Introduction: A Guided Tour through the Book

This chapter gives an overview of the content of the book. We follow the chapters in the sequence in which they appear, summarize key findings and theoretical arguments, and clarify the relationships between the chapters. Along the way, we explain some basic issues of overarching importance.

The book is divided into two parts: "Theory and Experiment" and "Background and Methods." The first part describes recent primary research findings about the visual system, along with cutting-edge theory and methodological considerations. The second part provides some of the more general neuroscientific and mathematical background needed for understanding the first part.

Although each chapter is independent, the first part, "Theory and Experiment," is designed to be read in sequence. The sequence roughly follows the stages of ventral-stream visual processing, which forms the focus of the book. Within this rough order, we placed closely related chapters together. We purposely interspersed theoretical and experimental chapters, and, within the latter, animal electrode recording and human fMRI studies. An overview of the chapters is given in figure I.1 and table I.1.

Localist and Distributed Codes

In chapter 1, Simon J. Thorpe reviews the debate about localist versus distributed neuronal coding in the context of recent experimental evidence. Early findings of neuronal selectivity to simple features at low levels of the visual hierarchy and to more complex features at higher levels suggested, by extrapolation, that there might be neurons that respond selectively to particular objects, such as one's grandmother. On a continuum of possible coding schemes from localist to distributed, this "grandmother cell" theory forms the localist pole. A code of grandmother cells could still have multiple neurons devoted to each object; the key feature is the high selectivity of the neurons. A grandmother-cell code is explicit in that no further processing is

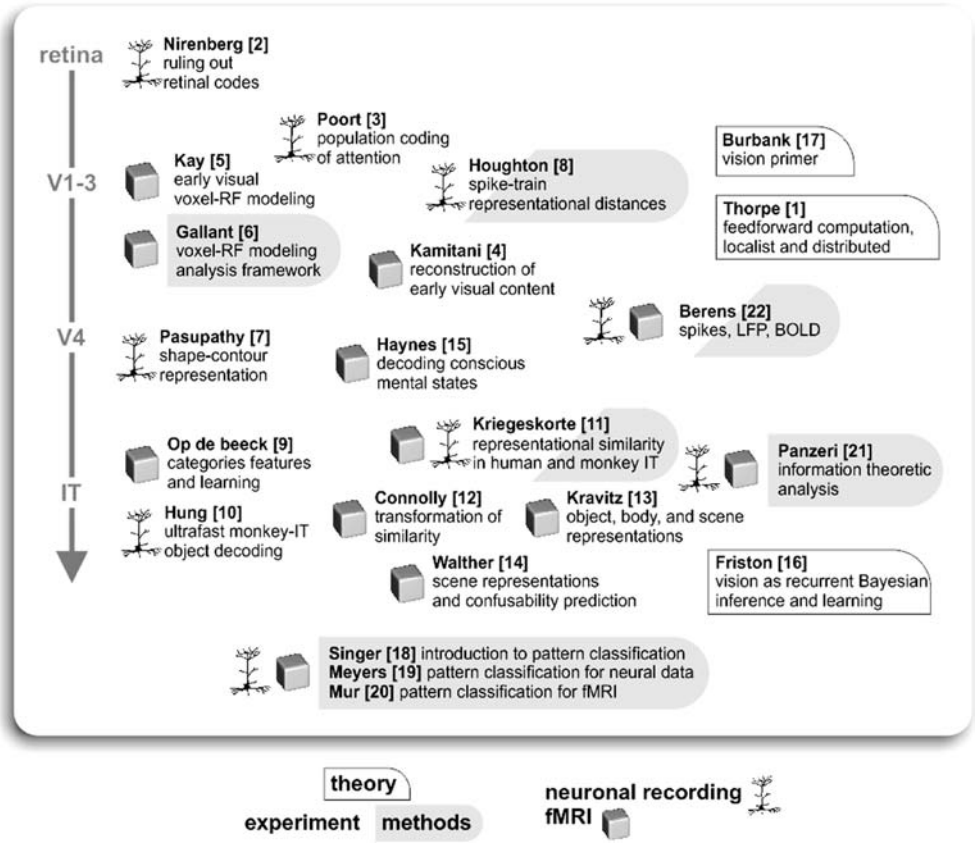

Figure I.1
Chapter overview. Along the vertical axis (arrow on the left), the chapters have been arranged roughly according to the stage of processing they focus on. Horizontally, chapters with a stronger focus on a particular stage of processing are closer to the axis on the left. Where possible, chapters related by other criteria are grouped together. For example, chapters 5 and 6 use the method of voxel-receptive-field modeling, while chapters 9 and 11–14 use the method of representational similarity analysis. Neuron and voxel icons label chapters using neuronal recordings and fMRI, respectively. Chapters focusing on theory, experiment, or methods have been visually indicated (see legend), with methods chapters marked by a gray underlay and experimental chapters with a strong methodological component marked by a partial gray underlay.

required to read out the code and conclude that a particular object is present. At the other end of the continuum is a distributed code, in which each neuron will respond to many different objects; thus, there is no single neuron that unequivocally indicates the presence of a particular object. In a distributed code, the information is in the *combination* of active neurons.

