Recommendation as Generalization: Evaluating Cognitive Models In the Wild

David D. Bourgin (ddbourgin@berkeley.edu),
Joshua T. Abbott (joshua.abbott@berkeley.edu),
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA, 94720

Abstract

The explosion of data generated during human interactions online presents an opportunity for cognitive scientists to evaluate their models on popular real-world tasks outside the confines of the laboratory. We demonstrate this approach by evaluating two cognitive models of generalization against two machine learning approaches to recommendation on an online dataset of over 100K human playlist selections. Across two experiments we demonstrate that a model from cognitive science can both be efficiently implemented at scale and can capture generalization trends in human recommendation judgments which neither machine learning model is capable of replicating. We use these results to illustrate the opportunity internet-scale datasets offer to cognitive scientists, as well as to underscore the importance of using insights from cognitive modeling to supplement the standard predictive-analytic approach taken by many existing machine learning approaches.

Keywords: cognitive modeling; recommender systems; big data

Introduction

Every day, terabytes of behavioral data are generated as people go about their daily lives online. Although many of these computer-mediated interactions differ from the tightly-controlled in-laboratory experiments common in psychological research, they offer insight into many of the same cognitive phenomena, often at a scale that would make even the most dutiful experimentalist blush. This new source of high throughput behavioral data, already the lifeblood of machine learning and AI researchers around the world, offers cognitive scientists a similar opportunity to evaluate cognitive models in the wild, at scale, with minimal investment. Moreover, the dearth of cognitive modeling techniques in current approaches to analyzing these behavioral datasets suggests an opportunity to re-establish the value of modeling minds as mediating influences on behavior in a domain that has been *de facto* colonized by computer science.

Product recommendation is a notable example of an applied task which can serve as a good test-bed for models from cognitive science. Not only is recommendation a prominent example of a fundamentally psychological task which has been translated into machine learning terms, but it also has become near ubiquitous in our daily interactions online. Indeed, automated approaches to recommendation have become one of the more prominent examples of the ways in which machine learning systems augment everyday human decision making.

Despite this enormous influence, however, studies from the recommendation system literature indicate that users are sensitive to the differences between human and automated recommendations, often favoring recommendations from other people. In two well-known papers, R. R. Sinha and Swearingen (2001) and S. Sinha, Rashmi, and Sinha (2001) report that on average users tend to prefer recommendations made by friends to those generated by automated recommendation systems, even if the identity of the recommender is not revealed. This general preference for human recommendations is not limited to close associates, either: Krishnan, Narayanashetty, Nathan, Davies, and Konstan (2008) report that even complete strangers are capable of outperforming automated recommendation systems for atypical user profiles. Human users appear particularly sensitive to algorithmic intervention in subjective domains: Logg (2017) reports that participants preferred recommendations of human experts to those of algorithms for subjective decisions, regardless of the domain. This finding is corroborated by Yeomans, Shah, Mullainathan, and Kleinberg (2017) who found that although automated recommendation systems often show superior empirical performance on a variety of information retrieval metrics, the majority of human users still prefer the recommendations of human experts in the domain of joke recommendation. In light of such findings, a natural question is whether we can identify systematic ways in which human and algorithmic recommendations deviate from one another, and if so, whether models designed to explicitly account for human cognition can help bridge this gap.

In the current paper, we underscore the potential online recommendation datasets have as cognitive test-beds by evaluating a version of the well-known Bayesian model of generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001) on hundreds of thousands of human judgments collected from a popular playlist-sharing website. We further illustrate the value that cognitive modeling techniques play in this new world by comparing the performance of the generalization model with that of two widely used machine learning approaches to recommendation. We conclude with a discussion of these results in the context of reintroducing principles of cognition into modern machine learning frameworks.

The plan for the rest of the paper is as follows. We begin by providing a brief overview of modern approaches to recommendation in both cognitive and computer science. We then outline a playlist completion task that will be the focus of the paper and describe two web experiments designed to collect fine-grained human recommendation judgments. We then examine the performance of two representative collaborative filtering models and two cognitive models on this task, evaluated using metrics from both cognitive and computer science.

