Automating the recoding, analysis, and interpretation pipeline using naturalistic visual scenes

April M. Schweinhart, Baxter S. Eaves Jr., and **Patrick Shafto**
Rutgers University - Newark
{april.schweinhart, baxter.eaves, patrick.shafto} @rutgers.edu

Abstract

Machine learning often focuses on how best to infer structure from data. Also important is the ability to convey that structure to human users. We investigate a system for automating quantification, analysis, and presentation of data to human users. We focus on the domain of natural scenes, an area in which human performance has been well explored, and can thus be used to inform choices of computational tools. Informed by perceptual science, we characterize a corpus of images in terms of the statistics of their orientation distributions. In two experiments, we compare mixture and topic models for analysis, and teaching-optimized versus average images for conveying model structure to people. Using a categorization task, in Experiment 1, we find that, when subclusters are overlapping and categorization difficult, examples selected to teach the category structure lead to improved categorization performance relative to examples closest to the mean. Experiment 2 further shows that mixture models outperformed topic models and teaching examples outperformed maximum likelihood. By leveraging cognitively natural machine learning methods to facilitate automatic analysis and summary of naturalistic data, this work has implications for conveying both typicality and variability of experiences in complex data.

1 Introduction

For decades the focus of machine learning has been to take a prepared data set and, given some parameters, infer structure from it. While it is important to find such patterns in data, it is also important to ensure that structure can be easily accessed by humans attempting to make sense of the data. Indeed, problems anywhere in the pipeline—from quantifying experiences, to inferring structure, to interpreting and acting on results—may to lead to incorrect outcomes. Thus, a critical problem for machine learning, data science, and artificial intelligence more generally is how to make choices about each step so that the people at the end of the data analysis pipeline understand the output and make correct decisions.

In this paper, we investigate automating the quantification, analysis, and interpretation pipeline. Solving this problem is a long term goal. Whereas typical machine learning and data science approaches rely on highly educated experts to implement and interpret analyses, we present a specific example of a general approach based on leveraging humans' uniquely powerful ability to learn from small amounts of data generated by teachers. A naive computational learner infers some structure in the data and a computational teacher selects a small subset of the original data that best convey that structure to humans learners. We compare teaching decisions to well-known alternative methods not motivated by human reasoning in the domain. Success on this project would greatly increase the accessibility of data-driven decision making by reducing the need for specific training.

We focus on a domain (natural scene perception) and task (categorization) that have been well studied in the human learning literature. This allows us to select methods of quantifying and analyzing data that are strongly informed by existing science. Specifically, we leverage known human competencies in perception, cognition and social learning. In two experiments, we investigate different machine learning methods—mixture models and topic models—and methods of summarizing their results—selecting examples through computational models of teaching or that capture the mean or maximum likelihood estimate. Teaching has computational support in the literatures on perception, cognition, and social learning, while choosing data close to the mean or that maximize likelihood do not.

The paper unfolds in three sections. First, we discuss foundational work in the areas of natural scene perception, categorization, and computational models of teaching. Second, we describe the pipeline for quantifying images, extracting categories from these data, and selecting images to teach the resulting categories. Third, we describe two experiments that investigate the performance of the approach with untrained learners. Experiment 1 focuses on the last step of the pipeline, the selection of images, via teaching or the maximum likelihood data, to communicate the results to the user. Experiment 2 additionally manipulates model used to infer structure from the data, comparing mixture models with topic models. The results show that when the problem is difficult, computational models of teaching outperform methods based on the mean or maximum likelihood. Results also show that a more cogni-

