Limitations to Optimal Search in Naturalistic Active Learning

Lisheng He (felix8.he@gmail.com)
SILC Business School, Shanghai University
Shanghai, China

Russell Richie (drrichie@sas.upenn.edu)
Children's Hospital of Philadelphia
Philadelphia, PA, USA

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA, USA

Abstract

We introduce a new empirical paradigm for studying naturalistic active learning, as well as new computational tools for jointly modeling algorithmic and rational theories of information search. Subjects in our task can ask questions and learn about hundreds of everyday items, but must retrieve queried items from memory. In order to maximize information gain, subjects need to retrieve sequences of dissimilar items. We find that subjects are not able to do this. Instead, associative memory mechanisms lead to the successive retrieval of similar items, an established memory effect known as semantic congruence. The extent of semantic congruence (and thus suboptimality) is unaffected by task instructions and incentives, though subjects are able to identify efficient query sequences when given a choice. Overall, our results indicate that subjects can distinguish between optimal and suboptimal search if explicitly asked to do so, but have difficulty implementing optimal search from memory. We conclude that associative memory processes place critical restrictions on people's ability to ask good questions in naturalistic active learning tasks.

Keywords: Active learning; Memory search; Computational modeling; Rational cognition

Introduction

People often choose what information they want to gather. This kind of learning is known as active learning, and has been the subject of intense study in recent years in several fields. Although there are many questions to ask about active learning, perhaps the most pressing question about active learning is this: how and why do people seek the particular information they seek?

Theories of rational cognition (Griffiths et al., 2010) provide an increasingly popular answer to this question. These theories propose that people search for information optimally; that is, they generate queries that provide the most information possible. The rational account of active learning has been successfully tested in many domains in psychology (for review see Coenen et al., 2019), however, one challenge for the rational account of active learning involves the role of semantic similarity in memory search. Optimal search often requires asking questions that are dissimilar to each other, as asking the same (or a similar) question repeatedly will usually provide the same (or similar) information. Consider, for example, a task in which the learner has to determine how much of a new nutrient there is in different food items. The learner can ask questions about each item sequentially (how much of the nutrient is there in a *strawberry*? how much in a *blueberry*? how much in an *egg*?) and must retrieve each item (*strawberry*, *blueberry*, *egg*) from memory prior to the query. As similar items usually have similar properties, for the questions to be maximally informative, the queried items must be as different to each other as possible. It is much better to follow up a query about *strawberry* with a query about *egg* than a query about *blueberry*.

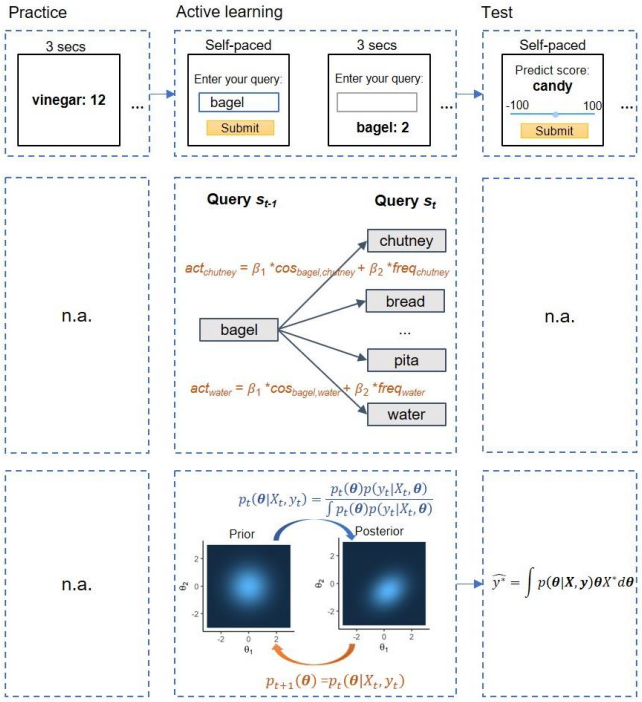

Figure 1. Experimental and modeling setup. Top: The practice, active learning, and test phases of the task. Middle: The Markov memory model which guides query generation in the learning phase. Bottom: The Bayesian model, which uses the information in the learning phase to give responses in the test phase.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

