Artificial and Biological Neural Systems

Theory Notes

Giovanni Valer

a.y. 2023/24

These notes are for the course "Artificial and Biological Neural Systems" held by professor Uri Hasson.
I have included all the topics seen during the lectures and tried my best to avoid mistakes. I have also included some digressions (i.e. non-mandatory topics), marking them as such and with a yellow background. That being said, a very broad range of topics was to be covered, hence I do not take responsibility for any mistake or imperfection that might be present; still, it would be much appreciated if you reported any of those, so that I can fix the notes accordingly. In the GitHub repository[1] *you can find the LaTex source code, where you can create issues and pull requests to include topics and corrections.*
I hope that this document will help in the study of this beautiful and exciting subject.

If you consider this material valuable for you in order to be prepared for the exam, consider offering me a coffee :)
Paypal: @GiovanniValer (`https://www.paypal.me/GiovanniValer`)

XOXO - jo

[1] TODO

Contents

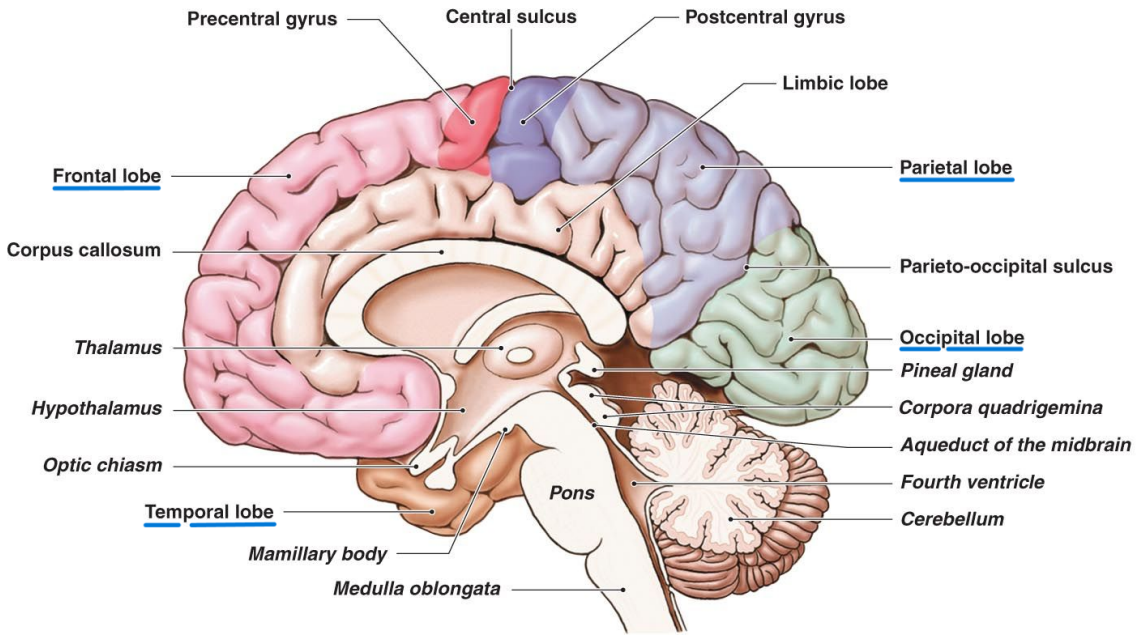
Chapter 1

Introduction to brain physiology and research methods

1.1 Basic facts about the human brain

The human brain is partitioned in many structures. The first one is the **cortex**, and in the following we provide the *traditional function specializations* ▲:

- **Temporal** lobe: **Auditory processing** (hearing and language)
- **Parietal** lobe: **Attention**, touch, saccade planning
- **Frontal** lobe: Planning, execution, **higher level cognition**, high level language processing and language production
- **Occipital** lobe: **Vision**: perception of visual features, categories and location



▲ Here we have the traditional function specializations, even though nowadays we know that all areas of the brain take part in more functions: a function is computed by the whole network. The auditory processing is a sort of exception: audio is processed only in the temporal lobe; however, it is still part of the network, as its functions can be *modified on demand*.

We then have the **subcortical structures**:

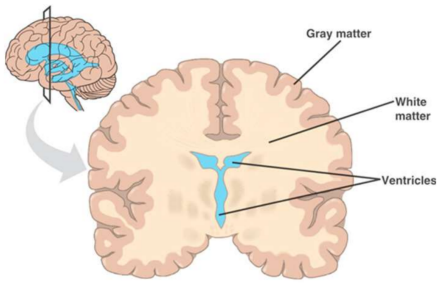
- **Cerebellum**: Fine motor control, implicated in emotional responses
- **Thalamus**: Major gateway for sensory processing. One of the final stops before sensory information arrives at the cortex

- **Hippocampus**: implicated in construction of memories; these are later transferred to other cortical regions
- **Corpus Callosum**: A main "highway" of white matter tracks that connects the two hemispheres

1.1.1 Gray matter

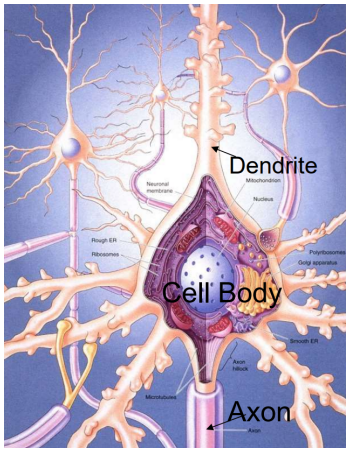
Gray matter (GM), which gets its name from its color, is the brain part that **performs computation**. The gray matter includes not only the cerebral cortex but also the cerebellum, basal ganglia, thalamus, and several other regions. Techniques for measuring brain activity often reflect the function of gray matter regions, which are responsible for the majority of cognitive processing. As a result, understanding the role and function of gray matter in the brain is essential for gaining insight into various neurological and psychiatric conditions.

