TechSights

A Shared Vision for Machine Learning in Neuroscience

Mai-Anh T. Vu,[1] **Tülay Adalı,**[8] **Demba Ba,**[9] **György Buzsáki,**[10] **David Carlson,**[3,4] **Katherine Heller,**[5] **Conor Liston,**[11] **Cynthia Rudin,**[6,7] **Vikaas S. Sohal,**[12] **Alik S. Widge,**[13] **Helen S. Mayberg,**[14] **Guillermo Sapiro,**[6] **and Kafui Dzirasa**[1,2]

[1]Department of Neurobiology, [2]Department of Psychiatry and Behavioral Sciences, [3]Department of Civil and Environmental Engineering, [4]Department of Biostatistics and Bioinformatics, [5]Department of Statistical Sciences, [6]Department of Electrical and Computer Engineering, [7]Department of Computer Science, Duke University, Durham, North Carolina 27710, [8]Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, Maryland 21250, [9]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, [10]Department of Neuroscience, New York University School of Medicine, New York, New York 10016, [11]Feil Family Brain and Mind Research Institute, Weill Cornell Medical College, New York, New York, 10065, [12]Department of Psychiatry and Weill Institute for Neuroscience, University of California–San Francisco, San Francisco, California 94158, [13]Department of Psychiatry, Massachusetts General Hospital, Charlestown, Massachusetts 02129, and [14]Department of Psychiatry, Neurology, and Radiology, Emory University, Atlanta, Georgia 30322

With ever-increasing advancements in technology, neuroscientists are able to collect data in greater volumes and with finer resolution. The bottleneck in understanding how the brain works is consequently shifting away from the amount and type of data we can collect and toward what we actually do with the data. There has been a growing interest in leveraging this vast volume of data across levels of analysis, measurement techniques, and experimental paradigms to gain more insight into brain function. Such efforts are visible at an international scale, with the emergence of big data neuroscience initiatives, such as the BRAIN initiative (Bargmann et al., 2014), the Human Brain Project, the Human Connectome Project, and the National Institute of Mental Health's Research Domain Criteria initiative. With these large-scale projects, much thought has been given to data-sharing across groups (Poldrack and Gorgolewski, 2014; Sejnowski et al., 2014); however, even with such data-sharing initiatives, funding mechanisms, and infrastructure, there still exists the challenge of how to cohesively integrate all the data. At multiple stages and levels of neuroscience investigation, machine learning holds great promise as an addition to the arsenal of analysis tools for discovering how the brain works.

Key words: machine learning; reinforcement learning; explainable artificial intelligence

Introduction

What is machine learning?

The term machine learning was coined by Arthur Samuel in 1959 to describe the subfield of computer science that involves the "programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning" (Samuel, 1959). In other words, the field investigates how computers can improve predictions, actions, decisions, or perceptions based on data and ongoing experience. The field of machine learning was driven by the development of algorithms for pattern recognition (e.g., an algorithm for filtering out unwanted marketing emails) and, in general, investigates the development of algorithms that can learn from and make predictions on data. These algorithms largely fall into a few dominant categories: supervised machine learning, unsupervised machine learning, and reinforcement learning.

In supervised machine learning, input data, or training data, have known labels, commonly supplied by human experts. The goal is to learn the relationship between the data and the labels such that the computer can predict the label of a previously unseen data item with accuracy comparable with the human expert. For instance, a training dataset could consist of a set of e-mails that are already classified as spam or not spam, and the goal of the computer algorithm is to settle on a model that can accurately label new incoming e-mail as "spam" or "not spam." As another example, in an fMRI study, Schuck et al. (2015) use a supervised machine learning classifier to classify the color of the stimuli seen by the subjects based on local fMRI brain activity. Examining the classifier accuracy over time and in different brain regions allowed them to infer where and when color was represented in the brain. Regression models, which learn relationships among variables, would fall into the category of supervised machine learning.

Reinforcement learning is a branch of supervised machine learning that has inspired and has been inspired by behaviorist psychology. The "classes" to be learned are actions that could be taken in response to a data item. Machines are trained to make

Received Sept. 25, 2017; revised Jan. 2, 2018; accepted Jan. 9, 2018.

This work was supported by National Institute of Mental Health Grant R01MH099192-05S2 to K.D., National Science Foundation GRFP to M.-A.T.V., and Duke Katherine Goodman Stern Fellowship to M.-A.T.V. Many of the themes presented in this manuscript reflect areas of consensus that emerged from the National Institutes of Mental Health-sponsored Explainable Artificial Intelligence Meeting on November 10, 2017 in Washington, DC (co-organized by Michele Ferrante, Guillermo Sapiro, and Helen S. Mayberg). We thank Shabnam Hakimi, Stephen D. Mague, R. Alison Adcock, Neil M. Gallagher, Austin Talbot, Sasha Burwell, and Rainbo Hultman for comments on this manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to either Dr. Kafui Dzirasa or Mai-Anh T. Vu, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, 421 Bryan Research Building, Box 3209, Durham, NC 27710, E-mail: kafui.dzirasa@duke.edu or mai.anh.vu@duke.edu.

