Chapter 41: Identifying stimulus-driven neural activity patterns in multi-patient intracranial recordings

Jeremy R. Manning
Dartmouth College, Hanover, NH, USA
jeremy.r.manning@dartmouth.edu

Note: This chapter is forthcoming in Intracranial EEG for Cognitive Neuroscience
Draft current as of February 7, 2022

Abstract. Identifying stimulus-driven neural activity patterns is critical for studying the neural basis of cognition. This can be particularly challenging in intracranial datasets, where electrode locations typically vary across patients. This chapter first presents an overview of the major challenges to identifying stimulus-driven neural activity patterns in the general case. Next, we will review several modality-specific considerations and approaches, along with a discussion of several issues that are particular to intracranial recordings. Against this backdrop, we will consider a variety of within-subject and across-subject approaches to identifying and modeling stimulus-driven neural activity patterns in multi-patient intracranial recordings. These approaches include generalized linear models, multivariate pattern analysis, representational similarity analysis, joint stimulus-activity models, hierarchical matrix factorization models, Gaussian process models, geometric alignment models, inter-subject correlations, and inter-subject functional correlations. Examples from the recent literature serve to illustrate the major concepts and provide the conceptual intuitions for each approach.

Keywords: stimulus-driven · multi-subject · signal processing · computational models · dynamics

41.1 Overview

Studying brain function often requires identifying brain responses to a given stimulus or set of stimuli. For some stimuli, and for some systems, this identification problem is relatively straightforward. For example, when a photopigment in a retinal photoreceptor absorbs light, this triggers a cascade of responses that is ultimately sent from the retina to other brain areas via the optic nerve [71]. In the general sense, however (i.e., for arbitrarily complex stimuli and arbitrary brain areas), the problem of identifying neural responses to known (or unknown) stimuli can be incredibly challenging [73].
41.1.1 Why is it challenging to identify stimulus-driven brain activity?

To illustrate the enormity of the challenge of identifying stimulus-driven brain responses in the general sense, it can be useful to start by considering what form a complete solution might take. First, we need some means of defining (and measuring) what brain activity is. For example, should we concern ourselves with measuring membrane potentials or firing rates of individual neurons? Or population activity in a given brain structure or network? And is it more appropriate to analyze or interpret activity patterns in the time domain (e.g., firing rate or voltage as a function of time), or in the frequency domain (e.g., characterizing the signal through the relative contributions of its constituent sinusoidal components at different frequencies)? Should we consider neurons and/or brain structures in isolation, or should we instead interpret each “unit” of activity within the context of the network(s) it participates in or contributes to? We discuss several different approaches to these questions (and their relative trade-offs) in Section 41.1.2.

Second, we need some means of characterizing (and ideally, quantifying) the stimulus itself. For a simple stimulus, such as a single photon of light, emitted from a known location in an otherwise completely dark room, constructing a sufficiently comprehensive model of the stimulus might be straightforward—and perhaps even trivial. For other stimuli, such as real-world experiences, constructing a comprehensive model of “what is happening” can be highly complex (at best). Essentially, building a stimulus model entails quantifying how different features, or properties, change over time. In our single-photon example, we might represent the stimulus as a timeseries of zeros (no photon present) and ones (photon present). For more complex stimuli, however, it may not even be clear what the features are. We discuss considerations and approaches to building explicit stimulus models in Section 41.1.3.

Characterizing brain activity and building stimulus models are each complex challenges in their own right. Linking the two provides its own set of additional challenges. We discuss these issues in Section 41.2.

41.1.2 How can we measure neural “activity” in the human brain?

The brain is a complex organ comprising myriad cell types that interconnect to form a vast network. When neuroscientists use the term brain activity, this can refer to a variety of possible physical and physiological phenomena. To contextualize what brain activity means, it can be helpful to first consider the brain’s structure and function.

Neurons are the dominant cell type in the brain; the adult human brain contains roughly 100 billion neurons. The cerebral cortex, commonly associated with high-level brain function, comprises roughly 80% of the adult human brain’s mass, but only roughly 20% of its neurons [59]. In general, brain activity refers to changes in cellular processes that neurons undergo. These processes can take many forms, including: rapid changes in membrane voltage that result in neurotransmitter release, called action potentials [63]; sub-threshold
After considering the various forms that brain (neural) activity can take, it can also be useful to define a relevant spatiotemporal scale (Fig. 1). For example, a biologist concerned with the structure and function of individual ion channels embedded in the neuronal cell membrane may be most interested in processes that happen over a span of picoseconds or nanoseconds (e.g., the amount of time it takes for an ion to pass through an ion channel, or the amount of time it takes for an ion channel to change its conformation). They may also be most interested in spatial resolutions on the orders of angstroms (e.g., the approximate scale of an individual ion channel). At another extreme, a neuroanatomist studying the comparative anatomy of different species of primates might be most interested in timescales on the order of decades (e.g., an entire lifetime) and spatial scales on the order of decimeters (e.g., the size of an adult brain). As summarized in Figure 1, different neuroimaging approaches are each associated with a range of temporal and spatial scales that they are best suited to measure.
If we are specifically interested in stimulus-driven neural activity, this implies focusing in on a limited range of spatiotemporal resolutions (gray shading in Fig. 1). Neuroimaging approaches that enable insights at those resolutions may provide particularly useful measures of stimulus-driven neural activity.

**Non-invasive approaches.** Neuroimaging approaches that rely on measurements taken using sensors that are placed without requiring surgery are referred to as **non-invasive**. In general, non-invasive neuroimaging entails placing one or more sensors on or near the subject’s head. Examples include **scalp electroencephalography** (EEG; i.e., recording voltages from small electrodes placed on the scalp); **magnetoencephalography** (MEG; i.e., measuring tiny changes in the magnetic field outside of the head caused by neuronal firing); and **functional magnetic resonance imaging** (fMRI; i.e., inferring changes in blood flow associated with neural activity using a powerful magnet placed around the head). A related approach, **magnetic resonance imaging** (MRI) uses strong magnetic fields, magnetic field gradients, and radio waves to produce a static anatomical image of the brain. Each of these approaches is widely used by neuroscientists interested in studying the neural basis of cognition and behavior. A benefit of relying on non-invasive neuroimaging is that these approaches are low-risk and may be safely used on healthy (non-patient) participants, and without the supervision of a physician. The main drawback of non-invasive neuroimaging is that, because these approaches all rely on sensors placed outside of the head, any relevant activity that is filtered out by the skull, or that is too weak to be measured from distant sensors, cannot be captured. This means that non-invasive neuroimaging approaches tend to have lower spatiotemporal resolution than invasive approaches.

**Invasive approaches.** Neuroimaging approaches that require surgery are referred to as **invasive**. Invasive in vivo techniques entail placing sensors directly on the surface of and/or in direct contact with deep structure inside of a living person’s brain. Examples include **intracranial electroencephalography** (iEEG; i.e., recording voltages from tiny wires implanted in the brain) and **electrocorticography** (ECoG; i.e., recording voltages from small electrodes lying directly on the brain’s cortical surface). Both iEEG and ECoG are similar to (non-invasive) EEG, in that all three approaches entail recording aggregate voltages from populations of many neurons. The key differences between these approaches are the locations and sizes of the electrodes. When sensors are larger and are placed far from the signal sources (i.e., neurons), as in EEG, the sensors pick up on relatively large populations of neurons that are spread over a large portion of the brain. When sensors are smaller and placed in direct contact with signal sources, as in ECoG and iEEG, the sensors pick up on smaller populations of neurons that are closer to the recording surface of the electrodes. When tiny **microwires** are used to generate iEEG recordings, it is even possible to record action potentials from individual neurons.
Invasive in vitro neuroimaging approaches entails taking measurements from brain slices or other structures that have been excised from an intact brain and placed in an isolated environment such as a petri dish. For example, patch clamp recordings use tiny glass micropipettes placed directly on the cell membrane to capture changes in membrane potential associated with the opening and closing of individual ion channels. In vitro approaches are not generally used to study stimulus-driven neural activity in humans, since excising the to-be-recorded tissues typically isolates the corresponding neurons from their sensory inputs. In vitro approaches also require destroying the to-be-recorded tissues, which presents ethical and safety concerns.

Because invasive neuroimaging approaches require physically cutting into the brain, they are not appropriate for use in non-patient populations. Rather, invasive recordings in humans are typically taken from neurosurgical patients who have their electrodes implanted as part of a treatment protocol. For example, people suffering from drug-resistant epilepsy may elect to undergo invasive monitoring (from implanted electrodes) in order to help neurologists localize the most likely source(s) of their seizures. During an extended hospital stay, the patients may elect to participate in research studies that are not directly related to their treatment, in the interest of advancing scientific knowledge by providing access to high quality recordings from their brain. Stimulus-driven responses in individual neurons or small circuits that unfold over sub-millisecond timescales can only be measured using invasive approaches like iEEG and ECoG. By providing measurements at both high spatial resolution and high temporal resolution, intracranial recordings can be ideally suited to studying stimulus-driven neural activity patterns (Fig. 1).

**Single-channel neural signals.** A single ECoG or iEEG electrode implanted in a patient’s brain measures changes in membrane potential (voltage) in individual neurons, other cells and signal sources, and populations of cells. Because neurons can most effectively transmit signals to other cells via action potentials, the timings of individual action potentials from a given neuron, or the firing rate of a neuron, can provide putative insights into that cell’s function (Fig. 2A). For example, if a neuron changes its firing rate when the patient is exposed to a particular stimulus or experience, this could suggest that the neuron plays some role in processing information pertaining to that stimulus or experience.

**Local field potentials** (LFPs) reflect the aggregate neuronal firing and sub-threshold changes in membrane potential across thousands of neurons near the recording surface of an electrode (Fig. 2B). When LFPs change during exposure to a stimulus or experience, this can suggest that the underlying population of neurons plays some role in processing that stimulus or experience. These changes may be aperiodic, as in Figure 2B, or sinusoidal, as in Figure 2C. Rhythmic (sinusoidal or periodic) changes in the LFP tend to reflect coordinated firing patterns across neurons in the population, whereas uncoordinated changes are reflected as changes in the volatility of the LFP [22, 36, 46, 94].
**Fig. 2. Measuring and processing single-channel neural signal.** All of the examples shown in this figure are constructed using simulated data. **A–D. Neural signals.** The illustrated examples show voltage (y-axis, arbitrary units) as a function of time (x-axis, arbitrary units).

**A. Neuronal firing.** The vertical lines illustrate the times at which an artificial neuron fired action potentials (spikes). The smooth curve shows the timeseries of firing rates, computed by convolving the spike timeseries with a Gaussian kernel.

**B. Local field potentials (LFPs).** LFPs reflect the aggregate neuronal firing and sub-threshold changes in membrane potential across thousands of neurons near the recording surface of an electrode.

**C. Oscillations.** When the local field potential exhibits sinusoidal fluctuations, this can reflect coordinated changes in membrane potential across a population of neurons.