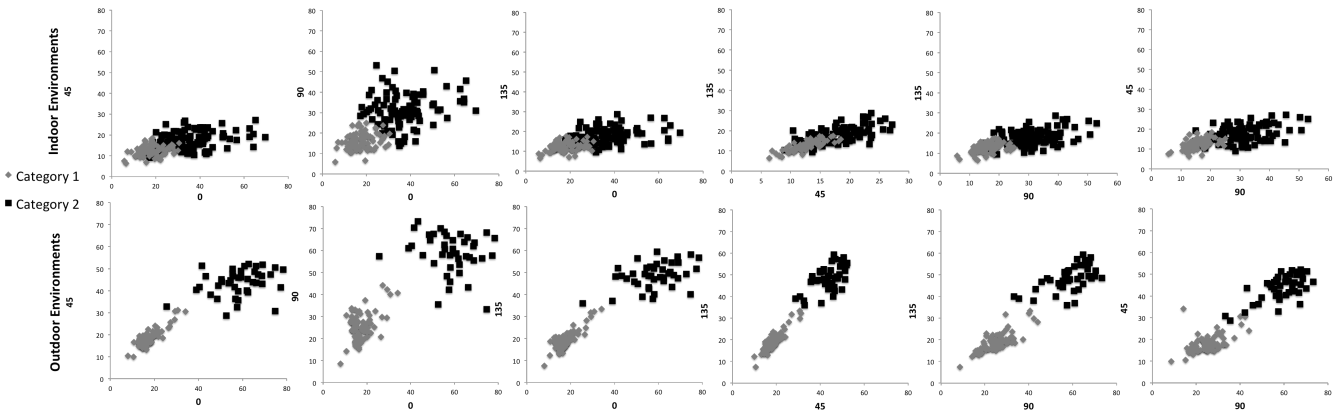

Figure 1: Scene category results. Orientation-orientation scatter plots of random samples from the target model. Different marker colors denote difference inner categories. The top row represents Indoor scenes and the bottom row represents Outdoor scenes. The indoor scene categories have considerably greater overlap than the natural scene categories.

tively natural representation of the domain—modeling scenes as mixture distributions—outperform a less cognitively natural, but otherwise effective, model in this visual domain.

2 Background

The natural-scene-category-teaching pipeline relies on findings from three literatures: natural scene perception, categorization, and computational models of teaching.

2.1 Natural scene perception

Natural scenes are semantically, structurally, and perceptually complex, and this complexity is decomposed by the human visual system, starting with low level features such as orientation. There is a characteristically biased distribution of oriented contours in natural scenes [1–3] and this anisotropy is reflected in the visual cortex at one of the earliest levels of visual processing [e.g. 4, 5]. Perception takes advantage of the regular anisotropy present in natural scenes [6] making it a logical structural property by which to quantify images as several previous image categorization approaches have done [7]. Therefore, it is sensible to quantify the structure of natural scenes in the orientation domain and determine if such structure can be taught to human users.

2.2 Categorization

There is a long history of behavioral research on human categorization (see [8, 9] for reviews). Anderson [10] [also 11] derived a model for learning an unknown number of categories, which was essentially a Dirichlet-process mixture model [12, 13]. Rasmussen [14] later proposed an efficient Gibbs sampling algorithm for this model. The Dirichlet process mixture model framework has since been widely adopted as a model of human category learning and in unsupervised machine learning, and has been used to model scene categories in images [7, 15]. Extending this previous work into communicating categories to human users is a logical next step.

2.3 Computational models of teaching

Computational models of teaching formalize the purposeful selection of examples whose goal is to enable the learner to infer the correct hypothesis [16–18]. Shafto & Goodman [16] introduced a Bayesian model of pedagogical data selection and learning, and used a simple teaching game to demonstrate that human teachers choose data consistently with the model and that human learners make stronger inferences from pedagogically-sampled data than from randomly-sampled data (data generated according to the true distribution; [19, 20]). More recently, Eaves Jr. *et al.* [21] employed advances in Monte Carlo approximation to facilitate tractable Bayesian teaching. Although enjoying considerable evidence in lab-based tasks where the target knowledge is selected by the experimenter, no prior work has investigated the possibility that this approach may be used to facilitate human learning from machine-derived knowledge.