For a population of n neurons, a localist single-neuron code can represent no more than n distinct objects, one for each neuron—and less if multiple neurons

Table I.1
Chapter content overview

	First Author, Last Author	Content Type	Regions	Brain-Activity Measurement	Content
1	Thorpe	Theory, model, exp.	Retina-IT	Electrode	Localist vs. distributed coding; spike-timing-dependent coding; plasticity
2	Nirenberg	Theory, exp., methods	Retina	In vitro recording	Ruling out retinal codes by comparing information between code and behavior
3	Poort, Roelfsema	Exp.	V1	Electrode	Decoding stimulus features and attentional states from V1 neurons
4	Kamitani	Exp., methods	V1-3, MT	fMRI	Decoding human early visual population codes and stimulus reconstruction
5	Kay	Methods, model, exp.	V1-4	fMRI	Voxel-receptive-field modeling for identification of natural images
6	Gallant, Wu	Methods, model, exp.	V1	fMRI	Methodological framework for voxel-receptive-field modeling
7	Pasupathy, Brincat	Exp.	V4, pIT	Electrode	Shape-contour representation by convex/concave curvature-feature combinations
8	Houghton, Victor	Theory, methods, exp.	—	Electrode	Measuring representational dissimilarity by spike-train edit distances
9	Op de Beeck	Exp., theory	IT	fMRI	Category modules vs. feature map; influences of task and learning
10	Hung, DiCarlo	Exp., theory	IT	Electrode	Decoding object category and identity at small latencies after stimulus onset; invariances
11	Kriegeskorte, Mur	Exp., theory, model, methods	IT	fMRI, electrode	Categoricality of object representation, comparing human and monkey; methods
12	Connolly, Haxby	Exp., theory, methods	IT	fMRI	Transformation of similarity across stages; advantages of pattern similarity analyses
13	Kravitz, Baker	Exp., theory	IT	fMRI	Object, body, and scene representations; position dependence
14	Walther, Fei-Fei	Exp., theory, methods	IT	fMRI	Distributed scene representations; decoding confusions predict behavioral confusions

Table I.1
(continued)

	First Author, Last Author	Content Type	Regions	Brain-Activity Measurement	Content
15	Haynes	Theory, methods	LGN-IT	fMRI	Decoding consciousness; uni- vs. multivariate neural correlates of consciousness
16	Friston	Theory, model, exp.	Retina-IT	fMRI	Visual system as hierarchical model for recurrent Bayesian inference and learning
17	Burbank, Kreiman	Theory tutorial	Retina-IT	—	Essentials of visual processing across stages of the visual hierarchy; dorsal/ventral stream
18	Singer, Kreiman	Methods tutorial	—	—	Introduction to statistical learning theory and pattern classification
19	Meyers, Kreiman	Methods tutorial	—	Electrode	Step-by-step tutorial on pattern classification for neural data
20	Mur, Kriegeskorte	Methods tutorial	—	fMRI	Step-by-step tutorial on pattern classification for fMRI data
21	Panzeri, Ince	Theory, methods	—	Electrode, fMRI	Information theoretic analysis of neuronal population codes
22	Berens, Tolias	Exp., methods	—	Electrode, fMRI	Relationship between spikes, local field potentials, and fMRI

redundantly code for the same object, as is commonly assumed. A distributed code can use combinations of neurons and code for a vast number of different objects (for binary responses, for example, there are 2^n distinct activity patterns). If the patterns used for representing objects are randomly chosen among the 2^n combinations, about half of the neurons will respond to any given object. A distributed code can also represent the stimuli with some redundancy, making it robust to damage to particular neurons. Moreover, it can represent the objects in terms of sensory or semantic properties, thus placing the objects in a multidimensional abstract space that reflects their relationships. Such an abstract space might emphasize behaviorally relevant similarities and differences in a graded or categorical manner. Although the signals indicating the presence of a particular object are distributed, the code may still be considered "explicit" if readout takes just a single step—for example, a downstream neuron that computes a linear combination of the neuronal population. (Such a downstream neuron would be a localist neuron.)

Note that what is called localist and distributed is fundamentally in the eye of the beholder, as it depends on the way the researcher thinks of the information to be represented. For example, consider the case of two neurons that encode the two-dimensional space of different jets of water. One neuron codes the amount of water per unit of time; the other the temperature of the water. A researcher who thinks of the space in terms of amount per unit of time and temperature will conclude that the code is localist. But a researcher who thinks of jets of water in terms of the amounts of cold and hot water per unit of time will conclude that the code is distributed. In practice, we tend to think of a code as localist if we can characterize each neuron's preferences in very simple terms; we think of the code as distributed if the description of the preference of a single neuron is complex and doesn't correspond to any concepts for describing the content that appear natural to us.

The "grandmother cell" theory did not initially have any direct empirical support. Findings of "grandmother" (or similarly highly selective) neurons were elusive. The failure to find such neurons, of course, doesn't prove that they don't exist. The idea of grandmother cells has also been criticized on theoretical grounds for failing to exploit the combinatorics. This led to a preference for more distributed coding schemes among many theorists. Indeed, distributed codes and multivariate analysis of the information they carry is a central theme of this book.

Sparse Distributed Codes

Despite the advantages of distributed codes, the appeal of highly selective single cells is not merely in the eye of the electrophysiologist who happens to record one cell at a time with a single electrode. The reason why more of the page you are reading is white than black may be the cost of ink. Similarly, the metabolic cost of neuronal activity creates an incentive for a code that is sparser (i.e., fewer cells responding to a particular object due to each cell's greater selectivity) than one that fully exploits the combinatorics. On the continuum between localist and distributed, the concept of a sparse code has emerged as a compromise that may best combine the advantages of both schemes. In a sparse code, few neurons respond to any given stimulus. And, conversely, few stimuli drive any given neuron.

