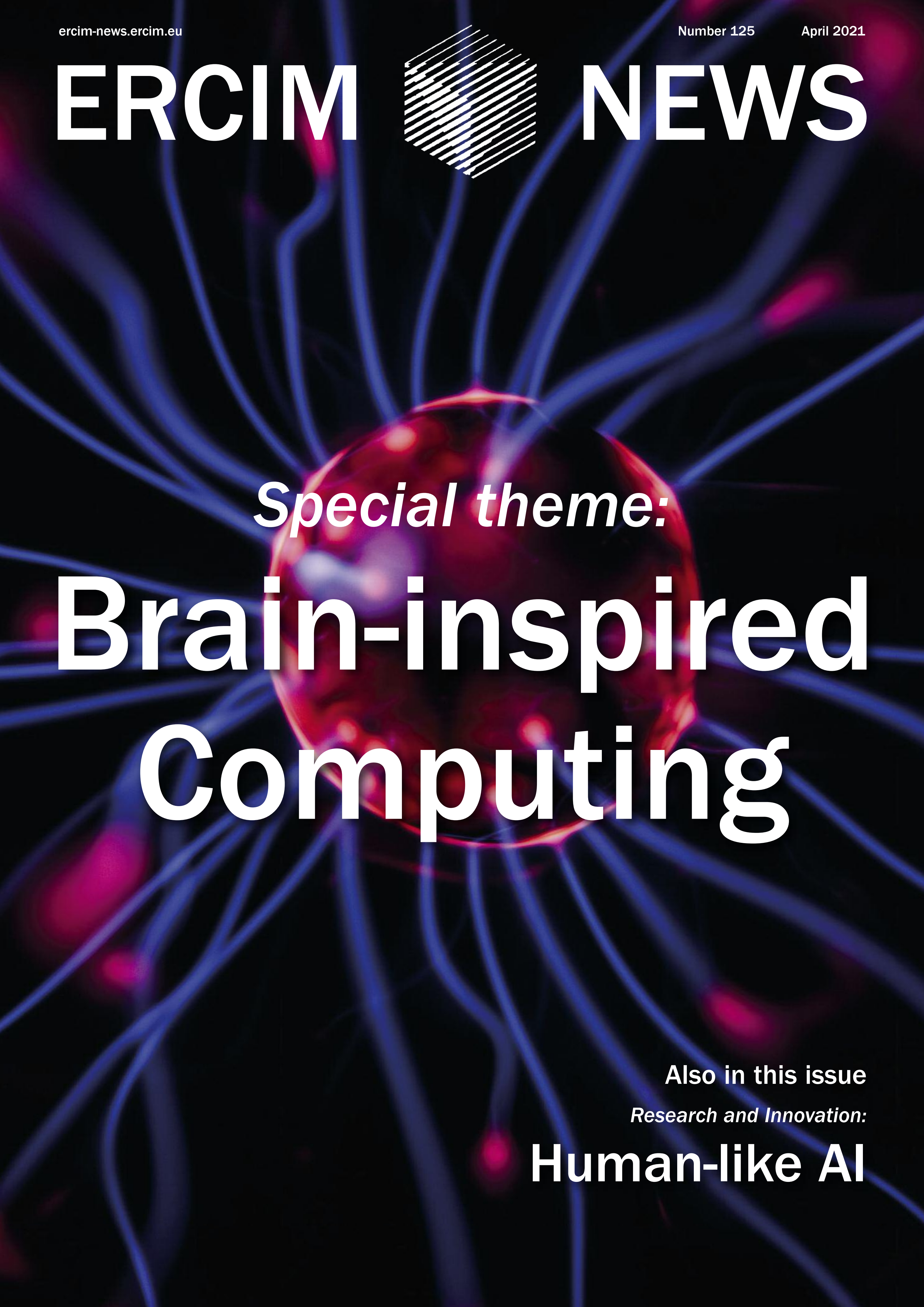
ercim-news.ercim.eu

Number 125 April 2021

ERCIM NEWS



*Special theme:*

# Brain-inspired Computing

Also in this issue

*Research and Innovation:*

## Human-like AI

## Editorial Information

*ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 6,000 printed copies and is also available online.*

*ERCIM News is published by ERCIM EEIG*
*BP 93, F-06902 Sophia Antipolis Cedex, France*
*+33 4 9238 5010, contact@ercim.eu*
*Director: Philipp Hoschka, ISSN 0926-4981*

***Contributions***
*Contributions should be submitted to the local editor of your country*

***ERCIM News online edition:*** *ercim-news.ercim.eu/*

***Next issue:***
*July 2021: Privacy-preserving Computation*

***Subscription***
*Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu*

*Cover photo by Josh Riemer on Unsplash.*

## JOINT ERCIM ACTIONS

## SPECIAL THEME

The special theme “Brain-inspired Computing” has been coordinated by the guest editors Robert Haas (IBM Research Europe) and Michael Pfeiffer (Bosch Center for Artificial Intelligence)

Introduction to the Special Theme

## RESEARCH AND INNOVATION

## ANNOUNCEMENTS

# ERCIM-JST Joint Symposium on Big Data and Artificial Intelligence

ERCIM and Japan Science and Technology Agency (JST) have organised the first joint symposium on Big Data and Artificial Intelligence, held on 18 and 19 February 2021. The online symposium aimed to present recent results of research conducted in the context of the JST CREST programme, as well as relevant research results from European institutions. The meeting provided the opportunity to the participants from Japan and Europe to familiarize themselves with recent research results and consider collaboration prospects that will arise in the context of the Horizon Europe framework programme or relevant initiatives from JST. Approximatively 80 invited participants attended the event.

Ms. Hiroko Tatesawa, Manager, ICT Group, Department of Strategic Basic Research, JST, introduced JST. JST, as one of the national Japanese funding agencies has about 1200 employees and an annual budget of 122.4 billion Yen (about 950 million Euro), of which about two thirds are allocated to funding programmes including strategic basic research, international collaborations, industry-acedemia collaboration and technology transfer. ERCIM President Dr. Björn Levin (RISE), then gave an overview on the activities of ERCIM.

With nine short talks, JST delivered an overview of CREST Program on Big Data Applications:

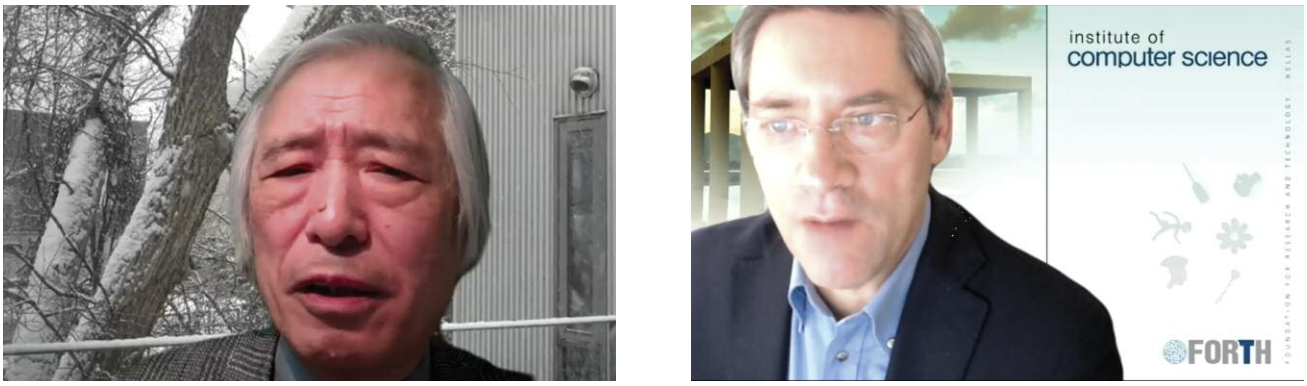
- Development of a knowledge-generating platform driven by big data in drug discovery through production processes by Dr. Makoto Taiji, (on Account for Kimito Funatsu (PI))
- Big data assimilation and AI: Creating new development in real-time weather prediction by Dr. Takemasa Miyoshi, RIKEN Center for Computational Science
- Establishing the advanced disaster reduction management system by fusion of real-time disaster simulation and big data assimilation by Prof. Shunichi Koshimura, International Research Institute of Disaster Science,Tohoku Univ.
- Exploring etiologies, sub-classification, and risk prediction of diseases based on big-data analysis of clinical and whole omics data in medicine, by Prof. Tatsuhiko Tsunoda, Graduate School of Science, Univ. of Tokyo
- Early detection and forecasting of pandemics using large-scale biological data: designing intervention strategy by Prof. Kimihito Ito (on Account for Hiroshi Nishiura (PI)), Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido Univ.
- Statistical computational cosmology with big astronomical data by Prof. Naoki Yoshida, Department of Physics/Kavli IPMU, Univ. of Tokyo
- Data-driven analysis of the mechanism of animal development by Shuichi Onami, RIKEN Center for Biosystems Dynamics Research
- Knowledge Discovery by Constructing AgriBigData by Masayuki Hirafuji, Graduate School of Agricultural and Life Sciences, Univ. of Tokyo
- Scientific Paper Analysis: Knowledge Discovery through Structural Document Understanding by Yuji Matsumoto, RIKEN Center for Advanced Intelligence Project.

The CREST programme is one of JST's major undertakings for stimulating achievement in fundamental science fields.

From ERCIM, three 20 minutes presentations were given:

- Scalable computing infrastructures: Keeping up with data growth by Prof. Angelos Bilas, FORTH
- Security, Privacy and Explainability for ML Models by Prof. Andreas Rauber, SBA Research and TU Wien and by Dr. Rudolf Mayer, SBA Research
- IoT - Big Data, Privacy: Resolving the contradiction by Rigo Wenning, Legal counsel, W3C.

The first day concluded with a presentation "AI and HPC" given by Prof. Satoshi Matsuoka, Director, RIKEN Center for Computational Science. In his talk, Prof. Matsuoka presented "Fugaku", the largest and fastest exascale supercomputer, how it is built and used for applications and experiments. These include for example the exploration of new drug candidates for COVID-19, a digital twin of the entire smart city of Fugaku as a shareable smart city R&D environ-



*Prof. Yuzuru Tanaka, Hokkaido University (left) and Prof. Dimitris Plexousakis, FORTH-ICS and University of Crete. Screenshots taken during the seminar.*

ment, and unprecedented deep learning scalability for AI developments.

The second day started with three more 20 minutes presentations from European researchers:

- Big Data Trends and Machine Learning in Medicinal Chemistry by Prof. Jürgen Bajorath, Prof., Univ. Bonn
- Data security and privacy in emerging scenarios by Prof. Sara Foresti, Univ. Milano
- Using Narratives to Make Sense of Data by Dr. Carlo Meghini, CNR.

Dr. Katsumi Emura, Director of AIP Network Laboratory, JST gave at talk "Promoting innovation with advanced research ecosystem - Introduction of JST AIP Network Laboratory". The AIP Network Laboratory is a virtual laboratory with teams in the fields of AI, big data, IoT, and cyber security under the JST Strategic Basic Research Program. This talk was followed by five short overviews of some other JST CREST Programs in the ICT area:

- Introduction for the research area of Core Technologies for Advanced Big Data Integration Program by Prof. Masaru Kitsuregawa, Director-General, National Institute of Informatics, Univ. of Tokyo
- A Future Society Allowing Human-Machine Harmonious Collaboration by Prof. Norihiro Hagita, Art Science Department, Osaka University of Arts
- Symbiotic Interaction Research toward Society of Human and AI by Prof. Kenji Mase Graduate School of Informatics, Nagoya Univ.
- Towards the Next Generation IoT Systems by Prof. Emeritus Hideyuki Tokuda, President, National Institute of Information and Communications Technology (NICT)
- Fundamental Information Technologies toward Innovative Social System Design by Prof. Sadao Kurohashi, Graduate School of Informatics. Kyoto Univ.

Prof. Masashi Sugiyama, Director, RIKEN Center for Advanced Intelligent Project, Univ. of Tokyo then gave a keynote talk entitled "Fundamental Machine Learning Research Opens Up New AI Society". He presented the RIKEN Center for Advanced Intelligence Project and machine learning as the core research challenge of current AI.

The last presentation of the symposium was given by Dr. Fosca Giannotti, CNR on "Explainable Machine Learning for Trustworthy AI".

The symposium was organised and chaired by Prof. emeritus Yuzuru Tanaka, Hokkaido University, Prof. emeritus Nicolas Spyratos, Université Paris - Saclay, and Prof. Dimitris Plexousakis, FORTH-ICS and University of Crete. They closed the symposium with the wish to continue the cooperation between ERCIM and JST.

**More information:**
Programme and links to slides and recorded presentations:
https://www.ercim.eu/events/ercim-jst-joint-symposium

**Please contact:**
Prof. Dimitris Plexousakis, Director FORTH-ICS, Greece

# ERCIM "Alain Bensoussan" Fellowship Programme

***The ERCIM PhD Fellowship Programme has been established as one of the premier activities of ERCIM. The programme is open to young researchers from all over the world. It focuses on a broad range of fields in Computer Science and Applied Mathematics.***

The fellowship scheme also helps young scientists to improve their knowledge of European research structures and networks and to gain more insight into the working conditions of leading European research institutions.

The fellowships are of 12 months duration (with a possible extension), spent in one of the ERCIM member institutes. Fellows can apply for second year in a different institute.

"The fellowship provided me a great platform to collaborate internationally with the proficient researchers working in my domain and sharpen my research skills. It has broadened my career options both in academia and industry and augmented the chances of getting my dream jobs.

Nancy AGARWAL
Former ERCIM Fellow

### Conditions

Candidates must:

- have obtained a PhD degree during the last eight years (prior to the year of the application deadline) or be in the last year of the thesis work with an outstanding academic record;
- be fluent in English;
- have completed their PhD before starting the grant;
- submit the following required documents: cv, list of publications, two scientific papers in English, contact details of two referees.

The fellows are appointed either by a stipend (an agreement for a research training programme) or a working contract. The type of contract and the monthly allowance/salary depends on the hosting institute.

### Application deadlines

Deadlines for applications are currently 30 April and 30 September each year.

Since its inception in 1991, over 500 fellows have passed through the programme. In 2020, 22 young scientists commenced an fellowship and 77 fellows have been hosted during the year. Since 2005, the programme is named in honour of Alain Bensoussan, former president of Inria, one of the three ERCIM founding institutes.

http://fellowship.ercim.eu

Introduction to the Special Theme

# Brain-inspired Computing –

by Robert Haas (IBM Research Europe) and Michael Pfeiffer (Bosch Center for Artificial Intelligence)

The origins of artificial intelligence (AI) can be traced back to the desire to build thinking machines, or electronic brains. A defining moment occurred back in 1958, when Frank Rosenblatt created the first artificial neuron that could learn by iteratively strengthening the weights of the most relevant inputs and decreasing others to achieve a desired output. The IBM 704, a computer the size of a room, was fed a series of punch cards and after 50 trials it learnt to distinguish cards marked on the left from cards marked on the right. This was the demonstration of the single-layer perceptron, or, according to its creator, "the first machine capable of having an original idea". [1]

Computation in brains and the creation of intelligent systems have been studied in a symbiotic fashion for many decades. This special theme highlights this enduring relationship, with contributions from some of the region's leading academic and industrial research laboratories. Ongoing collaboration between neuroscientists, cognitive scientists, AI researchers, and experts in disruptive computing hardware technologies and materials has made Europe a hotspot of brain-inspired computing research. The progress has been accelerated by EU-funded activities in the Future and Emerging Technologies (FET) programme, and more recently in the Human Brain Project. Over the last decade, similar large-scale interdisciplinary projects have started in the rest of the world, such as the BRAIN initiative in the US, and national initiatives in China, Canada and Australia. Brain-inspired computing projects from large industrial players such as IBM, Intel, Samsung and Huawei have focused on disruptive hardware technologies and systems. In technology roadmaps, brain-inspired computing is commonly seen as a future key enabler for AI on the edge.

Artificial neural networks (ANNs) have only vague similarities with biological neurons, which sparingly communicate using binary spikes at a low firing rate. However, compared with conventional ANNs, spiking neural networks (SNNs) require further innovation to be able to classify data accurately. Wozniak et al. (page 8) present the spiking neural unit (SNU), a model that can be introduced into deep learning architectures and trained to a high accuracy over a broad range of applications, including music and weather prediction. Yin et al. (page 9) show another application of backpropagation through time for tasks such ECG analysis and speech recognition, with a novel type of adaptive spiking recurrent neural network (SRNN) that achieves matching accuracy and an estimated 30 to 150-fold energy improvement over conventional RNNs. Masquelier (page 11) also exploits backpropagation through time, but accounts for the timing of spikes too, achieving comparable classification accuracy of speech commands compared with conventional deep neural networks.

SpiNNaker, described by Furber (page 12), and BrainScaleS, outlined by Schemmel (page 14), are the two neuromorphic computing systems that are being developed within the Human Brain Project. With a million cores representing neurons interconnected by a novel fabric optimised to transmit spikes, SpiNNaker allowed the first real-time simulation of a model of the early sensory cortex, estimated to be two to three times faster than when using GPUs or HPC systems. BrainScaleS introduced novel chips of analogue hardware circuits to represent each neuron, which are then configured in spiking neural networks, performing over 1000 times faster than real time, and leverage general-purpose cores closest to the analogue cores to control the plasticity of each neuron's synapses. This enables experimentation with various learning methods: For instance, Baumbach et al. (page 15) describe use cases that exploit these cores and simulate the environment of a bee looking for food, or perform reinforcement learning for the Pong game. In Göltz et al. (page 17), the timing of spikes is exploited, with an encoding representing more prominent features with earlier spikes, and training of the synaptic plasticity implemented by error backpropagation of first-time spikes.

How can modern AI benefit from neuroscience and cognitive science research? Alexandre et al. (page 19) present their interdisciplinary approach towards transferring neuroscientific findings to new models of AI. Two articles demonstrate how both hardware- and data-efficiency can be increased by following brain-inspired self-organisation principles. The SOMA project, presented by Girau et al. (page 20), applies both structural and synaptic neural plasticity principles to 3D cellular FPGA platforms. The multi-modal Reentrant Self-Organizing Map (ReSOM) model presented by Khacef et al. (page 22) highlights new opportunities to reduce the need for high volumes of labelled data by exploiting multi-modal associations. Ahmad et al. (page 24) introduce more plausible approaches towards plasticity to replace the well-known, but biologically unrealistic, backpropagation algorithm used for training deep neural networks. Nagy et al. (page 25) use results from human memory experiments to inform a new semantic compression technique for images, which captures the gist of visual memories. The highly efficient processing of the visual system is used as inspiration for a novel image and video compression technique by Doutsi et al. (page 27), exploiting the temporal dynamics of spiking neurons to detect characteristics of visual scenes.

A promising approach to brain-like computation, which could be applied to machine learning and robotics, is computing with very high-dimensional, unreliable, highly efficient, and widely distributed neuronal representations. Rahimi et al. (page 28) present an implementation of hyperdimensional computing on integrated memristive crossbars, showing how this computational paradigm is particularly well-suited for memristive hardware. Other examples of hardware–software co-design are described by Breiling et al. (page 30), who present the results of a national competition in Germany, in which entrants developed energy-efficient AI hardware for ECG classification.

Brains naturally combine signals from different modalities, which can help develop better sensor fusion algorithms for artificial autonomous systems. A neuromorphic implementation of a visual-auditory integration model using bio-inspired sensors for accurate and stable localisation is presented by Oess and Neumann (page 32). Similarly, Janotte et al. (page 34) describe a neuromorphic approach to both sensing and local processing of tactile signals. In their approach, event-driven sensors in the electronic skin encode tactile stimuli, and spiking neural networks learn to extract information that can be used for real time, low power control of robots or prostheses.

Computational neuroscience is an important source of information for brain-inspired computing. When trying to model the vast repertoire of learning rules at play at any time in different parts of the brain, manual modeling of plasticity rules beyond Hebbian learning and STDP becomes tedious. Mettler et al. (page 35) demonstrate how new mechanistic principles and rules for plasticity can be discovered by evolutionary algorithms, using an evolving-to-learn approach. A new mechanistic model of the cerebellar Purkinje neuron, by Nilsson and Jörntell (page 37), is matched to in-vivo recordings of spike communications and provides a surprisingly simple explanation as to what such cells actually do.

Europe is currently a leader in advancing neuromorphic computing technologies, but there is still a long road ahead to translate the research and technology development results into novel products. The NEUROTECH project, presented by Payvand et al. (page 39), has been very successful in building a community for European researchers in the field of neuromorphic computing. Among the highlights are monthly educational events on core themes such as event-driven sensing, and circuits for processing and learning. Projects such as NeuroAgents, described by Indiveri (page 40), aim to develop autonomous intelligent agents that show cognitive abilities on robotic platforms, entirely based on brain-inspired algorithms, sensors, and processors.

Although AI-based systems have reached performance levels beyond human capabilities in many specialised domains, such as image classification and game playing, it's important to note that there is still a long way to go for us to reach the flexibility and efficiency of biological brains in real-world domains. Building increasingly brain-like AI also requires re-thinking fundamental principles of computation: the classical von-Neumann model with its separation of computation and memory is not an efficient way to implement large scale biologically inspired neural networks. Recent progress in neuromorphic computing platforms and in-memory computing provides new opportunities for fast, energy-efficient, massively parallel, and continuously learning large-scale brain-inspired computing systems based on spiking neural networks. Furthermore, envisioning more advanced capabilities, such as making analogies and reasoning on problems is another step that will hinge on an interplay of ANNs and symbolic AI, a recent and very active field of research.

**Reference:**

[1] Melanie Lefkowitz, “Professor´s perceptron paved the way for AI – 60 years too soon”, Cornell Chronicle, https://kwz.me/h5j

**Please contact:**

Robert Haas
IBM Research Europe, Zurich
Research Laboratory, Switzerland
rha@zurich.ibm.com

Michael Pfeiffer
Bosch Center for Artificial
Intelligence, Renningen, Germany
michael.pfeiffer3@de.bosch.com

# Biologically Inspired Dynamics Enhance Deep Learning

by Stanisław Woźniak, Thomas Bohnstingl and Angeliki Pantazi (IBM Research – Europe)

***Deep learning has achieved outstanding success in several artificial intelligence (AI) tasks, resulting in human-like performance, albeit at a much higher power than the ~20 watts required by the human brain. We have developed an approach that incorporates biologically inspired neural dynamics into deep learning using a novel construct called spiking neural unit (SNU). Remarkably, these biological insights enabled SNU-based deep learning to even surpass the state-of-the-art performance while simultaneously enhancing the energy-efficiency of AI hardware implementations.***

The state-of-the-art deep learning is based on artificial neural networks (ANNs) that only take inspiration from biology to a very limited extent – primarily from its networked structure. This has several drawbacks, especially in terms of power consumption and energy efficiency [1]. Our group at IBM Research – Europe has been studying more realistic models of neural dynamics in the brain, known as spiking neural networks (SNNs). Since human brains are still much more capable than modern deep learning systems, SNNs have been widely considered as the most promising contender for the next generation of neural networks [1]. However, modelling and training complex SNNs has remained a challenge, which has led to the conviction among many researchers that the performance of ANNs is in practice superior to that of SNNs.

We have developed a novel spiking neural unit (SNU) that unifies the biologically inspired dynamics of spiking neurons with the research advances of ANNs and deep learning [2, L1]. In particular, we demonstrated that the leaky integrate and fire (LIF) evolution of the membrane potential of a biological neuron (Figure 1a) can be modelled, maintaining the exact equivalence, through a specific combination of recurrent artificial neurons forming jointly an SNU (Figure 1b). Therefore, SNUs can be flexibly incorporated into deep learning architectures and trained to high accuracy using supervised methods developed for ANNs, such as the backpropagation through time algorithm. We have compared SNUs and common deep

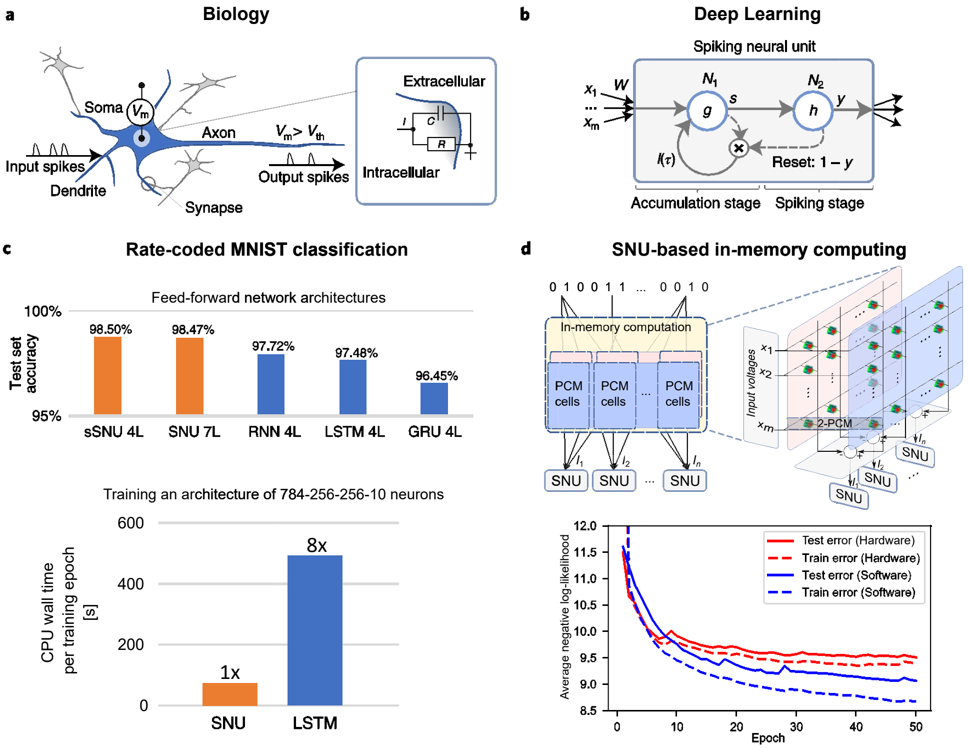


*Figure 1: Incorporating biologically inspired dynamics into deep learning: a. Biological neurons receive input spikes that are modulated by synaptic weights at the dendrites and accumulated into the membrane potential Vm in cell soma. This is typically modelled as a resistor-capacitor (RC) circuit. The output spikes are emitted through axons to downstream neurons. b. SNU models the spiking neural dynamics through two recurrent artificial neurons. N1 performs the accumulation and corresponds to the Vm. N2 controls the spike emission and resets the Vm. c. Digit classification for rate-coded MNIST dataset. In the upper part of the pane, feed-forward networks of common deep learning units are compared. Higher accuracy indicates improvement. In the lower part of the pane, the training time of SNU vs. LSTM is illustrated. d. The synaptic operations are accelerated in-memory through physical properties of the crossbar structure, illustrated in the upper part of the pane. We use two Phase Change Memory devices per synapse. The lower part of the pane contains a comparison of the average negative log-likelihood of software simulation vs. hardware experiment for the music prediction task using JSB dataset. Lower values correspond to higher-quality predictions. Figure adapted from [2].*

learning units, such as LSTMs and GRUs, on tasks including image classification (Figure 1c, upper part), language modelling, music prediction and weather prediction. Importantly, we demonstrated that although SNUs have the lowest number of parameters, they can surpass the accuracy of these units and provide a significant increase in speed (Figure 1c, lower part). Our results established the SNN state of the art with the best accuracy of 99.53% +/- 0.03%, for the handwritten digit classification task based on the rate-coded MNIST dataset using a convolutional neural network. Moreover, we even demonstrated the first-of-a-kind temporal generative adversarial network (GAN) based on an SNN.

Neural circuits in the human brain exhibit an additional set of functionalities observed on the neural dynamics level, where adaptive spiking thresholds play an important role as well as on the architectural level, where lateral inhibition between neurons is a commonly observed theme. Another advantage of the SNUs is that the developed framework can easily be extended to incorporate such neural functionalities. For example, we demonstrated that the LI-SNU variant, that implements the lateral inhibition, achieves digit classification performance that is on par with the original SNU while significantly reducing the required number of spikes.

The increasing adoption of ANNs has sparked the development of hardware accelerators to speed up the required computations. Because SNUs cast the dynamics of spiking neurons into the deep learning framework, they provide a systematic methodology for training SNNs using such AI accelerators. We demonstrated this with a highly efficient in-memory computing concept based on nanoscale phase-change memory devices (Figure 1d), where the spiking nature of SNNs leads to further energy savings [3]. Furthermore, we showed that SNNs are robust to operation with low-precision synaptic weights down to 4-bits of precision and can also cope with hardware imperfections such as noise and drift.

