Journal of Experimental Psychology: General

Recommendation as Generalization: Using Big Data to Evaluate Cognitive Models

David D. Bourgin, Joshua T. Abbott, Thomas L. Griffiths

Funded by:

- National Science Foundation

Abstract

The explosion of data generated during human interactions online presents an opportunity for psychologists to evaluate cognitive models outside the confines of the laboratory. Moreover, the size of these online datasets can allow researchers to construct far richer models than would be feasible with smaller in-lab behavioral data. In the current paper we illustrate this potential by evaluating three popular psychological models of generalization on two web-scale online datasets typically used to build automated recommendation systems. We show that each psychological model can be efficiently implemented at scale and in certain cases can capture trends in human judgments which standard recommendation systems from machine learning miss. We use these results to illustrate the opportunity internet-scale datasets offer to psychologists, 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, generalization, big data, machine learning

Document word count (approximate): 5129

Recommendation as Generalization: Using Big Data to Evaluate Cognitive Models

Every day, petabytes 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 psychology research, they offer insight into many of the same cognitive phenomena, often at a scale rarely seen in the behavioral sciences. This new source of high throughput behavioral data, already the lifeblood of machine learning and AI researchers, offers psychologists a similar opportunity to evaluate cognitive models in the wild, at scale, with minimal investment (Jones, 2016; Paxton & Griffiths, 2017). Moreover, the dearth of cognitive modeling techniques in current approaches to analyzing these behavioral datasets is an opportunity for psychologists to re-establish the value of modeling minds as mediating influences on behavior in a domain that has become a part of computer science by default.

Product recommendation is a notable example of an applied task which can serve as a test-bed for cognitive models. Not only has recommendation-guided search become near ubiquitous in our daily interactions online, but it is also a prominent example of a fundamentally psychological task which is now solved by machine learning systems. Indeed, automated approaches to recommendation have become one of the most common 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 people 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 people tend to prefer recommendations made by friends to those generated by automated recommendation systems. 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 jokes recommended by human experts. 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 the online recommendation datasets have as cognitive test-beds highlighting the relationship between recommendation and the psychological problem of generalization. We draw on this relationship to evaluate a version of the well-known Bayesian model of generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001) along with two other classic psychological models on hundreds of thousands of human judgments from two different recommendation domains. We illustrate the value that cognitive modeling techniques play in this new world by comparing the performance of these psychological models with that of widely-used machine learning approaches to recommendation. The results provide a test of the cognitive models at an unprecedented scale and level of realism, as well as a demonstration of how psychological theory can inform the interpretation of large behavioral data sets.

The plan for the rest of the paper is as follows. We begin by introducing the problem of recommendation, linking it to generalization, and providing a brief overview of modern approaches to recommendation in computer science and generalization in psychology. We outline a playlist completion task that will be the focus of the paper and describe two 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, across two separate recommendation domains. We conclude by discussing the implications of our findings.

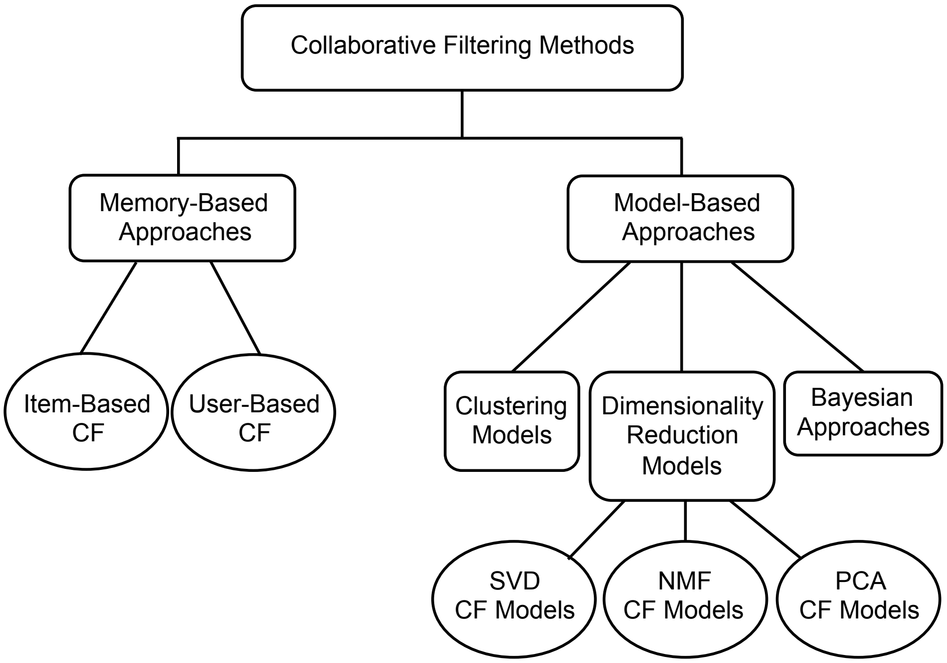


Figure 1. An overview of the relationships between popular approaches to collaborative filtering, emphasizing those approaches evaluated in the current paper. Notably, hybrid models like content-based CF approaches are not pictured.

Recommendation as a psychological problem

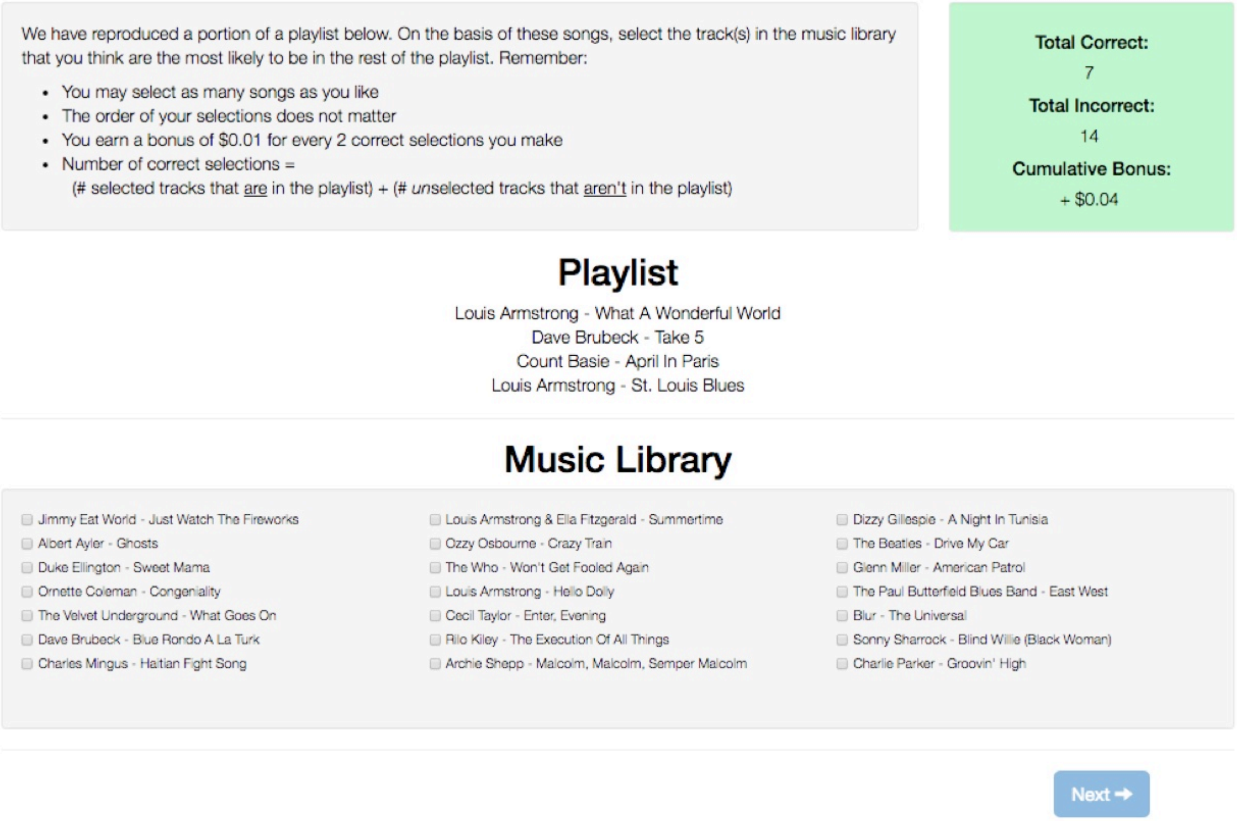
In the current paper we argue that the problem of recommendation in machine learning bears a formal resemblance to the classic problem of generalization in psychology. Below we offer an overview of psychological models of generalization and contrast these with models of recommendation from the machine learning literature.