**D. Phase coding.** The timing of an individual neuron’s action potentials with respect to oscillations in the local field potential can code information via the phase (angle) relative to a sine wave at the oscillation’s frequency.

**E. Spike-triggered average.** Phase coding may be identified by sampling the LFP before and after each spike and then averaging across all spikes. The spike-triggered average in this panel is computed using the artificial LFP displayed in Panel D.

**F. Power spectrum and broadband power.** Oscillatory contributions to the local field potential may be summarized as a power spectrum that shows the extent to which oscillations at each frequency contribute to the LFP. The power spectrum in this panel is computed using the artificial LFP displayed in Panel C. The underlying height of the power spectrum, called broadband power, may be estimated by using robust regression to fit a line to the power spectrum in log-log space. The area under the robust fit line may be used to estimate the firing rates of neurons in the underlying population [94].
When both spike timing information and LFP recordings are available, it is possible to examine whether a given neuron’s spikes are modulated according to the activity of the surrounding population. For example, phase-locked neurons tend to fire action potentials during a particular phase (angle) of an oscillation that appears in an LFP (Fig. 2D). Other neurons exhibit phase coding by changing their preferred phase according to properties of the stimulus, ongoing experience, or behavior. For these neurons, the phases (of LFP oscillations) at which spikes occur can carry additional information beyond firing rate alone. One way of characterizing phase-depending firing is to compute a spike-triggered average of the LFP preceding and proceeding each spike (Fig. 2E). When a neuron exhibits a phase preference, its spike-triggered average will look like an oscillation centered at the neuron’s preferred phase.

Although oscillations can sometimes be detected visually by examining a raw LFP recording, a more rigorous approach is to use signal processing methods to quantify the presence of oscillatory components of the LFP. A power spectrum (Fig. 2F) plots the power at each frequency—i.e., the extent to which oscillations at each frequency contribute to the LFP. When the LFP exhibits an oscillation, this appears as a peak (centered on the oscillation’s frequency) in the LFP’s power spectrum. An LFP may also exhibit multiple oscillations, which appear as multiple peaks in the power spectrum.

In addition to true (sinusoidal) oscillations in the LFP, the volatility of the LFP can also change its power spectrum. For example, an increase in the standard deviation of the LFP’s changes in voltages across successive timepoints will result in an increase in power at all frequencies. So-called broadband shifts in power can occur when the neurons in the underlying population change their firing rates.

Given the many ways to measure and characterize neural responses, which approach is best? The answer depends in part on what we hope to learn. For example, if we are interested in processes that we expect to depend on very precise timing and relatively simple neural computations, then neuron-centric signals like spike timing and firing rate may be especially promising. If we are instead interested in processes that we expect to depend on large-scale computations carried out by populations of thousands of neurons, then we may instead benefit from focusing on periodic and aperiodic features of local field potentials recorded from relatively large electrodes. In general, lower-level processes (e.g., signal transduction) tend to rely on smaller numbers of neurons and occur over shorter timescales. Approaches that operate over few neurons and that support high temporal resolution are often best-suited to studying these low-level processes. By contrast, high-level processes (e.g., scene understanding, complex planning, emotional processing) tend to rely on large populations of neurons and occur over relatively long timescales. Approaches that record from larger populations of neurons and that measure processes or changes that unfold over longer timescales are often best-suited to studying these high-level processes. While these general principles have tended to hold across many studies and recording modalities, it is worth noting some exceptions. For example,
single-neuron responses in humans [123] and non-human primates [117] can sometimes exhibit selectivity for high-level stimuli and semantic concepts.

**Units and patterns versus networks.** After identifying a set of signals that we think will be appropriate for studying our phenomenon or cognitive process of interest, the next key decisions regard whether we should treat those signals in isolation, or as part of a broader network. Essentially this comes down to a decision about how to combine signals and features within and across participants (Fig. 3).

Early single-neuron recordings (in cats and non-human primates) played a central role in Hubel and Wiesel’s Nobel Prize-winning work on mapping out receptive fields of visual cortical neurons [66, 67]. They mapped out the **receptive fields** of neurons in the primary visual cortex by measuring their firing rates as a function of the visual stimulus shown on the retina. In general, a neuron’s receptive field describes the stimulus to which it is maximally responsive. Hubel and Wiesel’s work showed that the primary visual cortex is organized into **orientation columns** of neurons whose receptive fields are tilted dark or light bars at a particular orientation relative to horizontal. Several decades later, researchers used high-field fMRI to show analogous orientation columns in human primary visual cortex [159]. The receptive fields of individual neurons can be enormously complex. In contrast to the simple stimuli preferred by primary sensory neurons, neurons in other brain regions can have receptive fields that correspond to high-level concepts. For example, hippocampal **place cells** fire preferentially when an animal travels to a particular location in an environment [40, 82, 150]. Other work has shown that some medial temporal lobe neurons appear to increase their firing in response to photographs of specific faces, animals, objects, or scenes [123].
When recordings from several neurons are available, the set of firing rates across the population can provide additional information beyond that contained in the firing rates of individual neurons \[4, 121\]. For example, if a single place cell responds to one area of an environment, a population of many place cells that each respond to a different location in the environment can provide a rich cognitive map.

Although they do not provide information about spike timing, macro-scale LFP recordings also reflect population-level neuronal activity. Multi-electrode LFP recordings from one or more brain areas can be especially informative. For example, multi-channel LFPs may be used to decode visual stimuli \[12, 13\], auditory stimuli \[13, 143\], speech production \[12, 122\], acute pain onset and intensity \[165\], and even semantic representations \[39, 97\].

Intracranial recordings (of individual neurons, populations of neurons, and LFPs) may also be considered within the context of the larger brain networks to which they belong or contribute \[11\]. One set of approaches to characterizing brain networks is informed by graph theory, a branch of mathematics concerned with characterizing network architectures, influence, and membership \[10, 15, 21, 64, 100, 129, 141, 145, 146\]. For example, a timeseries of recorded responses from multiple channels may be used to infer functional or causal interactions between the associated neural populations. After mapping out a network of pairwise connections between the responses, graph theoretic measures may be applied to estimate or compare the influence of a given channel or set of channels. Considering interactions can provide information beyond the responses of individual channels or patterns. For example, patterns of interactions between neurons or populations can show selective modulation in response to stimuli or features, even when the underlying individual neurons or populations do not appear responsive to the stimulus when considered in isolation \[127\].

**Static versus dynamic measures of brain activity.** When we attempt to discover the neural patterns associated with a particular stimulus or representation, we need to consider two fundamental questions about how the relevant patterns might change over time. The first question is whether brain representations are fundamentally stable. For example, each time you think of a concept, like the meaning of the word “automobile” do the brain areas relevant to representing that concept display the same basic activity patterns? Or do the neural representations of concepts change in meaningful ways over time, such that the representation of a concept looks fundamentally different each time we measure it? The second question is about whether representations themselves are static or dynamic. For example, when you think of the concept “cat,” does the entire representation essentially become activated as a single unit? Or do different components of the representation (e.g., “fur,” “mammal,” “whiskers,” “tail,” etc.) come online in sequence, perhaps in a stereotyped way that adds additional nuance or meaning?
Some of our conceptual knowledge, and presumably the underlying neural representations of that knowledge, is acquired over timescales on the order of several years. For example, as they develop, children acquire new representations of concepts and how they are related or organized [16, 85]. Changes in neural representations that occur over the course of years are unlikely to be captured by intracranial recordings, which are typically made over timescales on the order of days or weeks.

Another process that leads to changes in the neural representations is pattern separation. Pattern separation refers to the phenomenon of differentiating the neural representations of two or more related stimuli or concepts. For example, pattern separation can occur when we learn to identify and focus in on subtle differences between stimuli or concepts that initially seemed (nearly) identical. Pattern separation can occur over relatively short timescales, and can be identified using intracranial recordings [7, 90, 128, 160].

Although research on the development of conceptual knowledge and pattern separation shows that neural representations can change over time, there is also substantial evidence that neural representations are at least somewhat stable over timescales of hours to days, and even across different individuals [57, 101, 138]. For the most part, this body of work treats the neural representations of concepts and stimulus responses as essentially static. By assuming that the neural patterns evoked by a fixed stimulus are stable (within and across presentations), one can use a machine learning approach called pattern classification to learn mappings between neural patterns and stimulus labels [109]. Once these mappings are learned, they may be applied to new neural patterns to estimate the stimulus or “thoughts” associated with those neural patterns. This allows researchers to estimate the cognitive dynamics that occur during neuroimaging [31, 51, 93, 118, 119].

Some stimuli, such as words, images, pure tones, etc., are perceptually static. Most or all of the information in the stimulus is “made available” to our sensory systems at the same time. Certainly it may take some time for higher-order information to unfold in our minds, for example when we are presented with a complex or thought-provoking image. However, those dynamics are driven by internal processes rather than (directly) by the stimulus itself. Other stimuli, such as movies, motion sequences, and dynamic sounds like speech or music, are fundamentally dynamic. For example, if we were presented with only the average (across time) visual or auditory information in a popular feature-length film, we would be missing nearly all of the structure that made that film engaging or interesting. Dynamic stimuli, particularly naturalistic stimuli with rich spatiotemporal structure reminiscent of real-world experiences, can evoke dynamic neural responses that are often highly reliable across repeated presentations to a single participant participant, as well as across participants [26, 28, 43, 44, 53, 68, 69, 79, 80, 98, 105, 110]. While it is possible to temporally average across the timepoints of responses to some classes of dynamic stimuli while still achieving high reliability [105], this is not universally true. For example, in the domains of speech comprehension and speech
production, temporal information is a primary indicator of meaning. Studying neural responses to speech therefore requires considering how neural correlates of speech unfold over time [52][122].

Independent of whether a given stimulus is fundamentally static or dynamic, neural responses can also change according to which other stimuli were experienced nearby in time. For example, interpreting the neural response to ‘B’ in the sequence A B C D B might entail accounting for whether the given instance of ‘B’ is the one that follows ‘A’ or ‘D’. Randomizing stimulus order and averaging over repeated trials effectively removes this sort of contextual information. However, in some cases, the context in which a stimulus occurs—i.e., the set of other stimuli and thoughts that were experienced nearby in time—can play a critical role in how we process, interpret, and remember incoming information [92]. For example, priming participants using different cues can reliably bias them to interpret an ambiguous narrative in a particular way [162]. Accounting for these sorts of contextual effects often entails factoring in stimulus order or content to the corresponding analyses and models.

41.1.3 Building explicit stimulus models

Identifying neural responses to a stimulus requires formalizing what the stimulus in question is and when the participant is exposed to it. Broadly, this entails building explicit or implicit models that describe the composition and dynamics of the stimulus. The features, outputs, or predictions of these models may then be related to neural patterns.