The gray matter is on the perimeter of the brain, and its folded structure (with *gyri* and *sulci*, determined by the DNA) allows for higher surface. This also causes parts that are located close to each other in physical space, to be far away in cortical space.



1.1.2 White matter

White matter consists mainly of **long-range axon pathways**, or tracts, that **connect different regions of the brain**. Unlike gray matter, there is **no direct information processing** within white matter itself. The structure of white matter can change with learning because it reflects the long-range neural pathways that carry nerve impulses and facilitate the communication between different regions of the brain.



Damage to white matter, such as the loss of myelin in conditions like multiple sclerosis, can have significant impacts on communication between different regions of the brain, leading to various neurological and psychiatric symptoms. Therefore, understanding the role and function of white matter is critical for studying brain function and identifying potential targets for interventions in various neurological and psychiatric disorders.

1.1.3 The neuron

A neuron consists of **dendrites**, a **cell body** (aka **soma**) and an **axon**. Connections between neurons are called **synapses**, and typically occur on a neuron's dendrite (but in some cases also on soma), who receive synaptic signals. Synapses are **chemical**, **not electrical**. A neuron will *fire* (generate an action potential) depending on the number of signals it receives on its dendrites and their strength, which are *summed* in the neuron's body.

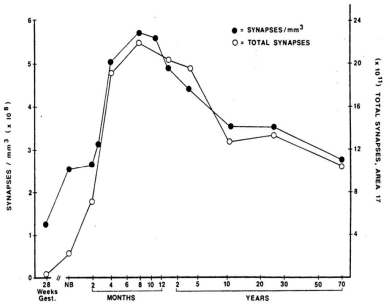
Digression: Synaptic communication

Synaptic communication is the process by which neurons communicate with each other through the release and reception of chemical signals called neurotransmitters. The postsynaptic neuron receives the signal, which is either excitatory (the neuron is depolarized and fires an action potential), or inhibitory (causing the neuron to hyperpolarize and become less likely to fire an action potential). Depolarization produces an *all-or-nothing* signal: **no partial firing**.

1.1.4 Neuroplasticity

Neuroplasticity refers to *network changes* over time, to adapt to the environment. Synapses can be *strengthened* or *pruned*.

Connections between neurons are consistently removed (or created) depending on use. A large percentage of neurons that develop die. A *hub* neuron, which receives inputs from many neurons is more likely to survive. When many neurons connect to a target neuron, this decreases their survival rate. This calibration is thought to be associated with generating an optimal degree of synaptic connection. Brain volume triples between birth and adulthood; this is mostly not due to addition of neurons, but to an increased number of connections (synapses), myelination of existing axons and greater dendritic branching.



Synaptic density and total synapses in visual cortex as a function of age.

1.2 Studying the human brain: tools and basics of experimentation

The human brain can be studied at different levels of organization, from systems and pathways to synapses and membranes.

- Systems and pathways: large-scale neural networks responsible for specific functions, such as sensory perception, motor control, and cognition. These may be topographically distributed.
- Circuits and neurons: networks of interconnected neurons that underlie information processing within the brain.
- Synapses and membranes: molecular and cellular mechanisms that govern the transmission of signals between neurons, such as the release of neurotransmitters and activation of ion channels.

Studying the brain at different scales provides insights into organization of function from the macroscopic to the microscopic level.

1.2.1 Basics of experimentation

There are several **non-invasive tools** available for studying brain activity and structure, including electroencephalography, structural imaging, functional magnetic resonance imaging (fMRI), and diffusion-weighted imaging. These tools provide insights into different aspects of brain function and structure: electrical activity of neurons, the structural connectivity of brain regions, and the metabolic activity (i.e. energy consumption) associated with specific tasks or behaviors.

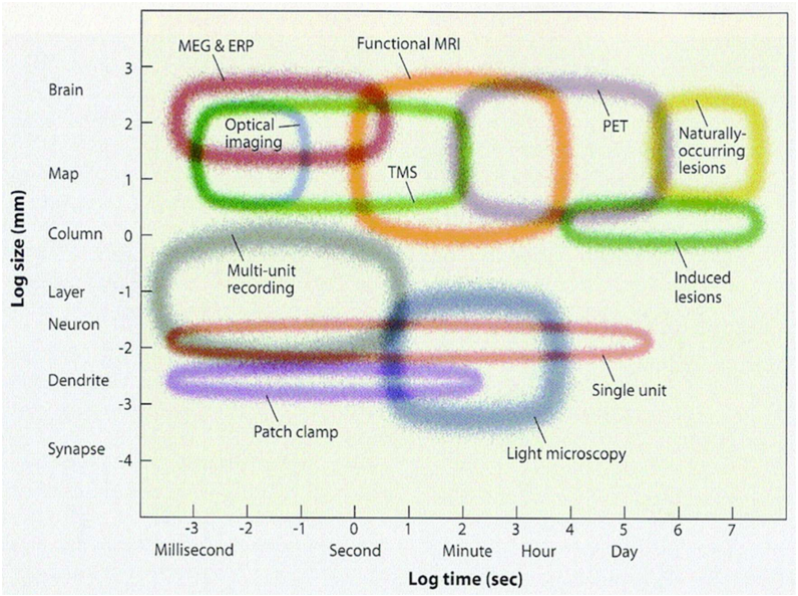


Figure 1.1: **Temporal** and **spatial resolution** of common tools. An additional dimension is **coverage**: how much of the brain is simultaneously observable by the tool.

Experimental procedures are used to analyze the data collected by these tools, which typically consists in conducting studies that involve manipulating variables of interest and collecting and analyzing data from participants. In neuroscience, an **experiment** or study is a systematic **investigation of a research hypothesis** that involves **manipulating variables and measuring their effects on some outcome of interest**. Conclusions from experiments are drawn by analyzing the data collected from participants and testing whether there are statistically significant differences between groups or conditions, typically using statistical methods to quantify the strength and direction of effects and assessing the probability that the observed effects are due to chance.