Recommendation

At their core, recommendation problems involve identifying items that a user is likely to find valuable based on their selection history. Examples include proposing songs to listen to given someone's listening history, suggesting movies to watch given a user's movie ratings, and recommending items to buy based on browsing history. In each case, a recommender is shown a *trace*, T, sampled from a hypothetical set of all items a user likes, U. On the basis of the items in the trace, a recommender must try to select additional items that are likely to also be in the set $U - T$.

Collaborative filtering (CF) approaches constitute the state-of-the-art in machine learning for generating automated recommendations. Algorithms in this class use data from many users to collaboratively generate predictions (filter) for items that may be of interest to a new user. Much of CF's success stems comes from its generality - it requires only a database of users and their item preferences to work - in contrast to content-based approaches which rely on task-specific user or item features.

CF algorithms can be organized by the amount of latent structure they assume in the user-item database (Figure 1). *Memory-based* CF algorithms operate directly on the user-item database, making minimal assumptions about the structure of the data (Su & Khoshgoftaar, 2009). Algorithms in this class predict preferences for new items based on the preferences of similar users or items, resulting in recommendation behavior that relies on only a subset of the dataset. In contrast, *model-based* CF algorithms use the user-item database to train an intermediate, lower-dimensional data model. Using this model to generate recommendations often results in more global recommendation decisions (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).



We have reproduced a portion of a playlist below. On the basis of these songs, select the track(s) in the music library that you think are the most likely to be in the rest of the playlist. Remember:

- You may select as many songs as you like
- The order of your selections does not matter
- You earn a bonus of $0.01 for every 2 correct selections you make
- Number of correct selections =
  (# selected tracks that are in the playlist) + (# *un*selected tracks that aren't in the playlist)

Total Correct:
7
Total Incorrect:
14
Cumulative Bonus:
+ $0.04

Playlist

Louis Armstrong - What A Wonderful World
Dave Brubeck - Take 5
Count Basie - April In Paris
Louis Armstrong - St. Louis Blues

Music Library

□ Jimmy Eat World - Just Watch The Fireworks
□ Albert Ayler - Ghosts
□ Duke Ellington - Sweet Mama
□ Ornette Coleman - Congeniality
□ The Velvet Underground - What Goes On
□ Dave Brubeck - Blue Rondo A La Turk
□ Charles Mingus - Haitian Fight Song
□ Louis Armstrong & Ella Fitzgerald - Summertime
□ Ozzy Osbourne - Crazy Train
□ The Who - Won't Get Fooled Again
□ Louis Armstrong - Hello Dolly
□ Cecil Taylor - Enter, Evening
□ Rilo Kiley - The Execution Of All Things
□ Archie Shepp - Malcolm, Malcolm, Semper Malcolm
□ Dizzy Gillespie - A Night In Tunisia
□ The Beatles - Drive My Car
□ Glenn Miller - American Patrol
□ The Paul Butterfield Blues Band - East West
□ Blur - The Universal
□ Sonny Sharrock - Blind Willie (Black Woman)
□ Charlie Parker - Groovin' High

Figure 2. Web interface for the recommendation task using the Art of the Mix (AOTM) data.

Recommendation as generalization

The recommendation problem above bears a direct relationship to the psychological problem of generalization in word learning (Xu & Tenenbaum, 2007), categorization (Nosofsky, 1986), and human similarity judgments (Tenenbaum & Griffiths, 2001). In the current paper, we explore this connection, arguing that human generalization constitutes a unifying psychological framework for automated recommendation.

In a traditional generalization task, individuals are shown examples of an unknown concept (e.g., a "fep") and are asked to identify other items that are likely to be in this conceptual category. Formally, a person observes a set of items $x = \{x_1, \ldots, x_n\}$, associated with a new category C and must compute $P(y \in C \mid x)$, the probability that a new item y

belongs to C given the items in x. Good generalizations correspond to selecting those items y that maximize $P(y \in C \mid x)$. Different cognitive models propose different approaches to representing items and combining them to produce generalization probabilities.

The correspondence between recommendation and concept generalization is straightforward: the recommendation trace, T, corresponds to the collection of category exemplars x, while the user's preference set U becomes the set of all items in the category, C. The generalization probability for a new item y can be treated as a recommendation score, where higher values suggest a greater probability that a user will respond positively to that item.[1] This correspondence is so great that in certain cases identical models have been proposed under different names across the generalization and recommendation literatures.[2]

Despite these similarities, however, there have not been any direct comparisons between CF and generalization models. In the next section we outline an experiment designed to evaluate models from each approach on a common task.

Experiment 1: Recommendation task

In our first experiment we compare popular psychological and machine learning models of recommendation with human recommendation profiles. This task bears a direct resemblance to many common machine recommendation tasks, including the recommendation of new items based on browsing history (e.g., Spotify's "Discover Weekly"), and the construction of new playlists using seeds (e.g., Apple Music's Genius recommendations). Rather than directly collect human recommendation profiles to train our models, we use two existing web recommendation datasets: Art of the Mix (AOTM;

[1] One notable difference is that rather than comparing a new item against items previously seen by the same user (as in a generalization study), recommendation problems compare the item to all items in a database, regardless of whether the user has encountered them.