What is a stimulus “model”? From an analytic perspective, describing a stimulus typically entails characterizing how different aspects of the stimulus change over time. If we are solely interested in the presence or absence of a stimulus, or the timings of a discrete sequence of trials (Fig. 4A), then the “stimulus” might be describable as a simple binary sequence (Fig. 4B). One could then examine or compare neural activity patterns recorded during the “on” timepoints versus “off” timepoints, or estimate how neural responses change during the transitions between those binary states. If stimuli are drawn from a well-defined set of categories, then category information may be conveyed using one binary sequence per category (Fig. 4C).

In other instances, we may be interested in understanding how neural responses relate to specific stimulus values, or how those values change over time. For example, we might describe the brightness or salience of a visual stimulus (or the loudness of an auditory stimulus, etc.) as a sequence of real-valued numbers (Fig. 5A). This could enable us to understand graded neural responses to the stimulus, such as how neural responses change as a function of the stimulus values. Event-triggered averages (analogous to spike-triggered averages such as Fig. 2E) can also provide insights into how the stimulus tended to change during a time window centered on a particular neural event (e.g., an action potential or the appearance of a specific activity pattern).
Fig. 4. Discrete stimulus timeseries. A. Example stimulus sequence. A succession of images are presented to the participant on a computer screen, interspersed by intervals of blank screen. Images are drawn from three categories: faces, outdoor scenes, and concrete objects. B. Onset and offset timing. Stimulus timing, but not stimulus category, may be conveyed using a single binary timeseries. C. Stimulus identity. Stimulus identity (e.g., category) may be represented using a single binary timeseries for each stimulus category or feature.

Some stimuli are best described by multivariate real-valued feature vectors whose elements (i.e., “features”) describe the absense, presence, or values of specific stimulus properties (Fig. 5B). Describing a stimulus as a timeseries of multivariate feature vectors can facilitate more nuanced mappings between those stimulus properties and different aspects of neural activity. For example, the firing rates of different neurons, or the patterns of power spectra across the electrodes in a given region of interest, might display different sensitivity to different stimulus features.

A fourth (general) way of describing how a stimulus changes over time is to use a multivariate timeseries with explicit event-level or trial-level boundaries (Fig. 5C). For example, in a movie, the scene cuts could constitute event boundaries—i.e., moments of transition where the stimulus features exhibit rapid “jumps” that are substantially larger than usual between-timepoint changes. Event boundaries can delineate changes in the focus of an ongoing conversation, scenes in a story or movie, environmental changes, or other transitions in the low-level or high-level content of the stimulus. Experimental trials can also be considered as a sort of event boundary. For example, a multivariate timeseries like that in Figure 5C could also be used to describe the content of a sequence of short video clips presented in succession. Within each clip, the features might change comparatively less than across clips.

As described next, there are many ways to define what the stimulus features are. This requires making assumptions about which aspects of the stimulus “matter” (e.g., in terms of evoking neural responses, predicting behaviors, etc.), and about how the moments of the stimulus timecourse should be matched up with moments of a neural recording. While these assumptions can have a large influence on the outcome of an analysis, there is unfortunately no universal way of describing or modeling stimuli (or of relating stimulus dynamics to neural responses). Rather, one must make informed decisions about how to proceed.
Fig. 5. Continuous stimulus timeseries. **A. Univariate real-valued stimulus.** The stimulus takes on any value at any timepoint. In this example the stimulus values are autocorrelated. **B. Multivariate real-valued stimulus.** The stimulus comprises multiple features, each of which can take on any value at any moment. When projected into 3D, the stimulus traces out a trajectory describing how the values of its features change over time. **C. Multivariate real-valued stimulus with event-level or trial-level dynamics.** This stimulus is similar to the one displayed in Panel B, but here the stimulus values exhibit occasional event boundaries.

Manual approaches. In trial-based experiments where stimuli (e.g., words, sounds, images, etc.) have well-defined onset and offset times, the stimulus onset and offset timeseries (e.g., Fig. 4B) can serve as a simple stimulus “model.” A binary sequence that solely describes stimulus onset and offset times ignores stimulus identity (e.g., the category or label) and stimulus features (e.g., the values of the corresponding feature vectors). In this way, modeling the stimulus using a binary sequence makes the implicit assumptions that (a) stimulus timing is the main factor of interest with respect to the associated neural responses and (b) stimulus identity and stimulus features may be safely ignored.

To model stimulus identity and timing (when both identity and timing information are well-defined, e.g. as in trial-based experiments), the stimulus may also be modeled using a set of binary timeseries (Fig. 4C). For example, one might define a separate binary timeseries describing the onsets and offsets of only one stimulus category or trial type. This approach makes the implicit assumption that the specific identities of different exemplars (e.g., within a stimulus category—such as face images of different people’s faces) may be safely ignored in favor of prioritizing coarser-scale information such as broad stimulus category or trial type labels.

When parameterized stimuli are constructed to vary along one or more explicit stimulus dimensions (e.g., visual stimuli that vary in brightness, contrast, spatial frequency, etc.), the values along each dimension and/or the parameters themselves can serve as a representation of the stimulus (e.g., Fig. 5). Each stimulus dimension (or parameter) may be represented by its own real-valued
timeseries. This approach makes the implicit assumption that the only relevant sources of variation in the stimulus are those characterized by the specified stimulus dimensions or parameters. All other stimulus features or properties are effectively ignored.

Some stimuli cannot be adequately characterized or described using their associated parameters and/or features that can be directly mapped onto specific physical or perceptual properties. For example, such stimuli might be best described using high-level perceptual or conceptual properties of the stimulus such as the presence of specific objects or high-level content, emotional tone, etc. When these properties may be readily judged or rated by human observers, normed ratings or judgements (typically collected by an independent set of participants) may be used as another means of quantifying stimulus features. These judgements may take the form of integer-valued or real-valued responses (e.g., rating how “happy” an image or sound is, on a particular scale) or binary “yes/no” judgements (e.g., indicating whether or not a tree is present in an image).

**Automated approaches.** Many types of stimuli, including natural images, text, complex sounds (e.g., speech, music, recordings of natural environments, etc.), movies, and others, cannot always be easily categorized or manually labeled or rated. In other cases, even if manually labeling stimuli might be possible in principle, in practice it may be too expensive in money or time to generate manual labels. Automated approaches to building stimulus models can scale to millions of stimuli and thousands of stimulus features.

One class of automated approaches to generating stimulus feature models entails applying probabilistic models and deep neural networks such as convolutional neural networks [84, 88], text embedding models [14, 17, 19, 24, 34, 37, 81, 99], transformers [20, 38, 125], and others [78] to the stimuli. The activations of the hidden or output layers of these networks may be used as feature vectors for the corresponding stimuli.

Another class of automated approaches includes visual entity tagging [29], image-to-text models [2, 6, 148], speech-to-text models [47, 74, 157, 166], and other algorithms for generating text data from non-text stimuli like still images, video, and sound. After training these models on large corpora of labeled examples, novel stimuli may be automatically tagged with text annotations. In turn, these annotations may be passed onto text embedding models to construct feature vectors for each stimulus. Alternatively, the annotations may be treated directly as stimulus features, for example by treating each unique keyword as a binary feature that is either present or absent in each stimulus.

**Human-in-the-loop** techniques [163] provide a balance between purely manual and purely automated methods. These techniques entail combining human feedback with classic machine learning approaches. For example, **multidimensional scaling** [23, 41, 151] may be applied to pairwise similarity judgements from human participants to derive n-dimensional feature vectors whose
Fig. 6. Within-patient versus across-patient electrode coverage. Each dot denotes the location of the recording surface of one neurosurgically implanted electrode. A. Example patient. The locations of the 169 electrodes implanted in one patient’s brain are displayed. B. Across-patient electrode locations. The locations of $n = 5023$ electrodes are displayed. Colors denote different patients (electrodes from $m = 53$ patients are displayed). Both panels: implantation locations are taken from [42] and filtered using a thresholding procedure to remove noisy signals reported by [111].

pairwise correlations or (inverse) distances are consistent with those judgements.

To facilitate comparisons or other analyses, it can be useful to partition continuous stimuli into discrete “states” or “events”. For example, hidden Markov models may be applied to a multidimensional timeseries of observations (e.g., feature vectors) to estimate the moments of transition that are interspersed between periods of relative stability [8, 61, 124]. Neural responses during different events, or at transition points, may then be examined further.

41.1.4 What are some modality-specific challenges to identifying stimulus-driven brain activity from intracranial recordings?

While intracranial recordings provide high spatiotemporal resolution data about neural activity (Fig. 1), the coverage afforded by intracranial recordings—i.e., the proportion of the brain volume captured across all electrodes—is relatively poor compared with popular non-invasive approaches like fMRI, MEG, and scalp EEG. The locations of electrodes implanted in a representative patient’s brain are shown in Figure 6A. Across the 169 electrodes, coverage is limited primarily to the left frontal and temporal lobes. The recordings from this example patient cannot provide direct information about activity outside of these regions.

The precise electrode numbers and implantation locations are determined by teams of clinicians whose primary goal is (typically) to locate the seizure focus for that patient. For some patients, the clinical team may have a relatively good sense of where the patient’s seizures likely originate. These patients may
be implanted with relatively few electrodes, spread over a relatively small area of the brain. For other patients where the seizure focus is less clear, they may be implanted with a larger number of electrodes spread throughout the brain. Taken together, these factors mean that each patient is implanted with a different number of electrodes, in different locations, according to a unique clinical plan. Whereas non-invasive recordings can often be easily aligned across people (since the sensor locations are typically held relatively constant across people), the alignment problem is highly non-trivial for intracranial recordings.

One benefit to having varied electrode locations across patients is that, for a sufficiently large number of patients, it becomes possible to obtain data from most of the brain (Fig. 6B). Full-brain maps and analyses may be obtained from intracranial recordings by constructing maps that are stitched together across patients. These maps obtained from intracranial recordings may then be compared to analogous maps obtained using other neuroimaging approaches to provide clues about the reliability of the across-patient findings. However, because full-brain maps derived from intracranial recordings typically require combining data across patients, it can be difficult to identify reliable within-patient effects, or to compare responses across patients.

41.2 Identifying stimulus-driven neural activity

Thus far, we have surveyed a variety of approaches for measuring or characterizing neural activity patterns and stimuli. When applied in conjunction, the result of these approaches is a set of two timeseries: one describing the patient’s neural responses and the other describing the stimulus the patient experienced as they exhibited those neural responses. The final step is to combine these characterizations in order to relate changes in neural activity to changes in the stimulus. Broadly, this combination step may be carried out within participant or across participant.

41.2.1 Within-participant approaches

Within-participant analyses are carried out on data from a single person. Identifying stimulus-driven neural activity using within-participant analyses typically entails combining data over time (e.g., across runs, conditions, trials, etc.). The objective is to estimate maps, patterns, or response profiles that are unique to each individual. These within-participant estimates may then be combined across participants to examine general tendencies in the population and/or individual-specific markers.

