Statistical Learning of Biological Structure in Human Brain Imaging

Inaugural-Dissertation

zur
Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt vor

Dr. med. Danilo Bzdok
aus Lauchhammer

Düsseldorf, September 2015

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. med. Simon Eickhoff
Koreferent: Prof. Dr. rer. nat. Stefan Conrad

Tag der mündlichen Prüfung: 25.01.2016

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

Santiago Ramón y Cajal

Peer-reviewed publications related to the present dissertation

Original papers

Bzdok D, Grisel O, Eickenberg M, Thirion B, Varoquaux G. Semi-supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data. *Advances in Neural Information Processing Systems*, 2015.

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

Bludau S*, **Bzdok D***, Gruber O, Kohn N, Riedl V, Palomero-Gallagher N, Mller V, Hoffstaedter F, Amunts K, 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, Laird AR, Fox PT, Eickhoff SB. Left inferior parietal lobe engagement in social cognition and language. Under review. *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, 81:381-92.

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 modelling. 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, in press.

Bzdok D & Eickhoff SB. The resting-state physiology of the human cerebral cortex. In: *Brain Mapping: An Encyclopedic Reference.* Oxford, 2015.

1 Introduction

1.1 Analytical and heuristic accesses to nature

The world around us is highly complex. A large proportion of human research efforts are undertaken in an *analytical* fashion expressed by "the unreasonable effectiveness of mathematics in the natural sciences" (Wigner, 1960). The author states that mathematical language is a powerful tool to describe, quantify, and predict phenomena in nature. He 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 regularities in the physical world, describe them by mathematical equations, and use these to their advantage. Starting from human-conceived axioms we have derived always more complicated properties of and relationships between mathematical objects by formal proofs (Connes, 2010). A logical pyramid of theoreoms is built that lead to always more general assertions. 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, 1992). 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 pineapple 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 imposing logical structure on the external world. It remains unclear 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, 2009). Even the most abstract mathematical concepts can be expressed by one and understood by another individual. Consequently, mathematical knowledge can be easily passed on across generations and geographical distances. One may also note that there is usually consensus among mathematicians about the architecture of their discipline. From an anthropological perspective, mathematical formalism for natural phenomena appears to be one of the most powerful tools and most defining properties of the human species (Dehaene, 2011). Indeed, Eugene Wigner concludes his praise of equations for the investigation of nature 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 [...]." (Wigner, 1960, p. 14).

However, the seeming perfection of mathematically described mechanisms in science has repeatedly been put into question formally and empirically. In formal approaches, many axiomatic systems, typically formulated by mathematics, have been shown to be incomplete and inherently contradictory because some true assertions cannot be proven (Gödel, 1931). That is, not every logical statement expressed by mathematics can be proven or rejected. This is not a weakness of mathematics itself but of the formal systems that are typically described by means of mathematical language[1]. Additionally, *Turing machines* are a tool of theoretical computer science that act on a stream of symbols by recourse to a set of known rules to emulate computer algorithms (Turing, 1936). Invention of these computer-program proxies was an important conceptual contribution to the development of modern computers. Yet, Turing's computer analogy was also instrumental to investigate the limits of formal instructions executed by later electrical computers (e.g., in form of complexity theory). Moreover, it is possible to define a real number with equidistributed digits ('Chaitin's constant Ω') that can however not be computed (Chaitin, 2006). This is also not a weakness of mathematics itself either but indicates limits of formal computation models that are described by mathematics. In sum, the seemingly exhaustive analytical systems that are enabled by the language of mathematics have been shown to be limited or incomplete in several formal approaches.

In empirical approaches "the unreasonable effectiveness of data" (Halevy et al., 2009) has challenged analytical rigor expressed by mathematical formalism alone as the best possible way to describe, quantify, and predict natural phenomena (LeCun et al., 2015; Pietsch, 2013). More concretely, three recent empirical observations might indicate a shift in discourse in some natural science branches:

1. Sophisticated, more accurate mathematical models can be outperformed by simple mathematical models that are provided with more input data. For instance, non-linear models may less well predict a complex non-linear phenomenon than simple linear models in high-noise and data-scarce scenarios.
2. Oversimplified models provided with much input data can outperform models that have been designed according to extensive domain expertise. For instance, complicated models drawn from human-made linguistic knowledge have been outperformed by linguistics-naïve order-1 hidden Markov models (HMM) combined with large text corpora.
3. There appears to be a minimal data threshold such that the data-derived models suddenly exhibit emergence properties. For instance, simple models have been much less successful in computational linguistics until a few years ago due to the much smaller text corpora.

Indeed, one prominent example for describing, quantifying, and predicting complex phenomena in a *heuristic* fashion is automatic language translation of human text and speech. Translation systems that have been built based on human-made grammar rules have perhaps never achieved satisfactory performance (Norvig, 2011). That is, analytical approaches that exploited thousands of book pages archiving linguistic domain expertise in form of deterministic grammatical, syntactical, and orthographical rules appeared insufficient

[1] Albert Einstein: "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality."

for building language models that correctly translate every-day communication between humans. Statistical machine translation was more successful by starting out based on the assumption that the 'true' rules underlying human language are not known. More specifically, probabilistic HMM are one of the frequently used heuristic approaches. The next word is predicted only by the one (order 1) or few (order n) preceding ones with equal transition probabilities (Bengio, 1999). 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 (i.e., Given the last 1 or n word[s], what are the probabilities of the next word?) can be easily *learned* from data by observing a preferably long stream of real-world language (i.e., a 'corpus'). In this way, the class of HMMs and similar models became a dominating feature of computational linguistics, while linguistic domain expertise lost always more in importance. 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 mechanistic link between a protein's *known* primary structure (i.e., 1D chain of amino acids) and its *unknown* tertiary structure (i.e., the combination of 3D foldings of the amino acid chain that subserves function). Among the ≈1,000,000 of existing proteins we only know the tertiary structure of ≈50,000 ones. The structural configuration is however necessary to identify the position of binding 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. That is, the sequence of amino acids determines the frequently unknown protein folding structure and the ensuing spectrum of protein-protein interaction in-vivo. 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 frequently unknown 3D structure as a black box. State-of-the-art neural network models 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 mechanisms 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. In the future, successful heuristic approaches will conceivably be always more successful in these and other probabilistic scenarios, including prediction of algorithmic trading in stock markets, political outcomes, optimization of advertisement strategies, and controlling self-driving cars.

In sum, humans have recently gained the capability to create heuristic algorithms that derive complex predictive models from large amounts of data. Observing a phenomenon many times might be sufficient to algorithmically extract the regularities of that phenomenon's behavior in nature. This has entailed a shift from purely mathematical treatment to giving up some human control to self-emergening mechanistic patterns (Jordan, 2015). In the absence of an analytical access, simple statistical models can thus automatically predict diverse natural phenomena depending on the quality and quantity of the available data resources.

1.2 Analytical and heuristic statistical approaches

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 methods have been developed rather independently (Breiman, 2001). They differ with regard to historical origin, mathematical foundation, and modelling goal.

The 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 to choose the model that best reflects the data. Data are then used to estimate the parameters of that prespecified model. Classical statistics has dominated research at the universities since its inception almost 100 years ago. Well known members of the CS family include for instance Student's t-test, ANOVA, and Chi-squared test. *Statistical hypothesis testing* has also been introduced at the beginning of the last century (Fisher, 1925; Neyman and Pearson, 1928). The same approach is still practiced today (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). Similarly, likelihood-ratio tests allow comparison of a null model and an alternative model to derive p-values of significantly different model likelihood given the data. Despite the prevailing presence of p-values in contemporary research practice, it was not 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. First, CS 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. Second, it suited perfectly the Popperian view of critical empiricism in academic discourse (Popper, 1934): 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 small samples and very few variables that can be grasped by plausible models with a small number of parameters chosen by the investigator in an analytical fashion.

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 (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 et al., 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 witnessed by the most basic terminology. In the CS literature, inputs to statistical modelling 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; Kuhn and Johnson, 2013). When evaluating whether a certain problem is a possible target for SL three requirements come into play (Abu-Mostafa et al., 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 existing analytical approaches probably create a better model).
3. We have data on the problem (the more, the better).

Since 1980 this regime led to a surge of new computer-intense statistical techniques that can be difficult to compute on a normal calculator and that are less concerned with making mathematically convenient assumptions to increase computational tractability (Efron and Tibshirani, 1991). This development has been flanked by changing properties of datasets that are always higher-dimensional (i.e., *wide data*, more features per observation) and based on larger samples (i.e., *long data*, more observations per dataset). The important implication is that the asymptotic guarantees in CS, typically related to the law of large numbers or the central limit theorem, do not provide mathematical backing of statistical methods applied in the high-dimensional regime (Giraud, 2014). Instead, SL provides often algorithmic solutions and theoretical justifications by incorporating additional structural constraints. This is well illustrated by the *bet on sparsity* (Hastie et al., 2015): The investigator should always expect only a subset of the features to be relevant because no existing statistical methods performs well in the dense scenario that assumes the entirety of the feature is relevant in the true model. The optimality of such structural constraint in model selection has very recently been shown by mathematical proofs. Generally, 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 (Bauer et al., 2015) and economic predictions (Manyika et al., 2011). In sum, statistical learning was mostly fashioned for problems with many variables in potentially large samples with mostly unknown data generating processes that are emulated by a mathematical function created from data by a machine in a heuristic fashion.

It is very important to appreciate that some statistical methods cannot be easily categorized by the CS-SL distinction. Although the two families of statistical methods can be easily conceptualized by a number of archetypical properties, statistical methods span a continuum between the two poles of CS and SL (Jordan, 2015, p. 61). For instance, models are often not exclusively deterministic because they incorporate a component that accounts for unexpected, noisy variation in the data. Further, the CS-SL distinction is orthogonal to the Bayesian-Frequentist difference and can be adopted in both these flavors (Efron, 2005). *Bayesian* statistics are optimistic in considering the data at hand conditioned on preselected distributional assumptions for the probabilistic model to perform direct inference based on the computed posterior distributions. *Frequentist* statistics are pessimistic in considering a distribution of possible data treating model parameters as unknown (i.e., pessimistic approach) to perform indirect inference by the differences between observation and model-derived data. Neither can *univariate* versus *multivariate* statistics relying on one versus several input variables be clearly grouped into either CS or SL. Statistical methods can also be distinguished into *parametric* and *non-parametric* groups (Bishop, 2006). Parametric approaches make explicit distributional assumptions (effectively acting as a type of filter or compression) and have a constant number of model parameters, whereas non-parametric approaches avoid predetermining its model structure (operating on the full data) and the model parameters increase explicitly or implicitly as a function of dataset size[2]. More interpretable parametric tests typically have higher power (i.e., require smaller datasets) and are less interpretable non-parametric tests typically incur higher computational load (i.e., due to dynamically increasing model complexity). Parametric models are exemplified by support vector machines, Lasso regression, and principal component analysis, while non-parametric models are exemplified by k-nearest neighbors, neural networks, and hierarchical clustering. Moreover, statistical models can be *discriminative* or *generative* (Ng and Jordan, 2003). Discriminative models focus on solving a supervised problem of predicting a class y by directly estimating $P(y|X)$ without capturing special structure in the data X. Typically more demanding in data quantity and computation resources, generative models estimate special structure by $P(y|X)$ from $P(X|y)$ and $P(y)$ and can thus produce synthetical examples $\tilde{X}$ for each class y. In general, neither CS nor SL can generally be considered superior. This is emphasized by the *no free lunch theorem* stating that no single statistical strategy can consistently do better in all circumstances (Wolpert, 1996). The investigator

[2] Non-parametric approaches are thus naturally prepared to capture emergence properties in larger dataset sizes (Jordan, 2015).

is challenged to choose the statistical approach that is best suited to the phenomenon under study and the research object at hand.

Regarding modelling goals, CS and SL exhibit various differences. CS typically aims at modelling 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 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 in imposing mathematical rigor on the phenomenon, whereas SL tends to be heuristic in 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). 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 prespecified 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 modelling by imposing an a priori model in a top-down manner, whereas SL implements algorithmic modelling 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 (Wigner, 1960) in a CS-contrained regime, while it is believed to be in the data (Halevy et al., 2009) in a SL-constrained regime.

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; Friston, 2010; Dayan et al., 1995; Gregory, 1980). Moreover, each probabilistc model can be viewed as a superclass of a deterministic model (Norvig, 2011). 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 curse of dimensionality

The increasing quantities of data to analyze put many research domains, including brain imaging, to the test (Poldrack and Gorgolewski, 2014). Today's functional neuroimaging methods offer very high resolution in space (especially fMRI and PET) and time (especially EEG and MEG) (Amunts et al., 2013; Buzsáki and Draguhn, 2004). Yet, the 'effective dimensions' (Jordan, 2015) of neuroimaging data remain elusive. The mere number of features per observation poses serious statistical challenges to the investigator. It is the neuroscientific version of what is generally referred to as *curse of dimensionality* (Bellman, 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 many variables.

The more dimensions an input space spans, the further the data points are away from each other (Hastie et al., 2011). The closest and furthest neighbors of a given point become always closer (Beyer et al., 1999). Counter-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. As the number of dimensions approaches infinity the volume of the hypersphere gets close to 0 and hence becomes irrelevant comparing to the hypercube volume. The probability mass of the hypercube will increase in its corners and decrease in its center. 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, 2006). Additionally, the target function is almost always unknown in statistical learning investigations. Hence, knowledge is frequently unavailable as to whether or not special structure exists in the input data that can be exploited. Knowledge of special structure of the phenomenon under study can simultaneously reduce *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 modelling 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, 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, statistical estimators 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 (Bach, 2014; 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 represents a hypothetical solution to the estimation problem) is hopefully reduced to a reasonable preselection (*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 tradeoff, 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 independent of 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. A 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 concerned with complexity restrictions to neurobiological reality as observed by fMRI scanning.

1.4 The human brain as a complex phenomenon

The human brain is a prime example of a black box that is complex, mysterious, and perhaps in part unknowable. It has repeatedly been proposed that the human brain might be the most complex object in the known universe (Kandel et al., 2013). With the language from above, the human brain might constitute a phenomon in nature that can perhaps *not* be perfectly grasped by purely analytical models expressed by mathematical formalism. More concretely, the *most pertinent structure* that should be assumed for the human brain, when measured by contemporary functional neuroimaging techniques (cf. next passages), is currently unknown. Hence, the brain imaging access to neuroscience can readily be framed as a problem of *representation learning* (Bengio et al., 2013). It is conceivable that this task can be solved without exhaustive neurobiological micro-/meso-/macro-level knowledge (Sandberg and Bostrom, 2008). This is always more supported by empirical evidence (Helmstaedter et al., 2013) and it is a contention that is embraced by the present dissertation.

Nature versus nurture. From a global perspective, the genetic difference between our genetic endowment 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 (Durham, 1991; Tomasello, 2009). That is, the human brain might be much more defined by the increasingly fast cultural evolution rather than 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 etc. pp. This form of learning is a very plausible and decisive property of intact tissue of the central nervous system. As a first challenge in neuroscience, 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 (Skinner, 1976; Dehaene and 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 optimization process (Vygotsky, 1978; Luhmann, 1984; Bengio, 2014).

Neural reuse. Given the acceleration in human cultural evolution (Virilio, 1997), 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, such as 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* or *neural recycling hypotheses* (Anderson, 2010; Dehaene and Cohen, 2007). Non-human primates are lacking many of the sophisticated mental operations that are crucially important for maintaining human societies (Mesulam, 1998; Tomasello et al., 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, 1978; Byrne and Whiten, 1988; Dunbar and Shultz, 2007). Yet, it is becoming increasingly clear that socio-affective processing in the human brain is probably realized by domain-general brain regions and networks that are not specific to processing social relationships (Bzdok et al., 2015; Behrens et al., 2009; Barrett and Satpute, 2013). These considerations entail a second challenge in neuroscience: It is probably impossible to know what purpose processing in a given neuronal population has originally evolved to serve. All that can be observed are external manifestations and correlative relationships of this latent biological purpose.

Inter-individual variability. 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 the cortical areas, as well as the history of their sensory inputs. The extent of a cortical 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 inter-individual differences in many (temporally transient) states and (temporally enduring) traits (Phelps and LeDoux, 2005; Sallet et al., 2011; Bickart et al., 2011). As a third challenge in neuroscience, it is currently unknown how inter-indivindividual differences in behavioral facets are mediated on the brain-level[3]. 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 and Rees, 2011). That is, there is no clear understanding of how this set of microstructures interact to solve neural computation problems, let alone their inter-individual differences. From a methodological perspective, volumetric modelling techniques conventionally employed in the neuroimaging field are naïve to many types of possible morphological differences between individuals. 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 in an individual (Ashburner and Klöppel, 2011).

Hemispheric asymmetry. Worth to be proposed as an independent challenge of neuroscience, the neurobiological purpose of inter-hemispheric 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). 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 and Shulman, 2002; Stephan et al., 2003; Price, 2010).

Observational versus interventional investigations. It is further unlikely that we will reach exhaustive understanding of the human brain by mere *observational*, as opposed to *interventional*, research methods [4] (Pearl, 2009). Purposely induced local 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, 1951), 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

[3] Santiago Ramón y Cajal wrote: '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.' (Ramón y Cajal, 1909/1911)

[4] This idea is reflected in Edward O. Wilson's words "Disturb Nature and see if she reveals a secret." (Wilson, 2013) as well as in G. M. Shepherd's words "Nothing in neuroscience makes sense except in the light of behavior." (Shepard, 1988)

from lesion findings constitutes yet another challenge to neuroscience. 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 et al., 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 limited by the existing variability 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 local, as opposed to distributed, differences in brain structure (Goodkind et al., 2015). More generally, it is still a matter of debate whether structure (i.e., locally specific micro- and chemoarchitecture), connectivity (i.e., short- and long-range 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).

The binding problem. 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 (Buzsáki et al., 2013). Perhaps since early reports of increasingly complex processing of neurons in the primary visual cortex (Hubel and Wiesel, 1965), neuroscientists tend to think information processing in the brain 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 the next processing step. Yet, brain oscillations are capable of predicting when this next processing step will occur. Oscillation measured by EEG and MEG techniques might be the currently most attractive access to the binding problem (Singer, 1999; Engel and Singer, 2001; Varela et al., 2001): There is very limited understanding of how environmental perturbation by multi-sensory stimulation is 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 (Buzsáki, 2006). Additionally, different oscillatory frequency bands flank each other in a same brain area in an interacting fashion (Le Van Quyen, 2011; Canolty et al., 2006). 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, low frequencies govern large-scale networks in the brain that, in turn, influence small local neuronal spaces with high-frequency oscillatory patterns.

Interpretational categories. From a philosophical perspective, the neuroscientist also 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 neuroscience, 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). The grammatical and lexical constraints of human language might be too tight to allow for unequivocal description of the diverse circumstances researchers encounter in scientific investigations. The meaning of language may be primarily defined by its practical use in concrete situations, rather than decontextualized abstractions necessarily shared by other individuals (Whorf, 1956). Words might not have an objective meaning equally accessible to and understood by everybody[5]. 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 (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, 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 repeatedly commited 'the error of the third kind' (Kimball, 1957): providing an accurate answer to an inadequate research question[6]. 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

[5] From a more sociological persepctive, also specialisation of human individuals alters consciousness (Habermas, 1981)

[6] 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.

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 plausible, that is, neurobiologically valid.

Epistemological limits. The last challenges to the neuroscientist mentioned in this dissertation introduction are of epistemological origin. Biology as a whole has a modest legacy in abstract theory (von Bertalanffy, 1968). This probably includes research on brain biology. In particular, the spectrum of permissible conclusions that can be drawn from neuroscientific investigations is strongly conditioned by each of the following questions (Carruthers, 2009; Dehaene, 2007; Bostrom, 2013):

1. Does the human brain offer sufficient computational resources to grasp, formalize, and predict itself?
2. Is the human mind capable to reflect on itself by directly contemplating itself via introspection or by indirectly contemplating an internalized self-model acquired through interaction with other individuals?
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?
4. How does the spectrum of existing ways to record data from brain systems inherently constrain the spectrum of possible scientific inferences?

Consequences. Taken together, there are many intricacies about neurobiology and the mosaic knowledge that we currently have about it. Despite ≈200 years of neuroscientific research, neuroscience is probably not even close to something like a unified theory of brain function (Friston, 2010; Bar, 2007; Kelso et al., 2013) that neuroscientists from different fields would be happy to consider. This caveat considerably complicates the formulation of precise, neurobiologically valid hypotheses that can be formally tested in targeted experimental studies. Therefore, it might be helpful to use heuristics-establishing statistical approaches for pattern discovery and inference, instead of having recourse to classical statistics alone. Deriving new knowledge about the brain exclusively by successive falsification of entirely human-conceived, intimately language-dependent, and dichotomically framed hypotheses might be viewed as hubris by some (Feyerabend, 1975; Cohen, 1994). The projects decribed in this dissertation are hence based on the assumption that an *exhaustive analytical understanding* of the brain might not be reached 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 many other domains, including language translation and drug discovery (cf. 1.1).

1.5 Functional magnetic resonance imaging (fMRI)

In the nineteenth century functional specialization in the Cortex cerebri of humans has been investigated predominantly by reports of circumscribed tissue damage (Harlow, 1848; Broca, 1865; Wernicke, 1881). Brain lesion studies together with brain stimulation during surgery (Penfield and Perot, 1963) and pharmacological intervention (Clark et al., 1970) were the mainstay of neuroscientific research for a long time, until they were complemented by axonal tracing studies for connectivity analysis in animals (Mesulam, 1976). Today, fMRI techniques are 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, (Smith et al., 2001)). fMRI enables the localization of activity changes in neuronal populations by means of measuring the accompanying changes in the oxy-to-deoxyhaemoglobin ratio in local draining veins (Roy and Sherrington, 1890; Ogawa et al., 1990). 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 and Raichle, 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 processing output. There is indeed 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, without being able to disentable excitatory and inhibitory neural activity (Logothetis and Wandell, 2004). As a central property of fMRI acquisitions, there is a tradeoff between coverage of sampled brain tissue, spatial and temporal resolution. For instance, augmenting the spatial resolution, while keeping brain coverage constant, deteriorates the temporal resolution. Finally, the regional responses in single individuals are routinely transformed into a standard brain space (i.e., 'spatial normalization' into the 'Talairach-Tournoux' or 'Montreal Neurological Institute' coordinate systems) for comparability and statistical analysis at the group-level (Talairach and Tournoux, 1988; Evans et al., 1992).

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 et al., 1999b). This is remarkable because fMRI typically tests differences in non-linear phenomena using linear methods. 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 restricted head movements. In this way, experimental 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 processing 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 measurement 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 ($\approx$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 probably 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 might not have been discovered without the advent of neuroimaging methods is the so-called *default mode network* (DMN) (Buckner et al., 2008). This particular resting-state network is a pure result of serendipity and is a focus of the present dissertation. 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 (Shulman et al., 1997). That is, the DMN was believed to increase neural activity in the idling, unconstrained mind and decrease activity during stimulus-driven, goal-directed tasks (Raichle 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 (Reivich et al., 1979; Raichle et al., 2001).

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 neuroscience is not even close to certain knowledge of what this might mean in 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., 1994). 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 predominantly 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 possibly elicit hidden quantities in neuroimaging data by providing 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 BOLD signal patterns in voxel sets can be localized. This extends the interpretational spectrum from mere increase/decrease of neural activity to the existence of complex combinations of activity changes distributed across voxels.
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 the interpretational spectrum is extended to the beginning, evolution, and end of distinct neural processes.
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 non-linearity as well as local versus distributed and isolated versus partially shared processing facets.

More generally, multivariate information inference is typically more potent than mass-univariate localization inference because the latter is inherently local and threshold-dependent (Friston et al., 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 resources, and the increasing trend for creating large data repositories (Poldrack and Gorgolewski, 2014).

More specifically, GLM-based and SL-based regimes in functional neuroimaging analysis can be conceptualized as complementary instances of *encoding models* and *decoding models* (Naselaris et al., 2011; Pedregosa et al., 2015):

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

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

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 linear mapping between brain space and feature space. Nevertheless, both encoding or decoding functions can be viewed as a prediction task since the mapping direction is irrelevant for deciding on a relationship between an activity X_t and a context y at time point t. 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 representations 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' (Brodersen, 2009), 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, the neuroimaging field is currently going through a transition from using parametric to always more non-parametric models (Russell and Norvig, 2010) and from data modelling to algorithmic modelling (Breiman, 2001). This paradigm shift has been prompted by an increasing drift towards higher-dimensional datasets (more features) that are more frequently collected in a collaborative fashion (more sites) with measurements from always more participants (more observations) (Kandel et al., 2013; Anonymous, 2014; Poldrack and Gorgolewski, 2014; Frackowiak and Markram, 2015). SL approaches could therefore be a useful alternative toolbox to complement methods derived from classical statistics. SL methods are typically 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 extensive neuroimaging data resources. The crucial challenges in imaging neuroscience might lie in enabling *mecha-*

nistic interpretability and understanding by models with generative, rather than descriminative, properties (Brodersen et al., 2011). An important step towards this goals is the question whether low-dimensional manifolds are embedded within the high-dimensional fMRI data.

2 Unsupervised modelling of brain regions

2.1 Motivation

The human brain is commonly assumed to be organized in distinct modules (Brodmann, 1909; Vogt and Vogt, 1919). These can be described according to structure, connectivity, and function. Cortical areas can be conceptualized as patches of the brain that differ from their neighbors in terms of their microarchitecture (e.g., cyto-, myelo- and receptorarchitecture), connectivity (i.e., set of input and output connections), and function (e.g., lesion-induced behavior or electrophysiological responses) (Van Essen, 1985; Felleman and Van Essen, 1991). The conjunction of *i)* input and output connectivity of a cortical area and *ii)* its local infrastructure is thought to crucially determine what classes of computational problems a cortical area can solve.

The correspondence between a cortical area and its characteristic axonal connectivity has prompted the class of connectivity-based parcellation (CBP) methods in neuroscience (Behrens et al., 2003; Wiegell et al., 2003). Capitalizing on the distinct connections of each area (Passingham et al., 2002), CBP divides a region of interest (ROI) into distinct subregions. Put differently, CBP performs a systematic summary of the Cortex cerebri by combining same neural tissue and separating different neural tissue according to an organizational criterion, namely, brain connectivity. The key idea is to first compute an individual connectivity profile for each individual voxel in the ROI. The voxel-wise connectivity profiles are then used to group the ROI voxels such that connectivity is similar for the voxels within a group and different between groups. That is, distinct clusters are identified in the ROI by long-range interaction patterns of the voxels in the ROI. On the one hand, previous investigations have demonstrated that CBP can reveal clusters that recover known microscopical boundaries of brain tissue (Behrens and Johansen-Berg, 2005). On the other hand, there are also reports showing that CBP may yield more fine-grained subdivisions than classical cytoarchitectonic mapping (Clos et al., 2013). Hence, CBP-derived brain modules may be viewed as 'functional areas', although these are outlined by connectivity clusters rather than functional clusters per se (Amunts et al., 2014).

Clustering uses a similarity measure to group a set of elements into subsets according to their measured similarity (Jain et al., 1999; Handl et al., 2005). In CBP, the clustering algorithm groups the voxels in the ROI into subsets according to similarity of their connectivity profiles, the heart of any CBP approach. As a result of the so-called 'no free lunch' theorem (Wolpert, 1996), no clustering algorithm performs optimally for all ROIs, types of connectivity information, and study motivations. In neuroimaging research, the k-means algorithm is likely to be the most popular clustering choice to divide a ROI into a preselected number of k non-overlapping subregions (Nanetti et al., 2009), although spectral and hierarchical clustering have also been used repeatedly.

K-means clustering (Lloyd, 1957; Forgy, 1965; Jain, 2010) depends on free parameters, including *i)* the cluster initialization, *ii)* the cluster number k, and *iii)* the tolerance for iteration stopping. Initially k voxels in the ROI are randomly chosen to represent the centers of the k desired clusters. Two steps are then iterated multiple times. First, the ROI voxels are assigned to the closest cluster center (i.e., 'centroid'), which equates with partitioning the ROI into k clusters. The the optimal parameter k is typically chosen by heuristic cluster validity criteria (cf. 2.3) Second, the k cluster centers are recomputed. As soon as the center needs to be shifted by less than the preset distance threshold (early stopping is often needed for the sake of time), the iterative process stops. Note that the final assignments of ROI voxels to particular clusters may vary with different cluster initializations and yield non-optimal solutions at local minima. Technically, this is because k-means represents a non-convex optimization problem. Neurobiologically, the same connectivity data may thus result in several stable solutions for the ROI parcellation at the same preset k (i.e., low reproducibility), as shown, for instance, in the human insula (Nanetti et al., 2009). Consequently, the algorithm is usually repeated many times.

In neuroimaging practice, k-means seems to perform best when the subregions in the ROI are expected to be *i)* few in number, *ii)* of similar size, and *iii)* featuring a roughly spherical shape on spatially correlated voxel-wise connectivity (Jain, 2010). Additionally, k-means clustering will converge in the majority of the cases (i.e., seldom early stopping by the tolerance parameter). In a CBP context, the same connectivity data can describe not only one but several stable solutions in ROI parcellation at the same preset k (i.e., low reproducibility), such as observed in the human insula (Nanetti et al., 2009). Consequently, the algorithm is conventionally applied many times since k-means fits idiosyncracies in data that may generalize poorly across participants. As a first practical consequence, the initialization of the cluster centers can be random (Hartigan and Wong, 1979) or based on prior knowledge (e.g., anatomical properties). As a second practical consequence, the final solution can be obtained by an averaging procedure or by selecting the centroids from the best solution (Nanetti et al., 2009). Further, the solutions for different selections of k (i.e., different number of clusters) are independent of each other. Repeating the clustering at different k's does not emulate a coherent hierarchy. That is, the solutions for ROI parcellation at each level (k) are independent of each other, which makes parent-children stratifications possible but not necessary[7].

Importantly, performing brain parcellation by CBP is not bound to a particular connectivity approach. Any method to measure brain connectivity can be employed that yields a connectivity profile for each voxel in the ROI. The functional connectivity modality in most frequent use is RSFC (Zhang and Raichle, 2010). It has recently been complemented by MACM (Robinson et al., 2010; Eickhoff et al., 2011). These connectivity measures reflect drastically different ways to conceive and quantify interneuronal communication between brain regions. Choosing one of them is as important as the ROI to investigate and has far-reaching implications for permissible conclusions from the identified clusters.

MACM operates under the assumption that functional connectivity between brain regions should entail

[7] This is contrary to divisive top-down or agglomerative bottom-up hierarchical clustering.

reliable coactivation (Robinson et al., 2010). It quantifies temporally related increase/decrease of neural activity in distant brain regions throughout various experimental tasks. This connectivity modality capitalizes on the increasing trend for large-scale data aggregation, exemplified by the BrainMap (Fox and Lancaster, 2002), Neurosynth (Yarkoni et al., 2011), and NeuroVault (Gorgolewski et al., 2014) neuroimaging databases. Caveats of MACM include *i)* reliance on very sparse activation information (i.e., peak coordinates of significant activation), which might entail missing information and biased sampling, *ii)* inability of participant-specific connectivity analysis, and *iii)* inheritance of the limitations from experimental neuroimaging studies. In spite of these limitations, the analysis of coactivation likelihoods represent an approach that is complementary to RSFC by focusing on the interactions during the performance of externally imposed tasks.

In contrast, functional connectivity can be measured in the human brain by RSFC correlations under the assumption that the coupling strengths between distant brain regions is measurable by correlation between time series of BOLD signal fluctuations outside of an experimental context (Biswal et al., 1995; Zhang and Raichle, 2010). It quantifies the correlative relationships between distant brain regions in participants idling in the MRI scanner. This is possible because interneuronal communication continues and is reflected by ongoing physiological fluctuations in the absence of an experimentally imposed cognitive set, i.e., during natural mind wandering, which can be measured using fMRI. RSFC signals have been shown to recover axonal connections and functional networks well-documented in monkey tracing studies and task-based neuroimaging studies. Despite initial skepticism, the consistency of RSFC results has been demonstrated repeatedly across participants, brain scans, time points, and other factors (Damoiseaux et al., 2006; Shehzad et al., 2009). RSFC thus provides proxies of dynamic neuronal interactions that might reflect mixtures of various cognitive processes and physiological factors (Smith et al., 2009; Mennes et al., 2013).

Task-related MACM and task-unrelated RSFC combined with CBP is an ideal set of tools to investigate the challenging task-rest behavior of the default mode network. The right and left temporo-parietal junctions (RTPJ/LTPJ), posteromedial cortex (PMC), and dorsomedial prefrontal cortex (dmPFC) have so far almost exclusively been studied as a cohesive brain network. There is a scarcity of targeted studies on functional segregation *within* this brain network As a first exception, the medial prefrontal cortex and the posterior cingulate cortex have been identified as network cores by clustering and graph-analytic analyses (Andrews-Hanna et al., 2010). As a second exception, an fMRI study revealed stronger recruitment of a so-called 'dorsal subnetwork' (dorsomedial prefrontal cortex and posterior cingulate cortex) during retrieval of emotional (but not neutral) autobiographical memories, whereas a resting condition without controlled task recruited a so-called 'ventral subnetwork' (ventral striatum, ventromedial prefrontal cortex extending into the subgenual anterior cingulate cortex) (Bado et al., 2014). These pieces of tentative evidence provide support for functional heterogeneity within parts of the DMN.

The human RTPJ of the DMN is a supramodal association area located at the border between the temporal and parietal lobes surrounding the posterior end of the Sylvian fissure. It is at times referred to as 'posterior inferior parietal lobule', 'Brodmann area 39', or 'posterior superior temporal sulcus'. The highly inconsistent neuroanatomical labeling epitomizes the lacking consensus on coordinates, micro- or macroanatomical landmarks that would topographically define the RTPJ (Déjerine, 1895; Mars et al., 2012). Put differently, 'temporo-parietal junction' is a vaguely defined term that is frequently used within various cognitive disciplines to refer to a certain *functional* (rather than anatomical) cortical module. An extensive body of work implies selectivity of the RTPJ for attentional processes, whereas a similarly extensive body of literature claims selectivity of the RTPJ for the higher-level processing of social information. That is, one line of research provides converging evidence for a key role of the RTPJ in attentional processes, whereas another line of research associates this region with social-cognitive processes.

A similar controversy is apparent in the human LTPJ. High-level social cognition tasks, on the one hand, typically modulate neural activity in the bilateral temporo-parietal junction of the parietal lobe but also the medial prefrontal cortex, and posteriomedial cortex. Language tasks, on the other hand, typically engage the inferior frontal gyrus, posterior superior temporal gyrus, and also the angular gyrus and supramarginal gyrus of the left parietal lobe. Hence, the LTPJ is probably commonly recruited in social cognition (Mar, 2011) and language (Binder et al., 2009) tasks. Yet, the neural correlates common to social cognition and language are currently underresearched. It is unclear whether both psychological functions engage the same anatomical subregions of the LTPJ. Similar to organizational questions about the RTPJ, it is hence open to debate whether different subregions in the LTPJ contribute to different (sub)processes underlying social cognition, language, and various other high-level cognitive processes.

