Brain Computation: A Computer Science Perspective

Wolfgang Maass, Christos H. Papadimitriou, Santosh Vempala, and Robert Legenstein

Institute for Theoretical Computer Science, Graz University of Technology, Graz, Austria
EECS, University of California, Berkeley, CA, USA
College of Computing, Georgia Tech, Atlanta, GA, USA
Institute for Theoretical Computer Science, Graz University of Technology, Graz, Austria
maass@igi.tugraz.at; christos@cs.berkeley.edu; vempala@gatech.edu; robert.legenstein@igi.tugraz.at

Abstract. The brain carries out tasks that are very demanding from a computational perspective, apparently powered by a mere 20 Watts. This fact has intrigued computer scientists for many decades, and is currently drawing many of them to the quest of acquiring a computational understanding of the brain. Yet, at present there is no productive interaction of computer scientists with neuroscientists in this quest. Research in computational neuroscience is advancing at a rapid pace, and the resulting abundance of facts and models makes it increasingly difficult for scientists from other fields to engage in brain research. The goal of this article is to provide — along with a few words of caution — background, up-to-date references on data and models in neuroscience, and open problems that appear to provide good opportunities for theoretical computer scientists to enter the fascinating field of brain computation.

1 Introduction

We have known since antiquity[1] that our brain gives rise to our perceptions, memories, thoughts and actions, and yet precisely how these phenomena arise remains the greatest scientific mystery and challenge of our time. This is despite massive, brilliant and accelerating progress in our understanding of the brain, its structure and molecular basis, its development and pathology, its neurons and its synapses, as well as the complex ways in which they are modified by experience[2].

[1] In the early 5th century BCE, Alcmaeon of Croton proclaimed the brain "the seat of intelligence," conjectured that it is connected to sensory organs through channels, and discovered and dissected the optical nerve. Disappointingly, in his response to Alcmaeon more than a century later, Aristotle argues instead that intelligence springs from the heart...

[2] [1] is a standard graduate and [2] a standard undergraduate textbook in Neuroscience, while [3] is a mathematical treatment of the subject.

How does the mind emerge from the brain? It seems very plausible, and has been strongly suggested over the decades [4–6], that the eventual answer to this question will be at least partly computational. We therefore believe that computer scientists, and theoreticians in particular, should work on this problem. And yet, despite important early connections between computer science and the study of the brain (see the brief historical account in Section 2), there is at present no community of computer theorists studying the brain[3]. Furthermore, there is no articulated suite of models, research questions, and early results in the interface between computer science and brain science, inviting computer scientists to participate in this grand quest[4]. This is significant, because such entry points have in the past marked the beginnings of successful interactions between computer science and other scientific disciplines, such as statistical physics [10], quantum physics [11, 12] and economics [13, 14].

This is the context and thrust of this paper. In Section 2 we give a brief historical overview of past interactions between computer science and the study of computational aspects of the brain, and we articulate David Marr's vision of computational research on the brain, *ca.* 1980. In Section 3 we discuss aspects of the methodology of the computational study of the brain, focusing on algorithms of the brain, abstract and simplified models of brain systems, and learning. In Section 4 we describe current work by our group on computational models for the formation, association, and binding of memories in the medial temporal lobe (MTL), a brain region believed to be involved with such activities. We conclude with an array of research questions and fronts.

2 History

The pioneers of computation were keenly interested in the brain. Turing saw the human brain as the archetype of computation [15], and later, famously, as an important challenge for computers [16]. Von Neumann in a posthumously published essay [17, 18] compares the brain with the computers of his time. He observes that the brain is larger in number of elements (still is, but it is getting close), but slower (much more so now); he notes the analogue nature, but digital operation, of neurons and synapses, acknowledges the key role played by biology and genes, and ponders the brain's architecture (having himself pioneered the computer's). Remarkably, he hypothesized already that the brain is likely to carry out computations on a statistical level with algorithms that are *"characterized by less logical and arithmetical depth that we are normally used to"*. McCulloch and Pitts [19] and later Rosenblatt [20] proposed stylized neuron-like elements as a possible basis of brain-inspired computation, initiating a rich research tradition which eventually brought us deep learning (on which more later).

[3] In contrast, there is a well developed theoretical field of investigation for the related field of Machine Learning, namely the COLT community.

[4] L. G. Valiant's work starting from the 1990s [7–9] is a notable exception discussed extensively later.

In 1980, computational neuroscience pioneer David Marr proposed an influential three-level approach to understanding brain computation [21]:

- At the *computational or behavioural level* (today we would call it *specificational*) one identifies the input-output behavior of the system being studied; we refer to this as the first level.
- At the *algorithmic level*, one seeks to understand the organizations and dynamics of the particular processes and representations used by the system; we refer to this as the second level.
- Finally, the *biological implementation level* entails identifying the biophysical elements (e.g., neurons and synapses) and molecular mechanisms employed by the system to realize the algorithm; we refer to this as the third level.

We shall use Marr's taxonomy as the basic framework of our discussion of computational approaches to the brain.

