Machine Learning for Clinical Psychology and Clinical Neuroscience

Marc N. Coutanche, Ph.D[1,2,3]* and Lauren S. Hallion, Ph.D.[1,3]

[1] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
[2] Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, USA
[3] Center for the Neural Basis of Cognition, Pittsburgh, PA, USA

* Corresponding author: marc.coutanche@pitt.edu

Citation:

Coutanche, M. N. & Hallion, L. S. (2020). Machine learning for clinical psychology and clinical neuroscience. In A. G. C. Wright and M. N. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology.* Cambridge, UK: Cambridge University Press.

Abstract:

A rapid growth in computational power and an increasing availability of large, publicly-accessible, multimodal datasets present new opportunities for psychology and neuroscience researchers to ask novel questions, and to approach old questions in novel ways. Studies of the personal characteristics, situation-specific factors, and sociocultural contexts that result in the onset, development, maintenance, and remission of psychopathology, are particularly well-suited to benefit from machine learning methods. However, introductory textbooks for machine learning rarely tailor their guidance to the needs of psychology and neuroscience researchers. Similarly, the traditional statistical training of clinical scientists often does not incorporate these approaches. This chapter acts as an introduction to machine learning for researchers in the fields of clinical psychology and clinical neuroscience. We discuss these methods, illustrated through real and hypothetical applications in the fields of clinical psychology and clinical neuroscience. We touch on study design, selecting appropriate techniques, how (and how not) to interpret results, and more, to aid researchers who are interested in applying machine learning methods to clinical science data.

Introduction

"In all its complexity, the question towards which all outcome research should ultimately be directed is the following: *What* treatment, by *whom*, is the most effective for *this* individual with *that* specific problem, and under *which* set of circumstances?" (Paul, 1967, p. 111)

This question, posed by Gordon Paul in 1967 (more recently distilled to its essence as, "what works for whom, and under what circumstances?"), has remained one of the foundational and as-yet unresolved problems of modern clinical psychology. Although the original quotation refers to questions of treatment selection, the spirit is broadly applicable as we seek to ask: what are the personal characteristics, sociocultural contexts, and situation-specific factors that result in the onset, development, maintenance, and remission of psychopathology? More recently, the question has also been extended to the level of the individual, with the goal of identifying the factors that are most relevant for understanding *this individual's* clinical presentation and treatment needs.

These types of questions form the core of much basic and applied research in the field of clinical psychology. Although many studies have made significant and meaningful progress toward answering one or two facets of these questions in isolation, questions of "what, for whom, under what circumstances, and why" remain largely unaddressed at an integrative level. Current challenges to answering these questions include small sample sizes, large and unwieldy bodies of research that have developed in relative isolation, and – most importantly for the purposes of this chapter – limited access to the statistical tools that would be best-suited to addressing these problems. Unfortunately, even the best efforts to implement **precision medicine** in a clinical psychology context have been constrained by a mismatch between the inherent complexity of these questions (e.g., Cohen & DeRubeis, 2018) and the relatively narrow "statistical toolbelt" that comes standard-issue in most clinical psychology training programs. In our experience, some of the most significant barriers to clinical psychology graduate students being able to use these particular advanced statistical methods include: 1) a limited availability of formal training opportunities that are geared to non-computer scientists; 2) informed decisions to de-prioritize methods that are not directly applicable to current research projects; and 3) textbooks that were written primarily with computer scientists or statisticians in mind, such that they assume a level of background knowledge that extends well beyond what most clinical psychology students have obtained in prior coursework.

The goal of this chapter is to help address some of the most significant barriers to entry with machine learning, with an eye toward questions of "what, for whom, under what circumstances, and why?" Our goal is to present a clear and concise introduction and how-to guide that covers basic principles and example applications of machine learning to questions in psychology.

What is Machine Learning?

The term "machine learning" refers to a range of mathematical techniques that leverage the computational power made available by modern computers to identify meaningful signals within large and often complex datasets. A common characteristic of these techniques is that they all identify patterns or relationships that are present in a given dataset and then extend those patterns to independent datasets to validate the model (Yarkoni & Westfall, 2017). The first phase, in which an algorithm is applied to describe patterns and relationships in one set of data, is described as the **learning** or **training** phase. The second phase, in which the resulting model is applied to a new dataset, is called the **testing** phase.

A common theme across techniques is a strong emphasis placed on **prediction**, or the extent to which a model that is trained on one set of data can successfully predict patterns and relationships in a new (untrained) dataset. As in traditional statistical approaches (e.g., linear regression), machine learning models are based on the patterns and relationships in data. However, whereas a traditional linear regression is evaluated on the basis of how well it accounts for patterns in an original dataset, a particular trained machine learning model is evaluated on the basis of how accurately it can predict patterns and relationships in new data. The ability of a model that was trained in one dataset to predict patterns in another is described as the model's ability to **generalize**.

Another unique and central feature of machine learning is its focus on multidimensionality. Although there are certainly good reasons to ask whether a given predictor is independently associated with an outcome absent the effects of all other constructs, this artificial scenario does little to capture the complex interplay that exists in reality. Instead, it may be more meaningful to ask how an outcome is predicted in the context of many predictors. This multidimensionality allows a model to identify subtle patterns in datasets, which have often been likened to **fingerprints**. Indeed, it has been suggested that in areas of psychology in which simple mechanistically understood models fail to successfully predict any behavior beyond a particular dataset, it is time to switch to prediction (rather than explanation alone) as a goal, even at the cost of interpretability (Yarkoni & Westfall, 2017).

Beyond Traditional Statistics: What Does Machine Learning Add?

The question of what "counts" as machine learning can be somewhat murky at times. This is because, at their core, some machine learning techniques rely on the same basic mathematical principles that are applied in traditional statistical analyses (e.g., logistic regression and the general linear model). In the definition of machine learning above, we suggested two key aspects that define and distinguish this set of methods, prediction and multidimensionality, which we expand on below.

Prediction and Generalization

In traditional statistical methods, a result (such as an *F* or *t* statistic) reflects the extent to which the model provides a strong fit to the current dataset. However, this value does not reflect the extent to which this *particular* model will apply beyond the dataset in which it was tested (Yarkoni & Westfall, 2017). For example, the specific pattern of beta coefficients from a linear regression analysis may not adequately describe a different set of patients or research participants. This lack of external validation has been identified as one major reason for recent failures to replicate particular biomarkers of psychiatric treatment outcome (Gillan & Whelan, 2017). In line with these observations, the recent push toward open science and reproducibility has led to an increased emphasis on the importance of replication as a criterion to evaluate the robustness of a given finding (Open Science Collaboration, 2015). However, the decision to conduct a replication study is nontrivial. Even a straightforward replication can require significant time, effort, and expense – none of which most independent scientists (let alone most graduate students) have to spare. Further, a failure to replicate can arise for a number of reasons, and as such can sometimes be difficult to interpret. A machine learning approach including **cross-validation** (discussed below) therefore offers a researcher several distinct advantages.

One major advantage is a machine learning approach's ability to quantify the generalizability of a "learned" pattern or algorithm to new data. Indeed, success is typically operationalized as the concordance between the values predicted from a previously trained model and actual values observed in different data. We discuss the best ways to determine which data-points are used for training versus testing below, but one common and intuitive approach is to **hold-out** anywhere from a single participant at a time (repeated iteratively for each participant) to up to half the sample. These held-out participants are considered the "test" dataset and are excluded from any analyses that are performed in the training dataset. This includes any imputation, transformation, or other data cleaning or preprocessing procedures that might use the entire dataset.

