NEUROTECH

Deliverable D2.6: NEUROTECH Roadmap and Benchmarks updated.

- Project: NEUROTECH, Grant number 824103
- Report dissemination level: public
- Report target delivery date (updated version): project month 30

Contributors:
Alice Mizrahi (Thales), Teodora Petrisor (Thales), Yulia Sandamirskaya (Intel), Steve Furber (University of Manchester), Michael Schmuker (University of Hertfordshire), Sabina Spiga (Consiglio Nazionale Delle Richerche), Bert Offrein (IBM), Melika Payvand (Universitat Zurich), Elisa Donati (Universitat Zurich).

Note about the updates compared to the first draft (Deliverable D2.2):
Taking advantage of discussion with the actors of the field, in particular during the different online events organised by the consortium, we have gathered more information and have been able to update the Roadmap in consequence. The whole document has been reviewed and improved by the consortium. Some sections have been significantly rewritten. In particular:
- A market analysis has been added
- The “Challenges” section has been completed with a description of paths for moving forward for each challenge, as well as with examples of concrete actions to undertake.
- The “Technology” section has been completed.
- A section about the different actors and their roles has been added.
- The axes of the next steps have been detailed and framed into timelines.

Summary

This deliverable is the final NEUROTECH Roadmap for Neuromorphic Computing Technologies. Neuromorphic computing is a booming interdisciplinary field which aims at building computing systems that take inspiration from the brain at the hardware level. In this document we start by defining the goals of the field. Then we explain the main applications of the field, which makes it so promising, and provide a market analysis. We present the main technologies that support neuromorphic computing. We explain the challenges that the field is currently facing and provide paths to overcome them. We provide guidelines toward industry adoption of neuromorphic computing. Finally we present timelines of the main directions in which we expect the field to progress.

Table of contents

1 Introduction:

1.1 The Goal of the Roadmap

By taking loose inspiration from the brain, artificial neural network algorithms have made tremendous progress in artificial intelligence. However, to unlock significant gains in terms of novel real-world capabilities, performance and efficiency, a more ambitious step needs to be taken: to develop a new technology that emulates neural computation directly at the hardware level.

Pursuing this goal, neuromorphic computing is a booming field bringing together several fields of research (electrical engineering, computer science, physics, material science, computational neuroscience). Because of its many applications, it has attracted the attention of industry as well as innovative SME.

The goal of this roadmap is to provide answers to the following questions:

- What is neuromorphic computing and what are its applications ?
- What are the challenges to the field?
- What steps are needed to enable industry adoption?

We expect this Roadmap to be useful to a wide range of audience, including the following:

- Researchers from various fields eager to link their research topic to neuromorphic computing will get an idea of the critical questions and challenges to be solved.
- Research departments of companies can be convinced to invest in neuromorphic computing and given an idea of where to start.
- Policy makers can be informed on the importance of the field and of the issues that need to be pushed for impact to be realized.

1.2 Approach to Developing a Roadmap

This roadmap was written and updated continuously during the project.
We used the live NEUROTECH Forum to collect opinions from experts and the community on several questions related to the NCT Roadmap. This was done during a guided panel discussion (recording is available) and with an online questionnaire, which the Forum participants were invited to fill-in. The members of the Industry Work Group were subsequently contacted to give their opinion.
The NEUROTECH online seminars series (both Educational and Industry events) were key opportunities to collect insights, in particular during the open panel sessions where current issues of the field were debated.

2 Neuromorphic Computing as a Goal

Our first action for developing a Roadmap was to improve the definition of neuromorphic computing. Rather than having an inside/outside boundary, we see neuromorphic computing as a goal towards which different directions converge. These directions are schematized in Figure 1. They correspond to features of the brain as a computer, which we seek inspiration from. These directions structure the roadmap of neuromorphic computing as they are the guiding principles of the field.

Each of these directions represent a breakthrough from the current computing paradigm. As such, Neuromorphic computing represents an extremely ambitious multi-disciplinary effort. Each direction will require significant advances in computing theory, architecture, device physics, software and algorithms.

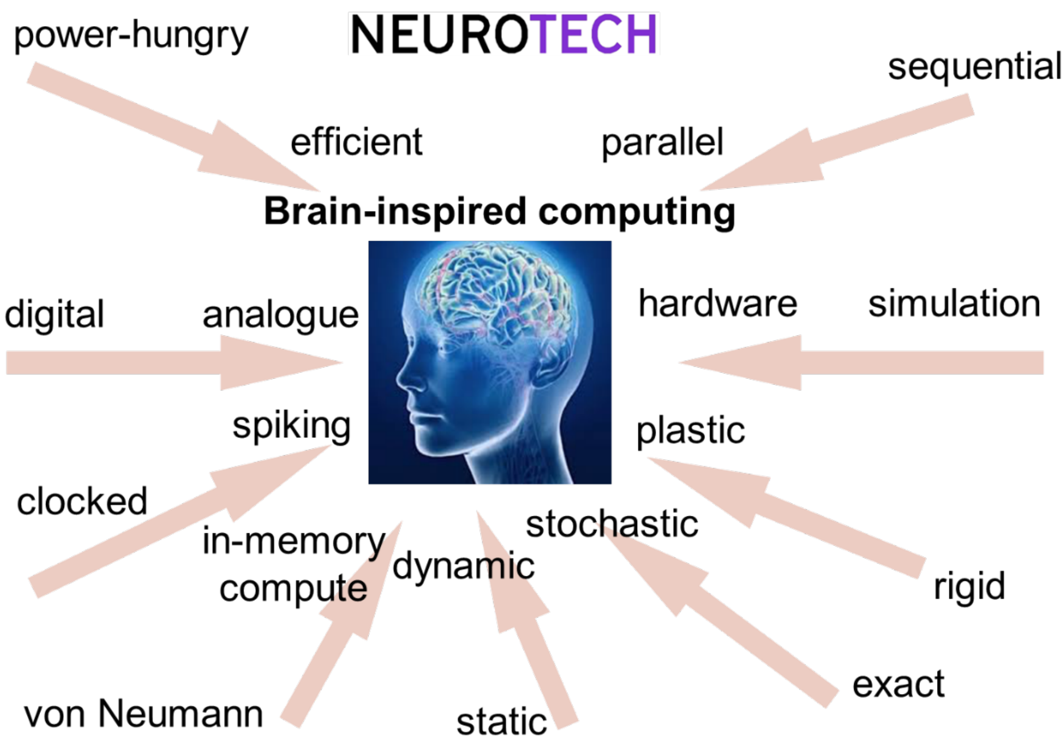


Figure 1 Neuromorphic computing as a goal

2.1 Hardware vs. Simulation

Taking inspiration from the brain for computing is already present in machine learning and artificial intelligence through artificial neural network algorithms. This abstract inspiration has already given rise to tremendous progress in image, video, audio and natural language processing, and to successful commercial applications. However, in order to unlock significant gains in terms of performance and efficiency, a more ambitious step needs to be taken: to build a new kind of computer, inspired from the brain at the hardware level. This is the goal of neuromorphic computing. We seek not just simulating artificial neural networks, but to actually build them.

2.2 Efficient vs. Power-hungry

Application wise, one key motivation for neuromorphic computing is to achieve significantly higher power efficiency than existing solutions. Artificial neural networks, when run on conventional hardware, consume a lot of energy. State of the art GPUs consume several hundreds of Watts, which limits the deployment of neural networks on embedded systems. Even supercomputers consuming a Mega Watt cannot emulate the whole human brain, which limits our ability to improve our understanding of the brain through such simulations. In comparison, the human brain consumes only 20 Watts. The energy efficiency of the brain is several hundreds of tera operations per second and per Watt, while existing solutions are limited to a few tera operations per second and per Watt. By building computers inspired from the brain at the hardware level, neuromorphic computing aspires to bridge this energy efficiency gap. For example, sensory computing in the brain achieves a large part of its efficiency by operating in an event-based manner, where signals are only sampled and transmitted when new information either arrives or is computed. Spiking architectures natively support this scheme and thus support efficiency gains through event-based processing. Nevertheless, it is our goal to present a broad perspective taking into account both artificial and spiking systems.