3 On Methodology

Can we hope to use Marr's method to discover the overarching algorithmic principle underlying all of brain computation, the coveted *algorithm run by the brain?* In articulating his three-level proposal, we believe that Marr was expecting the various systems in the brain (probably hundreds of them) to have each its own function and specification, and its own algorithm and hardware. One should expect *large-scale algorithmic heterogeneity* in the brain — a plethora of principles, methods, procedures, and representations — and one has to be prepared for the long haul of understanding them one by one. (But see [22, 23] for a recent principled attempt at a compilation of a broad range of elementary computational tasks at Marr's level.)

There is a subtlety in Marr's level two, where we infer the algorithm used by the system: We know from the theory of computation that there are infinitely many algorithms for the same task, and furthermore classical universality results [24, 25] imply that neuron-like systems can in principle implement any process and algorithm whatsoever. Showing that one particular algorithm accomplishing the level-one task can be implemented in the hardware of level three, or that a class of algorithms can be so implemented (see for example [26]), constitutes no evidence whatsoever that this algorithm or class is actually used at level two. To solve the second level problem, one needs to rely on experimental results revealing properties of the hardware (level three), and use these to restrict the unlimited repertoire of possible algorithms.

In fact, one may speculate that the algorithmic second level may in many cases end up being simply the computational behavior of the hardware/third level: *The algorithm vanishes,* essentially because the hardware is well adapted to (probably has co-evolved with) the task, and the inputs (from sensors or other parts of the brain) as well as the parameters of the chemical environment are adequate for driving the hardware in an essentially "algorithm-free" way. In other cases, the algorithm may be disappointingly opaque and lacking in a meaningful explanation, perhaps because it is the result of a long evolutionary process of parameter setting though trial and error; recurrent neural networks often appear to be like this.

Computational work of the brain must get inspiration from, and be meticulously cognizant[5] of, the tremendously rich and informative current experimental work in neuroscience. In fact, one particular strand of this work seems especially well suited to enlighten the computational study of of the brain: *Connectomics* [27, 28], the ongoing herculean effort to create detailed large-scale maps of all actual neurons and synapses of animal brains. Would this project, once successful, facilitate — even obviate — the computational study of the brain? In pondering this question, it is useful to remember deep learning: We currently have at our disposal a wide variety of artificial neural network architectures solving sophisticated problems, and *we know to the last detail* the precise structure, connectivity, and vast array of numerical parameters of these networks. And yet we are lacking a meaningful explication of how each of these systems solves the problem at hand. Further, one should keep in mind that a static connectome of the brain does not exist, at least for higher vertebrates such as mice. Instead synaptic connections in the brain are known to rewire themselves on a time scale of hours to days [29–31]. Hence, any connectome can only be a momentary snapshot of a dynamically changing brain structure, and brain computation has to be understood in the context of this dynamics.

Models. The study of the brain often employs *models* of the brain (or, more commonly, of parts thereof). Models are important and useful, but must be created and used with care. *Abstract models* create mathematical abstractions — that is, generalizations — of the realities of the brain or a subsystem thereof. In employing an abstract model, one must remember that it is a generalization; this means that *some but not all* of its specializations will be reasonable models of the brain. In addition, an abstract model may not be sufficiently abstract, in the sense that models of biological neural networks that take into account experimentally verified and functionally relevant features of biological neurons or synapses may *not* be specializations of the abstract model. For example, we know that weights of synapses are subject to use-dependent short-term plasticity; apparently every biological synapse has an individual short-term plasticity, which implies that its effective weight for the second spike in a spike train is smaller or larger than for the first one, and assumes yet another value for the third spike, depending on the interspike intervals and the specific type of synapse (see Section 1 of [32] for references). This feature of biological synapses does appear to be functionally relevant, and provides clues about the types of algorithms that can be implemented by biological networks of neurons. On the other hand, it sets such networks apart from Boolean circuits and artificial neural networks, which require that the parameters of the units remain stable between steps.

[5] The use of "killer adjectives" such as *biologically plausible* is a poor substitute for computational models and results informed by experimental knowledge.

Another genre of models are *simplified models.* Brain systems are often of tremendous complexity, and it is difficult and unwieldy to include all that is known from experiments in a single manageable model. In such cases, a *simplified model* can be invaluable for capturing the system's salient aspects, disregarding effects and interactions which seem largely inconsequential. However, in employing a simplified model one must remember what was thrown away, and in the end of the analysis go back to determine, for which kinds of predictions is the model suitable, and for which it is not. Simplified models are often further modified and implemented as *brain-inspired computational engines* for solving actual computational problems. This is of course valuable, but again one must remember that the success (or failure) of such engines may have little to teach us about the way brains work (deep learning comes again to mind).

Learning, Environments, and Language. One cannot engage in a computational study of the brain without considering how the brain is changed by the animal's experience — that is to say, how *learning*[6] happens in the brain. By "learning" one means changes occurring in the brain through interactions with other parts of the brain and, importantly, with the surrounding environment. Processes that implement learning are part of a large repertoire of plasticity processes that take place in the brain simultaneously at many different time scales, and whose function is only partially understood. Further, one cannot claim to understand the brain without also considering the brain's environment and its challenges. One subtlety here is that the environment is *affected* by the brain's activity — in the short term through motor action and animal interactions, in the longer term through design of the environment (dwellings, signs, etc.).

