# Efficient Data Compression Leads to Categorical Bias in Perception and Perceptual Memory

**Christopher J. Bates (cjbates@ur.rochester.edu)**

**Robert A. Jacobs (rjacobs@ur.rochester.edu)**
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY

## Abstract

Efficient data compression is essential for capacity-limited systems, such as biological memory. We hypothesize that the need for efficient data compression shapes biological perception and perceptual memory in many of the same ways that it shapes engineered systems. If true, then the tools that engineers use to analyze and design systems, namely rate-distortion theory (RDT), can profitably be used to understand perception and memory. To date, researchers have used deep neural networks to approximately implement RDT in high-dimensional spaces, but these implementations have been limited to tasks in which the sole goal is compression with respect to reconstruction error. Here, we introduce a new deep neural network architecture that approximately implements RDT in a task-general manner. An important property of our architecture is that it is trained "end-to-end", operating on raw perceptual input (e.g., pixels) rather than an intermediate level of abstraction, as is the case with most psychological models. We demonstrate that our framework can mimick categorical biases in perception and perceptual memory in several ways, and thus generates specific hypotheses that can be tested empirically in future work.

**Keywords: Perception; memory; deep neural networks; rate-distortion theory; categorical bias**

## Introduction

Biological cognitive systems are not infinite. For instance, it is commonly hypothesized that people have finite attentional and memory resources, and that these constraints limit what people can process and remember. In this regard, biological systems resemble engineered systems which are also capacity-limited. For any capacity-limited system, biological or engineered, efficient data compression is paramount. After all, a capacity-limited system attempting to achieve its goals should maximize the amount of information that it processes and stores, and this can be accomplished through efficient data compression. Of course, this raises the question of what one means by "efficient".

In engineered systems, resources (e.g., bandwidth, finite memory) are limited, and thus system designers allocate these resources so as to maximize a system's performance, a process referred to as "bit allocation" (Gersho & Gray, 1992). Consider the design of digital compression algorithms. For example, file sizes can be reduced by a substantial factor using JPEG (image) or MP3 (audio) compression while still maintaining enough fidelity for most applications. When thinking about how to best perform bit-allocation, engineers must consider several questions. Which data items are frequent, and thus should be encoded with short digital codes, and which data items are infrequent, and thus can be assigned longer codes? Which aspects of data items are important to task performance, and thus should be encoded with high fidelity via long codes, and which aspects are less task relevant, and thus can be encoded with lower fidelity via short codes? For example, frequencies beyond the range of the human ear are less important when compressing audio waveforms with MP3, and can be stored with less fidelity. To address these questions, engineers have developed rate-distortion theory (RDT), a sophisticated mathematical formalism based on information theory (Cover & Thomas, 1991).

Our goal in this paper is two-fold. First, although exact methods already exist for RDT analysis in low-dimensional spaces, approximate methods are needed for high-dimensional spaces. To date, researchers have used deep neural networks to approximately implement RDT in high-dimensional spaces, but these implementations have been limited to tasks in which the sole goal is data compression with respect to reconstruction error (e.g. Ballé, Laparra, & Simoncelli, 2016). An innovation of the research presented here is that we introduce a new deep neural network architecture that approximately implements RDT in a task-general manner. That is, our architecture discovers good data compressions even when the data will be used for regression, classification, recognition, or other tasks. An important property of our model is that it is trained "end-to-end", operating on raw perceptual input (e.g., pixels) rather than intermediate levels of abstraction (e.g., orientation, texture, shape), as is the case with most psychological models. In this way, our framework represents an early step toward scaling up models of perception and perceptual memory toward levels of complexity faced in real-world situations.

Our second goal is to present one important and previously uninvestigated implication of efficient data compression which can be compared against empirical phenomena in perception and perceptual memory. While in this paper we present only a qualitative comparison, future work can focus on more rigorous, empirical evaluations of the hypotheses that our modeling framework generates. Specifically, we examine the phenomenon of categorical bias, which we explain in more detail below.

## Principles of Efficient Data Compression and their Implications for Perception and Memory

This section examines important principles and implications of efficient data compression. We focus on one implication in particular, categorical bias, and draw a connection between

categorical bias in efficient compression and that found in perceptual memory.

