1 Computational Neuroscience

Neuroscience is a scientific area that aims to understand the nervous system, in particular, the brain. The brain is studied by researchers from different disciplines, such as anatomy, physiology, medicine, physics, biology, biochemistry, genetics, psychology, mathematics, and computer science. The challenging field of neuroscience must cross boundaries, and a multidisciplinary approach with the combined efforts from multiple scientists is necessary to understand the brain. Neuroscience has also given rise to other disciplines, including neuroeducation, neuroethics, neurolaw, neuroaesthetics, and neuromarketing, to name a few.

Computational neuroscience is an increasingly important branch of neuroscience that employs mathematical models, theoretical analyses, and abstractions of the brain. Computational neuroscience develops and tests hypotheses of brain mechanisms. Models are often analytically intractable. Models are compared to experimental data using carefully designed numerical experiments. **Data-driven computational neuroscience** employs statistical and computational models learned from data, obtained from disparate sources, such as the electrical activity of a neuron or a confocal microscopy image of a neuron, a recording of a neuronal population, magnetic resonance imaging of the whole brain, or microarray data from a patient with Parkinson's disease (PD). Models are evaluated and can be used to make predictions that must be experimentally verified.

This chapter starts with a basic introduction of the multilevel organization of the brain in Section 1.1 and some figures from the human brain in Section 1.2. Section 1.3 presents the main brain research initiatives worldwide. Section 1.4 covers helpful and recently developed neurotechnologies to understand the brain and record data. Data-driven computational neuroscience is elaborated in Section 1.5 from a statistical and machine learning perspective, in which data-sharing and bidirectional brain–computer benefits are described. Finally, the data sets employed throughout the book are described at length in Section 1.6.

1.1 The Multilevel Organization of the Brain

1.1.1 Multiscale Organization

The brain is a complex system whose functional and structural organization is characterized by a hierarchy of spatial and temporal scales (Bassett and Gazzaniga, 2011). The relationship between the mind and brain is far from being understood (Ascoli, 2015), but characterizing the structure of the brain and its organizing principles is a necessary

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

first step. Different levels of organization range from the molecular and cellular levels to networks, systems, and behavior. All of these levels form the physical and biological bases of cognition. Moreover, the structure within any given scale is organized into modules.

Neuroscientists – and hence societies, jobs, conferences, and publications – are usually categorized according to the organizational level they primarily address. However, these levels do not define independent domains, and their conceptual integration should be targeted. Therefore, all neuroscientists should have a basic understanding of the functions of the brain at different scales. Although the dependence on specialized and expensive equipment is a reason to investigate a single level of organization in experimental approaches, homogeneous computational tools are also available for theoretical approaches. Thus, integrative multilevel concepts addressing how one level constrains or informs another are lacking. For example, it is unknown how a mechanism at the genetic level influences the characteristics of large-scale systems, such as the behavior of an organism. An exception is the Hodgkin–Huxley formalism, which explains how the properties of membrane components determine the electrical behavior of entire neurons.

Nevertheless, cross-level integration is a difficult task. Relationships between phenomena at different levels are nonlinear and highly complex. These relationships are difficult to articulate in mathematical terms, and scientists have achieved the first step by investigating how variables at one level influence variables at another level. Unlike experimental approaches with technical or ethical limitations used to study some variables, computational approaches benefit from the free manipulation of parameters and complete reproducibility. Computational models can foster multilevel investigations. A multilevel computational model might integrate elements from compartmental neuron models, microcircuit representations of neuronal populations, and activity propagation in large-scale neuronal networks.

1.1.2 Spatial and Temporal Scaling

In the spatial domain, the brain has many levels of organization, ranging from the molecular level of a few Angstroms ($1Å = 10^{-10}$m), to the whole nervous system of over a meter. The **neuron** is a cell that is specialized for signal processing. Neurons generate electric potentials to transmit information to other cells via special connections called **synapses**. Mechanisms operating at the subcellular level play a role in information-processing capabilities. Neurons use cascades of biochemical reactions that must be understood at the molecular level, including the transcription of genetic information. The complexity of a single neuron makes computational models essential, and substantial progress has been achieved at this level.

Minicolumns are vertical columns that extend through the cortical layers of the brain. Minicolumns are the anatomical basis of columns, contain approximately 100 neurons and are 30 microns in diameter. More complex constructs include subareas (e.g., S2), areas (e.g., the somatosensory cortex), lobes (e.g., the temporal lobe), and the complete cerebral cortex.

Neurons connect to each other to form **neural circuits**. Networks of interconnected neurons exhibit complex behaviors and enable additional information-processing capabilities that are not present in a single neuron. This unique property of **emergence** of neural computation extends beyond the mere multiplication of single-processor capabilities. The

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

brain is also organized into higher-order levels. Networks with a specific architecture and specialized information-processing capabilities are included in larger structures or systems that interact and enable more complex information-processing tasks and new emergent properties. The central nervous system depends on the dynamic interaction of many specialized subsystems and the interaction of the brain with the environment.

In the temporal domain, the organization of the brain dynamically changes over multiple temporal scales. Inherent rhythms of brain activity vary in different frequencies (the highest frequency gamma band is >30 Hz, whereas the delta band is 1–2 Hz) and relate to different cognitive capacities. Learning and memory change neuronal connection patterns (through synaptic plasticity) on both short (seconds to minutes) and long (hours to days to months) timescales.

1.1.3 Modular Organization

Organization within a given scale is modular. Thus, the brain is decomposed into subsystems or modules. For example, anatomical modules in the spatial domain are present in cortical minicolumns or columns, whereas short- and long-term memory are the modules in the temporal domain. Within-module elements are more highly connected than between-module elements. This organization provides a compartmentalization that reduces the interdependence of modules and enhances robustness. Hierarchy and modularity together allow the formation of complex architectures of subsystems within subsystems with a high degree of functional specificity. Furthermore, modularity enables behavioral adaptation because each module functions and changes its function without negatively affecting the rest of the system.

The modular architecture has been more formally described by **complex network theory** applied to neuroimaging data (Sporns, 2010). The functional and structural hierarchical modularity of the connectivity of the human brain has been reported. Within these modular structures, brain regions perform different roles, as hubs with higher connectivity or as local processors. These regional roles are evident in both structural and functional connectivity networks and might have neurophysiological correlates. Each region displays different patterns of energetic activity and maintains different trajectories of synaptic development and redevelopment or plasticity. Hierarchical modularity is compatible with the minimization of energy consumption in developing and maintaining wiring, where most of the energy is used for the function of synapses. Physical constraints of wiring are also compatible with the spatial configuration of the observed connectivity (close regions interact strongly, whereas long-range anatomical connections or functional interactions connect very different modules).

The physical anatomical constraints of the human brain also constrain its function. Thus, two coherently active regions of the brain are often connected by a direct white matter pathway. Recently, researchers have attempted to map the wiring of the brain at different levels of spatial resolution (Section 1.3). Researchers have not yet clearly determined how structural connectivity might help predict function because a one-to-one relationship between structure and function is not plausible, and function appears to emerge from multiscale structures (a many-to-many mapping). Emergence occurs between multiple physical and functional levels. Causation seems to occur both upward and downward between multiple levels (located either in close proximity or distant regions) of

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

the brain, creating a complementary or mutually constraining environment of mental and physical functions. The brain is nonreducible, and its organization is not a simple sum of its parts.

1.2 The Human Brain

The human brain has a similar structure to the brains of other mammals, but it is larger than any other species in relation to body size, namely, it has the largest encephalization quotient. Much of the expansion is derived from the part of the brain called the cerebral cortex, which is a thick layer of neural tissue that covers most of the brain. This layer is folded in a way that increases the surface area to fit into the available volume. The cerebral cortex is divided into four lobes, called the frontal lobe, parietal lobe, temporal lobe, and occipital lobe (see Figure 1.1). Numerous cortical areas exist within each lobe, each of which is associated with a particular function, including vision, motor control, language, etc. The left and right sides of the cortex are broadly similar in shape, and most critical areas are replicated on both sides. However, some areas, particularly areas that are involved in language, show strong lateralization. In most people, the left hemisphere is dominant for language, whereas the right hemisphere is usually responsible for spatiotemporal reasoning.

Some quantitative measures of this organ are described below. The brain weighs between 1,200 and 2,000 grams in adults (see Figure 1.2) and between 350 and 400 grams in newborns and accounts for only 2% of an adult's total body weight. The average brain is 140 mm wide, 167 mm long, and 93 mm high. With a thickness of 1.5 to 4.5 mm and a total surface area of 2,500 cm^2 (Peters and Jones, 1984), the cerebral cortex accounts for greater than 80% of the brain mass but contains only 19% of all brain neurons (similar to other mammals). The total cerebral cortical volume is divided across the four lobes as follows: frontal lobe (41%), temporal lobe (22%), parietal lobe (19%), and occipital lobe (18%). As estimated using a novel quantitative tool called isotopic fractionation (Herculano-Houzel and Lent, 2005), the cerebral cortex contains 86 billion neurons (greater than the

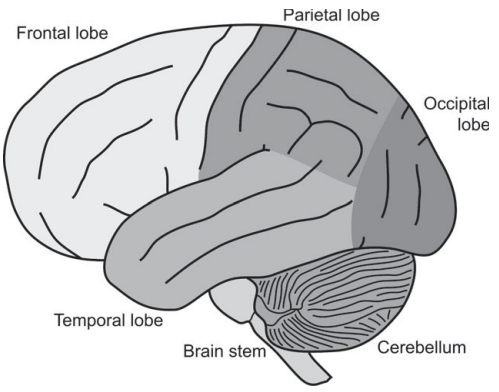
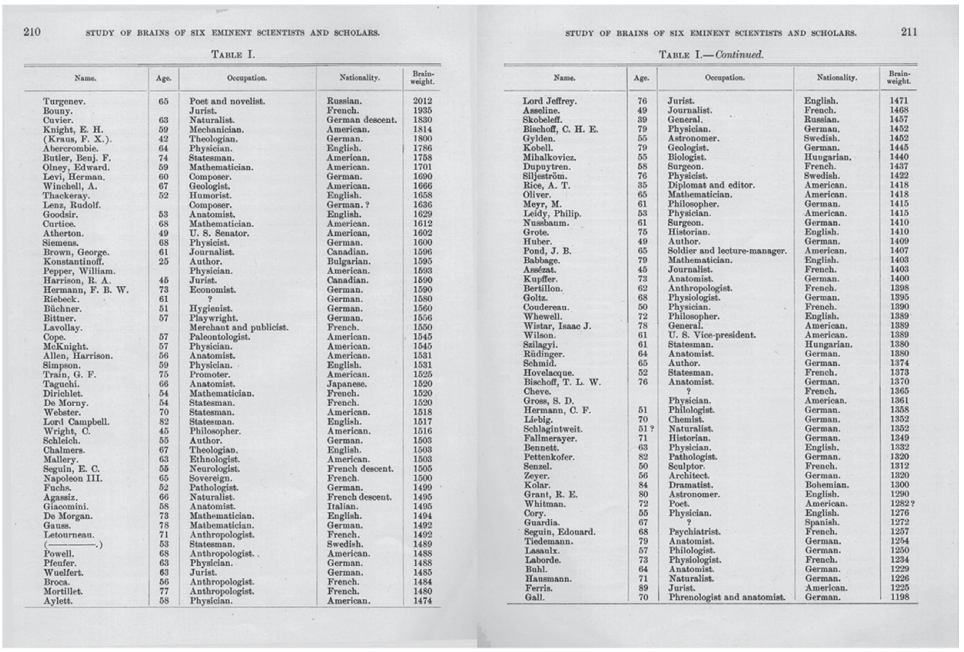


Figure 1.1 Lobes in the human cerebral cortex. Image taken from the Wikimedia Commons repository. For the color version, please refer to the plate section.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press



210 STUDY OF BRAINS OF SIX EMINENT SCIENTISTS AND SCHOLARS.

TABLE I.

Name.	Age.	Occupation.	Nationality.	Brain-weight.
Turgenev.	65	Poet and novelist.	Russian.	2012
Bouny.		Jurist.	French.	1935
Cuvier.	63	Naturalist.	German descent.	1830
Knight, E. H.	59	Mechanician.	American.	1814
(Kraus, F. X.).	42	Theologian.	German.	1800
Abercrombie.	64	Physician.	English.	1786
Butler, Benj. F.	74	Statesman.	American.	1758
Olney, Edward.	59	Mathematician.	American.	1701
Levi, Herman.	60	Composer.	German.	1690
Winchell, A.	67	Geologist.	American.	1666
Thackeray.	52	Humorist.	English.	1658
Lenz, Rudolf.		Composer.	German. ?	1636
Goodsir.	53	Anatomist.	English.	1629
Curtice.	68	Mathematician.	American.	1612
Atherton.	49	U. S. Senator.	American.	1602
Siemens.	68	Physicist.	German.	1600
Brown, George.	61	Journalist.	Canadian.	1596
Konstantinoff.	25	Author.	Bulgarian.	1595
Pepper, William.		Physician.	American.	1593
Harrison, R. A.	46	Jurist.	Canadian.	1590
Hermann, F. B. W.	73	Economist.	German.	1590
Riebeck.	61	?	German.	1580
Büchner.	51	Hygienist.	German.	1560
Bittner.	57	Playwright.	German.	1556
Lavollay.		Merchant and publicist.	French.	1550
Cope.	57	Paleontologist.	American.	1545
McKnight.	57	Physician.	American.	1545
Allen, Harrison.	56	Anatomist.	American.	1531
Simpson.	59	Physician.	English.	1531
Train, G. F.	75	Promoter.	American.	1525
Taguchi.	66	Anatomist.	Japanese.	1520
Dirichlet.	54	Mathematician.	French.	1520
De Morny.	54	Statesman.	French.	1520
Webster.	70	Statesman.	American.	1518
Lord Campbell.	82	Statesman.	English.	1517
Wright, C.	45	Philosopher.	American.	1516
Schleich.	55	Author.	German.	1503
Chalmers.	67	Theologian.	English.	1503
Mallery.	63	Ethnologist.	American.	1503
Seguin, E. C.	59	Neurologist.	French descent.	1505
Napoleon III.	65	Sovereign.	French.	1500
Fuchs.	52	Pathologist.	German.	1499
Agassiz.	66	Naturalist.	French descent.	1495
Giacomini.	58	Anatomist.	Italian.	1495
De Morgan.	73	Mathematician.	English.	1494
Gauss.	78	Mathematician.	German.	1492
Letourneau.	71	Anthropologist.	French.	1492
(————.)	53	Statesman.	Swedish.	1489
Powell.	68	Anthropologist.	American.	1488
Pfeufer.	63	Physician.	German.	1488
Wuelfert.	63	Jurist.	German.	1485
Broca.	56	Anthropologist.	French.	1484
Mortillet.	77	Anthropologist.	French.	1480
Aylett.	58	Physician.	American.	1474

STUDY OF BRAINS OF SIX EMINENT SCIENTISTS AND SCHOLARS. 211

TABLE I.—*Continued.*

Name.	Age.	Occupation.	Nationality.	Brain-weight.
Lord Jeffrey.	76	Jurist.	English.	1471
Asseline.	49	Journalist.	French.	1468
Skobeleff.	39	General.	Russian.	1457
Bischoff, C. H. E.	79	Physician.	German.	1452
Gylden.	55	Astronomer.	Swedish.	1452
Kobell.	79	Geologist.	German.	1445
Mihalkovicz.	55	Biologist.	Hungarian.	1440
Dupuytren.	58	Surgeon.	French.	1437
Siljeström.	76	Physicist.	Swedish.	1422
Rice, A. T.	35	Diplomat and editor.	American.	1418
Oliver.	65	Mathematician.	American.	1418
Meyr, M.	61	Philosopher.	German.	1415
Leidy, Philip.	53	Physician.	American.	1415
Nussbaum.	61	Surgeon.	German.	1410
Grote.	75	Historian.	English.	1410
Huber.	49	Author.	German.	1409
Pond, J. B.	65	Soldier and lecture-manager.	American.	1407
Babbage.	79	Mathematician.	English.	1403
Asézat.	45	Journalist.	French.	1403
Kupffer.	73	Anatomist.	German.	1400
Bertillon.	62	Anthropologist.	French.	1398
Goltz.	68	Physiologist.	German.	1395
Coudereau.	50	Physician.	French.	1390
Whewell.	72	Philosopher.	English.	1389
Wistar, Isaac J.	78	General.	American.	1389
Wilson.	61	U. S. Vice-president.	American.	1389
Szilagyi.	61	Statesman.	Hungarian.	1380
Rüdinger.	64	Anatomist.	German.	1380
Schmid.	65	Author.	German.	1374
Hovelacque.	52	Statesman.	French.	1373
Bischoff, T. L. W.	76	Anatomist.	German.	1370
Cheve.		?	French.	1365
Gross, S. D.		Physician.	American.	1361
Hermann, C. F.	51	Philologist.	German.	1358
Liebig.	70	Chemist.	German.	1352
Schlagintweit.	51 ?	Naturalist.	German.	1352
Fallmerayer.	71	Historian.	German.	1349
Bennett.	63	Physician.	English.	1332
Pettenkofer.	83	Pathologist.	German.	1320
Senzel.	50	Sculptor.	French.	1312
Zeyer.	56	Architect.	German.	1320
Kolar.	84	Dramatist.	Bohemian.	1300
Grant, R. E.	80	Astronomer.	English.	1290
Whitman.	72	Poet.	American.	1282?
Cory.	55	Physician.	English.	1276
Guardia.	67	?	Spanish.	1272
Seguin, Edouard.	68	Psychiatrist.	French.	1257
Tiedemann.	79	Anatomist.	German.	1254
Lasaulx.	57	Philologist.	German.	1250
Laborde.	73	Physiologist.	French.	1234
Buhl.	64	Anatomist.	German.	1229
Hausmann.	71	Naturalist.	German.	1226
Ferris.	89	Jurist.	American.	1225
Gall.	70	Phrenologist and anatomist.	German.	1198

Figure 1.2 Name, age, occupation, nationality, and brain weight (grams) of different personalities. Taken from Spitzka (1907).

number of all known stars in the universe) and 240 trillion synapses (Koch, 1999). The relationship between the body and brain size that applies to other primates is not true for humans, where the brain size is five to seven times larger than expected according to the body size.

