Campus Paris Saclay

FONDATION DE COOPERATION SCIENTIFIQUE

IDEX Paris-Saclay
Initiative Doctorale Interdisciplinaire 2014
Formulaire de dépôt

Candidat / NOM : Perez-Guevara **Prénom :** Martin Felipe
Date de naissance : 15/04/1988
Adresse mail : mperezguevara@gmail.com
Téléphone : 0611690472

Titre du projet de thèse (en Français) : Modèles Distribués du Langage et des Emotions Fondés sur des Etudes du Cerveau et des Agents Virtuels

Titre du projet de thèse (en Anglais) : Distributed Computational Models of Language and Emotion based on Brain Studies and Virtual Agents

Axe stratégique de rattachement du sujet dans l'IDI 2014 : *Les Sciences de la Vie à leurs interfaces*
Type de financement demandé à l'IDEX Paris-Saclay : *financement à 100%*

Directeur(s) de thèse :
(Remplir le tableau ci-dessous. Le premier nom sera considéré comme le Directeur de recherche à titre principal. Il doit être habilité à diriger une thèse de doctorat. Les noms suivants seront considérés comme co-directeurs.)

NOM / Prénom	Etablissement/ Grade	Unité de recherche de rattachement	% encadrement
Martin, Jean-Claude	Université Paris-Sud / PR	LIMSI-CNRS	45
Pallier, Christophe	DR CNRS	Directeur de l'équipe INSERM "Neuroimagerie du langage" UNICOG/NeuroSpin/INSERM/CEA	45
Le Scanff, Christine	Université Paris-Sud / PR	CIAMS	10

Unité(s) de Recherche où sera effectué le travail de thèse
(Compléter le tableau ci-dessous. L'unité de rattachement principal doit obligatoirement appartenir au périmètre scientifique de l'IDEX Paris-Saclay)

Unité de Recherche de rattachement principal	% de temps estimé
LIMSI-CNRS	45
UNICOG/NeuroSpin/INSERM/CEA	45
CIAMS, Université Paris Sud	10

Ecole Doctorale de rattachement du doctorant :
EDIPS

Etablissement d'enseignement supérieur dans lequel l'étudiant sera inscrit :
Université Paris Sud

Résumé du projet (en Français) :
(15 lignes maximum)
L'objectif de cette thèse est la définition d'un modèle informatique commun de deux fonctions cognitives peu étudiées conjointement et pourtant complémentaires : 1) communication langagière (perception et apprentissage d'un langage artificiel), et 2) perception de stimuli sociaux (expressions faciales d'agents virtuels). Le modèle s'inspira des travaux en modélisation distribuée de Paul Smolensky. Plusieurs mesures de complexité seront définies pour étudier ces fonctions cognitives et leurs interactions. Ces mesures seront ensuite utilisées pour montrer la validité des modèles que nous établirons d'après des études utilisant des techniques d'imagerie cérébrale et des études comportementales. Ce projet vise ainsi à faire progresser l'état de nos connaissances en neurosciences sur les fonctions cognitives et leurs interactions. La thèse permettra aussi des avancées sur les modèles informatiques de la cognition et des émotions. Enfin, le projet ouvre des perspectives sur la prise en compte de contextes réalistes en communication humaine et en communication homme-machine.

Project summary (in English) :
(15 lignes maximum)
The goal of this PhD thesis is to define a joint computational model of two complementary cognitive functions which are seldom considered together: 1) lingustic communication (perception and learning of artificial languages), and 2) perception of social stimuli (facial expressions of virtual agents). A model will be defined based on Paul Smolensky's framework. Several measures of computational complexity will be developed to study these specific cognitive functions and their interactions. These will be employed as a way to build evidence for the accuracy of the models by employing neuroimaging techniques during brain studies and behavioral studies. This project will not only advance the current knowledge in the considered cognitive functions and their interactions but will create a precedent for more ambitious integrative models of cognition and emotion. It will support neuroscience and social computing efforts in realizing studies in more realistic contexts in human-human and human-machine interactions.

Description du projet de thèse (en Français ou en Anglais) :
(Le candidat ou la candidate décrira son projet (5 pages maximum) en mentionnant en particulier : le contexte scientifique, social et sociétal du projet, l'impact attendu, le programme de recherche envisagé, les dispositifs qui seront utilisés, les modalités de financement prévues, les perspectives attendues en termes de développements futurs en recherche et applications.
Dans le même fichier, le candidat ou la candidate pourra aussi expliquer (2 pages maximum) sa motivation pour le sujet proposé, ainsi que les compétences acquises et à acquérir dans le cadre de la thèse, et les perspectives de carrière souhaitées.)

Distributed Computational Models of Language and Emotion based on Brain Studies and Virtual Agents

PhD Candidate: Martin Felipe Perez-Guevara

Advisors:

- Jean-Claude MARTIN (LIMSI-CNRS, PR Informatique Univ. Paris Sud):
  Social computing and virtual agents
  http://perso.limsi.fr/wiki/doku.php/martin/accueil
- Christophe PALLIER (NeuroSpin, UNICOG laboratory, INSERM / CEA):
  DR CNRS, Directeur de l'équipe INSERM "Neuroimagerie du langage"
  http://www.unicog.org/pm/pmwiki.php
- Christine LE SCANFF (CIAMS, Université Paris Sud, PR):
  Psychology of interindividual differences
  http://www.staps.u-psud.fr/fr/recherche/ciams/rime/christine_le_scanff.html

Scientific background

Computational models of cognition

As explained by Paul Smolensky (2006), the main object of study of cognitive science is the mind/brain. The current fundamental hypothesis in cognitive science is that cognition is computation. But the problem is to understand what kind of computation and how to link this with an appropriate cognitive architecture. Multiple studies in diverse cognitive fields but specially in language have gained a lot of explanatory power under the consideration of a symbolic cognitive architecture. Nonetheless such architecture does not cope well with the parallel distributed processing that seems to take place in the brain. It thus becomes crucial to link our understanding of symbolic representations with the distributed representations and processing that takes place in the brain. This link is formally given by Paul Smolensky in what he calls the "Integrated Connectionist/Symbolic Cognitive Architecture" (ICS).

An important aspect of ICS as a modeling framework is its representational power since it can account for any context free grammar (Smolensky, 2012). Moreover Paul Smolensky (2006) shows how other models of neural network representations like synchronous-firing, holographic reduced representations and recursive auto associative memories are just specific cases (typologies) of the more general activation vectors representations of this framework employing tensor calculus. This is also important because it allows for a more systematic, formal and objective integration of models of diverse cognitive functions. This is in contrast with simulations like Spaun of Eliasmith (2012) that reflect the state of the art of our understanding of neuronal dynamics and cognitive function, but still contain many subjective decisions about the modeling architecture such that general formal principles can not be easily extracted.

Cognitive functions to be integrated

Regarding the domains of language and mathematics, how to characterize symbolic representations and their visual processing have been extensively studied. Moreover a big part of Paul Smolensky's research is centered in representing and characterizing grammars, as defined by Noam Chomsky. This gives a starting point for more ambitious objectives on the study of how the structure of symbolic expressions is processed in the brain. The network involved in parsing sentences has been identified by Pallier, Devauchelle & Dehaene (2011) who observed increased activations in areas as a function of the size of syntactic constituents. In Mathematics,

Friedrich & Frederici (2009) have presented hierarchical or flat logical formulas to subjects and, from the contrast between the two, concluded that maths and language rely on different networks. Maruyama et al. (2012) have come to a similar conclusion by presenting subjects with more or less destructured arithmetic formulas. Moreover, the design of artificial grammars is a well studied tool employed to model and assess the capacity and limitations of humans and even animals to process different levels of syntactic processing in language, keeping this cognitive function as independent as possible from semantic processing. It has been shown by Friederici et al. (2006) that humans have the capacity to learn regular grammars and also by Rohrmeier (2012) that humans can learn context free grammars. Interestingly It has also been shown that different levels of grammar recruit different areas in the brain (Friederici et al. 2006). Nonetheless measures of complexity and models that would correlate empirical evidence from FMRI and MEG with computational resources consumption predictions are still not well established and differences in the processing of syntactic structure in language and mathematics is not completely understood even though both depend on written symbolic representations. Also reasons for why the brain is specialized for different levels of grammars or for different types of symbolic representations like natural languages and mathematics are not well depicted.

Although emotion is recognised as one of the main cognitive functions, the links between emotion, cognition and behaviors is quite complex. Social neuroscience is a rapidly growing field of research (Decety and Cacioppo 2011). Virtual agents are increasingly used as stimuli in these neuroscience studies (Costa et al. 2013). Since the parameters for their expressions and animation can be easily modified in real-time, they enable controlled studies about human perception and communication that would be difficult to control with real humans. They are also used for studying how much emotional expressions of faces might affect performance in perception tasks such as in the emotional version of the Stroop task (Compton et al. 2003). They can even be used to elicit stress with virtual juges in experimental protols such as the Trier Social Stress Test which was used in hundred of experimental studies and makes uses of a mathematical task (Kirschbaum et al. 93). Finally, the face itself can be seen as a structured stimulus and researchers in human-computer interaction as well as psychologists have been investigating how the perception of various areas of the face combine (Martin et al. 2006, Ekman & Friesen 1975) or how the brain processes congruent vs. incongruent spatial combinations of facial and bodily expressions (Meeren et al. 2013).

Social and societal context

Understanding cognition is one of the greatest challenges faced by humankind, showing fast and immense progress in the last century. This problem is at the heart of grasping and perhaps improving human capacities and interactions with other humans and the environment. Moreover, language is one of the most impressive cognitive functions. The capacity of humans to create standard common frameworks for the encoding, decoding and reproduction of information through symbolic representations has given birth to modern society and technological development thanks to the accumulation and transmission of knowledge. For this reason, studying language and its associated symbolic representations and structures is an important human endeavor. Nonetheless language is embedded in social interaction and both should be considered together.

Recent progress in social neuroscience (Decety and Cacioppo 2011) and social/affective computing (Scherer 2010) have encouraged pluridisciplinary approaches for the studies of emotions and the perception of their multimodal expressions. Such studies can not only improve our understanding of the social nature of our brain but also help us design efficient and intuitive human-computer interfaces featuring virtual agents. These virtual agents display intuitive human communication modalities (language, facial expressions) and are useful for the understanding and remediation of social cognitive disorders and encourage the design of accessible sociable human-computer interfaces (Brunet et al. 2014). Whereas in the past researchers have prefered to focus on a single cognitive function to study at a time, it is now acknowledged that cognition need to be studies in a more naturalistic context and that several cognitive functions need to be considered altogether. Moreover considering Interindividual differences and their psychological assessment are known to be relevant to explain differences in terms of language processing, emotion processing and even brain activation (Suttin et al. 2011), and thus should be considered to embrace a larger population of users and participants.

Expected results and Impact

- Help social neuroscience establish more realistic experiments and questions for which isolated studies of individual cognitive functions is not a feasible methodological path.
- Contribute to the employment and integration of modern neuroimaging tools to build evidence on the accuracy of computational neuronal models.
- Contribute to the abstractions necessary to experimentally compare different symbolic representational systems like natural language, mathematics and emotional expressions.
- Contribute to the understanding of employing virtual agents for training to assist pathologies (Brunet et al. 2014) and for applications like second language learning.

Research program

- Year 1: general modeling framework ; definition of the experimental protocols
- Year 2: experimenting with language ; experimenting with emotions and virtual agents
- Year 3: joint experimental study of language and emotions ; a posteriori model of joint cognitive functions ; writing of the PhD

Year 1 will be devoted to the understanding and extension of Paul Smolensky ICT framework to mathematically and computationally model cognitive function simultaneously from a connectionist and symbolic computation point of view. There are three important aspects to be determined in this regard:

- how different real biological neural architectures and models can be linked with the formalism of Smolensky, since the developed proposal have to be able to adapt to specifications of different neural dynamics and architectures, specially if diverse cognitive functions and cytoarchitectures have to be considered. If we want to approach predictions on the consumption of brain computational resources, we can not ignore previous biological evidence linked with the operation behind certain areas of the brain and their interactions. Paul Smolensky jumps ahead in this regard showing how alternative representations of neural models like synchronous firing, temporal models and holographic representations can be equivalently mapped to his framework.
- how to implement model integration with the framework. There are diverse mechanisms through which different cognitive functions might interact. For example they may interfere or complement each other, they may employ the same neuronal resources or different areas of the brain and they may show additivity or superadditivity patterns in brain activations. As exposed in the brain turing machine proposal, the composition of neural mechanisms is an important research question that is not broadly explored in the literature, even though simple computational steps of cognitive functions have been well characterized in many cases. Important considerations like the possibility of an actual sequential computation that would break distributed computation possibilities when integrating neural circuits have to be considered.
- how to establish a connection, possibly by employing information theory, between computational resources and complexity of cognitive function to derive predictions of neural activity and how to methodologically link evidence from experimental techniques to the mentioned predictions.

Year 2 will consist on applying the concepts developed during the first year to understand, separately, the modeling of the structure and learning of symbolic representations employed in language and the modeling of the perception of emotion and facial expressions displayed by virtual agents. The representation of these models will account for their future integration and for the specificities and constraints imposed by the Smolensky ICT framework.

- In the case of language there are two main aspects to explore. In the first place, the symbolic structure normally employed in language (considerations around natural language and mathematics), since this is necessary for the representational problem of the model. Then, in the second place, we will study grammars and more general structures in language. For this last part the employment of artificial grammars is particularly promising as an experimental tool, since they can be easily built and analyzed under theoretical considerations of the modeling framework. We consider as a possible experimental setup the verification of predictions of the model in the learning difficulty of diverse syntactic rules and structures based on complexity measures that contrast the learning mechanisms proposed by the

soft constraints of optimality theory implemented in Harmonic networks and the absolute constraints of the more traditional absolute constraints imposed by Chomsky's grammatical theory.

- In the case of facial expressions/emotions, we will study how facial expressions of emotions impact he perception and the learning of language. We will inspire from existing experimental protocols such as emotional stroop task and the Trier Social Stress Test. Static facial expressions (eg. facial expression above displayed artificial language sentences) and dynamic expressions (eg talking agent uttering the artificial language sentence) will be used as stimuli.