This optimal search strategy is the opposite of what researchers have observed in most recall tasks. Typically, when asked to retrieve items from memory, people generate sequences of semantically similar items, an effect known as semantic congruence. The semantic congruence effect is remarkably robust, and emerges across a variety of tasks including free association (De Deyne et al., 2019), free recall from lists (Howard & Kahana, 2002), semantic memory search (Bousfield & Sedgewick, 1944), and memory-based decision making (Aka & Bhatia, 2021). This is due to the associative structure of memory (Atkinson & Shiffrin, 1968). Retrieved items cue successive items based on their strength of association. Items that are similar are more associated with each other, which is why the retrieval of *strawberry* is more likely to cue *blueberry* than *egg*.

How is this conflict resolved in naturalistic active learning tasks? Are people able to search optimally and retrieve sequences of dissimilar items, or are they fundamentally constrained by the associative memory processes that lead to semantic congruence in other recall tasks?

Overview of Experimental Paradigm

Unfortunately, most studies on active learning are conducted under rarified conditions that do not require memory search. This is largely due to the difficulty in modeling naturalistic active learning, in which people can search over and ask questions about thousands of items and entities. Fortunately, recent work has shown the promise of distributed semantics models (DSMs) for solving this problem. DSMs use patterns of word-word co-occurrence in large collections of texts, to build vector representations of millions of real words and phrases. Words that are semantically similar, like *strawberry* or *blueberry*, tend to have similar distributions in text, and therefore end up with vector representations that are close to each other. For this reason, DSMs can describe many psychological phenomena (see Bhatia et al. 2019 for review), and, importantly, can predict semantic congruence effects in memory search.

We used DSMs to model memory search in a new naturalistic active learning task (Fig. 1, top panel). In the task, subjects (a) learned a novel property by querying 20 different entities in a category and getting feedback on those entities' property scores, and then (b) in the test phase predicted the scores of a fixed set of 20 test items. Prior to the active learning task, subjects participated in a practice phase, where they were presented five items and the corresponding property scores. Property scores for the entities were constructed by prespecified random linear functions on their word2vec (Mikolov et al., 2013) DSM vectors, giving similar items similar property scores. Exps. 1a, 2 and 3 implemented this task with 1,594 food items, while Exp. 1b implemented it with 1,734 animals. Additionally, Exp. 2 compared the queries in the active learning task with recall in a standard semantic memory search task. Exp. 3 provided detailed coaching on how to do well in the active learning task. Exp. 4 did not directly use the task but instead asked subjects to judge the optimality of search sequences in the task.

There were 396 subjects in Exps. 1a and 1b, 102 subjects in Exp. 2, 100 subjects in Exp. 3, and 48 subjects in Exp. 4. Subjects in all experiments were recruited from Prolific Academic and were US residents that were fluent in English. They were given a base payment of $2, and were given a bonus of $1.00 if their test performance (measured by RMSE) was in the top 10%, and $.50 if they were in the top 50%. Exps. 2 and 3 were pre-registered.

Semantic Congruence in Search

We used a computational model (Fig. 1 middle panel) to formally capture memory retrieval dynamics in the active learning phase (Exps. 1a, 1b, 2 and 3). In line with classic memory retrieval models, we assumed that subjects searched for a word among all candidate queries in the memory space S. Assuming the Markov property, the model predicted the switch from one query s_{t-1} to another query s_t using transition probabilities $\Pr[s_t|s_{t-1}]$, where $s_{t-1}, s_t \in S$ and $t \in \{2,3,...,T\}$ are the time steps. We allowed $\Pr[s_t|s_{t-1}]$ to be a function of item activation, which in turn depended on two key cognitive mechanisms -- semantic congruence and word frequency -- giving us: $\Pr[s_t|s_{t-1}] = \sigma(\beta_1 sim_{s_{t-1},s_t} + \beta_2 freq_{s_t})$, where sim_{s_{t-1},s_t} is the cosine-similarity between s_{t-1} and s_t, $freq_{s_t}$ represents the frequency (log-transformed) of candidate query s_t, and $\sigma(\cdot)$ is the softmax function that sets $\sum_{s_t \in S} \Pr[s_t|s_{t-1}] = 1$.

