M. J. KAHANA, Y. EZZYAT, T. D. PHAN, C. T. WEIDEMANN,

J. JACOBS

ELECTROPHYSIOLOGY OF HUMAN MEMORY

Contents

Preface

Advances in our ability to relate brain function to human behavior have radically transformed the study of human memory and cognition. Although handbooks and scholarly reviews have surveyed much of this work, I have long sought a textbook to teach students about the electrophysiology of human memory in a methdologically rigorous manner. In the present work, my coauthors and I have endeavored to introduce this field to both students and fellow researchers who want to gain an appreciation of both the methods and the results in this burgeoning field. We have limited our scope to the study of memory, which has itself become a huge discipline, but we recognize that one could have easily written multiple chapters covering allied topics such as perception, attention, language and reasoning.

In defining an organization for this material, we tried to keep the student in mind. After introducing memory and electrophysiology in the two foundational chapters, we work through methods of increasing, weaving both applications and worked problems into each chapter. We try to carefully explain each method before reviewing the articles that use that method. Because papers now employ mixed methodologies, we occasionally introduce methods briefly that we later cover in greater detail. By adding extensive side notes, we try to provide ancillary background without disrupting the overall flow of the text.

The present book was initially conceived through a reading group organized by trainees at the University of Pennsylvania in the summer of 2012. Eventually, it resulted in the development of a course entitled “Big Data, Memory and the Human Brain”. Originally we developed the methods using Matlab, but after a hiatus of a few years during which I was focused on the DARPA Restoring Active Memory project, we migrated the material to Python, benefiting from the extensive and freely-available libraries for advanced numerical processing available in that language. Several current and former trainees in the Computational Memory Lab at the University of Pennsylvania contributed to the design of this text, including

Dr. Nicole Long (U. of Virginia), Dr. Jeremy Manning (Dartmouth), Dr. Brad Lega (UTSW), Dr. Kareem Zaghloul (NINDS), Dr. Karl Healey (Michigan State University), Dr. John Burke (U. of Oklahoma), and Dr. Jim Kragel (University of Chicago). My coauthors and I initially divided major responsibilities for different chapters, but in the current version you can see evidence of each of our hands throughout the entire text.

This book is dedicated to our families . . .

Michael Kahana
Philadelphia, PA
March 2023

1
Paradigms, Principles, and Processes

A brief history

The study of human memory has a long and rich history spanning the wisdom literatures of ancient civilizations and continuing through the philosophical tradition of the British empiricist and associationist movements (Yates, 1966). The scientific study of human memory began in Germany in the late 19th century with the classic studies of Hermann Ebbinghaus (1850-1909) and George Elias Müller (1850-1934). These early studies demonstrated the power of experimental design and quantitative measurement in the study of memory, and introduced the classic "list learning" methodology, which has provided the basic experimental framework for the study of memory until this day (Ebbinghaus, 1885/1913; Müller & Pilzecker, 1900). Scientific research on human learning and memory quickly spread to the United States, where new laboratories sprung up at Harvard, Princeton, Cornell and Clark Universities. As in any new science, the early work was largely descriptive, aiming to characterize and quantify the basic data obtained from different experimental techniques (S. E. Newman, 1987).

Subsequent decades saw the emergence of several prominent "schools of psychology," each taking a different theoretical approach to the analysis of memory. The Gestalt school emphasized the role of mental representation and the organization (and reorganization) of knowledge as critical aspects of learning and memory (Köhler, 1947; Katona, 1951), whereas the associationist school emphasized the power of experience to create new webs of associations among pre-existing representations (Robinson, 1932; McGeoch, 1942). Influenced by the study of animal learning, many theorists emphasized the importance of reward-seeking behavior as the driver of learning and downplayed the idea that memory was anything more than a means of enabling animals to obtain rewards through action (Keller & Schoenfeld, 1950; McGeoch & Irion, 1952).

In the two decades following World-War II, information theory and the rise of computing machines provided a powerful new framework for thinking about memory (and cognition more generally) as a result of sequential stages of information processing. Here the idea is that the brain is a computing device that transduces sensory inputs into patterns of electrical activity and processes these patterns through sequential but possibly overlapping stages with the goal of achieving some behavioral objective. This new framework eventually transformed the field, sparking a "cognitive revolution," in

which theorists began to think of memory as one central part of a computing machine whose "software" could be decoded through analyses of accuracy and response time in well controlled experiments. Scientists could now peer into the "black box" (a term used to describe a computing machine whose inner mechanisms were beyond analysis) and hypothesize the underlying machinery supporting memory and cognition.

In parallel with the aforementioned developments in the cognitive analysis of human memory, scientists studying the biological mechanisms of the brain in animals developed techniques for measuring electrical signals in the brain while animals were engaged in complex behaviors. These developments, which will be further discussed in Chapter 2, set the stage for new theories of memory based on the idea that nerve cells (rather than microprocessors) were doing the computing operations supporting memory and cognition. This, in turn, led to the emergence of "connectionist" or "neural network" models of memory and learning in the 1980s, and the subsequent emergence of the field of computational neuroscience in the 1990s.

The late 20th century saw the emergence of the personal computer. This technological advance had transformative effects on both the experimental and theoretical analysis of human memory. Personal computers quickly filled psychology laboratories around the world, enabling students to rapidly design and deploy experiments that precisely controlled the presentation of lists of words or pictures and allowed researchers to easily collect a rich array of data on behavior, including the measurement of response time, mouse clicks, etc. As the graphics capabilities and processing power of the computers improved, researchers began developing new types of experiments, including those based on video-game technology (desktop virtual reality) that enabled the study of spatial memory within realistic 3D-rendered environments. Computing power also helped to fuel the development of computational models designed to account for the rich experimental data obtained in these laboratory experiments. With the rise of the internet in recent decades, researchers have now uploaded their experimental laboratory to the "cloud," enabling simultaneous data collection from large numbers of people as they perform game-like cognitive tasks controlled within their web browsers. This, in turn, has allowed scientists studying memory to examine the behaviors of people who are not university students, thus greatly expanding the applicability of their research to the world's diverse population.

The present work aims to introduce students to the use of electrophysiology in the study of human memory. This methdodology has played a key role in cognitive neuroscience in recent decades. Whereas memory research during the 20th century primarily advanced through the analysis of overt behavior (either accuracy, response time, or both), 21st-century research is accelerating through the use of technologies that allow us to peer into the brain as humans study and retrieve information. This technology includes both measures of bloodflow as measured non-invasively using functional-magnetic resonance imaging, and measures of electrical activity that can be obtained either non-invasively using scalp-recorded electromagnetic signals (EEG and MEG) and invasively by studying patients who have electrodes implanted in their brain for the treatment of various disorders. To manipulate electrophysiology, researchers can now use a variety of invasive and non-invasive electrical stimulation methods. Both recording and stimulation techniques have led to advances in our theoretical understanding of how

humans learn and remember information, and they are also setting the stage for new therapies that may help treat disorders of the brain that affect human cognition.

Laboratory paradigms

The standard paradigm for the study of human memory, whether using sophisticated brain recording techniques, or internet-based memory games, still involves the presentation of sequences of to-be-remembered items (*memoranda*) which research participants (*subjects*) are asked to *recognize* or *recall*. Here we briefly describe the basic methodology used in recognition and recall experiments. In later sections we will introduce important variants of these methods as well as some other techniques.

Recognition

Consider a police lineup in which the witness to a crime is later faced with the task of judging whether a suspect in the crime is the actual perpetrator. This is an example of a recognition task. Someone has experienced some information (e.g., having seen a person commit a crime) and must later decide whether a test item (the appearance of a suspect in the lineup) matches the memory of the person seen committing the crime. Here, the judicial system would very much like to understand what factors influence witness accuracy and whether the circumstances of the lineup can be engineered to improve witness accuracy (e.g., Coloff and Wixted, 2019).

Daily life contains numerous circumstances that call upon our ability to make recognition judgments. When someone waves at you on the bike path or running trail you may ask yourself whether you know the person. If you forget to bring your shopping list to the store, you may walk the aisles trying to recognize particular items as being among those you planned to purchase.

In studying recognition memory, scientists would like to exert greater control over the conditions prevailing both when an item is first studied and when memory for the item is later tested. Thus, in a laboratory recognition task, participants study a list of items presented in a highly controlled manner. These items are usually words, but other nonlinguistic stimuli such as complex scenes may also be used. During the test phase of the experiment, participants are shown a mixture of items that were presented as part of the study list and items that were not presented as part of the study list. For each item they must indicate whether or not it was on the just-studied list (Figure 1.1 illustrates the basic recognition procedure).

Recognition-memory judgments can be made for any attribute of an item, not just its membership on the list. For example, one might ask people to judge which of two items was most recently presented, how many times a given item was presented, or whether an item was presented in a male or female voice. Recognition can also be tested in a multiple-choice format. In this case, two (or more) test items are shown and participants have to judge which appeared on the list. Rather than asking subjects to make binary choice decisions, we can ask subjects to make ratings of confidence, or even to rank a set of choices based on their likelihood of having been studied on a list.

Recall

Recall is a form of memory in which we actually remember information that is distinct from the test cue itself. In taking an essay exam, a student must recall previously studied information that is relevant to the specific question (and hopefully organize that information in a coherent and insightful manner). In meeting an old friend, one would hope to be able to recall their name, when you last met, and other shared experiences. We have all experienced the frustration of trying to recall something that we feel we should remember, but at a given occasion we are unable to do so (and then, when it is hardly needed, the information easily pops into mind!).

In the laboratory, we can study recall in a variety of different ways. After studying a list of words, participants may be asked to recall the items freely, in any order (*free recall*), in order of study (*serial recall*), or even backwards. Figure 1.1 illustrates the free-recall task. Participants may also be probed to recall specific items. For example, the fifth word in the list may be given as a cue, and the participant may be asked to recall the prior or the subsequent item (*probed recall*).

One of the most widely used recall tasks involves the study of paired items followed by a memory test in which one member of each pair is shown as a cue for recall of its mate. As an example, you might study randomly paired common nouns, such as *carriage–dog*, *tree–pen*, *ribbon–diamond*, *horse–house*, etc. At the time of the test, you might first be asked to recall the word paired with *pen*, then the word paired with *ribbon*, and so on. This task is alternatively called *paired associates* or *cued recall*. Pairs can be probed either in the forward direction (e.g., *ribbon* as a cue to recall *diamond*) or in the backward direction (e.g., *pen* as a cue to recall *tree*). Although one might expect forward recall to be easier than backward recall, research has shown that people's ability to recall an individual pair does not depend on the order in which the pair was studied (Kahana, 2002). Yet, despite this fact, people can remember which item was first and which item was second (Mandler, Rabinowitz, & Simon, 1981).

Although words are the most commonly used stimuli in recognition and recall tasks, one can also use nonlinguistic materials, such as abstract pictures or shapes. One reason why memory researchers have emphasized the use of linguistic materials is that adults are very likely to attempt to code confusable nonlinguistic materials, such as complex shapes, in a linguistic manner. In the case of recall tasks, it is also difficult to measure the recall (reproduction) of nonverbal materials such as pictures. On the other hand, perceptual imagery plays a critical role in the encoding of linguistic materials, and visual imagery can greatly facilitate many forms of learning (Paivio, 1986).

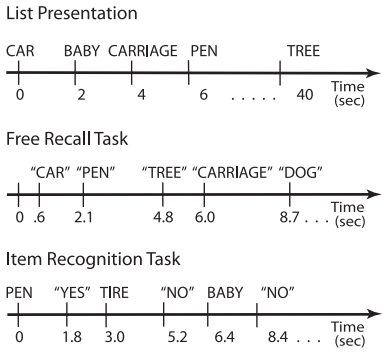


Figure 1.1: Recall and recognition tasks. Top: timing of item presentation. Middle: sample responses during a free-recall task. Bottom: sequence of test probes and *yes-no* responses in an item-recognition test. The words *pen* and *baby* appeared on the studied list, whereas the word *tire* had not. Quotes denote participants' responses.

Other paradigms

Although recognition, recall, and their variants are the methods most commonly used in the study of human memory, there are other important classes of memory paradigms that we will briefly introduce. Recognition and recall are often referred to as direct probes of memory because the subject is trying to consult their memory for the prior occurrence of an item. In other measures, memory may be proved indirectly, for example by having subjects perform a judgment task on an item that was previously encountered. For

example, in a *lexical decision* task subjects must decide, as quickly as possible, whether a sequence of letters is a valid word. Participants rarely make errors at this task, but they will respond more quickly to words that were recently encountered. In a *perceptual identification* task, participants see a degraded version of a word or picture that they have to identify. Again, identification is usually faster and more accurate if the item has recently been encountered, though there are important exceptions to this general observation (Huber, 2008). The direct–indirect distinction refers to the way in which memories are retrieved. In direct memory tests, participants are asked to recognize or recall an item as having occurred in a list of previously studied items. Indirect memory tests do not make reference to a previous study episode. Rather, they reveal the memorial consequences of the study episode on a person's ability to perform some other cognitive task, such as perceptual identification or fragment completion.

A related distinction applies to the study phase of a memory experiment. When presenting participants with a set of material for study, the experimenter may instruct them to *intentionally* learn the materials for a subsequent memory test. Alternatively, the material may be presented *incidentally* by using some cover story designed to make it less likely that participants will expect a memory test. For example, participants may be asked to rate the pleasantness of each word in a series or they may be asked to judge whether the words correspond to living or nonliving things. Although intentional learning is most often studied in the lab, incidental learning may be more typical of the operation of memory in our daily lives, where we experience a variety of information without necessarily trying to learn it for an upcoming test. Even if you are not trying to remember a specific detail of an experience, you may nonetheless store that detail in memory, and later, it may pop into your mind without any specific intention to retrieve it.

Principles

Certain very general principles govern the human memory system. Sometimes we might refer to these principles as "Laws", with the implication that a serious theory of memory should embody these principles, providing some deeper understanding or explanation of their generality and their interactions. These principles should also relate to neural mechanisms that would likely appear in measures of the brain's electrical activity, as discussed in subsequent chapters.

Interdependence of encoding and retrieval

Most memory tasks involve separate study and test phases. During the study phase, subjects *encode* new information into memory and during the test phase they *retrieve* the previously learned information. Researchers generally use the terms encoding and retrieval to refer to memory operations or processes, but sometimes for convenience they also use these terms to refer to the study and test phases of an experiment. However, this is not quite correct. As you can well imagine, while studying new information, subjects are also likely to retrieve prior knowledge about that information from memory and associate that prior knowledge with the new learning context. Similarly, when retrieving information from memory, subjects are also likely to be en-

coding the retrieved information into the test context. Indeed, research shows that the best way to learn something is to practice retrieving it (Karpicke & Roediger, 2008; Roediger, Putnam, & Smith, 2011). Similarly, successful learning frequently relies upon the retrieval of experiences associated with the to-be-learned item(s).

Recency

The principle or law of recency canonizes the observation that recent memories are more easily retrieved than memories of more remote experiences (Brown, 1824; Calkins, 1896). In his early studies, Ebbinghaus demonstrated that forgetting is rapid at first and then gradually slows. Figure 1.2 illustrates the forgetting function in three classic memory paradigms: paired-associate memory, free recall, and item recognition. In each of these three studies, forgetting is illustrated over the course of an experimental session, ranging from 10 minutes (A), to 40 minutes (B) to 70 minutes (C). In each case, we can see that forgetting continues over the course of the retention interval. In other experiments, people have studied forgetting effects over the course of many years, but under those circumstances, it is much harder to control for the mental activity of the subject during the retention interval.

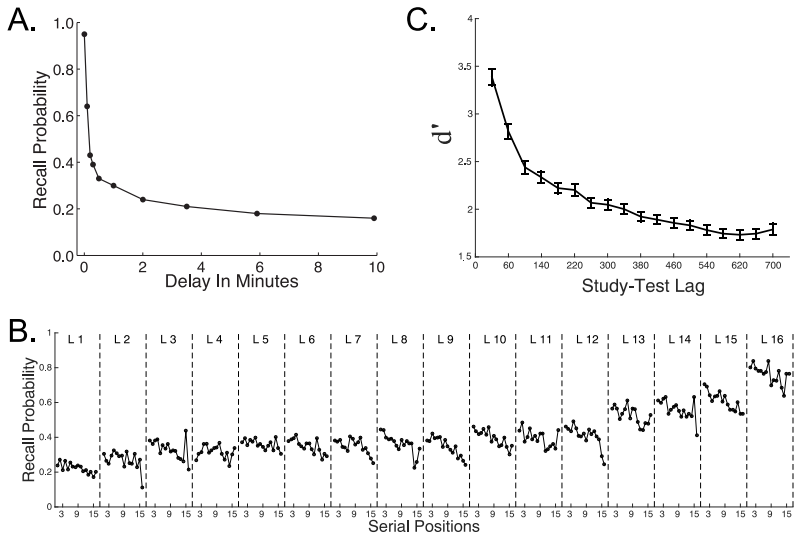


Figure 1.2: Forgetting in three classic memory paradigms. A. Memory for paired associates as a function of the amount of time (and number of pairs) intervening between study and test. Data from Rubin et al. (1999). **B.** Recall probability in final free recall as a function of both serial position and list position. More recent items appear towards the right. **C.** Recency in item recognition with performance measured using d', which is the z-normalized hit rate minus the z-normalized false alarm rate (Data from Experiment 1 of the PEERS study).

This seemingly obvious aspect of memory has generated more research and controversy than perhaps any other. One of the most important early findings was that the rate of forgetting depends critically on what happens between initial study and later memory tests (Kahana, 2012). Forgetting tends to be greatest when the interval between study and test (the *retention interval*) is filled with high levels of mental activity. In most experiments, forgetting of a given word pair is rapid because the retention interval is filled with studying and getting tested on other word pairs. In a classic study of forgetting over the course of sleep or waking activity, Jenkins and Dallen-

bach (1924) showed that subjects forget far more quickly when the retention interval is filled with waking activity than when subjects are asleep. Modern neuroscientific studies have focused considerable attention on the role of sleep in memory and an active debate currently centers on the question of whether sleep-related processes may actively help in the consolidation of memories learned during the day.

Short-term and long-term recency

The above figures illustrate recency on the scale of a single hour-long experimental session. Recency is also seen on a much faster scale by looking at serial position effects in recall of items within a single list. Here we find that the forgetting is very rapid. In the case of studying a list of 12 novel word-word pairs (paired-associate memory) if the last pair is tested immediately following study, it is almost always recalled correctly, and the next-to-last pair is recalled more than 50% of the time. Earlier pairs, however, are recalled less well, and the level of recall of these pairs decreases very slowly with increasing study-test retention intervals (Murdock, 1967). This can be seen for data from two different experiments, as shown in Figure 1.3.

The modality effect. The recency effect for paired associates is larger for auditorally presented lists than for lists that are presented visually. This *modality effect* is usually limited to the last 2 or 3 studied pairs (Murdock, 1972). The recency effect is also larger for auditorally presented items in other recall tasks, such as free and serial recall.

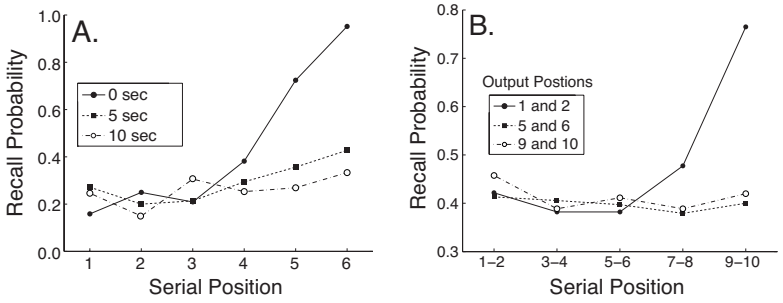


Figure 1.3: Short-term recency. People are very good at remembering the last couple of pairs in a list when they are tested immediately after study. **A.** Recency is greatly reduced after a delay of 5 sec (Murdock, 1967). **B.** Forgetting occurs during the recall phase itself. Items tested early in recall (output positions 1 and 2) show marked recency, whereas items tested late in recall (output positions 9 and 10) do not exhibit recency (Tulving & Arbuckle, 1963).

The Law of Primacy

When learning a series of items in a new context, subjects often exhibit superior memory for the first item in the series. This advantage often extends through to several early list items. We can readily observe this *primacy effect* in the free recall task, where subjects may recall the list items in any order. This is shown in Figure 1.4 for data obtained in the study of free recall. Subjects studied lists of 15 common nouns, which they subsequently recalled after a period of distraction. Here one can see a marked primacy effect in which recall is highest for the first list item and falls to a stable baseline by the fourth or fifth list item.

The classic explanation for primacy is that subjects who are highly motivated to recall as many items as possible share rehearsal time between the currently presented item and items presented in earlier serial positions (Brodie & Murdock, 1977; Tan & Ward, 2000). Experimental manipulations that promote rehearsal, such as providing subjects with longer interpresentation intervals, also enhance primacy. By contrast, conditions that discourage rehearsal, such as incidental encoding, fast presentation rates or interitem distraction, reduce primacy effects (Kahana, 2012).

The primacy effect is by no means specific to the free recall task. Primacy effects appear in delayed item recognition for both short and long lists. Paired-associate and probe-recall tasks also exhibit primacy effects, although the effects are often of smaller magnitude. Primacy effects also appear following event boundaries in both word lists and more naturalistic experiences As an example, (Polyn, Norman, & Kahana, 2009) asked subjects to study a list of twelve items in which they judged each item's size ("big" or "small") or pleasantness ("good" or "bad"). In some lists, subjects were asked to make the same judgment on every item. In others, subjects either made size judgments on the first six items and pleasantness judgments on the last six items, or vice versa. The lists in which subjects had to shift tasks midway through exhibited two primacy effects: a primacy effect for the beginning of the list, and an additional primacy effect after the task shift boundary. Thus, a change in the manner of item encoding resulted in improved memory for the first few items studied under the new encoding conditions.

While primacy effects exhibit near-universality across memory paradigms and subjects [1] , scientists are still debating the exact mechanisms underlying this phenomenon Although rehearsal of early list items constitutes an important factor underlying primacy in the free recall task, this mechanism cannot explain primacy observed under incidental learning conditions or in paradigms that make rehearsal very difficult. Enhanced recall for early list items would also result from a novelty-related boost in attention and/or reduce interference from items preceding a shift in context, as would be expected either at the beginning of a list of following an event boundary.

[1] **The vast majority of subjects exhibit a primacy effect** (Healey & Kahana, 2014).

Contiguity

Scholars throughout antiquity recognized that when we experience two items, *A* and *B*, in temporal succession, subsequently thinking of *A* appears to lead us to think of *B*, and vice versa. This recognition led to a rich body of theorizing about the association of ideas (e.g., Hume, 1739), and how these associative processes may explain the secrets of human cognition. Building upon this early philosophical tradition, Ebbinghaus (1885) sought to investigate associative processes experimentally through an examination of how much more quickly he could learn lists whose sequential structure was similar to previously learned lists. This early research led to the view that temporal contiguity was both necessary and sufficient for associative learning. Later scholars challenged these ideas by showing that without intention to learn the association between two contiguously experienced items, subjects exhibited very little evidence for associative learning (Thorndike, 1932; Hintzman, 2011). The problem, however, is that these studies may be simply demonstrating that the cognitive system can store more or less information, but it is still possible that any learning about two items will lead to some associative learning based on the contiguity principle.

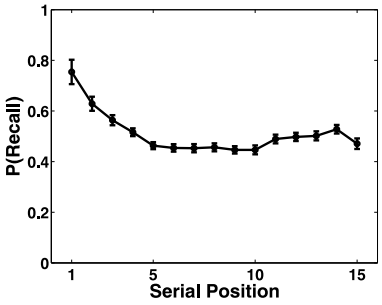


Figure 1.4: Primacy. Recall probability as a function of serial position. Error bars represent Loftus-Masson 95% confidence intervals (Loftus & Masson, 1994)

Whereas early research focused on explicit tests of associative memory, such as cued recall, more recent studies have used the free recall procedure to assess the role of contiguity in memory storage and retrieval. Because the order of recall reflects the order in which items come to mind, studies of recall transitions in free recall can help to reveal the organization of memory. Kahana (1996) asked how the probability of transitioning from an item studied in serial position *i* to an item studied in serial position *j* depends on the

$lag = j - i$) between the items.[2] This measure is called the *conditional-response probability as a function of lag*, or *lag-CRP*. Figure 1.5 shows lag-CRP functions obtained in three different variants of the free recall task: In immediate free recall (IFR), subjects begin recalling immediately following the final list item; in delayed free recall (DFR), subjects perform a demanding distractor task between the final list item and the recall period; in continual-distractor free recall (CDFR), subjects must perform a demanding distractor task following each and every list item.

This figure shows that in all three variants of the recall task, subjects make many more transitions among neighboring items than among items studied in more distant list positions. This is seen in the shape of the lag-CRP function, which decreases systematically as absolute lag increases, approaching an asymptotic value at moderate lags; the asymptotic value depends almost exclusively on list length, with lower asymptotic values for longer lists. The lag-CRP is also highly asymmetric, with transitions to neighbors being more likely in the forward than the backward direction. A final striking feature of the lag-CRP is the persistence of contiguity across time scales (Howard & Kahana, 1999). Although one might reasonably expect that requiring subjects to perform a demanding arithmetic task between items would sharply disrupt a subject's tendency to transition among neighboring items at retrieval, the data show otherwise; contiguity is preserved despite the disruption of the encoding process. Figure 1.6 illustrates the generality of the contiguity effect. Figure 1.6A-C shows that the contiguity effect appears robustly for both younger and older adults, for subjects of varying intellectual ability, and for both naïve and highly practiced subjects. Figure 1.6D-F shows that the contiguity effect also predicts confusions between different study pairs in a cued recall task, in errors made during probed recall of serial lists, and in tasks that do not depend on inter-item associations at all, such as picture recognition (see caption for details). Finally, long-range contiguity appears in many real-life memory tasks, such as recalling autobiographical memories (Moreton & Ward, 2010) and remembering news events (Uitvlugt & Healey, 2019).

[2] **To compute the conditional response probabilities**, one divides the frequency of transitions to a given lag by the possible transitions to that lag, excluding transitions that are outside of the bounds of the list or transitions to already recalled items. One can also do more sophisticated corrections for autocorrelations in goodness of encoding, as discussed more fully in (Healey, Long, & Kahana, 2019).

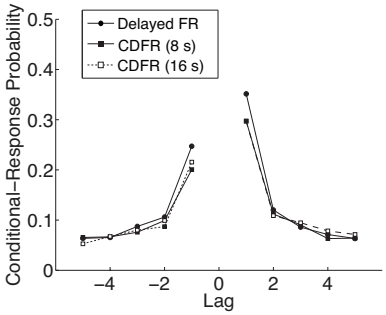


Figure 1.5: Contiguity in immediate, delayed and continual distractor free recall. The conditional-response probability as a function of lag exhibits a strong contiguity effect in both delayed and continual-distractor free recall (shown for distractors of 8 and 16-second duration). Positive values of lag correspond to forward recalls; negative values of lag correspond to backward recalls.

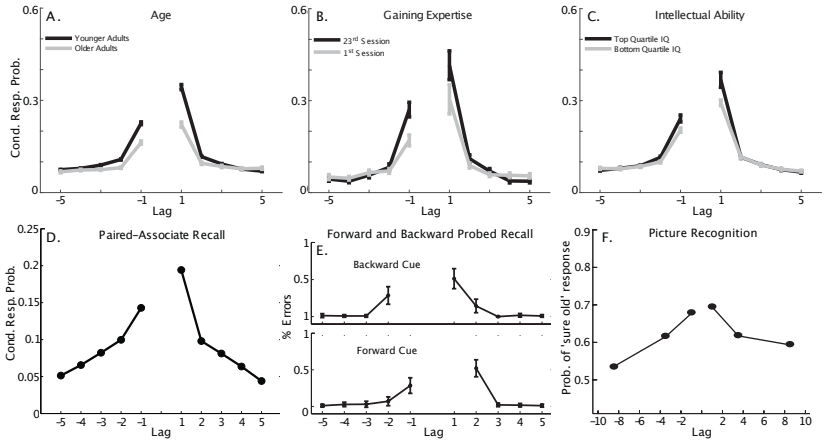


Figure 1.6: Universality of Temporal Contiguity. A. Older adults exhibit reduced contiguity, indicating impaired contextual retrieval. B. Massive practice increases contiguity effect (compare 1st and 23rd hour of recall practice). C. Higher-IQ subjects exhibit a stronger contiguity effect. D. The contiguity effect appears in conditional error gradients in cued recall, where subjects mistakenly recall items from pairs studied in nearby positions. E. When probed to recall the item that followed or preceded a cue item, subjects recall errors exhibit a contiguity effect both for forward and backward probes. F. The contiguity effect also appears in recognition memory. When successive test items come from nearby study list positions, subjects tendency to make successive high-confidence "old" responses exhibits a contiguity effect. See Healey et al (2019) for details.

Similarity

Whereas the contiguity effect illustrates the temporal organization of memories, it is also well known that subjects also make use of pre-existing semantic associations among list-items (Romney, Brewer, & Batchelder, 1993; Howard & Kahana, 2002b). This can be seen in people's tendency to make recall transitions among semantically related items, even in random-word lists that lack obvious semantic associates.

As one illustration of the *semantic similarity effect*, Figure 1.7 shows how the probability of making a recall transition among two items increases with their semantic relatedness.[3] This effect is evident even at low levels of semantic similarity when lists lack any strong associates or any obvious categorical organization, providing evidence that recall transitions are driven by the relative semantic strengths among the stored items (Howard & Kahana, 2002a; Howard, Addis, Jing, & Kahana, 2007; Long & Kahana, 2017).

As a further illustration of the law of similarity, we consider the everyday challenge of meeting a new person, hearing their name for the first time, and being able to subsequently recall their name when you next see them. Successful learning of name-face associations allows us to address people by their names, which is of great social significance. Sadly, many of us find it very difficult to perform these tasks without considerable effort, and even so, we often have difficulty remembering a name that once came to mind with great ease. Briefly reflecting on the challenge of learning to associate names and faces reveals several potential factors that make this task particularly difficult. First, associations between names and faces are very hard to elaborate in a meaningful way. Whereas with word pairs one can often come up with a verbal mediator or an interactive image, faces are far less amenable to such elaborations, and names are rarely related to faces in any meaningful way. Second, the attributes representing the appearance of a face are likely to differ from one encounter to the next. For example, the person may have styled his or her hair differently or may be wearing sunglasses or a hat. Here we consider a third factor that is less obvious but is theoretically quite interesting and important. More than virtually any stimuli studied in the laboratory, faces are extremely similar to one another. They all have the same basic shape, structure, organs, etc. When you consider a set of faces representing a single gender and age range and you remove all other superficial cues (e.g., hair styling, eyeglasses, etc.), it is amazing how visually similar they are. The fact that faces are structurally so similar means that for a given name-face pair, there will be other pairs whose faces are quite similar to that of the target pair. Pantelis, van Vugt, and Kahana (2007) asked whether such similarity relations could account for why some faces are significantly easier to learn than other faces.

To assess the similarities among synthetically generated faces, Pantelis and colleagues used a technique called *multidimensional scaling* (MDS, see, Steyvers, 2004, for a review of these methods). First, they used a behavioral study to assess the pairwise similarities among faces. MDS then uses the similarity matrix derived from these ratings (or other measurements) to create an N-dimensional vector representation of each face such that the distances between pairs of faces predicts the values in the similarity matrix. With MDS we begin with a one-dimensional solution and then move to progressively higher-dimensional solutions, asking at each step

[3] **Latent semantic analysis** assumes that words that are related in meaning tend to occur close together in texts. The method begins by taking a large corpus of text and counting the number of times that a given word i occurs in a given paragraph j. The resulting matrix, $L(i, j)$, has as many rows as there are words in the corpus and as many columns as there are paragraphs. A mathematical technique called *singular-value decomposition* is then used to transform the matrix in such a way as to reduce the number of columns while preserving the similarity structure among the rows. Semantic relatedness is measured by the cosine of the angle between vectors consisting of the entries in a particular pair of rows ($\cos\theta$). Completely unrelated words would have $\cos\theta \approx 0$, and strong associates would have $\cos\theta$ values between 0.4 and 1.0.

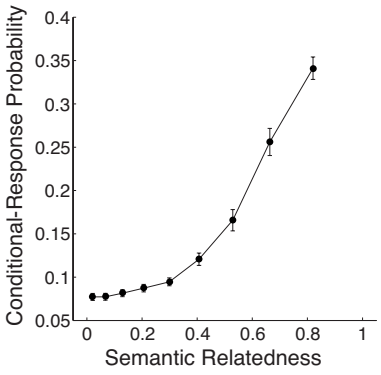


Figure 1.7: Semantic similarity effect. Subjects' recall transitions favor semantically similar items. Latent semantic analysis provided a measure of relatedness between 0 and 1 (Landauer and Dumais, 1997). Data from Kahana and Miller (2013).

whether the more complex model provided a substantial improvement in the model's ability to fit the similarity matrix. Pantelis et al (2007) found that a 4-dimensional solution (each face described by a set of 4 numbers) provided a very good fit to the similarity matrix. The output of the model was a vector representation of each face in a four-dimensional space, where each dimension roughly corresponded to one of the major attributes of the faces (e.g., the shape of the hairline; the width of the nose). Defining each face as a vector allowed them to calculate the similarities among any pair of faces.

Pantelis et al. hypothesized that people would have greater difficulty associating names with faces that had many "neighbors" in the four-dimensional face space (where neighbors are defined as the number of other faces that lie within a small radius around the target face). To test this hypothesis they asked participants to study eight faces that were paired with common American male names. At test, participants were presented with each of the eight studied faces, one at a time (see Figure 1.8A). As each face appeared, participants attempted to recall the name that was previously paired with the face. This study-testprocedure was repeated 10 times so that participants could learn the names of all of the faces.

When cued with a face at test, participants' accuracy at recalling the correct name decreased as the number of also-studied neighboring faces increased. RT data showed the inverse effect (see Figure 1.8B). Also, when participants made intralist intrusions, recalling a name paired with a different face on the target list, these errors tended to be names associated with faces that were similar to the target face (see Figure 1.8C).

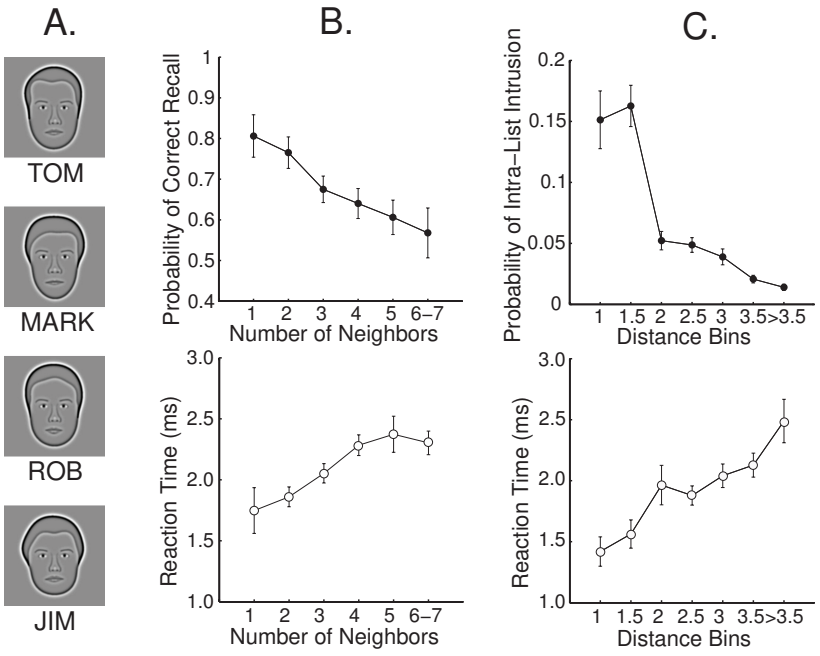


Figure 1.8: Similarity and memory for name-face associations. A. Examples of name-face pairs presented at study. **B.** Neighborhood Effect. The upper panel shows the probability of recalling the correct name when cued with a face at test, as a function of how many neighbors that face had within the study list. The lower panel shows response times for correctly recalling names of faces as a function of the number of neighbors. **C.** Each possible intrusion name corresponded with a study face, for which the distance from the cue face in four-dimensional face space was calculated. The upper panel shows the probability of making an intralist intrusion of a particular distance. The lower panel shows the response times for intralist intrusions of various distances from the cue face.

Repetition

Notwithstanding the well-known adage that "Practice makes perfect," the earliest scholars to study memory in the laboratory soon discovered that some forms of practice are better than others. Both Ebbinghaus (1885) and Müller and Schumann (1894) reported that repetition produced more rapid learning when the repeated trials were spaced apart as compared with when they were massed together.

Figure 1.9 illustrates the spacing effect in two classic recall paradigms: cued recall and free recall. Madigan (1969) had subjects study lists comprising a mix of once-presented (1P) and twice-presented (2P) items. The 2P items were either presented successively (spacing = 0), or with varying numbers of items separating the repetitions (2, 4, 8, 20, or 40). Immediately following presentation, subjects were asked to freely recall the words in any order. Figure 1.9A shows that increasing the separation of the 2P items improved their recall. Figure 1.9B illustrates a similar advantage for spaced practice in a paired-associate learning task. In this study, Glenberg (1976) had subjects study randomly-paired words with some pairs repeated at varying lags. He found an advantage for spaced practice that extended out to 20 or more intervening pairs.

Spacing effects have also been shown in real-world situations such as learning to type (Baddeley & Longman, 1978), learning foreign languages (Bahrick & Phelps, 1987), and remembering television commercials (Singh, Mishra, Bendapudi, & Linville, 1994). In the case of commercials, companies pay handsomely to present a brief message to a large audience in the hopes that the viewers will remember that message. Thus, it is not only important to produce a compelling message, it is also important to present the message in a way that maximizes viewers' memory for this incidentally encoded material. In a study by Singh et al. (1994), 413 people, including both young and older adults, viewed a television news program during which several commercials were presented. Some commercials were presented once, and some were repeated. The repeated commercials were separated by either one or four non-repeated commercials (spaced short and spaced long conditions). The researchers told participants that they were about to see a late-night news show taken from a network-affiliate station outside their own viewing area and that they were interested in knowing the participants' opinions about the news show. The following day, participants were brought back to the lab and tested on their memory for the commercials. They were given the name of a product category and asked to recall the brand name and claims made by the respective commercials. Participants' recall was 18% higher when the repeated commercials were spaced by four non-repeated commercials than when they were spaced by just one commercial. This advantage of spaced learning is one of the most robust and practical findings in the human memory literature.

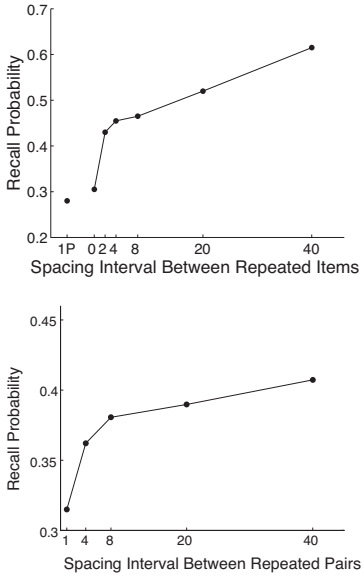


Figure 1.9: The beneficial effects of spaced practice. A. Probability of freely recalling words increases as a function of the spacing (lag) between their repeated presentations. 1P=once presented items (Madigan, 1969). **B.** Cued recall probability also increases as a function of the spacing of repetitions (Glenberg, 1976).

Processes

The foregoing analysis highlighted five very general principles of memory: the so-called laws of recency, primacy, contiguity, similarity and repetition. Any successful theory of the processes underlying memory encoding and retrieval must somehow embody these five principles.

Systems vs. Processes

Two major cognitive approaches predominate in the study of human memory. The first of these approaches, termed the *memory systems approach*, assumes that memory is not a singular entity, but rather is an umbrella term for a web of brain systems that each support different kinds of memory function. The goal of this approach is to determine what the different memory systems are and to identify the brain regions supporting each of these systems. Much as a biologist wants to classify the different types of organs that make up the body, the different types of cells that make up an organ, and the different types of processes that regulate cell function, memory researchers seek to identify the major subtypes of memory, the brain systems that give rise to them, and how these systems interact with one another. The second major cognitive approach involves the construction of computational models that describe the processes underlying memory function. Following this approach, researchers attempt to understand memory by specifying mathematical equations that characterize the encoding and retrieval processes underlying memory function. Having specified these equations, a computer is then used to solve the equations and thus generate predictions for what should happen in a memory experiment.[4] With these predictions in hand, scientists can ascertain where the model succeeds and where it fails to match the experimental data. These successes and failures give scientists ideas as to which of the model's assumptions are flawed, and how to fix them. Sometimes, the data needed to test a crucial aspect of a model do not exist. In these cases, an experiment must be designed to test the model. This interaction between data and models helps scientists create models of human memory that can account for a broader array of known data and that can also make better predictions about new memory phenomena that are yet to be observed.

During the past two decades a transformative development has taken place in the study of human memory. For the first century of memory research, models and theories of memory have been aimed at explaining experimental data on overt behavior, such as accuracy and RT in recall or recognition-memory tasks. Starting in the 1990s, a major emphasis in memory research has been to understand the biological mechanisms responsible for the encoding, storage, and retrieval of memories. In the next chapter we will turn to the biological foundations of human memory, which will set the stage for the analysis of electrophysiology through the rest of the book.

[4] In rare cases, one can solve the equations of a memory model analytically, as you would an algebra or calculus problem. However, most memory models are too complex to solve without the aid of a computer.

Representation and Search

A common assumption in many theories of memory is that our brains parse the incoming stream of experience into specific units which can be stored, retrieved, and combined into higher order units. This idea leads to the concept of *representation*—the idea that a given unit of memory can be represented by a mathematical object. Whereas early scholars conceived of each memory as an indivisible node in an associative network, subsequent scholars raised the possibility that each memory is actually a bundle of smaller units called elements, features, or attributes. We will refer to this as the *distributed representation hypothesis* (as distinct from a localist representation where a single unit represents a single complex memory). The distributed representation hypothesis can be traced back at least to the writings of the great

scientist/inventor Robert Hooke (1969), but these ideas also emerged in the work of philosopher/biologist Richard Semon (1923) and the writings of the early learning theorist Edwin Guthrie (1935).[5] The distributed representation recording a particular experience in memory is sometimes called the memory *trace* or, after Semon (1923), the *engram*. During the mid-20th century, mathematical learning theorists developed formal models of memory based on the ideas of distributed representations (e.g., Estes, 1955; Bower, 1972), and these models provided the foundation for important subsequent work on neural network theories of memory (McNaughton & Morris, 1987; Rumelhart, McClelland, & the PDP Research Group, 1986)

[5] Hintzman (2003) and Schacter (2001) provide fascinating discussions of the work of Hooke and Semon, respectively (see, also, Gomulicki [1953]).

For simple geometric forms (e.g., two-dimensional rectangles), the dimensions of items can be specified explicitly, and these physical dimensions often turn out to be the same as the psychological dimensions along which the items vary (e.g., Nosofsky, 1992; Kahana & Bennett, 1994). In the case of memorizing words or complex pictures, however, the theories assume that the items vary along a great number of physical and psychological dimensions whose identities may be difficult, or impossible, to fully identify. Latent Semantic Analysis, described above, provides one way of conceptualizing vector representations of words.

Another way to conceptualize the vectors representing complex stimuli is to think of them as characterizing the pattern of brain activity evoked by processing a given stimulus. Ultimately, all stimuli must be represented by the electrical activity of neurons in the brain. Some of these neurons may be very active, exhibiting a high firing rate, whereas others may be quiet. The firing rate of each neuron can be thought of as representing the value along the attribute coded by that particular cell (in reality a large number of neurons are likely involved in the coding of any attribute, not a single cell). Together, these cells can represent many kinds of information, including perceptual, contextual, and semantic aspects of an experience. Much as an image on a television screen is a pattern of brightness values distributed across the display, a memory may be thought of as a pattern of neural activation values distributed over a large array of neurons.

Assuming that some desired information is stored in memory, how does one find it? This *search problem* is particularly challenging in memory tasks where the to-be-remembered items are already well learned, such as the words in one's own language. In such tasks, one must still remember which words appeared on the list. To understand how people solve this search problem, one must not only specify the representation of the items and the processes governing the encoding of these representations in memory, but one must further specify how items are retrieved. Hypothesized retrieval mechanisms, which are central to all contemporary theories of memory, typically rely on the concept of association and the related concept of *cue-dependent retrieval*. In these theories of memory, different stored representations can evoke one another via associative connections formed both during the study of a target list and during one's prior experience with the studied items. These associations can be used to retrieve specific memories despite their being blended with many other non-targeted memories.

Multiple Traces

With distributed memories, each item is represented by a vector of values, one for each dimension or attribute, and the set of all items in memory can be thought of as an array where each row represents one dimension and each column represents a different item. Such an array of values is a *matrix*.

Here we introduce a simple yet powerful model for the learning process: a model in which each studied item lays down a new trace in a large and ever-growing array representing all of the traces stored in memory. This idea, which is referred to as the *multiple trace hypothesis*, implies that each encoded presentation of an item leaves its own memory trace (Hintzman, 1976; Moscovitch, Nadel, Winocur, Gilboa, & Rosenbaum, 2006). By allowing each studied item to lay down a unique trace, and by further assuming that each trace can consist of many attributes, we can easily accommodate the important idea that a given item (e.g., the word *cat*) will lay down somewhat different traces when studied on different occasions. Indeed, it seems strange to think that the encoding of a given word will be precisely the same at any two occasions. A much more natural assumption is that the stored attribute values vary based on the context in which a word occurred, such as the words that preceded it or the thoughts that it evoked.

The multiple trace hypothesis implies that the number of traces can increase without bound. Although the limits of information storage in the human brain are not currently known, it seems implausible for the brain to have an infinite storage capacity. The presence of an upper bound need not pose a problem for the multiple-trace hypothesis so long as traces can be lost/erased, similar traces can merge together, or the upper bound is large relative to the scale of human experience. We next consider how we can address the process of memory search in a model in which memories are stored in a matrix representation.

The matrix encompassing item-vectors $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$ is:

$$M = \begin{pmatrix} m_1(1) & m_2(1) & m_3(1) \\ m_1(2) & m_2(2) & m_3(2) \\ \vdots & \vdots & \vdots \\ m_1(N) & m_2(N) & m_3(N) \end{pmatrix}$$

Summed Similarity

When a person encodes a test item, we assume that it is converted into a vector representation, which can then be compared with all the vectors stored in the memory matrix. One way to recognize an item is to search the memory matrix, serially comparing the probe vector with each of the stored vectors. Although such a search process may be plausible for a very short list of items, it has difficulty accounting for the very rapid speed with which people can recognize test items drawn from long study lists.

As an alternative, one can search the memory matrix in *parallel*, comparing the probe item with each of the stored vectors at the same time. In carrying out such a parallel search, we would say *yes* if we found a perfect match between the test item and one of the stored memory vectors. However, if we required a perfect match to say *yes*, we may never say *yes* because a given item is likely to be encoded in a slightly different manner on any two occasions. Thus, when an item is encoded at study and at test, the representations will be very similar, but not identical. To circumvent this problem, we could accept partial matches so long as they exceeded some threshold of similarity. Alternatively, we could calculate the similarity for each comparison and sum these similarity values to determine the *global match* between the test probe and the contents of memory. Models that adopt this approach are called either *summed-similarity* or *global-matching* models.

Several versions of the summed-similarity approach have been proposed in the literature, and these models have proven to be quite successful in accounting for a wide range of recognition-memory data. Before we consider these models in greater detail, however, we need to address the issue of how to focus the search process on those items learned within a given context. In most memory experiments, this context is the most recently presented list.

Temporal Context

Laboratory studies have shown that changes in *situational context* between study and test can reduce performance on a range of memory tasks. These findings led to the view that contextual change is one of the major factors underlying forgetting (Carr, 1931; Hollingsworth, 1928; McGeoch, 1932; Robinson, 1932). Models of memory can account for this finding by allowing a subset of the attributes that represent an item to represent the situational context in which an item was learned, including its time and place of encoding. When context is not specifically manipulated in an experiment, by changing the location or other background attributes between study and test, we will use the term *temporal context* to refer to the contextual attributes that changed. Of course these attributes need not represent time per se; rather they represent any internal representation that accompanies our experience that varies with time, including background thoughts, affective states (i.e., are you happy or sad), and physiological variables (e.g., hunger, tiredness, anxiety).

By allowing for a dynamic representation of temporal context, items within a given list will have more overlap in their contextual attributes than items studied on different lists, or indeed items that were not part of an experiment (Bower, 1972; Crowder, 1976). If the contextual change between lists is sufficiently great, and if the context at the time of test is similar to the context encoded at the time of study, then recognition-memory judgments of the type described in the section on *Memory Paradigms* should largely reflect the presence or absence of the probe (test) item within the most recent (target) list, rather than the presence or absence of the probe item on earlier lists. This enables a multi-trace summed-attribute similarity model to account for many of the major findings concerning not only recognition memory tasks, but a number of related tasks involving memory judgments.

Summed-similarity Computations

Within the framework of attribute theory, two memories are identical if they share the same values along each attribute. Intuitively, the similarity of two vectors should decrease as the distance between them increases.[6] When two vectors are identical (i.e., the distance between them is zero), their similarity should be set to some maximal value. As the distance between the vectors increases toward ∞, their similarity should approach zero. The exponential decay function,

$$\text{similarity} = e^{-\tau \text{ distance}}$$

has exactly this property: it is equal to 1.0 when distance $= 0$ and it approaches zero as distance $\rightarrow \infty$. The variable τ (Tau) determines how quickly similarity decays with distance[7]. Armed with measures of similarity, we can sum the similarity between a test item and each of the stored vectors in

[6] The distance between two vectors, $\mathbf{m}_1$ and $\mathbf{m}_2$, is the length of the difference vector, $\mathbf{m}_1 - \mathbf{m}_2$, which is $\sqrt{\sum_{i=1}^{N} (m_1(i) - m_2(i))^2}$, where N indicates the number of attributes or dimensions.

[7] A related measure of similarity is the cosine of the angle between two vectors, $\cos\theta$. If the two vectors are identical, the angle will be zero and $\cos\theta = 1$. If the two vectors are perpendicular to one another, the angle will be 90° and $\cos\theta = 0$. If we were to select random values for each of the attributes, we would find that $\cos\theta = 0$ in expectation.

our memory matrix. Let us assume that a person says "yes, I remember that item" if the summed similarity exceeds a threshold value.

To formalize these arguments, suppose that $\mathbf{m}_i$ represents the ith item of an L-item list. Following study of the list, the matrix M would represent the list in memory.

$$M = \begin{pmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \mathbf{m}_3 & \dots & \mathbf{m}_L \end{pmatrix}$$

Now let $\mathbf{g}$ represent a test item, either a target (i.e., $\mathbf{g} = \mathbf{m}_i$ for some value of i) or a lure. The summed similarity between the test probe and the items stored in memory can be written as:

$$\sum_{i=1}^{L} \text{similarity}(\mathbf{g}, \mathbf{m}_i), \tag{1.1}$$

where the similarity between $\mathbf{g}$ and $\mathbf{m}_i$ is defined as:

$$e^{-\tau||\mathbf{g}-\mathbf{m}_i||} = e^{-\tau\sqrt{\sum_{j=1}^{N}(g(j)-m_i(j))^2}}. \tag{1.2}$$

The summed-similarity model dictates that a person will respond *yes* when the summed similarity between the test item and each of the stored items in memory (Equation 1.1) exceeds a threshold value C; otherwise the person responds *no*. The threshold value can be set to simultaneously maximize hits (correct *yes* responses to studied items) and minimize false alarms (incorrect *yes* responses to non-studied items).

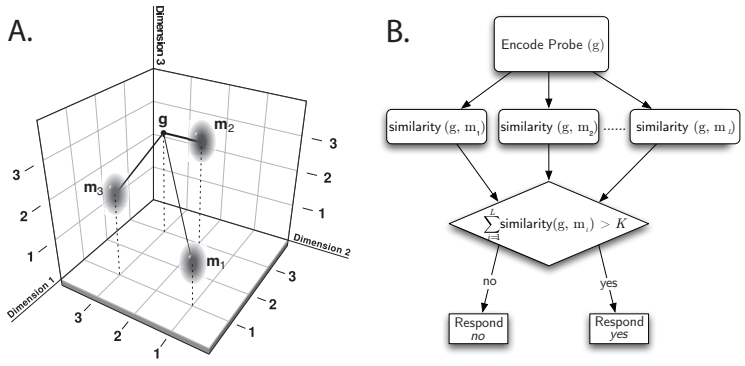


Figure 1.10: Illustration of the Summed Similarity Model. **A.** Schematic showing the memorial representations of three list items in a three-dimensional attribute space. Each item ($\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$) is depicted by a shaded ellipse representing the noise associated with the coding of the item. Also shown is the relatively noiseless representation of a nonstudied test probe ($\mathbf{g}$). **B.** Diagram of the information-processing stages in the summed similarity model. First, the test probe is encoded. Then, its similarity to the memorial representations of each list item is computed. If the summed similarity exceeds a threshold, C, the model responds *yes*.

Modeling Context Dynamics

Within the attribute similarity framework, we can account for the *Law of Recency* by assuming that the contextual attributes change slowly as each new item is presented (e.g., Estes, 1950, 1959). How might such a contextual-drift process work? Let us first denote the set of attributes representing context as the vector $\mathbf{t} = (t(1), t(2), \dots, t(N_{\text{context}}))$, where N_{context} is the number of attributes representing context. Now suppose that each time a new item is presented or remembered, we change the context vector by adding a small variable amount to each element. We can thus write down a simple Gaussian random-walk model for the context vector as:

$$\mathbf{t}_i = \mathbf{t}_{i-1} + \epsilon \tag{1.3}$$

where ϵ is a random vector whose elements are each drawn from a Normal, or Guassian, distribution, and where i is an index variable that counts each item presentation. Assuming that context changes gradually over the course of an experiment, the amount of change in context between the study of an item and its later test will increase with the number of items intervening between study and test. This is how context can be used to explain the recency effect. Recent targets will have higher summed similarity than remote targets (See Figure 1.11).

Summed-Similarity Summary

The summed-similarity framework provides a comprehensive account of many findings in the recognition memory literature. It embodies the principles of similarity, recency, and repetition mentioned above. The basic form of the model discussed here does not account for contiguity-based associations or the principle of primacy. More importantly, however, summed-similarity derives a single number (a scalar) to represent the output of the memory system. Such an approach fails to account for our ability to recall specific items or the cue-dependent nature of the recall process. The essential missing element is a process model of association. We next consider associative network models that account for both association and recall dynamics.

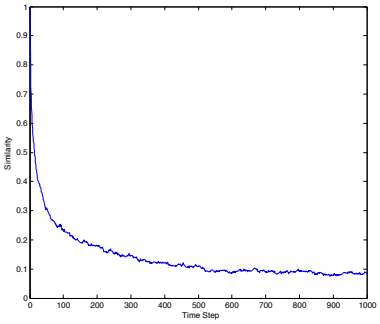


Figure 1.11: Contextual similarity decreases with each successively processed item. Similarity is defined as the $\cos\theta$ (**t**, **t**+τ) where τ is the number of time steps.

Neural Networks

A major advance in our understanding of associative memory took place in the 1970s when computer scientists, neuroscientists, and psychologists discovered mathematical methods for describing how associations could be represented in memory, and retrieved in a cue-dependent manner. These models of associative memory were inspired by the circuitry of the brain itself, where connections between neurons appear to play a crucial role in associative learning and recall. This class of models developed a strong following among scientists in diverse fields, and the research that grew out of this work is known as *connectionism* or *neural networks*.