2.3 Parallel vs. Sequential

One of the most impressive features of the human brain is its massive parallelism. Although each neuron computes at the millisecond scale (much slower than CMOS transistors which function below the nanosecond), the brain can perform 100 tera operations per second, orders of magnitude more than artificial neural networks on conventional computers. Parallel computing is a much studied topic beyond the scope of neuromorphic computing. However, parallel computing in conventional computer architectures is quite limited. Approaching the parallelism of the brain will require drastic changes in computer architectures. Moreover, it will require low power components so that they can all function simultaneously. Indeed, in current processors, the whole chip cannot function simultaneously because of power budget constraints.

2.4 In-memory Computing vs. von Neumann Architecture

Conventional computers rely on the von Neumann architecture, where memory and computing are physically separated. In consequence, a large part of the energy consumption and delays are due to the transfer of information between memory and computing parts, a phenomenon often referred to as the “von Neumann bottleneck”. In neural network algorithms, this issue is critical because huge numbers of parameters need to be stored and frequently addressed. The brain is extremely different in this regard: memory and computing are completely intertwined. The neurons, which compute, are connected by synapses, which carry the memory. Neuromorphic computing aims at bringing memory and computing together to achieve “in-memory computing”.

In-memory computing is being made possible through the development of emerging nanoscale memory devices. Various classes of such memories exist and will be discussed in this roadmap. Their common assets are that they are non-volatile, fast and low energy, can be read and written electrically or in some cases optically and can be monolithically integrated into CMOS chips.

2.5 Plastic vs. Rigid

Learning in the brain is made possible by its plasticity. The connections between neurons – the synapses – are not rigid but plastic, which means they can be modified. Learning, both in the brain and in artificial neural networks algorithms, corresponds to repetitive modification of the synapses until reaching a set of connections enabling the neural network to perform tasks accurately. In conventional computers, this is done by explicit modification of the memory banks storing the weights. Neuromorphic computing aims at building systems where weights are self-modified through local rules. Here again, the role of non-volatile memories intertwined with computing circuits is critical. Their dynamics makes it possible to implement bio-inspired learning rules. For instance, memristors can implement Spike Timing Dependent Plasticity, a bio-inspired rule for unsupervised learning.

2.6 Analogue vs. Digital

Conventional computers rely on digital encoding: voltages in the processor at the steady state only take two values, which represent 0 and 1. Transient intermediary values do not represent anything. All numbers are coded in binary, as a string of 0 and 1. In the brain, this is not the case. The electrical potential at the membranes of neurons can take continuous values, and so can the synaptic weights. Reproducing such behavior with digital encoding takes large circuits. Thus, using directly an analogue encoding could improve efficiency. Neuromorphic computing aims at using components with intrinsic analogue behavior mimicking the key functions of neurons and synapses. For neurons, this can be achieved by CMOS transistors used in an analogue regime and by emerging technologies such as spintronic nanodevices or photonics. For synapses, which also require non-volatility, emerging memories are a key enabler.

2.7 Dynamic vs. Static

Conventional computers use the steady state of their circuits to encode information. On the contrary, the brain is a complex dynamic system. Biological neurons are non-linear oscillators that emit spikes of voltage. They are coupled to each other and capable of collective behavior such as synchronization. There are also some indications that the brain functions at the critical point between order and chaos. Neuromorphic computing aims at emulating such dynamic behavior in order to go beyond the possibilities of static neural networks, in particular regarding learning. Here again, it is key to have circuits and components with intrinsic analogue dynamics emulating neural functions. Coupled oscillators can be achieved with CMOS ring oscillators, spintronic devices, metal-oxide sandwiches, photonics devices etc.

2.8 Spiking vs. Clocked

Conventional computers are run by a clock which sets the pace of all circuits. There is no such clock in the brain. Neurons emit and receive spikes in an asynchronous way. Neuromorphic aims at building computers built on these principles. By having activity only when necessary, energy consumption will be reduced.

2.9 Stochastic vs. Exact

Conventional computers aim at very high precision, coding numbers in 64 bits floating point precision. In the brain, this is far from the case as the biological environment is noisy and neurons and synapses exhibit variability and stochasticity. Resilience to such imprecision seems to be a key asset of neural networks. There are even suggestions that the brain uses noise for computing. Relaxing the constraints on the exactitude of components and computing steps will decrease energy consumption. Obtaining accurate results with approximate computing components and steps is a goal of neuromorphic computing. This will be crucial to be able to use components in their analogue regime, where noise and variability are more significant.

3 Applications “Pool”

Neuromorphic computing is both of scientific and practical interest. This illustrated by the fact that both academics and industrialists (from large groups to start-ups) are active in the field.

By definition, neuromorphic computing should provide solutions for problems where the brain is particularly efficient. Neuromorphic computing does not aim at replacing general computing. Rather, neuromorphic computing will be used in specialized chips that work together with general-purpose chips.

Here, we provide an overview of the most promising applications of neuromorphic computing. To select these applications, we have solicited input from the Work Groups – in particular the Industry group – as well as the Forum participants, both by email and during the panel discussions at the NEUROTECH Forum and at the online NEUROTECH events. These answers complete the results of the internal discussions of the consortium.

3.1 Artificial Intelligence on the Edge

Neuromorphic computing will provide systems capable of running state of the art artificial intelligence tasks – deep neural networks – while consuming little power and energy. The superior energy efficiency of event-based systems will enable applications of signal processing, inference and control at much lower power budgets than current technology. This opens the way to the deployment of artificial intelligence on the edge and in embedded systems, where consumption and size are critical.

Key applications are:

- Detection (always on sensor processing, very low latency and low power, ~10 microW)
- Recognition (could be triggered by ultralow-power detection, power: ~0.1mW)
- Situation awareness (semantic map of the environment, needs to be stored and updated online)

3.2 Sensor Processing

"Smart" sensors currently still rely heavily on computing centers where they send raw data to be processed and sent back. The ability for sensors to process information on site without data transfer would provide faster response as well as better security and privacy at a much lower total power consumption.

Neuromorphic computing could in particular be useful for the observation of sensory signals and decision to trigger further processing or an action made on the edge (bio-signal monitoring, fall detection, voice detection, etc.).

Neuromorphic computing will be a key enabler of an efficient and secure internet of things.

3.3 Health

Health is a field that is currently being transformed by neural networks, for instance for classifying tumor images into benign or malignant. Neuromorphic computing could bring further benefits, in particular for processing dynamical signals and time series. One example of application is ECG online evaluation.

The potential of neuromorphic computing for low power, small size chips performing artificial intelligence tasks can revolutionize biomedical sensors: implants could be capable of performing real time complex monitoring.

In health, the importance of data privacy is huge, making on-site processing of information even more critical.

3.4 Robotics

A natural application for neuromorphic computing is robotics. In particular, it could give rise to agile, compliant robots with HRI capabilities such as:

- Learning dynamical models
- Coordination of behavior
- Force control

Merging health with robotics is full of applications for neuromorphic computing. Smart pills capable of action in the body and prosthetics are two key examples.

Event-based systems for sensing, inference and control will enable lower latency and higher energy efficiency in robotic systems.

3.5 Optimization

Artificial neural networks use learning to solve large optimization problems. This has many applications outside what is usually thought of as cognitive tasks. This includes:

- Complex systems with many parameters
- High performance computing
- Thermodynamic simulation (which involves massive matrix-multiplication tasks which could be accelerated similar to NCT)

3.6 Natural Language Processing

Neuromorphic computing has the potential to process natural language and perform tasks such as translation and interpretation. It will be able to process speech in real time, from the raw dynamical data, to the reasoning on the extracted meaning.

3.7 Personal Assistants

Combining different applications of neuromorphic computing such as optimization and natural language processing will lead to more efficient personal assistants. These will be capable of time management and scheduling, but also of assistive robotics and care, in particular for elders.

3.8 Autonomous Vehicles

Combining robotics, sensory processing, optimization, and potentially natural language processing, autonomous vehicles have a strong need for neuromorphic computing. Many large industrial groups are working on the topic. One critical limitation of the autonomous car is the power consumption and size of the computing systems it relies on (several kW and a large space in the trunk).

