Statistical models for comprehensive meta-analyses of neuroimaging studies

Jérôme Dockès

université PARIS-SACLAY

NNT : 2019SACLT048

Thèse de doctorat

Modèles statistiques pour méta-analyses d'études de neuroimagerie.

Thèse de doctorat de l'Université Paris-Saclay
préparée à Télécom ParisTech

École doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Mathématiques et Informatique

Thèse présentée et soutenue à Palaiseau, le 2019-11-27, par

JÉRÔME DOCKÈS

Composition du Jury :

Angela Laird Professeur, Florida International University (Neuroinformatics and brain connectivity lab)	Rapporteur
Thomas Nichols Professeur, University of Oxford (Big Data Institute)	Rapporteur
Yves Grandvalet Directeur de Recherche, CNRS (Heudiasyc)	Examinateur
Lars Kai Hansen Professeur, DTU Compute (Cognitive Systems)	Examinateur
Pierre Zweigenbaum Directeur de Recherche, CNRS (LIMSI)	Président
Fabian Suchanek Directeur de Recherche, Télécom ParisTech (DIG)	Directeur de thèse
Gaël Varoquaux Directeur de Recherche, INRIA (Parietal)	Directeur de thèse
Bertrand Thirion Directeur de Recherche, INRIA (Parietal)	Invité

STATISTICAL MODELS FOR COMPREHENSIVE META-ANALYSES OF NEUROIMAGING STUDIES

JÉRÔME DOCKÈS

ABSTRACT

Brain mapping studies the spatial organization of the brain and associations between mental functions or diseases and anatomical structures. It relies on structural and functional neuroimaging techniques such as functional Magnetic Resonance Imaging (fMRI). The neuroimaging literature is growing rapidly, with thousands of publications every year. Extracting general and reliable knowledge about the brain from this huge amount of results is difficult. Indeed, individual studies lack statistical power and report many spurious findings. Even genuine effects are often specific to particular experimental settings and difficult to reproduce. Moreover, the mechanisms under study are very diverse, and terminology is not yet fully formalized. Therefore, integrating results across experiments and research domains is challenging.

Meta-analysis aggregates studies to identify consistent trends in reported associations between brain structure and behavior. The standard approach to meta-analysis starts by gathering a sample of studies that investigate a same mental process or disease. Then, a statistical test is performed at every voxel to delineate brain regions where there is a significant agreement among reported findings.

In this thesis, we develop a different kind of meta-analysis that focuses on prediction rather than hypothesis testing. We build predictive models that map textual descriptions of experiments, mental processes or diseases to anatomical regions in the brain. We use multivariate models to combine terms used to describe experiments or observations rather than studying each concept in isolation. We introduce models that learn from very few examples, thus extending the scope of meta-analysis to seldom-studied diseases and mental processes. Our supervised learning approach comes with a natural quantitative evaluation framework, and we conduct extensive experiments to validate and compare statistical models. To train these models, we collect and share the largest existing dataset of neuroimaging studies and stereotactic coordinates. This dataset contains the full text and locations of neurological observations for over 13 000 publications.

In the last part, we turn to decoding: inferring mental states from brain activity. We perform this task through meta-analysis of fMRI statistical maps collected from an online data repository. We use fMRI data to distinguish a wide range of mental conditions.

Standard meta-analysis is an essential tool to distinguish true discoveries from noise and artifacts. This thesis introduces methods for predictive meta-analysis, which complement the standard approach and help interpret neuroimaging results and formulate hypotheses or formal statistical priors.

RÉSUMÉ

La neuroimagerie permet d'étudier les liens entre la structure et le fonctionnement du cerveau. La littérature de neuroimagerie croît rapidement, avec des milliers de publications par an. Il est difficile d'extraire des connaissances sur le fonctionnement du cerveau de cette grande quantité de résultats. En effet, chaque étude manque de puissance statistique et peut reporter beaucoup de faux positifs. De plus, certains effets sont spécifiques à un protocole expérimental et difficiles à reproduire.

Les méta-analyses rassemblent plusieurs études pour identifier les associations entre structures anatomiques et processus cognitifs qui sont établies de manière consistente dans la littérature. Les méthodes classiques de méta-analyse commencent par constituer un échantillon d'études focalisées sur un même processus mental ou une même maladie. Ensuite, un test statistique est effectué pour chaque voxel, afin de délimiter les régions cérébrales dans lesquelles le nombre d'observations reportées est significatif.

Dans cette thèse, nous introduisons une nouvelle forme de méta-analyse, qui s'attache à construire des prédictions plutôt qu'à tester des hypothèses. Nous introduisons des modèles statistiques qui prédisent la distribution spatiale des observations neurologiques à partir de la description textuelle d'une expérience, d'un processus cognitif ou d'une maladie cérébrale. Nos modèles peuvent apprendre avec très peu d'exemples à associer une carte cérébrale à une fonction ou une maladie mentale. Notre approche est basée sur l'apprentissage statistique supervisé qui fournit un cadre classique pour évaluer et comparer les modèles. Pour entraîner nos modèles, nous construisons le plus grand jeu de données d'études de neuroimagerie et de coordonnées stéréotaxiques existant, qui rassemble plus de 13 000 publications. Dans la dernière partie, nous nous intéressons au décodage, qui consiste à prédire des états psychologiques à partir de l'activité cérébrale. Nous apprenons à décoder des images de cerveau à travers la méta-analyse de données d'IRM fonctionnelle (IRMf). Les images d'IRMf nous permettent de classifier un grand nombre d'états psychologiques.

La méta-analyse standard est un outil indispensable pour distinguer les vraies découvertes du bruit et des artefacts parmi les résultats publiés en neuroimagerie. Cette thèse introduit des méthodes adaptées à la méta-analyse prédictive. Cette approche est complémentaire de la méta-analyse standard, et aide à interpréter les résultats d'études de neuroimagerie ainsi qu'à formuler des hypothèses ou des a priori statistiques.

ACKNOWLEDGMENTS

I thank my supervisors Gaël Varoquaux, Fabian Suchanek and Bertrand Thirion for always being available and friendly, for their valuable guidance and for all they have taught me. I thank the members of the jury: Angela Laird, Thomas Nichols, Yves Grandvalet, Lars Kai Hansen, and Pierre Zweigenbaum for reviewing this thesis and for their kind and insightful comments. I thank all the members of the Parietal team for being great friends and colleagues and I will miss working in this group. I thank Russel Poldrack, Chris Gorgolewski and all the Poldrack Lab in Stanford University for welcoming me for a two-month stay and helping me improve my understanding of neuroimaging and its applications. Most of all, I thank my parents, my sisters, and my wife. Finally, I thank Digicosme and Digiteo for funding my thesis.

CONTENTS

ACRONYMS

ADHD Attention Deficit Hyperactivity Disorder

ALE Activation Likelihood Estimation

API Application Programming Interface

AUC Area Under Curve

BERT Bidirectional Encoder Representations from Transformers

BOLD Blood Oxygenation Level Dependent

BOW Bag-of-words

CBMA Coordinate-Based Meta-Analysis

csv comma-separated values

DTI Diffusion Tensor Imaging

EEG Electroencephalography

FAQ Frequently Asked Questions

FDR False Discovery Rate

FFA Fusiform Face Area

fMRI functional Magnetic Resonance Imaging

FoV Field of View

FWER Family-Wise Error Rate

FWHM Full Width at Half Maximum

GCLDA Generalized Correspondence Latent Dirichlet Allocation

GCV Generalized Cross-Validation

GLM General Linear Model

HCP Human Connectome Project

HRF Haemodynamic Response Function

i.i.d. independent identically distributed

IBC Individual Brain Charting

IBMA Image-Based Meta-Analysis

ICA Independent Component Analysis

ICBM International Consortium for Brain Mapping

ID Identitifier

INRIA Institut National de la Recherche en Informatique et Automatique

IP Internet Protocol

IRLS Iteratively Reweighted Least Squares

JATS Journal Article Tag Suite

KDA Kernel Density Analysis

KDE Kernel Density Estimation

LAD Least Absolute Deviations

LDA Latent Dirichlet Allocation

LOO Leave-One-Out

LTI Linear Time-Invariant

MA Modeled Activation

MEG Magnetoencephalography

MeSH Medical Subject Headings

MF Matrix Factorization

MKDA Multilevel Kernel Density Analysis

MNI Montreal Neurological Institute

MRI Magnetic Resonance Imaging

NCBI National Center for Biotechnology Information

NLM National Library of Medicine

NLP Natural Language Processing

NMF Non-negative Matrix Factorization

NNoD Neural Network over Dictionary

OLS Ordinary Least Squares

PCA Principal Component Analysis

pdf probability density function

PET Positron Emission Tomography

PMC PubMed Central

PMID PubMed ID

PPV Positive Predictive Value

RDF Resource Description Framework

ROC Receiver Operating Characteristic

ROI Region of Interest

RSVP Rapid-Serial-Visual-Presentation

SGNS Skip-Gram Negative Sampling

SKOS Simple Knowledge Organization System

SMA Supplementary Motor Area

SPARQL SPARQL Protocol and RDF Query Language

SPM Statistical Parametric Mapping

SVD Singular Value Decomposition

SVM Support Vector Machines

TFIDF Term Frequency · Inverse Document Frequency

TR repetition time

TV Total Variation

UID Unique Identitifier

VBM Voxel Based Morphometry

VBM Voxel-Based Morphometry

XHTML eXtensible HyperText Markup Language

XML eXtensible Markup Language

XSD XML Schema Definition

XSLT eXtensible Stylesheet Language Transformations

1 OVERVIEW

Neuroimaging techniques such as fMRI offer insights into the spatial organization of the brain, and the links between brain structure and behavior. The neuroimaging literature has been growing very rapidly, and now over 6 000 neuroimaging studies are published every year[1]. Extracting reliable knowledge from this huge amount of results is difficult, as terminology and experimental paradigms are not yet formalized, and reported findings contain many false discoveries. Therefore, important research efforts are dedicated to *meta-analysis*: pooling many published studies to perform statistical analysis of their findings. This enables identifying effects that are reported consistently across experiments. So far, meta-analysis has focused on sets of studies that investigate a common, well-defined psychological concept, such as a disease or a mental process. The objective is to detect a significant concordance of results within such homogeneous samples.

In this thesis, we develop a new framework for meta-analysis, which treats it as a prediction problem. The task is to use text, such as the description of a neuroimaging experiment, to predict the probable locations of the associated observations in the brain. Mapping behavior or phenotypes to brain structures is often called *encoding*. Our predictive framework is complementary to standard meta-analysis, which focuses on statistical inference and hypothesis testing.

In a supervised learning setting, we train multivariate regression models to map vector representations of neuroimaging reports to spatial densities over the brain. This approach is not restricted to well-defined and frequently investigated neurological phenomena. It opens the door to new applications of meta-analysis, to summarize and explore the literature, generate hypotheses and formal priors for new experiments, and ground the interpretation of neuroimaging results in a comprehensive view of published findings. In this chapter, we outline the structure of this manuscript.

1.1 LARGE-SCALE META-ANALYSIS

In this first part, we introduce brain mapping methods and the main functional neuroimaging modality considered in this thesis: functional Magnetic Resonance Imaging (fMRI). We describe the standard approach to collection and statistical analysis of fMRI data. We also

1 https://www.ncbi.nlm.nih.gov/pubmed/?term=%22neuroimaging%22

describe how neuroimaging results are reported, either in the form of statistical maps of the brain or of sets of coordinates for the locations of neurological observations. We then turn to meta-analysis: collecting reports of neuroimaging observations and analyzing them at the study level. We describe standard approaches to meta-analysis, and summarize how the work developed in this thesis departs from existing methods. The last chapter of this introduction explains how we collected the largest existing dataset of neuroimaging publications and associated brain locations. We rely on this dataset throughout the rest of this thesis to train and validate statistical models.

1.2 MAPPING TEXT TO BRAIN LOCATIONS

The second part of the manuscript builds a formal setting for supervised meta-analysis of neuroimaging studies. We describe our supervised learning framework, how we represent text and neuroimaging data and how we evaluate the generalization performance of statistical models. We also introduce our multivariate linear regression approach to mapping neurological terms to spatial densities over the brain. The supervised learning setting naturally provides us with tools for model selection and enables us to quantitatively compare models and data representations. Finally, we show that by embedding text into brain space, encoding models enable exploiting the large number of openly available neuroimaging reports that only contain text – i. e. no brain images nor stereotactic coordinates. This allows us to delineate neural support of diseases for which little imaging data is available.

1.3 MAPPING ARBITRARY QUERIES

In the third part, we extend our encoding models to leverage relationships binding the terms used in neuroscience reports. We also introduce a mechanism for feature selection suited to our particular setting – which involves a very high-dimensional output space as there is one dependent variable for each voxel in the brain. This enables the encoding models to handle a wide class of inputs, from single, seldom-used neuroimaging terms to full articles. Capturing relations between terms alleviates the problem of inconsistent terminology. It also enables models to cope with the fact that most terms of interest are very rare, occurring in less than 1% of articles. This scarcity is characteristic of high-cardinality sets of discrete objects, in which occurrence frequencies often follow a power law (a. k. a. Zipf law for the special case of word occurrences). Dealing with this curse of the power law, and learning from few observations, is a central theme of this thesis. This challenge arises in many other learning tasks, such as

recommendation or prediction tasks involving categorical variables. This third part also describes an extensive set of experiments that measure the generalization performance and the sample complexity of meta-analysis methods.

1.4 DECODING BRAIN IMAGES

In the last part of this thesis, we shift to a different problem. Instead of using text to predict brain images, we perform the reverse task. Given a statistical map of the brain, we predict the mental conditions it is associated with. We frame this as a supervised classification task, and use many neuroimaging studies from a large online repository to train non-linear decoding models. Unlike previous decoding studies, we strive to distinguish several dozens of labels, covering a wide range of mental states. Again, we are confronted with the problem of terms that are referenced in very few experiments. We show that it is possible to distinguish many different mental states from brain imaging data, and that with the large datasets that were gathered by the neuroimaging community in recent years, non-linear models begin to outperform the linear models that are typically used for decoding.

1.5 CONCLUSION

In the conclusion, we summarize our results and the difficulties that we met during the course of this thesis. We also present open data, open-source software, and an online platform for meta-analysis that resulted from our work.

Part I

LARGE-SCALE META-ANALYSIS

2 BRAIN MAPPING PRIMER

Brain mapping studies associations between mental functions or diseases and anatomical structures in the brain. In order to do so, it relies heavily on neuroimaging – techniques, such as Magnetic Resonance Imaging (MRI), that acquire images of the brain's structure, or record signals produced by its activity.

Brain mapping and neuroimaging have important applications in clinical neuroscience. For example, MRI can reveal atrophy of some brain structures in patients suffering from Alzheimer's disease. Such observations are useful for diagnosis and to understand the mechanisms that underlie the disease.

Neuroimaging also helps cognitive neuroscience to study the spatial organization of the brain. It is used to delineate the neural support of mental functions: identify regions or networks – parts of the brain – that are recruited to implement each mental process. For example, the first stages of processing visual information take place in the occipital lobe; the precentral gyrus is involved in planning and executing movements; Broca's area, in the frontal lobe, plays an important role in language comprehension and production. Many such associations are stable across individuals, and brain mapping can be undertaken at the individual or at the population level.

Functional neuroimaging techniques are particularly important for brain mapping. They provide direct or indirect measures of neural activity. They can thus be used to study how stimulating subjects in a particular way – such as showing them a picture of an object – or asking them to perform a certain task – such as reading a sentence out loud – correlates with their brain activity.

These techniques have provided many insights into the roles and interactions of brain structures during the past decades. Functional mapping of the brain is a vast research field that publishes thousands of articles every year. In this thesis, we develop methods to perform large-scale statistical analysis of the neuroimaging literature, in order to extract general and reliable knowledge about the brain from thousands of published studies.

In this chapter, we describe how neuroimaging studies acquire and analyze data, and how they represent and communicate their findings. In subsequent chapters, we will turn to meta-analysis. We will aggregate and analyze results from many studies, to extract reliable statistical summaries, and to build predictive models of the statistical links between brain structure and function.

To build an atlas of the human brain, and to associate mental functions or disorders with positions in this organ, we need a spatial reference. Moreover, in order to find links that hold for the human population – and not only for individual subjects – and to assemble findings from different neuroimaging studies, it must be possible to register individual brains to a common reference. Hence, before describing methods for statistical analysis of neuroimaging data, we begin with *spatial normalization*, and explain how brains of different individuals, scanned in laboratories across the world, can be embedded in a common spatial coordinate system.

2.1 TEMPLATES AND SPATIAL NORMALIZATION

Brains of different individuals have different shapes and sizes. Still, up to a nonlinear deformation, many aspects of their spatial organization are common, and brain mapping often aims to delineate regions that play similar roles in all individuals. In order to compare brain activity across individuals and experiments, and to find commonalities and differences, it is necessary to align brain images. Once images are aligned, positions in different images correspond to the same anatomical structures, and can be compared. This alignment is called *coregistration*. Images of different brains are translated, rotated, scaled, and deformed non-linearly until inter-individual variability in shape, size and position is sufficiently small (about 1 cm). Only then, pointwise comparison of brain images becomes meaningful. Even when we perform individual-level analysis, and process each individual's recorded brain activity separately, it can be beneficial to align the results obtained for several brains in order to assess similarities. This is typically done by registering all brain images to a standard brain *template*, that was obtained by coregistering and averaging anatomical images from many individuals.

2.1.1 The ICBM 152 template

The first reference used for spatial normalization was the 1988 Talairach and Tournoux atlas (Talairach and Tournoux, 1988), that was created by dissecting and labeling a single brain. Using this atlas is problematic for several reasons: *i)* it is based on the brain of a single 60-year old woman and is not representative of the whole population, *ii)* no MRI scan is available for this individual, which prevents the use of automatic registration techniques, and *iii)* only the left hemisphere was labeled, then reflected to obtain the right hemisphere, but there are some well-known asymmetries in normal brains. An in-depth discussion of the limitations of this spatial reference is provided in Devlin and Poldrack (2007).

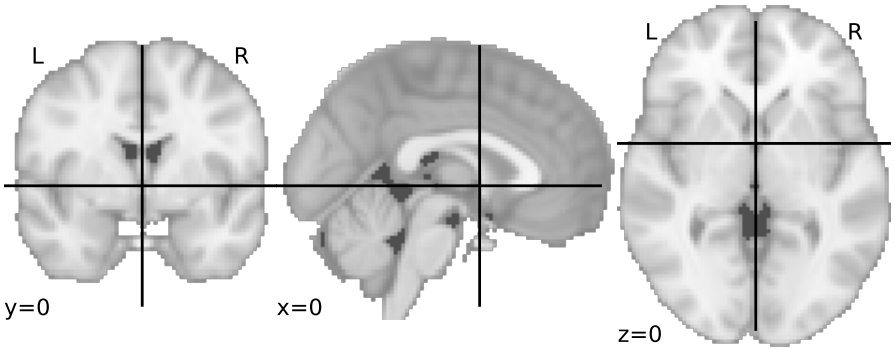


Figure 2.1: ICBM 152 template. This anatomical image was created at the MNI by coregistering and averaging brain scans of 152 adults. This template also defines a coordinate system. On this plot, the image is sliced in the x, y, and z directions at the origin of the coordinate system, and the cross shows the axes. The numbers at the bottom of each plot indicate the coordinates (in millimeters) at which the 3-dimensional image is sliced, i. e. $x = 0$, $y = 0$, $z = 0$. Once an image is registered to this template, locations in the image can be expressed in this standard spatial reference of the human brain.

The modern approach to spatial normalization is therefore to use an image-based template obtained by averaging brain scans from many individuals. The MNI 305 template (Evans et al., 1993) was created at the Montreal Neurological Institute (MNI) by registering 305 structural (anatomical) MRI images to the Talairach atlas brain. The current standard template is the ICBM 152 template (Fonov et al., 2009). It was created by registering 152 higher-resolution MRI images from the International Consortium for Brain Mapping (ICBM) project to the MNI 305 template and averaging them. The ICBM template is often referred to as the "MNI 152 template", or simply "the MNI template". Standard neuroimaging data analysis therefore typically entails registering all the acquired brain images to this standard (synthetic) brain. The ICBM 152 template is shown in Fig. 2.1.

2.1.2 Stereotactic coordinates

2.1.2.1 *The MNI 152 reference*

Once images are registered to the MNI template, brain locations can be expressed in terms of a standard spatial coordinate system, shown in Fig. 2.1. The different anatomical structures in the brain are aligned and can be found at the same coordinates.

2.1.2.2 *Positioning voxels*

Even when images are registered to a template and are therefore in the same coordinate space, they may have different resolutions or a different Field of View (FoV). Images contain measurements spaced on a regular spatial grid of *voxels* – the 3-dimensional equivalent of

pixels. An image is then described by: *i)* a data array, that holds a value for each voxel, and *ii)* an affine transformation, represented by an augmented matrix[1], that maps voxel indices to coordinates in the spatial coordinate system of the brain (in millimeters). Therefore, images can be downsampled, cropped, *etc.* , while remaining in the MNI space without losing the essential property that for each voxel, we can compute its MNI coordinate, and at that coordinate, the same anatomical structure is found in all images.

Hence, the MNI standard template and coordinate system enable communicating and comparing results across individual subjects, studies and labs. This can be done by providing full brain images registered to the template, or providing the coordinates of observations (for example of peaks of brain activity) in the standard coordinate system. This common space is therefore essential for brain mapping, and particularly for *meta-analysis* – pooling and analyzing results from several studies, which is the focus of this thesis.

2.2 STATISTICAL PARAMETRIC MAPPING

Brain mapping estimates statistical links between measurements made in the brain – for example by an MRI machine – and phenotypic or behavioral quantities of interest, such as a particular mental condition. The Statistical Parametric Mapping (SPM) approach (Penny et al., 2011) consists in obtaining many brain scans, and using this sample to compute a statistic for each voxel.

For example, in Voxel-Based Morphometry (VBM), we compute at each voxel a measure of the local gray matter density. We can obtain such images for a group of Alzheimer's disease patients and for a group of healthy controls. Then, for each voxel, we can compute the (rescaled) difference between the group means. We thus obtain one statistic for each voxel, which together form a *statistical map*. A statistical test can then be applied to identify voxels in which the observed statistic is significant – for example to reject, at some positions in the brain, the hypothesis that the local gray matter density is the same in patients as in healthy controls.

Functional neuroimaging techniques such as fMRI measure the brain activity. During a session, the brain activity of an individual is scanned repeatedly. For each voxel, we thus obtain a time series of measurements. For each time point, i. e. for each measurement, we also have variables that describe the mental condition of the subject – for example a binary variable indicating whether the subject is hearing a sound or not. The voxel time series can then be *regressed* on these behavioral descriptors. This yields one brain map of regression coefficients for each behavioral descriptor. Statistical tests can be applied, for example

1 https://en.wikipedia.org/wiki/Affine_transformation#Augmented_matrix

to identify regions in which hearing a sound has a significant effect on brain activity – most likely the auditory cortex.

The nature of the measurements, the data processing steps, and the topics of study vary between experiments and imaging modalities. Still, the SPM framework remains similar, and this type of analysis always results in statistical maps of the brain. In the next two sections, we describe in more detail the statistical analysis of fMRI data, which constitutes a canonical application of the SPM framework.

2.3 FUNCTIONAL MRI

Because of its high spatial resolution, fMRI is particularly adapted for brain mapping. The tools developed in this thesis rely mostly on statistical analysis of aggregated results from fMRI studies. Therefore, in this section and the next, we summarize how fMRI data are obtained and analyzed. Many aspects of the analysis of these data are shared with other modalities. This chapter only covers the basic notions needed to understand the rest of this thesis. For an extensive reference of fMRI analysis, see for example Poldrack, Mumford, and Nichols (2011).

2.3.1 The BOLD signal

fMRI provides an indirect measure of brain activity. The most common fMRI technique, introduced in Ogawa et al. (1990), exploits the fact that when neurons in a certain area of the brain become active, the blood flow to that area increases. This increased blood flow overcompensates the oxygen consumption of firing neurons, resulting in an increase of the local blood oxygenation. This change in oxygenation can be detected with an MRI machine, and this is the signal used in fMRI. It is called the Blood Oxygenation Level Dependent (BOLD) signal.

fMRI studies therefore record the BOLD signal in the brains of subjects in a scanner. Data analysis then aims to relate this indirect measure of brain activity to the mental processes that the subjects are instructed to perform, or the stimuli they receive.

2.3.2 The Haemodynamic Response Function

The change in blood oxygenation does not happen instantaneously when neuronal activity takes place. The increase in blood flow reaches its peak around 5 seconds after an impulse of neuronal activity, and lasts several seconds. The BOLD signal measured by the MRI machine is therefore the output of a filter, due to blood flow dynamics, that is applied to the signal we are truly interested in – the neural activity. This filter can be well approximated as a Linear Time-Invariant (LTI)

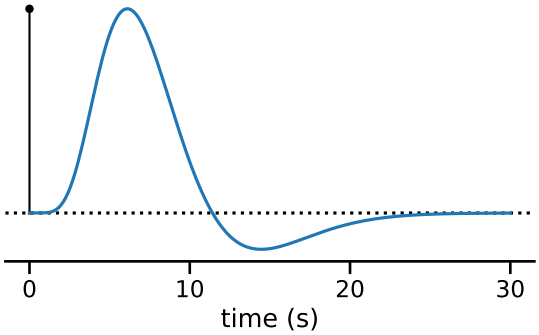


Figure 2.2: The HRF: the impulse response of the BOLD signal. The peak of the BOLD signal comes several seconds after a spike of neural activity, and is followed by an undershoot. Synthetic HRF obtained with the Nistats library (https://nistats.github.io)

filter. Its impulse response is called the Haemodynamic Response Function (HRF) and is shown in Fig. 2.2.

2.3.3 fMRI data

During an fMRI acquisition, the subject, or participant, lies in an MRI scanner. The scanner periodically records the BOLD signal in the brain of the participant. This is is typically done one 2-dimensional slice at a time, and the repetition time (TR) is around 2 seconds. During each cycle, a 3-dimensional image of the brain is acquired, which contains a BOLD signal measurement for each voxel (3-dimensional pixel). Therefore, the whole acquisition results in a 4-dimensional data array. For each voxel, we have one measurement per time point, and we may talk about voxel *time-series*.

2.3.4 Preprocessing

Before fMRI data can be analyzed, several preprocessing steps are applied: *i)* motion (movements of the subject in the scanner) is estimated and corrected, *ii)* images are co-registered, *iii)* images can be spatially smoothed, *iv)* a temporal registration can be applied to align events or conditions to which participants are subjected, and *v)* voxel time series are detrended. Spatial smoothing is often applied, usually with a Gaussian kernel, because it alleviates the issue of inter-subject variability, and it increases the statistical power of analyses by pooling information from neighboring voxels.

2.4 ANALYSIS OF TASK FMRI DATA

fMRI experiments are often categorized into two main setups: *task* fMRI, in which participants perform tasks or receive stimuli in the MRI scanner, and *resting-state* fMRI, in which participants remain idle in the scanner while their brain activity is recorded. Task fMRI is particularly relevant to this thesis, as it is an essential tool for brain mapping. Indeed, these experiments enable researchers to study the statistical link between the stimuli presented to the subjects and their brain activity. In this section, we sketch an overview of task fMRI data acquisition and analysis.

2.4.1 Experimental setup

In task fMRI, a scanner records the brain activity (the BOLD signal) of participants while they traverse a series of mental conditions. The goal is to find correlations between these mental conditions and activity in certain brain regions. Placing subjects in a mental condition can involve presenting them with certain stimuli, such as images of faces, music, or flickering checkerboards. It can also involve asking them to perform certain tasks, such as mental arithmetic, formulating a sentence, or remembering a story. Each time point can be described by the tasks and stimuli that are occurring at this instant: what is displayed on the screen, what sound is being played, whether the subject is resting or performing a task. For example, such a descriptor could take the value 1 if a face is shown on the screen and 0 otherwise. It is up to the researcher to decide which aspects of the stimuli and tasks are relevant, and to design these vectorial descriptors. The subsequent analysis aims to find correlations between these mental condition descriptors and the voxel time series – the BOLD signal recorded in each voxel at each cycle of the acquisition.

2.4.2 Modelling

The neural response is modeled as a linear function of the descriptors of tasks and stimuli (Friston et al., 1994). Remember that the BOLD signal, measured by the MRI scanner, is close to the result of a convolution of the neural activity signal with the HRF. We denote n the number of time points, m the number of voxels in the brain, and d the number of descriptors of the stimuli or mental condition. Each time point is then described by *i)* a 3-dimensional image, which is flattened to form a m-dimensional vector, and *ii)* a d-dimensional vector of mental condition descriptors. By vertically stacking these vectors to form matrices, we can write:

$$\mathbf{Y} = h \star \tilde{\mathbf{X}}\,\mathbf{B} + \mathbf{E}\,, \tag{2.1}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times m}$ is the BOLD signal, h is the HRF (assumed to be known in the simplest case), $\star$ denotes convolution, $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times d}$ are the mental condition descriptors, $\mathbf{B} \in \mathbb{R}^{d \times m}$ are the parameters of the linear model, that we need to estimate, and $\mathbf{E} \in \mathbb{R}^{n \times m}$ is unexplained noise. We strive to estimate $\mathbf{B}$, which links the measurements $\mathbf{Y}$ to the mental conditions $\tilde{\mathbf{X}}$. By computing the convolution

$$\mathbf{X} \triangleq h \star \tilde{\mathbf{X}} \,, \tag{2.2}$$

we can form a matrix of *regressors* $\mathbf{X} \in \mathbb{R}^{n \times d}$. This *design matrix* contains the descriptors of mental conditions, convolved with the HRF. It is also possible to add regressors corresponding to the convolution with derivatives of the HRF to capture small time displacements, which alleviates the need for temporal registration in preprocessing. Moreover, some regressors are included to capture the effect of nuisance parameters or confounds, such as motion parameters (movements in the scanner). Once the design matrix is built, we can write the General Linear Model (GLM):

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \,. \tag{2.3}$$

2.4.2.1 *Estimating model parameters*

In the most common case of a full rank design matrix, the linear model parameters $\mathbf{B}$ can be estimated with Ordinary Least Squares (OLS) regression:

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \,. \tag{2.4}$$

The covariance of $\hat{\mathbf{B}}_{:,j}$ for voxel (dependent variable) j is estimated by:

$$\hat{\mathbf{V}}_j = \hat{\sigma}_j^2 (\mathbf{X}^T\mathbf{X})^{-1} \,, \tag{2.5}$$

where

$$\hat{\sigma}_j^2 = \|\mathbf{Y}_{:,j} - \mathbf{X}\hat{\mathbf{B}}_{:,j}\|_2^2 \tag{2.6}$$

estimates the residual variance.

However, we should note that the BOLD noise is not independent identically distributed (i.i.d.), and in particular, it is temporally autocorrelated. Therefore, it is necessary to perform *prewhitening* before fitting the linear model. The prewhitening can be carried out by fitting a first linear model, using the residuals to estimate the covariance of the noise, and finding a basis $\mathbf{K}$ of $\mathbb{R}^n$ in which this covariance becomes the identity matrix. Each term in Eq. (2.3) is then premultiplied with $\mathbf{K}$:

$$\mathbf{K}\mathbf{Y} = \mathbf{K}\mathbf{X}\mathbf{B} + \mathbf{K}\mathbf{E} \,. \tag{2.7}$$

Once the prewhitening has been applied, $\mathbf{B}$ can be estimated with OLS.

2.4.3 Contrasts

The coefficient maps $\mathbf{B}$ estimate the effect of mental conditions on brain activity. However, each mental condition, described by stimuli or tasks, triggers a variety of mental processes in the brain. For example, pressing a button in response to a visual cue involves seeing and recognizing the stimulus, choosing the appropriate action, and moving to press the button (Poldrack and Yarkoni, 2016). Usually, we are interested in a subset of these mental processes, such as planning and executing the motor response, and want to isolate its effect.

A generally accepted hypothesis states that we can add a processing step to a task without causing important alterations to the other mental processes involved. This is sometimes referred to as the assumption of "pure insertion" (Sternberg, 1969). It then becomes possible to isolate the process of interest by contrasting carefully designed mental conditions that vary only with respect to that key process. For example, we could isolate the processes involved in the motor response by contrasting the condition in which participants press the button in response to a visual stimulus against a condition in which they see the stimulus but do not move. Based on the same logic, to isolate the effect of *recognizing* a face, we might contrast the condition "seeing a familiar face" versus the condition "seeing an unknown face". By doing so, we subtract the effects due to the process of visualizing a face and perceiving its features, and isolate the effect of the face *recognition* itself.

Applied to fMRI data, this subtraction logic leads us to compute differences, or more elaborate linear combinations, of the coefficients of the GLM that correspond to different mental conditions. As the coefficient matrix $\mathbf{B}$ has d columns, these linear combinations, known as *contrasts*, are expressed as vectors of $\mathbb{R}^d$.

2.4.4 Statistical inference

The effect size $\mathbf{B}$ does not have a meaningful scale. We are usually not interested in the effect size, but in discovering if the effect is statistically significant. For a contrast $\mathbf{c} \in \mathbb{R}^d$ on the model parameters, for each voxel $1 \leqslant j \leqslant m$,

$$\frac{\mathbf{c}^\top \hat{\mathbf{B}}_{:,j}}{\sqrt{\mathbf{c}^\top \hat{\mathbf{V}}_j \mathbf{c}}} \sim t_\nu \,, \tag{2.8}$$

where t_ν is the Student distribution with $\nu = n - d - 1$ degrees of freedom. By computing this statistic for each voxel, we can form *statistical maps*, or *contrast maps*, of the brain, in which each voxel holds a T statistic. For convenience, it is also common to convert the T statistics to Z statistics, to bypass the need to provide the degrees of freedom when sharing or interpreting the images. These brain

images holding one statistic in each voxel are often called T maps, Z maps, and β maps for the unnormalized GLM coefficients. We can compare the T (or Z) statistic at each voxel with quantiles of the Student (or normal) distribution to decide if the effect is statistically significant or not. Of course, other statistical tests are possible on the GLM parameters. For example, we can perform an F-test if we are interested in a multi-dimensional contrast. This can happen if some relevant aspect of a mental condition is captured by several descriptors, such as the one-hot encoding of a category of stimuli.

When performing such statistical tests, we should be aware that we are performing one test for each voxel, and there are hundreds of thousands of voxels in the brain. We should therefore correct for multiple comparisons; for example by using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), or the Bonferroni procedure to control the Family-Wise Error Rate (FWER) (Dunn, 1961). We can then compute a corrected significance threshold, and visualize a brain map of the T or Z statistics at each voxel for the contrast of interest. Some examples are shown in Fig. 2.3.

2.4.5 Group-level analysis

In some cases, we are not interested in the effect of a mental condition on a particular subject's brain activity, but in the population in general. Sometimes, we may also be interested in discovering how this effect varies from one population to another, such as patients suffering from a particular disease against healthy controls. Answering such questions, and drawing conclusions that hold for the population under study, is the goal of *group-level* analysis. A tutorial can be found in Thirion (2016). In order to find effects that are common across subjects, and generalize to the population, we must take into account two sources of variance: the *intra-subject* variance – the variance of the estimates of each subject's individual brain maps, and the *inter-subject* variance – the variability between subjects.

The simplest approach relies on a *random effects* model. This amounts to making the approximation that all subjects have the same level of intra-subject variance. Such an analysis begins by fitting a *subject-level* – also called *first-level* – GLM, as described in Section 2.4.2, for each subject enrolled in the study. In order to compare effects across individuals, their brain images must be registered to a common template, typically the MNI template (Section 2.1.1). This can be done before fitting the first-level GLM, or after – in this case the individuals' GLM coefficients, or β maps, are registered to the template. Then, a *second-level* – or *group-level* – GLM is fitted to regress the coefficients of individual first-level GLMs on descriptors of the individuals themselves. These regressors can indicate for example age, sex, or the presence of a

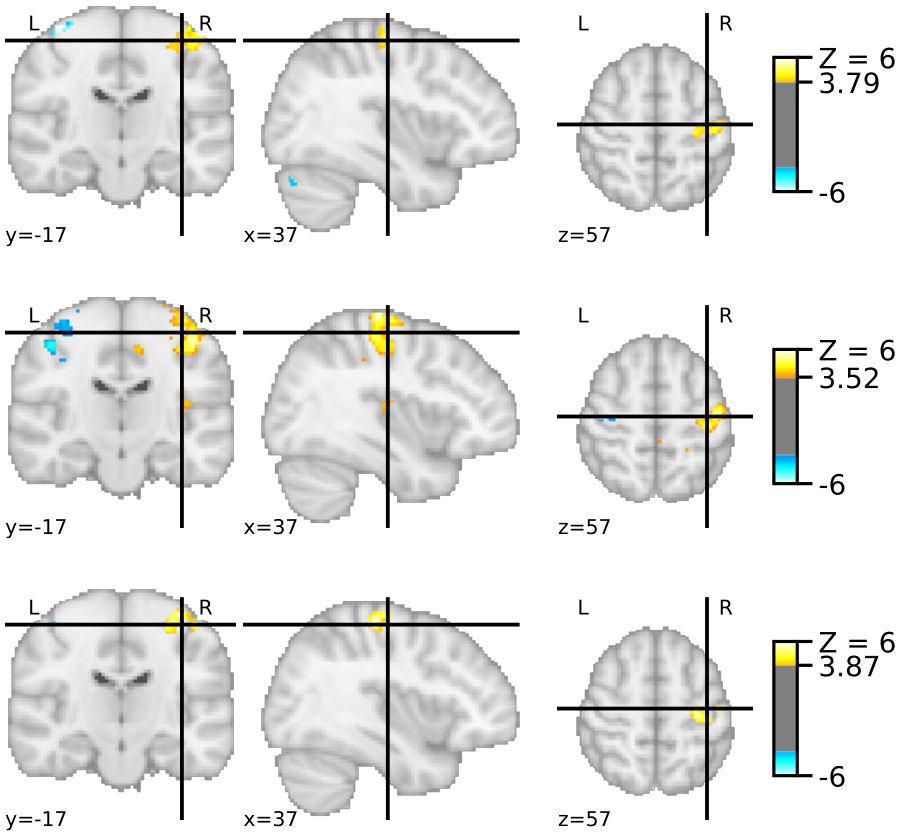


Figure 2.3: Subject-level example T maps: "left - right button press" contrast. These maps from Pinel et al. (2007), corresponding to three different subjects (participants), were obtained from a task where subjects had to press buttons either with their left hand or their right hand. A GLM is fitted to the BOLD time series, and these maps show T statistics for the contrast between the conditions "pressing a button with the left hand" and "pressing a button with the right hand", i. e. the difference between the GLM coefficients corresponding to these two conditions. We can see that this contrast results in positive values in the right motor cortex, which controls the left part of the body, and negative values in the left motor cortex, which controls the right part of the body, as expected. On this plot, we thresholded the maps to control the FDR at 5%, using the Benjamini-Hochberg procedure, as implemented by the Nistats library. We can also note that these images are registered to the MNI space, which enables easy comparison of the subjects, and that in this case brain activations are very similar across subjects. The numbers at the bottom of each plot indicate the coordinates at which the 3-dimensional image is sliced.

disease. In the simplest case, we are only interested in the mean effect of a mental condition in the population, and in this case the regressors for the second-level model reduce to a column of ones: $\mathbb{1}_N$, where N is the number of subjects. These subject descriptors form a second-level design matrix $\mathbf{Z} \in \mathbb{R}^{N \times a}$, where N is the number of subjects and a the number of regressors. For a particular first-level contrast, we can denote $\mathbf{C} \in \mathbb{R}^{N \times m}$ the stacked $\boldsymbol{\beta}$ maps (GLM coefficients) of the

individual subjects for this contrast. Then, the second-level GLM writes:

$$\mathbf{C} = \mathbf{Z}\mathbf{G} + \mathcal{E}\,, \tag{2.9}$$

where $\mathbf{G} \in \mathbb{R}^{a \times m}$ are the second-level model parameters, and the noise $\mathcal{E}$ is i.i.d. because we have assumed that all subjects have the same variance, and noise on the estimated β maps is independent from one subject to another. Then, the second-level model coefficients $\mathbf{G}$ and their variance can be estimated from a least-squares fit, in a similar way as what was done for the first-level GLM. Finally, it is possible to compute contrasts and test statistics for the second-level model coefficients, in the same way as for the first-level model. Thus, one obtains *group-level statistical maps*.

