NeuroImage 155 (2017) 549–564

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Inference in the age of big data: Future perspectives on neuroscience

Danilo Bzdok[a,b,c,d,*], B.T. Thomas Yeo[e,f,g,h,i]

[a] Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, 52072 Aachen, Germany
[b] JARA-BRAIN, Jülich-Aachen Research Alliance, Germany
[c] IRTG2150 - International Research Training Group, Germany
[d] Parietal team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France
[e] Department of Electrical and Computer Engineering, National University of Singapore, 119077 Singapore
[f] Clinical Imaging Research Centre, National University of Singapore, 117599 Singapore
[g] Singapore Institute for Neurotechnology, National University of Singapore, 117456 Singapore
[h] Memory Networks Programme, National University of Singapore, 119077 Singapore
[i] Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA

ARTICLE INFO

Keywords:
Systems biology
Epistemology
Hypothesis testing
High-dimensional statistics
Machine learning
Sample complexity

ABSTRACT

Neuroscience is undergoing faster changes than ever before. Over 100 years our field qualitatively described and invasively manipulated single or few organisms to gain anatomical, physiological, and pharmacological insights. In the last 10 years neuroscience spawned quantitative datasets of unprecedented breadth (e.g., microanatomy, synaptic connections, and optogenetic brain-behavior assays) and size (e.g., cognition, brain imaging, and genetics). While growing data availability and information granularity have been amply discussed, we direct attention to a less explored question: *How will the unprecedented data richness shape data analysis practices?* Statistical reasoning is becoming more important to distill neurobiological knowledge from healthy and pathological brain measurements. We argue that large-scale data analysis will use more statistical models that are non-parametric, generative, and mixing frequentist and Bayesian aspects, while supplementing classical hypothesis testing with out-of-sample predictions.

Introduction

During most of neuroscience history, before the emergence of genomics and brain imaging, new insights were "inferred" with little or no reliance on statistics. Qualitative, sometimes anecdotal reports have documented impairments after brain lesion (Harlow, 1848), microscopical inspection of stained tissue (Brodmann, 1909), electrical stimulation during neurosurgery (Penfield and Perot, 1963), targeted pharmacological intervention (Clark et al., 1970), and brain connections using neuron-transportable dyes (Mesulam, 1978). Connectivity analysis by axonal tracing studies in monkeys exemplifies biologically justified "inference" with many discoveries since the 60 s (Köbbert et al., 2000). A colored tracer substance is injected in vivo into source region A, uptaken by local neuronal receptors, and automatically transported in axons to target region B. This observation in *a single monkey* allows *extrapolating* a monosynaptical connection between region A and B to the *entire monkey species* (Mesulam, 2012). Instead, later brain-imaging technology propelled the data-intensive characterization of the mammalian brain and today readily quantifies axonal connections, cytoarchitectonic borders, myeloarchitectonic distributions, neurotransmitter receptors, and oscillatory coupling (Amunts et al., 2013; Frackowiak and Markram, 2015; Kandel et al., 2013; Van Essen et al., 2012). Following many new technologies to generate digitized yet noisy brain data, drawing insight from observations in the brain henceforth required assessment in the statistical arena.

In the quantitative sciences, the invention and application of statistical tools has always been dictated by changing contexts and domain questions (Efron and Hastie, 2016). The present paper will therefore examine how statistical choices are likely to change due to the progressively increasing granularity of digitized brain data. Massive data collection is a game changer in neuroscience (Kandel et al., 2013; Poldrack and Gorgolewski, 2014), and in many other public and private areas (House of Commons, 2016; Jordan et al., 2013; Manyika et al., 2011). There is a growing interest in and pressure for data sharing, open access, and building "big data" repositories (Frackowiak and Markram, 2015; Lichtman et al., 2014; Randlett et al., 2015). For instance, UK Biobank is a longitudinal population study dedicated to the genetic and environmental influence on mental disorders and other medical conditions (Allen et al., 2012; Miller et al., 2016). 500,000 enrolled volunteers undergo an extensive battery of clinical diagnostics

* Correspondence to: Department of Psychiatry, Psychotherapy and Psychosomatics, Pauwelsstraße 30, 52074 Aachen, Germany.
E-mail address: danilo.bzdok@rwth-aachen.de (D. Bzdok).

http://dx.doi.org/10.1016/j.neuroimage.2017.04.061
Received 20 August 2016; Received in revised form 25 April 2017; Accepted 25 April 2017
Available online 27 April 2017

from brain scans to bone density with a > 25 year follow-up. In the US, the Precision Medicine Initiative announced in 2015 to profile 1,000,000 individuals (Collins and Varmus, 2015). Targeted analysis of such national and international data collections may soon become the new normal in basic and clinical neuroscience. In this opinion paper, we will inspect the statistical scalability to the data-rich scenario from four different formal perspectives: i) parametric versus non-parametric models, ii) discriminative versus generative models, and iii) frequentist versus Bayesian models, as well as iv) classical hypothesis testing and out-of-sample generalization.

Towards adaptive models

Parametric models seek to capture underlying structure in data, which is representable with a fixed number of model parameters. For instance, many parametric models with Gaussianity assumptions will attempt to fit Gaussian densities regardless of the underlying data distribution. On the other hand, we think of *non-parametric* models as typically making weaker assumptions about the underlying data structure, such that the model complexity is data-driven, the *expressive capacity* does not saturate, the model structure can adapt flexibly, and the prediction can grow more sophisticated (see Box 1 for elaboration). Certain non-parametric models (e.g., Parzen window density estimation) will converge to the true underlying data distribution with sufficient data (although the amount of needed data might be astronomical). With increasing data samples, non-parametric models thus tend to make always-smaller error in capturing underlying structure in data (Devroye et al., 1996; Bickel et al., 2007). Relating these considerations back to the deluge of data from burgeoning neuroscience consortia, "the main concern is underfitting from the choice of an overly simplistic parametric model, rather than overfitting." (Ghahramani, 2015, p. 454). We therefore believe that non-parametric models have the potential to extract arbitrarily complex perceptual units, motor programs, and neural computations directly from healthy and diseased brain measurements.

In our opinion, the expressive capacity of many parametric models to capture cognitive and neurobiological processes is limited and cannot adaptively increase if more input data are provided. For instance, independent component analysis (ICA) is an often-used parametric model that extracts a set of macroscopic networks with coherent neural activity from brain recordings (Calhoun et al., 2001; Beckmann et al., 2009). Applied to human functional magnetic resonance imaging (fMRI) data, ICA reliably yields the default mode network, saliency network, dorsal attention network, and other canonical brain networks (Damoiseaux et al., 2006; Seeley et al., 2007; Smith et al., 2009). Standard ICA is parametric in the sense that the algorithm extracts a user-specified number of spatiotemporal network components, although the "true" number of macroscopic brain networks is unknown or might be ambiguous (Eickhoff et al., 2015). By coupling standard ICA with approximate Bayesian model selection (BMS), Beckmann and Smith (2004) allowed the number of components to flexibly adapt to brain data. The combination of parametric ICA and BMS yields an integrative modeling approach that exhibits the scaling property of non-parametric statistics (Goodfellow et al., 2016, p. 112; Ghahramani, 2015, p. 454): With increasing amount of input data, ICA with BMS adaptively calibrates the *model complexity* by potentially extracting more brain network components, thus enhancing the expressive power of classical ICA.

These advantages are inherent to *non-parametric* models that can potentially extract an always higher number of neural patterns that are *adaptively described by an always higher, theoretically infinite number of model parameters* as the amount of input data increase (Orbanz and Teh, 2011; Ghahramani, 2013). In doing so, we believe non-parametric models can potentially isolate representations of neurobiological phenomena that do not only improve quantitatively (e.g., increased statistical certainty) but also qualitatively (e.g., a much different, more detailed representation). We propose that *non-parametric models are hence more likely to extract neurobiological relationships that exclusively emerge in large brain datasets*. In contrast, parametric models are often more easily interpretable by the investigator, are more stable, and require less data to achieve a satisfactory model fit. Furthermore, parametric statistical tests are often more powerful, assuming the parametric assumptions are correct (cf. Friston, 2012; Eklund et al., 2016). These practical advantages are

Box 1: Parametric and non-parametric models

Contrary to common misunderstanding, both *parametric* and *non-parametric* statistical models involve parameters. 'Non-parametric' is typically defined in one of three different flavors (Bishop, 2006; Murphy, 2012; James et al., 2013): The first, perhaps most widespread meaning implies those statistical models that do not make explicit assumptions about a particular *probability distribution* (e.g., Gaussian distribution) from which the data have arisen. As a second and more general definition, non-parametric models do not assume that the *structure of the statistical model* is fixed. The third definition emphasizes that in non-parametric models, *the number of model parameters* increases explicitly or implicitly with the number of available data points (e.g., number of participants in the dataset). In contrast, the number of model parameters is fixed in parametric models and does not vary with sample size (Fig. 1). In its most extreme manifestation, non-parametric models might utilize larger memory than the actual input data themselves. Please note that the non-parametric scaling property of increasing model complexity with accumulating data can be obtained in different ways: i) a statistical model with infinitely many parameters or ii) a nested series of parametric models that can increase the number of parameters as needed (Ghahramani, 2015, page 454; Goodfellow et al., 2016, page 112).

The flexible non-parametric models include random forests (a special kind of decision-tree algorithm), boosting, nearest-neighbor algorithms (where complexity increases with the amount of input data), Gaussian Process methods, kernel support vector machines, kernel principal component analysis (kernel PCA), kernel ICA, kernel canonical correlation analysis, generalized additive models, and hierarchical clustering, as well as many forms of bootstrapping and other resampling procedures. Statistical models based on decision trees often constrain their size, which turns them into parametric models in practice. The more rigid parametric models include Gaussian mixture models, linear support vector machines, PCA, ICA, factor analysis, classical canonical correlation analysis, and k-means clustering, but also modern regression variants using sparsity or shrinkage regularization like Lasso, elastic net, and ridge regression.

Classical statistics has always had a strong preference for low-dimensional parametric models (Efron and Hastie, 2016). It is an advantage of parametric models to express the data compactly in often few model parameters. This increases interpretability, requires fewer data samples, has higher statistical power, and incurs lower computational load. Although the number of parameters in parametric models can be manually increased by the user, only non-parametric models have the inherent ability to automatically scale their *expressive capacity* with increasing data resources. Therefore, as the amount of neuroscience data continues to increase by leaps and bounds, parametric models might underfit the available data, while non-parametric models might discover increasingly complex representations that potentially yield novel neuroscientific insights.

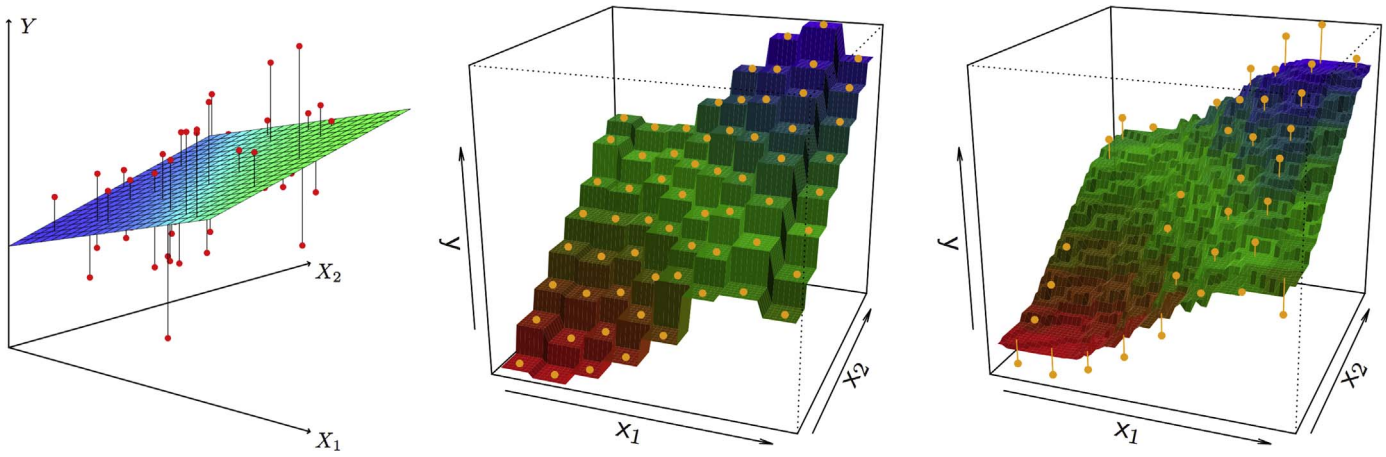

Fig. 1. Prediction based on parametric versus non-parametric regression. Fitted models that predict the continuous outcome Y based on the observed variables X_1 and X_2. *Left*: Ordinary linear regression finds the best plane to explain the outcome Y. *Middle/Right*: K-nearest neighbor regression predicts the same outcome Y based on K=1 (*middle*) or K=9 (*right*) closest data points in the available sample. Parametric linear regression cannot grow more complex than a plane (or hyperplane when there are more than two observed variables), resulting in big regions with identical predictions Y. Non-parametric nearest-neighbor regression can grow from a rough step-function regression surface (k=1) to a smoother and more complex regression surface (k=9) by incorporating more data. Non-parametric models can therefore outperform parametric alternatives in many data-rich scenarios (Ghahramani, 2015). Reused with permission from James et al. (2013).

paid for by the cost of more rigid models. We therefore believe that the strength of flexible non-parametric models to automatically adjust the number of model parameters will probably turn out to be a crucial property of statistical models used in data-rich neuroscience.