3.9 Smart Manufacturing

Industrial machines and processes can benefit greatly from neuromorphic computing. Optimization of a whole process or fabrication chain is one example. Robotics applications of neuromorphic computing will make fabrication more efficient. Neuromorphic computing can also provide solutions for anomaly detection in time series, automatization of controls and tests, design for manufacturing, defect detection and forecast, predictive maintenance of machines etc. These will make industry more sustainable.

3.10 Computational Neuroscience

Neuromorphic chips will be privileged systems to simulate biological neural networks. Thus, they could contribute to understanding the brain. This would bring massive novel knowledge but also provide new treatment for neurological diseases. It might also bring some light on how to achieve general intelligence.

4 Neuromorphic Computing Market Analysis

A careful market analysis, and prediction for Neuromorphic Computing’s share of it, is at this stage purely speculative as little consensus can be found between business analysts. This may be due to the lack of clear definitions of terms with very fuzzy frontiers between Artificial Intelligence, Deep Learning, Neuromorphic Architectures, Neural Chips. Sometimes all these are used interchangeably or even mixing the technological achievements or goals and applicative or use-case deployments. In addition, market analysts consider different criteria such as generated revenues, benefits, investments etc. Expected market growth may thus vary significantly based on these considerations. We refer the reader to the *aimultiple*[1] for an in-depth example of the different axes of analysis for the AI market predictions.

In the following, we attempt to give a first review of the neuromorphic market estimations. This is based on a study by Yole Development[2] that seem to have a similar understanding of the concept as we do, as well as a clear separation between neuromorphic computing and neuromorphic sensing, allowing for better identifying the main stakeholders, as shown in the timeline in Figure 2. Notice that the past five years witnessed the previously mentioned democratization tentative of neuromorphic technologies, as indicated by the increasing number of new actors. This comes with the diversification of the focus for each of these companies, instead of competing over the "unique" general-purpose architecture. In other words, the upcoming decade should be more marked by striving towards more and more specialized solutions for different purposes, and perhaps an idea of cohabitation with current technologies instead of their full replacement.

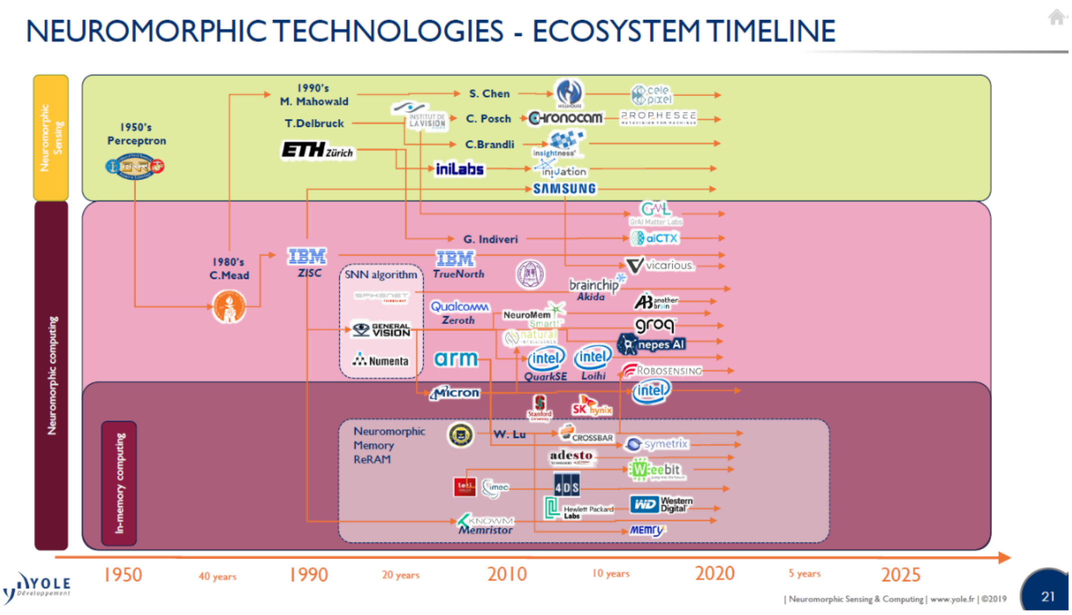
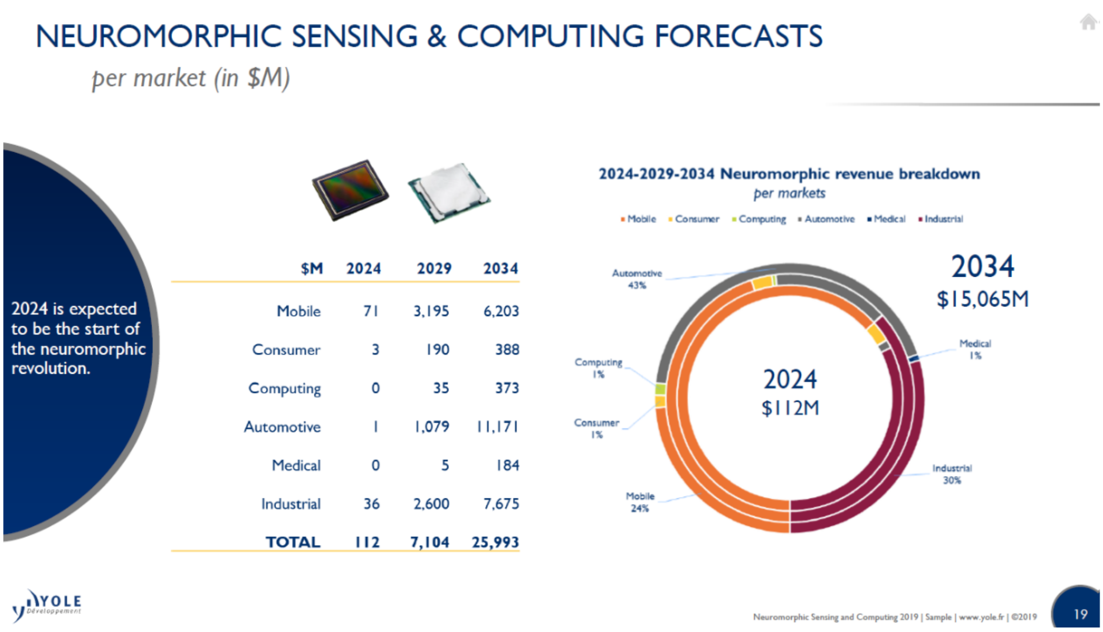


Figure 2: Evolution of the main actors in neuromorphic technologies (computing and sensing) during time.

Assuming that many of the open technical questions described in this roadmap are to be addressed successfully in the following decade or so, "2024 is expected to be the start of the neuromorphic revolution", in particular in terms of market pervasiveness, according to Yole Developpement 2019[2]. Their predictions, corroborated by other sources, and which, according to them, could also serve as a roadmap, are shown in Figure 3, for the cumulated neuromorphic computing (processors) and sensing activities.

[1] 131 Myth-Busting Statistics on Artificial Intelligence (AI) in 2021 (aimultiple.com)
[2] Yole Developpement, Neuromorphic Sensing Computing Industry Overview, 2019



$M	2024	2029	2034
Mobile	71	3,195	6,203
Consumer	3	190	388
Computing	0	35	373
Automotive	1	1,079	11,171
Medical	0	5	184
Industrial	36	2,600	7,675
TOTAL	**112**	**7,104**	**25,993**

Figure 3: market prediction by Yole Developpement, 2019, in the ideal situation where technological hurdles would all have been removed.

Consistent with this table, business analysis experts predict that the automotive domain is the fastest growing market[3] and by 2034 could take the lead in integrating neuromorphic technologies. Major neuromorphic chip manufacturers (*Intel Corporation, Qualcomm Incorporated, BrainChip Holdings Ltd., HRL Laboratories LLC, IBM Corporation*, …) aim to enhance their market reach, in particular through advanced driver assisted systems (ADAS) and autonomous vehicles[4]. Interestingly, by 2034 the projected usage for neuromorphic technologies, aside from the automotive domain, is no longer mostly on the mobile market. It should be overcome by the industrial activities. In light of these considerations, the Neurotech efforts to build a community concerned not only with the technological considerations of each of the involved stakeholders but also on working towards identifying the right applications and benchmarking activities take an entire new meaning.