Language is itself an important environment (since utterances are the input to a specialized yet overarching brain activity). This environment was designed from scratch, and, in evolutionary terms, *extremely* recently [34], at a time when the human brain had already been developed essentially to its present form. Human language is, so to speak, a last-minute adaptation. Furthermore, it has undergone its own vigorous evolutionary process over a window of very few thousands of generations. It seems natural to posit then that language has evolved to be well adapted to the human brain's strengths — for example, so it can be learned easily by babies. We believe that language is an especially important and opportune arena for the computational study of the brain and the mind.

4 Models of Memories and Cognitive Computation

Much current experimental work explores the nature and function of *memories:* the representation in the brain of distinct concepts, such as persons we know, places where we have been, or words we use. It is estimated that many tens of thousands of such memories are represented in the human brain, along with *associations* between them. We believe that memories, because of their discrete and

[6] In fact, Poggio [33] proposes that learning is so fundamental for brain computation so as to constitute an extra top level of Marr's hierarchy.

symbolic nature, and their close relationship with language, are an interesting place for theoretical computer scientists to start thinking about the brain.

Valiant's model. Leslie Valiant's *neuroidal model* was proposed in 1994 as a possible basis of a computational theory of the brain, and ultimately of cognition. He posits a random directed graph of neuroids (model neurons with discrete internal states) as nodes, and synapses as directed edges. Parameters of the neuroids and the synapses (e.g., internal state, threshold, strength, etc.) are modified in clocked discrete steps in a distributed, automaton-like manner. Valiant used this model to develop his theory of memory based on *items*. An item is a set of neurons whose simultaneous firing is coterminous with the subject thinking one particular thought (such as "apple"); items may or may not overlap, yielding two different models. Valiant defines Boolean-style operations on items: Join (e.g., "apple" may be joined with "green" to form a new item which will fire every time the two constituent items fire together) and Link (e.g., "apple" linked to the item representing the class "fruit"). The operations of Join and Link can be implemented within the neuroidal model by deterministic algorithms that switch between states of neurons and synapses, including synaptic weights and thresholds — the algorithms must switch rather arbitrarily between states in order to achieve the desired functionality — and by exploiting the random nature of the underlying directed graph to recruit and manipulate new neurons.[7]

Valiant's model was a brave and inspiring early attempt to make computational sense of the brain. In the two decades since the publication of [7], experimental neuroscience has provided much insight into various details of computation and plasticity (learning) of networks of neurons in the brain; some of these findings align well with the premises and predictions of Valiant's model, but others do not. Even though the complete rules for synaptic plasticity (the ways in which synaptic weights change in response to neural activity, effecting learning) are still not known, we now understand that Hebbian plasticity (changes in synaptic weights resulting from the near-simultaneous firing of neurons) can increase synaptic weights by some limited amount within a given time window, say, by 100% within a day; see e.g. [37], and furthermore there is a lot of variability in this respect among different synapses, and within the same synapse over time. Hence it cannot be assumed that synaptic weights can be set to an arbitrary and precise value during learning.

Similarly, as we discuss below, neural recordings both from the animal and the human brain [38] suggest that salient concepts are indeed encoded in the brain through distributed "assemblies" of neurons, so that a fair portion of the neurons in an assembly will fire whenever the corresponding concept is invoked. However, these assemblies are not static entities, since the concrete set of firing neurons

[7] Recently, Valiant's theory was extended by the introduction of the *predictive join*, or PJoin [35], a more algorithmically apt version of Join, which however is subject to the same criticism. It is an interesting question as to whether the conceptual primitives of Join, Link, PJoin, which enable rich computation [36], can be implemented in more realistic models.

varies substantially from trial to trial, presumably in dependence on the context, and, as we discuss below, the underlying set can be changed by experience. Also, even though, as we shall see, there is now evidence that associations somewhat akin to the ones predicted by Valiant's JOIN do happen in the human brain, such associations appear to be of a different nature and form than JOIN: Associations seem to be recorded by the assemblies "bleeding" into each other, as opposed to collaborating to create an altogether new assembly[8].

The Ison et al. experiment. In a very recent experiment [40], the formation of associations between memories in the human medial temporal lobe (MTL, a brain region with about a billion neurons in humans long thought to be crucial to the representation of memories) has been documented. They recorded from a few neurons[9] in the MTL of a human subject to whom many (over a hundred) pictures of known people and places were shown in a precise protocol. They found a particular neuron that fired every time the Eifel tower was shown, but not when Barack Obama was shown[10]. Then a combined image of the two was presented, and the neuron duly fired (as it always did when the Eifel tower was in sight). Remarkably, when a picture of Obama was presented, the neuron also fired: the subject had learned the connection, or *association,* between Obama and the Eifel tower! And the recorded neuron was a part of the representation of this association. The principle that associations between memory items are accompanied by overlaps in the corresponding assemblies was confirmed more recently also for longterm representations of associations [41].