In the following sections we will discuss the most important tools (namely EEG, structural imaging, fMRI), however there are many others:

- TMS (transcranial magnetic stimulation): induces a virtual lesion of a part of the brain (might be dangerous).
- MEG (Magnetoencephalography): measures the magnetic fields generated by neural activity directly.
- Patch clamp: records current from ion channels in cell membrane

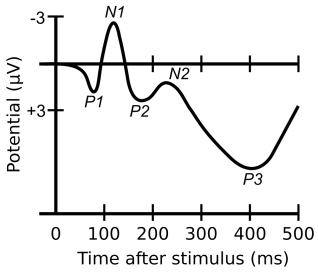
1.2.2 Electroencephalography (EEG)

EEG is a non-invasive method used to study patterns of brain activity with high temporal resolution. The principle behind EEG is that it is sensitive to very subtle changes in electric potentials below the sensors, which are propagated to the scalp. These changes reflect alterations in the electrical environment of thousands of neurons that fire in synchrony. Each EEG sensor gives one time series. It is difficult to pinpoint the brain regions causing the fluctuations, since the electromagnetic waves are dispersed by the scalp. However, the **timing of the signals is very precise**.

From EEG time series to ERP

Tasks can generate **stereotyped evoked potentials** that can be averaged (over all epochs ▲) to obtain an **Event-Related Potential**" (aka **Evoked-Response-Potential**). An ERP analysis quantifies electrical brain responses to events/stimuli based on time-locked EEG portions. This analysis can be used as the basis for more sophisticated analysis such as source localization.

▲ An epoch consists in the timespan [0ms, 500ms] after the stimulus presentation.



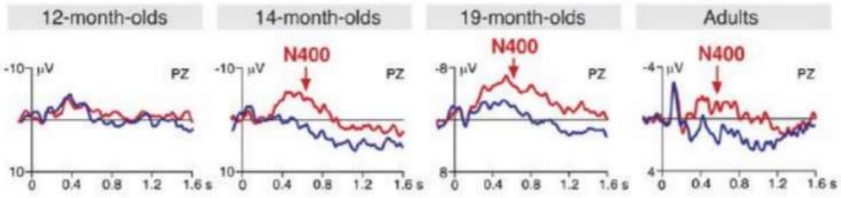
A waveform showing several ERP components. Notice the plot has negative voltages upward.

Neural activity is not the only activity causing electrical fluctuations. Muscles are also one of the main causes of fluctuations (e.g. every time we blink or move our eyes there are oscillations collected by EEG sensors). All these **noise** artifacts have to be independently measured and then removed from the data. The sensors placed near the eyes are indeed used to measure such noise. Moreover, since the interesting brain **signal is low**, while the **noise is high**, **many repetitions** of each condition are needed (the impact of noise on computing ERPs scales down as a function of the square root of the number of observations). This noise reduction applies to each timepoint measured. The need for many repetitions can be challenging and time-consuming, but it is necessary for obtaining reliable and statistically significant results. In addition to repetition, other techniques such as filtering and artifact rejection can also help to reduce noise in EEG recordings.

EEG also contains important information in form of frequency characteristics (cycling rate). The frequency differentiates sleep stages from awake, and drowsy from alert. Power plots can show the relative strength of each frequency, helping in *frequency band interpretation*. For instance, the alpha band in EEG, which has a frequency range of 8-12 Hz, is commonly observed during eyes-closed recordings and relaxed states. Alpha oscillations are associated with reduced communication between the cortex and thalamus. During externally oriented attention and stimulus processing, the alpha activity is suppressed.

ERP in practice

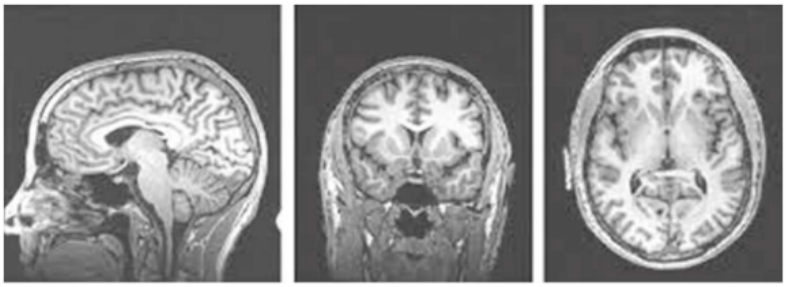
This experiment tries to understand when children develop the ability to predict language. It consists in presenting congruous (e.g. *pizza was too hot to eat*, in blue) or incongruous (e.g. *pizza was too hot to* ***sit***, in red) sentences.



Here we see the ERPs from a single sensor (the PZ one). We can notice the time point where the two functions diverge is when the word is processed by the brain (so we can understand how much it takes for a word to be processed).

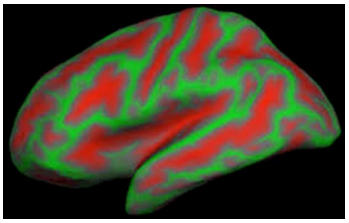
1.2.3 Structural imaging

Structural imaging involves collecting 3D images of the brain, similar to the images obtained in a medical setting.



This type of imaging can provide information about various aspects of brain structure, such as:

- Gray matter volume (i.e. the overall size of the gray matter in the brain)
- The density of gray matter in specific regions, which can approximate the concentration of neurons
- Cortical thickness at a resolution of a few millimeters
- Surface area of particular brain regions



Example of cortical thickness map.

Calculation of cortical thickness is usually done by converting the brain's 3D representation to a 2D sheet representation. We can look for correlation (covariance across different people) of structural cortical thickness between different areas of the brain.

