# Searching large hypothesis spaces by asking questions

**Alexander N. Cohen (alxcoh@gmail.com)**
Hunter College High School

**Brenden M. Lake (brenden@nyu.edu)**
Center for Data Science
New York University

## Abstract

One way people deal with uncertainty is by asking questions. A showcase of this ability is the classic 20 questions game where a player asks questions in search of a secret object. Previous studies using variants of this task have found that people are effective question-askers according to normative Bayesian metrics such as expected information gain. However, so far, the studies amenable to mathematical modeling have used only small sets of possible hypotheses that were provided explicitly to participants, far from the unbounded hypothesis spaces people often grapple with. Here, we study how people evaluate the quality of questions in an unrestricted 20 Questions task. We present a Bayesian model that utilizes a large data set of object-question pairs and expected information gain to select questions. This model provides good predictions regarding people's preferences and outperforms simpler alternatives.

**Keywords:** Bayesian modeling; active learning

People seem to ask rich and probing questions when faced with uncertainty. Whether someone is learning a new task, meeting a new person, listening to a presentation, or attending a press conference, people ask questions to better understand the state of the world – a form of "active learning" (Gureckis & Markant, 2012). Question asking is associated with several computational challenges, including selecting a question from a possibly infinite set of allowable options, and evaluating its quality against a large hypothesis space of possible world states. The scope of these challenges raises several key questions: Do people ask good questions? And if so, how do people effectively search over large numbers of questions and hypotheses?

The classic game of 20 Questions (20Qs) provides a window into this broader human ability. A round of 20Qs is played between a "game-master" and a "question-asker." The game-master thinks of an object and the question-asker asks up to 20 questions before guessing the identity of the object. The game-master answers each question with either "yes" or "no," with additional options such as "probably" or "probably not" available in variants of the game. Over the course of playing a game, an ideal question-asker would consider thousands of possible objects in the hypothesis space and select questions from an infinite set of options. Moreover, an effective player must continually update the plausibility of the hypotheses with each new piece of information. How do people manage to play this game? And do they ask effective questions as measured by normative metrics?

Previous work has found that people are surprisingly effective question-askers in modified 20Qs-style games with a limited set of possible objects (Eimas, 1970; Denney & Denney, 1973; Thornton, 1982; Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2015). People generally prefer to ask "constraint-seeking" questions such as "Is it alive?" rather than more specific "hypothesis-scanning" questions such as "Is it a frog?," especially when faced with large amounts of uncertainty. This ability seems to develop in childhood (Eimas, 1970; Ruggeri et al., 2015), decline in elderly populations (Denney & Denney, 1973), and can be disrupted when the feature statistics of the game do not match the real world (Nelson et al., 2014). In most cases, the preference for constraint-seeking questions is consistent with maximizing stepwise Expected Information Gain (EIG), an information theoretic measure with theoretical and empirical motivations. EIG is a measure of informativeness under a Bayesian framework, where higher scoring questions provide a larger reduction in the posterior entropy (effectively, providing information about the correct object). People ask questions predicted by this metric when reasoning in small (Eimas, 1970; Nelson et al., 2014; Ruggeri et al., 2015) or highly visual (Gureckis & Markant, 2009; Rothe, Lake, & Gureckis, 2016) hypothesis spaces, but the generality of this account remains an open question.

In this paper, we study how people evaluate questions in a larger and more naturalistic 20Qs task, which includes many broad classes of common objects (e.g., animals, artifacts, household objects, foods, etc). Using a large set of objects and questions as a rough proxy for semantic knowledge, we propose a Bayesian model for computing posterior belief over object hypotheses and the EIG of candidate questions. Human question preferences in both "complete" and "one-shot" 20Qs games are compared with this model along with various alternatives, providing an important test of how the EIG metric generalizes to richer and more naturalistic domains.

## Model of 20 Questions

A Bayesian framework is developed to reason and ask questions in a large hypothesis space of possible objects.

