HAL
archives-ouvertes.fr

Neuroimaging Research: From Null-Hypothesis Falsification to Out-of-sample Generalization

Danilo Bzdok, Gaël Varoquaux, Bertrand Thirion

▶ **To cite this version:**

Danilo Bzdok, Gaël Varoquaux, Bertrand Thirion. Neuroimaging Research: From Null-Hypothesis Falsification to Out-of-sample Generalization. Educational and Psychological Measurement, SAGE Publications, 2016. <hal-01338313>

HAL Id: hal-01338313
https://hal.archives-ouvertes.fr/hal-01338313

Submitted on 28 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Number of words in text: 2,700

Neuroimaging Research:

From Null-Hypothesis Falsification to Out-of-sample Generalization

Danilo Bzdok[1,2,3,*], Gaël Varoquaux[3], Bertrand Thirion[3]

1 Department of Psychiatry, Psychotherapy and Psychosomatics, Medical Faculty, RWTH Aachen, Germany
2 JARA, Translational Brain Medicine, Aachen, Germany
3 Parietal team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France

* Correspondance should be addressed to
Prof. Dr. Dr. Danilo Bzdok
Department for Psychiatry, Psychotherapy and Psychosomatics
Pauwelsstraße 30
52074 Aachen
Germany
mail: danilo[DOT]bzdok[AT]rwth-aachen[DOT]de

Abstract

Brain imaging technology has boosted the quantification of neurobiological phenomena underlying human mental operations and their disturbances. Since its inception, drawing inference on neurophysiological effects hinged on classical statistical methods, especially, the general linear model. The tens of thousands variables per brain scan were routinely tackled by independent statistical tests on each voxel. This circumvented the curse of dimensionality in exchange for neurobiologically imperfect observation units, a challenging multiple comparisons problem, and limited scaling to currently growing data repositories. Yet, the always-bigger information granularity of neuroimaging data repositories has lunched a rapidly increasing adoption of statistical learning algorithms. These scale naturally to high-dimensional data, extract models from data rather than prespecifying them, and are empirically evaluated for extrapolation to unseen data. The present paper portrays commonalities and differences between long-standing *classical inference* and upcoming *generalization inference* relevant for conducting neuroimaging research.

Keywords: neuroscience, statistical inference, epistemology, hypothesis testing, cross-validation

Introduction

While the brain-imaging domain has long been dominated by analysis approaches rooted in classical statistics, the changing dataset properties and increasing availability of statistical learning techniques have encouraged the more frequent use of statistical learning techniques in many quantitative domains (House of Common, 2016; Jordan et al., 2013; Manyika et al., 2011). This recent change in data analysis styles has engendered uneasiness and misunderstanding. The goal of the present paper is to disentangle classical inference by null-hypothesis testing and pattern recognition by out-of-sample generalization in neuroimaging with respect to their historical trajectories, conceptual frameworks, and interpretational differences.

During the past 15 years, neuroscientists have transitioned from exclusively qualitative reports of few patients with neurological brain lesions to quantitative lesion-symptom mapping on the voxel level in hundreds of patients (Bates et al., 2003). We have gone from manually staining and microscopically inspecting single brain slices to 3D models of neuroanatomy at micrometer scale (Amunts et al., 2013). We have also gone from individual experimental studies to the increasing possibility of automatized knowledge aggregation across thousands of previously isolated neuroimaging findings (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). Rather than laboriously collecting and publishing in-house data in a single paper, investigators are now routinely reanalyzing multi-modal data repositories managed by national, continental, and inter-continental consortia (Kandel, Markram, Matthews, Yuste, & Koch, 2013; Markram, 2012; Poldrack & Gorgolewski, 2014; Van Essen et al., 2012). The granularity of neuroimaging datasets is hence growing in terms of scanning resolution, sample size, and complexity of meta-information (S. Eickhoff, Turner, Nichols, & Van Horn, 2016; Van Horn & Toga, 2014). Consequently, the scope of neuroimaging analyses has expanded from the predominance of null-hypothesis testing to statistical-learning methods that are i) more data-driven by flexible models, ii) more often scalable to high-dimensional data, and iii) more heuristic by increased reliance on numerical optimization (S. B. Eickhoff, Thirion, Varoquaux, & Bzdok, 2015; Jordan & Mitchell, 2015; LeCun, Bengio, & Hinton, 2015). *Statistical learning* (Hastie,