1.2.4 Functional magnetic resonance imaging (fMRI)

Functional MRI is the method that helped most in uderstanding how the brain works. It consists in observing which parts of the brain are involved in **metabolic activity** when we do things like thinking or perceiving ▲.
FMRI uses a big magnet to affect protons in the brain and then measures how they behave as they return to their original state (areas that are involved in oxygen consumption have a different relaxation profile). By analyzing the patterns of proton behavior, we can identify which brain areas are more active during certain tasks. This method gives us very detailed information about where activity is happening in the brain and can help us understand how different regions contribute to complex processes like thinking and perception. With fMRI we get a time series for each **voxel** (typically a $3 \times 3 \times 3$ mm brain region).

▲ In neurology, metabolic activity is often used as a proxy for neural activity: active neurons require more energy to function and can increase their metabolic rate.

In our brain there are 60 to 70 thousands voxels. By chance (statistics) we will surely get some false positives. For this reason we undertake many experiments.

fMRI in an experimental context

The **signal** is defined as the measurable response to a stimulus.
Statistical detection theory:

- Understand relationship between stimulus and signal
- Describe noise properties statistically
- Devise methods to distinguish noise-only measurements from signal+noise measurements, and assess the methods' reliability

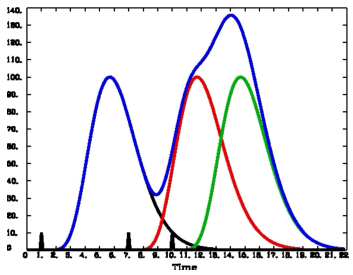
Stimulus-signal connection and *noise statistics* are complex and poorly characterized. We devise, a priori, a **mathematical model connecting stimulus** (or "activation") **to signal** (typically a regression model). We make an estimation of the **statistical model for the noise**, taking into account whether it is random, structured, oscillatory, etc. These two models are then combined to produce an **equation for measurements given signal+noise** (also often a regression model). The final result is an equation with few **free parameters that can be fitted to the data**.

FMRI analysis often fits a (convolved) activation model to **each voxel's time series separately** (*massively univariate analysis*). Pre-processing techniques are applied to reduce noise, including spatial smoothing across nearby voxels. The outcome of model fitting is a collection of parameters estimated from each voxel's data. The Activation Amplitude (Beta) is the most critical parameter, and it is related to the correlation between the model and activity. At the group level, the **voxel-level estimates are pooled together to reach a group-level conclusion per voxel**.

FMRI measures changes in neural activity, there are not absolute magnitudes. The baseline signal level in a voxel does not provide information about neural activity. An experiment using fMRI requires at least two conditions to detect changes in neural activity. Minimally, experiments use a "task" and "rest" condition. In such a two-task experiment intermixed with rest, a beta value is estimated per task, and their values are contrasted to determine the difference between the two tasks.

Hemodynamic response function (HRF)

The response measured by fMRI (called "Hemodynamic" because it is related to blood) is delayed since the blood requires time to flow. Notice this is not a simply shifted impulse response, as it is a smooth function. For this reason, fMRI tells us **where** the process takes place, but not **when**. Combining fMRI and EEG can be a solution.



1.2.5 Diffusion weighted imaging (DTI)

Diffusion weighted imaging (also known as **Diffusion MRI** or **Diffusion tensor imaging**), is used to examine the structure of **white matter fibers**. For each voxel, the preferred **direction** of diffusion and the **strength** of diffusion are estimated to determine white matter tracts. These connections are considered *hardwired* connections. They can be cross-referenced against functional connectivity.

Chapter 2

Overview of ML approaches to modeling cognitive neuroscience data

In this Chapter we have two papers on the topic of **cognitive neuroscience models**.

2.1 *Analyzing biological and artificial neural networks: challenges with opportunities for synergy?* Barrett et al. (2019)

Deep neural networks brought a revolution in the area of ML, with millions of parameters, no engineered features, and very high performance. We can see an **analogy with neuroscience**, as both fields need to:

- understand how neural networks, consisting of large numbers of interconnected elements, transform representations of stimuli across multiple processing stages to implement a wide range of complex computations and behaviours;
- describe and analyze very high dimensional data.

The analogies appear in four areas: Receptive fields, Ablation, Dimensionality reduction, and Representational geometries.

2.1.1 Receptive fields

Neurons in the human visual cortex are **specialized to process stimuli in specific spatial areas** (see ◇ Retinotopy map) **or certain types of features**. The neurons in the **initial processing** regions of the visual cortex have **small receptive fields**; sensitive to stimuli in small areas of visual space. As information is transmitted to **higher level** areas of visual processing, **receptive fields become larger**, enabling sensitivity to larger areas of space. These regions also encode more complex features, and there is evidence of "abstract" coding with invariance to small transformations. There are "**concept cells**" **sensitive to identity** of objects but **not to appearance**. For example, simple "repetition priming ▲" effect repeated exactly with same face but different orientation.

▲ Repetition priming refers to the change in responding to a word or an object as a result of a previous encounter with that same item, either in the same task or in a different task.

(Some) AI researchers also think that **DNN neurons may code for specific information**, which can be studied via receptive field analysis.
Some experiments investigate which types of images maximally activate a neuron. Other studies examine how receptive fields change in deeper layers (as we go deeper in the network, each element in the feature map is occupying more space in the visual field, through pooling, but *in what way are larger receptive fields more complex?*).