Overall, as shown in Figure 3, in three years' time the total part of the market for neuromorphic technologies should be expected to reach over one hundred million USD[5] and, a decade after, the predictions go to tens of billions of USD[6]. This is surely linked with the ubiquity of neural-based solutions at every stage of the producer-to-consumer chain, from semiconductors to fully deployed systems, thanks to their proven software superiority over other methods, and despite their resource limitations. Most probably, the biggest part of the AI market is accountable for deep learning technologies today, and in this landscape, the need for actual chips is becoming apparent.

[3] Mordor Intelligence - Neuromorphic Chip Market | Growth, Trends, Forecasts (2020 - 2025)
[4] Neuromorphic Chip Market Size, Status, Demand and Global Outlook 2020 (brainchipinc.com)
[5] Figures corroborated by MarketWatch: MarketWatch, Press Release, Feb 2021
[6] Neuromorphic Chips - Global Market Trajectory & Analytics (researchandmarkets.com) estimate the growth to 10.4 Billion USD by 2027, despite the COVID-19 crisis

Again, according to Yole Developpement, Figure 4 shows the business landscape as of 2019, highlighting the convergence of technological and application fields, in line with our vision for neuromorphic computing.

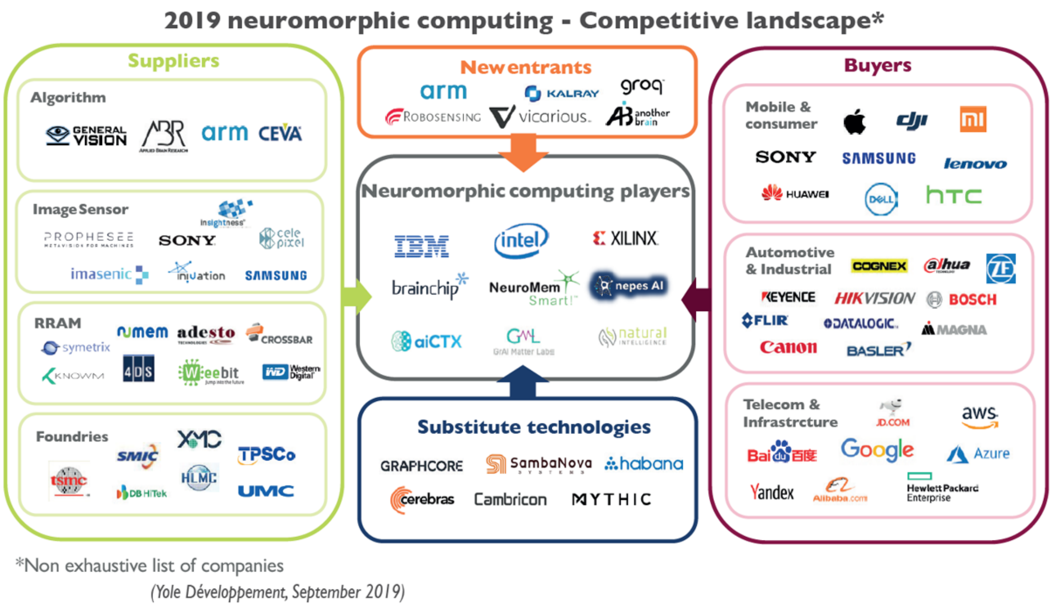


Figure 4: The convergence towards Neuromorphic Computing and its major actors, shows that the topic is pursued by large and small companies alike. This is true for all stages in the processing chain going from semiconductor technologies and sensors to the application deployment industries.

When joining progress in neural algorithms with the variety of markets addressed, it is clear that the application drive is undoubtedly pushing towards Artificial Intelligence adoption in most of the world's businesses today[7]. Yet this can only be achieved by handling the energy efficiency, latency and perhaps privacy issues, which induces a shift in semiconductor and hardware manufacturers' focus.

Currently, training neural networks occurs almost exclusively in data centers or offline architectures. With novel learning paradigms, more closely related to biology, such as unsupervised, local learning, a new step can be taken by brining training capability to more and more chips. In this respect, McKinsey&Company, acknowledges that the part of AI vs non-AI technologies (by AI meaning neural solutions) is growing[8], from 17 billion USD versus 223 billion USD for non-AI in 2020, up to an estimated 65 billion USD by 2025 versus 295 billion USD estimated for non-AI-related semiconductor available market. By total available semiconductor market McKinsey&Company includes: processors, memory and storage and excludes optics, discretes and micro-electrical-mechanical systems.

In same report they predict two main tendencies both at data centers and at the edge (Figure 5):

[7] Vision Spectra spring 2021 - Neuromorphic Processing Set to Propel Growth in AI
[8] AI hardware: Value creation for semiconductor companies | McKinsey

- the need for both inference and training is to grow rapidly in the next 3 to 5 years (left-hand side)
- there should be a shift in the preferred architectures (right-hand side), with ASICs becoming the predominant choice, especially at the edge.

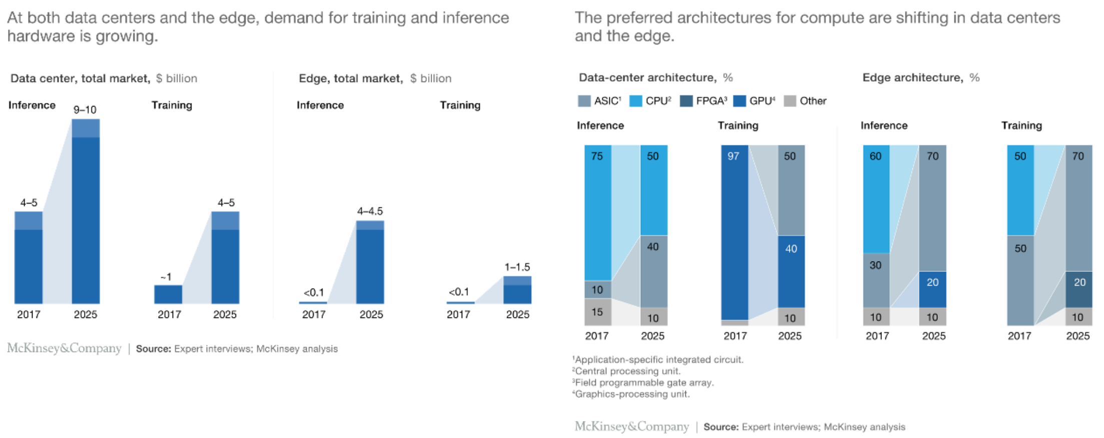


Figure 5: Market growth and circuit prevalence at data center and edge levels, according to McKinsey&Company

5 Technology: State of Art and Directions

The slow-down in the scaling of CMOS transistors (often referred to as the "end of Moore's law"), combined with the fact that the requirements of neuromorphic computing completely differ from conventional computing systems, have called for the use of new technologies for building neuromorphic chips. To a large extent, such platforms will build on the infrastructure, processes and silicon CMOS technology already available.

The involvement of novel technologies brings opportunities for neuromorphic computing, both in terms of functionalities (such as dynamical systems or memories) and efficiency (power consumption, size, speed etc.).
However, many of these technologies are not at the same maturity level as conventional digital CMOS transistors, which is a challenge for the development of neuromorphic chips, both for industries and academics.

Here we review the major technologies used for neuromorphic computing. In this first draft of the roadmap, we have identified the major technologies and key points to evaluate the assets and drawbacks of these technologies.

We list here the major technologies, from the most mature to the most exploratory.

5.1 Digital CMOS Technology

The mainstay of the semiconductor manufacturing industry, digital CMOS is well understood and delivers very consistent performance in volume manufacture. It can access the most advanced semiconductor technologies, which helps offset its intrinsic energy-efficiency disadvantages compared with analogue circuits. When applied to neuromorphic architectures, asynchronous, clocked and hybrid approaches to circuit timing can be used, and algorithms can be mapped into fixed (albeit highly parameterised and configurable) circuits for efficiency or into software for flexibility. Examples of the former include the IBM TrueNorth and Intel Loihi, and of the latter include NEUROTECH partner the University of Manchester’s SpiNNaker many-core neuromorphic computing platform[9].

For details on CMOS based neuromorphic systems, we direct the reader to the following reviews:

★ Furber, Steve. "Large-scale neuromorphic computing systems." *Journal of neural engineering* 13.5 (2016): 051001.

