STUDYING CONSCIOUS AND UNCONSCIOUS VISION WITH FMRI: THE BOLD PROMISE

Thesis by

Julien Dubois

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2013

(Defended May 29, 2013)

To my wife Christine,

whose face I consciously perceive every morning when I open my eyes,

and it makes me so happy.

ACKNOWLEDGEMENTS

Every single day I wake up and I am grateful. Yes, even on those days while I was writing this thesis, an infamously grueling experience. I have always been free to pursue my dream, even if it has been in constant evolution. When I was in high school, my dream was to study and save the tigers with an organization like the World Wildlife Fund. Then, as my studies went on, my dream evolved into wanting to solve the puzzle of how our planet got to be what it is (I got a bachelor's and master's degree in earth and atmospheric sciences). My dream evolved further into wanting to understand how life started (I got a master's degree in biochemistry, focused on the origins of life). But my dream was not done maturing yet. I gave it a little more time to do so, and went traveling around the world with my favorite person on this planet. During my travels, I came across Christof Koch's 2004 book, *The Quest for Consciousness*, and my dream finally entered puberty. I needed to understand the brain! I needed to understand what made us do everything that we did, know everything that we knew, feel everything that we felt, discover everything that we discovered.

There are so many people (and institutions) to acknowledge who have accompanied me and my dream during childhood. My parents, who provided a peaceful and loving home for me to grow in, made everyday life easy so I could simply concentrate on getting a great education, and never interfered with my choices. My brother and sister, with whom I grew up, contributed to forging my character. My many teachers who saw something in me and pushed me in the right direction. The Ecole Normale Supérieure, a haven for budding scientists, offered me freedom in the form of a four year fellowship to do what I loved doing: learn. Pr. Harold Helgeson welcomed me for my first long research endeavor, a six month stint in his theoretical biogeochemistry lab at UC Berkeley; I discovered California, and most importantly, met my wife-to-be. Then Pr. Pier Luigi Luisi welcomed me for another 6-month soul-searching stunt, working on the Minimal Cell

project in his laboratory in Rome. Christof put an end to my scientific childhood by responding to an email I sent to him after reading *The Quest* back in the spring of 2006. I had boldly asked him if I could come learn neuroscience in his lab and work for him for a year. To my advantage, I still had funding from the Ecole Normale Supérieure. He invited me to come to the lab and meet with him. When we met, I remember him asking: are you here to surf or to do neuroscience? Understandable reaction, I was very tan and blond from two months working as a divemaster on a small island off the coast of Honduras… but I was serious about learning neuroscience, and this is why you are reading this thesis today.

I have so many people to thank for what I have become over the six years of my dream's drawn-out puberty years.

First and foremost, my three main scientific mentors. Christof Koch, for supporting me through the hardships of failed experiments, providing a good working environment, and saving me the worries of looking for funding and other distracting matters. Christof is greater than life when it comes to scientific curiosity and motivation, and I already greatly miss his flamboyant, contagious energy. I caught Christof in a period of many transitions; I have the utmost respect for how he dealt with them, always righteous. Ralph Adolphs was the other major figure who accompanied me behind the curtains (and now officially in the spotlight), through my time at Caltech. Ralph was Nao's postdoc advisor, and from the onset I worked together with him. For every new project that I started, I sought out Ralph's invariably insightful advice. I always considered myself a member of his extended family – Ralph's lab truly feels like a family, with biweekly meetings, weekly social hour, monthly lab dinners, other social events such as movie nights, and of course the annual lab camping trip to Catalina Island. My two DoktorVaters (both German speakers, as it turns out) at Caltech were true mentors, two role models to look up to: Christof for extravagant, unstoppable genius; Ralph for grounded, meticulous genius. I had a third, no less important science and life mentor, and no less a genius, who was by my side most of

those years, though not officially: Rufin VanRullen. I worked on many research projects with Rufin over the past six years which are not part of this thesis, but contributed to making me the scientist that I am today. Rufin was with me from the very beginning; it is with him that I did my very first neuroscience project in the summer of 2006 before I came to work at Caltech. I can only aspire to be as good a scientist as he is – always thinking critically and oh so sharply, with a curiosity for the brain's inner workings that goes far beyond his own field of research.

Then there were my committee members, who helped shape this thesis in the last years of my adventure. I learned about fMRI experimental design from John O'Doherty back in 2006, when I took his class. Six years later, I still have much to learn from him. I had never interacted much with Doris Tsao, though I had much admiration for her work, until the last few months, when we discovered that we had some very similar interests and started collaborating on a fMRI decoding project; I enjoyed her "charge ahead" approach to research. Get it done. Mike Tyszka is probably one of the few people in the world that thoroughly understands everything about fMRI, and I have asked him many a question over the years about imaging sequences and the like. At some point I wanted to do some calibrated fMRI using breath holding in the scanner – I did not follow through with it, but Mike was always there to lend a hand. I truly was blessed with the nicest and most competent committee one could wish for.

So many other people should be mentioned, but this section would be way too lengthy. In general, I want to acknowledge everyone in the CNS department, professors, students, postdocs, etc.; and Tanya, of course, who really is the one running the show. I decided against listing the names of the many students and postdocs with whom I shared great experiences, fruitful discussions, good meals, exercise at the gym, barbecues, hikes, camping trips, etc. It would have been too long a list, and I would have felt bad for involuntarily leaving out some names. If you read those lines and think your name should probably have appeared here, you are right. I am thinking of you.

Life is not all about the workplace. When I moved here and started working at Caltech, I spent a wonderful three years living at Christine's parents' house; Tom and Margaret took me in as a child of their own, and always made me feel like I had not really left my parents' home. I found yet another set of parents in Mary and Steve, Christine's godparents, who have always taken such good care of me. I feel so grateful for the love that has surrounded me here, thousands of miles from my kin.

And then there is Christine. To me she is the proof that Plato's myth of the Androgyne may not be a myth after all. She is my other half, she makes me feel whole. I do not have to be someone else to accommodate her. She is the most loving, attentive, caring, sensitive, creative, intelligent, and fun human being that I have ever met. She has opened my eyes to many new ways to live an even fuller life, and there are still many, many more years for us to make the best of this strange odyssey.

Granted, I do not yet understand how our universe and our planet came about, how life on Earth started, or how consciousness arises from our brains; but I am extremely thankful that all these things happened.

ABSTRACT

Waking up from a dreamless sleep, I open my eyes, recognize my wife's face and am filled with joy. In this thesis, I used functional Magnetic Resonance Imaging (fMRI) to gain insights into the mechanisms involved in this seemingly simple daily occurrence, which poses at least three great challenges to neuroscience: how does conscious experience arise from the activity of the brain? How does the brain process visual input to the point of recognizing individual faces? How does the brain store semantic knowledge about people that we know? To start tackling the first question, I studied the neural correlates of unconscious processing of invisible faces. I was unable to image significant activations related to the processing of completely invisible faces, despite existing reports in the literature. I thus moved on to the next question and studied how recognition of a familiar person was achieved in the brain; I focused on finding invariant representations of person identity – representations that would be activated any time we think of a familiar person, read their name, see their picture, hear them talk, etc. There again, I could not find significant evidence for such representations with fMRI, even in regions where they had previously been found with single unit recordings in human patients (the Jennifer Aniston neurons). Faced with these null outcomes, the scope of my investigations eventually turned back towards the technique that I had been using, fMRI, and the recently praised analytical tools that I had been trusting, Multivariate Pattern Analysis. After a mostly disappointing attempt at replicating a strong single unit finding of a categorical response to animals in the right human amygdala with fMRI, I put fMRI decoding to an ultimate test with a unique dataset acquired in the macaque monkey.

There I showed a dissociation between the ability of fMRI to pick up face viewpoint information and its inability to pick up face identity information, which I mostly traced back to the poor clustering of identity selective units. Though fMRI decoding is a powerful new analytical tool, it does not rid fMRI of its inherent limitations as a hemodynamics-based measure.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

> "Bulk tissue technologies such as fMRI reliably identify which brain regions relate to vision, imagery, pain, or memory, a rebirth of phrenological thinking. Brain imaging tracks the power consumption of a million neurons, irrespective of whether they are excitatory or inhibitory, project locally or globally, are pyramidal neurons or spiny stellate cells. Unable to resolve details at the all-important circuit level, they are inadequate to the task at hand."
>
> Christof Koch, Confessions of a Romantic Reductionist (2012).[1]

"Unable to resolve details […] inadequate […]"

Christof omitted sharing these concerns when I started working on a fMRI project with Nao Tsuchiya trying to uncover the relationship between attention and consciousness in the fall of 2006. Somehow, I stuck with fMRI after I joined Caltech as a PhD student; most likely it was a combination of fascination, optimism and stubbornness that led me on this path. At the onset of my PhD, I was still quite naïve about fMRI, and simply thought of it as a really cool technique to measure brain activity. I believe that this is how most cognitive scientists think of fMRI; trying to be rather oblivious of the underlying complexity and abstracting themselves from details and concentrating on drawing conclusions. This is the reason why I needed to write the first chapter of this thesis; the glimpse that I offer into the physical bases of the fMRI signal will most likely leave the reader experiencing a queasy feeling about the whole venture. But the reader should not write the technique off right away; Nikos Logothetis, one of the leading experts in fMRI, wrote the following in his superb 2008 review[2] published in Nature:

> "[…] fMRI is not and will never be a mind reader, as some of the proponents of decoding-based methods suggest, nor is it a worthless and non-informative 'neophrenology' that is condemned to fail, as has been occasionally argued."

For now, let us thus downplay Christof's assertion and agree with Nikos that fMRI is not a worthless technology. This is the bet I made a few years ago.

"[…] the task at hand."

What is the task at hand, anyway? The dry version of it is that I have been trying to study how conscious visual experience arises from brain activity. The more colorful version is: what happens in my brain, when I open my eyes in the morning, look at my wife's face, recognize her and am filled with bliss? I have been seeking clues to answer this simple question through the use of fMRI. In the next chapters, I report failures and successes as they happened, to shed an (almost) unbiased light on studying the conscious and unconscious visual experience with fMRI. I report, in turn, on my ventures: whether invisible pictures of faces are processed unconsciously (Chapter 2); how the brain recognizes familiar people (Chapter 3); and finally, on two somewhat more methodologically oriented projects aimed at replicating strong single neuron findings with fMRI, one about a categorical response to pictures of animals in the right human amygdala, the other about face viewpoint and identity information in the macaque's face patches (Chapter 4). I relied heavily on fMRI throughout this thesis, but I am well aware of the importance of combining different recording techniques to draw unbiased conclusions about brain function; I bring this up in the discussion at the end of this thesis. I have, in fact, worked on other research questions in the past six years, which are not reported in this thesis, through my collaboration with Rufin VanRullen; with Rufin, I have mostly made use of electroencephalography and Transcranial Magnetic stimulation[3–5] (and psychophysics[6,7]). Thus I have a fairly good understanding of the advantages and drawbacks of the major techniques available for non-invasive studies of the brain. In the work that I present here, I focused on fMRI for two reasons: 1) it was the most readily available technique for me here at Caltech, and 2) I truly wanted to understand its limitations. I gained many insights into the technique itself, and some sparser insights on the question(s) I was asking with it.

Besides being a narrative of intertwined scientific projects that unfolded in the past years, this thesis is also, and perhaps primarily, the story of my childhood as a scientist. I feel like I may finally be coming of age. Finally! I enjoyed the process though. I hope you will too.

I. WHAT YOU MUST KNOW ABOUT FUNCTIONAL MAGNETIC RESONANCE IMAGING (FMRI)

My relationship with fMRI has had its share of ups and downs in the last few years. Perhaps this section should be entitled: "Things you wish you had known before starting fMRI research", or just as well, "Things you should periodically be reminded of when conducting fMRI research." Either way, there is no excuse for ignoring this chapter. Cognitive neuroscientists cannot all be MRI physicists, but we should still have a pretty good idea of what we are measuring to be informed users.

I asked Ralph Lee, the manager of MR education and technologies at the Center for Brain Imaging at Caltech, for some numbers. According to him, in the U.S.A. there are about 8,000 magnets. Out of those there may be 500 or so that are dedicated research magnets (the rest are for clinical use, however research universities may purchase time after hours for research). Perhaps about 12 sites in the U.S.A. are dedicated to neuroscience only. The number of functional MRI studies published yearly has grown exponentially since the technique was developed in 1990 (**Figure 1**), and we are at present at the downright alarming rate of about five studies published per day[8]! The troubling thing is that our understanding of the human brain is far from progressing at the same rate. In 2004, Nikos Logothetis and Brian Wandell wrote[9]:

> In the short period of time since its introduction, fMRI has evolved to become the most important method for investigating human brain function.

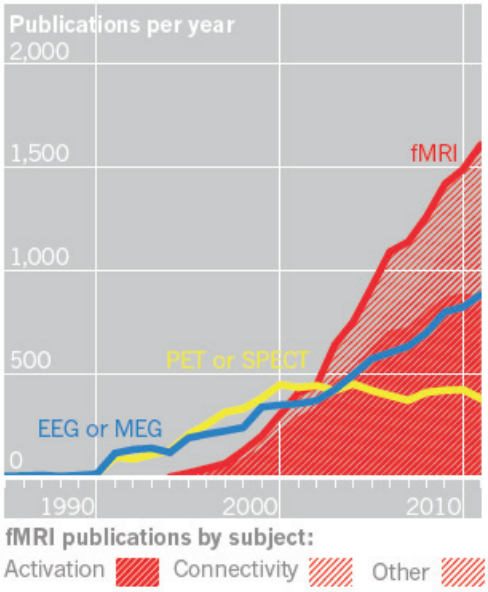


Figure 1 The rise of fMRI; data from ISI Web of Knowledge (fMRI = functional magnetic resonance imaging; PET=positron emission tomography; SPECT=single-photon emission computed tomography; EEG=electro-encephalography; MEG=magnetoencephalography). Reproduced from [8].

I succumbed to fashion, in some sense. This is why I want to make it very clear from the onset what fMRI is. You could probably gather much of the material that I go through in this chapter from various textbooks on the subject (I drew heavily from the excellent textbook by Huettel, Allen and Song[10] to compile this chapter, as well as a number of research and review papers [2,9,11–18]). This is a condensed tutorial, which summarizes what you should absolutely be aware of any time you read a fMRI study, and especially before you read the body of my work.

When scientists talk about their research, they blissfully omit many of the methodological details and technical challenges that they encountered in the process, and paint a pretty picture that far overreaches what can be concluded from the actual data – this is, of course, especially true when they talk to the public, but I find that it is often the case when they talk to their peers or write grant proposals. I am rather averse to this behavior, so here's the basic science, and you'll get plenty of reminders throughout this thesis of my obsession with honestly interpreting data – unfortunately, this may do my career some harm in the short term.

A. The Physics of BOLD fMRI

In the following, I provide a classical physics description of MRI, which is a simplified explanation of the phenomena at play; for a thorough understanding, one would need to resort to quantum mechanics, which is beyond the scope of this thesis.

1. MR signal generation

a. Spins

fMRI relies on hydrogen nuclei, ^{1}H. A hydrogen atom is composed of a proton (its nucleus) and an electron. The proton, under normal conditions, spins about itself. Because the proton is electrically charged, this rotation creates an electrical current, which in turn generates a torque (a turning force) when the proton is placed in a magnetic field: this is known as the magnetic moment. Because the proton has a mass, the rotation also gives rise to a non-zero angular momentum. These combined properties give the hydrogen nucleus the magnetic resonance property – I will refer to hydrogen nuclei as "spins" in the following. Note that other nuclei could potentially be used for magnetic resonance imaging (^{19}F, ^{31}P, ^{13}C, ^{23}Na, ^{17}O); all these nuclei have unpaired protons, hence spin. However, these are rare in the body/brain (rare isotopes), compared to ^{1}H which is omnipresent (body tissue contains 60-80% of water, H_2O…).

In a strong magnetic field, these spins will start a gyroscopic motion, called precession, about the axis of the magnetic field, at an angle determined by their angular momentum. There are two possible states for these precessing protons: parallel or antiparallel to the direction of the magnetic field. A spin can transition from the high-energy state (antiparallel) to the low-energy state (parallel), emitting a photon in the process; the energy of the photon is exactly the difference between the two quantized energy states. Conversely, a spin in the low energy state may absorb a photon and transition to the high-energy state. The frequency (ω_0) of the absorbed/emitted

electromagnetic energy ($E=h\,\omega_0$ where h is the Planck constant) can be shown to depend only on the magnetic field strength (B_0) and on the gyromagnetic ratio (γ, the ratio of the magnetic moment to the angular momentum, constant for a given nucleus): it is known as the Larmor frequency.

$$\omega_0 = \gamma B_0$$

For a 3T scanner, the Larmor frequency for hydrogen is approximately 127.74MHz (note that this is close to the FM and TV broadcast bands). Interestingly, the precession frequency (in rad/s) is also given by the Larmor frequency.

b. Net Magnetization

There are *many* (on the order of 7.5×10^{25}, just counting water molecules) hydrogen atoms in the average human brain. Under normal circumstances, these are oriented randomly, pointing in all directions, and thus do not lead to a substantial net magnetization. In the presence of a magnetic field, the spins will align, as described above; there will be a net magnetization (in the direction of the magnetic field), proportional to the difference in the number of spins in the parallel and antiparallel states. The higher the field strength, the higher the proportion of spins in the parallel, energetically more stable state (the energy difference between the two states increases linearly with field strength; this is known as the Zeeman effect).

c. Excitation and relaxation

The net magnetization, when tipped away from equilibrium, will start precessing around the main axis of the field at the Larmor frequency (as it relaxes to equilibrium). In a MRI scanner, a short excitation pulse at the Larmor frequency is applied by a transmitter coil to tip the net magnetization away from equilibrium.

The precession of the net magnetization vector can be picked up by a receiver coil; this is the MR signal. With time, the net magnetization vector relaxes back towards its equilibrium state via two mechanisms: longitudinal relaxation, whereby spins gradually returning to the low energy state lead to an increase in the longitudinal magnetization (back towards its equilibrium value) – the time constant of this process is called T1; and transverse relaxation, whereby the spins,

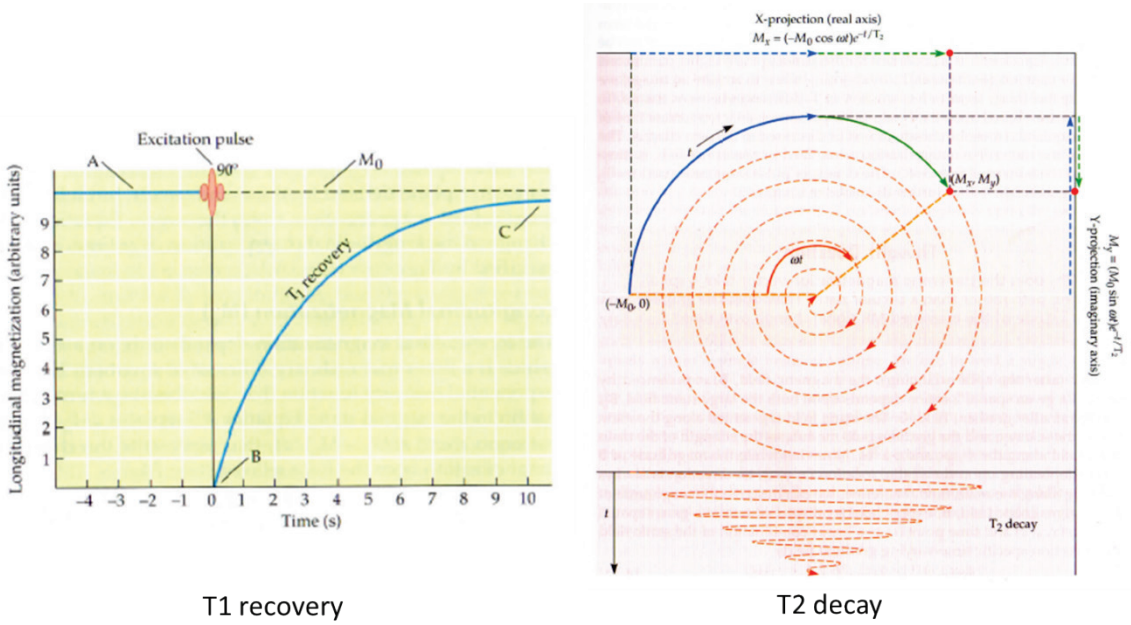


Figure 2 The changes in longitudinal (left) and transverse (right) magnetization over time, following an excitation pulse. Left: when fully recovered (A), the longitudinal magnetization is at its maximum value, as shown by the dotted line, and does not change over time. However, following an excitation pulse that tips the net magnetization into the transverse plane, there will be zero longitudinal magnetization (B). As time passes following excitation, the longitudinal magnetization recovers toward its maximum value (C). The time constant T1 governs this recovery process. Right: the magnetism in the transverse plane is a vector defined by its angle and magnitude. As time passes, its angle follows a circular motion with constant angular velocity ω, while its magnitude decays with time constant T2. These two components combine to form the inward spiral path shown (dashed lines). Shown at the top and right sides of the spiral path are its projections onto the x- and y- axes, respectively. Within each axis, the projection of the transverse magnetization is a one-dimensional oscillation, as illustrated by the blue and green lines. This oscillation is shown over time at the bottom of the figure, which illustrates the decaying MR signal. Reproduced from [10].

initially in phase as they precess, progressively get out of phase, leading to a decrease in the transverse magnetization (back towards zero) – the time constant of this process is called T2 (**Figure 2**). Transverse relaxation may sometimes be faster than predicted by the T2 time constant; local field inhomogeneities will cause spins to precess at slightly different frequencies,

hence getting out of phase more rapidly. The T2* time constant takes field inhomogeneity into account; naturally, T2* is always smaller than T2. These relaxation processes constrain how much MR signal can be picked up following a single excitation pulse.

2. MR image formation

We now know how to measure the net magnetization of a chunk of matter in a magnetic field. How do we generate a three-dimensional picture of the brain? This is achieved through the use of magnetic gradients (the magnetic field varies in space).

a. Slice selection

The application of a static gradient along the slice selection axis makes it so that spins along that axis have different Larmor frequencies. An excitation pulse centered at a given frequency will thus only affect the spins within a given slice of the volume. To excite a perfectly rectangular slice, a sinc-modulated electromagnetic pulse must be applied. Slice location and thickness are determined by, 1) the center frequency of the excitation pulse, 2) the bandwidth of the excitation pulse, and 3) the strength of the gradient field. Note that to increase resolution, strong gradients are required.

b. Spatial encoding

Once spins are excited within the selected slice, additional gradients are turned on for spatial encoding. In a typical gradient-echo sequence (which is what I used throughout this thesis), a first gradient is turned on in one dimension (before data acquisition), which results in the accumulation of a certain amount of phase offset; this is the phase-encoding gradient. A second gradient in the orthogonal direction is turned on during data acquisition, which changes the precession frequency of the spins; this is the frequency-encoding gradient.

c. Image formation

MR image formation relies on the formalism of k-space. k-space is the two-dimensional Fourier transform of a MR image; its complex values are sampled (at discrete points) during a MR measurement; it thus constitutes a temporary image space, in which data from digitized MR signals are stored during data acquisition. For instance, the previously described combination of a phase-encoding gradient before data acquisition and a frequency encoding gradient during data acquisition fills one line of k-space. When all lines of k-space are filled, the MR image can be retrieved with an inverse Fourier transform. An interesting consequence is that the field-of-view and resolution of the two-dimensional image of the slice depend, respectively, on the resolution and the field-of-view of two-dimensional k-space sampling.

3. MR contrasts and pulse sequences

a. T1- or T2-weighted?

There are two fundamental parameters in designing a pulse sequence: the repetition time (TR) is the time between two consecutive excitation pulses; the echo time (TE) is the time between the excitation pulse and data acquisition. The values of these two parameters dictate what type of image is collected; for instance, whether a T1-weighted image or a T2-weighted image.

b. Echo-Planar Imaging

Imaging the functioning brain requires being able to acquire images rapidly. Echo-Planar Imaging is a method in which the entire k-space is filled using rapid gradient switching following a single excitation pulse (in the classical gradient echo sequence described above, each line of k-space requires its own excitation pulse) (**Figure 3**). It is very taxing on the gradient hardware, and leads to several common artifacts such as signal loss (due to field inhomogeneities, e.g., at boundaries

between brain tissue and air-filled cavities) and geometric distortions (due to the long readout time for each excitation, small field variations in the image plane can cause distortions up to several pixels).

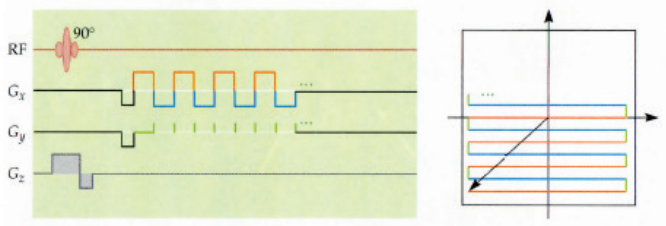


Figure 3 An echo planar imaging (EPI) pulse sequence (left) and the corresponding trajectory in k-space (right). Note that the directions of the gradients are changed rapidly over time to allow the back-and-forth trajectory through k-space. Reproduced from [10]

4. How can Blood Oxygen Level be picked up by MR?

Oxygen nuclei do not have the magnetic resonance property (there are no unpaired protons in the ^{12}O nucleus). Hence, measuring blood oxygen level directly with MR would not be possible. Hemoglobin, the molecule which carries oxygen in our blood, has a fortuitous property: oxygenated, it is diamagnetic (zero magnetic moment), while deoxygenated, it becomes paramagnetic. Fully deoxygenated blood has a magnetic susceptibility 20% higher than fully oxygenated blood. Deoxygenated blood will thus cause a local field inhomogeneity, leading to spin dephasing and faster transverse relaxation (**Figure 4**). In a T2* weighted image, a voxel's intensity will thus vary as a function of the oxygenation of the blood that flows through it.

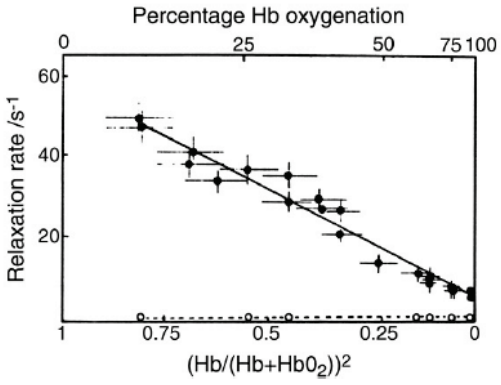


Figure 4 Effect of blood deoxygenation upon MR relaxation constants. Shown are the differential effects of blood deoxygenation upon transverse and longitudinal relaxation times, as expressed by the constants 1/T2 (filled circles) and 1/T1 (open circles). The x-axis indicated the square of the proportion of deoxygenated blood. Note that oxygenation increases from left to right. Clearly evident is the fact that 1/T2 decreases with increasing oxygenation; that is, the more deoxygenated hemoglobin that is present, the shorter the T2. Note that T1 is not affected by blood oxygenation level. Reproduced from [10,19].

B. The link between BOLD signal and neuronal activity

Information processing in the brain relies on communication between neurons. Each neuron can perform some basic operation on its inputs, and then propagates the output to other neurons; the main way (that we know of so far) for neurons to communicate over a distance is via action potentials. For instance, all of the information communicated from the retina to the rest of the nervous system is represented in the action potentials from ganglion cells. Consequently, the study of action potentials (via microelectrode recordings) has, for the past 50 years, been the gold standard for understanding computations performed in the brain. Is the Blood Oxygen Level Dependent (BOLD) signal related to the neural signal?

1. Increased activity → increased blood flow → deoxyHb flushed → increased signal

Neuronal activity requires energy in the form of Adenosine Tri-Phosphate (ATP). The brain does not store energy, hence ATP molecules must be synthesized on demand through the oxidation of glucose. Glucose and oxygen are supplied by a local increase in Cerebral Blood Flow (CBF).

This increase overcompensates the oxygen deficit resulting in an oversupply of oxygenated blood, and a corresponding increase in the BOLD signal (**Figure 5**).

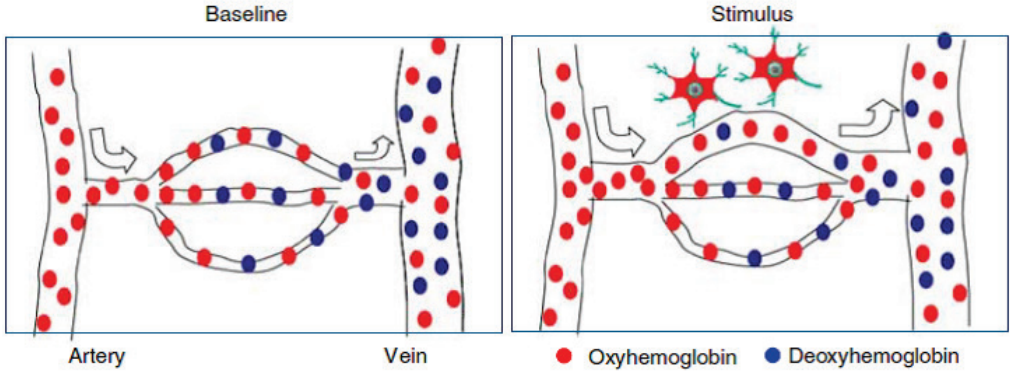


Figure 5. Vascular responses to neural activity. This schematic shows oxyhemoglobin (red dots) and deoxyhemoglobin (blue dots) in blood flowing through arteries, arterioles, capillaries, venules, and finally to veins. At prestimulus baseline conditions (left), blood oxygen saturation is ~100% in arteries, while it is ~60% in veins. Increases in neural activity (after stimulation, right) trigger an increase in blood velocity (indicated by the size of arrows) and dilation of vessels. The resulting increase in perfusion exceeds what is required by the increase in oxygen consumption rate. Reproduced from [18].

2. Devilish Detailed Mechanisms of Neurovascular coupling

a. Quick primer on brain vasculature

One often forgets how densely vascularized the brain is (that is, if you have never actually seen a brain). The brain consumes about 20% of blood oxygen (when the body is at rest); for an organ that represents only about 2-3% of body weight, this is quite spectacular. Two major arterial systems, the carotid arteries and the vertebral artery, supply oxygenated blood to the brain. These arteries (4-10mm in diameter) branch into smaller arteries, then into even smaller arterioles, and eventually into capillaries (5-10μm in diameter, corresponding to the width of a red blood cell). It is at the level of the capillaries that exchanges of oxygen, nutrients, and waste products occur. The distance between capillaries and neurons is likely less than 13μm (in the most densely vascularized areas of cortex), since the average intercapillary distance is about 25μm. Capillaries then coalesce into small venules, which collect into larger and larger veins, which drain into long venous channels (formed by the meningeal covering of the brain) called sinuses. The superior and

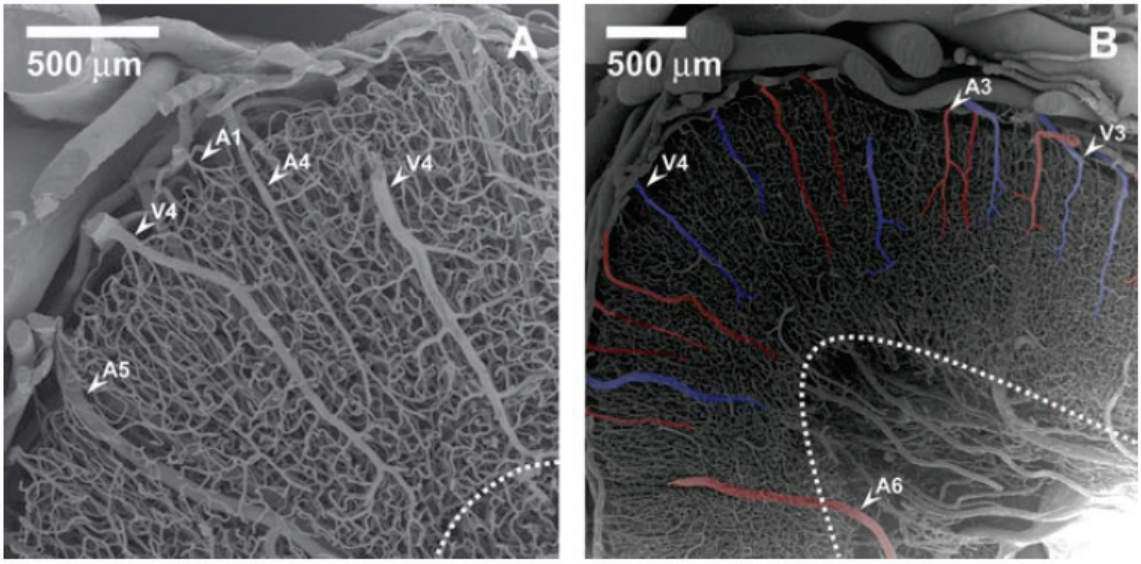


Figure 6 Scanning electron micrographs of a vascular corrosion cast from monkey visual cortex (superior temporal gyrus). Casts were cut and trimmed to allow a vertical view on the cortex. The gray-white matter demarcation line is shown as dashed line. Note the continuous orderly distribution of large vessels oriented perpendicularly to the cortical surface, their different length and branching patterns and the rather homogeneous mesh size and density of the capillary bed. (A=artery, B=vein). Reproduced from [13]

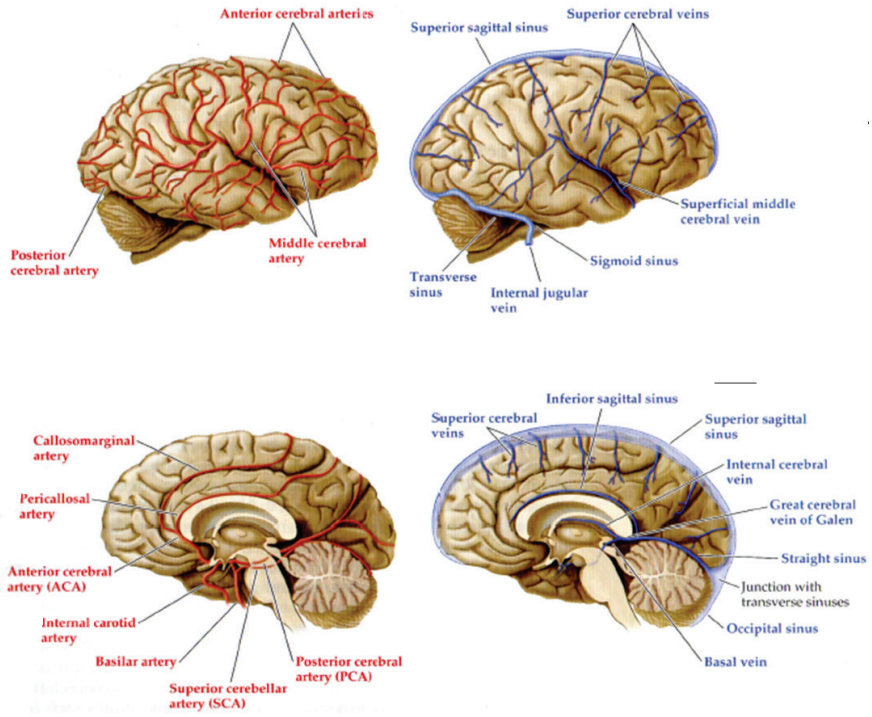


Figure 7 The arterial and venous organization of the cerebral vasculature. Shown are the lateral (top left) and medial (bottom left) views of the major arterial systems of the human brain. Blood is drained by a system of sinuses and veins, shown here in lateral (top right) and medial (bottom right) views. Reproduced from [10].

inferior sagittal sinuses drain into the transverse sinuses, which eventually form the jugular veins, exiting the skull and returning "dirty" blood to the heart (**Figure 7**).

b. Feedforward signaling (rather than metabolic feedback)

In the last few years, our understanding of neurovascular coupling has progressed rapidly. The traditional (intuitive) explanation of neurovascular coupling used to be that CBF was directly regulated by the increased metabolism and energy demand of neurons, through signals such as the fall in O_2 or glucose concentration, and the rise in CO_2 concentration (which dilates cerebral vessels; an increase in CO_2 leads to a decrease in pH, which is sensed by chemoreceptors, and vasodilation occurs to flush out excess CO_2). This metabolic feedback idea has been almost completely superseded. It is now believed that neurotransmitter-mediated signaling (mostly with glutamate, which is used in 90% of synapses in the brain) is the main factor in the regulation of CBF, and that a certain class of glial cells, the astrocytes, plays a major role. In short, glutamate mediated signaling leads to the release of vasoactive substances such as nitric oxide (NO) from neurons and arachidonic acid (+ derivatives) from astrocytes; these molecules can either have vasodilating or vasoconstricting effects, depending on local O_2 concentration. **Figure 8** represents a simplified scheme of our current, though still far from complete, understanding of glutamate-signaling mediated neurovascular coupling. Interestingly, another long-held belief was that blood flow was controlled solely at the level of arterioles, through the tone of smooth muscle surrounding those. As was recently discovered, the diameter of capillaries may be controlled by contractile cells called pericytes, which are present at roughly 50µm intervals along capillaries.

c. Oxygen oversupply: still not properly understood

As stated previously, the increase in blood flow leads to an oversupply of oxygen, which is the basis for BOLD fMRI. Many explanations were put forward to explain this mismatch. Among

these figured the rather complex "balloon model." It rests principally on the observation that the delivery system for oxygen (passive diffusion from capillaries to cells) is inefficient: as the velocity of blood inside capillaries increases, its transit time decreases, and the rate of oxygen delivery decreases nonlinearly. A disproportionate increase in blood flow becomes necessary to supply enough oxygen to cater to the increased demand. Unfortunately, experimental evidence against this hypothesis has since been reported.

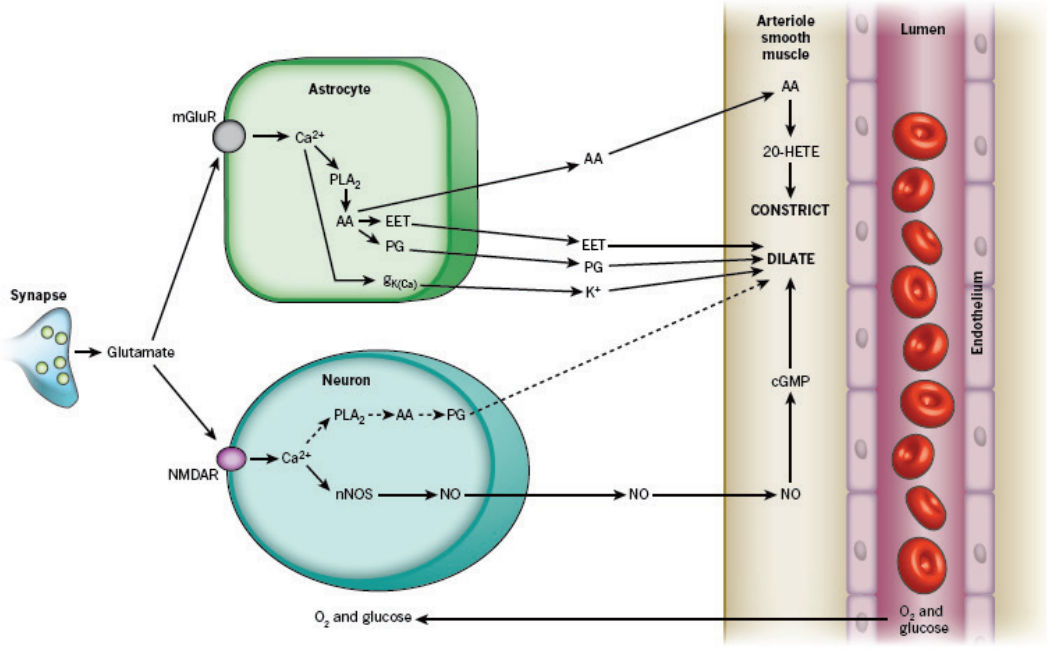


Figure 8 Major pathways by which glutamate regulates cerebral blood flow. Pathways from astrocytes and neurons (left) that regulate blood flow by sending messengers (arrows) to influence the smooth muscle around the arterioles that supply oxygen and glucose to the cells (right, shown as the vessel lumen surrounded by endothelial cells and smooth muscle). Reproduced from [16].

In the complex picture that is emerging, it is becoming clear that the relative importance of the different neurovascular coupling pathways (through neurons or astrocytes) is bound to differ between brain areas and between different neural networks in the same brain area. Also, as stated previously, differences in vascular density between different brain regions will lead to different coupling efficiencies. A study by Logothetis's team[13] demonstrated that the microvascular density

of primary visual cortex is higher than that of other visual areas; they suggest this feature may influence the signal-to-noise ratio of the hemodynamic signals and consequently increase the chances of detecting differences between conditions. This means that the relationship of the BOLD signal to underlying neural activity will be rather heterogeneous throughout the brain, a fortiori precluding direct comparisons between brain regions.

3. In practice: BOLD often correlates with LFPs

A series of studies conducted in sensory cortices of mammals supports a strong correlation between the Local Field Potential (LFP) and the BOLD signal. While the exact composition of the LFP (obtained by low-pass filtering extracellular recordings) itself is still a matter of investigation, the current understanding is that it reflects the excitatory/inhibitory postsynaptic potentials, together with dendritic hyperpolarization and intrinsic membrane oscillations; it is usually thought of as a measure of the input to neurons that are around the tip of the electrode, in the context of a typical extracellular recording. Since the input to neurons is bound to affect their output, it is generally assumed that neural activity indirectly drives the BOLD signal.

a. LFPs correlates with BOLD in sensory cortices…

In 2001, Logothetis and colleagues[20] recorded simultaneously the BOLD signal, the LFP (40-130Hz) and Multi-Unit spiking Activity (MUA) in the primary visual cortex of anesthetized monkeys viewing contrast gratings. They found a strong correlation between LFPs and BOLD, and a slightly weaker correlation (but still a robust one) between MUA and BOLD. It was also shown later that abolishing spiking in the visual cortex through pharmacological manipulation[14] still led to a robust correlation between LFPs and BOLD signal. The important take-home from all these studies (well summarized in [21]) is that the strong correlation that may sometimes be found between spiking activity and BOLD signal is due to a correlation between LFPs and BOLD

signal on the one hand, and LFPs and spiking on the other hand, rather than to a direct link between spiking and BOLD signal: BOLD and spike rate correlations cannot typically be assumed. Whether BOLD will reflect spike rate thus depends on whether LFPs and spike rate correlate.

b. … but not necessarily in other areas

As reviewed in [21], many studies have also failed to find a correlation between LFPs and BOLD; one region where it seems particularly difficult is the hippocampus. Ekstrom invokes local circuitry and local vasculature as possible causes for the absence of correlation. Whatever the reason, this is particularly disturbing for a project in which I invested much effort, that of finding fMRI evidence for "Jennifer Aniston neurons", which I describe in this thesis (page 88).

4. Spatial and temporal resolution of the BOLD signal

In pretty much any introductory course to neuroscience, you'll find a plot akin to that in **Figure 9**. It compares most of the techniques used in neuroscience, in terms of their spatial and temporal resolution. You see that fMRI has a decent spatial resolution compared to other non-invasive techniques such as EEG/MEG; however the temporal resolution is several orders of magnitude worse than for these techniques. In the following I quickly discuss what factors influence the spatial and temporal resolution of BOLD fMRI.

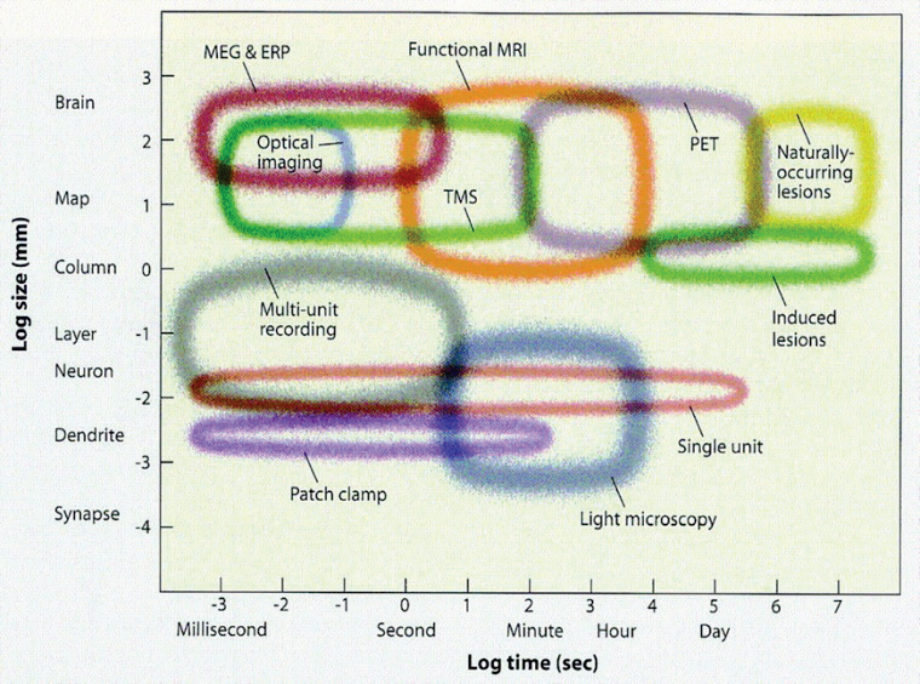


Figure 9 Spatial and temporal resolution of various methods for studying brain function (MEG=magnetoencephalography; ERP=event-related potential; fMRI=functional magnetic resonance imaging; PET=positive emission tomography; TMS=transcranial magnetic stimulation). Reproduced from [22].

a. Spatial resolution of BOLD fMRI

The main limiting factor in terms of spatial resolution for fMRI is that it measures a vascular response. For example, while the response should, in principle, be colocalized to the capillary beds, larger draining vessels may contribute to the signal; if this is the case, a BOLD contrast may be detected downstream of the active area, resulting in mislocalization and overestimation of the extent of activation. With the typical voxel size in most cognitive neuroscience studies this is not a serious problem, but it poses a limit to the resolution one can hope to achieve. Interestingly, a higher field strength leads to a lesser contribution of larger vessels[23]. Also, while the most common pulse sequence in cognitive neuroscience studies is Gradient-Echo Echo Planar Imaging, which is sensitive to all vessel sizes, a Spin-Echo sequence is less sensitive to larger vessels. A recent quantitative study of the vasculature of monkey primary visual cortex[13] concluded that the

ultimate spatial resolution of an imaging scheme based on the penetrating venous vessels would be around 0.70 mm^3; they also noted the possibility of a better resolution by imaging signals originating from arteries (feeding volume 0.44 mm3) or, better, from capillaries.

What is the resolution one can achieve? In 2001, researchers were able to resolve ocular dominance columns with fMRI at 4T[24] – these have a mean width of 1mm, and the fMRI voxels had an in-plane resolution of roughly 0.47x0.47mm^2. This is already impressive.

b. Temporal resolution

Vascular factors are also limiting the temporal resolution of BOLD fMRI; the vascular response is sluggish, and thus will always, at best, reflect a low pass filtered version of the underlying activity. Technically, it is also time consuming to sample a whole fMRI volume; with single-shot gradient-echo EPI, one slice can be acquired in roughly 40-50ms. One can imagine that technology will improve somewhat in the next years; for instance parallel imaging, which uses multiple coils to sample the image, already allows accelerated acquisition.

Despite its limitations in terms of spatial and temporal resolution, two major advantages of fMRI should be emphasized here: it can be done safely in live humans, and it allows imaging of the whole brain.

C. From ugly functional MR volumes to pretty colorful blobs

There now exist a couple of very good recent textbooks on fMRI data analysis [25,26], which you should read VERY carefully if you are thinking of toying with your own fMRI data (they were not available when I started). Here I describe briefly the steps that I typically use in my analyses to give you a flavor of how much processing goes into producing final results. My methods have evolved over the last six years, with an increased use of multiple software packages to perform

various operations. But in the end, these are pretty small variations around a main series of steps. I drew a schematic overview of the steps that I usually use in **Figure 10**.

At the end of a typical scanning session, you find yourself with about 1500-2000 dicom files, each file corresponding to one volume of fMRI data. A fMRI session is usually subdivided into many runs; between runs, the experimenter stops the scanner to give the subject (and the gradients) a few minutes' rest. Besides the functional runs, the experimenter may acquire a high resolution anatomical image (T1-weighted), as well as a field map–this is a special sequence which allows measurement of the magnetic field inhomogeneities.

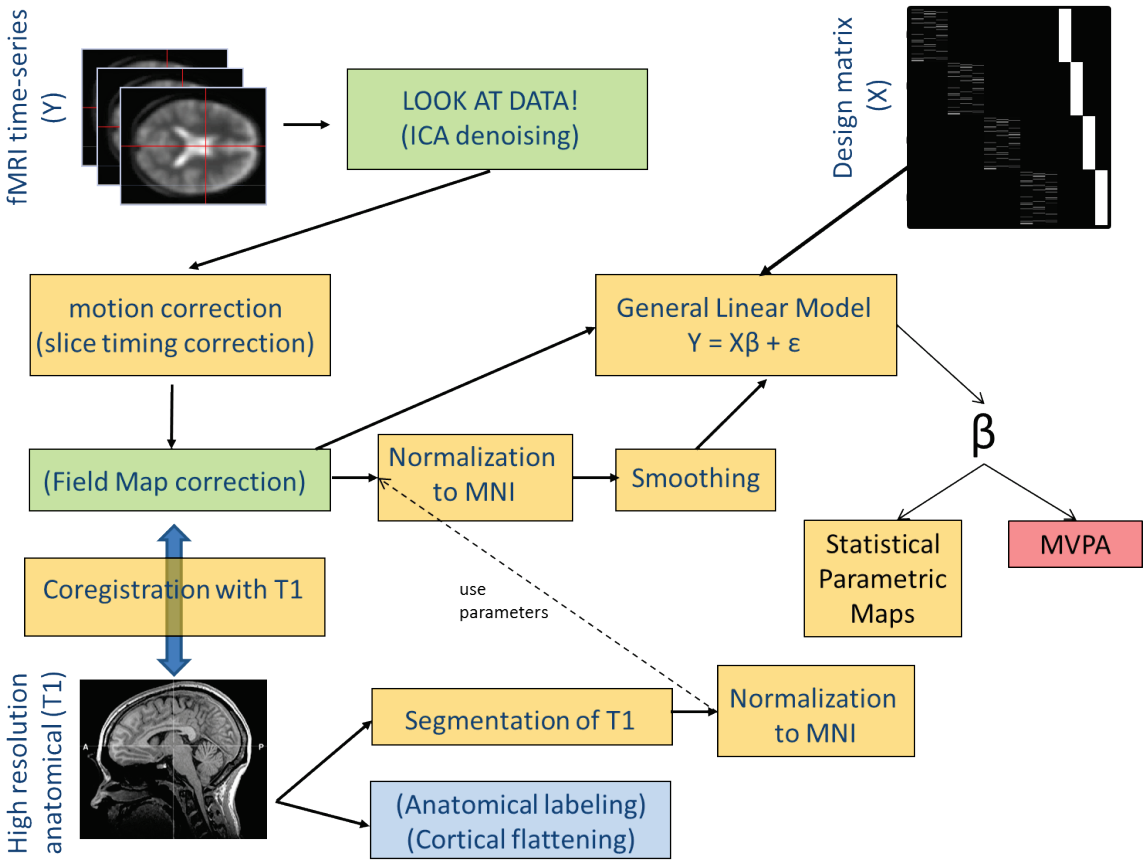


Figure 10 A rough schematic of my fMRI data processing pipeline. Colors represent the software suite that I usually use for each step (orange, SPM; green, FSL; blue, FREESURFER; red: custom software). If a step is between brackets, it means I do not always implement it. See text for more details on each step.

1. Preprocessing

a. Inspecting raw data, denoising with Independent Components Analysis (ICA)

The first and foremost preprocessing step is to look at the raw data. The importance of such simple quality assurance cannot be overemphasized. When you learn fMRI analysis, you can quickly get caught up into learning how to use fMRI preprocessing software, turning your brain off and treating it as a black box. Do not. Watch a movie of your raw data after you import it, and notice any strange artifacts. There is a SPM toolbox called "ArtRepair" which can be useful for viewing data and artifact rejection.