Neural network models of memory. Memories and their associations, especially in view of the experimental results just described, constitute a very concrete description at the first (specificational) level of Marr's framework, begging important questions about the third and second levels: How are memories represented in the animal MTL, how are these representations created, and how are they altered to record associations between memories?

We start by proposing an answer to the third-level problem: There are by now ample reasons to believe that *assemblies of neurons* play an important role in answering these questions. A *neuronal assembly* is a set of neurons that are likely to fire together, or at proximal times. It has not been established that the neurons in an assembly are interconnected by strong synaptic connections, but this is a reasonable hypothesis (in Valiant's model, intra-item connections do not matter). Assemblies were conjectured by Hebb [42] already in 1949 (who depicted them as Hamilton paths of strong synaptic connections). Since researchers have discovered in human subjects neurons responding to the Eifel tower or Jennifer

[8] Earlier experiments with rodents and monkeys did however find neurons that only responded to a specific combination of stimulus features but not to any of these features in isolation, see e.g. [39], supporting in this case Valiant's version.

[9] There were many human subjects, and a total of hundreds of recorded neurons, see [40] for details, but here we focus in this exposition on one subject and one neuron.

[10] Illustrating example.

Aniston [43, 40] by recording from only a few hundreds of randomly chosen neurons in MTL, and presenting a few hundreds of familiar stimuli, it is plausible that many more neurons (in the tens of thousands at least) respond consistently to this same stimulus. Further, it is tempting to assume that the reason these groups of neurons fire together after the image presentation is because they form an assembly. Neural computation in the rodent brain has also been found to be dominated by activations of assemblies of neurons, and in fact transiently active assemblies of neurons seem to have replaced attractors as the putative tokens of neural network activity, providing a link between single neurons and entities on the cognitive level [38]. However, a theory of neural computation with assemblies is still missing at this point.

How exactly does an assembly, corresponding to a particular memory, materialize in the MTL? And how are associations between two assemblies formed, in a way that explains the experiment in [40] (Obama causing the Eifel neuron to fire)? Ongoing simulations [44] demonstrate that a model neuronal system, with parameters for synaptic connectivity and plasticity of synaptic weights that are compatible with what we know about the MTL exhibits similar behavior:

- when presented with particular input patterns for long enough, neurons tend to form groups that fire consistently when the same pattern appears later; and
- when presented simultaneously with two such previously encountered patterns, some of the neurons in the two corresponding groups subsequently respond to both patterns.

Hence the formation of assemblies and the creation of associations between them can be reproduced in silico.

A theoretical model. It is difficult to model synaptic plasticity in a neural network so that the model (a) is consistent with experimental findings and (b) remains theoretically tractable. One approach used in the past is to analyze equilibrium points of the dynamics of synaptic weights in a network, see [45]. We have found that equilibrium analysis of a simplified model, along with a novel variant of random graph theory, can be applied to elucidate mathematically the emergence of assembly codes, and the formation of associations, in recurrent neural networks [46].

Equilibrium analysis of a linearized model of plasticity deals with the *expected behavior* of the synaptic weights and neurons in the system, predicting that the neurons in the assembly will be chosen at random, but with the neurons most affected by the stimulus assigned higher probability (such behavior was recently observed [47] in the formation of *olfactory* memories in the piriform cortex). To predict concrete behavior of the system and the formation of a stable assembly, we assume that the neural network (of pyramidal cells) is randomly and sparsely connected. This appears to be a reasonable simplified model in view of experimental data [48]. It appears that a plausible model is a $G_{n,p}$ [49] directed graph with an added bias for "pattern completions" [50] (such a model had been proposed for different purposes in [35]): Conditioned on the existence of edges (a, b) and (b, c), for example, edges (b, a) or (a, c) are many times more likely to exist than by chance defined by the baseline parameter p. Preliminary analysis indicates that such a model may succeed in predicting the formation of stable assemblies, and their modification (two assemblies shifting their support to form a large intersection) through the formation of associations in response to mixed stimuli.

Binding. A fundamental capability of the brain, especially the human brain, is to form and apply abstract rules. Such a rule could specify how to behave in a particular social context, how to pick up an object, or how to form a syntactically correct sentence. Applying such rules requires to bind temporarily a variable in an abstract rule to a concrete context. For example, a simple sentence may consist of a subject, a patient, and a verb, and these must be bound to specific words during sentence formation. Recently, evidence has been emerging from fMRI imaging of the human brain [51] about the processes that occur during this binding process. Binding is related to Valiant's LINK operation. However, that operation connects coequal memories, whereas binding involves an abstract concept (such as "verb," possibly represented not by an assembly but by a whole brain area as suggested by the results in [51]) bound to an ordinary memory.

