Statistical Learning of Biological Structure in the Human Brain

Dr. med. Danilo Bzdok

"But above all, master technique and produce original data; all the rest will follow."
Santiago Ramón y Cajal

Supervisors

Prof. Dr. rer.-nat. Stefan Conrad, Natural Science Faculty, HHU Düsseldorf, Germany

Prof. Dr. med. Simon Eickhoff, Medical Faculty, HHU Düsseldorf, Germany

Dr. Bertrand Thirion, INRIA, Saclay, France

Publications related to the present dissertation

Cumulative Impact Factor: ≈60

Original papers

Bzdok D, Grisel O, Eickenberg M, Thirion B, Varoquaux G. Semi-supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data. Under review at NIPS.

Bzdok D, Grisel O, Eickenberg M, Varoquaux G, Poupon C, Thirion B. Network-network architecture: Generative models of task activity patterns. Under review at Cerebral Cortex.

Bludau S*, **Bzdok D***, Gruber O, Kohn N, Riedl V, Mller V, Hoffstaedter F, Eickhoff SB. Medial prefrontal aberrations in major depressive disorder revealed by cytoarchitonically informed voxel-based morphometry. *American Journal of Psychiatry*, in press. *equal contributions

Bzdok D*, Hartwigsen G*, Reid A, Eickhoff SB. A hierarchy of the left inferior parietal lobe in social cognition and language. *Neuroscience and Biobehavioral Reviews*, in press. *equal contributions

Bzdok D, Heeger A, Langner R, Laird A, Fox P, Palomero-Gallagher, Vogt BA, Zilles K, Eickhoff SB. Subspecialization in the human posterior medial cortex. *Neuroimage*, in press.

Eickhoff SB, Laird AR, Fox PT, **Bzdok D***, Hensel L*. Functional segregation of the human dorsomedial prefrontal cortex. *Cerebral Cortex*, in press. *equal contributions

Bzdok D, Langner R, Schilbach L, Laird AR, Fox PT, Zilles K, Eickhoff SB. Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. *Neuroimage*, 2013.

Review and opinion papers

Eickhoff SB, Thirion B, Varoquaux G, **Bzdok D**. Connectivity-based parcellation: critique & implications. *Human Brain Mapping*, in press.

Eickhoff, SB & **Bzdok D**. Neuroimaging and modeling. Where is the road to clinical application? *Der Psychiater*, 2014, in press.

Eickhoff SB & **Bzdok D**. [Statistical meta-analyses in imaging neuroscience.] *Klinische Neurophysiologie*, 2013, 44:199-203.

Book chapters

Bzdok D & Eickhoff SB. Statistical learning of the neurobiological of schizophrenia. In: *The neurobiology of schizophrenia*, Springer, Heidelberg.

1 Introduction

1.1 Analytical and heuristic accesses to nature

The world around us is complex and volatile. A large proportion of human research efforts are undertaken in an *analytical* fashion based on the "the unreasonable effectiveness of mathematics in the natural sciences." This was phrased by the Hungarian-American physicist, mathematician, and Nobel laureate Eugene P. Wigner (1960). The language of mathematics is a powerful tool to describe, formalize, and predict phenomena in nature. The author emphasizes that it is not imperative that natural regularities exist in the world. He goes on to say that it might be even more surprising that humans can actually find these regularities and use them to their advantage. Similarly, Albert Einstein said: "The most incomprehensible, is that the world is comprehensible." Starting from human-conceived axiomes we have derived always more complicated properties of and relationships between mathematical objects by formal proofs (Connes A., "A view of mathematics"). A logical pyramid of theoreoms is built that lead to always more general asseraations. We also have detailed knowledge of the limitations of these mathematical assertions. On the one hand, an identical regularity can often be equally well described in very distant branches of mathematics. On the other hand, identical mathematical conclusions have reemerged from derivation of a priori unrelated assertions. Indeed, the same formal language has proofed very apt in the study of completely unrelated topics and diverging scientific disciplines; from the movements of celestial objects in the universe studied in astronomy to the metabolism pathways governing the inner life of the cell studied in biochemistry. Many rules about the world can thus be perfectly grasped (Hardy, "Apology"). As another example, Fibonacci numbers (1, 1, 2, 3, 5, 8, 13, etc.) reappear in many natural phenomena. The number of petals of a flower and the spirals of a pinapple tend to be Fibonacci sequences. The family tree of honey bees is also governed by Fibonacci regularities. Even the proportions of human finger bones follow this formalism. Knowledge of such mathematical regularities allows to impose logical structure on the external world. It remains an unresolved philosophical debate whether we have *discovered* or *invented* mathematics. Yet, there is probably no doubt that mathematical conceptualization evolves as a feature of human cultural evolution (Tomasello, 2001). Even the most abstract mathematical concepts can be exchanged between individuals. Consequently, this knowledge resource can be easily passed on across generations and geographical distances. One may note that there is usually consensus among mathematicians about the architecture of their discipline. From an anthropological perspective, mathematical formalism appears to be one of the most powerful tools and most defining properties of the human species (S. Dehaene, "The Number Sense"). Indeed, Eugene Wigner concludes his praise of equations with the following words: "The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research [...]." (1960, p. 14).

However, this dogma has repeatedly been challenged formally and empirically. In formal approaches mathematics were shown incomplete and inherently contradictory because, in any axiomatic system, some true assertions cannot be proven ("incompleteness theorems", Gödel, 1931). Additionally, it is possible to define a real number with equidistributed digits that can however not be computed ("Chaitin's constant Ω", Chaitin, 2006). In empirical approaches the omnipotence of mathematical equations has also been challenged as the best possible way to describe and predict nature (Halevy et al., 2009; Hinton/LeCun/Bengio, 2015; Pietsch, 201?). This shift in scientific discourse is made explicit by three recent empirical observations:

1. Sophisticated, more accurate models can be outperformed by simple models that are fit with massive training data.
2. Simple, inaccurate models trained with rich data can outperform models that have been designed according to extensive domain expertise.
3. There appears to be a minimal threshold of training data such that the derived models suddenly exhibit emergence properties.