Our work on SNU has bridged the ANN and the SNN worlds by incorporating biologically inspired neural dynamics into deep learning, enabling to benefit from both worlds. SNU allows SNNs to take direct advantage of recent deep learning advances, which enable easy scaling up and training deep SNNs to a high degree of accuracy. From the ANN perspective, SNU implements novel temporal dynamics for machine learning applications with fewer parameters and potentially higher performance than the state-of-the-art units. Finally, hardware accelerators can also benefit from SNUs, as they allow for a highly efficient implementation and unlock the potential of neuromorphic hardware by training deep SNNs to high accuracy. Our team is continuing the work towards developing the biologically inspired deep learning paradigm by also exploring ways of enhancing the learning algorithms with neuroscientific insights [4].

**Link:**
[L1] https://kwz.me/h5b

**References:**

[1] W. Maass: “Networks of spiking neurons: The third generation of neural network models,” Neural Networks, vol. 10, no. 9, pp. 1659–1671, 1997.
[2] S. Woźniak, et al.: “Deep learning incorporating biologically inspired neural dynamics and in-memory computing,” Nat Mach Intell, vol. 2, no. 6, pp. 325–336, Jun. 2020.
[3] A. Pantazi, et al.: “All-memristive neuromorphic computing with level-tuned neurons,” Nanotechnology, vol. 27, no. 35, p. 355205, 2016.
[4] T. Bohnstingl, et al.:, “Online Spatio-Temporal Learning in Deep Neural Networks,” https://arxiv.org/abs/2007.12723

**Please contact:**
Stanisław Woźniak, IBM Research – Europe, Switzerland
stw@zurich.ibm.com

# Effective and Efficient Spiking Recurrent Neural Networks

by Bojian Yin (CWI), Federico Corradi (IMEC Holst Centre) and Sander Bohté (CWI)

***Although inspired by biological brains, the neural networks that are the foundation of modern artificial intelligence (AI) use exponentially more energy than their counterparts in nature, and many local “edge” applications are energy constrained. New learning frameworks and more detailed, spiking neural models can be used to train high-performance spiking neural networks (SNNs) for significant and complex tasks, like speech recognition and EGC-analysis, where the spiking neurons communicate only sparingly. Theoretically, these networks outperform the energy efficiency of comparable classical artificial neural networks by two or three orders of magnitude.***

Modern deep neural networks have only vague similarities with the functioning of biological neurons in animals. As illustrated in Figure 1 (left), many details of biological neurons are abstracted in this model of neural computation, such as the spatial extent of real neurons, the variable nature and effect of real connections formed by synapses, and the means of communication: real neurons communicate not with analogue values, but with isomorphic pulses, or spikes, and they do so only sparingly.

As far back as 1997, Maass argued that mathematically, networks of spiking neurons could be constructed that would be at least as powerful as similar sized networks of standard analogue artificial neurons [1]. Additionally, the sparse nature of spike-based communication is likely key to the energy-efficiency of biological neuronal networks, an increasingly desirable property as artifi-

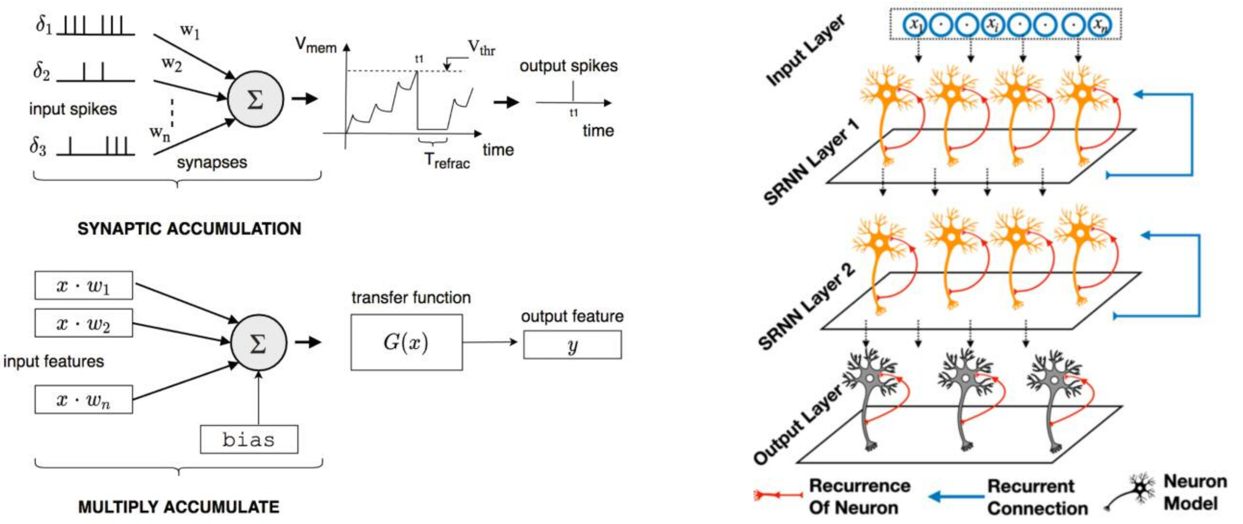


*Figure 1: Left: operation of spiking neurons compared to analogue neural units. Neurons communicate with spikes. Each input spike adds a weighted (and decaying) contribution to the internal state of the targeted neuron. When this state exceeds some threshold from below, a spike is emitted, and the internal state is reset. In the analogue neural unit, inputs x are weighted with corresponding weights w to add to the neuron's internal state, from which the neuron's output G(x) is computed. Right: illustration of a spiking recurrent neural network. Red connections denote the effective self-recurrence of spiking neurons due to the dynamic internal state.*

cial intelligence (AI) energy consumption is rapidly escalating and many edge-AI applications are energy-constrained. Until recently, however, the accuracy of spiking neural networks (SNNs) was poor compared to the performance of modern deep neural networks.

A principal obstacle for designing effective learning rules in SNNs has always been the discontinuous nature of the spiking mechanism: the workhorse of deep learning is error-backpropagation, which requires the estimation of the local gradient between cause and effect. For spiking neurons, this requires, for example, estimating how changing an input weight into a neuron affects the emission of spikes. This, however, is discontinuous: the spiking mechanism is triggered when the input into the neuron exceeds an internal threshold, and a small change in weight may be enough to just trigger a spike or prevent a spike from being emitted. To overcome the problem of learning with such discontinuous gradients, recent work [2] demonstrated a generic approach by using a "surrogate gradient". Importantly, the surrogate gradient approach enables the use of modern deep learning software frameworks like PyTorch and Tensorflow, including their efficient optimisers and GPU-support.

In joint work from CWI and IMEC [3], we demonstrate how surrogate gradients can be combined with efficient spiking neuron models in recurrent spiking neural networks (SRNNs) to train these networks to solve hard benchmarks using Back-Propagation-Through-Time: the SRNN is illustrated in Figure 1 (right). In particular, we showed how an adaptive version of the classical "leaky-integrate-and-fire" spiking neuron, the Adaptive LIF (ALIF) neuron, enables the networks to achieve high performance by co-training the internal parameters of the spiking neuron model that determine its temporal dynamics (the membrane-potential time-constant and the adaptation time-constant) together with the weights in the networks.

With these adaptive SRNNs, we achieve new state-of-the-art performance for SNNs on several tasks of inherent temporal nature, like ECG analysis and speech recognition (SHD), all while demonstrating highly sparse communication. Their performance also compared favourably to standard recurrent neural networks like LSTMs and approaches the current state of the art in deep learning. When comparing computational cost as a proxy for energy consumption, we calculated the required Multiply-Accumulate (MAC) and Accumulate (AC) operations for various networks. With a MAC being about 31x more energetically expensive compared to an AC, and with spiking neurons using ACs sparingly where MACs are used in standard neural networks, we demonstrate theoretical energy advantages for the SRNNs ranging from 30 to 150 for comparable accuracy (for more details, see [3]).

We believe this work demonstrates that SNNs can be compelling propositions for many edge-AI applications in the form of neuromorphic computing. At the same time, tunable spiking neuron parameters proved essential for achieving such accuracy, suggesting both novel venues for improving SNNs further and neuroscientific investigations into corresponding biological plasticity.

**Links:**

Project: NWO TTW programme "Efficient Deep Learning", Project 7, https://efficientdeeplearning.nl/
Pre-print: https://kwz.me/h5t
Code: https://kwz.me/h5d

**References:**

[1] W. Maass: "Fast sigmoidal networks via spiking neurons. Neural Computation", 9(2), 279-304, 1997.
[2] E.O.Neftci, H. Mostafa, F. Zenke: "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks", IEEE Signal Processing Magazine, 36(6), 51-63, 2019.
[3] B. Yin, F. Corradi, S.M. Bohté: "Effective and efficient computation with multiple-timescale spiking recurrent neural networks", in Int. Conf. on Neuromorphic Systems 2020 (pp. 1-8), 2020; Arxiv 2005.11633v1

**Please contact:**

Sander Bohte, CWI, The Netherlands
S.M.Bohte@cwi.nl

# Back-propagation Now Works in Spiking Neural Networks!

by Timothée Masquelier (UMR5549 CNRS – Université Toulouse 3)

***Back-propagation is THE learning algorithm behind the deep learning revolution. Until recently, it was not possible to use it in spiking neural networks (SNN), due to non-differentiability issues. But these issues can now be circumvented, signalling a new era for SNNs.***

Biological neurons use short electrical impulses called "spikes" to transmit information. The spike times, in addition to the spike rates, are known to play an important role in how neurons process information. Spiking neural networks (SNNs) are thus more biologically realistic than the artificial neural networks (ANNs) used in deep learning, and are arguably the most viable option if one wants to understand how the brain computes at the neuronal description level. But SNNs are also appealing for AI, especially for edge computing, since they are far less energy hungry than ANNs. Yet until recently, training SNNs with back-propagation (BP) was not possible, and this has been a major impediment to the use of SNNs.

Back-propagation (BP) is the main supervised learning algorithm in ANNs. Supervised learning works with examples for which the ground truth, or "label", is known, which defines the desired output of the network. The error, i.e., the distance between the actual and desired outputs, can be computed on these labelled examples. Gradient descent is used to find the parameters of the networks (e.g., the synaptic weights) that minimise this error. The strength of BP is to be able to compute the gradient of the error with respect to all the parameters in the intermediate "hidden" layers of the network, whereas the error is only measured in the output layer. This is done using a recurrent equation, which allows computation of the gradients in layer l-1 as a function of the gradients in layer l. The gradients in the output layer are straightforward to compute (since the error is measured there), and then the computation goes backward, until all gradients are known. BP thus solves the "credit assignment problem", i.e., it finds the optimal thing to do for the hidden layers. Since the number of layers is arbitrary, BP can work in very deep networks, which has led to the widely talked about deep learning revolution.

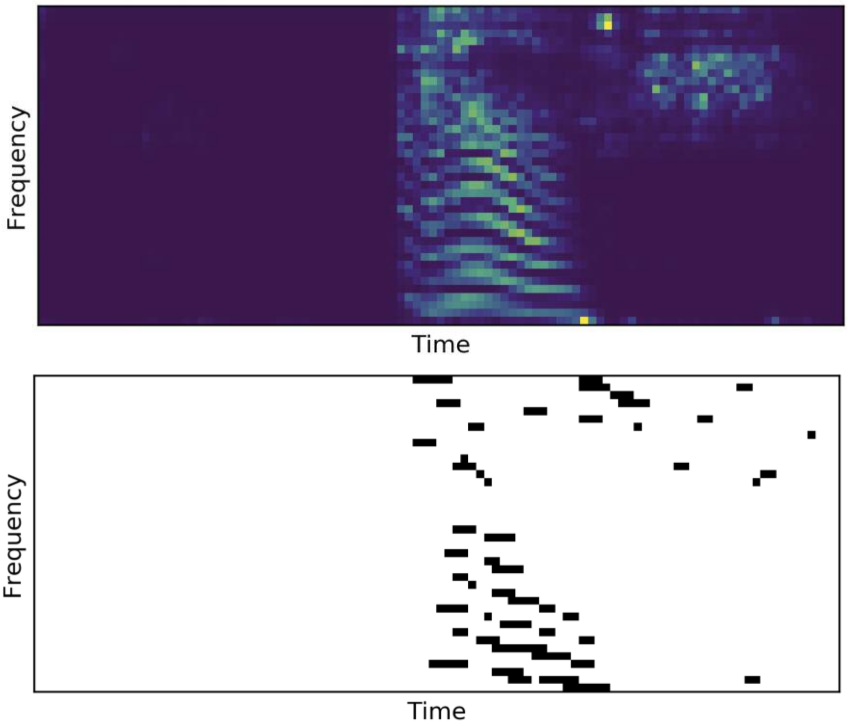


*Figure 1. (Top) Example of spectrogram (Mel filters) extracted for the word "off". (Bottom) Corresponding spike trains for one channel of the first layer.*

This has motivated us and others to train SNNs with BP. But unfortunately, it is not straightforward. To compute the gradients, BP requires differentiable activation functions, whereas spikes are "all-or-none" events, which cause discontinuities. Here we present two recent methods to circumvent this problem.

### S4NN: a latency-based backpropagation for static stimuli

The first method, S4NN, deals with static stimuli and rank-order-coding [1]. With this sort of coding, neurons can fire at most one spike: most activated neurons first, while less activated neurons fire later, or not at all. In particular, in the readout layer, the first neuron to fire determines the class of the stimulus. Each neuron has a single latency, and we demonstrated that the gradient of the loss with respect to this latency can be approximated, which allows estimation of the gradients of the loss with respect to all the weights, in a backward manner, akin to traditional BP. This approach reaches a good accuracy, although below the state-of-the-art: e.g., a test accuracy of 97.4% for the MNIST dataset. However, the neuron model we use, non-leaky integrate-and-fire, is simpler and more hardware friendly than the one used in all previous similar proposals.

### Surrogate Gradient Learning: a general approach

One of the main limitations of S4NN is the at-most-one-spike-per-neuron constraint. This constraint is acceptable for static stimuli (e.g., images), but not for those that are dynamic (e.g., videos, sounds): changes need to be encoded by additional spikes. Can BP still be used in this context? Yes, if the "surrogate gradient learning" (SGL) approach is used [2].

The most commonly used spiking neuron model, the leaky integrate-and-fire neuron, obeys a differential equation, which can be approximated using discrete time steps, leading to a recurrent relation for the potential. This relation can be computed using the recurrent neural network (RNN) formalism, and the training can be done using backpropagation through time, the reference algorithm for training RNNs. The firing threshold causes optimisation issues, but they can be overcome by using a "surrogate gradient". In short, in the forward pass of BP, the firing threshold is applied normally, using the Heaviside step function. But in the backward pass, we pretend that a sigmoid was used in the forward pass instead of the Heaviside function, and we use the derivative of this sigmoid for the gradient computation. This approximation works very well in practice. In addition, the training can be done using automatic-differentiation tools such as PyTorch or Tensorflow, which is very convenient.

We extended previous approaches by adding convolutional layers (see [3] for a similar approach). Convolutions can be done in space, time (which simulates conduction delays), or both. We validated our approach on a speech classification benchmark: the Google speech commands dataset. We managed to reach nearly state-of-the-art accuracy (94.5%) while maintaining low firing rates (about 5Hz, see Figure 1). Our study has just been accepted at the IEEE Spoken Language Technology Workshop [4]. Our code is based on PyTorch and is available in open source at [L1].

### Conclusion

We firmly believe that these results open a new era for SNNs, in which they will compete with conventional deep learning in terms of accuracy on challenging problems, while their implementation on neuromorphic chips could be much more efficient and use less power.

**Link:**
[L1] https://kwz.me/h5c

**References:**

[1] S. R. Kheradpisheh and T. Masquelier, "Temporal Backpropagation for Spiking Neural Networks with One Spike per Neuron," Int. J. Neural Syst., vol. 30, no. 06, p. 2050027, Jun. 2020.
[2] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate Gradient Learning in Spiking Neural Networks," IEEE Signal Process. Mag., vol. 36, no. October, pp. 51–63, 2019.
[3] S. Woźniak, A. Pantazi, T. Bohnstingl, and E. Eleftheriou, "Deep learning incorporating biologically inspired neural dynamics and in-memory computing," Nat. Mach. Intell., vol. 2, no. 6, pp. 325–336, 2020.
[4] T. Pellegrini, R. Zimmer, and T. Masquelier, "Low-activity supervised convolutional spiking neural networks applied to speech commands recognition," in IEEE Spoken Language Technology Workshop, 2021.

**Please contact:**

Timothée Masquelier
Centre de Recherche Cerveau et Cognition, UMR5549 CNRS – Université Toulouse 3, France
timothee.masquelier@cnrs.fr

# Building Brains

by Steve Furber (The University of Manchester)

***SpiNNaker – a Spiking Neural Network Architecture – is the world's largest neuromorphic computing system. The machine incorporates over a million ARM processors that are connected through a novel communications fabric to support large-scale models of brain regions that operate in biological real time, with the goal of contributing to the scientific Grand Challenge to understand the principles of operation of the brain as an information processing system.***

SpiNNaker is one of two neuromorphic computing systems offering an open service under the auspices of the EU Flagship Human Brain Project (HBP) as part of the EBRAINS research infrastructure [L1] – the other being the BrainScaleS machine developed by the University of Heidelberg in Germany. The machine has been 20 years in conception and 15 years in construction at the University of Manchester, UK, and has been supporting a service with a half-million core system since April 2016, upgraded to the full million cores since November 2018. Around 100 additional small SpiNNaker systems have been deployed in research labs around the world.

In many ways SpiNNaker resembles a massively-parallel high-performance computer (HPC), with two exceptions: (i) the processors used on SpiNNaker are small, efficient cores intended for embedded applications rather than the high-end powerful numerical processors used in HPC, and (ii) the communications fabric on SpiNNaker is optimised for carrying large numbers of small packets (where each packet conveys a neuron "spike") to many destinations, whereas HPC systems are optimised for large, fast data movements between two cores. The SpiNNaker communications architecture is very much motivated by the very high degree of connectivity found between neurons in the brain, where each neuron typically connects to many thousands of other neurons.

User access to SpiNNaker is typically mediated through the PyNN (Python Neural Networks) language, which is supported by a number of simulators, so models can be developed on any computer or laptop and then run on SpiNNaker, either through a web-based batch mode interface or using a Jupyter notebook. For real-time robot control it is generally necessary to have a SpiNNaker system collocated with the robot, though the host server in Manchester supports the HBP

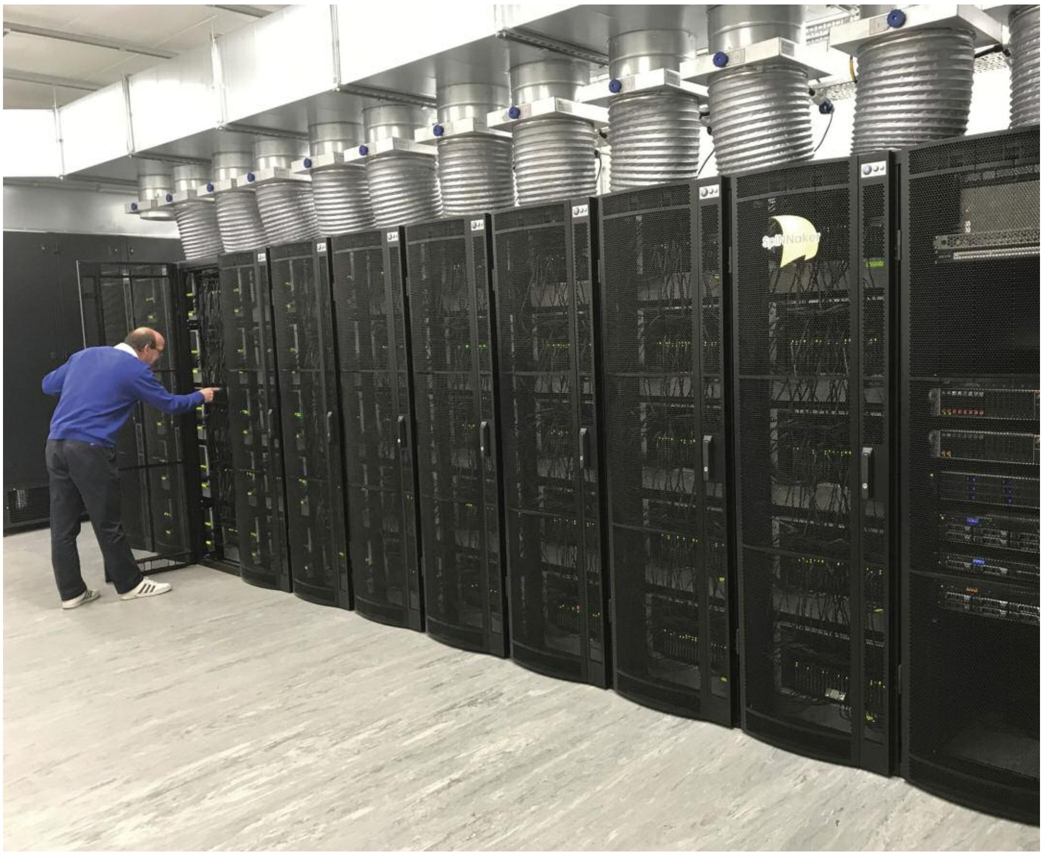


*Figure 1: The million-core SpiNNaker machine at Manchester. The machine occupies 10 rack cabinets, each with 120 SpiNNaker boards, power supplies, cooling and network switches. The 11th cabinet on the right contains servers that support configuration software and remote access to the machine.*

Neurorobotics platform for virtual neurorobotic control simulation.

Examples of brain models successfully run on the machine include a cortical microcircuit [1], where SpiNNaker was the first machine to achieve real time, and a hybrid model of the auditory system [2]. Further models are being developed in collaboration with partners within the HBP.

Even with a million processor cores, SpiNNaker cannot model systems approaching the scale of the full human brain with its (just under) one hundred billion neurons and several hundred trillion synapses, but it could potentially support models of the scale of the mouse brain, which is about a thousand times smaller than the human brain. However, biological data is relatively sparse, and putting together a full brain model even of an insect’s brain, which is a hundred to a thousand times smaller than a mouse brain, involves interpretation of, and extrapolation from, the available data, combined with more than a little guesswork! So progress towards a reasonably coherent description of information processing in the brain will be a long, slow haul, requiring diligent work from practical and theoretical neuroscientists, advances in brain imaging, and contributions from many other disciplines. SpiNNaker offers a platform for testing theories and hypotheses at scale as they emerge from such work, and also for investigating how concepts from brain science can be translated into advances in artificial intelligence with applications in the commercial domain.

The technology employed in the current SpiNNaker machine is now 10 years old, and its successor has been under development within the HBP as a collaboration between the University of Manchester and the Technical University of Dresden. SpiNNaker2 will be fabricated in 2021 and will deliver an order of magnitude increase in functional density and energy efficiency, taking SpiNNaker forward into its third decade, but with an increased emphasis on commercial applications alongside brain science.

The first 20 years of the SpiNNaker project have been documented in an Open Access book [3], including details of a number of applications and the current state of development of SpiNNaker2.

**Links:**

[L1] https://kwz.me/h5y
[L2] https://kwz.me/h5H

**References:**

[1] Oliver Rhodes et al, “Real-time cortical simulation on neuromorphic hardware”, Phil. Trans. R. Soc. A. 378:20190160. http://doi.org/10.1098/rsta.2019.0160

[2] Robert James et al, “Parallel Distribution of an Inner Hair Cell and Auditory Nerve Model for Real-Time Application”. IEEE Trans. Biomed. Circuits Syst. 12(5):1018-1026. doi: 10.1109/TBCAS.2018.2847562

[3] Steve Furber & Petrut Bogdan (eds.) (2020), “SpiNNaker: A Spiking Neural Network Architecture”, Boston-Delft: now publishers, http://dx.doi.org/10.1561/9781680836523

**Please contact:**

Steve Furber
The University of Manchester, UK
steve.furber@manchester.ac.uk

# The BrainScaleS Accelerated Analogue Neuromorphic Architecture

by Johannes Schemmel (Heidelberg University, Germany)

***By incorporating recent results from neuroscience into analogue neuromorphic hardware we can replicate biological learning mechanisms in silicon to advance bio-inspired artificial intelligence.***

Biology is still vastly ahead of our engineering approaches when we look at computer systems that cope with natural data. Machine learning and artificial intelligence increasingly use neural network technologies inspired by nature, but most of these concepts are founded on our understanding of biology from the 1950s. The BrainScaleS neuromorphic architecture builds upon contemporary biology to incorporate the knowledge of modern neuroscience within an electronic emulation of neurons and synapses [1].

Biological mechanisms are directly transferred to analogue electronic circuits, representing the signals in our brains as closely as possible with respect to the constraints of energy and area efficiency. Continuous signals like voltages and currents at cell membranes are directly represented as they evolve over time [2]. The communication between neurons can be modelled by either faithfully reproducing each action potential or by summarising them as firing rates, as is commonly done in contemporary artificial intelligence (AI).

Using the advantages of modern microelectronics, the circuits are operated at a much higher speed than the natural nervous system. Acceleration factors from 1000 to 100,000 have been achieved in previous implementations, thereby dramatically shortening execution times.

The BrainScaleS architecture consists of three pillars:
- microelectronics as the physical substrate for the emulation of neurons and synapses
- a software stack controlling the operation of a BrainScaleS neuromorphic system from the user interface down to the individual transistors
- training and learning algorithms supported by dedicated software and hardware throughout the system.

The analogue realisation of the network emulation does not only have the advantage of speed, but also consumes much less energy and is easier and therefore cheaper to fabricate than comparable digital implementations. It does not need the smallest possible transistor size to achieve a high energy efficiency. To configure the analogue circuits for their assigned tasks, they are accompanied by specialised microprocessors optimised for the efficient implementation of modern bio-inspired learning algorithms.

Suitable tasks range from fundamental neuroscientific research, like the validation of results from experimental observations in living tissue, to AI applications like data classification and pattern recognition.

These custom-developed plasticity processing units have access to a multitude of analogue variables characterising the state of the network. For example, the temporal correlation of action potentials is continuously measured by the analogue circuitry and can be used as input variables for learning algorithms [3]. This combination of analogue and digital hardware-based learning is called hybrid plasticity.

Figure 1 shows the recently finished BrainScaleS mobile system. A BrainScaleS ASIC is visible on the lower left side of the topmost circuit board. This version is suited for edge-based AI applications. Edge-based AI avoids the transfer of full sensor information to the data centre and thus enhances data privacy and has the potential to vastly reduce the energy consumption of the data centre and data communication infrastructure. An initial demonstration successfully detects signs of heart disease in EKG traces.

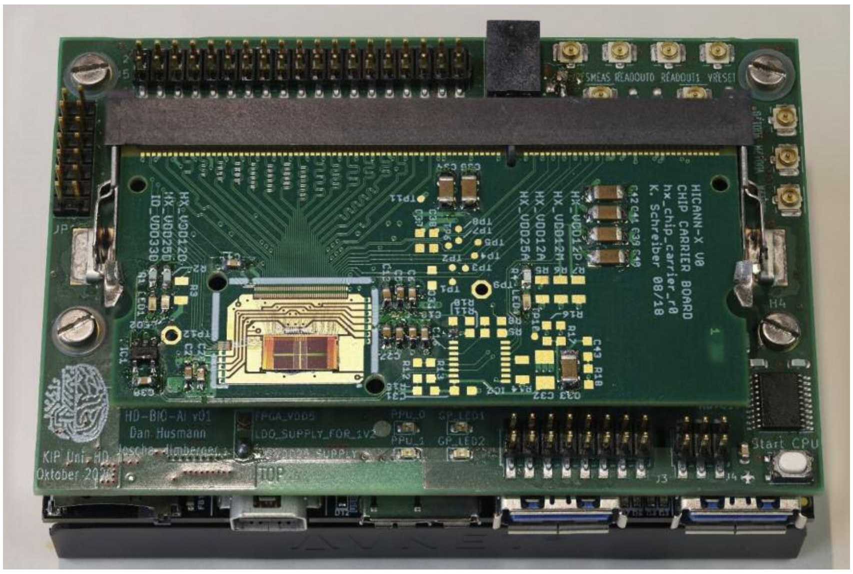


*Figure 1: The most recent addition to the BrainScaleS family of analogue accelerated neuromorphic hardware systems is a credit-card sized mobile platform for EdgeAI applications. The protective lid above the BrainScaleS ASIC has been removed for the photograph.*

The energy used to perform this operation is so low that a wearable medical device could work for years on battery power using the BrainScaleS ASIC.

The BrainScaleS architecture can be accessed and tested free of charge online at [L1].

This work has received funding by the EC (H2020) under grant agreements 785907 and 945539 (HBP) and the BMBF (HD-BIO-AI, 16ES1127)

**Link:** [L1] https://kwz.me/h5E

**References:**

[1] S. A. Aamir et al.: “A Mixed-Signal Structured AdEx Neuron for Accelerated Neuromorphic Cores,” IEEE Trans. Biomed. Circuits Syst., vol. 12, no. 5, pp. 1027–1037, Oct. 2018, doi: 10.1109/TBCAS.2018.2848203.