Hierarchical Bayesian model fitting provided both group- and individual-level estimation of β_1 and β_2 (Table 1). On the group level, cosine similarity had a strong positive effect on sequential memory search for food items in Exps. 1a, 2, and 3 and for animals in Exp. 1b. Likewise, frequent words/phrases were much more likely to be queried than infrequent ones for both foods in Exps. 1a, 2 and 3 and animals in Exp. 1b. The individual level estimation also suggested that a majority of our subjects displayed these tendencies. These results suggest that the underlying cognitive processes in our naturalistic active learning task resembled those underlying a typical memory task, processes which likely lead to sub-optimal queries.

Table 1: Mean and 95%CI of memory model parameters

	β_1	β_2
Exp. 1a	0.62 [0.57, 0.67]	0.97 [0.93, 1.02]
Exp. 1b	0.71 [0.66, 0.76]	0.93 [0.90, 0.97]
Exp. 2	0.68 [0.63, 0.74]	0.86 [0.80, 0.91]
Exp. 3	0.42 [0.36, 0.50]	0.87 [0.80, 0.94]

Bayesian Learning Model

Subjects learned about the target property from the scores given as feedback to queries. To formally capture this dynamic learning process, we assumed that the subjects were ideal Bayesian learners who took as input the scores y_t of their query (quantified by a DSM vector X_t) and learned a

linear mapping between them (Fig. 1, bottom panel). Subjects updated their belief of weights $\boldsymbol{\theta}$ that determined the linear mapping after observing each pair of X_t and y_t using the Bayes rule: $p_t(\boldsymbol{\theta}|X_t, y_t) = \frac{p_t(\boldsymbol{\theta})p(y_t|X_t,\boldsymbol{\theta})}{\int p_t(\boldsymbol{\theta})p(y_t|X_t,\boldsymbol{\theta})}$. In the test phase, we assumed that subjects made predictions on the test items based on the updated posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X}$ is the design matrix corresponding to their 20 queries in the active learning phase and $\boldsymbol{y}$ are the corresponding scores. The predicted scores at test can be written as $\widehat{y^*} = \int p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})\boldsymbol{\theta}X^* d\boldsymbol{\theta}$, where X^* is the vector corresponding to the test item.

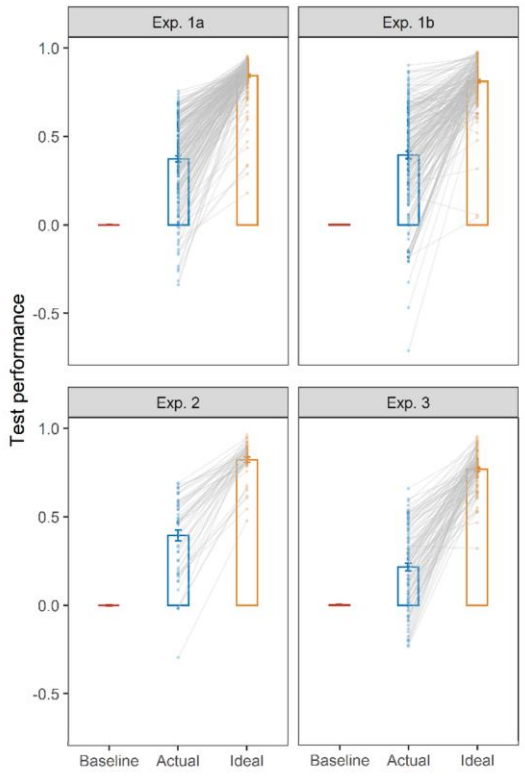

Figure 2. Baseline, actual and ideal test Pearson's R across experiments. Each point corresponds to a subject.

The ideal Bayesian learning model captured the between-subject variation in test performance (see Fig. 2 using Pearson's R as the measure). In the pooled analysis on the full datasets across experiments (Exps. 1a, 1b, 2 and 3; N = 546), we found that the predicted performance at test by the Bayesian learning model trained on each subject's queries was correlated with the actual performance of the subjects measured by both Pearson's R and root-mean-squared error (RMSE) (Pearson's R: r = 0.313, $p < 10^{-13}$; RMSE: r = 0.269, $p < 10^{-9}$). In separate analyses, the actual test Pearson's R was positively related to the predicted test Pearson's R by the ideal Bayesian learning model in Exps. 1a, 1b, and 2 (Exp. 1a: r = 0.418, $p < 10^{-9}$; Exp. 1b: r = 0.236, $p < .001$; Exp. 2: r = 0.528, $p < .0001$). In Exp. 3, in which participants were given coaching on how to generate efficient queries, we found that the Bayesian learning model failed to describe participant heterogeneity (r = 0.040, p = .696). Similar patterns emerged when we quantified test performance with RMSE (Exp. 1a: r = 0.219, p = .002; Exp. 1b: r = 0.166, p = .019; Exp. 2: r = 0.157, p = .277; Exp. 3: r = 0.021, p = .833).