One prominent example for modelling nature in such a *heuristic* fashion is automatic language translation of human text and speech. Translation systems based on human-made grammar rules have perhaps never achieved satisfactory success (P. Norvig, 2011, "On Chomsky"). That is, analytical approaches based on thousands of book pages archiving domain expertise in form of deterministic rules appeared insufficient for building language models that can cope with real-world settings. Statistical machine translation was more successful by implementing probabilistic hidden Markov models (HMM) as a heuristic approach. The next word or N-gram is predicted only by the one (order 1) or few (order n) preceding ones with equal transition probabilities (Bengio, 199X "Markovian Models"). This special case of a recurrent neural network computes the conditional probabilities of the next language element depending on the most recent history of these elements based on a dictionary of known elements. The transition matrix of human language can be easily *learned* from data by observing a preferably long stream of real-world language (i.e., a "corpus"). The class of HMMs thus became a dominanting feature of computational linguistics. Today, virtually all professional translation software solutions for both written and spoken language are enabled by heuristic statistical models.

As an example from human biology, we currently have only few means to predict the toxicity of environmental chemicals and potential effects of new drug compounds on health. The complex and unknown phenomenon in nature here pertains to the causal link from a protein's *known* primary structure (i.e., 1D chain of amino acids) to its *unknown* tertiary/quaternary structure (i.e., the combination of 3D foldings of the amino acid chain that subserves function). Among the millions of existing proteins we only know the tertiary/quaternary structure of approximately 30.000 ones. The structural configuration is however necessary to identify the position of bindings sites for protein-protein interaction. Knowledge of these sites

in 3D space is crucially important to infer that protein's interplay with the human body. In this case, the learning problem is to derive a computational model from massive pairs of known primary protein structure and known protein properties, including toxicity, by intentionally treating the 3D structure as a black box. State-of-the-art neural network algorithms have very recently solved this heuristic learning problem better than probably any previous approach in academia or industry (Dahl et al., 2014; Unterthiner et al., 2015). These investigators thus showed that biological activity can be reliably predicted from single amino acid chains even without recourse to biological domain expertise. There might currently not exist an analytical counterpart to such structure-function mapping of proteins. Similar probabilistic scenarios would include predicting political outcomes, optimizing advertisement strategies, algorithmic trading in stock markets, and controlling self-driving cars.

In sum, humans can create machines to derive algorithmic predictions from the data. Observing a phenomenon in nature a sufficient number of times might be sufficient to algorithmically extract the heuristics of its interaction behavior with the world. This has recently entailed a shift of attention from model complexity to data complexity and from purely mathematical treatment to giving up some human control to self-emergening patterns. In the absence of an analytical access, simple statistical models can thus automatically formalize diverse classes of natural phenomena depending on the quality and quantity of the available data resources.

1.2 Two cultures of statistical modelling

Statistics is a branch of mathematics that has arguably been the overall most successful information science. Statistics aims at extracting information from data about the mechanisms in nature that generated these data. Given its eclectic character, it may come as no surprise that statistics has developed both analytical and heuristic strategies to model regularities of phenomena in nature. Yet, analytical and heuristic statistical cultures have developed independently (Breiman, 2001). They differ with regard to historical origin, mathematical foundation, and modelling goal.

The overwhelming majority of statisticians follow an analytical regime by adhering to *classical statistics* (CS) for *data modelling*. They hold that the phenomenon under study can be viewed as a black box whose inner workings can be described by a small set of underlying variables. It is up to the statistician in charge to choose the model that best reflects nature. Data are then used to estimate the parameters of that pre-specified model. Classical statistics has dominated research at the universities for almost 90 years now. Well known members of the CS family include for instance Student's t-test, ANOVA, and Chi-squared test. *Statistical hypothesis testing* has been introduced in the beginning of the last century (Fisher, 1925; Neyman and Pearson, 1928). The same approach is still practiced today in its original form (Goodman, 1999). The ensuing *p-value* measures how likely it is to observe the data at hand assuming the non-preferred null hypothesis (H_0) to find indirect evidence for the preferred alternative hypothesis (H_1). Despite the prevailing

presence of p-values, it has not been conceived by Fisher as an acid test to judge existing versus non-existing effects in nature. Rather, the intention was a preliminary tool to filter which potential effects should be more explicitly tested (Nuzzo, 2014). Notably, the drawn conclusions may be wrong if the hand-selected model is a bad description of the natural phenomenon under study. Nevertheless, statistical hypothesis testing probably fit perfectly in its time of inception and adoption. In fact, it was designed for use with mechanical calculators (Efron and Tibshirani, 1991). Gaussian distributional assumptions have been very useful in many instances to reach mathematical convenience and, hence, computational tractability. Additionally, it suited perfectly the Popperian view of critical empiricism in academic discourse (Popper, 1935/2005 "Logik der Forschung"): scientific progress is to be made by continuous replacement of current hypotheses by always more explanatory hypotheses by means of *verification* and *falsification*. The rationale behind hypothesis falsification is that even a lot of evidence cannot confirm a given theory in an *inductive* way, while a single counter example is able to proof a theory wrong in a *deductive* way. In sum, classical statistics was mostly fashioned for problems with few data points that can be grasped by plausible models with a small number of parameters chosen by the investigator.

In contrast, only a small minority of statisticians follow a heuristic regime by adhering to *statistical learning* (SL) for *algorithmic modelling*. This statistical framework is frequently adopted by computer scientists, physicists, engineers, and others without formal statistical brackground that are typically working in industry rather than academia (cf. Daniel and Wood, 1971). They hold that natural phenomena can be studied by estimating regularities in the inputs and outputs to the black box without making assumptions about its internal "true" mechanisms. A statistical model is thus derived that expresses relationships between the input and output variables whose parameters are learned by training data (Abu-Mostafa, 2012). Put differently, a new function with potentially thousands of parameters is created that can predict the output from the input alone, without explicit programming model. The input data thus need to represent different variants of all relevant configurations of the examined phenomenon in nature. Well-known members of the SL family include for instance k-means clustering, Lasso/Ridge regression, and support vector machine classification. Please note that SL here summarizes the seemingly more specific terms "data-mining", "pattern recognition", "artificial intelligence", and "machine learning" that are often employed inconsistently. The independent historical origin of CS and SL families is even witnessed by the most basic terminology. In the CS literature inputs to statistical modeling are traditionally called *independent variables* or, more recently, *predictors*, while these are commonly referred to as *features* in the SL literature (Hastie et al., 2011). When evaluating whether a certain problem is a possible target for SL three requirements come into play (Abu-Mostafa, 2012):

1. A regularity exists (if there is no pattern, then it might still be worth trying SL).
2. The regularity cannot be formalized analytically (otherwise one can still apply SL, but it might not

create the best model).

