Using K-means Clustering for Out-of-Sample Predictions of Memory Retention

Florian Sense
f.sense@rug.nl / floriansense@gmail.com
InfiniteTactics, LLC, Beavercreek, OH, USA
Department of Experimental Psychology & Behavioral and Cognitive Neuroscience
University of Groningen, Groningen, The Netherlands

Michael Collins
collins.283@wright.edu
ORISE at Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Michael Krusmark, **Tiffany S. Jastrzembski**
{michael.krusmark.ctr, tiffany.jastrzembski}@us.af.mil
Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Abstract

In applied settings, computational models of memory have proven useful in making principled performance predictions. Specifically, historical data are used to derive model parameters in order to enable out-of-sample predictions. Parameters are typically fit to meaningful subsets of data. However, labels that demarcate what constitutes a "meaningful" subset are not always available. Here, we utilize a data-driven method to cluster past performance into subsets possessing statistical similarities. We contrast predictions from cluster-specific model parameters with predictions based on subsets that are artifacts of the experimental design. We show that cluster-based predictions are at least as accurate as the chosen baselines and highlight additional advantages of the data-driven approach.

Keywords: learning; memory; k-means clustering; computational model; prediction

Introduction

Psychological models of learning and forgetting have largely focused on explaining rather than predicting (Yarkoni & Westfall, 2017). That is, models are evaluated with regard to how well they fit data in various experimentally controlled conditions. To leverage such models' capabilities in applied settings, however, their ability to accurately predict new observations is paramount. One standard method for evaluating such predictions is to derive model parameters from historical data and apply them to new data (Shmueli, 2010).

Here, we focus on predicting individual, item-level responses. Deriving model parameters through standard fits to historical data raises the question of what data should the model be fit to optimize predictive validity? One method for parsing the data lies with estimating a unique set of parameters for each user's exposures to an item (i.e., a *trajectory*) to afford maximum flexibility in the model. This method comes at the cost of potentially overfitting the individual's performance and hindering out-of-sample predictive ability. Conversely, estimating a single set of parameters across all trajectories protects against overfitting the sample, but forfeits the ability to utilize more nuanced differences between trajectories that may be lost during data aggregation. Consequently, the ideal approach depends on identifying meaningful subsets of data from which parameters can be obtained that produce valid out-of-sample prediction.

As alluded to previously, there are various ways to parse historical data into meaningful subsets. In scientific contexts, natural subsets stem from experimental conditions. Similarly, data can be segmented at the level of the individual user. For the purpose of making predictions, the assumption for such segmentation is that parameters describing the average behavior either within a condition or for an individual user, can be used to predict performance on future observations. In naturalistic contexts, however, condition labels may be absent, vague, or inconsistent. Thus, the creation of subsets becomes a problem of data-driven dimensionality reduction. K-means clustering has been used with success to reduce high-dimensional skill representations, and has allowed for increased computational efficiency and enhanced interpretation (e.g., Ritter et al., 2009). Such data-driven methods detect patterns in the historical data, and the number and content of meaningful subsets are prescribed by the data themselves. Predictions are made under the assumption that homogeneous performance profiles in the historical data are likely to be similar in out-of-sample data.

In the current work, we assess the predictive capabilities of a computational process model in a standard paired-associates learning task that imposed strict experimental control on the repetition schedules. Specifically, we contrast the quality of predictions made using (a) reliable labels that are artifacts of the experimental design ("condition" and

"user"), or (b) a data-driven clustering method in which the subsets are derived directly from accuracy in the historical data.

Methods

Data

The data come from a multi-session paired-associate learning task in which 61 participants studied Japanese-English word pairs according to six tightly regimented schedules. An overview of the aggregate performance is provided in Figure 1. The six experimental conditions result from crossing two inter-trial-intervals (ITI; either 2 or 11 intervening trials) and three inter-session-intervals (ISI) between the first and second session that could be 0, 7, or 14 days—the 0-day ISI was a break of approximately five minutes between sessions. Both ITI and ISI were manipulated within-subject and participants studied five unique word-pairs in each condition (30 items total per user). The ISI between session two and three was always seven (plus/minus two) days.

The first repetition showed both the Japanese (cue) and English (response) word on screen. Participants typed the response to proceed. Accuracy for this response was set to 0 (in Figure 1 and reported analyses) to reflect the lack of prior knowledge of Japanese, which participants were screened for. All subsequent repetitions only showed the cue and were followed by corrective feedback.

