OXFORD

THIRD EDITION

# FUNDAMENTALS OF COMPUTATIONAL NEUROSCIENCE

THOMAS P. TRAPPENBERG

# Fundamentals of Computational Neuroscience

# Fundamentals of Computational Neuroscience Third Edition

Thomas P. Trappenberg
Dalhousie University

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries



First Edition published in 2002
Second Edition published in 2010
Third Edition published in 2023

Impression: 1



Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 100106, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2022943170

ISBN 978–0–19–286936–4

DOI: 10.1093/oso/9780192869364.001.0001

Printed in the UK by
Ashford Colour Press Ltd, Gosport, Hampshire

# Preface

Computational neuroscience is still a young and dynamically developing discipline, and some choice of topics and presentation style had to be made. This text introduces some fundamental concepts, with an emphasis on basic neuronal models and network properties. In contrast to the common research literature, this book is trying to paint the larger picture and tries to emphasize some of the concepts and assumptions for simplifications used in the scientific technique of modelling.

Computational neuroscience and Artificial Intelligence (AI) are close cousins. The term AI is said to be invented at the Dartmouth workshop in 1956 with many famous participants including psychiatrist Ross Ashby, the neurophysiologist Warren McCulloch who created one of the first mathematical neuron models, and Arthur Samuel, one of the pioneers in reinforcement learning. Computational models of neural systems such as models of neurons are much older, but connecting learning and cognitive systems created excitement over the possibility to better understand mind. The invention of learning machines has revolutionized many applications as recently seen in the dramatic progress of machine vision and natural language processing through deep learning.

While there has been much recent progress in machine learning, researchers in this area often wonder how the brain works. It sometimes seems that scientific progress oscillates between computational neuroscience and machine learning. For example, the progress of neural networks and statistical learning theory in the later 1980s and early 1990s was followed by enormous activities in computational neuroscience in the 1990s and early 2000s. For the last decade, deep learning has occupied an explosive growth in machine learning and data science, and now the time seems ripe for more renewed interest in looking more closely at the brain for inspirations to go deeper. This is fuelled by the increasing realization of limitations of deep learning, in particular with the challenge of learning semantic knowledge with limited data and the ability to transfer knowledge to situations that are not directly represented in the learning set.

In this new edition of my book, I tried to incorporate many of the recent lessons from deep learning. While there are excellent books on deep learning, our emphasis here is their connection to brain processing. An important aspect is thereby the concepts of representational learning and computation with uncertainties. Also, I now included gated recurrent neural networks that are becoming an important fundamental mechanisms when thinking about brain processing. While we will not be able to dive into all the recent progress, I hope that the text will guide further specific studies and research. Furthermore, it was important for me to streamline the existing text. I hope that I improved the readability of some of the text and even removed parts that seem less relevant to study the most basic fundamentals.

The themes included in this book are chosen to provide some path through the different levels of description of the brain. Chapter 1 provides a high-level overview and some fundamental questions about brain theories, a brief discussion about the

role of modelling, and some basic neuroscience facts that are useful to keep in mind for later use. We also review the essential scientific programming in Python and the basic mathematical and statistical concept used in the book. Chapters 2–4 focus on basic mechanisms and modelling of single neurons or population averages. This starts from a fairly detailed discussion of changes in the membrane potentials through ion channels, spike generations, and synaptic plasticity, with increasingly abstractions in the following chapters. Chapters 5–7 describe the information-processing capabilities of basic networks, including feedforward and competitive recurrent networks. The last part of the book describes some examples of combining such elementary networks as well as some examples of more system-level models of the brain.

Most models in the book are quite general and are aimed at illustrating basic mechanisms of information processing in the brain. In the research literature, the basic elements reviewed in this book are often combined in specific ways to model specific brain areas. Our hope is that the study of the basic models in this book will enable the reader to follow some of the recent research literature in computational neuroscience.

While we tried to emphasize some important concepts, we did not want to give the impression that the chosen path is the only direction in computational neuroscience. Therefore, we sometimes mention concepts without extensive discussion. These comments are intended to increase the reader's awareness of some issues and to provide some keywords to facilitate further literature searches. Also, while some examples of specific brain areas are mentioned in this book, a comprehensive review of models in computational neuroscience is beyond the scope of this text. We do not claim that this book covers all aspects of computational neuroscience nor do we claim it to be the only approach to this area, but we hope that it will contribute to the discussion.

## Mathematical formulas

This book includes mathematical formulas and concepts. We use mathematical language and concepts strictly as practical tools and to communicate ideas in contrast to using such formalism for mathematical proofs. We thereby tried to balance detailed mathematical notations with readability and communicating the basic concepts. From readers with less extensive training in such formal systems I ask for patience. We did not try to avoid mathematical formulations since such notations allow a brevity in communication that would be lengthy with plain written language. The chosen level of mathematical descriptions are mainly intended to be translated directly into programs and other quantitative evaluations.

There is no reason to be afraid of formulas, and it is important to see beyond the symbols and to understand their meaning. Many mathematics notations are invented to simplify descriptions. This includes the use of vectors and matrices, which will drastically shorten the specification of network models. We provide review chapters in the first part of the book to review such notations. We recommend some tutorials on such materials to allow students to move beyond these technicalities in the main text.

Most models in this book describe the change of a quantity with time, such as the change of a membrane potential after synaptic input or synaptic strength values over time during learning. Equations that describe such changes are called differential equations. A comprehensive knowledge of the theory of differential equations is not required for understanding this book. However, discussing the consequences of specific

differential equations and simulating them with computer programs is at the heart of this book. I hope our treatment will encourage a new look into a topic that sometimes seems overwhelming when treated in specialized classes. We will specifically become familiar with a simple yet telling example of a differential equation, that of a leaky integrator. A basic knowledge of the numerical approaches to solving differential equations is essential for this book and many other dynamic modelling approaches. Thus, we also include a review of differential equations and their numerical integration.

Another mathematical theory, that of random numbers, is also reviewed in the third chapter. The language of probability theory is very useful in computational neuroscience and should be taught in such a course. In neuroscience (as in other disciplines), we often get different values each time we perform a measurement, and random numbers describe such situations. We often think of these circumstances as noise, but it is also useful to think about random variables and statistics in terms of describing uncertainties. Indeed, it can be argued that learning and reasoning in uncertain circumstances is a fundamental requirement of the brain. We will argue that mental functions can be viewed as probabilistic reasoning.

### Programming examples

While this book includes a few examples of powerful analytical techniques to give the reader a flavour of some of the more elaborate theoretical studies, not every neuroscientist has to perform such calculations themselves. However, studying some of the general ideas behind these techniques is essential to be able to get support from those who specialize in such techniques. In particular, it is instructive when studying this book to perform some numerical experiments yourself. We therefore included an introduction to a modern programming environment that is very much suited for many of the models in neuroscience. Writing programs and creating advanced graphics can be learned easily within a short time, even without extensive prior programming knowledge.

The programs in this book are now provided in Python to improve accessibility and due to Python's increasing importance in machine learning and data science. While it was challenging to balance a scientist's approach of making minimalist and clean examples with common programming approaches, I hope that I found some balance. Comments in programs are often a good idea in complex software packages. However, the situation is different here. The programs are purposefully kept short and the expectation is that each line should be read and understood entirely. For example, we think that comments like `# assigning value b to variable a` to describe the code `a = b` should not be necessary. Instead, the reader should strive to be able to read the code directly. Comments in the program were therefore deliberately avoided except to explain some variable names to keep the variable names short, and some comments to structure the code. Many people have different styles of coding, and the style here tried deliberately to strive for compactness and simplicity. While it might be a new language for some, trying to understand each line in a program will help to master programming in a short time.

## References

This book does not provide a historical account of the development of ideas in computational neuroscience. Indeed, extensive references have been avoided where possible to concentrate on describing fundamental ideas. This is hence more consistent with course textbooks. References to the original research literature are only provided when following corresponding examples closely. The text is very much aimed at providing a starting place for further studies, and search engines will now easily provide further directions.

# Acknowledgements

Many friends and colleagues have contributed over the years to this book I am specifically thankful to Farzaneh Sheikhnezhad Fard, Alan Fine, Steve Grossberg, Alexander Hanuschkin, Geoffrey Hinton, Abraham Nunez, Kai Trappenberg, Nami Trappenberg, Jason Satel, Michael Schmitt, Dominic Standage, Fumio Yamazaki, and Si Wu.

# Contents

# I

# Background

# 1 Introduction and outlook

This introductory chapter is outlining the big picture. We define the scope of the computational neuroscience discussed in this book and outline some basic facts of brain organization and principles that we encounter in later chapters. This chapter includes a discussion on the role of scientific modelling in general and in neuroscience specifically. In addition, we outline a high-level theory of the brain as a predictive model of the world, and we outline some principles that will guide much of the discussions in this book.

## 1.1 What is computational neuroscience?

Computational or theoretical neuroscience uses distinct techniques and asks specific questions aimed at advancing our understanding of the nervous system. A brief definition might be:

> Computational neuroscience is the theoretical study of the brain used to uncover the principles and mechanisms that guide the development, organization, information processing and mental abilities of the nervous system.