Note that subjects' actual test performance was significantly better than the baseline model that assigned random scores at test. However actual test performance did not reach the accuracy levels predicted by the ideal Bayesian learning model (p's $< 10^{-12}$ in all experiments in Fig. 2), likely because of additional sources of noise during the test phase.

Effect of Semantic Congruence on Learning

The ideal Bayesian learning model also allowed us to evaluate query efficiency with Bayesian D-optimality, one of the most often used optimality criteria (Myung & Pitt, 2009). Mathematically, Bayesian D-optimality of the full set of queried items is the determinant of the Fisher information matrix $D = \det\{\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} + \Sigma^{-1}\}$, where $\boldsymbol{X}$ is the 11×20 design matrix corresponding to the 20 queried items, and Σ is the 11×11 covariance matrix prior to querying the items. At the beginning of the experiment, Σ is set as an identity matrix, corresponding to the standard multivariate normal prior distribution on $\boldsymbol{\theta}$. Intuitively, if the queried items are sparsely distributed in the space, the design matrix typically has a high Bayesian D-optimality. By contrast, if the queried items are close to one another, the design is likely to have a low Bayesian D-optimality.

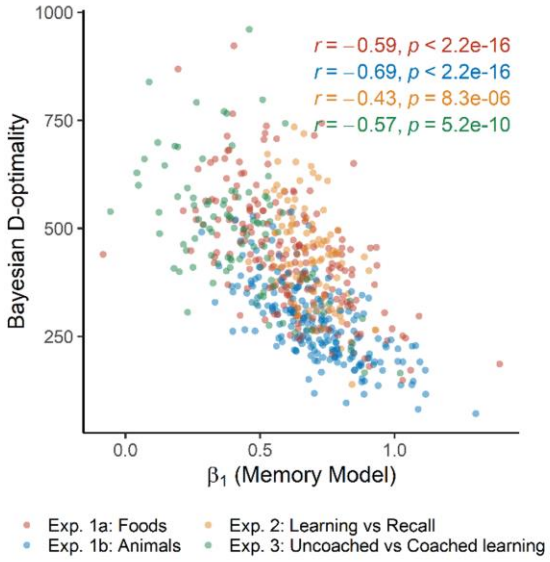

Figure 3. Correlations between semantic congruence parameter in memory model (β_1) and Bayesian D-optimality. Each point corresponds to a subject.

We correlated the Bayesian D-optimality of each subject's query sequence with the estimated degree of semantic congruence, β_1, in the memory model (Fig. 3). As expected, these two variables entail a strong tradeoff, such that subjects whose queries were more positively influenced by the

similarity with previous items had lower levels of Bayesian D-optimality.

Semantic *in*congruence was also associated with better test performance as predicted by the ideal Bayesian learning model. In a pooled analysis across Exps. 1a, 1b, 2 and 3 (N = 546), we found that β_1 was positively correlated with the ideal Bayesian learning model's predicted test RMSE (r = 0.446, $p < 10^{-15}$) and negatively with its predicted test Pearson's R (r = -0.093, p = .029). Additionally, subjects who displayed more semantic incongruence also performed better at test. β_1 was positively correlated with subjects' actual RMSE (r = 0.188, p < .001) and negatively correlated with subjects' actual Pearson's R (b = -0.212, p = .036) at test.