Generalized linear models and multivariate pattern analysis. Given a neural recordings (Sec. 41.1.2) and a stimulus model (Sec. 41.1.3), the two most widely used approaches to identifying stimulus-driven neural activity are generalized linear models and multivariate pattern analysis. Broadly, generalized linear
Identifying stimulus-driven neural activity

**models** (GLMs) are an approach to predicting a set of labels, \( Y \), typically represented by one or more feature vectors, from a set of an equal number of inputed observations, also represented by feature vectors, \( X \). Formally, we say that

\[
y_t = f(x_t, \beta),
\]

where \( y_t \) is the \( N \)-dimensional vector of output labels for observation \( t \), \( x_t \) is a the \( M \)-dimensional \( t \)th observation, and \( \beta \) is an \( M \times N \) matrix of weights. The **link function**, \( f(\cdot) \) takes as inputs a set of observations and weights and produces as output a transformed version of the \( N \)-dimensional vector \( \beta^T x_t \). For example, if the elements \( \beta^T x_t \) are Real-valued and lie within the interval \((-\infty, +\infty)\), then a sigmoidal link function (e.g., the logistic or hyperbolic tangent functions) would transform the elements to lie within the interval \((-1, +1)\). The “power” of the generalized linear model framework comes from the flexibility in how the link function may be defined. By choosing an appropriate link function, it is possible to take arbitrary Real-valued inputs and transform them into outputs that match a wide variety of useful formats—e.g., (unbounded) Real-valued outputs; probability-like values bounded between 0 and 1; indicator vectors (i.e., vectors where all values are 0s except for one element whose value is 1); binary-like values (i.e., where extreme values are “pulled” towards one of two boundaries, as in sigmoid functions); and many more. When \( X \) reflects neural activity and \( Y \) reflects the stimulus features during the corresponding moments, the fitted GLM weights (i.e., \( \beta \)) describe how different aspects of neural activity relate to different stimulus features. These fitted weights may also be used to “decode” stimulus features from new, previously unobserved, neural data.

**Multivariate pattern analysis** (MVPA) describes a second class of approaches for connecting stimulus features and neural activity. Like GLMs, the goal of MVPA is to predict a set of labels from a set of observations. Typically the “labels” comprise stimulus features and the “observations” comprise neural responses. GLMs are a special case of MVPA for which the output features reflect a (potentially transformed) linear combination of the input features. However, MVPA also includes a variety of other approaches for which the relations between input and output features are non-linear, and potentially even non-monotonic. The umbrella term for such algorithms is **pattern classifiers**, and includes GLMs, support vector machines, boosting, naive Bayes classifiers, nearest neighbor-based classifiers, and (deep) neural network-based classifiers, among others.

**Representational similarity analysis.** Comparing the **neural temporal correlation matrix** (i.e., the correlations between the neural patterns recorded at every pair of timepoints) to the **stimulus temporal correlation matrix** (i.e., the correlations between the stimulus feature vectors at every pair of timepoints) can reveal similarities and differences between how neural and stimulus feature change over time. **Representational similarity analysis** (RSA) entails computing the element-wise correlation between the upper triangles of the neural and stimulus correlation matrices. Following the logic of, to the extent that
neural patterns show a similar temporal correlation structure to stimulus features, we can interpret this as evidence that the neural and stimulus features are related.

RSA may also be carried out across a series of searchlights. Similar to how a searchlight illuminates a well-defined area of darkness, a searchlight analysis provides insights into the functionality or responses profiles of a focused region of interest. For each of a series of spherical volumes tiled throughout the brain, RSA may be performed for each volume by limiting the neural features under consideration to only those captured by electrodes within that sphere’s radius. This yields, for each sphere (i.e., searchlight) a single correlation coefficient between that sphere’s neural temporal correlation matrix and the stimulus temporal correlation matrix. Examining which searchlights displayed high versus low correlations can highlight which brain areas might represent the stimulus in a way consistent with a given stimulus model (i.e., the model used to construct or estimate the stimulus features).

A convenient property of RSA is that, unlike approaches like GLMs or MVPA, RSA does not require learning an explicit mapping between neural features and stimulus features. This is because RSA is driven solely by pattern similarities across timepoints, rather than by the specific properties of the patterns themselves. In this way, RSA can sometimes be a more sensitive way of identifying stimulus-driven neural patterns (e.g., compared with GLMs and MVPA). For example, high levels of noise during many of the exposures to a particular stimulus category will mean that neural decoding approaches will likely fail to effectively learn mappings between the neural and stimulus features for that category. However, RSA analyses effectively “average” over all timepoints, thereby highlighting aspects of neural activity with a similar temporal correlation structure to any stimulus features (even if the associations with a subset of the stimulus features are noisy).

Joint stimulus-activity models. Thus far, we have reviewed two approaches to identifying stimulus-driven neural activity. MVPA attempts to learn mappings between stimulus features and neural features, and RSA attempts to identify stimulus features and neural features that exhibit similar temporal correlation patterns. A third (related) approach entails building models that jointly consider the timecourses of neural and stimulus features. Whereas MVPA and RSA implicitly treat stimulus and neural features as a “ground truth,” joint stimulus-activity models allow the stimulus features and neural features to mutually inform each other. These models assume that, while the mappings between stimulus and neural features may be relatively stable, there may be some times when the stimulus features provide a more reliable signal and other times when neural features provide a more reliable signal.

Figure 7 provides some geometric intuitions for the idea of joint stimulus-activity models. First, consider how we might go about estimating the reliability of some neural responses we measure as people experience a given stimulus, such as watching a movie. One approach might be to expose a single partic-
Fig. 7. Joint geometric models of stimulus and neural features. A. Neural features from different individuals. Group-averaged trajectory of fMRI activity from ventral visual cortex split into two randomly-selected groups of subjects (group 1: \(n = 6\), group 2: \(n = 5\) watching the same movie [57]. B. A common geometric space for stimulus and neural features. Group-averaged trajectory of fMRI activity from ventral visual cortex and trajectory of movie (pixel intensities over time) hyperaligned [57] to a common space. In Panels A and B, the high-dimensional stimulus and neural trajectories have been projected onto three dimensions to facilitate visualization [62]. C. Interpreting coordinates in the common feature space. Stimuli (e.g., movie frames) may be reconstructed from ventral visual brain activity by mapping the coordinates of neural features in the common space onto coordinates in the stimulus space. Note: this figure is adapted from [62].

Geometrically, the neural features recorded from one individual, during one moment, can be conceptualized as a single point in a high dimensional feature space (whose dimensions each correspond to a single neural feature). Over time, as the individual’s neural activity changes, the successive activity patterns trace out a trajectory through this neural activity space. When the timecourses of neural activity patterns are similar across trials or individuals, this results in similar neural trajectory shapes (Fig. 7A).

The stimulus features from a single moment, as well as the timecourse describing how stimulus features change over time, may also be conceptualized as a trajectory through a (different) high dimensional feature space, whose dimensions each correspond to a single stimulus feature. It could be interesting to ask whether the neural and stimulus trajectories are similarly shaped, or whether there are particular moments or circumstances under which the shapes converge or diverge. However, because the dimensions of neural trajectories and stimulus trajectories do not (typically) match, an additional step is needed before such comparisons may be made. The procrustean transformation is a geometric transformation for bringing two sets of coordinates into an optimal point-by-point alignment. This entails computing the affine transformations (i.e., rotations, reflections, and scalings) that, when applied to coordinates in one set, minimize the average Euclidean distance between the corresponding coordinates in the second set. The resulting aligned coordinates may then be
directly compared, since the transformations map the coordinates in the first set into the same coordinate system (with the same dimensions) as the second set [57]. Mapping a neural trajectory into the stimulus feature space (Fig. 7B, C) provides a common coordinate system for describing both stimulus features and neural features.

We can use this geometric framework to conceptualize what it means to jointly model the stimulus and neural responses. Consider, for example, the stimulus and neural trajectories displayed in Figure 7B. While both trajectories look similar in some respects (e.g., they have roughly similar coarse-scale shapes), they also differ in potentially important ways (e.g., the stimulus trajectory is “spikier” than the neural trajectory). Which trajectory is “correct”? On one hand, the stimulus trajectory provides a relatively clean characterization of the stimulus that exactly reflects specific measurable aspects of what participants were exposed to. In this sense, the stimulus trajectory is not “corrupted” by measurement noise, inattention, or other factors unrelated to the stimulus itself. On the other hand, the stimulus trajectory is (by definition) a reflection only of the specific stimulus features that we, as the experimentalists, decided were likely to be important. Those features might be at best different and at worst unrelated to the stimulus properties that participants actually care about or respond to. In this sense, one could argue that neural responses reflect the most direct representation of aspects of the stimulus the brain is responding to, since those neural responses are uncorrupted by the experimentalists’ assumptions about which stimulus features are important. Taken together, it is clear that neither the stimulus trajectory nor the neural response trajectory, in isolation, provide a complete reflection of an individual’s internal mental representations of the stimulus. Instead, it might be most accurate to incorporate aspects of both the stimulus and neural trajectories. This joint stimulus-activity modeling approach acknowledges that the true representation(s) to which an individual’s brain is responding may lie somewhere between the stimulus and neural trajectories.

41.2.2 Across-participant approaches

Across-participant analyses are carried out on data from multiple individuals. Identifying stimulus-driven neural activity using across-participant analyses typically entails building an across-participant model or developing analyses that characterize similarities or differences in responses across participants. The objective is to estimate a single map, pattern, or response profile that is common across individuals. Some approaches also attempt to estimate individual differences that characterize how each individual’s responses differ from the group’s (aggregated) responses.

Across-participant models. Building across-participant models for identifying stimulus-driven neural activity requires defining a common representation for describing neural activity (and, potentially, linking neural features with stimulus features). Each participant’s data must first be mapped into the common
representation space. This may be carried out using anatomical [3, 60, 149], functional [27, 57, 58, 158], or other [130, 137] alignment methods. Next, the inference procedure (i.e., the algorithm for estimating model parameters from the observed data) must learn both local parameters (i.e., parameters that are specific to each individual) and global parameters (i.e., parameters that are shared across individuals). The final step is often to learn a mapping or linking function for connecting local and/or global parameters to stimulus features. This last step can be carried out within-participant (by learning mappings between local parameters and stimulus features) or across-participant (by learning mappings between global parameters and stimulus features).

Hierarchical matrix factorization models. Matrix factorization encompasses a family of mathematical approaches for decomposing a matrix into the product of several other matrices. This family includes a large number of machine learning models, including Topographic Factor Analysis (TFA) [96, 98], Topographic Latent Source Analysis (TLSA) [50], Principal Components Analysis (PCA) [114], Exploratory Factor Analysis (EFA) [144], and Independent Components Analysis (ICA) [33, 75], among others. Within the domain of neuroimaging, the general formulation is to first organize the neural feature vectors (from a single subject) into a $T$ by $N$ data matrix, $Y$ (where $T$ is the number of observations and $N$ is the number of neural features). We can then decompose $Y$ as follows:

$$Y \approx WF,$$

where $W$ is a $T$ by $K$ weight matrix (which describes how each of $K$ factors are activated for each observation) and $F$ is a $K$ by $N$ matrix of factor images (which describes how each factor maps onto the neural features). In the general case, there are infinitely many solutions for this decomposition. Different matrix factorization approaches converge on specific choices for $W$ and $F$ by placing different constraints on the forms the matrices must take or by choosing optimization metrics that emphasize different aspects of $Y$ to be preserved. For example, when $K << N$, the approximation of $Y$ via $WF$ will be inexact.