As seen in Figure 1, clear differences in the aggregate performance between the experimental conditions were observed. Acquisition is markedly better for short ITI word-pairs in the first session but long-term retention (repetition 21) is substantially better for long ITI items, especially at longer ISIs.

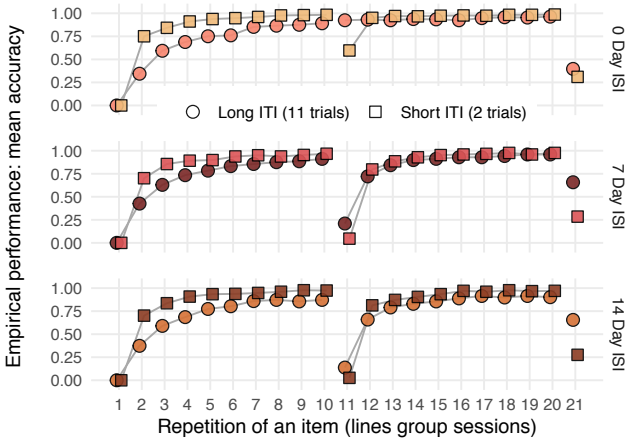

Figure 1. Mean performance at each repetition for the six experimental conditions. ITI = inter-trial-interval; ISI = inter-session-interval between first and second session.

The Predictive Performance Equation (PPE)

Given space limitations in the current format, we present a condensed overview of PPE's mechanisms and refer the interested reader to the extensive description in Walsh et al. (2018). Figure 2 summarizes how performance can be predicted (P) using timing information (t_i) in conjunction with four free parameters (m, b, τ, and s). To fit the model to empirical data—that is, to derive the best-fitting free parameters—two pieces of information are required: a timestamp and a performance metric. The best-fitting parameters are found by minimizing the sum-of-squares error between PPE's P and empirical performance.

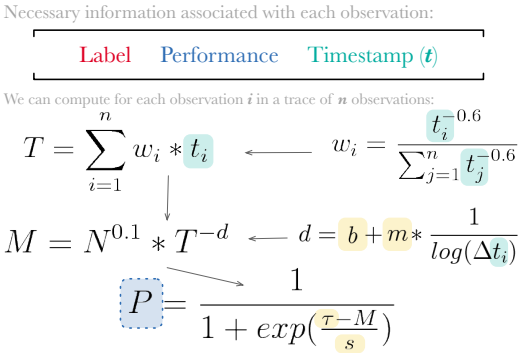

Figure 2. Predictive Performance Equation (PPE) mechanism and required information.

In practice, parameters are estimated for a relevant subset of data, which necessitates a third piece of information to be associated with each observation: a "label" that demarcates the subsets. The current work focusses on choosing an appropriate "label" that segments the data such that extracted parameters afford valid out-of-sample predictions. The following section will detail five approaches to utilizing existing or data-driven labels to predict performance.

Procedure

The data were first split into a *training* and a *test* set using an 80/20 split that ensured 80% of an individual user's trajectories were assigned to the training set (Yarkoni & Westfall, 2017). Parameters were extracted from the training set and predictions were generated for the test set. In the following, we will outline five approaches that differed in how they segmented the training data in order to derive PPE parameters (which are in turn used to generate predictions). They fall into two categories: *cluster-based* and *control* approaches. The former utilizes data-driven clustering to partition the data, while the latter relies on labels available as meta-data: user IDs and experimental conditions.

Cluster-based approaches All trajectories in the training data were treated in complete isolation (i.e., independent of the user or experimental condition they were associated with) and subjected to the k-means clustering algorithm (Hartigan & Wong, 1979). This approach originates from the signal processing literature, and aims to divide a set of observations into distinct clusters, such that individual observations belong to the cluster possessing the nearest mean. We settled on an eight-cluster solution because (a) the decrease in summed within-cluster sum of squares reached an asymptote at eight cluster centers, and (b) inspection of the cluster centroids

suggested that trajectories were grouped into behaviorally meaningful clusters (see Figure 3).