We propose that assemblies also play a prominent role during the binding of a variable to a context. Recent simulations [52] suggest that such binding operation can be implemented in a realistic neural model through so-called *assembly pointers.* Such pointer would connect an assembly representing "go" to a newly formed assembly within the intended brain area that represents the concept "verb", in a process similar to the assembly formation discussed above (with the "go" assembly now playing the role of the input stimulus).

Association Graphs. Occasionally, computational research on the brain will yield an interesting theoretical problem worthy of scrutiny through the methodology of theoretical computer science; we next describe briefly one such instance. As more and more memories and associations will be formed through life, an intricate network will be created [41], with intersections that are initially larger and then appear to shrink, and it would be of some interest to develop a theory of this aspect of cognition. It appears safe to assume that synaptic connections between the neurons of two assemblies A and B get strengthened when an association between the corresponding concepts is learned; this provides a plausible explanation for the previously described finding that both assemblies extend so that their intersection becomes larger (estimates range between a 4% and 40% of the size of a single assembly [41]). In an abstract model one can focus solely on these overlaps between associated assemblies, and ignore synaptic weights altogether. Such a network can be represented as an edge-weighted undirected graph (V, E, w) such that each vertex v is a memory, each edge $[u, v]$ is an association between memories u and v, and its weight w_{uv} represents the strength of this association, say the proportion of the neurons in the two assemblies that also lie in their intersection. We call such graphs *association graphs.*

One immediate question is, are all weighted graphs association graphs? The answer is trivially "yes" if no further assumptions are made, which can be shown through a straightforward modification of the Erdős construction of intersection graphs [53]. However, this construction may require that the size (number of neurons) of the assemblies/vertices differ considerably and that intersections are very small. What if we also insist that the assembly sizes are kept the same, or approximately so? This gives rise to an interesting theoretical problem. The requirement that the association graph be realized by intersecting assemblies by approximately equal size can be expressed as a linear program, whose variables are real numbers x_S representing the (normalized) number of neurons belonging to precisely all the assemblies in the set $S \subseteq V$. The constraints correspond to the vertices and the edges of the graph. One seeks to minimize the maximum relative difference between sizes of nodes. Interestingly, a related but more general problem had been addressed during the 1990s by philosophers [54].

It turns out that solving this linear program through the dual ellipsoid method is related to the *cut norm* ([55]), a well known deep subject in combinatorics. In collaboration with Nima Anari and Amin Saberi we have shown that the problem is in fact NP-hard, even to approximate within some n^{α} factor, but can be approximated in certain interesting special cases [56]. Another interesting variant is the one in which only the unweighted graph is given, with edges representing intersections of size above a threshold, while non-edges stand for intersections of size below a lower threshold. These first results suggest that not all association graphs can easily be embedded into the neural networks of the brain without causing missing or spurious associations. If this is the case, one might be able to relate this difficulty to particular deficiencies of the actual association graphs that are formed in the human brain.

There are many more questions and directions in connection to the graph-theoretic modeling of associations that seem worth exploring.

5 Open Questions

The purpose of the previous section was to describe ongoing work in just one possible direction — an important and opportune one, in our view — where methods that are common in theoretical computer science can support modeling, analyzing, and ultimately understanding brain function. The intended message of this article is that there are several such opportunities, not just in connection with memories but also with many other important questions and directions of research on brain computation; below is an assortment of such opportunities, starting with the ones closest to the described work.

- Given that cell assemblies seem central to our computational understanding of memory, defining them formally is of some importance. *What exactly is a cell assembly?* Is it a set of neurons whose connections through strong synapses cause them to fire simultaneously (and consistently in response to an input, or range of inputs), is it a pattern of firing activity of a set of neurons, or is it simply a distribution over a set of neurons? And, in each of these cases, in what sense and manner, and to what extent, is it transient? Obviously these questions require more experimental data.

- Neurons tend to have surprisingly different levels of activity (measured for example through their long-term average firing rate); this is true even for neurons of the same general type, e.g. pyramidal cells. Furthermore a few neurons are connected by really strong synapses while most are not [57]. These differences show up in statistical analyses as heavy-tailed distributions (often approximated by a log normal) of measurements such as long-term firing rates, synaptic weights, see e.g. [58, 59][11]. The question arises: what do these differences between neurons imply for the organization of neural computation? Do they point to an implicit hierarchical organization of neurons even within a single brain area, where more frequently firing neurons remember, process and transmit information in a coarser way — possibly even initialized through the genetic code — while less frequently firing neurons contribute refinements in a more flexible and experience-based manner?

- Another surprising invariant of neural activity in the awake brain is the scale-free (power law) distribution of avalanches of neural activity, i.e., of continuous episodes of neural activity within a patch of a brain area, or within larger brain areas, see e.g. [60, 61]. Scale-free distributed activity is commonly interpreted as a sign that the brain computes in a critical or near-critical regime [62]. Criticality of network dynamics could be an important clue for the large-scale organization of neural computations in the brain. However, several pieces of the puzzle are missing. Criticality is typically studied in deterministic dynamical system, while the brain is best modeled as a stochastic one; and we are not aware of a rigorous, computational understanding of criticality in dynamical systems. See [63, 64], and also [32], for references to first steps in this direction.