3 Data analysis pipeline

3.1 Quantifying images

The first stage in the pipeline involves quantifying the complex information in an image. To do this, we extract the orientation information using a previously developed image rotation method [see 2]. In this method, each frame is rotated to the orientation of interest and the amplitude of the cardinal orientations (horizontal and vertical) extracted and stored via fast Fourier transform filtering. Repeating this process at different orientations allows each image to be condensed into a series of 4 (Experiment 1) or 36 (Experiment 2) data points representing the amount of oriented structure in the environment at four primary orientations 0, 45, 90, and 135 degrees in global (Experiment 1) or local (Experiment 2) image regions. We processed 200 images for both outdoor and indoor environments. The resulting four-orientation data can be seen in Figure 1.

The second stage involves inferring structure (categories) from the orientation data so that it may be taught. We compare two methods for image categorization: the infinite Gaus-

sian mixture model (IGMM) and latent Dirichlet allocation (LDA).

Infinite Mixtures

In experiment 1, we represent scenes as categories in continuous, multidimensional amplitude space. We model these categories as multidimensional Gaussians with mean μ and covariance matrix Σ. Learners must learn how many categories there are, their means and covariance matrices, and must infer of which category each datum is a member. We capture this with the *infinite Gaussian mixture model* framework [IGMM 14].

Infinite mixtures allow for as few as one or as many as n mixture components (categories). The IGMM infers an assignment, z, of data to categories, which is assumed to follow a Dirichlet process–in this work, the Chinese restaurant process–prior with concentration parameter α, CRP(α) [22]. The likelihood of the data, x, is then

$$\ell(x \mid \theta) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i}). \tag{1}$$

where $\mathcal{N}(x; \mu, \Sigma)$ is the Gaussian (Normal) density of x given mean μ and covariance matrix Σ.

We place a conjugate, Normal inverse-Wishart prior on μ and Σ [23],

$$\Sigma \sim \text{Inverse-Wishart}_{\nu_0}(\Lambda_0^{-1}), \tag{2}$$

$$\mu|\Sigma \sim \mathcal{N}(\mu_0, \Sigma/\kappa_0). \tag{3}$$

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a bag-of-words model for inferring the topics in corpora [24, 25]. Topic models have been adopted for use with images [7, 15], by treating each image as a document composed of visual *words* from a number of visual *topics*, T, which we treat as categories. To generate D documents from a W-word vocabulary under LDA with parameters $\alpha, \beta \in (0, \infty)$,

for all topics, $t \in 1, \ldots, T$ **do**
  $\phi_t \sim \text{Dirichlet}_W(\beta)$
end for
for all documents, $d \in 1, \ldots, D$ **do**
  $\theta_d \sim \text{Dirichlet}_T(\alpha)$
  for $i \in \{1, \ldots, w_d\}$ **do**
    $z \sim \text{Discrete}(\theta_d)$
    $w_{d|i} \sim \text{Discrete}(\phi_z)$
  end for
end for

The likelihood of a corpus, C, given $\Phi = \{\phi_1, \ldots, \phi_T\}$ under LDA is

$$\ell(C|T, \Phi, \alpha, \beta) = \sum_z \left[\left(\prod_{d=1}^{D} \text{DirCat}(z_d|\alpha) \right) \prod_{i=1}^{n} \phi_{w_i}^{(z_i)} \right], \tag{4}$$

where DirCat denotes the Dirichlet-categorical distribution, the sum over z denotes the sum over all possible assignments of the n words in the corpus to topics, z_d indicates the assignment of words in document d, and $\phi_{w_i}^{(z_i)}$ indicates the probability of the i^{th} word under the topic to which it is assigned, z_i.

3.2 Bayesian teaching

Teaching implies choosing data, x, that lead a learner to a specific hypothesis, θ, which we shall refer to as the *target*. In a Bayesian setting, teaching means choosing data in proportion with their induced posterior density:

$$p_T(x \mid \theta) = \frac{p_L(\theta \mid x)}{\int p_L(\theta \mid x) dx} \propto \frac{\ell(x \mid \theta)}{m(x)}, \tag{5}$$

where $\ell(x \mid \theta)$ is the likelihood of x under θ, $m(x) = \int \ell(x \mid \theta)\pi(\theta)d\theta$ is the marginal likelihood of x, and the subscripts T and L denote probabilities from the teacher's and learner's perspective, respectively.