Alternatively, a researcher may approach the question of generalization from a more conservative or theoretically driven perspective. In this case, the researcher might consider whether a model that is trained on a dataset with certain characteristics (e.g., undergraduate students) can provide accurate predictions in a dataset that has different sample characteristics (e.g., treatment-seeking adults) or that involve the administration of a different treatment. The more “steps removed” between the training and testing datasets (that is, the more the training and testing datasets differ in research context, study participants, and so on), the stronger the test of generalization (see Gillan & Whalen, 2017 for a detailed consideration of this issue). From a theoretical perspective, discovering that a model performs better for some datasets than others can be informative, particularly if there are potentially meaningful differences between datasets. For example, Chekroud and colleagues (2016) applied machine learning with cross-validation to clinical trial data of response to a 12-week course of citalopram (a selective serotonin reuptake inhibitor; SSRI) in patients with depression (in the Sequenced Treatment Alternatives to Relieve Depression trial; STAR*D). Patient-reported variables were used to predict symptom remission with 64.6% accuracy in the STAR*D cohort. Reported clinical features such as insomnia and somatic complaints contributed to the model’s performance. The authors also examined the extent to which the model generalized to an independent clinical trial of response to escitalopram in a similar sample (Combining Medications to Enhance Depression Outcomes; COMED). Prediction was significantly above chance (59.6%) using the STAR*D model to predict escitalopram (versus placebo) response in this separate (COMED) trial. Notably, prediction accuracy dropped to near-chance (51.4%) when predicting response to a serotonin-norepinephrine reuptake inhibitor (SNRI; venlafaxine-mirtazapine) in the COMED trial: evidence that the model was specific to predicting response to SSRIs.

Multidimensionality

A further advantage of machine learning lies in its ability to account for the multidimensionality of psychological data. With a linear regression analysis, we typically ask if a particular variable (represented by a coefficient) is a significant predictor. For most machine learning techniques, the focus is less on "does this one variable predict the outcome?" and is closer to "does this suite of variables predict the outcome?" Successful solutions can include scenarios where individual variables only have marginal predictive ability but (when pooled together) form a highly predictive *set*. This point – of using overall prediction accuracy, rather than the coefficient for any one or more variables – is especially important because a given individual variable could contribute to successful predictions in a way that is not clearly revealed by inspecting beta weights (i.e., as in traditional linear regression). Instead, we consider unique predictor profiles, patterns, or fingerprints, without necessarily emphasizing specific details of that pattern.

As an illustrative example, one study found that individual symptoms did not differentiate unipolar from bipolar depression, but that the *pattern* of symptoms did (Perlis et at al., 2006). This concept has also played a significant role in the rapid development and application of machine learning to functional magnetic resonance imaging (fMRI) data. The discovery that *patterns* of brain activity can contain information about perceptual, cognitive and affective processes, when signals in individual voxels (or region averages) do not, continues to advance our understanding of brain systems with a degree of specificity that was not previously possible (Coutanche, 2013; Haxby et al., 2001; Tong and Pratte, 2012).

Considering multidimensional "profiles" of variables is not always intuitive, but it has nevertheless shown promise. For example, Astle and colleagues (2018) examined a heterogeneous sample of 530 children who were referred due to problems in attention, memory, language, or poor school performance. An unsupervised machine learning analysis (discussed further below) was conducted on performance across seven cognitive tasks. The authors employed Self Organizing Maps (SOMs; Kohonen, 1989), where an artificial neural network projects a multidimensional input space into two dimensions. The relevant output, a two-dimensional grid, represents the degree of similarity in the input data – here, task performance in individual children. This was combined with a ***k*-means clustering** approach (described in more detail below), which allocated the children to four empirically derived clusters based on their cognitive profiles. These clusters represented profiles corresponding to: 1) broad cognitive difficulties and significant reading, spelling and math problems; 2) age-typical cognitive and learning abilities; 3) working memory problems; or 4) phonological difficulties. Interestingly, the cognitive profiles of the children were not predicted by diagnostic status or the reason for their referral. The clusters did, however, correspond to differences in white matter brain networks measured through diffusion tensor MRI. These brain data had not been used to construct the SOMs or clusters, bolstering the validity of the detected clusters. This example identified data-driven neurocognitive dimensions underlying learning-related difficulties. Moreover, it highlights an advantage of recruiting and analyzing samples with varied, rather than "pure", deficits, which can often ignore people with comorbidity (a substantial portion of a population) and overemphasize similarity within groups. A focus on sampling along continuous dimensions in study design can provide new targets for approaches to early detection and intervention, as well as help to identify etiological mechanisms.

As we expand on below, machine learning techniques are also able to use a larger number of variables than standard statistical approaches. However, these approaches remain constrained by statistical power and are subject to the broad signal-to-noise and sample size considerations that also affect more traditional methods.

How Can Machine Learning Inform Clinical Theory?

The relevance and utility of machine learning techniques for addressing questions related to diagnosis, treatment selection, and prediction of critical behaviors is, in our opinion, fairly uncontroversial. A model that can predict how well an individual will respond to a treatment, or the likelihood that a given individual will die by suicide in a given 12-month period, would represent a significant and important advance with major clinical implications. However, one limitation of these techniques is that the full set of relationships between a model's predictors and its predictions are not always transparent. This limitation has resulted in machine learning approaches at times earning the (somewhat unflattering) "black box" descriptor. One response to this critique is that even the most opaque of black boxes can have its place. If an algorithm enhances prediction of important clinical information, it has value, irrespective of its transparency. However, the opacity of machine learning techniques does raise challenges when approached from a theoretical perspective. In this section, we consider whether, how, and to what extent, machine learning can be leveraged to inform theoretical models of psychopathology and its treatment.

Although theory-testing is not traditionally considered to be a strength of machine learning, there are nevertheless a number of ways in which these approaches can and have been creatively leveraged to advance theoretical understanding. For example, we have used a **feature ablation**/lesion approach (described below) in our own work (Hallion, Wright, Coutanche, Joormann, & Kusmierski, under review) to test theoretical predictions about the distinctions between different classes of perseverative thought (e.g., worry versus rumination). In that study, we identified features of thoughts that have been proposed to characterize and distinguish between different classes of perseverative thought (e.g., temporal orientation: worry is traditionally defined as future-oriented, whereas rumination is defined as present- or past-oriented; other features included valence, form, intrusiveness, and others). We asked participants to rate a sampling of their own thoughts according to each of those features. We then tested whether a model trained on those features could successfully predict scores on conventional self-report measures of worry, rumination, and other kinds of perseverative thought. Next, we examined the impact of ablating (removing) each feature that has been proposed as theoretically important for defining and distinguishing between the different types of perseverative thought. Within an ablation framework, a significant reduction in accuracy after removing a variable indicates that the variable contributed unique predictive power (i.e., offered predictive power that could not be obtained from the other variables in the model). Conducting this ablation process for each predictor in turn gives values for the relative influence of each variable. The analyses lent support to some theoretical distinctions between types of perseverative thought and raised challenges for others. For example, self-referential processing has been proposed to be a central feature of depressive rumination (Nolen-Hoeksema et al., 2008). Consistent with that account, ablating items related to self-referential processing significantly reduced the model's accuracy for predicting self-reported rumination, but did not impact accuracy for predicting other kinds of thought. However, although temporal orientation has previously been identified as the defining (and only reliable) difference between worry and rumination (Watkins, 2008), ablating items related to temporal orientation did not significantly or differentially impact the model's accuracy for predicting or discriminating between scores on traditional measures of worry and rumination. Altogether, these findings (along with a broader set of results not described here) pointed away from a categorical "subtype" model of perseverative thought and instead toward a fully dimensional model in which thoughts are most accurately described in terms of underlying features, rather than in terms of classes characterized *a priori* by the presence or absence of certain features.

This ablation approach could easily be adapted to examine the impact of including versus excluding any number of theory-derived predictors. Example questions might include:

- What is the impact of removing theoretically important (versus theoretically peripheral) predictors on predicting a symptom or syndrome of interest?
- Does including or removing a proposed risk factor change the model's ability to predict the development of psychopathology prospectively?

- Which baseline clinical features have the largest influence on the model's accuracy for predicting symptom trajectory or treatment response? (Inferred by an especially large drop in prediction when those features are removed.)
- Does the addition or removal of certain clinical characteristics improve the model's ability to discriminate between two closely related disorders (e.g., social anxiety disorder versus avoidant personality disorder) or constructs (e.g., fear versus anxiety)?