**Data set of 1000 objects.** We used a data set from Palatucci, Pomerleau, Hinton, and Mitchell (2009) of 1000 objects and 218 questions to construct the model. Objects span a variety of broad semantic classes including animals, insects, food, household items, tools, clothing, vehicles, sports, buildings, and other tangible nouns. It excludes specific people, specific places, proper nouns, ideas, verbs, etc. The set of questions concerns higher-level semantic categories (e.g., "Is it an animal?" or "Is it furniture?"), color ("Is it Yellow?"), shape and texture ("Is it long?," "Is it fuzzy?," "Is it bigger than a microwave oven?"), parts ("Does it have ears?"), actions ("Can it cause you pain?"), uses ("Can you play with it?"), common locations ("Does it live above ground?"), and emotions ("Does it make you happy?"). Each

of the 218,000 object-question pairs was evaluated on Amazon's Mechanical Turk using a five point scale from "definitely not" (coded as -1) to "definitely yes" (1). There were many different participants and each question was only answered once, necessitating the use of a noisy response model described in a subsequent section.

**Bayesian framework.** A Bayesian framework for 20Qs has been developed in prior work (Nelson et al., 2014; Ruggeri et al., 2015), and we extend this framework to model our large scale 20Qs task.

We use the following notation. The data set $\mathcal{D}$ contains 1000 objects $o \in O$ and 218 questions $f \in \mathcal{F}$ (which can also be viewed as "features"). The response for a particular object-question pair is a value on the five point scale $\mathcal{D}_{of} \in \mathcal{A} = \{-1, -0.5, 0, 0.5, 1\}$. During a specific game of 20Qs, the same questions and response scale as in $\mathcal{D}$ are used. The first question and its response are denoted $\{f_1, a_1\}$, the second question and response $\{f_2, a_2\}$, and so on where the information revealed so far is denoted as $\mathcal{K} = \{\{f_1, a_1\}, \{f_2, a_2\}, \ldots, \{f_n, a_n\}\}$ with $a_j \in \mathcal{A}$ and $f_j \in \mathcal{F}$. We also use the notation $\mathcal{K}_f$ and $\mathcal{K}_a$ to separately indicate the set of questions $\mathcal{K}_f = \{f_1, f_2, ..., f_n\}$ and their corresponding answers $\mathcal{K}_a = \{a_1, a_2, ..., a_n\}$.

During a game of 20Qs, Bayes' rule can be used to reason about the probability of each object given the questions and their responses so far,

$$P(o|\mathcal{K}_a; \mathcal{K}_f) = \frac{P(o)\prod_{j=1}^{n} P(a_j|o; f_j)}{\sum_{o' \in O} P(o')\prod_{j=1}^{n} P(a_j|o'; f_j)}. \quad (1)$$

For simplicity and for consistency with the behavioral study, a flat prior is used over the objects $P(o) = 1/1000$. Ideally, the likelihood $P(a_j|o; f_j)$ would be known and modeled separately for each feature-object pair to capture variability across different people in how they answer the same question. However this would require having many responses for each of the 218,000 object-question pairs, for which only one response is provided and collecting many more is unfeasible. Instead, we ran a separate Mechanical Turk experiment to estimate a shared noise model to use as proxy, described in a later section. The semicolon notation indicates that $\mathcal{K}_f$ is a parameter rather than random variables like $\mathcal{K}_a$ and $o$.

**Expected Information Gain (EIG).** In conjunction with the Bayesian model, EIG was used as the metric for deciding what question to ask next. The goal is to ask the question $f_{n+1}$ that maximizes the expected reduction in uncertainty (measured as Shannon entropy $H[\cdot]$ in the posterior distribution). The expectation is a weighted average over all possible answers $a_{n+1}$, such that

$$EIG(f_{n+1}) = \sum_{a_{n+1} \in \mathcal{A}} P(a_{n+1}|\mathcal{K}_a; f_{n+1}) \Big[ H[P(o|\mathcal{K}_a)] - H[P(o|a_{n+1}, \mathcal{K}_a; f_{n+1})] \Big]. \quad (2)$$

Note that we dropped an implicit dependence on $\mathcal{K}_f$ in each distribution. The posterior predictive distribution is,

$$P(a_{n+1}|\mathcal{K}_a; f_{n+1}, \mathcal{K}_f) = \sum_{o \in O} P(a_{n+1}|o; f_{n+1}) P(o|\mathcal{K}_a; \mathcal{K}_f).$$

**Likelihood model and response noise.** The likelihood $P(a|o; f)$ from Eq. 1 (dropping the subscript $j$ for convenience) requires modeling variability in how different people answer ($a$) the same question ($f$), in relation to a particular object-question pairing. This is needed because the gamemaster (as captured by the data set $\mathcal{D}$) may not entirely agree with the question-asker (a human participant or the Bayesian model) on how to answer a question. For instance, people may answer ambiguous questions differently, such as the pairing of "dog" and "Is it bigger than a loaf of bread?"

