European Journal of Personality, *Eur. J. Pers.* (2020)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2257

Personality Research and Assessment in the Era of Machine Learning

CLEMENS STACHL[1]*, FLORIAN PARGENT[2], SVEN HILBERT[3], GABRIELLA M. HARARI[1], RAMONA SCHOEDEL[2], SUMER VAID[1], SAMUEL D. GOSLING[4,5] and MARKUS BÜHNER[2]

[1]*Department of Communication, Stanford University, CA USA*
[2]*Department of Psychology, Psychological Methods and Assessment, Ludwig-Maximilians-Universität München, Germany*
[3]*Faculty of Psychology, Educational Science and Sport Science, University of Regensburg, Germany*
[4]*Department of Psychology, University of Texas at Austin, TX USA*
[5]*Melbourne School of Psychological Sciences, University of Melbourne, Australia*

Abstract: The increasing availability of high-dimensional, fine-grained data about human behaviour, gathered from mobile sensing studies and in the form of digital footprints, is poised to drastically alter the way personality psychologists perform research and undertake personality assessment. These new kinds and quantities of data raise important questions about how to analyse the data and interpret the results appropriately. Machine learning models are well suited to these kinds of data, allowing researchers to model highly complex relationships and to evaluate the generalizability and robustness of their results using resampling methods. The correct usage of machine learning models requires specialized methodological training that considers issues specific to this type of modelling. Here, we first provide a brief overview of past studies using machine learning in personality psychology. Second, we illustrate the main challenges that researchers face when building, interpreting, and validating machine learning models. Third, we discuss the evaluation of personality scales, derived using machine learning methods. Fourth, we highlight some key issues that arise from the use of latent variables in the modelling process. We conclude with an outlook on the future role of machine learning models in personality research and assessment. © 2020 The Authors. European Journal of Personality published by John Wiley & Sons Ltd on behalf of European Association of Personality Psychology

Key words: assessment; interpretability; machine learning; overfitting; personality

Over the past decade, a number of technological developments have allowed researchers to devise a range of new methods for collecting data in personality science. In particular, advances in consumer electronics (e.g. smartphones and wearables) and the subsequent development of mobile sensing methods (see Harari et al.,) have facilitated the longitudinal *in vivo* collection of highly detailed multidimensional data on behaviours and situations (Harari et al., 2016, 2018; Miller, 2012). In addition, behavioural residue harvested from websites and online social media platforms has proven to be a valuable source of data on behaviour linked to personality traits (Gosling & Mason, 2015; Wilson et al., 2012).

Alongside these advances in the collection and availability of such data, progress has also been made in the analytic methods that can be used to model these complex data. In particular, a multitude of new algorithms are available that use existing data to make predictions about new unseen data, to discover patterns, or to find groups of similar cases. This process is often referred to as *machine learning* (ML), which is just an umbrella term for a heterogeneous scientific field consisting of specialized sub-divisions like *predictive modelling*, *statistical learning* or *supervised learning*, *recommender systems*, and *unsupervised learning*. ML has catalysed the development of an astonishing array of technological advancements across many research fields, ranging from computer vision (Krizhevsky et al., 2017) and natural language processing (Devlin et al., 2018) to the prediction of acute kidney injury (Tomašev et al., 2019). Those breakthroughs have directly fuelled many practical applications such as autonomous vehicles (e.g. Waymo), personalization and recommender systems (e.g. Spotify and Netflix), and live translation of languages (e.g. DeepL and Google Translate). It is becoming increasingly clear that ML also has the potential to transform research and assessment in personality psychology. Many ML algorithms hold the advantage over classical approaches that they can handle vast datasets, including thousands of predictor variables, without succumbing to collinearity issues and violations of model assumptions. Moreover, it is possible for ML algorithms (when trained correctly) to recognize patterns in datasets of which humans are unaware and cannot even perceive. In the best-case scenario, the use of ML methods could lead to better, more objective, and automated personality assessments. However, past experiences have demonstrated that

*Correspondence to: Clemens Stachl, Department of Communication, Stanford University, Stanford, CA 94305, USA.
Email: stachl@stanford.edu
Stachl and Pargent contributed equally to this manuscript.

Received 14 August 2019
Revised 14 March 2020, Accepted 14 March 2020

much can go wrong in the application of ML when used for profiling or characterizing individuals (e.g. Grothoff & Porup, 2016).[1]

Thus, to effectively and safely use ML, researchers must first understand the basic principles of these methods. Here, we first provide a brief overview on how ML methods are currently used in personality psychology. Second, we discuss the main challenges that researchers face when using ML models in personality psychology. In particular, we emphasize important mechanisms that need to be understood to adequately build, interpret, and validate these methods and to critically evaluate the work of others. Most of these challenges are familiar to statisticians and ML engineers, yet are rarely addressed in articles targeted at applied researchers.[2]Third, we discuss the evaluation of personality scales, derived using ML methods, with regard to validity, reliability, and generalizability. Fourth, we highlight some key issues that arise from the use of latent variables in ML. Finally, we provide an outlook on the future use of ML methods in personality research and assessment.

MACHINE LEARNING IN PERSONALITY PSYCHOLOGY

Machine learning has been used in the private sector (e.g. to predict credit default) and in other disciplines (e.g. engineering) for many years, but applications in psychology are still rare. To date, just a handful of studies have used ML methods in the analysis of personality-relevant data, primarily focusing on the prediction of personality traits from different types of digital behavioural records (for a review, see Bleidorn & Hopwood, 2018). Recent reviews provide summaries of these and similar studies (Azucar et al., 2018; Settanni et al., 2018), so here we provide only a brief overview of the literature. Essentially, the research using ML models in personality falls into one of three categories, which we summarize as follows.

First, ML models have been used to predict individuals' Big Five personality traits from a wide range of data sources; these sources include digital footprints from social media platforms (e.g. Facebook Likes and status updates, Kosinski et al., 2013; Youyou et al., 2015), language samples (Park et al., 2015; Schwartz et al., 2013), spending records (Gladstone et al., 2019), music preferences (Nave et al., 2018), and mobile sensing data (Chittaranjan et al., 2013; De Montjoye et al., 2013; Hoppe et al., 2018; Mønsted et al., 2018; Schoedel et al., 2018; Stachl et al., 2019; W. Wang et al., 2018). More recently, researchers have started to apply unsupervised ML methods to identify other psychological constructs in digital data (Eichstaedt et al., 2018; Eisenberg et al., 2019; Schoedel et al.,).

Second, ML methods have been used to address methodological questions. For example, some studies have compared the relative effectiveness of using aggregated scale scores versus item-level data to predict life outcomes (Seeboth & Mõttus, 2018; Zweck et al., 2019), task performance, and self-report data (Eisenberg et al., 2019).

The third area in which ML approaches have been applied to personality data is the personalization of products and services through recommender systems. Personalization refers to the usage of information about the users of a system to adapt the functionalities or characteristics of the product or service to achieve a certain goal (e.g. Tkalcic et al., 2016, product recommendations on Amazon to facilitate purchase decisions). These adaptations are based on either the similarity of the user or objects to other users and objects (e.g. suggesting products based on similar products or based on purchases of users who also bought that product) or on predictive models (Aggarwal, 2016). A major motivation behind personalization is to reduce the amount of information with which a user is confronted by providing stimuli that are more suitable to the user's individual needs and interests (e.g. automatically rank movies by personal preference).

Personalization can be used to improve the usability and attractiveness of a product, a service, or a message, resulting in increased usage, higher satisfaction, loyalty, and acceptance. For example, the personalization of online advertisement campaigns can lead to more revenue and click-through rates (Boerman et al., 2017). The basic argument for the use of personality in recommender systems is that personality traits are known to be closely associated with individual differences in behaviour (e.g. Harari et al., 2019; Jackson et al., 2010; Stachl et al., 2019) and preferences (Nave et al., 2018; Randler et al., 2017; Youyou et al., 2015). Hence, adapting systems to user personality is an intuitive way to increase a system's attractiveness. Personality-based adaptions can be used to provide personalized visualizations (Schneider et al., 2017), to suggest music (Hu & Pu, 2010), and even to change the overall diversity of a recommender system itself (Wu et al., 2013). Most impressively, personality-based targeting has been shown to increase the effectiveness of marketing campaigns, leading to higher sales for personality-congruent advertisements (Matz et al., 2017).

Such personalization-based recommender systems have recently gained popularity as a result of the success of the efforts described earlier to predict personality from digital footprints (Settanni et al., 2018; Youyou et al., 2015), text (Park et al., 2015; Schwartz et al., 2013), and mobile sensing data (Stachl et al., 2019). It is valuable to compute users' personality scores because recommender systems often suffer from a lack of valid constructs on which to base their recommendations. Personality traits could solve this 'cold start' problem by using scientifically validated, relatively stable latent dimensions of individual differences as the basis of personalization systems (Hu & Pu, 2011). Comprehensive reviews of personality-based personalization and recommender systems can be found in Aggarwal (2016), Tkalcic et al. (2016), and

[1]This report describes how questionable training data were used in a random forest model to label Pakistani citizens as possible targets for drone strikes, directed at terrorists. We discuss this topic in detail in the section on the fairness of machine learning models.

[2]A detailed description of how ML models work is beyond the scope of this article, so we point readers interested in learning more to the excellent introductory (James et al., 2013; Yarkoni & Westfall, 2017) and advanced resources on the topic (Efron & Hastie, 2016; Hastie et al., 2009). For a detailed treatment of construct validity in the context of ML, we point readers to Bleidorn and Hopwood (2018).

Völkel et al. (2019). Personality psychologists are well placed to contribute to this active area of research.

In addition to the three domains of research noted earlier, a number of patterns can be discerned in the literature. One pattern concerns differences in the methods used by different disciplines; most psychological studies have used regularized linear regression models (e.g. LASSO) in their analyses (Eisenberg et al., 2019; Kosinski et al., 2013; Park et al., 2015; Schoedel et al., 2018; Schwartz et al., 2013; Settanni et al., 2018; Youyou et al., 2015), but research in computer science has tended to use more flexible, non-linear algorithms (Chittaranjan et al., 2013; De Montjoye et al., 2013; Mønsted et al., 2018; W. Wang et al., 2018).

Another striking pattern in the literature is the lack of connection between the work being performed in computer science and that being performed in psychology. In general, researchers in computer science and human–computer interaction were quicker than those in psychology to apply ML methods to the task of personality prediction. In fact, automated personality detection has emerged as its own separate field (Majumder et al., 2017; Mehta et al., 2020), in which the psychological literature is only cited with respect to the personality inventories that are used as target variables. As noted by Mønsted et al. (2018), some work coming out of this new field suffers from small, unrepresentative samples and questionable modelling practices (e.g. model overfitting, Chittaranjan et al., 2013; De Montjoye et al., 2013; W. Wang et al., 2018), which can undermine the validity and generalizability of their models.

To avoid a repetition of these problems in personality psychology, we use the present article to call for improvements in the training, application, reporting, and review of ML methods. To that end, we present a series of points to consider in setting best practices for the application of ML methods in personality psychology. To illustrate these points, we ground our discussion in examples taken from our own work on personality sensing (see Harari et al.,). Specifically, we draw on a study that applied ML methods to predict self-reported personality traits on the basis of mobile sensing data collected from smartphones (Stachl et al., 2019). This application is an example of *supervised machine learning* (also called *predictive modelling* or *statistical learning*), in which a statistical model is estimated (*trained*) to predict a criterion variable (*target*) based on several predictor variables (*features*). We focus on supervised ML, because it is currently the dominant method being used in psychological applications. Thus, the terms ML and predictive modelling will be used interchangeably throughout the remainder of this article. Figure 1 outlines central steps when performing a prototypical supervised machine learning study in personality psychology; the flowchart serves as a visual guide to how all topics discussed throughout the paper fit into the process of building, evaluating, and interpreting ML models in practice.