Although non-parametric models have been used in neuroimaging (e.g., Lashkari et al., 2012; Andersen et al., 2014), parametric models are today the predominant approach in neuroscience. Many big-sample studies (i.e., data from hundreds of animals or humans) currently apply the same parametric models as previous small-sample studies (i.e., a few dozen animals or humans). With increasing sample size, parametric analyses such as Student's t-test, F-test, ANOVA, linear regression, and Pearson's linear correlation on brain data from many hundred animals or humans yield a quantitatively increased certainty of statistical estimates (Button et al., 2013; Miller et al., 2016). However, we think that they might not necessarily improve the quality of neuroscientific insight gleaned from a sample with less observations.[1] In our opinion, an important caveat of parametric models manifests itself in their systematic inability to adaptively grow in complexity no matter how much brain data are collected and analyzed (Ghahramani, 2015).

In any classification setting where a statistical model distinguishes between two possible outcomes (e.g., healthy versus schizophrenic), a linear parametric model will always make predictions based on a separation between two classes by straight lines (or hyperplanes). Non-linear parametric models can be used to identify more complex structure in large datasets while keeping the model complexity (i.e., number of parameters) constant. By contrast, a *non-parametric* model can learn a non-linear decision boundary *whose shape grows more complex with more data*. In analogous fashion, classical hidden Markov models for time-series analysis and structure discovery (cf. example in next section) may get upgraded to infinite hidden Markov models with a theoretically unlimited number of hidden spatiotemporal components that can be estimated with increasing data samples. In non-parametric clustering (e.g., Pitman, 2006), the question of best cluster number can be reframed as optimal cluster granularity depending on data availability to allow the number of extracted clusters to grow organically with increasing sample size. Such non-parametric alternatives can automatically balance between model complexity (i.e., number of model parameters to be estimated) and parsimony (i.e., efficiency of expressing the brain phenomenon). *Finally, we believe that linear support vector machines as a current go-to choice for classification and regression (e.g.*, Knops et al., 2009; Jimura and Poldrack, 2012*) may be more often supplemented by non-parametric approaches, such as random-forest and nearest-neighbor-type algorithms (e.g.*, Ball et al., 2014; Haxby et al., 2011; Misaki et al., 2010*;* Pereira et al., 2011*), in future neuroscience studies.*

More broadly, many interesting phenomena in the brain are likely to be very complex. Fortunately, stochastic processes have been proposed that realize random variables over unlimited function spaces mapping from brain data to a certain target variable. As an important member, Gaussian Processes (GP) can be seen as infinite dimensional generalizations of the multivariate Gaussian distribution (Ghahramani, 2013; Orbanz et al., 2011). GPs (with exponential-type kernels) consist in specifying probability distributions on unknown functions with the aim to impose minimal a-priori assumptions on the learnable relationships and minimal constraints on the possible non-linear interactions (Rasmussen, 2006). Instead of fitting one parameter to each variable to predict a behavior or clinical outcome, such as in linear regression, GPs (with exponential-type kernels) can fit a collection of non-linear functions with theoretically unlimited expressive capacity to explain particularly complex brain-behavior associations. In our opinion, this can probably enhance predictive regression and classification in large-sample studies in neuroscience whenever the ground-truth model in nature is not linear and additive (cf. Ripke et al., 2013).

For instance, effective scaling to the high-dimensional scenario (i.e., p variables > n samples) was demonstrated by a GP regression model that could explain 70% of known missing heritable variability in yeast phenotypes (Sharp et al., 2016). This kind of statistical analysis is today usually performed by genome-wide association studies (GWAS) that are based on the parametric generalized linear model (GLM) (cf. Zhang et al., 2010; Hastie et al., 2015, pp. 31–32). GLM-based approaches have however often explained only small fractions of the total heritable genomic variation. GPs have demonstrated that emergent biological insight can be gained from complex non-additive interactions between gene locations (and thus potentially brain locations). These *higher-order non-linear interactions* frequently involved groups of ~20 locations (Sharp et al., 2016), while even trying to capture all possible pairwise gene-gene interactions is difficult for the much less flexible GLMs in usual GWAS investigations. In fact, the computational costs of GLM approaches typically scale exponentially as a function of the

[1] However, results gleaned from large data sample are less likely to suffer from power issues and are therefore more likely to be replicable.

interaction order (i.e., variable-variable interactions, variable-variable-variable interactions, etc.). Further, adding all combinations of non-linear interaction terms to a GLM can quickly lead to a scenario where the model parameters largely exceed the number of available samples, which makes it challenging to estimate a meaningful solution (Hastie et al., 2015, chapter 3). Current genetic studies therefore constrain statistical analysis, for instance, by considering only pairwise gene-gene interactions or by considering only a pre-selected subset of genetic locations (Ritchie et al., 2001). Compared to many parametric GLM approaches used in genome-wide studies, we think that non-parametric GPs could *more exhaustively search the space of higher-order non-linear interactions* (Rasmussen, 2006). In neuroscience, brain-imaging studies for instance have already profited from GP applications, such as in EEG (e.g., Zhong et al., 2008) and in fMRI (e.g., Marquand et al., 2010; Lorenz et al., 2016).

GP belongs to the broader family of kernel-based methods, which can provide statistical advantages by mapping brain variables to a richer variable space (Hofmann et al., 2008). Non-parametric classification or regression with kernels performs a preprocessing of the pairwise similarity between all observations in the form of a so-called *kernel matrix* (i.e., n samples x n samples). The advantage is that this does not require an explicit mapping from individual brain variables to the richer variable space (i.e., "kernel trick"). The statistical model plugs in the virtual variable space instead of the original input variables. This can lead to linear separability of complex neurobiological effects that are not linearly separable in the original variables. Statistical models endowed with a kernel inherit enriched transformation of the brain data with relevance to modern neuroscience (e.g., Marinazzo et al., 2011) because they can decrease the computational burden in the high-dimensional scenario. Such purposeful increase of input dimensionality and model complexity is useful for small to intermediate datasets (roughly $n < 100{,}000$ samples), but incurs high computation and memory costs in very large datasets (Goodfellow et al., 2016, chapter 5.9), where the kernel matrix can grow to terabytes sizes due to quadratic scaling with respect to the number of samples. Disadvantages of kernels include the inability to interpret contributions of individual variables and to distinguish informative variables from noise. Moreover, the goal of understanding brain function will probably involve several levels of neuroscientific analysis and kernels promise effective modality fusion to incorporate several different types of data (Eshaghi et al., 2015; Schrouff et al., 2016; Young et al., 2013; Zhang et al., 2011). This is because, mathematically, kernel addition equates with combining different data sources into a common data space. We believe that such genuine multi-modal integration can enable conjoint inference on behavioral outcomes, brain connectivity, function phenotypes, and genetic variability.

In sum, brain structure, function, connectivity, and genetics are high-dimensional in nature and thus difficult to understand for human intuition. By expressing brain phenomena in statistical models with a fixed number of parameters, parametric models are typically more interpretable, easier to implement, and faster to estimate. They are often the best choice in data-scarce scenarios, but can underfit in the "big data" scenario. In our opinion, exclusive reliance on parametric analysis may keep neuroscientists from discovering novel neurobiological insights that only come to the surface by allowing for more complex data representations in data-rich scenarios (Halevy et al., 2009; Jordan et al., 2013, p. 63). It was recently emphasized that "the best predictive performance is often obtained from highly flexible learning systems, especially when learning from large data sets. Flexible models can make better predictions because to a greater extent they allow data to 'speak for themselves'." (Ghahramani, 2015). Even if more complex statistical models do not always result in greater insight (Eliasmith et al., 2012), statistical approaches with non-parametric scaling behavior are naturally prepared to capture more sophisticated brain phenomena. This is because the complexity of statistical structure and thus potentially extracted neurobiological knowledge can grow without limit with the amount of available data samples.

Towards more interpretable models that extract biological structure

How statistical analysis scales to large datasets is also impacted by the distinction between *generative* and *discriminative* models. We emphasize that *generative models are more ambitious than discriminative models because generative models seek the ability to produce new data samples consistent with the original observations* (for technical details see Box 2). In contrast, discriminative models are only concerned with predicting a target variable. For instance, a discriminative model would focus on predicting the disease status of an individual based on his or her neuroimaging profile (e.g., Fan et al., 2008; Zhang et al., 2011), while a generative model would seek to generate the neuroimaging profile of an individual given his or her disease status (e.g., Zhang et al., 2016).

Generative models range from biophysically realistic models that attempt to mimic actual biological processes (Freyer et al., 2011; Deco et al., 2013) to more abstract statistical models that seek to extract meaningful biological structure (e.g., probabilistic ICA). While the more abstract generative models might not correspond to genuine biological mechanisms, the extracted structure can still be physiologically or biologically meaningful (e.g., fMRI brain networks extracted with probabilistic ICA). A major advantage of generative models is that their results are usually more interpretable than those of discriminative models (see excellent examples from Haufe et al., 2014). However, in order to produce realistic high-dimensional data examples (e.g., neuroimaging profiles), generative models might have to be considerably more complex than discriminative models that only seek to predict a single target variable (e.g., disease status). In these scenarios (e.g., Fig. 2), more data samples might be necessary for high-quality generative modeling. Therefore *with the increasing abundance of brain data in the neurosciences, a wider deployment of generative models will become more feasible and in our opinion, important for understanding the brain*.

Generative models can be used to jointly estimate a brain-behavior relationship and a hidden representation in the brain that is useful for explaining the target behavior. As an example from connectivity analysis, dynamic causal modeling (DCM; Friston et al., 2003) is a common approach to study 'effective connectivity' in brain imaging, which quantifies the functional influence that one brain region exerts on other brain regions. DCM is a generative model with neurobiological plausibility because it captures linear and non-linear interactions between neuronal populations together with a biophysical model of the hemodynamic response function. DCM affords an internal representation of how the investigator-designed external inputs (i.e., known changes in experimental manipulation) lead to unobserved states of neuronal populations (i.e., hidden neural activity in several brain regions), resulting in the *generation* of observed evoked brain-imaging signals. Hidden neuronal states can thus be derived from hemodynamic responses. In contrast, using support vector regression (SVR) to predict brain maturity from resting-state functional connectivity (Dosenbach et al., 2010) is a discriminative approach because it does not facilitate the generation of functional connectivity data from a participant's age. While SVR can predict age very well from brain measurements (Dosenbach et al., 2010), interpreting weights from discriminative models can be misleading (Haufe et al., 2014).

Generative models can help discover *how* environmental perception and motor execution are reflected in measured neural signals. It is a classic idea that sensory perception in humans and animals draws on

Box 2: Discriminative and generative models.

Formally, *discriminative* models try to find a direct mapping function *f* from features *x* to a target variable *y* (i.e., y = f(x)). In the probabilistic setting, this involves modeling the posterior probability *P(y|x)* directly. *Generative* models traditionally solve the prediction problem by estimating the joint distribution *P(x,y)* (Jebara, 2004; Bishop and Lasserre, 2007). The prediction *P(y|x)* can then be indirectly obtained by applying Bayes' rule. Consequently, generative models can in principle produce synthetic, never observed examples $(\tilde{x}, \tilde{y})$ by sampling from the estimated joint distribution *P(x,y)*. If the synthetic data $(\tilde{x}, \tilde{y})$ is indistinguishable from real data, this suggests that the generative model is of good quality. It is worth noting that certain new approaches, such as generative adversarial networks, do not explicitly estimate the data-generating distribution, but can still generate extremely realistic new observations (Goodfellow et al., 2014; Goodfellow et al., 2016, p. 645).

Discriminative models are often chosen for best possible prediction of a target variable *y* (e.g., behavioral phenotypes, age, performance or clinical scores) from features *x* (i.e., brain measurements). In contrast, generative models can also be used to predict target variable *y* from brain measurements x, although the primary goal is to model how to best synthesize x from *y* (Fig. 2). Furthermore, generative modeling can be performed without reference to the target variable, in which case the goal is to discover some hidden structure that can be used to "generate" the features *x* (i.e., generative unsupervised learning). Generative models can thus provide detailed insight into the brain by explicitly modeling the sources of variation from which certain observations in the brain have arisen. These sources of variation can range from being biophysically plausible (e.g., through neural mass modeling) to abstract statistical constructs that can still be biologically meaningful (e.g., components from probabilistic ICA). Because features unrelated to the target variable can be assigned high weights in discriminative models (Haufe et al., 2014), generative models tend to be more interpretable, which is an important advantage when the goal is scientific discovery.

Members of discriminative models include logistic regression, support vector machines, decision-tree algorithms like random forests or gradient boosted trees, and many neural-network algorithms. Generative models include linear and quadratic discriminant analysis, Naive Bayes, hidden Markov models, Gaussian mixture models, latent Dirichlet allocation, many dictionary learning methods, linear/latent factor models, ICA, PCA, probabilistic canonical correlation analysis (Bach and Jordan, 2005), as well as many non-parametric statistical models (Teh and Jordan, 2010) and certain modern neural-network algorithms, such as autoencoders (Kingma et al., 2014). It is worth noting that linear-regression-type techniques can be discriminative or generative. For instance, logistic regression is a discriminative model and its generative analog is linear discriminant analysis (Bouchard et al., 2004).