Tibshirani, & Friedman, 2001) henceforth comprises the umbrella of "machine learning", "data mining", "pattern recognition", "knowledge discovery", and "high-dimensional statistics".

The current shift from almost exclusive classical to increasingly solicited learning methods adds new categories of inference to the imaging neuroscientist's arsenal (Bzdok, 2016; B. Efron & Hastie, 2016). Indeed, classical statistics based on null-hypothesis falsification and statistical learning based on out-of-sample generalization capture partly diverging properties of the neurophenomenon under study (cf. Cohen, 1990; Gigerenzer & Murray, 1987). CS are mostly used in neuroimaging to answer *where* significant effects can be localized in the brain in an representational agenda, whereas SL are mostly used to answer *whether* relevant patterns can be found in neural activity in an informational agenda. Additionally, the quality of CS findings are judged based on explanatory power while SL findings are judged by predictive power, results of which may diverge in practice (Lo, Chernoff, Zheng, & Lo, 2015; Wu, Chen, Hastie, Sobel, & Lange, 2009). Statistical methods can be conceptualized as spanning a continuum between the two poles of classical statistics (CS) and statistical learning (SL) (B. Efron & Hastie, 2016; Jordan et al., 2013; p. 61), while their relationship has seldom been explicitly characterized in mathematical terms (see already Bradley Efron, 1978). Intuitively, the truth is believed to be in the model (cf. Wigner, 1960) in the CS regime, while it is believed to be in the data (cf. Halevy, Norvig, & Pereira, 2009) in a SL regime. It has indeed been previously stated that "one does not need to learn (from data) what one can infer (from the current model). Moreover, one does not need to infer what one can learn (intractable inferential procedures can be circumvented by collecting data)" (Jordan, 2010). Importantly, CS has mostly been fashioned for problems with small samples that can be grasped by plausible, handpicked models with a small number of parameters operating in an analytical fashion. SL was mostly fashioned for problems with many variables in potentially large samples with rare knowledge of the data-generating process emulated by a data-derived mathematical function by a machine in a heuristic fashion. Any choice of statistical method for a neurobiological investigation predetermines the spectrum of possible results and permissible

conclusions. This choice has recently become more difficult as SL models are turned into a trusted option of statistical machinery while models from CS have previously enjoyed monopoly.

From the perspective of the neuroimaging practitioner, both CS and SL have the common goal to extract neurobiological insight from brain scans. Both statistical regimes are applied to infer the mechanisms in nature that explain the neural correlates underlying psychological processes, relationships between questionnaires and brain connectivity measures, or the relevant differences in brain structure between normal and diseased populations. Both statistical families establish such brain-behavior associations by seeing through the noise and assessing a notion of how likely the relationship would be replicated in other data.

Despite these general similarities, CS and SL can be used to answer different questions in everyday neuroimaging analysis. First, tools from CS are the more natural choice when the investigator wishes to make a judgment about statistical relationships present in data collected in a retrospective fashion. SL tools are the natural choice when explicit judgment on future, yet-to-be-acquired behavioral and neuroimaging data is the aim of statistical modelling. Second, the CS framework has more explicitly evolved to answer questions about group differences (first group analyses were on potatoe varieties and bere brews), whereas the SL framework was more generally motivated by extrapolating patterns from input-output data (e.g., computer algorithms that learn board games or flying a helicopter from experience). In practice, however, SL has legitimately and successfully been used to find relevant neurobiological differences between different groups of individuals, for instance to find biomarkers (Gabrieli, Ghosh, & Whitfield-Gabrieli, 2015; Wager et al., 2013). Third, SL methods may turn out to be the more pertinent option when the investigator aims at predicting disease trajectories and drug responses in *single* brain acquisitions in a *single* individual. SL is therefore likely to play an increasingly important role for the upcoming trend of personalized medicine that is based on *intra-individual prediction* rather than *population-level inference*.

