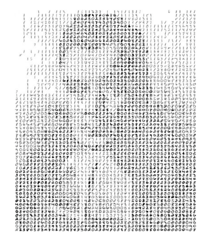
# Testing a Bayesian Measure of Representativeness Using a Large Image Database


Joshua T. Abbott joshua.abbott@berkeley.edu
Katherine A. Heller kheller@mit.edu
Zoubin Ghahramani zoubin@eng.cam.ac.uk
Thomas L. Griffiths tom_griffiths@berkeley.edu

## Overview

**Question:** How do people determine which elements of a set are most representative ?

**What we know:** Elements are representative when they are "good examples" of a concept. Quantified via Bayesian inference, a rational model takes the problem to be finding the best example that provides the most evidence for a target concept relative to possible alternatives.

**What we don't know:** The rational account of representativeness requires concepts to be pre-defined and has only been evaluated on simple, artificial stimuli in a laboratory setting. What if we don't know which distribution defines a concept? Can we test this psychological theory on more realistic stimuli?

**Our contribution:** We extend an existing Bayesian measure of representativeness to address the problem of deciding which objects in a set are the most representative. We show that the resulting model is related to an existing machine learning method known as Bayesian Sets, and apply this new measure to a large database of naturalistic images to determine which images are the most representative members of different sets.

## Representativeness

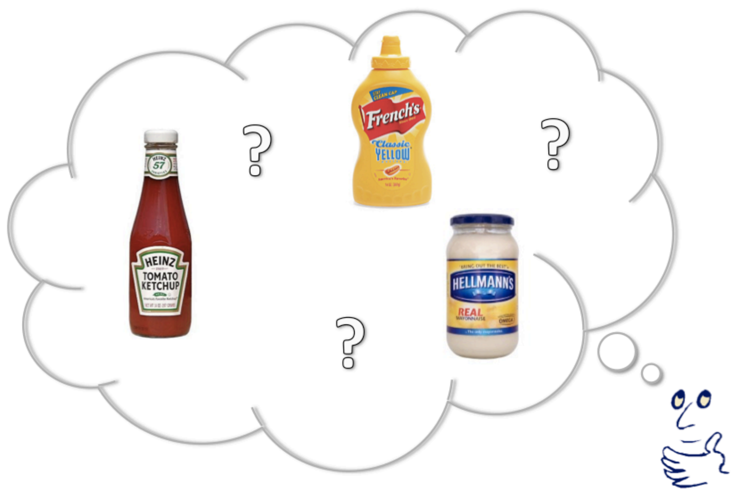
Why do people believe that the coin-flip sequence of **HHTHT** is more likely than **HHHHH** to be produced by a fair coin?

**HHTHT** is more *representative* of the output produced by a fair coin.

Given some observed data $d$ and hypothetical sources, $h_i$, we assume that a learner uses Bayesian inference to infer which hypothesis $h_i$ generated $d$. Tenenbaum and Griffiths (2001) defined the representativeness of $d$ for $h_i$ to be the evidence that $d$ provides in favor of a specific $h_i$ relative to its alternatives:

$$R(d,h_i) = \log \frac{P(d|h_i)}{\sum_{j\neq i} P(d|h_j)P(h_j)}$$

## Bayesian Sets

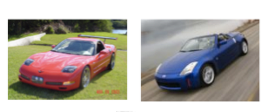


**What other items belong in this set?**

The problem addressed by Bayesian Sets is: you are given a few example items that represent some concept, or set, such as "condiment", and you want to find other items that belong to that same concept, here other "condiments."

Given a data collection $D$, and a subset of items $D_s = \{ \mathbf{d}_1, \ldots, \mathbf{d}_N \} \subset D$ representing a concept, the Bayesian Sets algorithm ranks each item $\mathbf{d}^*$ in $D$ by a model-based scoring criterion which compares the hypotheses that $\mathbf{d}^*$ does (numerator), or does not (denominator) belong to the same set as $D_s$:


data collection $D$



query $D_s$

$$\text{Bscore}(\mathbf{d}^*) = \frac{p(\mathbf{d}^*, \mathcal{D}_s)}{p(\mathbf{d}^*)p(\mathcal{D}_s)}$$

Ranking

Best → Worst


## Representativeness and Bayesian Sets

Given a set of items $D_s$ representing some concept, how representative is an element $\mathbf{d}^* \subset D_s$ of the whole set?

$$R(\mathbf{d}^*, \mathcal{D}_s) = \log \frac{P(\mathbf{d}^*|\mathcal{D}_s)}{\sum_{s\neq t} P(\mathbf{d}^*|\mathcal{D}_t)P(\mathcal{D}_t)}$$