In practice, the strength of generative models to jointly realize predictive modeling and a form of representation learning is often paid for by requiring more input data, possibly more computational resources and more model parameters to fit. The reason is that generative models need to take into account the joint distribution *P(x,y)*, which might be considerably more complex than the class posteriors *P(y|x)* (Fig. 2). The model performance can be further influenced by the additional assumptions of generative models compared to discriminative models (Bishop and Lasserre, 2007). In conclusion, generative models can improve interpretability but are frequently outperformed by discriminative models in prediction tasks, especially in cases with many samples (Ng and Jordan, 2002; Jebara and Meila, 2006; Xue et al., 2008) or many input variables (Kelleher, 2015, p. 516).

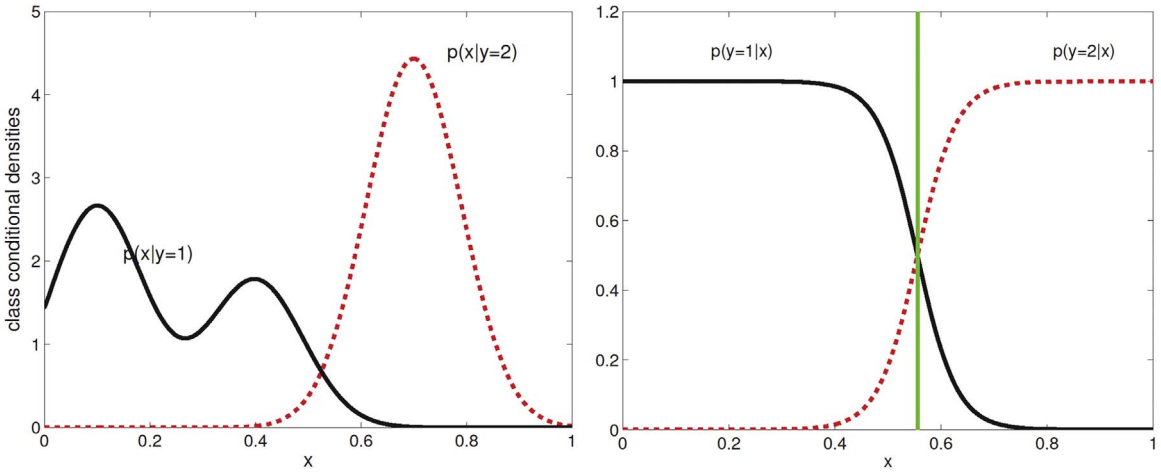

Fig. 2. Class-conditional densities can be more complex than class posteriors. To predict target class *y* from features *x*, generative models (*left*) estimate class conditional distributions *P(x|y=c)* and class priors *P(y=c)*, while discriminative models (*right*) estimate the posterior probability *P(y=c|x)* directly. In this example, the class conditional distributions *P(x|y=c)* are much more complex than the class posteriors *P(y=c|x)*. As such, an ideal generative model would have to be more complex (with more model parameters) than the ideal discriminative model in order to perform well in the prediction task. Hence, this more complex generative model would potentially require more training data to fit. However, the generative model can produce new unseen examples $(\tilde{x}, \tilde{y})$ and is typically more interpretable. Figure reused with permission from Murphy (2012).

the compositionality of environmental scenes into sensory primitives (Hubel and Wiesel, 1962). As a recent example of generative modeling in human auditory perception, the neural responses to diverse naturalistic sounds were stratified into distinct but spatially overlapping activity patterns (Norman-Haignere et al., 2015). The generative model discovered components of variation that captured selective tuning to frequency, pitch, and spectrotemporal modulation. Complex speech and music recruited anatomically distinct components suggesting the existence of distinct processing pathways for speech and music (Norman-Haignere et al., 2015).

Similarly, motor action on the ambient environment is probably assembled from a sequence of movement primitives (Wolpert et al., 2011) and sensorimotor learning is probably reliant on abstract internal representations. Both of these could be explicitly captured in

generative models but may become less evident using discriminative models (but see the success of Khaligh-Razavi et al., 2014; Yamins et al., 2014; Güçlü et al., 2015; Eickenberg et al., 2016 in understanding visual processing). There are already many promising applications of generative models in behavioral motor research (e.g., Acerbi et al., 2012; Franklin and Wolpert, 2011; Sing et al., 2009), but with much less frequent application to understanding the neural basis of motor action (but see Diedrichsen et al., 2005). As a computational approach to action choice, human social interaction has been described by a generative model that explicitly incorporated possible actions and expected subjective costs and rewards (Jara-Ettinger et al., 2016). This statistical model potentially allows an investigator to parse the observation of others' behavior and the derived conclusions on their beliefs, desires, and stable character traits. If agents act according to the generative model, the costs and rewards can be derived that were likely to have produced a given observed action. In neuroscience, the often less data-hungry discriminative models have so far been pervasive, while we expect generative models to grow in popularity along with greater data availability. We thus propose that generative models have the potential to carve perception, action, and cognition at their joints by statistically uncovering the relationships between their constituent neural elements.

Apart from sensory perception and motor execution, the possible interpretational gains of generative models have already been demonstrated in neuroscience studies on higher-order brain function. For instance, hidden Markov models have recently been applied to high-dimensional time-series data from magnetoencephalography (MEG) recordings (Baker et al., 2014). The employed generative models simultaneously inferred the spatial topography of the major brain networks subserving environmental responses and their cross-talk dynamics without making any a-priori assumptions about their anatomy. The model qualifies as generative because it takes into account the joint distribution over the neural activity time-series (i.e., "observed" variables[2]) and the underlying spatiotemporal components of variation (i.e., hidden variables). These model properties allowed the authors to argue that states of spatiotemporal coherence occur in 100-200ms time windows and that these functional coupling dynamics are faster than previously thought. As another example, neuroscientists often conceptualize psychological experiments as recruiting multiple neural processes supported by multiple brain regions (sometimes called 'multi-to-multi' mapping). This century-old notion (Walton and Paul, 1901) was recently expressed in the form of a generative model (Yeo et al., 2015). The author-topic model (Rosen-Zvi et al., 2010) was a natural choice because of its ability to derive unknown components of variation (i.e., cognitive primitives) whose constituent nodes (i.e., brain regions) can be shared to varying degrees among the discovered components. Applying the model to 10,449 neuroimaging experiments from the BrainMap database across 83 behavioral tasks revealed heterogeneity in the extent to which a given brain region participated in a variety of cognitive components and the extent to which a given cognitive component recruited a variety of brain regions. The results suggested that the human association cortex subserves diverse psychological tasks by flexible recruitment of functionally specialized networks whose constituent nodes are in part topographically overlapping.

Generative models are also useful for *representation learning* (Bengio et al., 2013), which pertains to extracting hidden "manifolds" (i.e., components of variation) directly from brain data. For instance, autoencoders (Hinton and Salakhutdinov, 2006; Goodfellow et al., 2016, chapter 14) are generative models that have been shown to generalize commonly employed representation discovery methods, including matrix decomposition techniques like ICA and PCA as well as clustering algorithms like k-means (Baldi and Hornik, 1989; Le et al., 2011). Applying generative autoencoder models to neural activity data opens the possibility to simultaneously extract local, non-overlapping components of variation (related to the notion of brain regions) and global, distributed components of variation (related to notion of brain networks) (Bzdok et al., 2015). Extracting an optimized region-network representation from brain data allows abandoning handpicked design of new summary variables from brain measurements (i.e., 'feature engineering'). Neurobiologically relevant representations can be revealed as sets of predictive patterns combining network components and region components that can together detect psychological tasks and disease processes. This happens without being constrained to either functional specialization into disjoint regions or functional integration by intertwined macroscopic networks (Sporns, 2013; Medaglia et al., 2015; Bzdok et al., 2017). The automatically discovered functional compartments, in turn, can be potentially utilized as features for supervised prediction.

In sum, we expect that generative models will be more readily exploited to discover hidden structure underlying brain measurements as data become more abundant. By exposing the low-dimensional structure embedded within high-dimensional brain measurements, generative models can provide more interpretable and more detailed insights into behavior and its disturbances (Stephan et al., 2017). However, "the more detailed and biologically realistic a model, the greater the challenges of parameter estimation and the danger of overfitting" (Stephan et al., 2015b). Additionally, generative models have been argued to be essential for *semi-supervised prediction from partially annotated data* (Bishop and Lasserre, 2007), yet another topic of growing importance (Bzdok et al., 2015). Moreover, a crucial next step in clinical neuroscience may lie in extracting underlying pathophysiological structure from brain measurements in mental disorders. Simply applying discriminative modeling strategies on psychiatric patients grouped by the diagnostic manuals DSM or ICD will likely recapitulate disease categories that are neither neurobiologically valid nor clinically predictive (Hyman, 2007; Insel et al., 2010). Ultimately, discriminative models may turn out to be less potent for reconstructing the neural implementation of information processing up to the level of 'decoding' mental content and thoughts directly from brain measurements.

Towards integration of traditional modeling regimes

The distinction between *Bayesian* and *frequentist* attitudes towards quantitative investigation (for technical details see Box 3) is well known in statistics (Freedman, 1995), and in neuroscience in particular (Friston et al., 2002; Stephan et al., 2009). Bayesian modeling emphasizes the importance of injecting a-priori assumptions into the data analysis, whereas frequentist modeling avoids the explicit introduction of prior beliefs. The Bayesian neuroscientist wants to discover statistical relationships that are calibrated on already existing knowledge deemed important by the investigator. In contrast, the frequentist neuroscientist wants to establish statistical relationships that are as objective and unconditioned by the investigator's expectations as possible. Note however that Bayesian approaches can employ flat or agnostic priors, while frequentist approaches can beconditioned on prior beliefs on the nature of the data distribution.

In the example of connectivity analysis, DCM is a *Bayesian* connectivity approach because experimentally induced connectivity changes are modeled under probabilistic priors on various biophysical parameters (e.g., resting oxygen extraction fraction, baseline coupling between regions and self-connection) governing the generative model of brain dynamics. In contrast, psychophysiological interaction (PPI) is a *frequentist* connectivity method because it seeks to model the changes in brain signals induced by experimental manipulations without placing probabilistic priors on neurophysiological properties of

[2] "Observed" is in quotations because the "observed" variables in this case corresponded to estimates of neuronal activity after beamforming and filtering the observed MEG data, rather than the original MEG data.

Box 3: **Frequentist and Bayesian models**.

In theory, the *frequentist* attitude aims at universally acceptable, investigator-independent conclusions on neurobiological processes by avoiding hand-selected priors on model parameters. The *Bayesian* attitude is more transparent in the unavoidable, necessarily subjective introduction of existing domain knowledge by specifying explicit model priors (Bishop, 2006; Murphy, 2012). Many frequentist approaches often achieve best-guess values by treating the model parameters as fixed unknown constants and input data as randomly generated conditioned on the model parameters (through the likelihood function). In Bayesian approaches, uncertainties in the estimation of model parameters are handled naturally by the computation of full posterior distributions and by marginalizing (i.e., summing or integrating) over random parameters of no interest. To this end, frequentist approaches often estimate a single set of model parameters by numerical optimization of the maximum likelihood. This single (point) estimate of the model parameters can potentially be used to predict new data. Unfortunately, this approach can lead to overfitting (Murphy, 2012, Chapter 2). In contrast, Bayesian approaches seek to estimate a posterior distribution over the space of model parameters. The posterior distribution can then be used to predict new data (i.e., by marginalizing over model parameters to compute the posterior predictive distribution), which provides protections against overfitting (Murphy, 2012, Chapter 2). The downside is that achieving posterior distributions of model parameters and integration over model parameters is generally much more difficult than achieving point estimates.

In practice, statistical models span a continuum between the extreme poles of frequentism and Bayesianism with many unexpected relations connecting the two paradigms (Bishop, 2006; Murphy, 2012). For instance, there are well-known frequentist approaches that perform model averaging, including bagging (Breiman, 1996) and boosting (Schapire, 1990). As another example, the bootstrap is a frequentist method for population-level inference of confidence intervals and non-parametric null-hypothesis testing (Efron, 1979). This procedure however readily lends itself to Bayesian interpretations and often agrees with the posterior distributions from Bayesian analysis under an uninformative prior (Hastie et al., 2001, chapter 8; Hastie et al., 2015, chapter 6). As another result of their many hidden relations, frequentist and Bayesian problem solutions can often be translated into each other. Many frequentist problems relying on gradient-based optimization can be recast as Bayesian integration problems using Langevin and Hamiltonian MCMC methods (Girolami and Calderhead, 2011). Conversely, many Bayesian integration problems can be recast as frequentist optimization problems using variational Bayesian approximation methods (Jordan et al., 1999). This makes a clear-cut distinction between frequentist and Bayesian statistics less compelling.

Important for data-intensive brain science, the frequentist-Bayesian tradeoff has a critical impact on the computational budget required for model estimation (Fig. 3). As a general tendency, the more one adheres to frequentist instead of Bayesian ideology, the less computationally expensive and the less technically involved are the statistical analyses. It is a widespread opinion that Bayesian models do not scale well to the data-rich setting, although there is currently insufficient work on the behavior of Bayesian methods in high-dimensional input data (Bishop and Lasserre, 2007; Jordan, 2011; Yang et al., 2016). While the purely frequentist approach often computes maximum likelihood estimation, the purely Bayesian approach seeks to sample from the full posterior probability distributions by computing asymptotically exact MCMC. Given their computational cost, MCMCs have mainly been used for low-dimensional problems with few input variables. Many non-deterministic MCMC variants suffer from i) difficulty in diagnosing convergence to the posterior distribution, ii) hard-to-control "random-walk" behavior, or iii) limited scaling to the high-dimensional setting (MacKay, 2003, chapter 29). Fortunately, the practical applicability of Bayesian methods has been greatly enhanced through the development of deterministic procedures for approximate inference such as variational Bayes and expectation propagation (Jordan et al., 1999; Minka, 2001; Bishop, 2006, chapter 19). Consequently, the different challenges of solving Bayesian posterior integrals motivated a rich spectrum of Bayesian-frequentist hybrid models (Efron, 2005) with an increasing trend towards *incorporating appealing Bayesian aspects into computationally cheaper frequentist models* (cf., Kingma et al., 2014; Sengupta et al., 2015, 2016; Mandt et al., 2017).