Teaching infinite mixture models

Given data $x = x_1, \ldots, x_n$ we wish to teach the assignment of data to categories, z, and the category means and covariance matrices. The IGMM framework assumes that learner knows only the prior parameters (μ_0, Λ_0, ν_0, and κ_0) and that all other quantities are unknown.

The teacher's target model, θ, consists of K means and covariance matrices, and an n-length assignment of data to categories. To draw data from $p_T(x \mid \theta)$, we employ random-walk Metropolis sampling [26, 27]. An initial set of n data are drawn from the target model, after which new data, x', are proposed by adding Gaussian noise to x. The new data are accepted ($x := x'$) according to the acceptance probability:

$$p(x' \mid x) := \min[A, 1], \quad A = \frac{\ell(x' \mid \theta)m(x)}{\ell(x \mid \theta)m(x')}. \tag{6}$$

To search for $\text{argmax}_x p_T(x \mid \theta)$ one may employ *simulated annealing* [28] by replacing A with $A^{1/T}$, such that T goes to zero with the number of Metropolis steps.

Exploiting conjugacy, we can calculate $m(x)$ exactly for a small number of data by enumerating over the set of possible assignment vectors, $z \in Z$, and for each z calculating the product of the marginal likelihoods of the data in each component given the prior parameters:

$$m(x) = \sum_{z \in Z} \text{CRP}(z; \alpha) \prod_{k=1}^{K_z} f(x_i : z_i = k \mid \Lambda_0, \mu_0, \nu_0, \kappa_0), \tag{7}$$

where K_z is the number of mixture components in z.

Teaching topic models

The number of topics is known under LDA, so we need only teach the learner $\Phi = \{\phi_1, \ldots, \phi_T\}$. We do not teach the assignment of visual words to visual topics, z, and thus marginalize over all possible z. The marginal likelihood for LDA is

$$\sum_z \left(\prod_{d=1}^{D} \text{DirCat}(z_d \mid \alpha) \right) \left(\prod_{t=1}^{T} \text{DirCat}(\{w_i : z_i = t\} \mid \beta) \right). \tag{8}$$

There are T^n terms in the sum over assignments of words to topics, thus neither Equation 4 nor Equation 8 can be computed exactly for most real-word problems. We estimate these quantities using sequential importance sampling [e.g., 21, 29].

4 Experiments

Different types of visual experience were collected by wearing a head mounted camera (NET CMOS iCube USB 3.0; 54.9o X 37.0o FOV) which sent an outgoing video feed to a laptop. Videos were recorded as observers walked around different types of environments for variable amounts of time (a nature preserve, inside a house, down-town in a city, around a University, etc). Subsequently, every 500th frame of the videos was taken as a representative sample of a given video and sample images were sorted into purely natural, outdoor scenes (no man-made structure) or scenes from indoor experience.

To derive a target distribution (means and covariance matrices of subcategories), we applied expectation maximization [EM; 30] to the orientation data from each setting (see Figure 1). EM found two categories for both indoor and outdoor images. Although each image comprises information about the amplitude of structure at specific orientations, there were qualitative visual implications of the choice of images used for teaching (see Figure 2).

The target visual topic model for LDA taken from the LDA sampler state, Φ, at the 1000^{th} iteration of Gibbs sampling. The number of topics was set to 2 to match the number of categories in the IGMM target model. The parameters, α and β, were set to maximize the probability of the images under two topics.

4.1 General Methods

To determine if our teaching model better conveyed the environmental data to humans we ran a series of psychophysical categorization tasks. If the teaching model captures cognitively natural aspects of the selection of evidence for learning, then we would expect this group to perform better than those provided examples that capture the center (mean) of the category distribution. Rather than have subjects categorize all possible images from the distribution, we focused on images that should be difficult to categorize – ambiguous images that lie somewhere between the two categories. We compared categorization of ambiguous images based on either one of the three best teaching pairs or one of the three image pairs that captured the central tendency of each inner category (the mean for Experiment 1; or most likely under the model in Experiment 2). By using multiple pairs of images for comparison, we sought to eliminate any effects of idiosyncratic semantic content (i.e. filing cabinets) in individual images. Participants were recruited through Amazon Mechanical Turk and paid for completing the task. Using a completely on-line categorization task allowed us to test the optimality of teaching categories to untrained observers. Participants were presented with a machine-selected exemplar pair (see examples in Figure 2) and 24 sequentially presented ambiguous images which they were asked to categorize as either category one (left) or category 2 (right). At least one additional image was presented as an *attention check*; one of the exemplar images was presented as a image to be categorized to eliminate subjects who were not paying attention to the task. The data from any subject (n = 43 total) who failed to correctly categorize the attention checks was not used in further analyses.