Neural-network models assume that each neuron (or node) is connected to many other neurons. Together, these neurons (nodes) form a highly interconnected network. In our model system, each neuron can be characterized by its *activation* value, and each connection is characterized by its strength value, or *weight*. At a given time, the vector of activations for the neurons in the network is called the *state vector* of the network. The state vector, which can represent the attributes of a memory, is not static. It changes from moment to moment as our mind wanders or as new information is being experienced. So, although the state vector can represent a previously stored memory vector, it does not store any memories. Rather, memories are stored in the connections between neurons, and these connections enable the network to recover previously stored memories. The connection strengths are also not static. They change during learning to allow the state vector to recall previously learned memories. An important property of neural network models is that they can store a large number of distinct memories in the pattern of connections among the nodes of the network.

Network Dynamics

Each node, or neuron, in a neural network gets input from other nodes and sends output to other nodes. A scalar value, termed activation, describes the output of each node. The activation of node i, denoted $a(i)$, is related to the input that it receives from other nodes in the network. The input that node i receives from node j is the product of $a(j)$ and the weight of the connections between i and j, denoted $w(i,j)$. Putting this together, we can write:

$$a(i) = g\left(\sum_{j=1}^{N} w(i,j)a(j)\right) \tag{1.4}$$

where the function $g(\text{input})$ transforms the unit's input into its activation value. For now we will assume that the activation of a unit equals its input (i.e., $g(\text{input}) = \text{input}$). This type of simple neural network is called a *Linear Associator*.[8] Equation 1.4, proposed by McCullough and Pitts (1943), is called the *dynamical rule* of the network. This rule provides a simplified description of the behavior of our artificial neurons, whereby activations map onto the neurons' firing rates and weights map onto the strength of the synaptic connections between the neurons (see Figure 1.12).

[8] In modeling the behavior of a neural network, it is useful to define a neuron's activation as being either above or below its average activation level. This is done by defining an activation of zero as the average activation of a neuron: positive activation values would denote above-average activity and negative activation values would denote below-average activity.

Hebbian Learning

Hebb (1949) proposed that learning results from changes in the connection strengths, or weights, between neurons. Specifically, he hypothesized that the weight of the connection between neurons i and j (denoted $w(i,j)$) is increased by the product of their activity at each time-step t. Mathematically, we can write the Hebb learning rule as:

$$w(i,j)_t = w(i,j)_{t-1} + a(i)_t a(j)_t, \tag{1.5}$$

where $a(i)_t$ denotes the activation of neuron i at time t. Although it is convenient to assume that all of the weights are initially set to zero (when $t = 0$), one could also assume that the weights are initially set to some random values.

According to the Hebb rule, if neurons i and j both have above-average (i.e., positive) activations, their connection is strengthened. Similarly, if they both have below-average (i.e., negative) activations, their connection is also strengthened. But, if one neuron's activation is positive while the others is negative, their connection is weakened.

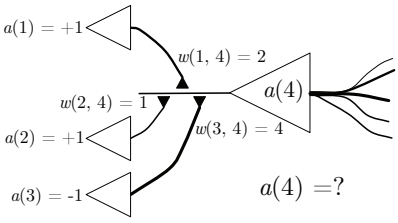


Figure 1.12: McCullough-Pitts model neuron. Neurons (open triangles) receive input from the left and send output to the right (this arrangement is arbitrary). The activation of Neuron 4, $a(4)$, is determined by the weighted sum of the inputs coming from Neurons 1–3 (in this example $g(-1) = -1$).

Modeling Recall

Consider two sets of neurons: one set represents the attributes of item A and the second set represents the attributes of item B. Let $a(i)$ denote the output of the ith unit representing A and let $b(j)$ denote the output of the jth unit representing B. (We again assume that items are vectors of attributes, but we now assume that each attribute value is coded by the firing rate of one of our simplified neurons). Suppose, further, that each neuron representing attributes of A is connected to each neuron representing attributes of B. We can now write the weight matrix as:[9]

$$w(i,j)_t = w(i,j)_{t-1} + a(i)_t b(j)_t. \tag{1.6}$$

[9] In matrix notation, the learning rule in Equation 1.6 is given by:

$$W_t = W_{t-1} + \mathbf{b}\mathbf{a}^\top$$

where $\mathbf{a}$ and $\mathbf{b}$ are column vectors.

To see how associative (cued) recall works, set the activations of the neurons representing A to their appropriate values and allow Equation 1.4 to determine the activations of the neurons representing item B. The output (activation) of the jth node of B is given by:

$$\tilde{b}(j) = \sum_{i=1}^{N} w(i,j)a(i) = \sum_{i=1}^{N} a(i)b(j)a(i) = b(j)\sum_{i=1}^{N} a(i)^2, \qquad (1.7)$$

where $\tilde{b}(j)$ is the new value of $b(j)$ after it has been updated by applying the dynamical rule. If we assume that each item vector, **a** and **b**, is of length one, then it is easy to show that $\tilde{\mathbf{b}} = \mathbf{b}$ because $\sqrt{\sum_{i=1}^{N} a(i)^2} = 1$.

Now if instead of just learning a single pair of items, we stored many pairs of items within the same neural network. It is easy to show that the network can still recall the correct item so long as all of the items are orthogonal vectors. Sparse, random vectors of high dimensionality will be very nearly orthogonal, so even a linear associator can accurately recover a good approximation of the target memory. If, however, you want to store and retrieve associations among similar items, as in the name-face study described above, you would need to move to a non-linear activation function to enable the network to iteratively converge to the correct target memory. This process if often called *deblurring*, and there are many types of networks that can achieve a high storage capacity for associative memories while still accurately recovering the target association.

Neural Network Summary

Inspired by the computational properties of actual neurons, neural network models allow memories to be evoked directly by a cue item without requiring a conventional search process. As such, the stored memories are referred to as being "content addressable." Neural networks store memories, defined as patterns of neural activity, in the strengths or weights of the connections between neurons. One of the most important and attractive features of neural networks is that they can store a multitude of memories in a single set of connections. As each new memory is experienced, the weights connecting the network's neurons are updated according to a learning rule. We focused on one such learning rule, known as Hebbian learning, whereby the weight between two neurons increases or decreases as a function of the product of the neurons' activations.

Rather than directly comparing a probe item with each of the stored memories, the retrieval (or recall) process in a neural network occurs as a natural process of the networks dynamics. When the network is not in a learning mode, each neuron's activity depends on the activity of the other neurons in the network, and the connection weights to those neurons. The simple dynamical rule illustrated in Figure 1.12 on the previous page and defined by Equation 1.4 on the preceding page allows the network to "recall" previously learned patterns.

As with summed-attribute-similarity models, one can incorporate contextual features into neural networks to distinguish between identical items encoded on distinct occasions or in different contexts. Assuming that context changes slowly over time, the network embodies the Law of Recency. However, without adding some additional assumptions or mechanisms, neu-

ral network models can not account for the beneficial effects of spacing on subsequent cued recall.

Organization of the Book

The present chapter introduced some of the major paradigms, principles and processes investigated by students of memory. List recognition and recall tasks, involving their distinct study and test phases, have fueled much of the theorizing on human memory and these tasks (and their variants) continue to occupy center stage. Subsequent chapters discuss other memory paradigms, including spatial learning and spatial memory, category learning, and probability learning. Citing data from recognition and recall tasks we documented five major laws of memory: recency, primacy, contiguity, similarity, and repetition. Finally, we surveyed some of the major process models of memory, including summed exemplar similarity models and connectionist network models. Absent from chapter 1 was any discussion of electrophysiology, which is the main focus of this book. The foundational background to the electrophysiological analysis of memory appears in Chapter 2, which introduces the subject of human electrophysiology.

Human electrophysiology is a big data affair. The size of a single dataset can often exceed one terrabyte. This book provides a hands-on tutorial to guide you through many of the key data science methods used in the analysis of human electrophysiological data. We begin by introducing the event-related potential technique, which allows researchers to identify electrophysiological signals that are time-locked to the presentation of a stimulus, or the execution of a motor response. Next, we introduce the reader to spectral analysis methods that have been used to uncover the role of oscillatory or rhythmic activity in human learning and memory. These techniques allow researchers to transform a complex time-varying signal from the time domain to the frequency domain and in doing so, relate signals such as an evoked potential to the alignment of oscillatory activity. We also discuss how spectral methods can be used to analyze non-oscillatory activity, which may also exhibit correlations with behavior. We next introduce multivariate approaches to analyzing electrophysiological data, focusing on regression and classification models that form the foundation of modern machine learning (ML). We show how these ML techniques have allowed researchers to decode the content of human thought, bringing extrasensory perception from fiction to reality. Chapter X then turns to the methods used in studying the functional connectivity underlying memory and cognition. Here we discuss spectral and statistical modeling methods used to assess functional connectivity. Subsequent chapters consider single-neuron recordings, brain stimulation and various modern data-analytic techniques.

2
Human Electrophysiology

Historical Background

Technological innovations that improve the measurement and quantitative analysis of data drives scientific progress. We can identify four developmental phases in the evolution of human electrophysiology that arose from specific technological advances. The first phase followed Hans Berger's famous discovery of the human electroencephalogram (EEG) in 1929. By applying electrodes to the scalp Berger was able to identify 10 Hz rhythmic activity overlying the human occipital cortex—the so-called alpha rhythm. Following the discovery of this so called alpha rhythm several other rhythms were identified, such as the 1-2 Hz (delta) rhythm associated with sleep and the 10-20 Hz rhythm overlying motor regions of the brain that increase during movement. By the late 1930s the use of the EEG had become widespread in clinical neurology and psychiatry, and many researchers began to explore the EEG correlates of both disease and normal cognitive function. For a fascinating and erudite analysis of research on brain rhythms, including many historical side notes, we refer the reader to Buzsaki (2004).

The emergence of the next major phase of electrophysiological research coincided with the widespread adoption of mainframe computing machines by major academic institutions. These computers made it feasible for researchers to analyze large electrophysiological datasets using statistical methods. (Psychophysiologists were early adopters of computing machines, such as the IBM 360 and the Digital Equipment Corporation PDP series of computers). Although scientists applied many data-analytic methods to EEG data, the evoked-potential method came to dominate research in this period, leading to the identification of time-locked modulations of the EEG, known as event-related potentials, or ERPs. One prominent example of these was the P300, a positive going ERP component appearing at around 300 milliseconds following an auditory stimulus. Chapter 3 describes the ERP method in detail, including discussion of the P300 and other components. The analysis of event-related potentials dominated the physiological study of human memory and cognition for several decades.

Whereas the second phase of research involved moving psychophysiological research out of the neurological clinic and into the psychological laboratory, the third phase involved a return to the clinic where researchers could gain access to invasive electrical recordings taken from neurosurgical patients being treated for drug-resistant epilepsy. Although researchers had

Figure 2.1: The discoverer of the human electroencephalogram, Hans Berger (1873 – 1941) was a professor of neurology at Jena, in Germany.

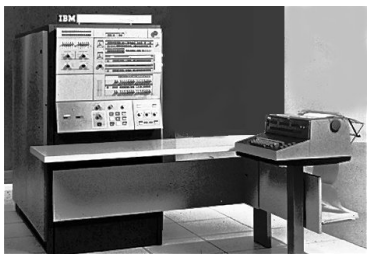


Figure 2.2: Photograph of the IBM360 mainframe computer.

been studying neurosurgical patients with epilepsy throughout the earlier periods, research in the clinic was not for the faint of heart. Until the late 1990s, EEG systems being used in clinics relied on archaic analog recording systems and lacked any convenient means of synchronizing between the recordings taken and the computers controlling cognitive tasks. Developers of these proprietary clinical systems did not aim to make their data easily readable or accessible to scientists. Few psychophysiologists were able to conduct research under these circumstances and the vast majority of data were therefore generated using scalp-EEG recording systems optimized for basic research and employed in studies of healthy individuals.

The third phase of human electrophysiological research involved a return to the clinical roots of the field, with researchers setting up controlled experiments in the epilepsy clinics to gain access to precious electrical recordings from deep brain structures, such as the hippocampus. These invasive brain recordings led to a re-evaluation of the rhythmic components of EEG signals. These developments served the critical role of bridging human and animal work on the electrophysiology of memory, leading to greater dialogue between neurobiologists studying memory in rodent models and cognitive neuroscientists studying memory in humans.

This third phase of research would not have been possible were it not for important technological developments in personal computing. Although the PC revolution began in the 1980s, it took another decade or so before these desktop machines replaced mainframe systems for most scientific computing needs. By the mid 1990s, advances in computer speed and the emergence of scientific computing platforms, such as MATLAB, allowed researchers to perform sophisticated quantitative analyses of EEG data with relative ease[1]. Researchers could use these new software tools instead of relying on proprietary systems mostly tailored to conducting ERP analyses. This in turn led to a dramatic increase in the sophistication of analytic methods, and in particular to a surge of interest in spectral analysis methods that we will discuss in Chapter 4. These spectral methods allowed researchers to return to the analysis of rhythmic brain activity with far greater rigor than was possible in the first half of the 20th century.

At the dawn of the 21st century there were dozens of research groups collecting and analyzing direct brain recordings in epilepsy centers around the world. These recordings included not only field potentials, but also signals from individual neurons, and researchers became very busy examining how these brain signals correlated with a host of perceptual and cognitive variables. This period also saw a number of striking demonstrations of conserved neural signals between animals and humans, particularly in the realm of spatial exploration. Researchers studying non-invasive scalp EEG also embraced spectral methods and the tools needed to perform these analyses were rapidly developed and freely disseminated.

Human brain recordings produce immense datasets, and the ability to make statistical inference on these data sets requires that we build multivariate models while avoiding overfitting our data. Although statistical methods for solving these problems have been known for decades, it is only in recent years that we have had the computing power to deploy these "machine learning" methods on large datasets. A principal goal of these methods is to ask whether the brain has information about some behavior or cognitive state, such that we can use the neural signals to make predictions in an indepen-

[1] EEG data form a series of voltage time series measured from different locations in the human brain or on the scalp. Spectral analysis methods may be used to measure both the presence of oscillations within a single time series as well as the phase locking of oscillations across two different voltage time series. Spectral analysis of EEG data reveals two important features of the time series of EEG data. First, EEG data, like most natural processes, has more energy at low than at high frequencies. This is a consequence of the autocorrelation within the time series; that is, the voltage at time t depends on the voltage at time $t-1$. The second feature of the time series is the presence of increased amplitude (or power) at specific frequencies. The presence of peaks at specific frequencies implies that, in addition to its generally slow-changing voltage pattern, the EEG time series can also contain oscillations at specific frequencies.

dent sample. This approach has significant practical utility in that it can be used to directly decode behavior or cognitive states from brain activity. Subsequent chapters both provide hands-on exercises involving these methods and discuss the application of these techniques in several cognitive domains.

As has been widely discussed (Millett, 2001), Hans Berger sought to find a physiological basis for mental telepathy. He pursued this goal for decades before he finally convinced his colleagues that he had found a reliable electrical signal emanating from the human brain. Scientists are now using multivariate statistical models to decode the variations in rhythmic activity throughout the brain to enable patients with severe motor impairments to communicate using brain signals, a feat not too far from the kind of mental telepathy that Hans Berger sought to understand.

Anatomy of an Experiment

Consider a list memory experiment in which subjects study items for a subsequent recognition or recall task. A researcher would normally program a computer to create the lists according to a set of pre-defined rules, present the lists either visually or auditorally to the subject, and record the responses. In these experiments subjects either give multiple-choice responses (e.g., Yes/No) or they provide vocal or typed responses to recall a word presented on a studied list. The computer would also create a log file to insure that all experimental events are recorded and time-stamped with an accurate algorithm to insure that the time stamps recorded to file closely match the true times at which the events occurred during the experiment. Using standard programming languages (e.g., Python, C++) coupled with some specialized experiment programming toolboxes, it is relatively straightforward to program a personal computer to control almost any experiment.

When you record physiological data, such as EEG signals from the scalp or direct recordings of field potentials from inside the brain (including single neuron responses), there is a separate device, also controlled by a (different) computer that manages the logging of the physiological signals. Any physiological study of memory requires an accurate method for aligning the physiological and behavioral data, so that one can create a final data structure with a single set of time stamps. [2]

When recording electrophysiological data there are two essential decisions that must be made at the outset: 1) The sampling rate at which the system will record the signals, and 2) the way that the electrical signals will be referenced. We briefly discuss each of these considerations below.

[2] Whereas EEG systems tailored for recording evoked potentials sometimes default to only record during specific behavioral epochs, it is much better to record continuous physiological data throughout the entire recording session. This can substantially improve your ability to filter the data, either to remove noise sources or identify slow rhythmic components in the data.

Sampling Rate. Higher sampling rates are akin to having more decimal places when you do math, and you may think that the higher the sampling rate the better. However, if the signals of interest are measured at the time scale of milliseconds there is no need to measure signals at nanosecond resolution. For scalp EEG signals, it is sufficient to record data at 1kHz (i.e., one sample per millisecond). However, for intracranial recordings with microwires, resolving the firing of individual neurons requires sampling rates of 10 kHz or higher.

Reference scheme. Neural activity supporting cognition produces the movement of charged particles (ions) in the brain. As such, the distribution

of charged particles, and their associated *electrical fields*, result in changes in the *voltage* between any two points. [3] Because EEG is the voltage difference between two sets of electrodes, the choice of reference scheme is critical. If the electrical fields between two locations is very similar, the voltage difference will be very small. If, however, there is a sudden change in the fields as you move across the cortex, this will produce a large voltage difference if measured by electrodes across this boundary. As such, if you are measuring spatially diffuse gradients in the electrical fields, you should choose a spatially distant and homogenous reference and if you want to measure spatially punctate gradients, you should choose a proximate reference.

For measuring large amplitude event-related signals reflecting the activity of large regions researchers frequently reference the EEG signals to either an average of all channels (average reference) or to a putatively neutral signal, such as that measured at the earlobes. For measuring high-frequency activity and single neuron responses from intracranial electrodes, researchers often choose a nearby electrode, or set of nearby electrodes, as the reference. Similarly, in clinical EEG studies used to localize epileptic spikes and seizures, neurologists also frequently use a bipolar reference scheme, referencing an electrode to its neighbor.

In the next several sections we briefly survey the major technologies for recording electrical signals in the human brain: Scalp EEG, magnetoencephalography, intracranial EEG, and neuronal recordings.

[3] Voltage is energy required to move a point charge between two locations. The reason it takes energy to move charge is because, as defined by Coulomb's law, charged particles exert a force on one another:

$$F_q = k_e \frac{q_1 \times q_2}{d_{1,2}^2},$$

where d is the distance between the charged particles, and k_e is the Coulomb constant. Thus, to move a point charge (i.e., a charge that is localized to an infinitesimal point in space) from coordinates $\vec{x}_1$ to $\vec{x}_2$ requires energy (=force×distance), and we define this difference in potential energy (because the point charge is assumed to have no mass) as *voltage*.

Electro- and magneto-encephalography

To record scalp EEG signals, we apply electrodes to the scalp of the head, either individually or by fitting a cap with integrated electrodes to the head. After applying the cap, or the individual electrodes, one applies gel (or sometimes salt water) to each electrode to improve the conductance from the scalp to the contact. Many modern systems provide amplification of the electrical signal at the site of each recording contact, and then send these signals to a system that records the signals at a particular sampling rate (e.g., 2 kHz).

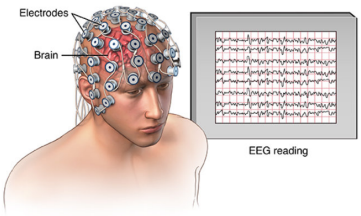


Figure 2.3: Illustration of a typical cap used for recording EEG signals, including an example of the EEG reading.

Magnetoencephalographic recordings require more complex equipment that measures the magnetic fields at the surface of the scalp. When ions move inside the brain they create currents, which induce magnetic fields that can be measured outside of the brain. These magnetic fields provide a view of currents that may be invisible to EEG measurements because currents produce magnetic fields that are orthogonal to the electrical fields. As such, by combining the measurement of magnetic and electrode fields, one can obtain a more complete view of the changing electrical signals in the brain. Researchers who record MEG (and EEG) signals often use algorithms to attempt to localize the sources of these signals to specific locations in the brain. There is much controversy, however, regarding the accuracy of these localizations. This is because the inverse problem (identifying the internal signals that produce a particular distribution of external potentials) cannot be uniquely solved and as such, the algorithms make very strong assumptions that may be unwarranted.

Intracranial Electroencephalography

Intracranial EEG recordings, which are also sometimes referred to as electrocorticographic (or ECoG) recordings, involve the implantation of electrodes into the brain itself, either in the form of (subdural) grids placed directly on the cortical surface (see Fig. 2.4) or depth electrodes placed into the brain's parenchyma. Because they measure brain activity with high spatial and temporal resolution, surgically implanted electrodes help physicians diagnose and treat neurological conditions such as epilepsy and Parkinson's disease. Here our focus is on ECoG recordings from patients undergoing invasive monitoring for drug-resistant epilepsy. In this procedure, surgeons implant ∼100–200 electrodes in widespread brain regions (Fig. 2.4A) to identify epileptic foci for potential surgical resection. Electrodes remain implanted throughout each patient's ∼1–3-week hospitalization. These electrodes include grid and strip electrodes (Fig. 2.4B,C), which record ECoG signals from the cortical surface, and depth electrodes (Fig. 2.4D), which penetrate the cortex to record field potentials from deep brain structures. In this text we will use the terms "ECoG" and "iEEG" interchangeably to refer to both surface and depth recordings. On occasion, surgeons implant microelectrodes, which record individual action potentials (Fig. 2.4E). We discuss these further below.

ECoG recordings measure brain activity directly with a resolution of $\sim 5\ mm^3$ (K. J. Miller, Sorensen, Ojemann, den Nijs, & Sporns, 2009). This high spatial resolution is a unique feature of ECoG compared to noninvasive methods like scalp electroencephalography (EEG) or magnetoencephalography (MEG). Noninvasive recordings, even with advanced localization algorithms, sometimes miss signals that are clearly visible with ECoG (Dalal et al., 2009). Furthermore, noninvasive techniques have difficulty isolating activity from deep brain structures and are relatively susceptible to muscle artifacts (Jerbi et al., 2009). Thus, ECoG is considered the clinical "gold standard" for accurately identifying seizure foci (Lachaux, Rudrauf, & Kahane, 2003; Crone, Sinai, & Korzeniewska, 2006). For the same reasons that ECoG recordings are useful to doctors, these data benefit researchers seeking to uncover the neural correlates of memory and cognition.

ECoG electrodes measure the combined synaptic activity across the local population of neurons, rather than recording individual action potentials (Logothetis, 2003; Crone et al., 2006). Due to this aggregation, ECoG recordings measure the electrical activity that is synchronized across these neurons, which often includes rhythmic activity such as neuronal oscillations. Oscillations appear as periodic changes over time in the voltage observed from an electrode (Fig. 2.4F) and can appear at frequencies from $<$0.1 Hz to 500 Hz. Spectral methods can help researchers to identify oscillatory activity in the time series of recorded voltages. These methods, which we discuss in detail in Chapter 4, can be used to measure the amplitude of the rhythmic component of a brain signal at different frequencies, as shown in the power spectrum in Figure 2.4G.

Microelectrode recordings

The most common method to measure *in vivo* neuronal activity is through extra-cellular recordings. In this method, a small micro-wire ($< 60\mu m$ diame-

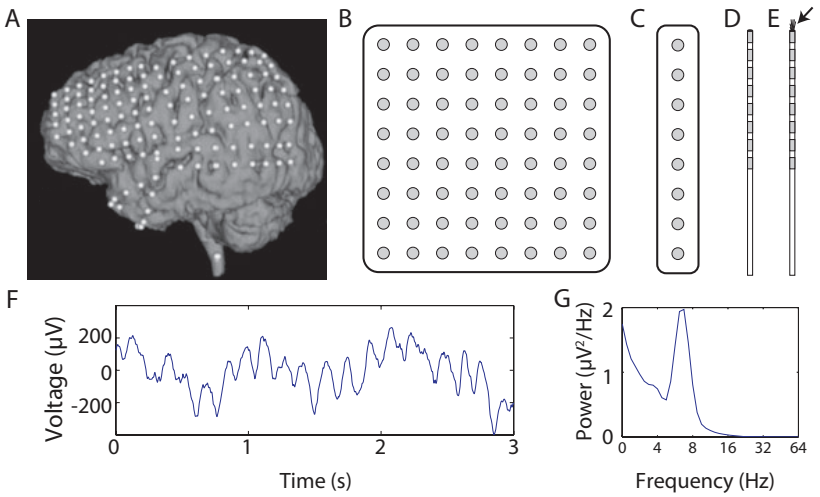


Figure 2.4: Intracranial EEG recordings. A. An MRI image of one patient's brain with the locations of implanted ECoG electrodes indicated with white dots. Modified, with permission, from (Towle et al., 2008). **B.** An illustration of an 8×8 electrode grid; gray shading indicates electrodes' conductive surfaces. (Illustrations not to scale). **C.** An illustration of an 8-electrode strip. **D.** A depth electrode with eight contacts. **E.** A depth electrode with microwires extending from the tip to record action potentials (marked by the arrow). **F.** A recording of ECoG activity from the right temporal gyrus. **G.** The power spectrum of the recording from Panel F, which shows that this trace exhibits a robust theta oscillation.

ter) is manually placed in the immediate vicinity of an intact neuron (Moxon & Nicolelis, 1999). Using single tungsten electrodes, micro-wire arrays and drivable tetrodes, extracellular neuronal recordings have long been a staple of animal electrophysiology studies. Indeed, if these electrodes are placed close enough to a neuron, activity from a single cell can be obtained (single-unit activity; SUA). Failing this, it remains possible to record multi-unit activity (MUA) from the collection of cells in a somewhat larger neighborhood around the recording site (Buzsáki, 2004; Gold, Henze, Koch, & Buzsaki, 2006). Both SUA and MUA recordings have contributed substantially to our understanding of the neural origin of cognitive functions.

Because single unit recordings pose substantial technical challenges in the clinic and they have yet to play a role in the treatment of epilepsy, only a handful of research-oriented clinical centers obtain these recordings from patients. To ethically obtain extracellular neuronal recordings in the clinic, one needs to use specialized depth electrodes such as the FDA-approved combined macro-micro depth electrodes designed by Itzhak Fried's team at UCLA (Behnke-Fried depth electrode and inner-wire bundle; Ad-Tech Medical, Racine WI) (Fried et al., 1999; Babb, Carr, & Crandall, 1973). This electrode is manufactured as two separate components. The clinical component (henceforth referred to as the Behnke-Fried depth; BFD) consists of eight standard cylindrical, depth macro-electrodes (90% platinum, 10% iridium alloy contacts, 1.3 mm in diameter, 0.8 mm in length) that are embedded on the surface of a silastic tube with a hollow lumen. The research component (henceforth referred to as the inner-wire bundle; IWB) runs through the lumen of the BFD and consists of eight micro-wires (platinum-iridium; 40 μm diameter; Teflon insulation) and one reference wire (an additional micro-wire stripped of insulation).

The fragile nature of the microwires poses a major challenge to the goal of obtaining quality unit recordings: A significant percentage break at some point during the patients hospitalization. For this reason it is essential to conduct impedance testing at various points post implantation. If the broken microwire is the reference channel, one can re-reference to a different

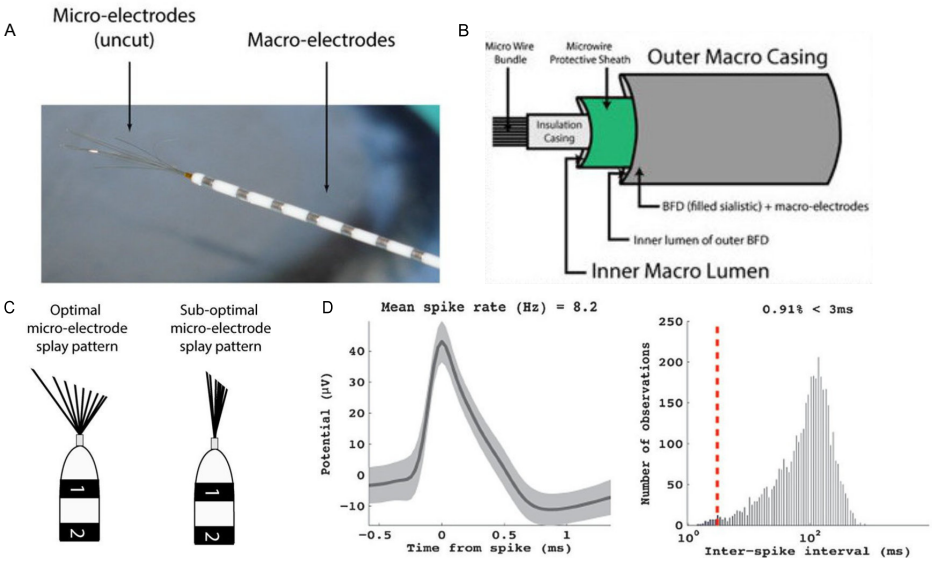


Figure 2.5: A. Projections of the micro-wires out of the end of the macro-electrode as received by the manufacturer (AdTech macro-micro electrode). **B.** Schematic of the outer and inner casing included in the macro-micro electrode. **C.** Cartoon example of optimal and sub-optimal micro-electrode splay patterns and staggered micro-wire lengths. **D.** The average waveform (left panel) and the distribution of inter-spike intervals (right panel) are shown.

microwire, insuring that the chosen reference is not recording any neuronal spikes. By taking adequate precautions to insure against broken electrodes (see Misra et al., 2014) one can obtain a yield of 2-3 neurons per depth electrode.

Researchers can choose one of several software packages developed to help isolate extra-cellular single unit activity from high-frequency microwire recordings (typically one uses a sampling rate in excess of 20 kHz for such analyses). These packages produce, as output, a series of spike times which one can easily convert into a continuous measure of the firing rate of each cell. Chapter 6 provides a more in-depth discussion of these techniques.

Why electrophysiology?

Supposing that our objective is to understand how memory works. Why study the brain? Many would answer by rejecting the premise behind this question. If the brain enables us to perform complex behaviors it is inherently interesting and we should try to understand it. Moreover, in cases of neurological and psychiatric diseases, understanding the brain has helped to advance diagnosis and treatment. Nonetheless, it would be valuable to consider how understanding electrophysiology can directly advance our understanding of cognition itself.

In answering this question consider first how we might understand cognition without electrophysiology. Here, the answer provided by traditional cognitive psychology is that we learn about the mind by studying behavior, and specifically we study the two primary dependent variables recorded during a memory task: accuracy and response time. Even cognitive psychologists, however, have found it very helpful to supplant these primary measures with additional behavioral measures whenever possible. For example, in the study of recall, the order of responses has helped elucidate the mechanisms of retrieval. In the study of recognition memory, researchers often ask subjects

to make confidence judgments ("How sure are you that the item was on the list"), or to judge the subjective quality of remembering ("Did you recollect specific details of the item's encoding or did it just feel familiar"). Whenever possible, cognitive psychologists have embraced more detailed information about the process of remembering and used that information to test theories of memory. One could reasonably argue that electrophysiological recordings should serve this role as well, providing millisecond-resolution data related to the processes of encoding, retention and retrieval. But to serve this role, researchers need to understand how to analyze and interpret the electrophysiological data.

As an example, consider the distinction between encoding and retrieval processes. During the study phase of an experiment we normally ask a subject to learn a list of memoranda (e.g., words or pictures) for a later test. There is no overt behavior measured during this stage, yet we know that encoding efficiency varies across both items and lists (e.g., Kahana, Rizzuto, & Schneider, 2005; Kahana, Aggarwal, & Phan, 2018). Without easy access to physiological measurements, cognitive psychologists developed elaborate behavioral methods to help uncover variability in encoding processes (e.g., Rundus, 1971). But by recording physiological activity during encoding, and relating that activity to subsequent memory performance, we can both identify those neural features measured during encoding that predict subsequent memory, and we can use those neural features to help characterize the previously unobserved variability in encoding processes.

An arguably even more powerful use of electrophysiological recordings would be to test theories of memory, but for this to happen, the theories need to make specific neural predictions. In some cases, neurobiologists studying brain recordings in animal models posit connections to memory phenomena that are not easily measured in animals. Subsequent sections describe such models, including the lLisman-Idiart-Jensen model of memory scanning. In that model, two brain oscillations work together to support *multiplexing* of short-term memories: enabling a small number of neural representations (of items, presumably) to repeat in serial order at the frequency of the theta rhythm. Cognitive neuroscientists have used both electrophysiological and haemodynamic (i.e., fMRI) recordings, in humans, to test the predictions of this model.

Mathematical models of human memory, developed to explain behavioral rather than neural data, frequently make predictions about the dynamics of internal representations. These internal representations had long been unobservable, but with modern recording techniques we can observe high-dimensional patterns of neural activity and study their dynamics during the course of a memory experiment. These studies have allowed us to identify putative representations in the brain that exhibit dynamics that match predictions from cognitive theories. Although it is hard to rule out cognitive theories with neural data (because you don't observe everything in the brain) it is nonetheless valuable to know which cognitive theories make predictions that can be related to neural processes. The success of one account over its competitor may reasonably lead to an increased investment in further experiments aimed at testing the more successful theory.

In his influential analysis of the visual system, David Marr (Marr, 1982) suggested a distinction between three levels of analysis of a complex system: the computational, the algorithmic and the implementational. Most complex

systems evolved to solve a problem, to compute something. Work at the computational level of analysis seeks to identify the problem that a system is trying to solve. The memory system, for example, needs to be able to deploy the relevant information from the past to solve the decision problems of the present. Much like a computer uses algorithms to solve any computational problem, the brain likely has its own algorithms to solve its computational problems. The development of memory models described in Chapter 1 illustrate the algorithmic approach to studying memory. Finally, the brain uses the nervous system to implement these algorithms. Viewed through the lens of Marr's analysis, electrophysiological data serve as an important bridge between the algorithmic level of cognitive theories of memory, and the implementational level of neural recordings. Having bridges between levels of analysis is essential if we are to someday understand how the brain gives rise to complex cognition and behavior.

Figure 2.6: David Marr (1945 – 1980) was a cognitive scientist at MIT known for his models of primate visual perception.

Neurons, Fields and Networks

The cognitive operations that contribute to episodic memory are instantiated in the patterns of activity produced by networks of neurons distributed throughout the brain. When neurons are active they generate *action potentials*, electrical events that alter the electric field in the neuron's extracellular environment. The action potentials of single neurons can be recorded using microelectrodes; larger electrodes implanted intracranially or non-invasively placed on the surface of the scalp can also be used to record changes in the aggregate electric field caused by the summed activity of many hundreds or thousands of neurons. Measuring these electric field changes is the basis of using intracranial and scalp electroencephalography to study memory function. Here, we first review some basic physical principles concerning the propagation of electrical charge that underlie the signals captured by these recording methods.

Neurons

Although there are many types of neurons in the nervous system, they share some common morphological characteristics. Neurons are made up of a cell body (*soma*) that contains the cell nucleus and performs the metabolic functions required by the rest of the cell. There are two types of processes that extend out of the soma and that enable communication between neurons: several *dendrites* and a single *axon*. The dendrites are short and branch out from the soma like trees; their role is to receive input signals from other neurons. The axon is generally much longer and is the neuron's output structure for sending signals to other neurons. The axon of a neuron extends out and terminates near the dendrites of a single or many other neurons. The axon of one neuron is separated from the dendrite(s) of another neuron by a small gap called the *synapse*.

Action potentials initiate at the soma (in a region called the axon hillock) and then travel down the axon via passive or saltatory conduction (described below). Communication between neurons depends on fast propagation of electrical signals witin a neuron from the axon hillock to the synapse, followed by slower chemical communication via the release of neurotransmitters into the synapse. The rapid propagation of the action potential is

accomplished by the flow of ions into and out of the cell through ion channels that span the cell membrane between the intracellular and extracellular media. Ion channels consist of proteins that open and close in response to specific electrical or chemical signals. The flow across ion channels occurs very rapidly, leading to large currents and changes in voltage across the cell membrane (Equation 2.1), compared to the resting state.

Resting membrane potential At rest (i.e. when a neuron is not in the middle of generating an action potential), there is a voltage difference across the cell membrane, meaning the difference in electric potential is non-zero between the inside and outside of the neuron [4]:

$$V_m = V_{intra} - V_{extra} \tag{2.2}$$

where V_x are the intra- and extracellular voltages. When V_m is measured in neurons at rest it is referred to as the *resting membrane potential* and is typically between -60 mV to -70 mV. Electrical signaling occurs when V_m deviates from its resting value due to opening and closing of ion channels that cause current to flow across the membrane. Deviations that cause V_m to become less negative (decreasing the voltage difference across the cell membrane) are called *depolarizations* and deviations that cause V_m to become more negative are called *hyperpolarizations*.

The resting membrane potential is important because it means that once voltage-gated ion channels open in response to depolarization inside the cell, the resting imbalance in voltage between the inside and outside of the cell will cause positively charged ions to flow into the cell. The flow of charged particles is passive, meaning that the cell does not expend energy to move ions across the cell membrane. The flow of charged particles also creates a change in the extracellular voltage that can be measured with microelectrode recordings.

Action potentials and synaptic transmission When the membrane potential of a neuron reaches a threshold of depolarization at the axon hillock, it triggers an action potential. The threshold-crossing depolarization leads to the opening of voltage-gated Na^+ channels in the area of the initial depolarization. The opened channels allow excess Na^+ outside the cell to flow into the cell, leading to further depolarization, which in turn causes the opening of additional voltage-gated channels nearby. This positive feedback loop occurs in such a way that Na^+ channels are sequentially opened along the length of the axon. The resultant propagation of the depolarization leads to current flow along the length of the axon in the extracellular space.

As the depolarization moves down the length of the axon, slower cellular processes begin to take effect in which K^+ ions flow out of the cell, leading to hyperpolarization and restoration of V_m back to the resting membrane potential. When the action potential propagates to the end of the axon, it causes the release of chemicals called *neurotransmitters* into the synapse between the end of the pre-synaptic axon and the post-synaptic dendrites of other cells. The released neurotransmitters can then bind to receptors in the post-synaptic neuron(s) and then, depending on whether the synapse is excitatory or inhibitory, cause either hyper- or depolarization of the post-synaptic cell. If the synapse is excitatory, the pre-synaptic action potential may thus evoke an action potential in the post-synaptic neuron.

[4] Current is the flow of electric charge through a medium, typically carried by electrons or ions. In the case of a neuron, current is induced when ions flow through voltage-gated channels as part of action potential propagation (see *Neurons*). The strength of current flow is dependent on the voltage difference between two points, where voltage refers to the difference in electric potential energy between two points. In a conductor, current and voltage are related directly through *Ohm's Law*

$$V = IR \tag{2.1}$$

which indicates that the current, I, present in a medium is a function of the voltage difference (V) between the two points and the resistance (R) of the medium. Analagous to the flow of water through a pipe, if the pipe is of small diameter (high resistance) then it will be difficult for much water to flow through the pipe, even if it is oriented at or close to vertical (high voltage). In contrast, the same vertical pipe orientation (voltage) could produce a much larger flow of water (current) in a pipe of larger diameter (low resistance).

Ensembles and their fields

THE EXTRACELLULAR FIELD Above, we described how action potentials alter the extracellular field during communication events between neurons. Here, we review other sources of current flow that alter the extracellular field, and how these lead to voltage fluctuations that are detectable with electrophysiological recording methods.

The extracellular field at any point is the sum or superposition of many types of ionic events occurring in the area near the recording electrode. When recording the extracellular field using scalp electroencephalography (EEG) or intracranial EEG, many small currents need to overlap in space and time in order to be detectable. For this reason it is generally accepted that the largest contributor to the extracellular field is synaptic activity because of the way that dendrites (and their synapses) cluster near the cell body. This allows the currents generated by the opening of many ion channels within a local area to aggregate, generating a *sink* at the point at which positive ions flow into the cell. In order to maintain conservation of charge, the sink generates a corresponding *source* and together the sink and source are referred to as a *dipole*. The cumulative effect of many dipoles in close spatial proximity contributes to measurable voltage fluctuations in the extracellular medium, and decays with distance (r) as $1/r^2$.

An important factor that impacts the strength of fluctuation in the extracellular field is the geometry of the dipole-generating neural elements. In general neurons in the cortex of the brain are arranged in layers such that the apical dendrites of neighboring pyramidal cells are oriented parallel to each other. This parallel organization is conducive to the superposition of neighboring dipoles from many active neurons, which leads typically to large signal fluctuations in cortical recordings. The other influence on the strength of the extracellular field is the temporal synchrony of the activity, which is necessary in order for the induced electrical events to superimpose.

Oscillations in the local-field potential

The neural events underlying fluctuations in the extracellular field must show some degree of temporal synchrony to be detectable in EEG signals measured at the scalp or even with large intracranial electrodes. As such, researchers have devoted significant effort to understand several prominent oscillations that are observed in the EEG recordings of animals and humans. Here we review some of these signals and their relation to cognitive processes.

Theta oscillations, place cells, and spatial memory

Because the hippocampal formation plays a critical role in learning and memory, researchers have long sought to understand how patterns of electrical activity measured in the rat hippocampus relate to animal behavior and cognition. Technological advances in the 1960s enabled researchers to measure these patterns of electrical activity in awake, behaving animals (e.g. Vanderwolf, 1969). One such pattern of electrical activity is the hippocampal theta rhythm—a 4-10 Hz oscillation that appears when an animal is alert and

interested in its surroundings (Bland, 1986). Although early work emphasized the relation of theta specifically to motor activity, subsequent studies demonstrated a broader role for theta in the hippocampal computations underlying spatial learning (Kahana, Seelig, & Madsen, 2001).

Theta's functional importance derives from several lines of evidence. First, theta appears to act as a windowing mechanism for synaptic plasticity, with synapses being strengthened when pre- and post-synaptic neurons fire at the peak of the theta rhythm, and synapses being weakened when neurons first at the trough of the theta rhythm (Huerta & Lisman, 1993; Hölscher, Anwyl, & Rowan, 1997).

If the phase of theta is crucial for experience-dependent changes in synaptic strength, then one might expect that stimulus events would produce a reset or phase shift in ongoing theta. Consistent with this hypothesis, the activity of hippocampal theta appears to be phase locked to stimuli when an animal is incentivized to maintain the stimulus information in memory (B. Givens, 1996). Taken together, these observations help to explain how important sensory input undergoes neural encoding (Hasselmo, Bodelon, & Wyble, 2002).

Second, theta's functional importance has been demonstrated through attempts to block theta. Theta can be blocked by lesioning a region in the brain known as the medial septum. Such lesions, in addition to blocking theta, produce severe impairments in memory function (e.g. see B. S. Givens and Olton (1990)). Although neither prior learning of spatial information nor hippocampal place representations are impaired by septal lesions, such lesions do impair the acquisition of new spatial information (Leutgeb & Mizumori, 1999). This evidence suggests that theta has a role in memory, but it is difficult to dissect the specific effect on theta from the concomitant cholinergic loss. Adaptive electric-field feedback, which can accentuate or minimize a specific frequency band in generated field potentials (Gluckman, Nguyen, Weinstein, & Schiff, 2001), may be used to directly assess the effects of theta manipulation in the rat.

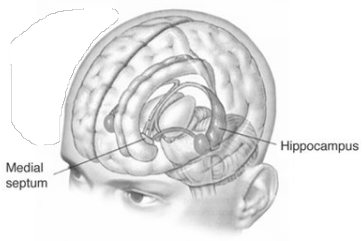


Figure 2.7: Illustration of deep brain structures including the medial septum and hippocampus.

Human Theta

By the 1990s, theta was extremely well described in rodent studies, yet there was scant evidence for any homologous rhythm in primates. Electrophysiological investigations in monkeys focused primarily on sensory and motor processes and their neural correlates, and no documented evidence for task-dependent theta in monkeys had been reported prior to 1999. Motivated by studies of theta activity during rodent navigation, Kahana, Sekuler, Caplan, Kirschen, and Madsen (1999) had patients with subdural grid electrodes learn to navigate through virtual, three-dimensional rendered mazes. Their recordings revealed clear oscillations in the unfiltered iEEG traces and demonstrated that, during maze navigation, intermittent bouts of theta activity appear with greater probability during longer mazes, even when controlling for degree of mastery (see Figure 2.8).

Caplan et al. (2001) showed that the effect of maze length on theta does not reflect the increased difficulty of encoding or retrieval at individual choice points. Rather, it reflects a global difference between long and short mazes. Caplan et al. found that gamma activity, but not theta, increased with increasing difficulty of individual choices at maze junctions. It was at

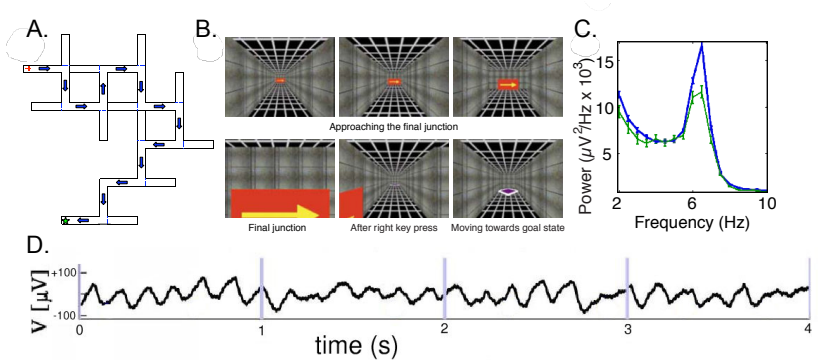


Figure 2.8: Theta oscillations during maze learning. A. Blueprint of a sample maze. Participants navigated T-junction mazes, from a starting point (+) to a goal position (⋆). **B.** Sample views during study phase. During the first four traversals, arrows denoted the correct path (*study*). Participants then repeatedly traversed the maze without arrows (*test*) until they traversed it three times consecutively without errors. Maze length was varied (short mazes had 6 junctions; long mazes had 12 junctions). **C.** Power spectra at an electrode in the inferior frontal gyrus indicating greater ∼6.5-Hz theta power for correct traversals of long mazes (thick blue) compared with traversals of short mazes (thin green). This effect was observed at multiple sites in each of five patients (Kahana et al., 1999; Caplan et al., 2001). **D.** Theta oscillations revealed by intracranial EEG recorded from an electrode on the inferior frontal gyrus during a virtual maze navigation task.

these difficult junctions in which the learning requirements would have been greatest.

Because subjects often learn T-mazes as a verbal sequence of left and right turns (Kirschen, Kahana, Sekuler, & Burack, 2000), it is difficult to make strong inferences about the role of theta in spatial processing from Kahana and Caplan's earlier studies. To help address this confound, Caplan et al. (2003) developed a task called "Yellow Cab" in which subjects play the role of a taxi driver, driving through a virtual town in search of passengers and delivering those passengers to their requested destinations (see Fig. 2.9a,b). Over repeated deliveries, passengers learn to find the shortest path between the random locations in which they find their passengers and the fixed locations of landmarks within the environment (Caplan et al., 2003; E. L. Newman et al., 2007).

Caplan et al. (2003) found increased theta activity at widespread cortical sites when subjects were moving (compared with periods in which they were still). Extending these results by comparing iEEG recordings from hippocampus and neocortex, Ekstrom et al. (2005) found increased theta activity in both regions when subjects were moving within the virtual town. Watrous, Fried, and Ekstrom (2011) further demonstrated that the frequency of human hippocampal theta increases with a subject's virtual speed, as in rodents. Finally, the amplitude of theta appears to positively correlate with performance in a virtual navigation task (Cornwell, Johnson, Holroyd, Carver, & Grillon, 2008).

The high amplitude theta activity reported during human maze learning appears very much like the theta seen in rodents during spatial exploration. Following the initial reports of task-dependent human theta, some investigators suggested that these observations may be specific to tasks that involve a spatial component (O'Keefe & Burgess, 1999); however, the discovery of rodent theta during non-spatial learning tasks (Hasselmo et al., 2002; Ekstrom, Meltzer, McNaughton, & Barnes, 2001) and numerous findings of task-dependent theta in non-spatial memory paradigms indicates that theta plays a far more general role in human cognition. Subsequent chapters discuss the extensive literature demonstrating the importance of theta-frequency activity in a wide range of cognitive tasks.

The ability of human intracranial recordings to record from deep brain structures allowed researchers to compare the properties of hippocampal

theta between humans and animals. As a result, researchers identified two important interspecies differences. In rodents, hippocampal theta oscillations reliably appear at 4–8 Hz(Buzsáki, 2005). However, in humans, hippocampal oscillations usually appear instead at 1–4 Hz (Bódizs et al., 2001; de Araujo, Baffa, & Wakai, 2002; Ekstrom et al., 2005; Jacobs, Kahana, Ekstrom, & Fried, 2007; Babiloni et al., 2008; Cornwell et al., 2008; Clemens et al., 2009). Furthermore, whereas rodent theta oscillations are routinely sustained for over ten seconds (O'Keefe & Recce, 1993; Buzsáki, 2005), human hippocampal oscillations usually appear only transiently (Caplan et al., 2003; Ekstrom et al., 2005) and sometimes not at all (Niedermeyer, 2008). Despite these differences, it seems that human 1–4-Hz hippocampal oscillations are functionally analogous to rodent 4–8-Hz theta. During navigation, both of these oscillations increase in amplitude during movement (Ekstrom et al., 2005; Buzsáki, 2005; Jacobs, Korolev, et al., 2010) and the phase of both oscillations modulates neuronal spiking (O'Keefe & Recce, 1993; Klausberger et al., 2003; Jacobs et al., 2007).

A neural GPS system

Advances in the ability to record extracellular action potentials from multiple hippocampal neurons while rodents performed complex navigation-based learning tasks led to the discovery of the neurons in the rodent hippocampus that represent the animals' spatial location within a given environmental (O'Keefe & Dostrovsky, 1971). These so-called *place cells* fire preferentially whenever the animal passes through a region of their environment, termed the *place field*. When navigating through an environment with an open layout, place cells fire without regard to the direction of movement, exhibiting the property of *omnidirectionality*; in contrast, place cells generally exhibit a preferred direction when the environmental layout consists of narrow paths that the animal would typically traverse in one of two directions, as in a T-maze. In the rodent hippocampus, approximately 25% of all pyramidal cells appear to represent information about the animals position within their environment.

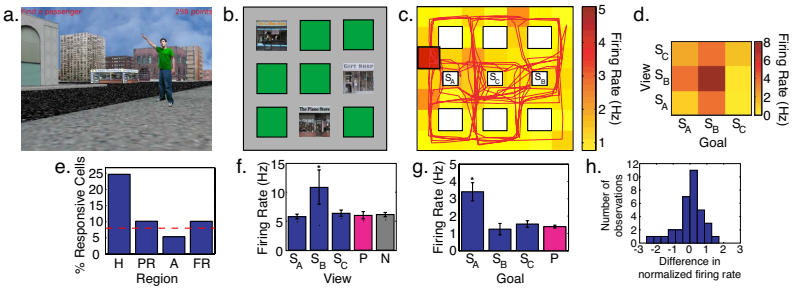


Figure 2.9: Cellular responses during spatial navigation. a. View from within Yellow Cab. **b.** Example town layout. **c.** Firing rate of a hippocampal neuron overlaid on map with participant's path shown in red. **d.** A conjunction cell that responded preferentially when the goal store (S_B) was also in view. **e.** Proportion of place cells in hippocampus (H), parahippocampal cortex (PR), amygdala (A), and frontal regions (FR). Red line indicates false-positive rate. **f.** A view cell in parahippocampal cortex that responded preferentially when viewing store B (S_B). P denotes responses to views of a passenger; N denotes neutral views (N) of neither stores nor passengers. **g.** A goal cell in the amygdala that responded preferentially when seeking store A (S_A). **h.** Regions of high firing included high numbers of traversals in different directions. The distribution of firing-rate differences across these traversals was centered on zero, indicating that cells responded in an omnidirectional manner.

The hippocampal theta rhythm appears to also play an important role in the neural coding of place. As a rat traverses a place field, hippocampal place cells fire at a progressively earlier phase of the ongoing theta oscillation (O'Keefe & Recce, 1993; Skaggs, McNaughton, Wilson, & Barnes, 1996). This information significantly improves accuracy in reconstructing the animal's position in space (Jensen & Lisman, 2000), providing additional support for the hypothesis that the phase of theta at which cells fire plays an important

role in the coding of place information in the rat hippocampus. Phase precession may also provide a more general mechanism for sequence learning in episodic memory outside of purely spatial domains (Buzsaki & Moser, 2013).

Ekstrom et al. (2003) asked whether neurons in the human hippocampus similarly encoded spatial information in a context-dependent manner. Using a variant of the Yellow-Cab taxi-driver game described above, Ekstrom et al examined the firing patterns of 378 neurons recorded across a sample of 7 neurosurgical patients. Figure 2.9c shows an example of a hippocampal cell whose firing rate peaked at its *place field* in a northwest region of the virtual town. Ekstrom et al observed significant place-selective neural activity in 39 cells, with the majority of place-responsive cells being found in the hippocampus (Fig. 2.9e). Ekstrom et al further observed that the majority of these cells responded in a direction-independent manner (Fig. 2.9h) mirroring the finding of omnidirectional place cells in the rat hippocampus (O'Keefe & Dostrovsky, 1971).

An advantage of virtual navigation tasks is that software is able to track which landmarks subjects could view during each frame in the game. Ekstrom et al used these data to determine how neurons responded to visual information, such as whether a target store was in view. Of theoretical interest was whether view-responsive cells exist in the same medial temporal lobe (MTL) brain network that appears to be involved in navigation. This analysis identified 41 cells that responded preferentially when participants viewed specific landmarks (Fig. 2.9f). Ekstrom et al observed a highly significant dissociation between the anatomical distributions of place and view cells, with place cells being most prevalent in the hippocampus and view cells being most prevalent in the parahippocampal region. This is consistent with a considerable body of functional neuroimaging evidence implicating the parahippocampal region in the processing of spatial scenes (Epstein & Kanwisher, 1998; Epstein, 2005).

In addition to studying the cellular correlates of place and view, the Yellow Cab paradigm provided the opportunity to explore the neural representation of goals. Because participants were asked to drive to one of three randomly chosen goal destinations on each trial, we were able to determine whether individual neurons were responsive to specific goal destinations independently of place or view. This analysis revealed 69 goal-responsive cells (Fig. 2.9g). Consistent with evidence from recent evidence in rodents, Ekstrom et al found these cells in frontal brain regions (Hok et al., 2005).

Finally, Ekstrom et al examined whether cells responded to conjunctions of these variables. For example, one might expect that some goal cells would respond most strongly when the goal destination comes into view. This can be seen in a right amygdalar cell that responded preferentially to a view of a given store when that store was the participant's intended destination (Fig. 2.9d). The authors observed significant numbers of neurons that responded to conjunctions of place and view, and to conjunctions of place and goal. The existence of these conjunctive cells was anticipated by Burgess, Becker, King, and O'Keefe's (2001) computational model of spatial navigation.

Building on the discovery of hippocampal place cells, which identify unique locations within a specific environment, a team led by husband and wife neurobiologists, Edvard and May-Britt Moser, discovered an amazing class of neurons, termed *grid cells* that fire when a rodent traverses locations

corresponding to the vertices of a triangular grid spanning the animal's environment. Unlike place cells, which reside primarily in the hippocampus, grid cells appear most prominently in the medial entorhinal cortex. By combining place and grid responses, the central nervous system is able to precisely represent an animal's location within an environment. Recording from neurosurgical patients as they performed a virtual navigation task in an open-field environment, Jacobs et al. (2013) reported grid-cell like responses in the human entorhinal cortex, and demonstrated that these cells exhibit similar response profiles as those reported in rodents. [5]

Place and grid cells exist as part of a broader network of neurons that encode other navigationally relevant variables including those related to local geometry (border cells), orientation (head direction cells), current navigational goals, and scene information. Current theories of spatial memory suggest that the hippocampus constructs a viewpoint-independent (allocentric) representation of the current environment based on not only place and grid cells, but also on inputs from parahippocampal cortex (scene information), parietal cortex (viewpoint-dependent egocentric information), temporal cortex (featural representations of environmental stimuli), and prefrontal cortex (goal- and route-planning information). The major challenge for the field is understanding how the brain uses these representations to navigate, reason about spatial relations, and flexibly distinguish between and exploit different spatial representations in a context-dependent manner. We take up some of these issues in a later chapter.