The appearance of the neocortex is a decisive event during the evolution of the mammalian telencephalon. Its activity is directly related to those capacities that distinguish humans from other mammals. By this reason, the neocortex can be considered as the most human part of the nervous system (DeFelipe, 2011).

1.2.1 Brain Sizes Variability for Different Species

Figure 1.3 shows the brain size for different mammalian species from human to mouse. The variability in brain weights is remarkable. For example, the insectivorous white-toothed pygmy brain weighs 0.060 g, while the heaviest brain corresponds to the sperm whale, with 9.200 kg on average. The brain of the Indian elephant weighs 6.900 kg, a similar quantity to the brain of the blue whale, the largest animal on Earth that has a body 20 times larger. By contrast, the gorilla and striped dolphin have similar body weights, although the gorilla's brain weight is less than half of the dolphin's.

The power law exponents that apply to the scaling of brain mass as a function of the number of neurons are 1,550, 1,016, and 1,056 for rodents, insectivores, and primates, respectively. The absolute number of neurons, in contrast to the body and brain size-centered view, has been proposed as the most relevant parameter for determining the cognitive abilities across species (Gazzaniga, 2008).

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

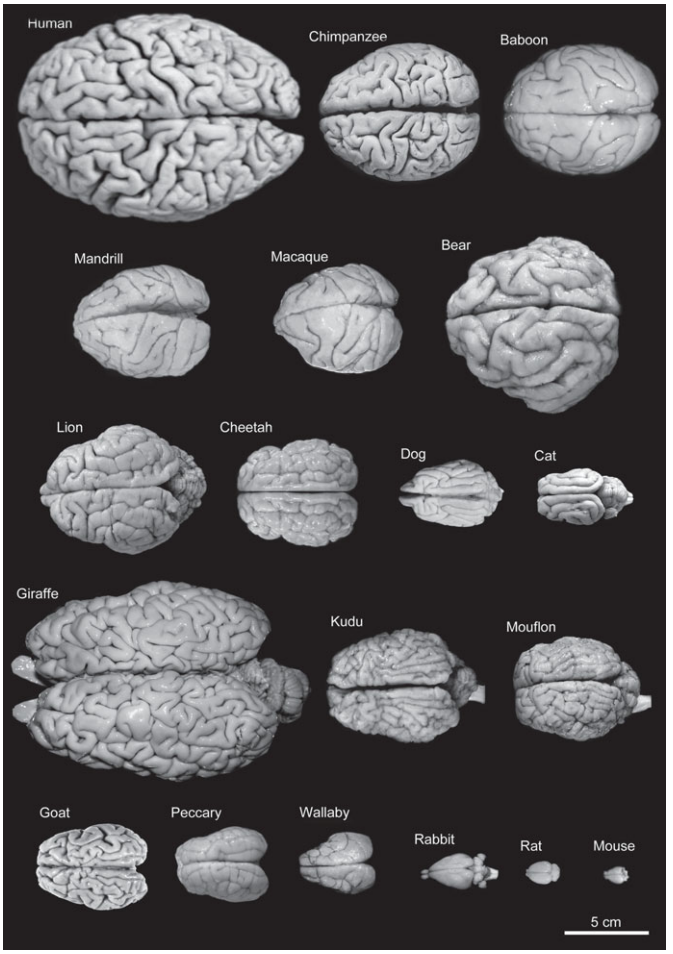


Figure 1.3 Variability of brain sizes for several mammals. Primates: human (1.176 kg), chimpanzee (273 g), baboon (151 g), mandrill (123 g), macaque (110 g). Carnivores: bear (289 g), lion (165 g), cheetah (119 g), dog (95 g), cat (32 g). Artiodactyls: giraffe (700 g), kudu (166 g), mouflon (118 g), goat (115 g), peccary (41 g). Marsupials: wallaby (28 g). Lagomorphs: rabbit (5.2 g). Rodents: rat (2.6 g), mouse (0.5 g). Image from DeFelipe (2011).

1.3 Brain Research Initiatives

In recent years, several brain-mapping initiatives have been initiated worldwide and are attempting to tackle one of the most fascinating challenges of the twenty-first century. Although these initiatives have different goals and areas of expertise, the common aim is to move closer to unlocking the elusive secrets of the human brain and to pursue myriad previously inaccessible scientific questions. This joint effort will require the merging of historically distinct scientific disciplines, such as engineering, chemistry, physics, and computer science, with neuroscience and psychology in so-called convergence science. This section describes the main goals of some of these initiatives developed by the European Commission, the United States, Japan, China, Canada, Korea, and Australia (see Figure 1.4).

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

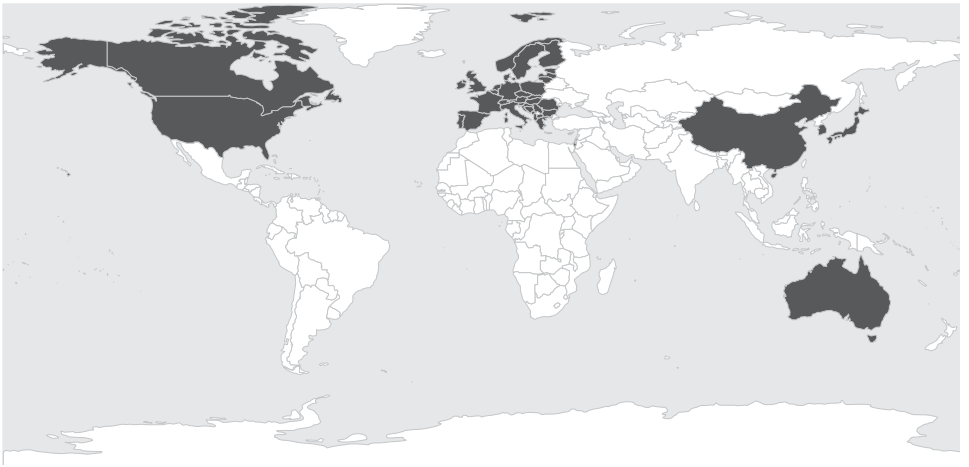


Figure 1.4 Regions around the world participating in brain research initiatives.

The **Human Brain Project** (HBP)[1] (Amunts et al., 2016), launched by the European Commission in 2013, is one of the first two Future and Emerging Technologies Flagship projects, a long-term (10 years) multinational research project. This project is based on the hypothesis that decoding the multiscale questions about the human brain requires deeper insights into structure and function of the brain at all levels of organization – from genes to the whole brain – with interdisciplinary experts, including neuroscientists, physicists, and mathematicians, as well as modern information and communication technologies (ICTs). More than 100 partner institutions in 19 countries in Europe collaborate by sharing data and tools. Four neuroscience subprojects, six research platforms, and two subprojects – ethics and society – and coordination and central services that cross-link all platforms and subprojects, as well as an education and training program that incorporates a multidisciplinary outlook, constitute the main elements of the HBP. The neuroscience subprojects aim to identify the organizational principles of spatial and temporal brain architecture and consist of (a) the organization of the mouse brain, (b) the organization of the human brain, (c) systems and cognitive neuroscience, and (d) theoretical neuroscience. The six research platforms are (a) the neuroinformatics platform, an effective ecosystem for software and data sharing; (b) the brain simulation platform, which aims to develop data-driven models and brain simulations at all scales; (c) the high-performance analytics and computing platform, which helps the neuroscience community compete using high-end supercomputers and systems for large-scale data analytics; (d) the medical informatics platform, which provides new diagnostic categories and new treatments for brain diseases as a result of the convergence between ICT, biology, and medicine; (e) the neuromorphic computing platform, which implements brain-like principles in machine learning and cognitive computing; and (f) the neurorobotics platform, which is developing simulating robots controlled by spiking neural networks.

The **Brain Research through Advancing Innovative Neurotechnologies Initiative** (BRAIN)[2] (Martin and Chun, 2016) was launched by President Obama in April 2013

[1] www.humanbrainproject.eu/en/.
[2] www.braininitiative.org.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

to "accelerate the development and application of new technologies that will enable researchers to produce dynamic pictures of the brain that show how individual brain cells and complex neural circuits interact at the speed of thought." The BRAIN Initiative is a public–private partnership including the National Institutes of Health (NIH), the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), the Intelligence Advanced Research Projects Activity (IARPA), the Food and Drug Administration (FDA), and over 20 private foundations, institutes, universities, companies, and international partners. The BRAIN Initiative seeks to obtain a better understanding of the inner workings of the human mind and to improve how we treat, prevent, and cure brain disorders. These goals will be accomplished by pursuing the following activities: (a) advancing neurotechnologies to enable scientists to monitor and modulate brain circuit activity; (b) facilitating dynamic imaging to produce a dynamic picture of brain function in real time; (c) exploring brain functionality to investigate how the brain records, processes, uses, scores, and retrieves vast quantities of information; (d) linking function and behavior by incorporating new theories and computational models; and (e) advancing consumer applications by developing safe and effective products for patients and consumers. In addition, the BRAIN Initiative provides a monthly report aimed at a general audience describing the potential applications of these tools in research or in clinical settings, with the objective of creating an environment that sustains the enthusiasm of scientists, the general public, and even policy-makers.

The **Brain Mapping by Integrated Neurotechnologies for Disease Studies** (Brain/MINDS)[3] (Okano et al., 2016) is a national brain project started by Japan in 2014 with a 10-year roadmap. Brain/MINDS has adopted a fundamentally different approach compared to BRAIN and HBP by focusing on accelerating the development of the common marmoset as a model for the exploration and discovery of knowledge-based strategies for the eradication of major brain diseases. The achievement of this goal requires (a) the development of a multiscale marmoset brain atlas and integrated data platform to support functional studies, (b) the generation of a genetically modified marmoset for experimental and preclinical studies, and (c) the creation of a clinical data center using translational biomarkers for the diagnosis and treatment of human brain diseases. Currently, 65 laboratories and 47 institutions in Japan and several partner countries collaborate on this project. The research is organized into four major groups: (a) structural and functional mapping of the marmoset brain, (b) development of innovative neurotechnologies for brain mapping, (c) human brain mapping and clinical research, and (d) advanced technology and application development. The marmoset was chosen as the experimental model because it maintains a unique phylogenetic position and because the analysis of certain higher cognitive behaviors is, in some cases, easier in a nonhuman primate than in rodents or other simple vertebrate models.

The **China Brain Project** (Poo et al., 2016) is a 15-year project (2016–2030) that focuses on macaques and whose central pillar is understanding the neural basis of human cognition to develop new preventive, diagnostic, and therapeutic approaches, as well as brain-inspired computing methods and systems that are considered essential to achieving more robust artificial intelligence. This approach is known as the "one body-two wings" scheme. Research investigating the neural circuits underlying the mechanisms of cognition

[3] www.brainminds.jp.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

will aim to understand human cognitive processes at different levels: from behavior to neural systems and circuits, to cells and molecules to the brain at the macroscopic and mesoscopic levels. One of the "wings" is the study of the pathogenic mechanisms and the development of effective and early diagnostic – at presymptomatic and prodromal stages – and therapeutic approaches for brain disorders with a developmental (e.g., autism and mental retardation), neuropsychiatric (e.g., depression and addiction), and neurodegenerative (e.g., Alzheimer's disease [AD] and PD) origin. This "wing" also refers to the efforts to provide a scientific basis for Chinese medicine, whose underlying mechanisms are largely unknown. The other "wing," brain-inspired computation, assumes that the human brain is currently the only truly generally intelligent system in nature that is capable of coping with different cognitive functions with extremely low energy consumption. Consequently, learning from the information processing mechanisms of the brain is clearly a promising method for building stronger and more general machine intelligence. One of the challenges of the China Brain Project is to make a general artificial intelligence that is capable of multitasking, learning, and self-adapting.

Brain Canada (Jabalpurwala, 2016)[4] is a project that was established by the Government of Canada in 2011 and expected to be completed in 6 years, but has been extended annually since 2016. Its vision is to understand the function of the brain in both healthy and disease states, improve lives, and achieve societal impacts. Brain Canada is achieving its goal by (a) increasing the scale and scope of funding to accelerate the pace of Canadian brain research; (b) creating a collective commitment to brain research across the public, private, and voluntary sectors; and (c) delivering transformative, original, and outstanding research programs. The approach is based on three main ideas: (a) one brain that considers this organ as a single complex system in which brain diseases and disorders often share common underlying mechanisms, such as cell loss, abnormal functioning of nerve cells, or chemical and molecular imbalances; (b) collaborative research encouraging high-risk, high-reward investigations that enable and support multidisciplinary teams; and (c) one community that includes governments, voluntary health organizations, philanthropists, business leaders, patients, caregivers, health administrators, clinicians, and, of course, researchers and their host institutions.

The **Korea Brain Initiative** (Jeong et al., 2016) was announced in May 2016, and the project was launched in 2018. The overall plan includes the development of novel neurotechnologies and the reinforcement of the neuroindustry with a vision to advance brain science by establishing and facilitating local, national, and global collaborative networks. The primary goal of the initiative is to foster neuroscience that improves the scientific understanding of the principles of higher brain functions to produce a new dynamic picture of healthy and diseased brains. Additional goals are to develop personalized treatments for mental and neurological disorders by extrapolating the concept of precision medicine and to stimulate collaboration among scientific institutes, academia and industry. The scope of the research project includes (a) constructing brain maps at multiple scales, (b) developing innovative neurotechnologies for brain mapping, (c) strengthening artificial intelligence-related research and development, and (d) developing personalized medicine for neurological disorders.

[4] www.braincanada.ca.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

A proposed **Australian Brain Initiative** is being developed by members of the Australian Brain Alliance.[5] This initiative was presented and discussed with the scientific community, government, and public in 2017. Its mission is to create an innovative and healthy nation by cracking the brain's code through an understanding of the mechanisms underlying the development of the neural circuitry, how the brain encodes and retrieves information, complex behaviors, and the adaptations to external and internal changes. The initiative is designed to address the following four challenges: (a) optimizing and restoring healthy brain function throughout life, (b) developing neural interfaces to record and control brain activity to restore its function, (c) understanding the neural basis of learning across the lifespan, and (d) delivering new insights into brain-inspired computing.

The **International Brain Initiative** was launched on September 19, 2016 in the United Nations' General Assembly in New York City with the United States, Argentina, Japan, and Germany as partners. One of the several goals for the initiative is to create universal brain-mapping tools. Two interesting ideas proposed at the meeting were the creation of (a) an International Brain Observatory, with tools such as powerful microscopes and supercomputing resources that scientists around the world could access, and (b) an International Brain Station that would automatically convert data from studies of the human brain or animal gene expression into standardized formats that would allow more people to analyze them (Reardon, 2016a). At the same time, the World Health Organization (WHO) wishes to ensure that the early discoveries and technological advances of the different brain initiatives are translated into tests and treatments for brain disorders, avoiding health disparities between developed and underdeveloped countries (Reardon, 2016b).

1.4 Neurotechnologies

Neurotechnology is the area of technology that includes every advance that helps researchers understand the brain. The field has only reached maturity in the last 20 years, particularly due to the advent of various brain imaging techniques. However, extensive research is still needed. In addition to visualizing the brain both for clinical and research purposes, which is the focus of this section, technologies are available that are designed to improve and repair brain functions. Drugs are available to control depression or sleep; improve motor coordination in patients with PD, Huntington's disease (HD), amyotrophic lateral sclerosis (ALS), and stroke; reduce epileptic episodes; and alleviate phantom pain perception, among others.

1.4.1 Visualizing a Single Neuron

Currently, fine-scale recordings of the electrical activity of a *single* neuron or a small group of neurons (even in living humans) is relatively easy. Needle-like electrodes are inserted into the brains of laboratory animals to stimulate neurons. In addition to being an invasive technique, this approach provides an incomplete picture of the whole brain and samples brain activity very sparsely. The probable multineuronal level of organization (system,

[5] www.brainalliance.org.au/.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

network, cellular, subcellular, and molecular levels) is not visualized using single-neuron recordings. New properties appear at each level of complexity.

The use of advanced microscopes is critical to examine neurons and their structures. **Confocal microscopy** (or confocal laser scanning microscopy), whose principle was patented in 1957 by Minsky (1988), one of the fathers of artificial intelligence, overcomes some limitations of traditional wide-field fluorescence microscopes. Confocal microscopes can examine thick samples, particularly samples with dense fluorescent staining of complex structures, where the relevant information may be hidden by out-of-focus haze. By adding a spatial pinhole placed at the confocal plane of the lens to eliminate out-of-focus light, the optical resolution and contrast are increased. Unlike a conventional microscope that can just view structures in the sample at the depth that the light penetrates, a confocal microscope only captures images at one depth level at a time. The depth of focus is controlled and limited. The surface of the sample is then scanned by moving either the sample or the light beam (horizontally), thereby reconstructing a 2D image at a specified depth. Next, vertical movements allow researchers to capture sets of images at different depths (optical sectioning), and 3D images of the sample are created using appropriate software.

The amount of magnification achieved by an optical microscope is limited by the wavelength of light. The shorter the wavelength of the light waves, the smaller the objects the microscope can see. The photons of visible light have a relatively large wavelength. However, electrons form waves with a much shorter length. This principle is the basis of **electron microscopy**, which uses a beam of accelerated electrons as a source of illumination instead of light. This technique allows researchers to examine tissues in greater detail – their ultrastructure – with magnifications of up to 10 million times, whereas most light microscopes achieve magnifications of less than 2,000x. At higher magnifications, the light waves start interfering with one another, and the images become blurry. The two most common types are the **transmission electron microscope** and the **scanning electron microscope** (SEM). Advances in microscopy for neuroscience are reviewed in Wilt et al. (2009).