Most papers in computational neuroscience journals follow one of two quite different principle directions. One direction is the use of computational methods to analyses data such as sorting spikes or to quantitatively test hypothesis. In this context, methods from AI (Artificial Intelligence) such as machine learning techniques are now often included as tools for data analytics. We will encounter such techniques, specifically that of neural networks and deep learning. However, our focus here is less on describing data analytics methods but rather to build models of brain functions to understand its processing capabilities. The type of computational neuroscience described in this book is hence mostly synonymous with theoretical neuroscience in that we develop and test hypotheses of the functional mechanisms of the brain.

We often use computer simulations in our studies, though 'computational' highlights more broadly our interested in the computational and information-processing aspects of brain functions. A main focus in this book is hence the development and evaluation of brain models, or models of specific functions of the brain. These are important to summarize knowledge, to quantify theories, and to test computational hypotheses. We focus thereby on fundamental mechanisms and mechanistic foundations which seem to be underlying brain processes. We also try to highlight some emerging principles of brain-style information processing. This book does claim a comprehensive theory of the mind. However, we hope that learning these fundamentals will be an important part of further developments.

*Fundamentals of Computational Neuroscience.* Third edition. Thomas P. Trappenberg,
Oxford University Press. © Oxford University Press 2023. DOI: 10.1093/oso/9780192869364.003.0001

### 1.1.1 Embedding within neuroscience

Computational neuroscience is a specialization within neuroscience. Neuroscience itself is a scientific area with many different aspects. Its aim is to understand the nervous system, in particular the *central nervous system* and the spine that we call the brain. The brain is studied in diverse disciplines such as physiology, psychology, medicine, computer science, and mathematics. Neuroscience emerged from the realization that interdisciplinary studies are vital to further our understanding of the brain. While considerable progress has been made in our understanding of brain functions, there are many open questions that we want to answer. What is the function of the brain and how does it achieve its task? What are the biological mechanisms involved? How is it organized? What are the information-processing principles used to solve complex tasks such as perception? How did the brain evolve? How does it change during the lifetime of organisms? What is the effect of damage to particular areas and the possibilities of rehabilitation? What are the origins of degenerative diseases and possible treatments? These are questions asked by neuroscientists in many different subfields, using a multitude of different research techniques.

Many techniques are employed in neuroscience to study the brain. Those techniques include genetic manipulations, recording of cell activities in cultured cells, brain slices, optical imaging; non-invasive functional imaging, psychophysical measurements; and computational simulations, to name but a few. Each of these techniques is complicated and laborious enough to justify a specialization of neuroscientists in particular techniques. Therefore, we speak of neurophysiologists, cognitive scientists, and anatomists. It is, however, vital for any neuroscientist to develop a basic understanding of all major techniques, so he or she can comprehend and utilize the contributions made within these specializations. Computational neuroscience is a relative new area of neuroscience with increasing importance. It fills an important role in quantifying theories based on the increasing amount of experimental discoveries. A basic comprehension of the contribution that computational neuroscience can make is becoming increasingly important for all neuroscientists.

Within computational neuroscience we often use computers, although other areas of neuroscience use computers. Our main reason for using computers is that the complexity of models in this area is often beyond analytical tractability. For such models we have to employ carefully designed numerical experiments to be able to compare the models to experimental data. However, we do not need to restrict our studies to this tool. Some models are analytically tractable or might be deliberately simplified to be analytically tractable. Such models often provide a deep and more controlled insight into the features of certain mechanisms and the reasons behind numerical findings.

Although computational neuroscience is theoretical by its very nature, it is important to bear in mind that models must be gauged on experimental data; they are otherwise useless for understanding the brain. Only experimental measurements of the real brain can verify ‘what’ the brain actually does. In contrast to the experimental domain, computational neuroscience tries to speculate ‘how’ the brain operates. Such speculations are developed into hypotheses, realized into models, evaluated analytically or numerically, and tested against experimental data. Also, models can often be used to make further predictions about the underlying phenomena.

## 1.2 Organization in the brain

Mental functions such as perception and learning motor skills are not accomplished by single neurons alone. These functions are an emerging property of specialized networks with many neurons that form the nervous system. The number of neurons in the central nervous system is estimated to be on the order of $10^{12}$, and it is demanding to explore such vast systems of neurons. Therefore, rather than trying to rebuild the brain in all its detail on a computer, we aim to understand the principal organization of brains and how networks of neuron-like elements can support and enable particular mental processes. Integration of neurons into networks with specific architectures seem to be essential for such skills. We will explore the computational abilities of several principal architectures of neural networks in this book.

A thorough knowledge of the anatomy of the brain areas we want to model is essential for any research that attempts to understand brain functions. However, although recent research has revealed many important facts about neural organization, it is still often difficult to specify all the components of a model on the basis of anatomical and physiological data alone, and plausible assumptions have to be made to bridge gaps in the knowledge. Even if we can draw on known details, it is often useful to make simplifying assumptions that enable computational tractability or the tracing of principal organizations sufficient for certain functionalities. It is beyond the scope of this book to describe all the details of neuronal organization, and more specialized books and research articles have to be consulted for specific brain areas. The aim of the following section is to outline a large variety of facts mainly to raise awareness of the many factors of structures and organizations in the brain. In computational neuroscience we have a constant struggle between incorporating as many details as possible while keeping models simple to illuminate the principles behind brain functions. We hope that this section will encourage more specific studies of brain anatomy.

### 1.2.1 Levels of organization in the brain

Models in computational neuroscience can target many different levels of descriptions. This in itself is a consequence of the fact that the nervous system has many levels of organization on spatial scales ranging from the molecular level of a few Angstrom ($1\text{Å} = 10^{-10}$m), to the whole nervous system on the scale of over a metre. Biological mechanisms on all these levels are important for the brain to function.

Different levels of organization in the nervous system are illustrated in Fig. 1.1. An important structure in the nervous system is the neuron, which is a cell that is specialized for signal processing. Depending on external conditions, neurons are able to generate electric potentials that are used to transmit information to other cells to which they are connected. Mechanisms on a subcellular level are important for such information processing capabilities. Neurons use cascades of biochemical reactions that have to be understood on a molecular level. These include, for example, the transcription of genetic information which influences information-processing in the nervous system. Many structures within neurons can be identified with specific functions. For example, mitochondria are structures important for the energy supply in the cell, and synapses mediate information transmission between cells. The complexity of a single neuron, and even isolated subcellular mechanisms, makes computational

studies essential for the development and verification of hypotheses. It is possible today to simulate morphologically reconstructed neurons in great detail, and there has been much progress in understanding important mechanisms on this level.

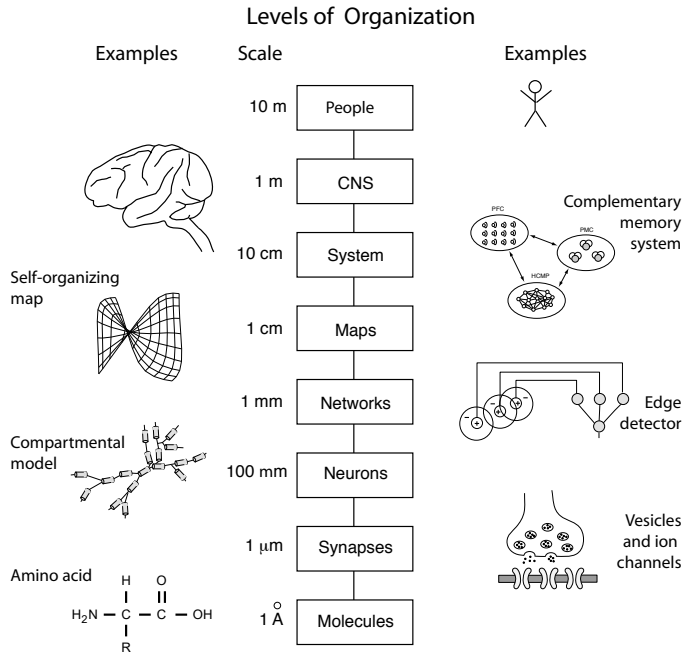


**Fig. 1.1** Some levels of organization in the central nervous system on different scales [adapted from Churchland and Sejnowski, *The computational brain*, MIT Press (1992)].

However, single neurons certainly do not tell the whole story. Neurons contact each other and thereby compose networks. A small number of interconnected neurons can exhibit complex behaviour and enable information-processing capabilities not present in a single neuron. Understanding networks of interacting neurons is a major domain in computational neuroscience. Networks have additional information-processing capabilities beyond that of single neurons, such as representing information in a distributed way. An example of a basic network is the edge detector formed from a centre-surround neuron as proposed by Hubble and Wiesel. The illustrated levels above the level labelled 'Networks' in Fig. 1.1 are also composed of networks, yet with increasing size and complexity. An example on the level termed 'Maps' in Fig. 1.1 is a self-oganizing topographic map, which is part of an important discussion in this book.

The organization does not stop at the map level. Networks with a specific architecture and specialized information-processing capabilities are composed into larger structures that are able to perform even more complex information-processing tasks. System-level models are important in understanding higher-order brain functions. The central nervous system depends strongly on the dynamic interaction of many specialized subsystems, and the interaction of the brain with the environment. Indeed, we will see later that active environmental interactions are essential for brain development and

function.