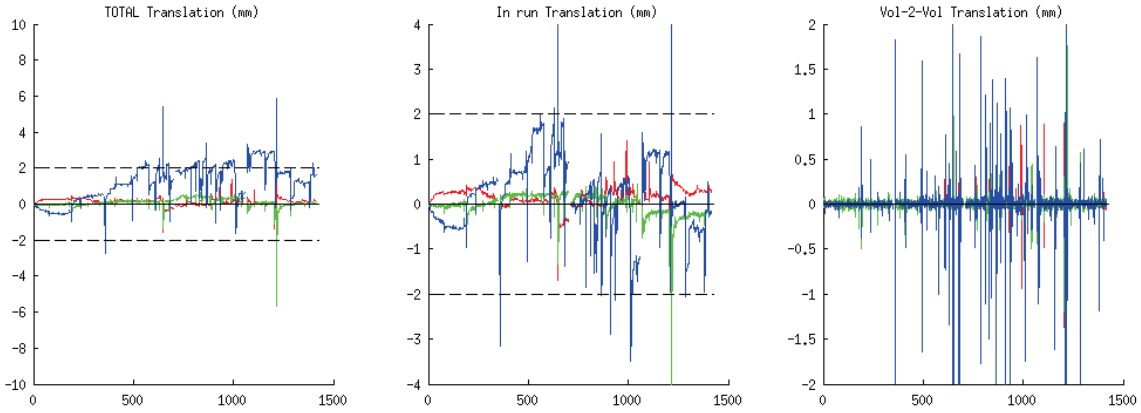
A step that can also be used at this stage is Independent Component Analysis (ICA) exploration. ICA is a method for separating a multivariate signal into additive components, using the assumption of mutual statistical independence for the underlying non-Gaussian source signals (blind source separation). There is noise in functional MRI data, such as physiological noise (breathing, cardiac rhythm) or artifacts due to the subject's motion (ring of activation; stripes due to spin history effects), that can be quite nicely separated from the rest of the signal through ICA. The FSL software package has a function called *melodic* which implements ICA and allows the inspection of individual components, then the removal of identified "noise" components from the time series. Identifying noise components is not simple; without good guidelines, the exercise is quite arbitrary. A recent paper attempted to establish those guidelines[27], and I have tried to follow it in my various attempts at using ICA denoising. The truth is, after spending much time including it in my pipeline, I have yet to find an improvement in my results using this additional preprocessing step.

b. Correcting for motion

Since it is difficult to get the IRB to approve the use of curare or other neuromuscular blocking drugs in the context of a research fMRI experiment (just kidding!), the subject lying in the scanner will inevitably move their head a little–in between runs, for sure, when the subject tries to get comfortable, swallows, breathes deeply, etc., but also during runs. Motion can cause a lot of problems in a fMRI dataset. Of course, it is rather easy to realign the volumes. This is done in SPM (the software package that I started with and stuck with, because once you develop a set of scripts that works more or less flawlessly, you just do not want to re-invest your time into another software package) through the *realign* function, or in FSL (the leading alternative to SPM, considered better by its dedicated users; so much so that some researchers who have been SPM users for years are considering making the switch) through the *mcflirt* utility. These functions assume a rigid body transformation, with six degrees of freedom (translation in x, y and z + rotation around x, y and z); the volumes in a time series are aligned to a reference volume (usually, a volume in the middle of the time series), using a simple cost function (least squares in the case of SPM, normalized correlation ratio in the case of FSL); then, they are resliced to match the reference volume voxel per voxel. This requires interpolation, which can be linear (fast, but leads to a lot of smoothing) or higher-order (such as sinc, spline, etc.).

While realigning the volumes with a rigid body transform may seem like it fixed everything since the volumes look so nicely aligned after this step, it did not. Why? One problem is that the protons that move into a voxel from a neighboring slice will have an excitation that does not match what the scanner expects. Another problem will arise in areas of susceptibility artifacts. Signal dropout and distortions in such regions depend on the angling of the slices with respect to the brain. A small amount of rotation can affect the shape of the brain in the reconstructed volume.

A.



B.

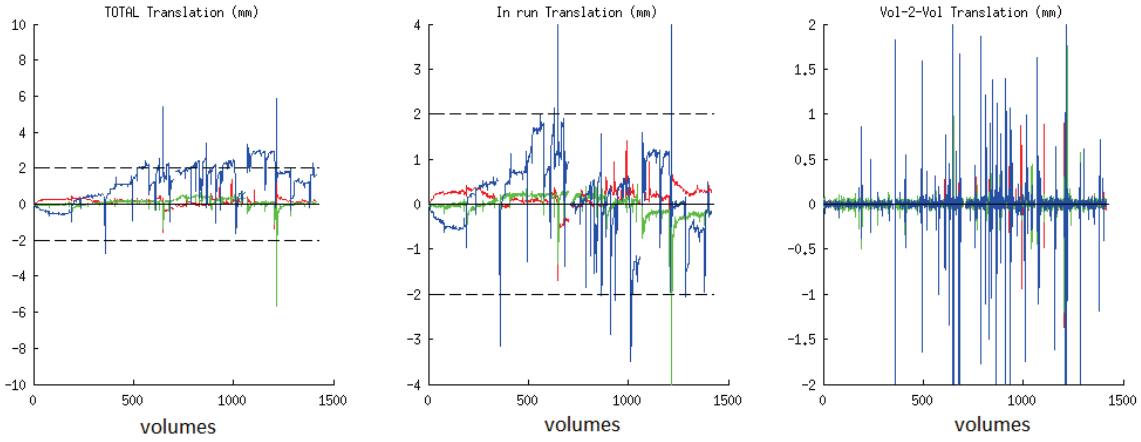


Figure 11 Motion parameters (translation component only, not rotation; red, x-axis i.e., moving head to the left or right;green, y-axis i.e., moving head back or forth; blue, z-axis i.e., moving head up or down) of two fMRI subjects, plotted over four runs of fMRI data acquisition. Dotted lines correspond to the voxel size. The first column represents the total translation, with reference to the first volume of the first run. The middle column shows translation referenced to the first volume of each run. The right column shows the translation between consecutive volumes. While all three plots are informative, the most problematic motion (most difficult to correct for) is the fast motion between consecutive volumes, best seen in the rightmost plot. A) a very jittery subject, B) a very still subject.

So the best practice is to pack your subject's head very tightly (but comfortably; I know, this is paradoxical) to prevent them from moving. In some imaging centers they use a bite bar; however, I was told that this may not be the best solution, since it may make the subjects swallow quite often, which will lead to some motion. One may think of other solutions, such as a bag that you

would fill with self-expanding foam, which would mold around the subject's head (I am serious, we discussed this with Mike at some point). In practice, at Caltech, I have been using some little foam pads, placed where they bother the subjects least. In **Figure 11**, you can see the motion parameters for two subjects, one really good one and one really bad one.

c. Slice timing correction

A fMRI volume is constructed one slice at a time; consequently, the exact acquisition time for each slice is different from the others. Most often, slices are acquired in an interleaved order. This means that slice 1 and slice 2 will be acquired about half a TR apart, i.e., in my case, about 1s apart (I use a TR of 2s in most of my experiments). Temporal interpolation can be used to correct for the different acquisition times, to the end of simulating that each volume was acquired instantly at a given time point. This is a built-in functionality of most fMRI processing software. While I used to apply slice timing correction religiously, after motion correction, I lately stopped doing so. The reason is that for a TR of 2s or less, it really does not make too much of a difference, and may instead introduce artifacts (due in particular to the interaction with motion correction). So I will not go into too much detail about it here.

d. Unwarping

Many researchers do not bother with this step; the fact is, it is not perfect, and may itself introduce additional noise in the data. However, when dealing with the temporal lobe, it becomes quite important to perform this step, and I now routinely use it in my analyses. The problem that one needs to correct for is that field inhomogeneities lead to voxel displacement in the reconstructed image (this was mentioned briefly in the first section of this chapter). A method exists to correct for these displacements, using a field map acquired during the scanning session. The details are beyond the scope of this thesis. While there exists a toolbox for SPM to perform

unwarping using a field map, I have found the FSL routines easier to implement. Initially, I was using *fugue*; but the latest version of FSL (v5) has a script called *epi_reg* which registers functional data to anatomical data, using the field map and a nice boundary based registration algorithm. This is what I use now.

e. Warping individual brains into a common space

In most studies, the researcher collects a number of subjects and averages the results from all subjects in order to make inferences about the population. In the case of whole brain analyses (when you look for activations throughout the imaged volume, rather than in specific areas), this is achieved by warping the brains of individual subjects to match a template brain. The original template brain is known as the Talairach brain, an atlas based on the dissection of a single brain[28]. Nowadays the template most commonly used is the MNI template (Montreal Neurological Institute), more especially the ICBM152 (International Consortium for Brain Mapping) based, as its name suggests, on an average of 152 brains, and roughly matched to the Talairach brain (though the coordinate correspondence is not straightforward). The process of warping individual brains to match the MNI template is called spatial normalization. Due to the smoothness of the template, the end result (the match between individual subjects' brains after normalization) is not perfect. Usually, researchers apply a fair amount of smoothing to their data when averaging across subjects (a Gaussian filter with full-width at half maximum of 8mm is common), which helps compensate the approximate anatomical match. Note that functional activations need not match exactly anatomically in different subjects, either. However, there exist other tools to further warp your subjects' brains and improve the final match. I have been working with a toolbox called ANTs (Advanced Normalization Tools) to this end, which uses diffeomorphic

transformations and allows you to create a template from your subject population* (**Figure 12**). Though, once again, it is unclear that this really improves the end result in terms of functional activations.

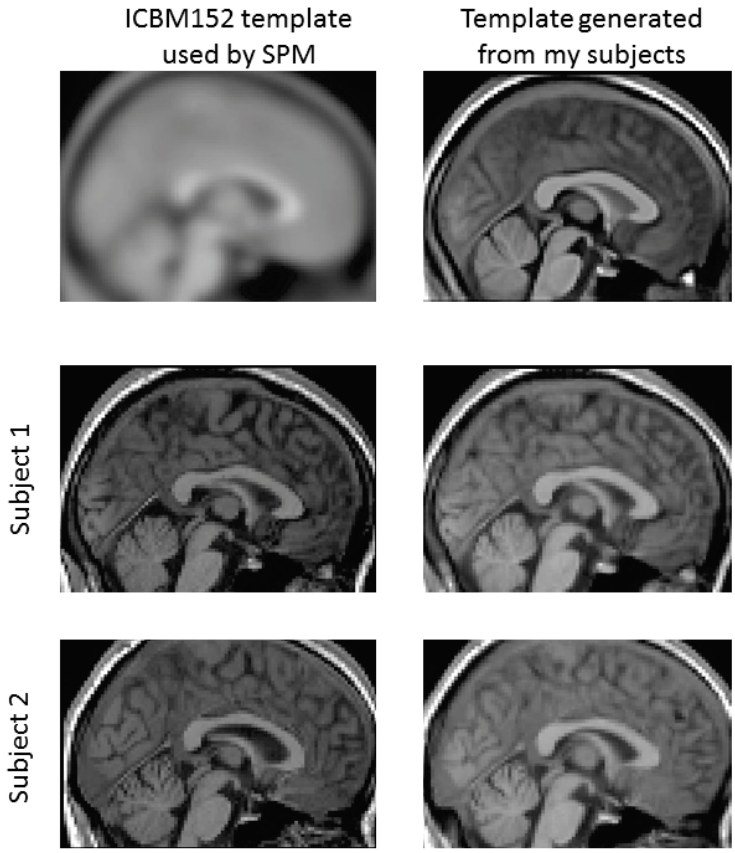


Figure 12 Comparison of two spatial normalization procedures. The left column shows the MNI template (top), and the result of spatial normalization of two representative subjects (middle and bottom). The right column shows a template generated from a population of 20 subjects with ANTs (top), and the result of matching the same two subjects to that template (middle and bottom). Note the increased accuracy in anatomical matching; for instance, the size of the ventricle under the corpus callosum.

Ideally, one should match individual brains on a functional rather than anatomical basis; indeed, it is unclear if specific functions will be subserved by the exact corresponding anatomical regions in different individuals. There have been recent developments in this field; researchers have tried to

* I just realized (as I am writing this and doing a bit of research) that there is a SPM toolbox called DARTEL that seems to do something similar… I had heard about DARTEL, but always thought that was what I was using when running SPM's normalize with segmentation. Oh well.

use the spatiotemporal pattern of responses to a given movie to match subjects' brains [29]. I have not tried to implement such a method in my own studies yet.

2. Canonical hemodynamic response function and general linear model

What does the BOLD signal look like, when neurons have a burst of activity (linked to the presentation of some stimulus)? The hemodynamic response is often described as sluggish: it does not increase instantaneously with neuronal activity nor returns to baseline immediately after the neurons do.

a. What does the Hemodynamic Response Function look like?

An abstraction that is used in most analyses of fMRI data is the ideal, noiseless response to an infinitesimally short stimulus, a typical trick in linear systems analysis. Though there have been

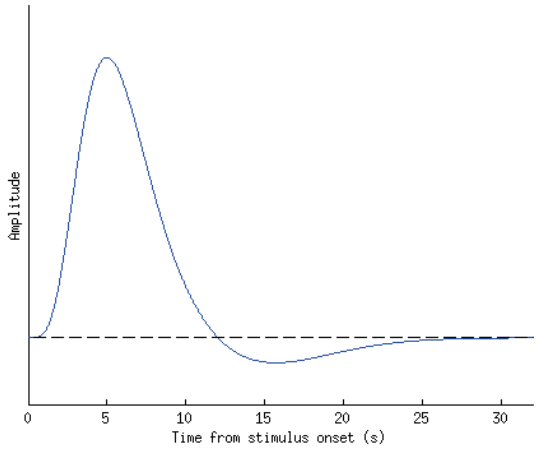


Figure 13 The canonical hemodynamic response function (HRF) in the SPM package. This is the hypothesized shape of the impulse response function of the system, i.e. the BOLD response to a neural event of infinitesimal duration. The predicted brain response to a given stimulus is obtained by convolving the stimulus's time course (usually, a box function) by this HRF, as is done in linear systems analysis. Generally, this canonical HRF is a decent fit to the true HRF for many normal subjects in many cortical and subcortical regions. Note the peak of the response around four to six seconds after stimulus onset, and the return to baseline over the next twenty seconds. Note the undershoot around fifteen seconds after stimulus onset.

studies showing that the shape of the hemodynamic response can vary quite wildly from one brain region to another in a single individual, and from one individual to another, still the majority of fMRI studies rely on a canonical HRF, applied blindly throughout the brain. In my case, I mostly make use of SPM for those analyses, and in **Figure 13** I plotted the default HRF used in that software package; it is a difference of gamma functions, with a peak response at six seconds and an undershoot at fifteen seconds from stimulus onset. These parameters were, I believe, chosen according to a study of the HRF in the primary visual cortex[30].

There are alternatives to using the canonical HRF as the expected shape of the BOLD signal when modeling your data. The canonical HRF has the highest bias (responses that conform to the canonical HRF are favored) but also yields the lowest variance in the estimates. A popular alternative approach is to complicate the model just slightly by including a temporal and a dispersion derivative of the canonical HRF as additional regressors in the General Linear Modeling (GLM) framework, which can account for some variations in the peak and duration of the HRF. If you are into reducing bias further (but increasing your variance), you can use a constrained basis set, a set of functions which captures some of the known aspects of the shape of the HRF whilst allowing some freedom to fit the data better. The final step in getting rid of all bias is to use a Finite Impulse Response (FIR) model, whereby each regressor models a time point in a specified peristimulus time window. I have looked at and played with these various ways to model the shape of the response, to some extent. In the end, I most often reverted to the canonical HRF. Though it may not always provide the best fit to the signal, it is also less likely to fit the noise… and there is a lot of noise in fMRI data!

b. Block and slow event-related designs

There are some experimental designs for which you do not really need to assume a shape for the hemodynamic response; if trials are spaced far enough apart, even the sluggish BOLD response

has time to go back to baseline, and you can use a simple ERP-like approach to analyzing the data – i.e., the typical analysis for EEG data, substract baseline and average trials. In this thesis there is some data that was analyzed that way, mostly the data acquired on macaque monkeys. Monkeys do not get sleepy or bored quite as easily as humans, and are highly motivated to keep fixating even when there is nothing on the screen so that they can get some juice and quench their thirst… but try this on humans, showing them an alternation of twenty-four seconds long blocks of images and twenty-four seconds long blank screens. They will either fall asleep or start thinking about all sorts of other things. This is why you usually have to speed things up a bit, which means that the BOLD responses of successive trials will overlap, and you have to get a little fancier in your extraction of responses to single trials.

c. Fast event-related designs and the General Linear Model

If trials in an experiment follow each other is quick succession, the hemodynamic responses to each of the two trials will roughly add up linearly. The linearity of the BOLD signal has been studied carefully by a few groups; if trials are more than two seconds apart[31], it is a rather good approximation. Convolving the canonical HRF with a boxcar function to model stimulus duration is also a pretty good approximation (if the duration of stimulation stays within reasonable limits, up to about six seconds, beyond which there is a clear deviation from linearity[32]).

The analysis of such designs usually relies on the General Linear Model. The researcher produces a model of what she thinks the BOLD signal would look like in brain regions that respond to her task first with boxcars that represent different events (onsets & durations), and then by convolving those boxcars with the canonical HRF (or another basis set). The result is known as a design matrix (**Figure 14**): it has as many rows as there are time points (i.e., fMRI volumes), and as many columns as there are conditions which the researcher deems will affect brain activity.

Some columns are usually included to denoise the signal (e.g., remove motion artifacts, physiological artifacts, low frequency drift, etc.) and are referred to as regressors of no interest.

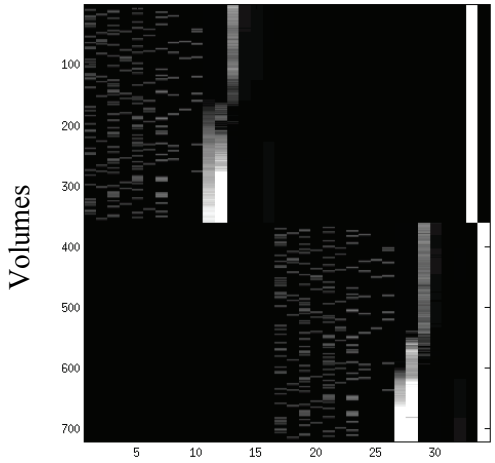


Figure 14 Example of a design matrix (X) for two fMRI runs of a fast event-related design experiment. Rows represent time (in units of fMRI volumes, i.e. with a resolution equal to the TR, the repetition time). Columns correspond to explanatory variables (a.k.a. regressors). Each run is modeled separately. The first run corresponds to volumes 1-360, the second run to volumes 361-720. Regressors 1-16 and 33 are used to model the first run, and 17-32 and 34 to model the second run. Note the motion regressors (11-16 and 27-32) and the constant regressor (33 and 34) for each run.

Once the design matrix (X) has been built, the researcher feeds it and her data (Y) into a multiple linear regression, which outputs parameter estimates (β) which correspond to the response amplitude for each condition in the model and an error term (ε):

$$Y = X\beta + \varepsilon$$

This is done for each voxel independently (this is often referred to as a mass-univariate approach). There are additional complications to running this multiple linear regression; one of them is that the BOLD signal is correlated in time, which violates the assumptions of linear regression. There are ways to deal with this, which are implemented in the software packages. But I will not go into any more details here as my main goal is to familiarize the reader with the main steps involved in analyzing fMRI data; this is not a fMRI analysis textbook!

3. Mass-univariate vs. multivariate statistics

a. Univariate analysis and the Statistical Parametric Map

Once the general linear model has been run, you can look at where in the brain condition A has a parameter estimate that is significantly different from zero…or, where condition A has a significantly larger parameter estimate than condition B…or…well, you can do a number of such things, relying mostly on T-tests and ANOVAs. And that is pretty much how you get a nice activation map which you can paint onto a coregistered anatomical volume.

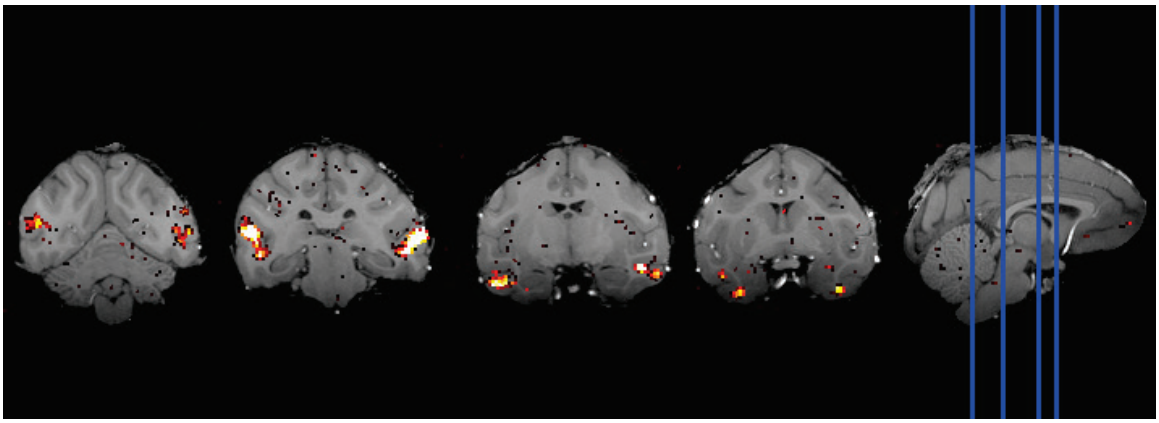


Figure 15 Example of a Statistical Parametric Map for the contrast (monkey and human faces) vs. (fruits, bodies, hands and objects) in a macaque monkey. Ten runs of a fMRI block design experiment were acquired in one session, featuring blocks of (familiar and unfamiliar) human faces, (familiar and unfamiliar) monkey faces, fruits, hands, bodies and objects. Hot spots correspond to areas that are significantly more activated by faces than by non-face categories. The statistical map was thresholded at $p<0.0001$. Four coronal sections are shown, corresponding to the blue lines drawn on the sagittal section to the right. From left to right, the hotspots correspond to face patches PL, ML/MF, AL/AF and AM.

b. Multivariate analysis

A rather recent development in the fMRI literature is the study of patterns of activation across multiple voxels, rather than just looking at the activation of single voxels or the mean activation of regions of interest. Since my thesis makes use of this method quite heavily (again, I was a fashion victim!), I describe the process in this section. I highly encourage the reader to refer to the

(too) many good papers describing the method and its tricks[33–42], should she want to conduct such analyses herself.

It makes good sense to be looking at patterns of activity, rather than averaging activity over a brain region and losing potentially meaningful information in the process. The usefulness of such a method was only convincingly demonstrated with the successful readout of orientation information from V1 (the earliest cortical visual area)[43,44], which predated the onset of my PhD adventure by about two years. The understanding back then was that we had found a silver bullet for the analysis of fMRI data, which could pick up underlying neural representations which should have been too fine to resolve given the size of fMRI voxels (**Figure 16**); this was the essence of the (in)famous hyperacuity interpretation of fMRI decoding. This interpretation made sense with what we knew of V1, with the orientation columns and all. Much debate surrounded this interpretation[45–51], fueled by the finding that smoothing prior to decoding barely affected classification performance, and recent fMRI evidence that a topographic map of orientation preference was present in human V1 at a much coarser scale (corresponding to the angular preference map that is used in retinotopic mapping studies[52] to find the borders of the different early visual areas).

Despite these challenges, fMRI decoding may still partly rely on hyperacuity[48,53,54]. It remains the most seductive interpretation of fMRI decoding, and as such, is deeply engrained in fMRI researchers' minds. Note that even if fMRI decoding does not rely on representations at spatial scales finer than that of fMRI voxels, it remains a very powerful technique for combining information from multiple loci in the brain, and is thus worth using.

The most common form of multivariate analysis performed in fMRI research is often referred to as fMRI decoding; which, in machine learning terms, can be described as pattern classification with supervised learning algorithms. How does it work?

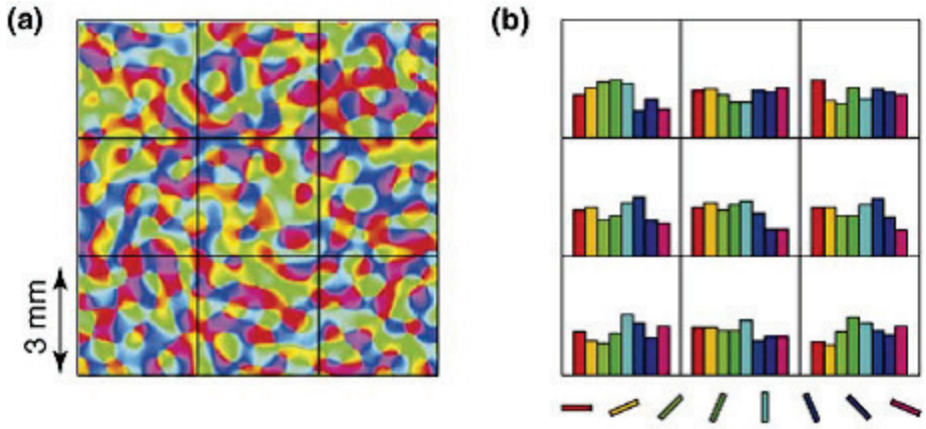


Figure 16 The hyperacuity interpretation of orientation decoding in V1. Part (a) shows a simulated orientation tuning map for a patch of visual cortex (different colors indicate different orientations), with a voxel-sized (3 mm) grid superimposed on the map. Part (b) shows the distribution of orientation selectivity values for each of the nine "voxels" shown in Part (a). Although all of the orientations are represented inside each voxel, the distribution of selectivity values is slightly different for each voxel. According to the hyperacuity interpretation, the classifier is able to exploit these small per-voxel irregularities in selectivity to decode orientation from multi-voxel patterns[43]. Reproduced from [35]

First, let me describe what a classifier is and what it does. There are many possible choices for a classifier, and I will not describe all of them thoroughly here. There are linear and non-linear classifiers; fMRI decoding usually uses the former, which seem to work just as well as non-linear classifiers on most datasets, are less prone to overfitting noisy datasets, and easier to interpret the results. Among the most popular linear classifiers are correlation-based classifiers (as introduced by Haxby[55] back in 2001), nearest-neighbor classifiers, Gaussian-naïve Bayes classifiers, Linear Discriminant Analysis (LDA), and linear Support Vector Machines (SVM). There have been a few attempts at comparing all these classifiers empirically on real data (e.g., [56]). My impression is that the former and the latter are used most often; the former for its extreme simplicity, the latter for its performance. I have almost always chosen to use linear SVMs in my ventures.

I am not going to go through the mathematics of linear Support Vector Machines here, there are many textbooks that you can refer to if you are interested (e.g., [57]). What SVMs do is to find a hyperplane, in the multidimensional space spanned by the data (each voxel, in the case of fMRI, is a dimension of that space), that maximizes the separation between examples from different

classes, while simultaneously keeping the errors low (the interplay between maximum margin and minimum error is controlled by a single parameter, C – the cost parameter). **Figure 17** shows an example in a two-dimensional space (two voxels), with two classes.

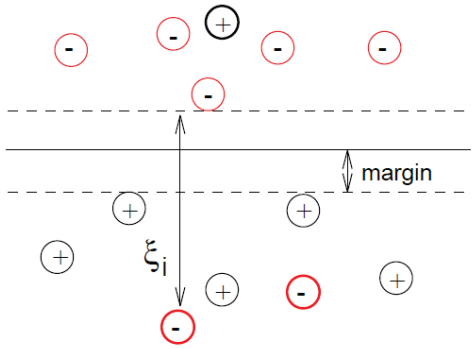


Figure 17 Graphical description of the problem a linear SVM solves, in two dimensions. The datapoints that we seek to classify belong to two classes, labeled as – and +. SVM finds the hyperplane (in the two-dimensional case, a line) that separates datapoints from the two classes as best as possible. The unique solution is given by the maximum margin constraint: the distance of the closest datapoints to the separating hyperplane is maximized. In this case (as in most real life datasets), there is no perfect solution; the dataset is not linearly separable, hence the classifier makes errors (circled in bold). The optimization that linear SVM solves includes a maximum margin constraint and a minimum error constraint; the cost parameter C controls the weight of the minimum error constraint (C >1 means that making few errors is more important than having a maximum margin).

Support Vector Machines solve binary problems exclusively. Multiclass problems (more than two conditions) have to be reformulated as a series of binary problems. The package that I use (LIBSVM, by Chih-Chung Chang and Chih-Jen Lin, http://www.csie.ntu.edu.tw/_cjlin/libsvm) uses an All vs. All scheme, meaning that each condition is compared in turn against each other condition, and the final results are determined by a voting procedure. There are other ways to reformulate multiclass problems, but it seems that All vs. All performs just as well as any of them.

How do you set things up for a fMRI decoding analysis? Pattern classification relies on a training phase, during which the classifier learns and finds the hyperplane that best separates the training data. Then the performance of the classifier on an independent, never-seen-before dataset is evaluated. If the hyperplane learned in the training phase allows classifying examples above

chance in the testing phase, the classifier is said to generalize; the information that it picked up in the training phase was meaningful.

The most common mistake when doing classification is to somehow take a peek at examples from the testing set during the training phase; the accuracy on the test examples will, in that case, be spuriously high. It is much easier to slip into this pitfall than you would expect, when you start tossing and turning your data to make it work. To control for this, I find it absolutely necessary to *always* repeat the entire procedure that led to a seemingly significant performance on shuffled data (i.e., randomizing the class labels before running the analysis). If you find yourself with above chance performance in the shuffled data, you did something wrong.

The optimal way to estimate the generalization performance of a classifier is to use a cross-validation procedure. Say you have a complete dataset, consisting of ten runs of fMRI data. One of the best ways to perform classification on such a set is to leave one run out (testing examples), use the other nine runs for training a classifier (the more training data, the better), and record the performance in the run that was left out. This procedure is repeated ten times, leaving each run out in turn for testing. In the end, the average of the ten accuracies that you recorded is a good estimate of generalization performance. I favor this procedure (leave-one-run-out) over other schemes (such as leave-one-example-out), especially when running a General Linear Model to estimate the responses to different conditions and using the parameter estimates for decoding; indeed, parameter estimates for different conditions within a run are not independent from each other.

This brings us to the question of what constitutes an example, what kind of data you should input to a classifier. If you have a (slow) block design or a slow event-related design, you could simply look at the percent signal change compared to baseline and feed this information to the classifier. If you have a rapid event-related design, the responses to successive trials add up and it becomes

necessary to perform some sort of deconvolution to retrieve the response to each condition; this is when you would use the GLM parameter estimates for input to your classifier. You could feed the β values directly, or you can use T-values, which take into account the variance of the parameter estimates; some have claimed that T-values give better results in general[34], and I usually use these.

There are a few more things to consider. One of them is that linear SVMs (and some other classifiers) need features that are not too wildly different in range. An important step is thus to normalize the input to the classifier. I usually normalize each dimension (voxel) using a z-score (taking out the mean and dividing by the standard deviation). A beginner's mistake here would be using the testing examples when computing the mean and standard deviation; you should only look at the training examples, and apply the same scaling to the test examples. Another method I often use is to separately z-score the data from each run, to account for potential differences between runs. Then, you can run the SVM algorithm. As I mentioned, linear SVMs have a cost parameter C, which it is recommended to train, in the machine learning community at least. You can train the C parameter with a cross-validation procedure within the training set by trying out different values of C, and choosing the one that leads to the best cross-validation accuracy in the training set. I have found that this is rarely beneficial in the case of fMRI decoding; it is time consuming, and likely to fit noise rather than signal. I usually do not bother and keep the cost parameter at its default value in LIBSVM, i.e., 1.

One last thing that you may want to do is perform feature selection, in the training phase. In theory, to train a linear classifier properly, you should have at least as many examples as you have dimensions. In fMRI, this rarely is the case. In practice through, linear SVMs are quite robust to the so-called "curse of dimensionality". To alleviate the curse however it is possible to first trim down the dimensions before running the algorithm. There are many ways to do so (for a review, see[58]). While in principle it is a good idea to perform feature selection, I find that it is

often misused. This is the step where I have noticed the most mistakes in the literature. For instance, how do you decide how many dimensions to keep once you have ranked the dimensions using a given criterion? You should definitely not decide based on your test accuracy. Again, I have found that feature selection, if properly applied, rarely improves results unless the number of voxels is really far too big for linear SVM to handle correctly with few examples (say, more than 1000-2000). If you are considering a larger dataset, you should definitely reduce it before attempting classification.

This concludes my attempt at making you an informed reader, and giving you a concise primer on the methods I use throughout this thesis. You can now safely go on, and perhaps, like me, wonder as you read through certain sections of this thesis, "How on earth could he think it would work?" As I said, this is a story of my childhood as a fMRI researcher. Learn from my experience, and you may save yourself a couple years of growing pains.

II. STUDYING THE UNCONSCIOUS PROCESSING OF INVISIBLE FACES WITH FMRI

AT A GLANCE

- I set out to study the effects of top down attention on the unconscious processing of faces and houses in the fMRI scanner.
  - ⇨ I did not find evidence of an effect of spatial attention.
  - ⇨ I did not find evidence of an effect of categorical attention (only a couple of pilot subjects).

- I felt the urge to check whether there was unconscious activation to invisible faces in the first place.
  - ⇨ I piloted a replication of Jiang & He's 2006 study, but could not replicate their result.
  - ⇨ I implemented a parametric variation of CFS mask contrast in the fMRI scanner; the data convincingly showed the absence of a univariate response to invisible faces in FFA. However, some papers published at the time that I was conducting this study made it somewhat redundant.

- I set out to compare sandwich masking and continuous flash suppression in a well-controlled behavioral design, using a famous/unfamiliar priming task.
  - ⇨ For priming with the same view, we did not find a significant difference between CFS and masking. We are still testing the different view condition.

- My experience is that studying unconscious processing is tricky business, and is taken somewhat lightly in some studies. I established a set of guidelines which should be carefully considered by researchers delving into this field.

As I mentioned in the Acknowledgements section, I joined Christof's lab after reading *The Quest for Consciousness*[59]. I wanted to work on the neural correlates of consciousness, and that is what I did, at least for the first couple of years.

The word consciousness can be understood at two levels: 1) being conscious vs. being unconscious (as you are when in a dreamless sleep); or 2) being conscious of a particular stimulus vs. not experiencing it. There is much fascinating research that is being done about the first level: what are the enabling factors of consciousness? Is a vegetative person conscious?[60] In the Koch lab however, we are more concerned with the second level, the contents of consciousness. In an awake, normally functioning brain, what are the characteristics of neural representations that lead to a conscious experience? Vision is a great modality to work with to address this question. Indeed, cognitive psychologists have come up with a large repertoire of techniques to present visual information to the brain, while preventing that information from eliciting a conscious percept. It is a complicated and somewhat messy literature. In the first section of this chapter, I briefly introduce the reader to the main techniques in that field and point out some of the difficulties that arise when studying the unconscious. Then, I go on to describe some of my own research, using fMRI to try and find evidence of unconscious processing for invisible faces, and characterize the modulation of that processing by attentional resources. Faced with many negative findings, I came to doubt existing reports of unconscious activity elicited by invisible faces in the fMRI scanner; I tried to replicate those under well controlled conditions, and I did confirm my hunch. In the next section, I report on a careful psychophysical comparison of two leading visual masking techniques in an effort to reconcile discrepant reports in the literature. I finally attempt to summarize what I have learned in the process, hoping that the insights I gained may serve future students of the unconscious.

A. Some background on the study of unconscious processing

1) Psychophysical magic: rendering visual stimuli invisible

When I introduce my research on unconscious processing, I almost invariably start with the Necker Cube[2] (**Figure 18**), a simple two-dimensional line drawing of all the edges of a three-dimensional cube, which the brain can interpret in one of two ways; when you look at this drawing, your conscious perception usually alternates between the two interpretations. The input to your visual system remains constant throughout; the switch occurs somewhere in your brain, where presumably both interpretations are represented at some level but only one of them can be consciously experienced at any point in time.

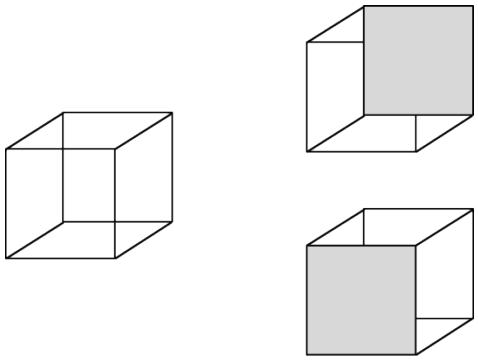


Figure 18 The Necker Cube (left) is a line drawing of a three-dimensional cube. The depth information in this drawing is ambiguous, leading to two plausible three-dimensional interpretations (right), which compete for awareness and usually alternate upon prolonged viewing.

One of the most striking phenomena in visual science, binocular rivalry, relies on the same principle, that of a constant physical input which has a constantly evolving interpretation. It consists in showing two different images to the two eyes, for instance a picture of Jennifer Aniston in the left eye, and a picture of Brad Pitt in the right eye. This is obviously a very unnatural setting, but the way the brain deals with it is quite striking; rather than averaging the inputs into a weird morph between Jen and Brad, the brain seems to push one input at a time to

[2] I have never used the Necker cube in my research.

the contents of consciousness. There is no question that both inputs are represented at early stages of visual processing; however, one of the inputs is silenced, at any point in time, and the other input is consciously experienced. What is the nature of the representation of the non-conscious input? How does it differ from the representation of the currently conscious input? This is a fascinating paradigm, and much effort has been spent (and is still spent) investigating it in great detail (for reviews, see for instance[61–64]).

The idea that some information is processed by our visual systems without reaching consciousness is both troubling and exciting. How much does visual processing without awareness influence our behavior? To answer this question, techniques were further developed to fully and controllably render visual stimuli invisible and study their processing by the visual system. The two techniques employed most often by psychophysicists nowadays to make visual stimuli invisible are masking (backward masking, BM), and continuous flash suppression (CFS). While BM has been around for a century, CFS was introduced rather recently, in 2005, by Nao and Christof [65]. I will describe both techniques here in turn.

Backward masking belongs to the larger family of visual masking methods. The most common set up for BM is fairly simple. The target image, which the experimenter wants to present unconsciously, is flashed rather briefly, usually less than 50 milliseconds. Then, directly following the offset of the target image (i.e., with a stimulus onset asynchrony of about 30-50 milliseconds), a brief mask image is presented (usually, somewhat longer duration than the target, maybe about 100 milliseconds) at the same location (this is paracontrast masking; sometimes one may prefer the mask to not overlap with the target, which is referred to as metacontrast masking, but it is less common in the unconscious processing literature). The mask image may be random or structured noise; for example, Sid Kouider uses overlaid upside down faces to mask face images most effectively[66] (**Figure 19**). In backward masking, the mask comes after the target image. Forward masking can be effective, too (the mask comes before the target), and can be

used in combination with backward masking for even stronger suppression (sandwich masking). The mask is most commonly presented in the same eye as the target (monoptic masking); dichoptic masking is not as commonly used in the unconscious processing literature.

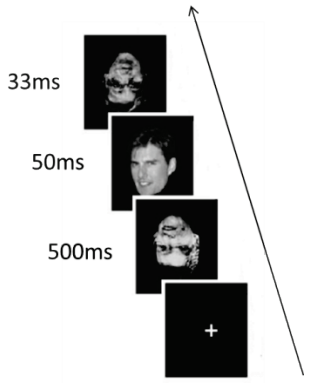


Figure 19 An example of sandwich masking (forward + backward). Tom Cruise's face is presented for 50ms, preceded and followed by masks made of overlaid upside down faces. The arrow represents time. Adapted from [66].

Continuous flash suppression was inspired by the basic phenomenon of binocular rivalry, which I described earlier. In binocular rivalry, if one of the images is higher in contrast, or more salient in some other way (colors, motion), it typically dominates perception for much longer periods than the competing image. CFS capitalizes on this observation and consists in presenting one eye with a high contrast, usually colorful pattern, which is updated periodically (every 100ms or so), while showing a low contrast, static picture in the other eye (**Figure 20**). In this way, one can achieve reliable suppression of the static image, for long durations (sometimes on the order of minutes). Armed with these tools, the next step is to establish the absence of conscious perception, before one can claim the existence of unconscious processing.

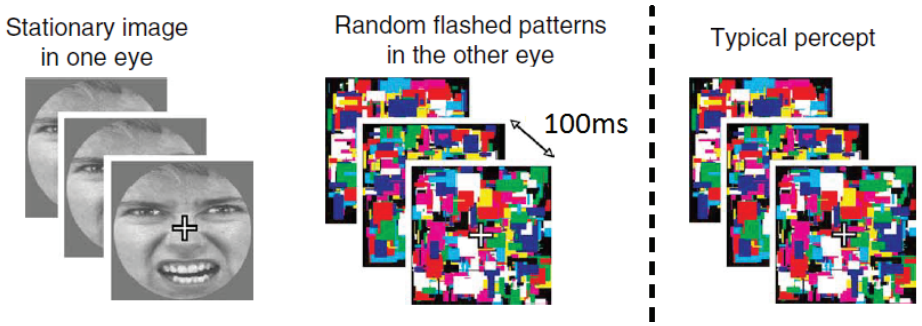


Figure 20 Continuous flash suppression (CFS) paradigm. A stationary gray stimulus is presented in one eye (left) while a series of different colored patterns are flashed in the other eye (center) every 100ms. Subjects fixate the central cross. Typically, subjects are only aware of the dynamic colored patterns and the stationary face remains invisible. Adapted from [65].

2) Establishing the absence of conscious perception

How do you establish the absence of conscious perception? It seems such a silly question, right? Simply ask subjects whether they saw the target or not, and you shall know whether it was consciously perceived.

Since researchers started venturing into the realm of unconscious processing, there has been a heated debate on whether invisibility should be established subjectively or objectively. Clearly, a first-person account seems most appropriate, since conscious experience is highly personal. The issue with subjective reports is that some subjects may be too conservative; in other words, they may report that they did not see a given stimulus, when they did, in fact, perceive some aspects of it consciously, simply not enough to be confident in saying what it was. The typical context in which our psychophysics experiments are run does not help. Knowing that the experimenter will be looking at their answers, the subject may be reluctant to report that they saw something when they only had a fairly vague percept; what will the experimenter think if they answer incorrectly while claiming that they have seen the stimulus? This criterion issue led some researchers to reject subjective reports altogether and concentrate on objective measures of consciousness. For instance, the subject may have to report whether a masked stimulus was a face or a house, and

whether she was conscious or not of the stimulus is inferred from her performance on this task. But what if unconscious processing leads to above chance performance, even when the subject was truly unaware of the stimulus? This is the essence of the complex interplay between "exclusiveness" on the one hand, and "exhaustiveness" on the other. A measure is exhaustive if it can detect any change in awareness, however small. But if we are to trust it to be a true measure of awareness it should also be exclusive, i.e., not classify any unconscious processing as conscious. There have been many discussions in the literature on the best way to measure conscious awareness (e.g., [67–73]), and the perfect method simply does not exist. However there seems to be a convergence towards a best practice, which I came up with on my own, and which I have seen many other groups come up with as well. Conscious awareness should be assessed on a trial-by-trial basis; a summary measure is not enough, as it cannot pick up trials when full invisibility was not achieved. Only a subjective measure can be used for single trial assessment; objective measures can only provide a summary of performance. Using a combination of an objective task (such as a forced choice about the invisible stimulus) and a subjective task (either confidence ratings, post-decision wagering, or visibility ratings such as the Partial Awareness Scale), one can implement trial selection and summary measures of conscious awareness that are independent of the subject's criterion (e.g., unequal variance Receiver Operating Curves, etc.). I will get back to these thorny issues as I discuss the current state of the literature and establish detailed guidelines at the end of this chapter.

B. Effects of attention on the processing of invisible faces

When I arrived at Caltech, I started working with Nao Tsuchiya on a question that he and Christof held dear, that of the relationship between attention and consciousness. Nao and Christof had observed that many researchers tended to conflate attention and consciousness (e.g., [74,75]). The confusion mainly stemmed from the rather loose working definitions that cognitive psychologists had for consciousness and attention. Consciousness, in this context, referred to the contents of

conscious experience. Attention (top-down, endogenous, voluntary attention; as opposed to bottom-up, exogenous, automatic attention, which is based on stimulus properties, such as saliency), on the other hand, referred to the mechanism that allowed certain information to be somehow processed more thoroughly by the brain, leading, for instance, to better performance in visual tasks. Until we have a clear mechanistic understanding of these two psychological constructs, the question of their interplay remains ill-constrained. Nao and Christof still thought it was timely to insist that attention and consciousness were entirely distinct brain processes[76]. They argued it should thus be possible to have four key situations: 1) consciousness and attention; 2) consciousness and no attention; 3) no consciousness and attention; 4) no consciousness and no attention. Finding examples in each of these four categories was Nao and Christof's main argument, and I reproduced the corresponding table from their 2007 review[76] (**Table 1**).

	Might not give rise to consciousness	Gives rise to consciousness
Top-down attention is not required	Formation of afterimages Rapid vision (<120 ms) Zombie behaviors	Pop-out in search Iconic memory Gist Animal and gender detection in dual tasks Partial reportability
Top-down attention is required	Priming Adaptation Visual search Thoughts	Working memory Detection and discrimination of unexpected and unfamiliar stimuli Full reportability

Table 1 A four-fold classification of unconscious percepts and behaviors. Reproduced from [76]

It is not my intent here to fully enter this discussion of what distinguishes attention and consciousness, whether they are entirely different brain processes or whether they rely on similar mechanisms to some extent. Briefly, when I started working on these matters, two of the situations listed in **Table 1** were rather incontestable: that attention and consciousness can co-occur, and that absence of attention and absence of consciousness can co-occur. The two situations that were more controversial were that of attention without consciousness, and consciousness without attention. My attempts were focused on demonstrating effects of attention without consciousness using brain activity measured by the fMRI scanner.

At the time, a very influential paper[77] had just been published in Sheng He's group, which demonstrated that invisible faces still elicited activity in the face selective areas of the brain (such as the fusiform face area, FFA, and the superior temporal sulcus, STS; see **Figure 21**). This boded well for our venture; invisible faces would elicit fMRI activity, which would be modulated by a proper attentional manipulation, and we would get a high impact publication. Let us see how things turned out.

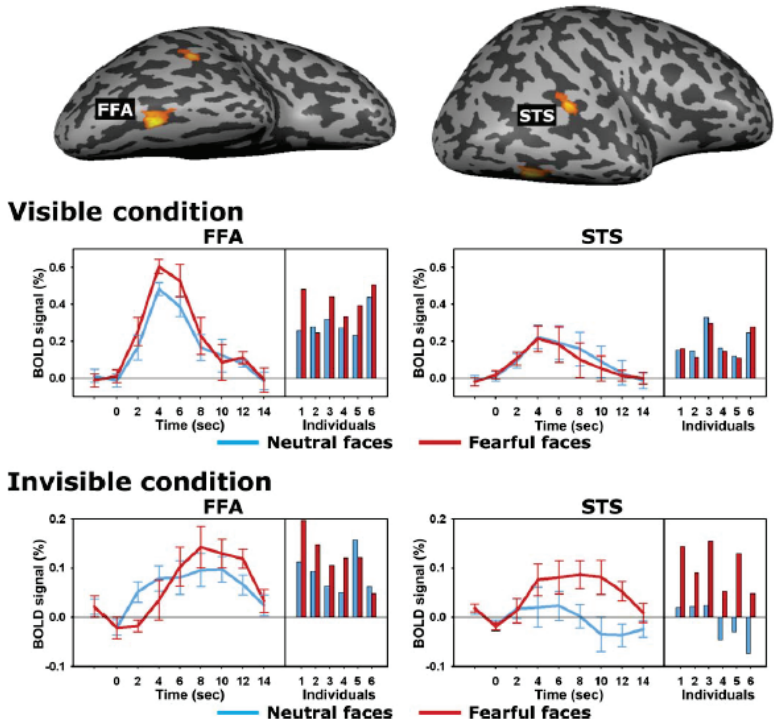


Figure 21 fMRI Responses of face-selective areas to both visible and invisible face images. (A) Face-selective areas (FFA and STS) were identified with an independent scan and depicted on the inflated right hemisphere of a representative observer. (B) Results for the visible condition. Each panel shows the time course averaged from six observers with scrambled faces as the baseline as well as the BOLD amplitude for each individual. Results from the left hemispheres are similar to the data shown here for the right hemispheres. Both the FFA and the STS had strong activations to visible neutral (blue curves and bars) and fearful (red curves and bars) faces. (C) Results for the invisible condition. Each panel shows the time course averaged across six observers and BOLD amplitude for each individual. Even when observers were not aware of the nature of the pictures presented in this condition, the FFA still showed substantial activation for both invisible neutral and fearful faces, whereas the STS only responded to invisible fearful faces. Error bars stand for standard error. Reproduced from [77].

1) The spatial attention experiment

There are different ways to pay attention. For vision, one way is to concentrate on one location in the visual field; this is referred to as spatial attention, and is arguably one of the most studied forms of attention. A metaphor that is often used in studies of spatial attention is that of a spotlight[78], illuminating a location while other locations are, comparatively, in the dark (spatial attention has also been described as a zoom-lens[79], to capture the relationship between the size of the attentional focus and the relative enhancement of processing at that location). Many questions remain on the mechanisms of spatial attention, such as whether two attentional spotlights can exist concurrently (a question that is dear to me, since I worked on it quite a bit, e.g., [5,6]). In the current context however, these are irrelevant details. It is enough to know that spatial attention can, on average and at the time scale of fMRI recordings, enhance processing at two distinct locations (as demonstrated in [80,81]).

The design that we used was heavily inspired from a published study by Vuilleumier and colleagues[82,83], and is represented in **Figure 22**. Four stimuli, two faces and two houses, are arranged in pairs, on the horizontal and vertical axes (e.g., two face pictures on the horizontal axis and two house pictures on the vertical axis). The face pictures could either both have neutral expressions, or both have fearful expressions. In Vuilleumier's experiments, the subject had to selectively attend to one of the two axes (horizontal or vertical) and decide whether the two images on that axis were the same or different. We replicated this setup as one of our experimental conditions. Our interest however was in presenting the face and house images unconsciously, and we used continuous flash suppression. Since a same-different task on invisible images could not be used any more to foster attentional selection (note that it could have been used as an objective measure of awareness), we added opaque red dots on top of the images and implemented a same-different task on the configuration of the dots.

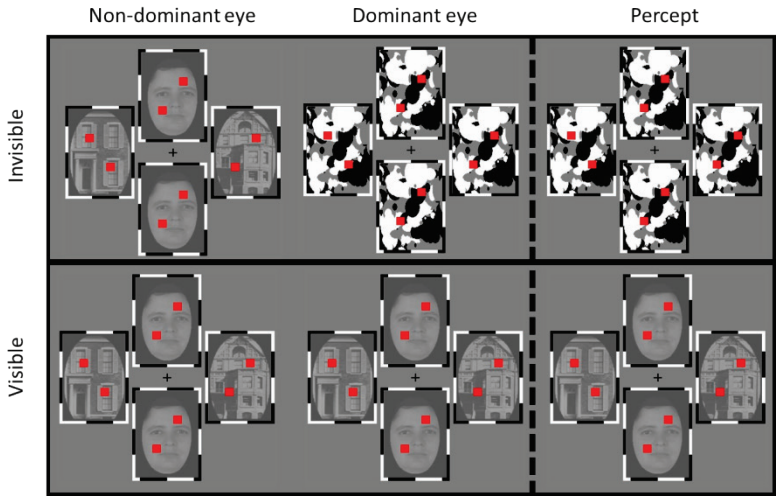


Figure 22 Stimulus display for the spatial attention experiment. The display was inspired by [82,83], except that we used continuous flash suppression to render stimuli invisible and study their unconscious processing. In the invisible condition (top), static face images were shown on one axis (here, vertical) and static house images on the other axis (here, horizontal), in the non-dominant eye (left column); while in the dominant eye, masks were shown in the four quadrants and dynamically changed every 100ms (middle column). The resulting percept, if suppression was strong enough, was simply of a series of random masks (right column). In the visible condition, the faces and houses were shown in both eyes, leading to their conscious perception. Red dots were superimposed to all images (whether masks or faces/houses); in some conditions, subjects had to perform an attentional task on the configuration of the dots (see text for more details).

To increase the signal-to-noise ratio, and thus our chances to observe effects of attention on the processing of invisible stimuli, we used a block design. Each block lasted 16 seconds and consisted of 11 trials; each trial consisted in the presentation of the display shown in **Figure 22**, for 0.5 seconds; trials were separated by a fixed intertrial interval, during which subjects were to give their response regarding the same/different dot configuration. At the end of each block, subjects were asked two additional questions: 1) whether the stimuli on the attended axis were faces or houses and 2) how confident they were, on a scale from one to four, where one: guessing and four: absolutely sure. Twenty-four blocks were presented during each fMRI run (divided equally into four conditions: neutral faces on the attended axis, fearful faces on the attended axis, neutral faces on the unattended axis, and fearful faces on the unattended axis). The same axis (either horizontal or vertical) was attended throughout a run.

I scanned four subjects on the finalized version of the paradigm, each of them undergoing three sessions of fMRI, for this experiment (I got a CBIC Discovery Grant for 20 hours of scanning to pilot this project). The first session consisted of six invisible runs, using continuous flash suppression. The second session and the third session consisted of six visible runs each. In one of the visible sessions, the task remained the same (same/different dot configuration task); in the other session, the task was the task used by Vuilleumier (same/different images). The stimuli were presented in the scanner with MR compatible goggles.