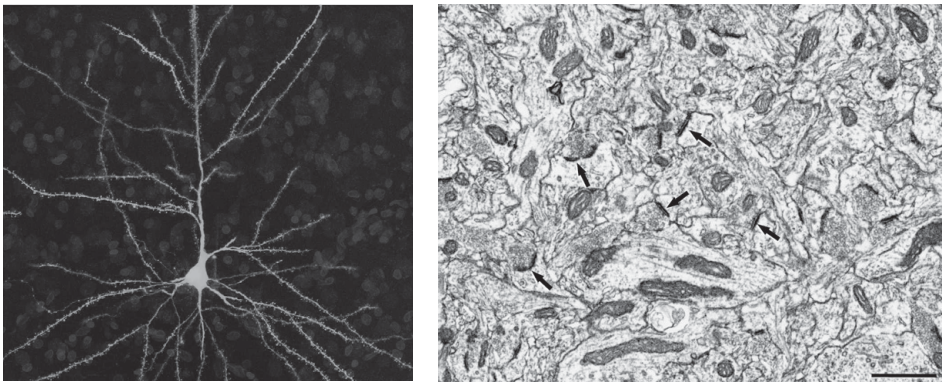
Figure 1.5 shows (a) a confocal microscopy image of an intracellular injected layer III pyramidal neuron in the human cingulate cortex, and (b) an electron microscopy image of synapses in the rat cerebral cortex.

1.4.2 Tracking Circuits of Neuronal Activity

However, we need to record neural activity across complete neural circuits. A map of the anatomical connections, or synapses, among neurons (the so-called **connectome**) is only a starting point that is unable to depict the constantly varying electrical activity underlying specific cognitive processes. The interesting issue is how a collection of neurons interact intricately and give rise to an emergent property (Yuste and Church, 2014). Moreover, due to their plasticity, neurons are continuously subjected to dynamic rearrangements.

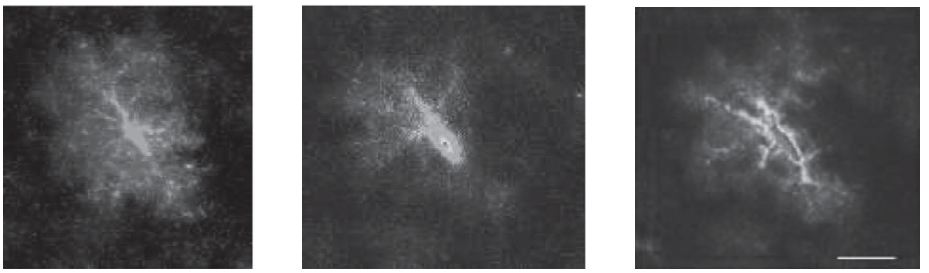
Large-scale recordings can be accomplished with the aid of nanotechnology; prototype arrays with more than 100,000 electrodes on a silicon base are able to track thousands of neurons in the retina. Stacking these arrays into 3D structures would multiply their scalability.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press



(a) (b)

Figure 1.5 (a) An intracellular injected layer III pyramidal neuron of the human cingulate cortex imaged with a Leica TCS 4D confocal scanning laser attached to a Leitz DMIRB fluorescence microscope. DAPI staining is presented in blue. Image kindly supplied by Javier DeFelipe (Benavides-Piccione et al., 2013). (b) SEM image of the rat cerebral cortex. Arrows indicate some asymmetric synapses. Scale bar: 1 μm. Image kindly supplied by Javier DeFelipe (Morales et al., 2011). For the color version, please refer to the plate section.



(a) (b) (c)

Figure 1.6 (a) Z-stack projection of a hippocampal astrocyte expressing the genetically encoded calcium indicator GCaMP6f. (b) Pseudocolor calcium images of the astrocyte depicted in (a) before electrical stimulation of the Schaffer collaterals. (c) The same image after stimulation. Scale bar: 20 μm. Images provided by Ana Covelo from the Alfonso Araque laboratory at the University of Minnesota. For the color version, please refer to the plate section.

In addition to electrical sensors (electrodes), new techniques for imaging neuronal activity are based on physics, chemistry, genetics, and engineering. Thus, in **calcium imaging**, cells are genetically engineered to fluoresce when calcium ions enter the neuron after it fires. The firing patterns of more than 1,000 neurons can be partially reconstructed in vitro or in vivo. This technique operates too slowly (limited time resolution compared to electrical recordings) to track the rapid firing of neurons. Additionally, new types of microscopes that show the simultaneous activity of neuronal populations in 3D are needed. Figure 1.6 shows a calcium image of a hippocampal astrocyte.

In **voltage imaging**, specific dyes alter the optical properties as the voltage of the neuronal membrane changes. The dyes are deposited on the neuron or across the cell membrane through genetic engineering. Although this technique is still in its infancy, the activity of every neuron in an entire circuit could be potentially recorded. Voltage sensors

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

may be composed of organic dyes, genetic indicators, or even nonbiological materials borrowed from nanotechnology (i.e., nanoparticle quantum dots or nanodiamonds), which are sensitive to neuronal activity.

The delivery and collection of light from neural circuits located deep below the surface of the brain is difficult. **Computational optics** may help control the fluorescence emitted from dyes when subsurface neurons fire in a similar manner to the method used by astronomers to correct image distortions due to atmospheric effects on starlight. New optical hardware includes two-photon imaging, high-numerical aperture objectives, and light-field cameras, to name a few. Techniques such as **microendoscopy** are used by neuroradiologists to penetrate further into the tissue and image deeper structures. In this case, a flexible tube with light guides is inserted into the femoral artery and moved to many parts of the body (including the brain). Based on **synthetic biology**, lab animals have been genetically engineered to synthesize a molecule ("molecular ticker tape") that changes when a neuron is activated, or artificial cells function as sentinels that patrol the body or are placed near a neuron to detect its firing through a nanosized circuit implanted in the artificial cells that wirelessly transmits the data to a nearby computer.

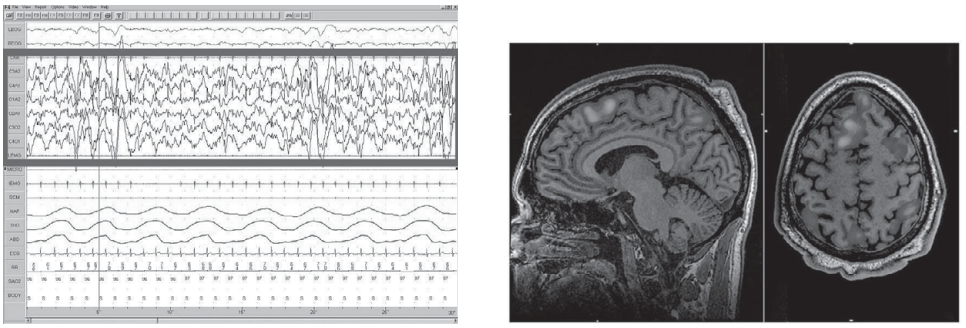
In addition to monitoring circuits of neuronal activity, the ability to freely activate and inactivate these circuits will help researchers determine the functions of the selected cells and control some forms of brain activity (e.g., epileptic seizures, Parkinsonian tremors, reward responses). Recent technologies for this application rely on optical signals, such as **optogenetics** and **optochemistry**. In the former approach, genetically engineered neurons produce light-sensitive proteins (bacteria- or algae-derived), causing them to either become activated or inactivated upon exposure to light of a particular wavelength through an optical fiber. In the latter approach, neurotransmitters are attached to light-sensitive chemicals that are activated upon exposure to light. These techniques are minimally invasive, provide great spatial and temporal single-cell resolution, and have been applied to living tissues.

1.4.3 Imaging Large Brain Regions

Methods have been developed to track the activity of neurons across the *whole* brain within the field called cognitive neuroscience. Thus, in **electroencephalography** (EEG), electrodes placed on the skull measure the coordinated activity of more than 100,000 neurons. The brainwave activity associated with neuronal depolarization is registered over a few milliseconds, but without identifying whether a specific neuron is active. Thus, EEG offers high temporal (real-time) resolution but poor spatial resolution because the precise origin of the signal is difficult to locate (Burle et al., 2015). Other electrical techniques include **magnetoencephalography** (MEG), which measures the magnetic fields produced by electrical activity in the brain and is less sensitive to the head geometry compared to EEG, and **electrocorticography** (ECoG), an invasive procedure requiring a surgical incision into the skull to implant the electrode grid.

Other techniques are based on metabolism and indirectly measure neuronal activity. Metabolic techniques are classically considered as having very good spatial resolution but rather poor temporal resolution, while the opposite trends are observed for electrophysiological techniques. The most widely used technique is **functional magnetic resonance imaging** (fMRI). This technique illuminates active brain areas in 3D maps, where each

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press



(a) (b)

Figure 1.7 (a) EEG image. Screen shot of a patient during slow wave sleep when he or she is snoring (stage 3); the image represents 30 seconds. The high amplitude EEG is highlighted in red. Public domain from the English Wikipedia. (b) An fMRI image obtained during working memory tasks. Freely available at `ideas.ted.com`. For the color version, please refer to the plate section.

(cerebral cortical) voxel is composed of more than 600,000 neurons. The high spatial resolution and low temporal resolution (seconds and minutes) of fMRI allow researchers to record changes in blood flow within voxels, see Glover (2011) for a review. **Positron emission tomography** (PET) analyzes changes in blood flow and the consumption of oxygen and glucose. **Single-photon emission computed tomography** (SPECT) also monitors blood flow. PET and SPECT have a high detection sensitivity, see Lu and Yuan (2015) for a review of both. **Near infrared spectroscopy** (NIRS) is an optical technique used to measure blood oxygenation in the brain. Light in the near infrared part of the spectrum is transmitted through the skull, and the extent to which the remerging light is attenuated is measured, which depends on blood oxygenation.

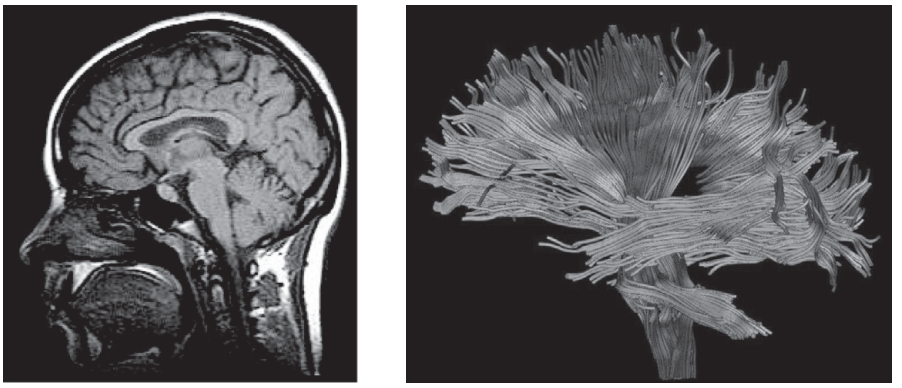
Figure 1.7(a) shows an EEG image from a patient during slow wave sleep; Figure 1.7(b) shows an fMRI image obtained during working memory tasks.

The previous techniques are functional and are predominantly used in cognitive neuroscience because they enable researchers to determine the location and timing of neural activity associated with performance on a cognitive task in patients with a disease and in healthy subjects. However, anatomical techniques are used differently, e.g., to localize neuropathies or to compare the size of specific brain structures between subjects through a volumetric analysis. A powerful technique is **magnetic resonance imaging** (MRI), which allows researchers to distinguish gray matter (neuronal cell bodies) from white matter (myelinated tracts). MRI visualizes anatomical structures based on the behavior of atoms in water (their protons) in a magnetic field. New anatomical techniques, such as **diffusion tensor imaging** (DTI), are designed to specifically visualize myelinated tracts. Figure 1.8 shows MRI and DTI.

Even when these techniques are used in combination in multimodal neuroimaging, e.g., EEG and MEG or EEG and fMRI (Uludağ and Roebroeck, 2014), their applications are limited. Using images of large brain regions, researchers can coarsely examine neuronal activity because they are unable to identify whether circuits are activated or inactivated.

Figure 1.9 presents the approximate temporal (x-axis) and spatial (y-axis) resolution achieved using the main neurotechnologies. Approaches and techniques span from the nanoscale to the macroscale in terms of spatial and temporal resolution.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press



(a) (b)

Figure 1.8 (a) MRI. Image available at Wikimedia Commons. (b) DTI of the lateral brain tractogram. Anonymous clinical image provided by Aaron G. Filler in Wikipedia. For the color version, please refer to the plate section.

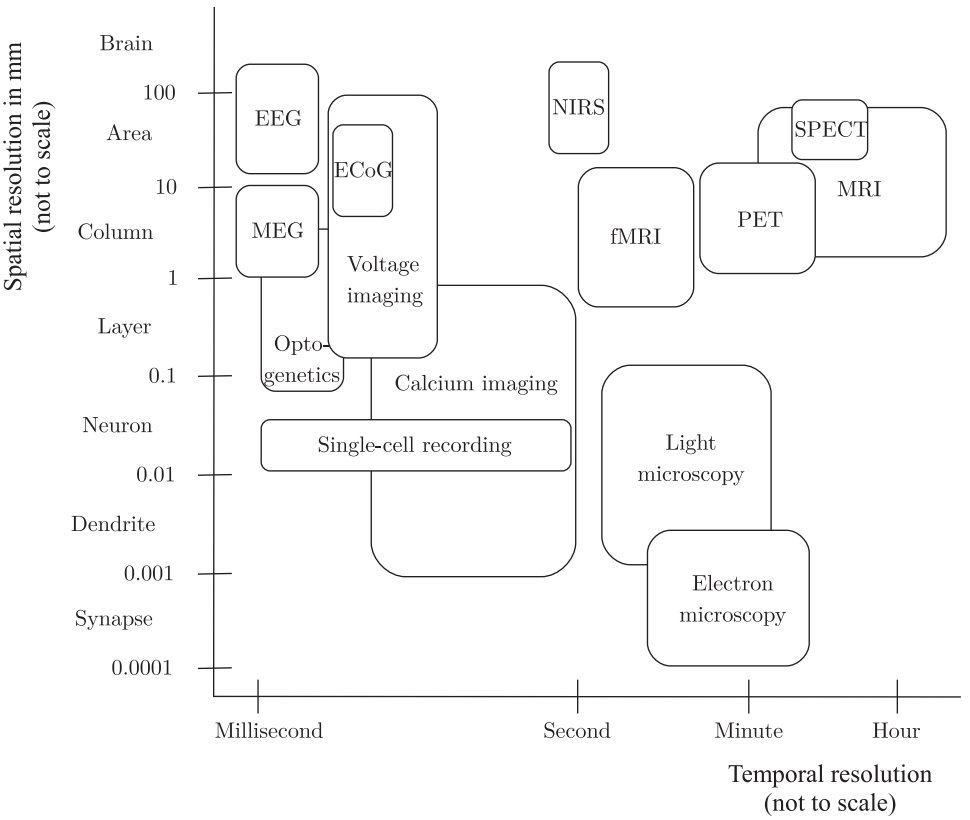


Figure 1.9 Spatial and temporal resolution of the main neurotechnologies. Acronyms explained within the text.

Nevertheless, new technologies that allow researchers to monitor, interpret, and alter the collective activity of vast neuronal populations (thousands or even millions of neurons) distributed across brain regions are required to improve our understanding of how the whole brain functions to drive thinking, behavior, cognition (perception, emotion, decision

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

making, consciousness, etc.), and the consequences of circuit malfunctions in patients with neurological disorders (schizophrenia, autism, AD, and PD). Large initiatives being conducted worldwide (Section 1.3) are attempting to advance the development of these technologies.

1.5 Data-Driven Computational Neuroscience

Hypothesis-driven research has been (and even still is) the standard approach used in the great majority of neuroscience projects. The null hypothesis must be clearly stated, data must be collected in a repeatable manner with a clear sampling design, and conclusions must be based on *p*-values (see Section 4.2). In a high percentage of cases, neuroscience research practices are still based on pre-Internet methods and employ the following steps: (a) experimental design, (b) data collection, (c) local storage of data, (d) inaccessible metadata, (e) data analysis using software installed on local computers, and (f) publication of a summary of the results.

The daily practice of brain science is only beginning to benefit from cloudification, a "**software as a service**" framework in which locally installed programs are replaced by web apps. **Cloud neuroscience** proposes that the data, code, and analytical results all live in the cloud together as a set of programs that run in a scalable manner and are accessible anywhere. Benefits of this approach include the simplification of global collaborations, the facilitation of open science, and the testing of a variety of models using the same data set, allowing neuroscientists to accelerate the discovery process. A crucial factor in this new view of neuroscience research is the so-called **data-driven** paradigm that amasses vast quantities of data for the automatic modeling of complex interactions between the brain and behavior, for example, and informs about the diagnosis and prevention of neurological disorders and psychiatric diseases. The access to anatomical, biochemical, connectivity, developmental, and gene expression (**ABCDE**) data will allow researchers to view the scientific process as a virtuous cycle, namely, a collective effort where each new experiment yields data, and after their analysis, new or refined models are created that suggest novel experiments and allow the cycle to be repeated if necessary (Neuro Cloud Consortium, 2016).

Neuroscience is becoming more data-centric, as increasing numbers of brain atlases, connectomes, and imaging data sets are being published. The number of projects that provide large data sets for testing a specific hypothesis and enabling data-intensive discovery is increasing (Akil et al., 2011). Three examples are described next. The first example is the Human Connectome Project[6] that takes advantage of high-throughput anatomical methods, such as resting-state functional magnetic resonance imaging (R-fMRI) for macroconnectome, or serial section electron microscopy for microconnectome. In the functional connectome, R-fMRI studies the brain at rest and reveals large-amplitude spontaneous low-frequency fluctuations in the fMRI signal that are temporally connected across functionally related areas and appear to show a universal architecture of positive and negative functional connections and interindividual variability (Biswal et al., 2010). The **1,000 Functional Connectomes Project** data set provides researchers

[6] www.humanconnectomeproject.org.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.1 Multiples of the unit byte for digital information

Unit of digital information	Previous unit	Bytes
1 kilobyte	1,024 bytes	10^3
1 megabyte	1,024 kilobytes	10^6
1 gigabyte	1,024 megabytes	10^9
1 terabyte	1,024 gigabytes	10^{12}
1 petabyte	1,024 terabytes	10^{15}
1 exabyte	1,024 petabytes	10^{18}
1 zettabyte	1,024 exabytes	10^{21}

the opportunity to simultaneously interrogate multiple functional circuits without the requirement for an a priori hypothesis. For the anatomical connectome (Lichtman et al., 2014), electron microscopy images at the nanometre level provide sufficient resolution to visualize the finest details of synaptic connections, showing all cells and all organelles. A substantial amount of data are generated from electron microscopy images. A single cubic millimeter of the rat cortex generates approximately 2 petabytes of data. See Table 1.1 for an idea of the amount of data that this number represents.