Although an individual researcher typically specializes in mechanisms of a certain scale, it is important for all neuroscientists to develop a basic understanding and appreciation of the functionalities of different scales in the brain. Computational neuroscience can help the investigations at all levels of description, and it is not surprising that computational neuroscientists investigate different types of models at different levels of description. Computational methods have long contributed to cellular neuroscience, and computational cognitive neuroscience is now a rapidly emerging field. The contributions of computational neuroscience are, in particular, important to understand non-linear interactions of subprocesses. Furthermore, it is important to comprehend the interactions between different levels of description, and computational methods have proven very useful in bridging the gap between physiological measurements and behavioural correlates.

### 1.2.2 Large-scale brain anatomy

The nervous system is distributed throughout the whole body. Some of the peripheral nervous system include sensors such as touch sensors or sensors for auditory signals. Some of those sensors like the eyes are in themselves already highly sophisticated neural systems, and the brainstem already processes sensory signals to produce fast responses such as reflexes. Of course, it is clear that more complex information processing can be achieved with the added complexity of the central nervous system that we usually call the brain (Fig. 1.2). The brain itself has a lot of structure in itself, such as subcortical midbrain areas that include structures that we will mention like the basal ganglia or the thalamus. Even within the cortex we can easily distinguish areas of the paleocortex and archicortex, which include structures like the amygdala, the secondary olfactory cortex, and the hippocampal formation. These cortical structures have mostly three or four layers of cortex compared to the six layers of the neocortex that coverse the outside of the mammalian brain. As the name indicates, the neocortex seems phylogenetically newer than the archicortex and the paleocortex, meaning that the neocorrtex developed later during evolution.

While the neocortex looks more homogeneous, regions of the neocortex are commonly divided into four lobes as illustrated in Fig. 1.2B, the occipital lobe at the rear of the head, the adjacent parietal lobe, the frontal lobe, and the temporal lobes at the flanks of the brain. Further subdivisions can be made, based on various criteria. For example, at the beginning of the twentieth century the German anatomist Korbinian Brodmann identified 52 cortical areas based on their cytoarchitecture, the distinctive occurrence of cell types and arrangements, which can be visualized with various staining techniques. Brodmann labelled the areas he found with numbers, as shown in Fig. 1.2B. Some of these subdivisions have since been refined, and letters following the number are commonly used to further specify some part of an area defined by Brodmann. Brodmann's cortical map is, however, not the only reference to cortical areas used in neuroscience. Other subdivisions and labels of cortical areas are based, for example, on functional correlates of brain areas. These include behavioural correlates of cortical areas as revealed by brain lesions or functional brain imaging, as well as neuronal response characteristics identified by electrophysiological recordings.

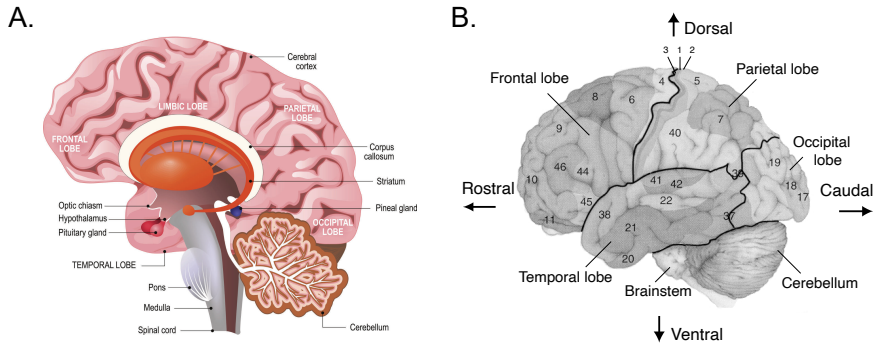


**Fig. 1.2** Outline of the lateral view of the human brain including the neocortex, cerebellum, and brainstem. The neocortex is divided into four lobes. The numbers correspond to Brodmann's classification of cortical areas. Directions are commonly stated as indicated in 1.2B.

It is, of course, of major interest to establish functional correlates of different cortical areas, a challenge that drives many physiological studies. We might speculate that the diverse functional specialization within the neocortex found with electrophysiological measurements is reflected in major structural differences among the different cortical areas to support specialized mental functions. It is therefore remarkable to realize that this is not the case. Instead, it is found that different areas of the neocortex have a remarkably common neuronal organization. All neocortical areas have anatomically distinguishable layers as discussed below. The differences in the cytoarchitecture, which have been used by Brodmann to map the cortex, are often only minor compared to the principal architecture within the neocortex, and these variations cannot account solely for the different functionalities associated with the different cortical areas.

The neocortex is different in this respect to older parts of the brain, such as the brainstem, where structural differences are much more pronounced. This is reflected in a variety of more easily distinguishable nuclei. We can often attribute specific low-level functions to each nucleus in the brainstem. In contrast to this, it seems that the cortex is an information-processing structure with more universal processing abilities that we speculate enable more flexible mental abilities. It is therefore most interesting to investigate the information-processing capabilities of neuronal networks with a neocortical architecture.

### 1.2.3 Hierarchical organization of cortex

A common feature of neocortex is that there are primary sensory areas in which basic features of sensory signals are represented, while other areas seem to support more complex representations or mental tasks. Let us highlight this common view of neocortex with the example of vision. The primary visual area that receives mayor input from the eyes lies in the caudal end of the occipital lobe and is called V1. Information is then transmitted to other visual areas in the occipital lobe before splitting into two major processing streams, the dorsal stream along a parietal to frontal pathway, and the ventral stream along the temporal lobe. It has been argued that the dorsal stream is specifically adapted to spatial processing, whereas the ventral stream is well equipped for object recognition. We will investigate a model of such what-and-where processing

later in the book. The main point here is that brain scientists try to identify functional specific areas and connections between these areas.

In order to understand how different brain areas work together it is important to establish the anatomical and functional connectivity between brain areas in more detail. Anatomical connections are not easy to establish as it is extremely difficult to follow the path of stained axons through the brain in brain slices (including the branches that can often have different pathways). This is a daunting task, though it has been done in isolated cases. There are other methods of establishing connectivities in the brain. These include the use of chemical substances that are transported by the neurons to target areas or from target areas to the origin. Functional connectivity patterns, in which we are particularly interested when studying how brain areas work together, can also be established with simultaneous stimulations and recordings in different brain areas. Such experiments show correlations in the firing patterns of neurons in different brain areas if they are functionally connected. Also, some large-scale functional brain organizations can be revealed by brain-imaging techniques such as functional magnetic resonance imaging (fMRI), which can highlight the areas involved in certain mental tasks. Such studies established clearly that different brain areas do not work in isolation. On the contrary, many specialized brain areas have to work together to solve complex mental tasks.

Some scientists, such as Van Essen and colleagues, have long tried to compile experimental data into connectivity maps similar to the one shown in Fig. 1.3. The specific example was produced by Claus C. Hilgetag, Mark A. O'Neill, and Malcolm P. Young. The researchers used a neuroinformatics approach. Neuroinformatics is specifically concerned with the collection and representation of experimental data in large databases to which modern data mining methods can be applied. Hilgetag and colleagues considered an algorithm that would evaluate many possible configurations, and they found a large set of possible connectivity patterns in the visual cortex satisfying most of the experimental constraints. Each box in Fig. 1.3 represents a cortical area that has been distinguished from other areas on different grounds, typically anatomical and functional. The solid pathways between these boxes represent known anatomical or functional connections. The order from bottom to the top indicates roughly the hierarchical order in which these brain areas are contacted in the information-processing stream, from primary visual areas establishing some basic representations in the brain to higher cortical areas that are involved in object recognition and the planning and execution of motor actions. The authors also took the two basic visual processing pathways in their representation into account, plotting brain areas of the dorsal stream on the left side and the ventral stream on the right side. Note that there are also interactions within these pathways.

Interestingly, most solutions of the numerical optimization problem have displayed some consistent hierarchical structures. All solutions found violated some of the experimental constraints (dashed line in Fig. 1.3), which is probably based on the inaccuracy of some of the experimental results. Also, the connections indicated are not unidirectional. It is well established that a brain area that sends an axon to another brain area also receives back-projections from the structures it sends to. Such back-projections are often in the same order of magnitude as the forward projections. Interesting examples, not included in Fig. 1.3, are so-called corticothalamic loops. The subcortical

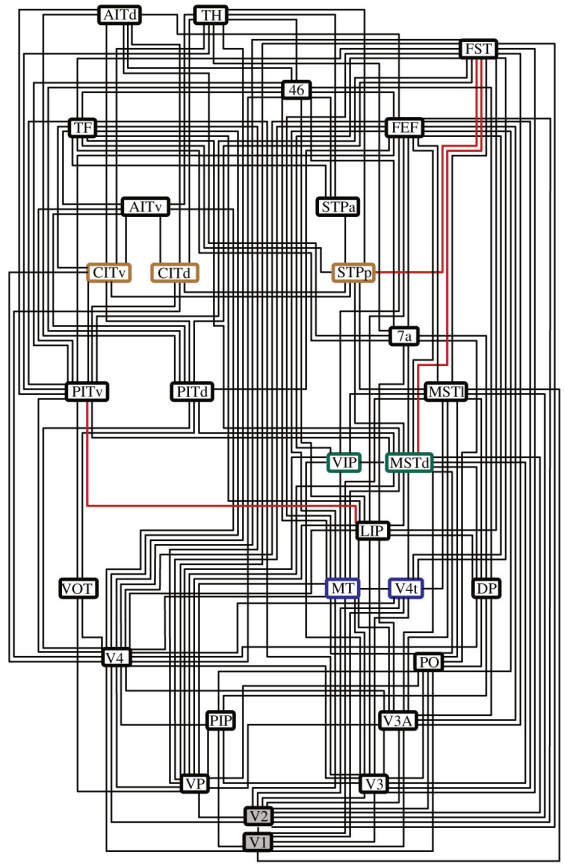


**Fig. 1.3** Example of a map of connectivity between cortical areas involved in visual processing [reprinted with permission from C. Hilgetag, M. O'Neill, and M. Young, *Philosophical Transactions of the Royal Society of London B* 355: 71–89 (2000)].