Classical inference

Neuroscientists have generated new insight for more than a century without strong reliance on statistical methodology by brain lesion reports (Broca, 1865; Harlow, 1848; Wernicke, 1881), microscopical inspection (Brodmann, 1909; Vogt & Vogt, 1919), and pharmacological intervention (Clark, Del Giudice, & Aghajanian, 1970). The advent of neuroimaging methods (Fox & Raichle, 1986) then enabled a much more easily quantifiable characterization of the brain systems underlying sensory, cognitive, or affective tasks. Ever since, topographical localization of neural activity effects was largely dominated by analysis approaches from CS, in particular the general linear model (GLM) (K. J. Friston et al., 1994). Widespread CS tools, such as standard deviation, p values, ANOVA, and hypothesis testing, had already been invented in the beginning of the 20th century and dominated statistical analysis in academia ever since (Cowles & Davis, 1982; Fisher & Mackenzie, 1923; Neyman & Pearson, 1933). In the neuroimaging domain, they were and still are routinely used in a mass-univariate regime by computing univariate statistics for activity observation in each voxel independently (K. J. Friston et al., 1994). It involves fitting beta coefficients corresponding to the columns of a *design matrix* (i.e., prespecified stimulus/task/behavior indicators, the independent variables) to a single voxel's imaging time series of measured neural activity changes (i.e., *dependent variable*) to obtain a beta coefficient per indicator. The ensuing multiple comparisons problem motivated more than two decades of methodological research (K.J. Friston, 2006; Thomas E. Nichols, 2012; Smith, Matthews, & Jezzard, 2001; Worsley, Evans, Marrett, & Neelin, 1992). It was early acknowledged that the unit of interest for the null hypothesis should be spatially neighboring voxel groups (Chumbley & Friston, 2009), which motivated the use of *random field theory* (Worsley et al., 1992) during model inference to alleviate the multiple comparisons problem.

CS operates by continuously replacing current hypotheses by always more pertinent hypotheses using *verification* and *falsification*. The rationale behind hypothesis falsification is that one counterexample can reject a theory by *deductive reasoning* (Goodman, 1999). The neuroscientist verbalizes two mutually exclusive hypotheses by domain-informed judgment. The investigator has

the agenda to disprove the null hypothesis because it only leaves the preferred alternative hypothesis as the new standard belief. If the data have a probability of <=5% to be true given the null hypothesis (P(result|H0)), it is conventionally evaluated to be significant. The *p value* denotes the conditional probability of obtaining an equal or more extreme test statistic provided that the null hypothesis H_0 is true at the prespecified significance threshold alpha (Anderson, Burnham, & Thompson, 2000). If the null hypothesis is not rejected, the test yields no conclusive result instead of a null result (Schmidt, 1996). In this way, classical hypothesis testing continuously replaces currently embraced hypotheses explaining a phenomenon in nature by better hypotheses with more empirical support in a Darwinian selection process. In the current "big-data" era, it is an important caveat that p values become better (i.e., lower) with increasing sample sizes (Berkson, 1938).