Generalization, or the validation of a model's success through prediction, can also be leveraged to test certain theoretical questions. Accuracy for predicting observations in a new dataset can both validate the original model and quantify the ability of this model to explain and predict observations from a set of data that differs in theoretically interesting ways. We discussed an example of generalization testing above in Chekroud et al. (2016), where a model trained on treatment response to citalopram was more accurate when making treatment predictions for escitalopram (which has a similar neural mechanism) versus an SNRI combination. Comparisons of generalizability can be extended to test theoretical ideas related to mechanism (do cognitive task outcomes improve prediction?), sample characteristics (does prediction vary by gender?), nosology (does the model distinguish between disorders?), and so on. By emphasizing the performance of a model in predicting outcomes in a new dataset, machine learning approaches are well-suited to testing specific, concrete, and testable predictions about the generality versus specificity of a given finding.

How Can Machine Learning Inform Clinical Neuroscience?

Machine learning techniques have a long history of application within a neuroimaging context. More recently, these techniques have been applied specifically to data from clinical samples and clinically relevant paradigms to enhance understanding of a wide variety of symptoms and disorders. These methods are more widely explored within those literatures so we do not provide a thorough treatment here (we refer the interested reader to Woo et al., 2017 for a review). We will instead touch briefly on some of the ways in which machine learning can be applied to neural data to make predictions about risk assessment, early detection, differential diagnosis, treatment response, and biological mechanisms.

Clinically predictive neural fingerprints can include a wide variety of neural signatures (see Sundermann et al., 2014 for a review), such as patterns of activity across voxels (Coutanche et al., 2011), resting-state connectivity (Du et al., 2012), task-based connectivity (Deshpande et al., 2013), and volumetric differences (Nouretdinov et al., 2011), among others. New developments in integrating machine learning with other brain measures, such as connectivity (Anzellotti & Coutanche, 2018), continue to present new targets for clinical investigations.

The availability of large datasets online, and prediction competitions (such as the attention deficit-hyperactivity (ADHD)-200 Global Competition; Bellec et al., 2017), have rapidly increased the number of classification studies focused on discriminating among patient groups or identifying transdiagnostic dimensional features. A prediction competition typically involves sharing a large dataset (including assigned labels, such as diagnostic status) with contestants who then develop (i.e., train) classifier models. A test dataset is then shared (without any labels). Contestants typically use their trained model to generate label predictions for the test dataset (e.g., diagnosis), which they submit to the contest organizers. The contest organizers can then grade each model based on the accuracy of its predictions. A welcome component of many recent contests is holding the test dataset in escrow, reducing the chance that final performance metrics might be inflated by accidental influence of test data.

In parallel, over the last ten years, the number of multi-site neuroimaging studies has risen significantly. These studies have several advantages to a researcher seeking to employ machine learning. Sample sizes in such studies can often number in the thousands, and having data from multiple sites allows models to be trained that are capable of generalizing to new sites and scanners (Sundermann et al., 2014).

An important consideration for the neuroimaging investigator is that a model can learn to use any features that are predictive of the relevant classification. On the one hand, this is why such models are so powerful, but on the other, this warrants caution, particularly because neuroimaging models can be influenced by clinically relevant factors that do not reflect the neural signals of interest, such as in-scanner head motion for predicting ADHD (Eloyan et al., 2012) or a strong effect of eye-blinks on predicting an autism spectrum disorder diagnosis (Eldridge et al., 2014). Such features are valid predictors for each respective disorder, but do not give insights into their neural underpinnings.

The application of machine learning techniques to understand the neural basis of psychopathology has yielded important insights. However, the utility of these techniques for matching individuals to treatments (which some might describe as the “holy grail” of clinical neuroscience) is as yet unclear. In their consideration of some of these issues, Gillan and Whelan (2017) suggest that efforts to develop a comprehensive computational neural model of specific clinical symptoms could result in a “dead end,” not because of methodological limitations, but because symptoms are rarely (if ever) pathognomonic. Just as a headache could result from any number of underlying pathologies, so too can many clinical symptoms. We do not believe that this equifinality renders futile any efforts to develop neural models of these symptoms, but it does raise theoretical, methodological, and interpretational challenges. Similarly, a classifier may be trained to differentiate between different disorders or syndromes, but if the *a priori* distinctions are artefactual or otherwise spurious, a classifier with perfect accuracy will nevertheless still fail to truly “carve nature at its joints.”

Down to Brass Tacks

An example application. Suppose we are broadly interested in understanding the cognitive and neural mechanisms that contribute to maintaining anxiety-related psychopathology. Let's further suppose that we have access to a large, multimodal dataset that includes an array of predictors and outcomes of interest. Our participants (cases) include over 1,000 individuals who responded to an advertisement for a research study on "worry and anxiety." Our variables include performance on a range of traditional neuropsychological and other cognitive tasks, resting state data from fMRI, self-report measures for a variety of anxiety symptoms, and DSM-5 diagnostic status as determined in a clinician-administered, semi-structured diagnostic interview. There are a multitude of potentially promising questions that we could ask using this dataset. As the chapter progresses, feel free to refer to these boxes to see example applications to this hypothetical dataset for the techniques we discuss.

A wide array of methods is available to the researcher wishing to apply machine learning to address clinical problems. These techniques can be broadly divided into **supervised learning** or **unsupervised learning** methods. Here, supervision (or lack thereof) refers to whether the training dataset includes labels or other information that corresponds to the criterion (or "correct answer"). An example of supervised learning, which we discuss further below, would be a **machine learning classifier**. In a classifier approach, the analyst assigns a label (e.g., diagnostic status) to each observation, which reflects the "true" class. This label could come from sources such as an expert's diagnosis or a genetic analysis, but (independent of the scientific truth) this class membership is treated as “true” by the classifier. The classifier model is trained to successfully separate observations into the classes using the values of the set of predictors (or "**features**").

Applying a supervised learning classifier. To continue our example, we might use a supervised learning model to predict diagnostic status from resting-state fMRI data and cognitive testing data. Our diagnostic interview data has labels for each participant (case): a Yes/No (binary) label for each diagnosis of interest (e.g., GAD; panic disorder; social anxiety disorder; specific phobia). A training dataset is first created – with fMRI data and cognitive testing data as predictors, and diagnostic status (Yes/No for each of the four diagnoses of interest) as outcomes. Using cross-validation (discussed further below), we could test how well a trained model can classify new cases (i.e., cases who were not included in the training dataset) with respect to their clinician-assigned diagnosis. One classification approach would be to ask separate (binary; 2-way) questions for each diagnosis (i.e., GAD – Yes/No; panic disorder – Yes/No; social anxiety disorder – Yes/No; specific phobia – Yes/No). In this case, we train and test separate models for each of the four diagnoses/outcome variables. Chance performance would be 50% accuracy for predicting each outcome.

In contrast, unsupervised models take a more data-driven approach, which does not assume that certain criterion (such as an observation’s class, or even the number of classes) is known. Instead, an unsupervised model is trained to provide the best fit to the underlying structure of a dataset. This can be useful when *a priori* classification schemes (e.g., a diagnostic taxonomy) may not accurately reflect scientific reality. An experimenter might dictate certain parameters (e.g., how many clusters to form), but the clusters themselves are derived without labels corresponding to one specific criterion or outcome of interest. This is not to say that unsupervised analyses are always preferable. The class labels provided in supervised approaches are often meaningful and can provide important information to guide the identification of a solution.

<u>Applying an unsupervised learning model.</u> In our ongoing example, we might seek to detect distinct profiles of cognitive and neural functioning within our sample. We could then examine how subjects with distinct profiles differ in their diagnoses or symptoms, and ask what this means for a particular diagnosis or symptom. In this case, the examined data would include our cognitive and neural variables, but without information about diagnostic status or symptoms (i.e., without labels). An unsupervised algorithm is used to detect robust patterns (such as clusters of cognitive and neural predictors) in the training set. The testing set can then be used to evaluate the reliability of these profiles for new cases. Alternatively, we could ask if the naturally emerging profiles of neurocognitive functioning can reliable predict differences in diagnoses or symptoms in independent cases.