A Mechanical Turk experiment was used to fit a response model by querying a subset of object-question pairs multiple times with different participants. The results were used to fit the model of response noise, estimated using a separate Bayesian analysis. Assume that the response model for any cell $\mathcal{D}_{of} \in \mathcal{A}$ is a unknown multinomial distribution $h_{of}$ over the five point scale. Let $h_{of} \in \mathcal{H}$ denote the possible set of multinomial distributions, where $\mathcal{H}$ is approximated by the empirical set of multinomials collected from Mechanical Turk. We assume that multinomials for new cells are drawn from the set of existing cells. The likelihood model for a question response then becomes

$$\begin{aligned} P(a|o; f) &= P(a|\mathcal{D}_{of}) \\ &= \sum_{h' \in \mathcal{H}} P(a|h_{of} = h') P(h_{of} = h'|\mathcal{D}_{of}) \\ &= \sum_{h' \in \mathcal{H}} P(a|h_{of} = h') \frac{P(\mathcal{D}_{of}|h_{of} = h') P(h_{of} = h')}{P(\mathcal{D}_{of})} \end{aligned}$$

which marginalizes over the uncertainty regarding the latent multinomial, given just a single sample from that multinomial $\mathcal{D}_{of}$. Both terms $P(a|h_{of} = h')$ and $P(\mathcal{D}_{of}|h_{of} = h')$ are just the probability of that response given the multinomial distribution represented by $h'$. A uniform prior $P(h_{of} = h')$ is used across the possible multinomials $h' \in \mathcal{H}$.

The set of multinomials $\mathcal{H}$ collected from Mechanical Turk consisted of 500 unique feature-question pairs. Fifty participants in the USA were asked to answer 100 questions randomly chosen from this set of 500 pairs, resulting in approximately 10 observations per cell. The corresponding set of 500 empirical multinomials was used to create $\mathcal{H}$. The fitted likelihood model is shown in Table 1.

Table 1: Fitted response model $P(a|\mathcal{D}_{of})$.

| | | $\mathcal{D}_{of}$ | | | | |
|---|---|---|---|---|---|---|
| | | -1.0 | -0.5 | 0.0 | 0.5 | 1.0 |
| $a$ | -1.0 | 0.704 | 0.311 | 0.185 | 0.126 | 0.062 |
| | -0.5 | 0.098 | 0.272 | 0.129 | 0.082 | 0.039 |
| | 0.0 | 0.097 | 0.214 | 0.380 | 0.201 | 0.099 |
| | 0.5 | 0.056 | 0.117 | 0.172 | 0.313 | 0.176 |
| | 1.0 | 0.044 | 0.086 | 0.133 | 0.278 | 0.623 |