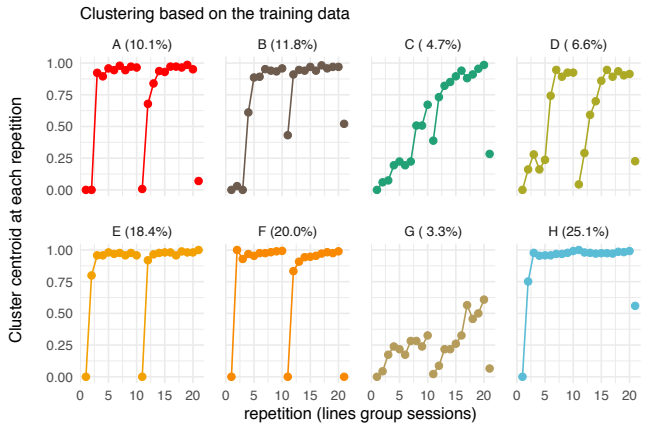

Figure 3. Cluster centroids at each repetition for the eight clusters derived from the training data. Panels' headings show the percentage of trajectories assigned to the cluster.

We then fit PPE to all trajectories in each cluster. This resulted in a unique set of the four free parameters for each data-derived cluster. Next, the trajectories in the test data were assigned to one of the clusters in Figure 3. Specifically, the root-mean-squared-error was computed between a trajectory—a vector of accuracy data—and each cluster's centroids—shown in Figure 3—and the trajectory was assigned to the cluster with the lowest RMSE. Lastly, using the PPE parameters associated with the cluster a trajectory was assigned to, predictions were generated for all trajectories in the test data.

In the Results section we will focus on evaluating the predictions for the 21st repetition specifically. Therefore, when assigning new trajectories to pre-defined clusters, we emulated a realistic scenario and assumed that only the first 20 repetitions of a trajectory were available when it must be assigned to a cluster.

The **cautious cluster-based approach** worked in the following way. Given that the last/withheld response could either be incorrect or correct, the cluster assignment was run twice; first, assuming that the incomplete trajectory would be completed with a final incorrect response, and next, assuming it would be completed with a final correct response. If the trajectory was assigned to the same cluster independent of the assumed last response, PPE predictions were generated using the parameters associated with the unambiguously assigned cluster. If the incomplete trajectory was assigned to different clusters depending on the assumed final response, two predictions were generated using the PPE parameters of the two assigned clusters. Those two predictions were averaged to yield a single, "cautious" prediction.

For comparison purposes, we additionally implemented two baseline approaches; a **blind cluster-based approach** in which only the incomplete trajectory (i.e., repetitions 1-20) was used for assignment to one of the eight clusters, and an **omniscient cluster-based approach** in which the complete trajectory was used as if the final repetition had already been observed. The *omniscient* approach mimicked the control approaches in the sense that it provided reliable "labels" for the membership of a trajectory (i.e., we know which experimental condition and which user a trajectory belongs to) and constituted a form of intentional data leakage.

Control approaches As outlined above, we compared the cluster-based approaches with two control approaches. Using the same training/test split of the data as described in the previous section, a unique set of PPE parameters was derived for (a) each experimental condition, and (b) each user. Therefore, the information available for a single condition or user provided the general data pattern for PPE equations to estimate its best-fitting parameters. The procedure was otherwise the same as for the cluster-based approaches. Using the *condition*'s/*user*'s PPE parameter estimates, a prediction was made for each trajectory in the test data.

Results

Before detailing and comparing the results for the control and cluster-based approaches, we should establish that the clusters derived from the training data (Figure 3) did not merely mimic the experimental conditions. Figure 4 verifies that this was not the case. However, some clusters were preferentially assigned to certain experimental conditions. For example, cluster E mostly covered ISIs of one or two weeks, cluster F was mostly assigned to Short ITI conditions, and cluster H was mainly assigned to the two 0 Day ISI conditions. However, most clusters were spread across (almost) all experimental conditions—no cluster exclusively mapped onto a single condition. This non-exclusivity corroborates that the data-driven clustering leveraged patterns in the data that were not directly dictated by the features of the experimental conditions. These differences in patterns between Figure 1 and Figure 3 highlights that the clustering algorithm isolates distinct performance profiles that are washed out by aggregating performance within a condition (clusters B, C, D, and G especially), potentially accounting for meaningful psychological variables.