In sum, the scalability of model estimation in the data-rich scenario is calibrated between frequentist numerical optimization and Bayesian numerical integration. High-dimensional data with many variables have been argued to motivate novel blends between less resource-demanding frequentist and more holistic Bayesian modeling aspects (Efron, 2005).

Method	Definition
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\|\boldsymbol{\theta})$
MAP estimation	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})$
ML-II (Empirical Bayes)	$\hat{\boldsymbol{\eta}} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})d\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\|\boldsymbol{\eta})$
MAP-II	$\hat{\boldsymbol{\eta}} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})d\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\eta}\|\mathcal{D}) \propto p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$

Fig. 3. Different shades of Bayesian inference. There is not one unique Bayesian formulation to perform statistical estimation. Rather, there are a variety of Bayesian frameworks. For instance, type-II maximum likelihood or empirical Bayes has genuine frequentist properties, does not specify a prior distribution before visiting the data, and is often used in non-Bayesian modeling. Generally, the more integrals that need to be solved or approximated in a given Bayesian formulation, the higher the computational budget needed for model estimation. Reused with permission from Murphy (2012).

brain dynamics (Friston et al., 1997). The Bayesian-frequentist distinction provides yet another angle in addition to the parametric/non-parametric and discriminative/generative perspectives on statistical models (Freedman, 1995; Roos et al. 2005; Gelman et al., 2014). Given the dominance of Bayesian statistics in the 19th century and frequentist statistics in the 20th century (Efron, 2005), one may wonder about their relative contributions in the 21st century. It is today unclear how well *fully* Bayesian models can scale to always bigger and more detailed brain data repositories. We hence speculate that *the many merits of Bayesian statistics in neuroscience research are most likely best exploited by integration with frequentist models that typically incur much lower computational burden*.

One appeal of Bayesian modeling is its intimate relationship to certain prominent hypotheses about both the workings of cognitive processes and their neural realizations. The Bayesian view of cognition is about placing expectation priors on the concepts that underlie perception and action in the ambient environment. After observing new environmental evidence, humans and intelligent animals may dynamically update the probabilistic priors on the concepts that could have produced the evidence. The goal of the neuroscience investigator would be to identify the algorithmic principles that govern how organisms solve problems of logical induction. Such an agenda is closely related to what David Marr termed the "computational" and “algorithmic” levels of brain function (Marr, 1982). Bayesian models have been argued to be an ideal choice to tackle three core questions in cognitive research (Tenenbaum et al., 2011): 1) How abstract knowledge drives learning from incomplete, noisy input, 2) How it is represented, and 3) How it is acquired? The probabilistic properties of Bayesian models are likely to be valuable for capturing uncertainty in

perception and decision-making, as well as the unavoidable presence of randomness that characterizes neuronal circuits (Faisal et al., 2008). As an example from computational psychology, Bayesian inference allowed for an explicit model of how intelligent organisms may learn new concepts from only single exposures to visual symbols (Lake et al., 2015). Each letter of an invented, never seen alphabet was represented as a combination of line stroke primitives. Bayesian inference allowed successfully browsing a large combinatorial space of stroke primitives most likely to have generated a given raw letter. The authors used a Bayesian non-parametric generative model that could even produce novel types of visual concepts by recombining parts of existing ones in creative ways. This model was also shown to outperform the discriminative, frequentist state-of-the-art model for object recognition (Lake et al., 2015). More generally, many aspects of the mind and brain can be recast as computational problems of inductive inference. Bayesian probabilistic models present a particularly attractive opportunity to decipher the mathematics of how intelligent organisms operate on and generalize from abstract concepts of world structure (Tervo et al., 2016).

When confronted with extensive brain data, we believe that the many desirable properties of Bayesian modeling and the relatively lower computational costs of frequentist models need to be balanced. In many imaging neuroscience applications, navigating the speed-accuracy tradeoff in Bayesian posterior inference has successfully reduced the computational burden. This tradeoff was achieved by using variational Bayes approximations, such as for Bayesian time-series analysis (Penny et al., 2003), model selection for group analysis (Stephan et al., 2009) and mixed-effects classification for imbalanced groups (Brodersen et al., 2013), as well as by adding constraints on macroscopic networks (Seghier and Friston, 2013) or neuronal fluctuations (Friston et al., 2014; Razi et al., 2015). In the case of *transdiagnostic* clinical neuroscience (Buckholtz and Meyer-Lindenberg, 2011; Goodkind et al., 2015; Insel and Cuthbert, 2015), hierarchical Bayesian models might gracefully handle the pervasive problem of *class imbalance* and provide certain levels of protection to selection bias (Murphy, 2012). Hierarchical Bayesian models can provide a parsimonious framework for introducing statistical dependences among multiple classes (e.g., disease groups), which might enable classes with small sample sizes (e.g., rare diseases) to borrow statistical strength from classes with larger sample sizes (e.g., related diseases). Finally, Bayesian statistics treat model parameters as random, allowing for more natural handling of *model parameter (and even structure) uncertainty* than in the frequentist regime where model parameters are assumed to be fixed (Ghahramani et al., 2013).

For instance, recent advances in non-parametric Bayesian methods (Orbanz and Teh, 2011) combined with extensive datasets promise forward progress in longstanding problems in cognitive and clinical neuroscience. As a key problem in cognition, neuroscientists have not agreed on a description system of mental operations (called 'taxonomy' or 'ontology') that would canonically motivate and operationalize their experiments (Barrett, 2009; Tenenbaum et al., 2011; Poldrack and Yarkoni, 2016). As a key problem in clinical neuroscience, partly shared neurobiological endo-phenotypes are today believed to contribute to the pathophysiology of various psychiatric and neurological diagnoses (called 'nosology') despite drastically different clinical exo-phenotypes (Brodersen et al., 2011; Hyman, 2007; Stephan et al., 2015a).

As an interesting observation, both these neuroscientific challenges can be statistically recast as latent factor problems (cf. Poldrack et al., 2012). In latent factor models (Ghahramani and Griffiths, 2006; Goodfellow et al., 2016, chapter 13), an underlying set of hidden components of variation are uncovered by assigning each observation in the brain to *each of the components to different degrees*. The same class of statistical models can potentially identify the unnamed building blocks underlying human cognition and the unknown neurobiological structure underlying diverse brain disorders. For instance, hierarchical Bayesian models were recently borrowed from the domain of text mining to estimate both a latent cognitive ontology (Yeo et al., 2015; Bertolero et al., 2015) and morphological atrophy subtypes in Alzheimer's disease (Zhang et al., 2016). Further, formal inference in non-parametric Bayesian models can potentially handle complexity in the brain by estimating the *number* of latent factors in cognition and disease using Chinese Restaurant Processes (Kemp et al., 2006; Pitman, 2006), the *relative implications* of latent causes in neurobiological observations using Indian Buffet Processes (Ghahramani and Griffiths, 2006), as well as deriving the *hierarchies* of cognitive primitives and disease endo-phenotypes using Hierarchical Dirichlet Processes (Teh et al., 2005). It is a particularly important (if not exclusive) possibility of cluster detection in the non-parametric Bayesian regime to allow each observation to participate in all clusters (e.g., Yeo et al., 2014; Moyer et al., 2015; Najafi et al., 2016). This contrasts the neurobiologically implausible 'winner-takes-all' assumption (e.g., each brain location is strictly assigned to only one cluster) of many widely used traditional clustering algorithms, including k-means, hierarchical, and ward clustering (e.g., Yeo et al., 2011; Craddock et al., 2012; Shen et al., 2013).

In sum, we propose that the statistical scalability of obtaining meaningful and accurate neuroscientific answers from extensive brain data should be balanced between the Bayesian and frequentist modeling agendas. Bayesian models enable explicitly informing model estimation by prior knowledge and they have many strengths regarding interpretational appeal, robustness to unequal group data, and in hierarchical statistical settings. While they can generalize better in the low-dimensional setting, scaling *fully* Bayesian models to handle high-dimensional data is challenging and an active area of research (cf. Breiman, 1997, Sengupta et al., 2015). Frequentist models, instead, are typically more modest in the required computation resources, are easier to use, and work faster out-of-the-box. Luckily, ingredients from both statistical regimes can be directly integrated by readjusting the modeling goal (Gopalan and Blei, 2013; Murphy, 2012, chapter 5; https://jasp-stats.org). The quantitative sciences therefore show a trend for novel blends of statistical models that are opportunistic in marrying Bayesian and frequentist advantages (Efron, 2005; Kingma et al., 2014). We predict that the recent emergence of extensive datasets in neuroscience will open a window of opportunity for exploring and exploiting more Bayesian-frequentist hybrid approaches (cf. Brodersen et al., 2011; Gilbert et al., 2016), which may for instance rely on empirical Bayes methods (Friston et al., 2016; Stephan et al., 2016). We expect that such developments will probably de-emphasize a strict dichotomy between the Bayesian and frequentist modeling philosophies in the neurosciences.

Towards diversification of statistical inference

Statistical inference is a heterogeneous notion that has recently been defined as the extraction of new knowledge from parameters in mathematical models fitted to data[3] (Jordan et al., 2013). We emphasize that classical *null-hypothesis testing* and modern *out-of-sample generalization* serve distinct statistical purposes and can be used together in practical data analysis. They perform different types of formal assessment for successful extrapolation of an effect beyond the data at hand that are embedded in different mathematical theories (for technical details see Box 4). Null-hypothesis testing evaluates whether observations are too extreme under the null hypothesis, whereas out-of-sample generalization evaluates how well fitted algorithms perform

[3] It is worth noting that in statistics, 'inference' typically refers to procedures, such as hypothesis testing and estimating conference intervals (performed within the same sample). By contrast, in machine learning, 'inference' typically refers to predicting information (e.g., hidden variables) of new data instances (i.e., out-of-sample). As such, the broader notion of inference (Jordan et al., 2013) encompasses both hypothesis testing and out-of-sample generalization.

Box 4: **Null-hypothesis testing and out-of-sample generalization**.

Statistical inference can be broadly defined as the extraction of new knowledge from parameters in mathematical models fitted to data (Jordan et al., 2013). *Classical inference* focuses on *in-sample* estimates by explained-variance metrics of the entire data sample (Fig. 4), while *pattern generalization* focuses on *out-of-sample* estimates by assessing prediction performance metrics on unseen data samples not used during model fitting (Friston, 2012 appendix). Therefore the mostly *retrospective* viewpoint of null-hypothesis testing can be contrasted with the mostly *prospective* viewpoint of the out-of-sample approach that seeks to learn a general principle from data examples and evaluate the result on unseen examples (cf. Goodman, 1999).

In classical inference, invented almost 100 years ago (cf. Fisher and Mackenzie, 1923; Neyman and Pearson, 1933), the scientist articulates two mutually exclusive hypotheses by domain-informed judgment with the agenda to disprove the null hypothesis embraced by the research community. A *p-value* is then computed that denotes the conditional probability of obtaining an equal or more extreme test statistic provided that the null hypothesis is correct. This conditional probability is conventionally set at the arbitrary significance threshold of alpha=0.05 (Wasserstein and Lazar, 2016). State-of-the-art hypotheses are continuously replaced by always more pertinent hypotheses using *verification* and *falsification* in a Darwinian process (Popper, 1935/2005). The classical framework of null-hypothesis falsification to infer new knowledge is still the go-to choice in many branches of neuroscience. Considering the data-rich scenario, it is an important problem that p-values intrinsically become better (i.e., lower) as the sample size increases because even very small effects will become significant (Berkson, 1938). Indeed, brain-behavior correlations of $r \approx 0.1$ were found to be statistically significant when considering a sample of $n = 5000$ participants even after correction for multiple comparisons (Miller et al., 2016). As reporting statistical significance alone becomes insufficient, it is now mandatory to report effect sizes in addition to or instead of p-values in certain scientific fields (Wasserstein and Lazar, 2016). Besides null-hypothesis testing, *asymptotic consistency guarantees* are a cornerstone of classical statistical theory (Fisher, 1922; Efron and Hastie, 2016). Many traditional statistics tools have been theoretically justified by demonstrating their convergence to the "truth" as the input data grow to infinity.

In contrast, out-of-sample generalization emerged much more recently as the fundamental statistical process underlying learning in animals, humans, and machines (Vapnik, 1989; Valiant 1984). It can be defined as testing whether an underlying complex pattern is learnable in a dataset (Bzdok, et al., 2016b). This inferential regime operates by necessary and sufficient conditions for generalization that have been formalized as *PAC learning* (probably approximately correct learning) from *computational complexity theory* (Valiant, 1984). This theoretical framework answers the question "Can we extrapolate a statistical relationship discovered in one set of data to another set of data in polynomial time?" Given a class of candidate functions defined by the statistical model (i.e., the hypothesis space), the PAC framework assesses the performance bounds of that model in selecting a function that is likely to yield the approximately correct result on the independent test data with high probability. The typical practical question of necessary minimum sample size is tied to the size of the hypothesis space (i.e., the number of theoretically learnable statistical relationships). Note that *PAC learnability* is a stricter statistical notion than consistency guarantees for a learning algorithm (Shalev-Shwartz et al., 2014, chapter 7).