[2] S. A. Aamir, Y. Stradmann, and P. Müller: “An accelerated lif neuronal network array for a large-scale mixed-signal neuromorphic architecture,” on Circuits and …, 2018, [Online]. Available: https://kwz.me/h5K.

[3] S. Billaudelle et al.: “Structural plasticity on an accelerated analog neuromorphic hardware system,” Neural Netw., vol. 133, pp. 11–20, Oct. 2020, doi: 10.1016/j.neunet.2020.09.024.

**Please contact:**
Johannes Schemmel
Heidelberg University, Germany,
schemmel@kip.uni-heidelberg.de
Björn Kindler
Heidelberg University, Germany,
kindler@kip.uni-heidelberg.de

# BrainScaleS: Greater Versatility for Neuromorphic Emulation

by Andreas Baumbach (Heidelberg University, University of Bern), Sebastian Billaudelle (Heidelberg University), Virginie Sabado (University of Bern) and Mihai A. Petrovici (University of Bern, Heidelberg University)

***Uncovering the mechanics of neural computation in living organisms is increasingly shaping the development of brain-inspired algorithms and silicon circuits in the domain of artificial information processing. Researchers from the European Human Brain project employ the BrainScaleS spike-based neuromorphic platform to implement a variety of brain-inspired computational paradigms, from insect navigation to probabilistic generative models, demonstrating an unprecedented degree of versatility in mixed-signal neuromorphic substrates.***

Unlike their machine learning counterparts, biological neurons interact primarily via electrical action potentials, known as “spikes”. The second generation of the BrainScaleS neuromorphic system [1, L6] implements up to 512 such spiking neurons, which can be near-arbitrarily connected. In contrast to classical simulations, where the simulation time increases for larger systems, this form of in-silico implementation replicates the physics of neurons rather than numerically solving the associated equations, enabling what could be described as perfect scaling, with the duration of an emulation being essentially independent of the size of the model. Moreover, the electronic circuits are configured to be approximately 1000 times faster than their biological counterparts. It is this acceleration that makes BrainScaleS-2 a powerful platform for biological research and potential applications. Importantly, the chip was also designed for energy efficiency, with a total nominal power consumption of 1 W. Coupled with its emulation speed, this means energy consumption is several orders of magnitude lower than state-of-the art conventional simulations. These features, including the on-chip learning capabilities of the second generation of the BrainScaleS architecture, have the potential to enable the widespread use of spiking neurons beyond the realm of neuroscience and into artificial intelligence (AI) applications. Here we showcase the system’s capabilities with a collection of five highly diverse emulation scenarios, from a small model of insect navigation, including the sensory apparatus and the environment, to fully fledged discriminatory and generative models of higher brain functions [1].

As a first use case, we present a model of the bee’s brain that reproduces the ability of bees to return to their nest’s location after exploring their environment for food (Figure 1). The on-chip general-purpose processor of BrainScaleS is used to simulate the environment and to stimulate the motor neurons triggering the exploration of the environment. The network activity stores the insect’s position, enabling autonomous navigation back to its nest. The on-chip environment simulation avoids delays typically introduced by the use of coprocessors, thus fully exploiting the speed of the accelerated platform for the experiment. In total, the emulated insect performs about 5.5 biological hours’ worth of exploration in 20 seconds.

While accelerated model execution is certainly an attractive feature, it is often the learning process that is prohibitively time-consuming. Using its on-chip general-purpose processor, the BrainScaleS system offers the option of fully embedded learning, thus maximising the benefit of its accelerated neuro-synaptic dynamics. Using the dedicated circuitry for spike-timing-dependent plasticity (STDP), the system can, for example, implement a neuromorphic agent playing a version of the game Pong. The general-purpose processor again simulates the game dynamics and, depending on the game outcome, provides the success signal used in the reinforcement learning. Together with the analogue STDP measurements, this reward signal allows the emulated agent to learn autonomously. The resulting system is an order of magnitude faster and three orders of magnitude more

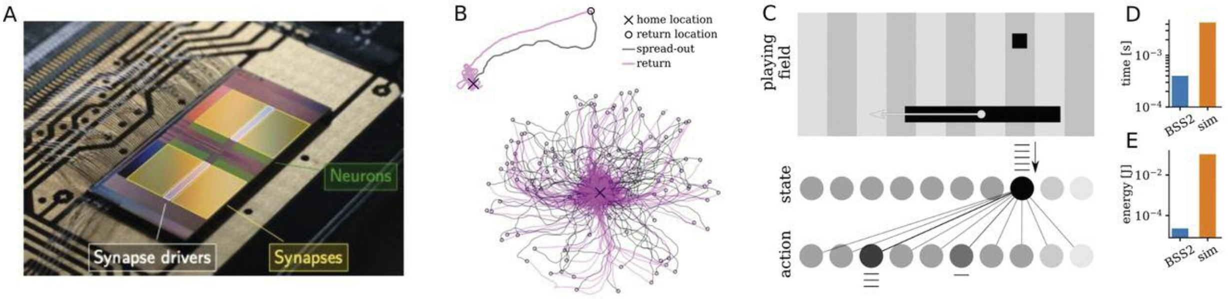


*Figure 1: (A) Photograph of the BrainScaleS chip with false-colour overlay of some of its components. (B) 100 paths travelled by an emulated insect swarming out from its nest, exploring the world (grey) and returning home (pink). (C) Emulated Pong agent and associated spiking network. Emulation time and energy consumption compared to state-of-the art simulations shown in (D) and (E).*

energy-efficient than an equivalent software simulation (Figure 1).

In a different example of embedded learning, a network is trained to discriminate between different species of iris flowers based on the shape of their petals using a series of receptors in a two-dimensional dataspace (Figure 2). For each neuron, the number of possible inputs is limited by the architecture and topology of the silicon substrate, much like the constraints imposed by biological tissue. Here too, the learning rule is based on spike-time similarities but enhanced by regularisation and pruning of weaker synaptic connections. On BrainScaleS, it is possible to route the output spikes of multiple neurons to each synaptic circuit. This allows the adaptation of the network connectome at runtime. In particular, synapses below a threshold strength are periodically removed and alternatives are instantiated. The label neurons on BrainScaleS then develop distinct receptive fields using the same circuitry. For this particular problem, it was possible to enforce a sparsity of nearly 90% without impacting the final classification performance (Figure 2).

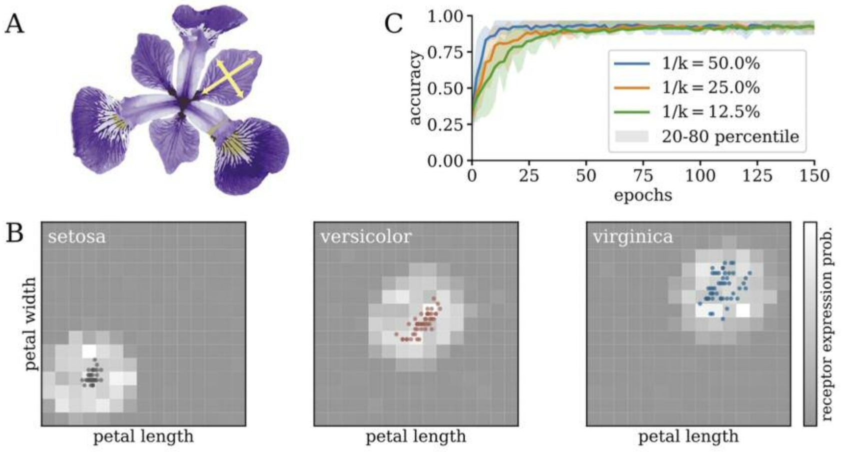


*Figure 2: (A) Different species of Iris can be distinguished by the shape of their flowers' petals. (B) Structural changes to the connectome of an emulated classifier organize appropriate receptive fields even for very sparse connectivity throughout the experiment. (C) Evolution of classification accuracy during training for different sparsity levels.*

The last two networks demonstrated in [1] hint at the more challenging nature of tasks that biological agents must solve in order to survive. One such task is being able to cope with incomplete or inaccurate sensory input. The Bayesian brain hypothesis posits that cortical activity instantiates probabilistic inference at multiple levels, thus providing a normative framework for how mammalian brains can perform this feat. As an exemplary implementation thereof, we showcase Bayesian inference in spiking networks on BrainScaleS (Figure 3). Similar systems have previously been used to generate and discriminate between hand-written digits or small-scale images of fashion articles [3, 4].

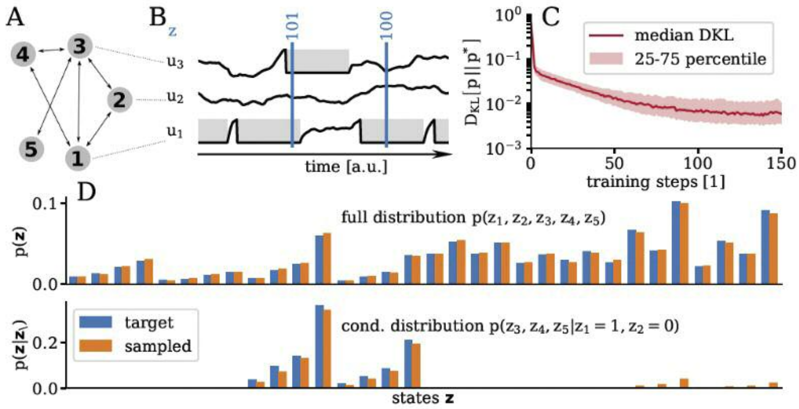


*Figure 3: Bayesian inference with spiking neural networks. (A, B) Neuronal spike patterns can be interpreted as binary vectors. (C) Neural dynamics can learn to sample from arbitrary probability distributions over binary spaces (D).*

The final network model was trained as a classifier for visual data using time-to-first-spike coding. This specific representation of information enabled it to classify up to 10,000 images per second using only 27 µJ per image. This particular network is described in more detail in its own dedicated article [L5].

These five networks embody a diverse set of computational paradigms and demonstrate the capabilities of the BrainScaleS architecture. Its dedicated spiking circuitry coupled with its versatile on-chip processing unit allows the emulation of a large variety of models in a particularly tight power and time envelope. As work continues on user-friendly access to its many features, BrainScaleS aims to support a wide range of users and use cases, from both the neuroscientific research community and the domain of industrial applications.

We would like to thank our collaborators Yannik Stradmann, Korbinian Schreiber, Benjamin Cramer, Dominik Dold, Julian Göltz, Akos F. Kungl, Timo C. Wunderlich, Andreas Hartel, Eric Müller, Oliver Breitwieser, Christian Mauch, Mitja Kleider, Andreas Grübl, David Stöckel, Christian Pehle, Arthur Heimbrecht, Philipp Spilger, Gerd Kiene, Vitali Karasenko, Walter Senn, Johannes Schemmel and Karlheinz Meier, as well as the Manfred Stärk foundation for ongoing support. This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).

**Links:**

[L1] https://youtu.be/_x3wqLFS278
[L2] https://kwz.me/h5S
[L3] https://kwz.me/h54
[L4] https://kwz.me/h50
[L5] https://kwz.me/h51
[L6] https://kwz.me/h52

**References:**

[1] S. Billaudelle, et al.: "Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate." 2020 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2020.

[2] T. Wunderlich, et al.: "Demonstrating advantages of neuromorphic computation: a pilot study," Frontiers in Neuroscience,vol. 13, p. 260, 2019.

[3] D. Dold, et al. "Stochasticity from function—why the bayesian brain may need no noise." Neural networks 119 (2019): 200-213.

[4] A. Kungl, F. Akos, et al.: "Accelerated physical emulation of Bayesian inference in spiking neural networks", Frontiers in neuroscience 13 (2019): 1201.

**Please contact:**

Andreas Baumbach, NeuroTMA group, Department of Physiology, University of Bern, Switzerland and Kirchhoff-Institute for Physics, Heidelberg University, Germany
andreas.baumbach@kip-uni-heidelberg.de

# Fast and Energy-efficient Deep Neuromorphic Learning

by Julian Göltz (Heidelberg University, University of Bern), Laura Kriener (University of Bern), Virginie Sabado (University of Bern) and Mihai A. Petrovici (University of Bern, Heidelberg University)

***Many neuromorphic platforms promise fast and energy-efficient emulation of spiking neural networks, but unlike artificial neural networks, spiking networks have lacked a powerful universal training algorithm for more challenging machine learning applications. Such a training scheme has recently been proposed and using it together with a biologically inspired form of information coding shows state-of-the-art results in terms of classification accuracy, speed and energy consumption.***

Spikes are the fundamental unit in which information is processed in mammalian brains, and a significant part of the information is encoded in the relative timing of these spikes. In contrast, the computational units of typical machine learning models output a numeric value without an accompanying time. This observation is, in a modified form, at the centre of a new approach: a network model and learning algorithm that can efficiently solve pattern recognition problems by making full use of the timing of spikes [1]. This quintessential reliance on spike-based communication perfectly synergises with efficient neuromorphic spiking-network emulators, such as the BrainScaleS-2 platform [2], thus being able to fully harness their speed and energy characteristics. This work is the result of a collaboration between neuromorphic engineers at the Heidelberg University and computational neuroscientists at the University of Bern, fostered by the European Human Brain Project.

In the implementation on BrainScaleS-2, to further enforce fast computation and to minimise resource requirements, an encoding was chosen where more prominent features are represented by earlier spikes, as seen in nature, e.g., in how nerves in the fingertips encode information about touch (Figure 1). From the perspective of an animal looking to survive, this choice of coding is particularly appealing, as actions must often be taken under time pressure. The biological imperative of short times-to-solution is similarly applicable to silicon, carrying with it an optimised usage of resources. In this model of neural computation, synaptic plasticity implements a version of error backpropagation on first-spike times, which we discuss in more detail below.

This algorithm was demonstrated on the BrainScaleS-2 neuromorphic platform using both an artificial dataset resembling the Yin-Yang symbol [3], as well as the real-world MNIST data set of handwritten digits. The Yin-Yang dataset highlights the universality of the algorithm and its interplay with first-spike coding in a small network, ensuring that training of this classification problem achieves highly accurate results (Figure 2). For the digit-recognition problem, an optimised implementation yields particularly compelling results: up to 10,000 images can be classified in less than a second at a runtime power consumption of only 270 mW, which translates to only 27 µJ per

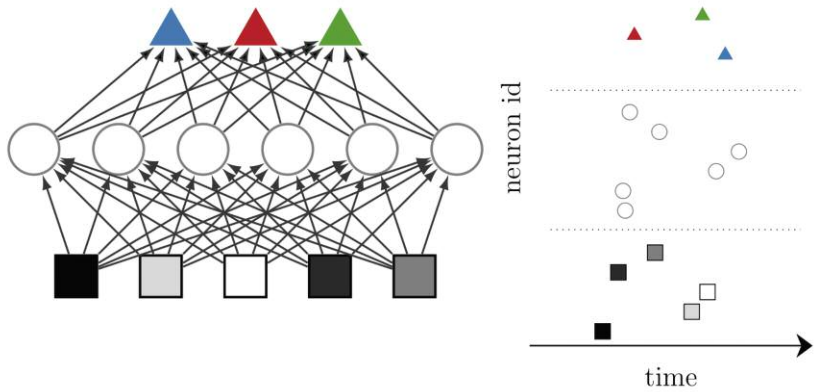


*Figure 1: (Left) Discriminator network consisting of neurons (squares, circles, and triangles) grouped in layers. Information is passed from the bottom to the top, e.g. pixel brightness of an image. Here, a darker pixel is represented by an earlier spike. (Right) Each neuron spikes no more than once, and the time at which it spikes encodes the information.*

image. For comparison, the power drawn by the BrainScaleS-2 chip for this application is about the same as a few LEDs.

The underlying learning algorithm is built on a rigorous derivation of the spike time in biologically inspired neuronal systems. This makes it possible to precisely quantify the effect of input spike times and connection strengths on later spikes, which in turn allows this effect to be computed throughout networks of multiple layers (Figure 1). The precise value of a single spike's effect is used on a host computer to calculate a change in the connectivity of neurons on the chip that improves the network's output. Crucially, we demonstrated that our approach is stable with regards to various forms of noise and deviations from the ideal model, which represents an essential prerequisite for physical computation, be it biological or artificial. This makes our algorithm suitable for implementation on a wide range of neuromorphic platforms.

Although these results are already highly competitive compared to other related neuromorphic realisations of spike-based classification (Figure 3), it is important to emphasise that the BrainScaleS-2 neuromorphic chip is not specifically optimised for our form of neural computation and learning but is rather a multi-purpose research device. It is likely that optimisation of the system, or hardware dedicated to classification alone will further exploit the algorithm's benefits. Even though the current BrainScaleS-2 generation is limited in size, the algorithm can scale up to larger systems. In particular, coupled with the intrinsically parallel nature of the accelerated hardware, a scaled-up version of our model would not require longer execution time, thus conserving its advantages when applied to larger, more complex data. We thus view our results as a successful proof-of-concept implementation, highlighting the advantages of sparse, but robust coding combined with fast, low-power silicon substrates, with intriguing potential for edge computing and neuroprosthetics.

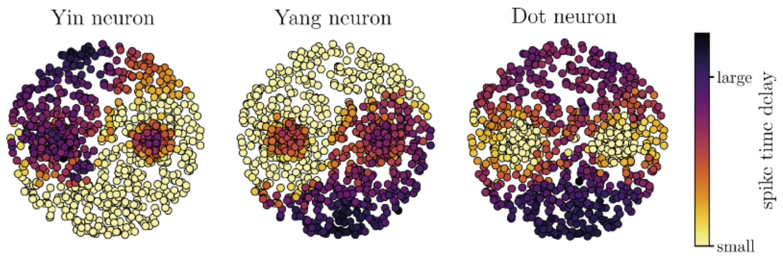


*Figure 2: Artificial dataset resembling the Yin Yang symbol [3], and the output of a network trained with our algorithm. The goal of each of the target neurons is to spike as early as possible (small delay, bright yellow color) when a data point, represented as a circle, is "in their area". One can see that the three neurons cover the correct areas in bright yellow.*

| platform | type | coding | network size/structure | energy per classification | classifications per second | test accuracy | reference |
|---|---|---|---|---|---|---|---|
| SpiNNaker | digital | rate | 764-600-500-10 | 3.3 mJ | 91 | 95.0 % | [A] |
| True North | digital | rate | CNN | 0.27 µJ | 1000 | 92.7 % | [B] |
| True North | digital | rate | CNN | 108 µJ | 1000 | 99.4 % | [B] |
| unnamed (Intel) | digital | temporal | 784-236-20-10 | 17.1 µJ | 6250 | 89.0 % | [C] |
| BrainScaleS-2 | mixed | temporal | 256-128-10 | 16 µJ | 10000 | 95.9 % | this work |

*Figure 3: Comparison with other spike-based neuromorphic classifiers on the MNIST data set, see [1] for details. [A]: E. Stromatias et al. 2015, [B]: S. Esser et al. 2015, [C]: G. Chen et al. 2018.*

We would like to thank our collaborators Andreas Baumbach, Sebastian Billaudelle, Oliver Breitwieser, Benjamin Cramer, Dominik Dold, Akos F. Kungl, Walter Senn, Johannes Schemmel and Karlheinz Meier, as well as the Manfred Stärk foundation for ongoing support. This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).

**Links:**

[L1] https://kwz.me/h53
[L2] https://kwz.me/h5S
[L3] https://kwz.me/h54

**References:**

[1] J. Göltz, L. Kriener, et al.: "Fast and deep: energy-efficient neuromorphic learning with first-spike times," arXiv:1912.11443, 2019.

[2] S. Billaudelle, et al.: "Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate," 2020 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2020.

[3] L. Kriener, et al.: "The Yin-Yang dataset," arXiv:2102.08211, 2021.

**Please contact:**

Julian Göltz, NeuroTMA group, Department of Physiology, University of Bern, Switzerland and Kirchhoff-Institute for Physics, Heidelberg University, Germany
julian.goeltz@kip.uni-heidelberg.de

# Higher Cognitive Functions in Bio-Inspired Artificial Intelligence

by Frédéric Alexandre, Xavier Hinaut, Nicolas Rougier and Thierry Viéville (Inria)

***Major algorithms from artificial intelligence (AI) lack higher cognitive functions such as problem solving and reasoning. By studying how these functions operate in the brain, we can develop a biologically informed cognitive computing; transferring our knowledge about architectural and learning principles in the brain to AI.***

Digital techniques in artificial intelligence (AI) have been making enormous progress and offer impressive performance for the cognitive functions they model. Deep learning has been primarily developed for pattern matching, and extensions like Long Short Term Memory (LSTM) networks can identify and predict temporal sequences. Adaptations to other domains, such as deep reinforcement learning, allow complex strategies of decision-making to be learnt to optimise cumulated rewards.

Contrary to human performance, these techniques require large training corpora and long training times. These challenges can be partly addressed by increasing data storage and computing power capacities. But another major flaw is less often mentioned: all the underlying cognitive functions are rather elementary and do not correspond to what is generally considered as human intelligence. The questions are thus: how is it possible to implement prospective reasoning and planning with these techniques? What about problem solving and creativity? Together with neuroscientists, we are studying the principles of brain organisation and memory network interactions (Figure 1). The main goal of our work, which takes place in the Mnemosyne Inria lab at the Bordeaux Neurocampus, is to transfer major experimental findings to new models of AI, capable of such higher cognitive functions.

We started by modelling the loops between the frontal cortex and the basal ganglia, known to play a major role in decision-making and in the acquisition of skills [1]. This has led us to propose that a major distinction should be made between two kinds of loops. Loops involving agranular areas of the frontal cortex, which have been well studied in rodents, are responsible for learning sensorimotor skills and stimulus-driven decision-making. These loops can be related to elementary cognitive functions described above, providing immediate responses to external cues, and internal reinforcement. In contrast, granular frontal regions present in primates, which have not been so well studied, are generally associated with what is called meta-cognition. Here, the same learning principles are applied to the history of performance of the elementary functions to decide which of these elementary functions are triggered, inhibited or updated, in a given context. This process yields contextual flexibility and rapid updating, as observed in higher cognitive functions.

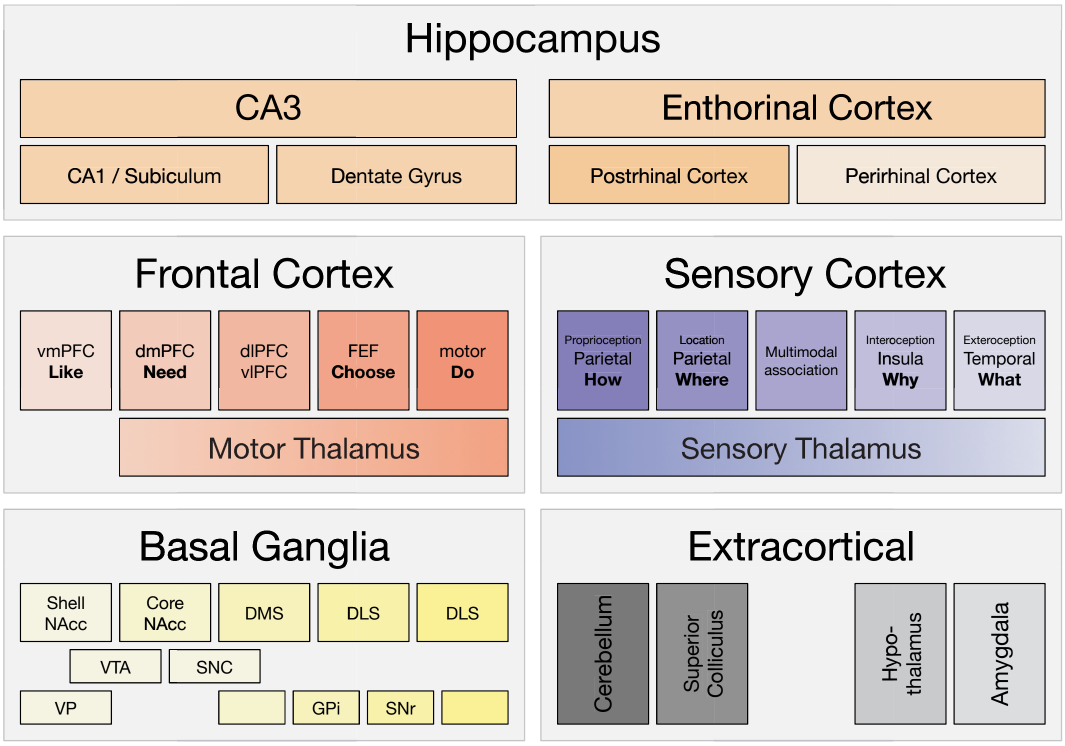


*Figure 1: This global architecture of the brain can be considered to be made of five cortico-basal loops and their main afferents from hippocampus and other extra-cortical structures. The three most central loops, which implement higher cognitive functions, include granular frontal areas.*

From a modelling point of view, this organisation of architecturally similar loops is very interesting because it implies that similar computing principles (implemented in the circuitry of cortico-basal loops) are exploited on different kinds of information to implement different kinds of behaviour (reflexive and reflective). We are, for example, investigating the possibility of basing decisions on the mechanisms of working memory [2], which represents a history of past activity, instead of decisions being made based on the activity itself. Likewise, decisions between different strategies can be made from levels of confidence estimated from previous experiences, instead of simply deciding from actual rewards. Still relying on architectural similarities, future work could be dedicated to the processing of temporal sequences: Projections of agranular frontal areas to the parietal cortex implement the temporal chaining of sensorimotor skills and homologous projections from the granular frontal areas are reported to play a role in sequential subgoal decomposition.

In addition, studying the projections from the hippocampus to the cortico-basal loops has given us important insights that can help us address another major problem related to prohibitive computational times for learning skills and decision criteria. The hippocampus is reported to be the place of episodic learning, storing past episodes and learning to replay them to train semantic and procedural memory. In neuroscience, this has been shown to be a major way to accelerate learning for the latter kinds of memory. We are currently working on computational models in this area, which are adapted to both improve learning time and decrease the size of the training corpora needed. In addition, the hippocampus is known to perform rapid and arbitrary binding of information. It has been proposed that the hippocampus in association with the frontal cortex can build prospective memory, also called imagination, which is fundamental for planning in the future. This is another area that we intend to study.

Here we have sketched a brief overview of the cerebral architecture responsible for higher cognitive functions and have explained that knowledge in neuroscience is already available to transfer computing principles to AI. Interestingly, this cerebral architecture is tightly associated with brain regions responsible for elementary cognitive functions related to sensorimotor processing and reinforcement learning. This research program can consequently be seen as an extension of existing models in AI. A more general reference to the nervous system [3] provides the anchoring of these models in the perceptual, bodily, emotional and motivational dimensions of behaviour, thus providing direct links for the robotic implementation of autonomous agents, which we are also investigating. On a different note, accessing higher cognitive functions also corresponds to manipulating structured knowledge representations and mechanisms of abstract reasoning, which are classically studied in symbolic AI. This opens unique perspectives into the study of the emergence of meaning and internal thought, which apply to the domains of natural language processing and educational science.

**Link:**
[L1] http://www.imn-bordeaux.org/en/teams/mnemosyne-mnemonic-synergy/

**References:**

[1] T. Boraud, A. Leblois, N.P. Rougier: “A natural history of skills”, Progress in Neurobiology, 171, 114 124, 2018. https://doi.org/10.1016/j.pneurobio.2018.08.003

[2] A. Strock, X. Hinaut, N.P. Rougier: “A Robust Model of Gated Working Memory”, Neural Computation, 32(1), 153 181, 2019 https://doi.org/10.1162/neco_a_01249

[3] F. Alexandre: “A Global framework for a systemic view of brain modeling”, Brain Inf. 8, 3 (2021). https://doi.org/10.1186/s40708-021-00126-4

**Please contact:**

Frederic Alexandre
Inria Bordeaux Sud-Ouest, France
Frederic.Alexandre@inria.fr

# Self-Organizing Machine Architecture

by Bernard Girau (Université de Lorraine), Benoît Miramond (Université Côte d’Azur), Nicolas Rougier (Inria Bordeaux) and Andres Upegui (University of Applied Sciences of Western Switzerland)

***SOMA is a collaborative project involving researchers in France and Switzerland, which aims to develop a computing machine with self-organising properties inspired by the functioning of the brain. The SOMA project addresses this challenge by lying at the intersection of four main research fields, namely adaptive reconfigurable computing, cellular computing, computational neuroscience, and neuromorphic engineering. In the framework of SOMA, we designed the SCALP platform, a 3D array of FPGAs and processors permitting to prototype and evaluate self-organisation mechanisms on physical cellular machines.***

The tremendous increase in transistor integration in recent years has reached the limits of classic Von Neuman architectures. Nonetheless, one major issue is the design and deployment of applications that cannot make optimal use of the available hardware resources. This limit is even more acute when we consider application domains where the system evolves under unknown and uncertain conditions, such as mobile robotics, IoT, autonomous vehicles or drones. In the end, we cannot foresee every possible context that the system will face during its lifetime, thus making it impossible to determine the optimal hardware substrate. Interestingly, the biological brain has solved this problem using a dedicated architecture and mechanisms that offer both adaptive and dynamic computations, namely, self-organisation [2].