Study Design

Here, we provide a brief overview of study designs that are well-suited to a machine learning analytic approach. These will often vary in whether they are between-

or within-subject designs, and whether we use categorical labels (classes) or continuous values.

Between or within-subject designs. Classifying between (or across) subjects involves training and testing a model using data from different sets of subjects, where a positive result reflects the model making successful predictions about new (untrained) participants.

An underappreciated alternative is to apply machine learning to within-subjects data (e.g., trial-by-trial performance in a cognitive task). Within-subject study designs that result in many observations per person (e.g., ecological momentary assessment, eye-tracking, computerized cognitive tasks, and neuroimaging), can be submitted to machine learning approaches to obtain a measure of the extent to which trial-types (e.g., a certain class of stimuli) are distinguishable within a given participant. This metric can in turn be related to an individual difference of interest.

For example, in an fMRI study of face perception in adolescents with autism spectrum disorder, we used patterns of activity in the fusiform gyrus to classify whether each participant was viewing a face or a house on a given trial (Coutanche, Thompson-Schill, & Schultz, 2011). We then related participants' "face versus house" classifier accuracies to their social scores on the Autism Diagnostic Observation Schedule (ADOS), a clinician-administered measure of symptom severity. Here, instead of attempting to classify participants, we were interested in relating an individual difference in symptom severity to the distinctiveness of their neural codes for faces. We found that neural activity patterns for faces and non-faces were more confusable in adolescents with more severe social impairments.

Predicting group membership. If we are classifying cases, how many (and what) classes should we include? When extending the principles of machine learning to classify three or more classes, there are several considerations to bear in mind. First, with multi-way (i.e., > 2 groups) classifications, although the overall accuracy speaks to discriminability among classes, a significant accuracy value does NOT mean that all the classes can be successfully predicted. This is analogous to the case of an omnibus ANOVA, where a significant *F*-test indicates that at least two groups differ, but does not tell us which (instead, post hoc tests are used to answer this).

Applying class predictions. Previously, we asked whether our model could predict the presence or absence of each of four diagnoses (GAD; panic disorder; social anxiety disorder; specific phobia). Instead, we might ask if a model can use predictors to discriminate *between* individuals with different diagnoses (i.e., 4-way classification). For illustrative purposes, imagine that each of our 1,000 participants meets criteria for one, and only one, disorder (unlikely, but such is the power of hypotheticals). In this case, the trained model would assign each "test" case to one of four classes (GAD OR panic disorder OR social anxiety OR specific phobia). Here, chance performance is 25% when the cases are distributed evenly across diagnoses (i.e., there is a 1 in 4 chance of classifying correctly on the basis of chance alone).

Applying caution to interpretations of accuracy. When conducting multi-way classifications (i.e., discriminating more than two groups), it is important to be cognizant that some predictors could be highly specific to one (and only one) outcome class. For example, suppose our model used “frequent uncued panic attacks” as a predictor. This predictor would have high accuracy for classifying cases as “panic disorder” or “not panic disorder”, but classification could be at chance for the other comparisons. In this example, panic disorder would be classified with (close to) 100% accuracy, while GAD, social anxiety disorder, and specific phobia would each be classified with 33.3% accuracy (because they are confusable with each other, but not with panic disorder). This particular scenario would give an *overall* classification performance of 50%, twice the level of chance (25%) and typically significant (see “Outputs and Interpretation” below for a discussion of testing statistical significance). Yet, it would not be entirely accurate to conclude that this model successfully discriminates between the four anxiety disorders. Instead, we have actually trained an excellent discriminator of one disorder. This example shows the importance of moving beyond classifier accuracy alone. Here, a **confusion matrix** (discussed further below) would give us insights into the structure of our multi-way classifier.

Predicting continuous and discrete outcomes. Classifiers often seem to get the machine learning headlines. The value of continuous models, however, is increasingly being recognized. Just like classifiers, these models are trained on a subset of the data and tested on a held-out set. In a continuous case, however, model success is determined from the difference (or prediction loss) between the true (continuous) values and the model's predictions. This is a form of regression analysis (e.g., support vector regression; SVR). One example is the prediction of age based on functional connectivity in resting-state fMRI data, giving measures of individual brain maturity (Dosenbach et al., 2010).

Applying models to predict continuous outcomes. As we learn more about the underlying structure of psychopathology, transdiagnostic approaches become increasingly important. Rather than attempting to classify individuals into discrete groups, we may be interested in developing a model to predict severity of a certain type of symptom (e.g., autonomic arousal; interpersonal difficulties). Self-report measures are generally well-suited to these questions. To continue our example, we might be interested in whether a model trained on neural and cognitive data can accurately predict how much (or how little) control an individual believes they have over their worry. Alternatively, we might be interested in whether our model is more (or less) effective for predicting the severity of cognitive symptoms (e.g., unwanted thought; difficulty concentrating) versus physical symptoms (e.g., muscle tension; autonomic arousal). Combined with a feature-ablation approach (discussed below), we might develop insights about mechanisms that underlie each symptom.

Data Analysis

Once you have decided on your design and dataset, how do you analyze your data?

Selecting features. A key advantage of machine learning models is their robustness to using a large number of "features" or predictors, relative to the number of observations. A standard linear regression model is quickly overwhelmed by a large number of predictors, but machine learning models are robust to a high predictor-to-observation ratio, including where predictors outnumber observations (as can happen in, for example, fMRI studies using hundreds or even thousands of voxels as predictors of one hundred observations for each condition). As the number of features rise, so does the chance of **overfitting**. Overfitting occurs when a large number of features lead a model to closely fit the unique idiosyncrasies of the training data, at the cost of limited generalizability to new (test) data. It is for this reason that machine learning techniques are rarely applied to the whole brain's set of voxels (which can easily number 60,000) without an intermediate feature selection step.

As the name suggests, "feature selection" aims to reduce the number of predictors in the final model, while optimizing the ability to detect a signal (i.e., to remove features with a minimal impact on the quality of prediction). Entire research programs are dedicated to analytically selecting a subset of available features (Guyon and Elisseeff, 2003) but it is valuable for a psychologist applying machine learning techniques to be familiar with a number of basic approaches.

One vital rule closely governs the choice and application of feature selection techniques: training and testing sets must always be independent. This is especially important during feature selection because an accidental violation of this rule can artificially inflate a model's (apparent) performance, even using data that is just noise. Consider a dataset of completely randomly generated values (i.e., noise) – with 10,000 observations and 1,000 predictors. We could randomly split these observations into two classes that we then attempt to classify. Ordinarily, we would expect to obtain chance performance from this pure-noise dataset. But now imagine that before classification, we quantify how “well” each variable separates the observations. Even by chance, some variables will have higher values in one condition than another. If we were to then choose the top 10 most “discriminating” variables, and classify our observations using these predictors, we would obtain above-chance performance. Clearly, this is not right – the dataset is pure noise – so how did we erroneously reach this conclusion? Our main error was to choose predictors (i.e., select features) based on the full dataset, including those observations that our model will test on. It is not surprising that our model could separate the classes because we picked those features that differed between them (by chance). Here, our model’s success is closer to a re-description of how we picked the variables (“top 10 most discriminative predictors”) than reflecting successful prediction. Fortunately, we can still select features without this error.

One possible approach to selecting features is to use a difference among groups/classes that is unrelated to the one we are classifying. For example, suppose we are classifying participants into one of two psychological disorders based on their responses to a number of surveys. The surveys that are ultimately selected as features might be chosen based on unrelated dimensions, such as reliability measures, from prior work. These dimensions might also be extractable from the data itself, though great care is needed to ensure this dimension is truly orthogonal (Kriegeskorte et al., 2009).

