REVIEW ARTICLE
https://doi.org/10.1038/s42256-019-0069-5
nature machine intelligence

Towards algorithmic analytics for large-scale datasets

Danilo Bzdok[1,2,3]*, Thomas E. Nichols[4,5] and Stephen M. Smith[4]

The traditional goal of quantitative analytics is to find simple, transparent models that generate explainable insights. In recent years, large-scale data acquisition enabled, for instance, by brain scanning and genomic profiling with microarray-type techniques, has prompted a wave of statistical inventions and innovative applications. Here we review some of the main trends in learning from 'big data' and provide examples from imaging neuroscience. Some main messages we find are that modern analysis approaches (1) tame complex data with parameter regularization and dimensionality-reduction strategies, (2) are increasingly backed up by empirical model validations rather than justified by mathematical proofs, (3) will compare against and build on open data and consortium repositories, as well as (4) often embrace more elaborate, less interpretable models to maximize prediction accuracy.

Tension is emerging in everyday data analysis in the biomedical sciences. Around the turn of the century, deployment of new measurement techniques, especially microarray-like techniques in genomics and brain scanning in neuroscience, have ignited data accumulation on a massive scale[1]. As one consequence, the amount of health-related data is expected to double several times per year starting from 2020[2]. Data-analytical methodology in turn has expanded more in the past two decades than any other point in history[3,4]. However, emerging opportunities to generate quantitative insight from accumulating data are adopted with hesitation in many empirical domains. Here, we portray the growing stack of algorithmic tools, illustrated with examples from the area of human neuroscience.

In many empirical sciences, classical statistics is still the dominant arsenal for deriving rigorous conclusions from data. A form of linear regression was already used by Gauss around 1795, and null-hypothesis testing emerged in the early twentieth century to formally assess the significance of tail-area statistics[3] (for an explanation of key terms used in this Review, see Tables 1 and 2). Closed-form textbook formulae were a necessity to avoid laborious paper-and-pencil calculations[5]. Electronic computations only became slowly available after the Second World War[3]. Hence, a new modelling approach was routinely validated by virtue of mathematical theory. Specifically, consistency theorems formally characterize how a particular analysis method behaves if the sample size increases indefinitely[6,7]. Much emphasis was put on straightforward linear models due to key advantages for understanding the relationships between carefully selected input variables. Simpler, less data-hungry models were also generally preferred because data acquisition was financially and logistically expensive for most of the past century. This is why experimental laboratory studies needed to be carefully planned in advance and research hypotheses had to be precisely defined beforehand[8,9]. From this traditional perspective of empirical investigation, re-analysing general-purpose data repositories would have been of little appeal. Moreover, it was only in the 1990s that desktop computer software for many machine-learning algorithms or Bayesian modelling approaches became widely available[3].

Impressive data aggregation in the early twenty-first century has given rise to a new kind of empirical research[10,11]. In many scientific domains, information has become cheaper and considerably more fine-grained and multifaceted, as well as freely available. This shift of context opens the door to the principled exploration of already acquired 'found' observational data. Increasingly, quantitative analysis tools are designed, evaluated and deployed ad hoc before complete formal analysis of their mathematical properties. Instead, empirical justifications are obtained from successful prediction performance in separate reference datasets[5,12]. More broadly, modern data analysis needs to negotiate trade-offs between statistical notions, such as effect uncertainty (for example, 'How sure are we about a detected effect?'), and computer-science notions, such as computational load and memory resources (for example, 'How expensive does the analysis become with an increasing number of input variables?')[10,13].

In this Review, we consider global trends in modern data analytics on large-scale datasets, including the presence of more variables, larger sample sizes, open data sources for analysis and assessment, and 'black box' prediction methods.

Global trends in empirical data analysis

As a looming culture clash, university education in various empirical fields is still focused on classical methods from a time of scarce data and limited computation. Blossoming data resources, however, entail a need for exploiting analytical techniques suited for today's data-rich setting. Getting back to our guiding example, imaging neuroscience has spawned increasingly wide and deep datasets over recent years. However, the adoption of analytical tools tailored for modern data is accelerating only recently[14,15]. The most commonly used methods from statistics and computing were not designed to solve the types of problems that data-rich scientists, including neuroscientists, are facing today. The present overview retraces this emerging transition from formally inspired modelling of a few hand-selected variables to learning complicated patterns from data with increasingly adaptive algorithms.

First, many empirical sciences are now generating detailed phenotypical descriptions of organisms and phenomena, such as the

[1]Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Aachen, Germany. [2]JARA, Translational Brain Medicine, Aachen, Germany. [3]Parietal Team, INRIA, Neurospin, CEA Saclay, Gif-sur-Yvette, France. [4]Wellcome Trust Centre for Integrative Neuroimaging (WIN-FMRIB), University of Oxford, Oxford, UK. [5]Big Data Institute, University of Oxford, Oxford, UK. *e-mail: danilo.bzdok@rwth-aachen.de

Table 1 | Glossary (A–G)

Technical term (example(s) from neuroscience)	Core intuition
Bagging[76]	'Wisdom of crowds' strategy to enhance predictive performance by averaging several outcome predictions from models that have been fitted to resampled versions of the same dataset.
Bias versus variance[77]	The bias–variance trade-off calibrates between losing information (bad fit to data at hand) or succumbing to noise (bad extrapolation to new data). A model with high bias tends to ignore relevant patterns in the data—underfitting due to low effective degrees of freedom. A model with high variance tends to extract arbitrary patterns in the data—overfitting due to high effective degrees of freedom.
Bayesian versus frequentist modelling[78–80]	Bayesian modelling assumes the model parameters to be random and the data to be fixed; vice versa for frequentist modelling. Consequently, Bayesian analysis provides certainty distributions for each model parameter value, while frequentist analysis yields a single best-guess value for each model parameter. Bayesian model posterior distributions are conditioned on the data at hand, while frequentist model estimation implicitly averages across other data one could have observed.
Canonical correlation analysis[81]	(Multivariate) pattern-discovery approach that extends the idea of PCA to two variable sets or two datasets. Mutual dependences are extracted as correlated linear combinations of these two data matrices.
Closed-form solutions	A mathematical formula that solves a problem 'in one shot' by a circumscribed set of computing operations (that is, non-iterative), always yielding the same result in the same amount of time.
Consistency theorems or asymptotic guarantees	A widely used class of mathematical proofs that study the properties of a given modelling approach by taking the number of data points to infinity. It is without asymptotic consistency guarantees that finite-sample theorems describe properties of modelling approaches as a function of the number of available data points.
Cross-validation[82,83]	A (non-parametric) sequential resampling procedure used as the gold standard to practically quantify the performance of predictive models to extrapolate discovered patterns to future data. First, model estimation is carried out by fitting the parameter values to the training data (in-sample). Second, if model hyperparameters need to be set, model selection can be carried out on another independent data split—validation data—to automatically tune towards a winning hyperparameter combination. Third, model evaluation then quantifies the pattern generalization based on predictive performance in independent hold-out data (out-of-sample). The overall process is repeated for different splits of the available data (usually five or ten times). Underfitting yields bad in- and bad out-of-sample generalization performance. Overfitting yields excellent in- and bad out-of-sample prediction accuracy.
Curse of dimensionality ('high dimensions')[84]	If data have abundant input variables, relative to the available number of data points, each such input dimension is populated with and represented by less data points. Hence, even classical (unregularized) linear regression can be over-parameterized. In this setting, common models have trouble finding patterns existing in the data and goodness-of-fit metrics (computed in-sample) may become impotent. Variance in model parameter estimation escalates and thus data overfitting becomes a core challenge.
Data augmentation	A heterogeneous group of ad hoc engineering tricks to repeatedly duplicate and modify the original data points while trying to keep their characteristics realistic. The increased effective sample size can allow for estimating more robust model parameters.
Deep neural-network algorithms ('deep learning')[85–87]	A growing class of pattern-learning algorithms that perform prediction based on a nonlinear, hierarchical, multilayer neural-network model. Deep neural-network algorithms are able to fit parameters of a particularly high number of nested nonlinear processing layers and have extreme freedom in fitting patterns in data.
Degrees of freedom	The number of separate pieces of information to be estimated from data. In the setting of classical linear regression, the degrees of freedom typically refer to the number of independent data points n minus the number of fitted model parameters p to estimate residual errors. This conception is starting to struggle or is difficult to compute for many modern adaptive modelling approaches.
Dimensionality reduction[88,89]	Breaking down the number of input variables to a (much) smaller number of quintessential summary variables. Examples include clustering approaches, such as k-means, to partition an array of input variables into typically few non-overlapping variable groups, and matrix decomposition approaches, such as PCA and independent component analysis, to extract new continuous representations spanning across input variables that may have partial overlap with each other.
Exchangeability[50]	A characteristic of the data that is a more general form of the independent-and-identically-distributed (i.i.d.) assumption. For example, null-hypothesis testing may be used to try to reject the hypothesis that males and females have the same average height. Here, exchangeability may be imposed by shuffling which height measurement belongs to male or female participants to assess whether the summary statistics differ given otherwise identical joint distributions among the input variables.
Ground truth[90]	The true pattern in nature to be approximated by using quantitative modelling of empirical measurements.