[5] John O'Keefe, Edvard Moser, and May-Britt Moser received the 2014 Nobel Memorial Prize for their characterization of the cellular networks underlying spatial cognition. This award followed the discovery of similar neural responses in numerous other species, including humans.

Do rhythms help to organize spikes?

Research in animals shows that brain oscillations provide a neuronal timing signal that allows neurons to encode information by spiking at a particular phase of an oscillation—a phenomenon called phase coding (O'Keefe & Recce, 1993; Fries, Nikolić, & Singer, 2007). To examine the prevalence and properties of phase coding in humans, Jacobs et al. (2007) examined how neurons in widespread regions varied their instantaneous firing rate according to the phase of ongoing oscillations. This work found that many neurons were *phase locked* to oscillations, a phenomenon in which they increased their firing rate at a particular phase of these oscillations. Figure 2.10A shows the activity of a neuron that exhibits this phenomenon by spiking just before the peak of the theta oscillation. The properties of neuronal phase locking varied between high- and low-frequency oscillations. Neurons phase locked to oscillations at frequencies slower than 10 Hz had various preferred phases, whereas neurons phase locked to oscillations faster than 10 Hz had preferred phases near the oscillation's trough. This indicates that oscillations faster than $\sim$10 Hz reveal specific times (the trough of the oscillation) when many neurons are active, whereas slower oscillations cannot predict population spike times with this level of precision.

In addition to examining the timing of individual action potentials, a different set of studies examined the relation between the rate of neuronal spiking and the amplitude of oscillatory activity. In some cases, neuronal firing rate is well predicted by the amplitude of simultaneous oscillations (Fig. 2.10B). However, the details of this relation dramatically vary according to the oscillation and brain region being examined. Oscillations at high frequencies ($>$10 Hz) in sensory cortex correlate positively with neuronal spiking

(Nir et al., 2007) and a similar, but weaker, pattern appears in hippocampus (Ekstrom, Suthana, Millett, Fried, & Bookheimer, 2009). In contrast, low-frequency oscillations exhibit varied correlations with single-neuron spiking: In neocortex, theta- and alpha-band oscillatory power is negatively correlated with neuronal spiking (Nir et al., 2007), but in hippocampus these oscillations do not correlate with spiking rate (Ekstrom et al., 2009).

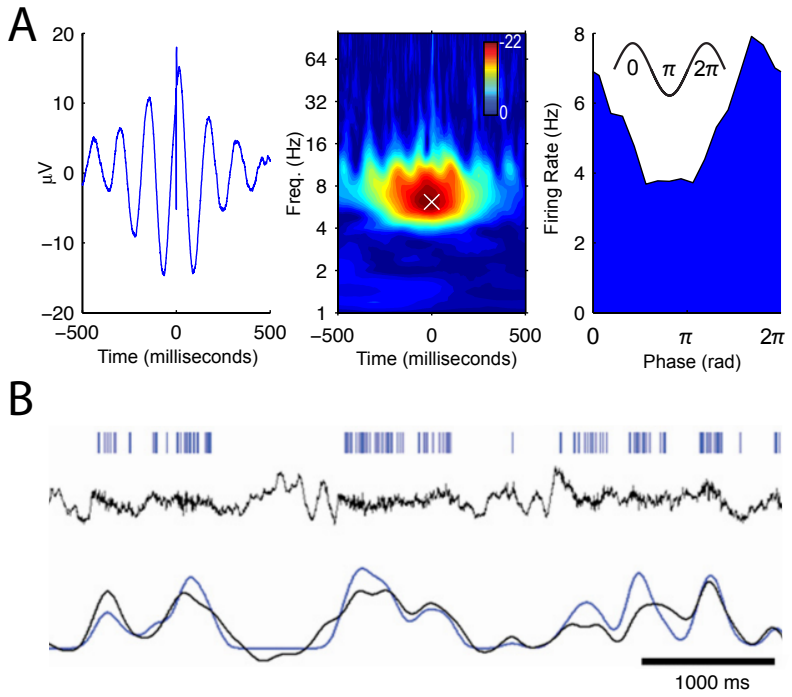


Figure 2.10: The relation between oscillatory brain activity and neuronal spiking. A. Activity of a neuron in the right superior temporal gyrus that spiked just before the peak of the theta oscillation. Left panel, average local-field potential (LFP) computed relative to each spike. Middle panel, z-score index of the phase uniformity at the time of each spike, as a function of frequency and time offset. White '$\times$' indicates frequency of peak phase locking. Right panel, firing rate of this cell as a function of instantaneous theta phase at the frequency of peak phase locking. Adapted from (Jacobs et al., 2007). **B.** The activity of a neuron in auditory cortex whose spiking was tightly coupled to the amplitude of simultaneous gamma oscillations ($r = 0.84$). Ticks in top row indicate action potentials. Middle row depicts the LFP signal filtered to only include frequencies below 130 Hz. Bottom row shows the correlation between LFP gamma power (black) and neuronal firing rate (blue). Adapted from (Nir et al., 2007).

"Broadband" EEG fluctuations as a marker of neural activity

Reports of strong correlations between neuronal firing and narrowband activity (i.e. oscillations) have supported the view that oscillations reflect synchronized spike timing in large neuronal ensembles (Singer & Gray, 1995; Logothetis, 2003; Fries et al., 2007). This follows in part from the temporal-binding hypothesis (von der Malsburg, 1981), which proposes that synchronized neural activity can solve the "binding problem" by linking multiple neuronal signals (Köhler, 1947; Koffka, 1935; Kanisza, 1979; Pal & Pal, 1993). However, an emerging body of research has shown that apparent correlations between spikes and gamma-band LFP activity are actually due to broadband LFP patterns, rather than band-specific oscillations (e.g., K. J. Miller et al., 2014; Burke, Ramayya, & Kahana, 2015).[6]

Using simultaneously recorded LFP and single-neuron activity, Manning, Jacobs, Fried, and Kahana (2009) investigated the relationship between broadband activity, narrowband oscillatory activity, and underlying neuronal spiking. Consistent with previous studies, they found a population of *narrowband-shift neurons*, which varied their firing in proportion to LFP power at specific frequency bands. Narrowband-shift neurons were present

[6] The $1/f$ spectrum (or pink noise spectrum) is instantiated in many systems in the natural world and refers to a characteristic of a signal in which the power spectral density of the signal is inversely proportional to frequency. $1/f$ noise has been observed in the flow of traffic on highways (Musha & Higuchi, 1976), the structure of DNA base sequences (Voss, 1992), and in the timeseries of errors that people make in perceptual memory tasks (Gilden, Thornton, & Mallon, 1995). Neural activity recorded using electrophysiological measures also follows a $1/f$ pattern, meaning that power at low frequencies is much higher than at low frequencies (also see Chapter 5).

throughout the brain, but were especially prevalent in the frontal cortex and amygdala. In addition, they observed a larger population of *broadband-shift neurons*, which varied their firing with the overall height of the LFP power spectrum at all frequencies. Broadband-shift neurons appeared in all examined brain regions, but were especially prevalent in the medial-temporal lobe. Broadband increases in LFP power were almost exclusively positively correlated with single-neuron firing, providing a robust estimate of neuronal firing. Below we describe this study in greater detail, primarily to illustrate the methods used in the analysis of both spectral EEG activity and neuronal spiking.

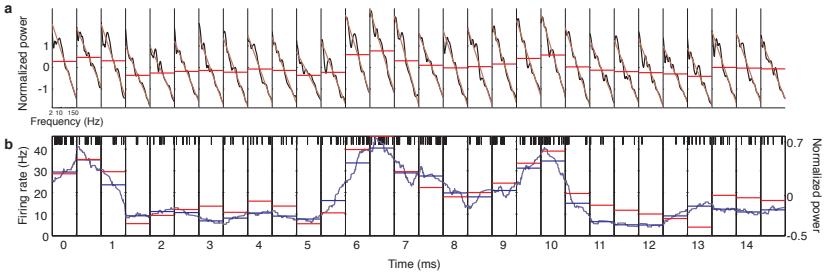


Figure 2.11: Neural spike times match broadband LFP fluctuations. Each box details the activity in one 500-ms epoch. **a.** This panel illustrates how various features of the LFP change over time. In each epoch, the black lines indicate the overall LFP power spectrum, brown lines indicate robust-fit lines, and the horizontal red lines indicate mean broadband powers. **b.** This panel illustrates changes in neuronal firing rate concurrent with changes in the LFP power spectrum. Black vertical ticks represent the times when individual spikes occured, dark blue lines indicates the smoothed firing rate (see *Methods*), and horizontal blue lines indicate mean firing rates in each epoch. Mean broadband power is shown in panel b (horizontal red lines) on a different scale (indicated at right).

Using recordings of 2,030 neurons from 20 neurosurgical patients Manning et al. (2009) determined how moment-to-moment variations in the local field potential related to simultaneous changes in the firing rates of nearby neurons. Spectral analysis methods provide a tool for transforming the time series of voltage activity (the EEG or local-field potential) into distinct frequency components. Even if there are no oscillations present in the neural activity, the spectrum will still have energy at varying frequencies. For a randomly varying voltage signal that is based, in part, on its prior state (i.e., an autocorrelated signal), the power spectrum will have more energy at lower frequencies than at higher frequencies, with power falling off as a function of the reciprocal of frequency, often raised to an exponent (i.e., $\text{Power}(f) \sim 1/f^{\alpha}$). An increase in overall variability in the signal would raise the power at all frequencies, conforming to the overall shape of the colored noise distribution. By contrast, a narrowband oscillation would appear as a peak in the power spectrum, rising above the background ($1/f^{\alpha}$) distribution.

Figure 2.11 illustrates the relation between broadband power shifts and spiking activity recorded from a sample electrode recorded in Manning et al's study. Panel A shows the normalized LFP power spectrum (black lines) and the mean broadband LFP power (red lines) for each of thirty consecutive 500-ms epochs. Panel B shows the neuron's spiking (black tick marks) and mean firing rate (blue lines) for these same epochs. Across these epochs, variations in broadband power were strongly correlated with simultaneous variations in the neuron's firing rate (Pearson's $r = 0.92$). Looking at the figure, one may observe that both broadband power and neuronal firing rate exhibited local maxima at 0.5 s, 4 s, 6.5 s, and 10 s, and both had local minima at 1.5 s and 5 s. Variations in LFP power were not limited to particular narrow frequency bands, but rather appeared as overall broadband shifts in

the entire power spectrum (brown lines in panel A).

To determine whether this pattern was robust across the entire recording session, Manning and colleagues examined broadband LFP power and firing rate for each of the half-second epochs recorded for this neuron. Data from each epoch appear as a point in Figure 2.12a, where the horizontal coordinate indicates the firing rate and the vertical coordinate indicates the normalized broadband power. Across the entire recording session, these points were clustered along the diagonal, indicating that neuronal firing rate was positively correlated with LFP broadband power (Pearson's $r = 0.6$). Figure 2.12b depicts this relation in a different manner, showing the mean LFP power spectra for each of five groups of epochs where this neuron had different firing rates (different colors in panel a). As this neuron's firing rate increased, the LFP power spectrum exhibited a proportional upward shift at *all* frequencies.

Manning et al next sought to identify all neurons whose firing rates varied with broadband power (as in the example above) or with narrow-band power. Because broadband power is influenced by each narrow frequency band, disambiguating broadband and narrowband effects is critical for understanding the relation between neuronal spiking and LFP activity. To identify neurons exhibiting each of these patterns, they fit a bivariate linear regression model to the relation between firing rate and measures of both broadband and narrowband LFP power. For each neuron, they computed the firing rate for each half-second epoch and they also computed LFP power measured at the same electrode at five narrow frequency bands: delta (2–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–150 Hz), in addition to computing broadband power. For each neuron they then performed a set of bivariate regressions where broadband power and the mean power in one narrow frequency band were simultaneously used to predict the neuron's instantaneous firing rate. The two β coefficients estimated in each regression indicate the contributions of broadband activity and this particular narrowband frequency band towards each neuron's firing rate.

Combining the results of all five regressions for each neuron, Manning et al designated a neuron as a *broadband-shift neuron* when all five β coefficients for the broadband predictor were significantly different from zero in the same direction. They designated a neuron as a *narrowband-shift neuron* if, across the five regressions, either (a) one and only one narrow-band β coefficient was significantly different from zero or (b) exactly two narrowband β coefficients at adjacent frequency bands (e.g. beta and gamma) were significantly different from zero in the same direction.

The broadband shift effect was remarkably unidirectional, with 92% of all broadband-shift neurons exhibiting this effect in a positive direction. In contrast, among narrowband-shift neurons, only 66% exhibited positive correlations. Figure 2.13 shows that broadband and gamma-band power were the two dominant LFP measures that positively correlated with firing rate. The proportions of neurons exhibiting these two effects were comparable to one another and were both significantly greater than the proportions of significant positive or negative correlations observed at other frequency bands

Assuming that brain function depends largely on the activity levels of neurons, the foregoing analysis provides valuable information on how to

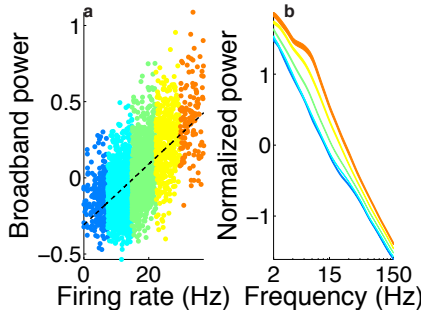


Figure 2.12: A representative neuron exhibiting a positive correlation between firing rate and broadband LFP power. a. Broadband power and firing rate for the neuron analyzed in the figure above. Each 500-ms epoch of the recording session is represented by one colored dot. The color of each dot represents its relative firing rate. Warm colors depict epochs with high firing rates, and cool colors indicate epochs with low firing rates. The dashed black line shows an ordinary least squares regression to these data. **b.** Average LFP power spectra for epochs with different firing rates. The same color scheme is used in both panels. As firing rate increases, the power spectrum exhibits a positive shift at all observed frequencies. The thickness of each line represents $\pm$ 1 SEM.

infer neuronal activity from changes in the local-field potential. Specifically, these analyses suggest that broadband increases in spectral power signal an increase in the firing rates of neurons in the region in which the LFP is being measured.

Gamma oscillations

Activity in the gamma band (typically defined in the 30-100 Hz range) reflects fluctuations in extracellular fields that are faster than those in the alpha band. Gamma activity typically increases during successful performance of a cognitive task and is, for example, enhanced during the learning of information that is later remembered relative to forgotten (Ezzyat et al., 2017). There is some debate about whether this type of increased gamma activity is truly oscillatory or instead reflects broadly asynchronous synaptic activity (Burke et al., 2015). Under the first scenario, synchronous gamma oscillations play a mechanistic role in cognitive processes such as memory encoding, in that the process depends on the coordinated firing of distinct neural networks (Jensen & Lisman, 2005); we discuss such a model of working memory encoding, the Lisman-Idiart-Jensen model, in more detail in the next section. Under the second scenario, increased activity in the gamma band reflects increased synaptic activity at the recorded location that is not synchronous. Under this view, increased gamma activity can be thought of as an index of the strength of underlying neural activity (Manning et al., 2009)–the implication is that gamma activation could reflect any number of underlying processes that are not memory specific. Instead, information about memory-specific processes is contained in the spatiotemporal pattern of gamma activity across the brain (see Chapter 3).

Theta-gamma interactions

In addition to the notion that oscillations at particular frequencies reflect (a)synchronous activity that *independently* support cognitive functions, there is also significant evidence for interactions between activity at different frequencies. One of the most prominent examples is theta-gamma coupling (Colgin, 2015). Typically this refers to fluctuations in gamma power that consistently occur at a particular point in the theta cycle (theta phase-locked changes in gamma power). For example, in the human hippocampus, increased gamma power was shown to occur preferentially at the trough of the theta cycle during the encoding of remembered compared to forgotten words (Lega, Burke, Jacobs, & Kahana, 2015). Long-term potentiation (LTP), the primary mechanism for altering the connection strength between neurons, is dependent on the phase of theta, suggesting a reason for gamma activity to be coupled to the theta cycle in hippcampus (Hasselmo & Eichenbaum, 2005).

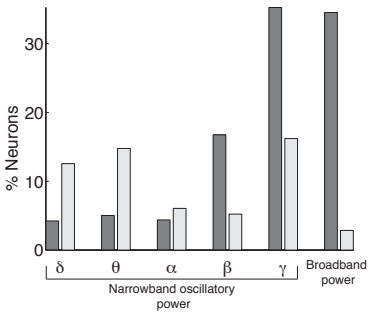


Figure 2.13: LFP components that predict firing rate. Dark gray bars indicate the percentage of neurons in each region that exhibited positive correlations between firing rate and a particular LFP feature; light gray bars show the percentage of neurons in each region that exhibited negative correlations. The bars on the left indicate the proportions of neurons whose firing rates were correlated with power in each narrow frequency band: delta (2–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–100 Hz). Each neuron may be counted in at most one direction (i.e., either positive or negative) per narrow frequency band. The bars on the right indicate the proportions of neurons whose firing rates were correlated with broadband power (i.e., broadband-shift neurons).

Electrophysiological models of memory

Oscillations play an important role in many neurophysiological models of human memory. By way of introduction to this literature, we will here review three major models that use frequency-specific neural activity to account for memory processes.

Lisman-Idiart-Jensen Model of Working Memory

Given the fact that neural signaling events such as action potentials occur on the order of tens or hundreds of milliseconds, a basic question about episodic memory that has concerned researchers is how these neural mechanisms can support the formation of associations between events that occur further apart in time. For example when listening to a professor's lecture, important information may be described over the course of tens of seconds or a minute. Although the information is spaced out in time, your brain is able to form a memory representation that associates all of the relevant information via strengthening of synaptic connections between neural assemblies representing each element of the memory. The fact that your brain can do this may not seem surprising, except that it is incompatible with long-term potentiation (LTP), the fundamental cellular mechanism that underlies hippocampal learning and memory (Markram, Lübke, Frotscher, & Sakmann, 1997), which serves to strengthen connections between cells that fire within $\sim$ 100 ms of each other.

The Lisman-Idiart-Jensen (LIJ) model (Jensen & Lisman, 2005) proposes that individual elements of a sequential memory trace are maintained in a cross-frequency working memory buffer. This maintenance occurs in the cortical structures (such as entorhinal cortex) that provide input to the hippocampus, where LTP is induced. Once per θ cycle, the network of cells that represents each item is activated. This is proposed to occur at the gamma frequency, with later items being represented at later sequential γ cycles.

Because γ has a frequency of roughly 30 to 120 Hz, this model has several qualities that make it attractive as an explanation for how working memory processes support long-term memory formation. The γ period roughly corresponds to the inter-item separation time that has been identified in the Sternberg task (Jensen & Lisman, 1998). There are also roughly seven γ cycles for each θ cycle, which implies a physiological limit of about seven items on the number of items that can maintained in working memory, consistent with classic work on working memory capacity limitations (G. A. Miller, 1956). Repeated activation of items at the γ frequency across multiple θ cycles also allows for items to be activated multiple times, increasing the strength of LTP.

Hasselmo Model of Hippocampal Encoding and Retrieval

Oscillatory synchronization has also been proposed to play a role in separating the processes involved in memory encoding and retrieval. This question is particularly relevant to theories of hippocampal function, since this structure is proposed to support both encoding and retrieval of episodic memories. However, the properties that would make the hippocampus a successful encoder of new memories, namely the ability to decorrelate similar inputs to create distinct memory representations, conflict with those necessary to retrieve memories given partial cues.

The Separate Phases of Encoding and Retrieval (SPEAR) model proposes that interactions between entorhinal cortex and hippocampus that support encoding and retrieval are instantiated at different phases of the theta rhythm. At the peak of theta, LTP is upregulated in the hippocampus, leading to association of presynaptic patterns in CA3 and postsynaptic patterns in CA1, both of which are driven by external input from entorhinal cortex.

In contrast at the trough of theta, activity generated by previously stored memories in CA3 will drive the response of CA1 and its output to entorhinal cortex. Because LTP is minimized at the theta trough, the pattern evoked by CA3 activity is not encoded as a new memory in CA1 and the system is biased for retrieval (Hasselmo et al., 2002).

Complementary learning systems

One class of neurophysiological model of episodic memory concerns the process by which memories undergo stabilization as a result of consolidation. Standard systems consolidation theory (Alvarez & Squire, 1994) was initially motivated by observations of temporally-graded retrograde amnesia, wherein patients show a gradient of forgetting of previously learned memories with those encoded just prior to the the amnesia-provoking injury showing the most forgetting, and memories experienced further back in one's personal history showing relatively normal forgetting. Systems consolidation theory proposes that newly-formed memory representations are encoded as patterns of activity in the hippocampus which, over time, come to be represented in cortical networks. The process by which this occurs is through information transfer from hippocampus to cortex, hypothesized to happen during offline rest periods and during sleep. Computationally, this process is necessary because the hippocampus and cortex are specialized to learn at different rates (i.e. synaptic plasticity occurs at different timescales) with the hippocampus able to rapidly acquire memory representations within a single trial, while cortex requires multiple presentations or exposures to patterns of information in order to acquire stable representations (Norman & O'Reilly, 2003). Physiologically, bursts of high-frequency activity in the hippocampus (sharp-wave ripples) are thought to reflect communication between the hippocampus and cortex, with low-frequency oscillations (e.g. θ) proposed to coordinate activity between hippocampus and cortex (Buzsáki, 2015; ?, ?; Siapas, Lubenov, & Wilson, 2005). Single-unit studies in rodents also provide evidence for systems consolidation theory, demonstrating replay during sleep of place cell sequences that were active during pre-sleep experiences

3
Event-related potentials

The event-related potential technique

We now turn to specific methods for using electrophysiological recordings to investigate cognitive processes. As we alluded to in Chapter 2, there are many approaches for analyzing brain activity and for relating it to other measures of interests (e.g., performance in a memory experiment). The calculation of event-related potentials (ERPs), however, has been such a dominant approach, that sometimes the term "ERP" is used interchangeably with "EEG" (e.g., by referring to an "ERP study" as if the ERP were the dependent measure rather than the EEG activity from which it is derived). The first reports of ERPs (then referred to as "evoked potentials")[1] in conscious humans appeared only 10 years after Hans Berger's publication describing the EEG in humans (Luck, 2014; Nisar & Yeap, 2014; Davis, 1939) and to this day ERPs remain the most popular approach for relating EEG data to performance in laboratory experiments.[2]

Figure 3.1 summarizes the ERP method. EEG activity surrounding a given event is made up of signal reflecting processing of that event as well as background activity ("noise") not related to the event. The goal is to separate signal from noise by sampling EEG activity relative to repeated exposures to the event of interest—the ERP is the average of the resulting time series. As a hypothetical example, imagine a recognition memory experiment of the sort illustrated in Figure 1.1 and suppose that processing of a probe item gives rise to three distinct deflections of the EEG activity (perhaps related to processing the visual features of a probe word, processing the meaning of the word, and initiating the process of determining whether the item had been studied, respectively). Imagine further that three distinct deflection of the EEG activity precede the execution of a response (perhaps these could correspond to the determination that it is time to execute a response, the finalization of the decision which response to make, and the initiation of a motor plan to execute the response). Figure 3.2 illustrates simulated background activity for a single trial i from such an experiment (n_i), simulated signals related to the stimulus and response onset on this trial (s_i), and the corresponding simulated EEG activity recorded at the scalp obtained by adding the signal to the background activity ($x_i = n_i + s_i$). Of course, in practice we will not be able to directly observe the background and signal activity. This example is somewhat contrived, but serves to illustrate the basic issue: on a single trial the underlying signal is so distorted by

[1] The first investigations of ERPs concerned early potentials that were thought to reflect exogenous (evoked) components, i.e., responses to the physical properties of a stimulus. With time, it became clear that internal states could profoundly affect brain activity in response to external events, leading researchers to propose additional endogenous (invoked) components reflecting psychological states (van Boxtel, 1998; Donchin, Ritter, & McCallum, 1979). The term "evoked potential" is still used, especially in clinical settings, to refer to a set of very characteristic potentials in the first $\approx$ 80 ms following a stimulus (Hillyard & Kutas, 1983). The distinction between evoked and invoked potentials, however, is not always clear, and we thus use the more neutral term ERP throughout.

[2] Our focus here is on EEG, but the same analysis approach is also widely used in magnetoencephalography (MEG) research where the resulting waveforms are known as "event-related fields" (ERFs). Beyond the laboratory, this approach also has a wide range of clinical applications and has even been applied to non-physiological time-series (e.g., event-related analyses of stock prices are known as "event studies"; MacKinlay, 1997).

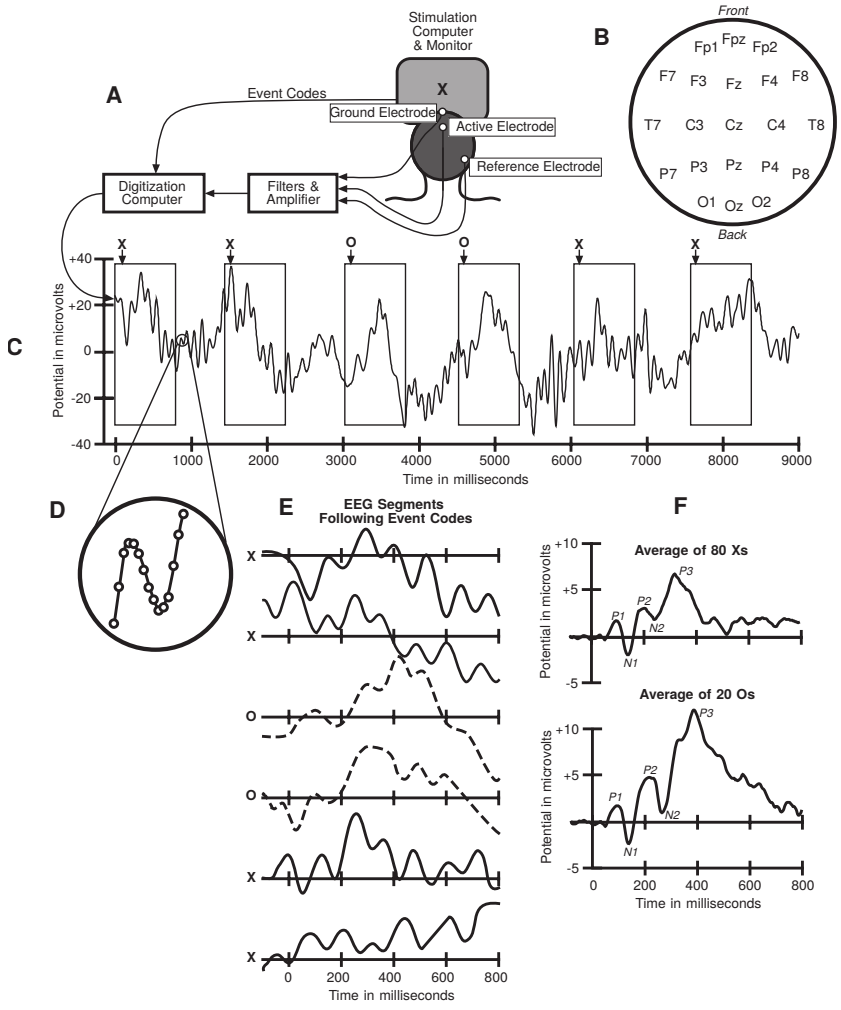


Figure 3.1: Example ERP experiment using the oddball paradigm. The subject viewed frequent Xs and infrequent Os presented on a computer monitor while the EEG was recorded from several active electrodes in conjunction with ground and reference electrodes (A). The electrodes were placed according to the International 10/20 System (B). Only a midline parietal electrode (Pz) is shown in panel A. The signals from the electrodes were filtered, amplified, and then sent to a digitization computer to be converted from a continuous analog signal into a discrete set of digital samples (D). Event codes were also sent from the stimulus presentation computer to the digitization computer, marking the onset time and identity of each stimulus and response. The raw EEG from the Pz electrode is shown over a period of 9 s (C). Each event code during this period is indicated by an arrow along with an X or an O, indicating the stimulus that was presented. Each rectangle shows a 900-ms epoch of EEG, beginning 100 ms prior to the onset of each stimulus. These epochs were extracted and then lined up with respect to stimulus onset (E), which is treated as 0 ms. Separate averages were then computed for the X and O epochs (F). Figure and caption from Luck (2014).

the ongoing background activity that it is effectively invisible in the EEG recording.

We refer to the deflections in the signal (**s**) as "components" and the goal of the ERP method is to estimate their properties from the recorded EEG activity (**x**). Our hypothetical example includes three components that relate to the processing of the probe item. As such, their timing should correlate strongly with the onset of the probe item but weakly with the onset of the subject's response. We further supposed three response-related components whose timing should be relatively invariant with respect to the response, but highly variable with respect to probe onset. To illustrate the power of the ERP method, we simulated a large number of such trials and computed stimulus-locked as well as response-locked ERPs (Figure 3.3). By averaging epochs locked to either stimulus or response onset, we can take advantage of the fact that background noise will cancel and estimate the corresponding components from the resulting ERP waveforms.

For this simulation, we generated random background activity and assumed that the signal associated with the three stimulus-related and the three response-related components consists of deflections made up of half

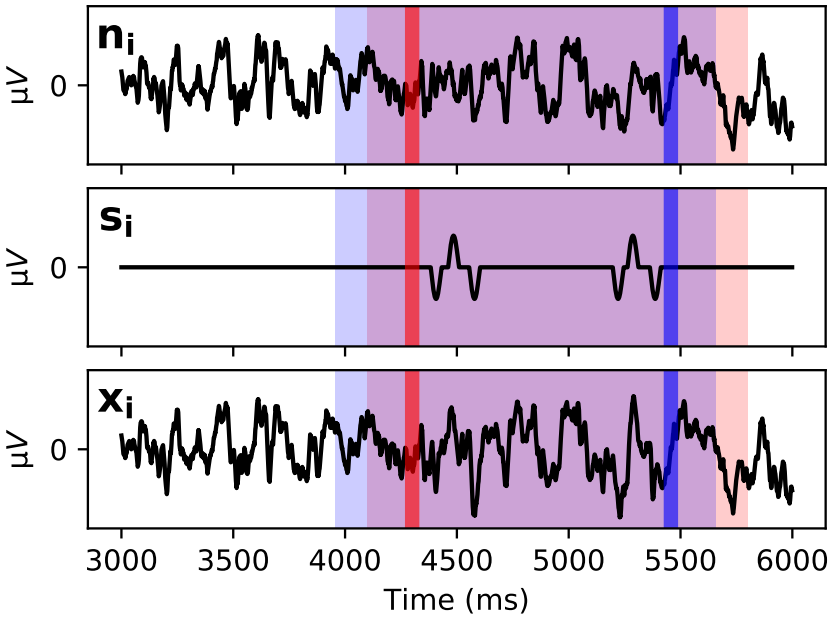


Figure 3.2: Simulated data illustrating the ERP method. Three seconds of simulated background activity and signal combine to form the observed EEG activity. Red and blue vertical lines indicate times of stimulus onset and response execution respectively. Data making up the corresponding event-related vectors $\mathbf{n}_i$, $\mathbf{s}_i$, and $\mathbf{x}_i$ are highlighted. In this example the stimulus-locked vectors contain samples from 200 ms before stimulus onset to 1500 ms after stimulus onset whereas the response-locked vectors contain data from 1500 ms before response execution until 200 ms after response execution.

a cycle of a sine wave with the first and last components causing a negative deflection and the middle component causing a positive deflection (see Figure 3.2). We assumed that response times varied uniformly between 300 and 1500 ms and further assumed that timing of the component most proximal to the respective event (stimulus onset for stimulus-related components and response onset for response-related components) varied uniformly between 60 and 100 ms relative to that event and that the timing of more distal components each independently varied in the same way relative to the timing of the previous component. Clearly none of these assumptions are particularly plausible, except for the fact that we expect some variability in biological systems. Despite the deliberate simplicity of our simulation, it serves to illustrate some basic properties of the ERP method.

Figure 3.3 shows the ERPs obtained by averaging various numbers (as indicated in the middle of each row) of stimulus-locked and response-locked epochs of simulated EEG activity. As the number of trials used to compute each ERP increases, a pattern of two negative deflections separated by a positive deflection emerges. The quality of the signal initially increases quite rapidly, but the increasing number of trials clearly has diminishing effects (note that the number of trials doubles from each row in Figure 3.3 to the next). This is a general property of the averaging procedure (discussed in more detail below) and important to keep in mind when attempting to minimize noise in the generation of ERPs.

Even though each epoch contained both a stimulus and a response, the response-locked components are not apparent in the stimulus-locked ERPs and neither are the stimulus-locked components in the response-locked ERPs. This illustrates that components whose variability is large relative to their extent will be lost in the averaging procedure (because across trials they will get averaged with adjacent components). The relatively small variabil-

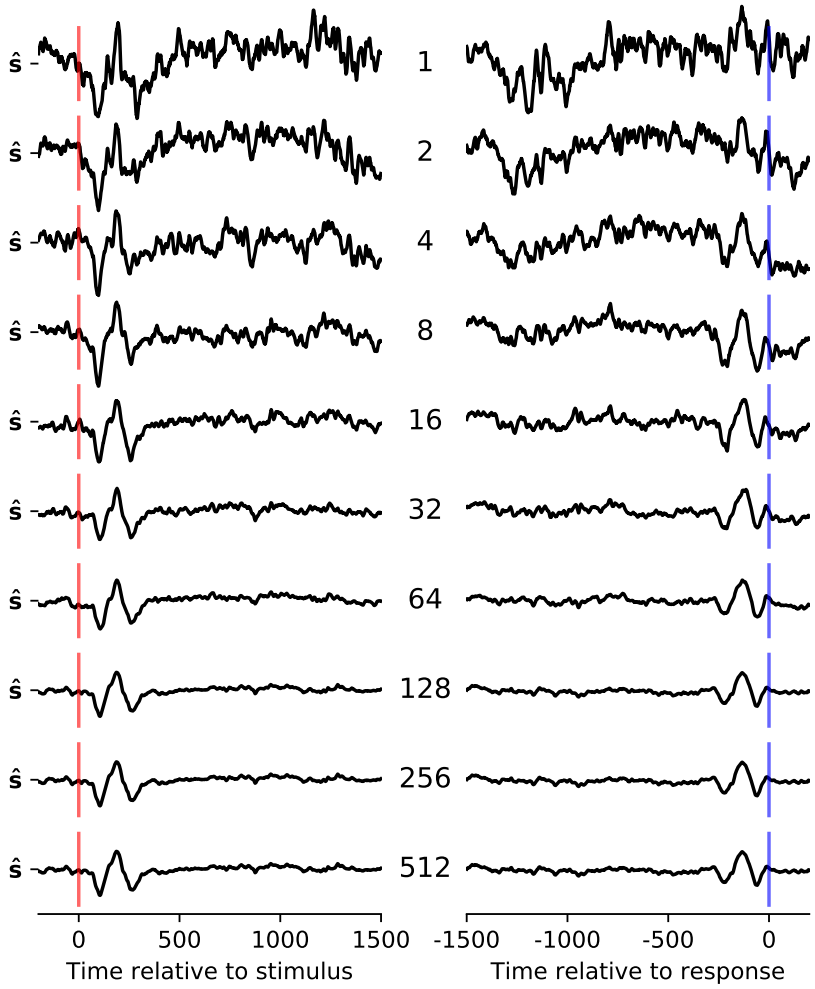


Figure 3.3: Simulated stimulus-locked and response-locked ERPs. Simulated stimulus-locked (left) and response-locked (right) ERPs obtained by averaging various numbers (indicated in the middle of each row) of simulated trials of the sort illustrated in Figure 3.2. Red and blue vertical lines indicate times of stimulus onset and response execution respectively. Note that the number of averaged trials doubles from each row to the next. $\hat{\mathbf{s}}$ indicates the estimated signal (i.e., ERP). See text for details of the simulation.

ity of the components that are related to the locking events results in them causing peaks and troughs in the ERP, but the resulting waveform is a distorted reflection of the underlying components. For example the fact that the timing of each simulated component varied independently with respect to the timing of the next more proximal component (or relative to the event offset in case of the most proximal component) means that the the variability of the components increased with distance from the locking event (because the joint variance of two independently varying events is equal to the sum of the individual variances). This results in a slightly attenuated amplitude and wider spread of the waveforms corresponding to the most distal component. Likewise, individual signals contained gaps between the negative and positive deflections, but these gaps do not appear in the average. Given that biological systems are inherently variable, the ERP waveforms can at best approximate the properties of the underlying components. Real ERPs exhibit fluctuations at relatively low frequencies with peaks and troughs growing increasingly broad as a function of time. Presumably this reflects similar processes with high-frequency fluctuations getting lost in the averaging due

to their variability and later components exhibiting relatively larger temporal variability than earlier ones. As we will discuss in more detail below, despite these limitations, ERPs reflect a remarkable number of psychologically relevant variables. In light of the above, it is however important to distinguish between ERP components and the ERP waveforms (which are a noisy reflection of those components surviving the averaging procedure), when using ERPs to inform psychological theory.

A formal background to the ERP method

The averaging procedure of the ERP method will only lead to a completely faithful representation of the underlying components under the following assumptions (Glaser & Ruchkin, 1976):

Linear combination of signal and noise: Just as the sound waves emitted by two speakers in the same room sum together in an audio recording, the assumption is that signal and noise simply sum together (rather than interact). Violations of this assumptions could arise from recording equipment, for example if the amplifier limits the signal to a maximum value which gets assigned to any sample exceeding this threshold.[3]

[3] This is known as "signal clipping".

Invariance of the signal: The signal must be identical for each repetition of the event for the average ERP to reflect the processing of individual events accurately. This assumption could be violated if participants habituate to an event, if the event is processed with variable latency and/or speed, or if the processes elicited by the event vary (e.g., a previously studied item used as recognition memory probe might trigger an elaborate recollective experience or more diffuse feelings of familiarity).

Noise is irregular: With respect the the event of interest, the contributions of noise to the recorded EEG activity must be irregular enough to be indistinguishable from independent samples of a random process. Events that occur at regular intervals could appear at consistent phases of oscillatory background activity, violating this assumption.

As the examples listed with each assumption (as well as our discussion above) suggest, it is unrealistic to hope that these assumptions are not violated in practice. Furthermore without independent means to distinguish between signal and noise, it is difficult to test these assumptions. However, careful experimental design can mitigate some potential problems (e.g., EEG studies usually contain randomly jittered delay periods to prevent the entrainment of background activity to experimental events) and, as we have discussed above, to the extent that violations are minor, ERPs will still provide useful information (Glaser & Ruchkin, 1976).

The EEG activity at a given channel (e.g., a pair of electrodes) can be expressed as a vector, $\mathbf{x}$, holding the recorded voltage samples in the order in which they were recorded. The following formal introduction to the ERP method and its properties follows closely that of Glaser and Ruchkin (1976) who provide additional details and derivations.

To generate an ERP, we partition the EEG activity to obtain a vector of EEG activity for each repetition, i, of the event of interest and assume that each such vector, $\mathbf{x}_i$, represents the sum of the signal, $\mathbf{s}$, and noise, $\mathbf{n}_i$: $\mathbf{x}_i = \mathbf{s} + \mathbf{n}_i$ (as indicated above, the signal is assumed to be invariant across repetitions of the event and thus does not have subscript indicating a par-

ticular instance of the event). The goal is to compute an estimate of $\mathbf{s}$, $\hat{\mathbf{s}}$, by averaging across the N repetitions of the event, yielding

$$\hat{\mathbf{s}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \mathbf{s} + \frac{1}{N}\sum_{i=1}^{N}\mathbf{n}_i \; . \tag{3.1}$$

Equation 3.1 defines the averaging at the heart of the ERP method. Without loss of generality we can assume that the expected value of the noise, $E\left[\mathbf{n}_i\right]$, is $\mathbf{0}$. We demonstrate that $\hat{\mathbf{s}}$ is an unbiased estimator of $\mathbf{s}$ by showing that its expected value, $E\left[\hat{\mathbf{s}}\right]$, is equal to $\mathbf{s}$:

$$E\left[\hat{\mathbf{s}}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\right] = \mathbf{s} + \frac{1}{N}\sum_{i=1}^{N}E\left[\mathbf{n}_i\right] = \mathbf{s}$$

The above discussion makes it clear that the precision of $\hat{\mathbf{s}}$ depends on the relative magnitudes of the signal and the average of the noise. Given that $\hat{\mathbf{s}}$ is an unbiased estimator of $\mathbf{s}$, we can expect that as N increases, $\hat{\mathbf{s}}$ approaches $\mathbf{s}$. We will now explore this relationship between the number of observations and the precision of $\hat{\mathbf{s}}$. The vector of standard errors of the estimated signal (i.e., the ERP), $\sigma_{\hat{\mathbf{s}}}$, contains the noise residual at each sample. Denoting the (diagonal) variance-covariance matrix of the noise, $E\left[\mathbf{n}_i\mathbf{n}_i^T\right]$, by Σ_n (the superscript T denotes the transpose; the off-diagonal elements of this matrix are zero due to the assumed independence of noise samples), we can first express the (also diagonal) variance-covariance matrix of the ERP, $\Sigma_{\hat{\mathbf{s}}}$, as a function of Σ_n and N and then obtain $\sigma_{\hat{\mathbf{s}}}$ by taking the element-wise square root of the main diagonal:

$$\begin{aligned}
\Sigma_{\hat{\mathbf{s}}} &= E\left[(\hat{\mathbf{s}} - \mathbf{s})(\hat{\mathbf{s}} - \mathbf{s})^T\right] \\
&= E\left[\sum_{i=1}^{N}\frac{\mathbf{n}_i}{N}\frac{\mathbf{n}_i^T}{N}\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}E\left[\mathbf{n}_i\mathbf{n}_i^T\right] \\
&= \frac{1}{N^2} \times N \times \Sigma_n
\end{aligned}$$

Thus

$$\sigma_{\hat{\mathbf{s}}} = \frac{\sigma_n}{\sqrt{N}} \; , \tag{3.2}$$

where $\sigma_n = \text{diag}\left(\Sigma_n\right)^{\circ\frac{1}{2}}$ and $\circ$ indicates that the square root is taken for each vector element.

The values in $\sigma_{\hat{\mathbf{s}}}$ denote the standard error at each sample and will vary to the extent that noise is nonstationary. It is worth taking a moment to reflect on what the above result means for efforts to increase the signal to noise ratio in ERP analyses. Clearly, all else being equal, more data leads to more precise estimates of the signal. The square root in the denominator of Equation 3.2, however, implies that to reduce noise by half, quadruple the amount of data is needed. This relationship highlights the importance of minimizing noise in the recording rather than relying solely on the averaging process to increase the signal to noise ratio (Luck, 2014).

Practical issues

The above example was deliberately simplistic to illustrate the basic issues associated with generating and interpreting ERPs. In practice, components can vary across many dimensions and be sensitive to a wide range of variables including stimulus properties, experience, or task demands. Even with the highly regular background activity and component properties in the simulation, however, we observed distortions in the average waveforms. In real EEG recordings, potentials can drift over time due to factors not related to brain activity which could cause serious distortions in the resulting ERPs if not appropriately countered (Luck, 2014). A generally effective strategy for eliminating this source of noise is the application of a *baseline correction*. The idea is that the period just prior (or, usually in the case of response-locked ERPs, just after) the event should be relatively free of event-related activity. One can thus center individual waveforms such that the average potential of the baseline activity is 0 μV. With this common preprocessing step, the amplitudes of the various deflections in the ERP waveforms represent the average difference from baseline rather than the average absolute potentials.

The inclusion of a (typically 100–200 ms) baseline period is also useful for assessing the success of the averaging procedure in eliminating background activity. To the extent that brain activity during the baseline period is not event-related, variability in the this period reflects noise and any deflections outside the baseline period that do not exceed those in the baseline period are unlikely to represent event-related activity (Luck, 2014; Woodman, 2010).

ERPs are frequently visualized by graphing individual waveforms with time on the abscissa and voltage on the ordinate (see, e.g., Figure 3.1F). It so happened that the first published ERPs (Davis, 1939) were plotted with what would conventionally be viewed as an inverted y-axis (i.e., positive voltages were plotted below zero and increased in the downward direction with voltages decreasing in the upwards direction). This quirk caught on with a large number of ERP researchers, but increasingly the trend is to follow the standard convention of having values increase in the upwards direction (Luck, 2014). As a result, it is important to pay careful attention to to orientation of the y-axis in published ERP waveforms and to clearly indicate the axis orientation when illustrating ERPs.

ERP components

ERP researchers have identified a wide range of components on the basis of systematic variations of ERP waveforms in response to experimental manipulations. Unfortunately naming conventions for these components are somewhat inconsistent. For many components the general pattern is that the name starts with the letter "P" or "N" to indicate a positive or negative deflection in the associated ERP waveform respectively[4] followed by a number that either indicates an ordinal position (as in "N2" for the second negative deflection; see also Figure 3.1F) or an approximate time in ms corresponding to the peak of the deflection (as in "P300" for a positive deflection peaking around 300 ms after the event onset).[5] Other components have more descriptive names such as the "lateralized readiness potential" (LRP), a response-related component that distinguishes which hand will execute a response or the contingent negative variation (CNV) a negative deflection thought to

[4] For magnetic fields recorded in magnetoencephalography experiments, the letter "M" is used instead.

[5] This labeling is not as intuitive as it might appear. The polarity of a component can vary with the placement of electrodes and choice of reference. Additionally, not all components are visible from all electrode sites and thus the ordinal indicator might not correspond to the deflections observed in a specific waveform. Similarly, the timing indicator need not be be very exact, especially for later components (e.g., positive deflections with a peak as late as 600 ms are often labeled as P300). A further complication arises from the fact that some (early) components are modality-specific and thus the label "P1" refers to different components depending on whether it was recorded relative to a visual or an auditory stimulus (Luck, 2014).

index anticipatory processes (Luck, 2014).

Earlier components tend to be highly sensitive to perceptual features of the eliciting stimuli whereas later components tend to covary more strongly with internal states (e.g., those that reflect task demands). One of the most widely studied components of the latter kind is the P300 (also known as P3 or P3b) which we briefly introduced in Chapter 2. One reason for the appeal of this component is that it is sensitive to experimental contingencies that require the categorization of the eliciting stimuli along task-defined dimensions. This makes this component useful for the study of a wide range of cognitive processes and allows it to establish an upper bound on the duration of the processes responsible for this categorization (Luck, 2014). An example is shown in Figure 3.1F where this component (labeled P3 in the figure) is considerably larger for rare stimuli.

In practice the distinction between different components is often tricky, because there can be considerable overlap (and variability) in their timing. Additionally, earlier deflections can have lasting effects on the ERP, making it difficult to unambiguously attribute ERP differences to later components when earlier differences are also apparent. For example, one might want to compare memory for faces with that for names by presenting images of faces and strings of letters spelling out the names on a computer screen with the instruction that these be memorized for a subsequent memory test. Comparing ERPs locked to the presentation of subsequently remembered faces with those locked to the onset of subsequently remembered names will reveal differences associated with the different perceptual properties of these stimulus types (e.g., it is known that that faces elicit an enhanced early negativity known as the N170 component) in addition to any differences associated with memory processes that are sensitive to the stimulus class. It is therefore important to structure comparisons between ERP waveforms such that the process of interest is isolated.

ERPs and human memory

In recognition memory tasks (see Figure 1.1) one can observe ERPs relative to the probe items (just as we proposed in our hypothetical example above). ERP waveforms distinguish between correctly identified old ("hits") and new ("correct rejection") probe items between around 300–500 ms at mid-frontal electrodes (this effect is sometimes called FN400 where the "F" prefix specifies the frontal scalp distribution) and between around 400–800 ms at left-parietal electrodes (see Figure 3.4; Wilding & Ranganath, 2012; Luck, 2014). Dissociations between these two components have prompted interpretations of these two components within the framework of dual-process theories of recognition memory that postulate that two distinct types of evidence drive recognition decisions: *familiarity* and *recollection* (Yonelinas, 2002; Yonelinas, Aly, Wang, & Koen, 2010; Malmberg, 2008). Familiarity refers to the diffuse sense that an item has been studied and the size of the mid-frontal old-new effect (FN400) is often interpreted as an electrophysiological index of this construct. Recollection, on the other hand, indicates the ability to retrieve contextual details associated with the study episode. The size of the left-parietal old-new effect is commonly thought to covary with recollective experience (Curran, 1999; Rugg & Curran, 2007; Wilding & Ranganath, 2012; Luck, 2014). Recent work by some of us, however, suggests that EEG

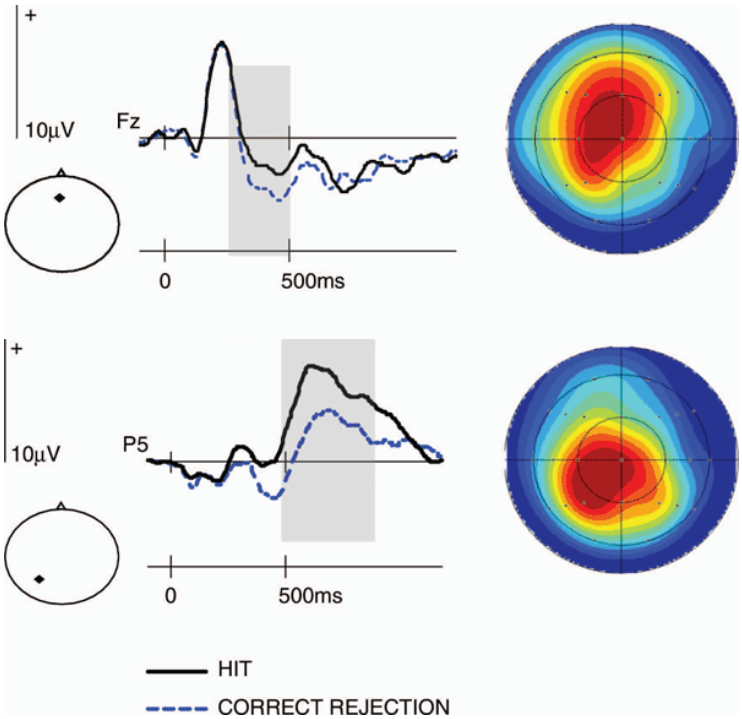


Figure 3.4: ERP old-new effects. An early mid-frontal (sometimes called FN400; top panel) and a late left-parietal (bottom panel) ERP old-new effect exhibiting greater positivity for hits (i.e., old items that were correctly recognized as old; black solid line) than for correct rejections (i.e., new items that were correctly classified as such; blue dashed line). The locations of the electrodes from which the ERPs were obtained (Fz and P5 respectively) are indicated in a small top down schematics of the head (the triangle at the top of the circle is a stylized nose and the dot inside the circle indicates the electrode location). On the right of each panel is a overhead scalp map (top is frontal) indicating the size of the difference between the ERP waveforms for hits and correct rejections across the entire scalp in the time windows highlighted in gray (300–500 ms and 500–800 ms respectively). Red indicates the largest difference and cooler colors indicate smaller differences. These plots confirm that the scalp distribution of the early effect is strongest at mid-frontal electrodes whereas the scalp distribution of the latter effect is more pronounced at left-parietal electrodes. Adapted from Wilding and Ranganath (2012).

activity reflects a unitary recognition signal combining the available sources of evidence distinguishing between old and new probe items (?, ?).

One can also study the processing of items during study as a function of subsequent memory. Such differences are known as *Dm* ("difference due to memory") or *subsequent memory effects*. Subsequent memory effects in ERPs typically do not manifest as distinct peaks, but instead as sustained positive deflections at centro-parietal electrodes for subsequently remembered items. These differences start around 400 ms after the onset of the study item and last several hundred milliseconds. Additional subsequent memory effects have been observed at left anterior electrodes and can be sensitive to the types of stimuli that are being remembered (Luck, 2014).

Experiments designed to investigate the properties of visual working memory, are particularly well suited to the ERP method. Trials in these experiments typically consist of a brief visual display followed by a fixed retention interval that ends with a test array (see Figure 3.5A). In an influential study Vogel and Machizawa (2004) identified what has become known as the Contralateral Delay Activity (CDA)—a sustained negative deflection of the ERP contralateral to the visual hemifield containing the memory set. In this experiment participants were asked to memorize colored squares in one hemifield of the display and, after a retention interval, were presented with a test array of squares that was either identical to the memory array or included a change to the color of one square in the cued hemifield (see Figure 3.5A). Figure 3.5B shows ERPs for electrodes ipsilateral and contralateral to the cued hemifield and Panel C of this figure indicates that the size of the difference between these waveforms rose with the size of the memory set,

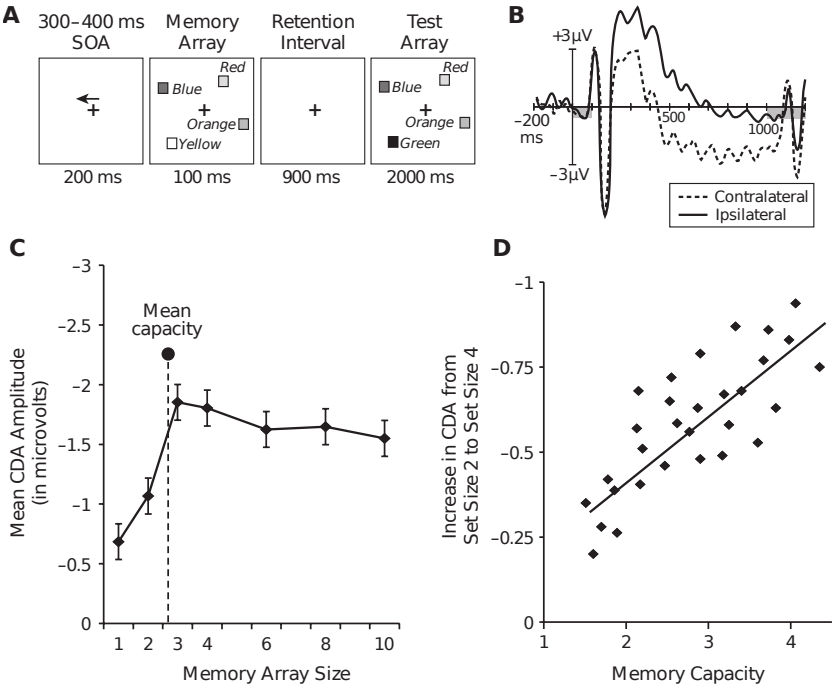


Figure 3.5: Working memory effects in ERPs (Vogel & Machizawa, 2004). (A) Design of the visual working memory task. An arrow cued the side on which square colors were to be memorized. (B) Grand average ERP waveforms for a memory array of size 4 (i.e., 4 items in the cued hemifield were to be remembered, 4 additional items in the uncued hemifield never changed). ERPs were time-locked to the onset of the memory array and averaged across electrodes at lateral occipital and posterior parietal locations. Gray rectangles indicate the presence of the memory and test array respectively. (C) Mean amplitude of the contralateral delay activity (CDA) between 300–900 ms after onset of the memory array as a function of the size of the memory array. Mean visual working memory capacity (2.8 items, estimated from response patterns) is indicated with a dashed vertical line. Error bars indicate 95% confidence intervals. (D) Visual working memory capacity plotted against increase in CDA amplitude between memory arrays of size 2 and 4 ($r = 0.78$). All ERPs and derived measures shown here are based on all trials, but follow-up work suggested that results are very similar if ERPs are only generated for trials with correct responses (Vogel, personal communication, Oct. 4, 2017). Adapted by Luck (2014) from Perez and Vogel (2012). Copyright 2012 Oxford University Press.

but only until memory capacity (derived from the patterns of errors in this task) was reached, ruling out that this effect was simply a response to the increased number of displayed squares. Vogel and Machizawa (2004) also related the size of the increase of the CDA between memory sets of 2 and 4 for each individual to the corresponding working memory capacity (derived from each individual's pattern of errors) and found a striking correlation between these measures (shown in Figure 3.5D).