Additionally, the human posteromedial cortex (PMC) has been functionally implicated in attention, language, and social cognition, as well as many other diverse psychological tasks. Perhaps due to the PMC's mosaic anatomical organization, attempts of global functional accounts range from covert reallocation of spatial attention (Gitelman et al., 1999), mediation between internal and external focus (Leech and Sharp, 2014), computation of environmental statistics (Pearson et al., 2009), and self-referential visuospatial imagery (Cavanna and Trimble, 2006) to modality-independent integration between emotional states and memories (Maddock, 1999). These proposed domain-spanning roles potentially explain its various domain-specific functional involvements potentially implemented in distinct functional modules of the PMC.

Finally, also distinct parts of the dmPFC are recruited during diverging high-level cognitive functions. The dmPFC is another associative brain region of the DMN, raising increasing suspicion to contain functionally heterogeneous subregions (Amodio and Frith, 2006; Gilbert et al., 2010). More generally, the anatomical subdivisions of the entire human medial prefrontal cortex (mPFC) were first mapped out by Brodmanns cytoarchitectonic studies (Brodmann, 1909). The cytoarchitectonical subdivision of the medial prefrontal cortex into BA 11 (orbitofrontal cortex), BA 10 (ventral mPFC and most of frontal pole), and BA 9 (dorsal mPFC and small part of frontal pole) might however not exhaustively capture regional functional heterogeneity in that region. As a rare example in fMRI research, topographically distinct parts of the human dmPFC have been shown to be recruited in social, emotional, memory, and attentional tasks from independent neuroimaging studies (Gilbert et al., 2010).

Consequently, the RTPJ, LTPJ, PMC, and dmPFC have diverging task-related functions and, thus perhaps, mosaic cytoarchitectures. Nevertheless, they have each been robustly implicated in the DMN (Buckner et al., 2008). Yet, the precise nature of neural processes realized in the DMN remains as elusive as its neurobiological organization. Viewed from the perspective of the curse of dimensionality (cf. introduction), adressing the functional heterogeneity of nodes in the DMN by means of clustering algorithms makes the following assumptions:

1. Neural activity in the DMN can be explained by a small set of local patterns (i.e., compositionality).
2. These can be described while ignoring extension of neural activity across various discrete local activity patterns (i.e., known cytoarchitectonic brain regions).
3. The latent clusters within DMN nodes derived from fMRI data describe distinct properties of underlying neurobiology.
4. The local patterns can be approximated by estimators that model the membership of node voxels to k latent components based on features of whole-brain connectivity.

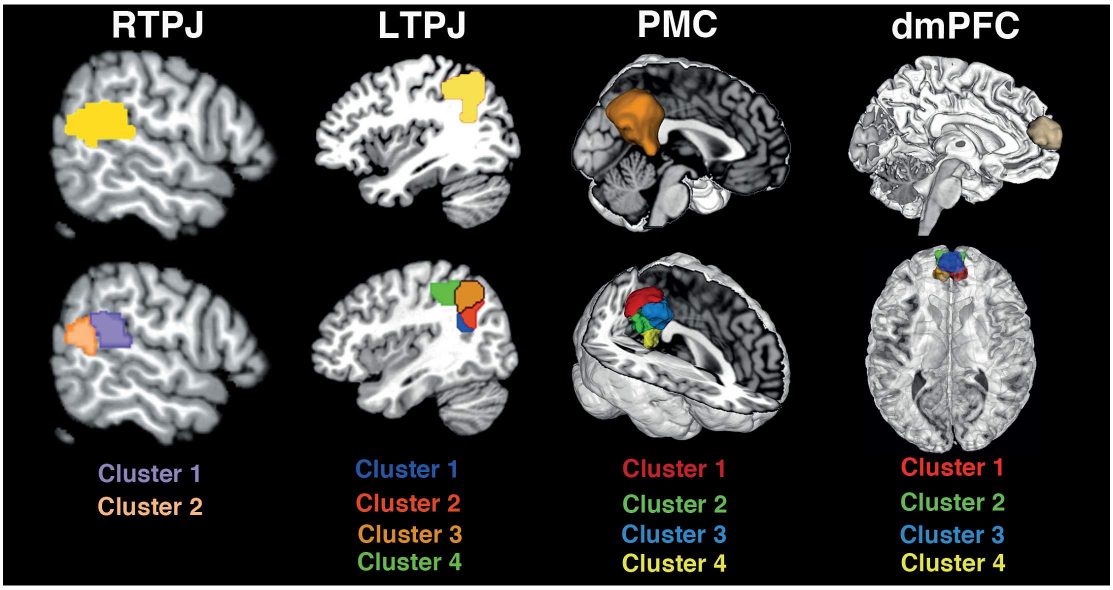

Figure 1: **Brain volumes of interest and connectivity-based parcellation results**

Top row: The topographic definitions of the regions of interests in the default mode network. *Bottom row:* Results of the best fitting clustering solution were projected onto the 3D brain surface.

2.2 Methodological approach

Seed region definition. The first decision in CBP analyses concerns the part of the brain to analyze. The ROI outlines the set of target gray-matter voxels that is to be automatically segregated into subregions. This important step operationalizes the investigation, which strongly impacts the overall outcome and interpretation. It is usually advisable to consolidate the location of the functional process of interest by means of image-based (Dehaene et al., 2003) or coordinate-based (Eickhoff et al., 2009) quantitative meta-analysis. The resulting meta-analytic ROI then statistically constrains the most robust location of activation convergence underlying the process of interest across various participants, study designs, and laboratories. CBP then explores the heterogeneity of connectivity patterns within this functionally defined region. Such composite functional ROIs allow extension of the interpretational space by asking: Are the different cognitive processes reflected by the different activations related to the same or different connectivity-defined modules in the human brain?

To both accommodate lacking neuroanatomical consensus and acknowledge the diverse functions ascribed to the RTPJ[8], the volume of interest for CBP (Fig. 1) was constructed by merging results of three meta-analyses of neuroimaging data on sustained attention (Langner and Eickhoff, 2012), sensorimotor control (Jakobs et al., 2012), and theory of mind (Bzdok et al., 2012b). This is because of the absence of commonly accepted neuroanatomical landmarks to define the location of this functional region. The first, most anterior, cluster resulted from a quantitative meta-analysis of neuroimaging experiments on sustained attention (Langner and Eickhoff, 2012), that is the capacity to stay focused on a particular task for extended time periods. A second cluster resulted from a quantitative meta-analysis of neuroimaging studies on sensorimotor control (Jakobs et al., 2012), that is the capacity to integrate exogenous stimuli and contextual information for behavioral response formation. The third, most posterior, cluster resulted from an ALE meta-analysis of neuroimaging experiments on theory of mind (Bzdok et al., 2012b), that is the capacity to model others thoughts, beliefs, and behavioral dispositions by abstract inference. Merging these activation clusters yielded a composite seed region that captures the various functional roles conventionally ascribed to the RTPJ in the neuroscientific literature.

In analogous fashion, the LTPJ seed region was constructed from coordinate-based meta-analysis of convergent neural activity in the left lateral parietal lobe during social and language task performance. High-level social processing was represented by a previous meta-analysis on theory of mind experiments (Bzdok et al., 2012b). General language processing was represented by a present meta-analysis on all language-associated terms (i.e., orthography, speech, syntax, semantics, and phonology) from the BrainMap database. The topographically converged activation in the lateral parietal cortex was then extracted from each meta-analysis

[8] Functionally-, rather than strictly neuroanatomically-, motivated terms were preferred for the seed regions, given that neural activation associated with the target cognitive processes routinely exceeds traditional macroscopical landmarks.

and merged into a composite region. After voxel-wise dilation and erosion, this ROI constrained subsequent CBP analyses.

The ROI comprising the PMC, in turn, was defined using neuroanatomical landmarks. Cytoarchitectonic information provided the superior borders, while macroanatomical structures of the standardized brain template guided the delineation of most other borders. Regarding the superior borders of the ROI, topographical information was provided by histological probability maps from the Jülich brain atlas (Zilles and Amunts, 2010). Based on regionally specific appearance of cortical layers, cell density, and cell types, the human cortex can be divided into a large number of cytoarchitectonically distinct brain regions. This structural segregation is an important indicator of functional heterogeneity. More specifically, the posterosuperior extent of the ROI was limited by the borders of the cytoarchitectonic areas 5M, 7A, and 7P (green line) (Scheperjans et al., 2008). Regarding the anterosuperior border, the ROI was drawn such as to border the dysgranular area 23d, as indicated by neuroanatomists (Vogt et al., 2006). For the remaining ROI borders, obvious macroanatomical structures served as topographical landmarks, including the splenium of the corpus callosum defining the rostral ROI border and the parietooccipital sulcus defining the ventral ROI border. This neuroanatomically defined region of interest, including the gray-matter of the precuneus, posterior cingulate cortex, and retrosplenial cortex within the PMC, served as the basis for subsequent CBP analyses.

Finally, the dmPFC ROI was formed by combining contrast analyses from two prior neuroimaging studies. Both used fMRI to compare brain activity underlying complex social judgments (trustworthiness and attractiveness) with emotional (happiness) and cognitive (age) judgments. In the first study (Bzdok et al., 2012a) visually presented facial stimuli were evaluated, whereas auditorily presented vocal stimuli of everyday sentences were evaluated in the second study (Hensel et al., 2015). In both studies, brain activity was found that was exclusively related to social judgments by parcelling out brain activity shared with emotional and cognitive judgments. Testing for topographical convergence between these two independent contrast analyses revealed the dorsomedial prefrontal cortex as the only region that featured specific brain activity related to complex social judgments congruently across visual and auditory stimuli. Subsequent to sagittal mirroring for symmetry, this ROI reflecting the part of the dmPFC that was specifically recruited by complex judgments provided the basis for subsequent CBP analyses.

Workflow. First, MACM was used to determine the coactivation profile of each voxel within the ROI. The seed voxels were then grouped based on similarities of their coactivation profiles by k-means clustering. Second, the most stable clustering solution was identified by the combination of different cluster stability metrics. Third, the whole-brain connectivity patterns of each derived cluster in the ROI were determined based on MACM and RSFC. These steps incorporated a data-guided framework to comprehensively characterize the four nodes of the default mode network.

Task-dependent functional connectivity (MACM). Delineation of whole-brain coactivation maps for each voxel of the ROI was performed based on the BrainMap database (Fox and Lancaster, 2002). The analysis will be limited to functional neuroimaging studies in the healthy human brain (no interventions, no group comparisons), which reported results as coordinates in stereotaxic standard space. These inclusion criteria yielded ≈8,000 eligible experiments at the time of analysis. Please note that all eligible BrainMap experiments have been considered because any pre-selection based on taxonomic categories would have constituted a strong a priori hypothesis about how brain networks are organized. However, it remains elusive how well psychological constructs, such as emotion and cognition, map on regional brain responses (Poldrack, 2006; Mesulam, 1998; Laird et al., 2009). The rationale of the coactivation analysis is to compute the convergence across (all foci of) those BrainMap experiments where the seed voxel in question is reported as active. One challenge in constructing voxel-wise coactivation maps is the limited number of experiments activating precisely at any particular seed voxel. Hence, pooling across the close spatial neighborhood has become the dominant approach in MACM analysis (Eickhoff et al., 2011) to enable a reliable delineation of task-based functional connectivity. Importantly, the extent of this spatial filter was systematically varied from including the closest 20 to 200 experiments in steps of two (Clos et al., 2013). That is, selection was performed for the sets of 20, 22, 24, ..., 198, 200 experiments reporting the closest activation at a given seed voxel (i.e., 91 filter sizes). This was implemented by calculating and subsequently sorting the Euclidean distances between a given seed voxel and any activation reported in BrainMap. Then, the x nearest activation foci (i.e., filter size) were associated with that seed voxel. The retrieved experiments were then used to compute the brain-wide coactivation profile of a given seed voxel for each of the 91 filter sizes. In particular, a coordinate-based meta-analysis was performed over all foci reported in these experiments to quantify their convergence. Since the experiments were identified by activation in or near a particular seed voxel, highest convergence was obviously found at the location of the seed. Convergence outside the seed, however, indicated coactivation across task-based functional neuroimaging experiments. These brain-wide coactivation patterns for each individual seed voxel were computed by activation likelihood estimation (ALE). The key idea behind ALE is to treat the foci reported in the associated experiments not as single points, but rather as centers for 3D Gaussian probability distributions that reflect the spatial uncertainty associated with neuroimaging results. Using the latest ALE implementation (Eickhoff et al., 2009; Turkeltaub et al., 2012; Eickhoff et al., 2012), the spatial extent of those Gaussian probability distributions was based on empirical estimates of between-participant and between-template variance of neuroimaging foci (Eickhoff et al., 2009). For each experiment, the probability distributions of all reported foci were then combined into a modeled activation (MA) map by the recently introduced 'non-additive' approach that prevents local summation effects (Turkeltaub et al., 2012). The voxel-wise union across the MA maps of all experiments associated with the current seed voxel then yielded an ALE score for each voxel of the brain that describes the coactivation probability of that particular location with the current seed voxel. The ALE scores of all voxels within gray matter (based on

10% probability according to the ICBM maps) were recorded before moving to the next voxel of the seed region. In sum, quantitative ALE meta-analysis over all foci reported in the experiments associated with the current seed voxel determined how likely any other voxel throughout the brain was to coactivate with that particular seed voxel. Notably, no threshold was applied to the ensuing coactivation maps at this point of analysis to retain the complete pattern of coactivation likelihood.

Task-independent functional connectivity (RSFC). Seed-voxel-wise whole-brain connectivity was likewise assessed using resting-state correlations as an independent modality of functional connectivity for cross-validation across different brain states. RSFC fMRI images were acquired in 100 healthy volunteers (50 female, mean age 45.2 years) without any record of neurological or psychiatric disorders. All participants gave written informed consent prior to entering the study, which had been approved by the ethics committee of the University of Bonn, Germany. Prior to the imaging session, participants were instructed to keep their eyes closed and just let their mind wander without thinking of anything in particular but not to fall asleep (which was confirmed in post-scan debriefing). For each participant, 300 RSFC EPI images were acquired using BOLD contrast [gradient-echo EPI pulse sequence, TR = 2.2s, TE = 30ms, flip angle = 90, in-plane resolution = 3.1 x 3.1mm2, 36 axial slices (3.1mm thickness) covering the entire brain]. The first four scans served as dummy images allowing for magnetic field saturation and were discarded prior to further processing using SPM8 (`www.fil.ion.ucl.ac.uk/spm`). The EPI images were first corrected for head movement by affine registration using a two-pass procedure. The mean EPI image for each participant was then spatially normalized to the MNI single participant template (Holmes et al., 1998) using the unified segmentation approach (Ashburner and Friston, 2005) and the ensuing deformation was applied to the individual EPI volumes. Finally, images were smoothed by a 5-mm FWHM Gaussian kernel to improve signal-to-noise ratio and compensate for residual anatomical variations. The time-series data of each individual seed voxel were processed as follows (Weissenbacher et al., 2009): In order to reduce spurious correlations, variance that could be explained by the following nuisance variables was removed: *i)* The six motion parameters derived from the image realignment, *ii)* the first derivative of the realignment parameters, *iii)* mean gray matter, white matter and CSF signal per time-point as obtained by averaging across voxels attributed to the respective tissue class in the SPM 8 segmentation and *iv)* coherent signal changes across the whole brain as reflected by the first five components of a principal component analysis decomposition of the whole-brain time series. All of these nuisance variables entered the model as first-order and - except for the principal components - also as second-order terms. Data were then band-pass filtered preserving frequencies between 0.01 and 0.08 Hz since meaningful resting-state correlations will predominantly be found in these frequencies given that the BOLD-response acts as a low-pass filter (Biswal et al., 1995). After temporal preprocessing, the times-series of each individual seed voxel were correlated with those of any other brain voxel. The ensuing correlation values were transformed into Fisher's Z-scores and recorded in a connectivity matrix. In sum, correlations between spontaneous metabolic fluctuations throughout the brain during mind-wandering in the absence of an externally structured task allowed quantifying the connectivity strength of the current seed voxel with any other voxel.

Connectivity-based segregation of DMN nodes. To identify possibly distinct subregions with unique connectivity patterns in the ROI at hand, CBP was performed based on MACM (Eickhoff et al., 2011) and RSFC (Kim et al., 2010) analyses. Independent for each of the two modalities, the brain-wide connectivity profiles for all seed voxels were combined into a NS x NB coactivation matrix, where NS denotes the number of seed voxels and NB the number of voxels in the reference brain volume. The most appropriate number of clusters in the respective ROI was then, analogous to previous CBP approaches (Behrens et al., 2003), determined in a NS x NS cross-correlation matrix. This matrix reflected how strongly the connectivity profiles of each pair of seed voxels correlated with each other. Given the use of 91 different filter sizes, this step resulted in 91 individual connectivity matrices, each representing the whole-brain connectivity of the seed voxels at a particular filter size. The parcellation of the VOI was performed using k-means clustering as implemented in Matlab with $k = 2, 3, .., 9$ using one minus the correlation between the connectivity patterns of seed voxels as a distance measure (i.e., correlation distance). This parcellation was performed for each of the 91 filter sizes independently, yielding 8 (k-means cluster solutions) x 91 (filter sizes) independent cluster solutions (Clos et al., 2013). K-means clustering is a non-hierarchical clustering method that uses an iterative algorithm to separate the seed region into a previously selected number of k non-overlapping clusters (Forgy, 1965; Hartigan and Wong, 1979). In all k-means applications, the distance measure was computed by 1 minus the sample correlation between points. K-means aims at minimizing the variance between elements within clusters and maximizing the variance between clusters by first computing the centroid of each cluster and subsequently reassigning voxels to the clusters such that their difference from the nearest centroid is minimal. For each of the 8 x 91 parcellations we recorded the best solutions from 100 replications with randomly placed initial centroids.

Characterization of the CBP-derived clusters: Connectivity. Following parcellation of the seed region based on regional heterogeneity in functional connectivity, additional MACM and RSFC analyses were performed on each of the ensuing clusters to characterize their whole-brain connectivity patterns. It is important to note that the above MACM and RSFC analyses assessed seed-voxel-wise connectivity patterns of individual seed voxels, while the overall connectivity pattern of a set of seed voxels was here assessed, i.e., the connectivity of the entire cluster. For the MACM analyses on the derived clusters, an ALE meta-analysis was performed across all BrainMap experiments featuring at least one focus of activation within the cluster using otherwise the same approach as described above. However, statistical inference was sought at this point, in contrast to the above MACM/RSFC analysis. To establish which regions were significantly coactivated with a particular cluster of voxels, ALE scores for the MACM analysis of this cluster were compared to a null-distribution that reflects a random spatial association between experiments, but regards the within-experiment distribution of foci as fixed. This random-effects inference assesses above-chance convergence between experiments. The observed ALE scores from the actual meta-analysis of experiments activating within a particular cluster were then tested against the ALE scores obtained under this null-distribution yielding a p-value based on the proportion of equal or higher random values. The resulting p-values were then thresholded at $p<0.05$ following cluster-level family-wise error correction for multiple comparisons (cluster-forming threshold at voxel-level: $p<0.001$). In addition to MACM analyses, RSFC analysis was also performed on the derived clusters. Time courses were extracted for all gray-matter voxels of a given cluster of the individual participant (Ashburner and Friston, 2005). The cluster time course was then expressed as the first eigenvariate of these voxels time courses. Pearson correlation coefficients between the time series of the CBP-derived ROI clusters and all other gray-matter voxels in the brain were computed to quantify RSFC. These voxel-wise correlation coefficients were then transformed into Fishers Z-scores and tested for consistency across participants by a one-sample t-test. The results of this random-effects analysis were then thresholded at $p<0.05$ following cluster-level family-wise error correction for multiple comparisons (cluster-forming threshold at voxel-level: $p<0.001$), analogous to MACM-derived cluster connectivity.

Characterization of the CBP-derived clusters: conjunction across MACM and RSFC results. To delineate areas showing task-dependent and task-independent functional connectivity with the derived subregions in the ROI, a conjunction analysis of the MACM and RSFC results was performed using the strict minimum statistics (Nichols et al., 2005). In practice, regions connected with the seed in both connectivity modalities were delineated by computing the intersection of the (cluster-level family-wise-error-corrected) connectivity maps from the two analyses detailed above. In this way, each derived cluster was associated with a network of areas that are congruently connected to that cluster across disparate (i.e., task-focused and mind-wandering) brain states.

2.3 Experimental results

The 'true' shape and number of clusters is unknown for most real-world clustering problems, and thus also in neuroimaging research. Finding an 'optimal' number of clusters represents an unresolved issue (i.e., *cluster validity problem*) in computer science, statistics, and machine learning (Jain, 2010; Handl et al., 2005). This has prompted the development of diverse heuristics (*cluster validity criteria*) to weigh the quality of the obtained clustering solutions. This heterogeneous set of evaluative measures is necessary because clustering algorithms will always find subregions in the investigator's ROI, whether these truly exist in nature or not. For all seed regions, these heuristic procedures to find the most neurobiologically pertinent clustering solution have indicated a most reasonable cluster number choice for each of the DMN nodes given MACM and RSFC connectivity data (Fig. 2).

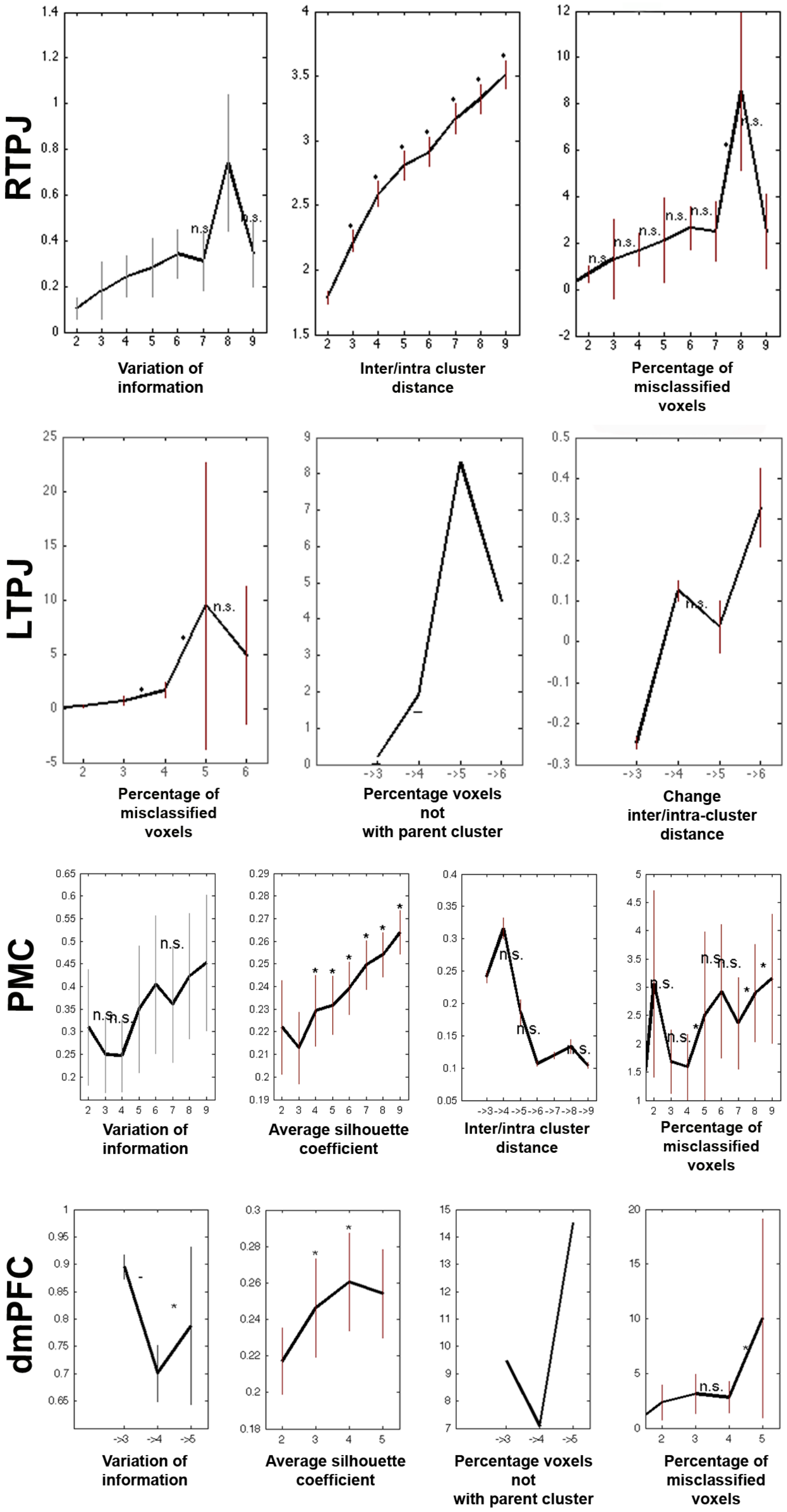

Figure 2: **Different clustering criteria for model selection**

Various different estimates of model fit have been applied to find the superior cluster solution. Note that only the computed cluster validation criteria are shown that converged to the respective optimal cluster number for each seed region.

Regarding the RTPJ node of the DMN, task-dependent connectivity (i.e., MACM) capturing neural activity during defined mental operations and task-independent connectivity (i.e., RSFC) capturing neural activity during task-free mind-wandering were computed for every single voxel within the seed region. For each of these two modalities, the connectivity profiles of all seed voxels were correlated with each other, yielding a symmetric matrix that indicated the similarity in whole-brain connectivity (for the respective modality) across any pair of seed voxels. k-means clustering of these two (MACM- and RSFC-derived) similarity matrices congruently indicated the presence of two distinct clusters within the RTPJ (Fig. 1). Thus, two independent modalities of functional connectivity, task-dependent MACM and task-independent RSFC, provided highly convergent two-cluster solutions with 90% of overlap. In contrast, these two modalities diverged strongly when attempting a more fine-grained clustering. In the three-cluster solution, only 38% of the seed voxels were assigned congruently across both analyses. The distinction into two clusters, therefore, represents the most robust regional differentiation within the RTPJ. Furthermore, the two-cluster solution was favored by three diverging cluster validity criteria (Fig. 2). First, the information-theoretic criterion variation of information yielded lowest values for two clusters. This minimum of variation thus reflected highest integrity of information in the two-cluster solution. Second, the inter/intra-cluster distance ratio, a cluster-separation criterion, was steadily increasing with the number of clusters without exhibiting any special structure. This reflected normal behavior of a data space that is increasingly compartmentalized and thus advocated the two-cluster solution for the sake of parsimony. Third, as a topological criterion, the percentage of misclassified voxels across filter sizes was lowest for two clusters. This indicated that lowest cluster number exhibited the least noise across the different filter sizes. Information-theoretic, cluster-separation, and topological criteria thus agreed on the two-cluster segregation of the RTPJ ROI as best fitting model given the data.

Task-constrained coactivations (MACM) and task-unconstrained time-series correlations (RSFC) congruently showed functional connectivity of the cluster 1 with the bilateral primary motor cortex, midcingulate cortex/supplementary motor area and anterior insula/inferior frontal gyrus. This set of areas resulting from congruent connectivity to the purple cluster 1 across both connectivity approaches could be referred to as 'anterior RTPJ network.' The orange cluster 2, in turn, featured congruent (across MACM and RSFC) functional connectivity with the bilateral inferior parietal cortex, precuneus, and right middle temporal gyrus - a 'posterior RTPJ network'. Importantly, a reciprocal relationship was found between the brain networks linked to cluster 1 and cluster 2, respectively, by assessing each clusters negative times-series correlations (Fig. 3). In particular, the bilateral midcingulate cortex/supplementary motor area and anterior insula/inferior frontal gyrus were not only positively coupled with cluster 1 across MACM and RSFC but were also negatively coupled with cluster 2 in the RSFC analysis. Conversely, bilateral inferior parietal cortex and precuneus (not including the posterior cingulate cortex) were positively coupled with cluster 2 and negatively coupled with cluster 1. From a neurophysiological perspective, one set of brain areas thus probably

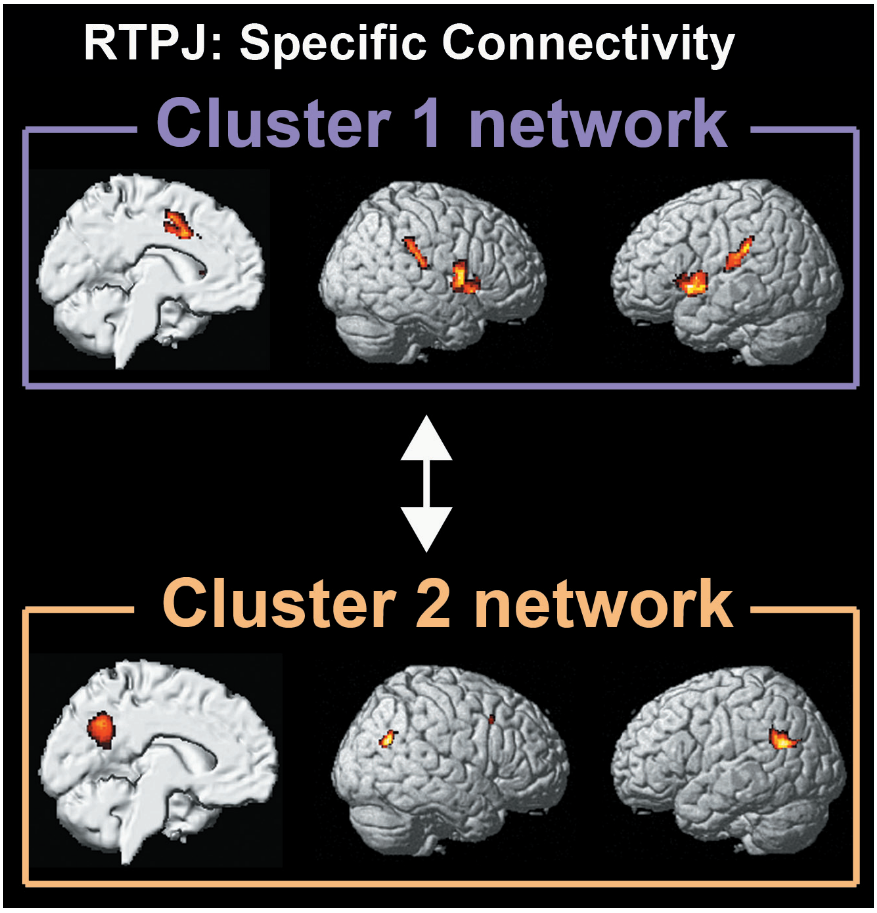

Figure 3: **Specific connectivity of the RTPJ clusters**
The midcingulate cortex/supplementary motor area and anterior insula/inferior frontal gyrus (*upper box*) demonstrated a positive functional relationship in MACM and RSFC analysis with purple cluster 1 as well as a negative one (RSFC-) with orange cluster 2. Conversely, bilateral TPJ and precuneus (lower box) were positively related (MACM, RSFC+) to the posterior RTPJ subregion and negatively related (RSFC-) to the anterior RTPJ subregion.

increases metabolic activity together with RTPJ cluster 1 and decreases activity with RTPJ cluster 2, while another set of brain areas shows the opposite pattern. These findings suggest a functional anti-correlation between the subregions discovered in the anterior and posterior RTPJ.

Regarding the LTPJ node of the DMN, several metrics were applied to weigh the various cluster solutions for the LTPJ ROI against each other (Fig. 2). First, as a topological criterion, the percentage of misclassified voxels across filter sizes was lowest for solutions up to four clusters. This indicated that low cluster numbers exhibited the least noise across the different filter sizes. Second, as another topological criterion, the percentage of voxels not related to the dominant parent cluster was lower in the four-cluster solution than for solutions with more clusters. Dividing the parietal ROI into four clusters thus contained relatively few regrouped voxels and therefore high continuity with their dominant parent cluster from the k-1 solution. Third, change of inter/intra-cluster ratio, a cluster-separation criterion, was higher for four clusters comparing to the three- and five-cluster solutions. This indicated that the four-cluster solution isolated each cluster well from the remaining ones. The four different measures of clustering quality thus unequivocally advocated the four-cluster solution as the best-fitting model and, thus probably, the most neurobiologically meaningful subdivision of the LTPJ ROI (Fig. 1).

Additionally, it can be instructive to consider the neighboring cluster solutions and their relations. In the three-cluster solution, dorsal aspects of the LTPJ ROI were separated into a single cluster, while ventral aspects of the ROI were separated into a rostroventral (most likely related to Wernicke's area) and a caudodorsal cluster. In the four-cluster solution, the former dorsal cluster was further subdivided into a bigger green rostromedial and a smaller caudolateral cluster. In the five-cluster solution, the former cluster was further subdivided into a medial and a lateral cluster. Note that k-means clustering was here applied independently several times to the same ROI. This procedure does not enforce hierarchically consistent cluster solutions. Nevertheless, the rostroventral and caudoventral clusters emerged independently with consistent topography in all three clustering analyses. This means that the regional heterogeneity in the whole-brain connectivity was more prominent for these clusters than for the clusters emerging from the green cluster. In other words, the two clusters in the ventral LTPJ ROI capture a more distinct connectional-functional segregation than the latter emerging clusters in the dorsal LTPJ ROI. This consolidates the usefulness of the four-cluster solution.