OPEN DATA, MATERIALS, AND CODE

Throughout the paper, we use data from the PhoneStudy mobile sensing dataset (Stachl et al., 2019). The dataset includes self-reported Big Five personality scores aggregated at the domain (5) and facet (30) levels, 1859 variables tapping real-world behaviour, and demographics (i.e. gender, age, and education). The Big Five personality traits were measured with the Big Five Structure Inventory (BFSI, Arendasy, 2009). The behavioural variables consist of a wide range of aggregated measures, obtained from smartphone sensing in the wild (e.g. calling behaviour and app usage). More information about this dataset can be found in Stachl et al. (2019). All data, materials, and code are available in the project's repository at https://osf.io/j9yrw/. To demonstrate the ML tools, we use packages from the extensive *mlr* universe in R (e.g. Binder, 2018; Bischl et al., 2016; Casalicchio et al., 2019; Molnar et al., 2018).

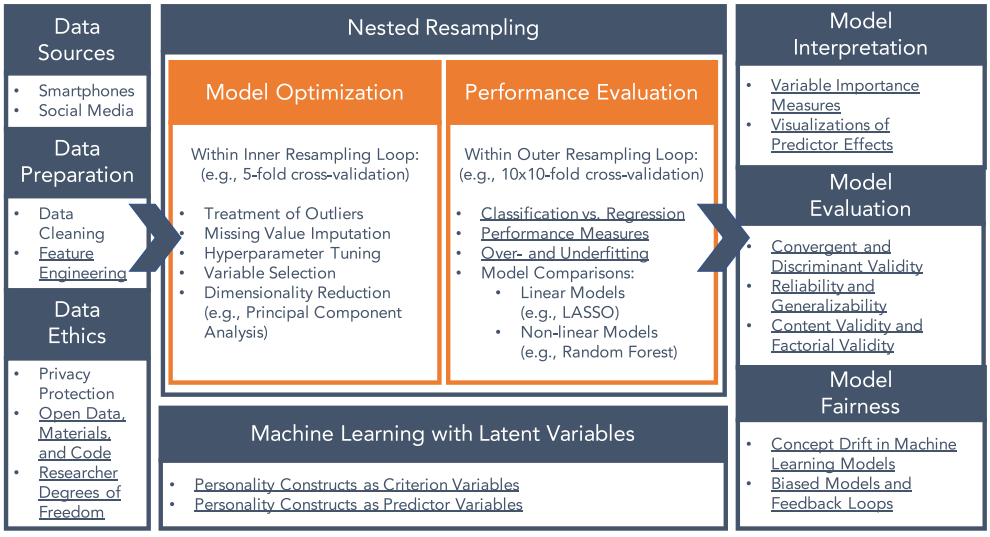

Figure 1. Schematic illustration of central steps in a prototypical supervised machine learning study in personality psychology. Underlined points are discussed in specific sections in this paper. Figure available at https://osf.io/j9yrw/, under a CC-BY4.0 license. [Colour figure can be viewed at wileyonlinelibrary.com]

BUILDING SUPERVISED MACHINE LEARNING MODELS

Feature engineering

After data cleaning, one of the most important and most difficult steps in building ML models is constructing the predictor variables. These *features* are mostly summary statistics that capture different aspects of certain variables like central tendency or variation (e.g. mean number incoming calls per day and entropy in app usage) and can be used in models to predict outcome variables (e.g. extraversion levels). Constructing features is necessary, either because most models cannot work on the raw data (e.g. time stamped event data) or because the raw data are not meaningful and lead to inferior predictive performance (e.g. individual smartphone apps are not used frequently enough by the majority of participants and need to be categorized, Stachl et al., 2017). Some ML algorithms automatically compute transformed features. For example, the kernel functions in *Support Vector Machines* perform non-linear transformations to predictor variables (Cortes & Vapnik, 1995; James et al., 2013). Improving the feature engineering is often more important than testing yet another predictive algorithm (Kuhn & Johnson, 2013, 2020). Accordingly, feature engineering is a crucial part of the iterative modelling process in ML (Kuhn & Johnson, 2020), and an increasing number of specialized R packages are available for this task (Au, 2019; Roque & Ram, 2019).

In feature engineering, domain scientists (i.e. personality researchers in the present case) can contribute tremendously to the success of a predictive model. In the case of personality psychology, this step involves 'translating' extant knowledge or assumptions from past research into predictor variables that contain variation in relation to a previously reported finding (e.g. frequency of communication app usage in Stachl et al. 2019; which was informed by Montag et al. 2015). Deriving features from the psychological literature is particularly valuable but does not preclude the inclusion of additional features that have not been previously reported. While unintuitive features (e.g. entropy of app usage, Stachl et al., 2019) can make it harder to understand the results, they can also boost the predictive performance of the model and, if they turn out to be predictive, can generate new hypotheses for consideration in future confirmatory research (cf. data mining). However, too many additional, uninformative predictors can also lower the model's performance for some algorithms, so it is reasonable to favour features that will contribute useful information to the model. The usefulness of a feature can be determined by trying out different feature sets and selecting the best one or using dimensionality reduction techniques (e.g. via principal component analysis); this process might seem counterintuitive to researchers coming from the classical modelling culture. So it is important to keep in mind the slightly different philosophy of the ML modelling culture, namely, to create a model that achieves optimal prediction performance on new data (Breiman, 2001b). To avoid the overestimation of model performance, decisions about the selection of important features must happen within the resampling process of the model. This issue is further discussed in the Nested resampling section.

Overfitting and underfitting

Personality psychologists are most familiar with using classical linear models to describe or 'explain' some variable of interest (Shmueli, 2010). In this approach, model quality is usually evaluated by how much variance in the criterion variable can be explained by the predictor variables in a dataset (i.e. 'in-sample' R^2) and whether diagnostics of residuals suggest that distributional assumptions of the model are met (Breiman, 2001b; Fox, 1991). In contrast, the primary question for the evaluation of supervised ML models is how well new, unseen data points can be predicted based on a model that has been estimated on a given dataset (Breiman, 2001b). Some scholars argue that this approach to modelling more directly resembles a central goal of empirical science: building models that can make predictions about yet unknown cases (Yarkoni & Westfall, 2017). Recently, researchers have also started to consider additional criteria for evaluating ML models (e.g. simplicity of a model, Molnar et al., 2019).

Whenever ML models are adopted by a new discipline, the first wave of publications is often plagued by a major issue (Saeb et al., 2017; Varma & Simon, 2006): overly optimistic estimates of predictive performance for applied models. As noted earlier, this issue has affected personality science too (Mønsted et al., 2018). A common challenge in ML is models that are overfitted to the specific characteristics of a single dataset (Cawley & Talbot, 2010). Overfitting occurs when a model incorporates random variation in a given dataset, that is not caused by the underlying, true relationship between predictors and criterion variables. The overfitted model only 'memorizes' the specific data points, rather than to capture the true underlying signal. This issue leads to models that are not descriptive of the data generating process in the population, so their predictive performance suffers when applied to new data, generated by the same process. Overfitting is particularly problematic in small samples and when using overly flexible models. Overly flexible models are said to have high variance, implying that model predictions could vary considerably when training the same algorithm on different samples from the same population. Some model classes like polynomials or decision trees can suffer from high variance in prediction, if they are not adjusted to reduce flexibility (e.g. by pruning decision trees).

Underfitting can also be problem; it occurs when an inflexible model is not able to account for the true complexity (e.g. non-linear effects and interactions) in the data and therefore cannot represent the systematic variance. Inflexible models (e.g. linear models with a low number of predictors) are said to have a high bias: some model predictions are wrong in a systematic way that is independent of the specific sample.[3]Similar to overfitting, underfitting causes lower predictive performance on new data than an adequately flexible

[3]Note that this is not the same /header in the "Biased Models and Hidden Feedback Loops" section.

model could achieve. The general goal in supervised ML is to achieve a good 'bias-variance trade-off', which means finding a model in which the interplay of bias and variance leads to the best possible predictive performance.

The basic principles behind model overfitting and underfitting are visually presented in Figure 2. Three different models are fitted to training data (black dots) and evaluated on new test data (white dots). In terms of model flexibility, the green function represents a simple linear regression model, the orange function represents a fifth degree polynomial, and the blue function represents a ninth degree polynomial. For this simulated example, we also know the true (data generating) function in the population, indicated by the dotted line. The plot shows that the orange model approximates the true function more closely compared with the others and its predictions have the highest predictive performance on the test data ($R^2_{test} = 0.94$). The blue model shows overfit by interpolating all observations from the training data and has a lower predictive performance on the test data ($R^2_{test} = 0.79$). Clearly, the linear (green) model underfitted the training data because it was not able to account for the complex (non-linear) pattern. It has the lowest prediction performance on the test data in this example ($R^2_{test} = 0.50$).

As noted by Yarkoni and Westfall (2017), the best guard against overfitting is the use of larger samples. However, large samples do not guard against underfitting because inflexible models will stay inflexible, no matter how much data are used to fit them. The true function in the population is unknown, and visual inspection of the fitted function is not possible in higher dimensions. So, in practice, we cannot detect improper models in an intuitive way, as in the example given earlier. The best approach to address underfitting is to test different models and to select the one with the best predictive performance on new, unseen data (e.g. high accuracy or low error). One important meta strategy for building algorithms with high predictive performance in many applied settings is to use ensemble models, which combine several simple models like decision trees to achieve a good bias-variance trade-off. In random forests, a good trade-off is achieved by reducing the high variance of deep trees, while gradient boosting reduces the high bias of shallow trees (see Breiman, 2001a; Friedman, 2001, for a detailed discussion of these methods).

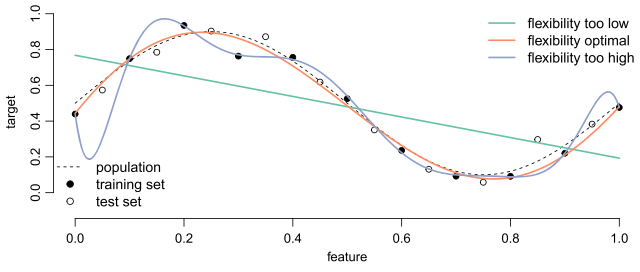

Figure 2. Schematic illustration of overfitting and underfitting based on three simulated models that use the same feature but have different model flexibility. Figure available at https://osf.io/j9yrw/, under a CC-BY4.0 license. [Colour figure can be viewed at wileyonlinelibrary.com] [Colour figure can be viewed at wileyonlinelibrary.com]

Nested resampling

Our overfitting example demonstrates why predictive performance estimates based on the training data (e.g. in-sample R^2) should not be used to estimate the predictive performance on new observations. Instead, resampling methods like *k*-fold cross-validation (Kohavi, 1995) should be used to repeatedly split the given dataset into a training set (on which the algorithm is trained) and a test set (on which predictions from the model built on the training set are used to compute an estimate of predictive performance). Unfortunately, this simple resampling strategy is not enough to prevent overly optimistic performance estimates in more complex modelling settings (Saeb et al., 2017; Varma & Simon, 2006).