I adjusted the contrast of the face and house stimuli to ensure invisibility. I did not use a staircase procedure or such; I simply started at a nominal contrast value (that I knew worked for most subjects, from pilot data) and if I observed that the subject started seeing stimuli, I decreased the contrast in the next run. Behaviorally (**Figure 23**), the manipulation worked as expected:

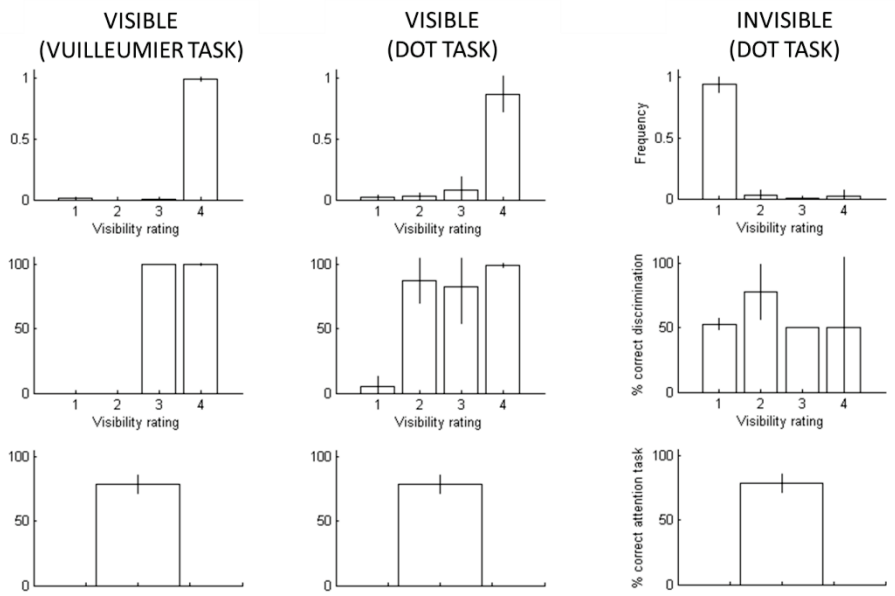


Figure 23 Behavior (average over four subjects) for the spatial attention experiment. The first row shows the distribution of confidence/visibility ratings, the second row the corresponding performance in the face/house objective discrimination task, and the last row corresponds to the performance in the main task. Columns are for the three sessions of fMRI data, with the three different paradigms described in the text. Note that performance on the main task is comparable across the three different paradigms.

invisibility was achieved with our CFS paradigm. Furthermore, the main task (either same/different dot configuration or same/different picture) was equally difficult in all three sessions.

For each of the four subjects, I also had two runs of a functional localizer, contrasting blocks of face pictures and blocks of house pictures, which allowed me to define the classical set of face selective areas (Occipital Face Area; Fusiform Face Area; functional Superior Temporal Sulcus), as well as the Parahippocampal Place Area and other posterior visual areas responding more to pictures of houses than faces.

Vuilleumier had reported that spatial attention enhanced the response to faces in the FFA (and that fearful faces elicited a higher activation than neutral faces)[82]. I replicated both of these findings, in the session where I used the same task as Vuilleumier (**Figure 24**, left). In the other

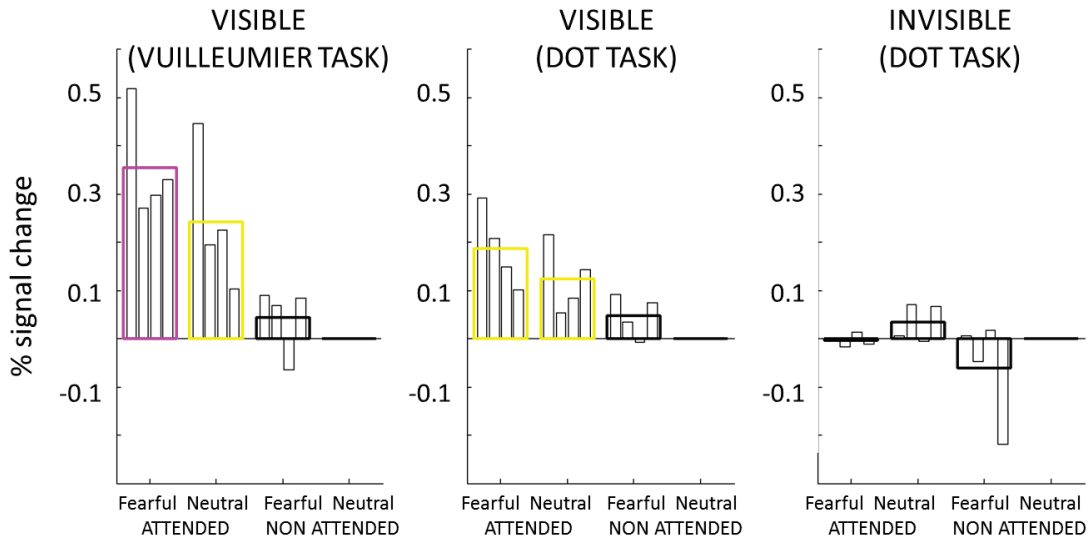


Figure 24 Effects of attention and emotion on the BOLD response in FFA (bilateral). All percent signal change values were computed against the baseline condition Neutral Faces Unattended (the rightmost condition). The larger bars represent the average across subjects, and the smaller bars are the values for each subject. Colors denote the significance of a (one-sided) t-test against 0 (yellow, $p<0.05$; magenta, $p<0.01$). Note that the visible condition with a dot task replicates the results of the visible condition with the original Vuilleumier task. Note also the absence of an increased activation to spatially attended faces in the invisible condition.

visible session, in which I used the dot configuration task, I still found an effect of attention on the BOLD response in FFA (**Figure 24**, center), though the effects were smaller in amplitude. Finally, in the invisible session, with the same dot configuration task, I did not find evidence of any effects of attention on FFA's response (**Figure 24**, right). Note that I rejected the trials for which subjective visibility was anything but one (guessing), a conservative criterion compared to standards in the literature.

Thus I could not evidence any effects of spatial attention on the processing of invisible faces in FFA, the most studied face responsive area in the human brain. I was no more successful in other face responsive regions such as the OFA and STS (which both show an effect of attention in the visible condition). I also conducted whole brain analyses for each subject. A second-level analysis is a bit premature, with only four subjects. However, I noticed an interesting pattern, which was present in three of the four subjects: the contrast attended faces vs. unattended faces yielded strong activations in the posterior, early visual areas, in the invisible condition (and only in the invisible condition; see **Figure 25**). At the moment, I am unsure how this can be explained. If anything, the house pictures used in our experiments elicited more activity in the early visual cortex than the face pictures; the effect is thus the opposite of what one would expect from a low-level perspective.

Spatial attention did not seem to enhance the processing of invisible faces in conditions where it usually enhances the processing of visible faces. Perhaps I should have expected this result. A behavioral study of the tilt aftereffect with invisible adaptors[84] (using CFS) had previously shown that spatial attention had no effect on adaptation while another form of top-down attention, feature-based attention, did (see also [85]). Feature-based attention describes the mechanism that allows you to, say, look for your daughter's red boots when trying to find her in a crowd of

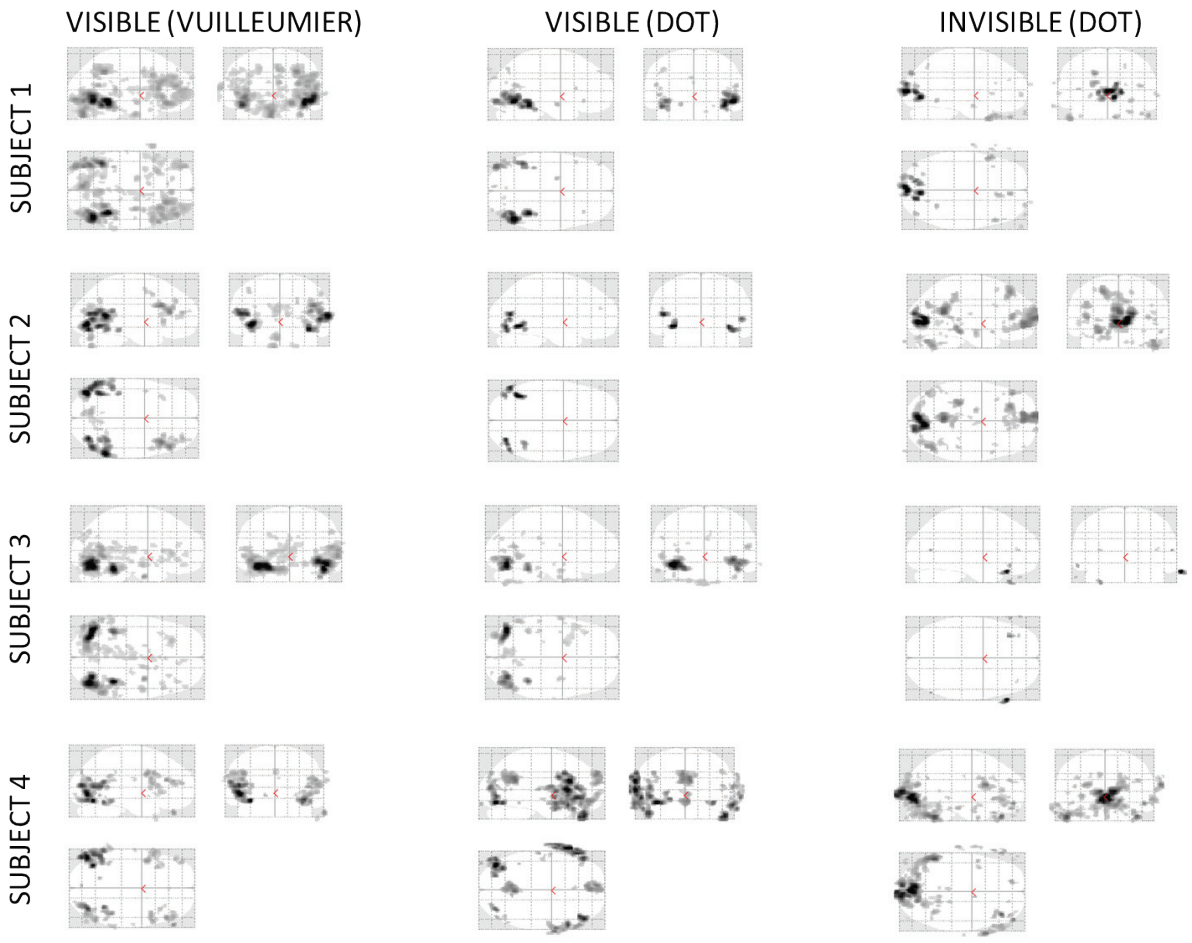


Figure 25 Whole brain statistical parametric maps, shown on the glass brain, for each of the four subjects in the spatial attention experiment for the contrast attended faces vs. unattended faces. Data was normalized to the MNI template, and smoothed with a 6mm kernel. Maps were thresholded at $p<0.001$ uncorrected. The glass brain is a useful representation in which the brain is transparent and you see activations at all depths for each section (top left, sagittal; top right, coronal; bottom left, axial). Note the fusiform activation in all subjects in the visible conditions (corresponding to the results of **Figure 24**). Note also the (puzzling) activation of the early visual cortex in the invisible condition, in three out of four subjects.

hyperactive kids. You can voluntarily attend to a chosen color, or to a certain orientation, etc.[86]. One characteristic of feature-based attention is that it can act on the whole visual field.

There is yet another form of endogenous attention, object-based attention[87], which describes focused attention to a given object. It is perhaps best exemplified in a situation like the face-vase illusion, in which an ambiguous drawing can be seen as a vase, or as two faces; you can attend selectively to one or the other interpretation of this drawing. The experiment that I just described

comprised an attentional manipulation beyond simple spatial selection; the Vuilleumier task, in which subjects were forced to attend to the images on which they had to make a decision, differed from the dot configuration task, in which subjects could completely ignore the irrelevant underlying images and concentrate on the dots. We saw that there was a difference in the resulting amount of attentional modulation in FFA, the Vuilleumier task leading to bigger effects (**Figure 24**). I surmise that this is the effect of object-based attention, on top of spatial attention effects.

With these considerations in mind, I designed a new task in view of harnessing the power of feature-based attention, and hoping for an effect on the processing of invisible stimuli.

2) The feature-based attention experiment

The final version of the display that I designed for a feature-based attention experiment is pictured in **Figure 26** (I went through a few iterations). In a central frame, two images were superimposed, a picture of a house and a picture of a face. The subject's task was to pay attention selectively either to the face or the house. The central display was updated every 1.5 seconds, and subjects had to perform a 1-back memory task, pressing a pre-assigned key if there was a repetition in the stream that they were attending to. Meanwhile, in the periphery, a face image was presented either to the left or to the right, and a house image was presented on the other side. These images were irrelevant to the task, however, the subject was informed of their existence. Each run consisted of 16 blocks, each lasting 16 seconds. At the end of each block, the subject was asked whether the face image was on the left or on the right in the periphery. Also she was asked to rate her confidence, on a scale from one to four (as in the previous experiment, we implemented both an objective and a subjective measure of awareness, for each block).

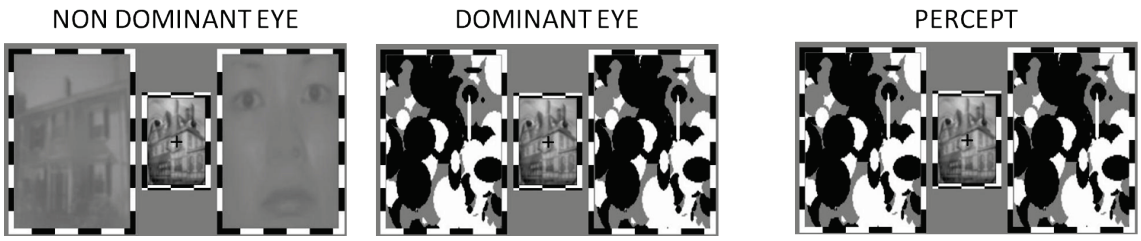


Figure 26 Basic display for the feature-based attention experiment. In a central frame (with a fixation cross in the center), an image of a face and an image of a house are shown in transparency, at the same contrast. This central image blend is updated every 1.5 seconds, during a 16 seconds long block, and the task of the subject is to report any repetitions in the attended stream (either face, or house). There are two peripheral frames. In the non-dominant eye (left column), one frame contains a static face image and the other a static house image. In the dominant eye (middle column), a series of masks are presented in each of the peripheral frames, updated every 100ms. The typical percept (right column) is of the masks in the periphery, and the image blend in the center.

What did I expect? If a subject was paying attention to a stream of faces in the center, which I considered to be a form of feature-based attention, my hypothesis was that this should lead to increased processing of faces elsewhere in the visual field; this was how feature-based attention worked with simpler features such as orientation, color[88], motion direction[88], etc.[3]. I realize as I am writing these lines that it is quite unclear whether one can talk of feature-based attention for anything but low-level features, but it did not dawn on me at that point. In any case, I expected to measure an increased activation in the early visual cortex of the contralateral hemisphere due to the increased processing of the peripheral face when central faces were attended (the same reasoning could apply to houses). I thought this should show up well in a couple of contrasts: "Attended Faces vs. Attended Houses, peripheral face on the left" should show an activation in the right early visual cortex, and a deactivation in the left; "Attended Faces vs. Attended Houses, peripheral face on the right" should lead to the opposite pattern.

[3] I should mention that, at the very onset of my ventures into the effects of attention on the processing of invisible stimuli, I started with motion stimuli, and was trying to replicate Melissa Saenz's feature-based attention result[88] with invisible motion; I was very new to psychophysics and failed to mask motion properly, hence abandoned the idea after a few months. I then started working with static stimuli, choosing faces and houses because they had been studied quite extensively with fMRI and specific brain regions seemed to be dedicated to their processing.

The trouble is that I made several design mistakes, such as having the "Attend to faces" and the "Attend to houses" instructions in different runs; in fMRI, you cannot easily contrast conditions that are presented in different runs, but I was not aware of this yet. Also, I was so confident that my feature-based attention prediction made sense, and so intent on getting right away to the final finding (the attentional modulation under invisible conditions), that I did not collect many visible runs. Visible runs were an important sanity check, to see if my prediction held with consciously seen stimuli. In the one subject for whom I did things right (all conditions present in each run, two visible runs), I did not find a conscious effect that matched my expectations (the effect is as expected in the right EVC, but opposite to expectations in the left EVC, see **Figure 27**). In the two subjects that I ran properly in the invisible condition, I did not find the hypothesized effects. Additionally, I encountered problems with CFS in the periphery; suppression seemed to break quite easily. This is unfortunately all I can report about this experiment, as I look back at it today. I thought I should mention it because the idea was interesting, and I spent quite a lot of time (two months) working on it; however I was a novice fMRI researcher and made too many mistakes.

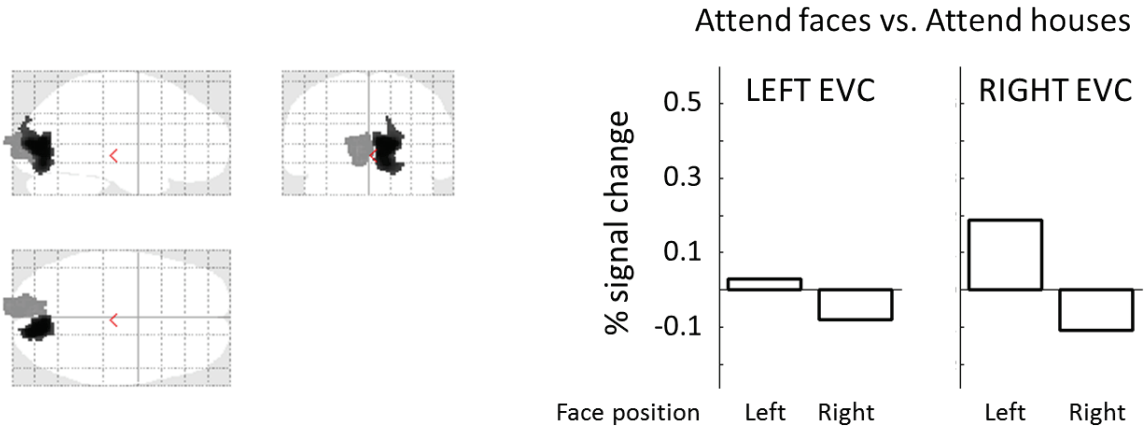


Figure 27 Checking predictions in the visible runs of the feature-based attention paradigm, in one subject. Left, results of a functional localizer consisting of alternating blocks of flashing checkerboards in the peripheral frames; dark gray, left visual field > right visual field; light gray, left visual field < right visual field ($p<10^{-5}$). Right, percent signal change for the contrasts discussed in the main text: Attention to faces vs. Attention to houses, when a face is presented on the left, and when a face is presented on the right. The pattern is as expected in the right early visual cortex (EVC): a face presented in the left visual field is enhanced by selective attention to the central stream of faces, and leads to a higher signal in the right EVC. However, it is against our prediction in the left EVC.

Following my perceived failures in evidencing effects of attention on invisible stimuli, I was beginning to doubt the very foundations of the experiments that I had been doing thus far. Was the evidence of fMRI activity to invisible faces, which I had taken for granted, a reproducible finding? What if I had been trying to modulate the processing of stimuli whose purported processing is not picked up by fMRI signals? I had to convince myself that fMRI could measure unconscious processing of faces under well controlled experimental conditions.

C. Revisiting the foundations: is there fMRI activity to invisible faces?

Jiang and He's finding[77] was initially accepted by the community as soon as it was published. But should it have been? I needed to check this for myself. In a first step, I attempted an almost exact replication of their experiment (with three subjects). Then, I addressed some concerns that I had about their design (namely, the control of invisibility) and conducted a full-fledged study to characterize the amount of processing that occurred at varying levels of visibility.

1) Replicating Jiang and He's study

I tried to follow the methods disclosed by Jiang and He[77], almost to the letter. Their experiment involved six subjects; I only used three subjects for my replication, in an attempt to save on precious fMRI scanning funds (you would be right in criticizing this replication attempt as being underpowered). The methods described by Jiang & He are as follows. An event-related fMRI design was implemented: during each two seconds long trial, subjects viewed a static face stimulus, which could be a fearful, neutral or scrambled face. Trials were separated by two seconds long intertrial intervals. A run of the experiment consisted of 48 trials; the order of presentation of fearful, neutral and scrambled faces was governed by pseudorandom m-sequences[89]. There were two visible and four invisible runs (see **Figure 28**). The subjects' task was to detect an occasional size change of the fixation cross. Continuous flash suppression (at

10Hz) was used to render faces invisible in the invisible runs; the face stimuli were presented binocularly in visible runs. The contrast of face stimuli was adjusted for each subject, so as to ensure invisibility. Participants underwent a 2-AFC experiment before scanning, in a fully comparable experimental situation. Trials consisted of two successive two seconds long temporal intervals, separated by a 500 milliseconds blank. An intact face was presented randomly in one of the two intervals, and a scrambled face in the other. Subjects had to decide whether the intact face was presented in the first or in the second interval. 100 trials were collected before the fMRI experiment, and 100 trials were collected after the fMRI experiment. Additionally, Jiang and He asked subjects whether they saw any face parts after each run of the invisible experiment.

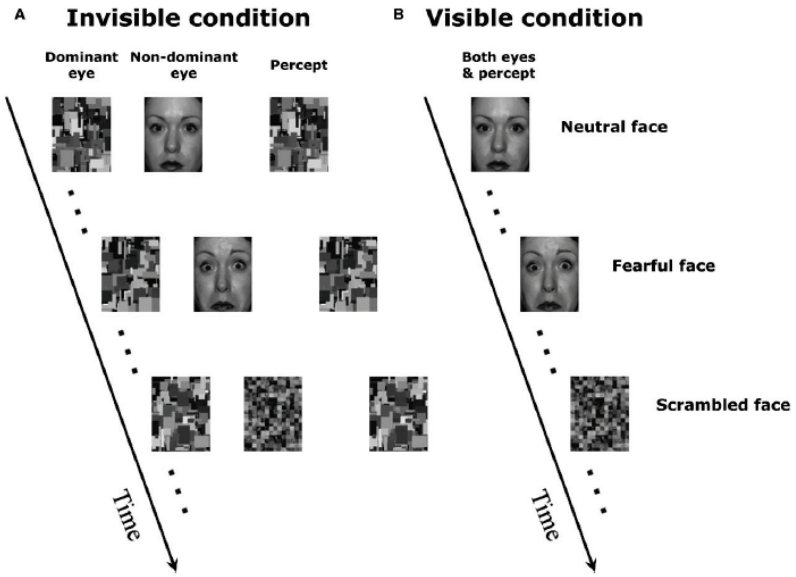


Figure 28 Design of the Jiang & He study, which I tried to replicate. (A) In the invisible condition, the intact face images with neutral and fearful expressions and the scrambled face images presented to the non-dominant eye can be completely suppressed from awareness by dynamic Mondrian patterns presented to the dominant eye because of interocular suppression. The suppression effectiveness was verified by objective behavioral experiments. (B) The visible condition was the same as the invisible condition except that the Mondrian patterns were not presented; instead, both eyes viewed the same face or scrambled face stimuli. Reproduced from [77].

In fact, the visibility task that I implemented was slightly different; instead of a temporal 2-AFC, I simply asked subjects whether they thought they saw a face (detection). This made the visibility assessment shorter (each trial is only two seconds, instead of 4.5 seconds). I told subjects that

there was a face in 50% of the trials; hence they should randomly guess "face" about 50% of the time even when they do not see anything (should experimenters suggest behavior?). The reason I chose this task is that it is better matched to the main experiment, with a single two seconds long stimulus presentation at each trial. I ran three subjects, 100 trials of the visibility task before and 100 trials after the fMRI experiment. I used Signal Detection Theory to compute d', and the results are shown in **Figure 29**. Interestingly I had used a higher contrast for one of the subjects, who did not seem to perceive anything in the practice trials; clearly, she adapted to the paradigm and by the end of the experiment, suppression was not effective any more. This subject should be discarded from further analysis (even by Jiang & He's standards, who claim that their six subjects were at chance before and after the fMRI session). It is worth noting that the other two subjects also exhibited a trend towards seeing more with more practice. I will come back to the important issue of threshold stability and practice in the next experiment.

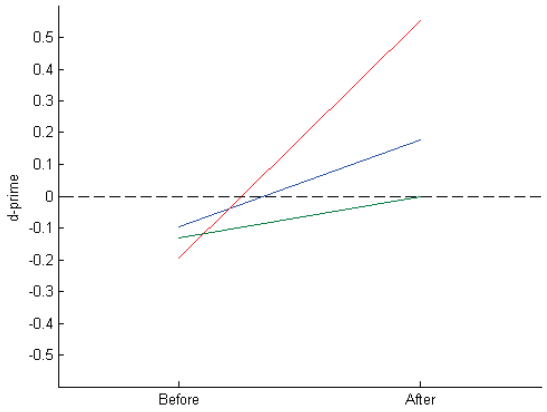


Figure 29 d' for the visibility experiment in my replication of Jiang & He, for three subjects (red, green and blue). All subjects were unaware of the stimuli before the fMRI experiment, as evidenced by a d' that is not greater than zero (it is unclear why d' is actually negative: if the subjects had no information, d' should be zero on average; this is likely to just be due to the small sample). However, by the end of the experiment, at least the red subject clearly perceived some of the stimuli, which is problematic if one wants to claim unconscious processing (see the section on good practices, page 81).

The other small difference between my design and the one reported by Jiang & He is the scrambling method that I used; instead of cutting images up into squares, I Fourier-transformed

the original images, then replaced the real phase values with random phase values, and reconstructed images with those random phase values. This procedure maintains the power in all frequencies and does not create sharp edges as the tessellation does.

I defined the fusiform face area in each subject using a separate functional localizer (blocks of faces and houses, contrast faces vs. houses). Then, I looked at the difference between parameter estimates (from a General Linear Model) for neutral and scrambled faces on the one hand, fearful and scrambled faces on the other hand, both in the visible runs and in the invisible runs. The results are plotted in **Figure 30**. Though I did replicate their finding of a larger response in FFA to visible fearful than visible neutral faces (this is a classical finding), I did not replicate the activation of FFA in the invisible condition which they found consistently in all subjects. Clearly, my study was underpowered, with only three subjects (one of whom was above chance in the post-experiment visibility test). However, I did not think that it was worth pursuing it any further – whether I replicated their result or not, I would not be able to publish it if I did not add an interesting twist to the paradigm.

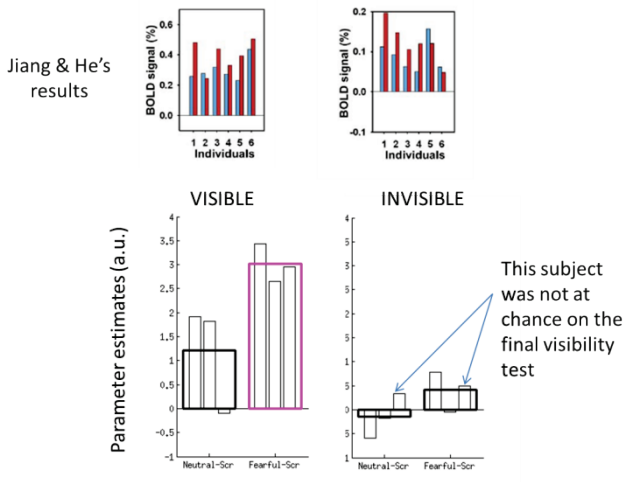


Figure 30 Results of my replication attempt for Jiang & He's experiment. Top, their results in the Fusiform Face Area (FFA), modified from [77]. Bottom, my results. I have no problem replicating a strong activation to visible neutral (left bar) and fearful (right bar) faces, as compared to scrambled faces. However, in the invisible condition, there is no clear activation to either neutral of fearful faces in FFA.

2) A parametric CFS experiment

The methods used by Jiang & He are not very conservative. Their visibility control relied on a display that was quite different from the main experimental display, and it was not sensitive to the presence of a small proportion of slightly visible trials in the main experiment; whether they actually rejected subjects on the basis of the subjective question, "Did you see any face parts in this run?", which they reportedly asked after each run, is not specified in the paper. I wanted to run an experiment in which I could track visibility on each trial, in order to convince myself that there could, indeed, be a fMRI activation in face responsive areas in response to invisible faces. I also wanted to see whether fMRI activity reflected visibility in a linear fashion, or in an all-or-nothing fashion (this question has been addressed by other groups with other techniques, e.g., the attentional blink[90]). I wanted to compare how different areas of the brain were affected differentially by stimulus strength and subjective visibility. Finally, I wanted to check whether CFS could be used repeatedly in the same subjects, and whether the visibility threshold would evolve over time. Here is how I eventually set things up, after a few iterations and aborted pilot experiments.

I recruited 11 subjects, most of which underwent two one hour and fifteen minutes long sessions of fMRI (two subjects only underwent one session). I aimed at having subjects performed eight runs of the main task, each run consisting of thirty-six trials. There were twelve conditions (hence each run presented each condition three times): four contrast levels for the masks, and three stimulus categories: fearful faces, houses, and nothing. Each trial consisted of a random ITI (between one and four seconds), followed by six seconds of stimulation (ten repetitions of a 300ms on, 300ms off sequence), followed by an objective task (was it a face or a house?), followed by a subjective visibility rating (how much of the stimulus did you see?).

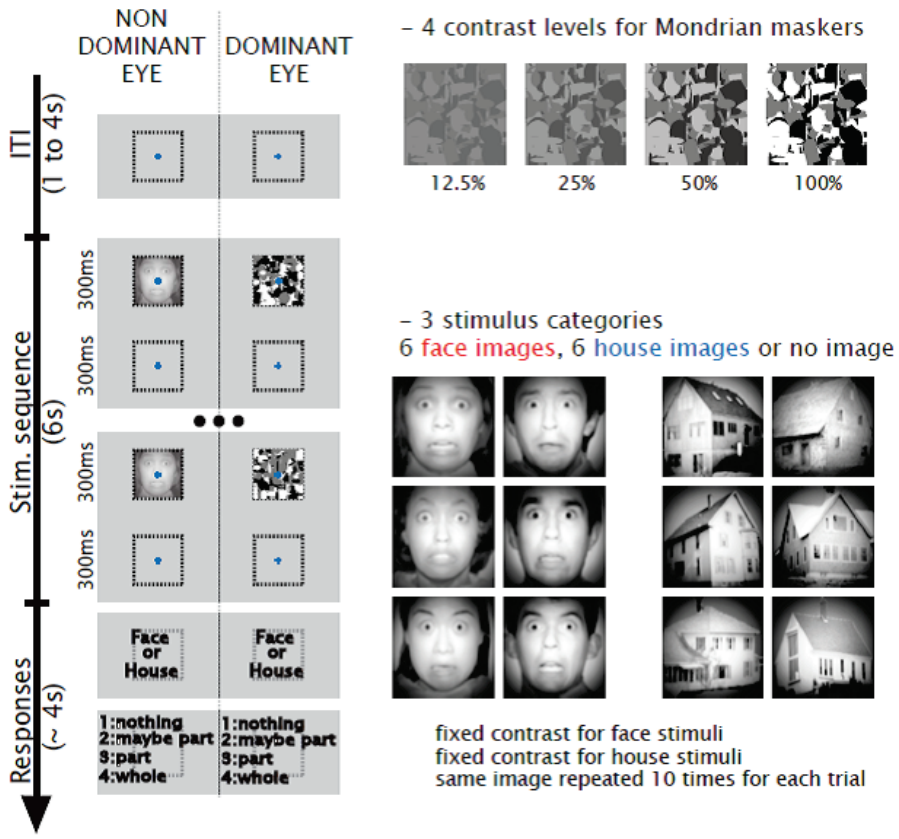


Figure 31 The paradigm for my parametric CFS experiment in the fMRI scanner. Left: temporal sequence of a block. The arrow represents time. After a random ITI (1<ITI<4 seconds), a 6 seconds long sequence is shown, consisting of 10 successive 300ms presentations of a static image in the non-dominant eye and a sequence of masks (changing every 100ms) in the dominant eye, separated by 300ms blank periods; the same static image is shown 10 times in a given block. At the end of this, two questions ensue: a two-alternative forced choice objective visibility task and a four-alternative forced choice subjective visibility task. Right, top: four possible mask contrasts (same mask contrast throughout a given block). Right, bottom: images shown in the non-dominant eye are fearful faces, houses (or nothing). See text for more details.

Each subject was first invited to the lab and I determined the contrasts for face and house images with a staircase procedure, so as to obtain a 70% correct performance. The mask contrast was fixed at a given value (which varied between subjects; I had not yet decided on the best value). My threshold determination procedure was based on a temporal 2-AFC, similar to the one described by Jiang & He[77]. A trial consisted of two 300ms intervals separated by 500ms. An image was shown in one of the intervals, and a scrambled image in the other; subjects were to decide which interval contained the meaningful image (a house or a face). The thresholds measured during this behavioral session were not intended to be used for the fMRI experiment;

rather, I wanted subjects to experience the display so that they would not be naïve. I also had the subjects perform a couple of "dummy runs" of the fMRI experiment outside the scanner to get used to the attentional task (reporting a size increase of the fixation cross).

I performed three threshold contrast determinations in the first session of fMRI (at the beginning, in the middle, and at the end); and two threshold determinations in the second session of fMRI (at the beginning and at the end). For a given subject, all determinations were done either at 0.125 mask contrast, or at 0.25 mask contrast. I observed that, on average, the threshold did not decrease with practice; if anything, there was a tendency for it to increase slightly over all subjects (**Figure 32**). There was much individual variability in the strength of masking by CFS: for some subjects, the contrast of stimuli had to be set to extremely low values. Also, I did observe a decreased effectiveness of CFS for some individual subjects (**Figure 33**).

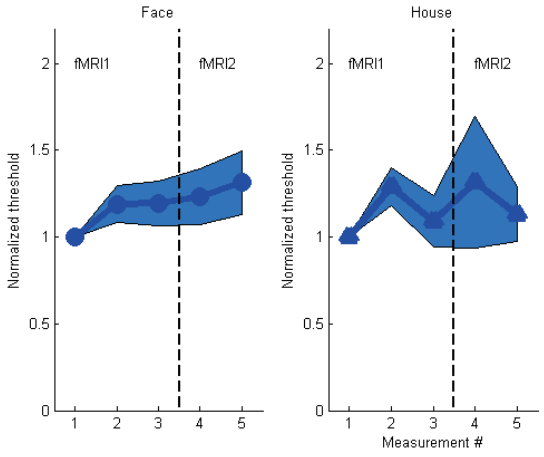


Figure 32 Staircase-determined thresholds for face (left plot) and house (right plot) images were roughly stable over time, across subjects. All subjects had prior experience with Continuous Flash Suppression through a practice session outside the scanner. Threshold measurements were performed five times for most subjects: during the first fMRI session, at the onset, in the middle, and at the end; during the second fMRI session, at the onset and at the end. The shaded area represents the standard error of the mean.

I picked contrasts according to the determination at the beginning of the first fMRI session and stuck with those contrasts throughout the fMRI experiment. Since I purposely varied visibility

with a parametric manipulation of mask contrast, and collected trial-by-trial visibility measures, I would be able to analyze the data even if visibility was higher than I wished it to be.

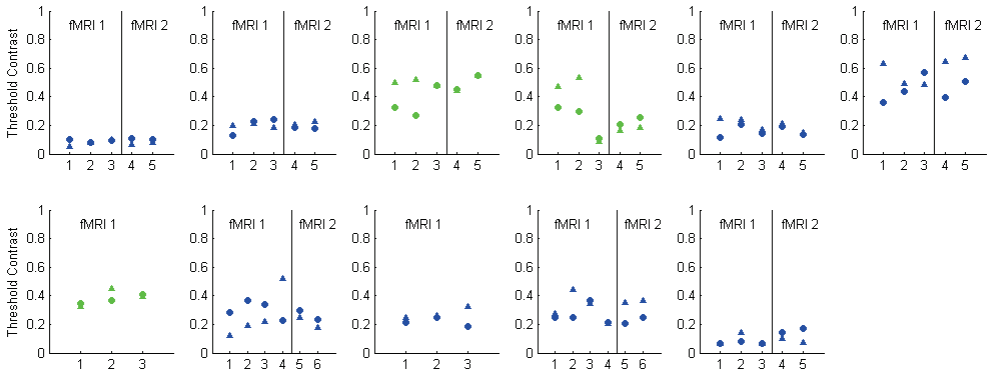


Figure 33 Individual data for the CFS threshold determination with a staircase procedure. Each plot represents a subject. Circles are for face images, triangles for house images. The symbols are green if the staircase threshold determination was done at a mask contrast of 0.25, and blue if done at 0.125. Note that the threshold did dramatically decrease for some individual subjects, for instance the 4th plot in the top row (a lower threshold means that suppression is less effective; the target contrast needs to be reduced).

The behavior of a typical subject (i.e., the best subject) during the fMRI experiment is shown in **Figure 34**.

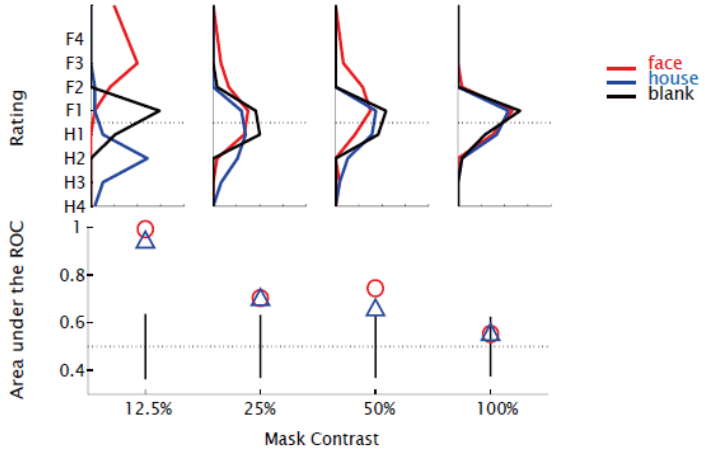


Figure 34 Example behavior for one subject in the main fMRI experiment, showing the effects of manipulating mask contrast on perception. Top panel: distribution of responses (F1: face, lowest confidence; H4: house, highest confidence). Lower panel: area under the ROC curve (red circle: face as signal, house as noise; blue triangle: house as signal, face as noise; black line: 95% confidence interval computed using distribution of responses to blank trials and assigning randomly face and house labels to each trial). This shows that our mask contrast manipulation had the desired effect of changing visibility (and objective performance).

For each subject, in addition to the main experiment, I ran some functional localizers. I was again interested in face and house responsive areas (faces: OFA, FFA, STS; houses: PPA), and thus used a block fMRI experiment, contrasting blocks of face images vs. blocks of house images. These regions were defined on the inflated cortical surface (following [91,92]) using Freesurfer. The results for the right hemisphere of one subject are shown in **Figure 35**.

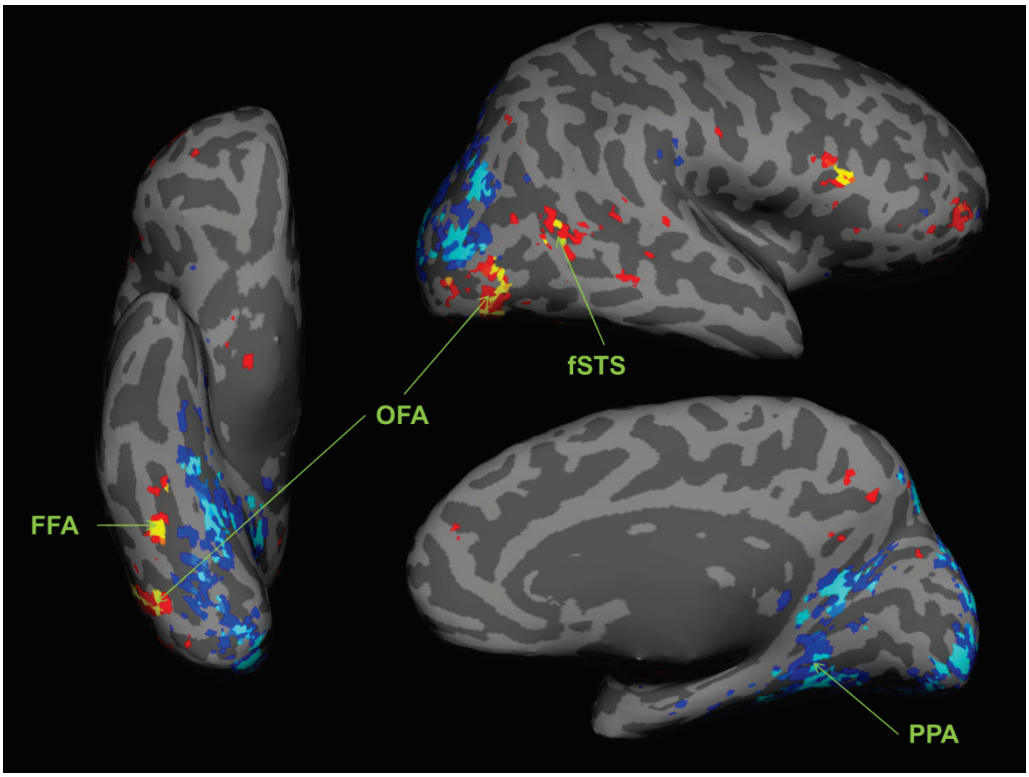


Figure 35 Face and House responsive areas, evidenced with the contrast faces vs. houses in a fMRI block design and "painted" on one subject's inflated brain. The dark gray areas represent sulci (i.e., valleys), the light gray gyri (i.e., hills). Statistical parametric map thresholded at $p<0.001$ uncorrected.

I was also interested in defining V1, and used a simple meridian mapping procedure to do so; the border between V1 and V2 corresponds to the vertical meridian, hence, by contrasting the activation to a flashing checkerboard on the vertical meridian vs. a flashing checkerboard on the horizontal meridian, this border is easily seen on the properly flattened (and cut) occipital cortex. Additionally, I ran a foveal localizer, by contrasting a flashing checkerboard in the center of the

visual field to a flashing checkerboard in the periphery. The results of the meridian mapping and foveal mapping experiments is shown in **Figure 36**. Freesurfer was used for flattening and cutting the cortex.

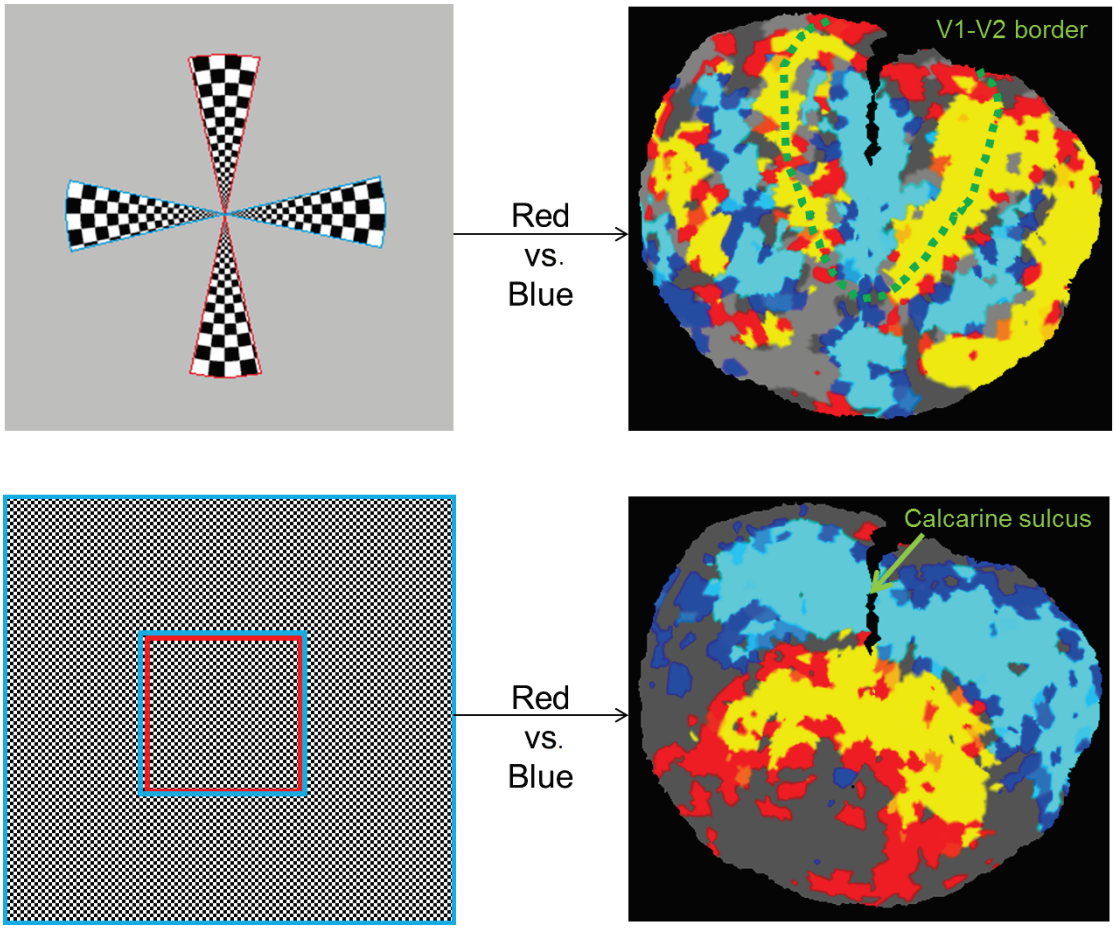


Figure 36 Retinotopy results for V1 definition. Left: stimuli. Right: thresholded statistical parametric map, $p<0.001$ uncorrected. The V1-V2 border is drawn by hand on the surface, following the maximum activation to the vertical meridian.

Finally, since Jiang & He reported that the amygdalae responded to fearful faces at the same level, whether faces were visible or invisible, I also defined the amygdalae in each subject, using automatic anatomical labeling with Freesurfer[93] (**Figure 37**).

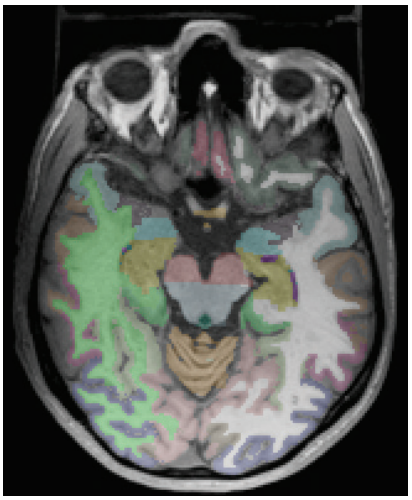


Figure 37 Automatic anatomical labels assigned by the recon-all procedure in Freesurfer[93]. In blue, the amygdalae.

Having defined all these areas, I was then able to look at the signal change related to mask contrast on the one hand and to subjective visibility on the other hand, in these areas. For the effect of mask contrast I performed the following analysis: for each subject, after running a general linear model with one regressor per trial, the average parameter estimate across the ROI for each trial was plotted as a function of mask contrast, and a Spearman correlation coefficient was computed (rank correlation: can capture monotonic, non-linear relationships). The effect of confidence was factored out by considering each confidence rating in turn, then averaging across confidence ratings to get the final correlation (**Figure 38**).

Out of the ROIs that I looked at, only V1 correlated significantly with mask contrast. Interestingly, additional analyses showed that the response in V1 did not seem to saturate, even at the highest mask contrast. The BOLD activation was always higher when a stimulus (face or house) was presented in the non-dominant eye then when nothing was shown (**Figure 39**).

The exact same procedure was used to look at the effect of visibility, factoring out the effect of mask contrast (**Figure 40**). Subjective visibility was reencoded from minus four to four, with negative values

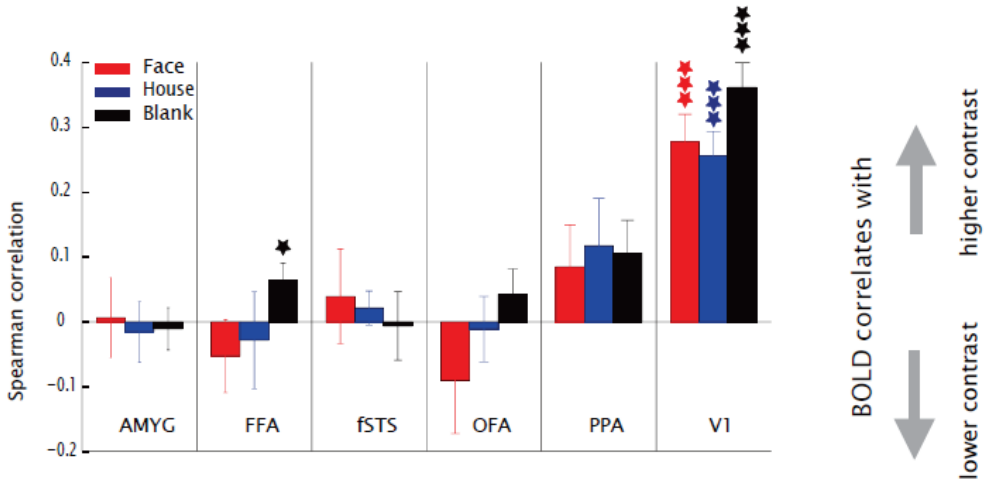


Figure 38 Correlation of BOLD activation with mask contrast (factoring out visibility). For each subject, average BOLD signal change in each ROI for each trial was plotted as a function of mask contrast (log scale), and the Spearman rank correlation coefficient was computed. The effect of confidence was factored out by considering each confidence rating in turn, then averaging across confidence ratings to get the final correlation. 1 star: $p<0.05$; 2 stars: $p<0.01$; 3 stars: $p<0.001$.

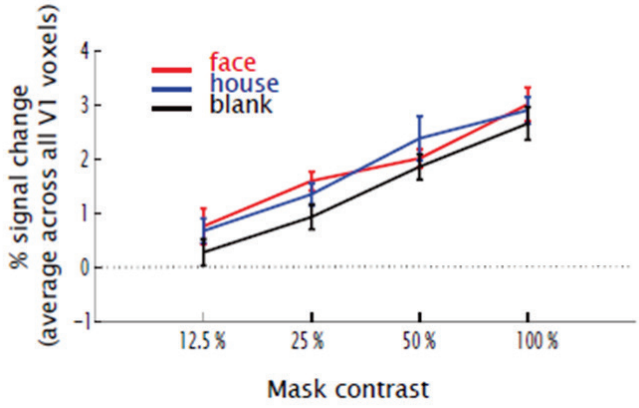


Figure 39 BOLD activity as a function of mask contrast in V1 (note the logarithmic scale on the x-axis). No main effect of mask contrast on difference between face trials and blank trials, or on difference between house trials and blank trials (1-way ANOVA)

corresponding to instances when the subject indicated that they saw a house and positive values for a face. There, I found that OFA and FFA correlated significantly with face visibility, whereas PPA correlated significantly with house visibility. These were the only significant correlations that I observed.

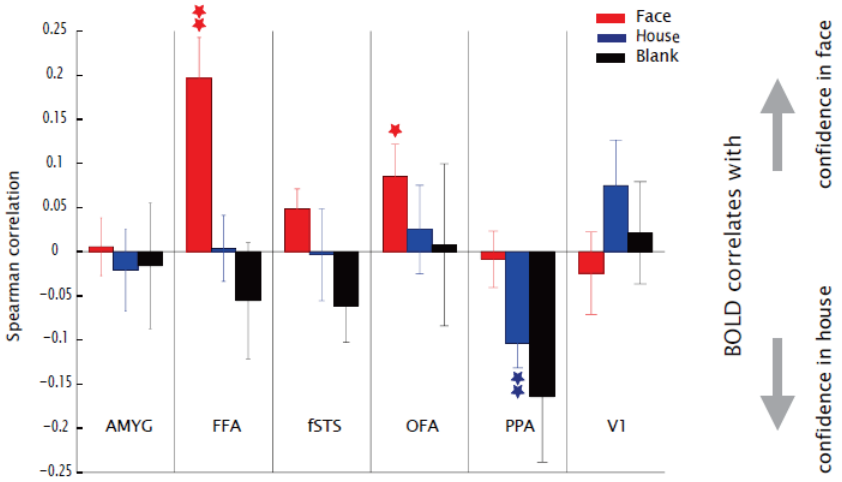


Figure 40 BOLD correlation with subjective visibility (factoring out the effects of mask contrast). For each subject, the average BOLD signal change for each ROI for each trial was plotted as a function of confidence rating, from -4 (house, high confidence) to 4 (face, high confidence). The effect of mask contrast was factored out by considering each mask contrast in turn, then averaging across mask contrasts to get the final correlation. 1 star: $p<0.05$; 2 stars: $p<0.01$; 3 stars: $p<0.001$.

I conducted some further analyses, focusing on the fusiform face area (arguably the most studied face responsive area). I declared mask contrast and subjective visibility as parametric regressors in a GLM, and looked at the corresponding parameter estimates. As SPM orthogonalizes parametric regressors with respect to each other when there are multiple, I performed this analysis declaring either mask contrast as the first parametric regressor, or subjective visibility as the first parametric regressor. I found that subjective visibility explained all the variance in FFA, leaving none to mask contrast. I plotted the parameter estimates averaged across FFA in the case when a face was presented, as a function of subjective visibility. I found that at a subjective visibility of 1 (seen nothing), there was no response in FFA. At a subjective visibility of two, the response was still not significant across the group of subjects. It took visibility three or four to get a sizeable response (**Figure 41**).

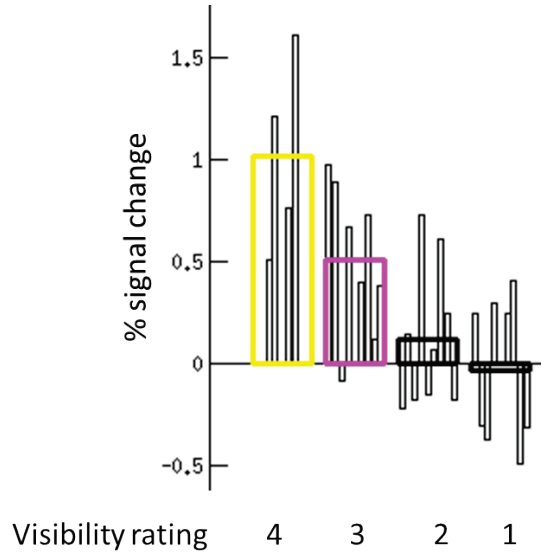


Figure 41 BOLD response to faces at different visibility ratings. The large bars represent average across subjects, the small bars individual subjects. There was no sizeable response on average in FFA voxels when faces were invisible. Only subjects for whom we had enough trials at each visibility level were included in the analysis (yellow, $p<0.05$; magenta, $p<0.01$).