Furthermore, a pattern generalization that is successful according to the PAC learning framework is, under mild conditions, also feasible according to the closely related notion of *Vapnik-Chervonenkis* (VC) *dimensions* (Vapnik, 1989) from *statistical learning theory* (Shalev-Shwartz et al., 2014). Analogously, the VC generalization bounds formally express the circumstances under which a class of functions is able to learn from a given finite amount of data to successfully predict a neurobiological phenomenon in unseen data (Hastie et al., 2001, chapter 7; Abu-Mostafa et al., 2012). The quantity of VC dimensions thus provides a probabilistic measure of whether a certain model is able to learn a distinction given a dataset. Good statistical models have *finite VC dimensions* as a sufficient (but not necessary) condition for successful approximation of the theoretically expected performance in unseen data. Note that finite VC dimensions imply PAC learnability according to the fundamental theorem of statistical learning (Shalev-Shwartz et al., 2014, theorem 6.7). Bad statistical models entertain a too large class of candidate functions (i.e., hypothesis space), which entails the inability for generalization conclusions on unseen data. As one of the biggest insights from statistical learning theory, the number of configurations that can result from a certain classification algorithm grows polynomially, while the error is luckily decreasing exponentially (Wasserman, 2013). In other words, in any intelligent organism or system, the opportunity to learn abstract patterns in the world eventually overweighs the difficulty of generalization to new observations. In practice, cross-validation procedures provide an accurate estimate of a model's "true" capacity to generalize to future data samples (Dwork et al., 2015).

on freshly sampled, independent data. In imaging neuroscience, the generalization performances of learning algorithms obtained from cross-validation procedures are frequently backed up by testing the null hypothesis of whether the achieved prediction performance is at chance level (Pereira et al., 2009) Box 5.

Drawing statistical inference on regional brain responses during controlled experiments has historically hinged on (parametric) classical null-hypothesis testing, but is increasingly flanked by out-of-sample generalization based on (non-parametric) cross-validation (Kriegeskorte, 2015a, 2015b; Bzdok, 2016; Yarkoni and Westfall, 2016). Classical inference measures the statistical significance associated with a relationship between typically few variables given a pre-specified model. For instance, t-tests are often used to evaluate whether the regional brain response, such as in the amygdala, is significantly different between healthy participants and psychiatric patients. In contrast, generalization inference empirically measures the robustness of patterns between typically many variables by testing how well an already fitted model extrapolates to unseen brain measurements (Hastie et al., 2001). In practice, cross-validation procedures are frequently used to quantify out-of-sample performance by an unbiased estimate of a model's capacity to generalize to data samples acquired in the future (Dwork et al., 2015; Varoquaux et al., 2016). This model assessment is done by a cycle of model fitting on a bigger subset of the available data (i.e., 'training set') and subsequent application of the trained model on the smaller remaining part of data (i.e., 'test set').

One may think that differences between the two ways of establishing neurobiological knowledge from brain measurements are mostly of technical relevance. Yet there is an often-overlooked misconception that statistical models with high explanatory power necessarily also exhibit high predictive power (Friedman, 2001; Lo et al., 2015; Shmueli, 2010; Wu et al., 2009). Put differently, a neurobiological effect assessed to be statistically significant by a p-value may some-

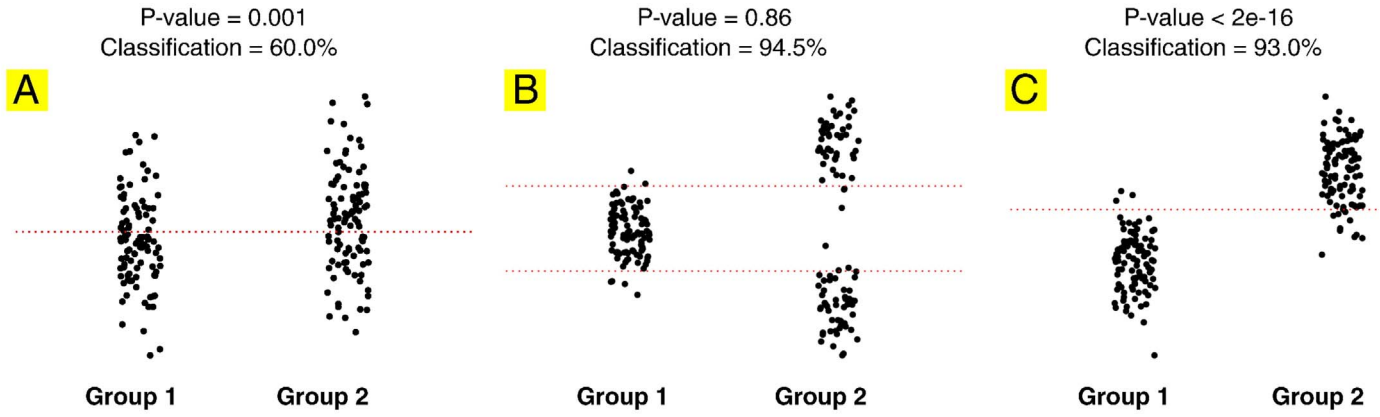

Fig. 4. Classical statistical inference and classification performance can lead to diverging conclusions. Differences between 100 brain measurements (data points) drawn from each of two groups are evaluated using two-sample t-tests ("P-value") and classification ("Classification"), where data points on either side of the dotted lines are predicted as being from different groups. In three cases with different data distributions, (A) t-test was statistically significant, while classification accuracy was poor, (B) t-test was not statistically significant, while classification accuracy was high, (C) t-test was statistically significant and classification accuracy was high. This toy example illustrates that null-hypothesis rejection and pattern recognition constitute two different statistical analyses that do not necessarily judge data distributions by the same aspects. Hence, group effects as assessed by significant p-values do not always entail a high classification performance, and vice versa. Figure reused with permission from Arbabshirani et al. (2017).

times not yield successful predictability based on cross-validation, and vice versa (cf. Fig. 4; Kriegeskorte et al., 2006). We also find it interesting to note that out-of-sample generalization with cross-validation puts the unavoidable theoretical modeling assumptions to an empirical test by directly assessing the model performance in unseen data (Kriegeskorte, 2015a, 2015b). In classical inference, the desired relevance of a statistical relationship in the general population remains grounded in formal mathematical proofs, typically without explicit evaluation on unseen data. Moreover, their many theoretical differences are more practically manifested in the high-dimensional setting where classical inference needs to address the multiple comparisons problem (i.e., accounting for many statistical inferences performed in parallel), whereas pattern generalization involves tackling the curse of dimensionality (i.e., difficulties of inferring relevant statistical structure in observations with thousands of variables) (Domingos, 2012; Friston et al., 2008; Huys et al., 2016). We therefore caution that care needs to be taken when combining both inferential regimes in practical data analysis (Bzdok, 2016; Yarkoni and Westfall, 2016).

We will now illustrate a case of "culture clash" between extrapolation based on classical inference and out-of-sample generalization. The issue has very recently gained momentum as *post-selection inference* in the statistics community (Taylor et al., 2015; Hastie et al., 2015, chapter 6.3; Efron and Hastie, 2016 chapter 16.6) and has a precursor in the neuroscientific literature as 'circular analysis' (Kriegeskorte et al., 2009; Vul et al., 2008): A neuroscientist wants to predict Alzheimer diagnosis from volumetric measurements in > 100,000 brain locations per brain scan by support vector machines with *sparsity*-inducing ℓ_1-penalization using cross-validation. Importantly, the sparsity assumption of the chosen model automatically chooses the minimal subset of variables necessary for classifying healthy versus diagnosed individuals by "silencing" the unimportant voxels with zero coefficients. In a second step, this investigator wants to test the statistical significance of the obtained non-zero voxel coefficients using classical inference to obtain p-values. In this adaptive case of initial variable selection and subsequent hypothesis testing, it is not appropriate to conduct an ordinary significance test (i.e., classical inference) on the automatically obtained sparse model coefficients (obtained from out-of-sample generalization). This would involve recasting a high-dimensional variable selection in the whole brain by one model into a setting where each brain voxel is assessed independently by many hypothesis tests (cf. Friston, 2012 appendix). Put differently, the t-test would ignore the fact that the sparse support vector machine had already visited the same data with the aim to reduce the number of variables to the most important ones (Wu et al., 2009). Applying t-tests on pre-selected variables also violates the assumption of classical statistical theory that the model is to be chosen before visiting the data. The issue in this data analysis scenario can be accounted for by the emerging tools for *post-selection inference* (Taylor and Tibshirani, 2015). These allow replacing the so-called *naive p-values* by *selection-adjusted p-values* for a set of variables that have previously been chosen to be meaningful predictors by another statistical model. This case study and similar clashes between inferential regimes will probably soon increase in the neurosciences and will encourage spurious findings if not handled appropriately (Gelman and Loken, 2014; Dwork et al., 2015).

Despite the pitfalls when combining classical inference and out-of-sample generalization, we stress that formal extrapolation determined by classical inference and pattern generalization have also been advantageously joined towards a given neuroscientific question. For instance, out-of-sample generalization estimated the relative functional contribution of the set of macroscopic brain networks (e.g., default mode network, saliency network, dorsal attention network) during a battery of psychological tasks (Bzdok et al., 2016). Classical ANOVA allowed for complementary information in finding the subsets of most explanatory networks for each psychological experiment in the task battery. Each contributed a different statistical insight into brain network function: Pattern generalization with cross-validation identified what combination of relative network recruitments best *predicts* the presence of a given psychological task in unseen brain scans. Classical statistical tools instead selected or altogether deselected which k network implications *explain most variance* with regard to a given psychological task.

More generally, the practice of performing formal cross-validation in *unseen data of the same kind* needs to be distinguished from performing informal extrapolation by showing that an effect discovered *in a first kind of data* (e.g., brain measurements) is exploited to make a new discovery *in a second kind of data* (e.g., behavioral, clinical, or genetic data). For instance, Latent Dirichlet allocation (Blei et al., 2003) was used to first find a nested hierarchy of brain volume atrophy

Box 5: **Misconceptions about "big data" in neuroscience**.

1) "**Big data is yet another hype.**"
   Massive data collection is transforming science, business, and government. In our opinion, this trend is only starting in neuroscience (Miller et al., 2016) and medicine (Collins et al., 2015). Given that brain function is barely understood, clinical care for most mental disorders often resorts to trial and error. Brain disorders have been estimated to cause ~€800 billion annual costs in Europe and to account for 13% of global burden of disease (Gustavsson et al., 2011; Mathers et al., 2008). Modern statistics applied to medical health records was estimated to create an annual value of ~US$300 billion in the US (Manyika et al., 2011) and £16 - £66 billion in the UK (House of Commons, 2016). Further, a workshop on health and analytics by the European Medicines Agency concluded that "by 2020, the amount of health-related data gathered in total will double every 73 days" (Nature Editorial, 2016). Medical care and biomedical research will probably be more and more driven by insight and intervention at the single-subject-level, rather than establishing and clinically translating group effects (Ashley, 2016). Enhanced exploitation of data can enable 'personalized medicine' to customize health care for individuals with the same disease by i) earlier detection and diagnosis of medical conditions before symptom onset, ii) predicting disease trajectories for effective patient stratification, and iii) finessing treatment decisions by predicting how well individual patients will respond to different drugs or therapies. We find it difficult to argue against the sustained benefits of statistically exploiting large data repositories in basic and clinical neuroscience. Yet, it may require readjusting the tension between data accessibility for the greater good of society and data privacy rights of every single citizen.
2) "**It is all about the data.**"
   The unconditional availability of high-quality datasets with rich meta-information is critical for neuroscience and keeps growing (Poldrack and Gorgolewski, 2014). Besides emphasizing the volume of accessible data, we believe that the central question should be what neuroscientists can actually do with it (Engert 2014; but see Anderson, 2008 and Halevy, 2009). In what ways do more brain data allow articulating and finding answers to new kinds of research hypotheses? We think that what is currently changing is the detail of knowledge that can be extracted about a given neurobiological phenomenon quantified in brain data. In our opinion, this will however require a symbiotic interplay between neuroscientific reasoning styles and statistical reasoning styles (Abbott, 2016; Goodman, 2016). The choice of statistical method constrains the spectrum of possible findings and permissible domain interpretations. Without improving statistical certainty of neuroscientific insight and without extending what can be concluded, we think that data collection initiatives will probably not live up to the considerable time, money, and human investments. Whether or not the promises of "big data" will be achieved intimately depends on the formulation of neuroscience questions and statistical model properties, which can fully leverage the unprecedented information granularity.
3) "**The more data, the better.**"
   Important neuroscientific insight has been and will be derived from hypothesis-driven, well-controlled *interventional* studies of small laboratory samples. Large consortium or population datasets typically recombine *observational* data (e.g., blood and metabolic samples, EEG, resting-state brain scans, and genetic sequencing) that were acquired without specific experimental aims. In our opinion, the more brain measurements are available, the more can potentially be learned about the brain given adequate statistical models. However, the more variables per observation are to be analyzed, the more difficult statistical modeling usually becomes. High resolution in space or time (corresponding to voxels, vertices, or time points) poses a serious statistical challenge as the so-called 'curse of dimensionality' (Hastie et al., 2001). The high-dimensional data scenario is frequently leading astray human intuition that is accustomed to regularities of a 3D world. In fact, with linear increase of variables captured in each observation, the necessary samples to populate these measurements grow exponentially, which complicates and incapacitates model estimation (Bishop, 2006). We believe that perhaps no existing statistical model would be able to yield satisfactory performance if the high-dimensional brain measurements did not have intrinsic structure leading to much lower 'effective dimensions' of interest. The tractability of model estimation in high dimensions is therefore likely to hinge on modeling approaches that can exploit the naturally existing biological compartments (e.g., brain regions and networks) in spatially and temporally fine-grained brain measurements.
4) "**The big data challenges can be tackled by hiring more staff with quantitative university degrees.**"
   Beside conceptual, statistical, and technical challenges, we believe that "big data" neuroscience also raises societal and educational issues. Making sense of extensive data collections will probably be hindered by a shortage of neuroscientists with the necessary quantitative talent. While educational opportunities for classical statistical methods are ubiquitous, systematic curricula for more modern machine learning methods currently exist at few universities (Cleveland, 2001; Donoho, 2015). Additionally, even students with a natural aptitude for mathematics and quantitative thinking typically require several years of practical experience to develop deep analytical skills (Barlow, 2013) that add to the load of traditional neuroscientific training. As a global phenomenon, 140,000 to 190,000 jobs in modern statistical analysis are expected to remain vacant in the US in 2018 due to severe talent gap (Manyika et al., 2011). This growing scarcity is also manifested in acquisitions of machine learning startups that frequently cost between $5 to $10 million per 'aqui-hired' data analyst (Henke et al., 2016). In fact, perhaps for the first time in history, the optimal skill set to become a successful (neuro)scientist is converging to the optimal skill set for a career in data-intensive industry. Many promising quantitative neuroscientists will be lured away to industry by higher salaries and better working conditions ('big data brain drain'; Vanderplas, 2013). In our opinion, the stakeholders in neuroscience research need to come up with an action plan to help close the talent gap in quantitative skills.
5) "**One can get by without programming skills.**"
   Analyzing large data collections to address neuroscientific questions requires many complicated and nested modeling choices. We would like to emphasize that the modeling choices are almost impossible to be performed by hand and exhaustively verbalized in paper publications. Automation by computer programming will become an essential toolkit addition for next-generation neuroscientists (Wilson et al., 2014). A scripted analysis pipeline defines a chain of experimental actions that can be infinitely copied for reuse in other laboratories.[4] Computational know-how manifested in programming code is increasingly shared with the international community and collaboratively evolves on social-coding platforms (e.g., www.github.com). In our opinion, the widespread adoption of script programming is likely to propel high-throughput statistical analysis, improve provenance tracking and reproducibility (Nosek et al., 2015), hence accelerating the pace of neuroscientific knowledge production.