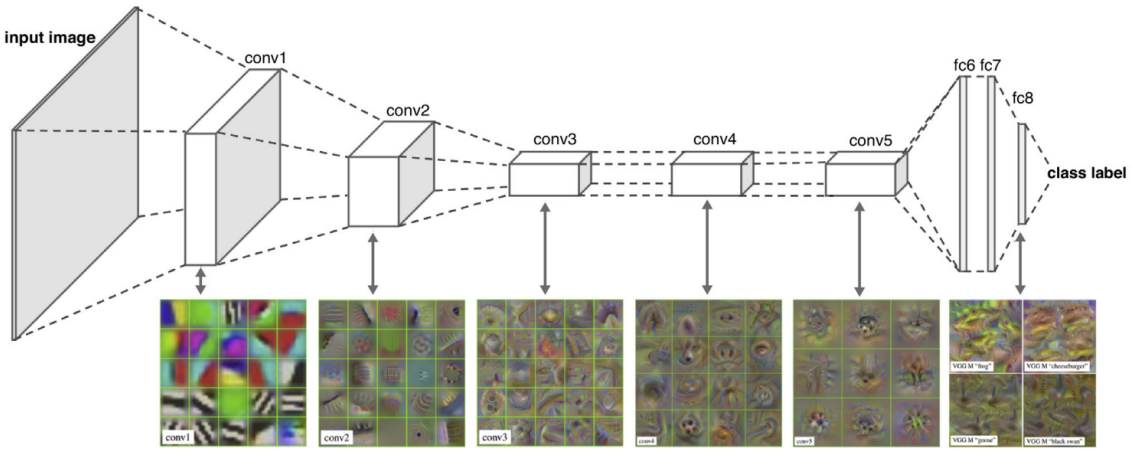
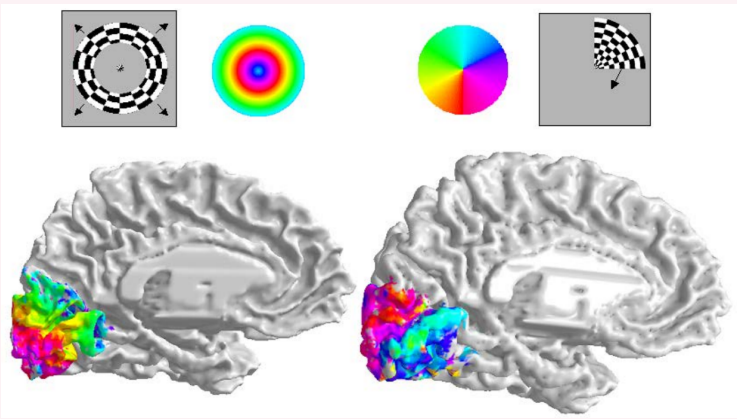


Figure 2.1: Top: how an input image is progressively processed by a CNN. Bottom: receptive fields for each layer are calculated using activation maximization (i.e. take an already trained CNN and create stimuli that produce the maximal activation in a given feature mapping in the network). For each layer, a single square corresponds to a feature map. We can notice `conv1` (the first layer) encodes large scale information; some shapes (e.g. circles) start appearing as early as `conv2`.

◆ **Retinotopy map**

The **occipital cortex** (**visual cortex**) is devoted to image processing. They found out that the coding in the brain is along two dimensions: **eccentricity** and **radial degree**. It is possible to get a map of such coding by applying fMRI to a relatively simple experiment:



Focus first on the experiment on the left. The person is looking at the center of the shape. While looking, the shape changes according to a defined temporal pattern (shrinking and expanding with a certain period). If there are neurons caring for a certain distance from the center, then they should fire at determinate time points. This is indeed the case, and we can track which neurons code for a certain eccentricity (we use colors to represent eccentricity). The experiment on the right is quite similar: while the person is looking at the center, the shape rotates around it.
Different people have almost all the same coding, and it is interesting to note how there are no "jumps" in the coding.

A similar thing can be done with the **primary auditory cortex**, in the **temporal lobe**, to get a **tonotopy map**.

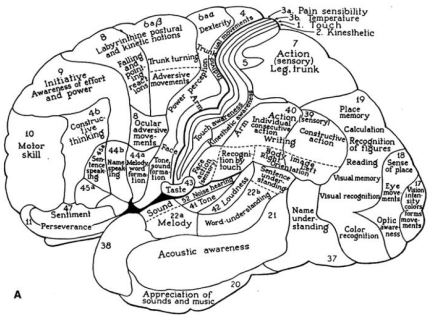
2.1.2 Ablation

Brain lesions (ablations) offer much information about potential function of brain areas. By mapping lesions to symptoms, we can understand which brain areas are important for given tasks.

Pruning and ablation causes performance deficits. However, thanks to neuroplasticity, the human brain can achieve again the same performance.

Speaking of artificial neural networks, the debate in the last years has seen three thesis on how DNNs encode information:

- distributed information (meaning we can remove some part without hardly impacting a single task),
- local encoding (meaning some neurons are super specialized and don't share encoding information),
- Modularity theory: not each single neuron is necessary, but some modules (groups of neurons) are necessary for specific tasks.



Kleist's functional brain map. This map is old and partially incorrect, but still interesting.

We ask ourselves, how do artificial neural networks change after ablation? The ablation (lesioning) analysis is applicable to DNNs. We can *silence* neurons and observe how this impacts the network output. Silencing of neurons is done via **structural pruning** (entirely removing a neuron with all its outgoing weights). Networks trained for generalization (out of sample prediction) are more robust to ablation than those trained on memorizing labels. Pruning and fine-tuning are active areas of research.

2.1.3 Dimensionality reduction

The brain codes information in a distributed manner, necessitating multivariate analysis:

- Multiple units encode information in the brain, leading to **redundancy** where two neurons may fire almost identically; or
- Information is coded in a **distributed** manner among multiple units (e.g. coding for 4 classes among 2 neurons, each coding $\{0, 1\}$); or
- **Correlation** among units could indicate that the **activity can be described in a lower dimensional space**.

It has been shown that, in DNNs, an object-by-feature matrix from fully connected layer can be **compressed by more than 80%** (e.g. from 4096 dimensions to 512) while maintaining almost all variance. This means, few low dimensions explain differences between images.

2.1.4 Representational geometries

To **understand representations**, we study how they are represented in different layers, how they change over time, and how are different embedding spaces related to each other.