An example of such a group-level statistical map is shown in Fig. 2.4. It shows statistics for a simple one-sample test, which corresponds to the case where we are simply interested in the mean effect of a particular contrast (in this case pressing a button with the left finger vs. pressing a button with the right finger) at the population level.

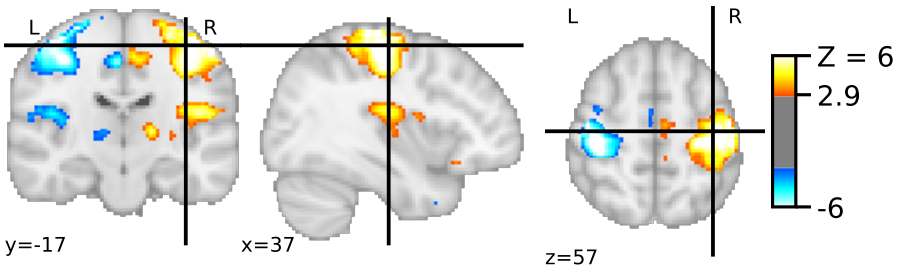


Figure 2.4: Group-level example Z map: population average of "left - right button press". This map was obtained by performing second-level analysis on the same data as Fig. 2.3 for 30 subjects, and for the same contrast: pressing a button with the left hand vs. the right hand. In this case we are interested in the mean effect in the population, hence the second-level design matrix is simply a column of ones. As was the case for individual subject maps, we see a contrast between left and right motor cortices. We also see the SMA and the secondary sensory cortex. This map was computed using Nistats.

2.4.6 Peak activation coordinates

To communicate the results of an fMRI study, it is possible to show plots of the resulting statistical maps (usually only the part above the significant threshold), as is done in Fig. 2.3 and Fig. 2.4. It is also possible (and recommended) to share the full unthresholded statistical maps on an online repository such as NeuroVault[2], after registering them to the MNI template. Finally, it is common to report *peak activation*

2 https://neurovault.org

coordinates: the locations of significant effects (brain activations) in MNI space.

To obtain peak activation coordinates, the researcher registers the statistical map to the MNI template. Then, she thresholds the map at the significance level. This thresholding forms *clusters*, connected components of above-threshold voxels in the brain. Within each cluster, she finds the position of the cluster maximum and reports its coordinates in the standard MNI coordinate system, along with the value of the test statistic at the corresponding position, and possibly other information such as the cluster size.

Tables of such peak activation coordinates appear in thousands of fMRI scientific publications. All software packages used for statistical analysis of fMRI data can produce reports that contain these tables. This allows us to extract a summary of the results of an article, in a standard spatial coordinate system, that enables comparison of findings across studies and laboratories, even when we do not have access to the actual statistical maps. These coordinates are therefore an essential resource. They are the basis of Coordinate-Based Meta-Analysis (CBMA), and we collected such coordinates from thousands of articles (Chapter 4) and used them throughout the rest of this thesis (Part II, Part III).

2.4.7 Decoding

We have seen how, after recording the BOLD signal while subjects receive stimuli or follow instructions, it is possible to find significant correlations between the presence of a mental condition and the activity of a particular brain region. This is done by regressing voxel time series on descriptors of mental conditions, as described in Sections 2.4.2 to 2.4.5. This process is called *encoding*: the fitted GLM encodes, or embeds, mental conditions into brain space.

Once these maps are obtained, we can ask which aspects of a map characterize the associated mental condition: given the statistical map, can we *predict* what is the associated mental condition? Using brain images to predict behavior or psychological conditions is called *decoding* (Cox and Savoy, 2003). Decoding models learn to identify brain activation patterns that are specific of a particular mental condition within a dataset. Therefore, they contain valuable information for brain mapping, and are complementary to encoding models (Naselaris et al., 2011). Performing decoding with very large output spaces, distinguishing a wide spectrum of mental conditions, enables formulating hypotheses or drawing conclusions about the cognitive role of particular brain regions or networks. This is known as *reverse inference* (Poldrack, 2006, 2011; Varoquaux and Poldrack, 2019). In Part IV, we train decoding models in a meta-analysis setting, relying on a dataset that regroups many different fMRI studies, tasks, and mental condi-

tions. Covering a wide range of mental conditions is a motivation for decoding meta-analyses and an important aspect of Part IV.

Decoding is a supervised learning problem. The inputs are brain images, and the targets are mental conditions. Depending on the descriptor of mental conditions that we are trying to predict, it can be a classification problem – e. g. classify maps associated with watching a face vs. watching a house as in Haxby et al. (2001), or a regression problem – e. g. predict the angle of a checkerboard. If we use a linear model, the coefficients for a particular output have the dimension of a brain map, and highlight the brain regions that are predictive of that mental condition. Being able to visualize and interpret these maps is thus a motivation for using linear models. An example of a decoding classifier's coefficient map is shown in Fig. 2.5.

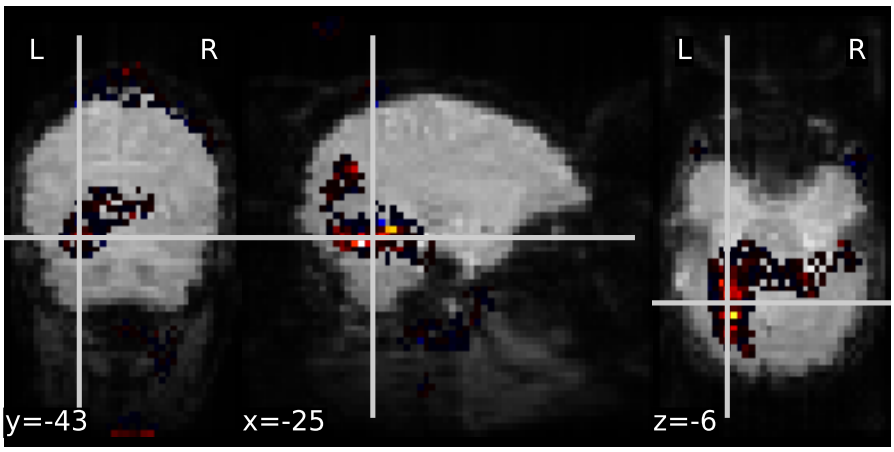


Figure 2.5: Decoding classifier weights: classifying "face" versus "house". Using data from Haxby et al. (2001), we train a logistic regression to distinguish the conditions "watching a picture of a face" and "watching a picture of a house". To cope with the high dimensionality of brain images, we apply a graph-net penalty on the classifier weights (Grosenick et al., 2013), which is why the coefficient map is sparse. In this case, the features provided to the decoder are not statistical maps, but directly the BOLD time series. This is possible without particular adjustments for this dataset because of its particularly simple block design with long durations. The resulting coefficient maps highlights the FFA, which is involved in face and object perception and recognition (Kanwisher, McDermott, and Chun, 1997). This map was obtained with the Nilearn Python library (`https://nilearn.github.io/`).

Usually, the input features are obtained from a first-level encoding GLM, fitted as in Section 2.4.2.1 (Mumford et al., 2012). We can choose to use the coefficients of the GLM (the β maps), or, more commonly, statistical contrast maps (T or Z maps, Section 2.4.4). However, it is also possible to predict mental conditions directly from the BOLD signal itself (Loula, Varoquaux, and Thirion, 2018).

Decoding is a difficult statistical learning problem, as the input features (brain maps) are very noisy and high-dimensional – there are hundreds of thousands of voxels in the brain, and sample sizes

are usually very small. Because of this high-dimensional setting, the supervised learning engines used for decoding are typically linear models, such as Support Vector Machines (SVM), logistic regression, or Linear Discriminant Analysis. Moreover, several strategies have been proposed to cope with the high dimensionality of brain images. These include feature selection (Kriegeskorte, Goebel, and Bandettini, 2006; Pereira, Mitchell, and Botvinick, 2009), dimensionality reduction based on decomposition methods such as Independent Component Analysis (ICA), Principal Component Analysis (PCA) or dictionary learning (Beckmann and Smith, 2004; Dohmatob et al., 2016; Mourao-Miranda et al., 2005; Varoquaux et al., 2010), projections on parcellations of the brain (Etzel, Gazzola, and Keysers, 2009; Hoyos-Idrobo et al., 2018), and several forms of regularization (Baldassarre, Mourao-Miranda, and Pontil, 2012; Gramfort, Thirion, and Varoquaux, 2013; Grosenick et al., 2013; Michel et al., 2011). Sparsity-inducing regularization terms, such as ℓ_1 or elastic-net penalties, are popular as they identify the relevant brain regions. Penalties that leverage the spatial structure of images – and the brain – such as Total Variation have also been successfully used.

2.5 CONCLUSION

In this chapter, we have introduced the objectives of *brain mapping* and described the analysis of fMRI data, an essential tool for this research field. In particular, we have seen how statistical analysis of fMRI data can provide maps of brain regions whose activity is significantly associated with a particular mental condition, and how these statistical maps can be registered to a standard spatial reference of the human brain, the MNI space. This is important, as most of this thesis (Part II and Part III) will focus on statistical analysis of *peak activations* – significant local maxima in statistical maps, whose MNI coordinates are reported in scientific publications.

We have also introduced the concept of *decoding*: predicting mental states, based on brain images such as encoding statistical maps. We will come back to decoding, in a large scale setting involving many conditions and fMRI studies, in Part IV.

Finally, we can note that the MNI space is used beyond reporting peak activation coordinates in task-fMRI encoding studies. The dataset we will use in subsequent chapters mostly contains fMRI peak activations, but it also includes some coordinates from other sources, which we briefly summarize here.

2.5.1 Other sources of activation coordinates

The stereotactic coordinates we will gather and analyze in subsequent parts can come from several sources, beyond task-fMRI.

2.5.1.1 *Resting-state fMRI*

We have described *task* fMRI, a paradigm where subjects perform tasks or receive stimuli while a scanner records their brain activity. Another paradigm is *resting-state* fMRI, where subjects lie idle in the scanner, without any stimuli or instructions (but hopefully without falling asleep). Resting-state fMRI enables studying brain connectivity: how the brain is spatially organized in functional networks of regions that communicate with each other. These correlation patterns across regions in the resting brain, or *connectomes*, have several uses, which include deriving biomarkers for prediction of behavior or diseases (Dadi et al., 2019). In studies involving resting-state and connectivity, it is also possible to report the locations of the regions considered (for example of a seed region whose connectivity to the rest of the brain is under study) in scientific publications.

2.5.1.2 *Other modalities*

Beyond fMRI, other modalities are extensively used for brain mapping. Positron Emission Tomography (PET) measures emissions from radiotracers that are injected in the bloodstream. Electroencephalography (EEG) and Magnetoencephalography (MEG) provide a direct measure of the brain's electrical activity. They are popular because they have a better time resolution than fMRI, even though their spatial resolution is worse and the analysis of their recordings involves a difficult source localization inverse problem.

VBM measures differences in local concentrations of brain tissue. For example, it can identify regions in which gray matter density is smaller in patients that suffer from a particular disease. As in fMRI, areas where a significant association exists can be identified with statistical tests and their positions can be reported in MNI space.

Another important modality is Diffusion Tensor Imaging (DTI), which probes the *structural* connectivity of the brain by tracking bundles of axons. Again, DTI studies can report the locations of regions where effects were found (for example significant differences in Fractional Anisotropy, which is thought to reflect important features of the white matter such as fiber density).

Finally, studies of injured patients, such as victims of strokes, may report the locations of lesions.

As long as images are registered to MNI space, results from all these modalities can be collected, pooled, and leveraged for meta-analysis, which is introduced in Chapter 3.

3 META-ANALYSIS

In Chapter 2, we have described standard statistical analysis of fMRI data, and how it can be used for brain mapping. Based on these techniques, a huge literature has grown: over 6 000 publications containing the term "neuroimaging" appear every year on PubMed[1]. Extracting reliable knowledge about the brain from this impressive amount of results poses some challenges. Most studies are conducted with few subjects, and many report results that cannot be reproduced. Moreover, many findings are specific to a particular experimental paradigm or protocol, and do not generalize to slightly different settings. Finally, many different aspects of cognition are studied almost in isolation, and integrating results across experiments and labs is difficult. Meta-analysis can address these limitations. It consists in pooling many neuroimaging studies and identifying consistently reported associations between brain regions and mental processes or diseases. In this chapter we give an overview of standard methods for meta-analysis. We also introduce a new approach, that we explore in the rest of this thesis, in Section 3.3.

3.1 A COMPLEX AND NOISY LITERATURE

In this section we summarize the main reasons why extracting comprehensive and reliable knowledge about the brain from the neuroimaging literature is difficult.

3.1.1 Power failure

Many neuroimaging studies are conducted with small sample sizes, involving less than 20 subjects. This causes a lack of statistical power and hinders reproducibility of results (Simmons, Nelson, and Simonsohn, 2011; Thirion et al., 2007). Low statistical power reduces the Positive Predictive Value (PPV), which is the probability that a positive finding reflects a true effect (Ioannidis, 2005). Publication bias and flexibility in data analysis contribute to excess significance and a high rate of false positives being reported in the literature (Button et al., 2013). Finally, some spurious findings are reported because of inadequate statistical methods, such as fixed-effect analyses that ignore the inter-subject variance, or lack of correction for multiple comparisons when

1 https://www.ncbi.nlm.nih.gov/pubmed/?term=%22neuroimaging%22

performing one statistical test for each voxel – i. e. mass univariate analysis (Poldrack et al., 2017; Thirion, 2016).

3.1.2 Lack of generalization

Besides false positives, genuine effects can be specific to an experimental setup, a particular set of stimuli or protocol, and not tied to the targeted cognitive processes (Westfall, Nichols, and Yarkoni, 2016). Thus, some studies report effects that are artifacts of an experimental protocol rather than general facts about the brain, and that fail to be reproduced in slightly different settings. Moreover, it is common for publications to extrapolate and over-interpret results.

3.1.3 Difficulty to integrate results

Even if true effects can be isolated from false positives, building a comprehensive brain atlas of mental processes and diseases entails the challenging task of integrating and summarizing genuine findings. There are many ill-defined and overlapping concepts in neuroscience, that are studied almost independently in different sub-fields. The wide range of mental processes and diseases under study, the diversity of experiments used to probe them, and the high variability of terminology make the integration of results very difficult (Fox and Lancaster, 2002; Poldrack and Yarkoni, 2016). Adapted statistical methods are therefore needed to extract general and reliable facts from the thousands of individual studies that have already been published.

3.2 META-ANALYSIS: TESTING CONSENSUS

One possible answer to the diversity and high false positive rate of neuroimaging results is meta-analysis. Meta-analysis consists in gathering many neuroimaging studies and analyzing their collective findings in a systematic way. Typically, the goal is to discover which associations between mental processes and brain regions are reported consistently across experiments. Meta-analysis can rely on brain images or on peak activation coordinates (Section 2.4.6).

In Image-Based Meta-Analysis (IBMA), we use the full statistical maps that each study computes from neuroimaging data (Section 2.4). IBMA collects unthresholded statistical maps from different studies and fits a study-level GLM, in a way that is similar to fitting second-level models as described in Section 2.4.5.

When feasible, IBMA is the preferred method (Salimi-Khorshidi et al., 2009). However, the vast majority of neuroimaging studies does not share unthresholded statistical maps. IBMA therefore cannot cover a representative portion of the literature, and for many topics of interest

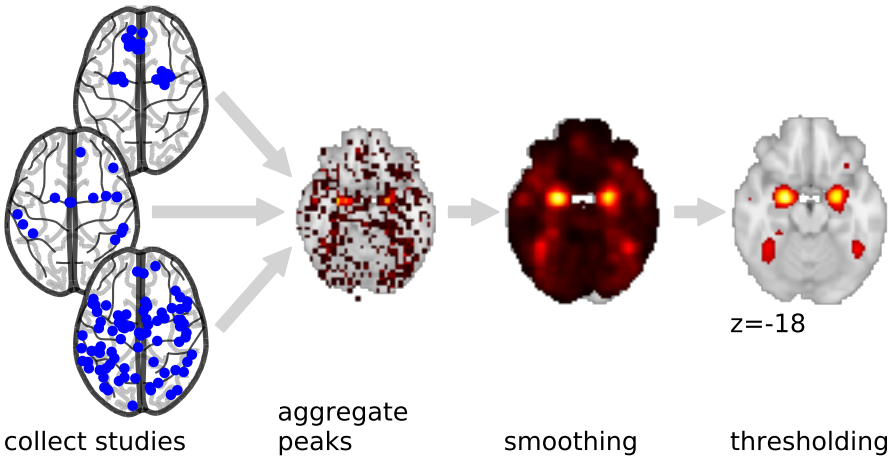


Figure 3.1: example CBMA for "emotion". This simplified illustration summarizes the CBMA process: we collect many studies that mention "emotion", extract their coordinates, estimate the coordinate density or the activation likelihood, and possibly threshold the resulting map at some significance level. On this illustration we see that the amygdala are highlighted by a meta-analysis for "emotion", as expected.

it is not (yet) possible to gather enough statistical maps for meta-analysis to achieve sufficient statistical power. Therefore, the most common approach is Coordinate-Based Meta-Analysis (CBMA), which relies on peak activation coordinates reported in scientific publications (Section 2.4.6).

3.2.1 Coordinate-Based Meta-Analysis

To perform Coordinate-Based Meta-Analysis (CBMA), we identify studies related to a topic of interest, such as "emotion". We aggregate activation coordinates reported in these studies and use them to compute either a probability of containing an activation, or an activation density, for each voxel in the brain. Then, a statistical test detects voxels in which this statistic significantly deviates from some null distribution. Usually, the null hypothesis is a uniform distribution of activations over the gray matter and the null distribution is estimated with Monte Carlo sampling.

CBMA methods differ in the way they aggregate coordinates and the meaning of the statistic they compute for each voxel, the kernel they use, and the way the null distribution is estimated. Still, they tend to follow the main steps summarized in Fig. 3.1. Here we summarize the two main approaches to this type of meta-analysis, Activation Likelihood Estimation (ALE) and Kernel Density Analysis (KDA). A more detailed description can be found in Wager, Lindquist, and Kaplan (2007).

3.2.1.1 *Activation Likelihood Estimation*

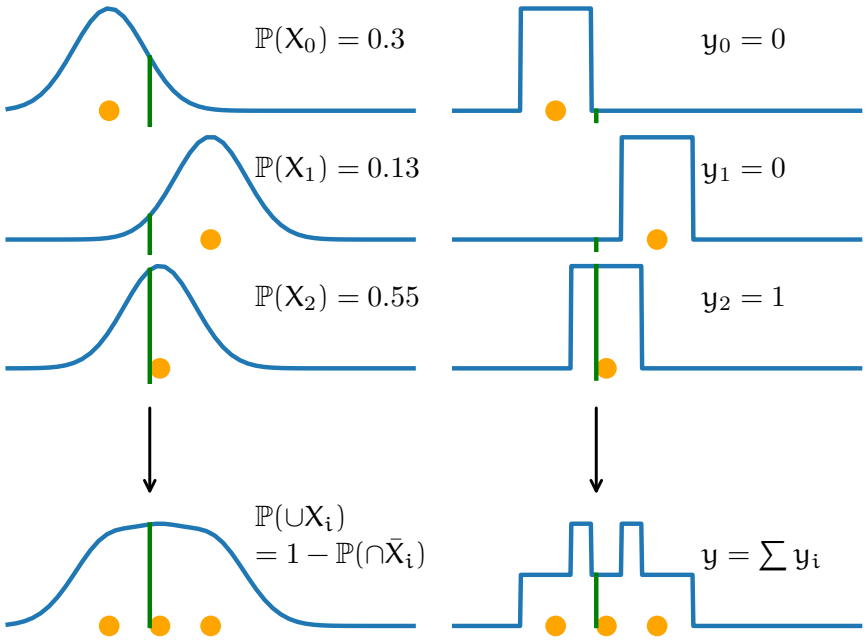
Activation Likelihood Estimation (ALE) considers that there is uncertainty about the position of brain activations (Laird, Fox, and Price, 2005; Turkeltaub et al., 2002): the location where an activation truly occurred can deviate slightly from the reported location. According to this model, the true location of an activation follows a Gaussian distribution, centered on the reported location. Analysis focuses on computing, for each voxel in the brain, the probability that at least one activation truly occurred in this voxel. This is the union of the probabilities associated with each activation. If there are n activations included in the meta-analysis and m voxels in the brain, we can denote X_{ij} the event that activation $i, 1 \leqslant i \leqslant n$ occurred in voxel $j, 1 \leqslant j \leqslant m$. The simplest form of ALE makes the important assumption that all activations are independent. Then the probability that at least one activation truly occurred in voxel j is

$$P\left(\bigcup_i X_{ij}\right) = 1 - \prod_i \left(1 - P(X_{ij})\right) . \tag{3.1}$$

ALE computes this probability for every voxel. This procedure is summarized in Fig. 3.2a.

To establish statistical significance, the most common approach is non-parametric, although some works such as Eickhoff et al. (2012) have proposed analytic estimators of the distribution of the test statistic in Eq. (3.1) under the null hypothesis. Usually, the null hypothesis is a uniform distribution of activations over the gray matter. Many (typically thousands) of sets of n (i.i.d.) locations are sampled from such a uniform distribution, and the same procedure (smoothing activations with a Gaussian kernel and computing the probability of the union of activations, Fig. 3.2a) is applied to each synthetic sample. This yields, for each voxel, a histogram of the activation likelihood under the null hypothesis, to which we can compare the value obtained from the real set of peak activation coordinates. Note that a uniform distribution of activations does not mean that the null distribution is the same for all voxels – for example due to smoothing and limit conditions, voxels close to the mask border may get lower activation likelihoods under the null hypothesis.

If we pool together all the coordinates from all studies, we ignore the inter-study variance and are performing a *fixed-effects* analysis. In this case, the discovered effects are specific to the set of considered studies and cannot be generalized to other studies. It is also possible to perform a *random-effects* analysis, where the procedure depicted in Fig. 3.2a is applied to each study or experiment separately, yielding one activation likelihood map, or Modeled Activation (MA), for each group of coordinates (Eickhoff et al., 2009). When computing the MA maps, the Gaussian kernel width can be adapted to the sample size of each study. Then, the procedure is repeated at the study level to



(a) ALE **(b)** KDA

Figure 3.2: CBMA methods. (a): ALE. Each activation yields a probability distribution over voxels. We compute the probability that at least one activation occurred in each voxel. (b): KDA. For each voxel we count the activations that were reported within a certain radius, obtaining a local density of activation. Both methods can be used in a fixed-effects analysis, where all coordinates from all studies are pooled, and a random-effects analysis, where the inter-study variance is taken into account.

compute the probability that each voxel was activated in at least one study. This random-effects analysis is reminiscent of the first and second level linear models used in standard fMRI analysis (Chapter 2). To compute the first-level maps, i. e. the MA maps, Turkeltaub et al. (2012) recommend to take the maximum of probabilities associated with each activation, rather than the union. This reduces the influence of the number and proximity of activations reported by individual studies. Moreover, they recommend to group activations by group of subjects rather than by experiment or study, so that subject groups participating in multiple experiments do not have a greater influence. The final ALE map is still computed as the union of MA maps, as in the standard ALE algorithm (Eickhoff et al., 2009).

3.2.1.2 *Kernel Density Analysis*

Kernel Density Analysis (KDA) (Wager, Jonides, and Reading, 2004; Wager and Smith, 2003; Wager et al., 2003) is another popular method for CBMA. It relies on Kernel Density Estimation (Scott, 2015; Silverman, 1986), usually using a hard ball for the kernel. A histogram of activations is computed over the brain and smoothed, yielding an esti-

mate of the density of activations. The resulting map is not interpreted as the probability that a voxel was activated at least once, as is the case in ALE, but as an expected number of activations within a certain radius of each voxel. The KDA method is depicted in Fig. 3.2b. The Monte Carlo method to obtain a null distribution for each voxel is the same as for ALE: sample many sets of coordinates from a uniform distribution and apply the procedure shown in Fig. 3.2b to each synthetic sample.

As ALE, KDA can be performed in a fixed-effects fashion, pooling together all the coordinates from all studies, or in a random-effects fashion. The latter approach is called Multilevel Kernel Density Analysis (MKDA) (Wager, Lindquist, and Kaplan, 2007). In MKDA, we start by computing one density for each study separately, before averaging over studies.

3.2.2 Large-scale and automated meta-analysis

CBMA methods such as ALE and KDA estimate the density of coordinates of a set of neuroimaging studies and compare it to the distribution of densities under a null hypothesis. Therefore, an essential step is the identification of neuroimaging studies related to the topic of interest and extraction of their peak activation coordinates. Software tools have continuously improved to make this process easier and more reproducible.

In a manual meta-analysis, we start by searching the literature for candidate publications which could be included, for example by querying PubMed[2]. Then, we inspect the retrieved studies to manually select those that are relevant to the topic of interest and should be included in the meta-analysis. This is difficult, time-consuming, and, to a large extent, subjective. Next, we manually extract the peak activation coordinates from each of the selected studies, which is tedious and error-prone. Finally, we can perform ALE or MKDA on the extracted coordinates.

This manual process limits the number of studies that can be included in the meta-analysis, hence its statistical power, and introduces variability across researchers who might select different studies. Therefore, tools and datasets have been created to enable more efficient and systematic meta-analysis. The most notable efforts are BrainMap (Laird, Lancaster, and Fox, 2005a) and NeuroSynth (Yarkoni et al., 2011), which we describe here. More details on the associated datasets are provided in Chapter 4.

2 https://www.ncbi.nlm.nih.gov/pubmed/

3.2.2.1 *BrainMap*

The BrainMap[3] database was created to enable efficient selection of studies and peak activation coordinates (Laird, Lancaster, and Fox, 2005a). It is a database of manually curated coordinates and annotations of studies. The annotations rely on a taxonomy (Fox et al., 2005) to provide rich and structured information about each study. With BrainMap, we can automatically select studies by relying on their structured metadata and the BrainMap taxonomy. This is done with the *Sleuth* software tool. We can then obtain the selected studies' peak activation coordinates, which have been manually extracted and verified, and perform ALE meta-analysis with the GingerALE software (Eickhoff et al., 2009).

BrainMap data is entered manually by neuroimaging researchers. The *Scribe* tool is used to enter the study annotations. All submitted data is examined by BrainMap staff before being added to the database. Therefore, the database features high-quality data. New studies still need to be annotated and their coordinates need to be extracted manually, but this is only done once, and in a systematic way.

3.2.2.2 *NeuroSynth*

As the neuroimaging literature's growth keeps accelerating, it can be difficult for manually curated databases to continue covering a representative proportion of publications. The NeuroSynth tool[4] was created to address this challenge (Yarkoni et al., 2011). It automatically downloads thousands of articles and extracts their peak activation coordinates. It enables users to perform meta-analyses on the resulting dataset. Studies in the NeuroSynth dataset are labelled with the abstracts of the corresponding publications. Instead of relying on structured metadata, as with BrainMap, studies are selected based on the occurrence of a single phrase, such as "working memory", in the abstracts. The coordinates contain more errors than in BrainMap, and the selection of studies is far less flexible, but NeuroSynth contains much more data and is very fast and easy to use: we only need to type the term of interest in a web interface. NeuroSynth thus enables powerful and fully automated meta-analysis.

An important aspect of NeuroSynth is that it replaces the non-parametric test of ALE and MKDA – in which the null hypothesis is usually a uniform distribution of activations over the gray matter – with a mass-univariate χ^2 test of independence between voxel activations and term occurrences in the studies' abstracts. This is computationally more efficient than the usual Monte Carlo method. Moreover, it tests a different null hypothesis and selects regions that are more specific of the term of interest. With NeuroSynth, the null hypothesis is not that

3 http://www.brainmap.org/
4 https://neurosynth.org/

activations are distributed uniformly, but that they are independent of the occurrence of the term of interest in the studies' abstracts. This yields more specific maps because regardless of the topic of study, the distribution of brain activations is not uniform, as seen in Fig. 3.3. Therefore, with enough data, a meta-analysis is likely to reject the null hypothesis of a uniform distribution in many areas of the brain, even if studies are selected randomly. Even in meta-analyses that do not have great statistical power, this phenomenon can be observed, which is one reason for only considering voxels inside the gray matter (another reason is to reduce the number of multiple comparisons). Testing the independence of a term's occurrence and a voxel's activation – as done by NeuroSynth – circumvents this issue.

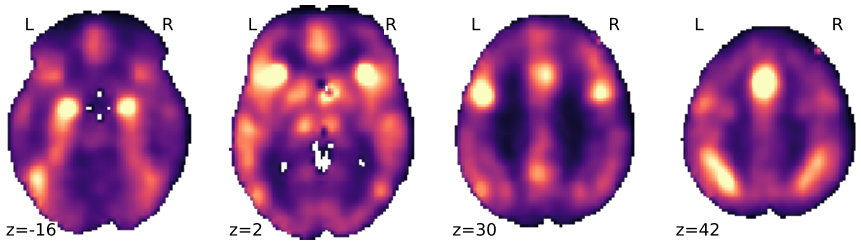


Figure 3.3: Average density computed across 13 000 neuroimaging studies. For each study, we estimate the density of its reported activations with a Gaussian KDE. Then, we average the densities across studies. As expected, the white matter contains far less activations, and it is often masked when performing meta-analysis. Within the gray matter, we see important variations: some regions such as the amygdala are often activated by neuroimaging experiments.

NeuroSynth obtains, for each study, *i)* its abstract and *ii)* its peak activation coordinates. For each study, an activation density map is obtained by convolving the peaks with a ball (of radius 10 mm), as in MKDA, and binarized. To perform a meta-analysis, NeuroSynth tests the independence between the activation of a particular voxel in a density map and the occurrence of the term of interest in the associated abstract. For a term of interest such as "emotion" and a given voxel (i, j, k), NeuroSynth builds a contingency table based on two binary criteria, resulting into four categories: *(i)* does the abstract contain the term "emotion"? and *(ii)* does the study report an activation within 10 mm of voxel (i, j, k)? Then, it applies Pearson's χ^2 test to this contingency table to test the null hypothesis that outcomes *(i)* and *(ii)* are independent. This test is repeated for every voxel (i, j, k) in the brain – as usual, the significance threshold must be corrected for multiple comparisons. An example result is shown in Fig. 3.4.

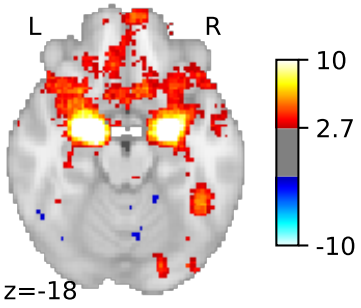


Figure 3.4: NeuroSynth map for "emotion". In above-threshold voxels, the null hypothesis of independence between brain activation and the occurrence of the term "emotion" in the study abstract is rejected. The χ^2 statistics have been converted to Z statistics for easier visualization. The map is thresholded controlling the FDR at 5%.

3.3 GOING BEYOND UNIVARIATE META-ANALYSIS

All existing meta-analysis methods, described in this chapter, rely on the same core paradigm: choosing a concept of interest, building a set of relevant studies, and establishing consensus in the associated brain images. This is true for both coordinate-based and image-based meta-analysis. Establishing consensus among published studies is essential to separate true neurological discoveries from noise and artifacts. However, the basis of meta-analysis – aggregating many studies and experiments for statistical analysis – also offers other possibilities.

Indeed, the high rate of false positives is only one of the challenges posed by the neuroimaging literature. Another important difficulty is the diversity of mental processes under study, of the concepts used to describe them, and of the experiments used to probe them. Current meta-analysis approaches enable integrating results across labs and experiments that investigate the same mental process or disease. However, using them to integrate results across domains and paradigms is more difficult. Indeed, they must build a set of studies that are relevant for a particular, well-identified mental process or disease, and it is difficult to weight selected studies based on their relevance to the topic of interest.

Existing methods do not provide mechanisms to combine results for different concepts. They perform in-sample statistical inference, and are not designed to interpolate between studies or make predictions for unseen data. This means that it is difficult to obtain results outside well-defined and frequently studied psychological concepts. These constraints are mostly due to the goal of standard meta-analysis and the need for a well-defined and meaningful null hypothesis.

Moreover, even for frequently studied concepts, building a set of relevant studies is difficult with existing fully automated frameworks such as NeuroSynth. Indeed, NeuroSynth relies on exact matching of single terms or expressions, whereas the terminology used in the neu-

roscience literature varies widely. NeuroSynth's univariate approach, based on the binary association of studies with a single expression, is therefore limited when we need to combine terms and synthesize a brain map for topics that are better described with a set of keywords, an abstract or a full paper. These limitations are discussed further in Part III.

3.3.1 Mapping text to brain regions

In this thesis, we introduce a new approach to meta-analysis. We use a large dataset of neuroimaging studies to build *predictive* models. The focus shifts from establishing consensus for a particular concept to *learning* multivariate mappings from mental diseases and psychological states to anatomical structures in the brain. This approach is complementary to methods such as ALE, MKDA or NeuroSynth: while standard meta-analysis performs statistical tests to evaluate consistency and trustworthiness of results among past studies, the new framework predicts, based on the description of an experiment or concept of interest, which brain areas are most likely to contain related neurological observations.

This paradigm shift enables new applications of meta-analysis. For example, it enables embedding any text related to neuroscience into brain space. This permits analyzing the spatial and anatomical information contained in the dozens of thousands of publications, case studies, and health records that are not labelled with images nor activation coordinates. The new framework also makes it possible to generate hypotheses for new experiments, or statistical priors for analyzing new neuroimaging results. For this application, it removes the restriction of meta-analysis to single, frequently-studied concepts. It also provides a natural way to evaluate and compare methods, by measuring their prediction performance on out-of-sample neuroimaging studies.

In Part II, we formalize this predictive framework, propose an adapted statistical model, quantify its performance and show how it can be used to analyze neuroimaging studies that do not provide images nor coordinates. In Part III, we improve the model described in Part II to extend the class of queries that it can map to brain regions. We also perform extensive experiments that show how the new model can compile meaningful results for challenging queries, such as rare terms or seldom-studied diseases. The evaluation methods introduced in Part III can also be useful to compare more traditional meta-analysis frameworks. We also introduce a new online tool for predictive meta-analysis of the neuroimaging literature[5]. Finally, in Part IV, we shift from encoding to decoding: predicting labels associated with neuroimaging studies, based on their statistical brain maps.

5 https://neuroquery.saclay.inria.fr

This image-based meta-analysis framework is the first to decode a wide variety of mental processes from brain images.

Most of this work relies on a new dataset that we created. It is the largest existing dataset of neuroimaging text and peak activation coordinates. Therefore, in Chapter 4, we start by describing this dataset and how we assembled it.

4 BUILDING A DATASET

In this chapter, we describe our dataset of neuroimaging publications and stereotactic activation coordinates, and explain how it is constructed. This dataset contains 13 498 full-text articles and the locations of their 418 772 peak activations. It is by far the largest dataset of its kind. Indeed, the NeuroSynth corpus (Yarkoni et al., 2011) contains a similar number of studies (14 371), but only exploits the article's abstracts. As our corpus contains the full articles, it amounts to around 20 times more text and provides a richer indexing of studies and their coordinates. Our dataset also contains a similar number of coordinates as the NeuroSynth dataset – which contains 448 255 after removing duplicates, but manual evaluation reveals that our coordinates are extracted with a higher precision (Section 4.4.3). We refer to our dataset as the NeuroQuery dataset, as it is distributed with the NeuroQuery tool described in Part III,

4.1 THE NEED FOR A NEW DATASET

Coordinate-based meta-analysis relies on estimating the spatial density of reported peak activations and finding the statistical link between this density and the description of mental conditions or disorders under study. Therefore, it is crucial to have access to a large dataset of labelled stereotactic coordinates. Coordinates are labelled either with manually curated meta-data or with text extracted from articles in which they were reported. Activation coordinates are a sparse and noisy signal, and many studies must be included for meta-analyses to have sufficient statistical power. The dataset size is therefore of paramount importance.

BrainMap was the first large-scale dataset to appear (Laird, Lancaster, and Fox, 2005a). It regroups coordinates that are manually entered and annotated according to a fixed taxonomy (Fox et al., 2005). According to the BrainMap website[1], in July 2019, the BrainMap database regroups 159 843 locations for 4 561 publications (grouping the functional and VBM databases). However, BrainMap does not contain the text of articles from which coordinates were extracted. Thus, only a restricted set of terms – included in the taxonomy and used in the annotations – can be mapped onto brain structures. Moreover, the amount of data increases slowly because the collection is not au-

1 http://www.brainmap.org

tomated and relies on a community effort (Laird, Lancaster, and Fox, 2005b).