[4] It is worth pointing out that running the same scripts might not necessarily lead to the same results because underlying software libraries (e.g., floating point libraries) might be different across computing platforms (linux versus windows). The use of containers might alleviate this issue (Poldrack et al., 2017).

endo-phenotypes in Alzheimer's disease. The clinical relevance of the atrophy endo-phenotypes was subsequently corroborated by revealing distinct decline trajectories in behavioral data on memory and executive function (Zhang et al., 2016). Additionally, informal extrapolation can also be performed based on *different kinds of neuroscientific methods* that address the same brain phenomenon. For instance, the neurobiological question "Are regions A and B connected?" can be confirmed by independent methods to quantify inter-regional coupling, such as structural and functional connectivities (Eickhoff et al., 2015). This is important because fMRI, EEG, MEG, fNIRS, and other brain imaging methods measure biological phenomena only indirectly. As such, complex processing and analysis methods are necessary to extract neuroscience discoveries from data. Extrapolating a demonstrated effect in a different modality (e.g., behavior, genetics, microbiomics) increases confidence that the findings reflect neurobiological reality. Combining different forms of validating discovered statistical relationships can therefore enhance the reproducibility of neurobiological findings (also see Nichols et al., 2017).

In sum, the leap from quantitative brain measurements to neurobiological knowledge is secured by statistical inference. We emphasize that there exists not one but several different types of statistical inference that can ask a certain neuroscientific question in different mathematical contexts that require differently nuanced neurobiological interpretations. Historically, classical inference was invented for problems with small samples that can be addressed by plausible, hand-picked models with a small number of parameters (Efron and Hastie, 2016). P-values and other classical *in-sample* estimates may therefore lose their ability to meaningfully evaluate model fit in data-rich neuroscience. Indeed, some authors emphasize that "one should *never* use sum of squared errors, p-values, R^2 statistics, or other classical measures of model fit on the training data as evidence of a good model fit in the high-dimensional setting." (James et al., 2013, p. 247, their emphasis). In contrast, we expect that out-of-sample generalization by successful cross-validation to independent data samples will be increasingly used given natural tuning to statistical estimations with more parameters and larger datasets. Moreover, out-of-sample generalization may be particularly important for a future of personalized psychiatry and neurology because cross-validated predictive models can be applied to and obtain answers from a *single patient* (Stephan et al., 2015b). Classical inference by null-hypothesis testing cannot typically produce such *intra-individual predictions* as it is constrained to using the entire data sample to test for (theoretical) extrapolation of an effect at the *population level* (Bzdok et al., 2016b; Arbabshirani et al., 2017). Ultimately, data richness will increasingly require preliminary dimensionality-reduction and feature-engineering procedures, such as k-means clustering and ICA decomposition, that do not themselves perform any type of statistical inference. We think that a back and forth between dimensionality-reducing data transformations, pattern generalization and hypothesis testing of the discovered candidate effects will become indispensable tools for understanding brain and behavior in the 21st century.

Towards deep learning models?

It is important to appreciate that some statistical models, especially modern deep neural network (DNN) algorithms, may not neatly fit into the traditional definitions of parametric versus non-parametric, discriminative versus generative, and frequentist versus Bayesian (Efron and Hastie, 2016, p. 446). DNNs excel at hierarchical non-linear classification or regression to automate feature extraction and capture higher-order statistical relationships (Schmidhuber, 2015; Goodfellow et al., 2016). Today's DNN models were recently enabled by the co-occurrence of i) increased data availability, ii) more computational resources to train always-larger DNNs, iii) a series of algorithmic advances.

More specifically, the parametric versus non-parametric distinction may become blurry for DNNs because of their high number of nested non-linear layers and possibly tens of millions of model parameters (cf. Bach, 2014; Mohamed et al., 2015; Efron and Hastie 2016; Goodfellow et al., 2016, chapter 6.2.1.2). On the one hand, DNNs practically correspond to the non-parametric notion in capturing always more complex structure with increasing input data as they utilize extremely large number of parameters and hence have a higher than necessary expressive capacity. On the other hand, DNNs do not formally satisfy the non-parametric notion of growing model parameters as data accumulate because the number of parameters is fixed. Similarly, the majority of current DNNs primarily qualify as discriminative statistical models. They can however use differentiable generator networks that take hidden variables as input to learn and draw samples from possible distributions over the data x determined by the model architecture (Goodfellow et al., 216, p. 684–686). Generative adversarial networks are an example of a discriminative-generative hybrid model, where a discriminative component distinguishes real data points as synthesized or real and its generative component aims to increase the error of the discriminative component (Goodfellow et al., 2014). Finally, many DNNs primarily qualify as frequentist models. They can however incorporate an unusual Bayesian component, such as by approximating the Bayesian posterior distribution using a separate deep inference network (Kingma et al., 2013; Kingma et al., 2014). Collectively, modern DNN approaches appear to often escape classical statistical notions.

Moreover, the tremendous success of recent DNNs in different application domains is partly due to sample sizes of $n > 1{,}000{,}000$ (LeCun et al., 2015; Jordan et al., 2015). In stark contrast, the reference datasets in brain imaging, today, reach between ~1000 participants (Human Connectome Project) and ~10,000 participants (UK Biobank Imaging), while genetic datasets are approaching the 100,000 participant margin for certain phenotypes (e.g., Psychiatric Genomics Consortium). Therefore, despite the growing literature applying DNNs to neuroscience applications (e.g., Kim et al., 2014; Plis et al., 2014; Khaligh-Razavi et al., 2014; Yamins et al., 2014; Zhang et al., 2015; Güçlü et al., 2015; Eickenberg et al., 2016; Jang et al., 2017), exploiting DNNs in neuroscience may be hindered by the brain data we currently have. Truly deploying DNNs for current neuroimaging resources would require a non-traditional formulation of neuroscience applications where for instance, the number of samples corresponds to the number of voxels or the number of time points, rather than the number of participants.

Concluding remarks and future perspectives

Following astronomy, particle physics, and genetics (Burns et al., 2014), massive data is the new reality in neuroscience and medicine. Rich datasets can extend the spectrum of possible findings and permissible conclusions about the brain. The progressively growing datasets and information granularity will, in our opinion, require a tectonic shift in data analysis practices (Bühlmann et al., 2016; Henke et al., 2016). Neuroscientists have to extend their modeling instincts towards quality of neurobiological insight that adaptively increases as data accumulate (Ghahramani, 2013; Efron and Hastie, 2016) and towards prediction on the single-subject level (Roberts et al., 2012; Arbabshirani et al., 2017; Stephan et al., 2017). Successfully adopting and flexibly switching between neuroscientific thought styles and statistical thought styles will probably turn into a precious key skill (Abbott, 2016; Goodman, 2016). We believe that next-generation PhD curricula should foster understanding of core statistical principles and include machine learning, computer programming, distributed multi-core processing, cloud computing, and advanced visualization (Akil et al., 2016; Vogelstein et al., 2016). Neuroscience is entering the era of large-scale data collection, curation, and collaboration (Poldrack and Gorgolewski, 2014) with a pressing need for statistical approaches tailored for the data-rich setting. These may frequently lie beyond the

Box 6: Trends box

- Neuroscience recently started collecting richly annotated, multi-modal datasets from hundreds and thousands of individuals managed by national, continental, and intercontinental consortia.
- Adaptive modeling approaches with non-parametric scaling can automatically increase model complexity (and potentially neurobiological insight) with increasing amount of data. We believe that non-parametric modeling strategies will therefore increasingly complement parametric statistical models.
- We believe that the widespread use of discriminative statistical models will be supplemented by more interpretable generative models that reveal biological insights into behavior and disease.
- It is our opinion that the tension between frequentist and Bayesian attitudes in statistical analysis may be relieved by hybrid models combining their advantages.
- While neurobiological knowledge is routinely inferred by null-hypothesis testing, we believe that the use of out-of-sample generalization by cross-validation is likely to grow in importance.

scope of the statistical repertoire cherished today. Analyzing extensive datasets with the most effective statistical techniques at our disposal would be an optimal use of public financial resources and our limited scientific efforts (Box 6).

Conflict of interest

We declare no conflict of interest.

Acknowledgements

We are indebted to Denis Engemann (Neurospin, Paris), Guillaume Dumas (Institut Pasteur, Paris), Daniele Marinazzo (Ghent University), Timo Dickscheid (Research Center Jülich), Klaus Willmes (RWTH Aachen University), Kristian Kersting (Technical University of Dortmund), and Dan Lurie (UC Berkeley) for insightful comments on a previous version of the manuscript. We thank Bertrand Thirion, Gaël Varoqaux, and Alexandre Gramfort (Neurospin, Paris) as well as Thomas Mühleisen (Research Center Jülich) for stimulating discussions. Further, we are grateful to the three reviewers for their very constructive feedback. YBTT is funded by Singapore MOE Tier 2 (MOE2014-T2-2-016), NUS Strategic Research (DPRT/944/09/14), NUS SOM Aspiration Fund (R185000271720), Singapore NMRC (CBRG/0088/2015), NUS YIA, and NRF Fellowship (NRF-NRFF2017-06). BD is funded by the Deutsche Forschungsgemeinschaft (DFG, BZ2/2-1, BZ2/3-1, and BZ2/4-1; International Research Training Group IRTG2150), Amazon AWS Research Grant, the German National Academic Foundation, and the START-Program of the Faculty of Medicine, RWTH Aachen.