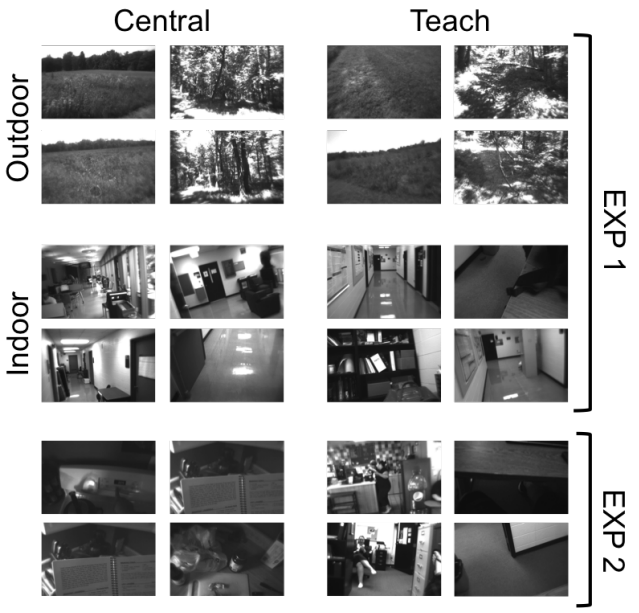

Figure 2: Examples of different exemplar pairs used in the categorization experiment for subject reference. The top row shows images used for outdoor scenes and the bottom row shows images used for indoor scenes. The left column shows the images that best capture the mean of the inner category distributions while the right column shows the example pairs picked by the model to teach the category.

4.2 Experiment 1

Experiment 1 focused on distinguishing indoor and outdoor scene types and determining if the teaching model provided better examples than images closest to the mean for each category. The ambiguous images in this experiment were chosen by calculating the Euclidean distance in orientation space each image lay from each inner category mean. The summed difference from each mean was then compared to the distance between the category means and the middle third of images closest to this value were labeled 'ambiguous'. A total of approximately 60 subjects were run in each of the 12 possible conditions (357 total). In order to minimize learning effects, the first four trials for trials were considered training and all results are based on performance on the last 18 images.

4.3 Results

In order to assess the results of the teaching model, we collapsed across the three exemplar pairs by first determining that there were no differences between them. Separate one-way ANOVAs were run and, while there were no differences for the indoor images, for outdoor images one pair in the teaching condition showed significantly lower performance than the other two runs $F(2, 124) = 54.26,\ p < .001$ (see Figure 3). Moreover, the standard deviation in this condition (.19) was more than twice any of the other 5 conditions (average = .093). Thirty percent of participants in this condition performed above chance level, correctly categorizing eleven images that were incorrectly categorized by most of

the participants who performed below chance. Interestingly, these eleven images were of open fields, which only the high-performers were grouping with the images selected for teaching that contained bodies of water. Our method for quantifying images based only on orientation content does not distinguish between fields and bodies of water (In Experiment 2 we explore a method that may map more closely to human perception by quantifying orientation in regions).

Subsequent analyses focus on the remaining two pairs of images for the teaching condition. To compensate for eliminating an entire pair of images from the outdoor condition, a random sample of equal size was extracted from the pooled total subjects who completed the task with indoor images. Further statistical analyses confirmed no differences across exemplar pairs between different runs within image type (Outdoor vs Indoor) and exemplar condition (Teach vs Mean) and thus the data from different exemplar pairs was pooled.

