COPPE
UFRJ
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia

A DESIGN STUDY ABOUT WINNOWING NEURONAL HYPOTHESES FROM THE VISUAL CORTEX

Pedro de Souza Asad

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Claudio Esperança

Rio de Janeiro
Dezembro de 2022

A DESIGN STUDY ABOUT WINNOWING NEURONAL HYPOTHESES
FROM THE VISUAL CORTEX

Pedro de Souza Asad

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Claudio Esperança

Aprovada por: Prof. Claudio Esperança
Prof. Ricardo Guerra Marroquim
Prof. Juliana Guimarães Martins Soares
Prof. Asla Medeiros e Sá
Prof. Carla Maria Dal Sasso Freitas

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2022

Asad, Pedro de Souza

A design study about winnowing neuronal hypotheses from the visual cortex/Pedro de Souza Asad. – Rio de Janeiro: UFRJ/COPPE, 2022.

XVIII, 190 p.: il.; 29, 7cm.

Orientador: Claudio Esperança

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2022.

Referências Bibliográficas: p. 162 – 176.

1. Visualização de informação. 2. Estudo de design. 3. Córtex visual. 4. Eletrofisiologia. 5. Aprendizado de máquina. I. Esperança, Claudio. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

To my grandparents

Acknowledgements

I would like to thank, in sort of an increasing order of intimacy, some central people in the quality of representatives of everyone whom I would otherwise miss or simply have no space to account for. It has been seven years after all, people have come and gone, disease has struck and lifted, and humankind lives on.

First, Donald Knuth and Leslie Lamport, who made TeX and LaTeX possible, as academic writing is ridiculously (more) cumbersome without them. The legions of open-source (creators, maintainers, and free folk), they who bestow us with the many flavors of GNU/Linux, the Python programming language and its thriving library ecosystem, and platforms like Git and Docker, the spells that empower us wizards to do the unthinkable — when we are not busy recompiling kernels. Alexandra Elbakyan, for believing so wholeheartedly in world of open scientific knowledge and living under a pirate flag in the name of that vision. Professors Anna Vilanova from the Eindhoven University of Technology, who reviewed the early designs of our visualization concept and provided much insightful advice, and Elmar Eisemann, head of the Computer Graphics and Visualization group at the Delft University of Technology, who made possible the singular, ground-shaking experience of being a stranger, in a strange land. All LFCOG members that participated in the user study, most notably professors Juliana Guimarães and Mario Fiorani, who opened their lab's doors and helped lay much of the foundational work for this thesis. My professors and mentors, Ricardo Marroquim and Claudio Esperança, who taught me many things in computer graphics and a few beyond it, who supported my studies in many ways, and who never ceased to believe this thesis would be finished (eventually). My relatives, friends, and colleagues, who persistently showed interest and enthusiasm for my studies — and consistently prompted me with the tormenting question of *when is it due?*

And last, but not least, I thank Luciany, spouse, friend, and lover, she who nourishes, supports, and walks along — and whom I think deserves a D.Sc. title of her own.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UM ESTUDO DE DESIGN SOBRE A FILTRAGEM DE HIPÓTESES NEURONAIS DO CÓRTEX VISUAL

Pedro de Souza Asad

Dezembro/2022

Orientador: Claudio Esperança

Programa: Engenharia de Sistemas e Computação

Esta tese trata da aplicação de métodos de visualização de informação à compreensão e ao aprimoramento do trabalho cognitivo de neurocientistas. Mais especificamente, relata os resultados de um estudo de design realizado com um grupo de seis pesquisadores, usando uma coleção de gravações eletrofisiológicas extracelulares do córtex visual primata obtida por eles. Apesar deste campo de pesquisa estar ativo há décadas, envolve configurações experimentais delicadas e o volume, ruído e complexidade dos dados implicam que analisar partes substanciais dos mesmos permanece desafiador e trabalhoso. No espírito da metodologia do estudo de design, uma modalidade central do campo da visualização de informação, exploramos uma destas coleções de dados com a ajuda de dois especialistas do domínio, interagindo com eles para entender quais perguntas eles têm sobre estes dados e como eles abordam sua investigação. Dado o desafio de filtrar o conjunto de observações neuronais hipotéticas, um aplicativo web para selecionar um conjunto de hipóteses limitado foi construído e quatro especialistas adicionais foram convidados a utilizar a ferramenta e relatar suas impressões. Analisamos comparativamente os dados de telemetria para determinar a regularidade e consistência de suas decisões e treinamos alguns modelos de aprendizado de máquina para aproximar suas escolhas. Descobrimos que seu processo de tomada de decisão é consideravelmente subjetivo e afetado por fatores aleatórios e difíceis de quantificar mas, no entanto, passível de aprimoramento semi-automático por tais modelos. A tese se encerra com uma intensa discussão de resultados quantitativos e qualitativos do estudo com usuários, com reflexões sobre nosso design e suas limitações e com um relato de possíveis direções para transformá-lo numa solução completa para interativamente limpar, explorar e consultar registros eletrofisiológicos do córtex visual.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

A DESIGN STUDY ABOUT WINNOWING NEURONAL HYPOTHESES FROM THE VISUAL CORTEX

Pedro de Souza Asad

December/2022

Advisor: Claudio Esperança

Department: Systems Engineering and Computer Science

This thesis concerns the application of information visualization methods to understand and enhance the cognitive work of neuroscientists. More specifically, it reports findings of a design study conducted with a group of six researchers, using a collection of extracellular electrophysiological recordings from the primate visual cortex obtained by them. Despite being an active research field for decades, it involves delicate experimental setups and the volume, noise, and complexity of the datasets imply that analyzing substantial portions of them remains challenging and time-consuming. In the spirit of the design study methodology, a core modality of the information visualization field, we explored such a collection of data with the aid of two domain experts, interacting with them to understand what inquiries they have about this data and how they go about investigating them. Given the challenge of filtering the hypothetical neuronal observations, a web application for winnowing a limited hypothesis set was built and four additional experts were invited to use the tool and report their impressions. We performed a comparative analysis of telemetry data to determine the regularity and consistency of their decision-making and trained a handful of machine learning models to approximate their choices. We found that their decision-making process is considerably subjective and affected by random, hard-to-quantify factors but nonetheless amenable to semi-automatic enhancement by such models. The thesis closes with an intensive discussion of quantitative and qualitative results from the user study, with reflections on our design and its limitations, and with an accounting of possible directions for extending it into a full-fledged solution for interactively cleaning, exploring, and querying electrophysiological recordings of the visual cortex.

Contents

List of Figures **xi**

List of Tables **xiv**

List of abbreviations **xvi**

1 Introduction **1**

1.1 Neuroscience background . . . 2

1.1.1 Neurons, spikes, spike trains, and firing rates . . . 3

1.2 Visual cortex . . . 7

1.2.1 Early visual system: aggregation of receptive fields . . . 8

1.2.2 Modeling receptive fields . . . 9

1.2.3 Stimulus, response, and functional attributes . . . 12

1.3 Extra-cellular electrophysiology . . . 12

1.3.1 Data acquisition . . . 13

1.4 Challenges . . . 15

1.5 Summary of contributions . . . 17

2 Data and methodologies **18**

2.1 The dataset and its preparation . . . 18

2.1.1 Visual stimulus types . . . 19

2.1.2 Spike detection and sorting . . . 22

2.1.3 Mapping of V1 and V2 receptive fields . . . 27

2.1.4 Functional attribute derivation . . . 28

2.1.5 Original dataset structure . . . 30

2.2 Design study literature review . . . 31

2.2.1 Conventions . . . 32

2.2.2 Design study definition . . . 32

2.2.3 Design frameworks . . . 33

2.2.4 Contributions of design studies . . . 39

2.2.5 Task analysis . . . 40

2.3 Fundamentals of machine learning . . . 42

2.3.1 AutoML 45

3 Design study 47

3.1 Methodology 48

3.2 Data abstraction 49

3.3 Cognitive work analysis, requirements, and tasks 55

3.3.1 Overall goals and requirements 55

3.3.2 Restricting the study scope 59

3.3.3 Abstract tasks 61

3.4 Data encoding 62

3.4.1 Informal study: preliminary encodings 62

3.4.2 Enhanced and new encodings 73

3.5 User study 77

3.5.1 Dataset selection 80

3.5.2 Participants profile 81

3.5.3 Web application 81

4 Quantitative and qualitative results 89

4.1 General results 89

4.1.1 General approval and rejection 89

4.1.2 Approval by anatomy and physiology 91

4.1.3 Parameters' influence 94

4.1.4 Influence on functional attributes 95

4.2 User behavior 98

4.2.1 Assignment completion 99

4.2.2 Session duration 99

4.2.3 Time-to-decision 102

4.3 Decision similarity 107

4.3.1 Overall similarity 107

4.3.2 Per-stimulus, per-decision similarity 107

4.3.3 Population vs. decision consensus 110

4.4 Decision predictability 112

4.4.1 Dataset preparation and model validation 113

4.4.2 Regression models compared 120

4.4.3 Ternary classification and ranking 120

4.5 Evaluation interviews 123

4.5.1 Statement implications 130

4.5.2 Debatable decisions 135

5 Discussion and conclusion 141
5.1 Design validation and limitations 141
5.1.1 Downstream path 141
5.1.2 Upstream path 144
5.2 Discussion of quantitative results 147
5.2.1 General results implications 147
5.2.2 User behavior implications 148
5.2.3 Decision similarity implications 149
5.2.4 Approval regression 150
5.3 Reflections 152
5.3.1 Domain problem characterization and task analysis difficulty . 153
5.3.2 Design study iterations 154
5.3.3 Design limitations 154
5.3.4 Time drainers and research focus 155
5.3.5 Technical aspects 156
5.4 Open questions and future directions 157
5.4.1 Cognitive biases and viz direction 157
5.4.2 Choosing parameters 158
5.4.3 Statistical significance 158
5.4.4 Neural coding and automatic winnowing 159
5.4.5 Improvements to encodings 159
5.4.6 Next steps for machine learners 160
5.5 Conclusion 161

References 162

A Data dictionary 177

B Feature engineering 181
B.1 Angular encoding 181
B.2 Min-max scaling 181
B.3 Standard scaling 182
B.4 Box-Cox power scaling 182
B.5 Yeo-Johson power scaling 182

C Machine learning model details 184

Index 187

List of Figures

1.1 General illustration of a neuron 4
1.2 Simulating the Hodgkin-Huxley model with different input currents . 6
1.3 Spike trains and firing rates 7
1.4 Visual system components and their receptive fields 8
1.5 An On-Off receptive field modeled as a Gabor function 10
1.6 A texture-like receptive field modeled as a Gabor function 10
1.7 The four lobes of the brain and the dual visual pathway 11
1.8 Overview of extracellular electrophysiological data acquisition 14

2.1 Dark moving bar stimulus schematics and intensity function 20
2.2 Drifting gratings stimulus schematics and intensity function 21
2.3 Overview of nested design study and validation framework 36

3.1 Data acquisition and derivation 52
3.2 Domain-specific workflow 57
3.3 Waveforms plot 63
3.4 Response map design used during initial user study 64
3.5 Early proposal for encoding summative responses to gratings stimuli . 65
3.6 Second proposal for encoding summative responses to gratings stimuli 67
3.7 Early overview panel for a recording site 69
3.8 Evidence of an orientation-selective RF 70
3.9 Evidence of a RF with no gratings tuning 70
3.10 Inconclusive RF evidence with weak gratings tuning 71
3.11 Inconclusive RF evidence with time-located orientation tuning 71
3.12 Inconclusive RF evidence with a soft spatial frequency selectivity . . 71
3.13 Evidence of large RF with time-located, inconclusive gratings response 71
3.14 Waveform summary of one summative and three individual observations 74
3.15 Response map plot 75
3.16 Receptive field polargram 75
3.17 Grating response encoding: direction, contrast, and frequency 76

3.18 Grating response encoding: direction, speed, and colors 77
3.19 Details-on-hovering and click-and-drag zooming response panels . . . 78
3.20 Gratings details-on-hovering and waveforms toggling 79
3.21 Observations by anatomy and physiology 82
3.22 Observation winnowing screen 86
3.23 Details of a moving bars observation 87
3.24 Details of a gratings observation 87
3.25 Web application login page 88
3.26 Web application dashboard 88

4.1 Mean approval by anatomical position 92
4.2 Major approvals by anatomical position 93
4.3 Correlation coefficients between parameterical and moving bars functional attributes by hemisphere 96
4.4 Correlation coefficients between parameterical and gratings functional attributes by hemisphere 97
4.5 Time spent by users on each screen they visited 100
4.6 Distribution of screen times per user 101
4.7 Time-to-decision distributions per stimulus and decision 104
4.8 When users opened the details view 106
4.9 Overall decision similarity 108
4.10 Decision similarity per decision and stimulus 109
4.11 Allocation of summative population by minimum consensus and stimulus 111
4.12 Correlation matrix of moving bars functional attributes 116
4.13 Correlation matrix of gratings functional attributes 116
4.14 PCA projections of first six components of moving bars functional features 118
4.15 PCA projections of first six components of gratings functional features 119
4.16 Train and test loss in all models per observation and stimulus type . . 122
4.17 Possible encoding of anatomical attributes using positional marks and scales at the top of the screen 132
4.18 Differences between moving bars summative and individual observations 1 136
4.19 Differences between moving bars summative and individual observations 2 137
4.20 Differences between gratings summative and individual observations . 138
4.21 Debatable moving bars decisions 139
4.22 Debatable gratings decisions 140

5.1 Outcome of design validation 142

C.1 Gradient boosting regression (GBR) learning curves 185
C.2 Regularized linear regression (RLR) learning curves 185
C.3 Support vector regression (SVR) learning curves 186

List of Tables

2.1 Secondary stimulus parameters values and variation groups 22

2.2 Waveform features used in spike sorting 27

3.1 Functional attributes common to all stimulus types 53

3.2 Moving bars functional attributes 54

3.3 Gratings functional attributes 54

3.4 Groups of gratings parameters varying with direction 68

3.5 Categories observed by domain experts when inspecting early encodings of functional attributes 70

4.1 Major and unanimous approval/rejection rates by stimulus type . . . 90

4.2 Major and unanimous approval/rejection rates by stimulus type . . . 90

4.3 Parameter influence on unanimous and majority approvals for gratings observations 95

4.4 Time-to-decision statistics per stimulus and action types 103

4.5 Kolmogorov-Smirnov test for per-user approval vs. rejection time-to-decisions distributions 105

4.6 Spearmann's rank-order correlation coefficients between consensus/disagreement and time-to-decision 106

4.7 Spearmann's rank-order correlation coefficients between experience and time-to-decision per decision type 107

4.8 Train-test split sizes for regression models 114

4.9 Model hyperparameters and their ranges 117

4.10 Ratio of cumulative variance retained by principal components in moving bars dataset 120

4.11 Ratio of cumulative variance retained by principal components in gratings dataset 120

4.12 Approval regression scores for all models 121

4.13 Approval regression scores for all models 121

4.14 Ternary classification metrics for all models 124

4.15 User statements gathered during evaluative interviews 126

A.1 Anatomical and physiological attributes 177
A.2 Parametrical attributes (spike detection) 178
A.3 Parametrical attributes (spike sorting) 178
A.4 Parametrical attributes (others) 178
A.5 Functional attributes (common) 179
A.6 Functional attributes (moving bars) 179
A.7 Functional attributes (gratings) 180

List of abbreviations

AD analog-to-digital. 14, 16

AI artificial intelligence. 146

ANN artificial neural network. 5, 13, 44

AutoML automated machine learning. 45, 112–115, 120, 150–152

CACW computer-assisted collaborative work. 60

CLI command line interface. 24

CV cross-validation. 114, 120, 151, 184

CVD circular variance for direction. 29, 72, 95, 160

CVO circular variance for orientation. 30, 72, 160

CWA cognitive work analysis. 40, 42, 55, 59

CytOx cytochrome oxidase. 13

DI direction index. 30, 72, 95

DNN deep neural network. 11

DSP digital signal processing. 22, 24

ECDF empirical cumulative distribution function. 98, 99, 102

ecephys extra-cellular electrophysiological. 1, 2, 12, 13, 15–18, 27, 34, 47, 49–51, 56, 58, 81, 157, 161

EDA exploratory data analysis. 58, 60, 143, 154, 155

EEG electroencephalography. 13

FWHM full width at half-maximum. 27

GBR gradient boosting regression. xiii, 113–115, 117, 120, 150–152, 184, 185

HCI human-computer interaction/interfaces. 31

IIR infinite impulse response. 25

IUDS inter-user decision similarity. 107, 108, 145, 147, 149

KFCV K-fold cross-validation. 184

KS Kolmogorov-Smirnov. 98, 99

LFCOG Laboratory of Cognitive Physiology. 1, 13, 19, 47, 55, 74, 81

LFP local field potential. 23, 25, 31

LGN lateral geniculate nucleus. 8, 9, 11

MEA multi-electrode array. 13, 16, 30, 31, 84, 91, 94, 155

ML machine learning. 2, 18, 42, 45, 112–115, 131, 144, 150, 152, 155, 160, 161

MLE maximum-likelihood estimation. 182, 183

MRI magnetic resonance imaging. 3, 13

MSE mean squared error. 113, 120, 121, 150–152

NWB neurodata without borders. 155

OI orientation index. 30, 72

PCA principal component analysis. 24, 26, 27, 51, 115, 117, 151, 181

PLX Plexon Inc. data files extension. 24, 26, 72, 95, 144, 155

PSTH peri-stimulus time histogram. 7

RBF radial basis function. 113

RF receptive field. 7–12, 16, 19, 20, 23, 27–29, 47, 54, 56, 62, 64, 69–72, 74, 75, 83, 94, 133–135, 139, 145–148, 157, 158

RGC retinal ganglion cell. 8, 9

RLR regularized linear regression. xiii, 113–115, 117, 120, 150, 151, 184, 185

SDK software development kit. 155

SGD stochastic gradient descent. 44

SNR signal-to-noise ratio. 64, 157

SPKC spike-continuous. 23–25, 31, 72, 95

SVG scalable vector graphics. 143

SVR support vector regression. xiii, 113–115, 117, 120, 151, 184, 186

TDD test-driven development. 144

TTD time-to-decision. 102, 106, 107, 145, 149

UI user interface. 83

viz information visualization. 1, 47, 49, 143, 146, 147, 153, 157, 158

WB wideband. 23–25, 31, 95

Chapter 1

Introduction

Cognitive neuroscience seeks to understand how cognitive capabilities arise from neuronal population dynamics by using methods from statistics and information theory to relate sensory stimulus to the observed electrical responses of these populations. The visual system is one of the most studied parts of the brain and some reasons for studying it include developing therapies for visual impairments, creating brain-machine interfaces, and even training artificial neuronal networks (ANNs) for object detection and tracking in robot vision.

Our interest in this topic found its way into a information visualization (viz) research project in collaboration with medical experts from the Laboratory of Cognitive Physiology (LFCOG), who have studied visual cortex over the years, and have thus recorded a considerable amount of neural activity in response to visual stimuli. One of these collections, which we used throughout this study, contains about 250 recording sessions of three macaque monkey subjects, totaling more than 1.5 TB of data. Meanwhile, viz is a prolific research field with lots of opportunity for interdisciplinary and applied research, and its methods have been applied successfully to many sub-areas of the medical sciences. Therefore, in the face of the immensity of this dataset, its idiosyncrasies, and the many open questions our collaborators posed, we set out to apply a design study methodology with one guiding vision: to augment the researchers' analytical power and data utilization capabilities.

In the present chapter, we lay out the neuroscience context necessary to understand our collaboration with the LFCOG and the challenges thence identified. We introduce many concepts from neuroscience, in particular the extra-cellular electrophysiological (ecephys) experimental framework, by which the aforementioned dataset was obtained. This chapter closes with an enumeration of the main challenges presented by the domain field and data, and of the contributions we offer.

Regarding the remainder of the thesis, Chapter 2 will discuss further details about the dataset and how we performed data preparation and derivation using domain-specific algorithms. It will also review the design study literature that laid out the foundation for our collaborative effort, and review a few machine learning (ML) concepts that are important for the quantitative analysis discussed later, as part of Chapter 4. Chapter 3 narrates our about developing encodings and identifying tasks for cleaning and exploring the ecephys dataset, which culminated in a web tool for judging the quality of neuronal responses observed to some visual stimulus types. Therefore, that chapter will also describe the that we performed as part of this tool's evaluation. Then, Chapter 4 discusses various quantitative and qualitative results of this user study. The former were obtained both by analyzing the tool's telemetry data, and by training a host of ML models to predict user decisions, and the latter were gathered during post-user study interviews with the participants. Finally, Chapter 5 will discuss implications of those results, threading the quantitative and qualitative aspects together in order to identify the design's strong points, limitations, and most promising future directions.

1.1 Neuroscience background

> The ultimate goal of neural science is to understand how the flow of electrical signals through neural circuits gives rise to mind [...] How do different patterns of interconnections give rise to different perceptions and motor acts?
>
> – KANDEL *et al.* [1]

The brain is probably the most complex organ developed by evolution. However, despite everything we known about its anatomy and physiology, we are only beginning to understand how function emerges from the interaction between its many constituent parts. The adult human brain has about 10^{11} nerve cells, each with thousands of connections, or synapses, to other cells, and they vary in many aspects, like their strength, the type of neurotransmitters that deliver signals, and whether they have an excitatory or inhibitory effect on the neuron that receives signals from the other (respectively called post-synaptic and pre-synaptic neurons) [1–3].

In one front, many researchers study brains from a *systems* approach, seeking to understand how smaller neuronal populations (the so called cortical columns) represent observed stimuli and cognitive states and how they perform computations on these representations [1, 3–21].

That line of investigation, which extends the framework for understanding the visual system advanced by famous neuroscientist David Marr [22] in the early 1980's, heavily relies on statistical and computational methods to infer quantifiable stimulus information from neuronal activity recorded during the presentation of carefully controlled stimuli, and on simulations of neuronal behavior models.

Meanwhile, others believe that the most important contemporary challenge for neuroscience is to obtain complete mappings of neural systems, *i.e.,* the listing of all neurons, their synapses, and related attributes (*e.g.,* the location of neurons, whether synapses are inhibitory or excitatory, and how strong they are), of an individual [2, 23]. This body of information, the so called connectome (in analogy with the now popular term *genome*), is believed by its proponents to be the ultimate level of detail from which neural function will be elucidated, and has emerged in more recent years thanks to the availability of ever more powerful microscopic scanning and reconstruction technologies [24].

Prominent researchers from systems neuroscience and connectomics are constantly debating the merits and flaws of both approaches [25, 26]. Most notably, systems neuroscience experiments are capable of observing neuronal population dynamics in great spatiotemporal detail, but only for populations of up to a few thousands of individuals (using techniques like optogenetics [27]), so they cannot provide an integrative picture of cognition. Conversely, connectomics studies can provide nearly complete and detailed topographical and topological maps of the brain, but the highest resolution is achieved by *post-mortem* studies that cannot provide accounts of live activity.

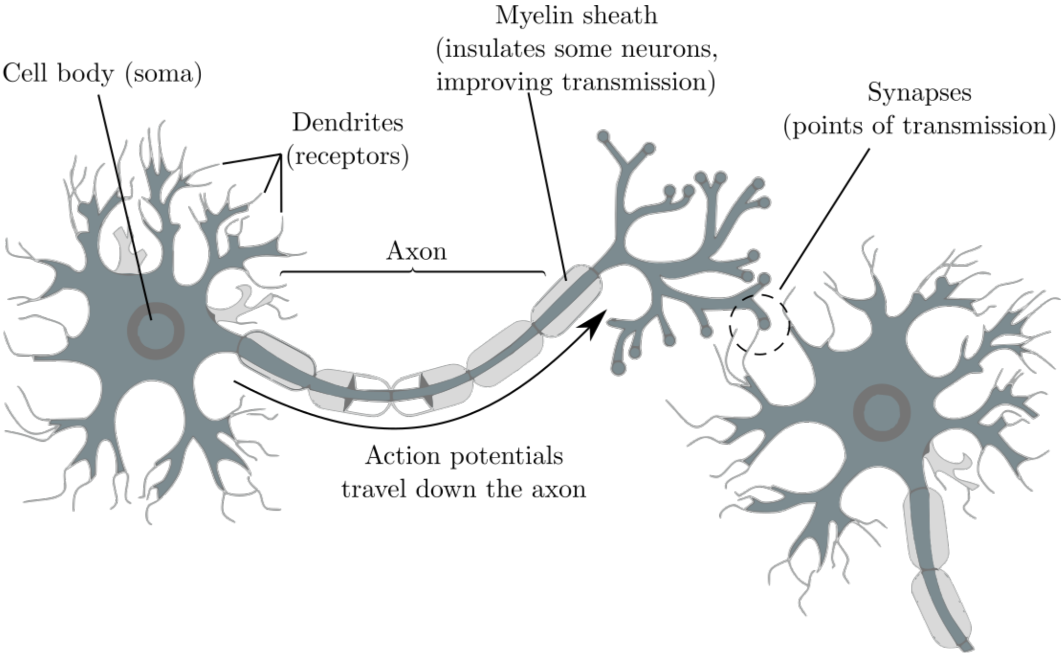
Naturally, attempts to close in the gap are made on both fronts. For instance, diffusion magnetic resonance imaging (MRI) allows to map *in vivo* connectomes, including from human beings [28], although the achieved resolution is at the macro scale [29]. Meanwhile, recent advances in electrode technology allow to record from a few hundred simultaneous neurons [30] per probe, even from freely moving mice [31]. In any case, simulating electrical activity in large-scale populations is currently not viable, with the exceptions of small nervous systems like that of *C. ellegans* (a worm species with a fairly regular connectome comprised of about 300 neurons [2]), which brings computational challenges onto the table. Finally, somewhere in the middle, an emerging field called network neuroscience blends concepts from network science and dynamic systems in order to study viable models of communication between brain regions like diffusion, routing (akin to artificial telecommunication networks), and hybrid models [32]. It attempts to bridge the gap between the other two areas by explaining how local computations can be transmitted and integrated at coarser levels. In any case, it is clear that only by combining several points of view will we be able to get a clearer picture of cognition.

1.1.1 Neurons, spikes, spike trains, and firing rates

Neurons and synapses are some of the most fundamental building blocks in the nervous system, and can be understood at various levels of detail. As illustrated in

Figure 1.1, neurons are always composed of four morphologically defined regions, although they may vary in size and arrangements [1]: soma (the cell body), dendrites, axon, and synapses.

Figure 1.1: General illustration of a neuron. Adapted from `https://commons.wikimedia.org/wiki/File:Components_of_neuron.jpg`, by Jennifer Walinga, released under CC-by-SA 4.0 license.



Besides the morphological description of the nervous system, DAYAN and ABBOT [3] identify three categories of nervous system models:

Descriptive Tersely describe what neurons and neural circuits do, without regards for how and why they do it. They summarize vast amounts of experimental data, usually in the form of probabilistic models.

Mechanistic Explain how the observed behavior of neurons and neuronal populations emerges from the interplay of its constituent parts, given precise anatomical and physiological descriptions.

Interpretive By using computational and information-theoretic principles, these models attempt to link the nervous system's behavioral features to the efficiency with which they might represent percepts and cognitive processes, so as to explain *why* they behave this way.

In this section, we briefly describe neurons and synapses from morphological, mechanistic, and descriptive viewpoints. Interpretive concerns will be addressed in Section 2.1 via the therein introduced concept of functional attributes.

From a simplified mechanistic viewpoint [3], dendrites capture input from and axons carry output to other neurons. A difference between intra/extracellular electric potential is kept by a complex interplay between numerous passive and active cell membrane structures (ion channels and ion pumps, respectively) that exchange various types of ions (like Na^+, Ca^{2+}, K^+, or Cl^-) between the intra/extracellular mediums[1]. In a simplified view[1], signaling from other cells, called pre-synaptic neurons, causes changes in some of these elements (like the opening and closing of some types of ion channels), leading to a temporary increase in membrane potential (the so-called action potential or spike) to be propagated down the axon, causing subsequent signaling of post-synaptic neurons.

The earliest precise mechanistic model of membrane potential is credited to HODGKIN and HUXLEY [33][1], who described the membrane potential by modeling various elements of the nervous cell using electric circuit elements, like capacitors, conductance indices, batteries, and current sources. This modeling yields a series of non-linear differential equations that predict the membrane's electric potential fluctuation over time as a function of current injection into the dendrites, effectively providing an account of spiking behavior, as illustrated in Figure 1.2. The Hodgkin-Huxley neuron is an example of the integrate-and-fire family of models, and even though other model families have been proposed to better account for phenomena like adaptation, bursting, and inhibitory rebounds, integrate-and-fire models are capable of predicting spike times as functions of input current very precisely [34]. Finally, we note that the perceptron, a simplified computational neuron model created by Frank Rosenblatt [35] after the Hodgkin-Huxley model, given slight modifications, forms the basis of artificial neural networks (ANNs) to this day.

Despite all the intricacies of mechanistic modeling, it is widely accepted that the precise shape of membrane potential fluctuations is not important for communication between neurons, and that the fundamental unit of communication is the spike itself [1, 3, 34]. Therefore, it is common to describe a neuron's activity by a spike train, *i.e.,* a sequence of $P \geq 0$ spike times: $(t_1, \ldots, t_P)$. The spike train may also be represented in functional form as a sum of Dirac delta functions

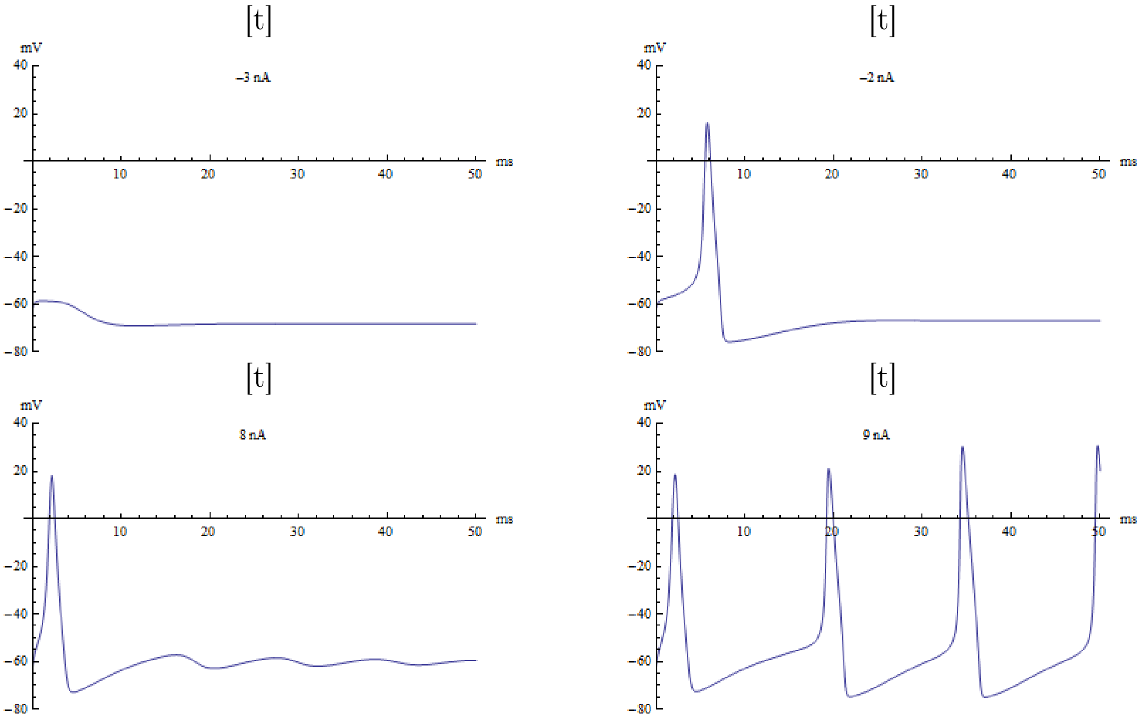
$$\rho(t) = \sum_{k=1}^{P} \delta(t - t_k) \tag{1.1}$$

or a as sum of Kronecker deltas

$$\rho(t) = \sum_{k=1}^{P} \delta[t - t_k] \tag{1.2}$$

[1] Who later received the Nobel prize for it, in 1963.

Figure 1.2: Simulating the Hodgkin-Huxley model with different input currents. As the injected current is increased, the system goes from equilibrium to a single spike, to a regime of cyclical spiking. Source: `https://commons.wikimedia.org/wiki/File:Hodgkins_Huxley_Plot.gif`, created by Alexander J. White, published under Creative Commons Attribution-Share Alike 3.0 Unported.



Mechanistic models may be able to predict membrane potential and spike times rather accurately *in vitro*, *i.e.,* with an isolated cell and controlled input current, however neurons have thousands of synapses and axons may span significant lengths of the brain, leading to intricate connection patterns and complex dynamics. As a result, *in situ* observations do not record the exact same spike train twice, given identical presentations of visual, auditory, olfactory, or tactile stimuli [34]. Therefore, from a descriptive perspective, it is more convenient to model neurons as stochastic event generators that fire spikes following a variable firing rate $r(t)$.

Firing rates are commonly estimated by taking the average of spike trains recorded during multiple trials, *i.e.,* equal-duration stimulus presentation sessions with all other factors kept constant. Given N trials with spike trains $\rho^{(j)}(t)$, $1 \leq j \leq N$, each with $P^{(j)}$ spikes, the resulting function is

$$\bar{r}(t) = \langle \rho(t) \rangle = \frac{1}{N} \sum_{j=1}^{N} \rho^{(j)}(t) \tag{1.3}$$

where the $\langle \cdot \rangle$ notation is a shorthand for the average of $\cdot$ over trials. More commonly, the spike trains are convolved with a kernel $K(t)$ in order to obtain a smooth function, which yields

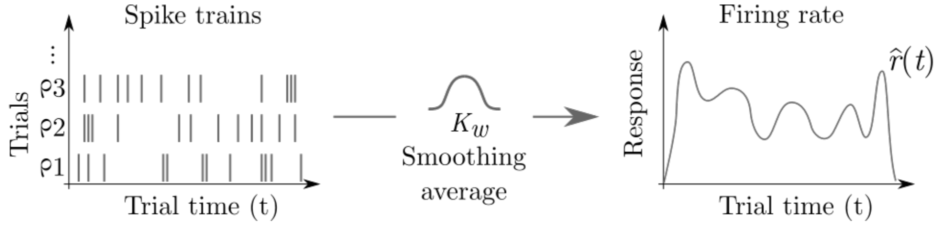
$$\bar{r}(t) = K(t) * \langle \rho(t) \rangle = \langle K(t) * S(t) \rangle = \left\langle \int_{-\infty}^{\infty} K(\tau) * \rho(t-\tau)\, d\tau \right\rangle \tag{1.4}$$

A Gaussian kernel $K_\sigma(t) = \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{t^2}{2\sigma^2}$ is frequently chosen for that purpose. Also noteworthy, the convolution with a box kernel

$$K_w(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq w \\ 0 & \text{otherwise} \end{cases}$$

results in a staircase function known as peri-stimulus time histogram (PSTH) that is convenient for representing responses in a vectorized form. Figure 1.3 illustrates the conversion from a collection of spike trains to a firing rate function using a bell-shaped (Gaussian) curve.

Figure 1.3: Spike trains and firing rates. On the left, spike trains ρ_j, each consisting of a time-series of impulse responses, are recorded for several trials (repetitions of a stimulus sequence). On the right, the trial-averaged spike train $\langle \rho \rangle$ is convolved a Gaussian kernel K_w to yield a smooth firing rate function, $\hat{r}(t)$.



On some occasions, it may be convenient to model neuronal firing via inhomogeneous Poisson processes, using the estimated firing rate as a time-varying arrival rate. That is a simplification, since neurons have refractory periods that make them incompatible with Poisson processes [34]. That simplification is specially useful when simulating neurons or computing information-theoretic quantities over distributions [3][36],

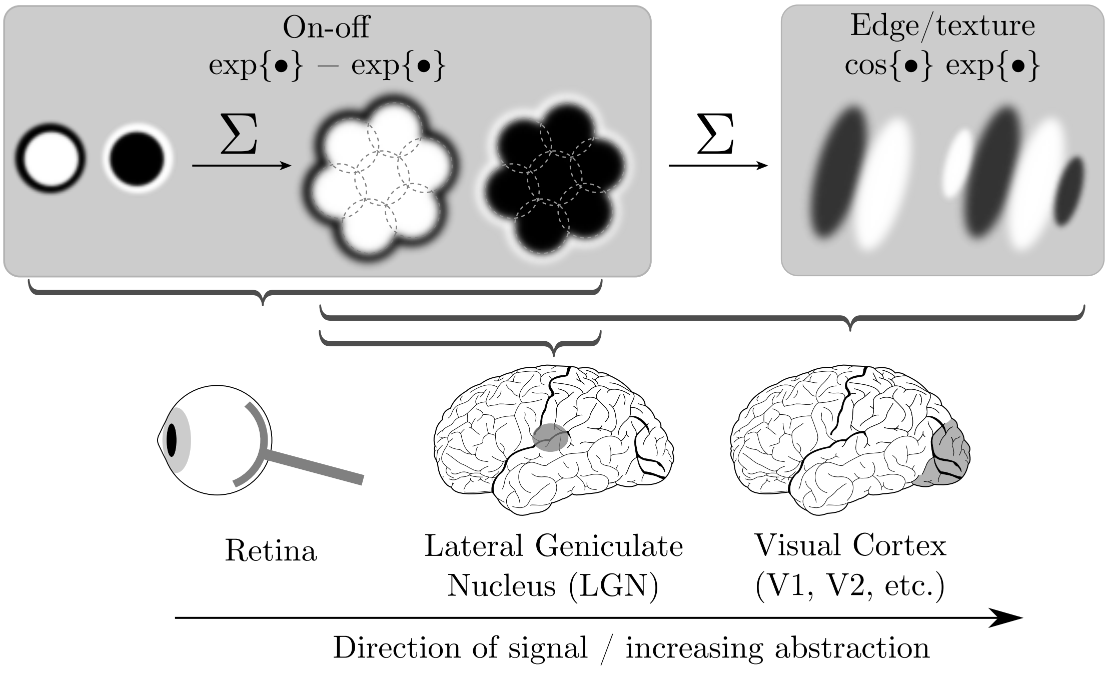
1.2 Visual cortex

One of the most studied brain subsystems is the mammalian visual cortex, widely credited with some of the most primitive stages of vision [37]. Neurons in the sensory cortex (where visual, auditory, and tactile percepts are processed) are specialized and tend to respond (that is, to spike) only to narrow ranges of stimuli. The core concept for characterizing these neurons' response properties is the receptive field (RF): it represents the spatiotemporal domain in the corresponding sensory organ where stimulation either excites or inhibits the cell [1]. In other words, it is the set of stimulus that either drives or silences the neuron. Mathematically, RFs are modeled as linear filters that predict the responses (that is, mean firing rates) of neurons to given stimuli [3]. In the case of visual cortex neurons, RFs may assume the shape of spots, edges, simple textures, be sensitive to colors, speed, direction, *etc.* The following subsection presents one possible mathematical formulation of RFs in the early visual system.

1.2.1 Early visual system: aggregation of receptive fields

The *early visual system* features the progressive aggregation of visual information, starting with simple photoreceptors in the retina all the way up to rudimentary texture-sensitive cells in the visual cortex [38]. Three anatomical elements commonly studied in this context are: the eye cells known as retinal ganglion cells (RGCs); the dual thalamic, relay regions known as lateral geniculate nucleus (LGN); and the posterior cortical region, known as visual cortex that is subdivided into various areas, like V1, V2 and others. The specialized cell types in each of these stages vision encode patterns light into progressively more abstract RFs, as indicated in Figure 1.4.

Figure 1.4: Visual system components and their receptive fields. RFs become increasingly bigger and more tessellated as neural signals traverse the early visual system components into higher order areas, as suggested by the summation symbol, $\sum$.


On-off
$\exp\{\bullet\} - \exp\{\bullet\}$
$\sum$
$\sum$
Edge/texture
$\cos\{\bullet\}\exp\{\bullet\}$
Retina
Lateral Geniculate Nucleus (LGN)
Visual Cortex (V1, V2, etc.)
Direction of signal / increasing abstraction

RGCs are usually associated with simple on-center or off-center RFs — that is, light spots surrounded by a dark background, or dark spots surrounded by a light background, respectively — and these cells are commonly sensitive to moving, rather than static light patterns. They forward their signals through the optic nerve to the next stop of the visual system, the LGN, in which neurons receive input from multiple On-Off and Off-On cells, effectively summing their responses into larger RFs. The summation of inputs is carried on in the earliest areas of the visual cortex, namely V1 and V2, resulting in elongated direction-selective RFs resembling edge detectors from image processing theory. All of these steps are illustrated in Figure 1.4. The identification of On-Off or Off-On direction-selective RGCs is usually credited to BARLOW *et al.* [39], as is the discovery of orientation-selective RFs in the cat's striate cortex credited to Hubel and Wiesel [40, 41].

1.2.2 Modeling receptive fields

Mathematically, RFs can be represented using several functions, but Gabor functions are frequently chosen because they are flexible enough to describe most RF shapes found in the early visual system [3]. Gabor functions are band-pass spatial filters corresponding to the product of a Gaussian kernel and a sinusoid, and have applications in image processing, texture synthesis and analysis, and image compression. A simplified Gabor function — aligned with the x axis and centered at the origin — can be described as

$$D(x,y) = \frac{C}{2\pi\sigma_x\sigma_y} \exp -\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2} \cos[2\pi(kx-\varphi)] \tag{1.5}$$

where σ_x, σ_y indicate the RF horizontal/vertical sizes, in degrees, k indicates its preferred spatial frequency, in cycles per degree, φ indicates its preferred spatial phase, in cycles, and C is a constant that represents the response strength, in Hertz. Note that Cartesian x, y coordinates around the fovea (*i.e.,* the center of the visual field) are usually represented in degrees by relying on the approximation that, for small θ, $\sin\theta \approx \theta$, such that polar/azimutal coordinates may be replaced with Cartesian ones. Figures 1.5 and 1.6 illustrate how both On-Off and texture-like RFs, respectively, may be described using these functions.

As previously mentioned, mathematical RF definitions may be interpreted as linear filters for predicting a cell's mean response given a visual stimulus. In other words, if the spatio-temporal function $I_{\mathbf{s}}(x,y,t) \mapsto \{0,1\}$ describes the normalized luminance in the visual field, then the expected mean response at time t would be given by

$$r(t) = r_0 + \iint D(x,y)\, I_s(x,y,t)\; dx\, dy \tag{1.6}$$

where r_0 is a basal firing rate that accounts for spontaneous firing. More complex RFs may be described by incorporating a temporal factor to account for sensitivity

Figure 1.5: An On-Off receptive field modeled as a Gabor function. Parameters: $k = 0.1$ rad°, $\varphi = 0.5$ rad, $\sigma_x = \sigma_y = 3°$, and $C = 2\pi\sigma_x\sigma_y$ Hz.

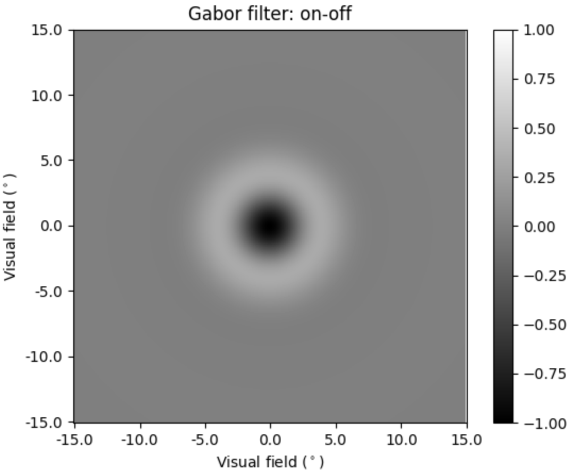
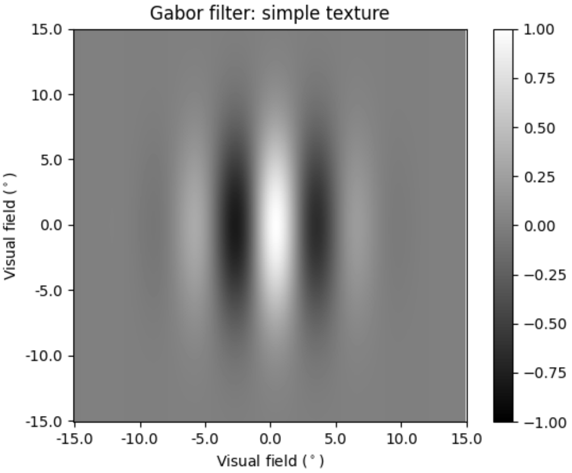


Figure 1.6: A texture-like receptive field modeled as a Gabor function. Parameters: $k = 0.15$ rad°, $\varphi = 0.08$ rad, and $\sigma_x = 4°$, $\sigma_y = 15°$, and $C = 2\pi\sigma_x\sigma_y$ Hz.



to motion and to motion in particular directions, and even by considering alternate colors, rather than simply luminance. In the next subsection, we will describe some stimulus types used to drive neurons in V2 without formally expanding their functional definitions or delving into why these stimuli are optimal for characterizing such RFs, keeping our presentation at a higher level. For a deeper discussion of those topics, see DAYAN and ABBOT [3].

A classical theory known as dual visual pathway, proposed by MISHKIN *et al.* [42, 43] and comprehensively reviewed by SERRE [37], pictured the later stages of processing past V1 through two separate *streams* or *pathways*:

Ventral stream Where the aggregation of earlier RFs eventually culminates in position-invariant, object identity representations ("what") in the temporal areas.

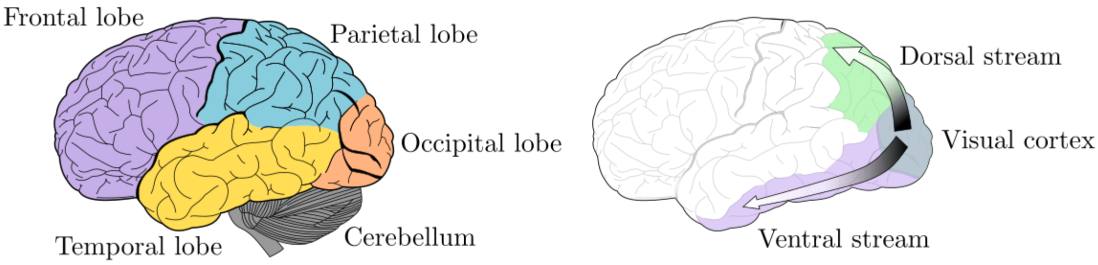
Dorsal stream Where, in a similar fashion, the aggregation of spatial information results in the representation of spatial relationships between objects, regardless of their identity ("where"), in the parietal areas.

They were mainly motivated by the apparently well-defined types of cognitive impairments suffered by accident victims and test subjects that had parts of one of these pathways severed or removed, while keeping the other functionality intact (*e.g.,* being able to navigate places and manipulate objects, but not to recognize their identity). However, these models have been considerably jeopardized, among other factors, due to:

- Numerous anatomical evidence of recurrent (feedback) connections both along and between these supposedly isolated and serial pathways, and corresponding functional studies showing sensitivity to spatial configuration in the temporal areas and to object identity in the parietal areas [44, 45],
- Existence of allegedly higher-level RFs as early as in the retina and LGN [46] and controversies on the preferential stimuli in post-V1 areas [47], and
- The phenomenon of neural modulation, in which cognitive states and top-down neurons regulate the RFs of putative early neurons [48].

Nonetheless, this model has laid the groundwork for future studies of the visual system, and the ventral stream, in particular, inspired the development of deep neural networks (DNNs) for object recognition that are now capable of very powerful object recognition [49, 50]. Figure 1.7 illustrates the aforementioned brain lobes and the flow of information along these streams.

Figure 1.7: The four lobes of the brain and the dual visual pathway. The main anatomical divisions of the brain are show on the left and the ventral and dorsal pathways are highlighted on the right. Adapted from SELKET [51].



1.2.3 Stimulus, response, and functional attributes

In order to characterize RFs in the sensory cortex, researchers present controlled stimuli to test subjects, record their neuronal responses (using a multitude of technologies), then correlate the latter with the former to obtain a *functional* description of the observed neurons, *i.e.,* a mathematical description of how they respond to given inputs. For instance, when studying the visual cortex, one may record spike trains $\left\{\rho^{(1)}, \ldots, \rho^{(N)}\right\}$ from a neuron while visual patterns are displayed on a screen, then use reverse correlation methods (*i.e.,* causal analysis) [3] to estimate the corresponding Gabor functions' parameters that more likely explain the observed responses. Those methods are based on the premise that the cell's mean firing rates can be predicted by linear-filtering the spatiotemporal function that describes the luminosity in the visual field, $I_{\mathbf{s}}(x, y, t)$, with the RFs impulse function, as described previously in Equation 1.6.

From this point, we will generically refer to any function $f(\rho^{(1)}, \ldots, \rho^{(N)})$ of the recorded responses as a functional attribute. Examples include the estimated firing rate, defined in Equation 1.4, and the estimated RF's preferred spatial frequency (see Equation 1.5). Most commonly, these functions depend only on the trial-averaged spike train or firing rate, rather than on trial-specific responses, which tend to be noisy. Depending on which parts of the visual cortex is studied, there will be more indicated functional attributes appropriate stimulus types [52].

1.3 Extra-cellular electrophysiology

A typical way of studying the visual cortex is to record neuronal responses to controlled visual stimuli using extracellular electrodes, an experimental framework and field of study known as ecephys, short for *extracellular electrophysiology*. Each electrode, inserted directly into the brain's tissue, but outside of nervous cells, captures the sharp fluctuations of electric potential when nearby neurons fire action potentials, more commonly dubbed as spikess. These traces are abstracted into time-series of discrete spiking events, averaged and smoothed into a firing rate, and the aforementioned technique of reverse correlation is applied in conjunction with statistical tests to determine if the neuron is indeed sensitive to the observed stimulus, or just fluctuates randomly [3]. Ecephys experiments like these have been conducted since the first half of the XX century, leading to the discovery of RFs in the retina [39, 53, 54] and visual cortex [40, 41].

Neuroscience-related research corresponds to 19% of research conducted in non-human primates, and besides neurophysiology, it contributes to advances in disease comprehension and treatment, such as Alzheimer's disease, Parkinson's disease, and

NeuroAIDS, for instance [55]. That is not to say those experiments are uncontroversial, after all they require invasive surgical procedures and frequently result in animal sacrifice [47, 56–61]. As such, carrying out such experiments requires approval by animal ethics regulation agencies (in most countries), besides careful preparation (*e.g.,* opening a shallow pit into the skull), and execution (*e.g.,* administering anesthetics) procedures. The advancement of non-invasive brain-scanning technologies, such as electroencephalography (EEG) and MRI, enable to study the human brain's connections and activity patterns in awake human subjects [24] in a non-invasive manner and, in doing so, they contribute to unveil the human brain's connectome [62–64], with a strong intersection with network science [65, 66] and ANNs inspired by the visual system [49]. However, the more traditional and invasive method of ecephys recordings is still a valuable source of information that captures neural activity at a unique spatiotemporal range (see fig|replace('label', 'devices')), yet unmatched by alternative techniques like *calcium imaging* [67] and optogenetics [27], therefore it remains a fundamental tool in solving the many existing contradictions and open questions about how visual processing is carried out in the mammalian, and most notably, the human brain [47, 48].

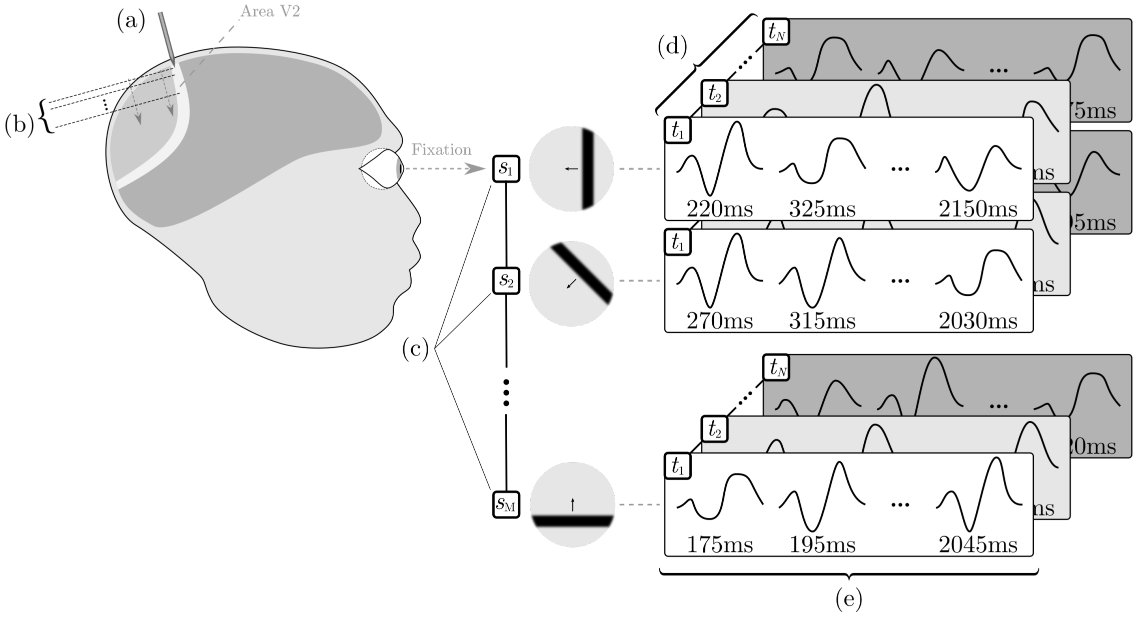
1.3.1 Data acquisition

The data acquisition process performed by the LFCOG's domain experts[68] employs a regular grid of electrodes, or multi-electrode array (MEA), inserted into a test subject's visual cortex to record extra-cellular fluctuation of electric potential, resulting from the activity of multiple cells. These electrodes record spike trains simultaneous to presentations of visual stimuli like the ones we described in Section 2.1.1. Recording sessions, involving trial-repeated presentations of all stimuli, are performed repeatedly by placing the MEA at multiple cortical positions (*i.e.,* increasing depths, possibly on both hemispheres). The smallest unit of information recorded by an electrode is that of a single spike, for which we know when it was detected, and its waveform, *i.e.,* a narrow snapshot of the electrical activity around the time of spiking. In practice, waveforms work like cell signature indicators that may be used to separate the multi-unit activity (that is, the summative response of multiple cells around the recording electrode's tip) back into individual putative cells, or single-units. Figure 1.8 illustrates the acquisition process.

Eventually, the neural tissue where the MEA was inserted is analyzed in a laboratory, being treated with preserving and reactant solutions, flattened, sliced, and analyzed in a microscope, allowing the identification of certain histo-chemical properties. In some works by the LFCOG, for instance [68, 69], the concentration of the cytochrome oxidase (CytOx) enzyme, which varies in bands that run orthogonal to the cortical surface, was identified and annotated for each electrode positioning, or recording site, and related to the response preferences of the corresponding recording sites. That is what we refer throughout this thesis as physiological attributes: complementary, qualitative information about a recording site that may present correlations of interest.

The recording process naturally involves the usage of specialized recording systems (hardware and software), like the Omniplex®; Neural Recording Data Acquisition System, which perform numerous tasks from signal amplification to analog-to-digital (AD) conversion, each with its own parameters. In this thesis, we will only be concerned with the steps that follow AD conversion, introduced in the next section.

Figure 1.8: Overview of extracellular electrophysiological data acquisition. The test subject is first anesthetized and its sight is fixated onto a screen with a calibrated illumination and coordinate system. An electrode matrix (a) is inserted into area V2 of the subject's brain at increasing depths (b), as multiple recording sessions are performed. During a session, each of the visual stimuli $\{s_1, ..., s_M\}$ (c) that comprise the previously designed stimulus protocol is presented on the screen in shuffled order, and this process is repeated a certain number of times, N. Each repetition t_j of a stimulus s_i is called a trial (d). The recording equipment (hardware and software components) preprocesses the electrical signals picked up by the electrodes generating, for each combination of electrode, stimulus, and trial, a variable-length time-series of responses (e) composed of spike times and spike waveforms (fixed-length snapshots of electrical activity around the time of spiking).



1.4 Challenges

The ecephys experimental framework presents a particular set of challenges to the observation of true neuronal behavior due to the inevitable interference caused by data acquisition and processing. It also has ethical implications, as the historical decisiveness of animal experimentation to the discovery of drugs and treatments is not universally accepted to outweigh risks to animal well-fare. Mice can be used as an alternative to primates in optogenetics-based studies with tiny sensors that allow long-term recording, but their anatomy is not as close to the human [70] and recording apparatus need to follow a compromise between animal well-fare and recording duration and extensiveness, which motivates the development of better equipment [71] and even the usage of larger species [70]. Long-term, less invasive recording is also possible [72] although it is known that these recordings tend to decrease in performance over time due to tissue damage [73]. Therefore, acknowledging the fact that ecephys experiments with primates are still relevant for advancing the field, one of core the motivations for this thesis is to increase the utilization of data gathered in past and future studies, in the hopes that they may be better justified, from a utilitarian perspective.

It is paramount that signals from a recording site do not contain actual individual responses but rather confounded responses (a summative signal) of multiple individuals. As we explained before, it is possible to separate these signals to some extent through spike sorting but that comes at the cost of increased data volume since one summative observation becomes multiple putative individual observations. Available methods[74] are parametric, therefore multiple solutions exist which further aggravates the problem of volume increase, even more so if we account for parameters affecting the spike detection phase. Those steps are usually only strictly necessary for carrying out posterior analyses, so the reproducibility of studies may be affected by simplified treatment and the comprehension of parameter effects may become hard to grasp. For instance, PERES *et al.* [68], who describe the data acquisition procedures used for collecting the V206 Dataset, only ever briefly mentions spike sorting, like so:

> Offline cell sorting was performed using the Plexon Offline Sorter software (Plexon Inc., Dallas, TX, RRID:SCR_000012). We applied Principal Component Analysis (PCA) to reduce the number of dimensions of our correlated variables. Subsequently, spike waveforms were clustered using the k-means algorithm (Webb, 2002). Finally, only the well isolated units were selected manually for further analysis.

There are multiple, hard-to-quantify sources of incidental noise (*e.g.,* progressive tissue damage, electrode tip wear, AC/DC line noise) that contaminate the signals and their conditional impact on the probability of approving summative or individual observations is unknown. Finally, many signals do not show clear selectivity to presented stimuli, so it's hard to distinguish if they contain useful yet unexpected information or just noise. We believe these traits make this problem domain ideal for a visual analysis treatment combining elements of exploration, interactivity, and semi-automated processing [75–77].