I presented some results for this project at the Society for Neuroscience meeting in Washington DC in 2008, with a poster entitled, "fMRI activation to visible and invisible faces and houses using continuous flash suppression with a confidence rating task." It got quite a bit of attention. I remember meeting Martin Hebart, a student of John-Dylan Haynes also working with CFS, and Guido Hesselmann, who at time was just getting started with CFS (and has since published very interesting studies using CFS, e.g., [94]; note the very tight similarities between that study and mine: even though I did not publish the results of this experiment in a journal, my efforts inspired others). It was then that Martin tipped me off about a study that his mentor had conducted together with Geraint Rees and Philipp Sterzer, and which was in press at Journal of Vision. I contacted Geraint and he sent a very nice reply, together with the manuscript[95]. They had implemented a trial-per-trial measure of visibility (objective + subjective), and reported my main finding, that of the absence of activation in FFA with a univariate analysis. However, they also reported that they could find evidence of unconscious processing using multivariate analysis.

I briefly attempted some fMRI decoding analysis with this data, and was not successful. It could be that I did not try hard enough or that the quality of my data was subpar. Geraint pointed out that they used high-resolution fMRI (1.5mm isotropic voxels), and that smoothing the data hurt their decoding performance. My 3mm isotropic voxels, and the slice angling which led to rather large signal dropout in many of the regions of interest (notably, close to FFA), likely hurt my ability to pick up patterns of activation.

Though I tried to convince myself that my study, with its good control of visibility and parametric design, could still find a niche and be published, this was a big blow. Additionally, in his email, Geraint pointed out that, "Our data have taken a-g-e-s to get into press though; for inexplicable reasons the reviewers thought the data were super-boring (and obvious) which is a bit weird." My interest with studying unconscious processing of faces in the fMRI scanner started withering.

The study of unconscious visual processing remains one of the best ways to study the neuronal processes that are necessary for visual consciousness. While I took a break from hands-on research in this area and started working on different projects (see next chapters), I tried to keep in touch with the literature. In the following section I describe some inconsistencies that I noticed in the literature and the realization that ensued: what if the classical techniques used to induce invisibility were not created equal[4]? In other words, it could be that masking and CFS, though they lead to the same subjective invisibility, do not let through the same amount of visual information to be processed unconsciously, and this may explain some of the controversies.

[4] I should be totally honest here and attribute this phrase to Sergey Fogelson; this was the title of his 2011 ECVP poster: "Not all suppressions are created equal: Categorical decoding of unconsciously presented stimuli varies with suppression paradigm."[182] I have not yet seen the corresponding published paper.

D. The extent of unconscious visual processing in the brain: quantifying differences between suppression techniques

To illustrate the current state of our understanding of the extent of unconscious processing in the brain, I will start with four quotes which illustrate the gap between two communities: the backward / forward masking aficionados, and the binocular rivalry / continuous flash suppression fans.

In 2007, Sid Kouider and Stanislas Dehaene wrote in a very thorough critical review of backward masking studies[96]:

> "Nowadays, while the existence of subliminal perception is no longer denied, the controversy has shifted to the depth of processing of invisible stimuli. While it is largely accepted that lower levels of processing (e.g. motor reflexes, sensory analysis) do not necessitate perceptual awareness, the existence of non-conscious computations at higher levels (e.g. semantic or inferential processing) remains debated."

But within the next four years, Stanislas Dehaene seemed to think that unconscious semantic processing was established clearly, as illustrated by the following assertion from his 2011 review with Jean-Pierre Changeux[97]:

> "Subliminal priming has now been convincingly demonstrated at visual, semantic and motor levels."

On the other hand, the world-leading experts working on binocular suppression techniques (binocular rivalry and CFS) mostly deny the existence of any processing beyond very low-level analysis. Blake and Logothetis were very clear about this matter back in 2002[62]:

> "Does priming occur if the priming stimulus is rendered invisible by binocular suppression? For visual tasks involving higher-level cognitive processes, including picture priming and semantic priming, the answer clearly is 'no' — suppression renders normally effective priming stimuli impotent. These results are not too surprising, for both of these priming paradigms call for relatively refined analyses of visual information, of the sort conventionally attributed to high-level visual processing outside the domain of early visual areas. Evidently, during suppression phases of rivalry, input to those processing stages is effectively blocked."

More recently, Timo Stein and Philipp Sterzer took care to review current findings with CFS quite thoroughly in one of their research papers[98]. They came to the following conclusion:

> "It has repeatedly been shown that priming effects triggered by high-level stimuli that are processed in ventral cortical areas, such as words, line drawings of objects and images of vehicles and animals are eliminated during interocular suppression."

I could go into a lot of detail about the evidence that these researchers rely on to draw such conclusions, and it would be extremely interesting to evaluate each and every original research study that they cite in a critical manner to decide whether the results can be trusted. I strongly believe that a lot of confusion has arisen in the unconscious processing literature from methods that do not properly control for visibility. I have come up with a list of criteria for evaluation of studies of unconscious processing, which I will describe at the end of this chapter.

It is important to emphasize that not everybody would agree with the quotes that I referenced here. A postdoc in my laboratory, Liad Mudrik, has published a paper (written during her PhD) in which she demonstrated the processing of incongruency under binocular suppression[99]. She painstakingly created images of people performing actions, in which she replaced the object of the action by another object which did not make sense in the specific context; for instance, replacing a phone handset with a banana. She found that incongruent images could reach awareness faster than congruent images (the original ones) using the breaking CFS paradigm, which consists in timing the duration of suppression by CFS. Recent controversies with the breaking CFS paradigm[100] made her a little uneasy, though, and she has been working on replicating her finding. She seems to have managed to do so with a BM paradigm which unfortunately does not clarify whether high-level processing is possible under binocular suppression.

The idea that suppression techniques may differ significantly in the amount of information that they let through under conditions of subjective invisibility is definitely in the air. I thought it was

my idea, but then I found it had been expressed by other authors. Interestingly, a postdoc in the lab, Nathan Faivre, who did his PhD in Paris with Sid Kouider, published a recent study[101] comparing BM, CFS and gaze-contingent crowding (the method that he used primarily in his thesis work). Ryota Kanai published an overview of different masking techniques, emphasizing the distinction between perceptual and attentional blindness; he classified both CFS and BM as perceptual blindness techniques. Finally there were two studies by Jorge Almeida and colleagues which implemented a comparison of unconscious categorical processing with BM and with CFS[102,103]. They found categorical priming for animal and tool images when suppressing primes with BM; only tools elicited priming when CFS was used to mask primes. Since, there have been some concerns about low-level confounds in these studies[104]. Though Almeida's idea was very good, I think that there is much room for improvement. For instance, in their CFS paradigm, primes were shown for 200ms and the target directly followed prime presentation; in their BM paradigm however, primes were only 30ms in duration, and there was a 100ms interstimulus interval before the target was shown. One should strive to reduce such differences between paradigms, in order to make strong claims about inherent differences. This is what I did when designing the following experiment.

The paradigm I came up with presented sandwich masking trials and continuous flash suppression trials in a random order; importantly, the subject could not know which it would be on any given trial. In fact, the subject was never told that there was such a manipulation; they believed that the masking technique was the same throughout the experiment. Masking trials and CFS trials were made to be as similar as possible. Of course, there had to be small differences as they were different techniques. The one parameter that I wanted to keep constant at all costs was the input energy, which included stimulus duration and stimulus contrast. Since sandwich masking can only work with brief stimuli (about 30-50ms), we chose a stimulus duration of 50ms. We fixed the input contrast to an arbitrarily chosen value (which worked well in pilot

experiments). The defining property of CFS being the dynamic succession of high contrast masks in the eye whose input the experimenter wants to favor, we came up with the design depicted in **Figure 42**.

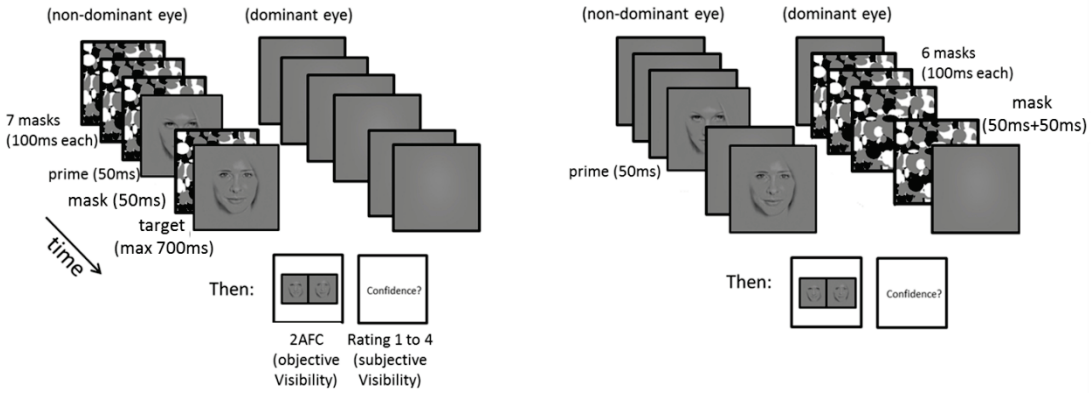


Figure 42 Our paradigm to compare sandwich masking (left) and continuous flash suppression (right). The displays are as equalized as possible, with the same total energy for visual stimulation (8 masks, a prime and a target). Figure adapted from Gregory Izatt's final SURF report (2012).

The total stimulus energy was exactly equalized between the two techniques: masks were refreshed every 100ms, and 8 masks were presented in total. It was only the sequence, and the eye in which the masks were presented, that changed. One issue that could be raised is that there was no dynamic component to CFS during the 50ms prime duration. We believe that we still harnessed the strength of CFS, by enforcing dominance through dynamic flashes in one eye; the duration of the masked stimulus seemed rather irrelevant to us in the context of trying to understand the technique, although the capacity of CFS to mask longer stimuli is an eminently desirable property and a reason for its success in the unconscious processing literature since its introduction.

This experiment was purely behavioral; the marker of unconscious processing that we were after was priming. We knew from the work of Sid Kouider[66] (amongst others) that faces showed priming effects when rendered invisible by sandwich masking, when using a famous/unfamiliar

task. That is to say, if a decision had to be made rapidly on a target face (whether it was famous or not famous), the response time (RT) was faster when the target was preceded by an invisible picture of the same face than when it was preceded by an invisible picture of a different face. Interestingly, the effect was not a purely low-level repetition priming effect: the same-view effect was obtained with a size change between the prime and the target (it is, however, unclear how much this controls for low-level confounds; Nathan Faivre and I have run a V1 model showing that reducing the size of an image by 20% is not quite enough), and the cross-view effect constitutes one more step in the generalization. The results from Kouider et al.[66] are shown in **Figure 43**. Note that the d-primes (measure of awareness) are skewed towards positive, and that the regression which is used to claim the existence of unconscious priming[105] is rather ugly (being brutally honest here).

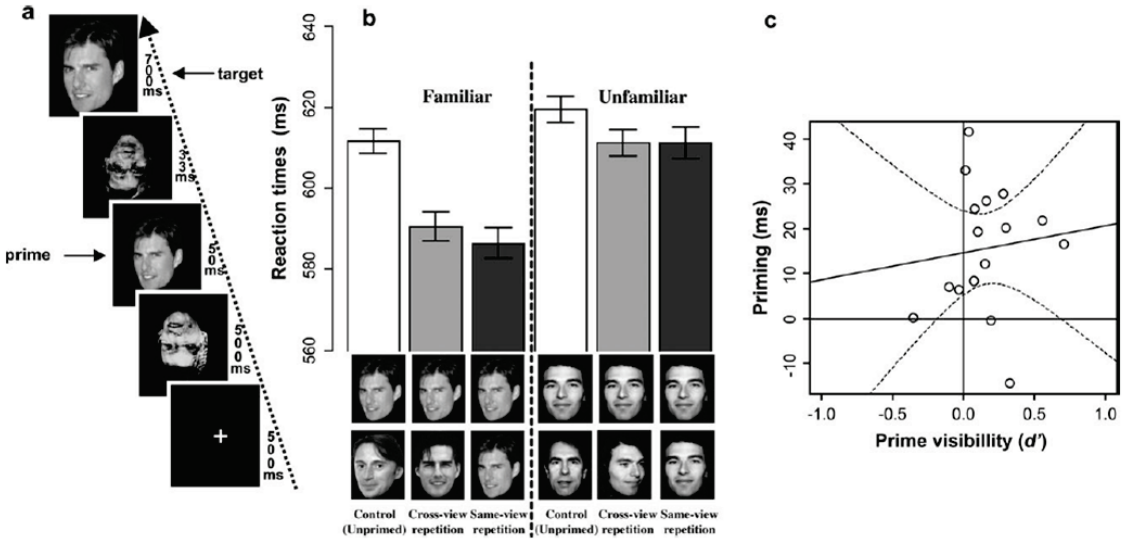


Figure 43 Schematic description of the subliminal face priming method and behavioral results in [66]. (a) Each trial consisted in the sequential presentation of a fixation cross, a forward mask, a prime, a backward mask and the target. Participants were presented with familiar and unfamiliar faces and were instructed to perform a fame-judgment task on the target. Masks were constructed from overlays of inverted faces. (b) Mean reaction times for the six priming conditions. The experiment involved a two-by-three factorial design including famous and nonfamous target faces preceded by a prime that could depict the same person in the same view (same-view conditions), the same person in a different view (cross- view conditions) or a different person (control condition). (c) Regression of priming on prime visibility. Each data point represents a participant. The regression functions (dotted lines indicate 95% confidence intervals) show the association between the global priming effect found for famous faces and prime visibility. Priming is interpreted as subliminal when the curve representing the lowest value in the confidence interval passes above the origin. Reproduced from [66].

The use of faces was motivated by Kouider's finding, an established behavioral finding to build upon, and by the possibility to equate low-level properties of these stimuli to a decent extent (at least, it is much easier than trying to equate the properties of tool pictures and animal pictures[104,106]). I advised a SURF (Summer Undergraduate Research Fellowship) student for the summer of 2012 as he developed a set of Matlab™ functions to realign and scale face images, mask the background and soften external features, and perform histogram and power spectrum matching. This set of functions can be used to work with any set of face images, and I used it myself in the latest version of my fMRI experiment on person identity (see Chapter 3). It may be useful to release it to the community as a toolbox.

There was one more component to this experiment. One way to change the strength of masking is to change the contrast of the masks (the contrast of the masked stimuli being fixed). Rather than titrating the mask contrast for each subject using a staircase procedure (as is often done in the literature), we used the same three mask contrasts for all subjects (as in [102]); this had the effect of making the experiment a little less boring (as the trials did not all look exactly the same), and it allowed us to see the effects of masking strength on unconscious processing (note that this approach is similar to the one that I used in the parametric fMRI experiment that I described previously).

Subjects who came in chose the most recognizable ten male and ten female famous faces, out of 30 male and 31 female faces available in our image database. They were instructed and performed some slowed down practice trials. Then they performed four to five runs of the experiment. Each run consisted of 144 trials, six repetitions for each condition: three mask contrasts, 0.02, 0.4 and 0.6; two masking methods, sandwich masking and CFS; two target categories, famous and unfamiliar; two target/prime relationships, related and unrelated. The size of the prime was reduced by 20 percent compared to the size of the target (11.1 vs. 13.5 degrees of visual angle). The only constraint on the order of trials that was implemented in the design was

that a given face could not be presented in consecutive trials, either as prime or target. Credit should go to Gregory Izatt for piloting the code that I had originally written, polishing it, and running all subjects.

As I write, only the same view priming experiment was run. Nathan, Greg and I joined forces for the analysis. We followed classical steps for the analysis of RT data. After discarding the first 15 trials allowing subjects to become acquainted with the task, we further discarded any reaction times less than 200ms and more than 1200ms. We also discarded trials for which subjects were wrong on the main task. Before looking at priming effects, we first ensured that the behavior of our subjects was acceptable, especially in terms of their subjective and objective measures of visibility, and their accuracy on the main task. In a first step, we were interested in reproducing a priming effect in well controlled conditions of invisibility; hence we selected only those trials when subjects reported a subjective visibility of one. Out of 20 subjects that were tested we discarded five, either because of their low accuracy on the main task, their above chance accuracy on the 2-AFC objective visibility task despite a subjective visibility rating of one, or outlier data (more than three standard deviations away from mean). All reaction time data was inverted (1/RT) before taking the mean, which is a common step in dealing with skewed reaction time distributions[107]; then the means were converted back to meaningful units by taking the inverse again. Our preliminary analysis of the data (**Figure 44**) shows that there is a priming effect both with sandwich masking and continuous flash suppression, in the same view condition, when pooling across famous and unknown targets. Interestingly, looking separately at unknown and famous targets, we could only find a significant priming effect (i.e., the arbitrary $p<0.05$ in a t-test against 0) in the sandwich masking condition, for unknown targets. The priming effect for the sandwich masking condition was significantly higher than for the CFS condition for unknown targets ($p=0.030$), and the difference between sandwich masking and CFS almost reached significance when grouping famous and unknown target trials ($p=0.066$). Note, however, that

separating conditions leads to very few datapoints in each condition (as few as 5-10 in some subjects) and therefore very unreliable estimates of mean reaction times.

We are in the process of collecting more data with this same view paradigm, and we are also starting to collect data for the different view condition which we hypothesize will lead to a stronger dissociation, whereby only sandwich masking will allow enough unconscious processing as to support a priming effect. This experiment is ongoing.

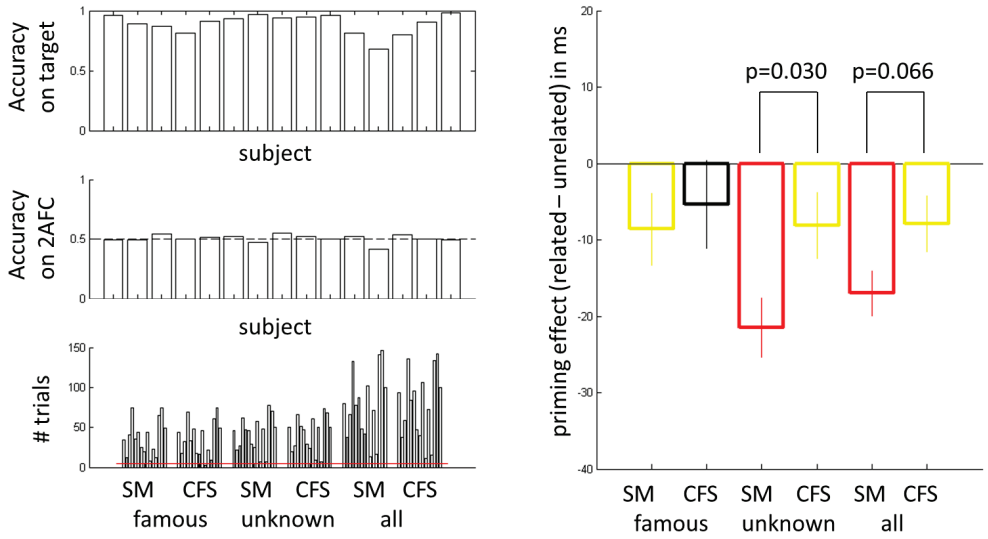


Figure 44 Preliminary results for same view priming in our comparison of sandwich masking (SM) and continuous flash suppression (CFS). This analysis only considers trials that were rated as visibility one, which leads to a variable number of trials across subjects. Left, from top to bottom: accuracy on the main task, whether the target face is famous or not famous; accuracy on the objective visibility task, whether the prime was the image shown on the left or on the right; number of trials for each condition (group of bars), for all subjects (in the same order as in the previous plots). Right: colors represent significance in a t-test against zero (red, $p<0.001$; yellow, $p<0.05$; black, $p>=0.05$); the error bars are s.e.m. As far as we can tell, there is no significant difference between repetition priming in the sandwich masking and continuous flash suppression conditions. We are now investigating cross-view priming.

E. Making sense of a rich, confusing, noisy literature: guidelines for studies of unconscious processing

A reason for my falling out of the unconscious processing field for a few years in the midst of my thesis work was my frustration with the lack of reproducibility of published results; perhaps the best example is my initial reliance on the Jiang & He report of fMRI activation to invisible faces,

which proved to be a non-reproducible finding (as per the published study by Sterzer, Rees and Haynes[95]; my own experiments; and some other unpublished attempts which I was informed of informally at conferences). I believe that time may be ripe for putting down in writing some guidelines for future studies of unconscious processing (in the same vein as the guidelines that Russ Poldrack published a few years ago for reporting fMRI studies[108]). Beyond obvious good practice rules for science in general (summarized in a cartoon by Neuroskeptic, **Figure 45**), I propose that the following points be addressed in all studies (from my experience with this field, and discussions I have had with colleagues over the years).

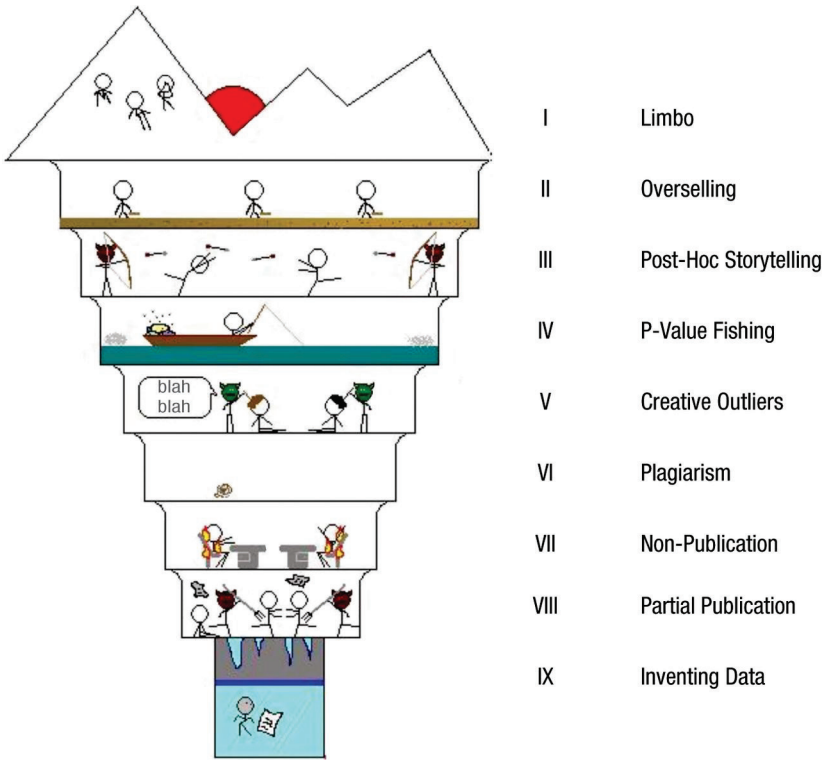


Figure 45 The Nine circles of scientific Hell, by Neuroskeptic. Reproduced from [109].

1) Controlling (in)visibility

It is obvious that you should implement a good measure of visibility when you intend to make a claim about unconscious processes. Attention should be given to the following points, and choices that are made should be reported fully.

a) When is visibility controlled?

- Before/after the main experiment

Visibility can be controlled either before or after the main experiment; it is even better if invisibility is established both before and after the experiment. It is then not too far-fetched to conclude that stimuli were invisible during the main experiment. Usually, when researchers implement these measures, they also informally ask subjects whether they saw anything during the main experiment and take a negative answer as further proof of invisibility.

- During the main experiment, with catch trials

Another option is to have a certain proportion of trials in which questions are asked about visibility (either objective, or subjective). This has the advantage of controlling visibility during the main experiment, under the exact conditions and brain state that the subject is in.

- Trial-by-trial

The most stringent control is to implement a visibility question at every single trial of the main experiment. The advantage is that the researcher can track any state change, which may have led the subject to detect the stimulus that was intended to be invisible; this also offers additional data for analysis, allowing to sort trials according to visibility (for subjective ratings at least). However, visibility questions add time to the experiment and they are very repetitive, which may lead the subjects to try and get rid of these questions and answer as quickly as possible, almost randomly.

As discussed earlier, I have been implementing and would recommend the latter option, the most thorough and conservative, despite the added boredom that subjects may experience. The "before only" approach should definitely be avoided, since adaptation to the technique being used (training effects), and also dark adaptation (most experiments are conducted in the dark, and the time constant of dark adaptation is rather long, on the order of 30 minutes), will likely lead to underestimating visibility. My strongest argument in favor of trial-by-trial measures is that brain state may fluctuate to a great extent within the duration of an experiment, at different time scales. There is now much evidence for different rhythms that affect perception: at a slow time scale (less than 2Hz), see for example [110]; at a more rapid time scale (~7Hz), see my own contributions [3,111]. It is thus of the utmost importance to have single trial measures that follow these otherwise hard to control fluctuations; the catch trial approach cannot address these fluctuations.

There are even more issues to be considered when one settles for the trial-by-trial measurement. The time elapsed between the presentation of the invisible stimulus and the visibility assessment may play a key role in the outcome of measures of awareness, as weak perceptual experiences may be very elusive and quickly dissipate after stimulus presentation.

b) Choice of a visibility measure

I already alluded to the debate between objective measure and subjective measure advocates, at the onset of this chapter. I concluded that it was probably best to use both, in an attempt to satisfy both criteria of exhaustiveness and exclusiveness. Here I delve a little deeper into the choices that are available, for subjective and objective measures.

- Choice of an objective measure

The objective measure is a question about the hidden stimulus, either a detection task (did you see a face? Yes/No), or a two- or N-forced choice discrimination task (2-AFC: was it a face or a house? Face/House). Performance is assessed on average over all trials (on any single trial, it is

possible to get the answer right without having seen anything, by chance), either with a simple accuracy assessment, but more meaningfully with using a signal detection approach, which yields a comparable measure for subjects who adopt different criteria (more/less conservative) or biases.

The main issue in the choice of an appropriate question is what level of information is important for the task at hand. Imagine that the stimulus that is being masked can be at one of two locations (e.g., left/right) and can belong to one of two categories (e.g., face/house). The task is a repetition priming task. Subjects may be able to report the location of the masked stimulus above chance, while being unable to report the category to which the stimulus belongs. Should the ability to report location be taken as evidence against unconscious processing of the category? This is the "partial awareness" problem[112]. Some authors argue that partial awareness should be ruled out to claim unconscious processing[69]. Others warn that the distinction should be made between measuring any conscious content vs. measuring the relevant conscious content[71]. I started out as a proponent of the first alternative; I wanted to rule out any conscious content. I now realize that this may have been too stringent a requirement. In fact, there are studies that addressed this empirically ([113], and a study in progress in our own lab, led by Nathan Faivre, Liad Mudrik and Hagar Sagiv).

My recommendation is thus the following: the objective task should estimate the visibility of the specific features from which the expected unconscious effect arises, e.g., orientation for a tilt aftereffect experiment, and fame for a fame priming experiment.

- Choice of a subjective measure

The simplest subjective measure is a binary choice: seen or not seen. Of course, we discussed that this is too crude, and too dependent on the subject's criterion (in a signal detection theory sense).

A measure that has been proposed is the Partial Awareness Scale (PAS)[114]. This is an intuitive, four-point scale: 1) No experience, 2) Brief glimpse, 3) Almost clear image, and 4) Absolutely

clear image. Note that this measure is directly related to subjective experience, and not to a judgment about an objective answer; in fact, it can be used independently of any other measure.

Another rather intuitive measure, used in combination with an objective task, is a confidence rating. This can only be used when there is a correct answer to the first, objective question. For example, the subject may have to decide whether the masked stimulus was a face or a house. Then, they can be asked for their confidence in their answer, e.g. on a four-point scale: 1) Guessing, 2) Vague intuition, 3) Pretty sure, and 4) Absolutely sure. Confidence ratings permit better data driven signal detection theory analysis, with an unequal variance model. The key difference with the PAS is that confidence ratings are a second-order judgment, a metacognitive assessment of whether one has enough information to answer the objective question accurately.

Finally, there was much excitement around a measure reintroduced by Persaud and colleagues in 2007[115,116], Post Decision Wagering (PDW) which was in fact introduced by Kunimoto and colleagues back in 2001[117]. Instead of a confidence rating, subjects are asked to bet either a low or a high wager on their response to an objective task. The idea is that subjects' drive to win will force them to use any information that they can access, thus providing a direct measure of awareness. Since, there has been much criticism of this original assertion. In fact, the dichotomous decision between a low and a high wager also involves a subject-specific criterion, which will depend on individual characteristics such as conservativeness, risk aversion, emotional arousal when betting, etc. [72]

My recommendation is to use the PAS or a confidence rating. Keep in mind that, per the previous discussion, the PAS may be too conservative if only the "no experience" trials are used… If an objective task is implemented, the confidence ratings seem the most intuitive and allow the specification of unequal variance signal detection theory models.

2) Controlling attention

This chapter started off with my efforts to find effects of attention on the processing of invisible stimuli (faces/houses) in the fMRI scanner. Though I was not very successful, there have by now been behavioral[118–120] and neural[121,122] demonstrations that attention does affect the processing of some invisible stimuli. Hence, attention should be controlled and reported in studies of unconscious processing. Of course, the most basic choice is whether subjects should be alerted to the presence of invisible stimuli at all. In most studies, this is the case especially if one implements a trial-by-trial measure of visibility. The next choices are whether to direct spatial and temporal attention to the stimuli.

a) Directing spatial attention to invisible stimuli

My experiment with spatial attention to faces did not yield positive effects, however, some behavioral studies have claimed that spatial attention affects the unconscious processing of non-face stimuli ([119,120], but see [84]). This evidence suggests the importance of directing spatial attention to invisible stimuli, in order to increase unconscious processing. It should be made clear in unconscious processing studies whether there was any control of spatial attention.

b) Directing temporal attention to invisible stimuli

I have mentioned the existence of attentional rhythms, at different scales, which influence perception. Considering this evidence[123–125], the predictability of the onset of invisible stimuli may be beneficial to their unconscious processing. It should be stated whether the timing was predictable or jittered, as this could be an important factor (Naccache demonstrated the importance of temporal attention in an implicit priming task[118]).

3) Catching the fleeting effects of unconscious processes

If there is unconscious processing, its effects may be short-lived. Typically, unconscious priming only affects subsequent stimuli in a rather short time window (100-200ms). Probing the effects of invisible processes, behaviorally or with imaging methods, should be done soon after the presentation of the invisible stimulus, and the exact time elapsed between stimulation and probing should be reported.

4) Ruling out alternate explanations

Before unconscious processing (at a high-level) is claimed, alternate explanations need to be carefully considered. Here are some example interpretations that should be rejected.

a) Low-level confounds

I am obsessed with low-level confounds, and I think it is for good reason (see Chapter 3). The truth is that there have to be some low-level differences between different stimuli; otherwise, they would be the same stimulus. The concern is in making sure that no systematic low-level difference can explain results just as well as a high-level interpretation. The typical example is the demonstration by Sakuraba and colleagues[104,106] that Almeida's finding of unconscious tool processing in the dorsal pathway[102,103] was really a more general artifact of the unconscious processing of oriented lines.

b) Action triggers

It has been argued that, in many cases, unconscious processing of stimuli occurs because the same stimuli were previously seen consciously. For instance, Eckstein and Henson[126] recently wrote:

> "When prime faces were never presented as visible probes within a test, priming was not reliable; when prime faces were also seen as probes, priming was only reliable if visible and masked presentations of faces were interleaved (not simply if primes had been visible in a previous session)."

I have been told that some researchers who tried to replicate the famous study by Jiang and colleagues with invisible nudes[127] (unconscious, sexual orientation dependent orienting of attention), only managed to do so when unconscious and conscious trials were interspersed; however this information is not reported in their paper.

The size of the stimulus set is also of the utmost importance: if only a few stimuli are used, the brain can easily build representations and associations for each stimulus, and claims of high-level categorical processing should be downplayed.

All this information (size of stimulus set, order of trials, presence of conscious trials, etc.) should be described carefully to foster reproducibility.

5) Effect size

The last recommendation, which is, of course, a general recommendation for any empirical finding, is to report effect size, number of subjects, number of trials per subject, and appropriate statistical tests. A small effect with a borderline significant p-value may not be reproducible[128]. There has been a great emphasis on this in psychology recently, as more and more studies fail the replication test[129–131].

Using these guidelines, an evaluation of all published results about unconscious processing is long overdue. Also, future studies should be mindful of these important choices, and report what they implemented in great detail.

III. STUDYING FAMILIAR PERSON RECOGNITION WITH FMRI

AT A GLANCE

- Pilot experiment: fMRI decoding of the identity of four TV series male characters using many short, silent movie clips.
  - ⇨ I can decode identity in the early visual cortex.

- Pilot experiment: fMRI decoding of the identity of four cartoon characters, using audio and movie clips.
  - ⇨ Again, I can decode identity in the early visual cortex; no evidence for significant decoding of identity across modalities (video <-> audio).

- fMRI decoding of the identity of three famous male actors, using pictures and written names.
  - ⇨ I found no evidence of decoding across modalities (pictures <-> written names), despite the use of a "supersubject" (multiple fMRI sessions to better train classifiers) in the last attempt.
  - ⇨ I recurrently find significant decoding of identity from written names, in the occipito-temporal lobes. This finding needs further investigation.

The previous chapter was dedicated mainly to studying the unconscious processing of invisible faces; finding out what the brain does outside the realm of our conscious experience is a fascinating endeavor. Overall, I found that fMRI did not provide strong evidence that invisible faces were processed unconsciously. It could, of course, simply be that the effect size is so small that it cannot reliably be extracted from noise. I thus thought that it may be wise to start looking at conscious perception, where larger effect sizes are expected, in order to have positive results to report in my thesis.

Since I had started developing a keen interest in the perception of faces, I decided to tackle the great unsolved problem of person recognition[5]. What is your brain doing as you recognize the face of a friend in a crowd, and retrieve information about them, what they do, how you know them, their name, the last time you saw them, etc.? How is the information about people that you know organized in your brain? There likely are different modules for different modalities (recognizing a name, recognizing a face, remembering past shared experiences), but how do these modules connect together to yield the integrated recognition of a person with all of their attributes? A catchy phrase I have used in presentations to summarize my research question is, "Where do you know who you know?" (I should say that this was inspired by a review paper on object recognition that I came across, entitled, "Where do you know *what* you know?"[132])

A. Some background on where you might know who you know

1. The state-of-the-art cognitive psychology model of person recognition

The understanding that we have today of person recognition relies heavily on psychophysical experiments and lesion studies that were summarized and interpreted about 25 years ago[133,134].

[5] Why I keep tackling unsolved problems that many people have had a stab at in the past is a good question. I may be setting myself up for failure over and over again. But who cares about the small questions, for which one already has a good guess at what the answer is? I hate reading a paper's abstract and wondering why anyone on earth would care.

Since then, not much progress has been made on the theoretical understanding of person recognition but this does not necessarily mean that the problem was considered solved. Here is an excerpt from an email conversation I had with Professor Mike Burton (the lead author of the 1990 paper[135] that I will describe in the following):

> "I'm struggling to think of any real developments on the semantic side of person recognition - though I'm not claiming our 1990 work is much of an account. I think what has really happened is that the focus of research has shifted, and now resides either in the visual front end, or in neural processes involved in face recognition. So I think people stopped looking at the questions we'd been addressing back then."

So, what were the conclusions from their 1990 paper? The essence of it is captured quite well in a single figure, a typical cognitive psychology box-and-arrow model of person recognition (**Figure 46**). The paper actually describes a computer implementation of this model, based on "interactive activation competition"[136] scheme which I think of as equivalent to some neuronal models with excitatory connections from one layer to another, and local inhibition within each layer (a very simplified version of what cortex is supposed to be doing).

Briefly, the person recognition system may take as input either a face or a name (written, or spoken, though the latter is not included in the model in **Figure 46**). It has dedicated modules to process these inputs, namely the Face Recognition Units (FRUs) and the Name Recognition Units (NRUs). These modules feed into a pool of Person Identity Nodes (PINs); any view of the face of Brad Pitt, or any instance of his written name, activates the corresponding Brad Pitt PIN. Recognition of familiarity purportedly happens at this stage. The Brad Pitt PIN is connected to a number of units in the Semantic Information Units (SIUs) pool, and this is how one retrieves what we know about familiar people. As regards naming, the model places names in the SIU pool. Note that other authors[137] have proposed to have several sub-pools of units within the SIUs. The name retrieval stage is implemented in the form of Lexical Output Units. Other critical

features of the model are that connections to the PINs are reciprocal, and all units within a given pool inhibit each other.

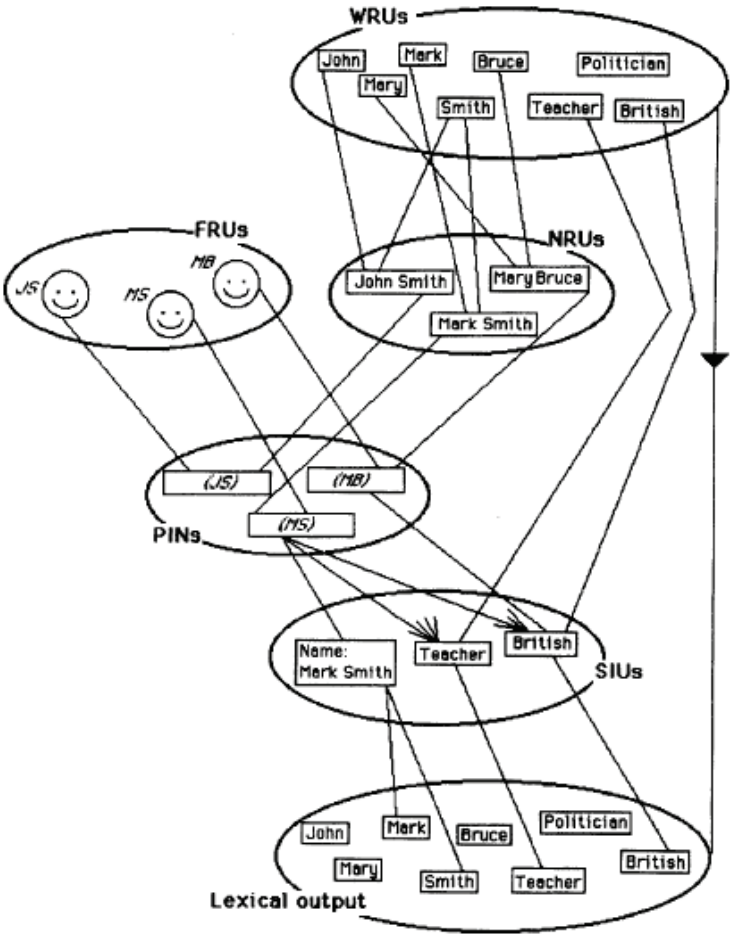


Figure 46 The Burton and Bruce IAC model of person recognition. FRUs (Face Recognition Units) and WRUs (Word Recognition Units) are the input units, which respond respectively when a face is seen and when a name is seen. Units in the FRU pool are tuned to specific identities; they perform invariant recognition of faces (in different conditions of illumination, different viewpoint, etc.). WRUs are tuned to words, and feed into NRUs (Name Recognition Units) which are tuned to names. Both FRUs and NRUs can activate (and reciprocally, be activated by) the corresponding PINs (Person Identity Nodes). Activation of a PIN corresponds to an amodal representation of the identity of an individual, and in Burton et al.'s framework, it is at this level that the feeling of familiarity arises. PINs are a sort of hub, mediating the retrieval of semantic information from sensory input (faces, names). Semantic information is stored in a pool of SIUs (Semantic Information Units); everything that is known about a given individual is stored in the SIUs, such as their for profession, nationality, whether they like strawberries, and their name (in this specific model; other models model the same as part of a separate pool). Note the reciprocal connections between SIUs and PINs; the feedback connections from SIUs to PINs are thought to mediate some semantic priming effects. Reproduced from [138].

What evidence does this model rely on? I'll go through a few of the psychophysical and lesion results here, to check that the model stands to good reason (for a review, refer to [139]). The

backbone of the model relies on the apparently sequential access to different types of information when a familiar face is encountered: 1) the face is recognized as familiar; 2) semantic information is retrieved about that person; 3) the name of the person is retrieved. This sequence has been evidenced from the types of errors made by people when recognizing faces, in diary studies (e.g., [140]) and in controlled laboratory experiments (e.g., [141]). For instance, it never happened that a participant could name a particular face without giving their occupation (or some other semantic attribute). In reaction time experiments, familiarity judgments are faster than occupation-based judgments, themselves usually faster than name-based judgments, in settings where the difficulty of the task is equated inasmuch as possible[142]. Additional evidence comes in the form of case studies of neuropsychological impairments: 1) patients suffering from prosopagnosia do not recognize known faces as familiar[143]; 2) some patients may experience known faces as familiar only, without any further semantic information[144]; 3) patients suffering from anomia have no trouble accessing semantic information, however they cannot retrieve the known person's name[145]. These patients also show an interesting pattern when they are presented with names rather than faces. In cases 1 and 3, patients have no trouble recognizing familiar names and retrieving semantic information. However, case 2 also showed an inability to retrieve semantic information from names, beyond the mere familiarity judgment. These observations are a motivation for the existence of a domain-independent semantic system, as described in the model (**Figure 46**). Studies of (conscious) repetition and semantic priming guided other features of the model. The reciprocal connections between PINs and SIUs and the shared SIUs for two familiar persons who share semantic information lead to semantic priming effects, which do not survive intervening presentations of unrelated familiar people because of within-pool inhibitory links. The strengthening of links between FRUs (or NRUs) and PINs by Hebbian learning can explain modality-specific, long-lasting repetition priming effects. The model accounts nicely for all these

results[6], and offers an interesting framework for thinking about person recognition. But what do we know about how the brain handles person recognition?

2. A set of brain areas concerned with face recognition

Many lesion and imaging studies have helped define and refine the set of brain regions that are involved in naming people and retrieving conceptual knowledge (about people, but also other object categories); for example, see [146–150]. These studies give some clues as to where in the brain the units hypothesized by Burton and Bruce may lie (e.g., the left anterior temporal lobe seems to play a key role in retrieving a person's name).

The question of familiar person recognition was reviewed by Gobbini and Haxby in the 2011 *Oxford Handbook of Face Perception*[151], which I shall try to summarize here briefly. In his classic model[152], updated a few times since (the current version is depicted in **Figure 47**), Haxby identifies some brain regions that are routinely found responding more to faces than to other objects as the core system for face perception: the Occipital Face Area (OFA), the Fusiform Face Area (FFA) and the posterior Superior Temporal Sulcus (pSTS). Note that he also includes other nearby regions which respond significantly to faces, however not maximally, as part of the core system; while these regions are not face selective, they have been shown to contribute to face perception to some extent[55]. Haxby hypothesized that there was an anatomical segregation between circuits dealing with invariant aspects for identification (e.g., the FFA), and changeable aspects for the recognition of emotions or eye gaze (e.g., the pSTS). The recognition of familiar faces may modulate activity in the core system (there has been much debate around FFA's role in coding identity), but it is mostly handled by modules in the extended system for Face Perception. Comparing faces with different levels of familiarity, a distributed set of areas shows up in neuroimaging studies. The medial prefrontal cortex and temporo-parietal junction are more active

[6] One cannot help but notice that most of the experimental evidence comes from the work of one research group… This is a bit disconcerting, but we will just have to go along with it for now.

for personally familiar faces than famous (or unknown) faces[153]. The posterior cingulate cortex and precuneus are activated by faces that are visually familiar, regardless of the amount of semantic information associated with them[154]. Amygdala and insula can also be modulated by the familiarity of faces; the amygdala generally responds more to faces of strangers, but in mothers it responds more to the face of their own child[151].

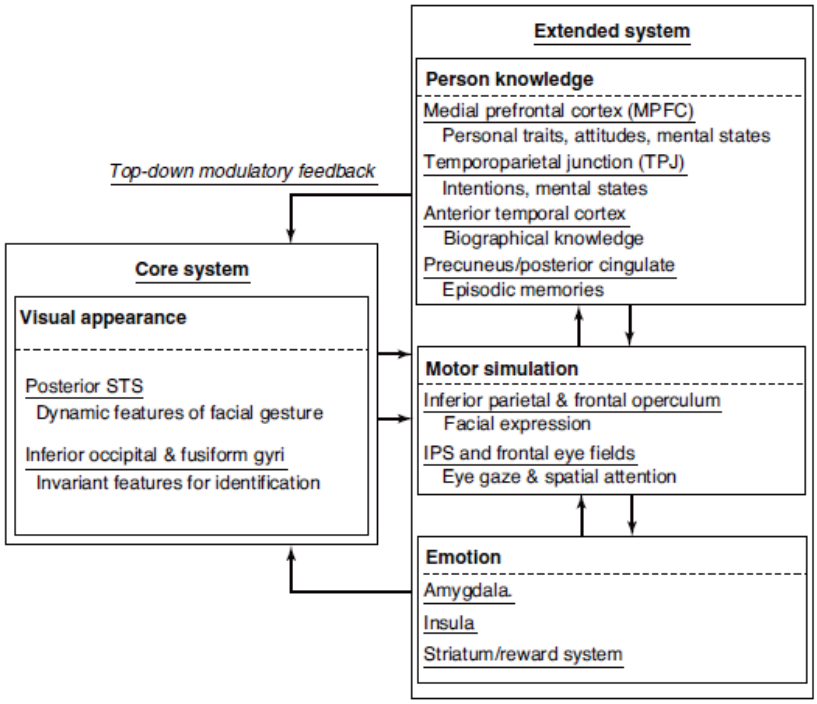


Figure 47 The Haxby & Gobbini cognitive neuroscience model for familiar face perception. This model divides brain areas that are involved in face perception into a Core System—occipitotemporal visual extrastriate areas that play a central role in the visual analysis of faces—and an Extended System—neural systems whose functions are not primarily visual but play critical roles in extracting information from faces. In the Core System, the authors emphasize a distinction between representation of invariant features that are critical for recognizing facial identity and representation of changeable features that are critical for facial gestures, such as expressions and eye gaze. They emphasize three sets of brain areas in the Extended System that are involved, respectively, in the representation of person knowledge, in action understanding (including gaze and attention), and in emotion. Familiar face recognition involves visual codes for familiar individuals in Core System areas in the fusiform, and possibly anterior temporal, cortex, along with the automatic activation of person knowledge and emotional responses. Facial expression involves visual codes in the STS, along with activation of representations of emotion and motor programs for producing expressions. Perception of eye gaze similarly involves visual codes in the STS, along with activation of brain areas for shifting attention and oculomotor control. Reproduced from [151].

That is, surprisingly, roughly the extent of our current neuroscientific understanding of familiar face perception; arguably, we do not understand very much at all, and the cognitive psychology

model is rather more pleasing as it provides a better account of mechanisms. The Haxby model is not incompatible with the Burton model presented in the previous section; the former is simply not focused on the possibility of other input modalities for person recognition, hence the question of the existence of a modality-free representation of person identity does not arise.

I have myself run some simple block-design fMRI experiments geared towards finding brain regions that are involved in recognizing familiar people. The simplest experiment one can think of is to contrast blocks presenting a sequence of pictures of famous people vs. block presenting a sequence of pictures of unknown people. I performed this basic experiment on a group of 19 subjects, as a twist to a standard faces vs. scenes localizer (**Figure 48**). The task was a basic one-back task for image repetitions. The univariate whole-brain results at the group level are shown in **Figure 49**. The precuneus and the medial orbitofrontal cortex were more active when seeing

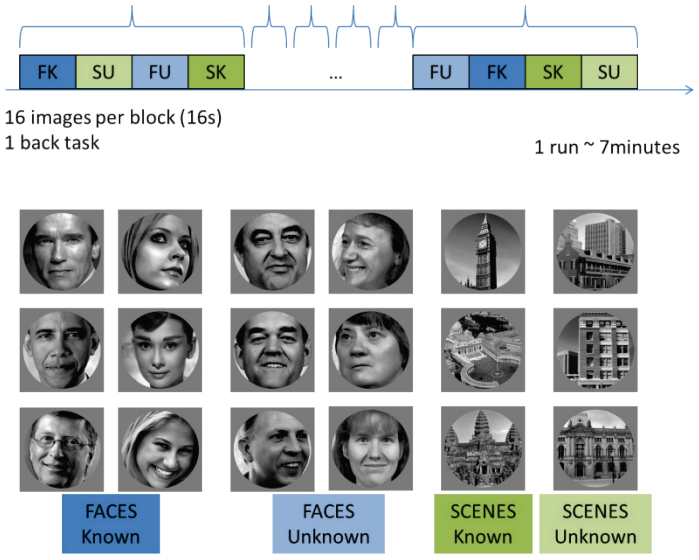


Figure 48 Design of our faces/scenes by famous/unknown fMRI paradigm for functional localization of face responsive areas and familiarity responsive areas. There were 24 blocks of 16 seconds, belonging to four conditions, which were shown in a pseudorandom order (top): FK (faces known), FU (faces unknown), SK (scenes known) and SU (scenes unknown). Within each block, a sequence of 16 images was shown; subjects performed a simple one-back memory task.

famous faces than when seeing unknown faces; additionally, the bilateral FFA and OFA showed enhanced activation as well. It is difficult to make much out of this simple paradigm; a host of processes are going on when subjects see a person they know, especially if they do not have a very engaging task to perform. The set of areas that I found are likely to be related to person recognition in some way, but further investigation and controls are needed to qualify this statement. I do not wish to attempt any sort of reverse inference here and go beyond what the data has to offer.

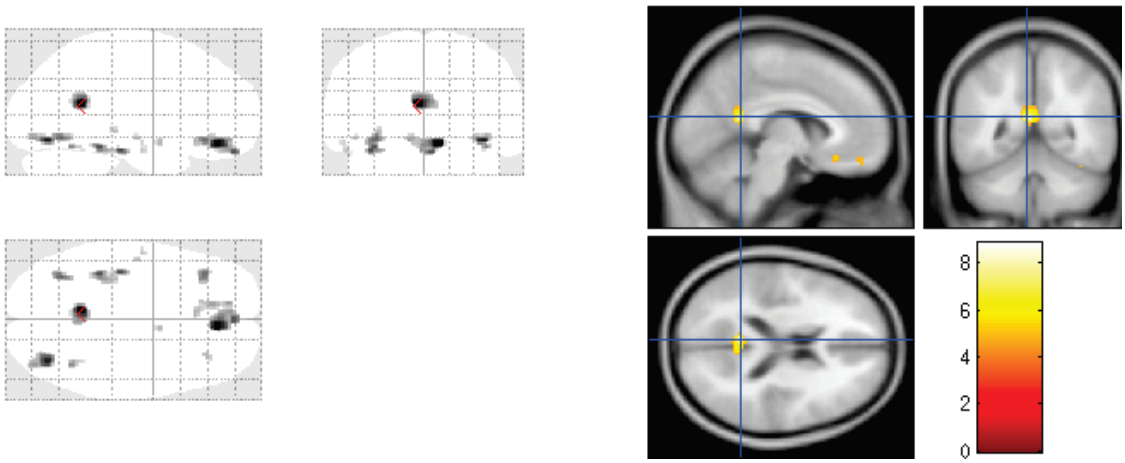


Figure 49 Whole brain results on a group of 19 subjects for the contrast faces known vs. faces unknown, shown on the glass brain (left) and on an orthographic projection (right) centered on the precuneus. Statistical map thresholded at $p<0.05$ (FDR corrected), with a minimum cluster size five voxels (using xjview).

I did design a more controlled experiment, in which I manipulated the task and presented both pictures and names of people (known, or unknown) to get a clearer picture of the network of areas that are engaged in person recognition, across modalities. Again, I used a block design experiment. Before each 16 second block, subjects were given a written instruction on the screen for the task that they were to perform. They had a single button to press, but that button meant something different for each block. There were three task categories: a gender task, a name task, and a profession task. There were four stimulus categories: famous faces, famous names, unknown faces, and unknown names. One run consisted in 24 blocks, distributed as described in

Figure 50. Each block consisted of eight stimuli. In the leftmost example block, in **Figure 50**, subjects had to press the button if the first initial (first letter of the first name) of the pictured famous person was between A and M. This task required name retrieval, whereas a gender task (press the button when the pictured famous person is a male) did not.