A complete visualization of the rat cortex requires an exabyte, which far exceeds the storage capability of any system that is currently available. A complete visualization of the human cortex, which is approximately 1,000 larger than the rodent cortex, will require a zettabyte to depict the anatomical features of 86 billion neurons communicating with each other via more than 250 trillion synapses. In terms of the time required to generate the data, electron microscopes can currently process several terabytes of data per hour, allowing a researcher to process a cubic millimeter of rodent brain in approximately 800 hours. A complete visualization of the mouse cortex will require at least a decade. In Yuste and Church (2014), the ability to monitor and optically control a large percentage of the 100,000 neurons in a fruit fly brain was predicted to be achieved in 2019, whereas these data will not be recorded in the mouse before 2024.

The second example is the **Neuroscience Information Framework** (NIF)[7] that provides access to more than 3,500 resources (data sets, tools, and materials), where some domains, such as electrophysiology and behavior, are underrepresented compared to genomics and neuroanatomy. The third example is the **Allen Institute for Brain Science**,[8] which maintains and curates more than 3,000 terabytes of data. However, the vast majority of data and metadata in neuroscience continue to remain inaccessible.

Akil et al. (2011) provide several recommendations for best practices in mining neuroscience data. (a) First, powerful tools must be developed to study the temporal and spatial changes in brain anatomy and activity. (b) The informatics infrastructure must be sufficiently flexible to incorporate new types of data. (c) Best practices for producing new neuroscience data must be defined. (d) A cultural shift in the field of neuroscience to allow data sharing is needed. (e) Community ontologies and identifiers are needed. (f) Data must be published in standardized table formats to facilitate data science. (g) Interdisciplinary research in fields such as computer science, machine learning, and visualization should

[7] neuinfo.org/.
[8] www.alleninstitute.org/.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

be fostered. (h) Educational strategies for the next generation of neuroscientists must be improved to ensure proficiency in data mining. These good practices will allow researchers to take advantage of the heterogeneity of neuroscience data with multiple spatial and temporal scales (Section 1.4) in all levels of neuronal organization: molecules (genotypes, protein interactions), cells (morphology and electrophysiology), cellular compartments (protein localization), brain regions, whole brain (functional and anatomical imaging), and organism (behavior) (French and Pavlidis, 2007) and to produce reproducible, collaborative, and large-scale neuroscience data needed for the twenty-first century. Ethical and legal issues must also be addressed (similar to the genome project).

1.5.1 Collecting Neuroscience Data

In addition to the obstacle of technology, the different ways laboratories work is another issue. Laboratories record different neurons in different brain regions from different animals performing different tasks. These disparate data are difficult to compare and combine. Modern neuroscience has shown a trend toward complementing the traditional, small laboratory culture by rewarding individual investigators with large, multidisciplinary teams using highly reproducible standards who are making all their methods, data, metadata, and software publicly available (Bouchard et al., 2016).

A survey by Tenopir et al. (2011) investigating **data-sharing** practices among scientists showed that approximately 50% of the respondents do not share data. However, 85% indicated an interest in having access to other researchers' data sets. In neuroscience, the culture of small laboratories that do not share data, metadata, or software is one of the causes of the **replication crisis**. The scandal highlighting bad scientific practices (Eklund et al., 2016) after the reanalysis of R-fMRI data from the public 1,000 Functional Connectomes Project may affect more than 3,000 published studies, almost all of which were funded by several national agencies. On the other hand, Wicherts et al. (2011) found that studies with accessible data tended to have fewer errors and more robust statistical and machine learning modeling approaches than studies where data sets were not available.

The culture shift in sharing data across laboratories is transforming "vertical" efforts, namely, applying single techniques to single problems in single species, into "horizontal" efforts, where the emphasis is placed on integrating data collected using a wide range of techniques (Sejnowski et al., 2014). These "horizontal" efforts will transform the current situation with many small models that encompass limited data sets and are more descriptive than explanatory. Sharing the **long-tail data** (Fergurson et al., 2014) is a way of obtaining large-scale data in neuroscience by merging small, granular data sets collected by individual laboratories in the course of day-to-day research. In the neurotrauma field, the traumatic brain injury (Maas et al., 2011) and spinal cord injury (Nielson et al., 2014) communities provide examples of the potential benefits of sharing long-tail neuroscience data.

Benefits of data sharing include (a) increased transparency and reproducibility of the results, (b) improving the research approach to employ the most recently developed experiments incorporating various research strategies, and (c) reduced economic costs derived from the lack of transparency and data inaccessibility. However, several reasons for the lack of motivation to share data also exist: (a) the competition to be the first to analyze the data set and to be recognized for publishing novel findings, (b) concerns regarding the

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

privacy of the human research participants, as the regulatory mechanisms for consent for use of data in the context of open access databases have not been completely established, (c) public–private partnerships involved in large-scale data projects can produce tensions derived from the different interests, and (d) problems derived from credit sharing in the academic community based on authorship status on publications. Leitner et al. (2016) provide a comparative study of citations of data published in neuroscience between papers sharing data sets and papers without this material, and the former achieve a significantly larger citation impact than the latter.

Examples of data-sharing initiatives in neuroscience include (a) **NeuroMorpho.Org** (Ascoli et al., 2017), which follows pragmatic recipes to ensure a win–win outcome for both researchers who are sharing and receiving data. The good practices for the roles of data curators are to (i) serve the end-users by developing a complement, rather than duplicating existing resources with a clear scientific need; (ii) adopt standard formats that maximize interoperability rather than proprietary formats; (iii) design intuitive ergonomics requiring only minimal instructions; (iv) solicit feedback from the users to improve functionality; (v) publish statistics on data access, downloads, and reuse; (vi) facilitate the contribution of data by assuming that conversion and standardization are the curators' job; (vii) use concise, consistent, and specific metadata annotation; (viii) publicly acknowledge the labs contributing data; and (ix) be patient and persistent in finding, requesting, and collating data, establish quality standards to maximize research utility, and diversify the experience of your team. (b) At the Allen Institute for Brain Science, large teams are generating complete, accurate and permanent resources for the mouse and human brain. The need for a highly specialized workforce that collaborates and submerges the ego to the needs of the group as a whole is a characteristic of team science at this institution, a condition for developing large-scale data and open science. Based on this philosophy, the scientific rewards are not in the promise of first or senior authorship, but in the participation in a historic mission at the frontier of science where new knowledge is generated to benefit all humans (Koch and Jones, 2016). (c) The national and international brain initiatives described in Section 1.3 are all also working within this data-sharing perspective.

The **people-powered research science** (PPRS) revolution is creating a global community of new "experts," proving that anyone with motivation and a computer is able to help scientists and accelerate scientific progress. PPRS is viewed as a kind of **citizen science** that is based on the Internet and has provided people who were previously excluded from academic science the opportunity to collect and generate data or to contribute as individuals or teams to analyze the data (Roskams and Popovic, 2016). Examples of PPRS in neuroscience include (a) **Eyewire** (Helmstaedter et al., 2013), where over 250,000 players from more than 140 different countries have been contributing to the first-ever 3D reconstructions of high-resolution networks of cells within the mouse retina since 2010; (b) **DREAM**, a crowd-sourcing approach in the form of challenges that have developed new models in the field of neuroscience based on machine learning to predict the future progression of ALS (Küffner et al., 2015); (c) **Mozak** (Roskams and Popovic, 2016), a gaming platform designed to rapidly accelerate our understanding of memory diversity by providing gradual training to become an expert neuron reconstructor; and (d) the **BigNeuron** project (Peng et al., 2015), a community effort that combines modern bioimaging informatics, recent improvements in labeling and microscopy, and the wide need for openness and standardization to provide a resource for the automated reconstruction of the

dendritic and axonal morphology of single neurons. While science has always benefited from standing on the shoulders of giants, the PPRS revolution will enable neuroscientists to stand on the shoulders of everyone.

1.5.2 Statistics and Machine Learning: The Two Cultures

This section discusses the similarities and differences between two disciplines, statistics and machine learning, whose objective is to analyze data by transforming them into computational models from which new knowledge and predictions are obtained.

1.5.2.1 Statistics

During the second half of the twentieth century, the nature of statistical practice changed with the advent of the computer. The progress in computer technology led to changes in statistical methodology and statistical ideas. Examples include the inversion of large matrices, iterative methods for the estimation of parameters, and the introduction of various types of resampling methods, such as bootstrapping, dynamic interactive visualization of data sets, and Bayesian statistics, a field that would not have been developed without the assistance of computers. Statistics was regarded as a mathematical discipline; however, it is currently regarded as a computational discipline focused on understanding the scientific problem and providing a correct interpretation (Hand, 2015).

The advent of computers also led to dramatic changes in statistical practice. The very easy use of software packages, particularly by statistically uninformed people, created risks of its own. This easy use has also contributed to trying many possible analyses instead of previously determining the proper method. An effective statistical analysis depends critically on understanding the scientific question, and thus automatic or rote strategies impose high risks. With the move to electronic rather than manual data collection, the dangers of multiple testing and overfitting became even greater.

The dominant statistical paradigm used in neuroscience studies is the **null hypothesis significance testing** (NHST), a hypothesis-driven approach (Section 4.2), in contrast with the data-driven approach advocated in this book (Section 1.5). The use of NHST in neuroscience must solve problems derived from the use of extremely small sample sizes and the so-called **pseudo-replication** that appears when non-independent samples are analyzed.

Nature journals published a set of guidelines for the correct reporting of statistical analyses that was presented by the *Nature of Neuroscience* journal (Nature Neuroscience Editorial, 2005). These guidelines are listed below. (a) Consult statistical experts to help design experiments and analyses, preferably before the data are collected. (b) Summarize all data sets with descriptive statistics before further analyses are performed. (c) Ensure that the statistical evidence is clearly described, providing information about what tests were used, how many samples were analyzed, the types of comparisons that were performed, the significance level found, etc. (d) Justify the choice of the analysis and confirm that the data conform to the assumptions underlying the tests (for example, most parametric tests require the data to be normally distributed). (e) Avoid the risk of false-positive results by using multiple comparisons tests (this situation is typically used for functional imaging data when multiple voxels are compared). Six years after these guidelines were published, Nieuwenhuis et al. (2011) found that in 157 of 513 behavioral, systems, and cognitive neuroscience articles published in *Science*, *Nature*, *Nature Neuroscience*, *Neuron*, and *The*

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Journal of Neuroscience during 2009 and 2010, the authors compared 2 experimental effects. In 79 papers, the authors used an incorrect procedure to compare the 2 effects. The most common error was to report the difference between their significance levels instead of the statistical significance (Section 4.2) of their differences. When the same study was performed on 120 cellular and molecular neuroscience articles published in *Nature Neuroscience* during the same 2 years, no single study used the correct statistical procedure to compare effect sizes.

These facts highlight the need for solid statistical training for neuroscience researchers, both in classical and novel statistical methods for data acquisition and analysis.[9] Each of the methods used to collect neural data from human and animal subjects, such as neuroimaging (radiography, fMRI, MEG, and PET), electrophysiology from multiples electrodes (EEG, ECoG, and spike trains), calcium imaging, optogenetics, and anatomical methods (diffusion imaging, electron microscopy, and fluorescent microscopy), produces data with its own set of statistical and analytical challenges.

Part II of this book contains three chapters devoted to the introduction of statistical methodology. Chapter 2 introduces several graphical and numerical representations of univariate, bivariate, or multivariate data, such as pie charts, barplots, histograms, summary statistics and principal component analysis. Chapter 3 presents probability theory and some of the most common univariate and multivariate distributions in both discrete and continuous domains. This chapter also describes methods for simulating random variables and basic concepts of information theory. Parameter estimation and hypothesis tests are presented in Chapter 4.

1.5.2.2 Machine Learning

We are living in an era of abundant data, and tools for searching, visualizing, modeling, and understanding large data sets are needed. These tools should be able to (a) faithfully capture the intrinsic uncertainty of the domain, (b) induce models from data in an automated and adaptive manner, (c) exhibit robustness against noisy and imprecise data, and (d) scale well to large data sets. Machine learning methods incorporate these characteristics by defining a space for possible models and developing learning-from-data procedures of model parameters and structures. Machine learning provides tools for extracting reliable and meaningful relationships and for generating accurate predictions and reliable decisions (Hinton, 2011); this approach is likely to be one of the most transformative technologies of the twenty-first century. Currently, humanity has a new way of deriving knowledge apart from evolution, experience, and culture, namely, from machines that are able to learn automatically.

Machine learning is the field in which researchers build computers that improve automatically through experience and is also viewed as systems that learn from data. Machine learning is considered an interdisciplinary field focusing on both the mathematical foundations and practical applications with connections to pattern recognition, data mining, adaptive control, statistical modeling, data analytics, data science, and artificial intelligence. Its applications cover a wide spectrum of topics, e.g., automatic speech recognition, computer vision (object, face, and handwriting recognition), information retrieval and web

[9] As far back as 1938, Wells (1938) wrote, "... a certain elementary training in statistical method is becoming as necessary for anyone living in this world of today as reading and writing ..."

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

searches, financial prediction and automated trading, Industry 4.0, sport analytics, medical diagnosis, personalized medicine, bioinformatics, neuroscience, etc.

Conceptually, machine learning algorithms are viewed as searching through a large space of candidate models of different types to identify the model that optimizes the previously established performance metric (Jordan and Mitchell, 2015). Machine learning algorithms vary (a) in the way they represent candidate models (naive Bayes, classification trees, logistic regression, finite-mixture models, etc.) and (b) in the way in which they search through this space of models (exact or heuristic optimization algorithms). A key scientific and practical goal is to theoretically characterize the capabilities of specific learning algorithms. First, obtain the shape of the decision boundary to discriminate between positive and negative instances (in a binary classification problem). Second, characterize the sample complexity (the amount of data required to learn accurately) and computational complexity (how much computation is required), and how both (sample and computational) depend on features of the learning algorithm.

The most widely used machine learning methods are **supervised learning** methods. Starting from a collection of $(\mathbf{x}, \mathbf{c})$ pairs, the goal is to produce a prediction $\mathbf{c}^*$ in response to a query $\mathbf{x}^*$. Different types of output $\mathbf{c}$ have been studied: (a) the simple **binary classification** problem, where $\mathbf{c}$ is a 1D vector, c, with two possible values; (b) **multi-class classification**, where c has R possible labels; (c) **multi-label classification**, where $\mathbf{c}$ is a d-dimensional vector that simultaneously adopts several of the d labels; and (d) general **structured prediction**, where $\mathbf{c}$ is a combinatorial object whose components may be required to satisfy some set of constraints. Supervised learning also includes cases in which $\mathbf{c}$ has real valued components (regression or multi-output regression problems) that are not covered in this book. Many forms of modeling (providing mappings from $\mathbf{x}^*$ to $\mathbf{c}^*$) exist: non-probabilistic classifiers (Chapter 7), such as k-nearest neighbors, classification trees, rule induction, artificial neural networks and support vector machines, and probabilistic classifiers (Chapter 8), such as discriminant analysis, logistic regression analyses, and Bayesian network classifiers.

Unsupervised learning or clustering is defined as the problem of determining the partitioning or grouping of similar data in the absence of explicit labels, as $\mathbf{c}$. Non-probabilistic clustering (Chapter 11) includes the topics of hierarchical and partitional clustering, whereas probabilistic clustering (Chapter 12) is mainly based on finite-mixture models and the expectation-maximization algorithm.

Domingos (2015) considers five groups of machine learning approaches with respect to how the machines should learn while extracting the maximum possible knowledge from the data: (a) the **symbolic approach** based on logic and philosophy, (b) the **connectionist approach**, with foundations in the way the brain functions using neuroscience as the basic theory and the backpropagation (of artificial neural networks) as the main algorithm, (c) the **evolutionist approach** based on evolutionary biology and using evolutionary computation (i.e., genetic algorithms) as its main algorithm, (d) the **Bayesian approach** that uses statistics and probabilistic inference as fundamentals and is based on the idea of adapting the world interpretation as new evidence arrives, and (e) the **analogist approach**, which searches in its memory for similar situations that worked properly in the past to solve the current problem, and k-nearest neighbors is its favorite algorithm.

Machine learning is a young, continuously expanding discipline in which new methods and algorithms are being developed daily, mainly within the new paradigm of

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

big data.[10] Large data sets require scalable algorithms whose time and space requirements are linearly or sublinearly correlated with the size of the problem (number of points or number of dimensions). Several opportunities and challenges in machine learning remain: (a) contrasting current machine learning approaches to naturally occurring systems (humans and animals) that in some situations can discover patterns from only one instance, something called one-shot learning in the jargon of artificial intelligence, (b) organizing different types of skills and types of knowledge (supervised and unsupervised) into a single-to-more-difficult sequence, (c) constructing lifelong or never-ending computer learners that continuously operate for years, (d) team-based learning inspired by people who often work on teams to collect and analyze data, and (e) mixed-initiative learning that provides new machine learning methods the ability to work collaboratively with humans.

Predictions made by machine learning systems (or artificial intelligence in general) must be accurate. In addition, humans should understand the whys of the recommended decisions. For example, if a system selects Miss Smith for a very risky neurological surgery, or if another system predicts that Mr. Jones's cognitive deterioration will be aggravated in the next 2 years, understanding what those decisions are based on is important.