Not only in ecephys but all over systems neuroscience, it seems that a major challenge is how to take advantage of the huge amount of data amassed in modern experiments [78]. In this scenario, data visualization arises as a promising direction for making sense out of the bulky, high-dimensional, and often noisy data recorded in different spatiotemporal scales. Nonetheless, visualization always involves abstraction, so that choosing a proper mapping of quantities to visual features that strikes the right balance between ease of understanding and richness of information, depends on knowing who is the target user (novice, or expert), what is the objective (to communicate key results, or to aid in exploration and insight discovery), and on following established principles [24].

The vocabulary of *big data*[79] defines big data by the presence of one or more of the five V's, that is, five concerns that impact data usage: volume, variety, veracity, velocity and value. In a similar fashion, we identified the following concerns in the ecephys field and, in particular, V2 Dataset:

Volume The amount of data to analyze is vast. That is a byproduct of many factors, like the elevated AD sampling rate, the number of electrodes and recording sites, or the combinatorial explosion of the space of preprocessing parameters.

Uncertainty Since many combinations of preprocessing parameters may be employed (*e.g.,* low-pass filters, spike detection thresholds, spike sorting features), one cannot be absolutely sure about the best choice of parameters for analyzing a given population. Could a set of parameters yield results closer to the true, underlying behavior, and still present fewer evidence for RFs according to the analysts's expectations? This concern is similar, in spirit, to *veracity.*

Sparsity The possibility of many recording sites not containing plausible RFs under any combination of parameters. For instance, PERES *et al.* [68] identified only 190 stimulus-driven neurons out of 721 single units in a total of 3 test subjects and various MEA insertion depths. This concern is similar, in spirit, to *value.*

One of the points addressed by this thesis, is that when a considerably large number of neurons is studied, applying a manual analysis workflow, which involves inspecting 2 and 3D plots of functional attributes, required for assessing quality and reducing uncertainty, on a per-cell basis is not scalable (that is, it hits the volume and sparsity walls). In Chapter 3, we will describe a collaborative study that began as an informal design study about facilitating the exploration of the V2 Dataset. That later evolved into a user study to investigate the level of agreement of a few domain experts when using *viz* to clean ecephys data, in an attempt to investigate whether it could be a viable approach to scale their current practices. The quantitative results of that study pertaining user behavior and decision similarity will be presented and discussed in Chapter 4, together with evidence for the viability of a semi-automated, machine learning-supported workflow for neuronal observation winnowing.

1.5 Summary of contributions

- We analyzed the domain-specific (ecephys studies of the primate visual cortex) work practices using the design study methodology and provided an account of it using the concept of user goals (Section 3.3)

- We assembled and perfected a number of encodings of neuronal behavior attributes that are already familiar to domain experts into a terse layout that allows multiple observations to be compared and judged (Section 3.4), showing that these representations lead to identification of relevant observation categories (Section 3.4.1)

- We developed (Section 3.4.2) and validated a new encoding for neuronal responses to sinusoidal drifting gratings stimuli

- We developed an online tool prototype for allowing domain experts to compare and judge multiple neuronal observations (Section 3.5)

- We studied expert decision dynamics in a specific dataset of V2 recordings, and discovered they often focus on conflicting evidence on the face of noise and uncertainty — an extensive account of decision behavior is provided in Sections 4.1-4.3 and a selection of debatable cases is presented in Section 4.5.2

- We showed that modifying preprocessing parameters can yield a great variety of results, with different approval statistics by users and therefore severe implications for how cognitive work is traditionally approached (Section 4.1.3)

- Finally, we showed that user decisions are amenable to approximation by machine learning, which may greatly enhance their performance by allowing neuronal observations to be ranked automatically and assigned a score that indicates if they need manual review (Section 4.4).

Chapter 2

Data and methodologies

This chapter introduces the raw materials, methods, and methodologies that were fundamental for developing this research. It provides necessary context for understanding the remaining of this thesis, and we advise the reader to at least skim through all sections, whilst placing greater attention on the areas that are less familiar to him/her. For instance, Ecephys data and algorithms (that is, our raw materials and basic processing methods) are covered in detail in Section 2.1 and should be easily comprehensible for readers from the medical sciences but less so for engineers and computer scientists, who will find its content helpful for successfully grasping the data semantics throughout the remainder of the thesis. In the following, Section 2.2 reviews design study literature, which is relevant for understanding the work methodology that we adopted towards our collaborators, and is fundamental for understanding Chapter 3 and the qualitative discussions in Chapter 4 and Chapter 5. Finally, Section 2.3 provides a brief summary of important ML concepts and methodologies that were foundational for our investigation of predictability of user decisions (discussed in Section 4.4) and the corresponding discussion in Chapter 5.

2.1 The dataset and its preparation

In this section, we describe the general characteristics of the dataset that are important for understanding the data abstraction discussion in Section 3.2. This includes describing how we processed the original data, which is an important information for readers from the problem domain, detailing the visual stimulus types used in the original experiments, and defining the fundamental functional attributes that relate to them. Nevertheless, for raw data acquisition procedures, which are completely outside the scope of this thesis, we refer the reader to PERES *et al.* [68] (subject *V206*). We will henceforth refer to this dataset as the V206 Dataset or simply *the dataset*. A complete listing (with notation and definitions) of all variables in the dataset is presented in Appendix A.

2.1.1 Visual stimulus types

A visual stimulus $\mathbf{s}$ may be characterized by its spatiotemporal *stimulus intensity function*, $I_{\mathbf{s}}(x, y, t)$, which associates a normalized luminosity level in the $[0, 1]$ range to each point (x, y) of the visual field, at time t. We indicate an individual stimulus by a vector symbol, like $\mathbf{s}$, to indicate the parametrical nature of stimuli. When discussing several stimuli $\{\mathbf{s}_1, \mathbf{s}_2, \ldots\}$ with a common intensity function that also varies with $\mathbf{s}_i$, we say that they constitute a *stimulus type* or *stimulus family.* Let us introduce two types of visual stimuli employed by the LFCOG during acquisition of the dataset.

Moving bars

These stimuli consist in light or dark bars moving at constant speed against a gray background, crossing a fovea-centered region of the visual field from one side to the other, as illustrated in Figure 2.1b, and is useful for studying both On-Off and texture-like RFs. The mathematical formulation of the stimulus intensity function, $I(x, y, t)$, for such a stimulus is

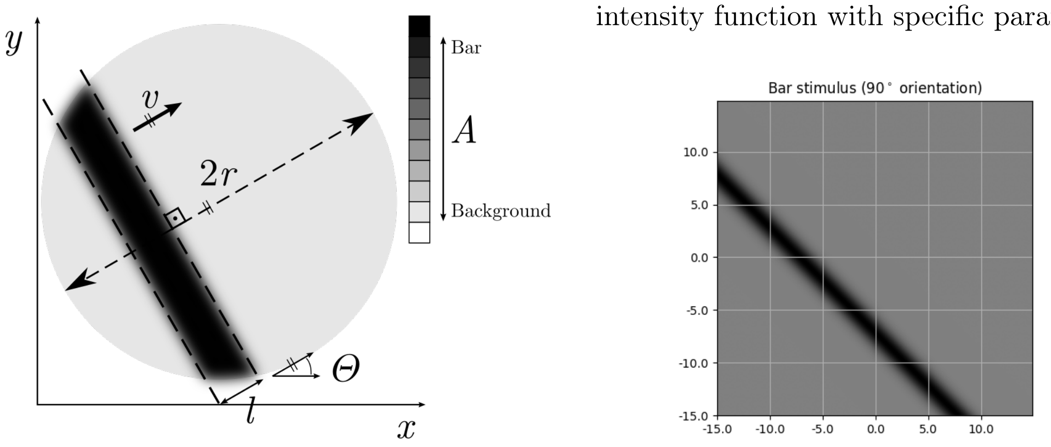
$$I(x, y, t) = \frac{1}{2} + \frac{1}{2}A\left\{1 - 2\exp\left[-\left(\frac{3(x\cos\Theta + y\sin\Theta + r - vt)}{l}\right)^2\right]\right\} \tag{2.1}$$

where $A \in [-1, 0) \cup (0, 1]$ represents the contrast between the bar and the background, Θ is the direction of bar movement (orientation is perpendicular) in radians, l is the bar width in degrees[1], $2r$ is the excursion — the distance covered by the bar — in degrees and v is the bar's constant speed in °/s. The stimulus duration, in seconds, is given by $T = \frac{2r}{v}$. Note that negative A values result in a dark bar and positive values result in a light bar.

Figure 2.1 provides an schematics view of the stimulus and an example with specific parameters. As a consequence of Equation 1.6 and Equation 2.1, it will elicit stronger responses when it crosses the boundaries of an On-Off RF, or the transitions between excitatory and inhibitory regions of texture-like RFs, when in the same orientation as their preferential one. Usually, when probing a neuron's RF, this stimulus will be parametrized with multiple, equally-spaced Θ values in the range $[0, 2\pi)$. Therefore, this stimulus type is useful for locating RF centers, estimating their sizes, preferential direction and orientation, and even for estimating their response latencies. From the description above, we may finally describe a moving bars stimulus $\mathbf{s}$ as a vector $\mathbf{s} = (r, A, \Theta, l, v)^\intercal$.

[1] At $\frac{1}{2}l$ away from the bar center in the direction of movement, the equivalent to 1.5σ away from a standard Gaussian's center, the bar contrast with the background is roughly $\frac{A}{10}$.

Figure 2.1: Dark moving bar stimulus schematics and intensity function. Coordinate axes represent the visual field around the fovea. Parameters used on (b): $A = -0.5$, $\Theta = 45°$, $l = 3°$, $v = 10°/\mathrm{s}$, $2r = 30°$, at time $t = 1^{\mathrm{s}}$ (a while before it crosses the center of the fovea).



(a) Moving bars stimulus schematics

(b) Spatio-temporal moving bars stimulus intensity function with specific parameter.

Drifting sinusoidal gratings

A drifting sinusoidal gratings stimulus (we call it *gratings*, for short) consists in a series of alternated light and dark stripes (sinusoidal gratings) that move perpendicular to their orientation at constant speed (drifting), as illustrated in Figure 2.2b, and is useful for studying both On-Off and texture-like RFs. The mathematical formulation of the stimulus intensity function, $I(x, y, t)$, for such a gratings stimulus is

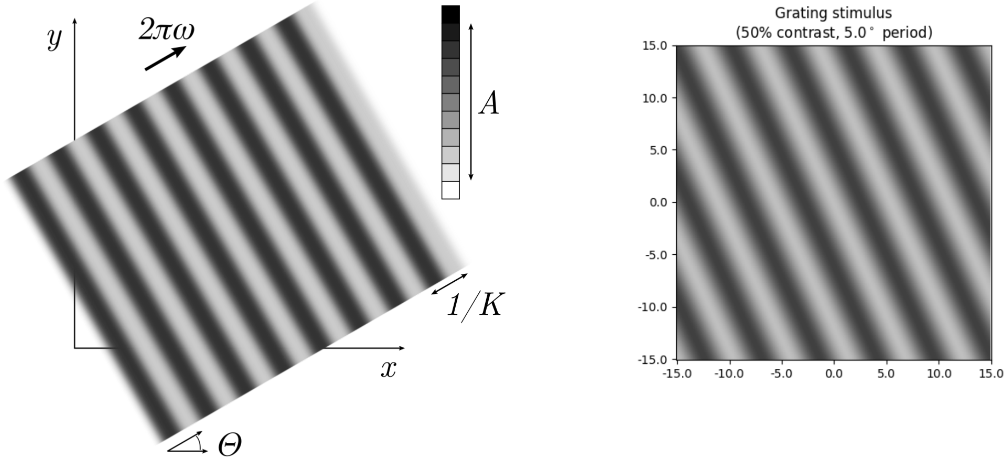
$$I(x, y, t) = \frac{1}{2} + \frac{1}{2} A \cos\left[2K\pi(x \cos\Theta + y \sin\Theta) - 2\omega\pi\, t\right] \tag{2.2}$$

where $A \in (0, 1]$ represents the contrast between the dark and light stripes, Θ is the direction of drifting (stripe orientation is perpendicular) in radians, K is the spatial frequency in radians per degree, and ω is the drifting speed in rad/s. Unlike the stimulus type, the excursion is not necessary for defining the stimulus intensity function. Figure 2.2 provides an schematics view of the stimulus and an example with specific parameters. As a consequence of Equations Equation 1.6 and Equation 2.2, it will elicit stronger responses when the stripes' orientation match the texture-like RFs own orientation, or when the RF has a omnidirectional On-Off shape. Since the stripes cover the entire visual field, a preferential stimulus will cause the neuron to spike repeatedly rather than in a located fashion, as a stimuli. When mapping a neuron's RF, this stimulus will be typically parametrized with multiple, equally-spaced Θ values in the range $[0, 2\pi)$, and a variety of values for the other parameters. Therefore, this stimulus type is useful for estimating preferential direction and orientation, spatial frequency, speed, and contrast of general texture-like RFs.

Figure 2.2: Drifting gratings stimulus schematics and intensity function. Coordinate axes represent the visual field around the fovea. Parameters used in (b): $A = 0.5$, $\Theta = 25°$, and $K = 1.5\text{rad}/°$ (and ω value not relevant).

(a) Sinusoidal drifting gratings stimulus schematics

(b) Spatio-temporal sinusoidal drifting gratings stimulus intensity function with specific parameters.



In the remainder of the text, we will abbreviate this stimulus type as simply gratings.

Although the definition of the stimulus intensity function given by Equation 2.2 models luminosity, making it suitable for black/white stripes, it can be extended for stripes of alternated colors, $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$ — where $\mathcal{C}$ is some colorspace — by interpreting it as a blending factor between $\mathbf{c}_1$ and $\mathbf{c}_2$. In that case, the stimulus luminosity on each point of the image plane depends on both the contrast and the chosen colors. With that in mind, we may finally describe a drifting sinusoidal gratings stimulus $\mathbf{s}$ as a vector $\mathbf{s} = (A, \Theta, K, \omega, \mathbf{c}_1, \mathbf{c}_2)^\intercal$.

Moving bars stimulus set details

The moving bars stimulus set covered a fovea-centered, 30°-wide circular area of the visual field in each eye. The 30° × 0.3° bar drifted at 10°/s (thus taking 3 s to swipe the entire screen) on 8 equally spaced, cardinal drifting directions, $\Theta \in \{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$. No inter-trial time was inserted and each stimulus was repeated over 10 trials, in a total of 80 trials, and approximately 4 minutes per recording session.

Gratings stimulus set details

The gratings stimulus set also covered a 30° area of the visual field. Stimulus parameters (direction Θ, contrast A, spatial frequency K, speed ω, and colors $\mathbf{c}$) assumed values according to variation groups, in which Θ plus up one other parameters varied while the others remained constant, except for one variation group where spatial frequency and drifting speed varied jointly. In all variation groups, Θ varies through 8 equally-spaced cardinal directions, that is

$$\Theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}$$

These variation groups, which include stimulus sets for studying sensitivity to all parameters. are described in Table 2.1,

Table 2.1: Secondary stimulus parameters values and variation groups. The values of each parameter — contrast, spatial frequency (SF), speed (Spd) and colors — that vary with motion direction while other secondary parameters remain constant are surrounded by boxes. The 3 rows at the bottom include tuples where spatial frequency and speed vary together. Colors are indicated as B/W (black and white) and Bl/Gn (blue and green).

Contrast	**SF** (1/◦)	**Spd. (Hz)**	**Colors**
0.06	0.5	3.0	B/W
0.12	0.5	3.0	B/W
0.5	0.5	3.0	B/W
1.0	0.5	3.0	B/W
1.0	1.0	3.0	B/W
1.0	2.0	3.0	B/W
1.0	1.0	1.0	B/W
1.0	1.0	3.0	Bl/Gn
1.0	1.0	10.0	B/W
1.0	1.0	30	B/W
1.0	0.5	1.5	B/W
1.0	2.0	6.0	B/W
1.0	4.0	12	B/W

Other than direction, the other parameters appeared in a total of 13 combinations, and one "neutral stimulus" consisting of a gray screen without any gratings was included, so a total of 105 stimuli were presented. Each trial lasted for 2 s, and 250 ms of 50% gray screen was inserted before and after stimulus presentation. With 10 trials per stimulus, each recording session lasted approximately 44 minutes.

2.1.2 Spike detection and sorting

Let us report general procedures performed in the area, which we applied to the raw data shared by our collaborators. Processing digitally sampled extracellular signals starts with the application of digital signal processing (DSP) techniques for separating recorded signals into multiple frequency bands that reflect different neuronal dynamics before spike detection can take place [74, 80]. Usually, the input raw signal, also known as wideband (WB), is first low-pass filtered to remove the local field potential (LFP) — a low frequency electric potential fluctuation that appears where cell bodies and axons are partially aligned, creating a dipole in the extracellular medium [81] — from the high-frequency, jagged wave known as spike-continuous (SPKC) signal. The latter contains the mixed activity of closer neurons, and spikes manifest as fast, high-amplitude variations that are said to *ride* that wave. Finally, spikes are isolated by applying a voltage amplitude threshold to the SPKC signal, usually set to a multiple $k\,\hat{\sigma}$ of the signal's estimated standard deviation $\hat{\sigma}$, with $k \in [3, 5]$ being typical values. After spikes are aligned by some criterion (*e.g.,* the time of threshold crossing, or of lowest voltage before repolarization) waveforms are extracted by clipping the SPKC signal inside a fixed range around the spike times.

The next step in processing is separating the multi-unit spike trains into single-unit spike trains, in a process known as spike sorting. REY *et al.* [74] defines the spike sorting problem, reviews existing approaches and their limitations. In summary, each neuron's waveform has a distinct shape, determined by its morphology [82] and an amplitude that depends on the neuron's distance to the recording electrode, which makes it possible, in theory, to separate the confounded responses back into the original neurons based on their appearance. The term single-unit is due to the uncertainty in this process, after all the resulting spike trains may feature missing, excess, or swapped spikes, therefore the resulting entities are not properly neurons, but neuronal behavior units. As an alternative, single-units are sometimes called *putative neurons*, since they represent possible cell behavior. Nonetheless, we will most often refer to multi- and single-units as *summative* and *individual observation*, respectively, in an attempt to abstract domain jargon and emphasize the most important features (we are dealing with uncertain observations, and they can reflect individuals or groups).

Separating individual spikes is important because close-by neurons do not always behave similarly, therefore not separating spikes could not only lead to imprecise identification of RFs but also to completely miss on complex dynamics like inhibition, competition, modulation, and sparse firing. One of the earliest techniques, based on template matching was described by ABELES and GOLDSTEIN [83]. Sorting can usually be improved by the application of stereotrodes [84] or tetrodes [85]. Those may compensate for cases in which spikes elicited by different cells are too similar to be distinguished, or when bursting neurons produce spikes with decreasing amplitudes. Early techniques used spike amplitudes, then template matching to separate spikes, but these approaches are either inefficient at handling scenarios with ambiguous spike amplitudes, or require inconvenient manual intervention. Nonetheless, spikes are indeed sometimes sorted manually FISCELLA *et al.* [8].

Later methods started to apply dimensionality reduction algorithms like principal component analysis (PCA) or wavelets onto waveforms features as a way to remove noise. Then cluster the low-dimensional points using manual cluster cutting or a clustering algorithm like k-means. The latter plays out well with the assumption that spike variability is due to additive noise and stationary Gaussian background noise — but requires a way of guessing the proper number of clusters.

Spike detection

Spike detection was implemented for WB and SPKC signals according to the Omniplex®; user guide [80]. The data Plexon Inc. data files extension (PLX) data files related to the moving bars stimulus only contained filtered spikes — that is, discrete event signals — but those related to the gratings stimulus contained both WB and SPKC signals — that is, continuous voltage signals — that can be processed to obtain spike trains. In the following paragraphs, we will describe the spike detection procedures employed on those data sources. All source-code relative to this phase is available in the `vizpike.pipeline.detection` Python module and command line interface (CLI)[2].

We refer the reader to SMITH [86] for a comprehensive introduction to the field of DSP. In summary, the basic definitions we need are: $x(t)$ is an input, time-dependent signal, $h(t)$ is a time-dependent transfer function (it describes how fast and intensely the system's output reacts to its input), $y(t)$ is an output, time-dependent signal produced by filtering x with h, that is

$$y(t) = (h * x)(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)\,d\tau$$

We follow the convention of using capital letters to denote the Laplace transforms of the time-dependent functions, which constitute functions of the frequency domain, *i.e.,* $X(s)$, $H(s)$, and $Y(s)$, more specifically,

$$X(s) = \int_{0}^{T} x(t)e^{-st}\,dt$$

where T in this context indicates the signal duration. Filters are rather commonly specified in terms of their frequency-domain transfer function, $H(s)$. In that case, by the convolution theorem, the filtering operation corresponds to

[2] Available at `https://gitlab.com/lcg/neuro/v2/vizpike/-/tree/develop/src/vizpike/pipeline/detection` via the `__init__.py` (module) and `__main__.py` (CLI) files, respectively.

$$Y(s) = H(s)\,X(s)$$

Producing a SPKC signal from an input WB signal involves filtering out low frequencies that correspond to the LFP. We did so by applying bidirectional filtering with a high-pass Bessel filter of order 2 and a 300 Hz cutoff frequency. By applying the filter twice (back and forth), the corresponding filter has zero phase and double the order of the original filter (therefore, the applied filter had an effective order of 4). The n-th order Bessel filter is an infinite impulse response (IIR) filter characterized by the following transfer function (in frequency domain):

$$H_n(s) = \frac{1}{a_0}\left[\sum_{k=0}^{n} a_k \left(\frac{s}{\omega_0}\right)^k\right]^{-1}$$

where

$$a_k = \frac{(2n-k)!}{2^{n-k}k!(n-k)!}, \text{ with} k = 0, 1, \ldots, n$$

and ω_0 is the cutoff frequency. Therefore, the filtered SPKC, time-dependent signal $V_{\text{SPKC}}(t)$ is computed as $V_{\text{SPKC}}(t) = (h * y)(T - t)$, where $y(t) = (h * V_{\text{WB}})(t)$, where T represents the signal duration, in this context.

Detecting candidate spikes from a SPKC signal using the threshold method [74] is described next. Given a threshold as a multiple $k\sigma$ of the signal's empirical standard deviation σ, and the signal's empirical mean μ, all sample times t_i where the threshold $\mu - k\sigma$ is crossed, that is

$$V(t_i) \geq \mu - k\sigma \text{ and } V(t_i + 1/\nu) < \mu - k\sigma$$

are considered to be candidate spike times. Then, waveforms $\mathbf{w}_i$ are extracted by clipping the signal between 8 points before t_i and 23 points past t_i, so the resulting waveforms are the vectors

$$\mathbf{w}_i = [V(t_i - 8/\nu) \cdots V(t_i + 23\nu)]^\intercal$$

The candidate spikes are then filtered based on their waveforms, being removed if they fail to pass the following chain of conditions:

- At least one sample must happen past threshold crossing (that is equivalent to require repolarization past spiking),

- The signal must be eventually positive before threshold crossing,

- The signal must contain a global minimum (valley) and a positive global maximum (peak) that are not at the edges (first or last samples),

- No samples may overlap with a previous waveform.

Finally, waveforms are clipped from the input signal again after realigning the spike times t_i (originally obtained from threshold crossing) at the waveforms' respective global minima, *i.e.,*

$$t_i \leftarrow \operatorname*{argmin}_{-8 \leq k < 23} V(t_i + k/\nu)$$

This contrasts with the spike trains directly recorded in the PLX data files, which keep spikes aligned at threshold crossing. We did so because that is one option available in Ple [80, 87] that is expected to improve feature extraction during spike sorting, described below.

This processing was conducted with $1 \leq k \leq 6$, which covers the most typical range of values [74]. Nevertheless, we decided to only use $k \in \{3, 5\}$ during the user study, as reported in Section 3.5.

Spike sorting

Spikes were sorted using the same algorithm described as "PCA + K-means scan" in Ple [80, 87], which are variants of the general workflow described in [74]. This workflow consists in using PCA to project waveform features into a 2- or 3-dimensional space that retains most of the variance while reducing noise, then clustering those points by similarity (translated into Euclidean proximity in the projected subspace) using the K-means clustering algorithm.

The Ple [80, 87] softwares use the waveforms themselves as input features to PCA but REY *et al.* [74] also reviews studies using derived features instead, hence we implemented both options (features are described in the next paragraph), besides processing all data files with both $d = 2$ and $d = 3$ projection dimensions. With either type of feature, we normalized the data by subtracting the mean and dividing by the standard deviation in each dimension — that process is local to each instance of spike detection. Regarding the number of clusters, which must be specified for the K-means algorithm, those softwares offer a *K-means scan* option that is not specified in their manuals. Therefore we applied the Davies-Bouldin score[88] which is one of the three more common indices used for evaluating clustering quality when the ground truth is unknown [89] (the other two are silhouette coefficient and the Calinski-Harabasz index). It is defined as

$$\frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{1}{\|\mu_i - \mu_j\|} \left(\frac{\sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|}{|C_i|} + \frac{\sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mu_j\|}{|C_j|} \right)$$

where k is the number of clusters, C_i is the i-th cluster, and μ_i is the i-th cluster's centroid.

The waveform features listed in Table 2.2 were computed from descriptions in Ple [87]. Notwithstanding they are not used as PCA features in the respective softawre but rather as secondary clues for manually adjusting parameters in various spike sorting methods. Numerous definitions are symmetrical for valleys (waveform minima) and peaks (waveform maxima), so we describe them only once. In the definitions, n represents the number of samples per waveform and $w_1, \dots, w_n$ indicate the samples. Where not qualified as *local*, peak and valley refer to the global maximum and minimum, respectively. Note that there exist cases where some of these features are ill-defined (for instance, when $\operatorname{argmax}_j w_j = n$, there is no peak). In order to obtain better precision, we upsampled all waveforms by a factor of 4x (from 32 to 125 samples) via cubic interpolation.

Table 2.2: Waveform features used in spike sorting

Feature	Definition
Valley/peak count	Number of local minima/maxima with negative/positive voltage values.
Valley/peak tick	$\operatorname{argmin}_j w_j$ and $\operatorname{argmax}_j w_j$.
Valley/peak magnitude	$\min_j w_j$ and $\max_j w_j$.
Valley/peak FWHM	The full width at half-maximum (FWHM) is the sample distance between the points where the waveform crosses half the respective extreme value (peak/valley).
Waveform area	$\sum_{j=1}^{n} \lVert w_j \rVert$
Waveform energy	$\sum_{j=1}^{n} w_j^2$
Waveform non-linear energy	$\sum_{j=1}^{n} w_j^2 - \sum_{j=1}^{n-2} w_j w_{j+2}$
Amplitude	$\max_j w_j - \min_j w_j$.
Peak-valley separation	$\operatorname{argmax}_j w_j - \operatorname{argmin}_j w_j$.

2.1.3 Mapping of V1 and V2 receptive fields

FIORANI *et al.* [56] presented a quantitative method for mapping RFs in the visual cortex. Before that, researchers had often determined RF location and limits qualitatively (that is, by manually relating response to stimulus[3]) since the seminal ecephys experiments investigating RFs in the visual cortex [39–41, 53, 54] In the proposed method, they combine a standard moving bars stimulus with back-projection, a digital processing technique commonly used in computerized tomography to reconstruct a 2D function from multiple 1D projections [90]. Traditional back-projection

[3] A short video strip that shows HUBEL and WIESEL [41] performing this procedure based on an audio representation of spike trains is available at https://www.youtube.com/watch?v=IOHayh06LJ4.

uses the inverse Radon transform [91] to map a usually dense , *i.e.,* a series of integrals of a function $f(x, y)$ along multiple sweeping radial directions, back to the original function. In a similar way, they presented multiple moving bar stimuli to a test subject and used the recorded responses to reconstruct the sensitivity at each point of the visual field, resulting in a matrix representation, dubbed response map, analogous to the theoretical Equation 1.5.

They employed a relatively low number of stimuli (6 or 8, typically) since many factors, like bar movement speed, covered amplitude of visual field, mapping resolution, and number of trials required for statistical reliability needed to be balanced out in order to keep the experiment sessions short, avoiding excessive tissue damage and allowing other stimulus types, such as the previously described sinusoidal drifting gratings. In short, the method involves the following steps:

1. For each movement direction, the stimulus intensity function (Equation 2.1) is convolved with the estimated mean firing rate, obtaining a 1D projection of the sensitivity along that direction

2. A 2D response map is generated by extrusion of the 1D profiles along the corresponding direction. Z-normalization is applied to allow discerning *background*, or *spontaneous* neuronal firing, from *foreground*, or *non-spontaneous* firing.

3. All 2D maps are averaged into a final response map, $z : [-E, E]^2 \to \mathbb{R}$, where $2E$ is the amplitude of the visual field that gets stimulated.

4. If the Z-normalized map peaks at a value of $z_{\max}$ then the RF's center is assumed to be at the peak's location, $\operatorname{argmax}_{(x,y)} z(x, y)$ and its area is defined as the set of points (x, y) containing the center and having $z(x, y) \geq \alpha z_{\max}$ and $\alpha z_{\max}$ is dubbed response cutoff.

2.1.4 Functional attribute derivation

Receptive field mapping

RFs were mapped using the method described in FIORANI *et al.* [56] and summarized in Section 2.1.3 with minor deviations. Our estimated response maps had a resolution of 512 × 512 pixels and we used a Gaussian kernel

$$K_\sigma(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{t^2}{\sigma^2}\right\}$$

with an aperture of $6\sigma = 150$ ms for estimating the basal firing rate $r_0[K]$. Latency was determined similarly to PERES *et al.* [68] but using binary search in the range of $[0, 200]$ ms with a maximum resolution of 5 ms to find the latency value that maximized the response map's peak.

We employed $\alpha = 0.5$ for computing the RF's region and contour (Section 2.1.3). In some cases, multiple connected component satisfying $z(x, y) \geq \alpha z_{\max}$ may be observable in the response map. In this case, we still consider the RF's area to be the connected component that contains the response peak $z_{\max}$ but derived an additional quantity, that we dub *response disparity*, defined as the difference between $z_{\max}$ and the largest peak in the remaining connected components. This quantity was not visualized during the user study but it was used for classifying neuronal observations in Section 4.4.

Besides the RF's center and contour, we derived a few more attributes from the response map, including but not limited to: perimeter, area, area of convex hull, and peak disparity (difference between response peak and second largest local minimum). Only a few of these attributes were encoded during the user study (see Section 3.5). Furthermore, Appendix A contains a complete list of derived attributes and derivation rules.

Gratings functional properties

Unlike the moving bars stimulus, which allows for localized interpretation of observed responses, sinusoidal gratings cover the entire stimulated portion of the visual field, such that the RF is stimulated during the entire duration of the trial. Therefore, the trial-averaged mean spike count $\bar{P}$ (or similarly, the mean firing-rate $\bar{R}$) and the tuning curves $R(s)$ become the most relevant indicators of selectivity. Nonetheless, a few additional functional attributes can be used to indicate preference for drifting direction and gratings orientations. We describe these attributes in the following.

Circular variance for direction (circular variance for direction (CVD)) is an index used to characterize the sensitivity to stimulus direction and is considered adequate for contrasting different neuronal populations because it is less sensitive to noise [92]*apud*[68]. It ranges from zero to one, with a value of zero indicating selectivity to a particular observed direction and a value of one indicating no selectivity between observed directions. It is defined as

$$\text{CVD} = 1 - \left\| \frac{\sum_\theta R_\theta e^{i\theta}}{\sum_\theta R_\theta} \right\| \tag{2.3}$$

where R_θ is the total response (*i.e.,* the mean spike count) observed with direction θ and the sum is computed over the set of all directions. Since the gratings stimulus set featured multiple parameter stimuli $(\theta, A, K, \omega, \mathbf{c})$ for a given θ, the sums were actually computed over $(\theta, A, K, \omega, \mathbf{c}) \in S$, where S is the set of all stimuli. In other words, all directions θ were weighed by responses to all stimuli in which they were the stimulus direction.

The circular variance for orientation (circular variance for orientation (CVO)) is similarly defined as

$$\mathrm{CVO} = 1 - \frac{\sum_{\varphi} R_{\varphi} e^{2i\varphi}}{\sum_{\varphi} R_{\varphi}} \tag{2.4}$$

but since $\varphi \in [0, 180°)$, the exponent in the numerator is $2i\varphi$ rather than $i\theta$, so as to map the set of orientations into the unit imaginary circle.

Two additional indices were also computed, based on PERES *et al.* [68]: direction index (DI) and orientation index (OI). The first is defined as

$$\mathrm{DI} = \frac{R_{\mathrm{max}} - R_{\mathrm{op}}}{R_{\mathrm{max}} + R_{\mathrm{op}}} \tag{2.5}$$

and the second as

$$\mathrm{OI} = \frac{R_{\mathrm{max}} - R_{\perp}}{R_{\mathrm{max}} + R_{\perp}} \tag{2.6}$$

where R_{max} represents the response in the direction/orientation with maximal response, R_{op} represents the response in the opposite direction and $R_{\perp}$ represents the response in the orthogonal orientation.

Besides the orientation/direction indices and circular variances we just described, we also computed *active* versions of these properties by discounting the observed spontaneous firing. We did so by computing the basal firing rate R_B as the mean firing rate during inter-trial periods and trials of the neutral stimulus (see Section 2.1.1). The mean firing rate was computed independently for each of those periods and they were averaged using their duration as a weighting factor. The final active indices were computed using the previously presented definitions, but replacing each response with its difference with R_B. Again, these indices were not visualized in the design study but we used them as features for predicting approval rates in Section 4.4.

2.1.5 Original dataset structure

Out of 3 test subjects (V202, V204, and V206), we only processed recordings of one of them (V206, a *Sapajus apella* adult male specimen). Responses were recorded using the MAP Software (RRID:SCR_003170) [4] by Plexon Inc. with 64 electrodes recording from area V2 in both hemispheres simultaneously — a 16x2 MEA on the left hemisphere, and a 8x4 MEA on the right hemisphere — and signals were

[4] See `https://scicrunch.org/resolver/RRID:SCR_003170` for the *research resource identifier* pertaining to this hardware/software system.

digitized at 40KHz. Various recording sessions were performed, usually alternating between moving bars (Equation 2.1) and drifting sinusoidal gratings (Equation 2.2) stimulus sets at each depth before pushing the MEAs deeper, sometimes with multiple repetitions of the same stimulus set and with independent movement of each MEA. Each session included 10 trial-shuffled presentations of each stimulus in the stimulus set. Furthermore, moving bars data files included only recorded spike trains for each electrode, whereas gratings files also included WB, LFP, and SPKC signals, beyond spike sorting performed with the Plexon Offline Sorter software (Plexon Inc., Dallas, TX, RRID:SCR_000012) [5]. We initially selected a subset of 4 moving bars-gratings recording session pairs in which the insertion depths coincided on both hemispheres. This selection included the left-by-right insertion depths of 500µm by 1500µm, 700µm by 1500µm, 1800µm by 2600µm, and 2200µm by 3000µm. Consequently, we processed eight data files containing double-hemisphere, simultaneously recorded signals with 32 electrodes by hemisphere. We processed the eight selected data files in PLX format[93] using a custom Python library [6] with a common range of parameters that we will describe later in this section.

2.2 Design study literature review

Research in information visualization is a thriving field with lots of opportunity for applied and inter-disciplinary research. However, despite the many new studies published every year in well-known visualization and human-computer interaction/interfaces (HCI) conferences such as IEEE Vis, EuroVis, CHI, and PacificVis, its major challenges have been, for many years, to produce not only original and aesthetically pleasing visualizations for specific domains but generalizable solutions and taxonomies that blend together into a truly scientific discipline, while properly accounting for the human factors that impact the validity of studies. Many works have focused on highlighting the challenges and pitfalls of designing and evaluating visualization systems, and the proposal of *design study* and *design evaluation* frameworks that aim to facilitate conducting new studies or evaluating and comparing existing ones, while unifying vocabulary and taxonomy, has become both a road map for applied research and a role model for high-level research by more experienced authors.

In this thesis, we report research on processing and visualization of visual cortex data. We mainly followed MUNZNER [94]'s nested model for visualization design and evaluation, BREHMER and MUNZNER [95]'s *action-target* task analysis

[5] See `https://scicrunch.org/resolver/SCR_000012` for the *research resource identifier* pertaining to this hardware/software system

[6] Documentation hosted at `https://lcg.gitlab.io/neuro/python-plx`.

framework, and MUNZNER [96]'s guidelines for attribute and interaction encoding, though a number of different works contributed ideas for our modeling and design choices. While not solely a thesis on information visualization, an iterative design study was a fundamental part of this research, therefore this section is concerned with discussing the associated design study and validation literature before we report our own work in the remainder of this chapter.

2.2.1 Conventions

Some words used in the *viz* literature are overloaded with meaning and may bear conflicting ideas in different studies. The word *task* is used with different meanings in the *viz* literature [94], sometimes referring to domain-specific activities, others to abstract operations, sometimes regarding low-level operations on data items (*e.g.,* locate anomalous samples), others pertaining to high-level objectives (*e.g.,* confirm hypothesis). MUNZNER [94], BREHMER and MUNZNER [95], MUNZNER [96] use the words *task* and *operation* to describe abstract actions performed by a user on abstract data and the word *problem* to describe a domain-specific activities, both on high and low level cases. LAM *et al.* [97] uses the word *goal* to describe abstract, high-level objectives a user has, and *task* for the low-level operations on abstract data listed by MUNZNER [94], BREHMER and MUNZNER [95], MUNZNER [96]. We use the words *goal* and *task* with LAM *et al.* [97]'s meanings and save the words *assignment* or *activity* to refer to visualization activities (comprised of multiple goals, each divisible into lower-level tasks) performed by the study participants during the user study reported in Section 3.5.

The terms domain problem and problem domain are commonplace in the *viz* literature, and since they are easily mistaken for each other, their distinction should be stressed: problem domain refers to a field of study to which a referred problem belongs, whereas domain problem is a particular problem inside an application field. Finally, we use the terms field study and user study (or lab study) to designate work performed inside the domain experts work environment (as reported by SHNEIDERMAN and PLAISANT [98]) and inside of a controlled setting, respectively.

2.2.2 Design study definition

According to SEDLMAIR *et al.* [99], a design study is a problem-driven research methodology comprised of the following elements/activities:

Real-world problem At the heart of a design study is a contribution toward solving a real-world problem. Hence real users and real data are mandatory.

Analysis Translating user requirements, jargon, data, and activities into an abstract form (usually further divided into data and task abstraction) is a necessary step to achieve generalizable results.

Design Creative but informed consideration of alternatives from a broad design space.

Validation Proving effectiveness of each contribution, from the very abstraction to results obtained by on-site usage of a tool.

Reflection Discussion of limitations and future work are what effectively turns the design activity into a research field.

At some point, the very fitness of information visualization for the given problem must be questioned, since a fully automatic or at least a semi-automatic solution could be available [99]. More precisely, when task clarity is crisp (that is, the analysis process can be formalized in detail) and information is explicit (that is, data is digitized and structured), an algorithmic approach (possibly based on machine learning) is more appropriate. When the information is solely in the experts' heads, nothing can be done (because there is not enough data). However, when either the task is somewhat fuzzy, or the information is somehow underrepresented in the computer, or a bit of both, the design study methodology becomes appropriate. Nonetheless, *viz* may always be used to communicate and inspect outcomes of information processing systems usage, even the fully automated ones.

2.2.3 Design frameworks

Abstract visualization design frameworks, such as MUNZNER [94, 96]'s *action-target task analysis framework*, provide a common abstract vocabulary that facilitates transferring knowledge between different domains, getting past domain-specific jargon, and choosing proper encodings based on existing work. The design process is rarely strictly linear [94, 99], so going back and forth between analysis stages may be done several times before deploying a new iteration of the proposed solution.

Alternatively, SEDLMAIR *et al.* [99] proposed a nine-step framework for conducting relevant and successful design studies, coarsely divided into three stages that go from collaboration winnowing and role definition (that is, to identify opportunities and determine if the project is promising and realistically realizable) to the core design study phase, which involves prototyping, building and deploying visualization systems, and conducting user studies and thence to the analysis phase, in which reflections are made and the study is published. The nine-stages in SEDLMAIR *et al.* [99] do not align perfectly the design stages proposed by MUNZNER [94], and they are in fact complementary approaches. While the first is more concerned with conducting successful research by listing a sequence of logical steps that must be followed in a study and identifying typical pitfalls that may be committed in each step, the second one is a nested abstraction and validation model that focuses on how to properly identify and validate contributions in different layers. Both studies anticipate the possibility of backtracking study steps but the nine-stage emphasizes it more strongly. All the same, these frameworks are not meant to be taken to heart but rather to provide guidance and prevent known mistakes and anti-patterns to be applied.

Nested design studies

In the nested framework, the output of one more abstract layer constitutes the input to the next, lower-level one. These layers also work as attention scopes for the modeling process and help identify different contributions. They are listed below from outer to inner:

Domain problem characterization The topmost layer involves understanding a specific application problem domain (ecephys in our case), mapping its vocabulary, methods, objectives, and challenges, so that a precise description of the data, algorithms, data analysis activities, and other non-functional requirements may be produced. Most of problem domain characterization for this thesis has been laid out in Chapter 2, although a listing of requirements is only explicitly elicited in Section 3.3, which discusses the next layer.

Task abstraction Map the inputs into a domain-agnostic list of low-level visualization operations on abstract data (*e.g.,* identify trends, locate outliers, derive distribution mean). This stages implicitly requires the data to be described in abstract terms as well, a process that is discussed in greater detail in MUNZNER [96]. LAM *et al.* [97] addresses some of the challenges in bridging higher-level visualization objectives (*e.g.,*discover observation, identify main cause, evaluate hypothesis), which they refer to as *goals*, to sequences of low-level tasks by proposing a few archetypical goals and listing sequences of corresponding tasks in existing works.

Data encoding and interaction design In this stage, visual representations and interaction paradigms are chosen to represent the data and to support performing the tasks that operate on them. The mapping from data items (*e.g.,* rows in a dataset or nodes in a graph) and their attributes (or, equivalently, *dimensions*) into visual elements that represent them is called *encoding.* Data items are usually represented by *marks* (*i.e.,* geometric entities such as dots, symbols, lines, and polygons) and their attributes are represented *channels* associated with the marks (*e.g.,* horizontal and vertical position, color, length, area, tilt). There are usually many ways of encoding the same set of attributes and a great deal of literature covers standard choices, best-practices and even the artistic factors related to it, such as the influential works of TUFTE [100] and WILKINSON [101]. We address data encoding and interaction design in Section 3.4.

Algorithm design The last and lowest-level step in the analysis cycle involves choosing proper algorithms to carry out the actual data encoding and interaction processing. Although it may at first sound like an exhausted research sub-field, many applications require real-time, costly operations such as multidimensional projections, matrix reordering, dynamic (often animated) graph layouts, and filtering/navigation in large datasets that constitute active research problems. Most data processing described in this thesis (Section 2.1) was performed *a priori* due to design study choices (Section 3.5).

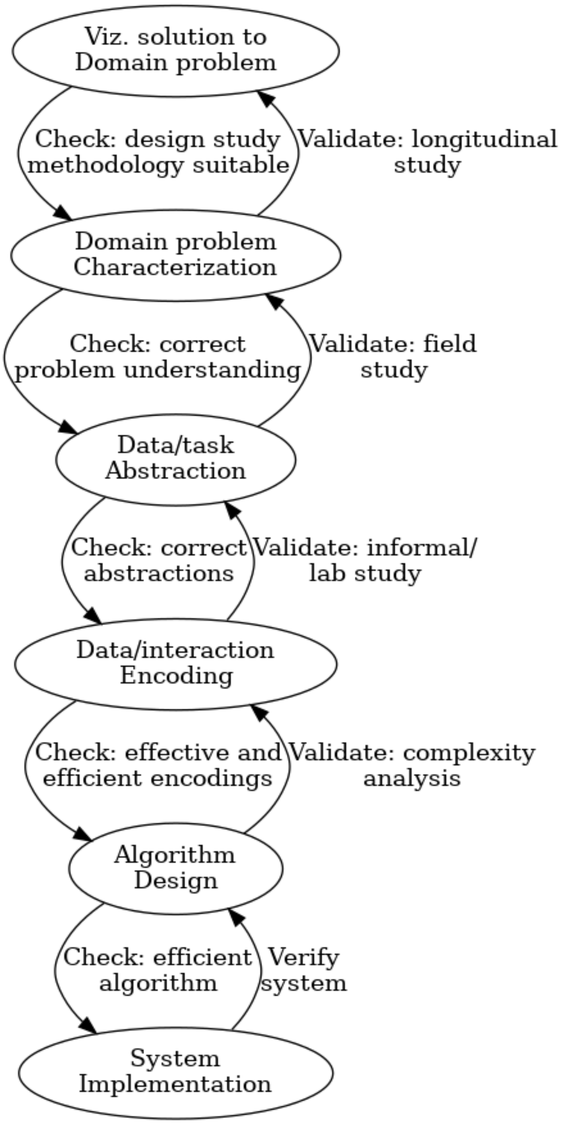
Implementation This is the level at which systems design techniques and actual programming tools are employed to materialize the design. Visualization and graphical user interface software libraries may be used to speed up the implementation of the application's main structure and more prosaic plots. Most challenges here are of technical rather than scientific merit, so this layer is typically not discussed by design studies.

Nested validation

The five scopes of attention and work for nested design studies also work as validation scopes. A design study will usually claim contributions in one or more of the four outer scopes, but each one of them has different inputs and outputs, so they will require different validation procedures. The implementation scope tends to present engineering challenges with lesser potential for novelty or academic publishing, and it usually requires *verification* rather than *validation* — *i.e.,* verifying an implementation is about checking that the built software meets its requirements correctly, whereas validating a design study is about checking that the requirements gathered exist and are relevant (but not necessarily complete) — therefore we do not discuss it.

Validating a scope requires that the scopes inside it are validated first. When no contributions are claimed at a scope and it uses techniques validated elsewhere, then its validation may be skipped altogether. A set of threats/pitfalls should be watched for before advancing scopes, while a more formal validation approach must be adopted when closing a scope, before validation can be resumed at the enclosing one. Since the nested process is rarely strictly linear, we can assume these checks and procedures are to be performed when crossing the boundary between two adjacent scopes, not when incidentally backtracking because of some change in priorities or requirements, for instance. Another way of seeing this is that it is *advisable* to perform downstream advancement checks while working on a project but *mandatory* to follow upstream validation when reporting and discussing a finished project. The diagram in Figure 2.3 illustrates the nested process, featuring downstream checks when descending scopes and upstream validation when ascending scopes.

Figure 2.3: Overview of nested design study and validation framework. In this framework[94] Design specification follows a downstream path from higher, more abstract levels down to lower, concrete levels, whilst observing some concerns. Validation follows the reverse upstream path.


Viz. solution to Domain problem
Check: design study methodology suitable
Validate: longitudinal study
Domain problem Characterization
Check: correct problem understanding
Validate: field study
Data/task Abstraction
Check: correct abstractions
Validate: informal/ lab study
Data/interaction Encoding
Check: effective and efficient encodings
Validate: complexity analysis
Algorithm Design
Check: efficient algorithm
Verify system
System Implementation

Now, let us present a more exhaustive list of the downstream checks and upstream validation procedures, combining advice from the nested framework and the nine-stage one. We will list the scopes from least to most enclosed and inform the outgoing procedures related to each.

Visualization solution to domain problem This scope is not identified in the original nested framework but it may be related to first three stages (the *pre-condition* phase) in the nine-stage framework. The main early advancement threats at this scope, reunited in Pitfalls 3-14, are committing to a collaboration when a *viz* problem does not exist (either the identified problems are not recurrent, existing methods are good enough, or a fully automated solution is possible), or there is not enough data, or the people involved are not fitting (they may lack expertise or access to the data, for example).

Domain problem characterization The early advancement threat is understanding the domain problem incorrectly. Pre-validation involves observing and interviewing the target users in order to correctly understand the problem, which requires appropriate observation techniques (like interviews, contextual inquiry, fly-on-the-wall, *etc.*), a right dose of domain understanding (not too little, nor too much), and proper interaction with domain experts (*i.e.,* enough time with the *right people* must be available). These correspond to Pitfalls 15-18.

On the opposite direction, once a system has been built and the data/task abstraction have been validated, adoption rates and other practices from the longitudinal study methodology[98] should be applied to make sure the constructed visualization system has ecological validity — that is, real users have successfully adopted the system for performing real tasks in the absence of direct interference by the visualization researchers. This is a reportedly harder form of validation, because many factors might influence the final adoption rate of a system and lead to false negatives. In particular, if work practices are very cemented and a visualization system provides borderline but not paradigm-shifting advancements, it could remain unadopted despite being valid. It might also be the case that secondary features (like the ability to export results to a standard format) are missing from the tool but present in a commercial software application. That is a failure of requirements gathering that may have slipped past other steps of the study and hinder tool adoption but by no means imply that the domain problem has been improperly characterized.

Data/task abstraction The premature advancement threats involve choosing incorrect abstractions, which should be countered by watching for Pitfalls 19 (incorrect level of abstraction) and, again, 15-18. Nonetheless, checking that abstractions are correct before advancing may be harder than checking that the problem was understood well since no tool exists as of yet. Adhering to data/task abstraction taxonomies [95–97] is helpful, but being able to do rapid prototyping at the next scope and past it (Pitfall 12), making it possible to iterate faster and restart from scratch if necessary, is a game changer. Otherwise, a big risk is assumed because a lot of time needs to be invested in subsequent scopes before validation becomes possible. Validation typically takes the form of field studies, which differ from laboratory ones by being performed in the domain experts' environment, to verify that the constructed system leads to effective results by real users operating on real data (this relates to Pitfall 25). Complementary anecdotal evidence of usefulness, usually in the form of interviews, may be collected. Pitfall 26 (considering positive feedback as sufficient proof of success) applies. These procedures may seem deceptively similar to those performed for the domain problem characterization scope but whereas that phase is concerned with measuring a tool's long-term adoption rate and aggregated value, this phase intends to make sure that users can make meaningful work with tool in their native environment, which may involve measuring time-and-error quantities, or evaluating the quality of insights produced. Therefore, a higher level of obtrusiveness is acceptable for the purpose of collecting evidence. In other words, late validation of domain problem characterization seeks to answer wether one has built the right tool (that is, tool for solving a useful, relevant problem), while late validation of data/task abstraction validation seeks to answer wether one has built the tool right.

Data/interaction encoding The preliminary threats involve choosing ineffective encoding/interaction techniques, or not properly exploring the design space. Advancing requires justification of chosen encodings, which assumes having enough knowledge of the *viz* literature (failure to do so is Pitfall 2). Mock-up sessions with *viz* experts are a way of pruning the design space and discarding bad alternatives. Pitfalls 20-23 relating to the proper consideration of the design space and rapid prototyping apply. Regarding the later, paper and wireframe protoypes, Wizard of Oz testing, and iterative informal user studies are ways of accelerating the process and detecting dead ends earlier.

Late validation depends on how innovative the proposed design is. When standard data/interaction encoding schemes are used, careful citation of the literature and justification of choices suffices. When new encodings are proposed, both informal user studies and laboratory studies that propose well-define tasks and measure time-and-error are appropriate.

Algorithm design The main advancement threat is choosing an expensive algorithm in terms of computation, memory, or both. Pre-validation is usually straightforward by means of computational complexity analysis. However, this is usually only relevant for a *viz* publication if a new data derivation or encoding scheme (like a multidimensional projection or graph layout algorithm) is being proposed. Late validation may usually amounts to measuring system time/memory.

System implementation That scope is usually not relevant from a *viz* viewpoint but there might exist contributions worth publishing in a software engineering journal or conference. Nonetheless Pitfalls 22 (non-rapid prototyping) and 23 (inadequate usability level) should be watched for, since they might impact upstream validation. Rather than validation, this level requires verification that the requirements (covering data processing, tasks, encodings, and algorithms) have been properly implemented.

2.2.4 Contributions of design studies

Contributions to viz a design study can make on different layers include [94]:

Domain and problem characterization Framing a specific, new domain problem as a *viz* problem.

Data/task abstraction Proposing new ways of performing an abstract task that go beyond a specific application domain (*e.g.,* finding trends in time series, locating outliers).

Data/interaction encoding Proposing new ways of representing information and interactions for new/existing data-types

Algorithm New ways of processing/deriving, or laying out data (*e.g.,* graph and matrix ordering, multidimensional projections, clustering)

As previously explained, validating the contributions at each of these levels presents different challenges. Validating algorithms is usually the easiest, as it amounts to proving it's correctness, theoretical optimality, and/or empirical performance using benchmarking datasets, all of which do not involve the human factor. Some specialists have historically called for more fields studies [94, 98], arguing that the field would benefit more from problem domain characterization studies. Not surprisingly, this is possibly the hardest validation case of all, since the success of *viz* solution usually depends on long-term, unobtrusive evaluations of tool adoption

— as done by SHNEIDERMAN and PLAISANT [98] for example — besides requiring that either all inner layers are validated as well (thus claiming additional contributions and making the study too broad and hard to understand), or that all techniques used for the inner layers are already validated by existing studies (which is a rather uncommon situation) [94]. Nonetheless, rather than just reporting findings and their known implications (commonplace in the natural sciences) the process of writing and reflecting about a design study is part of the process of making sense about the findings and revealing insights about future directions (a process more common in the social sciences) [99].

2.2.5 Task analysis

Task analysis is a broad term used for the process of understanding how specialists in a given problem domain accomplish their research goals and translating that perception into an abstract description formal The broader activity of getting acquainted with a domain and understanding its practices is sometimes referred to as cognitive work analysis (CWA), whence task analysis refers to modeling lower-level cognitive operations that work on well-defined data to produce certain outputs (like derived data). This activity usually involves translating domain jargon to general, abstract descriptions more familiar to computer science.

Asking users to introspect about their questions (what they want to know) and procedures (how they seek to answer it) is a notoriously insufficient requirements gathering technique [99, 102]. Therefore, design studies may approach the challenge of CWA in a variety of ways, some of which include:

Just talking Promoting formal meetings and encounters between *viz* specialists and domain experts with the aim of discussing work practices and collaboration goals, and eliciting requirements.

Fly-on-the-wall The *viz* researcher observes the daily labor activities of the domain experts and takes as many notes as possible without interfering or asking questions, in an attempt to observe untempered working conditions.

Contextual inquiry Unlike the previous method, this allows the *viz* researcher to interrupt the observed specialists to ask for clarification.

SEDLMAIR *et al.* [99] argues that contextual inquiries work better than the other two methods, because it allows to observe assumptions and procedures that specialists take for granted and would fail to report when *just talking*, while still obtaining clarification about steps that could be misinterpreted due to lack expertise and vocabulary when using *fly-on-the-wall.* Nonetheless, all of these approaches may let silent cognitive processes that unfold silently inside the observed worker's mind to slip unaccounted for. They criticize ethnography-based approaches, saying that a *viz* researcher should only learn so much about a particular domain as necessary to propose effective visualizations, at the risk of either spending too much time learning about the field or focusing on concerns beyond the scope of visualization at some point — this is a pitfall (PF-18) that I fell into.

Multi-level action-target task typology

The multi-level action-target task typology proposed by BREHMER and MUNZNER [95], MUNZNER [96] identifies three fundamental aspects of task analysis: *what*, *why*, and *how*. In this taxonomy, an abstract task corresponds to a complete specification of the *what* (input and output targets), *why* (action), and *how* (encoding). The *what* aspect corresponds to the inputs/outputs of each visualization task, and should be described in terms of abstract data types, like graphs, tables, or continuous sampling fields, for instance. We perform this data abstraction in Section 3.2. The *why* aspect is comprised of multiple levels of specification that translate the question a user is trying to answer through visualization into a general vocabulary. On a first level, a user could be either seeking to produce data (that is, to derive an attribute, to make an annotation, or to record their reasoning process), or to consume data (that is, to discover something, to present, or to aimlessly explore a dataset). On a second level, the way they search for information depends on wether they know what they are looking for, and where to look for it, resulting in four modalities, if we consider those as binary statements. On the third level, they distinguish between cases where a user wants to simply identify visualization targets, compare multiple such targets, or summarize them. The correct specification of the targets (inputs and outputs) is also fundamental for the subsequent step, as they could be looking for an specific attribute, trends, outliers, distributions, or relationships, for instance. Finally, the *how* aspect pertains to how the selected targets (attributes) and actions (operations) are encoded into a user interface.

One challenge with such frameworks is that they may seem overspecified and difficult to apply, and still miss on the abstractions necessary to translate some problem domains. After all, what's the point of following a difficult recipe if I may need to add unforeseen ingredients and steps? On this line of thought, LAM *et al.* [97] approached the challenge of bridging *lower-level* analysis tasks to *higher-level analysis goals*. They described an additional abstraction layer, consisting of higher-level analytic user goals, pondering that reasoning in terms of low level abstract tasks (e.g., identify outliers, sort, locate) is difficult in the absence of contextual information (for example, "what is the goal of doing a certain task w.r.t. the domain situation?" and "how does this task sum up to others in order to answer a certain question?"). They proposed to reverse the analysis direction by:

1. Identifying higher-level goals, like exploring and explaining, then 2. Unfold a goal into series of steps and loops that can be directly mapped to low level tasks, such as browsing, and comparing attributes.

This methodology was developed as a result of analyzing a number of IEEE VIS studies (2009-2015) and relating their task analyses to their goals, and serves as a complement, for instance, to the bottom-up *Action-Target task analysis typology* of MUNZNER [94], BREHMER and MUNZNER [95], MUNZNER [96].

Bridging from higher-level goals to tasks

No matter what form of CWA is employed, bridging from a broad, sometimes tacit understanding of work in a particular domain to an extensive list of lower-level visualization tasks operating on well-defined data types using a catalogue of battle-tested data encoding and interaction modes is notably difficult. That is why LAM *et al.* [97] proposed the identification of *user goals* as higher-level analysis objectives that can be mapped into sets of of lower-level tasks, like the action-target task typology [95, 96]. Their list is non-exhaustive, and we rely on the idea of *user goals* to frame an ecephy problem as a visualization problem, rather than using their particular taxonomy. After all the plethora of analysis frameworks and task/encoding taxonomies available in the literature cannot prevent the peculiarities of an application domain to creep in and prove existing methodologies insufficient.