Given the overlap between the connectivity profiles of the LTPJ clusters, parts of the brain have been investigated that were more strongly connected to a given cluster than the respective three other clusters). To this end, the brain regions were isolated that were selectively connected with a given cluster in contrast to all remaining clusters. For instance, to delineate the specific cluster connectivity of cluster 1, an AND conjunction was computed across the three difference maps (cluster 1 - cluster 2), (cluster 1 - cluster 3), and (cluster 1 - cluster 4). This procedure removed connectivity of cluster 1 that was shared with clusters 2, 3, and 4. This is because any voxel that is deemed to reflect specific connectivity of a given cluster had been determined to be statistically more associated with that cluster in three separate difference analyses with the respective three other clusters. According to MACM (Fig. 4), blue cluster 1 featured highest connectivity strength to the bilateral superior temporal gyrus (coinciding with Wernicke's area in the left hemisphere), superior temporal sulcus, inferior frontal gyrus, as well as aspects of the inferior parietal lobe. In the left hemisphere, cluster 1 was also specifically connected to the anteiror insula. These specific connectivity targets were confirmed by RSFC. Additionally, cluster 1 featured highest RSFC to the middle temporal gyrus, temporal pole, and mid/posterior cingulate cortex. The conjunction across specific MACM and RSFC corroborated the specific MACM profile of cluster 1, except for the left anterior insula and inferior frontal gyrus. Red cluster 2, according to MACM, demonstrated the highest connectivity strength to the bilateral ventromedial/frontopolar/dorsomedial prefrontal cortex, extending into the rostral anterior cingulate cortex,

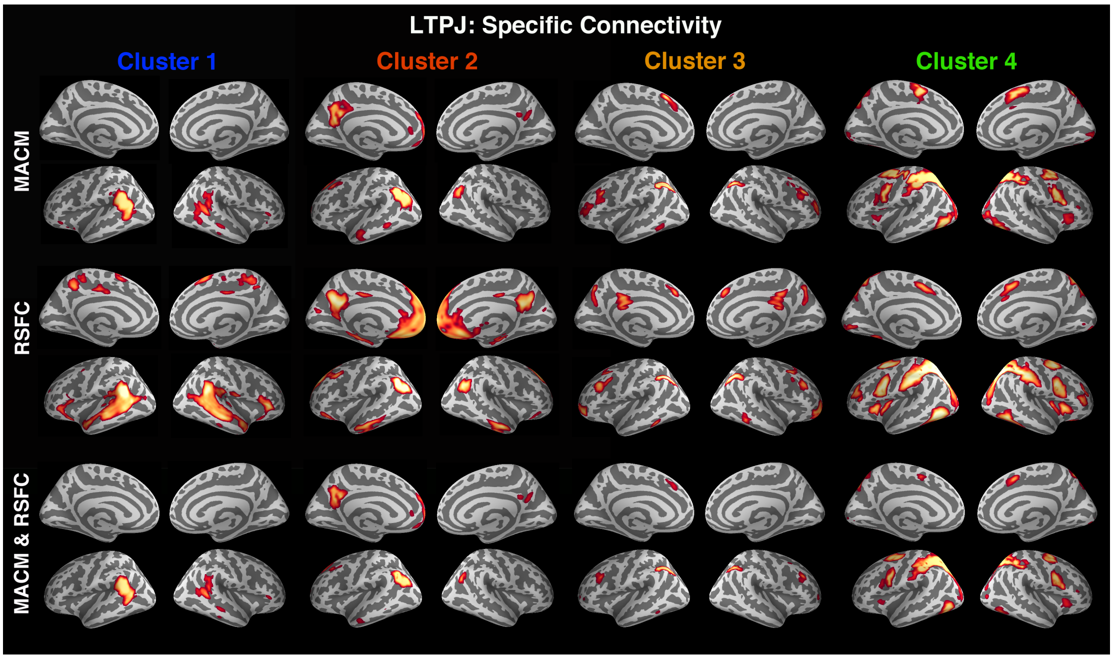

Figure 4: **Specific connectivity of the LTPJ clusters**
Specific functional connectivity patterns of the subregions of the four-cluster solution as determined using meta-analytic connectivity modelling (MACM; *top two rows*), resting-state connectivity (RSFC; *middle two rows*), and the conjunction of both methods (MACM & RSFC; *bottom two rows*). The significant results are rendered on left/right lateral and medial views of brain regions. Specific connectivity reflects stronger functional connectivity to a given cluster in the parietal ROI than to any of the three other clusters according to meta-analytic connectivity modelling (MACM).

posterior cingulate cortex, as well as aspects of the inferior parietal lobe. Specific connectivity in the left hemisphere was observed in the superior frontal gyrus and middle temporal gyrus. Notably, cluster 2 yielded the most widespread selective connectivity to highly associative brain regions among all four clusters. RSFC confirmed these specific connectivities by conjunction analysis and showed additional distributed results by individual analysis in the midcingulate, medial temporal, visual, and anterior-cingulate regions. Orange cluster 3 featured highest MACM coupling with the bilateral intraparietal sulcus and anterior aspects of dorsolateral prefrontal cortex. Specific connectivity in the left hemisphere was observed in left inferior temporal gyrus and anterior aspects of midcingulate cortex/supplementary motor cortex. Individual and conjunction RSFC analysis confirmed this set of regions. Yet, a part of the posterior cingulate cortex and the right inferior frontal gyrus were only revealed by specific RSFC. Green cluster 4 featured highest MACM connectivity to bilateral superior parietal lobe, posterior aspects of midcingulate cortex/supplementary motor area, and posterior aspects of dorsolateral prefrontal cortex, as well as anterior insula, primary visual cortex (including fusiform gyrus), and cerebellum (not shown). Indeed, specific RSFC confirmed this entire set of regions by conjunction analysis.

Regarding the PMC node of the DMN, several metrics were applied to weigh the various cluster solutions for the PMC ROI against each other (Fig. 2). First, the information-theoretic criterion 'variation of information' slightly decreased from three to four clusters and steeply increased from four to five clusters. This indicated that each cluster of the k-means clusterings became increasingly chaotic starting from five clusters. Second, the cluster-separation criterion 'silhouette coefficient' showed a positive bump at four clusters in the upward trend starting from three clusters. This indicated that clustering into four groups featured unexpectedly compact clusters, although compactness increases with the number of clusters. Third, the other cluster-separation criterion change of inter/intra-cluster ratio was highest for four clusters. This indicated that the four-cluster solution best isolated each cluster from the remaining ones. Fourth, as a topological criterion, the percentage of misclassified voxels across filter sizes was lowest for four clusters. This indicated that the four-cluster solution exhibited the least noise across the different filter sizes. The four different measures of clustering quality thus unequivocally advocated the four-cluster solution as the most robust differentiation in the PMC ROI. Cluster 1 emerged in the dorsocaudal part of the ROI in the parietal lobe (Fig. 1). Cluster 2 emerged in the ventral part of the PMC ROI in the posterior cingulate cortex. Cluster 3 emerged in the dorsorostral ROI in the posterior cingulate cortex, such as cluster 2. Finally, cluster 4 emerged in the ventrorostral ROI in and near to the retrosplenial cortex.

As to MACM (Fig. 5), red cluster 1 of the PMC ROI was specifically connected to the bilateral dorsolateral prefrontal cortex (at the caudal end of the middle frontal gyrus and superior frontal sulcus), intraparietal sulcus, and midcingulate cortex/supplementary motor area, as well as the right frontal eye field, temporo-parietal junction, and supramarginal gyrus. As to RSFC, cluster 1 was specifically connected to the right dorsolateral prefrontal cortex (similar topography as for MACM) extending into the frontal eye field, bilateral intraparietal sulcus, and right temporoparietal junction. Across MACM and RSFC, cluster 1 was congruently specifically connected to the bilateral intraparietal sulcus as well as the right frontal eye field and temporo-parietal junction. As to MACM, green cluster 2 was specifically connected to the bilateral vmPFC and inferior parietal cortex, as well as left inferior frontal gyrus. As to RSFC, cluster 2 was specifically connected to the bilateral vmPFC and inferior parietal cortex. Across MACM and RSFC, cluster 2 was congruently specifically connected to bilateral vmPFC and middle aspects of the left inferior parietal cortex. As to MACM, blue cluster 3 was specifically connected to the bilateral dmPFC, inferior parietal cortex, and posterior midcingulate cortex, as well as the left dorsolateral prefrontal cortex, superior parietal cortex, and posterior superior temporale sulcus. As to RSFC analyses, cluster 3 was specifically connected to the bilateral dmPFC, dorsolateral prefrontal cortex, inferior parietal cortex, posterior midcingulate cortex, middle temporal gyrus, temporal pole, and cerebellum (not shown). Across MACM and RSFC, cluster 3 was congruently specifically connected to the bilateral dmPFC, posterior midcingulate cortex, inferior parietal cortex, as well as the left caudal dorsolateral prefrontal cortex (close to the inferior frontal junction and ventral premotor cortex). As to MACM, yellow cluster 4 was specifically connected to bilateral anterior

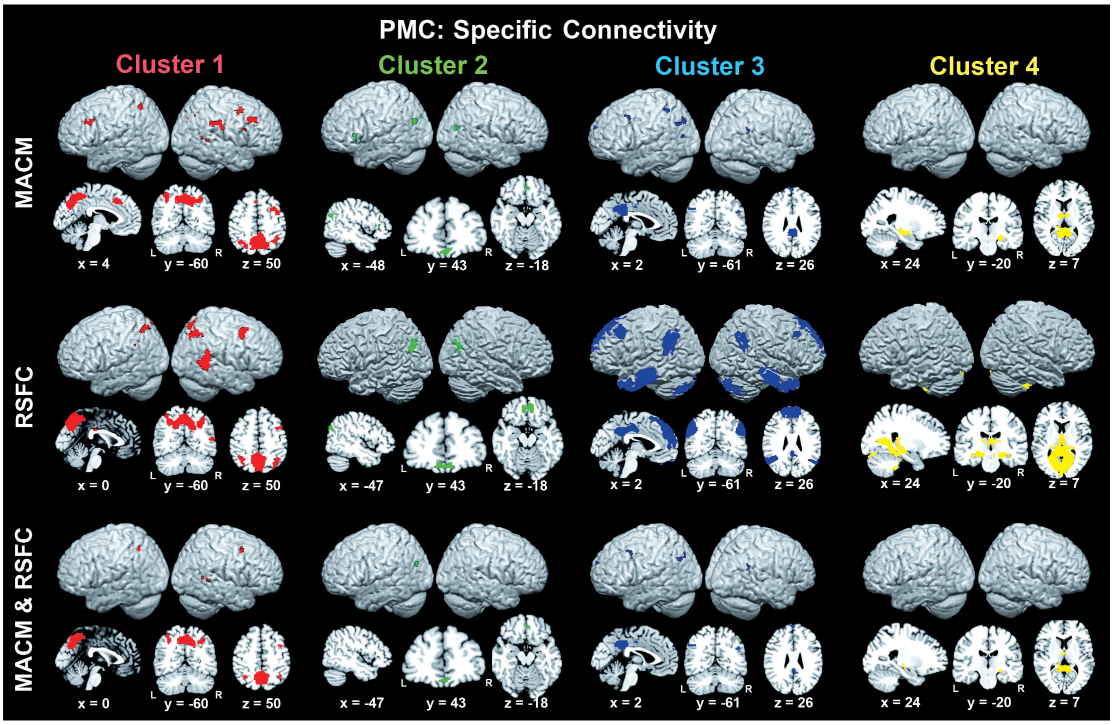

Figure 5: **Specific connectivity of the PMC clusters**
Depicts renderings as well as sagittal, coronal, and axial section views of brain regions more strongly functionally connected to a given cluster than to any of the three other clusters according to meta-analytic connectivity modelling (MACM; *top two rows*), resting-state functional connectivity (RSFC; *middle two rows*), and the conjunction across MACM and RSFC (*bottom two rows*). Coordinates in MNI space.

thalamus and right hippocampus. As to RSFC, cluster 4 was specifically connected to the bilateral hippocampus, thalamus, occipital lobe, and cerebellum (not shown), as well as left amygdala. Across MACM and RSFC, cluster 4 was congruently specifically connected to the bilateral anterior thalamus and right hippocampus.

To finally determine the optimal parcellation of the dmPFC ROI, metrics quantified model fit for comparison between cluster solutions. Information-theoretic, cluster separation, topological, and consistency criteria agreed in favoring the four-cluster solution as featuring the highest stability (Fig. 2). First, the information-theoretic criterion indicated variation of information to decrease from three to four clusters and to increase towards five clusters. This minimum of variation thus reflected highest integrity of information in the four-cluster solution. Second, as cluster separation criterion, the silhouette coefficient showed an increase from three to four clusters, whereas this metric decreased again in five clusters. Four clusters thus exhibited highest similarity among the voxels in each cluster. Third, as topological criterion, the percentage of voxels not related to the dominant parent cluster was minimal in the four-cluster solution. Four clusters thus contained the least amount of regrouped voxels and therefore highest continuity with their dominant parent cluster from the k-1 solution. Comparing to the three- and five-cluster solutions, the four-cluster solution showed highest hierarchical consistency. Fourth, as a consistency criterion, the percentage of misclassified voxels decreased in the four-cluster solution, whereas five clusters lost across-filter stability. In sum, these four diverging criteria measuring model fit all favored the four-cluster solution as the most stable segregation of the dmPFC ROI (Fig. 1).

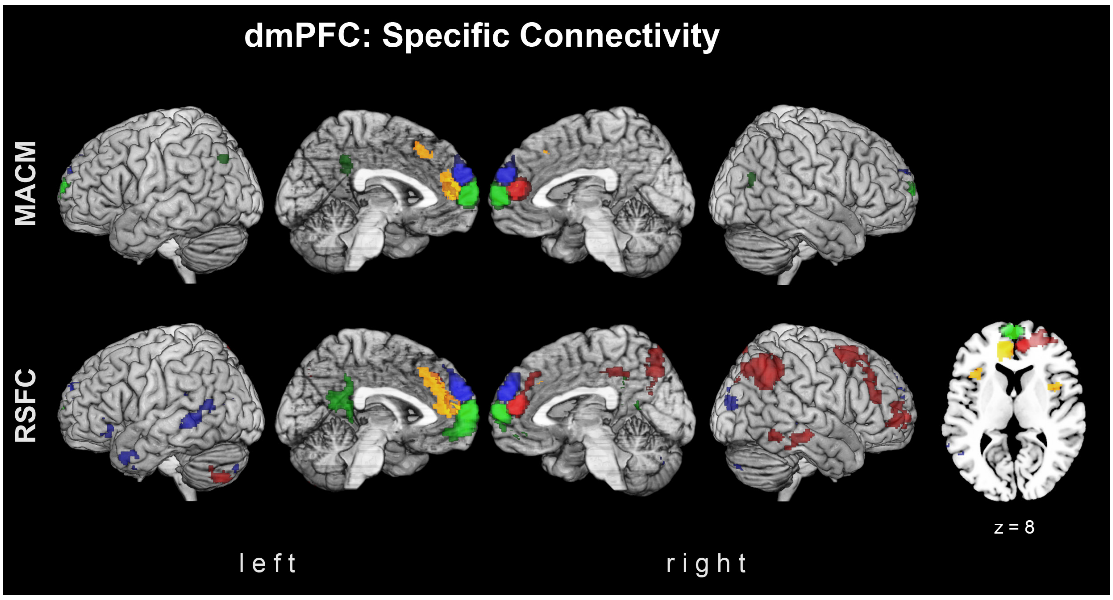

Figure 6: **Specific connectivity of the dmPFC clusters**
Depicting connectivity patterns which are stronger connected to a given dmPFC cluster, compared to all three other clusters based on meta-analytic connectivity modelling (MACM; *top row*) and resting-state functional connectivity (RSFC; *middle row*). Specific connectivity is rendered for caudal-right (red), rostroventral (green), rostrodorsal (blue), caudal-left (yellow) cluster.

In the MACM analysis (Fig. 6), specific connectivity patterns were only found for the green cluster 2 and yellow cluster 4. The green rostroventral cluster 2, more than all other dmPFC clusters, was connected to left posterior cingulate cortex and bilateral inferior parietal lobe, i.e., regions that form the core of what is known as the default mode network. The yellow caudal-left cluster 4, in turn, was specifically connected to the (left) anterior midcingulate cortex. In the RSFC analysis, rostroventral cluster 2 and caudal-left cluster 4 showed similar functional connectivity as in the specific MACM analyses. Rostroventral cluster 2 was specifically connected to the left posterior cingulate cortex. Cluster 4 featured specific connectivity to left anterior midcingulate cortex and - in addition to the MACM results - also the bilateral anterior insula. Furthermore, RSFC analyses also revealed specific connectivity for right-caudal cluster 1 and rostrodorsal cluster 3: Red cluster 1 was specifically connected to the right dorsolateral prefrontal cortex, superior parietal cortex, intraparietal sulcus, inferior parietal lobe, dorsal posterior cingulate cortex, and precuneus, middle temporal gyrus, and inferior temporal gyrus as well as left cerebellum (not shown). The blue rostrodorsal cluster 3 further showed specific connectivity to the left inferior frontal gyrus, temporal pole and bilateral middle temporal gyrus, middle occipital gyri and cerebella (not shown). It is noteworthy that RSFC yielded more distributed connections than MACM in all dmPFC clusters. RSFC thus also revealed more extensive specific connectivity patterns.

2.4 Discussion

From an anatomical perspective of brain segregation (Amunts et al., 2013), cortical areas are believed to be distinguishable from their neighbors by featuring a distinct microstructure, distinct connectivity, and distinct function. Brain function may follow naturally from brain structure and connectivity that conjointly enable locally specific neuronal computations (Passingham et al., 2002). As CBP is based on connectivity the defined clusters can be interpreted as 'functional areas' in a broad sense, but may not be directly interpretable as 'cortical areas' in a strict sense. In fact, the concepts of what constitutes a cortical area are mainly derived from studies of early sensory (Van Essen et al., 1992) and motor (Rizzolatti, 1987) brain systems. They may conceivably not be applicable to higher-level associative brain areas, such as the segregated nodes of the DMN. For instance, it is more and more difficult to relate the connectivity pattern of an area to its functional roles with increasing distance from sensory input processing (Mesulam, 1998; Yeo et al., 2011).

From a more methodological perspective of brain segregation, CBP does not address the question whether there is a 'true' neurobiological parcellation in a given brain region. CBP is employed to identify the clustering solution that is 'optimal' to describe the data. The question remains whether different parcellation results for the same ROI capture different resolutions or dimensions of an underlying neurobiological organization. The answer depends on the region of interest. For instance, previous CBP work on the insula (Nanetti et al., 2009; Kelly et al., 2012) and the right temporo-parietal junction (Mars et al., 2012) have indicated close agreement between the parcellations based on different connectivity modalities. In contrast, parcellations of the posteromedial cortex have diverged more strongly between existing CBP studies based on different connectivity modalities (Cauda et al., 2010; Zhang and Li, 2012). These observations corroborate the relevance of the conceptual differences between aspects of connectivity, such as RSFC, MACM, and others. Moreover, there is probably no such thing as 'the connectivity' for a particular brain volume of interest.

None of the connectivity modalities currently existing in neuroimaging provides axonal connectivity in stricto sensu (as gleaned from axonal tracing studies in monkeys). RSFC is task-unconstrained (i.e., task-independent) as opposed to task-constrained (i.e., task-dependent) MACM. Both RSFC and MACM identify temporal coincidence of neural signals in gray matter, that is, functional connectivity. Yet, functional connectivity between two regions may be mediated by a third region. That is, RSFC and MACM may be driven by indirect connections. This could however be alleviated by computing partial correlations, which is closer to direct interaction by summarizing conditional independences. In fact, regression-based estimators, such as dictionary learning, instead of standard clustering approaches, may be more robust to this issue as they take into account signal interactions. Additionally, RSFC and MACM are generally sensitive in delineating existing connections but more prone to false positives (Jbabdi and Behrens, 2013). Importantly, RSFC-CBP but not MACM-CBP can be conducted in individual participants. RSFC-CBP thus enables detecting inter-individual differences in regional functional organization, while MACM-CBP inevitably generalizes across various inter-individual variability sources in the sampled participant population.

Given the biological and methodological characteristics of the most frequently used functional connectivity measures, the following tentative distinction is suggested. MACM-CBP may reveal clusters that are probably functionally distinct modules even though the spectrum of brain functions is likely to be larger than what can be probed by neuroimaging techniques (Smith et al., 2009; Mennes et al., 2013). That is, task-based functional connectivity might be limited by human real-world behavior being richer than human in-scanner behavior. The neurobiological nature of RSFC-CBP derived clusters might be less certain. This is because the relation of resting-state correlations to anatomical connectivity, function, and the brain's housekeeping physiology is currently only incompletely understood.

Moreover, each time the investigator choses a clustering algorithm to be applied on a ROI, he or she accepts a number of implicit or explicit assumptions (Hastie et al., 2011). Therefore, any clustering algorithm unavoidably biases the resulting clustering solution with respect to the number, shape, relative sizes, hierarchy, and contiguity of the clusters. The inevitable assumptions and biases of a clustering algorithm motivate the choice of a particular one based on the aim of the study, the ROI, and the employed connectivity data. Moreover, using different connectional modalities it is possible to provide a valuable cross-confirmation of the clusters' biological relevance.

With regard to CS-SL distinction (cf. 1.2), assessing the significance of brain parcellation results is hard. This is particularly true in the strict sense of inferential statistics, rather a broader sense of 'interesting' or 'relevant' structure. The key problem in wanting to assess statistical significance of CBP results is the requirement of a null hypothesis to test against. Conceptually, a ROI clustering solution would hence be deemed statistically significant if it has a very low probability of being true under the null hypothesis that the investigator seeks to reject. Yet, such a null hypothesis is often difficult to formulate in clustering applications. Instead of inferential statistics, which test for a particular structure in the clustering results, exploratory statistics should instead discover and assess structure in the data (Efron and Tibshirani, 1991; Hastie et al., 2011). While it is true that statistical methods span a continuum between the two poles of inferential and exploratory statistics, comparing the importance or pertinence of clustering results from a CBP analysis is naturally situated more towards the latter. CBP hence represents an unsupervised statistical learning problem that is conventionally addressed by quantitatively comparing model fit using cluster validity criteria. It may therefore be seen as one instance of a current shift in neuroimaging away from classical inferential towards exploratory approaches, put differently, from voxel-level mappings to more global assessment of model fit or predictive power (O'Toole et al., 2007; Brodersen, 2009; Ashburner and Klöppel, 2011; Formisano and Kriegeskorte, 2012).

Importantly, as connectivity-derived clusters are mere descriptions of the data, meaning of the obtained clusters can only arise in the adoption of a neurobiological viewpoint. With regard to the RTPJ node of the DMN, the brain regions specifically connected to the cluster 1 (i.e., more strongly than any of the other clusters) have been proposed to compose a saliency network that detects relevant stimuli to guide behavior (Sridharan et al., 2008). Consistently, comprehensive across-study analyses (Dosenbach et al., 2006) attested a role of this network in tonically maintaining the task-imposed cognitive set or the task plan. The synopsis of present and previous evidence thus suggests that the cluster-1 network is central for sensorimotor control by integrating supramodal stimulus-guided attention and action initiation during externally structured tasks. On the other hand, prior research consistently implicated the cluster-2-connected brain regions in higher social processes, including perspective-taking (Mar, 2011), social judgments (Freeman et al., 2010), and imagination-driven empathy (Lamm et al., 2011). Further research also frequently implicated the cluster-2 network in memory processes, including autobiographical/episodic memory retrieval (Spreng et al., 2009) and semantic processing (Binder et al., 2009). Indeed, the topography of the specific cluster-2 regions corresponds to previous meta-analytic definitions of the well known 'default mode network'. Summing up the RTPJ node of the DMN, cluster 1 may be considered part of an externally oriented, stimulus-driven network that probably controls attention to salient events in our environment and reactions towards these. Conversely, cluster 2 appears to be part of an internally oriented, stimulus-independent network potentially involved in continuous memory-informed mental imagery to potentially predict plausible social events related to self. Although cluster 1 and cluster 2 were each implicated in many seemingly unrelated tasks by functional profiling, the difference between the two task groups could be parsimoniously explained by the required attention to either the external world or self. Summing up clustering results on the RTPJ node of the DMN, the presence of an anterior and posterior subregion with antagonistic connectivity and thus also functions emerged from application of the CBP toolbox.

With regard to the LTPJ node of the DMN, clusters emerging in the ventral versus dorsal ROI were more consistent across clustering analyses. The follow-up connectivity analyses consistently indicated a rostro-ventral versus caudo-ventral cluster in the inferior LTPJ ROI associated with lower- versus higher-level aspects, respectively, of both social and language regions. Clusters that emerged in the dorsal ROI, in turn, were connectionally related to domain-general attention and working-memory regions. The rostro-ventral cluster 1 exhibited specific connectivity (i.e., connectivity that is stronger with cluster 1 than any other cluster in the ROI) with the bilateral superior temporal gyrus and sulcus, inferior frontal gyrus, and regions in the left parietal lobe. These areas have previously been associated with general aspects of task processing and stimulus-response processing in social cognition and language tasks. In contrast to cluster 1, the caudo-ventral cluster 2 was specifically connected with the bilateral inferior parietal lobe, ventro- and dorsomedial prefrontal cortex (extending into the neighboring anterior cingulate cortex) and the posterior cingulate cortex, and left superior frontal gyrus and middle frontal gyrus. Previous studies suggested that these regions subserve high-level associative functions, including the default mode of brain function (Buckner et al., 2008). Analogous to the posterior subregion in RTPJ, the observed connectivity profile for cluster 2 converges with a set of brain regions underlying the default mode network (Uddin et al., 2010). The remaining two clusters in the dorsal ROI were characterized by highly similar connectivity profiles. Both the cluster 3 and the cluster 4 featured connections to areas previously associated with general cognitive control processes (Dosenbach et al., 2006; Clos et al., 2013). Summing up clustering of the LTPJ node of the DMN, social cognition and language related processing facets are unlikely to be clearly dissociable in the LTPJ based on large quantities of fMRI measurements. While cluster 1 and 2 were both congruently associated with social-cognitive and language regions, the data provide evidence for a gradient in the hierarchy of processing. Cluster 1 might predominantly subserve lower-level processing facets in social cognition and language and cluster 2 might be more engaged in higher-level facets of these processes. This concurs with the closely intertwined relationship between the development of social cognitive and language capabilities in children (Heyes and Frith, 2014).

With regard to the PMC node of the DMN, cluster 1 turned out to be the only connectivity-derived region neuroanatomically situated in the parietal lobe proper rather than cingulate cortex. This cluster derived from the PMC ROI was selectively connected to the right frontal eye field, bilateral intraparietal sulcus, and right temporo-parietal junction, congruently across task-related (MACM) and task-unrelated (RSFC) analyses. This connectivity pattern of cluster 1 corroborates that the precuneus is not part of the default mode network (Parvizi et al., 2006; Margulies et al., 2009). Instead, the cluster 1 is probably implicated in the internally or externally triggered, overt or covert allocation of attentional resources to internal or external information. The cluster 2 was specifically connected to the bilateral ventromedial prefrontal cortex and aspects of the left inferior parietal cortex across task-related MACM and task-unrelated RSFC. This ROI-derived cluster may thus be implicated in predominantly evaluating object features, as opposed to space features, in perceived or imagined visual stimuli, potentially informed by semantic concepts. The cluster 3 was specifically connected to the bilateral dorsomedial prefrontal cortex, posterior midcingulate cortex, and inferior parietal cortex as well as to the left dorsolateral prefrontal cortex, congruently across MACM and RSFC. In fact, this is in line with the well-known integration of the dPCC's in the dorsal visual stream - the 'where' system for spatial processing (Margulies et al., 2009; Vogt et al., 2006), in contrast to cluster 2's potential relation to the ventral 'what' system. Cluster 3 in the PMC ROI is implicated in overt and covert navigation of the self and body in real or imagined spatial environments. The cluster 4 was prominently connected to the right hippocampus and bilateral anterior thalamus, congruently across MACM and RSFC analyses. Also in monkeys, the retrosplenial cortex featured particularly strong axonal connections with these two brain regions (Kobayashi and Amaral, 2003, 2007; Parvizi et al., 2006; Vogt et al., 1987). Cluster 4 may thus be implicated in mediating between organism-centered (i.e., egocentric or from current view) and world-centered (i.e., allocentric or from a bird's eye view) perspective frames, which is a frequent feature of both processing memory and spatial scenes. Summing up the PMC node clustering and associated network patterns, mind-wandering at rest might be subserved by different neural computations implemented in the four derived components of the PMC. Note that 'mind-wandering' is used here in the broad sense of cognitive processes that are not controlled by a certain task, which makes no judgment about the content of these thought processes. Integrating the above cluster-wise interpretations, mind-wandering might be instantiated by internally navigating the self in mapped hypothetical, present, future, or past spaces (dorsal posterior cingulate cortex potentially mediating spatial scene elaboration, retrosplenial cortex potentially mediating the translation between ego and world oriented reference frames) to allow detecting behaviorally relevant features in those envisioned scenes (ventral posterior cingulate cortex potentially mediating self-relevance evaluation) by shifting the attentional focus between various feature and aspect contemplations (precuneus potentially mediating the covert reallocation of attentional resources between different internal representations). In sum, if these speculations hold some truth in them, the PMC's conceivable key role in the continuous environmental tracking in a generative, integrative process might explain both its highest energy consumption in the brain and its intimate coupling with conscious awareness.

Finally, with regard to the dmPFC node of the DMN, two rostral ones (dorsal and ventral) and two caudal ones (left and right) cluster have emerged. The rostral clusters, especially the rostroventral cluster, were strongly connected to the posterior cingulate cortex and the inferior parietal cortex. Among all four clusters, the cluster 2 was most consistently functionally coupled with the DMN across task-related and task-unrelated brain states. The present results agree with the currently scarce evidence for an important DMN node in the most frontopolar mPFC, most closely corresponding to the present cluster 2. Further, both rostral clusters featured functional connections to the amygdala and hippocampus in the limbic system associated with memory and social cognitive tasks. In contrast, the two caudal clusters were hemispherically divided and predominantly connected to the respective ipsilateral hemispheres. From a network perspective of brain function, the cluster 1 was connected to a right-lateralized frontoparietal attentional network (Corbetta et al., 2008). The cluster 4 was connected to the left aMCC and bilateral anterior insulae, proposed to correspond to a saliency network (Seeley et al., 2007). The results thus support a differentiation within the dmPFC by regionally distinct functional relations to limbic, attentional, and default mode networks. Moreover, it became apparant that the delineated dmPFC clusters exhibited prominent differences between task-related and task-unrelated functional coupling patterns. This observation contrasts previous bimodal studies of seed regions, such as in the nucleus accumbens (Cauda et al., 2010), where MACM and RSFC largely conformed. That is, functional connectivity of the brain's task state (MACM) and of the brain's resting state (RSFC) frequently exhibited more similarities than dissimilarities in the interaction pattern with the rest of the brain. Most similar to the RTPJ, also the dmPFC suggests itself as a candidate region for mediating between neural systems that are typically more active and less active during task performance, respectively. During tasks (MACM) the cluster 4 featured strongest connectivity with the DMN, whereas during mind-wandering (RSFC) this cluster featured strongest connectivity to the putative saliency network. This currently underappreciated neurobiological behavior is challenging to interpret given that the present data are purely observational, rather than interventional (cf. 1.4). Also this last of the four investigated DMN nodes therefore invigorates the increasingly recognized relationship between the physiological baseline of the human brain and an introspective psychological baseline implicated in continuous self-related social cognition and memory retrieval (Schilbach et al., 2008, 2012; Timmermans et al., 2012). More clearly, the sophisticated, essentially unknown computational problems solved by the nodes of the default mode network seems to range from task-free to task-constrained processing in the human brain.

In the future, these data-driven results can yield reliable cornerstones for a variety of consecutive neuroimaging analyses. Experimental methods requiring a-priori target regions can capitalize on CBP clusters to further characterize their behavioral implications by diverse viewpoints towards cross-modal functional mapping. This might include, but is not exclusive to, transcranial magnetic stimulation (TMS), dynamic causal modelling (DCM), structural equation models (SEM), Granger causality mapping (GCM), and ROI-based experimental fMRI analyses. From a broader perspective, CBP can thus enhance any neuroimaging technique reliant on prospective region definitions that critically hinges on proper fit of the priors. Ultimately, using clustering algorithms in neuroscience offers useful simplified views on brain regions that remain complex in nature.

3 Supervised modelling of brain networks

3.1 Motivation

There is considerable uncertainty about the most pertinent concept of functional brain architecture. Brain organization is often viewed from modular versus distributed perspectives. While the modular perspective emphasizes the segregation of the human cerebral cortex into functionally distinct cortical areas (cf. previous chapter), the distributed perspective emphasizes the functional interplay between widespred brain networks (i.e., sets of cortical areas). Systems neuroscience has indeed established the existence of a set of fluctuating yet robust brain networks in humans (Biswal et al., 2010; Shehzad et al., 2009). These are likely manifestations of electrophysiological oscillation frequencies (Hipp and Siegel, 2015). Their fundamental organizational role is further attested by continued covariation during sleep and anesthesia (Fox and Raichle, 2007).

Identical neural networks have repeatedly been found across cognitive domains using diverging methods. These observations prompted widely-adopted network notions, including the 'default mode network' (DMN; (Raichle et al., 2001)), 'salience network' (SN; (Seeley et al., 2007)), and 'dorsal attention network' (DAN; (Corbetta et al., 2008)). Developmentally, such large-scale networks emerge during late fetal growth (Doria et al., 2010), before cognitive capacities mature in childhood. In adult humans, nodes of a same cohesive network probably have more similar functional profiles than nodes from different networks (Anderson et al., 2013). Indeed, resting-state fluctuations *between large-scale networks* were observed to be less stable than coupling between regions of each network when assessed by intra-class correlation (Shehzad et al., 2009). Between-network connections were also less stable than intra-network connections when assessed by Kendall's coefficient of concordance. During task-unrelated random thought, network-network dynamics therefore appear to be more volatile across participants and brain scans than intra-network dynamics. This suggests that the *constellation* of relative network implications is an underappreciated unit of functional brain organization.

Consistently, the onset of a given cognitive task might induce characteristic changes in functional coupling of large-scale networks. The SN and DAN tend to display BOLD signal increases due to experimental stimulation, although the DMN often decreases. Whether stimulus-evoked compositions of such networks explain the majority (Smith et al., 2009) or only a fraction (Mennes et al., 2013) of overall task activity is currently unresolved. For instance, a working-memory task entailed increase in BOLD activity in DAN regions but decrease in default mode regions (Fransson, 2006). Notably, the functional connectivity did not change significantly within either DAN or DMN during this fMRI task. During auditory event transitions in another experimental fMRI study, both DAN and SN increased in activity, whereas the DMN decreased in activity (Sridharan et al., 2008). These changes of network-network constellation are probably mechanistically relevant for unfolding behavior. This idea is supported by evidence that proportional DMN recruitment impairs task performance believed to be subserved by other large-scale networks (Mason et al., 2007; Weissman et al., 2006). The mediation between canonical networks was tentatively proposed to involve the right anterior insula (Sridharan et al., 2008) and right temporo-parietal junction (Bzdok et al., 2013). Moreover, the relevance of network-network architectures possibly extends to psychiatric and neurological disorders (Seeley et al., 2009; Mennes et al., 2010). In particular, reciprocal coupling between DMN decrease and task-recruited networks was found to be absent in autism (Kennedy et al., 2006), reduced in schizophrenia (Whitfield-Gabrieli et al., 2009) and major depression (Hamilton et al., 2011), as well as frequency-altered in attention deficit hyperactivity disorder (Liddle et al., 2011).

In neuroimaging research, the notion of brain 'networks' (Mesulam, 2012) has frequently been investigated in rest data (Biswal et al., 2010) with a few exceptions of applications to task data (Smith et al., 2009; Mennes et al., 2013; Cole et al., 2014). The relationship between brain activity during task and rest has predominantly been addressed by three neuroscientific tools (Zhang and Raichle, 2010): targeted neuroimaging experiments, seed-based analysis of resting-state correlations, and independent component analysis (ICA). How much experimental neuroimaging studies can contribute to research on task-rest correspondence is open to question. A paradox is introduced when attempting to measure genuine mind-wandering as prompted by a task while lying in a brain imaging scanner (Krienen et al., 2010; Bado et al., 2014). Additionally, each neuroimaging study can tap on only a small subset of experimental tasks. This precludes synoptic assessment of human cognition as a whole. Seed-based connectivity investigations depend on the assumption that the seed-region definition successfully captures an underlying biological structure (Bullmore E and O Sporns 2009). Seed correlation patterns are typically mixtures of overlapping canonical networks and are significance-tested in voxel space that is orthogonal to network interpretations. Moreover, ICA decomposes neuroimaging activations into a set of independent topographical maps (Beckmann et al., 2005). There is no obvious way to relate the ensuing network maps to traditional psychological concepts (Yeo et al., 2014). Although the number of such independent components is unknown, conventionally only one or two choices are explored (Smith et al., 2009; Laird et al., 2009). One might draw two important conclusions. First, the neurobiological validity of seed-correlation and ICA results is rarely assessed by inferential rejection of a null hypothesis or supervised comparison against the ground truth. Second, ICA has so far been the almost exclusive choice for network decomposition in neuroimaging data. This has largely neglected alternative *latent variable models*, including principal component analysis (PCA), sparse PCA (SPCA), and factor analysis (FA).