brain. Investigators confronted with hundreds or thousands of quantitative measurements are also confronted with how to estimate statistical models with potentially hundreds or thousands of parameters. This new context questions the long-standing dogma that statistical analysis should strive to be maximally impartial. Instead, the unfolding analysis paradigm appears to ask, 'What is the most useful a priori knowledge that can inform and shape my quantitative analysis?'. The consequence is growing importance of bias-inducing regularization strategies and data transformations for dimensionality reduction, such as clustering and matrix decomposition.

Second, classical modelling tools have usually been trusted after their formal properties had been mathematically understood in detail. Increasing data availability is escalating the pace at which new quantitative methods are invented and re-purposed, even

Table 2 | Glossary (H–Z)

Technical term (example(s) from neuroscience)	Core intuition
Hierarchical multilevel regression[28,91]	Extension of classical linear regression, where the model parameters are also themselves modelled. Linear interactions are introduced by data-level regression parameters being regularized in groups towards upper-level model parameters to 'borrow statistical strength', such as between study sites. Can be carried out in the frequentist regime, but lends itself particularly well to Bayesian modelling. Linear hierarchical regression can more readily fit models with more parameters p than observations n.
Identifiability[35]	Whether the parameter values of a given model can be unambiguously estimated and thus meaningfully interpreted. The combination of data scenario and model properties may result in identifiability (highly valued in classical statistics) or non-identifiability (often a smaller concern in predictive machine-learning applications). Non-identifiable model parameters can result in very different fitted values despite identical prediction performance of the overall model.
k-means clustering[92]	A popular clustering algorithm that partitions the p input variables into k non-overlapping groups.
Markov chain Monte Carlo sampling[93]	An iterative sampling procedure for numerical approximation of challenging posterior integrals, such as those often arising in Bayesian statistics. Each random 'draw' yields one candidate set of model parameters that are jointly plausible as to how the data could have come about.
Maximum likelihood estimation[36]	Formal (parametric) framework on how to find one good set of model parameter values (assumed to be fixed) that maximize the plausibility of how the data may have come about given a pre-specified model. Ordinary linear regression and other classical approaches are special cases. MLE enjoys strong asymptotic guarantees, but can incur problems as the number of input variables p increase.
Multivariate versus univariate modelling[35,74,94,95]	Technically, univariate methods consider one variable at a time, whereas multivariate methods consider several, possibly many, variables at a time. In neuroscience applications, 'univariate' analysis has often been taken to refer to estimating effects for a single brain location, particular brain connection or specific gene at a time. Instead, 'multivariate' analysis would jointly assess patterns in many such biological measurements.
Parametric versus non-parametric[48]	A parametric approach explicitly assumes structure or a particular form of how the input variables relate to the output. A non-parametric approach tries to fully model the data themselves, for instance, by avoiding assuming Gaussian normality in the data.
Partial least squares[18,19]	(Multivariate) pattern-discovery approach aimed at decomposition of co-variation, similar to CCA. While partial least squares operates on the un-normalized co-variation, CCA acts on the data in a scale/unit-invariant fashion.
Permutation procedures[48,50]	A computation-intensive group of typically non-parametric resampling procedures, which can control error rates (including correction for multiple comparisons), while making few theoretical assumptions. For instance, such procedures enable computation of empirical distributions under some null hypothesis for significance testing in a much wider range of analysis scenarios.
Posterior predictive checks	In Bayesian modelling, generating new data (typically outcome predictions) from model parameter sets sampled from Markov chain Monte Carlo chains to assess discrepancies between an obtained probabilistic model and the actual data at hand.
Regularization/penalization/shrinkage/sparsification[96,97]	Bias is introduced on purpose in model estimation, for instance, to address the curse of dimensionality. As one widespread example, sparse modelling via L1 terms characteristically drives towards variable selection by encouraging exactly zero model parameter values (compare with LASSO regression). As another widely used example, L2 terms intentionally skew model parameter values to be closer to zero (compare with ridge regression).
Tail-area statistics[98]	Instead of some aggregate statistic (for example, mean, median, mode), interest lies in the shape of a data distribution, especially its extremes with low probability. Special interest was placed outside of the 95% interval assessing whether or not an observation exceeds two standard deviations as in null-hypothesis significance testing.
Variability[99]	A property of the data. For example, how does the volume of the amygdala really differ between individuals? The variability of a parameter estimate does not go to zero as the sample size approaches infinity, in contrast to uncertainty.
Uncertainty[99]	A property of the modelling approach. For example, how sure are we about the modelling estimate of amygdala volumes? The uncertainty of an estimated parameter value goes to zero as the number of data points n increases indefinitely, in contrast to variability. Frequentist standard deviations or error bars may mix aspects of variability and uncertainty, in contrast to Bayesian posterior density intervals.

before studying their theoretical properties. As such, the modern quantitative investigator has special interest in asking, 'How well does my obtained modelling solution hold up when directly evaluated in other sampled observations?'. Hence, extracted candidate models are more often grounded in empirical cross-validation or posterior predictive checks to judge the model's quality and practical usefulness.

Third, many empirical sciences have centred on careful planning and conducting of experiments in the laboratory. This predominance of acquiring expensive in-house datasets is re-balanced to ever-wider usage of openly accessible observational datasets. Investigators can increasingly ask 'How does my research question play out in existing consortium data?' or 'How does my newly developed method scale to open population datasets?'. These opportunities can improve the reproducibility of scientific claims and the comparability of quantitative approaches.

Fourth, empirical sciences, such as imaging neuroscience, capitalize on always more complex, sometimes untransparent, modelling approaches. These investigators may want to ask, 'How well can a powerful pattern-learning algorithm forecast outcomes from the natural phenomenon under study?'. On the one hand, by maximizing prediction accuracy on new data or settings, much harder problems can be tackled than before. On the other hand, uncompromising prediction studies may lose some interpretational grip on understanding the isolated role of each model parameter. The resulting interpretability trade-offs incur ethical and policy-related consequences for science, business and government.