STIM / TASK	Famous Faces FF	Famous Names FN	Unknown Faces UF	Unknown Names UN
Gender (Male/Female)	3 blocks	3 blocks	3 blocks	3 blocks
Profession (Actor/Singer)	3 blocks	3 blocks		
First Initial (A->M/N->Z)	3 blocks	3 blocks		

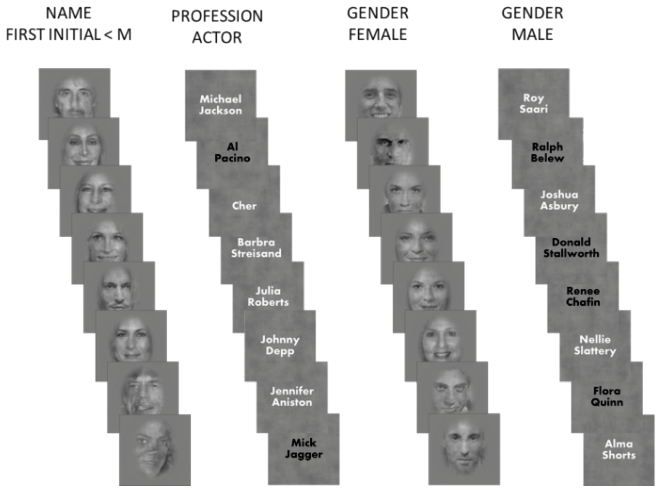


Figure 50 Design of the person identity network experiment. Top: table representing the number of blocks for each condition, in one run. Bottom: example blocks (the names written in black or white are an irrelevant feature). In a given 16 second block, a sequence of eight images was presented; in the leftmost example, the task is to press the button (only one button is given to the subject) whenever the first initial of the individual shown on the screen is between A and L (included). Hence, the perfect answer in this case would be, as represented by a binary vector: 1 1 1 1 1 1 0 0 [Al Pacino/Cher/Barbra Streisand/Julia Roberts/Johnny Depp/Jennifer Aniston/Mick Jagger/Michael Jackson].

A number of contrasts can be performed using this design, which are potentially informative with respect to the brain regions that play a role in various aspects of person recognition; some examples are: famous faces with a gender task vs. unknown faces with a gender task, to evidence regions containing Face Recognition Units or Person Identity Nodes; famous names with a gender task vs. unknown names with a gender task, to evidence regions containing Name

Recognition Units or Person Identity Nodes; but also famous faces with a profession task vs. famous faces with a gender task, which should involve Person Identity Nodes and Semantic Information Units; famous faces with a name task vs. famous faces with a gender task, which again should involve PINs and SIUs; etc.

I only ran this experiment on one subject (over six sessions, in the context of my supersubject experiment, cf. page 121). One region was identified multiple times, as being more active for famous people than unknown people (with a gender task), and also more active when performing a task on the name than when judging the gender of a famous face, in the posterior part of the left middle temporal gyrus (**Figure 51**). I also found evidence for enhanced activity in the medial prefrontal cortex when subjects performed a task on the occupation vs. the gender of a famous person. I think that the paradigm is promising and given the opportunity, one could run it in other subjects to investigate group-level effects.

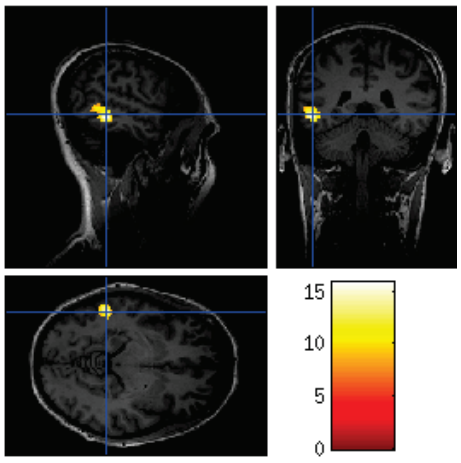


Figure 51 The posterior left middle temporal gyrus is more active when performing a task on the name of a famous person than judging their gender (from a picture). The SPM is shown on the anatomy of the only subject on whom we ran this experiment, and was thresholded aggressively here ($p<10^{-15}$, i.e. $T>8$) for a clean figure.

3. In-house favorite hypothesis: Jennifer Aniston neurons as Person Identity Nodes

In the Koch laboratory we are quite partial to the existence of a modality-free representation of person identity in the human brain. Christof has had the opportunity to work for over 10 years with surgeon Itzhak Fried at UCLA. Dr. Fried's specialty is in surgical treatment of epilepsy. Epilepsy is a common disease, affecting some 50 million people worldwide. Epilepsy is a chronic disease which consists of recurrent seizures, spurts of abnormal electrical activity in regions of brain, which may spread to the whole brain (generalized seizures are sometimes referred to as "grand mal"). Anticonvulsant drugs are usually enough to keep the condition under control and the only notable handicaps for most people who suffer from epilepsy is that they are not allowed to drive cars, and may hurt themselves involuntarily during seizures. However, in some cases medication is not enough to keep the seizures in check; in these cases, surgery may be an option, if the seizure can be traced back to a well-defined, dysfunctional piece of brain tissue. One of the most common refractory forms of epilepsy is medial temporal lobe epilepsy. At UCLA, patients who suffer from intractable epilepsy and are candidates for surgery are implanted with (up to 12) electrodes, which have six to seven contacts approximately 1.5mm wide with separations of 1.5-4mm. These contacts allow 24h monitoring of EEG data, and are used for triangulating the location of the epileptic focus. The monitoring period lasts as long as is necessary to record enough spontaneous seizures, typically between seven and ten days. The clinical electrodes are hollow, and Dr. Fried introduces microwires through their lumen, typically nine microwires, which spread apart rather randomly once they exit the macroelectrode shaft (they extend one to three millimiters from the end of the macroelectrode, as represented in **Figure 52**). These microwires can pick up the activity of single neurons, extracellularly. This is an incredible opportunity to study invasively the information represented in single neurons in the medial temporal lobes in the human brain. Patients who are rather bored as they wait for more than a week for seizures to occur are usually happy to take part to various experiments which

researchers have designed, while the activity of single neurons in their amygdalae, hippocampi, entorhinal and parahippocampal cortices is monitored (see **Figure 53**). Many more details can be found in Gabriel Kreiman's PhD thesis[155]; Gabriel had a central role in setting up the collaboration with Dr. Fried.

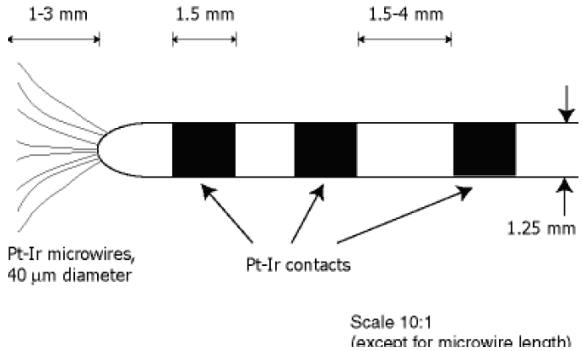


Figure 52 Schematic description of the type of electrodes most often used at UCLA in temporal lobe targets. Platinum/iridium contacts of approximately 1.5 mm length along the electrode are used to acquire clinical wide band EEG data. Through the lumen of the 1.25 mm diameter electrodes, 8 platinum/iridium microwires are inserted. Electrodes are fabricated at UCLA. Microwires extend 1 to 3 mm from the tip of the electrode, lying inside a cone with an opening angle of less than 45°. Reproduced from [155].

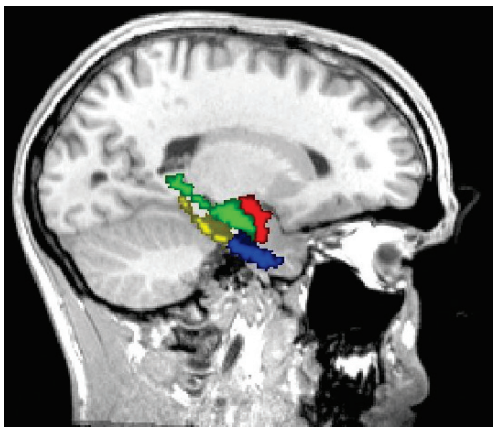


Figure 53 My brain with the regions usually targeted at UCLA color-labeled (Freesurfer recon-all). Luckily this is just for show, and I did not have to undergo the surgical procedure. Better yet, this picture was published in one of Christof's Scientific American Mind columns[156], hence this particular sagittal section of my brain is forever famous. (Red, amygdala; green, hippocampus; blue, entorhinal cortex; yellow, parahippocampal cortex).

One of the most compelling findings that arose from this collaboration is often referred to as the "Jennifer Aniston neuron". The original paper[157] that described the finding was authored by Rodrigo Quian-Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch and Itzhak Fried, and such neurons are still routinely found at UCLA (e.g., [158]). Note that there has not yet been any replication that we are aware of in another lab; however, very few labs in the world perform such recordings, and those that do may not be interested in replicating the finding.

Simply put, though, we find neurons in the medial temporal lobe that respond in an invariant manner to a familiar person (either personally familiar or famous), meaning that the neurons fire action potentials above their baseline rate when the patient sees any picture of the person, sees a cartoon of them, reads their name, or even hears their name. An example of one such cell is shown in **Figure 54**.

What are these neurons doing? What is their place in the person recognition network, if any? There is much debate around this question, but it is tempting to think of them as the neural substrates of Person Identity Nodes, referring back to Burton & Bruce's model (**Figure 46**).

With this background in mind (a bit less organized, admittedly), I decided to ask a very simple first question to start my venture into the problem of person recognition: could I find, with fMRI, evidence for invariant representations of famous people or characters? The reason I wanted to use fMRI is that it can be run time and time again in normal subjects, hence allowing to ask many different questions. This is not the case with single unit recordings at UCLA, for which the number of patients is limited, and the time with each patient is also limited. In the following section, I describe two pilot experiments that I ran and learned from.

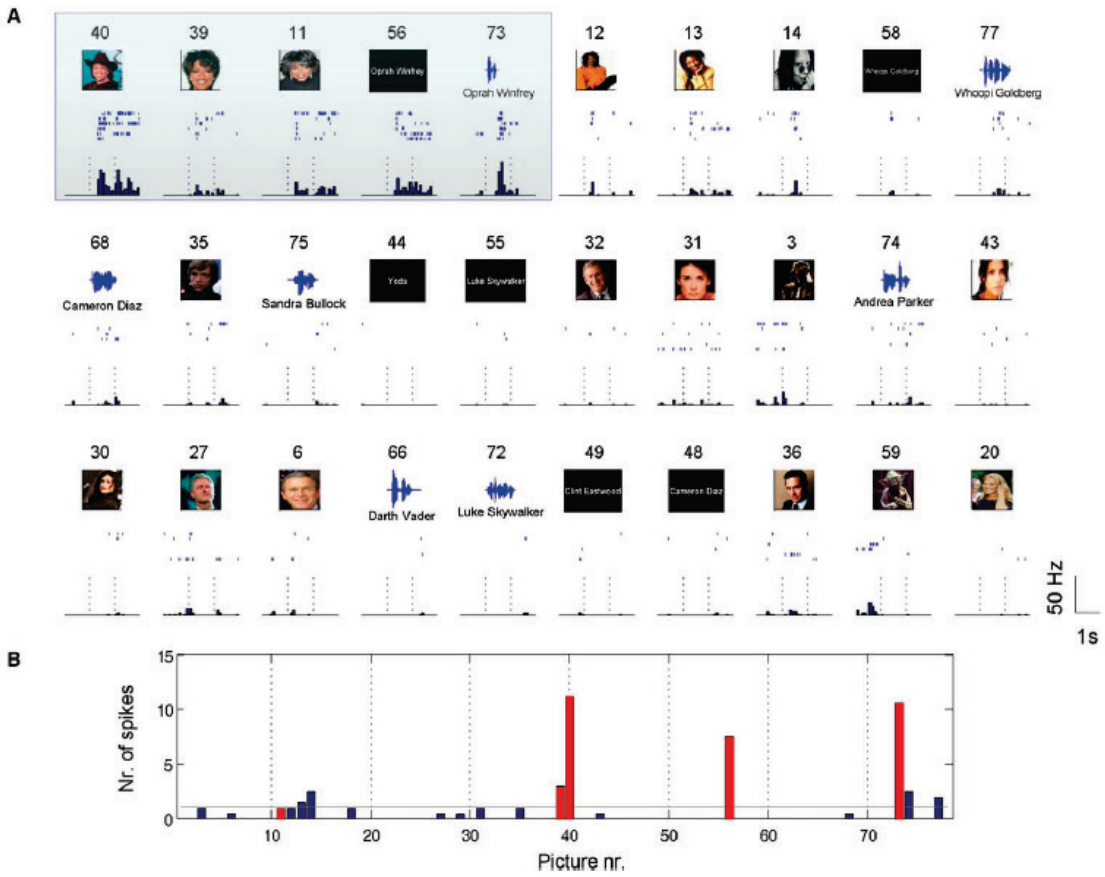


Figure 54 A single neuron responding selectively to Oprah Winfrey. (A) A neuron in the hippocampus that responded selectively to pictures of the television host Oprah Winfrey (stimulus 40, 39, and 11), as well as to her written (stimulus 56) and spoken (stimulus 73) name. To a lesser degree, the neuron also fired to Whoopi Goldberg. They were no responses to any other picture, sound, or text presentations. For space reasons, only the largest 30 (out of 78) responses are displayed. In each case the raster plots for the six trials, peristimulus time histograms (PSTH) and the corresponding pictures are shown. The vertical dotted lines mark picture onset and offset, one second apart. (B) Median number of spikes (across trials) for all stimuli. Presentations of Oprah Winfrey are marked with red bars. Stimulus numbers corresponds to the ones shown above each picture in (A). The gray horizontal line shows the five standard deviations above the baseline threshold used for defining significant responses Reproduced from [158]

B. Pilot experiments

1. Movie clips of four familiar characters

At the time when I was designing these experiments, it so happened that Christine and I had succumbed, in turn, to two TV series: the enigmatic *Lost*, and the explosive *24*. I chose two characters from each of those series: Jack Shepard and John Locke from *Lost*, and David Palmer and Jack Bauer from *24* (**Figure 55**). My reasoning was that they were all clearly different

individuals about whom Christine and I had lots of semantic information from watching many episodes; I should thus be able to individuate them from our brain activities, since they clearly were different in our minds. Also, as a fallback, in the event that I would not be able to tell these individuals apart from brain activity, I thought that I might be able to tell apart characters that came from the *Lost* world and characters that came from the *24* world; this was such obvious information subjectively that there surely would be an associated difference in brain activity.

Figure 55 Four characters for a first pilot experiment. Top: David Palmer, Jack Bauer (from *24*). Bottom: John Locke, Jack Shepard (from *Lost*).

I collected many two-second videos clips exclusively featuring one of these characters (approximately 100 clips for each character). I converted these to grayscale, and used a histogram matching procedure for each frame to remove obvious low level confounds, to some degree. I used a block design, with 16 second blocks. There were four block types, corresponding to the four characters. Each block was a succession of seven silent, grayscale movie clips. There were 20 blocks per run, plus four null blocks (structure of a run: [1234]0[1234]0[1234]0[1234]0[1234], where [] means random ordering). The task of the subject in the scanner was a simple 1-back task: if a video clip was seen twice in a row, the subject was to press a pre-assigned button. Scanning was done in a Siemens Tim Trio MRI scanner, outfitted with a brand new 32-channel head coil. 38 axial slices were acquired per volume, with a TR of 2500ms, a TE of 30ms, and 2mm isotropic voxels. In total, I acquired seven runs on Christine and ten runs on myself.

I performed decoding on this data, training a linear SVM classifier to discriminate the four characters on examples from all but one run and testing on the remaining run, and repeating this by using each run in turn as the test set (leave-one-run-out cross validation). Because the interstimulus interval between blocks was short, responses could not go back to baseline between successive blocks and I used a General Linear Model, modeling each block separately, to deconvolve the responses. The beta values were divided by their standard deviation before being assembled in a big matrix for decoding. I ran a searchlight throughout the brain to find areas informative with respect to identity. I soon found that there were many informative areas, and that decoding was extremely good in early visual areas in the occipital lobe (**Figure 56**), obviously

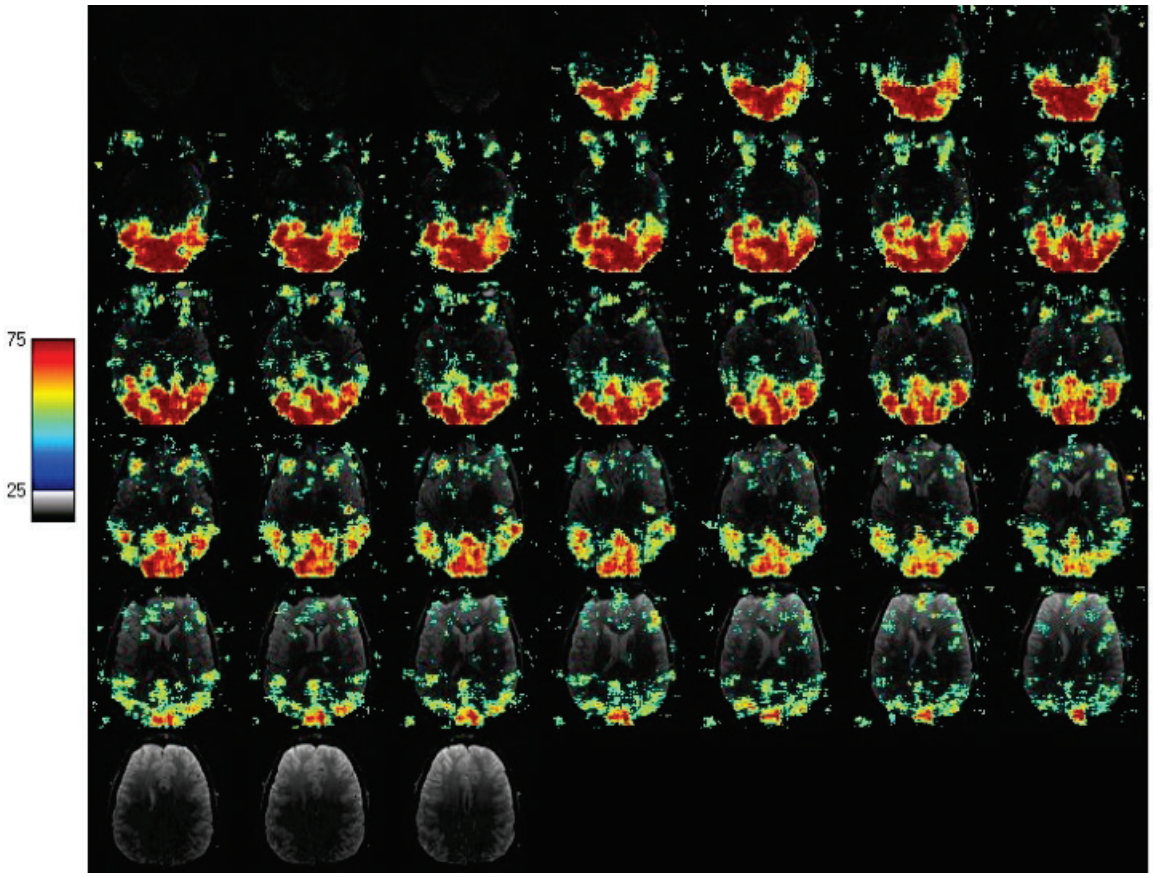


Figure 56 Decoding of identity in early visual cortex, in one subject. The performance of the classifier for each searchlight is overlaid on the EPI (sub-axial) slices, which are arranged from bottom to top; the left side correspond to the left side of the brain. Chance is 25%. The accuracy map was thresholded using FDR, with a threshold at 0.05.

pointing to the predominance of low level confounds, despite my (half-hearted?) attempt at equalizing the luminance histograms of the video clips. There are additional ways to control for low level confounds; one can for instance try to equalize the power spectra. In the case of movie clips, another low-level property that stands out in hindsight is the amount of motion in the clips; some characters have a tendency to move much more than others. I did not look too hard into quantifying and equalizing motion, though. For my purpose of finding regions with an invariant representation of person identity, this pilot experiment was very instructive: I needed to look at fMRI decoding across modalities, to be sure of the absence of low-level confounds.

2. Audio and movie clips of four familiar cartoon characters

There is something captivating about cartoon characters, and the concepts that they elicit are quite rich. Nerdy college kids are especially fond of them. Some of them may be able to write a ten-page biography of their favorite characters. Whether this is healthy is not my main concern here.

In order to find subjects who were very familiar with cartoon characters, and thus increase the likelihood of those characters eliciting a very powerful conceptual representation in these subjects, I conducted a survey in Christof's CNS120 class (which I TA-ed for a few years). I chose four cartoon series that Google thought were among the most popular in the U.S.A.: *Futurama*, *South Park*, *The Simpsons*, and *Family Guy*. I then sent an email survey to the class, and also asked them to forward it to friends at Caltech if they so desired. Aside from asking the surveyees for their age and gender, the main question was:

> Rate your FAMILIARITY with the following cartoons, from 0 to 4 stars.
>
> 0 star: you have never heard of it
>
> 1 star: you have heard of it but never watched it

2 stars: you have watched a few episodes of it, here and there, and know more or less what the show is about

3 stars: you have watched many episodes and are fairly well acquainted with the characters

4 stars: you have watched (almost) all episodes religiously and could write a ten-page essay about each of the main characters

The first surprise was how quickly students responded to this survey; I got 169 responses within 24 hours! The responses were almost evenly distributed between males and females. This attests to the power that cartoons exert on the Caltech student population. I looked specifically at whether there were students who gave a 4-star rating to more than one cartoon series. Being a fan of one series is probably fairly common, but being very knowledgeable about more than one series seemed a more unlikely situation. I represented the numbers of students who gave a 4-star rating to at least one of the cartoon series in a Venn diagram (for a fun visualization, see **Figure 57**). I was struck by finding that a sizeable number of subjects had given 4-star ratings to many of the cartoons. Four subjects actually gave 4-star ratings to all four cartoons! 11 subjects rated both *The Simpsons* and *Family Guy* with 4-stars. I thus had a very nice database of subjects to choose from for my experiments.

I decided to start with subjects who were highly familiar with both *The Simpsons* and *Family Guy*; the reason was that I could then pick the father and the son in both series (Peter/Stewie Griffin and Homer/Bart Simpson), which gave me a 2x2 design (father/son x Simpsons/Family Guy, see **Figure 58**). I thus set out to find where invariant representations of these four characters may lie (with the possibility of falling back to the concepts of father/son or Simpsons/Family Guy).

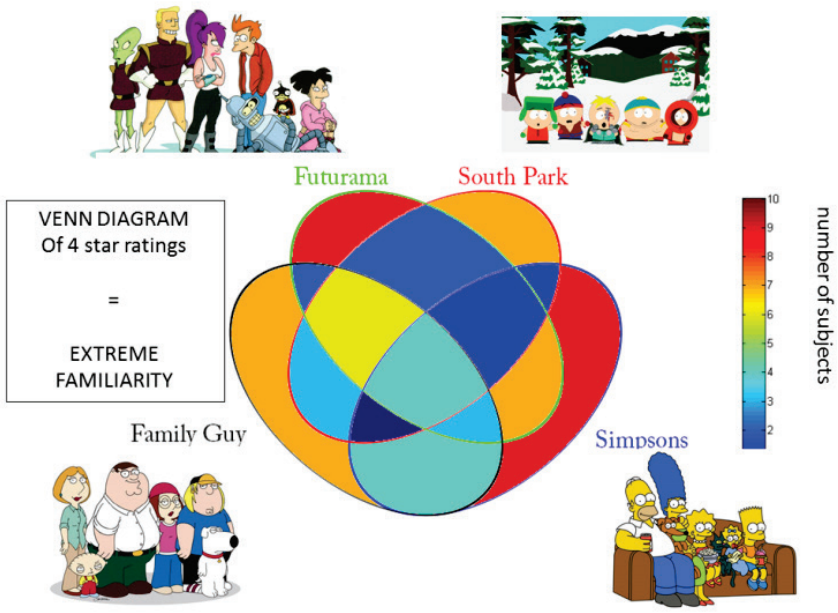


Figure 57 Caltech students love their cartoons. I performed a survey amongst Caltech undergraduates, asking them to rate from 1 to 4 their level of familiarity with each of four cartoons: Futurama, South Park, Family Guy, and the Simpsons. Here, I plot only the number of "4" responses ("4" meaning "you have watched (almost) all episodes religiously and could write a ten-page essay about each of the main characters"); out of 169 surveyees, a large number used the rating "4", and quite a few used it for two or more cartoons!

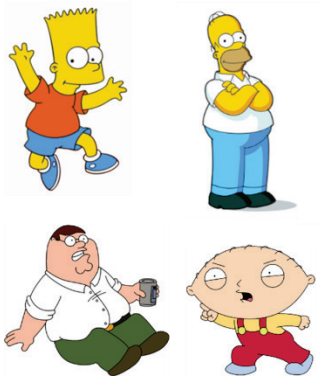


Figure 58 Cartoon characters used in the cartoon pilot. Top: Bart and Homer Simpson; Bottom: Peter and Stewie Griffin

As in the previous experiment, I collected video clips of the characters. This time, I collected five-second-long clips (a rather painstaking endeavor), making sure that each clip only featured the character of interest. I kept the audio track, separately, for all the video clips. Cartoon

characters have extremely distinct voices, and I decided to use that feature to investigate invariant representations and be freed from the curse of insidious low-level confounds.

I used a block design, as shown in **Figure 59**. Within a 15 second time window, I would show a five-second video clip and a five-second audio clip, separated by a five-second blank; the audio could lead or precede the video. Both clips were of the same character, but they only matched in half of the trials; in the other half, the audio was taken from a different video clip. The subject's task in the scanner was to determine whether it was a match, and press a button accordingly. These 15 second "blocks" were separated by nine second blank intervals. A run consisted of 20 blocks, which were presented is pseudo-random order ([1 2 3 4][1 2 3 4][1 2 3 4][1 2 3 4][1 2 3 4]).

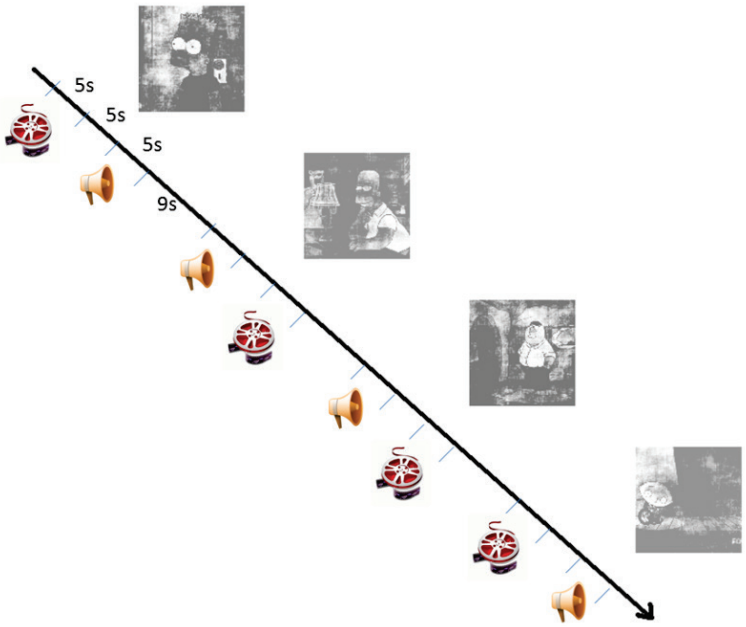


Figure 59 Experimental design for the cartoon pilot. The arrow represents time. A run consists of a succession of 15 second blocks, which themselves consist of a five second clip, a five second blank, then another five second clip. If the first clip is audio, the second is video, and vice-versa. Both clips in a block are of the same character. The task of the subject is to determine whether the audio and the video match. Blocks succeed each other, separated by nine second interblock intervals.

I ran one subject without preprocessing the clips in any way. Then, as in the previous experiment, I tried to reduce low level confounds in the video clips, by histogram matching AND power spectrum matching all frames of all movies (I used the recently published SHINE toolbox[159] developed in Fred Gosselin's lab, which packaged nicely things that I had coded myself in the past). This made the video clips look horrible, but at least it got rid of obvious confounds (**Figure 60**). As for the audio clips, I did not perform any preprocessing. I noticed after the fact that my video preprocessing made the task extremely hard; while the first subject's performance was at 75% (chance being 50%), the second subject was at around 40% (below chance? She likely got the instructions wrong, and was truly at 60%, which is much lower than the first subject).

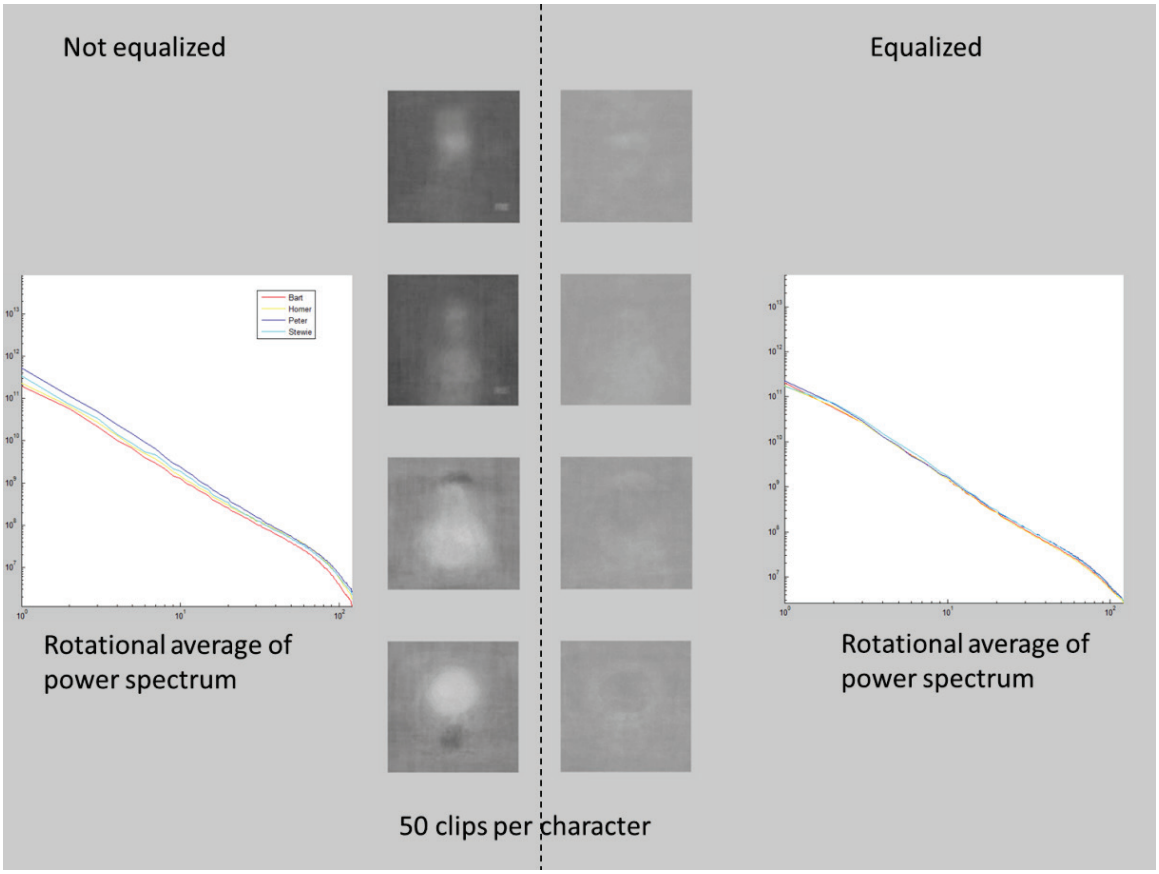


Figure 60 Effects of histogram matching and power spectrum equalization on mean images and power spectrum averages. Left: original clips (the four thumbnails correspond to the average of all frames for Bart, Homer, Peter and Stewie). Right: after equalization, obvious luminance and power-spectrum confounds are gone.

I collected eight runs of the experiment on each of two subjects, in the Siemens Tim Trio outfitted with a 12-channel coil (before that scanner was exclusively dedicated to primate research), using 30 slices and 1.8mm isotropic voxels (TR=2s, TE=30ms). I later repeated the experiment with one of these two subjects, in the other Siemens Tim Trio outfitted with a 32-channel coil, using 34 slices and 2mm isotropic voxels (TR=2s, TE=30ms). As in the previous experiment, I ran a General Linear Model and estimated the regressors for audio and video clips of the four characters. I then ran searchlight decoding throughout the fMRI volume, with a cubic searchlight (7x7x7 voxels). As expected, in the first subject, for whom I had not tried to remove low level confounds at all, I was able to decode the characters' identities from early visual cortex with high accuracy (training and testing on video clips, leave-one-run-out cross-validation). Despite the histogram equalization and power spectrum matching performed for the following subjects, early visual cortex still performed above chance on identity classification (**Figure 61**). I don't think I even managed to make a dent into the classification accuracy with my gimmicks. Interestingly, as well, I was able to decode identity from early auditory cortex, by training and testing on audio clips (**Figure 62**).

These results are comforting in that I was able to retrieve information about low level visual cues where I expected it, and information about low level auditory cues where I expected it. However, they are not too interesting; surely we already know that different sounds elicit different patterns in the early auditory cortex, and that different images elicit different patterns in the early visual cortex. The truly interesting question is whether there is any area of the brain where listening to Bart's voice and seeing Bart elicit similar activity. I attempted this across-modality decoding, training on video clips and testing on audio tracks, and vice versa. I reasoned that areas that truly support cross-modal decoding should also support within modality decoding, and used this principle as an additional warranty against false positives. There was a spot in the temporo-occipital part of the left inferior temporal gyrus where accuracy was weakly above chance for the

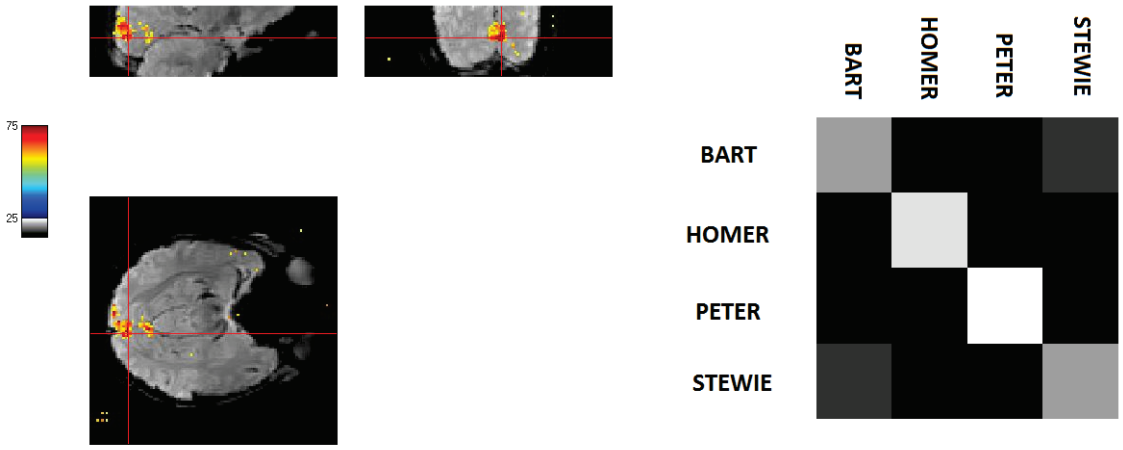


Figure 61 Decoding identity in early visual cortex from videos (after the attempt to remove low-level confounds with power spectrum and histogram matching of all video frames), in a single subject. Left: orthographic projection, centered on maximal accuracy (map is thresholded by FDR at q=0.05). Right: confusion matrix showing which identities seem to be easily discriminated from the others, at the location of the crosshair on the left (overall accuracy is ~75% as seen on the left side; Peter is easy to tell apart from the other characters, while Bart and Stewie are less differentiable at that location).

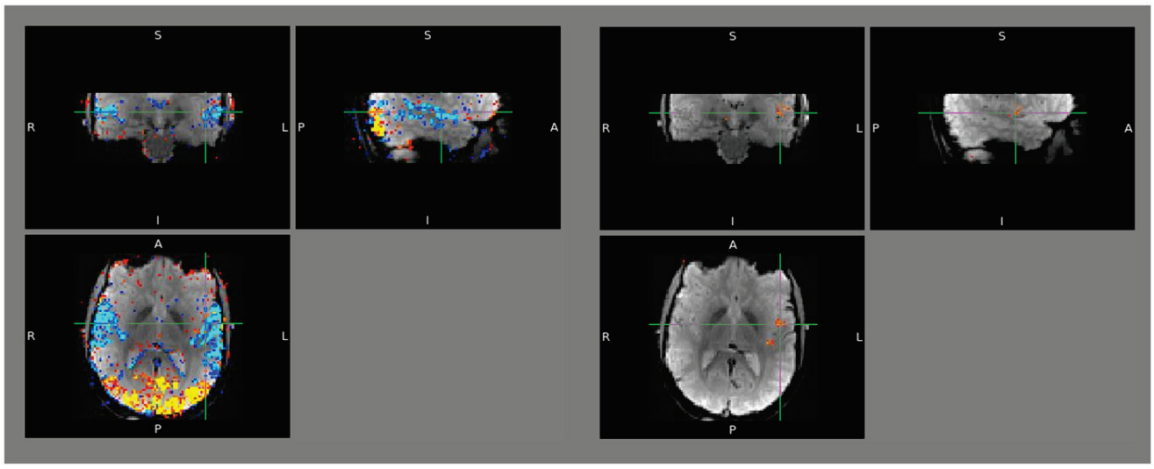


Figure 62 Decoding of identity from audio clips (voices). Left: the contrast video clips vs. audio clips, thresholded at p<0.001, is a good localizer for early visual (red-yellow) and early auditory (blue-light blue) cortices. Right: regions of above chance decoding of the identity from the voice (thresholded with FDR at q=0.05), at a location which corresponds to left early auditory cortex.

four decoding schemes (visual>visual, auditory>auditory, visual>auditory, auditory>visual), but it was so weak that I am pretty convinced it is a false positive, and I shall refrain from crowding this thesis with a figure that could mislead the reader.

In these pilot experiments, I learned a few important things. Firstly it is extremely hard to get rid of low level confounds, since I was able to decode which character was presented above chance from activity in the medial occipital lobe (early visual cortex), even after applying such techniques as power spectrum equalization and histogram matching. This is obviously always going to be a problem for anyone who claims to have found reliable decoding of identity; it is almost impossible to rule out low-level differences. I will come back to this issue in more depth in the following chapter. What I do want to point out here is the fuzzy definition of what one refers to as low level differences. Ultimately, two images that have no low-level differences at all are the same image; there has to be some information at the level of pixels for two images to be different. Hence, I believe that the only question that can be asked, which is impervious to such criticism, is whether there exists a representation of identity that is modality invariant, and the way to find it is to learn patterns in one modality and test whether they generalize to another. You will find that, although I am stating this now, I have been a bit schizophrenic with this assertion. I go back and forth between not caring at all about low-level confounds and caring a great deal too much (and making the stimuli ugly). On this note, I gave up on movie clips in the following; though they are definitely more engaging (as my subjects whom I subjected to hours of seeing still pictures in the scanner could attest), they are also more difficult to control; I have not given it too much thought, but it seems rather difficult to equalize the amount of motion in different movies without completely disrupting crucial information.

Secondly, using as many examples as possible may seem like a good idea to be able to train a classifier properly; however I observed that performance improves dramatically by averaging noisy examples rather than feeding them directly to the classifier. It is my experience that linear Support vector machines are surprisingly robust when faced with the curse of dimensionality. Again, though this has proved true empirically (and has been observed by other colleagues whom

I've shared my experience with), I have been known to forget and try averaging less examples together. I am sure I will never give up trying; however I shall try to stick to this principle inasmuch as possible in this thesis.

> ***It is better to average multiple noisy examples into one reliable example when training a classifier for fMRI decoding.***

Finally, we did not find very strong evidence for patterns that generalize from one modality to another in the second pilot. It could be for multiple reasons, of course, but I decided that it may have stemmed from my subjects not being engaged enough in thinking about the concepts. Arguably, my choices for tasks in the two pilots (a 1-back task for the first, and an audio-video matching task in the second) were suboptimal. This is why, in the main experiment that I describe in the following section, I used tasks that emphasized identification.

C. The full-fledged experiment: Brad Pitt, Matt Damon and Tom Cruise

For this experiment, I wanted to get back to a paradigm closer to what patients at UCLA do, i.e., simply see pictures and written names of famous actors for about one second each (while performing a simple face-non face task).

1. Stage 1: the basic design, a simple identification task, and seven subjects

I chose three good looking, ultra famous, white male actors: Brad Pitt, Matt Damon and Tom Cruise[7]. Most people can recognize these three actors and name several movies that they were in. For simplicity, I wanted to show the same actors to all subjects, rather than customizing the stimulus set for each subject and spending hours online looking at pictures of actors (I actually ended up, at some point, trying to customize the set of actors anyway. which points again to my

[7] Some colleagues have told me after the fact that Matt Damon was not quite a match for Brad Pitt or Tom Cruise, but I do not think that it matters too much.

schizophrenic behavior as a scientist). I collected 50 or so pictures of each actor, of which I ended up using 39. I also generated text, with the written names, in multiple fonts.

In order to have an engaging task that required identification, I switched to an event-related design. Lost in the plethora of recommendations in the literature to achieve maximal efficiency for event-related designs, I simply copied a design that had seemed to work wonders: Kay et al.[160] (the 2008 Nature paper on decoding natural images from V1, in Jack Gallant's group) used a m-sequence[89] of length 13, order 2 for their experiments. I created this m-sequence and filled it with my own stimuli. If it was efficient enough for a Nature paper, perhaps it would serve me well, too. The design is shown schematically in **Figure 63**. Stimulus onsets were synchronized with volume acquisition, and the use of null events effectively created some temporal jitter. There were three times as many picture trials as text trials (this originally stemmed from the fact that I wanted a lot of picture trials to learn from, and then I would test on text trials). The task was to press one of three buttons, at each trial, to identify who was presented (Brad Pitt, Matt Damon or Tom Cruise); the buttons were assigned and stayed the same throughout the experiment.

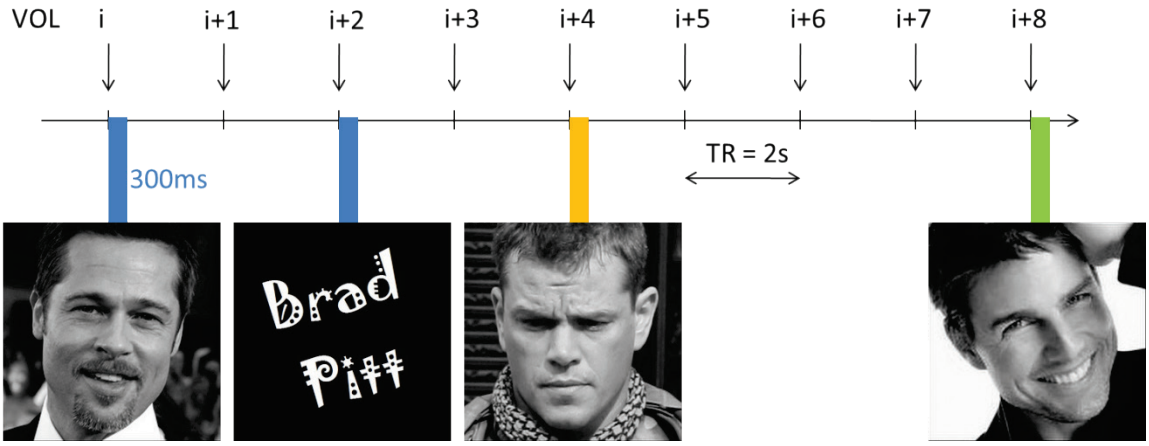


Figure 63 The (initial) design of the Brad Pitt, Matt Damon and Tom Cruise experiment. The arrow from left to right represents time, and is graduated in units of volumes (the repetition time is two seconds). Every other volume (stimulus onset asynchrony is 2xTR, i.e. four seconds), a static stimulus is shown on the screen, either a picture or a written name. There are 169 trials per run, including 13 null trials (nothing presented), 39 picture trials per actor, and 13 name trials per actor. The task of the subject is to press one of three buttons on each trial, corresponding to the identity of the actor that they are seeing. The three buttons were pre-assigned to the three identities and stayed the same throughout the experiment.

I scanned seven subjects in the Siemens Tim Trio with a 32 channel coil. 30 slices were acquired, in interleaved order, with 2mm isotropic voxels (TR 2s, TE 30ms). Slices covered the occipito-temporal lobes, and were roughly parallel to the AC-PC line (**Figure 64**). Note that the correct way might be to have the slices at a fixed angle compared to the B0 field, and make sure to have each subject lie in the scanner with approximately the same head tilt, for consistency between subjects.

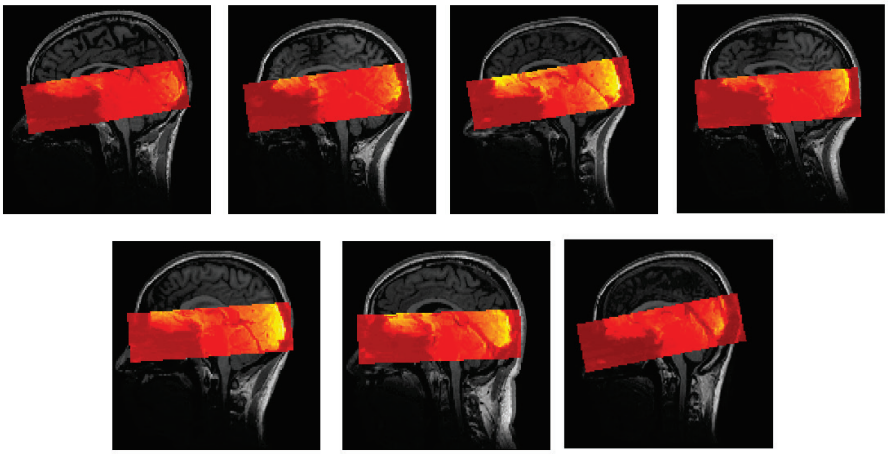


Figure 64 Imaging volume for seven subjects in the first version of the Brad Pitt/Matt Damon/Tom Cruise experiment. A sub-axial slice angle was used, covering the occipital and temporal lobes. The limited coverage is due to a fairly high resolutuon protocol: with a repetition time of two seconds (appropriate in a fast event-related design), 30 slices could be acquired at the 2mm isotropic resolution that we programmed (with the use of parallel imaging GRAPPA, and an acceleration factor of 2).

2. Stage 2: same design, one-back task on identity, 16 good subjects

I realized after analyzing the first seven subjects that my identification task created a motor confound. Indeed, if I could read out from fMRI signals which finger the subject was moving, I would be able to learn patterns that would generalize across modalities (pictures to names). And I found that I could read this information in the ipsilateral cerebellum (all subjects used their right hand) (red circles on **Figure 65**). Note that the accuracy was better when training was performed on pictures; this is because there were more trials with pictures (three times as many as trials with

names), hence the classifier had more data to capture relevant information. Another interesting feature in **Figure 65** is the presence of a spot of above chance decoding in the left anterior temporal lobe; this was a very exciting preliminary finding, given the involvement of the left anterior temporal lobe in disorders such as semantic dementia[147]. However, the threshold is VERY lenient ($p<0.01$, uncorrected), hence I should not elaborate on this any further (I did present this finding at the 2012 VSS meeting). The only way to see if these trends might hold was to run more subjects.

I changed the task for the next subjects. I used a one-back task on identity: subjects had just one button to press, and they would press it each time there was a repeat of identity from one trial to the next. This fixed the motor confound problem. I ran 19 new subjects. Note that the motor confound that I described previously is not really a huge problem, apart from leading to above chance decoding in motor regions. The regions that truly encode identity, across modalities, should still show above chance decoding, even if there is a motor confound. If we do find something, then we can worry about controls. This is another lesson I learned, though I have also forgotten about it countless times:

> ***it is not worth spending too much time thinking of every possible control experiment, if you do not yet have a result.***

I initially analyzed this new dataset separately because I wanted to see if any of the previously highlighted areas would show up in this completely independent set of subjects. This was my attempt at self-replication: if I get a similar result twice from two independent samples of subjects, it becomes quite convincing. There has been much press lately about problems in replicating scientific results[128,161,162], especially in our field; seminal papers, whose results are accepted by the community and inspire new research projects, sometimes may resist replication attempts. All scientists know, in the bottom of their hearts, why this is happening: most imaging studies are terribly underpowered[128]; and statistical tests are, more often than not, used abusively

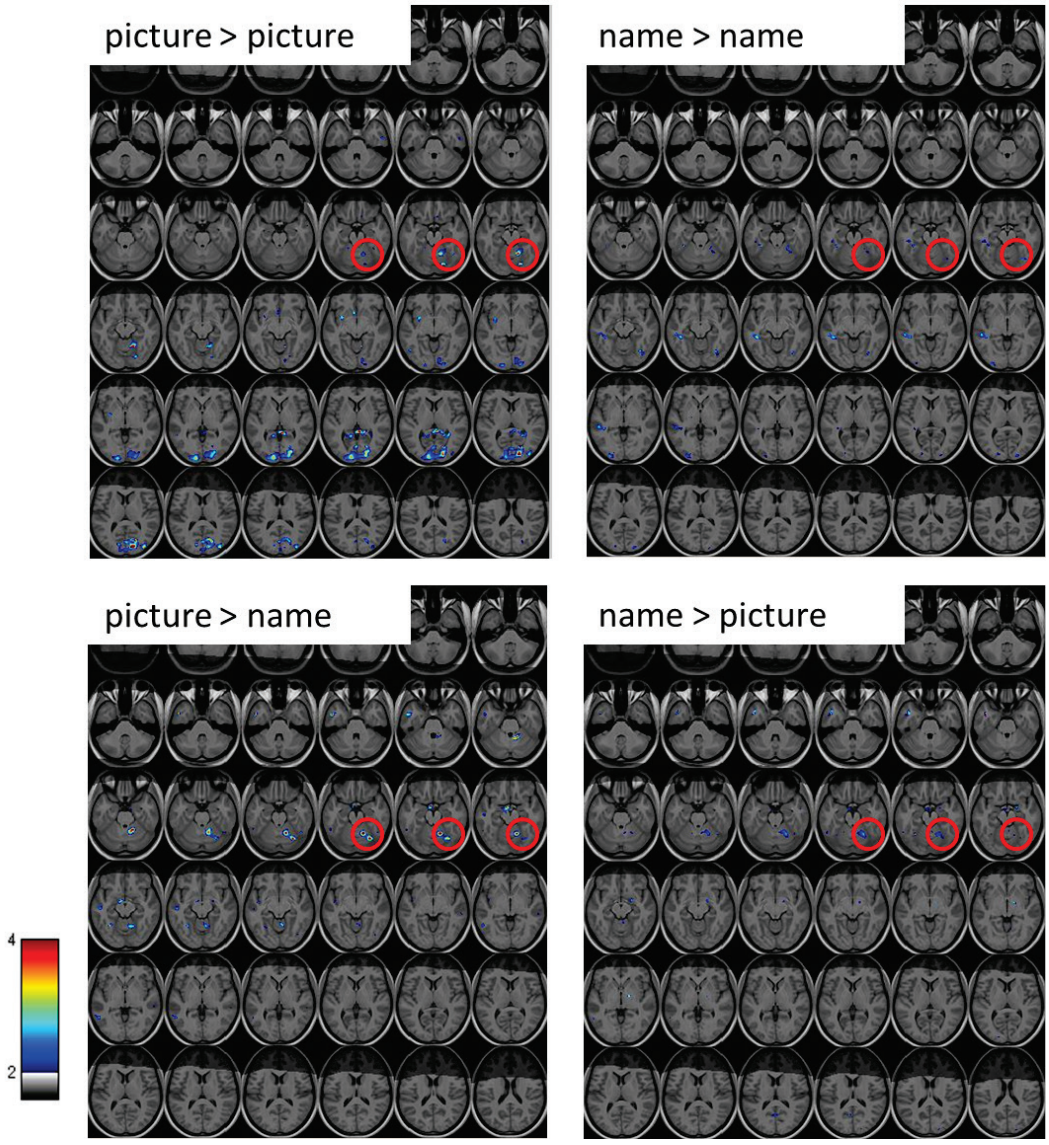


Figure 65 Searchlight decoding, represented as -log10(p) for a t-test across the seven subjects that performed the first version of the Brad Pitt/Matt Damon/Tom Cruise experiment, i.e. with an explicit identification task using three buttons. Note the above chance decoding in the right cerebellum, most likely due to a motor confound (decoding of finger movement). The notation "picture > name" means that the classifier is trained on (averages of) picture trials, and tested on (averages of) name trials. Results are shown on axial MNI slices, ordered from ventral to dorsal. Areas of the brain that are not covered in all subjects are shaded.

to confirm the researcher's pet hypothesis. If one statistical test (the one that is the most appropriate for the situation, assuming that the researcher knows anything about statistics, which is unfortunately a rather optimistic assumption) does not yield a significant result, most researchers will go down one of several possible murky paths: 1) acquire more and more data

until the test reaches significance, and stop as soon as it happens; 2) analyze the data in different ways until some statistic reaches a significant value; 3) tweak the test slightly, reject such and such "outlier" subject, until the test reaches significance. I have certainly been advised to, tempted to, and sometimes downright guilty of going down each of these paths (remember the Nine Circles of Scientific Hell[109], shown in **Figure 45**?). But in our field (cognitive psychology/cognitive neuroscience), this is standard procedure, and sadly any resistance may lead to a failed career. I dream of science that does not need statistics and results that are so clear that they do not need any hypocritic confirmation by a statistical test. This is why Nobel Prize winner Ernest Rutherford's quote resonates so well with me:

> "If your experiment needs statistics, you ought to have done a better experiment."