DOI:10.1523/JNEUROSCI.0508-17.2018

decisions through a dynamic trial-and-error process to maximize a desired outcome. The human expert no longer labels each item with the desired class (action) but instead creates a "scoring function" that tells the algorithm how good its move was. For example, a machine might have the goal of winning checkers games, and learn to select moves based on past interactions to maximize the chance of winning the checkers match (Sutton and Barto, 1998). In a typical situation, a scoring function only provides a reward or information based upon the outcome of a complete task after several actions (i.e., in the checkers match, only after a win/loss is achieved; in a Brain Computer Interface task, only when the objective is successfully obtained).

In unsupervised machine learning, on the other hand, the training data have no labels. The goal is to discover hidden structure in the data, perhaps by taking advantage of similarity or redundancy. A well-known example is principal component analysis, a statistical dimension reduction technique that exploits redundancy in the data using only second-order statistics. For unsupervised machine learning, input data might be composed of the symptom profiles of patients thought to have the same general neuropsychiatric illness but also to comprise meaningful heterogeneity. In this case, the goal of the model would be to group similar patients together, thus uncovering important structure within this diagnosis. This would be an example of a family of algorithms aimed at clustering. Critically, a major challenge with unsupervised machine learning algorithms for clustering or dimensionality reduction is understanding the features that make up the groups or reduced dimensions and transforming them into testable scientific hypotheses. For example, Drysdale et al. (2017) used machine learning to discover subtypes of depression based on fMRI functional connectivity, and then subsequently validated their findings via testing the follow-up hypothesis that a treatment modulating cortical connectivity would yield different outcomes among these subgroups.

These categories of machine learning differ in their inputs, outputs, and objectives, and thus encompass a powerful set of tools (e.g., classification, regression, clustering) that enable us to refine the ways we make predictions, make decisions, or discover structure from large sets of data. While machine learning has long been applied to the field of computational and theoretical neuroscience, its burgeoning role in broader cellular, systems, and cognitive neuroscience has been more recent, especially as statistical machine learning packages are being made available in standard analysis software. Accordingly, questions arise as to its place in empirical research ([No authors listed], 2014). Data-driven machine learning approaches are often directly contrasted to the more traditional hypothesis-driven approach, in which an experiment is undertaken to mathematically evaluate the plausibility of a concrete, falsifiable proposed explanation or model, given the observed data. These two approaches are often pitted against each other ([No authors listed], 2014). The question should not be whether one approach is better than the other, but rather how and when we can take advantage of these two complementary strategies.

Machine learning within the hypothesis-driven framework

Despite the distinction between data-driven and hypothesis-driven approaches, there are already many applications of machine learning within the hypothesis-driven framework. Many of these serve as ways to save time and effort, mitigate human biases, or make large datasets computationally tractable. For example, in MRI, image-processing algorithms allow for automated alignment of MRI scans from individual people to atlases (Jenkinson and Smith, 2001; Jenkinson et al., 2002; Andersson et al., 2007). This registration of individual scans to a unified atlas makes group-level spatial analyses and inferences possible. Other algorithms accomplish automated image segmentation, and allow for applications, such as parcellation of structural MRI scans into labeled regions (Fischl et al., 2004), identification of white matter tracts from diffusion MRI (O'Donnell et al., 2017), or identification of neuron structural boundaries in microscopic (EM) images (Jain et al., 2010). Image-processing algorithms can further be applied to video recordings to automate the measurement, identification, and categorization of animal behaviors (Anderson and Perona, 2014; Hong et al., 2015). Such tasks are otherwise accomplished manually, and the development of these technologies immensely reduces the required human time and effort, in turn enabling higher-throughput analysis pipelines. Beyond the added efficiency, with automated processes there is far less room for human subjectivity, biases, or error in coding of images or behaviors. Applied correctly, this can make results more objective, consistent, and reproducible.

In addition to these algorithms that aid in data processing, another example of an application of machine learning techniques within the hypothesis-driven framework is as a strategy for hypothesis testing. For example, in a study on hippocampal-prefrontal input and spatial working memory encoding, Spellman et al. (2015) optogenetically inhibited the ventral hippocampus (vHPC) projection to the medial prefrontal cortex (mPFC) during a spatial working memory task. From behavioral results, they drew the conclusion that input from vHPC to mPFC is critical for spatial cue encoding. To further test this hypothesis, they trained a classifier on the mPFC population firing rate to decode the spatial location of the animal's goal, and separately to decode the task phase. This classifier approach allowed them to quantify the reliability and strength of these mPFC neural representations. They were then able to statistically show that inhibition of the vHPC-to-mPFC signaling resulted in decreased classifier accuracy for spatial goal location but not for task phase. In other words, the ability to classify an outcome from the mPFC activity became a measure of how well that outcome was encoded in mPFC. This machine learning approach thus allowed them to test their hypothesis that these projections were specifically supporting working memory encoding of space, and not other task-relevant features. As another example, Paul et al. (2017) used supervised clustering to analyze single-cell transcriptomes of a set of previously anatomically and physiologically characterized cortical GABAergic neurons, and discovered that these categories indeed differ by a transcriptional architecture that encodes their synaptic communication patterns, confirming the subcategorization of these neurons. Thus, there are already many applications of machine learning that serve to bolster hypothesis-driven research, by automating aspects of data processing, or by yielding additional strategies for hypothesis testing.