structure called the thalamus was initially viewed as the major relay station through which sensory information projects to the cortex. However, it is becoming increasingly clear that the notion of a pure relay station is too simple as there are generally many more back-projections from the cortex to the thalamus compared to the forward projections between the thalamus and the cortex. Some estimates even indicate a number of back-projections that exceed the forward projections tenfold. The specific functional consequences of back-projections between the thalamus and the cortex as well as within the cortex itself are still not well understood. However, such structural features are consistent with reports of the influence of higher cortical areas on cell activities in primary sensory areas, for example, attentional effects in V1.

In the last decade there have been increasingly elaborate attempts to produce a more detailed wiring diagram of the brain, as so called connectome. One of the first full mapping of all neurons and connection has been done for a roundworm called Caenorhabditis elegans. This animal had therefore become one of the best-studied

model systems in neuroscience. The European Blue Brain Project attempts to map the mouse brain, and there are also attempts to map the human brain. A visualization of white matter connections from 20 subjects is shown in Fig. 1.4

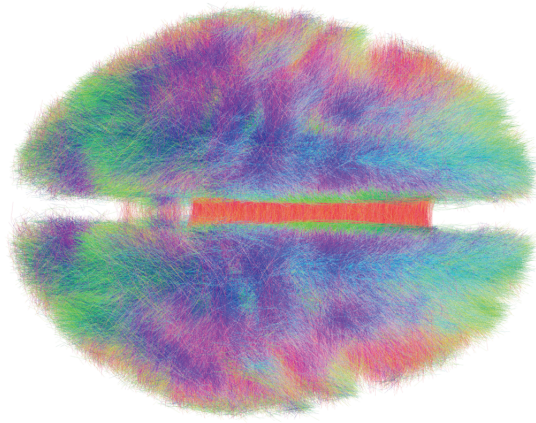

**Fig. 1.4** Example of a group-level connectome of human white matter. [A. Horn, D. Ostwald, M. Reisert, F. Blankenburg (November 2014). '*The structural-functional connectome and the default mode network of the human brain*'. *NeuroImage* 102(1): 142-{51].

### 1.2.4 Rapid data transmission in the brain

From the system level view of the brain it seems that there are many stages of processing in the brain so that achieving even basic tasks like object recognition could take a considerable time. An good illustration of how quickly information can be transmitted through the brain is provided by the research of Simon Thorpe and colleagues. They showed that human subjects are able to discriminate the presence or absence of specific object categories, such as animals or cars, in visual scenes that are presented for very short times, as short as 20 ms. The percentage of correct manual responses, which consisted of releasing a button only when an animal was present in a complex image that was presented for 20 ms, is shown for 15 subjects in Fig. 1.5A plotted against the mean reaction time for each subject. The experiment shows some trade-off between reaction time and recognition accuracy, but the important point to note here is the high level of performance for such short presentations of the images.

The ability to recognize objects with these short presentation times is not the only astonishing result in these experiments. The authors also recorded skull EEGs during the experiments. The event-related potential, averaged over frontal electrodes, is shown in Fig. 1.5B, separated for image presentations with and without animals. The average response is not different for the first 150 ms, but becomes markedly different thereafter. The response of the frontal cortex therefore already indicates a correct answer after 150 ms. This is remarkable because for such categorization tasks we know that neural activity has to pass through several layers of brain areas. Each neuron in the processing stream necessary for the categorization path must thus be able to process and pass on information in time intervals of the order of only 10–20 ms or so.

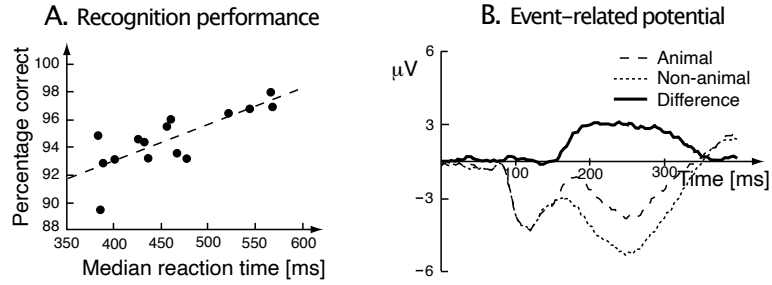


**Fig. 1.5** (A) Recognition performance of 15 subjects who had to identify the presence of an animal in a visual scene (presented for only 20 ms) versus their mean reaction time. (B) Event-related potential averaged over frontal electrodes across the 15 subjects [redrawn from Thorpe, Fize, and Marlow, *Nature* 381: 520–2 (1996)].

### 1.2.5 The layered structure of neocortex

Staining of cell bodies or neurites reveals a generally layered structure of the neocortex, as illustrated in Fig. 1.6. We include here some brief comments on experimental staining techniques to clarify assumptions and limitations of such techniques, since these techniques are often crucial for estimating parameters on which models are based. Many different staining techniques can be used to identify neurons or parts thereof. Some staining techniques, such as the Nissl stain, colour only the cell body and cannot be used to investigate dendritic or axonal organizations. The Golgi stain, based on a silver solution, can be used to visualize more parts of the neuron than those accessible by Nissl staining. When viewing illustrations of such stained tissues it is important to know that only a small percentage of neurons, on the order of only 1–2%, are stained by the Golgi staining method, and different neurons can have different receptivities to this stain. The appearance of neocortical slices visualized by different staining techniques is illustrated in Fig. 1.6A.

In addition to these traditional dyes there is now a variety of other staining techniques including direct intracellular dye injections reaching most parts of a neuron, anterograde staining that utilizes dyes that are taken up by the cell body and transported down the axons, and retrograde staining that utilizes dyes that are taken up by the terminal endings of axons and transported back to the cell body. The former two staining techniques can be used to identify the projection range of neurons, and the latter is useful to highlight the neurons that project into a particular brain area. Mastering such techniques and applying them carefully to get estimates of neuronal populations and dendritic or axonal organizations is a specialization within neuroscience on which computational neuroscientists rely heavily in order to develop biologically faithful models.

Historically, the neocortex is divided into six layers labelled with Roman numerals from I to VI, although more than six layers, commonly 10 including the white matter, can be identified and are included into the historical labelling scheme by further subdivisions. Layer IV is thereby subdivided into IVA, IVB, and IVC, and layer IVC is further subdivided into layers IVC$\alpha$ and IVC$\beta$. The extent, or thickness, of the layers varies throughout the neocortex up to a point where some layers are difficult to identify if not absent. Examples of stained slices from different areas within the

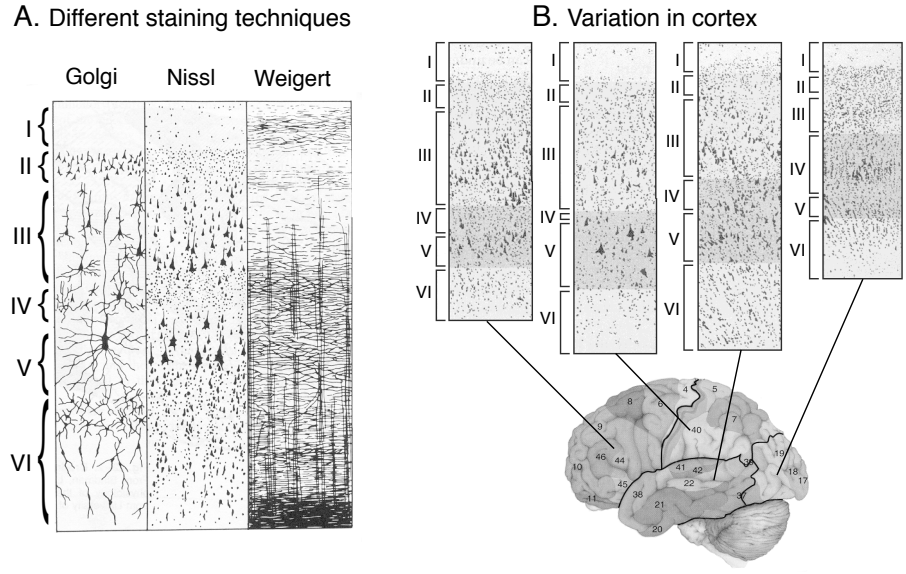


**Fig. 1.6** Examples of stained neocortical slices showing the layered structure of the neocortex. (A) Illustration of different staining techniques [adapted from Heimer, *The human brain and the spinal cord*, Springer, 2nd edition (1995)]. (B) Different sizes of cortical layers in different areas [adapted from Kandel, Schwartz, and Jessell, *Principles of neural science*, McGraw-Hill, 4th edition (2000)].

neocortex are shown in Fig. 1.6B. The visual appearance in the stained slices defining the layers is dependent on different populations of cell bodies and neurites. Layer I is easily distinguishable as it is mainly lacking in cell bodies and consists mainly of neurites. The other layers are marked by the domination of different cell types.