If the set of possible datasets being summed over in the denominator is large, the denominator will approximately equal P($\mathbf{d}^*$). Thus this measure of representativeness applied to sets of objects will closely approximate the logarithm of the Bayesian Sets score for dataset $D_s$

$$R(\mathbf{d}^*, \mathcal{D}_s) \approx \log \frac{P(\mathbf{d}^*|\mathcal{D}_s)}{P(\mathbf{d}^*)} = \log \frac{P(\mathbf{d}^*, \mathcal{D}_s)}{P(\mathbf{d}^*)P(\mathcal{D}_s)} = \log \text{Bscore}(\mathbf{d}^*)$$

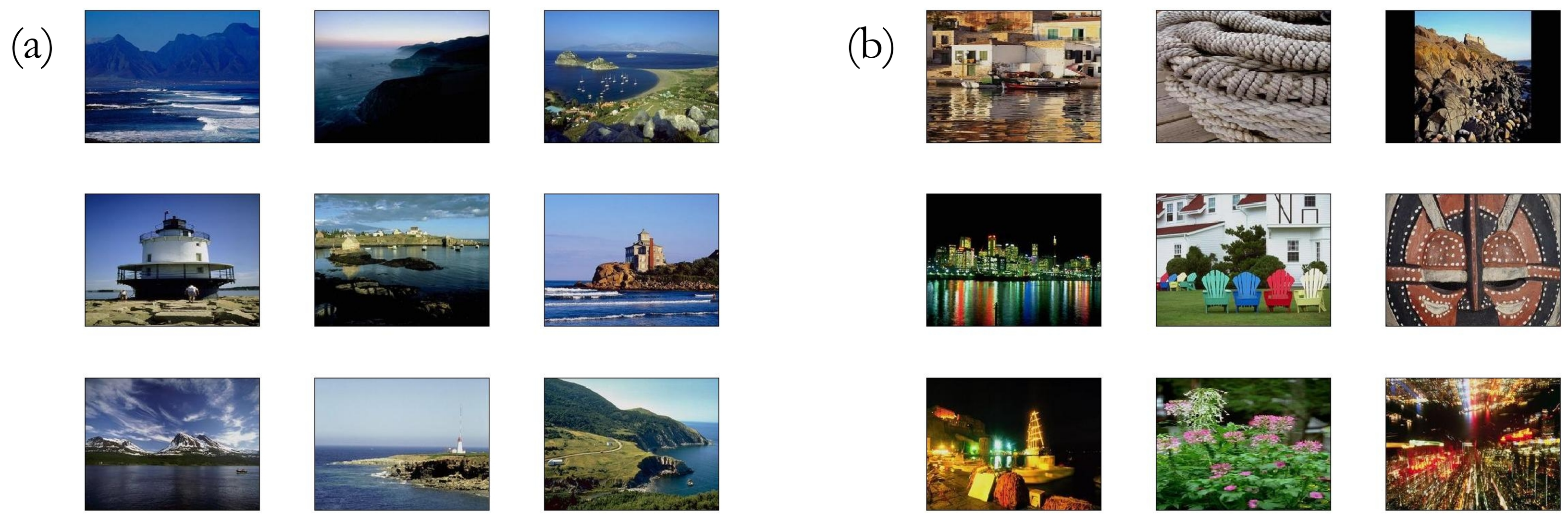
This relationship provides a link between the cognitive science literature on representativeness and the machine learning literature on information retrieval.

## Representativeness of Images

We evaluate our measure on a subset of the Corel database, partitioned into 50 labeled sets depicting unique categories, and compare our model against a likelihood model and two similarity models: a prototype model and an exemplar model.

We use a leave-one-out framework to perform fair comparisons across these different models. Given a set of images with a particular category label, we iterate through each image in the set and compute a score for how well this image represents the rest of the set.

**How representative is an image of a labeled set of images ?**




Results of the Bayesian model applied to the set of 299 images labeled **coast**. Panels (a) and (b) show the top nine and bottom nine ranked images, respectively.