Background

Research on human generalization provides a unifying psychological framework for studying automated recommendation, and, as we argue below, can be readily applied to model human recommendation judgments. Below we offer an overview of psychological models of generalization and contrast these with models of recommendation from the machine learning literature.

Bayesian models of human generalization

Probabilistic models of cognition have proved successful at capturing many high and low-level psychological processes. One of the earliest and most prominent examples of this approach is Shepard's (1987) Bayesian formulation of human generalization, which was demonstrated by Tenenbaum and Griffiths (2001) to encompass other popular set-theoretic approaches to similarity (e.g., Tversky, 1977). To study human generalization profiles, individuals are shown examples of previously unknown concepts (e.g., a "fep") and are asked to use this information to identify other items that are also likely to be in this conceptual category. For tasks of this sort, Shepard's model and its extensions (Tenenbaum & Griffiths, 2001) provide a close fit to human judgments across a range of domains. In the current paper, we draw a direct correspondence between generalization and recommendation, modeling recommendation as an instance of generalization in which the latent category to generalize is implied by a user's previously rated items.

The Bayesian formulation of the generalization problem is as follows: a person observes n examples $\mathbf{x}$ of a new concept C. The generalization problem is to compute the probability that a new item y belongs to C given the observations in $\mathbf{x}$, $P(y \in C|\mathbf{x})$. To calculate this probability, the participant can make use of a hypothesis space, $\mathcal{H}$, of potential concepts, each consisting of a set of items. We define a prior probability distribution over hypotheses, $P(h)$, and a likelihood function, $P(\mathbf{x}'|h)$, indicating the probability of observing the examples $\mathbf{x}'$ under hypothesized concept h. The probability of generalizing concept C to include item y, $P(y \in C|\mathbf{x})$ is then given by Bayesian model averaging:

$$P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x}) \quad (1)$$

where the probability of a hypothesized concept given the current examples, $P(h|\mathbf{x})$, is calculated via Bayes' rule:

$$P(h|\mathbf{x}) = \frac{P(\mathbf{x}|h)P(h)}{\sum_{h' \in \mathcal{H}} P(\mathbf{x}|h')P(h')} \quad (2)$$

and the probability of item y being in concept C given the hypothesis h is an indicator function, $P(y \in C|h) = \mathbb{1}_{y \in h}$, taking the value 1 when y is in the set picked out by h.

Collaborative filtering and recommendation systems

Collaborative filtering (CF) approaches constitute the state-of-the-art in machine learning and computer science for generating automated recommendation. Algorithms in this class are so-named because they use large amounts of data from many individual users to collectively (collaboratively) generate predictions (filter) for items that may be of interest to a new user. Much of the success of CF stems from its domain generality: algorithms in this class require only a database of users and their item preferences, distinguishing them from content-based approaches which rely on task-specific user or item features to augment the recommendation decision.

CF algorithms can be organized in terms of the amount of latent structure they assume when analyzing the user-item database. *Memory-based* CF algorithms operate directly on the user-item database, making minimal assumptions about the structure present in the data (Su & Khoshgoftaar, 2009). Algorithms in this class predict preferences for a new item by combining the preferences of the most similar users or items in the database, often resulting in recommendation behavior that is highly sensitive to the local variation in other highly similar users or items. In contrast, *model-based* CF algorithms use the user-item database to learn a model of the data which they then use to generate recommendations. Through their reliance on a data model, approaches in this category are often able to capture more global properties of a dataset relating to many or all entries in the database (Bell & Koren, 2007). Popular data modeling approaches include sequential decision models, dimensionality-reduction/latent factor models, and Bayesian networks (see Su & Khoshgoftaar (2009) for an overview).

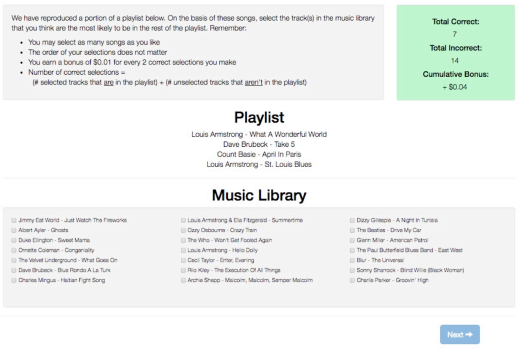

Figure 1: Web interface for playlist completion task.