It seems likely that neurophysiological recordings have been biased toward describing neurons that fire more rapidly and less selectively, making them easier to find while looking for responses. Consistent with this notion, unbiased neurophysiological recordings using electrode arrays tend to report high selectivities, suggesting sparse representations, in a variety of systems including the songbird vocal center, the mouse auditory cortex, and the human hippocampus.

Thorpe discusses additional arguments in favor of sparse coding. More recent evidence from neurophysiological recordings in the human medial temporal lobe suggests that there are neurons responding selectively to complex particular objects, for example, to Jennifer Aniston. Interestingly, the "Jennifer-Aniston cell" responded not just to one image, but to several images of the actress and even to the visual presentation of her name in writing. The cell did not respond to any other stimuli that the researchers tried. However, the relatively small number of stimuli and neurons that can be examined in such experiments (on the order of hundreds) suggests that neurons of this type might well respond to multiple particular objects. The "Jennifer-Aniston cell," then, might be more promiscuous than its exclusive preference for the actress among the sampled set of stimuli would suggest. Thorpe (citing Rafi Malach) refers to this as the "totem-pole cell" theory, where a cell has multiple distinct preferences like the faces on a totem pole.

It is important to note that descriptions like "Jennifer-Aniston cell" or "totem-pole cell" are likely to be caricatures that oversimplify the nature of these neurons. The underlying computations are more complex and much less well understood than those of early visual neurons.

In a distributed but sparse code, different objects are represented by largely disjoint sets of cells. This may render the code robust to interference between objects. Interference of multiple simultaneously present objects (i.e., the superposition of their representations) could create ambiguity in a maximally distributed code. Interference could also erase memories: If each neuron is activated by many different objects, then spike-timing-dependent plasticity might wash away a memory that is not reactivated over a long time. Highly selective neurons, Thorpe argues, could maintain a memory over decades without the need of reactivation. Their high selectivity would protect them from interference. He suggests that the brain might contain neuronal "dark matter," that is, neurons so selective that they may not fire for years and are virtually impossible to elicit a response from in a neurophysiological experiment.

Sampling Limitations: Few Stimuli, Few Response Channels

With current techniques, our chances are slim to activate neuronal "dark matter" or to ever find the other loves of the "Jennifer-Aniston cell." This reminds us of a basic challenge for our field: our limited ability to sample brain responses to visual stimuli. High-resolution imaging and advances in multi-electrode array recording have greatly increased the amount of information we can acquire about brain-activity patterns. However, our measurements will not fully capture the information present in neuronal activity patterns in the foreseeable future. The subsample we take always consists in a *tiny* proportion of the information that would be required

to fully describe the spatiotemporal activity pattern in a given brain region. Electrode recording and fMRI tap into population activity in fundamentally different ways (which we discuss further at the end of this overview). fMRI gives us a temporally smoothed and spatially strongly blurred (and locally distorted) depiction of activity (i.e., the hemodynamic response), with a single voxel reflecting the average activity across hundreds of thousands of neurons (and possibly other cell types). Neuronal recording gives us spatiotemporally precise information, but only for a vanishingly small subset of the neurons in the region of interest (and possibly biased toward certain neuronal types over others). In terms of information rates, fMRI and electrode recording are similarly coarse: An fMRI acquisition might provide us with, say, 100,000 channels sampled once per second, and an electrode array can record from, say, 100 channels sampled 1,000 times per second.

We subsample not only the response space but also the stimulus space. Typical studies only present hundreds of stimuli (give or take an order of magnitude). In fMRI, the stimuli are often grouped into just a handful of categories; and only category-average response patterns are analyzed. However, to characterize the high-dimensional continuous space of images, a much larger number of stimuli is needed. Consider a digital grayscale image defined by 64 × 64 pixels (4,096 pixels) with intensities ranging from 0 to 255 (a pretty small image by today's standards). The number of possible such images is huge: 256^{4096} ($\sim 10^{10,000}$). The more relevant subset of "natural" images is much smaller, but this subset is still huge and ill defined. To complicate matters, the concept of "visual object" is inherently vague and implies the prior theoretical assumption that scenes are somehow parsed into constituent objects.

Repeated presentations of the same stimulus sample help distinguish signal from noise in the responses. Noise inevitably corrupts our data to some degree. The number of responses sampled limits the complexity of the models we can fit to the data. A model that is realistically complex, given what we know about the brain, is often unrealistic to fit, given the amount of data we have. To fit such a model would be to pretend that the data provide more information than they do, and generalization of our predictions to new data sets would suffer (see discussion in chapters 18 and 19 about bias versus variance). Both subsampling of the response pattern and limited model complexity cause us to underestimate the stimulus information present in a brain region's activity patterns. Our estimates are therefore usually lower bounds on the information actually present.

Retina: Rate Code Ruled Out

Sheila Nirenberg describes an interesting exception to the rule of lower bounds on activity-pattern information (chapter 2). She describes a study in which an *upper*

bound could be estimated. Neuronal recordings performed in vitro captured the continuous activity of the *entire* retinal population representing the stimulus. Nirenberg and colleagues then tested different hypothetical codes, each of which was based on a different set of features of the spike trains (thus retaining a different subset of the total information). Because the recordings arguably captured the full population information, any code that retained less information than present in the animal's behavior (as assessed in vivo) could be ruled out. Spike-rate and spike-timing codes did not have all the information reflected in behavior, whereas a temporal-correlation code did the trick.