2.3 Fundamentals of machine learning

Let us review some fundamental concepts of ML, based on IZBICKI and DOS SANTOS [103]. Our context is the prediction of neuronal observation approval rates from functional attributes, that is, we mean to regress the former from the latter. The approval rate $y \in [0, 1] \subset \mathbb{R}$ corresponds to the output, or *target variable*, and the functional attributes $\mathbf{x} \in \mathbb{R}^n$ comprise the inputs, or *features.* A pair $(\mathbf{x}, y)$ is called an *observation* (in this section, we will use this word denote feature-target pairs that rather than neuronal observations). A dataset D is comprised of m observations, that is $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$.

A *regression function*, indicated by $r : \mathbb{R}^n \to \mathbb{R}$, approximates the value of the target variable with a precision that is quantified by a *loss function*, indicated by $l(r; \mathbf{x}, y) \mapsto \epsilon$, with $\epsilon \in \mathbb{R}$, that is usually required to be a non-negative, monotonically increasing function of $\|r(\mathbf{x}) - y\|$. A typical choice to satisfy that requirement, and the one adopted in this exposition, is the quadratic difference, or $l(r; \mathbf{x}, y) = (r(\mathbf{x}) - y)^2$. The so-called *cost function* represents the average value of the loss function over D, that is

$$L(r; D) = \mathbb{E}[l(\mathbf{x}, y) \mid (\mathbf{x}, y) \in D] = \frac{1}{m} \sum_{i=1}^{m} (r(\mathbf{x}_i) - y_i)^2$$

A *train-test split* of the dataset is a random partition of D into sets D_{train} and D_{test}, known as training and test sets, therefore these sets fulfill

$$D_{\text{train}} \cap D_{\text{test}} = \emptyset \tag{2.7}$$

$$D_{\text{train}} \cup D_{\text{test}} = D \tag{2.8}$$

This partition is fundamental for the process we will describe next.

Consider the *linear regression*, one of the simplest regression models, which can be expressed as

$$r(\mathbf{x}) = \theta_0 + \theta^{\intercal}\mathbf{x} = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$$

where the coefficients $\theta_0, \ldots, \theta_n$ are referred to as *model parameters.* To simplify notation, it is customary to indicate $\theta^{\intercal} = (\theta_0, \ldots, \theta_n)$ and consider that a leading 1 is prepended to $\mathbf{x}$, allowing us to write $r(\mathbf{x}) = \theta^{\intercal}\mathbf{x}$. In order to obtain a regression function that makes the best predictions, one seeks the value of $\theta \in \Theta \subset \mathbb{R}^{n+1}$ that minimizes the loss (that is the prediction error) over known inputs, that is $L(r; D)$, where Θ represents the parameter space. A *training algorithm* consists of an optimization algorithm that produces a sequence of parameter vectors $\left\{\theta^{(k)}\right\}_{k \geq 0}$ corresponding to regression functions $\left\{r^{(k)}\right\}_{k \geq 0}$ where it is desirable that $L\left(r^{(k)}; D_{\text{train}}\right) \geq L\left(r^{(k+1)}; D\right)$ although the model's performance, measured by L may oscillate over training epochs, indexed by k. Depending on the shape of the L function, we may either obtain a guaranteed a global minimum, or specify halting criteria (preferably at a local minimum) in order to return a solution $\hat{\theta} \leftrightarrow \hat{r}$. In the specific case of linear regression, the minimization of L has a closed-form solution. In general however, optimization algorithms use the gradient of L with respect to θ and one or more heuristics to search for incrementally better solutions.

Since we only have access to a finite sample of data, the solution's optimality is affected by factors such as

- Sample size m and sample quality,
- Dimensionality n of feature space,
- Complexity of $L(r; D)$, and

- Vastness of Θ,

therefore the training algorithm may produce a solution that scores well on a dataset D but not on a separate dataset D'. In essence, obtaining parameters

$$\hat{\theta} \approx \operatorname*{argmin}_{\theta \in \mathbb{R}^{n+1}} \mathbb{E}[l(r_\theta; \mathbf{x}, y) | (\mathbf{x}, y) \in D] \tag{2.9}$$

where the $\approx$ sign is there to highlight that, in general, we may not obtain a globally minimal solution, does not assure that $r_{\hat{\theta}}$ is also minimal for arbitrary observations $(\mathbf{x}, y) \notin D$. That is why the training algorithm evaluates only $L(r; D_{\text{train}})$, leaving $L(r; D_{\text{test}})$ for estimating the final cost, or *fitness*, of the solution found. In other words, the optimal regression function $\hat{r}$ minimizes the cost on the training set but its actual performance on unseen data is estimated using the test set. When $L(r; D_{\text{test}}) \lll L(r; D_{\text{train}})$, $r_{\hat{\theta}}$ is said to *overfit* the training data — it is particularly good at explaining an specific sample but not at extrapolating those results to an unseen sample — and when $L(r; D_{\text{train}}) \lll 0$ it is said to *underfit* the data — it is not powerful enough to even explain a specific sample. In the latter case, $L(r; D_{\text{test}})$ and $L(r; D_{\text{train}})$ might even have similar values.

As we mentioned before, linear regression allows for a closed-form solution of $\hat{\theta}$ on the right-hand side of Equation 2.9 but in the general case, we employ stochastic optimization methods for finding a candidate solution whilst avoiding overfitting. In that scenario, parameters that control the training process or the very structure of the parameter space Θ (*e.g.,* regularization coefficients, stochastic gradient descent (SGD)'s learning rate, or the number and size of layers in an ANN), known as *hyperparameters*, must be estimated prior to training. Thence, it is customary to save the test set to be used only for a final comparison of one or more trained models (in which case it is also known as *hold-out set*) and to estimate optimal hyperparameters by performing preliminary training sessions that are evaluated on the basis of an additional fraction of the training set set aside as a validation set. Alternatively, the method known as k-folding splits the training set into k equally sized portions, or *folds*, and for each tuple of hyperparameters, trains k models on each of the k subsets of $k-1$ folds, then estimates the fitness of the corresponding tuple by averaging over the cost of the k validation folds left out of each sub-model.

Finally, let us introduce three metrics that will be important for discussing the performance of majority-approval classifiers. A *classifier* $r(\mathbf{x})$ is a regression function that predicts a discrete rather than a continuous outcome variable Y. A *binary classifier* is such that $Y \in \{0, 1\}$, where $Y = 0$ is said to be *negative outcome* and $Y = 1$ is said to be *positive outcome.* A binary classifier's *precision*, P, is the ratio of predicted positive outcomes that are indeed positive, and its *recall*, R, the ratio of positive outcomes that are predicted as so. Formally,

$$P = \frac{\sum_i r(\mathbf{x_i}) y_i}{\sum_i r(\mathbf{x_i})} R = \frac{\sum_i r(\mathbf{x_i})}{\sum_i y_i} \tag{2.10}$$

where all sums are computed over D_{test}. Ideally, one seeks a binary classifier with $R = P = 1$, that is, a model that flags no *false positive outcomes* $(r(x_i) = 1, y_i = 0)$, nor *false negative outcomes* $(r(x_i) = 0, y_i = 1)$. It is customary to unify these metrics using a third one, the so-called *F1 score*, which is simply the harmonic mean of precision and recall:

$$F_1 = \frac{R\,P}{R + P} \tag{2.11}$$

Note that these metrics are commonplace when comparing binary classifier performance on the test set not during training. For training purposes, the most common loss function for binary classifiers is rather the binary cross-entropy.

2.3.1 AutoML

ML is thriving field with a long history of developments and a large arsenal of techniques. As such, training and evaluating a model that accurately predicts the value of target Y given observed features $\mathbf{X}$ with the best available knowledge requires procedures that vary considerably with model type, feature semantics, and dataset size, among others. To make the application of ML more accessible to a wider audience and to facilitate preliminary exploration of the solution space, the field of automated machine learning (AutoML) has emerged with the proposal of intelligently searching for promising models [104]. Some of the procedures typically performed by AutoML software packages include

- Data preparation (*e.g.,* dealing with missing values, splitting data into training and test sets),
- Feature engineering (*e.g.,* scaling, demeaning, encoding qualitative variables, and deriving new features by combining old ones, as in dimensionality reduction),
- Hyperparameter search (*i.e.,* searching parameters that affect learning or change the model structure itself),
- Training and evaluation,
- Search space pruning (*i.e.,* proactively aborting unpromising search branches based on previous results),
- Ensembling (*i.e.,* combining lower performance models to obtain a more powerful one),

- And neural architecture search (*i.e.,* a specialized type of hyperparameter search for neural networks).

Ultimately, the quality of results will depend on the algorithm's ability to select promising model classes and eliminate non-performant hyperparameter, the availability of computational power and search time, user-specified configurations (such as the ability to ignore certain classes of models, or hints about exploitable data properties), and naturally on the very volume, dimensionality, quality, and predictability of the data.

We employed an AutoML framework based on Bayesian optimization, Auto-sklearn [105], to test the potential of various regression methods performance. The software tested ensembles of various types of models (KNN, Adaboost, random forests, gradient boosting), for about 13 hours of searching with 2 CPUs allocated to each regression problem (moving bars and gratings stimulus types *vs.* summative and individual observation sets).

Chapter 3

Design study

Our collaboration with neuroscience domain experts from the LFCOG started out as series of informal encounters to discuss visualizations of functional attributes in the previously described (Section 2.1.4) dataset of neuronal observations of the primate V2, as a way to gain knowledge into the domain's goals and practices. As the encounters went by, it became increasingly clearer that winnowing plausible RF evidence from implausible observations was a key part of the process albeit not an analysis objective *per se*. Therefore, we proposed and developed a visualization web tool for performing an annotation task, with the goal of aiding domain experts in deriving a dataset of plausible neuronal observations from a larger mass of possibilities. That tool allowed domain experts to approve/reject neuronal observations selected from a slice of the V2 Dataset based on condensed functional attribute visualizations that were familiar to them, while juxtaposing observations derived with different parametric attributes, so that we could later provide insight on the influence of these parameters on approval rates and population distributions.

By describing the data acquisition and pre-processing procedures in the ecephys field, Section 2.1 already laid the ground for an abstract description of the V2 Dataset. Therefore, this chapter will start (Section 3.2) by abstracting the data, (*i.e.,* describing the main attributes, attribute categories, dimensions, and cardinalities) in the V2 Dataset before discussing the user goals in the terminology of LAM *et al.* [97]. After discussing these goals, which are a form of top-down analysis, it will unfold them into sequences of tasks (Section 3.3) in the context of the action-target design framework [94, 96]. Together, these two sections provide a domain problem characterization of neuronal population studies from ecephys recordings performed by our collaborators from LFCOG, in viz terminology. In the sequence, Section 3.4 discusses data and interaction encoding and how these evolved from an informal design study to a (laboratory) user study. These three sections correspond to the questions *what/why/how* of the action-target framework. Then, Section 3.5 reports the user study in itself, describing the proposed task, its assumptions and hypotheses, dataset and participant selection, the assembly of developed encodings into a web interface, and the telemetry process. The evaluation of this study in qualitative and quantitative terms, including its nested validation and usage data analysis, and a discussion about its limitations and future directions are the subjects of the next chapter.

3.1 Methodology

As previously reviewed, the nested design study framework [94, 96] identifies five main scopes of attention and work for design studies: domain problem characterization, data/task abstraction, data/interaction encoding, algorithm design, and implementation. The scopes are entered in order, producing outputs that serve as inputs to the next scope. The output of the domain problem characterization scope is an abstract description of a domain-specific problem, translating jargon into abstract terms more familiar to visualization and computer science. From that abstract description, the data/task abstraction can produce mathematical descriptions of data and cognitive operations (tasks) being performed on it. Then, proper encodings are chosen among existing techniques or new ones are developed to represent the abstract data/task descriptions, producing a specification of a user interface, including how data should be translated to pixels, animations, or other interactive features, and how the system should react to user actions. Thence, algorithms may be designed (notably when new encodings are proposed), and finally, an implementation is produced.

Realistically, these scopes are rarely entered once, in strict order, and do not constitute a full picture of a design study. Many more steps, from the very establishment of a collaboration between domain experts and *viz* researchers to the publishing of results are involved, and a design study will frequently backtrack to deal with changes in requirements, modeling, and deadlines [99]. For instance, it is reasonable to picture data being revisited many times to cope with the identification of new attributes (which might have been previously neglected) and even being performed in parallel with domain problem characterization. In fact, the risk of committing any fatal failures while running through the scopes fast-forward and building a system that is ultimately invalid is elevated, which is why MUNZNER [94] identifies preliminary checks to be performed before advancing scopes and SEDLMAIR *et al.* [99] lists pitfalls associated with each of a finer-grained, nine-stage design study framework. It is nonetheless useful to think of a design study in terms of these scopes because they set semantic boundaries that help us to identify different potential contributions and set preconditions for successful validation. We will describe our approach to risk prevention and validation later in this section.

As we mentioned, the nine-stage framework [99] paints a more detailed picture than the nested framework [94, 96] of the design study process by identifying issues related to collaboration establishment and results publishing. Nonetheless, we believe the terser format and the semantic boundaries of the nested framework are more appropriate for reporting a design study and discussing its validation. Therefore, the content of this chapter is organized in a way that seeks to reflect these scopes, while still reporting on insight gained from backtracking, with Section 3.2 discussing the data abstraction, Section 3.3 discussing domain problem characterization task abstraction, and Section 3.4 discussing data/interaction encoding. We use the concept of user goals by LAM *et al.* [97] to characterize the domain problem rather than discussing it in a separate section, since they provide a straight bridge between the ecephys background introduced in Chapter 2 and well-established abstract tasks [95, 96]. The section on data/interaction encoding reports on our preliminary, iterative informal user study, justifying some changes in encoding and some findings that helped restrict the domain problem characterization. The final form of the visualization tool is presented in Section 3.5, which also sets some evaluation goals addressed later in Chapter 5.

3.2 Data abstraction

In this section, we describe the structure of the V2 Dataset in viz terms. Let us briefly recall the V2 Dataset structure described in Section 2.1. At a *bare level*, it is composed of ecephys recordings acquired from visual cortex area V2 with multiple electrodes, over multiple experiments, with a variety of visual stimuli. It may be organized as tabular data with several indexing columns and it is possible to define the item (row) level in multiple ways, depending on the *entities* that we are interested in studying. The lowest possible level could be defined as that of a spike train (a time series) detected at a specified location in the visual cortex, during some repetition of a visual stimulus, given specific preprocessing parameters. On a higher level (a more aggregated one), we could define each row to contain the evidence of a hypothetical, individual neuron (single-unit in the domain jargon). That implies that lower-level attributes must be stacked into vectors or matrices (for instance, matrices of spike trains) or summarized in another way to become columns at this abstraction level. At an even coarser level, rows with identical acquisition conditions may be further aggregated, resulting in summative, hypothetical neuronal observations (multi-units in domain jargon). As we will describe soon, the two later abstractions are the most convenient to address the users' objectives. As the dataset itself may be parametrized by the application of domain-specific algorithms that act on the *bare data*, changing the algorithms's parameters generates new collections of rows without changing the bare data. Therefore, it can be seen as a virtually infinite table, with those algorithms' parameters acting as range indices. If the items are, indeed, hypothetical neuronal observations (either individual or summative), since the attributes affect the number and quality of generated hypotheses, if follows that parametrization is a key factor when analyzing populations, their distributions, and related uncertainties.

The attributes we just recalled may be grouped into the following categories.

Metadata Identification and temporal attributes, like the test subject name, and the experiment date/time, and the stimulus type.

Anatomical Specify the location of a recording in the brain (recording site). This is one of the core attribute types for the user goals.

Physiological Incidental characteristics of the anatomical recording sites. They are obtained by post-mortem analysis of the visual cortex tissue whence recordings took place. This is one of the core attribute types for the user goals.

Parametrical These are free parameters that the users may tweak in order to obtain different results. They may be further divided into: *detection parameters*, which affect how responses are obtained from raw signals; *sorting parameters*, which impact the splitting of detected responses into multiple individuals; and *functional parameters*, which affect the computation of functional attributes, described below. Parametrical attributes have a considerable impact on the data volume to be analyzed and the conclusions that may be drawn from visualizing that data. Their impact will be discussed in detail in Chapter 4.

Stimulus Provide a precise parameterized description of visual stimuli.

Chronological Specify when a stimulus repetition took place and for how long. Each repetition of a stimulus is called a trial in the ecephys domain.

Response Describe how, given an acquisition/signal processing parametrization, a hypothetical cell reacted to presented stimuli, on a bare level. Basically, the spike trains and corresponding waveforms. These are not directly visualized, but rather derived into the next category of attributes.

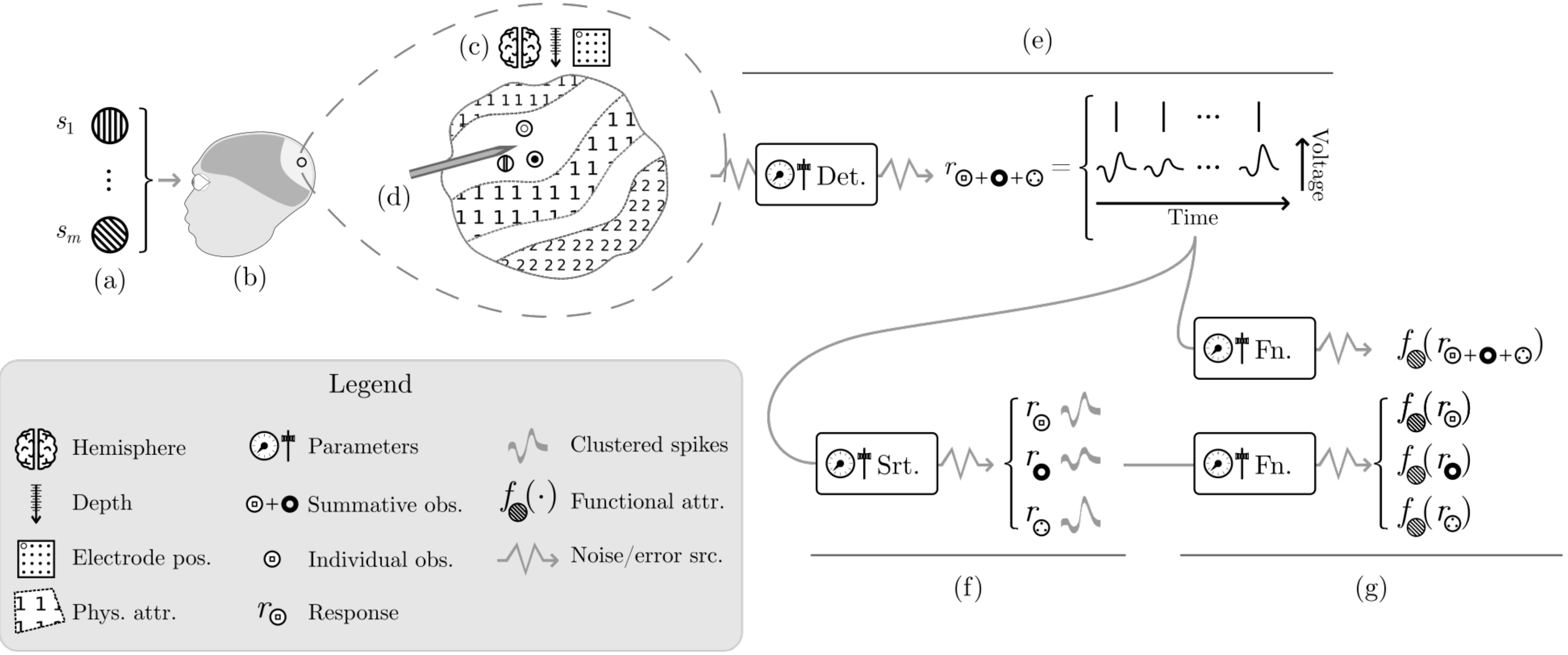
Functional Functions of the response that indicate how selective a hypothetical neuronal observation (either individual or summative) is to a particular set of stimuli. Many functional attributes summarize stimulus and response by representing the later as a function of the former, usually by aggregating over all possible chronological instantiations (trials) of a stimulus. This is one of the core attribute types for the user goals.

As we mentioned previously, the granularity of the dataset (*i.e.,* the definition of the row/item level) depends on the user goals. The basic study entity that the domain experts are interested in is that of a neuronal observation. Therefore, considering that most metadata attributes are constant in the dataset, and that functional attributes effectively summarize stimulus, chronological, and response attributes, we may finally abstract the dataset as a virtually infinite table, where anatomical and parametrical attributes are the indexing columns (or free variables), and physiological and functional attributes are the dependent columns/variables. Figure 3.1 summarizes the data acquisition and derivation phases, highlighting the attribute categories that we described.

With this modeling, by fully specifying anatomical, and parametrical attributes, one gets a collection of a few single-units acquired at identical conditions, or equivalently, their corresponding multi-unit. That is, the union of these attributes works as a compound *index* or *key* to multi-units. In order to break the aggregation and fully index individual single-units, an uppercase letter (A, B, C, ...) is assigned to each row in a multi-unit group by the spike sorting algorithm, in no particular order. This derived attribute has no deeper meaning other than distinguishing individuals, so it is not placed in any attribute category. The set of parametrical attributes has infinite cardinality and non-trivial topology, since some attributes are discrete (like the number of PCA dimensions used during spike sorting), others are continuous (such as the signal threshold for spike detection) and yet others are co-dependent (for instance, filter cutoff parameters only make sense if spikes are detected from a wideband signal). Therefore, this set of attributes may not be easily represented as a grid.

Due to the choices made in the user study (Section 3.5), only a subset of these attributes was selected from the original dataset, and their values were obfuscated to the study participants, so we will not be concerned with this issue. Nonetheless, the high cardinality aspect is an important feature to keep in mind when dealing with similar datasets produced in the ecephys field. Consider, for instance, that the average number of individual observations (single-units) found through spike sorting in this dataset is 3.85, and that there are 64 distinct recording sites (electrodes) and 39 recording sessions were performed in total. Then, adding a single extra parameterization to the search space during a visualization session would generate approximately 9600 new individual observations, on average.

Figure 3.1: Data acquisition and derivation. A visual stimulus set is presented multiple times, in shuffled order (a) to a test subject (b). Meanwhile, electrodes are positioned at certain anatomical positions, or *recording sites* (c), to record the signals of various individual cells (d). A spike detection algorithm is applied to derive a summative spike time series from raw signals (e). A spike sorting algorithm is applied to derive individual spike time series to approximate the behavior of original individuals (f). Functional attributes may be computed from both summative and individual responses as a way to study the sensory characteristics of recorded cell populations. Most acquisition and derivation steps are subject to some error due to physical interference or modeling simplification, as indicated by jagged lines and parameter knobs.



The response attributes have semantics of *matrices of time-series*. That is merely a formal consequence of our abstraction placing neuronal observations at the item level. Since spike trains are recorded for each repeated presentation (trial) t of all stimulus variations (*e.g.,* all directions of a moving bar) s, the time-series of response attributes need to be indexed by $[s, t]$. In fact, most of those attributes are never directly visualized (except for spike waveforms). It is just relevant to understand they pose an intermediate type of information that is used to derive the functional attributes, which are indeed visualized.

Table 3.1 presents functional attributes common to both stimulus types. Most properties have function semantics, which is expressed using the notation $A \to B$ to indicate a function from set A to set B. For instance, a common case is $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\} \to \mathbb{R}$, which indicates a function from stimuli to real numbers — these attributes are computed by averaging over trials (see Section 1.1.1) to get rid of the trial index in response matrices $R[\mathbf{s}_i, e_j]$. Table 3.2 lists moving bars functional attributes relevant for the design study, and Table 3.3 does the same for gratings. See Section 2.1.3 for details on how the response map is estimated from firing rates and how the other properties can be derived from it.

Table 3.1: Functional attributes common to all stimulus types. S indicates the stimulus set, T corresponds to trial duration, $\mathbb{R}_{\geq 0}$ indicates the non-negative real numbers, and n is an arbitrary positive integer.

Functional attribute	Domain	Description
Spike counts	$S \to \mathbb{R}_{\geq 0}$	Average number of spikes per stimulus.
Firing rates	$S \times [0, T] \to \mathbb{R}_{\geq 0}$	Per-stimulus time-varying firing rate function (see Section 1.1.1).
Basal firing rates	$\mathbb{R}_{\geq 0}$	Estimated spontaneous firing rate (r_0 term in Equation 1.6)
Active responses	$S \to \mathbb{R}$	Integral of firing rate function minus basal firing rate over trial duration.
Tuning curves	$\mathbb{R} \to \mathbb{R}_{\geq 0}$	Functions from stimulus parameters to mean firing rates.
Mean waveforms	$\mathbb{R}^n$	Sample-wise mean of waveforms.
Waveforms deviation	$\mathbb{R}^n$	Sample-wise standard deviation of waveforms.

Table 3.2: Moving bars functional attributes. $V_x \times V_y$ indicates the visual field domain and $\mathbb{R}_{\geq 0}$ indicates the non-negative real numbers.

Functional attribute	Domain	Description
Response map	$V_x \times V_y \to \mathbb{R}$	Z-normalized reconstructed response intensity over visual field.
RF center	2D point	RF center, estimated from response map peak.
RF boundary	2D polygon	RF boundary, estimated from vicinity of response map peak.
Perimeter, area, major axis, minor axis	$\mathbb{R}$	Geometric derivations of RF boundary.
RF latency	$\mathbb{R}_{\geq 0}$	Estimated delay between stimulus and response (volume under response surface).
Total response	$\mathbb{R}$	Integral of response map over visual field.
Polargram (of active firing)	$[0°, 360°) \to \mathbb{R}$	Fraction of active firing taking place within RF boundaries

Table 3.3: Gratings functional attributes

Functional attribute	Domain	Description
Preferred orientation	$[0°, 360°)$	Orientation with the highest average response
Preferred direction	$[0°, 360°)$	Direction with the highest average response
Orientation index	$(0, 1]$	Discrepancy between response in preferred orientation and its perpendicular orientation
Direction index	$(0, 1]$	Discrepancy between response in preferred direction and its opposite direction
Direction circular variance	$[0, 1]$	Variance of direction selectivity, with a value of 0 indicating selectivity to one exclusive direction and a value of 1 indicating omnidirectional response
Orientation circular variance	$[0, 1]$	Likewise for orientation
Active derivations	N/A	Active firing rate versions of all attributes above

3.3 Cognitive work analysis, requirements, and tasks

Our CWA was mainly based on think-aloud sessions with two focal participants and, occasionally, other members of the LFCOG that participated in the user study later on. We employed several discussion sessions presenting on-screen digital prototypes[1], as reported in Section 3.4.1, mostly with familiar representations of neuronal observations, and used those as departure points to discuss domain procedures, concepts, and investigation strategies. We found that to be the most affordable alternative to contextual inquiry in the face of scarce expert time, since it was only possible to meet the experts for a couple of hours a few times every year. Additionally, we got acquainted with academic writings of the group and other neuroscience literature (reported throughout Chapter 2), as recommended by LLOYD and DYKES [106]. In this section, we describe the user goals, their chaining, and the abstract visualization tasks that unfold from them.

3.3.1 Overall goals and requirements

From our observations (conversations and readings), we identified the following main activities in their workflow, as illustrated by Figure 3.2:

1. **Design experiments** for data collection — choosing stimulus sets and corresponding protocols to drive neuronal populations, based on prior knowledge about adequate stimulus types, and depending on the kind of analysis that is intended.

2. **Collect data** by

   - Preparing the laboratory, recording/stimulation equipment, and test subjects,
   - Recording summative responses (signals) at multiple anatomical locations (recording sites), and
   - (Occasionally) sacrificing test subjects before extracting neural tissue and performing histological analysis to determine physiological attributes at each recording site.

   The outputs of this phase vary but typically include raw signals, either as spike trains or wideband signals, and per-recording site physiological attributes.

[1] Original prototypes available at `https://lcg.gitlab.io/pasad/dsc/journal/f8d10c1c21/lab/E007/report.html`.

3. **Derive**

    - Putative summative/individual responses (spike trains) from per-recording site recordings, and consequently,
    - Corresponding neuronal observations and their functional attributes.

    That involves tweaking spike detection parameters (when wideband signals rather than preprocessed spike trains are used as inputs), performing spike sorting, and (possibly) applying other methods (like the RF mapping method described in Section 2.1.3) to obtain summative and individual functional attributes.

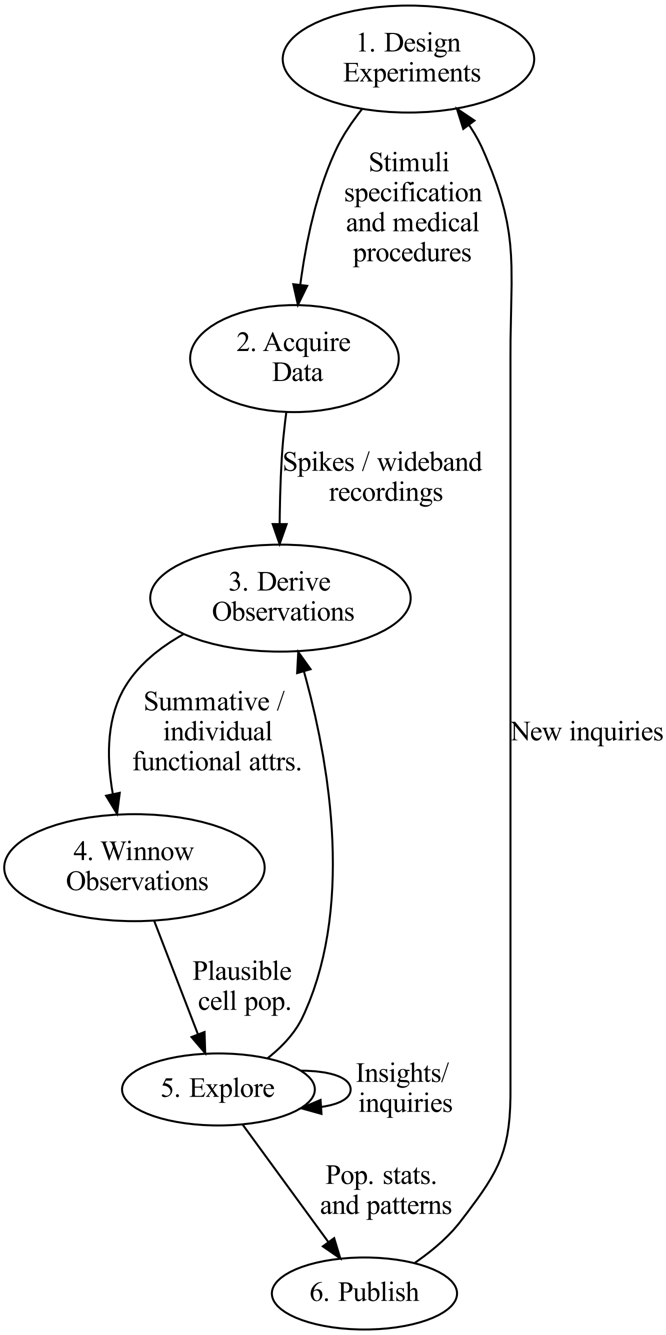
4. **Winnow** neuronal observation populations at various aggregation levels (*i.e.,* individual, summative, or arbitrary populations) based on available functional attributes. During this analysis, unpromising summative profiles are discarded, while the ones that seem composed of actually plausible individuals are inspected in more detail.

5. **Explore** and **characterize** population distributions (and **investigate** hypotheses). Single and joint distributions of attributes such as functional, anatomical, and physiological are compared in search for new or expected features and relations.

6. **Publish** summarized results, in the form of tables, plots, and numbers that encode population statistics.

Figure 3.1, presented in the previous section, focused more on the data acquisition and transformation aspects of the ecephys workflow, whereas in this section we are more concerned with how work is performed by specialists. Let us therefore look back at that figure and establish a few correspondences between the data gathering/processing steps illustrated there (which we will refer to by using the same letters as in that figure) and the work Activities described here, in Figure 3.2 (which we will refer to by their title). Designing experiments predates data acquisition but implies making decisions about things like the type and variety of stimuli (a) that will be presented, the species of test subjects (b) as well as the number of specimen, the anatomical locations surveyed (c), and the characteristics of electrode matrices (d). Acquiring data was implied by the transition from (c-d) to (e). Deriving observations involves, as we mentioned, spike detection and sorting and functional attribute computation, which were represented as steps (e-g). The ability to replay those steps depends on which signals and attributes are saved along the process. For instance, if only spike trains rather than wideband continuous electric potentials are

Figure 3.2: Domain-specific workflow. Activities 1-2 require planning ahead, as they condition what analyses will be possible down the road. Activities 3-5 constitute analytical user goals, in *viz* terminology, and may be iterated multiple times before the analysts deem results as satisfactory — a single reverse arrow is presented between activities Activity 6 involves more mundane actions like gathering most interesting results and planning paper layout but the visual presentation of results also constitutes an interesting locus for *viz* activities.



saved, then replaying spike detection is not viable, whereas the ability to perform spike sorting depends on having waveformss available. The remaining work Activities, starting with winnowing observations, will be performed from the computed functional attributes (g).

We now focus our attention on consequences and issues of the present modeling of ecephys work, focusing on Figure 3.2 and on the preceding description of activities once more. Our high-level description is not concerned with low-level operational details (*e.g.,* what equipment is used in Activity 2, or in which order the data is iterated while performing in Activities 3-4) and although it constitutes a logical sequence (the outputs of an activity are inputs to the next one), iterations are not precluded at any particular point. Nonetheless, iterating makes more sense in the places indicated by upstream arrows, such as deriving observations again because the partial results are sparse or seem ill-defined, or designing new experiments after completing a publishing cycle. The self-referential arrow involving exploration only emphasizes that exploratory data analysis (EDA) is in itself an iterative process that keeps going until either no more interesting patterns are apparent in the data or once enough evidence has been collected to support all proposed inquiries.

In terms of how suitable a design study methodology is to this problem, we first highlight that activities 1-2 are completely outside of the *viz* scope. Activity 3, as we previously discussed, depends on a finite set of parameters that may assume an infinite set of values, thus dramatically affecting the data volume to be analyzed during Activities 4-5. Very importantly, Activities 3-4 (and, to some extent 1-2 as well) — which are traditionally reported in *materials and methods* sections of domain-specific publications — are performed only because *they have to be*, being necessary mean to the ultimate goals represented by Activities 5-6. The fact that some but not all of the focal participants eventually developed analysis programs in Matlab for performing derivation activities only adds to that interpretation. We are not considering cross-dataset activities here, such as comparing different species, hemispheres, brain areas, or test subjects — these are all constant dimensions in our dataset (except for hemisphere), so we regard them as metadata — but nothing stops these aspects from being addressed by ecephys studies, which would have profound implications in terms of visualization design.

Some of the verbs used in our activity sequence (*derive*, *explore*, *characterize*), are employed in MUNZNER [96]'s action-target task typology to describe low-level visualization tasks (*i.e.,* cognitive/interactive operations on well-defined attributes encoded into visual representations, optionally producing outputs) but we use them with a higher-level meaning to describe work phases that may include several chained visualization tasks. For instance, the focal participants reported different procedures for finding plausible observations (activity 4), like browsing all recording sites from a given recording session, or browsing multiple depths of a single electrode (which goes through multiple recording sessions). Nonetheless, the activities show some fractal resemblance to MUNZNER [96]'s tasks, in that they have a *what-why-how* structure, *what* being the nouns and *why* being the verbs and the *how* being unspecified sequences of lower-level tasks. What we call *activity* is thus similar to the concept of *user goals* proposed by LAM *et al.* [97]. As a matter of fact, we initially attempted to use their taxonomy to describe the activities, but we found out their goals presented an intermediate abstraction level between activities and tasks. Therefore, in order to prevent overspecification, we refrained from using their terminology. However, we do follow their definition of *population*: a subset of data items (summative or individual neuronal observations) that may be obtained by attribute filtering or other selection mechanism.

Finally, an additional set of non-functional requirements might include

- **Maximizing** the size of the derived dataset while keeping high plausibility
- **Minimizing** repetitive labor
- **Reconciling** analyses by different users
- **Facilitating** the act of revisiting the data — it must be possible to reevaluate recording sites with poor results.

3.3.2 Restricting the study scope

The goals and requirements listed previously paint a very broad picture of the work performed by the domain experts. We initially envisioned an all-encompassing solution for satisfying all gathered requirements[2]. The envisioned solution would include several views of the dataset and the possibility of

- Interactively navigating throughout recording sites by filtering on anatomical, physiological, parametrical, or functional attributes, or a combination thereof comparing multiple selections of data (different recording sites and parameters for instance),
- Analyzing relationships between attributes in different categories,
- Annotating and saving selections of observations
- Allowing comparisons across different users' annotations and selections,
- Filling up missing pieces of data with simulated values and using those in comparative selections or summative analyses.

[2] This diagram presents a preliminary version of CWA conducted with the focal participants.

In summary, the envisioned system would combine elements of exploratory data analysis, computer-assisted collaborative work (CACW), visual analytics (multi-dimensional projections and machine learning), and simulation. While seemingly overspecified, the listed features are commonplace in general-purpose EDA and analytics software [75, 76].

Nevertheless, some reasons led us to focus our attention onto Activity 4 (winnowing observations). Aside from secondary pragmatical reasons, the greatest motivation for focusing on this locus of expert work is because it might have the greatest impact when considered in the context of the loop 3-5 (*i.e.,* derive, winnow, explore), since the ability to easily compare, rank, and judge observations (all of these being aspects of winnowing) would allow to tighten that loop and increase analyst awareness about how changing parameters leads to changes in availability of high-quality observations and, consequently, in population statistics. In contrast, when those activities are performed, as they currently are, in an isolated and chained fashion, the amount of work required for reprocessing and comparing alternative derivations (not to mention the temporal gap) impairs the realization of interesting relationships between parametrical and functional attributes. We may even argue that, in the worst case scenario, the need for manually keeping track of different versions of data files and applying different tools at each stage in order to replay the derive–winnow–explore cycle discourages experts from iterating at all, possibly leading to data underutilization.

Furthermore, restricting the work focus proved beneficial for the following secondary reasons:

Prototyping cost – such a complex system would require either complex Wizard of Oz prototyping, use of a generic visualization software that does not support encoding all of the attributes, or an entire development team.

Risks to evaluation – the identified requirements unfold into a rather large set of goals (in the sense of LAM *et al.* [97]), which in turn span several lower-level tasks (in the sense of MUNZNER [94], BREHMER and MUNZNER [95], MUNZNER [96]), therefore creating a complex validation scenario (considering the multi-layered view of SEDLMAIR *et al.* [99])

Computational costs – because of the challenges laid out in Section 1.4, allowing free-form exploration of the dataset would require either impractical preprocessing and storage of a large range of parametrical attribute plus fast retrieval and caching methods, or high-performance processing (possibly via parallelization in expensive hardware) for real-time tweaking of parametrical attributes. We believe those points may be addressed in a future study, as they would comprise contributions to the inner, algorithmic level of MUNZNER [94]'s nested design and validation model.

Data quality – finally, the informal design study conducted with the focal participants revealed, as narrated at the end of Section 3.4.1, that the original state of the V2 Dataset might require reprocessing before moving on to the exploratory aspect of visualization.

At some point, we considered explicitly encoding parametrical attributes in a user interface and allowing the experts to use them for browsing and comparing derivations, including aspects of Activity 4 in our study. However, besides the aforementioned reasons, we ultimately chose to obfuscate these attributes and to only encode functional attributes explicitly with the intent of measuring the relevance of parameters in user choices (see Section 4.1.3 for a discussion of that) and to avoiding biases towards specific parameters that could complicate this analysis.

Before moving on to a deeper discussion of the design study, the next subsection lists a few low-level tasks that unfold from user goals associated with Activities 3-5.

3.3.3 Abstract tasks

Some abstract tasks (in MUNZNER [96]'s sense) that may be related to derive–winnow–explore cycle include:

- **Search** a high-dimensional space (of parametrical attributes like spike detection and sorting options) for plausible neuronal observations at each recording site
- **Compare** different parametrical settings for each recording site or functional attributes of different selections of the entire population
- **Annotate** the best (or a few of the best) parametrical option(s) for each site — we emphasize that
    - There is not a clear definition of *best*
    - The high-dimensional space contains continuous dimensions, so it is not possible to precompute all points, furthermore tweaking the settings for a selection in order to reevaluate the findings must be a responsive operation
- **Overview** observations at the summative (multi-unit) level before visualizing them at the individual level
- **Browse/locate** a neuronal population for further analysis.

In particular, all tasks performed as part of the derivation goal can be seen as possessing a *recording* objective, since it must be possible to replay the exact same sequence of steps — *i.e.,* selecting a corpus of signals, deriving responses and functional attributes, deriving individual observations, and winnowing observations — and to branch it off at any point by changing intermediate parameters. On the technical side, combining step recording and caching would be necessary to make the system viable in a long timescale while achieving responsiveness, after all derivations might be slower than real-time and the amount of unique derivation sequences could grow infinitely. Regarding caching, numerous concerns would need to be addressed, such as expiration policy, time-to-leave, number of caching levels, *etc.*

3.4 Data encoding

In this section, we discuss the encodings of dataset attributes and the process of designing them by means of an informal user study conducted with focal participants. We report the issues and impressions raised during a series of encounters, and how we addressed them and produced improved versions that were combined into a web application employed in the user study. Abstractions discussed in Section 3.2 and Section 3.3 are presented in their final forms but this section includes a historical view of the study.

3.4.1 Informal study: preliminary encodings

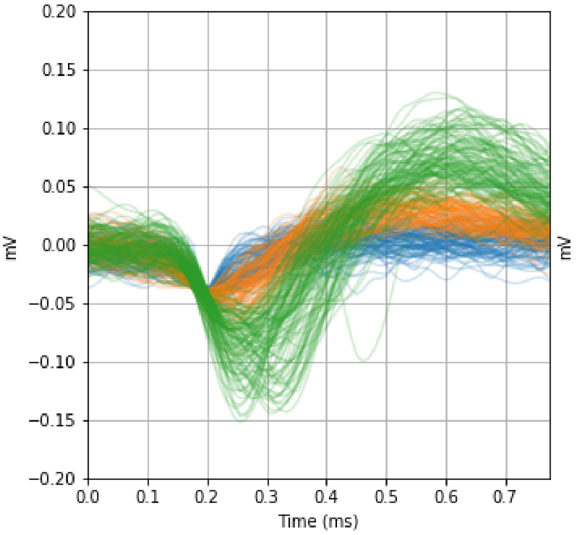
We met with the focal participants in a series of *in situ* and online encounters to discuss how to process, visualize, and interpret the recordings in the V2 Dataset, with the occasional presence of other participants of the late user study. From a chronological perspective, we started by re-implementing the RF mapping method described in Section 2.1.3 and applying it on a few moving bars recordings of the dataset to obtain enough material (tables, plots, statistics, *etc.*) to discuss in those meetings. The initial visualization prototypes were presented on static screens, and were focussed on encodings similar to those used in studies by the LFCOG [47, 56–61], like response maps and polargrams.

Over time, we started to introduce new plots, like waveform plots (common in the context of spike sorting, therefore usually seen as preliminary user goal), and gratings polar diagrams. While allowing domain experts to propose visualizations themselves is a common pitfall in design studies[99], we did this to reduce friction in initial meetings and make room for conversations about the dataset's features. This process allowed to identify a few fundamental attributes and encodings for conducting the neuronal hypothesis winnowing goal. We also discarded response attributes that provided somewhat useful information but cluttered the view as the static visualization panels started to grow large. This included, for instance, trial ranks that helped indicate when "a cell died" during a gratings experiment (more on that on Section 3.4.1) but provided no further clues otherwise. Next, we will discuss each encoding used throughout this phase of this study. Note that the data used in this phase only considered spike sorting results already included in the original dataset, so parametrical attributes are not discussed.

Waveforms

One response attribute that is common across all stimulus types is the set of spike waveforms. Each waveform $\mathbf{w}_i = (w_{i1}, \ldots, w_{in})$ is a vectorized representation of spike format (electric potential samples over a short sampling period) that lends itself to splitting a summative spike trains into a few multiple individual ones based on spike appearance, as described in Section 2.1.2. Commercial spike sorting tools, such as Plexon©'s Offline Sorter™or Omniplex® [80] typically plot waveforms by overlaying voltage-by-sample time line plots and using line colors to encode the observation's identity, as illustrated in Figure 3.3.

Figure 3.3: Waveforms plot. Representing a sample of 100 waveforms of every individual observation in the same summative group. The green waveforms show the highest resemblance to a stereotypical textbook spike waveform [1], followed by the orange ones, with smaller amplitude and longer repolarization intervals.



In addition to plotting each waveform with 50% transparency as a way of improving perception in high density and high overlay regions of the plane, each waveform was linearly upsampled from 32 to 125 points (twice doubling) and the number of waveforms per individual observation was restricted to 100. Despite these amendments, it is hard to ensure the visibility of individuals, specially when there are more than three of them inside the same group or when their waveforms are considerably similar. Therefore, Section 3.4.2 describes a summarized alternative.

Moving bars-specific encodings

The automatic RF mapping method by FIORANI *et al.* [56] described in Section 2.1.3 produces an intermediate functional attribute that is convenient for visual feature identification tasks: the response map, illustrated in Figure 3.4. Recall it is a heatmap that encodes the normalized reverse-correlated response strength at each point of the visual field, calculated by back-projecting the Z-normalized firing rate over all stimulus directions. Beyond obviously encoding point-wise responses, it in fact combines multiple features that follow from its definition and from the mapping algorithm itself:

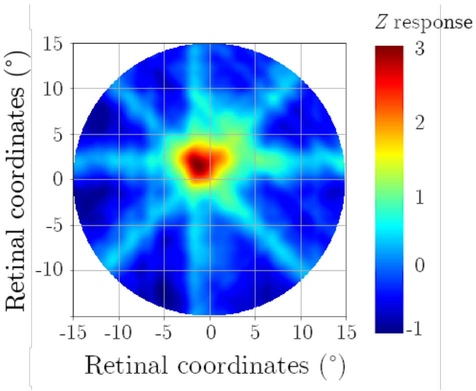
Location The RF is located around the peak in the response map. If there is no RF, than either there is no peak, or there are several local maxima with similar values.

Area Wether the RF is narrow or wide can be inferred from the surroundings of its peak.

Direction selectivity An omnidirectional RF appears as sharp peak surrounded by radial ridges of similar height, whereas a direction-selective RF shows shallower ridges perpendicular to directions in which its response is weaker.

Signal-to-noise ratio If the signal-to-noise ratio (SNR) is high, the non-RF regions of the map should feature low response

Figure 3.4: Response map design used during initial user study

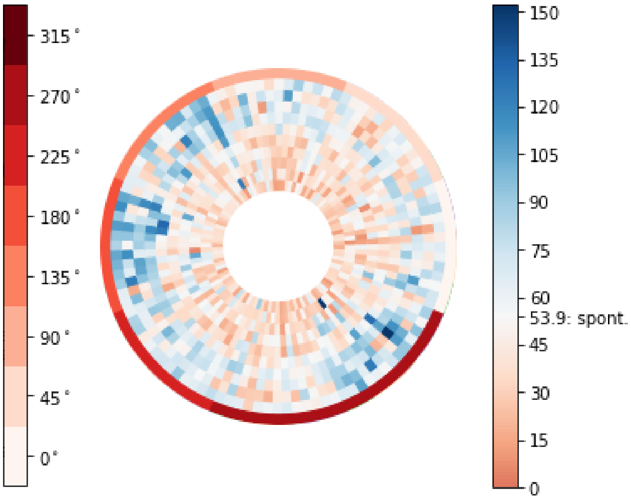


In the initial phase of they study, we encoded response maps using a spectral colorscale — which is highly discouraged for continuous data, specially signed data, due to its lack of inherent ordering, irregular luminance, hue-induced artifacts, and hostility for color-blind people [107] — only as a ways to reduce the number of new encodings presented to the domain experts, since they employed this scale in some of their works. Section 3.4.2 presents them with a diverging colorscale that was used in the final user study.

Gratings-specific encodings

We initially proposed to encode the responses to gratings stimuli by using circular sector plots, with each sector defined by a range of polar coordinates, $[r, r + \Delta r] \times [\theta, \theta + \Delta\theta]$, as illustrated in Figure 3.5. Stimuli were mapped into angles θ, stimulus repetitions (trials) were encoded into radii r, and the response intensity (either the spike count or the firing rate) was encoded into sector colors using a red-to-blue, sequential colorscale with its midpoint (white) set at the estimated basal firing rate.

Figure 3.5: Early proposal for encoding summative responses to gratings stimuli. Radii map to trials, angles to stimulus drifting direction, and sector colors to response intensity. Same-direction stimuli are displaced about the polar axis depending on other stimulus parameters to avoid overlapping, an additional outer ring encodes the drifting direction, and radial lines emphasize the separation between directions.



As described in Section 2.1, for each of the 8 cardinal drifting directions, the gratings stimulus set included 13 combinations of the other parameters targeted at studying different response properties, which motivated the adoption of drifting direction as the primary attribute for mapping stimuli onto the polar axis. More precisely, given a gratings stimulus $\mathbf{s} = (\Theta, A, K, \omega, \mathbf{c}_{\text{FG}}, \mathbf{c}_{\text{BG}})^\intercal$ and some ordering function of the non-direction parameters $o : (A, K, v, \mathbf{c}_{\text{FG}}, \mathbf{c}_{\text{BG}}) \to \{0, \ldots, 12\}$, then $\mathbf{s}$ was mapped to a starting angle of

$$\theta_{\text{start}}(\mathbf{s}) = \Theta + \frac{2\pi}{8 \cdot 13} o(A, K, \omega, \mathbf{c}_{\text{FG}}, \mathbf{c}_{\text{BG}}) - \frac{2\pi}{16}$$

and a final angle of

$$\theta_{\text{end}}(\mathbf{s}) = \theta_{\text{start}}(\mathbf{s}) + \frac{2\pi}{8 \cdot 13}$$

Consequently, any group of 13 stimuli with common drifting direction Θ was mapped into a Θ-centered angular slice $[\Theta - \frac{\pi}{8}, \Theta + \frac{\pi}{8}]$.

As for the radial component, per the adopted experimental protocol (described in Section 2.1), one trial of every 104 stimuli was presented in shuffled order before the next round of trials took place (with a different shuffling). Therefore, while we cannot infer anything about the ordering of stimulus trials with the same radius, given two sectors with midpoints (r, θ_1) and $(r + 1, \theta_2)$, we may affirm that (r, θ_1) happened before $(r + 1, \theta_2)$, which means the radial axis encodes temporal ordering of same-stimulus trials.

Care must be taken with circular sector plots to avoid reduced attention effects in the inner sectors due to their reduced area. We choose that design nonetheless because it allowed for a straightforward mapping between the angular channel and the semantics of drifting direction and because the number of trials was sufficiently small to remove the center portion of the polar plane without cluttering the view excessively.

The subsequent review of this design with the domain experts revealed a few insights — for instance, Figure 3.5 was interpreted as a summative observation with evident orientational preference — however it also elicited a few points of confusion, as the experts tended to:

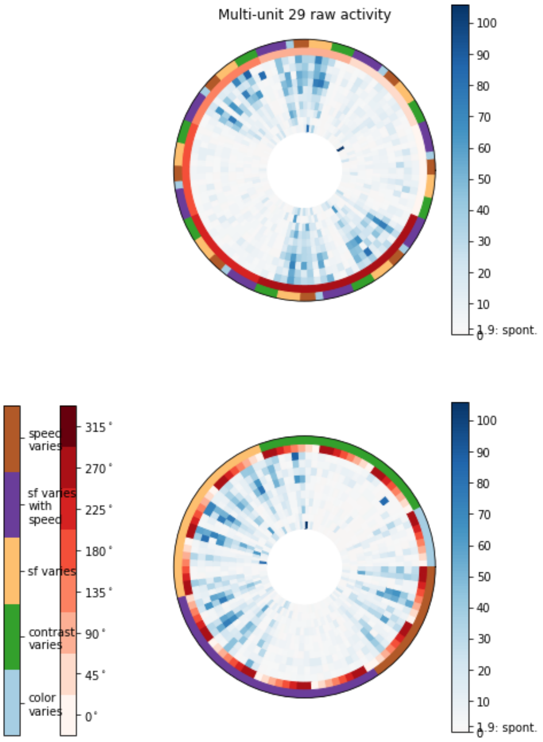
- Believe there were excess angles — they pondered 8 directions and 10 trials should have corresponded to 80 rather than 1040 sectors (10 for each for 104 stimuli), or

- Interpret displaced angles as intermediate drifting directions, rather than different parametrical variations of the same drifting direction, and

- Miss on information about the other stimulus parameters $(A, K, v, \mathbf{c}_{\text{FG}}, \mathbf{c}_{\text{BG}})$ not encoded in the diagram.

We pondered this design possibly failed because it had relied on the expert's memory about the stimulus set, ignoring the fact that they routinely work with a variety of sets designed for different purposes. In the following, we proposed an additional design with extra legends and multiple stimulus orderings to test if the ability to reorder stimuli (possibly in an interactive manner) could ease interpretation of the graphs and facilitate other insights. The result is shown in Figure 3.6: the top plot is the same as before, but the bottom one reorders stimuli based on other parameters first and drifting direction last. An additional outer ring was included in both cases to partially encode the other parameters using a qualitative colorscale.

Figure 3.6: Second proposal for encoding summative responses to gratings stimuli. The mapping from stimuli to polar angles remains the same in the top plot, but prioritizes other parameters over drifting direction in the bottom plot. This difference is emphasized by the outer rings, which cycle alternately between these plots. Variation groups encoded by the outer ring are further explained in Table 3.4.



However, the ordering of the bottom plot was reportedly hard to grasp because the cardinalities of secondary parameter relationships are not trivial in this stimulus set, as shown in Table 3.4 (a similar description was already given in Table 2.1 as well). The color of the outer ring encodes which of the non-direction parameters vary while others remain constant, which defines a *variation group*. Depending on the number of values of the varying parameter, there is a different number of direction cycles; for instance, as contrast varies over $\{0.06, 0.12, 0.5\}$ the cardinal directions cycle three times under the green segment of the outer ring corresponding to contrast variation. However, a fourth contrast value of 1.0 exists in the stimulus set and could be included into this variation group but that is also the contrast value that remains constant while the varying parameter is a different one, therefore the variation groups overlap. To worsen things, the *sf. varies with speed* variation group features two jointly varying parameters, K and ω.

The encoding of trials into the radial axis was initially considered interesting because it endowed the participants to raise hypotheses about data quality and consistency, as shown below, in Section 3.4.1. When the bottommost plot was included,

Table 3.4: Groups of gratings parameters varying with direction. For each of the eight equally-spaced cardinal drifting directions, there are thirteen combinations of the remaining parameters that correspond to the rows in this table, where the first four columns indicate the these parameters' values: cont. (contrast), SF (spatial frequency), spd. (speed), and background/foreground colors. With respect to colors, B/W indicates black/white and Bl./Gn. indicates dark blue/green For each parameter, there is a subset of rows where it assumes several values but all others have constant values, which defines a *variation group*. The last column indicates to which variation groups each row belongs.

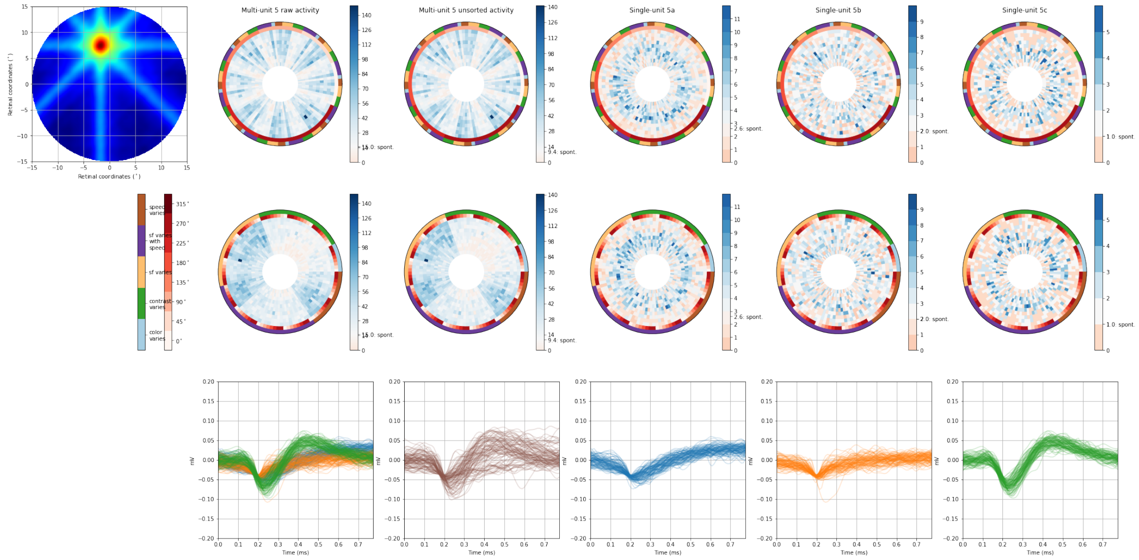
Cont.	**SF (1/°)**	**Spd. (Hz.)**	**Colors**	**Varying params.**
0.06	1	3	B/W	contrast
0.12	1	3	B/W	contrast
0.5	1	3	B/W	contrast
1	0.5	3	B/W	sf., sf./speed
1	1	3	B/W	contrast, sf., speed, sf./speed, color
1	2	3	B/W	sf., sf./speed
1	1	1	B/W	speed, sf./speed
1	1	10	B/W	speed, sf./speed
1	1	3	Bl./Gn.	color
1	0.5	1.5	B/W	sf./speed
1	2	6	B/W	sf./speed
1	4	12	B/W	sf./speed
1	1	30	B/W	speed, sf./speed

the same encodings as the previous design were kept to reduce cognitive effort but that resulted in loss of semantics for the polar axis. This design could indeed originate new conclusions, such as the hypothesis that the summative observation in Figure 3.6 is not only sensitive to a narrow range of orientations, but also selective to higher contrast values and lower drifting speeds. Nonetheless, the aforementioned issues motivated a total of gratings selectivity plots, discussed later in Section 3.4.2.