Hence the need to develop transparent, reliable, trustable, and explainable models avoiding the black boxes implicit in some machine learning paradigms. Among the paradigms presented in this book, classification trees, rule induction, Bayesian classifiers, Bayesian networks, and Markov networks stand out for their transparency, whereas random forests, support vector machines, artificial neural networks, and deep learning excel for their opacity and difficult interpretation. **Explainable artificial intelligence** is a trend that tries to develop systems that are interpretable for humans (Rudin, 2019).

Ethics is another aspect to take into account when developing intelligent systems in order to prevent them from having associated prejudices. Such prejudices can come either from the individual who has developed the system (and has conscious or unconscious preferences) or from the biases (for example, gender, race, age) implicit in the data set on which the automatic learning algorithm is applied, with problems both in the gathering or usage of data. Eliminating harmful biases is essential.

1.5.2.3 Statistics versus Machine Learning

This book is primarily concerned with the use of statistical methods and machine learning algorithms for transforming neuroscience data into computational models that are able to provide appropriate solutions for supervised and unsupervised classification problems (also called classification and clustering, respectively), as well as for discovering associations among the variables describing a problem. Statistics and machine learning are viewed as two different cultures for drawing conclusions from data (Breiman, 2001b). In both cultures, the input variables, $\mathbf{X}$, also known as predictor (machine learning) or independent (statistics) variables, are mapped to the output or response variables, $\mathbf{C}$, as illustrated in Figure 1.10.

[10] Four *v*'s define the main issues of big data: (i) volume of the data to be processed, requiring different storage and processing capabilities than traditionally; (ii) velocity, referring to the speed with which data is generated; (iii) variety of sources of the data to be processed, including structured, semi-structured, and unstructured data; and (iv) veracity, which refers to the good quality of the data being analyzed.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Figure 1.10 Transforming input variables, **X**, into output variables, **C**, through a process enacted by Nature.

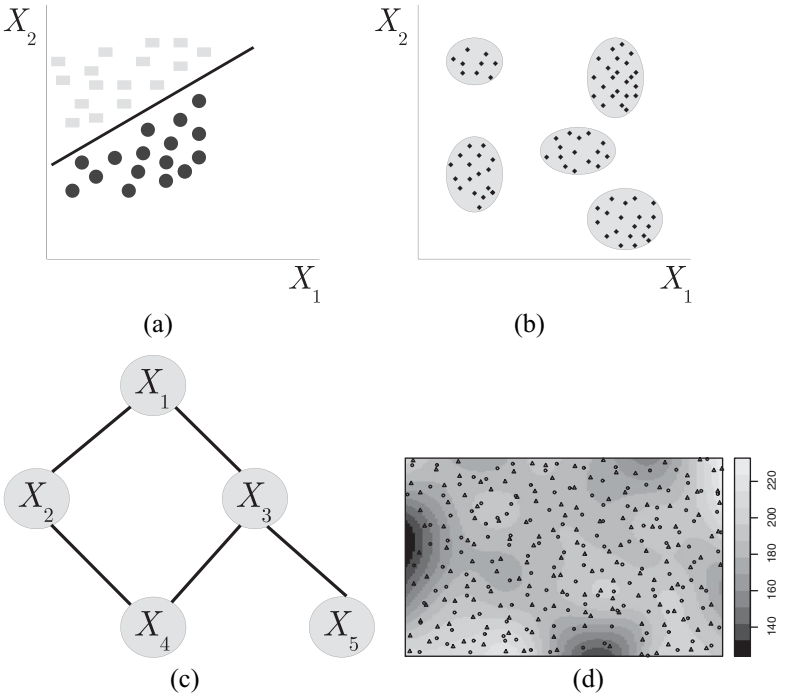


Figure 1.11 (a) Supervised classification (Part III). (b) Clustering (Part IV). (c) Discovery of associations (Part V). (d) Spatial statistics (Part VI).

In this book, we consider four main data analysis goals:

- **Prediction**: forecast the responses for future input variables. We consider discrete, both 1D and multidimensional, output variables.
- **Clustering**: merge similar inputs into groups.
- **Discovery of associations**: represent the relationships among the variables.
- **Spatial statistics**: study of the 2D/3D arrangement of points.

Figure 1.11 shows examples of each of these goals, which are developed throughout Parts III, IV, V, and VI of this book, respectively. All chapters in Part III assume that the output is unidimensional, c, except for Chapter 10, which considers a multidimensional output response, **c**. Chapters 11 and 12 in Part IV include situations where the output variable is hidden and should be determined by the model. Part V of the book, which is developed in Chapters 13 and 14, does not necessarily consider the distinction between input **x**, and output **c** and aims to discover the relationships among all variables. Chapter 15 in Part V includes exploratory data analysis and statistical modeling, as well as the simulation of spatial point processes.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

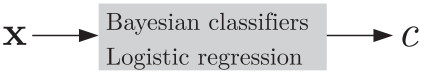


Figure 1.12 Examples of stochastic data models for prediction (supervised classification).

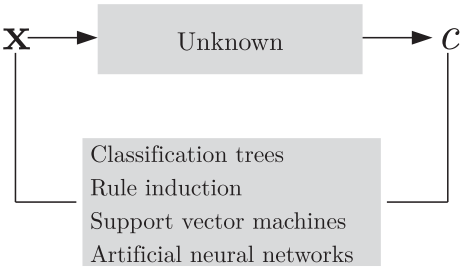


Figure 1.13 Examples of algorithm-based models for prediction (supervised classification).

The statistical methodology for modeling data is based on probability theory and assumes that a stochastic data model is responsible for mapping $\mathbf{x}$ to $\mathbf{c}$. In this approach, data analysis is based on the assumption of a stochastic data model for the inside of the box of Figure 1.10. A common assumption is that data are generated by independent draws from $\mathbf{c} = f(\mathbf{x}; \theta)$, where f is the mapping function specified by the model, and θ represents the parameters of the model (which are estimated from the data). Examples of stochastic data models for supervised classification are Bayesian classifiers and logistic regression, as illustrated in Figure 1.12.

Machine learning modeling considers the inside of the box to be complex and unknown. The approach is to find a mapping function $f(\mathbf{x})$, an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{c}$. Examples of algorithm-based models are classification trees, rules, support vector machines, and artificial neural networks, as shown in Figure 1.13.

Next we discuss the differences in some important issues between statistical and machine learning approaches. Table 1.2 presents some of these differences.

Model Assumptions. Statistical modeling requires a solid mathematical background and is based on its purest form on probability and measure theories, decision theory, Markov chains, asymptotic analysis, etc. The dependency of stochastic models on strong assumptions (e.g., multivariate Gaussian density, homoscedasticity, etc.) is justified by their mathematical malleability. In almost all real-world situations, however, these assumptions are not verified. In contrast, machine learning algorithms do not assume that conditions are met in the data set.

Model Selection. In statistics, the most appropriate model is selected as the model that best fits the data in terms of a measure related to the (penalized or marginal) likelihood. This process is usually enacted using a likelihood ratio-based hypothesis test comparing the fit between the current model and a candidate model. The candidate model will replace the current model only if the differences in likelihood are statistically significant. Search approaches include standard **forward selection** (starting from the empty model and incorporating the most informative variable at each step until there is no further improvement), **backward elimination** (starting from a model with all variables and deleting the worst variable at any time), and **stepwise selection** (where the inclusion and deletion stages are intermingled).

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.2 Main differences between statistics and machine learning approaches to data modeling

Statistics	Machine learning
Stochastic data model	Algorithm modeling
Model selection	Structure and parameter learning
Fitting	Learning
Likelihood ratio	Predictive accuracy
Forward, backward, stepwise	Metaheuristic
Collinearity	Feature subset selection
Bayesian approaches	Ensembles
Probabilistic output	Deterministic output

Modeling in the machine learning field relies on two ideas: (a) using a score measuring the goodness of the proposed model that is more directly related to the final aim; for example, if the main goal is prediction, the estimated classification accuracy, the F_1-measure, or the area under the ROC curve (see Section 5.2 for details) are score candidates; and (b) searching in the space of possible models using more sophisticated and intelligent metaheuristics, such as simulated annealing, tabu search, and genetic algorithms (Section 6.3.1). The use of these metaheuristics that require intensive computational algorithms is justified by the usually large cardinality (sometimes more than exponential in the number of variables) of the space of models.

In any case, both cultures agree on the principle of **Occam's razor** (Thorburn, 1915) or law of parsimony, that is, simpler models are better.

Feature Subset Selection. The term "curse of dimensionality" introduced by Bellman (1957) is associated with the need to find a small number of input variables in prediction models that contain most of the information required for the output variables. In statistics, these variables are selected based on the concept of collinearity and assumes that a previously fixed number of variables k of the total number of predictor variables n, $k < n$, should be chosen. The machine learning approach for feature subset selection is more computational and tries to find the best combination of input variables (Section 6.3) for the output variable in terms of predictive accuracy. From this perspective, the search is conducted in a space whose cardinality is given by 2^n.

More than One Model. Breiman (2001b) referred to the situation where different models with approximately the same predictive accuracy are produced, which is known as the Rashomon effect.[11] This effect is typically when the initial number of input variables is very large. Instead of choosing one of the models and discarding the rest, one possible alternative is to merge or combine the outputs of the different models, particularly if the models belong to different families and the errors are committed in different cases. In machine learning, this approach is known as **ensemble learning** (Chapter 9) and is considered a practical implementation of the full (or selective) Bayesian approaches to models developed in statistics.

Validation. The honest estimation of the generalization power of the predictive models (see Section 5.3) is conducted in a similar way in both cultures. Both the statistics and machine learning communities partition the data set into training and test sets, or use

[11] This effect takes its name from Akira Kurosawa's film *Rashomon*, in which a crime witnessed by four individuals is described in four mutually contradictory ways.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.3 Statistical and machine learning methods described in this book

Statistics	Machine learning
Feature selection (filter)	Feature selection (wrapper)
k-nearest neighbors	*k*-nearest neighbors
Classification trees	Classification trees
Discriminant analysis	Rule induction
Logistic regression	Artificial neural networks
Bayesian network classifiers	Support vector machines
Multidimensional classification	Metaclassifiers
Hierarchical clustering	Multi-label classification
Partitional clustering	
Probabilistic clustering	
Spatial statistics	

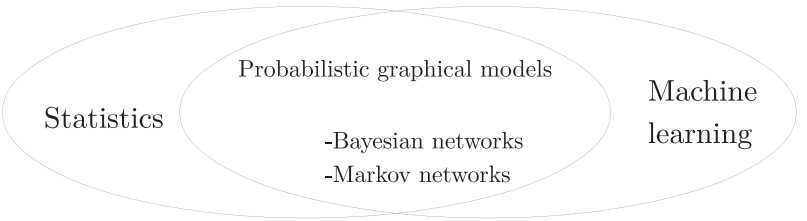


Figure 1.14 Probabilistic graphical models including Bayesian networks and Markov networks at the intersection between statistics and machine learning.

more sophisticated methods, including *k*-fold cross-validation, and bootstrapping. The probabilistic output provided by statistical models, which is richer than the deterministic response produced by almost all machine learning algorithms, indicates that statistical models can be validated with better measures, such as the Brier score.

Both the statistics and machine learning cultures are complementary rather than contradictory. Indeed, many machine learning algorithms incorporate statistics in their process, and some modern statistical modeling techniques have been developed by the machine learning community. An example is the probabilistic graphical models represented by Bayesian networks (Chapter 13) and Markov networks (Chapter 14) (Koller and Friedman, 2009), as illustrated in Figure 1.14.

Table 1.3 contains a list of other methods and algorithms covered in this book and organized as belonging to statistics or machine learning. Note that this organization is somewhat ambiguous, as *k*-nearest neighbors and classification trees were developed independently in both fields and constitutes an example of cooperation. *k*-nearest neighbor methods (Section 7.1) date back to the early 1950s (Fix and Hodges, 1951) and have been thoroughly investigated by the statistical community. Very similar algorithms, which are known as instance-based learning (Aha et al., 1991) and case-based reasoning (Kolodner, 1993), have been developed by the machine learning community. Classification trees (Section 7.2) were independently developed. They are designated as classification and regression trees (CART) in the statistics community (Breiman et al., 1984) and decision trees in the machine learning field (Quinlan, 1986). Seminal books comparing approaches from both cultures are Weiss and Kulikowski (1991) and Michie et al. (1994).

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

1.5.3 Brain-Inspired Machine Learning Methods and Hardware

The principles of the functions of the brain and the computer differ. The brain is able to work with novelty, complexity, and ambiguity, whereas computers are fast and very precise. However, the knowledge about the brain can serve as inspiration in the design of machines, both software and hardware, for example by applying the lessons learned from connectome graphs to making computers smarter.

In software, strategies used to unravel the algorithmic specializations of the sensory cerebral cortex are inspiring the next generation of high-performance machine learning. Brains exhibit a remarkable capacity for recognizing and learning about physical and abstract data that far exceeds the capabilities of the currently available state-of-the-art machine learning systems. A performance gap exists not only for high-level cognitive processes (e.g., understanding) but also for basic sensor information-processing tasks supporting these higher-level functions. Contemporary theories of cortical computing suggest that for sensory information-processing tasks, the brain employs algorithms with a limited set of computing primitives. These structural and/or functional motifs are used by one or more cortical area(s) to implement "core functions" of cortical algorithms that represent, transform data, and learn from data. A program of the American **IARPA** called **Machine Intelligence from Cortical Networks** (MICrONS) proposed in 2014 seeks to revolutionize machine learning by reverse-engineering the algorithms of the brain. MICrONS aims to design and implement novel machine learning algorithms that use the same "core functions" employed by the brain. Algorithms derived from the visual cortex will be tested on visual scenes, algorithms derived from olfactory cortex will be tested on olfactory cues, etc.

In hardware, **neuromorphic computing** takes its inspiration from observations of the complexity of the biological brain and considers brain knowledge as principles that can be applied to the design of hardware engineering systems. Seminal work on bio-inspired microelectronics culminated in the book published by Mead (1989). More recent examples of large-scale neuromorphic systems include (a) the **IBM TrueNorth** chip (Merolla et al., 2014), which is the outcome of a decade of work by the DARPA SYNAPSE program aimed at delivering a very dense, energy-efficient platform capable of supporting a range of cognitive applications, (b) **Neurogrid** (Benjamin et al., 2014), which is based on the heritage of Mead (1989), and uses subthreshold analogue circuits to model neuron and synapse dynamics in biological real time, (c) the **SpiNNAker** project (Furber et al., 2014), which has developed a massively parallel digital computer whose communication infrastructure is motivated by the objective of modeling large-scale spiking neural networks with connectivity similar to the brain in biological real time, and (d) the **BrainScaleS** neuromorphic system (Schemmel et al., 2010) developed at the University of Heidelberg with the aim of implementing physical models of neuronal processes.

1.6 Real Examples Discussed in This Book

The book is organized into the following six parts: I Introduction (Chapter 1), II Statistics (Chapters 2–4), III Supervised Classification (Chapters 5–10), IV Unsupervised Classifica-

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.4 Book organization

Chapter	Title	Data Set
2	Exploratory data analysis	1. Interneurons vs. pyramidal neurons
3	Probability theory and random variables	1. Interneurons vs. pyramidal neurons
4	Probabilistic inference	1. Interneurons vs. pyramidal neurons
6	Feature subset selection	2. GABAergic interneuron nomenclature
7	Non-probabilistic classifiers	1. Interneurons vs. pyramidal neurons
8	Probabilistic classifiers	1. Interneurons vs. pyramidal neurons
9	Metaclassifiers	1. Interneurons vs. pyramidal neurons
10	Multidimensional classifiers	3. Quality of life in Parkinson's disease
11	Non-probabilistic clustering	4. Dendritic spines
12	Probabilistic clustering	4. Dendritic spines
13	Bayesian networks	5. Basal dendritic trees
14	Markov networks	6. Brain connectivity
15	Spatial statistics	7. Spatial location of synapses in the neocortex

tion (Chapters 11–12), V Probabilistic Graphical Models (Chapters 13–14), and VI Spatial Statistics (Chapter 15). Some real-world neuroscience examples are used to illustrate the methods described in each chapter in Parts III to VI. Seven data sets were used for these illustrations, and they are described below, see Table 1.4.

1.6.1 Data Set 1: Interneurons versus Pyramidal Neurons

Discerning different neuronal cell types is an essential first step toward understanding neural circuits. Classical classifications of neuronal cell types used qualitative descriptors (de Nó, 1922), with nomenclature varying among researchers. Quantitative classifications using both unsupervised and (less frequently) supervised classification methods are more recent. They are actually necessary to obtain an objective set of descriptors for each cell type that most investigators agree upon.

Neocortical GABAergic interneurons are particularly difficult to distinguish and will be the main issue analyzed in Data Set 2 below. First, we will try to solve an easier problem: automatically distinguishing pyramidal cells from interneurons in the mouse neocortex based solely on their morphological features. These cells are the two principal neuronal types in the cerebral cortex (Ramón y Cajal, 1899), see Figure 1.15.

Obviously, the "ground truth," given by the presence or absence of an apical dendrite, is used to reliably label each cell as a pyramidal neuron `P` or interneuron `I`, but it is not included in the morphological features.