Some analytical methods allow one to reduce the dimensionality of a dataset in a way that preserves most of the variance. For example, a principal component analysis (PCA) can be used to reduce an original set of predictors to a smaller set of components that still explain a large part of the variance. This is a popular approach to feature selection when classifying fMRI data, as often 90% of the variance in the signal of hundreds of voxels can be captured in 10-30 new dimensions. Using these components reduces the number of predictors (thus reducing the risk of overfitting) while keeping much of the variance intact.

Another approach (not mutually exclusive with the first) is using cross-validation (described further below), which by design separates training from testing. The concern with the above example of predicting noise was that our predictors were selected based on their ability to discriminate the classes using *all* the data – including the testing set that was ultimately used to evaluate the model. The cross-validation structure allows us to avoid this situation because for each cross-validation iteration, we can select features based on training data only. Even if the contrast used to select features is identical to the ultimate contrast that we care about, the part of the dataset used to evaluate its predictive potential is held-out (independent), making above-chance classification no longer inevitable. Instead, the trained model must generalize to a subset of the data that was not involved in feature selection. An important consideration with using cross-validation in this manner is that the features selected can (and often will) differ across cross-validation iterations. This might not be a concern if optimizing accuracy is a primary goal, but it could present complications when asking which features are contributing to the final mean accuracy. One way to evaluate predictors in this case is to quantify (and possibly rank) how often a feature is selected across the full set of iterations.

Another form of selecting features is inherent to certain classification algorithms. Perhaps one of the simplest examples is **regularization**. As a model is fit during training, regularization can make certain solutions less likely by penalizing certain ways in which weights are assigned to features. For example, a least absolute shrinkage and selection operator (lasso) regression classifier tends to allocate weight values more sparsely by fixing some coefficients to zero, leading to fewer predictors contributing to predictions. Alternatively, a ridge regression classifier penalizes solutions that have many large positive or negative weights across predictors, which can be helpful for avoiding a situation where only one of several correlated predictors is disproportionately relied on. When using a classifier that requires a user to specify a parameter (such as a penalty value), this must be selected using independent data (such as piloting or prior work) or from the training (but never testing) data.

Cross-validation. A fundamental principle underlying machine learning is training a model on a subset of data, and then making predictions for an independent set. The manner in which we separate our training and testing sets is thus more than a question of logistics – it is fundamental to ensuring the validity of a model's performance. How should one divide data to best ensure this? A simple approach is to divide the dataset into halves randomly, train on the first half, and then test on the second. Although this split-half approach clearly separates the data, it is not the best approach from a statistical power perspective because it does not maximize the amount of training data. Cross-validation fills this role. Through cross-validation, a full sample is split into k subsamples (or **folds**). A classifier is trained on data in all-but-one (k-1) folds and tested on the remaining (held-out) fold. This is repeated across k-iterations with the data in each fold acting as the test set in turn.

Applying cross-validation. To conduct cross-validation for our example, we might choose to randomly assign each of our 1,000 cases to one of 10 separate datasets ("folds"), each containing 100 cases. In the first training iteration, the model would be trained on folds 2 - 10 (corresponding to cases 101 - 1,000) and then tested on fold 1 (cases 1 - 100). The second iteration would train on folds 1 and 3 - 10 (observations 1 - 100 and 201 - 1,000) and test on fold 2 (observations 101 - 200), and so on. By proceeding through ten iterations, every case is included in the test set once, but always remains independent from the training set.

In addition to maintaining training/testing independence, cross-validation helps maximize the amount of data used to train a model. This is particularly valuable because there is much to gain by increasing the percentage of data for training, with relatively little cost. This is most evident when considering how any single prediction is impacted by adding training versus testing data. The addition of new observations to a training set results in a more accurate model that is capable of making better predictions of new observations. In contrast, adding or removing a testing observation typically does not affect other predictions (an exception here is if a prior dependency is created, such as through *z*-scoring a test set).

Beyond maximizing the number of folds (and therefore the percentage of training data), what factors should we consider when deciding how to assign observations to folds? The most straightforward and intuitive approach would be to randomly assign cases to folds or use a simple leave-one-subject-out approach. In certain circumstances however, a different organizational structure might be preferred. For example, the organization of folds determines which datapoints are included in training and testing for each iteration, where datapoints in the same fold are kept together (e.g., all contributing to training or testing) and datapoints in different folds will at some point be split across training and testing. It is therefore important to ensure that to-be-split data are independent. For example, *z*-scoring could be conducted across observations within a fold, but doing the same with observations that cross folds could create dependencies between training and testing (i.e., the post-*z*-scoring values of the test set would then be influenced by values in the training set). If the data is collected from multiple sites, a "leave-one-site-out" approach can be a valuable structure. Successful predictions for held-out sites (e.g., predicting treatment outcomes for first-episode psychosis with approximately 70% accuracy in Koutsouleris et al., 2016) indicate that models are sufficiently robust to differences across sites, such as sample characteristics, geography, and others.

Another consideration relates to the balance (or ratio) of different classes, such as groups (in a between-subjects design) or trial-types (in a within-subjects design). Unbalanced datasets are common in clinical research, in large part because some groups of participants are easier to recruit than others. Missing data issues and flagged participants are another source of imbalance. Imbalanced samples can lead to difficulties in interpreting classifier performance. As a result, it is good practice to strive for equal representation of the classified groups, to the extent that is possible within the constraints of the research context.

Applying caution to unequal groups. To illustrate the difficulties with having imbalanced classes, suppose that in our example dataset, 700 participants meet criteria for GAD, while only 100 each meet criteria for panic disorder, social anxiety disorder, and specific phobia. Here, a model that always predicts that a given case should be classified as "GAD" would be accurate 70% of the time. This level of accuracy might seem impressive if the researcher is not accounting for the imbalanced dataset, but in reality, it does not reflect anything meaningful about each class (beyond frequency in this dataset).

Sometimes unbalanced designs are unavoidable, however, and when this is the case, solutions can be applied at the level of assessment, selection, and refinement of classifier models. In most cases, the most elegant solution is to balance the classes in each fold. One strategy is to randomly sub-sample observations in the less prevalent class to reach an equal number of datapoints in every class. To ensure a classification performance that is representative of the full sample, this can be repeated by randomly sampling the observations that are included from the smaller class each time.

Classifier and regression models. With dozens of classifier variants available, how should we choose which type of model to use? A key distinction among alternatives is whether a model is linear or nonlinear. Linear models can learn to distinguish observations of classes that are "linearly separable". With just one dimension, a linear separation would be equivalent to using a simple cut-off value. For two dimensions, a linear method of separation is a straight line.

Applying linear models. Suppose we are interested in predicting whether an individual was diagnosed with panic disorder during a diagnostic interview. Two likely relevant dimensions would be the frequency of uncued panic attacks in the past month (dimension 1), and the extent to which the individual fears or avoids future panic attacks (dimension 2). A higher value on each dimension is associated with a greater likelihood of diagnosis, so a straight diagonal line across the dimensions is one possible way to make this prediction (as such a line can separate higher from lower values in both dimensions). This is a linear solution.

As the number of dimensions increase beyond three, it can be difficult to visualize, but whereas a straight line can separate two dimensions, a straight plane is used for three, and a straight (multidimensional) hyperplane is used for more than three dimensions. However, so far these models have assumed a linear separation. If the line/plane/hyperplane must be curved to separate observations (such as when very high and very low –but not intermediate– predictor values are associated with an outcome), a linear model cannot learn the class distinction. Nonlinear models allow classes to be distinguished in ways beyond linear models, but also have a significant cost in interpretability: successful solutions can be difficult to visualize or understand. Interestingly, relatively few neuroimaging investigations have reported an advantage from using nonlinear, compared to linear, models (e.g., Kamitani & Tong, 2005), so a linear classifier such as a support vector machine (SVM) is a common choice.