Overview panel

The stimulus and functional attribute encodings described thus far were combined in a static overview panel, illustrated in Figure 3.7, for visualizing responses to moving bars and gratings stimuli in a single recording site. In the moving bars case, only the summative response map (Section 3.4.1) was represented but in the gratings case, both polar sector plots (Section 3.4.1), and waveforms plots (Section 3.4.1) were included. Furthermore, gratings activity was broken down by individual, with the leftmost column containing the summative observation, its consecutive column containing an *unsorted* observation (that is, the *observation* formed by all spikes discarded during the spike sorting process), and the following columns containing actual individual observations. In the next subsection, we report some hypotheses and issues raised by the study participants based on this type of panel.

Figure 3.7: Early overview panel for a recording site. A summative response map is shown on the top left, followed by four columns, each containing three rows of gratings activity plots. From top to bottom, polar sector plots followed by waveforms plots, and from left to right, summative activity, unsorted activity (corresponding to spikes that were discarded by the sorting procedure), followed by individual activity. Various detail panels are available online at `https://lcg.gitlab.io/pasad/dsc/journal/f8d10c1c21/lab/E007/reports`.



Early findings and hypotheses

Instances of the detail panel previously presented were generated for all recording sites in a selection of four gratings-moving bars recording pairs, and some panels were discussed with the study participants over videoconferencing sessions. Due to the sheer amount of screen space taken by the panels, in each discussed case, we present here only the summative activity rearranged into a single row, to the left of the respective response map[3]. Table 3.5 summarizes these cases and points to the relevant figures. Each figure is accompanied by a summary of the explanation provided by the focal participants.

Let us report and reflect on a few claims and concerns put forth by the domain experts, and a few realizations of our own. They confirmed and reiterated the importance of showing the following information: response maps for judging the existence of RFs, and waveforms for judging the overall quality of data. They also claimed that analyzing multi-unit activity (at least in the moving bars case) tends

[3] The full figures are available online at `https://lcg.gitlab.io/pasad/dsc/journal/f8d10c1c21/lab/E007/reports/report-223-246.run.html`.

Table 3.5: Categories observed by domain experts when inspecting early encodings of functional attributes

Figure	Description
Figure 3.8	Response map reveals a sharp RF and gratings plot reveal orientation-tuning.
Figure 3.9	Response map reveals a sharp RF but gratings plot reveal no tuning.
Figure 3.10	Response map reveals a sparse RF and gratings plot reveal orientation-tuning.
Figure 3.11	Response map reveals no RF but gratings plot reveals orientation-tuning.
Figure 3.12	Response map reveals no RF but gratings plot reveals omnidirectional tuning for spatial frequency and speed bands.
Figure 3.13	Response map reveals no RF and gratings plot suggests electrode drift or cell death amidst the experiment.

Figure 3.8: Evidence of an orientation-selective RF. The summative activity of two sorted units reveals orientation-selective RFs that respond to both moving bars (with a sharply-located center) and gratings stimuli. Only one other recording site in a total of 64 available in this experiment pair was found to have similar properties.

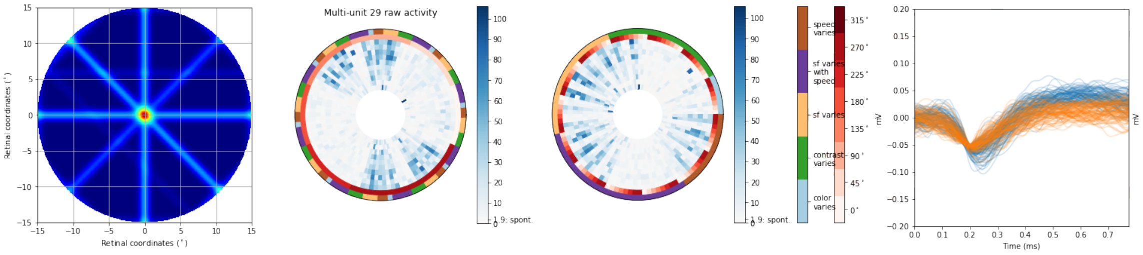
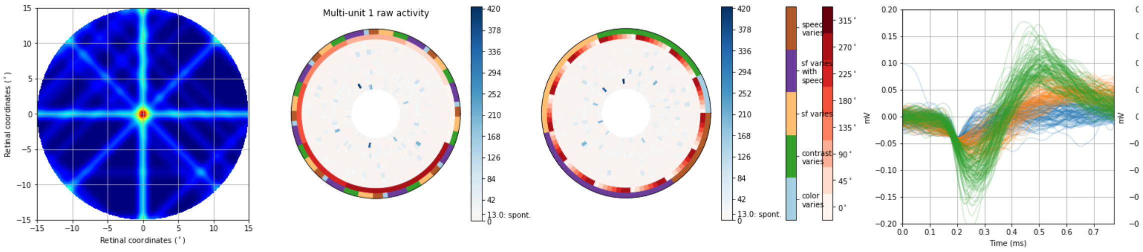


Figure 3.9: Evidence of a RF with no gratings tuning. Although RFs were found using moving bars stimuli and feasible waveforms were recorded during gratings stimulation, no tuning was observed with gratings, neither in single, nor in multi-unit activity. The salt and pepper pattern — that is, mostly low activity, with very few evenly spaced trials showing an elevated spike count — might hide a considerable amount of information, because it stretches the colorscale.



to give results that are consistent with later identified single-units, so in principle, it makes more sense to first overview the summative activity when performing a winnowing goal, and to eventually choose to visualize individual details.

Nonetheless, analyzing information in individual trials in the gratings polar sec-

Figure 3.10: Inconclusive RF evidence with weak gratings tuning. Response maps do not show conclusive RF locations but gratings activity is slightly orientation/speed/spatial frequency-selective.

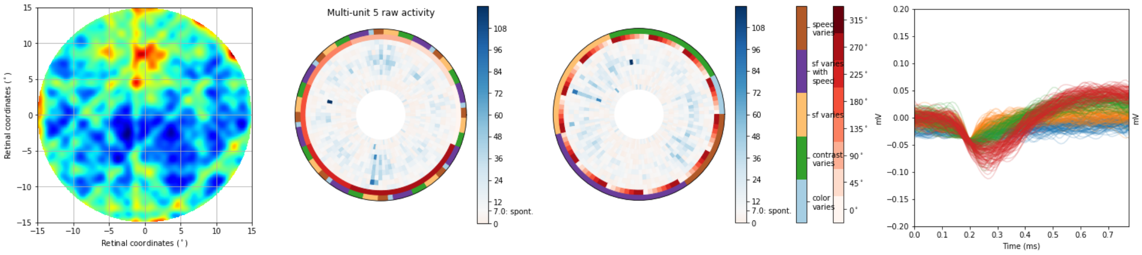


Figure 3.11: Inconclusive RF evidence with time-located orientation tuning. Response maps provide no RF indication but gratings activity is slightly orientation-selective, although higher response is only seen in the final trials.

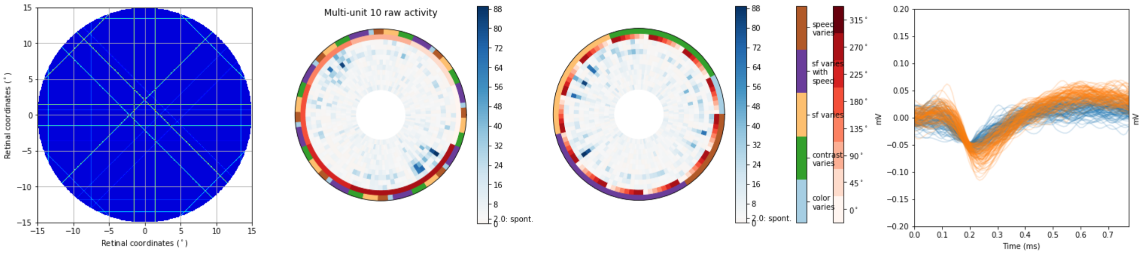


Figure 3.12: Inconclusive RF evidence with a soft spatial frequency selectivity. Response maps provide no RF indication but gratings activity is slightly selective to some ranges of spatial frequency and speed, regardless of drifting direction.

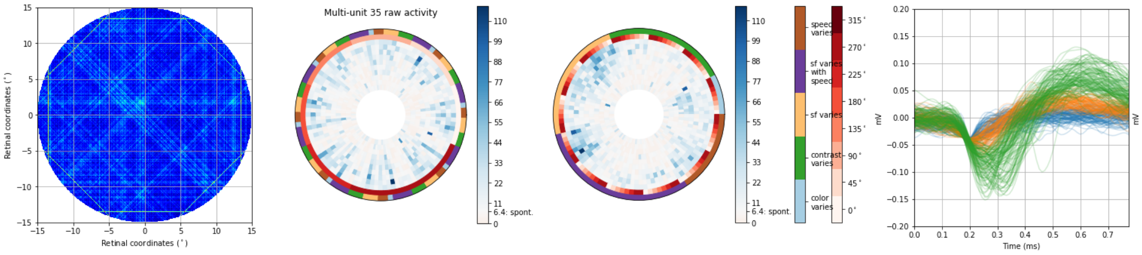
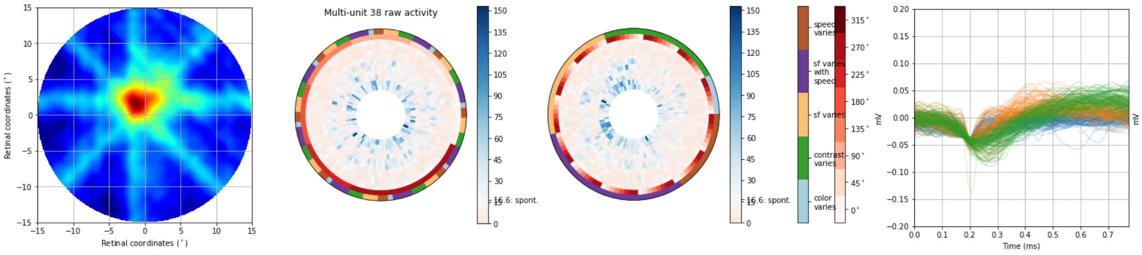


Figure 3.13: Evidence of large RF with time-located, inconclusive gratings response. This cell presents a well-located RF, given its response map. However, gratings activity suggests the electrode may have drifted or the cell may have died during the experiment, since responses drop after the first two trials.



tor plots proved to be challenging, for a few reasons. First, the encoding of more than three attributes (stimulus direction as angle, trial as radius, and response intensity as color) by using a partition of the angular channel with color marks to represent additional stimulus attributes (variation groups) was not easy to grasp. Second, it is already expected that trials fluctuate, therefore the ability to spot a seemingly rare condition like the previously mentioned "silence periods" did not pay off the cost of increasing the amount of information and making stimulus preferences hard to see. In other words, the "show as much information as possible" mantra did not come true.

The focal participants were remarkably impressed by the sparsity of plausible RF observations in the four pairs of recording sessions involving moving bars and gratings stimuli. We had not re-implemented spike sorting until this point, so the individual response profiles visualized so far used spike sorting information recorded in the PLX data files themselves. Despite the combined efforts we made to improve the curation of the V2 Dataset, the metadata available was not detailed enough to reconstruct all aspects of preprocessing, and the former graduate student that had been responsible for it had graduated and was no longer available to collaborate in this effort. Both participants raised serious concerns about the quality of spike sorting, and, User 4 emphasized that spike detection might have used poorly chosen SPKC signal thresholds (see Section 2.1.2). We later realized that in all cases analyzed by the specialists, most spikes responsible for noticeable features in the summative profiles were marked as discarded by the spike sorting procedure, which supported their concerns. That can be seen in Figure 3.7, where the motion direction-selectivity shown by the summative profile is completely retained in the unsorted panel and the actual individuals (5a and 5b) show no selectivity at all, with just a few high spike counts trials.

It was then that the user goals of deriving and winnowing neuronal observations (introduced in Section 3.3.1), allowing users to reprocess raw recordings and reevaluate the resulting populations, became evidently necessary for a successful *viz* system for this domain. After all, spike sorting is expected to be an error-prone phase, so counting on the possibility of needing to adjust it is more realistic than conceiving a downstream work methodology in which this step is performed only once early on, then never again.

We purposefully left some attributes out of the overview panel — *i.e.,* gratings indices like the CVO, CVD, OI, and DI, and RF latency (moving bars case) — because they were standalone scalars rather than 1D/2D functions that would not stand out in the face of other juxtaposed panels. On a few occasions, the domain experts asked us about the values of these attributes, so we wondered wether they could be really relevant for decision-making. At some point, we investigated the possibility of feeding all known functional attributes to multidimensional projection algorithms like k-means and t-SNE [108] as a way to locate clusters of look-alike neuronal activity, inspired by visualization tools of other application areas within the neuroscience field[109]. However, since the tests performed did not result in any meaningful categories that could resemble those identified manually in Table 3.5, these attributes remained excluded from the design space.

3.4.2 Enhanced and new encodings

We started our study with the typical *viz* premise that all available information should be presented. However, throughout the informal study, it became clear that the most interesting information was getting cluttered by excess, non-informative features, which led us to redesign some of the encodings. Meanwhile, we conducted a deeper task analysis of the domain (see Section 3.3) and concluded that a user study about the user goal of winnowing observations would be a promising contribution in the direction of facilitating the domain work and providing insights about the data characteristics. The user study is discussed in deeper detail in Section 3.5, while we will discuss below the encodings used for that study, and how they differ from the ones in the informal study.

Waveforms

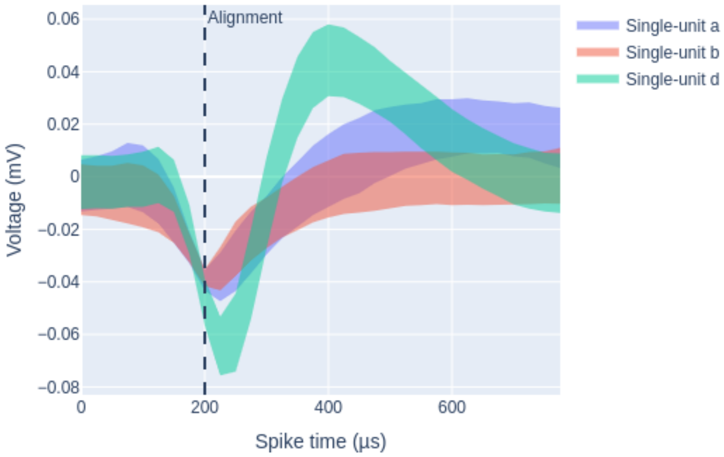
However informative the overlaying of waveforms may be, it tends to produce occluded views in some fairly frequent conditions, namely:

- With more than a few hundred spikes,
- In the presence of too many outlier spikes, or
- When the number of individuals is greater than two, or when their distinction is subtle.

Therefore, in order to present a cleaner visual while still providing clues on the quality of the sorting (in terms of separability and cohesiveness of clusters), we proposed to encode each individual observation's waveforms $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ by filled regions representing their most common shapes. Figure 3.14 illustrates the proposed alternative. Each filled region is defined between one standard deviation above and one standard deviation below the mean waveform. The waveform mean $\overline{\mathbf{w}}$ is computed by averaging the vectors, that is $\frac{1}{m}\sum_{i=1}^{m}\mathbf{w}_i$, and the waveform deviation $\mathbf{s}$ is computed as the vector of sample-wise standard deviations, that is $s_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(w_{ij} - \overline{w}_{ij})^2}$. Furthermore, transparency is added to the regions and a vertical dashed line indicates the sample time used for aligning the waveforms during spike detection (see Section 2.1.2).

Unlike other functional attribute encodings, it does not make sense to plot summative waveforms, since their filled regions would overlap all individual regions and provide too much of a spread region to be useful for any judgement.

Figure 3.14: Waveform summary of one summative and three individual observations. The upper and lower limits of each filled region are defined as one standard deviation above and one standard deviation below the mean waveforms of each observation.



Moving bars-specific encodings

We introduced few changes in relation to the previous response map design, namely: we changed the normalized response encoding to a blue-to-red, diverging colorscale, which encodes the response sign (excitatory *vs.* inhibitory) into hue (red *vs.* blue, respectively), and we plotted the RF contour, computed as the set of points around the peak with response greater or equal than 75% of the peak's value and annotated this value on the plot. The result is show in Figure 3.15.

So-called *polargrams* are also frequently used in studies by the LFCOG, so we included them in the web tool, without any changes to the original design [56]. Polargrams are polygons in polar coordinates, $\{(\theta_i, r_i)\}_i$, where the azimuth, θ_i, varies with stimulus drifting direction, and the radius, r_i, represents the proportion of spikes fired when the bar moves inside the RF's contour. The result is show in Figure 3.16.

Gratings polar diagrams

The representation of gratings responses was redesigned after the impressions collected in the informal user study. The main issues with the previous design (Section 3.4.1) were:

- Ambiguous/confusing encoding of stimulus variables into azimuths
- Excessive elements packed into small-area polar sectors
- Loss of semantics when changing stimulus ordering along the polar axis

Figure 3.15: Response map plot. It encodes the Z-normalized response $r_Z(x, y)$, estimated by the back-projection algorithm described in Section 2.1.3, in each point (x, y) of the visual field using a continuous diverging colorscale defined between -3 (dark blue, most inhibitory response) and 3 (dark red, most excitatory response). The isoline at the response threshold $r_\varepsilon = \frac{3}{4} \max_{(x,y)} r_z(x, y)$ that defines the RF is drawn with a solid black line on top of the map, and the value of r_ε is annotated near it with a line pointing to the RF's center.

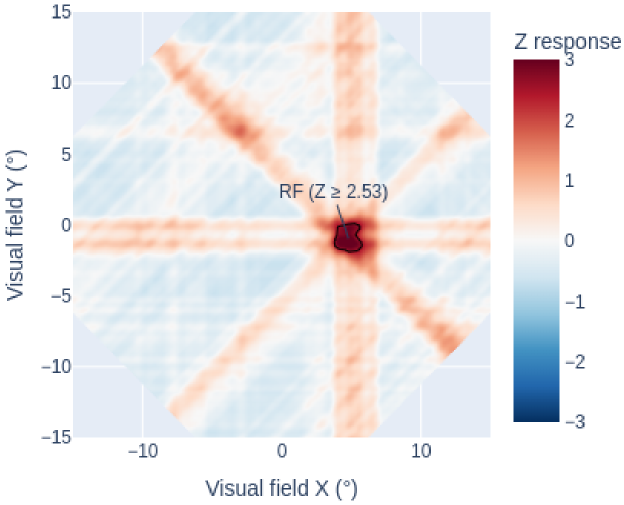
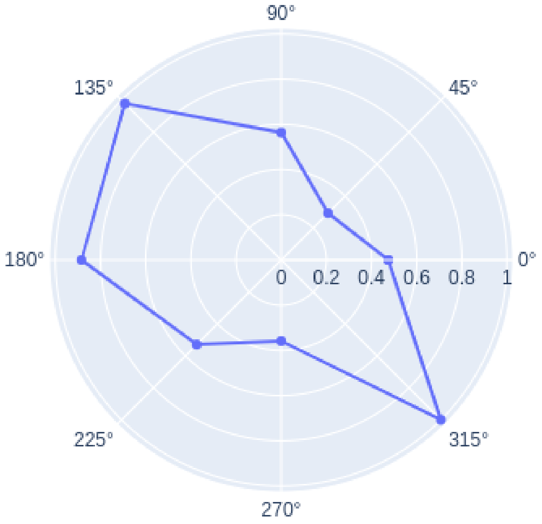


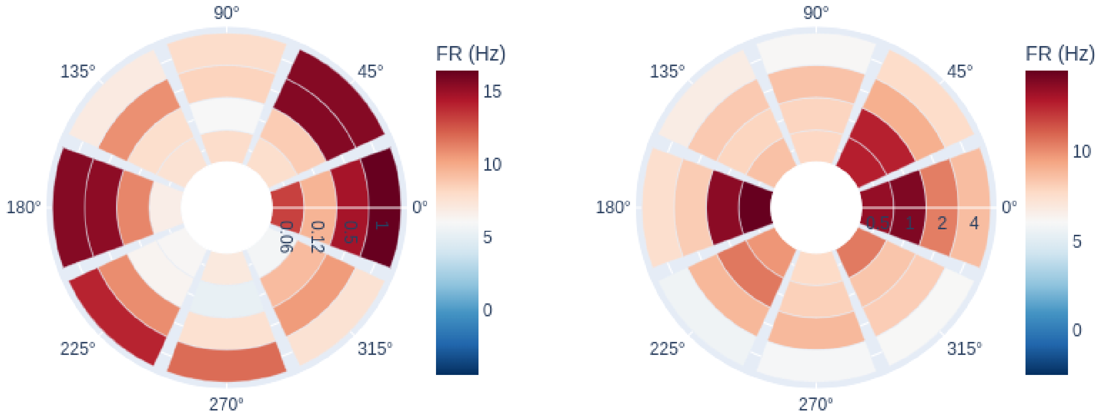
Figure 3.16: Receptive field polargram. Each point encodes the fraction of spikes that are fired while the bar passes through the estimated RF region relative to the total number of spikes fired when the bar moves in the respective direction. Those quantities are averaged over trials.



- Arguable value of mapping individual-trial spike counts rather than mean firing rates — despite the reported hypotheses of electrode drifts and cell death, the cases that motivated these were rarely seen and the domain experts argued that analyzing individual trials is rarely interesting due to the expected nature of inter-trial variability.

The proposed alternative, illustrated in Figures 3.17-3.18, employs four faceted plots that encode the secondary stimulus parameters into the radial axes while keeping drifting direction in the polar axes and representing the mean firing rate in the polar sectors' colors. By replacing per-trial spike counts with mean firing rates, these plots were turned into 2D tuning curves (discussed in Section 2.1.1), where tuning is represented by color, rather than height on the y-axis. Since the radial parameters are unevenly distributed across their respective ranges and sector area was not intended to encode anything, we laid them out as ordered categorical variables, *i.e.,* using equal radial steps. Due to parameter overlap across variation groups, as previously mentioned in Section 3.4.1, some sectors in different plots encode the same mean firing rate. In each plot, the colorscale was centered at the estimated basal firing rate, which causes the darker blue tones to encode physically impossible, negative values. Therefore, most plots are predominantly red, with blue tones indicating inhibitory responses, similar to response maps.

Figure 3.17: Grating response encoding: direction, contrast, and frequency



(a) Responses to gratings stimuli by drifting direction and contrast level

(b) Responses to gratings stimuli by drifting direction and spatial frequency

As a consequence of the proposed design,

- If the cell is direction/orientation-tuned, that should be consistent along all plots
- If the cell is selective to narrow ranges of the other parameters, that will be evident in their respective plots.

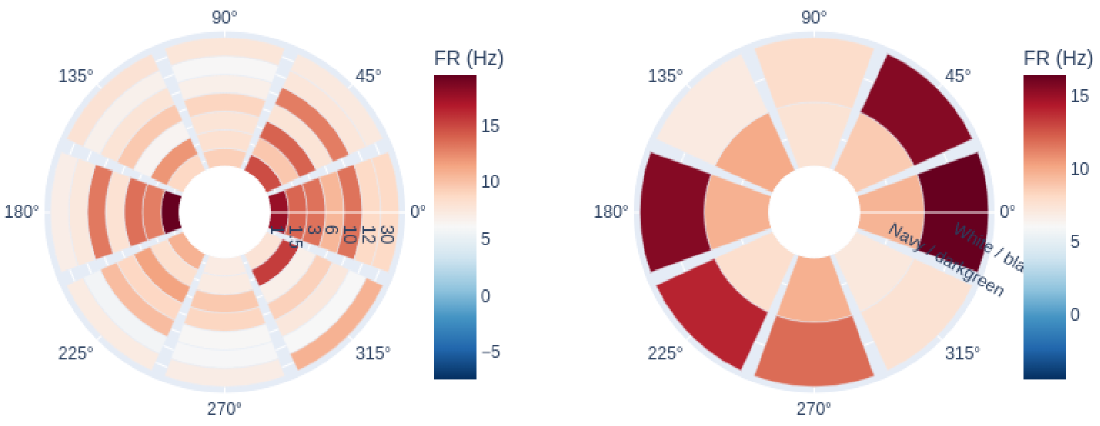
Downsides of this approach include:

- Grasping directional/orientational selectivity requires inspecting multiple plots.

Figure 3.18: Grating response encoding: direction, speed, and colors

(a) Responses to gratings stimuli by drifting direction and drifting speed

(b) Responses to gratings stimuli by drifting direction and stripe colors



- The stimulus set is not a dense Cartesian product of the parameter ranges, but if that were the case, the proposed faceting would not cover co-variation of parameters (that is, the design is not extensible to dense stimulus sets) — in fact, we already left out cases in which spatial frequency and speed vary together.

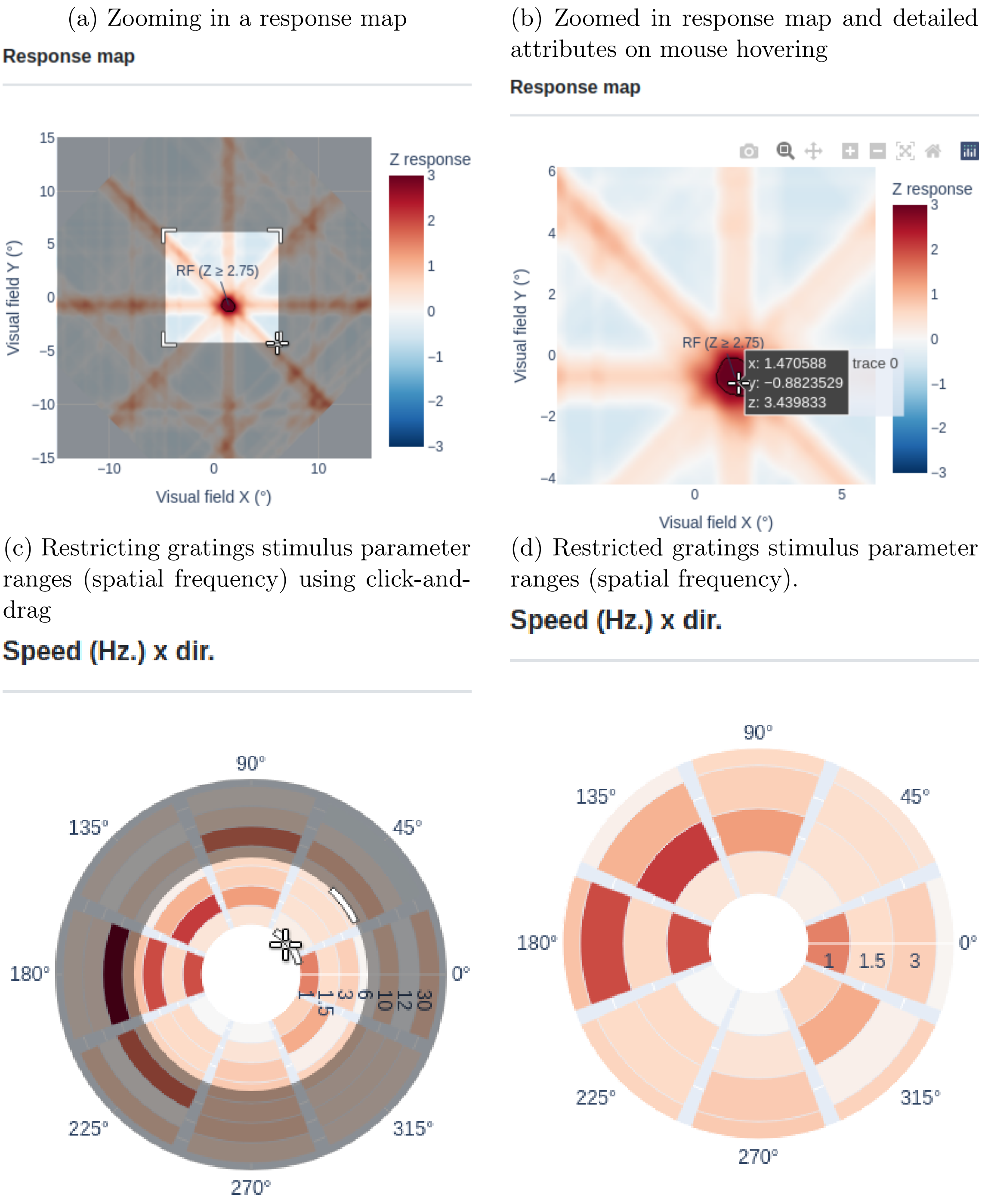
Responsiveness

All enhanced encodings of stimulus-response relationships just described were meant for usage in an interactive web application rather than in a static report (as the ones in the informal study). Therefore they were implemented using standard plot widgets from the Plotly graphing library that includes support for zooming and panning idioms, where applicable (Figures 3.19a-3.19b). Additionally, mouse-hovering over graphical elements, *i.e.*, pixels in response maps, points in polargrams, polar sectors in gratings plots, or curve points in waveforms plots, brought up a pop up window with encoded values and other attributes (*e.g.*, all stimulus parameter values, in gratings plots) — examples in Figure 3.19b and Figure 3.20a. In gratings plots, zooming and panning were replaced by domain range selection using click-and-drag, as illustrated in Figure 3.19c-Figure 3.19d. Finally, waveforms plots, the only type of plot in which individual rather than summative activity was presented, allowed to turn on/off individual waveform regions (Figure 3.20b).

3.5 User study

After the informal design study with the focal participants cast doubts on the quality of the initial preprocessing (data derivation) of the V2 Dataset, we decided to

Figure 3.19: Details-on-hovering and click-and-drag zooming response panels



(a) Zooming in a response map

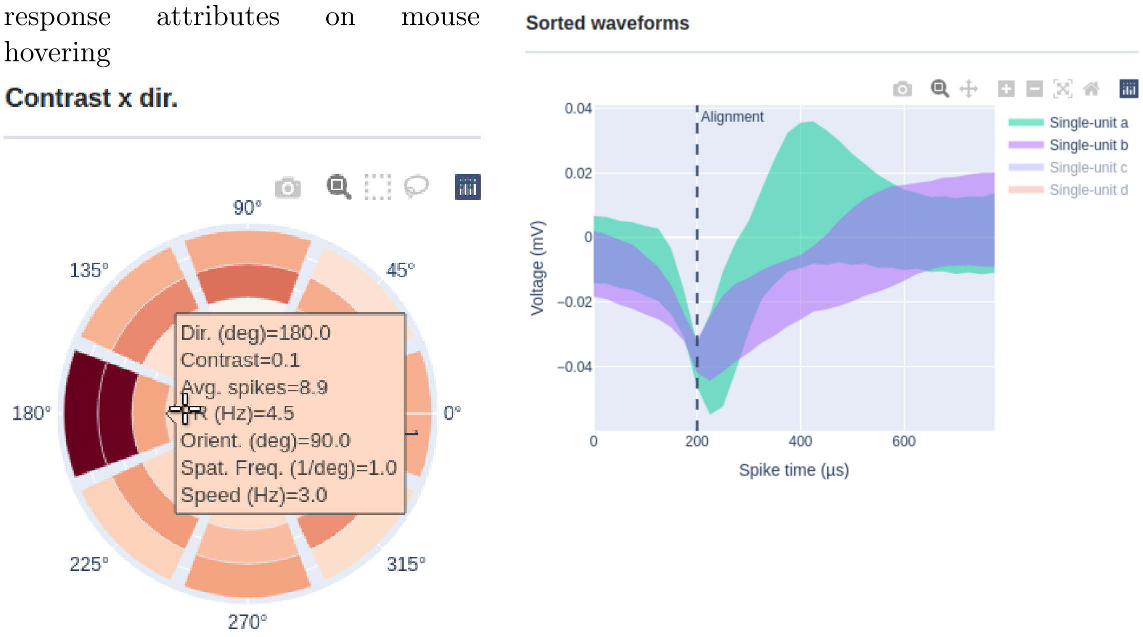
(b) Zoomed in response map and detailed attributes on mouse hovering

(c) Restricting gratings stimulus parameter ranges (spatial frequency) using click-and-drag

(d) Restricted gratings stimulus parameter ranges (spatial frequency).

focus on the deriving and winnowing goals (Figure 3.2). However, even the most modest selection of parametrical attributes would result in a handful of alternative hypothetical neuronal observations per anatomical location, and the dataset already contained 256 distinct positions (32 electrodes on each of two hemispheres, in four moving bars-gratings recording pairs), so the total number of hypotheses that a user would need to winnow before being able to conduct any population analysis (exploration goal) would be staggering. Therefore, instead of building a visualiza-

Figure 3.20: Gratings details-on-hovering and waveforms toggling



(a) Detailed gratings stimulus-response attributes on mouse hovering

(b) Turning on/off individual waveform regions.

Dir. (deg)=180.0
Contrast=0.1
Avg. spikes=8.9
FR (Hz)=4.5
Orient. (deg)=90.0
Spat. Freq. (1/deg)=1.0
Speed (Hz)=3.0

tion system allowing free navigation between the user goals (as previously discussed in Section 3.4.1 and illustrated in Figure 3.2), we assembled the encodings discussed in Section 3.4 into a web application for performing summative observation winnowing. By doing so, we nonetheless had to make a few assumptions, and in doing so, reduced the possible scope of validation (that is, we sacrificed any prospects of evaluating ecological validity), effectively conducting a laboratory study rather than a field study as originally intended.

Some implicit hypotheses worth mentioning:

- The selected attributes and their encodings are sufficient to lead to informed decisions by domain experts
- Screening can be done individually — *i.e.,* hiding contextual information and correlations — without loss of objective rigor by the participants.

The main goals of the user study can be stated as:

- Qualitatively validate the data abstraction and encoding,
- Determine the level of agreement between researchers,
- Find reasons for disagreement that could highlight, which functional attributes and their features are critical for decision-making,
- Estimate the uncertainty in downstream population characterization as a function of user disagreement, and

- Obtain a derived dataset for usage in a subsequent exploratory study (Activities 5-6 in Section 3.3.1).

Below, we detail how we selected a slice of the V2 Dataset for this study, then describe details about web application layout, access, and presentation.

3.5.1 Dataset selection

We selected the same slice of the dataset that had been previously discussed with the focal participants because we knew it contained interesting cases. We regarded any biases caused by the focal participants being shown revisiting this data as negligible since: the two studies were separated by around two years; we used different encodings, even if the most striking difference was the colorscale (as with response maps); and gratings encodings contained a different scope of information, and only resembled the previous designs *w.r.t* direction encoding

In order to extract statistical relevance for studying agreement between experts and the predictability of their decisions, both discussed in Chapter 4, we needed a fairly large number of decisions, without severely impacting the ability to attract participants. The initial processing of the dataset, described in Section 2.1 contained over 15000 summative observations and nearly 58000 individual observations but as explained in Section 3.2, the number of parameter combinations is virtually infinite, and the dataset may be viewed as a search space. Therefore we chose a single recording setting with constant depths in each hemisphere. The selected setting contained 64 electrodes (recording sites) but if we accounted for all parameters available for the chosen recording depths, we'd still get nearly 1500 summative observations and over 5600 individual observations. Hence we selected narrower parameter ranges for each stimulus type:

- Moving bars (4 parameter combinations) — two recording sessions on the same left-right hemisphere depth pair, using embedded spike detection with embedded spike sorting (that is, using derivations recorded in the original dataset), either employing waveforms samples or waveforms features as the sorting feature space.

- Gratings (6 parameter combinations) including: embedded spike detection with embedded spike sorting; embedded spike detection with reconstructed spike sorting (see Section 2.1.2) using waveforms samples as the sorting feature space; and reconstructed spike detection (see Section 2.1.2) on wideband/spike-continuous channels (3 or 5 sigma threshold) with reconstructed spike sorting (Section 2.1.2) using waveforms sample as sorting feature space.

In all cases, firing rates were computed by convolving mean spike trains with 150ms-wide Gaussian kernels, and basal firing rates were computed by stringing together all spike trains of inter-trial periods and neutral stimulus trials, in the gratings case. In some cases, no spikes resulted from the selected configuration, which resulted in fewer observation per stimulus type than the aforementioned totals. We ruled out winnowing at the individual level and allowed participants to make decisions only at the summative hypothesis level (that is, multi-units) to reduce the amount of work. Still, 603 observations had to be analyzed over 64 screens (around 9.4 observations per screen, on average, being 5.8 for gratings and 3.6 for moving bars).

The plots in Figure 3.21 break down the cardinality of the hypothesis sets by anatomical position using a matrix format. Each cell is colored according to its physiological attribute (CytOx band type).

3.5.2 Participants profile

Six users agreed to participate in the study, all of them being either active members of the LFCOG, or previous members of it. Five of those users were professors in ecephys departments and one of them was a graduate student. Two other ex-members of the laboratory were contacted but one did not have availability during the study period, and another one never responded.

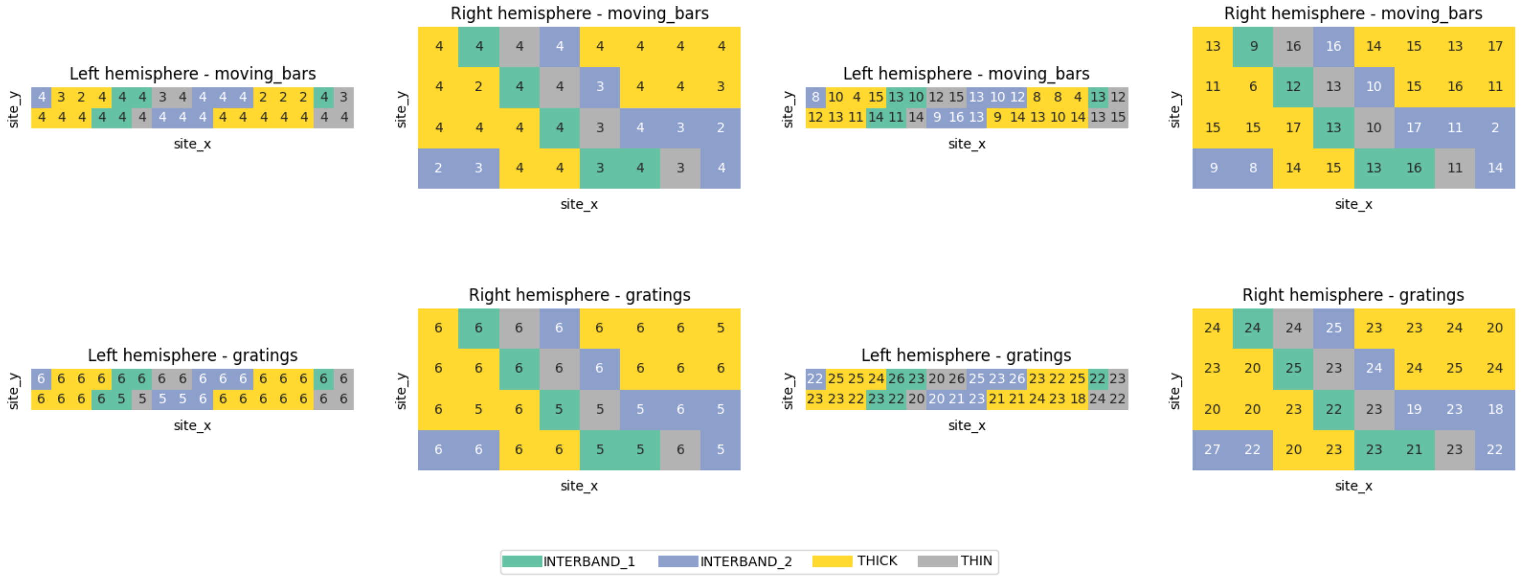
3.5.3 Web application

The improved visual encodings of stimulus and functional attributes discussed in Section 3.4 were combined into a web application[4] with the intent of allowing domain experts to screen a set of summative observations for further analysis, a goal that was presented in Section 3.3. As discussed previously in this section, by omitting and shuffling anatomical, parametrical, and physiological attributes, the application could reduce the effects of crosstalk, biases, and beliefs in the selection process.

The designed application featured one screen (Figure 3.22) for each recording site where overviews of all available summative observations were initially presented on two columns, one for the moving bars and another for the gratings stimulus types. Each row featured the available alternatives, corresponding to different parametrical attributes in shuffled order. Each alternative's identity was indicated by an opaque numerical identifier — corresponding to the alternative's index in a relational database — and the frame's color indicated the user's decision about that

[4] Made temporarily available between February 2022 and August 2022 at `https://vizpike.pasad.dev`. All source-code is open-sourced under the MIT license (`https://opensource.org/licenses/MIT`) and available at `https://gitlab.com/lcg/neuro/v2/vizpike`.

Figure 3.21: Observations by anatomy and physiology



alternative: green for approved, red for rejected, and gray for not yed decided. That component was presented on top of a stack of action buttons represented by icons:

- 🔍 opening up a details view with (approve, discard) in colored label
- ✓ approving the observation
- 🗑 discarding the observation

On the top, additional user interface (UI) components allowed logging out of the system, navigating to adjacent or unfinished screens, and assessing progress and current location in all screens that comprised the study. In the overview panel, most textual elements were stripped off from the plots as a way of fitting them all into a single full-screen-mode, 1920x1080 browser , enough for all screens in the user study — where the maximum amounts of moving bars and gratings alternatives per screen are 4 and 6, respectively — without scrolling, which is an undesirable UI element.

The details panel, illustrated in Figure 3.23 and Figure 3.24, was opened by clicking on details buttons and presented larger versions of the respective stimulus plots, with no textual elements (axis labels and ticks) removed.

In this design, not all functional attributes were encoded, whereas some of them were implicitly encoded, for instance:

- RF latency is one of few one-dimensional attributes, so it would not stand out amidst the plots,
- Mean firing rates, basal firing rates and spike counts are redundant in the presence of Z-normalized responses in the moving bars case,
- Spike counts are proportional to mean firing rates in the gratings case, therefore redundant,
- Tuning curves encode similar properties to what is conveyed by the relative strength between response map ridges and polargrams, therefore redundant,
- Gratings preferred directions and orientations are immediately conveyed by polar sector colors, and
- Other gratings indices, such as direction/orientation indices and circular variances are conveyed by the relative strengths of polar sectors centered in different azimuths.

We took additional measures to make collected usage statistics more reliable. First, all assignment items (individual screens) were shuffled for each participant, as well as the observations within them (that is, the top-to-bottom ordering of observations), as a way to reduce the impact of experience, fatigue, and positional anchoring bias on judging observations in and throughout assignment items. Since the assignment included only the winnowing goal, any contextual information that could lead to correlated decisions and mental shortcuts (*e.g.,* "the ten first electrodes seem to be really low-quality", or "the 3 sigma threshold is worse than 5 for gratings observations"), were omitted. In other words, the anatomical attributes (electrode positions, hemisphere, and MEA depth) corresponding to each screen and the parametrical attributes (relative to spike detection and sorting) associated with each observation were not only shuffled, but completely omitted.

The web application was made publicly available at https://vizpike.pasad.dev but accesses was restricted via username-password credentials shared privately via e-mail with each participant. Figure 3.25 shows the application login screen, and Figure 3.26 shows the user dashboard, presented after a successful login. The dashboard listed the tasks assigned to a user in a table, with each row containing the task's title, its screen count, completion level, and a link for continuing it from the first unfinished screen. All progress in screens was automatically saved at every user action.

Three instructional videos (in portuguese), in a total of approximately 28 minutes, were shared with each participant, advertising the purpose of the study, explaining each plot, how to navigate between assignment items (anatomical positions), and illustrating how to annotate summative observations as *approved* or *rejected*. These videos can be watched on the YouTube platform at the URLs below:

- Vizpike - research introduction
- Vizpike - presenting plots
- Vizpike - navigation and multi-unit selection

Data collection

The following metrics were computed by the system during user sessions:

- Login time
- Type (approval, rejection, detailing), time and target of actions
- Time of arriving/leaving each screen

These attributes are used for analyzing user behavior and computing metrics like the approval rate of each summative observation, the time required by a user to make a final decision about an observation, ant the number of times a user revisited their decision, for instance. These and other metrics of user behavior are discussed in Chapter 4.

Figure 3.22: Observation winnowing screen. From top to bottom, it contains: a login status text with a logout button; a navigation bar with text links for adjacent and unfinished recording sites; a progress bar indicating recording site completion (gray for unfinished, blue for finished, highlighted for current); and finally, on two separate columns, an overview of moving bars and gratings alternative summative observations. The browser was put on full-screen, no-toolbar mode, on a 1920x1080 monitor.

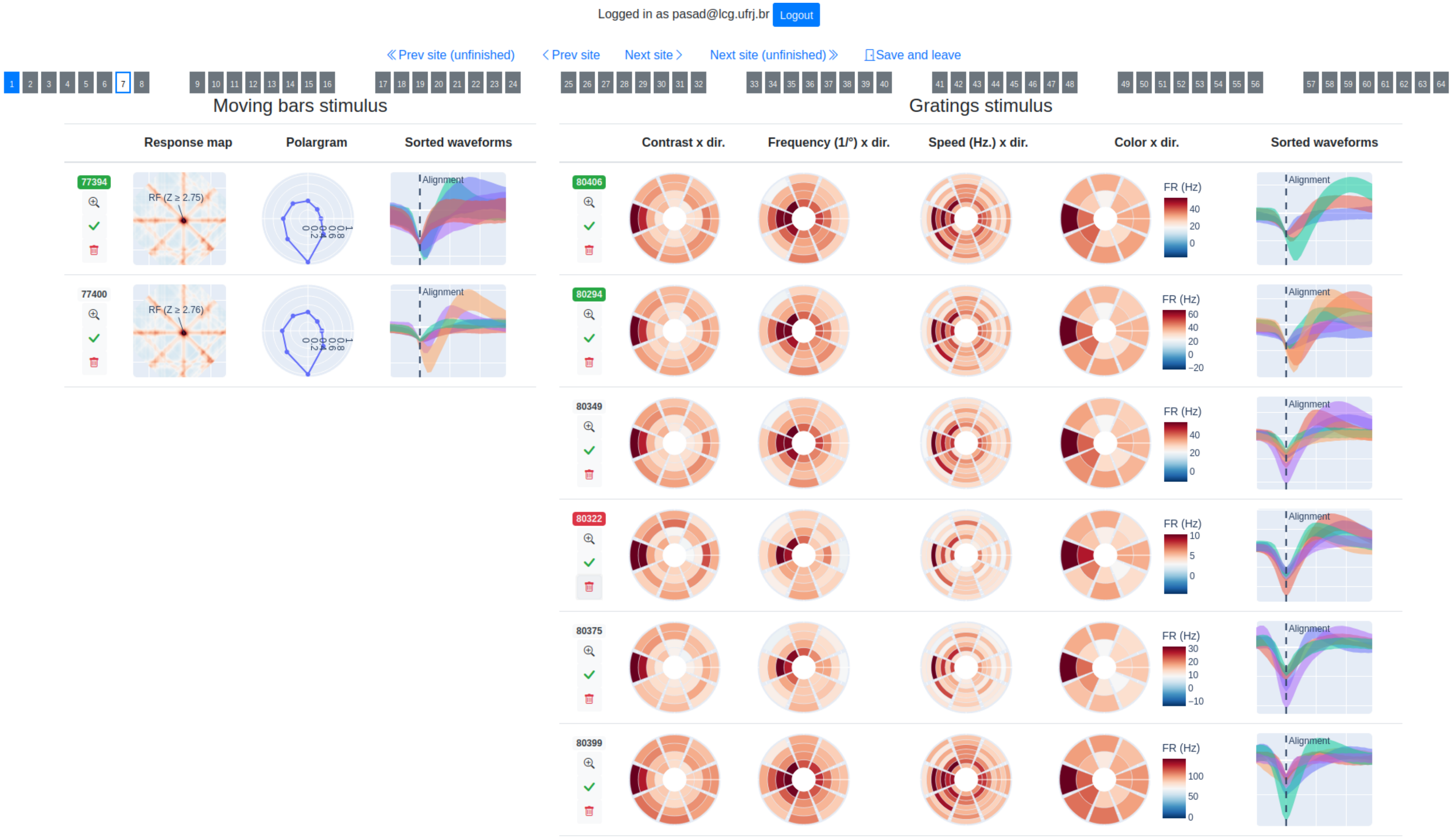
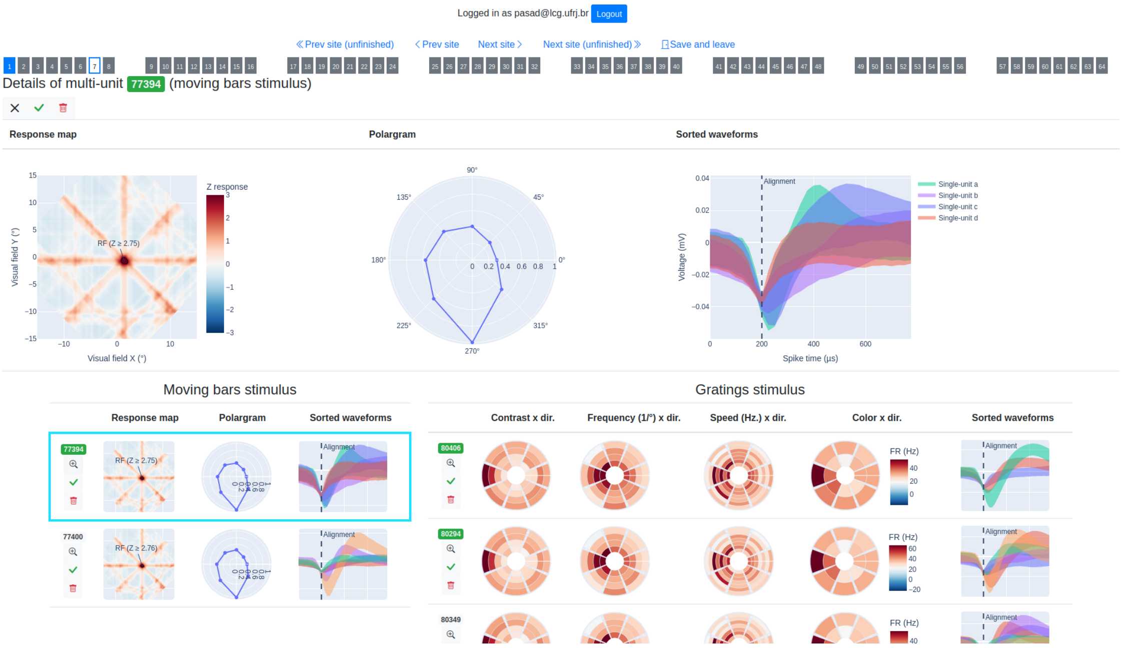


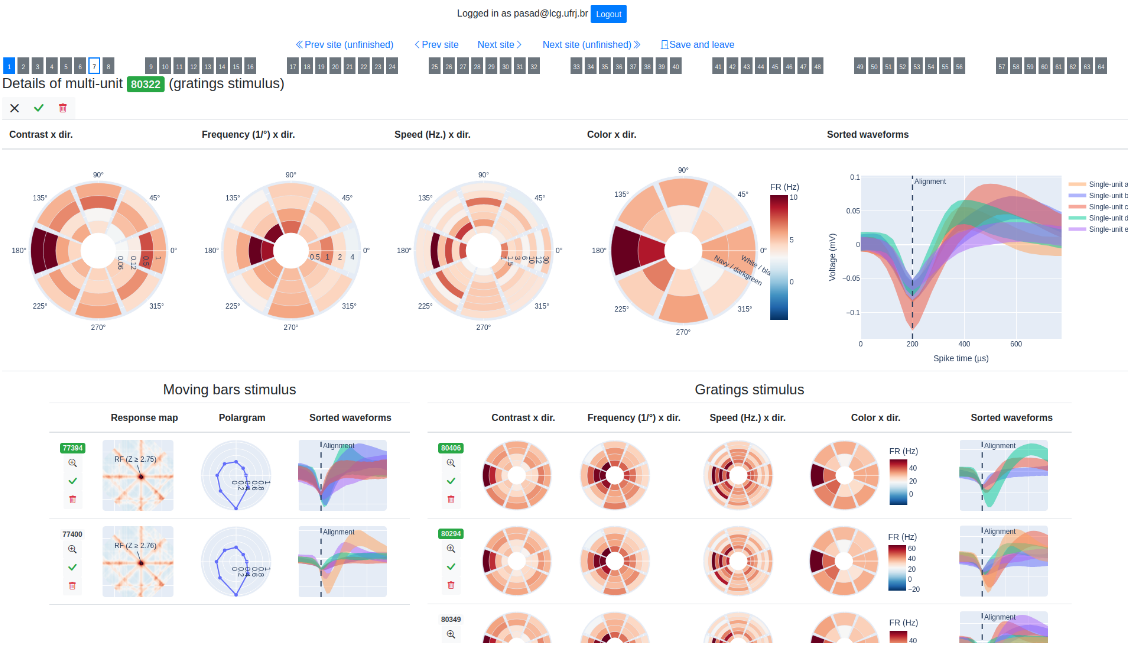
Figure 3.23: Details of a moving bars observation. The details panel is opened between the progress bar and the overview panel, and contains: a response map, a polargram, and a waveforms plot. The detailed observation is highlighted with a light blue frame on the overview panel.



Logged in as pasad@lcg.ufrj.br

Details of multi-unit 77394 (moving bars stimulus)

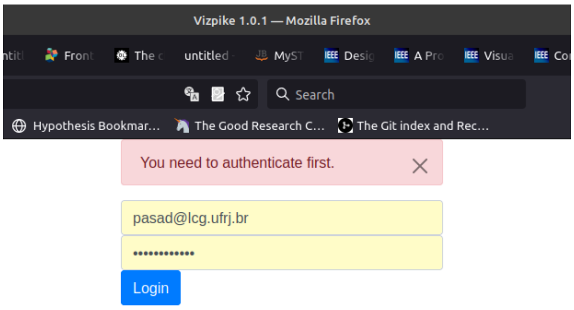
Figure 3.24: Details of a gratings observation. The details panel is opened between the progress bar and the overview panel, and contains: one polar sector plot for the responses to each pair of jointly varying stimulus parameters (from left to right, contrast by direction, spatial frequency by direction, speed by direction, and color by direction), and a waveforms plot. All sector plots are normalized the same way, so a single color bar is presented, besides the right-most plot.



Logged in as pasad@lcg.ufrj.br

Details of multi-unit 80322 (gratings stimulus)

Figure 3.25: Web application login page



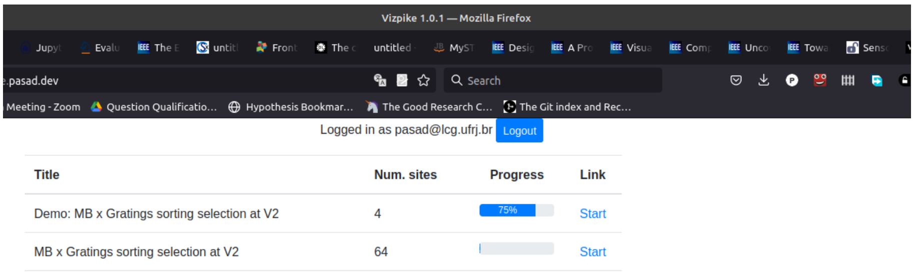
You need to authenticate first.

pasad@lcg.ufrj.br

••••••••••••

Login

Figure 3.26: Web application dashboard. Presented after logging into the system, this table contained the user's assignments, the number of screens (assignment items), their completion level, and a link to resume the assignment from its first unfinished item.



Logged in as pasad@lcg.ufrj.br Logout

Title	Num. sites	Progress	Link
Demo: MB x Gratings sorting selection at V2	4	75%	Start
MB x Gratings sorting selection at V2	64		Start

Chapter 4

Quantitative and qualitative results

This chapter presents and discusses qualitative and quantitative results of the user study reported in Chapter 3. Section 4.1 presents general statistics of the dataset implicitly derived by the winnowing activity (like the counts of approved *vs.* rejected observations), Section 4.2 focuses on user actions and behavior (like the time spent on each screen and the average number of actions required for making a decision), Section 4.3 discusses the similarities and differences in user decisions, setting the stage for a discussion of how to predict user decisions using machine learning methods, which is the subject of Section 4.4. Finally, Section 4.5 covers the evaluative interviews conducted with participants of the user study after they completed the assignments, corresponding to the qualitative part of the results.

4.1 General results

In this section, we detail and discuss numbers of the dataset implicitly derived by means of the user study. These results shed light on the overall quality and usability of the data.

4.1.1 General approval and rejection

Table 4.1 gives the rounded-off percentages of observations that were unanimously approved/rejected, or approved/rejected by the majority of users, or approved by exactly 50% of the users (branded as *disputed*), by stimulus type. The number of observations per stimulus is shown besides the stimulus name.

Let us first consider the naive stance that a 50% majority is an adequate criterion for saving/discarding an observation to/from a cleansed dataset used for posterior analysis (consider it inclusive on approval). Therefore, we are binary-classifying the data as majority-approved or majority-rejected, which leads to approving 58% of gratings observations and 32% of moving bars ones. It seems at first that gratings

Table 4.1: Major and unanimous approval/rejection rates by stimulus type

		Unanimous		**Major**		
Stimulus	**Obs.**	Approved	Rejected	Approved	Rejected	**Disputed**
Gratings	371	22%	11%	30%	31%	7%
Moving bars	232	15%	23%	14%	44%	3%

observations offer higher quality in general but it is noteworthy that 7% of their approvals are disputed, whereas moving bars dispute is less than half of that, in relative terms.

Let us now consider that the majority groups may contain false negatives or positives. Unanimous rejections corresponded to about 16% of the dataset and majority rejections corresponded to about 36% of the dataset, indicating that up to about half of the dataset might need to be discarded. Conversely, unanimous and majority approvals corresponded to and 19% and 24% of the dataset, respectively, meaning that one fifth of the data contains excellent observations and about one additional fourth might be usable. The bottom line is that by considering uncertainty in the majority groups, only 38% of the moving bars observations and 33% of the gratings ones (35% of the entire dataset) would be safely categorized as saved/discarded, requiring some disambiguation procedure to be employed on the rest of the data.

A final stance that we consider is presented by Table 4.2. This time, observations with at most one disagreeing vote are kept in the unanimously approved/rejected groups. That leads to an increase in “unanimous decisions” from 33% to 60% in the gratings case and from 38% to 75% in the moving bars case. Majority groups shrink accordingly while disputed groups retain their size. A disambiguation procedure remains necessary to deal with about 209 observations, or about 35% of the entire dataset.

Table 4.2: Major and unanimous approval/rejection rates by stimulus type. With a tolerance of up to one disagreement.