To obtain realistic estimates of the predictive performance on new data, any decisions regarding the modelling process that are based on information from the complete dataset (training and test data combined) must be repeated in the resampling scheme (i.e. *nested resampling*, Bischl et al., 2012; Varma & Simon, 2006). Common steps that are mistakenly performed on the complete dataset and not properly handled within resampling include variable selection (e.g. based on pairwise correlations between the criterion and the predictor variables), the reduction of the dimension of the predictor space (e.g. principal component analysis), or the setting of hyper-parameters in ML algorithms (e.g. the learning rate in gradient boosting, Friedman, 2001). Consistent with the observations of Mønsted et al. (2018), we have also encountered multiple instances of these mistakes in reviews and published papers. In the following example of variable selection, we show how the lack of resampling can lead to overly optimistic performance estimates.

To illustrate this issue, we predicted self-reported extraversion scores by using 1821 mobile sensing derived features in a random forest model (Table 1). In the first attempt, we used the complete dataset to select the 10 features with the highest correlation with the extraversion score. Using only these, we trained and evaluated this model on the same data. In-sample fits are often reported in publications in psychology and can be dangerously optimistic estimates of how well a model would generalize to new data (Yarkoni & Westfall, 2017). In our example, the flexible random forest yielded an in-sample value of $R^2 = 0.16$. In the second attempt, we not only used resampling (10 times repeated 10-fold cross-validation) but also embedded the variable selection procedure in the resampling process (Bischl et al., 2012; Varma

Table 1. Performance overestimation in variable selection

	R^2		*MSE*	
	M	*SD*	*M*	*SD*
In-sample performance	0.16		0.46	
Nested resampling	0.04	0.10	0.52	0.09

Note: In-sample performance: variable selection based on Pearson correlations prior to model fitting and evaluating performance based on the training data. Nested resampling: variable selection embedded within 10 times repeated 10-fold cross-validation. Mean and standard deviation of R^2 and *MSE* were computed across folds.

& Simon, 2006). This procedure resulted in a lower, but much more realistic estimate of the model's predictive performance ($R^2_M = 0.04$; $R^2_{SD} = 0.10$).

Even when using nested resampling, overly optimistic estimates can be generated when the best-performing algorithm is selected *post hoc* based on performance estimates. After trying out a large enough number of different algorithms, one might work better than others in a given dataset but does so only due to chance. A possible solution is the implementation of an additional resampling loop for the ML algorithms or a thorough inspection of the performance variability during resampling.

Performance measures

An issue related to the correct computation of realistic performance estimates with (nested) resampling is how to quantify predictive performance. When comparing a set of predictions $\hat{y}_i$ for observations $i = 1,\ldots, n$ with their respective observed criterion values y_i, the so-called performance measure defines what is meant by good predictions. In regression settings like the example earlier, the default measure in ML is the mean squared error, $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$. In many studies conducted by social scientists, the coefficient of determination $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, with $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, is reported instead, probably due to perceived familiarity with classical linear regression. In contrast to the MSE, which depends on the unit of the criterion variable, R^2 is a normalized measure with a value of one when all predictions are identical to their observed criterion values, and the natural baseline of zero, resulting when the mean criterion value would be used as a constant prediction. However, R^2 can also take on negative values, when evaluated on test observations that were not used to train the model. Such negative values are not an anomaly of the measure. Negative R^2 indicates that the model makes worse predictions than a simple model with constant mean prediction, illustrating the important notion that overfitted ML models can do far worse than a simple guess at the mean value that ignores all predictor variables. In our own experience, negative R^2 is a frequent reality when applying flexible models to small datasets in personality science (Pargent & Albert-von der Gönna, 2018; Schoedel et al., 2018). Apart from publication bias, one reason why such negative values rarely appear in publications might be that some researchers do not use the aforementioned formula but compute the squared Pearson correlation between the predictions and the observed criterion values (Schwartz et al., 2013). Unfortunately, these measures again diverge when predictions are computed on test data. In contrast to the general R^2, which can be roughly thought of as a normalized version of the MSE, the squared Pearson correlation is a special linear rank measure. The Pearson correlation between predictions and observed criterion values, which is reported in a substantial number of publications (e.g. Gladstone et al., 2019; Youyou et al., 2015), mainly captures whether observations with higher values in the outcome also receive higher predictions and vice versa. It only weakly reflects how much the predicted scores differ from the observed criterion values. An almost perfect correlation can be found even when predictions are off in absolute terms by a large degree. Note that in overfitted models with negative general R^2, the correlation between predictions and true outcome values can also be negative. However, squaring the correlation without the consideration of negative values would mistakenly suggest satisfactory performance in situations when the true performance is in fact lower.

The point that the correlation between predictions and observed criterion values does only register ranks should not imply that this property is not useful. In contrast, we would argue that there might be many practical applications in personality science where only the ranks matter. One example is recruiting, where a company might primarily be interested in the best applicants and not in how much they differ. The same might be true when personality assessments are used in personalized products and marketing services (Matz et al., 2017); in such cases, natural rank measures like the Spearman correlation or Kendall's τ coefficient might be most appropriate. Intuitively, ranking observations is a different and often a much easier task than accurate predictions in an absolute sense. Relying too heavily on absolute measures like the general R^2, the MSE, or more robust versions like the mean absolute error (MAE) might lead researchers to ignore models that are actually more suitable in practice. With an increasing adoption of ML methods in personality science, justifying which performance measures should be used to determine which models are best suited will be highly relevant. This requires personality scientists to be familiar with the practical consequences of choosing a certain performance measure.

In (binary) classification scenarios, where the criterion variable is a discrete factor variable, different challenges arise. The simplest performance measure here is the mean misclassification error, which is equal to the relative frequency of incorrect predictions. Unfortunately, this standard measure can be misleading in settings with imbalanced classes and also heavily depends on the concrete probability threshold, which is used to transform estimated class probabilities into predictions (Kuhn & Johnson, 2013). Alternatively, one can monitor and compare two performance measures (e.g. sensitivity and specificity) or use combined measures like F1 or the area under the curve (AUC). Measures independent of probability thresholds are the brier score or the AUC (for a comparison of different measures, see Ferri et al., 2009).[4]

Researcher degrees of freedom

The careful use of ML methods, as described earlier (e.g. nested resampling), somewhat guards against drawing false conclusions from data (Yarkoni & Westfall, 2017). However, we want to emphasize that more general principles of good scientific practice still apply too; personality scientists will,

[4]Most performance measures are easily adapted for multiclass classification, in which the criterion variable has more than two distinct values.

now more than ever, need large and representative samples to profit from the high flexibility of ML models and to obtain more precise performance estimates. Thanks to the new methods for gathering data (e.g. digital footprints and sensing data), obtaining large, diverse samples should be viable. In fact, personality researchers may soon be faced with the problem of how to deal with big datasets and data streams (Domingos & Hulten, 2000; Katal et al., 2013).

Correctly evaluated predictive models will provide more realistic estimates of how well the models generalize to new data. These realistic estimates could have the effect of drastically reducing many previously reported effect sizes to zero (Yarkoni & Westfall, 2017), as was the case in our illustration in the section on overfitting and underfitting. These lowered effect size estimates might create dangerous incentives to enhance reported performance by the use of researcher degrees of freedom (Sculley et al., 2018).

Even in the application of classical statistical analyses (Wicherts et al., 2016), researchers must make many analytic decisions; however, in the application of ML models, researchers have many times more decisions to make (e.g. in the selection of algorithm implementation, hyper-parameter settings, and resampling strategy, Pargent & Albert-von der Gönna, 2018). Therefore, more intensive and overarching efforts in open science practices will be necessary to ensure the integrity of findings from ML studies. These efforts should include the pre-registration of research and a clear labelling of exploratory versus confirmatory research (Jaeger & Halliday, 1998). Most importantly, the complete transparency of code and data (whenever publishing the data is possible) should be a requirement for ML analyses (Sculley et al., 2018). Finally, reporting standards for the use of ML models in psychological science should be improved. For example, details should be provided regarding the algorithms used (including the R or Python package), the exact type of resampling including fold aggregation procedure (e.g. pooling, mean, and median), and at least two different performance measures (a relative measure and an absolute measure). Also, as shown by Schoedel et al., , the exact way in which pre-processing was performed (e.g. imputation, transformations, variable selection, and whether it was performed within resampling or prior to it) should be made transparent.

Classification versus regression

One debatable way for researchers to boost the reported performance of ML models is by using classification instead of regression methods. In the analysis of our data reported earlier, we fitted a regression random forest, thus predicting continuous values for the outcome variable extraversion. However, a lot of work in personality computing has focused on predicting classes (i.e. 'low' vs. 'high') of personality traits, rather than continuous trait scores (Chittaranjan et al., 2013; De Montjoye et al., 2013; Majumder et al., 2017; Mønsted et al., 2018). This decision to focus on classes can pose a problem when the rationale for creating discrete classes is not fully transparent. In binary classification, the two classes are often generated around some fixed central tendency estimate (e.g. median), obtained from the sample under investigation (e.g. Chittaranjan et al., 2013). In some cases, an arbitrary dividing point is used (e.g. determine that the midpoint of a five-point rating scale is assigned to the 'low' vs. the 'high' class), leaving open the possibility that the decision was made to maximize reported performance.[5]

Classification problems are sometimes favoured over the prediction of continuous trait scores because they seem more intuitive and might suggest above chance performance in cases where regression models have not been successful. From a social science perspective, artificially constructed classification models have impeded theoretical progress for two main reasons: first, past research indicates that the distribution of trait scores in a population tends to be roughly Gaussian, such that most individuals will fall to the central tendency estimate of the scale (Schmitt et al., 2007). Hence, binary classification arbitrarily 'forces' a high/low distinction upon individuals with trait scores very close to the median, implying a greater separation between subjects than actually exists. Every measurement (i.e. individual personality score from a questionnaire) is error-prone, and the observed criterion values of each individual (value on the latent variable) may be close to but not exactly equal to the measured value. Hence, splitting a (normally distributed) sample at the median or mode will naturally result in the highest number of misclassifications due to measurement error. Therefore, the goal of automatically predicting personality trait scores in new data is made difficult, because the model is trained to classify traits based on a threshold that is likely to be idiosyncratic to the training dataset. A cut-off based on a large normative sample could be used instead. Second, comparing predictive performance across studies is very difficult when one study performed classification and the other performed regression, because performance measures from both settings cannot be easily compared. When classification is necessary to achieve satisfactory performance, this should be made transparent in the paper, and the data should be made available, so that other researchers are at least able to compute regression performance metrics themselves.

INTERPRETATION OF MACHINE LEARNING MODELS

Predictive models are often roughly separated into *good for prediction* (e.g. algorithmic, non-linear models like random forests) and *good for explanation* categories (e.g. classical, stochastic models like linear regression, Breiman, 2001b). Classical stochastic models come with a series of assumptions about the data generating process (e.g. linearity and homoscedasticity). Within the specific frame of these assumptions, predictions and model parameters have useful interpretations, and it is (relatively) intuitive to understand the assumed functional relationship. For example, it might be possible to mentally grasp a regression hyperplane, which is defined by a simple linear equation.

[5]When non-binary classification is used, the authors typically classify users based on the magnitude of the ordinal personality trait scale (De Montjoye et al., 2013; Mønsted et al., 2018).