To more systematically examine the tradeoff between semantic congruence in memory retrieval and optimal search in active learning, we simulated a semantic similarity-based retrieval strategy and compared its Bayesian D-optimality, as well as the predicted test performance by the ideal Bayesian learning model, with that of a number of other retrieval strategies (Fig. 5). The semantic similarity-based strategy produced the lowest query Bayesian D-optimality and achieved the worst test performance among all retrieval strategies. In stark contrast, the D-optimality Greedy strategy that kept selecting the most informative query according to the Bayesian D-optimality criterion always produced the lowest semantic congruence in queries and the best performance at test. Retrieval strategies that searched based on word frequency or based on random sampling achieved intermediate D-optimality and accuracy rates at test.

The subjects' actual query sequences were far from optimal, as compared with the D-optimality Greedy strategy (Fig. 5). They also displayed more semantic congruence, and achieved lower Bayesian D-optimality, than the random or frequency-based queries. Overall, these results, once again, suggest that subjects were unable to ask the most informative questions when the questions had to be generated from memory.

Effect of Task Demands

Do subjects' queries vary with task requirements? We examined this in two experiments. In Exp. 2, we compared an active learning condition with a traditional semantic memory search condition in which subjects were asked to list all foods that they could think of. As can be seen in Fig. 4, the conditions differed neither on semantic congruence (i.e.β_1) ($t_{96.7}$ = 0.818, p = .416), nor on Bayesian D-optimality of the queries ($t_{99.8}$ = 0.719, p = .474). In Exp. 3, we compared a coached learning condition, in which subjects were explicitly instructed to query more efficiently by sampling dissimilar items, with an uncoached condition (which was identical to Exp. 1a). The conditions did not differ on either semantic congruence ($t_{79.9}$ = 1.214, p = .229) or Bayesian D-optimality ($t_{93.0}$ = 1.018, p = .311). The conditions did not differ on their predictive performance at test either (test Pearson's R: $t_{88.9}$ = 1.05, p = .297; test RMSE: $t_{77.1}$ = 0.221, p = .826). Both experiments suggest that the semantic congruence effect is so strong that it cannot be moved by simple task demands.

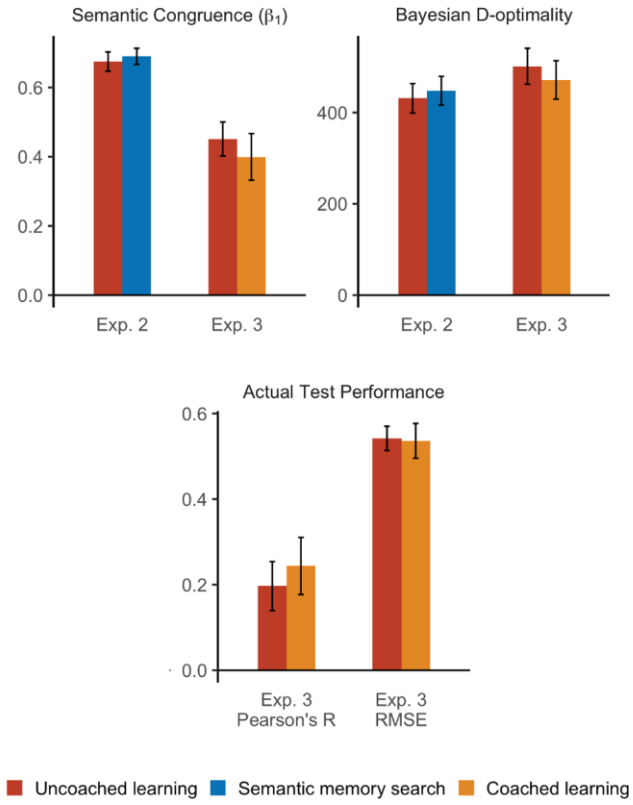

Figure 4. Between-condition comparisons of semantic congruence and Bayesian D-optimality of subjects' queries (Exps. 2 and 3) and actual test performance (Exp. 3). Error bars represent 95% confidence intervals.

Judging Optimality of Queries

Finally, we tested whether subjects could distinguish efficient query sequences from inefficient sequences. For this purpose, we gave subjects in Exp. 4 five pairs of subject-generated sequences. Most (32 of 48) subjects achieved greater than chance accuracy (p=.029 in a binomial test), and in fact the modal accuracy rate was 100%. On the pair level, three pairs achieved accuracy rates significantly higher than random (90%, 67% and 63% respectively, ps < .05) while the other two pairs were more difficult to judge and didn't pass the conventional significance threshold (48% and 52% respectively).