		Unanimous[1]		**Major**		
Stimulus	**Obs.**	Approved	Rejected	Approved	Rejected	**Disputed**
Gratings	371	33%	27%	19%	15%	7%
Moving bars	232	25%	50%	4%	18%	3%

Whatever the case, if we consider that user approval statistics in this dataset are minimally representative of other datasets, then non-unanimous decisions amount to a considerably portion of data, which has severe implications for the analysis. We will get back to this topic in Section 4.3 and Section 4.4.

[1] Tolerance of one vote off.

4.1.2 Approval by anatomy and physiology

Figure 4.1 shows the average approval rate per anatomical location $\mathbf{a}_i$ and stimulus type. Adjacent locations are encoded as adjacent squares and each matrix represents a hemisphere and its corresponding MEA arrangement for signal acquisition. The top row plots correspond to the moving bars stimulus type and the bottom ones to the gratings stimulus type. Some locations have mean approval close to zero indicating low availability of high quality observations in those locations, specially in the left hemisphere — in which exactly three locations feature zero mean approval of moving bars observations. In the left hemisphere there is a relatively weak correlation between approvals of gratings and moving bars observations on the same location (with Pearson's $r = 0.28$), but that relationship is more expressive in the right hemisphere (with $r = 0.62$).

Note that the differences in hemispheres could be caused by the very shape of the acquisition devices, since recording locations on the left hemisphere have a lower average number of neighboring locations, which may have unfavorably sampled the nearby neuronal populations. All the same, the higher correlation on the right hemisphere echos a statement made by the focal participants during the informal study, that moving bars observation quality is an indicator for gratings selectivity.

Now, instead of considering the mean approval rate, Figure 4.2 shows the ratio of observations that received major approval in each anatomical position $\mathbf{a}_i$, following the same conventions as the previous figure. By removing the often misleading concept of averages and accounting only for more popular observations (even though we know from the previous discussion this is a fuzzy line) we see that: some positions with low average approval indeed would have all of their observations discarded; whereas others with modest average approval would have most of their observations approved, or at least set aside for additional scrutiny.

In this interpretation, the spatial clustering of high approval locations becomes clearer, and a striking number of locations seem to offer no usable observations (40 in the moving bars case and 17 in the gratings case, counting both hemispheres). Correlation coefficients between the fraction of majority-approved observations of moving bars and gratings observations on the same sight are now less significant on the left hemisphere ($r = 0.11$) but still significant on the right one ($r = 0.55$). The number of locations featuring at least one majority-approved observation is 51 for gratings and 27 for moving bars stimuli, which stresses the concentration of higher-quality observations in few recording sites, in the latter case.

Finally, considering the physiological attribute of CytOx band type, there is not much to be said regarding the impact of this attribute on observation approval rates, since there very few recording sites per CytOx band type, limiting any conclusions

Figure 4.1: Mean approval by anatomical position

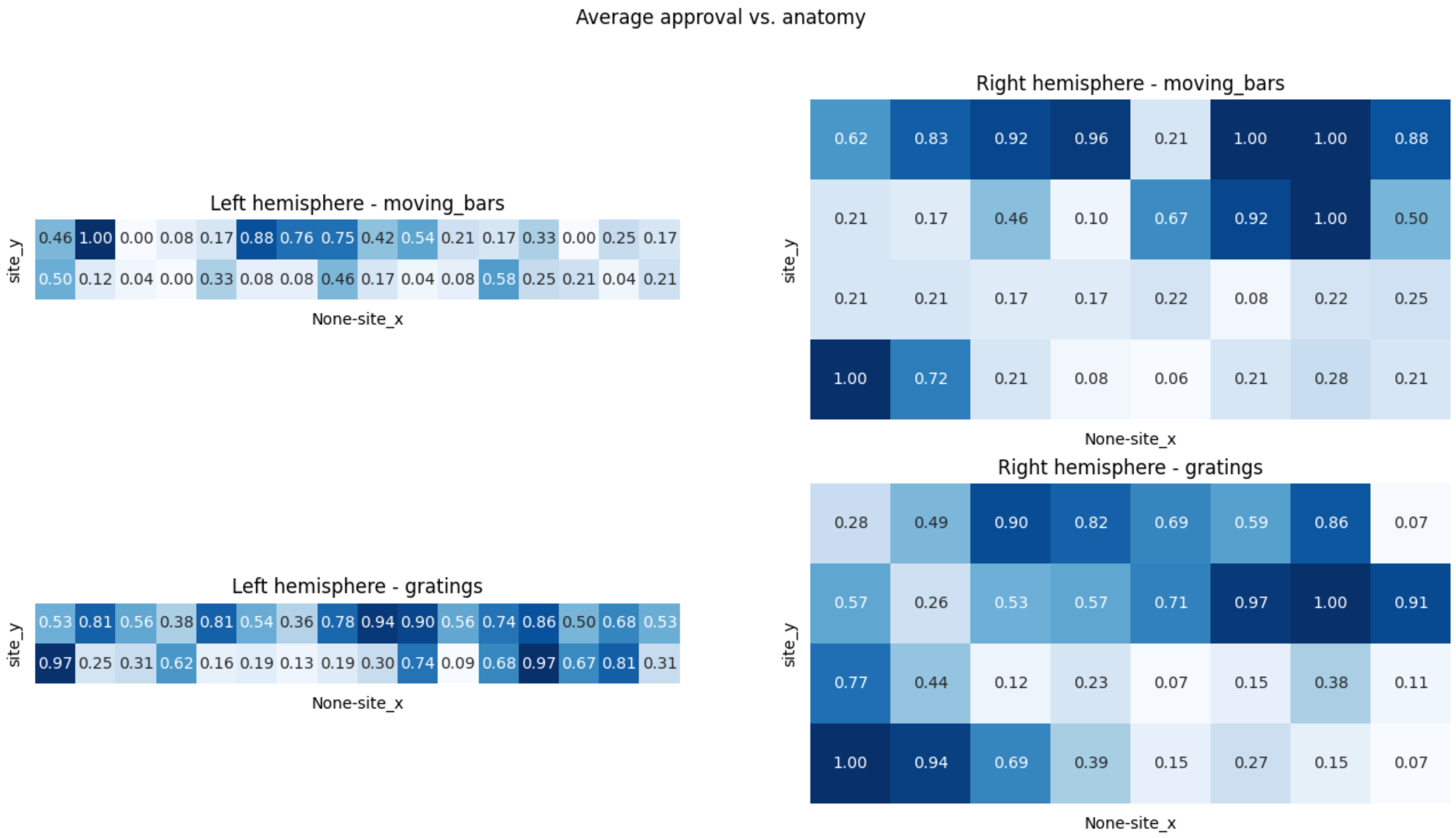
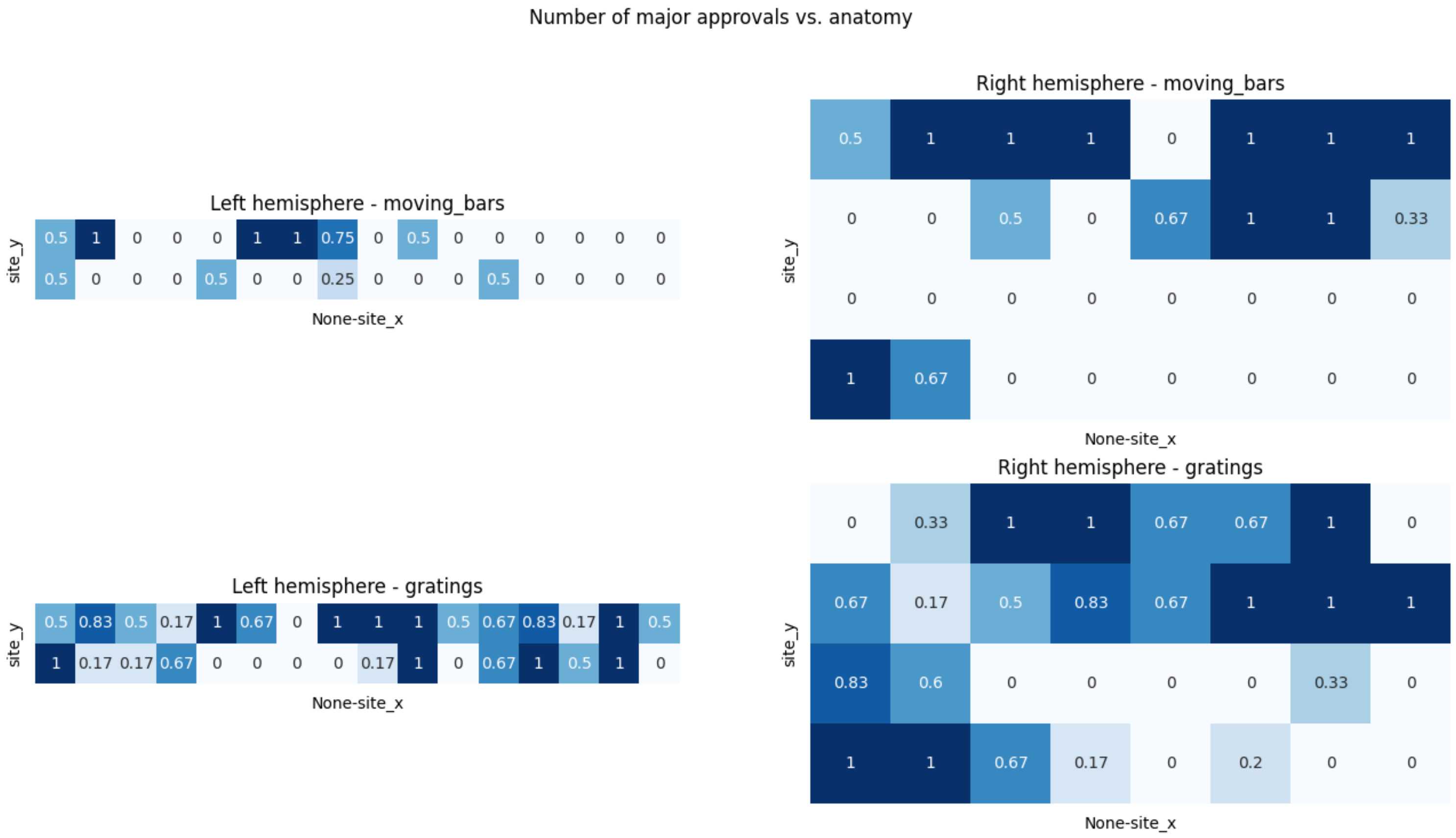


Figure 4.2: Major approvals by anatomical position



we might derived from correlations, beyond the fact that this attribute is already highly correlated with spatial positioning.

4.1.3 Parameters' influence

One of the hypotheses we set for investigating in the user study is whether parametrical attributes influence the approval rate of observations considerably. If confirmed, that would impact the selection of the parameter space when deriving the analysis dataset and, consequently, have a significant impact on the volume of data to be analyzed and, possibly, on the interpretation of the selected population.

We computed the influence of all available parameters on the approval rate of summative observations using the ϕ_K correlation coefficient [110]. This correlation coefficient ranges between 0 and 1 and unlike Pearson's r, Spearman's ρ, Cramér's V, and Kendall's τ, works with ordinal, categorical, and continuous variables, capturing non-linear dependencies, and resorting to Pearson's r in the case of bi-variate Gaussian data. Since many parameters are single-valued or have co-dependencies with others (*e.g.,* a signal level threshold for spike detection does not make sense if the input is a pre-detected spike train, rather than a wide-band signal), we used the $\phi_k = 1$ condition to rule out redundant parameters. Let us look at the results by stimulus type.

Moving bars

In the moving bars case, only two parameters had cardinality greater than one: the sorting feature space (see Section 2.1.2) and the recording session (see Section 2.1). This is mainly due to restrictions in the raw dataset, since wide-band signals were not recorded from the moving bars stimulus, thus restricting the pre-processing possibilities further than in the gratings case. The sorting feature space shows $\phi_k = 0$ with the approval rate. But the recording session shows $\phi_K = 0.2$, with a majority approval rate that varies from 39% to 27% between the first and second recordings, respectively.

We know from previous conversations with the focal participants that, on some occasions, notably when an *in situ* evaluation shows a small number of RFs, a stimulus set is repeated after softly shaking the MEA in place. Therefore, even though allegedly these sessions contain recordings of the same anatomical positions, they can be off by a couple of micrometers, so it is expected that they yield different approval rates. Interestingly, contrary to the expectations of the experimenter who decided to record a second take, in this case the first session gave the best results — that is, it resulted in more observations being majority-approved.

Gratings

In the gratings case, three independent variables stand out with substantial correlation with the approval rate: raw signal type (whether WB, SPKC, or pre-recorded spike train, with $\varphi_k = 0.38$); spike detection threshold (only applicable to the first two signal types, with $\varphi_k = 0.43$); and the waveform alignment mode (at threshold crossing or at global voltage minimum, with $\varphi_K = 0.18$). Table 4.3 summarizes the impact of these variables on the unanimous (one off tolerance) and majority approval/rejection rates. In

Table 4.3: Parameter influence on unanimous and majority approvals for gratings observations

		Unanimous[2]		**Majority**		
Variable	**Value**	Ap.	Rj.	Ap.	Rj.	**Disp.**
Spike detection	Original	11%	9%	7%	5%	2%
	SPKC	34%	30%	15%	16%	5%
	WB	32%	24%	20%	12%	11%
Waveform alignment	Valley	33%	27%	18%	14%	8%
	Thr. cross.	11%	9%	7%	5%	2%
Spike det. threshold	3.0	48%	14%	18%	15%	5%
	5.0	19%	40%	17%	14%	10%

4.1.4 Influence on functional attributes

Figures 4.3–4.4 show φ_k correlation coefficients [110] between parametrical and functional attributes of both stimulus types, by hemisphere. Much like for approval, experiment repetition has a non-negligible impact on some moving bars attributes. Gratings parametrical attributes have a more significant impact on their respective functional attributes. The relations between basal rate or total spikes and the spike detection threshold are trivial, since a higher threshold leads to less spikes being detected. Likewise, the relationship between the spike sorting method and the number of individuals is also trivial, since one of the methods consists in using a pre-recorded sorting information shipped with the original PLX data files. However, other correlations around 0.65 (involving CVD and DI, for instance) and others ranging approximately from 0.25 to 0.45 show substantial impact of parametrical attributes

[2] Tolerance of one vote off.

Figure 4.3: Correlation coefficients between parameterical and moving bars functional attributes by hemisphere

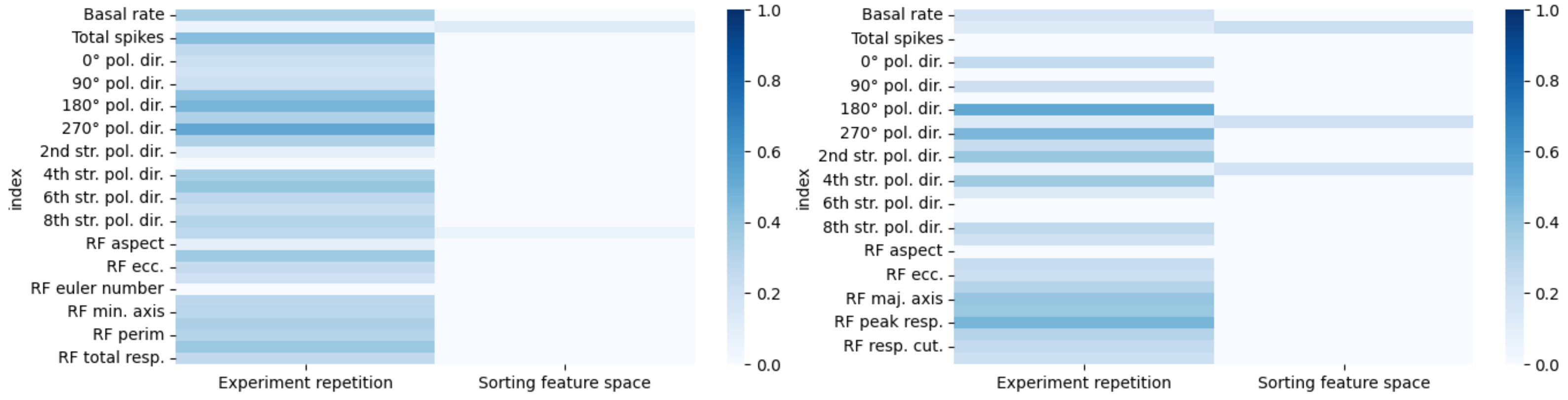
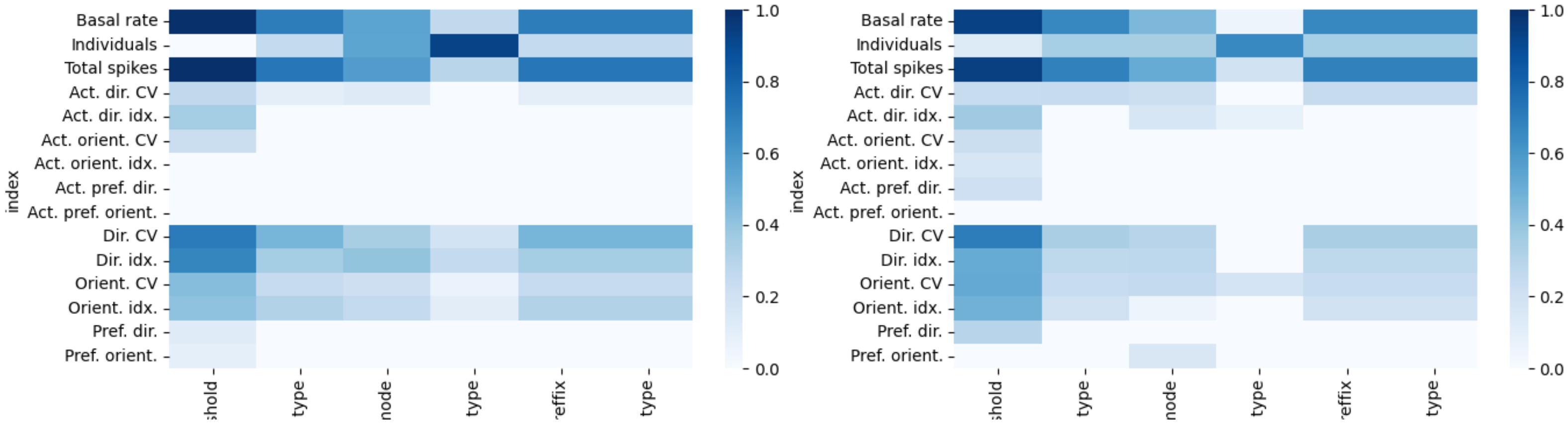


Figure 4.4: Correlation coefficients between parameterical and gratings functional attributes by hemisphere



4.2 User behavior

In this section, we show and discuss some measurements of user actions while performing the proposed activity. As explained in Section 3.5, three types of user actions concerning the targets were registered: approving or rejecting an observation, and opening its details view. Beyond its type, each recorded action was associated with two other pieces of information: a timestamp, and the target's identity. Furthermore, we recorded the time of arrival in each screen and the login date, allowing us to compute the time spent on each screen and to group actions based on their session (that is, everything taking place between two login events). As previously described, each of the $M = 64$ screens in the system corresponded to a single anatomical location $\mathbf{a}_i, 1 \leq i \leq M$, and featured one summative observation $H_{ij}, 1 \leq j \leq N$, for each of the preprocessing parameters $\mathbf{p}_j$ available for each stimulus type. Since both stimulus types are treated similarly, we will use this shorthand notation to address both cases, despite N and the very dimensionality and semantics of the parameter spaces being different in each case.

In this and future sections, we will makes frequent use of the two-sided, two-sample Kolmogorov-Smirnov (KS) test to compare empirical distributions. After all, it is a non-parametric test for assessing whether two independent samples originate from different distributions that is independent of the underlying distribution [111]. Given m samples $X_1, \ldots, X_m$ from a random variable X and n samples $Y_1, \ldots, Y_n$ from a random variable Y, with respective empirical cumulative distribution functions (ECDFs) $F(x)$ and $G(y)$, this test determines whether X and Y have different underlying distributions by verifying that

$$\sqrt{\frac{nm}{n+m}} D_{mn} > H^{-1}(1-p)$$

where $D_{m,n} = \max_x |F(x) - G(x)|$ is the maximum vertical distance between the empirical distributions, $1 - p$ is a confidence level for rejecting the null hypothesis (that X and Y are identically distributed), and $H^{-1}(x)$ is the inverse Kolmogorov-Smirnov distribution. Intuitively, as both sample sizes grow large, if X and Y have an identical, unknown distribution $T(x)$ the empirical distributions F, G should converge to T, and therefore it would become increasingly unlikely to draw samples resulting in empirical distributions too far apart. Although the KS statistic can be used for other purposes, like testing whether a sample comes from a particular reference distribution, or whether $\forall x \, F(x) \geq G(x)$, whenever we refer to the KS test or statistic, we mean the two-sided, two-sample version just defined.

4.2.1 Assignment completion

Five users completed the entire assignment, whereas User 8 judged most of the moving bars observations (97%) but only about half (52.8%) of the gratings ones. We will take this asymmetry into account whenever we compute any statistic over the population of observations, or when training machine learning models, later in Section 4.4.

4.2.2 Session duration

The plots in Figure 4.5 summarize how many times users logged into the system, how long they spent on each screen, and how long they spent in total before logging out. The top plot concatenates all same-session screen times as stacked bars without any filtering, whereas the following plot excludes screen times longer than 5 minutes, which likely represent pauses — after all it approximately corresponds to the 93.6% percentile of screen times. Users 3-6 performed the whole activity set in one long session, while users 7-8 logged in multiple times with shorter sessions. User 5 was fastest according to the bottom graph, performing the entire activity set in about 75 minutes (on average, about 1 minute and ten seconds per screen) while User 7 was the slowest. All the same users 3-6 spent a similar amount of time performing the entire activity set. User 8 is not comparable since they did not perform the entire activity set.

Next, Figure 4.6a shows the individual screen time distributions for each user, facilitating a direct comparison between scree time distributions. It illustrates the time needed to evaluate all parametric variations $\mathbf{p}_j$ of both stimulus types. Again, screen times above 5 minutes (which represent merely 2.6% of registered screen times) were considered outliers, thence purposefully omitted.

The variation in user distributions is considerable, suggesting different task execution patterns. For instance, User 5 shows $[Q1, Q3] = [16\text{s}, 30\text{s}]$, whereas User 3 shows $[Q1, Q3] = [37\text{s}, 137\text{s}]$, where $Q1, Q3$ represent the first and third quartiles, respectively. In part, this variation is explained by the number of times the users queried the details view, since Users 3 and 7, who show the largest median times, are the ones who performed this action more frequently. More specifically, User 3 opened the details view 33 times, User 7 did it 14 times, users 4 and 8 did it only twice, and User 5 never used it. Indeed, if we remove the screens in which the details view was opened at least once and the first action in each screen (which includes screen loading times), we obtain the distributions in Figure 4.6b, of which only User 3 ends up being noticeably different.

In terms of KS statistics, all pairs of users present values greater or equal than 0.2, with most corresponding p-values smaller than 1% — in other words, pairs of ECDFs

Figure 4.5: Time spent by users on each screen they visited. The vertical axis represents the user's login time and the boxes stacked along the horizontal axis represent how long they spent on each visited screen before logging out of the current session. In the bottom plot, screen visits longer than 4 minutes were excluded.

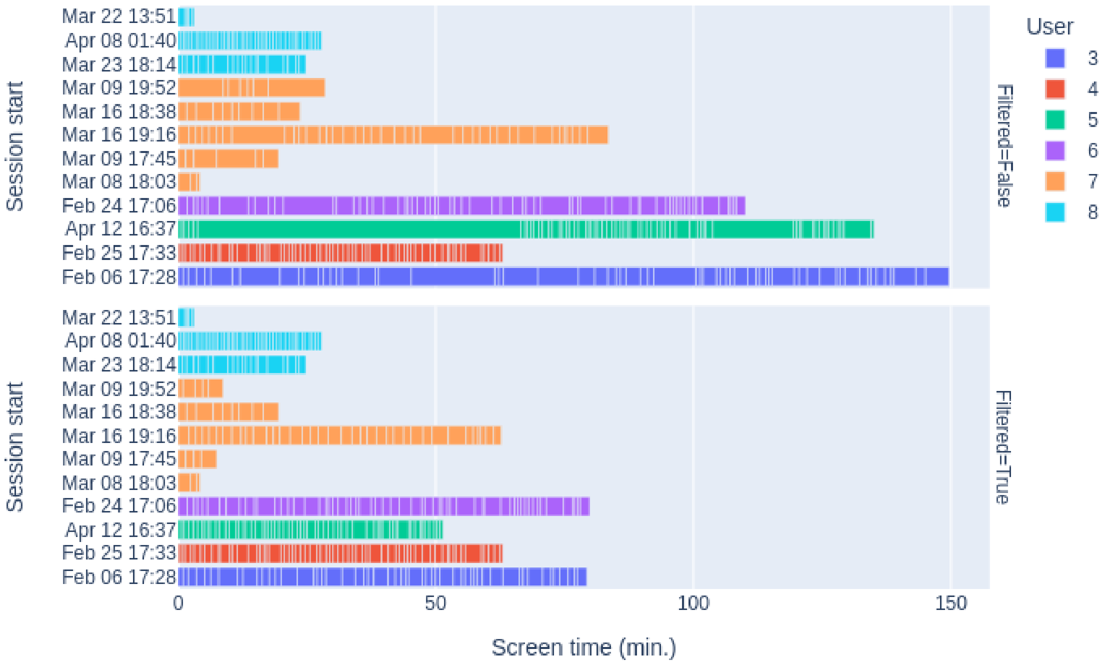
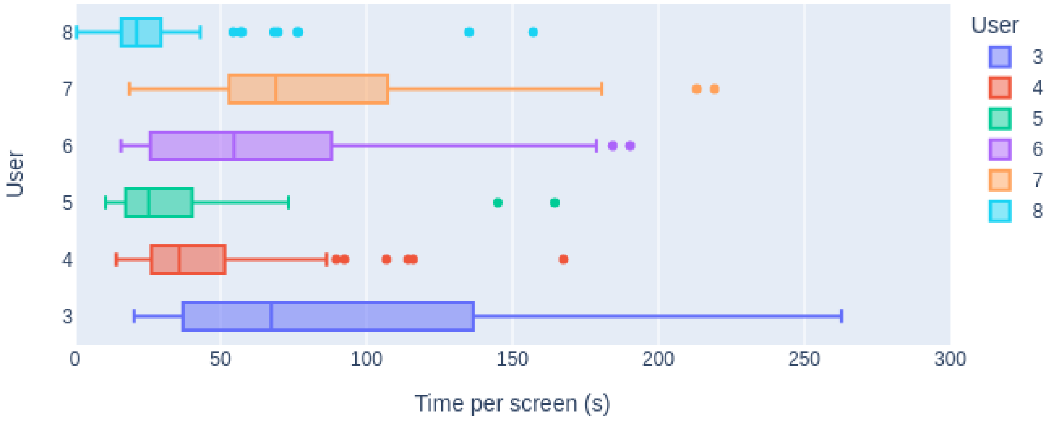
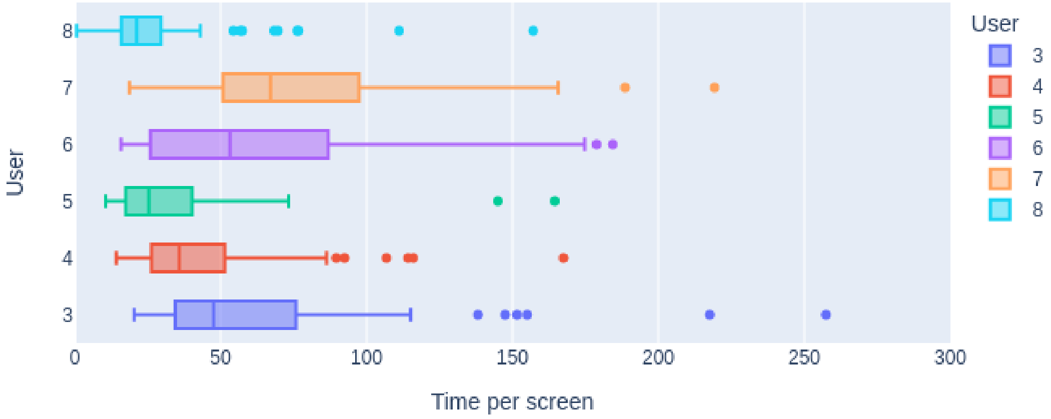


Figure 4.6: Distribution of screen times per user. The first action in each screen was excluded in both cases.

(a) Boxplots of screen times per user



(b) Boxplots of screen times per user excluding details viewing



differ, at some point by at least 20%, with a probability that equal distributions would give rise to such a result smaller than 1%. More specifically, the only cases where p is not at the third decimal place are: users 3/6, with KS $= 0.202$ and $p = 0.115$; users 3/7, with KS $= 0.219$ and $p = 0.074$; and users 5/8, with KS $= 0.219$ and $p = 0.068$.

The following analyses in this section will show that screen time variability is mostly a product of the variability in the time required to judge individual observations by the number of items per screen, and that other factors (like experience, details viewing, and divergence from consensus) do not account for much of it.

4.2.3 Time-to-decision

In general, users did not postpone their decisions in any given location very frequently. In more than 4 out of 5 times, all decisions pertaining a location were made in the first visit of its screen, with less than 30 cases, counting all users, in which a screen was completely skipped to the next one, and less than 50 cases in which a screen was partially fulfilled before navigating to the next one. Furthermore, they rarely spent more than 6 seconds after the last decision before moving on, which corresponds to less than one extra second per observation, suggesting they did not often perform extensive reevaluation of the work done on each screen before proceeding.

Let us analyze how long users took to make a decision (approval or rejection) about a particular observation, which we refer to as time-to-decision (TTD). To compute this metric, we first calculated the time-to-action as the difference between the timestamps of consecutive actions. Then, all actions related to a particular observation — which include viewing details, and possibly pressing approve/reject buttons multiple times — were aggregated to obtain the TTD. Table 4.4 provides TTD summary statistics, aggregated for each stimulus type and action (approve *vs.* reject). Decisions about moving bars observations were mostly taken in under 4 seconds, in adherence to the user reports (shown later in Section 4.5) that claim waveform plots and response maps allow for quick judgement of observation quality, even ignoring polargrams in some cases. Gratings observations, on the other hand, require more careful inspection of multiple tuning curves, which is reflected in a higher decision times in most percentiles. The maximum values are outliers that likely correspond to events such as a user temporarily interrupting work without logging out of the system.

Figure 4.7 provides detailed ECDFs for each user, stimulus and action types. User 5 was the fastest in most cases, making all of their moving bar rejections in under 10 s. Users fall into three broad groups with moving bars-related observations,

Table 4.4: Time-to-decision statistics per stimulus and action types

		TTD (s)						
		Min.	25%	50%	75%	95%	99%	Max.
Stimulus	**Action**							
Gratings	Approve	0.5	2.4	4.6	9.9	31.7	82.4	880.7
	Reject	0.3	1.9	4.2	9.5	27.8	76.4	3691.4
Moving bars	Approve	0.3	1.2	2.0	5.0	28.3	111.9	224.7
	Reject	0.0	1.1	1.5	3.0	10.3	60.1	103.4

considering how long they take to make 90% or more of their decisions: in less than 5 seconds (Users 4, 5, 6); between 10-15 seconds (User 8); and between 25-40 seconds (Users 3, 7).

- Users 4-6 took less than 5 seconds to make over 90% of their decisions, and less than 15 seconds to make the vast majority of them.
- User 8 was more cautious, taking 10-15 seconds to make 90% of their decisions, and 15-25 seconds tom make the remaining ones.
- Finally, users 3 and 7 took 25 seconds to make 90% of their decisions, and 25-40 seconds to make the remaining ones.

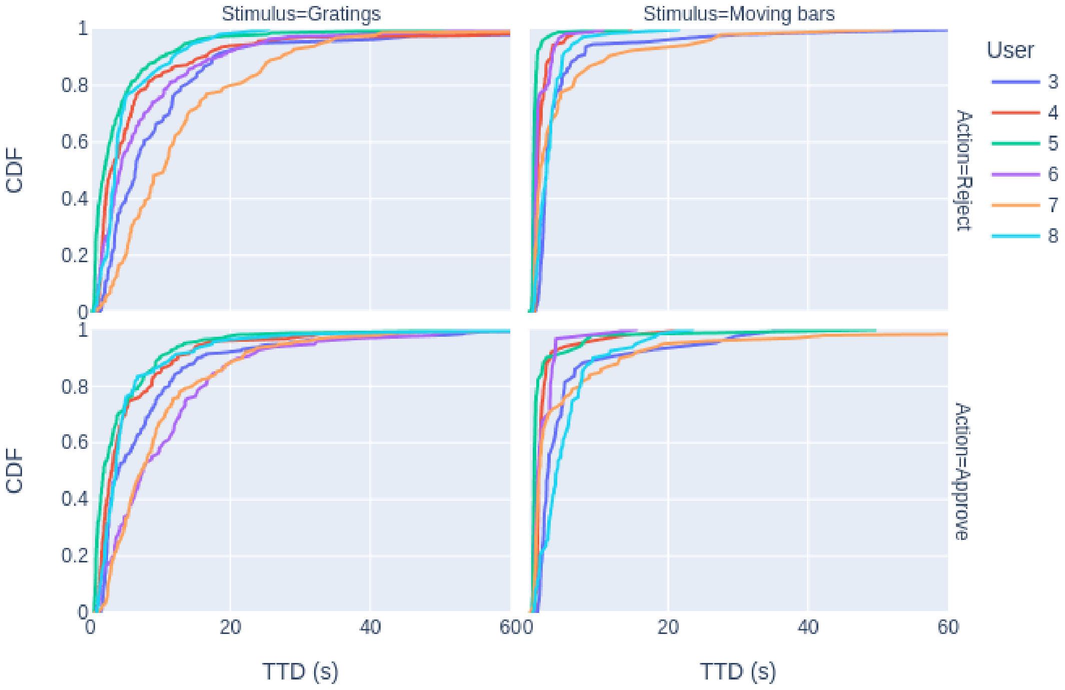
In general, users took longer with gratings-related observations, which was already expected due to the greater number of views associated with each, and again can be fit into three groups:

- Users 4 and 8 were most assertive, making about 90% of their decisions in up to 15 seconds, and most of the remaining ones in 15-30 seconds.
- Meanwhile users 3, 6, and 7 only made 60-75% of their decisions in under 15 seconds, up to 90% of them in 15-30 seconds, and up to 40 seconds to make the remaining 10%.
- User 5 shows a distinctive uniform distribution of time-to-action, which we have not identified the cause for.

In summary,

- For both stimulus types, some supposedly harder decisions may take twice as long as easier ones.
- Some user groups are able to make the vast majority ($\geq$ 99%) of their decisions
  - For moving bars observations in up to 15 seconds,

Figure 4.7: Time-to-decision distributions per stimulus and decision



- But gratings observations will take 25-30 seconds, in the hardest cases.

- Users showing fast decision-making for one stimulus type do not necessarily replicate that behavior for the other type.

Those points open up additional questions:

1. Do approval/rejection decisions have different time-to-action distributions?

2. How much are decisions affected by opening the details view?

3. Is there a relationship between time-to-decision and consensus/disagreement around respective observations?

We will address these three questions in the following paragraphs. Regarding the first question, Table 4.5 shows that some distribution are not easily distinguishable There, time-to-decision CDFs were computed separately for approved and rejected observations, and broken by user and stimulus, as before.

In the moving bars case, all users were consistently faster in rejecting than approving observations. For example, user 5 took less than 5 seconds to perform nearly all rejection decisions, compared to 10-15 seconds to perform nearly all approvals, and user 3 took less than 15 seconds to perform 90% of their rejections, compared to nearly 35 seconds to perform the same percentage of their approvals. Nonetheless, users 4 and 6's rejection distributions are not drastically sharper than their approval counterparts, and despite rejecting faster than approving, user 7 was more cautious when rejecting than remaining users, dedicating 25-40 seconds to the 10% hardest decisions. Interestingly, in the gratings case, there was not, in general, a strong trend to make one type of decision faster. In particular, users 4 and 8 made most of their decisions, of either type, in under 30 seconds, user 5's approval distribution is slightly non-uniform, contrary to its rejection distribution, but it both top at about 40 seconds, and user 7 was indeed faster when rejecting observations, as in the moving bars case, but not by much, since they took up to 30 seconds to make 90% of either type of decision. All the same, decisions about moving bars observations were faster in all cases, except for user 3's approvals.

Table 4.5: Kolmogorov-Smirnov test for per-user approval vs. rejection time-to-decisions distributions. Bold cells indicate cases where the p-value shows a considerable chance that approval and rejection follow the exact same distribution.

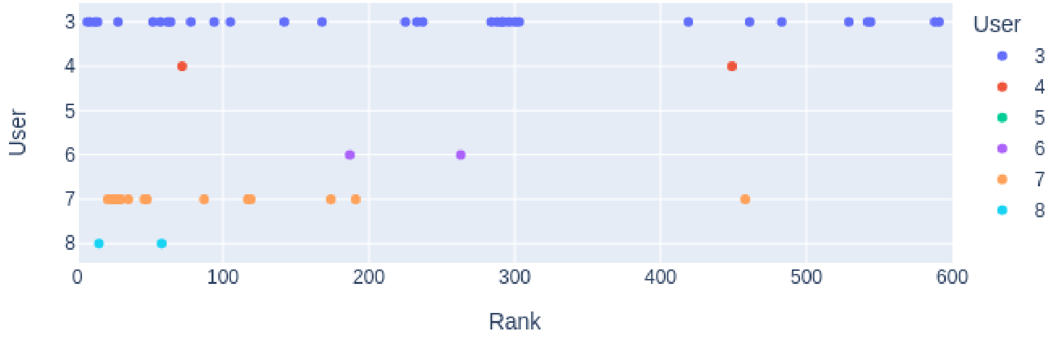
Stimulus	Gratings		Moving bars	
User	KS	p	KS	p
3	0.244	0.001	**0.143**	**0.481**
4	**0.087**	**0.497**	0.289	0.004
5	**0.120**	**0.151**	**0.187**	**0.145**
6	0.246	0.000	**0.249**	**0.070**
7	0.195	0.007	**0.091**	**0.892**
8	**0.106**	**0.710**	0.330	0.000

The distributions we just discussed are not considerably affected by repeated actions (caused by double clicks, for instance), or wavy decision patterns (*e.g.,* approve, reject, approve, *etc.*), since 95% of all decisions involve a single user action.

Impact of viewing details

Pertaining the influence of viewing details on the TTD, the user study contains insufficient data for doing any relevant estimations. After all, User 3 opened the details view more than all other users together but that still only happened in 17 (4.36%) of their gratings decisions and 16 (6.27%) of their moving bars decisions. User 5 did not open the details view at all and users 4/6 did not use it for their moving bars decisions. Figure 4.8 shows that User 3 opened the details view throughout the assignment, whereas User 7 concentrated most usage on the first 200 screens. Considering all 53 times any details panel was opened, only three observations were opened by any two different users, and all of the remaining 50 openings correspond to distinct observations.

Figure 4.8: When users opened the details view. The horizontal axis represents the order in which observations were judged by each user. Each dot represents an opening of the details view for the observation that was judged at that point.



Consensus and divergence

Finally, time-to-decision of particular observations is not strongly correlated with general consensus nor personal divergence around that particular observation. Table 4.6 shows Spearmann rank-order correlation coefficients for both cases. By *general consensus* we mean the fraction of users that approved a particular observation, and by *personal divergence* the fraction of users that disagree with a certain user pertaining a particular observation. Put another way, should we hypothesize that users' TTDs were influenced by whether target observations received a lot of overall support or whether the user diverged from the majority regarding the targets, there would not be a strong evidence in the data to support that claim. The only consistent indication is that these coefficients, with the exception of one, are positive, indicating that highly consensual and highly debatable observations tend to lead to longer decisions.

Table 4.6: Spearmann's rank-order correlation coefficients between consensus/disagreement and time-to-decision

User	3	4	5	6	7	8
Consensus	0.01	0.11	0.12	0.32	-0.07	0.13
Disagreement	0.26	0.20	0.16	0.18	0.16	0.18

Experience's influence

Finally, we consider the impact of user experience (that is, the time they have been using the system since login) on TTD. Table 4.7 shows Spearmann correlation coefficients between TTD and user experience for all approval/rejection decisions and for all users. All coefficients are negative and most have magnitudes below 0.2, suggesting a only slight tendency for faster decisions as experience increases. The strongest trends are associated with users 3 and 6, which tended to approve observations faster as they became experienced, and user 8, who tended to reject observations faster.

Table 4.7: Spearmann's rank-order correlation coefficients between experience and time-to-decision per decision type

User Decision	3	4	5	6	7	8
Approve	-0.28	-0.18	-0.10	-0.41	-0.08	-0.12
Reject	-0.06	-0.03	-0.11	-0.14	-0.02	-0.26

4.3 Decision similarity

4.3.1 Overall similarity

How similar are the decision records of different users? Figure 4.9 presents an overview of inter-user decision similarity (IUDS): for each user pair, we count the number of identical decisions they took and divide by the number of observations that both decided on, thus accounting for User 8's incomplete case and obtaining a number between 0 and 1. Formally, if A_i represents the set of observations approved by user i, R_i represents the set of observations rejected by them, and $D_i = A_i \cup R_i$, then

$$\text{IUDS}[i,j] = \frac{|(A_i \cap A_j) \cup (R_i \cap R_j)|}{|D_i \cap D_j|}$$

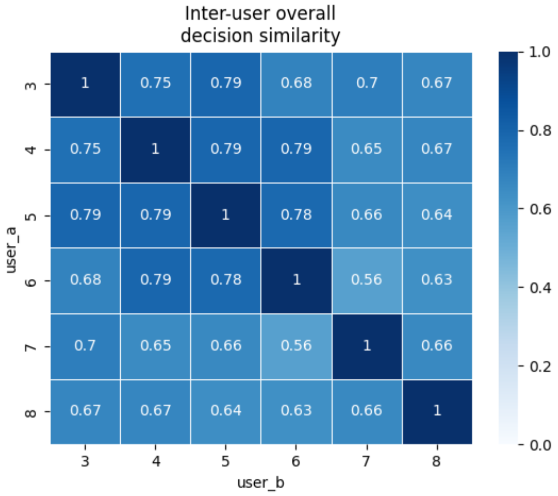
Therefore, from a frequentist point-of-view, IUDS$[i,j]$ is the probability of users i, j making the same decision about any observation, given that they both made a decision about it.

Most pairs of users disagree on about one third of observations (the median IUDS is 0.67). Users 6 and 7 are the most disagreeing, making different decisions almost half of the time. In summary, users judged observations in unique ways.

4.3.2 Per-stimulus, per-decision similarity

The next plot shows a breakdown of IUDS by decision and type of stimulus. Besides filtering observations by stimulus type, when we compute the approval and rejection-related indices we account only for the sets of approvals or rejections performed by

Figure 4.9: Overall decision similarity



any one of the users, respectively. In other words, the approval IUDS for users i, j is

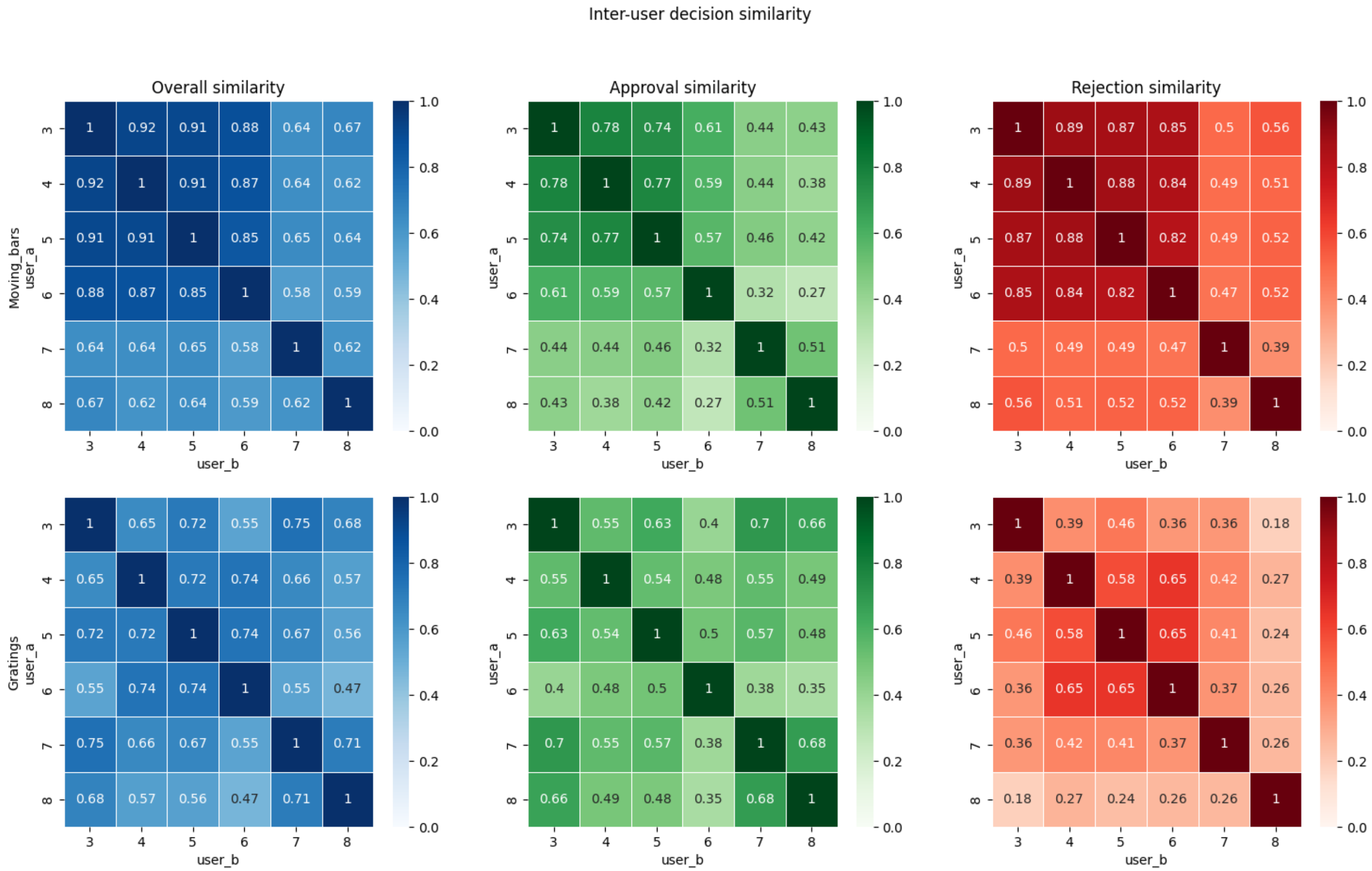
$$\text{IUDS}_A[i, j] = \frac{|(A_i \cap A_j)|}{|(A_i \cup A_j) \cap D_i \cap D_j|}$$

and their rejection IUDS is

$$\text{IUDS}_R[i, j] = \frac{|(R_i \cap R_j)|}{|(R_i \cup R_j) \cap D_i \cap D_j|}$$

Therefore, $\text{IUDS}_A[i, j]$ and $\text{IUDS}_R[i, j]$ represent the probability of users i, j respectively agreeing on an approval or rejection decision, both conditioned on the fact that one of them already made that decision. As a consequence of their definitions, these indices are not complementary $w.r.t$ to 1 or to the overall IUDS.

Figure 4.10: Decision similarity per decision and stimulus



We can see that moving bars decisions are overall more similar between different users than gratings decisions, with all user pairs agreeing on more than half of their decisions and users 3-5 forming a highly coherent block where pairs agree on at least 85% of their decisions. But when we consider the approval and rejection similarities separately, we obtain smaller indices with some user pairs agreeing upon as little as 27% of their approvals but agreeing upon at least half of their rejections (with the exception of users 7 and 8).

Gratings decision indices are more nuanced and overall less consistent than moving bars indices. Gratings approvals are frequently more similar than rejections, with more values above 0.5. Even if we disregard user 8 — a notable case, since they performed a partial task regarding gratings observations — it is still more common (7 out of 10 cases) for user pairs to agree with each on other on less than 50% of rejections. Interestingly, users 4-6 still represent a coherent group regarding rejections, compared to the moving bars case.

4.3.3 Population vs. decision consensus

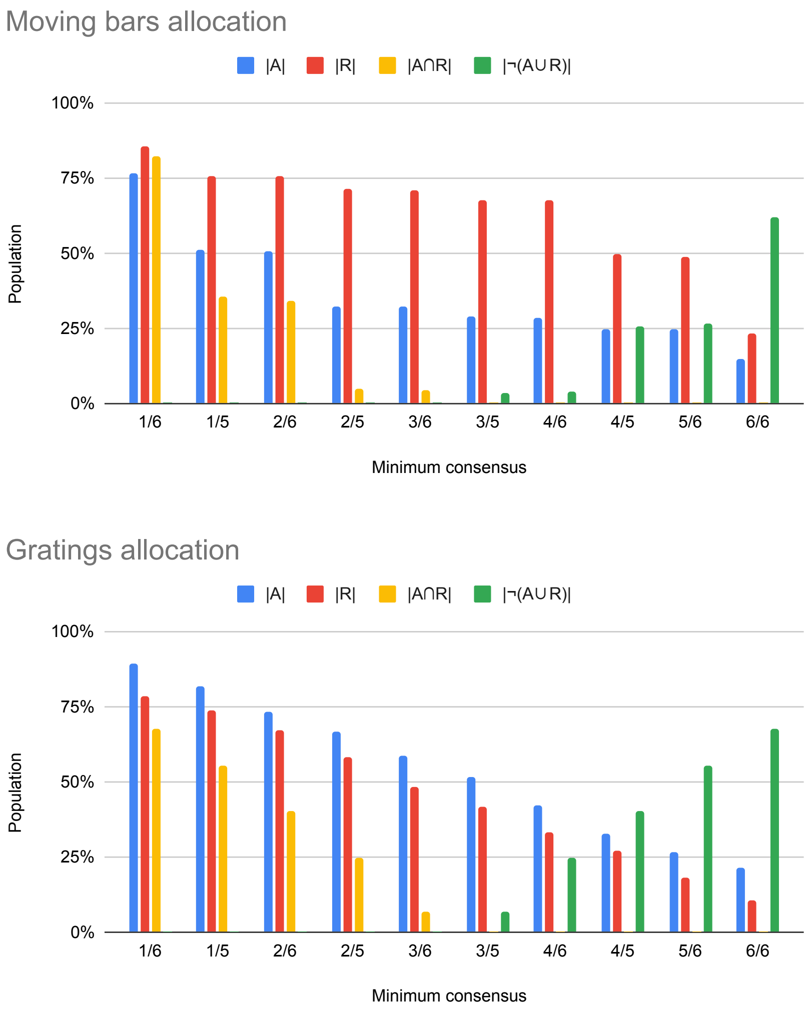
We define *consensus* to be the rate of agreement between all participants regarding a decision about an observation h. More specifically, the *approval consensus*, $c_A(h)$, is the fraction of users that approved h and *rejection consensus*, $c_R(h) = 1 - c_A(h)$, is the fraction of users that rejected h — both relative to the number of users that judged that observation. In a scenario of collective validation of observations, we could choose minimum consensus thresholds as a way to categorize observations as *finally approved* or *finally rejected*, indicated by the A and R sets, respectively. Formally,

$$A(c) = \{h \in H \mid c_A(h) \geq c\}\, R(c) = \{h \in H \mid c_R(h) \geq c\}$$

where H is the set of observations. Those definitions are convenient for treating the absence of a few decisions by user 8. However, they impose an ambiguity at exact values of c, since the number of participants is relatively small.

The plots in Figure 4.11 show how portions of the population would be allocated into the categories defined above as we varied c. They also include two additional categories: that of *ambiguous observations*, or $A \cap R$, and that of *undecided observations*, or $\overline{A \cup R}$. Note that these last two categories are mirrored about the 3/6 consensus, which is just a consequence of their definition: if we admit that a unit is approved or rejected with half or less than half of the votes, then we will have ambiguity in the classification, however if more than half of the votes are required for placing an observation in either of A or R sets, there will be no ambiguity but some observations will not match the criteria for entering any of the sets.

Figure 4.11: Allocation of summative population by minimum consensus and stimulus



These plots bring up a few interesting insights regarding the stimulus types. For gratings observations, the sizes of the $A(c)$ and $R(c)$ sets drop gradually as we increase c, which indicates there are not many observations that cause a divergence between a few specialists and the majority but rather that most observations have a uniform chance of gathering approvals, resulting in a nearly constant negative slope. Could this mean that gratings observations were overall harder to judge thus leading to looser decision criteria?

In contrast, moving bars observations show a completely different tendency, with sharp drops followed by plateaus in the size of the A set as we move from c =1/6 to $c = 1/5$ and from $c = 2/6$ to $c = 2/5$ and in the size of the R set as we move from c =4/6 to $c = 4/5$ and from $c = 5/6$ to $c = 6/5$. That suggests we could use alternative c thresholds for defining the A and R sets. For example, if we choose the limits of the sharpest drop prior unanimity, that is $A(2/5)$ and $R(4/5)$, then 32% of observations would be approved, 50% would be rejected, and 18% would be undecided but there would be no ambiguous decisions, as these sets are disjoint. In lay terms, that corresponds to being more optimistic towards approval (2 out of 5 votes would suffice for approval) and more cautious towards rejection (at least 4 out of 5 votes would be necessary for rejection).

One might ask what is the advantage of choosing different thresholds for each set. We ponder that choosing thresholds near drops in $|A(c)|$ and $|R(c)|$ could lead to a better appreciation of controversial decisions. In other words, by not merely relying on a blind majority rule, we could force more disagreeable observations into review — possibly by a different subset of participants — or requiring participants to provide justifications for their most controversial choices. That would be specially useful in a collaborative setting, as it would allow partitioning the observation sets and distributing them to different subsets of participants, reducing the total amount of work whilst requiring peer review in hard cases.

Finally, we considered whether a users' consensus adherence, *i.e.,* their decision similarity to all other users, could drift over time, as a possible indication of fatigue. We computed U_i's consensus adherence as the fraction of users that agree with U_i pertaining each observation judged, ordered these points by time, then smoothed the resulting series using a sliding Gaussian window with an aperture of 10 decisions. The fact that some users logged in multiple times was ignored when performing this convolution. We obtained the following Pearson correlation coefficients between decision rank and consensus adherence (rounded off to the third decimal place): -0.079 for User 3; 0.002 for User 4; -0.016 for User 5; -0.080 for User 6; -0.002 for User 7; and 0.037 for User 8. With these low magnitude correlations, and considering that anatomical positions were randomly ordered, we may conclude that activity time was not a determinant factor for users deviating from majority consensus.

4.4 Decision predictability

Section 4.1 showed how a significant portion of the observations in the dataset were rejected, with many anatomical positions showing very little or even no acceptance at all, and the previous section discussed how considerably dissimilar were the decisions made by different users. In this section, we study whether the functional attributes may be used to predict the mean approval of summative and individual observations. We aim to answer the question of whether a semi-automated winnowing system may use a collection of user decisions as a training set for regressing future approval rates.

For that purpose, we tested a few well-known ML models for predicting approval rates, besides using an AutoML software library that automatically tests and combines different model types. We employed dedicated models for each stimulus type, taking only functional attributes as input and completely ignoring anatomical, physiological, or parametrical attributes because:

- Doing otherwise would make for an unfair comparison with users' decisions, since that information was purposefully omitted from them, and
- These attributes are incidental to the dataset collection and construction, thus may reflect tendencies in the data that are not generalizable (*e.g.,* an electrode at a specific position provided lower quality recordings because it hit a vein, causing inflammation), therefore, even if they could improve prediction performance, that improvement would not be generalizable, and therefore not meaningful for answering the proposed question.

For both stimulus types, the summative and individual annotated datasets were randomly split into training and testing portions following a 3:1 ratio. Doing so is paramount when training any ML model, since its average performance on the training dataset is most probably better than in a batch of unseen samples [112]. Besides evaluating the regressors using the mean squared error (MSE) metric (the standard in regression problems), we also consider the performance of the predicted approval rate for classifying observations as *majority-approved* or *majority-rejected.* In other words, we also converted the trained regressors into classifiers.

4.4.1 Dataset preparation and model validation

The aim of this part of the study was to stablish soft bounds on the feasibility of predicting neuronal observation approval by a group of study participants. With that purpose we trained and evaluated regression models of the following classes:

- RLR, corresponding to the $r(\mathbf{x}) = \theta^\intercal \mathbf{x}$ predictor with a regularization term λ added to the cost function to avoid overfitting and penalize redundant features, that is $L(r; D) = \frac{1}{m} \sum \|y - r(\mathbf{x})\|^2 + \lambda \theta^\intercal \theta$
- GBR [113], which fits an ensemble of shallow regression trees (each considered a weak predictor) with the aim of generalizing better
- SVR [114], which augments the input features $\mathbf{x}$ by replacing them with functions, or *kernels* $(\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots)^\intercal$, that depend only on the Euclidean distance (radial basis function (RBF) kernel) or inner product (polynomial kernel) between $\mathbf{x}$ and respective support vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\} \subset \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, and
- AutoML ensembles comprised of several models classes[105].

The number of data points was not staggering in any of the specific datasets, so we did not concern with precise performance estimates but rather with assessing the very possibility of assisting user decisions using ML in an impactful manner. In that sense, we employed AutoML — which uses sophisticated hyperparameter search strategies to ensemble many models together whilst avoiding overfitting — to obtain a ceiling score and, following Occam's Razor principle [112], chose a few simpler model classes based on their characteristics:

- Linear regression for being the simplest regression model in conceptual terms,
- Gradient boosting for its ability to avoid overfitting and yet provide interpretable models,
- Support vector machines because their semantics are adherent to the problem — in a lower-dimensional analogy, consider that picking a few high-approval landmarks should be enough to trace a boundary around a core interval of functional attributes that favor approval, with observations becoming less favored as they move afar from the core.

Table 4.8 lists the size of D_{train}, D_{test} sizes in each of the studied datasets.

Observations	Stimulus	Train size	Test size	Features
Summative	Moving bars	173	58	23
Summative	Gratings	278	93	15
Individual	Moving bars	579	193	22
Individual	Gratings	1090	364	15

Table 4.8: Train-test split sizes for regression models

Furthermore, we evaluated hyperparameters using k-fold cross-validation (CV), with $k = 5$, over D_{train} in all cases. Fitting hyperparameters this way is necessary to avoid overly optimistic scores due to biasing them to D_{train} [115] and k-folds were the best choice given $\|D\|$ is not very large to consider a three-way partition into train-validation-test sets. GBR, RLR, and SVR hyperparameters are detailed in Table 4.9. For each model, the search space was explored exhaustively, by computing CV scores over the Cartesian product of all relevant hyperparameter options.