Deeper phenotyping always yields more variables

The brain sciences have recently been highlighted as the potentially most data-rich medical specialty[2]. In genomics and imaging genetics, jointly considering more than 1,000,000 single nucleotide polymorphisms typically exceeds the thousands of participants in the currently biggest human cohorts, for example, ref. [16]. In imaging neuroscience, a brain scan with commonly available resolution offers measurements from about 100,000–500,000 locations ('high-*p*' scenario with many variables). Yet, sample sizes have reached hundreds or thousands of participants ('low-*n*' scenario with few observations) only over recent years. In this context, too few sample observations from the participants may be available to allow for rigorous statements about each separate input dimension, such as a specific gene or a particular brain location. At the extreme, detailed neuroanatomical studies with measurements at micrometre resolution may be available in only one or a few participants[17]. Consequently, application of dimensionality-reduction techniques is becoming hard to avoid in various empirical sciences. The theme of reducing abundant multivariate information to the relevant essence is reflected in matrix decomposition techniques such as principal component analysis (PCA), canonical correlation analysis (CCA), partial least squares, independent component analysis and expanding tensor-decomposition techniques, as well as clustering techniques such as *k*-means[18,19]. In many workflows, it becomes an important pre-processing step to re-express the data in a simpler underlying form before applying the final data analysis model, such as linear regression[20].

In traditionally small datasets with few variables, computing ordinary linear regression analysis, as a special instance of maximum likelihood estimation (MLE), readily provides parameter estimates, with almost optimal precision[6]. Standard linear regression corresponds well to the traditional goals of statistics that valued impartiality to what may be expected in the data—unbiasedness. Indeed, even datasets with about 30–40 variables were still considered high-dimensional in the 1980s and 1990s[1]. Already in this setting, the formal theory backing up linear regression models starts to lose some of its optimality, although MLE successfully legitimized a series of classical statistics approaches still in pervasive use today[1]. Introducing an increasing number of input variables into a linear model usually leads to an increasing number of parameters to be fitted. Such larger models incur higher variance in estimating their model parameter values, for example, due to ambiguities in the fitting of partially redundant variables that carry similar information about the target outcome[20].

Even when adding more parameters to simple linear models, the expanded model capacity—with higher degrees of freedom—adds challenges to interpretation. With increasing number of model parameters, it becomes more difficult to clearly attribute the variance explained by each individual input variable[1]. Larger linear models are also more susceptible to picking up on idiosyncrasies and noise in data. In the high-dimensional scenario, hundreds or thousands of input variables (for example, brain region volumes, functional connectivity strengths or gene expression levels) can be submitted to model fitting. It becomes harder to tell, using classical goodness-of-fit tests, how well an obtained linear model actually encapsulates the data at hand. At the extreme, the number of measured variables exceeds the number of available samples or observations in many modern datasets[21]. Such data-rich settings hinder the reproducible identification of unique model parameter solutions. Such contexts render common linear-regression models non-identifiable, which makes the parameter value estimates difficult to interpret and can make model performance on new data poor[22].

The amount of information that can be gleaned from emerging high-dimensional datasets may sometimes remain low even if the sample size is increasing[23,24]. Probably no statistical approach performs well if thousands of input variables are truly individually informative about the outcome to be predicted[22]. Specially, with high dimensions, pre-assuming a more parsimonious underlying representation in the relevant variables (for example, identifying a smaller number of latent factors of variation) may be a pragmatic way to obtain useful and interpretable modelling solutions. Consequently, modern data often make it necessary to introduce some intentional bias into the data-analysis process. In addition to dimensionality-reduction techniques, a plethora of regularization strategies have flourished in response[22]. Often very simple extensions of traditional tools like linear regression can be effective in datasets with high-dimensional measurements. Such penalized linear models dedicated to many variables *p* are epitomized by the increasing adoptions of the ridge regression, least absolute shrinkage and selection operator (LASSO) and elastic net methods[22].

In particular, several classical analysis tools underwent a sparsification over the past 15–20 years: the introduced biasing assumption is that most input variables in the data are expected to be uninformative about the outcome. Sparsified model extensions were enthusiastically embraced by the machine-learning community[22]. Those often-frequentist modelling approaches try to learn from data which input variables can be ignored by encouraging a maximum of model parameters with exactly zero values. This penalized modelling regime assumes that effective model estimation should be skewed towards finding only a subset of the input variables to be relevant to the research question[22]. Besides sparsity-inducing regression via the LASSO and elastic net methods, these highly effective parsimony constraints recently motivated, for instance, sparse PCA, sparse CCA or sparse *k*-means.

Instead, more Bayesian-minded analysts may prefer to estimate the uncertainty of possible model parameter values and corresponding variable influence to be close to zero or not[25–27]. The investigator embracing Bayesian statistics intentionally biases model estimation by skewing parameter values towards existing knowledge expressed in prior distributions[28]. For instance, such approaches can capitalize on hierarchical dependence structure in the data that exists between the quantitative measurements, such as to share statistical strengths between individual outcomes pooled from different time points of a longitudinal study. In high dimensions, many Bayesian approaches may however suffer, as the prior probability distribution can take unexpected shapes, which may preclude the 'true' parameter values from being recovered. As a manifestation of this so-called curse of dimensionality, imposing prior knowledge by guiding model estimation to expect certain ranges of model parameter values more than others may become ineffective[29]. Even if probabilistic parameter distributions could be obtained on each input dimension, it would be challenging for a domain expert to interpret every single parameter[13]. Algorithmically, even modern approximate methods for Markov chain Monte Carlo sampling widely used to infer Bayesian models are susceptible to the consequences of the curse of dimensionality[30].

These side effects of rich multivariable phenotyping need to be tackled in an increasing number of modern neuroscience studies. Linear but flexible pattern-learning models have repeatedly yielded useful dimensionality reductions of high-dimensional subject descriptions and integration of different modalities of detailed measurements. In this spirit, CCA was used by Smith and colleagues to uncover population co-variation that links coupling measures of various brain networks and extensive phenotyping by a diversity of behavioural indicators[31]. Standard CCA can be viewed as