Matrix factorization measures: they are a good starting point to compare matrices and understand if there is some sort of covariance between the two. We compare representations across networks by comparing **object-by-feature matrices**. *Canonical Correlation Analysis* and *PLScorrelation* are two different factorization measures, they identify lower-level factors that capture and maximize the correlations/covariance between the datasets (note: these methods can be seen as "supervised" as they re-weight columns in both tables to maximize similarity).

Representational Similarity Analysis: it involves **comparing two similarity matrices**, often constructed from object-by-feature matrices. This yields an object-by-object similarity matrix. Representational Similarity Analysis does not factorize matrices. The principle is: describe how the objects cluster together (in the representational space) according to the matrix. To code for distances between objects, we need a distance matrix; and to get this a distance function is needed (e.g. Euclidean, Mahalanobis, etc.), indeed each row in the object-by-feature matrix is a vector. So we can **turn the object-by-feature matrix into an object-by-object distance matrix** (pairwise distance between objects). For human data (how human represent things), we usually don't get an object-by-feature matrices (we are provided instead an object-by-object matrix), so with RSA we can study the **relationship between human representation and machine representation**.

Linear regression: The machine representation in the neural network (*DNN embeddings*) can be used to predict brain activity (*neurobiological activation vectors*) using linear regression.

2.2 *What does the mind learn? A comparison of human and machine learning representations* Spicer and Sanborn (2019)

This paper reviews the modern machine learning techniques and their use in models of human mental representations, detailing three notable branches: spatial methods, logical methods and artificial neural networks.

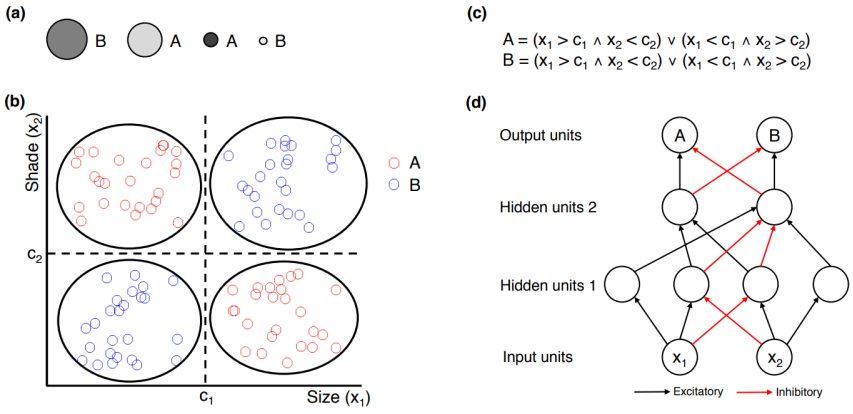


Figure 2.2: Illustration of the XOR classification task using size and shade: examples of stimuli in **(a)**. This problem can be solved by a spatial clustering method **(b)**, a logical Boolean method **(c)**, and an artificial neural network **(d)**.

2.2.1 Spatial methods

Spatial methods involve placing items in a **multidimensional space** and using their location to draw conclusions about **categorization**. Classification based on spatial methods can be determined by an item's location relative to a hyperplane or its similarity to different prototypes (means) or exemplars (centroids):

- **Prototype approaches** assume that learning is based on similarity to the center of a category (mean), which is stored after training (i.e. all examples are forgotten, the prototype only is stored).
- **Exemplar approaches** calculate similarity as a ratio between the similarity of an item to all items within a class (i...n) and the similarity of that item to all other items. This provides a fit per class and requires storing item-level information (i.e. all examples are kept).
- **Clustering** organizes items into groups (cohorts), with quality often quantified by the distance between items within and between clusters. Clustering can be either hard (each item belongs to only one cluster) or soft (items can have multiple memberships, potentially fuzzy).

Example of spatial method: Generalized Context Model (GCM)

According to the GCM, the probability that stimulus i is classified into category C_J is found by summing the similarity of i to all training exemplars of C_J and then dividing by the summed similarity of i to all training exemplars of all categories:

$$P(C_J|i) = \frac{\left(\sum_{j \in J} s_{ij}\right)^{\gamma}}{\sum_K \left(\sum_{k \in K} s_{ik}\right)^{\gamma}}$$

2.2.2 Logical methods and Artificial neural networks (ANNs)

Logical methods: concepts are based on a definition that is applied to the features of the object. One viable solution is searching for rules that maximizes discrimination between stimuli; the rules can be probabilistic (i.e., the rules, if not given, can be automatically computed/learned).

ANNs: do not make assumptions about the representations involved, but offer an implementation method.

2.2.3 Thoughts

Asking which model is the most accurate might be misleading: the answer depends on what area of science you work in. Therefore it is better to focus on whether a model offers "useful explorations of the ways in which human learning operates". The value depends not just on match to human behavior, but whether there is a need to understand the underlying representations. We have to consider **not just accuracy**, but **also confusion**. Today, very large emphasis is put on **explainability**: we want to explain the model behaviors (for naturally explainable models). And if it is not naturally explainable, we need to address this issue.

Chapter 3

Psychology of concepts and categories

In this Chapter we discuss about categories in sensation/perception (non-semantic categories, Section 3.1) and conceptual structure (words and concepts, Section 3.2).

3.1 Categories and categorical perception

Categorical perception is the phenomenon in which people perceive stimuli from different categories as more different from each other than stimuli from within the same category. This is useful as it introduces invariance in response with respect to a functionally defined category, allowing for rapid prediction, efficient memory, and compression.