★ Schuman, Catherine D., et al. "A survey of neuromorphic computing and neural networks in hardware." *arXiv preprint arXiv:1705.06963* (2017).

Today, the main challenges for digital CMOS neuromorphic platforms are :

- tools that raise the level of abstraction that users work with to the levels seen in mainstream AI, such as Tensorflow, PyTorch, etc.
- the development of effective training algorithms.

In the next few years we expect the following advances:

- increasing commercial use of digital neuromorphic technologies in edge applications.
- digital neuromorphic accelerators integrated into general-purpose computing chips, in a similar way to the integration of neural network accelerators today.

In the long term we expect the following advances:

- 3D stacked integration will support integration densities approaching that of the brain

5.2 Analogue/mixed-signal Technology

Event-based analogue mixed-signal neuromorphic technology combines the compact and low power features of analogue circuits with the robustness and low-latency ones of digital event-based asynchronous ones. The key feature of the mixed-signal design approach, compared to pure digital ones, is the ability to build systems able to carry out processing with stringent resources in terms of power and memory by (i) only dissipating power when the data is present, and (ii) processing the data on-line, as it sensed or streamed through the system, using circuits that have time constants matched to the dynamics of the sensory signals processed, and without needing to store data or state variables in memory. This technology is an enabler for the

[9] Furber, Steve, and Petruț Bogdan. "SpiNNaker-A Spiking Neural Network Architecture." (2020): 350..

applications requiring sub-mW always-on real-time processing of sensory signals, for example in edge computing, personalized medicine and Internet of Things domains. Examples of neuromorphic processors that follow this approach are the DYNAP (Dynamic Neuromorphic Asynchronous Processor) series of devices[10] developed by the UZH NEUROTECH members.

For details on CMOS based neuromorphic systems, we direct the reader to the following reviews:

- ★ Furber, Steve. "Large-scale neuromorphic computing systems." *Journal of neural engineering* 13.5 (2016): 051001.
- ★ Schuman, Catherine D., et al. "A survey of neuromorphic computing and neural networks in hardware." *arXiv preprint arXiv:1705.06963* (2017).

Today, the main challenges for analogue and mixed-signal CMOS neuromorphic platforms are :

- Variability in the characteristics of the computational units such as neurons and synapses
- Difficult technology scaling
- Large area consumption by capacitors which hold the state dynamics
- Large leakage current limits the minimum static power consumption and increases the noise level

In the next few years we expect the following advances:

- Silicon On Insulator (SOI) technologies are less prone to variability because of the better control on the conductive channel
- SOI technologies also have reduced leakage as a result of the removal of the drain-bulk and source-bulk junction diodes
- Also, SOI technologies have a body biasing option which can control the amount of leakage

In the long term we expect the following advances:

- Large density capacitance in advanced technologies which would reduce the area required for holding state dynamics
- Better design tools that make technology scaling easier for analog design
- With the help of algorithms, variability in analog design can be tolerated or exploited

5.3 Technologies Beyond CMOS

As the CMOS technology approaches its scaling limits, more attention is being devoted to the development of emerging devices, which provide high functionality in a small footprint. In particular, the members of the NEUROTECH network are at the forefront of the development of memristive device technologies, which are a broad class of devices whose resistance can be modified upon by electrical stimuli.

[10] Moradi et al. 'A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs)'. IEEE transactions on biomedical circuits and systems, 12(1), 106-122. (2018)

The resistance changes can last for short or long time scales, leading to a volatile or non-volatile memory effect, respectively. Memristive devices are based on a large variety of physical mechanisms, such as redox reactions and ion migration, phase transitions, spin-polarized tunnelling, and ferroelectric polarization, and they have the potential to meet the considerable demand for new devices that enable energy-efficient and area-efficient information processing beyond the von Neumann paradigm. The leading memristive technologies which are currently at high maturity level are those firstly developed as non-volatile memory devices for storage applications and then integrated in large arrays and in combination with CMOS, namely resistive random access memory (RRAM), Phase change memory (PCM), Ferroelectric memory (FeRAM), and magnetoresistive random access memories (MRAM). Recently, RRAM, PCM, FeRAM and spin-transfer torque MRAM have been receiving increasing interest for neuromorphic computing, and many hardware demonstrations have been reported at device, but also circuit and systems level. The results are promising and, despite the system-level integration still not being at the level of the fully CMOS-based ones, the field is improving very fast, and driven by the parallel advancement of these technologies and their CMOS integration for storage or in-memory computing applications. Furthermore, promising developments are underway towards new and less matures concepts which span from new materials (e.g. 2D, nanowires), metal-insulator transition (e.g. VO2-based), organic materials, spintronics (spin torque oscillators, domain walls, spin-waves, skyrmions) and photonics.

5.3.1 Synapse Implementation

The key features of artificial synapses are the ability to update their states given new information (learning, plasticity) and to store analogue information (memory). This can be implemented either with intrinsically analogue or multilevel devices (whether in RRAM, PCM and FeRAM devices, or using magnetic textures such as domain wall or skyrmions), or with binary stochastic devices (as demonstrated for filamentary RRAM, and STT-MRAM). In particular, NEUROTECH member CNR has shown RRAM to emulate analog synapses and demonstrated how this dynamics can be exploited to improve the memory lifetime of spiking neural networks based on mixed CMOS-RRAM architecture[11].

5.3.2 Neuron Implementation

The stochastic, volatility and non-linear properties of memristive device technologies are exploited to emulate neuronal behavior. Among promising technologies, we can mention FeRAM, VO2–based MIT devices, PCM, STT-MRAM, and spin-torque nano-oscillators (i.e. specific types of magnetic tunnel junctions, which can be driven into spontaneous microwave oscillations by an injected direct current). In particular,

[11] Brivio et al., ‘Extended memory lifetime in spiking neural networks employing memristive synapses with nonlinear conductance dynamics’, Nanotechnology 30(1):015102 (2019)

NEUROTECH member CNRS/Thales has shown how to use the non-linear dynamics of the latter for processing[12]

Several comprehensive reviews and books on the use of emerging technologies for neuromorphic computing have been written by the community. In particular:

★ Marković, D., Mizrahi, A., Querlioz, D., & Grollier, J. (2020). Physics for neuromorphic computing. *Nature Reviews Physics*, *2*(9), 499-510.
★ Spiga, S., Sebastian, A., Querlioz, D., & Rajendran, B. (Eds.). (2020). *Memristive Devices for Brain-Inspired Computing: From Materials, Devices, and Circuits to Applications-Computational Memory, Deep Learning, and Spiking Neural Networks*. Woodhead Publishing.
★ Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. (2020). Memory devices and applications for in-memory computing. *Nature nanotechnology*, *15*(7), 529-544.
★ Grollier, J., Querlioz, D., Camsari, K. Y., Everschor-Sitte, K., Fukami, S., & Stiles, M. D. (2020). Neuromorphic spintronics. *Nature electronics*, *3*(7), 360-370.
★ Shastri, B. J., Tait, A. N., de Lima, T. F., Pernice, W. H., Bhaskaran, H., Wright, C. D., & Prucnal, P. R. (2021). Photonics for artificial intelligence and neuromorphic computing. *Nature Photonics*, *15*(2), 102-114.

Today, the main challenges for beyond CMOS neuromorphic platforms are:
- Developing algorithms that take full advantage of device physics and are able to deal with device non-idealities (variability, synaptic programming, lack of available precision, etc.)
- Scaling-up to very large scale systems capable of performing real-life tasks
- Lack of availability of industrial quality devices for academics

In the next few years we expect the following advances:
- Larger hybrid CMOS/memristive systems using standard network topologies and algorithms
- Demonstrations of systems exploiting the physics of nano-devices for local self-learning
- Increase in devices performances (density, endurance, energy consumption)
- Increase in availability of emerging devices (PCRAM, MRAM, OxRAM…) in industrial fabrication processes as well as corresponding design libraries
- Better methods and tools for designing hybrid systems
- Better understanding of which technology is appropriate for which type of computing / application.
- Demonstration of computational building blocks with more emerging technologies (spin textures, quantum hardware, 2D materials, organic materials…)
- Larger scale and integrated realizations of photonics-based systems

[12] Romera et al. 'Vowel recognition with four coupled spin-torque nano-oscillators.'. Nature 563.7730: 230-234. (2018).