[2] Item-based collaborative filtering and the exemplar theory of categorization (Medin & Schaffer, 1978; Nosofsky, 1986) are a notable example. See Appendix for further details.

McFee & Lanckriet, 2012), a dataset of music playlists, and Goodbooks10k (Zajac, 2017), a dataset of reading lists.

Method

Stimuli.

AOTM dataset. Playlists for the music version of the recommendation task were sourced from the Art of the Mix (AOTM) dataset (McFee & Lanckriet, 2012). This dataset consisted of approximately 100K user-generated playlists over 120K unique songs, represented as a sparse, binary co-occurrence matrix $\mathbf{X}$ where rows corresponded to playlists and columns correspond to songs. We also allowed for the augmentation of this original set of playlist hypotheses, $\mathcal{H}$, with 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.[3] These additional playlists served as a “genre-bias” on the model predictions, encouraging the model to generalize along genre-boundaries. In many cases, inclusion of these playlists improved model fit to human data, reflecting the recommendation task’s emphasis on literary / musical genre. In the experiments below, we treated inclusion of these genre playlists in the hypothesis space as a hyperparameter.

Goodbooks dataset. Reading lists[4] for the literature version of the recommendation task were sourced from the `goodbooks-10k` dataset (Zajac, 2017). This dataset consisted of approximately 50K user ratings for 10K of the most popular books on the Goodbooks website.[5] To make this consistent with the AOTM data, we converted the dataset into a binary co-occurrence matrix. Just as with the AOTM data, we allowed for the original user playlists, $\mathcal{H}$, to be augmented with 28 additional “genre playlists”, where a genre playlist corresponded to the set of all books in $\mathbf{X}$ associated with a given literary

[3] `https://www.discogs.com/developers`

[4] In the rest of the paper we refer to these as “playlists“ to maintain consistency with the music data.

[5] `https://www.goodbooks.com`

genre, identified via user tags on the Goodbooks website.

Participants. A total of 47 (AOTM) and 40 (Goodbooks) participants on Amazon Mechanical Turk passed a pretest to assess their familiarity with the music or book genres used in the experiment, resulting in an average of 117 unique responses per cue condition. This provided enough statistical power to identify a product-moment correlation of $R = 0.35$ between model and human item recommendation probabilities at $\alpha = 0.05$ at $1 - \beta \geq 0.8$. Participants were paid $1.25 for their participation. All data was collected with the consent of the participants and approval from the UC Berkeley Committee for Protection of Human Subjects.

Design and procedure. We formulated our experiment as a playlist completion task: on each trial, a participant was shown a selection of one to five items from an unobserved playlist, and asked to select additional items that they thought would be in the playlist that generated these observations. Participants made their selections from a fixed library, which was constructed so as to always include seven "in playlist" items (i.e., items which were not in the selections shown to the participant but which were in the unseen playlist that generated them), seven "in genre" items (i.e., items which were not in the unseen playlist but which were in the same musical/literary genre), and seven "out-of-genre" items (i.e., items which were neither in the unobserved playlist nor its musical genre) (Figure 2). Participants were notified they would receive a bonus of $0.01 for every two correct selections they made on each trial to encourage them to select items judiciously. To reduce demand characteristics, we only reported a participant's accrued bonus *after* they submitted their selections, making it difficult for them to know which selections led to their overall earnings. To ensure that participants were familiar with the items in a playlist, we used performance on the genre pretest to determine which playlists to display to which participants.

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 random search (Bayesian generalization model) or grid search (all other models) using the online recommendation data, with behavior from the playlist completion task serving as the test set. For the best performing parameter settings for each model, see Appendix B.

Item-based CF model (Exemplar model). Item-based collaborative filtering, a classic memory-based approach, is one of the first techniques developed for generating recommendations at scale. Given a user's listening/reading history, the item-based CF model selects the k most similar items ("nearest neighbors") in the catalogue for each item in the user's history, ignoring any item that the user has already engaged with. For each item in this set, the recommendation score is defined as the average similarity between it and each of the items in the user's history.

Matrix factorization CF model. In addition to a memory-based CF algorithm, we also evaluated a popular model-based method: a dimensionality reduction approach based on nonnegative matrix-factorization (NMF) (Lee & Seung, 2001).[6] This approach represents the playlist-item database as a binary matrix, $\mathbf{X} \in \mathbb{Z}_2^{n \times m}$ where entry $x_{i,j}$ indicates whether item 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 least-squares objective $\frac{1}{2}||\mathbf{X} - \mathbf{WH}||^2_{\text{Fro}}$ where $|| \cdot ||_{\text{Fro}}$ indicates the Frobenius norm. Once $\mathbf{W}$ and $\mathbf{H}$ have been

[6] NMF was employed due to its success in previous recommendation settings (e.g., Zhang, Wang, Ford, & Makedon, 2006), and its relationship to feature-based object representations in psychology (Lee & Seung, 1999).

identified, completions for playlist i, $\mathbf{r}_i \in \mathbb{R}^j$, are generated via

$$\mathbf{r}_i = \mathbf{x}_i \mathbf{H}^\top \mathbf{H} \tag{1}$$

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).

Bayesian model of generalization. One of the most prominent models of generalization is Shepard's (1987) Bayesian formulation, which was demonstrated by Tenenbaum and Griffiths (2001) to encompass other popular set-theoretic approaches to similarity (e.g., Tversky, 1977). This approach defines a hypothesis space of potential concepts, $\mathcal{H}$, a prior probability distribution over hypotheses, $P(h)$, and a likelihood function, $P(x|h)$, giving the probability of observation x under hypothesis h. The probability of generalizing concept C to include item y, $P(y \in C|x)$ is then given by Bayesian model averaging:

$$P(y \in C|x) = \sum_{h \in \mathcal{H}} P(y \in C|h, x) P(h|x) \tag{2}$$

In the context of the playlist completion task, hypotheses, h, correspond to playlists, observations, x, correspond to the items in the partially observed playlist, and a concept, C, corresponds to the full playlist we wish to reproduce.

In the experiments below, we use a binarized NMF approximation of the the original playlist-item matrix as the hypothesis space for the Bayesian model.[7] We define a hierarchical prior over the hypotheses in this space, drawing inspiration from Tenenbaum (1999). A fraction $0 \leq 1 - \lambda_g \leq 1$ of the total probability was allocated to the original

[7] Using the non-negative factorization of the playlist-item matrix ensured that the Bayesian model operated on a similar feature space to the NMF CF approach. As a result, it is reasonable to consider the Bayesian model a hybrid approach which adds psychological bias to the representations derived via NMF.

playlists in $\mathcal{H}$ as a group, leaving λ_g to be distributed across the genre playlists. The 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 likelihood of a set of observed items x under a playlist h 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 item in the playlist occurred across the entire hypothesis space. Specifically, the likelihood was computed as

$$P(\mathbf{x}|h) = (1-\epsilon)P_{\text{size}} + \epsilon P_{\text{popularity}} \tag{3}$$

where

$$P_{\text{size}} = \begin{cases} 1/|h|^{|x|} & : x \subseteq h \\ 0 & : \text{otherwise} \end{cases} \tag{4}$$

and

$$P_{\text{popularity}} \propto \sum_{i \in h} |\{h' \in \mathcal{H} : i \in h'\}|. \tag{5}$$

In the above likelihood, the size term mirrored the form used in Tenenbaum (1999), while the popularity score was added to reflect the observation that human media consumption habits are particularly sensitive to social influences, most notably popularity (Salganik, Dodds, & Watts, 2006).