The soma of several neuronal types can be found in each neocortical layer, although the distribution can be used to mark the layers to some extent. As mentioned above, layer I is nearly completely lacking in cell bodies and consists mainly of neurites. Pyramidal cells can be found in most other layers of the neocortex. Layers II and III consist predominantly of small pyramidal cells, although the cells in layer III tend to be larger than those in layer II. Stellate neurons seem, in particular, concentrated around layer IV. In the upper part of this layer (IVA and IVB) one can find a mixture of medium-sized pyramidal cells and stellate cells, whereas the deeper layer (layer IVC) seems to be dominated by stellate neurons. Large pyramidal cells are found predominantly in layer V. A variety of cell types can be found in the deepest layer, layer VI. This includes Martinotti cells and also cells that have elongated cell bodies and are sometimes used to mark this layer. Such cells are sometimes called fusiform neurons.

### 1.2.6 Columnar organization and cortical modules

The neuronal organization in the neocortex discussed so far is mainly based on anatomical evidence. There is, in addition, an important functional organization in the neocor-

tex revealed by electrophysiological recordings. These experiments have shown that neurons in a small area of the cortex respond to similar features of an input stimulus. Hubel and Wiesel have investigated such organization in the primary visual (or striate) cortex. Neurons in this cortical area respond to visual bars moving in particular directions. More precisely, neurons in a small cortical column perpendicular to the layers and separated by around 30–100 $\mu$m respond to moving bars with a specific retinal position and orientation. These regions are called orientation columns. Separate from these arrangements are ocular dominance columns, cortical sections that respond preferentially to input from a particular eye (see Fig. 1.7A). The relations of orientation columns and ocular dominance columns are illustrated schematically in Fig. 1.7B. Neurons in small columns in other parts of the cortex also tend to respond to similar stimulus features. For example, cortical columns in the somatosensory cortex each respond to specific sensory modalities such as touch, temperature, or pain. The distribution of neurons with specific response characteristics is hence not purely random in the cortex, but there seems to be some form of organization.

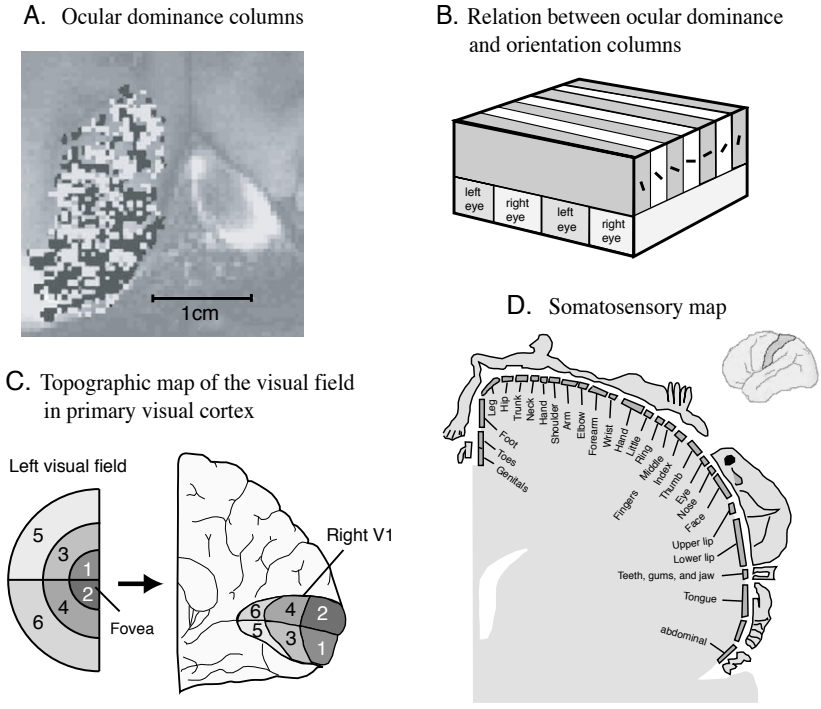


**Fig. 1.7** Columnar organization and topographic maps in the neocortex. (A) Ocular dominance columns as revealed by fMRI studies [from K. Cheng, R.A. Waggoner, and K. Tanaka, *Human ocular dominance columns as revealed by high-field functional magnetic resonance imaging*, *Neuron* 32: 359–74, (2001)]. (B) Schematic illustration of the relation between orientation and ocular dominance columns. (C) Topographic representation of the visual field in the primary visual cortex. (D) Topographic representation of touch-sensitive areas of the body in the somatosensory cortex. [(C) and (D) adapted from Kandel, Schwartz, and Jessell, *Principles of neural science*, McGraw-Hill, 4th edition (2000).]

Hubel and Wiesel called a collection of orientation columns representing a complete set of orientations a hypercolumn. They showed that adjacent hypercolumns in the striate cortex respond to visual input from adjacent retinal areas as illustrated in Fig. 1.7C. Central regions of the visual field are represented by a larger cortical area than peripheral areas. The mapping between the visual field and the cortical representation is therefore not area-preserving; the central visual area is over-represented, a feature that is called cortical magnification. However, the map preserves the relationships between adjacent points but not the area. Such maps are commonly labelled as topographic in the related literature. We will use this term in a general sense, meaning any map of a feature space with some systematic relations between points (features) on the map. For example, a tonotopic map, which is a map of sound representations, is topographic when adjacent frequencies are represented at adjacent locations in the map. A hypercolumn itself is a topographic map as it contains an ordered representation of orientation, and there are many more examples of such maps in cortex and in subcortical areas. One other example is illustrated in Fig. 1.7D, that of the somatosensory cortex which represents tactile input from different body parts. This cortical area represents again more sensitive areas with larger cortical areas. Chapter 8 explores mechanisms that can explain how such cortical organization can be formed though experience.

It is conceptually important that neurons in a small areas of the cortex respond to similar sensory stimuli as we can use this to simplify models of cortical organization and functions. For many models in this book it is therefore sufficient to represent the neurons in a certain area as a single unit, as discussed further in Chapter 5. With such population neurons it is then much easier to explore various brain mechanisms, such as the formation of topographic organizations or the transformation of representations.

### 1.2.7 Connectivity between neocortical layers

The connectivity pattern within the layered structure of neocortex is becoming increasingly important for computational models. Neurons in layer IV seem to receive a particularly large number of afferents through the white matter from subcortical and other cortical areas. This layer is therefore often viewed as an input layer. Layer V has many large pyramidal cells with axons extending into the white matter. This layer seems therefore to contribute largely to the output of cortical processing. As the white matter is the main pathway between remote cortical areas and, in particular, between cortical and subcortical areas, it is obvious to suggest that the information flowing through the white matter is, to a large extent, responsible for global information transmission in the brain. In contrast, pyramidal cells in layers II and III are thought to be largely responsible for long-range cortico-cortical tangential (lateral) connections. Martinotti cells in the deep layers of the neocortex have axons extending into layer I. These could be responsible for information transfer between adjacent cortical modules from which pyramidal neurons in the upper layers receive synaptic input. Stellate neurons, on the other hand, seem to be more local in their neuritic sphere. The smooth stellate cells are therefore candidates for inhibitory interneurons. Their role in the stabilization of cortical processing is an important issue that we will discuss in later sections of this book.

An outline of connectivity patterns within a small column of the neocortex is

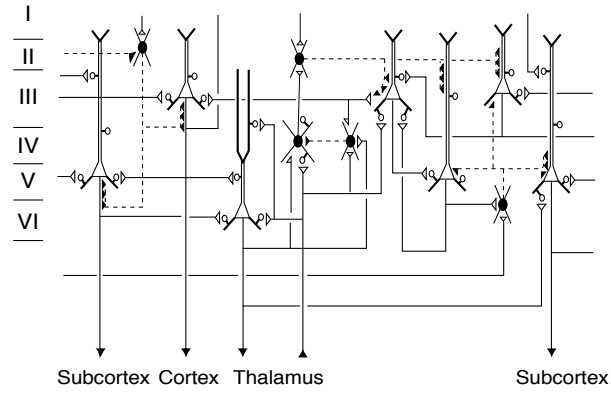


**Fig. 1.8** Schematic connectivity patterns between neurons in a cortical layer. Open cell bodies represent (spiny) excitatory neurons such as the pyramidal neuron and the spiny stellate neuron. Their axons are plotted with solid lines that end at open triangles that represent the axon terminal. The dendritic boutons are indicated by open circles. Inhibitory (smooth) stellate neurons have solid cell bodies and synaptic terminals, and the axons are represented by dashed lines [adapted from Douglas and Martin, in *Synaptic organization of the brain*, Shepherd (ed.), Oxford University Press (1990)].

summarized in Fig. 1.8. This scheme is, of course, only a rough approximation of the many details that are known experimentally. More detailed computational studies have still to be performed to understand the functional role of such organizations in more detail. An example of a model which incorporates laminar circuits is shown in Fig. 1.9. Grossberg and colleagues have now related such laminar models to many physiological and psychological findings, extending and unifying much of their earlier work. It is thereby interesting to note that, even on this level, cortical areas do not work in isolation. In the example shown it is important to consider the combined layered network of cortical area V1 and V2. Grossberg and collegaues showed that the deep layers (4–6) support thereby item storage, normalization of signals, and contrast enhancement, and that superficial layers (2/3) support grouping of information across processing channels, important factor in forming higher-order representations.