Unfortunately, studies of cortical visual population codes are faced with a more complicated situation, where our limited ability to measure the activity pattern (a small sample of neurons measured or voxels that blur the pattern) is compounded by multiple parallel pathways. For example, current technology does not allow us to record from all the neurons in V1 that respond to a particular stimulus. Moreover, if a given hypothetical code (e.g., a rate code) suggested the absence in V1 of stimulus information reflected in behavior, the code could still not be ruled out, because the information might enter the cortex by another route, bypassing V1. The other studies reviewed in this book, therefore, cannot rule out codes by Nirenberg's rigorous method. When population activity is subsampled, absence of evidence for particular information is not evidence of absence of this information. The focus, then, is on the positive results, that is, the information that can be shown to be present.

Early Visual Cortex: Stimulus Decoding and Reconstruction

In chapter 3, Jasper Poort, Arezoo Pooresmaeili, and Pieter R. Roelfsema describe a study showing that physical stimulus features as well as attentional states can be successfully decoded from multiple neurons in monkey V1. They find that stimulus features and attentional states are reflected in separate sets of neurons, demonstrating that V1 is not just a low-level stimulus-driven representation. The results of Poort and colleagues illustrate a simple synergistic effect of multiple neurons that even linear decoders can benefit from: noise cancelation. Neuron A may not respond to a particular stimulus feature and carry no information about that feature by itself. However, if its noise fluctuations are correlated with the noise of another neuron B which does respond to the feature, then subtracting the activity of A from B (with a suitable weight) can reduce the noise in B and allow better decoding. Such noise cancelation is automatically achieved with linear decoders, such as the Fisher linear discriminant. Although the decoding is based on a linear combination of the neurons, the information in the ensemble of neurons does not simply add up across neurons and cannot be fully appreciated by considering the neurons one by one.

Like Poort and colleagues, Yukiyasu Kamitani (chapter 4) describes studies decoding physical stimulus properties and attentional states from early visual cortex. However, Kamitani's studies use fMRI in humans to analyze the information in visual areas V1–4 and MT+. All these areas allowed significant decoding of motion direction. Grating orientation information, by contrast, was strongest in V1 and then gradually diminished in V2–4; it was not significant in MT+. Beyond stimulus features, Kamitani was able to decode which of two superimposed gratings a subject is paying attention to.

These findings are roughly consistent with results from monkey electrode recordings. Their generalization to human fMRI is significant because it was not previously thought that fMRI might be sensitive to fine-grained neuronal patterns, such as V1 orientation columns. The decodability of grating orientation from V1 voxel patterns is all the more surprising because Kamitani did not use high-resolution fMRI, but more standard $(3\text{mm})^3$ voxels. The chapter discusses a possible explanation for the apparent "hyperacuity" of fMRI: Each voxel may average across neurons preferring all orientations, but that does not mean that all orientations are exactly equally represented in the sample. If a slight bias in each voxel carries some information, then pattern analysis can recover it by combining the evidence across multiple voxels.

From decoding orientation and motion direction, Kamitani moves on to reconstruction of arbitrary small pixel shapes from early visual brain activity. This is a much harder feat, because of the need to generalize to novel instances from a large set of possible stimuli. In retinotopic mapping, we attempt to predict the response of each voxel separately as a function of the stimulus pattern. Conversely, we could attempt to reconstruct a pixel image by predicting each pixel from the response pattern. However, Kamitani predicts the presence of a stimulus feature extended over multiple stimulus pixels from multiple local response voxels. The decoded stimulus features are then combined to form the stimulus reconstruction. This multivariate-to-multivariate approach is key to the success of the reconstruction, suggesting that dependencies on both sides, among stimulus pixels and among response voxels, matter to the representation.

Early Visual Cortex: Encoding and Decoding Models

While Kamitani focuses on fMRI *decoding* models, the following two chapters describe how fMRI *encoding* models can be used to study visual representations. Kendrick N. Kay (chapter 5) gives an introduction to fMRI voxel-receptive-field modeling (also known as "population-receptive-field modeling"). In this technique, a separate computational model is fitted to predict the response of each voxel to

novel stimuli. Similar techniques have been applied to neuronal recording data to characterize each neuron's response behavior as a function of the visual stimulus. Kay argues in favor of voxel-receptive-field modeling by contrasting it against two more traditional methods of fMRI analysis: the investigation of response profiles across different stimuli (e.g., tuning curves or category-average activations) and pattern-classification decoding of population activity. He reviews a recent study, in which voxel-receptive-field modeling was used to predict early visual responses to natural images. The study confirms what is known about V1, namely that the representation can be modeled as a set of detectors of Gabor-like small visual features varying in location, orientation, and spatial frequency.