To illustrate how matrix factorization models may be constructed to capture multi-subject data, we can examine the details of two related models: hierarchical and non-hierarchical variants of the same matrix factorization model, topographic factor analysis. In its non-hierarchical framing, TFA specifies that each row of $F$ is parameterized by the center parameter, $\mu$, and the width parameter, $\lambda$, of a radial basis function. If a radial basis function has center $\mu$ and width $\lambda$, then its activity $RBF(r|\mu, \lambda)$ at location $r$ is:

$$RBF(r|\mu, \lambda) = \exp \left\{ -\frac{||r - \mu||^2}{\lambda} \right\}.$$

The factor images are filled in by evaluating each radial basis function, defined by the corresponding parameters for each factor, at the location(s) of each electrode or brain region of interest. In contrast to the factors obtained using PCA or ICA, TFA’s more constrained factors may be represented much more compactly;
Fig. 8. Topographic Factor Analysis. 

A. Spherical factors describe contiguous regions of similar activity. Each factor is represented as a radial basis function. A factor’s image may be constructed by evaluating its radial basis function anywhere within the brain volume. Level curves for several example factors fit to a synthetic 3D image are outlined in white; *x* denotes the factor centers projected onto the 2D slice displayed in the panel. 

B. Brain images are described by weighted sums of the factors’ images. After computing each factor’s image (using its radial basis function), arbitrary brain images may be approximated using weighted combinations of the images for each factor. The per-image weights may be used as a low-dimensional embedding of the original data. A 2D slice of the reconstruction for the image displayed in panel A demonstrates how contiguous clusters of locations are approximated using weighted activations of spherical factors. 

C. The global template serves as a prior for subject-specific parameters. The global template defines the numbers of factors, their locations, and their sizes, for the prototypical participant. Each individual participant’s parameters (factor locations and sizes) are fit using the global template as a prior. This provides a linking function between different participants’ factors, thereby enabling across-subject comparisons. A subset of the factors outlined in Panel A are displayed in the global template cartoon. The positions of these factors in each individual participant’s subject-specific template are displayed in different colors. Note: this figure is adapted from [86].
each factor corresponds to the structure or group of structures in the brain over which the factor spreads its mass (which is governed by $\mu$ and $\lambda$). TFA’s factors may be conceptualized as nodes located in 3D space whose activity patterns influence the observed brain data (Fig. 8A, B).

Hierarchical Topographic Factor Analysis (HTFA) works similarly to TFA, but places an additional constraint over the factors to bias all of the subjects to exhibit similar factors. Whereas TFA attempts to find the factors that best explain an individual subject’s data, HTFA also attempts to find the factors that are common across a group of subjects (Fig. 8C). This is important, because it allows the model to jointly consider data from multiple subjects.

HTFA handles multi-subject data by defining a global template, which describes in general where each radial basis function is placed, how wide it is, and how active its node tends to be. In addition to estimating how factors look and behave in general (across subjects), HTFA also estimates each individual’s subject-specific template, which describes each subject’s particular instantiations of each radial basis function (i.e. that subject’s radial basis function locations and widths) and the factor weights (i.e. the activities of each of that subject’s radial basis function factors in each of that subject’s observed neural activity patterns). Because the subject-specific templates are related to each other (hierarchically via the global template), a given factor’s radial basis function will tend to be located in about the same location, and be about as large, across all of the subject-specific templates. Because each subject has the same set of factors (albeit in slightly different locations and with slightly different sizes) we can run analyses that relate the factors across subjects.

The general approach of learning global and subject-specific factors may be applied to many matrix factorization models. For example, TLSA, PCA, and ICA each have hierarchical framings as well. The particular benefit of using (H)TFA to decompose and describe intracranial data is that the radial basis function factors may be evaluated at the unique locations of each individual patient’s electrodes, even though the electrode locations will differ across patients. And because the subject-specific templates are associated via the global template, aspects of the subject-specific templates may be compared across patients. For example, after using HTFA to learn the global and subject-specific templates, these templates may then be treated as neural features and examined in relation to stimulus features. Specifically, each patient’s $W$ matrix may be treated as neural features– but whereas the “raw” neural features in the original dataset will not be consistent across patients, the columns of each patient’s $W$ matrix may be directly combined or compared. In addition, the global and subject-specific factors (rows of $F$) may be examined to identify how the columns of $W$ map onto different brain areas or structures.

Gaussian process models. Gaussian process regression is an approach for estimating “missing” (unobserved) data by using related observed data. Gaussian process regression is particularly well-suited to applications where nearby datapoints are expected to take on similar values. For example, if we assume that nearby locations in the brain will exhibit similar neural activity patterns,
Fig. 9. Building across-patient models using Gaussian process regression. A. Electrode locations. Each dot reflects the location of a single electrode implanted in the brain of one patient. A held-out recording location from one patient is indicated in red, and the patient’s remaining electrodes are indicated in black. The electrodes from the remaining patients are colored by $k$-means cluster (computed using the full-brain correlation model shown in Panel D). B. Radial basis function kernel. Each electrode contributed by the patient (black) weights on the full set of locations under consideration (all dots in Panel A). The weights fall off with positional distance (in MNI152 space) according to a radial basis function. C. Per-patient correlation matrices. After computing the pairwise correlations between the recordings from each patient’s electrodes, correlations between all locations may be estimated using radial basis function-weighted averages. This yields one estimated full-brain correlation matrix for each patient. D. Merged correlation model. Combining the per-patient correlation matrices (Panel C) yields a single full-brain correlation model that captures information contributed by every patient. Here the rows and columns are sorted to reflect $k$-means clustering labels [using $k=7$, [161], whereby locations are grouped according to their correlations with the rest of the brain (i.e., rows of the matrix displayed in the panel). The boundaries denote the cluster groups. The rows and columns of Panel C have been sorted using the Panel D-derived cluster labels. E. Reconstructing activity throughout the brain. Given the observed recordings from the given patient (shown in black; held-out recording is shown in blue), along with a full-brain correlation model (Panel D), applying Gaussian process regression yields the most probable activity at the held-out location (red). Note: this figure is adapted from [111], and data are from [95, 97, 133–135].

we could use an approach like Gaussian process regression to estimate the most probable activity patterns from locations that were nearby (but not necessarily exactly overlapping with) the electrode implantation sites for that patient [111]. An overview of this technique is shown in Figure 9.

To build an across-subject model of neural activity patterns, we first need to define a set of locations in the brain to include in the combined model (Fig. 9A). These locations comprise the set of brain coordinates where we will want to estimate activity patterns for every patient. Next, we estimate a correlation model that describes how activity exhibited by each pair of locations is related. To estimate the correlation model, we first compute the pairwise correlations between activity recorded from each individual patients’ electrodes, and then we use spatial blurring (Fig. 9B) to interpolate those correlations over the full set of target locations in the model. This yields a single estimated correlation matrix for each patient (Fig. 9C). We can then use weighted averaging to combine the
patient-specific correlation matrices into a single correlation model (Fig. 9D). Essentially, an entry in a given patient’s individual correlation matrix will be weighted more heavily in the combined model if that patient had electrodes nearby to the pair of locations that entry reflects. Conceptually, this combined correlation model reflects “global” information from multiple patients, whereas each individual patient’s correlation matrix reflects “local” information from that patient alone. Given a correlation model (learned from multiple patients) and a set of recordings (observed from one patient), Gaussian process regression may be applied to reconstruct (i.e., estimate) neural activity at any location in the combined model— even if the given patient did not have any electrodes at that location (Fig. 9E).

Using Gaussian process regression to estimate full-brain activity patterns from a limited number of electrodes can be useful for identifying stimulus-driven neural activity. For example, whereas raw intracranial recordings are typically taken from different locations across patients (Fig. 6), the above approach may be used to estimate activity patterns at a common set of locations across people. Second, whereas raw intracranial recordings from a single patient typically lack full-brain coverage (Fig. 6A), the set of locations in the combined correlation model may be chosen to cover arbitrarily much of the brain, at arbitrarily high spatial resolution. In turn, this can enable researchers to train or apply other across-participant models, such as pattern classifiers, from different patients’ intracranial recordings [131].

Hyperalignment and the shared response model. Even when we record across subjects from an (ostensibly) overlapping set of locations or neural features, individual differences in stimulus-driven neural responses, behavior, internal representations, and even neuroanatomy can lead to different observed responses. When working with intracranial recordings, where the recording locations rarely overlap across people, and where non-standard neuroanatomical traits are relatively common, these factors are even more prevalent. Hierarchical matrix factorization and Gaussian process models make the simplifying assumption that different individual’s underlying neural representations are spatially similar. But what if the same functional representations are reflected by different spatial activity patterns across different people? Models that match up neural features primarily according to their spatial attributes will fail to capture or correctly identify (non-spatial) functional similarities across individuals. In contrast, functional alignment models attempt to discover functional overlap in neural activity patterns across individuals, even when the neural features across those individuals are incompatible or out of spatial alignment.

Hyperalignment [57] uses the procrustean transformation to align the neural trajectories of different individuals into a common feature space (Fig. 7). This entails computing the linear re-combination of neural features (for each individual) that brings the group’s neural trajectories into the closest point-by-point alignment. Because the procrustean transformation is invertible, neural features may be mapped between different individuals, or between specific individuals and the common feature space. The Shared Response Model [SRM; 30] is
similar to hyperalignment in that it provides a means of defining a common neural feature space that is shared across individuals. SRM extends hyperalignment by combining the alignment step with a dimensionality reduction step that attempts to specifically find a lower-dimensional common neural feature space.

Although hyperalignment and SRM are most often applied to fMRI data [58], in principle these models are modality-independent. For example, one recent study found that applying SRM to intracranial recordings, taken taken as patients watched a movie, revealed a set of shared components that co-varied with the affective content of the movie [158]. The study replicated (using intracranial recordings) several key findings from related fMRI work [25]. Another recent study, using intracranial recordings taken from the rodent hippocampus, demonstrated that hyperaligning neural features across animals enabled reliable across-individual decoding of the animals’ spatial locations [27].

Inter-subject correlation and inter-subject functional correlation. Inter-subject correlations [ISC; 54] and inter-subject functional correlations [ISFC; 140] entail computing the correlations between time-aligned signals recorded from different individuals as they perform a common task (Fig. 3). ISC operates on the same (or equivalent) neural features across individuals, and ISFC operates on different pairs of neural features across individuals.