We will select among relevant experimental tools like FMRI (BOLD and ASL), MEG, Eye-Tracking and virtual environments (Oculus Rift) to bring insight into the working of cognitive functions. The tools and techniques to consider will be those compatible with the infrastructure provided by LIMSI and NEUROSPIN as co-advisors of this thesis.

Year 3 consists in realizing the integration of the resulting models of the second year to explain and make predictions on brain activations and behavioral responses under the simultaneous realization of both cognitive tasks. The experimental tools that revealed to be appropriate in the second year will be kept and applied to the study of the influence of emotional facial expressions on the learning of artificial grammars will be studied with new experimental setups.

Equipment and software / dispositifs qui seront utilisés

- FMRI (better spatial volumetric resolution, more indirect measures of brain activity): BOLD (relative brain activation measures for better spatio-temporal resolution in intra subject analysis) ; ASL (absolute brain activation measures for better inter-subject analysis and tracking time evolution of tasks that involve learning)
- MEG (better time resolution, more direct measures of brain activity at the expense of spatial resolution)
- Eye tracking (Can be integrated with FMRI and MEG, or be employed alone for behavioral studies)
- Oculus Rift (Can be integrated with eye-tracking and provides control of stereoscopic virtual environments)

Funding / modalités de financement prévues

The salary of the PhD student is expected to be funded a 100% by the IDEX. This thesis will benefit from projects that are currently going on at the different partner sites. LIMSI-CNRS will provide its virtual human platform (MARC) and an oculus rift device. Several projects at LIMSI investigate the role of social companions for other application but without a focus on brain studies. There will be an opportunity to collaborate with these projects and fund some travel to conferences. UNICOG will provide access to its neuroimaging and eye-tracking platform.

Future directions

- Extension of the framework to the integration of more spontaneous / complex stimuli and other cognitive functions
- Study differences in symbolic representational systems like language in contrast to mathematics, focusing mainly on understanding differences in their learning.
- Mathematics learning with virtual agents; second language learning with virtual agents.
- Enhancement of human-machine interfaces with virtual agents and special symbolic representations.
- Advances in human-computer interfaces: Brain Computing Interfaces

References

Baker, M. (2001). The Atoms of Language. Basic Books.

Brunet et al. 2014. Advances in Virtual Agents and Affective Computing for the Understanding and Remediation of Social Cognitive Disorders. Research topic / special issue of Frontiers in Human Neuroscience.

Chomsky, N. (1957). Syntactic Structures. La Haye: Mouton.

Compton, Rebecca J.; Banich, Marie T.; Mohanty, Aprajita; Milham, Michael P.; Herrington, John; Miller, Gregory A.; Scalf, Paige E.; Webb, Andrew; Heller, Wendy (2003). "Paying attention to emotion: an fMRI investigation of cognitive and emotional Stroop tasks". Cognitive, Affective, & Behavioral Neuroscience 3 (2): 81–96.

Costa, T., Cauda, F., Crini, M., Tatu, M-K., Celeghin, A., de Gelder, B., Tamietto, M. (2013). Temporal and spatial neural dynamics in the perception of basic emotions from complex scenes. Social Cognitive and Affective Neuroscience

Decety, J. and J. T. Cacioppo (2011)The Oxford Handbook of Social Neuroscience. Oxford University Press.

Del Zotto, M., Deiber, M. P., Legrand, L. B., De Gelder, B., & Pegna, A. J. (2013). Emotional expressions modulate low α and β oscillations in a cortically blind patient. International Journal of Psychophysiology.

Ekman, P. & Friesen, W. V. (1975). Unmasking the face. A guide to recognizing emotions from facial clues. Englewood Cliffs, New Jersey: Prentice-Hall.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. science, 338(6111), 1202-1205.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: functional localization and structural connectivity. Proceedings of the National Academy of Sciences of the United States of America, 103(7), 2458-2463.

Friedrich, R. & Friederici, A.D.. (2009). Mathematical Logic in the Human Brain: Syntax. PLoS One 4 (5): e5599.

Grynszpan, O., Nadel, J., Martin, J. C., Simonin, J., Bailleul, P., Wang, Y., Gepner, D., Le Barillier, F., Constant, J. (2012). Self-monitoring of gaze in high functioning autism. Journal of Autism and Developmental Disorders, 42 (8), 1642-1650.

Jansen, A.R., Marriott, K. & Yelland, G.W. (2000). Constituent Structure in Mathematical Expressions.â In Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society: August 13-15, 2000, 22:238. Psychology Press.

Kirschbaum, C; Pirke, K M; Hellhammer, D H (1993). "The 'Trier Social Stress Test'—a tool for investigating psychobiological stress responses in a laboratory setting". Neuropsychobiology 28 (1-2): 76–81. PMID 8255414.

Martin, J.-C;, R. Niewiadomski, L. Devillers, S. Buisine, C. Pelachaud, Multimodal complex emotions: gesture expressivity and blended facial expressions, International Journal of Humanoid Robotics, special issue "Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids", 3(3), 269-292, 2006.

Maruyama, M., Pallier, C., Jobert, A., Sigman, M. & Dehaene, S. (2012). The Cortical Representation of Simple Mathematical Expressions. Neuroimage 61 (4): 1444-60.

Meeren, H.K.M., de Gelder, B., Ahlfors, S.P., Hämäläinen, M.S., Hadjikhani, N. (2013). Different Cortical Dynamics in Face and Body Perception: An MEG study. PLoS ONE 8(9): e71408. doi:10.1371/journal.pone.0071408

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical Representation of the Constituent Structure of Sentences. Proceedings of the National Academy of Science USA 108 (6): 2522-27.

Plate, T.A. (1995). Holographic Reduced Representations. IEEE Transactions on Neural Networks 6 (3): 623 -641.

Prince, A., Smolensky, P. (2008). Optimality Theory: Constraint interaction in generative grammar. John Wiley & Sons.

Rohrmeier, M., Fu, Q., & Dienes, Z. (2012). Implicit learning of recursive context-free grammars. PloS one, 7(10).

Schere, K. R., Tanja Banziger, Etienne Roesch. A Blueprint for Affective Computing: A sourcebook and manual (Series in Affective Science) 2010; Oxford University Press, USA

Smolensky, P, & Legendre, G. (2006). The Harmonic Mind. Vol. 1. MIT Press.

Smolensky, P. (2012). Symbolic functions from neural computation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 370(1971), 3543-3569.

Smolensky, P., & Legendre, G. (2006). The harmonic mind. Cambridge, MA: MIT Press.

Suttin, A., McCrae, R., Costa, P. (2011) The Neuroscience of Personality Traits: Descriptions and Prescriptions. In Decety, J. and J. T. Cacioppo.The Oxford Handbook of Social Neuroscience. Oxford University Press. pp 243-251.

Vagharchakian, L. Dehaene-Lambertz, G., Pallier, C. and Dehaene, S. (2012). A Temporal Bottleneck in the Language Comprehension Network. The Journal of Neuroscience 32 (26): 9089-9102.

Van den Stock, J., Vandenbulcke, M., Sinke, C., Goebel, R., & de Gelder, B. (2013). How affective information from faces and scenes interacts in the brain. Social cognitive and affective neuroscience. doi: 10.1093/scan/nst138

Zylberberg, A., Dehaene, S., Roelfsema, P. R., & Sigman, M. (2011). The human Turing machine: a neural framework for mental programs. Trends in cognitive sciences, 15(7), 293-300.

Martin Perez-Guevara
Project “Distributed Computational Models of Language and Emotion based on Brain Studies and Virtual Agents”

Motivation letter

My main research interest is to exploit the understanding of brain mechanisms, from a computational point of view, to develop human-machine interfaces that could assist the improvement of human cognition. This is a transversal and complex research issue, worth a life of research. I believe my studies in the applied mathematics master of complex systems science have provided me the basic tools and perspective to try to approach this interdisciplinary and ambitious topic of research.

I want to understand how symbolic and connectionist computation are linked and the stochastic dynamics that allow learning to take place in distributed representational systems. Moreover I want to discover the causes behind the bottlenecks that we can appreciate in symbolic information processing mechanisms like reading and to explore if they could be overcome by manipulating sensory input or other fundamental properties of neural networks. Finally I think that new hardware implementations aimed at augmented reality can drastically change the possibilities of sensory manipulation and facilitate sensory integration for learning.

To understand connectionist computation I think it is essential to study how neural networks can be framed in a computationally abstract theoretical framework and their connection with symbolic computation. For this, several proposals exist already, but I think Paul Smolensky's[1] is one of the most interesting to start exploring this topic. I got in touch with his framework, during my internship at NeuroSpin[2] on the syntax of symbolic representations, due to its interesting results on the analysis of grammars with optimality theory and their corresponding predictions on behavior. Nonetheless even Paul Smolensky's framework is just accounting for the representational and implementation problems of connectionist computation but is not addressing in enough detail, as I see it, the dynamical aspects that lead to the stability of computation in networks and the resilience, adaptability and multipurpose behavior that emerge from them.

I think that a connectionist representational framework can be extended with considerations that come from dynamical systems, stochastic processes and graph theory. In particular I believe the study of rare events with large deviation functions and of the universality classes that appear as a result of the study of stochastic processes could help explain many of the properties still

[1] Smolensky, P. (2012). Symbolic functions from neural computation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *370*(1971), 3543-3569.

[2] http://www-dsv.cea.fr/dsv/instituts/institut-d-imagerie-biomedicale-i2bm/services/neurospin-neurospin

not accounted for. Especially if appropriately integrated with the statistical properties of networks as they have been studied in graph theory. An important part of this belief comes from my experience working with Vivien Lecomte[3] on the application of methods like the cloning algorithm, that are normally employed to study non-stationarity and rare events on physical particle systems, to the study of the mechanisms that might lead to resilient memory storage on polychronous networks like those modeled by Izhikevich[4].

Moreover I think that topological properties of networks might be importantly linked to universal and convergent dynamics, even under big variations of the parameters of the modeled dynamics. Part of this thinking comes from the experience I had applying homology and general topological visualization tools to EEG datasets as part of my research training in Warwick University on BCI and neuro-imaging medical applications.

Nonetheless specific cognitive tasks, their models and approximations to biological measurements have to contribute to a general comprehension of connectionist computation and not just to the fine tuning of specific models. A state of the art network like Spaun[5] shows what is possible in terms of computational implementation and current intuition on the role of structure in neural networks but lacks a proper analytical presentation that could inform a more general model and allow the understanding of the system dynamics that lead to its successful behavioral responses.

Another important aspect of connectionist computation and neural models relies on understanding how some of the diverse indirect measures of neural activity could provide evidence for the accuracy of the representational models. Especially to try to infer general principles to be integrated into a common framework. My experience setting up and analyzing Bold-Fmri experiments, during my internship in NeuroSpin, gave me an important insight into the possibilities and limitations of these methods to infer brain dynamics. I recognize the challenge that is present in linking this and other methods with a theoretical neuronal modeling framework.

I am interested in understanding how sensory manipulation by itself can lead to changes in the dynamics of connectionist models. The multipurpose behavior and adaptation of neural networks could be exploited through inexpensive and safe virtual experimentation in contrast to the more hazardous and complex effects of pharmacological studies or invasive stimulation techniques. New devices that could be employed for virtual and

---

[3] Giardina, C., Kurchan, J., Lecomte, V., & Tailleur, J. (2011). Simulating rare events in dynamical processes. *Journal of statistical physics*, *145*(4), 787-811.

[4] Izhikevich, E. M. (2006). Polychronization: computation with spikes. *Neural computation*, *18*(2), 245-282.

[5] Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *science*, *338*(6111), 1202-1205.

augmented reality, like the Oculus Rift[6], offer a great opportunity to implement knowledge gained from theoretical and experimental neuroscience into education, communication and other important fields at the heart of human interaction.

Finally I think it is very important to understand the context in which information processing and more specifically language expression takes place. I believe fields like social neuroscience can greatly benefit from adequate modelling of sensory input and language from an information theory point of view. Since in this interdisciplinary areas many cognitive functions interact simultaneously and its difficult to simulate isolated events. Virtual agents and the study of emotional facial expressions as context for language is a great example of a multidisciplinary application that can lead to the exploration of important human activities like virtual second language learning.

I am convinced that conducting this project “Distributed Computational Models of Language and Emotion based on Brain Studies and Virtual Agents” for my doctoral studies will be a great opportunity to improve my current knowledge and will give me the foundational grounds upon which I would be able to start developing my more broad and ambitious research interests. In addition the team of associated supervisors will give me the opportunity to integrate and learn from the diverse areas of computer science, neuroscience and psychology that are necessary to successfully carry out my research ambitions.

Moreover I consider that the interdisciplinary skills that I have acquired in previous years of study along with a diverse background on social sciences, computer science and applied mathematics, will allow me to produce quality research and to successfully integrate the theoretical and experimental aspects to be explored. I believe this project will have a positive and direct impact in society thanks to a further development of scientific knowledge and to its straightforward and practical applications in fields like virtual second language learning.

[6] http://www.oculusvr.com/

MARTIN PEREZ-GUEVARA

Email: mperezguevara@gmail.com

Chambre 220, C.I.U.P Maison des Étudiants
de la'Asie du Sud-Est. 59B, Boulevard
Jourdan. 75014 Paris.

Telf. (33) 61 1690472

Personal Information:

Venezuelan, single, 26 years, born on April 15th 1988.

Education:

Palaiseau, France. September 2013

ECOLE POLYTECHNIQUE

Currently studying the second year of the Erasmus Mundus Masters in Complex Systems Science.

WARWICK UNIVERSITY ***Coventry, England. September 2012***

Finished the first year of the Erasmus Mundus Masters in Complex Systems Science with partial results with distinction

INTRODUCTION TO ARTIFICIAL INTELLIGENCE (https://www.ai-class.com/) ***Online Course. December 2011***
Online course developed in partnership with Stanford University
Given by Professors Dr. Sebastian Thrun and Dr. Peter Norvig

Successfully completed the course in the top 5% of the class with a score of 99.8%

UNIVERSIDAD METROPOLITANA ***Caracas, Venezuela. June 2011***

BA in Business Economics, Summa Cum Laude, top of class
Highest grade on dissertation based on artificial intelligence applied to capital markets.
Certificate of Competence in Business English

- Took additional courses on Mathematics and Programming
- Community service highest recognition for proposing and developing a finance introductory course for the ONG “Superatec”.