In general, finding important structure in data along the n most important directions of variation is frequently done using PCA (Shlens, 2014a; Bishop, 2006; Hastie et al., 2011). This analytical procedure reduces second-order correlation between the input features (referring to the second moment, i.e., covariance). PCA identifies the most meaningful basis (i.e., manifold) in order to rotate the data into a new coordinate system that reveals hidden meaningful structure. It is a data transformation that performs compression by exploiting correlations among the input variables. The ensuing first *principal component* expresses the direction of the largest variance corresponding to the largest eigenvalue. The subsequent principal component is derived along the second most important direction of variation under the constraint of being *orthogonal* to and thus uncorrelated with the first components etc. pp.[9] The assumptions made by PCA include *i)* linearity (complexity restrictions to the space of candidate bases), *ii)* large variance is an indicator of relevant structure, *iii)* the important modes of variation are orthogonal to each other, and *iv)* the input random variables follow a Gaussian distribution (use of PCA is discouraged otherwise). The same principal component space can be derived by maximizing the variance of the latent axes or minimizing the reconstruction error from the latent axes (Bishop, 2006). Once the PCA model has been fit to the data, any unseen data can be project onto (i.e., regressed against) the *principal component weights* of the n components (i.e., the explained-variance-ordered set of n basis vectors) to obtain *principal component loadings* (i.e., 'importance' along the coordinate system axes). In reexpressing a high-dimensional space in a low-dimensional space by eigenvalue theory, PCA is mathematically related to singular value decomposition (SVD) (Golub and Van Loan, 2012). SVD is a means from linear algebra for matrix decomposition by diagonalization to derive the singular values and singular vectors in the rectangular-matrix setting (i.e., not constrained to quadratic matrices such as general eigenvalue decomposition). PCA is further closely related to *whitening* procedures that perform Gaussian decorrelation transformation. It is often employed as a preprocessing step before actual data analysis (e.g., it is part of ICA, see passage below) that allows to treat all data dimensions equally. More specifically, whitening is a lossy operation (i.e., some information cannot be restored) that normalizes all dimensions to one (by dividing each component with the square root of its eigenvalue) and removes existing linear dependencies (by projecting the data onto the component eigenvectors). In imaging neuroscience using fMRI, PCA can readily be applied to series of brain images to impose orthogonality in both the components' time series and spatial patterns. In this context PCA yields the decomposition of BOLD time series into the spatial modes of highest variation (Friston et al., 1999a). It further rotates the fMRI signals into a space where the voxel values are less correlated with each other but conserves their important changes across brain scans. For instance, experimental fMRI activity has been projected onto the 40 PCA components to subsequently predict psychological tasks from individual brain scans (Carlson et al., 2003). Further, brain scans from participants exposed to stimuli depicting face and house pictures have also been decomposed into their distributed PCA patterns (Haxby et al., 2001). The first principal component explained 50% of the variance and located to the fusiform and parahippocampal gyri known to subserve the respective category perceptions (O'Toole et al., 2007). Up to now PCA has however mostly been applied to fMRI data in the aim of dimensionality reduction rather than for generation of neurobiologically interpretable network units.

[9] Algorithmically, somewhat similar to Gram-Schmidt orthonormalization.

Classical PCA may not always perform optimally in cases of few samples and many features (wide data), such as in typical fMRI studies (Viviani et al., 2005). Statistically, the sample eigenvectors and the population eigenvectors can diverge considerably in the high-dimensional setting (Johnstone, 2001). This statistical behavior is a result of the *dense* values of linear combinations of principal components that describe the original features. Many multivariate statistical models harness the curse of dimensionality by adding sparsity constraints to the optimization objective (Foucart and Rauhut, 2013). As a natural extension of PCA, sparse PCA derives *sparse principal components* with a maximum of zero weights to decrease dimensionality of the data projection vectors (Zou et al., 2006). Sparse PCA optimization thus weighs the tradeoff between correlation among the component loadings and sparsity of the component weights. The penalized decomposition of matrices can thus be viewed as an instance of automatic variable selection by imposing additional structural assumptions. Besides being better mathematically posed in high dimensions, less complex solutions also tend to be more interpretable. Importantly, the focus will be on sparsity in principal component weights, while an ℓ_1-penality can alternatively be imposed on the coefficients (i.e., 'loadings') of the principal components, such as in sparse dictionary learning[10] (Mairal et al., 2009). Linear combinations of feature *sub*sets can be extracted from neuroimaging data by reformulating PCA as a Lasso/Ridge/ElasticNet-type optimization problem (Tibshirani, 1996; Zou and Hastie, 2005). This can be achieved with or without additionally imposing orthogonality between the sparse principal components, although the orthogonality assumption can interfere with the sparsity constraint. In contrast to PCA, SPCA poses a hard nonconvex optimization problem (Hastie et al., 2015). Additionally, good SPCA implementations should sucessfully deal with multicollinearity between the input features (to avoid random, non-interpretable variable selection) and should be reducible to classical PCA (if the λ-coefficient of the ℓ_1-norm is set to zero). Sparse matrix decomposition of fMRI signals are also increasingly used in brain research (Varoquaux and Craddock, 2013). A sparse dictionary learning method has for instance been used to estimate a scalable atlas of functional brain regions from rest data (Varoquaux et al., 2011). This neurobiologically plausible statistical learning framework explicitly accounted for observation noise and estimated both subject- and group-level latent atlases. In sum, SPCA is a variant of PCA that derives linear combinations from subsets of the original voxels.

Furthermore, the interpretation of principal components has repeatedly been improved by rotation techniques (Jolliffe, 1995). Indeed, ICA reduces to first performing PCA/SVD (i.e., preliminary whitening) and subsequently searching for the rotation matrix yielding *independent components* (Hyvärinen and Oja, 2000; Bishop, 2006; Shlens, 2014b). Besides projection into an optimized coordinate system, PCA/SVD also accounts for noise estimation to which ICA is naïve. That is, successful application of ICA depends on preceding decorrelation of the data. It perform blind source separation and factor rotation by assuming that the

[10] Informally, sparse PCA can thus be viewed as the transpose of sparse dictionary learning.

data arise from a linear combination of (non-Gaussian) independent latent components. While the first step assumes Gaussian inputs and has a linear, analytical solution, the second step assumes non-Gaussian inputs and requires non-linear, numerical optimization to find the best rotation matrix. The first step accounts for second-order correlation, whereas the second step accounts for higher-order correlations. In contrast to PCA that learns directions of maximal variance, ICA learns components with a maximum of statistical independence, minimal mutual information, and maximal non-Gaussianity. The ensuing *mixing matrix* allows to deconvolve the linear transformation underlying factor combinations. The derived directions in the data are necessarily orthogonal in the whitened feature space (equal variance for each direction) but not in the original feature space (non-equal variance for each direction). In fact, ICA and PCA yield identical results if provided with Gaussian input data because uncorrelated Gaussian random variables are necessarily independent. That is, PCA alone can recover the unknown data-generating sources in cases of low higher-order correlation. In neuroimaging, ICA is by far the most frequently used decomposition method for times series of fMRI images (McKeown et al., 1998; Beckmann et al., 2005). Typically, ICA is applied by imaging neuroscientists to rest data of neural activity in the absence of a defined experimental task. The ensuing building blocks of brain activity are widely believed to represent macroscopical brain networks. These can be derived by their spatio-temporal fluctuation patterns without specifying an explicit a-priori hypothesis. Indeed, the networks patterns almost exclusively locate to the brain's gray matter, rather than yielding network nodes in white matter regions. Without an a-priori region, ICA can thus extract several 'networks' while separating out non-neuronal variance, for instance from movements and ventricles. These putative network units of brain organization have been compared with analoguous task networks (Smith et al., 2009) which exhibited a high degree of similarity. ICA was further capable of estimating strongly overlapping network components (e.g., with complementary hemispheric emphasis). Despite initial skepticism, such RS connectivity results were repeatedly shown to be consistent across participants, brain scans, time points, and other factors (Shehzad et al., 2009). ICA networks have for instance shown spatial structure that is robust to open versus closed eyes during image acquisition as well as to respiratory and cardiac sources of noise. In sum, ICA relies on the largest non-Gaussian directions to estimate a set of unobserved neuronal sources.

Finally, factor analysis models share many properties with PCA (Fabrigar et al., 1999; Beavers et al., 2013). There is actually an ongoing discussion whether or not FA and PCA are essentially identical models, yet they can produce different results. FA searches for hidden variables correlated with a maximum of the independent Gaussian inputs. It thus relies on covariance (i.e., second-order correlation) like PCA. FA analogously produces *latent factors* that are rotionally invariant, not unique, and cannot restore the original signals. Like PCA, FA is useful if the form of a latent factor space is relevant but not the coordinate system of its representation. Different from PCA, FA makes explicit assumptions about the form of the covariance matrix by expecting unequal variance of each factor (i.e., heteroscedasticity) rather than only ones on the diagonal (i.e., homoscedasticity). FA is aimed at quantifying *common variance*, whereas PCA yields latent variables of *maximal variance*. While PCA is a data transformation procedure from linear algebra, FA qualifies as a model-based approach. FA can in fact be used with both exploratory and inferential aims (Thompson, 2004). Additionally, FA does not yield explained variance scores for each of the latent factors which are thus not ordered like principal components. In contrast to the above models, the observed variables are modelled as linear combinations of latent factors with a noise term. The noise term takes the form of an independent error variance of each of the variables. More generally, FA models are close to low-rank approximation procedures. FA has been particularly popular in psychological research (Bortz, 2013). It has for instance been used to investigate the building blocks of human intelligence assuming both an overarching general and several task-specific contributors (Spearman, 1904). However, FA enjoys much less popularity in current neuroscience research.

Relating fMRI task data to sets of macroscopical latent patterns can enhance both neurobiological interpretability and statistical tractability. Interpretability can be increased by analyzing and discussing fMRI results along functional integration accounts of brain organization. Tractability can be increased by translating the high-dimensional voxel space into a lower-dimensional network space to render ensuing statistical analyses better behaved. More concretely, capturing whole-brain data by a lower-dimensional network representations can avoid overfitting in various function-approximation scenarios. In neuroimaging, the BOLD signal is measured in voxel units of typically $1 \times 1 \times 1 - 3 \times 3 \times 3mm^3$. Consequently, BOLD signal measurements in gray matter can range between tens of thousands and hundreds of thousands voxels per brain scan. Even simple linear estimators are therefore prone to overfitting in classification and regression tasks due to the high number of features. This combined problem of data at hand and algorithm to apply worsen as a function of voxel number. Some Gaussian density, for instance, is described by mean and covariance matrix. In a 1D space, this Gaussian distribution can be described by a single value for the mean and one for the variance. In a 3D space, however, the mean takes 3 values (linear growth with number of features) and the 3×3 covariance matrix takes 6 unique values (polynomial growth with number of features). In much more input variables, such as in fMRI voxel images, the number of parameters to be estimated will grow quadratically with the number of input dimensions. Put differently, the variance of parameter estimates does also go up with increasing dimensions of the input space (the bias remaining constant). More input variables greatly increase the complexity capacity of the estimator's function space. A bigger hypothesis set increases overfitting and decreases the probability of extrapolating to unseen datasets. There is moreover a combinatorial explosion of optimal variable subsets to build a decision function. In short, operating in voxel space is a curse-of-dimensionality problem. One conceivable remedy to this case of high dimensionality is the existence of low-dimensional subspaces in fMRI signals. Prominent properties of functional brain organization might be describable, measurable, and predictable in units of macroscopical brain networks. If the estimated 'intrinsic dimensions' of a brain image indeed carry most meaningful information, it should be possible to recover a close proxy of the original whole-brain image from knowledge of the compressed network information alone. For instance, not maximal-variance latent components but independent latent components might be most neurobiologically pertinent.

Neurobiologically, it remains elusive how these neurophysiological network phenomena relate to an individual's repertoire of mental operations. This calls for methodological approaches that go beyond completely unsupervised analyses of linear combinations of latent components in the brain at rest (i.e., without controlled task modulation). Viewed from the perspective of the curse of dimensionality (cf. introduction), the complexity restrictions that are imposed by decomposition models applied to fMRI data include:

1. Neural activity in the brain can be explained by a small set of global patterns (i.e., compositionality).
2. These can extend across various discrete local activity patterns (i.e., known cytoarchitectonic brain regions).
3. The latent components derived from fMRI data describe distinct properties of underlying neurobiology.
4. The global patterns can be approximated by estimators that model linear combinations of n latent components.

The present investigation targets network-network dynamics as a whole by multivariate statistical learning. A multi-step framework capitalized on two independent, large datasets (n=500 and n=81) with 18 typical neuroimaging tasks (Fig. 7). Each of these two task batteries attempted to cover the diversity of the human cognitive apparatus. In an unsupervised learning step (i.e., naive to task labels), a small number of representative brain networks was derived from extensive neuroimaging resources. This first step yielded explicit models of possible network patterns with minimal statistical and cognitive assumptions. In a supervised learning step (i.e., based on task labels), the dimensionality-reduced neuroimaging data were then submitted to an 18-task classification problem. For each task, this second step determined a plausible combination of the modes of variation in brain signals. This quantitative association with traditional psychological concepts made the neuroimaging results human-interpretable. In a validation step, task activation patterns were recovered from the respective component loadings in the learned statistical models. This third step evaluated how well formal network models can generate realistic task activation patterns. Capitalizing on network-network modelling, the relationship between task-engaged and not experimentally controlled brain states was quantitatively revisited.

3.2 Methodological approach

Dataset: HCP task data

The first of two task batteries was drawn from the Human Connectome Project (HCP). HCP is an international long-term project dedicated to network exploration (Van Essen et al., 2012). 500 HCP participants (2 removed for quality reasons) were without psychiatric or neurological history. Informed consent was obtained

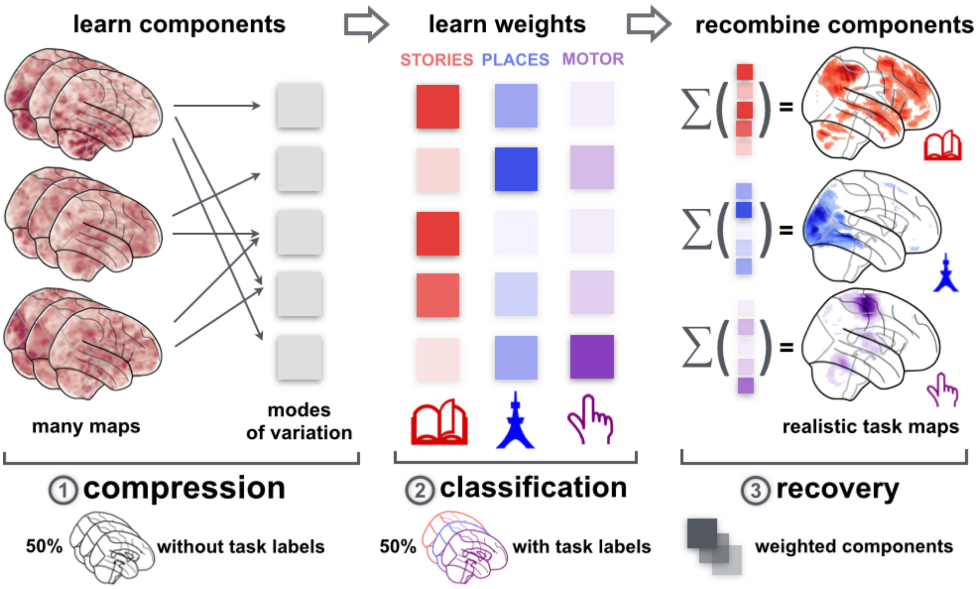

Figure 7: **Workflow**
(1) In two large neuroimaging datasets (HCP with n=500, ARCHI with n=81), the dominant spatial patterns of neural activity were discovered by decomposition methods. The repertoire of major neural networks in the human brain was hence computed without access to task information. (2) The explicit network definitions allowed reducing whole-brain activity maps from typical tasks into 40 component loadings per map. Statistical learning based on these biological features found a linear model to distinguish 18 tasks. A biophysiologically plausible network-network configuration was thus automatically derived for each task. (3) As face-validity, task activity maps were generated from the learned statistical models. The recovery performance quantified the biological meaningfulness of each model.

from all participants by the Washington University in St. Louis institutional review board. Network discovery is facilitated by probing experimental task paradigms that are known to tap on well-characterized neural networks. This was achieved by selecting established fMRI tasks that feature known suitability as localizers and reliability across participants (Barch et al., 2013). The HCP tasks were chosen by the external advisory board, consortium members, and HCP team from the National Institut of Mental Health (NIH). Over two image acquisition sessions, mostly block-design, but also event-related, paradigms were administered on 1) working memory/cognitive control processing, 2) incentive processing, 3) visual and somatosensory-motor processing, 4) language processing (semantic and phonological processing), 5) social cognition, 6) relational processing, and 7) emotional processing. All data were acquired on the same Siemens Skyra 3T scanner at Washington University. Whole-brain EPI acquisitions were acquired with a 32 channel head coil (TR=720ms, TE=33.1ms, flip angle=52, BW=2290Hz/Px, in-plane FOV=28 × 18cm, 72 slices, 2.0mm isotropic voxels). One task was run with right-to-left and one with left-to-right phase encoding. These analyses profited from the HCP minimally preprocessed pipeline (Glasser et al., 2013). This includes gradient unwarping, motion correction, fieldmap-based EPI distortion correction, brain-boundary-based registration of EPI to structural scan, non-linear (FNIRT) registration into MNI space, and grand-mean intensity normalization. The preprocessed maps were spatially smoothed by a Gaussian kernel of 4mm (FWHM). A general linear model (GLM) was implemented by FILM from the FSL suite with model regressors from convolution with a canonical hemodynamic response function and from temporal derivatives. It is important to note that HCP tasks were conceived to modulate activation in a maximum of different brain regions and neural systems. Indeed, excellent brain coverage was indicated by the summing the maps of whether or not a voxel showed a z-value bigger than 1.96 per task across participants (Barch et al., 2013). Note that statistical analysis was performed on the GLM-derived participant-level z-score maps in MNI space.

Dataset: ARCHI task data

The second task battery was drawn from the ARCHI dataset (Pinel et al., 2007). Importantly, it was conceived and acquired with the same goal for neural networks shared by a maximum of participants as the HCP task dataset. The diverse experimental tasks capture the cerebral basis of auditory and visual perception, motor action, reading, language comprehension and mental calculation. The approval was obtained from the regional ethical committee (Hopital de Bictre, France). 81 right-handed healthy participants (35 males and 46 females, 24 +/- 4.1 years, 3 not included in present analyses due to incomplete data) without psychiatric or neurological history participated in four fMRI sessions acquired under different experimental paradigms. The functional maps were warped in the MNI space and resampled at 3mm resolution. Contrasts were obtained using a GLM based on the specified task timing information, canonical hemodynamic response function, high-pass filtering and auto-regressive noise model. The tasks included 1) a *general localizer* paradigm that probes basic functions, such as button with the left or right hand, viewing horizontal and vertical checkerboards, reading and listening to short sentences, and mental computations (subractions); 2) a *social paradigm* that includes making covertly some inferences on short stories that involve false beliefs or not, viewing objects moving with or without a putative intention, listening to speech and non-speech sounds; 3) a *spatial mapping* paradigm that includes perfoming ocular saccades, grasping and orientation judgements on objects (the two different tasks were actually made on the same visual stimuli in order to charcterize graspin-specific activity), as well as judging whether a hand photograph was the left or right hand or was displaynig the pasl or back: again, the same input stimuli were presented twice in order to charcterize specific reponse to hand side judgement; 4) an *emotional paradigm* that include facial judgments of gender, trustworthiness and expression based on face photographs or photographs reduced to the eyes. While the first experimental paradigm is fast event-related design, the two others are made of small blocks of 5 to 7 seconds for each condition. The duration of the acquisitions was 1) 307s, 2) 489s, 3) 516s and 4) 436s. Visual stimuli were displayed in four 250ms epoches, separated by a 100ms intervals (i.e., 1.3s in total). Auditory stimuli were drawn from a recorded male voice (i.e., a total of 1.6s for motor instructions, 1.2-1.7s for sentences, and 1.2-1.3s for subtraction). The auditory or visual stimuli were shown to the participants for passive viewing or button response in event-related paradigms. Post-scan questions verified that the experimental tasks were understood and followed correctly. Based on these data, the following functional contrasts were computed and used in the following inference: 1) *basic localizer*, left versus right hand button, press, horisontal versus vertical checkerboard, auditory versus visual instructions, computation versus simple reading, motor tasks versus language and math tasks; 2) *social paradigm*: speech versus non-speech sounds, false belief- versus mechanistic kind of inference after visual or auditory presentation, interacting versus non-interacting figures in the movie clip; 3) *spatial mapping*: grasping versus orientation judgement, left/right hand versus hand side, effect of occular saccades; 4) *emotional paradigm*: gender judgement versus no task on face image and experssion, truthworthiness versus gender judgement and baseline. Whole-brain EPI data were acquired with the same Siemens Trio with a 32 channel head coil (TR=2400ms, TE=30ms, flip angle=60, in-plane FOV=19.2 × 19.2cm, 40 slices, 3.0mm isotropic voxels). A posterior-anterior phase encoding scheme was used for all images. Standard preprocessed was performed with Nipype (Gorgolewski et al., 2011), including slice timing, motion correction, alignment, and spatial normalization. Activation maps were spatially smoothed by a Gaussian kernel of 5mm (FWHM). Note that statistical analysis was performed on the GLM-derived participant-level z-score maps in MNI space, analogous to the HCP task data.

Dataset: HCP rest data

These two task datasets were complemented by HCP resting-state (RS) acquisitions of brain activity in the absence of experimental paradigms. In line with the goal of the present study, acquisition of these data was specifically aimed at the study of task-rest correspondence (Van Essen et al., 2012). The RS maps were acquired in two imaging sessions. Each participant contributed four time series (2 sessions each for left and right phase encoding) with 1,200 maps of multiband, gradient-echo planar imaging acquired during a period of 15min (TR=720 ms, TE=33.1 ms, flip angle=52, FOV=280 × 180mm, and 2.0mm isotropic voxels). Besides run duration, the task acquisitions were identical to the resting-state fMRI acquisitions for maximal compatibility between task and rest data. The unusually low TR allowed for effective increase in the signal-to-noise ratio. During map acquisition, participants fixated on a bright crosshair on a dark background. The analyses drew on "minimally preprocessed" rest data from 25 randomly selected healthy participants. Every participant who contributed scans to the present rest dataset was included with one left-right and one right-left session. Importantly, PCA was applied to each set of 1,200 rest maps for denoising and dimensionality reduction into 20 main modes of variation (Calhoun et al., 2001). That is, the 40 reduced rest maps from 25 randomly selected participants constituted the present rest dataset of 1000 concatenated, noise-cleaned rest activity maps.

In sum, the HCP task dataset incorporated 8,650 first-level activity maps from 18 diverse paradigms administered to 498 participants, while the ARCHI task dataset incorporated 1,404 first-level activity maps from

18 diverse paradigms administered to 78 participants. The HCP rest dataset, in turn, incorporated 1,000 activity maps drawn from different scanning sessions in 25 participants. The maps from all three datasets were downsampled to a common 53x63x46 space of 3mm isotropic voxels. The resampled activity maps were masked by a gray-matter probability of at least 10% according to tissue maps from the Intertional Consortium of Brain Mapping (ICBM). Each task or rest map was hence represented by 61,472 voxels representing Z values in gray matter.

Workflow

Data folding and model selection were performed in the following fashion. In the first step, one half of the task datasets (i.e., HCP and ARCHI) and the entire rest dataset were used for *unsupervised* (i.e., label-independent) discovery of latent structure by decomposition and clustering methods. The components of variation identified in the data allowed for feature engineering from biological structures. In the second step, *supervised* (i.e., label-dependent) classification algorithms were applied to the other half of the task activity maps to predict cognitive tasks. The winning models for classifying 18 cognitive tasks were compared by cross-validation. This framework is the gold standard to obtain an unbiased estimate of how well a trained classifier generalizes beyond the data samples at hand (Hastie et al., 2011; Pereira et al., 2009). The previously unseen half of the task data (i.e., 4,325 maps from HCP and 702 maps from ARCHI) were split into as many data folds as participants (i.e., 498 for HCP and 78 for ARCHI). In each fold, all task maps of a given participant were left out as test set for assessment of out-of-sample performance, while the task maps from the remaining participants served as training set for model estimation. In the third and last step, the averaged winning models were inverse-transformed into realistic task activity maps as face validity for the reduced representation of task activity patterns.

1. Unsupervised analysis layer

The first step uncovered hidden structure in large quantities of neural activity maps. To this end, each dataset (i.e., first half of task datasets and entirety of rest dataset) was decomposed into the 40 modes of variation. A large set of 40 “network archetypes” should ensure more parsimonious representation of the unknown target function, despite increased problem difficulty due to higher variance. Latent large-scale networks were computed by maximizing the intra-network homogeneity in neural activity and maximizing the between-network heterogeneity. Four different decomposition methods were selected that are in frequent general use.

Decomposition. Four different network decomposition techniques were used to test whether neuroscientific findings generalize across diverging methodological choices. First, *independent component analysis (ICA)* was used to unmix the multivariate BOLD signals into separate spatial components by minimizing their mutual information (Hyvärinen, 1999). This iterative blind source separation was realized by a parallel

FASTICA implementation (200 maximum iterations, per-iteration tolerance of 0.0001, initialized by a random mixing matrix, whitening). In neuroimaging research, ICA is frequently employed to separate out stable, statistically independent spatial patterns with identical time courses for the nodes of each extracted network (Beckmann et al., 2005). Second, *principal component analysis (PCA)* was used to capture the main directions of variation in the BOLD signals (Shlens, 2014a). This non-parametric, rotation-invariant method assumes orthogonal components to remove second-order dependencies (i.e., covariance) between the voxels by a change of basis (McKeown et al., 1998). An implementation of incremental PCA (Ross et al., 2008) performed batch-by-batch decomposition for low-rank approximation with efficient memory usage (batch size=100 activity maps, whitening). Third, *sparse PCA* was used to separate the BOLD signals into network components with few regions (Chennubhotla and Jepson, 2001), which scales well to large datasets. This decomposition method reformulates a PCA-like goal as a regression-type optimization problem constrained by an ℓ_1-penalty term. An implementation without orthogonality assumptions yielded spatially less distributed components to explain the variation in the data (1000 maximum iterations, per-iteration tolerance of $1 * 10^{-8}$, sparsity alpha=1, ridge-shrinkage at 0.01, Lasso path computed with coordinate descent) (Varoquaux et al., 2011). Fourth, *factor analysis (FA)* generalizes PCA by a heteroscedastic noise model (Bishop, 2006). Similar to PCA, FA performs a rotation-invariant low-rank approximation by identifying the latent space of variation (1000 maximum iterations, per-iteration tolerance of 0.01, 3 iterations in power method, randomized singular vector decomposition).

Note that all four decomposition methods implicitly assume large variance to be indicative of important structure. They reflect different ways to model sets of partially-overlapping major brain networks without access to any task information. All decompositions are soft in that each voxel with its BOLD signals can be part of the support in more than one network component. In most cases, the ensuing components are linear combinations of the original voxel signals.

Clustering. The decompositions into spatially overlapping network components were evaluated against clustering of the same data into non-overlapping regional components. That is, the similarity between voxels in neural activity changes across maps was indicative of coherent regional compartments, rather than distributed spatiotemporal networks. On the one hand, Ward clustering is a bottom-up hierarchical clustering approach (Johnson, 1967). A spatial constraint ensured that only neighboring voxels were incorporated into regional components of similar variation in BOLD signals (Thirion et al., 2014). On the other hand, k-means is a popular clustering approach (Lloyd, 1957). It divided the voxels into a preselected number of k regional components of variation without a spatial constraint (Nanetti et al., 2009).

Note that the clustering methods thus yielded a regional scaffold of functional brain architecture with emphasis on functional specialization, whereas the decomposition methods yielded a spatially distributed scaffold with emphasis on functional integration. The rationale was that four network-network models and two region-region models embodied quantitative proxies of functional integration versus functional specialization concepts of the human brain.

2. Supervised analysis layer

The extracted hidden components were subsequently used to reduce each task activity map to a much smaller number of component loadings. 61,472 voxels from each map's gray-matter voxels were thus condensed into 40 component loadings that quantify task-related network-network or region-region involvements. For network-network models, the presence of large-scale networks in every task map was determined by ridge regression based on a component design matrix (regularization alpha parameter=0.001, using Cholesky solver). 40 component loadings were thus computed by projection of each task map onto the 40 network components. For region-region models, 40 component loadings were computed by the cluster-wise mean of the BOLD signal. Task-related brain activity was thus summarized by brain regions of homogeneous activity changes across tasks. In both these approaches, the feature space for supervised classification had the form: #samples (i.e., 4,325 HCP task maps or 702 ARCHI task maps) × #features (40 component loadings). Linear support vector machines (SVM) were used to approximate the unknown target function that perfectly classifies the 18 cognitive tasks. This discriminative maximum-margin classifier was chosen for its simplicity, interpretability, and very good out-of-sample performance (Vapnik, 1996; Hanson and Halchenko, 2008). The feature space was preprocessed by normalization and mean-centering of each individual voxel due to sensibility of this procedure to scaling effects.

In particular, linear-SVM classification was performed with regularization by ℓ_1- and ℓ_2-penalty terms. Both shrinkage methods were used to search for parsimonious, more interpretable models (Hastie et al., 2011). First, ℓ_1-regularized SVM analyses conducted *i)* feature selection (i.e., select only relevant components) and *ii)* model estimation (i.e., determine what combination of components best disentangles the cognitive tasks) in an identical process. This approach introduced a maximum of zero-ed SVM weights for automatic determination of optimal model complexity. Please appreciate that prior dimensionality reduction (i.e., decomposition or clustering) yielded a set of new features that should be much less correlated than those of the initial voxel space (i.e., task and rest activity maps). From a neurobiological perspective, this approach acknowledged the assumption that smaller groups of, but not all, network components should be characteristic for the observed activity pattern of the various tasks. Hyper-parameter tuning was performed by a grid-search of "C" (7 steps between 10^{-3} and 10^{3}) and averaged the fold-wise model parameters (i.e., bagging; (Hastie et al., 2011)). The amount of sparsity was thus adapted to best predicting cognitive tasks. The collapsed SVM classification models thus represented the average hypothesis from that dataset with reduced variance but essentially unchanged bias. Second, the ℓ_2-regularized SVM analyses were used to shrink the model parameters *i)* towards zero and *ii)* towards each other. In so doing, all directions of variation were shrunken, but the low-variance directions were shrunken first. Comparing to ℓ_2-regularized

SVM, no implicit feature selection is performed as the ensuing parameters are typically small but non-zero. As explicit univariate component selection, ANOVA selected the most important k (i.e., 1, 5, 10, and 20) component loadings for each given task comparing to other tasks. That is, the multivariate classification problem was preceded by a univariate feature-selection procedure. In particular, the feature space (i.e., 40 component involvements) was reduced to the k component loadings most relevant for each task. The reduced feature space was subsequently fed into ℓ_2-penalized SVM to delineate their task-specific contribution. From a neurobiological perspective, this approach acknowledged the assumption that the balanced contribution of few large-scale networks should be sufficient to disentangle cognitive tasks.

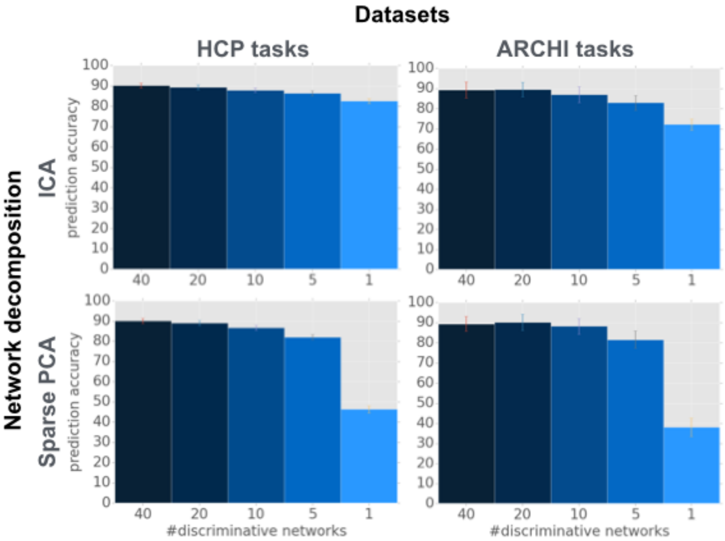

Figure 8: **Predictive accuracy of different network-network models**

Out-of-sample performance for 2 datasets and 2 decomposition methods as a function of considered network loadings per task. One half of the task data (i.e., 4,325 activity maps from HCP, 702 activity maps from ARCHI) were used for network discovery of 40 ICA and Sparse-PCA components. The network loadings of the previously unseen half of the task data (i.e., 4,325 HCP maps, 702 ARCHI maps) were then submitted to an 18-task classification problem. It was solved by ℓ_2-penalized SVM with automatic selection of the k most relevant networks for each task. To measure generalization performance, all task maps of one selected participant were left out in each cross-validation fold.

3. Validation layer

It was finally evaluated whether the obtained explicit models capture genuine properties of fMRI task activity. Complementing sparsity and predictive performance, this provides a face-validity criterion to disambiguate whether task-immanent aspects of neural activity or arbitrary descriminative aspects (e.g., structured noise, participant/scanner-related idiosyncracies) explain the models' success. To this end, artificial maps were generated from the explicit network-network or region-region models of each task. The 40 features describing large-scale network or region implications in a given task were transformed back into the original gray-matter voxel space with 61,472 voxels. In decomposition, the product was computed between the SVM coefficients and each component's support. In clustering, each region's mean activity was projected into each voxel of that region. Importantly, back-projection from an abstract, highly reduced space to the original task activity space should suceed as a function of the model's construct validity. For each of 18 tasks, Pearson's correlation between the model-derived activity map and the mean across first-level activity maps quantified the recovery performance. Consequently, the support recovery performance indexed the construct validity of quantitative network-network and region-region models.

Software implementation

Python was selected as scientific computing engine. Capitalizing on its open-source ecosystem helps enhance replicability, reusability, and provenance tracking. Nipy (Gorgolewski et al., 2011) performed basic analyses of functional neuroimaging data (`http://nipy.org/`). Scikit-learn (Pedregosa et al., 2011) provided efficient, unit-tested implementations of state-of-the-art statistical learning algorithms (`http://scikit-learn.org`). This general-purpose machine-learning library (Abraham et al., 2014) was interfaced with the neuroimaging-specific nilearn library for high-dimensional neuroimaging datasets (`http://github.com/nilearn/nilearn`). 3D visualization of brain maps was performed using PySurfer (`http://pysurfer.github.io/`).