- A further surprising feature of brain activity is that it is not input driven: the brain is almost as active when there is (seemingly) nothing to compute. For example, the neurons in the primary visual cortex (area V1) are almost as active as during visual processing as they are in complete darkness [65]. Since brain activity consumes a fair portion of the energy budget of an organism, it is unlikely that this spontaneously ongoing brain activity is just an accident, and highlights a clear organizational difference between computers and brains. A challenge for theoretical work is to understand the role of spontaneous activity in brain computation and learning.

- Another ubiquitous and mysterious feature of neural network activity in the brain is the prominence of stereotypical spatio-temporal firing patterns

[11] In fact, such lognormal distribution of synaptic weights can be predicted theoretically from a simplified model of plasticity.

of neurons that occur both during active processing of sensory stimuli and spontaneously, see e.g. [66–68]. These experimental data undermine theoretical models that are based on an orderly bottom-up organization of encoding and computational transformation, where individual neurons report through their firing the presence of a specific feature of a sensory stimulus, or a specific value of an analog feature (for example in so-called population codes). These puzzles are nicely described in [69] for the case of area V1, which is one of the brain areas where neural coding has been studied the most. The presence of stereotypical spatio-temporal firing patterns of neurons points to a more implicit coding and computing mechanisms, and better computing paradigms and computational models are needed.

- As we have discussed briefly, language appears to be a most attractive research arena for the computational study of the brain. Can we define a biologically plausible *small set of primitives* sufficient for language learning and generation? We feel that assemblies, associations, and binding may be of some relevance to this quest.

- *Visual invariants* are one of the mysteries of vision: How is it possible that a plethora of very different images and sensations (an object such as a person's face, and its various translations, rotations, zoom-ins and -outs, occlusions, etc., not to mention the person's last name, or voice) are mapped instantaneously and unambiguously to the same "memory"? We suspect that the processes of assembly formation and association may provide insight to this problem, see [70, 71] for experimental data and related theories.

- Randomness, its nature and utility, is one of the beloved research themes of Theoretical Computer Science. Valiant believes that random synaptic connections are an essential ingredient of the brain's power and versatility. Randomness is also ubiquitous everywhere in neural activity, resulting to a wide range of trial-to-trial variation in almost any brain experiment. It is essential to incorporate randomness in computational models of brain systems, and to understand its origins and function in the brain. We refer to sections 3 and 4 of [32] for references to related experimental data.

- The foundational understanding of the apparent power of deep learning is an important current challenge for Theoretical Computer Science. How does this quest relate to the brain? We refer to [72] for a discussion of related literature. Deep learning of some sort does happen in the brain (consider the visual cortex and the hierarchical processing through its areas, from V1 to V2 and V4 all the way to MT and beyond). But there are differences, and perhaps the most fundamental among them is the existence of *lateral and backward connections* between brain areas. What is their function, and how do they enhance learning?

- A complementary question is, *what replaces backpropagation in brain circuits?* The famous backpropagation algorithm that is used to efficiently optimize deep neural networks is incompatible with our understanding of brain connectivity, as it requires reciprocal connections with weight updates that are maintained to levels identical to those of the forward connections. An intriguing recent finding in this regard is the surprising learning capability of (rather shallow) neural networks in which, instead of backpropagation, feedback is carried out with *fixed random weights* [73].

6 Summary

We sketched the history, current status, and prospects of research interaction between computer scientists and neuroscientists in the quest of unraveling the organization of brain computation. We then focused on the specific question, how are memories and a web of associations between memories implemented in networks of neurons in the brain. This question appears to be especially well suited for contributions by theoretical computer scientists, since (a) a theory that is consistent with recent recordings from the human brain is missing; and (b) scaling and asymptotic analysis of model data structures and algorithms seem essential for understanding how the human brain can create and maintain an association web of tens of thousands of concepts. We concluded with a sprinkling of open questions, each accompanied by references to some of the most recent research articles and review papers in neuroscience. Since for most domains one cannot extract from the literature a single model or set of assumptions, familiarity with a diversity of models and experimental results is a prerequisite for any lasting contribution to our understanding of brain computation. Ultimately, an informed and fruitful dialogue and collaboration between computer scientists and neuroscientists may be the brightest hope we have for finally unraveling the mysteries of brain computation.

Acknowledgments

Written under partial support by the Human Brain Project of the European Union #604102 and #720270, and NSF grants CCF-1408635, EAGER-1415498 and CCF-1563838.