Statistical Learning

In the neuroimaging literature, it is seldom mentioned that the GLM would not have been solvable for unique solutions in the high-dimensional regime because the number of input variables p exceeded by far the number of samples n, which etnails an under-determined system of equations. This scenario incapacitates most statistical estimators from CS (cf. Giraud, 2014; Hastie, Tibshirani, & Wainwright, 2015). Regularization by shrinkage- and sparsity-inducing norms, such as in modern regression analysis via Lasso, ElasticNet, and RidgeRegression (cf. Hastie et al., 2015; Jenatton, Audibert, & Bach, 2011), emerged only later as a principled way to de-escalate the need for *dimensionality reduction* and to enable the tractability of the high-dimensional "p > n" case (Tibshirani, 1996). Note that the high-dimensional scenario is only challenging if the p variable are considered explanatory, not if they are considered as the data to be explained. Despite early approaches to "multivariate" brain-behavior associations (cf. K.J. Friston et al., 2008; Worsley, Poline, Friston, & Evans, 1997), the popularity of SL methods only peaked after being rebranded as "mind-reading", "brain decoding", and "multivariate pattern analysis" (Haynes & Rees, 2005; Kamitani & Tong, 2005). The conceptual appeal of this new access to the neural correlates of psychological and

pathophysiological processes was flanked by the availability of the necessary computing power and memory resources. Last but not least, there is an always-bigger interest in and pressure for data sharing, open access, and building "big-data" repositories in neuroscience (Devor et al., 2013; Gorgolewski et al., 2014; Kandel et al., 2013; Markram, 2012; Poldrack & Gorgolewski, 2014; Van Essen et al., 2012). As the dimensionality and complexity of neuroimaging datasets increases, neuroscientific investigations will probably benefit increasingly from SL methods and their variants adapted to the data-intense setting (e.g., Bzdok et al., 2016; Engemann & Gramfort, 2015; Kleiner, Talwalkar, Sarkar, & Jordan, 2012; Zou, Hastie, & Tibshirani, 2006).

It is important to emphasize that these SL algorithms came in a variety of flavors. They can be more easily interpretable (e.g., linear support vector machines) or less interpretable (e.g., kernel-based or ensemble-based models), be *parametric* with predefined model structure (e.g., hidden Markov models) or *non-parametric* with adaptive model structure (e.g., hierarchical dirichlet processes), and they can be grouped as *discriminative* (e.g., logistic regression) versus *generative* (e.g., latent factor models including independent component analysis).

The null-hypothesis testing framework in CS finds a close relative in the concept of *Vapnik-Chervonenkis dimensions* in *statistical learning theory*. The VC dimensions mathematically formalize the circumstances under which a pattern-learning algorithm can successfully distinguish between points and extrapolate to new examples (Vapnik, 1989, 1996). Note that this *inductive logic* to learn a general principle from examples contrasts the deductive logic of hypothesis falsification. In particular, VC dimensions provide a probabilistic measure of whether a certain model is able to learn a distinction given a dataset. As one of the most important results from SL theory, the number of configurations one can obtain from a classification algorithm grows polynomially, while the error is decreasing exponentially (Wasserman, 2013). Like degrees of freedom in null-hypothesis testing, the VC dimensions are unrelated to the *target function,* as the "true" mechanisms underlying the studied phenomenon in nature. Although the VC dimensions are the best formal concept to derive errors bounds in SL theory (Abu-Mostafa, Magdon-Ismail, & Lin, 2012), they can only be explicitly computed

for simple models. In practice, the out-of-sample performance is evaluated by *cross-validation.* This is the de facto standard to obtain an unbiased estimate of a model's capacity to generalize beyond the sample at hand (Bishop, 2006; Hastie et al., 2001). *Model assessment* is done by training on a bigger subset of the available data (i.e., *training set* for *in-sample performance*) and subsequent application of the trained model to the smaller remaining part of data (i.e., *test set* for *out-of-sample performance*), which is assumed to share the same distribution. Cross-validation thus permutes over the sample in data splits until the class label (i.e., categorical target variable) of each data point has been predicted once.