We tested categorization performance for the remaining teaching and mean pairs for outdoor and indoor images separately. Results plotted in Figure 3 show that participants were able to more easily categorize outdoor images than indoor images, $F(3, 319) = 20.12,\ p < .001$, which is consistent with the increased cluster separation of the outdoor categories (see Figure 1). Consequently, there were no differences in categorization performance based on the teaching exemplars as compared to the mean exemplars for outdoor images, $p = .90$; all participants performed well presumably because even the most ambiguous images were not especially difficult to categorize.

For indoor images, participants' categorization performance was significantly better for the teaching images relatively to the mean images, $p < .001$. These categories were less well separated in orientation space (see Figure 1). Consequently, the representative images selected for each category had a greater potential influence. Indeed, the teaching images, which are selected by the model to highlight the structure of the category *and* to contrast with the alternative category lead to better performance. Overall, the results of Experiment 1 indicate that for images whose categories are difficult to distinguish, the teaching model provides better exemplars for human category learners.

4.4 Experiment 2

Given the results of Experiment 1, Experiment 2 focused only on indoor images and used a higher-dimensional image quantification method. Inspired by the *spatial envelope* quantification of [7], we ran our global orientation analysis on nine sub regions of each image from Experiment 1. Each image is quantified into 36 orientation dimensions, corresponding to the 4 primary orientations (0, 45, 90, 135) in each of nine square image regions. We also sought to investigate whether the IGMM classification method outperformed Latent Dirichlet Allocation (LDA), which is less cognitively natural for categorization but is widely used for image analysis [7, 15]. The 36 dimensional data was fed directly into the IGMM, but was further quantized for the LDA model. Each image region was treated as a *word* and the *vocabulary* for indoor images was determined by K-means clustering. Each image

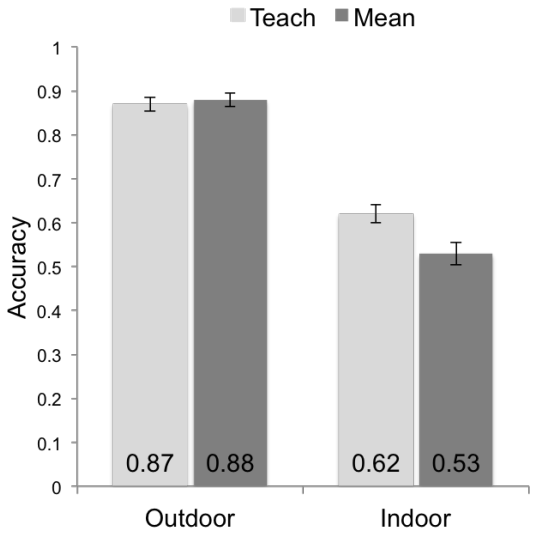

Figure 3: Results of psychophysical categorization experiments. Error bars represent two standard error of the mean.

contained nine regions summarized by four-dimensional continuous orientation data. The data from the nine regions of all of the two hundred images was assigned a cluster (word) by K-means (200 images times 9 data per image resulted in 1800 data to K-means). Elbow plots revealed that K=20 was the optimal vocabulary size. Each image was then a visual document composed of nine visual words from a 20-word vocabulary. To compare most directly with the IGMM we generated a target model with two topics by running LDA on the visual corpus. The images were then classified into one of two topics based on the higher percentage of words belonging to a given topic. Both the IGMM and LDA classification results were then fed into the teaching model to determine the best three image pairs for teaching each classification model. The highest likelihood pairs from each model were used for comparison. Ambiguous images were selected by finding their log-likelihood value under each category/topic (depending on the model), subtracting the two values, and finding the center third of images whose log likelihood difference score was closest to zero. This process led to 55 images under each model, 35 of which were identical across models. Approximately 23 subjects ran each of the 12 conditions for a total of 285 participants. Eleven were removed for incorrectly categorizing the attention checks. Preliminary analyses showed no learning affects and thus all 24 trials were included in the results.