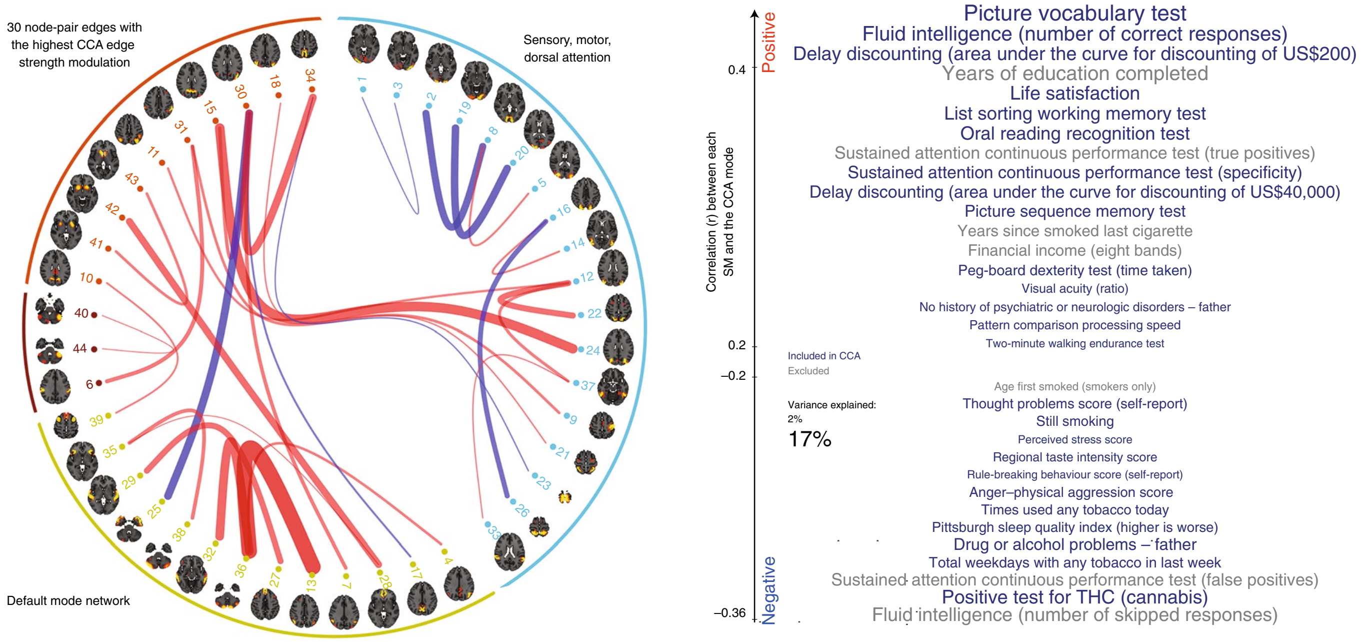

Fig. 1 | Significant population mode that relates patterns of inter-network connectivity to patterns within deep behavioural phenotyping. In the approximately 500 participant release of the Human Connectome Project, CCA lends itself particularly well to uncovering multimodal correspondences between brain and behaviour. Intrinsic network coupling fluctuations between 200 nodes were demonstrated to bear rich relationships with more than 100 cognitive assessments, demographic profiles and life-factor indicators. A functional connectivity fingerprint emerged with rich profiling of behavioural associations that varied along a global positive–negative axis with high intelligence, memory and cognition performances on the one end, and negative lifestyle measures and events on the other end. The brain regions exhibiting the strongest contributions to coherent connectivity changes were reminiscent of the default mode network, which is implicated in episodic memory and semantic capacity, mental scene construction and complex social reasoning, such as taking other people's perspective. Figure reproduced from ref. [31], Springer Nature Ltd.

reminiscent of classical statistics because this model is fitted based on MLE without deliberately imposing prior knowledge or bias that would guide parameter estimation. However, the same CCA method can be viewed to represent a prototypical approach suited for modern datasets because of in-built dimensionality reduction, avoidance of strong (parametric) assumptions about the distributions to be encountered in the data, the native ability to fuse two heterogeneous data modalities and acting in the high-variance regime due to the considerable degrees of freedom. CCA models have recently seen extensions and applications for sparse penalization[32], Bayesian modelling[33] and deep learning[34]. This multivariate method is complementary to so-called mass univariate approaches, which have been pervasively used in imaging neuroscience to study effects separately for each part of the brain[35,36].

In a recent CCA application in imaging neuroscience[31], one significant population mode of multimodal co-variation was obtained based on one robust set of canonical correlations. This multivariate brain–behaviour pattern extracted from rich phenotyping demonstrated a positive–negative axis: intelligence, memory and cognition tests and indices of life satisfaction on the positive end, and negative life-factor measures at the other end (Fig. 1). In addition, the functional connectivity weights emphasized prominent modulation of the brain's 'default mode network' (DMN). The doubly multivariate CCA technique was recently re-purposed to revisit the idea that the DMN subserves some of the most human-defining cognitive processes by pooling neural information across the cortical landscape[37]. Profiting from multimodal imaging data of 10,000 UK Biobank participants (Fig. 2), major nodes of the DMN were shown to explain variance in how canonical brain networks communicate with each other. This population neuroscience study[37] thus provided robust indicators that the biological role of the DMN may emerge from propagating brain-wide information flow to orchestrate the cortical network repertoire, potentially mediated by the right and left temporoparietal junction of the DMN (compare with ref. [38]). Sparsity to intentionally bias CCA estimation was recently used to provide a more complete understanding of the functional connectivity patterns of this major brain network during mind-wandering experience in humans[39]. Thus, imposing exactly zero relevance weights, certain random-thought behaviours among richly phenotyped experiences could be isolated to underlie functional connectivity signatures in the DMN.

Empirical model checks enabled by more samples

Classical statistics was conceived when datasets had modest sample size[3,8]. In the early twentieth century, a primary concern was to gather information from scarce data points to achieve reasonable confidence in the model estimates for meaningful parameter interpretation. In this data-scarce context, a key theoretical property to judge the usefulness of a given modelling approach was asymptotic consistency. This formal guarantee of model performance has certified a host of long-standing statistical approaches. This criterion quantifies whether model estimation converges to the true variable relationships as the number of observation samples increases, mimicking, for instance, availability of brain scans from an infinite number of participants.

However, increasingly flexible algorithmic approaches, with data-hungry deep neural-network algorithms as an extreme case, can simply memorize much of the provided observation samples in certain settings. That is, modern complex models may enjoy consistency guarantees, but be highly prone to seriously overfit the provided data[40]. This adaptiveness of flexible modelling approaches can lead to spuriously high performance when evaluated on the observations used for model fitting (in-sample performance). This scenario may change the role of consistency theorems if the goal is to

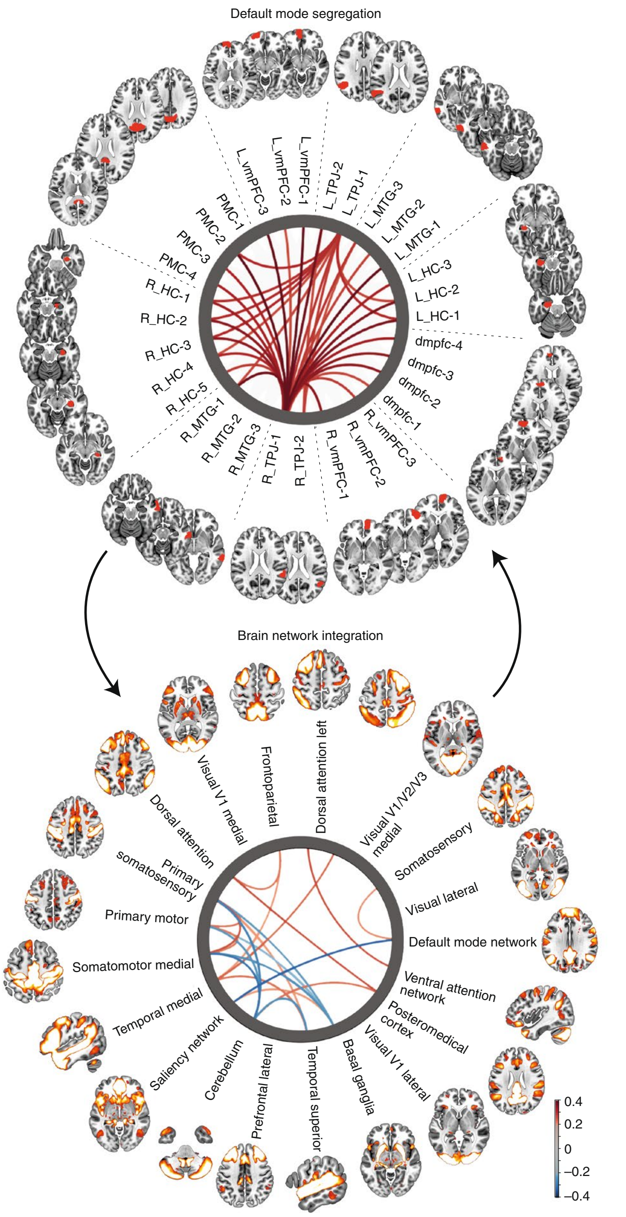

Fig. 2 | Strongest population mode that links intra-network connectivity patterns and inter-network connectivity patterns. In approximately 10,000 UK Biobank participants, CCA was used to identify robust correspondences between functional connectivity shifts inside a major brain network (top), and the default mode network and functional connectivity shifts between a set of major brain networks (bottom). This large-scale analysis made apparent that specific subregions inside the default mode network, namely, the right and left anterior temporoparietal junction, could play a dominant role in the process of global network reconfiguration in humans. v/dmPFC, ventro-/dorso-medial prefrontal cortex; PMC, posteromedial cortex; HC, hippocampus; MTG, middle temporal gyrus; TPJ, temporo-parietal junction; L, left; R, right. Figure reproduced from ref. [37], PNAS.

build models that perform well on observations to be sampled in the future, rather than the data sample at hand. Hence, in mathematical theory describing the convergence behaviour of many machine-learning models, finite-sample theorems are common where model performance is assessed as a function of the amount of observations available for model fitting, with its formal relation to the complexity of the chosen model and the ground-truth information density of the data rather than the raw number of input variables[23,24].