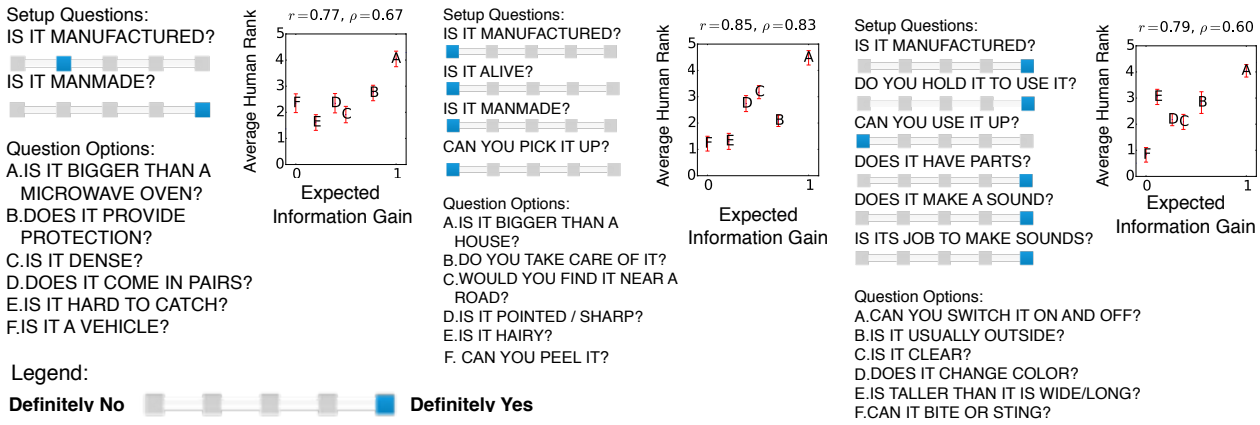


Figure 1: Three examples of one-shot 20 question games. Participants observed the Setup Questions with answers on a five point scale (see legend). They then ranked the Question Options in order of preference. In the scatter plots, the average human rank (error bars show $\pm 1$ s.e.) is compared with Bayesian expected information gain, which was normalized to fall between 0 and 1. The object chosen by the game-master for each game were Kitchen, Ground, and Tuba (from left to right).

## Experiment

A behavioral experiment was run on Amazon's Mechanical Turk using Psiturk (Gureckis et al., 2015) to compare human question asking with the Bayesian model and alternative models. Data was collected from 25 participants that reside in the USA, providing a base pay of $6 with an additional performance-based bonus. Two participants were not analyzed due to technical difficulties.

Participants acted as the question-asker while the computer plays the role of the game-master. Participants played two forms of 20Qs, including first a set of "complete games" and second a set of "one-shot games," each with a massive hypothesis space (any object from the set of 1000) and a fixed set of question choices. Complete games involved full games of 20Qs where each participant played unique games. One-shot games involved introducing participants to partially-finished 20Qs games (with some questions already answered) and then asking participants for their preferences regarding the next question (see examples in Fig. 1). Unlike the complete games, the one-shot games and associated question options were identical for each participant.

A number of checks were put in place to ensure that participants were engaged in the experiment. Before completing any of the full games, participants were asked to play a tutorial full game of 20Qs with artificial visual objects with four binary features. Participants were also given a short quiz after each segment of instructions, and correct answers were needed to proceed to the next section. This included instructions meant to convey the semantic space of possible objects (similar to the description in the section "Data set of 1000 objects"), although the precise set of 1000 objects remained unknown to participants. Participants were also told they should expect noise – that is, the computer responses are based on other people's answers (as usual in 20Qs) and thus they should expect occasional disagreements.

**Complete games.** Participants played five full games of 20Qs: one practice game and four real games. In each game, an object was randomly chosen by the computer from the set of 1000. At each step, participants were shown six questions they could ask (in random order), selected to evenly span a range of EIG values according to the Bayesian model.[1] Participants chose a question and then received the answer, after which they received a new set of questions to choose from, etc. Before receiving the answer choices at each step, they were also asked to type out their ideal question (which wasn't analyzed). In order to mitigate the possibility of multiple exposures to the same question, questions were presented as an option only once per full game. Also, the design discouraged multi-step planning strategies, as future question options were always unknown.

At the completion of the game, participants were asked to "guess" the identity of the hidden object. To make the task feasible, rather than selecting the correct object from the entire set of 1000 possibilities, people were shown a set of 20 objects that included the correct answer and 19 randomly chosen distractors. The 20 options were not shown until the game was completed, meaning no more questions could be asked (although the previous questions and answers could still be viewed). Participants were rewarded with a bonus for choosing the correct object from the set.

The potential bonus started at $0.50 and was decremented by $0.05 for each question they asked. Participants could choose to enter the guessing phase at any time instead of selecting another question. After nine questions the guessing phase began automatically. After guessing, they were shown the correct object.

[1] Possible questions were ranked by the model from best (1) to worst (218). Six questions were chosen based on their position in the ranked list (1, 37, 73, 109, 145, 181). If a question was presented previously as a possible choice, it was excluded from the ranking.