In deciding on the particular model used, the characteristics of the predictors and outcome variable are often important. If there are no unique properties of the employed predictors, a powerful linear classifier that is robust to overfitting, such as a SVM, might be recommended. This particular classifier can be trained relatively quickly and copes well with a large number of features relative to observations. If there is reason to believe that the predicting features explain very similar parts of the variance, a ridge regression classifier (discussed above) that is penalized for particularly extreme weights might be preferred. Some classifiers (such as SVMs) are binary (2-way only) though they can be assembled for multi-way classifications (Bishop, 2006), whereas others (such as Gaussian Naïve Bayes) are intrinsically multi-class. Perhaps the most significant property of a predicted variable is whether it has a categorical or continuous structure. Although classifiers of categorical outcomes are the most frequently employed form of machine learning, alternative (continuous) models, such as SVR, can predict variables where differences between values are meaningful (e.g., symptom severity, rather than categorical diagnoses). In this case, accuracy is not determined from correctly predicting a class, but from the distance between predicted and actual values.

Outputs and interpretation

The first output typically examined by an analyst is the accuracy of their model's predictions of the testing set. For a classifier, the model predicts the testing set's labels – whether an observation belongs to a particular class. Classification accuracy simply reflects the percentage of classifier predictions that are correct. Studies using cross-validation typically report the mean accuracy from all iterations, so that the overall accuracy reflects predictions made for every observation (as each observation would be held-out in the test set once). As with accuracy measured from a human behavioral experiment, a model's accuracy values can be submitted to various signal detection analyses. Measures such as d’, area under the curve (AUC), and others, are all applicable depending on the precise question being asked about a model's performance (for a primer on signal detection theory, see McNicol, 2005). For a continuous model, such as SVR, predictions are continuous and in the same scale as the outcome variable. Here, accuracy could be quantified in several different ways, including mean squared error, loss, or a correlation between the true and predicted values.

For machine learning analyses conducted within subjects, each participant receives their own accuracy value, giving a set of values for the whole group that can be tested against chance (such as through an ANOVA). An example of this would be predicting the emotional expression of different observed faces, based on an observer’s trial-by-trial eye-tracking. The ability to classify observed faces might differ based on the observer’s diagnostic status – an ANOVA that compares groups of individuals could be used to test this possibility.

When subjects are being predicted instead of trials (i.e., across-subject classification), we typically have a single overall accuracy value (e.g., 65% for predicting the group that subjects belong to). One can test the significance of this value through resampling techniques. In **permutation testing**, the labels of the observations are scrambled, then the same machine learning analysis (such as classification) is re-conducted using the misaligned labels. This is repeated thousands of times with a new label order each time. The resulting set of accuracy values can act as a null distribution of the outcome variable (i.e., a set of values that would be expected if the labels are actually uninformative). The accuracy obtained from the true analysis (with non-scrambled labels) is compared to this distribution to calculate a *p*-value (e.g., by asking if the true value falls within the top or bottom 2.5% of the distribution for a two-tailed test at $p < 0.05$), as well as confidence intervals. There are several considerations in deciding how to permute labels, such as ensuring that resampled values are "exchangeable" and that a sufficient number of permutations is run. Bootstrapping is a related technique that is commonly used to establish confidence intervals (discussed in Efron and Tibshirani, 1993).

The above output is typically only a first step. Multi-way classifications in particular can give above-chance performance due to a variety of possible combinations of discriminability and confusion between classes. One way to better understand a trained model's solution is to examine the confusion matrix. A confusion matrix breaks down a model's overall accuracy to reveal the kinds of errors being made. The confusion matrix organizes the predictions into their (true) classes, and displays the proportions of guesses (Figure 1). This allows the reader to examine which classes are more likely to be confused (hence “confusion matrix”). The proportion of guesses made in each class is often represented by a color scheme, which quickly communicates the pattern of easily confused classes

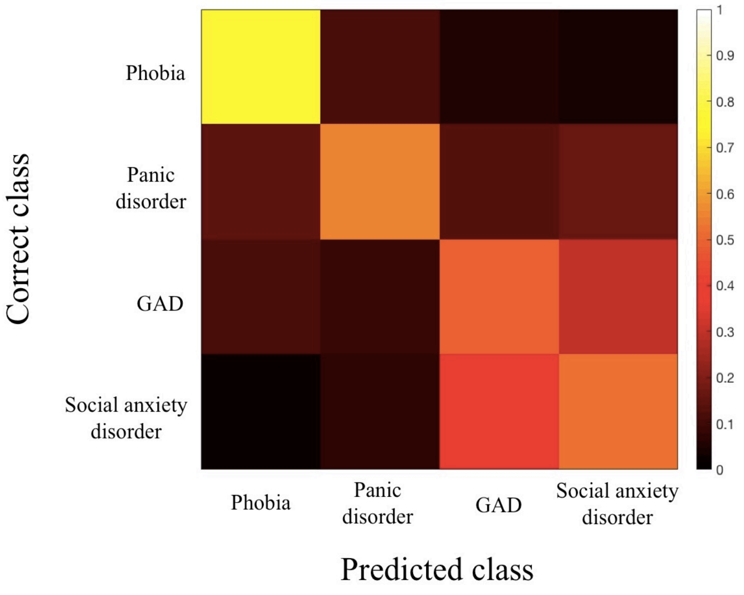

Figure 1: An example confusion matrix. The color-scale reflects proportions of class predictions, organized by the actual (correct) class. This hypothetical example is a classification of participants into one of four diagnoses (see box for discussion).

(e.g., third and fourth class in Figure 1). Because the diagonal represents correct guesses (e.g., when the observation is in the first class, correctly guessing "class one"), the diagonal of the matrix will often have higher values than off-diagonal parts if a classifier is successful.

Applying a confusion matrix to multi-way classification results. Imagine that we conduct a 4-way classification of cases into one of four diagnoses – GAD, panic disorder, specific phobia, or social anxiety disorder – based on validated self-report measures of perfectionism, depressive symptom severity, interpersonal concerns, and cued and uncued panic attacks. We obtain impressive classification accuracy (where chance is 25%) but wish to know how well or poorly our model discriminates different diagnoses from each other. A confusion matrix (Figure 1) shows the proportion of guesses made for each class. In our example, GAD and social anxiety disorder are more confusable by the model, suggesting greater similarity in their predictor profiles.

Understanding which classes are more or less confusable can be invaluable in understanding how a classifier is performing, and how classes relate to one another. The confusion matrix from a classifier can itself be statistically compared to a *behavioral* confusion matrix. For example, a researcher might be interested in how their model's successes and failures correspond to rates of misdiagnosis among clinicians. A confusion matrix of misdiagnosis can be created using relevant rates and then statistically compared (e.g., through a correlation) to test for similarity in how the disorders are distinguished or confused.

As well as examining overall performance, we might be interested in how features are contributing to a model's predictive power. The trained model's set of weights (or an "importance map") reflects how the model is using each predictor, which can give insights into the nature of the model. Caution is needed, however, in how the set of weights is interpreted. First, the weight given to a feature reflects its contribution as part of a pattern; a set of variables might each have a small weight, but as a set, they might still have substantial predictive power. Second, the model's weights reflect how this particular model was formed, but it is possible that similar prediction accuracies could be obtained with different sets of weights (i.e., the allocated weights are sufficient, but might not be necessary). Third, in cross-validation, a number (k) of models are being trained (one per iteration), giving one set of weights for every iteration. A common way to deal with this is to average the weight maps across iterations, or to train a model on the entire dataset purely for the purpose of inspecting the associated weights. It is important to note, however, that strictly speaking, we cannot then say that weights from a model fit to the entire dataset yielded the generated classification accuracy, since accuracy did not come from a model trained on all the data. Fourth, when a nonlinear model is used, the weight matrix cannot be easily understood.

In addition to visualizing the weights, it is possible to visualize the product of the weights with the corresponding inputs (e.g., Polyn et al., 2005), which can be easier to relate to the sign of the input data. For example, with a weight matrix alone, a negative weight can seem easy to interpret (lower values are more predictive of a class) but when the input data is itself negative for a particular class, this can become difficult (e.g., a negative predicting variable with a negative weight is actually a positive contribution of that variable). The weight-input product reflects both the input and the weight, removing this confusion. If a researcher is interested in the necessity of a given variable for successful prediction, or wishes to compare the relative predictive power of individual predictors, one elegant approach is ablating a predictor from the dataset, repeating the training and testing procedure, and comparing performance using this reduced model to performance with the full model. The prior consideration –that individual variables might not be particularly predictive, but can still contribute to an informative set– is still an important factor when interpreting the results of the ablation approach. Gotsopoulos and colleagues (2018) discuss these and other approaches for the classification of fMRI data (Gotsopoulos et al., 2018).