The rows in the data set contain 327 cells composed of 199 interneurons (60.86%) and 128 pyramidal cells (39.14%) from the cortex of PND 14 C57/B6 mice. All pyramidal neurons had clear apical dendrites. Many different subtypes of interneurons were identified and collected from several different laboratory studies. A complete description of methods used to prepare the brain slices and the histological procedures is provided in Guerra et al. (2011). Neuronal morphologies were reconstructed using Neurolucida (MicroBrightField [Glaser and Glaser, 1990]). The columns in the data set show 64 morphological features of

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

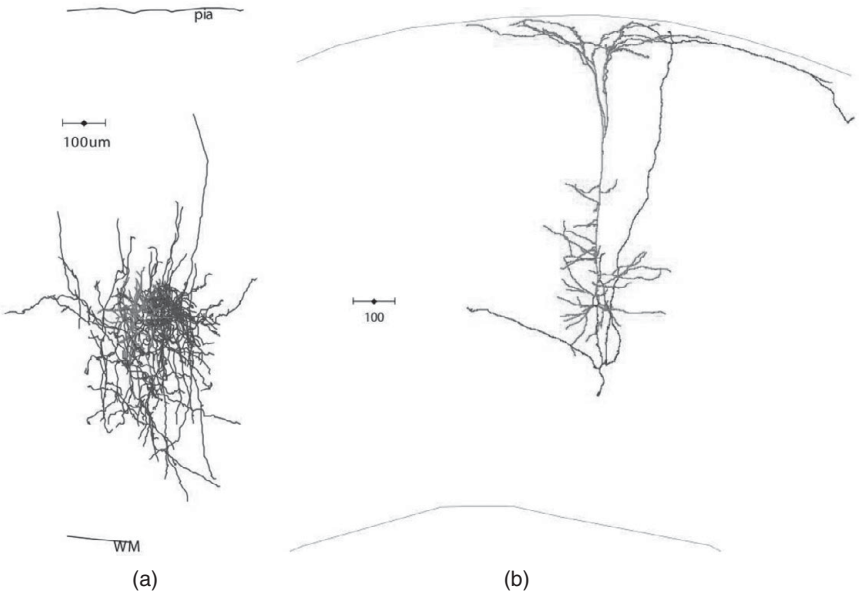


Figure 1.15 (a) Basket cell (interneuron). (b) Pyramidal cell. Both cells are located in the mouse neocortex, and their axonal arbor and their dendritic tree are shown in blue and red, respectively. Reprinted with permission from Guerra et al. (2011). For the color version, please refer to the plate section.

the reconstruction of each cell measured with the Neurolucida Explorer program, as well as the relative distance from the soma to the pia. Table 1.5 (somatic and axonal features) and Table 1.6 (dendritic features) list all 65 variables.

Some features, such as the somatic area and perimeter, number of axons and dendrites, axonal and dendritic length, axonal and dendritic branch angles, and number of axonal and dendritic nodes (branch points), are measured directly. Other features, such as the soma roundness, axonal and dendritic Sholl lengths (Sholl, 1953), convex hull analysis, and fractal analysis, are computed. Sholl length measures the radial distribution of the lengths of axonal or dendritic arbors around the soma. Concentric spheres centered on the soma are drawn at radius intervals of r μm. $r = 100$ μm for axons and $r = 50$ μm for dendrites. Then, the Sholl length at r μm is calculated as a fraction of the length of (axonal or dendritic) segments contained in the first Sholl ring divided by the total length of (axonal or dendritic) segments. The Sholl length at $2r$ μm is calculated as a fraction of the length of (axonal or dendritic) segments contained in the Sholl section from r to $2r$ (belonging to the second Sholl) divided by the total length of (axonal or dendritic) segments, and so forth. The convex hull analysis draws a 2D and 3D convex shape around the axons or dendrites. The area and perimeter of the 2D shape and the volume and surface area of the 3D shape are then calculated. The fractal analysis calculates the fractal dimension of the axon or dendrites using a linear regression analysis and therefore measures the extent to which the axonal or dendritic arbor fills the space. The relative distance from the soma to the pia is the ratio of the straight-line distance from the soma to the pia and the straight-line distance from the white matter to the pia. Thus, a value close to 0 (resp. 1) corresponds to a soma in a superficial (resp. deep) layer.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.5 Somatic and axonal features in Data Set 1

Feature	Description
Soma	
X_1	Somatic perimeter (μm)
X_2	Somatic area (μm^2)
X_3	Somatic aspect ratio
X_4	Somatic compactness
X_5	Somatic form factor
X_6	Somatic roundness
Axon	
X_7	Axonal node total (branching points)
X_8	Total axonal length (μm)
X_9	Total surface area of axon (μm^2)
X_{10}	Ratio of axonal length to surface area
X_{11}	Highest order axon segment
X_{12}	Axonal torsion ratio
X_{13}	Axonal planar angle ave
X_{14}	Axonal planar angle stdv
X_{15}	Axonal local angle ave
X_{16}	Axonal local angle stdv
X_{17}	Axonal spline angle ave
X_{18}	Axonal spline angle stdv
X_{19}	Ave tortuosity of axonal segments
X_{20}	Stdv of tortuosity of axonal segments
X_{21}	Axonal segment length ave
X_{22}	Axonal segment length stdv
X_{23}	Ave tortuosity of axonal nodes
X_{24}	Stdv of tortuosity of axonal nodes
X_{25}	Number axonal Sholl sections
X_{26}	Axonal Sholl length at 100 μm (fraction)
X_{27}	Axonal Sholl length at 200 μm (fraction)
X_{28}	Axonal Sholl length at 300 μm (fraction)
X_{29}	Axonal length density2
X_{30}	Axonal node density2
X_{31}	Convex hull axon area
X_{32}	Convex hull axon perimeter
X_{33}	Convex hull axon volume
X_{34}	Convex hull axon surface area
X_{35}	k-dim (fractal analysis)-axon

We are searching for a procedure that automatically distinguishes pyramidal cells from interneurons using this type of data, with cells characterized based on some morphological features and prior supervised information about the type of neuron (`P` or `I`).

An interesting issue is to identify which morphological features help distinguish the two neuron types i.e., to perform feature subset selection (Chapter 6).

Chapters 7 to 9 will show the performance of a battery of different supervised classification algorithms applied to this example. Note that procedures that omit this prior information aim to discover new or confirm some known hypotheses about subtypes of cells. This problem is tackled by unsupervised classification methods, which are described in

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.6 Dendritic features in Data Set 1; the relative distance to the pia is included at the bottom

Feature	Description
Dendrites	
X_{36}	Number of dendrites
X_{37}	Dendritic node total (branching points)
X_{38}	Total dendritic length (μm)
X_{39}	Ave length of dendrites (μm)
X_{40}	Total surface area of dendrites (μm^2)
X_{41}	Ratio of dendritic length to surface area
X_{42}	Highest order dendritic segment
X_{43}	Dendritic torsion ratio
X_{44}	Dendritic planar angle ave
X_{45}	Dendritic planar angle stdv
X_{46}	Dendritic local angle ave
X_{47}	Dendritic local angle stdv
X_{48}	Dendritic spline angle ave
X_{49}	Dendritic spline angle stdv
X_{50}	Ave tortuosity of dendritic segments
X_{51}	Stdv of tortuosity of dendritic segments
X_{52}	Dendritic segment length ave
X_{53}	Dendritic segment length stdv
X_{54}	Ave tortuosity of dendritic nodes
X_{55}	Stdv of tortuosity of dendritic nodes
X_{56}	Number of dendritic Sholl sections
X_{57}	Dendritic Sholl length at 50 μm (fraction)
X_{58}	Dendritic Sholl length at 100 μm (fraction)
X_{59}	Dendritic Sholl length at 150 μm (fraction)
X_{60}	Convex hull dendrite area
X_{61}	Convex hull dendrite perimeter
X_{62}	Convex hull dendrite volume
X_{63}	Convex hull dendrite surface area
X_{64}	k-dim (fractal analysis)-dendrites
X_{65}	Relative distance to pia

Chapters 11 and 12. Surprisingly, this unsupervised approach is used by most researchers to classify cortical neurons based on morphological, physiological and/or molecular features (Cauli et al., 2000; Kozloski et al., 2001; Wong et al., 2002; Tsiola et al., 2003; Andjelic et al., 2009; Helmstaedter et al., 2009a, 2009b; Karagiannis et al., 2009; McGarry et al., 2010; Battaglia et al., 2013; Helm et al., 2013; Perrenoud et al., 2013).

In addition to morphology, Ascoli et al. (2008) and Yuste et al. (2020) suggested a multimodal neuronal type definition, including physiological, molecular, and morphological features.

1.6.2 Data Set 2: GABAergic Interneuron Nomenclature

The data for this example have been borrowed from a paper by DeFelipe et al. (2013) that develops a methodology based on a new community-based strategy (crowd sourcing)

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

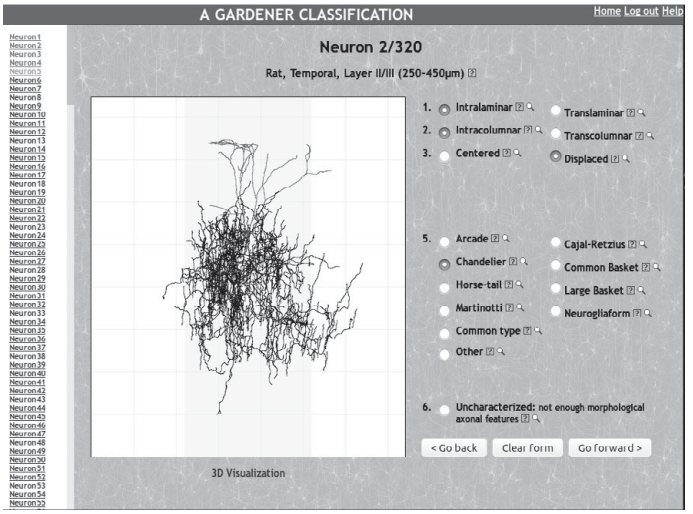


Figure 1.16 Screenshot of 1 of the 320 neurons included in the web-based interactive system. For each of the six class variables, the experts can select the most appropriate category describing the morphology of the neuron. For the color version, please refer to the plate section.

applied by a set of neuroanatomy experts with the objective of creating an accepted nomenclature for cortical GABAergic interneurons. The new methodology involves a web-based interactive system that enables experts to classify neurons with predetermined criteria (see Figure 1.16). Each expert has the option of classifying each neuron according to six class variables based on neuronal morphology. The first of these class variables refers to the geometric position of the neuron axonal arbor relative to cortical layers and includes the following categories: `intralaminar`, interneurons with axonal arbors that are mainly located in the layer of the parent soma, and `translaminar`, otherwise. The second class variable refers to the distribution of the axonal arbor relative to the size of cortical "columns" and covers two categories: `intracolumnar`, when the distance of the axonal arbors from the soma is not more than 300 µm in the horizontal dimension; and `transcolumnar`, if the neuron does not meet this constraint. The third class variable corresponds to the relative locations of the axonal and dendritic arbors and also includes two categories: `centered`, when the dendritic arbor is mainly located at the center of the axonal arborization, and `displaced`, otherwise. A fourth class is considered for interneurons categorized as translaminar and displaced. This variable has three possible categories: `ascending`, when the axonal arbor is mainly distributed toward the cortical surface; `descending`, when the distribution is mainly toward the white matter; and `both`, for neurons with axonal arbors distributed toward both the cortical surface and the white matter. The fifth class variable covers the common names of the cell types present in the literature (Jones and Peters, 1984): `arcade`, `Cajal-Retzius`, `chandelier`, `common basket`, `common type`, `horse-tail`, `large basket`, `Martinotti`, `neurogliaform`, and `other`. Finally, the sixth class variable examines whether a sufficient (or insufficient) number of morphological axonal characteristics is available to classify the interneuron and considers two categories: `characterized` and `uncharacterized`.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.7 Data set containing information for the 241 3D reconstructed interneurons. For each neuron, 2,886 morphological variables have been recorded, as well as the categorization provided by the 42 neuroanatomy experts into the 6 class variables, C_1 to C_6

				C_1			⋯	C_6		
Neuron	X_1	⋯	$X_{2,886}$	E1	⋯	E42	⋯	E1	⋯	E42
1	10.8	⋯	20.3	1	⋯	1	⋯	1	⋯	1
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
241	9.2	⋯	18.9	2	⋯	2	⋯	1	⋯	2

Forty-two neuroanatomy experts from different labs in Europe, the United States, Japan, and China participated in the experiment, providing their categorization for the 6 class variables. Branched structure, convex hull, Sholl, fractal, fan-in diagram, vertex, and branch angle morphological analyses were performed on each of the 241 neurons whose 3D reconstructions were available at NeuroMorpho.Org using the MicroBrightField Neurolucida package.

The initial data set (see Table 1.7) contains the 2,886 morphological variables described below, and the categorization provided by each of the 42 experts (E1, . . . , E42) according to the 6 class variables described in Figure 1.16 for each of the 241 interneurons.

In the original data set, in the Convex-Hull-2D block (see below), the number of intersections was measured in concentric spheres centered at the soma with increasing radii of 20 μm. Based on the advice of the neuroanatomists, this small radius was extended to 60 μm, resulting in a reduction of 300 variables. Thus, 2,586 final morphological variables were considered over which Chapter 6 will perform feature subset selection. The variables are organized into the following blocks:

- Box-Counting-Trees-Axons. The fractal dimension of the axon using the box-counting method (Mandelbrot, 1982). The fractal dimension is a quantity that indicates how completely the neuron fills the space. This value is measured by considering only the axonal arbor. The block contains a variable denoted as X_1.
- Box-Counting-Trees-Dendrites. The fractal dimension of the dendrites using the box-counting method. The block contains a variable denoted as X_2.
- Branch-Angle-Axon. We used planar, local, and spline angles that measure the direction of the branches at different levels. We computed the mean, standard deviation, and median of the three angles for the axonal arbor. Also, we measured these variables by dividing the data set according to the centrifugal order of the segments. The block contains variables from X_3 to X_{929}.
- Branch-Angle-Dendrite. Similarly, for the dendritic arbor, the result is variables from X_{930} to $X_{1,100}$.
- Cell-Bodies. The area, aspect ratio, compactness, convexity, contour size (maximum and minimum Feret), form factor, perimeter, roundness, and solidity of the soma. The block contains variables from $X_{1,101}$ to $X_{1,110}$.
- Convex-Hull-2D. This analysis measures the area and the perimeter of the 2D convex hull that includes the entire neuronal morphology. The block contains variables from $X_{1,111}$ to $X_{1,114}$.
- Convex-Hull-3D. This analysis measures the volume and the surface of the 3D convex hull that includes the entire neuronal morphology. The block contains variables from $X_{1,115}$ to $X_{1,118}$.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

- Neuron-Summary-Axon. The number of axonal endings, the total length of the axon, the mean length of the axonal trees, and the number of nodes (branching points) in the axon. The block contains variables from $X_{1,119}$ to $X_{1,122}$.
- Neuron-Summary-Dendrites. The number of endings, the number of nodes (branching points), the total length, and the mean length of each dendritic arbor. The block contains variables from $X_{1,123}$ to $X_{1,126}$.
- Polar-Axon. The polar histogram is a 360° projection of data that accounts for the neurite length and direction. In the projection, the length of a wedge is equivalent to the total length of neurites in a specified direction. A fan-in diagram is generated to further study axon directionality. This diagram is divided into sectors. The numbers in the polar variables represent the sector in which the variable is measured. This approach only considers the axonal arbor. The block contains variables from $X_{1,127}$ to $X_{1,162}$.
- Polar-Dendrite. This analysis only considers the dendritic arbor. The block contains variables from $X_{1,163}$ to $X_{1,198}$.
- Segment-Axons. The total, mean, median, and standard deviation of the length of the segments belonging to the axonal arbor of the neuron. Also, we measured these variables by dividing the data set according to the centrifugal order of the segments. The block contains variables from $X_{1,199}$ to $X_{1,614}$.
- Segment-Dendrites. Similarly, for dendrites, the block contains variables from $X_{1,615}$ to X_{1694}.
- Sholl-Axon. The number of intersections in concentric spheres centered at the soma with increasing radii of 60 μm. The analysis also includes the number of endings, nodes, and the total length of the segments included in those spheres. Only the axonal arbor is analyzed. The block contains variables from $X_{1,695}$ to $X_{1,806}$.
- Sholl-Dendrite. Similarly, for dendrites, the block contains variables from $X_{1,807}$ to $X_{1,846}$.
- Tree-Totals-Axon. The number of endings and the number of segments of the axonal arborization. These variables were also measured according to the centrifugal order. The block contains variables from $X_{1,847}$ to $X_{2,052}$.
- Tree-Totals-Dendrite. Similarly, for dendrites, the block contains variables from $X_{2,053}$ to $X_{2,090}$.
- Vertex-Axon. Vertex analysis of the connectivity of the nodes in the branches is performed to describe the topological and metric properties of the axonal arbor. The block contains variables from $X_{2,091}$ to $X_{2,506}$.
- Vertex-Dendrite. Similarly, for dendrites, the block contains variables from $X_{2,507}$ to $X_{2,586}$.