3. We have data on the problem (the more, the better).

This regime led to a surge of new computer-intensive statistical techniques since 1980 that can be difficult to compute on a normal calculator and that are less concerned with mathematical tractability (Efron, 1991). This development has been flanked by changing properties of datasets that are always higher-dimensional (i.e., more features per observation) and based on larger samples (i.e., more observations). This is a trend that is not specific to neuroimaging research but also takes places in other scientific disciplines, including but not exclusive to weather forecasting and economic predictions (Manyika et al., 2011). In sum, statistical learning was mostly fashioned for problems with many data points with largely unknown data generating processes that are emulated by a mathematical function created en passant by a machine.

Importantly, some statistical methods cannot be easily categorized by the CS-SL distinction. Statistical methods do, in fact, span a continuum between the two poles of CS and SL (Jordan/Frontiers in Massive Data, p. 61). Nevertheless, the two families of statistical methods can be easily distinguished by a number of archetypical properties. Bayesian statistics are however orthogonal to the CS-SL distinction and can be adopted in both methodological families in various flavors. Neither can the terms univariate versus multivariate (i.e., relying on one versus more than one input variable) be clearly grouped into either CS or SL. More generally, neither CS nor SL can generally be considered superior. This is captured by the *no free lunch theorem* stating that no single statistical strategy can consistently do better in all circumstances (Wolpert, 1996). The challenge relies in choosing the statistical approach that is best suited to the neurobiological phenomenon under study and the neuroscientific research object at hand.

Regarding modelling goals, CS and SL exhibit various differences. CS typically aims at modeling the black box by making a set of accurate assumptions about its content, e.g. the type of signal distribution. Contrarily, SL typically aims at finding any way to model the output of the black box from its input while making the least assumptions possible (Abu-Mostafa et al., 2012). In CS the phenomenon is therefore treated as partly known (i.e., the stochastic processes that generated the data), whereas in SL the phenomenon is treated as complex, completely unknown, and partly unknowable. It is in this way that CS tends to be analytical (i.e., imposing mathematical rigor on the phenomenon), whereas SL tends to be heuristic (i.e., finding useful approximations to the phenomenon). CS assumes a given statistical model at the beginning of the investigation, whereas in SL the model is generated in the process of the statistical investigation. In more formal terms, CS therefore closely relates to parametric statistics for *confirmatory* data analysis, whereas SL closely relates to non-parametric statistics for *exploratory* data analysis (Tukey, 1977 "Exploratory data analysis"). In more practical terms, CS is typically applied to experimental data that were generated the investigator controlled the variables of interest (i.e., the system under studied is perturbed), while SL is typically applied to observational data without such structured influence by the investigator (i.e., the

system is left unperturbed) (Domingos, 2012). The work unit for CS is the quantified significance associated with a statistical relationship between few variables given a pre-specified model. The work unit for SL is the quantified robustness of patterns between many variables or, more generally, the robustness of *special structure* in the data (Hastie et al., 2011). CS therefore tests for a particular structure in the data, whereas SL explores and discovers structure in the data. Formally, CS implements data modeling by imposing an a priori model in a top-down manner, whereas SL implements algorithmic modeling by fitting a model as a function of the data at hand in a bottom-up manner. Intuitively, the "truth" is believed to be in the model (cf. Wigner, 1960) in a CS-contrained world, while it is believed to be in the data (cf. Halevy et al., 2009) in a SL-constrained world.

As a drastically oversimplified, yet useful, conclusion, CS preassumes and tests *a model for the data*, whereas SL learns *a model from the data*. Indeed, both human and computer learning are theoretically more conceivable in a probabilistic rather than deterministic sense (Abu-Mostafa et al., 2012; Dayan et al., 1995; Friston, 2010; Gregory, 1980). Moreover, each probabilistc model can be viewed as a superclass of a deterministic model (P. Norvig, "On Chomsky"). Taken together, CS assumes that the data behave according to known mechanisms, whereas SL exploits computer algorithms to avoid the a-priori specifications of data mechanism.

1.3 The human brain as a complex phenomenon in nature

The human brain is a prime example of a black box that is complex, mysterious, and perhaps in part unknowable. It is frequently proposed that the human brain might be the most complex object in the known universe (Nature editorial, october 2014). With the language from above, the human brain might constitute a phenomon in nature that can perhaps *not* be perfectly grasped by mathematical formalism alone. More concretely, the *most pertinent structure* that we should assume for the human brain, when measured by contemporary functional neuroimaging techniques (cf. next passages), is currently unknown. Hence, the neuroimaging access to neuroscience can readily be framed as a problem of *representation learning* (Bengio, 2014). It is conceivable that this task can be solved without exhaustive neurobiological micro-/meso-/macro-level knowledge (Bostrom, 2014). This is always more supported by empirical evidence (e.g., Helmstaedter, M. et al. "Connectomic reconstruction" 2013 Nature) and it is a contention that is embraced by the present disseration.

From a global perspective, the genetic difference between our genetic equipment and that of our closest ancestors, the non-human primate, turns out to be strikingly small. This has encouraged the conviction that one or very few key genetic adaptations in the primate lineage have unchained an avalanche of cognitive and cultural inventions that led up to today's civilization (Tomasello, 2001). That is, the human species might be much more defined by the increasingly fast cultural evolution rather the ramifications of slow biological evolution. Crucial cognitive improvements, such as the emergence of verbal language, might have

fueled cultural improvements that, in turn, enabled further cognitive improvements and inventions et cetera pp. This form of *online learning* is a very plausible and decisive property of intact tissue of the central nervous system. As a first challenge in brain science, it might therefore be impossible to cleanly dissect the nature-nurture interplay into independent contributing factors that act during phylogeny (i.e., development of the species) and ontogeny (i.e., development of an individual organism). In this sense, investigating the limits between "nature" and "culture" in the human brain might equate with asking a paradoxical question (Dehaene & Cohen, 2007). Instead, a necessary factor for the high level of abstraction in human culture might have precisely been the inextricability, due to bidirectional influence, of neurobiological plasticity and relentless cultural exchange between human individuals in a non-stop, autopoietic optimization process (Vygotsky 1978, "Mind in Society"; Luhmann 1984, "Soziale Systeme"; Bengio 2013, "Evolving Culture").