**One-shot games.** After the full games, participants completed ten one-shot games of 20Qs. Each of these games was pre-generated and the same games were shown to each participant in random order. These one-shot games followed the same structure as the full games, except that participants were incrementally shown a number of questions and answers, as if someone else was playing the game for them. To make sure they were following along, they were asked to reconstruct the answers to two questions as a quiz before continuing to the next stage.

After the quiz, participants were asked to choose between six questions (see Fig. 1). Rather than selecting only the best question as in the full games, they were asked to rank the questions in order of preference. After ranking, the game was over, the top question was answered, and an analogous guessing phase occurred. A bonus of \$0.20 was rewarded for correctly guessing the object.

The one-shot games varied in depth: the number of previous questions shown before participants ask a new question. There was 1 trial of depth 0 (participants chose the first question), 3 trials of depth 2, 3 trials of depth 4, and 3 trials of depth 6 (see Fig. 1 for examples). Deeper games were not included, since we expected people would have trouble processing that much information on each trial. The one-shot games were generated by recursively querying the model. Each question option appeared only once across the set of one-shot games.

## Alternative Models

In addition to the Bayesian model that maximizes EIG (henceforth "full Bayesian model" or "EIG"), we evaluated a range of alternatives.

**Expected Utility (EU).** The EU model uses the full Bayesian machinery to choose the question that maximizes expected "reward gain" rather than "information gain," metrics that often make similar predictions (Markant & Gureckis, 2012). EU maximizes the expected bonus in the guessing phase, assuming it occurs immediately after the next question is answered. EU is defined like EIG except that the entropy function $H[\cdot]$ (Eq. 2) is replaced by a function that computes the expected bonus (and multiplied by $-1$). In the bonus phase, we assume the EU model selects the object (of 20) with highest posterior probability. The probability that its choice is correct (according to model belief), marginalizing over all possible correct objects and all possible sets of random distractors, can be computed exactly with combinatorics.

**Context Insensitive (CI).** The CI model is a simplified Bayesian model that ignores the current game state (previous questions and answers). It chooses questions to maximize EIG while assuming the current knowledge state is the prior (uniform). The CI model tests whether people ask context-sensitive questions or just choose questions based on a set of pre-computed "good" questions.

**Random Subset (RS).** The RS model is a simplified Bayesian model that does not consider the whole hypothesis space. Instead, RS considers only $K$ naively chosen hypotheses, related to the 20Qs model of Navarro and Perfors (2011). Rather than full Bayesian updating (Eq. 1), all but $K$ randomly chosen objects are assigned zero prior probability (and thus zero posterior probability). As with the full Bayesian model, questioned are selected to maximize EIG but with this simplified posterior. A simulated experiment with RS used 25 simulated participants, each with a different random subset of the hypotheses (also re-sampled for each trial). We ran a total of 100 simulated experiments, and the reported statistics (e.g., correlation coefficients) were computed and then averaged over simulated experiments. Different values of $K$ were explored where increasing $K$ converges to the full model.

**Familiarity.** Although steps were taken to mitigate question repetition in the behavioral experiment, including using unique question options within each full game and not repeating options across one-shot games, a potential design issue is that some questions were seen by participants more than once (either as choices or previously revealed information). For instance, the high-value early question "Is it manufactured?" appeared many times as revealed information (Fig. 1). The Familiarity model attempts to explain the value of a question by the number of times people previously saw it.

## Results and Discussion

**Analysis of one-shot games.** We analyzed the ability of each model to predict human question preferences on the one-shot games. For the human data, a question quality score was computed as the average rank position across the 25 participants, where 0 is the worst question and 5 is the best question. The correlation between these human quality scores and the model scores were computed for each game individually (both Pearson ($r$) and Spearman rank ($\rho$) are reported).

Overall, full Bayesian EIG provides the best account of people's preferences. The model predictions for each one-shot game are shown in Fig. 2, and several individual games can be examined in detail in Fig. 1. The average correlation between EIG and human quality scores was $r = 0.777$ and $\rho = 0.718$. At first glance, it appears that games at depth 6 show a weaker correlation on average, but this may be due to one game with an unconventional selected object (a "brake").