3.3 Experimental results

It was formally tested whether neural activity patterns measured with fMRI in humans are largely explained by changes in cohesive network units. Whole-brain activity maps were expressed as a linear combination of 40 spatiotemporally coherent patterns (i.e., components). The distributed BOLD signals from 3D voxel space were thus reduced to 40 component loadings in a network space. This low-dimensional summary of the constellation of brain network involvements was the target of the present investigations.

The predictive accuracy of network-network models was first assessed across methodological choices and datasets. ICA and Sparse PCA were used to translate task activity maps into sets of network loadings. ℓ_1-penalized support vector machines (SVM) with C-parameter tuning on independent component analysis (ICA) and Sparse principal component analysis (PCA) loadings correctly detected 18 tasks in 90% to 93% of the time. Note that chance level is at 5.6% in an 18-class scenario. The recall ranged between 90% and 93%, while the precision ranged between 87% and 90%. ICA and Sparse PCA yielded higher model sparsity than other decomposition methods (i.e., PCA and factor analysis). Few network components turned out to be sufficient to correctly describe activity maps by the learned network sets. More generally, the observation of many zeros among the averaged model parameters suggests that the "true" decomposition of task activity maps is a linear combination of few network components. In sum, 18 diverse cognitive tasks could be very well distinguished solely based on 40 loadings of large-scale networks. The approach has been validated using four different decomposition methods in both HCP (Human Connectome Project) and ARCHI datasets. In

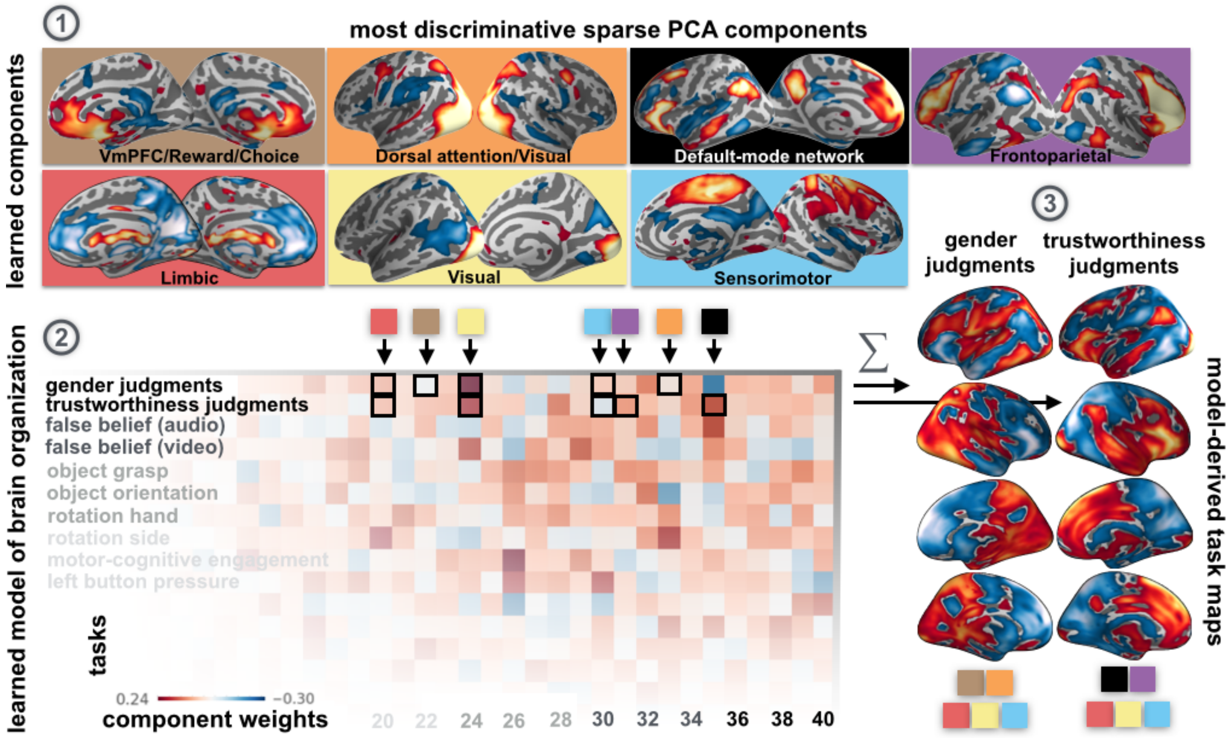

Figure 9: **Network-network composition underlying two experimental conditions from ARCHI**
(1) Examples of 40 large-scale networks drawn from task-unrelated resting-state fluctuations using Sparse PCA decomposition. This enabled translation of whole-brain task activity maps from the ARCHI dataset into 40 component loadings per map. (2) These measures of network implication served as basis for statistical learning of a quantitative model that disambiguates activity maps from the 18 task classes. In the depicted matrix, each weight represents the automatically determined importance of the corresponding large-scale network for a given cognitive task. Highlighted weights (*black framed*) correspond to the five networks (*colored boxes*) identified as most distinctive for the first two ARCHI tasks. (3) The explicit model is exploited to generate a whole-brain activity map for each of the 18 tasks. This is exemplified here by gender judgments and trustworthiness judgments on facial stimuli. Three resting-state networks (cf. 1) related to visual, emotional, and motor regions were selected as discriminative for both tasks (*red, yellow, blue boxes*). Each task was further associated with two specific large-scale networks. Consistent with published experimental fMRI studies (Bzdok et al., 2012a), the default mode network (*black box*), implicated in higher-order social processing, was significantly associated with trustworthiness judgments but not gender judgments on faces.

sum, the extraction of network sets by ICA and Sparse PCA achieved the highest prediction accuracies, precision and recall scores, as well as the most parsimonious model parameters. ICA and Sparse PCA hence identify underlying networks that allowed the most efficient quantification of distinctive neurobiological features.

After validating the proportional implication of large-scale networks as a salient property of task activity, the focus was on the interpretability of the network-network models (Fig. 8 and 11). The 18-task classification problem was solved by ℓ_2-penalized SVM of the selected k most important network loadings for each task (k =

40, 20, 10, 5, and 1). Importantly, the determined single most important ICA or Sparse-PCA network loading per task yielded classification success between 83% (FastICA, standard deviation [SD] computed across participant-wise data splits: 1.3) and 46% (SparsePCA, SD=1.7) for HCP as well as between 72% (FastICA, SD=3.0) and 38% (SparsePCA, SD=4.7) for ARCHI. Increasing the number of k discriminative networks per task rapidly saturated the predictive accuracy. The classification accuracy was virtually identical when knowing all 40 or the 20 most distinctive network loadings, still comparable to 10 and 5 loadings, in both datasets. In sum, experimentally evoked neural activity patterns can be well predicted by the relative implication of 5 large-scale networks per task.

It was then evaluated whether the classification models were fit for purpose. It was quantified to what extent the winning explicit models capture genuine properties of fMRI task activity. This provided a face-validity criterion to disambiguate whether task-immanent aspects of neural activity or arbitrary discriminative aspects (e.g., structured noise, participant/scanner-related idiosyncracies) explain the models' success. Whole-brain activity maps were thus generated from the network-network models for each task. The model-derived activity maps were then Pearson correlated with the mean first-level activity maps as a measure of support recovery performance. For ICA decomposition, the mean linear correlation across 18 tasks reached ρ=0.81 (HCP, SD=0.20) and ρ=0.88 (ARCHI, SD=0.07). For Sparse-PCA decomposition, in turn, mean correlations reached ρ=0.69 (HCP, SD=0.21) and ρ=0.70 (ARCHI, SD=0.15). These findings were confirmed against negative tests by pseudo network-network models derived from decomposition of Gaussian noise. The corresponding correlation analyses between first-level task maps and model-derived task maps ranged between ρ=0.32 (SD=0.07) and ρ=0.25 (SD=0.06). In sum, fMRI-data-derived neurobiological networks allowed significantly better reconstruction into the original voxel activity space than randomized fake network templates. As an alternative to network-network models, whole-brain task activity was also captured by region-region models (supplementary methods). Capturing task activity in local clusters units, rather than distributed network units, was less successful in reconstructing task-specific neural activity.

Finally, network-network models were put into practice by testing explicit hypotheses about functional brain architecture. It was formally tested whether given task activity patterns can be accounted for by network sets obtained from a diverging task battery and from the task-free resting brain (Fig. 9 and 10). To this end, recovery performance was compared between network sets learned from non-identical task data and from rest data. Learning network-network models from a non-identical task battery (i.e., networks from HCP task data and classification in ARCHI task data, or vice versa) yielded recovery performances between ρ=0.50 (SD=0.17) and ρ=0.43 (SD=0.17) across two decompositions and datasets. Importantly, learning network-networks models from task-unrelated resting-state correlations yielded very similar recovery performances between ρ=0.51 (SD=0.15) and ρ=0.46 (SD=0.11). Assessed by independent t-tests, recovery performance based on network sets from different task data was in no instance significantly better than recovery from networks discovered in task-unrelated rest data. This suggests that the network ecosystem of the mind-

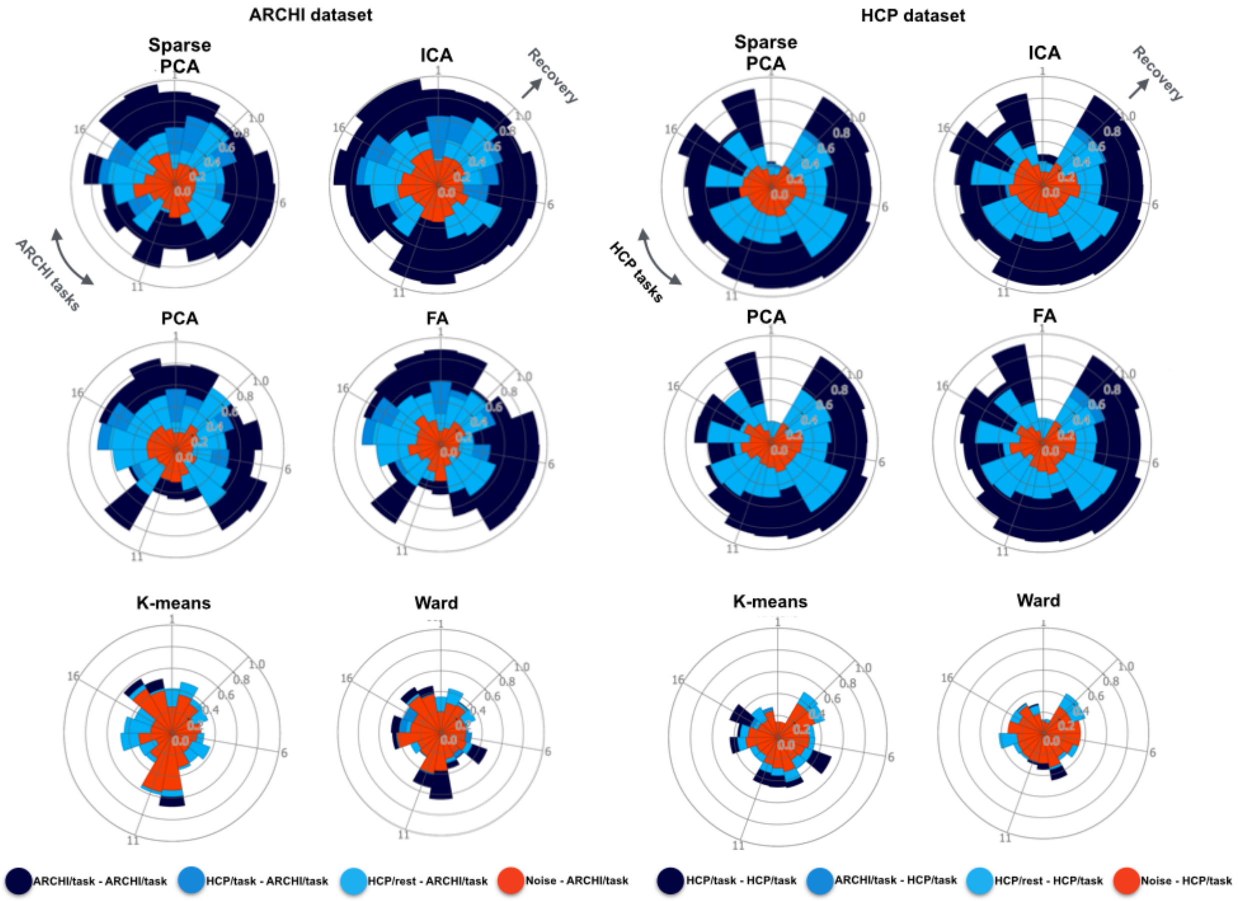

Figure 10: **Recovery performance of task activity patterns across decomposition and clusterings**
Four different network-network models (*upper and middle row*) were computed by decomposition based on sparse principal component analysis (sparse PCA), independent component analysis (ICA), PCA, and factor analysis (FA). They capture functional brain architecture with emphasis on functional integration as opposed to regional specialization. Two different region-region models (*lower row*) were computed based on ward clustering (regions always spatially compact) and k-means clustering (no spatial constraint). They capture functional brain architecture with emphasis on regional specialization. The recovery performance of all 18 tasks (*radial columns*) is measured by the Pearson correlation ρ between the model-derived task activity maps and the average first-level task map. As a first conclusion, modelling task-specific neural activity appears to be more successful based on functional network units than on functional region units. Additionally, network and region dictionaries were derived from i) identical task-data (as positive test; *dark blue*), ii) non-identical task-data (*medium blue*), iii) resting-state data (*light blue*), and iv) Gaussian noise (as negative test, *red*). As a second conclusion, task-derived and rest-derived network dictionaries are similarly successful in recovering whole-brain activity during divering experimental tasks.

wandering brain is recruited in response to specific environmental challenges.

3.4 Discussion

A principled approach is proposed to modelling brain activity during fMRI tasks in sets of network units. Typical activity patterns evoked by experimental tasks are shown to be combinations of standard net-

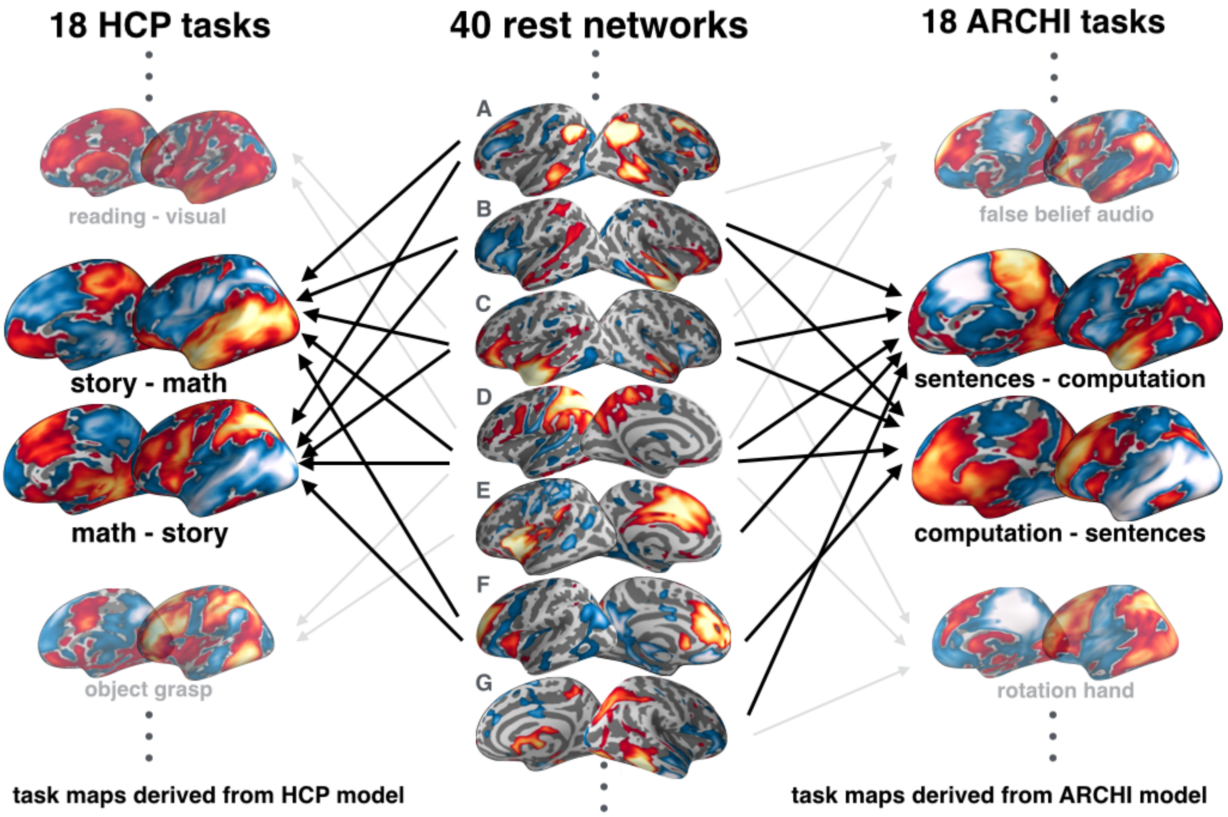

Figure 11: **Task-rest correspondence**

As a use case for network-network modelling, a fundamental question of brain organization was examined: the correspondence between task and rest architecture. *Middle column*: Examples of resting-state networks (RSN) derived from decomposition using sparse PCA of HCP rest data. RSN B and C might related to semantics processing in the anterior temporal lobe (Visser et al., 2010), RSN D covers extended parts of the parietal cortex, while RSN E and F appear to be variants of the so-called "salience" network (Seeley et al., 2007). *Left/Right column*: Examples of task-specific neural activity generated from network-network models of the HCP/ARCHI task batteries. The arrows indicate what rest networks were automatically ranked top-five in distinguishing 1 given task from the 17 other tasks. Although the task sets are different in HCP and ARCHI, "story versus math" and "sentences versus computation" were the most similar cognitive contrasts in both datasets. For these 4 contrasts, the model-derived task maps are highly similar. Consequently, 2 independent classification problems in 2 independent datasets with a 6-fold difference in sample size resulted in 2 independent explicit models that, nevertheless, generated highly comparable task-specific maps.

work patterns. Despite diverging model assumptions (i.e., orthogonality, sparsity, independence, and heteroscedasticity), all used decomposition methods allowed for generative models that quantify combinations of spatiotemporal activity patterns. Neurobiologically interpretable stratifications have thus been computed of large-scale network recruitment specific to cognitive tasks without recourse to cognitive theory. Decomposition and reconstruction of atomic network ensembles has been validated in two large task fMRI datasets, each aimed at capturing human cognition comprehensively. Although network-network dynamics are not directly observable using fMRI techniques, the underlying organization can be formally modeled in order to test explicit hypotheses.

The configuration of large-scale networks was identified here as a characteristic property of fMRI tasks. Decomposition into 40 networks allowed distinguishing 18 typical tasks in up to 90% (HCP dataset) and 93% (ARCHI dataset). Selecting only the loading of each task's single most distinctive of 40 total networks, classification still scored at up to 83% (HCP) and 72% (ARCHI). In particular, ICA and Sparse PCA appeared more successful than PCA and FA in searching the space of plausible network-network architectures. Neurobiological pertinence was indicated by overall best task classification scores, highest parsimony in model parameters, and model-derived recovery of most realistic task maps. This suggests that the underlying assumptions of these statistical models reverse-engineer critical properties of human network constellation. In this way, present results would favor large-scale patterns to parsimoniously distribute in space (reflected by SparsePCA models) and vary independently (reflected by ICA models), whereas orthogonality (reflected by PCA) and different variances (reflected by FA) seem less important. Notably, attempting reconstructing statistically coherent source signals operates with a network definition that is not strictly neurobiological (Mesulam, 2012). Nevertheless, a huge proportion of task activity can be explained by a combination of the brain's main functional networks. This suggests that contrast approaches of experimental fMRI studies do not necessarily elicit idiosycratic activity patterns, but mainly a recombination of the always same underlying networks. Indeed, a limited set of interacting entities generates an infinity of different observations in numerous natural systems, formalized in General System Theory (von Bertalanffy, 1950). This contention lends itself to a diverse set of neuroscientific questions.

The feasability of constructing and comparing network-network models was demonstrated in targeted experiments. The study juxtaposed the competing "dichotomy" and "manifold" hypotheses on the neuroarchitectural difference between task behaviors and the idlying brain (Buckner et al., 2008). The dichotomic view advocates functional antagonism between so-called "task-positive" and "task-negative" neural networks. Task-positive networks are believed to instantiate exteroceptive, environment-oriented mind sets to maintain task-constrained stimulus evaluation and response. Task-negative networks are believed to instantiate interoceptive, environment-detached mind sets to maintain adaptive mental imagery. The dichotomic view receives support from the following observations: *a*) the DMN consistently decreases in neural activity during various neuroimaging tasks (Shulman et al., 1997; Laird et al., 2009), *b*) activity in task-positive brain regions was consistently anti-correlated with activity in task-negative regions (Fox et al., 2005), and *c*) the spectrum of activity patterns in subcortical, limbic, and primary sensorimotor areas was richer at task then rest (Mennes et al., 2013). Hence, the dichotomic view predicts that component extraction from rest data will yield a dictionary of network definitions insufficient to delineate task-specific network compositions. In stark contrast, the "manifold" view advocates a fluctuating equilibrium of functionally distinct large-scale networks that is perturbed at task onset. This favors an identical ensemble of neural networks underlying functional brain architecture in active and resting brain states. This view, in turn, receives support from the following observations: *a*) seed-region-based resting-state correlations frequently recover task-typical networks (Biswal et al., 1995), *b*) separate decomposition of task and rest activity maps yielded a number of topographically similar networks (Smith et al., 2009), and *c*) only 11% of whole-brain connectivity patterns always shifted at onset of different tasks (Cole et al., 2014). Hence, the manifold view predicts that component extraction from rest and task will be similarly performant in delineating task-specific network compositions. This formally tested dichotomy against manifold hypothesis by learning independent network-network architectures based on non-identical task and rest data. Across four diverging decomposition methods, recovery performance from task-derived network dictionaries was in no instance significantly better than rest-derived network dictionaries. This indicator of equal network repertoires across task and rest favors the manifold perspective on task-rest correspondence. Consistent with Smith and colleagues (2009), explicit network-network models challenge the frequently embraced separation into "extrinsic" and "intrinsic" brain systems (Golland et al., 2007). Our approach reconciles the seemingly contradictory contention that activation patterns in certain brain areas are richer during task then rest (Mennes et al., 2013). The same set of canonical networks might give rise to task-evoked network compositions seldomly observed at rest. In sum, the conducted network-network experiment suggests that task-constrained onset of cognitive processes modulates an equivalent repertoire of large-scale brain networks.

Algorithmically, SPCA and ICA yielded higher reconstruction performances than PCA and FA across cognitive tasks and datasets. This suggests SPCA as a viable alternative to ICA-based image decomposition that currently has a monopol in neuroimaging research. Both SPCA and ICA thus appear equally useful for identification of latent brain networks in fMRI data and exploitation of such discovered macroscopical structures for feature engineering. What statistical properties might underlie the superiority of SPCA and ICA as well as the inferiority of PCA and FA?

Both PCA and FA relie on the covariance structure in the data to determine the latent components/factors. The relatively poorer performance suggests that second-order dependencies in fMRI are insufficient to allow optimal reconstruction of the original data from the low-dimensional 'network' representation. These statistical properties of models for network decomposition might not lend themselves particularly well to applications in imaging neuroscience. Enforcing maximal decorrelation of the discovered modes of variation might make it difficult to actually grasp 'function integration.' Several neurobiological activity patterns probably reflect higher-level processing that is rather remote from sensory input from the external world (Mesulam, 1998; Dehaene et al., 1998; Spreng et al., 2009). Such associative, integrative neural processes are naturally cognitive-domain-overarching and therefore implicated in a number of diverging mental tasks. PCA and FA however minimize the correlation of such higher-order, integrative processes with lower-level, more domain-specific neural processes manifested as spatiotemporally coherent activity pattern. Put differently, PCA and FA appear to impose a form of exclusivity on network-network implications that is unlikely to be neurobiologically valid. In stark contrast, the employed SPCA and ICA implementations allow more liberty for correlatedness between the discovered latent network components.

More specifically, the statistical property of orthogonality might have had particularly detrimental effects on fMRI analyses in both the temporal and spatial domain (Beckmann et al., 2005; Friston, 1998). Temporal coincidence of neurophysiological events might entail artificial coupling of neurobiological processes that are actually distinct. For instance, experimental-stimulus-induced and systematic-head-motion-induced neural activity could get fused into a single latent component. Similarly, externally triggered, simultaneous activity in the ventral and dorsal visual stream (Ungerleider and Mishkin, 1982; Ungerleider and Haxby, 1994) might be erroneously modelled as a common neurobiological process. Spatial coincidence of neurophysiological events will also be forced to belong to one or very few latent components. For instance, neural activity in the salience network (Yarkoni et al., 2011) and the default mode network (Laird et al., 2009) is differentially modulated by various cognitive tasks. Such task-overarching neurobiological processes are unlikely to be modelled and thus interpreted correctly by statistical models that impose an orthogonality constraint. In fact, PCA only requires the component number n as parameter choice and produces an answer regardless of the nature of the data-generating process. In neuroimaging practice, the orthogonality assumptions appears impertinent and arbitrary (Friston et al., 1999a) but is probably less deleterious in the spatial than the temporal dimension (Beckmann et al., 2005). That is, different resting-state networks with largely overlapping spatial structures (e.g., hemispherically homologous brain networks) can be cleanly decomposed as a function of the distinctness of their BOLD time series. More fundamentally, PCA elicits an intriguing discrepancy between its theoretical performance and its real-world performance in fMRI data. On the one hand, it can be mathematically proven that PCA provides the optimal data reduction under squared error loss (Shlens, 2014a). On the other hand, raw data recorded from nature hardly ever exhibit orthogonal directions of variation (Hastie et al., 2011). Experimental paradigms as designed by cognitive psychologists are an exception to this rule. They introduce artificial orthogonality into their measurements by contrasting experimental 'conditions' that intentionally diverge according to a variable of interest (Viviani et al., 2005). In sum, although PCA thus provides the most 'effective' compressed representation, it did not yield the most interpretable one in the present neuroimaging analyses.

In contrast, the independence assumption of ICA and the decorrelation-sparsity tradeoff of SPCA are less concerned by these caveats as they do not completely abolish correlation between the latent network representations. In this sense, the present results suggest that SPCA and ICA may constitute more neurobiologically valid models of network decomposition than PCA and FA. Put differently, one voxel can be more readily part of the support vector of more than one component (i.e., distinct brain networks) in ICA and SPCA than in PCA and FA. This appears to be the case although only ICA can restore the original signals, whereas a truthful reconstruction is less likely in SPCA due to the constraint for parsimony.

Apart from these many dissimilarities, ICA, SPCA, PCA, and FA all constitute latent variable models that describe fMRI time series by a linear mixture of separate data-generating processes. None of them can actually distinguish between between common variance and unique variance. While all four decomposition procedures reexpress the data as *linear* combinations of latent distributed variables, only the ICA procedure incorporates an element of *non-linearity* in its estimation process. As one important conclusion, due to the linearity of network modelling, they cannot investigate one of the key properties of the functional integration account of brain organization: the *interactions* between large-scale activation patterns in a statistical sense. That is, none of the used decomposition models can genuinely capture complicated effects *between* latent brain networks. This ignores that the brain is a highly non-linear system (Kandel and Schwartz, 2000). Nevertheless, ICA, SPCA, PCA, and FA have all been able to find some intrinsic low-dimensional subspaces in fMRI data that carry the majority of meaningful information. All four were able to recover a close proxy of the original whole-brain images from knowledge of the compressed network information alone. As another important conclusion, these results advocate a low *intrinsic dimensionality* of conventional fMRI data that has been evidenced here across two independent datasets and diverging model assumptions. Compression into a chosen network space can indeed recover voxel information in the whole brain from very low-dimensional spaces. Note however that none of the four latent variable models is able to indicate a 'right' number of network building blocks as they do not perform any explicit or implicit rank estimation. The 'true' number of basis functions is typically unknown in real-world settings, including neuroimaging research. Intriguingly, the reduced spaces found by PCA, SPCA, and ICA (potentially also FA) can be generalized by autoencoder architectures (cf. next chapter).

Taken together, systems neuroscience has transitioned the interpretational focus only recently from regions to networks. The composition of large-scale network recruitment qualified here as meaningful mechanism underlying fMRI experiments. Task-specific network compositions might thus intimately relate to psychological notions of mental operations. This organizational principle might extend to other species given existence of large-scale networks in monkeys (Mantini et al., 2011) and rats (Lu et al., 2012). Moreover, clinical research corroborated DMN dysfunction in various brain disorders (Broyd et al., 2009; Whitfield-Gabrieli and Ford, 2012). As a tempting alternative hypothesis for future research, not disturbance in the DMN itself but its relation to other superordinate networks might be disorder-specific. Ultimately, the present investigation exposes network-network architecture as an important neural mechanisms that maintains human cognition.

4 Semisupervised modelling for structure discovery and structure inference

4.1 Motivation

The 'true' intrinsic dimensionality of conventional fMRI data is unknown. Yet, the previous chapter provides pieces of evidence for dominating low-dimensional subspaces in BOLD signals. This is relevant to the neuroimaging investigator because the performance of SL models is strongly dependent on the chosen representation of the features. In cases of low a-priori certainty about the structure of the phenomenon (especially in case of brain-behavior and brain-concept correspondences), much engineering effort may be dedicated to data preprocessing pipelines. In practice, human-made representations are often used to manually reduce the data dimensionality. This preliminary procedure can considerably facilitate the model estimation process by enhancing the bias-variance tradeoff (i.e., typically small increase in model bias, much reduced variance due to lower model complexity) under the condition that the chosen feature representations are neurobiologically valid (cf. introduction). Yet, hand-crafted feature engineering does increase the work load and may be impossible for many neuroscientific questions due to unknown ground truth. An attractive alternative might therefore consist in casting the scenario as a *representation learning* problem. It is concerned with discovering the most pertinent feature configurations, agglomerations, and transformations to describe the explanatory variation in the data (Bengio et al., 2013). It revolves around the question where to assume probability mass in a high-dimensional setting. That is, a function is learned that maps from the input space into an (unknown) optimized feature representation space. Exploiting the discovered *manifolds* of condensed representation, the input data can be projected into a new, low-dimensional coordinate system that spans directions transversing the neighborhoods of highest probability mass (Bengio et al., 2013). It turns out that neural network models constitute a class of computational architectures (Schmidhuber, 2015) that lend themselves particularly well to *representation* and *manifold learning* tasks (Bengio, 2009; LeCun et al., 2015). Among such layered non-linear networks, *autoencoders* (AE) appear particularly attractive. Although the advantages of AE models have been known for several decades (Bourlard and Kamp, 1988; Hertz et al., 1991; Hinton and Zemel, 1994), the feasibility of effectively training them has been demonstrated only a few years ago (Hinton and Salakhutdinov, 2006). These unsupervised automatic feature extractors can approximate the manifold distributions dormant in high-dimensional BOLD signals as an alternative to parcellating brain region (chapter 2) and quantifying brain network stratifications (chapter 3). *Rather than making the a-priori assumption of brain division into discrete local features (corresponding to the notion of anatomical brain areas) or continuous global features (corresponding to the notion of functional brain networks), the local and/or global observational units are statistically estimated from the neuroimaging data.* The AE is a feedforward neural network that is typically composed of an input layer, one single hidden layer, and an output layer. Its prototypical architecture follows

$$\hat{\mathbf{x}} = \sigma(\mathbf{W_1}\sigma(\mathbf{W_0}\mathbf{x} + \mathbf{b_0}) + \mathbf{b_1}), \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ denotes the vector with d input variables per observations, $\mathbf{W_0}$ is the weight matrix for projection from input space into the hidden space (*encoder module*), $\mathbf{W_1}$ is the weight matrix for back-projection from these hidden units to the output space (*decoder module*), σ is a non-linear function (typically sigmoid $\sigma = \frac{1}{1+e^{-x}}$, possibly also hyperbolic tangent $\sigma = \frac{\sinh x}{\cosh x} = \frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}}$ or rectified linear unit $\sigma = \max(0, x)$), $\hat{\mathbf{x}} \in \mathbb{R}^{\mathbf{d}}$ is the corrupted reconstruction of $\mathbf{x}$, $\mathbf{b_0}$ and $\mathbf{b_1}$ are bias vectors. The number of hidden units h is reflected by the row number of the W_0 matrix, while the projection functions for each of the h hidden components are represented by each matrix row's values. Critically, a *bottleneck* with $h < d$ units imposes a rank constraint on $\mathbf{W_0}$. The encoder module is necessary for training and application of this model, whereas the decoder module can be discarded after model estimation. Point estimates of the model parameters $\{\mathbf{W_0}, \mathbf{W_1}, \mathbf{b_0}, \mathbf{b_1}\}$ of the AE are obtained by minimizing the reconstruction error

$$\mathbf{E} = \|\hat{\mathbf{x}} - \mathbf{x}\|^2. \tag{4}$$

The squared ℓ_2-norm of the discrepancy between AE input and output serves as objective function for empirical risk minimization. This simple learning model is capable of condensing diverse types of input data into local (brain-area-like) and global (distributed-pattern-like) representations via reconstruction under compression prior (Hinton and Salakhutdinov, 2006; Ranzato et al., 2006; Vincent et al., 2010). That is, a comprehensive representation is computed for the input in an unsupervised setting (Bengio et al., 2013; Olshausen and Field, 1996). The small bottleneck layer represents a latent code that is sufficient to recover an approximation to the original input data. A conversion is thus identified that allows translating high-dimensional data to a low-dimensional embedding. The ensuing transformation matrix maps from the original coordinate systems with many axes to a new coordinate system with few, but more explanatory, axes. The AE architecture thus allows for automatic determination of the most expressive manifolds in the data, including functional neuroimaging data. In fact, as theoretical results, adding additional computation layers to the AE does always improve a lower bound on the log probability assigned to training data under mild assumptions (Hinton et al., 2006). Addionally, a linear AE can be shown to be better with regard to the best possible training reconstruction error given a certain number of hidden units under mild assumptions[11]. Imposing sparsity constraints on the optimization problem, the regularized AE variants preclude the possibility of viewing the latent code as a low-dimensional coordinate system while maintaining the important directions of variation. Furthermore, viewing AEs are a single-hidden-layer neural network with a non-decreasing homogeneous activation function (Bach, 2014), it can be observed that *i)* the model is parameterized by a compact topological space, *ii)* is able to approximate any measurable target function (i.e., bigger function space than that of polynomials), *iii)* sparsity-inducing norms do not alter the optimization properties (since all norms are equivalent on $\mathbb{R}^n$), and *iv)* such norm-induced variable selection

[11] Mathematical proof by Hugo Larochelle in 'Neural networks: Autoencoder - linear autoencoder' (`www.youtube.com/watch?v=xq-I0Rl8mt0`).

can determine the number of hidden units automatically, while *v)* the model solves a parametric problem in case of constant unit numbers (i.e., model parameters do not increase with increasing sample size). In sum, AEs are unsupervised learning machines that represent a rare opportunity as automated preprocessors and feature extractors in medical image analysis, instead of picking image features by hand. This instance of flexible dimensionality reduction can be useful for feature engineering in the aim of visualization as well as classification and regression.