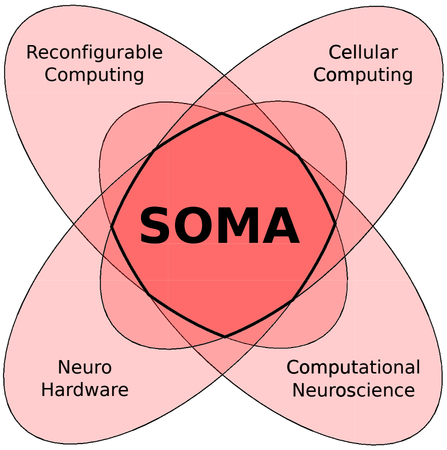


*Figure 1: SOMA is a convergence point between past research approaches toward new computation paradigms: adaptive reconfigurable architecture, cellular computing, computational neuroscience, and neuromorphic hardware.*

However, even though neuro-biological systems have often been a source of inspiration for computer science, the transcription of self-organisation at the hardware level is not straightforward, presenting several challenges. We are working on coupling this new computational paradigm with an underlying conventional systolic architecture [1]. We use self-organised neural populations on a cellular machine where local routing resources are not separated from computational resources; this ensures natural scalability and adaptability as well as a better performance/power consumption trade-off compared to other conventional embedded solutions. This new computing framework may indeed represent a viable integration of neuromorphic computing into the classical Von Neumann architecture and could endow these hardware systems with novel adaptive properties [3].

## Cortical plasticity and cellular computing in hardware

This objective led us to study a desirable property from the brain that encompasses all others: cortical plasticity. This term refers to one of the main developmental properties of the brain where the organisation of its structure (structural plasticity) and the learning of the environment (synaptic plasticity) develop simultaneously toward an optimal computing efficiency. This developmental process is only made possible by some key features: the ability to focus on relevant information, representation of information in a sparse manner, distributed data processing and organisation fitting the nature of data, leading to a better efficiency and robustness. Our goal is to understand and design the first artificial blocks that are involved in these principles of plasticity. Hence, transposing plasticity and its underlying blocks into hardware is a step towards to defining a substrate of computation endowed with self-organisation properties that stem from the learning of incoming data.

The neural principles of plasticity may not be sufficient to ensure that such a substrate of computation is scalable enough in the face of future massively parallel devices. We anticipate that the expected properties of such alternative computing devices could emerge from a close interaction between cellular computing (decentralisation and hardware compliant massive parallelism) and neural computation (self-organisation and adaptation). We also believe that neuro-cellular algorithmics and hardware design are so tightly related that these two aspects should be studied together. Therefore, we propose to combine neural adaptivity and cellular computing efficiency through a neuro-cellular approach of synaptic and structural self-organisation that defines a fully decentralised control layer for neuromorphic reconfigurable hardware. To achieve this goal, the project brings together neuroscientists, computer science researchers, hardware architects and micro-electronics designers to explore the concepts of a Self-Organizing Machine Architecture: SOMA (see Figure 1). This self-organisation property is already being studied in various fields of computer science, but we are investigating it in an entirely new context, applying perspectives from computational neuroscience to the design of reconfigurable microelectronic circuits. The project focuses on the blocks that will pave the way in the long term for smart computing substrates, exceeding the limits of current technology.

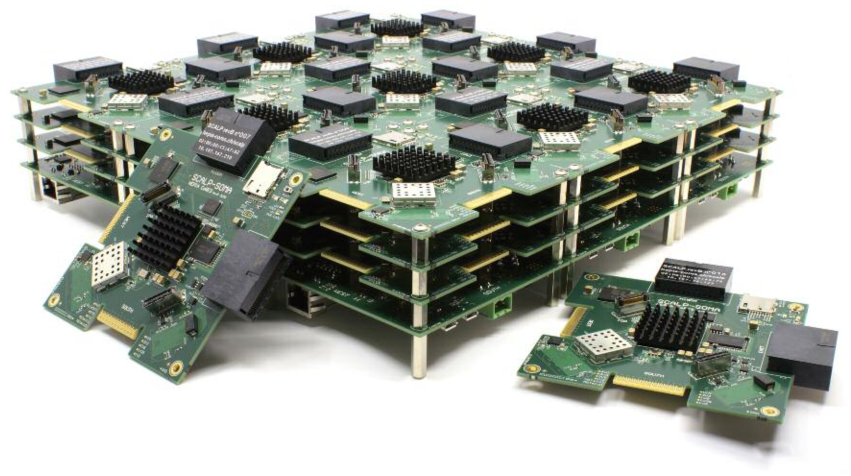


*Figure 2: The SCALP platform, a set of FPGAs and processors with 3D topology, was designed to evaluate self-organisation mechanisms on cellular machines. Algorithms based on cellular self-organising maps are the basis of the self-adaptation properties.*

## Convergence of research fields

Previous work has explored the possibility of using neural self-organising models to control task allocation on parallel substrates [1], while adapting neural computational paradigms to cellular constraints. Adaptive reconfigurable computing focuses on virtualisation of reconfigurable hardware, runtime resource management, dynamic partial reconfiguration, and self-adaptive architectures. Cellular approaches of distributed computing result in decentralised models that are particularly well adapted to hardware implementations. However, cellular computation still lacks adaptation and learning properties. This gap may be filled with the help of computational neuroscience and neuromorphic engineering through the definition of models that exhibit properties like unsupervised learning, self-adaptation, self-organisation, and fault tolerance, which are of particular interest for efficient computing in embedded and autonomous systems. However, these properties only emerge from large fully connected neural maps that result in intensive synaptic communications.

Our self-organising models are deployed on the Self-configurable 3-D Cellular multi-FPGA Adaptive Platform (SCALP) (Figure 2), which has been developed in the framework of the SOMA project. SCALP is a multi-FPGA hardware platform that enables the creation of prototype 3D NoC architectures with dynamic topologies. A node is composed of a Xilinx Zynq SoC (dual-core ARM Cortex-A9 @866 MHz + Artix-7 programmable logic with 74,000 cells), 2 Gb DDR3 SDRAM, a 5-port Ethernet switch, and a PLL. The inherent cellular scalable architecture of SCALP, coupled with its enhanced interfaces, provides the ideal computation platform for implementing cellular neuromorphic architectures by imposing real physical connectivity constraints. Also, a real 3D machine architecture (instead of a simulated one) can better handle scalability issues when modelling dynamic bio-inspired 3D neural connectivity. We have already proposed such models using both dynamical sprouting and pruning of synapses inside a self-organising map and a method to migrate neurons between clusters to dynamically reassign neurons from one task to another. These methods provide dynamic structural and computational resource allocations, inspired by the brain's structural and functional plasticity, and are currently being deployed onto the SCALP platform.

**Links:**
[L1] https://kwz.me/h55
[L2] https://kwz.me/h5

**References:**

[1] L. Rodriguez, B. Miramond, B. Granado "Toward a sparse self-organizing map for neuromorphic architectures", ACM JETC, 11 (4), 1-25, 2015.

[2] G.I. Detorakis, N.P. Rougier: "A Neural Field Model of the Somatosensory Cortex: Formation, Maintenance and Reorganization of Ordered Topographic Maps", PLoS ONE 7, 7, e40257, 2012.

[3] A. Upegui, et al.: "The Perplexus bio-inspired reconfigurable circuit", in Proc. of the Second NASA/ESA Conference on Adaptive Hardware and Systems, 2007.

**Please contact:**

Nicolas Rougier, Inria Bordeaux Sud-Ouest, France,
Nicolas.Rougier@inria.fr
+33 7 82 50 31 10

# Reentrant Self-Organizing Map: Toward Brain-Inspired Multimodal Association

by Lyes Khacef (University of Groningen), Laurent Rodriguez and Benoît Miramond (Université Côte d'Azur, CNRS)

***Local plasticity mechanisms enable our brains to self-organize, both in structure and function, in order to adapt to the environment. This unique property is the inspiration for this study: we propose a brain-inspired computational model for self-organization, then discuss its impact on the classification accuracy and the energy-efficiency of an unsupervised multimodal association task.***

Our brain-inspired computing approach attempts to simultaneously reconsider AI and von Neumann's architecture. Both are formidable tools responsible for digital and societal revolutions, but also intellectual bottlenecks linked to the ever-present desire to ensure the system is under control. The brain remains our only reference in terms of intelligence: we are still learning about its functioning, but it seems to be built on a very different paradigm in which its developmental autonomy gives it an efficiency that we haven't yet attained in computing.

Our research focuses on the cortical plasticity that is the fundamental mechanism enabling the self-organization of the brain, which in turn leads to the emergence of consistent representations of the world. According to the neurobiologist F. Varela, self-organization can be defined as a global behaviour emerging from local and dynamic interactions, i.e., unifying structure and function in a single process: the plasticity mechanism. It is hence the key to our ability to build our representation of the environment based on our experiences, so that we may adapt to it. It is also the basis of an extremely interesting characteristic of the human brain: multimodal association.

In fact, most processes and phenomena in the natural environment are expressed under different physical guises, which we refer to as different modalities. Multimodality is considered a fundamental principle for the development of embodied intelligence, as pointed out by the neuroscientist A. Damasio, who proposed the Convergence-Divergence Zone framework [1]. Such a framework models the neural mechanisms of memorisation and recollection. Despite the diversity of the sensory modalities, such as sight, sound and touch, the brain arrives at similar representations and concepts (convergence). On the other hand, biological observations show that one modality can activate the internal representation of another modality. For example, when watching a specific lip movement without any sound, the activity pattern induced in the early visual cortices activates in early auditory cortices the representation of the sound that usually accompanies the lip movement (divergence).

Here we summarise our work on the Reentrant Self-Organizing Map (ReSOM) [2], a brain-inspired computational neural system based on the reentry theory from G. Edelman [3] and J.P. Changeux using Kohonen Self-Organizing Maps (SOMs) and Hebbian-like learning to perform multimodal association (see Figure 1).

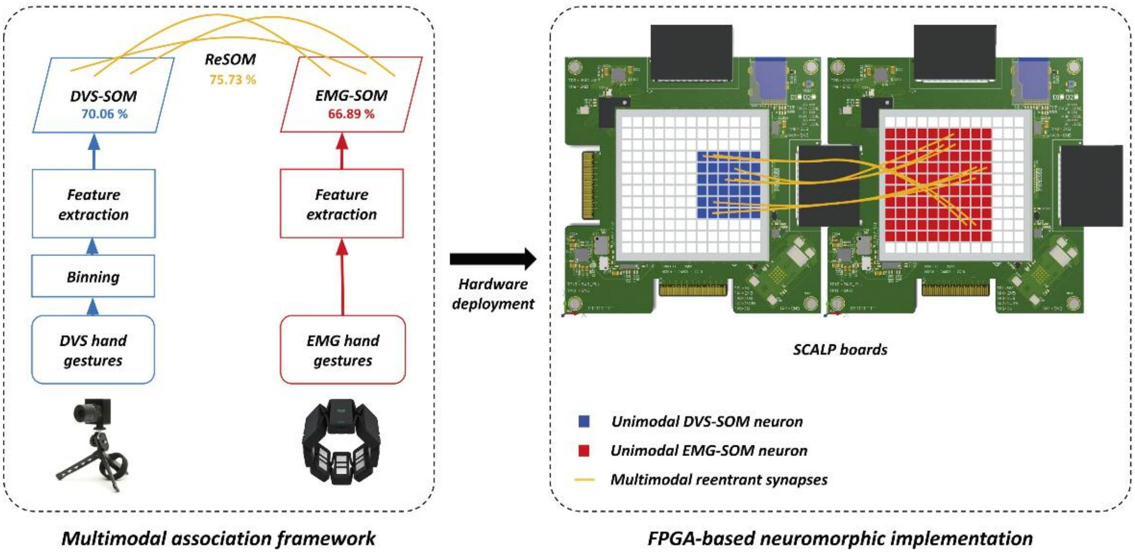


*Figure 1: Reentrant Self-Organizing Map: (left) Processing pipeline from data acquisition at input to multimodal association for decision making at the output with unimodal and multimodal accuracies for a hand gestures recognition task based on a DVS camera and EMG sensor; (right) FPGA-based neuromorphic implementation of the proposed self-organizing artificial neural network on multiple SCALP boards for real-time and energy-efficient processing.*

## ReSOM model

The brain's plasticity can be divided into two distinct forms: (i) structural plasticity, which, according to the Selection Neural Groups Theory [3], changes the neurons' connectivity by sprouting (creating) or pruning (deleting) synaptic connections, and (ii) synaptic plasticity that modifies (increases or decreases) the existing synaptic strength. We explore both mechanisms for multimodal association through Hebbian-like learning. In the resulting network, the excitement of one part spreads to all the others and a fragment of memory is enough to awaken the entire memorised experience. The network becomes both a detector and a producer of signals.

First, the unimodal learning is performed independently for each modality using the SOM, a brain-inspired artificial neural network that learns in an unsupervised manner (without labels). Then, based on co-occurrent multimodal inputs, the neurons of different SOMs create and reinforce the reentrant multimodal association via sprouting and Hebbian-like learning. At the end of the multimodal binding, the neural group selection is made, and each neuron prunes up to 90% of the possible connections to keep only the strongest ones. The third step is then to give sense to these self-associating groups of neurons. This is made by labelling one of the SOMs maps using very few labels (typically 1%), so that each neuron is assigned the class it represents. The fourth step is to label the entire network (the other maps) by using the divergent activity from the first labelled map. This way, the system breaks with the general principle of classical machine learning by exploiting the strength of the multimodal association and takes advantage of the coherence of the data from its experience to build in an incremental way a robust representation of the environment. From an application point of view, this means that the system only needs few annotations from a single modality to label the maps of all the other modalities. Finally, once the multimodal learning is done and all neurons from all SOMs are labelled, the system computes the convergence of the information from all the modalities to achieve a better representation of the multi-sensory input. This global behaviour emerges from local interactions among connected neurons.

## Results and discussion

Our experiments [2] show that the divergence labelling leads to approximately the same unimodal accuracy as when using labels, while the convergence mechanism leads to a gain in the multimodal accuracy of +8.03% for a written/spoken digits database [L1] and +5.67% for a DVS/EMG hand gestures database [L2]. We also gained +5.75% when associating visual hand gestures with spoken digits, illustrating the McGurk effect. Indeed, studies in cognitive and developmental psychology show that spoken labels and auditory modality in general add complementary information that improves object categorisation.

In summary, the ReSOM model exploits the natural complementarity between different modalities so that they complete each other and improve multimodal classification. Furthermore, it induces a form of hardware plasticity where the system's topology is not fixed by the user but learned along the system's experience through self-organization. It reduces the inter-map communication and thus reduces the system's energy consumption. This result could open up a whole lot of new directions, inspired by the brain's plasticity, for future designs and implementations of self-organizing hardware architectures in autonomous systems such as vehicles, robots, drones or even cortical prosthesis.

**Links:**

[L1] https://zenodo.org/record/4452953
[L2] https://zenodo.org/record/3663616

**References:**

[1] Damasio: "Time-locked multiregional retroactivation: A systems level proposal for the neural substrates of recall and recognition".

[2] Khacef et al.: "Brain-Inspired Self-Organization with Cellular Neuromorphic Computing for Multimodal Unsupervised Learning".

[3] Edelman: "Group selection and phasic reentrant signaling a theory of higher brain function".

**Please contact:**

Lyes Khacef, University of Groningen, Netherlands, l.khacef@rug.nl
Prof. Benoît Miramond, Université Côte d'Azur, France,
benoit.miramond@univ-cotedazur.fr

# Brain-inspired Learning Drives Advances in Neuromorphic Computing

by Nasir Ahmad (Radboud University), Bodo Rueckauer (University of Zurich and ETH Zurich) and Marcel van Gerven (Radboud University)

***The success of deep learning is founded on learning rules with biologically implausible properties, entailing high memory and energy costs. At the Donders Institute in Nijmegen, NL, we have developed GAIT-Prop, a learning method for large-scale neural networks that alleviates some of the biologically unrealistic attributes of conventional deep learning. By localising weight updates in space and time, our method reduces computational complexity and illustrates how powerful learning rules can be implemented within the constraints on connectivity and communication present in the brain.***

The exponential scaling of modern compute power (Moore’s law) and data storage capacity (Kryder’s law), as well as the collection and curation of big datasets, have been key drivers of the recent deep learning (DL) revolution in artificial intelligence (AI). This revolution, which makes use of artificial neural network (ANN) models that are trained on dedicated GPU clusters, has afforded important breakthroughs in science and technology, ranging from protein folding prediction to vehicle automation. At the same time, several outstanding challenges prohibit the use of DL technology in resource-bounded applications.

Among these issues, the quest for low-latency, low-power devices with on-demand computation and adaptability has become a field of competition. A number of approaches have emerged with candidate solutions to these problems. These include highly efficient hardware specifically designed to carry out the core tensor operations which compose ANN models (especially for mobile devices), cloud-based compute farms to supply on-demand compute to internet-connected systems (held back by access and relevant privacy concerns) and more.

Brain-inspired methods have also emerged within this sphere. After all, the mammalian brain is a prime example of a low-power and highly flexible information processing system. Neuromorphic computing is the name of the field dedicated to instantiating brain-inspired computational architectures within devices. In general, neuromorphic processors feature the co-location of memory and compute, in contrast to traditional von-Neumann architectures that are used by modern computers. Other key features include asynchronous communication of the sub-processors (there is no global controller of the system), and data-driven computation (computing only takes place with significant changes in the input). A number of companies and academic research groups are actively pursuing the development of such neuromorphic processors (Intel’s Loihi, IBM’s TrueNorth, SpiNNaker, and BrainScaleS, to name a few) [1]. These developments progress apace.

We expect that brain-inspired learning rules can become a major driver of future innovation for on-board neuromorphic learning. In particular, the above described architectural design

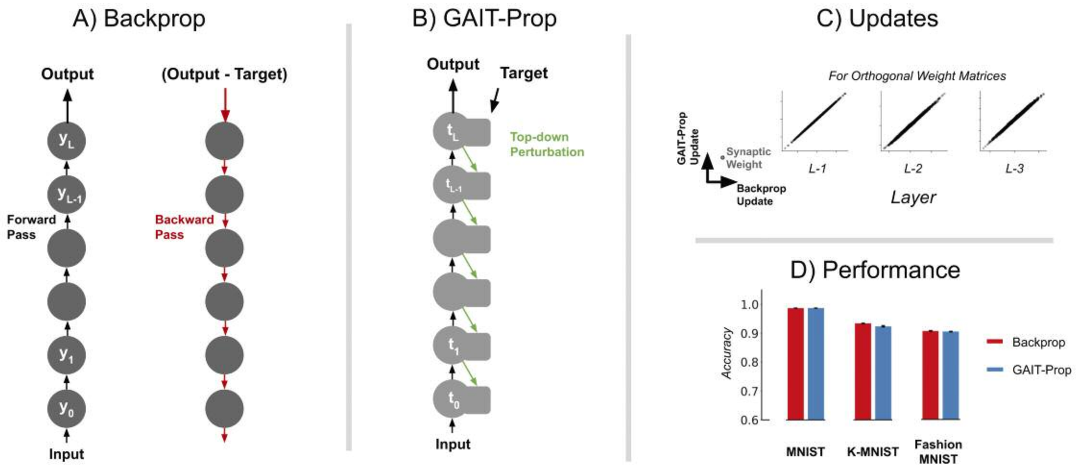


*Figure 1: Comparison of standard backpropagation with our proposed GAIT-Prop method. In sections A) and B), circles indicate layers of a deep neural network. A) In backpropagation, all neuron activations $y_i$ in each layer i of the network need to be stored during a forward pass, and then the network activity halted so that that weights can be updated in a separate backward pass based on a global error signal. B) In GAIT-Prop, a top-down perturbation circuit is described which can transmit target values ti required to compute weight updates locally at each unit by making use of the dynamics of the system. C) Under particular constraints, GAIT-Prop produces identical weight updates compared to Backprop. D) This is also exhibited during training where on a range of benchmark datasets (MNIST, KMNIST, and Fashion MNIST) GAIT-Prop matches the performance of backpropagation.*

choices for neuromorphic chips precisely match the constraints faced by neurons in brains (local memory and compute, asynchronous communication, data-driven computation and more).

Unfortunately, the computations required to carry out the traditional gradient-based training of ANNs (known as backpropagation of error) break the properties of both neuromorphic architectures and real neural circuits. Error computations in the typical format require non-local information, which implies that the memory distributed across the sub-processing nodes would need to communicate in a global fashion (Figure 1A). For this reason alone, the traditional methods for backpropagation of error are undesirable for neuromorphic "on-chip" training. Furthermore, computations associated with learning and inference (i.e., the application of the ANN) are carried out in separate phases, leading to an undesirable "blocking" phenomenon. By comparison, the brain does not appear to require non-local computations for learning. Thus, by finding solutions to brain-inspired learning, we might arrive at solutions to "on-chip" training of neuromorphic computing.

Recent developments in brain-inspired learning have produced methods which meet these requirements [2]. In particular, our group recently developed a method (GAIT-Prop [3]) to describe learning in ANN models. GAIT-Prop relies on the same system dynamics during inference and training (Figure 1B) such that no additional machinery is required for gradient-based learning. When the system is provided with an indication of the "desired" output of the system (a training signal), it makes use of theoretically determined inter-neuron connectivity to propagate this desired signal across the network structure through the activities of the network units. The change in activity can then be used by each individual neuron to carry out relevant updates.

Importantly, under some limited constraints, the updates produced by the GAIT-Prop algorithm precisely match those of the very powerful backpropagation of error method (Figure 1C). This ensures that we can achieve matching performance despite the local and distributed nature of the GAIT-Prop algorithm (Figure 1D). Our algorithm also provided for understanding how a desired network output can be translated into target outputs for every node of an ANN system. Since our method relies on error signals being carried within the dynamics of individual units of the network (requiring no specific "error" nodes) it requires less computational machinery to accomplish learning. This feature is ideal for neuromorphic systems as it ensures simplicity of node dynamics while enabling high accuracy.

We see our approach and extensions thereof, in which systems that learn are close to indistinguishable in their dynamics to the systems carrying out inference computations, as an important step in the development of future neuromorphic systems, as it mitigates the complexities associated with standard learning algorithms. Systems equipped with this capability could be embedded in mobile devices and would be capable of learning with data locally, also thereby reducing privacy concerns which are common in an era of cloud-computing and mass data storage.

**References**

[1] M. Bouvier et al.: "Spiking neural networks hardware implementations and challenges: A survey", ACM JETC, 2019.
[2] T.P. Lillicrap et al.: "Backpropagation and the brain", Nat Rev Neurosci, 2020.
[3] N. Ahmad et al.: "GAIT-prop: A biologically plausible learning rule derived from backpropagation of error", NeurIPS, 2020.

**Please contact:**

Nasir Ahmad
Radboud University, Nijmegen
N.Ahmad@donders.ru.nl

# Memory Failures Provide Clues for more Efficient Compression

by Dávid G. Nagy, Csenge Fráter and Gergő Orbán (Wigner Research Center for Physics)

***Efficient compression algorithms for visual data lose information for curbing storage capacity requirements. An implicit optimisation goal for constructing a successful compression algorithm is to keep compression artifacts unnoticed, i.e., reconstructions should appear to the human eye to be identical to the original data. Understanding what aspects of stimulus statistics human perception and memory are sensitive to can be illuminating for the next generation of compression algorithms. New machine learning technologies promise fresh insights into how to chart the sensitivity of memory to misleading distortions and consequently lay down the principles for efficient data compression.***

Humans are prone to errors when it comes to recollecting details about past experiences. Much research has addressed the questions of which details our memory chooses to store and which are systematically discarded. Until recently we have not had methods to learn the complex statistics to which memory has adapted to (natural statistics) so there is little data available about how these systematic failures link to natural stimulus structure.

Researchers at the Wigner Research Center for Physics, Budapest, have addressed this challenge using variational autoencoders, a new method in machine learning that learns a latent variable generative model of the data statistics in an unsupervised manner [1]. Latent variable generative models aim to identify the features that contribute to the generation of the data and every single data point is encoded as a combi-

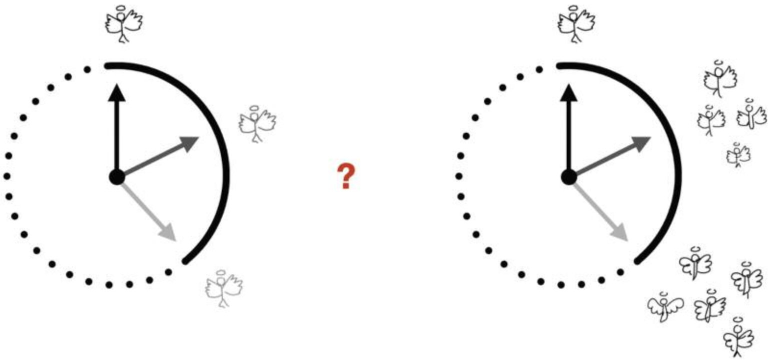


*Figure 1: The dynamics of forgetting: memory undergoes changes and information is systematically discarded over time, which provides insights into the sensitivities of memory. An account of information loss would be to simply lose memory through "fading": uniformly losing precision when reconstructing episodes from memory. Instead of such non-specific information loss, compression with latent-variable generative models implies that reconstruction errors reflect uncertainty in latent features. As the delay between encoding and recollection increases, latent variable representation of the stimulus is reshaped and we can capture signatures of lower-rate compressions.*

nation of these features. The proposed semantic compression idea establishes how to distinguish compression artefacts that go unnoticed from those that are more easily detected: artefacts that induce changes in the latent features are more easily noticed than artefacts of equal strength that leave the latent features unchanged. The theory of semantic compression was put to a test by revisiting results from human memory experiments in a wide array of domains (recollection of words, chess board configurations, or drawings) and showing that semantic compression can predict how memory fails under a variety of conditions.

Using generative models to encode complex stimuli offers new perspectives for compression. The power of these generative models lies in the fact that the latent features discovered by the generative model provide a concise description of the data and are inherently empowered by reconstruction capabilities [2]: in contrast to more traditional deep learning models, variational autoencoders are optimised for efficient reconstruction ability. Nonlinear encoding is thus a way to build ultra-low bit-rate data compression technologies. This "generative compression" idea is boosted by two critical factors that concern two qualitatively different aspects of data compression. First, recent work in the field has demonstrated a tight link between variational autoencoders and the theory of lossy compression [3]. This link demonstrates that VAEs are optimised for the same training objective as the theory that establishes the optimality criterion for lossy compression. In fact, a continuum of optimal compressions can be established, depending on the allowed storage capacity, thus different generative models can be constructed for different needs. Second, research by Nagy et al. provides support that the structure of errors that generative compression is sensitive to is similar to the structure of errors that memory is sensitive to. Theory claims that this is not simply a lucky coincidence: the basis of it is that the generative model is trained on stimulus statistics that are designed to approximate the statistics that the human brain was adapted to.

In the context of human memory, the study highlights a fascinating view. We are all too aware that memories undergo disappointing dynamics that result in what we call "forgetting". Is the gradual process of forgetting best understood as a simple process of fading? The answer the authors provide could be more appropriately described as memories become less elaborate as time passes: the theory of lossy compression establishes a harsh trade-off between storage allowance and the richness of the retained details. As time passes it is assumed that the capacity allocated for a memory decreases and details are lost with stronger compression. And what are the details that remain? Those are again determined by stimulus statistics: features that are closer to the gist of the memory prevail while specific details disappear.

**References:**

[1] D.G. Nagy, B. Török, G. Orbán: "Optimal forgetting: Semantic compression of episodic memories", PLoS Comp Biol, e1008367, 2020.

[2] I. Higgins I, et al.: "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework", Proc. of ICLR, 2018.

[3] A. Alemi, et al.: "An information-theoretic analysis of deep latent-variable models", arXiv:1711:00464.

**Please contact:**

Gergő Orbán, Computational Systems Neuroscience Lab, Dept. of Computational Sciences, Wigner Research Center for Physics, Hungary
orban.gergo@winger.hu

# Neuronal Communication Process Opens New Directions in Image and Video Compression Systems

by Effrosyni Doutsi (FORTH-ICS), Marc Antonini (I3S/CNRS) and Panagiotis Tsakalides (University of Crete and FORTH/ICS)

***The 3D ultra-high-resolution world that is captured by the visual system is sensed, processed and transferred through a dense network of tiny cells, called neurons. An understanding of neuronal communication has the potential to open new horizons for the development of ground-breaking image and video compression systems. A recently proposed neuro-inspired compression system promises to change the framework of the current state-of-the-art compression algorithms.***

Over the last decade, the technological development of cameras and multimedia devices has increased dramatically to meet societal needs. The significant progress of these technologies has ushered in the big data era, accompanied by serious challenges, including the tremendous increase in volume and variety of measurements. Although most big data challenges are being addressed with paradigm shifts in machine learning (ML) technologies, where a limited set of observations and associated annotations are utilised for training models to automatically extract knowledge from raw data, little has been done about the disruptive upgrade of storage efficiency and compression capacity of existing algorithms.