Quantifying properties of ERP waveforms

To allow ERPs to inform psychological theory, it is necessary to quantify how they vary with time and experimental condition. The early parts of ERPs are usually characterized by distinct peaks and troughs and some studies attempt to quantify corresponding amplitudes and/or latency. This is trickier than it might appear. For example, simply selecting the maximum in a time-window could produce spurious results if noise in the waveform causes high amplitude blips (which may not coincide with the peak of the component). Likewise, if the time window is not carefully chosen, the maximum might correspond to a value in the rise towards a subsequent peak rather than representing the summit of the component of interest. More sophisticated procedures for estimating the amplitude and latency of peaks and troughs exists, but it is worth bearing in mind that there is nothing inherently special about these points. Alternative measures, such as when the waveform first deviates from baseline, can be better suited for relating components to underlying processes and, indeed, are often used, especially when the aim is to put an upper bound on the timing of a specific process associated with a given ERP component (for a comprehensive review of different measures to quantify properties of ERP waveforms, see Chapter 9 of Luck, 2014).

A particularly popular measure is the area between the waveform and o μV (or between two waveforms) within a given time interval. An important reason for the popularity of the area measure is that it is relatively robust to small fluctuations in the waveforms, but for later components that no longer exhibit clearly defined peaks and troughs, there are few alternatives (see Figure 3.5 for an example of the use of the area measure to study the electrophysiology of working memory).

The role of ERPs in memory research

ERPs are particularly well suited for studying neural activity immediately following (or preceding) well-defined events. Encoding and retrieval processes, however, can extend over relatively long time periods and are not always easy to link to specific events. For example, the presentation of a study item may prompt the retrieval of a previously studied item and deliberate rehearsal processes are difficult to control (see Chapter 1 for a discussion of these issues). As indicated above, ERPs tend to reflect low frequency deflections of the EEG activity, yet spectral features of the EEG activity that are usually lost in the ERP have also been shown to contain information about episodic memory processes (Nyhus & Curran, 2010; Jacobs, Hwang, Curran, & Kahana, 2006).

Despite these limitations, ERPs have been reliably linked to performance in memory tasks and are almost certainly the most popular method in investigations of the electrophysiology of human memory. Undoubtedly part of this popularity is due to the the simplicity of the computations that are required for the generation of ERPs and to the substantial prior literature establishing the properties of ERPs in a wide range of experimental contexts. The rapid increase in the availability of more powerful computational resources, however, has led to an increasing use of methods that investigate neural activity on a trial-by-trial basis and that also consider spectral features that are lost in the generation of ERPs. It is likely that the use of these methods (which will be the foci of subsequent chapters) will continue to gain in popularity and perhaps even surpass the use of ERPs in the coming decades. As, we hope, will become obvious in the course of reading this book, the various methods for investigating neural activity have complementary strengths and weaknesses. A comprehensive understanding of the electrophysiology of human memory is therefore likely to depend on the results of investigations using a wide range of methods and the aim of this and the coming chapters is to highlight those that have to date been proven to be most promising for this endeavor. Before we turn to other univariate methods for analyzing brain activity, however, we provide a brief overview over statistical issues that arise in the analysis of electrophysiological data and the main methods for addressing them.

Example: ERPs in short-term item recognition

Students of memory have devised a variety of methods for probing short-term memory (see Postle & Oberauer, in press; Foster, Vogel, & Awh, in press, for reviews). Here we consider Sternberg's memory scanning task. Other experimental paradigms used to probe short-term memory (sometimes called *Working Memory* include immediate serial recall (as in recalling a once-

presented phone number in forward order), *N*-back recognition (matching a current test item to an item presented *N* items before, and visual working memory (matching a multi-attribute perceptual stimulus to an array of previously presented stimuli).

Saul Sternberg originally developed this task to study reaction times in memory search (Sternberg, 1966) and early data supported the idea that in some cases subjects perform a sequential serial scan of the contents of a short-term memory buffer (see Chapter 1). Sternberg gave his subjects a short list of items and then, following a brief delay, subjects saw a test probe and responded to indicate whether or not it was on the list. As an example of one such list, a participant might study a list of four consonants (*B*, *M*, *F*, and *S*). After a 1- to 2-sec delay, the participant would judge whether the letter *M* was in the just-presented list. With short lists, comprising one to six items, practiced participants made very few errors (often less than 3%) and responded rapidly (often in less than half a second).

Models of the cognitive processes involved in short-term recognition focus on the period immediately following the presentation of the test, or probe, item. At that moment, the subject must interrogate their memory for the just presented list which they have somehow stored in their memory system. Sternberg proposed that subjects compare the test item to each stored memory in serial (sequential) fashion and that they respond after making all of the comparisons (exhaustive search). Other theorists have suggests different models of the task, including models that allow for partial matches between the test item and the stored exemplars in memory CITES and models that assume that subjects do not have perfect separation between the most recent list and other material in memory CITES.

Statistical analyses of electrophysiological data

Several practical issues arise when determining the parameters for the analysis of ERP waveforms and other types of electrophysiological data. For example, to determine the time window(s) for which the area under an ERP waveform should be computed, it might be tempting to plot the ERPs, and to determine the time window(s) on the basis of the shape of these waveforms (e.g., if the difference between the ERPs for two experimental conditions appears particularly large between 450 and 650 ms after event onset, one might wish to calculate statistics on the area between the corresponding ERPs within these two time points). Until not very long ago, this approach was quasi-accepted practice and, unfortunately, the associated problems are still not universally appreciated. In short, the issue with picking analysis parameters on the basis of explorations of the data that are to be analyzed is that it severely inflates the probability of finding an effect when none is present(Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Kilner, 2013).[6] A related concept is that of *overfitting* which we will discuss in detail in Chapter 5. In null hypothesis significance testing (NHST), the *p*-value indicates the probability of obtaining the observed results (or more extreme results) given that the null hypothesis is true.[7] The validity of this *p*-value, however, depends on unbiased sampling of the data. In cases where the time window (or other analysis parameters) are informed by explorations of the same data set, the probability of observing an effect that is at least as large as the observed one, given that the null hypothesis is true, may be substantially larger than

[6] This approach to data analysis is reminiscent of the proverbial Texas sharpshooter who shoots at a wall and then paints a bull's-eye around the bullet hole.

[7] If this definition of the *p*-value surprises you, you may have been led astray by ubiquitous erroneous definitions that sometimes even make it into textbooks. The *p*-value is emphatically not the probability that the null hypothesis is true, an index of the likelihood that the results will (fail) to replicate, or some variation on these or similar themes.

the p-value indicates.

One way to avoid this problem is to determine analysis parameters *a priori*, for example on the basis of an independent data set. A complementary approach is to test multiple analysis parameters, for example by testing multiple time windows (determined *a priori*!) or doing away with time windows altogether and calculating separate statistics for each sample. Because of the complexity of electrophysiological data sets (for each event, many samples are often collected across a large range of sensors), such data sets are routinely subjected to multiple statistical tests. This raises related statistical issues which we will discuss next.

Multiple comparisons

In NHST, a threshold, α, is set (most commonly to $\alpha = 0.05$) such that results with p-values that fall below this threshold are deemed statistically significant (i.e., the null hypothesis is rejected). In this framework one can make two types of errors (see Table 3.1): A Type I error refers to erroneously rejecting the null hypothesis when there is no effect. In contrast, a Type II error refers to an erroneous failure to reject the null hypothesis when an effect is present.[8] The α-level directly controls the Type I error rate for a given statistical test, but, all else being equal, reducing the Type I error rate increases the Type II error rate and vice versa. When conducting multiple statistical tests, it is usually desirable to either control the probability that any of the tests leads to a Type I error (this is the *family-wise error rate*, FWER) or to limit the total proportion of Type I errors across all tests (this is the *false discovery rate*, FDR).

To illustrate the issue, imagine attempting to rappel from a burning building using a makeshift rope made out of bed sheets. If you knew that each knot joining two bedsheets independently had a 5% chance of failure, would you prefer using 2 long bed sheets or 4 short bed sheets to make the rope for your escape? With no other potential points of failure, it should be clear that a rope with fewer knots is the safer escape option. In this case, assuming a failure of the rope is fatal, you will have a 95% chance of making it out alive with 1 knot, but a less than 86% ($0.95^3 \times 100$) chance of survival with 3. With 14 knots or more, the rope is more likely to fail than to hold despite each individual knot only having a 5% failure rate. Likewise, imagine calculating ERPs for 2 experimental conditions and running separate statistical tests for the difference between them at each sample across participants. Data from a 1 s period sampled at 1000 Hz would produce a FWER very close to 1.0 ($1 - 0.95^{1000}$ to be exact) if the tests were independent. Below we will discuss different methods for controlling the FWER or the FDR across multiple statistical tests. With the exception of the Bonferroni correction, which is not generally used when correcting for large numbers of tests, these methods capitalize on dependencies between the statistical tests[9] to limit either FWER or FDR without disproportionately increasing the Type II error rate.

Many experimental designs that include multiple independent variables are routinely analyzed within an Analysis of Variance (ANOVA) framework. A common misconception is the idea that the use of an ANOVA instead of, say, calculating individual t-tests avoids such issues with multiple comparisons. It is true that the ANOVA controls Type I error levels for each statistical test so that the number of factor levels does not affect the Type I er-

[8] This situation is analogous to that in recognition memory tests where one can erroneously identify a new item as old or fail to recognize an old item as old. Rather than referring to these errors as Type-I and Type-II errors respectively, they are usually labeled as "false alarms" and "misses" in the context of recognition memory tests or other binary choice tasks.

	Fail to reject H_0	Reject H_0
H_o is true	✓	Type I error (α)
H_0 is false	Type II error (β)	✓

Table 3.1: Decision table for an individual null hypothesis significance test. The significance threshold α controls the Type I error rate. The inverse of the Type II error rate ($1 - \beta$) is known as the power of a statistical test. For multiple statistical tests controlling the FWER limits the probability of making at most 1 Type I error across the family of statistical tests whereas controlling the FDR limits the expected proportion of Type I errors among all significant results (i.e., the proportion of erroneous rejections of H_0 among all rejections of H_0 across the family of statistical tests).

[9] In electrophysiological data, measures are usually highly correlated across nearby time points and sensors introducing substantial dependencies between associated statistical tests.

ror level for the test of the corresponding main effect. However, in ANOVAs we typically compute multiple statistical tests and the Type I error rate for this family of tests is not controlled. An ANOVA with 2 factors usually involves tests of the 2 main effects and a test of their interaction for a total of 3 statistical tests. The addition of just 1 additional factor more than doubles the number of statistical tests to 7 (3 main effects + 3 2-way interactions + 1 3-way interaction). When ANOVAs are used in analyses of electrophysiological data, it is not uncommon to add additional factors (e.g., for electrode locations across the anterior-posterior and/or the left-right dimensions, or for multiple time windows).[10] With 4 factors (corresponding to a total of 15 statistical tests) or more, the FWER exceeds 0.5 making it more likely than not that at least one of these tests produces a significant result even when all null hypotheses are true. It is therefore important not to be fooled into believing that issues of multiple comparisons do not apply when computing "only one" ANOVA. Especially for complex ANOVA designs with 4 or more factors, it can be prudent to apply the kinds of controls of the FWER or FDR that we introduce below (see Bishop, 2014, for a more thorough discussion of these issues).

[10] Such use cases usually violate the ANOVA assumption that samples at different factor levels are independent from each other.

Bonferroni and Holm correction to control the FWER

The simplest way to control the FWER is to divide the uncorrected significance threshold α by the number of statistical tests, m, to compute a corrected threshold value α/m on which the statistical decisions are based; this procedure is known as the Bonferroni correction. For 100 statistical tests and $\alpha = .05$, the corrected significance threshold would be $0.05/100 = 0.0005$ and only tests with p-values below this threshold would be considered significant at the (family-wise) .05 threshold.[11] The Bonferroni correction does not consider dependencies between the statistical tests and is therefore very conservative in situations such as those discussed in this chapter. This reduces the power to detect effects in the data, especially among a large number of tests, and therefore the Bonferroni correction is not typically used for electrophysiological data.

A less conservative alternative to the Bonferroni correction is a procedure developed by Holm (1979). For this correction, all p-values are ordered from smallest to largest and one finds the smallest index, i in the resulting list of p-values for which $p_i > \alpha/(m - i + 1)$, where m is the number of statistical tests. The null hypotheses corresponding to the p-values at index $i - 1$ or below are rejected. This procedure can also be used to calculate adjusted p-values ($\tilde{p}$) as follows (Yekutieli & Benjamini, 1999):

$$\tilde{p}_i = \max_{k=1,\ldots,i} \Big\{ \min((m - k + 1) \times p_k, 1) \Big\}.$$

Table 3.2 summarizes the threshold values at each step of both FWER control procedures. Even though Holm's (1979) procedure is less conservative than the Bonferroni correction, it still leads to very low significance thresholds when the number of statistical tests is large (see Figure 3.6 for an illustration) and thus limits the power to find effects. In practice it is therefore often better to limit the FDR instead of the FWER or to use non-parametric shuffling procedures to limit the FWER, both of which we will discuss below.

[11] Proof that the Bonferroni correction controls the FWER: Suppose that we are conducting a total of m statistical tests. Let p_i, $i \in \{1, \ldots, m\}$ be the corresponding p-values and I_0 the set of indices of the true null hypotheses. The FWER is the probability of making at least one Type I error,

$$\begin{aligned} \text{FWER} &= P(\bigcup_{I_0} \{p_i \leq \alpha/m\}) \\ &\leq \sum_{i \in I_0} P(p_i \leq \alpha/m) \\ &\leq |I_0| \times \alpha/m \leq \alpha \end{aligned}$$

Bonferroni	Holm
α/m	α/m
α/m	$\alpha/(m-1)$
$\vdots$	$\vdots$
α/m	α

Table 3.2: Table of significance thresholds at each step of two FWER control procedures. p-values are sorted from lowest to highest and are compared to the corresponding significance thresholds at each step.

Controlling the False Discovery Rate

In situation where a single Type I error is less problematic than the failure of a single knot in the hypothetical scenario sketched out above, controlling the FDR instead of the FWER can often represent a good compromise between limiting the number of both Type I and Type II errors. Controlling the FDR constitutes *weak control* of the Type I error rate, because, on average, we would expect 5% of the statistically significant tests to result in a Type I error. Thus FWER control implies FDR control (but not vice versa) and when all null hypotheses are true, controlling the FWER and the FDR are equivalent (Dudoit, Shaffer, & Boldrick, 2003).

The two main methods for controlling the FDR are that by Benjamini and Hochberg (1995) and that by Benjamini and Yekutieli (2001). The former is less conservative (and more popular), but has been shown to control the FDR only when the individual tests are not negatively correlated (Benjamini & Yekutieli, 2001). For this procedure p-values are sorted from smallest to largest and the largest index i in the corresponding list of sorted p-values for which $p_i \leq (i/m) \times \alpha$ is identified. The null hypotheses at this and lower indices are rejected. Adjusted p-values ($\tilde{p}$) are calculated as follows (Yekutieli & Benjamini, 1999):

$$\tilde{p}_i = \min_{k=i,\cdots,m} \left\{ \min(m/k \times p_k, 1) \right\}$$

The assumption that statistical tests are not negatively correlated can be problematic. For example, some electrodes might be placed at opposite ends of a dipole, which could lead positive deflections at one electrode to coincide with negative deflections at the other. An alternative method to control the FDR that is valid even for these cases is to identify the largest index i in the list of sorted p-values for which $p_i \leq (i/(m \times \sum_{j=1}^{m} 1/j) \times \alpha$ (Benjamini & Yekutieli, 2001). As with the other procedure, null hypotheses at this and lower indices are rejected. The adjusted p-values ($\tilde{p}$) for Benjamini and Yekutieli's (2001) FDR correction is as follows:

$$\tilde{p}_i = \min_{k=i,\cdots,m} \left\{ \min(\sum_{j=1}^{m} \frac{1}{j} m/k p_k, 1) \right\}$$

Table 3.3 summarizes the significance thresholds for these two FDR correction procedures. These thresholds differ by the harmonic sum term $\sum_{j=1}^{m} 1/j$, which is approximately $\log(m)$ when m is large. In practice, the method by Benjamini and Yekutieli (2001) is seldom used, because it leads to substantially higher levels of Type II errors than that proposed by Benjamini and Hochberg (1995). Furthermore, Benjamini and Hochberg's procedure appears to be relatively robust to violations of the assumptions that tests are not negatively correlated (Clarke & Hall, 2009; Groppe, Urbach, & Kutas, 2011). Figure 3.6 shows significance thresholds for each comparison across the different methods for controlling the FWER and the FDR discussed so far.

Permutation Tests

Above we provided the definition of a p-value as the probability of obtaining the observed (or more extreme) results given that the null hypothesis is true. The idea of permutation tests is to calculate this probability non-parametrically (i.e., without making any assumptions about the shape of the

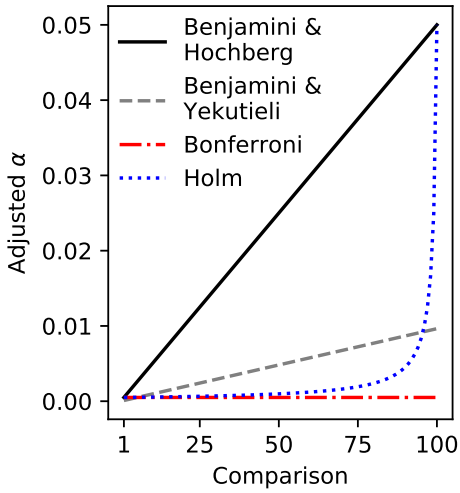


Figure 3.6: Threshold functions under different multiple comparison procedures. Significance threshold are shown for each of 100 tested hypethesis (sorted from lowest to highest p-value). The threshold for the Bonferroni correction is constant across comparisons, whereas thresholds for the FDR control procedure proposed by Benjamini and Hochberg (1995) increase linearly with threshold. Corresponding thresholds for the other procedures fall between these two functions.

Benjamini & Hochberg	Benjamini & Yekutieli
α/m	$\alpha/(m\sum_{j=1}^{m} 1/j)$
$2\alpha/m$	$2\alpha/(m\sum_{j=1}^{m} 1/j)$
$\vdots$	$\vdots$
α	$\alpha/(\sum_{j=1}^{m} 1/j)$

Table 3.3: Table of significance thresholds at each step of two FDR control procedures. p-values are sorted from lowest to highest and are compared to the corresponding significance thresholds at each step.

sampling distribution) by generating many new samples from the observed data under the assumption that the null hypothesis holds. Imagine observing the area under a specific portion of the ERP waveform in 10 individuals who were each subjected to two experimental conditions. If the null hypothesis that there is no difference between the experimental conditions is true, we can permute the condition labels to generate many hypothetical data sets that come from the same distribution as the actually observed data. For each individual, there are two possible assignments of condition labels to conditions (the correct assignment, and the permuted assignment) leading to a total of 1024 (2^{10}) possible permutations, one of which corresponds to the actually observed data set (this permutation test is a non-parametric analog to a paired t-test, taking into account that every individual provides data for both conditions). For each permutation we could calculate a t-statistic of the difference in area under the two ERP waveforms corresponding to the respective condition labels. If the null hypothesis is true, the difference in areas for the actually observed condition labels should come from the same distribution as the differences for the permuted condition labels and we would expect the t-statistic for the actually observed difference to fall near the center of the distribution of t-statistics corresponding to the differences for the permuted condition labels. The p-value for this permutation test simply corresponds to the proportion of t-statistics in this distribution that are at least as extreme as the actually observed t-statistic. Thus, if the actually observed t-statistic is among the $\alpha \times 100\%$ most extreme t-values, we can reject the null hypothesis that the areas for the two experimental conditions are identical.

With larger number of observations, it can be impractical to compute the distribution of the relevant statistic for each possible permutation. It is reasonable in these cases to instead use a large number of random permutations. Whatever the number of permutation, the corresponding p-vale is determined by the rank of the actually observed statistic within the permutation distribution and thus the number of permutations determines the resolution of the resulting p-value: For n permutations, the smallest p-value we can compute is $1/n$.

This approach is quite flexible and not limited to, say, the computation of t-statistics for each permutation. It is, however, easy to bias the results by not setting up the permutation procedure carefully. For example, in the above example, one might have wanted to directly permute the vector of 20 condition labels across all individuals.[12] These permutation would include many cases where both observations from an individual are assigned the same condition label and thus the difference statistic computed for each permutation would reflect a substantial amount of between-subject variance (i.e., this permutation test would not correspond to a paired t-test, because it does not take into account dependencies between multiple observations from the same individual).

So far we have described how to set up a permutation test for an individual hypothesis. For testing multiple hypotheses using the permutation test, one can apply the same procedures for controlling the FWER or the FDR to the p-values from a family of permutation tests. However, there is a way to incorporate control of the FWER directly into the permutation procedure. For each permutation one can construct the distribution of the *maximal statistic* (Nichols & Holmes, 2001) and then compare the statistic for the actually observed data against this distribution. For example, consider the example

[12] This would result in 20! (more than two quintillion) possible permutation and thus would be good example for a situation where it would be advisable to use a random sample of all possible permutations.

above, but instead of comparing two areas, we are interested in calculating the difference between the two ERP waveforms at each sample over a 1 s period sampled at 1000 Hz. For each permutation we can now calculate 1000 difference statistics (one for each sample) and retain the largest. We then compare all difference statistics for the actually observed data to this distribution of maximal statistics. Calculating the *p*-values by determining the proportion of more extreme statistics in this distribution of maximal statistics controls the FWER.

Again, it is easy to bias the results if one does not set up the permutation procedure carefully. For example, in the above use-case, one might be tempted to permute the condition labels separately for each sample across the 1 s period. In this case the difference statistic at each sample would be computed from independent assignments to condition labels and the distribution of maximal statistics would reflect between-subject variability in the auto-correlational structure of the time series.

Maris and Oostenveld (2007) have proposed additional refinements to the permutation procedure that capitalize on the auto-correlational structure in electrophysiological data, rather than just equate it for each permutation. The basic idea is that psychologically or electrophysiologically meaningful effects should be clustered (for example in time or space), whereas isolated effects are more likely due to noise. A difference in two ERP waveforms, for example, that is confined to a single sample (corresponding to 1 ms, for a sampling rate of 1000 Hz) is highly suspicious, because we would expect meaningful effects to last longer. This cluster-based permutation procedure requires the specification of the expected cluster size which needs to be done *a priori* (i.e., without being informed by the data that is being analyzed; see above). Naturally, if multiple cluster sizes are being considered, additional steps are required to control the FWER or the FDR as explained above.

*Bootstrap Methods**

A resampling method that is closely related to permutation tests is the bootstrap method (Efron & Tibshirani, 1993). Because permutation tests rely on the *exchangeability assumption* (i.e., condition labels can be exchanged under the null hypothesis), they test whether two distributions are identical and cannot test specific moments (such as means) in isolation. Usually this is not a problem—if an experimental manipulation truly has no effect, we would expect associated waveforms to be identical in all respects and not, for example, differ in variance even though the means are identical. However, it can be useful to estimate statistics of the data set, without relying on the *exchangeability assumption* or assumptions about the shape of the sampling distribution.

The bootstrap method consists of repeatedly sampling (with replacement) from each condition and calculating the statistic of interest based on these samples. Let us reconsider our above example with 10 participants for each of whom we have two ERP waveforms corresponding to distinct experimental conditions. If we repeatedly draw (with replacement!) 10 samples from the data set and calculate a *t*-statistic on the difference for each of these draws, we can calculate the bootstrapped *p*-value by comparing the *t*-statistic for the actual data with the distribution of bootstrapped *t*-statistics (the *p*-value corresponds to the proportion of bootstrapped *t*-statistics that

are at least as extreme as the actual *t*-statistic). In each iteration we could also simply compute the differences; the standard deviation across these bootstrapped differences would correspond to the bootstrap estimate of the standard error of this difference.

The bootstrap estimate is asymptotically correct, in the sense that as the original sample size approaches the population size (which may be infinite) the bootstrap sampling distribution approaches the population sampling distribution. To deal with multiple comparisons, the bootstrap method can be also be used to estimate a maximal statistic or multiple bootstrapped *p*-values can be subjected to procedures to control the FWER or the FDR as explained above. Westfall and Young (1993) provide a comprehensive review and formal exposition of resampling methods for testing multiple hypotheses.

4
Time-frequency decomposition methods

Memory-related neural oscillations

Consider the neural and cognitive processes that unfold between the perception of a test item and its classification as "old" or "new" in a recognition memory test. Such a judgment requires the interplay of perceptual, memory, decision, and response processes. Some of these processes align closely to stimulus or response events (e.g., early perceptual processes unfolding directly after the presentation of a target stimulus, or motor processes immediately preceding a response). Such time-locked processes would appear when analyzing the evoked potential, using methods described in Chapter 3. But one may also seek to understand memory-related neural activity between these early perceptual and late response processes. As this activity will not be precisely time-locked to the cue or response event, time averaging in an event-related potential analysis will obscure these signals.

We can identify neural correlates of behavior that do not time-lock to stimulus and response events in either the time or the frequency domain. In the time domain, we can measure the properties of the voltage series, such as its mean and standard deviation. However, brain activity often exhibits rhythmic components or oscillations. Chapter 2 discussed several examples of rhythmic activity related to cognitive processes, including the 3-8 Hz theta rhythm which exhibits striking correlations with both spatial cognition and verbal memory (J. F. Miller et al., 2018; Herweg, Solomon, & Kahana, 2020; Rudoler, Herweg, & Kahana, 2021). To determine whether memory-related variables influence the presence of theta activity, we need some way of extracting rhythmic components from a time series. The time-frequency decomposition methods introduced here have wide-ranging applicability throughout the physical and life sciences. We will draw upon these methods extensively in subsequent chapters.

The Fourier Transform

One of the most important mathematical insights in history came to us from the French mathematician Jean Baptiste Joseph Fourier (1768 - 1830). In 1807 Fourier presented his work *Théorie analytique de la chaleur* (*The Analytical Theory of Heat*) in which, among other things, he claimed that an infinite sum of sinusoids, varying in their frequencies and phase-shifts, could

constitute *any* continuous periodic signal. This claim built upon the work of other mathematicians (e.g., Leonhard Euler), but Fourier is credited for recognizing its broad impact and potential utility. Fourier's radical idea implies that one can decompose any arbitrary continuous period signal into its constituent parts, allowing us to understand complicated and seemingly irregular signals in terms of well-behaved sine waves. In Fourier's case, he used this idea to solve the heat equation, for which there existed at the time no general solution (although solutions did exist for special cases in which the heat source was a regular function such as a sinusoid). Fourier's had the insight of representing complicated heat source functions as the sum of simple sinusoids, which allowed him to define a general solution to the heat equation as the sum of solutions to the simpler functions. As is often the case in academia, Fourier's new idea faced resistance from established figures, including Joseph Louis Lagrange (1736-1813), who argued that it was not possible to use sinusoids to perfectly represent functions with discontinuous slopes such as square and sawtooth waves (while this is true for a finite number of component sinusoids, it is possible to derive arbitrarily close approximations). Luckily for us, Fourier published his work 15 years later, and *Fourier analysis* has become one of the most widely used analytic tools in science and engineering.

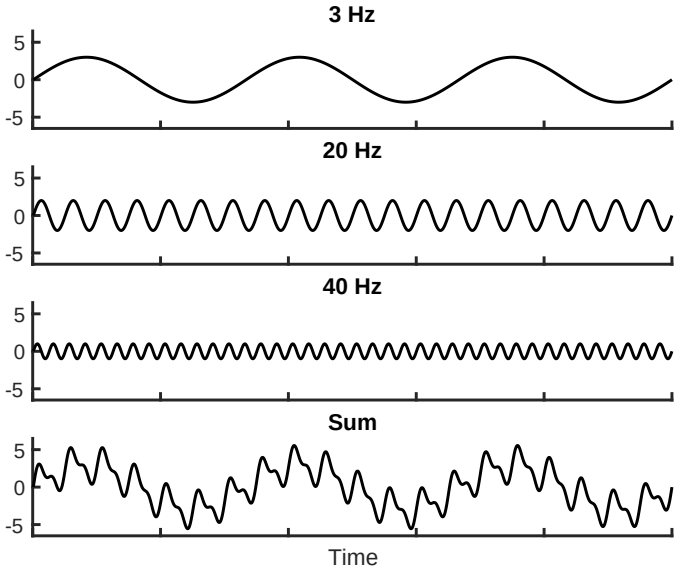


Figure 4.1: Constructing a signal by summing sine waves. Summing 3, 20, and 40 Hz sinusoids, with amplitudes 3, 2, and 1 (arbitrary units) produces a complex waveform. Fourier methods allow us to reverse this summation, discovering the sinusoidal components within a complex signal.

Figure 4.1 illustrates how a complex periodic signal emerges from summing a few sinusoidal oscillations. The large amplitude low frequency (3 Hz) component contributes most to the shape of the function, while the smaller amplitude high frequency (20 and 40 Hz) components yield smaller fluctuations. Here we see how the superposition of sinusoids can lead to functions that vary in complicated ways. The value of the Fourier transform is in its ability to go in the reverse direction: to reveal the underlying structure of signals, such as electrophysiological brain data, whose patterns are less obvious in the underlying time series.

Decomposing real-world signals into representations based on sinusoids

is the foundation of the branch of electrical engineering known as signal processing. By uncovering the sinusoidal components of signals, one can develop tools to enhance the contributions of desired components and attenuate the contributions non-desired components. The process of modifying a signal in such a way is called *filtering* and is one major application of Fourier analysis. For example, audio signals can be filtered to attenuate frequencies that carry noise, with minimal impact on the frequencies that carry speakers' voices. Image processing illustrates another major application of Fourier methods. Visual images can be represented as two-dimensional signals, and decomposing image signals with methods based on the Fourier transform can provide information about what underlying frequencies contribute the most information to the appearance of the full image. This is the basis of image compression methods which seek to minimize the storage requirements for digital images while maintaining the fidelity of the picture being represented. In this and later chapters we will make similar use of Fourier-based methods to perform spectral decomposition of electrophysiological data. We will use this approach to both remove the contributions of undesirable noise, and to derive insights into the mechanisms of memory function in the brain.

Sinusoids

Before introducing the Fourier transform, we review the various mathematical notations for sinusoids. For an oscillation of frequency f in Hz, where $f = 1/T$ for oscillation time period T in seconds, we can define an angular frequency $\omega = 2\pi f$ in radians per second. Two other parameters describe the oscillation at each frequency: amplitude (r), which is one-half the peak-to-trough size of a sinusoidal wave; and phase (ϕ), the wave's offset in radians relative to an arbitrary starting time $t = 0$:

$$r \sin(\omega t + \phi) \tag{4.1}$$

Alternatively, one can write the same wave as a function of *cosine*:

$$r \cos\left(\omega t + \phi + \frac{\pi}{2}\right) \tag{4.2}$$

where the cosine wave is phase-shifted by $\frac{\pi}{2}$ relative to the sine wave. Although $\sin \omega t$ and $\cos \omega t$ are the standard notations for sinusoidal waves, any sinusoid at any phase can be written as the sum of a sine and cosine of the same frequency:

$$A \sin(\omega t + \phi) = B \sin \omega t + C \cos \omega t \tag{4.3}$$

where $A = \sqrt{B^2 + C^2}$ and $\phi = \text{atan2}(B, C)$ (see, Figure 4.2 for a graphical illustration). The function atan2 is the version of the arctangent function that correctly handles negative values of a, providing the full range of -180° to 180°, or $-\pi$ to π, as possible phase values.

Sinusoids are often depicted as waves oscillating over time but they can also be thought of as representing the motion of a point around the circumference of a circle. If we consider a circle of radius r centered on the origin, we can draw a vector $\mathbf{v}$ from the origin to the edge of the circle. If we imagine rotating $\mathbf{v}$ at frequency (speed) ω counterclockwise around the circle, the projection of $\mathbf{v}$ onto the x and y-axes over time traces out the $A \cos \omega t$ and $A \sin \omega t$ functions, respectively (where $A = r$).

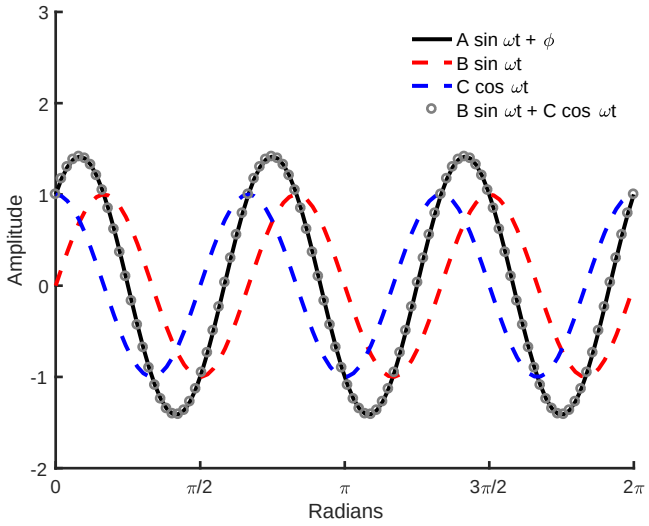


Figure 4.2: Sinusoid represented as the sum of sin **and** cos **waves.** A graphical representation of Equation 4.3 shows that the sum of $B \sin \omega t$ (dashed red line) and $C \cos \omega t$ (dashed blue line), shown in gray circles is equal to the sinusoid $A \sin(\omega t + \phi)$ (solid black line).

Complex Numbers

The preceding section presented two key ideas: (1) any sinusoid of frequency ω can be written as the sum of a sine and a cosine at the same frequency ω; (2) sinusoids can represent rotation around a circle. Both of these ideas are helpful for understanding why *complex sinusoids* are so important for the Fourier transform and other spectral decomposition methods. Complex sinusoids simplify much of the mathematics required for Fourier analysis; so before introducing the Fourier transform, we first briefly review complex notation.

A complex number is the sum of two parts, taking the form $a + bi$ where a is the *real* part, b is the *imaginary* part, and i is the imaginary number $\sqrt{-1}$. [1] The operators $\Re[]$ and $\Im[]$ are used when separating a complex number into its real and imaginary parts. Thus, the real part of the complex number $3 + 4i$ would be extracted by writing $Re[3 + 4i] = 3$, while the imaginary part would be extracted by writing $Im[3 + 4i] = 4$. In equations, complex numbers are usually represented by a single variable, i.e. $z = 3 + 4i$, which highlights an important property of complex numbers, namely that they are a convenient way to represent two quantities within a single variable.

Complex numbers can be represented in terms of their real and imaginary parts, a and b, by using a two-dimensional plane in which the x-axis is the *real* axis and the y-axis is the *imaginary* axis. Figure 4.3 shows the number $z = 3 + 4i$ graphed in this way. The red dot shows the position of z as a point in the 2D plane and shows how it is composed of real part $a = 3$ and imaginary part $b = 4$. We can also see from Figure 4.3 that the same complex number (i.e. the same point in the complex plane) can be represented by two alternative parameters, r and ϕ. r is the *magnitude* of the complex number and ϕ is the angle subtended by r and the real axis. r and ϕ can be found for

[1] The traditional notation from mathematics is to use i to denote $\sqrt{-1}$, while in electrical engineering j is used to distinguish $\sqrt{-1}$ from electric *current*, which is represented by i. Here we will use the convention from mathematics.

any complex number $z = a + bi$ using trigonometry:

$$\begin{aligned} r &= \sqrt{Re[z]^2 + Im[z]^2} = \sqrt{a^2 + b^2} \\ \phi &= \text{atan2}\left(b, a\right). \end{aligned} \tag{4.4}$$

Expressing a complex number with r and ϕ is called using *polar notation*. We can also define the relation between polar notation and the notation of the complex plane, $a + bi$, also referred to as *rectangular notation*:

$$\begin{aligned} Re[A] &= r\cos(\phi) \\ Im[A] &= r\sin(\phi) \end{aligned} \tag{4.5}$$

By comparing Equations 4.4 and 4.5 to Equation 4.3, you may begin to see how a sinusoid could naturally be represented as a location in the complex plane, but you may also wonder why it would ever be useful to take a single sin or cos function and split it into the sum of both a sin and cos. As we will see in a moment, there is a mathematical relationship between complex numbers represented as sinusoids and complex *exponential* functions that is foundational to the analysis of sinusoidal functions in science and engineering.

By performing some substitution, we see that

$$a + bi = r\cos(\phi) + ir\sin(\phi) = r(\cos(\phi) + i\sin(\phi)), \tag{4.6}$$

which gives us the relation between a complex number's rectangular and polar notations. We can now introduce an important definition in complex analysis, Euler's Relation:

$$e^{i\phi} = \cos(\phi) + i\sin(\phi) \tag{4.7}$$

One of the reasons that Euler's Relation is such a critical tool in science and engineering is that, by allowing us to recast sinusoids as exponentials, Euler's Relation simplifies many of the mathematical operations needed to analyze sinusoidal functions. [2] For example, multiplying two sinusoids requires the application of several trigonometric identities, whereas doing so with exponential functions requires only adding the angles of the respective sinusoids and multiplying their magnitudes:

$$r_1 e^{i\phi_1} r_2 e^{i\phi_2} = r_1 r_2 e^{i(\phi_1 + \phi_2)} \tag{4.8}$$

As we will see in the next section, the complex exponential, $e^{i\phi}$ plays an important role in methods for spectral decomposition such as the complex Fourier transform.

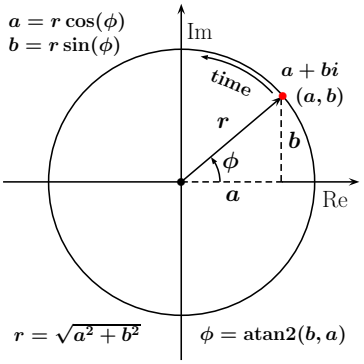


Figure 4.3: The complex plane. The complex number $z = a + bi$ is shown on the complex plane. The x- and y-axes are used to plot the *real* and *imaginary* parts of a complex number as a single point on 2D space (red dot). The magnitude r of the vector from the origin to $(a,\ b)$ is equal to $\sqrt{a^2 + b^2}$. The angle subtended by z and the real axis is ϕ, which will become important later in our discussion of how to extract phase information from oscillatory EEG activity. As fixed amplitude oscillation progresses in time, the complex values move counterclockwise around the complex plane in the circle shown, with the phase progressing periodically from 0 to 2π and around again.

[2] For the interested reader, Equation 4.7 can be shown to be true using the Taylor series expansion of the exponential function.

The Fourier Transform

The classic continuous Fourier Transform (FT) is defined in integral form for a time-domain function $f(t)$ as:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i2\pi\omega t}dt, \tag{4.9}$$

producing a frequency-domain function F over angular frequency ω. This continuous Fourier transform has high utility for dealing with analytical expressions but does not directly deal with the type of time-series array data

found in electrophysiology analysis. We must switch to a discrete representation that can operate on discrete array elements.

The analogous Discrete Fourier Transform (DFT) of the time-series signal contained in an array x is given by

$$X[\omega] = \sum_{n=0}^{N-1} x[n] e^{-i2\pi\omega n/N} , \tag{4.10}$$

where N is the number of data points in the array, ω is the frequency being evaluated in units of the inverse time length of the array, and $X[\omega]$ is the resulting Fourier coefficient of the time-series $x[n]$ at frequency ω. The output $X[\omega]$ is the frequency-domain representation of the signal $x[n]$ at frequency ω. We can think of $X[\omega]$ as the magnitude of a sinusoid of frequency ω that is present in $x[n]$. [3]

[3] The definition of the Fourier transform given by Equation 4.10 is called the complex *discrete Fourier transform* (DFT). This is the Fourier transform used in spectral analysis of discrete signals, i.e. signals such as EEG that are recorded and sampled using digital electronics.

The frequency-domain values $X[\omega]$ provide a complete representation of the original time-series data $x[n]$, which can be demonstrated by the fact that the inverse DFT can recreate the original data by the following operation:

$$x[n] = \frac{1}{N} \sum_{n=0}^{N-1} X[\omega] e^{i2\pi\omega n/N} , \tag{4.11}$$

This clarifies that the frequency-domain representation does not summarize the data, but rather helps us to identify oscillatory components present in the original time series.

Equation 4.10 combined with equation 4.7 express the basic idea of the Fourier transform, which is that the frequency-domain representation $X[\omega]$ is obtained from the time-domain signal $x[n]$ by a point-wise multiplication with a pair of sinusoids of frequency ω defined over the length of $x[n]$ (N samples). The sum of this point-wise multiplication is then taken over all N samples. Readers with some background in linear algebra will be familiar with another name for the point-wise multiplication/summation operation: the *dot product*. Just as the dot product provides a measure of the "similarity" of two vectors in space, the Fourier transform measures the similarity between the time-series signal and each sinusoid of frequency ω. Using the relation illustrated in Figure 4.3, this similarity can be expressed as either the complex values $a + bi$ (a from the cosine term, b from the sine term) that come out of the DFT, or equivalently as a phase (ϕ) and amplitude (r).

An algorithm known as the Fast Fourier Transform (FFT) performs the DFT with far greater computationally efficiency than the direct computation shown in equation 4.10. The FFT algorithm can perform a full DFT in a runtime that scales with the length of data times the log base 2 of the data length, and consequently remains efficient with larger datasets. Many software libraries provide optimized code for applying the FFT to large data arrays.

The top panel of Figure 4.4 shows a signal created by adding together 3, 20, and 40 Hz sinusoids. The result of the Fourier transform of this signal is plotted in the bottom panel. The frequency of the recovered sinusoid is shown on the x-axis, while its amplitude is on the y-axis. Since the summed signal was constructed exactly from sinusoids at 3, 20 and 40 Hz (and nothing else), there are peaks at each of those frequencies, while all other frequencies are zero. The height of the peak for each frequency corresponds to its amplitude in the original signal. You will notice that the x-axis contains both positive and negative frequencies—this arises from the symmetries

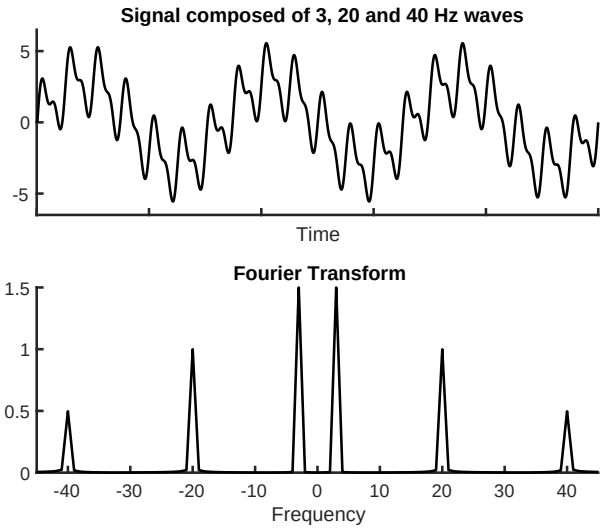


Figure 4.4: The output of the Fourier transform. The top panel shows a signal created by adding together sinusoids of frequencies 3, 20, and 40 Hz (see Figure 4.1). The bottom panel shows the Fourier transform of this signal and illustrates how the transform recovers the frequencies and amplitudes of the sinusoids from which the signal is composed (x-axis restricted to ± 40 for the purposes of visualization). However, this representation does not show phase information, which we discuss in more detail below.

of the sinusoidal function as a nuanced consequence of beat frequencies, similar to the pattern illustrated in Figure 4.5d, and means that a sinusoid of frequency ω that contributes to a signal will have its energy (amplitude) split between the ω and $-\omega$ sides of the x-axis when its Fourier transform is graphed. This representation of putting the 0 in the center is referred to as a Fourier transform "shift". In the data coming straight out of a DFT calculation, the 0 frequency is the first element, followed by the positive frequencies in increasing order. The corresponding negative frequencies begin at the last element with each more negative one toward the middle. In practice, the negative side of the spectrum is often omitted from visualization and the positive side is plotted with the amplitudes doubled to account for the contribution of the negative frequencies. In Figure 4.4 we graphed the shifted Fourier transform of our signal for f in the range $[-45,\ 45]$, but in fact the DFT of a signal $x[n]$ of length N produces N equally-spaced frequencies n/T for n from 0 through $N-1$, where T is the full time period covered by the input data. As described above, this is equivalent in the shifted expression to N equally-spaced frequencies with 0 in the center and a step size of $1/T$ going positively and negatively. The 0th frequency corresponds to the sum of all N values in x, and has a 0 value only in the special case of zero-centered inputs as shown here. The special frequency $1/T$ is referred to as the first harmonic, and corresponds to the frequency of sinusoids with a period of all N input elements. $2/T$ is referred to as the second harmonic, and so on with third and fourth.

Consequently, the DFT approach yields a fixed set of frequencies determined by the time period covered by the input length. If a specific set of target frequencies is desired, the resulting frequencies from a DFT can be adjusted by increasing or decreasing the amount of input data used for each DFT calculation. If additional data is not available and decreases cannot be made, small frequency adjustments can in principle be made by adding zeros or mirroring data at the edges, but this risks edge ringing effects and compromising some of the data integrity, and should be approached with caution.

The Sampling Theorem

Imagine that you collect EEG data from a participant in a memory experiment and you are interested in using the Fourier transform to investigate the amplitude of a 250 Hz oscillation. Will you be able to use the Fourier transform to decompose the EEG time-series to extract the amplitude of this sinusoid? While one can adjust the frequency resolution of the FT result by adjusting the length as described above, doing so does not alter the highest frequencies obtainable. This maximum frequency observable is instead determined by the temporal step size of the input data.

Like many signals that one encounters in nature, the voltage fluctuations that are measured by EEG electrodes can in principle vary continuously over time and take on an arbitrary number of values at intermediate time points. Unlike the natural world, digital electronic devices like computers cannot represent a continuous spectrum of varying values and must instead store information in finite and discrete form. Therefore, any continuous signal that is recorded and stored in a digital device is a discrete and finite representation of the original continuous signal. The frequency at which values of the continuous signal are digitally recorded is called the sampling rate of the signal and is typically expressed in units of Hertz (Hz; i.e., number of samples per second).

The sampling rate is the key parameter that determines how closely the digital signal resembles the original continuous signal that it represents. How high does the sampling rate need to be in order to produce a digital signal that is "close enough" to the original continuous signal? The *sampling theorem* provides an answer to this question. [4] The sampling theorem states that a continuous signal can be perfectly reconstructed from its digital representation if it contains no information at frequencies greater than one-half the sampling rate of the signal. The minimum sampling frequency that allows perfect reconstruction of a signal is called the *Nyquist* rate.

[4] The sampling theorem is sometimes referred to as the Shannon sampling theorem or the Nyquist sampling theorem after the authors who are credited with its discovery.

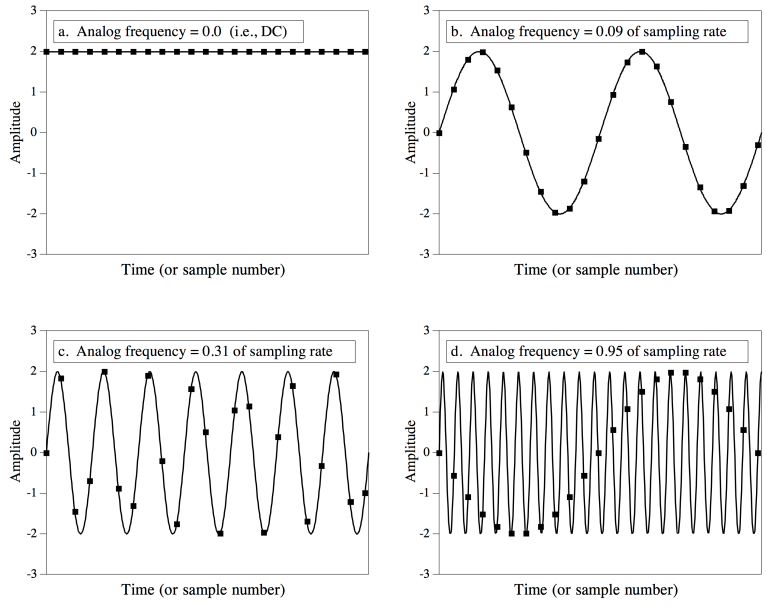


Figure 4.5: Illustration of the Sampling Theorem. In each panel the smooth curve represents a continuous signal and the squares represent the points at which the signal is digitally sampled. (a) A sinusoid of frequency 0 (straight line) can be perfectly reconstructed from the set of samples. (b) This sinusoid has a frequency that is 0.9 the sampling rate, as would be the case for a 90 Hz oscillation sampled at 1000 Hz. The spacing of the samples produces a good representation of the original sinusoid. (c) For a sinusoid with a frequency equal to 0.31 the sampling rate, one can reconstruct the original signal from the samples, although this is less obviously true than in (b). (d) When a sinusoid at frequency 0.95 times the sampling rate is sampled, it masquerades as a sinusoid of lower frequency, a phenomenon known as *aliasing*. Figure from Steven W. Smith *Digital Signal Processing*, 2003.

The sampling theorem means that if you are interested in using the Fourier transform to measure 250 Hz oscillations in your EEG data, then, in principle, you must sample the data at no lower than 500 Hz. However, measuring oscillatory activity at 250 Hz with a sampling rate of 500 Hz means that you will be relying on just two samples per cycle to provide information about 250 Hz activity, which is ill-advised in practice because EEG recordings (and all real signals) are noisy, and certain phase alignments between the signal being sampled and the set of sample times will suppress the signal. It is therefore better to use a sampling rate somewhat higher than the Nyquist rate to increase the number of observations per cycle used in estimating oscillatory activity. Since the frequency domain signal support of an instrument is approximately given by the target frequency's normalized amplitude in the FT of the sampling windows, we can estimate how good the higher sampling rates are. For some practical examples of this, a sampling rate of 3 times the target oscillation frequency will capture about 83% of the information available for that frequency in the underlying signal, while a sampling rate of 4 times the oscillation will capture about 90% of the information available.

Figure 4.5 illustrates the consequences of using the same sampling frequency to sample four different continuous signals that vary in their frequency content. In (a–c), the sampling rate is greater than twice the frequency of the signal, which means the original signal can be reconstructed from the digitized data. Contrast this with (d), where the frequency of the signal is nearly as high as the sampling rate. In this case the higher frequency waveform takes the appearance of a lower frequency sinusoid in the sampled time-series, a phenomenon called *aliasing* or referred to as a beat frequency.

The discussion above has explained the principles of a Fourier transform and how to extract and interpret the frequency content of a signal using this approach. As explained, the Fourier transform result is a fully reversible expression of the entire time-series signal, and therefore contains all of the frequency information. However, the manner in which this frequency information is presented might not match the desired model of an analyst. For example, if a 7 Hz oscillation starts and concludes in the middle of the time series, a Fourier transform will represent the starting and concluding of this oscillation by constructing a wave packet of amplitudes distributed across the frequencies around 7 Hz with the pattern of phases of these frequency components determining the position of the oscillation in time. Therefore, while this does in fact contain the transient nature of the oscillation (and will allow for reconstruction of the original signal), it is not represented in a useful form for analysis of transient oscillatory signals. Experiments designed to study cognitive processes such as memory encoding and retrieval are carried out with the explicit goal of comparing how the frequency content of the EEG signal varies in response to experimental events. Because of these limits on the situations in which a direct Fourier transform provides a good representation of the desired oscillatory properties, other methods have been developed that provide time-varying estimates of the frequency content of a signal. These methods form the basis for the next sections.

Short-time Fourier Transform

The logic of the short-time Fourier transform (STFT) is that instead of applying the Fourier transform to the entire time-series, the full length of the series is broken into smaller time windows, and the Fourier transform is applied to each window separately. Doing so leads to an estimate of the frequency content of a signal for each time window. The resulting *time-frequency* representation will be sensitive to changes in frequency content that happen over time, for example in response to experimental events.

Two important parameters to consider when using STFT to analyze data are (1) the size of the window and (2) the degree of overlap between consecutive windows. The window size determines the tradeoff between resolving time and frequency information with the analysis. Selecting a window that is brief will lead to estimates that are well localized in the time domain, meaning they will be derived from a relatively short interval, but this will lead to estimates that are poorly localized in the frequency domain. This is because the frequency resolution of the Fourier transform of noisy data is determined by the length of the analyzed signal (see previous section) so a window of shorter length (with fewer samples) will produce estimates at fewer frequencies. A more problematic issue is that the frequency estimates will be of poor quality because the window length does not provide enough cycles of lower frequency oscillations to obtain good levels of signal-to-noise. By taking longer windows, the signal-to-noise ratio of the frequency estimates will be improved, but at the expense of the ability to localize frequency fluctuations in time. Thus, the window size for the STFT should be selected to allow sufficient estimation of the lowest frequency of interest—one rule of thumb is to use a window long enough to include at least three cycles of the lowest frequency of interest. For example, an analysis where the lowest frequency is 6 Hz should be done with a window size of at least 500 ms. However, selecting a window length that is sufficient to obtain good estimates of the frequency content at the lowest frequency of interest leads to another problem which is that a window long enough to include a few cycles of the lowest frequency may be too long to resolve changes in activity at higher frequencies. In the 1 s that are needed to obtain reliable estimates of the 3 Hz oscillation, there could be prominent fluctuations in activity at, say, 100 Hz (see Chapter 3 for discussion of how such high-frequency activity relates to memory). These fluctuations will be reduced to a single estimate of 100 Hz activity for the entire 1 s window. This trade-off between resolution in the time domain (knowing when things happened) and resolution in the frequency domain (knowing what frequencies are present) applies to all of the spectral methods that we have discussed thus far, as well as those that we will encounter later.

Windowing Functions

Since the short timescale FTs have fewer repetitions of target oscillation frequencies, the effects of edges on the resulting values are enhanced. The fundamental assumption of the finite DFT is that a signal is composed of periodic signals of the resulting frequencies. For most natural data attempting to isolate frequencies which are only approximately matching the target frequency, this assumption is in fact wrong. Whenever the frequencies of the

underlying oscillations in the signal are slightly less than or more than the frequencies resulting from the transform, which is almost always true for signals not forcibly synchronized with the analysis window, this results in a periodic discontinuity between the values at the starting and ending edges of the signal, which appear to the DFT like a sharp transition in the signal. The FT of a sharp transition has components in all frequencies, and therefore any periodic discontinuities at the edges will produce anomalous amplitude and phase impacts, often referred to as ringing artifacts, in all frequencies of the transform. The magnitude of this effect is stronger at short timescales when there are few proper repetitions contributing to the signal compared to the amplitude of the periodic-boundary transition at the edges.

Since these edge effects disrupting the spectral information in the phase and amplitude values are often unacceptable for an analysis, a set of solutions have been devised involving using windowing functions. One approach to this is to attenuate the amplitude of the time-series data near the beginning and end of each time window being analyzed. An illustration of this is provided in Figure 4.6, where the *Hann* window defined by

$$w[n] = \frac{1}{N} \sin^2 \left(\frac{\pi n}{N-1} \right) \tag{4.12}$$

is applied by pointwise multiplication with the time-series.

Usage of the Hann window in this manner guarantees that the periodic boundary issues described above are addressed, as the edges of the signal being analyzed are attenuated to zero. Furthermore, a detailed analysis of the Hann window's structure reveals that it does a good job of preserving the oscillatory dynamics present in the non-attenuated portion. Many other competing windowing functions are also available with similar properties, each with subtle differences in the results obtained. However, to observe gradual temporal dynamics in the oscillatory behavior, this particular approach requires generating a very large number of copies of overlapping segments of the time-series data, attenuating each one separately, and then processing each with a DFT. This motivates the consideration of an alternative approach, where instead of modifying the time-series, the sinusoids themselves are attenuated.

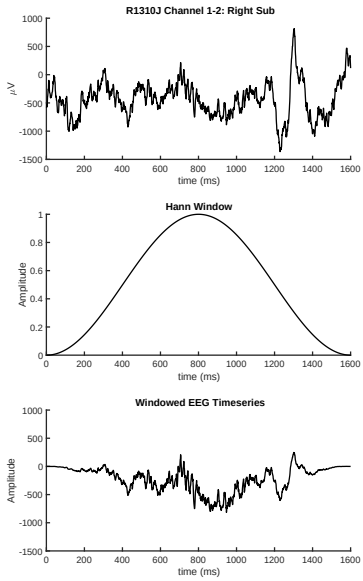


Figure 4.6: Windowing a time-series. *Top*, an example EEG time-series collected from the right subiculum of an intracranial EEG subject performing the FR1 task. *Middle*, the Hann windowing function. *Bottom*, windowing the EEG signal with the Hann window tapers the time-series.