With the increasing sample sizes of modern datasets, empirical evaluation procedures are becoming attractive to vouch for model quality beyond those participants or observations used for model estimation, to new, independent data points. In fact, even a simple, inflexible model with few parameters can often overfit the available observations. This is because not every aspect of the measured data usually reflects the phenomenon of interest[25]. For instance, magnetoencephalographic brain measurements of neural activity responses can be influenced by passing trains hundreds of metres away or by other electromagnetic fluctuations that happen to occur in the environment. Here, the model used may not have optimally fitted to the intended purpose in the participant sample at hand, even if the model enjoys the theoretical consistency guarantee to approach the true statistical relationship with unlimited amounts of data[23]. Moreover, at a given sample size of, say, 1,000 brain scans, it is possible that a purposefully biased model has already converged closer to the ground-truth solution based on the limited number of available observations than an unbiased model that is theoretically ensured to be correct in unlimited observations or participants.

As a consequence of staggering increase in sample size, modern quantitative analyses can increasingly be backed up by data-dependent optimality criteria rather than relying mostly on formal optimality guarantees. In the machine-learning community, resampling and permutation schemes are popular, typically non-parametric tools to glean further information from the data themselves. As an important example, cross-validation procedures[41,42] repeatedly split the available observations to assess the discrepancy between potentially overly optimistic model performance on the training data used for model estimation and the previously unseen test data. This empirical model check can now be increasingly used to approximate the expected model performance in observations or participants yet-to-be-observed in the future[3,20]. An additional validation data split (internal to the training data) routinely serves for tuning any algorithm hyper-parameters to the data at hand, such as to optimize the strength of inducing zero parameters by sparse regularization.

Similarly, empirical permutation procedures allow non-parametric null-hypothesis testing based on exchangeability assumptions. This practical re-implementation of classical statistical inference is more general and flexible than what can typically be achieved by the common assumptions of 'independent and identically distributed'[29]. In addition, bagging can improve prediction performance on new data based on data resampling and averaging hundreds of model solutions[43]. Moreover, bootstrapping can bestow population uncertainty intervals around almost any frequentist statistical approach, derived directly from the available observations themselves[44]. These empirical model checks are based on repeatedly resampling the data at hand, which yields more truthful results with more observations.

In a similar data-guided fashion, Bayesian approaches commonly re-adjust prior assumptions consecutively to enhance model estimation. The practical performance of each candidate model can be evaluated using posterior predictive checks that generate new data from candidate sets of posterior parameter distributions[29]. As Bayesian estimation conditions are based on the provided participants or observations, their fully specified probability intervals for each model parameter are valid for any sample size. These confidence bounds naturally tend to become always narrower as the amount of available observations grows. Moreover, as the influence of the imposed prior knowledge gradually wanes with increasing sample size, the means of the inferred posterior parameter distributions ultimately converge with frequentist estimates of a particular parameter value (from MLE).

In imaging neuroscience, the recent surge in sample size led to re-evaluation of some established means to draw rigorous conclusions from brain measurements. In a sample of approximately 5,000

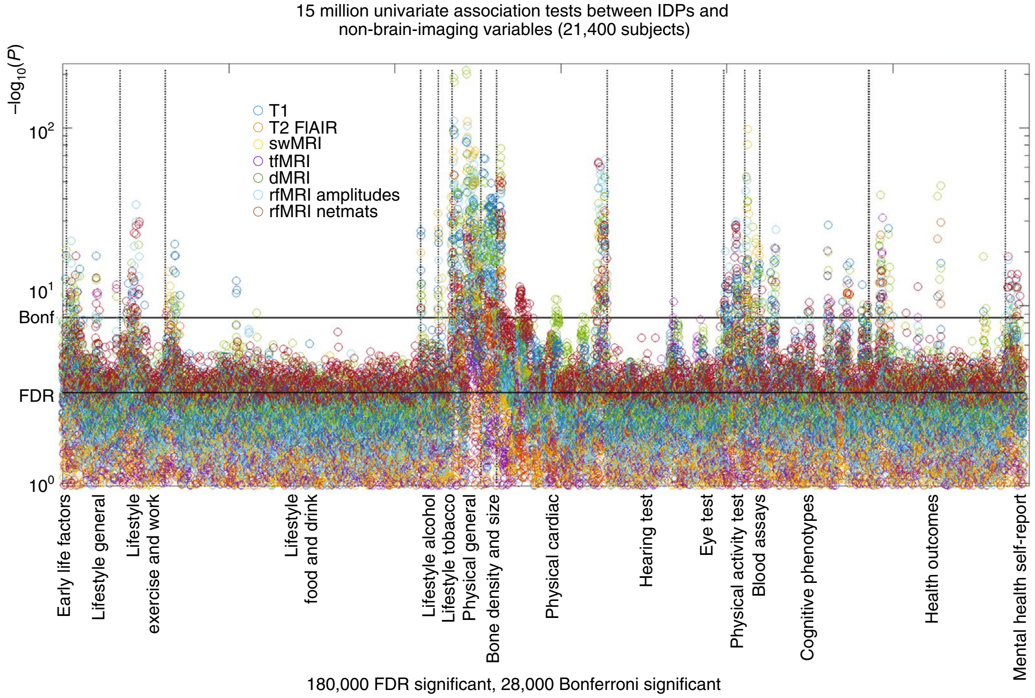

Fig. 3 | Relevance of population associations between six brain-imaging modalities and thousands of behavioural phenotypes. For approximately 21,000 UK Biobank participants, this Manhatten plot depicts results from approximately 15 million cross-subject association tests (each colour indicates a different neuroimaging modality). The horizontal dashed lines indicate significance after correction for multiple statistical comparisons based on Bonferroni's more stringent method (Bonf, top line) or more modern false discovery rate (FDR, bottom line). Even after accounting for family-wise error, approximately 28,000 (Bonferroni) or approximately 180,000 (FDR) brain–behaviour associations remained statistically significant at the population level. These results demonstrate the rich relationships between different brain tissue measurements and extensive phenotyping of many thousands of individuals. Figure computed analogous to previous study on the UK Biobank[44]; see there for details.

UK Biobank participants, Pearson correlation analyses between a behavioural phenotype and a brain imaging feature with a correlation coefficient r of ~0.1 were found to be statistically significant for the most part (Fig. 3). This was even the case after correction for multiple comparisons[45], which was anticipated long ago[46]. In this univariate-flavoured approach, reporting effect estimates as interesting based on P values alone may become insufficient if many observations are included in the analysis. This calls for systematic reporting of effect sizes (that is, model parameter values) and other importance metrics such as prediction performance computed from cross-validation procedures[8,47]. Further, having larger participant samples, combined with more complicated multivariate analysis settings, has propelled the use of non-parametric null-hypothesis testing schemes based on more flexible exchangeability assumptions and data resampling schemes[48,49]. For instance, also in imaging neuroscience, statistical significance is increasingly drawn by generating a to-be-tested empirical null distribution directly from the data themselves. For example, by shuffling which brain scan is labelled as male versus female to make statements about statistically distinguishable sex differences in the brain[50]. The more participants' data are available in a neuroscience dataset, the more reliable conclusions from these resampling procedures can become[51].