However, a big downside of linear models is that reality has more than once proven to be complex and is often non-linear (Benson & Campbell, J. P., 2007; Cucina & Vasilopoulos, 2005; Stachl et al., 2019). In prediction-focused benchmark experiments, linear models often perform worse than more flexible algorithms. Algorithmic models can reflect more complex patterns in data but generally lack directly interpretable parameters. For example, there is no simple equation, which can be written down to describe the algorithmic procedure by which a random forest computes its predictions. Especially for ensemble models and deep neural networks, there is no straightforward picture of which functional relationship has been learned by the model. By refraining from restrictive assumptions about the data generative process, algorithmic models are trading in their out-of-the-box interpretability for an increase in prediction performance.

Machine learning sometimes has a bad reputation because of the limited interpretability of such *black box* models. If you asked personality scientists whether they would prefer predictive or interpretable models, the answer would probably be 'all of the above'. This desire for interpretability triggered the development of new methods to extract useful information from the *black boxes*' inner workings. In this section, we provide short summaries of some model-agnostic methods and provide some suggestions for how they could be used; for readers wanting more extensive information, we refer them to the comprehensive *Interpretable Machine Learning* book (Molnar, 2019), as a good starting point. We also highlight that an alternative strategy to explaining black box models, which we do not discuss further, is to search for interpretable models with comparable predictive performance (Rudin, 2019).

Variable importance measures

Several methods have been developed to better understand how predictions in ML models are made (Doshi-Velez & Kim, 2017; Guidotti et al., 2018). As explained by Yarkoni and Westfall (2017), the importance of single predictors or groups of predictors can be assessed by comparing the predictive performance of models trained with and without them. These analyses are computationally demanding for large predictor sets, so variable importance measures have been developed to approximate them. One generic metric is permutation importance. It was originally proposed by Breiman (2001a) for random forest models, but the method is in fact model agnostic (Fisher et al., 2018). The principle behind permutation importance is relatively straightforward: values in the variables of interest are shuffled across observations (i.e. permuted) before prediction. The greater the decline in predictive performance in comparison with predictions with the original unshuffled version of the variable, the higher the importance of it. However, unlike standardized β-coefficients from linear regression models, permutation importance estimates do not represent the unique contribution of predictor variables under the assumption that all other predictor variables remain constant. Rather, this metric quantifies the marginal impact of a predictor including interactions with all other predictors in the model. In the case of correlated predictors, this procedure can reduce the indicated importance of inter-correlated variables in comparison with a model that would contain only one of those predictors (Nicodemus et al., 2010).[6]Hence, depending on the research question, it might be useful to also consider alternative measures of variable importance with a conditional interpretation (Strobl et al., 2008). Note that the calculation of variable importance measures can be performed on independent test data or on the complete dataset. Especially in small samples, the latter approach can lead to unrealistic results. For a more extensive discussion of this unsolved issue, see Molnar (2019).[7]

Getting back to personality psychology, we fitted another random forest model to predict extraversion scores with a subset of the predictor variables (communication-related variables). We then calculated permutation importance measures for these variables. In Figure 3, predictor variables are ranked by their permutation importance (loss in mean absolute error). For example, the variable daily mean number of phone ringing events (*daily_mean_num_call_ring*) seems particularly important in this model. These estimates can be used in personality science to (i) compare the model with theoretical assumptions (see Evaluation of Machine Learning-Based Personality Scales section on the evaluation of machine learning models), (ii) create starting points for the extension of existing theories, and (iii) generate new hypotheses for future research.

Visualizations of predictor effects

To better understand the influence of predictors on the predictions of a criterion, it can be helpful to visualize the marginal effects in a plot. Partial dependence plots (Friedman, 2001), individual conditional expectation plots (ICE; Goldstein et al., 2015), and accumulated local effect plots (ALEs Apley, 2016) are common methods to achieve this. Partial dependence plots cannot handle correlated predictors very well, which is why we use ALE plots in our example. Figure 4 displays the importance of the *daily mean number of outgoing calls* in a predictive model for the personality trait extraversion. Unlike variable importance measures, the plot shows *how* predictions of the outcome variable (extraversion) change with regard to regions of the predictor variable. For our example, the plot suggests that predicted extraversion increases up to a mean number of about three outgoing calls per day.[8]The visualization therefore provides additional information beyond variable importance measures and can help researchers understand how exactly the predictor variables and the criterion variable are theoretically related.

Research in personality psychology is focused on the understanding and predicting of systematic differences in

[6]This phenomenon is also known from classical regression and structural equation models (McFatter, 1979).
[7]For the sake of completeness, we included syntax to compute permutation importance based on both the complete and test data in the project's repository at https://osf.io/j9yrw/.
[8]Note the minuscule effect size of this single predictor variable. The first and third quartiles of the extraversion variable are $Q_1 = -0.52$ and $Q_3 = 0.50$.

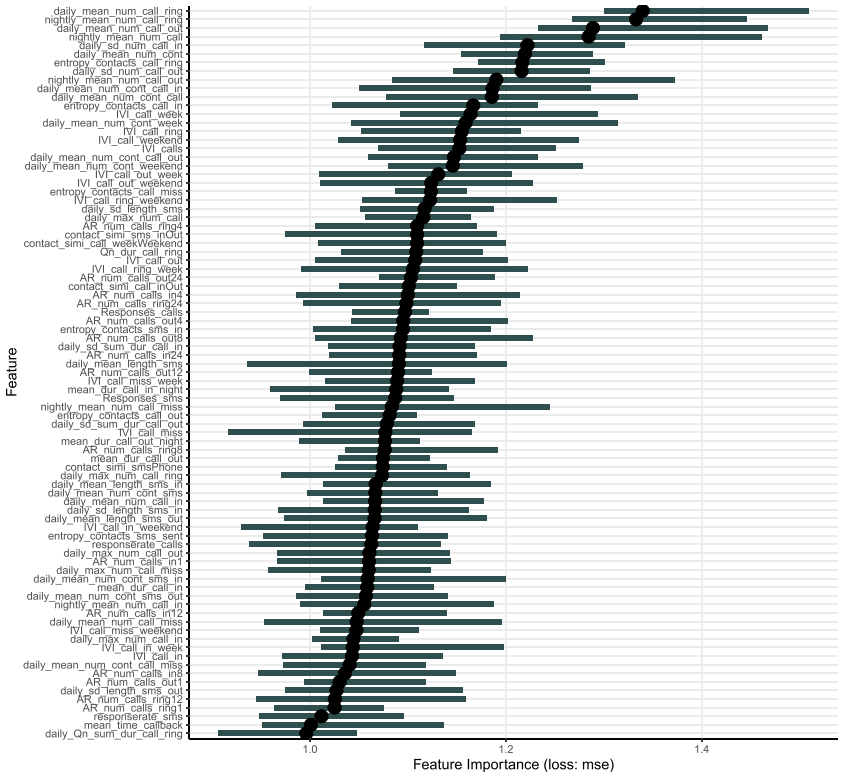

Figure 3. Permutation variable importance. Importance measures were obtained with 10 repetitions. Measure is reduction in mean squared error. More information on the displayed variables can be found in Stachl et al. (2019). Figure available at https://osf.io/j9yrw/, under at a CC-BY4.0 license. [Colour figure can be viewed at wileyonlinelibrary.com]

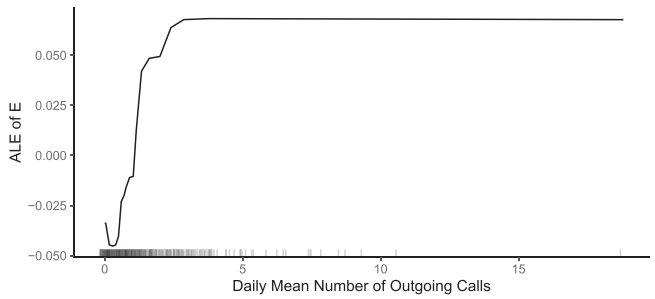

Figure 4. ALE plot visualizing the change in mean predicted values of extraversion with regard to the daily mean number of outgoing calls. Figure available at https://osf.io/j9yrw/, under a CC-BY4.0 license

individuals' mental and cognitive states, traits, the related processes that cause these differences, and their variation across time and situations (personality dynamics Funder, 2006; Rauthmann et al., 2015). As demonstrated in Figures 3 and 4, variable importance and ALE plots can help researchers quantify and visualize the impact of features in predictions of flexible, non-linear models. Researchers using ML methods in personality psychology should report variable importance and predictor effects in addition to measures of predictive performance. They are not equivalent to parameters from stochastic models but enrich reports of successful predictive models with information about how these models work and what information the predictions are based on. This information not only helps to make psychology a more data-focused and prediction-focused science (Yarkoni & Westfall, 2017) but also extend and challenge our knowledge about existing theoretical constructs, such as behavioural manifestations of personality traits (Stachl et al., 2019). More broadly, the scientific investigation of natural phenomena has traditionally been viewed as a strictly deductive process, dictated by the rigorous testing of pre-specified hypotheses. In contrast to this view stands the inductive approach to scientific reasoning: the data-driven creation of knowledge in a bottom-up process. The high dimensionality of new types of data and the fact that for many traceable behaviours no *a priori* hypotheses exist render a purely deductive scientific approach unsuitable in many cases or at least inefficient. A more iterative and alternating process between inductive and deductive scientific practices has been called for Mahmoodi et al. (2017). Using ML algorithms together with methods of model interpretation could help to complete the circle between prediction and explanation in personality psychology (Shmueli, 2010).

FAIRNESS OF MACHINE LEARNING MODELS

In addition to helping with theory development, increasing the interpretability of ML methods can also help make personality psychology more relevant in practical contexts. As noted earlier, many flexible ML models can be difficult to understand. However, using these models for personality-based personnel assessment in organizations will require a good understanding of how the applied models make decisions and which information they use to do so.

This step is necessary to avoid algorithmic discrimination (Kusner & Loftus, 2020; Sweeney, 2013), to comply with legal requirements (Goodman & Flaxman, 2016), and to decide whether an algorithm uses only the 'right' information for its predictions to arrive at decisions that are fair. When ML models are treated as black boxes and little to no attention is paid to the construction and inner workings of the models, unwanted information can make its way into the predictions. For example, a model that aims to classify pictures into huskies and wolves could have high accuracy while only looking for the presence of snow in the pictures (Ribeiro et al., 2016). Furthermore, the information used in ML models can become outdated over time, and the predictiveness of once impactful variables can deteriorate. For example, how people's online behaviour is related to personality might be changing over time as different things become trendy (Kosinski et al., 2014). Hence, the initial and continuous validation of a model's functionality is crucial for its application in practical settings. Model validation can be intricate and has often been neglected in applied ML contexts, often with serious consequences (Dastin, 2018). However, model validation is extremely relevant for the application of ML in the field of personality psychology. Most studies using ML in personality research have not focused on the analysis of the inner workings of their models (see Settanni et al., 2018). Thus, in the following sections, we discuss two key aspects of model validity (concept drift and biases), and we highlight their importance through examples.

Concept drift in machine learning models

Personality science could benefit a great deal from using interpretable ML methods in research as well as in practical applications like personnel selection. However, it is not certain that the performance of a trained model will remain constant during its lifecycle. This phenomenon is called 'concept drift' and describes how the prediction error of a trained model increases over the application period of the model, possibly without being noticed. In the case of personality assessment, an example could be the progressively decreasing accuracy of a model predicting personality scores. This decreased accuracy could, in turn, lead to unfavourable consequences in practice (e.g. declining effectiveness of some recruitment process). Lu et al. (2018) describe a range of different types of drift, such as gradual, but also sudden, and recurring drifts.