Let us get back to reality now, after this short rant. What did the maps look like, averaged over the new set of subjects? I discarded some subjects on the basis of their motion parameters (one of the subjects moved more than two centimeters and I lost the anterior temporal lobes!). Also, some of the remaining subjects did not show good "picture > picture" decoding in the occipital cortex, which I decided to be reason enough to exclude them. I thus kept a total of 13 subjects out of the 19. The p-values for a t-test against chance level are shown in **Figure 66**. As you can see, there was not much more than previously in terms of cross-modal decoding; unfortunately, the spot that we saw in the first set of seven subjects in the left anterior temporal lobe did not show up in this new group of subjects.

I literally spent *months* on this data, turning it in all possible directions. Looking back I feel like it was a waste of my time, but perhaps a better way to look at it is that this was a way for me to further build my expertise. I must have preprocessed the raw EPIs at least five times. Perhaps the motion correction could be improved? Perhaps slice timing correction would help? Maybe I should not include the motion parameters in the General Linear Model? Maybe I should not

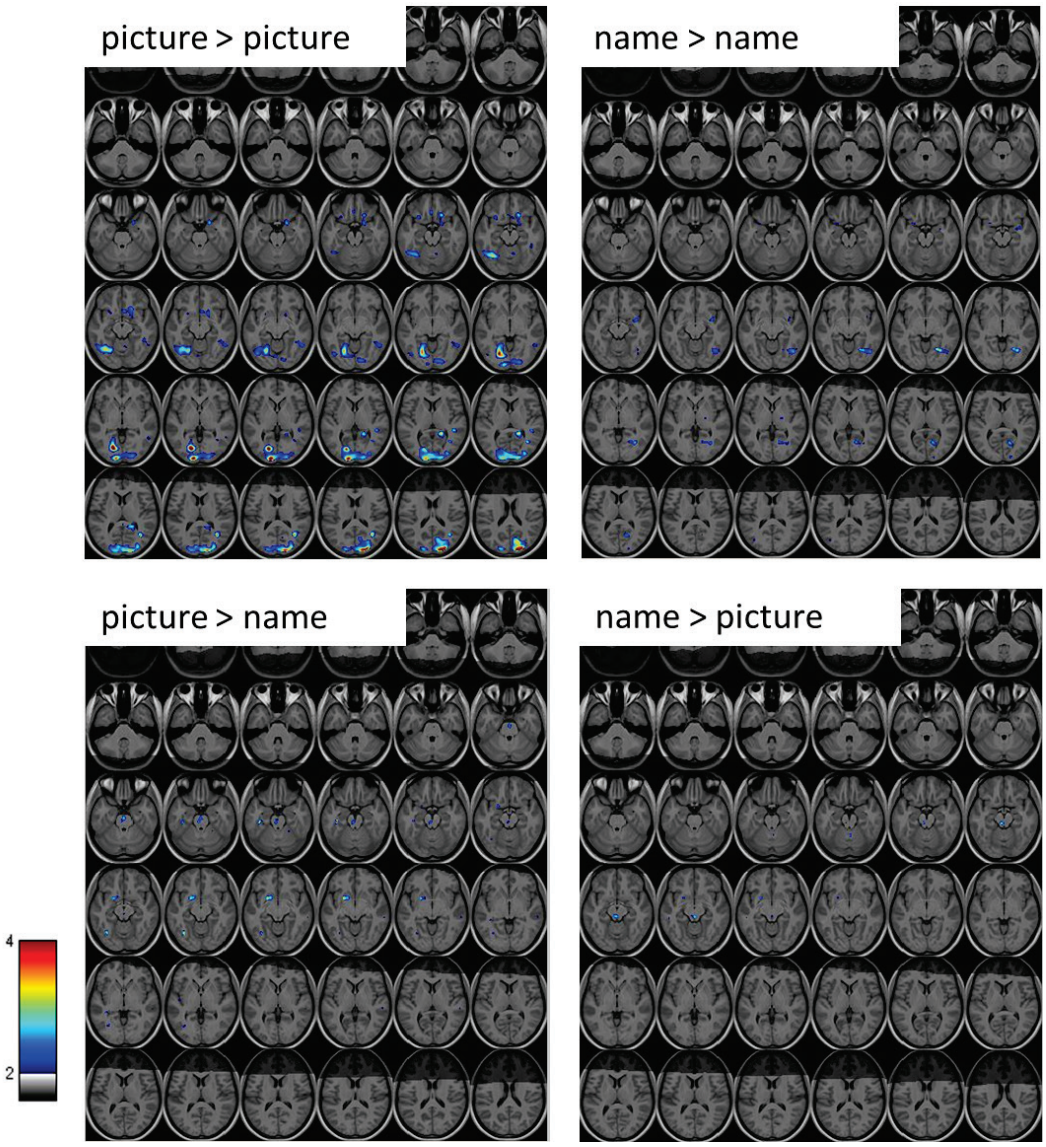


Figure 66 Searchlight decoding, represented as -log10(p) for a t-test across the 13 subjects that performed the second version of the Brad Pitt/Matt Damon/Tom Cruise experiment, i.e., with the one-back task on identity. No area seems to support decoding within AND across modalities.

include the key presses? How about implementing my own GLM? What happens if I correct for distortions (due to inhomogeneities in the magnetic field)? What if I remove noisy components after running an Independent Components Analysis? Maybe the problem lies in the normalization procedure. What if I try to carefully match my subjects' anatomies (using ANTs, Advanced Normalization Tools, which rely on diffeomorphic transformations)? What if instead of relying

on whole brain analyses, I define anatomical regions of interest? What about functional regions of interest (I had run separate scans to locate face responsive areas, and also areas that respond more to famous faces than unknown faces, as described on page 93)? Or maybe there was a problem with my decoding methods? Would feature selection (filter and wrapper approaches) help? Should I average less trials together to create examples (here we go, back to my schizophrenia!)? Should I tweak the classifier's parameters? I have run so many slightly different flavors of this decoding analysis, and looking back I wonder how I could be so obstinate. Needless to say, I will spare you most of the details and associated, innumerable figures.

This is not the first time that I get stuck running very slight different analyses on a resistant dataset, losing track of the big picture. The truth is that this is rather pointless. I am hoping that writing it in black and white will seal this lesson in my brain once and for all. John-Dylan Haynes visited Caltech while I was in the midst of these endless analyses, and I had breakfast with him at the Athenaeum. He gave me what I now consider to be one of the best pieces of advice I was ever given for my research, "go for the big effects." Do not get trapped in experiments that do not deliver such effects.

The big effects should show up even with suboptimal analyses.

In John-Dylan's lab, when they do MVPA they always use the same classifier with the same parameters (a linear Support Vector Machine with cost parameter 1). It should work to some extent. If it does not work at all, it is pretty much not worth continuing. In fMRI decoding, there simply is not enough data to properly train classifiers, let alone do cross-validations in the training set to pick the "perfect" parameters. All that you fit with these fancy techniques is noise. All the machine learning trickery is only useful if you have enough data to learn from, and this is never the case with fMRI.

3. Stage 3: the supersubject experiment

But I was stubborn and wanted to try one last thing. To train a classifier properly, you need enough reliable data. In my previous approach, I only had a few examples for each subject (four runs of data, yielding a total of 12 picture example and four name examples per identity after averaging all examples within each run). Also, I was worried about suboptimal matching of anatomical regions between subjects, after normalization. So I thought that the solution would be to run many sessions of fMRI data with one subject; there would be much more data to train a classifier, and I would not have any trouble realigning the subject's brain with itself from one session to another (or so I thought!). It also dawned on me that the one-back task on identity might have been causing trouble. Indeed, it required subjects to think of whom they had seen in the previous trial in order to correctly answer whether it was the same identity or not. This means that the representation of the identity from the previous trial would linger on during the current trial, thus muddling the identity-specific patterns that we were looking for. It is not clear whether this is the reason why the experiment failed in stage two, but it was a potential explanation and reason enough to rekindle the hope that we would find identity-specific, modality-independent patterns. Scientists never give up, you see (note that running a different experiment is different from grinding a dataset over and over again). I thus implemented a few tweaks to the experiment. Even though I did not really want to look at within modality decoding, I aligned the faces of all the actors and matched the histograms, using a set of routines that I had developed with my SURF student Gregory Izatt on the backward masking vs. continuous flash suppression project (page 72). I also wanted to cram some more trials in, and reduced the duration of a trial from four seconds to three seconds. Importantly, I equalized the number of picture and name trials; it would indeed be nice to have the same power going either direction (picture>name and name>picture). Finally, I came back to an identification task, which I implemented on 50% on trials, with a trial

specific key assignment (to avoid motor confounds) which would appear on screen when an answer was required.

I submitted my dear wife Christine to the ordeal; I needed a subject who would be motivated enough to withstand this boring experiment without falling asleep. She completed three runs of the main experiment twice a day, three days in a row, for a total of 18 runs. As she was getting a bit bored in the end, I added a twist on the last day: I substituted half of the written name trials with spoken names (which I generated through a European website which charges a small fee for a nice range of computer generated voices, http://www.acapela-group.com), and half of the picture trials with movie names clipped from the movie posters. Otherwise the experiment had the exact same structure. This was obviously not a gratuitous twist; if the decoding from pictures to written names worked well, I would be able to test whether the same patterns supported decoding for yet other modalities. Examples of the stimuli are shown in **Figure 67**.

Figure 67 Stimuli used in the last version of the Brad Pitt/ Matt Damon/ Tom Cruise experiment (the "supersubject" version). Only pictures and written names were used for 14 out of 18 runs. Movie names (clipped from posters) and spoken names (not shown, obviously!) were used in the four remaining runs (replacing half the picture trials and half the written name trials.

As in the previous two stages of this experiment, I first ran a searchlight throughout the brain, trying to find informative regions. I ran the four versions: train and test on pictures, train and test on names, train on pictures test on names, train on names test on pictures. The results are shown in **Figure 68**. The maps were thresholded such that the False Discovery Rate is less than 5% (the p-values were obtained from a binomial chance distribution). No voxels survived correction in the cross-modal maps. Interestingly, performance was very good for name>name decoding in the occipito-temporal cortex bilaterally (the regions of good decoding roughly correspond to regions responding more to text than pictures, in a univariate contrast; including, e.g., the Visual Word Form Area). These areas are strong candidates for word recognition units or name recognition units. It is unclear how much of this performance could be attributed to low-level confounds, and it is difficult to make a very strong claim with just three names. It is a promising result that could be pursued. Interestingly, in a recent report, Adrian Nestor (in Marlene Behrmann's group) was unable to decode individual pseudowords with a searchlight analysis[163].

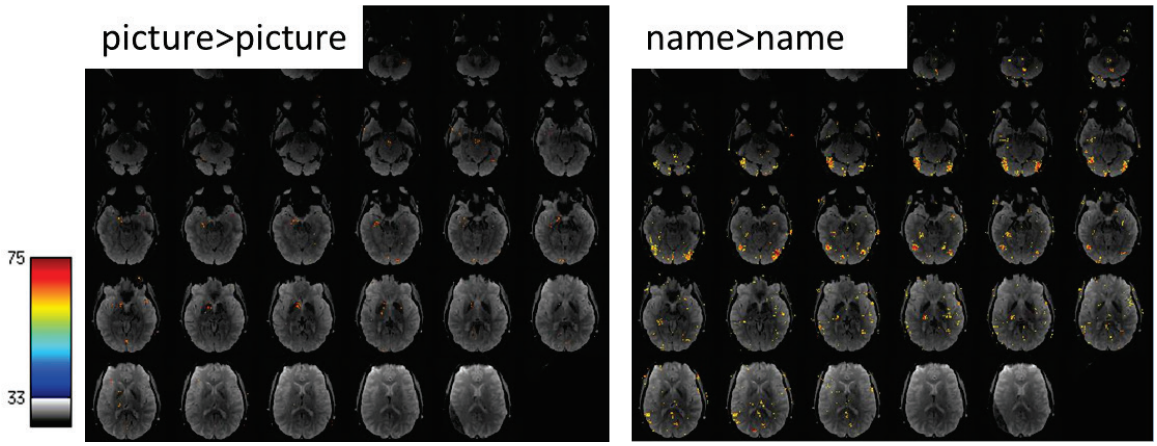


Figure 68 Searchlight decoding results for the last version of the Brad Pitt / Matt Damon / Tom Cruise experiment (the "supersubject" experiment). The picture>name and name>picture decoding schemes did not yield any decoding above the FDR threshold of q=0.05, hence are not represented here. The finding of significant decoding in the bilateral occipito-temporal cortices in this subject is a finding worth pursuing further, to understand for instance whether it corresponds to underlying Word Recognition Units or Name Recognition Units.

The positive result for decoding individual names tells us that we achieved good statistical power in this experiment. Since we could not evidence any crossmodal decoding, it also tells us that it is time to give it up. Anecdotally, I, of course, tried decoding in regions of interest (such as the hippocampi); I tried smoothing the data, I tried feature selection, etc. I am rather convinced that this data did not support crossmodal decoding.

I started thinking of this project in 2010, and the first subject was acquired on June 7th of that year. This means that I just summarized a project that occupied my thoughts for about two years in roughly 30 pages. According to Christof:

> "This is the very essence of research; it is difficult, and discovering something truly original and novel is rare, certainly much rarer than you would expect based on the literature. You have to do many things that after the fact appear pointless…"

…and such things I did. However, I came out with a true expertise on multivariate analysis of fMRI data, having tried pretty much anything that can be tried, and in the end realizing that fancy is not the name of the game for fMRI decoding. If it does not work with a simple approach, it is unlikely to work with a complicated one unless some cheating is inadvertently introduced (peeking at the test set). On many occasions I have spent hours implementing a new way to do the decoding, hours running it on dozens of cores, finally had a result. This got my hopes up for the front page of a good journal, only to find a few days later that I had in fact cheated without meaning to. In this vein, I have seen many papers that claim to increase their decoding accuracy with some feature selection scheme. I daresay that, more often than not, the accuracy only increases because they choose the number of features such that accuracy is maximal in the test set. This is plainly wrong, and a blatant case of double dipping[164]. If I had only one piece of advice for newbies, it would be this: use a simple linear classifier that is somewhat regularized,

and do not try too hard. Your efforts should go into thinking harder about ways to understand what your above chance decoding means in terms of what the brain is doing.

So what about invariant person recognition? We have not been able find any corroborating evidence for the Jennifer Aniston neurons with fMRI, nor have we been able to find any other strong candidate regions for the Person Identity Nodes (within the slab of occipito-temporal cortex that our imaging sequence covered). It thus remains unclear how our knowledge about the people that we know is organized in the brain, how regions that recognize faces are linked with regions that recognize written names, etc. It may be that fMRI is simply not the right tool for these investigations, despite the promise that MVPA (fMRI decoding) can offer a window on representations at scales smaller than the voxel size. As I will argue in the next chapter, a fMRI signal cannot arise without some clustering of neurons with similar tuning properties and this may be a reason for our negative findings here.

Perhaps I should have listened to Andreas Kleinschmidt, whom I met at the VSS conference in May 2010. We were sitting at a table by the pool at the Naples Grande hotel (now the Waldorf Astoria). I talked to him about my idea of finding “Jennifer Aniston fMRI patterns” and he was very skeptical, if not downright certain that it could not work. I barely knew him and his work back then and ignored what he said. I even thought to myself: well, since he is so highly dubious that I will succeed, it should be a nice contribution to the literature if I do! Unfortunately, I did not.

IV. REPLICATING SINGLE NEURON FINDINGS WITH FMRI

AT A GLANCE

- Guided by a strong single unit finding at UCLA, I looked for a categorical response to animals in the right amygdala with fMRI, using pictures from the IAPS database in one experiment (animals / not animals), and the image set used in patients at UCLA in a second experiment (animals / faces / landmarks).
    - ⇨ Weak finding of a categorical response to animals in the right amygdala, in the IAPS image set. No significant finding for the original UCLA patient image set.
    - ⇨ Strong categorical response to faces, as compared to response to landmarks, in the fMRI dataset; there is no such indication in the single unit data, either in terms of spikes or Low Frequency Potentials (LFPs).

- I tried to decode identity and viewpoint for a set of 20 faces (four identities x five viewpoints) in the macaque brain with fMRI, in areas where populations of single neurons represented this information.
    - ⇨ No significant decoding of identity where single neurons represent that information; however, fMRI decoding does pick up viewpoint information.

In the previous chapter, my thinking was heavily guided by a rather solid finding (at least in our lab) from single neuron recordings in the human brain (the Jennifer Aniston neurons). We failed to uncover information in fMRI activity that would corroborate this finding. In the first chapter I took time to introduce fMRI in great detail, and I made it quite clear that it was anything but a perfect readout of the underlying neuronal activity. In this chapter, I describe how I pursued two strong findings from single neuron recordings and put them to the test in fMRI experiments.

First, Florian Mormann (now a Professor at the University of Bonn, Germany) and Simon Kornblith (now a PhD student at MIT) were conducting a meta-analysis of the last 10 years of data collected at UCLA, and found that neurons in the right amygdala responded to animals significantly more than to other categories such as faces, objects and landmarks; they asked me to help design and run a fMRI experiment, to see if we could find evidence of a categorical response to animals in the right amygdala.

Second, Doris Tsao approached me with an orphaned dataset she had collected with a summer student, Archy de Berker (now a PhD student at University College London), in view of checking whether face identity and face viewpoint information that are differentially represented in the face patches of macaque monkeys (according to single unit recordings) would be picked up with fMRI.

I report here on these two attempts to replicate single neuron findings with fMRI, and what I learned in the process.

A. A categorical response to animals in the human right amygdala.

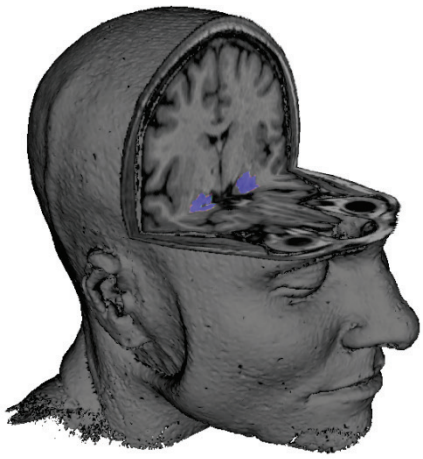


Figure 69 My wife Christine's reconstructed head, with her bilateral amygdalae labeled (image produced with Slicer; automatic anatomical labelling with Freesurfer). Note: the dip in front of her right ear is due to the headphones she was wearing in the scanner.

1. Single neurons in the human right amygdala respond more strongly to animal pictures than to other categories

I did not contribute to the analysis of the single neuron data, however I find it important to describe it here in detail in order to give substance to the finding. Furthermore, since I was an author on the paper[165] that came out of this study, I did proofread the manuscript several times and commented on the techniques used here.

The single unit data was recorded from the MTL in 41 neurosurgical patients who were undergoing epilepsy monitoring, as described in the previous chapter (34 right handed; 23 male; 18–54 years old). Subjects were sitting comfortably in bed while they viewed ca. 100 images per session on an LCD monitor (1 second each, with six repetitions in pseudorandom order). Stimulus sets contained images of persons (grand average 72%), animals (10%), landmarks (15%), or objects (3%). During 111 experimental sessions, a total of 3598 neurons (2153 multi-units, 1445 single units) were recorded from in the amygdala (1239 units), hippocampus (1397 units) and

entorhinal cortex (962 units). Of these, 460 units (131 in amygdala, 218 in hippocampus, 111 in entorhinal cortex) responded significantly to one or more of the presented stimuli.

Neurons in the amygdala responded predominantly to pictures of animals but not to pictures of other stimulus categories (**Figure 70**). The composition of the stimulus sets and the regional yield of units varied across sessions. To compare neuronal responsiveness across regions and categories, Florian and Simon counted all instances where an image from a given stimulus category was presented to a neuron from a given MTL region. Based on these cumulative presentation counts, they then calculated the overall percentage of significant neuronal responses for each MTL region for each stimulus category (e.g., if 10% of amygdala neurons responded to 10% of all animal pictures in each session, the resulting response percentage would be 1%). This allowed a straightforward comparison of response percentages across regions and categories. To statistically assess differences in responsiveness to the four stimulus categories, they used the Mantel-Haenszel chi square test, a generalized version of Pearson's chi square test for two-by-four contingency tables, stratified for different subjects. Comparison of response percentages for different stimulus categories in the three MTL regions showed a highly significant preference in the responses of amygdala neurons for animal images, while responsiveness in hippocampus exhibited no significant difference between stimulus categories and entorhinal cortex showed a preference for all other stimulus categories over persons (**Figure 71a**).

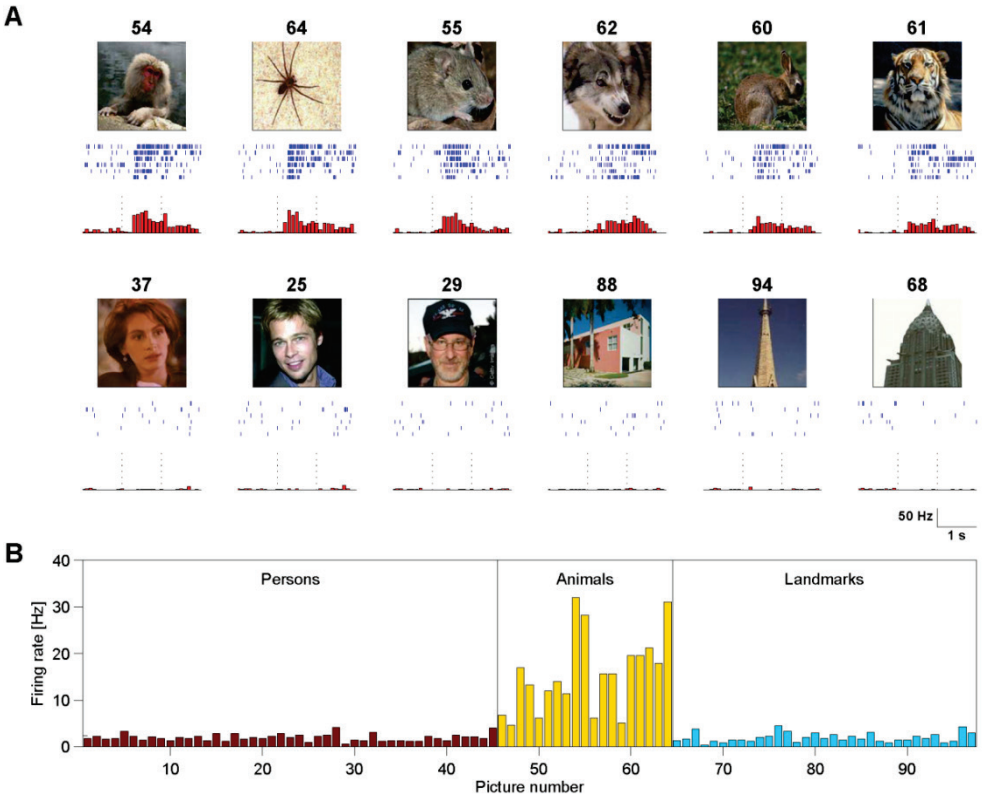


Figure 70 A single unit in the amygdala activated by animal pictures. A: Responses of a neuron in the right amygdala to pictures from different stimulus categories, presented in randomized order. For each picture, the corresponding raster plots (order of trials from top to bottom) and peristimulus time histograms are given. Vertical dashed lines indicate image onset and offset (one second apart). B: The mean response firing rates of this neuron between image onset and offset across six presentations for all individual pictures. Pictures of persons, animals and landmarks are denoted by brown, yellow and cyan bars, respectively. Reproduced from [165].

The preferential response of amygdala neurons to animal stimuli could have been caused by two different effects. On the one hand, the number of units that respond to animals could be particularly high. On the other hand, those units that respond to animals could do so with a lower within-category selectivity, i.e., they could respond to a higher portion of animal stimuli than what is observed for units responding to other categories. In the data, Florian and Simon found a combination of both effects. Although animal pictures constituted only 10% of the stimulus material, they comprised 24% of all responsive units. Furthermore, units that responded to

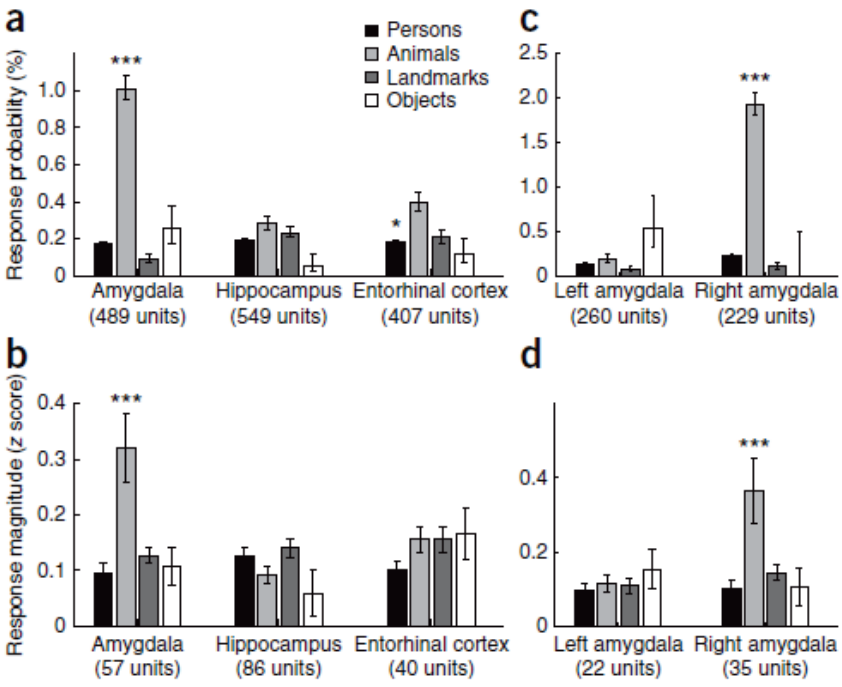


Figure 71 Amygdala neurons respond preferentially to animal pictures. (a) Response probabilities of neurons in different MTL regions to different stimulus categories revealed significant preferences in the amygdala ($P<10^{-15}$, main effect of increased responses to animals at ~1%) and entorhinal cortex ($P<0.03$, main effect of decreased responses to persons), but not in the hippocampus. (b) Mean response magnitudes of all responsive neurons showed increased response activity of amygdala neurons to animals ($P<10^{-5}$). (c,d) The animal preference in both response probability and magnitude was seen only in the right amygdala ($P<10^{-15}$ and $P<0.0005$, respectively). Error bars denote binomial 68% confidence intervals (a,b) and s.e.m. (c,d). *$P < 0.05$, ***$P < 0.001$. Reproduced from [165].

animals responded to 2.9 animal pictures on average, whereas units that responded to one of the other categories responded to an average of 1.9 stimuli.

To control for potential confounding effects, they evaluated several additional factors in the stimulus set. Since the animals in the stimulus set were usually depicted with body parts, whereas many of the persons were shown as face only, they tested whether the difference in responses to animals and persons was actually a difference in response to faces versus bodies by contrasting the response percentages for persons with or without body parts, but found no significant effect (p=0.95). Secondly, they tested for effects of familiarity of the stimuli since the animals were not personally known to the subjects, whereas 95% of the persons were familiar. Responsiveness to unknown and familiar persons did not differ statistically in the amygdala (p=0.98). They also

evaluated food items as a subcategory of objects potentially related to reward-processing, but found no significant preference of amygdala neurons for this stimulus category as compared to persons or landmarks (p=0.47).

To test for laterality in the electrophysiological amygdala data, Florian and Simon performed the category preference analysis separately for the left and right amygdala. They found that the preferential neuronal responses to animals were exclusive to the right amygdala (**Figure 71c**).

To rule out an influence of the underlying pathology in the patients, they repeated the category preference analysis after excluding any amygdala neurons that were located in the hemisphere containing the epileptic focus. The observed laterality was confirmed to be independent of the side of the epileptic focus. To test for an effect of handedness on the observed laterality, they performed the statistics separately for the 33 right-handed and the eight left-handed patients. Both groups showed a significant category effect for animals in the right, but not in the left amygdala ($p<10^{-10}$ and p=0.01, respectively), indicating that these two functional asymmetries are not related to each other.

To test whether specific response patterns of amygdala neurons to animals are also present at the population level, Florian and Simon analyzed how images are segregated by response patterns, following a categorization technique previously applied to monkey inferotemporal cortex neurons. For a set of 201 amygdala units (96 in the left, 105 in the right amygdala) recorded during ten sessions from eight patients, they compared response patterns to a set of 57 stimuli that were presented to each of these units. Representational dissimilarity matrices reflecting the dissimilarity between every pair of these 57 stimuli showed a specific response pattern to animals in the right, but not the left amygdala that differed from the response patterns to persons and landmarks (**Figure 72**). Furthermore, hierarchical clustering analysis, which groups the stimuli based on similar response patterns without any use of category information, showed that animals

and non-animals indeed form distinguishable clusters in the population code of the right amygdala (**Figure 72***)*.

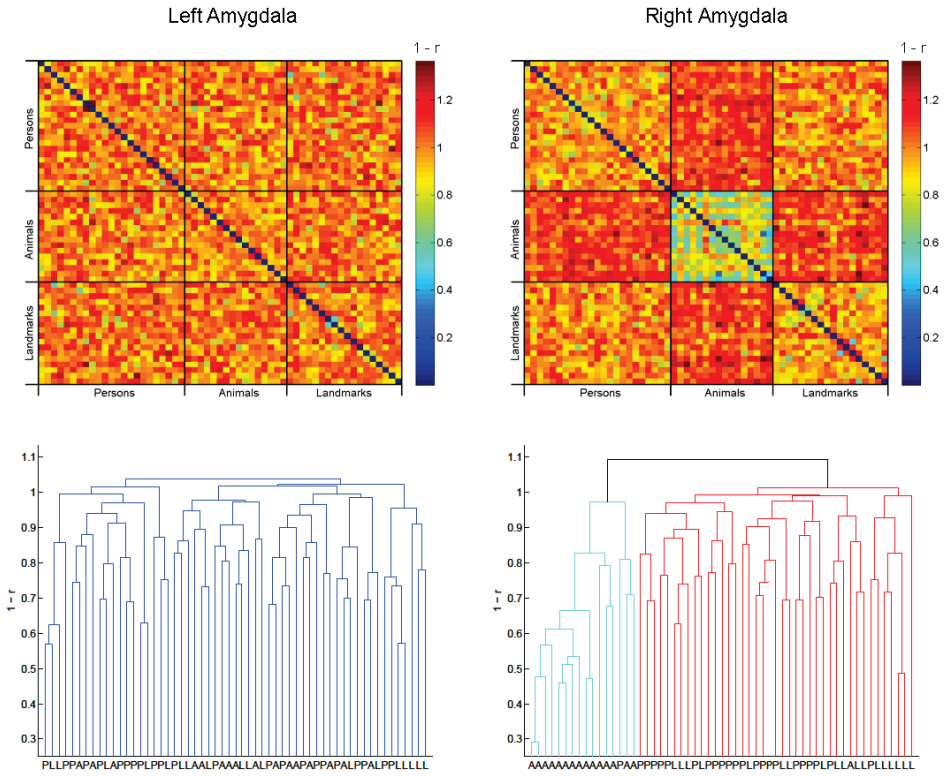


Figure 72 A specific category response to animals in the right amygdala at the population level. (a) For a set of 201 amygdala units (96 left, 105 right) that were all presented with the same 57 stimuli (23 persons, 16 animals, 18 landmarks), we constructed representational dissimilarity matrices by determining the dissimilarity in evoked response patterns for each pair of stimuli (as 1 – *r* from the Pearson correlation across units). (b) Hierarchical cluster analysis automatically grouped stimuli with similar response patterns together into clusters. In the right amygdala, this unsupervised procedure yielded a cluster that contained all animal stimuli, whereas no such category effect was found in the left amygdala.

2. A fMRI experiment to replicate and control aspects of the single unit findings

a. The official story

The animal images that elicited neuronal responses in the amygdala consisted of 23 pictures and contained not only aversive animals (e.g., spider, snake), but also cute animals (e.g., squirrel, rabbit, panda). To test whether emotional valence and arousal had an effect on the amygdala's

responsiveness, Florian and Simon examined the possible influence of valence and arousal ratings on amygdala responses to these 23 stimuli and found no evidence for any effect of arousal or valence, indicating that the observed responses were indeed category-specific to animals and not related to the emotional content of the pictures. To further corroborate the hypothesis that the human amygdala is specifically involved in the processing of animals and that this effect is independent of the emotional valence or arousal caused by the stimuli, we decided to conduct a fMRI study. We selected 60 animal pictures from the International Affective Picture System (IAPS)[166] and matched these with non-animal IAPS pictures having a similar emotional valence and arousal. The 60 images in each of the two categories were further divided into four groups having high and low emotional valence and high and low arousal, respectively. The resulting eight groups of images were presented in a blocked design while subjects performed a 1-back memory task. There were 15 distinct images in each block; each image was presented for 800ms after a 200ms blank screen, and there were three randomly chosen repetitions, so that 18 images were shown per block. A run consisted of 16 blocks (two times each of the eight blocks). The blocks alternated between the animal and non-animal categories but valence / arousal were randomized. Each run began and ended with a ten second gray fixation screen, for a total of ($16 \times 18s + 20s = 308s$) approximately five minutes per run. After the task, subjects rated the 120 images outside the scanner for valence and arousal to confirm that there was indeed no significant difference in these dimensions between the animal and control pictures.

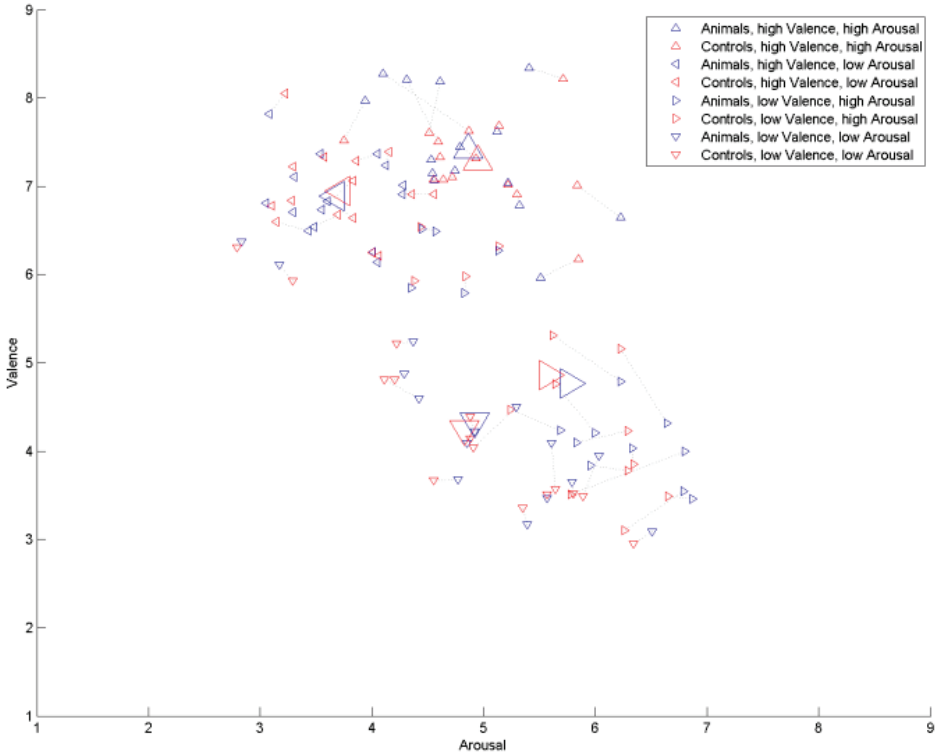


Figure 73 Stimulus sets for the fMRI paradigm. 60 pairs of animal and non-animal stimuli, taken from the IAPS picture set and matched for emotional valence and arousal, were divided into four groups of low and high valence and low and high arousal, respectively, and presented to ten subjects in a 3.0 T Siemens Magnetom Trio Scanner. The average values for each of the eight groups are represented by large triangles.

I analyzed the BOLD response extracted from the left and right amygdala using a conventional general linear model and calculated the contrasts between animals and non-animals, high and low valence, and high and low arousal, respectively. Group analysis revealed a localized cluster of voxels activated by animals in the right amygdala. The significance of this cluster at an overall level of 0.02 was confirmed after correction for multiple comparisons based on the total volume of both amygdalae. Additional positive and negative activations can be seen bilaterally in the lateral occipital cortex and the occipito-temporal junction and represent activity along the ventral object recognition pathway.

These results were published[165].

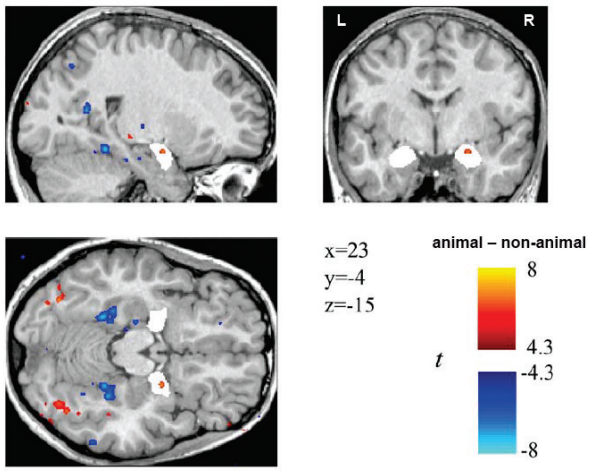


Figure 74 Thresholded statistical parametric map for the contrast animal vs. non animal. Group analysis of ten subjects using a standard general linear model (GLM) showed a cluster of voxels in the right amygdala (MNI coordinates x=23; y=-4; z=-15) that responded more strongly to animal than to non-animal pictures (P<0.001, uncorrected; P=0.02 after small-volume correction based on the total volume of both amygdalae). This animal vs. non-animal contrast is independent of emotional valence and arousal since stimuli from both categories were matched for these emotional dimensions. Reproduced from[165].

b. A look behind the scenes

This fMRI experiment took about a year, from conception to running, and many months for analysis; it ended up as supplementary Figure 11 in a Brief Communication. It was a good idea, and we tried hard, but it just did not work well enough to be able to make a big claim. The final figure (**Figure 74**) showed one aspect of the results.

We acquired more data and conducted more analyses than described in the previous section, and here I wish to disclose them, perhaps even make some sense of them, which is not an easy endeavor. Unfortunately, selective reporting is a major reason why a lot of published experimental results are not reproducible, and I am pretty sure it is quite a common practice in most fields of science.

We scanned more subjects than we reported in the figure. Initially, we scanned 12 subjects, using a rather standard imaging protocol (**Figure 75**, left), with 3mm isotropic voxels and sub-axial slices (slightly tilted forward, for maximum coverage of the brain). We noticed a fair amount of dropout close to the amygdala, the classic black hole found in most EPI studies, due to magnetic field inhomogeneities at the air-tissue interface (for a visual of how the dropout affects the borders of the amygdala regions in one subject, see **Figure 76**). Our result (the one in the published supplementary figure) is based on the analysis of those subjects. We were not quite satisfied with it and decided to try and improve our imaging to get better signal in the amygdala, and perhaps a more significant result. I thus scanned a couple of subjects and explored variations in the imaging parameters, roughly guided by some publications in the domain [167,168] and interacting with Mike Tyszka. I found that 2mm voxels and coronal slices reduced the dropout dramatically, and I ended up scanning 14 subjects with this new imaging protocol (**Figure 75**, right).

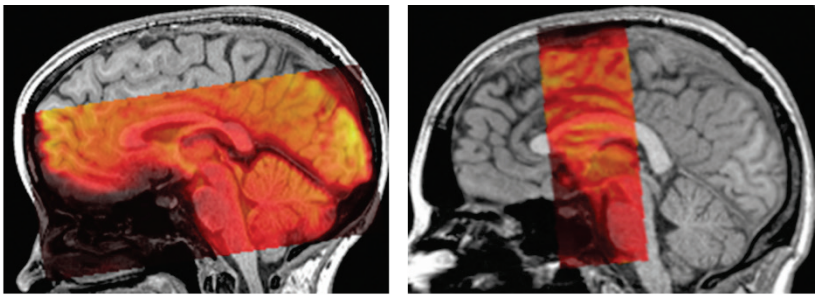


Figure 75 Comparison of the two scanning sequences that we used. On the left, the sequence used for the published results. On the right, an "optimized" sequence to avoid dropout in the amygdala region. Note the higher resolution (2mm isotropic vs. 3mm isotropic), but much reduced imaging volume (24 slices with 128mm field of view vs. 32 slices with a 192mm field of view), in the improved sequence.

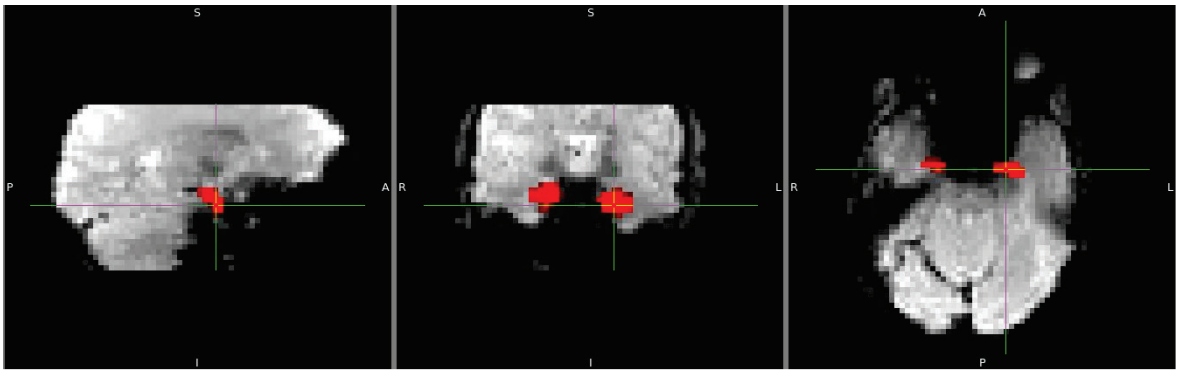


Figure 76 Dropout due to magnetic field inhomogeneities affects signal very close to the amygdala (labeled in red) in a typical subject scanner with the 3mm isotropic, original fMRI protocol.

In the end, we were quite pleased to see that my protocol tweaks succeeded: the amygdalae were much less affected by dropout in the second group of subjects (as can be seen from the absence of a "tail" towards lower signal in **Figure 77**).

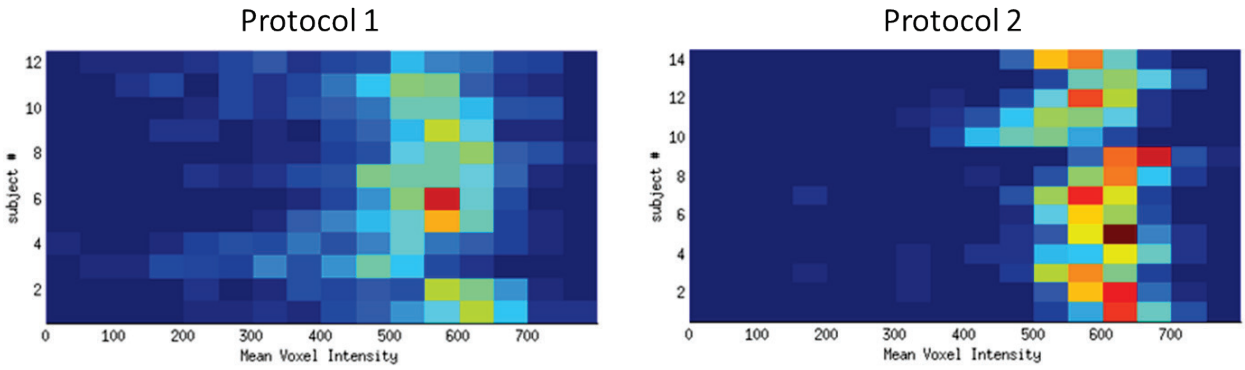


Figure 77 The mean fMRI signal in the amygdala, binned for each subject. Protocol 2 shows less of a tail towards lower values, hence is less affected by signal dropout, as desired.

We also conducted many more analyses than the voxelwise analysis reported in the figure (**Figure 74**). I spent several weeks looking at Region of Interest analysis, i.e., averaging all voxels from the left amygdala on the one hand, and all voxels from the right amygdala on the other hand. I tried using anatomical labeling in individual subjects (with Freesurfer), anatomical labels on the normalized brain (Juelich atlas, Harvard-Oxford atlas available in FSL), and the SPM anatomy

toolbox labels. I tried different preprocessing schemes, different statistical tests for inference, etc. It would be difficult (and arbitrary) to pick just one of the many analyses I attempted; in essence, though, the pattern that emerged most often was as follows: a significant effect of arousal (high arousal blocks of images elicited more activity in the bilateral amygdala than low arousal blocks); no effect of valence; and a trend towards a higher response to blocks of animal images than blocks of non-animal images, bilaterally. I could, of course, have picked a specific analysis that led to a significant animal vs. non animal contrast in the right amygdala and not in the left and reported it, but this would have been dishonest; clearly, the ROI analyses did not support a lateralization of categorical processing for animals in the amygdala. Note that the ROI analyses may not be the best way to go in the amygdala: since the amygdala is a very heterogeneous structure, with several subnuclei which likely perform different functions, averaging contrasts across the whole structure may hide some local effects. This is why we ended up solely reporting the voxelwise analysis. For illustrative purposes, **Figure 78** shows the results of one arbitrarily chosen version of the ROI analyses; the ROIs were defined individually with Freesurfer, and I looked at the differences in beta parameters (corrected to reflect percent signal change). Note the most significant result is the effect of arousal in the IAPS experiment (for a description of the fMRI experiment with the UCLA image set, see below).

In contrast to ROI analyses, a voxelwise group analysis depends critically on voxels being well aligned from one subject to another; this is a major difficulty, when using whole brain normalization, which leads to imperfect matching between subjects. At the time, I did not attempt a better normalization targeted to the amygdalae. However, note that yet another issue lies in coregistering the functional images to the structural image; EPIs suffer from distortions due to magnetic field inhomogeneities. At the time, I did not know much about this, and did not collect in-session B0 maps as part of the scanning protocol; hence, I could not apply any distortion correction. Note that I performed the voxelwise analyses on normalized data only, without

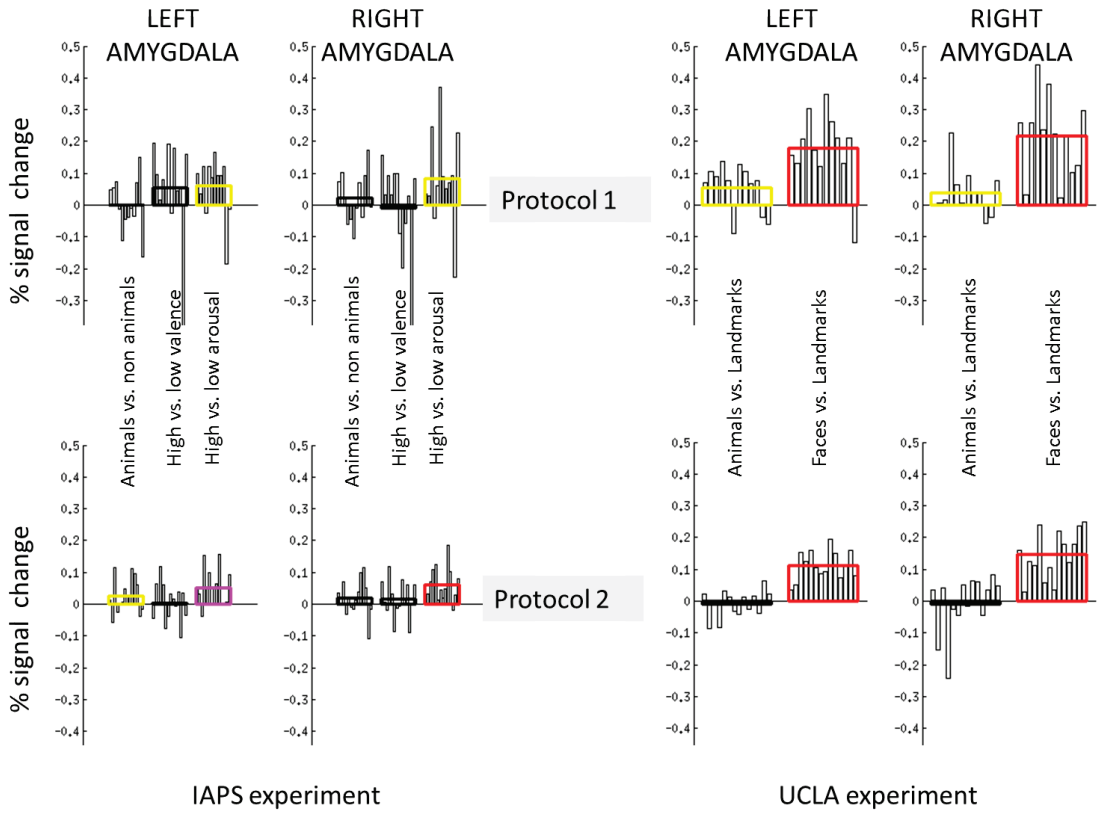


Figure 78 ROI analyses, based in the mean activation in the left and right amygdalae, for both sets of subjects (top, the original 13 subjects scanned at 3mm isotropic; bottom, the new set of 14 subjects, scanned at 2mm isotropic), and both image sets (left, the IAPS image set; and right, the UCLA image set). The large bars represent the average across subjects, and the small bars are for individual subjects. The p-values for a one-sided t-test against zero are color coded (yellow: $p<0.05$; magenta: $p<0.01$; red: $p<0.001$). The only consistent finding in the IAPS experiment at the ROI level (average across all left and right amygdala voxels, respectively) is a higher activation for high-arousal than low-arousal images; no significant categorical effect can be seen (except in the left amygdala, in the second set of subjects). With the UCLA image set, there is a very significant activation to faces in the left and right amygdalae, compared to landmarks. There is also a weak activation to animals, as compared to landmarks, which only transpires in the first set of subjects however.

additional smoothing. Smoothing is commonly applied to alleviate the misregistration problems, when performing group level analysis. However, smoothing can be detrimental if the smoothing kernel is larger than the true size of the activated region. In my case, I found that smoothing worsened the right amygdala's activation that I was finding with the animals vs. non animals contrast.

Conducting a similar whole brain analysis in the new group of subjects, I found that our result, the cluster of significant voxels in the right amygdala, did not replicate at the same threshold. However, I was pleased to find that by lowering the threshold, there was a positive cluster at the exact same coordinates as previously (**Figure 79**); we had thus replicated our finding in some fashion. The lesser power may have stemmed from multiple factors: perhaps the registration was less accurate in the new group of subjects, due to more distortions; perhaps the new scanning parameters, while they decreased the dropout, were not as good for detecting BOLD responses.

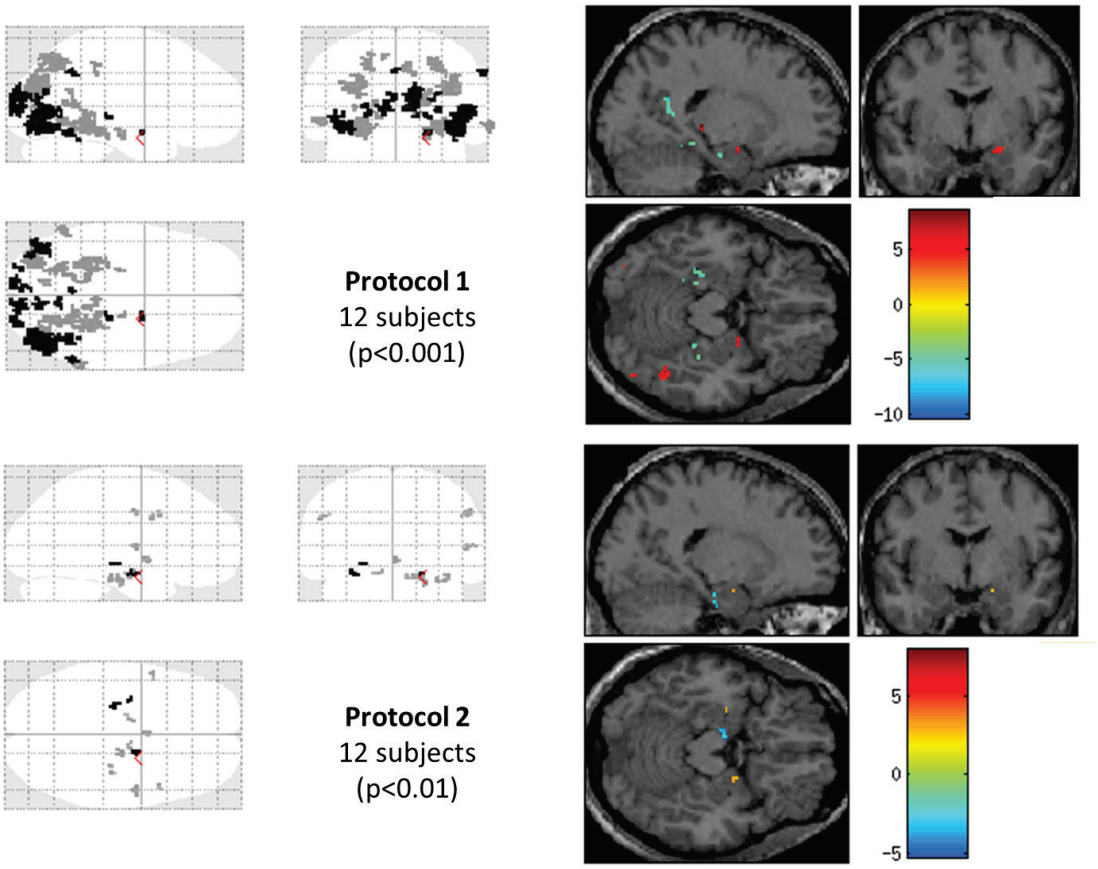


Figure 79 Statistical parametric maps for the contrast animals vs. non animals, in the experiment using the IAPS image set, in the two set of subjects (Protocol 1, 3mm isotropic; Protocol 2, 2mm isotropic). These are whole brain, group results (in MNI space). On the left, a glass brain representation, and on the right, an orthographic projection on the normalized anatomy of one subject, centered at MNI coordinates [20 -4 -18]mm. Note that the threshold is different for the two protocols. Nevertheless, this shows that the original cluster of voxels found in the right amygdala somewhat replicates in the second, independent set of subjects.