NeuroSynth (Yarkoni et al., 2011), an even larger dataset, was created by automatically downloading articles and extracting peak activation coordinates. In July 2019, NeuroSynth contains 448 255 unique locations for 14 371 studies. It also contains the term frequencies for 3 228 terms (1 335 are actually used in the NeuroSynth online tool[2]), based on the abstracts of the studies.

For this thesis, we ruled out using the BrainMap database because *i)* it is much smaller than NeuroSynth, *ii)* it does not contain the text of the publications, and *iii)* it is not openly shared. We considered using the NeuroSynth dataset. However, it only contains term frequencies for abstracts, without the articles' full text. This results in a shallow description of the studies, based on a very short text (around 20 times smaller than the full article). As a result, many important terms occur rarely: they seldom appear in abstracts, and can be associated with very few studies. For such terms, meta-analysis lacks statistical power. When the full text is available, the number of term occurrences – associations between a term and a study – increases significantly (Fig. 4.1). Therefore, the dataset is richer, terms are better described by their set of associated studies, and meta-analyses have more statistical power. In addition, as publications are subject to copyright and restricted distribution, NeuroSynth only provides term frequencies for a fixed vocabulary and not the text they are extracted from. Finally, NeuroSynth's coordinate extraction procedure can be improved, as discussed in Section 4.4.3.

To overcome these limitations we create a new dataset, the NeuroQuery dataset. It contains a similar number of studies as the NeuroSynth dataset, but provides the full text rather than the abstracts only. We also create a new coordinate extraction tool, producing less noisy brain activation data.

4.2 GATHERING ARTICLES

To build our dataset, we first need to collect a large number of scientific articles related to neuroimaging, from which we can extract text and activation coordinates. Our primary goal is to use the full text of the articles. Thus, it is important to retrieve them in a well-structured format, in order to extract the relevant sections. NeuroSynth obtains the full-text articles by scraping HTML pages from several scientific journals' websites. Then, NeuroSynth extracts coordinates from tables in these pages. However, while it is relatively easy to find tables, it is more challenging to extract the article's text and separate it from other content, such as advertisements, links to other articles, references, or

2 http://neurosynth.org

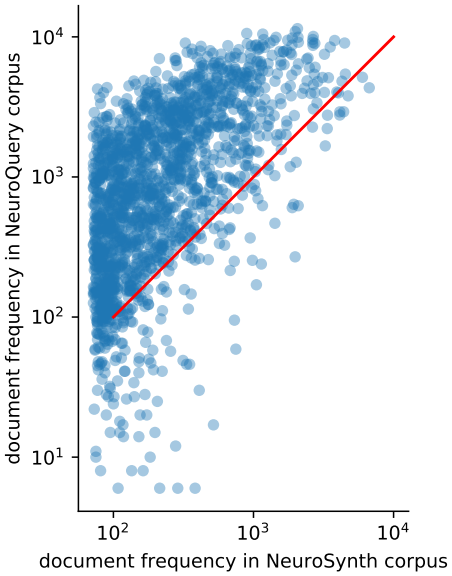


Figure 4.1: Document frequencies (number of documents in which a word appears) for terms from the NeuroSynth vocabulary, in our corpus and in NeuroSynth. Words occur in fewer documents in the NeuroSynth corpus because it only contains abstracts. Even when considering only terms present in the NeuroSynth vocabulary, our corpus contains over 3M term-study associations – 4.6 times more than NeuroSynth.

acknowledgments. This extraction becomes a daunting task if it has to be done for many journals' websites with different layouts and markup conventions. Furthermore, within each journal, the structure of the page can vary. As a result, the pages downloaded by NeuroSynth are not usable beyond coordinate extraction. To obtain text, NeuroSynth downloads separately the abstracts from the PubMed Central API[3] using BioPython[4]. Additionally, websites can change their structure frequently, so a collection of several scrapers is brittle and difficult to maintain.

4.2.1 Data sources

Rather than scraping HTML pages from individual journals' websites, we download articles in bulk from platforms that aggregate articles from many journals and distribute them in a structured format. We focus on PubMed Central (PMC) and Elsevier. PMC, one of the National Center for Biotechnology Information (NCBI) databases, contains over 5M articles. It is the largest free full-text archive of medical publications[5]. Elsevier is a commercial publisher. We use Elsevier because *i)* it contains several important journals which are not free and therefore

3 https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch
4 https://biopython.org/DIST/docs/api/Bio.Entrez-module.html
5 Note that PMC is not the same database as PubMed; PubMed tracks citations while PMC archives full-text articles: https://www.nlm.nih.gov/bsd/difference.html.

not accessible through PMC, *ii)* our institution (INRIA) has an account that gives us access to Elsevier content, and *iii)* Elsevier exposes an API which allows downloading articles in bulk in a uniform XML language[6].

We should note that some important neuroscience journals, such as the Nature journals, are almost absent from our two data sources. Thus, there is a selection bias with respect to the whole population of neuroscience publications. We hypothesize that this does not introduce an important bias in the reported links from mental functions or disorders to brain structures but this assumption has not been checked.

4.2.2 Data collection architecture

All downloaded data is in XML format. In this first downloading step, almost no transformation is performed. We store downloaded elements in documents that also contain some provenance information, such as the platform they were downloaded from and the query used to retrieve them. We validate the results using XSD. Documents that are not valid are discarded. Valid documents are stored in XML files. As a consequence, we can index them with a appropriate tools, and explore and query our corpus with XQuery. We use BaseX[7] for this (Grün, Holupirek, and Scholl, 2007).

In order to discover and download articles, we rely on the Entrez Programming Utilities, or E-Utilities[8] (Sayers, 2009). They are a set of server-side programs that provide a programmatic interface to NCBI resources. We use four E-Utilities: *i)* ESearch for discovering articles that match a query, *ii)* ESummary for obtaining meta-data, *iii)* ELink for URLs where the articles themselves may be found, and *iv)* EFetch to download the full-text articles when they are available in PMC. We also use the Elsevier API to retrieve some full-text articles. An overview of the data collection architecture is shown in Fig. 4.2.

4.2.2.1 *Discovering articles and full-text providers*

Using ESearch, we query the PubMed database to discover relevant articles. We search either for all articles from a particular journal or based on keywords such as "fMRI" or "brain mapping". The UIDs – i. e. PMIDs, as we query the PubMed database – of matching articles are stored on the Entrez History server. Using the keys that identify our search results on the history server, we query ESummary to retrieve articles' meta-data in batches. This meta-data contains information such as the title, journal, and publication type. Next, we upload the PMIDs to Elink, which lists the available URLs to providers of the full text or other information about the articles. We use ELink results to

6 https://dev.elsevier.com/api_docs.html
7 http://basex.org/
8 https://www.ncbi.nlm.nih.gov/books/NBK25501/

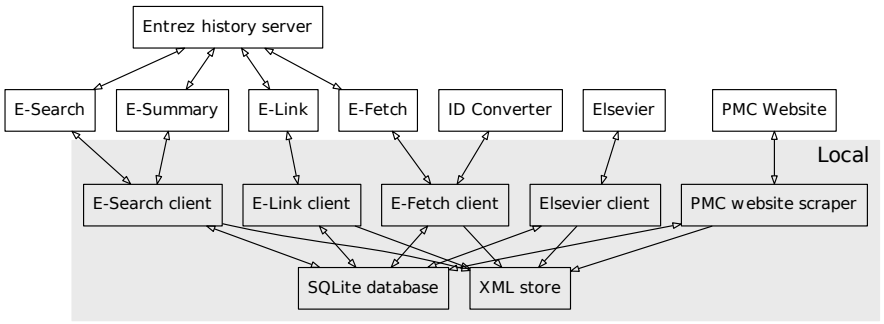


Figure 4.2: Data collection architecture. Several clients gather information from web services and store it in XML files. E-Search and E-Summary provide PMIDs and metadata about articles that match a query such as a set of journals. E-link provides lists of URLs from which articles may be downloaded. Based on E-link results, the full text can be downloaded from E-Fetch, Elsevier, or the PMC website. The ID Converter converts from PubMed IDs to PMC IDs. Search results can be stored in the Entrez history server, which avoids uploading them when re-using them with different Entrez utilities. The different clients share data such as IDs of articles to download or URLs where they might be found in an SQLite database. All documents are validated with XSD before entering the XML store and documents that fail to validate are discarded.

discover which articles may be available on PMC or on the Elsevier platform.

4.2.2.2 *Downloading articles*

The procedure described in Section 4.2.2.1 yields the PMIDs of relevant publications and the platforms from which they are available. Next, we can download the articles' full text.

PMC OPEN ACCESS Articles in the Open Access subset of PMC are downloaded in batches using Efetch. First, PubMed IDs are converted to PMC IDs using the NCBI's ID converter[9]. The resulting PMC IDs are uploaded to the history server. Using ESearch, we create a new result set on the history server by adding the "open access" filter. Finally, we use Efetch and the keys for this result set to retrieve the batches of full-text articles in XML format. Some articles are in the PMC database and freely available, but not in "open access", meaning that they cannot be downloaded through the API due to restrictions imposed by their publisher. We download these articles by scraping the PMC web interface.

PMC For articles that are in the PMC database but not accessible in XML format from the API, we download the corresponding page from

9 https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/

the PMC website[10]. We limit our tool to one request every 2 seconds to avoid causing an important load on the server. As the website tries to prevent automatic tools from downloading articles, it is necessary to send requests with headers that seem like they could come from a web browser and to change one's apparent IP address regularly. The resulting pages are obtained in XHTML format. As they are not meant for programmatic use, they are more difficult to exploit than results retrieved from an API. However, in this first downloading step, we only make sure they are well-formed. We do so by parsing them with the BeautifulSoup[11] Python library (Richardson, 2007), trying to fix them with this tool if they are malformed, and discarding them if they cannot be fixed. Scraping a website is far less efficient than querying an Application Programming Interface (API). Unlike NeuroSynth, we only perform this operation for one website, whose pages are in eXtensible HyperText Markup Language (XHTML) format and have a relatively uniform structure.

ELSEVIER When Elink information indicates that an article is available in Elsevier, we download it from the Elsevier Article Retrieval API[12] using INRIA's authentication information. As Elsevier limits such downloads to a fixed number of articles per week, we schedule the download over several weeks.

4.2.2.3 *Asynchronous data collection*

Data is collected by five asynchronous processes, one for each queried web service: one process searches for articles and downloads their meta-data with ESearch and ESummary, one downloads link information from Elink, and the other three download the full text articles from Elsevier, the PMC API, and the PMC website, respectively. These processes create XML files, but also store and share relevant information (such as the discovered articles and URLs, or the PMIDs of articles which have been downloaded from each platform) in an SQLite[13] database – which is a single binary file (Owens, 2006).

4.3 TRANSFORMING ARTICLES

The downloaded articles are written in one of three XML languages. Those downloaded from the PMC website are in XHTML, those that come from Elsevier are in this publisher's "`xocs`" language[14], and those that come from the PMC API use the Journal Archiving and

10 https://www.ncbi.nlm.nih.gov/pmc/
11 https://www.crummy.com/software/BeautifulSoup/bs4/doc/
12 https://dev.elsevier.com/documentation/ArticleRetrievalAPI.wadl
13 https://www.sqlite.org/index.html
14 https://schema.elsevier.com/dtds/document/fulltext/xcr/xocs-article.xsd

Interchange Tag Set[15] from the Journal Article Tag Suite (JATS), created by the National Library of Medicine (NLM). In order to treat them uniformly in the rest of the pipeline, the second step is to convert all articles to a single language. We use the Archiving Tag Set because many of the documents we download already use this tag set. In addition, it is flexible, permissive, well documented, and used in many applications. We use eXtensible Stylesheet Language Transformations (XSLT) to translate documents from XHTML or Elsevier's `xocs` to the Archiving Tag Set. Translated documents are validated using the schema provided by NCBI[16]. At the end of this step, we obtain over 149 000 full-text journal articles (with unique PMIDs) that validate against the Archiving Tag Set's schema.

4.4 EXTRACTING TABLES AND COORDINATES

Once all articles are in a single XML language, extracting their text or specific parts such as the abstract or keywords is straightforward. The remaining challenge is to extract and normalize the tables and then identify and extract the peak activation coordinates.

4.4.1 Extracting tables

We use XSLT to extract tables, translate them to the XHTML table format, normalize them, and store them in a separate XML file for each article. Tables for the articles obtained from the PMC website are not embedded in the XHTML page, but have to be retrieved separately. We extract the table IDs from the article and download the web pages that contain the tables from the PMC website.

4.4.1.1 *Translating CALS to XHTML*

The Archiving Tag Set allows tables to be described using either the XHTML 1.1 table model or the OASIS XML Exchange table model, which is based on the CALS table model[17]. Elsevier uses its own table model, based on CALS. We transform Elsevier tables to CALS during the translation of all articles to JATS (Section 4.3). When extracting tables from our articles, we translate them to the XHTML 1.1 table model using stylesheets provided by docbook[18].

15 https://jats.nlm.nih.gov/archiving/
16 https://jats.nlm.nih.gov/archiving/1.2/xsd.html
17 https://jats.nlm.nih.gov/archiving/tag-library/1.2/element/table.html
18 https://docbook.org/tools/

4.4.1.2 *Normalizing tables*

Cells in tables can span several rows and columns. NeuroSynth's chaotic parsing of tables sometimes results in coordinates from different triplets getting mixed up or part of the table missing (e. g. PMIDs 15488911, 15862201). When extracting a table, we normalize it by splitting cells that span several rows or columns and duplicating these cells' content; the normalized table has the shape of a matrix. The original table and the normalized version are stored together. Finally, all unicode characters that represent "+" or "-" signs (such as `&#x2212;` "MINUS SIGN") are mapped to their ASCII equivalents, "+" (`&#x2b;` "PLUS SIGN") or "-" (`&#x2d;` "HYPHEN MINUS") using XSLT's `character-map` element. While this may seem obvious, many coordinates lose their "-" sign or disappear during NeuroSynth's coordinate extraction because it overlooks this issue (e. g. PMIDs 15528081, 15589105), or because of a similar issue with whitespace.

4.4.2 Extracting and filtering coordinates

Once tables are isolated, in XHTML format, and their rows and columns are well aligned, the last step is to find and extract peak activation coordinates. This extraction is not performed in XSLT but in Python. Heuristics find columns containing either single coordinates or triplets of coordinates based on their header and the cells' content. A heuristic detects when the coordinates extracted from a table are probably not stereotactic peak activation coordinates, either because many of them lie outside a standard brain mask or because the group of coordinates as a whole fits a normal distribution too well. In such a case, the whole table is discarded. Finally, coordinates are stored in a csv file. This file also contains, for each coordinate triplet, the identifier and label of its table, as well as the identifier of its document of origin. Out of the 149 000 downloaded and formatted articles, 13 498 contain coordinates that were extracted by this process, resulting in a total of 418 772 locations.

All the extracted coordinates are treated as coordinates in the MNI space (Section 2.1.1), even though some articles still refer to the Talairach space. The precision of extracted coordinates could be improved by detecting which reference is used and transforming Talairach coordinates to MNI coordinates. However, differences between the two coordinate systems are at most of the order of 1 cm, and much smaller in most of the brain. This is comparable to inter-subject variability after spatial normalization and to the size of Gaussian kernels typically used to smooth images. Moreover, the alignment of brain images does not only depend on the used template but also on the registration method, and there is no perfect transformation from Talairach to MNI space (Lancaster et al., 2007). Therefore, treating all

	False positives	False negatives
NeuroSynth	20	28
NeuroQuery	3	8

Table 4.1: Number of extracted coordinate sets that contain at least one error of each type, out of 40 manually annotated articles. The articles are chosen from those on which NeuroSynth and NeuroQuery disagree – the most likely to contain errors.

coordinates uniformly is acceptable, but better handling of Talairach coordinates is a clear direction for improving the NeuroQuery dataset.

4.4.3 Coordinate extraction quality

We evaluate the coordinate extraction process by checking the articles for which NeuroSynth and NeuroQuery disagree. Out of 8 692 articles in the intersection of both corpora, the extracted coordinates differ (for at least one coordinate) in 1 961 (i. e. in 23% of articles). We select the first 40 articles (sorted by PMID) and check the extracted coordinates manually. As shown in Table 4.1, our method extracts false coordinates from far fewer articles: 3 out of 40 articles have at least one false location in our dataset, against 20 for NeuroSynth. While these numbers may seem high, note that errors are far less likely to occur in articles for which both methods extract exactly the same locations.

4.5 BUILDING THE VOCABULARY

Another essential resource is our vocabulary: the set of words or phrases used to tokenize text and parametrize its representation in a vector space. Using a vocabulary comprising only words or expressions relevant to neuroscience reduces the amount of noise and the dimensionality of the vector representations. In addition, we want to recognize some phrases, or "n-grams", as expressions that designate a single entity – for example "white matter", "Parkinson's disease", or "Default Mode Network". Therefore, we do not use all the words found in our corpus as is often done. Instead, we build a vocabulary of terms related to neuroscience by leveraging several manually curated ontologies or controlled vocabularies.

4.5.1 Vocabulary sources

Our vocabulary comprises five important lexicons of neuroscience, based on community efforts: the subset of Medical Subject Headings

(MeSH) (Lipscomb, 2000) dedicated to neuroscience and psychology[19], Cognitive Atlas (Poldrack and Yarkoni, 2016), NeuroNames (Bowden and Martin, 1995), and NIF (Gardner et al., 2008). We also include all the terms and bigrams used by NeuroSynth (Yarkoni et al., 2011). Finally, the vocabulary also comprises the labels of 12 anatomical atlases, listed in Appendix A.1.

CognitiveAtlas[20] is an ontology of cognitive processes, tasks and phenotypes. We include it in our vocabulary and we also use it in Chapter 10 to annotate brain images with cognitive labels.

4.5.2 Vocabulary extraction

We obtain the ontologies as Resource Description Framework (RDF)[21] graphs, which we parse using the `rdflib` library[22]. We query each ontology with SPARQL to construct a simplified graph based on a subset of the Simple Knowledge Organization System (SKOS)[23] RDF vocabulary. We merge the resulting graphs. We also annotate some labels as belonging to a category – either anatomy, pathology, or psychology – based on the label's source (for example NeuroNames is only concerned with anatomy, the MeSH tree is organized by topic, etc.). Although this graph could be useful in itself, it is not exploited in the rest of this thesis, as we focused more on continuous associations between terms derived from co-occurrence statistics than on ontology relations. Hence, we do not describe this graph at length. In order to construct our vocabulary, we simply gather all the objects of `skos:literalForm` relations.

Extracting all literal forms yields a list of strings, to which we add all the terms from NeuroSynth's vocabulary. All these strings are then tokenized to identify the individual words that constitute them. When the tokenization yields several words, we include both the full expression (the n-gram) and the individual words in the vocabulary. We thus obtain 40 149 terms and expressions. Many expressions extracted from the ontologies are very rare, especially when they contain several words, such as "accessory medullary lamina globus pallidus". We discard all the terms and expressions that occur in less than 5 / 10 000 articles. The resulting vocabulary contains 7 547 terms and expressions related to neuroscience.

We show the number of terms extracted from each source and the size of the different vocabularies' intersections in Table 4.2. These manually curated vocabularies display relatively low overlaps, revealing that there is still no fully established terminology in neuroscience. This constitutes a motivation for associating peak locations with text

19 MeSH are the terms used by PubMed to index articles
20 http://www.cognitiveatlas.org/
21 https://www.w3.org/TR/rdf11-concepts/
22 https://rdflib.readthedocs.io/en/stable/
23 http://www.w3.org/TR/skos-reference

% of ↓ contained in →	Cognitive Atlas (895)	MeSH (21287)	NeuroNames (7146)	NIF (6912)	NeuroSynth (1310)	NeuroQuery (7547)
Cognitive Atlas	100%	14%	0%	3%	14%	68%
MeSH	1%	100%	3%	4%	1%	9%
NeuroNames	0%	9%	100%	29%	1%	10%
NIF	0%	12%	30%	100%	1%	10%
NeuroSynth	9%	14%	5%	5%	100%	98%
NeuroQuery	8%	25%	9%	9%	17%	100%

Table 4.2: Intersections of controlled vocabularies. NeuroQuery contains all the terms from the other vocabularies that occur in more than 5 out of 10 000 articles. "MeSH" corresponds to the branches of PubMed's MEdical Subject Headings related to neurology, psychology, or neuroanatomy. The NeuroSynth vocabulary is extracted from the NeuroSynth corpus and manually curated. Terms from the NeuroSynth vocabulary that are not included in our vocabulary are bigrams such as "gyrus stg", which are part of longer expressions – "supplementary temporal gyrus (stg)" – and are tokenized differently in our text preprocessing (see Appendix B.4 for the list of discarded NeuroSynth terms and Appendix B.3 for details on tokenization).

(abstracts or full articles), rather than a few entries from a taxonomy as in BrainMap. Indeed, taxonomies are incomplete and do not always reflect the terminology used in practice by researchers. For example, 32% of labels from the CognitiveAtlas ontology are used in less than 0.05% of articles from our corpus.

4.6 SUMMARY OF COLLECTED DATA

The data collection described in this chapter provides us with important resources, which we exploit to train the models described in the rest of this thesis: *i)* over 149K full-text journal articles related to neuroscience – 13K of which contain peak activation coordinates – all translated into the same structured format and validated; *ii)* over 418K peak activation coordinates for more than 13K articles; *iii)* a vocabulary of 7 547 terms related to neuroscience, each of them occurring in at least 6 articles from which we extracted coordinates. This dataset is the largest of its kind. In the rest of this thesis we focus on the set of 13K articles from which we extracted peak locations.

4.6.1 Text

In terms of raw amount of text, this corpus is 20 times larger than NeuroSynth's. Combined with our vocabulary, it yields over 5.5M

occurrences of a unique term in an article. This is over 5 times more than the word occurrence counts distributed by NeuroSynth[24]. When considering only terms in NeuroSynth's vocabulary, the corpus still contains over 3M term-study associations, 4.6 times more than NeuroSynth. For the rest of this thesis, using this larger corpus will result in denser representations, higher statistical power, and coverage of a wider vocabulary.

4.6.2 Coordinates

The set of extracted coordinates is almost the size of NeuroSynth's – which is 7% larger with 448 255 coordinates) – and is less noisy. To compare coordinate extractions, we manually annotated a small set of articles for which NeuroSynth's coordinates differ from ours. Compared with NeuroSynth, our extraction method reduces the number of articles with incorrect coordinates (false positives) by a factor of 7. It also reduces the number of articles with missing coordinates (false negatives) by a factor of 3 (Table 4.1). For subsequent work, this will result in less noisy brain activation data.

4.6.3 Sharing data

We cannot share the full text of the articles because of copyright laws, but the vocabulary, extracted coordinates, and term occurrence counts for the whole corpus are freely available online[25].

The dataset presented in this chapter is exploited in Part II and Part III. Having text and the locations of neurological observations for thousands of studies enables us to learn the associations between mental processes or diseases and brain structures in a supervised learning setting.

24 https://github.com/neurosynth/neurosynth-data
25 https://github.com/neuroquery/neuroquery_data

Part II

MAPPING TEXT TO BRAIN LOCATIONS

5 TEXT TO SPATIAL DENSITIES

In the next two chapters, we build the framework for predictive meta-analysis. Our goal is to map text onto brain regions: given a document, such as an article or a case report, we want to synthesize a brain map of the associated anatomical structures. Translating text into brain maps can provide access to under-exploited information contained in medical publications, case reports and health records. Indeed, neuroimaging observations are usually reported and stored in text documents. Even in the academic literature, most publications do not provide the data that they used or the statistical maps that they computed. Being able to extract spatial information from such documents can yield new insights. Therefore, we need a principled way to map, or *encode*, documents onto locations in the brain.

Encoding whole documents departs from standard meta-analyses. We do not aim to perform a statistical test to evaluate the consistence of results reported in different studies. Instead, we frame encoding as a prediction problem: given text, we can *predict* the spatial distribution of the observations that it describes. We are not computing test statistics but predictions for out-of-sample neuroimaging studies. In this setting, we can compare predicted distributions with the locations truly reported in neuroimaging studies. Encoding then becomes a supervised learning problem, and we can benefit from a standard framework to estimate the mapping from text to brain locations, and to measure performance and compare models.

In this chapter, we formalize the problem of mapping text documents to brain images and propose a regression model. We refer to this model as "text-to-brain". In Chapter 6, we apply this approach to our dataset, present quantitative and qualitative assessment of text encoding models, and show how such models can be used to analyze a large corpus of pure text.

5.1 PROBLEM SETTING

We want to map text, such as articles and case reports, to brain regions. More precisely, we want to predict the spatial density of the observations that the text describes. One reason for focusing on this spatial density is that it does not depend on the number of observations reported in each study. The number of reported coordinates depends on the number of computed contrasts and on details of the data analysis procedure, and it is not meaningful when pooling a wide variety of

studies. To learn the mapping from text to spatial densities, we use the neuroimaging studies from our dataset, described in Chapter 4. For each study, we have the text extracted from the article, and a set of peak activation coordinates.

We denote S a study, associated with a text T, and a set of $c \in \mathbb{N}_{>0}$ peak locations $\mathcal{L} = \{l_a \in \mathbb{R}^3, a = 1 \dots c\}$. The peak locations are realizations of a continuous random variable that has a probability density function (pdf) p over the brain. Our goal is to predict this pdf, based on the text of the study. We denote q our predicted pdf. A predicted density q should be close to the true pdf p, or take high values at the coordinates actually reported in the study: the prediction is good if $\prod_{l \in \mathcal{L}} q(l)$ is large.

In a supervised learning setting, we start from our collection of studies $\{S = (T, \mathcal{L})\}$, with T the text and $\mathcal{L}$ the locations. Building the prediction engine then entails the choice of a model relating the density p to the text T, the choice of a loss, or data-fit term, and some regularization on the model parameters. We now detail how we make each of these choices to construct a prediction.

5.2 REGRESSION MODEL

We start by modelling how the spatial density p of a study depends on the text T. First, we choose a representation for the text. Then, we define our model space – how we represent and parametrize our predicted spatial density, and the class of functions to which belongs the mapping from T to the prediction q.

5.2.1 Representing text: TFIDF features

To make a prediction based on the text T of a study, we start by building a vector representation of T. We represent a document by its Term Frequency · Inverse Document Frequency (TFIDF) features (Salton and Buckley, 1988), which are reweighted Bag-of-words features. A TFIDF representation is a vector in which each entry corresponds to the (reweighted) frequency of occurrence of a particular term. The *term frequency*, tf, of a word in a document is the number of times the word occurs, divided by the total number of words in the document. The *document frequency*, df, of a word in a corpus is the proportion of documents in which it appears. The *inverse document frequency*, idf, is defined as:

$$\text{idf}(w) = -\log(\text{df}) + 1 = -\log \frac{|\{ i \mid w \text{ occurs in } T_i \}|}{n} + 1 \,, \quad (5.1)$$

where n is the number of documents in the corpus and $|\cdot|$ is the cardinality. Term frequencies are reweighted by their idf, so that frequent words, which occur in many documents (such as "results" or "brain"),

are given less importance. Indeed, such words are usually not very informative.

TFIDF features exploit a fixed vocabulary – each dimension is associated with a particular word. We denote d the size of the vocabulary (i. e. the dimension of the TFIDF vector space), and $\mathbf{x} \in \mathbb{R}^d$ a vector of TFIDF features: it is the input on which we base our prediction.

5.2.2 Representing spatial distributions

In order to construct a predicted density q over the brain, we rely on a brain parcellation – a finite partition $\mathcal{R}$ of $\mathbb{R}^3$:

$$\mathcal{R} = \{\mathcal{R}_0\} \cup \{\mathcal{R}_k,\ k = 1 \dots m\} \quad . \tag{5.2}$$

This partition contains a background region, $\mathcal{R}_0$: it is the region of $\mathbb{R}^3$ that lies outside of the brain. Inside the brain, the partition can be either a regular grid of voxels or a brain atlas, which is a division into anatomical regions. These brain regions are considered to be homogeneous. The predicted density q is then piece-wise constant over each region of this brain parcellation, and equal to 0 outside of the brain (in $\mathcal{R}_0$). We will therefore write q as a linear combination of the indicator functions of the atlas regions (excluding the background). To do so, we define the normalized region indicator functions:

$$r_k = \frac{\mathbb{I}_k}{\|\mathbb{I}_k\|_1},\ k = 1 \dots m \quad , \tag{5.3}$$

and then q as a linear combination of these functions:

$$q = \sum_{k=1}^{m} \alpha_k r_k \quad . \tag{5.4}$$

Defining q as a function of the input TFIDF vector $\mathbf{x}$ then reduces to defining each weight $\{\alpha_k,\ k = 1 \dots m\}$ as a function of $\mathbf{x}$.

Using a brain parcellation is a form of regularization: constraining the prediction to be in the span of $\{r_k\}$ reduces the size of the model space. Fine partitions, such as atlases with many regions or voxel grids, yield models with more expressive power, but more likely to overfit. Choosing an atlas thus amounts to a bias-variance trade-off.

5.2.3 Linear model

We model the weight of each atlas region as a linear response to $\mathbf{x}$:

$$q(z) = \sum_{t=1}^{d} \sum_{k=1}^{m} \mathbf{x}_t \mathbf{B}_{t,k} r_k(z) \quad \forall z \in \mathbb{R}^3 \tag{5.5}$$

where $\mathbf{B} \in \mathbb{R}^{d \times m}$ are model parameters, which we will learn. Note that for some values of $\mathbf{x}$ and $\mathbf{B}$, this linear approximation can yield values for q which are not a valid pdf (for example if they do not sum to 1). Therefore, once the model is trained, when we make a prediction, we coerce it into a density in a post-processing step, by taking its positive part and normalizing it.

5.2.4 Simple baseline: label-constrained encoder.

A simple approach to mapping text onto brain maps is to look for atlas labels in the text, and ignore interactions between terms. To apply this rule-based approach, we choose one atlas, whose regions are labelled with meaningful names, such as "occipital cortex". Then, the probability assigned to a region is chosen to be proportional to the frequency of its label in the text. For example, if the atlas contains a region labelled "parietal lobe", the predicted density in this region is proportional to the number of occurrences of "parietal lobe" in the text. In this case, the vocabulary is the set of atlas labels; the number of features d is equal to the number of regression targets m and the matrix of regression coefficients is square. The vocabulary can be ordered so that for every k, the word w_k is the label of the region $\mathcal{R}_k$, and in this case $\mathbf{B}$ is constrained to be diagonal. We call this model the *label-constrained encoder*, and we use it as one of our baselines.

5.3 FITTING THE MODEL

We fit the coefficients $\mathbf{B}$ of our model by empirical risk minimization. We define an error function, $\mathcal{E}(p, q)$, that measures the mismatch between a predicted density q and the true density p of a neuroimaging study's peak activations. Then, we set $\mathbf{B}$ to minimize the average of this error (augmented with a regularization term) over the example studies in our training dataset.

5.3.1 Estimation of the true density

We do not have access to the true pdf, p, from which the peak locations of each study in our training set were sampled. Therefore, we use an estimate of this density, which we denote $\hat{p}$. By construction of our prediction q, the best approximation of p that we can obtain belongs to the span of our brain regions $\{r_k\}$. Hence, we build our estimate $\hat{p}$ in this space: $\hat{p}$ is uniform over each voxel or atlas region. When the brain is divided into small voxels, there are few coordinates for a large number of regions. Therefore, we use Gaussian KDE (Scott, 2015; Silverman, 1986; Wand, 1994). Peak locations are binned to obtain voxel counts, which are then convolved with a Gaussian kernel. When

using an atlas with larger brain regions, we estimate the density using a histogram, with the atlas regions acting as the bins: we simply set the probability of a region to be proportional to the number of coordinates that it contains.

Whether we use KDE or a histogram, the estimated density $\hat{p}$ is a linear combination of $\{r_k\}$ and we denote $\{\hat{y}_k\}$ the coefficients:

$$\hat{p} = \sum_{k=1}^{m} \hat{y}_k r_k \ . \tag{5.6}$$

5.3.2 Loss functions

The next step is to define our data-fit term – the error $\mathcal{E}(\hat{p}, q)$ that we minimize over the training sample to set the model parameters by empirical risk minimization. We consider two common distances. The simplest choice is the ℓ_2 distance, $\|\hat{p} - q\|_2^2$, which is often used and has the appeal of being differentiable everywhere and yielding a very tractable optimization problem.

The second option we consider is a statistical distance , the Total Variation (TV). Remember that p is the density of a probability distribution $\mathcal{P}$. Denoting $\mathcal{U}$ the finite σ-algebra of all unions of elements of the partition $\mathcal{R} = \{\mathcal{R}_k, k = 0 \dots m\}$ of $\mathbb{R}^3$, $\mathcal{P}$ constitutes a measure on $\mathcal{U}$ characterized by:

$$\mathcal{P}(\mathcal{R}_k) = \int_{\mathcal{R}_k} p(z) dz \quad . \tag{5.7}$$

In the same way, q is the density of a probability distribution $\mathcal{Q}$. Then, the TV between $\mathcal{P}$ and $\mathcal{Q}$ is defined by:

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) = \sup_{\mathcal{A} \in \mathcal{U}} |\mathcal{P}(\mathcal{A}) - \mathcal{Q}(\mathcal{A})| \quad . \tag{5.8}$$

A classical result, stated for example in Gibbs and Su (2002), shows that this supremum – which in our case is a maximum, as $\mathcal{U}$ is finite – is attained by taking $\mathcal{A}$ to be the union of $\{\, \mathcal{R}_k \mid \mathcal{P}(\mathcal{R}_k) > \mathcal{Q}(\mathcal{R}_k) \,\}$ (or its complementary) and:

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} \sum_{k=1}^{m} |\mathcal{P}(\mathcal{R}_k) - \mathcal{Q}(\mathcal{R}_k)| \tag{5.9}$$

$$= \frac{1}{2} \int_{\mathbb{R}^3} |p(z) - q(z)| \, dz \quad . \tag{5.10}$$

The TV is half of the ℓ_1 distance between the pdfs. $\|\hat{p} - q\|_1$ is therefore a natural choice for our loss.

5.3.3 Decomposing the loss function

Let us call v_k the volume of the region r_k, i.e. the size of its support:

$$v_k \triangleq \|\mathbb{I}_k\|_1, \ k = 1 \dots m \quad . \tag{5.11}$$

Remember from Eq. (5.3) that $r_k = \frac{1}{\nu_k} \mathbb{I}_k$. Our loss can now be written as a sum over regions (see Appendix B.1 for details):

$$\mathcal{E}(\hat{p}, q) = \int_{\mathbb{R}^3} \delta(\hat{p}(z) - q(z)) dz \tag{5.12}$$

$$= \sum_{k=1}^{m} \nu_k \delta \left(\frac{\hat{y}_k}{\nu_k} - \frac{\sum_{t=1}^{d} x_t B_{t,k}}{\nu_k} \right) \tag{5.13}$$

Where δ is either the absolute value or the square function.

5.4 SOLVING THE ESTIMATION PROBLEMS

To set the model parameters $\mathbf{B}$, we use n example studies from our dataset $\{\mathcal{S}_i = (\mathcal{T}_i, \mathcal{L}_i),\ i = 1 \dots n\}$. We learn $\mathbf{B}$ by minimizing the empirical risk on $\{S_i\}$, augmented with an ℓ_2 penalty on $\mathbf{B}$. We add to the previous notations the index i of each example: p_i, q_i. $\hat{\mathbf{Y}} \in \mathbb{R}^{n \times m}$ is the matrix of dependent variables, whose i^{th} row is $\hat{\mathbf{Y}}_i \in \mathbb{R}^m$. $\hat{\mathbf{Y}}_i$ holds the estimated density in each brain region for study i. $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix, whose i^{th} row is $\mathbf{X}_i \in \mathbb{R}^d$. $\mathbf{X}_i$ holds the TFIDF features extracted from the text of study i.

5.4.1 Case $\delta = \ell_2$: ridge regression

When we choose the ℓ_2 loss, the empirical risk is

$$\sum_{i=1}^{n} \sum_{k=1}^{m} \left(\frac{\hat{\mathbf{Y}}_{i,k}}{\sqrt{\nu_k}} - \sum_{t=1}^{d} \frac{1}{\sqrt{\nu_k}} \mathbf{X}_{i,t} \mathbf{B}_{t,k} \right)^2 . \tag{5.14}$$

Defining $\tilde{\mathbf{Y}}_{:,k} = \frac{\hat{\mathbf{Y}}_{:,k}}{\sqrt{(\nu_k)}}$ and $\tilde{\mathbf{B}}_{:,k} = \frac{\mathbf{B}_{:,k}}{\sqrt{(\nu_k)}}$, and adding an ℓ_2 penalty, the problem becomes:

$$\underset{\tilde{\mathbf{B}}}{\operatorname{argmin}} \left(\|\tilde{\mathbf{Y}} - \mathbf{X}\tilde{\mathbf{B}}\|_F^2 + \lambda \|\tilde{\mathbf{B}}\|_F^2 \right) \tag{5.15}$$

where $\lambda \in \mathbb{R}_+$. This is the least-squares ridge regression predicting $\hat{p}$ expressed in the orthonormal basis of our search space $\{\frac{r_k}{\|r_k\|_2}\}$.

In order to set the hyperparameter λ, we use Generalized Cross-Validation (GCV) (Rifkin and Lippert, 2007). With this efficient scheme, a Singular Value Decomposition (SVD) of the design matrix is computed once. Then, for each hyperparameter, both the model coefficients and the Leave-One-Out (LOO) cross-validation errors can be computed very cheaply. With our large number of dependent variables, the computational cost is essentially that of computing the product $\mathbf{U}\,\mathbf{D}\,\mathbf{U}^T\,\mathbf{Y}$, where $\mathbf{U}$ are the left singular vectors of $\mathbf{X}$ and $\mathbf{D} \in \mathbb{R}^{d \times d}$ is diagonal, once for each hyperparameter on the regularization path – see Rifkin and Lippert (2007) for details.