Convolution and Wavelet Transforms

One of the most often used methods for spectral decomposition in electrophysiology data is the Morlet wavelet transform. This technique addresses many of the limitations discussed in the previous sections. Before covering the Morlet wavelet, we first introduce *convolution*: a mathematical operation that computes the similarity over time between two functions when one is reversed, by summing the dot product for various relative offsets in time.

Convolution and filtering basics

An intuitive way to think about discrete convolution on array data is to imagine starting with two time-series, inverting one of them from left to right and sliding it along the other time-series, incrementing one sample at a time. For each increment, convolution quantifies the dot product between those two signals—convolution is therefore a measure of how the similarity between

two functions changes over time. Convolution is ubiquitous in the analysis of linear systems where the goal is to specify a system's predicted output given an arbitrary input function and an "impulse response" characterizing the system's measurement. Thus, another way to think about convolution is that it measures the effect of applying one function (the system's temporal response to each data value) to another (the input signal).

Discrete convolution of two time-series x and h is defined as:

$$x[n] * h[n] = \sum_{m=0}^{N-1} x[m]h[n-m] \tag{4.13}$$

The array $x[n]$ is often called the signal while $h[n]$ is called the kernel, serving the role of the impulse response described above. Because the $x[m]h[n-m]$ part computes the dot product between two signals for different time shifts m, which are then summed, the output of convolution is another time-series that measures the similarity between $x[n]$ and $h[n]$ across time. The output of $x[n] * h[n]$ will therefore be largest at the timepoints where $x[n]$ and the reverse of $h[n]$ are most similar. Note that for representational simplicity Equation 4.13 did not address the array boundaries, as the standard continuous definition of convolution is defined on functions of infinite range. Next we will show efficient calculation of this for discrete data with Fourier methods, which are periodic, which means for actual calculations the definition Equation 4.13 should be regarded as having periodic boundaries where e.g. $x[-1]$ access $x[N-1]$ and $x[N+2]$ access $x[2]$.

Convolution is an operation defined on two signals in the time domain but it can also be expressed in terms of the frequency domain representations of the two signals:

$$x[n] * h[n] = \mathcal{F}^{-1}\{\mathcal{F}\{x[n]\} \cdot \mathcal{F}\{h[n]\}\} \tag{4.14}$$

where $\mathcal{F}\{x\}$ denotes the Fourier transform of x and $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. Equation 4.14 is called the *Convolution Theorem* and states that convolving two signals in the time domain is equivalent to point-wise multiplication of the same two signals in the frequency domain and then applying the inverse Fourier transform. Practically, this property of convolution is useful because convolution in time is a very slow operation to carry out—the point-wise multiplication and summation operation has to be repeated for each timestep of overlap between x and h. Contrast this with the frequency domain version, in which point-wise multiplication is carried out only once. Because of the existence of algorithms like the fast Fourier transform (FFT), which can very efficiently compute the Fourier transform and its inverse, performing three Fourier transform operations in conjunction with a single multiplication operation can be orders of magnitude faster than computing the dot product repeatedly over time, especially for very long signals.

Morlet wavelet transform

The Morlet wavelet transform is a widely used spectral decomposition method that overcomes some of the limitations of strictly Fourier-based methods. Two key benefits of using wavelets are (1) the ability to extract time-resolved estimates of the frequency content of a signal; and (2) the use

of kernels that scale as a function of frequency of interest. In this section, we describe the mathematics that underlie Morlet wavelets and describe how to use them to extract power and phase information from a signal.

A simple way to describe a wavelet is that it is a sinusoid windowed by a Gaussian:

$$\begin{aligned} h[n,k] &= Ae^{-\frac{(n-\mu)^2}{2\sigma^2}} e^{\frac{-i2\pi kn}{N}} \\ A &= \frac{1}{(\sigma\sqrt{\pi})^{1/2}} , \end{aligned} \tag{4.15}$$

where N is the length of the signal, ω is the frequency of the wavelet, and μ, σ are the mean and standard deviation of the Gaussian. Here, σ is defined as follows:

$$\sigma = \frac{c_n}{2\pi k} , \tag{4.16}$$

where c_n is the wave number or cycle count for the wavelet and affects the number of cycles that are present in the wavelet before it tapers to near zero. Figure 4.7 gives an example of windowing a sinusoid with a Gaussian to produce a wavelet. Changing the wave number c_n changes the tradeoff between frequency and time resolution in the output of the wavelet transform. Increasing c_n will create a wavelet with more cycles, which means that the convolution between the wavelet and the time-series at any particular time-point will be influenced by data from a longer time interval. Including data over a longer period of time in this way will increase the resolution of the convolved signal in the frequency domain but will do so at the expense of resolution in the time domain. The structure of windowing a sinusoid by a Gaussian for a Morlet Wavelet is obtained as an optimal wavelet solution when one attempts to minimize the product of the standard deviations of time and frequency, so this has become a widely used wavelet choice.

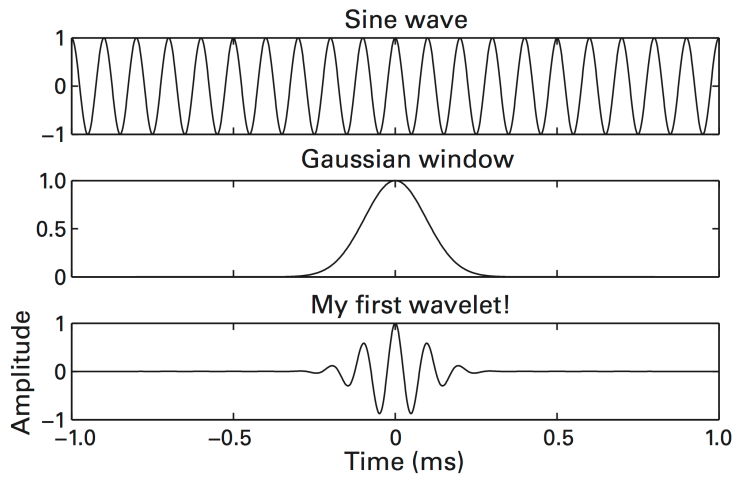


Figure 4.7: Constructing a wavelet. *Top*, a sinusoid. *Middle*, a Gaussian window. *Bottom*, pointwise multiplying a Gaussian window with a sinusoid to produce a wavelet. Figure reproduced from Michael X. Cohen's *Analyzing Neural time-series Data*, 2014.

Wavelet decomposition results from convolving the wavelet with the EEG time-series. Because convolution performed with an FFT assumes a periodic signal, the output can contain edge artifacts (as with the STFT case described above). However, wavelet analysis limits the impact of edge artifacts to the

resulting time values in which significantly non-zero elements of the wavelet are shifted toward the edges. This provides a clear guide to the buffer size required at the beginning and end of the signal of interest to remove edge artifacts, as shown in Figure 4.8. With the definitions used in Equations 4.15 and 4.16 the produced wavelets for frequency f drop to less than 1% at $\frac{c_n}{2f}$ from the midpoint, giving them a region of interest that stays within a temporal width of c_n/f. Buffers of length c_n/f for the lowest frequency f (the one which extends the furthest in time) correspond to the Gaussian envelope dropping to about 1 over 500 million, which is a good estimate for the scale of reduction in the edge effects compared to not using buffers. Removing buffers after the wavelet decomposition results in 'clean' time-frequency data within the target interval.

While buffers constructed out of real adjacent data are the most authentic choice, there are several cases where this is not ideal. One is during time-sensitive real-time calculation where subsequent data is not yet available. Another is when adjacent data is expected to contain abrupt physiological changes, such as from the onset of a stimulus, that one wants to explicitly exclude from the analyzed result. To address these cases it is possible to use *mirror buffers*, which consist of taking the real data right before the end excluding the last element, flipping it around in time, and appending it to the end. The corresponding operation can also be done at the beginning, taking real data right after the beginning, excluding the first element, flipping it around in time, and inserting it at the beginning. This mirror buffer operation could enhance transient power fluctuations up or down at the ends and can introduce a slight discontinuity in the slope, but broadly speaking it does an adequate job of preserving the general frequency dynamics of the underlying signal without introducing major systematic artifacts.

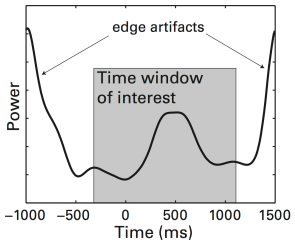


Figure 4.8: Edge Artifacts and Buffers. Convolution of an EEG time-series and a wavelet leads to *edge artifacts* which are a consequence of zero-padding in the convolution operation and do not reflect true EEG signal. To deal with this issue when analyzing data, additional segments of time called *buffers* are added to the window of interest before the convolution and then discarded before proceeding with additional analysis.

An alternative formulation for specifying the widths of each wavelet is given by Cohen (2019) with wavelet h_w as:

$$h_w[n,\omega] = Ae^{-\frac{4\ln(2)(n-\mu)^2}{w^2}} e^{\frac{-i2\pi\omega n}{N}}, \tag{4.17}$$

where the new parameter w takes the roll of σ, but with w being the full-width half-max (FWHM) of the wavelet. The time width w is kept identical for the wavelets at all frequencies. This results in different numbers of oscillatory repetitions at different frequencies, but also results in the wavelets at each frequency covering the same time range, which provides some intuitive benefits for analysis. This leads to a relationship connecting the FWHM w to the wave number c_n for each frequency of:

$$w = \frac{c_n\sqrt{2\ln(2)}}{\pi f}. \tag{4.18}$$

Complete Wavelets

The formulation in Equation 4.15 is referred to as a *standard* Morlet Wavelet, but it has some characteristics which make it hazardous for electrophysiology analyses when the wave number c_n is low. If an analysis is performed with a standard Morlet Wavelet with low c_n on a flat non-oscillating signal offset from zero, the sinusoidal terms will be zero, but the cosine terms will be non-zero! Carried further, this results in the non-oscillating signal yielding a non-zero amplitude measure, and a phase value which will either

be 0 or 180 degrees, depending on the original signal being above or below zero. For analyses which attempt to extract small consistent signals amidst a background, this can result in significant false observations that look like real results. Because low frequency oscillations appear like an offset in higher frequencies, this causes low frequency amplitudes to bleed into higher frequencies. Phase locking and phase consistency analyses are particularly subject to error from this because the biasing of phase values toward 0 and 180 degrees creates the artificial appearance of phase locking even if none is present.

While this effect is formally non-zero for all values of c_n in standard wavelets, it becomes very small for c_n of 5 or more. However, there is an alternative of using *complete* Morlet Wavelets which eliminates these artifacts, and will work reliably for both large and small c_n. The origin of the effect in standard wavelets occurs in the cosine term because the symmetric cosine is reduced in amplitude by the Gaussian envelope (see Figure 4.7) in an amplitude-biased manner, reducing negative-going parts of the wavelet more than positive-going parts of the wavelet. This gives the cosine wavelet a non-zero integral, or sum of the values in an array containing the wavelet, which means that taking the dot-product of it with a systematically offset signal will be non-zero.

The complete Morlet Wavelet resolves this by satisfying what is called the *admissibility criteria* and applying an offset to the cosine oscillations used to construct the wavelet, restoring the integral to zero:

$$\begin{aligned} h[n,\omega] &= A_c e^{-\frac{(n-\mu)^2}{2\sigma^2}} \left(e^{\frac{-i2\pi\omega n}{N}} - e^{-\frac{1}{2}\sigma^2} \right) \\ A_c &= \left(1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\sigma^2} \right)^{-1/2} . \end{aligned} \tag{4.19}$$

Here σ retains the same form as given by Equation 4.16. The change to the wavelets from this correction factor is slight and nuanced, and they appear visually similar. With large c_n they appear almost identical, and with low c_n the sine term is identical, while the center of the cosine term is very slightly shifted downward, while both still go to zero the same way far from the center. The other principles explained remain the same with complete wavelets, except that the phase and amplitude values can be relied upon for statistical tests with c_n values of less than 5.

Some software packages providing Morlet Wavelet functionality will only offer the standard wavelets, because these issues are less often important for applications such as images and audio. The prominent software packages for electrophysiology typically do offer complete wavelets, but sometimes this requires explicitly enabling them, so analysts should be attentive to this need.

Amplitude, Power, and Phase Information

Convolving a complex Morlet wavelet with an EEG time-series results in a time-series of complex numbers. How does one use this time-series to extract information about sinusoidal activity? Remember that these complex numbers, $a + bi$ are sinusoids represented in the complex plane by magnitude r and phase angle ϕ. Thus, extracting the amplitude of the sinusoid at each timepoint is simply a matter of computing $r = \sqrt{a^2 + b^2}$ and $\phi = \text{atan2}\,(b, a)$.

Although to this point we have graphed the output of spectral decom-

positions, such as the Fourier transform, in terms of amplitude, it is in fact customary to graph the *power* of a signal:

$$\text{Power} = \text{Amplitude}^2 \tag{4.20}$$

Natural signals tend to have far greater power at low than at high frequencies (see next section). Thus we usually work with log-transformed power, and often normalize the resulting values (e.g., by transforming power into *decibel* units or by z-transforming log-power values).

Applications of time-frequency analyses.

We saw several applications of time-frequency analysis in Chapter 2. Figure 2.4G illustrated the power spectrum of the single raw voltage trace shown in Figure 2.4F, with a large peak evident in the 4-8Hz theta-frequency band. Figure 2.8 illustrated a power spectrum for recordings from a single electrode taken while subjects navigated easy vs. hard mazes in a virtual environment. Here, too, we see a large peak in the theta frequency range. Both of these analyses applied wavelet transformations to the time series of recorded voltages.

In addition to revealing narrow-band oscillations in the field potential, spectral decomposition methods can also provide information on non-oscillatory (broadband) aspects of the time series. Manning et al's (2009)'s analysis of intracranial recordings obtained while patients played a spatial learning game called *Yellow Cab* illustrates how fluctuations in broadband power (power averaged across a very wide range of frequencies) predicted the firing rates of individual neurons (Figures 2.11 and 2.12). These broadband EEG fluctuations more consistently and reliably correlated with neuronal firing than did narrow-band oscillatory activity at any of the examined frequencies.

Measures of phase consistency

Broadly speaking, consistent changes in amplitude can reveal that activity is occurring for an event in a particular region of the brain, but a consistency of phase or of phase relationships can reveal nuances of how signals are propagating. Phase is a form of temporal information revealing for each time point where in the oscillation period a particular frequency is. For a time t, a phase of 0 for frequency f corresponds to the extracted oscillation being at its peak value at time t. Similarly, a phase of 90 degrees corresponds to that oscillation passing through its middle point in a downward trajectory. If an oscillating signal at one point A is driving an oscillating signal at another point B, then B will be delayed slightly in reaching the same peak value. This propagation delay will result in B having a negative-shifted phase with respect to A at the same time point.

This sort of relationship of the consistency of the phase difference between electrode contacts is what is observed in phase-locking measurements. These measurements indicate connectivity, or the propagation of oscillatory signals between regions of the brain, which will be discussed more in Chapter 8.

Aside from a consistent phase relationship between two measured signals, it is also possible to observe a consistent phase relationship with respect to

an external event, such as with theta reset. In a properly designed experiment with event timing jitter, this sort of consistent phase with respect to an event time becomes a clear indication of a causal response that has propagated to the electrode contact as a reaction to the external stimulus.

von Mises distribution

If the phase of a sample of data has no correlation with the time point at which it was examined, then it will sample from the *circular uniform distribution*, meaning the circular histogram of the values will approach a circle with large amounts of data, and have noise deviations away from the circle for smaller amounts of data. In contrast, if there is some sort of concentration of phase values due to an underlying consistency, the circular analog to the normal distribution which becomes a good starting model is called the *von Mises distribution*. This distribution is approximately a wrapped normal distribution, and serves as a good first-order approximation for distributions of consistent phase with an uncertainty.

Circular Statistics

When starting with a sample of phase values, it is instructive to consider them as existing in a circular histogram, or polar plot. This circular histogram is equivalent to the complex plane representation shown in Figure 4.3, which means each phase value has a corresponding real and imaginary component, denoted for the unit circle as $a = \cos\phi$ and $b = \sin\phi$. A standard method of assessing a magnitude of consistency of the signal is the *mean resultant length* or *r-bar*, which consists of summing those real and imaginary components, and dividing the resultant vector magnitude by N, the number of phases, as expressed by (Fisher, 1993):

$$\bar{r} = \frac{1}{N}\sqrt{\left(\sum_{i=0}^{n-1}\cos\phi_i\right)^2 + \left(\sum_{0}^{n-1}\sin\phi_i\right)^2}. \tag{4.21}$$

The $\bar{r}$ values range from 0 to 1, where 0 is a completely uniform sample of phase values, and 1 is all samples having the same phase value

The direction of the vector for $\bar{r}$ provides the central tendency of the phase, and is obtainable as with the other complex plane vectors, by:

$$\overline{\phi} = atan2\left(\left(\sum_{i=0}^{n-1}\sin\phi_i\right), \left(\sum_{i=0}^{n-1}\cos\phi_i\right)\right). \tag{4.22}$$

It is then necessary to assess the confidence in this central tendency of phase, for which we can first calculate a measure of the spread in values, the *sample circular dispersion*, δ, by (Fisher, 1993):

$$\delta = \frac{1 - \frac{1}{N}\sqrt{\left(\sum_{i=0}^{n-1}\cos 2\phi_i\right)^2 + \left(\sum_{0}^{n-1}\sin 2\phi_i\right)^2}}{2\bar{r}}. \tag{4.23}$$

With this one can calculate a $100(1-\alpha)\%$ confidence interval $\overline{\phi} \pm \phi_{CI}$, given by (Fisher, 1993):

$$\phi_{CI} = \arcsin z_{\frac{\alpha}{2}}\frac{\delta}{N}, \tag{4.24}$$

where $z_{\frac{\alpha}{2}}$ is a z-score for obtaining the two-tailed probability of α that the value is outside the range. For example, for a 95% confidence interval, $z_{\frac{0.05}{2}}$ is approximately 1.9604. Note that $\bar{r}$, δ, and ϕ_{CI} all have an intrinsic dependency on sample size, so comparisons between these in phase analyses should only be done between phase samples of the same size.

Perhaps use Rizzuto 2006 as a nice example of the above methods. The material for this can be found in the document working_mem.tex

Statistical methods in TF analysis

The power spectral density

A ubiquitous pattern in natural time series is that the associated power-spectrum declines with frequency such that power at any given frequency is approximately $1/f^{\alpha}$, where f denotes frequency and α is a parameter that describes how quickly power declines with frequency.[5] EEG data exhibit this pattern, as shown for an example signal in Figure 4.9. This has the practical implication that in analyses of EEG data, it is important to normalize the baseline power across low and high frequencies (see next section) in order to facilitate comparisons of relative changes in power values across frequencies. As we will see in Chapter 5, matching the range spanned by different features is important for ensuring that regression-based classification methods learn equally from low and high frequency features.

[5] Often times the exponent α is omitted and the spectrum is simply labeled $1/f$. Random noise that has a $1/f$ spectrum is called *pink noise* (distinct from *white noise* which has a flat frequency spectrum).

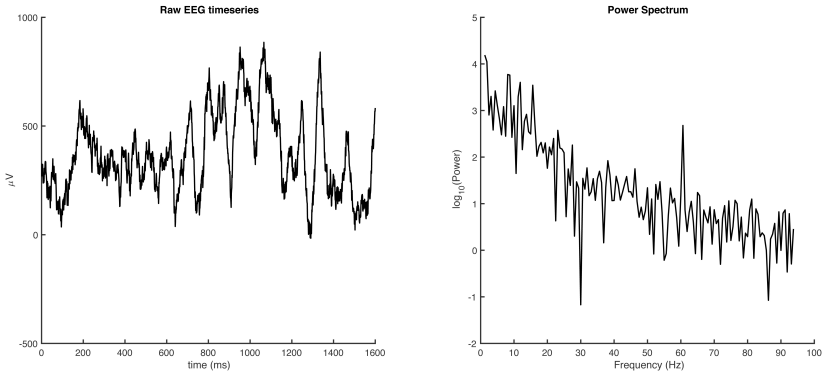


Figure 4.9: Example of $1/f$ power spectrum Data collected from a bipolar pair of intracranial electrodes implanted in the left middle temporal gyrus of a subject performing the FR1 task. The data were collected for the 0-1600 ms encoding interval relative to word onset for a single word event. *Left*, the raw EEG time-series. *Right*, the power spectrum for this event, which exhibits the $1/f$ pattern.

Baseline Normalization

EEG spectral power generally obeys a $1/f$ power law, indicating that lower frequencies tend to have higher power than higher frequencies. To facilitate the statistical analyses of EEG signals, baseline normalization techniques are often used to transform spectral powers at different frequencies onto the same scale. There are several ways to perform baseline correction in time-frequency analyses. The baseline value is typically taken to be the average power of the pre-event period (e.g., pre-encoding). One method is to calculate the percent change from the baseline of the signal by taking the difference between the signal and the baseline, and dividing this difference by the baseline. Another method is to log-transform the signal by taking

the $\log_{10}$ of the quotient of the signal to the baseline, converting the signal into decibel units (dB). A third method involves calculating the z-score of the signal using the distribution of values in the baseline period by taking the difference between the signal and the baseline and dividing it by the standard deviation of the baseline. Table 4.1 summarizes these methods.

Percent-change-normalization and z-normalization bring the variables of interest onto the same scale and eliminate their units. Log-transformation eliminates both the units and reduces the variability of the data by bringing the outliers in the original distributions closer to the rest of the observations. The percent-change-normalization of spectral power of EEG signals typically results in skewed distributions, a property that is not desirable for many parametric statistical analyses. The z-normalization is sensitive to outliers in the baseline distribution and can produce unreliable transformed variables. One should carefully check for outliers in the baseline distribution before applying the z-transformation. The log-transform does not suffer from the aforementioned disadvantages of the other methods.

Method	Formula
%Δ	$(\mathbf{s}_{f,t} - \bar{\mathbf{b}}_f)/\bar{\mathbf{b}}_f$
dB	$10 \times \log_{10}(\mathbf{s}/\bar{\mathbf{b}}_f)$
z	$(\mathbf{s}_{f,t} - \bar{\mathbf{b}}_f)/\sigma_{\bar{\mathbf{b}}_f}$

Table 4.1: Baseline Normalization Schemes. %Δ: percent change, dB: decibel, z: z-transform. $\mathbf{s}_{f,t}$ indicates the signal vector (power values) at frequency f and time t. $\bar{\mathbf{b}}_f$ indicates the average power at frequency f during the baseline period and $\sigma_{\mathbf{b}_f}$ indicates the respective standard deviation.

Researchers traditionally employ parametric methods, which typically assume normality of the distribution from which the data is drawn (also known as the data-generating distribution), to analyze the difference between variables of interest (e.g., power values, ERP amplitudes) under different conditions. However, the normality assumption on the generative distributions of powers is often too restrictive and easily violated in practice. The normality of the *sampling distribution* of the test statistic (e.g., t-statistic) is sufficient even though the data-generating distribution might not be normal. Fortunately, the approximate normality of a typical test statistic such as the mean is guaranteed by the central limit theorem provided that the sample size is large enough and that the data-generating distribution does not deviate too much from normality. In other words, if the data-generating distribution is exactly normal, then the sampling distribution is normal regardless of the sample size. It does not take a large sample size ($N \geq 30$ is the typical rule of thumb) for the sampling distribution to approach normality if the data-generating distribution is close to normal and symmetric. On the other hand, if the data-generating distribution is extremely skewed (e.g., a Bernoulli distribution with 0.99 probability on 0 and 0.01 on 1), it would take a large amount of data (hundreds or even thousands of samples) before the sampling distribution is close to being normal. As a result, transformations (such as the logarithm) that reduce the variability and bring the distribution of the data closer to normality are desirable when one uses the parametric statistical framework. The log-transformation is particularly well-suited for powers because the distributions of powers are approximately distributed as chi-squared, which is positively skewed. As a result, taking the logarithm of power results in a distribution that is close to normality. In addition, log-transformed power exhibits a negative linear trend with increasing frequency. This relationship can be used to study broad-spectrum properties using the regression framework.

One should be mindful about what is being tested after the transformations of variables. For example, testing the mean of the log-transform of power is not equivalent to testing the mean of power.[6] As a result, we cannot make a statement about the mean of the original distribution using a statistical test on the log-transformed data.

[6] The difference comes from the fact that the logarithm of mean is not the same as the mean of logarithms. By Jensen's inequality and the log function being concave,

$$\frac{\log(x_1) + \cdots + \log(x_n)}{n} \leq \log(\frac{x_1 + \cdots x_n}{n})$$

The equality occurs if and only if $x_i = x_j, \forall i, j \in \{1, \cdots n\}$, which does not happen in practice.

Attenuating noise in the raw data

Standard approaches to analyzing both scalp and intracranial EEG data typically include multiple steps for preprocessing, including using statistical thresholds to identify channels or epochs that contain raw signals that are corrupted by noise. A common approach is to calculate statistical measures such as variance/standard deviation or kurtosis of the raw EEG signals. Epochs or channels that show outlying values relative to other epochs/channels are then removed before conducting further analysis. However, there is evidence that such preprocessing approaches are unlikely to improve the researcher's ability to identify brain-behavior correlations of interest (Meisler, Kahana, & Ezzyat, 2019). Figure 4.10 shows a systematic analysis of the effect of such methods on the intracranial FR1 task. The analysis shows that that removing putatively noisy data does not increase *t*-statistics in the subsequent memory analysis.

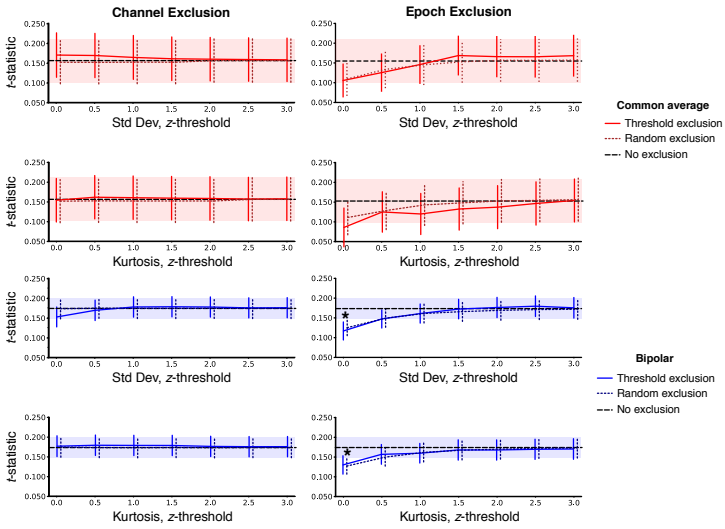


Figure 4.10: Effects of statistical exclusion of epochs and channels. A systematic analysis of statistical thresholding for identifying noisy data showed that removing data does not increase estimates of the subsequent memory effect.

Another class of noise removal methods involves review of raw data by researchers trained to identify abnormal physiological signals, for example epileptic activity. As shown in Figure 4.11 (left panel) such manual removal did not reliably increase univariate estimates of the subsequent memory effect. In the case of multivariate classification of subsequent memory, removing epochs significantly reduced classifier performance. This suggests that the cost of having fewer training observations was not outweighed by any benefit associated with removing epochs with potentially epileptic data.

Across all sets of analyses the preprocessing step that did significantly increase statistical power in the subsequent memory analysis was the use of bipolar referencing, compared to common average referencing. Bipolar referencing schemes use neighboring electrodes to reference one another, creating a new array of 'virtual' channels across all pairs in a participant's montage. While it is clear that such an approach increases the strength of the subsequent memory analysis of high-frequency activity in iEEG and in broadband classification, an open question concerns the effects of bipolar referencing in the analysis of lower frequency signals alone, as well as in data

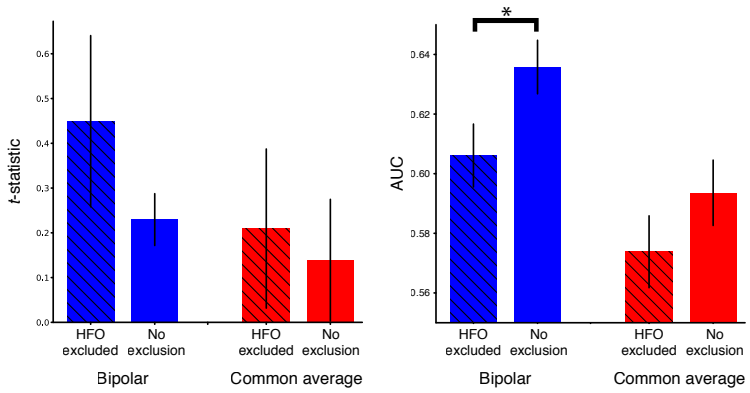


Figure 4.11: Effects of manual exclusion of epochs. Manual review and removal of epochs exhibiting high-frequency activity consistent with epilepsy did not increase estimates of the subsequent memory effect. In the case of multivariate classification (right panel) using bipolar referenced data, manual exclusion of such epochs reliably impaired classifier performance.

collected non-invasively at the scalp.

Multiple comparisons correction in time-frequency analyses

In Chapter 3, we addressed the multiple-comparisons issue associated with testing event-related potentials that involve many channels and time points. In addition to channels and time points, time-frequency analyses have frequency as an extra dimension. As a result, one faces a larger multiple-comparisons problem when testing spectral power. Fortunately, methods introduced in Chapter 3 to control for the family wise error rate or the false discovery rate can be used to mitigate this issue.

5
Regression and Classification

This chapter introduces multivariate approaches for studying the electrophysiology of human memory. Multivariate methods partition into those which aim to uncover structure in unlabeled data (unsupervised learning) and those that learn a mapping between features of the data and some labeled outcome (supervised learning). In this chapter, we focus on supervised learning methods that predict outcomes measured on interval (regression) or nominal (classification) scales.[1]

In studying human memory, we might ask what factors predict performance in a list memory task such as 'free recall'. We can approach this question as a regression problem by considering lists, sessions, or even subjects as the unit of analysis. For example, we might ask what features of neural activity during the study period predict the number of recalled items in the subsequent free recall period. Alternatively, we might be interested in relating the neural activity during the presentation of individual items during the study period to each item's subsequent recall status (because recall status is a binary, categorical variable, we refer to this as a classification problem).

In the standard univariate approach we ask whether each of a (potentially large) number of neural features (e.g., power in a particular frequency band, measured at a specific electrode and in a given time interval) reliably predicts subsequent recall performance, after correcting for multiple comparisons as described in Chapter 3. For the multivariate approach, we ask a different question: Given a set of neural features, how accurately can we predict recall performance? Here we are not primarily asking about the importance of specific features; rather, we are asking how well the ensemble of features can predict our outcome variable. To answer this *prediction* question, we need to evaluate a single model that includes information from all the features. In the case of multiple linear regression, we try to predict recall performance by computing a weighted sum of neural features whose weighting coefficients minimize some objective function (e.g., the deviation between observed and predicted values). As we discuss below, we can use a very similar approach for classification.

The multivariate prediction approach entails a conceptual leap from asking which features underlie function to asking how the distribution of features predicts function. By addressing this problem, we can exploit signals that univariate methods may miss. Specifically, univariate methods may lead us to believe that only a subset of features (e.g., activity in one particular brain region) provides information about performance in a given

[1] Measures at interval scales are those where the differences between units are constant (e.g., the difference in temperature between 10°C and 11°C is the same as that between 20°C and 21°C) whereas nominal scales refer to arbitrary assignments of labels to outcome (e.g., labeling recalled items as "1" and not recalled items as "0"; ?, ?).

task, whereas there may be a lot of collective information distributed across features even when they individually fail to meet a univariate significance threshold (Haxby et al., 2001). Consider another key difference: If several distinct features, X_1, X_2, and X_3 each individually predict the same outcome Y then the univariate approach does not distinguish between cases where X_1, X_2, and X_3 also predict one another and cases where each one provides unique information. By contrast, the multivariate approach will select among these features, finding the linear combination that best predicts Y, which may only use one of the three predictor variables. In traditional regression approaches, as frequently used in the social sciences, standard textbooks admonish students against using correlated independent (X) variables. When building a model using correlated variables, common guidance would have us build composite or factor scores from the correlated predictor variables. This would allow us to interpret the coefficients on these factors by reducing their colinearity.

Linear Regression

Linear regression models a linear relationship between the predictors (such as features of brain activity) and outcomes (such as recall performance):

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon \,, \tag{5.1}$$

where $X = (X_1, \cdots, X_p)$ is a vector of predictors and Y is the outcome variable.[2]

Thus, for each sample, i of n observations $(x_1, y_1), \cdots (x_n, y_n)$ we have

$$y_i = \beta_0 + \beta_1 X_{1i} + \cdots \beta_p X_{pi} + \epsilon_i. \tag{5.2}$$

Using matrix notation, we can compactly write the linear system in (5.2) as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{5.3}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ & & & & \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}_{n\times(p+1)}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}_{(p+1)\times 1}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}. \tag{5.4}$$

The X matrix is known as the *design matrix* and β is the vector of parameters (predictors) of the linear model.

To fit this model to the data, we need to estimate β by optimizing an *objective function*.[3] The objective function usually quantifies the mismatch between model predictions and actual outcomes. There are many ways to quantify this mismatch, but in practice, the sum of the squared differences between actual and predicted outcomes is often used for its computational convenience:[4]

$$l(\beta) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2, \tag{5.5}$$

[2] A common interpretation the β coefficients is that β_j reflects the average effect on Y for every unit increase in X_j. However, when predictors are correlated, as is usually the case in practice, β_j represents the additional contribution of x_j on y after accounting for the contributions of other predictors (Hastie, Tibshirani, & Friedman, 2009).

[3] When the goal is to minimize the objective function it can also be referred to as *loss function*.

[4] Using the squared differences rather than, say, the absolute differences makes the objective function differentiable. This feature allows for efficient minimization of the loss function.

where x_i is the i^{th} row of the design matrix $\mathbf{X}$. Estimating β by minimizing this objective function is known as the *least squares (LS) estimate*:

$$\hat{\beta} = \arg\min_{\beta} (y - X\beta)'(y - X\beta) \tag{5.6}$$

Differentiating the objective function with respect to β, we obtain the *normal equations*

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{5.7}$$

If the matrix $X'X$ is of full rank (non-singular), the unique solution is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{5.8}$$

The LS estimator $\hat{\beta}$ is an unbiased estimate of β.[5] In addition, the *Gauss-Markov*'s theorem asserts that among all linear unbiased estimates, the least squares estimates of β have the smallest variance. Given a new observation x_0 of X, we predict the value of Y to be

$$\hat{y}_0 = x_0'\hat{\beta}. \tag{5.12}$$

[5] Proof that the LS estimator is unbiased.

$$E[\hat{\beta}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}] \tag{5.9}$$
$$= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}(\mathbf{X}\beta + \epsilon)] \tag{5.10}$$
$$= \beta + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\epsilon]. \tag{5.11}$$

If $\mathbf{X}$ and ϵ are uncorrelated and $E[\epsilon] = 0$, then $E[\hat{\beta}] = \beta$

Model evaluation

After fitting a linear model to the data, one might ask: how well does the model fit the data? A common metric that is used for assessing the performance of a linear model is the coefficient of determination, R^2, which is defined to be the proportion of variance explained by the model. Mathematically,

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{5.13}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares ($\bar{y}$ denotes the average) and $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares ($\hat{y}_i$ denotes the model prediction). It can be shown that for a simple linear regression model with one predictor, R^2 is the square of the correlation between Y and X, and thus can range between 0 and 1.[6] If $R^2 = 1$, the model perfectly explains the variability of the response variable. If $R^2 = 0$, then the model does not explain any variability of the response variable. R^2 does not take into account model complexity, and a more complex model will produce a lower residual sum of squares (and hence a higher R^2). An alternative measure of model fit that takes into account model complexity is the *adjusted* R^2, defined as $R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$, where n is the number of observations and p is the number of predictors. Thus the adjustment consists of dividing the sums of squares by their respective degrees of freedom.

[6] Negative values are possible for out-of-sample predictions commonly used for cross-validation, as discussed below.

Logistic Regression

We next consider how to extend the regression framework to categorical outcomes. For exanple, we may wish to use features of brain activity to predict a categorical variable, such as whether a word is recalled or recognized. We refer to these models as classification models, but the distinction between regression and classification is not as clear-cut as it might appear. A case in point is the logistic regression model, the focus of this section, which is generally used for classification. When trying to classify a binary variable (e.g.,

recall vs. failure to recall) it is convenient to assign the values 0 and 1 to the two categories. With this assignment, it might be tempting to use the linear model discussed above to predict these binary outcomes. The output of the linear model, however is not bounded which could let to predictions that are very different from the two valid values and would be difficult to translate to a binary prediction. Rather than predicting the binary outcome directly as in (5.3), we can instead predict a transformed outcome using a link function, $h(\cdot)$:

$$dh(\mathbf{y}) = \mathbf{X}\beta + \epsilon, \tag{5.14}$$

A logistic regression uses the *logit* transform as its link function:

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right), \tag{5.15}$$

which has a range from $-\infty$ to ∞ and thus the left-hand side of (5.14) will never be out of bounds. We can think of the linear model discussed above as using an identity link function, highlighting the close connection between linear and logistic regression.[7]

If we refer to the category with the value 1 as the "target category," we can model the probability of the target category as a linear function of the predictors using the logit link function:[8]

$$\log\frac{p(x)}{1-p(x)} = x'\beta \tag{5.16}$$

If we solve for $p(x)$ in terms of x and β, we obtain

$$p(x) = \frac{1}{1+\exp(-x'\beta)} \tag{5.17}$$

The function $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the *sigmoid function* and thus the predicted probability of the target category in a logistic regression model is a sigmoid function of the linear function of the predictors. Figure 5.1 illustrates the sigmoid function.

[7] Other link functions are also possible, for example log-linear models, sometimes used to model count data, use $h(x) = log(x)$.

[8] Of course, which category is labeled the "target category" is arbitrary and one would predict the complementary probability if the labels were reversed.

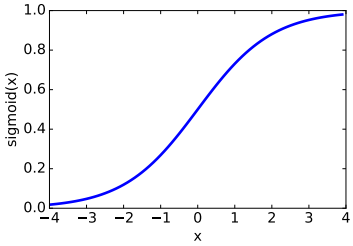


Figure 5.1: The sigmoid function takes in real inputs and returns outputs in the range [0,1] .

Objective function

Suppose we have a training data set $(x_1, y_1), \cdots, (x_n, y_n)$. For each pair (x_i, y_i), the probability of observing y_i given x_i is:

$$P(y = y_i \mid X = x_i) = p_i^{y_i}(1-p_i)^{1-y_i}, \tag{5.18}$$

where $p_i = \frac{1}{1+\exp(-x_i'\beta)}$. The loss function is defined to be the log-likelihood function:[9]

$$\begin{aligned} l(\beta) &= \log\prod_{i=1}^{N} P(Y = y_i \mid X = x_i) \\ &= \sum_{i=1}^{N}\{y_i \log p_i + (1-y_i)\log(1-p_i)\}(\text{ negative cross-entropy function }) \\ &= \sum_{i=1}^{N}\{y_i x_i'\beta - \log(1+\exp(x_i'\beta))\} \end{aligned} \tag{5.19}$$

[9] The squared-error loss function used in linear regression is not convex when applying logit link function and thus difficult to optimize.

A larger log-likelihood indicates a better fit and we thus maximize the log-likelihood function l. The optimal solution $\hat{\beta}$ for the objective function l does not have a closed form. Therefore, we need to solve for β numerically. Hastie et al. (2009) describe numerical methods used to estimate β.

Once we obtain our estimate $\hat{\beta}$, the prediction for a new observation x_0 is given by:

$$\hat{y}_0 = \frac{1}{1 + \exp(-x_0'\beta)}. \tag{5.20}$$

That is, we predict the response for x_0 has a probability $\hat{y}_0$ of observing the target class.

Model evaluation

When evaluating the performance of a classifier, it is tempting to calculate the proportion of accurate classifications.[10] It is easy to show, however, that accuracy alone is not particularly informative. Consider classifying brain activity during encoding with respect to subsequent recall. If a classifier is described as having correctly predicted subsequent recall for 85% of the considered encoding events, it might appear to have extracted meaningful encoding-related brain activity. Such a level of performance could also be achieved, however, if the classifier always predicted subsequent recall and 85% of studied words were indeed recalled. Whereas this classifier correctly predicted subsequent memory for each item that was later recalled, it got every failed recall wrong. When assessing classifier performance, it is therefore important to consider the different types of errors a classifier can make.

[10] A binary classification decision can be derived from a continuous classifier output using a threshold, e.g., by generating a prediction for the target class whenever its probability is at least 0.5.

Chapter 3 introduced the concept of Type I and Type II errors in the context of statistical tests. Just as there are four different possible outcomes in hypothesis testing (two types of correct decisions and two types of errors; see Table 3.1) there are four types of possible outcomes in any binary classification task. A confusion matrix specifies all possible outcomes and their counts or probabilities (see Table **??**). *true positives* or *hits* denote correct target classifications; *false positives* or *false alarms* denote incorrect target classifications. Because we know the total number of target (n_T) and non-target ($n_{\neg T}$) items, the hit- and false-alarm rates fully specify the confusion matrix. Thus, a comprehensive assessment of classifier performance can be achieved by taking both of these measures into account. There are many indices of classifier performance that differ in the details of how these measures combine and whether they only consider binary predictions or the continuous classifier output. The next section presents details of the area under the receiver operating characteristic (ROC) function measure of classification performance. This function, known as the AUC, is also commonly calculated to characterize the performance of human classifiers in recognition memory experiments.

	Target	Non-target
"yes"	true positives (hits)	false positives (false alarms)
"no"	false negatives (misses)	true negatives (correct rejections)
Total	n_T	$n_{\neg T}$

Table 5.1: Confusion matrix for a binary classification task; "yes" and "no" labels indicate classifier predictions for and against the target class respectively with "target" and "non-target" labels reflecting the true class membership. The *hit rate* is number of true positives, divided by n_T and the *false alarm* rate is the number of false positives divided by $n_{\neg T}$. In the machine learning literature, the hit rate is sometimes referred to as *recall* and in the context of medical tests it is known as *sensitivity* (with *specificity* corresponding to the correct rejection rate).

Receiver operating characteristic functions

All binary classification tasks share the basic structure described above. As such, the problem of assessing classifier performance is quite general, with applications across a wide range of fields, including engineering, psychology, and statistics. Signal Detection Theory (Green & Swets, 1966) offers a framework for analyzing performance in detection tasks built on work studying the characteristics of radar receiver operators during World War II. As the

gain on a radar receiver is increased, it is increasingly sensitive to relevant objects (e.g., enemy aircraft), but also increasingly likely to display spurious signals (Streiner & Cairney, 2007). Receiver operating characteristic (ROC) functions describe the performance of a classifier as the threshold for target identification is varied by relating false positive rates to true positive rates.

Just as the probability of a false alarm increases for radar receiver operators as the gain on the radar receiver increases, psychological variables determine the balance between true and false positive responses in human classifiers. In memory research, binary classification is the predominant method for assessing recognition memory (see, Chapters 1 and 3). The probability of endorsing a probe item as previously studied (and thus the hit and false alarm rates) vary with variables such as the base rates of previously studied and new probe items and the reward structure of the task (e.g., setting the reward for hits to be higher than the loss associated with making false alarms can increase the proportion of probe items endorsed as studied). A logistic-regression classifier will generate a predicted probability of a target label, and we can use this probability to make a binary classification decision. As the threshold for endorsing the target class decreases from 1.0 to 0.0 the balance of hits and false-alarms associated with these thresholds traces out the ROC function.

The area under the ROC curve (AUC) provides a useful index of classifier performance that makes minimal assumptions about the underlying data-generating process. Classifiers that cannot distinguish between the two classes produce completely overlapping distributions of predictions; hence the resulting ROC function falls on the main diagonal in ROC space (because hit and false alarm rates are identical at each threshold). The resulting AUC is 0.5 whereas perfect separation of the two classes produces an AUC of 1.0.[11]. Generally, the AUC reflects the probability that a randomly chosen target instance will be associated with a higher predicted probability than a randomly chosen non-target instance, making it a directly interpretable ("normalized") index of classifier performance. Furthermore, the AUC is the statistic computed for the Wilcoxon test of ranks and directly related to the Gini coefficient. Fawcett (2006) presents a thorough overview of ROC analyses that illustrates these properties and, among other things, addresses generalizations to multi-class classification problems, comparisons to other methods for assessing classifier performance, and statistical issues associated with ROC data.

Overfitting, Underfitting and The Bias-variance Decomposition

In applications that use neural measures to predict performance in a psychological experiment, the number of predictors usually exceeds the number of observations (e.g., recall performance might be modeled by power in a range of frequencies at each available electrode). As such, it is likely that a substantial fraction of the modeled relationship between predictors and outcomes will thus not generalize to independent test data. In such cases, we say that the model overfits the training data. One way to avoid overfitting is to constrain the parameters of the model to avoid individual predictors from having large effects—a technique known as *regularization*. A model that is too constrained, however, is prone to underfitting, a failure to pick up on relations between predictors and outcomes. Figure 5.2 illustrates an example

[11] In case of substantial overfitting, performance on a held-out data set can sometimes fall below 0.5

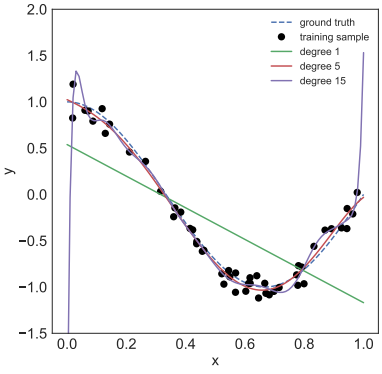


Figure 5.2: An example of polynomial regression with degree equal to 1,5, or 15. The training data set consists of 30 observations simulated according to $y = \cos(1.5\pi x) + \epsilon$, where $\epsilon \sim \mathrm{N}(0, 0.01)$. The figure shows that a linear function badly *underfits* the training data. A polynomial of degree 5 provides a very good fit to the data. A higher degree polynomial, such as one with 15 degrees, *overfits* the training data because it learns the noisy component of the data.

of fitting a function with an overly complex model (overfitting), a model that is too simple (underfitting), and a model that provides a reasonable fit to the data. The concepts of overfitting and underfitting can be described precisely by the two theoretical properties of the estimated function $\hat{f}$: *bias* and *variance*. Consider a regression problem: $y = f(X) + \epsilon$, where $\mathrm{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2$ and suppose that we have estimated a prediction function $\hat{f}$ to fit the true function f by applying a learning algorithm to a training data set.

The prediction error of a new point x_0 with $y_0 = f(x_0) + \epsilon_0$, under the squared error loss is given by:

$$\begin{aligned}\mathrm{E}[y_0 - \hat{f}(x_0)]^2 &= \mathrm{E}[f(x_0) - \mathrm{E}\hat{f}(x_0) + E\hat{f}(x_0) - \hat{f}(x_0) + \epsilon_0]^2 \\ &= [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 \\ &\quad + \mathrm{E}[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\ &\quad + \mathrm{Var}[\epsilon^2] \\ &= \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}.\end{aligned}$$

Deriving the second equation from the first requires expansion of the square and algebraic simplification. The term $\mathrm{E}\hat{f}(x_0) - f(x_0)$ is called the bias of the prediction function $\hat{f}$ since it measures the average deviation from the ground truth. $\mathrm{E}[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2$ is called the variance of $\hat{f}$ since it measures how much the estimate $\hat{f}$ changes with new training data. $\mathrm{Var}[\epsilon^2]$ constitutes irreducible error due to noise; it cannot be eliminated regardless of how well we are able to estimate f. In general, we want a learning algorithm that has low bias and low variance, but in practice there tends to be a trade-off such that reducing variance increases bias, and vice versa.

Cross Validation

If we were to fit polynomial functions of increasing complexity to the data shown in Figure 5.2 the deviation between the observed and predicted values ("training error") would decrease. However, if we were to make predictions about new data generated in the same way, the resulting "testing error" would increase once the complexity of the model exceeds an optimal level (indicating overfitting). Figure 5.2 illustrates this relationship between model complexity and training and testing errors.

In general, we are less interested in accurately describing the observed data (including the noise inherent in our measurement) than we are in being able to generalize to future observations generated in the same way. The training error is therefore a poor measure of model success and we thus need to calculate a test error for data that was not used to fit the model. Electrophysiological data from humans engaged in memory tasks is expensive to collect and therefore it is not usually practical to hold out a large part of the data for the calculation of a testing error. Cross-validation refers to a class of approaches to calculate a testing error while making efficient use of the data.

One standard approach is the k-fold cross-validation that forms k partitions out of the full data set . Each of these partitions is held out once with the model being trained on the remaining $k-1$ partitions of the data as illustrated in Figure 5.4. We are thus able to make full use of the entire data set by repeatedly fitting the model on a subset of the data and evaluating it on the held-out partition. In applying this approach to electrophysiological data

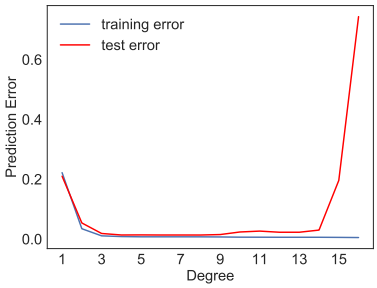


Figure 5.3: The blue line represents the training error, and the red line the test error. The test error is computed using 100 simulated observations other than the training sample. The linear model is inadequate, demonstrating both high training error (underfitting) and high test error. On the other extreme, a polynomial with a high degree learns the training data too well (overfitting) and fails to generalize to independent test data. The optimal degree is around five in this case.

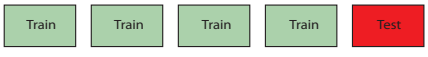


Figure 5.4: Five-fold cross-validation . In this example, we train the machine learning algorithm on the first four folds (green) and then test its performance on the hold-out fold (red).

one must take great care to insure that one's training algorithm cannot see any parts of the data from the held-out partition. Even a very small degree of overlap between training and testing data will completely invalidate the cross-validation procedure. This means that one must not do any normalization that includes data from the held-out partition and one must ensure that any spectral filters do not cross the boundaries between "folds". [12] If $K = N$, the sample size, we call this procedure *leave-one-out* cross-validation because each observation is a fold itself.

A better approach for cross-validating electrophysiological data involves partitioning the data according to the task structure. If one can collect N sessions of data from the same participant then one can train a model on $N - 1$ sessions and test the model on the held-out-session, repeating the process for each session as one would with the k-fold technique. If one cannot do this at the level of sessions, then one could do this at the level of lists so long as one is careful to maintain complete separation between training and held-out lists.

[12] Formally, for the k^{th} fold, we train a model on the remaining $K - 1$ folds to obtain an estimator $\hat{f}^{-k}$ and then compute the testing error for the held-out fold:

$$\mathrm{CV}_{error}(\hat{f}^{-k}) = \frac{1}{n_k} \sum_{i \in C_k} l(y_i, \hat{f}^{-k}(x_i)) \,, \tag{5.21}$$

where C_k and n_k denote the set of indices and the number of observations in the k^{th} fold. The cross-validated estimate of the testing error is given by

$$\mathrm{CV}_{error}(\hat{f}) = \sum_{i=1}^{K} \frac{n_k}{N} \mathrm{CV}_{error}(\hat{f}^{-k}). \tag{5.22}$$

That is we take a weighted average of the fold-based errors to obtain the cross-validated error.

Nested cross validation

When we need to estimate meta-parameters of the model (e.g., how many features to include as predictors or how to regularize the model to limit complexity) we must evaluate the performance of the model for these different settings on data that was not used for training (a validation set) and then evaluate the model with the final meta-parameter values on an independent testing set. With ample data, we can partition the total dataset into separate training, validation, and testing sets. But in the case of electrophysiological data, we usually need to rely on cross-validation. One way to adapt cross-validation to accommodate both validation and testing is to implement a nested cross-validation procedure. In the case of k-fold, or leave-one session out, cross-validation, we would first hold out one session/fold as a testing set (as explained above) and then repeatedly hold out one of the remaining sessions/folds as a validation set. Thus, in the case of 5-fold/session cross-validation, we would have 5×4 iterations of the cross-validation procedure.

Regularization and Model Selection

A common problem in machine learning is overfitting, which occurs when a model learns not only the signal but also the noisy component of the training data set. As a result, the model fails to predict unseen future data (test set). In Figure 5.2, the polynomial regression model of degree 15 overfits the training set by learning its idiosyncratic features, and the shape of the predicted function $\hat{f}$ is heavily influenced by the noisy component in the training set. In the context of regression analysis, when there are many correlated predictors in the model, the estimates of the coefficients will be poorly determined and have high variance due to the design matrix being nearly singular. In such cases, the regression can yield highly correlated coefficients with large magnitudes; these coefficients can cancel each other out to generate reasonable predicted outcome values. A common way to combat overfitting is to utilize regularization techniques that penalize large coefficients. Regularization techniques are ubiquitous in machine learning and can be applied to complicated models such as logistic regression, trees, deep neural net-

works, etc. In this section, we will introduce regularization in the context of regression.

The penalized regression model can be formulated as follows:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\beta)^2 + \lambda \text{Penalty}(\beta). \tag{5.23}$$

That is, we find $\hat{\beta}$ that minimizes the sum of squared error and the penalty imposed by the value of β.[13] One of the most common penalty terms for β is the L_q norm: $||\beta||_q^q = \beta_1^q + \cdots + \beta_p^q$. The learning model is *Lasso* regression when $q = 1$ and *Ridge* regression when $q = 2$. The penalty term λ controls how much we penalize large coefficients. Consider two extreme cases: When $\lambda = 0$, we do not penalize large coefficients at all, and the estimated $\hat{\beta}$ is our usual least-squares coefficient. On the other hand, as $\lambda \to \infty$, the estimated regression coefficients approach o because the algorithm severely penalizes non-zero coefficients.

Regularization reduces the variances of the coefficients by constraining their sizes. However, it comes with a price. Regularization introduces bias into the estimates. To elucidate this point, consider L_2 (ridge) penalized regression. The coefficient vector $\hat{\beta}$ minimizes the following objective function:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\beta)^2 + \lambda \beta'\beta. \tag{5.24}$$

We can rewrite Equation 5.25 in a compact matrix form,

$$\hat{\beta} = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \beta'\beta \tag{5.25}$$

where $\mathbf{y}$, $\mathbf{X}$ are defined as above. The ridge regression has a closed-form solution:

$$\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1} X'y \tag{5.26}$$

where I is a $p \times p$ identify matrix. Again, the ridge solution is a linear function of $\mathbf{y}$. When $\lambda = 0$, we recover the usual least square solution. When $\lambda > 0$, the ridge solution is no longer an unbiased estimator of β due to the λI term in the solution. As explained above bias and variance trade off and as λ increases, the variance in the ridge solution also decreases, at the cost of increased bias.

Features in regression and classification models

Electrophysiological studies of memory often use spectral power at several frequency bands and multiple recording electrodes as neural features for decoding brain states. As discussed in Chapter 4, power varies widely across different frequencies and typically decreases with increasing frequency according to the relation $power \propto 1/f^{\alpha}$. Because regularization penalizes large parameter values, it is important to normalize the neural features so that they have the same mean and standard deviation. This can be achieved with a simple z-transform, by calculating the mean and standard deviation for each feature (e.g., each frequency at each electrode), subtracting the respective means, and dividing by the respective standard deviation.

In theory, there is no limit to the number of features we can extract from electrophysiological recordings, but more complex models will lead to

[13] The penalized objective function for the logistic regression model is

$$l(\beta) = \sum_{i=1}^{N} \{y_i x_i'\beta - \log(1 + \exp(x_i'\beta))\} + \lambda \text{Penalty}(\beta) .$$

greater overfitting. Above, we discussed regularization as a way to limit a model's complexity. An alternative approach is to limit the number of features (feature selection) or to transform the features into a lower-dimensional space (feature engineering). By limiting the parameter values, regularization limits how much a feature can contribute to the model. An extreme version of this is to set a parameter to 0, thus removing the corresponding feature from the model. When using an L_1 norm for regularization, the optimization procedure may exhibit this behavior and completely remove some features from the model.[14] There are many other ways to restrict the number of parameters, for example, based on prior knowledge about which features are important or based on separate statistical tests. Features extracted from electrophysiological recordings often exhibit high correlations with one another (e.g., between features obtained from nearby electrodes), and thus dimensionality reduction techniques, such as principal components analysis (PCA) can help reduce the number of parameters that need to be estimated from the data. (see e.g., ?, ?; for an introduction to these methods).