Kay's study is an example of a general fMRI methodology developed in the lab of Jack Gallant (the senior author of the study). Jack L. Gallant, Shinji Nishimoto, Thomas Naselaris, and Michael C. K. Wu (chapter 6) present this general methodology, which combines encoding (i.e., voxel-receptive-field) and decoding models. First, each of a number of computational models is fitted to each voxel on the basis of measured responses to as many natural stimuli as possible. Then the performance of each model (how much of the non-noise response variance it explains) is assessed by comparing measured to predicted responses for novel stimuli not used in fitting the model. The direction in which a model operates (encoding or decoding) is irrelevant to the goal of detecting a dependency between stimulus and response pattern (a point elaborated upon by Marieke Mur and Nikolaus Kriegeskorte in chapter 20). However, Gallant's discussion suggests that the direction of the model predictions should match the direction of the information flow in the system: If we are modeling the relationship between stimulus and brain response, an encoding approach allows us to use computational models of brain information processing (rather than generic statistical models as are typically used for decoding, which are not meant to mimic brain function). The computational models can be evaluated by the amount of response variance they explain. Decoding models, on the other hand, are well suited for investigating readout of a representation by other brain regions and relating population activity to behavioral responses. For example, if the noise component of a region's brain activity predicts the noise component of a behavioral response (e.g., categorization errors; see chapter 14), this suggests that the region may be part of the pathway that computes the behavioral responses.

Midlevel Vision: Curvature Representation in V4 and Posterior IT

Moving up the visual hierarchy, Anitha Pasupathy and Scott L. Brincat (chapter 7), explore the representation of visual shapes between the initial cortical stage of V1 and V2 and higher-level object representations in inferior temporal (IT) cortex. At this intermediate level, we expect the representational features to be more complex

than Gabor filters or moving edges, but less complex than the types of features often found to drive IT cells. Pasupathy and Brincat review a study that explores the representation of object shape by electrode recordings of single-neuron responses to sample stimuli from a continuous parameterized space of binary closed shapes. Results suggest that a V4 neuron represents the presence of a particular curvature at a particular angular position of a closed shape's contour. A posterior IT neuron appears to combine multiple V4 responses and represent the presence of a combination of convex and concave curvatures at particular angular positions. The pattern of responses of either region allowed the decoding of the stimulus (as a position within the parameterized stimulus space). This study nicely illustrates how we can begin to quantitatively and mechanistically understand the transformations that take place along the ventral visual stream.

What Aspect of Brain Activity Serves to "Represent" Mental Content?

When we analyze information represented in patterns of activity, we usually make assumptions about what aspect of the activity patterns serves to represent the information in the context of the brain's information processing. A popular assumption is that spiking rates of neurons carry the information represented by the pattern. While there is a lot of evidence that spike rates are an important part of the picture, experiments like those Nirenberg describes in chapter 2 show that we miss functionally relevant information if we consider only spike rates.

Conor Houghton and Jonathan Victor (chapter 8) consider the general question of how we should measure the "representational distance" between two spatiotemporal neuronal activity patterns. In a theoretical chapter at the interface between mathematics and neuroscience, they consider metrics of dissimilarity comparing activity patterns that consist in multiple neurons' spike trains. The aim is to find out which metric captures the functionally relevant differences between activity patterns. Houghton and Victor focus on "edit distances" (including the "earth mover's distance"), which measure the distance between two patterns in terms of the "work" (i.e., the total amount of changes) required to transform one pattern into another. Jonathan Victor had previously proposed metrics to characterize the distance between single-neuron spike trains. Here this work is extended to populations of neurons, suggesting a rigorous and systematic approach to understanding neuronal coding.

Inferior Temporal Cortex: A Map of Complex Object Features

Moving farther down the ventral stream, Hans P. Op de Beeck discusses high-level object representations in inferior temporal (IT) cortex in the monkey and in the

human (chapter 9). This is the first chapter to review the findings of macroscopic regions selective for object categories (including faces and places). Face-selective neurons had been found in monkey-IT electrode recordings decades earlier. However, the clustering of such responses in macroscopic regions found in consistent anatomical locations along the ventral stream was discovered by fMRI, first in humans and later in monkeys. It has been suggested that these regions are "areas" or "modules," terms that imply well-defined anatomical and functional boundaries, which have yet to be demonstrated.

The proposition that the higher-level ventral stream might be composed of category-selective (i.e., semantic) modules sparked a new debate about localist versus distributed coding within the fMRI community. The new debate in fMRI concerned a larger spatial scale (the overall activation of entire brain regions, not single neurons) and also a larger representational scale (the regions represented categories, not particular objects). Nonetheless, the theoretical arguments are analogous at both scales. Just like the functional role of highly selective single neurons remains contentious, it has yet to be resolved whether the higher ventral stream consists of a set of distinct category modules or a continuous map of visual and/or semantic object features.

Op de Beeck argues that the finding of category-selective regions might be accommodated under a continuous-feature-map model. He reviews evidence suggesting that the feature map reflects the perceptual similarity space and subjective interpretations of the visual stimuli, and that it can be altered by visual experience.

Chou Hung and James DiCarlo (chapter 10) describe a study in which they repeatedly presented seventy-seven grayscale object images in rapid succession (a different image every 200 ms) while sequentially recording from more than three hundred locations in monkey anterior IT. The images were from eight categories, including monkey and human faces, bodies, and inanimate objects.

Single-cell responses to object images have been studied intensely for decades, showing that single neurons exhibit only weak object-category selectivity and limited tolerance to accidental properties. From a computational perspective, however, the more relevant question is what information can be read out from the neuronal population activity by downstream neurons. Single-neuron analyses can only hint at the answer. Hung and DiCarlo therefore analyzed the response patterns across object scales and locations by linear decoding. This approach provides a lower-bound estimate (as explained above) on the information available for immediate biologically plausible readout.