ADVANCED INSTITUTE OF FINANCE ***Caracas, Venezuela. Sept 2007***

Preparatory course for stock brokers and investment advisors

UNIVERSIDAD SIMÓN BOLIVAR ***Caracas, Venezuela. Sept 2005- June 2006***

Year of study in Electronics Engineering with outstanding grades 5.00/5.00

COLEGIO SANTIAGO DE LEÓN DE CARACAS ***Caracas, Venezuela. June 2005***

High school degree on “science”, top of class

- Received honors recognition (silver) for academic history.

Professional Experience:

Orsay, France. currently

Research internship
INSERM-CEA neuroimaging unit U992 / NeuroSpin, under supervision of Dr. Christophe Pallier

Neurospin is a research centre dedicated to neuroimaging. It was opened on January 1st 2007 and belongs to Life Science Direction at CEA (but has strong links with the Matter Science Division since its conception).

- Currently I do research on the processing of syntactic structures in language and mathematics in the brain.

Contact information: Dr. Christophe Pallier, Head of the language neuroimaging unit.

Coventry, England. 06/13-08/13

Research internship
Institute of Digital Healthcare / Neuro-engineering lab, under supervision of Professor Christopher James

The Insititute of Digital Healthcare is a five year partnership with NHS, WMG and Warwick Medical School which aims to improve people's health and wellbeing through the development, evaluation and use of innovative digital technologies and services.

- I did research on persistence homology applied to the analysis of time series from EEG activity of patients with epilepsy. The main focus of the internship was on the development of visualization techniques employing latest algorithms of topological data analysis and on developing algorithms to explore the possibilities of unsupervised learning on time series.

Contact information: Professor Christopher James, Director at the Institute of Digital Healthcare.
C.james@warwick.ac.uk. (44) 24 7615 1261

Researcher and Software Engineer and Developer. ***Caracas, Venezuela. 08/11-08/12***
Cinet Producciones S.A. - Research and Development Department.

This company is devoted to the development of audiovisual materials aimed at interactive and cinematographic applications. Also gives expert advice to companies on internal and external communications to boost processes efficiency based on optimal human interactions and communications.

- I did research on ways to implement game mechanics on different in-company processes and services with the objective of developing and implementing web tools intended to boost activities performance with efficient communications and engagement.

Contact information: Ing. Adalberto Gabaldon, Associate Director (adalberto.gabaldon@gmail.com, +58-426-5157010)

Software Engineer and Developer (on project) ***Caracas, Venezuela. 05/11-08/11***
Multus Investments/Consulting. - Research and Marketing Departments.

- I managed a project aimed to develop a web application designed for the analysis and presentation of global banking data. The application was implemented on Google's cloud infrastructure and most of the software was developed in the Java programming language employing Google Web Toolkit as the development framework.

Contact information: Ernesto Moreno, MA.
Multus Investments Position - Portfolio Manager (ernesto.moreno@multusinvestment.com , +1-305-9891361, +1-786-3501142).
Multus Consulting Position - Director (ernesto.moreno@multusconsulting.com, +58-212-2859875, +58-424-1318483)

Research Internship ***Caracas, Venezuela. 07/10-11/10***
Multus Investments/Consulting. - Research Department.

This Company takes care of clients on North and South America providing financial advice and portfolio management services. It has been a Registered Investment Advisor (RIA) in the state of Florida (U.S.A.) since November 2010.

- I assisted with the development of stock prices predictive models employing supervised neural networks and other artificial intelligence algorithms.
- I developed an application that connects to the EEUU SEC FTP servers to automatically download financial data packages from US public companies.

Contact information: Ernesto Moreno, MA.
Multus Investments Position - Portfolio Manager (ernesto.moreno@multusinvestment.com , +1-305-9891361, +1-786-3501142).
Multus Consulting Position - Director (ernesto.moreno@multusconsulting.com, +58-212-2859875, +58-424-1318483)

Skills:

- Languages: Spanish (mother tongue), English (Fluent – IELTS 8.0), French (Basic),
- Programming languages: Java (Advanced), Netlogo platform,Visual Basic for Application (intermediate), C (intermediate), Python (basic), Bash (basic), Javascript (basic), Haskell (basic), MPS (meta-programming framework)
- Wolfram Mathematica programming (intermediate) and Matlab programming (intermediate)
- Web applications development (Employing Google's cloud infrastructure and tools)
- Very creative when participating in projects and research. (Good at relating ideas and thinking outside the box)
- Fast learner with the ability to work under severe pressure.
- Very organized and disciplined.

Achievements:

- Received an Erasmus Mundus Scholarship in 2012 to study the Erasmus Mundus Masters in Complex Systems Science.
- Published BA's dissertation article on the arbitraged research magazine Anales of Universidad Metropolitana, Vol. 12. N° 1. 2012.
- Presented BA's dissertation article in Universidad Metropolitana's Intelectual Creation Congress (Congreso de Creación Intelectual de la Universidad Metropolitana)

Other interests:

Study of Computational Neuroscience, Cognitive Psychology, Philosophy, Economics, Sociology, Complexity and System's Science, Mathematics, Software development, Programming paradigms, Pedagogy and Gamification applications; Tango and Latin dancing; Bujinkan (martial art); Weight lifting; Running.

ÉCOLE POLYTECHNIQUE
ParisTech
Direction des Études
Le Directeur

RELEVE DE NOTES
TRANSCRIPT OF RECORDS

Nom de l'élève
Name of the Student : Perez-Guevara Truskowski (Martin)

Année scolaire :
Academic Year 2013 / 2014

Master de l'Ecole Polytechnique
Master of the Ecole Polytechnique

ECTS	Intitulé du cours - *Course Title*	Note - *Mark*
6	HSS650A-Theoretical analysis of complex systems	A
5	HSS650B-Opens problems seminars, Journal club	A
6	HSS650E-Multiscale approach to learning	A
6	HSS650F-Complex systems made simple	A
4	HSS656-Therapeutic evaluation: Thinking in Systems	A
4	HSS650K-Networks:Empirical results, Modeling & Dyn. Processes	A

Palaiseau, le 31 Janvier 2014
Directeur des Etudes - *Dean of Studies*
Ecole Polytechnique
Joaquim Nassar

ÉCOLE POLYTECHNIQUE . F - 91128 PALAISEAU CEDEX
Tél. : (33) 01 69 33 33 00 - Fax. : (33) 01 69 33 38 51
Mél. : direction.etudes@polytechnique.fr

centre for complexity science

THE UNIVERSITY OF WARWICK

Thursday 28th November 2013

To Whom It May Concern,

This letter confirms that Mr Martin Felipe Perez-Guevara is currently registered as a full time MSc Student of the two year Erasmus Mundus Masters in Complex Systems Science at the University of Warwick. Mr Perez-Guevara commenced his study on the Erasmus Mundus Masters with us at the University of Warwick on the 24th September 2012. Mr Perez-Guevara is expected to complete his study on the Erasmus Mundus Masters in Complex Systems Science on the 30th September 2014.

The Erasmus Mundus Masters in Complex Systems Science is a two year (24 month) course. The the study is not separated into individual semesters, and students are expected to complete the two year period of study in one continuous period. As a result of this no formal transcript can be issued for the study undertaken on this course until the end of the full two year period of study. However, please find below details of the marks that Mr Perez-Guevara has obtained on the modules he has studied to date, the results of which have been confirmed by the Department.

CO901- **Self Organisation and Emergence-** 74
CO903- **Complexity and Chaos in Dynamical Systems** -73
CO907- **Quantifying Uncertainty and Correlation in Complex Systems -**75
CO905- **Stochastic Processes** -75
CO902- **Probabilistic and Statistical Inference** -73
CO904- **Statistical Mechanics and its applications to Complex Systems** -71
MA4J5- **Structures of Complex Systems** -82
Miniproject- **"Exploring BCI Paradigms for Augmented Reality Applications"** - 74

Please don't hesitate to contact me if you have any questions.

With best regards,

Miss Jen Bowskill
Administrator

UNIVERSITY OF WARWICK
COMPLEXITY SCIENCE
DOCTORAL TRAINING CENTRE

Jen Bowskill
Administrator for Complexity Science DTC
Complexity Science Centre
Zeeman Building
The University of Warwick
Coventry CV4 7AL • United Kingdom
Tel: +44 (0)2476 5523673
: +44 (0)2476 1 50866
Email: complexity@warwick.ac.uk

CENTRE NATIONAL
DE LA
RECHERCHE SCIENTIFIQUE

LABORATOIRE D'**I**NFORMATIQUE
POUR LA **M**ECANIQUE
ET LES **S**CIENCES DE L'**I**NGENIEUR

La Directrice-Adjointe
Tel : (33) 1 69 85 80 88
email : anne.vilnat@limsi.fr

Lettre de soutien au projet de thèse interdisciplinaire :

Distributed Computational Models of Language and Emotion based on Brain Studies and Virtual Agents

Le but de ce projet est de faire intervenir des agents virtuels (développés au LIMSI) dans les études sur le cerveau menées à Neurospin, en étudiant les différences entre les individus dont le CIAMS est spécialiste. Ce projet vise donc à développer une approche pluridisciplinaire associant l'utilisation des technologies virtuelles pour l'étude de la cognition et de la perception (LIMSI-CNRS/groupe CPU), des modèles et observations du cerveau à travers différentes techniques (UNICOG / INSERM / CEA / NEUROSPIN), et les principales théories de la personnalité et des différences interindividuelles (CIAMS / Université Paris Sud). Il se situe ainsi pleinement dans l'axe « Sciences de la Vie à leurs interfaces » de l'Idex.

Il s'agit d'un projet clairement pluridisciplinaire, alliant l'informatique, la psychologie et les neuro-sciences. Le laboratoire soutient depuis longtemps ces approches pluridisciplinaires, tant avec le CIAMS au sein de l'université Paris-Sud, qu'avec les équipes du CEA, dans le cadre de la future Université Paris-Saclay. Nous donnons donc un avis très favorable à cette demande.

Fait à Orsay le 21 mai 2014

Jean-Claude MARTIN
Professeur en Informatique
Université Paris-Sud
Responsable du groupe CPU du LIMSI
LIMSI-CNRS, BP 133
91403 Orsay Cedex, FRANCE
Tel. : 06 84 21 62 05 Fax : 01 69 85 80 88
e-mail : martin@limsi.fr
Web : http://www.limsi.fr/Individu/martin/

Orsay, le Mercredi 21 Mai 2014

Recommandation pour la candidature de Martin Felipe Perez-Guevara

Martin Felipe PEREZ-GUEVARA effectue actuellement son stage de M2R et Sciences de la Complexité à l'Ecole Polytechnique où il effectue des recherches sur la modélisation du cerveau à l'aide de techniques d'imagerie cérébrale.

Il est extrêmement motivé par les interactions entre informatique et neurosciences via la définition de modèles computationels. Ses compétences et son expertise sur ces deux domaines ainsi que sa motivation pour le sujet de la thèse qu'il a défini avec nous en font un candidat rare et idéal pour une collaboration pluridisciplinaire alliant informatique (LIMSI), neurosciences (NEUROSPIN) et psychologie de la personnalité (CIAMS).

Sa thèse nous permettra aussi de renforcer en interne les interactions entre enseignants-chercheurs en informatique et en psychologie au sein du groupe « Cognition Perception Usages » que je dirige.

Pour toutes ces raisons, je soutiens très chaleureusement la candidature de Martin Felipe PEREZ-GUEVARA à une allocation de thèse IDI.

Jean-Claude MARTIN
Professeur en Informatique à l'Université Paris-Sud
Responsable du groupe Cognition Perception Use au LIMSI-CNRS

Inserm
Institut national
de la santé et de la recherche médicale
Unité 992 de Neuroimagerie Cognitive

COMMISSARIAT
À L'ÉNERGIE ATOMIQUE

May 14th, 2014

Martin Felipe Perez-Guevara est un modélisateur qui propose de s'attaquer à la question de l'encodage des représentations des structures complexes dans le cerveau. Il effectue actuellement son stage de M2 de sciences de la complexité dans mon équipe, ce qui lui a permis notamment de se former à l' IRM fonctionnelle. Martin Felipe fait preuve de beaucoup d'énergie et d'enthousiasme, ainsi que d'une grande capacité d'apprentissage et de creativité.

L'un des objectifs du projet de thèse défendu par Martin Felipe est de développer des modèles reliant activité cérébrale et complexité des structures représentées, puis de tester empiriquement les prédictions de ces modèles, notamment en les comparant avec des mesures obtenues avec des techniques d'imagerie cérébrale. Le point de départ de la modélisation utilisera le cadre théorique de Paul Smolensky, qui est à ce jour la proposition la plus détaillée concernant l'encodage des structures. Cette question étant extrêment pertinente pour le langage, j'ai accepté de co-diriger ce travail de thèse. De plus, la réalisation de ce projet apporterait une dimension de modélisation qui n'existe pratiquement pas dans mons équipe, où les étudiants sont des biologistes, des psychologues ou des linguistes.

Christophe Pallier,
DR CNRS
Directeur de l'équipe INSERM "Neuroimagerie du langage"

Inserm U562, CEA/SACLAY/NeuroSpin, Bât. 145, Point courrier 156, 91191 GIF/Yvette Cedex
Téléphone : 33 (0)1 69 08 79 34 • Fax : 33 (0)1 69 08 79 73
Courrier électronique: Christophe@pallier.org: • Site web : http://www.unicog.org

Martin Felipe Perez-Guevara

Project “Distributed Computational Models of Language and Emotion based on Brain Studies and Virtual Agents”

Master internships report

During my master I had to participate in two research internships. In both of them I realized research in the field of neuroscience. In the first one, I spent three months at the Institute of Digital Healthcare in Warwick University in 2013. There I employed new mathematical methods to analyze EEG signals to contribute in the field of Brain-Computer interfaces and Healthcare (also worked on the analysis of EEG datasets of epilepsy). While in the second one I worked on experimental designs aimed at understanding structure in language employing neuroimaging techniques.
I am currently doing a 6 months internship at NeuroSpin (INSERM/CEA, Saclay) under the supervision of Dr. Christophe Pallier. This internship, started in January 7th 2014, is required for the last period of my master (Erasmus Mundus Master in Complex Systems Science), organized between Warwick University in England, Gothemburg and Chalmers in Sweden and Ecole Polytechnique in France. After finishing the internship the 7th of July I would effectively finish the requisites to receive a degree M2 from Warwick University and Ecole Polytechnique.