The full results and the fits for the alternative models are shown in Table 2. EU has the second best fit with an average $r = 0.717$ and $\rho = 0.643$. While similar to EIG, its performance can degrade with depth when it becomes confident it can get the bonus, leading to weak preferences for subsequent questions. The CI model dramatically degraded with depth (as expected), achieving an average correlation of $r = 0.367$ and $\rho = 0.366$. This suggest that people are not just memorizing a set of "good questions" and asking those regardless of context. Finally, the RS model showed a strong average correlation with the human ratings ($r = 0.712$ and $\rho = 0.650$) when considering $K = 20$ objects but fared poorly when con-

Table 2: Pearson ($r$) and Spearman ($\rho$) correlations between human and model preferences in the one-shot games.

| Game Number | Depth | Object | Full Bayesian $r$ | Full Bayesian $\rho$ | Expected Utility $r$ | Expected Utility $\rho$ | Context Insensitive $r$ | Context Insensitive $\rho$ | Random Subset k=20 $r$ | Random Subset k=20 $\rho$ | Random Subset k=5 $r$ | Random Subset k=5 $\rho$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | Blanket | 0.874 | 0.829 | 0.723 | 0.771 | 0.874 | 0.829 | 0.864 | 0.827 | 0.847 | 0.805 |
| 2 | 2 | Kitchen | 0.767 | 0.667 | 0.825 | 0.667 | 0.597 | 0.522 | 0.737 | 0.669 | 0.690 | 0.640 |
| 3 | 2 | Step | 0.825 | 0.829 | 0.825 | 0.826 | 0.815 | 0.829 | 0.835 | 0.822 | 0.816 | 0.792 |
| 4 | 2 | Almond | 0.839 | 0.812 | 0.820 | 0.812 | 0.876 | 0.812 | 0.840 | 0.806 | 0.843 | 0.802 |
| 5 | 4 | Scarecrow | 0.834 | 0.812 | 0.907 | 0.812 | 0.679 | 0.734 | 0.765 | 0.809 | 0.709 | 0.753 |
| 6 | 4 | Ground | 0.850 | 0.829 | 0.796 | 0.829 | 0.238 | 0.486 | 0.630 | 0.626 | 0.118 | 0.162 |
| 7 | 4 | Cricket | 0.928 | 0.943 | 0.927 | 0.943 | -0.480 | -0.314 | 0.854 | 0.816 | -0.315 | -0.198 |
| 8 | 6 | Brake | 0.320 | 0.257 | 0.526 | 0.257 | 0.216 | 0.257 | 0.356 | 0.326 | 0.267 | 0.213 |
| 9 | 6 | Tuba | 0.795 | 0.600 | 0.064 | -0.086 | -0.044 | -0.371 | 0.558 | 0.302 | -0.061 | -0.140 |
| 10 | 6 | Mop | 0.739 | 0.600 | 0.763 | 0.600 | -0.104 | -0.143 | 0.687 | 0.507 | 0.457 | 0.300 |
| Average | 2 | | 0.810 | 0.769 | 0.823 | 0.768 | 0.763 | 0.721 | 0.804 | 0.766 | 0.783 | 0.745 |
| Average | 4 | | 0.871 | 0.906 | 0.877 | 0.861 | 0.146 | 0.308 | 0.750 | 0.750 | 0.171 | 0.239 |
| Average | 6 | | 0.618 | 0.486 | 0.451 | 0.257 | 0.023 | -0.086 | 0.534 | 0.378 | 0.132 | 0.124 |
| Average | All | | 0.777 | 0.718 | 0.717 | 0.643 | 0.367 | 0.366 | 0.712 | 0.650 | 0.437 | 0.412 |