All physically-realized systems are finite, and thus have finite limits on processing and storage capacities. For people, this implies that faulty perception and memory—what engineers refer to as "lossy compression"—is inevitable. If perception and memory cannot be perfect, can they at least be as good as possible given their capacity limits? This question has been explored in the context of low-level perception ("efficient coding"; see Barlow, 1961; Simoncelli & Olshausen, 2001), and researchers have found that low-level perceptual representations tend to be highly efficient with respect to the statistics of the environment.

Here, we focus on explaining higher-level sensory perception from the standpoint of efficient data compression. As we show in our results and analyses below, abstraction and categorization may be data-efficient strategies in many capacity-limited situations. There is strong empirical evidence that people employ these strategies in memory. For instance, research suggests visual working memory (VWM) avails of a wide array of summary statistics (e.g. Brady & Tenenbaum, 2013; Brady, Konkle, & Alvarez, 2009; Sims, 2016; Mathy & Feldman, 2012). In addition, various forms of abstract conceptual structures have been studied extensively in the context of long-term memory (LTM), such as schemas and scripts (Bartlett & Burt, 1933; Schank & Abelson, 1977).

A central assumption for our analysis below on categorical bias is that memory traces decay. Evidence for decay can be found in many experiments, including iconic visual memory and VWM (e.g. Sperling, 1960; Luck, 2008). We account for the decay of individual memory traces by hypothesizing that memory is biased toward representing recent information because recent information tends to be more task-relevant (Anderson, 1991). Consequently, memory engages in a form of adaptive bit-allocation in which fewer resources are devoted to older perceptual traces (suggesting that these traces are recoded in more compact and abstract ways over time) until so few resources are devoted to a trace that, effectively, the trace has fully decayed. This process frees up resources that can then be used to encode new information.

We propose that this reallocation happens both across and within memory subsystems. Within a subsystem (e.g. visual short-term memory), an individual trace tends to lose information over time to decay. Across systems, decay rates for individual traces vary. First, at stimulus offset, highly-detailed sensory information decays very rapidly. Next, sensory (e.g. iconic) memory representations are less detailed (more categorical) and decay more slowly. Short-term or working memory representations contain still less detail about the stimulus, are even more categorical and abstract, and decay more slowly than those of sensory memory. Finally, LTM contains the least amount of detail about the originally-observed stimulus, is the most categorical and abstract, and decays slowest.

For a well-designed system with limited storage, making decay rates proportional to information content is an efficient strategy—abstract representations (e.g those found in LTM) have low information content, and therefore can be retained "cheaply". As an analogy, imagine you are trying to make room on a full hard drive. It would be efficient to first remove large video files, before worrying about much smaller text files. Because highly abstract traces can be retained cheaply, LTM can accrue and store a large amount of traces over time. By contrast, working or sensory memory subsystems contain more detailed representations, and therefore cannot keep as many traces concurrently.

Consistent with our theory, experimental findings indicate that nearly all subsystems are influenced by a mix of perceptual and conceptual factors, but that the balance tilts more in favor of the conceptual the longer something is held in memory. Irwin (1991, 1992) demonstrated that iconic memory maintained more visual detail about an array of dots than VWM, whereas VWM representations seemed to be more abstract, coding information in a way that was robust to spatial translations. Brady and Alvarez (2011) found that observers' memories for the size of an object are systematically biased toward the mean of the object's category (see also Hemmer & Steyvers, 2009). Several experiments also indicate that memories for spatial location are biased toward spatial "prototypes" (Huttenlocher, Hedges, Corrigan, & Crawford, 2004; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Newcombe, & Sandberg, 1994). VWM representations not only encode "gist" or summary statistics (Oliva, 2005) over low-level visual features and textures, they also summarize high-level constructs such as the emotion of a face (Haberman & Whitney, 2007, 2009).