### 1.2.8 Cortical parameters

It is not feasible, and not a major scientific focus, to extract a detailed wiring diagram of the brain. Even an estimation of cortical parameters, such as the number of neurons in a cortical area, the number of connections and their physiological strength, the composition of an area with neuronal types, etc., is often not easy to extract experimentally. In addition, most of the experimental estimations can only be made for particularly favourable cases from which we have to generalize. The generalizations of such experimental studies are often obscured by considerable variations of such numbers within different cortical areas and between different species. Also, the estimation of such parameters varies considerably with different experimental techniques. You might therefore ask yourself how we can build biologically faithful network models of the brain without the necessary experimental support.

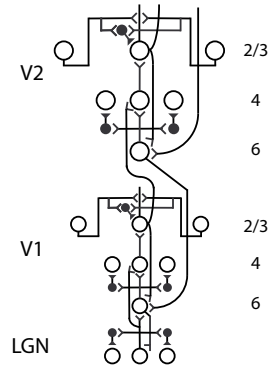


**Fig. 1.9** Example of laminar model for the early visual system which attributes specific information processing abilities to the laminar circuits. [With permission from S. Grossberg, see Grossberg, *Spatial Vision* 12: 163–86 (1999).]

The answer is that we have to approach the study of brain networks from different angles. We will study in this book primarily general network architectures and study the general computational capabilities of such networks. These studies reveal, as we will see throughout the book, that many computational abilities of the networks do not depend critically on specific details and are hence present in a large variety of networks within certain classes. Furthermore, we will discuss mechanisms that guide the development and fine tuning of networks to achieve specific computational tasks. We therefore approach the study of the brain from the perspective of extracting general principles that guide the organization of the brain as well as revealing the computational consequences of classes of structurally related networks.

To explain brain functions we have, of course, to concentrate on the classes of networks that are consistent with brain networks. Predictions of models therefore have to be tested carefully with experiments on the real brain. Biologically faithful models can also be guided by experimental estimations of general cortical organization. In Table 1.1 we summarize some rough estimates of neocortical parameters that are good to be aware of when discussing biologically faithful network models. The values presented only indicate an order of magnitude, which is good to keep in mind when developing very general models, and we will see that it is already instructive to study models with some very crude approximations of cortical organization. More specific estimates for specific cortical areas that are modelled should, of course, be taken into account for more specific studies.

How much of the detail of neocortical organization is necessary to explain certain brain functions is difficult to assess and has to be considered for each specific question. Some specifics are certainly essential for very detailed explanations, while we can gain a lot of insight into some information-processing principles in the brain from very general organizational principles. There are also good reasons to believe that the brain itself has to work within general architectures in contrast to very detailed specific architectures coded, for example, genetically. A generally accepted hypothesis is that the brain architecture is based on genetically coded organizational principles on which self-organization mechanisms and experience-based learning act to fine-tune

**Table 1.1** Some rough estimates of neocortical parameters [see for example Abeles, *Corticonics: neural circuits of the cerebral cortex*, Cambridge University Press (1991)]

| Variable | Value |
|---|---|
| Neuronal density | 40,000/mm$^3$ |
| Neuronal composition: | |
| Pyramidal | 75% |
| Smooth stellate | 15% |
| Spiny stellate | 10% |
| Synaptic density | $8 \times 10^8$/mm$^3$ |
| Synapses per neuron | 1000–20,000 |
| Distribution of synaptic types on pyramidal cell | |
| Inhibitory synapses | 10% |
| Excitatory synapses from remote sources | 45% |
| Excitatory synapses from local sources | 45% |
| Asynchronous gain (relative synaptic efficiency) | 0.003–0.2 |
| Time duration of spike | $\sim$ 1 ms |
| Velocity of spike (myelinated axon of 0.02 mm diameter) | 120 m/s |
| Length of axon | few mm to $\sim$ 1 m |
| Synaptic cleft | 20 nm |
| Synaptic transmission delay due to diffusion | 0.6 ms |

the organization to achieve accurate and flexible behaviour.

## 1.3 What is a model?

Modelling is an integral part of many scientific disciplines, and neuroscience is no exception. The more complex a system is, the more we have to make simplifications and build example systems to provide insights into aspects of the complex system under investigation. The term 'model' appears frequently in many scientific papers, and describes a vast variety of constructs. Some papers present a single formula as a model, some papers fill several pages with computer code, and some describe with words a hypothetical system. We need to understand what a model is, and, in particular, what the purposes of models are.

It is important to distinguish a model from a hypothesis or a theory. Scientists develop hypotheses of the underlying mechanisms of a system that have to be tested against reality. In order to test a specific feature of a hypothesis, we build a model that can be evaluated. Sometimes we try to mimic real systems under artificial means, in order to be able to test the systems by different conditions, or to make measurements that would not be possible in a 'real' system. A model is hence a simplification of a system in order to test particular aspects of the system or hypothesis. A brief explanation of a model, which is useful to remember throughout this book and in research, is:

> A model is an abstractions of a real-world system to demonstrate particular features of, or investigate specific questions about, the system. Or, in more

scientific terms, a model is a quantification of a hypothesis to investigate the hypothesis.

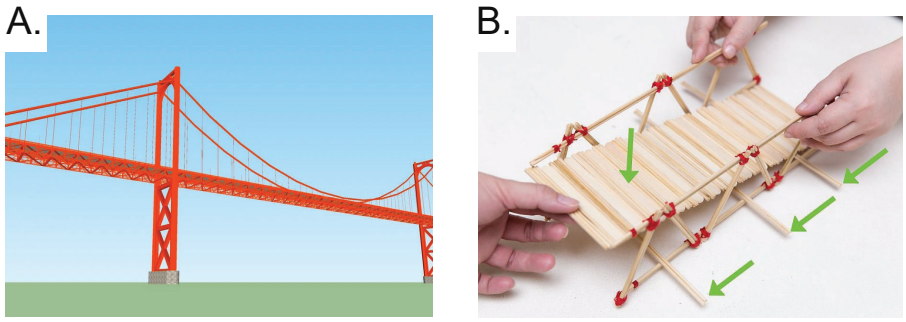


**Fig. 1.10** (A) A computer model of a building which gives a three-dimensional impression of the design [image courtesy of 3dwarehouse.sketchup.com]. (B) A physical model of a bridge that can be used to test the statics [image courtesy of wikiHow.com].

A good example of this is the use of models in the field of architecture and structural engineering. Small-scale paper models of buildings, or computer graphics (Fig. 1.10) generated with sophisticated three-dimensional graphics packages, can be used to get a first impression of the physical appearance and aesthetic composition of a design. Or we can build a model to test some statics of a design. A model has a particular purpose attached to it and has to be viewed in this light. A model is not a recreation of the 'real' thing. The paper model of a house cannot be used to test the stability of the construction, a purpose for which a building engineer uses different models. In such models, it is important to scale down physical properties of the building materials regardless of the physical appearance, such as the colour of the building.

### 1.3.1 Phenomenological and explanatory models

In science, we typically represent experimental data in the form of graphs, and then seek to describe these data points with mathematical functions (Fig. 1.11). An example of this is the 'modelling' of response properties (tuning curves) of neurons in the lateral geniculate nucleus (LGN), which can be fitted with a specific class of functions called Gabor functions by adjusting the parameters of these functions. Gabor functions are therefore said to be 'models' of the receptive fields in the LGN. Of course, this phenomenological model does not tell us anything about the biophysical mechanisms underlying the formation of receptive fields and why cells respond in this particular way, so such a 'model' seems rather limited. Nevertheless, it can be useful to have a functional description of the response properties of LGN cells. Such parametric models are a shorthand description of experimental data that can be used in various ways. For example, if we want to study a model of the primary visual cortex, to which these cells project, then it is much easier to use the parametric form of LGN responses as input to the cortical model, rather than including further complicated models of the earlier visual pathway in detail.

As scientists, we want to find the roots of natural phenomena. The explanations

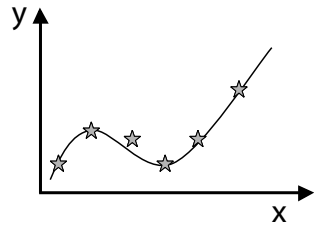


**Fig. 1.11** The graph shows some data points, plotted with star symbols, such as data derived from experimental measurements. Also shown is a model represented by the line, and this curve fits the data reasonable well. The curve can be derived from a simple mathematical formula that fits the data points (phenomenological model), or result from more detailed models of the underlying system.

we are seeking are usually deeper than merely parameterizing experimental data with specific functions. Most of the models in this book are intended to capture processes that are thought of as being the basis of the information-processing capabilities of the brain. This includes models of single neurons, networks of neurons, and specific architectures capturing brain organizations. Models are used to study specific aspects of a hypothesis or theory, but can also help to interpret experimental data.