Prototype model. Finally, we evaluated an implementation of the prototype theory of categorization (Reed, 1972). We define a prototypical playlist, x_{proto}, to be a set containing those items that are present in the majority of playlists that contain at least one item from the set of observations, x. Following Abbott, Austerweil, and Griffiths (2012), the generalization score for a new item y is defined to be

$$\text{Pscore}(y|x) = \exp\left\{-\lambda_p \ \text{dist}(y,\ x_{\text{proto}})\right\} \tag{6}$$

where $\text{dist}(\cdot,\cdot)$ is the Hamming distance between the vector representations of its arguments and λ_p is a free parameter.

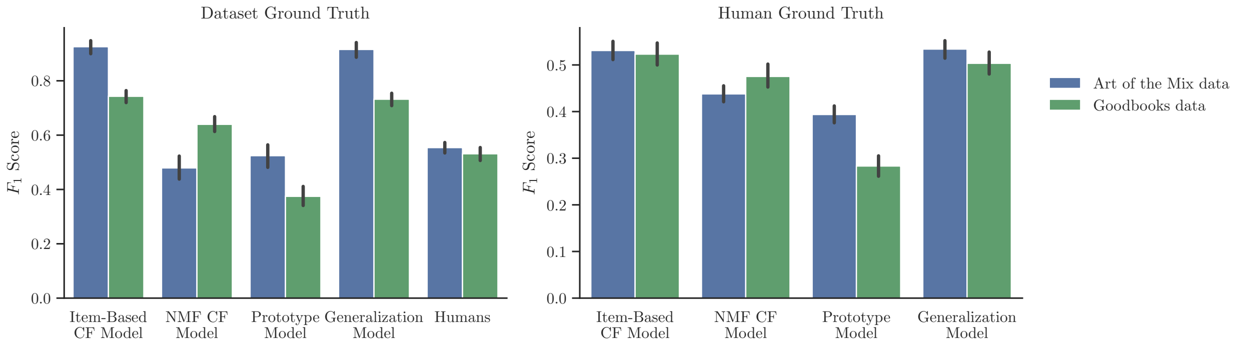


Figure 3. Model F_1 scores on the playlist completion task. *A*. Playlist ground-truth F_1 scores. This metric reflects the ability of a model to accurately identify all positive examples of items in the unobserved full playlist, and none of the items outside of it. *B*. Human ground-truth F_1 scores. This metric reflects the model's ability to select the same items as humans do on each playlist problem without selecting anything else.

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 frequencies for in-playlist, in-genre, and out-of-genre items as a function of the number of seed items displayed. The F_1 score is a common measure of performance, calculated as the harmonic mean of recall (the ratio of hits to hits plus misses) and precision (the ratio of hits to hits plus false alarms) (Rijsbergen, 1979). Each version we report 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 psychology. The correlation in recommendation probabilities provides a finer-grained analysis of a model's capacity to reproduce human recommendation profiles, as it captures trends in generalization performance as a function of item category, rather than simply in aggregate.

Parameters for each model were fit separately to each evaluation metric via grid search on the online datasets, using the Mechanical Turk data for evaluation. The best parameter values for each model on each metric are listed in Appendix B.

When evaluated against the playlist ground truths from both the AOTM and Goodbooks dataset, we found that the models varied widely in their capacity to accurately predict items in the test set. Indeed, both the item-based CF model and Bayesian generalization model showed strong generalization performance on the held-out playlists (Figure 3, left panel). In contrast, the matrix factorization model struggled to accurately predict the held-out data, likely due to its reliance on an intermediate data model (the latent factor matrices, $\mathbf{W}$ and $\mathbf{H}$) during prediction rather than on the dataset itself. This too is likely to explain the poor human performance on this metric, as the psychological representations of the items under consideration are unlikely to correspond directly with any specific playlists in the current datasets.

Table 1

Model correlations with the per-problem human ground-truth F_1 scores in Experiment 1.

Dataset	Item-based CF Model	NMF CF Model	Prototype Model	Generalization Model
AOTM	$R = 0.06, p = 0.473$	$R = 0.13, p = 0.119$	$R = -0.25, p = 0.002$	$R = 0.42, p < 0.001$
Goodbooks	$R = 0.37, p < 0.001$	$R = 0.45, p < 0.001$	$R = 0.23, p = 0.008$	$R = 0.43, p < 0.001$

In addition to calculating the average F_1 across all problems as in Figure 3, we also evaluated each model's correlation with the Mechanical Turker's average playlist ground-truth F_1 scores on a per-problem basis. This metric reflects a more nuanced evaluation of each model's fit to the Mechanical Turk data, ensuring that it can reproduce similar recommendation tendencies on *each recommendation problem* rather than simply in aggregate. On this metric, we found that the Bayesian generalization model achieved the best fit on the AOTM data and was marginally outperformed by the NMF CF model on the Goodbooks dataset (Table 1). Interestingly, the prototype model showed a significant *negative* correlation on this metric for the AOTM data, but a positive correlation on the

Goodbooks dataset. These results are consistent with the prototype model's low aggregate fit to human raters (as seen in Figure 3), and suggest that it tends to achieve better predictions on AOTM problems that humans have more difficulty with (and vice versa). This effect may be related to the difference in size between AOTM and Goodbooks playlists. In the AOTM case, shorter average playlists produce sparser prototypes, which in turn result in more items receiving the same or similar recommendation scores for a given problem. This more diffuse recommendation profile (see the Prototype model's near-equal generalization probabilities for in-playlist and in-genre items in Figure 4) will potentially prove advantageous on ambiguous problems (e.g., when there are fewer seeds) that prove difficult for humans, but not be able to capitalize on the added information present when the number of seeds increases, and human generalizations improve.

A similar picture emerged when we evaluated the models in terms of their ability to reproduce human ratings. On both datasets, there existed significant differences between models ($F(4, 1182) = 160.2, p < 0.001$). In each case the Bayesian and item-based CF models were indistinguishable under Tukey's HSD (AOTM: $M = 0.53$, $SD = 0.24$ and $M = 0.53$, $SD = 0.24$, respectively; Goodbooks: $M = 0.50$, $SD = 0.23$ and $M = 0.52$, $SD = 0.23$), while both the matrix factorization and prototype models showed significantly lower performance (AOTM: $M = 0.44$, $SD = 0.22$ and $M = 0.39$, $SD = 0.22$; Goodbooks: $M = 0.48$, $SD = 0.23$ and $M = 0.28$, $SD = 0.21$). On both datasets, the prototype model was significantly worse than any of the other models considered.