Visual LTM representations appear to be even more abstract. Konkle, Brady, Alvarez, and Oliva (2010) performed a visual LTM experiment in which subjects studied images of real-world objects drawn from different categories. Subjects studied between one and 16 exemplars per category, and later performed memory recognition test trials. It was found that as the number of exemplars from a category increased during study, memory performance decreased. Further analysis revealed that the conceptual distinctiveness of a category—low when category exemplars belong to the same subcategories and high when exemplars belong to different subcategories—is correlated with visual LTM performance but perceptual distinctiveness is not. The authors concluded that "observers' capacity to remember visual information in long-term memory depends more on conceptual structure than perceptual distinctiveness" (Konkle et al., 2010, p. 558).

To understand how abstraction results from efficient compression, it is important to understand the two central principles of RDT, which we name the "Prior Knowledge Principle" and the "Task-Dependency Principle". Now, we will briefly explain each principle and intuitively how each one can give rise independently to categorical representations.

**Prior Knowledge Principle:** Prior or domain knowledge is crucial to designing information-efficient systems. Accurate knowledge of stimulus statistics allows an agent to form

efficient representational codes given a limited capacity. To code a stimulus efficiently, a code must be designed using knowledge of the statistics of the to-be-coded items. Consider Morse code which is an algorithm for encoding letters of the alphabet as binary signals ("dots" and "dashes"). The designers of this code realized that they could increase its efficiency (i.e., decrease average code length) using knowledge of letter frequencies by assigning the shortest binary sequences to the most frequently transmitted letters. The more "peaky" the frequency of letters, the less information messages convey, and the shorter codes can be on average. For example, if 90% of the English language consisted of the letter 'e', then messages could be coded much more compactly on average than with real English in which e's are not nearly so frequent.

In many domains, the stimulus prior (i.e. distribution over stimuli) is highly peaked around several values. For example, if the set of stimuli consists of many photographs of various apples and bananas, this would constitute two different peaks (or modes) in the space of images around apples and bananas respectively. Efficient data compression predicts that these types of "modal" stimulus distributions will result in categorical bias. Specifically, as memory capacity is decreased (e.g. when decaying from short-term to LTM), representations should be attracted to one of the two modes, resulting in categorical bias.

**Task-Dependency Principle:** In addition to prior knowledge, for a code to be optimal, it must also take into account the current behavioral goals (or task) of an agent. Codes should allocate resources according to how an agent will use the encoded information. In particular, if it is costly to an agent to confuse stimulus values $x$ and $y$, then codes should be designed so that these values are easily discriminated, even if this means a loss of precision for other discriminations.

As was the case with prior knowledge, efficient data compression predicts that certain behavioral goals will result in categorical bias. Namely, if effective behavior depends on making category distinctions, then when capacity is decreased, efficient codes should become more biased toward category prototypes, even when the stimulus prior is uniform. Thus, efficient data compression produces two distinct hypotheses for the existence of categorical bias. Either it results from modalities in the stimulus prior or from behavioral goals. These hypotheses may be evaluated in future work.

In the next section, we present the RDT formalism in order to make the prior knowledge and task-dependency principles mathematically precise. Then, we will demonstrate in simulation that each principle can indeed give rise to categorical bias.

## Overview of Rate-Distortion Theory

Information theory addresses the problem of how to send a message over a noisy channel (e.g., a telephone wire) as quickly as possible without losing too much information. How much information can be sent per unit time (or per symbol) is the information 'rate' of a channel. Rate-distortion theory focuses on the case when the capacity (or rate) is too low to send the signal perfectly for a particular application (e.g., trying to hold a video conference with a slow internet connection). In this situation, one's goal is to design a channel that minimizes the average cost-weighted error (or distortion) in transmission, subject to the capacity limitation. Crucially, the optimization depends on two factors: (i) the prior distribution over inputs to the channel, and (ii) how the transmitted signal will be used after transmission. The first factor is important because common inputs should be transmitted with greater fidelity than uncommon inputs. The second factor is important because, depending on the application, some kinds of errors may be more costly than others.