The BRIEFING project [L1] aims to mimic the intelligence of the brain in terms of compression. The research is inspired by the great capacity of the visual system to process and encode visual information in an energy-efficient and very compact yet informative code, which is propagated to the visual cortex where the final decisions are made.

If one considers the visual system as a mechanism that processes the visual stimulus, it seems an intelligent and very efficient model to mimic. Indeed, the visual system consumes low power, it deals with high resolution dynamic signals (109 bits per second) and it transforms and encodes the visual stimulus in a dynamic way far beyond the current compression standards. During recent decades, there has been significant research into understanding how the visual system works, the structure and the role of each layer and individual cell that lies along the visual pathway, and how the huge volume of visual information is propagated and compacted through the nerve cells before reaching the visual cortex. Some very interesting models that approximate neural behaviour have been widely used for image processing applications, including compression. The biggest challenge however, is that the brain uses the neural code to learn, analyse and make decisions without reconstructing the input visual stimulus.

There are several proven benefits to applying neuroscience models to compression architectures. We developed a neuro-inspired compression mechanism by using the Leaky Integrate-an-Fire (LIF) model, which is considered to be the simplest model that approximates neuronal activity, in order to transform an image into a code of spikes [1]. The great advantage of the LIF model is that the code of spikes is generated as time evolves, in a dynamic manner. An intuitive explanation for the origin of this

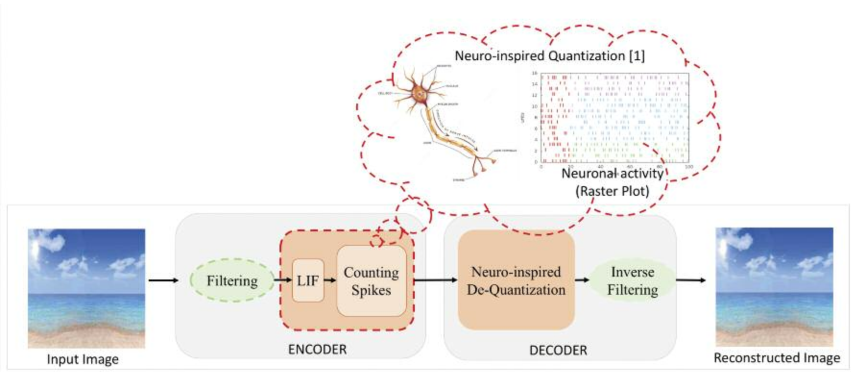


*Figure 1: An illustration of the neuro-inspired compression mechanism that enables efficient reduction of the number of bits required to store an input image using the Leaky Integrate-and-Fire (LIF) model as an approximation of the neuronal spiking activity. According to this ground-breaking architecture, an input image can be transformed into a sequence of spikes which are utilised to store and/or transmit the signal. The interpretation of the spike sequence based on signal processing techniques leads to high reconstruction quality results.*

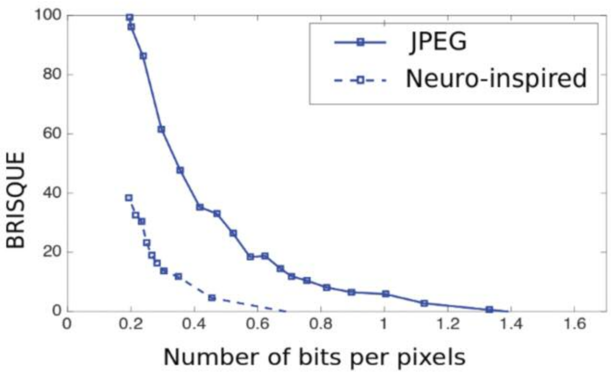


*Figure 2: This graph shows that the BRISQUE algorithm that has been trained to detect the natural characteristics of visual scenes is able to detect far more of these characteristics within images that have been compressed using the neuro-inspired compression than the JPEG standards. The BRISQUE scores are typically between 0 and 100, where the lower the score the better the natural characteristics of the visual scene.*

performance is that the longer the visual stimulus exists in front of a viewer, the better it is perceived. Similarly, the longer the LIF model is allowed to produce spikes, the more robust is the code. This behaviour is far beyond the state-of-the-art image and video compression architectures that process and encode the visual stimulus immediately and simultaneously without considering any time parameters. However, taking advantage of the time is very important, especially when considering a video stream that is a sequence of images each of which exists for a given time.

Another interesting aspect is that a neuro-inspired compression mechanism can preserve important features in order to characterise the content of the visual scene. These features are necessary for several image analysis tasks, such as object detection and/or classification. When the memory capacity or the bandwidth of the communication channel are limited it is very important to transmit the most meaningful information. In other words, one needs to find the best trade-off between the compression ratio and the image quality (rate-distortion). We have proven that neuro-inspired compression is much more valuable than the state-of-the-art such as JPEG and/or JPEG2000, which both cause drastic changes in these features [2]. More specifically, we evaluated the aforementioned models using the BRISQUE algorithm [3], a convolutional neural network that has been pre-trained in order to recognise natural characteristics within the visual scene. As a first step, we compressed a group of images with the same compression ratio using both the neuro-inspired mechanism and the JPEG standard. Then, we fed the CNN with the compressed images and we observed that it was able to detect far more natural characteristics within images that had been compressed by the neuro-inspired mechanism than JPEG standard.

This project, funded by the French Government within the framework of the “Make Our Planet Green Again” call, is a collaboration between the Institute of Compute Science (ICS) at Foundation for Research and Technology – Hellas (FORTH) and the Laboratory of Informatics, Signals and Systems Sophia Antipolis (I3S) at French National Centre for Scientific Research (CNRS)

**Link:**
[L1] https://kwz.me/h57

**References:**

[1] E. Doutsi, et al.: “Neuro-inspired Quantization”, 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 689-693.

[2] E. Doutsi, L. Fillatre and M. Antonini: “Efficiency of the bio-inspired Leaky Integrate-and-Fire neuron for signal coding”, 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1-5.

[3] A. Mittal, A. K. Moorthy, and A. C. Bovik: “No-reference image quality assessment in the spatial domain”, IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

Please contact:
Effrosyni Doutsi, ICS-FORTH, Greece
edoutsi@ics.forth.gr

# Fulfilling Brain-inspired Hyperdimensional Computing with In-memory Computing

by Abbas Rahimi, Manuel Le Gallo and Abu Sebastian (IBM Research Europe)

***Hyperdimensional computing (HDC) takes inspiration from the size of the brain’s circuits, to compute with points of a hyperdimensional space that thrives on randomness and mediocre components. We have developed a complete in-memory HDC system in which all operations are implemented on noisy memristive crossbar arrays while exhibiting extreme robustness and energy-efficiency for various classification tasks such as language recognition, news classification, and hand gesture recognition.***

A cursory examination of the human brain shows: (i) the neural circuits are very large (there can be tens of thousands of fan-ins and fan-outs); (ii) activity is widely distributed within a circuit and among different circuits; (iii) individual neurons need not be highly reliable; and (iv) brains operate with very little energy. These characteristics are in total contrast to the way traditional computers are built and operate. Therefore, to approach such intelligent, robust, and energy-efficient biological computing systems, we need to rethink and focus on alternative models of computing, such as hyperdimensional computing (HDC) [1][2].

The difference between traditional computing and HDC is apparent in the elements that the machine computes with. In traditional computing, the elements are Booleans, numbers, and memory pointers. In HDC they are multicomponent vectors, or tuples, where neither an individual component nor a subset thereof has a specific meaning: a single component of a vector and the entire vector represent the same thing. Furthermore, the vectors are very wide: the number of components is in the thousands. These properties are based on the observation that key aspects of human memory, perception, and cognition can be explained by the mathematical properties of hyperdimensional spaces comprising high-dimensional binary vectors known as hypervectors [1]. Hypervectors are defined as d-dimensional (where d ≥ 1,000) (pseudo)random vectors with independent and identically distributed (i.i.d.) components. When the dimensionality is in the thousands, a huge number of quasi-orthogonal hypervectors exist. This allows HDC to combine such hypervectors into new hypervectors using well-defined vector space operations, defined such that the resulting hypervector is unique, and with the same dimension.

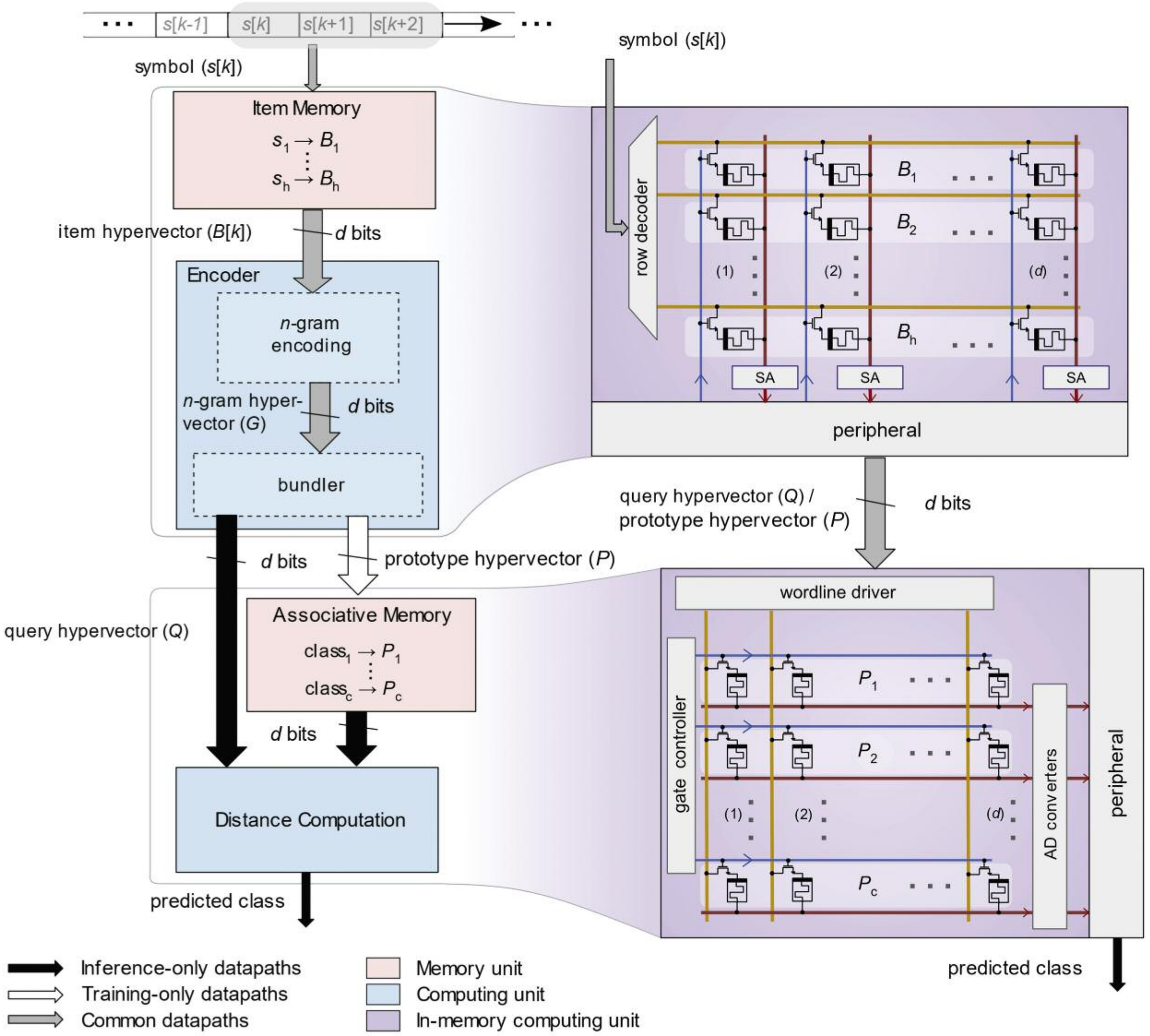


*Figure 1: The concept of in-memory HDC. A schematic of the concept of in-memory HDC showing the essential steps associated with HDC (left) and how they are realized using in-memory computing (right). An item memory (IM) stores h, d-dimensional basis hypervectors that correspond to the symbols associated with a classification problem. During learning, based on a labelled training dataset, a designed encoder performs dimensionality-preserving mathematical manipulations on the basis hypervectors to produce c, d-dimensional prototype hypervectors that are stored in an AM. During classification, the same encoder generates a query hypervector based on a test example. Subsequently, an AM search is performed between the query hypervector and the hypervectors stored in the AM to determine the class to which the test example belongs. In in-memory HDC, both the IM and AM are mapped onto crossbar arrays of memristive devices. The mathematical operations associated with encoding and AM search are performed in place by exploiting in-memory read logic and dot-product operations, respectively. A dimensionality of d = 10,000 is used. SA, sense amplifier; AD converters, analog-to-digital converters are adapted from [6].*

HDC has been employed in a range of applications, including traditional computing, machine learning, cognitive computing, and robotics [3]. It has shown significant promise in machine learning applications that involve temporal patterns, such as text classification, biomedical signal processing, multimodal sensor fusion, and distributed sensors [4]. A key advantage is that the training algorithm in HDC works in one or only a few shots: that is, object categories are learned from one or a few examples, and in a single pass over the training data as opposed to many repetitive iterations in the deep learning models [4].

HDC begins by representing symbols with i.i.d. hypervectors that are combined by nearly i.i.d.-preserving operations, namely binding, bundling, and permutation, and then stored in associative memories to be recalled, matched, decomposed, or reasoned about. Manipulation and comparison of these large patterns results in a bottleneck when implemented on the conventional von Neumann computer architectures. On the other hand, the chain of operations implies that failure in a component of a hypervector is not contagious leading to robust computational framework. For instance, when unrelated objects are represented by quasi-orthogonal 10,000-bit vectors, more than a third of the bits of a vector can be flipped by randomness, device variations, defects, and noise, and the faulty vector can still be identified with the correct one, as it is closer to the original error-free vector than to any unrelated vector chosen so far, with near certainty. Therefore, the inherent robustness and the need for manipulations of large patterns stored in memory make HDC particularly well suited to emerging computing paradigms such as in-memory computing or computational memory based on emerging nanoscale resistive memory or memristive devices [5].

In the past few years, we have been working towards designing and optimising a complete integrated in-memory HDC system in which all the operations of HDC are implemented on two planar memristive crossbars together with peripheral digital CMOS circuits. We have been devising a way of performing hypervector binding entirely within a first memristive crossbar using an in-memory read logic operation and hypervector bundling near the crossbar with CMOS logic. These key operations of HDC cooperatively encode hypervectors with high precision, while eliminating the need to repeatedly program (i.e., write) the memristive devices. Unlike previous work, this approach matches the limited endurance of memristive devices and scales well with 10,000-dimensional hypervectors, making this work the largest experimental demonstration of HDC with memristive hardware to date [6].

In our architecture, shown in Figure 1, an associative memory search is performed using a second memristive crossbar for in-memory dot-product operations on the encoded output hypervectors from the first crossbar, realising the full functionality of the HDC system. Our combination of analog in-memory computing with CMOS logic allows continual functioning of the memristive crossbars with desired accuracy for a wide range of multiclass classification tasks, including language classification, news classification, and hand gesture recognition from electromyography signals. We verify the integrated inference functionality of the system through large-scale mixed hardware/software experiments, in which hypervectors are encoded in 760,000 hardware phase-change memory devices performing analog in-memory computing. Our experiments achieve comparable accuracies to the software baselines and surpass those reported in previous work. Furthermore, a complete system-level design of the in-memory HDC architecture synthesized using 65 nm CMOS technology demonstrates a greater than six-fold end-to-end reduction in energy compared with a dedicated digital CMOS implementation. More details can be found in our paper published in Nature Electronics [6].

**References:**

[1] P. Kanerva, Hyperdimensional computing: an introduction to computing in distributed representation with high-dimensional random vectors. Cogn. Comput., 2009.

[2] P. Kanerva, “Computing with High-Dimensional Vectors,” IEEE Design & Test, 2019.

[3] A. Mitrokhin, et al. Learning sensorimotor control with neuromorphic sensors: toward hyperdimensional active perception. Science Robotics, 2019.

[4] A. Rahimi, et al. Efficient biosignal processing using hyperdimensional computing: network templates for combined learning and classification of ExG signals. Proc. IEEE, 2019.

[5] A. Sebastian, et al. Memory devices and applications for in-memory computing. Nature Nanotechnology, 2020.

[6] G. Karunaratne, et al. In-memory hyperdimensional computing. Nature Electronics, 2020.

**Please contact:**

Abbas Rahimi, IBM Research Europe,
Säumerstrasse 4, 8803 Rüschlikon,
Switzerland
+41 44 724 8303, abr@zurich.ibm.com

# E = AI$^2$

by Marco Breiling, Bijoy Kundu (Fraunhofer Institute for Integrated Circuits IIS) and Marc Reichenbach (Friedrich-Alexander-Universität Erlangen-Nürnberg)

***How small can we make the energy consumed by an artificial intelligence (AI) algorithm plus associated neuromorphic computing hardware for a given task? That was the theme of a German national competition on AI hardware-acceleration, which aimed to foster disruptive innovation. Twenty-seven academic teams, each made up of one or two partners from universities and research institutes, applied to enter the competition. Two of the eleven teams that were selected to enter were Fraunhofer IIS: ADELIA and Lo3-ML (the latter together with Friedrich-Alexander-University Erlangen-Nürnberg - FAU) [L1]. Finally Lo3-ML was one of the four national winners awarded by the German research minister Anja Karliczek for best energy efficiency.***

A national competition, sponsored by the German research ministry BMBF [L2], challenged teams of researchers to classify two-minute-long electro-cardiogram (ECG) recordings as healthy or showing signs of atrial fibrillation. The two teams involving Fraunhofer IIS ran completely independently of each other and followed completely different approaches. “ADELIA” implemented a mixed-signal neural network (NN) accelerator, which is primarily made of crossbar arrays of novel analogue processing units (see Figure 1) and standard SRAM cells for storage, while “Lo3-ML” designed a basically digital accelerator with non-volatile Resistive RAM (RRAM) cells, for which analogue write-read-circuits were developed.

In the ADELIA project, Fraunhofer IIS developed an analogue NN accelerator with a mixed-signal frontend. The NN is trained using quantized weights of seven levels (quasi-3bit), quantized gain (4 bit) and offset (5 bit) of a batch normalisation (BN), and a customised ReLU activation function (AF). Crossbar arrays made of novel analogue processing units (APUs) perform the primary computations of the NN: the multiply and accumulate operations (MACs). These APUs contain resistors and switches plus standard CMOS SRAM cells, which hold the trained

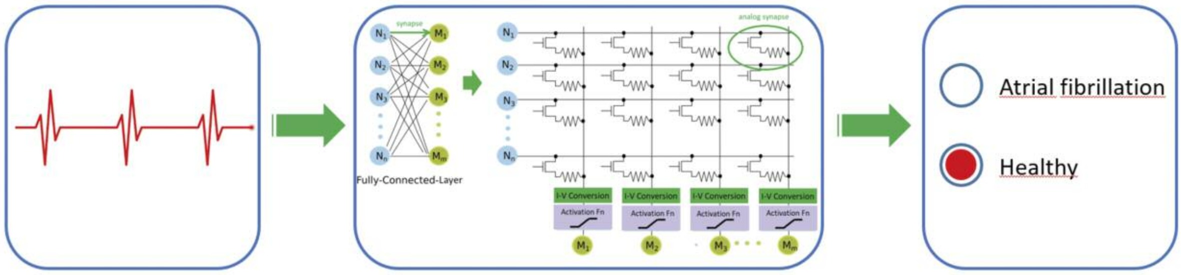


*Figure 1: An analogue deep learning accelerator uses an analogue crossbar for energy-efficient analogue matrix multiplication, which is at the core of the neural network calculation for classifying the ECG signals. (Sources: Fraunhofer IIS and Creative Commons CC0).*

weights after chip start-up. The non-linear activation circuits drive (in analogue) their following NN layer, thus avoiding any conversion overhead from ADCs and DACs. Hardware-software co-design was heavily employed during training and circuit implementation of the NN. While the NN training iteratively took into account several hardware constraints such as quantized weights, quantized gains and offsets of BN, customised AF, and some component variations, the NN circuit was generated automatically from the software model using a tool developed during the project. Thus a fast analogue CMOS based energy efficient NN accelerator was developed that harnesses the parallel processing of analogue crossbar computing without inter layer data converters and without the need for memristor technology.

The second project, Lo3-ML, was a collaboration between Fraunhofer IIS and the chairs for Computer Architecture and Electronics Engineering at FAU. The project built upon their previous work with non-volatile RRAMs [2]. These allow the accelerator core of the chip to be powered down while idle (which is most of the time) to save energy, while the pre-processing core is always awake and collects relatively slowly incoming data (e.g., at low sampling frequency). Once sufficient data is present, the accelerator core is woken up and does a quick processing before powering down again. This scheme saves up to 95% of energy compared to a conventional architecture, where both cores are always on. An RRAM cell can store binary or multi-level values. A trade-off analysis showed that using ternary values both in the RRAM cells and as the weights of a severely quantized NN offered the best energy trade-off. However, for such extreme quantization we had to develop dedicated hardware-aware training algorithms, as corresponding tools were not available at that time. Moreover, to cope with this quantization, a Binary-Shift Batch Norm (BSBN) using only the shifting of values was introduced. The design and implementation of the chip was done iteratively. In order to achieve quick turn-around cycles, the design space exploration for both NN and HW architecture was partially automated [1] including, for example, automatic generation of a bit-true simulation.

The projects were realised by intensive work in small teams over just 15 months. Despite operating as completely independent projects, both ended up selecting similar NNs with multiple convolutional layers followed by fully connected layers. The achieved metrics are 450 nJ per inference for ADELIA on a 22 nm Global Foundries-FDSOI technology and 270 nJ for Lo3-ML on a 130 nm IHP process. ADELIA produced one of the very few analogue NN accelerators worldwide that are actually ready for chip implementation. Mixed-signal NN acceleration is generally considered as very energy-efficient and hence promising for the next generation of ultra-low-power AI accelerators. Although the task in the project was ECG analysis, the developed architectures can also be used for other applications and could be easily extended for more complex tasks. The design-flows developed in the two projects will be combined in the future and will serve as the basis for a highly automated design space exploration tool jointly for NN topology and HW architecture.

**Links:**

[L1] https://kwz.me/h58
[L2] https://kwz.me/h59

**References:**

[1] J. Knödtel, M. Fritscher, et al.: "A Model-to-Netlist Pipline for Parametrized Testing of DNN Accelerators based on Systolic Arrays with Multibit Emerging Memories", MOCAST Conf., 2020.

[2] M. Fritscher , J. Knödtel, et al.: "Simulating large neural networks embedding MLC RRAM as weight storage considering device variations", Latin America Symposium on Circuits and System, 2021.

**Please contact:**
Marco Breiling
Fraunhofer Institute for Integrated Circuits IIS, Germany
marco.breiling@iis.fraunhofer.de

# Brain-inspired Visual-Auditory Integration Yielding Near Optimal Performance – Modelling and Neuromorphic Algorithms

by Timo Oess and Heiko Neumann (Ulm University)

***Audition equips us with a 360-degree far-reaching sense to enable rough but fast target detection in the environment. However, it lacks the precision of vision when more precise localisation is required. Integrating signals from both modalities to a multisensory audio-visual signal leads to concise and robust perception of the environment. We present a brain-inspired neuromorphic modelling approach that integrates auditory and visual signals coming from neuromorphic sensors to perform multisensory stimulus localisation in real time.***

Multimodal signal integration improves perceptual robustness and fault tolerance, increasing the information gain compared to merely superimposing input streams. In addition, if one sensor fails, the other can compensate for the lost information. In biology, integration of sensory signals from different modalities in different species is a crucial survival factor and it has been demonstrated that such integration is performed in a Bayesian optimal fashion However, multimodal integration is not trivial, since signals from different modalities typically are represented in distinct reference frames and one needs to align these frames to relate inferred locations to the spatial surround. An example is the integration of audio and visual signals. The auditory space is computed by level or time differences of binaural input whereas the visual space is derived from positions related to the sensory receptor surfaces (retina, camera). Robust alignment of such distinct spaces is a key step to exploit the information gain provided by multisensory integration.

In a joint project [L1] between the Institutes of Applied Cognitive Psychology and Neural Information Processing at Ulm University, we investigated the processing of auditory and visual input streams for object localisation with the aim of developing biologically inspired neural network models for multimodal object localisation based on optimal visual-auditory integration as seen in humans and animals. In particular, we developed models of neuron populations and their mutual interactions in different brain areas that are involved in visual and auditory object localisation. These models were behaviourally validated and, in a subsequent step, implemented on IBM's brain-inspired neurosynaptic chip TrueNorth [3]. There, they ran in real-time and controlled a robotic test platform to solve target-driven orientation tasks.

In contrast to vision with its topographic representation of space, auditory space is represented tonotopically, with incoming sound signals decomposed into their frequency components. Horizontal sound source positions cannot directly be assigned a location relative to the head direction. This is because this position is computed from different cues, caused by (i) the temporal and intensity difference of the signals arriving at the left and right ear and (ii) the frequency modulation induced by the head and shape of the ears. Thus, to spatially localise a sound source an agent needs to process auditory signals over a cascade of several stages. We modelled these stages utilising biologically plausible neural mechanisms that

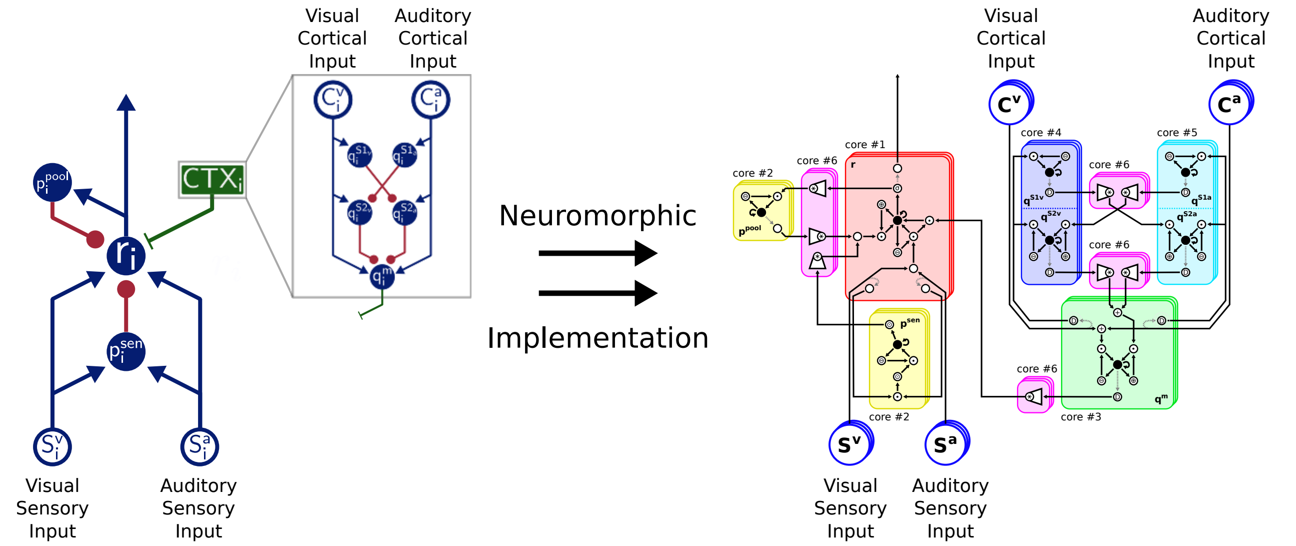


*Figure 1: Model Architectures. Left panel, architecture of the multisensory integration model. Blue lines indicate excitatory connections, green lines indicate modulatory connections, and red lines indicate inhibitory connections. Green box is modulatory cortical signal that is elaborated in grey box. Filled circles represent model neurons, hollow circles indicate inputs to the model. Letters indicate the name of neurons and inputs. Right panel, arrangement of the model architecture on the TrueNorth neurosynpatic chip. Note the similar hierarchy of the architectures with differences only in the fine structure. Adapted from [2], Creative Commons license 4.0 (https://creativecommons.org/licenses/by/4.0/)*

facilitate subsequent deployment on neuromorphic hardware. In particular, we constructed a model for binaural integration of sound signals encoding the interaural level difference of signals from the left and right ear [1]. The model incorporates synaptic adaptation to dynamically optimise estimations of sound sources in the horizontal plane. The elevation of the source in the vertical plane is separately estimated by a model that demonstrates performance enhancement by binaural signal integration. These estimates jointly define a 2D map of auditory space. This map needs to be calibrated to spatial positions in the environment of the agent. Such a calibration is accomplished by learning a registered auditory position map which is guided by visual representations in retinal coordinates. Connection weights are adapted by correlated activities where learning is triggered by a third factor leading to robust map alignment but is sufficiently flexible to adapt to altered sensory inputs.