Input features were normalized to either lie in the $[0, 1]$ range or to have zero mean and unit-variance (procedures described in Appendix B) for the RLR and SVR classes but not for GBR, which does not require normalized features, nor AutoML, since the software package performs feature transformations itself. Except for the AutoML model, handled by its software package, all models were trained and evaluated using the Scikit-learn Python library [116].

Dimensionality reduction

In high-dimensional datasets, there are often correlated dimensions, which hinders the performance of ML algorithms since correlated features tend to overshadow the contributions of uncorrelated ones, specially in the presence of regularization. Therefore, it is customary to reduce input dimensionality, most commonly through PCA, which maps n input features into $d \leq n$ uncorrelated features with unit-variance. That allows us to save computational power, and reduce model complexity at the expense of model interpretability, since the input features are now linear combinations of the (already scaled) original dimensions.

The AutoML package automatically preprocesses data, which already includes scaling features, as we previously mentioned, and reducing their dimensionality. Furthermore, for GBR classifiers we may include a feature subsampling hyperparameter which leads to randomly choosing a subset of features when training base regression trees, therefore it is not necessary to scale or dimensionality-reduce input features. Consequently, a PCA transformation of scaled functional attributes was only applied to obtain input features for RLR and SVR.

Figures 4.12-4.13 show detailed correlation matrix plots of the target variable and the input functional attributes, providing visual clues of the relative amount of redundancy in the datasets and hinting at the predictive power of the available dimensions. Figures 4.14-4.15, also in the appendix, contain scatter plots of the six first principal component dimensions of each dataset. Tables 4.10-4.11 below show the cumulative amount of variance that is accounted for by principal components of the normalized feature sets after applying a PCA transformation. Only two in every component is shown, for brevity. Same-stimulus summative and individual datasets have similar characteristics, with cumulative variance ratios that differ by at most 10^{-3} past the 9th component, suggesting that a substantial number of dimensions might be discarded. Nonetheless, rather than choosing the number of dimensions kept off heuristics based on the cumulative amount of variance, we included the number of kept components as a hyperparameter of the RLR and SVR regressors.

Figure 4.12: Correlation matrix of moving bars functional attributes

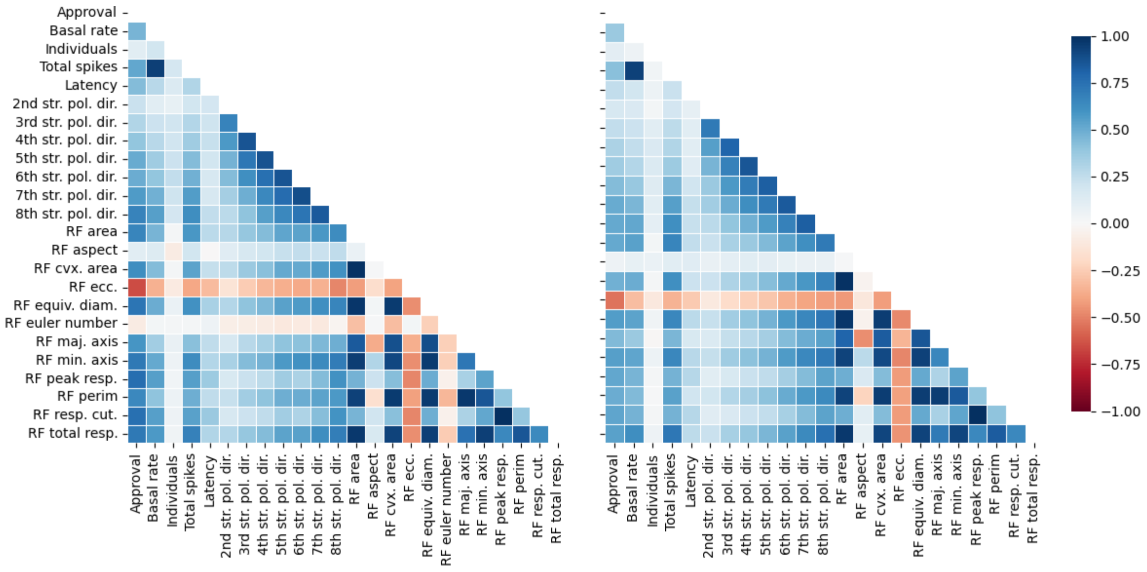
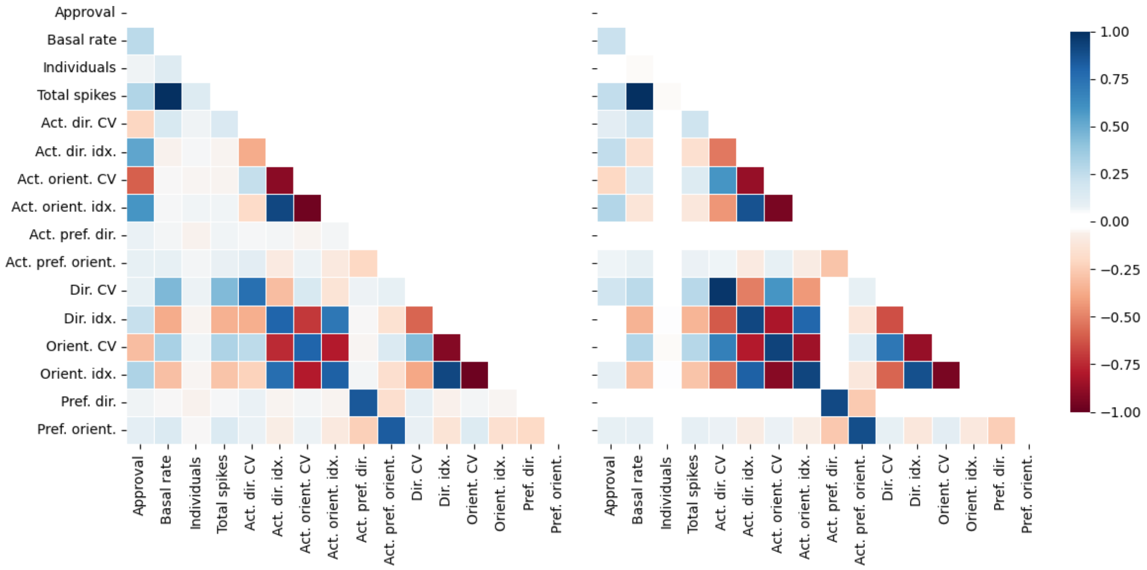


Figure 4.13: Correlation matrix of gratings functional attributes



Models	(Symbol) description	Domain	Observation
RLR,SVR	(d) Retained PCA dimensions	$2 \leq d \leq n$	n depends on stimulus type
RLR	(λ) Regularization constant	$\left\{10^k : -3 \leq k \leq 3\right\}$	Logarithmically-spaced values
SVR	(C) Regularization constant	$\left\{10^k : -3 \leq k \leq 3\right\}$	Logarithmically-spaced values, roughly equivalent to $1/\lambda$
SVR	(ϵ) Tol. margin from pred.	$\left\{10^k : -3 \leq k \leq -1\right\}$	Logarithmically-spaced values
SVR	Kernel type	Linear, polynomial, RBF, or sigmoid	
SVR	Polynomial kernel degree	$\{2, 3, 4, 5, 6, 7\}$	
SVR	Shrinking	$\{0, 1\}$	
SVR	(γ) Kernel coefficient	$\{1/(n\,\mathbb{V}[X]), 1/n\}$	
SVR	Max solver iterations	200 thousand	
GBR	(α) Learning rate	$\left\{10^k : -3 \leq k \leq 3\right\}$	Logarithmically-spaced values
GBR	(k) Max. num. of estimators	$\{50, 100, \ldots, 500\}$	
GBR	(q) Subsampling factor	$(0, 1]$	
GBR	Split quality measure	Friedman MSE, MSE	
GBR	Min. samples per split	$\{2.5\%, 5\%, 10\%, \text{unlimited}\}$	
GBR	Min. samples per leaf	$\{1.25\%, 2.5\%, 5\%, \text{unlimited}\}$	
GBR	Max. depth	$\{1, ..., 9\}$	

Table 4.9: Model hyperparameters and their ranges

Figure 4.14: PCA projections of first six components of moving bars functional features

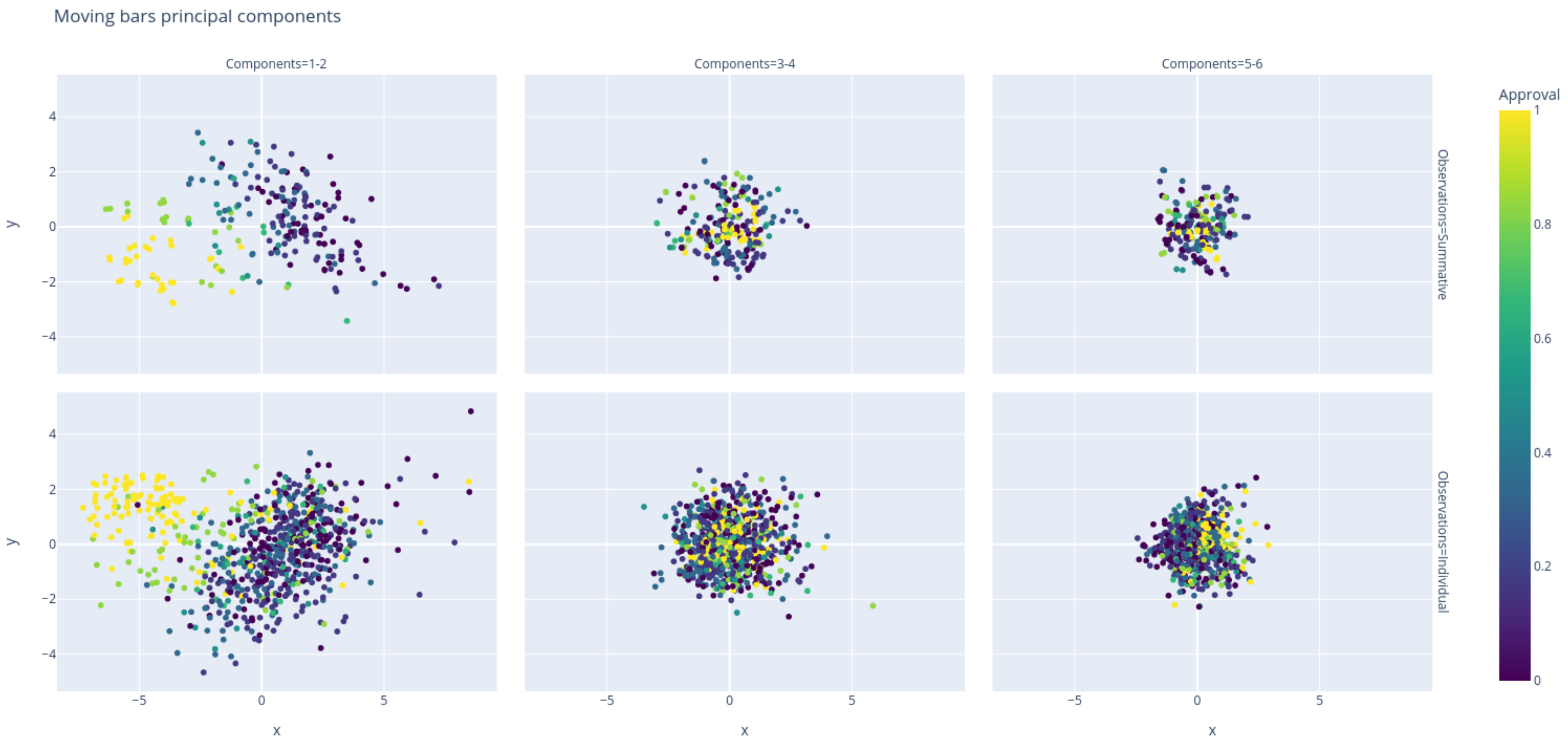


Figure 4.15: PCA projections of first six components of gratings functional features

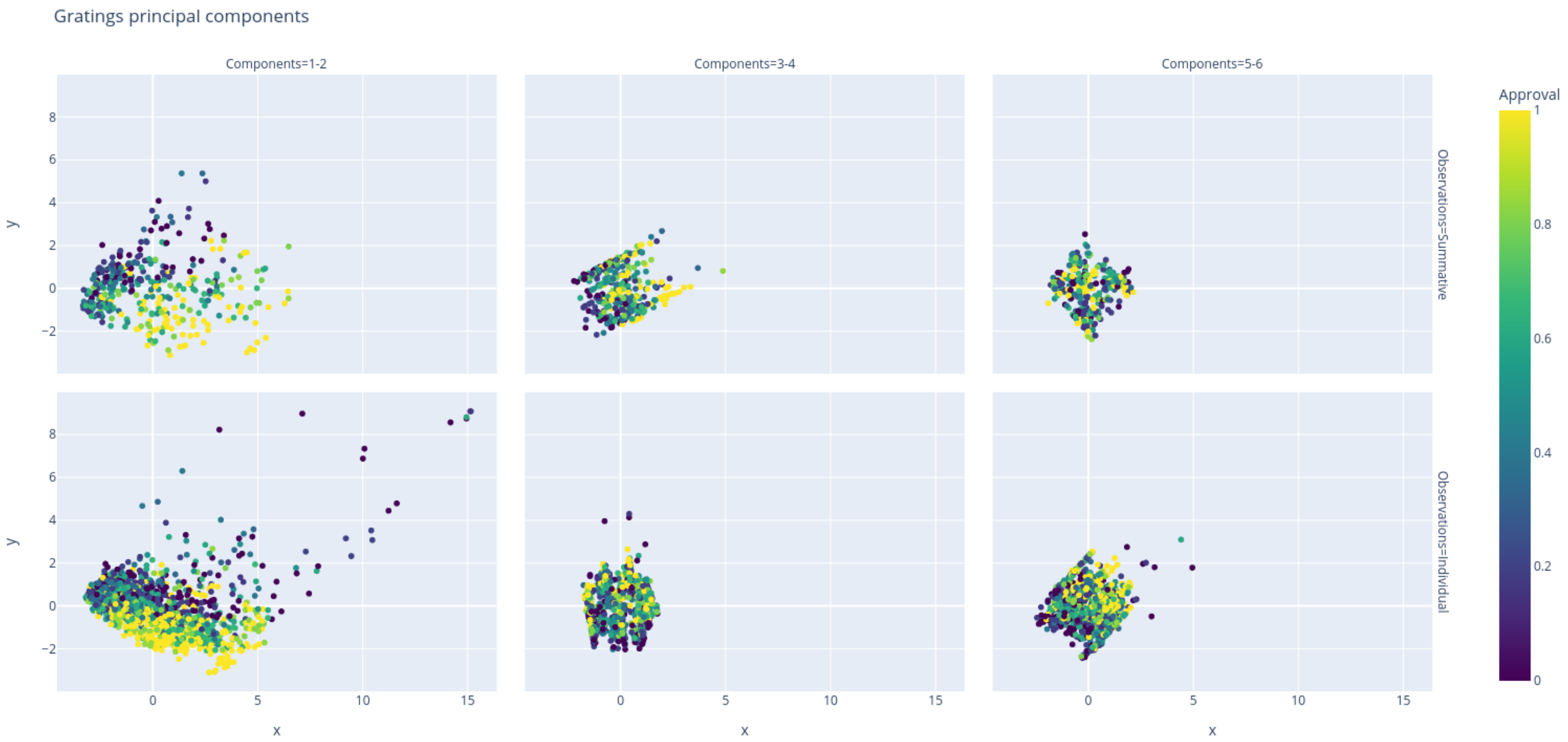


Table 4.10: Ratio of cumulative variance retained by principal components in moving bars dataset

Obs.	PC1	PC3	PC5	PC7	PC9	PC11	PC13	PC15	PC17	PC19
Sum.	0.597	0.834	0.937	0.984	0.992	0.995	0.997	0.998	0.999	1.0
Ind.	0.555	0.816	0.93	0.981	0.988	0.994	0.997	0.998	0.999	1.0

Table 4.11: Ratio of cumulative variance retained by principal components in gratings dataset

Obs.	PC1	PC3	PC5	PC7	PC9	PC11	PC13	PC15	PC17	PC19
Sum.	0.447	0.692	0.838	0.927	0.953	0.97	0.984	0.993	0.998	1.0
Ind.	0.489	0.704	0.853	0.946	0.971	0.982	0.991	0.996	0.999	1.

4.4.2 Regression models compared

Table 4.12 summarizes the performance of the best model of each family, in terms of training/test scores (approval rate MSE), and majority approval precision, recall, and F1 scores. For the majority approval classifiers, given that $r(\mathbf{x}) \approx y$ is the regressed approval probability, we chose approval decision thresholds $r(\mathbf{x}) \geq y_0$ to maximize the F1 score, not necessarily $y_0 = 1/2$. For details about the CV-selected hyperparameters. of the optimal models, see Table 4.13, which follows the same notation as Table 4.9.

Train and test MSE metrics are also plotted as per-model overlaid bars in Figure 4.16 Except for the gratings RLR models, train score is always better than test score. In terms of approval rate prediction, moving bars models are more precise than gratings models and models for individual observations are nearly as performant as models for summative observations, with the exception of gratings RLR. The worst performance, in general, corresponds to SVR models, and ensembles produced by AutoML and GBR are nearly equivalent.

4.4.3 Ternary classification and ranking

Let us consider two additional applications of regressed approval rates:

- Observation ranking — that is, given two competing observations, possibly with identical anatomical attributes but with slightly different parametrical attributes, which observation has a higher approval probability?

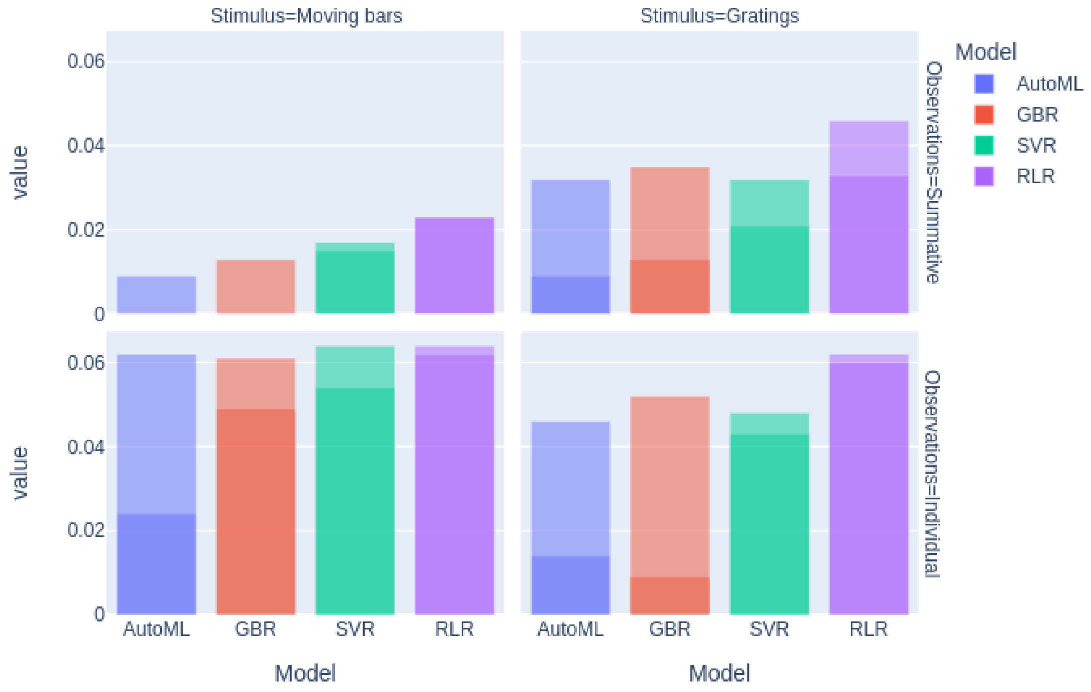
Table 4.12: Approval regression scores for all models. $L(D_{\text{train}})$ and $L(D_{\text{test}})$ represent MSE on the train and test datasets, respectively. P and R represent precision and recall scores, and p indicates the optimal decision threshold for maximizing the F1 score, also shown. For each sub-problem, bold numbers highlight the best metric achievable, considering differences only in the third decimal place to be ties.

		Model	$L(D_{\text{train}})$	$L(D_{\text{test}})$	$p \geq$	F1	P	R
Summative	MB	AutoML	0.0	**0.009**	0.4	**1.0**	**1.0**	**1.0**
		GBR	0.0	0.013	0.425	0.97	0.941	**1.0**
		SVR	0.015	0.017	0.325	0.914	0.842	**1.0**
		RLR	0.023	0.023	0.375	0.914	0.842	**1.0**
	Gratings	AutoML	0.009	**0.032**	0.3	0.86	0.754	**1.0**
		SVR	0.021	**0.032**	0.45	0.857	0.8	0.923
		RLR	0.046	0.033	0.475	**0.889**	**0.857**	0.923
		GBR	0.013	0.035	0.4	0.87	0.839	0.904
Individual	MB	GBR	0.049	**0.061**	0.4	0.683	0.672	0.694
		AutoML	0.024	0.062	0.35	0.712	0.671	**0.758**
		RLR	0.064	0.062	0.45	**0.738**	0.75	0.726
		SVR	0.054	0.064	0.45	0.704	**0.826**	0.613
	Gratings	AutoML	0.014	**0.046**	0.375	0.828	0.735	0.948
		SVR	0.043	0.048	0.375	**0.841**	0.742	**0.972**
		GBR	0.009	0.052	0.4	0.817	0.747	0.901
		RLR	0.062	0.06	0.475	0.795	**0.765**	0.826

Table 4.13: Approval regression scores for all models. $L(D_{\text{train}})$ and $L(D_{\text{test}})$ represent MSE on the train and test datasets, respectively. P and R represent precision and recall scores, and p indicates the optimal decision threshold for maximizing the F1 score, also shown. For each sub-problem, bold numbers highlight the best metric achievable, considering differences only in the third decimal place to be ties.

		Model	Details
Summative	MB	GBR	450 estimators; 23 features; 3 nodes deep
		SVR	RBF kernel; $d = 18$; $C = 0.316$; $\epsilon = 0.100$; $k = 95$
		RLR	$d = 20; \lambda = 3.162$
	Gratings	SVR	RBF kernel; $d = 14$ $C = 0.316$; $\epsilon = 0.010$; $k = 259$
		RLR	$d = 10$; $\lambda = 10.0$
		GBR	50 estimators; 15 features; 3 nodes deep
Individual	MB	GBR	100 estimators; 22 features; 7 nodes deep
		RLR	$d = 20$ $\lambda = 1.0$
		SVR	RBF kernel; $d =$
	Gratings	SVR	RBF kernel; $d = 14$ $C = 0.316$; $\epsilon = 0.100$; $k = 715$
		GBR	200 estimators; 15 features; 7 nodes deep
		RLR	$d = 14$; $\lambda = 31.623$

Figure 4.16: Train and test loss in all models per observation and stimulus type. Train and test metrics are represented by overlaid semi-transparent bars along the x-axis, with the smaller metrics always corresponding to train performance. Rows contain plots for summative (top) and individual (bottom) observations and columns contain plots for the moving bars and gratings stimulus types.



- Detecting observations that require further inspection — how precisely could we detect observations that have intermediate approval rates, such that they are not evidently

The first application simply amounts to comparing the regressed approval rates of both observations. A reasonable scoring metric for a model's ability to rank observations is, given observations pairs

$$\{[(\mathbf{x}_{i1}, y_{i1}), (\mathbf{x}_{i2}, y_{i2})]\}_{i>0}$$

to count how many times the predicate

$$[y_{i1} < y_{i2} \implies r(\mathbf{x}_{i1}) < r(\mathbf{x}_{i2})] \vee [y_{i1} > y_{i2} \implies r(\mathbf{x}_{i1}) > r(\mathbf{x}_{i2})]$$

is true and divide that by the number of such pairs.

The second application requires precisely defining the three target classes, labeled as $\mathfrak{A}$ for *certainly approved*, $\mathfrak{R}$ for *certainly rejected*, and $\mathfrak{M}$ for *must review*. However, as will be reported in Section 4.5, some study participants declared changing their minds about certain observations during the evaluative interviews, which just confirms a basic intuition that user decisions are subject to noise. Furthermore, some observations were judged by only 5 of 6 total participants. Therefore, let us assume a conservative stance towards majority decisions, with minimal tolerance for disagreement, and postulate that

$$c(y) = \begin{cases} \mathfrak{R}, \text{ if } y \leq 1/5 \\ \mathfrak{M}, \text{ if } 1/5 < y < 4/5 \\ \mathfrak{A}, \text{ if } y \geq 4/5 \end{cases} \tag{4.1}$$

where $c(y)$ is a generic classification model based on the approval rate y. This way, certain approval/rejection tolerate at most one disagreeing vote in every five (consequently, also one in every six) and everything in between is considered to require further inspection. The critical classification metric for $\mathfrak{A}, \mathfrak{R}$ is precision, since false positives lead to saving low quality observations for further analysis or discarding relatively good ones while false negatives lead to an increase of expert time. Conversely, the critical metric for $\mathfrak{M}$ is recall.

Table 4.14 shows observation ranking and classification metrics for the same regression models that we discussed earlier, in this section. Ranking pairs were obtained from the respective test datasets by gathering all pairs of observations with identical anatomical attributes (in essence, observations from the same electrode). Classification metrics were computed by assuming $c(y)$ as the true label and $c(r(\mathbf{x}))$ as the predicted one. The thresholds for inclusion in the extreme classes were tweaked in order to maximize their precision, so they may differ from Equation 4.1.

4.5 Evaluation interviews

Five of the six study participants were interviewed after completing the proposed activity, in sessions that lasted between 45 and 90 minutes. Although the sessions did not follow a strict question-answer format, the following themes were addressed in all interviews:

- Usability issues (*e.g.,* page loading time, responsiveness, layout)
- Encoding issues (*e.g.,* confusing plots, missing attributes, improper scales)
- Goal/task issues (*e.g.,* clarity of goal, decision difficulty, exhaustion)
- Data idiosyncrasies (*e.g.,* quality, coherence)

Table 4.14: Ternary classification metrics for all models. $P_{\mathfrak{A}}$ and $P_{\mathfrak{R}}$ represent precision for the $\mathfrak{A}$ and $\mathfrak{R}$, respectively, and $R_{\mathfrak{M}}$ represents recall for the $\mathfrak{M}$ class. The p columns contain the optimal probability thresholds for inclusion in the $\mathfrak{A}$ and $\mathfrak{R}$, an may vary from Equation 4.1. For each sub-problem, bold numbers highlight the best metric achievable, considering differences only in the third decimal place to be ties.

						Ranking			
		Model	$P_{\mathfrak{R}}$	$P_{\mathfrak{A}}$	$R_{\mathfrak{M}}$	Pairs	Score	$p_{\mathfrak{R}} \leq$	$p_{\mathfrak{A}} \geq$
Summative	MB	SVR	0.923	**1.0**	0.667	3	0.667	0.35	0.6
		RLR	0.921	**1.0**	0.667	3	0.333	0.35	0.625
		GBR	0.943	**1.0**	0.778	3	**1.0**	0.3	0.675
		AutoML	**0.971**	**1.0**	**0.889**	3	**1.0**	0.3	0.675
	Gratings	GBR	0.731	0.867	0.686	21	0.571	0.3	0.7
		AutoML	0.75	**0.871**	0.714	21	0.571	0.3	0.7
		SVR	0.762	**0.871**	0.743	21	**0.667**	0.3	0.7
		RLR	**0.9**	0.867	**0.857**	21	0.619	0.325	0.7
Individual	MB	SVR	0.655	0.903	0.382	59	0.424	0.3	0.6
		GBR	0.654	**0.963**	0.418	59	**0.475**	0.325	0.7
		AutoML	**0.701**	0.933	0.473	59	0.458	0.3	0.675
		RLR	**0.706**	**0.962**	**0.582**	59	**0.475**	0.325	0.7
	Gratings	GBR	0.672	0.731	0.726	346	0.723	0.3	0.7
		RLR	0.643	0.736	0.732	346	**0.731**	0.325	0.7
		SVR	0.723	0.765	0.771	346	**0.734**	0.325	0.7
		AutoML	**0.739**	**0.819**	**0.803**	346	0.723	0.325	0.7

- Decision factors and strategies (*e.g.,* what attributes and features were fundamental for decision-making, or what sequence of cognitive steps lead to making decisions)

- Reasons for disagreeing with other participants in a few illustrated cases.

Those themes were pursuit in a non-structured manner, so that affirmations outside of them could be advanced by the interviewees. Each statement (*i.e.,* affirmations, questions, suggestions) made by them was written down by the interviewer (the author of this thesis) on time and we later open-coded these statements. After three coding rounds, we came up with 24 distinct statements grouped into 7 categories. The categories, each identified by a single letter, are listed below:

Data idiosyncrasies (D) Inconsistencies or quality issues in the data that hindered the participant's trust in the tool or the data itself.

Encoding issues (E) Elements that were incorrectly assumed to encode an attribute or, conversely, encodings that were not perceived.

Decision factors (F) Specific structures in the data, or queries that were important for decision-making.

Perceptions (P) General perceptions about decision-making and goal/task performance that were commented but not elaborated as clearly as *Factors*.

Suggestions (S) Changes in encoding or attribute selection suggested for making the panels more informative.

Usability issues (U) Problems in using the interface effectively.

Whishes (W) Degrees of freedom and interaction processes the participants expected to find in the tool in order to explore the dataset freely and/or to perform alternative goals.

In some cases, different participants made contradictory statements, so we encoded their position regarding each statement using the numbers -1 (disagreed), 0 (did not mention it), and 1 (agreed). Whenever a participant made a statement (either in its affirmative or negative form), we say the participant *voted* for it. Table 4.15 briefly describes the 24 statements, each identified by a letter that indicates its category and a subscript that abbreviates its description, and followed by a short description of the statement and the votes cast by each participant. The table is ordered by decreasing order of votes, then consensus, such that the last row contains the most debatable statement.

Code	Title	U3	U4	U6	U7	U8
$\mathrm{W}_{\mathrm{PopCtx}}$	Missed contextual information and navigational features for the appreciation of hypotheses and their relationships	1	1	1	1	1
$\mathrm{P}_{\mathrm{Sys}}$	Perceived system as useful	1	1	1	1	1
F_{FR}	Reported firing rates in gratings diagrams ground decisions and break ties	1	-1	1	-1	1
$\mathrm{E}_{\mathrm{RowAttr}}$	Assumed vertical stacking encoded domain attributes	1	1	1	1	0
$\mathrm{F}_{\mathrm{RFPeak}}$	Reported moving bars decision strongly influenced by receptive field's peak-to-local-maxima disparity	1	1	1	1	0
$\mathrm{F}_{\mathrm{RFZ}}$	Reported moving bars decision strongly influenced by receptive field's Z-response cuttoff	1	1	1	1	0
$\mathrm{S}_{\mathrm{FRZ}}$	Suggested firing rates in gratings diagrams should be Z-normalized	0	1	-1	1	1
$\mathrm{F}_{\mathrm{RFPol}}$	Reported moving bars decision strongly influenced by polargram	1	0	-1	1	1
$\mathrm{F}_{\mathrm{RMDec}}$	Perceived response maps enabled fast and reliable decisions about moving bars hypotheses	1	0	1	1	0
$\mathrm{W}_{\mathrm{HypSel}}$	Missed the ability to update panels for single hypothesis selection	0	1	1	1	0
$\mathrm{F}_{\mathrm{WFNoise}}$	Reported waveform diagrams enable spotting noisy data and confirm rejections	1	0	1	1	0
$\mathrm{P}_{\mathrm{GrHard}}$	Complained gratings selectivity was harder to interpret	0	1	-1	1	0
$\mathrm{F}_{\mathrm{WFDec}}$	Reported waveform diagrams indicate individual unfeasibility	1	-1	1	0	0
$\mathrm{E}_{\mathrm{HypId}}$	Assumed hypothesis labels encoded domain attributes	1	0	1	0	0
$\mathrm{P}_{\mathrm{Tired}}$	Perceived task repetition as tiresome and poised to cause decision criteria drift over time	1	1	0	0	0
$\mathrm{F}_{\mathrm{RFSz}}$	Reported moving bars decision strongly influenced by receptive field's size	1	0	0	1	0
$\mathrm{D}_{\mathrm{PolRM}}$	Complained polargrams do not match response maps and should not contain so many null vertices	0	-1	0	1	0
$\mathrm{P}_{\mathrm{ActConf}}$	Perceived the activity as unclear	0	0	0	0	1
$\mathrm{E}_{\mathrm{RMDisp}}$	Complained the disparity in moving bars hypotheses was caused by processing errors or bad data selection	0	0	0	1	0
$\mathrm{U}_{\mathrm{PolMis}}$	Complained polargrams did not load properly for some screens	0	0	0	0	1
$\mathrm{F}_{\mathrm{RFCent}}$	Reported moving bars decision strongly influenced by receptive field's centrality	0	0	0	1	0
$\mathrm{W}_{\mathrm{UpDat}}$	Missed the ability to upload their own files into the system	0	0	0	1	0
$\mathrm{U}_{\mathrm{PltSz}}$	Complained fixed plot sizes caused interface overlays in smaller screens	1	0	0	0	0
$\mathrm{E}_{\mathrm{RowHyp}}$	Assumed relationship between same-row hypotheses from different stimuli	1	0	0	0	0

Table 4.15: User statements gathered during evaluative interviews

Below, we provide longer explanations about each statement. In the cases where additional insight was provided by certain users (regarding why they agree or disagree with the statement), it is listed below the description.

D_{PolRM} – Complained polargrams do not match response maps and should not contain so many null vertices. Users claimed that polargrams seemed incorrectly calculated due to the sheer amount of null vertices they contained.

> **User 4:** User justified the observed patterns on the basis of low quality data and automatic estimation without filtering, that is: few spikes → low firing rates → noisy response maps → ill-defined receptive fields → lack of activity on multiple directions.

E_{RMDisp} – Complained the disparity in moving bars hypotheses was caused by processing errors or bad data selection. Since there were two recordings of the moving bars stimulus, the evidence could be strikingly different. They were confused by this disparity and assumed it could be either: (a) bad data processing/selection; (b) different recording sites mixed together.

F_{RFSz} – Reported moving bars decision strongly influenced by receptive field's size.

F_{RMDec} – Perceived response maps enabled fast and reliable decisions about moving bars hypotheses. Z-normalized responses, receptive field contour, size, shape, and Z threshold are key properties of hypothesis and their encoding in response maps allows for quick and realiable decisions.

F_{WFDec} – Reported waveform diagrams indicate individual unfeasibility. The summative activity represented by other diagrams may seem feasible, but the waveforms diagram shows that individual hypotheses are not.

> **User 4:** Spike detection is not perfect, therefore when other diagrams indicate functional selectivity, waveforms should not lead to discarding a summative hypothesis.
>
> **User 8:** The semantics of the waveform diagram was not clear.

F_{WFNoise} – Reported waveform diagrams enable spotting noisy data and confirm rejections. Noisy spike patterns (due to low detection thresholds or AC/DC interference, for example) are easily spottable in waveform diagrams, which grounds rejection decisions further.

F_{RFZ} – Reported moving bars decision strongly influenced by receptive field's Z-response cuttoff.

$\mathrm{F_{FR}}$ – Reported firing rates in gratings diagrams ground decisions and break ties. Firing rates represented in gratings diagrams are important to confirm feasibility of decisions or to break ties.

$\mathrm{F_{RFPol}}$ – Reported moving bars decision strongly influenced by polargram.

> **User 6:** The user reported unfamiliarity with these diagrams and claimed they did not influence their decision.

$\mathrm{F_{RFCent}}$ – Reported moving bars decision strongly influenced by receptive field's centrality.

$\mathrm{F_{RFPeak}}$ – Reported moving bars decision strongly influenced by receptive field's peak-to-local-maxima disparity.

$\mathrm{E_{HypId}}$ – Assumed hypothesis labels encoded domain attributes. The presence of numeric database IDs in hypothesis labels confused users. They guessed the meaning of the IDs in the context of their field (like electrode insertion depths, for instance).

$\mathrm{E_{RowAttr}}$ – Assumed vertical stacking encoded domain attributes. Similarly to hypothesis IDs, users interpreted vertical stacking of summative hypotheses as an encoding of positional attributes (like electrode depth, or simply different recording sites).

$\mathrm{E_{RowHyp}}$ – Assumed relationship between same-row hypotheses from different stimuli. Despite each screen presenting varied, usually unmatched quantities of hypotheses for both stimulus types, users assumed that hypotheses in the same row were related (in the same recording site, for instance) and therefore it made sense to judge them mutually.

$\mathrm{P_{ActConf}}$ – Perceived the activity as unclear. The objective of the proposed activity was not clear.

$\mathrm{P_{GrHard}}$ – Complained gratings selectivity was harder to interpret. Participants considered gratings selectivity harder to interpret, specially due to the amount of plots.

> **User 4:** They suggested using a different encoding, similar to multi-colored polargrama, but did not elaborate on it.
>
> **User 7:** Gratings-selective populations will sometimes exibit selectivity to orthogonal direction as a way to enhance contrast. This leads to seemingly unselective profiles when visualized at the summative level. That leads to $\mathrm{W_{HypSel}}$.

P_{Tired} – Perceived task repetition as tiresome and poised to cause decision criteria drift over time. Users reported becoming tired with the study's duration and said their decision criteria and attention to details probably changed over time.

P_{Sys} – Perceived system as useful. Users reported a sense of satisfaction with the system, claiming that it reunited most relevant information for the winnowing task and made the winnowing process easier, albeit repetitive.

User 6: Bad data was very easy to spot and discard immediately.

User 8: Understanding the activity's goal was not easy.

S_{FRZ} – Suggested firing rates in gratings diagrams should be Z-normalized. Since brain operates on varied but usually low firing rates, comparing low absolute values is meaningless, therefore firing rates should be Z-normalized in gratings diagrams.

User 6: Although Z normalization is useful, very low firing rates are unrealiable.

User 8: Alternating between Hertz and normalized would be useful.

U_{PolMis} – Complained polargrams did not load properly for some screens. An issue where polargrams would not load properly for some screens was reported only by this user.

User 8: It was possible to perform the activity despite the absent information but it became more difficult.

U_{PltSz} – Complained fixed plot sizes caused interface overlays in smaller screens. Since all plots had fixed sizes, the interface did not fit entirely in screens with resolution lower than 1920 x 1080 pixels, causing partial panel overlays.

W_{PopCtx} – Missed contextual information and navigational features for the appreciation of hypotheses and their relationships. Since domain experts expect data locality to reflect population consistency (for example, similar functional properties along electrode's path, or among neighboring electrodes) but this information was completely hidden, they felt lacking relevant information for judging task items in a holistic way and even interpreted some same-screen hypotheses with very different properties to be indicators of ill-defined populations. They also expressed the willingness to navigate through the dataset while performing the task. This sparks discussion about the suitability of the task abstraction: what is the properis way to design a tool for population search, top-down or bottom-up (as we did)?

User 4: Suggested using a notation for indicating some of these attributes, for example: L1200μm#2 (left-hemisphere, 1200μm deep, channel 2).

User 8: Reported difficulty in understanding navigation.

W_{HypSel} – Missed the ability to update panels for single hypothesis selection. Users missed the ability to update the diagrams by toggling contributions of individual hypotheses (for instance, in response to toggling them in the waveforms diagram), specially for the gratings hypotheses. That would have serious implications for the task abstraction itself, since allowing to screen hypotheses at either the summative or individual levels would become reasonable.

W_{UpDat} – Missed the ability to upload their own files into the system.

4.5.1 Statement implications

Let us now reflect on the statements presented. We will revisit the main conclusions in Section 5 when discussing limitations and future directions of our design study but for now we will look at them in more detail. We will discuss design implications and limitations by referring to statements that support them, using the (A/B) notation besides statement codes to indicate that a statement was made by A participants and that its negation was made by B participants.

General reception

The unanimous statement that the system is useful (P_{Sys} 5/0) suggests that correct data/task abstractions and effective (but not necessarily efficient) data/interaction encodings were chosen. Such a statement cannot be taken at face value, specially due to the close relationship between the author and the interviewees, which may bias feedbacks towards an encouraging tone [99]. In particular, it cannot be assumed to validate domain problem characterization, since only a longitudinal study could confirm or deny the users' perception of usefulness. As a matter of fact, User 8 claimed difficulty in grasping the objectives of the proposed activity, which probably stems from the fact that it was a proposed activity, rather than they using it on the wild. Nonetheless, considering that users were able to conduct meaningful work and to perceive it as so, and that their decisions were indeed fairly regular (as shown by the quantitative discussion in Section 4.3), we may interpret this as an indication of relevant data/task abstraction.

It is important to highlight that some users declared becoming exhausted and interpreted this as a cause for reduced attention over time (P_{Tired} 2/0). When we consider the presented data was but a small slice of a much larger collection, it becomes clear that semi-automated decision making should be pursuit, perhaps requiring a few initial decisions by domain experts to bootstrap a ML model like the ones described in Section 4.4.

Finally, the wishes for

- Isolating individual observations from summative panels like response maps and gratings polar sector plots in order to winnow them independently ($\mathrm{W}_{\mathrm{HypSel}}$ 3/0), and

- Uploading custom datasets ($\mathrm{W}_{\mathrm{UpDat}}$ 1/0)

constitute a possible set of mandatory features (the first one indeed a filtering task) for a system enabling users to conduct the winnowing goal. Complementary to $\mathrm{W}_{\mathrm{UpDat}}$ we also believe some form of data saving/export feature, despite not being mentioned, would be required.

Tool limitations

The tool was shown to be limited in a few ways. We purposefully shuffled and omitted anatomical/parametrical attributes to prevent confirmation and anchoring biases [117] when judging observations — that is, to prevent users from systematically favoring certain attribute values for making decisions based on pre-existing expectations about which parametrical attributes give better results (confirmation bias), or on opinions formed after appreciating just a few early or more evident samples (anchoring bias). However, it led to all users feeling that the lack of contextual information (in the form parametrical/anatomical attributes) made using the system less appreciable ($\mathrm{W}_{\mathrm{PopCtx}}$ 5/0). Users showed interest in combining the **Derive**, **Winnow**, and **Explore** goals in a single tool, whereby they could tweak parameters, obtain new populations, compare them, then take further action. This would naturally bring in tasks with searching, filtering, and comparison semantics, as outlined in Section 3.3, and that would either require these attributes to be made explicit, or to be derived into agnostic attributes (that is, obfuscated) to enable searching without incurring in (or at least diminishing the effect of) the aforementioned biases.

Attribute omission likely also led users to believe that the vertical positioning of observations ($\mathrm{E}_{\mathrm{RowAttr}}$ 4/0) and the placement of moving bars/gratings observations at the same screen height ($\mathrm{E}_{\mathrm{RowHyp}}$ 1/0) encoded some attribute or relationship, which they did not. Anatomical attributes might have been visually encoded on a side or top panel, and provided that the sequence of anatomical positions was still shuffled through screens, this might not have contributed to any biases, and yet increased the users' awareness of the context. Figure 4.17 illustrates how this could have been implemented, using a combination of styled marks on simple positional channels to encode brain hemisphere, 2D electrode position, and electrode depth.

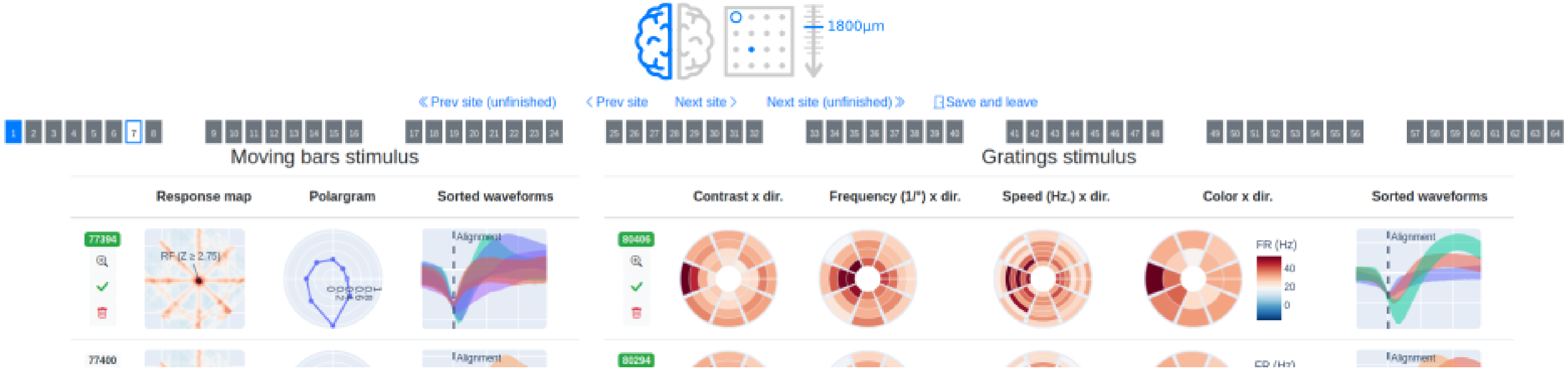


Figure 4.17: Possible encoding of anatomical attributes using positional marks and scales at the top of the screen

On the usability spectrum, two other important, already mentioned features were missed: the ability to update summative views, like response maps and gratings diagrams, to arbitrary selections of individual observations in the waveforms view ($\mathrm{W}_{\mathrm{HypSel}}$ 3/0) and the ability to upload their own files to the system ($\mathrm{W}_{\mathrm{UpDat}}$ 1/0). The second is a rather positive sign, as it demonstrates the interest necessary for a field test, and the first might be easily solved by placing the individual, togglable legend markers from the waveforms view in evidence and linking them to the update of the summative views.

All the same, their urge to winnow individual observations beyond respective summative groups shows the success of the waveforms view, as it was the sight of different profiles in it that had them try toggling the individuals and realizing the missing feature. Some toggling cases were analyzed in a Wizard of Oz fashion during the interviews in response to $\mathrm{W}_{\mathrm{HypSel}}$. We will present some of these cases again in Section 5.

Decisions about moving bars observations

Most clues used to identify RFs in moving bars observations could be quickly found in the response map plot ($\mathrm{F}_{\mathrm{RMDec}}$ 3/0), as confirmed by the following additional decision factor statements:

- Higher Z-response cutoffs indicate higher feasibility ($\mathrm{F}_{\mathrm{RFZ}}$ 4/0) — although well-defined thresholds may be hard to specify with noisy data, attributes with statistical meaning, specially within a broader context, help to ground decisions more firmly.
- RF sizes are important indicators of their quality ($\mathrm{F}_{\mathrm{RFSz}}$ 2/0)
- An RF's viability diminishes with its distance from the visual field's center ($\mathrm{F}_{\mathrm{RFCent}}$ 1/0)
- The sharper a response map's peak in comparison to other local maxima, the more likely it is to be a true RF ($\mathrm{F}_{\mathrm{RFPeak}}$ 4/0)

Polargrams were also reported to be very important for decisions ($\mathrm{F}_{\mathrm{RFPol}}$ 3/1) but were mentioned less frequently (yielding only one decision factor and one idiosyncrasy statement) and not as unanimously as response maps (since one of the participants did not use them). User 8 reported issues with them not loading properly ($\mathrm{U}_{\mathrm{PolMiss}}$ 1/0) and said that it did not prevent performing the task but made it more difficult (User 8's observations in $\mathrm{P}_{\mathrm{Sys}}$). Therefore, we need to consider that maybe only User 6 did not make use of them and that User 4 may have simply not mentioned their importance yet relied on them. Finally, User 7 raised concerns about data quality and the system's correctness after seeing a few null vertices in polargrams but User 4 commented on those simply being an artifact of low spike counts. That bears the question of wether lower-level (less aggregated and processed) functional attributes, like spike trains should be made available in the details panel.

Decisions about gratings observations

Gratings observations raise a few points, since they were reported by some but not all users as being harder to interpret than moving bars ones ($\mathrm{P}_{\mathrm{GrHard}}$ 2/1) but did not earn many statements overall. User 4 even began to articulate an alternative encoding proposal based on closed polygonal curves (similar to polargrams) and User 7 suggested that the chosen encodings may not be ideal for summative observations, since individuals with orthogonal preferences would seem like a non-selective summative observation.

The encoding chosen for the design study was certainly an improvement over the informal study version, since that one condensed far too much information about the stimulus in a single plot and, far worse, used color coding to indicate quantitative stimulus variables. However, based on the discussion above, we conclude that the encoding of gratings functional attributes should be redesigned. Furthermore, that would be a proper case for a quantitative usability study based on trial-and-error measurements for determining the best solution from a pool of proposals.

On a different matter, the choice of firing rate as the encoded response attribute also sparks discussion, since two pairs of users have contrasting views on wether it is better to visualize the very firing rates or their z-normalized counterparts (F_{FR} 2/2 and $\mathrm{S}_{\mathrm{FRZ}}$ 2/1). The most conservative reasoning is to compute both attributes and allow the user to toggle between them, or to encode both using the same color-scale, since they are proportional.

Peculiar individuals hidden in summative observations

Regarding the $\mathrm{W}_{\mathrm{HypSel}}$ (3/0) and $\mathrm{P}_{\mathrm{GrHard}}$ (2/1) statements again, we gathered a few examples of summative observations with individual observations that look particularly different. Figures 4.18-4.19 shows examples of summative moving bars details panels on top, followed by individual details panels for each of its constituent observations. The summative panel shows the presence of a direction-selective RF, as indicated by its almond-shaped polargram and mildly pronounced peak (with $Z \geq 1.58$) but the unfolding of its individuals shows somewhat different profiles.

In Figure 4.18, we can see that individuals b and d have similarly shaped waveforms but with different amplitudes that suggest they are indeed separate individuals. Their peaks are not as outstanding and incontestable as the summative's, since other local maxima have similar values in the response map, and with d's area being atypically non-convex but their polargrams maintain a consistent profile of direction selectivity. Individual c has a pronounced peak relative to the background ($Z \geq 1.98$) but it is not outstanding, since other local maxima have similar values and equally small areas.

Meanwhile, in Figure 4.19, individuals c and d show pronounced and outstanding response peaks and orientation selectivity, so they look plausible. However individual b's hypothetical RF center is an artifact that lies at the edge of the visual field and its waveforms do not have a typical shape, so it should be discarded.

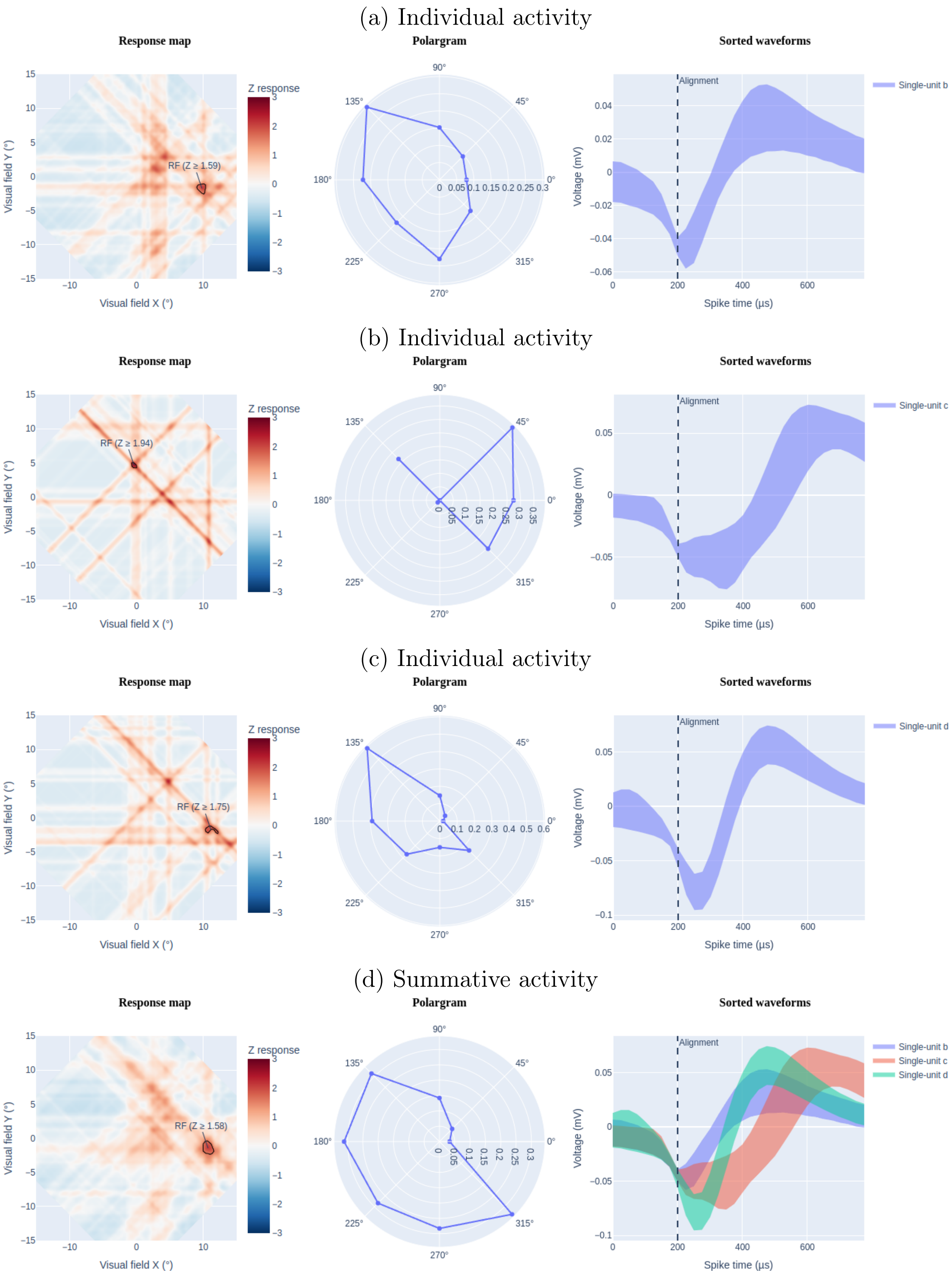
Finally, let us present some cases of gratings responses, illustrated in Figure 4.20. Summative movement in horizontal (0-180) and ascendant diagonal (45-225) directions prefers higher contrast (≥ 0.5) and lower frequencies (≤ 1 Hz) responds better to black and white than to blue and green gratings Individuals a/b/d show similar preferences Individual c shows a more confounded pattern, with less obvious direction selectivity and no clear contrast rang. Since it counts with considerably lower response levels than the others, it is probably composed of leftover spikes. Individual e shows a similar pattern to a/b/d although considerably more confounded but since its response is comparable in terms of magnitude it could result from combining two individuals that should be further divided.

4.5.2 Debatable decisions

During the interviews, users were also confronted with a few of their decisions that diverged from the majority. The interviewer asked about two to three observations of each stimulus type in each of the following categories: approved by majority but rejected by interviewee; and rejected by majority but approved by interviewee. After being presented with the selected observations, they were asked to explain their decision reasoning. Figure 4.21 shows three cases of moving bars stimulus and Figure 4.22 shows four cases of gratings stimulus, with justifications provided in the captions.

In the figures, the main justification for rejecting moving bars observations is the RF being small (mentioned three times), and the response map peak's distinctiveness is also mentioned once. With gratings observations, justifications are more varied but "selectivity for stimulus attributes" was a generic statement that appeared three times, once concerning specifically motion direction and another concerning gratings orientation. Waveform shapes were also mentioned.