Machine learning beyond the hypothesis-driven framework

Machine learning has applications well beyond the hypothesis-driven framework. The more exploratory data-driven approach allows us to explore data in a way that is less limited by our hypothesis space. After all, experiments are only as useful as the hypotheses that they are designed to test, and full hypothesis testing on a drastically expanding dimensionality is intractable. To this end, machine learning methods allow us to extract from our data the dimensions that explain the most variance or even to learn a data-driven taxonomy. For example, data-driven video analysis of behavior may not only serve as an automated replacement for human behavioral coding, but if an unsupervised approach is used, may even generate new behavioral classifications,

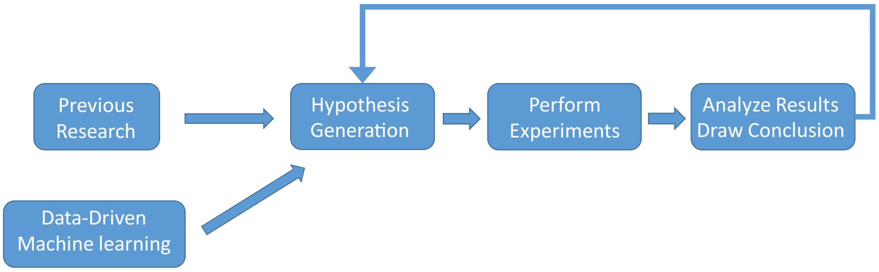


Figure 1. Model for data-driven science supporting hypothesis-driven science. Within the framework of hypothesis-driven science, machine learning can be used to generate hypotheses to be subsequently tested.

unlimited by human a priori behavioral classification or labels (Anderson and Perona, 2014). In fMRI, data-driven algorithms have been used to parcellate the brain based on fMRI functional connectivity data, yielding a functionally relevant fMRI atlas free of the constraints of a priori brain parcellations and labels (Craddock et al., 2012). In another fMRI study, Chang et al., (2015) used machine learning to identify a sensitive and specific neural signature of affective responses to aversive images that was unresponsive to physical pain, thus allowing them to infer neural components differentiating negative emotion from pain, "providing a basis for new, brain-based taxonomies of affective processes." Along these lines, independent component analysis and classification algorithms have been used to infer neural networks, decode brain states, or separate noise from signal (Jung et al., 2001; Thomaset al., 2002; Zuo et al., 2010; Lemm et al., 2011; Calhoun et al., 2014; Whitmore and Lin, 2016). Such strategies for capitalizing on the high dimensionality and multivariate nature of data are certainly not unique to neuroscience; the entire field of bioinformatics has emerged from this idea and is deeply rooted in machine learning (Larrañaga et al., 2006; Libbrecht and Noble, 2015).

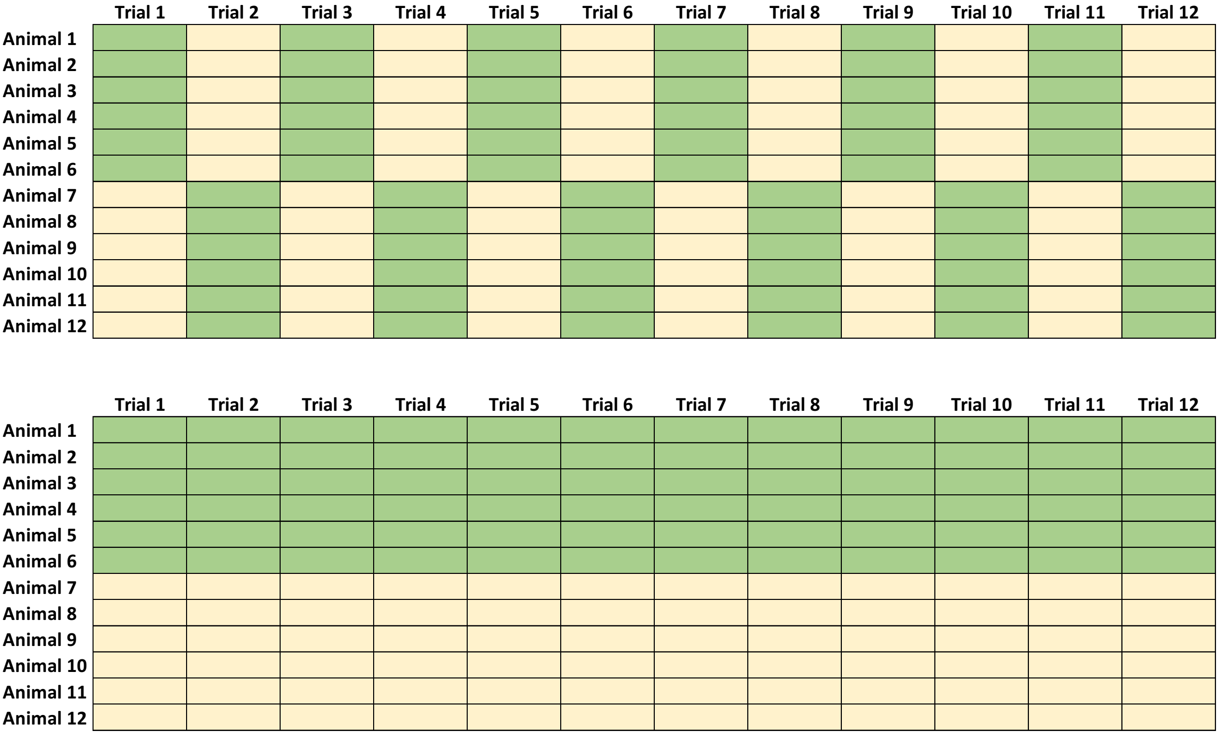
By revealing structure in the data, such machine learning approaches may yield new, testable hypotheses. In a study examining the neural mechanisms underlying stress-induced behavioral adaptation, D. Carlson et al. (2017) recorded cellular activity and local field potentials from multiple brain regions. Using a supervised machine-learning approach, they found that cellular activity in two of the recorded brain regions (infralimbic cortex and medial dorsal thalamus) showed adaptation across repeated exposure to stress. This led to a series of follow-up experiments to investigate whether and how these two regions were connected. The two regions were found to be functionally connected via cross-frequency phase coupling, which was further confirmed via an optogenetic manipulation experiment. In this case, machine learning revealed potential predictors, generated a relevant hypothesis space, and led to the design of a (successful) confirmatory experiment.