5.4.2 Case $\delta = \ell_1$: LAD regression

When we choose the ℓ_1 loss, the empirical risk becomes

$$\sum_{i=1}^{n}\sum_{k=1}^{m}|\hat{\mathbf{Y}}_{i,k} - \sum_{t=1}^{d}\mathbf{X}_{i,t}\mathbf{B}_{t,k}| \tag{5.16}$$

This problem is known as Least Absolute Deviations (LAD) regression, a particular case of quantile regression (Chen and Wei, 2005; Koenker and Bassett Jr, 1978). Unlike ℓ_2 regression, which provides an estimate of the conditional mean of the target variable, ℓ_1 provides an estimate of the median. Quantile regression is more robust to outliers and better-suited than least-squares when the noise is not normally distributed (Koenker and Bassett Jr, 1978). Adding an ℓ_2 penalty, we have the minimization problem:

$$\hat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}}\left(\|\hat{\mathbf{Y}} - \mathbf{X}\,\mathbf{B}\|_1 + \lambda\,\|\mathbf{B}\|_F^2\right) \tag{5.17}$$

Unpenalized quantile regression is often written as a linear program and solved with the simplex algorithm (Koenker and d'Orey, 1994), Iteratively Reweighted Least Squares, or interior point methods (Portnoy and Koenker, 1997). Yi and Huang (2017) use a coordinate-descent to solve a differentiable approximation of the quantile loss (the Huber loss) with elastic-net penalty. Here, we minimize Eq. (5.17) via its dual formulation (c.f. supplementary material):

$$\hat{\boldsymbol{\nu}} = \underset{\boldsymbol{\nu}}{\operatorname{argmax}}\left(\operatorname{Tr}(\boldsymbol{\nu}^T\hat{\mathbf{Y}} - \frac{1}{4\lambda}\boldsymbol{\nu}^T\mathbf{X}\,\mathbf{X}^T\boldsymbol{\nu})\right) \quad \text{s.t.} \quad \|\boldsymbol{\nu}\|_\infty \leqslant 1, \tag{5.18}$$

where $\boldsymbol{\nu} \in \mathbb{R}^{n\times m}$ and Tr is the trace operator. The primal solution is given by $\hat{\mathbf{B}} = \frac{\mathbf{X}^T\hat{\boldsymbol{\nu}}}{2\lambda}$. As the dual loss is differentiable everywhere and the constraints are *bound* constraints, we can use an efficient quasi-Newton method such as L-BFGS-B (Byrd et al., 1995) to solve Eq. (5.18). The hyperparameter λ is set by cross-validation on the training set. We use warm-start on the regularization path (decreasing values for λ) to initialize each problem.

5.4.3 Fitting the label-constrained encoder

For the label-constrained encoder (Section 5.2.4), the coefficient matrix is constrained to be diagonal. Fitting this model amounts to fitting one univariate regression for each region – finding the coefficient that links the density in each region to the frequency of occurrence of its label. This can be done with either loss; since this model is meant to be a simple baseline, in our experiments we only use the least-squares version.

5.5 VALIDATION METRIC

Our supervised learning setting provides a natural way to quantify performance and conduct model selection: using cross-validation to measure the test error on left-out studies. While training models, we use an estimate $\hat{p}$ of the true spatial density p that we want to predict. However, to measure performance, we want to use only the actual observations: the peak coordinates extracted from neuroimaging studies. Indeed, the loss we compute during training is dependent on the density estimate, thus does not allow us to compare performance across density estimators (varying atlas choice or Gaussian kernel bandwidth). Moreover, density estimation is part of our modelling pipeline, and measuring performance on unprocessed observations is more meaningful. Therefore, the metric we choose is the average log-likelihood with respect to peak locations that were actually reported in test articles, i. e. the average of the log of q at reported locations. Since the log diverges in 0, before computing this metric, we make sure the predicted density is strictly positive everywhere in the brain. To do so, we define $\tilde{q}$, a mixture of a uniform density over the whole brain, and the normalized positive part of the regression model output:

$$\tilde{q} = (1 - \epsilon)q + \epsilon \sum_{k=1}^{m} \frac{\mathbb{I}_k}{v_k} , \tag{5.19}$$

and we use $\tilde{q}$ as our prediction. This amounts to considering that in the test coordinates, a proportion ϵ are noise – false positives in the studies or errors in the coordinate extraction process. The score for a study $S_i = (\mathcal{T}_i, \mathcal{L}_i)$, with locations $\mathcal{L}_i = \{l_{i,a}, a = 1, \dots c_i\}$ is then:

$$\frac{1}{c_i} \sum_{a=1}^{c_i} \log(\tilde{q}_i(l_{i,a})) \tag{5.20}$$

and we average this score over all the studies in the test set.

In Chapter 6, we use this metric to quantitatively compare models and brain parcellations, and we also present some qualitative results.

6 TEXT-TO-BRAIN EXPERIMENTS

In Chapter 5, we introduced a supervised learning framework to map text onto spatial densities over the brain. In this chapter, we apply this framework to the dataset described in Chapter 4. We measure quantitative performance and compare models in Section 6.2. In Section 6.3, we inspect the coefficients of a trained linear model. In Section 6.4, we show an example use case for such a tool: deriving the brain regions associated with neurological diseases, based on textual reports only.

6.1 EXPERIMENTAL SETTING

We rely on our dataset of over 13 000 studies and their peak activation coordinates. In Section 6.4, we also exploit our larger corpus of 130 000 studies for which we do not have peak activation coordinates (Chapter 4).

TFIDF representations depend on the choice of a vocabulary. The terms that carry the most valuable spatial information are those related to anatomy, which directly describe the locations of observations. To obtain good predictions, in this chapter, we restrict the vocabulary described in Section 4.5 to the subset that is related to anatomy, resulting in around 1000 neuroanatomical terms. Here, we are encoding full journal articles, which describe their observations in terms of anatomical regions. Restricting the vocabulary to these very informative features is therefore beneficial. However, for shorter documents, that may not use anatomical terms, such a small and specialized vocabulary could result in very sparse or empty TFIDF representations. Therefore, we will revisit this choice in Part III.

Our TFIDF representation for a study is the uniform average of the normalized TFIDF vectors for its title, abstract, full text, and keywords. Therefore, all parts of the article are taken into account, but a word that occurs in the title is more important than a word the article body (since the title is shorter).

Remember that we are training regression models to predict brain maps from TFIDF features. We have a training sample of n neuroimaging studies. Each study is associated with d TFIDF features, corresponding to a vocabulary of d words: $\mathbf{X} \in \mathbb{R}^{n \times d}$. We use these features to predict m target variables, that are the densities in each region in our partition of the brain: $\mathbf{Y} \in \mathbb{R}^{n \times m}$, where m is the number of regions. With our choice of corpus and vocabulary, the dimensions of the data used in this section are: the number of samples $n \approx 13\,000$, the

number of features $d \approx 1\,000$, and the number of outputs m depends on the brain parcellation used and ranges from 22 (Harvard-Oxford subcortical atlas) to over 28 000 (4 mm voxel grid).

When using Gaussian KDE, we use a diagonal bandwidth matrix, with diagonal (h^2, h^2, h^2), where $h = 1$. The corresponding Full Width at Half Maximum (FWHM) is given by:

$$\text{FWHM} = 2\,\delta\,h\,\sqrt{2\,\ln(2)}\,, \tag{6.1}$$

where $\delta = 4\,\text{mm}$ is the voxel size. With our bandwidth choice this results in a FWHM of 9.4 mm, which is in the range of kernel sizes typically employed to smooth fMRI data.

6.2 PREDICTION PERFORMANCE

This first experiment compares models and measures their prediction performance on left-out data. We perform 100 folds of shuffle-split cross-validation, keeping 10% (i. e. 1 300) of the studies in the test set. To assess the impact of the brain parcellation $\{\mathcal{R}_k\}$, we compare several anatomical atlases and a regular grid of 4-mm cubic voxels. We also compare several regression models. We include two naive baselines. The first is the *label-constrained encoder*. It is a simple rule-based approach that looks for the occurrence of the exact labels of a single atlas in the text (Section 5.2.4). It is included to check if such a discrete heuristic based on manually segmented and labelled brain regions performs better than a more data-driven approach. The second baseline is a sanity check: it simply predicts the average of the training set. It can be considered to measure chance level. Finally, we measure the performance of the linear model described in Chapter 5, using the least-squares and LAD losses.

The results are presented on Fig. 6.1. For all models, voxel-wise encoding performs better than any atlas. Large atlas regions regularize too much. Despite its higher dimensionality, voxel-wise encoding learns better representations of anatomical terms. The label-constrained model performs poorly, often below chance level. This seems to indicate that covering a large vocabulary and taking into account interactions between words is important. Moreover, the regions associated with anatomical terms are more accurately described with continuous densities than with discrete segmentations, and even atlases often disagree (Bohland et al., 2009). Therefore, *learning* the mappings from terms to regions is more reliable than relying on any particular atlas. Finally, for voxel-wise encoding, ℓ_1 regression outperforms ℓ_2. The best encoder is therefore learned using LAD regression and a voxel partition.

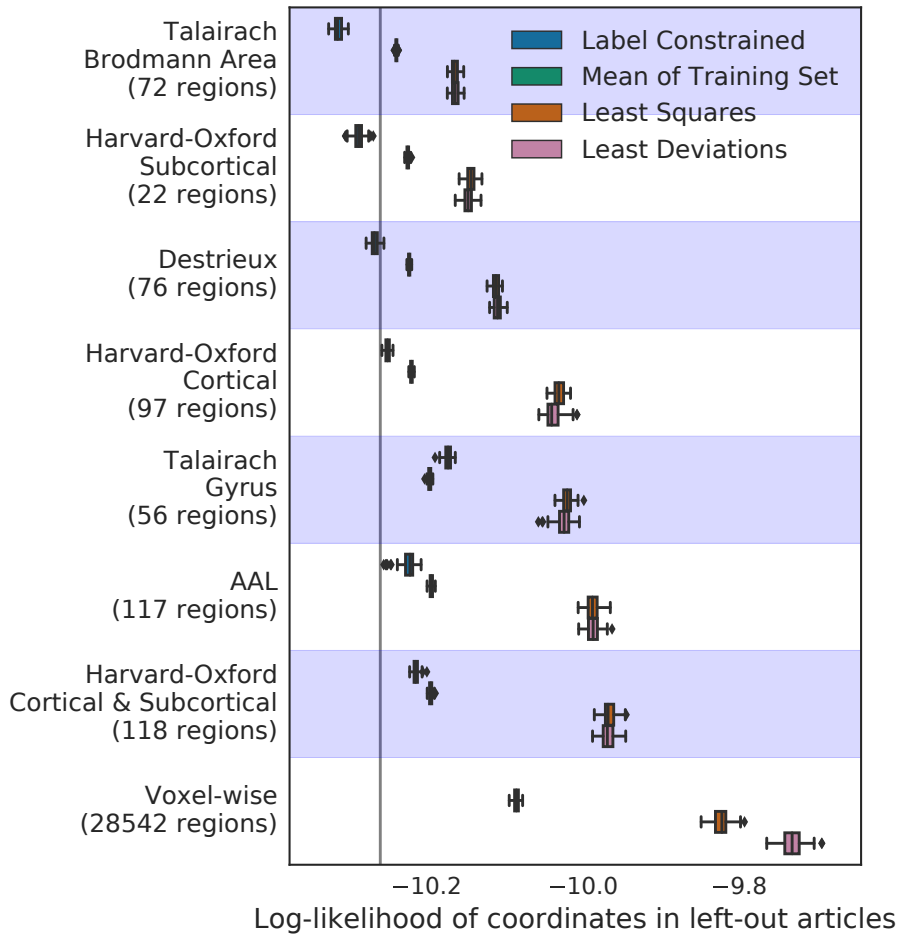


Figure 6.1: Log-Likelihood of coordinates reported by left-out articles in the predicted distribution . The vertical line represents the test log-likelihood for a uniform distribution over the brain. Voxel-wise encoding is better than relying on any atlas. In this setting, ℓ_1 regression significantly outperforms least-squares.

6.3 MODEL INSPECTION

The coefficients of the linear regression (rows of $\mathbf{B}$) are the brain maps that the model associates with each anatomical term. They are close to what neuroanatomists would expect (see for example Figs. 6.2 and 6.3).

6.4 META-ANALYSIS OF A TEXT-ONLY CORPUS

Once trained, the encoding models described in this chapter can extract spatial information from unannotated text. As an illustration, suppose that we want to use our corpus of 130 000 unlabelled articles (from which we could not extract peak locations) to estimate the distribution of observations associated with certain neurological diseases, such as aphasia. This type of meta-analysis should be feasible, since

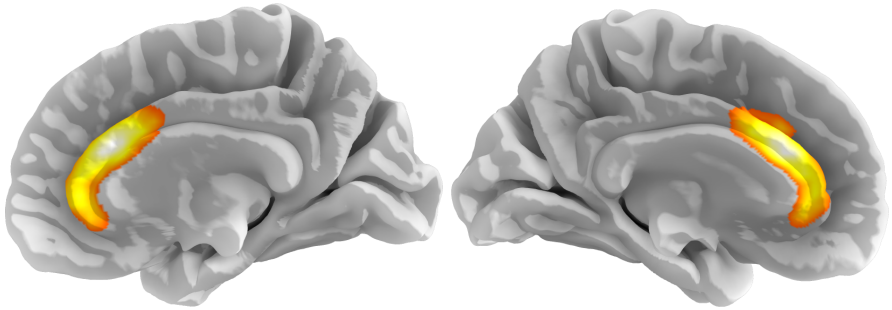


Figure 6.2: regression coefficient for "anterior cingulate"

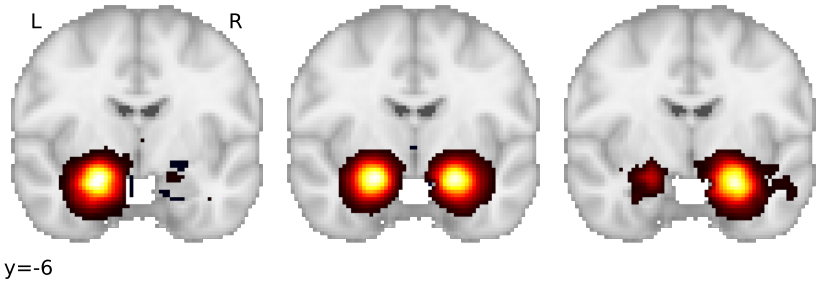


Figure 6.3: regression coefficients for "left amygdala", "amygdala", and "right amygdala"

in the corpus, several articles study aphasia and provide the names of anatomical structures that they examined, or in which they made observations (such as finding a lesion). However, this information is provided as free text, which we cannot use directly to compute statistics and identify locations in the brain.

6.4.1 Estimating densities for text-only documents

Trying to extract this spatial information with a rule-based approach and a predefined brain atlas is unreliable, as shown in Section 6.2. A better approach is to use a statistical encoding model that maps text to locations in the brain, trained on our dataset of 13 000 labelled articles. We therefore start by training our model, which learns over 1 000 mappings from terms to brain locations, such as the ones shown in Figs. 6.2 and 6.3. We denote $\hat{\mathbf{B}} \in \mathbb{R}^{d \times m}$ the coefficients of the trained model. We then use the trained model to transform into brain maps the articles of the unlabelled corpus – the corpus which provides no activation coordinates. Thus, the unstructured spatial information that these articles contain, which was out of reach before, is extracted in the form of quantitative images: spatial densities over the brain. Denoting N the number of documents in the text-only corpus and $\mathbf{C} \in \mathbb{R}^{N \times d}$ their TFIDF features, we obtain the matrix of predicted densities as $\mathbf{C}\hat{\mathbf{B}} \in \mathbb{R}^{N \times m}$.

6.4.2 Average density of relevant documents

We can then compute the average brain map for articles that mention "aphasia", weighted by the frequency of this term in their text. We denote $\mathbf{f} \in \mathbb{R}^N$ the TFIDF of the term of interest in each of the documents; we use these frequencies as a measure of relevance of each document to the query. The average map for relevant documents is computed as:

$$\bar{\mathbf{y}}_{\mathbf{f}} \triangleq \frac{1}{\|\mathbf{f}\|_1} \mathbf{f}^\top \mathbf{C}\, \hat{\mathbf{B}} \,. \tag{6.2}$$

We are interested in regions where the density of observations is higher in articles that are related to aphasia than in other articles. To identify these regions, we compute the ratio between the average density for documents related to aphasia and the average density across the whole corpus. In log space, this amounts to computing the difference:

$$\log(\bar{\mathbf{y}}_{\mathbf{f}}) - \log(\bar{\mathbf{y}}_{\mathbb{1}}) = \log(\frac{1}{\|\mathbf{f}\|_1} \mathbf{f}^\top \mathbf{C}\, \hat{\mathbf{B}}) - \log(\frac{1}{N} \mathbb{1}^\top \mathbf{C}\, \hat{\mathbf{B}}) \,. \tag{6.3}$$

The resulting contrast map for the example of "aphasia" is shown in Fig. 6.4, and correctly highlights regions affected by this disorder (Damasio, 1992). Brain maps obtained for other examples, Huntington's disease and Parkinson's disease, are shown respectively in Fig. 6.6 and Fig. 6.5.

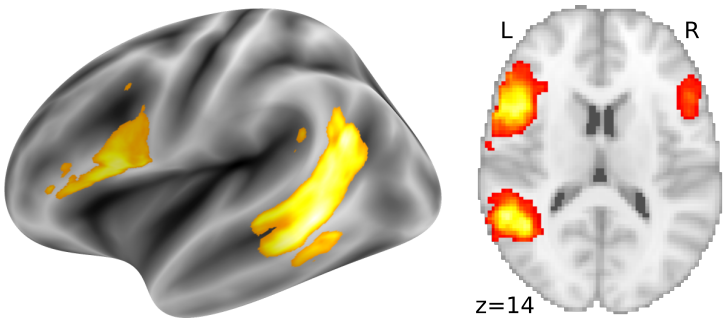


Figure 6.4: Map obtained for "aphasia", centered on Broca's and Wernicke's areas, in agreement with the literature (Damasio, 1992).

6.4.3 Transfer learning through term co-occurrences

In the experiment we described in this section, learning the mapping from text to distributions over the brain enables us extract spatial information from articles in the form of brain maps, and compute statistics on these structured representations. Contrasting the average density associated with a disease against the mean density for the whole corpus allows us to associate a meta-analytic brain map with

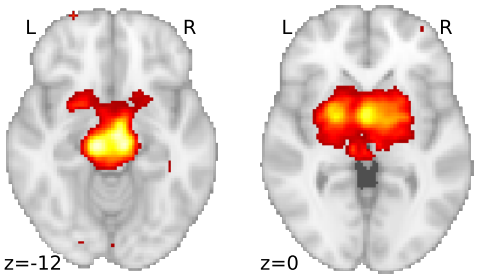


Figure 6.5: Map obtained for "Parkinson", highlighting the thalamus, basal ganglia, and other subcortical nuclei involved in Parkinson's disease (Davie, 2008).

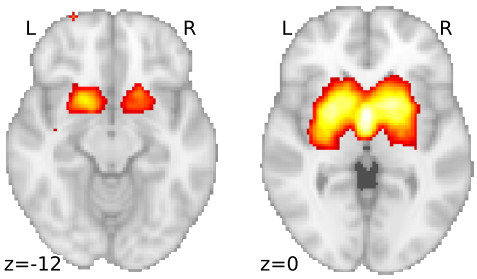


Figure 6.6: Map obtained for "Huntington", highlighting the striatum, caudate, putamen, basal ganglia, that are involved in Huntington's disease (Walker, 2007)

the disease, which recovers established domain knowledge (Damasio, 1992; Davie, 2008; Walker, 2007).

The corpus analyzed to study the disease only contains text. What enables us to recover maps for words that are not seen during the supervised training is their *co-occurrence*, in the unlabelled corpus, with anatomical terms whose spatial distribution is captured by the encoding model.

Beyond this simple example, this statistical framework for extracting spatial information from text could be applied to other datasets, such as collections of patient health records. Moreover, although we do not explore this further here, methods used for traditional meta-analysis could be applied to the predicted densities to form a null hypothesis and perform an actual statistical test.

6.5 CONCLUSION

We have introduced a statistical framework to translate textual description of studies into spatial distributions over the brain. This framework provides a natural metric to evaluate and compare models. Quantitative evaluation reveals that for this task, a voxel-wise discretization of the brain and Kernel Density Estimation are more appropriate than computing histograms based on larger atlas regions. Moreover,

high-dimensional term-frequency representations and multivariate models outperform a rule-based approach based on manually segmented atlases. When using KDE, penalized Least Absolute Deviations regression (a form of robust regression) yields a performance gain over ridge regression.

Learning to map text to brain locations enables using a corpus of pure text (containing no images or coordinates) to conduct meta-analyses in brain space. Applied to several diseases, this text-based meta-analysis synthesizes accurate brain maps that reflect the domain knowledge. This transfer from the labelled corpus to the unannotated text that describes a disease is due to the co-occurrence, in the unlabelled corpus, of the disease name with well-encoded anatomical terms. In Part III, we will further exploit this idea. Leveraging co-occurrence statistics will allow us to extend our encoding framework from mapping the text of neuroimaging articles to mapping arbitrary queries related to brain anatomy, function or diseases.

The work presented in Part II was published in Dockès et al. (2018).

Part III

MAPPING ARBITRARY QUERIES

7 THE NEUROQUERY MODEL

In Part I, we described standard meta-analysis methods, which find consistent associations between mental processes or disorders and brain structures across the neuroimaging literature. In Part II, we introduced text-to-brain, a predictive framework that maps neuroscience articles to distributions over the brain. Here, we improve the model described in Part II, and enable it to handle a larger vocabulary and very short documents, or even single terms. We create a general framework, called "NeuroQuery", that can encode into brain maps arbitrary queries, from single terms to full articles. This new model can handle a much larger vocabulary than previous approaches, including very rare terms that are challenging for text-to-brain and for which standard meta-analysis lacks statistical power. NeuroQuery can be used as an online tool: https://neuroquery.saclay.inria.fr.

7.1 EXTENDING THE SCOPE OF META-ANALYSIS

For many relevant queries – descriptions of mental functions, tasks, or diseases, neither standard meta-analysis methods such as ALE or NeuroSynth nor the text-to-brain model introduced in Part II can provide a satisfying brain map. Indeed, both approaches rely on a restricted vocabulary.

Text-to-brain depends on the terms that parametrize its TFIDF feature extraction. Documents that contain none of those terms cannot be encoded. If a document does not contain any term from text-to-brain's vocabulary, the TFIDF representation for this document is empty and the model cannot make a prediction. This is unlikely to happen with full articles, which contain many words, but can happen if we want a brain map for a short keyword search or for a single term. The TFIDF vocabulary can be extended, but each new term introduces a (potentially noisy) covariate when training the model. Hence, there is a trade-off between vocabulary size and prediction accuracy. Indeed, if the vocabulary is very large, inputs are very high-dimensional and contain many uninformative features, which hinders the estimation of model parameters.

Standard meta-analysis methods, such as NeuroSynth and ALE, are based on a different approach. These methods build a set of studies that investigate a mental process or a brain disorder. In the case of NeuroSynth, relevant studies are identified by matching a single term in their abstract. Standard methods then perform a statistical test to

identify regions where activations are consistently reported in the selected studies. They are therefore meant to produce statistical maps for well-identified mental processes that have been investigated in many neuroimaging studies. Standard meta-analysis is not meant to interpolate between studies, nor to make predictions about topics of interest that have seldom been investigated or require assembling and weighting results from very diverse studies. Moreover, the main fully automated framework, NeuroSynth, can only handle mental processes or diseases that can be completely described with a single phrase, such as "face recognition".

In this chapter we introduce NeuroQuery. It extends the class of queries for which we can obtain maps of the most relevant brain regions – regions that are likely to contain related neurological observations. We are still in the predictive setting of Part II – and not testing hypotheses as in standard meta-analysis. We start by presenting an example application which would be challenging for both text-to-brain and standard meta-analysis.

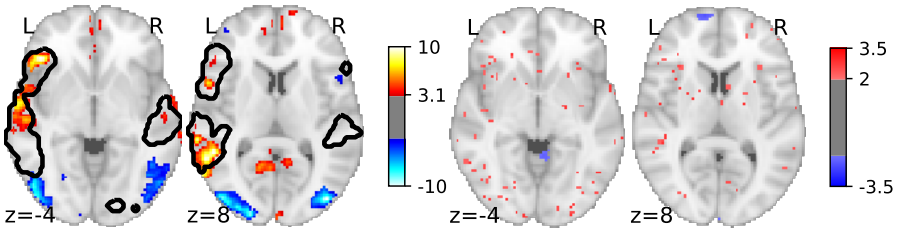
7.1.1 Motivating example

One of the simplest applications of meta-analysis is to put brain locations identified in a neuroimaging study in perspective with the literature. As an example, consider a study of the neural support of the translation of orthography to phonology, studied with visual stimuli by Pinho et al. (2018). Analyzing the results of this study builds upon an experimental contrast labeled by the authors as "Read pseudo-words vs. consonant strings", shown in Fig. 7.1. Given this description, what prior hypotheses arise from the literature for this contrast? Conversely, given the statistical map resulting from the experiment, how to compare it with previous reports on similar tasks? Meta-analysis seems the tool of choice.

Standard meta-analysis would struggle to provide an answer. Indeed, the current meta-analytic paradigm requires the practitioner to select a set of studies that are included in the meta-analysis, usually with equal weights. Existing methods provide no objective way to continuously weight studies according to their relevance to the topic of interest. In the case of "Read pseudo-words vs. consonant strings", which studies from the literature should be included? Even with the corpus of 14 000 full-text articles described in Chapter 4, only 29 studies contain all 5 words from the contrast description, which leads to a noisy and under-powered meta-analytic map (shown in Fig. 7.1b).

To avoid relying on the contrast name, which can be seen as too short and terse, it could be beneficial to do a meta-analysis based on the page-long task description[1]. This would require combining even more terms, which precludes selecting studies that contain all of

1 https://project.inria.fr/IBC/files/2019/03/documentation.pdf



(a) IBC map and NeuroQuery contour **(b)** ALE on contrast definition

Figure 7.1: Studying the contrast "Read pseudo words vs. consonant strings". (a): Group-level map from IBC for the contrast "Read pseudo-words vs. consonant strings" and contour of NeuroQuery map obtained from this query. The NeuroQuery map was obtained directly from the contrast description in the dataset's documentation, without needing to manually select studies for the meta-analysis nor convert this description to a string pattern usable by existing automatic meta-analysis tools. The contour corresponds to the arbitrary level 3.1. The map from which the contour is drawn, as well as a NeuroQuery map for the page-long description of the RSVP language task, are shown in Fig. 8.1. (b): GingerALE (Eickhoff et al., 2009) map for 29 studies that contain all terms from the IBC contrast description: too few studies can be automatically selected.

them. A more manual selection may help to identify relevant studies. But manual or semi-automated meta-analysis is more difficult and time-consuming. Moreover, some topics of interest may not have been investigated by themselves, or only in very few studies. For example, some rare diseases, or a task involving a combination of mental processes that have been studied separately, but never – or rarely – together. To compile a brain map from the literature for such queries, we need to interpolate between studies that are only partly related to the query. Standard meta-analytic methods lack an automatic way to measure the relevance of studies to a question, and to interpolate between them. This prevents them from answering new questions, or questions that cannot be formulated simply.

It would also be difficult to encode the contrast label using the text-to-brain model described in Part II. To make a prediction for "pseudo-words vs. consonant strings" with text-to-brain, we would need to include those terms in its vocabulary. Adding many terms to the vocabulary makes the input features more high-dimensional and could deteriorate the model's predictions, as mentioned in Section 7.1. Moreover, a query as short as a contrast label would result in an extremely sparse TFIDF vector, far from the distribution of the text-to-brain training documents. Therefore, there would be no guarantee of an accurate prediction. In this chapter, we will extend the text-to-brain model to address these difficulties.

7.1.2 Contributions of this chapter

The approach we present here, NeuroQuery, can assemble results from the literature into a brain map based on an arbitrary query. As text-to-brain (Chapter 5), NeuroQuery uses a multivariate model of the statistical link between multiple terms and corresponding brain locations. It is fitted on full-text publications. Thus, it learns to weight and combine terms to predict the brain locations most likely to be reported in a study. It can predict a brain map given any query related to neuroscience – not only single words, but also, for example, abstracts or full papers. With an extensive quantitative evaluation, we show in Section 8.5 that it retains text-to-brain's capacity to generalize to unseen studies, and approximates well the results of actual experimental data collection.

But unlike text-to-brain, NeuroQuery also models the semantic relations that underlie the vocabulary of neuroscience. Existing automatic meta-analysis methods such as NeuroSynth only include studies that explicitly reference a term chosen by the user. On the contrary, NeuroQuery leverages term co-occurrence statistics to model the semantic similarities that bind the terms used in the literature. Thus, it makes better use of the available information, and can recover biologically plausible brain maps where other methods lack statistical power. In particular, it can encode rare terms, as shown in Section 8.3 and Section 8.4. This semantic smoothing also makes it less sensitive to slight variations in terminology, as can be seen in Fig. 8.3. Interchangeable or very closely related terms yield similar brain maps through NeuroQuery, whereas the choice of an exact word form can change the results of traditional meta-analysis. This is central as there is no universally established vocabulary to describe mental processes and disorders. Indeed, the various controlled vocabularies built for neuroscience ontologies or meta-analysis are very weakly overlapping (Table 4.2). The similarities captured by NeuroQuery can also help researchers to navigate related neuroscience concepts while exploring their associations with brain activity.

In Part II, text-to-brain encoded documents through the anatomical terms that they contain. As the focus of most meta-analyses is brain function or diseases, here we use the whole vocabulary of over 7 000 terms collected in Chapter 4. To cope with the resulting sparse and high-dimensional inputs, we introduce a feature selection and adaptive regularization method, suited to regression problems with large output spaces. Moreover, while text-to-brain predicts densities over the brain, NeuroQuery can also estimate the variance of its predictions, and output brain maps of Z statistics. Although these brain maps remain predictions, and cannot be used for hypothesis testing, rescaling them in this way is more practical for some applications, such as defining a Region of Interest (ROI).

Beyond the scope of traditional meta-analysis, NeuroQuery extracts from the literature a coherent statistical summary of evidence accumulated by neuroimaging research. It can be used to explore the domain knowledge across sub-fields, generate new hypotheses, and construct quantitative priors or ROIs for future studies, or put in perspective results of an experiment. NeuroQuery is easily usable online[2], and can be openly downloaded for offline use[3]. We start by describing the statistical model behind NeuroQuery in this chapter. In Chapter 8, we present a qualitative and quantitative assessment of the new possibilities it offers.

7.2 OVERVIEW OF THE NEUROQUERY MODEL

NeuroQuery is an encoding model that maps text to a statistical map of associated brain regions. The main components of the model are an estimate of similarities between terms, derived from co-occurrence statistics, and a regression model that links term occurrences to neural activations. To generate a brain map, NeuroQuery first uses the estimated semantic associations to map the query onto a set of keywords that are well encoded in the brain. Then, it transforms the resulting representation into brain space using its regression model (Fig. 7.2). This model can thus be understood as a reduced rank regression, where the low-dimensional representation is a distribution of weights over keywords selected for their strong link with brain activity. In the rest of this chapter, we describe how we estimate semantic relations, select keywords, and map them onto brain activations.

7.3 REVISITING THE TEXT-ONLY META-ANALYSIS

Like text-to-brain, NeuroQuery represents text with TFIDF vectors. The first step of the NeuroQuery pipeline consists in *smoothing* these term-frequency representations: adding weight to terms that are related to the query but do not occur explicitly. Many expressions are challenging for existing meta-analysis frameworks, because they are too rare, polysemic, or have a low correlation with brain activity. Rare words are problematic because peak activation coordinates are a very weak signal: from each article we extract little information about the associated brain activity. Therefore existing frameworks typically need to see a term in many studies (hundreds in the case of NeuroSynth, see Section 8.4) in order to detect a pattern in peak activations. Term co-occurrences are more consistent and reliable, and capture semantic relationships (Turney and Pantel, 2010). It may require dozens or

2 https://neuroquery.saclay.inria.fr
3 https://github.com/neuroquery/neuroquery

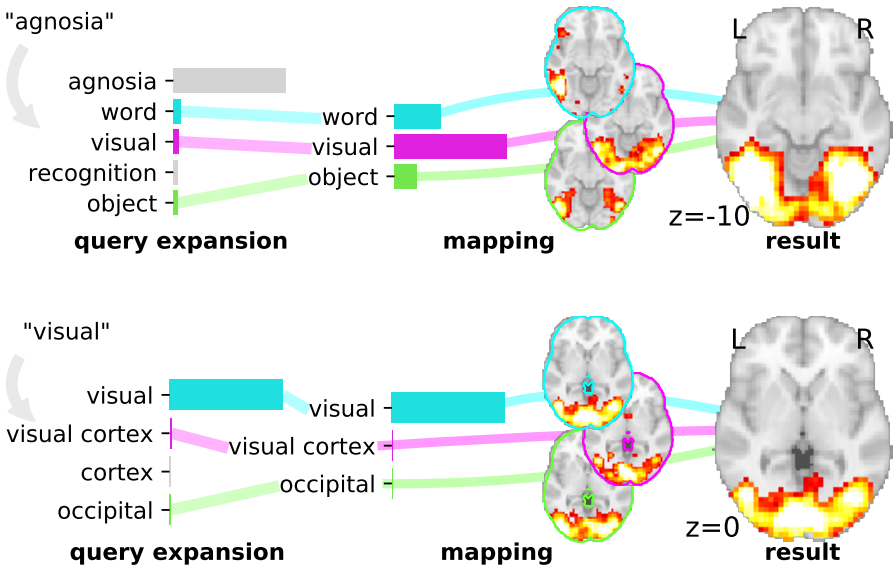


Figure 7.2: Overview of the NeuroQuery model: two examples of how association maps are constructed for the terms "agnosia" and "visual". The query is expanded by adding weights to related terms. The resulting vector is projected on the subspace spanned by the smaller vocabulary selected during supervised feature selection. Those well-encoded terms are shown in color. Finally, it is mapped onto brain space through the regression model. When the word, as "visual", has a strong association with brain activity and is selected as a regressor, the smoothing has limited effect.

hundreds of studies to detect a peak activation pattern for "aphasia", but with a few examples we notice that it often appears close to "language". By leveraging this information, NeuroQuery recovers maps for terms that are too rare to be mapped reliably.

We have already seen an example of exploiting term co-occurrences in Section 6.4, where we perform meta-analysis on a corpus of pure text by averaging the predicted densities for studies relevant to particular diseases. Here, we revisit this example and show that averaging predictions for relevant studies is equivalent to smoothing an input TFIDF vector by multiplying it with the covariance of TFIDF features across the corpus.

In Section 6.4, we started from a trained regression model and a matrix of TFIDF features for a corpus. The corpus contains n neuroimaging studies. The TFIDF matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ contains the term frequencies of d words in each of those studies[4]: $\mathbf{X}_{i,j}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant d$ is the TFIDF of term j in study i.

The regression coefficients are denoted $\hat{\mathbf{B}} \in \mathbb{R}^{d \times m}$, where m is the number of voxels in the brain. These coefficients result from

4 Regarding notation: in Section 6.4, we were dealing with two corpora. One was labelled with peak coordinates, whereas the second one contained only text. To distinguish them, we denoted $\mathbf{C}$ the TFIDF features for the unlabelled corpus. For simplicity, let us suppose here that there is only one corpus, with TFIDF features $\mathbf{X}$.

regressing the activity of each voxel in the brain on the TFIDF features: $\hat{\mathbf{B}}$ transforms a term-frequency vector $\mathbf{x}$ into a brain map $\hat{\mathbf{y}} = \hat{\mathbf{B}}^T \mathbf{x}$.

For a particular query term, such as "aphasia", we compute the average prediction, weighted by the relevance $\mathbf{f} \in \mathbb{R}^n_{\geqslant 0}$ of documents in the corpus:

$$\bar{\mathbf{y}}_{\mathbf{f}} \triangleq \frac{1}{\|\mathbf{f}\|_1} \mathbf{f}^T \mathbf{X} \hat{\mathbf{B}} \in \mathbb{R}^m \,. \tag{7.1}$$

In Eq. (7.1), $\mathbf{X}\hat{\mathbf{B}}$ computes the predictions for all the documents in the corpus, and taking the product by $\|\mathbf{f}\|_1^{-1}\mathbf{f}^T$ computes the average of these predictions, weighted by their relevance $\mathbf{f}$ to the term of interest.

If we extend the TFIDF vocabulary to include the term of interest, the relevance $\mathbf{f}_i$ of document i can be computed as the inner product of its TFIDF representation $\mathbf{x}_i$, and the basis vector $\mathbf{e}_j$, where j is the index of the term of interest in the vocabulary, and $\{e_j, j = 1 \dots d\}$ is the natural basis of the TFIDF feature space. More generally, for any (non-empty) query represented by the TFIDF vector $\mathbf{q}$, the inner product $\mathbf{x}_i^T \mathbf{q}$ is a simple relevance score for the document $\mathbf{x}_i$. The weighting vector can thus be chosen to be:

$$\mathbf{f}(\mathbf{q}) = \mathbf{X}\,\mathbf{q} \,. \tag{7.2}$$

Note that since $\mathbf{q}$ and $\mathbf{X}$ contain TFIDF features, they are non-negative, and $\|\mathbf{X}\,\mathbf{q}\|_1 > 0$ unless $\mathbf{q}$ is the empty query, which we do not consider. Then, the average density for the query $\mathbf{q}$ can be computed as:

$$\bar{\mathbf{y}}_{\mathbf{f}(\mathbf{q})} = \frac{1}{\|\mathbf{X}\,\mathbf{q}\|_1} \mathbf{q}^T \mathbf{X}^T \mathbf{X} \hat{\mathbf{B}} \,. \tag{7.3}$$

In Eq. (7.3), $\mathbf{q}^T\mathbf{X}^T$ measures the relevance of documents, and $\mathbf{X}\hat{\mathbf{B}}$ computes their associated brain map. However, we can also group factors differently to provide another interpretation of Eq. (7.3). Ignoring the multiplicative constant $\|\mathbf{X}\,\mathbf{q}\|_1^{-1}$, we notice that the TFIDF vector $\mathbf{q}$ representing the query is multiplied by a smoothing matrix $\mathbf{X}^T\mathbf{X}$ (the covariance of the corpus), then mapped onto brain space through the regression coefficients $\hat{\mathbf{B}}$. Computing the average prediction for documents similar to $\mathbf{q}$ thus amounts to smoothing $\mathbf{q}$ by the covariance of the TFIDF features.