[14] for L_q norms with $q > 1$, parameter values can become arbitrarily close to 0 but never exactly 0.

Applications

Below we consider several applications of standard and logistic regression models in the analysis of data from memory experiments.

Application I: Predicting list level recall performance.

In a typical free recall experiment, participants study multiple lists of items, each followed by a recall period during which they attempt to recall as many items as they can from the just-studied list (see Chapter 1). Even if one endeavor to equate the study lists on variables that could affect recall performance, such as list length and item difficulty, the proportion of recalled items can vary dramatically from list to list (Kahana et al., 2018). This variability in performance appears within individual research subjects, each of whom exhibits better or worse memory from moment to moment and from day to day. This application evaluates the extent to which we can predict list-level recall performance from brain activity during the presentation of the study lists for an individual participant in Experiment 4 of the *Penn Electrophysiology of Encoding and Retrieval (PEERS)*. We calculated power across 15 frequencies ranging from 2-200 Hz from scalp EEG recordings as subjects studied common English nouns. We averaged the log and z-transformed power values across all study words in each list, and we used the list-average-z-scored powers across frequencies and electrodes as features in a ridge regression model. Because the proportion of recalled items must fall between 0 and 1, we trained the ridge regression model to predict the logit-transformed proportion of recalled words ($\text{logit}\left(p(\text{rec})\right)$).[15]

For this example, we fixed the regularization parameter and evaluated the performance of our ridge regression model using a leave-one-session-out cross-validation procedure. Figure 5.5 plots the actual list-level recall performance (ranging from no recalled words to perfect recall of all 24 words) against the predictions from our ridge regression model (both back-transformed from logits to probabilities for convenience). Whereas the range of predicted list-level recall performance is noticeably smaller than that of the actual list-level recall performance, and there is considerable variability in the

[15] $\text{logit}\left(p(\text{rec})\right) = \log\left(\frac{p(\text{rec})}{1-p(\text{rec})}\right)$. To avoid invalid values for lists in which all or none of the words were recalled we set $p(\text{rec})$ for those cases to 0.999 and 0.001 respectively.

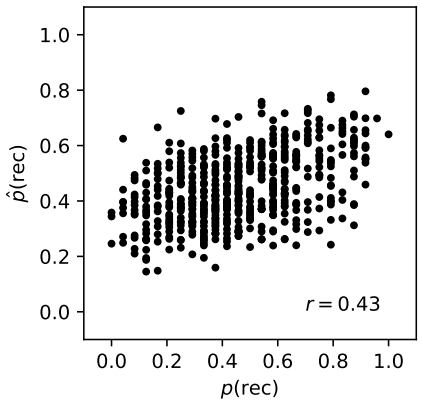


Figure 5.5: Actual and predicted list-level recall performance. The probability of list-level recall ($p(\text{rec})$) for all lists across all sessions for participant 344 in the PEERS Experiment 4, plotted against corresponding predicted list-level recall performance ($\hat{p}(\text{rec})$) from the ridge regression model described in the text. The correlation between the actual and predicted list-level performance across the 576 lists is indicated in the lower left.

predicted performance for each level of actual performance, one sees a clear relationship between actual and predicted recall performance. Figure 5.5 indicates the correlation between actual and predicted recall performance ($r = 0.43$); another popular metric is the coefficient of determinant R^2 which can be calculated as $1 - \frac{\text{RSS}}{\text{TSS}}$ where RSS and TSS are the residual and total sum of squares respectively.For our example, the RSS and TSS were 691 and 848, respectively, leading to $R^2 = 0.19$ (which corresponds closely to the squared correlation between actual and predicted recall performance). For a comprehensive example of using brain activity to predict list-level recall performance, see Weidemann and Kahana (2021).

Application II: Predicting subsequent memory for individual items.

We can also use the techniques described in this chapter to investigate the extent to which neural activity during item encoding predicts subsequent recall. Given the numerous determinants of successful recall, it is unlikely that measures of brain activity during encoding alone would predict recall with high precision. Nevertheless, it is likely that brain activity during encoding contains some information about how well the item is processed and later remembered. It is not clear a priori, however, what the relevant signals would be. Furthermore, diagnostic signals might be weak in isolation and only able to predict subsequent recall reliably in concert with other such features—an ideal use-case for a classifier that learns the importance of each feature from the training data.

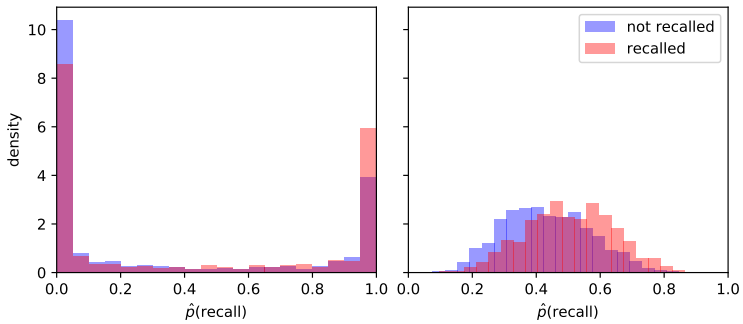


Figure 5.6: Normalized densities of predicted probabilities of subsequent recall from penalized logistic regression classifiers. The classifiers were trained on iEEG activity during encoding in a single patient for two different levels of regularization. At the lower level of regularization (left; $\lambda = 1$), predicted probabilities tend to be extreme, whereas less extreme predicted probabilities are more common with stronger regularization (right; $\lambda = 1000$). Despite considerable overlap between the distributions at both levels of regularization, it is clear that predicted probabilities of subsequent recall tended to be larger for items that were later recalled indicating that the classifiers were able to distinguish encoding events on the basis of subsequent recall.

We trained a logistic regression classifier on average power in a range of frequencies between 5 and 80 Hz during the presentation of each word in the encoding phase of a free recall task (the FR1 experiments deascribed in the Appendix). Figure 5.6 illustrates predicted probabilities of subsequently recalled and not recalled items, for one participant and two levels of regularization. It is clear from the figure that the level of regularization has a strong effect on the distribution of predicted probabilities and the overlap between these distributions for the target and non-target class. The predicted probabilities tend to be larger for the target class than for the non-target class indicating that the classifier extracted features that distinguish between the two classes. The figure also suggests that higher levels of regularization leads to greater separation between the distributions of predicted class probabilities. We can quantify this intuition by examining the ROC functions shown in Figure **??**. For comprehensive examples of using brain activity to predict recall for individual items, see Weidemann et al. (2019).

Application 3: Comparisons among machine learning methods

Here we apply five commonly-used machine learning methods to the basic problem of predicting recall of items based upon neural features recorded during the study phase of an experiment. The data for this example consists of 20 subjects who performed at least three sessions of a free-recall tasks. We trained classifiers to discriminate between good versus bad memory-encoding using spectral features derived from activity recorded at many brain locations. We used a cross-validation approach in which we train the classifier on $N - 1$ sessions and then test it on the remaining session, repeating this until we obtain out-of-sample predictions for every session. We then compute an ROC curve relating true and false positives as a function of the criterion used to assign regression output to responses labels.

All five algorithms reliably classified subsequent memory based on brain activity during item encoding. Figure **??** illustrates the area under the ROC curve for each classification algorithm. Although we obtained the best performance for the L2-penalized logistic regression classifier, that does not mean that this classifier generally outperforms the others. The differences between classifier performance more likely reflects the statistical structure of the noise in the features being used. Thus, with different neural features we may have observed a different ranking between the algorithms.

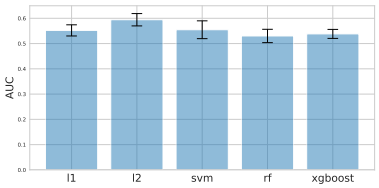


Figure 5.8: Classification Performance for Various Machine Learning Algorithms

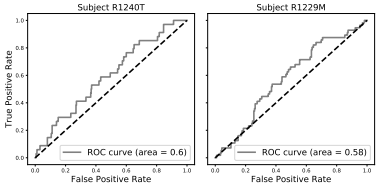


Figure 5.9: Sample ROC curves (gray) of L2 penalized logistic classifiers for 2 intracranial subjects who performed a free-recall task. The dashed black lines indicate ROC curves of at-chance classifiers.

Summary and Future Directions

We presented standard and logistic regression models to illustrate the general multivariate approach for relating brain activity to performance in psychological experiments. We have found these models useful in our analyses of the electrophysiological basis of human memory. Still, they are only some of the simplest examples drawn from a large set of statistical models that can help us to relate brain activity to cognitive function. Application 3 briefly demonstrated how applying other machine learning methods to the same dataset yields similar levels of classification performance.

In all of the examples presented here, however, we used spectral power at several (eight) frequency bands and at each electrode (or bipolar pair of intracranial electrodes) as our "features" for decoding brain states. Prior work using univariate methods had already established spectral features as statistically reliable correlates of memory performance. But we could have made other choices. Machine learning is model fitting, and the more features one uses, the more complex the model becomes and the more likely the model will overfit the data, leading to poor generalization. Penalization and cross validation help, but sometimes you can use prior knowledge about the data, from other studies or from the statistical structure of the data itself, to improve feature selection.

If, for example, you knew that two features were extremely highly correlated, and that each was subject to independent sources of noise, then including both features would be unlikely to benefit classifier generalization enough to warrant the overfitting that would occur by fitting to each features noise during the training phase. In this case, you would prefer to throw out redundant features. One systematic technique for doing this involves transforming your original features vectors into another vector space of lower dimensionality that has less redundancy between features. You might start

by examining the correlation matrix between your features and look at the highest correlations. You could write a computer program that systematically asks how well each variable can be predicted by the other variables and then throw out those variables with the highest R^2 values. But this would also involve overfitting and would be highly sensitive to the order in which you iteratively discard variables. A better method would be to use a dimensionality reduction technique, such as principal components analysis (PCA) to the data and then use the PCA-derived features in the classifier. Although we do not present results of this approach here, prior work by our group has found this to provide highly reliable classificaiton performance, but not numerically superior to the performance achieved by the standard L^2 logistic regression classifier.

6
Recoding and Analyzing Individual Neurons

Introduction: Why measure single-neuron activity in humans?

Neurons likely represent the fundamental unit of information processing in the brains of both human and non-human animals. As such, we would like to understand the relations between neural firing and complex behavior, as these relations will likely offer insights into the mechanistic bases of human cognition. A range of laboratory neuroscience studies had examined single-neuron firing during a range of behaviors and showed that neuronal spiking activity contains a wealth of behaviorally relevant information. These studies have shown that neurons across the brain vary their firing rates in relation to behavioral events, including signals related to perception, motor, and memory processes. However, most existent studies on the brain's single-neuron firing patterns come from laboratory studies of animals. This leaves a substantial gap in our understanding because many cognitive processes are difficult or perhaps impossible to study without humans.

Fortunately, in recent years it has been possible to address this gap, with the development of microelectrodes that are deployable in neurosurgical patients (Fried et al., 1999). These electrodes, which extend from the tips of standard clinical depth electrodes, have a small conductive recording surface ($\sim$40 μm diameter) that allows them to record the spiking activity of individual neurons or sometimes small neuron groups (Quiroga, Nadasdy, & Ben-Shaul, 2004). The activity from these cells is thus much more spatially precise compared to the larger-scale signals that are obtained from the macroelectrodes used for conventional intracranial EEG studies, which measure activity from $\sim 5 \times 10^6$ cells (J. Miller, Polyn, & Kahana, 2007).

Cluster cutting: Distinguishing individual neurons from microwire recordings

Once a microwire electrode is implanted and connected to a recording amplifier, it immediately begins recording electrical activity from the surrounding tissue. Given their high impedance, these electrodes often sample signals from a radius of roughly 100–200μm. Depending on the anatomical organization and neuropil density in the precise location where each electrode is implanted, this means that a microelectrode may measure the activity of multiple neurons. Fortunately, in many cases the waveforms of individual

neurons will appear with different shapes. This is a result of the complex spatial 3-D distribution of the voltage and currents in the immediate area surrounding each neuron (?, ?). Through a procedure commonly referred to as "cluster cutting" it is possible to differentiate among these waveforms to estimate which specific waveforms belong to unique neurons.

Figure **??** illustrates the cluster cutting procedure (?, ?). Beginning with a recording of the voltage-time series from one microelectrode (typically with a sampling rate of at least $\sim$ 10kHz), the purpose of this procedure is to obtain the times when individual action potentials occurred and to label each spike according to the neuron, or "cluster," from which it came. The first step in this process is to take the raw voltage time series and filter it to include only frequencies above $\sim$4kHz, which is the band where action potentials show up. Second, detecting timepoints when the amplitude of the filtered signal exceeds a threshold (usually 3–5 standard deviations) identifies possible extracellular action potentials.

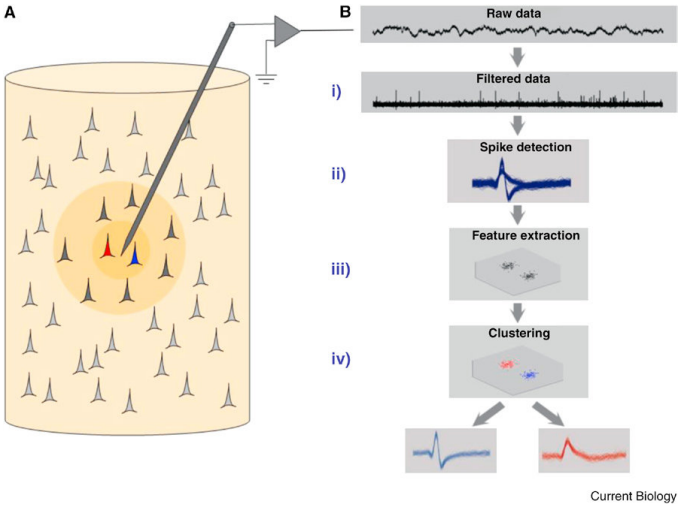


Figure 6.1: Figure from Quoriga 2012 spike sorting article

Following thresholding, the identified timepoints consist of large-amplitude fluctuations that have a duration roughly comparable to the 1–2-ms duration of an action potential. The following steps in cluster cutting are designed to distinguish whether these fluctuations reflect true neuronal action potentials or noise. If they are true action potentials, then this process tries to distinguish which spike waveforms came from different neurons on the basis of differences in waveform shape. To accomplish these goals, the next step of cluster cutting involves extracting various features from the shape of each neuron, such as identifying the height of its upward and downward peaks, as well as spectral measures of its waveform such as its frequency and principal components. Together these measures provide a quantitative summary for the waveform shape from each action potential. This provides a visual guide so as to which neurons have similar shapes and thus are likely to come from the same underlying neuron. This procedure is generally performed by creating a scatter plot of the features from individual spikes, which generally creates a series of visually discernible clusters—these plots are the cause of

the name "cluster cutting."

Finally, based on the clusters that appear in the scatter plot of spike waveforms, the final steps are to label the clusters that represent distinct neurons, as well as to distinguish the cells that seem to reflect noise clusters. If one finds highly distinct (i.e., spatially separable) clusters in feature space, then we can have high confidence that those clusters reflect different neurons. However, one often finds considerable overlap between clusters. This lack of separation creates situations where the experimenter has to use their own judgment to estimate whether two point clouds reflect different neurons. A related issue concerns determining which identified waveforms reflect true neuron waveforms versus artifacts or background noise. One technique used to distinguish action potential waveforms caused by noise is by conducting spectral analysis of the time course of the action potentials from one cluster. This can identify, for example, potential spike waveforms that consistently appear at a frequency of 60 Hz or its harmonics, which would provide a strong indication that an apparent cluster reflected noise.

The above description of cluster cutting focuses on the mostly manual steps that must be performed as part of this process. In addition, several software packages can facilitate this process, such as Wave_Clus, Combinato (), or MountainSort, and others. However, even with assistive software, Cluster cutting is an imperfect science that usually requires manual guidance. Nonetheless, this procedure constitutes a key element of single-neuron physiology.

Following cluster cutting, the next step for most analyses is to compare the activity of each neuron to a person's behavior. Although there are many ways to assess a neuron's activity from its spiking, the most common method is to measure the frequency, or firing rate, of a neuron's action potentials in a given interval and to measure how this quantity varies in proportion to a subject's simultaneous behavior following sections give several examples of how researchers have measured changes in neuronal firing rates to characterize cells that encode activity related to memory or spatial processing. Together, the studies summarized below show that we now have evidence for neurons in the human hippocampal formation whose activities represent all the key types of features that comprise episodic memories: neural correlates of concepts, of space, and of time. Episodic memory involves linking these three representations together and allowing an input representation to pattern-complete the missing features of a memory.

Single-neuron codes for concepts

Because the hippocampus is a common target region for microwire recordings, the focus of many studies is to examine how human single-neuron activity relates to high-level cognitive processes, including memory. The wodrk in this area compared the firing rates of hippocampal cells as subjects recognized and processed individual stimuli, with a goal of identifying neurons whose firing rates distinguished the specific content of a viewed item. Early work in this area began by comparing neuronal firing rates as neurosurgical patients with implanted microwires viewed words on a computer screen. By comparing firing rates as subjects viewed different words, ? (?) reported cells whose firing rates significantly changed according to the identity of a viewed word (Fig. **??**A). Neurons responded when subjects viewed

some words but not others. This work raised the intriguing idea that human hippocampal neurons exhibit sensitivity to abstract representational patterns. This led to significant follow-up work exploring the specific nature of these representations.

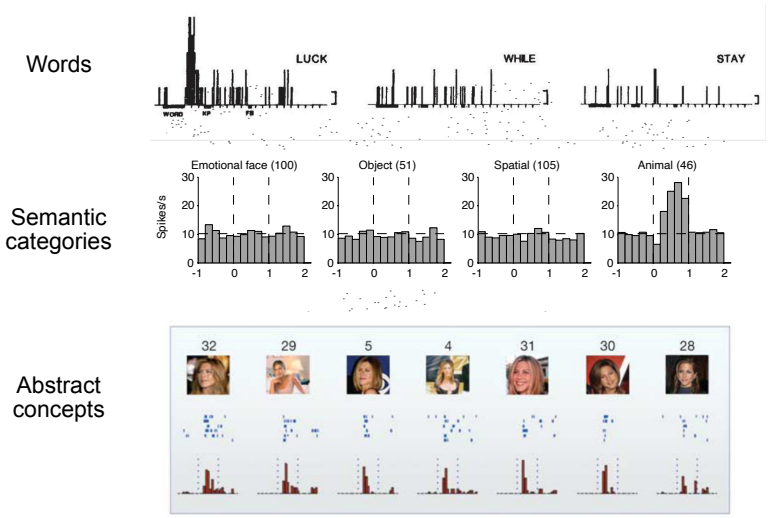


Figure 6.2: (a) word-specific neurons in memory Heit, Halgren, category specific neurons (Kreiman 2000), (c) Jennifer Aniston cells

Kreiman, Koch, and Fried (2000) asked whether the activity of hippocampal neurons responded to the semantic organization of items, such as whether individual neurons specifically responded to sets of images or words from related categories, as opposed to random stimuli. Figure **??**B shows one figure from this study, illustrating the activity of a neuron with an increased firing rate when a subject viewed an image of an animal but not objects from other categories. These results emphasized the degree to which neurons in the human hippocampal formation exhibit abstract coding properties, with their firing rates correlating with high-level features of a subject's cognitive state, in contrast to neurons from sensory regions that represented perceptual properties of viewed items (?, ?).

A line of research beginning with Quiroga, Reddy, Kreiman, Koch, and Fried (2005) advanced the hypothesis that neurons in the human hippocampal formation represent abstract representations of a person's cognitive state. By having subjects view different images of the same person, these investigators found that individual hippocampal neurons activated consistently to the identity of a target individual regardless of the specific photographic representation being viewed (see e.g., Figure **??**C, which illustrates the activity of a neuron that responded to various images of the actress Jennifer Aniston). The images that caused this neuron to activate were strikingly different, and in fact, the same cells activated even when hearing the actress's name.

Across these studies, the pattern of results supports the view that individual cells in the hippocampal system activate to represent abstract features of a person's cognitive state rather than those related to low-level percepts. Because the hippocampus is known to be vital for memory encoding (Scoville & Milner, 1957) and has widespread cortical projections (), these findings help explain the nature of the brain-wide networks that support memory

retrieval by indicating that the hippocampus is activated by highly abstract neuronal representations that then may reinstate brain-wide patterns that contain more detailed signals related to sensory information (?, ?).

Spatially-selective neural responses during virtual navigation

In addition to its known role in episodic memory encoding, the hippocampus and surrounding structures also play a role in spatial navigation. Recordings of hippocampal neurons in rats identified "place cells," each of which typically show low background firing rates but show high spiking activity when the rat is located at a particular location in a spatial environment (O'Keefe & Dostrovsky, 1971). In the nearby entorhinal cortex, neurons behave as "grid cells", activating when a rodent occupies one of many spatial locations arrayed across a spatial environment as if occupying the vertices of a tesselating series of equilateral triangle (Hafting, Fyhn, Molden, Moser, & Moser, 2005). Studies in humans sought to confirm that this phenomenon existed in humans to confirm that this potentially important interspecies similarity related to spatial navigation and cognition (Ekstrom et al., 2003; Jacobs et al., 2013).

Although neurosurgical patients with implanted electrodes are normally confined to their hospital beds, it is possible to study neuronal responses related to spatial cognition using virtual reality. By performing a spatial memory task embedded within a computer-controlled 3D virtual environment, it revealed how neuronal firing rates varied as a function of a subject's virtual location and direction during navigation in virtual reality environments. These studies successfully identified neurons in the human hippocampus and entorhinal cortex that behaved as place and grid cells by activating when the subject was located at one or many locations across an environment, respectively (Fig. **??**A,B). Furthermore, these studies also identified "head direction" cells, whose firing rate varied as a function of the direction that the subject was pointed in a virtual environment (Figure. **??**C), mirroring findings from rodents (?, ?).

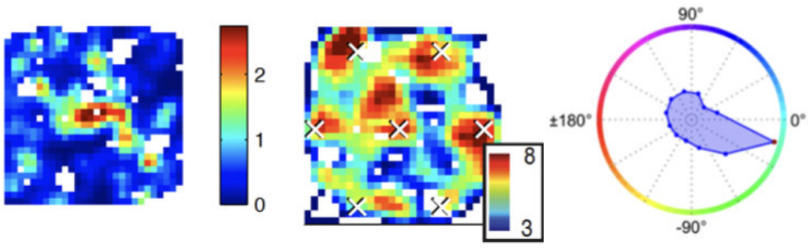


Figure 6.3: Examples of human place (A), grid (B), and direction cells (C). A and B from Jacobs et al., 2013; C from Kunz et al (2021).

More complex human spatial cell types

Beyond confirming that humans, like simpler animals, have neurons such as place and grid cells that faithfully activate according to a person's own location, more recent work has shown that during navigation humans also show novel spatial firing patterns that were not observed previously in other

animals. A more recent study by Tsitsiklis et al. (2020) showed that humans also have neurons whose firing rates correspond to remote locations, in contrast to place cells that represent an animal's own current location. These findings were made by having subjects perform a spatial task where they are directed to follow a specific path during navigation, as seen in Figure **??**A. During this task, the firing of ∼20% of hippocampal neurons significantly changed their firing rate as a function of the location where the subject was trying to go, marked in this task by a treasure chest. In contrast to how a standard place cell would be expected to behave, these cells did not respond to the subject's own position. These results suggested that the nature of a behavioral task and its demands are key factors in determining the responses of individual cells in the hippocampal formation, with neurons varying their firing characteristics to represent either a local or remote location according to task demands.

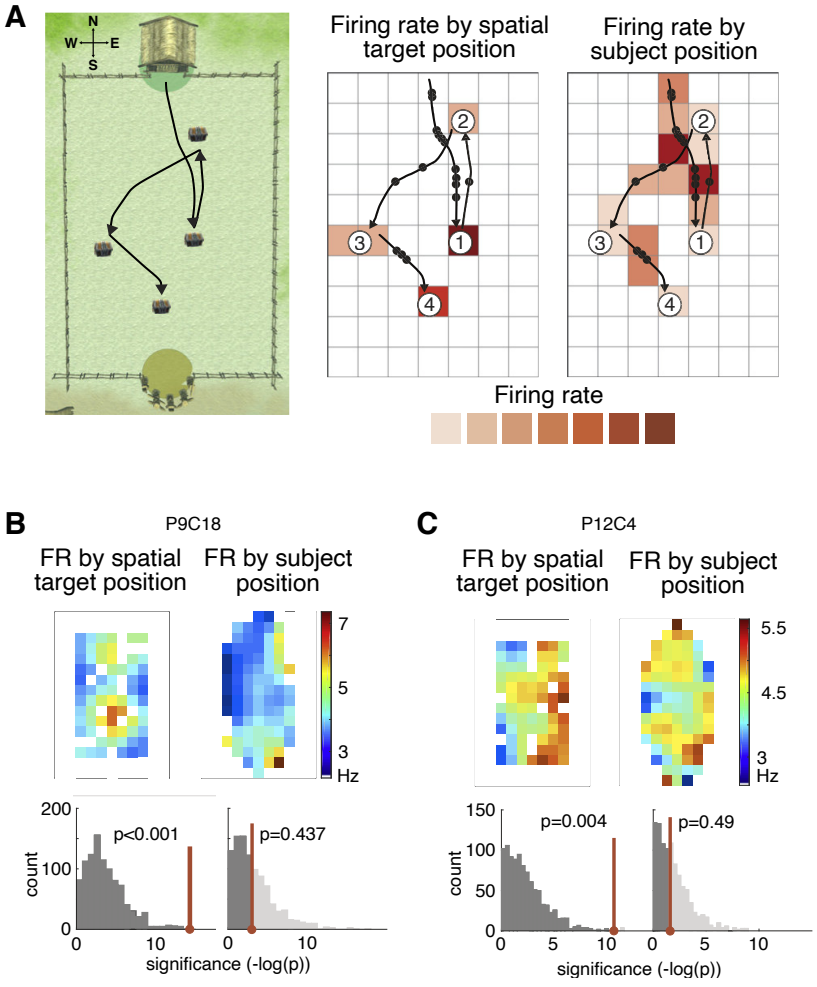


Figure 6.4: Neuron that represents remote spatial targets from Tsitsiklis et al. 2020

For a human neural representation of space to be relevant to support behavior, it might be useful for the particular locations represented by each cell to vary according to task demands. A recent study found evidence for

exactly this kind of pattern in humans by showing that spatially modulated neurons in the human entorhinal cortex changed the spatial tuning of their firing according to a subject's memory state (Qasim et al., 2019). In this study, subjects performed a memory task where on each trial they were first given the name of a target object and then moved through a virtual environment. The environment contained four invisible target objects, which were located at different hidden unmarked locations. Subjects were instructed to press a button when they were located at the position of the cue object.

Thus, on each trial of the task, the subject was focused on a different object and hidden location. To test whether this differing focus modulates neuronal firing patterns, the firing rates of individual neurons were then measured as a function of the subject's location in the environment as before (Fig. **??**) with the exception that each cue condition was measured separately. By performing this procedure, in addition to identifying place cells, which responded at fixed spatial locations as in earlier studies, this study identified a new phenomenon called "memory trace" cells (Fig. **??**). These cells shifted the locations of their firing fields between trials of the task where subjects were cued on different remembered locations. This result is important because it shows that the hippocampal representation of an environment is not a static representation but can be transformed according to memory or other behavioral demands. This study suggests that an important area of future research is to identify the factors that cause human neurons to change their spatial firing patterns and to distinguish these factors from signals seen in simpler animals like rodents.

Time cells

In addition to single-neuron correlates of space and item information, as described above, a final aspect of episodic memory is the ability to associate events with particular moments in one's day or life. By examining recordings of single-neuron activity from subjects performing memory tasks, these studies tested for neural correlates of time by comparing how the firing rates of individual neurons changed throughout the course of memory encoding intervals. Using the spatial memory paradigm described above (Figure **??**), Tsitsiklis et al. (2020) identified neurons whose firing rates varied as a function of the order of the item in a list they were trying to learn. Individual hippocampal neurons showed increased firing rates when the subjects learned items at particular item positions (Fig. **??**a,b). Across cells, there was the greatest representation of items at the beginning of each list (Fig. **??**c), which may relate to behavioral findings concerning "primacy" items by showing that they have stronger and more distinctive neuronal representations.

A related set of findings comes from a study of single-neuron firing during the encoding phase of the free recall task (Umbach et al., 2020). Here, the firing rates of individual cells were measured as a function of time during encoding. The results demonstrated the existence of "time cells" in the human medial temporal lobe, each of which activated at a particular time point during the encoding interval (Fig. **??**d). Individual time cells activate at different moments during these encoding intervals (although, again, there is an over-representation of the beginning of the interval), which suggests that the neural representation from the population of time cells could provide a

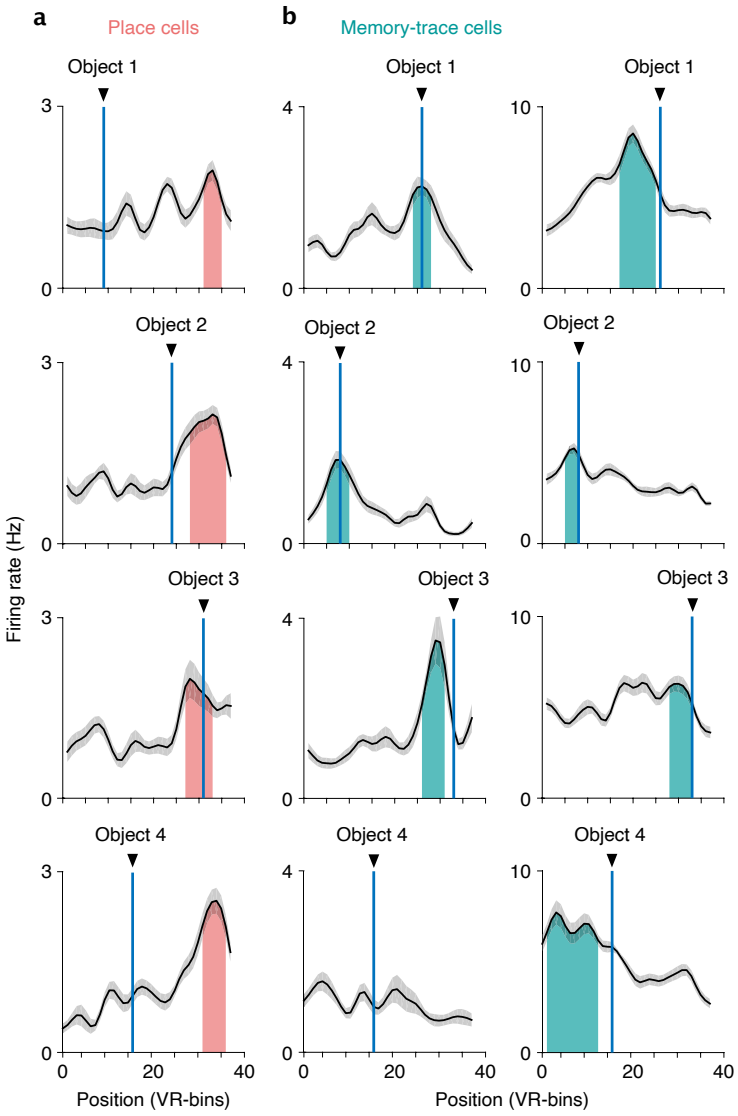


Figure 6.5: Neuron that represents remembered object locations from Qasim et al. 2019

)

distinctive neural pattern that uniquely differentiates separate moments to support memory encoding and retrieval.

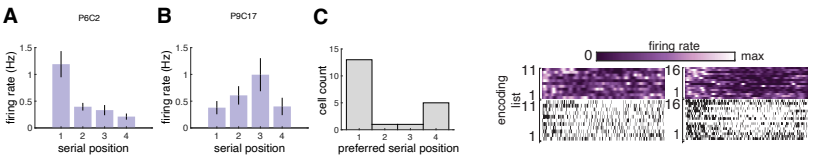


Figure 6.6: Neurons whose firing rates represent time in the human MTL. (A) Neurons that represent specific moments during list learning (Tsitsiklis et al 2020). (B) neurons that activate to represent times during list learning (Umbach et al., 2020)

Conclusion

In summary, single-neuron recordings from the human medial temporal lobe show that the firing rates of individual cells during behavior correspond to an array of task-relevant variables. In many cases, these findings replicate findings in animals, as well as sometimes extending those results. Human single-neuron firing patterns often show highly abstract representations during complex behaviors, which is one way that the single-neuron firing patterns in the human hippocampus differ from those in other sensory- and motor-related regions. Notably, the neural representations in the hippocampus contain some of the key building blocks of episodic memories, including representations of the semantic content related to a new cognitive state, elapsed time, and location in space. Thus, these single-neuron firing patterns could be instrumental in allowing our brains to form memories. Future research on neuronal firing patterns has the potential for identifying new types of neuronal activity related to cognition, by having human subjects perform richer and ever more complex behavioral tasks.

Recordings of in the human basal ganglia

Whereas the preceding sections focused on memory- and navigation-related neuronal responses in the human medial-temporal lobe (MTL), a large scientific literature has also implicated some of the brain's deepest structures, known as the basal ganglia, in various forms of learning and memory. Here we examine the responses of neurons in these structures during a collection of tasks in which people learn probabilistic associations through repeated trials with feedback. The probabilistic nature of the tasks makes learning the associations very difficult without experimenter feedback, or "reinforcement". Hence we will refer to these as reinforcement learning tasks.

Assuming that learning new information entails some cost to the organism, one would expect a brain that is optimized for efficient learning to encode more on trials that provide new information than on trials with highly predictable outcomes. In line with this basic principle of efficient learning, studies in non-human primates observed that neurons in a basal ganglia subregion known as the *substantia nigra* (SN) fill in the story here. In these studies describe studies. explain SN dopamine connection.

Zaghloul et al (2009) asked whether the human SN similarly encoded

the expectation of a reward. They measured intra-operative activity of SN neurons using microelectrode recordings in ten PD patients undergoing DBS surgery of the STN while they engaged in a probability learning task. Participants repeatedly drew (virtual) cards from one of two decks. Cards from one deck yielded positive feedback 65% of the time, whereas cards from the other deck were rewarded in only 35% of the trials. The experimenters instructed subjects that the decks differed in their probability of yielding a reward and asked them to adjust their choices to maximize reward (Zaghloul et al., 2009).

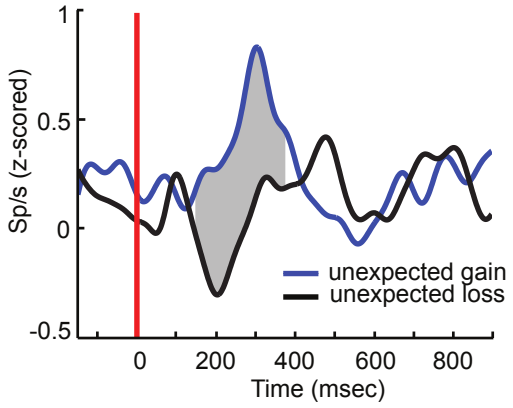


Figure 6.7: *Normalized firing rates for unexpected gains and losses.* Red line indicates feedback onset. The gray region marks the 225 ms interval between 150 and 375 ms after feedback onset. Traces represent activity from 15 SN cells recorded from ten participants (Zaghloul et al., 2009).

Zaghloul et al sought to measure the activity of SN dopaminergic neurons in response to unexpected rewards and the unexpected absence of rewards. They classified feedback as expected and unexpected based on a model that estimated the expected reward from a given deck as a function of reward history (Zaghloul et al., 2009). Specifically, they assumed that the effect of past reinforcements decreases with time based on a power function, used in previous quantitative models to describe decay in episodic memory (cf., Rubin & Wenzel, 1996; Wixted & Ebbesen, 1991). [1] Zaghloul et al used this model of expected reward to classify the feedback associated with each trial as unexpected gain, unexpected loss, expected gain, or expected loss.

To determine how SN neurons encode behavioral feedback across participants, they examined pooled activity for dopaminergic neurons in response to unexpected gains and losses (Zaghloul et al., 2009). [2] During the period between 150 and 375 ms after feedback onset, spike rates in response to unexpected gains were significantly greater than spike rates in response to unexpected losses ($p < 0.001$; Figure **??**). By responding to unexpected rewards, these putatively dopaminergic cells encode new information that likely helps participants maximize reward in the probability learning task.

Decision conflict and the sub-thalamic nucleus

An organisms survival depends on its ability to select actions that maximize value among competing alternatives. Researchers have extensively studied the neural basis of action selection in the context of sensorimotor tasks where an animal must choose between two competing stimuli. These studies reveal that some cortical neurons gradually increase their firing rates when integrating sensory evidence, signaling action selection once a threshold is exceeded.

[1] According to their model, the expected reward, $E_d[n]$, for a particular deck, d, on the nth trial was defined as:

$$E_d[n] = 0.5 + 0.5 \sum_{i=1}^{n-1} R_d[n-i] \alpha i^{-\tau} \qquad n = 2, \ldots, N \,, \tag{6.1}$$

with $E_d[1] = 0.5$ and $R_d[n]$ coding the feedback for choosing deck d on the nth trial (out of a total of N trials) as one for positive feedback and negative one for negative feedback ($R_d[n] = 0$ for trials when deck d was not selected). Finally, τ is a parameter determining how quickly the power function falls off and α is set such that the weights of the power function approximate one over infinite trials for a given τ). Based on the best fitting τs, participants selected the deck with the higher expected reward on 74.9% of the trials.

[2] They defined dopaminergic neurons [SAY HOW]

However, it is unclear how the brain adjusts firing rate thresholds to optimize action selection. The basal ganglia, with their widespread and direct connections to the cortex, may help resolve these issues. This central structure can theoretically adjust threshold criteria and efficiently convey information between separate cortical regions.

A large body of research has implicated the subthalamic nucleus (STN) in the encoding of decision conflict, which arises when an individual encounters competing response options or when determining the optimal choice is uncertain. This process is vital for assessing the potential outcomes of various actions and adjusting response strategies accordingly. To address this issue,

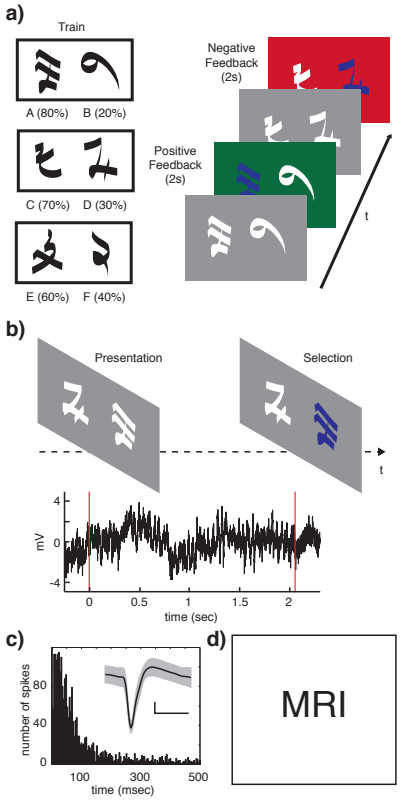


Figure 6.8: *Probability learning and decision task.* **(a)** Probability learning task. Three pairs of symbols (AB, CD, EF) appear in random order while microelectrodes recordi signals in the substantia nigra. Subjects must choose one of the two stimuli on each trial. The experimenters assign a random reward rate to each symbol, which remains fixed throughout the experiment. Probabilistic feedback follows each choice for 2 seconds, with a green screen and cash register sound indicating positive feedback and a red screen with an audible buzz indicating negative feedback. **(b)** In the decision task, researchers present participants with combinations of all symbols and instruct them to choose one of the two symbols presented on each trial. After each selection, the screen turns grey, and participants do not receive any corrective feedback. Microelectrodes capture STN activity during the decision task. The bottom panel represents a typical STN recording during a single trial, with red lines indicating stimulus presentation and button selection. **(c)** Histogram of interspike intervals from one spike cluster. (*Inset*) he black line represents the mean waveform of this spike cluster, while the grey area shows the standard deviation. The scale bar represents 10 mV and 0.5 ms. Scale bar represents 10 mV and 0.5 ms.

Zaghloul et al. (2012) addressed this issue by recording individual neurons in the human STN during a learning task designed to induce decision conflict. They employed a task involving a training and testing phase with multiple probability comparisons (L. M. Frank, Stanley, & Brown, 2004; M. Frank, Samanta, Moustafa, & Sherman, 2007). During the training phase, subjects viewed three pairs of symbols (AB, CD, EF) in random order, and participants were instructed to choose one of the two stimuli on each trial (Fig. **??**A). Subjects made selections by pressing the button of a handheld controller with either the left or right hand. The three stimulus pairs were characterized by different relative rates of reward (80% vs 20%; 70% vs 30%; 60% vs 40%). Reward rates associated with each symbol were determined

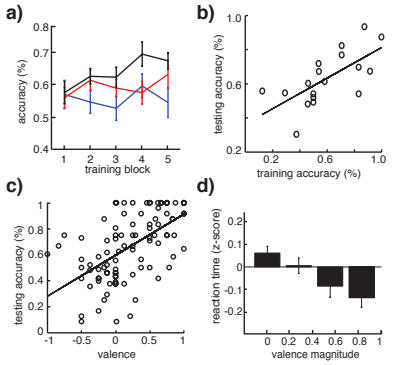


Figure 6.9: *Behavioral performance.* **(a)** Learning rates are quantified by dividing the total number of training trials into five equally sized blocks and determining how often participants correctly choose the symbol with the higher *a priori* reward rate for each block. Black, red, and blue traces represent mean accuracy across all participants for the AB, CD, and EF symbol pairs respectively. Error bars represent *SE*. **(b)** Each point represents the overall accuracy during the final block of training mapped to the overall accuracy throughout testing for each of the 18 experiments. Black line demonstrates the best fit of a linear regression. **(c)** Each point represents the valence associated with each symbol pair presented during testing mapped to how often the participant chose the symbol with the higher *a priori* reward rate during testing for all pairs for all 18 experiments. Valence for a symbol pair is determined by the difference in reward expectation for each symbol during the final block of training (see Methods). Per our convention, positive valence for a symbol pair suggests that the participant ascribes greater reward expectation to the symbol with the higher *a priori* reward rate. Black line demonstrates the best fit of a linear regression. **(d)** Bars represent the mean z-scored response times associated with one of four equally populated bins of valence magnitude. Data is pooled across all trials across all participants for illustrative purposes. Error bars represent *SE*, respectively. Red line indicates stimulus onset. Bars represent mean z-scored spike response averaged across all trials across all experiments. Error bars represent *SE*. **(d)** Average spike activity between 100 and 400 ms after stimulus onset for high, medium, and low decision conflict. Bars represent average of pooled z-scored spike responses for this period, averaged across all trials across all experiments. Black bars represent pooled data divided into levels of decision conflict based on valence (see text). Striped bars represent pooled data, divided into levels of decision conflict based on response time. Error bars represent *SE*.

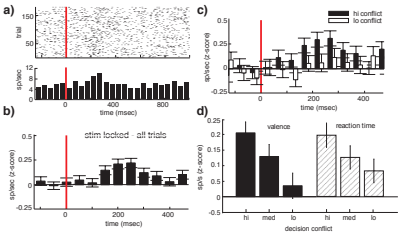
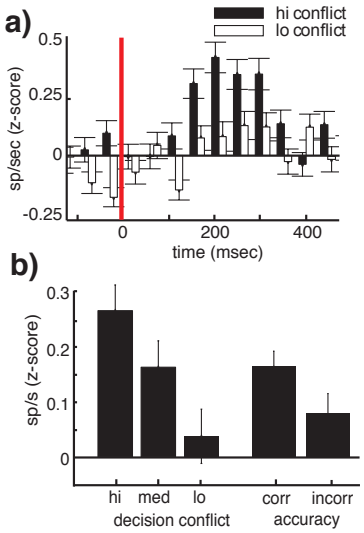


Figure 6.10: *Spike activity during the decision task.* **(a)** Raster plot (top) and histogram (bottom) captured from a single STN cluster during the decision task. Red line indicates stimulus onset. Raster plot demonstrates spike activity during all 180 trials of the decision task. Histogram presents average spike activity (z-scored) across all recorded clusters in response to all decision trials. Red line indicates stimulus onset. Spike responses are z-scored for each trial and averaged across all trials across all experiments. **(b)** Pooled spike activity across all recorded clusters in response to all high conflict decision trials during accurate choices (black line) and all low conflict decision trials during accurate choices (red line, respectively). Red line indicates stimulus onset. Bars represent mean z-scored spike responses averaged across all trials across all experiments. Error bars represent *SE*. **(c)** Average spike activity between 100 and 400 ms after stimulus onset for high, medium, and low decision conflict during accurate trials (left). Average spike activity between 100 and 400 ms after stimulus onset for correct and incorrect trials (right). Bars represent average of pooled z-scored spike responses for this period, averaged across all trials across all experiments. Error bars represent *SE*.



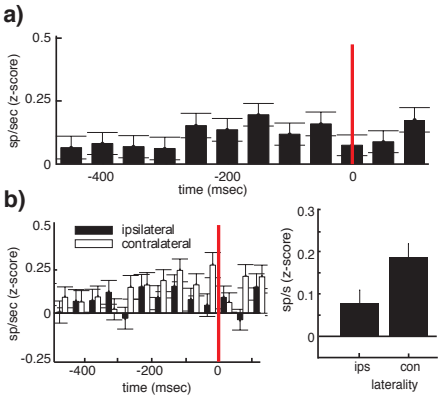


Figure 6.12: *Spike activity during accurate trials.* **(a)** Pooled spike activity across all recorded clusters in response to all decision trials time locked to button selection. Red line indicates the time participants made a choice with a button press. Spike responses are *z*-scored for each trial and averaged across all trials across all experiments. Error bars represent *SE*. **(b)** Pooled spike activity across all recorded clusters in response to all trials associated with ipsilateral and contralateral button presses (black and white bars, respectively; left). Red line indicates the time participants made a choice with a button press. Average spike activity between 0 and 300 ms before button press for trials associated with contralateral and ipsilateral button presses (right). Bars represent mean *z*-scored spike responses averaged across all trials across all experiments. Error bars represent *SE*.

randomly and fixed throughout the experiment. Probabilistic feedback followed each choice. In the event of positive feedback, the selection screen turned green, and an audible ring of a cash register was presented. In the event of negative feedback, the selection screen turned red, and an audible buzz was presented. Each trial consisted of presentation of the stimuli, participant choice, and a two second display of feedback. Over the course of the training phase of the experiment, participants learn the underlying probability structure.

In a subsequent testing phase of the experiment (Fig. **??**B), participants were presented with combinations of all symbols, including novel combinations, and instructed to choose one of the two symbols presented on each trial. Each pair was presented up to 12 times in random order. Participants completed on average 153.7 trials (SE 11.7) during the testing portion of the experiment. After each choice, the selection screen turned grey. No feedback indicating whether the choice was correct was presented. During the training phase of the task, participants developed a hierarchy of reward expectations associated with each symbol, reflected in how often they select that symbol at the end of training (see Methods). Depending on the relative expectations between each symbol, each trial during the testing phase of the task can be characterized by a level of decision conflict. Low decision conflict occurs when comparing a symbol with relatively high reward expectation to a symbol with low reward expectation, whereas high decision conflict occurs when comparing two symbols with similar levels of reward expectation.

Over the course of the training portion of the task, participants exhibited a significant improvement in accuracy for the AB symbol pairs between the first and final blocks of training (Fig. **??**A; t(34)=2.11, $p = .04$)). The smaller differences in *a priori* reward rates for the CD and EF symbol pair led to smaller, not statistically significant, improvements in accuracy (t(34)=1.08, p= .14; t(34)=−0.29, $p = .61$). To assess to what extent performance in the training phase predicted performance in the testing phase (reflecting a use of the contingencies learned during the training phase in the subsequent testing phase), we regressed accuracy during the testing phase on accuracy during the training phase (across all symbol pairs) for all participants and found a significant relationship between training and testing accuracy ($b = 0.45$, $t(16) = 3.89$, $p = .0013$, $r^2 = 0.49$).

We investigated the extent to which participants used their assessment of reward expectation to inform their decisions during the testing phase (Fig. **??**C). We define valence for a symbol pair as the difference in a participant's reward expectations for the two symbols, as derived from the choice responses in the last training block (see Methods). As per our convention, positive (negative) valence for a symbol pair corresponds to the the participant ascribing greater (lower) reward expectation to the symbol with the higher *a priori* reward rate. Hence, if a symbol pair has a valence of 1 (-1), the participant should always select the symbol with the higher (lower) *a priori* reward rate during presentations of this pair. Linear regression demonstrated a significant relationship between valence and testing accuracy ($b = 0.32$, $t(99) = 7.10$, $p < .001$, $r^2 = 0.34$) when combining behavioral data from all participants for all combinations of symbol pairs.

To confirm that the magnitude of the valence reflects decision conflict, we measured the response time associated with every trial in the decision task. Since response time should increase with decision conflict, we expected a negative relationship between response time and valence magnitude. The mean response time across all experiments between the presentation of stimuli and button press was 1660 ms (SE 174 ms). We averaged z-scored response times for all trials associated with one of four equally spaced bins of valence magnitude across participants (Fig. **??**D). Response time was inversely related to valence magnitude ($p < .001$, permutation test, see Methods), suggesting that valence magnitude is an appropriate surrogate for decision conflict.

We extracted and sorted single-unit activity captured from STN microelectrode recordings to find 38 uniquely identified spike clusters (2.1 clusters per recording (SE 0.29)). We excluded spike clusters with median firing rates below 1 Hz from our analysis, under the assumption that they were over-splitting artifacts. We thus retained 27 spike clusters (1.5 spike clusters per recording (SE 0.31)). Average recorded waveforms and a histogram of interspike intervals from one spike cluster are shown in Fig. **??**C.

Representative spike activity recorded from a single STN cluster in a single participant during the decision task, time locked to the presentation of symbol pairs, is shown in Fig.**??**A (see Supplementary Fig. 1 for a second representative example). There is a clear increase in spike activity around 200 ms following the presentation of symbol pairs (t=0), suggesting increased STN activity associated with decision processes. Spike activity across all recorded spike clusters confirmed this response (Fig. **??**B). Following the presentation of symbol pairs, as participants initiated the decision process, there was a consistent and significant increase in spike activity between 100 and 400 ms compared to baseline activity (t(26)=2.04, $p = .052$, t-test across clusters; $p < .001$,permutation test; see Methods).

To determine whether the increase in spike activity was related to the level of decision conflict, we divided valence magnitude into three equally spaced bins for each participant and pooled spike activity for each bin. Across participant average pooled histograms for the smallest and largest valence magnitudes, corresponding to the highest and lowest decision conflict respectively, are shown in Fig. **??**C. There is an increase in spike activity between 100 and 400 ms after symbol pair presentation, but this increase is greater for trials associated with higher decision conflict. To confirm this difference, we averaged spike activity between 100 and 400 ms and found

a consistent and statistically significant difference between spike activity associated with high and low decision conflict, as determined by valence magnitude (t(26)=2.36, $p = .026$; $p = .008$, permutation test; Fig.**??**D). Spike activity associated with high decision conflict was also significantly greater than activity associated with medium decision conflict (t(26)=2.71, $p = .013$; $p = .005$, permutation test), although spike activity associated with medium decision conflict was not significantly different than that associated with low decision conflict (t(26)=-0.90, $p = .811$; $p = .057$, permutation test).

We examined whether this difference in spiking activity between levels of decision conflict was modulated by the accuracy of choice. We identified trials when participants chose symbols with the higher *a priori* probability as accurate choices. Overall, spike rates were not significantly higher during accurate choices than inaccurate choices during the period between 100 and 400 ms after stimulus (t(26)=.136, $p = .45$; $p = .44$, permutation test; Fig.**??**B). During accurate choices, however, spike activity was significantly modulated by decision conflict during the period 100 to 400 ms after stimulus (Fig. **??**A, B). Spike activity associated with high conflict trials was significantly greater than low and medium conflict trials during correct choices (t(26)=3.46, $p = .002$ and t(26)=2.39, $p = .026$ respectively; $p < .001$ and $p < .001$ respectively, permutation test). These effects were consistent when we defined accurate trials as those trials where participants chose symbols with the higher reward expectation as determined by training). Conversely, during incorrect choices, there was not a significant difference in spiking activity between high and low decision conflict trials (t(26)=-0.39, $p = .65$; $p = .38$, permutation test).

To examine how STN activity is modulated at the time of decision, we investigated spiking rates for all cells time locked to the moment of selection, when participants pressed the left or right button, for each trial (Fig. **??**A). During the 300 ms window immediately preceding the moment of selection, there remained a significantly higher level of spike activity compared to baseline (t(26)=2.31, $p = .028$; $p < .001$, permutation test). This activity rapidly decreased as participants made their selection.

To examine the role of STN spike activity in modulating ipsilateral and contralateral motor function, we examined the relationship between spiking rates and the laterality of selection.Trials recorded from the right STN were designated ipsilateral when the right button was selected, and contralateral when the left button was selected. During the 300 ms window immediately preceding the moment of selection, there was greater spike activity during trials involving contralateral button presses compared to trials involving ipsilateral selections (t(26)=2.45, $p = .021$; $p = .019$, permutation test; Fig. **??**B).

7
Geometric Similarity

This chapter uses insights from Euclidean geometry to draw conclusions about cognitive states revealed through patterns of neural activity. The general framework considers cognitive states as being instantiated in patterns of activity; we assume that patterns of neural activity, measured using multi-channel electrophysiology, reflect these cognitive processes and the representations they act upon.

Long before the dawn of cognitive neuroscience, scholars asked whether the neural activity underlying particular cognitive functions localized to specific brain regions or distributed across widespread neural networks.[1] In support of the idea that cognitive processes do not localize to specific brain regions, Karl Spencer Lashley found that the performance of rats navigating mazes deteriorated as a function of the extent of cortical lesions across a wide range of lesion targets (?, ?).[2] More recent evidence supports the notion that much of cognitive function relies on patterns of highly distributed neural activity. Indeed, in what has become a classic study, Haxby et al. (2001) showed that neural activity reliably distinguished between different visual categories even when neural activity from regions that responded maximally to those categories was excluded. Likewise, neural activity in regions that responded maximally to one category also reliably distinguished between other categories.

If the features of neural activity supporting specific aspects of cognition appear to be distributed across widespread brain regions, we can ask how the similarity of these patterns relate to aspects of behavior. The geometric similarity methods introduced in this chapter relate similarity structures in patterns of neural activity to external similarity structures (e.g., similarities between experimental stimuli or responses). By thinking of neural recordings from many sites as reflecting a pattern in N-dimensional space (where N is the number of features), we can measure the distance (or, inversely, the similarity) between patterns of neural recordings collected at different times, and use such measurements to derive insight into the cognitive processes associated with engagement in experimental tasks. These methods have a substantial history in the literature on human neuroscience using functional magnetic resonance imaging (Haxby et al., 2001; Kriegeskorte et al., 2008) and have also become popular in analyzing human electrophysiological data.

[1] Propagating a highly localized view of brain function, Descartes implicated the Pineal Gland as the seat of the soul and as responsible for a wide range of cognitive function, including memory: "Descartes' mechanical explanation of memory was as follows. The pores or gaps lying between the tiny fibers of the substance of the brain may become wider as a result of the flow of animal spirits through them. This changes the pattern in which the spirits will later flow through the brain and in this way figures may be 'preserved in such a way that the ideas which were previously on the gland can be formed again long afterwards without requiring the presence of the objects to which they correspond. And this is what memory consists in' (AT XI:177, CSM I:107)." (?, ?).