Experiment 1: Playlist completion task

In this section, we outline a web-based experimental task designed to collect fine-grained human recommendation judgments on a familiar recommendation task. We set up our experimental task as a playlist completion problem: participants are shown a *trace*, $\mathbf{x}_{\text{trace}}$, from an unobserved full playlist $\mathbf{x}_{\text{full}}$, where $\mathbf{x}_{\text{trace}}$ is a list of a subset of the songs that appear in $\mathbf{x}_{\text{full}}$. On the basis of the songs in the trace, participants are asked to select additional songs from a music library which are likely to also be in the playlist which contains $\mathbf{x}_{\text{trace}}$. This task bears a direct resemblance to many common music recommendation tasks, including the recommendation of new songs based on listening history (e.g., Spotify's "Discover Weekly"), and

the construction of new playlists using seeds (e.g., Apple Music's Genius recommendations).

Methods

A total of $n = 51$ participants on Amazon Mechanical Turk completed a pretest to assess musical fluency. Upon passing the pretest, participants were shown traces from playlists deemed to be consistent with their musical expertise. The number of songs in each playlist trace was varied from one to five. Additionally, the music library available on each question was constructed to always include seven "in playlist" songs which were not in $\mathbf{x}_{\text{trace}}$ but which were in $\mathbf{x}_{\text{full}}$, seven "in genre" songs which were not in $\mathbf{x}_{\text{full}}$ but which were in the same musical genre, and seven "out-of-genre" songs which were neither in $\mathbf{x}_{\text{full}}$ nor in its musical genre. Participants received a bonus of $0.01 for every two correct selections they made on each playlist trial.

Playlists for the experimental task were sourced from the Art of the Mix dataset (McFee & Lanckriet, 2012). The dataset also provided the hypothesis space for the recommendation algorithms discussed below, represented as a sparse, binary co-occurrence matrix $\mathbf{X}$ of approximately 100K playlists by 120K songs. We augmented the original hypothesis space, $\mathcal{H}$, to include 10 additional "genre playlists", where a genre playlist corresponded to the set of all songs in $\mathbf{X}$ associated with a given music genre, as identified via the Discogs API.[1] We refer to the set of original playlists augmented with the genre playlists as $\mathcal{H}^+$.

Model specifications

We evaluate the performance of three computational models of recommendation against human judgments on the above playlist completion task. Each model was selected to provide representative coverage of the diversity of recommendation approaches across cognitive and computer science. Parameterizations for each model were arrived at independently via grid-search for each evaluation metric.

Item-based CF model Item-based CF models are one of the most prominent memory-based CF approaches, having been used extensively in industrial applications, including amongst others Amazon.com's product recommendation engine (Sarwar, Karypis, Konstan, & Riedl, 2001; Linden, Smith, & York, 2003). Item-based CF algorithms operate by computing the pairwise similarities between all items in the user-item database and combining these ratings for each entry in a user's preference history. These aggregated similarity scores are then used to identify the unrated items for each user that are most similar to their previous selections. It is of note that this formulation has a direct correspondence with the exemplar theory of categorization from the cognitive science literature (Medin & Schaffer, 1978; Nosofsky, 1986).

For the binary $n \times m$ playlist-song matrix used in the playlist completion task, we employed a cosine similarity function to measure the pairwise similarities between the binary song columns, $\mathbf{x}_j \in \mathbb{Z}_2^n$. This produced a symmetric matrix of similarity ratings, $\mathbf{S} \in \mathbb{R}^{m \times m}$. We averaged over rows to produce the recommendation scores for user i, $\mathbf{r}_i$:

$$\mathbf{r}_i = \frac{\sum_j \mathbf{s}_j \mathbb{1}_{j \in \mathbf{x}_i}}{\sum_j \mathbb{1}_{j \in \mathbf{x}_i}} \tag{3}$$

where $\mathbf{s}_j$ is the j^{th} row in $\mathbf{S}$ and $\mathbb{1}_{j \in \mathbf{x}_i}$ is an indicator function which is 1 when song j is in playlist i, and 0 otherwise. In the current experiment we found that using a cosine similarity function to generate $\mathbf{S}$ and only averaging over the top-2 most similar items resulted in the best fit to the human data.