References

Abbott, A., 2016. US mental-health chief: psychiatry must get serious about mathematics. Nat. News, (October 26).
Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H.T., 2012. Learning From Data. AMLBook, California.
Acerbi, L., Wolpert, D.M., Vijayakumar, S., 2012. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. PLoS Comput. Biol. 8, e1002771.
Akil, H., Balice-Gordon, R., Cardozo, D.L., Koroshetz, W., Norris, S.M.P., Sherer, T., Sherman, S.M., Thiels, E., 2016. Neuroscience training for the 21st century. Neuron 90, 917–926.
Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., 2012. UK Biobank: current status and what it means for epidemiology. Health Policy Technol. 1, 123–126.
Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.E., Bludau, S., Bazin, P.L., Lewis, L.B., Oros-Peusquens, A.M., Shah, N.J., Lippert, T., Zilles, K., Evans, A.C., 2013. BigBrain: an ultrahigh-resolution 3D human brain model. Science 340, 1472–1475.
Anderson, C., 2008. The end of theory: the data deluge makes the scientific method obsolete. Wired Mag. 16, 07.
Andersen, K.W., Madsen, K.H., Siebner, H.R., Schmidt, M.N., Mørup, M., Hansen, L.K., 2014. Non-parametric Bayesian graph models reveal community structure in resting state fMRI. NeuroImage *100*, 301–315.
Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017b. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145, 137–165.
Ashley, E.A., 2016. Towards precision medicine. Nat. Rev. Genet. 17.9, 507–522.
Bach, F.B., Jordan, M.I., 2005. A probabilistic interpretation of canonical correlation analysis.
Bach, F.B., 2014. Breaking the curse of dimensionality with convex neural networks. arXiv preprint arXiv 1412, 8690.
Baker, A.P., Brookes, M.J., Rezek, I.A., Smith, S.M., Behrens, T., Smith, P.J.P., Woolrich, M., 2014. Fast transient networks in spontaneous human brain activity. eLife 3, e01867.
Baldi, P., Hornik, K., 1989. Neural networks and principal component analysis: learning from examples without local minima. Neural Netw. 2, 53–58.
Ball, T.M., Stein, M.B., Ramsawh, H.J., Campbell-Sills, L., Paulus, M.P., 2014. Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology 39, 1254–1261.
Barlow, M., 2013. The Culture of Big Data. O'Reilly Media, Inc.
Barrett, L.F., 2009. The future of psychology: connecting mind to brain. Perspect. Psychol. Sci. 4, 326–339.
Beckmann, C.F., Mackay, C.E., Filippini, N., Smith, S.M., 2009. Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. NeuroImage, 47.
Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: a review and new perspectives. IEEE Trans. pattern Anal. Mach. Intell. 35.8, 1798–1828.
Berkson, J., 1938. Some difficulties of interpretation encountered in the application of the chi-square test. J. Am. Stat. Assoc. 33, 526–536.
Bertolero, M.A., Yeo, B.T.T., D'Esposito, M., 2015. The modular and integrative functional architecture of the human brain. Proc. Natl. Acad. Sci. USA 112, E6798–E6807.
Bickel, P.J., Doksum, K.A., 2007. Mathematical Statistics: Basic Ideas and Selected Topics. Pearson, Upper Saddle River, NJ.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, Heidelberg.
Bishop, C.M., Lasserre, J., 2007. Generative or discriminative? Getting the best of both worlds. Bayesian Stat. 8, 3–24.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 1997. No Bayesians in foxholes. IEEE Expert 12, 21–24.
Brodersen, K.H., Daunizeau, J., Mathys, C., Chumbley, J.R., Buhmann, J.M., Stephan, K.E., 2013. Variational Bayesian mixed-effects inference for classification studies. NeuroImage 76, 345–361.
Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7, e1002079.
Brodmann, K., 1909. Vergleichende Lokalisationslehre der Großhirnrinde, (Barth, Leipzig).
Bouchard, G., Triggs, B., 2004. . The tradeoff between generative and discriminative classifiers. In: Proceedings of the 16th IASC International Symposium on Computational Statistics (COMPSTAT'04), pp. 721–728.
Buckholtz, J.W., Meyer-Lindenberg, A., 2012. Psychopathology and the human connectome: toward a transdiagnostic model of risk for mental illness. Neuron 74, 990–1004.
Burns, R., Vogelstein, J.T., Szalay, A.S., 2014. From cosmos to connectomes: the evolution of data-intensive science. Neuron 83.6, 1249–1252.
Bühlmann, P., Drineas, P., Kane, M., van der Laan, M., 2016. Handbook of Big Data. CRC Press.
Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.
Bzdok, D., 2016. Classical Statistics and Statistical Learning in Imaging Neuroscience. arXiv Prepr. arXiv 1603, 01857.
Bzdok, D., Eickenberg, M., Grisel, O., Thirion, B., Varoquaux, G., 2015. Semi-supervised factored logistic regression for high-dimensional neuroimaging data. Advances in Neural Information Processing Systems, 3330–3338.
Bzdok, D., Varoquaux, G., Grisel, O., Eickenberg, M., Poupon, C., Thirion, B., 2016. Formal models of the network co-occurrence underlying mental operations. PLoS

Comput. Biol. http://dx.doi.org/10.1371/journal.pcbi.1004994.
Bzdok, D., Varoquaux, G., Thirion, B., 2016b. Neuroimaging research: From nullhypothesis falsification to out-of-sample generalization. Educ. Psychol. Meas., (0013164416667982).
Bzdok, D., Eickenberg, M., Varoquaux, G., Thirion, B., 2017. Hierarchical region-network sparsity for high-dimensional inference in brain imaging. Inf. Process. Med. Imaging (IPMI).
Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001. A method for making group inferences from functional MRI data using independent component analysis. Hum. Brain Mapp. 14, 140–151.
Clark, W.G., Del Giudice, J., Aghajanian, G.K., 1970. Principles of Psychopharmacology: A Textbook for Physicians, Medical Students, and Behavioral Scientists. Academic Press Inc.
Cleveland, W.S., 2001. Data science: an action plan for expanding the technical areas of the field of statistics. Int. Stat. Rev. 69, 21–26.
Collins, F.S., Varmus, H., 2015. A new initiative on precision medicine. New Engl. J. Med. 372.9, 793–795.
Craddock, R.C., James, G.A., Holtzheimer, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human. Brain Mapp. 33 (8), 1914–1928.
Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F., 2006. Consistent resting-state networks across healthy subjects. Proc. Natl. Acad. Sci. 103 (37), 13848–13853.
Deco, Gustavo, Ponce-Alvarez, Adrián, Mantini, Dante, Romani, Gian Luca, Hagmann, Patric, Corbetta, Maurizio, 2013. Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. J. Neurosci. 33 (27), 11239–11252.
Devroye, L., Györfi, L., 1996. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.
Diedrichsen, J., Hashambhoy, Y., Rane, T., Shadmehr, R., 2005. Neural correlates of reach errors. J. Neurosci. 25, 9919–9931.
Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55, 78–87.
Donoho, D., 2015. 50 years of Data Science. Tukey Centennial workshop.
Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361.
Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A., 2015. The reusable holdout: preserving validity in adaptive data analysis. Science 349, 636–638.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat., 1–26.
Efron, B., 2005. Modern Science and the Bayesian-frequentist ControversyDivision of Biostatistics. Stanford University.
Efron, B., Hastie, T., 2016. Computer-Age Statistical Inference. Cambridge University Press.
Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2016. Seeing it all: convolutional network layers map the function of the human visual system. NeuroImage.
Eickhoff, S.B., Thirion, B., Varoquaux, G., Bzdok, D., 2015. Connectivity-based parcellation: critique and implications. Hum. Brain Mapp.
Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci., 201602413.
Eliasmith, C., Stewart, T.C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., Rasmussen, D., 2012. A large-scale model of the functioning brain. Science 338, 1202–1205.
Engert, F., 2014. The big data problem: turning maps into knowledge. Neuron 83, 1246–1248.
Eshaghi, A., Riyahi-Alam, S., Saeedi, R., Roostaei, T., Nazeri, A., Aghsaei, A., Pakravan, M., 2015. Classification algorithms with multi-modal data fusion could accurately distinguish neuromyelitis optica from multiple sclerosis. NeuroImage: Clin. 7, 306–314.
Faisal, A.A., Selen, L.P., Wolpert, D.M., 2008. Noise in the nervous system. Nat. Rev. Neurosci. 9.4, 292–303.
Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., Alzheimer's Disease Neuroimaging Initiative, 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage *39* (4), 1731–1743.
Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. Lond. Ser. A 222, 309–368.
Fisher, R.A., Mackenzie, W.A., 1923. Studies in crop variation. II. The manurial response of different potato varieties. J. Agric. Sci. 13, 311–320.
Frackowiak, R., Markram, H., 2015. The future of human cerebral cartography: a novel approach. Philos. Trans. R. Soc. Lond. B: Biol. Sci. 370, 20140171.
Franklin, D.W., Wolpert, D.M., 2011. Computational mechanisms of sensorimotor control. Neuron 72, 425–442.
Freedman, D., 1995. Some issues in the foundation of statistics. Found. Sci. 1, 19–39.
Freyer, Frank, James, A. Roberts, Becker, Robert, Robinson, Peter A., Ritter, Petra, Breakspear, Michael, 2011. Biophysical mechanisms of multistability in resting-state cortical rhythms. J. Neurosci. 31 (17), 6353–6361.
Friedman, Jerome H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232.
Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6, 218–229.
Friston, K.J., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. NeuroImage 39, 181–205.
Friston, K.J., Glaser, D.E., Henson, R.N., Kiebel, S., Phillips, C., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: applications. NeuroImage 16, 484–512.
Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.
Friston, K.J., 2012. Ten ironic rules for non-statistical reviewers. Neuroimage 61, 1300–1310.
Friston, K.J., Kahan, J., Biswal, B., Razi, A., 2014. A DCM for resting state fMRI. NeuroImage 94, 396–407.
Friston, K.J., Litvak, V., Oswal, A., Razi, A., Stephan, K.E., van Wijk, B.C.M., Ziegler, G., Zeidman, P., 2016. Bayesian model reduction and empirical Bayes for group (DCM) studies. NeuroImage 128, 413–431.
Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. Nature 521, 452–459.
Ghahramani, Z., 2013. Bayesian non-parametrics and the probabilistic approach to modelling. Philos. Trans. R. Soc. A 371, 20110553.
Ghahramani, Z., Griffiths, T.L., 2006. Infinite latent feature models and the Indian buffet process. NIPS, 475–482.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, USA.
Gelman, A., Loken, E., 2014. The Statistical Crisis in Science Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. Am. Sci. 102 (6), 460–465.
Gilbert, J.R., Symmonds, M., Hanna, M.G., Dolan, R.J., Friston, K.J., Moran, R.J., 2016. Profiling neuronal ion channelopathies with non-invasive brain imaging and dynamic causal models: case studies of single gene mutations. NeuroImage 124, 43–53.
Girolami, Mark, Calderhead, Ben, 2011. Riemann manifold langevin and hamiltonian monte carlo methods. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 73.2, 123–214.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst., 2672–2680.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, USA.
Goodkind, M., Eickhoff, S.B., Oathes, D.J., Jiang, Y., Chang, A., Jones-Hagata, L.B., Ortega, B.N., Zaiko, Y.V., Roach, E.L., Korgaonkar, M.S., Grieve, S.M., Galatzer-Levy, I., Fox, P.T., Etkin, A., 2015. Identification of a common neurobiological substrate for mental illness. JAMA Psychiatry.
Goodman, S.N., 1999. Toward evidence-based medical statistics. 1: the P value fallacy. Ann. Intern. Med. 130, 995–1004.
Goodman, S.N., 2016. Aligning statistical and scientific reasoning. Science 352 (6290), 1180–1181.
Gopalan, P.K., Blei, D.M., 2013. Efficient discovery of overlapping communities in massive networks. Proc. Natl. Acad. Sci. USA 110, 14534–14539.
Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.
Halevy, A., Norvig, P., Pereira, F., 2009. The unreasonable effectiveness of data. Intell. Syst., IEEE 24, 8–12.
Harlow, J.M., 1848. Passage of an iron rod through the head. Boston Med. Surg. J. 39, 389–393.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning. Springer Ser. Stat., (Heidelberg, Germany).
Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The lasso and Generalizations. CRC Press.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110.
Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., Sethupathy, G., 2016. The Age of Analytics: Competing in a Data-Driven World (Technical report). McKinsey Global Institute.
Hinton, G.E., Salakhutdinov, Ruslan R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.
Hofmann, T., Schölkopf, B., Smola, A.J., 2008. Kernel methods in machine learning. Ann. Stat., 1171–1220.
House of Commons, 2016. The Big Data Dilemma (S.a.T.). Committee on Applied and Theoretical Statistics, UK.
Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–154.
Huys, Q.J.M., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19, 404–413.
Hyman, S.E., 2007. Can neuroscience be integrated into the DSM-V? Nat. Rev. Neurosci. 8, 725–732.
Insel, T.R., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang, P., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167, 748–751.
Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? Precisely. Science 348, 499–500.
Jang, H., Plis, S.M., Calhoun, V.D., Lee, J.-H., 2017. Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: evaluation using sensorimotor tasks. NeuroImage 145, 314–328.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.
Jara-Ettinger, J., Gweon, H., Schulz, L.E., Tenenbaum, J.B., 2016. The naïve utility calculus: computational principles underlying commonsense psychology. Trends Cogn. Sci. 20, 589–604.
Jebara, T., 2004. Machine Learning: Discriminative and Generative. Kluwer, Dordrecht.
Jebara, T., Meila, M., 2006. Machine learning: discriminative and generative. Math. Intell. 28, 67–69.
Jimura, K., Poldrack, R.A., 2012. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. Neuropsychologia 50, 544–552.