Another incentive for introducing AEs into neuroimaging research relies in their ability to emulate common unsupervised decomposition approaches, including PCA, SPCA, and ICA (cf. chapter 3). Indeed, the largest axes of variance identified by (truncated, since rank-reduced) PCA are also identified by AEs with one linear hidden layer and a squared error loss (Baldi and Hornik, 1989). While dimensionality reduction was traditionally based on PCA, generalized data decomposition by AE has outperformed PCA in a number of different unsupervised learning applications (Hinton and Salakhutdinov, 2006). In fact, the space spanned by the h hidden units of the AE is identical to the space spanned by the h principal singular values/vectors with possible rotional skewing. In contrast to PCA, the variance directions learned by the AE are typically unit-variance scaled and are not necessarily orthogonal. More specifically, the surface of the reconstruction error $\mathbf{E}$ of the linear AE equipped by squarred error loss exhibits unique local and global minima that reproduce the principal components of the training data (Baldi and Hornik, 1989). Subspaces determined by PCA also result from affine variants of this linear AE incorporating bias units for the encoding and decoder modules. Furthermore, the linear AE yields SPCA-like results if ℓ_1-shrinkage terms are added to the matrix and bias weights in the optimization objective. Besides searching the axes of main variances in the data, those solutions in the function space are encouraged that enable a maximum of sparsity (i.e., number of zero weights) despite optimized reconstruction. This model incorporates an intentional bias that encourages discovery of particularly robust input features. Moreover, AEs can also yield ICA-like results in case of tying the weight matrices $\mathbf{W_0}$ and $\mathbf{W_1}$ (i.e., $\mathbf{W_0} = \mathbf{W_1}^T$) as well as adding a non-linear convex function at the first layer (Le et al., 2011). When adding for instance a sigmoid non-linearity to the activations of the encoder layer, the AE can still operate in the PCA space by abiding by the linear regime of the sigmoid function, yet it can extend beyond the PCA-induced space as well (Bourlard and Kamp, 1988; Japkowicz et al., 2000). More generally, PCA-like linear AEs, SPCA-like sparse linear AEs, and ICA-like non-linear AEs can be shown to be mathematically equivalent under certain conditions (Le et al., 2011). Further, all three can also be viewed as directed graphical models (Wainwright and Jordan, 2008). The AE would sample from the learned low-dimensional factors (i.e., rows of $\mathbf{W_0}$ matrix) and subsequently obtain the input with added noise. Yet, the three AE architectures diverge regarding the bias imposed on the hidden layer and the inference of that layer. From the perspective of dictionary learning (Mairal et al., 2010), the first layer represents projectors to the discovered set of basis functions which are combined by the second layer to perform reconstruction of the input data (Bach, 2014; Olshausen and Field, 1996). Further conceptual accesses to understanding the mechanisms of AE models include for instance generative models and information theory (Vincent et al., 2010, 2008).

Besides solving representation learning tasks by flexible projection onto the main modes of variation in a heuristic fashion, AEs have very recently shown good performance in *semisupervised learning tasks*. The semisupervised regime aims at better classification and regression by using (an abundance of) unlabeled data in addition to the actually labeled data. In contrast to purely supervised learning techniques, they benefit from unlabeled data samples during model estimation. The main motivation relies in the usually low availability and high costs of meticulously labeling data, while, in many domains, unlabeled data can be collected fast, with low costs, and without domain expertise. For instance, a variational Bayesian AE enabled semisupervised generative models implemented in deep neural networks that achieved 97.41% accuracy based on only 600 labeled and 49,400 unlabeled training samples from the MNIST dataset of digit images[12] (Kingma and Welling, 2013; Kingma et al., 2014). Similarly, semisupervised deep autoencoders conditioned by a bottleneck with 10 'supervised' and 2 'unsupervised' layer units achieved up to 99.06% accuracy on the MNIST dataset (Cheung et al., 2014). These semisupervised classification scores are promising given that the best classification performances are hardly better using state-of-the-art supervised models on the fully labeled dataset. Conceptually, semisupervised learning can be divided into *transductive learning* and *inductive learning*. Transductive learning is concerned with label inference for the unlabeled samples among the provided partly labeled data, whereas inductive learning is concerned with inference of a mapping from data to labels. The distinction can be viewed from the perspective of a class-room analogy where a teacher presents some solved examples for an upcoming exam but also provides some unsolved examples (Zhu and Goldberg, 2009). Transductive learnings would pertain to solving the remaining examples as a take-home exam, whereas inductive learning would pertain to improving on the unsolved examples as they are typical of the upcoming in-class exam. Viewed from the perspective of the curse of dimensionality (cf. introduction), the complexity restrictions that are imposed by semisupervised AE architectures (Zhu, 2006; Seeger, 2000; Kingma et al., 2014) include:

1. $P(X)$ and $P(y|X)$ do share distribution parameters (unlabeled data are not necessarily beneficial for a given supervised learning problem).
2. A low-dimensional manifold exists in the data at hand (i.e., compositionality).
3. Not all variables of X are necessary to obain an optimized solution to the learning problem (parameterized by sparsity-inducing ℓ_1-norms).
4. If two data samples exhibit a small distance on the low-dimensional manifold, they should be identically labeled.
5. The identically labeled data samples form hard clusters (rather than soft, partly overlapping clusters).

[12] `http://yann.lecun.com/exdb/mnist/`

Taken together, semisupervised representation learning by AEs is able to tell class-relevant directions of variation apart from class-irrelevant ones.

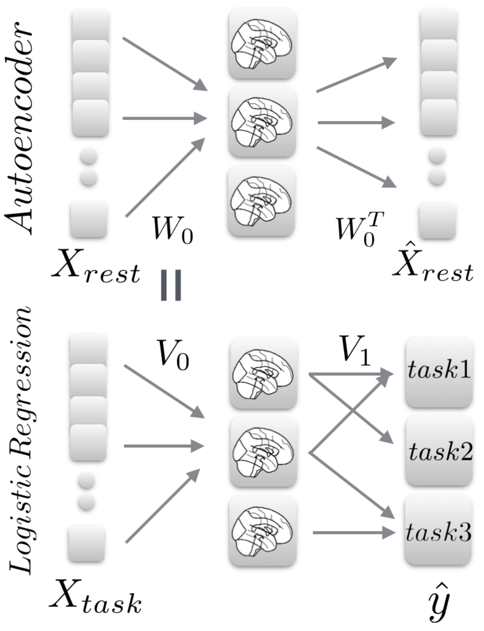

Figure 12: **Model architecture** Linear autoencoders find an optimized compression of 79,941 brain voxels into n unknown activity patterns by improving reconstruction from them. The decomposition matrix equates with the bottleneck of a factored logistic regression. Supervised multi-class learning on task data (X_{task}) can thus be guided by unsupervised decomposition of rest data (X_{rest}).

Indeed, existing analysis methods for neuroimaging research are almost exclusively concerned with either discovering neurobiological structure or assessing the neural correlates associated with mental tasks. To *discover structure*, on the one hand, spatial distributions of neural activity structure across time, ICA is often used in fMRI studies (Beckmann et al., 2005). It decomposes the BOLD signals into the primary modes of variation. Similarly, SPCA has been used to separate BOLD signals into parsimonious network components (Varoquaux et al., 2011). Network discovery by applying ICA or SPCA is typically performed on task-unrelated (i.e., *unlabeled*) "resting-state" data. These capture brain dynamics during ongoing random thought without controlled environmental stimulation. In fact, a large portion of the BOLD signal variation is known not to correlate with a particular behavior, stimulus, or experimental task (Fox and Raichle, 2007). To *test structure*, on the other hand, the neural correlates underlying mental tasks, the GLM is the dominant approach (Friston et al., 1994). The contribution of individual brain voxels is estimated according to a design matrix of experimental tasks. Alternatively, psychophysiological interactions (PPI) elucidate the influence of one brain region on another conditioned by experimental tasks (Friston et al., 1997). As a last example, an increasing number of neuroimaging studies model experimental tasks by training classification algorithms on brain signals (Poldrack et al., 2009). All these methods are applied to task-associated (i.e., *labeled*) data that capture brain dynamics during stimulus-guided behavior. As an important conclusion, existing supervised neuroimaging analyses cannot exploit the abundance of easily acquired resting-state data (Biswal et al., 2010). *These may allow better discovery of the manifold of brain states due to the increasingly acknowledged high task-rest similarities of neural activity patterns* (Smith et al., 2009; Cole et al., 2014). The integration of brain-network discovery into supervised classification can yield a semi-supervised learning framework. The most relevant neurobiological structure should hence be identified for the prediction problem at hand. In this last of three studies, an autoencoder will be fit to (unlabeled) rest data and integrated as a rank-reducing bottleneck into a multinomial logistic regression fit to (labeled) task data. This can solve a compound statistical problem of unsupervised data representation and supervised classification that have been previously studied in isolation. Neurobiologically, this allows delineating a low-dimensional manifold of brain network patterns and then distinguishing mental tasks by their most discriminative linear combinations. Theoretically, a reduction in model variance should be achieved by resting-state autoencoders that privilege the most neurobiologically valid models in the hypothesis set. The reduced hypothesis set entails a concomitant reduction in VC dimensions, which increases the probability for out-of-sample generalization (Vapnik, 1996). Practically, semisupervised learning can address the fact that neuroimaging research frequently suffers from scarcity of labeled data. This has limited the spectrum of representations that has been extracted from GLM analyses based on few-participant samples.

4.2 Methodological approach

A hybrid model combines an autoencoder and factored logistic regression in a semisupervised regime (Fig. 12). Different but complementary neurobiological properties can thus be conjointly exploited in an approach that is

1. *mixed:* by using rest and task data,
2. *factored:* by performing brain network decomposition, and
3. *multi-task:* by capitalizing on neural representations shared across mental operations.

Data. As the currently biggest openly-accessible reference dataset, the Human Connectome Project (HCP) was chosen as data resource (Barch et al., 2013). Neuroimaging task data with labels of ongoing cognitive processes were drawn from 500 healthy HCP participants (cf. methods section of chapter 3). The HCP data have been used exclusively in standardized form (i.e., mean-centering and unit-variance scaling) given that it is often useful, sometimes necessary and virtually never detrimental (Kuhn and Johnson, 2013). No further preprocessing has been applied on the HCP data. 18 HCP tasks were selected that are known to elicit reliable neural activity across participants (Table 1). In sum, the HCP task data incorporated 8,650 first-level activity maps from 18 diverse paradigms administered to 498 participants (2 removed due to incomplete data). All maps were resampled to a common 60 × 72 × 60 space of 3mm isotropic voxels and gray-matter masked (at least 10% tissue probability). The supervised analyses were thus based on labeled HCP task maps with 79,941 voxels of interest representing z-values in gray matter.

These labeled data were complemented by unlabeled activity maps from HCP acquisitions of unconstrained resting-state activity (Smith et al., 2013). These reflect brain activity in the absence of controlled thought. In sum, the HCP rest data concatenated 8,000 unlabeled, noise-cleaned rest maps with 40 brain maps from each of 200 randomly selected participants.

Cognitive Task	**Stimuli**	**Instruction for participants**
1 Reward	Card game	Guess the number of a mystery card for gain/loss of money
2 Punish		
3 Shapes	Shape pictures	Decide which of two shapes matches another shape geometrically
4 Faces	Face pictures	Decide which of two faces matches another face emotionally
5 Random	Videos with objects	Decide whether the objects act randomly or intentionally
6 Theory of mind		
7 Mathematics	Spoken numbers	Complete addition and subtraction problems
8 Language	Auditory stories	Choose answer about the topic of the story
9 Tongue movement	Visual cues	Move tongue
10 Food movement		Squeezing of the left or right toe
11 Hand movement		Tapping of the left or right finger
12 Matching	Shapes with textures	Decide whether two objects match in shape or texture
13 Relations		Decide whether object pairs differ both along either shape or texture
14 View Bodies	Pictures	Passive watching
15 View Faces	Pictures	Passive watching
16 View Places	Pictures	Passive watching
17 View Tools	Pictures	Passive watching
18 Two-Back	Various pictures	Indicate whether current stimulus is the same as two items earlier

Table 1: **Description of psychological tasks to predict.**

A further focus relied on the utility of the optimized low-rank projection in one task dataset for dimensionality reduction in another task dataset. To this end, the HCP-derived network decompositions were used as preliminary step in the classification problem of another large sample. The ARCHI dataset (Pinel et al., 2007) provides activity maps from diverse experimental tasks, including auditory and visual perception, motor action, reading, language comprehension and mental calculation. Analogous to HCP data, the second task dataset thus incorporated 1,404 labeled, grey-matter masked, and z-scored activity maps from 18 diverse tasks acquired in 78 participants.

Linear autoencoder. The labeled and unlabeled data were fed into a linear statistical model composed of an autoencoder and dimensionality-reducing logistic regression. The affine autoencoder takes the input $\mathbf{x}$, projects it into a coordinate system of latent representations $\mathbf{z}$ and reconstructs it back to $\mathbf{x'}$ by

$$\mathbf{z} = \mathbf{W_0 x} + \mathbf{b_0} \qquad\qquad \mathbf{x'} = \mathbf{W_1 z} + \mathbf{b_1}, \tag{5}$$

where $\mathbf{x} \in \mathbb{R}^\mathbf{d}$ denotes the vector of $d = 79{,}941$ voxel values from each rest map, $\mathbf{z} \in \mathbb{R}^\mathbf{n}$ is the n-dimensional hidden state (i.e., distributed neural activity patterns), and $\mathbf{x'} \in \mathbb{R}^\mathbf{d}$ is the reconstruction vector of the original activity map from the hidden variables. Further, $\mathbf{W_0}$ denotes the weight matrix that transforms from input space into the hidden space (encoder), $\mathbf{W_1}$ is the weight matrix for back-projection from the hidden variables to the output space (decoder). $\mathbf{b_0}$ and $\mathbf{b_1}$ are corresponding bias vectors. The model parameters $\mathbf{W_0}, \mathbf{b_0}, \mathbf{b_1}$ are found by minimizing the expected squared reconstruction error

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{AE}}(\mathbf{x})\right] = \mathbb{E}\left[\|\mathbf{x} - \mathbf{W_1}(\mathbf{W_0}\mathbf{x} + \mathbf{b_0}) + \mathbf{b_1})\|^2\right]. \tag{6}$$

This reconstruction error criterion equates with maximizing a lower bound on the mutual information between input and the learned representation. Here $\mathbf{W_0}$ and $\mathbf{W_1}$ were chosen to be tied, i.e. $\mathbf{W_0} = \mathbf{W_1^T}$. Consequently, the learned weights are forced to take a two-fold function: That of signal *analysis* and that of signal *synthesis*. The first layer *analyzes* the data to obtain the cleanest latent representation, while the second layer represents building blocks from which to *synthesize* the data using the latent activations. Tying these processes together makes the analysis layer interpretable and pulls all non-zero singular values towards 1. Non-linearities were not applied to the activations in the first layer. Finally, imposing shrinkage terms on the optimization objective of this AE avoided learning the identity function as final solution that would trivially maximizes mutual information between $\mathbf{x}$ and $\mathbf{x}'$ (Vincent et al., 2010).

Factored logistic regression. The factored logistic regression model is probably best described as a variant of a multinomial logistic regression. Specifically, the weight matrix is replaced by the product of two weight matrices with a common latent dimension. The latter is typically much lower than the dimension of the data. Alternatively, this model can be viewed as a single-hidden-layer feedforward neural network with a linear activation function for the hidden layer and a softmax function on the output layer. As the dimension of the hidden layer is much lower than the input layer, this architecture is sometimes referred to as a 'linear bottleneck' in the literature. The probability of an input $\mathbf{x}$ to belong to a class $i \in \{1, \ldots, l\}$ is given by

$$P(Y = i|\mathbf{x}; \mathbf{V_0}, \mathbf{V_1}, \mathbf{c_0}, \mathbf{c_1}) = \text{softmax}_i(f_{\mathcal{LR}}(\mathbf{x})), \tag{7}$$

where $f_{\mathcal{LR}}(\mathbf{x}) = \mathbf{V_1}(\mathbf{V_0}\mathbf{x} + \mathbf{c_0}) + \mathbf{c_1}$ computes multinomial logits and $\text{softmax}_i(x) = \exp(x_i)/\sum_j \exp(x_j)$. The matrix $\mathbf{V_0} \in \mathbb{R}^{\mathbf{dxn}}$ transforms the input $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ into n latent components and the matrix $\mathbf{V_1} \in \mathbb{R}^{\mathbf{nxl}}$ projects the latent components onto hyperplanes that reflect l label probabilities. $\mathbf{c_0}$ and $\mathbf{c_1}$ are bias vectors. The loss function is given by

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{LR}}(\mathbf{x}, \mathbf{y})\right] \approx \frac{1}{N_{X_{task}}} \sum_{k=0}^{N_{X_{task}}} \log(P(Y = y^{(k)}|\mathbf{x^{(k)}}; \mathbf{V_0}, \mathbf{V_1}, \mathbf{c_0}, \mathbf{c_1}). \tag{8}$$

Layer combination. Importantly, the optimization problem of the linear autoencoder and the factored logistic regression are linked in two ways. First, their transformation matrices mapping from input to the latent space are tied

$$\mathbf{V_0} = \mathbf{W_0}. \tag{9}$$

Hence, a compression of the 79,941 voxel values into n unknown components was searched that represent a latent code optimized for *both* rest and task activity data. Second, the objectives of the autoencoder and the factored logistic regression are interpolated in the common loss function

$$\mathcal{L}(\theta, \lambda) = \lambda \mathcal{L}_{\mathcal{LR}} + (1 - \lambda) \frac{1}{N_{X_{rest}}} \mathcal{L}_{\mathcal{AE}} + \Omega(\theta). \tag{10}$$

In so doing, the combined model parameters $\theta = \{\mathbf{V_0}, \mathbf{V_1}, \mathbf{c_0}, \mathbf{c_1}, \mathbf{b_0}, \mathbf{b_1}\}$ were searched with respect to the (unsupervised) reconstruction error and the (supervised) task classification. On the one hand, minimization of negative log-likelihood $\mathcal{L}_{\mathcal{LR}}$ as loss ensures decomposition of the X_{task} data samples according to the most classification-relevant hidden factors. On the other hand, minimization of the reconstruction-error loss $\mathcal{L}_{\mathcal{AE}}$ pushes variation of the X_{task} data-sample projections towards the manifolds underlying X_{rest}. $\mathcal{L}_{\mathcal{AE}}$ is devided by $N_{X_{rest}}$ to equilibrate both loss terms to the same order of magnitude. $\Omega(\theta)$ represents an ElasticNet-type regularization (Zou and Hastie, 2005; Ng, 2004) that combines ℓ_1 and ℓ_2 penalty terms $\forall p \in \theta$.

Optimization. The parameters of the system were learned by gradient descent optimization (Boyd and Vandenberghe, 2004; Bubeck, 2014). These gradients were obtained by using the chain rule to backpropagate error derivatives, first, through the LR layer and, then, through the factored decomposition with AE. The *rmsprop* solver was chosen (Tieleman and Hinton, 2012) as refined variation of stochastic gradient descent and a minibatch alternative to rprop. Rmsprop dictates an *adaptive learning rate* for each model parameter by scaled gradients from a running average. Step-size adaption is performed as parameter-by-parameter updates with multiplicative rate increases/decreases if the last two gradient signs agreed/disagreed. Momentum was not added as it typically does not lead to improvements when using rmsprop (Tieleman and Hinton, 2012). The batch size was set to 100 (given much expected redundancy within X_{rest} and X_{task}), matrix parameters were initalized by Gaussian random values multiplied by 0.004 (i.e., gain), and bias parameters were initalized to 0.

The normalization factor and the update rule for θ are given by

$$\begin{aligned} \mathbf{v^{(t+1)}} &= \rho \mathbf{v^{(t)}} + (1 - \rho) \left(\nabla_\theta f(x^{(t)}, y^{(t)}, \theta^{(t)}) \right)^2 \\ \theta^{(t+1)} &= \theta^{(t)} + \alpha \frac{\nabla_\theta f(x^{(t)}, y^{(t)}, \theta^{(t)})}{\sqrt{\mathbf{v^{(t+1)}}} + \epsilon}, \end{aligned} \tag{11}$$

where f is the loss function computed on a minibatch sample at timestep t, $0 < \rho < 1$ constitutes the decay rate. ρ was set to 0.9 to deemphasize the magnitude of the gradient. Further, α is the learning rate and ϵ a global damping factor. The hyperparameter α was set to 0.00001 by prior studies and ϵ was set to 10^{-6}. Note that experimentation with other solvers (stochastic gradient descent, adadelta, and adagrad) was done but found that *rmsprop* converged faster and with similar or higher generalization performance.

Hints. In fact, the constraint by a rest-data autoencoder is closely related to the notion of *hints* (Abu-Mostafa, 1994). Rather than regularization in a strict sense, its purpose is to introduce prior information on *known* properties of the *unknown* target function f. Rather than only relying on input-output pairs in the learning process, the hypothesis set was narrowed to the biologically most plausible solutions. That is, the search space was reduced in a way that is compatible with the expected representation of BOLD activity signals.

Implementation. The analyses were performed in Python. We used *nilearn* to handle the large quantities of neuroimaging data (Abraham et al., 2014) and *Theano* for automatic, numerically stable differentiation of symbolic computation graphs (Bastien et al., 2012; Bergstra et al., 2010).

4.3 Experimental results

Serial versus parallel structure discovery and classification. It was first tested whether there is a substantial advantage in combining unsupervised decomposition and supervised classification learning. The approach was benchmarked against performing data reduction on the (unlabeled) first half of the HCP task data by PCA, SPCA, ICA, and AE ($n = 5, 20, 50, 100$ components) and learning classification models in the (labeled) second half by ordinary logistic regression. PCA reduced the dimensionality of the task data by finding orthogonal network components using a change of basis (whitening). SPCA separated the task-related BOLD signals into network components with few regions by a regression-type optimization problem constrained by ℓ_1 penalty (no orthogonality assumptions, 1000 maximum iterations, per-iteration tolerance of 10^{-8}, $\alpha = 1$). ICA performed iterative blind source separation by a parallel FASTICA implementation (200 maximum iterations, per-iteration tolerance of 0.0001, initialized by random mixing matrix, whitening). AE found a code of latent representations by optimizing projection into a bottleneck (500 iterations, same implementation as below for rest data). The second half of the task data was projected onto the latent components discovered in its first half. Only the ensuing component loadings were submitted to ordinary logistic regression (one hidden layer, $\ell_1 = 0.1$, $\ell_2 = 0.1$, 500 iterations). These serial two-step approaches were compared against parallel decomposition and classification by SSFLogReg (two hidden layers, $\lambda = 1$, $\ell_1 = 0.1$, $\ell_2 = 0.1$, 500 iterations). Importantly, all trained classification models were tested on a large, unseen test set (20% of data) in the present analyses. Across choices for n, SSFLogReg achieved more than 95% out-of-sample accuracy, whereas supervised learning based on PCA, SPCA, ICA, and AE loadings ranged from 32% to 87% (Table 2). Intentionally disregarding rest data here, fLR without AE ($\lambda = 1$) was used to test for the superiority of parallel over serial structure discovery and classification. This experiment indeed established the advantage of directly searching for classification-relevant structure in the fMRI data, rather than solving the supervised and unsupervised problems independently. This effect was particularly pronounced when assuming few hidden dimensions.

n	PCA + LogReg	SPCA + LogReg	ICA + LogReg	AE + LogReg	SSFLogReg
5	45.1 %	32.2 %	37.5 %	44.2 %	**95.7%**
20	78.1 %	78.2 %	81.0 %	63.2 %	**97.3%**
50	81.7 %	84.0 %	84.2 %	77.0 %	**97.6%**
100	81.3 %	82.2 %	87.3 %	76.6 %	**97.4%**

Table 2: **Serial versus parallel dimensionality reduction and classification.** Chance is at 5.6%.

Model performance. SSFLogReg was subsequently trained (500 epochs) across parameter choices for the hidden components ($n = 5, 20, 100$) and the balance between autoencoder and logistic regression ($\lambda = 0, 0.25, 0.5, 0.75, 1$). Assuming 5 latent directions of variation should yield models with higher bias and smaller variance than SSFLogReg with 100 latent directions. Given the 18-class problem of HCP, setting λ to 0 consistently yields generalization performance at chance-level (5.6%) because only the unsupervised layer of the estimator is optimized. At each epoch (i.e., iteration over the data), the out-of-sample performance of the trained classifier was assessed on 20% of unseen HCP data. Additionally, the "out-of-study performance" of the learned decomposition ($\mathbf{W_0}$) was assessed by using it as dimensionality reduction of an independent labeled dataset (i.e., ARCHI) and conducting ordinary logistic regression on the ensuing component loadings.

	n = *5*					n = *20*					n = *100*				
	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
Out-of-sample accuracy	*6.0%*	88.9%	95.1%	**96.5%**	95.7%	*5.5%*	97.4%	**97.8%**	97.3%	97.3%	*6.1%*	97.2%	97.0%	**97.8%**	97.4%
Precision (mean)	*5.9%*	87.0%	94.9%	**96.3%**	95.4%	*5.1%*	**97.4%**	97.1%	97.0%	97.0%	*5.9%*	96.9%	96.5%	**97.5%**	96.9%
Recall (mean)	*5.6%*	88.3%	95.2%	**96.6%**	95.7%	*4.6%*	**97.5%**	97.5%	97.4%	97.4%	*7.2%*	97.2%	97.2%	**97.9%**	97.4%
F1 score (mean)	*4.1%*	86.6%	94.9%	**96.4%**	95.4%	*3.8%*	**97.4%**	97.2%	97.1%	97.1%	*5.3%*	97.0%	96.7%	**97.7%**	97.2%
Out-of-study accuracy	*39.4%*	60.8%	54.3%	60.7%	**62.9%**	*77.0%*	79.7%	**81.9%**	79.7%	79.4%	79.2%	**82.2%**	81.7%	81.3%	*75.8%*

Table 3: **Performance of SSFLogReg across a coarse grid of model parameter choices.** Chance is at 5.6%.

Three noteworthy observations could be made (Table 3). First, the most supervised estimator ($\lambda = 1$) achieved in no instance the best accuracy, precision, recall, or f1 scores on HCP data. Classification by SSFLogReg is therefore facilitated by imposing structure from the unlabeled rest data. Second, the higher the number of latent components n, the higher the out-of-study performance with small values of λ. This suggests that the presence of more rest-data-inspired hidden components results in more effective feature representations in unrelated task data. Third, for $n = 20$ and 100 (but not 5) the purely rest-data-trained decomposition matrix ($\lambda = 0$) resulted in noninferior out-of-study performance of 77.0% and 79.2%, respectively (Table 3). This confirms that guiding model learning by task-unrelated structure extracts features of general relevance beyond the supervised problem at hand.

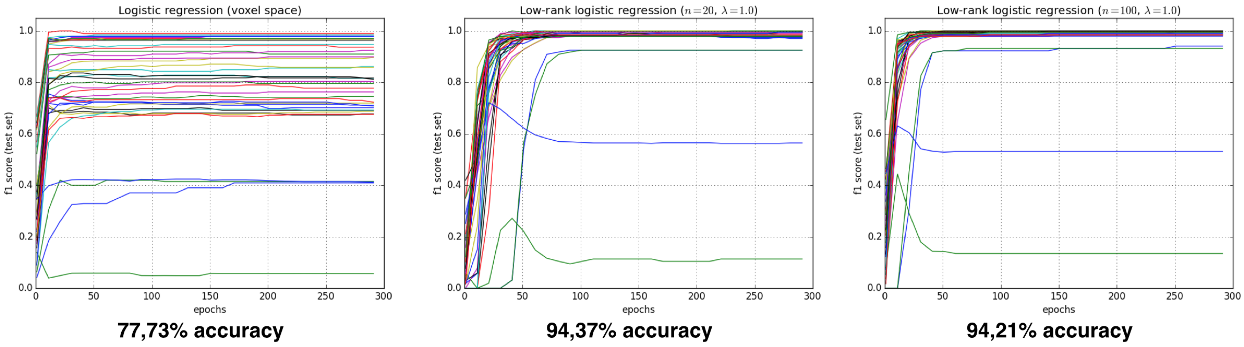

Figure 13: **Accuracy of 38-task w/o bottleneck classificaton across epochs** Depicts the f1 scores of 38 class predictions for each of 300 minibatch iterations (chance is at 2.6%). Multinomial logistic regression operating in voxel space without bottleneck (*left plot*) as well as with 20-component (*middle plot*) and 100-component (*right plot*) bottleneck. The curves show per-task f1 performance on the validation set. The three accuracy scores (*bottom*) reflect the percentage of correctly classified samples in the validation set by the final model. Hence, introduction of low-dimensional factors into the logistic regression model showed at least 15% better class separability.

Individual effects of dimensionality reduction and rest data. The impact of introducing a bottleneck layer was first quantified disregarding the autoencoder. To this end, ordinary logistic regression was juxtaposed with SSFLogReg with $\lambda = 1$. For this experiment, the difficulty of the classification problem was increased by including data from all 38 HCP tasks. Indeed, increased class separability in component space, as compared to voxel space, entails differences in generalization performance of $\approx 17\%$ (Fig. 13). Notably, the cognitive tasks on reward and punishment processing are among the least predicted with ordinary but well predicted with SSFLogReg (tasks 3 and 4 in Fig. 14). These experimental conditions have been reported to exhibit highly similar neural activity patterns in GLM analyses of that dataset (Barch et al., 2013). Consequently, also local activity differences (in the striatum and visual cortex in this case) can be successfully captured by brain-network modelling.

The impact of rest structure was then contemplated by modulating its influence ($\lambda = 0.25, 0.5, 0.75$) in data-scarce and data-rich settings ($n = 20$, $\ell_1 = 0.1$, $\ell_2 = 0.1$). At the beginning of every epoch, 2000 task and 2000 rest maps were drawn with replacement from same amounts of task and rest maps. In frequently encountered data-scarce scenarios, both training and validation scores improved as we depart from the most supervised model ($\lambda = 1$). This suggests that the autoencoder of SSFLogReg does not behave like a classical regularization but more like transfer learning. This effect was particularly pronounced in the data-scarce scenario, yet still present in the data-rich scenario (Fig. 15).

Feature identification. It was finally examined whether the models were *fit for purpose* (Fig. 16 and 17). To this end, Pearson's correlation was computed between the classifier weights and the averaged neural activity map for each of the 18 tasks. Ordinary logistic regression thus yielded a mean correlation of $\rho = 0.28$

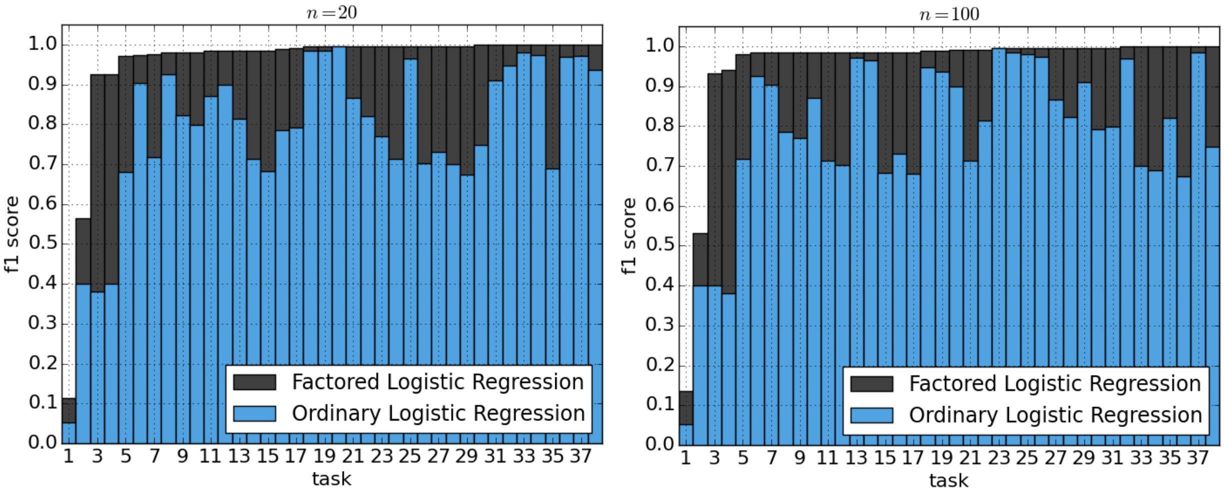

Figure 14: **Effect of bottleneck in a 38-task classificaton problem** Depicts the f1 score for each of 38 class predictions (chance is at 2.6%) sorted from lowest to highest. Multinomial logistic regression operating in voxel space (*blue bars*) was compared to SSFLogReg operating in 20 (*left plot*) and 100 (*right plot*) latent modes (*grey bars*). Autoencoder or rest data were not used for these analyses ($\lambda = 1$). Ordinary logistic regression yielded 77.7% accuracy out of sample, while SSFLogReg scored at 94.4% ($n = 20$) and 94.2% ($n = 100$). Hence, compressing the voxel data into a component space for classification achieves higher task separability.

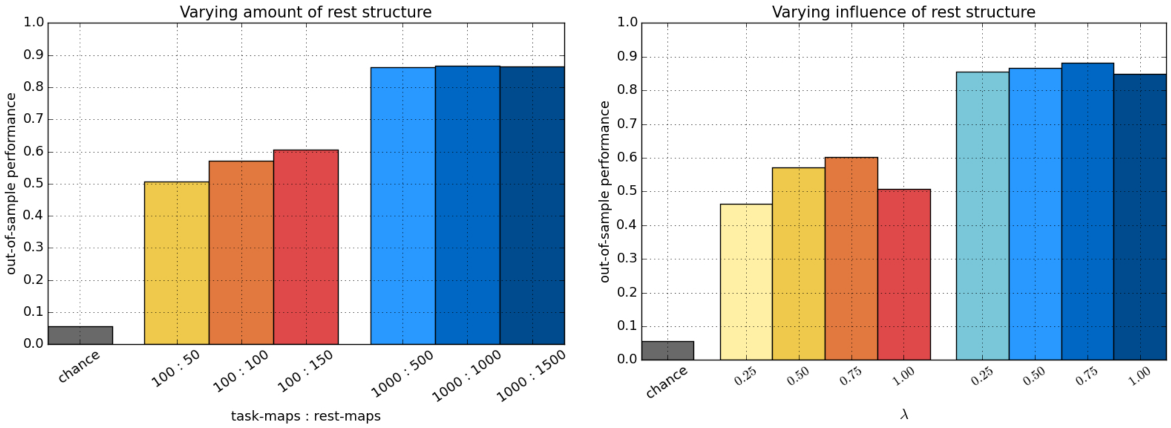

Figure 15: **Effect of rest structure in data-scarcity versus data-richness** Model performance of SSFLogReg ($n = 20$, $\ell_1 = 0.1$, $\ell_2 = 0.1$) as a function of varying amount (*left plot*) and influence (*right plot*) of rest data. In the data-scarce setting 100 task and 100 rest maps (*hot colors*) composed the dataset, whereas in the data-rich setting 1000 task and 1000 task maps were used (*cold colors*). To examine varying amount of rest structure, a given number of task maps was flanked by less, same or more rest maps. To examine varying influence of rest structure, the impact of the rest-autoencoder layer was modulated for different choises of λ. Gradient-descent optimization was always performed on 2000 task and 2000 rest maps drawn with replacement from the dataset. At the begining of each epoch, these were drawn with replacement from a pool of 100 or 1000 different task and rest maps, respectively. Especially the *cold* bars in the data-scarce scenarios suggests a 'transfer learning'-type effect (Pan and Yang, 2010). Chance is at 5.6%.

across tasks. For SSFLogReg ($\lambda = 0.25, 0.5, 0.75, 1$), a per-class-weight map was computed by matrix multiplication of the two inner layers. Feature identification performance thus ranged between $\rho = 0.35$ and $\rho = 0.55$ for $n = 5$, between $\rho = 0.59$ and $\rho = 0.69$ for $n = 20$, and between $\rho = 0.58$ and $\rho = 0.69$ for $n = 100$. Consequently, SSFLogReg puts higher absolute weights on relevant structure. This reflects an increased signal-to-noise ratio, in part explained by the more BOLD-typical local contiguity. Conversely, SSFLogReg puts lower probability mass on irrelevant structure. Despite lower interpretability of the results from ordinary logistic regression, the salt-and-pepper-like weight maps were sufficient for good classification performance.