The data set presented in Table 1.7 can be transformed in several alternative ways to reach a consensus among the responses of the 42 experts. After separately considering each of the 6 class variables, one possibility is to establish a consensus threshold, the simplest version of which would coincide with the majority vote. In this case, the category assigned to each neuron will be the one with highest frequency. Another possibility is to select those neurons with a given category frequency greater than 21 (half of the number of experts) from the data set. The last option is to use the information provided by the experts in terms of the relative frequency of each of the possible categories. This option is illustrated in Table 1.8 for the class variable C_5, representing the common usage of the interneuron names in the literature, encoded from 1 to 10.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.8 Data set containing information for the 241 3D reconstructed interneurons. For each neuron, 2,586 morphological variables have been recorded, as well as a probabilistic label for C_5, from the categories provided by the 42 neuroanatomists

				C_5									
Neuron	X_1	$\cdots$	$X_{2,586}$	1	2	3	4	5	6	7	8	9	10
1	10.8	$\cdots$	20.3	0.08	0.14	0.04	0.32	0.02	0.13	0.01	0.05	0.11	0.10
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
241	9.2	$\cdots$	18.9	0.07	0.10	0.05	0.23	0.08	0.21	0.03	0.06	0.12	0.05

1.6.3 Data Set 3: Quality of Life in Parkinson's Disease

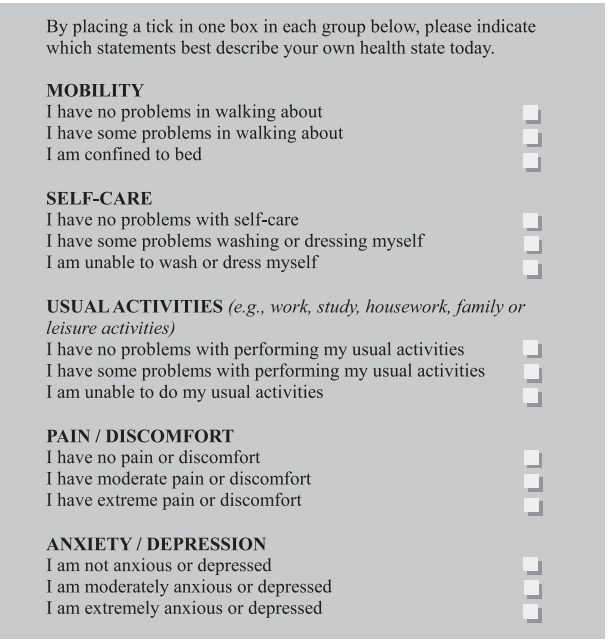
The data for this example were derived from a study by Borchani et al. (2012) that attempted to predict the European Quality of Life-5 Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39). The EQ-5D is a generic health-related quality-of-life measure that is used in general populations and patients with any disorder. EQ-5D contains five items (Figure 1.17), namely Mobility, Self-care, Usual activities, Pain/Discomfort, and Anxiety/Depression. Each item has three possible responses: `no problems`, `some problems`, and `severe problems`. However, PDQ-39 (Figure 1.18) is a specific instrument that is widely used in individuals with PD to capture patients' perceptions of their illness. PDQ-39 measures the severity and degree of disability in patients with PD using 39 questions covering 8 dimensions (see Table 1.9): mobility, activities of daily living, emotional well-being, stigma, social support, cognitions, communication, and bodily discomfort. Each question is scored on a five-point scale: `never`, `occasionally`, `sometimes`, `often`, and `always`.

The analyzed data set includes 488 patients with PD, each of whom was characterized with 39 predictor variables (PDQ-39) and 5 variables to be predicted (EQ-5D). The objective is to learn a multidimensional classifier that is able to assign the 5 classes of the EQ-5D to each patient. As these 5 classes are believed to be interrelated, the solution of learning 5 unidimensional supervised classification models does not appear to be appropriate. Therefore, an approach based on multidimensional classification should be adopted. Chapter 10 will be devoted to this.

1.6.4 Data Set 4: Dendritic Spines

This example analyzes dendritic spines, which were first described by Cajal in 1888. Dendrites of a single neuron can contain hundreds or thousands of spines. Although their exact functions remain unclear (reviewed in Yuste [2010]), the morphology of dendritic spines appears to be critical for their functions. Pyramidal neuron spines are the targets of most excitatory synapses in the cerebral cortex. The shape of the dendritic spines may determine their synaptic strength and learning rules. Quantitative analyses have revealed strong correlations between spine morphological variables and the synaptic structure: (a) the spine head volume and total spine volume in the neocortex are positively correlated with the area of the post-synaptic density, with a remarkably small variance (Arellano et al., 2007). This area is correlated with the number of presynaptic vesicles, the number

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press



By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

MOBILITY
I have no problems in walking about
I have some problems in walking about
I am confined to bed

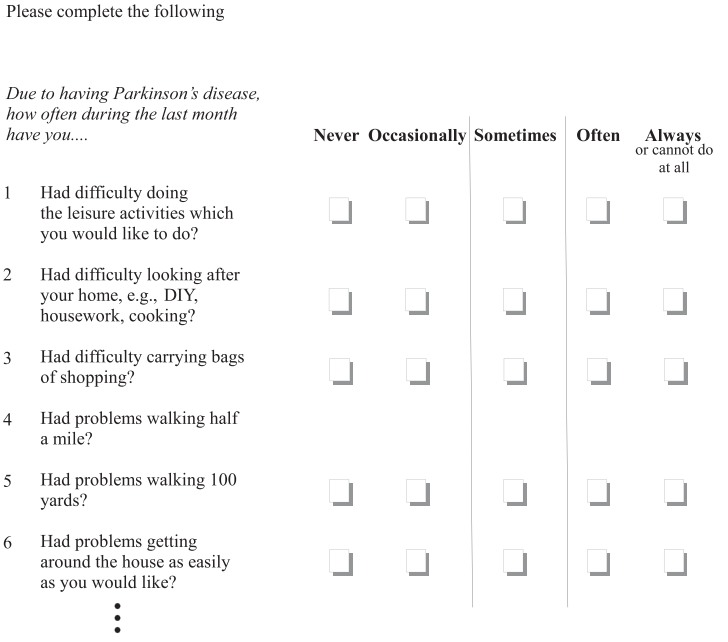
SELF-CARE
I have no problems with self-care
I have some problems washing or dressing myself
I am unable to wash or dress myself

USUAL ACTIVITIES *(e.g., work, study, housework, family or leisure activities)*
I have no problems with performing my usual activities
I have some problems with performing my usual activities
I am unable to do my usual activities

PAIN / DISCOMFORT
I have no pain or discomfort
I have moderate pain or discomfort
I have extreme pain or discomfort

ANXIETY / DEPRESSION
I am not anxious or depressed
I am moderately anxious or depressed
I am extremely anxious or depressed

Figure 1.17 Five classes of the EQ-5D quality-of-life measure.



Please complete the following

Due to having Parkinson's disease, how often during the last month have you....	Never	Occasionally	Sometimes	Often	Always or cannot do at all
1 Had difficulty doing the leisure activities which you would like to do?					
2 Had difficulty looking after your home, e.g., DIY, housework, cooking?					
3 Had difficulty carrying bags of shopping?					
4 Had problems walking half a mile?					
5 Had problems walking 100 yards?					
6 Had problems getting around the house as easily as you would like?					

Figure 1.18 Six questions of the PDQ-39 questionnaire.

of postsynaptic receptors, and the ready-releasable pool of transmitter, (b) the length of the spine neck is proportional to the extent of biochemical and electrical isolation of the spine from its parent dendrite (Harris and Stevens, 1989; Nusser et al., 1998; Yuste et al., 2000), (c) larger spines can generate larger synaptic currents than smaller spines (Matsuzaki et al.,

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.9 The PDQ-39 items

Mobility:	
pdq1	Had difficulty doing the leisure activities you would like to do
pdq2	Had difficulty looking after your home, e.g., DIY, housework, cooking
pdq3	Had difficulty carrying bags of shopping
pdq4	Had problems walking half a mile
pdq5	Had problems walking 100 yards
pdq6	Had problems getting around the house as easily as you would like
pdq7	Had problems getting around in public
pdq8	Needed someone else to accompany you when you went out
pdq9	Felt frightened or worried about falling over in public
pdq10	Been confined to the house more than you would like
Activities of daily living:	
pdq11	Had difficulty washing yourself
pdq12	Had difficulty dressing yourself
pdq13	Had problems doing up buttons or shoe laces
pdq14	Had problems writing clearly
pdq15	Had difficulty cutting up your food
pdq16	Had difficulty holding a drink without spilling it
Emotional well-being:	
pdq17	Felt depressed
pdq18	Felt isolated and lonely
pdq19	Felt weepy or tearful
pdq20	Felt angry or bitter
pdq21	Felt anxious
pdq22	Felt worried about your future
Stigma:	
pdq23	Felt you had to conceal your Parkinson's from people
pdq24	Avoided situations which involve eating or drinking in public
pdq25	Felt embarrassed in public due to having PD
pdq26	Felt worried by other people's reaction to you
Social support:	
pdq27	Had problems with your close personal relationships
pdq28	Lacked support in the ways you need from your spouse or partner
pdq29	Lacked support in the ways you need from your family or close friends
Cognitions:	
pdq30	Unexpectedly fallen asleep during the day
pdq31	Had problems with your concentration, e.g., when reading or watching TV
pdq32	Felt your memory was bad
pdq33	Had distressing dreams or hallucinations
Communication:	
pdq34	Had difficulty with your speech
pdq35	Felt unable to communicate with people properly
pdq36	Felt ignored by people
Bodily discomfort:	
pdq37	Had painful muscle cramps or spasms
pdq38	Had aches and pains in your joints or body
pdq39	Felt unpleasantly hot or cold

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

2004), and (d) dendritic spines are dynamic structures with fluctuations in volume that appear to have important implications for cognition and memory (Dunaevsky et al., 1999; Matus, 2000; Kasai et al., 2010).

Dendritic spines present a wide variety of morphologies, particularly in the human cortex (Benavides-Piccione et al., 2013). Spines are highly motile and can undergo reshaping, even in the adult. In fact, the loss or alteration of these structures has been described in the pathogenesis of major neurological disorders. Thus, a statistical analysis of spine morphology is indispensable for providing formal support for these and other hypotheses.

Although different morphology-based classifications of spines have been proposed, the one that is still most widely used today categorizes spines into three essential types: thin, mushroom, and stubby (Peters and Kaiserman-Abramof, 1970). This classification relies solely on a visual inspection of microscopy images and focuses on the head-to-neck diameter ratio, length-to-head diameter ratio, and head diameter. However, researchers have also argued that the large diversity of spines portrays a continuum of morphologies rather than the existence of discrete classes (Arellano et al., 2007). Therefore, a detailed description of morphologies is needed to identify clusters of human spines that share similar characteristics. This approach will likely require a certain probability of cluster membership for a given spine, capturing that continuum.

The data set contains 2,000 individually 3D reconstructed dendritic spines from layer III pyramidal neurons located in the cingulate cortex of a human male age 40 years and constitutes a random sample extracted from the pool of more than 4,500 spines of this individual analyzed in Luengo-Sanchez et al. (2018). Eight hundred eighty-six spines (44.30%) were located on apical dendrites, whereas the remaining 1,114 spines (55.70%) were located on basal dendrites. The tissue was obtained at autopsy (2–3 h postmortem). Apical and basal dendrites were then scanned using confocal microscopy and completely reconstructed in 3D using a methodology detailed elsewhere (Benavides-Piccione et al., 2013). Then, for each spine, a particular threshold was selected to constitute a solid surface that exactly matched its contour, see Figure 1.19.

An important issue is to extract a set of variables describing the 3D spine shapes. These variables should be sufficiently representative to summarize the shape and sufficiently meaningful to be easily interpreted by domain experts. Graph-based techniques handle both global (coarser) and local (more detailed) features. The techniques extract a geometric meaning from a 3D shape using a graph showing the linkage of shape components. Of the many existing shape-matching methods (Tangelder and Veltkamp, 2008), we were partially inspired by the concept of the Reeb graph defined by a geodesic distance (length of the shortest path along the surface of the model) for this data set. A skeletal structure of the 3D model in the form of a graph (Figure 1.20) is built. This graph captures the global topology of the shape. Then, variables are attached to each graph node to consider local features.

Spines were approximated by a continuous surface composed of a sequence of seven sections S_i (coaxial tubular-shaped) with heights $h_i, i = 1, \ldots, 7$. In each section, curves defining top, T_i, and bottom, B_i, regions were assumed to be ellipses with major (T_i^R, B_i^R) and minor (T_i^r, B_i^r) radii or axes. Thus, B_i^R is the major radius of the ellipse separating sections S_{i-1} and S_i. The surface was required to be continuous, and therefore coherence constraints were imposed on adjacent sections: $B_i^R = T_{i-1}^R, B_i^r = T_{i-1}^r, \forall i = 1, \ldots, 7$. Ratios between sections S_i and S_j, denoted by $\varphi_{ij} = \frac{B_j^R B_j^r}{B_i^R B_i^r}$, provide information about the widening

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

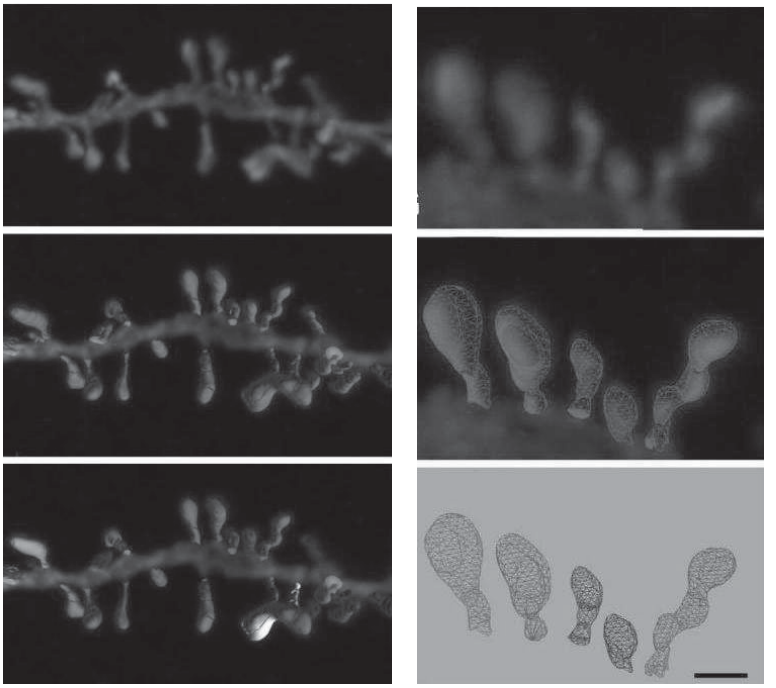


Figure 1.19 From top to bottom and from left to right: confocal microscopy z projection image of a dendritic segment from a horizontally projecting basal dendrite of an intracellular injected layer III pyramidal neuron of the human cingulate cortex (age 40 years). The complete morphology of each dendritic spine is reconstructed in 3D. Estimation of the spine volume values using color codes (blue-white: 0.0–0.8 μm^3). Ultimately, for each individual spine, 3D triangular meshes are output. Scales are: 2.5 μm (figures to the left) and 1 μm (figures to the right). Reprinted with permission from Luengo-Sanchez et al. (2018). For the color version, please refer to the plate section.

or narrowing along the spine. Three ratios, φ_{24}, φ_{26}, and φ_{46}, were considered. The growing direction of each ellipse of the spine is related to the mean direction of a region and is measured with the cosine of the azimuth angle, $cos(\phi_i)$, and by the polar angle θ_i for each ellipse, except the first one. The direction Φ of the perpendicular vector to the ith ellipse is called the instant direction and is determined with its azimuth angle and its polar angle, Θ_i, namely, the inclination of the vector perpendicular to the ellipse with respect to the z-axis. The volume of each region, V_i, is an approximation of the volume between two consecutive ellipses and computed from the convex hull of T_i and B_i. The volume of the spine, V, denotes the total volume of the spine and is computed as $V = \sum_{i=1}^{7} V_i$. In summary, the 54 features characterizing each spine (36 morphological features and 18 features necessary for their subsequent simulation) are shown in Table 1.10.

Figure 1.20 illustrates the meaning of the different morphological and simulation features presented in Table 1.10. Chapters 11 and 12 will find groups of these spines.

1.6.5 Data Set 5: Basal Dendritic Trees

This example concerns dendritic morphology, which is essential for understanding connectivity and the functional roles of neurons. Specifically, pyramidal neurons represent key

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.10 The 54 features (36 morphological features and 18 features for simulation purposes) characterizing each spine in Data Set 4

Type	Feature	Description	Number
Morphological	h_1–h_7	Height of each section	7
Morphological	B_2^R–B_7^R	Major axis of the ellipse of each section	6
Morphological	B_2^r–B_7^r	Minor axis of the ellipse of each section	6
Morphological	φ_{24}, φ_{26}, φ_{46}	Ratio between sections	3
Morphological	$\cos(\phi_2)$–$\cos(\phi_7)$	Cosine of the azimuth angle of the growing direction	6
Simulation	θ_2–θ_7	Polar angle of the growing direction of each ellipse	6
Simulation	Φ_2–Φ_7	Direction of the perpendicular vector to the ellipse	6
Simulation	Θ_2–Θ_7	Inclination of the vector perpendicular to the ellipse	6
Morphological	V_1–V_7	Volume of each region	7
Morphological	V	Volume of the spine	1

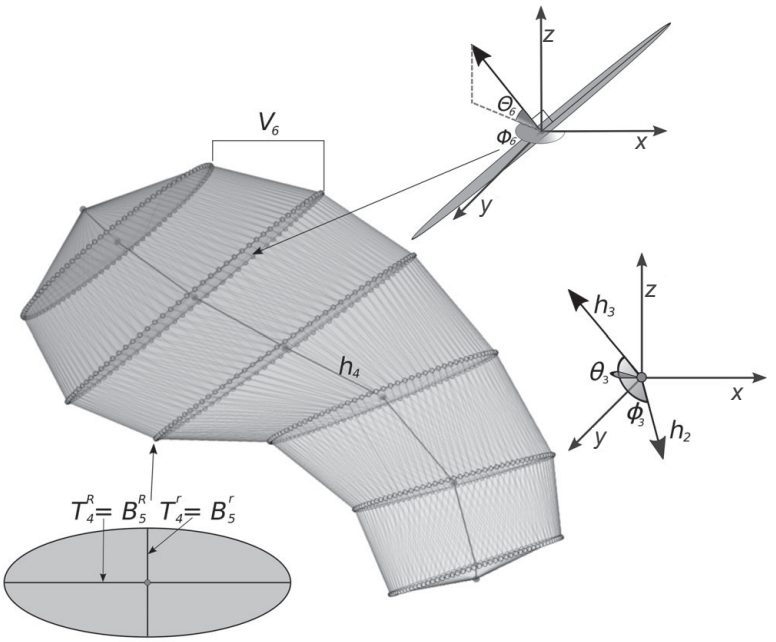


Figure 1.20 Illustration of the features used to characterize the spines. Observe the points at the centroids of the ellipses. They are connected by vectors whose lengths are denoted by h_i. Each ellipse is defined by its centroid, major axis ($T_{i-1}^R = B_i^R$), and minor axis ($T_{i-1}^r = B_i^r$). From the vectors connecting the centroids of the ellipses, angles ϕ_i, θ_i, and Φ_i, Θ_i are computed. The volumes of each section, V_i, are added for computing the volume V of the whole spine.

elements in the functional organization of the cerebral cortex, as they are the most frequent neuronal type (70–85%) and the main source of cortical excitatory synapses. The structure of the dendritic tree of pyramidal neurons affects the process of integration, and its size influences the topographic sampling map and the mixing of inputs (Wen et al., 2009). The branching patterns of the dendritic trees are related to synaptic input processing (Koch and Segev, 2000; Häusser and Mel, 2003) and affect the electrical behavior of the neurons (Mainen and Sejnowski, 1996; Vetter et al., 2001; Chen, 2009).