The category (among 8) and identity (among 77) of an image could be decoded with high accuracy (94 percent and 70 percent correct, respectively), far above chance level. Once fitted, a linear decoder generalized reasonably well across sub-

stantial scale (2 octaves) and small position changes (4 deg visual angle). The decoder also generalized to novel category exemplars (i.e., exemplars not used in fitting), and worked well even when based on a 12.5-ms temporal window (capturing just 0–2 spikes per neuron) at 125-ms latency. Category and identity information appeared to be concentrated in the same set of neurons, and both types of information appeared at about the same latency (around 100 ms after stimulus onset, as revealed by a sliding temporal-window decoding analysis). Hung and DiCarlo found only minimal task and training effects at the level of the population. This is in contrast to some earlier studies, which focused on changes in particular neurons during more attention-demanding tasks. From a methodological perspective, Hung and DiCarlo's study is exemplary for addressing a wide range of basic questions, by applying a large number of well-motivated pattern-information analyses to population response patterns elicited by a set of object stimuli.

Representational Similarity Structure of IT Object Representations

Classifier decoding can address how well a set of *predefined* categories can be read out, but not whether the representation is inherently organized by those categories. Nikolaus Kriegeskorte and Marieke Mur (chapter 11) review a study of the similarity structure of the IT representations of 92 object images in humans, monkeys, and computational models. Kriegeskorte and Mur show that the response patterns elicited by the ninety-two objects form clusters corresponding to conventional categories. The two main clusters correspond to animate and inanimate objects; the animates are further subdivided into faces and bodies. The response-pattern dissimilarity matrices reveal a striking match of the structure of the representation between human and monkey. In both species, IT appears to emphasize the same basic categorical divisions. Moreover, even within categories the dissimilarity structure is correlated between human and monkey. IT object similarity was not well accounted for by several computational models designed to mimic either low-level features (e.g., pixel images, processed versions of the images, features modeling V1 simple and complex cells) or more complex (e.g., natural image patch) features thought to reside in IT. This suggests that the IT features might be optimized to emphasize particular behaviorally important category distinctions.

In terms of methods, the chapter shows that studying the similarity structure of response patterns to a sizable set of visual stimuli ("representational similarity analysis") can allow us to discover the organization of the representational space and to compare it between species, even when different measurement techniques are used (here, fMRI in humans and cell recordings in monkeys). Like voxel-receptive-field modeling (see chapters 5 and 6, discussed earlier), this technique

allows us to incorporate computational models of brain information processing into the analysis of population response patterns, so as to directly test the models.

Andrew C. Connolly, M. Ida Gobbini, and James V. Haxby (chapter 12) discuss three virtues of studying object similarity structure: it provides an abstract characterization of representational content, can be estimated on the basis of different data sources, and can help us understand the transformation of the representational space across stages of processing. They describe a human fMRI study of the similarity structure of category-average response patterns and how it is transformed across stages of processing from early visual to ventral temporal cortex. The similarity structure in early visual cortex can be accounted for by low-level features. It is then gradually transformed from early visual cortex, through the lateral occipital region, to ventral temporal cortex. Ventral temporal cortex emphasizes categorical distinctions.

Connolly and colleagues also report that the replicability of the similarity structure of the category-average response patterns increases gradually from early visual cortex to ventral temporal cortex. This may reflect the fact that category-average patterns are less distinct in early visual cortex. Similarity structure was found to be replicable in all three brain regions, within as well as across subjects. Replicability did not strongly depend on the number of voxels included in the region of interest (100–1,000 voxels, selected by visual responsiveness).

The theme of representational similarity analysis continues in the chapter by Dwight J. Kravitz, Annie W.-Y. Chan, and Chris I. Baker (chapter 13), who review three related human fMRI studies of ventral-stream object representations. The first study shows that the object representations in ventral-stream regions are highly dependent on the retinal position of the object. Despite the larger receptive fields found in inferior temporal cortex (compared to early visual regions), these high-level object representations are not entirely position invariant. The second study shows that particular images of body parts are most distinctly represented in body-selective regions when they are presented in a "natural" retinal position—assuming central fixation of a body as a whole (e.g., right torso front view in the left visual field). This suggests a role for visual experience in shaping position-dependent high-level object representations. The third study addresses the representation of scenes and suggests that the major categorical distinction emphasized by scene-selective cortex is that between open (e.g., outdoor) and closed (e.g., indoor) scenes. In terms of methods, Kravitz and colleagues emphasize the usefulness of ungrouped-events designs (i.e., designs that do not assume a grouping of the stimuli a priori) and they describe a straightforward and very powerful, split-half approach to representational similarity analysis.

The representation of scenes in the human brain is explored further in the chapter by Dirk B. Walther, Diane M. Beck, and Li Fei-Fei (chapter 14). These authors investigate the pattern representations of subcategories of scenes (including mountains, forests, highways, and buildings) with fMRI in humans. They relate the confus-

ability of the brain response patterns (when linearly decoded) to behavioral confusions among the subcategories. This shows that early visual representations, though they distinguish scene subcategories, do not reflect behavioral confusions, while representations in higher-level object- and scene-selective regions do. In terms of methods, this chapter introduces the attractive method of relating confusions (a particular type of error) between behavioral classification tasks and response-pattern classification analyses, so as to assess to what extent a given region might contribute to a perceptual decision process.