Whereas much of the cognitive science literature uses the number of remembered "items" as a measure of memory capacity, information theory defines channel capacity as the mutual information between the input distribution and the output distribution. That is, if you know what comes out of a channel, how much information does that give you about what was inserted into the channel? If mutual information is high (high capacity), then the outputs tell you a lot about the inputs, but if it is low (low capacity), then the channel does not transmit as much information. The mutual information $I(x;y)$ for discrete random variables $x$ and $y$ is given by:

$$I(x;y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\ p(y)}. \tag{1}$$

In the case of memory, sensory stimuli (e.g., pixel values) can be regarded as inputs to a channel, and neural codes are the channel's outputs (e.g., firing rates, changes to synaptic weights). The capacity of memory is the mutual information between the stimulus distribution and the neural code.

RDT seeks to find the conditional probability distribution of channel outputs (neural codes, denoted $\hat{x}$) given inputs (sensory stimuli, denoted $x$) that minimizes an error or distortion function $d(x,\hat{x})$ without exceeding an upper limit $C$ on mutual information. For example, the distortion could be defined as the squared difference between the channel input and output, $(x-\hat{x})^2$. Mathematically, this minimization is the following constrained optimization problem:

$$\begin{aligned} Q^* = \underset{p(\hat{x}|x)}{\arg\min} & \sum_{x,\hat{x}} p(x)\ p(\hat{x}|x)\ d(x,\hat{x}) \\ & \text{subject to } I(x;\hat{x}) \leq C \end{aligned} \tag{2}$$

where $Q^*$ is the optimal channel distribution.

## Rate-Distortion Theory and Categorical Bias

Above, we described abstract or categorical representations as being an efficient strategy for compression, and pointed to evidence that human cognition makes use of this strategy. Furthermore, we noted that as the average information-content of memory traces decreases, the degree of categorical bias increases. We suggested that LTM might be viewed as using highly-compressed and categorical compressions, whereas perception uses less-compressed, less-categorical

compressions. For example, suppose you view an image of an apple. At short delays, you may remember that it was a red apple, at a longer delay, you may only remember that it was an apple, and perhaps at still longer delays, you may only remember that you saw a fruit. At long delays, categorical bias is large, because your memory for one apple is very similar to your memory for a different apple. Here, we demonstrate this phenomenon in simulation. We use a toy, one-dimensional domain in which it is possible to find the optimal lossy compression. In experiments below, we use approximate methods to extend this result to high-dimensional spaces, closer to the level of complexity that real brains must cope with.

As mentioned above, lossy compression can produce categorical bias when the stimulus prior is modal or when the loss function penalizes miscategorizations. Figure 1 demonstrates categorical bias effects in each case for unidimensional stimuli. The top panel (A) shows the case of a modal prior and squared error loss for $d$, while the bottom panel (B) shows the case of a uniform prior and categorical loss for $d$. According to the categorical loss, there is high cost to misremembering a stimulus that belongs to category $A$ as one that belongs to category $B$, but low cost to misremembering a stimulus as another member of the same category. For example, consider plants that can be grouped as edible or poisonous. Misremembering a poisonous plant as an edible plant has high cost, whereas misremembering an edible plant as a different edible plant has low cost.

Figure 1A and B illustrate that channels optimized for a modal prior or a categorical loss, respectively, yield strong categorical bias at low capacity, but little at higher capacity. In the top rows of each (low capacity), $p(\hat{x}|x)$ is nearly identical for all values of $x = x_0$ within a category, but differs for two $x_0$ from different categories. In both A and B, categorical bias arises because values closer to the modes are "safer" when capacity is low and transmission errors are likely. On the other hand, at high capacity (bottom rows), $p(\hat{x}|x)$ is tightly peaked around the true input $x_0$ in both cases. In experiments below, for brevity we only elicit categorical bias via the distortion function (panel B).