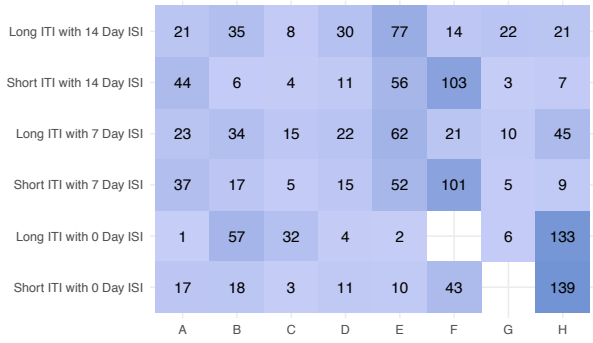

Figure 4. The overlap between clusters and experimental conditions is shown by tallying how frequently a trajectory from a given condition was assigned to each cluster.

Control approaches

The first control approach uses the *condition*-constrained PPE parameters to make predictions for the test data. These predictions are shown for the 21st repetition in Figure 5. Experimental conditions are plotted against performance, where the gray dots indicate the actual accuracy on the final repetition and the colored dots show the model's predictions. Observed accuracy was jittered both horizontally and vertically to create a visual impression of the relative number of responses in each condition. The predicted performance was only jittered horizontally to preserve the precise values (the same applies to Figure 8 and Figure 9).

The *condition*-based predictions shown in Figure 5 were highly homogenous within a condition. Visual comparison with the final repetition in Figure 1 suggests that predictions closely matched the aggregate empirical data. Closer inspection—a detailed description of which is beyond the scope of the current report—confirmed that PPE fit the six empirical trajectories in Figure 1 very well. The low variance in predicted values can probably be attributed to the very similar temporal inputs. As described above, PPE requires two inputs and one of those—the timestamps—should be very alike within each condition. No matter what the exact values of the second input to PPE (i.e., accuracy), the predictions are largely identical within a condition. As a result, the *condition*-based predictions resemble the empirical means but are rather non-committal (that is, closer to 0.5 than either 0 or 1).

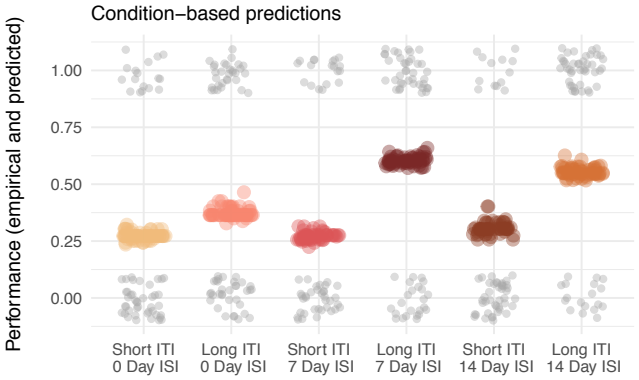

Figure 5. Predicted performance (colored) compared with actual performance (gray) for the *condition*-based approach.

The user-based predictions are shown in Figure 6. Predicted performance (colors and shapes matching Figure 1) is contrasted with actual accuracy (gray; no jitter) for each user. Users are ordered on the x-axis according to their actual performance (overall number correct; ordering is consistent across panels).

Two patterns surface in Figure 6. First, the *user*-based predictions showed much larger variability than the *condition*-based predictions: Scanning horizontally across the figure's panels indicates that predictions are scattered across most of the y-axis in every condition (between-user variability) and scanning vertically, we see that both low and high performers—at the left and right ends of the x-axes, respectively—exhibit substantial within-user variability. This is likely because the per-user set of PPE parameters is used to generate predictions for new trajectories of varying schedules. This is in stark contrast with the *condition*-based approach, in which all trajectories within a condition have very homogeneous temporal structures. Second, the predictions steadily increase from the left to the right side of Figure 6. This implies a correlation between predicted and observed performance that is confirmed by a significant positive point-biserial correlation ($r = 0.405$, $t(357) = 8.361$, $p < 0.001$) computed across all conditions.

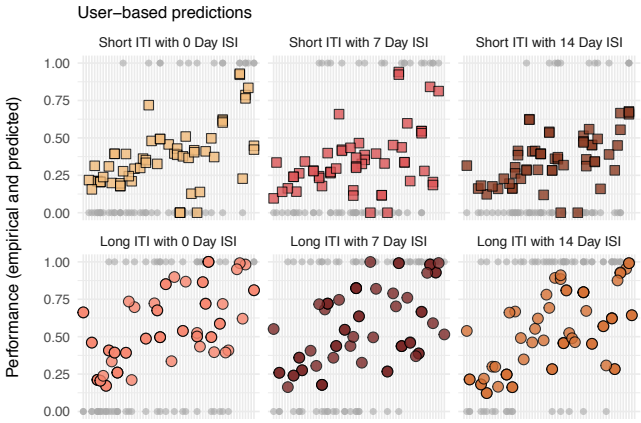

Figure 6. Predicted performance (colored) compared with actual performance (gray) for the *user*-based approach.