Table 2

Overall model correlation with average human recommendation probabilities, stratified by recommendation level and number of seeds.

Dataset	Item-based CF Model	NMF CF Model	Prototype Model	Generalization Model
AOTM	$R = 0.93, p < 0.001$	$R = 0.94, p < 0.001$	$R = 0.94, p < 0.001$	$R = 0.99, p < 0.001$
Goodbooks	$R = 0.97, p < 0.001$	$R = 0.98, p < 0.001$	$R = 0.63, p = 0.012$	$R = 0.99, p < 0.001$

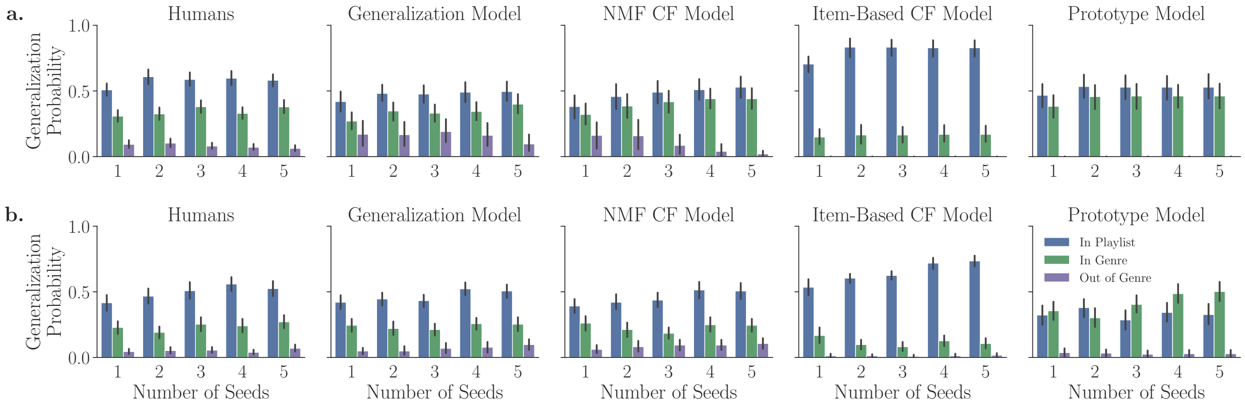


Figure 4. Human and model recommendation probabilities as a function of item category and number of seeds for (a) Art of the Mix and (b) Goodbooks datasets.

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 seeds (Figure 4). Drawing inspiration from Xu and Tenenbaum (2007), recommendations were broken down into in-playlist, in-genre, and out-of-genre items, allowing us to calculate the model's tendency to generalize at each level. Whereas the human ground-truth F_1 scores indicated that both the item-based CF and Bayesian generalization models were approximately equally capable of reproducing human selection profiles, this finer-grained analysis revealed that the relatively coarse human-ground truth 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-item 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 (Table 2).

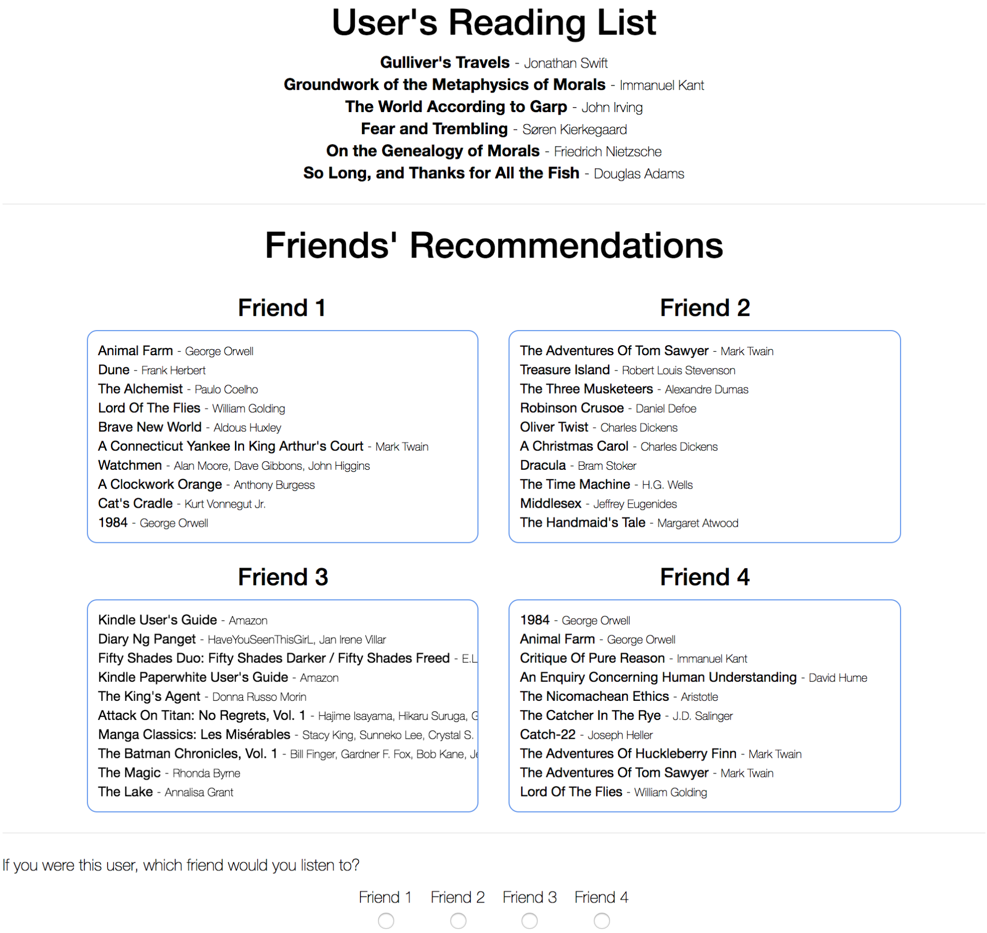
Finally, in order to account for the different number of parameters across the different models under consideration, we computed the Bayesian information criterion (BIC; Schwarz et al., 1978) for each model and performed a model comparison. The probability of an item x_i under the model with parameters θ was computed from the model's score, s_i, using the Luce choice rule (Luce, 1959):

$$p(x_i|\theta) := \frac{e^{\beta s_i}}{\sum_{j \neq i} e^{\beta s_j}}$$

where β is a tuning parameter that was optimized independently for each model.

On the AOTM dataset, the difference in BIC between the Bayesian model (the best performing) and the NMF model (the second best performing model) was 929.08. According to the scale proposed in Kass and Raftery (1995), this difference constitutes "very strong" evidence in favor of the Bayesian generalization model. In contrast, on the Goodbooks dataset the difference in BIC between the NMF model and the Bayesian model was 64.89, providing very strong evidence in favor of the NMF model.