How we know categorical representation exists

A demonstrating experiment consists in:

- selecting a set of stimuli that uniformly covers a certain physical domain (e.g. sound freq 100Hz-8000Hz),
- select an objective distance measure so that the space is partitioned in intervals; e.g. distance in frequency space (applicable to both sounds and colors),
- select a method for operationalizing human similarity (e.g. similarity judgments, generalization, confusion [same/different]),
- in one procedure, assign all stimuli to categories (e.g. assign all stimuli to color names); in a second, obtain similarity judgments for within-category vs. between-category pairs, or ask for categorization, and evaluate if the boundary is fuzzy or not (e.g. how much objects are considered similar when belonging to the same category, and when to different categories)

3.1.1 Categorical perception in audition

In auditory stimuli, there is discrimination between speech sounds. People have a sharper discrimination boundary between sounds that are perceived as belonging to different phonetic categories than between sounds that are perceived as belonging to the same category. For experimenting, we use as objective dimension the **Voice Onset Time** (VOT) of consonants (i.e. the timespan between the start of the consonant and the start of sound emitted by vocal cords). The discrimination performance is simply a "same/different" judgment. We present consonants such as /b/ and /p/. A fixed-size physical difference in VOT, that is easily discriminated when it straddles the boundary between two categories (labeled as /b/ or /p/), produces *chance* discrimination performance when both tokens come from the same category (either both /b/ or both /p/); results are in Figure 3.1.
Within phonetic category, two **different stimuli sound the same** (see ◇ Eimas et al. (1971)).

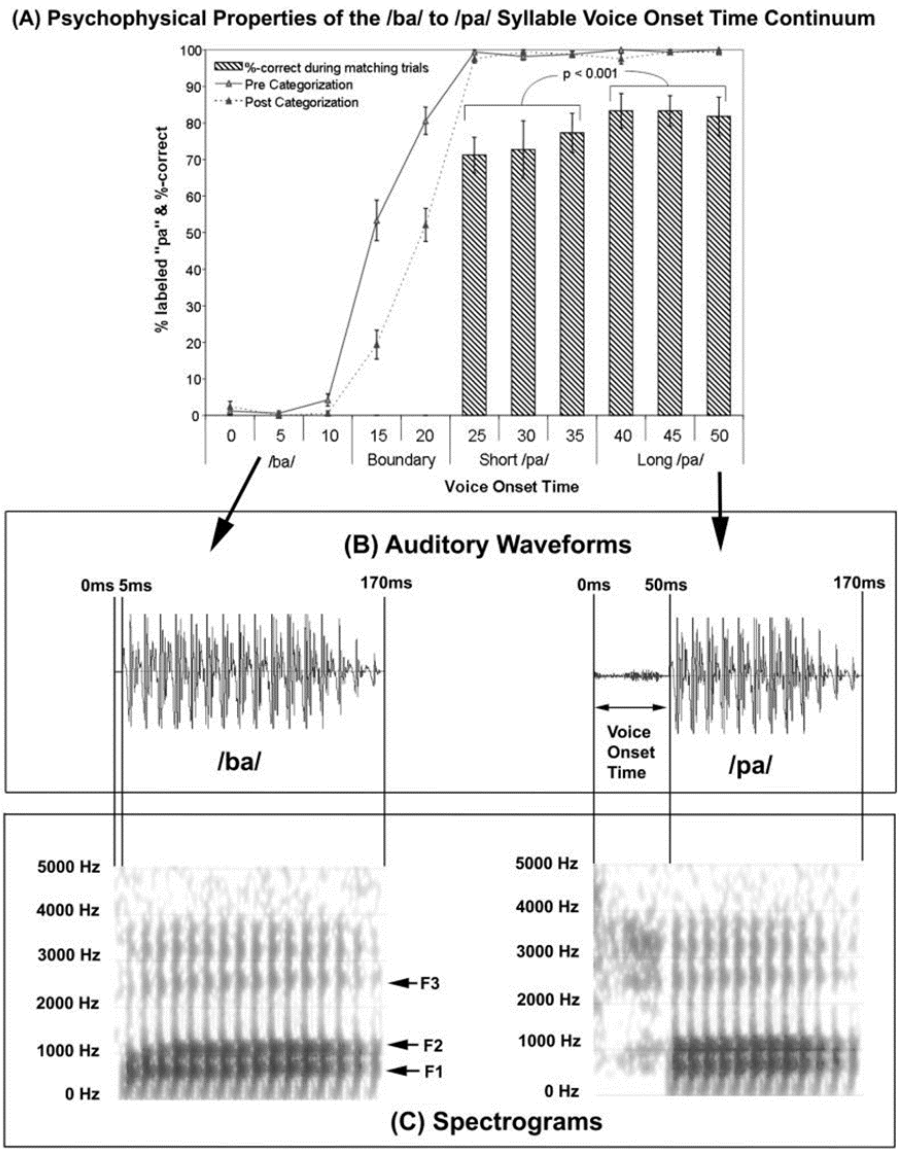


Figure 3.1: For the experiment, they created a range of stimuli with VOT between 5ms and 50ms (which are the standard values for respectively /ba/ and /pa/). In **(A)** are the results, with the percentage of responders hearing a /pa/ sound. We see the shift is between 10ms and 25ms

Prior categories aid online processing via categorical perception, in a sort of *experience-dependent learning*; related phenomena are:

- *Change deafness*, Vitevitch (2003): participants repeat words presented by a speaker. Halfway through study, the speaker changed. Only 40% of participants noticed the change.
- *Sine-wave speech*: phonetic categories/expectations can be considered *priors* on sounds, impacting whether a stimulus is perceived as speech (e.g. if you previously listen to a voice and then hear a sound which is not speech, but has some correlation to the previous voice, you can "hear" the words).
- *McGurk Effect*: the categorization of sounds is not only an auditory task, since our brain combines multimodal inputs (e.g. also from vision).

◆ **Eimas et al. (1971)**

They observed 4-month old sucking rate on pacifier (notice that a higher rate is interpreted as more surprise/interest). They examined the rate as function of relation between current and previously heard stimuli, in particular they presented two stimuli with VOT differing by 20ms. In one condition (labeled "D") the difference straddled (on two sides of) the border of a phonetic boundary (stimuli perceived as "b" and "p" by adults). In another condition "S" they belonged to the same phonetic category.