Overall, while subjects in Exps. 1-3 generally queried in a suboptimal manner, subjects in Exp. 4 showed the ability to distinguish between more and less efficient query sets. This is consistent with the idea that subjects can evaluate optimality to some degree, but are unable to consider optimality when generating queries due to universal, powerful constraints on memory retrieval, particularly semantic similarity.

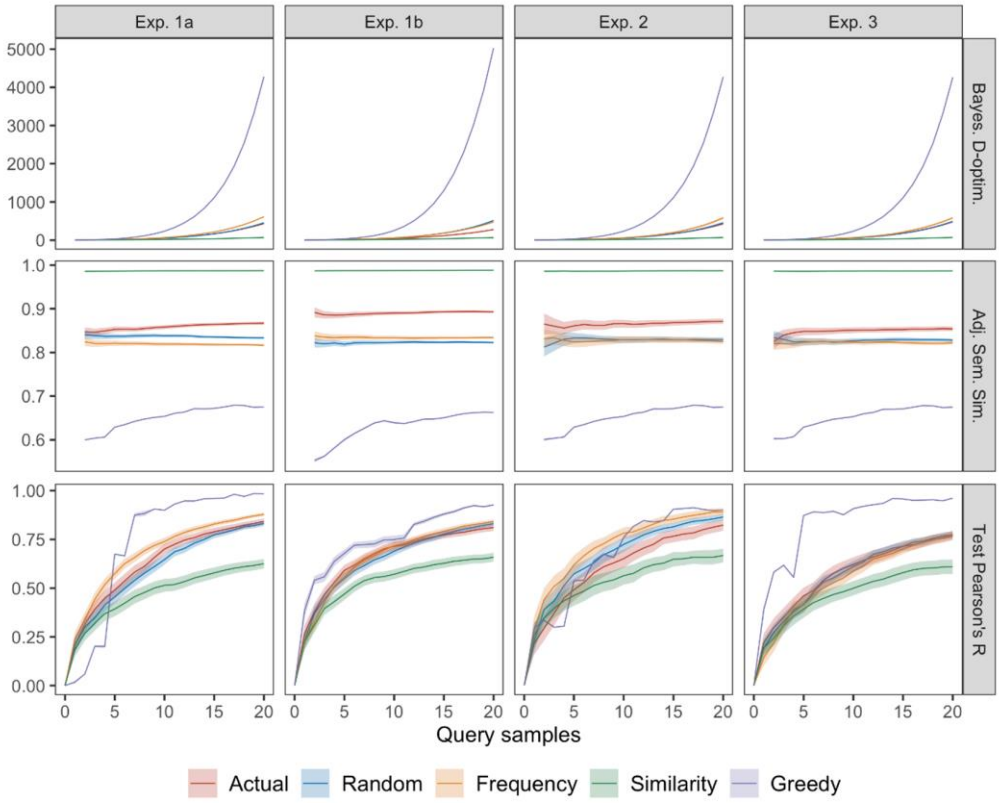

Figure 5. Properties of a Bayesian learning model that makes predictions based on simulated queries. Actual query uses the subjects' actual sequences of queries. Random query randomly selects from all the items with no replacement. Frequency-based query randomly selects from the top-100 most frequent items with no replacement. Similarity-based query keeps selecting the most cosine-similar item that hasn't been queried before. Greedy query keeps selecting the item with the highest Bayesian D-optimality. Shaded bands are 95% CIs.

Discussion

Contrary to the popular optimality hypothesis that claims that people can ask efficient questions that maximize expected information gain, we found that subjects failed to generate optimal inquiries and that the suboptimality was largely due to associative memory search (Exps. 1a & 1b). Additional pre-registered experiments showed that subjects' querying behavior was no more optimal – or less similarity-driven – in our active learning task than a traditional semantic search task (Exp. 2), and no more optimal or less similarity-driven when directly told to query more optimally by querying *dissimilar* items (Exp. 3), suggesting that memory-based active learning is at the mercy of extremely stubborn memory constraints, which are difficult to alleviate by task instructions. A final experiment showed that subjects can distinguish between the more and less optimal query sets, suggesting that subjects understand what optimality entails, but that memory constraints make the spontaneous generation of optimal queries from memory difficult.