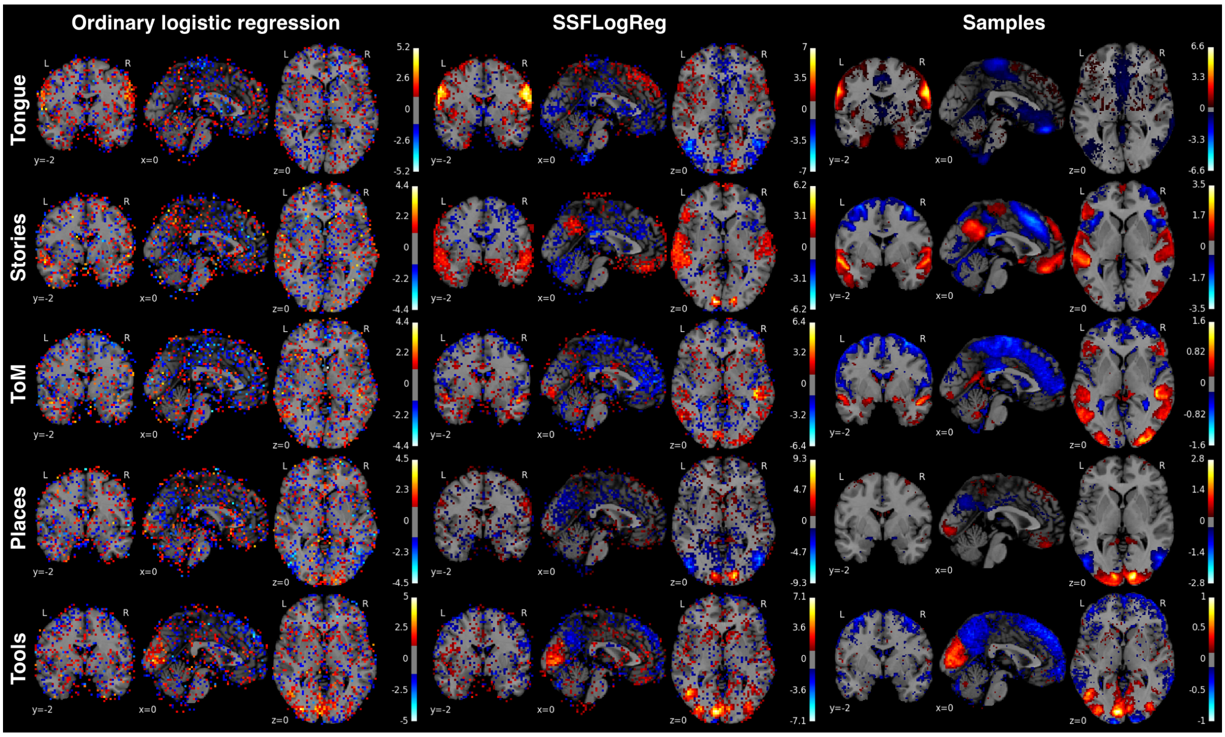

Figure 16: **Classification weight maps** The voxel predictors corresponding to 5 exemplary (of 18 total) psychological tasks (*rows*) from the HCP dataset (Barch et al., 2013). *Left column:* multinomial logistic regression (same implementation but without bottleneck or autoencoder), *middle column:* SSFLogReg ($n = 20$ latent components, $\lambda = 0.5$, $\ell_1 = 0.1$, $\ell_2 = 0.1$), *right column:* voxel-wise average across all samples of whole-brain activity maps from each task. SSFLogReg *a*) puts higher absolute weights on relevant structure, *b*) lower ones on irrelevant structure, and *c*) yields BOLD-typical local contiguity (without enforcing an explicit spatial prior). All values are z-scored and thresholded at the 75^{th} percentile.

Hence, SSFLogReg yielded class weights that were much more similar to features of the respective training samples for all choices of n and λ. SSFLogReg therefore captures genuine properties of task activity patterns, rather than participant- or study-specific artefacts.

Miscellaneous observations. For the sake of completeness, modifications of the statistical model are reported informally that did not improve generalization performance. *a*) Introducing stochasticity into model learning

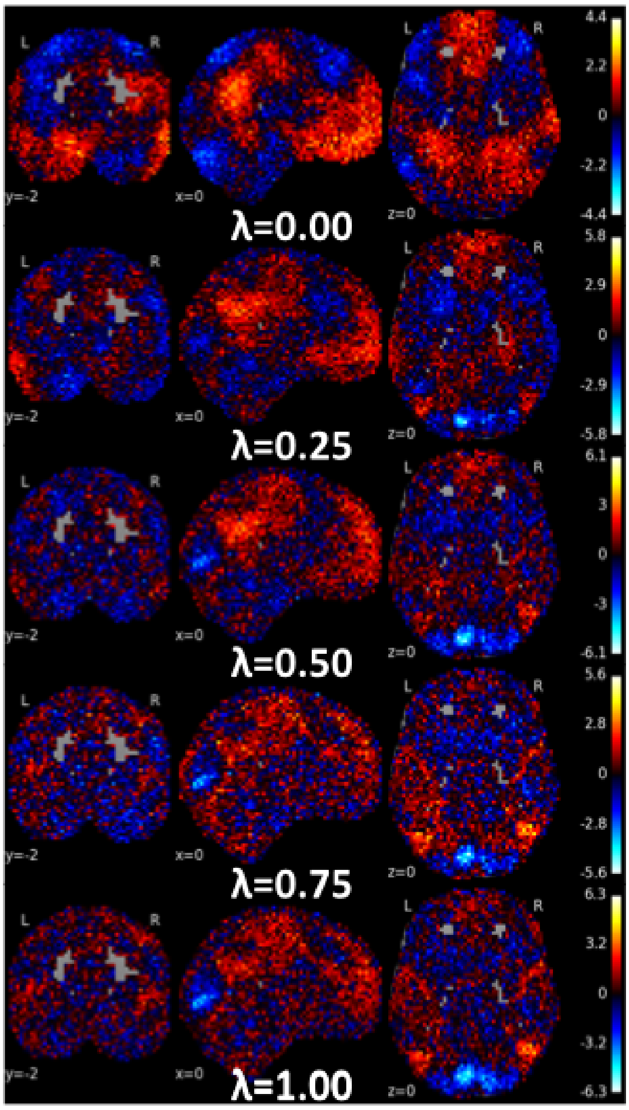

Figure 17: **Weight maps of a same hidden factor ranging from unsupervised to supervised regime** One of the n factors from the hidden layer ($\mathbf{W_0}$) was plotted for the same data (full HCP dataset) and the same model choices ($n = 20, \ell_1 = 0.1$, $\ell_2 = 0.1$) along a λ-grid between purely unsupervised ($\lambda = 0.0$, *top row*) and purely supervised ($\lambda = 1.0$, *bottom row*) settings. As qualitative evidence, a slow transition from rest- to task-typical brain networks was observed in brain space. Although difficult to quantify, rest network elements appear to get 'reassembled' to latent factors of the LR. This increased confidence that the improved model performance of rest-informed fLR is not only an arbitrary effect of spatially smooth noise. All values are z-scored.

by input corruption of $\mathbf{X_{task}}$ deteriorated model performance in all scenarios. Adding *b*) rectified linear units (ReLU) to $\mathbf{W_0}$ or other commonly used non-linearities (*c*) sigmoid, *d*) softplus, *e*) hyperbolic tangent) all led to decreased classification accuracies, probably due to sample size limits. Further, *f*) "pretraining" of the bottleneck $\mathbf{W_0}$ (i.e., non-random initialization) by either corresponding PCA, SPCA or ICA loadings did not exhibit improved accuracies, neither did *h*) autoencoder pretraining. Moreover, introducing an additional *h*) overcomplete layer (100 units) after the bottleneck was not advantageous. Finally, imposing either *i*) only ℓ_1 or *k*) only ℓ_2 penalty terms was disadvantageous in all tested cases. This favored ElasticNet regularization chosen in the above analyses.

4.4 Discussion

The field of imaging neuroscience is still largely divided into studies that *either* use unsupervised methods (especially PCA, SPCA, and ICA) to find 'resting-state networks' in task-unrelated data *or* use supervised methods (GLM, logistic regression, SVMs, etc.) in experimental data to locate the neural correlates underlying psychological processes (Smith et al., 2013; Barch et al., 2013). However, recent neuroimaging studies provide evidence for a close correspondence between resting-state correlations and task-constrained neural activity (Smith et al., 2009; Cole et al., 2014). It might be useful in a number of analysis settings to combine the unsupervised discovery of special structure and the supervised prediction of mental tasks. To this end, the last project has proposed an alternative model that facilitates identification of predictive brain regions of defined mental operations by conjoint learning in those modes of variation that are most relevant to the multi-task problem at hand. This model combines flexible brain-network discovery by AEs applied to rest data and mental task prediction enabled by factored logistic regression applied to task data.

In particular, three pieces of evidence attest to the advantages of the undertaken SL approach. First, the proposed AE/fLR-hybrid classified very well 18 psychological tasks from the HCP reference dataset. The out-of-sample performance ranged between 95.7% and 97.6% accuracy across choices for different numbers of latent variance components (n=5/20/50/100). These scores outperformed the *serial* approach that initially applies PCA/SPCA/ICA-reduction of rest data into the main components of variation and subsequent LR task classification on the component loadings, which yielde out-of-sample accuracies ranging between 32.2% and 87.3%. Second, the benefitting knowledge of rest structure has improved the universality of the learned factorization machines. That is, the already trained encoder layer was used for preprocessing whole-brain maps into dimensionality-reduced representations of an *independent* dataset. The task-separability was consequently enhanced in that second dataset with with out-of-dataset performance increases from 79.4% to 81.9% (n=20) and from 75.8% to 82.2% (n=100) in 18 *unseen* tasks. Third, the integrated AE/fLR has also considerably enhanced the feature identification. When computing the voxelwise Pearson correlation between the model weights and class maps averaged across participants, the correspondence was improved from r=0.28 (ordinary LR) up to r=0.55 (n=5), r=0.69 (n=20), and r=0.69 (n=100). Exploiting rest and task data by the AE/fLR-model thus improved the predictive weight maps for interpretability by cognitive neuroscientists, besides objective gains in model performance. In sum, recent neuroscientific evidence for similar neural activity patterns during task and rest were incorporated into a domain-specific learning model for joint dimensionality reduction and mental task classification. It enables mapping of traditional psychological concepts on combinations of brain networks shared across diverse psychological tasks.

Using the flexibility of factored models, the low-dimensional representation has been learned from high-dimensional voxel brain space that is most important for prediction of cognitive task sets. The higher generalization accuracy and feature recovery, comparing to ordinary logistic regression, corroborate the conclusion from chapter 3: fMRI data probably concentrate near a low-dimensional manifold of characteristic brain activity combinations. That is, the neural activation pattern underlying defined mental operations can be expressed by a linear combination of a small number of activity components. While the activation components had been static in that earlier study, the most relevant n activation components have here been learned for the supervised classification goal at hand. Besides yielding higher classification scores, SSFLogReg has therefore solved a representation learning problem as an *auxiliary learning task*. On the one hand, the characteristic combination of the learned components has been quantified for optimized prediction of each class of cognitive processes. On the other hand, however, a useful devision of brain activity into n continuous overlapping representation components has been obtained as new observational units for neural activity. In stark contrast to fitting linear classification models (e.g., logistic regression and SVMs) to brain voxels, the close integration of backpropagation and layered neural networks enables reverse-engineering of the most explanatory neurobiological representations for a given brain-behavior or brain-concept associations. Such representation learning can probably not be achieved by the unsupervised (e.g., PCA, SPCA, ICA, clustering algorithms) or supervised (e.g., GLM, psychophysiological interactions) statistical models that are currently available in the neuromaging field. From the perspective of cognitive neuroscience, the brain activity recruited during an individual's appraisal of face versus house visual stimuli, for instance, is probably best distinguished in locally constrained, different overlapping activation patterns in the bilateral fusiform face area of the visual cortex (Kanwisher et al., 1997; Gauthier et al., 1999). However, optimized representations to distinguish attention-related and language-related mental processes are more likely to distinguish global activation patterns of similar brain regions divided into the left and right brain hemisphere (Stephan et al., 2003; Corbetta and Shulman, 2002; Dronkers et al., 2004). As a final contrasting example, neurobiological distinction of high-level mental operations related to semantic processing (Mitchell et al., 2008) or processing others' mind states (Van Overwalle, 2009) is probably best achieved by bilateral macroscopical activation patterns. In this way, the proposed flexible *generative models* can formally grasp data-generating neurobiological mechanisms by estimating the parameters of the joint distributions $P(Y, X) = P(X|Y)P(Y)$. This is infeasable with *discrimintative models* that directly estimate the posterior probability $P(Y|X)$ of the unknown target function (Ng and Jordan, 2003; Bishop and Lasserre, 2007).

Additionally, SSFLogReg incorporates semisupervised components into the model structure (i.e., classification in partly labeled data). Importantly, the possible outcomes of adding unlabeled data samples to a supervised estimation problem range from beneficial to detrimental effects depending on the correctness of the model assumptions about $P(X)$ and $P(X|Y)$ (Zhu and Goldberg, 2009). In the present study, there are three pieces of evidence for the usefulness of rest data. First, the purely supervised fLR model ($\lambda = 1$, without autoencoder or rest data) achieved in no instance the best accuracy, precision, recall or f1 scores on HCP data across component choices (n=5/20/100). Second, the decomposition matrix from purely unsupervised autoencoder learning on rest data ($\lambda = 0$) was applied as a data transformation procedure in an independent dataset (ARCHI). This resting-state-activity-enabled dimensionality reduction for classifying 18 independent tasks increased the accuracies from 77% to 81.9% (n=20) and from 79.2% to 82.2% (n=100). Third, rest-informed classification has improved the out-of-sample performance by 10% in the data-scarce setting (100 task and 100 rest maps) and by 2% in the data-rich setting (1000 task and 1000 rest maps). These beneficial semisupervised outcomes therefore suggest correct model assumptions about the relationship between $P(x)$ and $P(x|y)$. But how does this application of semisupervision to neuroimaging data relate to the notions of *imputation* (i.e., completing missing labels for data samples), *transductive* learning (i.e., predict labels of unlabeled data items in the training set) and *inductive* learning (i.e., predict labels in unseen test data) (Zhu and Goldberg, 2009; Kingma et al., 2014; Zhu, 2006)? SSFLogReg can be viewed as an instance of imputation since it infers the missing labels of the held-out test set. Unlike typical semisupervised learning, however, the present approach does not assume that the labeled and unlabeled data samples have been sampled from the same data-generating process. That is, SSFLogReg does not assume $\mathbf{X_{task}}$ and $\mathbf{X_{rest}}$ to represent sampled observations of the same underlying population phenomenon. Consequently, SSFLogReg did not have the goal to infer labels of that differently generated part of the data (i.e., transductive semisupervision) but it does improve performance on a supervised task on unseen data (i.e., inductive semisupervision). Nevertheless, the present approach emphasizes formally incorporating existing knowledge on the general structure of the studied phenomenon and is thus probably less well described along the transductive-inductive distinction (Zhu, 2006). More globally, there is an increasing trend to initiate data collaborations and share large data sets from various neuroimaging methods (Poldrack and Gorgolewski, 2014; Frackowiak and Markram, 2015). The relative paucity of meticulously labeled brain maps stands in stark contrast to the increasing abundance of unlabeled brain maps. This available pool of massive imaging data from brain structure and function can be exploited by semisupervised learning techniques to improve classification and regression tasks by incorporating a-priori knowledge of normal neurobiology structure.

From a machine-learning perspective, factorization of the logistic regression weights can be viewed as transforming a *multi-class classification problem* into a *multi-task learning problem.* While 'multi-class classification' refers to categorizing unlabeled data samples into more than two classes (Bishop, 2006), 'multi-task learning' refers to better solving several related problems by exploiting their commonly shared structure (Caruana, 1997). The improved model performance suggests that many different psychological processes do recruit similar underlying macroscopial units of brain organization. Using the main variations in the data enables solving classification and regression problems along the main manifolds underlying the question at hand (Vincent et al., 2010). Automatized feature-engineering machines in form of autoencoder layers might therefore be a useful adoption in various neuroimaging analyses. Besides increased performance, these models are more interpretable by automatically learning a mapping to and from a brain-network space. This domain-specific learning algorithm encourages departure from the artificial and statistically less attractive voxel space. Neurobiologically, brain activity underlying defined mental operations can be explained by linear combinations of the main activity patterns. Flexible extraction of fundamental building blocks of brain organization might be a relevant stepping stone for discovering the building blocks (i.e., 'cognitive primitives') of human thought.

In the future, it might be interesting to extend SSFLogReg from a bottleneck layer capturing primarily manifestations of functional integration to a first hidden layer that can also capture manifestations of functional segregation of brain organization. That is, AEs have shown satisfactory performance in recovering general representational compartments in fMRI data. They are however naïve to the boundaries of microscopically distinguishable brain regions (cf. chapter 2). It has become established in neuroscience that the human brain is organized in discrete functional modules (Brodmann, 1909; Vogt and Vogt, 1919) that define widely accepted boundaries based on microarchitectural properties (e.g., cyto-, myelo- and receptorarchitecture). As one possible solution, a stacked hidden layer could contain one unit group learning brain-network-like projections and one unit group learning brain-region-like projections. The BOLD signal of a given voxel could thus contribute to both the explanatory spectrum of functional segregation and integration. Such a stacked bottleneck layer could be further improved by *structured sparsity* that inserts prior knowledge into the learning problem by encouraging practically likely sparsity patterns using penality terms (Huang et al., 2011). More specifically, one could impose group-wise sparsity (i.e., encourage dropping all region or all network representations) and within-group sparsity (i.e., encourage dropping a maximum of individual region/network units). The new hyperparameter that mediates between group-wise and within-group sparsity can be set by cross-validation with a coarse parameter grid, analoguous to common practice in sparse-group Lasso (Simon et al., 2013) and elastic-net regularization (Zou and Hastie, 2005). Moreover, as a natural extention of ℓ_1 penalties, the order of network patterns could be implicitly estimated by introducing a *trace norm* on the input decomposition matrix (Recht et al., 2010; Bach, 2008; Mishra et al., 2013). This penalty term sums the singular values as a convex surrogate for matrix rank in the aim of reducing the projection space to less basis functions. The ensuing complexity reduction should allow for more advanced variable selection by finding the most explanatory axes of the low-dimensional coordinate sytem. Neurobiologically, the most useful number of global network patterns could thus be determined automatically.

Taken together, automatic reduction of brain maps to their neurobiological essence may be essential to reduce computational costs and increase interpretability in the high-dimensional neuroimaging setting. After training, the encoder module allows very fast inference because data transformation into an optimized representation does not require any estimation procedure. Initiatives for data collection are rapidly increasing in neuroscience (Poldrack and Gorgolewski, 2014). These promise structured integration of neuroscientific knowledge accumulating in databases. Tractability by condensed feature representations can avoid the ill-posed problem of learning the full distribution of activity patterns. This is not only relevant to the multi-class challenges spanning the human cognitive space (Schwartz et al., 2013) but also the multi-modal combination with high-resolution 3D models of brain anatomy (Amunts et al., 2013) and high-throughput genomics (Need and Goldstein, 2010). The biggest socioeconomic potential may lie in across-hospital clinical studies that predict disease trajectories and drug responses in psychiatric/neurological populations (Frackowiak and Markram, 2015; Gustavsson et al., 2011).

5 Conclusion

Two major views on human brain organization are functional specialization by distinct cortical areas and functional integration by long-range connections of macroscopical networks. These complementary perspectives on brain architecture have been successfully recast in a machine-learning regime. Unknown target functions have been approximated by numerical optimization in the high dimensionsal scenario. Functional specialization was reframed as searching for discrete local patterns in neural activity fluctuations by clustering procedures, while functional integration was reframed as searching overlapping global activity patterns by matrix decomposition procedures. Both these neurobiologically motivated restrictions to complexity circumvented the curse of dimensionality and allowed for useful, simplified views on brain function. Finally, unsupervised structure discovery and supervised structure inference have been incorporated into a new, unified statistical estimator by means of autoencoders and neural network algorithms. The discovered existence of low-dimensional subspaces in fMRI data has been exploited for projection along the optimized directions of variation and supervised inference in the learned embedding space. Domain knowledge has thus improved the performance of supervised statistical models by imposing known biological structure on the estimation process. The statistical-learning access to brain imaging has allowed for a detailed characterization of the complicated relationship between the human brain in a goal-directed task state and an idlying rest state. Questions revolving around this task-rest correspondance may indeed not lend themselves particularly well to the cognitive-theory-driven experiments that typically dominate functional neuroimaging research. Ultimately, many unresolved challenges remain in systems neuroscience. Recent advances in heuristic gradient-based methods for approximating complex regularities in neurobiologial phenomena represent an attractive alternative to traditionally used purely analytical methods rooted in classical statistics. Neuroscientific questions can be reformulated as machine-learning problems and subsequently translated back into neuroscientific practice and interpretation. This inter-disciplinary research agenda has the potential to extend the spectrum of possible questions and permissible conclusions about the brain.

6 Acknowledgments

Many people and institutions have contributed tremendously to the present doctoral thesis. I am greatly indebted to my first supervisor Prof. Dr. Simon Eickhoff at the Faculty of Medicine, Heinrich-Heine University, Düsseldorf, Germany. The combination of tight supervision, unusual flexibility, and momentum allowed me to progress in a unique academic environment. Much thanks to the entire team at the Institute of Clinical Neurosciences and Medical Psychology, Düsseldorf, Germany, as well as at the Institute of Neuroscience and Medicine 1, Forschungszentrum Jülich, Germany, for their support, especially during my scarce presence when attending mathematics classes.

I am also greatly indebted to my second supervisor Prof. Dr. Stefan Conrad at the Faculty of Natural Sciences, Heinrich-Heine University, Düsseldorf, Germany. Thanks for being open to the interdisciplinary endeavor of supervising a PhD student with a medicine background, for allowing me to attend the stimulating data-mining classes, and for always being available to my questions.

As my third thesis supervisor, I am also particularly thankful to Dr. Bertrand Thirion at INRIA, Saclay, France. Thanks for readily hosting me as a visiting PhD student during three research stays in Parietal group at Neurospin, CEA, Gif-sur-Yvettes, France, the last of which lasted 12 months. Thanks for dedicating so much time to my many questions about statistics and the biology-algorithm interaction in imaging neuroscience. The routinely copious discussions with him as well as Olivier Grisel, Michael Eickenberg, and Elvis Dohmatob have enabled me to adopt a statistical *Denkstil*, which is invaluable. Gaël Varoquaux, Loïc Estève, and Alexandre Abraham never hesitated to come out of nowhere and tap on my fingers when Python code risked not to be fully PEP-8 compliant or to degrade towards fatal design choices. I was happy to be indoctrinated with the code of conduct in scientific computing in Python. Thanks to all the people at Parietal and Neurospin for the warm welcome and the fantastic atmosphere. All this would have been less fun without Régine Bricquet who rear-covered at the frontiers of paperwork and logistics.

Last but not least, I very much thank my family and friends for refreshing moments during the last three years and for understanding my decision against becoming a medical practitioner. Thanks to Paris for being such a marvellous place to exist. Moreover, this intellectual journey of intense interdisciplinarity would have been impossible without the financial and ideological backbone of the German National Merit Foundation.

7 References

A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*, 8:14, 2014.

Y. S. Abu-Mostafa. Learning from hints. *J Complex*, 10(1):165–178, 1994.

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H. T. Lin. *Learning from data.* AMLBook, California, 2012.

D. M. Amodio and C. D. Frith. Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci*, 7(4): 268–77, 2006.

K. Amunts, C. Lepage, L. Borgeat, H. Mohlberg, T. Dickscheid, M. E. Rousseau, S. Bludau, P. L. Bazin, L. B. Lewis, A. M. Oros-Peusquens, et al. Bigbrain: an ultrahigh-resolution 3d human brain model. *Science*, 340(6139):1472–1475, 2013.

K. Amunts, M. J. Hawrylycz, D. C. Van Essen, J. D. Van Horn, N. Harel, J. B. Poline, F. De Martino, J. G. Bjaalie, G. Dehaene-Lambertz, and S. Dehaene. Interoperable atlases of the human brain. *Neuroimage*, 99:525–532, 2014.

M. L. Anderson. Neural reuse: A fundamental organizational principle of the brain. *Behav Brain Sci*, 33(04):245–266, 2010.

M. L. Anderson, J. Kinnison, and L. Pessoa. Describing functional diversity of brain regions and brain networks. *Neuroimage*, 73:5058, 2013.

J. R. Andrews-Hanna, J. S. Reidler, J. Sepulcre, R. Poulin, and R. L. Buckner. Functional-anatomic fractionation of the brain's default network. *Neuron*, 65(4):550–62, 2010.

A. Anonymous. Focus on big data. *Nat Neurosci*, 17(11):1429–1429, 2014.

J. Ashburner and K. J. Friston. Unified segmentation. *Neuroimage*, 26(3):839–51, 2005.

J. Ashburner and S. Klöppel. Multivariate models of inter-subject anatomical variability. *Neuroimage*, 56(2):422–39, 2011.

F. Bach. Consistency of trace norm minimization. *J Mach Learn Res*, 9:1019–1048, 2008.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *arXiv preprint arXiv:1412.8690*, 2014.

P. Bado, A. Engel, R. Oliveira-Souza, I. E. Bramati, F. F. Paiva, R. Basilio, J. R. Sato, F. Tovar-Moll, and J. Moll. Functional dissociation of ventral frontal and dorsomedial default mode network components during resting state and emotional autobiographical recall. *Human Brain Mapp*, 35(7):3302–3313, 2014.

P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

M. Bar. The proactive brain: using analogies and associations to generate predictions. *Trends Cogn Sci*, 11(7):280–9, 2007.

D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, and C. Feldt. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

L. F. Barrett and A. B. Satpute. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Curr Opin Neurol*, 23(3):361–372, 2013.

F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

A. S. Beavers, J. W. Lounsbury, J. K. Richards, S. W. Huck, G. J. Skolits, and S. L. Esquivel. Practical considerations for using exploratory factor analysis in educational research. *Practical assessment, research & evaluation*, 18(6):1–13, 2013.

C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith. Investigations into resting-state connectivity using independent component analysis. *Philos Trans R Soc Lond B Biol Sci*, 360(1457):1001–13, 2005.

T. E. Behrens and H. Johansen-Berg. Relating connectional architecture to grey matter function using diffusion imaging. *Phil Trans R Soc B*, 360:903911, 2005.

T. E. Behrens, H. Johansen-Berg, M. W. Woolrich, S. M. Smith, C. A. Wheeler-Kingshott, P. A. Boulby, G. J. Barker, E. L.

Sillery, K. Sheehan, O. Ciccarelli, A. J. Thompson, J. M. Brady, and P. M. Matthews. Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat Neurosci*, 6(7):750–7, 2003.

T. E. J. Behrens, L. T. Hunt, and M. F. S. Rushworth. The computation of social behavior. *Science*, 324(5931):1160–1164, 2009.

R. Bellman. *Adaptive control processes: a guided tour*, volume 4. Princeton University Press, 1961.

Y. Bengio. Markovian models for sequential data. *Neural Computing Surveys*, 2:129–162, 1999.

Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

Y. Bengio. Deep learning and cultural evolution. In *Proceedings of the 2014 conference companion on genetic and evolutionary computation companion*, pages 1–2. ACM, 2014.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a cpu and gpu math expression compiler. *Proceedings of the Python for scientific computing conference (SciPy)*, 4:3, 2010.

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Database TheoryICDT99*, pages 217–235. Springer, 1999.

K. C. Bickart, C. I. Wright, R. J. Dautoff, B. C. Dickerson, and L. Feldman-Barrett. Amygdala volume and social network size in humans. *Nat Neurosci*, 14(2):163–164, 2011.

J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb Cortex*, 19(12):2767–96, 2009.

C. M. Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006. 4.

C. M. Bishop and J. Lasserre. Generative or discriminative? getting the best of both worlds. *Bayesian Statistics*, 8:3–24, 2007.

B. B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magn Reson Med*, 34(4):537–41, 1995.

B. B. Biswal, M. Mennes, X. N. Zuo, S. Gohel, and C. et al. Kelly. Toward discovery science of human brain function. *Proc Natl Acad Sci U S A*, 107(10):4734–9, 2010.

J. Bortz. *Statistik: Für Sozialwissenschaftler*. Springer-Verlag, 2013.

N. Bostrom. *Anthropic bias: Observation selection effects in science and philosophy*. Routledge, 2013.

H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

M. Brett, I. S. Johnsrude, and A. M. Owen. The problem of functional localization in the human brain. *Nat Rev Neurosci*, 3 (3):243–9, 2002.

P. Broca. Sur la faculté du language articulaire. *Bulletins et Memoires de la Societé d'Anthropologie de Paris*, 6:377–393, 1865.

K. H. Brodersen. Decoding mental activity from neuroimaging data the science behind mind-reading. *The New Collection, Oxford*, 4:50–61, 2009.

K. H. Brodersen, T. M. Schofield, A. P. Leff, C. S. Ong, E. I. Lomakina, J. M. Buhmann, and K. E. Stephan. Generative embedding for model-based classification of fmri data. *PLoS Comput Biol*, 7(6):e1002079, 2011.

K. Brodmann. *Vergleichende Lokalisationslehre der Groshirnrinde*. 1909.

S. J. Broyd, C. Demanuele, S. Debener, S. K. Helps, C. J. James, and E. J. Sonuga-Barke. Default-mode brain dysfunction in mental disorders: a systematic review. *Neurosci Biobehav Rev*, 33(3):279–96, 2009.

S. Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter. The brain's default network: anatomy, function, and relevance to disease. *Ann N Y Acad Sci*, 1124:1–38, 2008.

G. Buzsáki. *Rhythms of the Brain*. Oxford University Press, 2006.

G. Buzsáki and A. Draguhn. Neuronal oscillations in cortical networks. *Science*, 304(5679):1926–1929, 2004.

G. Buzsáki, N. Logothetis, and W. Singer. Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. *Neuron*, 80(3):751–764, 2013.

R. W. Byrne and A. Whiten. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford University Press, Oxford, 1988.

D. Bzdok, R. Langner, F. Hoffstaedter, B. I. Turetsky, K. Zilles, and S. B. Eickhoff. The modular neuroarchitecture of social judgments on faces. *Cereb Cortex*, 22(4):951–61, 2012a.

D. Bzdok, L. Schilbach, K. Vogeley, K. Schneider, A. R. Laird, R. Langner, and S. B. Eickhoff. Parsing the neural correlates of moral cognition: Ale meta-analysis on morality, theory of mind, and empathy. *Brain Struct Funct*, 217(4):783–796, 2012b.

D. Bzdok, R. Langner, L. Schilbach, O. Jakobs, C. Roski, S. Caspers, A. R. Laird, P. T. Fox, K. Zilles, and S. B. Eickhoff. Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. *Neuroimage*, 81:381392, 2013.

D. Bzdok, D. Gross, and S. B Eickhoff. The neurobiology of moral cognition: Relation to theory of mind, empathy, and mind-wandering. In *Handbook of Neuroethics*, pages 127–148. Springer, 2015.

V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human Brain Mapp*, 14(3):140–151, 2001.

R. T. Canolty, E. Edwards, S. S. Dalal, M. Soltani, S. S. Nagarajan, H. E. Kirsch, M. S. Berger, N. M. Barbaro, and R. T. Knight. High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313(5793):1626–1628, 2006.

T. Carlson, P. R. Schrater, and S. He. Patterns of activity in the categorical representations of objects. *J Cognitive Neurosci*, 15(5):704–717, 2003.

P. Carruthers. How we know our own minds: The relationship between mindreading and metacognition. *Behav Brain Sci*, 32 (2):1–62, 2009.

R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

F. Cauda, G. Geminiani, F. D'Agata, K. Sacco, S. Duca, A. P. Bagshaw, and A. E. Cavanna. Functional connectivity of the posteromedial cortex. *PLoS One*, 5(9), 2010.

A. E. Cavanna and M. R. Trimble. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129 (Pt 3):564–83, 2006.

G. Chaitin. The limits of reason. *Scientific American*, 294(3):74–81, 2006.

C. Chennubhotla and A. Jepson. Sparse pca. extracting multi-scale structure from data. *Eighth IEEE International Conference on Computer Vision, 2001*, 1:641–647, 2001.

B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

W. G. Clark, J. Del Giudice, and G. K. Aghajanian. *Principles of psychopharmacology: a textbook for physicians, medical students, and behavioral scientists*. Academic Press Inc, 1970.

M. Clos, K. Amunts, A. R. Laird, P. T. Fox, and S. B. Eickhoff. Tackling the multifunctional nature of broca's region meta-analytically: Co-activation-based parcellation of area 44. *Neuroimage*, 83C:174–188, 2013.

J. Cohen. The earth is round. pages 997–1003, 1994.

M. W. Cole, D. S. Bassettf, J. D. Power, T. S. Braver, and S. E. Petersen. Intrinsic and task-evoked network architectures of the human brain. *Neuron*, 83c:238251, 2014.

A. Connes. A view of mathematics. *available as a pdf file from http://www. alainconnes. org/en/downloads. php*, 2010.

M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci*, 3(3): 201–215, 2002.

M. Corbetta, G. Patel, and G. L. Shulman. The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58(3):306–24, 2008.

G. E. Dahl, N. Jaitly, and R. Salakhutdinov. Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*, 2014.

J. S. Damoiseaux, S. A. Rombouts, F. Barkhof, P. Scheltens, C. J. Stam, S. M. Smith, and C. F. Beckmann. Consistent resting-state networks across healthy subjects. *Proc Natl Acad Sci U S A*, 103(37):13848–53, 2006.