We chose to report solely on the first dataset, mostly because the addition of the second dataset complicated matters (and there is only so much you can say in the legend of a supplementary figure!).

There was, however, also a slightly larger complication. We ran two different experiments: the one based on the IAPS images, which I described, and also another one, based on the images from the UCLA set. The original animal images from the single-neuron recordings study were divided into three categories: animals, persons, and landmarks, with 15 stimuli per category. We used a block design, with 15 distinct images in each block; each image was presented for 800ms after a 200ms blank screen in a 1-back attention task with three randomly chosen repetitions, so that 18 images were shown per block. A run consisted of 15 blocks (five times each of the three blocks). Each run began and ended with a ten second gray fixation screen, for a total of (15×18s + 20s = 290s) approximately five minutes per run. At the end of the fMRI session, all images were presented outside the scanner and participants were asked to indicate how positive/negative (valence) and how exciting (arousal) they thought each image was on a nine-point Likert-type scale, ranging from one to nine.

This second experiment, while it was not as well designed as the IAPS experiment to control for valence and arousal, provided a nice additional check of whether we would replicate single neuron findings, i.e., a higher response to animals than other categories (faces and landmarks) in the right amygdala. Unfortunately, the right amygdala cluster did not show in the animals vs. landmarks contrast. Worse, there was a significant cluster in the approximate same location, in the left amygdala, for the first group of subjects (which did not replicate in the second group of subjects however). The non-replication of our cluster in the right amygdala could have stemmed from the contrast with landmarks; landmarks may themselves elicit more hemodynamic activity in the right amygdala than in the left amygdala. This does jeopardize our previous finding to some extent, but it may also be due to the presence of another category of stimuli in the design:

faces! While the face category did not seem to elicit a particularly high response in either amygdala in the single unit data (refer back to **Figure 71**), faces have repeatedly been found to be a very potent activator of the amygdala in fMRI studies. We replicated this well-established result easily in this experiment: the contrast faces vs. landmarks led to very significant activations in the bilateral amygdalae, in both groups of subjects (**Figure 80**). Note that, in keeping with the rest of the analyses, I did not perform any smoothing here, but some smoothing would make these activations even more significant (since they are quite large, smoothing enhances them rather than drowning them). A ROI analysis (which you may already have peeked at in **Figure 78**, right) also

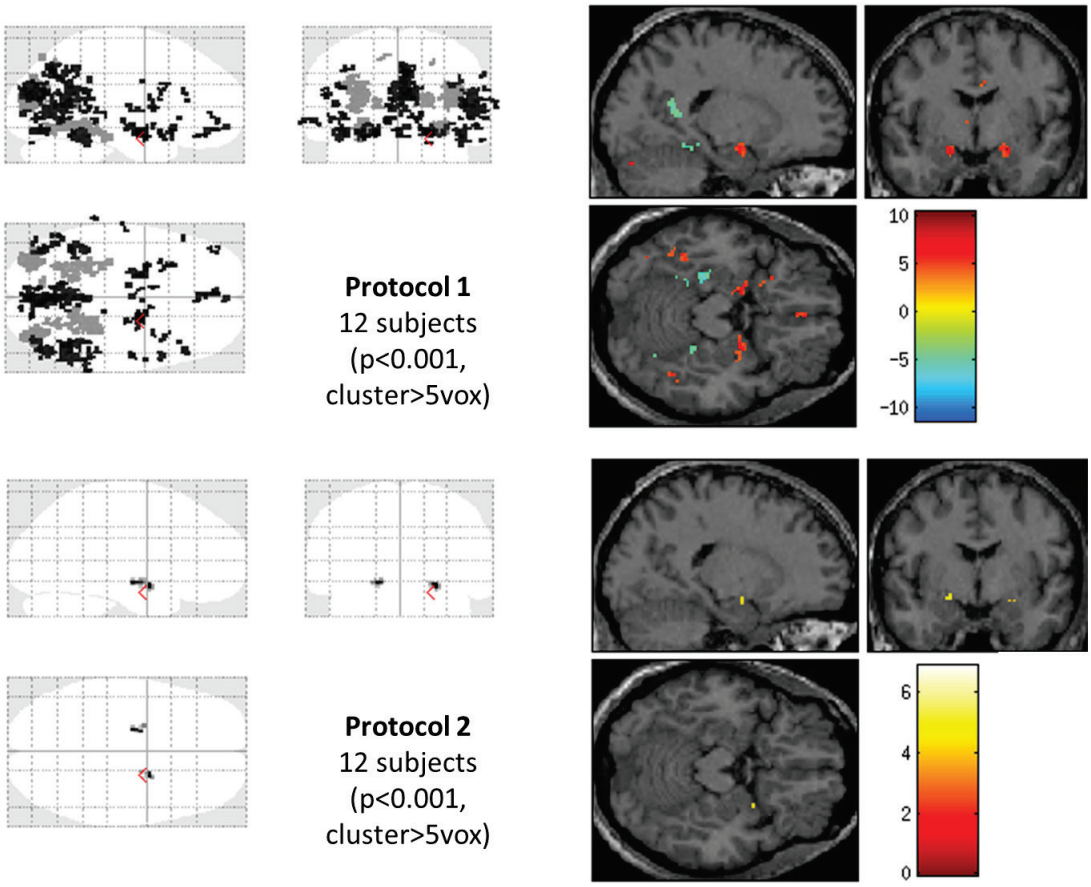


Figure 80 Statistical parametric maps for the contrast faces vs. landmarks in the experiment based on the UCLA image set. In both independent sets of subjects, we note a clear bilateral activation of the amygdalae (the orthographic projections are centered at MNI coordinates: [20 -4 -18]mm).

shows a very significant activation bilaterally for the faces vs. landmarks contrast, in both groups of subjects.

This is yet another reminder that fMRI and electrophysiology are not always in agreement. What does it mean though, to have an overwhelming fMRI activation, but no clear neuronal responses? It can maybe be understood in terms of fMRI measuring incoming information (subthreshold activity), rather than neural output (action potentials), as I alluded to in the first chapter (page 17). It has been reported before that local field potentials (LFPs) are a better predictor of fMRI activity. Interestingly, we did have a look at the LFPs from the single neuron recordings at UCLA to see if they matched fMRI better; however we did not find it to be the case, as can be seen in **Figure 81**, which shows the event related potentials (ERPs) averaged across all channels in the amygdalae.

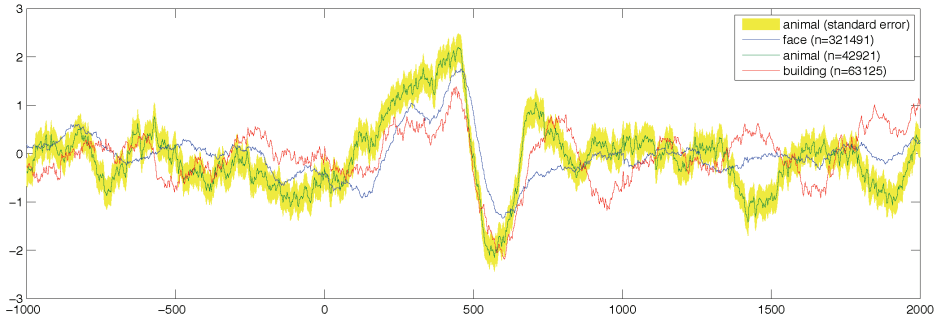


Figure 81 The average event related potentials (ERPs) computed from the Low Frequency Potentials, averaged across all amygdala channels, for three stimulus categories (faces, animals and buildings). The amplitude of the positive peak at 450ms for the animal ERP is larger than for other stimulus categories, but not significantly. The overwhelming response to faces picked up by fMRI does not show in the LFP analysis.

What do we make of all this? It is unclear what we can really make of the cluster that we found in the right amygdala; to me, it remains a borderline finding, too weak to do much further work with. The only truly incontestable fMRI activation is that to faces, and as we discussed, this is at odds with what the neurons recorded from in epileptic patients at UCLA seem to be doing.

Since a key component of this thesis is the use of multivariate analyses for fMRI decoding, I did try and perform MVPA on this data. While I was able to decode almost perfectly the category that was presented from activity in the ventral pathway (inferior temporal and fusiform gyri, etc.), decoding in the amygdalae did not yield any significant decoding, either with wholebrain searchlight approaches or with ROI analysis. The conception that MVPA will magically read out underlying neuronal information where univariate analysis fails needs a reality check. This is what I hope to bring in the next section.

B. Face identity and viewpoint information in the macaque face patch system

My ventures trying to replicate electrophysiological findings with fMRI have not met much success thus far. In the previous chapter, I tried relentlessly to find invariant representations of familiar person identity, yet when I looked specifically in the hippocampus, even the Jennifer Aniston neurons mischievously shunned my attempts. Another significant finding in human single units, that of a categorical response to animals in the right amygdala, also made itself scarce and even morphed into something else (a response to faces) when examined with fMRI. What sort of representations can fMRI pick up? What makes the Jennifer Aniston neurons and the categorical response to animals in the amygdala invisible to hemodynamics?

An amazing opportunity to start answering these questions arose on a sunny afternoon (well, I do not quite remember what the weather was like in truth, but sunny afternoon is a fairly likely assumption in Pasadena, CA). I had just had a meeting with my committee members, embarrassingly telling them about the many hoops I had been jumping through with the invariant identity project, without any tangible results yet. I was trying to get excited about the result I had from decoding identity from names, and starting to think of how I could exploit that finding and perhaps get some interesting insights into the brain's repository for names of familiar people. But in truth, I was quite disappointed that my efforts for the last several months had not met any

success yet. A few hours after the meeting, I got an email from Doris. She offered to let me have a look at a dataset she had collected with a SURF student, Archy de Berker, two summers ago; the experiment was aimed at trying to decode face identity and face viewpoint from fMRI data in the macaque monkey. The uniqueness of this dataset was that she had previously recorded electrophysiological data from some of the face patches (in other macaque monkeys), and hence she had single unit data for the same stimuli, a ground truth against which to compare the performance of fMRI decoding.

After getting a bit more information, I was hooked; this was a perfect opportunity for me to apply my skills, it was a nice new fresh dataset (well, two years old, but fresh for me), and perhaps it would offer me some insights on my failures. I was so excited that I had a preliminary analysis of the data within a week, which pretty much formed the core of the story that I am about to tell. Of course, refining the story was a bit more time consuming. This also made me realize how pleasant it is to have clean data, in which you have almost no doubt what is happening, and your only concern becomes to present it in the most intelligible and truthful way. But without further ado, here is the story.

1. Face viewpoint and identity information in single units

Six bilateral face patches can be evidenced with fMRI in the macaque monkey[169], which are heavily connected[170]: PL, a posterior patch; ML and MF, two patches in the middle temporal lobe; AL and AF, two patches in the anterior lateral temporal lobe; and AM, a patch in the anterior medial temporal lobe (**Figure 82**). These are very reproducible from one monkey to the next; Doris uses a functional localizer comprising eight blocks per run and repeated roughly 10 to 15 times during a fMRI session; the blocks are: human faces (unfamiliar), human faces (familiar), monkey faces (unfamiliar), monkey faces (familiar), fruits, bodies, hands, and technological objects.

Doris targeted ML and MF for electrophysiological recordings, and found that they comprised a majority (97%) of face selective cells[171]. In 2010, Doris and Winrich Freiwald published a beautiful study in Science[172], in which they described the face viewpoint and face identity tuning of neurons in the different face patches.

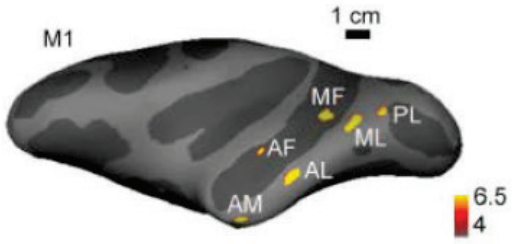


Figure 82 Face patches evidenced with fMRI in monkey M1. The face localizer fMRI experiment consists of blocks of human faces, monkey faces, fruits, bodies, hands, and technological objects. The contrast faces vs. fruits, hands, bodies, objects is used to find areas of the monkey brain that are more responsive to faces than to other categories. Reproduced from [172].

I had the opportunity to take a look at this dataset for myself and apply multivariate analyses to it, combining the information for multiple neurons to try and decode face viewpoint and face identity. The data had been acquired in three different monkeys (I will refer to them here as M1, M2 and M3, to preserve their anonymity), over the course of four years. The number of neurons recorded from each monkey, and which patches were recorded from, is summarized in **Table 2**. I actually went through the lab notebooks for all of these recordings (deciphering notes taken almost ten years ago), to do an analysis of clustering. It made me realize how important good bookkeeping is for experimental science; that someone can go through your notes and make sense of them many years later is quite a feat, and I shall aspire to being this disciplined when I run experiments (I currently am having a hard time piecing together what I have done five years ago I can hardly imagine someone else taking up this task!).

Spend the time needed to thoroughly document everything you are doing so that you can go back to your data and analyses many years later and make sense of them (almost) instantly.

MONKEY	PATCH	# UNITS	# SESSIONS	COLLECTED BETWEEN
M1	AL	37	12	02/09/2007 – 10/07/2007
	AM	134	23	05/08/2007 – 11/01/2007
	MF	83	7	08/21/2008 – 08/29/2008
M2	AL	177	59	03/10/2006 – 10/14/2007
	AM	59	14	10/20/2008 – 11/07/2008
M3	ML	57	26	09/10/2004 – 11/07/2005

Table 2 Summary of single unit recordings reported in [172].

The monkeys were passively fixating (they were rewarded periodically for their good fixation with juice) while being presented with a set of 200 images, the "face views" dataset. This image set comprised pictures of 25 human individuals, taken from eight different vantage points: left full profile, left half profile, frontal, right half profile, right full profile, top, bottom, and back of the head (**Figure 83**). One out of 25 of the individuals was familiar to the monkeys (Doris herself).

Figure 83 The eight viewpoints in the face views dataset, used by Winrich Freiwald and Doris Tsao in [172].

For each patch in turn, I selected all the single, well isolated units that had seen each picture at least three times, which yielded respectively 68 neurons in ML/MF, 105 neurons in AL, and 159 neurons in AM. I randomly picked three trials for each image for neurons that had seen the images more than three times. I did not throw out any neurons based on their visual responsiveness or face selectivity (which they did in [172]), as I preferred to have a complete,

unbiased dataset (even though it remained inherently biased, since in the first place experimenters chose to record from certain neurons which had the properties that they were looking for). The properties of single units in ML and MF are, as far as we understand, very similar; hence we pooled data from these two patches. Note also that I pooled neurons over different monkeys and over different sessions, which may seem rather artificial; however, it is the only way to truly build a population of neurons, since only two to three neurons can be recorded from, simultaneously, with the fine tungsten electrode used in these experiments. I defined the response of a neuron as the firing rate in the [50ms 200ms] window after stimulus onset (each image was shown for 200ms). I defined the baseline as the firing rate in the [0ms 50ms] window. The baseline was removed from responses, and responses were normalized to the range [0 1] by dividing by the maximum response of each given neuron across all presentations. This follows the approach described in [172].

I applied the same methods I have been applying for multivariate pattern analysis of fMRI data, namely linear support vector machine classification. I implemented a three-fold cross-validation, whereby I kept one example for each image as the test set, and trained on the remaining two examples from each image, at each fold. I tried to decode, in turn, the viewpoint and the identity of faces. The results for the full image set are presented in **Figure 84**. The bar graphs (**Figure 84**, top) report the fraction of test examples that were correctly labeled; chance level is marked by a dashed line, and a 95% confidence interval derived from a permutation test is plotted as a vertical line (1000 surrogate accuracies, obtained with the same procedure but with shuffled class labels). Viewpoint can be decoded above chance level in all three studied patches. So can identity, but there is a clear increase in classification accuracy from the most posterior patch (ML/MF) to the most anterior patch (AM). More insight can gained by looking at the confusion matrices (**Figure 84**, bottom). The rows represent the true labels of the test examples (for viewpoint, the order is the same as in **Figure 83**); the columns represent the predicted labels of the test examples. For

instance, it can be seen that the population of neurons in ML/MF confuses the full and half profiles. In contrast, the population in AL confuses right and left full profiles (mirror symmetry). In AM, there is much confusion between all profile views, but they can still be distinguished from the frontal view quite easily.

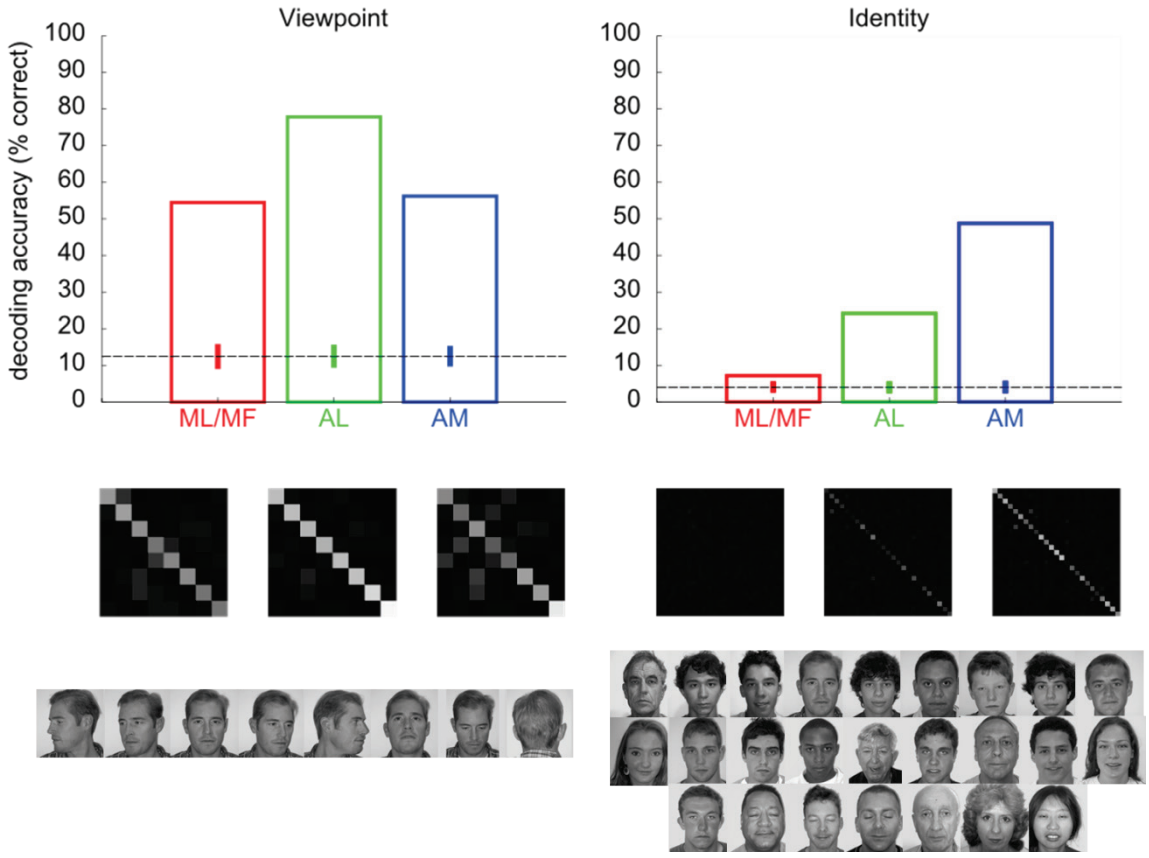


Figure 84 Single unit decoding of viewpoint and identity, for the full face views dataset. The set comprised eight viewpoints and 25 identities. Left: viewpoint decoding. The bars represent the accuracy of multiclass decoding in the three face patches. Chance level is the dashed line (100/8=12.5%); the 95% confidence intervals from a permutation test (1000 surrogates) are shown as colored vertical lines, roughly centered at chance level. Below, the confusion matrices are shown for each patch; rows represent the true labels, and columns the labels predicted by the classifier (the order of labels in the confusion matrix is shown below). Correct classifications thus fall along the diagonal. Right: identity decoding.

Doris and Archy picked four identities (randomly I was told) from the face views image set, and five viewpoints (left full, left half, frontal, right half, right full), to perform a fMRI experiment. The set of 20 images that they ended up with is shown in **Figure 85**.

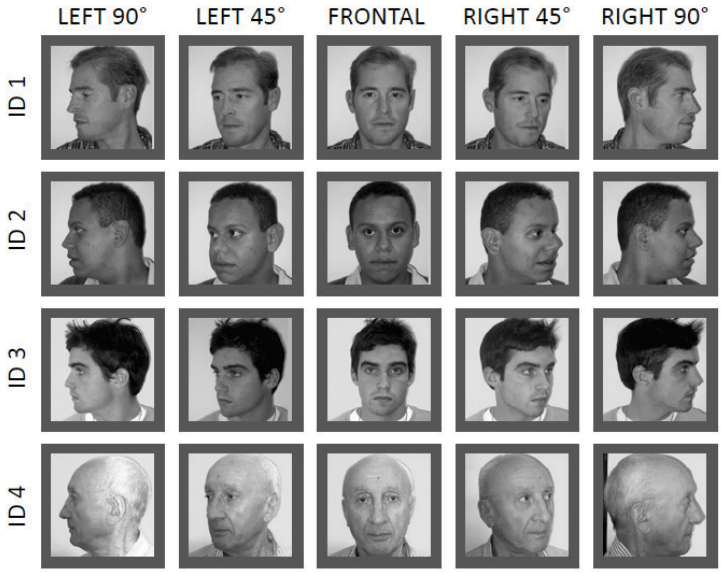


Figure 85 The 20 images from the face views image set which were used in the fMRI experiments. Four male identities are pictured at five different viewpoints.

I repeated the same analysis, but kept to these 20 images. The results are shown in **Figure 88A** and **Figure 89A**. These results represent a ground truth against which I could compare the fMRI decoding results.

The fMRI experiments were run on different monkeys (two monkeys, M4 and M5). This could be seen as a weak point in this study; our strong claim is that we know what the underlying units are doing, but since the single unit data comes from other monkeys we cannot be absolutely sure. Of course, experience has taught us that these face patches are very reproducible from one monkey to the next in the fMRI data, and there is no reason to think that the neurons in those patches should do anything different from one monkey to the next. To add credence to this claim, I analyzed a dataset of neurons that had been recorded from the AM patch of one of the monkeys who participated in the fMRI experiments (M5). 28 units were recorded from while M5 was presented with a large set of face images, including a set of 16 familiar identities at three different viewpoints (straight, half right profile and full right profile). A linear Support Vector Machine analysis (**Figure 86**) shows that the population is informative with respect to viewpoint (more

specifically, it can distinguish the frontal view apart from the profile views quite easily), and with respect to identity. The rather low performance for identity, compared to the accuracies obtained previously, is likely due to the limited number of units that we had to work with (28 here, vs. 159 previously).

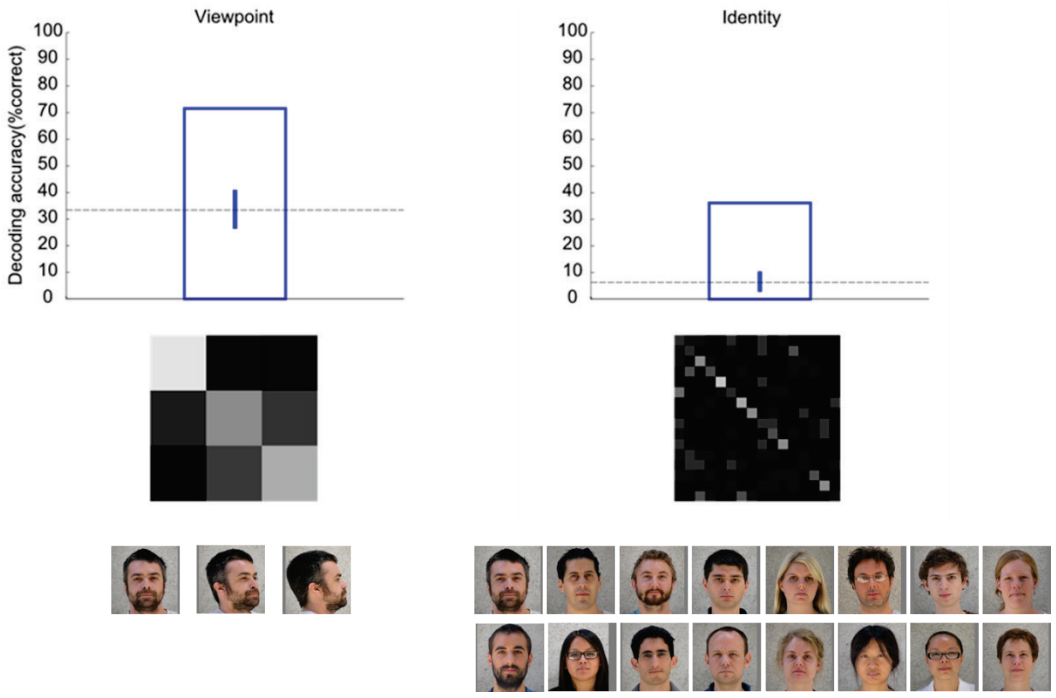


Figure 86 Decoding of viewpoint and identity in M5's AM patch, from 28 single units, with a dataset comprising 16 familiar individuals pictured at three different viewpoints. Left: viewpoint decoding. Right: identity decoding. Same layout as in **Figure 84**.

2. fMRI MVPA retrieves viewpoint information but fails for identity

The design of the fMRI experiment was very simple. Monkeys were scanned in a 3T Siemens Tim Trio. This was a different scanner from the one that we used for human research, but the exact same model (actually, after this new scanner was installed it was used for human research for a few months; some of the data that I reported in this thesis for human subjects came from the same scanner). Monkeys passively viewed images on a screen, while their eye position was monitored using an infrared eye tracking system (ISCAN), and a juice reward was delivered

every two to four seconds if fixation was properly maintained. The fixation spot size was 0.13° in diameter. We used a multi-echo sequence (EPI, TR 2s, TE 30ms, 96 x 96 matrix, 1 mm isotropic resolution). Note that this is much higher resolution than the typical human fMRI experiment (in this thesis I have mostly presented human data acquired at $2x2x2mm^3$, which is considered high resolution for human fMRI). In combination with a concomitantly acquired fieldmap, this allowed high fidelity reconstruction by undistorting most of the B0-field inhomogeneities. MION (Monocrystalline Iron Oxide Nanocolloid) contrast agent was used to improve signal/noise ratio[173] (**Figure 87**) (Doris does this systematically); MION is one type of superparamagnetic iron oxide, which enhances proton relaxation, and has the desirable effect of reducing the $T2/T2^*$ relaxation time.

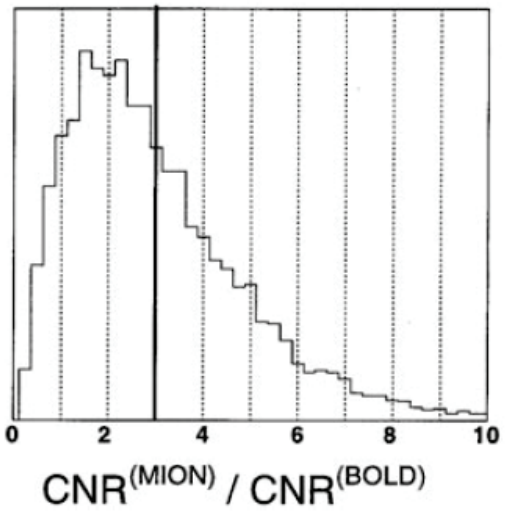


Figure 87 The Contrast to Noise Ratio (CNR) using MION compared to using BOLD (no contrast agent); on average, the CNR is three times higher with MION. Reproduced from[173].

Images presented on the screen spanned 4.7° of visual angle. 24-second blocks of a gray background alternated with 24-second blocks of one of the 20 images in our image set (**Figure 85**). During a given image block, the same image was presented throughout, and its position was jittered slightly (0.2 deg) every two seconds to prevent visual adaptation. We got ten good fMRI

runs for M4 and M5. During each run we presented ten images, hence it took two runs to present all 20 images. The order of images was fixed (run A: id2, straight; id1, left full; id4, straight; id3, straight; id2, left half; id4, left half; id3, right half; id2, left full; id1, right full; id4, right half; run B: id3, right full; id1, straight; id1, left half; id1, right half; id4, left full; id3, left full; id2, right half; id4, right full; id2, right full; id3, left half). A functional localizer to define the face patches was run in a separate fMRI session, as described earlier. After preprocessing (realignment and distortion correction with Freesurfer) I analyzed the data in the classical way with a GLM. I used a custom hemodynamic response function, since MION leads to a differently shaped hemodynamic response than BOLD fMRI (parameters estimated by J. Mandeville):

$$hrf = \left(\frac{t-\delta}{\tau}\right)^{\alpha} e^{-\frac{t-\delta}{\tau}} \text{ with } \delta = 0, \tau = 8 \text{ and } \alpha = 0.3.$$

The contrast faces (monkey and human) vs. other categories (fruits, bodies, hands and objects) yielded clear activations, which I thresholded at $p<0.0001$ to define the face patches. The numbers of voxels in each patch are shown in **Table 3**.

	Left hemisphere			Right hemisphere		
	ML/MF	AL	AM	ML/MF	AL	AM
M4	165	95	27	181	140	68
M5	326	156	58	472	156	58

Table 3 Size of clusters at $p<0.0001$ used to define the face patches in M4 and M5 (number of voxels)

The EPI data for the MVPA experiment was realigned to the first run and corrected for distortions caused by magnetic field inhomogeneities using Freesurfer. The short TR (two seconds) did not warrant the application of slice timing correction. Following the approach used in Frank Tong group (and described for example in [174]), we first detrended the time course of each voxel in each run using a second order polynomial, then z-scored the signal (using the mean and standard deviation across time). Then, we took the average of time points 16, 18, 20, 22, 24 and 26 seconds (which encompasses the peak of the fMRI response) as the signal for each block. We

extracted the signal for each block at each voxel (n_{vox} = 96 x 96 x 54), thus populating a (n_{blocks} x n_{vox}) matrix. We then selected the columns of this matrix that corresponded to each functionally defined region of interest (face patches), or to each searchlight for whole brain analyses, and fed this data to a classifier. We found that the SVM classifier performed above chance for viewpoint classification in all three patches for both monkeys (**Figure 88**). Critically, the confusion matrices were in very good agreement with the confusion matrices that we obtained with the single units.

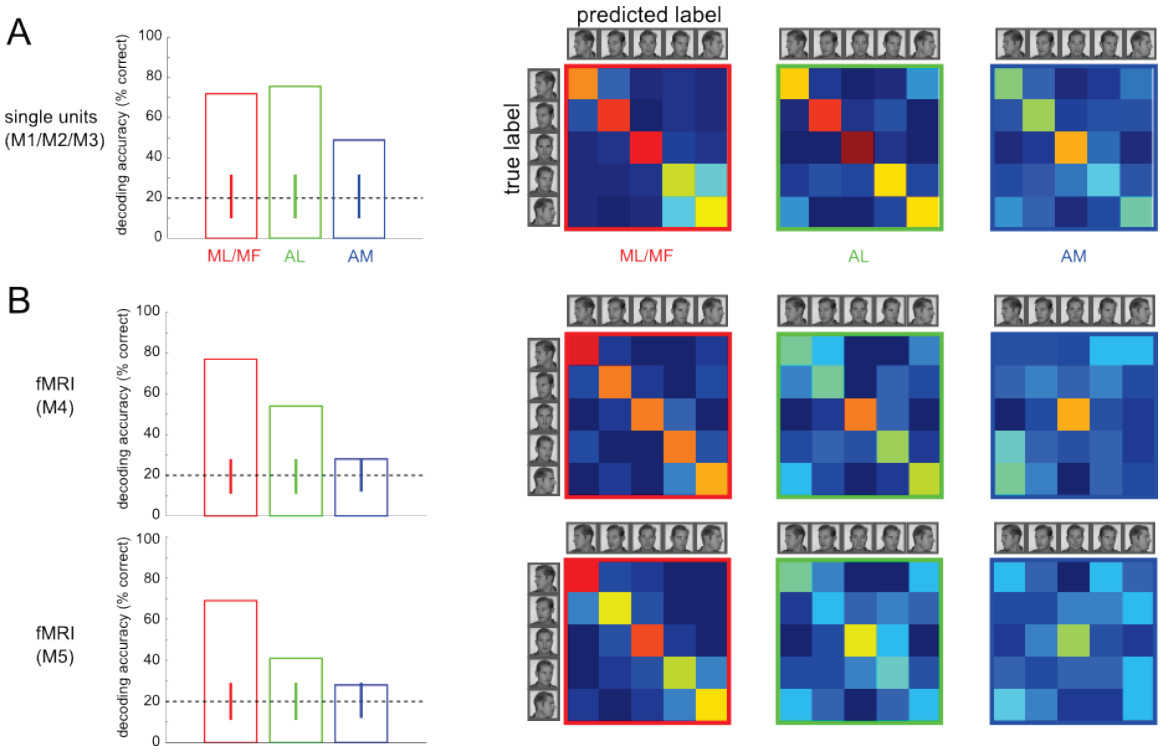


Figure 88 Decoding of viewpoint with single unit and fMRI data. A) Left, decoding accuracy in the three face patches using the data from the single unit recordings. Chance is indicated with a dashed black line. The 95% interval from a permutation test (1000 surrogates) is shown as a vertical line for each patch. Right, confusion matrices for each patch. Rows represent the true labels (ordered from full left to full right profile) and columns represent the predicted labels. B) Same as A, using fMRI data. The results are shown separately for the two monkeys, M4 and M5. Note the nice correspondence between the confusion matrices of the single unit data and fMRI in ML/MF and AL, especially the mirror symmetry in AL (whereby left and right profile representations are difficult to tell from each other) (I quantified the correspondence with Spearman correlation coefficients ρ, and assessed their significance with a permutation test; M4: ML/MF ρ=0.705, $p<10^{-3}$, AL ρ=0.780, $p<10^{-3}$, AM ρ=0.584, $p=3.9 \times 10^{-3}$; M5:ML/MF ρ=0.728, $p<10^{-3}$, AL ρ=0.785, $p<10^{-3}$, AM ρ=0.349, p=0.073).

This showed that viewpoint information, where present in single units, could be read out with fMRI. At last, fMRI retrieved information represented in the underlying neurons! Note the nice

mirror selectivity in AL in the fMRI decoding results, which we had found already in the single unit data.

As for identity, the neural information resisted my attempts to access it with fMRI (**Figure 89**). I tried many tricks, growing and shrinking the face patches and using various flavors of feature selection, to no avail. While 159 cells in AM had enough information to perform almost perfectly at identity classification, this could not be retrieved with fMRI MVPA.

For another take on this, I conducted a representational similarity analysis[175], comparing the information contents of the single unit recordings and of the fMRI data. The first step was to generate representational dissimilarity matrices (RDMs) for the single unit recordings and for the fMRI data by computing the distances between the responses elicited by the different images in the image set. The distance measure most classically used is based on the Pearson correlation. I also investigated a new distance measure, based on the distance to the separating hyperplane in a linear SVM one against one classification; the rationale is that such a measure is less sensitive to uninformative dimensions (correlation weighs all dimensions equally), and finds the maximum achievable distance (with a linear classifier). The RDMs for the two types of recordings and with the two distance measures are shown in **Figure 90** (only monkey M4 is shown for the fMRI).

Before proceeding further, I had noticed that the images appeared to have a fair amount of low level differences, which may confound our conclusions. Most obviously, the pictures of the older man (ID 4) were significantly brighter than the other identities. Note that since these were the images used in the single neuron recordings, it would not have made sense to try and equalize them in the fMRI experiments. We could only acknowledge the confound and try to compensate for it. To quantify the low level differences in a biologically relevant way, I ran a computational model of V1 (the first stage of the H-MAX model, described in [176]) on the image set. First, following the approach used by Tim Kietzmann[174], a PhD student with Frank Tong, whom I met

at VSS, I foveated the images (taking into account the visual angle that they covered in the experiment, and using a toolbox available at http://svi.cps.utexas.edu/). Then, I convolved the

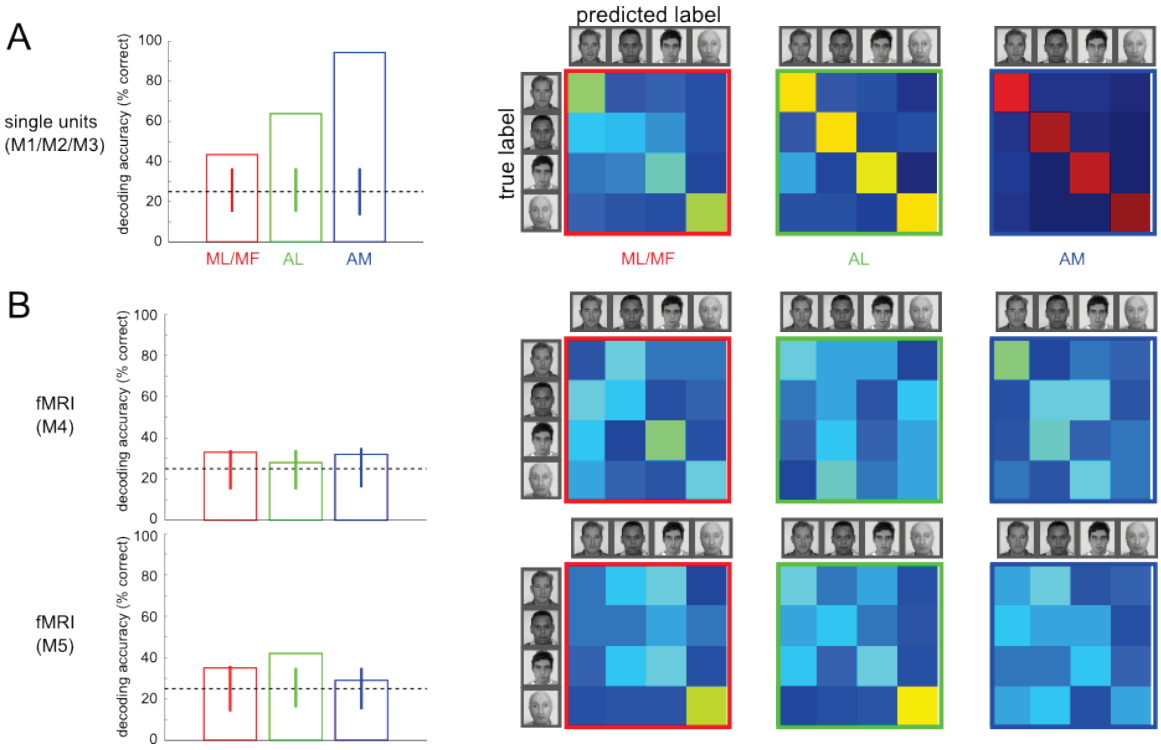


Figure 89 Decoding of identity with single unit and fMRI data. A) Left, decoding accuracy in the three face patches using the data from the single unit recordings. Chance is indicated with a dashed black line. The 95% interval from a permutation test is shown as a vertical line for each patch. Right, confusion matrices for each patch. Rows represent the true labels (ordered from full left to full right profile) and columns represent the predicted labels. B) Same as A, using fMRI data. The results are shown separately for the two monkeys, M4 and M5. Note the very good classification of ID 4 in ML/MF and AL in the single unit data and M5's fMRI data, which I attribute to obvious low level differences. Importantly, there is no significant retrieval of identity information in AM in the fMRI data, whereas AM represents identity almost perfectly in the single unit data.

foveated images with a set of gabors, at different scales and orientations, which can be seen in **Figure 91A**. The results of the convolutions were concatenated into a huge vector, and a Pearson correlation based RDM was generated (**Figure 91B**). I was not surprised to find that ID 4 really stood out, which was best seen with the RDM arranged by identity rather than viewpoint (as it was in **Figure 90**).

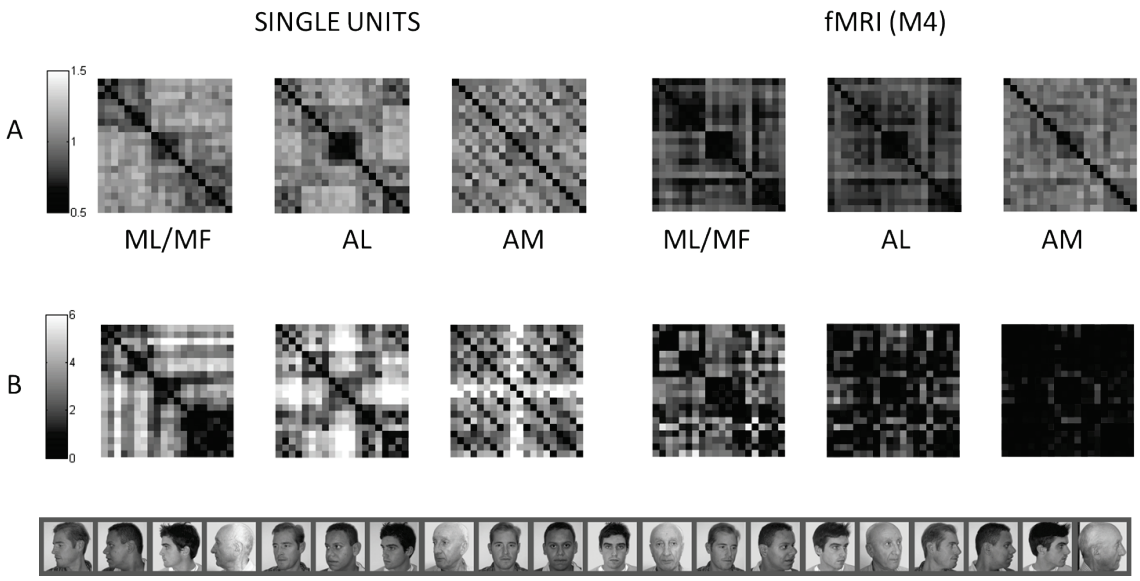


Figure 90 Representational Dissimilarity Matrices (RDMs). The distances between all pairs of images (in the order depicted at the bottom) are computed from the single unit data (left) and from the fMRI data of monkey M4 (right): A) using a Pearson correlation based distance measure, as in [175] and B) using the distance from the separating hyperplane in a linear SVM one vs. one decoding framework (measure that I introduced). Diagonal values were set to zero. Some patterns emerge clearly from these RDMs, such as the mirror symmetry in AL (darker top right and bottom left corners), and the identity coding in AM (dark diagonal stripes).

The representational dissimilarity obtained with this V1 model allowed us to try and correct for low-level differences, by regressing out the V1 model RDM from the single unit RDM and the fMRI RDM, respectively, before quantifying their similarity. This remains a hack, since the V1 model[176] I used is a simplification of the output of V1; but it captured some of the low-level information present in both experimental RDMs. So were the single unit and fMRI RDMs related? Of course they were; it can be seen by eye already in **Figure 90**. I computed the Spearman correlation between the upper triangular parts (without the diagonal) of the corrected single unit and fMRI RDMs (**Figure 92**). However, the most critical aspect was to understand whether the positive correlation came from identity information or viewpoint information. To determine to what extent identity and viewpoint information contributed to the correlation, I selectively shuffled either the identity information (e.g., exchanging the labels of "id1, left full" and "id4, left full" keeps viewpoint information while disrupting identity information) or the

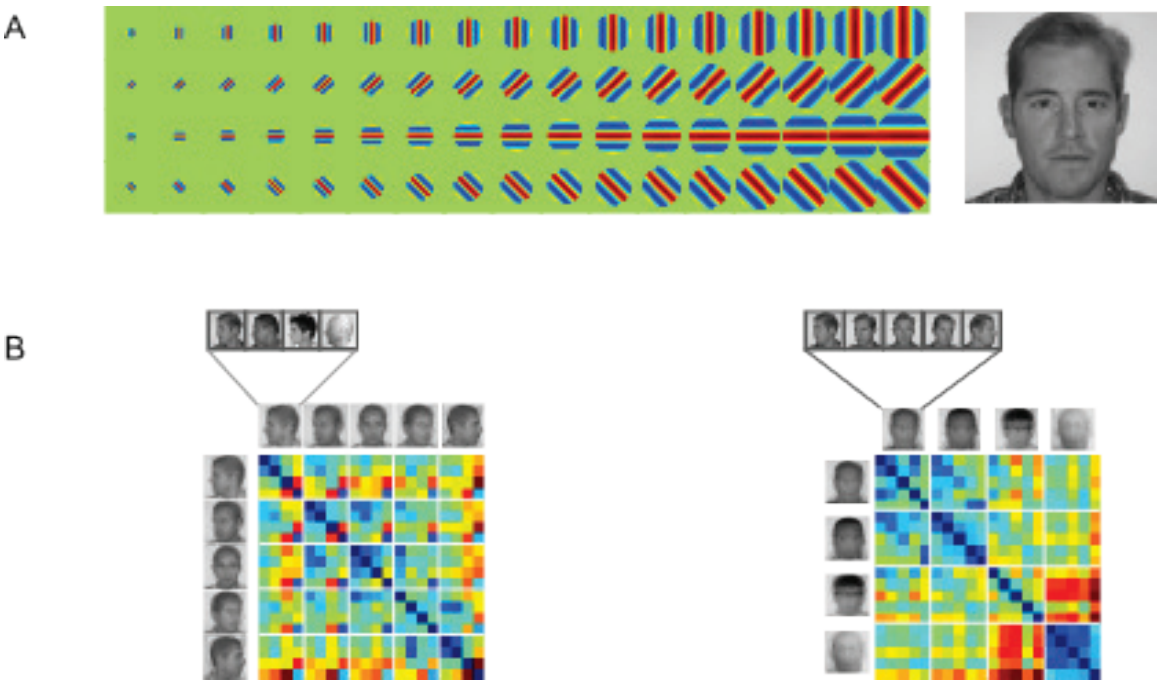


Figure 91 Simple model of V1 and corresponding representational dissimilarity matrix. A) Left, the bank of gabor filters at 17 scales and four orientations constituting the V1 model[176]. Right, a face stimulus from the experiment, shown at the same scale as the gabor filters. B) Left, representation dissimilarity matrix (distance metric is Pearson correlation based), sorted by viewpoint. Right, representational dissimilarity matrix sorted by identity. I am thankful to Tim Kietzmann for sharing some Matlab code with me to perform this analysis[174].

viewpoint information. I report the average correlations obtained over 1000 such selective randomizations in **Figure 92**. We found that the correlation between RDMs appeared to be driven almost entirely by viewpoint information: interfering with this information led to correlations that were no different from chance, whereas interfering with identity information hardly affected the correlation.

These two approaches (linear SVM classification and representational similarity analysis) converged to show that viewpoint information is well represented in the fMRI data, but identity information was very difficult to pick up (I'll stick with saying very difficult rather than impossible, because with methods evolving it may happen in the future… also, as we say where I come from: "Impossible n'est pas français!").

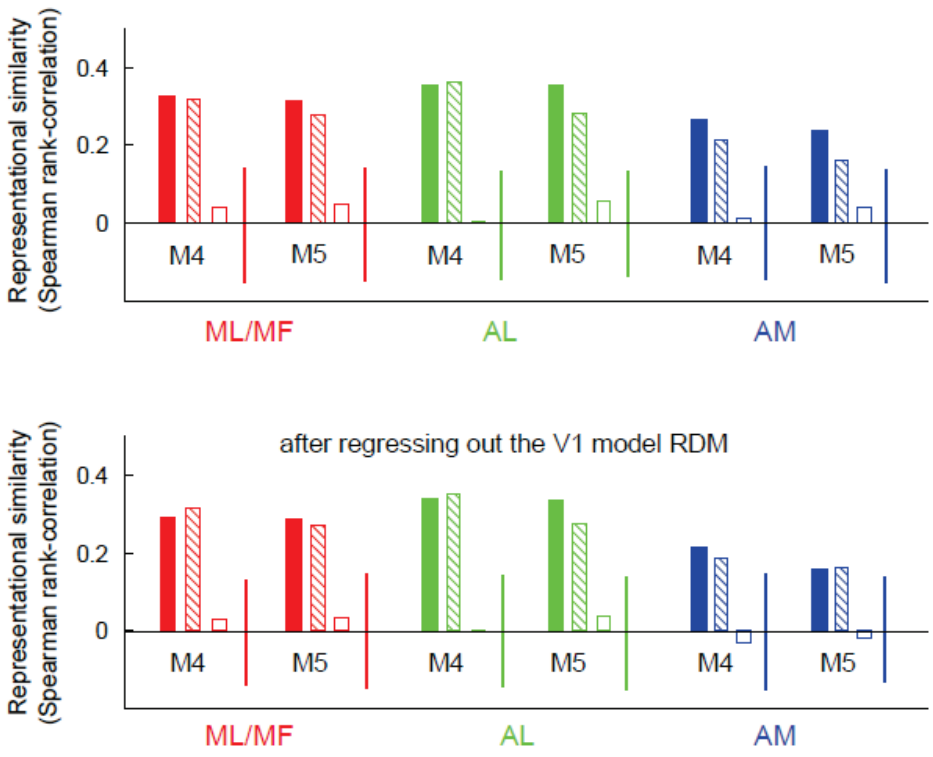


Figure 92 Top: Spearman rank correlation between single unit RDM and fMRI RDMs (filled bars), before regressing out the V1 model RDM. After selectively shuffling identity information, the Spearman correlation was almost unchanged (hatched bars). However, selectively shuffling viewpoint information severely disrupted the Spearman correlation (empty bars). The vertical lines are the 95% confidence interval obtained with a permutation test (1000 surrogates). Bottom: same, after regressing out the V1 model RDM.

3. Clustering is key

Our results so far beg for an explanation: why is the decoding of identity with fMRI MVPA so much harder than the decoding of viewpoint? A few reasons may come to mind.

fMRI typically yields noisier measurements in the anterior temporal lobes than in more posterior cortical areas; since invariant identity information lies mostly in anterior areas, while viewpoint information lies in posterior areas, this could explain the dissociation that we observed. I quantified the functional signal-to-noise ratio (fSNR, defined within a GLM framework as the magnitude of the fMRI signal change divided by the standard deviation of the residuals across time) for each patch in both monkeys (**Figure 93**). While I found that the fSNR generally

decreased as one moved from ML/MF to AL then AM, the fSNR for AM in both monkeys was comparable to the fSNR for ML/MF and AL in M5. As shown earlier, the rather low fSNR in M5's ML/MF and AL did not prevent the classifier from performance well above chance for viewpoint classification. Additionally, AL was shown to carry significant information both about viewpoint and identity in the single unit data; only viewpoint information appeared to be retrievable with fMRI, hence the reason for the failure of identity decoding is to be found elsewhere.

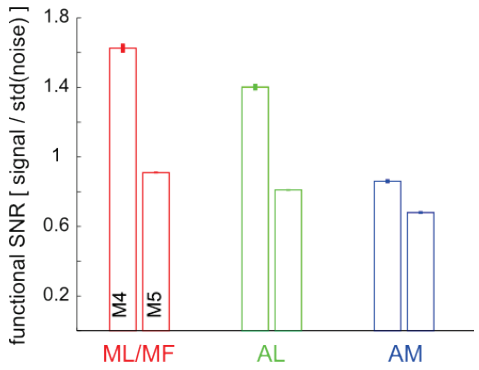


Figure 93 Functional signal-to-noise ratio (fSNR) in the regions of interest, for M4 and M5 (mean across ten functional runs, and standard error). The functional SNR was computed using the output of a General Linear Model, as the average of parameter estimates for face block regressors divided by the standard deviation of the residuals. Note that fSNR generally decreases from posterior (ML/MF) to anterior (AM) areas. However, the fSNR in AM for monkey M4 is comparable to the fSNR is ML/MF and AL for monkey M5; since these two areas supported decent decoding in M5, fSNR cannot be the main cause for the poor decoding of identity in AM.

The signal measured in fMRI is hemodynamic, and an obvious prerequisite for a sizeable hemodynamic response is that enough neurons be active in a given area. With this in mind, we looked at the sparseness of neural representations for identity and viewpoint in our three regions of interest. I computed the Gini index[177] on the basis of the normalized average responses to each image in the face views set, for all neurons in a given patch (**c**). The normalized responses represent how strongly each neuron respond to each image, as a fraction of their maximal response (in the image set); the response is sparse if only a few neurons respond near their maximal rate.

$$G = 1 - 2\sum_{k=1}^{N} \frac{c_{(k)}}{\|c\|}\left(\frac{N-k+\frac{1}{2}}{N}\right) \text{ for ordered data, } c_{(1)} \leq c_{(2)} \leq \cdots \leq c_{(N)}.$$

We found (**Figure 94**) that the sparseness of both identity and viewpoint representations (as computed after averaging responses across viewpoints and identities, respectively) increased significantly from ML-MF to AM; the representation of viewpoint was slightly sparser than the representation of identity in AL. Hence, though sparseness may be a factor in the difficulty to retrieve information from AM's fMRI patterns, it still cannot account for the discrepancy in retrieving information about viewpoint vs. identity in AL.