Relationship between classical inference and statistical learning in neuroimaging research

There is an *often-overlooked misconception that models with high explanatory power do necessarily exhibit high predictive power* (Lo et al., 2015; Wu et al., 2009). On a general basis, *CS and SL do not judge findings by the same aspects of evidence* (Lo et al., 2015; Shmueli, 2010). In neuroimaging papers based on classical hypothesis-driven inference p values (and less often confidence intervals) are ubiquitously reported. It has been previously emphasized (K.J. Friston, 2012) that p values and effect sizes reflect in-sample estimates in a retrospective inference regime (CS). These metrics find an analogue in out-of-sample estimates issued from cross-validation in a prospective prediction regime (SL). While the retrospective (CS) versus prospective (SL) distinction is an important property of the conceptual frameworks, fitted CS model can also be used in a prospective aim and SL models are routinely estimated based on existing data in practice. In-sample effect sizes are typically an *optimistic* estimate of the "true" effect size (inflated by high significance thresholds), whereas out-of-sample effect sizes can be *unbiased* estimates of the "true" effect size. Note however that in-sample estimates are unbiased and thus not optimistic if the "true" model is known. In-sample effect sizes are a priori unbiased estimates of the true effects. However, in current practice, these effects are often selected and measured on the same dataset, yielding to optimistic bias. Reporting instead effects sizes estimated on the test set does not suffer from such biases.

When looking at neuroimaging research through the CS lens, statistical estimation revolves around solving the *multiple comparisons problem* (Thomas E. Nichols, 2012; T.E. Nichols & Hayasaka, 2003). From the SL stance, however, it is the *curse of dimensionality* and *overfitting* that statistical analyses need to tackle (Domingos, 2012; K.J. Friston et al., 2008). In typical neuroimaging studies, CS methods typically test one hypothesis many times (i.e., the null hypothesis), whereas SL methods typically search through thousands of different hypotheses in a single process (i.e., walking through the function space by numerical optimization) (MacKay, 2003, chapter 41). The high voxel resolution of common brain scans offers parallel measurements of >100,000 brain locations. In a mass-univariate regime, such as after fitting voxel-wise GLMs, the same statistical test is applied >100,000 times. The more often the investigator tests a hypothesis of relevance for a brain location, the more locations will be falsely detected as relevant (false positive, Type I error), especially in the noisy neuroimaging data. The issue consists in too many simultaneous statistical inferences. From a general perspective, all dimensions in the data (i.e., voxel variables) are implicitly treated as equally important and no neighborhoods of most expected variation are statistically exploited (Hastie et al., 2001). Hence, the absence of complexity restrictions during the statistical modelling of neuroimaging data takes a heavy toll at the final inference step.

As an intriguing hybrid analysis between classical and learning regimes, the "searchlight" approach for "pattern-information analysis" enabled *whole-brain* assessment of *local* neighborhoods of predictive patterns of neural activity fluctuations (N. Kriegeskorte, Goebel, & Bandettini, 2006). It combines the generalization of signals carried in sets of local voxel signals (i.e., whehter?; unit of observation: voxel groups) but nevertheless quantifies the brain-behavior associations for each brain voxel at the brain-global level akin to mass-univariate analyses (i.e., where?; unit of observation: single voxels of the whole brain).

In neuroimaging research, statistical analysis grounded in CS and SL is closely related to *encoding models* and *decoding models*, respectively (Nikolaus Kriegeskorte, 2011; Naselaris, Kay, Nishimoto, & Gallant, 2011; Pedregosa, Eickenberg, Ciuciu, Thirion, & Gramfort, 2015). Encoding models regress