The Bayesian model's inability to outperform the NMF model on the Goodbooks data when controlling for differences in the number of model parameters may reflect a fundamental difference in user intent between the two datasets. Whereas the Art of the Mix website encouraged users to submit themed playlists and curated mixtapes, each playlist in the Goodbooks dataset reflected the entire set of books a user rated. In many cases this included books the user did not enjoy alongside books they did, along with mixtures of multiple genres/interests within the same reading list. One consequence of more heterogenous playlists in the Goodbooks dataset is that a psychological model of generalization will be less equipped to model the data effectively.



User's Reading List

Gulliver's Travels - Jonathan Swift
Groundwork of the Metaphysics of Morals - Immanuel Kant
The World According to Garp - John Irving
Fear and Trembling - Søren Kierkegaard
On the Genealogy of Morals - Friedrich Nietzsche
So Long, and Thanks for All the Fish - Douglas Adams

Friends' Recommendations

Friend 1

Animal Farm - George Orwell
Dune - Frank Herbert
The Alchemist - Paulo Coelho
Lord Of The Flies - William Golding
Brave New World - Aldous Huxley
A Connecticut Yankee In King Arthur's Court - Mark Twain
Watchmen - Alan Moore, Dave Gibbons, John Higgins
A Clockwork Orange - Anthony Burgess
Cat's Cradle - Kurt Vonnegut Jr.
1984 - George Orwell

Friend 2

The Adventures Of Tom Sawyer - Mark Twain
Treasure Island - Robert Louis Stevenson
The Three Musketeers - Alexandre Dumas
Robinson Crusoe - Daniel Defoe
Oliver Twist - Charles Dickens
A Christmas Carol - Charles Dickens
Dracula - Bram Stoker
The Time Machine - H.G. Wells
Middlesex - Jeffrey Eugenides
The Handmaid's Tale - Margaret Atwood

Friend 3

Kindle User's Guide - Amazon
Diary Ng Panget - HaveYouSeenThisGirL, Jan Irene Villar
Fifty Shades Duo: Fifty Shades Darker / Fifty Shades Freed - E.L
Kindle Paperwhite User's Guide - Amazon
The King's Agent - Donna Russo Morin
Attack On Titan: No Regrets, Vol. 1 - Hajime Isayama, Hikaru Suruga, G
Manga Classics: Les Misérables - Stacy King, Sunneko Lee, Crystal S.
The Batman Chronicles, Vol. 1 - Bill Finger, Gardner F. Fox, Bob Kane, J
The Magic - Rhonda Byrne
The Lake - Annalisa Grant

Friend 4

1984 - George Orwell
Animal Farm - George Orwell
Critique Of Pure Reason - Immanuel Kant
An Enquiry Concerning Human Understanding - David Hume
The Nicomachean Ethics - Aristotle
The Catcher In The Rye - J.D. Salinger
Catch-22 - Joseph Heller
The Adventures Of Huckleberry Finn - Mark Twain
The Adventures Of Tom Sawyer - Mark Twain
Lord Of The Flies - William Golding

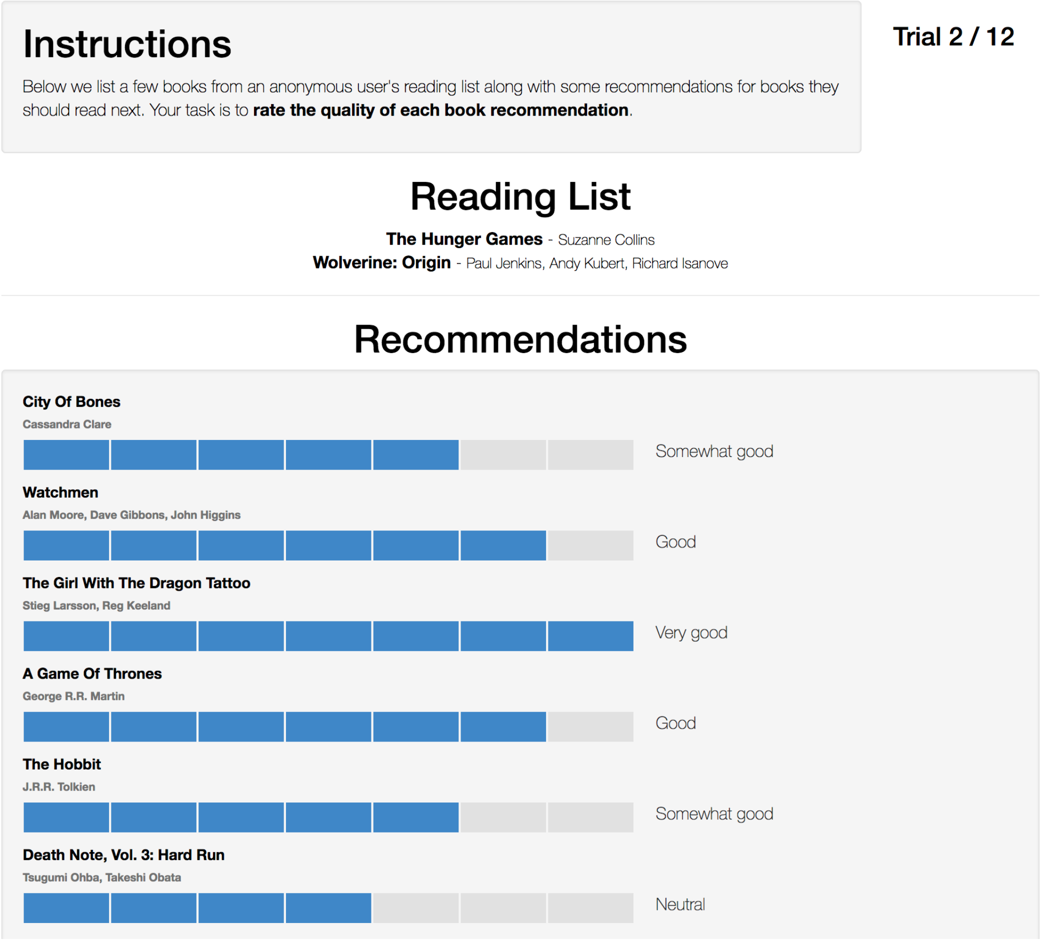
If you were this user, which friend would you listen to?

Friend 1 ○ Friend 2 ○ Friend 3 ○ Friend 4 ○

Figure 5. Web interface for the aggregate version of the rating task, displaying a trial on the Goodbooks data.

Experiment 2: Rating model recommendations

The results of Experiment 1 indicate that a model from psychology – the Bayesian model of generalization – is capable of performing as well or better than several collaborative filtering approaches on a playlist completion task. Further, both the Bayesian and matrix factorization models showed strong fits to observed human recommendation behavior on the AOTM and Goodbooks datasets. A natural next question is whether people – potential users of these recommendation systems – are also sensitive to the differences between models.



Instructions

Below we list a few books from an anonymous user's reading list along with some recommendations for books they should read next. Your task is to **rate the quality of each book recommendation**.