While certainly powerful as a brute force approach to hypothesis generation, data-driven machine learning approaches may also yield more direct insight into brain function. After all, the problems solved by the brain have strong parallels to the problems solved by machine learning. For example, we as humans must be able to make sense of all the multisensory information coming into the brain to make relevant inferences, such as whether a person we are talking to may be angry (Wegrzyn et al., 2015). Similarly, machine learning algorithms must learn structure from large multidimensional data; moreover, it can be shown that allowing data from multiple modalities to fully interact and inform each other leads to more powerful biomarkers (Levin-Schwartz et al., 2017). Such parallels have made for incredibly powerful cross-pollination between machine learning and neuroscience. For example, reinforcement learning first emerged in computer science in the 1950s (Bellman, 1957, 2010; Sutton and Barto, 1998). Quantitative models of reinforcement learning emerged to describe error-based learning (Bush and Mosteller, 1951; Rescorla and Wagner, 1972; Mackintosh, 1975; Pearce and Hall, 1980). Then in a seminal paper, Schultz et al. (1997) showed neuronal evidence for reward prediction error, a key element of reinforcement learning, in the brain, invigorating a whole branch of neuroscience (Niv, 2009; Gershman and Daw, 2017). Another example is that of deep learning, a family of machine learning algorithms aimed at learning data representations; early artificial neural networks were inspired by neurobiology (McCulloch and Pitts, 1990). The deep learning field continued to advance, and developments made on the computational front have now inspired hypotheses on the neuroscience front (Marblestone et al., 2016). Thus, because of the similarity of the problems being solved by machine learning algorithms and the brain, statistical and computational developments can inform neuroscience and yield new theories of brain function.

Validation of machine learning results

A common criticism of data-driven approaches is that they can be void of mechanism and thus can limit inferences and interpretation (T. Carlson et al., 2017). To return to the study on neural adaptation to repeated stress exposure (D. Carlson et al., 2017), the discovery that the cell firing in the infralambic cortex and medial dorsal thalamus was related to the behavior provided little insight into the mechanism per se. What were the cells doing? How were the regions connected? In which direction does the information flow? Does the behavior drive the activity or vice versa? Only through our follow-up experiments was it determined that the two regions were functionally connected through cross-frequency phase coupling. Further optogenetic manipulation experiments revealed that the changes observed in this circuitry were part of a compensatory mechanism in response to repeated stress exposure.

As illustrated, one way to overcome the limitations of each of these approaches is to design experiments that leverage the advantages of both approaches (Fig. 1). A scientific study could be divided into two phases. The first phase would be aimed at exploration, discovery, and ultimately hypothesis generation. For instance, in an animal study, initial experiments would focus on collecting large, broad datasets, such as cellular activity, local field potentials, motion, behavior, etc., from a mouse as it undergoes a contextual manipulation. Machine learning approaches could identify relationships among the dimensions in ways that relate to the mouse's physiology, behavior, and context, yielding specific hypotheses regarding these relationships. The second phase of the experiment would then be designed to test these concrete hypotheses, using a variety of techniques for biological manipulation. Viral strategies enable us to generate mice with specific and conditional genetic mutations. Technologies, such as optogenetics and designer receptors exclusively activated by designer drugs (DREADDs), allow us to manipulate the activity of specific cell types, and in the case of optogenetics, with precise timing and frequency (Boyden et al., 2005; Armbruster et al., 2007). Such



	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10	Trial 11	Trial 12
Animal 1	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨
Animal 2	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨
Animal 3	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨
Animal 4	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨
Animal 5	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨
Animal 6	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨
Animal 7	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩
Animal 8	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩
Animal 9	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩
Animal 10	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩
Animal 11	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩
Animal 12	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10	Trial 11	Trial 12
Animal 1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Animal 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Animal 3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Animal 4	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Animal 5	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Animal 6	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Animal 7	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Animal 8	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Animal 9	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Animal 10	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Animal 11	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Animal 12	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨

Figure 2. Hold-out trial versus out-of-sample model validation. Validation commonly accomplished within animal. For example, a model might be trained on subsets of each animal's data (top, green) and tested on the remainder of data from the same animal (top, yellow). Here, we propose training on a subset of the animals (bottom, green), and testing on an independent set of animals (bottom, yellow).

technologies allow us to test more refined hypotheses and discover more specific mechanisms underlying brain function, as a recent multisite *in vivo* recording study showed by linking large-scale neural dynamics to stress-related behavioral dysfunction (Hultman et al., 2016). In human subjects, a similar chain of events could be performed. From machine learning on a high-dimensional fMRI or EEG dataset, one could generate hypotheses about functional encoding of behavior, then manipulate the putative encoding through focused brain stimulation. Invasive and noninvasive stimulation techniques are far from the precision of optogenetics, but there is emerging evidence that they can be used to change specific features of brain activity (Smart et al., 2015; Philip et al., 2017; Widge et al., 2017).