Given this recent acceleration in cultural evolution (cf. Paul Virilio, "Open Skype"), it might be rather unlikely that the human brain has developed dedicated neuronal populations to subserve the panoply of novel behaviors. Rather, evolutionarily recent mental skills (e.g., reading and writing, explicit pedagogy, and symbolic mathematics) are realized by recombining low-level circuits that initially developed for other functional roles. This view has become known as "neural reuse" and "neural recycling" hypotheses (Anderson, 2010; Dehaene & Cohen, 2007). Non-human primates are lacking many of the sophisticated mental operations that are crucially important for maintaining human societies (Mesulam, 1998; Tomasello, 2003). In fact, the "social brain hypothesis" states that our computationally powerful brains are not an adaptation to solve problems posed by the physical environment, but for successfully coping with increasingly complex human social systems (Humphrey, 1984; Byrne et al., 1988; Dunbar and Shultz, 2007). Yet, it is becoming increasingly clear that socialaffective processing in the human brain is probably realized by domain-general brain regions and networks not specific to maintaining social interactions (Bzdok et al., 2015 "Neurobiology of Morality"; Behrens et al., 2009 "Computation"; Barret et al., 2013). These considerations entail a second challenge in brain science: It is probably impossible to know what purpose neural processing in a given part of the brain has originally evolved to serve. We can only observe external manifestations and correlative relationships of this latent biological purpose.

Importantly, no two human brains are alike. Quite the opposite, they differ with regard to the morphology of gyri and sulci, the topology of cytoarchitectonically and chemoarchitectonically distinguishable brain areas, the axonal connections linking these brain areas, as well as the history of their sensory inputs. The extent of a brain area and its inter-individual variability can be quantitatively examined with its relation to cognition and behavior, that is, performance in psychological tasks in the healthy or diseased brain. For instance, the volume of the amygdala is linked to interindividual differences in memory performance as well as many other (temporally transient) states and (temporally enduring) traits. As third challenge in brain science, it is currently unknown how interindividual differences in behavioral facets are mediated on the brain-level. The renowed neuroanatomist Santiago Ramón y Cajal wrote (1909): "The complexity of

the nervous system is so great, its various association systems and cell masses so numerous and complex, and challenging, that understanding will forever lie beyond our most committed efforts." More specifically, it remains largely elusive whether distinct behavioral differences between individuals are associated with changes of cell bodies, dendrites, axonal connections, and/or glial cells (Kanai et al., 2011). That is, we do not have clear understanding of how this set of microstructures interact to solve neural computation problems, let alone their interindividual differences. From a methodological perspective, volumetric modelling techniques conventionally employed in the neuroimaging field are naïve to many types of possible morphological differences. For instance, it is currently difficult to statistically grasp inversely proportional left and right hemisphere volumes or a medical condition that randomly affects either the left or the right brain per individual (Ashburner et al., 2011).

Worth to be proposed as an independent challenge of brain science, the secret of interhemispheric asymmetry is yet to be unveiled. The connectivity differences between the left and right brain are for instance currently underresearched. They are even hardly known in the monkey (Stephan, 2007) that usually serves a fallback system for human connectivity investigations (Mesulam, 2012 "The evolving"). In humans, the majority of homologous brain areas feature direct anatomical connections. Nevertheless, as two textbook examples, why the language and attention processes typically lateralize to the left and right hemisphere, respectively, is currently understood only in modest fragments (Corbetta 2000; Stephan et al., 2003; Price et al., 2010).

It is further unlikely that we will reach exhaustive understanding of the human brain by mere *observational*, as opposed to *interventional*, classes of research methods (cf. J. Pearl, 2000 "Causality"). This idea is reflected in Edward O. Wilson's words "disturb Nature and see if she reveals a secret" as well as in G. M. Shepherd's words "Nothing in neuroscience makes sense except in the light of behavior." Purposely induced focal lesions of brain tissue in rats have early been systemically related to resulting differences in behavioral performance indices (Franz and Lashley, 1917). In hamsters, cats, and monkeys, decortication entails only small sensory or motor effects, while such tissue impairments of the neocortex in humans result in much more pronounced and less reversible functional deficits (Lashley, 1952; MacLean, 1982), which points to increasing corticalization of brain function. In humans, brain lesion studies have been the most common approaches to localize brain functions until about 20 years ago. However, inferring neurobiological insight from lesion findings constitutes yet another challenge to brain science. It constitute an overly simplistic conclusion that changes in behavior after destroying brain tissue in a circumscribed brain area directly reveals functional roles of that brain area (Young, 2000). It is a limitation of these studies that they attempt to derive the *normal* function of an area from the effects of *damage* to that area. First, the destroyed brain area might primarily subserve inhibitory effects, such that abolition can increase neural processing subserved in remote areas mediated by network connections. Second, a large fraction of human lesion cases are stroke patients. The spectrum of lesion patterns found in these populations is however seriously limited by the

existing spectrum of brain vessel anatomy (e.g., the majority of ischemic strokes affect the Arteria cerebri media). Third, there is probably not a single psychiatric disorder that would be characterized by very *focal* (as opposed to distributed) differences in brain structure (cf. Goodkind, 2015 JAMA). More generally, it is still a matter of debate whether structure (i.e., locally specific micro- and chemoarchitecture), connectivity (i.e., short- and long-rang axonal targets), and function (i.e., lesion-induced behavioral changes) reflect three viewpoints on the same heterogeneity of a particular brain area (Passingham et al., 2002; Kelly et al., 2012 Neuroimage).