Besides the illustrated cases, we also highlight the following facts. User 3 could not explain one of their outstanding rejections and one of their outstanding approvals. User 4 was confronted with three cases were they had approved the obser-


(a) Individual activity

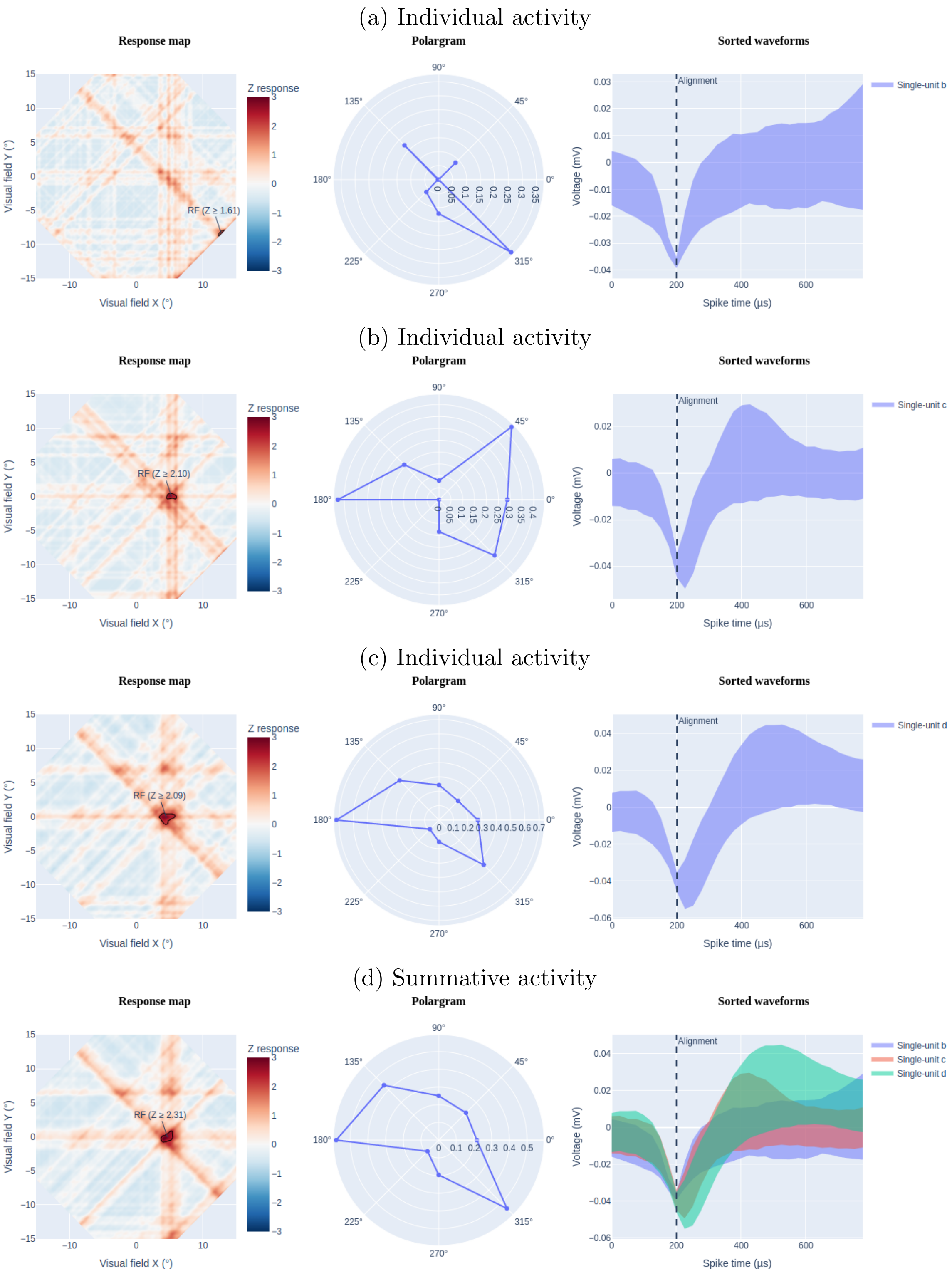
(b) Individual activity

(c) Individual activity

(d) Summative activity

Figure 4.18: Differences between moving bars summative and individual observations 1

vation but could not explain why, saying they would reject them if analyzing them right now and attributing this confusion to fatigue or rush. User 8 was confronted with one outstanding rejection and simply said “it looks fine and I would approve it


(a) Individual activity

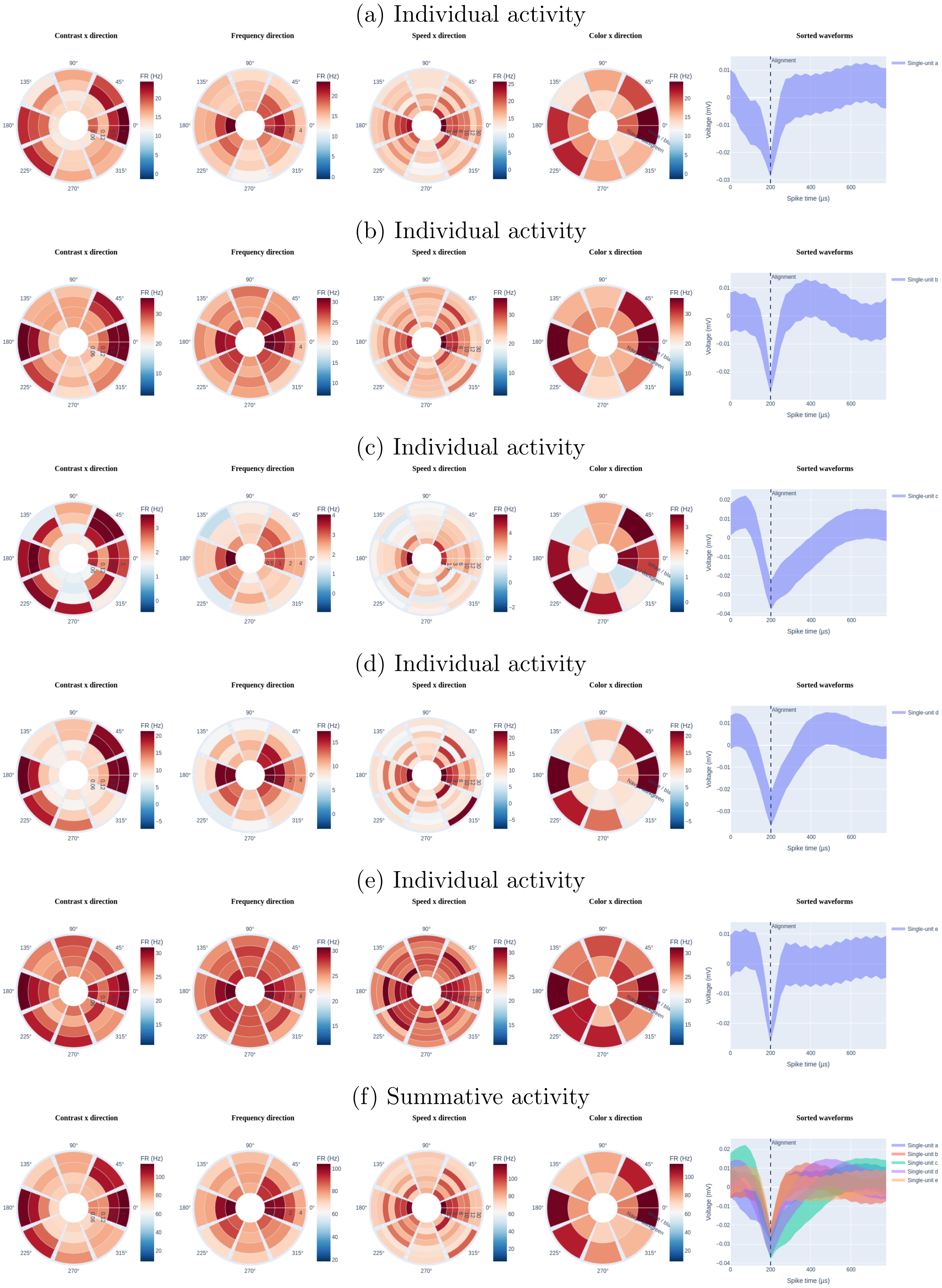
(b) Individual activity

(c) Individual activity

(d) Summative activity

Figure 4.19: Differences between moving bars summative and individual observations 2

now”. User 6 explained that one of their rejections was motivated by a comparison with a very alternative from the same screen, leading to the conclusion that the latter seemed more plausible.



(a) Individual activity

(b) Individual activity

(c) Individual activity

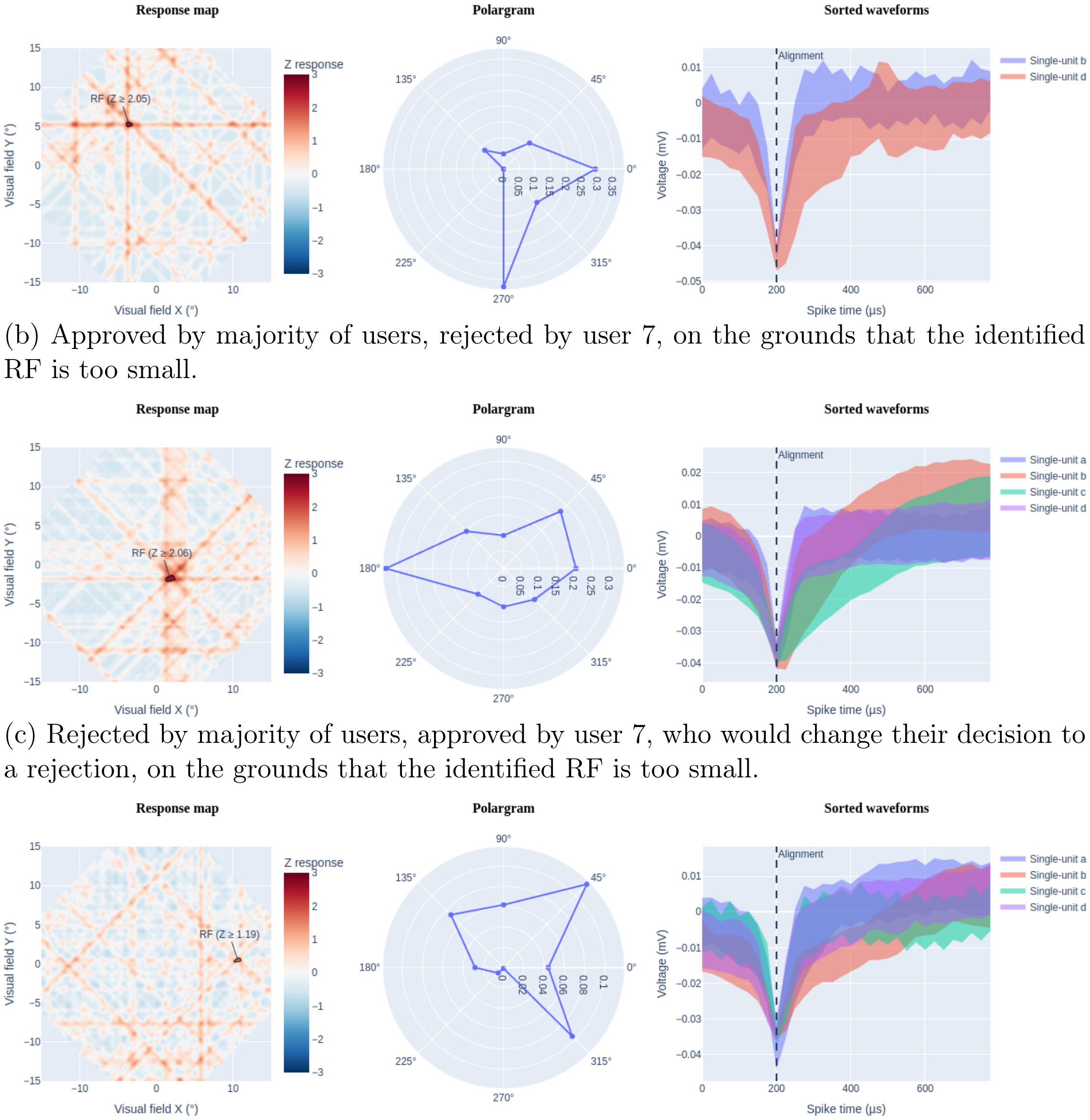
(d) Individual activity

(e) Individual activity

(f) Summative activity

Figure 4.20: Differences between gratings summative and individual observations

Figure 4.21: Debatable moving bars decisions

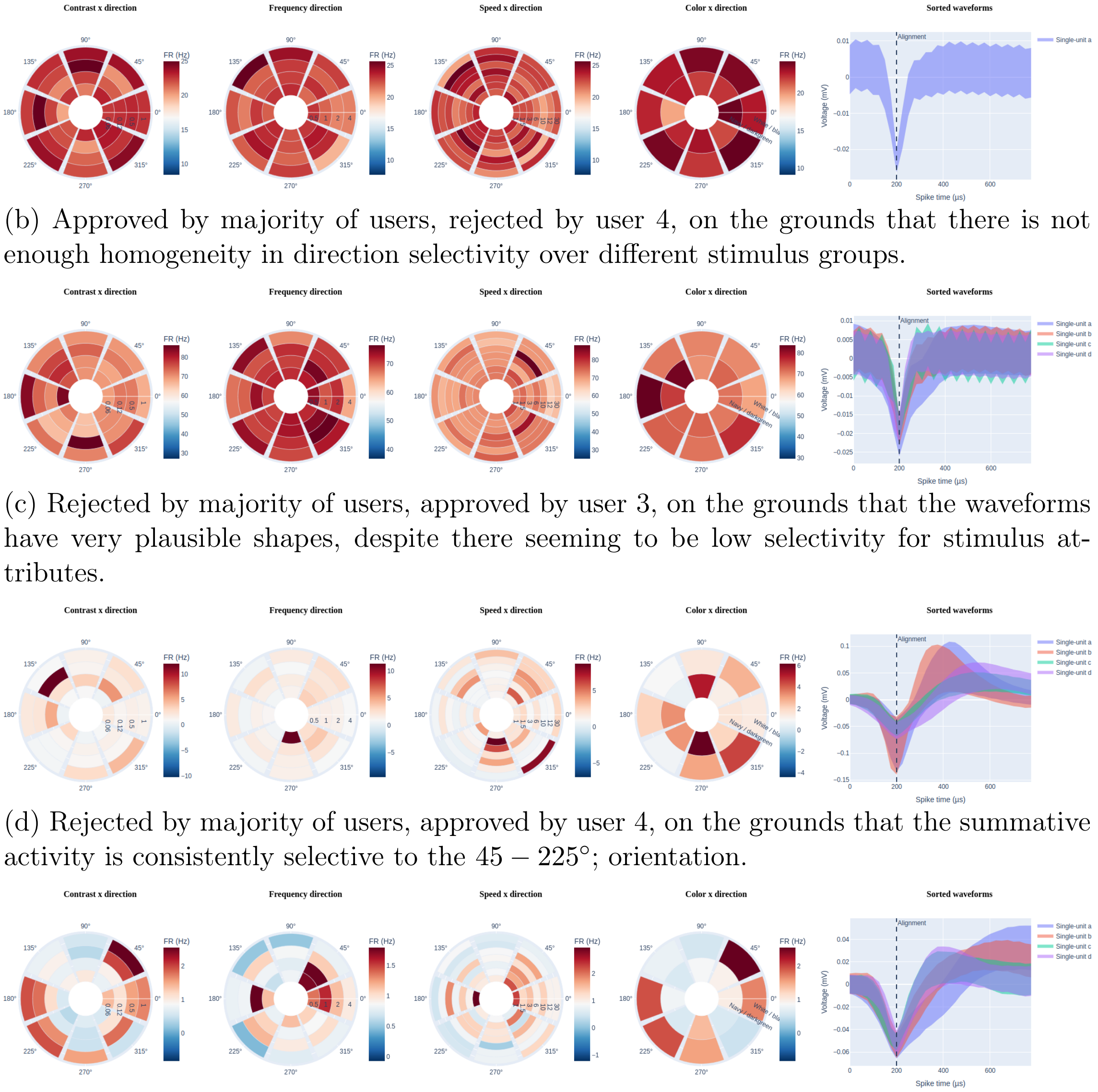


(a) Approved by majority of users, rejected by user 3, on the grounds that the identified RF is too small and its response peak is not relevantly distinct from other local maxima.

(b) Approved by majority of users, rejected by user 7, on the grounds that the identified RF is too small.

(c) Rejected by majority of users, approved by user 7, who would change their decision to a rejection, on the grounds that the identified RF is too small.

Figure 4.22: Debatable gratings decisions



(a) Approved by majority of users, rejected by user 3, on the grounds that there is only one constituent individual that shows no selectivity for stimulus attributes.

(b) Approved by majority of users, rejected by user 4, on the grounds that there is not enough homogeneity in direction selectivity over different stimulus groups.

(c) Rejected by majority of users, approved by user 3, on the grounds that the waveforms have very plausible shapes, despite there seeming to be low selectivity for stimulus attributes.

(d) Rejected by majority of users, approved by user 4, on the grounds that the summative activity is consistently selective to the $45 - 225°$; orientation.

Chapter 5

Discussion and conclusion

5.1 Design validation and limitations

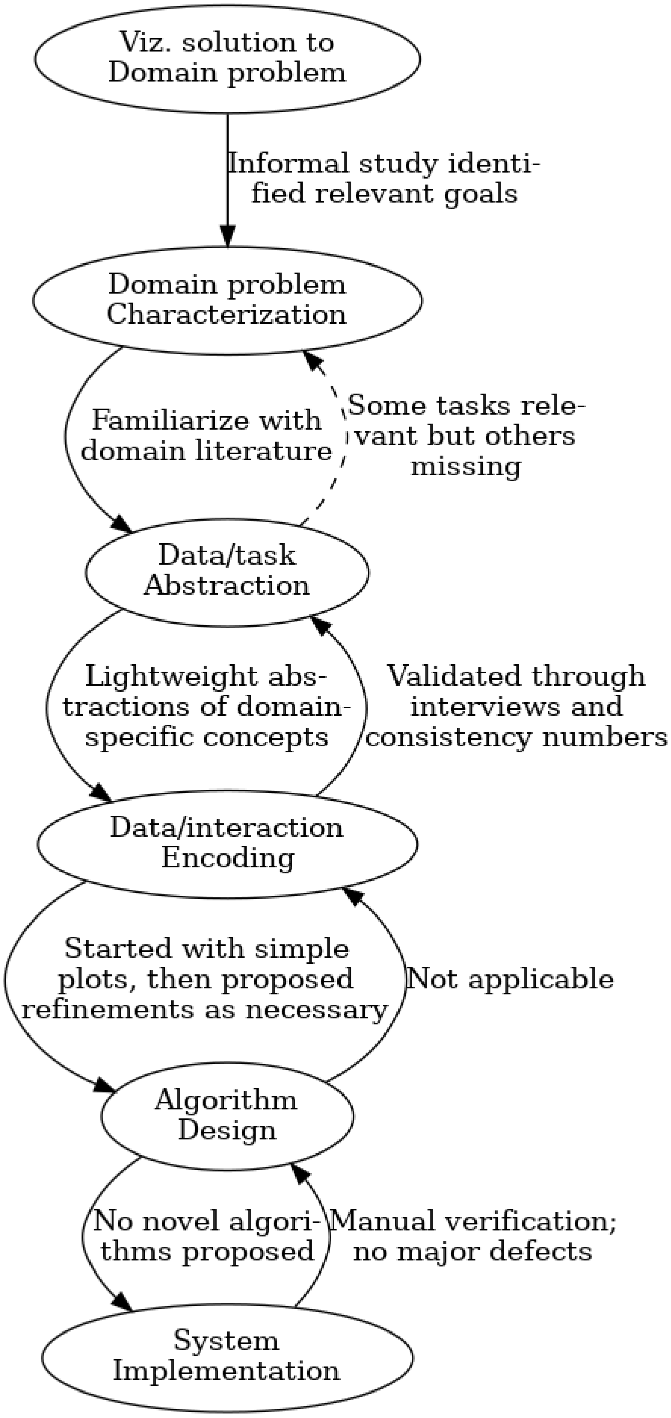
Let us recall the nested validation model described in Section 2.2.3. Here, we identify how we approached the downstream and upstream validation steps, justifying our views and identifying steps that were not performed. The diagram in Figure 5.1 provides an overview and the following text provides details, first following in the downstream direction, then returning in the reverse direction. Throughout this discussion, we will refer to some design study pitfalls identified by SEDLMAIR *et al.* [99] using the same *PF-N* notation as in their paper but the first occurrence will always be stated in full text.

5.1.1 Downstream path

Informal user study importance

The informal user study was a long-lived iterative process comprised of several encounters that discussed problems, datasets, data processing and visualizations — in total, we performed between ten and twenty encounters, some via web conferencing and others *in situ*, spread over four years. Therefore, it was a cornerstone for downstream validation of most steps prior to algorithm design. When considering visualization as a potential solution to a domain problem, the first encounters made it clear that there existed analysis scalability and data quality issues, as introduced in Section 1.4, and that these were amenable to a design study. In terms of domain problem characterization, these encounters contributed to get familiar with domain literature — starting with GATTASS *et al.* [47], FIORANI *et al.* [56], PERES *et al.* [68], then DAYAN and ABBOT [3], GERSTNER *et al.* [34], REY *et al.* [74], among others cited in Chapter 2 — which was fundamental for modeling the data (see Section 3.2 and Appendix A), perform dataset curation, and implement data pro-

Figure 5.1: Outcome of design validation


Viz. solution to Domain problem
Informal study identified relevant goals
Domain problem Characterization
Familiarize with domain literature
Some tasks relevant but others missing
Data/task Abstraction
Lightweight abstractions of domain-specific concepts
Validated through interviews and consistency numbers
Data/interaction Encoding
Started with simple plots, then proposed refinements as necessary
Not applicable
Algorithm Design
No novel algorithms proposed
Manual verification; no major defects
System Implementation

cessing/derivation routines (Section 2.1.1). When the time came for choosing data encoding schemes, we started from plots used in that literature, before proposing additional encodings — *i.e.,* two iterations of gratings activity panels and waveform plots, as reported in Section 3.4, — then arranging all elements into the web interface used for the design study (Section 3.5).

Data/task abstraction and design study scope restriction

Regarding data/task abstraction, we initially attempted to follow a task analysis framework [96] to reduce the risk of proposing uncanny tasks and encodings. However, this framework ended up providing overly abstract concepts (like searching for known targets *vs.* browsing, locating trends and outliers, comparing distributions, *etc.*) that would be interesting for an EDA design focused on multiple of the user goals we catalogued in Section 3.3.1. Even with the help of a higher-level framework for identifying user goals [97], it was still challenging to map the diffuse, very domain-specific analytical tasks into sets of lower-level visualization tasks. However, once we identified the higher-level steps in their workflow — namely, **designing** experiments, **collecting** data, **deriving** putative summative responses, **winnowing** neuronal observation populations, **exploring** and **characterizing** population distributions, and **investigating** hypotheses — and found out that they had concerns about the quality of their derivations (*i.e.,* concerns about spike detection and sorting by third party tools), we decided to narrow down our study scope to the derivation/winnowing phases, as discussed in Section 3.3.2. That also contributed to reduce prototyping and engineering costs besides risks to validation, since an EDA system covering all the steps from derivation to investigation would involve an immense number of tasks and encodings, on the viz side, and more sophisticated data modeling, storage, processing, and caching, on the engineering side.

Brief mention about algorithms and implementation

Finally, the last step in the downstream direction was verifying that chosen algorithms had acceptable runtime and memory asymptotic complexities before moving on to implement the system. That was unnecessary in our case, since we did not propose any new visualization or neuroscience data processing algorithms, relying on existing methods for data processing (see Section 2.1.1 and Section 2.1) and on an open-source library [118] for rendering scalable vector graphics (SVG) graphs in a web environment. Implementation considerations regarding the efficiency of information retrieval, rendering, and caching, to name a few, are interesting software engineering problems that exist inside of/below the node labeled "System Implementation" but these are outside of this thesis' scope, so we will not discuss them here.

Refer to Section 5.3.5 for a brief acknowledgment of technical aspects. We do highlight that performance issues are an important concern for an extended exploratory design, as we will briefly discuss later on.

5.1.2 Upstream path

System verification and algorithmic validation

Now, let us now recurse back the steps in the upstream direction. Starting at system verification, the lower-level dataset processing and derivation routines were unit-tested — although not in a strict test-driven development (TDD) approach [119] but rather partially — to ensure its correctness. Data processing and derivation routines, including PLX file loading, derived data saving, array and matrix operations, signal processing, projection and clustering algorithms, and ML algorithms were all implemented on top of widespread open-source software libraries [116, 120–123]. The web application, implemented on top of another open-source framework[118], was not verified using formal testing techniques, only manual testing by the author, including state verification of its data caching and storage layers before and after performing various combinations of tracked user actions — *i.e.,* logging in and out, approving and rejecting an observation, opening the details panel, and navigating between screens. Nonetheless, during the user study only one defect was reported by one of the study participants, involving polargrams eventually not being rendered (see the $\mathrm{U}_{\mathrm{PolMiss}}$ statement in Section 4.5). Other than that, we assume that full coverage of summative observations by at least one action per user who completed the proposed activity set in the database — that is, for each user, at least one action per observation was registered — is enough indication that no major defects exist in relation to telemetry. Validating algorithm design, as explained, was not necessary.

Interviews and quantitative results regarding encoding and tasks

The user study and the post-activity evaluation interviews were the main tools at our disposal for validating data/interaction encoding and data/task abstraction, although the ideal for the latter would be performing a field study. Time-and-error measurements were out of question since there was no baseline methodology to compare against, not to mention that neuronal observations are inherently noisy, thence marking user decisions as correct or incorrect would be problematic from start. Therefore, we must rely on a combination of user statements and quantitative results to make inferences about the effectiveness of chosen encodings and the relevance of the observation winnowing goal. Let us revisit the quantitative and qualitative results to discuss their implications in more detail.

The plots chosen to encode functional attributes themselves were adequate, as enforced by the many user statements about decision factors ($\mathrm{F_{FR}}$, $\mathrm{F_{RFPol}}$, $\mathrm{F_{RFCent}}$, $\mathrm{F_{RFPeak}}$, $\mathrm{F_{RFSz}}$, $\mathrm{F_{RFZ}}$, $\mathrm{F_{WFNoise}}$, $\mathrm{F_{WFDec}}$, and $\mathrm{F_{RMDec}}$) and confirmed by the similarity in user decisions. That is not to say they are the most effective (which again is a hard verification in the absence of an error measurement) nor that they are free of issues. For instance, gratings panels were perceived by two users as hard to interpret ($\mathrm{P_{GrHard}}$), and the suggestion for presenting firing rates in a z-normalized scale ($\mathrm{S_{FRZ}}$), or at least making it possible to alternate between absolute and normalized firing rates is critical for design refinement. As a matter of fact, most of the aforementioned decision factors, with the exception of $\mathrm{F_{FR}}$, regard moving bars observations or waveforms and moving bars had much better IUDS than gratings. However, during the users' confrontation with conflicting results, they mentioned "selectivity for stimulus attributes", "direction selectivity", and "orientation selectivity" when discussing their positions regarding gratings observations, and $\mathrm{P_{GrHard}}$ was actually denied by one participant. Therefore, we may conjecture that the difficulty of some users is due to: their unfamiliarity with the gratings encodings, since they are novel proposals, unlike response maps, polargrams, and even waveform plots; the greater complexity of this stimulus type, since it contains more parameters to which the observation may be selective. That point of view is reinforced by the slower TTD distributions of gratings observations. All the same, the quantitative results — namely the IUDS indices and regression model performances — about both moving bars and gratings observations show that user decisions are regular enough to form clusters of think-alike users and to be predictable with substantial accuracy, which can only imply that the respective encodings were able to carry enough information for users to behave as such.

Nonetheless, the encoding of gratings functional attributes was questioned by User 7 (as a comment to $\mathrm{W_{HypSel}}$) to be fragile to the presence of perpendicularly direction/orientation-selective cells in the same summative profile, favoring the impression that the summative observation is not selective, where in fact its individuals are complementary selective. That is a point of attention for gratings encodings, suggesting that they might be more appropriate for comparing individual observations than rating the quality of summative observations, where they might lead to "false negatives".

The most challenging points regarding encodings are not related to the presentation of neuronal selectivity and RFs quality but rather to the arrangement of alternative observations and navigation. That was highlighted by user statements such as $\mathrm{E_{HypId}}$, $\mathrm{E_{RowAttr}}$, $\mathrm{E_{RowHyp}}$, where users assumed that observation labels, vertical stacking of same-stimulus observations, and horizontal arrangement of different-stimulus observations, respectively, encoded some sort of attribute or relationship when they did not. Navigation through anatomical attributes and selection of derivation parameters were purposefully obfuscated to prevent selection and anchoring biases, among others, and to reduce the effect of fatigue on user decisions (since screens were shuffled between users), allowing us to make a subsequent analysis of decision similarity and predictability (see Section 4.3 and Section 4.4) with greater confidence.

Abstract tasks and viz solution to a domain problem

On the upstream direction, the most appropriate course of validation for task abstraction would normally be a field study, letting users apply the system to perform work with real data in their native environment. We instead conducted a user study (or a laboratory study) consisting of a predefined list of activities to be performed on a selection of data of their own. We did so because we intended to verify the applicability of machine learning to reduce the amount of work required to winnow observations (topic of Section 4.4). Although the scope of conclusions is narrower in this case, we took care to propose a usage scenario that is incidental to their workflow (after all they have to winnow observations before analyzing functional attribute populations), and the quantitative results confirmed that the amount of work may be reduced by a semi-automated learning approach, although a compromise needs to be made between precision (of classifying or ranking observations) and efficiency (of reducing the amount of manual intervention).

By considering the favorable statements gathered during the interviews in conjunction with the quantitative data, we might conclude that the winnowing goal is indeed relevant to the domain experts and that lower-level tasks involving the identification and comparison of RFs for the included stimulus types can be successfully performed by the proposed user interface elements. Nonetheless, missing functionality like the ability to toggle individual observations ($\mathrm{W}_{\mathrm{HypSel}}$) and to upload data files and download analysis results ($\mathrm{W}_{\mathrm{UpDat}}$), not to mention the aforementioned issues with missing anatomical/parametrical attributes, confirm that the built tool does not completely cover the derivation/winnowing goals. Furthermore, the fact that User 8 declared feeling confused by the proposed activity ($\mathrm{P}_{\mathrm{ActConf}}$) and that Users 3 and 4 declared becoming tired as they performed the activity set only served to stress how augmentation by artificial intelligence (AI) would be an essential feature for a viz system covering goals all the way down from derivation to investigation.

Consequently, given that our system only allowed users to perform a predefined summative observation winnowing activity on a slice of the data, and not to tweak parameters, upload their own files, or winnow individual observations — as noted by themselves during evaluative interviews (see Section 4.5) — it is certain that they would not be interested in using the final a tool in a realistic scenario. Since the aforementioned requirements are missing for an ecological validity study to be performed, we can safely conclude that tasks abstraction is incomplete and that our contributions lie before the solution to a viz problem.

5.2 Discussion of quantitative results

5.2.1 General results implications

Unanimous decisions covered 33% of the gratings observations and 38% of the moving bars observations, including both unanimous approvals and rejections. Classifying the remaining observations is tricky because the specialists grounded their decisions on multiple factors (*e.g.,* the absolute size or sharpness of a RF, the presence of plausible waveforms, or polargram shapes) but they weighed them in differently, even in a contradictory manner, and not even in a strictly self-consistent manner, since they changed a few of their decisions during the evaluative interviews (Section 4.5).

Applying a majority rule to binary-classify observations as “good” or “bad”, as we initially did in Section 4.4, is possible but not necessarily the best choice. After all, we saw that a significant portion of observations caused a 3 × 3 tie (7% of gratings and 3% of moving bars) and that different thresholds could be used to classify moving bars observations as a way to both partition the work better among multiple participants and to force reconsideration and discussion of more controversial decisions (Section 4.3.3).

In the gratings case, the decision process was overall more difficult, as reported by the participants themselves (Section 4.5.1, specially $\mathrm{P}_{\mathrm{GrHard}}$) and confirmed by lower IUDS indices (Section 4.4) and the linear profile of approved population size as function of the approval threshold (Figure 4.11). Nonetheless, if we chose a strict majority rule, considerable portions of the dataset would be discarded (42% of gratings and 68% of moving bars observations).

To a great extent, the quality issues in the dataset used for the user study seem to be localized (Section 4.1.2), after all a few anatomical positions concentrated most of the high-approval observations (about 12 and 17 out of 64 areas contain only majority-approved observations of moving bars and gratings types, respectively) with a great many (40 and 17, respectively) featuring only major rejections, and yet many others (12 and 30, respectively) featuring variable amounts of approvals/rejections, with a clear tendency for spatial clustering of high approval observations. We did not observe a strong relationship between approval rates and the physiological attribute (Section 4.1.2) but it should be emphasized that the dataset consists of a single recording, therefore relevant associations could still exist in other recording sessions. In relation to parametrical attributes (Section 4.1.3), the approval of moving bars observations was considerably better in the first recording session but not significantly sensitive to the spike sorting method, whereas signal type, waveform alignment, and spike detection thresholds had meaningful impacts on gratings observations although this impact remains elusive due to the low cardinality of these dimensions.

And what does the lower average approval rates for moving bars observations tell us? Some users have stated that absence of RFs makes them ignore gratings activity altogether but they approved many gratings observations where no corresponding moving bars ones were approved. It is difficult to conclude anything here, since they were instructed to winnow observations from different stimuli independently from each other, therefore, maybe they just followed the instructions. Alternatively, they may have been more cautious with gratings observations, avoiding to discard some of them when uncertain. Nonetheless, the independence between moving bars and gratings decisions shows that the data may present interesting patterns in one, none, or both stimulus types.

5.2.2 User behavior implications

Four users performed the entire activity set in one single session, while two others logged into the system up to four times, nonetheless distributing their total time unevenly between these sessions (Figure 4.5). The slower user took 1 minute and 10 seconds per anatomical location and the fastest took less than 30 seconds, on average (Figure 4.6). The users who participated in the informal user study were not particularly faster or slower than the others.

Most decisions (about 98.47% for moving bars and 96.82% for gratings) were taken without opening the details view (Section 4.2.3), suggesting that despite missing legends and having lower definition, the chosen encodings were as informative in their smaller versions as in the larger ones, which facilitates packing and achievin higher information-to-ink ratio. That was true even for gratings decisions, which were reportedly harder to make and more controversial than moving bars ones. As a matter of fact, User 5 did not open the details view at all, while Users 4/6 only opened the details view twice for gratings decisions and never for moving bars ones. Although User 3 made heavier usage of details, that still only represented 4.36% of their gratings and 6.27% of their moving bars decisions.

Most of the time (82%), users made a decision about all observations presented in a screen before moving on and whenever they did not, at most one additional visit to the same screen was required (Section 4.2.3). Furthermore, they rarely took more than 5 seconds to reevaluate their choices before moving on (Section 4.2.3).

Regarding the TTD (Section 4.2.3), three quarters of decisions about moving bars observation could be made in up to 5 seconds for approvals and 3 seconds for rejections, whereas three quarters of decisions about gratings observations could be made in up to 10 seconds, regardless of being approval or rejection decisions. The time spent on each observation varied considerably between different users, and the same user tended to vary several seconds depending on stimulus type and decision outcome. We could not find convincing evidence that overall consensus or personal divergence from it have a meaningful impact on the time spent on each observation, other than a few Spearmann correlation coefficients with substantial values that might suggest some users take longer to decide about more debatable observations. There is also no conclusive evidence that experience makes the assignment any faster. Therefore, it is not clear what makes them spend time on each observation.

An important limitation of this analysis is that we cannot guarantee that users did not shift their attention through different observations on each same screen. Even if we enhanced telemetry by registering mouse hovering events, that would still miss on what they are visually focusing on. One possible direction would be regressing normalized TTDs from on a combination of functional attributes and decision outcomes (*i.e.,* approved or rejected). If a regression model achieved reasonable performance on this task, we would have an indication of what makes it easier/harder (read faster/slower) to judge an observation, except for a user-dependent scale factor. Finally, although viewing details corresponded to increased TTD, we cannot infer any relationships accurately due to the low incidence of viewing details.

5.2.3 Decision similarity implications

The IUDS indices revealed considerable disagreement between users. Moving bars observations showed the highest overall agreement, with a cluster of users (3-6) being very consistent in their rejections (above 5/6) but less so in their approvals (above 3/5). Gratings observations were subject to a lot more disagreement, with all but three rejection indices above 1/2 and approval indices above 1/2 in the average. Furthermore, moving bars approval and rejection IUDS were positively considerably correlated but gratings were not considerably correlated. These facts and numbers confirm decisions are indeed very subjective and influenced by contrasting perceptions of what constitutes "good" observations. If the decisions showed a high degree of consistency, we might consider a collaborative tool design where an observation would not need to be evaluated so many times by different users. Therefore, in the absence of more objective criteria for deciding when observations should be accepted or rejected, a compromise must be made between specialist time and careful appreciation of data idiosyncrasies. More will be said about this on the next section.

5.2.4 Approval regression

We used AutoML to investigate the potential for rating observations automatically based on their functional attributes, essentially the same information as the study participants themselves had access to — actually to the whole set of functional attributes, whereas users only visualized a subset of them. Using ML to augment domain expert productivity requires the appreciation of a few subtleties.

First, what exactly is the system going to augment? If multiple observations are to be compared and ranked, so that a user may selectively inspect a few options based on their scores, then regression is a proper choice. However, if the objective is to simply save expert time by detecting *evidently bad* and *evidently good* observations that can be discarded/saved right away, thus setting aside the rest for irrevocable manual inspection, then classification is the proper choice.

Second, how confident can we be on our data? Our user study did not produce a staggering amount of data points, and the interviews demonstrated that users were not strict, in a formal sense, when making decisions, possibly drifting from an ideal decision function over time because of fatigue, distraction, experience, visualization quality, or other random/unaccounted for effects. Therefore, observation approval rates are, at best, noisy proxies for observation quality (from a regression perspective), and categorizing observations as *approved*, *rejected*, or *requires inspection* requires establishing thresholds on top of these noisy indicators.

Summative moving bars

For summative moving bars regression, the best performance (AutoML) gave a MSE of 0.009, which corresponds to an average error of approximately 0.095 in the regressed approval rate, while the worst model's (RLR) MSE corresponds to a mean absolute error of 0.152. Therefore, the average error in summative moving bars regression was less than flipping 1/6 votes for any model. The AutoML binary and ternary classifiers delivered near perfect performance and the non-AutoML ones also delivered promising performances, in the binary case with perfect recall and precision no worse than 5/6 (that is, flagging false majority-approved observations up to 1/6 of times), and with even better precision numbers in the ternary case, where conversely at least 2/3 of problematic cases can be expected to be identified and routed for designated users to analyze. In terms of ranking efficiency, it was impossible to draw any meaningful conclusions, since only 3 pairs of observations were eligible for comparison out 58 samples in the test set.

According to Table 4.13, the GBR model has overfit the training set by choosing a number of estimators greater than the number of points, which led to perfect score on the training set and implies its hyperparameters should be reviewed despite the reasonable performance on the test set compared to the AutoML baseline. Other than that, learning curves in Appendix C suggest that the RLR and SVR models have topped their performance given the available features, and could perhaps be improved by adding extra features to compensate for their higher bias. Since the CV-selected hyperparameters for these models actually discarded features via PCA and we made all functional attributes available to them, these models are likely to be indicating the true performance given the inherent noise of approval rates. If that is the case, this could only be remedied by inviting additional participants to improve the stability of the approval rates. A higher number of data points would also contribute to increase confidence in the discussed numbers.

Summative gratings

In the summative gratings case, all models achieved basically the same performance, with MSEs that correspond to an average error of 0.179, which sits between flipping 1/6 and 1/5 votes. The binary and ternary classifiers delivered less ideal performances although still surprising given the issues already discussed about this stimulus type. Surprisingly, the best compromise between precision and recall was achieved by the RLR classifier, with $P = 0.857$ and $R = 0.923$ in the binary case, and $P_{\mathfrak{R}} = 0.9$, $P_{\mathfrak{A}} = 0.867$ and $R_{\mathfrak{M}} = 0.857$ in the ternary case. Ranking by the AutoML and GBR models was nearly as effective as flipping a coin, while RLR and SVR were just marginally better (0.619 and 0.667 accuracy, respectively). That can be explained by the higher MSE, as it takes less that the mean absolute error to flip the ranks of two observations with an approval difference of only 1/6 votes. Differences in classifier performance despite the similar regression error are probably due to each model's ability to faithfully represent the decision boundaries in the respective feature space.

Again, the CV-selected hyperparameters for each model shown in Table 4.13 provides relevant insights to complement the learning curves in Appendix C. The GBR model was not overfit this time, as shown by its train MSE evolution but it possibly requires more samples to generalize better and improve validation performance. The number of estimators (50) was not so large as for moving bars (450) but the step size (50) should be reduced during CV since it is too large compared to the number of points. The closing of training and CV loss curves for RLR and SVR models suggest a high-bias (underfitting) scenario but if we consider that all models actually had similar performance and that AutoML is likely to present a baseline, then the base error we observe may just be a consequence of inherent problem difficulty. After all, we known from other quantitative/qualitative results that gratings observations were harder to judge, so these models are probably just hitting the best possible performance.

Individual models

All individual moving bars models had similar regression performance, where the best MSE (achieved by GBR) of 0.061 corresponds to a staggering mean absolute error of 0.247. As for gratings models, MSE values were slightly more varied but still performed all too similarly when we consider the mean absolute errors (between 0.214 and 0.245). Curiously, gratings models performed better as binary classifiers, as ranking machines, and also as ternary classifiers (considering the F1 score of all metrics shown in Table 4.14). Actually, some individual ranking machines performed better than their summative counterparts but that may be explained, in part, by greater numbers of evaluation ranking pairs.

Pinpoint conclusions are harder with these models, because assigning the same approval rate to all individuals in the same summative group inserted an unknown amount of imprecision. Most learning curves (with the exception of gratings GBR) point to a high bias scenario, which reinforces that point of view.

Final thoughts on models

In all cases, the AutoML model provided the best regression score, or was second-best by a MSE difference of merely 0.001, whereas other models alternated ranks throughout problems. In the summative gratings and individual moving bars cases, all models had basically the same regression performance, which suggests a saturation of the learning problem. The first scenario likely took place because gratings observations were overall harder to judge and caused wavier decision criteria to be applied, whereas the second was probably due to the aforementioned inherent artificial noise.

Summative moving bars provided the best regression and classification performances, followed by gratings, with taller baselines due to theirs inherent difficulty. The elusive performance of individual regressors and classifiers shows that a dataset of individual observation quality is required for providing more accurate accounts of the ML potential in the individual case.

5.3 Reflections

Developing the work reported in this thesis was an opportunity to learn many things in the fields of information visualization, neuroscience, and data science. We dedicate this section to reflecting about pitfalls of the design study process that may have hindered its progress, and to account for facts that have worked positively for its conclusion. We refer whenever possible to the pitfalls in SEDLMAIR *et al.* [99] using the same PF-N notation as in their paper but also writing out full-text descriptions uppon first occurrence. This discussion is however less concerned with formal aspects and speaks more openly about the author's and advisor's own experience and views.

5.3.1 Domain problem characterization and task analysis difficulty

Doing proper task analysis proved to be challenging. Design study and task analysis frameworks like BREHMER and MUNZNER [95], MUNZNER [96] are ongoing efforts by the viz community to provide researchers with more solid foundations for evaluating work or conducting novel research. But despite the years of accumulated knowledge, the abstract tasks are sometimes "too abstract" and hard to apply, which is why LAM *et al.* [97] proposed the concept of user goals. At some point, we believed to possess a reasonable understanding of the domain problem but could still not list a coherent and self-contained set of abstract tasks operating on/producing well-defined inputs/outputs. We attempted to follow the action-target task typology very strictly but that resulted in an endless inventory of chained abstract tasks like

> *browse anatomical attributes to find summative response profiles → compare two summative response profiles → find peaks in 2D response distribution*

that did not seem to find their way into a terse design specification.

We believe one of the root causes was precisely PF-16 (expecting just talking or fly on wall to work) since we held many interviews and discussions about collaboration directions with the domain experts but never engaged in *in situ* contextual inquiries or even fly on wall to observe them doing work rather than listening to them explain how they work. As a consequence, we ended up focusing on their analysis goals and assumed that we could choose encodings that supported whatever lower-level tasks were required to achieve those goals. The ultimate consequence was PF-19 (too little abstraction), which led to very domain-specific work that does not bring transferable knowledge to viz in general.

Nonetheless, since task abstraction walks hand-in-hand with data abstraction, the strive to find abstract tasks at least helped us understand the data very well, and to come up with a categorization of attributes as *metadata*, *stimulus*, *anatomical*, *physiological*, *parametrical*, or *functional*, that was important for designing the web tool and the winnowing activity, and that remains relevant for future work (as these attributes play different roles in their workflow).

5.3.2 Design study iterations

The informal user study was a great way of identifying issues with proposed encodings and to refine them. The ability to iterate fast over a few months with biweekly-spaced encounters in which we showed wireframe prototypes and discussed impressions helped avoid the risks of PF-22 (non-rapid prototyping) for a while, yet we feel it should also have been applied during the process of building the final tool. The latter was a several-month stretch without any additional communication with the study participants, and additional encounters might have helped to identify issues like the superposition of individuals in gratings polar plots, or the confusion caused by the obfuscation of anatomical attributes.

To some extent, that was a result of the current state of the art of visualization software components. We will comment more about technical difficulties in Section 5.3.5 bur for short, building a responsive web application that efficiently displays sophisticated plots and records user actions requires advanced software development skills and considerably many human-hours, where the first is not the focus of a researcher and the second must be divided with other research activities. The fact that we had little time to finish the tool in this scenario and little input from our collaborators during this phase was therefore the cause for PF-20 (prematurely committing to a design after considering a small design space).

5.3.3 Design limitations

We initially aimed at building an EDA solution to allow formulating and investigating hypotheses involving anatomical, physiological, and functional attributes. Once data quality and preprocessing issues were identified, we expanded the scope to include parametrical attributes, and our vision for the tool started to encompass support for data derivation and filtering. Eventually, the design study we conducted focused only on judging observation quality from a few opaque parametrical choices. Should the proposed abstractions and encodings work for that purpose, the next step would be integrating them into a larger scale, perhaps explicitly encoding parameter choices and allowing for customization, or including additional recording sessions and allowing users to navigate through them based on some encoding of anatomical attributes.

Some of our initial designs, which were left out of this thesis, focused on plotting various classes of attributes using parallel coordinates and allowing users to freely browse sets of individual observations based on their anatomical locations. However, we ultimately concluded that, given the dataset's size, this would lead to an immense exploration space, such that comparing findings of different users would be harder. By proposing a (laboratory) user study, we presented users with an assignment rather than empowering them with a tool, as they were unable to upload data files, download analysis results, or appreciate all the attributes they are interested in, therefore incurring in PF-23 (too little usability). In a sense, we brought up this point to highlight that we recognize usability to be a high-priority attack vector for future work, since tailoring the tool for a specific assignment was a conscious decision to validate the observation judgment task, the proposed encodings, and the applicability of ML to facilitate observation judgment.

5.3.4 Time drainers and research focus

One important time drainer was the necessity to understand and process the PLX file format and to perform dataset curation. After all, in its initial state the dataset was a collection of binary files (containing raw signals and, sometimes, spike sorting results) with a naming convention that indicated some attributes (*i.e.,* subject name, stimulus type, MEA depth on both hemispheres, and session repetition), lacking any relational structure to allow querying attributes, row sets, or relationships, for instance. This situation was similar to PF-14, where *no real data is available*, except in this case data *was available* but *was not ready.*

Another considerable time sink was implementation effort. We can account some of that to changes in the research team composition, as two other students (a graduate and an undergraduate) who participated in the study besides the author left while the EDA design was still on table. The aforementioned PF-22 (non-rapid prototyping) probably also played its role, as the focus during the intermediate phases should not have been on application building but rather on producing cheap and expendable prototypes. In very early stages of research, we even considered converting the PLX data files into the neurodata without borders (NWB) format [124] which would allows us to benefit from software development kits (SDKs), and visualizations produced by the Allen Institute for Brain Science's Brain Observatory project [125]. However, once it became clear that data curation was required anyway, we decided to build a custom data file format akin to NWB for storing recorded signals and use a relational database for the remaining attributes (either original or derived). We did not return to this point later, since dataset preparation became our stopping point.

The third and perhaps most ensnaring time drainer is represented by PF-18 (learning too much of their problems/language). The author became interested in the field of neural coding and the problem of neural correlations, considering if those could be applied to the V2 Dataset. However, that field did not match neither our expertise, nor our collaborators', which led to a considerable time spent on surveying the field rather than applying known-to-work methodologies.

5.3.5 Technical aspects

We implemented multi-layered caching to speed up loading times, preventing any computation in real time. Page loading was kept within 5 seconds, and details panels could be shown with delay inferior to one second. The fact that all attribute derivations were fixed allowed us to precompute them and to cache all plots in main memory, which eliminated concerns about computation time and kept page loading times relatively low. Nonetheless, the use of a scripted programming language and of a low-tier rented web server would certainly be limiting factors for a system that allowed users to freely change derivation parameters.

The technical aspects of implementing an information system for visualizing the data using the proposed encodings and performing the modeled tasks have not been emphasized so far in this text. However, we believe these should be at least briefly considered, as the *viz* literature seems to be rather complacent of the effort required to even prototype these systems. Despite the recommendations for building paper prototypes, the costs of moving to the next step and actually building an information system for performing visualizations tasks often requires a team of developers with system development skills rather than a few graduate students with other responsibilities and a focus on publication. The core concerns that we had to address during development of this system were:

- Reproducibility of data processing
- Storage of original and derived datasets
- Processing power for deriving data and producing plots
- Web deployment
- Application responsiveness

The open-source code of the system is available at `https://gitlab.com/lcg/neuro/v2/vizpike`. It is versioned in a Git repository with over 700 commits and over 47000 lines of code/text in its final version. Secondary repositories developed for supporting this thesis have amassed, in total, nearly 3000 commits and 730000 lines of code/text. That material includes configuration files, source code files, and markup, covering software libraries, system components, research journals, experiment notebooks, and general documentation. Gitlab CI/CD technology was used to facilitate and speed up the deployment of new system versions and Docker-based containerization was employed to maximize portability. Various open-source software libraries were fundamental for implementing this system [116, 120–123, 126, 127].

5.4 Open questions and future directions

5.4.1 Cognitive biases and viz direction

Cognitive biases are failures in human rationality (*e.g.,* memory, perception, judgment, and logic) that occur systematically in the human population (usually without our knowledge but also in spite of our awareness and will to avoid them) that prevent us from making precise assessments of a situation or objectively better decisions [117]. It is a long-studied theme in psychology but only very recently did it begin to gain attention in information visualization. We believe this field is important for addressing the following question:

> Should ecephys visualization systems be top-down or bottom-up?

The domain experts' ultimate goals are to characterize neuronal populations in terms of functional, anatomical, and physiological attributes, and to investigate hypotheses revolving around their joint distributions. When they approach the data (whether using a viz methodology or not), they carry expectations set by prior knowledge in the field, for instance (the three first elements were stated during interviews):

- Functional attributes are continuous along an electrode's penetration path (commented multiple times during informal user study encounters),
- The presence of perpendicularly orientation/direction-selective cells to improve resolution/contrast (stated by User 7 during interviews),
- A strong RF evidence from moving bars stimulus increases the chances of finding gratings-selectivity,
- Modular and cyclical distributions of functional attribute across physiological boundaries [68], and

among others. Therefore, a top-down system for population characterization may induce confirmation bias by leading them to discard populations that do not conform to their expectations (*e.g.,* low cohesiveness among observations in neighboring recording sites). Likewise, a bottom-up system may induce other types of biases, by encouraging to set a high SNR threshold that leads to discarding hypotheses that albeit noisy, are consistent with their neighbors or alternative observations, thus distorting the resulting functional attribute distributions.

Section 4.1.2 showed that the availability of high-quality RFs varies immensely with anatomical positioning, Section 4.1.3 showed that functional attribute distributions were considerably affected by parametrical attributes. The latter enforces the need for additional studies about functional uncertainty due to parametrical diversity, and both facts together make a good case for a comparison between a top-down and a bottom-down viz tool.

5.4.2 Choosing parameters

Considering the population characterization goal, it would be fundamental for users to explore how functional attributes vary with the available parameters. We presented a simple analysis in terms of φ_k correlation coefficients (Section 4.1.3) that only highlighted the relationships between these attribute categories but the user interface should support users in performing search-and-compare tasks for answering questions like "what parameters yield the sharpest RFs?", or "how does spike detection threshold influence direction selectivity?", for example. They could rely on this task as a sort of round-two winnowing, whereby they could elect optimal parameters, thus yielding a final population of observations to actually characterize and explore. Furthermore, there should be ways of linking this exploration to other parameters, like anatomical and physiological ones, after all it might make sense to choose constant or similar parameters per electrode or insertion depth, for instance, rather than allowing completely different setting for each recording site.

5.4.3 Statistical significance

We did not apply statistical significance tests for filtering observations, like the Student's t-test for relevance of non-spontaneous activity *vs.* baseline activity [68]. In that study, they reported only 190 out of 721 observations (about 26%) passing the $p \leq 5\%$ test. Keeping in mind that the $p \leq 5\%$ rule should be applied with care when making statements about the (in)existence of relationships between variables [128], we believe that statistical significance values like that could be included in the visualized attributes, if not applied directly to the filtering of data, as a way to bring an additional dimension for comparing alternative observations. One possibility would be searching for alternative observations using p values as a discriminatory attribute, for example, as a heatmap of p values on dimensionality-reduced 2D parameter axes. Another possible application of statistical significance tests would be grouping same-location observations that failed to pass a test for distinct firing rate distributions, so that users could focus on analyzing fewer alternative observations during the overview cycle.

5.4.4 Neural coding and automatic winnowing

A great deal of neuroscience research, including studies of the visual cortex, focuses on the issues of how information is represented, processed, stored, retrieved, and transmitted in the brain, a field generally known as *neural coding.* Neural coding borrows concepts from information theory, like Fisher information, to quantify the reliability with which stimulus parameters can be inferred (using statistical methods) from an observed population's response [19].

It is a established fact that average firing rates are correlated in populations of neighboring neurons sharing synaptic inputs and that these correlations vary with stimulus [9]. Whether these correlations enhance or impair the efficiency of the neural code for representing stimulus information was a topic of intense debate for years but with higher resolution recording technologies, many recent studies started to show that correlated, stimulus-dependent variability of neural responses in local neural circuits contributes to more precisely inferring stimulus from response [4, 5, 7, 8].

For a while, we dedicated ourselves to survey this field and even began to compute and analyze correlations in the V2 Dataset but changed our focus to the design study before obtaining results that could be integrated into the visualization. We believe this theme deserves further investigation, as we conjecture that a maximally informative selection of observations — that is, choosing a set of observations that maximizes an information measure, thus maximizing precision of stimulus inference — would constitute a relevant attribute for the domain experts to use when deriving, comparing, and winnowing observations. It might even become a feature for decision regression and, most optimistically, a criterion for automatic election of optimal parameters, hence freeing the users to only pursue their population characterization and investigation goals.

5.4.5 Improvements to encodings

Gratings encodings were problematic for visualizing summative profiles in some cases, due to the presence of perpendicularly orientation/direction-selective individuals in the same summative profile. These encodings of firing rates might be kept for inspecting individuals but different encodings (possibly of a firing rate-derived attribute or other attributes) should be developed for representing the summative activity in a way that avoids superposition while still making it clear that selectivity exists inside the summative profile.

Moreover, a few topic improvements that we consider valuable are:

- Highlighting linked sectors in different gratings plots could be leveraged for different purposes, like emphasizing same-orientation/direction responses in various parameter combinations, since these attributes are expected to show consistent responses.

- As mentioned before, allowing to toggle individuals in each summative profile plot, which could be achieved either by clicking or hovering over individuals' legends in the waveforms plot, for example.
- Visualizing additional attributes that were excluded from the present encodings, including latency CVD, and CVO, *etc.* using a complementary panel with a parallel coordinates plot, for instance.
- As a minor change to waveforms plots, the waveform regions could be overlaid with individual waveforms when mouse-hovered, providing both a summary and a detailed view of spike shapes.

5.4.6 Next steps for machine learners

Training and evaluation of ML models was an important part of this thesis. There are a few points we believe should be addressed in subsequent studies:

- Train ternary classifiers directly, rather than building classifiers out of approval rate regressors. After all, by optimizing a proper classification loss function, these models could obtain better performance than our current results.
- Apply the trained summative observation classification models to other recordings of the V2 Dataset, verifying if applicability to other *corpora* of data is possible, or if very different distributions would require retraining/adjustments to feature engineering.
- Study the viability of continuous learning setup using transfer learning techniques [129], which might help with the previous topic and open the doors to a more scalable tool design, whereby users would only ever need to rank a few observations that the models classified with less certainty.
- Increase the number of samples to improve the accuracy and confidence of the performance measures and over/underfitting analyses presented.
- Considering the ranking application, training proper ranking models rather than relying on point-wise approval regression may lead to better performance [130].

5.5 Conclusion

The design study has come a long way since the initial contact with our collaborators. We started from an open-ended proposal of exploring ecephys data from the primate visual cortex as a way of investigating relationships between anatomical, physiological, and functional attributes, detoured to computing neural signal correlations and information measures on the hopes of uncovering unseen facts from the dataset, then turned our attention back to cleaning an initial mass of bi-stimulus neuronal observations.

The latter was accomplished by building a web-based tool and inviting six users to participate in a laboratory activity where we recorded their decisions and usage patterns. We found out that a combination of moving bars functional attribute encodings familiar to them combined with new gratings functional attribute encodings proposed by us allowed them to make reasonably consistent decisions about observations. The overall quality of the dataset, besides the obfuscation of anatomical and physiological attributes impacted negatively on user experience but they were still able to advance cohesive points about the decision factors used when judging observations.

After training four ML models, we were able to regress the users' average approval rate per summative observation and to classify observations according to whether they received majority approval (binary classification), or near unanimous approval/rejection (ternary classifier). While observation associated with the moving bars stimulus showed better overall performance, gratings observations could also be predicted despite a few shortcomings of the proposed encodings, which tend to become cluttered in some cases. Regressing and classifying approval rates of individuals was attempted but the lack of trustworthy data did not allow for robust conclusions. These results showed that ML is applicable to this problem as a way to reduce repetitive human effort, although tighter bounds on the scale of this reduction would benefit from additional samples.

Scientific research is an iterative process and design studies are no exception. We listed many open questions and future directions we would like to have addressed before time ran out. Most prominent among these are the application of trained ML models to other *corpora* of data to test the potential for a semi-automated system where the user goals of of preprocessing, filtering, and interpreting the neural signals can be performed in an integrated fashion, rather than in a cascaded workflow, besides developing the necessary encodings and tasks that would support such a system.

References

[1] KANDEL, E. R., SCHWARTZ, J. H., JESSELL, T. M., et al. *Principles of Neural Science*. 5 ed. 120 South Riverside Plaza, Chicago, IL 60606, USA, McGraw-Hill, 2013. ISBN: 978-0-07-181001-2.

[2] SEUNG, H. S. *Connectome: How the Brain's Wiring Makes Us Who We Are*. 1 ed. 9205 Southpark Center Loop. Orlando, FL 32819, USA, Houghton Mifflin Harcourt Trade, 2012. ISBN: 978-0547508184.

[3] DAYAN, P., ABBOT, L. F. *Theoretical Neuroscience: Computation and Mathematical Modeling of Neural Systems*. 1 ed. 1 Rogers St, Cambridge, MA 02142, USA, MIT Press, 2001. ISBN: 9780262541855.

[4] BÁNYAI, M., LAZAR, A., KLEIN, L., et al. "Stimulus complexity shapes response correlations in primary visual cortex", *Proceedings of the National Academy of Sciences*, v. 116, n. 7, pp. 2723–2732, 2 2019. ISSN: 0027-8424. doi: 10.1073/pnas.1816766116. Available at: <http://dx.doi.org/10.1073/pnas.1816766116>.

[5] FRANKE, F., FISCELLA, M., SEVELEV, M., et al. "Structures of Neural Correlation and How They Favor Coding", *Neuron*, v. 89, n. 2, pp. 409–422, 1 2016. ISSN: 0896-6273. doi: 10.1016/j.neuron.2015.12.037. Available at: <http://dx.doi.org/10.1016/j.neuron.2015.12.037>.

[6] ORBÁN, G., BERKES, P., FISER, J., et al. "Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex." *Neuron*, v. 92, n. 2, pp. 530–543, 10 2016. ISSN: 1097-4199. doi: 10.1016/j.neuron.2016.09.038.

[7] ZYLBERBERG, J., CAFARO, J., TURNER, M. H., et al. "Direction-Selective Circuits Shape Noise to Ensure a Precise Population Code", *Neuron*, v. 89, n. 2, pp. 369–383, 2016. ISSN: 1097-4199. doi: 10.1016/j.neuron.2015.11.019. Available at: <http://dx.doi.org/10.1016/j.neuron.2015.11.019>.