Second internship (Partial report / Experimental pilot description appended)

Main problem, background and state of the art

The main problem I explored in my internship is that of the identification of the brain mechanisms and areas related with syntactic processing on symbolic representations and more specifically on mathematical expressions.
On the language domain, evidence for the cognitive processes behind syntactic structures can be tracked down to priming studies demonstrating how people tend to reuse syntactic constructions when they are previously exposed to them (Bock et al., 2006; Branigan, 2006). Also Noppeney & Price (2004) reported that syntactic priming caused a repetition suppression effect in a region of the anterior left temporal lobe. However other studies revealed contradictory evidence like Devauchelle et al (2009) and Humphries et al. (2006).
Other parts of the brain have also been related to syntactic processing and manipulations, like the left inferior frontal gyrus, including Broca's area (Tettamanti et al., 2002; Musso et al., 2003; Ben-Shachar et al., 2004, 2003). Interestingly Tettamanti et al (2002) and Musso et al. (2003) taught grammatical and ungrammatical rules to participants during FMRI scanning. The grammatical stimuli depended on hierarchical structures while the ungrammatical one only had sequential properties. It turned out to be the case that subjects learned the rules behind both types of stimuli but the grammatical ones show more activation in the left inferior frontal gyrus. On the other hand, evidence for the participation of Broca's area in the processing of hierarchical linguistic structures come from studies employing artificial grammars (Friederici et al., 2006a; Bahlmann et al., 2008), in which the subjects were trained in two languages, one with embedded constituents and the other only with serial structure. It turned out that only participants that learned the embedded constituents structures showed increased activation in Broca's area.
Moreover the mentioned experiments show that the brain respond to syntactic manipulations but do not reveal how nested constituents could be encoded in the brain. Pallier et al. (2009) tested the hypothesis that tree structure is encoded by an increasingly complex distributed cell assembly pattern in concordance with the complexity of a processed tree structure. He provided evidence for this relationship by looking at the bold response of 12 word sequences, where the size of the

syntactic constituent was manipulated.
Nonetheless on the case of mathematical formulas, that clearly also have a structure of nested constituents, little research exists about their parsing. Jansen et al. (2003) showed that well-formed formulas are better memorized than ill-formed ones. Also they showed in later research that the latencies of fixation increase at the end of mathematical constituents, in a similar fashion to text (Jansen et al., 2007). This suggested that mathematical formulas rely on similar mechanisms as language even though it has other forms of presentation, sometimes two dimensional. Moreover, most work on mathematics employing neuro-imaging techniques have focused on numerosity, except for the work of Friedrich & Friederici (2009) who studied complexity in mathematical expressions in a similar way to studies that compare sentences and lists of words. They were surprised to find limited activation in frontal areas, since their earlier work on language revealed more posterior regions.
Then several important questions arise regarding the encoding, complexity and processing of mathematical syntactic structures in the brain. How much overlap should we expect between the brain regions responsible for language and mathematical processing, specially when we consider syntactic manipulations? Can we find areas that would respond exclusively to syntactic manipulations in mathematical formulas? How mathematical syntactic manipulations are processed in the brain, can we show areas that respond to increasing complexity? These are the state of the art questions that I approached during the internship.

Approaching the problem with diverse experimental designs

As mentioned in the previous section there are three main state of the art questions that I am approaching in the internship.

1. What are the brain areas that respond to mathematical formulas processing?
2. What are the brain areas that respond more specifically to syntactic manipulations in mathematical formulas? Are there areas that reflect automatic parsing?
3. Can we find areas that respond to complexity in number of constituents in mathematical formulas?

The first and third questions have been approached by recent experiments in the lab. Unpublished results on general research on the processing of mathematical concepts written on words and mathematical formulas seem to recruit additional areas that are not employed by natural language. It seems that there are important non overlapping regions.
Moreover Maruyama et al. In the paper “The cortical representation of simple mathematical expressions” published in 2012 reported results from experiments at the lab that supported the notion of non overlapping areas for mathematical processing and indicated areas that responded to complexity in the number of constituents in formulas. Nonetheless certain weaknesses were identified in the results due to the employment of a demanding memory task during the presentation of the designed formulas. Due to the memory task the complexity effect could be confounded with more difficulty in the memory task, moreover automatic parsing areas could not be identified since the parsing of formulas is forced by the task.
The second question on the other hand has not been directly tackled by current experiments, specially because of the complications behind exploring syntactic manipulations in mathematical formulas. This is due to the fact that the possible syntactic manipulations are very restricted if one wants to keep the grammaticality of the formulas and also the number of constituents. Moreover possible changes in syntactic structure are correlated with number of characters and visual effects due to the introduction of parentheses.
I worked on the development of two experiments to contribute to the depicted questions and specially to questions 2 and 3:

1. The first experiment is focused on question 3. It consisted on a remake of Maruyama's experiment with the same stimuli that consisted on formulas that had the same length with different number of constituents embedded in non grammatical character sequences. But in this case we completely got rid of the memory task, so the stimuli is simply presented for very quick intervals of time (200ms) that are even faster than possible eye saccades. The task would be a simple detection instruction to ensure attention but without forcing the parsing of the stimulus. We would expect to confirm decreasing complexity effects reported in Maruyama's paper, observe the network of brain activations related only with the appreciation of the mathematical formulas and moreover be sure that this areas would be related with automatic processes since parsing was not forced. Moreover we extended Maruyama's stimuli to also include formulas with symbols since the original stimuli only contained numbers, so that we could differentiate areas that responded to complexity for symbolic and numeric representations in formulas. More details on the experiment design can be read in Appendix A.

2. The second experiment is focused on question 2. I proposed to implement a syntactic supra-conscious priming design on mathematical formulas. The idea is that by having differences in the shape of formulas but not in the syntactic tree structure we could isolate the effect on brain activations of syntactic manipulations. Due to the repetition suppression effect that can be seen in the hemodynamic response function we expect to test the hypothesis that an area involved in automatic parsing processing would show more activation for pairs of formulas with different tree structure than with the same tree structure. Moreover we include non formula stimuli with the same characters and similar visual shape manipulations as the ones induced in the real formulas to control for areas that might be responding to mere visual priming effect. Again on this experiment, as in the first one, the task during each trial was a simple detection task that would need interfere with formula parsing and would not force parsing, but would just ensure attention to the stimuli. In addition we considered two levels of formula complexity, expecting that a simple contrast should confirm some areas as revealed in Maruyama's experiment. More details on the experiment design can be read in Appendix A.

Results and perspectives

Up to this date, only a pilot with two subjects have been run, due to the high costs associate with running FMRI experiments. Small modifications might be included in the definite experimental design and a complete subject population will be tested to form publishable quality results. Nonetheless the current pilot already reveals promising and interesting patterns that suggest we are close to answering the questions of interest.

1. Regarding the first experiment, Maruyama's revisited. We seem to confirm as expected some and possibly all the areas where a decreasing complexity effect was detected. As reported in Maruyama's paper. Moreover we seem to be able to see additional areas of activation for formulas containing numbers than symbols. We found a lateralization effect in the processing of formulas that emphasized a left branching in the syntactic structure and we seem to have an additional area reflecting increasing complexity in the orbital area that was not discovered with the original Maruyama paradigm. But more data needs to be collected to certainly support any findings.

2. Regarding the second experiment, the priming trees. Automatic parsing areas seem to be

confirmed in parietal and frontal regions thanks to the hypothesized syntax effect. There seem to be different effects related to different syntactic manipulations like changes in the symmetry of branching against changes in the top structure of the tree. Moreover a partial complexity effect might be appreciated in occipital areas. More complicated effects related with specificities on syntactic manipulations and other characteristics of the trees are suggested by the data but the significance is too low to report them with responsibility. But definitely more data needs to be collected to strongly support any claims on the suggested findings.

We expect that running the experiments on a complete population of subjects will reveal important significant effects that would account as strong evidence to back up an answer to the proposed questions. Considering further research, in the case of the complexity of trees we would like to test complexity in number of constituents without the repairing effects that might be taking place under the Maruyama's stimuli paradigm due to the ungrammatical portions in lower level formulas. Moreover we would like to find a way to test for the composition of Mathematical constituents and link this we a complexity effect. Going even further, we have the idea of employing Paul Smolensky's framework of connectionist computation to model mathematical symbolic representations, complexity and composition. So we can test modeled predictions of the inner workings of the neural circuits responsible for the parsing of mathematical formulas and calculation with new neuroimaging experiments and experimental paradigms still not explored in the field.

Summary of internship experience

During the current internship I have been trained in experimental design for language studies employing Bold-Fmri. Moreover I have been trained in the preprocessing, processing and analysis techniques necessary to interpret Fmri studies. Even though not at the same level of detail, I have also been exposed to learn about other neuroimaging techniques like MEG and ASL-Fmri, since my lab team works with both technologies.
Its important to note that in the lab the techniques employed are cutting edge and many of the newer tools publicly available in the field are supplied by common research partners inside the lab, like the pypreprocess and pyhrf tools. Moreover I have been exposed to new estimation techniques developed by partners in the lab like the Joint-Detection Estimation method, which was recently published and allows with bayesian analysis to not only recover the stimuli HRF related amplitude like in standard GLM analysis but also to estimate the HRF shape.
In the field of language I had the opportunity to work on problems in the state of the art on, by studying how syntax is processed in the brain for natural language and mathematical representations. My current work has allowed me to get in touch with surprising and interesting phenomena of communication, learning and information processing in the brain. In the last months I have gained plenty of experience and insight into the field of cognitive neuroscience and awareness of the different activities involved in this type of research.

First internship (Summary / complete official report appended)

During my first internship, with respect to the neuroscientific field, I was trained in Brain-computer interfaces and EEG technologies. Moreover I got in touch with healthcare applications of neuroscience, since my internship was conducted at the neuroengineering lab of the Institute of digital healthcare in Warwick University. In addition I had a strong training on applied mathematics and in particular in the area of topological data analysis, which is a promising new field for understanding big data sets and assisting unsupervised learning.
The project that I proposed for this short internship was to develop a framework that would assist

the development of BCIs based on new mathematical tools in the field of topological data analysis. The main idea is that this mathematical methods would help identify features of the EEG signal associated with certain events in an unsupervised way. To look for event related potential patterns that go beyond traditional established paradigms that can be quite slow and have low accuracy for the user. The idea is that the framework would help identify features of the space depicted by the EEG signals that, if could not improve the accuracy, at least would contribute to the speed of the interfaces thanks to the employment of more natural and fast occurring events.

The results of the partially implemented framework on the application of the new methods revealed to be promising, although more exploration was necessary from the mathematical and experimental point of view in diverse signal sets and mathematical features of the space depicted by the signals. Nonetheless the project established a strong precedent on the application of a mathematical tool that is still far from being a standard implementation in neurosciences and many other field, since its quite recent and complex.

The details of the work on this internship can be quite extensive but are well depicted in the research report presented for the master at the end of year M1 in Warwick University. The research report of the internship is appended to this file with the title “Exploratory framework on EEG signals for the development of BCIs”.

Appendix A. Experiments of pilot “Mathformula2” of second master internship

Mathformula2 consists of two main experiments, a remake with some modifications of maruyama's stimuli and a presentation of supra-conscious priming formulas.

Maruyama remake: 2 sessions, 11:23 min each with 432 scans

Stimuli

- The stimuli presented had the same style as in Maruyama. Partially scrambled formulas of 11 characters. Formula generation was inferred, so that the templates for formulas were extracted from Maruyama stimuli in database. Formulas always contain 4 variable symbols in random order, three binary operators + or -, with at least one of each, and four paranthesis. Balanced on the amount of times a type of character appears in each position for each type of formula was seeked as possible.

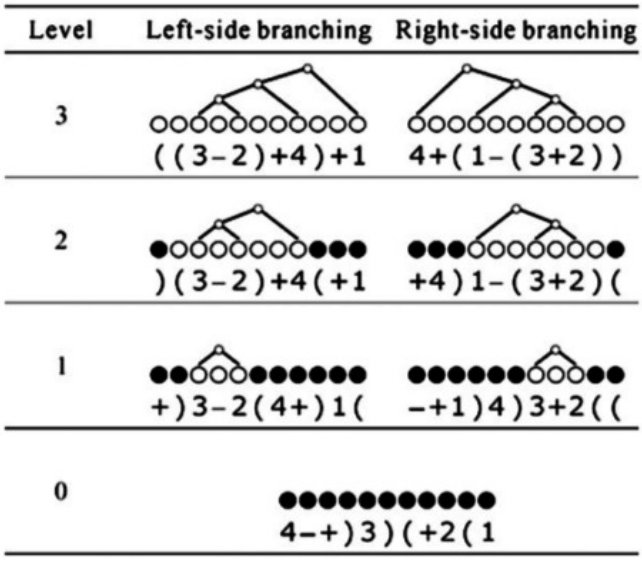
- The only important modification was that in one session the letters a,b,c and d were employed instead of the numbers 1,2,3 and 4.

- In total there were 160 stimuli. 40 for each level category and in the case of levels with branching, 20 were left-sided and 20 right-sided.

- From each level 4 stimuli were selected randomly to introduce an @ symbol in a uniformly randomly selected position of the formula. So that 10% the stimuli was employed for the task.

Next we present an image illustrating the stimuli and a table showing an example of the newly included symbol formulas in contrast with the number ones.