If we do not want to include new terms in the features used to learn the supervised regression $\hat{\mathbf{B}}$, we can always project the smoothed representation back to the original, smaller vocabulary. In this case, we use two vocabularies. One large vocabulary, of size d, is used for smoothing, and includes the term(s) of interest – "aphasia" in this example. A smaller vocabulary of size $u < d$, included in the large one, is used for learning the supervised regression $\hat{\mathbf{B}} \in \mathbb{R}^{u \times m}$. In Section 6.4, this smaller vocabulary was restricted to anatomical terms. Denoting $\mathbf{P} \in \mathbb{R}^{u \times d}$ the projection matrix that selects the terms present in the small vocabulary, the brain map can be computed as

$$\bar{\mathbf{y}}_{\mathbf{f}(\mathbf{q})} = \frac{1}{\|\mathbf{X}\,\mathbf{q}\|_1} \mathbf{q}^T \mathbf{X}^T \mathbf{X} \mathbf{P}^T \hat{\mathbf{B}} \,. \tag{7.4}$$

This is the approach of NeuroQuery: *i)* extract TFIDF features corresponding to a large vocabulary, *ii)* smooth the extracted features – here the smoothed features are $\|\mathbf{X}\,\mathbf{q}\|_1^{-1}\,\mathbf{q}^\top\,\mathbf{X}^\top\,\mathbf{X}$, *iii)* project the resulting representation on a subset of the large vocabulary, chosen because it provides good features for predicting brain activity, and finally *iv)* map the projected TFIDF onto the brain through a trained regression model (Fig. 7.2). Based on the encouraging results obtained in Section 6.4, we will revisit each step of this procedure: smoothing (Section 7.4), selecting terms included as features for the supervised regression, and learning the regression coefficients (Section 7.5).

7.4 BUILDING THE SMOOTHING MATRIX

In order to smooth the sparse input features, we exploit the covariance of our training corpus. We rely on Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999). We use an NMF of $\mathbf{X} \in \mathbb{R}^{n\times d}$ to compute a low-rank approximation of the covariance $\mathbf{X}^\top\,\mathbf{X} \in \mathbb{R}^{d\times d}$. Thus, we obtain a denoised term co-occurrence matrix, which measures the strength of association between pairs of terms. We start by computing an approximate factorization of the corpus TFIDF matrix $\mathbf{X}$:

$$\mathbf{U}, \mathbf{V} = \underset{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}_{\geqslant 0}\\ \mathbf{V}\in\mathbb{R}^{k\times d}_{\geqslant 0}}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{U}\,\mathbf{V}\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \gamma(\|\mathbf{U}\|_1 + \|\mathbf{V}\|_1)\,, \quad (7.5)$$

where $k < d$ is a pre-defined constant (set to $k = 300$ in our experiments). Computing this factorization amounts to describing each document in the corpus as a linear mixture of k latent factors, or *topics*. In Natural Language Processing (NLP), similar decomposition methods are referred to as *topic modelling* (Blei, Ng, and Jordan, 2003; Deerwester et al., 1990). The hyperparameters $\lambda = 0.1$ and $\gamma = 0.01$ are set by evaluating the reconstruction error, sparsity of the similarity matrix, and extracted topics (rows of $\mathbf{V}$) on an unlabelled (separate) corpus. We find that the NeuroQuery model as a whole is not very sensitive to these hyperparameters and we obtain similar results for a range of different values.

Eq. (7.5) is a well-known problem. We solve it with a coordinate-descent algorithm described in Cichocki and Phan (2009) and implemented in `scikit-learn`[5] (Pedregosa et al., 2011). Then, let $\mathbf{N} \in \mathbb{R}^{k\times k}$ be the diagonal matrix containing the Euclidean norms of the columns of $\mathbf{U}$, i. e. such that $\mathbf{N}_{ii} = \|\mathbf{U}_{:,i}\|_2$ and let $\tilde{\mathbf{V}} = \mathbf{N}\,\mathbf{V}$. We define the

5 https://scikit-learn.org

word similarity matrix $\mathbf{A} = \tilde{\mathbf{V}}^T \tilde{\mathbf{V}} \in \mathbb{R}^{d \times d}$. This matrix is a denoised, low-rank approximation of the corpus covariance. Indeed,

$$\begin{aligned} \mathbf{X}^T \mathbf{X} &\approx (\mathbf{U}\mathbf{V})^T \mathbf{U}\mathbf{V} && (7.6)\\ &= \mathbf{V}^T \mathbf{N}^T \left(\mathbf{U}\mathbf{N}^{-1}\right)^T \mathbf{U}\mathbf{N}^{-1}\mathbf{N}\mathbf{V} && (7.7)\\ &\approx \tilde{\mathbf{V}}^T \tilde{\mathbf{V}}\,. && (7.8) \end{aligned}$$

The last approximation is justified by the fact that the columns of $\mathbf{U} \in \mathbb{R}^{n \times k}$ are almost orthogonal, and $\mathbf{U}^T\mathbf{U}$ is almost a diagonal matrix. This is what we observe in practice, and is due to the fact that $n \approx 13\,000$ is much larger than $k = 300$, and that to minimize the reconstruction error in Eq. (7.5) the columns of $\mathbf{U}$ have an incentive to span a large subspace of $\mathbb{R}^n$.

The similarity matrix $\mathbf{A}$ contains the inner products of the low-dimensional embeddings of the terms in our vocabulary. We form the matrix $\mathbf{T}$ by dividing the rows of $\mathbf{A}$ by their ℓ_1 norm:

$$\mathbf{T}_{i,j} = \frac{\mathbf{A}_{i,j}}{\|\mathbf{A}_{i,:}\|_1} \;\forall\, i = 1 \ldots d,\; j = 1 \ldots d\,. \tag{7.9}$$

This normalization ensures that terms that have many neighbors are not given more importance in the smoothed representation. The smoothing matrix that we use is then defined as:

$$\mathbf{S} = (1 - \alpha)\,\mathbf{I} + \alpha\,\mathbf{T}\,, \tag{7.10}$$

with $0 < \alpha < 1$ (in our experiments α is set to 0.1). This smoothing matrix is a mixture of the identity matrix and the term associations $\mathbf{T}$. Thus, the amount of smoothing is controlled and terms actually present in the query have a higher weight than terms introduced by the query expansion. This prevents degrading performance for documents which contain well-encoded terms, which obtain good prediction even without smoothing. This explains why in Fig. 7.2, the prediction for "visual" relies mostly on the regression coefficient for this exact term, whereas the prediction for "agnosia" relies on coefficients of terms that are *related* to "agnosia".

The smoothed representation for a query $\mathbf{q}$ becomes: $\mathbf{S}^T\mathbf{q}$, where $\mathbf{q} \in \mathbb{R}^d$ is the TFIDF representation of the query in large vocabulary space, and $\mathbf{S} \in \mathbb{R}^{d \times d}$ is the smoothing matrix.

7.5 MULTI-OUTPUT FEATURE SELECTION

When building an encoder for full articles in Part II, we used only anatomical terms to extract TFIDF features for documents. However, now that we are building a meta-analysis tool, we must be able to exploit a more diverse vocabulary, as most queries will focus on mental processes or diseases, and contain no anatomical terms. Hence, we

consider the whole vocabulary collected in Chapter 4, which contains over 7 000 terms related to anatomy, psychology and psychiatry.

As a result, TFIDF representations become more high-dimensional, and include some uninformative terms such as "magnetoencephalography". Such sparse, uninformative features can degrade the performance of the regression model. As considering only anatomical terms is no longer satisfactory, there is no easy way to decide which terms should be used as features for the supervised text-to-brain regression. We need a data-driven feature selection procedure. Moreover, among terms that we do want to keep in the vocabulary, some words display a much stronger correlation with brain activity than others. For example, "auditory" is well correlated with activations in the auditory areas, whereas "working memory" has a lower signal-to-noise ratio. Therefore, once we have selected the most informative terms, we must adapt the regularization applied to each feature, so that the most noisy coefficients get penalized more.

7.5.1 Identifying informative terms

Many existing methods for feature selection are not adapted to our case, because: *i)* the design matrix $\mathbf{X}$ is very sparse, and more importantly *ii)* we want to select the same features for $\approx 28\,000$ outputs (each voxel in the brain is a dependent variable). We therefore introduce a new reweighted ridge regression and feature selection procedure.

Our approach is based on the observation that when fitting a ridge regression with a uniform regularization, the most informative words are associated with large coefficients for many voxels. We start by fitting a ridge regression with uniform regularization. We obtain one statistical map of the brain for every feature (every term in the vocabulary). The maps are rescaled to reduce the importance of coefficients with a high variance. We then compute the squared ℓ_2 norms of these brain maps across voxels. These norms are a good proxy for the importance of each feature. Terms associated with large norms explain well the activity of many voxels and tend to be helpful features. We rely on these brain map norms to determine which features are selected and what regularization is applied. The feature selection and adaptive regularization are described in detail in the rest of this section.

7.5.2 Z scores for ridge regression coefficients

We keep the notations of Part II. Our design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ holds TFIDF features for d terms in n studies. There are m dependent variables, one for each voxel in the brain, which form $\mathbf{Y} \in \mathbb{R}^{n \times m}$.

The first ridge regression fit yields coefficients $\hat{\mathbf{B}}^{(0)}$:

$$\hat{\mathbf{B}}^{(0)} = \underset{\mathbf{B} \in \mathbb{R}^{d \times m}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\,\mathbf{B}\|_F^2 + \lambda \|\mathbf{B}\|_F^2, \tag{7.11}$$

where $\lambda \in \mathbb{R}_{>0}$ is a hyperparameter set with Generalized Cross-Validation (Rifkin and Lippert, 2007). We then compute an estimate of the variance of these coefficients. The approach is similar to the one presented in Gaonkar and Davatzikos (2012) for the case of SVMs. A simple estimator can be obtained by noting that the coefficients of a ridge regression are a linear function of the dependent variables. Indeed, solving Eq. (7.11) yields:

$$\hat{\mathbf{B}}^{(0)} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \,. \tag{7.12}$$

Defining

$$\mathbf{M} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T \,, \tag{7.13}$$

a simple estimate of the coefficients' variance is then:

$$\widehat{\mathrm{Var}}(\hat{\mathbf{B}}^{(0)}_{:,j}) = \mathbf{M}\,\widehat{\mathrm{Var}}(\mathbf{Y}_{:,j})\,\mathbf{M}^T \,, \tag{7.14}$$

where $\mathbf{Y}_{:,j}$ indicates the j^{th} column of matrix $\mathbf{Y}$. The squared residuals provide a plug-in estimate of $\mathrm{Var}(\mathbf{Y}_{:,j})$, considering that samples are i.i.d. and assuming that voxels are independent. We thus obtain a brain map of Z scores for each term in the vocabulary: for $i \in \{1, \ldots, d\}$,

$$\hat{\mathbf{z}}_i \triangleq \frac{\hat{\mathbf{B}}^{(0)}_{i,:}}{\hat{\boldsymbol{\sigma}}_i} \,, \tag{7.15}$$

where $\hat{\boldsymbol{\sigma}}_i \in \mathbb{R}^m$, $\hat{\sigma}_{ij}, 1 \leqslant j \leqslant m$ is an estimate of the standard deviation of $\hat{\mathbf{B}}^{(0)}_{ij}$, and division is performed element-wise.

7.5.3 Reweighted ridge matrix

Once we have a Z-map for each term, we summarize these maps by computing their squared Euclidean norm. In practice, we smooth the Z scores: $\hat{\mathbf{z}}_i$ in Eq. (7.15) is replaced by

$$\hat{\boldsymbol{\zeta}}_i = \frac{\hat{\mathbf{B}}^{(0)}_{i,:}}{\hat{\boldsymbol{\sigma}}_i + \delta} \,, \tag{7.16}$$

where δ is a constant offset. The offset δ allows us to interpolate between basing the regularization on the Z scores, or on the raw coefficients, i. e. the β-maps. We obtain better results with a large value for δ, such as the mean variance of all the regression coefficients. This prevents selecting terms only because they have a very small estimated variance in some voxels.

Next, we compute the mean μ and standard deviation s of $\{\|\hat{\boldsymbol{\zeta}}_i\|_2^2, i = 1 \ldots d\}$, and set an arbitrary cutoff

$$c = \mu + 2s \,. \tag{7.17}$$

All features i such that $\|\hat{\zeta}_i\|_2^2 \leqslant c + \epsilon$, where ϵ is a small margin to avoid division by zero in Eq. (7.19), are discarded. In practice we set ϵ to 0.001. The value of ϵ is not important, because features that are not discarded but have their ζ norm close to c get very heavily penalized in Eq. (7.19) and have coefficients very close to 0.

We denote $u < d$ the number of features that remain in the selected vocabulary. We denote $\phi : \{1 \dots u\} \to \{1 \dots d\}$ the strictly increasing mapping that reindexes the features by keeping only the u selected terms: $\phi(\{1 \dots u\})$ is the set of selected features. We denote $\mathbf{P} \in \mathbb{R}^{u \times d}$ the corresponding projection matrix:

$$\mathbf{P}^T_{:,i} = \boldsymbol{e}_{\phi(i)} \ \forall i \in \{1 \dots u\}\,, \tag{7.18}$$

where $\{\boldsymbol{e}_i, i = 1 \dots d\}$ is the natural basis of $\mathbb{R}^d$. The regularization for the selected features is then set to

$$w_i = \frac{1}{\|\hat{\zeta}_{\phi(i)}\|_2^2 - c}\,. \tag{7.19}$$

Finally, we define the diagonal matrix $\boldsymbol{W} \in \mathbb{R}^{u \times u}$ such that $\boldsymbol{W}_{ii} = w_i$ and fit a new set of coefficients $\hat{\mathbf{B}} \in \mathbb{R}^{u \times m}$ with this new ridge matrix.

7.5.4 Fitting the reweighted ridge regression

The reweighted ridge regression problem writes:

$$\hat{\mathbf{B}} = \underset{\mathbf{B} \in \mathbb{R}^{u \times m}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\,\mathbf{P}^T\,\mathbf{B}\|_F^2 + \gamma \operatorname{Tr}(\mathbf{B}^T\,\boldsymbol{W}\,\mathbf{B})\,, \tag{7.20}$$

Where $\gamma \in \mathbb{R}_{>0}$ is a new hyperparameter, that is again set by GCV. With a change of variables this becomes equivalent to solving the usual ridge regression problem:

$$\hat{\Gamma} = \underset{\Gamma}{\operatorname{argmin}} \|\mathbf{Y} - \tilde{\mathbf{X}}\,\Gamma\|_F^2 + \gamma\,\|\Gamma\|_F^2\,, \tag{7.21}$$

where $\tilde{\mathbf{X}} = \mathbf{X}\,\mathbf{P}^T\,\boldsymbol{W}^{-\frac{1}{2}}$ and we recover $\hat{\mathbf{B}}$ as $\hat{\mathbf{B}} = \boldsymbol{W}^{-\frac{1}{2}}\,\hat{\Gamma}$.

The variance of the parameters $\hat{\mathbf{B}}$ can be estimated as in Eq. (7.14) – without applying the smoothing of Eq. (7.16).

7.5.5 Summary of the regression with adaptive regularization

The whole procedure for feature selection and adaptive regularization is summarized in Algorithm 1.

In practice, the feature selection keeps $u \approx 200$ features. It has a very low computational cost compared to other feature selection schemes. The computational cost is that of fitting two ridge regressions (and the second one is fitted with a much smaller number of features). Moreover, the feature selection also reduces computation at prediction time, which is useful because we deploy an online tool based on the NeuroQuery model[6] (Section 8.7). Using a low-rank factorization of

6 https://neuroquery.saclay.inria.fr

Algorithm 1: Reweighted Ridge Regression

input: TFIDF features $\mathbf{X}$, brain activation densities $\mathbf{Y}$, regularization hyperparameters Λ, variance smoothing parameter δ

use GCV to compute the best hyperparameter $\lambda \in \Lambda$ and $\hat{\mathbf{B}}^{(0)} = \operatorname{argmin}_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_F^2$;

compute variance estimates $\hat{\sigma}_i^2$ as in Eq. (7.14);

$\hat{\zeta}_i \leftarrow \frac{\hat{\mathbf{B}}_i^{(0)}}{\hat{\sigma}_i + \delta}\ \forall i \in \{1 \dots d\}$;

compute c according to Eq. (7.17) ;

define ϕ the reindexing that selects features i such that $\|\hat{\zeta}_i\|_2^2 > c + \epsilon$;

define $\mathbf{P} \in \mathbb{R}^{u \times d}$ the projection matrix for ϕ as in Eq. (7.18) ;

$w_i \leftarrow \frac{1}{\|\hat{\zeta}_{\phi(i)}\|_2^2 - c}\ \forall i \in \{1 \dots u\}$;

$\boldsymbol{W} \leftarrow \operatorname{diag}(w_i,\ i = 1 \dots d)$;

use GCV to compute the best hyperparameter $\gamma \in \Lambda$ and $\hat{\mathbf{B}} = \operatorname{argmin}_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{P}^T\mathbf{B}\|_F^2 + \gamma \operatorname{Tr}(\mathbf{B}^T\boldsymbol{W}\mathbf{B})$;

return $\hat{\mathbf{B}}$, $\widehat{\operatorname{Var}}(\hat{\mathbf{B}})$, γ, $\mathbf{P}$, $\boldsymbol{W}$

the covariance of the corpus TFIDF (Section 7.4), rather than the full covariance matrix, also reduces the computational cost of making a prediction. The feature selection procedure, despite being heuristic, yields satisfying results on our dataset and problem (Chapter 8), and on toy synthetic data (Appendix C.1).

7.6 SMOOTHING, PROJECTION, MAPPING

To make a prediction, NeuroQuery combines the semantic smoothing described in Section 7.4 and the linear regression of brain activations described in Section 7.5. To encode a new document or query, the text is expanded, or smoothed, by adding weight to related terms using the semantic similarity matrix. The resulting smoothed representation is projected onto the reduced vocabulary of selected keywords, then mapped onto the brain through the linear regression coefficients (Fig. 7.2). A prediction thus writes:

$$\hat{\mathbf{y}} = \mathbf{x}^T\, \mathbf{S}\, \mathbf{P}^T\, \hat{\mathbf{B}} \tag{7.22}$$

$$= \mathbf{x}^T ((1 - \alpha)\, \mathbf{I} + \alpha\, \mathbf{T})\, \mathbf{P}^T\, \hat{\mathbf{B}}\ . \tag{7.23}$$

The rank of this linear model is the size u of the restricted vocabulary that was found to be reliably mapped to the brain. Compared with other latent factor models, this 2-layer linear model is easily interpretable, as each dimension (both of the input and the latent space) is associated with a term from our vocabulary.

In addition, NeuroQuery uses an estimate of the voxel-level variance of association, using the approach detailed in Section 7.5.2 and Eq. (7.14), and reports a map of Z statistics (The smoothing of Eq. (7.16) is only used for feature selection and is not applied when computing output maps). However, we should remember that these maps represent the prediction of a multivariate model for the smoothed representation of a query, which is treated as a fixed quantity. Therefore these maps must not be interpreted as direct associations between a term and brain activity, and are not used to test any meaningful null hypothesis. In particular, they have a different meaning than ALE, MKDA or NeuroSynth maps.

8 NEUROQUERY EXPERIMENTS

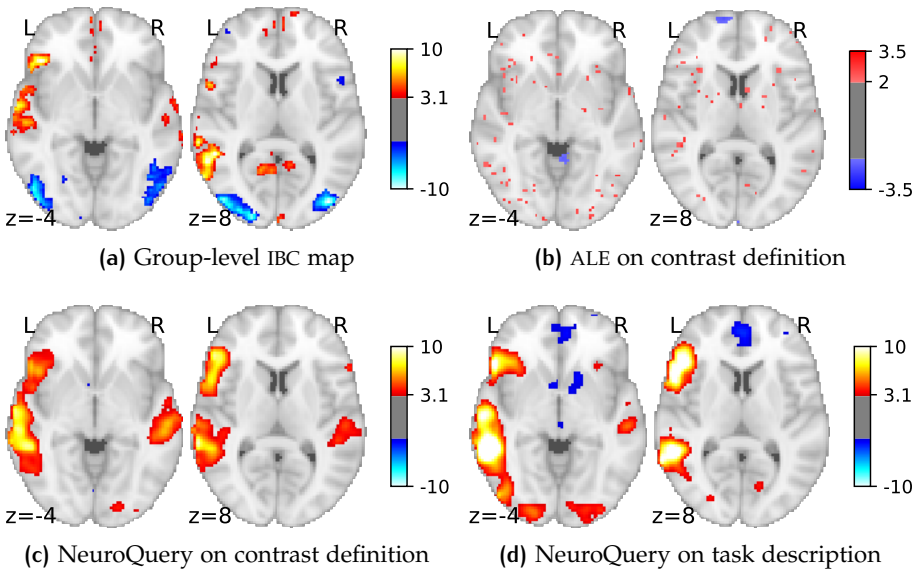
In Chapter 7, we described the NeuroQuery model and how it performs smoothing of sparse input features and adaptive regularization. Here we explore the possibilities offered by NeuroQuery. We start by presenting in more detail the motivating example based on the IBC fMRI dataset from Section 7.1. Then, we evaluate how NeuroQuery copes with queries that are challenging for other methods, including rare terms. Finally, we measure prediction performance with cross-validation, as was done for text-to-brain in Chapter 6.

8.1 ILLUSTRATION: INTERPRETING FMRI MAPS

After running an fMRI experiment, it is common to compare the computed contrasts to what is known from the existing literature, and even use prior knowledge to assess whether some activations are not specific to the targeted mental process, but due to experimental artifacts such as the stimulus modality. It is also possible to introduce prior knowledge earlier in the study and choose a Region of Interest (ROI) before running the experiment. This is usually done based on the expertise of the researcher, which is hard to formalize and reproduce. With NeuroQuery, it is easy to capture the domain knowledge and perform these comparisons or ROI selections in a principled way.

As an example, on Fig. 8.1, we take a contrast from the RSVP language task (Humphries et al., 2006; Pinho et al., 2018) in the IBC dataset. In the articles' methods, it is described as "Read pseudo-words vs. consonant strings". The corresponding group-level map from the IBC dataset is shown in Fig. 8.1a.

In theory, meta-analysis could tell us whether such a map is consistent with the literature. However, selecting relevant studies is difficult. Selecting all studies that contain all terms from the contrast name for an ALE meta-analysis lacks statistical power, as seen in Fig. 8.1b. We can obtain a brain map from NeuroQuery by simply transforming the contrast description, without any manual intervention. This map, shown on Fig. 8.1c, covers most of the activations from the group-level map: activations revealed by the fMRI study are consistent with what could be expected from the literature. Note that the NeuroQuery map is a prediction of regions where activations are likely to occur, and thus has a different interpretation than ALE, MKDA or NeuroSynth maps. We can also query NeuroQuery with the task description to obtain a map for the whole task, rather than a map that is specific to



(a) Group-level IBC map **(b)** ALE on contrast definition

(c) NeuroQuery on contrast definition **(d)** NeuroQuery on task description

Figure 8.1: Using meta-analysis to interpret fMRI maps Example of the "Read pseudo-words vs. consonant strings" contrast, derived from the RSVP language task in the IBC dataset. (a): the group-level map obtained from the actual fMRI data from IBC. (b): ALE map using the 29 studies in our corpus that contain all 5 terms from the contrast name – computed with GingerALE (Eickhoff et al., 2009). (c): NeuroQuery map obtained from the contrast name. (d): NeuroQuery map obtained from the page-long RSVP task description in the IBC dataset documentation: https://project.inria.fr/IBC/files/2019/03/documentation.pdf

a particular contrast. This map is shown in Fig. 8.1d, and highlights areas related to language, which is the focus of the RSVP language task (Pinho et al., 2018).

8.2 BASELINE METHODS

In the next three sections we compare the NeuroQuery model with existing automated meta-analysis methods, investigate how it handles terms that are challenging for the current state of the art, and quantitatively evaluate its predictive performance. We compare NeuroQuery with NeuroSynth (Yarkoni et al., 2011), the most well-known open meta-analytic tool, and with Generalized Correspondence Latent Dirichlet Allocation (GCLDA) (Rubin et al., 2017).

NeuroSynth computes meta-analytic map for single terms by performing a test of independence between the occurrence of a given term in the abstract of a study, and the fact that the study reports an activation within 10 mm of a given voxel. It is described in more

detail in Section 3.2.2.2 and Section 4.1. In our experiments, we use the authors' implementation, trained on their dataset[1].

GCLDA is an extension of Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). It models each study as a mixture of latent variables, called "topics". Each topic is associated with *i)* a distribution of weights over words, as in the original LDA model, and *ii)* a mixture of 3-dimensional Gaussians, which describes a density over the brain. In the author's experiments, each topic is associated with two Gaussians, whose means can be constrained to be symmetric across the sagittal plane. GCLDA is an important baseline because it is the only multivariate meta-analytic model that we are aware of. However, the maps it produces have a low spatial resolution because it models brain activations as a mixture of a small number of Gaussians. Moreover, it takes several days to train and a dozen of seconds to produce a map at test time, and is thus unsuitable for building an online and responsive tool like NeuroSynth or NeuroQuery. In our experiments, we use the implementation of GCLDA available at https://github.com/tsalo/gclda.

8.3 MAPPING CHALLENGING QUERIES

By combining term similarities and an additive encoding model, NeuroQuery can accurately map rare or difficult terms for which standard meta-analysis lacks statistical power. A few examples are shown in Fig. 8.2.

Moreover, capturing the relationships between terms results in consistent meta-analytic maps for similar terms. For instance, "calculation", "computation", "arithmetic", and "addition" are all related terms that are associated with similar maps by NeuroQuery, as seen in Fig. 8.3. On the contrary, NeuroSynth studies these terms in isolation, and thus suffers from a lack of statistical power and produces empty, or nearly empty, maps for some of these terms.

Finally, we can note that although the reweighted ridge regression procedure selects relatively few features, the neural support of these ≈ 200 terms covers most of the brain (Appendix C.2).

8.4 MEASURING SAMPLE COMPLEXITY

To measure how well methods cope with rare terms, we conduct meta-analyses on subsampled corpora. We mask occurrences of some terms to make them artificially rare, and use the maps obtained from the full corpus as a reference. We choose a set of relatively frequent and well-mapped terms, such as "language", for which NeuroQuery and

1 https://github.com/neurosynth/neurosynth

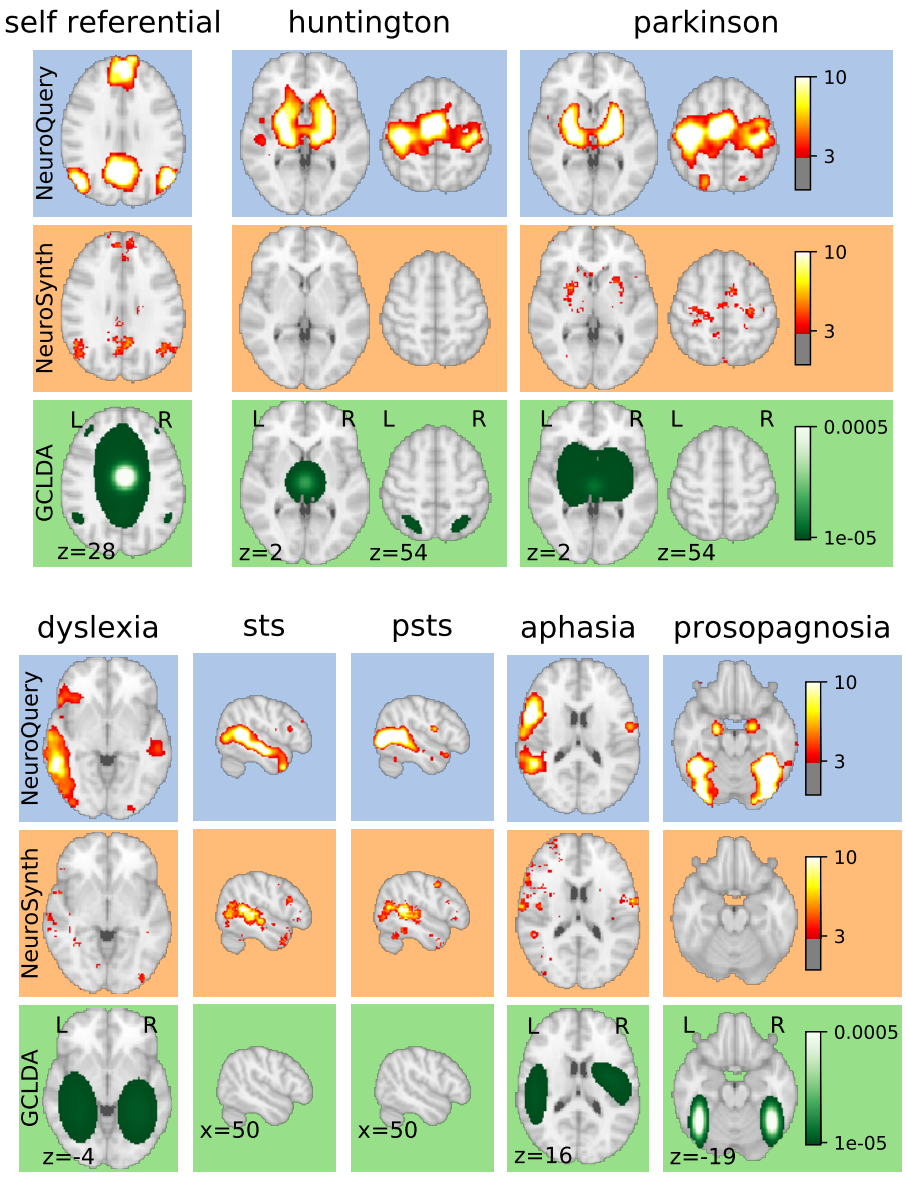


Figure 8.2: Example of maps obtained for a given term, compared across different large-scale meta-analysis frameworks. "GCLDA" has low spatial resolution and produces inaccurate maps. For relatively straightforward terms like "psts" (posterior superior temporal sulcus), NeuroSynth and NeuroQuery give consistent results. For terms that are more rare or difficult to map like "dyslexia", only NeuroQuery generates usable brain maps.

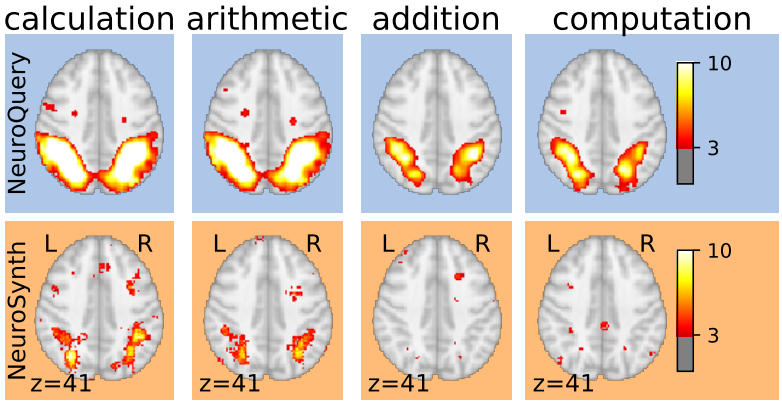


Figure 8.3: Maps obtained for a few words related to mental arithmetic. By correctly capturing the fact that these words are related, NeuroQuery can use its map for easier words like "calculation" and "arithmetic" to encode terms like "computation" and "addition" that are difficult for meta-analysis.

NeuroSynth (trained on a full corpus) give consistent results. For each of those terms, we construct a series of corpora in which the word becomes more and more rare: starting from a full corpus, we erase randomly the word from many documents until it occurs at most in $2^{13} = 8912$ articles, then $2^{12} = 4096$, and so on.

For many terms, with dozens of examples or even less, NeuroQuery can produce maps that are qualitatively and quantitatively close to the maps it obtains for the full corpus (and to NeuroSynth's full-corpus maps). NeuroSynth typically needs hundreds of examples to obtain similar results, as seen in Figs. 8.4 and 8.5. Document frequencies roughly follow a power law (Piantadosi, 2014), meaning that most words are very rare. Despite discarding terms that occur in less than 6 articles, half the terms in our vocabulary occur in less than 76 articles out of 13 000 (see Fig. 8.6). Reducing the number of studies required to map well a term (a. k. a. sample complexity of the meta-analysis model) therefore greatly widens the vocabulary that can be studied by meta-analysis.

8.5 PREDICTION PERFORMANCE

Unlike standard meta-analysis methods, that compute in-sample statistics, NeuroQuery is a predictive model and can produce brain maps for out-of-sample neuroimaging studies. This enables us to quantitatively assess its generalization performance. Here we check that NeuroQuery captures reliable links from concepts to brain activity – associations that generalize to new, unseen neuroimaging studies. We do this with 16-fold shuffle-split cross-validation. After fitting a NeuroQuery model on 90% of the corpus, for each document in the left-out test set (around 1 300), we encode it, normalize the predicted

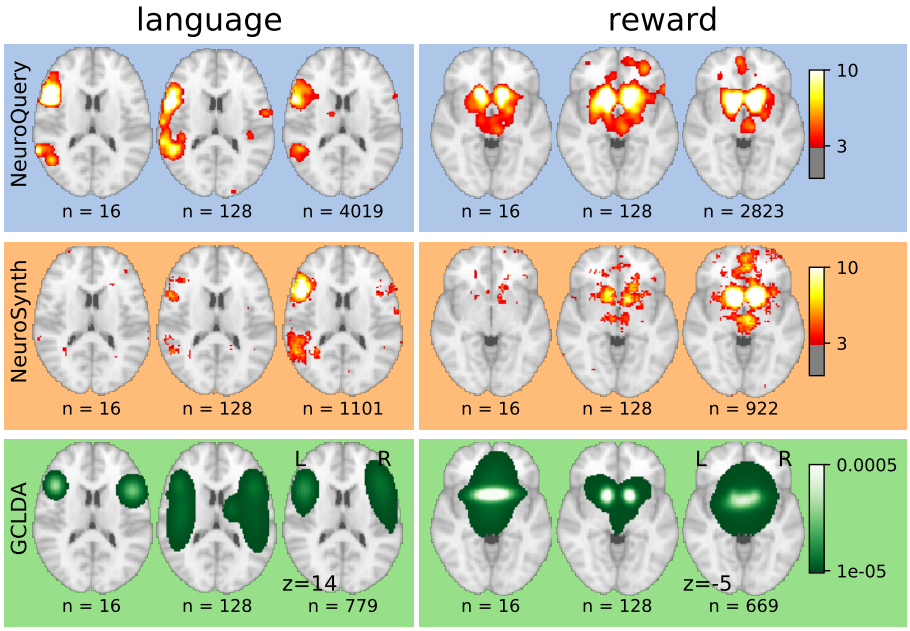


Figure 8.4: Learning good maps from few studies. Maps obtained from subsampled corpora, in which the encoded word appears in 16 and 128 documents, and from the full corpus. NeuroQuery needs less examples to learn a sensible brain map.

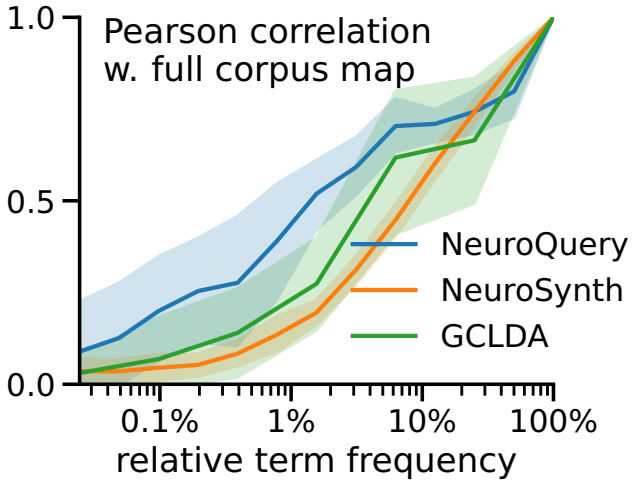


Figure 8.5: Convergence of maps toward their value for the full corpus, as the number of occurrences increases. Averaged over 13 words: "language", "auditory", "emotional", "hand", "face", "default mode", "putamen", "hippocampus", "reward", "spatial", "amygdala", "sentence", "memory".

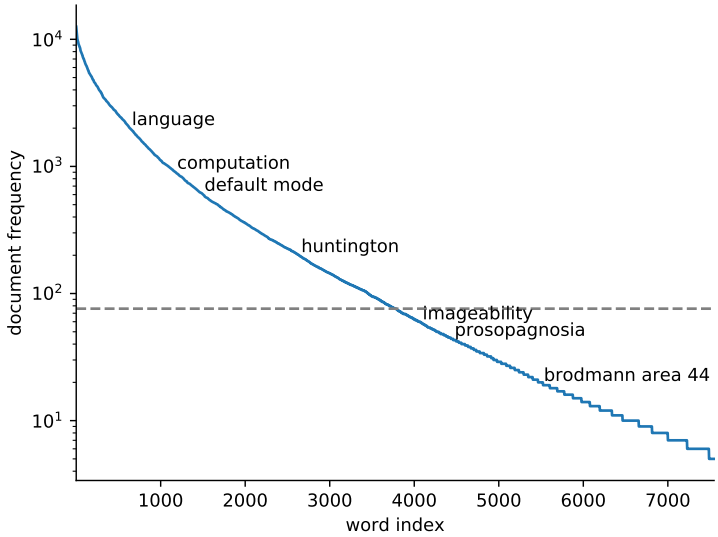


Figure 8.6: Most terms occur in few documents: we plot the document frequencies for terms in our vocabulary, sorted in decreasing order. While some terms are very frequent, occurring in over 12 000 articles, most are very rare: half occur in less than 76 (out of 14 000) articles.

brain map to coerce it into a probability density, and compute the average log-likelihood with respect to the coordinates reported in the article. The procedure is then repeated 16 times.

The results are presented in Fig. 8.7a. We also perform this procedure with NeuroSynth and GCLDA. NeuroSynth does not perform well for this task. Indeed, the NeuroSynth model is designed for single-phrase meta-analysis, and does not have a mechanism to combine words and encode a full document. Moreover, it is a tool for in-sample statistical inference, which is not well suited for out-of sample prediction. GCLDA performs significantly better than chance, but still worse than a simple ridge regression baseline. This can be explained by the unrealistic modelling of brain activations as a mixture of a few Gaussians, which results in low spatial resolution, and by the difficulty to perform posterior inference for GCLDA. To check that the good performance is not only due to our choice of likelihood model, we also use another metric, introduced in Mitchell et al., 2008. It tests the ability of the meta-analytic model to match the text of a left-out study with its brain map. For each article in the test set, we draw randomly another one and check whether the predicted map is closer to the correct map or to the random negative example. More than 72% of the time, NeuroQuery's output has a higher Pearson correlation with the correct map (Fig. 8.7b) than with the negative example.