To compute ISC for a particular neural feature, we first isolate that feature’s timeseries in each individual’s brain. (If the same neural feature is not present across individuals, ISC may be performed after employing another approach to equating or mapping between neural features across individuals.) Next, for a single “reference” individual, we correlate their timeseries (for the given neural feature) with the timeseries for the same neural feature averaged across all other individuals. This yields a single correlation coefficient for that individual, for the given neural feature. Repeating this calculation using each individual in turn as the reference yields one correlation coefficient for each individual. Finally, we average the correlations across individuals to obtain a single ISC value for the given neural feature.

Computing ISFC is similar to computing ISC. However, whereas ISC correlates the same neural feature across individuals, ISFC correlates different neural features across individuals. The result is a symmetric matrix of correlation values that summarize how (on average, across individuals) each pair of neural features are correlated.

ISC and ISFC are particularly effective at capturing stimulus-driven activity patterns during naturalistic tasks (e.g., story listening, movie viewing, natural conversation, etc.) when constructing a reliable model of the stimulus time-course can be challenging [139]. Effectively, ISC and ISFC treat the average signals recorded from other participants as a “model” of the stimulus dynamics. Since non-stimulus-driven activity patterns are not expected to be correlated across people, ISC and ISFC are designed to specifically identify timecourses of stimulus-driven neural patterns. Although these approaches are most com-
monly applied to non-invasive recordings, they have been successfully applied to intracranial recordings as well [55, 65, 104, 120].

41.3 Summary and concluding remarks

Identifying stimulus-driven neural activity requires selecting an appropriate recording modality and experimental paradigm, defining neural (Sec. 41.1.2) and stimulus (Sec. 41.1.3) features, and then building explicit or implicit linking functions between neural and stimulus features (Sec. 41.2). We reviewed two general strategies for building these links: within-participant approaches and across-participant approaches.

Within-participant approaches include generalized linear models, multivariate pattern analysis, representational similarity analysis, and joint stimulus-activity models. These approaches each attempt to identify individual-specific maps, patterns, and/or response profiles. Across-participant approaches include hierarchical matrix factorization models, Gaussian process models, geometric alignment models, inter-subject correlation, and inter-subject functional correlation. These approaches each attempt to identify stimulus-driven neural activity that is similar across individuals.

We also identified several challenges that are unique to intracranial recordings. These challenges primarily stem from two factors. First, building across-participant models requires accounting for differences in electrode placement, number of electrodes, and electrode type, across individuals (Fig. 6). Second, because intracranial electrodes must be implanted surgically, the subject population in human intracranial experiments is limited to neurosurgical patients with serious neurological symptoms such as drug-resistant epilepsy. These symptoms often result from brain abnormalities (e.g., trauma or other forms of physical damage, developmental abnormalities, and/or other structural or functional issues). These issues provide challenges both to comparing findings across individuals within an intracranial experiment, and also to generalizing any findings to the broader population.

As a field, cognitive neuroscience is still decades away from being able to link neural and stimulus features at high levels of detail. This is partly due to recording quality (even in high-fidelity modalities like intracranial recordings) and coverage, and partly due to insufficient quality or fidelity of stimulus models and decoding algorithms. Nevertheless, insights into the associations between stimuli and neural responses can help to elucidate the neural basis of cognition in ways that behavior alone cannot. For example, when behaviors are ambiguous (e.g., a response could convey several meanings, a response could arise from several equally reasonable or likely cognitive processes, etc.) or when there are no behaviors for a given cognitive phenomenon (e.g., forgetting, unshared internal thoughts, etc.), additional signal is needed to resolve those ambiguities. In addition, understanding the neural underpinnings of cognition requires measuring neural activity in some form. Recent developments in natural language processing and deep learning, along with advances in tools for
more easily constructing neural, stimulus, and decoding models [e.g., 1, 49, 113] suggest a bright future for this important area of neuroscientific inquiry.
Bibliography

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: a system for large-scale machine learning. Proceedings of the USENIX Symposium on Operating Systems Design and Implementation pp. 265–283 (2016)

[2] de Andrade, D.O.S., Maia, L.F., de Figueiredo, H.F., Viana, W., Trinta, F., de Souza Baptista, C.: Photo annotation: a survey. Multimedia Tools and Applications 77, 423–457 (2018)

[3] Archip, N., Clatz, O., Whalen, S., Kacher, D., Fedorov, A., Kot, A., Chrischohoides, N., Jolesz, F., Golby, A., Black, P.M., Warfield, S.K.: Non-rigid alignment of pre-operative MRI, fMRI, and DT-MRI with intra-operative MRI for enhanced visualization and navigation in image-guided neurosurgery. NeuroImage 35(2), 609–624 (2007)

[4] Averbeck, B.B., Latham, P.E., Pouget, A.: Neural correlations, population coding and computation. Nature Reviews Neuroscience 7(5), 358–366 (2006)

[5] Azevedo, F.A.C., Carvalho, L.R.B., Grinberg, L.T., Farfel, J.M., Ferretti, R.E.L., Leite, R.E.P., Filho, W.J., Lent, R., Herculano-Houzel, S.: Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. The Journal of Comparative Neurology 513(5), 532–541 (2009)

[6] Baboud, L., Čadik, M., Eisemann, E., Seidel, H.P.: Automatic photo-to-terrain alignment for the annotation of mountain pictures. IEEE Xplore p. doi.org/10.1109/CVPR.2011.5995727 (2011)

[7] Bakker, A., Kirwan, C.B., Miller, M., Stark, C.E.L.: Pattern separation in the human hippocampal CA3 and dentate gyrus. Science 319(5870), 1640–1642 (2008)

[8] Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A.: Discovering event structure in continuous narrative perception and memory. Neuron 95(3), 709–721 (2017)

[9] Barrs, B.A.: The mystery and magic of glia: a perspective on their roles in health and disease. Neuron 60(3), 430–440 (2008)

[10] Bassett, D.E., Bullmore, E.D.: Small-world brain networks. The Neuroscientist 12(6), 512–523 (2006)

[11] Bassett, D.S., Sporns, O.: Network neuroscience. Nature Neuroscience 20(3), 353–364 (2017)

[12] Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M.A., Logothetis, N.K.: Local field potentials and spiking activity in primary visual cortex convey independent information about natural visual stimuli. The Journal of Neuroscience (2008)
[13] Belitski, A., Panzeri, S., Magri, C., Logothetis, N.K., Kayser, C.: Sensory information in local field potentials and spikes from visual and auditory cortices: time scales and frequency bands. Journal of Computational Neuroscience 29, 533–545 (2010)

[14] Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. Journal of Machine Learning Research 3, 1137–1155 (2003)

[15] Betzel, R.F., Medaglia, J.D., Kahn, A.E., Soffer, J., Schonhaut, D.R., Bassett, D.S.: Inter-regional ECoG correlations predicted by communication dynamics, geometry, and correlated gene expression. arXiv 1706.06088 (2017)

[16] Blaye, A., Bernard-Peyron, V., Paour, J.L., Bonthoux, F.: Category flexibility in children: distinguishing response flexibility from conceptual flexibility; the protracted development of taxonomic representations. European Journal of Developmental Psychology 3(2), 163–188 (2006)

[17] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)

[18] Boser, B.E., Guyon, I.M., Vapnick, V.: A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, ACM (1992)

[19] Brown, P.F., deSouza, P.V., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics 18(4), 467–480 (1992)

[20] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. arXiv 2005.14165 (2020)

[21] Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience 10(3), 186–198 (2009)

[22] Buzsáki, G.: Rhythms of the brain. Oxford University Press, New York, NY (2006)

[23] Carroll, J.D., Chang, J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition. Psychometrika 35(3), 283–319 (1970)

[24] Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzweil, R.: Universal sentence encoder. arXiv 1803.11175 (2018)

[25] Chang, L.J., Jolly, E., Cheong, J.H., Rapuano, K., Greenstein, N., Chen, P.H.A., Manning, J.R.: Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. Science Advances 7(17), doi.org/10.1126/sciadv.abf7129 (2021)

[26] Chang, L.J., Manning, J.R., Baldassano, C., de la Vega, A., Fleetwood, G., Geerlings, L., Haxby, J., Lahnakoski, J., Parkinson, C., Shappell, H., Shim, W.M., Wager, T., Yarkoni, T., Yeshurun, Y., Finn, E.: Naturalistic data analysis. Zenodo doi.org/10.5281/zenodo.3937848 (2020)
[27] Chen, H.T., Manning, J.R., van der Meer, M.A.A.: Between-subject prediction reveals a shared representational geometry in the rodent hippocampus. Current Biology 31(19), 4293–4304 (2021)

[28] Chen, J., Leong, Y.C., Honey, C.J., Yong, C.H., Norman, K.A., Hasson, U.: Shared memories reveal shared structure in neural activity across individuals. Nature Neuroscience 20(1), 115 (2017)

[29] Chen, M., Zheng, A., Weinberger, K.: Fast image tagging. Proceedings of Machine Learning Research 28(3), 1274–1282 (2013)

[30] Chen, P.H., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., Ramadje, P.J.: A reduced-dimension fMRI shared response model. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 460–468. Curran Associates, Incorporated (2015)

[31] Chiu, Y.C., Wang, T.H., Beck, D.M., Lewis-Peacock, J.A., Sahakyan, L.: Separation of item and context in item-method directed forgetting. NeuroImage 235, 117983 (2021)

[32] Churchland, P.A., Sejnowski, T.J.: Perspectives on cognitive neuroscience. Science 242(4879), 741–745 (1988)

[33] Comon, P., Jutten, C., Herault, J.: Blind separation of sources, part II: problems statement. Signal Processing 24(1), 11–20 (1991)

[34] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv 1705.02364 (2018)

[35] Cortes, C., Vapnick, V.: Support vector networks. Machine Learning 20, 273–297 (1995)

[36] Crone, N.E., Korzeniewska, A., Franaszczuk, P.J.: Cortical gamma responses: searching high and low. International Journal of Psychophysicsiology 79(1), 9–15 (2011)

[37] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)

[38] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv 1810.04805 (2018)

[39] Dezfooli, M.P., Daliri, M.R.: Single-trial decoding from local field potential using bag of word representation. Brain Topography 33, 10–21 (2020)

[40] Ekstrom, A.D., Kahana, M.J., Caplan, J.B., Fields, T.A., Isham, E.A., Newman, E.L., Fried, I.: Cellular networks underlying human spatial navigation. Nature 425, 184–187 (2003)

[41] Ennis, D.M., Palen, J., Mullen, K.: A multidimensional stochastic theory of similarity. Journal of Mathematical Psychology 32, 449–465 (1988)