the data against a design matrix with potentially many explanatory columns of stimulus (e.g., face versus house pictures), task (e.g., to evalute or to attend), or behavioral (e.g., age or gender) indicators by fitting general linear models. In contrast, decoding models typically predict these indicators by training and testing classification algorithms on different splits from the whole dataset. In CS parlance, the encoding model fits the neural activity data by the *beta coefficients*, the *dependent variables*, according to the indicators in the *design matrix* columns, the *independent variables*. An explanation for decoding models in SL jargon would be that the *model weights* of a *classifier* are fitted on the *training set* of the *input data* to *predict* the *class labels*, the *target variables,* and are subsequently evaluated on the *test set* by *cross-validation* to obtain their *out-of-sample generalization performance.* Put differently, a GLM fits coefficients of stimulus/task/behavior indicators on neural activity data for each voxel separately given the design matrix (Naselaris et al., 2011), while classifiers predict entries of the design matrix for all voxels simultaneously given the neural activity data (Pereira, Mitchell, & Botvinick, 2009). A key difference between CS-mediated encoding models and SL-mediated decoding models thus pertains to the *direction of inference* between brain space and indicator space (K.J. Friston et al., 2008; Varoquaux & Thirion, 2014). These considerations also reveal the intimate relationship of CS models to the notion of so-called *forward inference,* while SL relate to formal *reverse inference* in functional neuroimaging (S. B. Eickhoff et al., 2011; Poldrack, 2006; Varoquaux & Thirion, 2014; Yarkoni et al., 2011). *Forward inference* relates to encoding models by testing the probability of observing activity given knowledge of a psychological process, while *reverse inference* relates to brain "decoding" by testing the probability of a psychological process being present given knowledge of activation in a brain location.

This is contrasted by the high-dimensional SL regime, where the initial model choice by the investigator determines the complexity restrictions to all data dimensions (i.e., not single voxels) that are imposed explicitly or implicitly by the model structure. Model choice predisposes existing but unknown low-dimensional neighborhoods in the full voxel space to achieve the prediction task. Here, the toll is taken at the beginning because there are so many different alternative model choices that

would impose a different set of complexity constraints. For instance, signals from "brain regions" are likely to be well approximated by models that impose discrete, locally constant compartments on the data (e.g., k-means or spatially constrained Ward clustering). Tuning model choice to signals from macroscopical "brain networks" should impose overlapping, locally continuous data compartments (e.g., independent component analysis (Beckmann, DeLuca, Devlin, & Smith, 2005) or sparse principal component analysis (Zou et al., 2006)). Knowledge of such *effective dimensions* in the neuroimaging data is a rare opportunity to simultaneously reduce the model bias *and* model variance, despite their typically inverse relationship. Statistical models that overcome the curse of dimensionality typically incorporate an explicit or implicit metric for such anisotropic neighborhoods in the data (Bach, 2014; Bzdok, Eickenberg, Grisel, Thirion, & Varoquaux, 2015; Hastie et al., 2001). Viewed from the bias-variance tradeoff, this successfully calibrates the sweet spot between underfitting and overfitting. Viewed from statistical learning theory, the VC dimensions can be reduced and thus the generalization performance increased. Applying a model without such complexity restrictions to high-dimensional brain data, generalization becomes difficult to impossible because all directions in the data are treated equally in with isotropic structure. At the root of the problem, data samples seem to be evenly distributed in high-dimensional data scenarios (Bellman, 1961). The learning algorithm will not be able to see through the noise and will thus overfit. In fact, these considerations explain why the multiple comparisons problem is closely linked to encoding studies and overfitting is more closely related to decoding studies (K.J. Friston et al., 2008). Moreover, it offers explanations as to why analyzing neural activity in a region of interest, rather than the whole brain, simultaneously alleviates both the multiple comparisons problem (called "small volume correction" in CS studies) and the overfitting problem (called "feature selection" in SL studies).

Conclusion

Statistical inference is a hetergeneous concept; not only in the imaging neurosciences. Historically, the invention and application of statistical methods has always been driven by practical necessity. The constantly growing neuroimaging repositories might therefore necessitate a categorical change in statistical methodology. Classical statistics has mostly been invented for small-sample experimental studies, but is alone insufficient for the large-scale analysis of internationally acquired, multi-modal neuroimaging repositories with complex meta-information. It is suitable to analyse marginal statistics, i.e. the signal present in each brain image location, irrespective of others (encoding). In contrast, SL consider ensembles of brain image locations, which opens the way to conditional interpretation (information carried by a voxel, given the others) and discriminative reasoning (diagnosis and decoding).