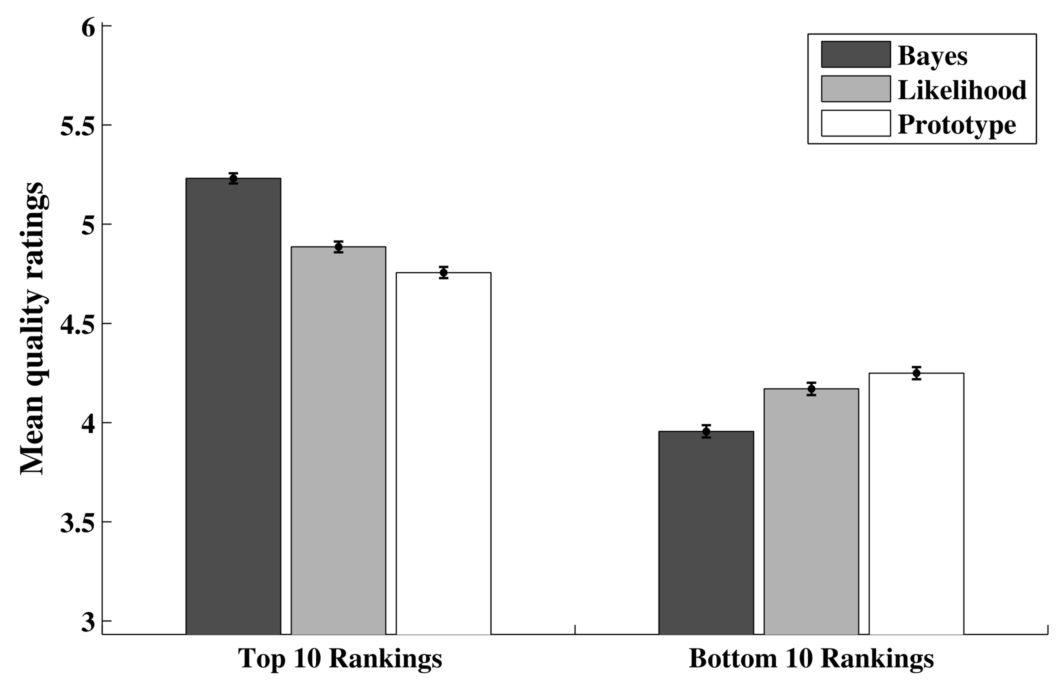
## Modeling Human Ratings

Given a set of images provided with a category label, how do people determine which images are good or bad examples of that category?

**Method:** A total of 500 participants (10 per 50 categories) were recruited via Amazon Mechanical Turk. Participants were shown a series of images and asked to rate how good an example each image was of an assigned category on a scale of 1 (bad) to 7 (good).

**Stimuli:** A union of the top 10 and bottom 10 ranked images for each of the 50 categories for the Bayesian, likelihood, and prototype model.

**Results:** Once the human ratings were collected, we computed the mean rating for each image and the mean of the top 10 and bottom 10 results for each algorithm used to create the stimuli. We additionally computed bounds for the ratings based on the optimal set of top 10 and bottom 10 results per category. The vertical axis is bounded by these optimal ratings.

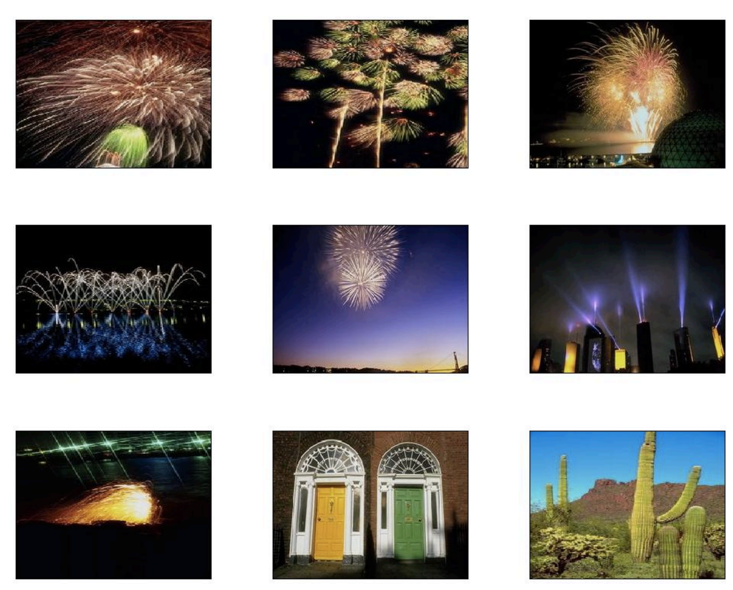



## Outlier Detection

Representativeness of items in sets provides a novel method of finding outliers in sets.

To test this idea, we injected outlier images into the 50 category sets and measured how low these images were ranked by each model.

| Model | Avg. Outlier Position |
|---|---|
| Bayesian | 0.805 |
| Likelihood | 0.779 |
| Prototype | 0.734 |
| Exemplar | 0.734 |



The bottom nine ranked images for the set labeled **fireworks** with two outliers.

## Conclusions

- Extended an existing Bayesian model of representativeness to handle sets of items and exploited the relationship with Bayesian Sets to evaluate on a large database of naturalistic images. The results provide strong evidence for this characterization of representativeness.
- Closer integration of methods from cognitive science and machine learning: (1) first large-scale experimental comparison of the Bayesian Sets algorithm to human judgments and (2) first evaluation of Bayesian measure of representativeness in context of a real applied problem.