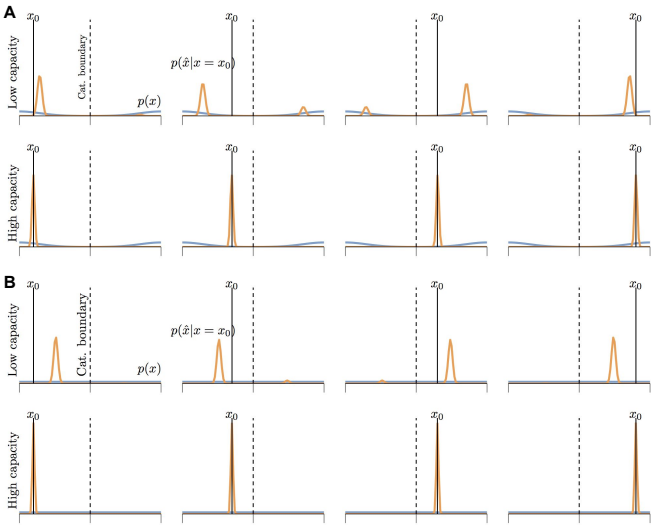


Figure 1: Illustration of how categorical bias can be explained via the prior (A) or the distortion function $d$ (B). Horizontal axes plot stimulus space, vertical axes plot probability, dotted vertical line is the category boundary, solid vertical line marks the true stimulus value ($x = x_0$), and orange line plots output distribution $p(\hat{x}|x)$. Input distribution $p(x)$ is given by the blue line. Top and bottom rows in A and B show results for low and high capacity channels, respectively. In A, distortion function was squared error and $p(x)$ was bimodal. In B, distortion function was a weighted sum between a pure categorical loss and a square-error loss with weights of 1 and 0.001, respectively, and $p(x)$ was uniform.

## RDT Neural Networks

Although RDT can be implemented exactly to find optimal compressions for problems using low-dimensional stimuli, it is too computationally expensive to be used with high-dimensional stimuli. Therefore, researchers have considered approximate implementations based on deep neural networks. To date, however, these implementations have been limited to tasks in which the sole goal is data (e.g., image) compression (e.g. Ballé et al., 2016). In this section, we introduce a new deep neural network architecture that approximately implements RDT in a task-general manner. In other words, our architecture discovers good data compressions even when the data will be used for regression, classification, recognition, or other tasks. Like previous RDT neural network implementations, our architecture is trained "end-to-end", meaning that it operates on raw sensory input (e.g., pixel values) rather than intermediate levels of abstraction (e.g., orientation, texture, shape), as is the case with most psychological models. The combination of end-to-end operation and task generality represents an important step toward scaling up models of perception and perceptual memory toward levels of complexity faced in real-world situations.

**Rate-distortion (RD) Autoencoders:** A key component of our models is the "autoencoder", parameterized models (e.g., neural networks) that map inputs to themselves subject to an information bottleneck. This bottleneck "forces" a model to find a more abstract, latent representation of the data. These abstract representations can then be used in subsequent tasks. Conventional neural network autoencoders consist of one or more 'encoder' layers, a middle 'latent' layer, and one or more 'decoder' layers. The latent layer typically has many fewer units than there are input dimensions, effectively reducing the dimensionality of the representation.

RD autoencoders differ from traditional autoencoders in that (i) they have a stochastic latent layer, and therefore a clear probabilistic interpretation, and (ii) a regularization term is added to the training objective function which acts to constrain how much information is represented in the latent units. If the coefficient on this term is high, then the network will seek a highly compressed latent representation. In our experiments, the latent unit activations are our models' "memory" of an input. Several variants of the rate-distortion autoencoder have been proposed, but here we choose the β-

Variational Autoencoder (β-VAE; Alemi et al., 2018).

**Architecture:** The models for all experiments presented here are defined by deep feedforward neural networks. Our general architecture (see Figure 2) consists of two modules: a β-VAE autoencoder and a decision module. The decision module takes as input the memory code (i.e., the activations of the latent units in the autoencoder) and optionally a task-related "probe" image, and outputs a decision variable. For example, in a change-detection task, the input to the autoencoder would be a target image, the input to the decision module would be a probe image and memory representation of the target, and the output of the decision module would be the probability that the probe is different than the target. Correspondingly, the training objective function has three terms, which can all be weighted differently to achieve different tradeoffs, corresponding to: (1) the distortion (or error) of the autoencoder's image reconstruction, (2) the information capacity of the memory representation, and (3) the decision error. Crucially, we can manipulate what kind of information is encoded in memory by varying how much reconstruction error is weighted relative to decision error during training, as well as how one kind of decision error is weighted relative to others (e.g., up-weighting errors along one stimulus dimension relative to other dimensions).