In chapter 15, John-Dylan Haynes discusses how fMRI studies of consciousness can benefit from pattern-information analyses. A central theme in empirical consciousness research is the search for neural correlates of consciousness (NCCs). Classical fMRI studies on NCCs have focused on univariate correlations between regional-average activation and some aspect of consciousness. For example, regional-average activation in area hMT+/V5 has been shown to be related to conscious percepts of visual motion. However, finding a regional-average-activation NCC, does not address whether the specific content of the conscious percept (e.g., the direction of the motion) is encoded in the brain region in question. Combining the idea of an NCC with multivariate population decoding can allow us to relate specific conscious percepts (e.g., upward visual motion flow) to specific patterns of brain activity (e.g., a particular population pattern in hMT+/V5) in human fMRI. Beyond the realm of consciousness, we return to this point at a more general level in chapter 20, where we consider how classical fMRI studies use regional-average activation to infer the "involvement" of a brain region in some task component, whereas pattern-information fMRI studies promise to reveal a region's representational content, whether the organism is conscious of that content or not.

Vision as a Hierarchical Model for Inferring Causes by Recurrent Bayesian Inference

In chapter 16, the final chapter of the "Theory and Experiment" section, Karl Friston outlines a comprehensive mathematical theory of perceptual processing. The chapter starts by reviewing the theory of probabilistic population codes. A population code is probabilistic if the activity pattern represents not just one particular state of the external world, but an entire probability distribution of possible states. On one hand, bistable perceptual phenomena (e.g., binocular rivalry) suggest that the visual system, when faced with ambiguous input, chooses one possible interpretation (and explores alternatives only sequentially in time). On the other hand, there is evidence for a probabilistic representation of confidence. These findings suggest a code that is probabilistic but unimodal. Friston argues that the purpose of vision is to infer the causes of the visual input (e.g., the objects in the world that cause the light

patterns falling on the retina), and that different regions represent causes at different levels of abstraction. He interprets the hierarchy of visual regions as a hierarchical statistical model of the causes of visual input. The model combines top-down and bottom-up processing to arrive at an interpretation of the input. The top-down component consists in prediction of the sensory input from hypotheses about its causes (or prediction of lower-level causes from higher-level causes). The predicted information is "explained away" by subtracting its representation out at each stage, so that the remaining bottom-up signals convey the prediction errors, that is the component of the input that requires further processing to be accommodated in the final interpretation of the input. Friston suggests that perceptual inference and learning can proceed by an empirical Bayesian mechanism. The chapter closes by reviewing some initial evidence in support of the model.

In the second part of the book, "Background and Methods," we collect chapters that provide essential background knowledge for understanding the first part. These chapters describe the neuroscientific background, the mathematical methods, and the different ways of measuring brain-activity patterns.

A Primer on Vision

In chapter 17, Kendra Burbank and Gabriel Kreiman give a general introduction to the primate visual system, which will be a useful entry point for researchers from other fields. They describe the cortical visual hierarchy, in which simple local image features are detected first, before signals converge for analysis of more complex and more global features. In low-level (or "early") representations, neurons respond to simple generic local stimulus features such as edges and the cortical map is retinotopically organized, with each neuron responsive to inputs from a small patch of the retina (known as the neuron's "receptive field"). In higher-level regions, neurons respond to more complex, larger stimulus features that occur in natural images and are less sensitive to the precise retinal position of the features (i.e., larger receptive fields). The system can be globally divided into a ventral stream and dorsal stream, where the ventral "what" stream (the focus of this book) appears to represent what the object is (object recognition) and the dorsal "where" stream appears to represent spatial relationships and motion.

Tools for Analyzing Population Codes: Statistical Learning and Information Theory

Jed Singer and Gabriel Kreiman (chapter 18) give a general introduction to statistical learning and pattern classification. This chapter should provide a useful entry

point for neuroscientists. Statistical learning is a field at the interface between statistics, computer science, artificial intelligence, and computational neuroscience, which provides important tools for analysis of brain-activity patterns. Moreover, some of its algorithms can serve as models of brain information processing (e.g., artificial neural networks) or are inspired by the brain at some level of abstraction. A key technique is pattern classification, where a set of training patterns is used to define a model that divides a multivariate space of possible input patterns into regions corresponding to different classes. The simplest case is linear classification, where a hyperplane is used to divide the space. In pattern classification as in other statistical pursuits, more complex models (i.e., models with more parameters to be fitted to the data) can overfit the data. A model is overfitted if it represents noise-dominated fine-scale features of the data.

Overfitting has a depressing and important consequence: a complex model can perform worse at prediction than a simple model, even when the complex model is correct and the simple model is incorrect. The complex correct model will be more easily "confused" by the noise (i.e., overfitted to the data), while the simple model may gain more from its stability than it loses from being somewhat incorrect. This can happen even if the complex model subsumes the simple model as a special case. The phenomenon is also known as the bias-variance tradeoff: The simple model in our example has an incorrect bias, but it performs better because of its lower variance (i.e., noise dependence). As scientists, we like our models "as simple as possible, but no simpler," as Albert Einstein said. Real-life prediction from limited data, however, favors a healthy dose of oversimplification.