C. Daniel and F. S. Wood. Fitting equations to data. 1971.

P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural Comput*, 7(5):889–904, 1995.

S. Dehaene. A few steps toward a science of mental life. *Mind, Brain, and Education*, 1(1):28–47, 2007.

S. Dehaene. *The number sense: How the mind creates mathematics.* Oxford University Press, 2011.

S. Dehaene and L. Cohen. Cultural recycling of cortical maps. *Neuron*, 56(2):384–398, 2007.

S. Dehaene, M. Kerszberg, and J. P. Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci U S A*, 95(24):14529–14534, 1998.

S. Dehaene, M. Piazza, P. Pinel, and L. Cohen. Three parietal circuits for number processing. *Cogn Neuropsychol*, 20(3): 487–506, 2003.

J. J. Déjerine. *Anatomie des centres nerveux*, volume 1. Rueff, 1895.

P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

V. Doria, C. F. Beckmann, T. Arichia, N. Merchanta, M. Groppoa, F.E. Turkheimerb, S.J. Counsella, M. Murgasovad, P. Aljabard, R.G. Nunesa, D.J. Larkmana, G. Reese, and A. D. Edwards. Emergence of resting state networks in the preterm human brain. *Proc Natl Acad Sci U S A*, 107(46):20015–20020, 2010.

N. U. Dosenbach, K. M. Visscher, E. D. Palmer, F. M. Miezin, K. K. Wenger, H. C. Kang, E. D. Burgund, A. L. Grimes, B. L. Schlaggar, and S. E. Petersen. A core system for the implementation of task sets. *Neuron*, 50(5):799–812, 2006.

N. F. Dronkers, D. P. Wilkins, R. D. Van Valin, B. B. Redfern, and J. J. Jaeger. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1):145–177, 2004.

R. I. M. Dunbar and S. Shultz. Evolution in the social brain. *Science*, 317(5843):1344–1347, 2007.

W. H. Durham. *Coevolution: Genes, culture, and human diversity.* Stanford University Press, 1991.

B. Efron. *Modern science and the Bayesian-frequentist controversy.* Division of Biostatistics, Stanford University, 2005.

B. Efron and R. Tibshirani. Statistical data analysis in the computer age. *Science*, 253(5018):390–395, 1991.

S. B. Eickhoff, A. R. Laird, C. Grefkes, L. E. Wang, K. Zilles, and P. T. Fox. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Hum Brain Mapp.*, 30(9):2907–2926, 2009.

S. B. Eickhoff, D. Bzdok, A. R. Laird, C. Roski, S. Caspers, K. Zilles, and P. T. Fox. Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. *Neuroimage*, 57(3):938–49, 2011.

S. B. Eickhoff, D. Bzdok, A. R. Laird, F. Kurth, and P. T. Fox. Activation likelihood estimation meta-analysis revisited. *Neuroimage*, 59(3):2349–61, 2012.

A. K. Engel and W. Singer. Temporal binding and the neural correlates of sensory awareness. *Trends Cogn Sci*, 5(1):16–25, 2001.

S. A. Engel, G. H. Glover, and B. A. Wandell. Retinotopic organization in human visual cortex and the spatial precision of functional mri. *Cereb Cortex*, 7(2):181–192, 1997.

A. C. Evans, D. L. Collins, and B. Milner. An mri-based stereotactic atlas from 250 young normal subjects. *Soc Neurosci*

Abstr, 18(408), 1992.

L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods*, 4(3):272, 1999.

D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, 1(1): 1–47, 1991.

P. Feyerabend. *Against method: Outline of an anarchistic theory of knowledge. Atlantic Highlands.* 1975.

R. A. Fisher. *Statistical methods for research workers.* Genesis Publishing Pvt Ltd, 1925.

E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

E. Formisano and N. Kriegeskorte. Seeing patterns through the hemodynamic veil–the future of pattern-information fmri. *Neuroimage*, 62(2):1249–56, 2012.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Springer, 2013.

D. F. Fox and M. E. Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci*, 8:700–711, 2007.

M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci U S A*, 102(27):9673–8, 2005.

P. T. Fox and J. L. Lancaster. Opinion: Mapping context and content: the brainmap model. *Nat Rev Neurosci*, 3(4):319–21, 2002.

P. T. Fox and M. E. Raichle. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during so matosensory stimulation in human subjects. *Proc Natl Acad Sci U S A*, 83:1140–1144, 1986.

R. Frackowiak and H. Markram. The future of human cerebral cartography: a novel approach. *Philos Trans R Soc Lond B Biol Sci*, 370(1668), 2015.

P. Fransson. How default is the default mode of brain function? further evidence from intrinsic bold signal fluctuations. *Neuropsychologia*, 44:28362845, 2006.

S. I. Franz and K. S. Lashley. The retention of habits by the rat after destruction of the frontal parts of the cerebrum. *Psycholobiology*, 1:3–18, 1917.

J. B. Freeman, D. Schiller, N.O. Rule, and N. Ambady. The neural origins of superficial and individuated judgments about ingroup and outgroup members. *Hum Brain Mapp*, 31:150–159, 2010.

K. J. Friston. Modes or models: a critique on independent component analysis for fmri. *Trends Cogn Sci*, 2(10):373–375, 1998.

K. J. Friston. The free-energy principle: a unified brain theory? *Nat Rev Neurosci*, 11(2):127–38, 2010.

K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp*, 2(4):189–210, 1994.

K. J. Friston, C. Buechel, G. R. Fink, J. Morris, E. Rolls, and R. J. Dolan. Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage*, 6(3):218–29, 1997.

K. J. Friston, J. Phillips, D. Chawla, and C. Buechel. Revealing interactions among brain systems with nonlinear pca. *Human Brain Mapp*, 8(2-3):92–97, 1999a.

K. J. Friston, E. Zarahn, O. Josephs, R. N. A. Henson, and A. M. Dale. Stochastic designs in event-related fmri. *Neuroimage*, 10(5):607–619, 1999b.

K. J. Friston, C. Chu, J. Mourao-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner. Bayesian decoding of brain images. *Neuroimage*, 39(1):181–205, 2008.

M. Fukunaga, S. G. Horovitz, P. van Gelderen, J. A. de Zwart, J. M. Jansma, V. N. Ikonomidou, R. Chu, R. H. R. Deckers, D. A. Leopold, and J. H. Duyn. Large-amplitude, spatially correlated fluctuations in bold fmri signals during extended rest and early sleep stages. *Magnetic Resonance Imaging*, 24(8):979–992, 2006.

I. Gauthier, M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat Neurosci*, 2(6):568–573, 1999.

S. J. Gilbert, R. N. Henson, and J. S. Simons. The scale of functional specialization within human prefrontal cortex. *J Neurosci*, 30(4):1233–7, 2010.

C. Giraud. *Introduction to High-Dimensional Statistics.* CRC Press, 2014.

D. R. Gitelman, A. C. Nobre, T. B. Parrish, K. S. LaBar, Y. H. Kim, J. R. Meyer, and M. Mesulam. A large-scale distributed network for covert spatial attention: further anatomical delineation based on stringent behavioural and cognitive controls. *Brain*, 122 (Pt 6):1093–106, 1999.

M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, and J. R. Polimeni. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

K. Gödel. über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für Mathematik und Physik*, 38(1):173–198, 1931.

Y. Golland, S. Bentin, H. Gelbard, Y. Benjamini, R. Heller, Y. Nir, U. Hasson, and R. Malach. Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. *Cereb Cortex*, 17(4):766–77, 2007.

G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

M. Goodkind, S. B. Eickhoff, D. J. Oathes, Y. Jiang, A. Chang, L. B. Jones-Hagata, B. N. Ortega, Y. V. Zaiko, E. L. Roach, M. S. Korgaonkar, et al. Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry*, 72(4): 305–315, 2015.

K. J. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform*, 5:13, 2011.

K. J. Gorgolewski, G. Varoquaux, G. Rivera, Y. Schwarz, S. S. Ghosh, C. Maumet, T. E. Nichols, R. A. Poldrack, J. P. Poline, Y. Yarkoni, and D. S. Margulies. Neurovault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. page in press, 2014.

R. L. Gregory. Perceptions as hypotheses. *Philos Trans R Soc Lond B Biol Sci*, 290(1038):181–197, 1980.

A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni, et al. Cost of disorders of the brain in europe 2010. *Eur Neuropsychopharmacol*, 21(10):718–779, 2011.

J. Habermas. *Theorie des kommunikativen Handelns.* Frankfurt am Main, 1981.

A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

J. P. Hamilton, D. J. Furman, C. Chang, M. E. Thomason, E. Dennis, and I. H. Gotlib. Default-mode and task-positive network activity in major depressive disorder: Implications for adaptive and maladaptive rumination. *Biol Psychiatry*, 70 (4):327333, 2011.

J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15): 3201–12, 2005.

S. J. Hanson and Y. O. Halchenko. Brain reading using full brain support vectormachines for object recognition: There is no face identication area. *Neural Comput*, 20:486–503, 2008.

G. H. Hardy. *A mathematician's apology.* Cambridge University Press, 1992.

J. M. Harlow. Passage of an iron rod through the head. *Boston Medical and Surgical Journal*, 39(20):389–393, 1848.

J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *J Appl Stat*, 28:100–108, 1979.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics, Heidelberg, Germany, 2011.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, 2015.

J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

J. D. Haynes and G. Rees. Predicting the orientation of invisible stimuli from acitvity in human primary visual cortex. *Nat Neurosci*, 8(5):686–691, 2005.

M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174, 2013.

L. Hensel, D. Bzdok, V. I. Müller, K. Zilles, and S. B. Eickhoff. Neural correlates of explicit social judgments on vocal stimuli. *Cerebral Cortex*, 2015.

J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation*, volume 1. Basic Books, 1991.

C. M. Heyes and C. D. Frith. The cultural evolution of mind reading. *Science*, 344(6190), 2014.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length, and helmholtz free energy. *NIPS*, pages 3–3, 1994.

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput*, 18(7):1527–1554, 2006.

J. F. Hipp and M. Siegel. Bold fmri correlation reflects frequency-specific neuronal correlation. *Curr Biol*, 2015.

C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans. Enhancement of mr images using registration for signal averaging. *J Comput Assist Tomogr*, 22(2):324–33, 1998.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *J Mach Learn Res*, 12:3371–3412, 2011.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Neurophysiol*, 28(2):229–289, 1965.

N. K. Humphrey. *The social function of intellect.*, pages 303–317. Cambridge University Press, Cambridge, 1978.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

A. Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651666, 2010.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACN Computing Surveys*, 31(3):264–323, 1999.

O. Jakobs, R. Langner, S. Caspers, C. Roski, E. C. Cieslik, K. Zilles, A. R. Laird, P. T. Fox, and S. B. Eickhoff. Across-study and within-subject functional connectivity of a right temporo-parietal junction subregion involved in stimulus-context integration. *Neuroimage*, 60(4):2389–2398, 2012.

N. Japkowicz, S. J. Hanson, and M. Gluck. Nonlinear autoassociation is not equivalent to pca. *Neural Comput*, 12(3):531–545, 2000.

S. Jbabdi and T. E. Behrens. Long-range connectomics. *Ann N Y Acad Sci*, 1305:83–93, 2013.

S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241254, 1967.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat*, pages 295–327, 2001.

I. T. Jolliffe. Rotation of principal components: choice of normalization constraints. *J Appl Stat*, 22(1):29–35, 1995.

M. I. Jordan. *Frontiers in massive data analysis.* National Academies Report, 2015.

Y. Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nat Neurosci*, 8(5):679–685, 2005.

R. Kanai and G. Rees. The structural basis of inter-individual differences in human behaviour and cognition. *Nat Rev Neurosci*, 12(4):231–42, 2011.

E. R. Kandel and J. H. Schwartz. *Principles of Neural Science*, volume 4. McGraw-Hill New York, 2000.

E. R. Kandel, H. Markram, P. Matthews, R. Yuste, and C. Koch. Neuroscience thinks big (and collaboratively). *Nat Rev*

Neurosci, 14(9):659–664, 2013.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–4311, 1997.

C. Kelly, R. Toro, A. Di Martino, P. Cox, C. L.and Bellec, F. X. Castellanos, and M. P. Milham. A convergent functional architecture of the insula emerges across imaging modalities. *Neuroimage*, 61(4):1129–1142, 2012.

J. A. S. Kelso, G. Dumas, and E. Tognoli. Outline of a general theory of behavior and brain coordination. *Neural Networks*, 37:120–131, 2013.

D. P. Kennedy, E. Redcay, and E. Courchesne. Failing to deactivate: Resting functional abnormalities in autism. *Proc Natl Acad Sci U S A*, 103(21):82758280, 2006.

J. H. Kim, J. M. Lee, H. J. Jo, S. H. Kim, J. H. Lee, S. T. Kim, S. W. Seo, R. W. Cox, D. L. Na, S. I. Kim, and Z. S. Saad. Defining functional sma and pre-sma subregions in human mfc using resting state fmri: functional connectivity-based parcellation method. *Neuroimage*, 49(3):2375–86, 2010.

A. W. Kimball. Errors of the third kind in statistical consulting. *JASA*, 52(278):133–142, 1957.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *NIPS*, pages 3581–3589, 2014.

Y. Kobayashi and D. G. Amaral. Macaque monkey retrosplenial cortex: Ii. cortical afferents. *J Comp Neurol*, 466(1):48–79, 2003.

Y. Kobayashi and D. G. Amaral. Macaque monkey retrosplenial cortex: Iii. cortical efferents. *J Comp Neurol*, 502(5):810–33, 2007.

N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *Proc Natl Acad Sci U S A*, 103 (10):3863–8, 2006.

F. M. Krienen, P. C. Tu, and R. L. Buckner. Clan mentality: evidence that the medial prefrontal cortex responds to close others. *J Neurosci*, 30(41):13906–15, 2010.

M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, 2013.

A. R. Laird, S. B. Eickhoff, K. Li, D. A. Robin, D. C. Glahn, and P. T. Fox. Investigating the functional heterogeneity of the default mode network using coordinate-based meta-analytic modeling. *J Neurosci*, 29(46):14496–505, 2009.

C. Lamm, J. Decety, and T. Singer. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*, 54(3):2492–502, 2011.

R. Langner and S. Eickhoff. A meta-analytic review of the neural mechanisms of vigilant attention. *Psychol Bull*, page in press, 2012.

K. S. Lashley. Functional interpretation of anatomic patterns. *Research publications-Association for Research in Nervous and Mental Disease*, 30:529–547, 1951.

Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *NIPS*, pages 1017–1025, 2011.

M. Le Van Quyen. The brainweb of cross-scale interactions. *New ideas in psychology*, 29(2):57–63, 2011.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

R. Leech and D. J. Sharp. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(Pt 1):12–32, 2014.

E. B. Liddle, C. Hollis, M. J. Batty, M. J. Groom, J. J. Totman, M. Liotti, G. Scerif, and P. F. Liddle. Task-related default mode network modulation and inhibitory control in adhd: effects of motivation and methylphenidate. *Journal of Child Psychology and Psychiatry*, 52(7):761771, 2011.

S.P. Lloyd. Least squares quantization in pcm. *IEEE Trans Inf Theory*, 28:129–137, 1957.

N. K. Logothetis and B. A. Wandell. Interpreting the bold signal. *Annu Rev Physiol*, 66:735–69, 2004.

N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–7, 2001.

H. Lu, Q. Zou, H. Gu, M. E. Raichle, E. A. Stein, and Y. Yang. Rat brains also have a default mode network. *Proc Natl Acad Sci U S A*, 109(10):39793984, 2012.

N. Luhmann. *Soziale Systeme*. Suhrkamp Frankfurt am Main, 1984.

R. J. Maddock. The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain. *Trends Neurosci*, 22(7):310–6, 1999.

J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2009.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J Mach Learn Res*, 11:19–60, 2010.

D. Mantini, A. Gerits, K. Nelissen, J. B. Durand, O. Joly, L. Simone, H. Sawamura, C. Wardak, G. A. Orban, R. L. Buckner, and W. Vanduffel. Default mode of brain function in monkeys. *J Neurosci*, 31(36):12954–62, 2011.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

R. A. Mar. The neural bases of social cognition and story comprehension. *Annu Rev Psychol*, 62:103–34, 2011.

D. S. Margulies, J. L. Vincent, C. Kelly, G. Lohmann, L. Q. Uddin, B. B. Biswal, A. Villringer, F. X. Castellanos, M. P. Milham, and M. Petrides. Precuneus shares intrinsic functional architecture in humans and monkeys. *Proc Natl Acad Sci U S A*, 106(47):20069–74, 2009.

R. B. Mars, J. Sallet, U. Schuffelgen, S. Jbabdi, I. Toni, and M. F. Rushworth. Connectivity-based subdivisions of the human right "temporoparietal junction area": Evidence for different areas participating in different cortical networks. *Cereb Cortex*, 22(8):1894–1903, 2012.

M. F. Mason, M. I. Norton, J. D. Van Horn, D. M. Wegner, S. T. Grafton, and C. N. Macrae. Wandering minds: the default network and stimulus-independent thought. *Science*, 315:393–395, 2007.

M. J. McKeown, T. P. Jung, S. Makeig, G. Brown, S. S. Kindermann, T. W. Lee, and T. J. Sejnowski. Spatially independent activity patterns in functional mri data during the stroop color-naming task. *Proc Natl Acad Sci U S A*, 95(3):803–810, 1998.

M. Mennes, C. Kelly, X. N. Zuo, A. Di Martino, B. B. Biswal, X. F. Castellanos, and M. P. Milham. Inter-individual differences in resting state functional connectivity predict task-induced bold activity. *Neuroimage*, 50(4):16901701, 2010.

M. Mennes, C. Kelly, S. Colcombe, F. X. Castellanos, and M. P. Milham. The extrinsic and intrinsic functional architectures of the human brain are not equivalent. *Cereb Cortex*, 23:223–229, 2013.

M. Mesulam. A horseradish peroxidase method for the identification of the efferents of acetyl cholinesterase-containing neurons. *J Histochem Cytochem*, 24(12):1281–5, 1976.

M. M. Mesulam. From sensation to cognition. *Brain*, 121:1013–52, 1998.

M. M. Mesulam. The evolving landscape of human cortical connectivity: facts and inferences. *Neuroimage*, 62(4):2182–2189, 2012.

B. Mishra, G. Meyer, F. Bach, and R. Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

M. Mur, P. A. Bandettini, and N. Kriegeskorte. Revealing representational content with pattern-information fmri–an introductory guide. *Soc Cogn Affect Neurosci*, 4(1):101–9, 2009.

L. Nanetti, L. Cerliani, V. Gazzola, R. Renken, and C. Keysers. Group analyses of connectivity-based cortical parcellation using repeated k-means clustering. *Neuroimage*, 47(4):1666–77, 2009.

T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–10, 2011.

A. C. Need and D. B. Goldstein. Whole genome association studies in complex diseases: where do we stand? *Dialogues in Clinical Neuroscience*, 12(1):37, 2010.

J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, pages 175–240, 1928.

A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

A. Y. Ng and M. I. Jordan. On discriminative vs. *Generative Classifiers: A Comparison of Logistic Regression and Naïve Bayes, NIPS*, 15, 2003.

T. Nichols, M. Brett, J. Andersson, T. Wager, and J. B. Poline. Valid conjunction inference with the minimum statistic. *Neuroimage*, 25(3):653–60, 2005.

P. Norvig. On chomsky and the two cultures of statistical learning. *Author Homepage*, 2011.

R. Nuzzo. Scientific method: statistical errors. *Nature*, 506(7487):150–2, 2014.

S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A*, 87(24):9868–72, 1990.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

A. J. O'Toole, F. Jiang, H. Abdi, N. Pénard, J. P. Dunlop, and M. A. Parent. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J Cognitive Neurosci*, 19(11): 1735–1752, 2007.

S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10): 1345–1359, 2010.

J. Parvizi, G. W. Van Hoesen, J. Buckwalter, and A. Damasio. Neural connections of the posteromedial cortex in the macaque. *Proc Natl Acad Sci U S A*, 103(5):1563–8, 2006.

R. E. Passingham, K. E. Stephan, and R. Kotter. The anatomical basis of functional localization in the cortex. *Nat Rev Neurosci*, 3(8):606–16, 2002.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. M. Pearson, B. Y. Hayden, S. Raghavachari, and M. L. Platt. Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Curr Biol*, 19(18):1532–7, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *J Mach Learn Res*, 12:2825–2830, 2011.

F. Pedregosa, M. Eickenberg, P. Ciuciu, B. Thirion, and A. Gramfort. Data-driven hrf estimation for encoding and decoding models. *Neuroimage*, 104:209–220, 2015.

W. Penfield and P. Perot. The brain's record of auditory and visual experience. *Brain*, 86(4):595–696, 1963.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45:199–209, 2009.

L. Pessoa. On the relationship between emotion and cognition. *Nat Rev Neurosci*, 9(2):148–58, 2008.

E. A. Phelps and J. E. LeDoux. Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron*, 48(2):175–187, 2005.

W. Pietsch. Big data–the new science of complexity. 2013.

P. Pinel, B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J. B. Poline, and S. Dehaene. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neurosci*, 8:91, 2007.

N. Pisapia, M. Turatto, P. Lin, J. Jovicich, and A. Caramazza. Unconscious priming instructions modulate activity in default and executive networks of the human brain. *Cereb Cortex*, 22(3):639–49, 2012.

R. A. Poldrack. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci*, 10(2):59–63, 2006.

R. A. Poldrack. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72 (5):692–7, 2011.

R. A. Poldrack and K. J. Gorgolewski. Making big data open: data sharing in neuroimaging. *Nat Neurosci*, 17(11):1510–1517, 2014.

R. A. Poldrack, Y. O. Halchenko, and S. J. Hanson. Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychol Sci*, 20(11):1364–72, 2009.

K. R. Popper. *Logik der Forschung. Wien*, volume 308. 1934.

C. J. Price. The anatomy of language: a review of 100 fmri studies published in 2009. *Ann N Y Acad Sci*, 1191:62–88, 2010.

M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. A default mode of brain function. *Proc Natl Acad Sci U S A*, 98(2):676–82, 2001.

S. Ramón y Cajal. Histologie du systeme nerveux de lhomme et des vertébrés. *Maloine, Paris*, pages 655–664, 1909/1911.

M. A. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *NIPS*, pages 1137–1144, 2006.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

M. Reivich, D. Kuhl, A. Wolf, J. Greenberg, M. Phelps, T. Ido, V. Casella, J. Fowler, E. Hoffman, A. Alavi, P. Som, and L. Sokoloff. The [18f]fluorodeoxyglucose method for the measurement of local cerebral glucose utilization in man. *Circ Res*, 44(1):127–37, 1979.

G. Rizzolatti. Functional organization of inferior area 6. *Ciba Found Symp*, 132:171–86, 1987.

J. L. Robinson, A. R. Laird, D. C. Glahn, W. R. Lovallo, and P. T. Fox. Metaanalytic connectivity modeling: delineating the functional connectivity of the human amygdala. *Hum Brain Mapp*, 31(2):173–84, 2010.

D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang. Incremental learning for robust visual tracking. *Int J Comput Vis*, 77(1-3): 125–141, 2008.

C. S. Roy and C. S. Sherrington. On the regulation of the blood supply of the brain. *J Physiol*, 11:85–108, 1890.

S. Russell and P. Norvig. *Intelligence artificielle: Avec plus de 500 exercices*. Pearson Education France, 2010.

J. Sallet, R. B. Mars, M. P. Noonan, J. L. Andersson, J. X. Oreilly, S. Jbabdi, P. L. Croxson, M. Jenkinson, K. L. Miller, and M. F. S. Rushworth. Social network size affects neural circuits in macaques. *Science*, 334(6056):697–700, 2011.

A. Sandberg and N. Bostrom. Whole brain emulation. 2008.

F. Scheperjans, K. Hermann, S. B. Eickhoff, K. Amunts, A. Schleicher, and K. Zilles. Observer-independent cytoarchitectonic mapping of the human superior parietal cortex. *Cereb Cortex*, 18(4):846–67, 2008.

L. Schilbach, S. B. Eickhoff, A. Rotarska-Jagiela, G. R. Fink, and K. Vogeley. Minds at rest? social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Conscious Cogn*, 17(2):457–67, 2008.

L. Schilbach, D. Bzdok, B. Timmermans, P. T. Fox, A. R. Laird, K. Vogeley, and S. B. Eickhoff. Introspective minds: Using ale meta-analyses to study commonalities in the neural correlates of emotional processing, social unconstrained cognition. *PLoS One*, 7(2), 2012.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

Y. Schwartz, B. Thirion, and G. Varoquaux. Mapping cognitive ontologies to and from the brain. *NIPS*, 2013.

M. Seeger. Learning with labeled and unlabeled data. Technical report, 2000.

W. W. Seeley, V. Menon, A. F. Schatzberg, J. Keller, G. H. Glover, H. Kenna, A. L. Reiss, and M. D. Greicius. Dissociable intrinsic connectivity networks for salience processing and executive control. *J Neurosci*, 27(9):2349–2356, 2007.

W. W. Seeley, R.K. Crawford, J. Zhou, B.L. Miller, and M.D. Greicius. Neurodegenerative diseases target large-scale human brain networks. *Neuron*, 62:4252, 2009.

Z. Shehzad, A. M. Kelly, P. T. Reiss, D. G. Gee, K. Gotimer, L. Q. Uddin, S. H. Lee, D. S. Margulies, A. K. Roy, B. B. Biswal, E. Petkova, F. X. Castellanos, and M. P. Milham. The resting brain: unconstrained yet reliable. *Cereb Cortex*, 19(10): 2209–29, 2009.

G. M. Shepard. *Neurobiology.* Oxford University Press, 1988.

J. Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014a.

J. Shlens. A tutorial on independent component analysis. *arXiv preprint arXiv:1404.2986*, 2014b.

G. L. Shulman, J. A. Fiez, M. Corbetta, R. L. Buckner, F. M. Miezin, M. E. Raichle, and S. E. Petersen. Common blood flow changes across visual tasks .2. decreases in cerebral cortex. *J Cognitive Neurosci*, 9(5):648–663, 1997.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *J Comp Graph Stat*, 22(2):231–245, 2013.

W. Singer. Neurobiology: Striving for coherence. *Nature*, 397(6718):391–393, 1999.

B. F. Skinner. *About behaviorism.* Vintage, 1976.

S. M. Smith, P. M. Matthews, and P. Jezzard. *Functional MRI: an introduction to methods.* Oxford University Press, 2001.

S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proc Natl Acad Sci U S A*, 106(31):13040–5, 2009.

S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, et al. Resting-state fmri in the human connectome project. *Neuroimage*, 80:144–168, 2013.

C. Spearman. "general intelligence," objectively determined and measured. *Am J Psychol*, 15(2):201–292, 1904.

R. N. Spreng, R. A. Mar, and A. S. N. Kim. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *J Cognitive Neurosci*, 21(3):489–510, 2009.

D. Sridharan, D. J. Levitin, and V. Menon. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc Natl Acad Sci U S A*, 105(34):12569–74, 2008.

K. E. Stephan. On the role of general system theory for functional neuroimaging. *J Anat*, 205(6):443–70, 2004.

K. E. Stephan, J. C. Marshall, K. J. Friston, J. B. Rowe, A. Ritzl, K. Zilles, and G. R. Fink. Lateralized cognitive processes and lateralized task control in the human brain. *Science*, 301(5631):384–386, 2003.

J. Talairach and P. Tournoux. *Co-planar stereotaxic atlas of the human brain.* Thieme, New York, 1988.

B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J. B. Poline, D. Lebihan, and S. Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–16, 2006.

B. Thirion, G. Varoquaux, E. Dohmatob, and J. B. Poline. Which fmri clustering gives good brain parcellations? *Front Neurosci*, 8:167, 2014.

B. Thompson. *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association, 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc: Series B*, pages 267–288, 1996.

T. Tieleman and G. Hinton. Lecture 6.5rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

B. Timmermans, L. Schilbach, A. Pasquali, and A. Cleeremans. Higher order thoughts in action: consciousness as an unconscious re-description process. *Philos Trans R Soc Lond B Biol Sci*, 367(1594):1412–23, 2012.

M. Tomasello. *The Cultural Origins of Human Cognition.* Harvard University Press, 2009.

M. Tomasello, J. Call, and B. Hare. Chimpanzees understand psychological states - the question is which ones and to what extent. *Trends Cogn Sci*, 7(4):153–156, 2003.

J. W. Tukey. *Exploratory Data Analysis.* Pearson, New Jersey, 1977.

A. M. Turing. On computable numbers, with an application to the entscheidungsproblem. *J Math*, 58(345-363):5, 1936.

P. E. Turkeltaub, S. B. Eickhoff, A. R. Laird, M. Fox, M. Wiener, and P. Fox. Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Hum Brain Mapp*, 33(1):1–13, 2012.

L. Q. Uddin, K. Supekar, H. Amin, E. Rykhlevskaia, D. A. Nguyen, M. D. Greicius, and V. Menon. Dissociable connectivity within human angular gyrus and intraparietal sulcus: evidence from functional and structural connectivity. *Cereb Cortex*, 20(11):2636–46, 2010.

L. G Ungerleider and J. V. Haxby. What and where in the human brain. *Curr Opin Neurol*, 4(2):157–165, 1994.

L. G. Ungerleider and M. Mishkin. *Two cortical visual systems.*, page 549586. MIT Press, Cambridge, 1982.

T. Unterthiner, A. Mayr, G. Klambauer, and S. Hochreiter. Toxicity prediction using deep learning. *arXiv preprint arXiv:1503.01445*, 2015.

D. C. Van Essen. *Functional organization of primate visual cortex*, pages 259–329. Plenum Press, New York, 1985.

D. C. Van Essen, C. H. Anderson, and D. J. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–23, 1992.

D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, and S. W. Curtiss. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

F. Van Overwalle. Social cognition and the brain: a meta-analysis. *Human Brain Mapp*, 30(3):829–858, 2009.

F. Van Overwalle. A dissociation between social mentalizing and general reasoning. *Neuroimage*, 54(2):1589–1599, 2011.

V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1996.

F. Varela, J. P. Lachaux, E. Rodriguez, and J. Martinerie. The brainweb: phase synchronization and large-scale integration. *Nat Rev Neurosci*, 2(4):229–239, 2001.

G. Varoquaux and R. C. Craddock. Learning and comparing functional connectomes across subjects. *Neuroimage*, 80:405–15, 2013.

G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. *Information Processing in Medical Imaging*, pages 562–573, 2011.

P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res*, 11:3371–3408, 2010.

P. Virilio. Open sky. *London and New York: Verso*, 1997.

M. Visser, E. Jefferies, and M. A. Lambon Ralph. Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *J Cogn Neurosci*, 22(6):1083–94, 2010.

R. Viviani, G. Grön, and M. Spitzer. Functional principal component analysis of fmri data. *Human Brain Mapp*, 24(2):109–129, 2005.

B. A. Vogt, D. N. Pandya, and D. L. Rosene. Cingulate cortex of the rhesus monkey: I. cytoarchitecture and thalamic afferents. *J Comp Neurol*, 262(2):256–70, 1987.

B. A. Vogt, L. Vogt, and S. Laureys. Cytology and functionally correlated circuits of human posterior cingulate areas. *Neuroimage*, 29(2):452–66, 2006.

C. Vogt and O. Vogt. *Allgemeine Ergebnisse unserer Hirnforschung*, volume 21. JA Barth, 1919.

L. von Bertalanffy. An outline of general system theory. *Br J Philos Sci*, 1:13891164, 1950.

L. von Bertalanffy. *General system theory: Foundations, development, applications.* Braziller. New York, 1968.

L. S. Vygotsky. *Mind in society: The development of higher mental process.* Cambridge, MA: Harvard University Press, 1978.

P. Watzlawick, J. H. Beavin, and D. A. Jackson. *Pragmatics of Human Communication: A study of international patterns, pathologies, and paradoxes.* Norton, 1967.

A. Weissenbacher, C. Kasess, F. Gerstl, R. Lanzenberger, E. Moser, and C. Windischberger. Correlations and anticorrelations in resting-state functional connectivity mri: a quantitative comparison of preprocessing strategies. *Neuroimage*, 47(4):1408–16, 2009.

D. H. Weissman, K. C. Roberts, K. M. Visscher, and M. G. Woldorff. The neural bases of momentary lapses in attention. *Nat Neurosci*, 9(7):971–978, 2006.

C. Wernicke. Die akute haemorrhagische polioencephalitis superior. *Lehrbuch Der Gehirnkrankheiten Für Aerzte Und Studirende, Bd II*, 2:229–242, 1881.

S. Whitfield-Gabrieli and J. M. Ford. Default mode network activity and connectivity in psychopathology. *Annu Rev Clin Psychol*, 8:49–76, 2012.

S. Whitfield-Gabrieli, H. W. Thermenos, S. Milanovic, M. T. Tsuang, S. V. Faraone, R. W. McCarley, M. E. Shenton, A. I. Green, A. Nieto-Castanon, P. LaViolette, J. Wojcik, J. D. Gabrieli, and L. J. Seidman. Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proc Natl Acad Sci U S A*, 106(4):1279–84, 2009.

B. L. Whorf. *Language, Thought, and Reality.* Technology Press of MIT, 1956.

M. R. Wiegell, D. S. Tuch, H. B. Larsson, and V. J. Wedeen. Automatic segmentation of thalamic nuclei from diffusion tensor magnetic resonance imaging. *Neuroimage*, 19(2 Pt 1):391–401, 2003.

E. P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. richard courant lecture in mathematical sciences delivered at new york university, may 11, 1959. *Communications on pure and applied mathematics*, 13(1):1–14, 1960.

E. O. Wilson. *Letters to a young scientist.* WW Norton & Company, 2013.

L. Wittgenstein. *Philosophische Untersuchungen.* Blackwell, 1953.

D. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Comput*, 8:13411390, 1996.

T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods*, 8(8):665–70, 2011.

B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zollei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol*, 106(3):1125–65, 2011.

B. T. Yeo, F. M. Krienen, S. B. Eickhoff, S. N. Yaakub, P. T. Fox, R. L. Buckner, C. L. Asplund, and M. W. Chee. Functional specialization and flexibility in human association cortex. *Cereb Cortex*, 2014.

M. P. Young, C. C. Hilgetag, and J. W. Scannell. On imputing function to structure from the behavioural effects of brain lesions. *Philos Trans R Soc Lond B Biol Sci*, 355(1393):147–61, 2000.

D. Zhang and M. E. Raichle. Disease and the brain's dark energy. *Nat Rev Neurol*, 6(1):15–28, 2010.

S. Zhang and C. Li. Functional connectivity mapping of the human precuneus by resting state fmri. *Neuroimage*, 59(4): 3548–3562, 2012.

X. Zhu. Semi-supervised learning literature survey. 2006.

X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

K. Zilles and K. Amunts. Centenary of brodmann's map–conception and fate. *Nat Rev Neurosci*, 11(2):139–45, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J R Stat Soc: Series B*, 67(2):301–320, 2005.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J Comp Graph Stat*, 15(2):265–286, 2006.