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project). The study was further supported by the Deutsche Forschungsgemeinschaft (DFG, BZ2/2-1 and BZ2/3-1 to D.B.; International Research Training Group IRTG2150), Amazon AWS Research Grant (D.B.), the German National Academic Foundation (D.B.), the START-Program of the Faculty of Medicine, RWTH Aachen (D.B.), and and the MetaMRI associated team (B.T., G.V.).

References

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

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M. E., . . . Evans, A. C. (2013). BigBrain: an ultrahigh-resolution 3D human brain model. *Science, 340*(6139), 1472-1475. doi: 10.1126/science.1235381

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *The journal of wildlife management*, 912-923.

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

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., & Dronkers, N. F. (2003). Voxel-based lesion-symptom mapping. *Nat Neurosci, 6*(5), 448-450. doi: 10.1038/nn1050

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos Trans R Soc Lond B Biol Sci, 360*(1457), 1001-1013. doi: 10.1098/rstb.2005.1634

Bellman, R. E. (1961). *Adaptive control processes: a guided tour* (Vol. 4): Princeton University Press.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*(203), 526-536.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Heidelberg: Springer.

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

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Großhirnrinde*. Leipzig: Barth.

Bzdok, D. (2016). Classical Statistics and Statistical Learning in Imaging Neuroscience. *arXiv preprint arXiv:1603.01857*.

Bzdok, D., Eickenberg, M., Grisel, O., Thirion, B., & Varoquaux, G. (2015). *Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data.* Paper presented at the Advances in Neural Information Processing Systems.

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*, DOI: 10.1371/journal.pcbi.1004994.

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage, 44*, 62-70.

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.

Cohen, J. (1990). Things I have learned (so far). *American psychologist, 45*(12), 1304.

Cowles, M., & Davis, C. (1982). On the Origins of the .05 Level of Statistical Significance. *American Psychologist, 37*(5), 553-558.

Devor, A., Bandettini, P. A., Boas, D. A., Bower, J. M., Buxton, R. B., Cohen, L. B., . . . Franceschini, M. A. (2013). The challenge of connecting the dots in the BRAIN. *Neuron, 80*(2), 270-274.

Domingos, P. (2012). A Few Useful Things to Know about Machine Learning. *Communications of the ACM, 55*(10), 78-87.

Efron, B. (1978). Controversies in the foundations of statistics. *American Mathematical Monthly*, 231-246.

Efron, B., & Hastie, T. (2016). *Computer-Age Statistical Inference*.

Eickhoff, S., Turner, J. A., Nichols, T. E., & Van Horn, J. D. (2016). Sharing the wealth: Neuroimaging data repositories. *NeuroImage, 124*(FZJ-2015-06893), 1065–1068.

Eickhoff, S. B., Bzdok, D., Laird, A. R., Roski, C., Caspers, S., Zilles, K., & Fox, P. T. (2011). Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. *Neuroimage, 57*(3), 938-949. doi: S1053-8119(11)00509-X [pii]

10.1016/j.neuroimage.2011.05.021

Eickhoff, S. B., Thirion, B., Varoquaux, G., & Bzdok, D. (2015). Connectivity-based parcellation: Critique and implications. *Hum Brain Mapp*. doi: 10.1002/hbm.22933

Engemann, D. A., & Gramfort, A. (2015). Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage, 108*, 328-342.

Fisher, R. A., & Mackenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *The Journal of Agricultural Science, 13*(03), 311-320.

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

Friston, K. J. (2006). *Statistical parametric mapping: The analysis of functional brain images.* Amsterdam: Academic Press.

Friston, K. J. (2012). Ten ironic rules for non-statistical reviewers. *Neuroimage, 61*(4), 1300-1310. doi: 10.1016/j.neuroimage.2012.04.018

Friston, K. J., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., & Ashburner, J. (2008). Bayesian decoding of brain images. *Neuroimage, 39*(1), 181-205. doi: 10.1016/j.neuroimage.2007.08.013

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

Gabrieli, J. D., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron, 85*(1), 11-26. doi: 10.1016/j.neuron.2014.10.047

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. NJ: Erlbaum: Hillsdale.