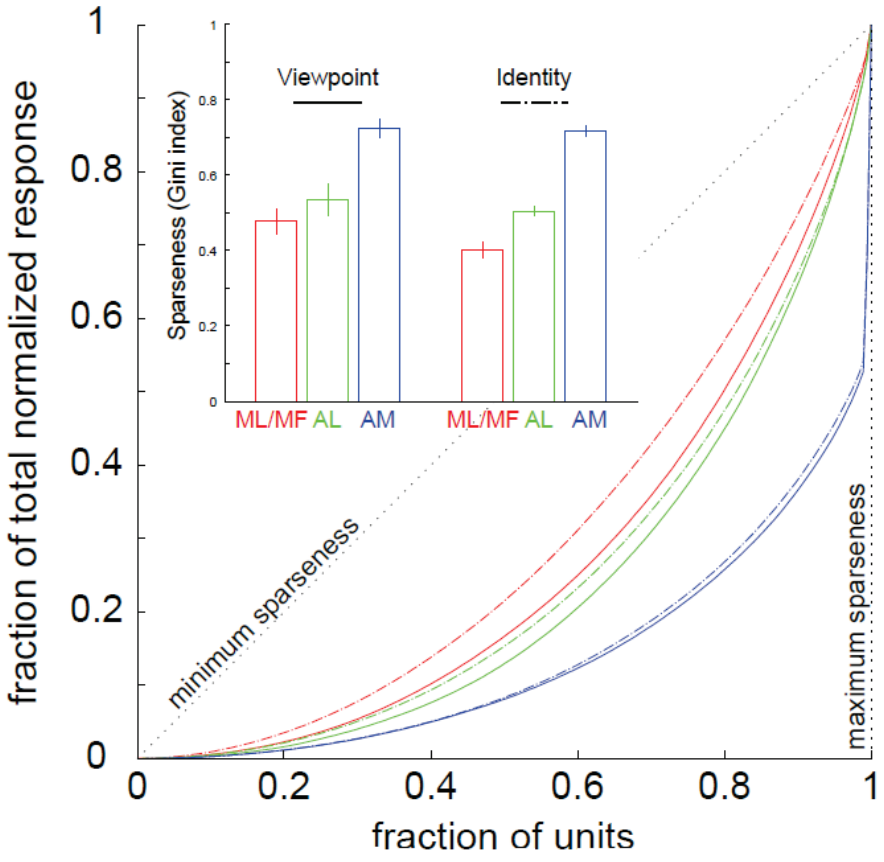


Figure 94 Sparseness of the neuronal representations of viewpoint (solid lines) and identity (dash-dotted lines), computed from the single unit data (using all faces in the face views image set). The Gini index (bar plot, inset) corresponds to twice the area below the diagonal when plotting the fraction of the total response against the fraction of units (main plot). Sparseness increases from posterior to anterior regions, but identity representations are no sparser than viewpoint representations.

Even if a code is sparse, it may still give rise to a sizeable hemodynamic response if the active units are tightly clustered. If informative units are scattered rather than concentrated, it is less likely that the hemodynamics will carry much information about the underlying representation. I looked at whether units recorded within 1mm of each other had similar tuning to viewpoint and identity in each of the three regions of interest (**Figure 95**). I found strong clustering in AL for viewpoint, but no evidence for clustering of identity tuning; in AM, clustering was weak for viewpoint and identity, compared to viewpoint clustering in AL. I believe that this is the main reason why identity information fails to be picked up with fMRI.

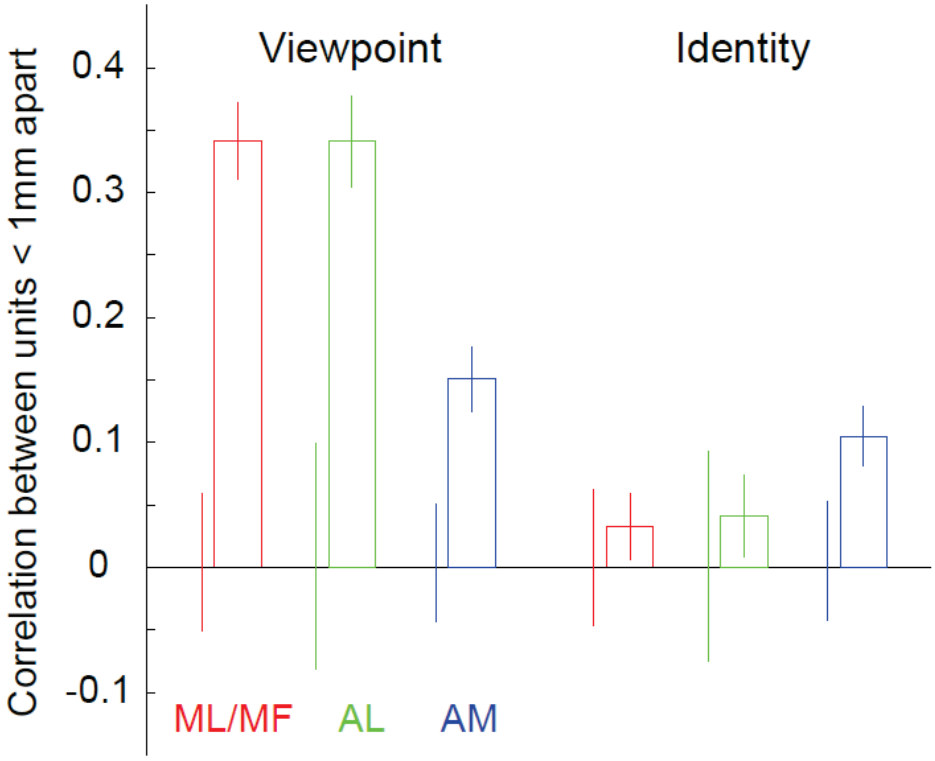


Figure 95 Clustering of single unit responses. The correlation of responses of neighboring units (≤1mm) was assessed, across viewpoints and across identities, in the three regions of interest. Viewpoint selectivity in ML/MF and AL is very clustered, while identity selectivity does not show above chance clustering. In AM, both viewpoint and identity selectivity are clustered, but to a much lesser extent than in ML/MF or AL. A 95% confidence interval for the distribution of chance was estimated with a permutation test (1000 surrogates) and plotted as vertical lines. Error bars are s.e.m.

4. What about fMR adaptation?

Before the advent of fMRI multivariate pattern analysis, the technique for studying underlying neuronal representations with fMRI was the fMR adaptation paradigm[178]. The oversimplified idea is as follows: imagine that a fMRI voxel contains two populations of neurons equal in size, one that is tuned to Brad Pitt, and one that is tuned to Tom Cruise (to come back to a cherished example). If you show pictures of Brad Pitt, the response of the voxel will be at a given level Y. If you show pictures of Tom Cruise, the response of the voxel will be at roughly the same level Y, because just as many neurons (but, different ones) are active for that condition. Hence, a simple fMRI contrast between pictures of Brad Pitt and pictures of Tom Cruise will not lead to any sizeable effect in this voxel, even though the underlying neuronal populations perfectly distinguish the two actors. Then comes fMR adaptation; by repeatedly presenting pictures of Brad Pitt, you can adapt the neuronal population coding for Brad Pitt. Hence, a subsequent trial in which you show Brad Pitt again will lead to a rather weakened fMRI activation. However, a picture of Tom Cruise should still lead to an activation of the same magnitude as before adaptation, since the population tuned to Tom Cruise has not been adapted. This is a very clever way to determine whether there are populations of neurons tuned to certain stimuli in a fMRI voxel.

Doris and Archy also ran a fMRI experiment in view of studying fMR adaptation's ability to retrieve viewpoint and identity information in the same two monkeys, M4 and M5. It was again a fMRI block design, with 24-second blocks of faces separated by 24-second block of fixation. There were four block types: A) Fixed Identity, Fixed Viewpoint; B) Variable Identity, Fixed Viewpoint; C) Fixed Identity, Variable Viewpoint; and D) Variable Identity, Variable Viewpoint (**Figure 96**, top). According to the logic of fMR adaptation, if a voxel comprises neurons that are sensitive to the identity of the faces presented, then the haemodynamic response should be larger in conditions B and D than in A or C; similarly, if a voxel comprises neurons that are sensitive to

viewpoint, we should see a release from adaptation in blocks C and D compared to blocks A and B (**Figure 96**, bottom). The blocks were presented in a counter-balanced manner, with every identity and orientation presented in each block type. We collected 10 good runs for M4 and 12 for M5.

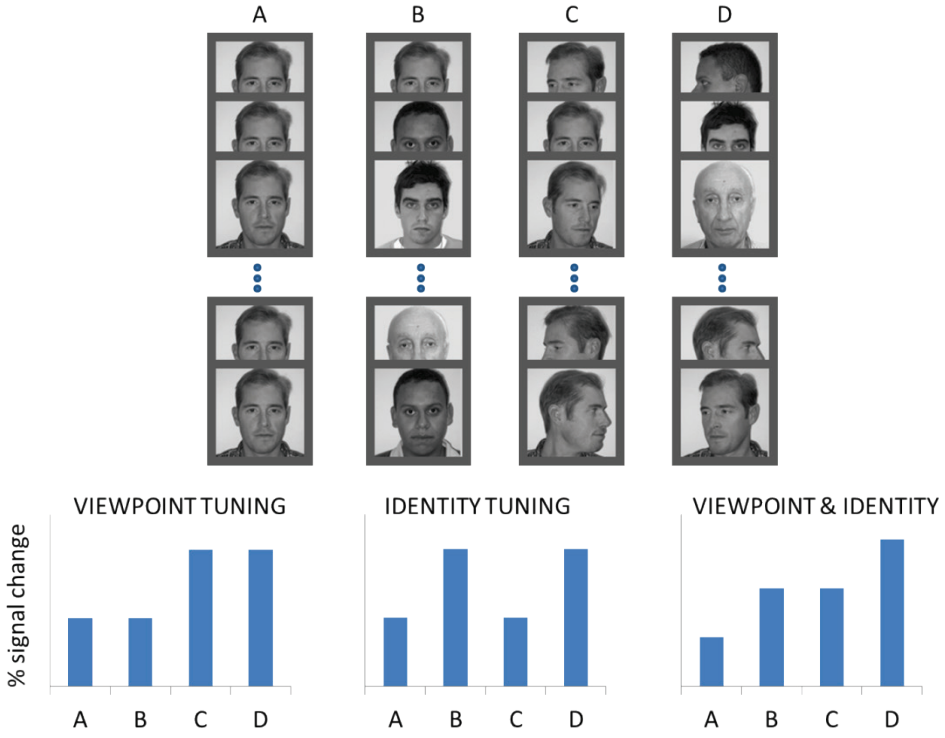


Figure 96 fMR adaptation paradigm. Top: each run had four block types. A=same identity, same viewpoint; B=different identity, same viewpoint; C=same identity, different viewpoint; D=different identity, different viewpoint. Bottom: predictions. If there are underlying neuronal populations that are tuned to different viewpoints within a voxel, the response to blocks with the same viewpoint throughout (A and B) should be less than the response to blocks with varying viewpoint (C and D). Similarly, if there are underlying neuronal populations tuned to identity, one would expect same identity blocks (A and C) to yield lower activations than varying identity blocks (B and D). Finally, one can imagine a mixed situation in which there are both populations tuned to different identities and to different viewpoints. In that case, we expect A to show the most adaptation, i.e., the lowest activation, and D to show the largest activation, while B and C will be somewhere in between (not necessarily equal).

The data was analyzed in a classical fashion, with a General Linear Model and mass-univariate approach (again, the HRF was a custom HRF for MION fMRI). I performed both whole-brain analyses and ROI analyses in the face patches. The results for ROI analyses for both monkeys M4 and M5 are plotted in **Figure 97**. In essence, we did not find any significant differences between

conditions that made sense with our expectations and were consistent across both monkeys. We were forced to conclude that our fMR adaptation paradigm failed to uncover either of the viewpoint and identity information. However, this may be a power limitation for this study rather than a true negative result.

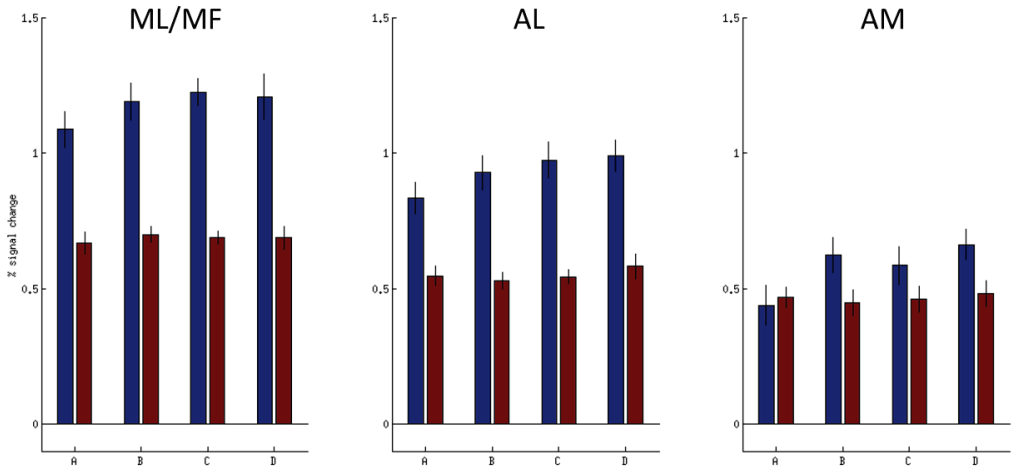


Figure 97 ROI analysis of fMR adaptation experiment for both monkeys (M4: blue; M5: red). The error bars are the s.e.m. across runs. M4 may show results that are compatible with the predictions (mixed tuning in AL, identity tuning in AM), but M5 does not. Our paradigm is likely underpowered.

Discrepancies between fMRI MVPA results and fMR adaptation results have been reported elsewhere[179,180]. I do not wish to claim that I have demonstrated such a discrepancy here, however. Our adaptation paradigm was likely too simplistic, for example, failing to control attention, which has great effects on fMRI activations. Perhaps a better way to perform adaptation would have been to selectively adapt a population, for instance with a B block (constant viewpoint), and then present a test trial (either the same viewpoint, or a different viewpoint) and compare the responses to these two test trials. I believe this may be worth investigating further (and I may do so).

How much more knowledgeable are we about the ability of fMRI to retrieve underlying neural representations? In the second section of this chapter, we finally reported a positive result: that fMRI decoding picked up viewpoint information, there where it was available in the underlying single unit populations. This is a great achievement. It also allowed us to start uncovering some key factors contributing to the successful transfer of a neural code to hemodynamic signals; mainly, the clustering of neurons tuned to the same properties appears to be essential. Surely, it makes sense; but I believe that this empirical demonstration is an important contribution.

What this also brings forward is that fMRI MVPA is not magical. When it was introduced and popularized, around the start of my PhD, it was almost presented as a technique that would allow fMRI to read out any underlying neural representation. How many times, at the end of a fMRI talk, have I heard the question, “Have you tried decoding?” Granted, fMRI MVPA is a fairly sensitive technique and it may pick up some information in distributed patterns where univariate analyses fail. But it still relies on the hemodynamic signal, and some representations simply cannot elicit a sizeable hemodynamic response. Period.

V. THE BOLD PROMISE

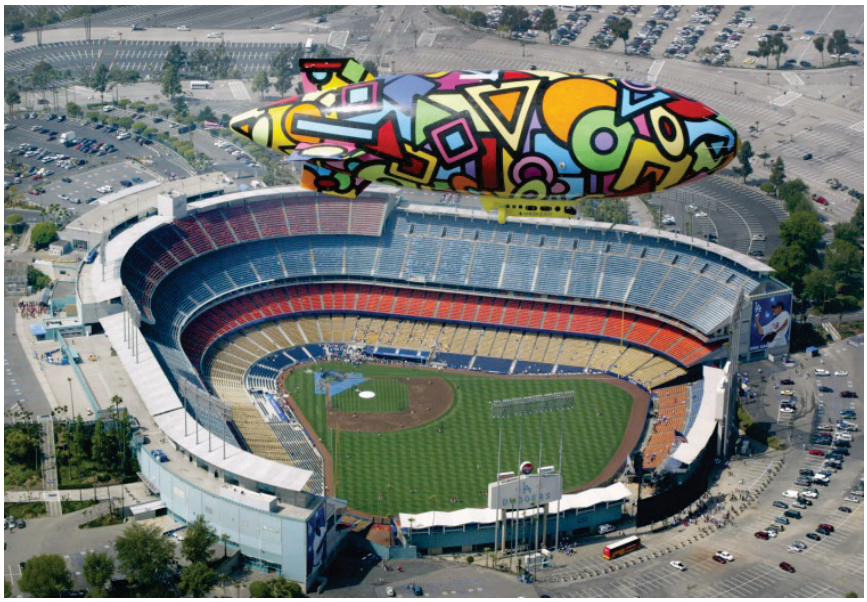


Figure 98 A colorful blimp flying over Dodger Stadium. A metaphor for bulk tissue technologies, used by Christof in his Quest[59] . Source: Getty Images.

> "Attending a poster session at a recent meeting, I was reminded of the old adage 'To the man who has only a hammer, the whole world looks like a nail.' In this case, however, instead of a hammer we had a magnetic resonance imaging (MRI) machine and instead of nails we had studies. Many of the studies summarized in the posters did not seem to be designed to answer questions about the functioning of the brain; neither did they seem to bear on specific questions about the roles of particular brain regions. Rather, they could best be described as 'exploratory.' People were asked to engage in some task while the activity in their brains was monitored, and this activity was then interpreted post hoc." [181]

These are the first sentences of a 1999 paper by S. M. Kosslyn, and there is much truth to them. MRI technology is really cool, and the classical visuals that come out of fMRI analyses, blobs of activity on a gorgeous anatomical picture of the brain, are mesmerizing. As MRI technology has become widely available for brain research, it is important to take a step back and think of what we can learn from this technique. I myself was guilty of trying to use the same hammer in the

various experiments that I conducted, hoping that it would be a sufficient tool. For my defense it was a brand new shiny hammer, the fMRI decoding hammer, which had worked wonders for other people in the testing phase.

When I came to work with Christof, I wanted to study the neural bases of consciousness. That is when I was handed the hammer for the "neural" part. It took me a little while to get accustomed to using it, and I did not reflect right away on what it was good for; I did my best to study the owner's manual in-depth first. The first nail that I hit with it was a fascinating question, that of the effects of attention on unconscious processing. Understanding consciousness is the question that comes back most often when people (lay people and neuroscientists alike) are asked what the biggest unsolved mystery about the brain is. In a recent talk for the Caltech Alumni association, Ralph suggested that it is unclear that we will ever understand consciousness. It is quite clear to me now that the answer, if it comes at all, will not come from the single-handed use of fMRI in humans. Christof's decision to start working on a slightly simpler system (the mouse brain rather than the human brain), for which invasive techniques are an option and an array of genetic manipulations are available, makes a lot of sense for his Quest.

Initially, however, I did not think my hammer was to blame; rather, I decided that the whole field of unconscious processing presented too many methodological challenges, and that the literature was fraught with non-replicable findings. My focus shifted to questions that I thought would be less controversial. I kept using my hammer; since I had invested much time already learning to use it, I wanted to get some return on my investment. Performing fMRI at Caltech is made so easy that it would be a shame not too take advantage of it: the facilities are available 24 hours a day, seven days a week, and I had the training to operate the scanner myself, the only requirement being to have another person in the console room (and an account to charge scanner hours to). One of the strengths of fMRI is to provide a picture of what is happening in the whole brain at a fairly decent spatial and temporal resolution, hence it is well suited for studying networks of brain

areas and their interactions. When I undertook the person recognition project, I had a few areas of interest in mind from previous literature (hippocampus, anterior temporal lobes), where I expected to be able to read out information about person identity; but I also wanted to start understanding how they interacted. Of course, as you know if you have read the previous chapters, these ideas were not too successful; fMRI could not pick up invariant representations of person identity to start with, hence I could not delve into further questions.

This is when I started doubting the power of my hammer. Why had fMRI, with its cool decoding add-on, been failing me so far? When Florian showed me the highly significant finding in the human right amygdala of a response to the animal category that was larger than the response to other categories of images, I thought that that finding would surely be reproducible with fMRI, more so than the Jennifer Aniston neuron finding. Unfortunately, the fMRI evidence for a categorical response to animals in the right amygdala was weak, at best. Before throwing my hammer out the window in frustration, I had the opportunity to put fMRI to the test again when Doris approached me with a beautiful single cell finding in macaque monkeys: the differential representation of face identity and viewpoint in the different face patches[172], and an already collected fMRI dataset that needed to be analyzed. There, I showed a dissociation in the ability of fMRI to pick up the representations of identity and viewpoint; I was finally able to evidence some of the factors that contributed to a neural representation being picked up by fMRI, e.g. clustering of like-tuned neurons. "Elementary my dear Watson[§§];" fMRI measures hemodynamic activity.

By now, the reader should have a fair understanding of what fMRI measures; I went through the physics of the fMRI signal in the first chapter, then described the application of fMRI in a few settings throughout this thesis, and put it to the test against measurements of single neuron activity. fMRI measures a hemodynamic signal whose spatial and temporal specificity are

[§§] As I was looking up this phrase, I found out that it was never uttered by Sherlock Holmes in any of Sir Arthur Conan Doyle's written works. I thought it was worth a footnote.

constrained by biological and physical factors; most importantly, fMRI reflects mass activation. Whatever technical developments the future holds for fMRI, this constraint will remain. Right now, with the 3T scanner available at Caltech outfitted with a volume head coil, a typical high-resolution protocol covers a rather thin slab of the brain with two millimeter isotropic voxels. The volume of each voxel is thus $8mm^3$; since the average density of neurons (in the cortex) is 20,000 to 30,000 per cubic millimeter (see [2]), each voxel contains about 200,000 neurons. The other important consideration is that fMRI cannot differentiate function-specific processing from neuromodulation, or bottom-up from top-down signals, or excitatory from inhibitory activity; fMRI reflects metabolism (and some blood volume modulations that are a direct consequence of neural signaling).

Christof used a sobering metaphor in *The Quest*, which stuck with me and which I transcribe here as I remember it and as it fits my argument. Performing fMRI (or EEG, for that matter) is a bit like flying a blimp over a large stadium (**Figure 98**) during a soccer game (I realize that Dodger Stadium is geared towards baseball; however, after nine years in the US I still have no idea what baseball is about, so I prefer to talk about soccer), and recording what is happening with a microphone. When a goal is scored, the microphone will definitely pick up a huge clamor ascending from the stadium, and you will be able to follow what is happening from this very coarse perspective. Clamors will not only arise when a goal is scored, unfortunately; any exciting action–a great performance by a goalkeeper, a foul leading to a player getting hurt–are all events that would lead to large mass activations of the crowd. You certainly cannot hope to follow the game and understand the rules of soccer using only the microphone installed on the blimp. But it is also true that listening exclusively to one or two of the spectators will not be any better suited for the task at hand, since they may be chatting about something that has nothing to do with the game (this is what single neuron recordings with electrodes would amount to in the context of this metaphor). If the task at hand is to follow the game and understand the rules of soccer, most of

the conversations happening between the spectators are irrelevant. What you would need to do is find a camera, zoom in on the field, track the trajectories of the players and of the ball, and concurrently listen with your microphone for clamors coming up from the stadium. Only then could you get a fairly decent idea. Note the use of multiple recording devices, each offering crucial information that the other devices cannot pick up. In cognitive neuroscience, there are multiple techniques to choose from, ranging from non-invasive monitoring (fMRI, EEG, MEG) and non-invasive perturbations (TMS, tDCS), to invasive monitoring (electrode/array recordings, optical dyes, PET) and invasive perturbations (optogenetics, micro-/macro-stimulation); some of which can only be performed in animals.

How do I know what information should be collected to follow the soccer game and understand the rules of soccer? The trick is that I already know how soccer works, and I know what you should be looking for to understand it. The overarching question in this thesis was to understand how a pattern of impinging photons on the retina could lead to the conscious percept of the face of a loved one. Evidently, no one knows the answer for sure ahead of time, as I did in the case of the soccer stadium metaphor. Hence, what we need is a good hypothesis, possibly derived from a model or a theory, which makes predictions that are falsifiable (Popper still rules!) or at least incompatible with the predictions of a concurrent theory. This is the essence of Kosslyn's quote at the start of this chapter; too often, fMRI studies are exploratory in nature and simply aim at seeing a set of brain areas light up, followed by a post-hoc interpretation. This is the reason why the publishing of five new fMRI studies each day has not dramatically advanced our understanding of brain function very much yet. In the case of our investigation of person recognition, we had a cognitive model derived from the results of many behavioral studies consisting of various sets of units interacting in a certain way. We looked for evidence of the existence of the Person Identity Nodes, which unfortunately we did not find with fMRI; however, we eventually attributed this to a limitation of hemodynamics rather than a falsification of the

model. Single unit recordings do find cells that have the properties one would expect of the theorized Person Identity Nodes (Jennifer Aniston neurons). In retrospect, it is almost evident that fMRI was bound to fail when applied to the study of the sparse neural representations of person identity in the hippocampus. However, it may still have a role to play, for instance in looking at the connectivity between the hippocampus and the region potentially containing Name Recognition Units (see section starting on page 121) during person recognition.

I am finally realizing that I have mostly been hitting the wrong nails with my hammer. It is not that fMRI is a worthless technology (with or without the decoding add-on); it is simply that I have been trying to use it as if it were a direct measure of neural activity, and asking questions that other techniques would have been more appropriate for. I do not think that the first chapter of this thesis had quite sunk in my brain before I wrote it down, even though I studied its contents many times in the past six years. The truth is that I am far from being the only one who makes this mistake. fMRI is increasingly accessible and many cognitive neuroscience labs count it as one of their main tools; I am certain that many use it as just another way to measure neural activity, without giving it much more thought. The advent of fMRI decoding came with many promises, and nowadays many believe that fMRI can read fine patterns of neural activity. I fell for the bold promise of Multivariate Pattern Analysis when it started getting popular six years ago. I have shown (see page 161) that fMRI decoding does not rid fMRI of its inherent limitations. A serious reflection on what questions are worth investigating with fMRI, given its constraints and strengths, is needed. Especially before starting yet another fMRI experiment, one should establish whether the outcome will lead to an advancement of our understanding of the brain. Is this not our ultimate goal, after all?

REFERENCES

1. Koch, C. *Confessions of a Romantic Reductionist*. (MIT Press: 2012).

2. Logothetis, N. K. What we can do and what we cannot do with fMRI. *Nature* **453**, 869–78 (2008).

3. Busch, N. A., Dubois, J. & VanRullen, R. The phase of ongoing EEG oscillations predicts visual perception. *The Journal of neuroscience* **29**, 7869–76 (2009).

4. VanRullen, R., Busch, N., Drewes, J. & Dubois, J. Ongoing EEG phase as a trial-by-trial predictor of perceptual and attentional variability. *Frontiers in Psychology* **2**, 1–9 (2011).

5. Dubois, J., Macdonald, J. & VanRullen, R. Broadband frequency tagging: Reevaluating the sustained division of the attentional spotlight at high temporal resolution. *Vision Sciences Society Annual Meeting* (2010).

6. Dubois, J., Hamker, F. H. & VanRullen, R. Attentional selection of noncontiguous locations : The spotlight is only transiently " split ". *Journal of Vision* **9**, 1–11 (2009).

7. Dubois, J. & Vanrullen, R. Visual trails: do the doors of perception open periodically? *PLoS biology* **9**, e1001056 (2011).

8. Smith, K. fMRI 2.0. *Nature* **484**, 4–6 (2012).

9. Logothetis, N. K. & Wandell, B. A. Interpreting the BOLD signal. *Annual review of physiology* **66**, 735–69 (2004).

10. Huettel, S. A., Song, A. W. & McCarthy, G. *Functional Magnetic Resonance Imaging*. (2004).

11. Lauritzen, M. Reading vascular changes in brain imaging: is dendritic calcium the key? *Nature reviews Neuroscience* **6**, 77–85 (2005).

12. Goense, J. B. M. & Logothetis, N. K. Neurophysiology of the BOLD fMRI signal in awake monkeys. *Current biology* **18**, 631–40 (2008).

13. Weber, B., Keller, A. L., Reichold, J. & Logothetis, N. K. The microvascular system of the striate and extrastriate visual cortex of the macaque. *Cerebral cortex* **18**, 2318–30 (2008).

14. Rauch, A., Rainer, G. & Logothetis, N. K. The effect of a serotonin-induced dissociation between spiking and perisynaptic activity on BOLD functional MRI. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6759–64 (2008).

15. Attwell, D. & Iadecola, C. The neural basis of functional brain imaging signals. *Trends in neurosciences* **25**, 621–5 (2002).

16. Attwell, D. *et al.* Glial and neuronal control of brain blood flow. *Nature* **468**, 232–43 (2010).

17. Petzold, G. C. & Murthy, V. N. Role of astrocytes in neurovascular coupling. *Neuron* **71**, 782–97 (2011).

18. Kim, S.-G. & Ogawa, S. Biophysical and physiological origins of blood oxygenation level-dependent fMRI signals. *Journal of cerebral blood flow and metabolism* **32**, 1188–206 (2012).

19. Thulborn, K. R. My starting point: the discovery of an NMR method for measuring blood oxygenation using the transverse relaxation time of blood water. *NeuroImage* **62**, 589–93 (2012).

20. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, 150–7 (2001).

21. Ekstrom, A. How and when the fMRI BOLD signal relates to underlying neural activity: the danger in dissociation. *Brain research reviews* **62**, 233–44 (2010).

22. Gazzaniga, M., Ivry, R. & Mangun, G. *Cognitive neuroscience: the biology of the mind*. (2008).

23. Ogawa, S., Menon, R. S., Kim, S. G. & Ugurbil, K. On the characteristics of functional magnetic resonance imaging of the brain. *Annual review of biophysics and biomolecular structure* **27**, 447–74 (1998).

24. Cheng, K., Waggoner, R. a & Tanaka, K. Human ocular dominance columns as revealed by high-field functional magnetic resonance imaging. *Neuron* **32**, 359–74 (2001).

25. Ashby, F. G. *Statistical Analysis of fMRI data*. (MIT Press: 2011).

26. Poldrack, R. A., Mumford, J. A. & Nichols, T. E. *Handbook of functional mri data analysis*. (Cambridge University Press: 2011).

27. Kelly, R. E. *et al.* Visual inspection of independent components: defining a procedure for artifact removal from fMRI data. *Journal of neuroscience methods* **189**, 233–45 (2010).

28. Talairach, J. & Tournoux, P. *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging*. (1988).

29. Haxby, J. V *et al.* A Common , High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron* **72**, 404–416 (2011).

30. Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. Linear systems analysis of functional magnetic resonance imaging in human V1. *The Journal of neuroscience* **16**, 4207–21 (1996).

31. Wager, T. D., Vazquez, A., Hernandez, L. & Noll, D. C. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage* **25**, 206–18 (2005).

32. Glover, G. H. Deconvolution of impulse response in event-related BOLD fMRI. *Neuroimage* **9**, 416–429 (1999).

33. Haynes, J. & Rees, G. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* **7**, 523–534 (2006).

34. Mur, M., Bandettini, P. A. & Kriegeskorte, N. Revealing representational content with pattern-information fMRI--an introductory guide. *Soc Cogn Affect Neurosci* **4**, 101–9 (2009).

35. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* **10**, 424–430 (2006).

36. O'Toole, A. J. *et al.* Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J Cogn Neurosci* **19**, 1735–1752 (2007).

37. Pereira, F., Mitchell, T. & Botvinick, M. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* **45**, S199–S209 (2009).

38. Tong, F. & Pratte, M. S. Decoding patterns of human brain activity. *Annu Rev Psychol* **63**, 483–509 (2012).

39. Schwarzkopf, D. S. & Rees, G. Pattern classification using functional magnetic resonance imaging. *Wiley Interdiscip Rev Cogn Sci* **2**, 568–579 (2011).

40. Kragel, P. A., Carter, R. M. & Huettel, S. A. What makes a pattern? Matching decoding methods to data in multivariate pattern analysis. *Front Neurosci* **6**, 162 (2012).

41. Misaki, M., Kim, Y., Bandettini, P. A. & Kriegeskorte, N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage* **53**, 103–18 (2010).

42. Chadwick, M. J., Bonnici, H. M. & Maguire, E. A. Decoding information in the human hippocampus: A user's guide. *Neuropsychologia* **50**, 3107–3121 (2012).

43. Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nat Neurosci* **8**, 679–685 (2005).

44. Haynes, J.-D. & Rees, G. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci* **8**, 686–691 (2005).

45. Op de Beeck, H. P. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses. *NeuroImage* **49**, 1943–1948 (2010).

46. Kamitani, Y. & Sawahata, Y. Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. *NeuroImage* **49**, 1949–1952 (2010).

47. Op de Beeck, H. P. Probing the mysterious underpinnings of multi-voxel fMRI analyses. *NeuroImage* **50**, 567–571 (2010).

48. Kriegeskorte, N., Cusack, R. & Bandettini, P. A. How does an fMRI voxel sample the neuronal activity pattern: Compact-kernel or complex spatiotemporal filter? *NeuroImage* **49**, 1965–1976 (2010).

49. Shmuel, A., Chaimow, D., Raddatz, G., Ugurbil, K. & Yacoub, E. Mechanisms underlying decoding at 7 T : Ocular dominance columns , broad structures , and macroscopic blood vessels in V1 convey information on the stimulated eye. *NeuroImage* **49**, 1957–1964 (2010).

50. Swisher, J. D. *et al.* Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. *J Neurosci* **30**, 325–330 (2010).

51. Freeman, J., Brouwer, G. J., Heeger, D. J. & Merriam, E. P. Orientation decoding depends on maps, not columns. *J Neurosci* **31**, 4792–804 (2011).

52. Engel, S. A., Glover, G. H. & Wandell, B. A. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. Cortex* **7**, 181–192 (1997).

53. Harrison, S. A. & Tong, F. Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**, 632–5 (2009).

54. Mannion, D. J., McDonald, J. S. & Clifford, C. W. G. Discrimination of the local orientation structure of spiral Glass patterns early in human visual cortex. *NeuroImage* **46**, 511–515 (2009).

55. Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).

56. Pereira, F. & Botvinick, M. Information mapping with pattern classifiers: a comparative study. *NeuroImage* **56**, 476–96 (2011).

57. Cristianini, N. & Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*. (Cambridge University Press: 2000).

58. Guyon, I. & Elisseeff, A. An Introduction to Variable and Feature Selection. *J Mach Learn Res* **3**, 1157–1182 (2003).

59. Koch, C. *The Quest for Consciousness: a neurobiological approach*. (Roberts & Co.: 2004).

60. Sorger, B. *et al. Another kind of "BOLD Response": answering multiple-choice questions via online decoded single-trial brain signals. Progress in brain research* **177**, 275–92 (Elsevier: 2009).

61. Blake, R. A Primer on Binocular Rivalry, Including Current Controversies. *Brain and Mind* **2**, 5–38 (2001).

62. Blake, R. & Logothetis, N. K. Visual competition. *Nature reviews Neuroscience* **3**, 13–21 (2002).

63. Tong, F., Meng, M. & Blake, R. Neural bases of binocular rivalry. *Trends in cognitive sciences* **10**, 502–11 (2006).

64. Maier, A., Panagiotaropoulos, T. I., Tsuchiya, N. & Keliris, G. A. Introduction to research topic - binocular rivalry: a gateway to studying consciousness. *Frontiers in human neuroscience* **6**, 263 (2012).

65. Tsuchiya, N. & Koch, C. Continuous flash suppression reduces negative afterimages. *Nature neuroscience* **8**, 1096–101 (2005).

66. Kouider, S., Eger, E., Dolan, R. & Henson, R. N. Activity in face-responsive brain regions is modulated by invisible, attended faces: evidence from masked priming. *Cerebral cortex* **19**, 13–23 (2009).

67. Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M. & Pessoa, L. Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in cognitive sciences* **12**, 314–321 (2008).

68. Timmermans, B., Sandberg, K., Cleeremans, A. & Overgaard, M. Partial awareness distinguishes between measuring conscious perception and conscious content: Reply to Dienes and Seth. *Consciousness and Cognition* **19**, 1081–1083 (2010).

69. Sandberg, K., Timmermans, B., Overgaard, M. & Cleeremans, A. Measuring consciousness: Is one measure better than the other? *Consciousness and cognition* **19**, 1069–1078 (2010).

70. Overgaard, M., Timmermans, B. & Sandberg, K. Optimizing subjective measures of consciousness. *Consciousness and Cognition* **19**, 682–684 (2010).

71. Dienes, Z. & Seth, A. K. Measuring any conscious content versus measuring the relevant conscious content: comment on Sandberg et al. *Consciousness and cognition* **19**, 1079–80; discussion 1081–3 (2010).

72. Dienes, Z. & Seth, A. Gambling on the unconscious: a comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and cognition* **19**, 674–81 (2010).

73. Overgaard, M. & Sandberg, K. Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **367**, 1287–96 (2012).

74. Posner, M. Attention: the mechanisms of consciousness. *Proceedings of the National Academy of Sciences* **91**, 7398–7403 (1994).

75. Merikle, P. M. & Joordens, S. Parallels between perception without attention and perception without awareness. *Consciousness and cognition* **6**, 219–36 (1997).

76. Koch, C. & Tsuchiya, N. Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences* **11**, 16–22 (2007).

77. Jiang, Y. & He, S. Cortical responses to invisible faces: dissociating subsystems for facial-information processing. *Current Biology* **16**, 2023–2029 (2006).

78. Posner, M. I., Snyder, C. R. & Davidson, B. J. Attention and the detection of signals. *Journal of experimental psychology* **109**, 160–74 (1980).

79. Eriksen, C. W. & St James, J. D. Visual attention within and around the field of focal attention: a zoom lens model. *Perception & psychophysics* **40**, 225–40 (1986).

80. McMains, S. A. & Somers, D. C. Multiple spotlights of attentional selection in human visual cortex. *Neuron* **42**, 677–686 (2004).

81. McMains, S. A. & Somers, D. C. Processing efficiency of divided spatial attention mechanisms in human visual cortex. *The Journal of neuroscience* **25**, 9444–8 (2005).

82. Vuilleumier, P., Armony, J. L., Driver, J. & Dolan, R. J. Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron* **30**, 829–841 (2001).

83. Vuilleumier, P., Richardson, M. P., Armony, J. L., Driver, J. & Dolan, R. J. Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature neuroscience* **7**, 1271–1278 (2004).

84. Kanai, R., Tsuchiya, N. & Verstraten, F. A. J. The scope and limits of top-down attention in unconscious visual processing. *Current Biology* **16**, 2332–2336 (2006).

85. Jehee, J. F. M., Brady, D. K. & Tong, F. Attention improves encoding of task-relevant features in the human visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **31**, 8210–9 (2011).

86. Maunsell, J. H. R. & Treue, S. Feature-based attention in visual cortex. *Trends in neurosciences* **29**, 317–22 (2006).

87. Chen, Z. Object-based attention: a tutorial review. *Attention, perception & psychophysics* **74**, 784–802 (2012).

88. Saenz, M., Buracas, G. T. & Boynton, G. M. Global effects of feature-based attention in human visual cortex. *Nature neuroscience* **5**, 631–2 (2002).

89. Buračas, G. T. & Boynton, G. M. Efficient Design of Event-Related fMRI Experiments Using M-Sequences. *NeuroImage* **16**, 801–813 (2002).

90. Sergent, C. & Dehaene, S. Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological science* **15**, 720–8 (2004).

91. Kanwisher, N. & Yovel, G. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **361**, 2109–28 (2006).

92. Spiridon, M., Fischl, B. & Kanwisher, N. Location and Spatial Profile of Category-Specific Regions in Human Extrastriate Cortex. *Human Brain Mapping* **89**, 77– 89 (2006).

93. Fischl, B. *et al.* Whole brain segmentation automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).

94. Hesselmann, G., Hebart, M. & Malach, R. Differential BOLD Activity Associated with Subjective and Objective Reports during "Blindsight" in Normal Observers. *The Journal of neuroscience* **31**, 12936–44 (2011).

95. Sterzer, P., Haynes, J.-D. & Rees, G. Fine-scale activity patterns in high-level visual areas encode the category of invisible objects. *Journal of Vision* **8**, 1–12 (2008).

96. Kouider, S. & Dehaene, S. Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **362**, 857–75 (2007).

97. Dehaene, S. & Changeux, J.-P. Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–27 (2011).

98. Stein, T. & Sterzer, P. High-level face shape adaptation depends on visual awareness: Evidence from continuous flash suppression. *Journal of Vision* **11**, 1–14 (2011).

99. Mudrik, L., Breska, A., Lamy, D. & Deouell, L. Y. Integration without awareness: expanding the limits of unconscious processing. *Psychological science* **22**, 764–70 (2011).

100. Stein, T., Hebart, M. N. & Sterzer, P. Breaking Continuous Flash Suppression: A New Measure of Unconscious Processing during Interocular Suppression? *Frontiers in human neuroscience* **5**, 167 (2011).

101. Faivre, N., Berthet, V. & Kouider, S. Nonconscious influences from emotional faces: a comparison of visual crowding, masking, and continuous flash suppression. *Frontiers in psychology* **3**, 129 (2012).

102. Almeida, J., Mahon, B. Z., Nakayama, K. & Caramazza, A. Unconscious processing dissociates along categorical lines. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 15214–8 (2008).

103. Almeida, J., Mahon, B. Z. & Caramazza, A. The role of the dorsal visual processing stream in tool identification. *Psychological science* **21**, 772–8 (2010).

104. Sakuraba, S., Sakai, S., Yamanaka, M., Yokosawa, K. & Hirayama, K. Does the Human Dorsal Stream Really Process a Category for Tools? *Journal of Neuroscience* **32**, 3949–3953 (2012).

105. Draine, S. C. & Greenwald, A. G. Replicable unconscious semantic priming. *Journal of Experimental Psychology: General* **127**, 286 (1998).

106. Hebart, M. N. & Hesselmann, G. What Visual Information is Processed in the Human Dorsal Stream? *Journal of Neuroscience* **32**, 8107–8109 (2012).

107. Whelan, R. Effective analysis of reaction time data. *The Psychological Record* **58**, 475–482 (2010).

108. Poldrack, R. A. *et al.* Guidelines for reporting an fMRI study. *NeuroImage* **40**, 409–14 (2008).

109. Neuroskeptic The Nine Circles of Scientific Hell. *Perspectives on Psychological Science* **7**, 643–644 (2012).

110. Monto, S., Palva, S., Voipio, J. & Palva, J. M. Very slow EEG fluctuations predict the dynamics of stimulus detection and oscillation amplitudes in humans. *J. Neurosci.* **28**, 8268–8272 (2008).

111. VanRullen, R. & Dubois, J. The psychophysics of brain rhythms. *Frontiers in psychology* **2**, 1–10 (2011).

112. Kouider, S. & Dupoux, E. Partial Awareness Creates the "Illusion" of Subliminal Semantic Priming. *Psychological Science* **15**, 75–81 (2004).

113. Hong, S. & Blake, R. Interocular suppression differentially affects achromatic and chromatic mechanisms. *Attention, Perception, & Psychophysics* **71**, 403–411 (2009).

114. Ramsoy, T. Z. & Overgaard, M. Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences* **3**, 1–23 (2004).

115. Persaud, N., McLeod, P. & Cowey, A. Post-decision wagering objectively measures awareness. *Nature neuroscience* **10**, 257–61 (2007).

116. Koch, C. & Preuschoff, K. Betting the house on consciousness. *Nature neuroscience* **10**, 140–1 (2007).

117. Kunimoto, C., Miller, J. & Pashler, H. Confidence and accuracy of near-threshold discrimination responses. *Consciousness and cognition* **10**, 294–340 (2001).

118. Naccache, L., Blandin, E. & Dehaene, S. Unconscious masked priming depends on temporal attention. *Psychological Science* **13**, 416–424 (2002).

119. Finkbeiner, M. & Palermo, R. The role of spatial attention in nonconscious processing: a comparison of face and nonface stimuli. *Psychological science* **20**, 42–51 (2009).

120. Van Boxtel, J. J. A., Tsuchiya, N. & Koch, C. Opposing effects of attention and consciousness on afterimages. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8883–8 (2010).

121. Bahrami, B., Lavie, N. & Rees, G. Attentional load modulates responses of human primary visual cortex to invisible stimuli. *Current biology : CB* **17**, 509–13 (2007).

122. Watanabe, M. *et al.* Attention but not Awareness Modulates the BOLD Signal in the Human V1 during Binocular Suppression. *Science* **334**, 829–831 (2011).

123. Busch, N. A. & VanRullen, R. Spontaneous EEG oscillations reveal periodic sampling of visual attention. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16048–53 (2010).

124. Mathewson, K. E., Gratton, G., Fabiani, M., Beck, D. M. & Ro, T. To see or not to see: prestimulus alpha phase predicts visual awareness. *The Journal of neuroscience* **29**, 2725–32 (2009).

125. Mathewson, K. E., Fabiani, M., Gratton, G., Beck, D. M. & Lleras, A. Rescuing stimuli from invisibility: Inducing a momentary release from visual masking with pre-target entrainment. *Cognition* **115**, 186–91 (2010).

126. Eckstein, D. & Henson, R. N. Stimulus/response learning in masked congruency priming of faces: Evidence for covert mental classifications? *The Quarterly journal of experimental psychology.* **65**, 92–120 (2012).

127. Jiang, Y., Costello, P., Fang, F., Huang, M. & He, S. A gender- and sexual orientation-dependent spatial attentional effect of invisible images. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17048–52 (2006).

128. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* (2013).

129. Abbott, A. Disputed results a fresh blow for social psychology. *Nature* **497**, 16 (2013).

130. Shanks, D. R. *et al.* Priming Intelligent Behavior: An Elusive Phenomenon. *PLoS ONE* **8**, e56515 (2013).

131. Pashler, H. & Wagenmakers, E.-J. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science* **7**, 528–530 (2012).

132. Patterson, K., Nestor, P. J. & Rogers, T. T. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature reviews Neuroscience* **8**, 976–87 (2007).

133. Bruce, V. & Young, A. Understanding face recognition. *British journal of psychology* (1986).

134. Young, A. & Bruce, V. Understanding person perception. *British Journal of Psychology* **102**, 959–974 (2011).

135. Burton, A. M., Bruce, V. & Johnston, R. A. Understanding face recognition with an interactive activation model. *The British journal of psychology* **81**, 361–380 (1990).

136. McClelland, J. & Rumelhart, D. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological review* **88**, 375–407 (1981).

137. Bredart, S., Valentine, T., Calder, A. & Gassi, L. An interactive activation model of face naming. *The Quarterly journal of experimental psychology.* **48**, 466–486 (1995).

138. Burton, A. & Bruce, V. Naming faces and naming names: Exploring an interactive activation model of person recognition. *Memory* **1**, 457–480 (1993).

139. Bruce, V. & Young, A. W. *Face Perception*. 481 (Psychology Press: New York, 2012).

140. Young, A. W., Hay, D. C. & Ellis, A. W. The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British journal of psychology* **76**, 495–523 (1985).

141. Hay, D. C., Young, A. W. & Ellis, A. W. Routes through the face recognition system. *The Quarterly journal of experimental psychology.* **43**, 761–791 (1991).

142. Young, A., Ellis, A. & Flude, B. Accessing stored information about familiar people. *Psychological Research* **50**, 111–115 (1988).

143. De Haan, E. H., Young, A. W. & Newcombe, F. Face recognition without awareness. *Cognitive Neuropsychology* **4**, 385–415 (1987).

144. De Haan, E. H., Young, A. W. & Newcombe, F. A dissociation between the sense of familiarity and access to semantic information concerning familiar people. *European Journal of Cognitive Psychology* **3**, 51–67 (1991).

145. Flude, B., Ellis, A. & Kay, J. Face processing and name retrieval in an anomic aphasic: Names are stored separately from semantic information about familiar people. *Brain and cognition* **11**, 60–72 (1989).

146. Damasio, H., Tranel, D., Grabowski, T., Adolphs, R. & Damasio, a Neural systems behind word and concept retrieval. *Cognition* **92**, 179–229 (2004).

147. Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D. & Damasio, A. R. A neural basis for lexical retrieval. *Nature* **380**, 499–505 (1996).

148. Martin, A. The representation of object concepts in the brain. *Annual review of psychology* **58**, 25–45 (2007).

149. Barsalou, L. W. Cognitive and Neural Contributions to Understanding the Conceptual System. *Current Directions in Psychological Science* **17**, 91–95 (2008).

150. Semenza, C. The Neuropsychology of Proper Names. *Mind & Language* **24**, 347–369 (2009).

151. Haxby, J. V & Gobbini, M. I. Distributed Neural Systems for Face Perception. *Oxford Handbook of face perception* 93–110 (2011).

152. Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. The distributed human neural system for face perception. *Trends in cognitive sciences* **4**, 223–232 (2000).

153. Gobbini, M. I., Leibenluft, E., Santiago, N. & Haxby, J. V Social and emotional attachment in the neural representation of faces. *NeuroImage* **22**, 1628–35 (2004).

154. Gobbini, M. I. & Haxby, J. V Neural response to the visual familiarity of faces. *Brain research bulletin* **71**, 76–82 (2006).

155. Kreiman, G. On the neuronal activity in the human brain during visual recognition, imagery and binocular rivalry. *PhD thesis, California Institute of Technology, Pasadena, CA*. (2002).

156. Koch, C. Being John Malkovich: Personal Control of Individual Brain Cells. *Scientific American Mind* (2011).

157. Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).

158. Quiroga, R. Q., Kraskov, A., Koch, C., Fried, I. & Quian Quiroga, R. Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain. *Current Biology* **19**, 1–28 (2009).

159. Willenbockel, V., Sadr, J. & Fiset, D. Controlling low-level image properties: the SHINE toolbox. *Behavior Research Methods* **42**, 671–684 (2010).

160. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).

161. Ioannidis, J. P. A. Why most published research findings are false. *PLoS medicine* **2**, e124 (2005).

162. Lehrer, J. The truth wears off. *The New Yorker* **12**, 1–7 (2010).

163. Nestor, A., Behrmann, M. & Plaut, D. C. The Neural Basis of Visual Word Form Processing: A Multivariate Investigation. *Cerebral Cortex* 1–12 (2012).

164. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience* **12**, 535–40 (2009).

165. Mormann, F. *et al.* A category-specific response to animals in the right human amygdala. *Nature Neuroscience* 1–3 (2011).doi:10.1038/nn.2899

166. Lang, P. J., Bradley, M. M. & Cuthbert, B. N. *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. (1997).

167. Chen, N.-K., Dickey, C. C., Yoo, S.-S., Guttmann, C. R. . & Panych, L. P. Selection of voxel size and slice orientation for fMRI in the presence of susceptibility field gradients: application to imaging of the amygdala. *NeuroImage* **19**, 817–825 (2003).

168. Robinson, S., Windischberger, C., Rauscher, A. & Moser, E. Optimized 3 T EPI of the amygdalae. *NeuroImage* **22**, 203–10 (2004).

169. Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B. & Tootell, R. B. H. Faces and objects in macaque cerebral cortex. *Nat Neurosci* **6**, 989–995 (2003).

170. Moeller, S., Freiwald, W. A. & Tsao, D. Y. Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science* **320**, 1355–9 (2008).

171. Tsao, D., Freiwald, W., Tootell, R. & Livingstone, M. A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).

172. Freiwald, W. A. & Tsao, D. Y. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–851 (2010).

173. Leite, F. P. *et al.* Repeated fMRI using iron oxide contrast agent in awake, behaving macaques at 3 Tesla. *NeuroImage* **16**, 283–294 (2002).

174. Kietzmann, T. C., Swisher, J. D., Konig, P. & Tong, F. Prevalence of Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Visual Pathways. *J Neurosci* **32**, 11763–11772 (2012).

175. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2**, 4 (2008).

176. Serre, T. & Riesenhuber, M. Realistic Modeling of Simple and Complex Cell Tuning in the HMAX Model , and Implications for Invariant Object Recognition in Cortex. (2004).

177. Hurley, N. & Rickard, S. Comparing measures of sparsity. *IEEE Trans Inf Theory* **55**, 4723–4741 (2009).

178. Grill-Spector, K. & Malach, R. fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol* **107**, 293–321 (2001).

179. Drucker, D. M. & Aguirre, G. K. Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cereb Cortex* **19**, 2269–2280 (2009).

180. Epstein, R. A. & Morgan, L. K. Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis. *Neuropsychologia* **50**, 530–543 (2012).

181. Kosslyn, S. M. If neuroimaging is the answer, what is the question? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **354**, 1283–94 (1999).

182. Fogelson, S., Miller, K., Kohler, P., Granger, R. & Tse, P. U. Not all suppressions are created equal: Categorical decoding of unconsciously presented stimuli varies with suppression paradigm. *Perception ECVP Abstract Supplement* 144 (2011).