What are the possible reasons for such a decline in predictive accuracy? Technology and culture are evolving at a rapid pace such that the purpose of technical devices and the way we interact with them are constantly changing. Consequently, the information structure in the data resulting from such measurement devices is also changing. If variables from mobile sensing are used as features in ML models and the personality-related information embedded in these digital records changes over time, predictions of the model might gradually drift. For example, the type of apps that extraverted people use might change over time. Thus, performance deteriorates if the model is not retrained at appropriate intervals. Slow and gradual drifts are particularly likely to go unnoticed (Baena-Garcıa et al., 2006). These changes can be particularly impactful if the direct relationship with the criterion is affected. For example, Matz et al. (2017) speculated that 'liking' the TV series *Game of Thrones* on a social network in 2017, when it had been established in mainstream culture, might be differentially indicative of individual personality traits like openness to experiences, compared with having 'liked' it during its first season. The start of the final season in 2019, which was received with mixed feelings by the fanbase, might mark another concept drift in likes.

Over the years, mobile phone usage has undergone steep and continuous change. Originally, mobile phones were mostly used for calls on the go, but now modern smartphones serve as powerful mobile computers offering a vast range of functionalities. Forms of communication have become more diverse, and the resulting digital records and their importance for predicting personality have changed. For example, calls and text messages are increasingly being replaced by audio and video messages, managed by various apps (Lu et al., 2018). This drift can be demonstrated in the mobile sensing data from the PhoneStudy project. In Figure 5, the frequency of some communication-related variables are plotted by the year of collection (study 1: 2014/15, $n = 137$, study 2: 2015/16, $n = 248$, study 3: 2017/18, $n = 279$; Stachl et al., 2019). The data suggest that from 2014 to 2018, the daily mean number of outgoing and incoming text messages decreased whereas the daily mean number of communication and social media app usage increased. Furthermore, the Spearman correlations between extraversion and the daily mean number of messaging app usage decreased from $r_s = 0.24, CI_{95\%} = [0.08, 0.39]$ in study 1, to $r_s = 0.14, CI_{95\%} = [0.01, 0.26]$ in study 2, and $r_s = 0.14 CI_{95\%} = [0.03, 0.26]$ in study 3. This effect illustrates how the predictive performance of models trained in 2014 using these data might deteriorate over time. Note that the descriptive tendency shown in this pedagogical example could also be caused by other factors so we discourage any substantive generalizations, based on these findings.

The changing relevance of input variables poses a challenge for (interpretable) ML applications in personality science. The formulation of persistent theories is often the primary goal in personality, so future research should address the question of how models based on rapidly changing technological indicators can become more robust against changes in digital behaviours. Several approaches have been proposed, such as implementing a control mechanism in terms of a wrapper or an online algorithm (Baena-Garcıa et al., 2006; Gama & Castillo, 2006). These incremental rule-based models use decision rules to directly detect a drift in the incoming data stream (Deckert, 2013). Most methods compare the performance of a single model on different time windows and adapt the model when the change in error rate passes a threshold (Lu et al., 2018), but Klinkenberg (2005) proposed selecting the best-performing model each time the drift control is implemented.

One possible way to overcome the problem of drift might be to group single events into categories, such as 'messages', regardless of messages' form (text or voice) or content. We suggest that researchers invest time and effort to finding

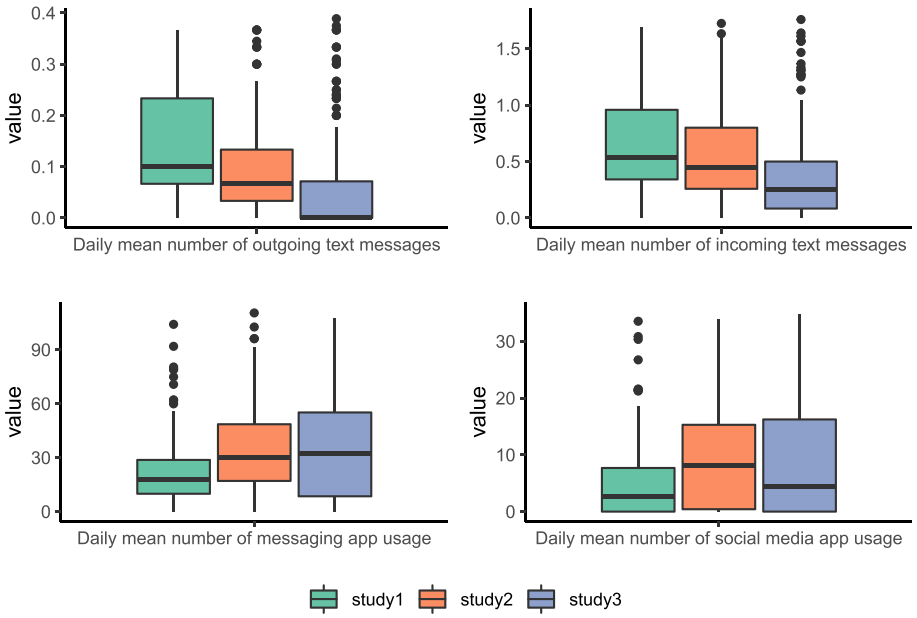

Figure 5. Boxplots showing the distribution of the daily mean number of communication-related variables separated by study. Outliers more than three times the standard deviation from the median were excluded. Figure available at https://osf.io/j9yrw/, under a CC-BY4.0 license. [Colour figure can be viewed at wileyonlinelibrary.com]

persistent and stable digital behavioural dimensions when working on theoretical models in personality psychology. However, if the primary goal is the application of ML models (e.g. in personality assessment), we need to be ready to take into account the changes in human behaviour in a rapidly evolving digitized world.

Similar to the process by which people naturally update and improve their personality inferences about others over time, there might come a paradigm shift, from fixed models that only work well at a given point in time to continuously learning models that are frequently re-evaluated and retrained when the performance deteriorates (Gonfalonieri, 2019). Of course, this issue is not unique to ML models; the meaning of traditional questionnaire items (e.g. 'I like going dancing in the ballroom' vs. 'I like going dancing at nightclubs' as an assumed indicator for extraversion) also changes over time. However, we assume that in many cases, the life expectancy of meaningful questionnaire items will be higher than that of indicators from digital data.

Biased models and feedback loops

In some cases, concept drift can be caused by the model itself; so-called (hidden) feedback loops are a particular aspect of applying ML models that are iteratively trained as new data are being gathered (cf. online-learning). These loops occur when the application of an algorithm has an effect on the data they are fed to learn from. They can cause simple bugs, like a font that keeps expanding endlessly (Sculley et al., 2014), but may also cause serious harm in situations where social inequalities are deepened by an algorithm that uses them as features (Kusner & Loftus, 2020). An interesting example is the set of models that send the police to areas where the most crimes are reported, leading to even more crimes being reported (because there are so many police units to report them), resulting in yet more units sent to this area, and so on (see Lum & Isaac, 2016).

Hidden feedback loops can pose a grave danger to the benefits of predictive models. Therefore, algorithmic transparency is essential, which is reflected in calls for a rigorous science of interpretable ML (Doshi-Velez & Kim, 2017) and to ensure that ML models are aligned with human values (Irving & Askell, 2019). *Fair machine learning* has grown as an important research field dedicated to dealing with the unwanted effects of algorithms, whether they are of an ethical, financial, or scientific nature (see Chouldechova & Roth, 2018). Hidden feedback loops are important here because they are concerned with differential effects of algorithms on sub-populations or underrepresented groups (Liu et al., 2018), which are often discriminated against. In October 2018, news outlets reported that the online-retailer Amazon Inc. abandoned a project on using ML in their hiring process, when serious discrimination against women was discovered in an algorithm trained on texts from resumes of previously successful employees at the company (Dastin, 2018). Investigators noticed that the algorithm learned that many current and past employees had male first names and consequently used that information for the selection of prospective employees. If such an algorithm would be applied, even more men would be employed, which could lead to a working environment in which female employees are considered even less effective, thereby further strengthening the bias in later iterations of the model. In personality psychology, dangerous feedback loops could also arise in the context of proposed policy changes targeted at altering citizens' personalities based on research on the effectiveness of personality interventions (Bleidorn et al., 2019). For example, individual personality trait levels could be used for personalization efforts in adaptive systems. People with low predicted scores in extraversion could be exposed to content that would increase

system usage (e.g. screen time), but doing so would also decrease social interaction (e.g. highly immersive single player games). Consequently, this intervention could cause even lower predicted scores in extraversion.

Moreover, even if a sensitive factor (e.g. gender or race) is eliminated from the training data, it is not guaranteed that a new model will produce unbiased predictions. The removed information can often be inferred based on a combination of harmless features (Ingold & Soper, 2016). Often, the only reliable option to ensure the absence of bias is to explicitly compare the predictions for meaningful groups of observations (e.g. comparing the predictions for male and female applicants). Recently, researchers in artificial intelligence have realized that psychologists and other social scientists with an expertise in experimental design could play an important role in monitoring ML models experimentally (Irving & Askell, 2019). Based on effective new methods for interpretable ML, this could ensure that important ethical constraints and requirements are met. Naturally, in applications like personnel selection, this experimental process would require a detailed understanding of human psychology and personality assessment; personality psychologists would be well placed to contribute their knowledge about individual differences to this process.

Personality psychologist might encounter ethically problematic applications of ML sooner than expected. Overly optimistic claims about the success of ML models in predicting personality (Mønsted et al., 2018) have already given rise to an ever-growing number of IT start-ups. These start-ups sell the promise of predicting just about any outcome of interest, including individual personality trait levels. Past research indeed suggests that personality prediction might be possible to some degree (Park et al., 2015; Schwartz et al., 2013), but for many commercial products, it is unclear how well these systems actually work and whether they have been tested against the problematic biases described earlier. A recent example of a commercial psychological test that promises to predict job-relevant characteristics (including personality traits like emotional stability, sense of responsibility, and goal orientation) is the PRECIRE JobFit; all predictors are based on speech samples from an automated phone interview. By publishing an official test review (Schmidt-Atzert et al., 2019), the Testkuratorium of the German Psychological Society might have inadvertently legitimized the use of the test in personality assessment and recruiting. Although the method received a comparably bad rating in the review, this could be problematic, because the manual did not report sufficient information on the applied ML algorithms and on how their performance was evaluated. Based on the information in the review, none of the dangers of bias outlined in this section seem to be addressed by the manual (which is also not publicly available). We hope this example will lead to a discussion on the general evaluation process of psychometric tests (i.e. transparency), with special considerations for ML-based tests. If personality psychologists remain unfamiliar with basic ML principles, our discipline will be poorly placed to advice industrial partners about which assessment tools should be used in responsible recruiting practices and about the advantages and disadvantages of the new methods.

EVALUATION OF MACHINE LEARNING-BASED PERSONALITY SCALES

At the moment, most applications of ML methods in personality research are models trained on new types of indicators (e.g. smartphone logging data) to predict scores from established personality inventories (e.g. the BFSI). Bleidorn and Hopwood (2018) review and evaluate early studies from this line of research, which they call *Machine Learning Personality Assessment*. They raise the important question of how personality scales derived from this framework should be evaluated by the scientific community. Bleidorn and Hopwood (2018) recommend directly transferring Cronbach and Meehl's (1955) familiar construct validity framework to the validation of ML-based personality scales. To determine what role the construct validation framework should play, it is important to consider the intended goal and use-case of a newly constructed psychometric scale. If the predictions of an ML-based personality scale are to be used as real *measures* of a latent variable from classical or probabilistic test theory, with the main difference being that the indicators come from some digital device instead of a questionnaire, then the ML-based scale should be evaluated using criteria similar to those used for traditional scales (i.e. construct validation). However, there are also applications for which the ML-based scale is trained directly on a concrete criterion (e.g. job fit), instead of predicting a questionnaire score; in such cases, classical psychometric properties should play only a secondary role, and the potential of the ML-based scale should not be restricted by holding the scale to the standards required of a traditional questionnaire-based scale. The following psychometric properties play a different role for ML-based scales.