Each area in the brain exhibits activation patterns of neuronal populations with oscillatory regularities. These oscillatory circles and their associated behaviors are highly preserved in mammalian evolution (Buzsaki, 2013 "Scaling brain size"). Perhaps since Hubel and Wiesel's (1965) description of increasingly complex processing of neurons in the primary visual cortex neuroscientists tend to think information processing as serial sequences of sensory bottom-up and modulating top-down information streams. Axonal feedforward and feedback connections are indeed a very good predictor of *what* the next processing step is. Yet, brain oscillations are capable of predicting *when* this next processing step will occur. Oscillation measured by EEG and MEG techniques might be the most attractive access to another challenge in brain science: *the binding problem* (Singer, 1999; Engel, 2001; Varela, 2001). We are far from understanding how environmental perturbation by multi-sensory stimulation is coherently integrated and linked with prior experience into a holistic higher-order percept via spatially distributed and temporally coherent electrophysiological activity. In animals, oscillatory but not spiking activity of neuronal populations appears to be closely associated with sensory input processing. The interpretation of oscillation findings is however demanding. This is because they simultaneously reflect a maintenance equilibrium, sensitivity to external stimuli, and formation of processing outputs. For instance, perception of environmental stimuli is an intrinsically probabilistic process with nonidentical results depending on the state of ongoing oscillatory circles. Additionally, different "rhythms" (i.e., frequency bands) flank each other in a same brain area in an interacting fashion. The same rythm can reflect different categories of computational processes in different brain areas and networks. Some brain structures are characterized by specific rythms that may not be found in the rest of the brain. Different frequency bands can subserve a same cognitive process, while different cognitive processes can be realized by the same frequency bands. Finally, high frequencies govern large-scale networks in the brain that, in turn, influence small local neuronal spaces with slow oscillatory patterns.

Also from a philosophical perspective the neuroscientist faces problems when articulating observations of phenomena in the brain. For instance, brain areas or experimental effects are frequently described according to "emotional" versus "cognitive" interpretational categories. However, this class of judgments implicitly preassumes the neurobiological validity of traditional psychological categories. That is, it assumes that those two concepts have a discrete representation in measurable neurobiology. Yet, as another major challenge to brain science, it remains elusive how and to what extent psychological terms, such as "emotion"

and "cognition" (Pessoa, 2008; Van Overwalle, 2011), map onto regional brain responses (Laird et al., 2009; Mesulam, 1998; Poldrack, 2006). Potentially unjustified a-priori hypotheses are imposed on the organization of human brain systems. It should hence be carefully called into question what terms are an adequate word choice to refer to discrete neurobiological processes. More globally, confusion introduced by human language itself is at the origin of many scientific problems (Wittgenstein, 1953/2001 "Philosophical Investigations"). The grammatical and lexical constraints of human language might be too tight to allow for unequivocal description of the diverse circumstances humans encounter in science and ever-day life. According to Wittgenstein the meaning of language is primarily defined by its practical use in concrete situations, rather than decontextualized abstractions necessarily pre-shared by interlocutors. Words might not have an objective meaning equally accessible to and understood by everybody (e.g., also specialisation alters consciousness according to Habermas, 1984/87 "Theorie des kommunikativen Handelns" volume 1/2). This is all the more the case for language descriptions of phenomena that do not occur in every-day reality. In this sense, discussing subtleties of abstract neurobiological concepts, which can hardly be practically experienced, are frequent subject to ambiguity, thus leading to unnoticed misunderstanding and unresolvable paradoxes (cf. Bostrom, 2002; Watzlawick et al., 1967). Biological processes in the brain are an instance of such not directly experienceable phenomena underdetermined by human language that entail interpretative conundrum. More concretely, there is still no community-wide consensus on a comprehensive description system of human mental operations (Poldrack, 2006 and 2011). This has caused considerable heterogeneity in how neuroimaging experiments have been motivated and conducted. Moreover, it resulted in frequently inconsistent findings that are difficult to reconcile conceptually. Statistically, rather than falsely rejecting (i.e., type I error) or falsely accepting (i.e., type II error) the null hypothesis, previous experimental fMRI studies motivated by preassumed psychological categories might have commited "the error of the third kind" (Kimball et al., 1957): providing an accurate answer to an inadequate research question. It might be more useful to strive towards "an approximate answer to the right question" (John W. Tukey) given that "all models are wrong" (George Box) anyways. In sum, cognitive neuroscience has so far heavily relied on concepts historically inherited from traditional, non-neurobiological scientific disciplines. These considerations are especially relevant to investigations whose conclusions heavily rely on CS. Statistical hypothesis testing makes the strong implicit assumption that the semantic concepts used to formulate the null and alternative hypotheses are "true" (i.e., neurobiologically congruent).

The last challenges to the neuroscientist mentioned here are of epistemological origin. Biology as a whole has a modest legacy in abstract theory. This probably includes the history of the biology of the brain. In particular, the spectrum of permissible conclusions that can be drawn from neuroscientific investigations is strongly conditioned by the following three questions (Carruthers, 2009; Dehaene, MBE 2007):

1. Does the human brain offer sufficient computational resources to grasp, formalize, and predict itself?

2. Is the human mind capable to reflect upon itself by directly contemplating itself via introspection or by indirectly contemplating an internalized self-model acquired through interaction with others?
3. To what extent is the self-reflexive description of the phenomenology of the human mind by the human mind itself immanently limited and paradoxical?

Taken together, there are many intricacies about neurobiology and the mosaic knowledge that we currently have about it. Despite ≈200 years of neuroscience, we are probably not even close to something like a unified theory of brain function that neuroscientists from different fields would accept (cf. Friston, 2010 "Free energy principle"; Bar, 2009 "Predictions"). This caveat considerably complicates the formulation of precise, neurobiologically valid hypotheses that can be experimentally tested in targeted studies. Therefore, it might be helpful to use heuristics-establishing statistical approaches for pattern discovery instead of classical statistics alone. Discovering the mystery of the brain *exclusively* by successive falsification of entirely human-conceived, intimately language-dependent, and dichotomically framed hypotheses might be viewed as hubris by some (cf. Cajal, 1909; Cohen, 1994). Therefore, the present dissertation is built on the assumption that we might not reach an *exhaustive analytical understanding* of the brain any time soon and that a more pragmatic access may rely in the *heuristic approximation of brain mechanisms* by statistical learning models. Such an attempt to learn patterns from data would follow the same direction as recent research developments in language translation and drug discovery (cf. 1.1).