Trial 2 / 12

Reading List

The Hunger Games - Suzanne Collins
Wolverine: Origin - Paul Jenkins, Andy Kubert, Richard Isanove

Recommendations

City Of Bones
Cassandra Clare
Somewhat good

Watchmen
Alan Moore, Dave Gibbons, John Higgins
Good

The Girl With The Dragon Tattoo
Stieg Larsson, Reg Keeland
Very good

A Game Of Thrones
George R.R. Martin
Good

The Hobbit
J.R.R. Tolkien
Somewhat good

Death Note, Vol. 3: Hard Run
Tsugumi Ohba, Takeshi Obata
Neutral

Figure 6. Web interface for the individual version of the rating task, displaying a trial on the Goodbooks data.

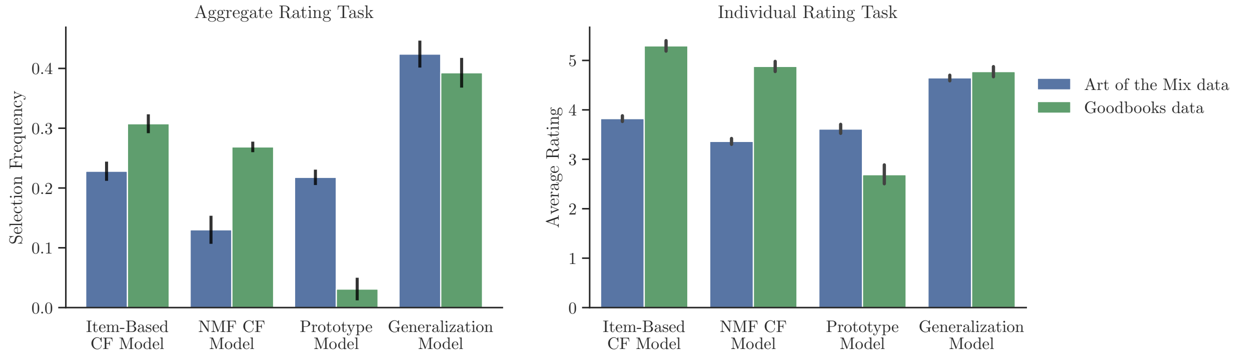


Figure 7. Human ratings for each version of the rating task from Experiment 2.

Methods

Participants. Participants were recruited via Amazon Mechanical Turk and paid \$2.50 in the individual rating condition and \$1.00 in the aggregate rating condition. After dropping participants that did not pass the genre pretest or otherwise failed to complete the experiment there were a total of $n = 124$ subjects remaining in the AOTM aggregate condition, $n = 79$ subjects in the Goodbooks aggregate condition, $n = 64$ subjects in the AOTM individual rating condition, and $n = 46$ subjects in the Goodbooks individual rating condition. As above, all data was collected with the consent of the participants and approval from the UC Berkeley Committee for Protection of Human Subjects.

Stimuli. To generate new stimuli for our second experiment we created “hybrid playlists” consisting of an equal mixture of items from two randomly selected playlists from the AOTM or Goodbooks data in Experiment 1. We used these hybrid playlists within the recommendation framework from Experiment 1 to generate the top recommendations from each modeling approach.

Participants in the aggregate rating condition were shown the top 10 recommendations of each model (each grouped and labeled as the recommendations of a different “friend”) and a “user reading list“ consisting of two, four, or six items from the hybrid playlist. The model recommendations were filtered to remove any items and artists already appearing in the user reading list (Figure 5).

Participants in the individual rating condition were shown a “reading list“ consisting of two, four, or six items from a hybrid playlist, along with the shuffled union of the top 10 recommendations for each model on that playlist. The model recommendations were filtered to remove any items and artists appearing in the reading list (Figure 6).

Design and procedure. On each trial we provided participants with two, four, or six seed items from a hybrid playlist and asked them to rate the quality of the recommendations produced by each model either in aggregate or individually. Participants in the aggregate rating condition were told that an anonymous user had recently listened to / read the seed items and asked four of their friends to provide recommendations for things to listen to / read next. Participants were then asked to select which friend they thought provided the best recommendations. Participants completed a maximum of 12 rating trials contingent on their performance on a genre pretest.

Participants in the individual rating condition were asked to rate each one of the items in the recommendation list in terms of how likely each was to be in the playlist containing the observed seed items. Participants completed a maximum of 12 rating trials contingent on their performance on a genre pretest.

Results

In the individual version of the rating task, we found that participants significantly favored songs recommended by the Bayesian generalization model in comparison to any of the other models evaluated for the AOTM data ($F(3, 8396) = 113.397$, $p < 0.001$; Figure 7, right panel). There were no significant interactions between the recommendation rank and the model. On the Goodbooks data, however, we found that participants rated the individual recommendations of the item-based CF model significantly higher on average than the prototype, Bayesian, or matrix CF models ($F(3, 3534) = 216.08$, $p < 0.001$). Moreover, there was no significant difference in average individual ratings for the matrix CF and the Bayesian model on the Goodbooks data.

In the aggregate version of the rating task, a more robust trend emerged (Figure 7, left panel). Here, the collective recommendations of the Bayesian model of generalization were selected as best more often than the recommendations of any of the other models across both datasets (AOTM: $\chi^2(3, N = 684) = 126.28$, $p < 0.001$; Goodbooks: $\chi^2(3, N = 387) = 111.45$, $p < 0.001$). 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 suggest that users may also favor recommendations from models which reproduce larger proportions of human recommendation behavior.

Discussion

Online behavioral datasets offer an opportunity to evaluate theories of cognition at a scale rarely seen in traditional laboratory studies. To demonstrate this potential, we began by evaluating representative models from both the cognitive modeling and machine learning literatures on two massive online datasets. Using data from user interactions online we were able to capture people's spontaneous behavior spanning over a decade. Notably, both datasets had more than 100,000 unique observations and could be leveraged without incurring any experimentation cost–properties rarely seen in traditional laboratory studies.

To illustrate how web datasets can come to bear on psychological questions we outlined the correspondence between many popular online activities and more fundamental psychological processes. As a concrete example, we drew a connection between recommendation and generalization, and showed that we could evaluate psychological models of word-learning and concept acquisition alongside machine learning approaches on web recommendation datasets. We found that a psychological model, the Bayesian model of generalization, performed competitively in comparison to standard collaborative filtering approaches across a collection of psychological and information-retrieval metrics, even out-performing other models on a music recommendation dataset. Moreover, we found that human experts significantly favored the recommendations from the Bayesian model over other recommendation approaches, underscoring the value of taking the psychological dimension of common machine learning tasks seriously.

The opportunity to reformulate standard machine learning applications in terms of human cognition is an exciting avenue for researchers in both machine learning and psychology. As we have argued above, many online tasks engage directly with fundamental cognitive abilities. Traditionally the data generated during these online interactions have been handled by engineers and computer scientists, but our results indicate the value that can 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 influence current approaches to predicting behavior by explicitly engaging with the psychological origins of the data (Griffiths, 2015; Jones, 2016; Paxton & Griffiths, 2017).