[42] Ezyat, Y., Kragel, J.E., Burke, J.F., Levy, D.F., Lyalenko, A., Wanda, P., O’Sullivan, L., Hurley, K.B., Busygin, S., Pedisich, I., Sperling, M.R., Worrall, G.A., Kuczewicz, M.T., Davis, K.A., Lucas, T.H., Inman, C.S., Lega, B.C., Jobst, B.C., Sheth, S.A., Zaghoul, K., Jutras, M.J., Stein, J.M., Das, S.R., Gorniak, R., Rizzuto, D.S., Kahana, M.J.: Direct brain stimulation
modulates encoding states and memory performance in humans. Current Biology 27, 1–8 (2017)

[43] Finn, E.S.: Is it time to put rest to rest? Trends in Cognitive Sciences doi.org/10.1016/j.tics.2021.09.005 (2021)

[44] Finn, E.S., Bandettini, P.A.: Movie-watching outperforms rest for functional connectivity-based prediction of behavior. NeuroImage 235, 117963 (2021)

[45] Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: consistency properties. Tech. rep., USAF School of Aviation Medicine (1951)

[46] Fries, P., Nikolić, D., Singer, W.: The gamma cycle. Trends in Neurosciences 30(7), 309–316 (2007)

[47] Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C.: Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech and Audio Processing 12(4), 401–408 (2004)

[48] Gelman, A., Hill, J.: Data analysis using regression and multi-level/hierarchical models. Cambridge University Press (2007)

[49] Gelman, A., Lee, D., Guo, J.: Stan: a probabilistic programming language for Bayesian inference and optimization. Journal of Educational and Behavioral Statistics 20(5), 530–543 (2015)

[50] Gershman, S.J., Blei, D.M., Pereira, F., Norman, K.A.: A topographic latent source model for fMRI data. NeuroImage 57, 89–100 (2011)

[51] Gershman, S.J., Schapiro, A.C., Hupbach, A., Norman, K.A.: Neural context reinstatement predicts memory misattribution. The Journal of Neuroscience 33(20), 8590–8595 (2013)

[52] Hamilton, L.S., Oganian, Y., Hall, J., Chang, E.F.: Parallel and distributed encoding of speech across human auditory cortex. Cell 184(18), 4626–4639 (2021)

[53] Hasson, U., Malach, R., Heeger, D.J.: Reliability of cortical activity during natural stimulation. Trends in Cognitive Sciences 14, 40–48 (2010)

[54] Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R.: Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640 (2004)

[55] Haufe, S., DeGuzman, P., Henin, S., Arcaro, M., Honey, C.J., Hasson, U., Parra, L.C.: Elucidating relations between fMRI, ECoG, and EEG through a common natural stimulus. NeuroImage 179(1), 79–91 (2018)

[56] Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430 (2001)

[57] Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J.: A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416 (2011)

[58] Haxby, J.V., Guntupalli, J.S., Nastase, S.A., Ma, F.: Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. eLife 9, e56601 (2020)
[59] Herculano-Houzel, S.: The human brain in numbers: a linearly scaled-up primate brain. Frontiers in Human Neuroscience 3(31), doi.org/10.3389/neuro.09.031.2009 (2009)

[60] Hermes, D., Miller, K.J., Noordmans, H.J., Vansteensel, M.J., Ramsey, N.F.: Automated electocorticographic electrode localization on individually rendered brain surfaces. Journal of Neuroscience Methods 185(2), 293–298 (2010)

[61] Heusser, A.C., Fitzpatrick, P.C., Manning, J.R.: Geometric models reveal behavioral and neural signatures of transforming naturalistic experiences into episodic memories. Nature Human Behavior 5, 905–919 (2021)

[62] Heusser, A.C., Ziman, K., Owen, L.L.W., Manning, J.R.: HyperTools: a Python toolbox for gaining geometric insights into high-dimensional data. Journal of Machine Learning Research 18(152), 1–6 (2018)

[63] Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerve. Journal of Physiology 117, 500–544 (1952)

[64] Honey, C.J., Kötter, R., Breakspear, M., Sporns, O.: Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proceedings of the National Academy of Sciences, USA 104(24), 10240–10245 (2007)

[65] Honey, C.J., Thesen, T., Donner, T.H., Silbert, L.J., Carlson, C.E., Devinsky, O., Doyle, J.C., Rubin, N., Heeger, D.J., Hasson, U.: Slow cortical dynamics and the accumulation of information over long timescales. Neuron 76, 423–434 (2012)

[66] Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. The Journal of Physiology 160(1), 106 (1962)

[67] Hubel, D.H., Wiesel, T.N.: Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology 195(1), 1 (1968)

[68] Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L.: Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458 (2016)

[69] Huth, A.G., Nisimoto, S., Vu, A.T., Gallant, J.L.: A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76(6), 1210–1224 (2012)

[70] Huxter, J., Allen, K., Csicsvari, J.: Theta phase-specific codes for two-dimensional position, trajectory and heading in the hippocampus. Nature Neuroscience 11(5), 587–594 (2008)

[71] Jacobs, G.H.: Color vision. In: Encyclopedia of Biological Chemistry III (Third Edition). Elsevier (2021)

[72] Jacobs, J., Kahana, M.J., Ekstrom, A.D., Fried, I.: Brain oscillations control timing of single-neuron activity in humans. The Journal of Neuroscience 27(14), 3839–3844 (2007)

[73] Jonas, E., Kording, K.P.: Could a neuroscientist understand a microprocessor? PLoS Computational Biology 13(1), e1005268 (2017)
[74] Jones, D.A., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D.A., Zissman, M.A.: Measuring the readability of automatic speech-to-text transcripts. Proceedings of the Eurospeech Conference pp. 1585–1588 (2003)

[75] Jutten, C., Herault, J.: Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Signal Processing 24(1), 1–10 (1991)

[76] Kamondi, A., Acsady, L., Wang, X.J., Buzsáki, G.: Theta oscillations in somata and dendrites of hippocampal pyramidal cells in vivo: activity-dependent phase-precession of action potentials. Hippocampus 8, 244–261 (1998)

[77] Kandel, E.R., Schwartz, J.H., Jessell, T.M.: Principles of neural science. Appleton and Lange (2000)

[78] Kharitonov, E., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T.A., Rivière, M., Mohamed, A., Dupoux, E., Hsu, W.N.: Text-free prosody-aware generative spoken language modeling. arXiv 2109.03264 (2021)

[79] Khosla, M., Ngo, G.H., Jamison, K., Kuceveski, A., Sabuncu, M.R.: Cortical response to naturalistic stimuli is largely predictable with deep neural networks. Science Advances 8(22), doi.org/10.1126/sciadv.abe7547 (2021)

[80] Ki, J.J., Kelly, S.P., Parra, L.C.: Attention strongly modulates reliability of neural responses to naturalistic narrative stimuli. The Journal of Neuroscience 36(10), 3092–3101 (2016)

[81] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems pp. 1097–1105 (2012)

[82] Knierim, J.J., Kudrimoti, H.S., McNaughton, B.L.: Place cells, head direction cells, and the learning of landmark stability. The Journal of Neuroscience 15(3), 1648–1659 (1995)

[83] Kriegeskorte, N., Mur, M., Bandettini, P.: Representational similarity analysis— connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2, 1–28 (2008)

[84] Kroll, J.F., Sunderman, G.: Cognitive processes in second language learners and bilinguals: the development of lexical and conceptual representations. In: Doughty, C.J. (ed.) The Handbook of Second Language Acquisition. Wiley (2003)

[85] Kumar, M., Anderson, M.J., Antony, J.W., Baldassano, C., Brooks, P.P., Cai, M.B., Chen, P.H., Ellis, C.T., Henselman-Petrusek, G., Huberdeau, D., Hutchinson, J.B., Li, Y.P., Lu, Q., Manning, J.R., Mennen, A.C., Nastase, S.A., Richard, H., Schapiro, A.C., Schuck, N.W., Shvartsman, M., Sundaram, N., Suo, D., Turek, J.S., Turner, D., Vo, V.A., Wallace, G., Wang, Y., Williams, J.A., Zhang, H., Zhu, X., Capota, M., Cohen, J.D., Hasson, U., Li, K., Ramadge, P.J., Turk-Browne, N.B., Willke, T.L., Norman, K.A.: BrainIAK: the brain image analysis kit. Aperture In press (2021)
[87] Landauer, T.K., Dumais, S.T.: A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240 (1997)

[88] LeCun, Y., Bengio, Y.: Convolutional neural networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks 3361(10), 1–14 (1995)

[89] Lee, K.H., Kung, S.Y., Verma, N.: Low-energy formulations of support vector machine kernel functions for biomedical sensor applications. Journal of Signal Processing Systems 69(3), 339–349 (2012)

[90] Leutgeb, J.K., Leutgeb, S., Moser, M.B., Moser, E.I.: Pattern separation in the dentate gyrus and CA3 of the hippocampus. Science 315(5814), 961 (2007)

[91] Lisman, J.: The theta/gamma discrete phase code occurring during the hippocampal phase precession may be a more general brain coding scheme. Hippocampus 15, 913–922 (2005)

[92] Manning, J.R.: Context reinstatement. In: Kahana, M.J., Wagner, A.D. (eds.) Handbook of Human Memory. Oxford University Press (2020)

[93] Manning, J.R., Hultert, J.C., Williams, J., Piloto, L., Sahakyan, L., Norman, K.A.: A neural signature of contextually mediated intentional forgetting. Psychonomic Bulletin and Review 23(5), 1534–1542 (2016)

[94] Manning, J.R., Jacobs, J., Fried, I., Kahana, M.J.: Broadband shifts in LFP power spectra are correlated with single-neuron spiking in humans. The Journal of Neuroscience 29(43), 13613–13620 (2009)

[95] Manning, J.R., Polyn, S.M., Baltuch, G., Litt, B., Kahana, M.J.: Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. Proceedings of the National Academy of Sciences, USA 108(31), 12893–12897 (2011)

[96] Manning, J.R., Ranganath, R., Norman, K.A., Blei, D.M.: Topographic factor analysis: a Bayesian model for inferring brain networks from neural data. PLoS One 9(5), e94914 (2014)

[97] Manning, J.R., Sperling, M.R., Sharan, A., Rosenberg, E.A., Kahana, M.J.: Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clustering during memory search. The Journal of Neuroscience 32(26), 8871–8878 (2012)

[98] Manning, J.R., Zhu, X., Willke, T.L., Ranganath, R., Stachenfeld, K., Has- son, U., Blei, D.M., Norman, K.A.: A probabilistic approach to discovering dynamic full-brain functional connectivity patterns. NeuroImage 180, 243–252 (2018)

[99] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv 1301.3781 (2013)

[100] Mikolov, T., Yih, W.T., Zweig, G.: Linguistic regularities in continuous space word representations. Proceedings of the National Association for Computational Linguistics pp. 746–751 (2013)