Stimuli Example:



Examples for left-side branching and amount of stimuli per category. (160 stimuli in total / 16 from which are altered for the task)

Category	Ex. numbers	Ex. symbols	Amount
Level 3	((2+3)-4)+1	((a+b)-d)+c	40
Level 2	)(4+3)-1(2-	)(a+d)-c(b-	40
Level 1	2)1-4((+)3-)	)d+b(-c-(a	40
Level 0	4)3-+(1()2-	+)-c)d+(a(b	40
Task	3+(-@2)4)(1	b(-+d)c@)-a	16

Task

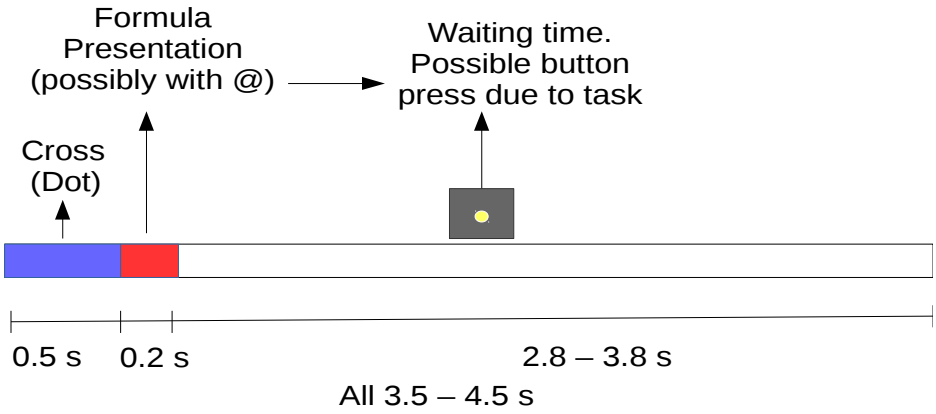
- The task consisted in pressing a button given to the right hand whenever a formula with an

@ was detected.

Presentation

- The total time of each trial is 4s with a continuous jitter of 500ms
- cross is presented to subject for 500ms
- formula stimuli is presented to subject for 200ms
- There is a maximum response time given of 2 seconds after presentation of stimuli
- In total 3.3 seconds are given after stimuli presentation with a continuous jitter of 500ms.

Diagram of stimuli presentation and task



Priming formulas (Priming trees): 2 sessions, 12:23 min each with 472 scans

Stimuli:

- We consider only valid formulas representing certain selected syntactic (tree) structures. In total there are six types of formula structures. Considering var for variable, bin for binary operator and una for unary operator, the employed formula structures are presented in the next table. Every formula contains 2 parentheses.

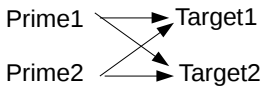
Formula type	Example	Structure	# operators	# variables	# characters
Una left (1)	log(y)+x	una(var)binvar	2	2	8
Una right (2)	x*sin(y)	varbinuna(var)	2	2	8
Una top (3)	tan(a-b)	una(varbinvar)	2	2	8
Only una (4)	tan(y)	una(var)	1	1	6
Only bin (5)	(y+x)	(varbinvar)	1	2	5
Large Scramble (6)	b)tan*a(	var)unabinvar(	2	2	8
Small Scramble una (7)	x)(sin	var)(una	1	1	6
Small Scramble bin (8)	y)x(+	var)var(bin	1	2	5

- The symbols considered to uniformly randomly fill the formulas were:
  - a, b, x and y for variables

    - +,-, and * for binary operators
    - log, sin, cos and tan for unary operators
- In a trial a pair of formulas is presented as prime and target, only formulas with the same number of operators are paired. So that valid pairs are combinations of trees 1,2 and 3, combinations of trees 4 and 5 and the pairs of the same type for 6,7 and 8. The possible pairs are shown in the next table.

Pair	Prime example	Target example
(1,1)	cos(a)+b	tan(y)*x
(1,2)	tan(x)+y	b*cos(a)
(1,3)	sin(a)+b	cos(x*y)
(2,1)	x+tan(y)	cos(b)*a
(2,2)	b*cos(a)	x+tan(y)
(2,3)	x-tan(y)	a*log(b)
(3,1)	sin(a+b)	cos(x)*y
(3,2)	tan(y+x)	a-cos(b)
(3,3)	log(a*b)	tan(x-y)
(4,4)	sin(a)	log(x)
(4,5)	log(y)	(b-a)
(5,4)	(b*a)	tan(y)
(5,5)	(a*b)	(x-y)
(6,6)	y)cos-x(	b)log+a(
(7,7)	y)(sin	a)(cos
(8,8)	y)x(-	a)b(*

- In every case the pair of formulas differed in the symbols included in the tree structure:
    - Variables in one tree were a and b, while in the other tree they were x and y
    - Unary operator in one tree had to be any other unary operator than the one in the other tree
    - Binary operator in one tree had to be any other binary operator than the one in the other tree
- A priming effect is looked for regarding syntactic structure so the following diagram is enforced across trials in the presentation of prime/target to discover areas for which there is a higher response when formulas that differ only in tree structure are presented together in contrast to formulas with the same tree structure. (Looking for syntax areas)



Pair	Prime	Target
(1,1)	log(a)-b	cos(x)*y
(2,1)	a-log(b)	cos(x)*y
(1,2)	log(a)-b	x*cos(y)
(2,2)	a-log(b)	x*cos(y)
(1,1)	cos(x)*y	log(a)-b
(2,1)	cos(x)*y	a-log(b)
(1,2)	x*cos(y)	log(a)-b
(2,2)	x*cos(y)	a-log(b)

- Besides the implementation of the correct priming scheme, in this case the relationship of structure between prime and target is symmetric (order do not matter). So to eliminate possible confounds of the order in the presentation of symbols the same scheme is repeated with the same stimuli but in an inverted order.

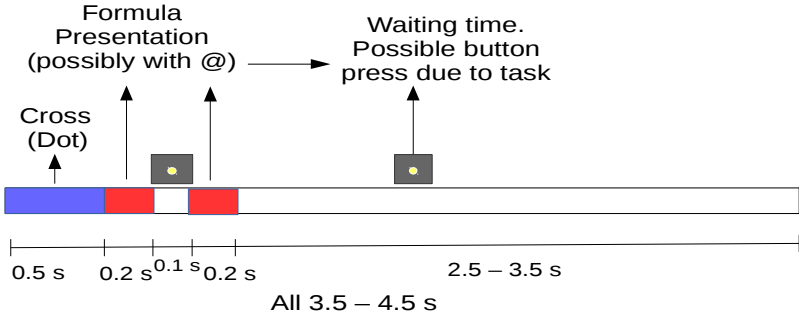
Task

- The task consisted in pressing a button given to the right hand whenever a formula with an @ was detected. Any of the two formulas presented could contain the @. (The idea of the task is avoid explicit parsing, so that automatic syntax areas could be revealed).

Presentation

- The total time of each trial is 4s with a continuous jitter of 500ms
- cross is presented to subject for 500ms
- formula prime is presented to subject for 200ms
- The screen is blank for 100ms
- formula target is presented to subject for 200ms
- There is a maximum additional response time given of 2 seconds after presentation of the second formula. If the subject presses the button at any time after the presentation of the first formula, it is also recorded.
- In total 3 seconds are given after stimuli presentation with a continuous jitter of 500ms.

Diagram of stimuli presentation and task



Exploratory framework on EEG signals for the development of BCIs

Martin Perez-Guevara
Supervisor: Christopher James

ABSTRACT: Finding appropiate spatio-temporal features of electroencephalography (EEG) signals to build a brain computer interface (BCI) is an extremely complex and challenging problem. In this study, motivated by new perspectives on the brain computer interface's research community and new methods developed in topological data analysis, we propose a framework to explore features in the EEG signal domain in an unsupervised way, such that a subject aided by machine learning and self labeling could establish the basis of a completely personalized and accurate BCI. In this study, we show the viability of an exploratory framework methodology based mainly on the detection instead of classification perspective and the Mapper method that belongs to the domain of persistent homology, and discuss further research and development necessary to implement the framework.

1. INTRODUCTION

Electroencephalography (EEG) is a recording of the electrical activity around the scalp. Specifically, as explained in Piotr Olejniczak review [1], it is a graphic time series representation of the difference in voltage between two different cerebral locations. The obtained signal is influenced by diverse factors like the electrical conductive properties of the tissues that lie in the middle of the electrical source and the electrode employed to measure the potentials, the conductive properties of the electrode itself and the orientation of the potential source in the cortex.

The EEG is made possible thanks to the current flow that passes through the tissues between the source of electrical activity inside the brain and the recording electrode. But then EEG only provides a two-dimensional picture (surface in the scalp) of a three dimensional process (volume of the brain), which degenerates into the inverse problem, since the electrical source can not be uniquely determined from EEG and so different tasks extracted from brain activity might not be easy to differentiate. Also the bone tissue and skin tissue with the influence of the environment add additional noise to the signal that further complicates analyzing EEG recordings.

Even though EEG presents the inverse problem and a low signal to noise ratio, it is still employed for its extremely high time resolution. Its non invasive nature and recent developments in commercial products, like the Emotiv headset , that lower the costs of the hardware. Turning EEG into an accessible solution for diverse applications like the NeuroPhone system proposed by researchers in Darthmouth College [2].

In this study, EEG will be employed as the main tool to explore the construction of a Brain-computer Interface (BCI), which consists on interpreting the signals obtained from the brain to control commands in a computer program, for example selecting characters to write a word.

Most BCI applications developed so far are aimed towards aiding disabled people. In a similar fashion to eye glaze devices, BCI applications allow writing, selecting items on a screen or controlling the direction of a wheelchair. Moreover the most successful algorithms relay on motor imagery and event based potentials, since particularly paraplegic people can use these intentions without the interference of muscle activity on the EEG signal.

The P300 Wave, as presented by Picton [3], is a great example of a context specific feature that can be extracted from the waveforms of EEG activity signals to detect the conscious intention of a subject to select an improbable target. Donchin et al. [4], show how the P300 wave effect can be employed to implement a BCI for spelling with high accuracies that only depends on visual stimuli and the intention of a subject. This particular implementation, commonly known as the the P300 speller in the BCI community, is one of the promising illustrations of the potential and possibilities of BCI employing only EEG to record brain activity.

However, Schalk et al. [6], identify in their study two important issues regarding the construction of a BCI, the signal identification problem and the signal identification paradox.

The signal identification problem explains that the selection of EEG signals and their location, originated by specific brain activity, is not completely understood yet (there is essentially no theoretical basis), since even the fundamental processes behind brain activity are not well comprehended.

The BCI signal identification problem is fundamentally different to a normal classification problem in the sense that data classes are not easily defined a priori to consequently select the features. It has been only empirically shown sometimes that particular mental tasks have particular effects on specific brain signals, and still the definition of the tasks and signal features to implement some kind of classification is difficult, suboptimal and poorly defined. Moreover the identified signal features are normally subject-dependent and non-stationary.

So it is expected that multiple alternative features, like the P300 wave, could still be identified and employed for BCI applications. Furthermore, motor imagery and tasks normally defined to implement a BCI for disabled people could not be

useful when considering completely healthy individuals in diverse contexts aimed at machine control, critical applications and augmented reality.

This unexplored domain of signal features and defined tasks pose an interesting challenge that might be addressed by personalized unsupervised learning. In this study, topological data analysis will be presented as an approach to aid the exploration of the features of EEG signals under different brain activity states. Particularly the Mapper method [5] will be implemented as a tool to explore and visualize the signals in conjunction with hard and soft clustering algorithms.

On the other hand, the signal identification paradox is due to the fact that there is no a priori basis for selecting mental tasks and signal features, so the possible choices increase with increasing signal fidelity and the latter improves by defining and discriminating more classes of brain activity (signal specificity), which means that also the identification procedure and algorithmic training increase in complexity.

Moreover signals might change in time and under learning and interaction conditions in ways that are difficult to identify to retrain the algorithmic classifiers. Under this scenario of increasing complexity in the dynamics of the feature/task space, it is possible that the BCI performance may degrade even with better signal recording.

As an anwer to the mentioned paradox, Schalk et al. [6] propose the SIGFRIED (SIGnal modeling For Real-time Identification and Event Detection) methodology. Which consist on the perspective of detecting events that are unlikely to belong to a general class which is uninteresting for control, like a resting state, to then use the unlikelihood of events as a control parameter in the construction of a BCI. This approach greatly simplifies the collection of labeled information on an exploratory framework and will be employed in this study in conjunction with the Mapper method to try to reveal interesting areas on the feature space that might be discriminated to build control mechanisms on a BCI.

Furthermore there is an interesting phenomena appreciated during the development of BCI with different subjects under similar experimental conditions called BCI illiteracy. This consist on the inability of the algorithms that were able to build accurate classifiers based on the identified relation between signal features and tasks to work on a non-negligible portion of the subjects population (between 15% and 30%), as explained by Vidaurre et al.. [7]. It is possible that employing personalized unsupervised learning to detect relevant signal features for each subject will address this problem, as there seems to be in many cases no completely universal solution to relate specific brain states and feature signals.

Considering the potential of EEG and all the mentioned challenges to build a BCI, the main proposal of this study will consist on developing an exploratory framework based on personalized unsupervised learning, topological data analysis and the detection instead of classification perspective to tackle the emphasized problems while at the same time looking for new insights in the identification of useful EEG signal features.

2. TOPOLOGICAL DATA ANALYSIS

Nowadays data is being produced at increasing rates thanks to new experimental methods and developments in high power computing. Furthermore the nature of data is changing, now it is more high-dimensional and noisier with more missing parts then ever. As explained by Gunnar Carlsson [8], developments in geometry and topology might be employed to bring to light many informative features of this kind of datasets for which conventional methods, dependant on specific metrics or low dimensional spaces, might fail.

According to Gunnar [8] there are several key points regarding data analysis that justify the use of geometric and topological methods. For instance: When to obtain knowledge about how data is organized in a large scale is desirable, like finding out patterns or clusters that gives insights about the dataset; That metrics are not theoretically justified in many domains of problems like in biology; That spaces and their coordinates are not natural in any sense and are also commonly unjustified; And that summaries over the whole range of parameters when analyzing data can be more valuable than individual choices, like keeping the whole dendogram when realizing hierarchical clustering procedures to analyze data.