Matrix factorization CF model In addition to the instance of memory-based CF algorithm described above, we also evaluated a popular model-based algorithm: a dimensionality reduction approach based on nonnegative matrix-factorization (NMF) (Lee & Seung, 2001). This approach represents the playlist-song database as a binary matrix, $\mathbf{X} \in \mathbb{Z}_2^{n \times m}$ where entry $x_{i,j}$ contains whether song j appeared in playlist i, and identifies non-negative low rank factors, $\mathbf{W} \in \mathbb{R}_+^{n \times k}$ and $\mathbf{H} \in \mathbb{R}_+^{k \times m}$ whose product approximates $\mathbf{X}$. NMF identifies the factor matrices $\mathbf{W}$ and $\mathbf{H}$ via coordinate descent on the objective

$$f = \frac{1}{2} ||\mathbf{X} - \mathbf{WH}||_{Fro}^2 \tag{4}$$

where $||\cdot||_{Fro}$ indicates the Frobenius norm. Once $\mathbf{W}$ and $\mathbf{H}$ have been identified, completions for playlist i, $\mathbf{r}_i \in \mathbb{R}^j$, are generated via

$$\mathbf{r}_i = \mathbf{x}_i \mathbf{H}^\top \mathbf{H} \tag{5}$$

where $\mathbf{x}_i \in \mathbb{Z}_2^m$ is the binary row vector corresponding to the current preferences for playlist i in the database. Matrix factorization approaches like NMF are one of the most commonly used versions of model-based CF, in part due to their ease of implementation and scalability (Su & Khoshgoftaar, 2009). In the current experiment, we found that using between 50 and 100 latent factors resulted in the best model performance.

Bayesian model of generalization In the context of the playlist completion task, hypotheses, h, correspond to rows in the playlist-song matrix $\mathbf{X}$, observations, $\mathbf{x}$, correspond to the binary vector of songs in the partially observed playlist we wish to complete $\mathbf{x}_{\text{trace}}$, and C corresponds to the unknown fully observed playlist we are attempting to reproduce.[2]

In the experiments below, we use a hierarchical prior over the augmented hypothesis space $\mathcal{H}^+$, drawing inspiration from Tenenbaum (1999). A fraction $1 - \lambda_g$ of the total probability was allocated to the original playlists in $\mathcal{H}$ as a group, leaving λ_g to be distributed across the genre playlists. The

[1] https://www.discogs.com/developers

[2] For convenience in the equations below we represent hypothesis h as the set of nonzero column indices for the corresponding row of the playlist-song matrix, i.e., $h_i = \{j : x_{ij} = 1\}$.

λ_g probability was distributed uniformly across the genre hypotheses, while the $1-\lambda_g$ probability was distributed over the original playlists as a function of the playlist size according to an Erlang distribution, $p(h) \propto (|h|/\sigma^2)\exp\{-|h|/\sigma\}$. The mixture parameter, λ_g, controlling the influence of the genre playlists, was set to 0.1, while the Erlang parameter σ was set to 150, favoring larger playlists. These settings were arrived at independently via grid search for the stated objective.

The likelihood term, $P(\mathbf{x}|h)$ in the Bayesian generalization model was defined as a mixture distribution with weight ε balancing the influence of the size of the playlist under consideration with a popularity term measuring how many times each song in the playlist occurred across the original hypothesis space $\mathcal{H}$. Specifically, the likelihood was computed as

$$P(\mathbf{x}|h) = (1-\varepsilon)P_{\text{size}} + \varepsilon P_{\text{popularity}} \tag{6}$$

where

$$P_{\text{size}} = \begin{cases} 1/|h|^{\|\mathbf{x}\|_2} & : i \subseteq h \;\forall\; i : x_i = 1 \\ 0 & : \text{otherwise} \end{cases} \tag{7}$$

and

$$P_{\text{popularity}} \propto \sum_{i \in h} |\{h' \in \mathcal{H} : i \in h'\}|. \tag{8}$$

Prototype model Finally, we evaluate an implementation of the prototype theory of categorization (Reed, 1972). We define a prototypical playlist, $\mathbf{x}_{\text{proto}}$, to be a set containing those songs that are present in the majority of playlists in the database consistent with at least one song in the set of observations, $\mathbf{x}$. Following Abbott, Austerweil, and Griffiths (2012), the generalization score for a new song y was defined to be

$$\text{Pscore}(y|\mathbf{x}) = \exp\left\{-\lambda_p \,\text{dist}(y,\, \mathbf{x}_{\text{proto}})\right\} \tag{9}$$

where $\text{dist}(\cdot,\cdot)$ is the Hamming distance between the vector representations of its arguments and λ_p is a free parameter whose optimal value ranged between between 1 and 25 for the experiments reported below.