[101] Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A.: Predicting human brain activity associated with the meanings of nouns. Science 320(5880), 1191 (2008)
[102] Mnih, A., Hinton, G.: A scalable hierarchical distributed language model. Advances in Neural Information Processing Systems (2009)
[103] Mota, B., Herculano-Houzel, S.: All brains are made of this: a fundamental building block of brain matter with matching neuronal and glial masses. Frontiers in Neuroanatomy 8(127), doi.org/10.3389/fnana.2014.00127 (2014)
[104] Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., Malach, R.: Coupling between neuronal firing, field potentials, and fMRI in human auditory cortex. Science 309(5736), 951–954 (2005)
[105] Nastase, S.A., Halchenko, Y.O., Connolly, A.C., Gobbini, M.I., Haxby, J.V.: Neural responses to naturalistic clips of behaving animals in two different task contexts. Frontiers in Neuroscience 12(316), doi.org/103389/fnins.2018.00316 (2018)
[106] Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. Journal of the Royal Statistical Society 135(3), 370–384 (1972)
[107] Nelson, D.L., McEvoy, C.L., Schreiber, T.A.: The University of South Florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, and Computers 36(3), 402–407 (2004)
[108] Noble, W.S.: Support vector machine applications in computational biology. Kernel Methods in Computational Biology pp. 71–92 (2004)
[109] Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10(9), 424–430 (2006)
[110] Owen, L.L.W., Chang, T.H., Manning, J.R.: High-level cognition during story listening is reflected in high-order dynamic correlations in neural activity patterns. Nature Communications 12(5728), doi.org/10.1038/s41467–021–25876–x (2021)
[111] Owen, L.L.W., Muntianu, T.A., Heusser, A.C., Daly, P., Scangos, K.W., Manning, J.R.: A Gaussian process model of human electrocorticographic data. Cerebral Cortex 30(10), 5333–5345 (2020)
[112] Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F.: Reconstructing speech from human auditory cortex. PLoS Biology 10(1), e1001251 (2012)
[113] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems pp. 8026–8037 (2019)
[114] Pearson, K.: On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572 (1901)
[115] Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2014)
[116] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv 1802.05365 (2018)
[117] Poldrack, R.A., Wagner, A.D., Prull, M.W., Desmond, J.E., Glover, G.H., Gabrieli, J.D.: Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. NeuroImage 10, 15–35 (1999)
[118] Polyn, S.M., Miller, J.F., Koskin, V.S., Morton, N.W., Kahana, M.J.: Tracking category-related neural patterns during free recall using scalp EEG. Society for Neuroscience (2007)
[119] Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A.: Category-specific cortical activity precedes retrieval during memory search. Science 310, 1963–1966 (2005)
[120] Potes, C., Brunner, P., Gunduz, A., Knight, R.T., Schalk, G.: Spatial and temporal relationships of electrocorticographic alpha and gamma activity during auditory processing. NeuroImage 97, 188–195 (2014)
[121] Pouget, A., Dayan, P., Zemel, R.: Information processing with population codes. Nature Reviews Neuroscience 1, 125–132 (2000)
[122] Proix, T., Saa, J.D., Christen, A., Martin, S., Pasley, B.N., Knight, R.T., Tian, X., Poeppe, D., Doyle, W.K., Devinsky, O., Arnal, L.H., Mégévand, P., Giraud, A.L.: Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. Nature Communications 13(48), doi.org/10.1038/s41467–021–27725–3 (2022)
[123] Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I.: Invariant visual representation by single neurons in the human brain. Nature 435(23), 1102–1107 (2005)
[124] Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
[125] Radford, A., Wu, J., Child, R., Luo, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog 1(8) (2019)
[126] Rasmussen, C.E.: Gaussian processes for machine learning. MIT Press (2006)
[127] Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., Fusi, S.: The importance of mixed selectivity in complex cognitive tasks. Nature (2013)
[128] Rolls, E.T.: The mechanisms for pattern completion and pattern separation in the hippocampus. Frontiers in Systems Neuroscience 7(4), 1–21 (2013)
[129] Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: uses and interpretations. NeuroImage 52, 1059–1069 (2010)
[130] Sabuncu, M.R., Singer, B.D., Conroy, B., Bryan, R.E., Ramadje, P.J., Haxby, J.V.: Function-based intersubject alignment of human cortical anatomy. Cerebral Cortex 20(1), 130–140 (2010)
[131] Scangos, K.W., Khambhati, A.N., Daly, P.M., Owen, L.L.W., Manning, J.R., Ambrose, J.B., Austin, E., Dawes, H.E., Krystal, A.D., Chang, E.F.: Biomarkers of depression symptoms defined by direct intracranial neurophysiology. Frontiers in Human Neuroscience In press (2021)
[132] Schapire, R.E.: The boosting approach to machine learning: an overview. Nonlinear Estimation and Classification pp. 149–171 (2003)

[133] Sederberg, P.B., Kahana, M.J., Howard, M.W., Donner, E.J., Madsen, J.R.: Theta and gamma oscillations during encoding predict subsequent recall. The Journal of Neuroscience 23(34), 10809–10814 (2003)

[134] Sederberg, P.B., Schulze-Bonhage, A., Madsen, J.R., Bromfield, E.B., Litt, B., Brandt, A., Kahana, M.J.: Gamma oscillations distinguish true from false memories. Psychological Science 18(11), 927–932 (2007)

[135] Sederberg, P.B., Schulze-Bonhage, A., Madsen, J.R., Bromfield, E.B., McCarthy, D.C., Brandt, A., Tully, M.S., Kahana, M.J.: Hippocampal and neocortical gamma oscillations predict memory formation in humans. Cerebral Cortex 17(5), 1190–1196 (2007)

[136] Sejnowski, T.J., Churchland, P.S., Movshon, J.A.: Putting big data to good use in neuroscience. Nature Neuroscience 17, 1440–1441 (2014)

[137] Sekihara, K., Nagarajan, S.S., Poeppel, D., Marantz, A., Miyashita, Y.: Application of an MEG eigenspace beamformer to reconstructing spatio-temporal activities of neural sources. Human Brain Mapping 15(4), 199–215 (2002)

[138] Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., Just, M.A.: Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS One e1394, 1–9 (2008)

[139] Simony, E., Chang, C.: Analysis of stimulus-induced brain dynamics during naturalistic paradigms. NeuroImage 216, 116461 (2020)

[140] Simony, E., Honey, C.J., Chen, J., Hasson, U.: Dynamic reconfiguration of the default mode network during narrative comprehension. Nature Communications 7(12141), 1–13 (2016)

[141] Sizemore, A.E., Giusti, C., Kahn, A., Vettel, J.M., Betzel, R.F., Bassett, D.S.: Cliques and cavities in the human connectome. Journal of Computational Neuroscience 44(1), 115–145 (2018)

[142] Skaggs, W.E., McNaughton, B.L., Wilson, M.A., Barnes, C.A.: Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. Hippocampus 6, 149–172 (1996)

[143] Smith, E., Kellis, S., House, P., Greger, B.: Decoding stimulus identity from multi-unit activity and local field potentials along the ventral auditory stream in the awake primate: implications for cortical neural prostheses. Journal of Neural Engineering 10(1), 16010 (2013)

[144] Spearman, C.: General intelligence, objectively determined and measured. American Journal of Psychology 15, 201–292 (1904)

[145] Sporns, O., Betzel, R.F.: Modular brain networks. Annual Review of Psychology 67, 613–640 (2016)

[146] Sporns, O., Honey, C.J.: Small worlds inside big brains. Proceedings of the National Academy of Sciences, USA 103(51), 19219–19220 (2006)

[147] Steyvers, M., Shiffrin, R.M., Nelson, D.L.: Word association spaces for predicting semantic similarity effects in episodic memory. In: Healy, A.F. (ed.) Cognitive Psychology and Its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer. American Psychological Association, Washington, DC (2004)
[148] Stone, Z., Zickler, T., Darrell, T.: Autotagging Facebook: social network context improves photo annotation. IEEE Xplore p. doi.org/10.1109/CVPRW.2008.4562956 (2018)

[149] Studholme, C., Constable, R.T., Duncan, J.S.: Accurate alignment of functional EPI data to anatomical MRI using a physics-based distortion model. IEEE Transactions on Medical Imaging 19(11), 1115–1127 (2000)

[150] Tolman, E.C.: Cognitive maps in rats and men. Psychological Review 55, 189–208 (1948)

[151] Torgerson, W.S.: Theory and methods of scaling. Wiley, New York, NY (1958)

[152] Turner, B.M., Forstmann, B.U., Wagenmakers, E.J., Brown, S.D., Sederberg, P.B., Steyvers, M.: A Bayesian framework for simultaneously modeling neural and behavioral data. NeuroImage 72, 193–206 (2013)

[153] Wan, E.A.: Neural network classification: a Bayesian interpretation. IEEE Transactions on Neural Networks pp. 303–305 (1990)

[154] Wang, Y., Guo, Y.: A hierarchical independent component analysis model for longitudinal neuroimaging studies. NeuroImage 189(1), 380–400 (2019)

[155] Westerhuis, J.A., Kourti, T., MacGregor, J.F.: Analysis of multiblock and hierarchical PCA and PLS models. Journal of Chemometrics 12, 301–321 (1998)

[156] Wieting, J., Kiela, D.: No training required: exploring random encoders for sentence classification. arXiv 1901.10444 (2019)

[157] Wilpon, J.G., Rabiner, L.R., Lee, C.H., Goldman, E.R.: Automatic recognition of keywords in unconstrained speech using hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing 38(11), 1870–1878 (1990)

[158] Xie, T., Cheong, J.H., Manning, J.R., Brandt, A.M., Aronson, J.P., Jobst, B.C., Bujarski, K.A., Chang, L.J.: Minimal functional alignment of ventromedial prefrontal cortex intracranial EEG signals during naturalistic viewing. bioRxiv doi.org/10.1101/2021.05.10.443308 (2021)

[159] Yacoub, E., Harel, N., Ügurbil, K.: High-field fMRI unveils orientation columns in humans. Proceedings of the National Academy of Sciences, USA 105(30), 10607–10612 (2008)

[160] Yassa, M.A., Stark, C.E.L.: Pattern separation in the hippocampus. Trends in Neurosciences 34(10), 515–525 (2011)

[161] Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zollei, L., Polimieni, J.R., Fischl, B., Liu, H., Buckner, R.L.: The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of Neurophysiology 106(3), 1125–1165 (2011)

[162] Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C.J., Hasson, U.: Same story, different story: the neural representation of interpretive frameworks. Psychological Science 28(3), 307–319 (2017)

[163] Zanzotto, F.M.: Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research 64, 243–252 (2019)
[164] Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of the International Conference on Computational Linguistics. pp. 2335–2344 (2014)

[165] Zhang, Q., Xiao, Z., Huang, C., Hu, S., Kulkarni, P., Martinez, E., Tong, A.P., Garg, A., Zhou, H., Chen, Z., Wang, J.: Local field potential decoding of the onset and intensity of acute pain in rats. Scientific Reports 8(8299), doi.org/10.1038/s41598–018–26527–w (2018)

[166] Ziman, K., Heusser, A.C., Fitzpatrick, P.C., Field, C.E., Manning, J.R.: Is automatic speech-to-text transcription ready for use in psychological experiments? Behavior Research Methods 50, 2597–2605 (2018)