When will the sample sizes be sufficient for fitting and evaluating deep-learning approaches in the area of imaging neuroscience? Experts recently proposed a general rule of thumb[52]: in various application areas, $n = 5{,}000$ samples per category to be distinguished were often necessary to achieve relevant model prediction performance. However, datasets with $n > 10{,}000{,}000$ samples were repeatedly necessary to exceed human-level performance. The low sample (n)-to-variables (p) ratio in today's neuroscience datasets may still hamper the potential of deep-learning techniques. Some current shortcomings on data availability in imaging neuroscience may be alleviated by data augmentation strategies, using deep neural-network algorithms whose parameters were already estimated on independent data and other tricks[52].

Open data become a test bed and reference point

The modus operandi in many empirical sciences is still to collect and analyse in-house data for publication in one paper. Various kinds of questions simply cannot be asked quantitatively using one small dataset, such as extracting links between a human's genetic blueprint and her vast diversity of behaviours[53]. Often, genetics amasses data from participant samples collaboratively to chase small effects in multisite consortia as a confederated research endeavour (for example, Psychiatric Genomics Consortium). There are always more incentives and maturing practices to accumulate, curate and distribute data for exploration, knowledge generation and intervention[54]. This trend reverberates in various empirical research communities and is reinforced by data-sharing mandates increasingly specified by funding agencies[55].

Availability of rich open datasets enables using and intersecting data in unexpected ways, fuels continuous development of novel multivariate pattern-learning techniques and renders new research questions actionable. A trusted community dataset can provide a common test bed for those analysis methods as well as benchmarks to compare against a set of state-of-the-art methods in different processing pipelines. As an early tradition in machine learning, the Modified National Institute of Standards and Technology (MNIST) database established itself as a community-wide dataset with 70,000 images of scanned handwritten digits '0' to '9'. New approaches are expected to beat the globally recorded status quo and to compare against human performance in the task of number detection[52].

Kaggle-like competitions are also gaining momentum, where a data-analysis challenge is announced and a larger portion of an existing dataset is provided for model development (www.kaggle.com). At the end of the competition, the modelling solutions from each team are evaluated and ranked on the hidden part of the dataset (see refs. [56,57] for examples from neuroimaging). Such a prime example of healthy competition starts to show high efficacy to crowd-source novel analysis strategies to solve global challenges in biomedicine, as well as in business and government. Open data sharing is also an opportunity to dramatically reduce research costs. More broadly, in the future, new data-analysis methods will perhaps be validated empirically based on statistical performance on shared datasets across diverse existing studies and across various workflows[5,12].

In imaging neuroscience, the majority of the large-scale data initiatives so far have been retrospective collections of independently acquired data from different research centres[58]. Such data repositories can vary considerably in key properties, such as data quality and quality-control procedures. Across-site heterogeneity may explain why, counterintuitively, predictive model performance has been repeatedly reported to decrease as the available data increase[59]. As an ambitious attempt to create a large-scale neuroimaging dataset, the ENIGMA consortium launched in 2009 to centrally orchestrate research projects and recruitment of participating groups by providing analysis pipelines and quality-control protocols. Several thousand participants were characterized with different imaging modalities and genetic profiling. A smaller number of data initiatives realized prospectively planned collections with agreed-on standards for data acquisition. Ensuing repositories offer higher data comparability due to strengthened efforts to, among many others, calibrate acquisition conditions, staff training or travelling experts. The Human Connectome Project was launched in 2009[60] to promote insight into human brain connectivity by providing extensive multimodal measurements of approximately 1,200 healthy adults (aged 22–35), including approximately 300 twin pairs. For each participant, the project gathered structural, functional and diffusion magnetic resonance imaging, genotyping data, as well as a variety of more than 400 demographic, behavioural and lifestyle indicators. With genetic profiling and an extensive variety of phenotyping descriptors, UK Biobank Imaging is even more comprehensive. This data collection initiative set out in 2006 to gather genetic and environmental (for example, nutrition, lifestyle, medications) data of approximately 500,000 volunteers (aged 40–69) and is currently the world's largest biomedical dataset. Its brain and body imaging extension was launched in 2014 (with the brain imaging gathering structural, functional, diffusion and susceptibility-weighted magnetic resonance imaging for approximately 100,000 participants by 2022)[45].

Compared with more established application domains of machine learning, large datasets of human populations pose additional challenges. Extensive phenotypical profiling calls for a careful balance of trust between protecting each participant's privacy and providing rich open biomedical datasets to the larger research community. UK Biobank participants may be given the possibility to opt out of sharing records and retroactively deny consent at any time. Industry can boost health-related big-data analytics by offering computing infrastructure as well as data gathering. However, possible conflicts of interest need to be taken into consideration. As public and media perception is very important, transparent presentation, that is both enthusiastic about the benefits of a study and completely honest, is crucial to avoid inaccurate negative messages being promulgated.

Powerful 'black box' predictions supplement simple models

As a core value of classical data analysis, insight is maximized by assuming linear additivity in how the input variables relate to each other and to the output prediction[6,12]. A traditional goal of statistics is to cleanly isolate the (univariate) effects of 'special' variables on an outcome, such as a risk factor or a treatment response. All components of the model were supposed to be readily understandable by the investigator. The input variables were typically meticulously hand-picked and chosen to have meaningful units based on existing domain knowledge. This analysis paradigm of generating subject-matter understanding from 'introspecting' isolated variable relationships has contributed tremendously to scientific progress in the twentieth century[3]. However, this explainable modelling regime may also have exhausted the repertoire of natural phenomena that can be usefully described and understood by straightforward linear modelling (but see ref. [61]).

In many empirical sciences, including imaging neuroscience, investigators started moving towards more complex modelling approaches and analysis pipelines. Expanding data resources is a prerequisite to estimating flexible, highly adaptive models that have a larger capacity to represent convoluted relationships between variables, such as hierarchical dependencies and higher-order nonlinearity[62]. As more data become available, empirical scientists can now bring to bear more flexible models. In certain cases, the price one may have to pay is that some aspects of the estimated model remain partly opaque to human intuition, pushing investigators to give up on uncompromised model transparency. As an early hint, Bayesian hierarchical modelling can gain traction on complicated datasets that handle nested data settings (for example, brain scans from participants in different cities) with many more model parameters than input variables. These extensions of classical linear models allow for integrating disparate information sources, sharing statistical strengths between variability sources, de-escalating concerns of class imbalance and selection bias, and estimating full uncertainty distributions[29]. As a side effect of increased model complexity, however, not every single parameter value of such a Bayesian hierarchical model may merit equal attention for scientific interpretation.

As a continuation of this theme, in adaptive machine-learning algorithms, and especially in deep neural-network algorithms, much emphasis is put on the output of a model. This change of focus is why identifiability may receive lesser attention in certain studies, although model interpretability was key in classical statistics. Take for example a neurosurgeon who wants to remove brain tissue without impairing language. By relying on a linear model, she predicts outcome in a language task from neural activity measured across the cortex[35]. If quite different model parameter solutions yield an identical prediction accuracy, the model is not identifiable. This fitted linear model cannot be physiologically interpreted as a brain map indicating where tissue resection is safer to preserve language capacity. Different candidate models would have other parameters with small absolute values that can suggest diverging brain locations to be less implicated in language processes, which hampers the classical goal towards mechanistic explanation.

Instead, the predictive analyst would typically neglect such arbitrary values of estimated model parameters and prioritize successful prediction of language performance. This neuroscience example illustrates that parameter interpretation is often more challenging in multivariate models optimized for prediction performance[3]. The difficulty in explaining the role of individual input variables is even bigger in current deep neural-network architectures[63]. Here, the output predictions can result from highly nonlinear processing cascades from the input variables. There may be little hope to exhaustively understand every single one of the thousands or millions of model parameters in some of today's machine-learning models[64,65]. Although some remedies have been recently proposed[66,67], these largely provide understandable simplifications of or linear approximations to the actual nonlinear prediction function.