Regardless of whether we use these machine learning approaches to generate hypotheses or to learn something about the computations being performed in the brain, our success will depend heavily on the model space we choose. A common aphorism in statistics states, "All models are wrong, but some are useful" (Box, 1979). In other words, no model we come up with can fully describe what is happening in a given system; however, a model can nonetheless be useful, capturing some amount of truth or making reliable predictions. So how do we know which models are useful? Validation is key. For example, although Drysdale et al. (2017) were able to determine patterns of neural activity that clustered with behavioral symptoms and treatment responses in a cohort of depressed subjects, the scientific and potential clinical value of their finding would have been dramatically diminished if their model failed to predict which treatments to which a new patient with depression would respond. Simply put, we can find regularities in nearly any dataset, but the real test is whether those regularities or predictions hold up using additional (nonoverlapping) datasets handled in the same manner (Woo et al., 2017). Generally, this level of validation is performed within research groups, where model predictions generated on a subset of data acquired during an experiment are validated on another subset of data that was held out during the model generation stage. For example, cellular firing data acquired for animals during a selected subset of experimental days may be used to develop the machine learning models, whereas data from other days may be used for model validation (Fig. 2, top). The limitation of this strategy is that, while the model solution may apply to the specific group of animals used for experimental testing, it may not extrapolate to new animals. Thus, it is more optimal for the model validation to be performed out of subject. In this scenario, the model would be developed using all of the trials from a specific set of experimental animals, then validated using another set of experimental animals (Fig. 2, bottom). The next level of validation is determining whether similar findings will hold up across research groups. Indeed, Drysdale et al. (2017) replicated their initial findings from their unsupervised clustering analysis in a completely separate dataset, thus adding confidence that these findings are robust. This level of validation has always been the gold standard for both data- and hypothesis-driven scientific discovery.

Explainable artificial intelligence: a vision for machine learning

With proper validation, machine learning has great promise both within and beyond the hypothesis-driven experimental framework. Nonetheless, machine learning further holds the potential to generate unifying models of brain function and behavior. Maximizing this potential of machine learning in neuroscience will require a different type of validation approach that emphasizes interpretability and generalizability. In other words, do the machine learning-discovered models capture fundamental principles of brain function and reflect causative phenomenon that

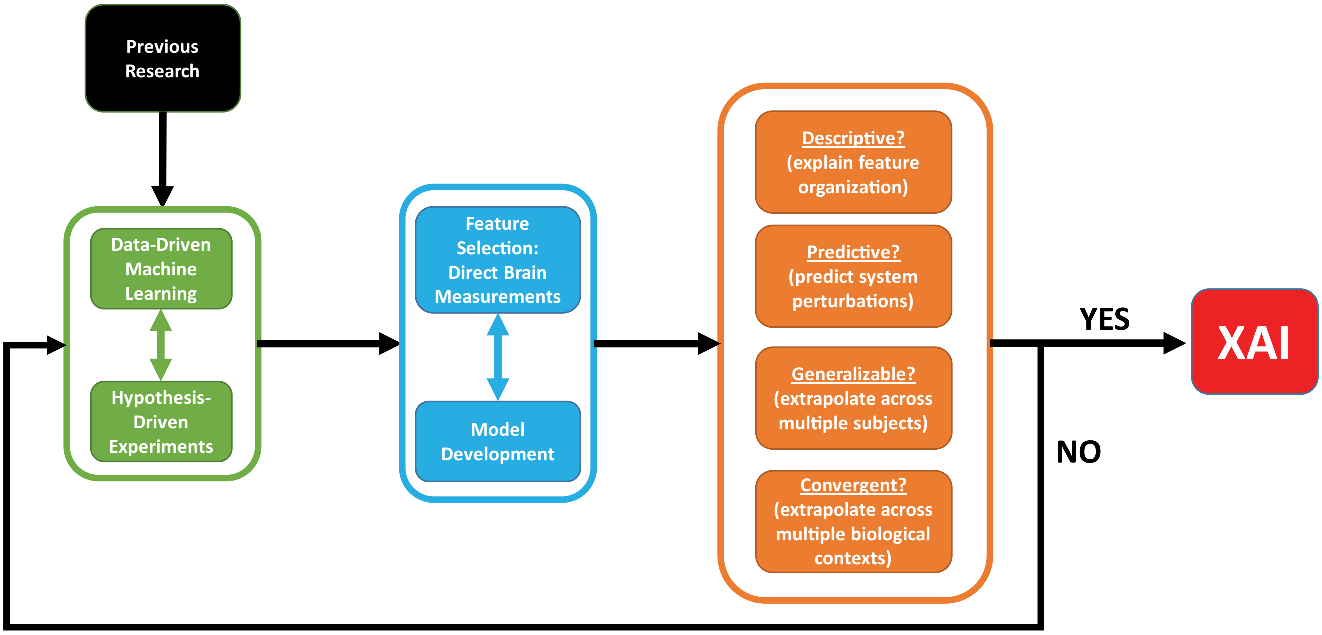

Previous Research
Data-Driven Machine Learning
Hypothesis-Driven Experiments
Feature Selection: Direct Brain Measurements
Model Development
Descriptive? (explain feature organization)
Predictive? (predict system perturbations)
Generalizable? (extrapolate across multiple subjects)
Convergent? (extrapolate across multiple biological contexts)
YES
NO
XAI