4.5 Results

Again, we collapsed across the three exemplar pairs by first determining that there were no differences between those used within conditions (Teach vs. Likelihood). Separate one-way ANOVAs determined that the only significant difference between exemplar pairs was between pair 1 and pair 2 in the GMM likelihood condition ($F(2,68) = 3.59$, $p = 0.03$). However, neither pair 1 nor pair 2 was significantly different from pair 3 and thus all three exemplar pairs were collapsed into

overall teaching and likelihood conditions for each model. The overall 2 (model) by 2 (condition) way between subjects ANOVA showed significant main effects of teaching and model, but no interaction: $F(1,270) = 8.93$, $p = .003$, $F(1,270) = 24.87$, $p < .001$, and $F(1,274) = 0.93$, $p = 0.34$ respectively. As can be seen in Figure 4, the main effect of teaching is driven predominately by the higher accuracy scores in the GMM condition; the LDA condition shows no difference between teaching and likelihood exemplars. Accuracy scores were also significantly higher under the GMM model than the LDA model, in general. This suggests an additive improvement in performance for analysis with the mixture model and example selection with the teaching algorithm for these images on this task. Notably, performance in the GMM-teach condition is similar to that in Experiment 1, suggesting that characterizing images at the four orientations (cardinals and obliques) was sufficient to capture information about the orientation distribution in the images.

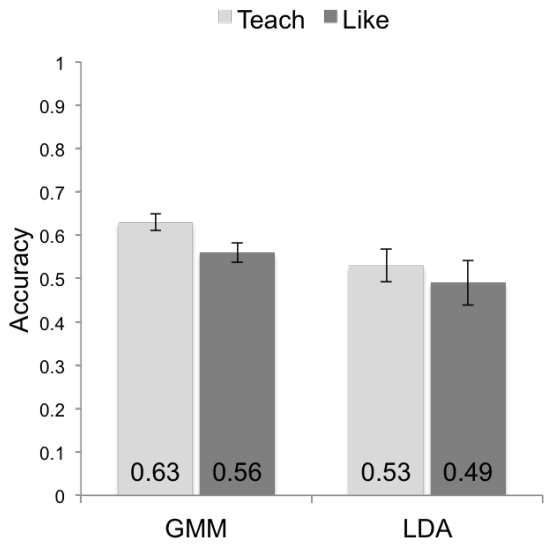

Figure 4: Results of Experiment 2. Error bars represent two standard error of the mean.

5 Conclusion

We presented an approach to optimizing the data analysis pipeline to minimize required expertise/training in data analytics in order to make informed decisions and increase accuracy. We leveraged known human competencies in perception, cognition and social learning as well as information classification methods from machine learning. We illustrated the approach in a domain and on a task where human competencies are well-investigated—scene perception and categorization, respectively—which allowed us to select established solutions to the problems of quantifying and analyzing the data. We presented an experimental investigation into the use of social learning methods, specifically a computational formalization of teaching, to provide a generic method of translating analytic results into human-understandable format. Because these experiments relied on the entire data analysis pipeline, our experiments necessarily tested both the efficacy of the computational model of teaching and the methods of quantifying and analyzing data.

Our results showed that human performance was significantly greater than chance for the two problems tested and that performance was related to the difficulty of the categorization. Results also showed that the computational teaching method performed well, exhibiting specific gains when the data analysis problem was hard. In this particularly difficult condition, the mean images failed to communicate the necessary distinction between categories. This failure demonstrates how a loss of information anywhere in the data analysis pipeline can indeed lead to incorrect outcomes. This work is a step forward in solving the problem of how to make choices along the pipeline such that users at the end are able to make informative decisions about the data without extensive training. It should be noted, however, that these results are specific to this domain of visual teaching with a relatively small sample size. The primary conclusion of this work is that complete automation and optimization of the data analysis pipeline is possible as long as one chooses a psychologically appropriate data model.