1.4 The curse of dimensionality

Not only neurobiological and conceptual challenges, but also the increasing quantities of analyzed data put neuroscientific research to the test (Gorgolewski & Poldrack, NNR). Today's neuroimaging methods offer very high resolution in space (especially fMRI and PET) and time (especially EEG and MEG) (Amunts et al., 2014 Science; Buzsaki & Draguhn, 2009 Science). The mere number of features poses serious statistical challenges to the investigator. It is the neuroscientific version of what Richard Bellman called the *curse of dimensionality* (1961). At the root of the problem, all data samples look virtually identical in high-dimensional data scenarios. Accustomed to regularities in 3D neighborhoods, human intuition is often led astray in how data behave in input spaces with an extreme number of variables.

The more dimensions an input space spans, the further the data points are away from each other (Hastie et al., 2011). Conter-intuitively, measuring the distance between a randomly selected data point and its closest uniformly distributed neighbors, reveals a shell-like occurrence probability of these neighbors, rather than a centered probability mass. Put differently, when approximating a hypersphere by a surrounding hypercube, the probability mass of the hypercube would almost entirely lie outside the hypersphere (Domingos, 2012). Put in yet another way, a space divided into isotropic units grows exponentially in the unit number with linearly increasing dimensionality. As the main practical conclusion, the amount of data necessary to populate these units also grows exponentially with linearly increasing input variables (Bishop, PRML).

Additionally, the target function is almost always unknown in statistical learning investigations. Hence, we frequently have no knowledge of whether or not special structure may exist in the input data that can be exploited. Knowledge of special structure of the phenomenon under study can reduce both *bias* (i.e., difference between the target function and the average of the function space derivable from a model) and *variance* (i.e., difference between the best approximating function from the function space and the average of the function space). This is a rare opportunity in SL because increasing, for instance, the model complexity typically increases the variance and lowers the bias, and vice versa. In particular, the problem of overfitting in SL has an immediate relationship with the multiple-comparisons problem in CS (Domingos, 2012). The *bias-variance decomposition* captures the fundamental tradeoff in statistical modeling between approximating the behavior of the studied phenomenon and generalizing to newly generated data describing that behavior.

A peacefully coexisting conceptual framework exists in SL that is independent of the unknown target function. The *Vapnik-Chervonenkis (VC) dimensions* formalize the circumstances under which learning processes can be successful (Vapnik, 1989, 1996). This comprises any instance of learning from a number of observations to derive heuristic rules that capture properties of phenomena in nature, including learning in humans and machines. Formally, the VC dimensions measure the complexity capacity of a class of approximating functions (i.e., the function space). Practically, good models have finite VC dimensions and are therefore capable to generalize to new data. Bad models have infinite VC dimensions that are unable to make generalization conclusions on unseen data, regardless of data quantity.

More concretely, SL approaches that incorporate locally varying functions in small *isotropic* neighborhoods will fail to generalize in high-dimensional data scenarios. SL approaches that overcome the curse of dimensionality typically incorporate an explicit or implicit metric for *anisotropic* neighborhoods (Hastie et al., 2011). It is the *hyperparameters* that govern the smoothing behavior of the imposed local neighborhoods. In so doing, the *hypothesis set* (i.e., each function in the function space represent a hypothetical solution to the estimation problem) is hopefully reduced to a reasonable pre-selection (cf. *regularization*). Guiding the statistical estimation process by complexity restrictions can alleviate the curse of dimensionality. First, we can deliberately exclude members of the hypothesis set. Viewed from the bias-variance trade-off, this calibrates the sweet spot between underfitting and overfitting. Viewed from Vapnik's statistical learning theory, the VC dimensions can be reduced and thus the generalization performance increased. Second, there is an infinity of possiblities to restrict the hypothesis set. Yet, these choices are typically guided by external knowledge beyond the data at hand. Third, different complexity restrictions typically lead to different best approximating functions.

In sum, the choice of any statistical method contraints the spectrum of possible results and of permissible interpretations. Any scientific discovery in the brain is only valid in the context of the complexity restrictions that have been imposed on the neurobiological phenomenon of interest. No single statistical strategy, be

it SL, CS, or other, can consistently do better in all neuroscientific investigations (Wolpert, 1996). The present dissertation is hence dedicated to the juggling with complexity restrictions to neurobiological reality as observed by fMRI scanning.

1.5 Imaging neuroscience

Functional specialization in the Cortex cerebri of humans has been investigated in the nineteenth century predominantly by lesion reports (Harlow, 1848, 1868; Broca, 1865; Wernicke 1881 "Die acute, hmorrhagische Poliencephalitis superior"). Brain lesion studies and brain stimulation during surgery were the mainstay of neuroscientific research for a long time, until they were complemented by axonal tracing studies for connectivity analysis in animals (cf. Mesulam, 1976). Today, functional magnetic resonance imaging (fMRI) is the most frequently chosen approach for non-invasive, in-vivo brain research in humans, counting more than 1,000 new neuroimaging publications per year. The impact of fMRI is explained by the availability of brain scanners in medical institutions, its non-invasivness, and its significant spatial resolution (1-2 mm, Engel et al., 1997) and temporal resolution (a few seconds, Jezzard 200X). fMRI enables the localization of neural activity changes at the synapse by means of measuring the accompanying changes in the oxy-to-deoxyhaemoglobin ratio in local draining veins (Roy et al., 1890; Ogawa et al. 1990/1993). For instance, onset of vibratory stimulation of a participants' hand entails regional accumulation of metabolic equivalents that cause regional blood flow increase ("neurovascular coupling") in the contralateral somatosensory cortex (Fox et al., 1986). In particular, the measured BOLD (blood oxygen-level dependent) signal exhibits an initial dip after the onset of neural activity increase that is attributed to the fast local increase in deoxyhemoglobin. The ensuing hyperperfusion and the thus generated (relative) hyperoxygenation then dictate the BOLD signal shape (i.e., "hemodynamic response function"). It is slightly different across the brain regions of an individual, across individuals, and probably across different tasks. Neural activation is finally followed by re-inhibition of blood flow observable as an undershoot at the end of the BOLD signal (Logothetis et al., 2001). Juxtaposing neural activity and corresponding BOLD signals, the BOLD signal is at least one order of magnitude noisier, scales roughly linearly with neural activity, and is better predicted by local field potentials than multi-unit spiking activity. The BOLD signal is possibly more associated with input to and processing in a local neuronal population rather than its output. There is thus no clear-cut quantitative relation between the spike rate of neuronal populations and the ensuing BOLD response. Rather, the BOLD signal reflects a mixture of transient spikes and continuous membrane potentials (Logothetis et al., 2004). As a central property, there is a tradeoff between coverage of sampled brain tissue, spatial and temporal resolution. For instance, augmenting the spatial resolution, while keeping brain coverage contant, deteriorates the temporal resolution. Finally, the regional responses in single individuals are transformed into a standard brain space (i.e., "spatial normalization" into the "Talairach-Tournoux" [cite] or "Montreal Neurological Institute" [cite] coordinate systems) for comparability and statistical analysis on the group-level.