A major role of models in science is to illustrate principles underlying natural phenomena on a conceptual level. Sometimes the level of simplification and abstraction is so crude that scientist talk about toy models, although this terminology undermines slightly the importance of such models in science. The simplifications made in such models might be necessary to employ analytical methods to analyse such models at a depth that is not possible in more realistic models. The educational importance of these toy models should not be underestimated, in particular, in demonstrating principal mechanisms of natural phenomena. It is easy to complicate things, but the real scientific challenge is to simplify theoretical concepts.

The current state of neuroscience, often still exploratory in nature, frequently makes it difficult to find the right level of abstraction to properly investigate hypotheses. Some models in computational neuroscience have certainly been too abstract to justify claims derived from them. On the other hand, there is a great danger in keeping too many details that are not essential for the scientific argument. Models are intended to simplify experimental data, and thereby to identify which details of the biology are essential to explain particular aspects of a system. Modelling the brain is not a contest in recreating the brain in all its details on a computer. It is questionable how much more we would comprehend the functionality of the brain with such complex and detailed models. Models must be carefully constructed, and reasonable simplification is the result of careful scientific investigations and insight into natural processes. It is not always possible to justify all assumptions adequately, but it is important to at least clarify which assumptions have been made and to argue about them. The purpose of models is to better comprehend the functionality of complex systems, and simplicity should be a major guide in designing appropriate models. This philosophy is sometimes called the principle of parsimony, also known as Occam's razor. Basically, we want the model as simple as possible, while still capturing the main aspects of data that the model should capture. We will see this principle in action throughout this book, and discuss the issue

specifically in conjunction with machine learning in Chapter 6.

### 1.3.2 Models in computational neuroscience

Where do we start when trying to explain brain functions? Do we have to re-build the brain in its entirety with computational techniques in order to understand it? Currently, many neuroscientists are collaborating in an ambitious project, called the Blue Brain Project, to implement in a computer, with as much detail as possible, neocortical circuits. This type of detailed project can show us which emergent phenomena can be expected in the brain, or where gaps in our knowledge require further investigation. However, even if it was be possible to simulate a whole brain on a computer, with all the details from biochemistry to large-scale organization, this would not explicitly mean that we have better explanations of brain functions. What we are looking for, at least in this book, is a better comprehension of brain mechanisms on explanatory levels. It is therefore important to learn about the art of abstraction, making suitable simplifications to a system without abolishing the important features we want to comprehend.

The level that is most appropriate for the investigation and explanatory abstraction depends on the scientific question. For example, epileptic seizures are known to be caused by synchronization of whole brain areas, and an understanding of brain dynamics on a systems level is important to comprehend such disorders. This should, of course, not exclude the search for causes on much smaller scales. We know that Parkinson's disease is caused by the death of dopaminergic neurons in the substantia nigra, and the role of dopamine in the initiation of motor actions is important to comprehend the full scale of impairments, and to develop better methods of coping with such conditions. The causes of the cell death of dopaminergic neurons are still not known, and it is important to find the reasons, possibly on a genetic level in this case, to enable a real treatment of Parkinson's disease. Therefore, various levels must be investigated, and we have to learn to make connections between the different levels in order to follow how the low-level circumstances, such as mechanisms on a genetic level or biochemical processes in neurons, can influence the characteristics of large-scale systems, such as behaviour of an organism.

## 1.4 Is there a brain theory?

It is often said that the brain is the most complex system that we know of in nature, and understanding how it works is a large, if not impossible, task. It is true that understanding how the brain works seems difficult when trying to reverse engineer it. However, reverse engineering a system can even be difficult for systems that we created in the first place. For example, imagine measuring the varying states of a transistor in a computer while the computer is running a word processor. Such measurements give important clues from which it should be possible to discover regularities when certain operations are repeated on the computer. Also, such measurements make it possible to discover important principles that must be at work in the digital computer system, such as the discrete nature of information representations. However, it seems a daunting task to recreate a computer program such as a word processor from these data, even if we were able to analyse a large number of measurements of many parts

of the computer at the same time. The direct reverse engineering of the brain from data seems even more challenging, given the biological nature of the object. However, we can use the data to understand the principles of brain processing, and we can then use this knowledge to build brain-style information-processing systems. This approach is taken this book.

Correspondingly, there has been some shift in the research approach of the neuroscience community. The past decades have been an era in neuroscience marked by a flood of explorations. Recordings with micro-electrodes from single cells contributed significantly to this exploration, and searching for response properties of neurons is at least as exciting as it must have been for explorers such as Marco Polo to discover new lands. New brain-imaging techniques, such as functional magnetic resonance imaging (fMRI), make it possible to monitor living brains of subjects performing specific mental tasks. Explorations of brain functions with such techniques have been essential in advancing our knowledge in neuroscience. However, while a mountain of data has been gathered for brain functions, using a multitude of techniques, we are now slowly entering a new phase in neuroscience, that of formulating more quantitative hypotheses of brain functions. This shift in the focus of neuroscience research demands some more specific experimental analysis and more dedicated tests of such hypothesis. It is increasingly important to formulate a quantitative hypotheses, and possible alternatives, in such a way that the hypothesis can be tested experimentally.

This new era of neuroscience sounds a lot like quantitative scientific areas such as chemistry or physics, and while a quantitative analysis will generate new breakthroughs in our understanding of brain functions, the ultimate question is if there can be a brain theory. We could take the position that in order to understand the brain we need to understand all of the structural details and the current state of a particular brain. This is no different in other scientific areas such as physics. For example, to completely describe the physics of an individual aeroplane we have to know the precise location and form of each nut and bolt and all other structural details up to the amount of dirt on the wings and the details of the air it is flying through. Another example is that of a pot of boiling water, which consists of a lot of individual molecules in a very dynamic state. Measuring all the microscopic details in the last two examples seems impossible. Yet, we have a fairly good understanding of the process of boiling water and why an aeroplane can fly. This was not possible overnight but is the result of dedicated scientific research over the past few centuries. Important for the success in physics, chemistry, and other scientific disciplines, was the realization of the right level of description or the right level of abstraction of a problem. For example, we learned to describe the average behaviour of the molecules in an ideal gas, which gave us the fundamentals of thermodynamics and ultimately led to a better understanding of the mechanisms of flowing air that enables engineers to construct more efficient aeroplane. We know today about the essential quantum nature of atomic and subatomic interactions, but a description of Mount Everest on this level is not reasonable. There are geological theories of mountain formation that are more appropriate than employing quantum theory to these questions.

There is no reason to conjecture that the brain cannot be tractable with similar scientific rigour to that developed in other disciplines. The brain is certainly more complex than a gas of weakly interacting atoms. However, there are very fundamental

questions that we can attack; for example, how a network stores memories that can be recalled in an associative way. Indeed, we have made considerable progress with this question, considering our understanding of associative networks discussed in this book. Another fundamental question is why the brain is relatively stable, while still being able to adapt to novel environments. Brain theories of this kind are now emerging, and some of these theories will be discussed in this book. It may be too early to talk about a single brain theory, but there is no reason to suspect that theories will advance our understanding of brain functions. Indeed, I think that major breakthroughs have already been made, which need to be brought to a wider audience. The goal of this book is to contribute to this endeavour.

### 1.4.1 Emergence and adaptation

Standard computers, such as PCs and workstations, have one or more central processors. Each processor is rather complex, with specialized hardware and microprograms implementing a variety of functions, such as loading data into registers, adding, multiplying, and comparing data, as well as communicating with external devices. These basic functions can be executed by instructions, which are binary data loaded into a special interpreter module. Complicated data processing can be achieved by writing, often lengthy, programs, which are instructions representing a collection of the basic processor functions. When solving a task with a computer, we have to instruct the machine to follow precisely all the steps that we determined beforehand would solve a particular problem. The sophistication of computers basically reflects the ingenuity of the programmer.

In contrast, information processing in the brain is very different in several respects. The brain employs simpler processing elements than computers, but lots of them. To explore information processing in networks of neurons, we will mainly use very simple abstractions of real neurons, which we call nodes to stress this drastic simplification. These fundamental processing units can be implemented in hardware, or simulated on a standard computer; for the discussions in this book this does not make a difference. We keep the functionality of nodes as simple as possible for the sake of employing lots of them, typically hundreds, thousands, or even more. The usage of many parallel working processors has motivated the term parallel distributed processing in this area. However, with this term one is tempted to think that the processes are independent because only processes that are independent can be processed on different processors in parallel. In contrast to this, a major ingredient of information processing in the brain is the interaction of neurons, and the interaction of neurons is accomplished by assembling them into large networks.

It is the interaction of nodes that enables processing abilities not present in single nodes. Such capabilities are good examples of emergent properties in rule-based systems. Emergence is the single most defining property of neural computation, distinguishing it from parallel computing in classical computer science, which is mainly designed to speed up processing by distributing independent algorithmic threads. Interacting systems can have unique properties beyond the mere multiplication of single processor capabilities. It is these types of abilities we want to explore and utilize with neural networks. These system properties are labelled as emergent to stress that we did not encode these properties directly into the system. To better appreciate this, we

distinguish the description of a system on two levels, the level of basic rules defining the system, and the level of description aimed at understanding the consequences of such rules. In the study of neural networks we are interested in understanding the consequences of interacting nodes.