Results

We evaluated each model using three different evaluation criteria: F_1 score using the fully observed playlist data as ground-truth, F_1 score using *human* selections as ground truth, and model correlations with human selection probabilities across recommendation levels (in playlist, in genre, out of genre). The F_1 score is a common measure of a test's accuracy, calculated as the harmonic mean of a test's recall (its ability to correctly identify all true positives) and its precision (its tendency to produce false positives). Each F_1 version captures a different aspect of the recommendation performance: the playlist ground-truth F_1 score is one of the standard evaluation criteria within machine learning and information retrieval, while human ground truth F_1 scores are closer to model evaluation metrics used in cognitive science. The correlation in recommendation probabilities is a novel metric which provides a finer-grained analysis of a model's capacity to reproduce human recommendation profiles.

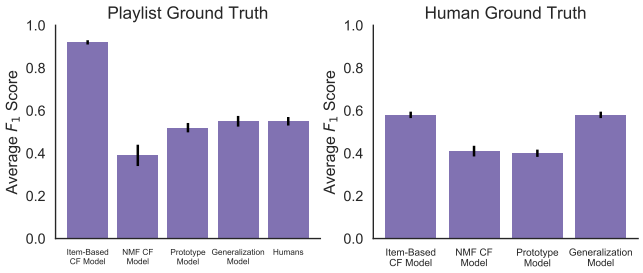

Figure 2: Model F_1 scores on the playlist completion task. *A*. Playlist ground-truth F_1 scores. This metric reflects the ability of a model to accurately identify all positive examples of songs in the unobserved full playlist, and none of the songs outside of it. *B*. Human ground-truth F_1 scores. This metric reflects the model's ability to select the same songs as humans do on each playlist problem, and not to select anything else.

The parameters for each model were fit separately to each evaluation metric via grid search. The item-based CF model performed best when averaging over the nearest 2 neighbors for both of the F_1 metrics. Similarly, the NMF model scored highest on each F_1 metric using $k = 100$ latent factors. The Bayesian model achieved marginally higher performance with different settings of ε on the playlist ground-truth vs. human ground-truth metrics ($\varepsilon = 1 \times 10^{-6}$ and $\varepsilon = 0.0001$, respectively), as did the prototype model ($\lambda_p = 25$ and $\lambda_p = 1$).

When evaluated against the playlist ground truths from the AOTM dataset, we found that both the matrix factorization and the item-based CF models differed significantly from the performance of humans and the two cognitive science models (Figure 2). Indeed, the item-based CF model showed a strong tendency to simply reproduce the source playlist, due in large part to its direct reliance on the raw playlist-song matrix, while the matrix factorization model showed lower overall fit due to its use of an intermediate data model (the latent factor matrices, $\mathbf{W}$ and $\mathbf{H}$). In contrast, both the prototype and Bayesian generalization models showed F_1 scores similar to humans.

When we evaluated the models in terms of their ability to reproduce human ratings, a slightly different picture emerged: while both the prototype and matrix factorization model had difficulty reproducing human judgments, the Bayesian generalization model and item-based CF model performed similarly (Figure 2).