Based on local changes in cerebral blood flow, experimental fMRI has provided insight into the cerebral localization of specific tasks related to sensory processing, motor actions, and affective functions (Brett et al., 2002). This is achieved by performing fMRI on an individual that lies inside the scanner magnet while attending and responding to psychological tasks, compared to the absence of that task. Usually, the neural correlates of a given task (i.e., a mental process of interest) are isolated by subtraction of the activation measured during a closely related task (i.e., control task) that is supposed not to evoke the mental process of interest. This relies on the principle of "pure insertion" that cognitive subtraction between the psychological processes of both target and control tasks is possible due to large absence of interaction between them. Although this assumption may not be tenable in many practical cases, the principle of direct task comparison has been widely adopted since it has been shown to be neurobiologically useful, as well as statistically robust and reproducible (Friston, Zarahn, Josephs, Henson, Dale, 1999). In many instances, analysis and interpretation of the brain imaging data is often performed by integrating additional behavioral data (e.g., task reaction times in the simplest case). Dozens of scans of a same experimental task that cover metabolic changes in the whole brain are acquired for enhanced sensitivity. The spectrum of neuroimaging-compatible tasks is practically only limited by the scanner surroundings and the interdiction of head movements. In this way, fMRI tasks have revealed the location patterns of various regionally specific effects in health and disease.

In contrast, in the absence of task (i.e., during mind wandering), the human brain is not at rest. While most fMRI studies focus on the minority of neural activity changes conditioned by external stimulation, increasing attention is devoted to the majority of neural activity patterns that underlie the biochemical maintenance of the neural "house-keeping" architecture. That is, the BOLD signal can also be measured in a task-unconstrained fashion by probing participants that lie in the scanner without following a defined psychological task. Participants are instructed to think of nothing in particular let their minds go, and leave their eyes open/closed or look at a fixation cross. During mind wandering humans typically mentally shift between various heterogeneous types of thoughts, memories, and predictions. This is why resting BOLD patterns are believed to reflect the repertoire of cognitive operations that the human brain can perform (Smith et al., 2009). From a neurophysiological perspective, intra- and inter-neuronal activity continues in the human brain's resting functionality. The resting-state BOLD signal reflects fluctuations in physiological signals recorded in the absence of task as reflected in a voxels' time courses. Importantly, the (small) amplitude of the resting-state signal is modulated by transient psychological states (e.g., arousal, attention, and alertness), but also cardiac and respiratory influences. Indeed, the decomposability of this signal measurment into independent components suggests a set of distinct influences rather than one coherent signal pattern (Fukunaga et al., 2006). More specifically, evidence exists in favor of a neuron-, metabolism-, vasculature-, and oxygen-driven genesis of the resting-state BOLD signal. More specifically, correlation analysis can detect temporal coincidence in the spontaneous, slow fluctuations (0.01 - 0.1 Hz) of rest BOLD.

This is taken as a measure of functional coordination between topographically distant parts of the brain. Measuring these coherent spatiotemporal couplings in resting-state BOLD fluctuations yields a set of robust neural networks. It led to the discovery of a set of so-called *resting-state networks*. In sum, the biggest fraction of the various brain signals does not correlate with a particular behavior, stimulus, or experimental task. These partially uncouple in a task setting, but the relative change is small. It is commonly agreed that the variability in the RS signal is related to the individual's (unconstrained) mental operations. It likely represents a physiological instantiation of a human beings' default mental repertoire.

A property of the brain that we might not have discovered without the advent of neuroimaging methods is the so-called *default mode network* (DMN). The present dissertation is closely related to this particular resting-state network that is a pure result of serendipity (Shulman et al., 1997; Gusnard et al., 2001). 15 years ago, the soon to be called DMN was initially proposed to be exclusive in decreasing neural activity consistently during experimental paradigms requiring stimulus-guided behavior. That is, the DMN was believed to increase neural activity in the idling, unconstrained mind and decrease activity during stimulus-driven, goal-directed tasks (Gusnard et al., 2001). On a macro-scale, the metabolic baseline turnover is not equally distributed across the brain. Interestingly, the brain areas of the DMN include the hot spots of highest metabolic consumption that locate, first, to the posterior cingulate cortex extending into the adjacent retrosplenial cortex and precuneus and, second, to the medial prefrontal cortex extending into the anterior cingulate cortex (Raichle et al., 2001; Reivich et al., 1979). It was later even argued that this network is systematically anti-correlated with brain regions more active during task performance (Fox, et al., 2005). Indeed, goal-directed task performance improves with increased activity in saliency-related areas and decreased activity in default-mode areas (Weissman et al., 2006). Conversely, increased activity in DMN areas were linked to increased occurrence of task-independent thoughts (i.e., mind-wandering) during task execution (Mason et al., 2007). Two fMRI studies employing Granger causality analysis further corroborated the anti-correlation by indicating negative influence of the default-mode on the saliency network (Pisapia et al., 2012) and vice versa (Sridharan et al., 2008). This anti-correlation was recently challenged by repeated reports of brain regions exhibiting both task-constrained and task-unconstrained increases in neural activity (Buckner, et al., 2008). More specifically, the DMN is now known to consistently increase neural activity during a small set of complex cognitive tasks, including the contemplation of others and ones own mind states, spatial navigation, as well as scene construction processes when envisioning past, fictitious, and future events (Spreng, et al., 2009); more generally, envisioning situations detached from reality. It was speculated that the human brain might have evolved to, by default, predict environmental events using mental imagery. Constructing detached probabilistic scenes could thus influence perception and behavior by estimating saliency and action outcomes. This would invigorate a possible relationship between the physiological baseline of the human brain and an introspective psychological baseline (Schilbach et al., 2008). In sum, the DMN routinely defies neuroscientific intuitions and challenges established methods.