In the long term we expect the following advances:

- Automated tools for the development and design of hybrid chips including emerging devices
- Possibility of commercial fabrication of hybrid chips
- Adapted algorithms
- Dedicated applications where novel technologies are required
- Fully functional large scale platforms merging different technologies

6 Challenges for Neuromorphic Computing

In order to unlock its potential and provide the applications described above, neuromorphic computing must overcome several challenges. Discussions within the consortium, the work groups, at the Forum and at online events have allowed us to come up with a list of limitations and challenges that neuromorphic computing currently faces.

Neuromorphic computing is mostly a recent field of study. Although some work had started in the early days of computing, the recent progress both in artificial intelligence and in emerging technologies has brought a new boom in neuromorphic computing. This has opened the door to many subfields, technologies and research directions. This novelty of the field also implies a lack of maturity, which comes with challenges that can be classified into four main categories.

For each challenge, we provide paths for moving forward as well as examples of concrete actions taken.

6.1 Lack of Theoretical Foundations

■ Neuromorphic Computing in General

There are no clear theoretical foundations for neuromorphic computing. It is neither clear how exactly the brain works, nor which aspects of this working should be emulated by neuromorphic computing.
In particular, event-based algorithms for signal processing, inference and control need to be developed and validated with regards to their reliability, efficiency and feasibility in neuromorphic systems.

■ Learning

Learning is a crucial element of computing systems inspired from the brain. In software artificial neural networks, it consumes huge amounts of data, time and energy. Neuromorphic computing aims at finding better approaches. However, these are still lacking clear solutions.

In particular, neuromorphic computing aims at developing:

- Training approaches using only local information
- Better training & optimization of spiking neural networks

- Efficient online learning
- Better understanding of unsupervised learning

Relationship to Novel Substrates and Architectures

Neuromorphic algorithms and architectures must be co-designed with their substrate. Theoretical foundations on how to achieve this are lacking. New bio-inspired concepts should be selected and optimized for their compatibility with electronic implementation. Computational models and algorithms for non-Von Neumann architectures (beyond neural networks, non-linear oscillator networks, Ising machines, optimizers, etc...). must be developed. The scalability concepts and laws (equivalent of Denard scaling for neuromorphic computing) are lacking and would be useful.

Paths for Moving Forward

- Algorithm development for Spiking Neural Networks and particularly for time-based computation and using sparsity.
- Investigate what is the correct abstraction level for neural computation (neurons and synapses vs. populations vs. cortical columns).
- Dedicated research on these topics and collaborations with neuroscientists and computer scientists must be conducted. These must keep in mind how to translate theoretical findings in usable hardware.
- Access to the most technologically mature neuromorphic computing platforms will enable researchers to test novel algorithms.
- *Examples of concrete action taken:* the 8th Annual Neuro-Inspired Computational Elements workshop, organized by NEUROTECH partner University of Heidelberg on March 16-19th 2021, provided dedicated tutorials to neuromorphic computing platforms BrainScales 2, Dynapse, SpiNNaker and Loihi. Furthermore, it featured many pieces of work tackling the development of hardware friendly bio-inspired learning algorithms. Videos of the presentations can be found on the YouTube channel of the workshop. The 6th Educational NEUROTECH online event (April 13th 2021) focused on learning with spiking networks.

6.2 Lack of Technological Maturity

Novel Technologies Themselves

Technologies beyond CMOS transistors in the digital regime suffer from low maturity. Some examples of such issues are: variability in analogue CMOS circuits, lack of endurance in memristive switching devices, difficulty to achieve analogue non-volatile memories.

Accessibility

Neuromorphic systems and devices are hard to access. The community should work on making hardware available, packaged for use, reliable and affordable.

The development of versatile neuromorphic building blocks to be integrated into larger systems is a possibility.

- **Paths for Moving Forward**

- To address this issue, the community should work both on material and device development and on novel computational paradigms that function in spite (or even thanks to) the issues faced by emerging technologies.
- It is critical that the development of novel technologies is pursued not only by academics but also by the semiconductor industries so that neuromorphic systems can be brought to maturity as well as mass produced.
- *Examples of concrete actions taken:* funding targeting the push of neuromorphic computing technologies to higher TRL, such as the Joint Undertaking ECSEL Andante project ; initiatives federating actors around specific technologies, such as the Spintronic Factory co-founded by Thales.

6.3 Lack of Standardized Tools and Benchmarks

Lack of Whole Stack from Hardware to Software

Conventional computing has benefited from multi-decade development of the stack from hardware to software. This is not yet the case for neuromorphic computing. The different layers of the stack are not independent or well defined. Knowledge of the whole stack is important to develop neuromorphic systems. Working on the maturity of the stack would make it easier to address each issue and facilitate scaling up of systems to more complex networks and tasks.

Lack of Tools for Development

There are not yet standard tools for developing, inspecting and debugging neuromorphic systems. For instance, having a tool comparable to TensorFlow, a platform for deep learning that comprises a collection of deep learning algorithms and architectures that could be used for spiking neural networks would be of great use. Actors in the field are increasingly aware of this situation and have initiated this direction (ex: *Applied Brain Research*) but this effort needs to get much more global in the community. Open source tools are much more likely to be adopted by academics and industrials alike. Similarly, tools that support and facilitate the deep inspection of event-based spiking networks during run-time would be of great help in developing algorithms for large-scale neuromorphic systems.

Lack of Benchmarks and Targeted Applications

Neuromorphic computing is not necessarily efficient for the same applications as conventional software neural networks. New standard applications and benchmarks are still lacking for neuromorphic computing. As a consequence corresponding datasets are also lacking.

☐ **Paths for Moving Forward**

- Putting together and sharing datasets
- Funding projects that produce the relevant tools
- Active communication within subfields around benchmarking
- *Example of concrete action taken:* Sharing datasets relevant to neuromorphic computing, in particular event-based, on the Neurotech portal (https://neurotechai.eu/resources/datasets/)

6.4 Lack of a Solid Community

A strong and well identified community is critical for a scientific field of study, especially for a new and growing field. In the case of neuromorphic computing, this need is especially important but also complex to achieve. This is due to the heterogeneity and interdisciplinarity of the field. Neuromorphic computing brings together actors from computer science, neuroscience, physics, electronic engineering, material science and more. Academics, industrials and SME are involved. Such a diversity is a huge opportunity for the field both on the scientific and human sides. However, it requires special effort to make people from such different backgrounds communicate and collaborate.

While community networks and events can self-organize in more narrow and mature fields, this should not be expected for neuromorphic computing, where a conscious action is needed. The Neurotech consortium and the resulting events and actions are a first step. As a striking example, many forum participants confided that this was the first neuromorphic computing event they had the opportunity to attend. More educational materials are also required to keep the community up to date with developments that are not in their core expertise, as well as to involve new actors.

It is critical that such actions continue to be encouraged, both at the individual and institutional levels.

☐ **Paths for Moving Forward**

- Organize in person and well as virtual gatherings that are inclusive to the various subfields and actors of the community.
- Emphasize educational content that can put everyone on the same page and playing field regarding knowledge.
- Encourage collaborative projects where different disciplines and actors are present.
- Encourage interactions between relevant projects and the community.
- Encourage relevant training in academic programs, especially at graduate level.
- *Example of concrete actions taken:* educational content such as the Neurotech Educational online seminars, organization of summer schools open to all career levels, production of handbooks, organization of recurrent gathering to create continuity, interactions around European projects such as the Neurotech Science and Technology Workgroup Webinar

7 Needs for Adoption by Industry

Despite the importance and large span of applications for neuromorphic computing, a number of roadblocks need to be overcome in order to achieve adoption by industry. Here we present the main roadblocks described by industrial actors and provide paths for moving forward.

7.1 Applications

The community needs to find some “killer apps” that will demonstrate the potential of neuromorphic computing.

These demonstrations should highlight the fact the neuromorphic chips are competitive with existing solutions and in particular software based deep neural networks. There are two ways for this:

- Solve questions that deep learning is not capable of solving (such as reasoning, generalization to multiple tasks, catastrophic forgetting etc.).
- Solve questions that deep learning is capable of solving but doing it more efficiently that accelerators can do.