Identifying clusters on the basis of latent structure

Unsupervised learning approaches aim to discover latent structures in a dataset, without explicit guidance such as a set of labels. Instead, clusters or factors are discovered based on the predicting variables, in a training portion of the dataset. This structure can then be validated in a testing set, allowing a researcher to ask whether the discovered structure is robust, as well as giving the opportunity to test whether the structure conforms to one or more hypotheses.

Two of the most popular clustering techniques are K-mean clustering, and hierarchical clustering. In K-mean clustering, datapoints are assigned to one of K clusters based on distance in multidimensional space. Multiple solutions (differing numbers of clusters, K) can often accurately reflect a dataset, where latent structures are present at differing degrees of coarseness / specificity. The number of clusters (K) can be specified according to an existing theory, or from the data. One popular data-driven way of selecting K is through the “elbow method”, where the percentage of variance explained is plotted against the number of clusters. Adding a cluster has diminishing returns on the variance explained, so the analyst can look for the point at which the explained variance decreases significantly less than it has before (i.e., the plot resembles an elbow-shape) to select K.

Hierarchical clustering is another approach. This technique allocates each datapoint to its own cluster, before merging nearby clusters until they form a hierarchical branching tree-structure. This approach can reveal multiple degrees of specificity for a given clustered dataset (at different levels of the hierarchy), allowing one to observe multiple structures within a dataset. An example of this being applied is in Chekroud et al. (2017), who used hierarchical clustering to group baseline symptoms in patients with depression using data from the Quick Inventory of Depressive Symptomatology (QIDS-SR) and clinician-rated Hamilton Depression (HAM-D) rating scale. Three clusters (“Core emotional”, “sleep (insomnia)”, “atypical”) were identified, which were differentially responsive to antidepressant treatments. In another study, Drysdale et al. (2017) applied hierarchical clustering to resting-state fMRI connectivity data in a large multi-site sample of patients with depression. The researchers identified four subtypes based on resting-state connectivity values. The dissociated subtypes had differing clinical symptom profiles and varied in their responsiveness to repetitive transcranial magnetic stimulation therapy.

Common pitfalls

There are a number of important considerations for psychology researchers seeking to apply machine learning techniques to their data. First, although the prediction framework gives confidence in a model's validity in new data, we are still always constrained by the variability (or lack thereof) that is present in the data. Characteristics such as the demographics of our sample, the noise structure of data from a particular device (e.g., MRI scanner), particular nuances of how a clinical team verifies diagnoses, and more, will often affect data quality. It often requires an explicit effort, resources, and often collaborations, to apply trained models to applicable datasets that have been collected by different teams, with different samples. For example, "leave-one-site-out" cross-validation often produces lower accuracies than leave-one-subject-out, indicative of the effect that such characteristics have on model performance. One way to increase the ability of a model to generalize is to expose it to observations with broad variance (i.e., recruiting a sample that is diverse across relevant characteristics). A model trained with such variance will be more robust at predicting the signal in new data (for a similar principle in neuroimaging, see Coutanche & Thompson-Schill, 2012).

With computational power improving every year, and increasing access to computing clusters, analyses that once took hours can now take minutes or even seconds. Nonetheless, a researcher can still find that a planned analysis requires days to complete. In addition to "housekeeping" to make a code more efficient, it can be important to consider the influence that certain analysis choices (such as leave-one-subject-out classification) have on the time required. Combining a 1000-fold classification (for 1,000 subjects) with techniques that measure the impact of excluding features (feature ablation, discussed above), or with nested cross-validation (in which another cross-validation is run inside the training data, usually to select optimal classification parameters), can quickly compound the required computation time. In such cases, changing the fold structure to, for example, leave-10-subjects-out is a quick way to reduce computation time (as it reduces the number of models to be trained by a factor of 10), while preserving independence between training and testing.

Looking ahead

A number of recent analyses have focused on relating treatment outcomes to underlying dimensions rather than diagnostic status alone. For example, Webb and colleagues (2018) examined endophenotypes in Major Depressive Disorder to predict responsiveness to sertraline versus placebo. Better outcomes were predicted by greater depression severity, higher neuroticism, older age, less impairment in cognitive control, and employment. Successful prediction in this case serves the dual purpose of identifying how patients might be assigned to treatments and giving insights into dimensions that underlie depression and treatment-induced change. Relatedly, efforts to recruit for a diverse array of symptoms (e.g., Astle et al., 2018) show promise for developing models that apply to the real heterogeneity present within the population, and avoid artificially inflating reliability within examined groups.

Another promising direction is to combine information across modalities in the same model to improve predictive power. The diagnostic signal present in behavioral self-report, clinicians' reports, fMRI data, genotypes, and more, are very likely to explain distinct portions of individual variance. Bringing these diverse predictors into the same model presents both a challenge for interpretation and an opportunity for detecting meaningful differences across individuals.

The translation of findings from the laboratory to medical practice has its own set of challenges. As models are developed, the question of practicality (in terms of relative cost, efficiency, etc.) becomes important (Cohen & DeRubeis, 2018; Gillan & Whelan, 2017). The precise role taken by machine-informed decisions on treatment or diagnosis has yet to be determined. Rather than replacing clinician judgment altogether, any generalizable and successful model is more likely to act as a supplement to clinical intuitions and experience (Linden, 2012). The degree of influence of any such model is likely to be greater for particular questions or decisions that are frequently difficult for today's clinicians to determine alone, such as the identification of Unipolar versus Bipolar Depression (e.g., Perlis et al., 2006). If particular models are employed in clinical practice, it will also be important to maintain a suitably open line of communication between commercial entities and independent scientists regarding the precise basis for any treatment recommendations. For example, patents have been filed in relation to predicting treatment response in depression using machine learning techniques (e.g., Chekroud et al., 2017, Patent WO2017210502A1). Likewise, improving treatment predictions depends on an effective relationship between data scientists and commercial entities involved in dissemination.

Summary

The recent application of machine learning techniques in clinical psychology has presented exciting opportunities and a daunting training challenge for psychology researchers. Fortunately, the number of free learning resources available to new machine learning investigators grows every day (see Further Resources below for some jumping-off points). We hope that the research and approaches reviewed above illustrate that – more than simply providing an increase in sensitivity – machine learning approaches provide an opportunity to select treatments, refine diagnoses, understand cognitive and neural bases for symptoms and clinical dimensions, and more. In this sense, the most powerful tool in the investigator's box remains her/his creativity and willingness to apply these techniques to problems in new ways. As more clinical investigators master the approaches we have discussed, new opportunities will arise to combine their content expertise with machine learning, leading to research studies and applications that are truly unique and groundbreaking. We look forward to enjoying your work.

Further Resources

The modern machine learning researcher has a plethora of resources to choose from. You should choose a resource that bests suits your own learning preferences, while operating at the right level of theory and practical application. There is no substitute for getting your hands on data, so we recommend applying what you learn, even with a sample dataset, whenever possible. Below we have listed some compilations, along with some of our personal favorites

Compilations of resources and online guides

- Machine Learning for Humans

  A series and free e-book that provides simple, plain-English explanations accompanied by math, code, and real-world examples: https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12
- My Curated List of AI and Machine Learning Resources from Around the Web

  An enormous list of noteworthy machine learning blogs, video courses, code, forum discussions and more: https://medium.com/machine-learning-in-practice/my-curated-list-of-ai-and-machine-learning-resources-from-around-the-web-9a97823b8524

Books

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R* (1st edition). New York: Springer.
- Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data* (1st edition). Cambridge; New York: Cambridge University Press.