Convergent and discriminant validity

Probably the most important criterion for assessing the quality of an ML-based personality scale is its performance in predicting the personality questionnaire it has been trained to predict. Adopting the traditional construct validation framework, Bleidorn and Hopwood (2018) describe this performance as *convergent validity*. However, it is important to distinguish between correlations with the primary target variable in a supervised ML task and correlations with external measures of the target construct (usually a different questionnaire) that the model was not explicitly trained to predict. High correlations with convergent external measures are much more impressive and are differentially informative about the usefulness of the scale. This distinction becomes even more important considering that ML with multiple outputs allows researchers to optimize the performance with regard to more than one target variables.[9] The same strategy could also be used to modify the model's loss function to explicitly discourage associations with constructs expected to be distinct from the measured construct of interest (akin to

[9] See Tomašev et al. (2019) for a current example with several secondary targets, whose addition led to a significant increase in predictive performance for the primary target.

promoting discriminant validity in the classical construct validity framework). However, this option should be used with care because low correlations with theoretically distinct constructs should not, by themselves, be considered a measure of quality. For example, there could be cases in which personality science proposes constructs that are not distinct at all, when evaluated on an empirical basis that is not biased by the subjective process of item construction in classical tests. To account for these unique opportunities of ML-based scales, we propose that the terms *internal convergent validity* and *internal discriminant validity* be used to refer to correlations of model predictions with target variables included into the loss function during training and that *external convergent validity* and *external discriminant validity* be used to refer to the traditional correlations with external constructs, which were not considered during training.

Reliability and generalizability

Reliability refers to the amount of variance in the true scores of a test in relation to its total variance. The variance of the true score can only be determined by observing repeated test administrations across different instances like time or (subsets of) items. Measures of internal consistency like Cronbach'sα (Cronbach, 1951) should not be used for ML-based personality scales because (i) those measures estimate the reliability of the sum score of individual indicators, which is not how the actual predictions are computed for ML models (e.g. random forests) and (ii) we do not expect all variables in ML-based personality scales to be exchangeable indicators for a single latent variable. We agree with Bleidorn and Hopwood (2018) that estimates of retest reliability can and should be computed for ML-based personality scales. This procedure is quite straightforward for predictive models with features that are based on aggregated statistics (e.g. average call frequency per day) and does not even require additional data collection; for example, for a model predicting personality with social media language, Park et al. (2015) report correlations between predictions based on text from different time intervals. We know from classical test theory, that the correlation between test repetitions can be interpreted as reliability only if the parallel measurement model holds (Steyer, 2001). Whenever more than two test repetitions can be computed as in Park et al. (2015), structural equation modelling should be used to (i) test whether predictions from different time intervals can be considered unidimensional indicators of a common latent variable, (ii) compare different measurement models, and (iii) use the appropriate reliability measure for a model with appropriate fit.

Generalizability takes reliability one step further by focusing on the adequacy of the model in new contexts (Bleidorn & Hopwood, 2018). As we have noted in this paper, ML research, with its heavy focus on out-of-sample performance, experimentally testing model predictions for hidden bias, and continuously validating models during their lifecycle, could be an excellent role model for psychological science (Yarkoni & Westfall, 2017). Those principles should be embraced for both ML-based personality scales and traditional ones.

Content validity and factorial validity

Bleidorn and Hopwood (2018) propose using expert ratings to determine which predictor variables are in line with personality theory and using only those ratings to train ML-based personality scales. We agree that effective feature engineering is usually based on theoretic knowledge of personality constructs, but we do not think that a theoretic and intuitive interpretation of features should be a necessity. In fact, one of the biggest potentials of ML-based personality scales is their capacity to reflect a more realistic structure of human personality that is most certainly more complex than models currently used in personality research. Hence, we are concerned that an over-reliance on traditional notions of content validity could be detrimental to the ultimate performance of new innovative scales.

Similarly, the sole examination of the linear factor structure of ML-based personality scales (Bleidorn & Hopwood, 2018) might fall short of realizing the full potential of these methods. We agree that this approach could be a useful application of interpretable ML to explain model predictions; however, an intuitively interpretable factor structure should not be a necessary psychometric property of a personality measure, particularly if it has been deliberately designed to optimize internal convergent validity to detect the potentially complex non-linear structure of digital indicators of personality. Thus, when the primary goal is to construct scales with a well-defined structure from theory-derived interpretable indicators, psychometric models that can incorporate non-linear interactions between person covariates (e.g. Brandmaier et al., 2013) might be a more suitable framework compared with classical ML models.

MACHINE LEARNINGWITH LATENT VARIABLES

In contrast to the usage of ML in many areas, applications in psychological research face the unique challenge of latent variables (e.g. personality traits) that cannot be measured directly. Psychometric models are commonly used to infer these latent variables from indirect indicators like questionnaire items. All well-established personality models currently rely on questionnaire data to measure human personality. Consequently, when using ML in personality research and assessment, two common scenarios arise: in ML models, personality constructs are used either as criterion variable or as predictor variables. We will discuss important implications of both settings.

Personality constructs as criterion variables

As noted earlier, personality constructs are often predicted based on potentially interesting predictor sets (Settanni et al., 2018). In this scenario, a single numeric descriptor of individual personality scores is used as the criterion variable in a supervised ML task. The simplest and most frequently used measures are the arithmetic mean or sum score of the complete set of items of the personality questionnaire, which are theorized to measure the dimension of interest (e.g.

Youyou et al., 2015). However, this simple approach ignores the error inherent to psychological measurements. Since the development of Spearman's (1904) famous formula of correction for attenuation, scholars have been aware that simple correlations between manifest test scores will underestimate the true association between the presumably underlying latent variables. With substantial amounts of measurement error contaminating most psychological measures, this discrepancy can become quite large. Thus, using the sum score as the criterion can lead to serious underestimation of the performance theoretically possible by an ML model.

One solution to this problem that takes measurement error into account is to instead use trait estimates from a psychometric model (instead of sum scores) as the criterion. Stachl et al. (2019) implemented this solution, using estimates from a partial credit model (Masters, 1982), which is the item response model that was used in the normative sample of the BFSI (Arendasy, 2009). The general strategy of substituting a manifest criterion variable in ML by the output of a theoretically crafted statistical model is not uncommon and has already been used in numerous applications (e.g. Hothorn & Jung, 2014).

One problem of this more sophisticated strategy is that the ML model is unaware of the estimation error of the psychometric model that generated the latent trait estimates. This problem can be solved by integrating the psychometric measurement model into the ML algorithm that is used to make the predictions. Consider that simple item response models like the Rasch model are in fact generalized linear models (Bürkner, 2019). Hence, the models could be incorporated in artificial neural networks (Goodfellow et al., 2016), which include generalized linear models as a special case. Yeung (2019) recently demonstrated the viability of this procedure, employing the Rasch model as an output function in a deep neural network in the context of *knowledge tracing* (i.e. predicting which questions are solved by participants in massive online courses, Hernández-Blanco et al., 2019).[10]

The correction for attenuation formula, which is used in many ML applications in personality science (e.g. Gladstone et al., 2019; Youyou et al., 2015), implies that using manifest sum scores as criterion variable leads to underestimation of predictive performance. However, overestimation is also possible when systematic method bias is contained in the psychological measurements. There is a huge psychometric literature suggesting that questionnaire responses contain not just trait information but are also consistently influenced by response styles (Jackson & Messick, 1958), which are stable trait-like individual differences in how people use the response categories of a questionnaire (Wetzel et al., 2016). If these response styles are correlated with the psychological trait of interest, which has been suggested by some empirical studies (Naemi et al., 2009; Zettler et al., 2016), flexible ML algorithms might inadvertently model these tendencies. This would pose a big problem because ML models would then also predict this method bias instead of predicting just the latent personality trait they are supposed to predict. With increasing amounts of data, overly optimistic performance estimates of ML models might be even more confounded by method bias than classical modelling approaches. Thus, the possibility that personality scores might not be unidimensional and might contain method bias should always be considered before using the correction for attenuation formula to adjust predictive performance in ML.

[10]Similar to Mehta et al. (2020), the knowledge tracing literature serves as an example of a task traditionally located in educational psychology that is now increasingly solved with state-of-the-art ML technology (Piech et al., 2015).

Personality constructs as predictor variables

A slightly different situation arises when psychological constructs are used as predictor variables in ML analyses. Here, the problem is not that the performance estimates might be biased but that it might be possible to increase predictive performance by clever feature engineering. When using personality measurements from questionnaires as predictors, an important question is which degree of aggregation should be used in the analyses. In personality psychology, this is sometimes referred to as the fidelity-bandwidth dilemma (Cronbach & Gleser, 1957; Hogan & Roberts, 1996). One theoretical approach to explaining and dealing with the phenomenon is Brunswik's (1956) lens model, which describes the aggregation and disaggregation of indicators to predict psychological constructs of different granularity. Many psychologists would naturally use the domain sum scores based on the item responses as predictors because this practice reflects their psychometric methods training. However, many ML algorithms can handle multiple predictor variables simultaneously. Hence, each item could be used as a separate predictor (for a demonstration, see Pargent & Albert-von der Gönna, 2018). The aggregation of item scores also involves some loss of information, so their individual use could lead to better predictions; supporting this idea, some recent studies have found small but consistent increases in predictive performance when using items instead of sum scores (Seeboth & Mõttus, 2018; Zweck et al., 2019).

Not computing summary statistics delegates the task of separating true signal and noise from the practitioner to the ML algorithm. Instead of weighting predictors based on some theoretical measurement model (i.e. the sum score as the simplest example), the ML model learns appropriate weights based on the data. This procedure might yield superior results when sample sizes are large enough, but the model-based approach might be more effective if the sample is small. Similar to the setting with personality constructs as criterion variable, the use of trait estimates from psychometric models as predictors might be a useful strategy. We have already noted that systematic method bias in questionnaires, such as response styles, can be meaningfully associated with the criterion variable. Thus, combining psychometric modelling of method bias (e.g. Böckenholt & Meiser, 2017; Jin & Wang, 2014; Tutz et al., 2018) with flexible ML algorithms, capable of modelling non-linear effects and interactions, might even allow us to use the peculiarities of psychological measurements to increase predictive performance. Preliminary attempts to include separate indicators of participants' extreme response styles have not been successful (Pargent,

2017), but similar strategies might still have an impact in different settings or in larger samples.