Cluster-based approaches

A central step in the cluster-based approaches is the assignment of trajectories from the test data to one of the clusters derived from the training data (see Figure 3). The only difference between the *omniscient* and *blind* approaches is whether they did or did not (respectively) take the last repetition into account when assigning a trajectory to a cluster. The confusion matrix in Figure 7 shows that under these different assumptions, the same trajectory can be assigned to different clusters. Most confusions occur between clusters A and E as well as F and E—specifically, the *blind* approach preferentially assigns trajectories to either A or F, many of which are assigned to cluster E by the *omniscient* approach. On the other hand, all assignments to clusters G and H are unambiguous—no matter whether the last observation is known or not. The confusion matrix suggests that ambiguous assignments in the *cautious* approach will primarily arise if a trajectory most closely matches cluster A, E, or F (a pattern confirmed in Figure 9).

The *blind* cluster-based predictions (not shown) look very similar to the *omniscient* predictions (Figure 8). While the *condition*-constrained predictions tightly center around the empirical means, the *blind* and *omniscient* cluster-based predictions emulate a cluster's centroid on the last repetition but exhibit higher variance—likely because the same model

parameters are used across trajectories that have very different temporal structures (cf. Figure 4).

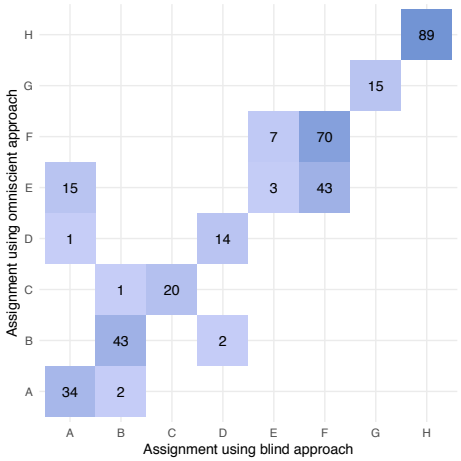

Figure 7. Confusion matrix comparing assignment of new trajectories for the *blind* and *omniscient* approaches.

Consequently, the primary difference between the *blind* and *omniscient* approach is a result of the differential assignments apparent in Figure 7: The *blind* approach assigned 113 trajectories to cluster F, while the *omniscient* approach only assigned 77 trajectories to the same cluster (7 of which the *blind* approach assigned to cluster E). Conversely, the *omniscient* approach assigned 60 trajectories to cluster E, only 3 of which were also assigned to cluster E by the *blind* approach. The remaining omnisciently assigned E trajectories were blindly assigned to either A or F.

One stark difference between the *omniscient* cluster-based predictions and the *condition*-based predictions is that the former commits to more extreme predictions: In clusters A, F, and G, for example, performance is predicted to be poor with great certainty for virtually all observations in that cluster (and the reverse is true for cluster E; Figure 8). For clusters whose centroid on the 21st repetition is at neither boundary (e.g., clusters B, D, and H), on the other hand, predicted performance is more spread.

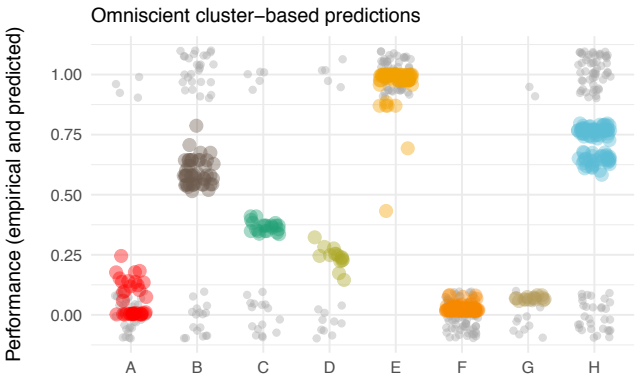

Figure 8. Predicted performance (colored) compared with actual performance (gray) for the *omniscient* cluster-based approach.