[2] Lashley proposed that cortical tissue could take over functionality from damaged areas — a feature he termed "equipotentiality". He qualified this potential for neural plasticity by proposing the "law of mass action whereby the efficiency of performance of an entire complex function may be reduced in proportion to the extent of brain injury within an area whose parts are not more specialized for one component of the function than for another" (?, ?, p. 25).

Representational and Neural Similarity

Geometric similarity serves as a fundamental concept in models of human perception and cognition. As you watch the suitcases come along on the conveyer belt, you think you see your bag; but just as you are about to pick it up, you realize it belongs to someone else. Just as similarity determines perceptual discriminability it also determines the interference among items in memory and the ability to generalize among memories. Chapter 1 reviewed models of recognition memory based on the idea that the brain computes the sum of the similarities between a test item and the items stored in memory. If this summed similarity exceeds a threshold, subjects endorse the test item. Models of recall also assume similarities determine the organization of recall and the nature of intrusion errors.

Cognitive models make specific assumptions about similarity that guide the behavioral predictions of the model. But overt behavior only offers intermittent snapshots of the activity of the mind. Neural recordings, however, can provide valuable information about internal cognitive states that do not assert themselves in overt behavioral measures. Here we discuss methods scientists have used to study the neural basis of memory and cognition by measuring neural similarity—i.e., the similarity of neural data measured at two given time points.

Consider our example of identifying whether a suitcase seen on the conveyor belt matches your memory of your own suitcase. As you fixate your eyes on a target suitcase, you presumably represent the visual image in some multidimensional representation that includes information such as the size, color, and shape of the suitcase. You can also retrieve your memory of the same features that characterize your own suitcase. Now you can compute the similarity between the percept and the features retrieved from memory and thereby decide whether to pick up the suitcase.

Let us assume, for the sake of argument, that for each feature value $f_i(t)$ in some perceptual/memory space, F, there exists a neural measurement $b_i(t)$ that perfectly correlates with the feature value $f_i(t)$. Now even if this is true, there is no guarantee that any of your neural sensors, which record signals $\tilde{b}_i(t)$, will be measuring $b_i(t)$. Yet, the ability to reliably decode information about perceptual features on F from neural recordings on B suggests that these recordings possess some statistical relationship to the hypothesized features. Suppose, then, that the observable state of the brain at time t, as measured across the deployed sensors, $\tilde{B}(t)$, has some predictive relationship with the representations assumed under a cognitive model. If our cognitive model predicts that $F(t_j)$ should more closely resemble $F(t_k)$ than $F(t_r)$, then we can ask whether the same similarity ranking holds on $\tilde{B}$.

We, therefore, move from analyzing one feature at a time to analyzing N features at a time, where N could be the number of channels, the number of frequencies $\times$ channels, or any other set of features. The pattern of neural activity across N features constitutes a vector in the N-dimensional feature space. Consider the example of a set of recordings of spectral power in the γ frequency band at N channels across M observations. The set of M observations constitutes a matrix where each column contains the pattern of γ power across channels. A typical univariate approach to analyzing such data might be to use parametric statistics to aggregate the power values within each electrode (across M columns) and compare power between the N

electrodes.

In contrast, we could instead consider the set of N power values as a pattern that reflects the state of the participant's brain at each timepoint, $\mathbf{P}_t$. The N values that comprise $\mathbf{P}$ at each of M timepoints can be thought of as vectors corresponding to points in N dimensional space. By thinking of the patterns as vectors, we can then begin to consider distances between vectors as reflecting how similar or different the brain is at different points in time, and what such similarity (or lack thereof) tells us about the cognitive operations the brain is performing.

When applying geometric similarity methods to neural data, the core assumption is that quantifying the "proximity" between observations reveals how the brain processes or represents information. Several methods can be used to measure "proximity" between neural patterns, which we review below.

Distance-based methods

Euclidean distance. Perhaps the simplest way to measure the proximity between two points is to calculate the Euclidean distance between them. Given two observations A and B defined in N-dimensional space, the Euclidean distance between A and B is given by

$$D = \sqrt{\sum_{i=1}^{N} (A_i - B_i)^2} \tag{7.1}$$

Fig. 7.1A illustrates an example using three observations in a two-dimensional space. The Euclidean distance from A to B and A to C is $\sqrt{8} \approx 2.828$ for both pairs of points. Our goal in computing the distance between the observations is to characterize their similarity, e.g., the similarity of reference point A to observations, B and C. Using Euclidean distance, the points B and C are equally similar (or dissimilar) to A. Here, we have defined A, B, C in an arbitrary two-dimensional space. One could imagine a scenario in which Measure 1 and Measure 2 reflect a real-world measurement, for instance, power in the 3-8 Hz theta and > 40 Hz gamma frequency bands as measured from an electrophysiological recording.

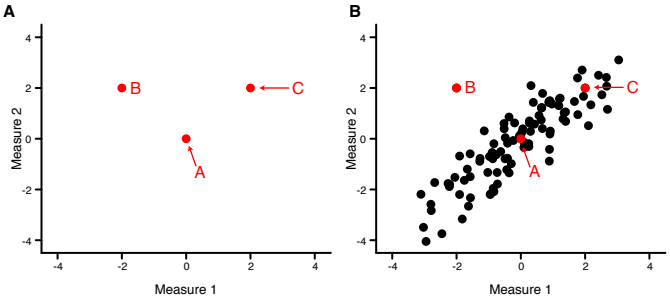


Figure 7.1: A. An example showing three points in a two-dimensional space. **B**. The same three points appear embedded in a distribution of 100 observations.

Mahalanobis distance. We typically want to make comparisons among a sample of observations derived from some underlying distribution. Fig. 7.1B shows a toy example of 100 observations drawn from a hypothetical dataset.

The points A, B, C appear overlaid in the same locations as in Panel A. Here Measures 1 and 2 exhibit a strong positive correlation.

When we used Euclidean distance to calculate similarity, B and C appeared equidistant from A leading to the conclusion that they are equally similar to A. However, now that we have many more observations, it is clear that there is structure in our dataset that we are missing when using Euclidean distance to measure similarity. Assume that the observations in Fig. 7.1B reflect the underlying distribution of the data. In this case, collecting many more observations would reveal that observations near C occur more often than those near B.

The Mahalanobis metric measures distance between multivariate points while accounting for the correlational structure of the data. The idea is to account for situations like that in Fig. 7.1B, in which we would like to calculate the (dis)similarity between observations using a multivariate distance metric, but we would like to use the overall distribution of the data to scale our estimate of the distance of each observation. The Mahalanobis distance normalizes the Euclidean distances between points by the covariance matrix $\mathbf{S}$ of the data:

$$D_M = \sqrt{(\mathbf{a} - \mathbf{b})^T \mathbf{S}^{-1} (\mathbf{a} - \mathbf{b})} \tag{7.2}$$

Calculating the Mahalanobis distance between our pairs of points yields $D_M(A, B) = 4.786$ and $D_M(A, C) = 1.376$; we can see that, unlike Euclidean distance, the Mahalanobis distance is larger between A and B and smaller between A and C, reflecting the fact that B is a more unusual observation than is C, given the distribution of the data.

Angle- and correlation-based methods

Cosine similarity. Another way to measure the similarity between two vectors is to calculate the angle between them. We can define the *cosine similarity* between two vectors $\mathbf{A}$ and $\mathbf{B}$ as follows:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{7.3}$$

where $\mathbf{A} \cdot \mathbf{B}$ is the *dot product* between $\mathbf{A}$ and $\mathbf{B}$ and $\|\mathbf{A}\|$ is the magnitude of $\mathbf{A}$. [3]

[3] The dot product of two vectors is given by $\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{N} A_i B_i$. The magnitude of a vector is given by $\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{N} A_i^2}$

Context-based theories of memory propose that a slowly drifting representation of recent experience is used by the memory system as a glue to bind new experiences into coherent memory traces. One prediction of such a theory is that the similarity between neural states should gradually decrease as time elapses. If one were interested in testing such a prediction using human electrophysiological recordings, such a question suits geometric similarity methods.

By considering the pattern of neural activity across channels as reflecting a point in high-dimensional space, Manning et al. (2011) calculated the change in pairwise neural similarity between items during memory encoding as a function of inter-item distance. Using cosine similarity to compare such patterns, Manning et al. (2011) showed that inter-item similarity falls off as a function of the lag between pairs of items. This result shows that as more time elapses between the encoding of two items, the similarity between the

neural patterns reflecting encoding of each item goes down; alternatively, the distance between the points in N-dimensional space increases (Fig. 7.2A).

Manning et al. further used cosine similarity to measure similarity between neural patterns during *recall* of items and the patterns of items that had flanked the recalled item previously during *encoding*. As predicted by context theories of memory, the pattern associated with the recalled item showed the highest similarity to the encoding patterns for items that had immediately followed and preceded it at encoding; similarity for items at greater forward and backward lags decreased as a function of lag (Fig. 7.2B).

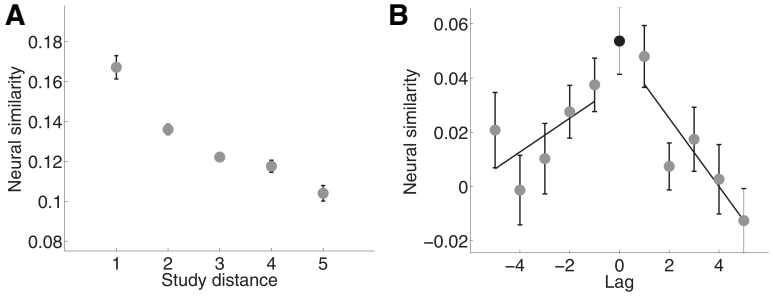


Figure 7.2: An example of applying cosine similarity to intracranial EEG data during a free recall task. **A:** Similarity between neural encoding patterns decreases as a function of the lag between the items, a key prediction of retrieved-context theory. **B:** Similarity decreases with lag when computed between the pattern evoked by recall of item i and encoding patterns for items that preceded and followed i. From Manning et al. (2011).

Manning et al. (2011) aimed to characterize how similarity between patterns varied as a function of inter-item distance, to test the predictions of context-based accounts of episodic memory. Other work has used Pearson correlation as a similarity measure to test predictions of reactivation-based models of memory encoding and consolidation. In rodents, there is a large body of evidence suggesting that recent experiences are replayed or reactivated to facilitate consolidation processes. Such studies use sequential firing of hippocampal place cells that are active when an animal traverses an environment. In humans, reactivation has been difficult to measure directly because of the difficulty in identifying individual features of neural activity that are as tightly coupled to the external environment as rodent place cells.

One way to circumvent this limitation is to use geometric similarity methods to compare multivariate patterns of neural activity. By comparing the similarity between electrophysiological patterns before and after event boundaries, one study showed evidence in humans for reactivation (?, ?). The authors reasoned that if reactivation of a recent experience occurs at event boundaries, then the similarity between neural patterns before and after a boundary should be higher than in a control condition in which the inter-pattern lag was matched but an event boundary did not occur.

This study illustrates another rationale for using similarity measures in lieu of a classification approach. In this case, the prediction is that the information that is reactivated at each event boundary should be unique to what was recently experienced. By computing a similarity measure for each trial and then aggregating the similarity values across trials, if individual similarity values are consistently higher for one condition than another, this should be evident in the aggregate measure.

Correlation. Another angle-based measure of similarity that is closely related to cosine similarity is correlation. In the case of mean-zero variables, the cosine and correlation measures become computationally identical.

Statistical considerations. Geometric similarity methods produce measurements that are not normally distributed, making them unsuitable for direct analysis using standard parametric methods (e.g. *t*-test and ANOVA). Computing the correlation or cosine similarity between patterns, for example, produces values in the range [-1,1]. To address this issue researchers use transformations, such as Fisher's *r*-to-*z* transformation, to produce dependent measures that do not violate the assumption of parametric statistics. Given a sample correlation ρ computed between two patterns, Fisher's transformation is given by:

$$z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \tag{7.4}$$

This is also known as the inverse hyperbolic tangent transformation. Using this transformation to 'normalize' correlations (or any other similarity/distance measure bounded between [-1,1]), one can treat the transformed observations like other normally distributed data.

Feature Engineering and Dimensionality Reduction

Manning et al's study of context reinstatment, described above and in Figure 7.2, used a technique called principal components analysis to transform a high-dimensional space composed of spectral powers across electrodes into a lower dimensional space of orthogonal components. PCA constructs orthogonal linear combinations of the initial high-dimensional feature space, with each linear component accounting for the maximum residual variance. As is common in applications of PCA, Manning et al used a statistical criterion (Kaiser, 1960) to determine how many principal components to use in their analysis.

Previously we discussed the merits of using the Mahalanobis distance to represent distances in a manner that reflects the correlational structure of the neural features. An advantage of transforming the data using PCA is that each resultant feature is orthogonal to the other features, thus allowing for standard Euclidean or $\cos\theta$ distance measures to capture the similarities among the feature vectors.

Because Manning et al specifically aimed to evaluate predictions of retrieved-context theory, they wanted to evaluate the similarities of the brain's contextual representation. Not knowing which neural features may represent the context of a given memory, they took an additional step often referred to as feature engineering. Rather than simply using the first N principal components (with N determined by the Kaiser criterion), they sought to evaluate similarities among features that possess a key characteristic of context, namely being temporally autocorrelated. As such, they specifically selected those PCA features that exhibited a positive autocorrelation whose consistency across lists met a statistical consistency threshold of $p < 0.10$ (see Manning et al's Supplementary Materials and Methods for details on the calculation of this *p*-value).

Semantic Organization of Memories

Retrieved context theories argue that successful recall results from a search process that reinstates patterns of neural activity present during item en-

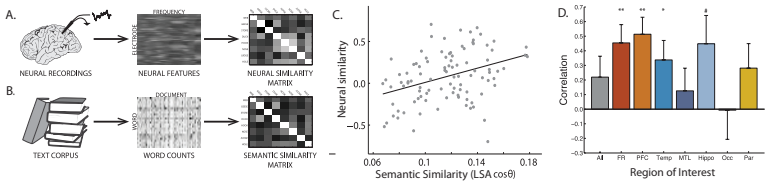


Figure 7.3: Neural similarity predicts semantic clustering. A,B. Illustration of feature vector construction and similarity matrices for neural and text corpus data. Neural recordings are processed into sets of neural features. Each feature is the mean power at a specific frequency, recorded from a single electrode during a particular presentation or recall. Selecting PCA-derived components that predicted LSA semantic similarity at encoding we constructed a neural similarity matrix for all pairs of recalled words. **C**. The correlation between neural and semantic similarity from right matrices in A,B obtained from a representative subject. **D**. Across-subject correlation between semantic clustering during recall and the correlation between neural and semantic similarity matrices. "All" indicates this correlation for all electrodes in the dataset; other bars are for specific ROIs as in Fig 4. # denotes $p < .1$, $*$ denotes $p < .05$, and $**$ denotes $p < .01$.

coding (Kahana, 2020). These models make detailed predictions about the patterns of neural reinstatement that should relate to memory performance. Specifically, these theories argue that recalling of a past event not only recovers features of the event itself but also recovers information associated with other events that occurred nearby in time. The events surrounding a target event, and the thoughts they evoke, may be considered to represent a context for the target event, helping to distinguish that event from similar events experienced at different times. The ability to reinstate this contextual information during memory search has been considered a hallmark of episodic, or event-based, memory.

In the above example, we describe the application of geometric similarity methods to the analysis of neural context reinstatement (Manning et al., 2011). Their study compared the oscillatory activity recorded when the ith studied word was recalled to the activity recorded when the word, and neighboring words on the list (serial positions $i + lag$), were studied. To test the context-reinstatement hypothesis they compared the feature vectors associated with each recall event with the feature vectors associated with the neighbors of the recalled word in the study sequence. For each correctly recalled word, they calculated the similarity between the feature vector associated with the recall event and the feature vectors associated with each of the studied items. This allowed them to calculate the correlation between neural similarity and lag, for both positive and negative lags (Fig. **??**A). Consistent with the context-reinstatement hypothesis, both effects were statistically reliable, with neural similarity decreasing with increasing absolute lags in both positive and negative directions This result would not be expected on the basis of the structure of the lists themselves, as each item was randomly drawn from a pool of common nouns. The decrease in neural similarity with lag mirrors the behavioral contiguity effect seen in the order of these participants' recalls Thus, in remembering a list item, people not only revive the pattern of brain activity that occurred during the item's prior encoding, as has been documented (e.g., Gelbard-Sagiv, Mukamel, Harel, Malach, & Fried, 2008), but they also revive the brain activity associated with neighboring items. Yaffe et al. (2014) replicated the Manning et al reinstatement analysis in a paired-associate learning data set.

In verbal memory experiments, semantic associations also serve as a major influence on the organization of recall (Howard & Kahana, 2002b; Howard, Venkatadass, Norman, & Kahana, 2007). To uncover the neural representation of the semantic organization in free recall Manning, Sperling, Sharan, Rosenberg, and Kahana (2012) used geometric similarity methods to identify spectral features during encoding and retrieval that vary systematically with the semantic relatedness of the studied items (as measured by LSA). These features were projected back into the space of neural activity and were used in computing a neural similarity matrix for each participant that in-

dexed the cosine similarity between the neural representations of each item (Fig. 7.3A&B). A second cosine similarity matrix was then constructed for the semantic similarity between items using LSA values. They then computed the correlation between these matrices for each patient (N=46) as an estimate of the degree to which each patient's neural activity patterns reflected the semantic relationships between studied words (Fig. 7.3D). They found that the correlation between the neural and semantic similarity matrices was related to semantic clustering across patients, in particular in electrodes in prefrontal cortex, temporal lobe and hippocampus (Fig. 7.3C).

Gelbard-Sagiv et al. (2008) asked whether reactivation occurs at the level of individual neurons. Recording neuronal responses in neurosurgical patients as they performed a free-recall task, they identified neurons that increased their firing rate when people viewed movie clips featuring famous people, characters, or animals. They found that neurons that responded when viewing a movie clip also responded when people were free recalling the movies they had seen. This effect was observed significantly for neurons in the hippocampus and other MTL regions.

Reactivation of spatial context

Just as the temporal or semantic characteristics can serve as context for a set of studied items, the spatial attributes of an environment can also serve as context for material experienced within that environment (Smith, 1988). To elucidate the joint contributions of temporal and spatial information to memory search, J. F. Miller, Lazarus, Polyn, and Kahana (2013) asked subjects to play the role of a delivery person in a 3D-rendered virtual town (Figure 7.4A,B). In a first phase of this experiment, subjects became familiar with the town layout, as well as the locations of the stores, by navigating to each of the target stores in succession. Subjects then began a series of "Delivery Days," each of which involved delivering a series of 12 objects, one to each of 12 stores. At the end of the delivery day, the screen went blank and the subjects were cued to freely recall all of the delivered objects. During each delivery day, subjects were cued to navigate to a series of randomly chosen, trial-unique, stores, and upon arrival at each store they were informed of the identity of the object they had delivered. After the final delivery day, subjects were asked to freely recall all of stores in the town (this was an additional probe of spatial memory). Miller et al. (2013) observed significant temporal clustering for the recalled items, significant spatial clustering for the recalled stores, and significant spatial clustering in the order of recalled items themselves (Figure 7.4C). This latter effect provides insight into how memories are organized within a spatiotemporal context by demonstrating spatial organization of nonspatial memories embedded in a spatial context.

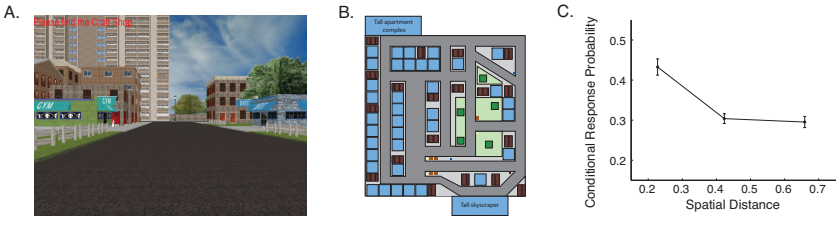


Figure 7.4: Spatial clustering in free recall of objects delivered in a virtual town. A. View a subject may see when navigating through the virtual town. **B.** Overhead map of the town. **C.** Objects that were delivered to nearby locations tend to cluster during recall, as seen in the conditional response probability as a function of spatial distance within the town. Data from Miller et al. (2013).

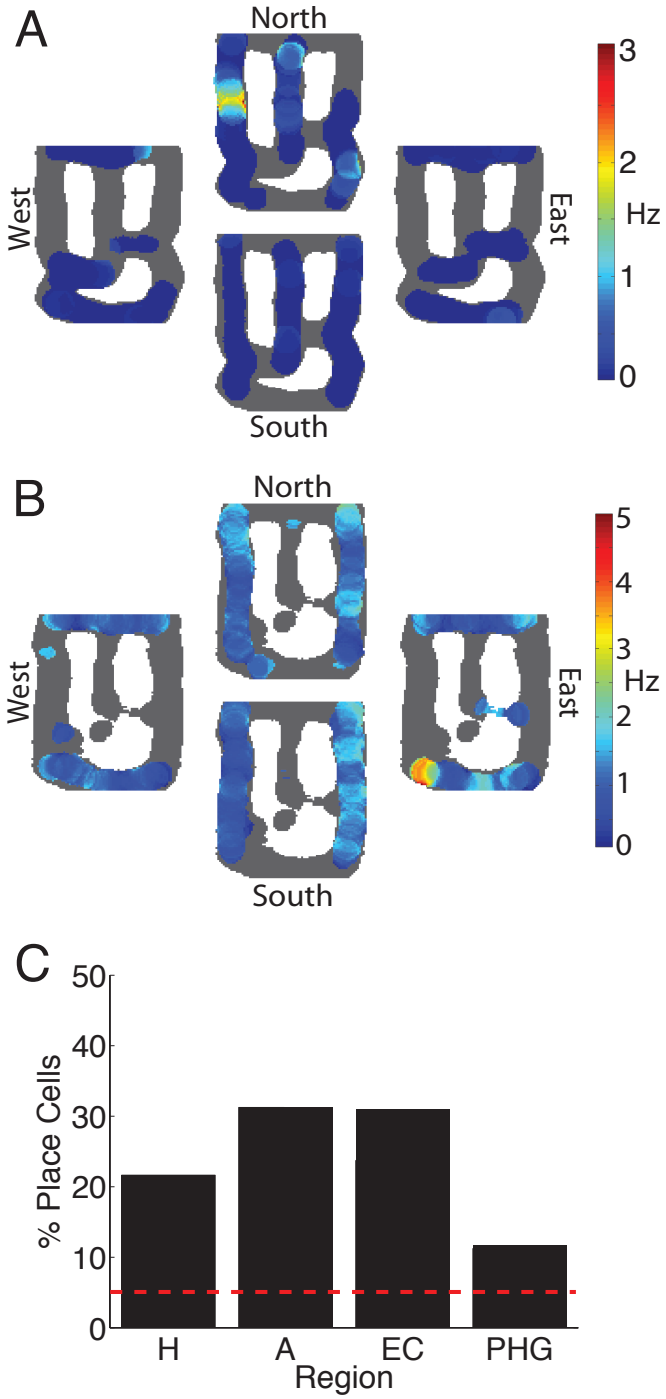


Figure 7.5: Place-responsive cells. **A**. Firing rate map for a hippocampal neuron responsive to northward traversals, plotted separately for each cardinal direction. Grey represents all areas traversed by the subject, regardless of direction of travel. **B.** An entorhinal neuron responsive to eastward traversals. **C.** The regional distribution of place-responsive cells (among 362 neurons). Red line indicates a false positive rate of 5%.

In a study of neural pattern similarity, Miller et al (2013B) had patients with implanted hybrid microwire and macrowire electrodes play a very similar game to the one described above. They first identified patterns of neuronal activity that represented subjects' location within the virtual town. In many species of mammals, neurons primarily in the hippocampal formation respond preferentially when the animal traverses a given region within a spatial environment (O'Keefe & Dostrovsky, 1971; Muller, Bostock, Taube, & Kubie, 1994). These so-called place cells have also been identified in humans as they perform virtual navigation tasks similar to the one studied here (Ekstrom et al., 2003; Jacobs, Kahana, Ekstrom, Mollison, & Fried, 2010). Miller et al identified place-responsive cells as the neurons that exhibited significantly increased firing at a particular location in the virtual environment. Figure 7.5A depicts the activity of one example place-responsive cell, which increased its firing rate when the subject was positioned on the left side of the virtual environment and facing north. Like this cell, the majority of the identified place-responsive cells exhibited direction dependence (72%) and did not exhibit significant place fields when direction of traversal was not taken into account. This finding is similar to prior findings of directionally oriented place cells in environments with clearly defined routes, in contrast to open environments, where omnidirectional place cells predominate (Muller et al., 1994; Ekstrom et al., 2003). Figure 7.5B shows the firing rate of a place-responsive cell from the entorhinal cortex, which activated in the south part of the environment during eastward movements. They identified 88 place-responsive cells, comprising 24.3% of all observed neurons. There were significant numbers of place-responsive cells throughout all MTL subregions studied (binomial test with $p < 0.05$ for each region, Figure 7.5C).

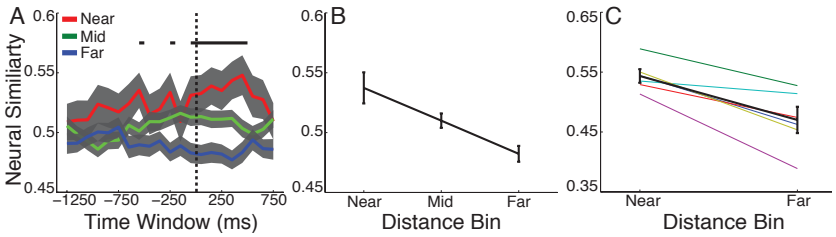


Figure 7.6: A. The average neural similarity for near, middle, and far spatial distance bins between ensemble place-responsive cell activity during navigation and place-responsive cell activity during item recall as a function of time relative to recall onset, computed in overlapping 500-ms time windows (x-axis values indicate the center of the time window). Shaded regions indicate $\pm 1 S.E.M.$ across recalled items. Significant timepoints are indicated with a horizontal bar (determined by ANOVAs at each time point with a false discovery rate adjusted significance threshold of 0.009). **B.** The average neural similarity for near, middle, and far spatial distance bins within the period of -300–700 ms relative to recall onset. Error bars indicate $\pm 1 S.E.M.$ across recalled items. **C.** The neural similarity for near and far spatial distance bins for each of the included subjects (thin colored lines) and the subject average (thick black line) within the period of -300–700 ms relative to recall onset. Error bars indicate standard error of the mean across subjects.

We found significant spatial context reinstatement surrounding the time of item vocalization (timecourse illustrated in Figure 7.6A). The level of neural similarity between recall activity and navigation activity appears ordered as expected from the spatial context reinstatement hypothesis, with areas of the environment near an item's encoding location exhibiting the highest similarity scores, intermediate spatial distances exhibiting middling similarity scores, and far spatial distances exhibiting the lowest similarity scores (this effect being strongest in the -300 to 700 ms interval illustrated in Figure 7.6B). An ANOVA indicated a significant effect of distance bin on the level of neural similarity ($F(2,300) = 8.7$, $p < .0001$). Performing this latter analysis across subjects rather than recall events revealed a similar result, with the neural similarity within the near distance bin being significantly greater than neural similarity within the far distance bin (Figure 7.6C, $t(5) = 5.0$, $p = .004$). These analyses show that during the spontaneous recall of an item, place-responsive neurons exhibit firing patterns similar to those

they showed during exploration of the region of the town where the item was previously delivered. Thus, one observes that recalling an episodic memory involves recovery of its spatial context, as seen in the activity of place-responsive cells in the human hippocampal formation and surrounding MTL regions.

Other applications

- Yaffe et al (2014) PNAS, Xie et al (2020) Nat Human Behavior.
- Foelkarts 2018 J. Neuro single unit reinstatement
- Lohnas et al. (2022)
- Verschure (2019) Nat Comm. 11 patients. Not great data.

8
Connectivity and Interactivity

In a previous chapter, we used the regression method to find an optimal linear model relating a collection of neural features—such as power at different frequencies and brain sites—to some measure of behavior, such as successfully encoding an item in a memory experiment. This regression approach illustrated how we could build predictive models of human behavior by linearly combining the influence of many neural features. Here we go beyond the analysis of individual features to consider how features of neural activity *interact* in their relation to behavior and cognition.

What do we mean by an *interaction* between features? In a statistical sense, an interaction of two variables simply means that the influence of an independent variable x_i on a dependent variable y depends on the level of another independent variable x_j. Consider the simple case of just two pairs of recording contacts, perhaps one pair of electrodes in the left hippocampus and a second pair in the right hippocampus. Let $v_L(t)$ and $v_R(t)$ represent these two voltage series and let $m(t)$ represent the brain's propensity to effectively store new memories at a given moment [1]. We can separately estimate the correlations between $m(t)$ and transformations of the voltage series, e.g. spectral power estimated at a given frequency, $f = 100Hz$, denoted $P_f[v_L(t)]$ and between $P_f[v_R(t)]$. The correlations we estimate will tell us the degree to $100Hz$ power in the left and right hippocampi predict mnemonic ability. Using multiple regression we can further ask what linear combination of these two features best predicts $m(t)$ by solving the simple bivariate regression model:

$$m(t) = \beta_0 + \beta_1 P_{f=100}[v_L(t)] + \beta_2 P_{f=100}[v_L(t)] + \epsilon$$

Implicit in the above equation is the idea that time-variation in $P_{100}[v(t)]$ relates to time variation in $m(t)$. The interaction hypothesis states that the influence of $P_{100}[v_L(t)]$ depends on the value of $P_{100}[v_R(t)]$, such that a larger realization of $P_{100}[v_R(t)]$ leads to a bigger or smaller effect of $P_{100}[v_L(t)]$ on $m(t)$. Mathematically, we can capture this dependency by adding a multiplicative term to the above expression:

$$m(t) = \beta_0 + \beta_1 P_{100}[v_L(t)] + \beta_2 P_{100}[v_L(t)] + \beta_{12} P_{100}[v_L(t)] P_{100}[v_R(t)] + \epsilon$$

In this model, we can estimate the coefficients β_1, β_2, and $\beta_1 2$ that optimize the goodness of fit between the model and the data. To the extent that including the multiplicative term $P_{100}[v_L(t)] \times P_{100}[v_R(t)]$ leads to better model prediction, we would say that the variables reliably interact with one another.

[1] In this example we will eschew the question of how our memory state variable, $m(t)$, relates to measured memory performance, such as successful free recall or recognition memory.

In the above example, we created a product for each interval of interest by averaging the powers across the entire interval. Assuming that we have zero-centered our features (e.g., by z-scoring them), we can think of this interaction as distinguishing trials with the same sign (high-high and low-low) and different sign features (high-low and low-high). A positive beta weight on the interaction implies that high-high states signal good memory more than the sum of the effect predicted based on trials with high-zero features and zero-high features.

This simple case has only two features. But suppose that for each trial of an experiment, we extract M spectral power features for each of N recording channels, giving us $N \times M$ features. Each unique pair of features will create a product term, leading to $\frac{N^2M^2-NM}{2}$ product features. One may ask whether this feature set will be too large to estimate using the regression methods described in Chapter 6. Theoretically, one should be able to design a regularization scheme that can handle arbitrarily large feature sets (REF). In practice, however, when feature sets become sufficiently large, being slightly off with the regularization can result in extreme overfitting, and finding the right hyperparameters may not be practicable.

Thinking about the data generating process can provide valuable insights into modeling the interactions between activity at different brain locations. Consider the theory implicit in the preceding analysis. By testing for an interaction between signals recorded in different brain locations as predictors of good memory we that predict successful memory we

Adjacency matrix

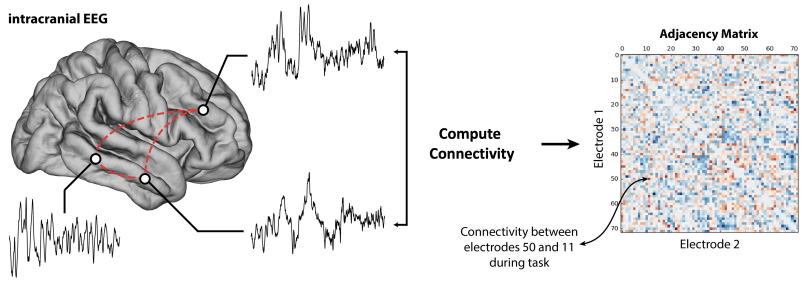


Figure 8.1: Computing an adjacency matrix.

Neural measures of adjacency

Functional vs. structural connectivity

Spectral measures: Coherence and Synchrony

Application 1:

pairwise similarity metrics applied to FR data, like in Ethan's two papers. We could either use phase synchronization, or coherence, or amplitude envelope

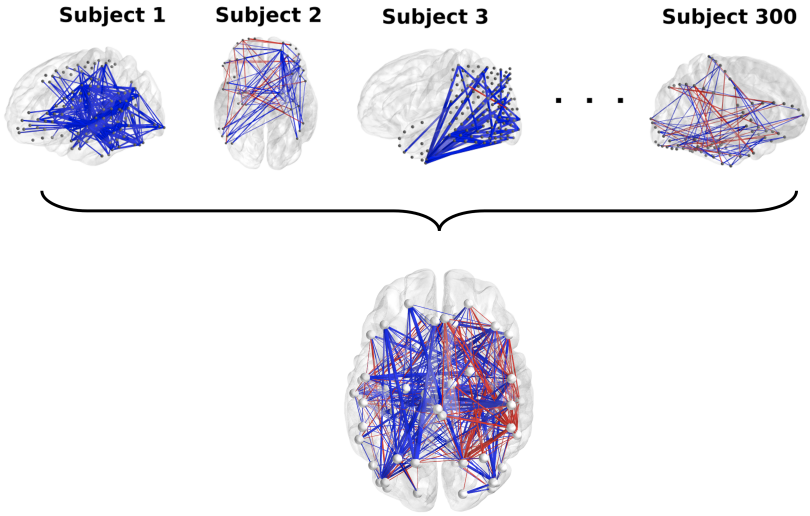


Figure 8.2: Computing an adjacency matrix.

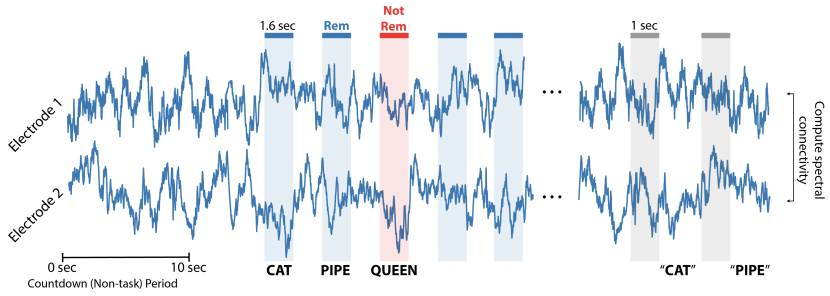


Figure 8.3: Computing an adjacency matrix.

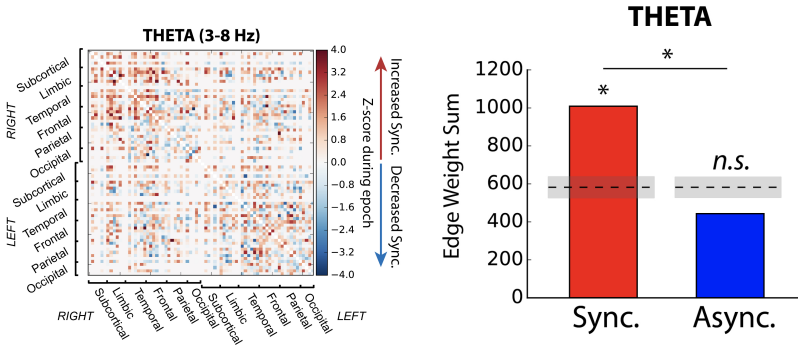


Figure 8.4: Computing an adjacency matrix.

correlation, or lagged correlation. No beauty contest. Lets go with coherence.
Same 20 subjects as Tung has identified.

MTL electrodes only. Divided into 10 regions (2 $\times$ 5).

Pairwise coherence for recalled vs. non-recalled items at encoding.

Correlation measures: Contemporaneous vs. lagged

Model-based methods

Application 2:

Exercise: Evaluate order in ECoG FR1 dataset

Run AR(p), p = 1,..8, on all FR1 subjects on all encoding events

Same analysis as application 1 but using AR model and SV model.

Traveling waves

Ask Josh to draft something

Relevant lab papers

- Burke, J. F., Zaghloul, K. A., Jacobs, J., Williams, R. B., Sperling, M. R., Sharan, A. D., et al. (2013). Synchronous and asynchronous theta and gamma activity during episodic memory formation. Journal of Neuroscience, 33(1), 292âĂŞ304.
- Solomon 2017 Nature Communications
- Solomon 2018 Nature Communications
- Solomon 2019 Current Biology
- Dani Bassett modeling paper with our lab (deterministic dynamical model)
- Tung Phan et al 2019 eLife (probabilistic dynamical model)

9
Brain Stimulation

In their classic annual review article on the state of the field, (Tulving & Madigan, 1970) bemoaned the slow rate of progress, whimsically noting that "Once man achieves the control over the erasure and transmission of memory by means of biological or chemical methods, psychologists armed with memory drums, F tables, and even on-line computers will have become superfluous in the same sense as philosophers became superfluous with the advancement of modern science; they will be permitted to talk, about memory, if they wish, but nobody will take them seriously." One of us (Kahana) remembers chuckling when he first read this passage in 1990, as he could not have imagined then that serious scientists would foresee a time when the external modulation of memories was even a remote possibility. The present reader will recognize that neural modulation has become a major therapeutic approach in the treatment of neurological disease and that the outrageous prospects raised by Tulving and Madigan now appear to be within arm's reach.

Using Stimulation to Establish Causality

In previous chapters we have seen how diverse brain regions exhibit differential activity across various manipulations of memory encoding and retrieval. Although it is tempting to interpret these contrasts as reflecting the true engagement of these signals in the manipulated memory processes, this need not be the case. To establish a causal link between neural activity and memory function we must conduct experiments that directly manipulate neural function and assess the effects of these manipulations on behavior and cognition. Whereas historically such manipulations had typically relied on lesioning a part of the brain, we now have a wide array of tools available to exogenously manipulate and modulate the nervous system. One approach that dovetails nicely with the invasive recording studies reviewed above is direct electrical stimulation.

Neurologists routinely use direct electrical stimulation of the brain for functional mapping of eloquent cortex. In these clinical mapping studies, stimulation reversibly lesions a small brain area (via electrical depolarization). This allows physicians to determine whether this lesioning impairs some crucial cognitive ability, such as speech or motor function. Functional mapping via electrical stimulation thus allows neurosurgeons to identify the boundaries of brain areas to be avoided during resective surgeries used to

treat intractable epilepsy or brain tumors. More recently, brain stimulation has also been used to disrupt pathological signals in the brain, including seizures, or even to restore healthy brain function. We turn to this prospect in the next section.

Penfield's Patients

The field of human direct brain stimulation began with the work of Wilder Graves Penfield (1891-1976). Born in Montana, Penfield went on to establish and direct the Montreal Neurological Institute at McGill University. His work revolutionized the field of neurosurgery, in particular for the treatment of epilepsy. Penfield developed the functional mapping technique of electrically stimulating the brains of awake patients under local anesthesia in the operating room. At the time, neurologists knew that patients often have a stereotyped perceptual disturbance, known as an 'aura,' just prior to experiencing a seizure. Penfield reasoned that if stimulation of a particular brain area evoked a patient's aura, it might indicate that the underlying stimulated tissue contributed to the patient's seizures. Beyond introducing the important clinical technique of functional mapping, Penfield surreptitiously discovered that electrical stimulation of certain brain regions, particularly within the temporal lobes, could evoke memory-like phenomena.

Penfield describes what the patients experienced as follows:

> . . . the phenomenon is sometimes extensive and elaborate, sometimes fragmentary. It may include the sights and sounds and the accompanying emotions of a period of time, and the patient usually recognizes it spontaneously as coming from his past (Penfield & Perot, 1963).

Penfield also noted that stimulation applied over years evoked the same experience when it was applied to the same part of the cortex. These observations, collected over decades of research, showed that direct brain stimulation could evoke autobiographical memories, laying the groundwork for more recent studies designed to test causal mechanisms of memory retrieval. Penfield also suggested that stimulation might modulate the brain's ability to encode new information.

Figure 9.1: Wilder Graves Penfield. Among other major accomplishments, the American-Canadian neurosurgeon developed the technique of stimulation mapping in epilepsy surgery. In the process, he paved the way for the use of direct brain stimulation for studying cognitive processes. Image courtesy of Wikipedia.

How stimulation can evoke memories

In the years following Penfield's pioneering discoveries, several researchers published studies showing how stimulation could evoke recollection (Gloor, Olivier, Quesney, Andermann, & Horowitz, 1982; Halgren, Walter, Cherlow, & Crandall, 1978; Ojemann, 1991). One recent study aimed to build upon this prior work by relating the effect of stimulation on memory to the normal patterns of physiological activity observed at the stimulation site (Jacobs, Lega, & Anderson, 2012). This approach is important for determining stimulation's mechanism of action, which remains poorly understood. In this report of one patient, Jacobs et al. applied bipolar stimulation to a pair of adjacent electrodes located in the inferior temporal lobe. They found that following stimulation, the patient consistently recalled memories from high school (Figure 9.2).

Although interesting that the authors showed that stimulation of this region of temporal lobe reliably produced memories of high school, the study's

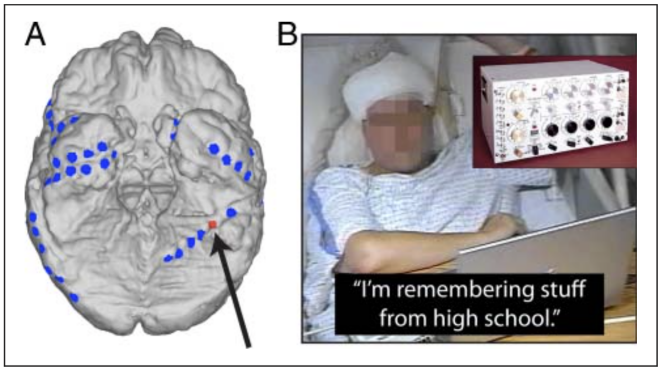


Figure 9.2: Stimulation evoked memories of high school. (A) The surface reconstruction of the brain of a patient who underwent direct electrical stimulation to the location indicated with the red dot. The patient's other recording electrodes are rendered as blue dots (ventral brain view). (B) An anonymized depiction of the patient and an example response given during the experiment. Figure courtesy of (Jacobs et al., 2012).

other important contribution was its analysis of the electrophysiology of this region during unstimulated retrieval of high school memories. To assess this, the researchers recorded intracranial EEG activity while the patient retrieved information about their experience from high school without stimulation. As a measure of the activity of the underlying neural populations (see Chapters 2 and 3), the authors examined spectral power in the high-frequency range (60–150 Hz) at the stimulated electrode during retrieval. The authors found that the stimulated electrode showed a significant decrease in high-frequency power during retrieval of memories from high school compared to non-high school memory retrieval (Figure 9.3). Given that stimulation of this region specifically evoked memories of high school, the findings suggest that stimulation that mimics the typical electrophysiological activity of the stimulated site can recapitulate the retrieval operations carried out by the site.

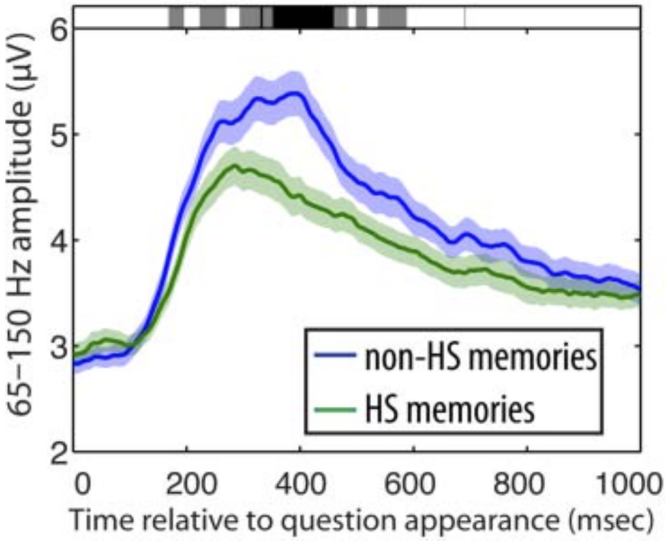


Figure 9.3: High-frequency power during memory retrieval. When the patient was cued to retrieve memories from high school, high-frequency power was significantly lower than during retrieval of non-high school memories, suggesting that stimulation of this region specifically instated neural activity evoked by high school memory retrieval. Figure courtesy of (Jacobs et al., 2012).

Stimulation of the medial temporal lobes

Unlike Jacobs et al., (2012), most studies that have examined direct brain stimulation for memory modulation have targeted the hippocampus and medial temporal lobe structures under the assumption that these brain regions play a key role in memory encoding, retention, and retrieval. These studies typically used an *open-loop* design in which stimulation is applied to the target structure of interest irrespective of ongoing neural activity. Researchers then measure stimulation's effects on memory performance and, sometimes, physiology. The results of this literature have been mixed, with some studies showing memory facilitation with stimulation of the medial temporal lobes and other studies showing memory disruption. As we will see later in the chapter, this inconsistency in the literature may be due both to the selection of stimulation target as well as to the open-loop fashion in which stimulation was delivered.

Several studies have used stimulation of the hippocampus and medial temporal lobes to disrupt various stages of the memory process. Single-pulse stimulation of the hippocampus has been shown to disrupt memory encoding and retrieval (Halgren, Wilson, & Stapleton, 1985) in a visual recognition task. In a sample of four patients, single-pulse stimulation was applied at the onset of images at either encoding, retrieval, both, or neither. Numerically, stimulation at both encoding and retrieval led to worse memory performance than stimulation at either encoding or retrieval. They applied stimulation to various medial temporal lobe structures (hippocampus, parahippocampal cortex, and amygdala) and their findings suggested that applying brief stimulation to these regions could reliably impair memory function.

Merkow et al. (2017) also examined how stimulating the medial temporal lobe during different phases of the memory process leads to changes in performance. While patients performed a free recall task the researchers applied 50 Hz stimulation during either the encoding, arithmetic distractor, or recall phases of each list. The authors found that overall stimulation led subjects to recall fewer words and that this effect was largest for the lists where stimulation was applied during the arithmetic distractor period. The data are consistent with other work reviewed here showing stimulation-related disruption in memory performance. The data are also consistent with a contextual change-based account in which stimulation between encoding and retrieval led subjects' internal context representations to drift more rapidly. If this were true, retrieved-context models of memory would predict that memory should be worse in these cases because the context during the recall period, which is used to cue items for retrieval, is more dissimilar relative to the encoding period.

In another study of stimulation's effect on memory, Suthana et al (2012) applied 50 Hz stimulation to the entorhinal cortex during a spatial memory task. Participants (N = 6) played a computer game in which they had to learn to navigate around a virtual environment (Suthana et al., 2012). During the learning phase, they applied stimulation to white matter proximate regions within the entorhinal cortex. They then measured the latency and path length that characterized subject responses to cues during the test phase. The researchers found that subjects could more quickly and efficiently locate targets learned during stimulation blocks, suggesting that stimulation during encoding led to improved spatial learning.

However, an attempted replication of Suthana et al's finding showed that stimulation of the entorhinal cortex (and medial temporal lobes generally) during encoding reliably *disrupted* memory performance (Jacobs et al., 2016). Using a similar spatial navigation task that included more observations per subject, these researchers found that entorhinal stimulation impaired both spatial and verbal memory performance. Stimulation of the hippocampus also disrupted memory performance.

Stimulation outside of the medial temporal lobes

Perrine et al. (1994) found that stimulation of lateral temporal cortex disrupted memory in three out of nine patients who did not have very significant baseline memory impairment. These results are only suggestive as stimulation was applied during different task phases and locations in different patients.

Clark et al., 1999

Ezzyat et al 2017 open loop

Closed-loop neuromodulation

Ezzyat et al., closed loop

Microstimulation of reward learning

Ramayya J. Neuro

Ramayya Frontiers

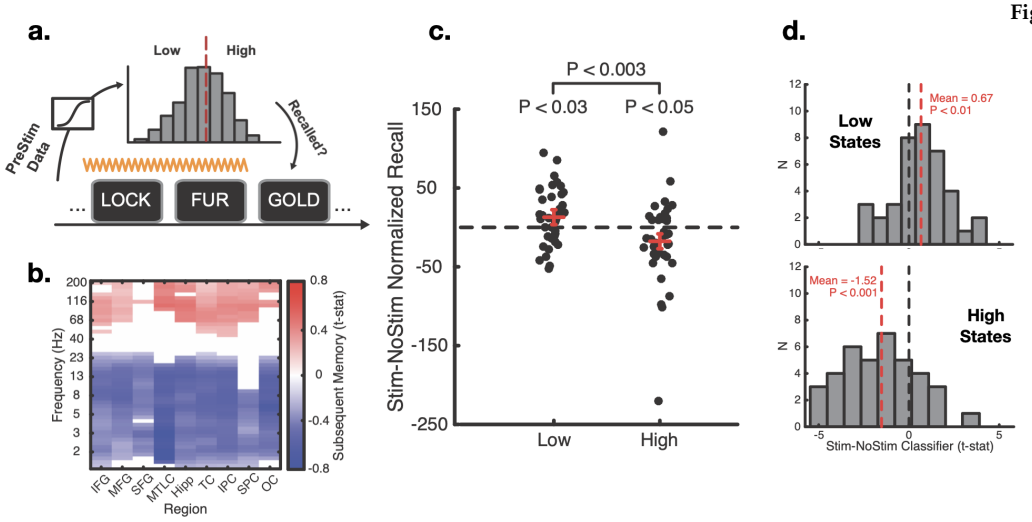


Figure 9.4: XXX

Figure 9.5: XXX

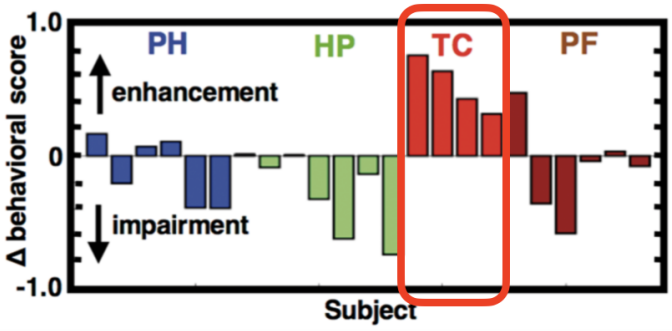


Figure 9.6: XXX

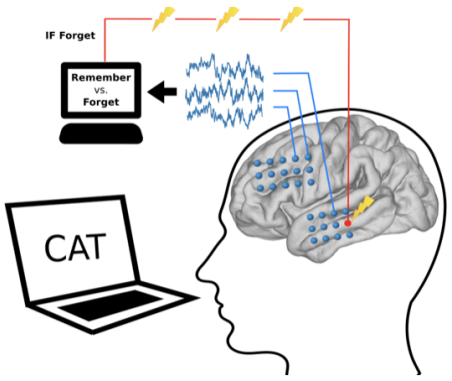
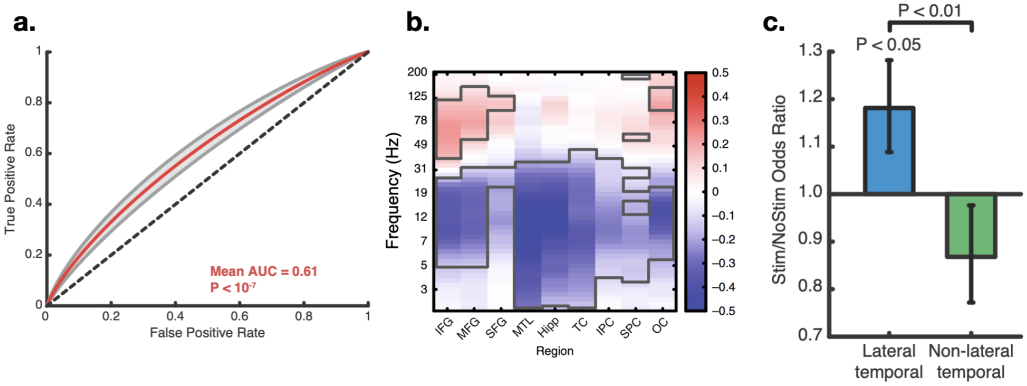


Figure 9.7: XXX



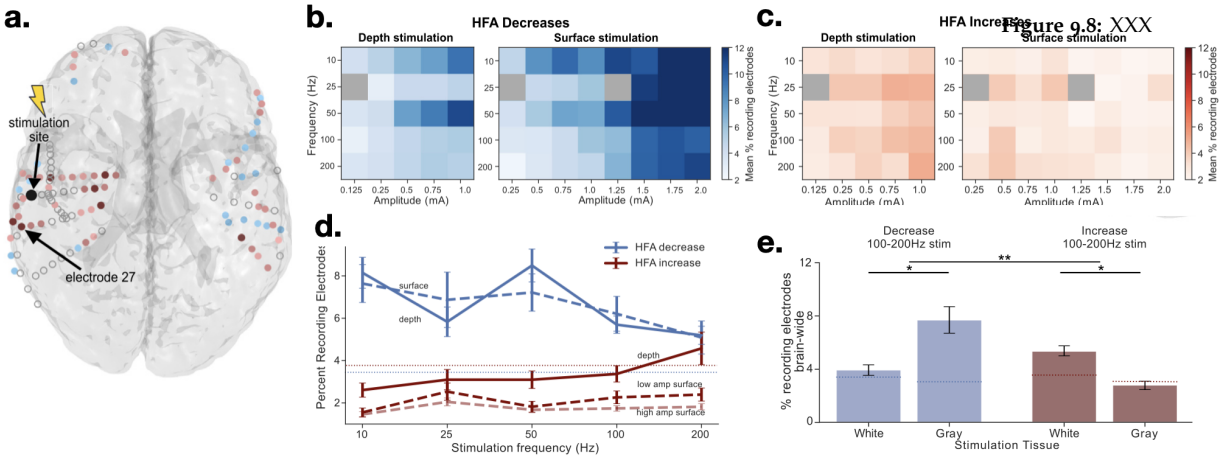


Figure 9.8: XXX

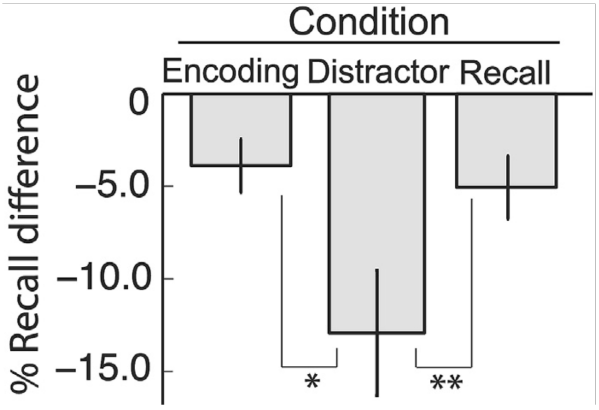


Figure 9.9: XXX

A
Description of Datasets

This appendix describes four collections of electrophysiological datasets: studies involving (1) scalp EEG recordings taken during memory tests, (2) intracranial EEG recordings, (3) intracranial electrical stimulation and recordings, and (4) single-neuron recordings. Each of the major sections describes multiple studies belonging to each of these categories.

Scalp EEG Datasets

PEERS: The Penn Electrophysiology of Encoding and Retrieval Study

While Ebbinghaus' classic memory research involved extensive self-experimentation, most memory studies in the last century have relied on single-session experiments with small samples. In contrast, the Penn Electrophysiology of Encoding and Retrieval Study (PEERS) aimed to collect high-resolution, within-subject data from a large number of subjects performing various episodic memory tasks. Between 2008 and 2018, PEERS gathered data from over 300 subjects in more than 7,000 sessions, focusing on trial-level data, individual differences, and continuous EEG data during memory encoding and retrieval. Below we describe the methods of the five PEERS datasets and their online availability.