Figure 3. XAI. Here we present a vision for leveraging machine learning toward developing unified models. The criteria for models achieving XAI are that they must be based on measurable brain biology and be descriptive, predictive, generalizable, and convergent.

extrapolate across multiple biologically relevant contexts? Explainable artificial intelligence (XAI) emphasizes the development of more interpretable, explainable models and should be the ultimate goal of big data-neuroscience (Fig. 3). To achieve XAI, machine learning models must accomplish a broad set of comprehensive goals. First, models must be based on measurements of brain biology, such as cell firing, local field potential oscillations, protein levels, BOLD MRI signal, and scalp EEG. Second, models must explain how measured features are organized relative to each other (explainable generative models, which provide a probabilistic description of how the complex observed data can be synthesized from simpler, explainable properties). Explainability, or interpretability, is crucial; although a very complex model might yield more accurate predictions than a simpler model, a high priority for XAI is understanding how the model works and how the variables interact. Third, models must predict specific outcomes, such as behavioral or physiological changes in response to perturbations on the system (predictive models and discriminative, or conditional, models). Fourth, models must extrapolate across multiple subjects (generalizable models). Fifth, models must extrapolate across multiple biological contexts (convergent models). An example of an explainable model is the Krebs cycle, which describes how multiple enzymes and substrates are organized together to generate energy. Additionally, this model of energy production extrapolates across individuals, and across many biological models of health and disease.

XAI fits well with the National Institute of Mental Health's Research Domain Criteria framework, which strives to integrate many levels of explanation to better understand basic dimensions of human brain function underlying behavior. An XAI model may, for example, explain how gene expression, cell firing, and/or oscillations across multiple brain regions are organized within a biological network. The dependencies inferred among these neural network components should show predictable changes, given an experimental perturbation of gene expression, or cellular excitability, or oscillatory synchrony. The model would also explain how this network relates to a behavior, such as sociability, and it would extrapolate across multiple individuals. Finally, this model would explain why normal social behavior is disrupted in seemingly distinct clinical phenomena, such as depression and autism. Thus, XAI could comprehensively explain phenomena at multiple levels and their relationships with each other, and demonstrate that this explanation withstands hypothesis-based perturbations and validation.

Developing these explainable models will require big datasets. One potential strategy for acquiring these datasets could involve longitudinal observations in a relatively small number of subjects (i.e., in the hundreds range). This approach would ultimately facilitate many repeated observations of the brain during behavior. The explainable models would then be built using within-subject variance, isolating the relationship between brain function and behavior relative to a drifting biological baseline. The advantage of this approach is that it can initially be implemented by a smaller number of research groups. Another potential strategy could be to collect time-limited data across a much larger number of subjects. (i.e., in the tens of thousands to millions range). The explainable models would then be built using across-subject variance. This latter approach would likely require that many research groups collaborate in the data collection phase. A critical first step would be to align collection methods, standards, and paradigms across a broad research community, as many within the neuroimaging community are already doing (Poldrack et al., 2013; Gorgolewski et al., 2016).

Because developing these explainable models will ultimately also require observations and hypothesis testing in both human and animal studies, directed efforts to build transdisciplinary teams made up of neuroscience researchers, clinicians, and data scientists with varying levels of analytical expertise are warranted. These directed efforts may include developing novel funding mechanisms, or revising current peer review processes to prioritize grant applications that include both human and animal studies. Nevertheless, it will not simply be sufficient to build teams that include expertise in genetics, cellular/molecular neuroscience, systems neuroscience, cognitive neuroscience, treatment paradigms including pharmacology and neuromodulation, behavior quantification in health and disease, statistics, and machine learning. Rather, a unique new group of scientists capable of bridging the broad gaps between these disciplines will be needed to yield the promise of XAI. These scientific "integrators" will need expertise in multiple disciplines, enabling them to successfully translate across the intellectual and cultural boundaries that exist between fields. These boundaries may exist at the level of logic constructs (e.g., the difference between frequentist and Bayesian statistics), or at the level of simple language. For example, the word "model" has a different connotation for each of the fields described above.

So where do we find these scientific integrators? Simply put, we must train them. The medical scientist training program, in which students are trained as both basic scientists and clinicians, is a long-standing example of an approach for developing scientific integrators. Along these lines, our nation's neuroscience leadership has highlighted the important role that psychiatrists cross-trained in engineering/mathematics will play in advancing neuroscience, and grant mechanisms that foster postdoctoral cross-training in neuroscience and data science have recently been promoted through the national BRAIN Initiative (National Institutes of Mental Health, 2017). Finally, we must continue to adapt our neuroscience ecosystem to promote studies that advance XAI. For example, federal agencies can optimize grant review processes both by promoting the broad participation of data scientists and scientific integrators in peer review panels, and by educating peer reviewers on the strengths of data-driven science. There is without doubt still room and a necessary scientific role for traditional hypothesis-driven experiments; but if we are to expand neuroscience to incorporate big data and XAI, then we must allow for and encourage interdisciplinary integrative peer review as well.