Consequently, for increasingly popular, powerful prediction models, classical (parametric) inference may become more challenging to obtain statistically significant *P* values. In complicated nonlinear models, it is partly infeasible to assess the 'trueness' of an effect of individual input variables, as an exclusive path for scientific knowledge creation[9,68,69]. These developments do not belittle the importance of working theories in guiding the cumulative construction of scientific understanding. However, there are certain hard problems in empirical research where estimating complex 'black box' models may be one of the very few viable solutions. This is probably the case in weather forecasting, perhaps also in some areas of neuroscience. Assessing which aspects of a phenomenon have been successfully captured or inadvertently ignored in an estimated model will probably more often rely on predictive simulation of new data or querying the obtained model for predictions on unseen participants or observations in the twenty-first century. As a side effect of optimizing uncompromised prediction performance, some neuroscience applications may move away from the goal of causal discovery or even move away from cumulative creation of scientific knowledge[70]. Aiming for crude prediction performance estranges the investigator from asking 'why?'—the reason behind statistical relationships between certain input variables. Today, there is still no commonly agreed-on framework for causal inference.

Instead of carving out new biological mechanisms in nature, prediction metrics can capture how well an estimated complex 'black box' model can 'imitate' or 'reproduce' the studied phenomenon. The Bayesian-frequentist debate, which ignited much controversy in twentieth-century statistics (for example, ref. [71]), may give way to a new antagonistic discourse. One candidate dilemma is classical statistical inference in interpretable models versus prediction accuracy of complicated natural phenomena. Particularly flexible analysis approaches can be applied to quantitatively describe particularly complex systems, such as the human brain in health and disease. In many settings, empirical scientists may have to prioritize 'providing insight' (that is, classical statistical inference targeted at single input variables) against 'accurately modelling the world' (that is, model prediction outputs)[72]. The implied domain interpretability trade-offs will have important consequences for ethical considerations and policy-making[73].

In imaging neuroscience, the prediction–inference antagonism has surfaced as whether or not the prediction accuracies of increasingly used multivariate pattern-learning approaches and machine-learning algorithms should undergo post-hoc statistical significance testing (compare with refs. [47,59,74]). This discussion appears to highlight a culture clash between different data-analysis communities seldom in contact before[12]. Neuroimaging and other empirical academic fields have been dominated by a decade-long legacy of initial linear-regression-type estimation and subsequent statistical null-hypothesis testing. Since its inception, machine learning, however, has put a premium on prediction performance as 'hard currency'[12,20]. This is especially the case given the immediate practical relevance for various data-intensive industries, such as recommendation systems or micro-targeted customer advertisement[75]. In general, input variables that do enhance prediction accuracy do not always declare to be statistically significant[14,25,58]. Conversely, variables that are assessed to be statistically significant can be useless for the goal of prediction in new data in certain cases (compare with ref. [69]). To recapitulate these diverging modelling notions of importance, validating a built machine-learning model based on the metric of successful out-of-sample prediction is based on extracting patterns in the training data and evaluating how well these identified relationships extrapolate to independent observations drawn from the same distribution. In contrast, classical null-hypothesis significance testing pursues a different analytical goal in asking the question whether an obtained prediction accuracy exceeds two standard deviations of happening by chance under some null hypothesis.

Conclusion

Historically, innovation and changing practices in quantitative analytics have been shaped by trends in application domains. The recent advent of massive data in neuroscience and biomedicine is ushering towards larger revisions in everyday analytical practices. (1) Unconstrained linear regression models have been a workhorse in twentieth-century empirical research. However, in the twenty-first century, analysis that biases model estimation by parameter regularization or involves dimensionality-reducing transformations may become ubiquitous as extensive datasets become more available. (2) Many classical models have routinely been backed up by consistency theorems, emulating infinite sample size to characterize the model's quality for converging to a good parameter solution. Assessing model quality based on the variance explained in the fitted data will probably be increasingly supplemented by empirical validation procedures, such as prediction accuracy in untouched data. (3) Many empirical sciences such as neuroscience may transition from the predominance of a priori planned, gathered and published experimental data to conducting more re-analyses of freely available data resources with deep and wide phenotyping. Traditionally, scientific value was seen in unique private datasets. Now, creative modelling strategies may become key to making the most of mushrooming open datasets. (4) Powerful predictive models may not easily lend themselves to exhaustive understanding of all extracted variable–variable relationships. However, the democratization and feasibility of advanced pattern-learning algorithms may enable quantifying always more sophisticated phenomena in nature. Hence, an important dilemma arises between classical relevance claims about single model parameters and successful 'black box' predictions of complicated natural phenomena. These megatrends in data analytics will hopefully propel the scientific description of particularly complex systems, epitomized by the human brain in health and disease.

Received: 19 April 2018; Accepted: 5 June 2019;
Published online: 9 July 2019

References

1. Efron, B. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* Vol. 1 (Cambridge Univ. Press, 2012).
2. *Nature* **539**, 467–468 (2016).
3. Efron, B. & Hastie, T. *Computer-Age Statistical Inference* (Cambridge Univ. Press, 2016).
4. Jordan, M. I. On statistics, computation and scalability. *Bernoulli* **19**, 1378–1390 (2013).
5. Donoho, D. 50 years of data science. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).
6. Casella, G. & Berger, R. L. *Statistical Inference* Vol. 2 (Duxbury, 2002).
7. Efron, B. & Tibshirani, R. J. Statistical data analysis in the computer age. *Science* **253**, 390–395 (1991).
8. Nuzzo, R. Scientific method: statistical errors. *Nature* **506**, 150–152 (2014).
9. Wasserstein, R. L. & Lazar, N. A. The ASA's statement on *P*-values: context, process, and purpose. *Am. Stat.* **70**, 129–133 (2016).
10. Blei, D. M. & Smyth, P. Science and data science. *Proc. Natl Acad. Sci. USA* **114**, 8689–8692 (2017).
11. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
12. Breiman, L. Statistical modeling: the two cultures. *Stat. Sci.* **16**, 199–231 (2001).
13. Jordan, M. I. et al. *Frontiers in Massive Data Analysis* (The National Academies Press, 2013).
14. Bzdok, D. & Yeo, B. T. T. Inference in the age of big data: future perspectives on neuroscience. *NeuroImage* **155**, 549–564 (2017).
15. Smith, S. M. & Nichols, T. E. Statistical challenges in "big data" human neuroimaging. *Neuron* **97**, 263–268 (2018).
16. Elliott, L. T. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).
17. Amunts, K. et al. BigBrain: an ultrahigh-resolution 3D human brain model. *Science* **340**, 1472–1475 (2013).
18. McIntosh, A. R. & Mišić, B. Multivariate statistical analyses for neuroimaging data. *Annu. Rev. Psychol.* **64**, 499–525 (2013).
19. McIntosh, A., Bookstein, F., Haxby, J. V. & Grady, C. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage* **3**, 143–157 (1996).
20. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2001).