Software and packages

- R - https://www.r-project.org/
- Python - https://www.python.org/
- The Python scikit learn machine learning toolbox - https://scikit-learn.org/stable/
- MATLAB - https://www.mathworks.com/products/matlab.html
- Octave - https://www.gnu.org/software/octave/

Neuroimaging-specific resources

- The Princeton MVPA MATLAB toolbox - https://github.com/PrincetonUniversity/princeton-mvpa-toolbox
- PyMVPA - http://www.pymvpa.org/
- Nilearn - https://nilearn.github.io/
- Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109.
- Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*(1 Suppl), S199-209.

Bibliography

Anzellotti, S., & Coutanche, M. N. (2018). Beyond Functional Connectivity: Investigating Networks of Multivariate Representations. *Trends in Cognitive Sciences*, *22*(3), 258–269.

Astle, D. E., Bathelt, J., CALM Team, & Holmes, J. (2018). Remapping the cognitive and neural profiles of children who struggle at school. *Developmental Science*, e12747.

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., & Craddock, R. C. (2017). The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage*, *144*(Pt B), 275–286.

Chekroud, A., Krystal, J. H., Gueorguiva, R., & Chandra, A. (2017). *WO2017210502A1*. World Intellectual Property Organization. Retrieved from https://patents.google.com/patent/WO2017210502A1/en

Chekroud, Adam M., Gueorguieva, R., Krumholz, H. M., Trivedi, M. H., Krystal, J. H., & McCarthy, G. (2017). Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach. *JAMA Psychiatry*, *74*(4), 370–378.

Chekroud, Adam Mourad, Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., … Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet. Psychiatry*, *3*(3), 243–250.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology*, *14*, 209–236.

Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: Why, how, and what does it tell us? *Cognitive, Affective & Behavioral Neuroscience*, *13*(3), 667–673.

Coutanche, M. N., & Thompson-Schill, S. L. (2012). The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. *NeuroImage*, *61*(4), 1113–1119.

Coutanche, M. N., Thompson-Schill, S. L., & Schultz, R. T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage*, *57*(1), 113–123.

Deshpande, G., Libero, L. E., Sreenivasan, K. R., Deshpande, H. D., & Kana, R. K. (2013). Identification of neural connectivity signatures of autism using machine learning. *Frontiers in Human Neuroscience*, *7*, 670.

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., … Schlaggar, B. L. (2010). Prediction of individual brain maturity using fMRI. *Science (New York, N.Y.)*, *329*(5997), 1358–1361.

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., … Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, *23*(1), 28–38.

Du, W., Calhoun, V. D., Li, H., Ma, S., Eichele, T., Kiehl, K. A., … Adali, T. (2012). High classification accuracy for schizophrenia with rest and task FMRI data. *Frontiers in Human Neuroscience*, *6*, 145.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap* (1 edition). New York: Chapman and Hall/CRC.

Eldridge, J., Lane, A. E., Belkin, M., & Dennis, S. (2014). Robust features for the automatic identification of autism spectrum disorder in children. *Journal of Neurodevelopmental Disorders*, *6*(1), 12.

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., … Caffo, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, *6*.

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, *18*, 34–42.

Gotsopoulos, A., Saarimäki, H., Glerean, E., Jääskeläinen, I. P., Sams, M., Nummenmaa, L., & Lampinen, J. (2018). Reproducibility of importance extraction methods in neural network based fMRI classification. *NeuroImage*, *181*, 44–54.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hallion, L.S., Wright, A.G.C., Coutanche, M.N., Joormann, J., and Kusmierski, S.N. (submitted). An empirically-derived taxonomy of perseverative thought: Evidence for a dimensional approach to classification.

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

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

Kohonen, T. (1989). *Self-Organization and Associative Memory* (3rd ed.). Berlin Heidelberg: Springer-Verlag.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., … Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry*, *3*(10), 935–946.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neurosci*, *12*(5), 535–540.

Linden, D. E. J. (2012). The Challenges and Promise of Neuroimaging in Psychiatry. *Neuron*, *73*(1), 8–22.

McNicol, D. (2005). *A Primer of Signal Detection Theory*. Psychology Press.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking Rumination. *Perspectives on Psychological Science*, *3*(5), 400–424.

Nouretdinov, I., Costafreda, S. G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., & Fu, C. H. Y. (2011). Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. *NeuroImage*, *56*(2), 809–813.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251).

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*(2), 109–118.

Perlis, R. H., Brown, E., Baker, R. W., & Nierenberg, A. A. (2006). Clinical features of bipolar depression versus major depressive disorder in large multicenter trials. *The American Journal of Psychiatry*, *163*(2), 225–231.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966.

Silva, R.F. et al. The tenth annual MLSP competition: schizophrenia classification challenge. IEEE Int. Workshop Mach. Learn. Signal Process. 1–6 (2014).

Sundermann, B., Herr, D., Schwindt, W., & Pfleiderer, B. (2014). Multivariate classification of blood oxygen level-dependent FMRI data with diagnostic intention: a clinical perspective. *AJNR. American Journal of Neuroradiology*, *35*(5), 848–855.

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology*, *63*, 483–509.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin*, *134*(2), 163–206.

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., … Pizzagalli, D. A. (2018). Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychological Medicine*, 1–10.

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, *20*(3), 365–377.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Glossary

Confusion matrix: A matrix that presents a classifier's guesses, organized by each (true) class. This allows us to examine the classes that a classifier is more or less likely to misclassify (i.e., confuse). Often visualized with colors reflecting the number of guesses given for each class.

Cross-validation: A process by which a model is trained on all-but-one (k-1) subsamples of a dataset ("folds") and then tested on the subsample held-out from training. Iterations of training and testing are conducted to hold-out each fold in turn (see: holding-out; fold).

Feature: A predictor that is used as part of a machine learning model. Each observation will typically have values on a large set of features.

Feature ablation: A method for quantifying the unique predictive contribution of individual or subsets of predictors ("features") to a model's success by systematically removing feature(s) and measuring the subsequent drop in model accuracy (see: Feature).

Fingerprint: The weighted pattern of predictors that is obtained during the training phase for a given model.

Fold: A subsample of a dataset that is held-out for testing during cross-validation (see: cross-validation).

Generalization: The accuracy with which a machine learning model that is trained on one dataset can predict patterns and relationships in a new (untrained) dataset.

Holding-out: Excluding one or more cases from a training set for testing a model.

***k*-means clustering:** A machine learning approach that groups data into groups (clusters) based on similarity in the values of their features. The number of clusters (*k*) is determined by the analyst (see: unsupervised learning).

Learning: A process by which an algorithm identifies patterns or relationships that are present in a given dataset (see: training).

Machine learning classifier: A computational model that is trained to separate data (such as observations) based on labels that are typically assigned by the analyst (such as diagnostic status). The model then gives class (label) predictions for separate data based on the patterns it was able to extract from the trained data.

Overfitting: The act of a model being trained so precisely to the training dataset that it becomes less able (or even unable) to generalize to independent test data (which will almost never have exactly the same values as a training set). Often caused by having too many features relative to observations (see feature; regularization).

Permutation testing: A statistical tool allowing one to compare a result such as classification accuracy to a null distribution generated by repeatedly running an analysis with shuffled labels each time. Comparing the true accuracy to this distribution can give a *p*-value, confidence intervals, and other useful properties.

Precision medicine: The broad effort to match patients to treatments on the basis of personal or contextual factors.

Regularization: A technique for reducing the chance of overfitting a model to the training data by constraining the types of solutions that are learned (through how weights are allocated to features).

Supervised learning: A type of machine learning model for which a label (or other information; i.e., “correct” answer) is assigned to each class or observation in a training set.

Testing: The process by which a model developed on one dataset (or part of a dataset) is applied to a new (untrained) dataset in order to assess the validity or generalizability of the model (see: cross-validation).

Training: The process by which a model is exposed to a dataset and identifies patterns or relationships within this data (see: learning)

Unsupervised learning: A process in which a model is trained on data without labels (i.e., the analyst does not teach the model the “right answer”).