Our results offer several takeaways for psychologists. Most generally, we draw attention to the link between fundamental cognitive phenomena studied by behavioral scientists and common online activities. This link opens the door for researchers to use existing online behavioral datasets to advance theories of cognition. We demonstrate one such example by fitting several well-known models from the generalization literature on two online recommendation datasets, allowing us to evaluate their performance in a noisier, more naturalistic setting. Additionally, by comparing the performance of the cognitive models to popular approaches from machine learning, we demonstrate how the abundance of online behavioral data can open the door for greater correspondence between theory-based and data-driven modeling approaches. Finally, by showing that a cognitive model can perform as well or better than two popular approaches from the machine learning literature, we illustrate the potential for psychologists to make inroads on tasks that have traditionally been studied under the umbrella of computer science.

Context of Research. The current work arose out of our discussions of the psychological aspects of many standard tasks in machine learning and information retrieval. Each of the authors has been involved in efforts to expand and scale cognitive modeling techniques to internet-scale data. In the future we hope to find new opportunities for researchers in academia and industry to collaborate on curating behavioral datasets that can be used to answer fundamental questions about the mind.

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 conference 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.

Bourgin, D. D., Abbott, J. T., & Griffiths, T. L. (2018). Recommendation as generalization: Evaluating cognitive models in the wild. In *Proceedings of the 40th annual conference of the cognitive science society* (Vol. 40).

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

Jones, M. N. (Ed.). (2016). *Big data in cognitive science.* New York, NY: Routledge.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

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. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556–562).

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.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York, NY: John Wiley & Sons.

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.

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638.

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

Rijsbergen, C. J. V. (1979). *Information retrieval* (2nd ed.). Newton, MA, USA: Butterworth-Heinemann.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–856.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

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 *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. In *Advances in neural information processing systems* (pp. 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*.

Zajac, Z. (2017). Goodbooks-10k: a new dataset for book recommendations. *FastML*. http://fastml.com/goodbooks-10k.

Zhang, S., Wang, W., Ford, J., & Makedon, F. (2006). Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the 2006 siam international conference on data mining* (pp. 549–553).

Appendix A: Correspondence between exemplar model and item-based CF

For a user U_k, an item-based collaborative filtering model defines score for a new item y to be

$$\mathrm{score}(y \mid U_k) \propto \sum_{j \in S_y} R_k(j) \cdot \mathrm{sim}(y, j) \tag{7}$$

where $R_k(j)$ is user k's rating on item j, $\mathrm{sim}(y, j)$ is the similarity between item y and item j, and S_y is the set of IDs for the m most similar items to item y that have also been rated by user U_k.

The generalized context model (GCM; Nosofsky, 1986), a popular formulation of the exemplar model, defines the probability that a participant will associate a new item y with category C_k after observing $X = \{e_1, \ldots e_N\}$ positive exemplars as

$$P(y \in C_k | X) \propto b_k \left(\sum_{j=1}^{N} W_{k,j} \cdot \mathrm{sim}(y, j) \right)^{\gamma} \tag{8}$$

where, $W_{j,k}$ is the memory strength between category C_k and exemplar j, $b_k > 0$ is a response bias for category k, $\mathrm{sim}(y, j)$ is the similarity between y and exemplar j, and $\gamma > 0$ is a response scaling parameter.

The correspondence between models is straightforward: concepts, C, and memory strengths, W, in the GCM correspond to users, U, and ratings, R, in item-based CF, while the additional parameters b_k and γ in the GCM are set to 1 for item-based CF. If we assume that a user has rated all N in the database and $m = N$, the item-based CF model is equivalent to the GCM.

Appendix B: Model parameters

All model parameters were arrived at independently via random search (Bayesian generalization model) or grid search (all other models) against the stated objective. The parameter grids / grid ranges for each model were:

Bayesian Generalization Model:

$\sigma \sim$ Uniform(1, 1000)

$\epsilon \sim$ Uniform(0, 0.5)

$\lambda_g \sim$ Uniform(0, 0.7)

NMF number of latent factors: $[2, 5, 10, 25, 50, 100, 150, 200, 250]$

NMF Binarization threshold $\sim$ LogUniform(0, 0.1)

Item-based CF Model:

Number of nearest neighbors: $[100, 200, 400, 600, 800, 1000, \text{All}]$

Include genre playlists: [True, False]

Prototype Model:

λ_p: $[1, 10, 15, 25, 50, 75, 100]$

Include genre playlists: [True, False]

NMF CF Model:

Number of latent factors: $[2, 5, 10, 25, 50, 100, 150, 200, 250]$

Include genre playlists: [True, False]

AOTM Dataset. For human ground-truth F_1 correlations (Table 1), the best performance for the item-based CF model was found by averaging over all items in the hypothesis space without including genre playlists. The best matrix factorization CF model used 5 latent factors and did not include genre playlists. The best Bayesian generalization model used $\sigma = 861$, $\epsilon = 0.2899158150127287$, and $\lambda_g = 0.3275493039121179$, operating on the NMF-transformed feature space with 50 latent factors and a binarization threshold of $9.680514395205266 \times 10^{-5}$. The best prototype model used $\lambda_p = 15$ and did not include genre playlists.

For human recommendation probability correlations (Table 2), the best performance for the item-based CF model was found by averaging over the top 100 most similar items, including genre playlists in the overall hypothesis space. The best matrix factorization CF model used 10 latent factors and included genre playlists. The best Bayesian generalization model used $\sigma = 309$, $\epsilon = 0.2553692504953126$, and $\lambda_g = 0.44184686010121443$, operating on the NMF-transformed feature space with 100 latent factors and a binarization threshold of 1×10^{-12}. The best prototype model used $\lambda_p = 15$ and included genre playlists in the hypothesis space.

Goodbooks Dataset. For human ground-truth F_1 correlations (Table 1), the best performance for the item-based CF model was found by averaging over the top 100 most similar items and including genre playlists in the hypothesis space. The best matrix factorization CF model used 200 latent factors and also included genre playlists. The best Bayesian generalization model used $\sigma = 576$, $\epsilon = 0.3548974755071327$, and $\lambda_g = 0.6492645958009968$, operating on the NMF-transformed feature space with ten latent factors and no binarization. The best prototype model used $\lambda_p = 1$ and included genre playlists in the hypothesis space.

For human recommendation probability correlations (Table 2), the best performance for the item-based CF model was found by averaging over the top 100 most similar items, including genre playlists in the overall hypothesis space. The best matrix factorization CF model used 250 latent factors and did not include genre playlists. The best Bayesian generalization model used $\sigma = 983$, $\epsilon = 0.323958956941186$, and $\lambda_g = 0.1423489705102997$, operating on the NMF-transformed feature space with 20 latent factors and a binarization threshold of 0.04270991600377532. The best prototype model used $\lambda_p = 1$ and included genre playlists.