[8] FISCELLA, M., FRANKE, F., FARROW, K., et al. "Visual coding with a population of direction-selective neurons." *Journal of neurophysiology*, v. 114, n. 4, pp. 2485–99, 10 2015. ISSN: 1522-1598. doi: 10.1152/jn.00919.2014. Available at: <http://dx.doi.org/10.1152/jn.00919.2014>.

[9] LIN, I. C., OKUN, M., CARANDINI, M., et al. "The Nature of Shared Cortical Variability", *Neuron*, v. 87, n. 3, pp. 645–657, 8 2015. ISSN: 1097-4199. doi: 10.1016/j.neuron.2015.06.035. Available at: <http://dx.doi.org/10.1016/j.neuron.2015.06.035>.

[10] KANITSCHEIDER, I., COEN-CAGLI, R., KOHN, A., et al. "Measuring Fisher Information Accurately in Correlated Neural Populations", *PLoS Computational Biology*, v. 11, n. 6, pp. 1–27, 2015. ISSN: 15537358. doi: 10.1371/journal.pcbi.1004218.

[11] ECKER, A. S., BERENS, P., COTTON, R. J., et al. "State Dependence of Noise Correlations in Macaque Primary Visual Cortex", *Neuron*, v. 82, n. 1, pp. 235–248, 4 2014. ISSN: 0896-6273. doi: 10.1016/j.neuron.2014.02.006.

[12] HU, Y., ZYLBERBERG, J., SHEA-BROWN, E. "The Sign Rule and Beyond: Boundary Effects, Flexibility, and Noise Correlations in Neural Population Codes", *PLoS Computational Biology*, v. 10, n. 2, pp. 1–22, 2 2014. ISSN: 1553-7358. doi: 10.1371/journal.pcbi.1003469. Available at: <http://dx.doi.org/10.1371/journal.pcbi.1003469>.

[13] COHEN, M. R., KOHN, A. "Measuring and interpreting neuronal correlations", *Nature Neuroscience*, v. 14, n. 7, pp. 811–819, 6 2011. ISSN: 1097-6256. doi: 10.1038/nn.2842.

[14] GRAF, A. B. A., KOHN, A., JAZAYERI, M., et al. "Decoding the activity of neuronal populations in macaque primary visual cortex", *Nature Neuroscience*, v. 14, n. 2, pp. 239–247, 2011. doi: 10.1038/nn.2733.

[15] CHURCHLAND, M. M., YU, B. M., CUNNINGHAM, J. P., et al. "Stimulus onset quenches neural variability: a widespread cortical phenomenon", *Nature Neuroscience*, v. 13, n. 3, pp. 369–378, 3 2010. ISSN: 1097-6256. doi: 10.1038/nn.2501.

[16] ECKER, A. S., BERENS, P., KELIRIS, G. A., et al. "Decorrelated neuronal firing in cortical microcircuits", *Science*, v. 327, n. 5965, pp. 584–587, 2010. ISSN: 00368075. doi: 10.1126/science.1179867.

[17] QUIAN QUIROGA, R., KREIMAN, G. “Measuring sparseness in the brain: Comment on Bowers (2009).” *Psychological Review*, v. 117, n. 1, pp. 291–297, 2010. ISSN: 1939-1471. doi: 10.1037/a0016917. Available at: <`http://doi.apa.org/getdoi.cfm?doi=10.1037/a0016917`>.

[18] BOWERS, J. S. “On the Biological Plausibility of Grandmother Cells: Implications for Neural Network Theories in Psychology and Neuroscience”, *Psychological Review*, v. 116, n. 1, pp. 220–251, 2009. ISSN: 0033295X. doi: 10.1037/a0014462.

[19] AVERBECK, B. B., LATHAM, P. E., POUGET, A. “Neural correlations, population coding and computation”, *Nature Reviews Neuroscience*, v. 7, n. 5, pp. 358–366, 5 2006. ISSN: 1471-003X. doi: 10.1038/nrn1888. Available at: <`http://dx.doi.org/10.1038/nrn1888`>.

[20] MONTANI, F., KOHN, A., SMITH, M. A., et al. “The Role of Correlations in Direction and Contrast Coding in the Primary Visual Cortex”, *Journal of Neuroscience*, v. 27, n. 9, pp. 2338–2348, 2007. ISSN: 0270-6474. doi: 10.1523/JNEUROSCI.3417-06.2007.

[21] SOMPOLINSKY, H., YOON, H., KANG, K., et al. “Population coding in neuronal systems with correlated noise”, *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, v. 64, n. 5 I, pp. 051904/1–051904/11, 2001. ISSN: 15393755. doi: 10.1103/PhysRevE.64.051904.

[22] MARR, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York, NY, USA, Henry Holt and Co., Inc., 1982. ISBN: 0716715678.

[23] LICHTMAN, J. W., PFISTER, H., SHAVIT, N. “The big data challenges of connectomics”, *Nature Neuroscience*, v. 17, pp. 1448–1454, 2014. ISSN: 1546-1726. doi: 10.1038/nn.3837. Available at: <`https://doi.org/10.1038/nn.3837`>.

[24] GOLDSTONE, R. L., PESTILLI, F., BÖRNER, K. “Self-portraits of the brain: Cognitive science, data visualization, and communicating brain structure and function”, *Trends in Cognitive Sciences*, v. 19, n. 8, pp. 462–474, 2015. ISSN: 1879307X. doi: 10.1016/j.tics.2015.05.012.

[25] MOVSHON, A., HELMSTAEDTER, M. “The Great Debate: Connectomics”. 2016. Available at: <`ttps://www.youtube.com/watch?v=uSbNRyY2QH0`>.

[26] MOVSHON, A., SEUNT, S. "Connectomics: Sebastian Seung vs. Tony Movshon". 2012. Available at: <https://www.youtube.com/watch?v=q4KrhDZQ088>.

[27] DEISSEROTH, K. "Optogenetics", *Nature Methods*, v. 8, n. 1, pp. 26–29, jan 2011. ISSN: 1548-7091. doi: 10.1038/nmeth.f.324. Available at: <http://www.nature.com/articles/nmeth.f.324>.

[28] SHI, Y., TOGA, A. W. "Connectome imaging for mapping human brain pathways", *Molecular Psychiatry*, v. 22, n. 9, pp. 1230–1240, 2017. ISSN: 14765578. doi: 10.1038/mp.2017.92. Available at: <http://dx.doi.org/10.1038/mp.2017.92>.

[29] SOTIROPOULOS, S. N., ZALESKY, A. "Building connectomes using diffusion MRI: why, how and but", *NMR in Biomedicine*, v. 32, n. 4, pp. 1–23, apr 2019. ISSN: 0952-3480. doi: 10.1002/nbm.3752. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/nbm.3752>.

[30] JUN, J. J., STEINMETZ, N. A., SIEGLE, J. H., et al. "Fully integrated silicon probes for high-density recording of neural activity", *Nature*, v. 551, n. 7679, pp. 232–236, Nov 2017. ISSN: 1476-4687. doi: 10.1038/nature24636.

[31] JUAVINETT, A. L., BEKHEET, G., CHURCHLAND, A. K. "Chronically implanted neuropixels probes enable high-yield recordings in freely moving mice", *eLife*, v. 8, pp. 1–17, 2019. ISSN: 2050084X. doi: 10.7554/eLife.47188.

[32] AVENA-KOENIGSBERGER, A., MISIC, B., SPORNS, O. "Communication dynamics in complex brain networks", *Nature Reviews Neuroscience*, v. 19, n. 1, pp. 17–33, 2017. ISSN: 1471-003X. doi: 10.1038/nrn.2017.149. Available at: <http://www.nature.com/doifinder/10.1038/nrn.2017.149>.

[33] HODGKIN, A. L., HUXLEY, A. F. "A quantitative description of membrane current and its application to conduction and excitation in nerve", *The Journal of Physiology*, v. 117, n. 4, pp. 500–544, 1952. doi: https://doi.org/10.1113/jphysiol.1952.sp004764. Available at: <https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1952.sp004764>.

[34] GERSTNER, W., KISTLER, W. M., NAUD, R., et al. *Neuronal Dynamics: From single neurons to networks and models of cognition.* 1 ed. Cambridge University Press & Assessment, Shaftesbury Road, Cambridge,

CB2 8EA, United Kingdom, Cambridge University Press, 2014. ISBN: 9781107635197.

[35] ROSENBLATT, F. “The perceptron: A probabilistic model for information storage and organization in the brain”, *Psychological Review*, v. 65, pp. 386–408, 1958. doi: 10.1037/h0042519.

[36] RAJDL, K., LÁNSKÝ, P., KOSTAL, L. “Fano Factor: A Potentially Useful Information”, *Frontiers in Computational Neuroscience*, v. 14, 2020.

[37] SERRE, T. “Hierarchical Models of the Visual System”. In: *Encyclopedia of Computational Neuroscience*, Springer New York, pp. 1–12, New York, NY, 2014. ISBN: 978-1-4614-7320-6. doi: 10.1007/978-1-4614-7320-6_345-1. Available at: <`http://link.springer.com/10.1007/978-1-4614-7320-6{_}345-1`>.

[38] SPILLMANN, L. “Receptive fields of visual neurons: the early years”, *Perception*, v. 43, n. 11, pp. 1145–1176, 2014.

[39] BARLOW, H. B., HILL, R. M., LEVICK, W. R. “Retinal Ganglion Cells Responding Selectively to Direction and Speed of Image Motion in the Rabbit”, *The Journal of physiology*, v. 173, pp. 377–407, oct 1964. ISSN: 0022-3751. Available at: <`http://www.ncbi.nlm.nih.gov/pubmed/14220259http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1368915`>.

[40] HUBEL, D. H., WIESEL, T. N. “Receptive fields of single neurones in the cat’s striate cortex”, *The Journal of physiology*, v. 148, n. 3, pp. 574–591, 1959.

[41] HUBEL, D. H., WIESEL, T. N. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”, *The Journal of physiology*, v. 160, n. 1, pp. 106–154, 1962.

[42] MISHKIN, M., UNGERLEIDER, L. G., MACKO, K. A. “Object vision and spatial vision: two cortical pathways”. v. 6, pp. 414–417, 1983. doi: https://doi.org/10.1016/0166-2236(83)90190-X. Available at: <`https://www.sciencedirect.com/science/article/pii/016622368390190X`>.

[43] MISHKIN, M., UNGERLEIDER, L. G., MACKO, K. A. “Object vision and spatial vision: two cortical pathways”, *Trends in Neurosciences*, v. 6, pp. 414 – 417, 1983. ISSN: 0166-2236. doi: https://doi.org/10.1016/0166-2236(83)90190-X. Available at: <`http://www.sciencedirect.com/science/article/pii/016622368390190X`>.

[44] KRAVITZ, D. J., SALEEM, K. S., BAKER, C. I., et al. "The ventral visual pathway: an expanded neural framework for the processing of object quality." *Trends in cognitive sciences*, v. 17, n. 1, pp. 26–49, 1 2013. ISSN: 1879-307X. doi: 10.1016/j.tics.2012.10.011.

[45] KRAVITZ, D. J., SALEEM, K. S., BAKER, C. I., et al. "A new neural framework for visuospatial processing", *Nature Reviews Neuroscience*, v. 12, n. 4, pp. 217–230, 2011. ISSN: 1471003X. doi: 10.1038/nrn3008. Available at: <`http://dx.doi.org/10.1038/nrn3008`>.

[46] NIELL, C. M. "Vision: More than expected in the early visual system", *Current Biology*, v. 23, n. 16, pp. R681–R684, 2013. ISSN: 09609822. doi: 10.1016/j.cub.2013.07.049. Available at: <`http://dx.doi.org/10.1016/j.cub.2013.07.049`>.

[47] GATTASS, R., LIMA, B., SOARES, J. G., et al. "Controversies about the visual areas located at the anterior border of area V2 in primates", *Visual Neuroscience*, v. 32, n. 2015, pp. E019, oct 2015. ISSN: 0952-5238. doi: 10.1017/S0952523815000188. Available at: <`http://www.journals.cambridge.org/abstract{_}S0952523815000188`>.

[48] FAIRHALL, A. "The receptive field is dead. Long live the receptive field?" *Current Opinion in Neurobiology*, v. 25, pp. ix – xii, 2014. ISSN: 0959-4388. doi: https://doi.org/10.1016/j.conb.2014.02.001. Available at: <`http://www.sciencedirect.com/science/article/pii/S0959438814000361`>.

[49] COX, D. D., DEAN, T. "Neural networks and neuroscience-inspired computer vision", *Current Biology*, v. 24, n. 18, pp. R921–R929, 2014. ISSN: 09609822. doi: 10.1016/j.cub.2014.08.026. Available at: <`http://dx.doi.org/10.1016/j.cub.2014.08.026`>.

[50] POGGIO, T., MUTCH, J., LEIBO, J., et al. "The computational magic of the ventral stream: Sketch of a theory (and why some deep architectures work)." *Mit-Csail-Tr-2012-035*, pp. 1–122, 2012.

[51] SELKET. "Ventral-dorsal streams". `https://commons.wikimedia.org/wiki/File:Ventral-dorsal_streams.svg`. Licensed as CC-by-SA-3.0 (`http://creativecommons.org/licenses/by-sa/3.0/`). Accessed on Nov. 15th 2022., 2007.

[52] SHARPEE, T. O. "Computational Identification of Receptive Fields", *Annual Review of Neuroscience*, 2013. ISSN: 0147-006X. doi: 10.1146/annurev-neuro-062012-170253.

[53] BARLOW, H. B., HILL, R. M. "Selective sensitivity to direction of movement in ganglion cells of the rabbit retina", *Science (New York, N.Y.)*, v. 139, n. 3553, pp. 412–4, feb 1963. ISSN: 0036-8075. Available at: <`http://www.ncbi.nlm.nih.gov/pubmed/13966712`>.

[54] HARTLINE, H. K. "The response of single optic nerve fibers of the vertebrate eye to illumination of the retina", *American Journal of Physiology-Legacy Content*, v. 121, n. 2, pp. 400–415, 1938.

[55] CAPITANIO, J. P., EMBORG, M. E. "Contributions of non-human primates to neuroscience research", *The Lancet*, v. 371, n. 9618, pp. 1126–1135, 2008. ISSN: 01406736. doi: 10.1016/S0140-6736(08)60489-4.

[56] FIORANI, M., AZZI, J. C., SOARES, J. G., et al. "Automatic mapping of visual cortex receptive fields: A fast and precise algorithm", *Journal of Neuroscience Methods*, v. 221, pp. 112–126, 2014. ISSN: 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2013.09.012. Available at: <`https://www.sciencedirect.com/science/article/pii/S0165027013003233`>.

[57] DA SILVA, R. P. *Propriedades funcionais das bandas de citocromo oxidase no córtex visual secundário (V2) de primatas (Sapajus apella)*. Tese de Doutorado, Federal University of Rio de Janeiro, 2017.

[58] AZZI, J. C. B., GATTASS, R., LIMA, B., et al. "Precise visuotopic organization of the blind spot representation in primate V1", *Journal of Neurophysiology*, v. 113, n. 10, pp. 3588–3599, 2015. doi: 10.1152/jn.00418.2014. Available at: <`https://doi.org/10.1152/jn.00418.2014`>. PMID: 25761953.

[59] JANSEN-AMORIM, A., FIORANI, M., GATTASS, R. "GABA-induced inactivation of Cebus apella V2 neurons: effects on orientation tuning and direction selectivity", *Brazilian Journal of Medical and Biological Research*, v. 46, n. 7, pp. 589–600, jul 2013. ISSN: 0100-879X. doi: 10.1590/1414-431X20132859. Available at: <`http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S0100-879X2013000700589{&}lng=en{&}tlng=en`>.

[60] JANSEN-AMORIM, A. K., FIORANI, M., GATTASS, R. "GABA inactivation of area V4 changes receptive-field properties of V2 neurons in Cebus mon-

keys", *Experimental Neurology*, v. 235, n. 2, pp. 553–562, jun 2012. ISSN: 00144886. doi: 10.1016/j.expneurol.2012.03.008. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0014488612001069>.

[61] JANSEN-AMORIM, A. K., LIMA, B., FIORANI, M., et al. "GABA inactivation of visual area MT modifies the responsiveness and direction selectivity of V2 neurons in Cebus monkeys", *Visual Neuroscience*, v. 28, n. 06, pp. 513–527, nov 2011. ISSN: 0952-5238. doi: 10.1017/S0952523811000411. Available at: <http://www.journals.cambridge.org/abstract{_}S0952523811000411>.

[62] MARGULIES, D. S., BÖTTGER, J., WATANABE, A., et al. "Visualizing the human connectome", *NeuroImage*, v. 80, pp. 445–461, 2013. ISSN: 10538119. doi: 10.1016/j.neuroimage.2013.04.111. Available at: <http://dx.doi.org/10.1016/j.neuroimage.2013.04.111>.

[63] SMITH, S. M., BECKMANN, C. F., ANDERSSON, J., et al. "Resting-state fMRI in the Human Connectome Project", *NeuroImage*, v. 80, pp. 144–168, 2013. ISSN: 10538119. doi: 10.1016/j.neuroimage.2013.05.039.

[64] BISWAL, B. B., MENNES, M., ZUO, X.-N., et al. "Toward discovery science of human brain function", *Proceedings of the National Academy of Sciences*, v. 107, n. 10, pp. 4734–4739, 2010. ISSN: 0027-8424. doi: 10.1073/pnas.0911855107. Available at: <http://www.pnas.org/cgi/doi/10.1073/pnas.0911855107>.

[65] BASSETT, D. S., SPORNS, O. "Network neuroscience", *Nature Neuroscience*, v. 20, n. 3, pp. 353–364, 2017. ISSN: 1097-6256. doi: 10.1038/nn.4502. Available at: <http://www.nature.com/doifinder/10.1038/nn.4502>.

[66] SPORNS, O. "Contributions and challenges for network models in cognitive neuroscience", *Nature Neuroscience*, v. 17, n. 5, pp. 652–660, 2014. ISSN: 15461726. doi: 10.1038/nn.3690. Available at: <http://dx.doi.org/10.1038/nn.3690>.

[67] STEVENSON, I. H. I., KORDING, K. K. P. "How advances in neural recording affect data analysis." *Nature neuroscience*, v. 14, n. 2, pp. 139–42, 2011. ISSN: 1546-1726. doi: 10.1038/nn.2731.How. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3410539{&}tool=pmcentrez{&}rendertype=abstract{%}5Cnhttp://www.nature.com/neuro/journal/v14/n2/abs/nn.2731.html>.

[68] PERES, R., SOARES, J. G. M., LIMA, B., et al. "Neuronal response properties across cytochrome oxidase stripes in primate V2", *Journal of Comparative Neurology*, v. 527, n. 3, pp. 651–667, feb 2019. ISSN: 00219967. doi: 10.1002/cne.24518. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/cne.24518>.

[69] MARCONDES, M., ROSA, M. G., FIORANI, M., et al. "Distribution of cytochrome oxidase-rich patches in human primary visual cortex", *Journal of Comparative Neurology*, v. 527, n. 3, pp. 614–624, feb 2019. ISSN: 0021-9967. doi: 10.1002/cne.24435. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/cne.24435>.

[70] PERENTOS, N., KRSTULOVIC, M., MORTON, A. J. "Deep brain electrophysiology in freely moving sheep", *Scientific Reports*, v. 32, n. 4, pp. 763–774.e4, jan 2022.

[71] JIANG, Z., HUXTER, J. R., BOWYER, S. A., et al. "TaiNi: Maximizing research output whilst improving animals' welfare in neurophysiology experiments", *Scientific Reports*, v. 7, n. 1, pp. 8086, aug 2017.

[72] HOMER, M. L., NURMIKKO, A. V., DONOGHUE, J. P., et al. "Sensors and Decoding for Intracortical Brain Computer Interfaces", *Annual Review of Biomedical Engineering*, v. 15, n. 1, pp. 383–405, jul 2013. ISSN: 1523-9829. doi: 10.1146/annurev-bioeng-071910-124640. Available at: <https://www.annualreviews.org/doi/10.1146/annurev-bioeng-071910-124640>.

[73] MAHAJAN, S., HERMANN, J. K., BEDELL, H. W., et al. "Toward Standardization of Electrophysiology and Computational Tissue Strain in Rodent Intracortical Microelectrode Models", *Frontiers in Bioengineering and Biotechnology*, v. 8, may 2020. ISSN: 2296-4185. doi: 10.3389/fbioe.2020.00416. Available at: <https://www.frontiersin.org/articles/10.3389/fbioe.2020.00416https://www.frontiersin.org/article/10.3389/fbioe.2020.00416/full>.

[74] REY, H. G., PEDREIRA, C., QUIAN QUIROGA, R. "Past, present and future of spike sorting techniques", *Brain Research Bulletin*, v. 119, pp. 106–117, 2015. ISSN: 18732747. doi: 10.1016/j.brainresbull.2015.04.007. Available at: <http://dx.doi.org/10.1016/j.brainresbull.2015.04.007>.

[75] CHEN, M., EBERT, D. S. "An Ontological Framework for Supporting the Design and Evaluation of Visual Analytics Systems", *Computer Graphics*

Forum, v. 38, n. 3, pp. 131–144, jun 2019. ISSN: 0167-7055. doi: 10.1111/cgf.13677. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/cgf.13677>.

[76] BATTLE, L., HEER, J. "Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau", *Computer Graphics Forum*, v. 38, n. 3, pp. 145–159, jun 2019. ISSN: 0167-7055. doi: 10.1111/cgf.13678. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/cgf.13678>.

[77] KEIM, D. A., MANSMANN, F., SCHNEIDEWIND, J., et al. "Visual Analytics: Scope and Challenges". In: *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, pp. 76–90, Berlin, Heidelberg, Springer Berlin Heidelberg, 2008. ISBN: 978-3-540-71080-6. doi: 10.1007/978-3-540-71080-6_6.

[78] SEJNOWSKI, T. J., CHURCHLAND, P. S., MOVSHON, J. A. "Putting big data to good use in neuroscience", *Nature Neuroscience*, v. 17, n. 11, pp. 1440–1441, 2014. ISSN: 1097-6256. doi: 10.1038/nn.3839. Available at: <http://www.nature.com/doifinder/10.1038/nn.3839>.

[79] GUPTA, S., KAR, A. K., BAABDULLAH, A., et al. "Big data with cognitive computing: A review for the future", *International Journal of Information Management*, v. 42, pp. 78–89, 2018. ISSN: 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2018.06.005. Available at: <https://www.sciencedirect.com/science/article/pii/S0268401218304110>.

[80] *OmniPlex® User Guide*, 17 ed. Plexon Inc., 6500 Greenville Avenue, Suite 700, Dallas, Texas 75206, USA, 10 2018. Obtained at https://plexon.com/wp-content/uploads/2017/06/OmniPlex-User-Guide.pdf. Last access on 2019-01-23.

[81] BUZSÁKI, G., ANASTASSIOU, C. A., KOCH, C. "The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes", *Nature Reviews Neuroscience*, v. 13, n. 6, pp. 407–420, jun 2012. ISSN: 1471-003X. doi: 10.1038/nrn3241. Available at: <http://www.nature.com/articles/nrn3241>.

[82] GOLD, C., HENZE, D. A., KOCH, C., et al. "On the Origin of the Extracellular Action Potential Waveform: A Modeling Study", *Journal of Neurophysiology*, v. 95, n. 5, pp. 3113–3128, may 2006. ISSN: 0022-3077. doi: 10.1152/jn.00979.2005. Available at: <https://www.physiology.org/doi/10.1152/jn.00979.2005>.

[83] ABELES, M., GOLDSTEIN, M. "Multispike train analysis", *Proceedings of the IEEE*, v. 65, n. 5, pp. 762–773, 1977. doi: 10.1109/PROC.1977.10559.

[84] MCNAUGHTON, B. L., O'KEEFE, J., BARNES, C. A. "The stereotrode: A new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records", *Journal of Neuroscience Methods*, v. 8, n. 4, pp. 391–397, aug 1983. ISSN: 01650270. doi: 10.1016/0165-0270(83)90097-3. Available at: <`https://linkinghub.elsevier.com/retrieve/pii/0165027083900973`>.

[85] GRAY, C. M., MALDONADO, P. E., WILSON, M., et al. "Tetrodes markedly improve the reliability and yield of multiple single-unit isolation from multi-unit recordings in cat striate cortex", *Journal of Neuroscience Methods*, v. 63, n. 1-2, pp. 43–54, dec 1995. ISSN: 01650270. doi: 10.1016/0165-0270(95)00085-2. Available at: <`https://linkinghub.elsevier.com/retrieve/pii/0165027095000852`>.

[86] SMITH, S. W. "The Scientist and Engineer's Guide to Digital Signal Processing". `https://www.analog.com/en/education/education-library/scientist_engineers_guide.html`, 1999.

[87] *Offline Sorter^{TM}*, 4.5.0 ed. Plexon Inc., 6500 Greenville Avenue, Suite 700, Dallas, Texas 75206, USA, 6 2020. Obtained at `https://plexon.com/wp-content/uploads/2020/01/Offline-Sorter-v4-User-Guide.pdf`. Last access on 2022-07-16.

[88] DAVIES, D. L., BOULDIN, D. W. "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. PAMI-1, n. 2, pp. 224–227, apr 1979. ISSN: 0162-8828. doi: 10.1109/TPAMI.1979.4766909. Available at: <`http://ieeexplore.ieee.org/document/4766909/`>.

[89] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., et al. "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, v. 12, pp. 2825–2830, 2011.

[90] KAK, A. C., SLANEY, M. *Principles of computerized tomographic imaging.* 345 East 47th Street, New York, NY 10017-2394, United Stated of America, IEEE Publishing, 2001. ISBN: 0-87942-198-3. Available online at http://www.slaney.org/pct/pct-toc.html. Last visited on 2022-07-22.

[91] DURRANI, T. S., BISSET, D. "The Radon transform and its properties", *Geophysics*, v. 49, n. 8, pp. 1180–1187, 1984.

[92] MAZUREK, M., KAGER, M., VAN HOOSER, S. D. "Robust quantification of orientation selectivity and direction selectivity", *Front Neural Circuits*, v. 8, pp. 92, aug 2014.

[93] *Reading PLX and DDT files with Matlab*. Plexon Inc., 6500 Greenville Avenue, Suite 700, Dallas, Texas 75206, USA, 11 2005. Contained in the package available at `https://plexon.com/wp-content/uploads/2017/08/OmniPlex-and-MAP-Offline-SDK-Bundle_0.zip`. Last access on 2019-01-23.

[94] MUNZNER, T. "A Nested Model for Visualization Design and Validation", *IEEE Transactions on Visualization and Computer Graphics*, v. 15, n. 6, pp. 921–928, 2009. doi: 10.1109/TVCG.2009.111.

[95] BREHMER, M., MUNZNER, T. "A Multi-Level Typology of Abstract Visualization Tasks", *IEEE Transactions on Visualization and Computer Graphics*, v. 19, n. 12, pp. 2376–2385, 2013. doi: 10.1109/TVCG.2013.124.

[96] MUNZNER, T. *Visualization Analysis and Design*. 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742, United States, CRC Press, 2014. ISBN: 9781466508910.

[97] LAM, H., TORY, M., MUNZNER, T. "Bridging from Goals to Tasks with Design Study Analysis Reports", *IEEE Transactions on Visualization and Computer Graphics*, v. 24, n. 1, pp. 435–445, 2017. doi: 10.1109/TVCG.2017.2744319.

[98] SHNEIDERMAN, B., PLAISANT, C. "Strategies for Evaluating Information Visualization Tools: Multi-Dimensional in-Depth Long-Term Case Studies". In: *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, p. 1–7, New York, NY, USA, 2006. Association for Computing Machinery. ISBN: 1595935622. doi: 10.1145/1168149.1168158. Available at: <`https://doi.org/10.1145/1168149.1168158`>.

[99] SEDLMAIR, M., MEYER, M., MUNZNER, T. "Design Study Methodology: Reflections from the Trenches and the Stacks", *IEEE Transactions on Visualization and Computer Graphics*, v. 18, n. 12, pp. 2431–2440, 2012. doi: 10.1109/TVCG.2012.213.

[100] TUFTE, E. R. "The visual display of quantitative information", *The Journal for Healthcare Quality (JHQ)*, v. 7, n. 3, pp. 15, 1985.

[101] WILKINSON, L. "The grammar of graphics". In: *Handbook of computational statistics*, Springer, pp. 375–414, Berlin, Heidelberg, 2012.

[102] PRETORIUS, A. J., WIJK, J. J. V. "What Does the User Want to See? What do the Data Want to Be?" *Information Visualization*, v. 8, n. 3, pp. 153–166, 2009. doi: 10.1057/ivs.2009.13.

[103] IZBICKI, R., DOS SANTOS, T. M. *Aprendizado de máquina: uma abordagem estatítica.* 1 ed. São Carlos, SP, Brazil, Independent publishing. Available online at `http://www.rizbicki.ufscar.br/AME.pdf`, 2020. ISBN: 978-65-00-02410-4.

[104] HE, X., ZHAO, K., CHU, X. "AutoML: A survey of the state-of-the-art", *Knowledge-Based Systems*, v. 212, pp. 106622, 2021. ISSN: 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.106622. Available at: <`https://www.sciencedirect.com/science/article/pii/S0950705120307516`>.

[105] FEURER, M., KLEIN, A., EGGENSPERGER, KATHARINA SPRINGENBERG, J., et al. "Efficient and Robust Automated Machine Learning". In: *Advances in Neural Information Processing Systems 28 (2015)*, pp. 2962–2970, 2015.

[106] LLOYD, D., DYKES, J. "Human-centered approaches in geovisualization design: investigating multiple methods through a long-term case study", *IEEE Trans Vis Comput Graph*, v. 17, n. 12, pp. 2498–2507, 12 2011.

[107] BORLAND, D., TAYLOR, R. M. "Rainbow color map (still) considered harmful", *IEEE Computer Graphics and Applications*, v. 27, n. 2, pp. 14–17, 2007. ISSN: 02721716. doi: 10.1109/MCG.2007.323435.

[108] VAN DER MAATEN, L., HINTON, G. "Visualizing data using t-SNE", *Journal of machine learning research*, v. 9, n. Nov, pp. 2579–2605, 2008.

[109] HUISMAN, S. M. H., VAN LEW, B., MAHFOUZ, A., et al. "BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome", *Nucleic Acids Research*, v. 45, n. 10, pp. 1–11, jan 2017. ISSN: 13624962. doi: 10.1093/nar/gkx046. Available at: <`http://graphics.tudelft.nl/Publications-new/2017/HVMPHMVRL17`>.

[110] BAAK, M., KOOPMAN, R., SNOEK, H., et al. "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics". 2018. Available at: <`https://arxiv.org/abs/1811.11440`>.

[111] DEGROOT, M. H., SCHERVISH, M. J. *Probability and Statistics.* 4 ed. 75 Arlington Street, Suite 300, Boston, MA 02116, United States, Addison-Wesley, 2012. ISBN: 0-321-50046-6.

[112] MOSTAFA, Y. A., MAGDON-ISMAIL, M., LIN, H.-T. *Learning from Data: a Short Course.* United States, AMLBook.com, 2012. ISBN: 1600490069.

[113] FRIEDMAN, J. H. "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, v. 29, n. 5, pp. 1189–1232, 2001. doi: 10.1214/aos/1013203451.

[114] SMOLA, A. J., SCHÖLKOPF, B. "A tutorial on support vector regression", *Statistics and Computing*, v. 14, n. 3, pp. 199–222, aug 2004. ISSN: 0960-3174. doi: 10.1023/B:STCO.0000035301.49549.88.

[115] CAWLEY, G. C., TALBOT, N. L. "On over-fitting in model selection and subsequent selection bias in performance evaluation", *Journal of Machine Learning Research*, v. 11, pp. 2079–2107, 2010. ISSN: 15324435.

[116] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., et al. "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, v. 12, pp. 2825–2830, 2011.

[117] DIMARA, E., FRANCONERI, S., PLAISANT, C., et al. "A Task-Based Taxonomy of Cognitive Biases for Information Visualization", *IEEE Transactions on Visualization and Computer Graphics*, v. 26, n. 2, pp. 1413–1432, feb 2020. ISSN: 1077-2626. doi: 10.1109/TVCG.2018.2872577. Available at: <`https://ieeexplore.ieee.org/document/8476234/`>.

[118] DEVELOPMENT TEAM, P. "Dash". `https://dash.plotly.com/`. Last visited on 2022-07-24., 2015–2022.

[119] BECK. *Test Driven Development: By Example.* USA, Addison-Wesley Longman Publishing Co., Inc., 2002. ISBN: 0321146530.

[120] HARRIS, C. R., MILLMAN, K. J., VAN DER WALT, S. J., et al. "Array programming with NumPy", *Nature*, v. 585, n. 7825, pp. 357–362, 7 2020. doi: 10.1038/s41586-020-2649-2. Available at: <`https://doi.org/10.1038/s41586-020-2649-2`>.

[121] VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", *Nature Methods*, v. 17, pp. 261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[122] PANDAS DEVELOPMENT TEAM, T. “pandas-dev/pandas: Pandas”. 2 2020. Available at: <https://doi.org/10.5281/zenodo.3509134>.

[123] WES MCKINNEY. “Data Structures for Statistical Computing in Python”. In: Stéfan van der Walt, Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference*, pp. 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[124] RÜBEL, O., TRITT, A., DICHTER, B., et al. “NWB:N 2.0: An Accessible Data Standard for Neurophysiology”, *bioRxiv*, p. 523035, 2019. ISSN: 2692-8205.

[125] INSTITUTE, A. B. “Allen Brain Observatory”. http://observatory.brain-map.org, 2016.

[126] KLUYVER, T., RAGAN-KELLEY, B., PÉREZ, F., et al. “Jupyter Notebooks ? a publishing format for reproducible computational workflows”. In: Loizides, F., Scmidt, B. (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87–90. IOS Press, 2016. Available at: <https://eprints.soton.ac.uk/403913/>.

[127] FEURER, M., EGGENSPERGER, K., FALKNER, S., et al. “Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning”, *arXiv:2007.04074 [cs.LG]*, 2020.

[128] WASSERSTEIN, R. L., SCHIRM, A. L., LAZAR, N. A. “Moving to a World Beyond “p $<$ 0.05””, *American Statistician*, v. 73, n. sup1, pp. 1–19, 2019. ISSN: 15372731. doi: 10.1080/00031305.2019.1583913. Available at: <https://doi.org/10.1080/00031305.2018.1583913>.

[129] WEISS, K., KHOSHGOFTAAR, T. M., WANG, D. “A survey of transfer learning”, *Journal of Big Data*, v. 3, n. 1, pp. 9, dec 2016. ISSN: 2196-1115. doi: 10.1186/s40537-016-0043-6.

[130] RAHANGDALE, A., RAUT, S. “Machine Learning Methods for Ranking”, *International Journal of Software Engineering and Knowledge Engineering*, v. 29, n. 6, pp. 729–761, 2019. ISSN: 02181940. doi: 10.1142/S021819401930001X.

[131] YEO, I. N., JOHNSON, R. A. “A new family of power transformations to improve normality or symmetry”, *Biometrika*, v. 87, n. 4, pp. 954–959, 2000. ISSN: 00063444. doi: 10.1093/biomet/87.4.954.

Appendix A

Data dictionary

Here, we describe all attributes in the dataset. Table A.1 describes anatomical and physiological attributes. Table A.2 describes parametrical attributes related to spike detection. Table A.3 describes parametrical attributes related to spike sorting. Table A.4 describes other parametrical attributes. Table A.5 describes functional attributes. Table A.6 describes functional attributes related to the moving bars stimulus type. Table A.7 describes functional attributes related to the gratings stimulus type.

Table A.1: Anatomical and physiological attributes

Name	Domain
Brain area	$\{\text{V2}\}$
Brain hemisphere	$\{\text{Left}, \text{Right}\}$
Recording depth	$[0, \infty)$ μm
MEA X	$\mathbb{N}$
MEA Y	$\mathbb{N}$
CytOx band	$\{\text{Thick}, \text{Think}, \text{InterbandsI}, \text{InterbandII}\}$

Table A.2: Parametrical attributes (spike detection)

Name	Domain
Highpass filter cutoff	$[0, \infty)$ Hz
Highpass filter order	$\mathbb{N}$
Highpass filter family	$\{\text{Bessel}\}$
Highpass filtering mode	$\{\text{Bidirectional}\}$
Max. overlapping waveform samples	$\mathbb{N}$
Waveform samples after alignment	$\mathbb{N}$
Waveform samples before alignment	$\mathbb{N}$
Raw signal sampling	$[0, \infty)$ Hz
Spike detection threshold	$[0, \infty)$
Raw signal type	$\{\}$
Waveform alignment mode	$\{Thresholdcrossing, Firstvalley\}$
Waveform samples	$\mathbb{N}$
Waveform sampling rate	$[0, \infty)$ Hz

Table A.3: Parametrical attributes (spike sorting)

Name	Domain
Sorting clustering algorithm	$\{PCA\}$
Sorting feature space	$\{Waveformsamples, Waveformfeatures\}$
Sorting maximum clusters	$\mathbb{N}$
Sorting projection dimensions	$\mathbb{N}$
Sorting projection algorithm	$\{\}$
Sorting type	$\{Embedded, Reimplemented\}$
Sorting upsampling	$\mathbb{N}$

Table A.4: Parametrical attributes (others)

long_label	Domain
Experiment repetition	$\{N\}$
Source signal preffix	$\{SPK, SPKC, WB\}$
Pre-threshold samples	$\{N\}$
Source signal sampling	$[0, \infty)$ $\{Hz\}$
Source signal type	$\{Spiketrain, Spike - continuous, Wideband\}$
Waveform alignment mode	$\{Thresholdcrossing, Firstvalley\}$
Waveform samples	$\{N\}$
Waveform sampling rate	$[0, \infty)$ $\{Hz\}$

Table A.5: Functional attributes (common)

Name	Domain
Basal rate	$[0, \infty)$ Hz
Individuals	$\mathbb{Z}_{\geq 0}$
Total spikes	$\mathbb{Z}_{\geq 0}$

Table A.6: Functional attributes (moving bars)

Name	Domain
Response map	$[-15°, 15°]^2 \rightarrow (-\infty, \infty)$
RF area	$[0, \infty)^{\circ^2}$
RF aspect ratio	$[0, 1]$
RF convex area	$[0, \infty)^{\circ^2}$
RF eccentricity	$[0, \infty)^\circ$
RF equivalent diameter	$[0, \infty)^\circ$
RF euler number	$\mathbb{Z}_{\geq 0}$
RF major axis	$[0, \infty)^\circ$
RF minor axis	$[0, \infty)^\circ$
RF peak disparity	$[0, \infty)$
RF peak response	$(-\infty, \infty)$
RF perimeter	$[0, \infty)^\circ$
RF response cutoff	$[0, \infty)$
RF total response	$(-\infty, \infty)$
Latency	$[0, \infty)$ ms
0° polargram direction	$[0, 1]$
45° polargram direction	$[0, 1]$
90° polargram direction	$[0, 1]$
135° polargram direction	$[0, 1]$
180° polargram direction	$[0, 1]$
225° polargram direction	$[0, 1]$
270° polargram direction	$[0, 1]$
315° polargram direction	$[0, 1]$
1st strongest polargram direction	$[0, 1]$
2nd strongest polargram direction	$[0, 1]$
3rd strongest polargram direction	$[0, 1]$
4th strongest polargram direction	$[0, 1]$
5th strongest polargram direction	$[0, 1]$
6th strongest polargram direction	$[0, 1]$
7th strongest polargram direction	$[0, 1]$
8th strongest polargram direction	$[0, 1]$

Table A.7: Functional attributes (gratings)

Name	Domain
Active direction CV	$[0, 1]$
Active direction index	$[0, 1]$
Active orientation CV	$[0, 1]$
Active orientation index	$[0, 1]$
Active preferred direction	$[0, \infty)$
Active preferred orientation	$[0, 180)^\circ$
Direction CV	$[0, 1]$
Direction index	$[0, 1]$
Orientation CV	$[0, 1]$
Orientation index	$[0, 1]$
Preferred direction	$[0, \infty)$
Preferred orientation	$[0, 180)^\circ$

Appendix B

Feature engineering

In this section, we detail the procedures for turning the functional attributes from the V2 Dataset into scaled features suitable for making predictions — that is, into unit-less scalars mostly contained in the $[-1, 1]$ and $[0, 1]$ ranges. In the end, we also present feature correlation matrices and scatter plots of PCA-projected features. Refer to the data dictionary in Appendix A for their domains and categories. For each of the transformations described below, assume we are transforming each scalar variable $x \in X$ independently of other variables. Summations and minimum/maximum computations iterate over the set of values found in the dataset, X.

B.1 Angular encoding

The variables in this group have angular semantics, therefore the easiest way to map them into an Euclidean setting is to replace each variable $x \in [0, 360)^\circ$ by a pair of variables,

$$x \mapsto (\cos x, \sin x)$$

This is the only transformation in which a dimension is replaced by two additional dimensions. These variables include:

This mapping produces two uncorrelated scalars in the $[-1, 1]$ range.

B.2 Min-max scaling

These variables either belong to a closed interval $[a, b]$ but without characteristic distribution shapes, or belong to a semi-closed interval like $[a, \infty]$. We transform them using

$$x \mapsto \frac{x - \min X}{\max X - \min X}$$

which lies in the $[0, 1]$ range. The following variables belong to this group:

B.3 Standard scaling

These variables have bell-shaped distributions, so we map them according to

$$x \mapsto \frac{x - \bar{X}}{S(X)}$$

where $\bar{X}$ is the mean value of the set X and $S(X)$ is its standard deviation. The resulting variables belong to open ended intervals $[-\infty, \infty]$ but, by definition, they have zero mean and unit variance. The following variables belong to this category:

B.4 Box-Cox power scaling

We use the Box-Cox power transformation [131] with variables that have strictly positive values and a distribution that resembles a log-normal one. The mapping follows

$$x \mapsto \qquad x' = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0, \end{cases} \tag{B.1}$$

where the parameter λ is computed by maximum-likelihood estimation (MLE) — that is, by maximizing the log-likelihood distribution of (λ, μ, σ) given x', assuming that $X' \sim \mathcal{N}(\mu; \sigma)$. The resulting distributions are therefore approximately normal. The following variables were treated this way:

B.5 Yeo-Johson power scaling

The Box-Cox transform fails should x assume negative values. Therefore the Yeo-Johson power transform [131] may be used to obtain normally distributed features. The modified transform is given as

$$x \mapsto \qquad x' = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0, x \geq 0, \\ \log{(x+1)} & \text{if } \lambda = 0, x \geq 0 \\ -\dfrac{(-x+1)^{2-\lambda}-1}{2-\lambda} & \text{if } \lambda \neq 2, x < 0, \\ -\log{(-x+1)} & \text{if } \lambda = 2, x < 0 \end{cases} \tag{B.2}$$

and, once again, λ is estimated through MLE. The following variables were treated this way:

Appendix C

Machine learning model details

Here, we present a few details about the machine learning models discussed in Section 4.4. Figures C.1, C.2, and C.3 contain learning curves of GBR, RLR, and SVR models. Learning curves show the training and validation costs as functions of $|D_{\text{train}}|$, and constitute useful tools for identifying overfitting and underfitting scenarios and for deciding on whether to collect additional data, adjust regularization, obtain new features, or consider other models [112]. At each point of the x-axis, the train/validation scores were computed using K-fold cross-validation (KFCV) on a training set slice of the corresponding size, with $k = 5$ folds. Dots correspond to CV score averages, and the shaded areas around them indicate the standard deviation of scores.

Figure C.1: GBR learning curves

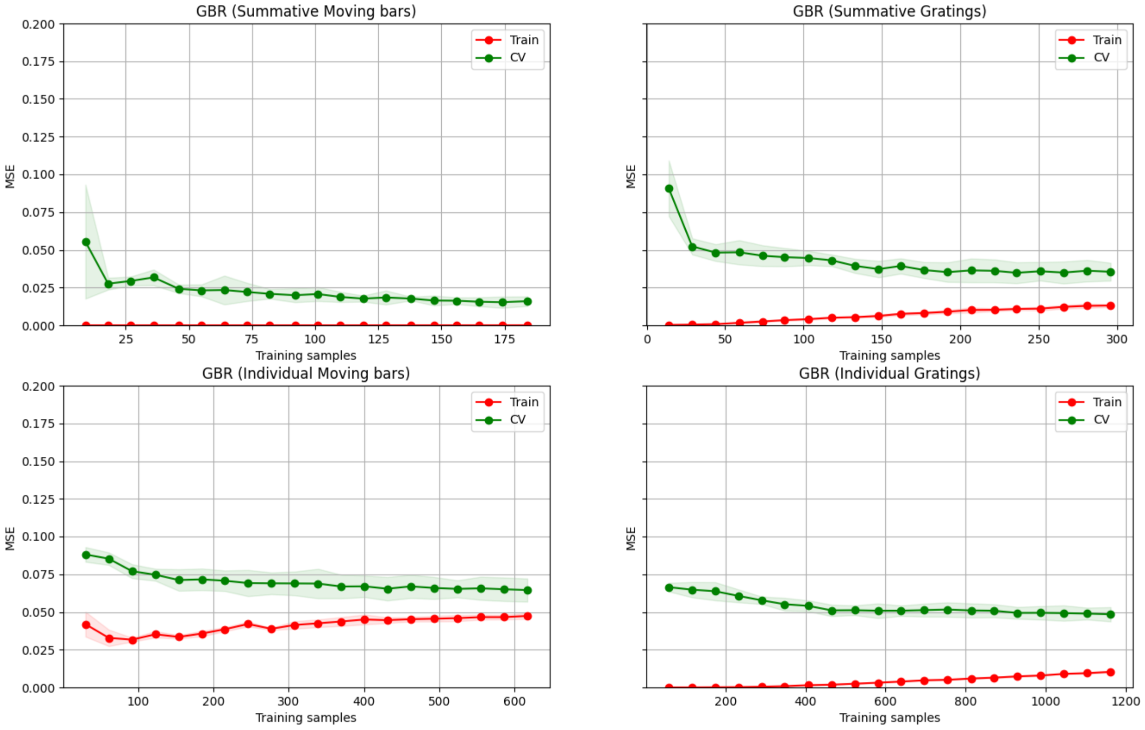


Figure C.2: RLR learning curves

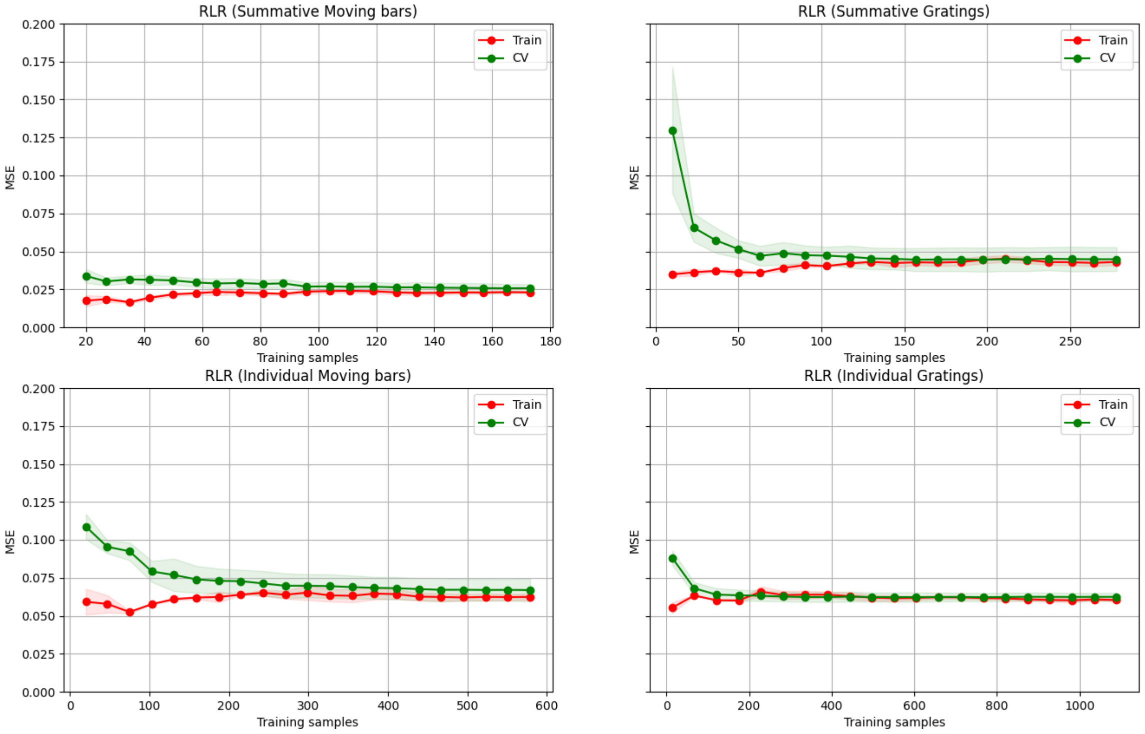
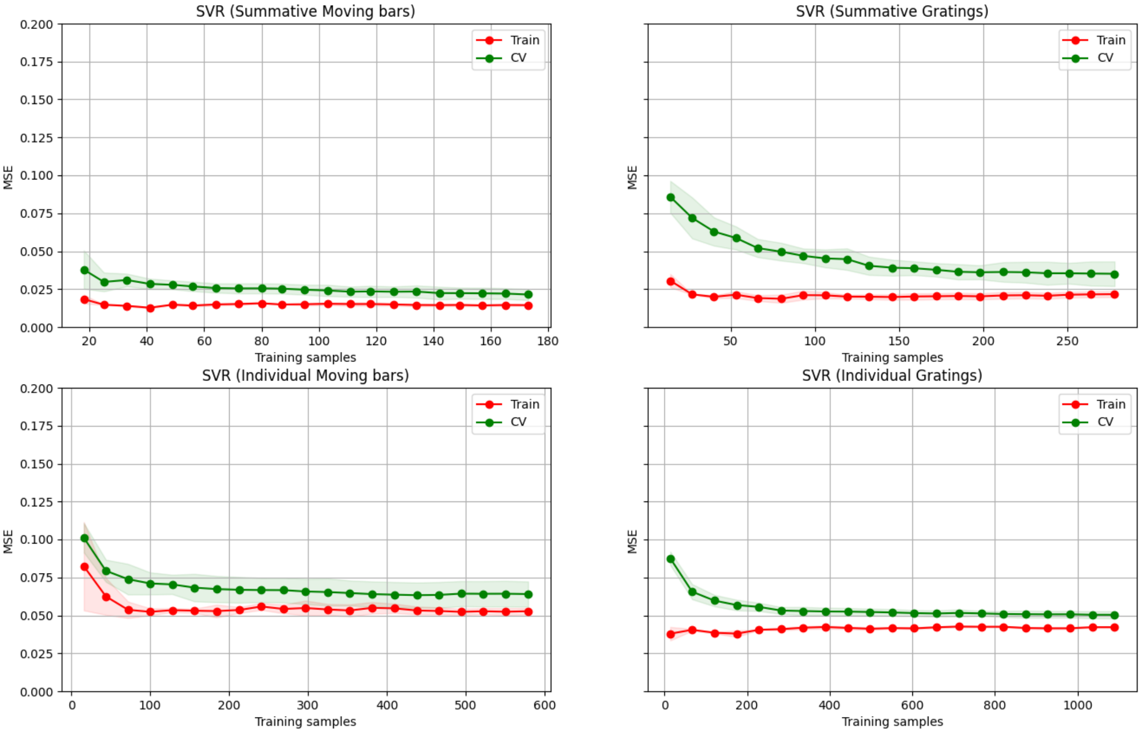


Figure C.3: SVR learning curves



Index

abstract visualization task, 55
action potential, 5, 12
analysis goal, 41
anatomical attribute, 51, 56, 120, 123, 146, 157, 177
attribute, 34, 62
attribute encoding, 68
axon, 5

basal firing rate, 9, 53, 65, 76, 83
binary cross-entropy, 45

Calinski-Harabasz index, 26
chronological attribute, 51
classifier, 113
connected component, 29
connectome, 3, 13
contextual inquiry, 55
continuous diverging colorscale, 75
correlation matrix, 115
cortical column, 2
cost function, 42
CytOx band type, 91

data abstraction, 37, 48, 144
data encoding, 34, 48, 49, 144
Davies-Bouldin score, 26
dendrite, 5
design study, 1, 2, 18, 30, 32, 35, 134, 141, 159
diverging colorscale, 65, 74
domain expert, 38, 40, 47, 48, 51, 59, 65, 66, 69, 72, 75, 81, 131, 150, 157
domain problem, 32, 37, 39, 49
domain problem characterization, 34, 47–49, 141
domain-specific, 49
dual visual pathway, 10
dynamic systems, 3

ecephy, 42
ecological validity, 37
excursion, 19

F1 score, 45, 120
feature, 42
field study, 32, 38, 79, 144, 146
field test, 133
firing rate, 6–8, 12, 53, 64, 65, 83, 134
  average, 159
  basal, 28, 30
  estimated mean, 28
  mean, 30
  trial-averaged, 12
firing-rate
  mean, 29
Fisher information, 159
focal participant, 55, 58, 61, 62, 77, 80, 94
fovea, 9, 19–21
functional attribute, 4, 12, 16, 18, 29, 42, 47, 50, 51, 53, 54, 56, 58, 60–62, 64, 68, 72, 73, 79, 81,

83, 112, 114, 115, 134, 145, 149, 150, 157, 177, 181

Gabor function, 9, 12
gratings, 20, 21, 24, 31, 63, 68, 69, 72, 77, 78, 91, 99, 120, 122, 134, 135, 143, 157, 177
  sinusoidal, 29
    drifting, 20, 31
gratings polar sector plot, 131
ground truth, 26

hyperparameter, 44, 46, 114, 115
hyperparameters, 120

individual, 63
individual observation, 15, 62, 63, 73, 74, 120, 134
informal design study, 47, 61, 77
informal study, 91, 134
informal user study, 38, 49, 74, 141
integrate-and-fire, 5
interaction design, 34
interaction encoding, 48, 49, 144

K-means, 26
Kolmogorov-Smirnov
  distribution
    inverse, 98

lab study, 32
laboratory study, 38, 79
learning curves, 184
learning rate, 44
linear regression, 43, 44
longitudinal study, 37
loss function, 42

mean firing rate, 53, 76, 83
moving bars, 19–21, 24, 31, 62, 68, 69, 72, 78, 91, 99, 120, 122, 133, 135, 145, 148, 157, 177
  stimulus, 29
multi-unit, 13, 23, 49, 51, 81
multi-units, 51

nervous system, 3
network neuroscience, 3
network science, 3, 13
neural coding, 155, 159
neural correlation, 155
neuron, 3, 4, 12, 49
  post-synaptic, 5
  pre-synaptic, 5
neuronal hypothesis, 62, 157
neuronal observation, 29, 42, 47, 50, 53, 72, 78, 113, 123, 144
  winnowing, 17
neuronal population, 2, 3, 47, 55

observation, 63, 102–104
  individual, 23
  summative, 23
observation winnowing, 144
Occam's Razor principle, 114
optogenetics, 13
overfitting, 44, 113, 114

parameter space, 43
parametric attribute, 47
parametrical attribute, 51, 60, 61, 63, 78, 81, 94, 120, 146, 157, 177
perceptron, 5
physiological attribute, 14, 51, 55, 56, 91, 147, 157, 177
polargram, 62, 77, 83, 87, 134, 144, 147
polargrams, 74
population characterization, 157
post-synaptic, 2
pre-synaptic, 2
precision, 45

precision score, 44, 120, 123
principal component, 115
problem domain, 18, 32, 40, 41
problem domain characterization, 39

qualitative colorscale, 67

recall, 45
recall score, 44, 120, 123
recording site, 14–16, 50, 55, 56, 58, 68–70, 81, 86, 157
regression function, 42
regressor, 113
regularization coefficient, 44
requirements gathering, 40
response attribute, 51, 62, 63
response cutoff, 28
response map, 28, 29, 53, 54, 62, 64, 69, 71, 74, 76, 77, 80, 87, 102, 131, 133, 135
retina, 8, 11, 12
reverse correlation, 12, 64

sequential colorscale, 65
silhouette coefficient, 26
single-unit, 13, 23, 49, 51
sinogram, 28
sinusoidal drifting gratings, 28
spike, 5, 26
spike count, 29, 65, 72, 76
  mean, 29
spike detection, 23, 51, 56, 58, 72, 73, 84, 177
spike detection threshold, 95
spike sorting, 15, 23, 26, 31, 51, 56, 58, 62, 63, 72, 84, 148, 155, 177
spike train, 5–7, 12, 13, 23, 24, 27, 49, 50, 53, 55, 56, 63, 95, 134
  trial-averaged, 12
spikes, 12
spontaneous firing, 9, 30
stimuli, 49, 53
stimulus, 2, 12, 19, 20, 50, 53, 157
stimulus attribute, 51, 81
stimulus intensity function, 19, 20
stimulus type, 53
striate cortex, 9
study participant, 32, 51, 69, 123, 150, 154
study participants, 113, 122, 144
summative, 49
summative observation, 15, 66, 68, 74, 81, 98, 120, 134
summative observation winnowing, 79
summative observations, 134
support vector, 113
synapse, 2–4

target variable, 42
task abstraction, 34, 37, 48, 49, 144
task analysis, 40
test dataset, 123
test set, 43–45, 150
time-to-action, 102, 103
train-test split, 43
training set, 43, 44, 184
transfer function, 24
trial, 6, 12, 14, 50, 72
  -averaged, 29
tuning curve, 29, 76

upsampling, 63
user, 50
user goal, 17, 41, 47, 49, 50, 55, 62, 73, 79, 143
  neuronal population characterization, 157
user study, 2, 26, 29, 32, 47, 55, 62, 65, 89, 94, 144, 146, 147, 150

V2 Dataset, 16, 17, 47, 49, 61, 62, 72, 77, 80, 89, 155, 159, 160, 181
V206 Dataset, 15, 18
validation set, 44
ventral stream, 11
visual cortex, 1, 7–9, 12, 13, 27, 49, 50, 159
visual encoding, 81
visual field, 9
visual stimulus, 49

waveform, 13, 23, 27, 50, 63, 73, 87, 143, 147
waveforms, 58, 69, 77
web application, 79, 81
wideband, 51, 55
winnowing goal, 62
Wizard of Oz, 60

Z-normalized, 64