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

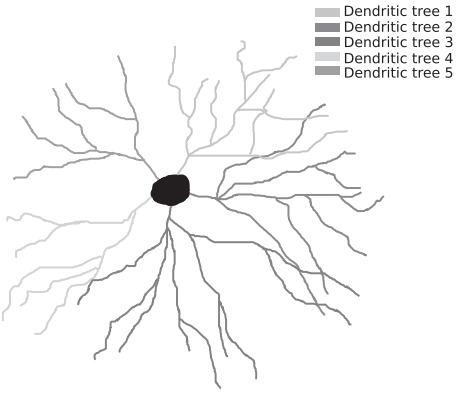


Figure 1.21 Basal dendritic arbor of a pyramidal neuron, where each dendritic tree is drawn in a different tone. Adapted from López-Cruz et al. (2011).

Researchers do not understand how and why vastly different arbor shapes form. In the last two decades, computational stochastic models have been used to measure relevant parameters of real neuronal arbors. These models and their subsequent use for simulating virtual neuron morphologies may help identify the basic structures and important features of neuronal classes.

Neuronal processes are not always easy to trace, and data on the complete dendritic tree of real neurons are rather scarce. However, in pyramidal neurons, the whole basal dendritic arbor – about 90% of the dendritic length in neurons from layers II/III and V (Larkman, 1991) – has been completely reconstructed in single horizontal sections (Elston and Rosa, 1997). This information is valuable for validating the simulated virtual neurons.

Thus, Data Set 5 includes a set of 3D reconstructions of 90 pyramidal neurons from the mouse neocortex (two BC57 black mice, 2 months old). The neurons were located in layer III of three different cortical regions: the secondary motor cortex (M2), secondary somatosensory cortex (S2), and lateral secondary visual cortex and association temporal cortex (V2L/TeA). The whole basal dendritic trees of the neurons were traced using the Neurolucida package (Glaser and Glaser, 1990). The tissue preparation and injection process are detailed in Benavides-Piccione et al. (2006). The reconstructions are publicly available at `www.neuromorpho.org` (Ascoli, 2007) as part of the DeFelipe laboratory archive.

Each basal dendritic arbor is composed of approximately 6 main trunks, which we will call dendritic trees, see Figure 1.21. One hundred four dendritic trees were observed in M2, 103 in S2, and 156 in V2L/TeA.

A segment is the straight line between two branching points. For each pair of sibling segments, a set of 41 morphological variables is measured from the 3D reconstructions of real dendrites, as described in López-Cruz et al. (2011). We distinguish two types of variables: first, construction variables that define the morphology of a segment (segment length, orientation, and bifurcation) and are necessary to incrementally build the virtual dendritic trees; and second, evidence variables that measure the part of the morphology of the dendritic tree located below a pair of sibling segments. Table 1.11 lists the 41 variables.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.11 Construction (C) and evidence (E) morphological variables for Data Set 5

No.	Type	Variable	No.	Type	Variable
1	E	Subtree degree (no. endings)	22	E	Neighbor distance
2	E	Subtree no. bifurcations (no. nodes)	23	E	Neighbor inclination
3	E	Subtree total length	24	E	Neighbor azimuth
4	E	Subtree width	25	E	Neighbor extension
5	E	Subtree height	26	E	Neighbor angle
6	E	Subtree depth	27	E	Parent segment length
7	E	Subtree box volume	28	E	Parent segment inclination
8	E	Subtree max distance between nodes	29	E	Parent segment azimuth
9	E	Subtree max distance to soma	30	E	Root segment length
10	E	Subtree max length	31	E	Root segment inclination
11	E	Subtree min length	32	E	Root segment azimuth
12	E	Subtree max order	33	E	Segment centrifugal order
13	E	Subtree min order	34	C	Left segment length
14	E	Subdendrite length	35	C	Left segment inclination
15	E	Subdendrite width	36	C	Left segment azimuth
16	E	Subdendrite height	37	C	Left segment bifurcates
17	E	Subdendrite depth	38	C	Right/root segment length
18	E	Subdendrite box volume	39	C	Right/root segment inclination
19	E	Subdendrite distance to soma	40	C	Right/root segment azimuth
20	E	Subdendrite inclination	41	C	Right/root segment bifurcates
21	E	Subdendrite azimuth			

More specifically, evidence variables describe the context of the segment and how the tree is constructed. These variables include information about the subtree (variables 1–13), subdendrite (variables 14–21) and nearest segment (variables 22–26). The centrifugal order (or branch order) of a segment is the number of bifurcations along the path to the soma. For a given pair of sibling segments with an order a, the subtree is the part of the dendritic tree including all the segments with an order less than a. Likewise, the subdendrite is the path from the soma to the sibling segments' branching point. Figure 1.22 shows a pair of sibling segments with a centrifugal order value of 5 (gray lines), its subtree (gray area), and subdendrite (dotted area). Finally, the nearest segment refers to the segment in the dendritic tree that does not belong to the subdendrite (neighboring segment in Figure 1.22).

Parent segment morphological variables (27–29) and root segment morphological variables (30–32), as well as the centrifugal order of the segment (variable 33), complete the set of evidence variables.

Construction variables specify the segment morphology (variables 34–41), whether the segments (left/right) branch, and the spherical coordinates of each segment end point.

Chapter 13 will deal with these variables and use their statistical distributions to automatically find their relationships in shaping the dendritic tree structure. Then, a simulation algorithm will sample the distributions to output virtual dendrites that should be indistinguishable from real dendrites.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

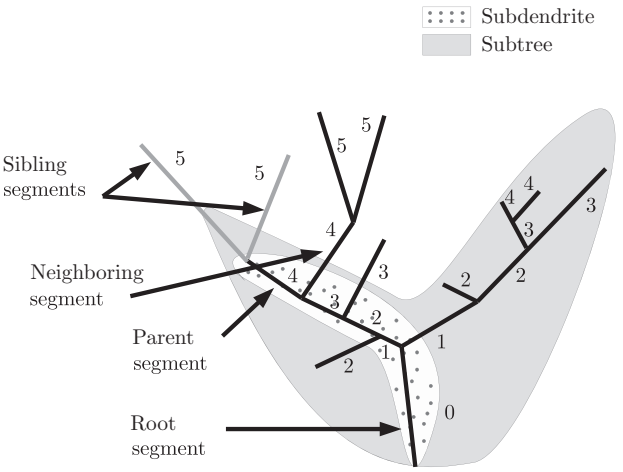


Figure 1.22 Subtree (gray area), subdendrite (dotted area), neighboring, parent and root segments for two sibling segments (gray lines). The numbers refer to the centrifugal order of the segments. Adapted from López-Cruz et al. (2011).

1.6.6 Data Set 6: Brain Connectivity

This data set, which was obtained from Huang et al. (2010), includes fluorodeoxyglucose PET images from 232 subjects: 49 with AD, 116 with mild cognitive impairment (MCI), and 67 normal controls (NCs). All images were downloaded from the Alzheimer's disease neuroimaging initiative (ADNI) database[12] to define the progression of AD. Image preprocessing, which is mainly performed with Statistical Parametric Mapping software (Wellcome Department of Cognitive Neurology[13]), ultimately yields average PET measurements for 42 anatomical volumes of interest, which are known to be most frequently affected by AD. These brain regions include the frontal, parietal, occipital, and temporal lobes. Table 1.12 lists the 42 variables.

In Chapter 14, we will use these data to identify functional brain connectivity networks for the three different types of subjects: patients with AD, patients with MCI, and NCs. Functional connectivity refers to the coherence of the activities among distinct brain regions (Horwitz, 2003). We search for statistical dependencies between different brain regions. Higher cognition is derived from the interactions between different brain regions rather than individual regions working independently. Thus, patients with AD, who are characterized by a global cognitive decline, may exhibit abnormal functional brain connectivity patterns. For example, the hippocampus and other regions in the brains of patients with AD exhibit reduced functional connectivity (Grady et al., 2001; Wang et al., 2007; Supekar et al., 2008), whereas increased connectivity has been observed between the frontal lobe and other regions in the brains of subjects with early AD and MCI, which is interpreted as a compensatory reallocation or recruitment of cognitive resources (Gould et al., 2006).

[12] adni.loni.usc.edu/.
[13] www.fil.ion.ucl.ac.uk/spm.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

Table 1.12 Variables representing anatomical volumes of interest for Data Set 6 (L = left hemisphere, R = right hemisphere)

Frontal lobe		Parietal lobe		Occipital lobe		Temporal lobe	
Var.	Description	Var.	Description	Var.	Description	Var.	Description
X_1	Frontal_Sup_L	X_{13}	Parietal_Sup_L	X_{21}	Occipital_Sup_L	X_{27}	Temporal_Sup_L
X_2	Frontal_Sup_R	X_{14}	Parietal_Sup_R	X_{22}	Occipital_Sup_R	X_{28}	Temporal_Sup_R
X_3	Frontal_Mid_L	X_{15}	Parietal_Inf_L	X_{23}	Occipital_Mid_L	X_{29}	Temporal_Pole_Sup_L
X_4	Frontal_Mid_R	X_{16}	Parietal_Inf_R	X_{24}	Occipital_Mid_R	X_{30}	Temporal_Pole_Sup_R
X_5	Frontal_Sup_Medial_L	X_{17}	Precuneus_L	X_{25}	Occipital_Inf_L	X_{31}	Temporal_Mid_L
X_6	Frontal_Sup_Medial_R	X_{18}	Precuneus_R	X_{26}	Occipital_Inf_R	X_{32}	Temporal_Mid_R
X_7	Frontal_Mid_Orb_L	X_{19}	Cingulum_Post_L			X_{33}	Temporal_Pole_Mid_L
X_8	Frontal_Mid_Orb_R	X_{20}	Cingulum_Post_R			X_{34}	Temporal_Pole_Mid_R
X_9	Rectus_L					X_{35}	Temporal_Inf_L
X_{10}	Rectus_R					X_{36}	Temporal_Inf_R
X_{11}	Cingulum_Ant_L					X_{37}	Fusiform_L
X_{12}	Cingulum_Ant_R					X_{38}	Fusiform_R
						X_{39}	Hippocampus_L
						X_{40}	Hippocampus_R
						X_{41}	ParaHippocampal_L
						X_{42}	ParaHippocampal_R

1.6.7 Data Set 7: Spatial Location of Synapses in the Neocortex

One major issue in cortical circuitry is to determine the spatial distribution of synapses and whether synaptic connections are specific (DeFelipe et al., 2002). Two major morphological types of synapses have been identified: asymmetric and symmetric (Gray, 1959). The major sources of asymmetric synapses are spiny neurons (pyramidal and spiny nonpyramidal cells) and extrinsic cortical afferents, whereas the vast majority of symmetric synapses are formed by the population of aspiny or sparsely spiny interneurons.

The state-of-the-art methods for obtaining 3D data from which to estimate the spatial distribution, size, and number of synapses from ultrathin sections of brain tissue are based on serial reconstructions (Bock et al., 2011). The development of automated electron microscopy techniques has attempted to overcome the extremely time-consuming and difficult task of reconstructing large volumes of tissue (Briggman and Denk, 2006).

The tissues in this example (Merchán-Pérez et al., 2014) were obtained using a new dual-beam electron microscope that combines a focused ion beam (FIB) column and a SEM. The FIB column mills thin layers of material as a result of the collision of the gallium ion beam with the tissue. The SEM is then applied to the milled surface, obtaining a backscattered electron image. This milling/imaging process is automatically repeated to obtain a large series of images that represent a 3D sample of the tissue. Image resolution in the *xy* plane was 3.7 nm/pixel. The *z*-axis resolution (section thickness) was 20 nm.

Three male Wistar rats sacrificed on postnatal day 14 were used for this study. Animals were administered a lethal intraperitoneal injection of sodium pentobarbital (40 mg/kg) and were intracardially perfused with 2% paraformaldehyde and 2.5% glutaraldehyde in 0.1 M phosphate buffer. All animals were handled in accordance with the guidelines for animal research established in the European Union Directive 2010/63/EU, and all procedures were approved by the local ethics committee of the Spanish National Research Council.

Ten different samples of the neuropil in layer III of the somatosensory cortex were obtained from three different animals. All samples selected for FIB milling/SEM imaging were located at the mid-depth of layer III. After applying a 3D unbiased counting frame and correcting for tissue shrinkage (Merchán-Pérez et al., 2009), the volume of tissue

Table 1.13 Descriptive characteristics of the 10 samples of the neuropil in layer III

Sample no. and animal identification	Number of synapses per cubic micron	Mean distance to nearest neighbor (nm) ± sd	Mean Feret's diameter of synaptic junctions (nm) ± sd
1 (W31)	0.9857	519.55 ± 136.35	377.19 ± 159.63
2 (W31)	0.6936	594.07 ± 192.28	462.18 ± 177.52
3 (W33)	0.9279	537.43 ± 159.20	437.62 ± 168.04
4 (W33)	1.0088	537.39 ± 157.70	414.22 ± 169.04
5 (W33)	0.9474	597.30 ± 174.02	466.03 ± 215.91
6 (W33)	0.9399	533.21 ± 163.29	423.38 ± 169.83
7 (W33)	0.9881	487.17 ± 172.30	397.29 ± 168.22
8 (W35)	0.7997	568.21 ± 178.51	427.90 ± 168.15
9 (W35)	1.1267	501.38 ± 156.97	378.35 ± 166.60
10 (W35)	1.0178	523.74 ± 150.36	405.43 ± 175.62
All samples	0.9399	535.78 ± 166.81	417.06 ± 175.97

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press

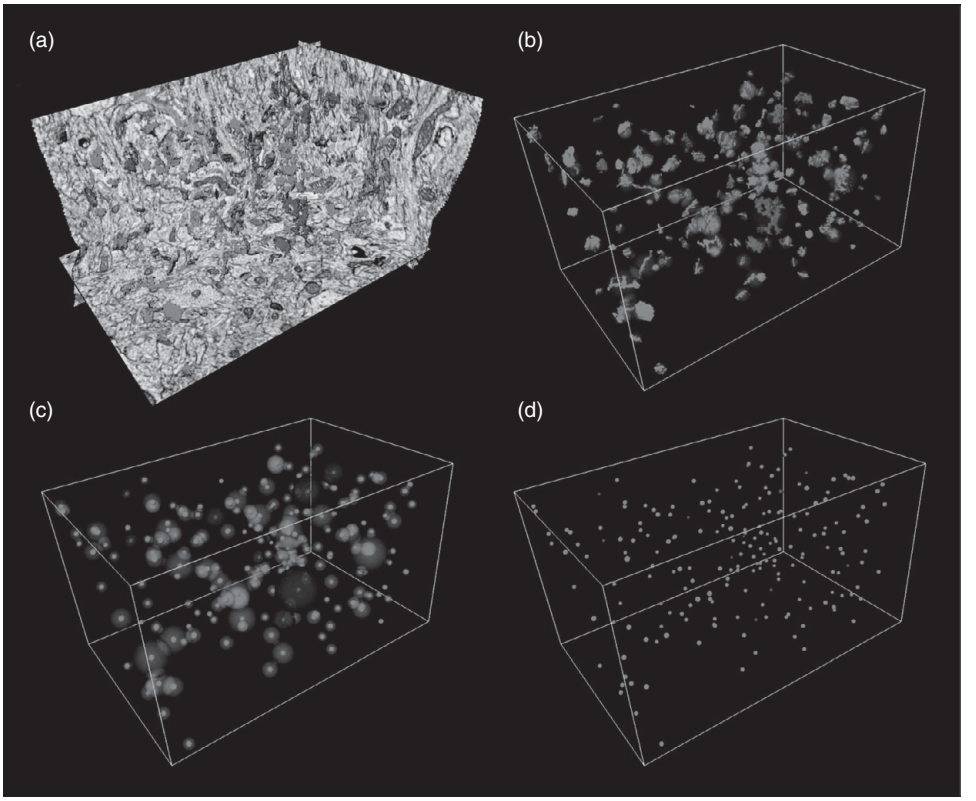


Figure 1.23 Example of a tissue volume whose dimensions are $7.16 \times 4.58 \times 3.98$ μm. (a) Asymmetric and symmetric synaptic junctions are shown in green and red, respectively. (b) Segmentation of the synaptic functions. (c) The smallest spheres circumscribing the synaptic junction used to calculate its Feret's diameter. (d) The centroids of the spheres. Image from Merchán-Pérez et al. (2014) reproduced with permission. For the color version, please refer to the plate section.

analyzed from each sample ranged from 149.13 to 247.58 μm^3. Synaptic junctions within these volumes were visualized, automatically segmented, and reconstructed in three dimensions using Espina software (Morales et al., 2011). There were 1,695 synaptic junctions.

Table 1.13 contains the following information about the 10 samples: animal identification; densities of synapses, which were calculated by dividing the actual number of synaptic junctions by the volume of the counting frame; mean distance to the nearest neighboring synapses, which was calculated as the distance between the centroids of the synaptic junctions; and the mean Feret's diameter of synaptic junctions, which was computed as the diameter of the smallest sphere circumscribing the synaptic junction. Centroids that were located closer to the boundaries of the counting frame than to any other centroid were excluded from the calculations because their nearest neighbor might be outside the counting frame at an unknown distance.

Figure 1.23 displays the segmentation of the reconstructed synaptic functions, the smallest sphere containing each of the synapses, and the geometric centers, or centroids, of the spheres for 1 of the 10 tissues. These centroids were determined to indicate the spatial position of the synapses. The spatial statistical analysis of this data set is presented in Chapter 15.

https://doi.org/10.1017/9781108642989.003 Published online by Cambridge University Press