Giraud, C. (2014). *Introduction to high-dimensional statistics*: CRC Press.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine, 130*(12), 995-1004.

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., . . . Margulies, D. S. (2014). NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. in press.

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *Intelligent Systems, IEEE, 24*(2), 8-12.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. Heidelberg, Germany: Springer Series in Statistics.

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

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

House of Common, S. a. T. (2016). *The big data dilemma*. UK: Committee on Applied and Theoretical Statistics.

Jenatton, R., Audibert, J.-Y., & Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research, 12*, 2777-2824.

Jordan, M. I. (2010). Bayesian nonparametric learning: Expressive priors for intelligent systems. *Heuristics, Probability and Causality: A Tribute to Judea Pearl, 11*, 167-185.

Jordan, M. I., 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. (2013). *Frontiers in Massive Data Analysis*. Washington, D.C.: The National Academies Press.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

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

Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R., & Koch, C. (2013). Neuroscience thinks big (and collaboratively). *Nature Reviews Neuroscience, 14*(9), 659-664.

Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. (2012). The big data bootstrap. *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

Kriegeskorte, N. (2011). Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage, 56*(2), 411-421.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proc Natl Acad Sci U S A, 103*(10), 3863-3868. doi: 10.1073/pnas.0600244103

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. doi: 10.1038/nature14539

Lo, A., Chernoff, H., Zheng, T., & Lo, S. H. (2015). Why significant variables aren't automatically good predictors. *Proc Natl Acad Sci U S A, 112*(45), 13892-13897. doi: 10.1073/pnas.1518285112

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*: Cambridge university press.

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*.

Markram, H. (2012). The human brain project. *Scientific American, 306*(6), 50-55.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage, 56*(2), 400-410. doi: 10.1016/j.neuroimage.2010.07.073

Neyman, J., & Pearson, E. S. (1933). On the Problem of the most Efficient Tests for Statistical Hypotheses. *Phil. Trans. R. Soc. A, 231*, 289-337.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage, 62*(2), 811-815.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res, 12*(5), 419-446.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *NeuroImage, 104*, 209-220.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage, 45*, 199-209. doi: 10.1016/j.neuroimage.2008.11.007

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci, 10*(2), 59-63. doi: S1364-6613(05)00336-0 [pii]

10.1016/j.tics.2005.12.004

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat Neurosci, 17*(11), 1510-1517. doi: 10.1038/nn.3818

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological methods, 1*(2), 115.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, 289-310.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., . . . Consortium, W. U.-M. H. (2012). The Human Connectome Project: a data acquisition perspective. *Neuroimage, 62*(4), 2222-2231. doi: 10.1016/j.neuroimage.2012.02.018

Van Horn, J. D., & Toga, A. W. (2014). Human neuroimaging as a "Big Data" science. *Brain Imaging Behav, 8*(2), 323-331. doi: 10.1007/s11682-013-9255-y

Vapnik, V. N. (1989). *Statistical Learning Theory*. New York: Wiley-Interscience.

Vapnik, V. N. (1996). *The nature of statistical learning theory.* New York: Springer.

Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience, 3*(1), 28.

Vogt, C., & Vogt, O. (1919). Allgemeinere Ergebnisse unserer Hirnforschung. *Journal für Psychologie und Neurologie, 25*, 279-461.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine, 368*(15), 1388-1397.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*: Springer Science & Business Media.

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

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics, 13*, 1-14.

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism, 12*, 900-900.

Worsley, K. J., Poline, J.-B., Friston, K. J., & Evans, A. C. (1997). Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage, 6*(4), 305-319.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics, 25*(6), 714-721.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods, 8*(8), 665-670. doi: 10.1038/nmeth.1635

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics, 15*(2), 265-286.