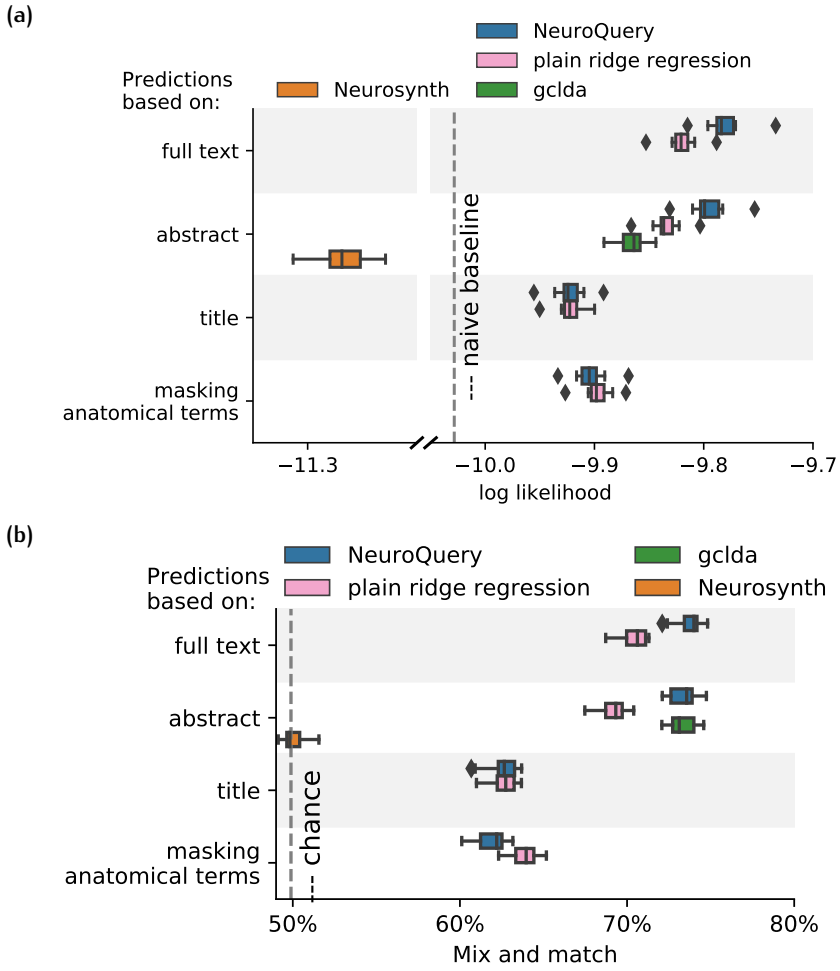


Figure 8.7: Explaining coordinates reported in unseen studies. (a): log-likelihood for coordinates reported in test articles, relative to log-likelihood of a naive baseline that predicts the average density of the training set. NeuroQuery outperforms GCLDA, NeuroSynth, and a ridge regression baseline. Note that NeuroSynth is not designed to make encoding predictions for full documents, which is why it does not perform well on this task. (b): "mix and match", or "leave-two-out" metric (Mitchell et al., 2008): how often is the predicted map closer to the true map than to a map randomly drawn from the test set? For this metric GCLDA's lack of spatial resolution is not detrimental: the prediction does not need to accurately reconstruct the true map, but only be closer to it than to other studies' maps.

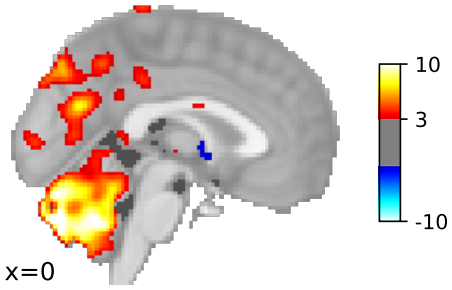


Figure 8.8: Example failure: prediction for "ADHD".

8.6 EXAMPLE FAILURE

In some rare cases, the use of semantic similarities can yield overconfident results. As NeuroQuery is explicit about how it builds a prediction, its failures are also easily detected. For example, failures can happen when a term has no closely related neighbors that correlate well with brain activity. "ADHD" is very similar to "attention deficit hyperactivity disorder", "hyperactivity", "inattention", but none of these terms are selected as features by the model because their link with brain activity is relatively loose. Hence, for "ADHD", the model builds its prediction on terms that are not very closely related and produces a map that highlights mostly the cerebellum, shown in Fig. 8.8. Although Attention Deficit Hyperactivity Disorder (ADHD) seems to be associated with differences in the volume and connectivity of the cerebellum, this map is incomplete and misleading (Konrad and Eickhoff, 2010). However, the failure is easy to detect for a user, based on the list of terms produced by NeuroQuery to support the prediction. In this particular case, the online tool also issues a warning stating that results may not be reliable. Being able to explore the internals of the model – the terms it uses to make a prediction and their individual coefficients, their weight, and the related studies and their coordinates, is therefore essential and limits the risk of interpreting NeuroQuery results incorrectly.

8.7 USING NEUROQUERY

NeuroQuery can easily be used online: `https://neuroquery.saclay.inria.fr`. Users can enter free text in a search box (rather than select a single term from a list as is the case with NeuroSynth) and discover which terms, neuroimaging publications, and brain regions are related to their query. NeuroQuery is also available as an open-source Python package that can be easily installed on all platforms: `https://github.com/neuroquery/neuroquery`. We hope this will allow advanced users to integrate NeuroQuery in other applications, extend it, and test its limitations. The package allows training new NeuroQuery models

as well as downloading and using a pre-trained model. Finally, the data used to train our NeuroQuery model is freely available: https://github.com/neuroquery/neuroquery_data. This repository contains the vocabulary list and document frequencies, TFIDF features for our whole corpus, and peak activation coordinates.

8.8 CONCLUSION: USABLE META-ANALYSIS TOOLS

NeuroQuery departs from existing meta-analytic frameworks by performing prediction rather than hypothesis testing, and by describing neuroscience topics of interest by continuous combinations of concepts rather than matching publications for exact terms. It is thus complementary of previously existing methods such as ALE, MKDA and NeuroSynth. As it combines multiple terms and interpolates between available studies, it predicts brain locations for new studies. It can predict accurate brain maps for terms that are too rare for standard meta-analysis – and especially existing automated tools such as NeuroSynth.

As it forms a comprehensive statistical summary of the literature, NeuroQuery has many potential applications in neuroscience research. *i)* In preparation of new functional neuroimaging studies, it helps to formulate hypotheses, to define formal priors, ROIs, or even to preregister experiments. *ii)* To analyze neuroimaging results, it enables comparing them with trends in the literature. Statistical maps directly derived from the description of the experiment ground objective and formal comparisons. *iii)* It can also help literature reviews by finding individual studies that are related to the query and their reported activations. *iv)* Finally, by navigating through the links to similar terms or studies, it can be used in an exploratory way, discovering relationships between neuroscience concepts as well as their neural correlates (Yeo et al., 2014). NeuroQuery can readily be used online, and one can also install it and download its data.

Compiling results across studies and laboratories is essential for the progress of human brain mapping (Yarkoni et al., 2010). Mental processes are difficult to isolate, and individual studies may report results that do not generalize. Thus, tools are needed to denoise and summarize knowledge accumulated across a large number of studies. Such tools must be usable in practice and match the needs of researchers who exploit them to study human brain function and disorders. NeuroSynth took a huge step in this direction by enabling anyone to perform, in a few seconds, a fully automated meta-analysis across thousands of studies, for an important number of isolated terms. Still, users are faced with the difficult task of mapping their question to a single term from the NeuroSynth vocabulary, which cannot always be done in a meaningful way. If the selected term is not popular enough,

the resulting map also risks being unusable for lack of statistical power. NeuroQuery provides brain maps for arbitrary queries –from seldom-studied terms to free-text descriptions of experimental protocols. Thus, it enables applying fully-automated and quantitative meta-analysis in situations where only semi-manual and subjective solutions were available. NeuroQuery therefore helps the feasibility of principled and quantitative accumulation of knowledge in neuroscience.

Part IV

DECODING BRAIN IMAGES

9 DECODING

In the next two chapters we present work in collaboration with Romuald Menuet. He is the first author of this study, wrote the implementation of the models described here, and performed the numerical experiments. I was involved in the project since the beginning and participated in design choices (from high-level decisions to implementation details) and writing the resulting publication.

In this part of the manuscript, we depart from previous chapters in two ways. First, we do not use reported peak activation coordinates anymore; we use full statistical maps of the brain. Instead of extracting the brain activation data from published articles, we download contrast maps for many studies from the NeuroVault[1] (Gorgolewski et al., 2015) online repository. Second, we switch from encoding to the reverse task: given a statistical map of the brain, which mental processes can we associate with it? This task is often called *decoding*. We are now mapping brain structures to behavioral annotations, whereas in the previous chapters we were mapping behavioral terms to brain structures. The NeuroVault repository contains dozens of thousands of statistical maps, and metadata that describes the experiments that produced these maps. We use this metadata to extract several labels for each brain image, and train statistical models to predict those labels, in a supervised learning setting. In this chapter, we formalize this decoding problem, present our approach, and describe the metrics that we use to evaluate methods. In Chapter 10, we detail how we obtain and process data from NeuroVault, describe our experiments, and present results.

9.1 PROBLEM SETTING

We train a model to predict which labels are relevant to describe a statistical map of the brain. We are in a multi-label classification setting. We have a set of n brain images, each represented by a vector in $\mathbb{R}^m$, where m is the number of voxels in the brain. These brain images are the inputs, from which we must predict cognitive labels. We have a fixed set of possible labels L that can be used to describe an image, such as "language processing" or "finger tapping". One or several labels from this predefined set are attached to each image. The labels for an image can be represented by their one-hot encoding,

1 https://neurovault.org

a vector $\mathbf{y} \in \{0,1\}^{|L|}$, such that $\mathbf{y}_l$ is 1 if label l is true, and 0 otherwise. These one-hot encodings thus form the classification targets $\mathbf{Y} \in \{0,1\}^{n \times |L|}$. We want to train a statistical model to predict which labels will be attached to a given image. To do so, we select an appropriate vector representation for input brain images (Section 9.2), and a model that links the occurrence of each label to these input vectors (Section 9.3.1). We set the model parameters with empirical risk minimization, which entails choosing a data-fit term and a regularization term (Sections 9.3.2 and 9.3.3).

9.2 REPRESENTING STATISTICAL BRAIN MAPS

fMRI statistical maps are very high-dimensional: there are typically over 200 000 voxels inside of the brain. Moreover, statistical maps contain some noise, and the location of activations can be influenced by registration and other parts of the analysis pipeline (Carp, 2012). As the number of available statistical maps is limited, it is therefore beneficial to reduce the dimension of input feature vectors before feeding them to a decoder. Many methods for feature selection and dimensionality reduction of fMRI statistical maps have been presented and compared in the literature.

One of the most successful approaches is to use an unsupervised decomposition method such as PCA, ICA or dictionary learning (Olshausen and Field, 1997) to learn a set of *spatial components*, which are brain maps that span a meaningful subspace of $\mathbb{R}^m$ (Thirion et al., 2014). Once such a basis is learned, fMRI maps can be projected on the spanned subspace and represented as a linear combination of these components. In extensive comparisons (Dadi et al., 2019; Mensch et al., 2017a), dictionary learning has obtained the best performances among other unsupervised decomposition methods, both in terms of reconstruction error and as a feature extraction method for supervised tasks such as decoding or predicting phenotypic variables. We therefore use dictionary learning to learn the basis in which we represent our brain maps.

The dictionary is learned from fMRI data. In order to learn such a decomposition, N fMRI maps, coming from different subjects, tasks and time points, are flattened into vectors and stacked. Thus, we form a matrix $\mathbf{S} \in \mathbb{R}^{N \times m}$, in which each row is a brain map, and each column corresponds to a voxel in the brain. Extracting the components then relies on factorizing this data matrix. Matrix Factorization has been extensively used in this way to learn unsupervised low-dimensional representations in diverse settings (Lee and Seung, 1999). We start by choosing a number of components $p \in \mathbb{N}$: the dimension of the basis that we will learn. We then decompose the stacked fMRI maps $\mathbf{S}$ as

$$\mathbf{S} \approx \mathbf{A}\,\mathbf{D}\,, \tag{9.1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times p}$ are the *loadings* of the N fMRI training maps, and $\mathbf{D} \in \mathbb{R}^{p \times m}$ are the p spatial components that form the *dictionary*. The loadings $\mathbf{A}$ and dictionary $\mathbf{D}$ are learned by minimizing the reconstruction error $\|\mathbf{S} - \mathbf{A}\,\mathbf{D}\|_F^2$, augmented with regularization terms, possibly under constraints. Formulations of dictionary learning differ in their choice of regularization terms (such as ℓ_1 or ℓ_2 penalties on $\mathbf{A}$ or $\mathbf{D}$) and constraints (such as constraining $\mathbf{A}$ or $\mathbf{D}$ to be non-negative, or to be in the ℓ_1 or ℓ_2 unit ball of $\mathbb{R}^m$).

In our experiments, we use a dictionary trained by solving the non-negative matrix factorization problem

$$\mathbf{A}, \mathbf{D} = \underset{\substack{\mathbf{D} \in \mathcal{C} \\ \mathbf{A} \in \mathbb{R}^{N \times p}}}{\operatorname{argmin}} \|\,\mathbf{S} - \mathbf{A}\,\mathbf{D}\,\|_F^2 + \lambda\,\|\,\mathbf{A}\,\|_F^2\,, \tag{9.2}$$

where the constraint $\mathcal{C} = \{\,\mathbf{D} \in \mathbb{R}^{p \times m},\ \|\mathbf{D}_j\|_1 \leqslant 1,\ \mathbf{D}_j \geqslant 0\,\}$ forces the dictionary components to be in the simplex of $\mathbb{R}^{p \times m}$, which results in sparse and positive components. The dictionary is trained on 27 task fMRI studies downloaded from OpenNeuro[2]. Learning decompositions from such a huge dataset was made possible by recent optimization methods for matrix factorization (Mensch et al., 2016, 2017b) and we use the implementation provided by the authors[3], with λ set to 0.001.

Once we are equipped with a dictionary $\mathbf{D} \in \mathbb{R}^{p \times m}$, we can obtain the low-dimensional representation of an fMRI map $\mathbf{s} \in \mathbb{R}^m$ by regressing it on the dictionary components with least-squares regression. We obtain its low-dimensional representation as:

$$\mathbf{x} = \underset{\mathbf{x} \in \mathbb{R}^p}{\operatorname{argmin}} \|\,\mathbf{s} - \mathbf{D}^T\,\mathbf{x}\,\|_2^2\,. \tag{9.3}$$

These *loadings*, or *embeddings*, $\mathbf{x} \in \mathbb{R}^p$ of the brain maps in our dataset are used as the input features of the decoding model. They form the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ of our decoding supervised classification task.

Projecting the original maps on a dictionary yields a dimension reduction from $\approx 10^5$ voxels to $\approx 10^3$ components. This projection reduces the variance of input features and estimated model parameters. It also makes training models much faster, enabling more extensive model selection and validation.

We use dictionaries of 3 different resolutions (128, 512 and 1024 components) to embed the original voxel activations in spaces of lower dimension. We also evaluate the performances when we use the stacking of those 3 embeddings as features to decode. For those 4 resolutions, we evaluate performance by keeping either all the component loadings or only their positive part – i. e. setting negative loadings to 0. Because of the positivity constraint on dictionary components in Eq. (9.2), non-negative loadings represent a non-negative brain map.

2 https://openneuro.org/

3 https://github.com/arthurmensch/modl

This can help decoding, as the negative part of statistical maps is often related to control conditions, which are seldom represented in our annotations (Section 10.1.2.1).

9.3 CLASSIFICATION MODEL

In this section, we describe the decoding model that we use to predict the one-hot encodings of cognitive labels, which form a matrix of targets $\mathbf{Y} \in \{0, 1\}^{n \times |L|}$, from the loadings of brain maps on dictionary components, which form a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$.

9.3.1 Model

Decoding commonly relies on high-dimensional multivariate linear models (Naselaris et al., 2011). Indeed, fMRI maps are noisy high-dimensional data and, aside from some rare large-scale studies such as Human Connectome Project (HCP), they involve few subjects and thus have small sample sizes (Bzdok and Yeo, 2017; Poldrack et al., 2017). Hence, non-linear models, that are more expressive, tend to overfit the noise in the data (Mensch et al., 2017a). As we are using a very large dataset that aggregates several studies (Section 10.1), and reduce the input dimension by projecting images on a dictionary, we can explore fully connected neural networks as decoding models. These models are particularly flexible in terms of regularization and architecture. We train them to jointly score all the possible labels. The input features of these networks are the compressed representations of brain images: the loadings of images on dictionary components. Such models are related to the factored logistic regressions of Bzdok et al. (2015), but differ in the use of an alternate basis for the initial projection and deeper models for the classification. As the brain maps are projected on dictionary learning components, and the supervised learner is a neural network, we call these models Neural Network over Dictionary (NNoD). We use up to 3 hidden layers. For the hidden units, we consider the identity activation function (resulting in a linear model) and the rectifier activation function $z \mapsto \max(z, 0)$. Output units are discussed in Section 9.3.2 and regularization in Section 9.3.3.

9.3.2 Data-fit term

The form of the loss function depends on how we model the occurrence of labels: whether we treat each label as a separate task, or want to predict a categorical distribution over all possible labels. In both cases, the corresponding model is trained with maximum likelihood. We denote f the function learned by the neural network up to the output

activation function. For example, if the simplest case where there are no hidden layers – i. e. logistic regression – f can be written:

$$f(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b}\,, \tag{9.4}$$

where $\mathbf{W} \in \mathbb{R}^{|L| \times p}$ and $\mathbf{b} \in \mathbb{R}^{|L|}$ are the coefficients of the model. Then, the model's prediction is:

$$\hat{\mathbf{y}} = g(f(\mathbf{x}))\,, \tag{9.5}$$

where g is the output activation function and $\mathbf{x} \in \mathbb{R}^p$ is the input vector of dictionary loadings.

9.3.2.1 *Multi-task setting*

We can consider labels separately, and treat the computation of a score for each label as a different classification task. In this case the output distribution for each label is a Bernoulli distribution and we use sigmoid output units. Then the prediction $\hat{\mathbf{y}} \in \mathbb{R}^{|L|}$ is such that:

$$\hat{\mathbf{y}}_l = \sigma(f(\mathbf{x})_l) \tag{9.6}$$

where σ is the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}\,. \tag{9.7}$$

The cost function in this setting is then:

$$\mathcal{L}_B(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{l \in L} \mathbf{y}_l \log(\hat{\mathbf{y}}_l) + (1 - \mathbf{y}_l) \log(1 - \hat{\mathbf{y}}_l)\,, \tag{9.8}$$

where B stands for "Bernoulli".

9.3.2.2 *Multi-class setting*

A slightly different approach is to introduce a tighter coupling between the labels and consider that they are, to some extent, exclusive. In this case, instead of producing one Bernoulli parameter for each label, the model produces the parameter of a Multinoulli distribution, or categorical distribution, over all the labels: a vector in the probability simplex of $\mathbb{R}^{|\mathcal{L}|}$, where $|\mathcal{L}|$ is the number of possible labels. In this case, a softmax function is applied to the output layer of the network. The prediction becomes:

$$\hat{\mathbf{y}} = \text{softmax}(f(\mathbf{x})) \tag{9.9}$$

where

$$\text{softmax}(\mathbf{z})_l = \frac{\exp(\mathbf{z}_l)}{\sum_l \exp(\mathbf{z}_l)}\,. \tag{9.10}$$

In this case the cost function becomes

$$\mathcal{L}_M(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_l \frac{\mathbf{y}_l}{\sum_l \mathbf{y}_l} \log(\hat{\mathbf{y}}_l)\,, \tag{9.11}$$

where M stands for "Multinoulli".

9.3.3 Regularization

We use an elastic-net penalty (Zou and Hastie, 2005) on the model coefficients. This regularization penalizes a combination of the square and absolute value of the parameters. Moreover, we apply dropout (Srivastava et al., 2014) on the input and hidden layers. We apply the same penalty and level of dropout on all layers.

9.3.4 Model selection

We select the amount of regularization (elastic-net penalty), amount of dropout, type of hidden unit and width of the hidden layers by cross-validation. To do so, we perform a randomized search for these hyperparameters on the training data. The 2-layer perceptron using the stacked dictionary loadings gives the best results. For this model, the input dictionary loadings have dimension $1024 + 512 + 128 = 1664$, the hidden layer has width 300, and a dropout of 0.2 is applied on the hidden layer (i. e. elements have a 0.2 probability of being set to 0). The same penalty is applied on ℓ_1 and ℓ_2 norms of coefficients and is set to 0.001. The activation function is the rectifier.

Results presented in Chapter 10 use a held-out test set (Section 10.1.3).

9.4 EVALUATION

We use unthresholded statistical brain maps from NeuroVault, labelled with terms from CognitiveAtlas. CognitiveAtlas (Poldrack and Yarkoni, 2016) is a large ontology of cognitive processes and tasks. We label each image with one or several CognitiveAtlas terms and learn to predict these labels. Details on NeuroVault and CognitiveAtlas are provided in Section 10.1. The number of true labels varies widely from one statistical map to another, depending on the way metadata fields were filled (Section 10.1.2). Hence, we prefer to use metrics that are not strongly influenced by the number of true labels. No decoding study has used the same set of labels as we do. Still, some works such as Varoquaux et al., 2018, Yarkoni et al., 2011 and Rubin et al., 2016, have targeted label sets that overlap with ours. To validate decoding performance, we need metrics that support comparison on the intersections of the sets of labels used by these different studies. Another important issue is that the frequency of labels varies greatly, with many labels used for a few images only, and a few very frequent labels such as "visual perception". We choose the metrics to ensure that they measure the performance of models for all the labels: capturing only the most frequent terms should not suffice to obtain a high score. Based on these criteria, we use two different metrics that are both

independent from label prevalence and are computed on a per-label (instead of per-sample) basis.

9.4.1 Pseudo Recall at k

For each label, we estimate the probability that, given a map for which this label is true, the decoder will rank it in the first k labels. Then, we average this score across labels to obtain our metric. This is slightly different from Recall at k, which estimates the expected number of true labels among the first k. We compute an average over all labels, rather than over all samples, because standard Recall at k would give too much importance to the few very frequent labels. Therefore, we call our metric *Pseudo Recall at* k. We denote $\mathbf{Y} \in \{0,1\}^{n \times |L|}$ the true labels and $\hat{\mathbf{Y}} \in \{0,1\}^{n \times |L|}$ the predicted labels for the n images in our dataset. We denote by $\hat{r}_i$ the ranking of the labels induced by the prediction $\hat{\mathbf{Y}}_i$ for sample i. It is a bijective function $\hat{r}_i : \{1, \dots |L|\} \to \{1, \dots |L|\}$ such that:

$$\hat{\mathbf{Y}}_{i,l} > \hat{\mathbf{Y}}_{i,l'} \implies \hat{r}_i(l) < \hat{r}_i(l') \,. \tag{9.12}$$

For example, if the label l obtained the highest score for sample i, $\hat{r}_i(l) = 1$. Then, we define the pseudo-Recall at k:

$$\text{Pseudo-Recall@k} = \frac{1}{|L|} \sum_{l \in L} \frac{|\{i, \hat{r}_i(l) \leqslant k, \mathbf{Y}_{i,l} = 1\}|}{|\{i, \mathbf{Y}_{i,l} = 1\}|} \,, \tag{9.13}$$

where $|\cdot|$ is the cardinality. For comparison, the Recall at k would be defined as:

$$\text{Recall@k} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\{l, \hat{r}_i(l) \leqslant k, \mathbf{Y}_{i,l} = 1\}|}{|\{l, \mathbf{Y}_{i,l} = 1\}|} \,. \tag{9.14}$$

Note that for simplicity we ignored the possibility of ties here; as $\hat{\mathbf{y}}$ are continuous values ties do not happen in practice. Pseudo Recall at k is easy to interpret and would make sense to evaluate a tool that suggests annotations for statistical maps. However, it does not enable comparing decoders that work with different sets of labels. Moreover, it only considers the induced ranking. Therefore, in order to obtain scores for individual labels, we also consider a widely used binary classification metric: the area under the Receiver Operating Characteristic (ROC) curve.

9.4.2 Receiver Operating Characteristic (ROC)

For a given label l, the Area under the ROC curve, which we will simply refer to as Area Under Curve (AUC) in the following, is defined as:

$$\underset{l}{\text{AUC}} = \mathbb{P}(\hat{\mathbf{Y}}_{i,l} > \hat{\mathbf{Y}}_{j,l} \mid \mathbf{Y}_{i,l} = 1,\ \mathbf{Y}_{j,l} = 0) \,. \tag{9.15}$$

We can obtain a global score by averaging this metric over all labels.

$$\text{AUC} = \frac{1}{|L|} \sum_{l \in L} \mathbb{P}(\hat{Y}_{i,l} > \hat{Y}_{j,l} \mid Y_{i,l} = 1,\ Y_{j,l} = 0)\,. \tag{9.16}$$

In a multi-label setting, averaging metrics such as the AUC over labels is sometimes computed with weights, for example proportional to the frequency of each label. Here, we want to weight all labels equally. In particular, we do not want to give more importance to frequent labels.

Having described our decoding models and appropriate metrics to evaluate them, in the next chapter we perform experiments on a large dataset of statistical maps annotated with cognitive labels.

10 DECODING EXPERIMENTS

In this chapter, we apply the decoding framework introduced in Chapter 9 to a large dataset that aggregates statistical maps from many neuroimaging studies, labelled with dozens of cognitive concepts. We start by describing our dataset in Section 10.1, and the dictionary on which we project fMRI maps in Section 10.2. We then present qualitative and quantitative results in Sections 10.3 and 10.4.

10.1 DATASET

Here we describe how we build a dataset to train supervised decoders that predict cognitive labels from unthresholded fMRI statistical maps (Chapter 9). We detail in Section 10.1.1 how we obtain statistical maps and metadata from the NeuroVault online repository (Gorgolewski et al., 2015). In Section 10.1.2, we explain how we use terms from the CognitiveAtlas ontology (Poldrack et al., 2011) to label brain images for supervised learning.

10.1.1 Neurovault

We collect statistical maps from the NeuroVault online platform[1] (Gorgolewski et al., 2015). NeuroVault is the largest existing repository of fMRI statistical maps and can provide material for Image-Based Meta-Analysis (IBMA). NeuroVault images can be explored and visualized via the website. The platform also exposes an API[2]. At the beginning of this thesis, we wrote a client for this API and integrated it in the Python library Nilearn[3] (Abraham et al., 2014). We use this client to download all the brain maps available on NeuroVault and their metadata. In January 2019 and after removing exact duplicates, we retrieved from NeuroVault over 60 000 brain images, coming from many different sources and adding up to more than 60 gigabytes of compressed data.

Images on NeuroVault are organized into *collections*. The data we downloaded is divided into 1 522 collections. Collections can regroup images for several tasks and usually from one task many contrasts can be derived. A collection can contain first-level contrasts (for individual

1 https://neurovault.org/
2 https://neurovault.org/api-docs
3 https://nilearn.github.io/

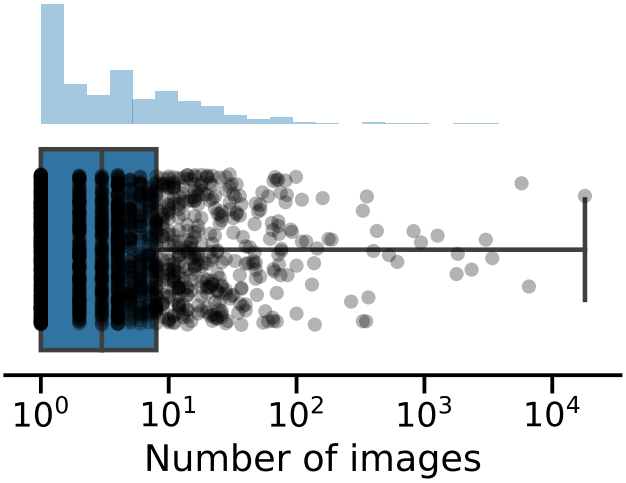


Figure 10.1: Distribution of neurovault collection sizes. Each dot represents a collection and shows the number of statistical maps it contains. Most collections contain only a few images, but some contain several thousands.

subjects), or second-level (group-level) contrasts (Section 2.4). Therefore, depending on the number of contrasts probed, on the nature of the statistical maps (subject-level or group-level), and the number of participants, the size of collections varies widely, ranging from 1 contrast to over 18 000. The distribution of collection sizes is long-tailed, with many small collections – the median size is 3 contrasts – and some very large collections (Fig. 10.1)

10.1.1.1 *Selecting NeuroVault maps*

NeuroVault aims to store "unthresholded statistical maps, parcellations, and atlases produced by MRI and PET studies"[4]. These images can correspond to human adults, children or other species. They can cover the whole brain or be restricted to a smaller ROI. Some collections also provide anatomical images. Moreover, in practice, many of the uploaded statistical maps are thresholded. For this study, we are only interested in whole-brain unthresholded statistical maps resulting from fMRI experiments on human adults.

We therefore select a subset of NeuroVault to train our models. To do so, we take into account the metadata as well as the image values. Relevant metadata fields for this purpose include `is_valid`, `is_thresholded`, `modality`, `not_mni` (indicating that the image is not registered to MNI space), `image_type` (indicating if the image is a statistical map or an atlas), `map_type` (partly redundant with `image_type`, can indicate the statistic such as Z or T, or take values such as "anatomical" or "ROI/mask"), and `subject_species`. Unfortunately, the metadata is not always accurate. For example, many statistical maps are annotated as unthresholded but have obviously been thresholded.

4 https://neurovault.org/

Errors can happen when uploading maps and many images do not contain sensible values. Therefore, after an initial filtering based on metadata, we also perform an automatic selection based on the image values. We exclude images that contain too many exact zeros, as they are either thresholded or the result of ROI analyses. We also exclude images that contain very large values, indicating that they are probably not Z nor T statistics. This selection yields around 54,000 maps, which represents a large fraction of the NeuroVault statistical maps.

10.1.2 Labelling statistical brain maps

Maps from NeuroVault come with metadata. Some fields in the metadata, such as `contrast_definition` or `task`, describe the experiment behind the image. Using NeuroVault metadata, we annotate our training images with one or more labels from the CognitiveAtlas ontology[5] (Poldrack et al., 2011). CognitiveAtlas is an online cognitive neuroscience knowledge base. It provides a description for more than 800 cognitive concepts and 700 experimental tasks, with some relationships between concepts as well as between concepts and tasks. CognitiveAtlas is one of the knowledge bases that were used to assemble the vocabulary described in Section 4.5 (see Table 4.2).

10.1.2.1 *Label extraction*

To label an image, we start by pruning the corresponding metadata. Some terms in the metadata are related to control conditions, with contrast definitions such as "reading vs. listening". Unfortunately, control conditions are not consistently represented in the metadata, and completely absent for many images. Hence, they cannot be extracted reliably. Therefore, we discard information related to control conditions – text that appears after patterns such as ">", "vs. ", "-", "versus", *etc.* in the `contrast_definition` or `name` metadata fields. This motivates considering only the positive part of the statistical maps (Section 9.2), since negative activations are often related to the control condition. Once control conditions have been removed, we extract all CognitiveAtlas concept labels from the remaining metadata. We discard extremely rare labels – those that occur in less than 10 images (out of over 29,000). Moreover, when the occurrences of two labels have a Pearson correlation above 0.95, we only keep the most frequent one.

10.1.2.2 *Limitations of exact label matching*

The labels that result from the simple extraction of exact matches described in Section 10.1.2.1 contain some errors.

5 http://www.cognitiveatlas.org/

STRING FORM VARIATIONS CognitiveAtlas does not distinguish entities or concepts from their labels. Terms in the CognitiveAtlas ontology are the full denominations of tasks or mental conditions, which can be quite long and do not always match the terminology that is used in practice. On the contrary, NeuroVault metadata is unconstrained and uncurated. The terminology used in this metadata is not always the same as in CognitiveAtlas. Moreover, most of the relevant metadata is entered by hand, and not validated. Hence, abbreviations and typos are frequent. For example, some collections use "right hand", "r. hand" or "RH" to indicate that subjects had to use their right hand (e. g. to press a button). The corresponding CognitiveAtlas term is "right hand response execution". Exact matching of CognitiveAtlas therefore limits the number of studies that we can exploit.

SPURIOUS LABELS The metadata fields in NeuroVault are not used consistently. For many images, the annotations include labels that describe the collection, or the overall task or experiment, rather than the particular contrast they are attached to. For example, 786 maps corresponding to the "shape recognition" condition display the label "emotion" in their annotations, because the corresponding study uses this condition as a baseline in an emotional task. This introduces false positives in the extracted labels.

LABEL STRUCTURE Many mental conditions are related. For example, performing "auditory sentence comprehension" implies performing "auditory perception" and "language comprehension". Some are also very similar, or at least used interchangeably in NeuroVault annotations. For example, "audition" and "auditory perception" are distinct CognitiveAtlas concepts (and not directly related in CognitiveAtlas), but unrestricted manual annotations on NeuroVault do not make such distinctions consistently and for our purposes they should be treated as synonyms. Ignoring these relations causes false negatives (missing labels) in our annotations.

10.1.2.3 *Improving cognitive labels*

To mitigate the issues described in Section 10.1.2.2, we design a set of heuristics to enrich and correct the raw annotations extracted from NeuroVault metadata.

DEDICATED LABEL EXTRACTORS FOR LARGE COLLECTIONS As shown in Fig. 10.1, the distribution of NeuroVault collection sizes has a long tail and contains some very large collections, such as HCP. Within a single collection, annotations are more consistent than across the whole NeuroVault repository. It is therefore worthwhile to examine the few very large collections and their use of metadata fields, in order to design dedicated rules to extract labels for their images.

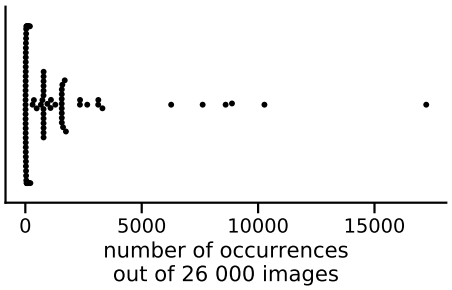


Figure 10.2: Number of images for which each label is used. Each dot is one of the 96 labels that we learn to decode and shows how many times it occurs in the dataset. Most labels occur in less 7 out of every 1 000 images: we must rely on few training examples despite having a large dataset.

Indeed, doing so can improve the training labels for thousands of statistical maps. We improve our training annotations by adding 123 rules to extract labels from the largest collections, fix some mistakes and handle some frequent synonyms.

MISSING LABEL INFERENCE An important difficulty is that many labels are very rare. Even among the 96 labels that we selected for training, most are used to tag only a few images (Fig. 10.2). However, we can slightly alleviate this scarcity by using the hierarchical structure of cognitive concepts to infer missing labels.

To exploit the structure of cognitive concepts, we assert a set of *related* and *broader* relations to infer some missing labels. For example, if an image is labelled with "speech production", the label "language" is added if it is missing. Or if an image is labelled with one of "audition" and "auditory perception", the other label is added. These relations form a directed graph, a small part of which is shown in Fig. 10.3.

We build this new ad-hoc graph rather than using CognitiveAtlas relations because CognitiveAtlas is too incomplete for missing label inference. Indeed, many concepts in CognitiveAtlas are disconnected from the graph, and some important relations are missing. For example, "auditory sentence comprehension" is not related to any other auditory concept. Moreover, the graph we construct aims to reflect the practical use of the most frequent terms in NeuroVault metadata, and capture which terms are used in similar contexts on this platform. This is far less ambitious than constructing an ontology that reflects the accumulated knowledge of cognitive science.

Using a hand-made discrete graph is a different approach than relying on continuous similarities, estimated from raw data, as was done in the context of encoding in Part III. In the decoding study presented here, we obtain better results (i. e. a higher decoding AUC on the IBC test data) with the rule-based approach described in this section. This can be explained by the fact that here we are in a more controlled setting, and decode a restricted set of labels. Moreover,

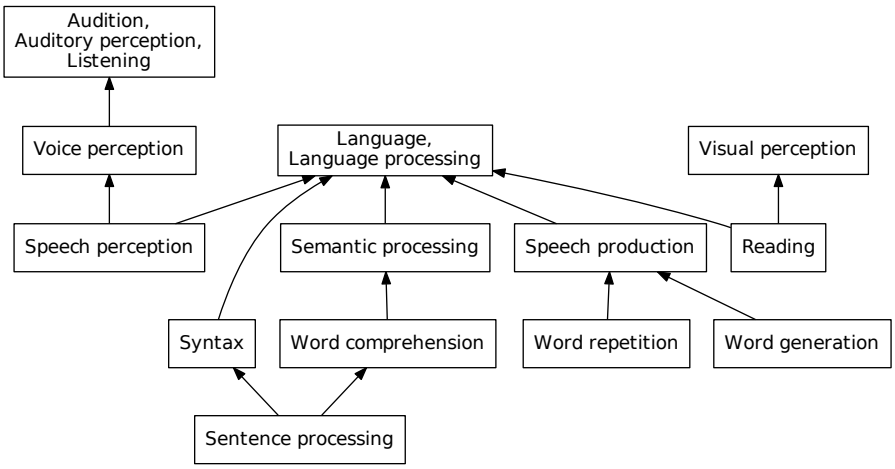


Figure 10.3: Missing labels inference graph. Here we illustrate a part of our label completion graph, for some labels related to language. The arrows assert implication. For example, "Reading" implies "Language" and "Visual perception". When several expressions appear in a box, separated by commas, they are treated as synonyms: for example "Language" and "Language processing" are merged.

obtaining appropriate continuous similarities for this study would be challenging, as NeuroVault metadata is of a very different nature than free text: there is a large distribution shift from data used to train unsupervised word representations to NeuroVault annotations.