Then topology turns out to be a good candidate to address the mentioned issues, since, as stated by Gunnar [8], it is the branch of mathematics which deals with qualitative geometric information. In general it is the study of connectivity information, so topological methodologies like homology can help study the datasets. Also the geometric properties studied by topology are less sensisitive to the choice of metrics and do not depend on chosen coordinates on specific spaces.

Furthermore, the idea of building summaries over complete domains of parameters, when analyzing datasets, involves the notion of functoriality that is at the heart of algebraic topology and allows the computation of homological invariants from local information. Also it is known that information about topological spaces can be learned by simplicial approximation.

A great example of how topology can be applied to data analysis can be appreciated in the research done by Gunnar et al. [9], where they found a new type of breast cancer from microarray data, with a significant biological signature, that was completely ignored by previously employed clustering algorithms. This discovery was done in a completely unsupervised way by looking at the shape of the data obtained with the application of the Mapper method [5] and the selection of functions representative of the problem at hand, like the abnormality of the cancer tissue. Then the discovery was confirmed by going back to the specific clustered points of the dataset and studying their relation.

Moreover the mentioned research was one of the main motivations to apply topological data analysis and particularly the Mapper method in this study. Since EEG signal features can

constitute a very high dimensional and still quite unexplored dataset, like the microarray dataset employed to understand cancer tissue.

2.1 Details on the application of topological data analysis for this study

Point clouds, understood as a finite set of points for which a distance function applies, are the main objects to which the geometric and topological techniques are applied. As explained by Gunnar [8], one can think of point clouds as finite samples taken from a geometric object, perhaps with noise. The notion of point clouds is quite abstract and therefore any set of points defined in an n-dimensional space for which a distance function can be defined is a candidate for topological data analysis.

Then the features of a finite segment of an EEG signal, which is a discrete multidimensional time series, can be represented as a point of data. This implies that by considering consecutive segments of the EEG signal, one can obtain point clouds representative of brain activity and so of sets of tasks behind that brain activity. In this study the features of overlapped segments of the same length of an EEG signal will be considered to form the point clouds.

Once a point cloud is defined, one needs to represent somehow its topology to be able to apply any of the tools of homology to find homological invariants that will give us an insight on the properties of the underlying object analyzed. As explained by Gunnar [8], intuitively, a simplicial complex structure on a space is a representation of the space as a union of points, intervals, triangles and higher dimensional analogues. And it turns out that a simplicial complex provides a particularly simple combinatorial way to represent topological spaces.

There are several methods to build simplicial complexes to approximate the topology of the space represented by the point cloud, like the Cech, Vietoris-Rips and Witness complexes methods. But in this study clustering algorithms will be employed as explained in the Mapper method [5].

Clustering algorithms take a finite metric space as an input and produce as output a partition of the underlying space, where the subspaces defined by the partition are considered as clusters. In the context of a metric space, this means that points inside a cluster are nearer to each other than to points in different clusters.

As stated by Gunnar [8], clustering should be thought of as the statistical counterpart to the geometric construction of the path-connected components of a space, which is the fundamental block upon which algebraic topology is based. This justifies the use of clustering as an alternative tool for the construction of simplicial complexes as outlined in the Mapper method [5].

Nonetheless an important problem in the construction of the simplicial complexes under any method is the selection of the values of the parameters that in each case will result on the simplicial complex that best approximate the true topology of the object underlying the point cloud. This problem can be addressed by the ideas behind persistent homology. Specifically by looking at all the simplicial complexes defined by the whole range of values of the parameters considered and analyzing how the topological properties vary as the parameters' values change.

Robert Ghrist [10] surveys throughly how persistent homology can be employed on diverse point cloud datasets, by employing a representation of the induced algebraic characterizations called barcodes. Moreover Balakrishnan et al. [11] introduce some statistical ideas to persistent homology to separate short lived topological properties considered as "topological noise" from the real approximated "topological signal" with measures of statistical confidence.

This opens up the possibility of exploring the brain processes behind the EEG signals in this study in an unsupervised way under general considerations of time (by progressively considering new datapoints in time) and space (by considering features of multiple combinations of electrode locations on the scalp at a given time).

2.2 Details on the Mapper method

The Mapper method, presented by Singh et al. [5], allows the representation of a point cloud dataset as a simplicial complex. It is based on the idea of partial clustering of the data guided by a set of meaningful functions defined on the data.

After applying the method to the dataset, the obtained simplicial complexes can be analyzed with the techniques of persistent homology to reveal qualitative information about the underlying object represented by the point cloud dataset. Moreover the simplicial complexes generated by this method can be used to visualize and interpret the dataset directly thanks to the meaning assigned to the functions that guide the partial clustering.

The Mapper method steps, represented in Figure 1, can be summarized as:

1. Determining the point cloud dataset.
2. Selecting a small set of meaningful functions to map the points to a low dimensional space.
3. Segmenting the low dimensional space into intervals of length "l" overlapped with percentage "o".
4. Generating the subdatasets corresponding to the defined intervals in the low dimensional space.
5. Applying a clustering algorithm with automatic detection of the number of clusters in each subdataset to obtain the nodes of the simplicial complex.
6. Evaluating the intersections of clusters belonging to consecutive overlapped intervals to obtain the connections of the simplicial complex.
7. Building the simplicial complex for further analysis
8. In the case of further visualization of the simplicial complex, define additional visual properties to ana-

lyze the dataset, like color or size of the nodes.

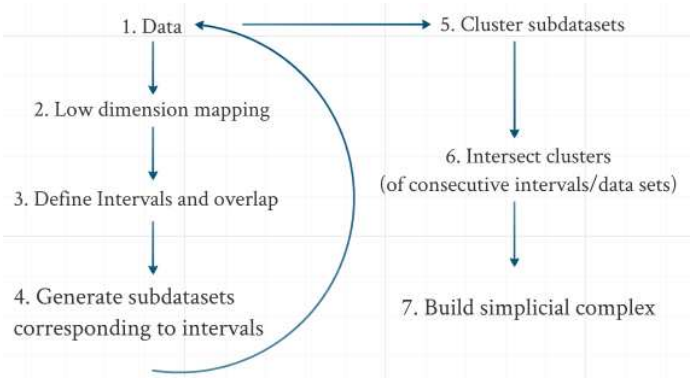


Fig. 1. Diagram of the Mapper method

The visual properties that can be established to have a more informative representation of the dataset when visualizing the resulting simplicial complex are numerous. For example one can define the color to the nodes to represent the average value of the meaningful functions used for the low dimensional mapping or define the size of the nodes to represent the proportion of points belonging to each cluster with respect to the whole dataset.

In the case of this study, coloring the nodes, using pie charts to replace the nodes and maintaining spatial configurations of the nodes, when drawing the simplicial complex, are the visual properties that will be defined to get an informative view on the structure of the brain processes and tasks underlying the EEG signals.

Furthermore, a more comprehensive explanation of the mapper method along with its application on multiple examples of trivial and non trivial point cloud datasets belonging to diverse problems on shape and object recognition can be found in the paper published by Singh et al. [5].

3. SIGFRIED (SIGNAL MODELING FOR REAL-TIME IDENTIFICATION AND EVENT DETECTION)

As was pointed out in the introduction, the BCI classification problem is peculiar in the sense that not only the features that represent classes but also the classes themselves have to be found and defined. This is an important problem, since there is no fundamental theoretical basis to define the tasks that are expected to generate changes on the signals recorded from brain activity and as greater signal fidelity is desired then more complex task definitions are also required.

Schalk et al. [6] explains the signal identification problem and paradox and how they greatly increase the time cost of developing and implementing brain computer interfaces.This delay the adoption of BCI technology even though there has been important advances in the production of mass consumption inexpensive devices for EEG.

The time and effort that a subject requires to train the algorithmic classifiers grows fast with the complexity of the task definitions even when details about the relationship between the signal and the task are known. Moreover the procedures based on specific features of EEG signals might not work for all subjects. Furthermore, to complicate even more the problem, signal features have shown to be non stationary and highly sensitive to feedback conditions imposed by interacting with the digital interfaces.

SIGFRIED constitutes an anwer to the mentioned difficulties because it only needs a small sample of only one reference class to be able to discriminate other classes. It might also use more than one class as reference and the resulting output in any case is a continuous feature that can be employed as an input for computer commands. In addition the methods behind it are not expensive computationally and easy to implement.

Schalk et al. [6] propose to define a rest category that could be used as a main reference to detect and analyze non rest activity that might be used as input for a BCI. In this study this suggestion will be taken into account to create a meaningful function to map the point cloud of the EEG signal to the low dimensional space in the Mapper method. This will allow to interpret the structure of the simplicial complex in terms of non rest or extreme events and to see if some distinction can be made between labeled tasks with respect to specific EEG signal features.

3.1 Details on SIGFRIED

SIGFRIED can be summarized in the following steps:

1. Specify signal features that will constitute the representation of brain activity in the segments of EEG signals.
2. Retrieve a labeled sample of the desired reference class. In this study a class representing an approximated rest state will be employed as reference.
3. Fit a Gaussian mixture model to the reference class.
4. Compute loglikelihood of each data point with respect to the fitted Gaussian mixture model.
5. Employ the loglikelihood of points as a continuous detection signal. In this study this measure will be employed to give meaning to the structure of the simplicial complex obtained from the application of the Mapper method.

In the original proposal of Schalk et al. [6], the Gaussian mixture model (GMM) is fitted to the reference class by employing and Expectation-Maximization procedure complemented with the Akaike Information Criterion to automatically determine the number of gaussian distributions in the mixture.

But in this study a more promising approach called free split/merge expectation maximization (FSMEM), presented by Wagenaar [12] and developed as an extension to the work of Ueda et al. [13], will be used to fit the GMM with an automatic detection of the number of Gaussian distributions in the mixture. In addition this approach will also be employed as a soft clustering alternative when determining the clusters of the

subdatasets corresponding to intervals of the low dimensional mapping when implementing the Mapper method.

4. EXPLORATORY FRAMEWORK (PARTIALLY DEVELOPED)

The main idea behind this study is to propose a framework that would allow any subject to set up a BCI by exploring the personalized and dynamic relationship between his self defined tasks on specific contexts of action and the EEG signals features' space derived from brain activity.

There are many challenges behind this idea. The first one is to counteract the signal identification problem and paradox. This is the main motivation to adopt the detection instead of classification paradigm and apply some of the ideas behind SIGFRIED.

The second challenge is determining the moment at which a fundamental change on a feature of the EEG signal has taken place in a high time resolution and continous signal setting. This can be addressed by considering multiple time scales on the segments of the EEG signal with a relative time point in common. That translates into two possibilities: an even higher dimensional space characterizing a point in time by the features of different segment lengths in the signal or studying the persistent topological properties of different time scales. Then topological data analysis can play an important role in the development of the framework.

The third challenge is that labeling can not be exact because of the fast and noisy changes on the EEG signal. Even in the current most carefully set up experiments with labeling, a big segment is considered in which the action that should generate a change in a signal feature takes place. But the exact segment of the signal that should correspond to the realized task is quite difficult to define. This motivates considering multiple time scales, unsupervised learning and a probabilistic perspective on the likelihood of the cloud points to try to find patterns in the EEG signal.

The fourth challenge is that the feature space can become very high dimensional thanks to spatial resolution (electrodes distributed around the scalp) and the huge amount of possible features that can be obtained from a signal. This again motivates employing topological data analysis as a way to explore the high dimensional nature of this dataset without trying to make too many assumptions about which features are important from the start.

The fifth challenge would be to implement computationally efficient unsupervised learning algorithms to assist the subject. Since the feedback of the framework and the desired BCI should work as fastest as possible to give the subject an intuitive and viable experience. This motivates the exploration of clustering algorithms and methods optimized for the specific nature of the EEG signal dataset, like a stream collaborative clustering.

Finally it is of most importance to create an interface intuitive enough so the subject can easily learn to navigate and comprehend the feedback from the machine learning algorithms that assist him to be able to successfully discriminate and choose areas of the feature space to stablish the BCI control structure. It is also important for the framework to allow the implementation, in the future, of any fundamental breakthrough on neuroscience or on the condition and activities of the subject. This can be achieved by exploiting the visualization capacities of the mapper method and minimizing assumptions during the exploration process.

The framework steps can be summarized as:

1. Settting up the online/offline exploration data generation
2. Exploring as much fundamentally different features of the EEG signal as possible.
3. Exploring the possible meaningful functions that can be defined for detection of unlikely events which can constitute the BCI control structure and the visualization tools.
4. Implementing the Mapper method with considerations on the partition of the low dimensional mapping and the clustering algorithms that would be optimal for the nature of the EEG signal and the BCI construction problem.
5. Implementing bayesian methods to asses any task classification and confidence statistics on the topological persistent properties of the dataset.
6. Computing the persistent topological properties of the dataset.
7. Visualizing and interacting with the exploration framework to build the BCI.

The main objective of this study is to confirm, up to step four, the viability of the framework to find structure in the EEG signal features. So the development of the rest of the framework and its optimization can be justified as necessary further research.

4.1 Online/Offline exploration data setup

In this study, we employ a dataset taken from the BCI competition IV[1] that is actually part of a bigger dataset from a study of Blankertz et al. [14]. The selected dataset is labeled and contains a category that can be understood as rest before intructions to execute a task like moving the right or left hand or foot are presented to subjects.

We considered the training dataset belonging to subject A, consisting of 200 trials for which the subject was randomly asked to move his left hand or foot after a period of rest. In each trial, the first 2 seconds consist of a white screen, then a cross appear in the center of the screen for the next 2 seconds and finally the instruction to move the left hand or foot appears over the cross and remains for 4 seconds. An example of a trial with the time segments categorized can be appreciated in Figure 2.