The results also highlighted known limitations in our data analysis pipeline. Specifically, we quantified images exclusively in terms of orientation content—one of the earliest steps of visual processing. This underestimates people's categorization abilities and our results revealed that while our approach performed better in general, there were specific cases where this quantification hampered decision making. Two notable instances are in the overall accuracy—we used only the most difficult images—and the case of semantic differences in outdoor images, which represent information that was not available to the models. Given the known limitations of this approach, we take this to be a promising negative result and an area for future work. Other areas for future work include generalization to domains that are less perceptually natural (e.g. radiography), to a broader array of representative decision making tasks, and exploring cases where analysis uncertainty is passed through to the decision maker. Regardless, our results indicate that creating data analysis pipelines based on human perceptual, cognitive, and social learning capacities is possible and a potentially fruitful direction for future research.

6 Acknowledgements

This work was support in part by NSF awards DRL-1149116 and CHS-1524888 to P.S. Undergraduate RA, Yunier Consuegra, programmed the Mechanical Turk experiments.

References

1. Hansen, B. C. & Essock, E. a. A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of vision* **4,** 1044–1060 (2004).

2. Schweinhart, A. M. & Essock, E. a. Structural content in paintings: Artists overregularize oriented content of paintings relative to the typical natural scene bias. *Perception* **42,** 1311–1332 (2013).
3. Wainwright, M. J. Visual adaptation as optimal information transmission. *Vision Research* **39,** 3960–3974 (1999).
4. Furmanski, C. S. & Engel, S. A. An oblique effect in human primary visual cortex. *Nature neuroscience* **3,** 535–536 (2000).
5. Sun, P. *et al.* Demonstration of tuning to stimulus orientation in the human visual cortex: a high-resolution fMRI study with a novel continuous and periodic stimulation paradigm. *Cerebral Cortex,* bhs149 (2012).
6. Hansen, B. C. & Essock, E. A. A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision* **4,** 5–5 (2004).
7. Oliva, A. & Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* **42,** 145–175 (2001).
8. Richler, J. J. & Palmeri, T. J. Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science* **5,** 75–94 (2014).
9. Ashby, F. G. & Maddox, W. T. Human category learning. *Annu. Rev. Psychol.* **56,** 149–178 (2005).
10. Anderson, J. The adaptive nature of human categorization. *Psychological Review* **98,** 409 (1991).
11. Anderson, J. R. & Matessa, M. Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning* **9,** 275–308 (1992).
12. Antoniak, C. E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics,* 1152–1174 (1974).
13. Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The annals of statistics,* 209–230 (1973).
14. Rasmussen, C. The infinite Gaussian mixture model. *Advances in neural information processing,* 554–560 (2000).
15. Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A. & Freeman, W. T. *Discovering objects and their location in images* in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* **1** (2005), 370–377.
16. Shafto, P. & Goodman, N. D. *Teaching games: Statistical sampling assumptions for learning in pedagogical situations* in *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (2008).
17. Shafto, P., Goodman, N. D. & Frank, M. C. Learning From Others: The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science* **7,** 341–351 (June 2012).
18. Shafto, P., Goodman, N. D. & Griffiths, T. L. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology* **71C,** 55–89 (Mar. 2014).
19. Bonawitz, E. *et al.* The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition* **120,** 322–30 (Sept. 2011).
20. Zhu, X. Machine Teaching for Bayesian Learners in the Exponential Family. *Advances in Neural Information Processing Systems,* 1–9 (2013).
21. Eaves Jr., B. S., Schweinhart, A. & Shafto, P. in *Big Data in Cognitive Science* (ed Jones, M.) (Psychology Press, New York, NY, in press).
22. Aldous, D. Exchangeability and related topics. *Ecole dÉté de Probabilités de SaintFlour XIII* **1117,** 1–198 (1985).
23. Murphy, K. P. *Conjugate Bayesian analysis of the Gaussian distribution* tech. rep. (University of British Columbia, 2007).
24. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3,** 993–1022 (2003).
25. Griffiths, T. L. & Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* **101 Suppl,** 5228–35 (2004).
26. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics* **21,** 1087–1092 (1953).
27. Hastings, W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109 (1970).
28. Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., *et al.* Optimization by simulated annealing. *science* **220,** 671–680 (1983).
29. Maceachern, S. N., Clyde, M. & Liu, J. S. Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics* **27,** 251–267 (1999).
30. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the royal statistical society. Series B (methodological)* **39,** 1–38 (1977).