10.1.3 Held-out test data

To measure prediction performance, we set aside a large and well-annotated collection: the Individual Brain Charting (IBC) dataset (Pinho et al., 2018). Using this collection as our test set ensures that there is no overlap between training and testing data. This would be more difficult to ensure if we performed cross-validation, as some collections or parts of collections can be duplicated on NeuroVault. Indeed, collections can be tied to publications that rely on common experiments, and provide duplicated or almost identical images. These duplicates or quasi-duplicates can be very difficult to detect. For example, two independent studies can perform the same analysis on the same data, obtaining similar but slightly different results, and upload the statistical maps in separate collections and with different image names. The IBC collection contains over 6 500 contrasts, which come from 12 individual subjects. As it is one of the largest collections and was uploaded very recently, there is little risk that it is duplicated on NeuroVault.

10.2 DICTIONARY SELECTION

For our low-dimensional representations of brain maps, we consider dictionaries of different resolutions (see Section 9.2): 128, 512 or 1024 components. We also consider a multi-resolution representation obtained by stacking the loadings over the components of these three dictionaries. Contours of the dictionary components are shown in Fig. 10.4. For simplicity, we choose the representation that yields the

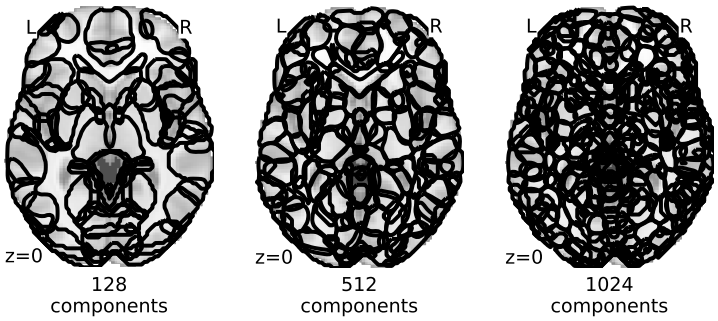


Figure 10.4: Dictionary components: outlines of the thresholded components of the three dictionaries we use. Input maps are regressed on these stacked components to obtain the dictionary *loadings*, which are the input features of the neural network.

best average AUC on the IBC dataset – at the risk of overfitting the dictionary choice. The multi-resolution representation yields the best predictions. Moreover, using only the positive part of statistical maps yields a performance gain. This may be due to the fact that we ignore labels related to control conditions or baselines (Section 10.1.2.1).

10.3 INSPECTING TRAINED MODELS

As we use non-linear models (Section 9.3.1), the model coefficients do not provide a direct mapping from brain images to cognitive labels. Still, we can perform a sensitivity analysis on our decoding model, by automatic differentiation of the model's output over its inputs. For each label, we compute the gradient with respect to the input features (dictionary component loadings), averaged over images in the training set. The sensitivity analysis reveals which brain regions are characteristic of a particular cognitive label. Some examples are shown in Figs. 10.6 and D.1.

We can see that the decoding model captures biologically plausible links between mental processes and brain activity. However, it also captures some biases of the neuroimaging data accumulated on NeuroVault, related to common practices, choices of stimuli and response media, and choices of control conditions. Indeed, due to some repeated standard experimental protocols designed to probe

specific cognitive processes, the model sometimes learns to recognize those protocols instead of more fundamental associations between the targeted processes and structures in the brain. For example, in the data we collected from NeuroVault "emotion perception" is usually probed by showing pictures of faces expressing different emotions. The model therefore learns to use activations in areas related to face perception, such as the FFA, to predict the label "emotion" (Fig. 10.5). Experimental protocols and stimulus modalities can therefore be important confounders, despite the large number of independent studies in our training dataset. The correlation of "emotion perception" with activity in the FFA would not generalize to a new study that would use a different stimulus, such as spoken descriptions of emotions.

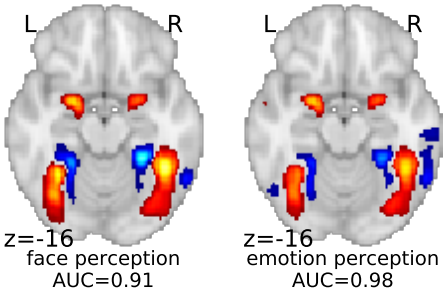


Figure 10.5: Face and emotion perception. The decoder correctly captures areas related to face perception, such as the FFA, to predict the occurrence of the label "face perception". However, we see that its decoding map for "emotion perception" is almost identical, and shows areas related to face perception rather than emotion perception. This is due to the fact that in our training dataset, visual stimuli are used to probe emotion perception, and maps that are labelled with "emotion perception" often contain activations related to face perception. Therefore, we must be careful when interpreting the decoding maps, as they capture the biases of the training data. Of course, this is also true of encoding or any other analysis method.

10.4 PREDICTION PERFORMANCE

10.4.1 Predicting exactly matched labels

In this section we present quantitative results with the raw labels, matched exactly in NeuroVault metadata, as described in Section 10.1.2.1. Hence, we do not use the rules described in Section 10.1.2.3. With this exact matching, 96 unique CognitiveAtlas concepts are found in the metadata of at least 10 NeuroVault images. We train our models using these 96 labels and the 26 000 NeuroVault images that are annotated with at least one of them. Only 37 of these labels are found in IBC annotations with this labelling procedure. Performance is therefore measured on 6 500 maps from the IBC study and 37 labels.

10.4.1.1 *Model comparison*

We compare the performance for different models in this setting in Table 10.1. The models we consider obtain similar results. Still, the non-

Output distribution	Hidden layers	AUC	Pseudo Recall at 10
Bernoulli	0	0.80	0.42
Bernoulli	1	**0.81**	0.36
Bernoulli	3	0.79	0.34
Multinoulli	0	0.80	0.46
Multinoulli	1	0.80	0.38
Multinoulli	3	0.77	0.30

Table 10.1: NNoD performance on IBC — 37 original concepts.

linear Bernoulli model (a 2-layer perceptron with a sigmoid output activation function, see Section 9.3.2) achieves a slightly higher AUC (averaged over concepts). The best model is trained with an elastic net penalty (combined ℓ_1 and ℓ_2 regularization on the model weights of both layers), while applying dropout on the input and hidden layers (see Section 9.3.3). We also compare our decoder with pre-trained NeuroSynth and GCLDA models in Table 10.2. To do so, we compute AUC on labels that appear both in IBC CognitiveAtlas annotations, and on either NeuroSynth's or GCLDA's vocabulary.

10.4.1.2 *Detailed scores for the best model*

The AUC and some example decoding maps are shown in Fig. 10.6: our decoding model achieves higher-than-chance predictions for 35 out of 37 considered labels. The AUC metric seems meaningful. Indeed terms for which the AUC is high, such as "left finger response execution", tend to have plausible decoding maps, whereas terms that are not decoded above chance level, such as "working memory", do not capture relevant brain regions in their decoding maps. The detailed AUC scores for each term and model are shown in Fig. D.2.

Model	Common labels	Model AUC	NNoD AUC
GCLDA	20	0.53	0.73
NeuroSynth	14	0.58	0.71

Table 10.2: Comparison with GCLDA and NeuroSynth. The AUC (both for the considered model and for NNoD) is averaged over CognitiveAtlas concepts found both in IBC annotations and in the considered model's vocabulary.

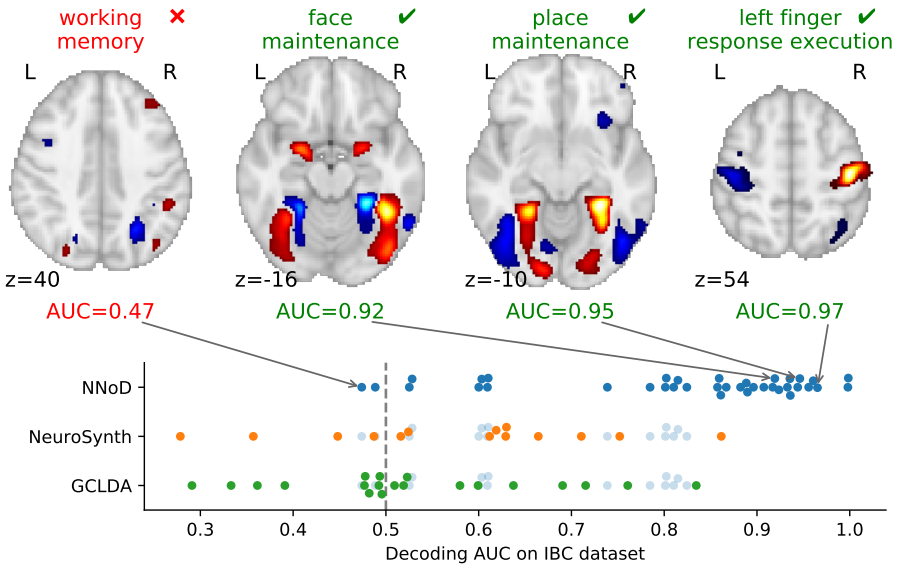


Figure 10.6: Decoding exactly-matched labels. We evaluate the AUC of our NNoD model on 37 labels matched in the IBC dataset (test set), after training it to decode 96 labels on all the other collections. Using pre-trained GCLDA and NeuroSynth models, we compare our results for the labels that also appear in the vocabulary recognized by these models (NNoD AUCs for terms in the vocabulary intersections are shown in light blue). Detailed scores for each label are shown in Appendix D and show that NNoD outperforms other methods for most labels. On the top, we show decoding maps for some example terms. Terms that are well decoded such as "place maintenance" have meaningful maps, whereas terms such as "working memory" whose neural correlates are poorly captured get low AUC scores. As the decoding maps do not have a meaningful scale, we threshold them arbitrarily at the 95^{th} percentile for visualization.

10.4.2 Predicting expanded labels

In this section we present results obtained with the improved labels, derived from CognitiveAtlas terms that are matched in NeuroVault annotations using the heuristics described in Section 10.1.2.3: collection-specific rules and missing label inference. Using these rules, some missing labels are inferred and the annotations become less sparse. Therefore, in this setting, 106 CognitiveAtlas labels occur in the annotations (despite merging 27 pairs of synonyms) of the 26 000 training maps. Among 106 labels, 51 are found in the enriched dataset annotations of the 6 500 IBC maps, that we use once again for validation.

10.4.2.1 *Model comparison*

The AUC scores are presented in Table 10.3. With these new annotations, once again the considered architectures yield similar performance, with the non-linear Bernoulli model (multi-class, sigmoid output activation function) obtaining a slightly higher AUC. In Table 10.4,

Output distribution	Hidden layers	AUC	Pseudo Recall at 10
Bernoulli	0	0.82	0.63
Bernoulli	1	**0.84**	0.66
Bernoulli	3	0.83	0.64
Multinoulli	0	0.80	0.63
Multinoulli	1	0.83	0.65
Multinoulli	3	0.81	0.63

Table 10.3: NNoD AUC on the IBC dataset, averaged 51 enriched concepts.

Model	Common labels	Model AUC	NNoD AUC
GCLDA	31	0.54	0.81
NeuroSynth	23	0.62	0.80

Table 10.4: Model comparison using enriched labels.

we show the comparison with NeuroSynth and GCLDA on labels that are recognized by these models and present in IBC annotations.

10.4.2.2 *Detailed scores for the best model*

The results are similar to those obtained with the exactly-matched labels, but with a slight improvement for NNoD, NeuroSynth and GCLDA, and a larger number of mental conditions, since more labels are found in the IBC dataset. All terms are decoded above chance level. Example decoding maps and decoding AUC scores for each term are shown in Fig. D.1 and Fig. D.2.

10.5 CONCLUSION

In this chapter, we have presented the first study to decode a wide variety of cognitive processes from functional brain images. Our statistical models achieve good classification performance, as measured by the AUC, for 90% of several dozen mental conditions found in the IBC dataset, which provides an extensive coverage of human cognition.

10.5.1 Large scale decoding

We have used thousands of statistical maps coming from many different sources, and of different natures. We found that the heterogeneity of these images is not a roadblock. In particular, models were trained on a dataset that contains Z-maps, T-maps and univariate beta-maps, and mixes subject-level and group-level maps. Still, these models provided good predictions for the subject-level Z-maps of the IBC dataset.

Removing T-maps from the training data slightly decreased performance, indicating that exploiting more data is more important than making the distinction between Z and T statistics (we did not try to use the `number_of_subjects` field from NeuroVault to transform T-maps into Z-maps).

10.5.2 Non-linear decoding architectures

On the considered feature and label spaces, non-linear models performed slightly better than linear ones for decoding (best linear AUC = 0.82, best non-linear AUC = 0.84), on a test set containing thousands of brain images. To our knowledge, even if this increase is minor, this is the first time that those models proved useful to increase decoding performance on fMRI images. Our approach also confirms that good decoding performance can be reached through aggressive dimensionality reduction, obtained from unsupervised decompositions. This has already been observed in several studies such as Dadi et al. (2019) and Mensch et al. (2017a). We did not try training models on the original voxels maps as it would have incurred a very important computational cost, given the size of our dataset.

10.5.3 Handling heterogeneous annotations

One of the most difficult aspects of this work was to construct usable annotations for our training images. CognitiveAtlas is the largest ontology of human cognition, and NeuroVault is the largest online repository of statistical maps of the brain. Still, less than 100 CognitiveAtlas concepts (out of over 800) could be found in NeuroVault metadata. Moreover, for most studies the NeuroVault annotations are very scarce and inconsistent. This is not surprising, as providing rich annotations for neuroimaging studies in a uniform way is a huge challenge. Even selecting the relevant vocabulary is a daunting task (Poldrack and Yarkoni, 2016). Still, better describing the semantic content, and the experiments behind shared data is crucial. In this study we found this to be much more important than detailed information about the acquisition settings or the analysis pipeline.

As is usually the case when the set of tags is unrestricted, NeuroVault annotations are very sparse and contain a myriad of very rare labels (Fig. 10.2). We explored several ways to densify these annotations and infer missing labels. We considered *smoothing* the label matrix, relying on term co-occurrences in large corpora, as was successfully done in an encoding context in Part III. For this, we used co-occurrences in documents of the NeuroQuery corpus (Chapter 4), but also word embeddings trained with Skip-Gram Negative

Sampling (SGNS) on larger corpora, such as dumps of Wikipedia[6] or PubMed Central[7]. These approaches did not improve prediction scores.

In the experiments presented here, we were more successful with rule-based completion of the labels than with continuous data-driven approaches (Section 10.1.2.3). This indicates that discrete graphs and high-dimensional embeddings can both be useful tools to cope with noisy and sparse data. In this study, although we decoded a larger and more varied set of mental conditions than what had been attempted before, we could control the terms that we include in our label set. This is different from the setting of Part II and Part III, where input features were extracted from free text or arbitrary queries, and models had to cope with a much wider vocabulary. Moreover, although they are mostly uncurated, NeuroVault metadata remain structured annotations. They are more restricted, and of a different nature, than free text. Therefore, building dedicated rules to exploit the structure of these annotations is beneficial.

Finally, although we did not explore this further, it is likely that statistical and rule-based label completion can be complementary approaches, with statistical associations capturing general similarities, and dedicated rules yielding a performance gain for specific tasks.

10.5.4 Capturing bias

Although NeuroVault is a large repository and its maps come from many different sources, some spurious associations between mental conditions and brain regions arise from experiments, stimuli and response modalities that are used very frequently in functional neuroimaging. Meta-analysis captures these systematic associations. Although carefully crafted baselines and contrasts alleviate this issue, mental processes can only be isolated by probing them in many different ways in order to discover which links with brain activity can be reproduced in varied settings. Meta-analysis allows us to measure these effects across studies and laboratories, and possibly identify processes such as "emotion perception" that are targeted very uniformly.

10.5.5 Combining encoding and decoding

A possible extension of this work is to combine it with coordinate-based meta-analyses, based on the literature. Encoding maps obtained with CBMA could be used to regularize decoding models, especially for labels that are seldom used in NeuroVault annotations. Indeed, meta-analysis on the literature can reliably cover a much larger vocabulary.

6 https://meta.wikimedia.org/wiki/Data_dumps
7 https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

Part V

CONCLUSION

11 CONCLUSION

In this thesis, we propose a new approach to meta-analysis of neuroimaging studies, based on predictive modelling. This framework enables embedding rich descriptions of experiments, cognitive processes or neurological diseases into brain space. Here we summarize some of the difficulties that we met, possible directions for future research and resources that resulted from our work.

11.1 PRACTICAL CHALLENGES

Comprehensive meta-analysis faces many challenges. An important one is the rarity of many neuroscience terms. This scarcity of observations is typical of machine learning problems that involve text or other high-cardinality sets of discrete objects. Several methods help us alleviate this problem. Collecting more data, and data of a higher quality, gives us access to more observations with a higher signal-to-noise ratio. Considering terms not in isolation, but in the context of their relationships, allows us to guide predictions for rare terms with the mappings of their better-understood neighbors. Adapted feature selection methods, either based on domain knowledge (e. g. considering only anatomical terms) or data-driven, keep the dimension and sparsity of input features under control, and prevent drowning the signal in noise coming from uninformative terms.

Rare terms or labels also raise the issue of adapted metrics and validation methods. Validation scores can be inflated by good predictions on frequent items, which are not always the most important. Ground truth to validate predictions for rare items is scarce. One possible approach is to make some terms artificially rare by masking them in randomly sampled documents. More generally, training and testing models on subsampled data, subsampling either articles, sections of articles, or terms based on their document frequency or category, helps us understand the strengths and limitations of our methods and improve them.

Regarding aspects that are more specific to neuroimaging, one of the major difficulties is access to good-quality data. Very few studies share their data on standard online platforms. Therefore, comprehensive meta-analysis is limited to peak activation coordinates, at the cost of a huge loss of information. Moreover, the brain images available on online platforms are poorly annotated. Standards for annotations mostly focus on details of the acquisition and analysis procedure,

which are easy to describe. Neuroscience still lacks a standardized way to describe the cognitive processes, behavioral or phenotypic data that underlie statistical maps. This is an important roadblock for meta-analysis. Finally, despite efforts towards openly shared data, accessing descriptions of experimental protocols and results to perform meta-analysis is difficult. Publications and data can only be considered to be openly shared if they can be downloaded automatically and in standard formats, and this is not often the case.

11.2 FUTURE DIRECTIONS

Having efficient encoding models that can handle a large class of queries opens new perspectives for meta-analysis. For example, we explored very briefly the possibility to use comprehensive mappings from text to brain regions, learned from the literature, to regularize models trained on specific tasks. The simplest example is through the definition of an ROI. In a small experiment, the performance of a decoder classifying different language tasks was improved by restricting the input features to an ROI selected through meta-analysis. The fact that our encoding models can handle any text, without needing to manually map queries to a restricted set of single terms, makes it possible to generalize this regularization strategy to other encoding or decoding tasks. Beyond the definition of an ROI, meta-analytic encoding mappings could regularize decoding coefficients by shrinking them towards the encoding maps, or by modulating other forms of regularization. Meta-analytic maps could also provide priors for small-sample encoding studies, thus reducing the degrees of freedom of encoding models, increasing statistical power, and reducing the variance of effect estimates.

To improve the encoding models presented in this thesis, future work could also focus on modelling noise on the design matrix, i. e. the TFIDF features. In regression settings the design matrix is often treated as a fixed quantity, but this may not be justified when dealing with features that represent text. Indeed, there are many different ways to describe an experiment, and the exact choice of terms can be considered as random. Noise on the input features may be the main source of errors in the encoding model's coefficients and predictions. Taking it into account might lead to better uncertainty estimation and better regularization strategies.

11.3 WHAT DID NOT WORK

During this thesis, we also explored several ideas that did not improve results.

11.3.1 More complex Natural Language Processing

To deal with the sparsity and high dimensionality of text, a common approach is to use *word embeddings*, trained for example with SGNS (Mikolov et al., 2013) or more modern techniques such as language models or Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). Replacing TFIDF features with word embeddings trained with SGNS degrades the performance of our encoding models, and hinders their interpretability. We obtain better results by smoothing the high-dimensional inputs and performing feature selection. A similar idea is to use word embeddings or other criteria to define similarities between documents and exploit them with kernel ridge regression. Again, we obtain better results by performing regression in the TFIDF feature space.

11.3.2 More specific text extraction

Fully automated meta-analytic tools such as NeuroSynth and NeuroQuery often raise the question that one article can provide results on different experiments, and that pooling coordinates from all activation maps can bias results. As we have the articles in XML format, we can conduct encoding experiments where we treat each table of stereotactic coordinates separately. Each table is then associated with specific text, such as the text of paragraphs that contain a hyperlink to it and the table caption. This does not improve predictions, suggesting that using denser and less noisy data is more important than isolating specific portions of the text. For other questions raised by fully automated meta-analysis, we recommend reading the NeuroSynth FAQ[1].

11.3.3 Compressing encoding targets

Reducing the dimension of brain maps for encoding, either through the use of atlases or parcellations, could reduce computational cost and denoise brain activations. We considered several atlases, data-driven parcellations, and dictionary loadings. All these approaches degrade performance and we obtain the best results by constructing the target maps with Gaussian KDE in the high-dimensional voxel space.

11.3.4 Data-driven smoothing of decoding labels

Using either encoding maps or similarities based on co-occurrence statistics for label inference in the decoding setting does not improve the classification scores of our pipeline. Instead, we are more suc-

1 https://neurosynth.org/faq/

cessful when curating and completing labels with a heuristic and rule-based approach. We hope that as ontologies continue to grow and annotations continue to improve, such an automatic curation can be performed in a more principled way.

11.4 RESOURCES RESULTING FROM THIS THESIS

Here we list resources that we created and shared during this thesis and that can be useful for other projects related to meta-analysis, machine learning and neuroimaging.

11.4.1 Data

In https://github.com/neuroquery/neuroquery_data, we share data and a pre-trained encoding model. This repository contains vocabulary files with document frequencies, term counts and TFIDF features for all the documents in our dataset, and extracted peak activation coordinates. We only share term frequency matrices, and not the text of articles, to comply with copyright regulations. The repository also contains a pre-trained encoding model that can be downloaded, used offline and integrated in software applications. This model is provided as a directory containing its vocabulary, term similarities, and regression coefficients. Keeping these resources separate helps reusability.

11.4.2 Software

In https://github.com/neuroquery/neuroquery, we share tested and documented code to train and use our encoding model. This library contains functions to automatically download and use the data and pre-trained model provided in the data repository. It also enables easily training new models from scratch, which helps the possibility to replicate and extend our experiments. The repository also contains some didactic examples that can be run online[2].

11.4.3 Online meta-analysis

In collaboration with Hande Gözükan and Romain Primet, we developed the NeuroQuery online platform: https://neuroquery.saclay.inria.fr. Users can enter a query and see a brain map of the most relevant regions, a list of related terms, and a list of related publications. The brain map can be downloaded in a standard format. The

2 https://mybinder.org/v2/gh/neuroquery/neuroquery.git/master?filepath=examples

model used in the online platform is provided in the data repository, and it is easy to download it and access the same functionality offline and programmatically.

Future work on this platform may involve better integration with other services such as NeuroSynth and CognitiveAtlas.

11.4.4 Contributions to other projects

During this thesis, I also participated in larger open-source projects. I have been actively contributing to Nilearn[3] throughout the duration of the thesis. I have also contributed to other Python packages such as Scikit-learn[4] and Nistats[5].

11.5 FINAL NOTE

This thesis introduces predictive meta-analysis. This approach studies the links that tie anatomical structures to mental processes or brain diseases in a supervised learning setting. In the encoding setting, our models use the description of an experiment, mental condition or disease to predict the probable locations of associated neurological observations. In the decoding setting, we analyze brain maps to predict the associated mental processes.

Predictive meta-analysis is complementary to standard methods, which focus on hypothesis testing and identifying effects that are reported consistently. Meta-analytic supervised encoding models help prepare neuroimaging studies – by predicting possible results, interpret statistical brain maps and interactively explore the literature. They can generalize to new combinations of terms. Our models can handle queries such as single terms or full articles. This makes them useful beyond the analysis of scientific neuroimaging publications, as they can embed into brain space reports that do not provide brain images nor stereotactic coordinates. We also propose a validation framework and metrics to evaluate and compare meta-analytic encoding and decoding models.

We provide an open-source implementation of our encoding models and expose a trained model through a web application. We also share a large dataset of term counts and stereotactic coordinates. We thus hope that predictive meta-analysis will be a useful addition to the neuroimaging community's toolbox of software packages and statistical methods.

3 https://nilearn.github.io/
4 https://scikit-learn.org/stable/
5 https://nistats.github.io/

REFERENCES

REFERENCES

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux (2014). "Machine learning for neuroimaging with scikit-learn." In: *Frontiers in neuroinformatics* 8, p. 14.

Baldassarre, Luca, Janaina Mourao-Miranda, and Massimiliano Pontil (2012). "Structured sparsity models for brain decoding from fMRI data." In: *2012 Second International Workshop on Pattern Recognition in NeuroImaging*. IEEE, pp. 5–8.

Beckmann, Christian F and Stephen M Smith (2004). "Probabilistic independent component analysis for functional magnetic resonance imaging." In: *IEEE transactions on medical imaging* 23.2, pp. 137–152.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation." In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Bohland, Jason, Hemant Bokil, Cara Allen, and Partha Mitra (2009). "The brain atlas concordance problem: quantitative comparison of anatomical parcellations." In: *PloS one* 4.9, e7200.

Bouma, Gerlof (2009). "Normalized (pointwise) mutual information in collocation extraction." In: *Proceedings of GSCL*, pp. 31–40.

Bowden, Douglas M and Richard F Martin (1995). "NeuroNames brain hierarchy." In: *Neuroimage* 2.1, pp. 63–83.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò (2013). "Power failure: why small sample size undermines the reliability of neuroscience." In: *Nature Reviews Neuroscience* 14.5, p. 365.

Byrd, Richard H, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu (1995). "A limited memory algorithm for bound constrained optimization." In: *SIAM J on Sci Comp* 16.5, pp. 1190–1208.

Bzdok, Danilo and BT Thomas Yeo (2017). "Inference in the age of big data: Future perspectives on neuroscience." In: *Neuroimage* 155, pp. 549–564.

Bzdok, Danilo, Michael Eickenberg, Olivier Grisel, Bertrand Thirion, and Gaël Varoquaux (2015). "Semi-supervised factored logistic regression for high-dimensional neuroimaging data." In: *Advances in neural information processing systems*, pp. 3348–3356.

Carp, Joshua (2012). "On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments." In: *Frontiers in neuroscience* 6, p. 149.

Chen, Colin and Ying Wei (2005). "Computational issues for quantile regression." In: *Sankhyā: The Indian Journal of Statistics*, pp. 399–417.

Cichocki, Andrzej and Anh-Huy Phan (2009). "Fast local algorithms for large scale nonnegative matrix and tensor factorizations." In: *IEICE transactions on fundamentals of electronics, communications and computer sciences* 92.3, pp. 708–721.

Cohen, William W, Pradeep Ravikumar, Stephen E Fienberg, et al. (2003). "A Comparison of String Distance Metrics for Name-Matching Tasks." In: *IIWeb*. Vol. 2003, pp. 73–78.

Cox, David D and Robert L Savoy (2003). "Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex." In: *Neuroimage* 19.2, pp. 261–270.

Dadi, Kamalaker, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, Gaël Varoquaux, Alzheimer's Disease Neuroimaging Initiative, et al. (2019). "Benchmarking functional connectome-based predictive models for resting-state fMRI." In: *Neuroimage* 192, pp. 115–134.

Damasio, Antonio R (1992). "Aphasia." In: *New England Journal of Medicine* 326.8, pp. 531–539.

Davie, Charles Anthony (2008). "A review of Parkinson's disease." In: *Br med bul* 86, p. 109.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis." In: *Journal of the American society for information science* 41.6, pp. 391–407.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805*.

Devlin, Joseph T and Russell A Poldrack (2007). "In praise of tedious anatomy." In: *Neuroimage* 37.4, pp. 1033–1041.

Dockès, Jérôme, Demian Wassermann, Russell Poldrack, Fabian Suchanek, Bertrand Thirion, and Gaël Varoquaux (2018). "Text to brain: predicting the spatial distribution of neuroimaging observations from text reports." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 584–592.

Dohmatob, Elvis, Arthur Mensch, Gael Varoquaux, and Bertrand Thirion (2016). "Learning brain regions via large-scale online structured sparse dictionary learning." In: *Advances in Neural Information Processing Systems*, pp. 4610–4618.

Dunn, Olive Jean (1961). "Multiple comparisons among means." In: *Journal of the American statistical association* 56.293, pp. 52–64.

Eickhoff, Simon B, Angela R Laird, Christian Grefkes, Ling E Wang, Karl Zilles, and Peter T Fox (2009). "Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty." In: *Human brain mapping* 30.9, pp. 2907–2926.

Eickhoff, Simon B, Danilo Bzdok, Angela R Laird, Florian Kurth, and Peter T Fox (2012). "Activation likelihood estimation meta-analysis revisited." In: *Neuroimage* 59.3, pp. 2349–2361.

Etzel, Joset A, Valeria Gazzola, and Christian Keysers (2009). "An introduction to anatomical ROI-based fMRI classification analysis." In: *Brain research* 1282, pp. 114–125.

Evans, Alan C, D Louis Collins, SR Mills, ED Brown, RL Kelly, and Terry M Peters (1993). "3D statistical neuroanatomical models from 305 MRI volumes." In: *1993 IEEE conference record nuclear science symposium and medical imaging conference*. IEEE, pp. 1813–1817.

Fonov, Vladimir S, Alan C Evans, Robert C McKinstry, CR Almli, and DL Collins (2009). "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood." In: *NeuroImage* 47, S102.

Fox, Peter T and Jack L Lancaster (2002). "Mapping context and content: the BrainMap model." In: *Nature Reviews Neuroscience* 3.4, p. 319.

Fox, Peter T, Angela R Laird, Sarabeth P Fox, P Mickle Fox, Angela M Uecker, Michelle Crank, Sandra F Koenig, and Jack L Lancaster (2005). "BrainMap taxonomy of experimental design: description and evaluation." In: *Human brain mapping* 25.1, pp. 185–198.

Friston, Karl J, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak (1994). "Statistical parametric maps in functional imaging: a general linear approach." In: *Human brain mapping* 2.4, pp. 189–210.

Gaonkar, B. and C. Davatzikos (2012). "Deriving statistical significance maps for SVM based image classification and group comparisons." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 723–730.

Gardner, Daniel, Huda Akil, Giorgio A Ascoli, Douglas M Bowden, William Bug, Duncan E Donohue, David H Goldberg, Bernice Grafstein, Jeffrey S Grethe, Amarnath Gupta, et al. (2008). "The neuroscience information framework: a data and knowledge environment for neuroscience." In: *Neuroinformatics* 6.3, pp. 149–160.

Gibbs, Alison L and Francis Edward Su (2002). "On choosing and bounding probability metrics." In: *International statistical review* 70.3, pp. 419–435.

Gorgolewski, Krzysztof J, Gael Varoquaux, Gabriel Rivera, Yannick Schwarz, Satrajit S Ghosh, Camille Maumet, Vanessa V Sochat, Thomas E Nichols, Russell A Poldrack, Jean-Baptiste Poline, et al. (2015). "NeuroVault. org: a web-based repository for collecting and

sharing unthresholded statistical maps of the human brain." In: *Frontiers in neuroinformatics* 9, p. 8.

Gramfort, Alexandre, Bertrand Thirion, and Gaël Varoquaux (2013). "Identifying predictive regions from fMRI with TV-L1 prior." In: *2013 International Workshop on Pattern Recognition in Neuroimaging*. IEEE, pp. 17–20.

Grosenick, Logan, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor (2013). "Interpretable whole-brain prediction analysis with GraphNet." In: *NeuroImage* 72, pp. 304–321.

Grün, Christian, Alexander Holupirek, and Marc H Scholl (2007). *Visually exploring and querying XML with BaseX*.

Haxby, James V, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex." In: *Science* 293.5539, pp. 2425–2430.

Hoyos-Idrobo, Andrés, Gaël Varoquaux, Yannick Schwartz, and Bertrand Thirion (2018). "FReM–scalable and stable decoding with fast regularized ensemble of models." In: *NeuroImage* 180, pp. 160–172.

Humphries, Colin, Jeffrey R Binder, David A Medler, and Einat Liebenthal (2006). "Syntactic and semantic modulation of neural activity during auditory sentence comprehension." In: *Journal of cognitive neuroscience* 18.4, pp. 665–679.

Ioannidis, John PA (2005). "Why most published research findings are false." In: *PLoS medicine* 2.8, e124.

Kanwisher, Nancy, Josh McDermott, and Marvin M Chun (1997). "The fusiform face area: a module in human extrastriate cortex specialized for face perception." In: *Journal of neuroscience* 17.11, pp. 4302–4311.

Koenker, Roger and Gilbert Bassett Jr (1978). "Regression quantiles." In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Koenker, Roger and Vasco d'Orey (1994). "Remark AS R92: A remark on algorithm AS 229: Computing dual regression quantiles and regression rank scores." In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 43.2, pp. 410–414.

Konrad, Kerstin and Simon B Eickhoff (2010). "Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder." In: *Human brain mapping* 31.6, pp. 904–916.

Kriegeskorte, Nikolaus, Rainer Goebel, and Peter Bandettini (2006). "Information-based functional brain mapping." In: *Proceedings of the National Academy of Sciences* 103.10, pp. 3863–3868.

Laird, Angela R, P Mickle Fox, Cathy J Price, et al. (2005). "ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts." In: *Hum brain map* 25, p. 155.

Laird, Angela R, Jack J Lancaster, and Peter T Fox (2005a). "Brainmap." In: *Neuroinformatics* 3.1, pp. 65–77.

Laird, Angela R, Jack L Lancaster, and Peter T Fox (2005b). "The social evolution of a human brain mapping database." In: *Neuroinformatics* 3.1, pp. 65–78.

Lancaster, Jack L, Diana Tordesillas-Gutiérrez, Michael Martinez, Felipe Salinas, Alan Evans, Karl Zilles, John C Mazziotta, and Peter T Fox (2007). "Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template." In: *Human brain mapping* 28.11, pp. 1194–1205.

Lee, Daniel D and H Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization." In: *Nature* 401.6755, p. 788.

Lipscomb, Carolyn E (2000). "Medical subject headings (MeSH)." In: *Bulletin of the Medical Library Association* 88.3, p. 265.

Liu, Han, Mark Palatucci, and Jian Zhang (2009). "Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery." In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, pp. 649–656.

Loula, João, Gaël Varoquaux, and Bertrand Thirion (2018). "Decoding fMRI activity in the time domain improves classification performance." In: *NeuroImage* 180, pp. 203–210.

Mensch, Arthur, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux (2016). "Dictionary learning for massive matrix factorization." In: *International Conference on Machine Learning*, pp. 1737–1746.

Mensch, Arthur, Julien Mairal, Danilo Bzdok, Bertrand Thirion, and Gaël Varoquaux (2017a). "Learning neural representations of human cognition across many fMRI studies." In: *Advances in Neural Information Processing Systems*, pp. 5883–5893.

Mensch, Arthur, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux (2017b). "Stochastic subsampling for factorizing huge matrices." In: *IEEE Transactions on Signal Processing* 66.1, pp. 113–128.

Michel, Vincent, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, and Bertrand Thirion (2011). "Total variation regularization for fMRI-based prediction of behavior." In: *IEEE transactions on medical imaging* 30.7, pp. 1328–1340.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems*, pp. 3111–3119.

Mitchell, Tom M, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just (2008). "Predicting human brain activity associated with the meanings of nouns." In: *science* 320.5880, pp. 1191–1195.

Mourao-Miranda, Janaina, Arun LW Bokde, Christine Born, Harald Hampel, and Martin Stetter (2005). "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data." In: *NeuroImage* 28.4, pp. 980–995.

Mumford, Jeanette A, Benjamin O Turner, F Gregory Ashby, and Russell A Poldrack (2012). "Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses." In: *Neuroimage* 59.3, pp. 2636–2643.

Naselaris, Thomas, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant (2011). "Encoding and decoding in fMRI." In: *Neuroimage* 56.2, pp. 400–410.

Ogawa, Seiji, Tso-Ming Lee, Alan R Kay, and David W Tank (1990). "Brain magnetic resonance imaging with contrast dependent on blood oxygenation." In: *proceedings of the National Academy of Sciences* 87.24, pp. 9868–9872.

Olshausen, Bruno A and David J Field (1997). "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* 37.23, pp. 3311–3325.

Owens, Mike (2006). *The definitive guide to SQLite*. Apress.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). "Scikit-learn: Machine learning in Python." In: *Journal of machine learning research* 12.Oct, pp. 2825–2830.

Penny, William D, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

Pereira, Francisco, Tom Mitchell, and Matthew Botvinick (2009). "Machine learning classifiers and fMRI: a tutorial overview." In: *Neuroimage* 45.1, S199–S209.

Piantadosi, Steven T (2014). "Zipf's word frequency law in natural language: A critical review and future directions." In: *Psychonomic bulletin & review* 21.5, pp. 1112–1130.

Pinel, Philippe, Bertrand Thirion, Sébastien Meriaux, Antoinette Jobert, Julien Serres, Denis Le Bihan, Jean-Baptiste Poline, and Stanislas Dehaene (2007). "Fast reproducible identification and large-scale databasing of individual functional cognitive networks." In: *BMC neuroscience* 8.1, p. 91.

Pinho, Ana Luísa, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, et al. (2018). "Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping." In: *Scientific data* 5, p. 180105.

Poldrack, Russell A (2006). "Can cognitive processes be inferred from neuroimaging data?" In: *Trends in cognitive sciences* 10.2, pp. 59–63.

– (2011). "Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding." In: *Neuron* 72.5, pp. 692–697.

Poldrack, Russell A, Jeanette A Mumford, and Thomas E Nichols (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Poldrack, Russell A and Tal Yarkoni (2016). "From brain maps to cognitive ontologies: informatics and the search for mental structure." In: *Annual review of psychology* 67, pp. 587–612.

Poldrack, Russell A, Aniket Kittur, Donald Kalar, Eric Miller, Christian Seppa, Yolanda Gil, D Stott Parker, Fred W Sabb, and Robert M Bilder (2011). "The cognitive atlas: toward a knowledge foundation for cognitive neuroscience." In: *Frontiers in neuroinformatics* 5, p. 17.

Poldrack, Russell A, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni (2017). "Scanning the horizon: towards transparent and reproducible neuroimaging research." In: *Nature Reviews Neuroscience* 18.2, p. 115.

Portnoy, Stephen, Roger Koenker, et al. (1997). "The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators." In: *Statistical Science* 12.4, pp. 279–300.