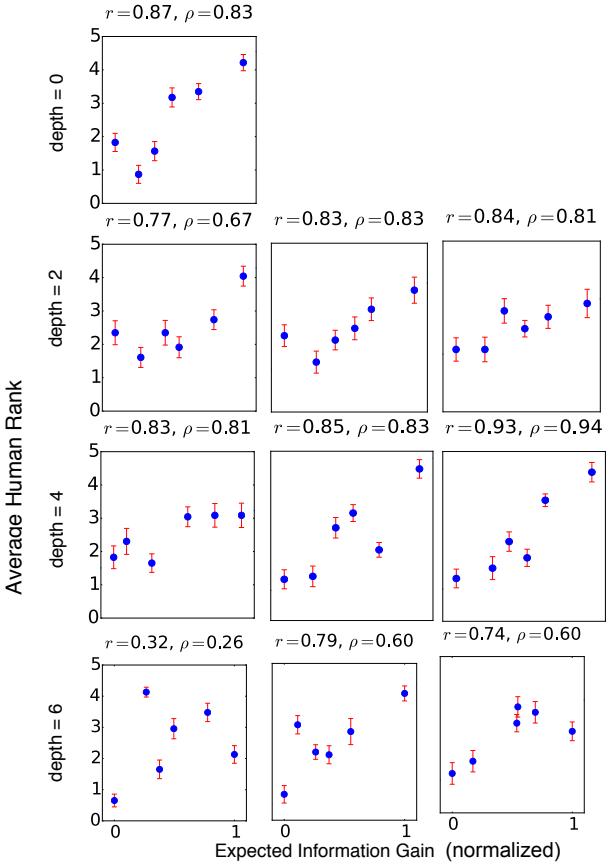


Figure 2: Bayesian model fits for one-shot games, with the different games organized by depth. The average human rank score (error bars show $\pm 1$ s.e.) is compared with expected information gain (normalized between 0 and 1).

sidering $K = 5$ objects ($r = 0.437$ and $\rho = 0.412$), suggesting that while people may approximate the full Bayesian inference by considering a random subset of hypotheses, the subsets may need to be of considerable size. RS performance across a wider range of subset sizes is shown in Fig. 3.

The Familiarity model had to be analyzed differently. Since different participants saw each question a different number of times, the model correlation was computed separately for each participant's ranking. The correlation values averaged across participants as well as one-shot games were low (average $r = 0.0799$, $\rho = 0.0806$), but it varied substantially by depth since early questions are more likely to be repeated (depth 0, $r = 0.454$; depth 2, $r = 0.168$; depth 4, $r = -0.0479$; and depth 6, $r = -0.005$). While this statistic lacks the power of the previous calculations, the EIG model correlations computed in the same way yielded a higher average fit ($r = 0.384$, $\rho = 0.383$), suggesting that question familiarity was not the driving factor in question choice.

**Analysis of complete games.** We also analyzed the ability of the models to predict participant choices in the complete 20Qs games. The same analyses used in the one-shot games could not be repeated for these complete games for several reasons. First, participants chose to ask just one question at each step rather than rank a list of options. Second, each participant played a unique set of games, unlike the pre-generated one-shot games for which responses could be aggregated. In an accuracy analysis, the models selected the question with maximum score, and these choices were compared to the participant choices. The EIG model performed best at 28.4% correct where chance is 16.7%, followed by EU (27.0%) and CI (23.6%). The RS models had accuracy levels of 21.5% ($K = 5$) and 25.8% ($K = 20$). All of the models were above baseline, but no model succeeded in predicting a majority of choices. These results point to the challenges of predicting individual responses in the complete 20Qs games.

## Conclusions

People ask questions when faced with uncertainty, seemly undaunted by very large hypothesis spaces. Previous work has used Bayesian modeling and the Expected Information Gain (EIG) to explain how people assign value in 20 Questions (20Qs), but these studies have considered restricted hypotheses spaces that can be presented all at once to participants, such as the children's game "Guess who" (Nelson et

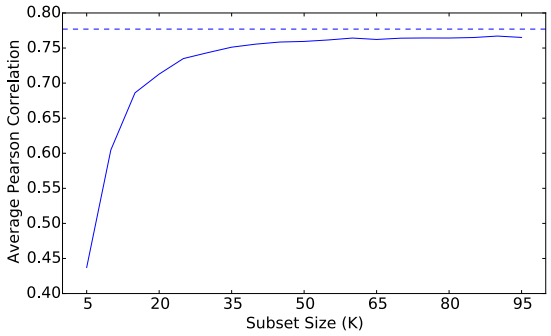


Figure 3: Average correlation ($r$) between human preferences on the one-shot games and the Random Subset model for different subset sizes $K$ (solid line; same as Table 2 last row). For comparison, the full Bayesian model is shown as the dotted line.

al., 2014). It is unclear how people assign value to questions in very large hypothesis spaces that they are unlikely to represent explicitly and in their entirety.