Each of the five PEERS experiments involved multiple sessions of memory tasks with EEG recording. Each experiment involved some variant of a free recall task. Experiments 1-3 included encoding task manipulations, variation in distractor conditions, and end-of-session recognition and final-recall tests. Subjects completed these first three PEERS experiments across 20 sessions. In addition, this cohort also completed two sessions of neuropsychological tests. PEERS Experiment 4 sought to maximize the statistical power of data collected in a delayed recall-task without any encoding task manipulations. We estimated, based on the earlier PEERS studies, that subjects would not be able to complete more than 24 two-hour long sessions in a single term. With the goal of maximizing our statistical power we thus recruited subjects for a 24-session experiment, striving to enroll 10 subjects at the start of each term (we typically completed around 7). We collected vocal responses which we annotated (offline) for accuracy and response times. We conducted a fifth PEERS experiment designed to control for pre-motor correlates of retrieval. This was the last PEERS study completed prior to the start of the COVID-19 pandemic.

Due to the very substantial investment of time and resources, each subject first participated in a screening session to ensure that they understood the demands of the experiment prior to signing on for the full experiment. Below we provide a concise description of the experimental methods. Table 1 gives the number of subjects who completed each experiment. Additional procedural details appear in an online appendix at `memory.psych.upenn.edu`

PEERS experiment	*N*	Sessions	Dates
Preliminary Experiment	∼730	1	2010-2019
Exp. 1: Immed. recall + task manip. Final-free recall. Recognition.	172	7	2010-2014
Exp. 2: Recall + distractors. Final-free recall. Recognition.	157	7-9	2010-2014
Exp. 3: Exp 1 + externalized recall. Final-free recall. Recognition	60 (IFR), 92 (EFR)	4 (IFR), 6 (EFR)	2010-2014
Exp. 4: Delayed recall	98	24	2014-2018
Exp. 5: Long-delay recall + pre-motor control	57	10	2019-2020

Table A.1: *Demographic Information for PEERS Studies*

PEERS Experiments 1 and 3

Because Experiments 1 and 3 were virtually identical, we describe their methods together. As illustrated in Figure **??**A, each session comprised a series of 16 immediate free recall trials, each involving a unique list of 16 visually-presented words. Each session ended with a recognition test (yellow box). Half of the sessions were randomly chosen to include a final free recall test before recognition (in final free recall, subjects attempt to recall as many words as they can remember from all 16 lists) Experiment 3 differed from Experiment 1 in that a subset of subjects received *externalized free recall* instructions. In externalized recall (Kahana, Dolan, Sauder, & Wingfield, 2005) subjects verbalized all words that came to mind at the time of test, even if they thought those words did not occur in the most recent list or had already been recalled during the current recall period, and to press the spacebar following any such error.

Subjects encountered three types of lists: (1) No-task lists, which they studied with the generic instruction of trying to learn the items for a subsequent test, (2) task lists, where each item appeared concurrently with a cue indicating one of two judgments (size or animacy) the subject should make for that word, and (3) task shift lists, where subjects alternated between size and animacy tasks every 2-6 items within each list. The size task asked subjects âĂIJWill this item fit into a shoebox?âĂİ; the animacy task asked subjects âĂIJDoes this word refer to something living or not living?âĂİ The current task was indicated by the color, font, and case of the presented item. Each session included 12 task lists and four no-task lists. The first session of PEERS Experiment 1 included equal numbers of size, animacy, and task-shift lists; subsequent sessions included three size, three animacy, and six task-shift lists. We constructed a pool of 1,638 words for use in PEERS1-3. Based on the results of a prior norming study, only words that were clear in meaning and that could be reliably judged in the size and animacy encoding

tasks were included in the pool.

PEERS Experiment 2

Experiment 2 introduced a within-subject, within-session, distractor manipulation (Figure **??**B). In addition to immediate free recall trials, as in Experiments 1 and 3, this experiment introduced delayed free recall and continual distractor free recall, with distractor intervals of varying duration. In each distractor interval, subjects solved math problems of the form $A + B + C =?$, where A, B, and C were positive, single-digit integers. When a math problem appeared, subjects typed the sum as quickly as possible consistent with high accuracy (they received a monetary bonus based on the speed and accuracy of their responses). For the distractor intervals in the first two lists, one list had a distractor period following the last word presentation for 8 s and the other had an 8 s distractor period prior to and following each word presentation. In the remaining 10 lists, subjects performed free recall with five possible durations for the between-item and end-of-list distractor tasks, such that two lists had each of the five conditions. As listed here, the first number indicates the between-list distractor duration and the second number indicates the end-of-list distractor, both in seconds: 0-0, 0-8, 0-16, 8-8, 16-16. A 0 s distractor refers to the typical, non-filled duration intervals as described for Experiments 1 and 3. Subjects encoded all items using either a size or an animacy judgment task. Session one included seven size-judgment lists and seven animacy judgment lists. Subsequent sessions included six task-shift lists, three size-task lists and three animacy-task lists.

PEERS Experiment 4

This experiment sought to simplify the methodology used in previous experiments, focusing exclusively on delayed free recall. Here, each of 98 subjects completed 24 sessions of delayed free recall. Each session consisted of 24 trials, with each trial containing a list of 24 individually presented words followed by a 24-second distractor period (see Figure **??**C). A random half of the lists (excluding the first list) were preceded by a 24-second, distractor-filled delay. A free recall test followed the post-list distractor on each list.

The word pool for this experiment consisted of a 576-word subset of the 1638-word pool used in a previous PEERS experiment, and subjects saw the same 576 words (24 lists $\times$ 24 items) on each of sessions 1 through 23 with the ordering of words randomized for each session. The 24th session introduced a set of novel words, as described in the Appendix. Subjects were given a short break (approximately 5 minutes) after every 8 lists in a session.

PEERS Experiment 5

The fifth PEERS experiment sought to contrast neural correlates of retrieval following a very long delay, with neural correlates of retrieval of a just presented single item. During each of the first five sessions, subjects quietly read each of the 576 words used in Experiment 4. After reading each word, they waited 1 sec (or longer) before saying the word aloud. These 576 immediate recall trials occurred in 24 blocks of 24 items, each preceded by a countdown, thus mimicking the 24-list structure of Experiment 4.

On each trial, a black screen was shown for a jittered 1000–1600 ms (uniformly distributed), after which a single word appeared onscreen in white text for 1200–1800 ms (uniformly distributed). Following presentation, the screen went blank again and subjects were instructed to pause briefly, and then vocalize the word they had just seen. If they began speaking within 1.0 second of word offset, the message âĂIJToo fast.âĂİ appeared on the screen in red text. By avoiding these messages subjects could increase the size of their bonus payment. After the subject finished speaking, a tone sounded, marking the end of the current trial. Speech was detected using a volume amplitude threshold. In addition to the 10-second countdown between blocks, two 2-minute mid-session breaks were administered after block eight and block 16.

At the start of session six, subjects were given a surprise free recall task in which they were instructed to recall as many words as possible from the previous sessions in any order, while also vocalizing any additional words that come to mind in their attempt to recall these items (Externalized recall instructions: Kahana, Dolan, et al., 2005; Lohnas, Polyn, & Kahana, 2015; Zaromb et al., 2006). We administered this long-delay recall task as the start of each of the sessions 6 through 10, giving subjects 10 minutes to recall as many of the 576 words as they could remember. After this free recall test, subjects continued with the same immediate recall task as in earlier sessions. Subjects saw the same 576 words in each of their 10 sessions, but the ordering of these words was randomized for each session. Although subjects saw the 576 words across multiple sessions, the only information identifying these words as belonging to the target list was their occurrence within the context of our experiment, thus making this a test of long-term episodic memory.

Additional Methodological Details

In each of the PEERS experiments, subjects received a base salary for their participation. In addition, they received a modest bonus for performance and a separate bonus for completing all of the sessions. The performance bonus varied slightly across experiments, but it incentivized subjects for achieving high levels of recall while maintaining a high-level of performance on the arithmetic distractor tasks. In addition, we provided a bonus to subjects for maintaining a low blink rate during critical item presentation events.

Subjects were tested using one of two recording systems: the Electrical Geodesics HydroCel Geodesic Sensor Net (HCGSN) system and the BioSemi ActiveTwo system. Subjects with IDs between LTP093 and LTP330 were tested using the EGI system; subjects with IDs LTP331 and higher were tested using the BioSemi system. In both systems alignment between behavioral and EEG data was achieved by sending trigger pulses over a TTL channel on the EEG recording. Every 800–1200 ms (uniformly distributed) the experiment's software sent an electrical pulse over the TTL channel and recorded a timestamp at the time the pulse was sent.

[1] EGI system specifications

[1] EEG measurements were recorded using Geodesic Sensor Nets (GSN; Netstation 4.3 acquisition environment, from Electrical Geodesics, Inc.). The GSN provided 129 standardized electrode placements across participants. All channels were digitized at a sampling rate of 500 Hz, and the signal from the caps was amplified via either the Net Amps 200 or 300 amplifier. Recordings

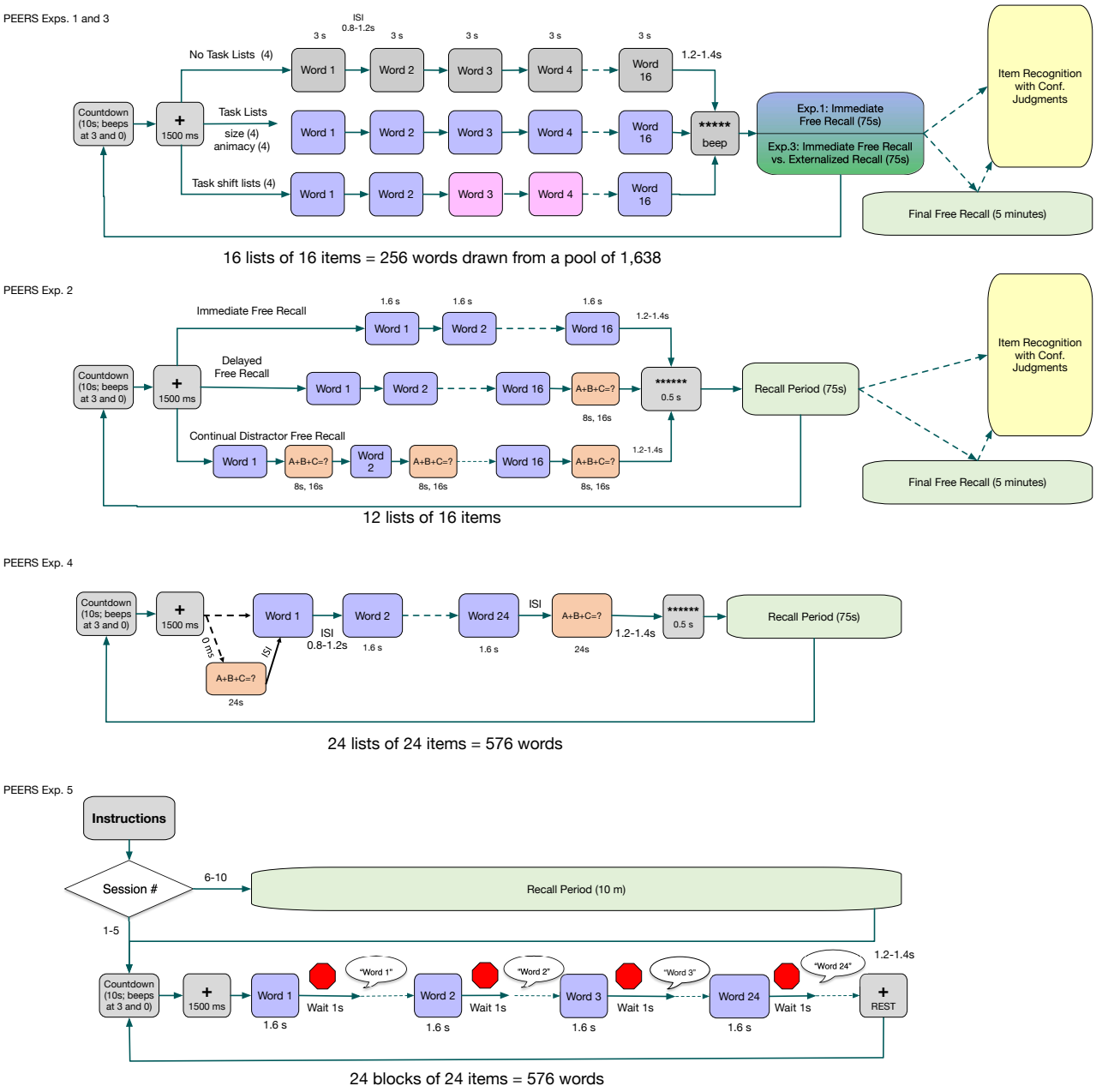


Figure A.1: Schematic of PEERS methods. The same group of subjects took part in Experiments 1-3, across 20 experimental sessions. Experiments 4 and 5 involved separate subject groups, recruited in later years of the project. Each experiment involved some form of a free recall task, and Experiments 1-3 also included recognition and final-free recall tasks. Experiment 5 only included final free recall. For Experiments 1-3, subjects either studied items without a specific encoding task, or judged items' size or animacy. The color of the word bubbles in the first row of the schematic indicates

were referenced to Cz (electrode 129 on the GSN), and later re-referenced to the average of all electrodes To help identify eyeblink and other movement artifacts, electrooculogram (EOG) activity was monitored bipolarly using right and left electrode pairs on the GSN (electrodes 8 and 126 on the right; electrodes 25 and 127 on the left).

[2] EEG measurements were recorded using BioSemi ActiveTwo systems alongside BioSemi's ActiView recording software. Recordings were made from 128 standardized electrode placements across participants. The signal was amplified and digitized by an ActiveTwo AD-box at a sampling rate of 2048 Hz. The ActiveTwo system uses a feedback loop of two electrodes in place of an ordinary ground electrode (Luck, 2014, Box 5.3): a passive *driven right leg* (DRL) electrode injects a small current to move the the potential of the subject towards that of the amplifier circut and an active *common mode sense* (CMS) electrode serves the function of an ordinary ground electrode (the potentials between all other electrodes and the CMS electrode are recorded). Referencing takes place offline (we referenced to the average of all electrodes). To help identify eyeblink and other movement artifacts, electrooculogram (EOG) activity was monitored using additional electrode pairs at each eye.

[2] Biosemi system specifications

Data Availability

The PEERS data, both behavior and electrophysiology, is freely available as OpenNeuro Dataset `ds004395` (`https://openneuro.org/datasets/ds004395/`; CITE). Data can be downloaded directly through the OpenNeuro web interface or by using their command line utility tool. The dataset has its own Digital Object Identifier (DOI) and citation tools are available on the dataset webpage. The Computational Memory Lab website also provides a detailed methods description of each of the PEERS studies described above: `http://memory.psych.upenn.edu/PEERS`.

Intracranial Recordings

In some rare clinical indications, neurosurgeons will implant electrodes into the brain's of patients and record brain activity while the patients remain awake. During these intracranial recording studies, researchers will sometimes ask patients to perform cognitive tasks thereby relating their brain recordings to task variables and behavioral performance. Here we describe a large series of experiments involving brain recordings during a variety of memory tasks, with data collected across ten medical centers. These studies formed part of a multi-center project, funded by the *Defense Advanced Research Projects Agency* and led by a team of scientists at the University of Pennsylvania. The long-term goal of this project was to develop technologies capable of restoring active memory.

Penn Restoring Active Memory (PRAM) Project

The PRAM project included five major experiments: Free recall of 'unrelated' word lists (FR), free recall of semantically categorized word lists (CatFR), paired-associate memory (PAL), recall of spatial locations during active navigation (YC), and cued recall of object-location associations (TH). Below

we describe the basic methods used in each study and provide references to the original papers that contain more detailed methods descriptions.

Free Recall of 'unrelated' and semantically organized word lists (FR and CatFR)

In these experiments, subjects contributed multiple sessions of delayed free recall of either unrelated or categorically organized word lists (Experiment Codes FR1 and CatFR1, respectively). Subjects were neurosurgical patients undergoing intracranial electroencephalographic monitoring as part of clinical treatment for drug-resistant epilepsy. Data were collected as part of a multi-center project designed to assess the effects of electrical stimulation on memory-related brain function. [3] Electrophysiological data were collected from electrodes implanted subdurally on the cortical surface as well as deep within the brain parenchyma[4]. In each case, the clinical team determined the placement of the electrodes so as to best localize epileptogenic regions. Subdural contacts were arranged in both strip and grid configurations with an inter-contact spacing of 10 mm. Most subjects also had temporal lobe depth electrodes with 5 mm inter-contact spacing. Intracranial data were referenced to a common contact placed either intracranially, on the scalp or mastoid process. We provide both these raw data as well as data rereferenced using a bipolar montage (Burke et al., 2014).
[5]

Paired-associate learning (PAL)

To investigate the neural basis of human associative learning and retrieval, cognitive neuroscientists have subjects perform the classic verbal paired associate task – a task long used in the psychological study of memory and learning. Here we describe the methods used in the paired-associate learning (PAL) task investigated by Dr. K. Zaghloul and colleagues in numerous papers (Greenberg, Burke, Haque, Kahana, & Zaghloul, 2015; Yaffe et al., 2014). In this procedure, each list comprised six pairs of words selected randomly and without replacement from the same set of 300 words described in the FR task description (above). Subjects studied word pairs in anticipation of a delayed cued recall test. Preceding each word pair, a fixation cross appeared for a jittered 250âĂŞ300 ms pre-stimulus interval followed by a blank inter-stimulus interval (ISI) between 500âĂŞ750 ms. Following this pre-stimulus interval, the two words appeared stacked in the center of the screen for 4000 ms. A blank ISI of 1000 ms jitter followed pair presentation.

Following presentation of the sixth study pair, and before the recall test, subjects performed the same arithmetic distractor task as described in the FR and CatFR tasks above. In the test phase, one randomly chosen word from each of the six studied pairs appeared as a cue for recall of its mate. The cue word was chosen randomly and could either have been the first or second member of a given pair. As the cue word appeared, subjects attempted to vocally recall the correct target word. Each cue word appeared for 4000 ms followed by a blank ISI of 1000 ms, providing subjects with a total of 5000 ms as the maximum recall time window. Subject could vocalize their response any time during the recall period after cue presentation. They were instructed to vocalize âĂŸPASSâĂŹ if they could not retrieve the target word. After

[3] Data were collected at the following centers: Thomas Jefferson University Hospital (Philadelphia, PA), Mayo Clinic (Rochester, MN), Hospital of the University of Pennsylvania (Philadelphia, PA), Emory University Hospital (Atlanta, GA), University of Texas Southwestern Medical Center (Dallas, TX), Dartmouth-Hitchcock Medical Center (Lebanon, NH), Columbia University Medical Center (New York, NY), National Institutes of Health (Bethesda, MD), and University of Washington Medical Center (Seattle, WA). The research protocol was approved by the IRB at each hospital and informed consent was obtained from each participant.

[4] Anatomical localization of electrode placement was accomplished using a two step process. First, hippocampal subfields and MTL cortices were automatically labeled in a pre-implant 2 mm thick coronal T2-weighted MRI using the automatic segmentation of hippocampal subfields (ASHS) multi-atlas segmentation method (Yushkevich et al., 2015). Following this automatic labeling procedure, a post-implant CT scan was co-registered with the MRI using Advanced Normalization Tools (Avants, Epstein, Grossman, & Gee, 2008). Electrodes that are visible in the CT were then localized within the MTL subregions by a pair of neuroradiologists with expertise in MTL anatomy. The whole brain cortical surface was also obtained from a pre-implant volumetric T1-weighted MRI using Freesurfer (Fischl et al., 2004), and subdural electrodes were separately co-registered and localized to cortical regions using an energy minimization algorithm (Dykstra et al., 2012).

[5] For the bi-polar montage, we identified all pairs of immediately adjacent contacts on every depth, strip and grid and took to difference between the signals recorded in each pair. The resulting bipolar timeseries was treated as a virtual electrode.

the study, trained research assistants manually annotated each recorded response as correct, intrusion, or pass. A response would be designated as âĂŸPASSâĂŹ when no vocalization was made or when the participant vocalized the word âĂŸPASS.âĂŹ We also recorded participantsâĂŹ response times as the time lapsed before they made the first vocalization following the probe onset. A single experimental session typically comprised 150 total word pairs, or trials (i.e., 25 lists).

YC

JOSH???

TH

JOSH???

RepFR2

Riley???

Courier

Deepti – Courier versions 1, 2, 3 and 4. Table with differences???

Microwire-recorded single neuron datasets

Chapter 5 discusses special methodologies for recording individual neurons in the human brain as subjects perform memory tasks. Here we describe several publicly available datasets involving single-neuron recordings during memory tasks.

Yellow-Cab Experiments

Manning et al's 20 subject dataset

YC Train

Goldmine

Courier

Rutishauser datasets

Intracranial Brain Stimulation Datasets

Chapter 9 discusses how memory scientists have been using direct electrical stimulation both to manipulate memory function and to study networks supporting memory function. Here we describe several publicly available datasets on brain stimulation and memory gathered by under the PRAM project.

Open Loop stimulation Experiments

FR2 and CatFR2
- PAL2
- YC2

Closed-loop Stimulation Experiments

FR3/5 and CatFR3/5
- PAL3/5
- TH3/5
- TH

Open-loop Parameter Search Experiments

PS1: Amplitude and Duration
- PS2: Amplitude and Frequency
- PS3: Amplitude and Theta-Burst Frequency
- Location Search

References

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences, USA*, *91*(15), 7041–7045.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41.

Babb, T. L., Carr, E., & Crandall, P. H. (1973). Analysis of extracellular firing patterns of deep temporal lobe structures in man. *Electroencephalography and Clinical Neurophysiology*, *34*(3), 247-257.

Babiloni, C., Vecchio, F., Mirabella, G., Buttiglione, M., Sebastiano, F., Picardi, A., ... others (2008). Hippocampal, amygdala, and neocortical synchronization of theta rhythms is related to an immediate recall during Rey auditory verbal learning test. *Human Brain Mapping*, *30*(7), 2077–2089.

Baddeley, A. D., & Longman, D. J. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, *21(8)*, 627-635.

Bahrick, H. P., & Phelps, E. (1987). Retention of spanish vocabulary over 8 years. *Journal of Experimental Pyschology: Learning, Memory, and Cognition*, *13(2)*, 344-349.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, Series B*, *57*, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bishop, D. V. M. (2014, 5). Interpreting unexpected significant findings. Retrieved from `https://figshare.com/articles/Interpreting_unexpected_significant_findings/1030406` doi: 10.6084/m9.figshare.1030406.v1

Bland, B. H. (1986). The physiology and pharmacology of hippocampal formation theta rhythms. *Prog. Neurobiol.*, *26*, 1–54.

Bódizs, R., Kántor, S., Szabó, G., Szũcs, A., Erõss, L., & Halász, P. (2001). Rhythmic hippocampal slow oscillation characterizes REM sleep in humans. *Hippocampus*, 747–753.

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 85–121). New York: John Wiley and Sons.

Brodie, D. A., & Murdock, B. B. (1977). Effects of presentation time on nominal and functional serial position curves in free recall. *Journal of Verbal Learning and Verbal Behavior*, *16*, 185-200. doi: 10.1016/S0022-5371(77)80046-7

Brown, T. (1824). *Lectures on the philosophy of the human mind*. Philadelphia, PA, USA: John Grigg and William P. Bason.

Burgess, N., Becker, S., King, J., & O'Keefe. (2001). Memory for events and their spatial context: models and experiments. *Philosophical Transcripts of the Royal Society London B: Biological Sciences*, *356*, 1493-1503.

Burke, J. F., Long, N. M., Zaghloul, K. A., Sharan, A. D., Sperling, M. R., & Kahana, M. J. (2014). Human intracranial high-frequency activity maps episodic memory formation in space and time. *NeuroImage*, *85*, 834–843. doi: 10.1016/j.neuroimage.2013.06.067

Burke, J. F., Ramayya, A. G., & Kahana, M. J. (2015). Human intracranial high-frequency activity during memory processing: Neural oscillations or stochastic volatility? *Current Opinion in Neurobiology*, *31*, 104-110. doi: https://doi.org/10.1016/j.conb.2014.09.003

Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. *Nature Neuroscience*, *7*, 446–451.

Buzsáki, G. (2005). Theta rhythm of navigation: Link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus*, *15*, 827–840.

Buzsáki, G. (2015, sep). Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, *25*(10), 1073–1188. Retrieved from `https://doi.org/10.1002%2Fhipo.22488` doi: 10.1002/hipo.22488

Buzsaki, G., & Moser, E. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, *16*(2), 130–138.

Calkins, M. W. (1896). Association: An essay analytic and experimental. *Psychological Review Monographs Supplement*, *1*(2). Retrieved from `http://psychclassics.yorku.ca/Calkins/Assoc/`

Caplan, J. B., Madsen, J. R., Raghavachari, S., & Kahana, M. J. (2001). Distinct patterns of brain oscillations underlie two basic parameters of human maze learning. *Journal of Neurophysiology*, *86*,

368–380.

Caplan, J. B., Madsen, J. R., Schulze-Bonhage, A., Aschenbrenner-Scheibe, R., Newman, E. L., & Kahana, M. J. (2003). Human theta oscillations related to sensorimotor integration and spatial learning. *Journal of Neuroscience*, *23*(11), 4726–4736.

Carr, H. A. (1931). The laws of association. *Psychological Review*, *38*, 212-228.

Clarke, S., & Hall, P. (2009). Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, 332–358. doi: 10.1214/07-aos557

Clemens, Z., Weiss, B., Szűcs, A., Erőss, L., Rásonyi, G., & Halász, P. (2009). Phase coupling between rhythmic slow activity and gamma characterizes mesiotemporal rapid-eye-movement sleep in humans. *Neuroscience*, *163*(1), 388–396.

Cohen, M. X. (2019). A better way to define and describe morlet wavelets for time-frequency analysis. *NeuroImage*, *199*, 81–86.

Colgin, L. L. (2015). Theta–gamma coupling in the entorhinal–hippocampal system. *Current opinion in neurobiology*, *31*, 45–50.

Cornwell, B., Johnson, L., Holroyd, T., Carver, F., & Grillon, C. (2008). Human hippocampal and parahippocampal theta during goal-directed spatial navigation predicts performance on a virtual Morris water maze. *Journal of Neuroscience*, *28*(23), 5983–5990.

Crone, N. E., Sinai, A., & Korzeniewska, A. (2006). High-frequency gamma oscillations and human brain mapping with electrocorticography. *Progress in Brain Research*, *159*, 275 - 295.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Curran, T. (1999). The electrophysiology of incidental and intentional retrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia*, *37*, 771–785.

Dalal, S., Baillet, S., Adam, C., Ducorps, A., Schwartz, D., Jerbi, K., … Lachaux, J. (2009). Simultaneous MEG and intracranial EEG recordings during attentive reading. *Neuroimage*, *45*(4), 1289–1304.

Davis, P. A. (1939). Effects of acoustic stimuli on the waking human brain. *Journal of Neurophysiology*, *2*, 494–499.

de Araujo, D. B., Baffa, O., & Wakai, R. T. (2002). Theta oscillations and human navigation: A magnetoencephalography study. *Journal of Cognitive Neuroscience*, *14*(1), 70–78.

Donchin, E., Ritter, W., & McCallum, W. C. (1979). Cognitive psychophysiology: The endogenous components of the ERP. In *Event-related brain potentials in man (behavioral biology series)* (pp. 349–412). Academic Press Inc. Retrieved from `https://www.amazon.com/Event-related-Brain-Potentials`

-Behavioral-biology/dp/0121551504?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0121551504

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 71–103. doi: 10.1214/ss/1056397487

Dykstra, A. R., Chan, A. M., Quinn, B. T., Zepeda, R., Keller, C. J., Cormier, J., ... Cash, S. S. (2012). Individualized localization and cortical surface-based registration of intracranial electrodes. *NeuroImage*, *59*(4), 3563–3570.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ekstrom, A. D., Caplan, J., Ho, E., Shattuck, K., Fried, I., & Kahana, M. (2005). Human hippocampal theta activity during virtual navigation. *Hippocampus*, *15*, 881–889.

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*, 184–187.

Ekstrom, A. D., Meltzer, J., McNaughton, B., & Barnes, C. (2001). NMDA receptor antagonism blocks experience-dependent expansion of hippocampal place fields. *Neuron*, *31*, 631–638.

Ekstrom, A. D., Suthana, N., Millett, D., Fried, I., & Bookheimer, S. (2009). Correlation between BOLD fMRI and theta-band local field potentials in the human hippocampal area. *Journal of Neurophysiology*, *101*(5), 2668.

Epstein, R. A. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*, 954–978.

Epstein, R. A., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*, 598–601.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94-107.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*, 145-154. doi: 10.1037/h0048509

Estes, W. K. (1959). *Component and pattern models with markovian interpretations. studies in mathematical learning theory*. Stanford, CA: Stanford University Press.

Ezzyat, Y., Kragel, J. E., Burke, J. F., Levy, D. F., Lyalenko, A., Wanda, P. A., ... Kahana, M. J. (2017). Direct brain stimulation modulates encoding states and memory performance in humans. *Current Biology*, *27*(9), 1251–1258. doi: 10.1016/j.cub.2017.03.028

Fawcett, T. (2006, jun). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. Retrieved from `https://doi.org/10.1016%2Fj.patrec.2005.10.010` doi: 10.1016/j.patrec.2005.10.010

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., ... others (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, *14*(1), 11–22.

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge, England: Cambridge University Press.

Foster, J. J., Vogel, E. K., & Awh, E. (in press). Oxford handbook of human memory. In M. J. Kahana & A. D. Wagner (Eds.), (2nd ed., chap. Working Memory as Persistent Neural Activity). Oxford, U. K.: Oxford University Press.

Frank, L. M., Stanley, G., & Brown, E. (2004). Hippocampal plasticity across multiple days of exposure to novel environments. *Journal of Neuroscience*, *24*(35), 7681–7689.

Frank, M., Samanta, J., Moustafa, A., & Sherman, S. (2007). Hold your horses: Impulsivity , deep brain stimulation, and medication in parkinsonism. *Science*, *318*, 1309–1312.

Fried, I., Wilson, C., Maidment, N., Engel, J. J., Behnke, E., Fields, T., ... Ackerson, L. (1999). Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. *Journal of Neurosurgery*, *91*, 697–705.

Fries, P., Nikolić, D., & Singer, W. (2007). The gamma cycle. *Trends in Neurosciences*, *30*(7), 309–316.

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., & Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, *3*, 96–101.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f noise in human cognition. *Science*, *267*(5205), 1837–1839.

Givens, B. (1996). Stimulus-evoked resetting of the dentate theta rhythm: relation to working memory. *Neuroreport*, *8*, 159-163.

Givens, B. S., & Olton, D. S. (1990). Cholinergic and GABAergic modulation of medial septal area: effect on working memory. *Behavioral Neuroscience*, *104*, 849-55.

Glaser, E. M., & Ruchkin, D. S. (1976). *Principles of neurobioligical signal analysis*. New York: Academic Press.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*, 1-16.

Gloor, P., Olivier, A., Quesney, L., Andermann, F., & Horowitz, S. (1982). The role of the limbic system in experiential phenomena of temporal lobe epilepsy. *Annals of Neurology*, *12*(2), 129–144.

Gluckman, B. J., Nguyen, H., Weinstein, S. L., & Schiff, S. (2001).

Adaptive electric field control of epileptic seizures. , *21*(2), 590-600.

Gold, C., Henze, D., Koch, C., & Buzsaki, G. (2006). On the Origin of the Extracellular Action Potential Waveform: A Modeling Study. *Journal of Neurophysiology*, *95*(5), 3113–3128.

Gomulicki, B. R. (1953). The development and present stat us of the trace t heor y of memory. *British Journal of Psychology Monograph Supplements*, *29*, 1-91.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: Wiley.

Greenberg, J. A., Burke, J. F., Haque, R., Kahana, M. J., & Zaghloul, K. A. (2015). Decreases in theta and increases in high frequency activity underlie associative memory encoding. *NeuroImage*, *114*, 257–263. doi: 10.1016/j.neuroimage.2015.03.077

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. doi: 10.1111/j.1469-8986.2011.01273.x

Guthrie, E. R. (1935). *The psychology of learning*. Harper.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*, 801–806. doi: 10.1038/nature03721

Halgren, E., Walter, R., Cherlow, D., & Crandall, P. (1978). Mental phenomena evoked by electrical stimulation of the human hippocampal formation and amygdala. *Brain*, *101*(1), 83.

Halgren, E., Wilson, C. L., & Stapleton, J. M. (1985, July). Human medial temporal-lobe stimulation disrupts both formation and retrieval of recent memories. *Brain and Cognition*, *4*(3), 287–95.

Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. , *14*, 793–817.

Hasselmo, M. E., & Eichenbaum, H. (2005). Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural Networks*, *18*(9), 1172–1190. doi: 10.1016/j.neunet.2005.08.007

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer. Retrieved from `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425-2429.

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed

by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*(2), 575–596. doi: 10.1037/a0033715

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, *26*(3), 699–720. doi: 10.3758/s13423-018-1537-3

Hebb, D. O. (1949). *Organization of behavior*. New York: Wiley.

Herweg, N. A., Solomon, E. A., & Kahana, M. J. (2020). Theta oscillations in human memory. *Trends in Cognitive Science*, *24*(3), 208-227. doi: https://doi.org/10.1016/j.tics.2019.12.006

Hillyard, S. A., & Kutas, M. (1983). Electrophysiology of cognitive processing. *Annual Review of Psychology*, *34*, 33–61.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The psychology of learning and memory* (p. 47-91). New York: Academic Press.

Hintzman, D. L. (2003). Robert hooke's model of memory. *Psychonomic Bulletin & Review*, *10(1)*, 3-14.

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth". *Perspectives on Psychological Perspectives on Psychological Science*, *6*(3), 253-271.

Hok, V., Save, E., Lenck-Santini, P. P., & Poucet, B. (2005). Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proceedings of the National Academy of Sciences, USA*, *102*(12), 4602–4607.

Hollingsworth, H. L. (1928). *Psychology: Its facts and principles*. D. Appleton and Company.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Hölscher, C., Anwyl, R., & Rowan, M. J. (1997). Stimulation on the positive phase of hippocampal theta rhythm induces long-term potentiation that can be depotentiated by stimulation on the negative phase in area CA1 *in vivo*. *Journal of Neuroscience*, *17*, 6470–6477.

Hooke, R. (1969). *The posthumous works of Robert Hooke: With a new introduction by Richard S. Westfall*. Johnson Reprint Corp.

Howard, M. W., Addis, K. A., Jing, B., & Kahana, M. J. (2007). Semantic structure and episodic memory. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 121–141). Mahwah, NJ: Laurence Erlbaum and Associates.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923–941. doi:

10.1037/0278-7393.25.4.923

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299. doi: 10.1006/jmps.2001.1388

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*(1), 85–98. doi: 10.1006/jmla.2001.2798

Howard, M. W., Venkatadass, V., Norman, K. A., & Kahana, M. J. (2007). Associative processes in immediate recency. *Memory & Cognition*, *35*(7), 1700–1711. doi: 10.3758/BF03193503

Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental Psychology: General*, *137*(2), 324-347.

Huerta, P. T., & Lisman, J. E. (1993). Heightened synaptic plasticity of hippocampal CA1 neurons during a cholinergically induced rhythmic state. *Nature*, *364*(6439), 723–725.

Jacobs, J., Hwang, G., Curran, T., & Kahana, M. J. (2006). EEG oscillations and recognition memory: Theta correlates of memory retrieval and decision making. *NeuroImage*, *15*(2), 978–87. doi: https://doi.org/10.1016/j.neuroimage.2006.02.018

Jacobs, J., Kahana, M. J., Ekstrom, A. D., & Fried, I. (2007). Brain oscillations control timing of single-neuron activity in humans. *Journal of Neuroscience*, *27*(14), 3839–3844.

Jacobs, J., Kahana, M. J., Ekstrom, A. D., Mollison, M. V., & Fried, I. (2010). A sense of direction in human entorhinal cortex. *Proceedings of the National Academy of Sciences*, *107*(14), 6487–6482.

Jacobs, J., Korolev, I., Caplan, J., Ekstrom, A., Litt, B., Baltuch, G., ... Kahana, M. (2010). Right-lateralized brain oscillations in human spatial navigation. *Journal of Cognitive Neuroscience*, *22*(5), 824–836.

Jacobs, J., Lega, B., & Anderson, C. (2012). Explaining how brain stimulation can evoke memories. *Journal of Cognitive Neuroscience*, *24*(3), 553–563.

Jacobs, J., Miller, J., Lee, S. A., Coffey, T., Watrous, A. J., Sperling, M. R., ... Rizzuto, D. S. (2016, December). Direct electrical stimulation of the human entorhinal region and hippocampus impairs memory. *Neuron*, *92*(5), 983-990. doi: https://doi.org/10.1016/j.neuron.2016.10.062

Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X., ... Kahana, M. J. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, *16*(9), 1188–1190. doi: 10.1038/nn.3466

Jenkins, & Dallenbach, K. M. (1924). Oblivescence during sleep and waking. *The American Journal of Psychology*, *35*, 605–612.

Jensen, O., & Lisman, J. E. (1998). An oscillatory short-term memory buffer model can account for data on the Sternberg task. *J. Neuroscience*, *18*, 10688–10699.

Jensen, O., & Lisman, J. E. (2000). Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. *J. Neurophysiol.*, *83*, 2602–2609.

Jensen, O., & Lisman, J. E. (2005, Feb). Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends Neurosci*, *28*(2), 67–72. doi: 10.1016/j.tins.2004.12.001

Jerbi, K., Freyermuth, S., Dalal, S., Kahane, P., Bertrand, O., Berthoz, A., & Lachaux, J. (2009). Saccade related gamma-band activity in intracerebral EEG: dissociating neural from ocular muscle activity. *Brain Topography*, *22*(1), 18–23.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109. doi: 10.3758/BF03197276

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition*, *30*, 823–840.

Kahana, M. J. (2012). *Foundations of human memory*. New York, NY: Oxford University Press.

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(1), 107–138. doi: 10.1146/annurev-psych-010418-103358

Kahana, M. J., Aggarwal, E. V., & Phan, T. D. (2018). The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1857–1863. doi: 10.1037/xlm0000553

Kahana, M. J., & Bennett, P. J. (1994). Classification and perceived similarity of compound gratings that differ in relative spatial phase. *Perception & Psychophysics*, *55*(6), 642-656.

Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *Journal of Gerontology: Psychological Sciences*, *60*(2), 92–97. doi: 10.1093/geronb/60.2.P92

Kahana, M. J., Rizzuto, D. S., & Schneider, A. (2005). Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 933-953.

Kahana, M. J., Seelig, D., & Madsen, J. R. (2001). Theta returns. *Current Opinion in Neurobiology*, *11*, 739–744.

Kahana, M. J., Sekuler, R., Caplan, J. B., Kirschen, M., & Madsen, J. R. (1999). Human theta oscillations exhibit task dependence during virtual maze navigation. *Nature*, *399*, 781–784.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141 - 151.

Kanisza, G. (1979). *The organization of vision*. New York: Praeger.

Karpicke, J. D., & Roediger, H. L. (2008, February). The critical importance of retrieval for learning. *Science*, *319*, 966-968.

Katona, G. (1951). *Psychological analysis of economic behavior*. New York: McGraw-Hill.

Keller, F. S., & Schoenfeld, W. N. (1950). *Principles of psychology: A systematic text in the science of behavior*. New York: Appleton-Century-Crofts.

Kilner, J. M. (2013). Bias in a common eeg and meg statistical analysis and how to avoid it. *Clinical Neurophysiology*, *124*(10), 2062–2063.

Kirschen, M. P., Kahana, M. J., Sekuler, R., & Burack, B. (2000). Optic flow helps humans learn to navigate through synthetic environments. *Perception*, *29*, 801–818.

Klausberger, T., Magill, P. J., Marton, L. F., Roberts, J. D., Cobden, P. M., Buzsáki, G., & Somogyi, P. (2003). Brain-state- and cell-type-specific firing of hippocampal interneurons in vivo. *Nature*, *421*, 844–848.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace and World.

Köhler, W. (1947). *Gestalt psychology: The definitive statement of gestalt theory*. New York: Liveright.

Kreiman, G., Koch, C., & Fried, I. (2000). Imagery neurons in the human brain. *Nature*, *408*, 357–360.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*, 1126–1141.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. doi: 10.1038/nn.2303

Lachaux, J. P., Rudrauf, D., & Kahane, P. (2003). Intracranial EEG and human brain mapping. *Journal of Physiology-Paris*, *97*(4–6), 613–628.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lega, B., Burke, J. F., Jacobs, J., & Kahana, M. J. (2015). Slow theta-to-gamma phase amplitude coupling in human hippocampus

supports the formation of new episodic memories. *Cerebral Cortex*, *26*(1), 268-278.

Leutgeb, S., & Mizumori, S. J. (1999, Aug). Excitotoxic septal lesions result in spatial memory deficits and altered flexibility of hippocampal single-unit representations. *J Neurosci*, *19*(15), 6661-6672.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. doi: 10.3758/BF03210951

Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *Journal of Neuroscience*, *23*(10), 3963 - 3971.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337–363. doi: 10.1037/a0039036

Long, N. M., & Kahana, M. J. (2017). Modulation of task demands suggests that semantic processing interferes with the formation of episodic associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 167–176. doi: 10.1037/xlm0000300

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). The MIT Press. Retrieved from `http://www.ebook.de/de/product/21836692/steven_j_luck_an_introduction_to_the_event_related_potential_technique.html`

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, *35*, 13-39.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior*, *8*, 828-835.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384. doi: 10.1016/j.cogpsych.2008.02.004

Mandler, G., Rabinowitz, J. C., & Simon, R. A. (1981). Coordinate organization: The holistic representation of word pairs. *American Journal of Psychology*, *92*, 209-222.

Manning, J. R., Jacobs, J., Fried, I., & Kahana, M. J. (2009). Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *Journal of Neuroscience*, *29*(43), 13613–13620. doi: 10.1523/JNEUROSCI.2041-09.2009

Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, *108*(31), 12893–12897. doi: 10.1073/

pnas.1015174108

Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., & Kahana, M. J. (2012). Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clustering during memory search. *Journal of Neuroscience*, *32*(26), 8871–8878. doi: 10.1523/JNEUROSCI.5321-11.2012

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, *275*(5297), 213–215.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co., Inc. New York, NY, USA.

McCullough, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115-133.

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*, 352-70. doi: 10.1037/h0069819

McGeoch, J. A. (1942). *The psychology of human learning: An introduction*. New York: Longmans.

McGeoch, J. A., & Irion, A. L. (1952). *The psychology of human learning, 2nd edition*. New York: Longmans, Green and Co.

McNaughton, B. L., & Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, *10*, 408-415.

Meisler, S. L., Kahana, M. J., & Ezzyat, Y. (2019). Does data cleaning improve brain state classification? *Journal of Neuroscience Methods*, *328*. doi: https://doi.org/10.1016/j.jneumeth.2019.108421

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Miller, J., Polyn, S., & Kahana, M. (2007). Clustering by spatial proximity during memory search. Society for Mathematical Psychology conference. Irvine, CA.

Miller, J. F., Lazarus, E., Polyn, S. M., & Kahana, M. J. (2013). Spatial clustering during memory search. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*(3), 773–781.

Miller, J. F., Watrous, A. J., Tsitsiklis, M., Lee, S. A., Sheth, S. A., Schevon, C. A., ... Jacobs, J. (2018). Lateralized hippocampal oscillations underlie distinct aspects of human spatial memory and navigation. *Nature communications*, *9*(1), 2423.

Miller, K. J., Honey, C. J., Hermes, D., Rao, R. P., den Nijs, M., &

Ojemann, J. G. (2014, January). Broadband changes in the cortical surface potential track activation of functionally diverse neuronal populations. *NeuroImage*, *85*, 711–720.

Miller, K. J., Sorensen, L. B., Ojemann, J. G., den Nijs, M., & Sporns, O. (2009). Power-law scaling in the brain surface electric potential. *PLoS Computational Biology*, *5*(12).

Millett, D. (2001). Hans Berger: From Psychic Energy to the EEG. In *Perspectives in Biology and Medicine* (Vol. 44, p. 522-542). The Johns Hopkins University Press. doi: 10.1353/pbm.2001.0070

Misra, A., Burke, J., Ramayya, A., Jacobs, J., Sperling, M., Moxon, K., ... Sharan, A. (2014). Methods for implantation of micro-wire bundles and optimization of single/multi-unit recordings from human mesial temporal lobe. *Journal of neural engineering*, *11*(2), 026013.

Moreton, B. J., & Ward, G. (2010). Time scale similarity and long-term memory for autobiographical events. *Psychonomic Bulletin & Review*, *17*(4), 510–515.

Moscovitch, M., Nadel, L., Winocur, G., Gilboa, A., & Rosenbaum, R. S. (2006). The cognitive neuroscience of remote episodic, semantic and spatial memory. *Current Opinion in Neurobiology. Special Issue: Cognitive neuroscience.*, *16*, 179-190.

Moxon, K. A., & Nicolelis, M. A. L. (1999). Multiple-site recording electrodes. In *Methods Simultaneous Neuronal Ensemble Recordings* (25th - 45th ed.). Boca Raton, FL: CRC Press.

Müller, G. E., & Pilzecker, A. (1900). Experimental contributions to memory theory. *Zeitschrift für Psychologie Eganzungsband*, *1*, 1-300.

Müller, G. E., & Schumann, F. (1894). Experimentelle beiträge zur untersuchung des gedächtnisses. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, *6*(80-191), 257–339.

Muller, R., Bostock, E., Taube, J., & Kubie, J. (1994). On the directional firing properties of hippocampal place cells. *Journal of Neuroscience*, *14*, 7235.

Murdock, B. B. (1967). Recent developments in short-term memory. *British Journal of Psychology*, *58*(3-4), 421-433. doi: 10.1111/j.2044-8295.1967.tb01099.x

Murdock, B. B. (1972). Short-term memory. In G. H. Bower (Ed.), *The psychology of learning and motivation:advances in reasearch and theory.* (Vol. 5, p. 67-127). New York: Academic Press.

Musha, T., & Higuchi, H. (1976). The 1/f fluctuation of a traffic current on an expressway. *Japanese Journal of Applied Physics*, *15*(7), 1271.

Newman, E. L., Caplan, J. B., Kirschen, M. P., Korolev, I. O., Sekuler, R., & Kahana, M. J. (2007). Learning your way around town:

How virtual taxicab drivers learn to use both layout and landmark information. *Cognition*, *104*(2), 231–253.

Newman, S. E. (1987). Ebbinghaus' "on memory": Some effects on early american research. In D. S. Gorfein & R. R. Hoffman (Eds.), *Memory and learning: The ebbinghaus centennial conference* (p. 77-87). Hillsdale, NJ, England: Lawrence Erlbaum and Associates.

Nichols, T. E., & Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*, 1–25.

Niedermeyer, E. (2008). Hippocampic theta rhythm. *Clinical EEG and neuroscience: official journal of the EEG and Clinical Neuroscience Society (ENCS)*, *39*(4), 191.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I., & Malach, R. (2007). Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Current Biology*, *17*(15), 1275–1285.

Nisar, H., & Yeap, K. H. (2014). Introduction. In N. Kamel & A. S. Malik (Eds.), *EEG/ERP analysis: Methods and applications* (pp. 1–20). CRC Press. Retrieved from `https://www.amazon.com/EEG-ERP-Analysis-Methods-Applications-ebook/dp/B00OKUG384?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=B00OKUG384`

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*, *110*, 611–646.

Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (p. 363-394). Hillsdale, New Jersey: Lawrence Erlbaum and Associates.

Nyhus, E., & Curran, T. (2010). Functional role of gamma and theta oscillations in episodic memory. *Neuroscience & Biobehavioral Reviews*, *34*(7), 1023-1035.

Ojemann, G. (1991). Cortical organization of language. *Journal of Neuroscience*, *11*(8), 2281.

O'Keefe, J., & Burgess, N. (1999). Theta activity, virtual navigation and the human hippocampus. *Trends Cogn Sci*, *3*(11), 403-406.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*, 171–175.

O'Keefe, J., & Recce, M. L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*,

3, 317–30.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.

Pal, N. R., & Pal, S. K. (1993). A review of image segmentation techniques. *Pattern Recognition Letters*, *26*, 1277 - 1294.

Penfield, W., & Perot, P. (1963). The brain's record of auditory and visual experience. *Brain*, *86*(4), 595–696.

Perez, V. B., & Vogel, E. K. (2012). What ERPs can tel us about working memory. In S. J. Luck & E. S. Kappenman (Eds.), *The oxford handbook of event-related potential components* (pp. 361–372). New York: Oxford University Press. doi: 10.1093/oxfordhb/9780195374148.013.0180

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. doi: 10.1037/a0014420

Postle, B. R., & Oberauer, K. (in press). Oxford handbook of human memory. In M. J. Kahana & A. D. Wagner (Eds.), (2nd ed., chap. Working Memory). Oxford, U. K.: Oxford University Press.

Qasim, S. E., Miller, J. F., Inman, C., Gross, R. E., Willie, J., Bradley, L., ... Joshua, J. (2019). Memory retrieval modulates spatial tuning of single neurons in the human entorhinal cortex. *Nature Neuroscience*, 1-9.

Quiroga, R. Q., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, *16*, 1661–1687.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(23), 1102–1107.

Robinson, E. S. (1932). *Association theory to-day; an essay in systematic psychology*. New York: The Century Co.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of Learning and Motivation: Cognition in Education* (Vol. 55, pp. 1–36). Oxford: Elsevier.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*, 28-34.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Pyschology: Learning, Memory, and Cognition*, *25(5)*, 1161-1176.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*, 734-760.

Rudoler, J. H., Herweg, N. A., & Kahana, M. J. (2021). Oscillatory and fractal biomarkers of human memory. In *Context and Episodic Memory Symposium.* Philadelphia, PA.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends Cogn Sci*, *11*(6), 251–257.

Rumelhart, D., McClelland, J., & the PDP Research Group. (1986). *Parallel distributed processing*. MIT Press.

Rundus, D. (1971). An analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*, 63-77. doi: 10.1037/h0031185

Schacter, D. (2001). *Forgotten ideas, neglected pioneers: Richard semon and the story of memory*. New York, NY: US: Psychology Press.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgreery, and Psychiatry*, *20*, 11-21.

Semon, R. W. (1923). *Mnemic psychology (b. duffy, trans.)*. London: George Allen & Unwin (Original work published 1909).

Siapas, A., Lubenov, E., & Wilson, M. (2005). Prefrontal phase locking to hippocampal theta oscillations. *Neuron*, *46*, 141–151.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci*, *18*, 555–586. doi: 10.1146/annurev.ne.18.030195.003011

Singh, S. N., Mishra, S., Bendapudi, N., & Linville, D. (1994). Enhancing memory of television commercials through message spacing. *Journal of Marketing Research*, *31*, 384-392.

Skaggs, W. E., McNaughton, B. L., Wilson, M. A., & Barnes, C. A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, *6*, 149–172.

Smith, S. M. (1988). Environmental context-dependent memory. In G. M. Davies & D. M. Thomson (Eds.), *Memory in context: Context in memory.* (p. 13-34). Oxford, England: John Wiley & Sons.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652–654.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? an introduction to receiver operating characteristics curves. *The Canadian Journal of Psychiatry*, *52*, 121–128. doi: 10.1177/070674370705200210

Suthana, N., Haneef, Z., Stern, J., Mukamel, R., Behnke, E., Knowlton, B., & Fried, I. (2012). Memory enhancement and deep-brain stimulation of the entorhinal area. *The New England Journal of Medicine*, *366*, 502–10.

Tan, L., & Ward, G. (2000). A recency-based account of the primacy

effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1589–1625. doi: 10.1037/0278-7393.26.6.1589

Thorndike, E. L. (1932). *The fundamentals of learning*. New York: Bureau of Publications, Teachers College.

Towle, V., Yoon, H., Castelle, M., Edgar, J., Biassou, N., Frim, D., ... Kohrman, M. (2008). ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. *Brain*, *131*(8), 2013.

Tsitsiklis, M., Miller, J., Qasim, S. E., Inman, C. S., Gross, R. E., Willie, J. T., ... Jacobs, J. (2020). Single-neuron representations of spatial targets in humans. *Current Biology*, *30*(2), 245–253. doi: 10.1016/j.cub.2019.11.048

Tulving, E., & Arbuckle, T. Y. (1963). Sources of intratrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *1*, 321-334.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology*, *21*, 437-484.

Uitvlugt, M. G., & Healey, M. K. (2019). Temporal proximity links unrelated news events in memory. *Psychological Science*, *30*(1), 92–104. doi: 10.1177/0956797618808474

Umbach, G., Kantak, P., Jacobs, J., Kahana, M. J., Pfeiffer, B. E., Sperling, M., & Lega, B. (2020). Time cells in the human hippocampus and entorhinal cortex support episodic memory. *PNAS*, *117*(45), 28463-28474. doi: https://doi.org/10.1073/pnas.2013250117

van Boxtel, G. J. M. (1998). Computational and statistical methods for analyzing event-related potential data. *Behavior Research Methods, Instruments, & Computers*, *30*, 87–102. doi: 10.3758/bf03209419

Vanderwolf, C. (1969). Hippocampal electrical activity and voluntary movement of the rat. *Electroencephalography and Clinical Neurophysiology*, *26*, 407–418.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*, 748–751.

von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Report No. 81-2). Göttingen, Germany: Max-Planck-Institute for Biophysical Chemistry.

Voss, R. F. (1992). Evolution of long-range fractal correlations and 1/f noise in dna base sequences. *Physical review letters*, *68*(25), 3805.

Wachter, J. A., & Kahana, M. J. (Submitted). A retrieved-context theory of financial decisions. *Submitted*. doi: 10.3386/w26200

Watrous, A. J., Fried, I., & Ekstrom, A. D. (2011). Behavioral corre-

lates of human hippocampal delta and theta oscillations during navigation. *Journal of Neurophysiology*, *105*(4), 1747–1755.

Weidemann, C. T., & Kahana, M. J. (2021). Neural measures of subsequent memory reflect endogenous variability in cognitive function. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(4), 641–651. doi: 10.1037/xlm0000966

Weidemann, C. T., Kragel, J. E., Lega, B. C., Worrell, G. A., Sperling, M. R., Sharan, A. D., ... Kahana, M. J. (2019). Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *Journal of Experimental Psychology: General*, *148*(1), 1–12. doi: 10.1037/xge0000480

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). John Wiley & Sons. doi: 10.2307/1270279

Wilding, E. L., & Ranganath, C. (2012). Electrophysiological correlates of episodic memory processes. In S. J. Luck & E. S. Kappenman (Eds.), *The oxford handbook of event-related potential components.* New York: Oxford University Press. doi: 10.1093/oxfordhb/9780195374148.013.0187

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*, 409-415.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, *72*, 2031–2046. doi: 10.3758/bf03196680

Yaffe, R. B., Kerr, M. S., Damera, S., Sarma, S. V., Inati, S. K., & Zaghloul, K. A. (2014). Reinstatement of distributed cortical oscillations occurs with precise spatiotemporal dynamics during successful memory retrieval. *Proceedings of the National Academy of Sciences*, *111*(52), 18727–8732.

Yates, F. A. (1966). *The art of memory.* London, England: Routledge and Kegan Paul.

Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Pnanning and Inference*, *82*, 171–196. doi: 10.1016/s0378-3758(99)00041-5

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441-517.

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*, 1178–1194. doi: 10.1002/hipo.20864

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje,

E. C., ... Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping*, *36*(1), 258–287.

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2009). Human Substantia Nigra neurons encode unexpected financial rewards. *Science*, *323*, 1496–1499.

Zaghloul, K. A., Weidemann, C. T., Lega, B. C., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2012). Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. *The Journal of Neuroscience*, *32*(7), 2453–2460.

Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., & Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 792–804. doi: 10.1037/0278-7393.32.4.792

Index