Neuroimaging research on the DMN corroborated that this particular network consistently decreases activity during externally focused mental tasks and typically increases activity during a small set of internally focused mental tasks. It may reflect unfocused every-day mind wandering in form of continuous environmental tracking in a generative, integrative process. But we are not even close to certain knowledge of what this might mean en detail.

1.6 Statistical learning approaches in brain imaging

Everyday neuroimaging practice is still largely dominated by analysis approaches drawn from classical statistics. Much of the success of cognitive neuroscience since the 1990'ies has been implemented in the mass-univariate analysis of neuroimaging data using the general linear model (GLM). The GLM treats each volumetric pixel of brain scans (i.e., voxel) as independent to perform serial univariate statistics (Friston et al., 1995). Univariate approaches are recognized to be an excellent test for topographical localization of neural activity, i.e., a differential increase or decrease of neural activity in individual brain voxels.

SL approaches promise to extend this representational agenda of fMRI investigations (i.e., analysis of activation localizations) to an informational agenda (i.e., analysis of information patterns) (Kriegeskorte et al., 2006; Mur et al., 2009). SL approaches can elicit hidden quantities in neuroimaging data by providing new pieces of evidence to four questions (Brodersen, 2009; Pereira et al., 2009):

1. *Where* an information category is neurally processed? As SL techniques are inherently multivariate, the coherent patterns of BOLD signal in voxel sets are localized. This extends the interpretational spectrum from mere increase/decrease of neural activity to the existence of complex combinations of distributed activity changes.
2. *Whether* a given information category is encoded by neural activity? This extends the interpretational spectrum to topographically similar but neurally distinct processes that potentially underlie different psychological concepts.
3. *When* an information category is generated, processed, and bound? When applying SL to BOLD time series, for instance, starting from experimental stimulus onset, the interpretational spectrum is extended to the evolution of predictive performance in the time dimension.
4. *How* an information category is neurally processed? The interpretational spectrum is extended to computational properties of the neural processes, including for instance, linearity versus nonlinearity as well as local versus distributed and isolated versus partially shared computational facets.

More generally, multivariate information inference is typically more potent than mass-univariate localization inference because the latter is inherently focal and threshold-dependent (Friston 2008). The popularity and adoption of SL methods in neuroimaging has steadily increased since the attempt of "mind-reading" or "decoding" cognitive processes from neural activity patterns (Haynes and Rees, 2005; Kamitani and Tong,

2005). The conceptual appeal has been complemented by recent advances in computing power, memory resource, and the increasing trend for creating large data repositories (Poldrack and Gorgolewski, 2014).

More specially, GLM-based and SL-based analysis regimes in functional neuroimaging can be conceptualized as complementary instances of *encoding models* and *decoding models* (Naselaris et al., 2008):

$$f\colon\ y_t \rightarrow \mathbf{X_t} \tag{1}$$

$$g\colon\ \mathbf{X_t} \rightarrow y_t \tag{2}$$

where (1) represents a (voxelwise) GLM as an encoding function and (2) for instance a (brainwise) linear classifier as a decoding function, $X_t \in \mathbb{R}^d$ a 3D matrix of voxel values holding BOLD signals in brain space, $y_t \in \mathbb{N}^n$ a set of indicators of a psychological task or mental context, and $t \in \mathbb{N}$ a time series of brain scans. In a related vocabulary, the encoding function is the basis for *forward inference*, testing the probability of observing neural activity in brain regions given knowledge of the psychological process (Yarkoni et al., 2011). The decoding function, in turn, is the basis for *backward inference*, testing the probability of a psychological process being present given knowledge of neural activity in brain regions. A main difference between encoding and decoding models pertains to the direction of linearly mapping between brain space and feature space. Nevertheless, both encoding or decoding functions can be viewed as a prediction task since, for deciding on a relationship between an activity X_t and a context y at time point t, the mapping direction is irrelevant. Encoding models are superior to decoding models for establishing which processing facets are preferentially represented within brain regions. Encoding models can also be easily compared to one another, whereas inference about brain represetations according to decoding models reduces to model comparison. An important advantage of decoding models is that they lend themselves more naturally to examining the correspondence between brain activity in brain regions and indices of behavioral performance. Decoding models are also more flexible than encoding models in allowing "identification" (Kay et al., 2008), inferring a stimulus or task from a finite set based on brain activity, and "reconstruction" (Miyawaki et al., 2008; Thirion et al., 2006), restoring a stimulus or task from brain activity. In sum, the classical functional neuroimaging localization might be a weak choice to perform inference on structure-function relationships without formal modelling (Stephan, 2004), whereas decoding models are readily applicable for establishing complex structure between high-dimensional neuroimaging data and variables of interest.

Taken together, as a complementary methodological family, SL approaches are characterized by a) making the least assumptions possible, b) being more motivated by computational models rather than cognitive theory, and c) automatically mining structured knowledge from data resources. Even a small-group fMRI study qualifies as a high-dimensional statistical problem. A transition from parametric to non-parametric modeling (e.g. Russell & Norvig 2010, Ch. 18.8) and from data to algorithmic models (Breiman 2001) has been prompted by an ongoing drift towards more data-driven (i.e., fewer assumptions), higher-dimensional

(i.e., more features per observation) neuroimaging analyses on larger samples of neuroimaging data (i.e., more observations). Epiphenominal of the current clash between CS and SL techniques in the neuroimaging field, there is much controversy about how they interact in everyday research practice (e.g., whether or not cross-validated classification accuracies need to be validated by p-values). The crucial challenges in imaging neuroscience might however lie in *mechanistic interpretability and understanding* by way of generative, rather than descriminative, SL models. (Brodersen et al., 2011). An important step towards this goals is the question whether low-dimensional manifolds are embedded within the high-dimensional neuroimaging data.