Such calibrated internal world representations establish accurate and stable multimodal object representations. We modelled multisensory integration neurons that receive excitatory and inhibitory inputs from unimodal auditory and visual neurons, respectively, as well as feedback from cortex [2]. Such feedback projections facilitate multisensory integration (MSI) responses and lead to commonly observed properties like inverse effectiveness, within-modality suppression, and the spatial principle of neural activity in multisensory neurons. In our model, these properties emerge from network dynamics without specific receptive field tuning. A sketch of the functionality of the circuit is shown in Figure 1. We have further investigated how multimodal signals are integrated and how cortical modulation signals affect it. Our modelling investigations demonstrate that near-optimal Bayesian integration of visual and auditory signals can be accomplished with a significant contribution by active cortical feedback projections. In conclusion, the results shed new light how recurrent feedback processing supports near-optimal perceptual inference in cue integration by adaptively enhancing the coherence of representations.

Having analysed the functional properties of our models for audio object localisation and multisensory integration, we deployed the neural models on a neuromorphic robotic platform (Figure 2). The platform is equipped with bio-inspired sensors (DVS and bio-inspired gamma-tone filters) for energy efficient processing of audio and visual signals, respectively, and a neuromorphic processing unit. We defined an event-based processing pipeline from sensory perception up to stages of subcortical multisensory integration. Performance evaluation for real world inputs demonstrate that the neural framework runs in real time and is robust against interferences making it suitable for robotic applications with low-energy consumption.

Further evaluations of this platform are planned, that comprise investigating the ability of the model to flexibly react to altered sensory inputs by, e.g., modifying the position of the microphones. Attention mechanisms will additionally enable a selective enhancement of multisensory signals during active search. Deeper understanding of such principles and testing their function and behaviour on robots provides the basis for developing advanced systems to self-organise stable orientation, localisation, and recognition performance.

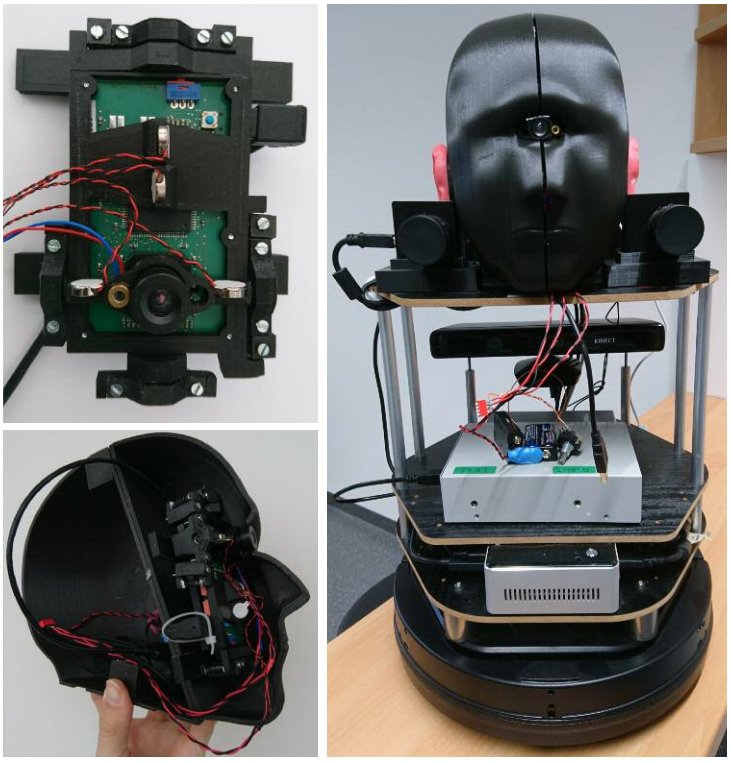


*Figure 2: Left top, eDVS in mounted, vibrational frame driven by vibration motors to induce micro saccades for detection of static objects. Left bottom, left side of 3D printed head with eDVS frame and microphone in ear channel. Right side, complete head mounted on robotic platform.*

**Link:**

[L1] https://www.uni-ulm.de/in/neuroinformatik/forschung/schwerpunkte/va-morph/

**References:**

[1] T. Oess, M. O. Ernst, H. Neumann: "Computational principles of neural adaptation for binaural signal integration", PLoS Comput Biol (2020) 16(7): e1008020. https://doi.org/10.1371/journal.pcbi.1008020

[2] T. Oess, et al.: "From near-optimal Bayesian integration to neuromorphic hardware: a neural network model of multisensory integration", Front Neurorobot (2020) 14:29. doi: 10.3389/fnbot.2020.00029

[3] P. A. Merolla, et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface", Science, vol. 345, no. 6197, pp. 668-673, Aug 2014.

**Please contact:**

Timo Oess, Heiko Neumann
Dept. Applied Cognitive Psychology;
Inst. for Neural Information Processing, Ulm University, Germany
{timo.oess | heiko.neumann}@uni-ulm.de

# Touch in Robots: A Neuromorphic Approach

by Ella Janotte (Italian Institute of Technology), Michele Mastella, Elisabetta Chicca (University of Groningen) and Chiara Bartolozzi (Italian Institute of Technology)

***In nature, touch is a fundamental sense. This should also be true for robots and prosthetic devices. In this project we aim to emulate the biological principles of tactile sensing and to apply it to artificial autonomous systems.***

Babies are born with a grasping reflex, triggered when something touches their palms. Thanks to this reflex, they are able to hold on to fingers and, later on, manually explore objects and their surroundings. This simple fact shows the importance of biological touch for the understanding of the environment. However, artificial touch is less prominent than vision: even tasks such as manipulation, which require tactile information for slip detection, grip strength modulation and active exploration, are widely dominated by vision-based algorithms. There are many reasons for the underrepresentation of tactile sensing, starting from the challenges posed by the physical integration of robust tactile sensing technologies in robots. Here, we focus on the problem of the large amount of data generated by e-skin systems that strongly limits their application on autonomous agents which require low power and data-efficient sensors. A promising solution is the use of event-driven e-skin and on-chip spiking neural networks for local pre-processing of the tactile signal [1].

## Motivation

E-skins must cover large surfaces while achieving high spatial resolution and enabling the detection of wide bandwidth stimuli, resulting in the generation of a large data stream. In the H2020 NeuTouch [L1] project, we draw inspiration from the solutions adopted by human skin.

Coupled with non-uniform spatial sampling (denser at the fingertips and sparser on the body), tactile information can be sampled in an event-driven way, i.e., upon contact, or upon the detection of a change in contact. This reduces the amount of data to be processed and, if merged with on-chip spiking neural networks for processing, supports the development of efficient tactile systems for robotics and prosthetics.

## Neuromorphic sensors

Mechanoreceptors of hairless human skin can be roughly divided into two groups: slowly and rapidly adapting. Slowly adapting afferents encode stimulus intensity while rapidly adapting ones respond to changes in intensity. In both cases, tactile afferents generate a series of digital pulses (action potentials, or spikes) upon contact. This can be applied to an artificial e-skin, implementing neuromorphic, or event-driven, sensors' readout.

Like neuromorphic vision sensors, the signal is sampled individually and asynchronously, at the detection of a change in the sensing element's analogue value. Initially, the encoding strategy can be based on emitting an event (a digital voltage pulse) when the measured signal changes by a given amount with respect to the value at the previous event. We will then study more sophisticated encoding based on local circuits that emulate the slow and fast adaptive afferents. Events are transmitted off-chip asynchronously, via AER-protocol, identifying the sensing element that observed the change. In this representation, time represents itself and the temporal event pattern contains the stimulus information. Thus, the sensor remains idle in periods of no change, avoiding the production of redundant data, while not being limited by a fixed sampling rate if changes happen fast.

We aim to exploit the advantages of event-driven sensing to create a neuro-

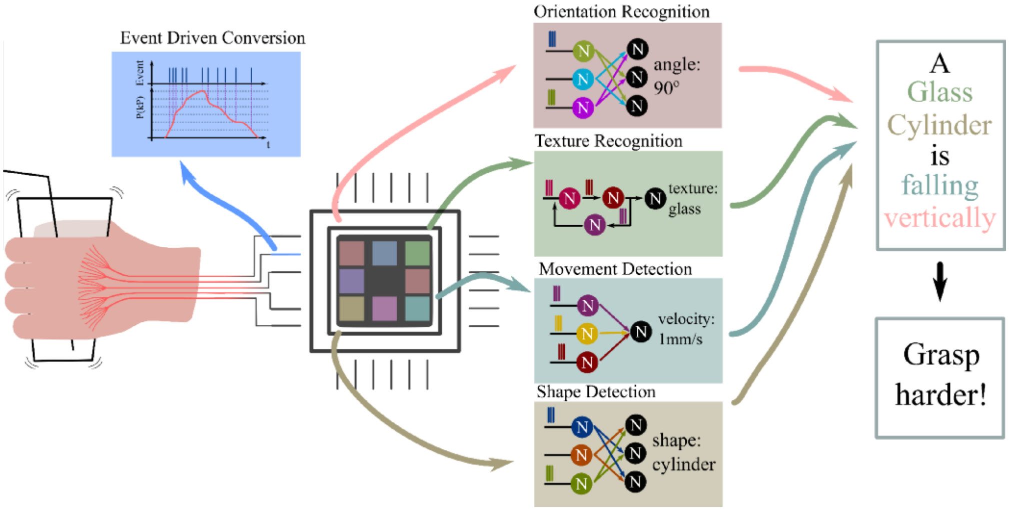


*Figure 1: A graphical representation of the desired outcome of our project. The realised architecture takes information from biologically inspired sensors, interfacing with the environment. The outcoming data are translated into spikes using event-driven circuits and provide input to different parts in the electronic chip. These different parts are responsible for analysing the incoming spikes and delivering information about environmental properties of objects. The responses are then used to generate an approximation about what is happening in the surroundings and impact the reaction of the autonomous agent.*

morphic e-skin that produces a sparse spiking output, to solve the engineering problem of data bandwidth for robotic e-skin. The signal can be then processed by traditional perception modules. However, the spike-based encoding calls for the implementation of spiking neural networks for extracting information.

### Spiking neural networks

The spiking asynchronous nature of neuromorphic tactile encoding paves the way to the use of spiking neural networks (SNNs) to infer information about the tactile stimulus. SNNs use neuron and synapse models that more closely match the behaviour of biology, using spatio-temporal sequences of spike to encode and decode information. Synaptic strength and the connectivity of the networks are shaped by experience, through learning.

Examples of neuromorphic tactile systems that couple event-driven sensing with SNN have been developed for orientation detection, where the coincident activations of several event-driven sensors is used to understand the angle at which a bar, pressed on the skin, is tilted [2]. Different sensors' outputs are joined together and connected to a neuron, when the neuron spikes it signals that those sensors were active together. Another example is the recognition of textures, which can be done by recreating a spiking neural network that senses frequencies [3]. A novel architecture, composed only of neurons and synapses organised in a recurrent fashion, can spot these frequencies and signal the texture of a given material.

These networks can be built in neuromorphic mixed-mode subthreshold CMOS technology, to emulate SNNs using the same technological processes of traditional chips, but exploiting an extremely low-power region of the transistor's behaviour. By embedding the networks directly on silicon using this strategy, we aim for low power consumption, enabled by the neuron's ability to consume energy only when active and to interface directly with event-driven data.

### Conclusion

Our goal is to equip autonomous agents, such as robots or prosthesis, with the sense of touch, using the neuromorphic approach both for sensing and processing. We will employ event-driven sensors to capture temporal information in stimuli and to encode it in spikes and spiking neural networks to process data in real time, with low power consumption. The network will be implemented on a silicon technology, delivering the first neuromorphic chip for touch.

These two novel paradigms have several advantages that will result in a structure capable of exploring the environment with bioinspired and efficient architectures. This combined approach can greatly enhance the world of autonomous agents.

**Link:**
[L1] https://neutouch.eu/

**References:**

[1] C. Bartolozzi, L. Natale, L., Nori, G. Metta: "Robots with a sense of touch", Nature Materials 15, 921–925 (2016).
[2] A. Dabbous, et. al.: "Artificial Bio-inspired Tactile Receptive Fields for Edge Orientation Classification", ISCAS (2021) [in press].
[3] M. Mastella, E. Chicca: "A Hardware-friendly Neuromorphic Spiking Neural Network for Frequency Detection and Fine Texture Decoding", ISCAS (2021) [in press].

**Please contact:**

Ella Janotte, Italian Institute of Technology, iCub facility, Genoa, Italy.
ella.janotte@iit.it

Michele Mastella, BICS Lab, Zernike Inst Adv Mat, University of Groningen, Netherlands
m.mastella@rug.nl

# Uncovering Neuronal Learning Principles through Artificial Evolution

by Henrik D. Mettler (University of Bern), Virginie Sabado (University of Bern), Walter Senn (University of Bern), Mihai A. Petrovici (University of Bern and Heidelberg University) and Jakob Jordan (University of Bern)

***Despite years of progress, we still lack a complete understanding of learning and memory. We leverage optimisation algorithms inspired by natural evolution to discover phenomenological models of brain plasticity and thus uncover clues to the underlying computational principles. We hope this accelerates progress towards deep insights into information processing in biological systems with immanent potential for the development of powerful artificial learning machines.***

What is the most powerful computing device in your home? Maybe your laptop or smartphone spring to mind, or possibly your new home automation system. Think again! With a power consumption of just 10W, our brains can extract complex information from a high-throughput stream of sensory inputs like no other known system. But what enables this squishy mass of intertwined cells to perform the required computations? Neurons, capable of exchanging signals via short electrical pulses called "action potentials" or simply "spikes", are the main carriers and processors of information in the brain. It is the organised activity of several billions of neurons arranged in intricate networks, that underlies the sophisticated behaviour of humans and other animals. However, as physics has long known, the nature of matter is determined by the interaction of its constituents. For neural networks, both biological and artificial, the interaction between neurons is mediated by synapses, and it is their evolution over time that allows sophisticated behaviour to be learned in the first place.

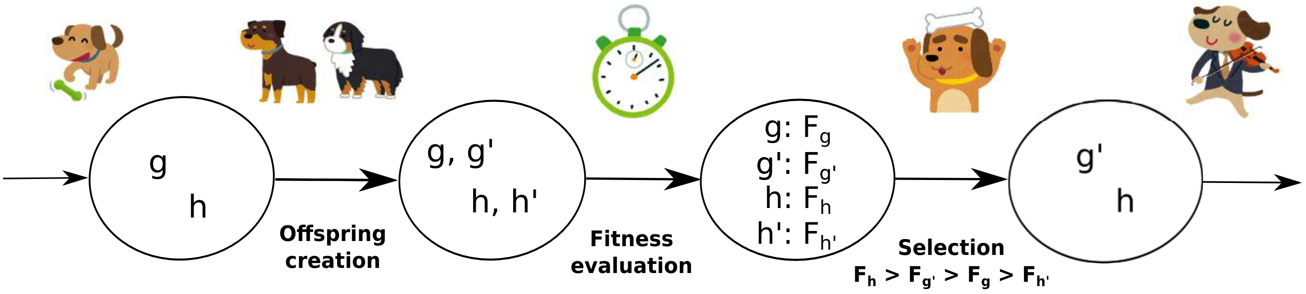


*Figure 1: Schematic overview of our evolutionary algorithm. From an initial population of synaptic plasticity rules (g, h), new solutions (offspring, g', h') are created by mutations. Each rule is then evaluated using a particular network architecture on a predefined task, resulting in a fitness value ($F_g$, $F_g$', $F_h$, $F_h$'). High-scoring rules are selected to become the new parent population and the process is repeated until a plasticity rule reaches a target fitness.*

Synaptic plasticity describes how and why changes in synaptic strengths take place. Early work on uncovering general principles governing synaptic plasticity goes back to the 1950s, with one of the most well-publicized results being often summarised by the mnemonic “what fires together, wires together”. However, this purely correlation-driven learning is but one aspect of a much richer repertoire of dynamics espoused by biological synapses. More recent theories of synaptic plasticity have therefore incorporated additional external signals like errors or rewards. These theories connect the systems-level perspective (“what should the system do?”) with the network level (“which dynamics should govern neuronal and synaptic quantities?”). Unfortunately, the design of plasticity rules remains a laborious, manual process and the set of possible rules is large, as the continuous development of new models suggests.

In the Neuro-TMA group at the Department of Physiology, University of Bern we are working on supporting this manual process with powerful automated search methods. We leverage modern evolutionary algorithms to discover various suitable plasticity models that allow a simulated neuronal network architecture to solve synthetic tasks from a specific family, for example to navigate towards a goal position in an artificial two-dimensional environment. In particular, we use genetic programming, an algorithm for searching through mathematical expressions loosely inspired by natural evolution (Figure 1), to generate human-interpretable models. This assures our discoveries are amenable to intuitive understanding, fundamental for successful communication and human-guided generalisation. Furthermore, this interpretability allows us to extract the key interactions between biophysical variables giving rise to plasticity. Such insights provide hints about the underlying biophysical processes and also suggest new approaches for experimental neuroscience (Figure 2).

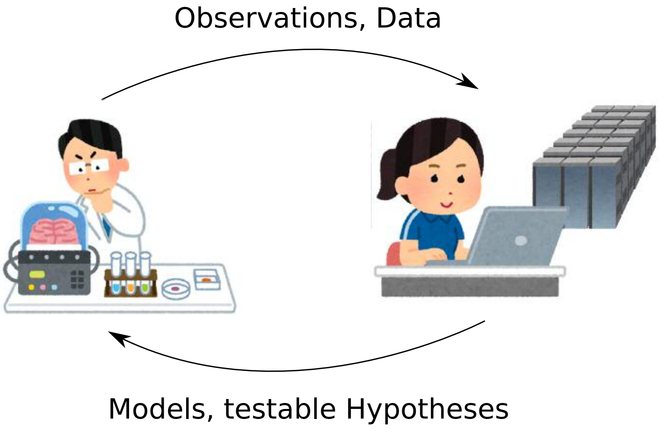


*Figure 2: Symbiotic interaction between experimental and theoretical/computational neuroscience. Experimental neuroscientists provide observations about single neuron and network behaviours. Theoretical neuroscientists develop models to explain the data and develop experimentally testable hypotheses, for example about the time evolution of neuronal firing rates due to ongoing synaptic plasticity.*

Two of our recent manuscripts have highlighted the potential of our evolving-to-learn (E2L) approach by applying it to typical learning scenarios in both spiking and rate-based neuronal network models. In [1], we discovered previously unknown mechanisms for learning efficiently from rewards, recovered efficient gradient-descent methods for learning from errors, and uncovered various functionally equivalent spike-timing-dependent-plasticity rules with tuned homeostatic mechanisms. In [2], we demonstrated how E2L can incorporate statistical properties of the dataset to evolve plasticity rules that learn faster than some of their more general, manually derived counterparts.

Since our approach requires a large number of neuronal network simulations, we make use of modern HPC infrastructure, such as Piz Daint at the Swiss National Supercomputing Centre (CSCS) as well as high-performance software for the simulation of neuronal networks [3] and from the Scientific Python ecosystem. To support the specific needs of our research we have developed an open-source pure-Python library for genetic programming [L1]. We believe that the open nature of such community codes holds significant potential to accelerate scientific progress in the computational sciences.

In the future, we will explore the potential of neuromorphic systems, dedicated hardware for the accelerated simulation of neuronal network models. To this end, we collaborate closely with hardware experts at the Universities of Heidelberg, Manchester and Sussex.

In summary, our E2L approach represents a powerful addition to the neuroscientist's toolbox. By accelerating the design of mechanistic models of synaptic plasticity, it will contribute not only new and computationally powerful learning rules, but, importantly, also experimentally testable hypotheses for synaptic plasticity in biological neuronal networks. This effectual loop between theory and experiments will hopefully go a long way towards unlocking the mysteries of learning and memory in healthy and diseased brains.

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3). We would like to thank Maximilian Schmidt for our continuous collaboration and the Manfred Stärk foundation for ongoing support.

**Links:**
[L1] https://kwz.me/h5f
[L2] https://kwz.me/h5S
[L3] (Graphics) www.irasutoya.com

**References:**

[1] J. Jordan, et al.: "Evolving to learn: discovering interpretable plasticity rules for spiking networks", 2020. arXiv:q-bio.NC/2005.14149

[2] H.D. Mettler et al.: "Evolving Neuronal Plasticity Rules using Cartesian Genetic Programming", 202, arXiv: cs.NE/2102.04312.

[3] J. Jordan, et al.: "Extremely Scalable Spiking Neuronal Network Simulation Code: From Laptops to Exascale Computers, Frontiers in Neuroinformatics, 12, 2018, doi:10.3389/fninf.2018.00002

**Please contact:**

Henrik D. Mettler,
NeuroTMA group, Department of Physiology, University of Bern
henrik.mettler@unibe.ch

# What Neurons Do – and Don't Do

by Martin Nilsson (RISE Research Institutes of Sweden) and Henrik Jörntell (Lund University, Department of Experimental Medical Science)

***Biology-inspired computing is often based on spiking networks, but can we improve efficiency by going to higher levels of abstraction? To do this, we need to explain the precise meaning of the spike trains that biological neurons use for mutual communication. In a cooperation between RISE and Lund University, we found a spectacular match between a mechanistic, theoretical model having only three parameters on the one hand, and in vivo neuron recordings on the other, providing a clear picture of exactly what biological neurons "do", i.e., communicate to each other.***

Most neuron models are empirical or phenomenological because this allows a comfortable match with experimental data. If nothing else works, additional parameters can be added to the model until it fits data sufficiently well. The disadvantage of this approach is that an empirical model cannot *explain* the neuron – we cannot escape the uneasiness of perhaps having missed some important hidden feature of the neuron. Proper explanation requires a mechanistic model, which is instead based on the neuron's underlying biophysical mechanisms. However, it is hard to find a mechanistic model at an appropriate level of detail that matches data well. What we ultimately want, of course, is to find a simple mechanistic model that matches the data as well as any empirical model.

We struggled for considerable time to find such a mechanistic model of the cerebellar Purkinje neuron (Figure 1a), which we use as a model neuron system. Biological experiments revealed, at an early stage, that the neuron low-pass filters input heavily, so clearly, the cause of the high-frequency component of the interspike interval variability could not be the input, but was to be found locally. The breakthrough came with the mathematical solution of the long-standing first-passage time problem for stochastic processes with moving boundaries [1]. This method enabled us to solve the model equations in an instant and allowed us to correct and perfect the model using a large amount of experimental data.

We eventually found that the neuron's spiking can be accurately characterised by a simple mechanistic model using only three free parameters. Crucially, we found that the neuron model necessarily requires three compartments and must be stochastic. The division into three compartments is shown in Figure 1b, having a distal compartment consisting of the dendrite portions far from soma; a proximal compartment consisting of soma and nearby portions of the dendrites; and an axon initial segment compartment consisting of the initial part of the axon. One way to describe the model is as trisecting the classical Hodgkin-Huxley model by inserting two axial resistors and observing the stochastic behaviour of ion channels in the proximal compartment.

From the theoretical model we could compute the theoretical probability distribution of the interspike intervals (ISIs). By comparing this with the ISI histograms (and better, the kernel density estimators) of long (1,000–100,000 ISIs) in vivo recordings, the accuracy was consistently surprisingly high (Figure 1c). Using the matching to compute model parameters as an inverse problem, we found that the error was within a factor two from the Cramér-Rao lower bound for all recordings.

It seems that the distal compartment is responsible for integrating the input; the proximal compartment generates a ramp and samples it, and the axon initial segment compartment detects a threshold passage and generates the spike.

We have concluded that the neuron function appears rather straightforward, and that this indicates that there is potential to proceed beyond the spiking level towards higher levels of abstraction, even for biological neurons. The fundamentally inherent stochasticity of the neuron is unavoidable, and this must be taken into account, but there is no need to worry excessively about hidden yet undiscovered neuron features that would disrupt our view of what neurons are capable of. The reason is the nearly perfect match between model and data; the match cannot be significantly improved even if the model is elaborated, which we show using the Cramér-Rao lower bound.

The major limitation is that we assume stationary input. This is by design, because we want to eliminate uncontrollable error sources such as cortical input. In the cerebellum, this can be achieved experimentally by decerebration. However, by observing the distal compartment's low-pass filtering properties, it is straightforward to generalize the model to accept non-stationary input using a pseudo-stationary approach.

So, what do the neurons do, then? In brief, it turns out that the Purkinje neurons first soft-threshold the internal potential, and then encode it using pulse frequency modulation, dithered by channel noise to reduce distortion. And this is it! One of the three free parameters is the input, and the other two correspond to the soft-threshold function's gain (slope) and offset (bias), respectively. Please refer to [2] for details, or to [L1] for a popular description without formulas.

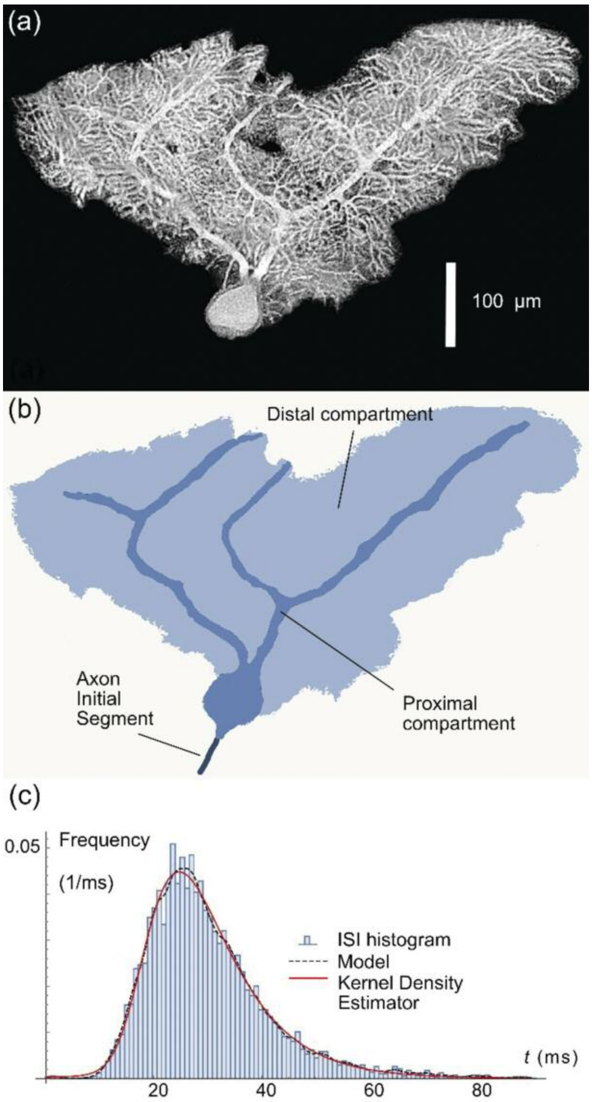


*Figure 1: (a) Confocal photomicrograph of Purkinje neuron. (b) Proposed division of neuron into three compartments. (c) Example of the match between theoretical model (red solid trace) and experimental interspike-interval histogram (blue bars; black dashed trace for kernel density estimator). Image credits: CC BY 4.0 [2].*

From a technical point of view, it is interesting to note that the soft-thresholding function is nearly identical to the rectifying linear unit (ReLU), and even more so to the exponential, or smooth soft-thresholding function that has recently received much attention in the machine learning field.

The next step is to investigate the implications for ensembles of neurons. Is it possible to formulate an abstraction which treats such assemblies of neurons as a single unit without considering each neuron independently?

This research was partially funded by the European Union FP6 IST Cognitive Systems Initiative research project SENSOPAC, "SENSOrimotor structuring of Perception and Action for emerging Cognition", and the FP7 ICT Cognitive Systems and Robotics research project THE, "The Hand Embodied", under grant agreements 028056 and 248587, respectively.

**Link:**

[L1] https://www.drnil.com/#neurons-doing-what (retrieved 2021-03-16)

**References:**

[1] M. Nilsson: "The moving-eigenvalue method: Hitting time for Itô processes and moving boundaries", Journal of Physics A: Mathematical and Theoretical (Open Access), October 2020. DOI: 10.1088/1751-8121/ab9c59

[2] M. Nilsson and H. Jörntell: "Channel current fluctuations conclusively explain neuronal encoding of internal potential into spike trains", Physical Review E (Open Access), February 2021. DOI: 10.1103/PhysRevE.103.022407

**Please contact:**

Martin Nilsson, RISE Research Institutes of Sweden,
martin.nilsson@ri.se

# NEUROTECH - A European Community of Experts on Neuromorphic Technologies

by Melika Payvand, Elisa Donati and Giacomo Indiveri (University of Zurich)

***Neuromorphic Computing Technology (NCT) is becoming a reality in Europe thanks to a coordinated effort to unite the EU researchers and stakeholders interested in neuroscience, artificial intelligence, and nanoscale technologies.***

Artificial intelligence (AI) has been achieving impressive results in a wide range of tasks. Progress in AI is producing ever more complex and powerful algorithms. However, the process of training and executing these algorithms on standard computing technologies consumes vast amounts of energy and is not sustainable in the long term. A promising approach to building a new generation of AI computing technologies that can dramatically reduce power consumption is "neuromorphic computing and engineering". This ground-breaking approach draws inspiration from biological neural processing systems. Today's neuromorphic research community in Europe is already leading the state of the art. The NEUROTECH project has been extremely successful in fostering this community, promoting its growth, and engaging with stakeholders to enable the uptake of this technology in industry globally. The NEUROTECH services, webinars, forums, educational and public outreach activities are attracting interest from both research institutions and small, medium, and large enterprises worldwide.