In brain science, pattern classification is used to "decode" population activity patterns, that is, to predict stimuli from response patterns. This is the most widely used approach to multivariate analysis of population codes. Tutorial introductions to this method are given by Ethan Meyers and Gabriel Kreiman for neural data (chapter 19) and by Marieke Mur and Nikolaus Kriegeskorte for fMRI data (chapter 20). These chapters provide step-by-step guides and discuss the neuroscientific motivation of particular analysis choices.

Pattern analyses are needed to detect information interactively encoded by multiple responses. In addition, they combine the evidence across multiple responses, thus boosting statistical power and providing useful summary measures. The combination of evidence would be useful even if interactive information were absent. These advantages apply to both neuronal and fMRI data, but in different ways. Single-neuron studies miss interactively encoded information, and perhaps also effects that are weak and widely distributed. However, they can still contribute to our understanding of population codes within a brain region. Arguably, most of what we know about population codes today has been learned from single-neuron studies.

The single-voxel scenario is quite different, as discussed by Mur and Kriegeskorte. In addition to the hemodynamic nature of the fMRI signal and its low spatial resolution, single-voxel fMRI analyses have very little power because of the physiological and instrumental noise and because of the need to account for multiple testing carried out across many voxels. As we make the voxels smaller to pick up more fine-grained activity patterns within a region, we get (1) *more* and (2) *noisier* voxels. The combination of weaker effects and stronger correction for multiple tests leaves single-voxel analysis severely underpowered. Pattern-information analysis recovers power by combining the evidence across voxels. Classical fMRI studies have used regional averaging (or smoothing) to boost power. This approach enables us to detect overall regional activations at the cost of missing fine-grained pattern information. Regional-average activation is taken to indicate the "involvement" of a region in a task component (or in the processing of a stimulus category). However, the region remains a black box with respect to its internal processes and representations. The pattern-information approach promises to enable us to look into each region and reveal its representational content, even with fMRI.

Whether we use neuronal recordings or fMRI, we wish to reveal the information the code carries. If pattern classification provides above-chance decoding of the stimuli, then we know that there is mutual information between the stimulus and the response pattern. However, pattern classification is limited by the assumptions of the classification model. Moreover, the categorical nature of the output (i.e., predefined classes) leads to a loss of probabilistic information about class membership and does not address the representation of continuous stimulus properties. It would be desirable to detect stimulus information in a less biased fashion and to quantify its amount in bits.

Stefano Panzeri and Robin A. A. Ince (chapter 21) describe a framework for information theoretic analysis of population codes. Information theory can help us understand the relationships between neurons and how they jointly represent behaviorally relevant stimulus properties. If the neurons carry independent information, the population information is the sum of the information values for single neurons. To the extent that different neurons carry redundant information, the population information will be less than that sum. To the extent that the neurons synergistically encode information, the population information can be greater than the sum. The case of synergistic information was described earlier in the context of chapter 3: If neurons A and B share noise, but not signal, A can be used to cancel B's noise. Subtracting out the noise improves the signal-to-noise ratio and increases the information. Panzeri and Ince place these effects in a general mathematical framework, in which the mutual information between the stimulus and the population response pattern is decomposed into additive components, which correspond to the sum of the information values for single neurons and the synergistic offset

(which can be positive or negative and is further decomposed into signal- and noise-related subcomponents).

The abstract beauty of the mathematical concept of information lies in its generality. In empirical neuroscience, the necessarily finite amount of data requires us to sacrifice some of the generality in favor of stable estimates (i.e., to reduce the error variance of our estimates by accepting some bias). However, information theory is key to the investigation of population coding not only at the level of data analysis, but also at the level of neuroscientific theory.

What We Measure with Electrode Recordings and fMRI

The experimental studies described in this book relied on brain-activity data from electrode recordings and fMRI. We can analyze the response patterns from these measurement techniques with the same mathematical methods, and there is evidence that they suggest a broadly consistent view of brain function (e.g., chapter 11). However, fMRI and electrode recordings measure fundamentally different aspects of brain activity. Moreover, the two kinds of signal have been shown to be dissociated in certain situations. The final chapter by Philipp Berens, Nikos K. Logothetis, and Andreas S. Tolias (chapter 22) reviews the relationship between neuronal spiking, local field potentials, and the blood-oxygen-level-dependent (BOLD) fMRI signal, which reflects the local hemodynamic response thought to serve the function of adjusting the energy supply for neuronal activity.

Neuronal spikes represent the output signal of neurons. They are sharp and short events, and thus reflected mainly in the high temporal-frequency band of the electrical signal recorded with an invasive extracellular electrode in the brain. The high band (e.g., >600 Hz) of electrode recordings reflects spikes of multiple neurons very close to the electrode's tip (<200 micrometers away) and is known as the multi-unit activity (MUA).

The low temporal-frequency band (e.g., <200 Hz) of electrode recordings is known as the local field potential (LFP). Compared to the MUA, the LFP is a more complex composite of multiple processes. It appears to reflect the summed excitatory and inhibitory synaptic activity in a more extended region around the tip of the electrode (approaching the spatial scale of high-resolution-fMRI voxels). The LFP is therefore thought to reflect the input and local processing of a region, whereas the MUA is thought to reflect the spiking output. The LFP is also more strongly correlated with the BOLD fMRI signal than the MUA. Berens and colleagues describe what is currently known about the highly complex relationships among these three very different kinds of brain-activity measurement.