Our results stand in stark contrast with the large body of work that finds optimal search in active learning. The theory that people acquire information optimally has been very successful in explaining human inquiry in several domains. However, most prior studies use fairly simple, artificial stimuli, and do not require subjects to generate queries from memory. We thus suggest that the scope of the optimality hypothesis in explaining human active learning may be more limited than previously thought. Indeed, we suspect that any setting in which subjects must formulate sequences of queries in natural language will probably be constrained by memory processes, particularly the similarity-driven associative memory search.

Although associative memory processes curtail optimal active learning, that does not mean that people's memory processes are inherently flawed. Rather, memory serves multiple cognitive functions and the associative biases documented in this paper may reflect optimal tradeoffs between diverging task demands. Indeed, many researchers have argued that association or similarity-driven memory search is part of an optimal system for semantic memory retrieval (Hills et al. 2012). Related work has shown that associative memory processes implicated in judgment and decision biases are adaptive in that they often lead to accurate inference and generalization with minimal cognitive cost (Bhatia, 2017; Tenenbaum & Griffiths, 2001). Regulating these processes in active learning tasks may be too effortful, and people may be optimally trading off performance with the cognitive cost required to succeed in our task (Lieder &

Griffiths, 2020). This theory predicts that even though we were unable to reduce semantic congruence and increase optimal search through coaching, performance may improve with higher incentives or practice. Testing these predictions is an important topic for future work.

Other future directions include the refinement of our memory and learning models. For example, subjects in our study learned about novel target properties. Yet they came into the experiments with idiosyncratic knowledge about food items or animals. Thus, it is likely they held different prior belief about the novel target properties. Since prior belief is not the focus of this paper, we assumed all subjects held the same prior belief in the experiments. In future work, the shape of prior belief can be set as free parameters and the same framework can be used to derive the prior representation of target properties in a given domain. Individual differences in this regard can be revealed. The Bayesian learning model also assumes that subjects maintain a distribution of belief over multiple hypotheses (possible coefficients on the latent representations). However, other research suggests that in a closely related – and not even as complex – active category learning setting, subjects maintain a single hypothesis at a time (Markant & Gureckis, 2013). Previous research also reveals other simple heuristics, such as the split-half heuristic (Navarro & Perfors, 2011) and the likelihood difference heuristic (Nelson, 2005), in active learning tasks. It is possible that such heuristics play a role in the query search in our active learning tasks and, therefore, can be considered in the modeling of algorithmic processes in future research.

Our work contributes to the emerging body of research that offers researchers a naturalistic search domain to study active learning. Additionally, our computational models integrate insights from several fields, and are able to jointly describe both algorithmic memory search processes (which we have specified using a Markov random walk model) as well as the optimality or suboptimality of these search processes for active learning. In this way, our paper presents a powerful new research paradigm for naturalistic active learning. There has been an increasing interest in porting computational cognitive models beyond abstract lab stimuli, to attempt to describe everyday cognition. This has been driven by the availability of new machine learning models that offer quantitative representations for natural entities (see Bhatia & Aka, in press for a review), as well as the growing demand from policy makers and practitioners for theory-driven behavioral and cognitive insights. Our research is part of this trend, and we look forward to future work that applies established algorithmic and rational theories of cognition to rich stimuli sets to better understand human cognition and behavior in the wild.

References

Aka, A. & Bhatia, S. (2021). What I like is what I remember: memory modulation and preferential choice. *Journal of Experimental Psychology: General*, 150(10), 2175–2184.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of Learning and Motivation* (Vol. 2, pp. 89-195). Academic Press

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, 124(1), 1-20.

Bhatia, S. & Aka, A. (in press). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, 29, 31-36.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, 30(2), 149-165.

Coenen, A., Nelson, J.D., & Gureckis, T.M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 26(5), 1548-1587

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3), 987-1006.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357-364.

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval?. *Journal of Memory and Language*, 46(1), 85-98.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119(2), 431-440.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, e1.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology-General*, 143(1), 94–122.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems*

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116(3), 499-518.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, 118(1), 120-134

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112(4), 979-999.

Tenenbaum, J. B., & Griffiths, T. L. (2001, August). The rational basis of representativeness. *In Proceedings of the 23rd annual conference of the Cognitive Science Society*.