21. Giraud, C. *Introduction to High-dimensional Statistics* (CRC Press, 2014).
22. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations* (CRC Press, 2015).
23. Mohri, M., Talwalkar, A. & Rostamizadeh, A. *Foundations of Machine Learning* (Adaptive Computation and Machine Learning Series, MIT Press, 2012).
24. Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms* (Cambridge Univ. Press, 2014).
25. McElreath, R. *Statistical Rethinking* (Chapman & Hall/CRC, 2015).
26. Kruschke, J. K. *Doing Bayesian Data Analysis* (Elsevier, 2011).
27. Wipf, D. P. & Nagarajan, S. S. Sparse estimation using general likelihoods and non-factorial priors. In *Advances in Neural Information Processing Systems* 1625–1632 (NIPS, 2008).
28. Chen, G. et al. Handling multiplicity in neuroimaging through Bayesian lenses with multilevel modeling. *Neuroinformatics* https://doi.org/10.1007/s12021-018-9409-6 (2018).
29. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* Vol. 2 (Chapman & Hall/CRC, 2014).
30. MacKay, D. J. C. *Information Theory, Inference and Learning Algorithms* (Cambridge Univ. Press, 2003).
31. Smith, S. M. et al. A positive–negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.* **18**, 1565–1567 (2015).
32. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).
33. Virtanen, S., Klami, A. & Kaski, S. Bayesian CCA via group sparsity. In *Proc. 28th International Conference on International Conference on Machine Learning* (eds Getoor, L. & Scheffer, T.) 457–464 (Omnipress, 2011).
34. Andrew, G., Arora, R., Bilmes, J. & Livescu, K. Deep canonical correlation analysis. In *International Conference on Machine Learning* 1247–1255 (PMLR, 2013).
35. Haufe, S. et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* **87**, 96–110 (2014).
36. Friston, K. J. et al. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1994).
37. Kernbach, J. M. et al. Subspecialization within default mode nodes characterized in 10,000 UK Biobank participants. *Proc. Natl Acad. Sci. USA* **115**, 12295–12300 (2018).
38. Bzdok, D. et al. Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. *NeuroImage* **81**, 381–392 (2013).
39. Wang, H.-T. et al. Dimensions of experience: exploring the heterogeneity of the wandering mind. *Psychol. Sci.* **29**, 56–71 (2018).
40. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. Preprint at *arXiv* https://arxiv.org/abs/1611.03530 (2016).
41. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Series B* **36**, 111–147 (1974).
42. Geisser, S. The predictive sample reuse method with applications. *J. Am. Stat. Assoc.* **70**, 320–328 (1975).
43. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
44. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (CRC Press, 1994).
45. Miller, K. L. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523 (2016).
46. Berkson, J. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Stat. Assoc.* **33**, 526–536 (1938).
47. Bzdok, D. Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.* **11**, 543 (2017).
48. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).
49. Winkler, A. M. et al. Non-parametric combination and related permutation tests for neuroimaging. *Hum. Brain Mapp.* **37**, 1486–1511 (2016).
50. Ge, T., Yeo, B. T. T. & Winkler, A. A brief overview of permutation testing with examples. *Organization for Human Brain Mapping* https://www.ohbmbrainmappingblog.com/blog/a-brief-overview-of-permutation-testing-with-examples (2018).
51. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2017).
52. Goodfellow, I. J., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
53. Medland, S. E., Jahanshad, N., Neale, B. M. & Thompson, P. M. Whole-genome analyses of whole-brain data: working within an expanded search space. *Nat. Neurosci.* **17**, 791–800 (2014).
54. Leonelli, S. *Data-centric Biology: A Philosophical Study* (Univ. Chicago Press, 2016).
55. Poldrack, R. A. & Gorgolewski, K. J. Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* **17**, 1510–1517 (2014).
56. Bron, E. E. et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. *NeuroImage* **111**, 562–579 (2015).
57. Sarica, A., Cerasa, A., Quattrone, A. & Calhoun, V. Editorial on special issue: machine learning on MCI. *J. Neurosci. methods* **302**, 1 (2018).
58. Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage* **145**, 137–165 (2017).
59. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
60. Van Essen, D. C. et al. The Human Connectome Project: a data acquisition perspective. *NeuroImage* **62**, 2222–2231 (2012).
61. Petkova, E. et al. Statistical analysis plan for stage 1 EMBARC (Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care) study. *Contemp. Clin. Trials Commun.* **6**, 22–30 (2017).
62. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459 (2015).
63. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
64. Shmueli, G. To explain or to predict? *Stat. Sci.* **25**, 289–310 (2010).
65. Harrell, F. Is medicine mesmerized by machine learning? *Statistical Thinking* http://www.fharrell.com/post/medml/ (2019).
66. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 4765–4774 (NIPS, 2017).
67. Chen, J., Song, L., Wainwright, M. J. & Jordan, M. I. Learning to explain: an information-theoretic perspective on model interpretation. Preprint at https://arxiv.org/abs/1802.07814 (2018).
68. Szucs, D. & Ioannidis, J. When null hypothesis significance testing is unsuitable for research: a reassessment. *Front. Hum. Neurosci.* **11**, 390 (2017).
69. Bzdok, D. & Ioannidis, J. P. A. Exploration, inference and prediction in neuroscience and biomedicine. *Trends Neurosci.* **42**, 251–262 (2019).
70. Pearl, J. & Mackenzie, D. *The Book of Why: The New Science of Cause and Effect* (Basic Books, 2018).
71. Efron, B. Why isn't everyone a Bayesian? *Am. Stat.* **40**, 1–5 (1986).
72. Norvig, P. On chomsky and the two cultures of statistical learning. *Peter Norvig* http://norvig.com/chomsky.html (2011).
73. O'Neil, C. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).
74. Haynes, J.-D. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron* **87**, 257–270 (2015).
75. Henke, N. et al. *The Age of Analytics: Competing in a Data-driven World* Technical Report (McKinsey Global Institute, 2016).
76. Hoyos-Idrobo, A., Varoquaux, G., Schwartz, Y. & Thirion, B. FReM—scalable and stable decoding with fast regularized ensemble of models. *NeuroImage* **180**, 160–172 (2018).
77. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2016).
78. Friston, K. J. et al. Classical and Bayesian inference in neuroimaging: applications. *NeuroImage* **16**, 484–512 (2002).
79. Friston, K. J. et al. Classical and Bayesian inference in neuroimaging: theory. *NeuroImage* **16**, 465–483 (2002).
80. Körding, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
81. Friston, K. J., Liddle, P. F., Frith, C. D., Hirsch, S. R. & Frackowiak, R. S. J. The left medial temporal region and schizophrenia. *Brain* **115**, 367–382 (1992).
82. Varoquaux, G. et al. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* **145**, 166–179 (2017).
83. Pereira, F., Mitchell, T. & Botvinick, M. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* **45**, 199–209 (2009).
84. Allen, E. A., Erhardt, E. B. & Calhoun, V. D. Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron* **74**, 603–608 (2012).
85. Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 94 (2016).
86. Plis, S. M. et al. Deep learning for neuroimaging: a validation study. *Front. Neurosci.* **8**, 299 (2014).
87. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
88. Doria, V. et al. Emergence of resting state networks in the preterm human brain. *Proc. Natl Acad. Sci. USA* **107**, 20015–20020 (2010).
89. Sui, J. et al. A CCA+ ICA based model for multi-task brain imaging data fusion and its application to schizophrenia. *NeuroImage* **51**, 123–134 (2010).
90. Jonas, E. & Kording, K. P. Could a neuroscientist understand a microprocessor? *PLoS Comput. Biol.* **13**, e1005268 (2017).
91. Dai, T. & Guo, Y., Alzheimer's Disease Neuroimaging Initiative. Predicting individual brain functional connectivity using a Bayesian hierarchical model. *NeuroImage* **147**, 772–787 (2017).
92. Eickhoff, S. B., Thirion, B., Varoquaux, G. & Bzdok, D. Connectivity-based parcellation: critique and implications. *Hum. Brain Mapp.* **36**, 4771–4792 (2015).

93. Woolrich, M. W. Bayesian inference in FMRI. *NeuroImage* **62**, 801–810 (2012).
94. Haxby, J. V. et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
95. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103**, 3863–3868 (2006).
96. Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W. & Strother, S. C. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognit.* **45**, 2085–2100 (2012).
97. Baldassarre, L., Pontil, M. & Mourão-Miranda, J. Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding. *Front. Neurosci.* **11**, 62 (2017).
98. Woo, C. W., Krishnan, A. & Wager, T. D. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage* **91**, 412–419 (2014).
99. Faisal, A. A., Selen, L. P. & Wolpert, D. M. Noise in the nervous system. *Nat. Rev. Neurosci.* **9**, 292–303 (2008).

Competing interests

The authors declare no competing interests.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.