For the *cautious* cluster-based approach, the equivalent of Figure 5 and Figure 8 is more complicated because a given trajectory does not necessarily have a single cluster associated with it. As explained in the Methods, each trajectory is assigned to a cluster twice; once assuming the last response will be incorrect and once assuming it will be correct. In 52.1% of the trajectories, this resulted in the same assigned cluster, as revealed in the lower panel of Figure 9. Not surprisingly, for trajectories that were *cautiously* assigned to the same clusters, the pattern is very similar to the *omniscient* approach (cf. Figure 8). The upper panel in Figure 9 shows the trajectories that were assigned to different trajectories depending on the assumed last response. A data point's location in the upper panel is determined by the cluster assigned when the last response is assumed to be incorrect, while its color is determined by the cluster assigned if the last response is assumed to be correct.

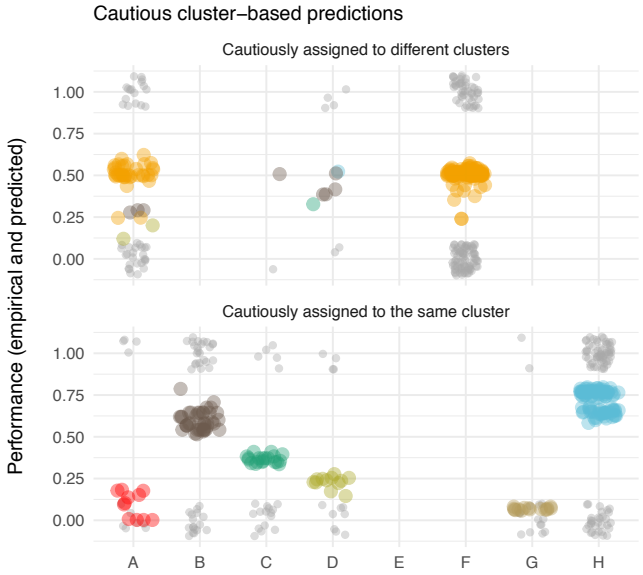

Figure 9. Predicted performance (colored) compared with actual performance (gray) for the *cautious* cluster-based approach.

Several interesting patterns emanate from Figure 9. First, not a single trajectory has been assigned to cluster E, independent of whether the last response was assumed to be correct or not. Second, not a single trajectory was unambiguously assigned to cluster F—however, the majority of ambiguous assignments (71.5%) are a conflict between assignment to cluster F (assuming the last repetition is incorrect) and cluster E (assuming it is correct). Similarly, 21.5% of ambiguous assignments are to clusters A and E. This confirms what we derived from the confusion matrix: Trajectories most confused by *blind* and *omniscient* assignments will present the largest challenge to the *cautious* approach. Third, in these cases of differential A-E and F-E assignments, the resulting predictions (orange points in upper panel of Figure 9) center around 0.5 because cluster A and F

predict near-floor and cluster E predicts near-ceiling performance, which produces a *cautiously* uncertain prediction of roughly 0.5.

In an attempt to summarize and compare the predictions made for the 21st repetition in the test data, Table 1 lists three model fit statistics for each of the five approaches. The area under the ROC curve (AUC; Fawcett, 2006) can be interpreted as the probability that the predicted performance will be ranked higher for a randomly chosen correct response than a randomly chosen incorrect response. This probabilistic interpretation highlights that the AUC is based on the relative rank of predictions rather than the distance from the "truth." The root-mean-squared-error (RMSE), on the other hand, quantifies the absolute distance between prediction and "truth." The logarithmic loss is expressed on an open-ended scale and imposes a harsh penalty on incorrect predictions. Each of these statistics emphasizes different dimensions of the quality of an approach's predictions and condense the nuanced effects discussed above into a single number for easy comparison.

Table 1. Model fit statistics for repetition 21 in the test data.

Approach	AUC†	RMSE‡	Log loss‡
Condition	0.676	0.475	15.5
User	0.736	0.457	13.7
Blind clusters	0.655	0.536	14.2
Omniscient clusters	0.902	0.351	8.3
Cautious clusters	0.681	0.485	15.6

Discussion

The main goal of the current work was to explore a data-driven method to segment performance profiles for purposes of estimating model parameters to produce valid out-of-sample predictions. Out-of-sample predictions generated from variations of data-driven clustering methods were contrasted with out-of-sample predictions generated from fits to each subset indicated by reliable labels (experimental conditions and users). While summary fit statistics suggest that the condition and clustering approaches performed approximately equally well (Table 1), we believe that the clustering-based approach provides a more nuanced picture that is revealed only when the predictions are inspected closely (Figure 5 through Figure 9).