To further explore the capacity of each model to fit human judgments, we looked at model correlations with the average human recommendation probability, stratified by recommendation level and the number of cues (Figure 3). Drawing inspiration from Xu and Tenenbaum (2007), recommendations were broken down into in-playlist, in-genre, and out-of-genre songs, allowing us to calculate the model's tendency to generalize at each level. Whereas the human ground-truth F_1

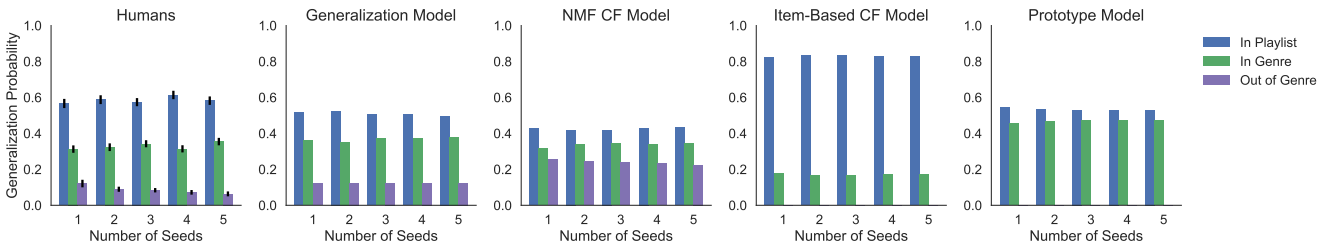

Figure 3: Human and model recommendation probabilities as a function of song category and number of cues.

scores indicated that the item-based CF and Bayesian generalization models were equally capable of reproducing human selection profiles, this finer-grained analysis revealed that the relatively coarse F_1 metric masks significant differences in the two models' generalization behavior. Qualitatively, the item-based CF model was heavily biased towards selecting in-playlist items at the expense of generalizing beyond the specific playlists in the playlist-song database. The prototype model was slightly less strict in its generalizations, producing comparable amounts of in playlist and in-genre recommendations, but refusing to generalize out-of-genre. This behavior put both models at odds with humans, who exhibited the characteristic exponential decay in generalization tendency from in-playlist to in-genre to out-of-genre. Importantly, the Bayesian generalization model did the best at reproducing this tendency, while the matrix factorization model showed less distinction overall between the different recommendation levels. Quantitatively, the Bayesian generalization model's recommendation gradients showed the highest correlation with the human selection data, $R = 0.9814, p < 0.001$, when compared against the other models (NMF CF: $R = 0.9254, p < 0.001$, Item-based CF: $R = 0.8705, p < 0.001$, Prototype model: $R = 0.9206, p < 0.001$).

Experiment 2: Rating model recommendations

The results of Experiment 1 indicate that, despite showing high marks along traditional information retrieval metrics of success, two modern collaborative filtering approaches fail to capture important properties of human recommendation behavior in a playlist completion task. In contrast, a model from computational cognitive science – the Bayesian model of generalization – showed a strong fit to human recommendation profiles. A natural next question is whether users are also sensitive to these differences.

Methods

In a second experiment we generated new "hybrid playlists" consisting of an equal mixture of songs from two randomly selected playlists from Experiment 1. On each trial we provided participants with two, four, or six cue songs from a hybrid playlist, and instructed them to rate each model's top 10 recommendations in terms of how likely they were to be in the playlist containing the observed cues. We recruited 50 participants via Amazon Mechanical Turk and ran them in this rating task. Of these participants, 18 failed to pass the music pretest on any genre, leaving a total of $n = 32$ participants in the final study. Of these 32, each participant completed an average of six rating trials.

Results

Aggregating human ratings across cue conditions, we found that participants significantly favored songs recommended by the Bayesian generalization model in comparison to any of the other models evaluated ($F(3, 8396) = 113.397$, $p < 0.001$; Figure 4). There were no interactions between the recommendation rank and the model. Indeed, in line with findings that users are both sensitive to differences between human and model recommendations and favor those recommendations made by humans, our results indicate that users also favor recommendations from models which reproduce larger proportions of human recommendation behavior.

Discussion

Internet-scale behavioral datasets offer an opportunity to evaluate theories of cognition at a scale rarely seen in traditional in-lab studies. We demonstrate this potential using two experiments. We began by evaluating two representative models from both the cognitive science and machine learning literatures on a dataset compiled from people's interactions online. Notably, these data reflected people's spontaneous interactions in their natural social environment, spanned a period of over a decade, had more than 100,000 unique observations, and was completely free. We found that that two models from the collaborative filtering literature as well as a popular model of categorization failed to reproduce important aspects of human recommendation behavior, while a model of human generalization showed a superior fit to these naturalistic human generalization patterns. In a second experiment we found that participants significantly favored the recommendations from the generalization model over both the categorization and machine learning models, bolstering the empirical validity of the generalization model and demonstrating the value of taking the psychological component of these tasks seriously.