1 Dataset published on: http://www.bbci.de/competition/iv/desc_1.html

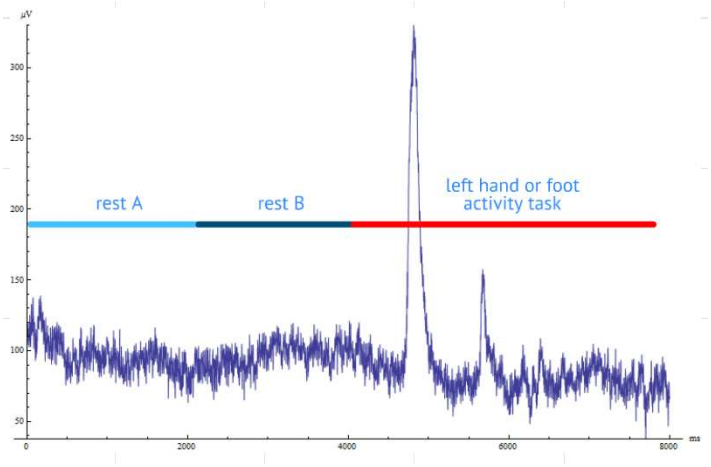


Fig. 2. Trial example with categories of time segments indicated

The EEG was setup with 59 channels (electrodes around the scalp), the signal have a time resolution of 1000Hz, which means a 1000 samples per second and was band-pass filtered between 0.5 and 200 Hz.There are in total four categories that are accurately labeled, two types of rest, left hand movement and foot movement.

Then we considered time segments of 300ms sampled consecutively from the signal of every trial, every 40ms in the case of the rest categories and every 20ms in the case of the movement categories. The length was selected to be 300ms since it is long enough to contain the big perturbations observed in the signals during the movement task.

So we ended up with a point cloud of 50200 points. Of these, 15200 points belonged to rest categories, 17500 points to left hand movement and 17500 points to foot movement. With a dimensionality of 17700 that correspond to the 300 time measures of the time series of each of the 59 channels (electrodes).

We did not consider a point every millisecond in this first approach of the framework for practical computational constraints in time and memory, since we would end up clustering datasets with millions of points in that case. Also signals are not expected to change radically from 1ms to the next so we would just be oversampling segments of signals with almost identical features.

4.2 Exploring features of EEG signals

The list of spatio-temporal features that have been considered for EEG signals in the literature is numerous. McFarland et al. [15] makes a good review of the most popular ones and their application on BCI. From these the Fourier Transform is number one on the category of temporal features and will be employed in this study to show the capacity of the framework to capture structure in the EEG dataset.

To test the framework, the average of the power spectrum, computed with the Fast Fourier Transform in Mathematica, in the frequency bands Beta (13-30Hz), Gamma A (30-100Hz) and Gamma B (100-200Hz) was considered for each datapoint. An example of the power spectrum of a datapoint and the corresponding band frequencies can be appreciated in Figure 3.

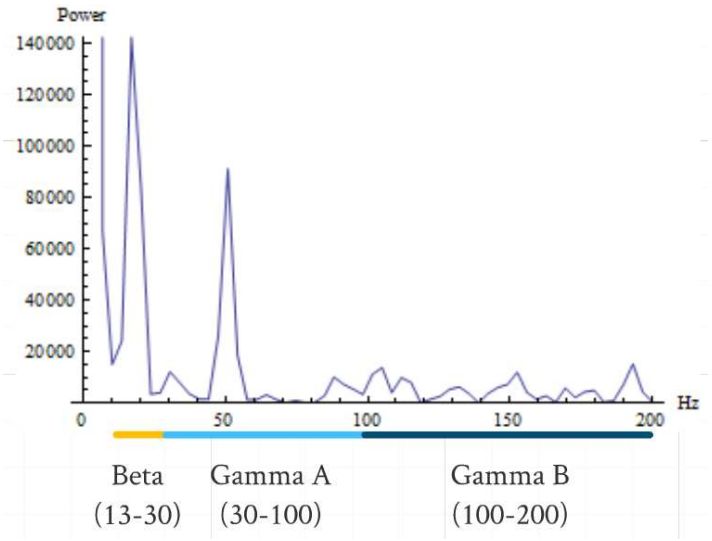


Fig. 3. Power Spectrum example with frequency bands indicated

Moreover we considered the averaged signal of the trials for each movement task.Then a subset of 12 channels was selected from the 59 channels set, based on the greatest difference between the average of the power spectrum in the mentioned frequency bands of the two tasks. The selected channels were "AF3", "Fz", "CFC5", "C5", "CCP5", "CP5", "P4", "P6", "PO1", "PO2", "O1" and "O2". How the difference on the average of power for the different tasks' averaged signal in the Beta band can be appreciated in Figure 4, as an example of the channel selection, although the Gamma bands were also considered.

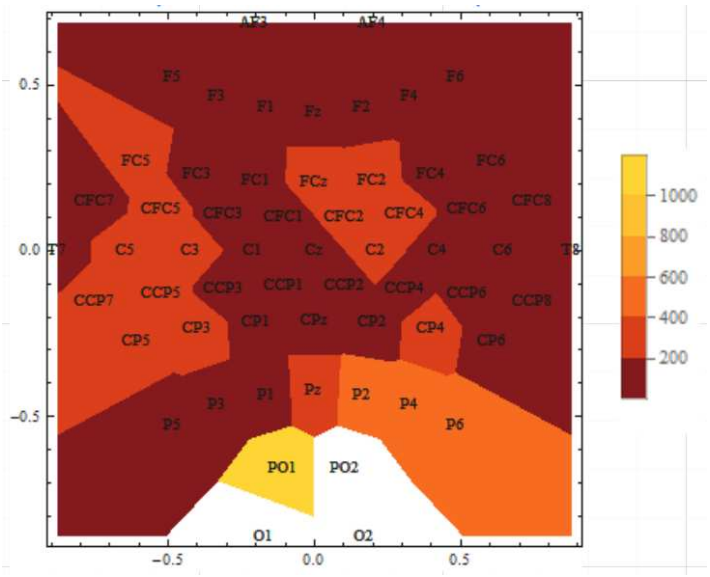


Fig. 4. Difference in the averaged power spectrum of the two movement tasks' averaged signal for the Beta frequency band (The channels are arranged as the 2d projection of their positions in the scalp) / (One can see the big difference in the channels PO1, PO2, O1 and O2, so these were part of the selected set of channels)

Finally each datapoint is represented by the average of the three frequency bands for each of the 12 selected channels, which results in a 36 dimensional representation of the cloud points. However in future research multiple additional features should be considered simultaneously. We are limiting the application of the framework to this feature extraction technique and to a smaller number of preselected channels, for practical computational constraints in the dimensionality of the data, particularly for working with soft clustering based on fitting a gaussian mixture model.

4.3 Exploring meaningful functions for detection

In this study we will explore the EEG signal point cloud from two different perspectives. The first one is implementing Schalk et al. [6] proposal to fit a GMM to rest categories to then be able to discriminate samples of other classes based on their loglikelihood. In this way we would be characterizing extreme events of activity as a continuous function.

The second one consists on considering notions of complexity as defined by Christopher James et al. [16] to characterize the brain activity. Particularly we will present the Fisher's information measure and contrast it with the perspective of SIGFRIED when analyzing the existence of structure in the EEG signal dataset.

In figure 5 the frequency of the loglikelihood of the points in the case of SIGFRIED and of the logarithm of the complexity measure can be appreciated. It is noticeable that the loglikelihood of SIGFRIED greatly separates a small portion of the

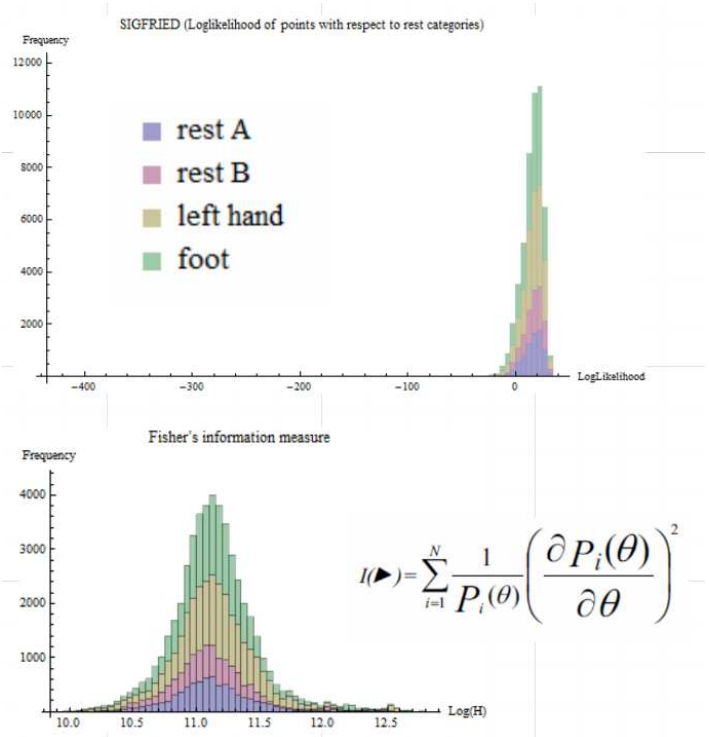


Fig. 5. Histogram of distribution of point categories for SIGFRIED and the information measure

points from the mountain of points likely to belong to rest activity, the loglikelihood of some points is even lower than -400 in contrast with the main distribution centered above 0. This makes a big contrast with the distribution of the values of the Fisher's information measure that seems to approximate more a normal distribution with short tails for all categories, which means that the complexity captured by the measure does not necessarily relate to the idea of rest vs non rest activity.

As an additional clarification, the Fisher's information measure is obtained by: First constructing a matrix with consecutive overlapped subsegments of length 100 ms belonging to the segment of 300ms that represents the EEG signal point for each channel; Then stacking the matrices of all channels to create a global matrix for the point; After a singular value decomposition (SVD) is employed to get the singular values; Finally the singular values are used as probabilities in the information measure formula. The formula for the Fisher's information measure can be seen in Figure 5 and an illustration of the computation of complexity measures taken from the work of Christopher James et al. [16] can be appreciated in Figure 6.

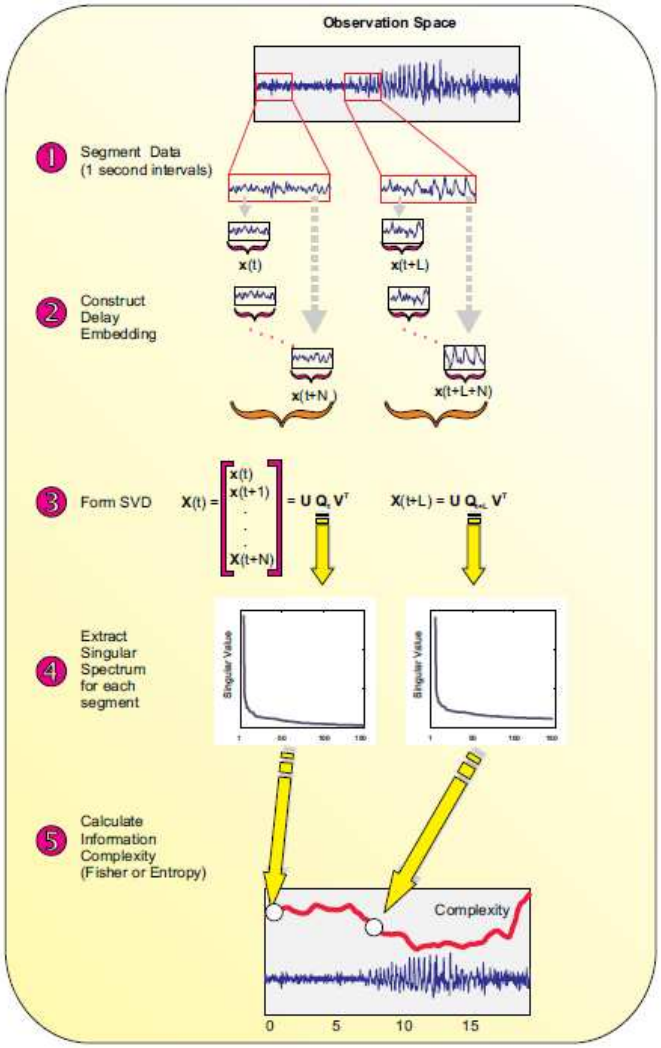


Fig. 6. Illustration of the computation of complexity measures

4.4 Implementing the Mapper method

The implementation of the mapper method realized in this study, considering the same structure of steps presented before, can be summarized as:

1. The dataset is formed by the cloud of points of segments of the EEG signal, also represented by a 36 dimensional vector that encodes the information of the average power spectrum for 3 different band frequencies for 12 different electrode locations in the scalp.
2. The SIGFRIED methodology and the Fisher's information measure will be employed as the meaningful functions to map the point cloud to a low dimensional space.
3. The low dimensional mapping will be partitioned by different number of intervals and varied overlaps to appreciate changes on the data structure due to resolution considerations.
4. The subdatasets corresponding to the defined intervals in the low dimensional space are generated.
5. Two different clustering algorithms will be employed with the Mapper method. The first one, as proposed by the Mapper method [5], is single linkage clustering [17] with the addition of an automatic detection on the number of clusters employing the Silhouettes statistic [18]. The second one is fitting a GMM with FSMEM[2] [12] as in the SIGFRIED methodology to explore the application of soft clustering for further Bayesian analysis in posterior research. In this way we obtain the nodes of the simplicial complexes.
6. The intersections of clusters belonging to consecutive overlapped intervals are evaluated to obtain the connections of the simplicial complexes.
7. The simplicial complexes are built for further analysis.
8. Pie charts will be employed instead of nodes to represent the proportion of the categories of points inside every node. Also a coloring of nodes will be employed as an alternative representation to show the value of the meaningful functions associated with the node's intervals. Moreover the simplicial complexes will be spatially arranged sometimes in such a way that it can be interpreted from the perspective of the meaningful function and the pie charts at the same time.

4.5 About the obtained simplicial complexes

The final result of the mapper method are the simplicial complexes on which persistent homology can be applied to reveal the most relevant and persistent topological properties of the underlying object, which in this case is the brain processes and tasks represented by the EEG signal features.

But before applying persistent homology, it is important to confirm that some interesting structure is being captured by the Mapper method and the functions defined to implement it. So considering that the power spectrum of white noise approximates a constant value we can model how the simplicial complex of completely unstructured white noise would look like and compare it with the structures of the simplicial complexes that we get with the proposed functions and features of the EEG signal.