See Figure 3.2 for more details.

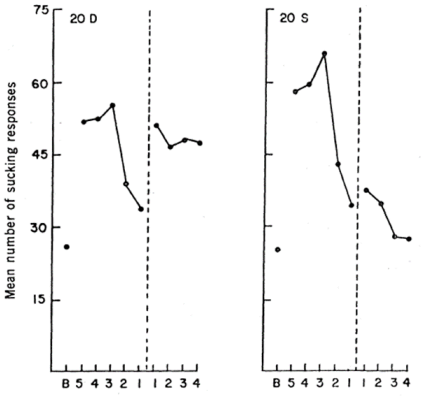
Brain areas coding for **low-level** representations are **not influenced** by such categorization, while those coding for high-level representations do.

3.1.2 Categorical perception in vision

In categorical perception for color, discrimination of items that cross category boundaries is better (faster, more accurate) than when the items are within the same color category. Notice that color category is **linguistic** ▲. For example, it is easier to distinguish between a green stimulus and a blue stimulus than between two stimuli within the same category (two shades of green), who are **spaced at the same distance**.

▲ Practical note: color differences in terms of discriminability can be equated across between-category and within-category comparisons by using the *Commision Internationale de L'Eclairage* (CIE) values.

Color categories are not universal, and thus **categorical perception depends on language** (see △ Robertson et al. (2000)).



(a) (b)

Figure 3.2: 5 min of habituation precede a 20 ms VOT change, either within the same **(a)** or across **(b)** phonetic category; habituation consists in hearing the /ba/ sound. This proves how 4-month-old children already have the auditory **categorization** enabling them to distinguish between /b/ and /p/ sounds.

◆ **Robertson et al. (2000)**

A stone-age tribe Berinmo uses "*nol*" as the color name that in English fall under both green and blue, so they have no categorical perception at the boundary between green and blue (no boundary in their language). On the contrary, they have a category boundary between "*nol*" and "*wor*" that does not exist in English as both sides are green. Berinmo people exhibit better discrimination of 32 cross-category items than 32 within-category items at the boundary between *nol* and *wor*. English speakers do not show categorical perception at this boundary.

3.2 Conceptual structure

3.2.1 Theory on words and concepts

The typical approach is to assume that words are associated with concepts or a network of conceptual representations. We need to first have a good theory of what the conceptual structure is like, and then we can see how this structure is used to represent meaning when referred to by words. Ultimately, a word is a sound or written pattern, and it is generally assumed that a single word corresponds to a single concept (this assumption would help with word-embedding models), but there are complications.

There is evidence proving that *form-to-meaning mapping* is not 1-to-1. This evidence is called **polysemy**: a phenomenon where a single word can have multiple meanings depending on the context in which it is used, such as "cinema" which can refer to different things in different contexts (e.g. *American cinema is naïve* vs *This cinema is ugly*).

According to Murphy, it is impossible for a single concept to fully capture the meaning of a polysemous word; the process of meaning extension is not a result of *natural change* (chaining) but a matter of **online derivation**. Online derivation does not prevent information from being also stored somewhere, at least for a short time, in order to understand the nearby context more effectively. Indeed, Klein and Murphy (2001) found evidence suggesting that the different senses of polysemous words can be stored, as they observed priming effects when a word was used twice in the same sense, and interference effects when the sense was switched.
Another possibility is that a word specifies a set of potentials that is then refined by context to determine which sense is intended.

We found out that **words are not concepts**. Such distinction between *meaning* and *lexical form* is also proven by *anomia*. It is a type of *aphasia* that results in the inability to retrieve the lexical form of a concept for production, even though the ability to recognize or define the term is maintained.

3.2.2 Representing the meaning of concepts in the brain

Chapter 4

Modeling conceptual organization

4.1 Modeling typicality

Deep Neural Networks Predict Category Typicality Ratings for Images
Lake et al. (2015)

TODO

4.2 Words-as-features as models of cognition

Predicting Human Brain Activity Associated with the Meanings of Nouns
Mitchell et al. (2008)

TODO

4.3 Modeling similarity spaces expressed in human behavior and brain responses

TODO

Chapter 5

Modeling human representational geometry

TODO

Chapter 6

Human Memory and Learning

TODO

Chapter 7

Language

TODO

Chapter 8

Attention

TODO

Bibliography

David GT Barrett, Ari S Morcos, and Jakob H Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55: 55–64, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2019.01.007. URL `https://www.sciencedirect.com/science/article/pii/S0959438818301569`. Machine Learning, Big Data, and Neuroscience.

Brenden M. Lake, Wojciech Zaremba, Rob Fergus, and Todd M. Gureckis. Deep neural networks predict category typicality ratings for images. In David C. Noelle, Rick Dale, Anne Warlaumont, Jeff Yoshimi, Teenie Matlock, Carolyn D. Jennings, and Paul P. Maglio, editors, *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015*, Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, pages 1243–1248. The Cognitive Science Society, 2015. Publisher Copyright: © Cognitive Science Society, CogSci 2015.All rights reserved.; 37th Annual Meeting of the Cognitive Science Society: Mind, Technology, and Society, CogSci 2015 ; Conference date: 23-07-2015 Through 25-07-2015.

Tom Michael Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai min Kevin Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191 – 1195, 2008. URL `https://api.semanticscholar.org/CorpusID:6105164`.

Jake Spicer and Adam N Sanborn. What does the mind learn? a comparison of human and machine learning representations. *Current Opinion in Neurobiology*, 55:97–102, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2019.02.004. URL `https://www.sciencedirect.com/science/article/pii/S095943881830103X`. Machine Learning, Big Data, and Neuroscience.