References

1. Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J.: Principles of Neural Science. Volume 5th Edition. McGraw-Hill (2013)
2. Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W.C., LaMantia, A.S., White, L.E.: Neuroscience. Volume 5th Edition. Sinauer Associates, Inc (2011)
3. Dayan, P., Abbott, L.F.: Theoretical neuroscience. Volume 10. Cambridge, MA: MIT Press (2001)

4. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co., Inc., New York, NY, USA (1982)
5. Valiant, L.G.: A theory of the learnable. Commun. ACM **27**(11) (1984) 1134–1142
6. Hawkins, J., Blakeslee, S.: On Intelligence. Times Books (2004)
7. Valiant, L.G.: Circuits of the mind. Oxford University Press (1994)
8. Valiant, L.G.: A neuroidal architecture for cognitive computation. J. ACM **47**(5) (2000) 854–882
9. Valiant, L.G.: Memorization and association on a realistic neural model. Neural Computation **17**(3) (2005) 527–555
10. Jerrum, M., Sinclair, A.: Polynomial-time approximation algorithms for the Ising model. SIAM Journal on computing **22**(5) (1993) 1087–1116
11. Yao, A.C.C.: Quantum circuit complexity. In: Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on, IEEE (1993) 352–361
12. Bernstein, E., Vazirani, U.: Quantum complexity theory. SIAM Journal on Computing **26**(5) (1997) 1411–1473
13. Papadimitriou, C.: Algorithms, games, and the internet. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing, ACM (2001) 749–753
14. Nisan, N., Ronen, A.: Algorithmic mechanism design. In: Proceedings of the thirty-first annual ACM symposium on Theory of computing, ACM (1999) 129–140
15. Turing, A.M.: On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London mathematical society **2**(1) (1937) 230–265
16. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236) (1950) 433–460
17. Neumann, J.v.: The Computer and the Brain. Yale University Press (1958)
18. Neumann, J.v., Burks, A.W.: Theory of self-reproducing automata. University of Illinois Press (1966)
19. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics **5**(4) (1943) 115–133
20. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6) (1958) 386
21. Marr, D.C., Poggio, T.: From understanding computation to understanding neural circuits. Technical Report AI-M-357, Massachusetts Institute of Technology (Cambridge, MA US) (1976)
22. Marcus, G.F.: The algebraic mind: Integrating connectionism and cognitive science. MIT press (2003)
23. Marcus, G.F., Marblestone, A., Dean, T.: The atoms of neural computation. Science **346**(6209) (2014) 551–552
24. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks **2**(5) (1989) 359–366
25. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information theory **39**(3) (1993) 930–945
26. Eliasmith, C., Anderson, C.H.: Neural engineering: Computation, representation, and dynamics in neurobiological systems. MIT press (2004)
27. Seung, H.S.: Neuroscience: towards functional connectomics. Nature **471**(7337) (2011) 170–172
28. Lichtman, J.W., Livet, J., Sanes, J.R.: A technicolour approach to the connectome. Nature Reviews Neuroscience **9**(6) (2008) 417–422

29. Holtmaat, A., Svoboda, K.: Experience-dependent structural synaptic plasticity in the mammalian brain. Nature Reviews Neuroscience **10**(9) (2009) 647–658
30. Minerbi, A., Kahana, R., Goldfeld, L., Kaufman, M., Marom, S., Ziv, N.E.: Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity. PLoS Biology **7**(6) (2009) e1000136
31. Kasai, H., Fukuda, M., Watanabe, S., Hayashi-Takagi, A., Noguchi, J.: Structural dynamics of dendritic spines in memory and cognition. Trends in neurosciences **33**(3) (2010) 121–129
32. Maass, W.: Searching for principles of brain computation. Current Opinion in Behavioral Sciences (Special Issue on Computational Modelling) **11** (2016) 81–92
33. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience **2**(11) (1999) 1019–1025
34. Berwick, R.C., Chomsky, N.: Why only us: Language and evolution. MIT Press (2016)
35. Papadimitriou, C.H., Vempala, S.S.: Cortical learning via prediction. In: Proc. of COLT. (2015)
36. Papadimitriou, C.H., Petti, S., Vempala, S.: Cortical computation via iterative constructions. In: Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016. (2016) 1357–1375
37. Froemke, R.C., Debanne, D., Bi, G.Q.: Temporal modulation of spike-timing-dependent plasticity. Frontiers in Synaptic Neuroscience (17 June 2010) https://doi.org/10.3389/fnsyn.2010.00019
38. Buzsaki, G.: Neural syntax: cell assemblies, synapsembles, and readers. Neuron **68**(3) (2010)
39. Komorowski, R.W., Manns, J.R., Eichenbaum, H.: Robust conjunctive itemplace coding by hippocampal neurons parallels learning what happens where. The Journal of Neuroscience **29**(31) (2009) 9918–9929
40. Ison, M.J., Quiroga, R.Q., Fried, I.: Rapid encoding of new memories by individual neurons in the human brain. Neuron **87**(1) (2015) 220–230
41. De Falco, E., Ison, M.J., Fried, I., Quiroga, R.Q.: Long-term coding of personal and universal associations underlying the memory web in the human brain. Nature Communications **7** (2016) 13408
42. Hebb, D.O.: The organization of behavior: A neuropsychological theory. Wiley, New York (1949)
43. Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I.: Invariant visual representation by single neurons in the human brain. Nature **435**(7045) (2005) 1102–1107
44. Pokorny, C., Ison, M., Legenstein, R., Papadimitriou, C., Vempala, S., Maass, W.: Memory-specific assemblies and associations between assemblies emerge in a cortical microcircuit model through STDP (in preparation)
45. Nessler, B., Pfeiffer, M., Buesing, L., Maass, W.: Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. PLOS Computational Biology **9**(4) (2013) e1003037
46. Papadimitriou, C., Petti, S., Maass, W., Vempala, S.: Dynamic networks and the emergence of assemblies (in preparation)
47. Franks, K.M., Russo, M.J., Sosulski, D.L., Mulligan, A.A., Siegelbaum, S.A., Axel, R.: Recurrent circuitry dynamically shapes the activation of piriform cortex. Neuron **72**(1) (2011) 49–56
48. Wang, X.J., Kennedy, H.: Brain structure and dynamics across scales: in search of rules. Current Opinion in Neurobiology **37** (2016) 92–98