Scientific explanations were dominated in the past by the formulation of a set of principles and rules that govern a system. A system can be defined by a set of rules, like in a game. In science, we assume that natural systems are governed by a finite set of rules. The search for these basic rules, or fundamental laws as they are called in this case, was the central scientific quest for centuries. It is not easy to determine such laws, but enormous progress has been made nevertheless. Newton's laws defining classical mechanics, or the Maxwell equations of electromagnetism are beautiful examples of fundamental laws in nature. We do not have a theory of the brain on this level, and some have argued that we might never find a simple set of rules explaining brain functions. However, in many scientific disciplines we are beginning to realize that even with a given set of rules we might still not have a sufficient understanding of the underlying systems. This is analogous to the idea that knowing the rules of the card game Bridge is not sufficient to be a good Bridge player.

Rules define a system completely, and it can therefore be argued that all the properties of a system are encoded in the rules. However, we have to realize that even a small set of rules can generate a multitude of behaviours of the systems, which might be difficult to understand from the basic rules due to the presence of emergent properties. A different level of description might then be more appropriate. For example, thermodynamics can describe appropriately the macroscopic behaviour of systems of many weakly interacting particles, even though the systems are governed by other microscopic rules. On the other hand, there are emergent properties in Newtonian systems that are not well described by classical thermodynamics, such as turbulent fluids. A deeper understanding of emergent properties is becoming a central topic in the science of the twenty-first century.

The importance of emergent properties in networks of simple processing elements is not the only extension of traditional information-processing approaches that we think are crucial for understanding brain functions. Another important ingredient is that the brain is an example of an adaptive system. In our context, we define adaptation as the ability of a system to adjust its response to external stimuli in dependence of the environment states and the expectations of the system. Humans are a good example of systems that have mastered adaptive abilities and learning. Adaptation and learning is an area that has attracted a lot of interest in the engineering community. The reason for this is that learning systems, which are systems that are able to change their behaviour based on examples in order to solve information-processing demands, have the potential to solve problems where traditional algorithmic methods have not been able to produce sufficient results. Adaptation has two major virtues. One is, as just mentioned, the promise to solve information-processing problems for which explicit algorithms are not yet known. A second virtue is our aim to build systems that can cope with continuously changing environments. A lot of research in the area of neural networks is dedicated to the understanding of learning. Engineering applications of neural networks are not bound by biological plausibility. In contrast to this, we concentrate in this book on biologically plausible learning mechanisms that

can help us to comprehend the functionality of the brain.

### 1.4.2 Levels of analysis

David Marr and Tomaso Poggio realized the usefulness of distinguishing between different levels of description when explaining processes, which Marr later explains beautifully in his posthumously published book *Vision*. One issue that Marr raised in his book was that different people need different kinds of explanations. Explanations of brain functions that are satisfactory for a non-specialist can be insufficient for a physiologist or psychologist who is trying to make sense of specific observations in his or her expriments. A computer engineer may require different types of explanations of brain functions again, to be able to implement specific information-processing algorithms. Marr also stressed the difference between understanding computers and understanding computation. Knowing how bits are represented and transformed in a computer is far from understanding high-level applications such as the World Wide Web.

Besides these important considerations for any explanatory theory, Marr made an important distinction between different levels of analysis. These are summarized in Marr's book as:

1. **Computational theory:** What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
2. **Representation and algorithm:** How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
3. **Hardware implementation:** How can the representation and algorithm be realized physically?

The most abstract and general level, that of a computational theory, is concerned with what a process is trying to achieve, and what the principal approach to the solution is. In his book, Marr discusses the example of a cash register, which is adding up the price of goods brought to the checkout counter. This explanation answers the question of what the cash register is doing, but we need also to ask why the cash register is doing it (adding the numbers) in this specific way. The answer to this question is that the rules of adding numbers, rather than multiplying them, encapsulates what we think is appropriate for this process; that prices of individual goods should just add up, independent, for example, of the order in which the goods are presented. Marr assigns great importance to this level and explains in his book:

> To phrase the matter in another way, an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and hardware) in which it is embodied.

Having a clear theory of what the brain, or specific brain processes, are trying to accomplish can be a powerful research guide and needs to be considered more in neuroscientific research.

The next level of description is concerned with how the computation, specified on the computational level, is realized on an algorithmic level. Marr considered brain research as an information-processing problem, and he clearly was aware of the duality between representing and processing information. That is, in order to process information, this information has first to be represented in a specific form. For example,

numbers can be represented in binary form or with Roman numbers. Representations are important since the specific algorithm used for implementing a process usually depends on the representation. Many different representations are possible, and there are usually many algorithms for each type of representation which can accomplish the task. Different representations can drastically influence algorithms, and different algorithms can have quite different properties, such as robustness or efficiency. Theories in computational neuroscience have strongly focused on representations, on which specific algorithms are built. It should also be mentioned that there are examples that use the arguments in the opposite direction, such as deriving representations from constraints on algorithms, such as robustness. Clearly, the brain must use specific representations, and specific algorithms, and it is the goal of computational neuroscience to help find them.

As a side note, in the definition of this representational and algorithmic level, Marr speaks specifically of input and output representations, and the process of transforming, or mapping, input to output. We will soon discuss why this mapping view is not adequate for brain functions. In a later section of his book, Marr discusses the difficulty in finding invariant features from changing data. This is an example that is difficult in a mapping framework but which can be resolved easily in the computational theory of the brain presented below. But this criticism should not distract from the importance of this level of explanation.

The third level of Marr's scheme is concerned with the physical realization of the algorithms and representation. This is the level to which neuronal biophysics and physiology can speak most directly. Again, there might be several possible physical realizations of a specific algorithm, and we want to understand the physical realizations in the brain. This level of explanation does not only back up the level of explanations above, but is also important in its own right for specific interventions. The different levels are weakly related, and inspiration and constrains from different levels of analysis can guide research on other levels.

Since this book is intended to be an introductory guide, we follow mainly a bottom-up approach, through learning first about some physical properties of the nervous system, how information is represented in the brain, and some algorithmic theories of solving specific problems. A broad knowledge of such issues is necessary for any advanced work in neuroscience. However, as Marr suggested, it is important to guide research more specifically through computational theories. We therefore outline in the following section a broad computational theory of the brain to which we return in more detail in the final chapter.

## 1.5 A computational theory of the brain

### 1.5.1 Why do we have brains?

One of the first questions we might want to ask is why we have a brain at all. While this question has likely been asked and answered in many different ways, the take of Daniel Wolpert is particularly illuminating. Wolpert studies sensorimotor control; how brains are able to supervise appropriate movements in a complex sensory world. He notes that plants live stable lives without a brain and he also tells the story of a little

sea creature, called the sea squirt, which is hatched with a small nervous system and swims through the ocean until it settles down on some rock, after which it digests its brain. Hence, it seems that animals need a brain to move.

We can go further and ask why we want to move and why it seems that increasingly complex brains are developed even though we there are much simpler creatures that move. Moving around can help to find food and sexual partners to enable survival of its species and to make evolutionary progress. Mathematically, we view this as maximizing an objective function. With this view, we can imagine that nervous systems developed into more and more sophisticated systems that help to achieve the evolutionary objectives of moving organisms. The human brain is thereby, arguably, the most developed example of such evolved systems, even inventing machines to travel. Thus, a more formal answer to what the brain is doing might be:

> The brain produces goal-directed behaviour to maximize our probability of survival.

If individual brains exist to maximize the organism's survival capacity, then why could it happen that a teenager gets drunk and falls off a cliff? The reason is that maximizing survival probability acts on a population, not on an individual level. The brain implements a specific strategy for each individual, which itself may not be perfect but necessary to explore new solutions. Experimenting with the unknown is essential for finding new, and ultimately better, solutions, so that the adventurous behaviour of a teenager can ultimately help our society.

### 1.5.2 The anticipating brain

In order for the brain to produces goal-directed behaviour to maximize our probability of survival, or in short, for the brain to make good decisions, we argue that there are two principle strategies. One is to build a reactive system that learns from previous rewards in the environment. Such habitual learning and decision-making is widespread in adaptive agents including humans. If we touch a hot stove we learn fast to avoid this in the future without deliberative effort. Such a system is central in fast decision-making and is sometimes called System 1.

The other strategy, that of system 2, is to build a model of the world that can be used to calculate good decisions. Such a deliberative system is of course much more complex and requires a much more elaborate information-processing system. Such a system is likely much more complex to learn, and decisions have to be calculated and therefore require more mental effort and are likely more time consuming. However, a huge advantage is that such a system is able to generalize better to more novel situations with causal relations.

Both systems must be able to function with uncertainties. This is important as the real world has either hidden causes or is too complex to comprehend or too difficult to sense in all its details. We will encounter more concrete discussions of the different decision systems in Chapter 11, and we stress some of the probabilistic underpinnings throughout the book. Here we outline some important ingredients for building both systems and the view of the brain as an anticipatory decision system.

Mechanistically, we can start describing brain as a decisions system that senses the world and then produce appropriate actions to achieve some objectives. In a