NEUROTECH is an EU coordination and support action that aims to generate mutual awareness among organisations, national networks, projects, and initiatives in neuromorphic computing and engineering, and to promote the exchange of tools, solutions, ideas, and, where possible, IP among all EU stakeholders. The participating institutions are the University of Zurich (Switzerland, coordinator), University of Manchester (UK), Heidelberg University (Germany), University of Groningen (Netherlands), Italian Institute of Technology (IIT, Italy), University of Bordeaux (France), University of Hertfordshire (UK), THALES SA (France), IBM Research GmbH (Switzerland), Consiglio Nazionale Delle Ricerche (CNR, Italy), Interuniversitair Micro-Electronica Centrum vzw (IMEC, Belgium) and Commissariat a l'Energie Automatique et aux Energies Alternatives (CEA LETI, France).

The project is a coordination action that organises a variety of events around the topic of NCTs. The various events include monthly educational activities, where experts are invited to answer a specific question on NCTs, e.g., supporting technologies for neuromorphic processing, event-driven sensors, and processors, and circuits for online learning. Every event is recorded and uploaded in an online channel, which acts as a collection of introductory material for students and researchers who want to get involved in neuromorphic computing. The project is composed of four subgroups: Industry, Bridge, Science, and Ethics, which aim to create a connection with industry and other communities, and to define guidelines for ethics in neuromorphic computing and engineering. Each workgroup organises periodic webinars and panel discussions to increase awareness within the community and with a large audience. With the exception of an initial forum, held in Leuven, Belgium, in 2019, all these events are currently being held online.

The NEUROTECH web portal [L1] aims to create a web presence for the project that can be used to make both

We create and lead the **Neuromorphic Computing Technology community in Europe**, by catalysing research and collaboration.

the NCT community and potential partners outside the core community aware of the project and its plans. In addition, the web portal collects useful resources for the community, such as datasets, references (papers, books, courses), and deliverables. The web portal also provides a Roadmap [1] that defines neuromorphic computing and its next steps, a state-of-the-art document that collects knowledge about the latest breakthroughs in NCTs, and a cartography document that collects links from different national and international networks and organisations in Europe to facilitate the growth of the NEUROTECH network. It also provides information about upcoming events – both organised and supported by the project and other relevant events worldwide, and provides a feature for subscribing to the NEUROTECH mailing list through which relevant news about the NCT is announced. NEUROTECH Forum II that was held online on March 15th 2021 provided an overview on the opportunities of AI, technology trends, and prospects. Dedicated discussions were held on enabling SME's to step into neuromorphic computing and AI as well as on AI induced ethical questions. The forum attracted more than 1000 views, with 150 attendees at all times for the entire day from all over the world.

The project started in November 2018 and will continue until the end of 2022. Future activities of the project revolve around the growth of the community and bridge between fields that can influence and benefit from NCT. These activities include, but are not limited to, industrial panel discussions about NCT opportunities and challenges, scientific discussions about new discoveries and the development of materials and algorithms that empower NCT, and discussions on the ethical aspects of creating NCT and its applications.

The upcoming events include the educational events that happen monthly, and the different workgroup activities that are scheduled regularly. For example, the Science workgroup event is planned on April 27th for bringing together the coordinators of the recently accepted neuromorphic EU projects to answer the question How novel technologies can boost neuromorphic computing? A view from European project consortia.

**Link:**
L1] http://neurotechai.eu

**Reference:**

[1] E. Donati et al.: "Neuromorphic technology in Europe : Brain-inspired technologies are advancing apace across Europe and are poised to help accelerate the AI revolution", The Innovation Platform, 2020

**Please contact:**

Melika Payvand, Elisa Donati, Giacomo Indiveri, University of Zurich, Switzerland
+41 44 635 3024
(melika, elisa, giacomo)@ini.uzh.ch

# NeuroAgents – Autonomous Intelligent Agents that Interact with the Environment in Real Time

by Giacomo Indiveri (University of Zurich and ETH Zurich)

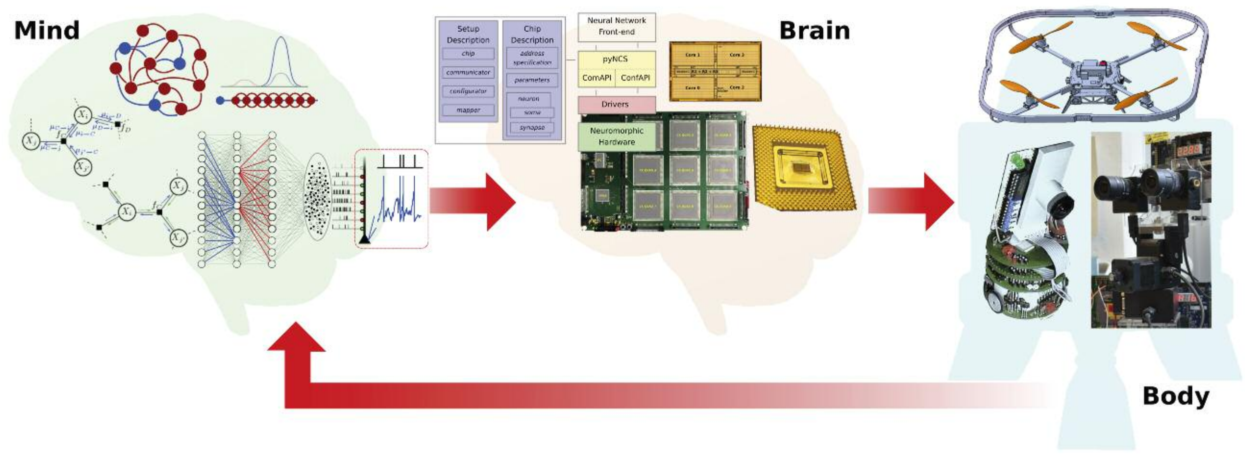
***Artificial intelligence systems might beat you in a game of Go, but they still have serious shortcomings when they are required to interact with the real world. The NeuroAgents project is developing autonomous intelligent agents that can express cognitive abilities while interacting with the environment.***

Tremendous progress is being made globally in machine learning, artificial intelligence (AI), and deep learning. In parallel, especially in Europe, there is a strong convergence of AI algorithms, neuroscience basic research and technology development.

However, despite this remarkable progress, artificial systems are still unable to compete with biological systems in tasks that involve processing sensory data acquired in real time in complex settings and interacting with the environment in a closed-loop setup. The NeuroAgents project is bridging this gap with basic and applied research in "neuromorphic" computing and engineering. Our team is combining the recent advancements in machine learning and neural computation with the latest developments in microelectronic and emerging memory technologies. We are designing autonomous systems that can express robust cognitive behaviour while interacting with the environment, through the physics of their computing substrate. In particular, the project is making inroads in understanding how to build "neuromorphic cognitive agents", i.e., fully fledged artificial systems that can carry out complex tasks in uncontrolled environments, with low-level adaptive sensory processing abilities and high-level goal-oriented decision-making skills.

The project has two ambitious, tightly interlinked goals: (i) to study the principles of computation used by animal brains for building a disruptive neuromorphic computing technology, and (ii) to build electronic signal processing systems that use the same physics of computation used by biological neurons to better understand how animal brains compute.

The project started in September 2017 and is currently entering its final stages of research. The studies of neural computation and AI algorithms (the "mind" of the project) paralleled the design of a new neuromorphic processor chip (the "brain" of the project), which has been recently delivered and is now being tested. Meanwhile the "mind" algorithms and architectures were mapped on older generation neuromorphic processors produced in the previous



NeuroP ERC project, and interfaced with robotic actuators (the "body" part of this project).

The techniques employed combine the study of neural dynamics in recurrent networks of neurons measured from neuroscience experiments, with the simulation of spiking neural processing architectures on computers, and with the design of mixed-signal analogue/digital electronic circuits that emulate the properties of real neurons and synapses [1,2]. The theoretical neuroscience studies focus on perception, memory storage, decision-making and motor planning. The software simulations take the high-level results obtained from the theory and implement them using software spiking neural network simulators. The software is designed to incorporate the "features" of the electronic circuits used into the simulated neural networks. These features include restrictions such as limited resolution, noise, or the requirement to clip signals to only positive values (e.g., because voltages cannot go below the "ground" reference level and currents can only flow in one direction). The microelectronic techniques focus on the design of spike-based learning and plasticity mechanisms at multiple time scales, and of asynchronous event-based routing schemes for building large-scale networks of spiking neurons. The robotics activities include testing the models and the spiking neural network chips with neuromorphic sensors and motorised platforms (e.g., in active vision stereo setups) [3].

Future activities of the project involve connecting all the pieces together (i.e., mind, brain, and body), and applying the results and the know-how gained to practical "edge-computing" applications, i.e., applications in which the agent needs to respond to the data that is being sensed in real-time, with low power and without having to resort to cloud-based computing resources.

The final goal of developing brain-inspired sensing, processing, and cognitive architectures, all implemented with spiking neural network chips, is indeed within reach and the NeuroAgents project promises to produce very interesting results.

NeuroAgents is an ERC Consolidator project (No. 724295) hosted at the Institute of Neuroinformatics at the University of Zurich and ETH Zurich, Switzerland. The project has led to collaboration with many EU partners, including IBM Research Zurich; IIT Genova, Italy; CEA-LETI Grenoble, France; the University of Groningen, the Netherlands; Newcastle University, UK; or the SME SynSense AG, Zurich Switzerland.

**References:**

[1] G. Indiveri and Y. Sandamirskaya: "The importance of space and time for signal processing in neuromorphic agents: the challenge of developing low-power, autonomous agents that interact with the environment", IEEE Signal Processing Magazine 36.6 (2019): 16-28.

[2] A. Rubino, et al.: "Ultra-Low-Power FDSOI Neural Circuits for Extreme-Edge Neuromorphic Intelligence", IEEE Transactions on Circuits and Systems I: Regular Papers, 2020.

[3] N. Risi, et al: "A spike-based neuromorphic architecture of stereo vision", Frontiers in neurorobotics 14, (2020): 93.

**Please contact:**

Giacomo Indivieri
University of Zurich
giacomo@ini.uzh.ch

# Human-like AI

by Dave Raggett, (W3C/ERCIM)

***Human-like general purpose AI will dramatically change how we work, how we communicate, and how we see and understand ourselves. It is key to the prosperity of post-industrial societies as human populations shrink to a sustainable level. It will further enable us to safely exploit the resources of the solar system given the extremely harsh environment of outer space.***

Human-like AI seeks to mimic human memory, reasoning, feelings and learning, inspired by decades of advances across the cognitive sciences, and over 500 million years of neural evolution since the emergence of multicellular life. Human-like AI can be realised on conventional computer hardware, complementing deep learning with artificial neural networks, and exploited for practical applications and ideas for further research.

The aim is to create cognitive agents that are knowledgeable, general purpose, collaborative, empathic, sociable and trustworthy. Metacognition and past experience will be used for reasoning about new situations. Continuous learning will be driven by curiosity about the unexpected. Cognitive agents will be self-aware in respect to current state, goals and actions, as well as being aware of the beliefs, desires and intents of others (i.e. embodying a theory of mind). Cognitive agents will be multilingual, interacting with people using their own languages.

Human-like AI will catalyse changes in how we live and work, supporting human-machine collaboration to boost productivity, either as disembodied agents or by powering robots to help us in the physical world and beyond. Human-like AI will help people with cognitive or physical disabilities – which in practice means most of us when we are old.

We sometimes hear claims about existential risks from strong AI as it learns to improve itself resulting in a superintelligence that rapidly evolves and nolonger cares about human welfare. To avoid this fate, we need to focus on responsible AI that learns and applies human values. Learning from human behaviour avoids the peril of unforeseen effects from prescribed rules of behaviour.

Human-like AI is being incubated in the W3C Cognitive AI Community Group [L1], along with a growing suite of web-based demos, including counting, decision trees, industrial robots, smart homes, natural language, self-driving cars, a browser sandbox and test suite, and an open-source JavaScript chunks library. The approach is based upon the cognitive architecture depicted in Figure 1:

- Perception interprets sensory data and places the resulting models into the cortex. Cognitive rules can set the context for perception, and direct attention as needed. Events are signalled by queuing chunks to cognitive buffers to trigger rules describing the appropriate behaviour. A prioritised first-in first-out queue is used to avoid missing closely spaced events.
- Emotion is about cognitive control and prioritising what's important. The limbic system provides rapid assessment of situations without the delays incurred in deliberative thought. This is sometimes referred to as System 1 vs System 2.
- Cognition is slower and more deliberate thought, involving sequential execution of rules to carry out particular tasks, including the means to invoke graph algorithms in the cortex, and to invoke operations involving other cognitive systems. Thought can be expressed at many different levels of abstraction.
- Action is about carrying out actions initiated under conscious control, leaving the mind free to work on other things. An example is playing a musical

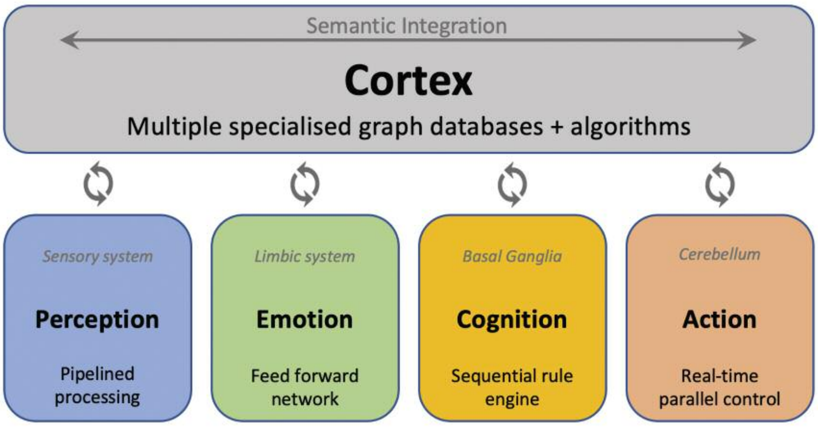


*Figure 1: Cognitive architecture with multiple cognitive circuits equivalent to a shared blackboard.*

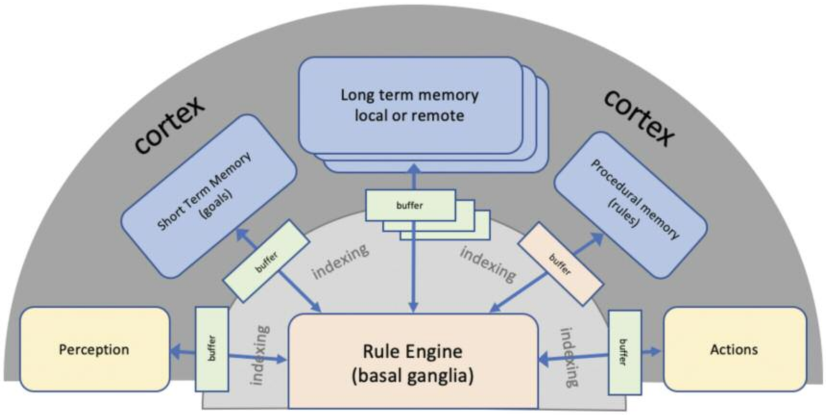


*Figure 2: Cortico-basal ganglia circuit for cognition.*

instrument where muscle memory is needed to control your finger placements as thinking explicitly about each finger would be far too slow.

Zooming in on cognition, we have the following architecture, which derives from work by John Anderson on ACT-R [L2, 1]. The buffers each hold a single chunk, where each chunk is equivalent to the concurrent firing pattern of the bundle of nerve fibres connecting to a given cortical region. This works in an analogous way to HTTP, with buffers acting as HTTP clients and the cortical modules as HTTP servers. The rule engine sequentially selects rules matching the buffers and either updates them directly or invokes cortical operations that asynchronously update the buffers.

A lightweight syntax has been devised for chunks as collections of properties for literals or names of other chunks, and equivalent to n-ary relationships in RDF. Chunks provide a combination of symbolic and sub-symbolic approaches, with graphs + statistics + rules + algorithms. Chunk modules support stochastic recall analogous to web search. Chunks enable explainable AI/ML and learning with smaller datasets using prior knowledge and past experience.

Symbolic AI suffers from the bottleneck caused by reliance on manual knowledge engineering. To overcome this challenge, human-like AI will mimic how children learn in the classroom and playground. Natural language is key to human-agent collaboration, including teaching agents new skills. Human languages are complex yet easily learned by children [2], and we need to emulate that for scalable AI. Semantics is represented as chunk-based knowledge graphs in contrast to computational linguistics and deep learning, which use large statistics as a weak surrogate. Human-like AI doesn't reason with logic or statistical models, but rather with mental models of examples and the use of metaphors and analogies, taking inspiration from human studies by Philip Johnson-Laird [L3, 3].

Humans are good at mimicking each other's behaviour. For example: babies learning to smile socially at the age of 6 to 12 weeks. Language involves a similar process of mimicry with shared rules and statistics for generation and understanding. Work on cognitive natural language processing is focusing on the end-to-end communication of meaning, with constrained working memory and incremental processing, avoiding backtracking through concurrent processing of syntax and semantics.

Work on human-like AI is still in its infancy but is already providing fresh insights for building AI systems, combining ideas from multiple disciplines. It is time to give computers a human touch!

An expanded version of this article is available [L4]. The described work has been supported by the Horizon 2020 project Boost 4.0 (big data in smart factories).

**Links:**

[L1] https://www.w3.org/community/cogai/
[L2] http://act-r.psy.cmu.edu/about/
[L3] https://www.pnas.org/content/108/50/19862
[L4] https://www.w3.org/2021/Human-like-AI-article-raggett-long.pdf
[L5] https://www.w3.org/2021/digital-transformation-2021-03-17.pdf

**References:**

[1] J. R. Anderson: "How Can the Human Mind Occur in the Physical Universe?", Oxford University Press, 2007, https://doi.org/10.1093/acprof:oso/9780195324259.001.0001
[2] W. O'Grady: "How Children Learn Language", Cambridge University Press, 2012, https://doi.org/10.1017/CBO9780511791192
[3] P. Johnson-Laird "How We Reason", Oxford University Press, 2012, https://doi.org/10.1093/acprof:oso/9780199551330.001.0001

**Please contact:**

Dave Raggett, W3C/ERCIM
dsr@w3.org

# Graph-based Management of Neuroscience Data: Representation, Integration and Analysis

by Maren Parnas Gulnes (University of Oslo / SINTEF AS), Ahmet Soylu (OsloMet – Oslo Metropolitan University) and Dumitru Roman (SINTEF AS)

***Advances in technology have allowed the amount of neuroscience data collected during brain research to increase significantly over the past decade. Neuroscience data is currently spread across a variety of sources, typically provisioned through ad-hoc and non-standard approaches and formats, and it often has no connection to other relevant data sources. This makes it difficult for researchers to understand and use neuroscience and related data. A graph-based approach could make the data more accessible.***

A graph-based approach for representing, analysing, and accessing brain-related data [1] could be used to integrate various disparate data sources and improve the understandability and usability of neuroscience data. Graph data models and associated graph database management systems provide performance, flexibility, and agility, and open up the possibility of using well-established graph analytics solutions; however, there is limited research on graph-based data representation as a mechanism for the integration, analysis, and reuse of neuroscience data.

We applied our proposed approach to a unique dataset of quantitative neuroanatomical data about the murine basal ganglia – a group of nuclei in the brain essential for processing information related to movement. The murine basal ganglia dataset consists of quantitative neuroanatomical data about basal ganglia found in healthy rats and mice, collected from more than 200 research papers and data repositories [2]. The dataset contains three distinct information types: quantitations (counts), distributions, and cell morphologies. The counts and distributions relate to either entire cells or specific parts of the cell, while the morphologies describe the cell's physical structure. The dataset's primary purpose is for researchers to find and compare neuroanatomical information about the basal ganglia brain regions.

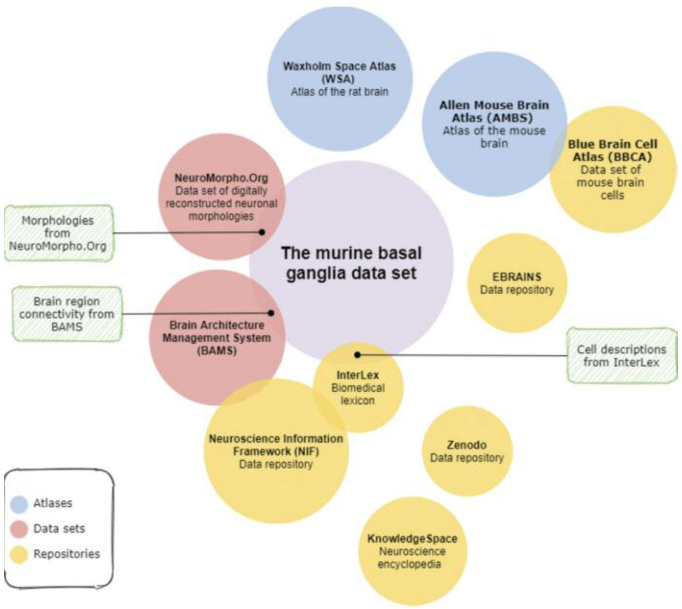

Waxholm Space Atlas (WSA)
Atlas of the rat brain
Allen Mouse Brain Atlas (AMBS)
Atlas of the mouse brain
Blue Brain Cell Atlas (BBCA)
Data set of mouse brain cells
NeuroMorpho.Org
Data set of digitally reconstructed neuronal morphologies
Morphologies from NeuroMorpho.Org
The murine basal ganglia data set
EBRAINS
Data repository
Brain region connectivity from BAMS
Brain Architecture Management System (BAMS)
Cell descriptions from InterLex
InterLex
Biomedical lexicon
Neuroscience Information Framework (NIF)
Data repository
Zenodo
Data repository
KnowledgeSpace
Neuroscience encyclopedia
Atlases
Data sets
Repositories

*Figure 1: An overview of initiatives investigated for overlap with the murine basal ganglia dataset.*

To identify datasets that overlap with the murine basal ganglia dataset for integration purposes, we evaluated a set of related data sources, including repositories, atlases, and publicly available data, against the following criteria: (i) serves data programmatically; (ii) contains data related to the basal ganglia; and (iii) provides data that could be connected to murine basal ganglia. Figure 1 summarises the results of our investigation; Brain Architecture Management System (BAMS) [L1], InterLex [L2], and NeuroMorpho.Org [L3] matched the specified criteria.

We designed and implemented a graph model for the murine basal ganglia dataset and migrated the data from the relational database into a NoSQL graph database [3]. Further, we designed and implemented the integration of data from

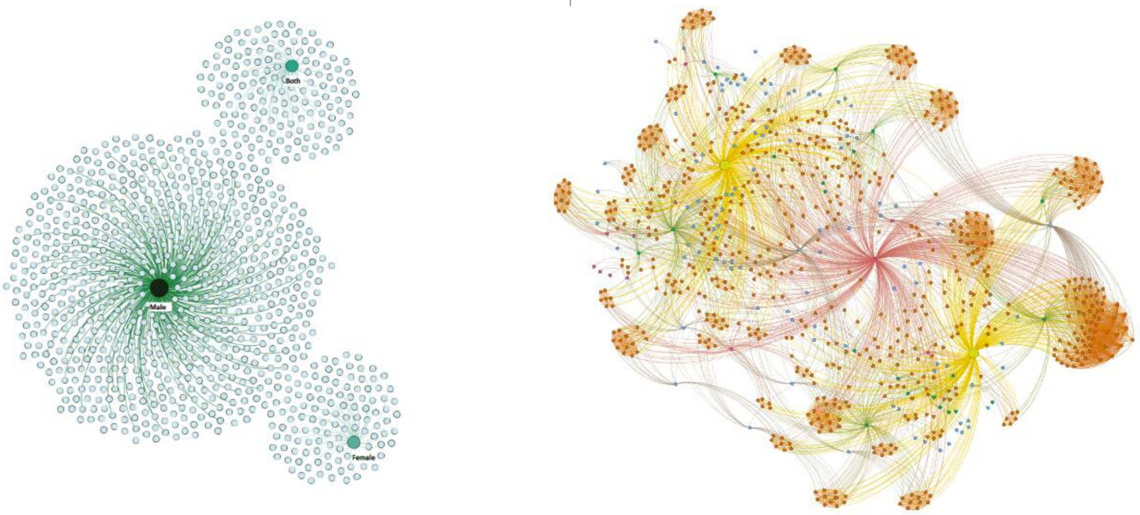


*Figure 2: (a) Relationship between the dataset analyses and the sex and (b) the dataset analyses with related nodes.*

related neuroscience data sources and the technical solution for graph-based data analysis. To provide web-based access to the graph data, we designed and implemented a web application and API. The main components of our approach were: (i) the common graph model based on a native graph database management system; (ii) integration of identified external datasets through an extract-transform-load process; (iii) graph analytics to analyse the graph data, based on existing graph algorithms and visualisation methods; (iv) web-based data access interface for the data based on the graph model. The database generated in this study consists of 9,539 distinct nodes with 46 distinct node labels, 29,807 distinct relationships, and 66 distinct relationship types [L4].

We conducted exploratory and confirmatory data analyses on the data. The former aims to obtain general information about the data and the latter to answer specific questions. For the exploratory analysis, we used community detection algorithms to investigate the graph data structure (Label propagation and Louvain algorithms), a centrality algorithm to find influential nodes in a graph (PageRank), a node similarity algorithm to compare nodes (available in Neo4j), and graph visualisations (ForcedAtlas2) to investigate the general data structure. A consultation with a neuroscience expert revealed that the results include interesting expected and unexpected findings as well as already known findings. For example, Figure 2 (a) shows that males (largest green node) are studied far more often than females. In the confirmatory data analysis part, we aimed to find similar analyses based on a specific criterion using a node similarity algorithm. For example, we searched for studies (i.e., analyses) that investigated the same cell type in the same brain region and with the same object of interest. Figure 2 (b) presents the studies (in orange) in the dataset connected to the specified nodes and species. The yellow nodes represent the two species in the dataset, and the central node in the middle is the cell type “neurons”.

The results and our experience in representing, integrating, and analysing basal ganglia data show that a graph-based approach can be an effective solution, and that the approach should be further considered for management of various types of neuroscience data.

**Links:**
[L1] https://bams1.org
[L2] https://scicrunch.org/scicrunch/interlex/dashboard
[L3] http://neuromorpho.org
[L4] https://github.com/marenpg/jupyter_basal_ganglia

**References:**

[1] R. Angles and C. Gutiérrez: “Survey of graph database models”, ACM Computing Surveys, vol. 40, no. 1, pp. 1–39, 2008.
[2] I.E. Bjerke, et al.: “Database of literature derived cellular measurements from the murine basal ganglia”, Scientific data, vol. 7, no. 1, pp. 1–14, 2020.
[3] R. Cattell: “Scalable SQL and NoSQL data stores”, Sigmod Record, vol. 39, no. 4, pp. 12–27, 2011.

**Please contact:**

Dumitru Roman, SINTEF AS, Norway
dumitru.roman@sintef.no

# The ICARUS Ontology: A General Aviation Ontology

by Joanna Georgiou, Chrysovalantis Christodoulou, George Pallis and Marios Dikaiakos (University of Cyprus)

***A key challenge in the aviation industry is managing aviation data, which are complex and often derived from heterogeneous data sources. ICARUS Ontology is a domain-specific ontology that addresses this challenge by enhancing the semantic description and integration of the various ICARUS platform assets.***

The current digital revolution has heavily influenced the way we manage data. For the past few years every human activity, process, and interaction has been digitised, resulting in an exponential increase in the amount of data produced. There is huge potential for useful information to be extracted from this data, perhaps enabling us to discover new ways of optimising processes, find innovative solutions, and improve decision-making. However, managing enormous amounts of data from numerous, heterogeneous sources that do not share common schemas or standards poses many challenges: data integration and linking remains a major concern.

The procedure of integrating and linking data can be expensive and is often underrated, especially for small and medium-sized enterprises (SMEs), which may lack the required expertise and be unable to invest the necessary time and resources to understand and share the digital information derived from their operational systems. Usually, data models are created in a manner that can handle information by encoding the structure, format, constraints, and relationships with real-world entities.