In figure 7, an important difference between the structure of white noise and the methods employed to analyze the EEG signal dataset can be observed. Moreover there is also an important difference between the information measure and the SIGFRIED perspective on the structure of the generated simplicial complex. The nodes in figure 7 are represented as pie charts showing the proportion of the labeled categories that were clustered inside each node, and the spatial arrangement of the simplicial complexes is aligned with the notion of increaing complexity or decreasing loglikelihood as corresponds to the underlying perspective. In this case all the low dimensional mappings where partitioned in 20 intervals with a 50% overlap.

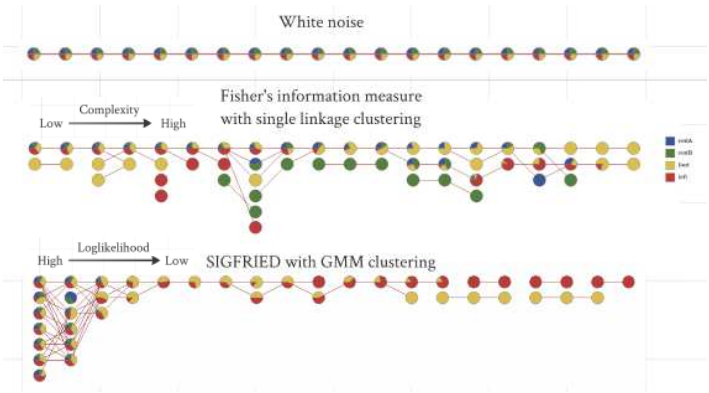


Fig. 7. Comparison of the structure of data in the simplicial complexes produced under the fisher's information and SIGFRIED perspective against white noise.

It seems that SIGFRIED effectively improves the discrimination capacity of the framework to separate the rest from non rest tasks and then also improves the capacity to further find important differences between the non rest tasks in an unsupervised way. But on the other hand the information measure is showing more patterns and structures that although might not be directly connected with the defined tasks, might give some important insight into some brain processes or different tasks in a different context, which is also important if we are exploring the feature space on an unsupervised way.

This might imply that both functions, SIGFRIED and the information measures, can be useful and perhaps complementary. Suggesting the possibility of combining them on a two dimensional mapping of the dataset when projecting the point cloud on a low dimensional space to partition it during the implementation of the Mapper method

Furthermore, in Figure 8, it is possible to see the rich structures that arise at a different level of resolution (considering 40 intervals with 50% overlap) in the case of the information

[2] Matlab code obtained from : http://www.mathworks.com/ matlabcentral/fileexchange/22711-free-split-and-merge-expectation-maximization-for-multivariate-gaussian-mixture

measure perspective. This confirms the potential of the proposed methods in the exploratory framework to discover structures in the EEG signal, possibly allowing the desired development of a BCI and a better comprehension of particular brain processes for a specific subject.

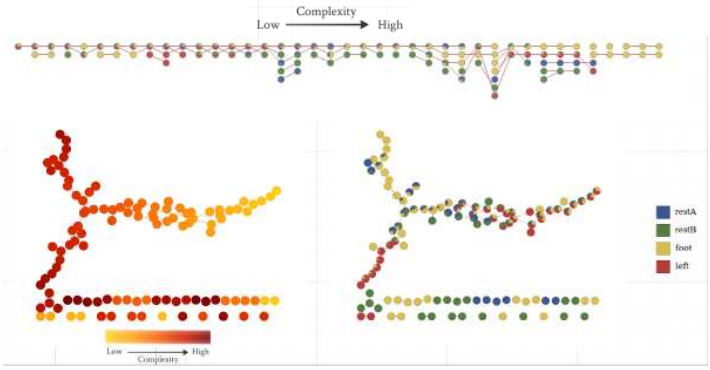


Fig. 8. Simplicial complex represented visually in different ways, produced under the Fisher's information measure and single linkage clustering.

In addition to Figure 8, Figure A1, in Appendix A, shows the important structural changes that can be seen in the generated simplicial complexes at different resolutions (different partitions on the low dimensional mapping and percentage of overlap). This shows the importance of applying the tools of persistent homology to be able to establish which of the appreciated topological features of the complexes are actually approximating the underlying object to the point cloud.

5. NECESSARY FURTHER DEVELOPMENT OF THE FRAMEWORK

After checking the potential of the Mapper method and in general of the notions of persistence homology, and of the perspective of detection instead of classification to explore EEG signals to build a BCI, there is still plenty to develop to complete a preliminary full implementation of the framework.

As part of the topics that need further development we can consider mainly:

1. Extending the use of Bayesian methods for detection and classification in framework, so that confidence measures can be taken at the different stages of the methodology.
2. Implementing clustering mechanisms optimized for the nature of the EEG signal dataset and the persistent homology techniques. Lets consider the high throughput nature of the EEG signal, the need to cluster different time scales and the possibility of understanding channels as diferent populations with similar features. Then it would be reasonable to propose the implementation of a stream collaborative soft clustering inspired on the works of Song et al. [19] and Pedrycz et al. [20]. In addition it would be interesting to consider the ideas of Chazal et al. [21] on clustering also based on persistent homology.
3. Implementing the persistent homology analysis to characterize the persistent topological properties of the EEG signals in time, space and resolution.
4. Implementing additional techniques to extract information from the simplicial complexes, like the appearance and persistence of branching structures that are not captured by homology.
5. Extending the number of meaningufl features for detection and for the clustering of EEG signal's point cloud.
6. Defining the visualization methods that will be employed to receive feedback and interact with the framework once it is implemented. To employ a two dimensional meaningful mapping on the Mapper method seems like a good alternative in contrast with the current one dimensional mapping.
7. Finally most of the algorithms and computations at different scales can run in parallel, so it would be crucial to exploit GPU parallelization to turn the framework into a fast and responsive interface for a BCI.

6. FINAL REMARKS

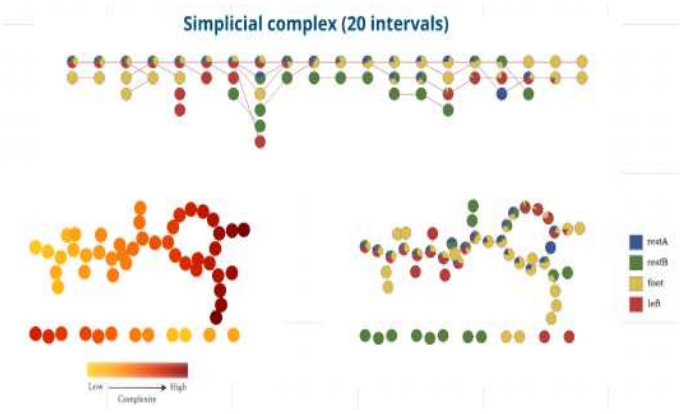
This study have presented important challenges on the construction of BCI and at the same time proposed a framework that might encode a solution to them, based on new mathematical methodologies and new paradigms in the BCI community.

The preliminary results of the potential of the proposed framework and its methods to characterize EEG signals, to understand the relationship between tasks and brain activity and very likely to construct in an easier and more accurate way a brain computer interface, seem very promising.

Nonetheless there is still quite a lot ahead to develop and explore before claiming the usefulness of the framework or settling down for the specific algorithms, methods and perspectives that assisted the different steps of the proposed exploratory framework.

7. APPENDIX A (SIMPLICIAL COMPLEXES)

In the following Figure A1, the effect of different levels of resolution due to changes on the number of intervals and the percentage of the overlap during the implementation of the Mapper method can be appreciated.



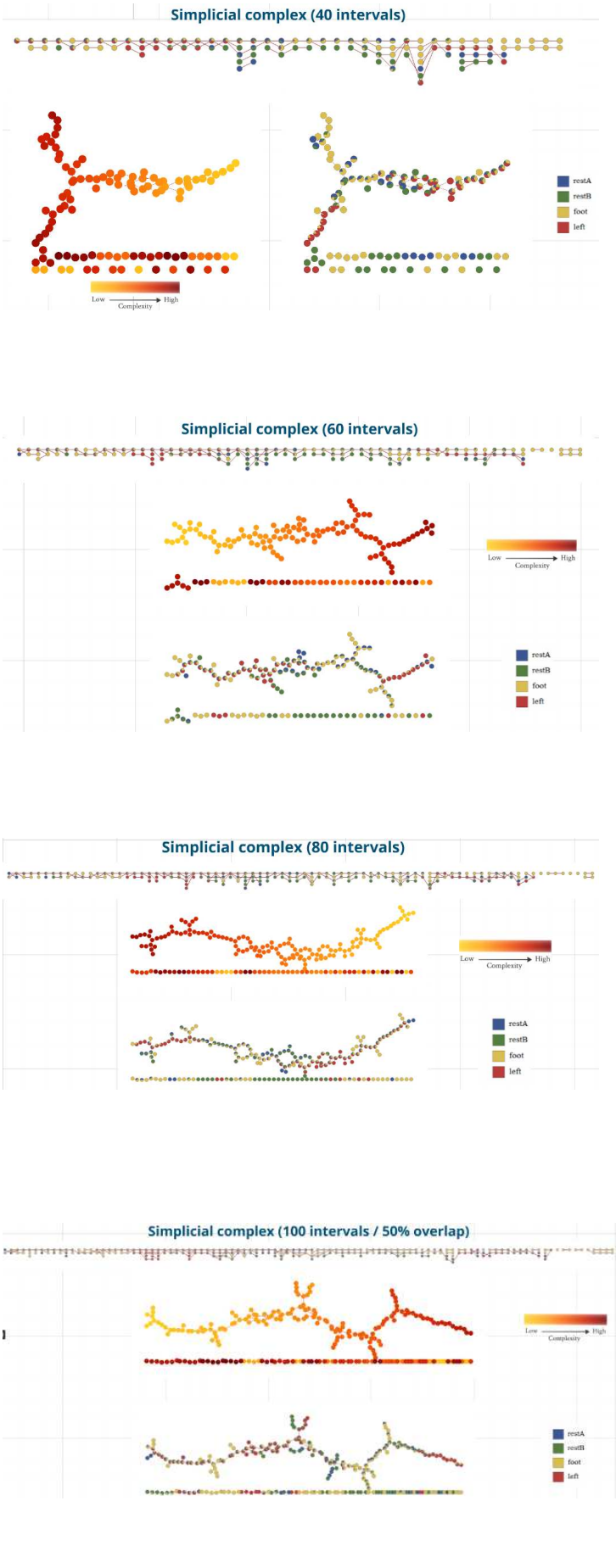


Fig. A1. Simplicial complexes based on the Fisher's information measure and single linkage clustering, with different resolutions of intervals with a 50% overlap followed by different resolutions of overlap with a 100 intervals.

8. ACKNOWLEDGEMENT

First of all, thanks to Christopher James for his dedication and commitment as supervisor of this miniproject. Thanks to Simon Davies and Sylvester Rozario from the Neuroengineering lab for their support. And thanks to the professors of the complexity centre for their time and help, particularly to professors Robert Mackay, Charo I del Genio and Colm Connaughton.

9. REFERENCES

[1] Olejniczak, P. (2006). Neurophysiologic basis of EEG. Journal of clinical neurophysiology, 23(3), 186-189.

[2] Campbell, A., Choudhury, T., Hu, S., Lu, H., Mukerjee, M. K., Rabbi, M., & Raizada, R. D. (2010, August). NeuroPhone: brain-mobile phone interface using a wireless EEG headset. In Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds (pp. 3-8). ACM.

[3] Picton, T. W. (1992). The P300 wave of the human event-related potential.Journal of clinical neurophysiology, 9(4), 456-479.

[4] Donchin, E., Spencer, K. M., & Wijesinghe, R. (2000). The mental prosthesis: assessing the speed of a P300-based brain-computer interface. Rehabilitation Engineering, IEEE Transactions on, 8(2), 174-179.

[5] Singh, G., Mémoli, F., & Carlsson, G. (2007, September). Topological methods for the analysis of high dimensional data sets and 3d object recognition. InEurographics Symposium on Point-Based Graphics (Vol. 22). The Eurographics Association.

[6] Schalk, G., Brunner, P., Gerhardt, L. A., Bischof, H., & Wolpaw, J. R. (2008). Brain–computer interfaces (BCIs): detection instead of classification. Journal of neuroscience methods, 167(1), 51-62.

[7] Vidaurre, C., & Blankertz, B. (2010). Towards a cure for BCI illiteracy. Brain topography, 23(2), 194-198.

[8] Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.

[9] Nicolau, M., Levine, A. J., & Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences, 108(17), 7265-7270.

[10] Ghrist, R. (2008). Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society, 45(1), 61-75.

[11] Balakrishnan, S., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., & Wasserman, L. (2013). Statistical Inference For Persistent Homology. arXiv preprint arXiv:1303.7117.

[12] Daniel Wagenaar (2000). FSMEM for MoG. http://www.danielwagenaar.net/res/papers/00-Wage2.pdf

[13] Ueda, N., Nakano, R., Ghahramani, Z., & Hinton, G. E. (1998). Split and merge EM algorithm for improving Gaussian mixture density estimates. In Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop (pp. 274-283). IEEE.

[14] Blankertz, B., Dornhege, G., Krauledat, M., Müller, K. R., & Curio, G. (2007). The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects. NeuroImage, 37(2), 539-550.

[15] McFarland, D. J., Anderson, C. W., Muller, K. R., Schlogl, A., & Krusienski, D. J. (2006). BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation. Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 14(2), 135-138.

[16] Lowe, D., James, C. J., & Germuska, R. (2001). Tracking complexity characteristics of the wake brain state.

[17] Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

[18] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[19] Song, M., & Wang, H. (2005, March). Highly efficient incremental estimation of gaussian mixture models for online data stream clustering. In Defense and Security (pp. 174-183). International Society for Optics and Photonics.

[20] Pedrycz, W., & Rai, P. (2008). Collaborative clustering with the use of Fuzzy C-Means and its quantification. Fuzzy Sets and Systems, 159(18), 2399-2427.

[21] Chazal, F., Guibas, L. J., Oudot, S. Y., & Skraba, P. (2011, June). Persistence-based clustering in Riemannian manifolds. In Proceedings of the 27th annual ACM symposium on Computational Geometry (pp. 97-106). ACM.