The data used here were highly structured. This has two consequences relevant for the current work. First, the condition labels provided a great deal of implicit information since the experimental manipulation created temporally distinct repetition schedules. This temporal homogeneity in the trajectories associated with each *condition* resulted in largely invariant predictions for the 21st repetition (Figure 5). Conversely, grouping by user, we observed that predictions varied significantly within a user but there was a positive correlation between predicted and observed accuracy. Both the *condition* and *user* approach results suggest that fitting PPE to a subset of data works as intended and PPE capitalizes on information that is coherent within a given subset (i.e., the temporal regularities within each *condition* or the ability of a *user* across various timings).

The second consequence of using these highly structured data pertains to the fixed number of repetitions. Knowing that trajectories from the test data would inevitably match trajectories in the training data in terms of their length allowed us to extract a single set of clusters (Figure 3). In naturalistic data, trajectories will undoubtedly vary in length, potentially complicating the derivation of clusters and the assignment of new trajectories to clusters. If the dataset of interest is large enough, a potential solution would be to bin trajectories by their lengths and derive clusters for each bin. For example, the clusters in Figure 3 could be used for trajectories that contain roughly 20 observations, and another set of clusters could be derived and referenced for trajectories that contain >30 or <10 observations.

The exploration of the data-driven *cautious* clustering approach presented here has both advantages and disadvantages related to its assumption that the to-be-predicted response in a trajectory could be either correct or incorrect. One downside is that if this assumption leads to the same trajectory being assigned to different clusters (top panel Figure 9) the resulting *cautious* prediction is essentially always a non-committal 50%. This might be partially due to the specific data used here, which yielded two highly confusable clusters that make opposite predictions for the final repetition (E and F, see Figure 3). In this scenario, a non-committal prediction could be considered sensible. One advantage of the *cautious* approach is that in the reverse scenario, unambiguous assignment independent of the anticipated response, predictions are fairly accurate and more confident than in the *condition* approach.

An additional advantage of clustering the data from an experimental study is a descriptive overview that adds nuance to other forms of aggregation. Here, for example, each *trajectory* was treated in complete isolation, independent of the user or experimental condition it was associated with. Figure 3 suggests that the majority of trajectories are assigned to clusters with very high performance that mostly differ on repetitions 11 and 21 (cf. clusters E, F, and H) but that there are also clusters that capture slower (clusters B, C, D) or incomplete learning (cluster G). In conjunction with the mapping provided in Figure 4 (and its equivalent for user-cluster mappings), the clusters constitute an especially useful a descriptive tool. It is possible that distinct clusters map onto psychological variables (mnemonic strategies, fatigue, etc.) outside the scope of the conducted study. Another potential extension of the current work would stem from utilizing alternative clustering approaches (see Berkhin, 2006 for an extensive survey).

Due to space constraints, we did not report in more detail the fit to the first 20 repetitions and instead chose to highlight the predictive accuracy on the 21st repetition. Given the train/test split procedure, however, the model actually predicted *all* observations in the test set. A closer

investigation of the fit to the first 20 repetitions—especially the expected dip in performance on repetition 11—would be a natural extension of the current work. Similarly, an in-depth analysis and discussion of the estimated model parameters should prove productive.

We believe the data-driven clustering approach presented here has utility in applied scenarios for which theoretical assumptions about meaningful subsets of data are hard to make or necessary meta-data are unavailable (similar to, for example, Ayers, Nugent, & Dean, 2008). Additionally, clustering reveals patterns in the data obscured by aggregation along conventional dimensions (e.g., Figure 1 masks the information in Figure 3, particularly if Figure 4 is also considered). Therefore, the procedure outlined here should be repeated and refined across both experimental and naturalistic datasets in order to better isolate settings for fruitful application.

References

Ayers, E., Nugent, R., & Dean, N. (2008). Skill set profile clustering based on student capability vectors computed from online tutoring data. Proceedings of the 1st International Conference on Educational Data Mining, Montreal, Canada.

Berkhin, P. (2006). A survey of clustering data mining techniques. *In Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Hartigan, J. A. & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R. C., & Towle, B. (2009). Reducing the Knowledge Tracing Space. *International Working Group on Educational Data Mining*.

Shmueli, G. (2010). To explain or to predict?. *Statistical Science*, *25*(3), 289-310.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive science*, *42*, 644-691.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.