Finding these applications requires:

- Interactions between research actors and end users.
- Increasing TOPS/mm2 and TOPS/W by orders of magnitude for conventional neural networks (CNN, LSTM, FC, ...)
- Clear benchmarking of existing and proposed solutions, close to real applications.
- If neuromorphic computing cannot compete with software neural networks in general, finding areas where it can., e.g. in the domain of low-latency event-based systems.
- Performing demos on niche tasks.

7.2 Maturity

Neuromorphic computing is still immature technologically. Steps to make it more usable will require:

- Increasing the technology readiness level of the beyond-CMOS technologies
- Improving our understanding of neuromorphic computing to avoid a black-box situation
- Definition and theorization of algorithms and computing/programming paradigms that use neuromorphic computing, for instance spiking neural networks, for performing engineering tasks.
- Improving the scalability of devices and architectures

7.3 Ease of use

In order to be adopted by industrials beyond pure research and development, neuromorphic computing should be easy to use. This requires:

- More tools and infrastructure for development and debugging
- Development of reliable compiler software stacks
- Design of user-friendly GUIs that can help end-users to write neuromorphic networks, such as spiking neural networks, that performs practical tasks.
  Training people to have knowledge the whole stack (materials, devices, systems, algorithms, applications)
- Providing easier access to existing systems and platforms
- Developing methods for easier training and programming
  Tacking the large amount of data needed for the training by developing systems that require less data and making more data available
- Catering to the development of communities to make skill transfer and collaboration easier.

8 Actors and Roles

The growth of the field requires the synergy of different actors, filling different roles.

One role is to conduct the research, from fundamental to applied. This role is filled by academic laboratories, Research and Technology Organizations, and the research departments of companies. Technology providers and in particular foundries have a critical role for enabling other institutions to conduct research. It is important that these actors work together for a smooth increase of TRL of neuromorphic systems. There is a large number (over a hundred) of collaborative research projects funded by the EU. A list and cartography of these projects can be found in the Neurotech Deliverable D1.2.

Another role is to use neuromorphic systems for applications. Many companies are interested in the applications offered by neuromorphic systems, either to use them themselves or to include them in their products. It is critical that research actors communicate with them to provide them information about the state of the art and possibilities, as well as get quantitative requirements for applications, so as to have target KPIs for the development of systems.

Some companies are in the interesting position of being able to both conduct research and then develop and sell applications. Large companies (*Intel, IBM, Qualcomm, Hewlett Packard, Samsung Electronics*, ...) that have the ambition to be key players in the field can afford to invest a large effort in the development of neuromorphic systems. They can also afford to pursue the development of different technologies and applications in parallel. Their involvement will be critical to increase the maturity of systems and reduce their time to market. On the other hand, start-ups have the advantage of being flexible and willing to take risks on novel technologies and disruptive algorithms. They can quickly prototype proof-of-concept demonstrators and concentrate on emerging applications to make their breach on the market.

The list of SMEs searching to develop and commercialize neuromorphic technologies, both CMOS and hybrid technologies and targeting wide markets keeps growing. Without aiming for exhaustiveness, we list some of the very promising ones below.

Neural Chips and platforms:

- **SynSense** (former AiCTX) – a spinoff from IniLabs (Univ. of Zurich, ETHZ Zurich) developing the DYNAP class of processors aiming for IoT and medical signal analysis among others (https://www.synsense-neuromorphic.com)
- **GrAI Matter Labs** – develops a very low-power spiking processor for edge applications aiming for markets such as drones, industrial automation, AR/VR, robotics (https://www.graimatterlabs.ai)
- **LightON** – a photonics startup developing the Optical Processing Unit co-processor for massively parallel machine learning computations (https://lighton.ai/)
- **BrainChip** – develops the AKIDA spiking neural processor, also for edge applications (https://brainchipinc.com/)
- **Rain Neuromorphics** – a startup developing a physics-based analog neuromorphic processor (https://rain.ai)
- **AnotherBrain** – focusing on energy-efficient chipset with self-learning, explainable algorithms (https://anotherbrain.ai/)
- **Applied Brain Research** inc. – develops a multi-HW platform and the software codesign and synthesis suite called "Nengo", endowed with different learning algorithms and neural models ranging from conventional computing to spiking ASICs, FPGAs and controllers (https://appliedbrainresearch.com/)

Dynamic Vision Sensors:

- **Prophesee** – produces an event-driven vision sensor and SDK for very low-power data-efficient computer vision systems (www.prophesee.ai)
- **IniVation** – another spin-off from iniLabs producing also an event-driven vision sensor – announced mid-april 2021 as the first neuromorphic sensor to embark a satellite and reach space (https://inivation.com/)

More broadly, we have compiled a list of companies with interest in neuromorphic computing, published in Deliverable D2.4.

The role of the Neurotech consortium is to federate these actors and create synergies. Our main actions taken include:

- Organization of Forums for gathering actors and discussing issues
- Educational seminars to create common knowledge in the field as well as provide introductory materials for newcomers
- Work group events to discuss current issues. The Industry workgroup presents companies working the field, the Science and Technology workgroup create

synergies between projects, the Ethics workgroup provides a platform to debate issues around ethical concerns

- Provide resources for the field: videos of presentations, datasets for computing tasks, references of key publications
- Keep up to date with the European projects in the field as well as with the companies active or interested in the field

9 Outlook and Next Steps

We have investigated the challenges to neuromorphic technology as well as collected opinion from industrial actors (leveraging the Forums and the Industry Work Group) about the need for industry adoption. From these intersecting concerns, we extract 5 axes of growth for the field. Each axis corresponds to a timeline that we present here. Note that these timelines are meant to be modified and completed as the field evolves.

9.1 Field and Community Maturity

This axis is linked to both the novelty of the field and the fact that it is intrinsically multidisciplinary. It corresponds to the “Lack of solid community” challenge as well as an observed request from industry to have more information and educational material about NCT.

The first step was to better define neuromorphic computing. We laid out our vision of the field in the present deliverable and published it in a perspective article[13].

The next steps include:

- Providing educational material about NCT, available to all.
- Organizing events with people from different subfields and backgrounds
- Encouragement from funding agencies for collaborative NCT research project where different disciplines are merged

The timeline below illustrates key steps ahead:

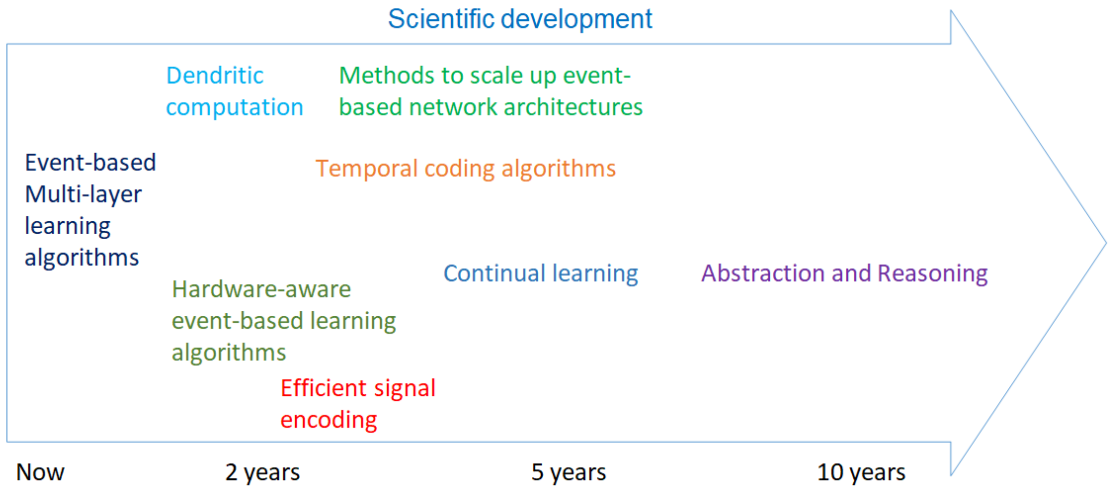
[13]Donati, E., et al. "Neuromorphic technology in Europe: Brain-inspired technologies are advancing apace across Europe and are poised to help accelerate the AI revolution." *The Innovation Platform* (2020).