Machine learning thus holds great promise in advancing the field of neuroscience, not as a replacement for hypothesis-driven research, but in conjunction with it. Machine learning tools can bolster large-scale hypothesis generation, and they have the potential to reveal interactions, structure, and mechanisms of brain and behavior. Importantly, given the dangers of spurious findings or explanations void of mechanism, care must be taken to ensure the utility of such an approach. It is with proper replication, validation, and hypothesis-driven confirmation that machine learning analysis approaches will fulfill the great promise they hold, allowing us to make greater strides toward understanding how the brain works.

References

Anderson DJ, Perona P (2014) Toward a science of computational ethology. Neuron 84:18–31. CrossRef Medline

Andersson JL, Jenkinson M, Smith S (2007) Non-linear registration aka spatial normalization: FMRIB Technical Report TR07JA2. https://www.fmrib.ox.ac.uk/datasets/techrep/tr07ja2/tr07ja2.pdf.

Armbruster BN, Li X, Pausch MH, Herlitze S, Roth BL (2007) Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. Proc Natl Acad Sci U S A 104:5163–5168. CrossRef Medline

Bargmann C, Newsome W, Anderson A, Brown E (2014) BRAIN 2025: a scientific vision. Bethesda, MD: Brain Research through Advancing Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee.

Bellman R (1957) A Markovian decision process. J Math Mech 6:679–684.

Bellman R (2010) Dynamic programming. Princeton, NJ: Princeton UP.

Box GE (1979) Robustness in the strategy of scientific model building (MRC-TSR-1954). Madison, WI: Wisconsin University, Madison Mathematics Research Center.

Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K (2005) Millisecond-timescale, genetically targeted optical control of neural activity. Nat Neurosci 8:1263–1268. CrossRef Medline

Bush RR, Mosteller F (1951) A mathematical model for simple learning. Psychol Rev 58:313–323. CrossRef Medline

Calhoun VD, Miller R, Pearlson G, Adalı T (2014) The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. Neuron 84:262–274. CrossRef Medline

Carlson D, David LK, Gallagher NM, Vu MT, Shirley M, Hultman R, Wang J, Burrus C, McClung CA, Kumar S, Carin L, Mague SD, Dzirasa K (2017) Dynamically timed stimulation of corticolimbic circuitry activates a stress-compensatory pathway. Biol Psychiatry 82:904–913. CrossRef Medline

Carlson T, Goddard E, Kaplan DM, Klein C, Ritchie JB (2017) Ghosts in machine learning for cognitive neuroscience: moving from data to theory. Neuroimage. Advance online publication. Retrieved Aug. 6, 2017. CrossRef Medline

Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD (2015) A sensitive and specific neural signature for picture-induced negative affect. PLoS Biol 13:e1002180. CrossRef Medline

Craddock RC, James GA, Holtzheimer PE 3rd, Hu XP, Mayberg HS (2012) A whole brain fMRI atlas generated via spatially constrained spectral clustering. Hum Brain Mapp 33:1914–1928. CrossRef Medline

Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, Fetcho RN, Zebley B, Oathes DJ, Etkin A, Schatzberg AF, Sudheimer K, Keller J, Mayberg HS, Gunning FM, Alexopoulos GS, Fox MD, Pascual-Leone A, Voss HU, Casey BJ, et al. (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23:28–38. CrossRef Medline

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, Caviness V, Makris N, Rosen B, Dale AM (2004) Automatically parcellating the human cerebral cortex. Cereb Cortex 14:11–22. CrossRef Medline

[No authors listed] (2014) Focus on big data. Nat Neurosci 17:1429.

Gershman SJ, Daw ND (2017) Reinforcement learning and episodic memory in humans and animals: an integrative framework. Annu Rev Psychol 68:101–128. CrossRef Medline

Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, Flandin G, Ghosh SS, Glatard T, Halchenko YO, Handwerker DA, Hanke M, Keator D, Li X, Michael Z, Maumet C, Nichols BN, Nichols TE, Pellman J, Poline JB, et al. (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 3:160044. CrossRef Medline

Hong W, Kennedy A, Burgos-Artizzu XP, Zelikowsky M, Navonne SG, Perona P, Anderson DJ (2015) Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. Proc Natl Acad Sci U S A 112:E5351–E5360. CrossRef Medline

Hultman R, Mague SD, Li Q, Katz BM, Michel N, Lin L, Wang J, David LK, Blount C, Chandy R, Carlson D, Ulrich K, Carin L, Dunson D, Kumar S, Deisseroth K, Moore SD, Dzirasa K (2016) Dysregulation of prefrontal cortex-mediated slow-evolving limbic dynamics drives stress-induced emotional pathology. Neuron 91:439–452. CrossRef Medline

Jain V, Seung HS, Turaga SC (2010) Machines that learn to segment images: a crucial technology for connectomics. Curr Opin Neurobiol 20:653–666. CrossRef Medline

Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156. CrossRef Medline

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841. CrossRef Medline

Jung TP, Makeig S, McKeown MJ, Bell AJ, Lee TW, Sejnowski TJ (2001) Imaging brain dynamics using independent component analysis. Proc IEEE 89:1107–1122. CrossRef Medline

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, Robles V (2006) Machine learning in bioinformatics. Brief Bioinform 7:86–112. CrossRef Medline

Lemm S, Blankertz B, Dickhaus T, Müller KR (2011) Introduction to machine learning for brain imaging. Neuroimage 56:387–399. CrossRef Medline

Levin-Schwartz Y, Calhoun VD, Adalı T (2017) Quantifying the interaction and contribution of multiple datasets in fusion: application to the detection of schizophrenia. IEEE Trans Med Imag 36:1385–1395. CrossRef Medline

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16:321–332. CrossRef Medline

Mackintosh NJ (1975) A theory of attention: variations in the associability of stimuli with reinforcement. Psychol Rev 82:276–298. CrossRef

Marblestone AH, Wayne G, Kording KP (2016) Toward an integration of deep learning and neuroscience. Front Comput Neurosci 10:94. CrossRef Medline

McCulloch WS, Pitts W (1990) A logical calculus of the ideas imminent in nervous activity. Bull Math Biol 52:99–115; discussion 73–97. Medline

Niv Y (2009) Reinforcement learning in the brain. J Math Psychol 53:139–154. CrossRef

O'Donnell LJ, Suter Y, Rigolo L, Kahali P, Zhang F, Norton I, Albi A, Olubiyi O, Meola A, Essayed WI, Unadkat P, Ciris PA, Wells WM 3rd, Rathi Y, Westin CF, Golby AJ (2017) Automated white matter fiber tract identification in patients with brain tumors. Neuroimage Clin 13:138–153. CrossRef Medline

Paul A, Crow M, Raudales R, He M, Gillis J, Huang ZJ (2017) Transcriptional architecture of synaptic communication delineates GABAergic neuron identity. Cell 171:522–539.e20. CrossRef Medline

Pearce JM, Hall G (1980) A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87:532–552. CrossRef Medline

Philip NS, Nelson BG, Frohlich F, Lim KO, Widge AS, Carpenter LL (2017) Low-intensity transcranial current stimulation in psychiatry. Am J Psychiatry 174:628–639. CrossRef Medline

Poldrack RA, Gorgolewski KJ (2014) Making big data open: data sharing in neuroimaging. Nat Neurosci 17:1510–1517. CrossRef Medline

Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, Cumba C, Koyejo O, Milham MP (2013) Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neurosci 7:12. CrossRef Medline

Rescorla RA, Wagner A (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement, Vol 2: Current research and theory. (pp. 64–69) New York, NY: Appleton-CenturyCrofts, 1972.

National Institutes of Mental Health, NIH (2017) BRAIN Initiative Fellows: Ruth L. Kirschstein National Research Service Award Individual Postdoctoral Fellowship F32. https://grants.nih.gov/grants/guide/rfa-files/RFA-MH-17-250.html. Accessed November 24, 2017. Advance online publication.

Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3:210–229. CrossRef

Schuck NW, Gaschler R, Wenke D, Heinzle J, Frensch PA, Haynes JD, Reverberi C (2015) Medial prefrontal cortex predicts internally driven strategy shifts. Neuron 86:331–340. CrossRef Medline

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599. CrossRef Medline

Sejnowski TJ, Churchland PS, Movshon JA (2014) Putting big data to good use in neuroscience. Nat Neurosci 17:1440–1441. CrossRef Medline

Smart OL, Tiruvadi VR, Mayberg HS (2015) Multimodal approaches to define network oscillations in depression. Biol Psychiatry 77:1061–1070. CrossRef Medline

Spellman T, Rigotti M, Ahmari SE, Fusi S, Gogos JA, Gordon JA (2015) Hippocampal-prefrontal input supports spatial encoding in working memory. Nature 522:309–314. CrossRef Medline

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction, Vol 1. Cambridge, MA: Massachusetts Institute of Technology.

Thomas CG, Harshman RA, Menon RS (2002) Noise reduction in BOLD-based fMRI using component analysis. Neuroimage 17:1521–1537. CrossRef Medline

Wegrzyn M, Bruckhaus I, Kissler J (2015) Categorical perception of fear and anger expressions in whole, masked and composite faces. PloS One 10: e0134790. CrossRef Medline

Whitmore NW, Lin SC (2016) Unmasking local activity within local field potentials (LFPs) by removing distal electrical signals using independent component analysis. Neuroimage 132:79–92. CrossRef Medline

Widge AS, Ellard KK, Paulk AC, Basu I, Yousefi A, Zorowitz S, Gilmour A, Afzal A, Deckersbach T, Cash SS, Kramer MA, Eden UT, Dougherty DD, Eskandar EN (2017) Treating refractory mental illness with closed-loop brain stimulation: progress towards a patient-specific transdiagnostic approach. Exp Neurol 287:461–472. CrossRef Medline

Woo CW, Chang LJ, Lindquist MA, Wager TD (2017) Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 20: 365–377. CrossRef Medline

Zuo XN, Kelly C, Adelstein JS, Klein DF, Castellanos FX, Milham MP (2010) Reliable intrinsic connectivity networks: test-retest evaluation using ICA and dual regression approach. Neuroimage 49:2163–2177. CrossRef Medline