More generally, the opportunity to reformulate standard machine learning applications in terms of human cognition is an exciting avenue for both machine learning researchers and cognitive scientists. Every day, behavioral data from millions

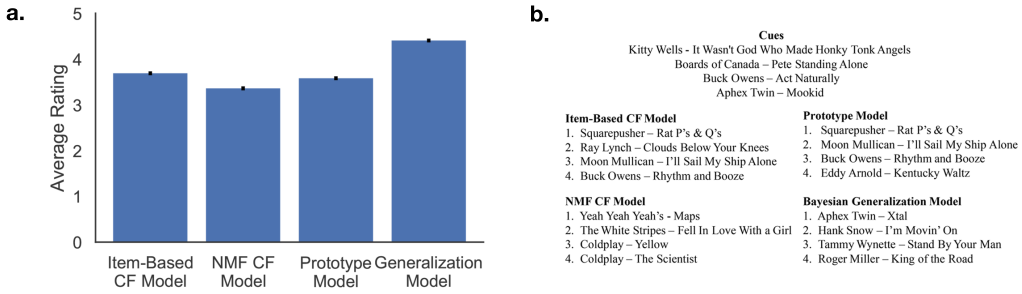

Figure 4: *A.* Average ratings for the top 10 recommendations produced by each model. Error bars reflect ± 1 SEM. *B.* Top four recommendations produced by each model on a sample problem.

of people is dutifully coded, anonymized, and stored as they engage in a variety of online tasks. Many of these tasks engage directly with fundamental cognitive abilities like categorization, generalization, and semantic understanding. While traditionally the data generated during these interactions has been handled by engineers and computer scientists, the above results indicate the potential gains that come from modeling the psychological aspects of these tasks directly. Just as the cognitive revolution in psychology demonstrated the necessity of incorporating mental states as mediating factors in behavior, so too can computational models of cognition revolutionize the current machine learning approaches to behavioral modeling by explicitly engaging with the psychological origins of the data (Griffiths, 2015).

Our results offer several takeaways for cognitive scientists. By using a web-scale recommendation dataset to fit models from cognitive science, we demonstrate how behavioral data from the web can be used to advance theories of cognition, providing an avenue for modelers interested in testing their theories on noisier, more realistic tasks. In addition, by showing that a cognitive model outperforms two popular approaches from the machine learning literature on a recommendation task, we illustrate the importance of incorporating principles from cognitive modeling in a domain that has been *de facto* colonized by computer science. Finally, and perhaps most importantly, we demonstrate how the influx of online behavioral data can help narrow the gap between the more theory-based approaches in cognitive science and the more data-driven approaches in machine learning.

Acknowledgments. This work was supported by NSF grant SMA-1338541.

References

Abbott, J., Austerweil, J., & Griffiths, T. (2012). Constructing a hypothesis space from the web for large-scale Bayesian word learning. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (Vol. 34).

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, *9*(2), 75–79.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Krishnan, V., Narayanashetty, P. K., Nathan, M., Davies, R. T., & Konstan, J. A. (2008). Who predicts better?: Results from an online study comparing humans and an online recommender system. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp. 211–218).

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, *14*, 556–562.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80.

Logg, J. M. (2017). *Theory of machine: When do people rely on algorithms?* (Tech. Rep. No. 17-086). Harvard Business School NOM Unit Working Paper.

McFee, B., & Lanckriet, G. R. (2012). Hypergraph models of playlist dialects. In *Proceedings of the 13th International Society for Music Information Retrieval* (pp. 343–348).

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 285–295).

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sinha, R. R., & Swearingen, K. (2001). Comparing recommendations made by online systems and friends. In *Proceedings of the DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries* (Vol. 106).

Sinha, S., Rashmi, K. S., & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. In *Proceedings of the ACM SIGIR 2001 Workshop on Recommender Systems* (pp. 24–33).

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *4*, 1-19.

Tenenbaum, J. B. (1999). Rules and similarity in concept learning. *Advances in Neural Information Processing Systems*, *12*, 59–65.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2017). Making sense of recommendations. *Management Science*.