Richardson, Leonard (2007). "Beautiful soup documentation." In: *April*.

Rifkin, Ryan M and Ross A Lippert (2007). "Notes on regularized least squares." In:

Rubin, Timothy N, Oluwasanmi Koyejo, Krzysztof J Gorgolewski, Michael N Jones, Russell A Poldrack, and Tal Yarkoni (2016). "Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition." In: *bioRxiv*, p. 059618.

– (2017). "Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition." In: *PLoS computational biology* 13.10, e1005649.

Salimi-Khorshidi, Gholamreza, Stephen M Smith, John R Keltner, Tor D Wager, and Thomas E Nichols (2009). "Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies." In: *Neuroimage* 45.3, pp. 810–823.

Salton, Gerard and Christopher Buckley (1988). "Term-weighting approaches in automatic text retrieval." In: *Information processing & management* 24.5, pp. 513–523.

Sayers, Eric (2009). "The E-utilities in-depth: parameters, syntax and more." In: *Entrez Programming Utilities Help [Internet]*.

Scott, David W (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, Bernard W (1986). *Density estimation for statistics and data analysis*. Vol. 26. CRC press.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn (2011). "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." In: *Psychological science* 22.11, pp. 1359–1366.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting." In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Sternberg, Saul (1969). "Memory-scanning: Mental processes revealed by reaction-time experiments." In: *American scientist* 57.4, pp. 421–457.

Talairach, J and P Tournoux (1988). "Co-planar stereotaxic atlas of the human brain. 1988." In: *New York: Theime*.

Thirion, Bertrand (2016). "Functional neuroimaging group studies." In: *Handbook of Neuroimaging Data Analysis*, pp. 335–354.

Thirion, Bertrand, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, and Jean-Baptiste Poline (2007). "Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses." In: *Neuroimage* 35.1, pp. 105–120.

Thirion, Bertrand, Gaël Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline (2014). "Which fMRI clustering gives good brain parcellations?" In: *Frontiers in neuroscience* 8, p. 167.

Turkeltaub, Peter E, Guinevere F Eden, Karen M Jones, and Thomas A Zeffiro (2002). "Meta-analysis of the functional neuroanatomy of single-word reading: method and validation." In: *Neuroimage* 16.3, pp. 765–780.

Turkeltaub, Peter E, Simon B Eickhoff, Angela R Laird, Mick Fox, Martin Wiener, and Peter Fox (2012). "Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses." In: *Human brain mapping* 33.1, pp. 1–13.

Turney, Peter D and Patrick Pantel (2010). "From frequency to meaning: Vector space models of semantics." In: *Journal of artificial intelligence research* 37, pp. 141–188.

Varoquaux, Gaël and Russell A Poldrack (2019). "Predictive models avoid excessive reductionism in cognitive neuroimaging." In: *Current opinion in neurobiology* 55, pp. 1–6.

Varoquaux, Gaël, Sepideh Sadaghiani, Philippe Pinel, Andreas Kleinschmidt, Jean-Baptiste Poline, and Bertrand Thirion (2010). "A group model for stable multi-subject ICA on fMRI datasets." In: *Neuroimage* 51.1, pp. 288–299.

Varoquaux, Gaël, Yannick Schwartz, Russell A Poldrack, Baptiste Gauthier, Danilo Bzdok, Jean-Baptiste Poline, and Bertrand Thirion (2018). "Atlases of cognition with large-scale human brain mapping." In: *PLoS computational biology* 14.11, e1006565.

Wager, Tor D, John Jonides, and Susan Reading (2004). "Neuroimaging studies of shifting attention: a meta-analysis." In: *Neuroimage* 22.4, pp. 1679–1693.

Wager, Tor D, Martin Lindquist, and Lauren Kaplan (2007). "Meta-analysis of functional neuroimaging data: current and future di-

rections." In: *Social cognitive and affective neuroscience* 2.2, pp. 150–158.

Wager, Tor D and Edward E Smith (2003). "Neuroimaging studies of working memory." In: *Cognitive, Affective, & Behavioral Neuroscience* 3.4, pp. 255–274.

Wager, Tor D, K Luan Phan, Israel Liberzon, and Stephan F Taylor (2003). "Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging." In: *Neuroimage* 19.3, pp. 513–531.

Walker, Francis O (2007). "Huntington's disease." In: *The Lancet* 369.9557, pp. 218–228.

Wand, MP (1994). "Fast computation of multivariate kernel estimators." In: *J Comp Graph Stat* 3.4, pp. 433–445.

Westfall, Jacob, Thomas E Nichols, and Tal Yarkoni (2016). "Fixing the stimulus-as-fixed-effect fallacy in task fMRI." In: *Wellcome open research* 1.

Yarkoni, Tal, Russell A Poldrack, David C Van Essen, and Tor D Wager (2010). "Cognitive neuroscience 2.0: building a cumulative science of human brain function." In: *Trends in cognitive sciences* 14.11, pp. 489–496.

Yarkoni, Tal, Russell A Poldrack, Thomas E Nichols, David C Van Essen, and Tor D Wager (2011). "Large-scale automated synthesis of human functional neuroimaging data." In: *Nature methods* 8.8, p. 665.

Yeo, BT Thomas, Fenna M Krienen, Simon B Eickhoff, Siti N Yaakub, Peter T Fox, Randy L Buckner, Christopher L Asplund, and Michael WL Chee (2014). "Functional specialization and flexibility in human association cortex." In: *Cerebral cortex* 25.10, pp. 3654–3672.

Yi, Congrui and Jian Huang (2017). "Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression." In: *J Comp Graph Stat* 26, p. 547.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net." In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.

APPENDICES

A DATASET DETAILS

A.1 ATLASES

List of the atlases whose labels were included in the vocabulary.

name	URL
talairach	http://www.talairach.org/talairach.nii
harvard_oxford	http://www.nitrc.org/frs/download.php/7700/HarvardOxford.tgz
destrieux	https://www.nitrc.org/frs/download.php/7739/destrieux2009.tgz
aal	http://www.gin.cnrs.fr/AAL-217
JHU-labels	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#JHU-labels
Striatum-Structural	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#Striatum-Structural
STN	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#STN
Striatum-Connectivity-7sub	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#Striatum-Connectivity-7sub
Juelich	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#Juelich
MNI	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#MNI
JHU-tracts	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#JHU-tracts
Thalamus	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases#Thalamus

B PEAK DENSITY PREDICTION

B.1 DECOMPOSITION OF THE LOSS FUNCTION

Here we detail how the text-to-brain loss function can be written as a sum over brain regions.

$$\mathcal{E}(\hat{p}, q) = \int_{\mathbb{R}^3} \delta(\hat{p}(z) - q(z))\, dz \tag{B.1}$$

$$= \int_{\mathbb{R}^3} \delta(\hat{p}(z) - q(z)) \sum_{k=0}^{m} \mathbb{I}_k(z)\, dz \tag{B.2}$$

$$= \int_{\mathbb{R}^3} \delta(\hat{p}(z) - q(z)) \sum_{k=1}^{m} \mathbb{I}_k(z)\, dz \tag{B.3}$$

because $\hat{p}$ and q are null in the background and $\delta(0,0) = 0$. Moreover, $\mathbb{I}_k \neq 0 \implies r_{k'} = 0 \ \ \forall k' \neq k$, so:

$$\mathcal{E}(\hat{p}, q) = \sum_{k=1}^{m} \int_{\mathbb{R}^3} \delta\left(\hat{y}_k r_k(z) - \sum_{t=1}^{d} x_t B_{t,k} r_k(z)\right) \mathbb{I}_k(z)\, dz \tag{B.4}$$

$$= \sum_{k=1}^{m} \int_{\mathbb{R}^3} \delta\left(\frac{\hat{y}_k}{v_k} - \frac{\sum_{t=1}^{d} x_t B_{t,k}}{v_k}\right) \mathbb{I}_k(z)\, dz \tag{B.5}$$

$$= \sum_{k=1}^{m} v_k \delta\left(\frac{\hat{y}_k}{v_k} - \frac{\sum_{t=1}^{d} x_t B_{t,k}}{v_k}\right) . \tag{B.6}$$

B.2 ℓ_2-PENALIZED ℓ_1 REGRESSION

We have the minimization problem:

$$\hat{B} = \underset{B}{\operatorname{argmin}} \left(\|\hat{Y} - X B\|_1 + \lambda \|B\|_F^2 \right) \tag{B.7}$$

where $X = (x_i) \in \mathbb{R}^{n \times d}$ and $\hat{Y} = (\hat{y}_i) \in \mathbb{R}^{n \times m}$ and $\lambda \in \mathbb{R}_+$.
The problem is equivalent to:

$$\underset{Z,B}{\operatorname{argmin}} \left(\|Z\|_1 + \lambda \|B\|_F^2 \right) \tag{B.8}$$

$$\text{s.t. } \hat{Y} - X B - Z = 0 . \tag{B.9}$$

Introducing the dual variable $\nu \in \mathbb{R}^{n \times m}$, the Lagrangian is:

$$L(Z, B, \nu) = \|Z\|_1 + \lambda \|B\|_F^2 + \operatorname{Tr}(\nu^T (\hat{Y} - X B - Z)) . \tag{B.10}$$

The derivative with respect to $\mathbf{B}$ is

$$2\lambda \mathbf{B} - \mathbf{X}^\mathsf{T} \boldsymbol{\nu} \,, \tag{B.11}$$

so minimizing with respect to $\mathbf{B}$ yields $\mathbf{B} = \frac{\mathbf{X}^\mathsf{T}\boldsymbol{\nu}}{2\lambda}$ and

$$\min_{\mathbf{B}} L(\mathbf{Z}, \mathbf{B}, \boldsymbol{\nu}) = \|\mathbf{Z}\|_1 + \operatorname{Tr}(\boldsymbol{\nu}^\mathsf{T}\hat{\mathbf{Y}} - \boldsymbol{\nu}^\mathsf{T}\mathbf{Z} - \frac{1}{4\lambda}\boldsymbol{\nu}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{\nu}) \,. \tag{B.12}$$

The dual norm of the l_1 norm is l_∞, so minimizing with respect to $\mathbf{Z}$ we get the Lagrange dual function

$$g(\boldsymbol{\nu}) = \min_{\mathbf{Z},\mathbf{B}} L(\mathbf{Z}, \mathbf{B}, \boldsymbol{\nu}) \tag{B.13}$$

$$= \begin{cases} \operatorname{Tr}(\boldsymbol{\nu}^\mathsf{T}\hat{\mathbf{Y}} - \frac{1}{4\lambda}\boldsymbol{\nu}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{\nu}) & \text{if } \|\boldsymbol{\nu}\|_\infty \leqslant 1 \\ -\infty & \text{otherwise} \,. \end{cases} \tag{B.14}$$

Hence the dual problem is:

$$\hat{\boldsymbol{\nu}} = \operatorname*{argmax}_{\boldsymbol{\nu}} \left(\operatorname{Tr}(\boldsymbol{\nu}^\mathsf{T}\hat{\mathbf{Y}} - \frac{1}{4\lambda}\boldsymbol{\nu}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{\nu}) \right) \quad \text{s.t. } \|\boldsymbol{\nu}\|_\infty \leqslant 1 \,. \tag{B.15}$$

g is differentiable; its gradient is

$$\nabla_g(\boldsymbol{\nu}) = \hat{\mathbf{Y}} - \frac{1}{2\lambda}\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{\nu} \,. \tag{B.16}$$

We solve this problem using an efficient algorithm: L-BFGS-B (Byrd et al., 1995). Then we obtain the primal solution as $\hat{\mathbf{B}} = \frac{\mathbf{X}^\mathsf{T}\hat{\boldsymbol{\nu}}}{2\lambda}$.

This approach can easily be generalized to the quantile loss for any quantile $\rho \in (0, 1)$. The regularized quantile regression problem, for a quantile ρ, can be written by separating the positive and negative parts of the error:

$$\operatorname*{argmin}_{\mathbf{U},\mathbf{V},\mathbf{B}} \; \rho\,\mathbf{U} + (1-\rho)\,\mathbf{V} + \lambda\,\|\mathbf{B}\|_F^2 \tag{B.17}$$

$$\text{s.t.} \quad \mathbf{U} \succeq 0 \tag{B.18}$$

$$\mathbf{V} \succeq 0 \tag{B.19}$$

$$\mathbf{Y} - \mathbf{X}\mathbf{B} - \mathbf{U} + \mathbf{V} = 0 \,. \tag{B.20}$$

A similar derivation as the one used for the LAD problem yields the dual quantile regression problem:

$$\operatorname*{argmax}_{\boldsymbol{\nu}} \; \boldsymbol{\nu}^\mathsf{T}\mathbf{Y} - \frac{1}{4\lambda}\boldsymbol{\nu}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{\nu} \tag{B.21}$$

$$\text{s. t.} \quad -\boldsymbol{\nu} + \rho \succeq 0 \tag{B.22}$$

$$\boldsymbol{\nu} - \rho + 1 \succeq 0 \,. \tag{B.23}$$

This problem can thus be solved with the same approach as the one we used for LAD regression.

B.3 DETAILS ON TOKENIZATION

The first step of Term Frequency · Inverse Document Frequency (TFIDF) feature extraction is tokenization: transforming a string of characters into a list of entries from our vocabulary. Here we summarize the tokenization pipeline that we implemented. It follows the usual steps of text preprocessing for information extraction. To make explanations concrete, we consider tokenizing an example (meaningless) document:

```
"Lesions in the supplementary motor areas (and the
auditory cortex)"  .
```

B.3.1 Extracting words

The first step is to transform a string of characters into a list of words. This is done by matching a simple regular expression. For our example document, this step yields the list:

```
["Lesions", "in", "the", "supplementary", "motor",
"areas", "and", "the", "auditory", "cortex"]  .
```

B.3.2 Standardizing words

Extracted words are then standardized. In all our experiments, they are converted to lower case. We also experimented with stemming or lemmatizing them, but found available stemmers, such as implementations of the Porter stemmer, to be too aggressive in some cases. This may be due to the fact that we are considering a very specialized vocabulary. The above list becomes

```
["lesions", "in", "the", "supplementary", "motor",
"areas", "and", "the", "auditory", "cortex"]  .
```

If we are using a Porter stemmer, it becomes

```
["lesion", "in", "the", "supplementari", "motor",
"area", "and", "the", "auditori", "cortex"]  .
```

B.3.3 Removing stop words

Some stop words, such as “the” and “and”, are removed from the list of extracted words. The filtered list for our example is:

```
["lesions", "supplementary", "motor", "areas",
"auditory", "cortex"]  .
```

B.3.4 Extracting phrases

Once we have a list of (standardized and filtered) words, the next step is to extract collocations. Collocations are sequences of words that often appear together, such as "anterior cingulate cortex". Collocations are sometimes called *n-grams*; for example "anterior cingulate cortex" is a 3-gram, or trigram. Once the previous steps have created a list of words, we group them into phrases (possibly of length 1).

In other information extraction settings, the vocabulary is often built from the text corpus itself. Collocations are then discovered by considering all the collocations that occur at least once in the whole corpus, and selecting the most relevant ones. This selection can be a simple threshold on frequency. It can also be based on more elaborate strategies, such as keeping n-grams that occur much more often than would be expected if the words they contain occurred independently (Bouma, 2009; Mikolov et al., 2013). When discovering n-grams in this way, the number of n-grams to consider grows very rapidly with n, the maximum number of words in a vocabulary entry.

Our vocabulary, on the other hand, comes from curated vocabulary lists such as the Medical Subject Headings. It includes some relatively long expressions, such as "caudal anterior cingulate cortex" or "mild traumatic brain injury". We do not consider all possible sub-parts of these expressions. Our greedy tokenization method extracts full expressions sequentially, rather than building all possible pairs, triplets, and quadruplets of words. We build, for each word, a graph of the terms that can follow it to form an expression in our vocabulary. An example for the term "supplementary" is shown in Fig. B.1. When a term is found in the text, we consume the following words, matching the longest sequence that results in an expression in the vocabulary. After grouping adjacent terms into phrases, our example becomes

```
[["lesions"], ["supplementary", "motor", "areas"],
["auditory", "cortex"]]  .
```

Other tokenizing methods might have produced

```
[["lesion"], ["supplementary"], ["motor"],
["area"], ["supplementary", "motor"], ["motor",
"area"], ["supplementary", "motor", "area"]].
```

If we want to add the sub-parts of each expression, it can be easily done once the TFIDF features are extracted, by multiplying the (sparse) TFIDF matrix with the sparse matrix $\mathbf{S} \in \mathbb{R}^{v \times v}$ (where v is the vocabulary size) such that $\mathbf{S}_{ij} = 1$ if phrase i is a sub-part of phrase j and 0 otherwise.

This different tokenization method explains why some terms from the NeuroSynth vocabulary appear too rare to be included in ours. They are pairs of terms that are the end of longer expressions,

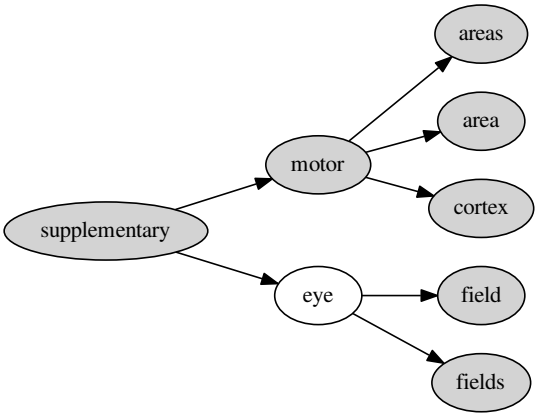


Figure B.1: tokenization paths after "supplementary". Paths ending with a filled (gray) node correspond to expressions in the vocabulary. Paths ending with a white node are not in the vocabulary. When extracting phrases, we match the longest sequence that ends in a gray node. For example, "supplementary eye field" is parsed as `[["supplementary", "eye", "field"]]`, but "supplementary eye" is parsed as `[["supplementary"], ["eye"]]` .

or spurious collocation detections. For example, “superior temporal gyrus (stg)” is tokenized with our method as `[["superior", "temporal", "gyrus" ], ["stg"]]` , which is why the collocation `["gyrus", "stg"]`, from NeuroSynth’s vocabulary, almost never occurs and is not included in our vocabulary. The complete list of expressions present in NeuroSynth’s vocabulary but not ours is provided in Appendix B.4.

B.3.5 Merging synonyms

Finally, expressions that are very similar are merged to remove duplicates due to small spelling variations and reduce the dimensionality of the TFIDF features. To discover similar expressions, we rely on the Jaro-Winkler distance (Cohen, Ravikumar, and Fienberg, 2003). If two expressions have a very small Jaro-Winkler distance, the least frequent is replaced by the most frequent one. For example, “brain stem” is replaced by “brainstem” and “movements” is replaced by “movement”. This is a less aggressive alternative to stemming for improving robustness to spelling variability. After this last step, our tokenized document finally becomes

```
[["lesion"], ["supplementary", "motor", "area"],
["auditory", "cortex"]]  .
```

Counting the occurrences of each unique phrase then yields the Bag-of-words (BOW) features from which the TFIDF features are derived.

B.4 DISCARDED NEUROSYNTH TERMS

Although all NeuroSynth terms are considered for inclusion in our vocabulary, some are found too rare and are discarded:

> gyrus stg, cortex vmpfc, gyrus posterior, gyrus insula, cortex ppc, gyrus medial, cortex vlpfc, gyrus mfg, sulcus sts, incentive delay, gyrus cerebellum, memory wm, deficit hyperactivity, gyrus anterior, compulsive disorder, hyperactivity disorder, disorder ocd, mind tom, face ffa, cortex dacc, stress disorder, cortex mpfc, frontal eye, cortex dorsolateral, gyrus ifg, gyrus superior, motor sma, disorder ptsd, lobule ipl

These terms are all bigrams. They are present in NeuroSynth's vocabulary but not ours because of differences in the tokenization process. For example "gyrus stg" occurs in the expression "superior temporal gyrus (stg)", which is parsed as `[[ "superior", "temporal", "gyrus" ], [ "stg"]]` in our case.

C NEUROQUERY DETAILS

C.1 REWEIGHTED RIDGE ON TOY DATASET

As a sanity check for our reweighted ridge regression model, we apply it to a toy synthetic dataset. We generate a small dataset. We choose a number of samples $n = 260$, a number of targets $m = 560$, a total number of regressors $d = 160$ to be roughly proportional to the size of our real data. We set a smaller number of informative features $u < d$ arbitrarily to 10. We build a design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, and a true coefficient matrix $\mathbf{B}^* \in \mathbb{R}^{u \times m}$. We generate

$$\mathbf{Y} \in \mathbb{R}^{n \times m} = \mathbf{X}_{:,:u}\, \mathbf{B} + \mathbf{E}\,, \tag{C.1}$$

where $\mathbf{E} \in \mathbb{R}^{n \times m}$ is Gaussian noise. For all the outputs, only the first u features are informative. The design matrix, regression coefficients and dependent variables are created with `scikit-learn`'s `make_regression` (Pedregosa et al., 2011). We can check that the reweighted ridge outperforms a usual ridge regression, as well as a multi-task lasso (Liu, Palatucci, and Zhang, 2009) (whose hyperparameter is set with an inner cross-validation loop), in Fig. C.1. It also

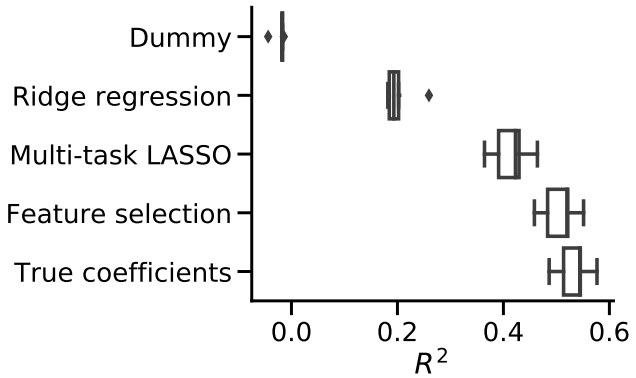


Figure C.1: Cross-validation R^2 scores. We see that on this dataset with many uninformative features, the reweighted ridge with feature selection and the multi-task lasso both outperform a uniform ridge regression baseline. "Dummy" predicts the mean of the training set and measures chance level.

correctly recovers the support of the coefficients (Fig. C.2). The multi-task lasso is included here for reference, but it would be extremely costly to fit to our real problem, which is 50 times bigger than this toy example.

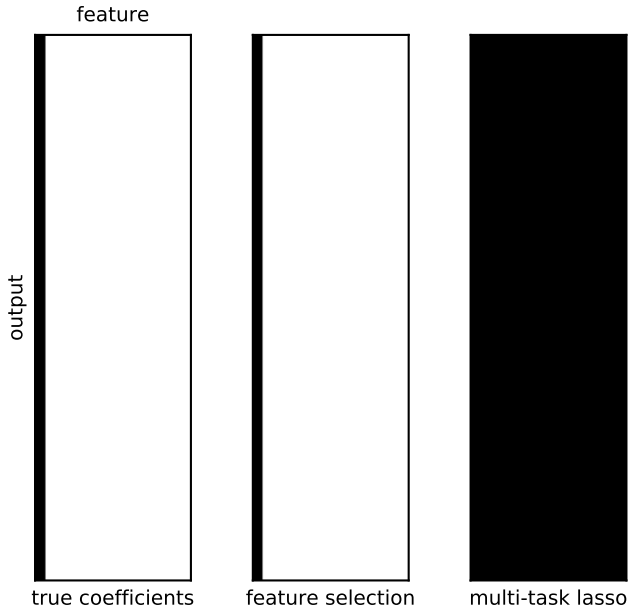


Figure C.2: Support of coefficients for the true coefficients, the feature selection procedure, and the multi-task lasso. Each row corresponds to an output and each column corresponds to a feature. Non-zero entries are shown in black. We see that on this toy example the feature selection procedure correctly recovers the true coefficients' support.

C.2 NEUROQUERY BRAIN COVERAGE

In Fig. C.3, we assess how well the brain is covered by the few terms chosen by our feature selection procedure. We threshold the supervised regression coefficients and count, for each voxel, the number of terms (coefficients) for which it is above the threshold. We can see that most brain areas are in the support of several brain maps (i. e. covered by several selected terms).

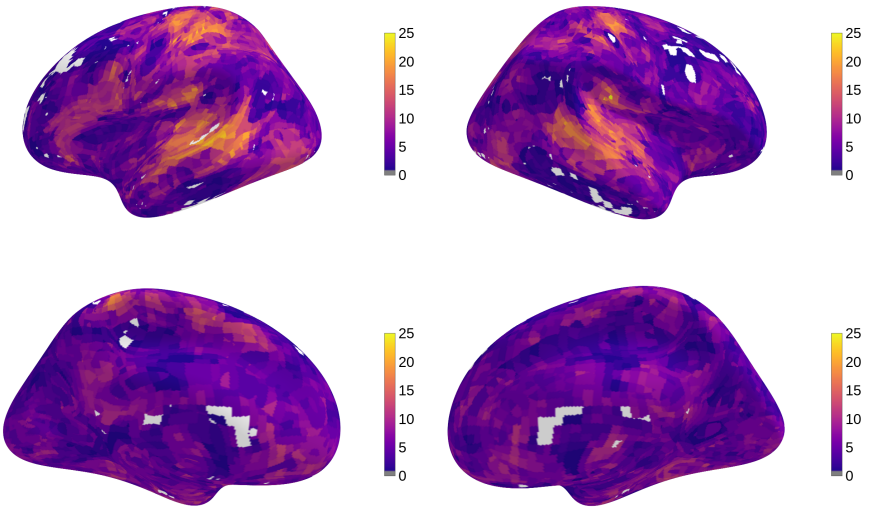


Figure C.3: Brain coverage: we consider only the Z maps of the $\approx$ 200 terms selected by the supervised regression. We threshold them to control the FDR at 5% across all voxels and all terms, and count, for each voxel, the number of terms for which it is above the threshold. We see that most brain areas are covered by several terms, and that only a small part of the brain (in white) is not covered.

D | DECODING RESULTS

D.1 EXAMPLE MAPS WITH ENRICHED LABELS

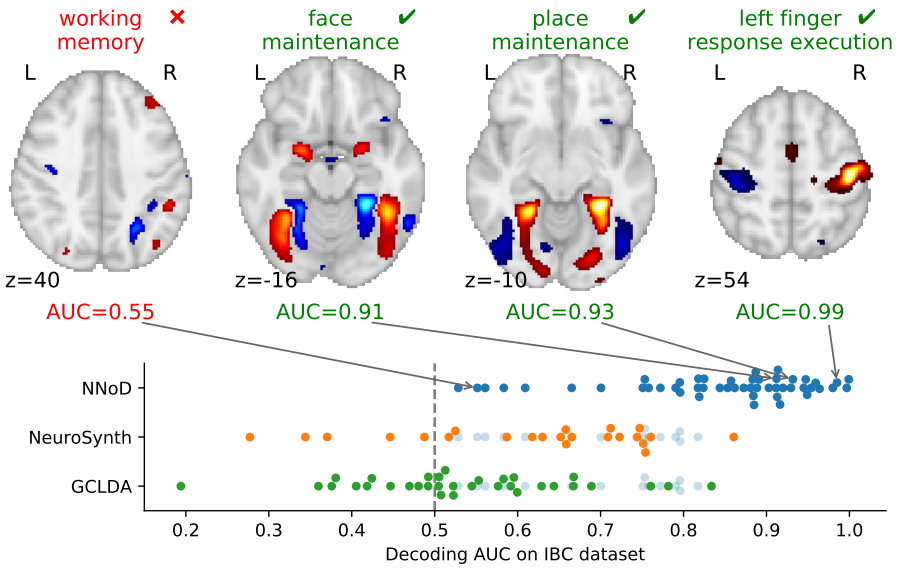


Figure D.1: Performances on enriched concepts. Compared with original labels decoding, after enriching our training dataset with heuristics, a similar model manages to decode more labels (51 in the IBC dataset) with a slightly better accuracy.

D.2 DETAILED SCORES PER LABEL

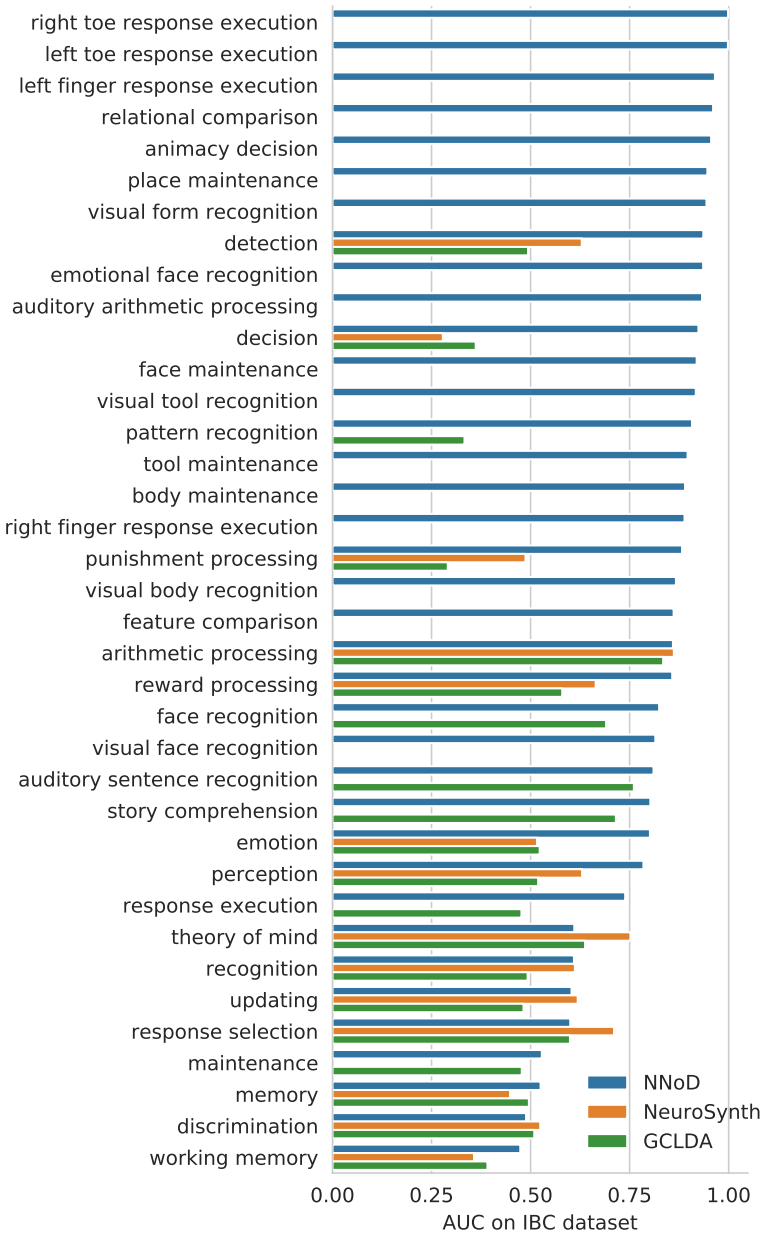


Figure D.2: Classification AUC for each term on the IBC test set. Most terms are decoded far above chance level. We also report the AUC of NeuroSynth and GCLDA models, pre-trained on the NeuroSynth dataset of coordinates, for the terms in their vocabulary.

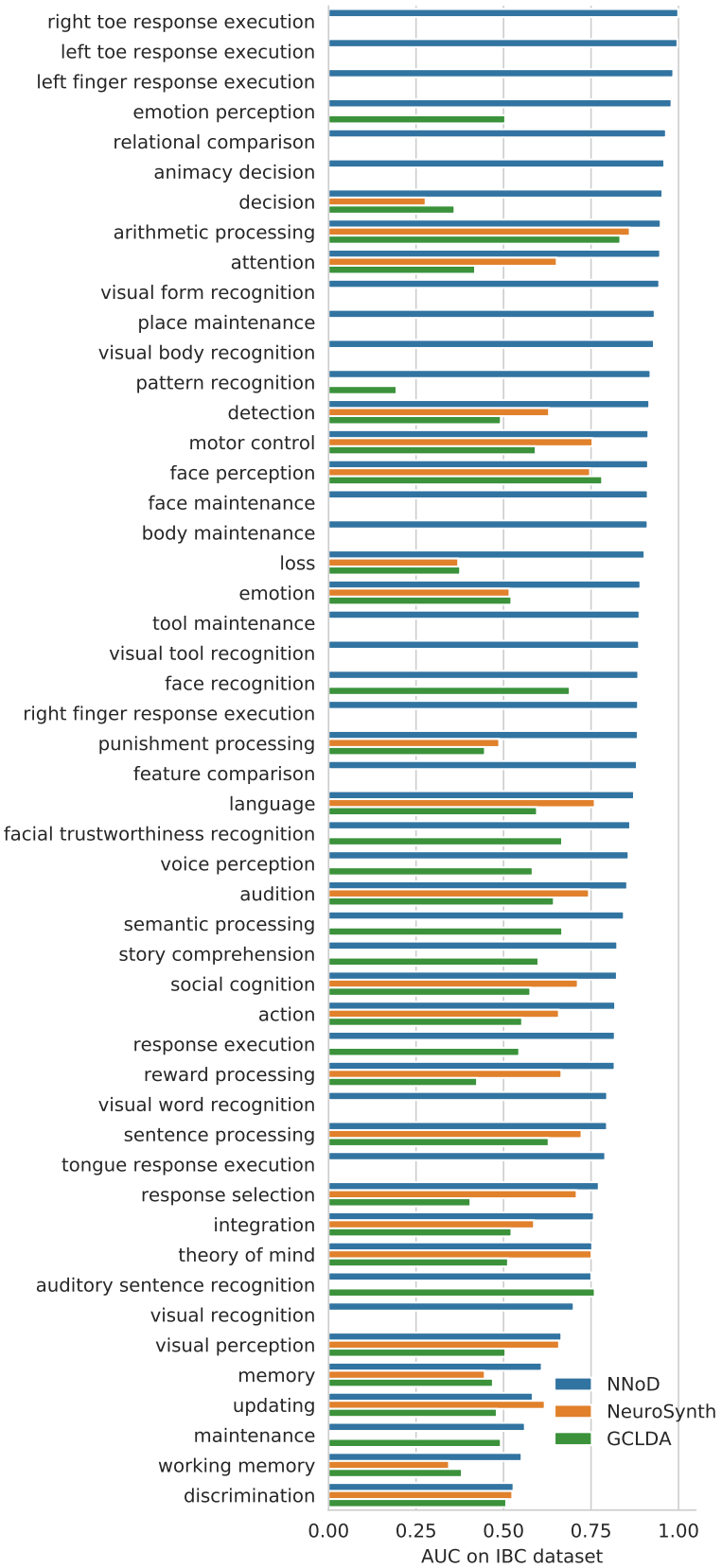


Figure D.3: Classification AUC when using our heuristics to enrich the labels. All models perform slightly better than with the exactly-matched labels, and more terms are found in the IBC test set after label inference.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Modèles statistiques pour méta-analyses d'études de neuroimagerie.

Mots clés : Méta-analyse, apprentissage statistique, neuroimagerie

Résumé : La neuroimagerie permet d'étudier les liens entre la structure et le fonctionnement du cerveau. Des milliers d'études de neuroimagerie sont publiées chaque année. Il est difficile d'exploiter cette grande quantité de résultats. En effet, chaque étude manque de puissance statistique et peut reporter beaucoup de faux positifs. De plus, certains effets sont spécifiques à un protocole expérimental et difficiles à reproduire.
Les méta-analyses rassemblent plusieurs études pour identifier les associations entre structures anatomiques et processus cognitifs qui sont établies de manière consistente dans la littérature. Les méthodes classiques de méta-analyse commencent par constituer un échantillon d'études focalisées sur un même processus mental ou une même maladie. Ensuite, un test statistique permet de délimiter les régions cérébrales dans lesquelles le nombre d'observations reportées est significatif.
Dans cette thèse, nous introduisons une nouvelle forme de méta-analyse, qui s'attache à construire des prédictions plutôt qu'à tester des hypothèses. Nous introduisons des modèles statistiques qui prédisent la distribution spatiale des observations neurologiques à partir de la description textuelle d'une expérience, d'un processus cognitif ou d'une maladie cérébrale. Notre approche est basée sur l'apprentissage statistique supervisé qui fournit un cadre classique pour évaluer et comparer les modèles. Nous construisons le plus grand jeu de données d'études de neuroimagerie et de coordonnées stéréotaxiques existant, qui rassemble plus de 13 000 publications. Dans la dernière partie, nous nous intéressons au décodage: prédire des états psychologiques à partir de l'activité cérébrale.
La méta-analyse standard est un outil indispensable pour distinguer les vraies découvertes du bruit et des artefacts parmi les résultats publiés en neuroimagerie. Cette thèse introduit des méthodes adaptées à la méta-analyse prédictive. Cette approche est complémentaire de la méta-analyse standard, et aide à interpréter les résultats d'études de neuroimagerie ainsi qu'à formuler des hypothèses ou des a priori statistiques.

Title : Statistical models for comprehensive meta-analyses of neuroimaging studies

Keywords : Meta-analysis, statistical learning, neuroimaging

Abstract :
Thousands of neuroimaging studies are published every year. Exploiting this huge amount of results is difficult. Indeed, individual studies lack statistical power and report many spurious findings. Even genuine effects are often specific to particular experimental settings and difficult to reproduce.
Meta-analysis aggregates studies to identify consistent trends in reported associations between brain structure and behavior. The standard approach to meta-analysis starts by gathering a sample of studies that investigate a same mental process or disease. Then, a statistical test delineates brain regions where there is a significant agreement among reported findings.
In this thesis, we develop a different kind of meta-analysis that focuses on prediction rather than hypothesis testing. We build predictive models that map textual descriptions of experiments, mental processes or diseases to anatomical regions in the brain. Our supervised learning approach comes with a natural quantitative evaluation framework, and we conduct extensive experiments to validate and compare statistical models. We collect and share the largest existing dataset of neuroimaging studies and stereotactic coordinates. This dataset contains the full text and locations of neurological observations for over 13 000 publications.
In the last part, we turn to decoding: inferring mental states from brain activity. We perform this task through meta-analysis of fMRI statistical maps collected from an online data repository. We use fMRI data to distinguish a wide range of mental conditions.
Standard meta-analysis is an essential tool to distinguish true discoveries from noise and artifacts. This thesis introduces methods for predictive meta-analysis, which complement the standard approach and help interpret neuroimaging results and formulate hypotheses or formal statistical priors.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France