In the context of supervised ML, strategies to transform the original item responses into more meaningful indicators of psychological constructs would be considered feature engineering. In the last decade, *deep learning* (Goodfellow et al., 2016) has emerged based on the promise of creating models that can perform effective feature engineering automatically, when fed with huge quantities of data. So psychologists may benefit from collaborations with deep learning specialists to develop neural network architectures, specifically designed for the special properties of psychological questionnaire items. Unfortunately, most current datasets in psychology are much too small for deep learning models to be effective. However, the ML community has developed methods of *unsupervised learning* to find meaningful structures, based on large datasets in which the criterion variable of interest is not included. The structure is then transferred to models trained on the smaller datasets in which the criterion is available. This is a common strategy in computer vision or natural language processing, where networks trained on gigantic datasets can be reused in new applications (e.g. Devlin et al., 2018). In psychological research, Y. Wang and Kosinski (2018) were probably the first to use such a strategy, by training a neural network on a large dataset of facial images and employing the trained model in a second step to predict sexual orientation in a smaller, carefully collected lab dataset. The same method could also be applied to big datasets of traditional questionnaire data like the Big Five Project[11]or the Attitudes, Identities, and Individual Differences Study (Hussey et al., 2019).

Among other ML methods, deep auto-encoders could be a promising method for extracting a general structure of personality factors (see Goodfellow et al., 2016, for a description of the method and Liu & Zhu, 2016, for a rare psychological application of auto-encoders). Auto-encoders can be thought of as a highly non-linear variant of principal component analysis. The complex representations of personality dimensions resulting from such models could then be applied to smaller datasets, which include the same personality questionnaire on which the auto-encoder has been trained, but would additionally include new criterion variables of interest.

OUTLOOK AND CONCLUSION

In this article, we have discussed a number of important methodological challenges and highlighted some potential pitfalls that need to be considered in the application of ML models. Nevertheless, we are convinced that central ML concepts, such as resampling, out-of-sample error evaluation (e.g. via cross-validation), and methods of interpretable ML (e.g. ALE plots), can contribute to the robustness and generalizability of studies in personality psychology. In particular, we see two primary ways in which ML methods will play a decisive role in personality research and assessment in the near future.

First, ML methods will act as a useful addition to the researcher's toolbox of methods. Along with the advent of large, fine-grained datasets, ML will help researchers handle their complexity and high dimensionality. Unregularized linear models will quickly reach their limits due to factors such as multicollinearity, but more flexible models are capable of using complex data to make predictions. If evaluated correctly, ML methods can also show which variables provide the most predictive value, informing the development and validation of theories in personality psychology; methods of interpretable ML should play a particularly important role in this process. Using a process of continuous refinement, large numbers of digital and behavioural indicators could be used to predict a wide range of personality traits. The most predictive indicators could then be used after new data are collected to build an updated model, contributing to the creation of more cumulative knowledge in the discipline (Eisenberg et al., 2019). The use of ML models will also make it easier to compare new studies to research from other disciplines (assuming the precautions noted earlier heeded, such as correct categorization and stable information content of predictor variables). For example, a lot of work in the areas of human–computer interaction, computer science, and engineering have used ML models to investigate human behaviour in relation to the use of technology (Baeza-Yates et al., 2015; Eiband et al., 2019). Interdisciplinary research on personality could be vital to achieve technological breakthroughs with high societal impact.

Second, ML methods could allow insights from personality psychology to be translated to practical applications in a more reliable way. We have seen how cross-validated models can provide a more realistic estimate (in contrast to in-sample fit statistics) of how well the predictiveness of models is likely to generalize to new data; the high prevalence of cross-validation in industrial projects, where big money is lost if models do not actually perform and scale, might be another clue to its effectiveness. Hence, generalizing models (even with small effects) could increase the relevance of personality psychology in applied contexts. Relatedly, psychologists will be confronted with the situation that in practice, predictions can often be made without the availability of an explanation and beyond the context of an established theory (Yarkoni & Westfall, 2017). In other areas such as natural language processing, genetics, or bioinformatics, this practice has led to the successful development of models and indirectly to generating new scientific insights (Shmueli, 2010).

The usage of ML methods in psychological research is expected to increase sharply in the near future and cutting edge applications of ML will require collaborations with data scientists. So it will be necessary for researchers in personality psychology to equip themselves with both the terminology and the methodology of ML. At the same time, personality psychologists are well placed to play a decisive role in the prospective development of fair and understandable ML methodologies (Irving & Askell, 2019) that respect that personality constructs are latent variables. Knowledge of these methods will pave the way for a fruitful implementation of ML models in the field of psychological research and is set to lead to a better understanding of personality.

[11]https://www.thebigfiveproject.com.

REFERENCES

Aggarwal, C. C. (2016). *Recommender systems. [electronic book]: The textbook*. Cham:Springer.

Apley, D. W. (2016). *Visualizing the effects of predictor variables in black box supervised learning models*. Retrieved from https://arxiv.org/abs/1612.08468

Arendasy, M. (2009). *BFSI: Big-Five Struktur-Inventar (Test & Manual)*. Mödling: SCHUHFRIED GmbH.

Au, S. Q. (2019). *Fxtract: Feature extraction from grouped data*. Retrieved from. https://CRAN.R-project.org/package=fxtract

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Baena-Garcıa, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavalda, R., & Morales-Bueno, R. (2006). *Early drift detection method*. In *Fourth International Workshop on Knowledge Discovery from Data Streams*, *6*, Berlin, Germany, pp. 77–86.

Baeza-Yates, R., Jiang, D., Silvestri, F., & Harrison, B. (2015). *Predicting the next app that you are going to use*. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*,ACM Press, New York, New York, USA, pp. 285–294. https://doi.org/10.1145/2684822.2685302

Benson, M. J., & Campbell, J. P. (2007). To be, or not to be, linear: An expanded representation of personality and its relationship to leadership performance. *International Journal of Selection and Assessment*, *15*(2), 232–249. https://doi.org/10.1111/j.1468-2389.2007.00384.x

Binder, M. (2018). *MlrCPO: Composable preprocessing operators and pipelines for machine learning*. Retrieved from. https://CRAN.R-project.org/package=mlrCPO

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., & Jones, Z. M.(2016). Mlr: Machine learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5. Retrieved from https://jmlr.org/papers/v17/15-066.html

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, *20*(2), 249–275. https://doi.org/10.1162/EVCO_a_00069

Bleidorn, W., Hill, P., Back, M., Denissen, J. J. A., Hennecke, M., Hopwood, C., Jokela, M., …, & Roberts, B. (2019). The policy relevance of personality traits. *Am Psychol*, *74*, 1056–1067.

Bleidorn, W., & Hopwood, C. J. (2018). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203. https://doi.org/10.1177/1088868318772990

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70* (1), 159–181. https://doi.org/10.1111/bmsp.12086

Boerman, S. C., Kruikemeier, S., & Borgesius, F. J. Z. (2017). Online behavioral advertising: A literature review and research agenda. *Journal of Advertising*, *46*(3), 363–376. https://doi.org/10.1080/00913367.2017.1339368

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71. https://doi.org/10.1037/a0030001

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215. https://doi.org/10.2307/2676681

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley:University of California Press.

Bürkner, P.-C. (2019). *Bayesian item response modelling in R with brms and Stan*. arXiv Preprint arXiv:1905.09501.

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. In M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, & G. Ifrim (Eds.), *Machine learning and knowledge discovery in databases*. Cham:Springer International Publishing, pp. 655–670.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107. Retrieved from https://www.jmlr.org/papers/v11/cawley10a.html

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433–450. https://doi.org/10.1007/s00779-011-0490-1

Chouldechova, A., & Roth, A. (2018). *The frontiers of fairness in machine learning*. arXiv Preprint arXiv.1810.08810.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Gleser, G. C (1957). *Psychological tests and personnel decisions*.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Cucina, J. M., & Vasilopoulos, N. L. (2005). Nonlinear personality–performance relationships and the spurious moderating effects of traitedness. *Journal of Personality*, *73*(1), 227–260. https://doi.org/10.1111/j.1467-6494.2004.00309.x

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Towards Data Science. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

De Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). *Predicting personality using novel mobile phone-based metrics*. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*,Springer-Verlag, Berlin, Heidelberg, pp. 48–55. https://doi.org/10.1007/978-3-642-37210-0_6

Deckert, M. (2013). Incremental rule-based learners for handling concept drift: An overview. *Foundations of Computing and Decision Sciences*, *38*(1), 35–65. https://doi.org/10.2478/v10209-011-0020-y

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv Preprint arXiv.1810.04805.

Domingos, P., & Hulten, G. (2000). *Mining high-speed data streams*. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA. pp. 71–80. https://doi.org/10.1145/347090.347107

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv Preprint arXiv.1702.08608.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*, Vol. *5*:New York, NY, USA: Cambridge University Press.

Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019). *When people and algorithms meet: User-reported problems in intelligent everyday applications*. In *Proceedings of the 24th international conference on intelligent user interfaces*, ACM, New York, NY, USA. https://doi.org/10.1145/3301275.3302262

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., …, & Schwartz, H.A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203–11208. https://doi.org/10.1073/PNAS.1802331115

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), 2319. https://doi.org/10.1038/s41467-019-10301-1

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Fisher, A., Rudin, C., & Dominici, F. (2018). *Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective*. Retrieved from https://arxiv.org/abs/1801.01489

Fox, J. (1991). *Regression diagnostics: An introduction*. (Vol. 79). Sage.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://www.jstor.org/stable/2699986

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, *40*(1), 21–34. https://doi.org/10.1016/j.jrp.2005.08.003

Gama, J., & Castillo, G. (2006). Learning with local drift detection. In X. Li, O. R. Zaïane, & Z. Li (Eds.), *Advanced data mining and applications*. Berlin, Heidelberg:Springer Berlin Heidelberg, pp. 42–55.

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological Science*, *30*(7), 1087–1096. https://doi.org/10.1177/0956797619849435

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Gonfalonieri, A. (2019). *Why machine learning models degrade in production*. Towards Data Science. Retrieved from https://towardsdatascience.com/why-machine-learning-models-degrade-in-production-d0f2108e9214

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA, USA:MIT Press.

Goodman, B., & Flaxman, S. (2016). *European Union regulations on algorithmic decision-making and a "right to explanation"*. Retrieved from https://arxiv.org/abs/1606.08813

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*(1), 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Grothoff, C., & Porup, J. (2016). *The NSA's SKYNET program may be killing thousands of innocent people*. Towards Data Science. Retrieved from https://arstechnica.com/information-technology/2016/02/the-nsas-skynet-program-may-be-killing-thousands-of-innocent-people/

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Harari, G. M., Müller, S. R., Gosling, S. D., Harari, G. M., Müller, S. R., & Gosling, S. D. (2018). Naturalistic assessment of situations using mobile sensing methods, *The Oxford handbook of psychological situations*. Oxford, United Kingdom: Oxford University Presslocation. https://doi.org/10.1093/oxfordhb/9780190263348.013.14

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., …, & Gosling, S. D.(2019). Sensing sociability: Individual differences in young adults' conversation, calling, texting and app use behaviors in daily life. *Journal of Personality and Social Psychology.* https://doi.org/10.1037/pspp0000245

Harari, G, M., Vaid, S., Müller, S. R., Stachl, C., Marrero, Z., Schoedel, R., Bühner, M., & Gosling, S. D. (in press). *Personality Sensing for Theory Development and Assessment in the Digital Age*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning data mining, inference, and prediction* (2nd ed.). New York, NY, USA:Springer Science & Business Media. https://doi.org/10.1007/978-0-387-84858-7

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, *2019*, 1–22. https://doi.org/10.1155/2019/1306039

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity–bandwidth trade-off. *Journal of Organizational Behavior*, *17*(6), 627–637.