Here we studied 20Qs in a very large hypothesis space – unbounded, from the perspective of participants – by relying on people's pre-existing semantic knowledge. While this creates difficulties for formal modeling, we presented a Bayesian model that incorporated a large data set of objects and question responses. In a series of one-shot and full games, people's ranking and selection of questions were best predicted by the full Bayesian model and expected information gain. On the one-shot games, this model achieved relatively strong average correlations with the aggregate human preferences ($r = 0.78$ and $\rho = 0.71$), providing a better explanation for the data than a range of alternatives, including a simple form of approximate Bayesian inference that considers small random subsets of the hypothesis space (Navarro & Perfors, 2011). Our results are consistent with the view of people as "good question-askers," although predicting individual responses in the full 20Qs games remains a challenge.

Many questions remain for future work. How does the Bayesian model perform on unusual games when a rare object is selected (e.g., brake or scarecrow) instead of a more common object (e.g., blanket or apple)? Our framework could also be used to explore how changes to the prior influence choice, comparing scenarios where the secret object was selected by different processes (e.g., by a child, an adult, or a computer). An additional challenge is to explain how people generate questions in more natural, free-form tasks rather than choosing from a pre-selected set (Rothe et al., 2016).

While our computational-level analysis will help constrain future algorithmic accounts of this ability (Marr, 1982), it does not itself offer a concrete algorithm for how people evaluate question quality in large hypothesis spaces. While our analyses suggest people consider more than just a few hypotheses, it seems implausible they are considering (even implicitly) hundreds or thousands of hypotheses when less demanding and still accurate inference strategies may exist.

An intriguing possibility is that people may aim to discover the right higher-level category rather than the specific item, at least during early stages of a 20Qs game. People could use pre-existing semantic categories such as animals, plants, living things, artifacts, vehicles, tools, etc. as these intermediate goals. Compared to the full Bayesian account, this alternative algorithm predicts that people devalue questions that cross-cut the high-level categories they are considering. For instance, while "Does it move?" is often a good question according to the full Bayesian model, it applies to subsets of both living things and artifacts, cutting across the salient hierarchical structure in the hypothesis space and thus potentially devaluing it relative to other questions. The predictions of this account can be tested empirically.

Future work will explore this and other algorithms for approximate Bayesian inference, as well as individual difference data, to further our understanding of how people search large hypothesis spaces by asking questions.

**Acknowledgments.** We thank M. Palatucci and Intel for their data set, and Todd Gureckis for helpful comments on an earlier draft. This research was supported by the Moore-Sloan Data Science Environment at NYU.

## References

Denney, D. R., & Denney, N. W. (1973). The use of classification for problem solving: A comparison of middle and old age. *Developmental Psychology*, *9*(2), 275–278.

Eimas, P. D. (1970). Information processing in problem solving as a function of developmental level and stimulus saliency. *Developmental Psychology*, *2*(2), 224–229.

Gureckis, T. M., & Markant, D. B. (2009). Active Learning Strategies in a Spatial Concept Learning Game. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Gureckis, T. M., & Markant, D. B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Gureckis, T. M., Martin, J., McDonnell, J., Alexander, R. S., Markant, D. B., Coenen, A., ... Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*.

Markant, D., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Marr, D. C. (1982). *Vision*. San Francisco, CA: W.H. Freeman and Company.

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., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Palatucci, M., Pomerleau, D., Hinton, G., & Mitchell, T. (2009). Zero-Shot Learning with Semantic Output Codes. In Y. Bengio, D. Schuurmans, & J. Lafferty (Eds.), *Advances in Neural Information Processing Systems (NIPS).*

Rothe, A., Lake, B. M., & Gureckis, T. M. (2016). Asking and evaluating natural language questions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Thornton, S. (1982). Challenging "Early Competence": A Process Oriented Analysis of Children's Classifying. *Cognitive Science*, *6*, 77–100.