Jordan, M.I., 2011. A message from the President: the era of big data. ISBA Bull. 18, 1–3.
Jordan, M.I., 2013. Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, Council, N.R.Frontiers in MassiveData Analysis. The National Academies Press, Washington, D.C.
Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An introduction to variational methods for graphical models. Mach. Learn. 37, 183–233.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349, 255–260.
Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416.
Kandel, E.R., Markram, H., Matthews, P.M., Yuste, R., Koch, C., 2013. Neuroscience thinks big (and collaboratively). Nat. Rev. Neurosci. 14, 659–664.
Kelleher, J.D., Mac Namee, B., D'Arcy, A., 2015. Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press.
Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N., 2006. Learning Systems of Concepts with an Infinite Relational Model. Aaai, 5.
Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915.
Kim, J., Calhoun, V.D., Shim, E., Lee, J.-H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. NeuroImage 124, 127–146.
Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. Proceedings of the International Conference on Learning Representations (ICLR).
Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M., 2014. Semi-supervised learning with deep generative models. Adv. Neural Inf. Process. Syst., 3581–3589.
Knops, A., Thirion, B., Hubbard, E.M., Michel, V., Dehaene, S., 2009. Recruitment of an area involved in eye movements during mental arithmetic. Science 324, 1583–1585.
Köbbert, C., Apps, R., Bechmann, I., Lanciego, J.L., Mey, J., Thanos, S., 2000. Current concepts in neuroanatomical tracing. Progress. Neurobiol. 62, 327–351.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA 103, 3863–3868.
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12, 535–540.
Kriegeskorte, N., 2015a. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vision. Sci. 1, 417–446.
Kriegeskorte, N., 2015b. Cross-validation in brain imaging analysis. bioRxiv, 017418.
Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B., 2015. Human-level concept learning through probabilistic program induction. Science 350, 1332–1338.
Lashkari, D., Sridharan, R., Vul, E., Hsieh, P.J., Kanwisher, N., Golland, P., 2012. Search for patterns of functional specificity in the brain: a nonparametric hierarchical Bayesian model for group fMRI data. NeuroImage 59 (2), 1348–1368.
Le, Q.V., Karpenko, A., Ngiam, J., Ng, A.Y., 2011. ICA with reconstruction cost for efficient overcomplete feature learning. Adv. neural Inf. Process. Syst., 1017–1025.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lichtman, J.W., Pfister, H., Shavit, N., 2014. The big data challenges of connectomics. Nat. Neurosci. 17, 1448–1454.
Lo, A., Chernoff, H., Zheng, T., Lo, S.H., 2015. Why significant variables aren't automatically good predictors. Proc. Natl. Acad. Sci. USA 112, 13892–13897.
MacKay, D.J.C., 2003. Information Theory, Inference and Learning Algorithms. Cambridge university press.
Mandt, S., Hoffman, M.D., Blei, D.M., 2017. Stochastic gradient descent as approximate Bayesian inference. Eprint arXive 1704, 04289.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A., 2011. Big data: The Next Frontier for Innovation, Competition, and Productivity (Technical report). McKinsey Global Institute.
Marinazzo, D., Liao, W., Chen, H., Stramaglia, S., 2011. Nonlinear connectivity by Granger causality. NeuroImage 58, 330–338.
Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., Mourão-Miranda, J., 2010. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage *49* (3), 2178–2189.
Marr, D.C., 1982. Vision. W.H. Freeman and Company, San Francisco, CA.
Medaglia, J.D., Lynall, M.E., Bassett, D.S., 2015. Cognitive network neuroscience. J Cogn. Neurosci. 27, 1471–1491.
Mesulam, M., 2012. The evolving landscape of human cortical connectivity: facts and inferences. NeuroImage 62, 2182–2189.
Mesulam, M.M., 1978. Tetramethyl benzidine for horseradish peroxidase neurohistochemistry: a non-carcinogenic blue reaction product with superior sensitivity for visualizing neural afferents and efferents. J. Histochem. Cytochem. 26, 106–117.
Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci.
Minka, T.P., 2001. Expectation propagation for approximate Bayesian inference. In: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp. 362–369.
Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53, 103–118.
Mohamed, S., 2015. A statistical view of deep learning.
Moyer D., Gutman B., Prasad G., Faskowitz J., Ver Steeg G., Thompson P., 2015. Blockmodels for connectome analysis. In: Proceedings of the 11th International Symposium on Medical Information Processing and Analysis (SIPAIM 2015), International Society for Optics and Photonics, 96810A–96810A.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT press.
Najafi, M., McMenamin, B.W., Simon, J.Z., Pessoa, L., 2016. Overlapping communities reveal rich structure in large-scale brain networks during rest and task conditions. NeuroImage. 135, 92–106.
Nature Editorial, 2016. The power of big data must be harnessed for medical progress. November, 24.
Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests for statistical hypotheses. Philos. Trans. R. Soc. A: Math., Phys. Eng. Sci. 231, 289–337.
Ng, A.Y., Jordan, M.I., 2002. On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. Adv. neural Inf. Process. Syst. 14, 841.
Nichols, T.E., Das, S., Eickhoff, S.B., Evans, A.C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M.P., Poldrack, R.A., Poline, J.-B., 2016. Best practices in data analysis and sharing in neuroimaging using MRI. bioRxiv, 054262.
Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–1296.
Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Yarkoni, T., 2015. Promoting an open research culture. Science 348, 1422–1425.
Orbanz, P., Teh, Y.W., 2011. Bayesian nonparametric modelsEncyclopedia of Machine Learning. Springer, 81–89.
Penfield, W., Perot, P., 1963. The Brain? S record of auditory and visual experience. Brain 86, 595–696.
Penny, W., Kiebel, S., Friston, K., 2003. Variational Bayesian inference for fMRI time series. NeuroImage 19, 727–741.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, 199–209.
Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: a comparative study. Neuroimage 56, 476–496.
Pitman, J., 2006. Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002. Springer.
Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J.A., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8, 229.
Poldrack, R.A., Gorgolewski, K.J., 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci. 17, 1510–1517.
Poldrack, R.A., Mumford, J.A., Schonberg, T., Kalar, D., Barman, B., Yarkoni, T., 2012. Discovering relations between mind, brain, and mental disorders using topic mapping. PLoS Comput. Biol. 8, e1002707.
Poldrack, R.A., Yarkoni, T., 2016. From brain maps to cognitive ontologies: informatics and the search for mental structure. Annu. Rev. Psychol. 67, 587–612.
Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.-B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci.
Popper, K., 1935/2005. Logik der Forschung 11th ed. Mohr Siebeck, Tübingen.
Randlett, O., Wee, C.L., Naumann, E.A., Nnaemeka, O., Schoppik, D., Fitzgerald, J.E., Portugues, R., Lacoste, A.M.B., Riegler, C., Engert, F., Schier, A.F., 2015. Whole-brain activity mapping onto a zebrafish brain atlas. Nat. Methods 12, 1039–1046.
Rasmussen, C.E., 2006. Gaussian Process. Mach. Learn.
Razi, A., Kahan, J., Rees, G., Friston, K.J., 2015. Construct validation of a DCM for resting state fMRI. NeuroImage 106, 1–14.
Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H.R., Boomsma, D.I., Cichon, S., 2013. A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry 18, 497–511.
Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H., 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Human. Genet. 69, 138–147.
Roberts, N.J., Vogelstein, J.T., Parmigiani, G., Kinzler, K.W., Vogelstein, B., Velculescu, V.E., 2012. The predictive capacity of personal genome sequencing. Sci. Transl. Med. 4, (133ra158-133ra158).
Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., Tirri, H., 2005. On discriminative Bayesian network classifiers and logistic regression. Mach. Learn. 59, 267–296.
Rosen-Zvi, Michal, Chemudugunta, Chaitanya, Griffiths, Thomas, Smyth, Padhraic, Steyvers, Mark, 2010. Learning author-topic models from text corpora. ACM Trans. Inf. Syst. (TOIS) 28 (1), 4.
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117.
Schrouff, J., Mourão-Miranda, J., Phillips, C., Parvizi, J., 2016. Decoding intracranial EEG data with multiple kernel learning method. J. Neurosci. Methods 261, 19–28.
Seeley, W.W., Menon, V., Schatzberg, A.F., Keller, J., Glover, G.H., Kenna, H., Reiss, A.L., Greicius, M.D., 2007. Dissociable intrinsic connectivity networks for salience processing and executive control. J. Neurosci. 27, 2349–2356.
Seghier, M.L., Friston, K.J., 2013. Network discovery with large DCMs. NeuroImage 68, 181–191.
Sengupta, B., Friston, K.J., Penny, W.D., 2015. Gradient-free MCMC methods for dynamic causal modelling. NeuroImage 112, 375–381.
Sengupta, B., Friston, K.J., Penny, W.D., 2016. Gradient-based MCMC samplers for dynamic causal modelling. NeuroImage 125, 1107–1118.
Shalev-Shwartz, S., Ben-David, S., 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.
Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5, 197–227.
Sharp, K., Wiegerinck, W., Arias-Vasquez, A., Franke, B., Marchini, J., Albers, C.A., Kappen, H.J., 2016. Explaining missing heritability using Gaussian process

regression. bioRxiv, 040576.
Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. NeuroImage *82*, 403–415.
Shmueli, G., 2010. To explain or to predict? Stat. Sci., 289–310.
Sing, G.C., Joiner, W.M., Nanayakkara, T., Brayanov, J.B., Smith, M.A., 2009. Primitives for motor adaptation reflect correlated neural tuning to position and velocity. Neuron 64, 575–589.
Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., Beckmann, C.F., 2009. Correspondence of the brain's functional architecture during activation and rest. Proc. Natl. Acad. Sci. USA 106, 13040–13045.
Sporns, O., 2013. Network attributes for segregation and integration in the human brain. Curr. Opin. Neurobiol. 23, 162–171.
Stephan, K.E., Binder, E.B., Breakspear, M., Dayan, P., Johnstone, E.C., Meyer-Lindenberg, A., Schnyder, U., Wang, X.-J., Bach, D.R., Fletcher, P.C., 2015a. Charting the landscape of priority problems in psychiatry, part 2: pathogenesis and aetiology. Lancet Psychiatry.
Stephan, K.E., Iglesias, S., Heinzle, J., Diaconescu, A.O., 2015b. Translational perspectives for computational neuroimaging. Neuron 87, 716–732.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. NeuroImage 46, 1004–1017.
Stephan, K.E., Schlagenhauf, F., Huys, Q.J.M., Raman, S., Aponte, E.A., Brodersen, K.H., Rigoux, L., Moran, R.J., Daunizeau, J., Dolan, R.J., 2017. Computational neuroimaging strategies for single patient predictions. NeuroImage 145, 180–199.
Taylor, J., Tibshirani, R.J., 2015. Statistical learning and selective inference. Proc. Natl. Acad. Sci. USA 112, 7629–7634.
Teh, Y.W., Jordan, M.I., 2010. Hierarchical Bayesian nonparametric models with applications. Bayesian Nonparametrics, 1.
Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2005. Sharing Clusters Among Related Groups: hierarchical Dirichlet Processes. Adv. neural Inf. Process. Syst.
Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: statistics, structure, and abstraction. Science 331, 1279–1285.
Tervo, D.G.R., Tenenbaum, J.B., Gershman, S.J., 2016. Toward the neural implementation of structure learning. Curr. Opin. Neurobiol. 37, 99–105.
Valiant, L.G., 1984. A theory of the learnable. Commun. ACM 27.11, 1134–1142.
Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., Consortium, W.U.-M.H., 2012. The Human Connectome Project: a data acquisition perspective. NeuroImage 62, 2222–2231.
Vanderplas, J., 2013. The big data brain drain: why science is in trouble. Blog post.
Vapnik, V.N., 1989. Statistical learning theory. Wiley-Interscience, New York.
Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2016. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage.
Vogelstein, J.T., Amunts, K., Andreou, A., Angelaki, D., Ascoli, G., Bargmann, C., Burns, R., Cali, C., Chance, F., Chun, M., 2016. To the Cloud! A grassroots proposal to accelerate brain science discovery. Neuron, press.
Vul, E., Harris, C., Winkielman, P., Pashler, H., 2008. Voodoo correlations in social neuroscience. Psychol. Sci.
Walton, G.L., Paul, W.E., 1901. Contribution to the study of the cortical sensory areas. Brain 24, 430–452.
Wasserman, L., 2013. All of Statistics: A Concise Course in Statistical Inference. Springer Science & Business Media.
Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. Am. Stat., (00-00).
Wilson, G., Aruliah, D.A., Brown, C.T., Hong, N.P.C., Davis, M., Guy, R.T., Waugh, B., 2014. Best practices for scientific computing. PLoS Biol. 12, e1001745.
Wolpert, D.M., Diedrichsen, J., Flanagan, J.R., 2011. Principles of sensorimotor learning. Nat. Rev. Neurosci. 12, 739–751.
Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K., 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25, 714–721.
Xue, J.-H., Titterington, D.M., 2008. Comment on “On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes”. Neural Process. Lett. 28, 169–187.
Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. 111, 8619–8624.
Yang, Y., Wainwright, M.J., Jordan, M.I., 2016. On the computational complexity of high-dimensional Bayesian variable selection. Ann. Stat. 44, 2497–2532.
Yarkoni, T., Westfall, J., , 2016. Choosing prediction over explanation in psychology: Lessons from machine learning.
Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zollei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol. 106, 1125–1165.
Yeo, B.T.T., Krienen, F.M., Chee, M.W.L., Buckner, R.L., 2014. Estimates of segregation and overlap of functional connectivity networks in the human cerebral cortex. NeuroImage 88, 212–227.
Yeo, B.T.T., Krienen, F.M., Eickhoff, S.B., Yaakub, S.N., Fox, P.T., Buckner, R.L., Asplund, C.L., Chee, M.W., 2015. Functional specialization and flexibility in human association cortex. Cereb. Cortex 25, 3654–3672.
Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., Alzheimer's Disease Neuroimaging Initiative, 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clin. 2, 735–745.
Zhang, X., Mormino, E.C., Sun, N., Sperling, R.A., Sabuncu, M.R., Yeo, B.T.T., 2016. Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. Proc. Natl. Acad. Sci. USA 113.42, E6535–E6544.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Alzheimer's Disease Neuroimaging Initiative, 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55, 856–867.
Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42, 355–360.
Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D., 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 108, 214–224.
Zhong, M., Lotte, F., Girolami, M., Lécuyer, A., 2008. Classifying EEG for brain computer interfaces using Gaussian processes. Pattern Recognit. Lett. 29, 354–359.