9.2 Scientific Development

This axis is the core of NCT research. The next steps include:

- Stronger theoretical foundations
- Development of learning mechanisms

The timeline below provides an overview of the key steps ahead.



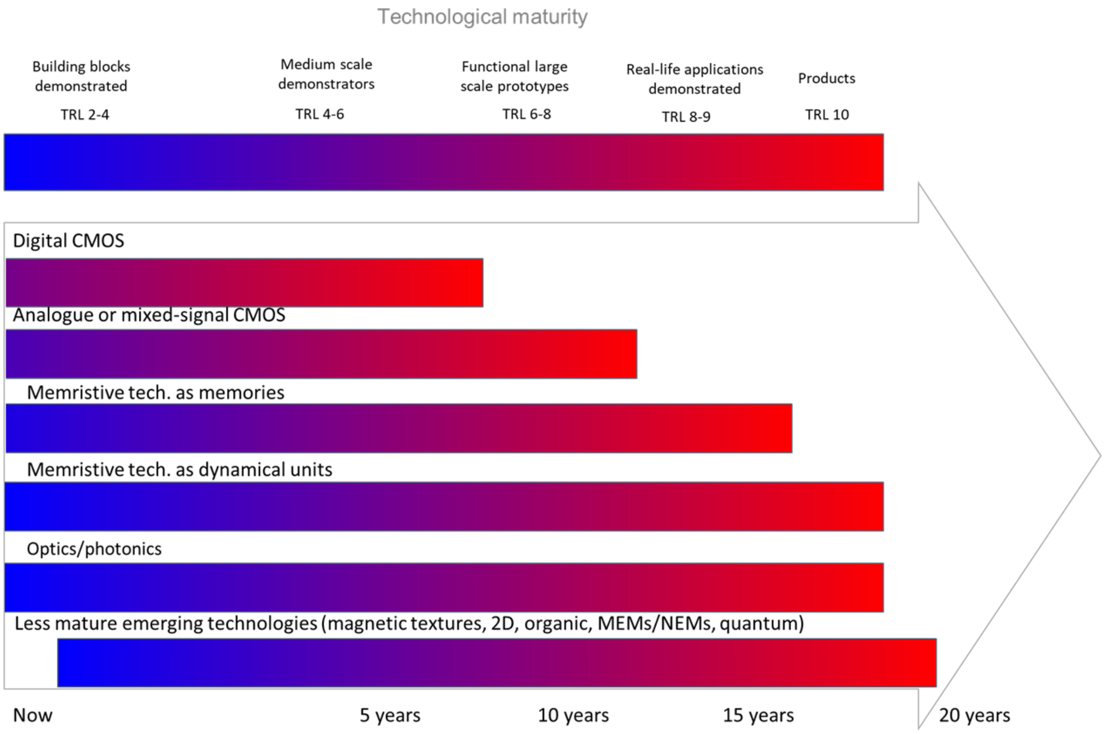
9.3 Technological Maturity

This axis was both identified by the actors and the field and the end-users as a critical need. We have identified the most popular hardware subfields (see above) and produced a state of the art, leveraging the Science Work Group, available in Deliverable D1.2.

Note that investment from large companies into specific technologies will speed up the maturity increase significantly. This applies to the quality of the devices themselves as well as their availability and to the tools to develop systems using the technology.

It is thus of critical importance that the technology providers are made to see the interest of investing in novel technologies. One important path for this is incentives from policy makers to EU technology providers to develop NCT.

The timeline below highlights the predicted maturity evolution of the main technologies:



9.4 Tools for Development and Use

Both actors within NCT research and end-users are in need of specific tools for development and use. These include dedicated benchmarks, datasets, automated design tools, programming libraries, etc.

To favor uptake and the building of a larger community of users and stakeholders, the neuromorphic community should commit to the design of modular and reusable sensing and computing modules, starting from the standardization of the protocol of communication, i.e., Address Event Representation. Open-source implementations of algorithms and datasets sharing will push the progress of the field, also using common datasets for benchmarking. A milestone in this path is the definition of benchmarks that are valuable to validate neuromorphic systems. Commonly used datasets for SNN architectures and learning rules are derived from standard machine learning for vision, namely the event-driven version of the MNIST (N-MNIST). While it is important to compare against the mainstream community, these datasets do not capture the spatiotemporal analysis capability of SNN, nor can they test their adaptation and noise robustness features. While the neuromorphic vision community is tackling this issue

with the acquisition of *ad-hoc* datasets (e.g., the DVS-gestures dataset), the neuromorphic community at large needs benchmarks for the hardware platforms, cognition, action planning, control and execution modules and for the fully integrated system.
At the same time, neuromorphic technologies require interface when need to be interfaced with application, i.e., robotics. Working with streams of events, instead of static frames or batch, requires the development of *ad-hoc* interfaces and software libraries for handling the events. Currently, open-source JAVA and libraries -- ROS and YARP -- have been developed within two of the main robotic middlewares. However, they require contributions from a large community to grow and reach the maturity needed for successful adoption in robotics.

7.5 Applications

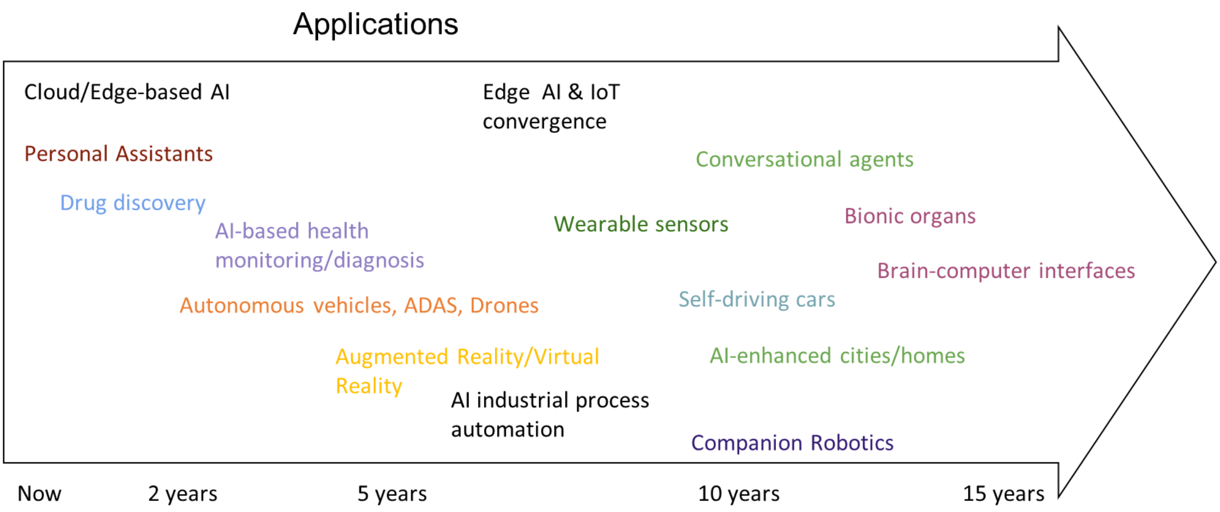
Application scope: The scope itself of neuromorphic computing applications will evolve. The evolution is from Narrow AI, to Broad AI, to General AI.

Narrow AI is focused on addressing very focused tasks (such as buying a book with a voice-based device) based on “common knowledge.” That’s the reason narrow AI is scaling very quickly in the consumer world where there are a lot of common tasks and data to train these systems.

Broad AI is about integrating AI within a specific business process of an enterprise where you need business- and enterprise-specific knowledge and data to train this type of system.

General AI refers to machines that can perform any intellectual task a human can.

Application type: Applications will evolve from cloud based to AI to AI at the edge to AI everywhere. The Schematic below presents the main application steps ahead.



Conclusion

Neuromorphic computing is a booming field with extremely promising applications, but which faces challenges due to its novelty and interdisciplinarity. The Neurotech consortium brings together actors from academic research, RTOs and industry. This first roadmap serves as a starting point to federate the field by stating clear goals as well as identifying challenges and paths to overcome them. In the next few years, we will be guided by this roadmap and will in turn update it as we move forward.