49. Erdős, P., Renyi, A.: On the evolution of random graphs. Publ. Math. Inst. Hungary. Acad. Sci. **5** (1960) 17–61
50. Guzman, S., J Schlö gl, A., Frotscher, M., Jonas, P.: Synaptic mechanisms of pattern completion in the hippocampal ca3 network. Science **353**(6304) (September 2016) 1117–1123
51. Frankland, S.M., Greene, J.D.: An architecture for encoding sentence meaning in left mid-superior temporal cortex. Proceedings of the National Academy of Sciences **112**(37) (2015) 11732–11737
52. Legenstein, R., Papadimitriou, C.H., Vempala, S., Maass, W.: Assembly pointers for variable binding in networks of spiking neurons. arXiv preprint arXiv:1611.03698 (2016)
53. P. Erdős, A.G., Posa, L.: The representation of graphs by set intersections. Canadian J. Math. **18** (1966) 106–112
54. Pitowsky, I.: Correlation polytopes: their geometry and complexity. Mathematical Programming **50**(1) (1991) 395–414
55. Alon, N., Naor, A.: Approximating the cut-norm via Grothendieck's inequality. SIAM J. Comput. **35**(4) (2006) 787–803
56. Anari, N., Maass, W., Papadimitriou, C., Saberi, A.: Association graphs (in preparation)
57. Song, S., Sjöström, P.J., Reigl, M., Nelson, S., Chklovskii, D.B.: Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biology **3**(3) (2005) e68
58. Buzsaki, G., Mizuseki, K.: The log-dynamic brain: how skewed distributions affect network operations. Nat Rev Neurosci **15**(4) (2014) 264–278
59. Grosmark, A.D., Buzsaki, G.: Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. Science **351**(6280) (2016) 1440–1443
60. Beggs, J.M., Plenz, D.: Neuronal avalanches in neocortical circuits. The Journal of Neuroscience **23**(35) (2003) 11167–11177
61. Bellay, T., Klaus, A., Seshadriand, S., Plenz, D.: Irregular spiking of pyramidal neurons organizes as scale-invariant neuronal avalanches in the awake state. eLife **4** (2015) e07224
62. Priesemann, V., Wibral, M., Valderrama, M., Pröpper, R., Le Van Quyen, M., Geisel, T., Triesch, J., Nikolic, D., Munk, M.H.: Spike avalanches in vivo suggest a driven, slightly subcritical brain state. Front. Syst. Neurosci **8** (2014)
63. Legenstein, R., Maass, W.: Edge of chaos and prediction of computational performance for neural circuit models. Neural Networks **20**(3) (2007) 323–334
64. Legenstein, R., Maass, W.: What makes a dynamical system computationally powerful. New directions in statistical signal processing: From systems to brain (2007) 127–154
65. Fiser, J., Chiu, C., Weliky, M.: Small modulation of ongoing cortical dynamics by sensory input during natural vision. Nature **431** (2004) 573–583
66. Luczak, A., Barthó, P., Harris, K.D.: Spontaneous events outline the realm of possible sensory responses in neocortical populations. Neuron **62**(3) (2009) 413–425
67. Bathellier, B., Ushakova, L., Rumpel, S.: Discrete neocortical dynamics predict behavioral categorization of sounds. Neuron **76**(2) (2012) 435–449
68. Miller, J.e.K., Ayzenshtat, I., Carrillo-Reid, L., Yuste, R.: Visual stimuli recruit intrinsically generated cortical ensembles. Proceedings of the National Academy of Sciences **111**(38) (2014) E4053–E4061
69. Olshausen, B.A., Field, D.J.: How close are we to understanding V1? Neural computation **17**(8) (2005) 1665–1699

70. DiCarlo, J.J., Cox, D.D.: Untangling invariant object recognition. Trends in cognitive sciences **11**(8) (2007) 333–341
71. Cox, D.D.: Do we understand high-level vision? Current opinion in neurobiology **25** (2014) 187–193
72. Marblestone, A.H., Wayne, G., Kording, K.P.: Toward an integration of deep learning and neuroscience. Frontiers in Computational Neuroscience **10** (2016)
73. Lillicrap, T.P., Cownden, D., Tweed, D.B., Akerman, C.J.: Random feedback weights support learning in deep neural networks. arXiv preprint arXiv:1411.0247 (2014)