The challenge of managing big data is apparent in various industry domains, including the aviation industry. Unfortunately, aviation data providers use very distinct data models that can vary across different dimensions [1], like data encoding format, data field naming, data semantics, spatial and temporal resolution, and measurement unit conventions. To enhance the integration of data in a data platform for the aviation industry, we designed and introduced the ICARUS ontology [2].

The ICARUS [L1] platform helps stakeholders that are connected to the aviation industry by providing them with a system to share, collect or exchange datasets, knowledge, and skills. Through these services, the ICARUS project seeks to help stakeholders gain better perspectives, optimise operations, and increase customer safety and satisfaction. Additionally, it aims to offer a variety of user-friendly assets, such as datasets, algorithms, usage analytics and tools. To manage the complex and dynamic evolution of these assets, we sought to develop an ontology to describe various information resources that could be integrated and managed by the ICARUS platform. Ontologies can act as valuable tools to describe and define, in a formal and

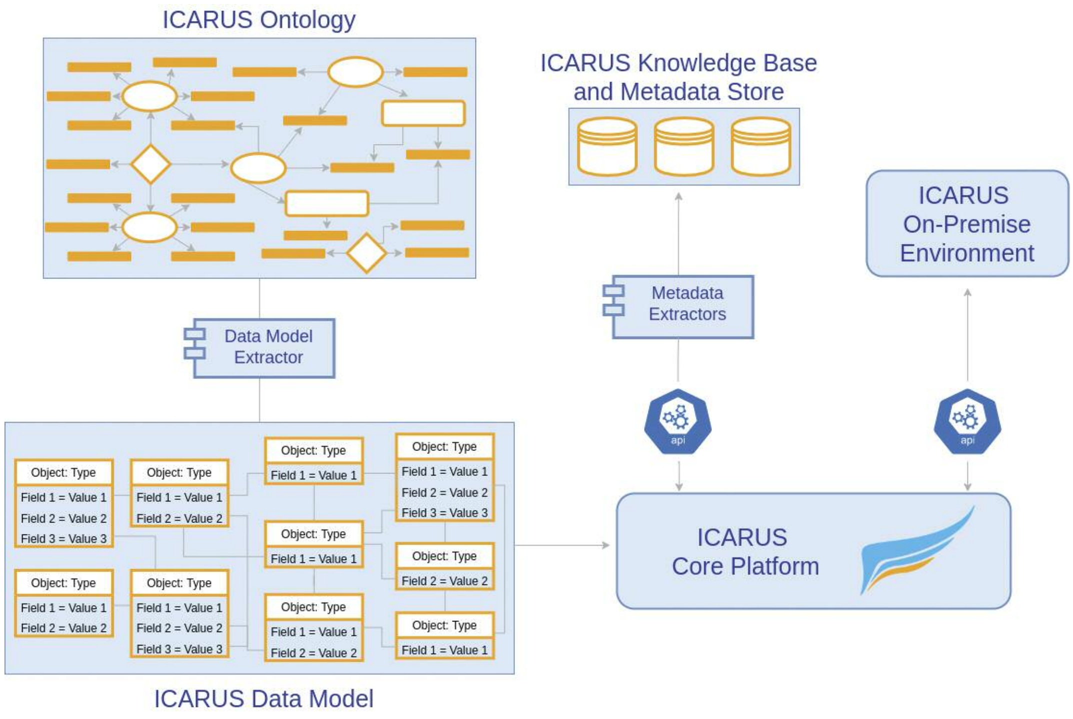


*Figure 1: Ontology and knowledge-base extension of the ICARUS platform architecture.*

explicit manner, relationships and constraints, and for knowledge representation, data integration, and decision making.

The ICARUS ontology, as is depicted in Figure 1, enables: (i) the integration and description of numerous sources of heterogeneous aviation data; (ii) the semantic representation of the metadata generated by the ICARUS platform and the knowledge-based storage, which is dynamically updated and accessed through queries. More specifically, the ICARUS ontology extracts, stores, and represents semantically in the knowledge-based storage all of the platform's datasets and operations. It represents other modules of the ICARUS platform semantically, like deployed algorithms, user interactions, service assets, and their popularity. In addition, the ontology keeps the ICARUS data model and knowledge-based storage up-to-date, while facilitating the ongoing integration of new datasets and services. In this way, it also enables searches and queries of various heterogeneous data sources that are available on the ICARUS platform. Lastly, the ontology provides an application-programming interface to feed valuable knowledge into the ICARUS recommendation engine algorithms.

The ontology can be applied in various scenarios, including the management of epidemic data.  Epidemics, such as the COVID-19 pandemic, present a significant challenge for health organisations when it comes to as locating, collecting and integrating accurate airline and human mobility data with adequate geographical coverage and resolution. If such challenges are addressed well, this can lead to improved epidemic forecasts [3] and potentially enable better estimates of anticipated relative losses of revenue under various scenarios by estimating the reduction of passengers in the airline mobility network. The ICARUS ontology can semantically combine epidemic and aviation-related data for data analytics and epidemic predictions. These analytics and predictions can be extracted from the ontology by completing three key stages: (i) using the ontology to collect aviation and health related data based on concepts / entities stored in the ontology; (ii) combining the datasets based on their relationships as described in the ontology; and (iii) using SPARQL queries to derive new information and insights from the combined datasets.

The main innovative aspects of the ICARUS ontology are: (i) the reuse of existing domain ontologies while adding new concepts, definitions, and relations in order to create a new integrated domain ontology for aviation; and (ii) the use of multiple layers to represent both metadata and aviation-specific concepts. With this approach, the ontology's domain can also be used for other purposes, for instance, general data marketplaces.

The ICARUS ontology can be utilised to perform various analytic SPARQL queries and extract knowledge. It can be found in our github repository [L2] and it is intended to be used by:

- data providers and consumers, who can be stakeholders that are directly or indirectly connected in the aviation value chain industry;
- people from the IT industry who are supporting the aviation value chain. (e.g., IT companies, web entrepreneurs, and software engineers);
- research organisations and universities studying the aviation ecosystem;
- the general public (e.g., passengers);
- other data marketplaces.

ICARUS is a three-and-a-half year project (started 1 January 2018), funded by the European Union's Horizon 2020 research and innovation programme. This work is partially supported by the EU Commission in terms of ICARUS 780792 (H2020-ICT-2016-2017) project and by the Cyprus Research Promotion Foundation in terms of COMPLEMENTARY/0916/0010 project. The authors are

solely responsible for the contents of this publication and can in no way be taken to reflect the views of the European Commission.

**Links:**
[L1] https://www.icarus2020.aero/
[L2] https://github.com/UCY-LINC-LAB/icarus-ontology

**References:**
[1] R. M Keller: "Ontologies for aviation data management", in 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pages 1–9. IEEE, 2016.
[2] D. Stefanidis, et al.: "The ICARUS Ontology: A general aviation ontology developed using a multi-layer approach", in Proc. of the 10th International Conference on Web Intelligence, Mining and Semantics (WIMS), 2020.
[3] C. Nicolaides, et al.: "Hand-Hygiene Mitigation Strategies Against Global Disease Spreading through the Air Transportation Network", Risk Analysis, vol. 40, no. 4, pp. 723–40, 2020.

**Please contact:**
George Pallis, University of Cyprus, Cyprus
gpallis@cs.ucy.ac.cy

# Security Management and the Slow Adoption of Blockchains

by Peter Kieseberg, Simon Tjoa and Herfried Geyer (St. Pölten University of Applied Sciences)

***Blockchains offer a valuable set of tools to provide resilient distributed solutions for applications that require a high level of integrity. Still, apart from being used to store cryptocurrency data, blockchains are not used in many other areas. This article gives an overview of the major obstacles to the uptake of blockchains, from an information security perspective.***

Security management is one of the most important parts of any resilient security strategy for real-life systems. A defender must be able to detect incidents, act coherently and according to best practices, and to learn from attackers in order to develop mitigation strategies for the future.

Several standards and best practices have been devised; the best known of which is arguably the ISO27k family [1]. Standardisation is a particularly important process in security management, not only to address technical issues like compatibility with other organisations, but also as an insurance for the decision-makers within those companies, for when large breaches occur.

Blockchains as ledgers for cryptographic currencies currently exist in a rather special environment, since through their development and the original stakeholders, they typically do not have to adhere to the same levels of accountability and privacy that traditional IT systems do. This approach may work for a system that exists outside of the traditional financial mechanisms, but the scenario is different in the traditional IT landscape, like banking or land registers. In these environments, the inherent risks of this more "free" approach will not be accepted by typical end-users, or even be illegal. Some experts believe that a major reason for the slow acceptance of novel blockchain applications outside the purely speculative area of cryptocurrencies is that trust, accountability and thorough control over the data on the chain are currently unavailable for blockchain mechanisms [2].

One major problem when introducing blockchain-based systems is that they are usually treated in the same way as any other form of IT system with respect to security management. However, they possess several features that differ fundamentally, depending on the exact type of blockchain (e.g., public vs. private, type of consensus mechanism), as well as the implementation (see Figure 1). One major difference is the highly distributed nature of the systems, which is otherwise generally restricted to niche applications or highly unregulated applications like file sharing and bit torrent. Furthermore, unlike most other applications, blockchain-based systems lack a mechanism to remove data from the chain; on the contrary, most blockchain systems do not support the deletion of data at all. This is in stark contrast to most filesharing applications, where data is rendered inaccessible after a while, depending on the interest of the other participants, and at some point, typically dwindles into obscurity. In blockchain applications on the other hand we must assume that the attacker might be in possession of any arbitrary selection of old versions of the chain, including all its content. This is especially important when considering the issue of access control. When it comes to providing cryptographic access control, unlike in normal distributed systems it is not possible to update protection by re-encrypting all data on the chain, as the attacker can always be in the possession of old versions.

Another trust-related problem, for which a solution would be a key to the widespread adoption of blockchains, is the issue of redaction of wrong information entered into the chain. While there already exist methods for marking this information as invalid, it cannot be removed altogether from the

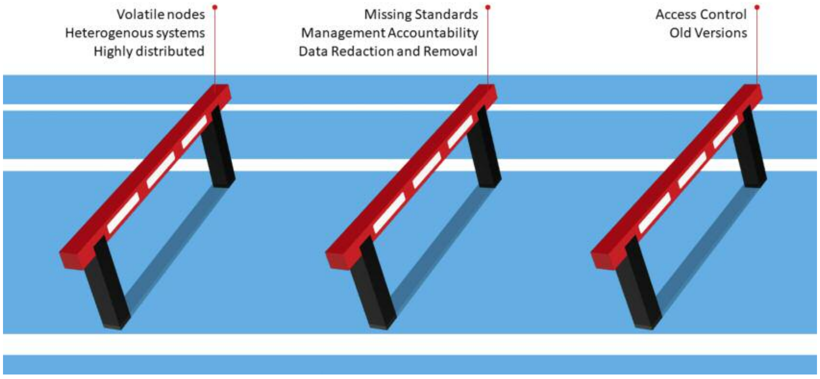


*Figure 1: Key security management concerns in the blockchain ecosystem.*

chain. This is especially important when illegal content like child pornography is introduced into a chain, where full and un-recoverable deletion is required. While there are solutions to this problem, they typically do not remove the illegal data from the chain but merely render it inaccessible. Managing the system as a responsible entity thus can even result in legal problems. Furthermore, the nature of these distributed systems is highly volatile, with nodes entering and leaving, new nodes being introduced and old ones never operating again. This makes asset management very complex, especially since it is unclear to the system provider, which nodes work together, or might belong to the same entities.

Last, but not least, another key issue from a security management perspective is  heterogeneous systems: Most companies already run traditional IT systems and won't change all of it to blockchains at once, which means that blockchain-based systems need to interface with traditional programs that use totally different approaches for deciding when a transaction is valid, how data is stored and access managed. Thus, these interfaces would pose an important attack surface, especially since there are currently no standardised solutions to securely integrate arbitrary blockchains into complex traditional systems [3].

Security management is a major issue in IT security, as many practical problems in securing complex environments cannot be reduced to pure technical issues. Furthermore, technical solutions are often rendered useless because it is impossible to introduce them to the system. Security management is thus one of the most important aspects in security, and given the complexity and special requirements of blockchains, novel approaches are required.

In the JRC for Blockchain Technologies and Security Management [L1], we tackle these issues and provide best practices for adopting blockchain-based systems in addition to traditional IT systems. Our focus is on the issues of trust and access control, but we also provide best practices for dealing with illegal content from the design phase onwards. Furthermore, we provide insights into how to adopt standards from the ISO27k family as applied to the particularities of blockchain-based systems.

**Link:**
[L1] https://research.fhstp.ac.at/projekte/josef-ressel-zentrum-fuer-blockchain-technologien-sicherheitsmanagement

**References:**

[1] ISO/IEC 27001:2017.  Information Technology - Security Techniques – Information Security Management Systems – Requirements.

[2] L. König, et al.: “The Risks of the Blockchain A Review on Current Vulnerabilities and Attacks. J. Internet Serv. Inf. Secur., 10, 110–127, 2020.

[3] L. König, et al.: “Comparing Blockchain Standards and Recommendations. Future Internet, 12(12), p.222, 2020.

**Please contact:**

Peter Kieseberg
St. Pölten University of Applied Sciences, Austria
peter.kieseberg@fhstp.ac.at

# Trick the System: Towards Understanding Automatic Speech Recognition Systems

by Karla Markert (Fraunhofer AISEC)

***Automatic speech recognition systems are designed to transcribe audio data to text. The increasing use of such technologies makes them an attractive target for cyberattacks, for example via “adversarial examples”. This article provides a short introduction to adversarial examples in speech recognition and some background on current challenges in research.***

As their name suggests, neural networks are inspired by the human brain: just like children learn the abstract notion of “an apple” from a collection of objects referred to as “apples”, neural networks automatically “learn” underlying structures and patterns from data. In classification tasks, this learning process consists of feeding a network with some labelled training data, aiming to find a function to describe the relation between data and label. For example, a speech recognition system presented with audio files and transcriptions can learn how to transcribe new, previously unprocessed speech data. In recent decades, research has paved the way to teaching neural networks how to recognise faces, classify street signs, and produce texts. However, just as humans can be fooled by optical or acoustic illusions [L0], neural networks can also be tricked to misclassify input data, even when they would be easy to correctly classify for a human [1].

Here we discuss two popular network architectures for speech recognition, explain how they mimic human speech perception, and how they can be tricked by data manipulations inaudible to a human listener. We discuss why designing neural networks that are robust to such aberrations is hard, and which research questions might help us make improvements.

Speech recognition can be based on different statistical models. Nowadays neural networks are a very common approach. Automatic speech recognition (ASR) models can be realised end-to-end by one single neural network or in a hybrid fashion. In the latter case, deep neural networks are combined with hidden Markov models.

When training an end-to-end model like Deepspeech [L1] or Lingvo [L2], the system is only provided with the audio data and its final transcription. Internally, the audio files are often pre-processed to “Mel-frequency cepstral coefficients”. In this case, the audio signal is cut into pieces, called frames, which are in return decomposed into frequency bins approximately summing up to the original input. This pre-processing step reflects how the inner ear transmits sounds to the brain. A recurrent neural network then transcribes every frame to a character without any additional information. Using a specific loss function, sequences of frame-wise transcriptions

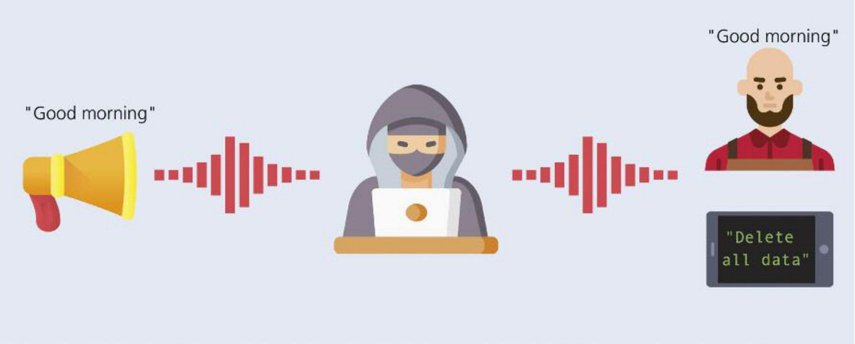


*Figure 1: Adversarial examples can fool speech recognition systems while being imperceptible to the human. This enables far-reaching attacks on all connected devices.*

are turned into words. This process mimics our brain understanding words from a series of sounds.

On the other hand, hybrid models like Kaldi [L3] consists of different submodels that are trained separately and require additional expert knowledge. The audio data is preprocessed in a similar way to above. The acoustic model consists of a neural network that turns these audio features into phonemes (provided by an expert). Subsequently, the phonemes are turned into words by a language model that also accounts for language-specific information such as word distributions and probabilities.

Despite all their differences in the mathematical setup, both approaches are susceptible to adversarial examples. An adversarial example is an audio file manipulated in such a way that it can fool the recognition system, with the manipulation being inaudible to a human listener, as depicted in Figure 1. It can thus happen that, for a given audio file, a human hears “good morning” while the ASR model understands “delete all data”. Clearly, this is particularly threatening in sensitive environments like connected industries or smart homes. In these settings, the voice assistant can be misused to control all connected devices without the owner’s awareness. Current techniques enable hackers to craft nearly imperceptible adversarial examples, some of which even allow for playing over the air.

Adversarial attacks are constructed per model and can be designed in a white-box setting, where the model is known to the attacker, as well as in a black-box setting, where the attacker can only observe the model’s output (the transcription, in the case of ASR systems). There are different explanations for why these attacks are possible and why one can manipulate input data to be classified as a chosen target: the neural network over- or underfits the data, or it just learns data attributes that are imperceptible to humans. With respect to images, it has been shown that some adversarial examples are even transferable between different models trained to perform the same task (e.g., object classification). In contrast, adversarial examples in the speech domain exhibit far less transferability [2].

Over the years, different mitigation approaches have been proposed. So far, the state-of-the-art method of defending against adversarial attacks is to include adversarial examples in the training data set [L4]. However, this requires a lot of computation: computing adversarial examples, retraining, recomputing, retraining. Current research addresses questions regarding the interpretability of speech recognition systems, the transferability of audio adversarial examples between different models, the design of detection methods for adversarial examples or of new tools to measure and improve robustness, especially in the language domain. Here, the feature set (audio), the human perception (psychoacoustic), and the learning models (recurrent neural networks) differ from the image domain, on which most previous research has focussed so far.

In a way, acoustic illusions and audio adversarial examples are similar: the perception of the human or the neural network, respectively, is fooled. Interestingly, very rarely can the human and the machine both be fooled at the same time [3]. Rather, even when the ASR system is very accurate in mimicking human understanding, it is still susceptible to manipulations elusive to the human ear. Fortunately, however, such attacks still need careful crafting and can only work under suitable conditions. Thus, currently, “good morning” still means “good morning” in most cases.

**Links:**

[L1] https://www.newscientist.com/article/dn13355-sound-effects-five-great-auditory-illusions/
[L2] https://github.com/mozilla/DeepSpeech
[L3] https://github.com/tensorflow/lingvo
[L4] https://github.com/kaldi-asr/kaldi
[L5] Lessons Learned from Evaluating the Robustness of Neural Networks to Adversarial Examples. USENIX Security (invited talk), 2019.
https://www.youtube.com/watch?v=ZncTqqkFipE

**References:**

[1] I.J. Goodfellow, J. Shlens, C. Szegedi: “Explaining and harnessing adversarial examples”, arXiv preprint arXiv:1412.6572, 2014.
[2] H. Abdullah et al.: “SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems”, arXiv e-prints, 2020, arXiv: 2007.06622.
[3] G. F. El Sayed et al.: “Adversarial examples that fool both computer vision and time-limited humans”,  arXiv preprint arXiv:1802.08195, 2018.

**Please contact:**

Karla Markert
Fraunhofer Institute for Applied and Integrated Security AISEC, Germany
karla.markert@aisec.fraunhofer.de

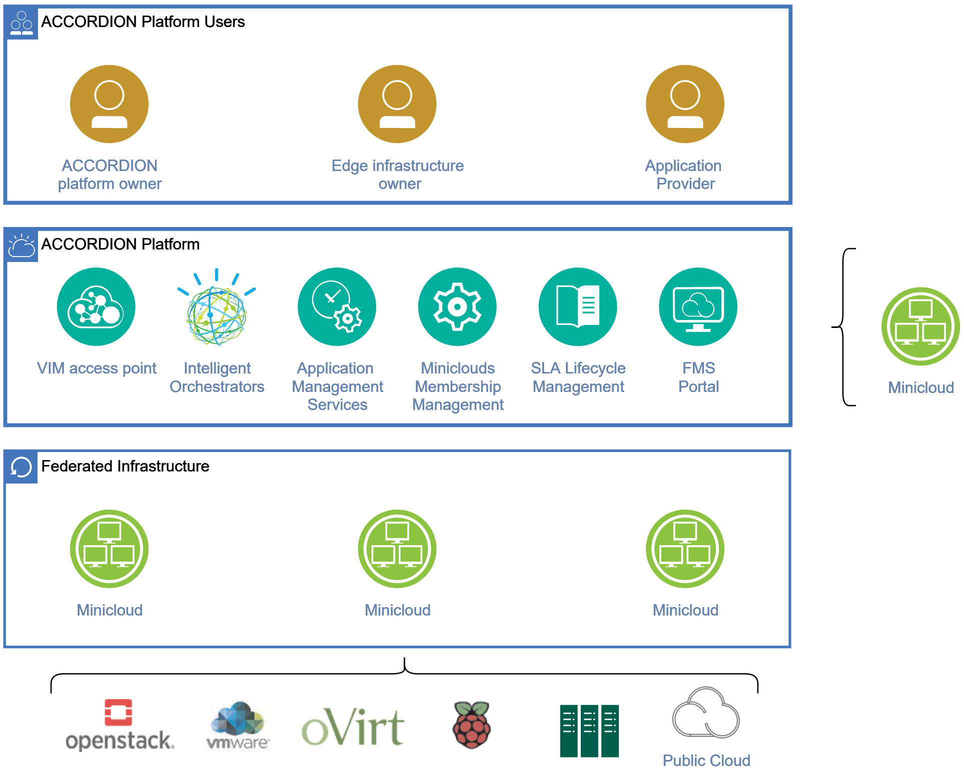
# ACCORDION: Edge Computing for NextGen Applications

by Patrizio Dazzi (ISTI-CNR)

***Cloud computing has played a vital role in the digital revolution. Clouds enable consumers and businesses to use applications without dealing with local installations and the associated complexity. However, a big class of applications is currently being blocked because of their dependency on on-site infrastructures or specialised end-devices but also because they are too latency-sensitive or data-dependent to be moved to the public cloud.***

These Next Generation (NextGen) applications would benefit from an advanced infrastructure with ubiquitous presence, unblocking them from fixed geographies. Current edge computing implementations support only specific geography and architectures, which means that the scope of local resources and infrastructures are also restricted to the needs of certain edge-enabled applications. Existing solutions lead to user lock-in and, overall, have a negative impact on the open diffusion of edge computing. They actually hinder the exploitation of the ubiquitous presence of edge infrastructure, limiting the edge-enablement of the NextGen applications. The shift towards edge computing, supported and promoted by the ACCORDION project aims to (i) limit vendor lock-in situations that trap SMEs into committing  to the services offered by big vendors, but also (ii) leverage the vast pool of local, potentially specialised, resources of the SME landscape, and (iii) mitigate the slow adoption rate of new technologies by many SMEs. In this context, ACCORDION aspires to:

- provide an open, low-latency, privacy-preserving, secure and robust virtualised infrastructure; and
- provide an application management framework, tailored to the needs of the SME's skillset, that will enable the deployment of NextGen applications on top of this infrastructure. This method allows us to reap the benefits of edge computing while exploiting the set of heterogeneous resources that can embed such computing facilities.



*ACCORDION Platform architecture.*

ACCORDION's key concept is aligned with the recent developments in the Multi-Access Edge Computing (MEC) front, extending the interest to both network and computation resources. ACCORDION aims to provide a platform to address the needs of NextGen applications by properly exploiting edge resources. The platform is intended to provide an integrated approach dedicated to both developers and local infrastructure owners. It will encompass frameworks for application development, solutions for an adaptive and robust cloud/edge infrastructure continuum, and the abstraction of widely heterogeneous pools.

ACCORDION, coordinated by ISTI-CNR, started in early 2020, so we are just starting the second year of research. In line with our original plans, the first year of the project focused largely on analysing the state of the art, the selection of base technologies and the definition of the overall architecture of the system: moving from the high-level conceptual one to a more concrete one, in which the key modules of the system along with the interactions among them have been defined.

We will soon deliver the core modules of the ACCORDION platform, to be followed by a complete release of the integrated ACCORDION platform. Once the integrated platform is ready, a first release of ACCORDION use cases will be properly tailored to be run on top of it.

**Links:**
[L1] https://cordis.europa.eu/project/id/871793
[L2] https://www.accordion-project.eu

**Please contact:**
Patrizio Dazzi, Project Coordinator, ISTI-CNR, Italy
patrizio.dazzi@isti.cnr.it

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Call for Proposals

## Dagstuhl Seminars and Perspectives Workshops

***Schloss Dagstuhl – Leibniz-Zentrum für Informatik is accepting proposals for scientific seminars/workshops in all areas of computer science, in particular also in connection with other fields.***

If accepted the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/ administrative work, and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Thanks to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular, details about event form and setup as well as the proposal form and the proposing process can be found on

**https://www.dagstuhl.de/dsproposal**

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

### Important Dates

- This submission round:
  April 1 to April 15, 2021
  Seminar dates: In 2022/2023
- Next submission round:
  October 15 to November 1, 2021
- Seminar dates: In 2023.

## W3C Workshop on Wide Color Gamut and High Dynamic Range for the Web

Since 1996, color on the Web has been locked into a narrow-gamut, low dynamic-range colorspace called sRGB.

Display technologies have vastly improved since the bulky, cathode-ray tube displays of the 1990s. Content for the Web needs to be adaptable for different gamuts, different peak luminances, and a very wide range of viewing conditions. To understand what next steps are envisioned to enable Wide Color Gamut and High Dynamic Range on the Open Web platform, W3C organizes a virtual workshop in April-May 2021 with pre-recorded talks and interactive sessions from browser vendors, content creators, color scientists and other experts.

**Link:** https://kwz.me/h5q

## W3C Workshop on Smart Cities

Heavily used by various industries and services (including payments/commerce, media distribution, video conferencing, connected cars, etc.), the Web is becoming a promising platform for IoT interoperability. In the context of the current global sanitary context, there are great expectations for smarter and easier integration of various technologies from multiple vendors related to IoT devices and Web services. W3C organizes a W3C workshop on 25 June 2021 to discuss applications and use cases of W3C's Web of Things standards for smart city services.

**Link:**
https://w3c.github.io/smartcities-workshop/

ERCIM
European Research Consortium
for Informatics and Mathematics

## Horizon Europe Project Management

A European project can be a richly rewarding tool for pushing your research or innovation activities to the state-of-the-art and beyond. Through ERCIM, our member institutes have participated in more than 90 projects funded by the European Commission in the ICT domain, by carrying out joint research activities while the ERCIM Office successfully manages the complexity of the project administration, finances and outreach.

### Horizon Europe:How can you get involved?

The ERCIM Office has recognized expertise in a full range of services,including:

- Identification of funding opportunities
- Recruitment of project partners (within ERCIM and through ournetworks)
- Proposal writing and project negotiation
- Contractual and consortium management
- Communications and systems support
- Organization of attractive events, from team meetings to large-scale workshops and conferences
- Support for the dissemination of results.

### How does it work in practice?

Contact the ERCIM Office to present your project idea and a panelof experts within the ERCIM Science Task Group will review youridea and provide recommendations. Based on this feedback, theERCIM Office will decide whether to commit to help producing yourproposal. Note that having at least one ERCIM member involvedis mandatory for the ERCIM Office to engage in a project.If the ERCIM Office expresses its interest to participate, it willassist the project consortium as described above, either as projectcoordinator or project partner.

**Please contact:**
Peter Kunz, ERCIM Office
peter.kunz@ercim.eu

ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.

ERCIM is the European Host of the World Wide Web Consortium.

Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
www.iit.cnr.it

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
www.cwi.nl

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
www.fnr.lu

Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
www.ics.forth.gr

Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
www.iuk.fraunhofer.de

INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal
www.inesc.pt

Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
www.inria.fr

I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
www.isi.gr

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

SBA Research gGmbH
Floragasse 7, 1040 Wien, Austria
www.sba-research.org/

SIMULA
PO Box 134
1325 Lysaker, Norway
www.simula.no

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
www.sztaki.hu/

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
www.cs.ucy.ac.cy/

Universty of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
www.mimuw.edu.pl/

VTT Technical Research Centre of Finland Ltd
PO Box 1000
FIN-02044 VTT, Finland
www.vttresearch.com