Hoppe, S., Loetscher, T., Morey, S. A., & Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, *12*, 105. https://doi.org/10.3389/fnhum.2018.00105

Hothorn, T., & Jung, H. H. (2014). RandomForest4Life: A random forest for predicting ALS disease progression. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, *15*(5-6), 444–452. https://doi.org/10.3109/21678421.2014.893361

Hu, R., & Pu, P. (2010). A study on user perception of personality-based recommender systems, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, Vol. *6075*, Berlin, Germany, EU:LNCS, pp. 291–302. https://doi.org/10.1007/978-3-642-13470-8_27

Hu, R., & Pu, P. (2011). *Enhancing collaborative filtering systems with personality information*. In *RecSys'11 - Proceedings of the 5th ACM Conference on Recommender Systems*, New York, NY, USA, pp. 197–204. https://doi.org/10.1145/2043932.2043969

Hussey, I., Hughes, S., Lai, C. K., Ebersole, C. R., Axt, J., & Nosek, B. A (2019). *The attitudes, identities, and individual differences (AIID) study and dataset*:OSF. https://doi.org/10.17605/OSF.IO/PCJWF

Ingold, D., & Soper, S. (2016). *Amazon doesn't consider the race of its customers. Should it?* Retrieved from https://www.bloomberg.com/graphics/2016-amazon-same-day/

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*, *4*, e14. https://doi.org/10.23915/distill.00014

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*(4), 243. https://doi.org/10.1037/h0045996

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511. https://doi.org/10.1016/j.jrp.2010.06.005

Jaeger, R. G., & Halliday, T. R. (1998). On confirmatory versus exploratory research. *Herpetologica*, *54*, S64–S66.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, *103*, 440. https://doi.org/10.1007/978-1-4614-7138-7

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138. https://doi.org/10.1177/0013164413498876

Katal, A., Wazid, M., & Goudar, R. H. (2013). *Big data: Issues, challenges, tools and good practices*. In *2013 6th International Conference on Contemporary Computing, IC3 2013*, Noida, India, pp. 404–409. https://doi.org/10.1109/IC3.2013.6612229

Klinkenberg, R. (2005). Meta-learning, model selection, and example selection in machine learning domains with concept drift. In M. Bauer, B. Brandherm, J. Fürnkranz, G. Grieser, A. Hotho, A. Jedlitschka, & A. Kröner (Eds.), *Lernen, Wissensentdeckung und Adaptivitat, LWA*, (pp. 164–171). Saarbrücken, Germany: DFKI.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In *Proceedings of the*

14th international joint conference on artificial intelligence, *2*, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp. 1137–1143. https://dl.acm.org/citation.cfm?id=1643031.1643047

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, *95*(3), 357–380. https://doi.org/10.1007/s10994-013-5415-y

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*,*60*(6), 84–90. https://doi.org/10.1145/3065386

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*, Vol. *26*. New York:Springer.

Kuhn, M., & Johnson, K. (2020). *Feature Engineering and Selection*.New York, NY, USA: Chapman and Hall/CRC. https://doi.org/10.1201/9781315108230

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, *578*(7793), 34–36. https://doi.org/10.1038/d41586-020-00274-3

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). *Delayed impact of fair machine learning*. arXiv Preprint arXiv.1803.04383.

Liu, X., & Zhu, T. (2016). Deep learning for constructing microblog behavior representation to identify social media user's personality. *PeerJ Computer Science*, *e81*, 2. https://doi.org/10.7717/peerj-cs.81

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13* (5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mahmoodi, J., Leckelt, M., van Zalk, M., Geukes, K., & Back, M. (2017). Big Data approaches in social and behavioral science: Four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, *18*, 57–62. https://doi.org/10.1016/J.COBEHA.2017.07.001

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, *32*(2), 74–79. https://doi.org/10.1109/MIS.2017.23

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114

McFatter, R. M. (1979). The use of structural equation models in interpreting regression equations including suppressor and enhancer variables. *Applied Psychological Measurement*, *3*(1), 123–135. https://doi.org/10.1177/014662167900300113

Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2020). Recent trends in deep learning based personality detection. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-019-09770-z

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. https://doi.org/10.1177/1745691612441215

Molnar, C. (2019). *Interpretable machine learning (1st ed., p. 318)*. Retrieved from https://christophm.github.io/interpretable-ml-book/

Molnar, C., Bischl, B., & Casalicchio, G. (2018). iml: An R package for interpretable machine learning. *JOSS*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Molnar, C., Casalicchio, G., & Bischl, B (2019). *Quantifying interpretability of arbitrary machine learning models through functional decomposition*. Retrieved from https://arxiv.org/abs/1904.03867

Mønsted, B., Mollgaard, A., & Mathiesen, J. (2018). Phone-based metric as a predictor for basic personality traits. *Journal of Research in Personality*, *74*, 16–22. https://doi.org/10.1016/J.JRP.2017.12.004

Montag, C., Blaszkiewicz, K., Sariyska, R., Lachmann, B., Andone, I., Trendafilov, B., …, & Markowetz, A. (2015). Smartphone usage in the 21st century: Who is active on WhatsApp? *BMC Research Notes*, *8*(1), 331. https://doi.org/10.1186/s13104-015-1280-z

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261–286.https://doi.org/10.1111/j.1467-6494.2008.00545.x

Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., & Rentfrow, J. (2018). Musical preferences predict personality: Evidence from active listening and Facebook Likes. *Psychological Science*, *29*(7), 1145–1158. https://doi.org/10.1177/0956797618761659

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(1), 110. https://doi.org/10.1186/1471-2105-11-110

Pargent, F. (2017). *Detection, avoidance, and compensation—Three studies on extreme response style*,Doctoral thesis, Ludwig-Maximilians-Universität München, Germany, EU, Retrieved from https://nbn-resolving.de/urn:nbn:de:bvb:19-211562

Pargent, F., & Albert-von der Gönna, J. (2018). Predictive modeling with psychological panel data. *Zeitschrift Für Psychologie*, *226*(4), 246–258. https://doi.org/10.1027/2151-2604/a000343

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., …, & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., &Sohl-Dickstein, J., (2015). *Deep Knowledge Tracing*. Proceedings of the 28th International Conference on Neural Information Processing Systems -Volume *1*, pp. 505–51. Cambridge, MA, USA: MIT Press. https://papers.nips.cc/paper/5654-deep-knowledge-tracing.pdf

Randler, C., Schredl, M., & Göritz, A. S. (2017). Chronotype, sleep behavior, and the big five personality factors. *SAGE Open*, *7*(3), 1–9. https://doi.org/10.1177/2158244017728321

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality*, *29*(3), 363–381. https://doi.org/10.1002/per.1994

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Retrieved from https://arxiv.org/abs/1602.04938

Roque, N., & Ram, N. (2019). tsfeaturex: An R package for automating time series feature extraction. *Journal of Open Source Software*, *4*(37), 1279. https://doi.org/10.21105/joss.01279

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience*, *6*(5), 1–9. https://doi.org/10.1093/gigascience/gix019

Schmidt-Atzert, L., Künecke, J., & Zimmermann, J. (2019). TBS-DTK-Rezension: PRECIRE JobFit. *Psychologische Rundschau*, *70*(4), 299–301. https://doi.org/10.1026/0033-3042/a000459

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits:

Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, *38*(2), 173–212. https://doi.org/10.1177/0022022106297299

Schneider, H., Schauer, K., Stachl, C., & Butz, A. (2017). Your data, your vis: Personalizing personal data visualizations, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, Vol. *10515 LNCS*. Cham:Springer, pp. 374–392. https://doi.org/10.1007/978-3-319-67687-6_25

Schoedel, R., Au, Q., Völkel, S. T., Lehmann, F., Becker, D., Bühner, M., …, & Stachl, C. (2018). Digital footprints of sensation seeking. *Zeitschrift Für Psychologie*, *226*(4), 232–245. https://doi.org/10.1027/2151-2604/a000342

Schoedel, R., Pargent, F., Au, Q., Völkel, S. T., Schuwerk, T., Bühner, M., & Stachl, C. (in press). *To Challenge the Morning Lark and the Night Owl: Using Smartphone Sensing Data to Investigate Day-Night Behavior Patterns*.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., …, & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., …, & Young, M. (2014). *Machine learning: The high interest credit card of technical debt*. In *SE4ML: Software Engineering for Machine Learning (Nips 2014 Workshop)*.

Sculley, D., Snoek, J., Wiltschko, A., & Rahimi, A (2018). *Winner's curse? On pace, progress, and empirical rigor*. ICLR.

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, *32*(3), 186–201. https://doi.org/10.1002/per.2147

Settanni, M., Azucar, D., & Marengo, D. (2018). Predicting individual characteristics from digital traces on social media: A meta-analysis. *Cyberpsychology, Behavior, and Social Networking*, *21*(4), 217–228. https://doi.org/10.1089/cyber.2017.0384

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25* (3), 289–310. https://doi.org/10.1214/10-STS330

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *100* (3/4), 441–471. Retrieved from https://www.jstor.org/stable/1422689

Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., …, & Bühner, M. (2019). *Behavioral patterns in smartphone usage predict big five personality traits. OSF*. https://doi.org/10.17605/OSF.IO/KQJHR

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., …, & Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701–722. https://doi.org/10.1002/per.2113

Steyer, R. (2001). Classical (Psychometric) Test Theory. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences*. (Second Edition, pp. 785–791). https://doi.org/10.1016/B978-0-08-097086-8.44006-7

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*. *56*(5), 44–54

Tkalcic, M., Carolis, B. D., Gemmis, M. D., Odi, A., & Košir, A. (2016). In Tkalčič, M., De Carolis, B., de Gemmis, M., Odić, A., & Košir, A. (Eds.), *Emotions and personality in personalized services*. Switzerland:Springer International Publishing, pp. 3–11. https://doi.org/10.1007/978-3-319-31413-6

Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., …, & Mohamed, S.(2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, *572*(7767), 116–119. https://doi.org/10.1038/s41586-019-1390-1

Tutz, G., Schauberger, G., & Berger, M. (2018). Response styles in the partial credit model. *Applied Psychological Measurement*, *42* (6), 407–427. https://doi.org/10.1177/0146621617748322

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91

Völkel, S. T., Schödel, R., Buschek, D., Stachl, C., Au, Q., Bischl, B., …, & Hussmann, H. (2019). Opportunities and challenges of utilizing personality traits for personalization in HCI: Towards a shared perspective from HCI and psychology, *Personalized human-computer interaction*. (pp. 31–63). Oldenbourg, Germany:De Gruyter.

Wang, W., Harari, G. M., Wang, R., Müller, S. R., Mirjafari, S., Masaba, K., & Campbell, A. T.(2018). Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–21. https://doi.org/10.1145/3264951

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257. https://doi.org/10.1037/pspa0000098

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*(3), 279–291. https://doi.org/10.1177/1073191115583714

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(3), 203–220. https://doi.org/10.1177/1745691612442904

Wu, W., Chen, L., & He, L. (2013). *Using personality to adjust diversity in recommender systems*. In *HT 2013 - Proceedings of the 24th ACM Conference on Hypertext and Social Media*, Paris, France, pp. 225–229. https://doi.org/10.1145/2481492.2481521

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yeung, C.-K. (2019). *Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory*. arXiv preprint arXiv:1904.11738.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112* (4), 1036–040. https://doi.org/10.1073/pnas.1418680112

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports. *Journal of Personality*, *84*(4), 461–472. https://doi.org/10.1111/jopy.12172

Zweck, B. M., Pargent, F., & Bühner, M. (2019). *Prediction of health-related outcomes and turnover intention with the Munich employee health questionnaire (MEHQ). PsyArXiv*. https://doi.org/10.31234/osf.io/bgu6m