**Implementation Details:** Specific architectural choices for both experiments discussed below were standard within the neural network literature, and no specific fine-tuning was required to produce our results. In Experiment 1, we chose standard fully-connected layers with 'tanh' activation functions. The encoder and decoder both had two hidden layers, and the decision module had one. The latent layer and all hidden layers had 500 units. However, results were relatively insensitive to the choices of number of hidden units and layers, as long as the number of units was large. In Experiment 2, the encoder was composed of four $3 \times 3$ convolutional layers (32, 64, 64, and 64 filters for each layer, respectively), followed by a fully-connected layer with 1000 units. There were 1000 latent (memory) units. The decoder mirrored the encoder, except that convolutional layers were replaced with standard convolution-transpose layers. All hidden units used rectified-linear activations (ReLU). Again, a range of architectural choices can produce similar results. Finally, the decision module output was a single sigmoidal unit in Experiment 1, while in Experiment 2, the output was a softmax layer with one output unit for each of the three categories. All networks were trained with the "Adam" optimization algorithm.

**Training sets:** For Experiment 1, the dataset consists of images of an artificial plant-like object which we varied along two dimensions: leaf width and leaf angle. Images were converted to gray scale, down-sampled, and cropped to a size of $120 \times 120$ pixels. The space was discretized to 100 values along each dimension, for a total of 10,000 unique stimuli.

For Experiment 2, we used the Fruits-360 database[1]. We chose a subset of the classes to train on, specifically apples, tomatoes, and bananas. We augmented the dataset during training by randomly zooming and cropping inputs, as well as flipping the inputs horizontally at random. All images were resized to $112 \times 112$ pixels.

[1] https://github.com/Horea94/Fruit-Images-Dataset

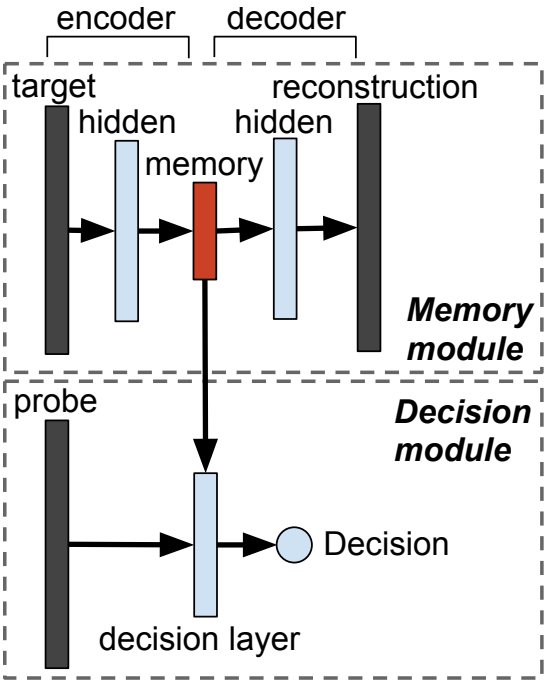


Figure 2: Schematic of the general model architecture. Dark gray boxes represent a vector of pixel values, while other boxes represent layers (or a set of layers) in the network. Layer that represents the memory code is in red.

**Experiment 1: Artificial Images:** Experiment 1 used the artificial plants dataset to demonstrate that the categorical bias effect depicted in Figure 1 extends to models operating in high-dimensional pixel-space. We show that, as expected, when a limited-capacity network is highly penalized for miscategorizing a stimulus, its memories exhibit categorical bias.

We trained the architecture on the full plants dataset. Following panel B in Figure 1, the training objective function was a mixture of pixel reconstruction error and categorical error, with a high relative coefficient on the latter. Specifically, the decision module was tasked with deciding whether the target image (input into the autoencoder) was the same category as a subsequent randomly-chosen probe image (input to the decision module). Given the high penalty for miscategorization, the optimal strategy for a model with very little capacity is to store little more than the category label. Figure 3 demonstrates this outcome by plotting target image reconstructions (outputs from the decoder) corresponding to a range of possible inputs. At low capacity (top panel), reconstructions of exemplars to the left of the category boundary are all nearly identical, and reconstructions of exemplars to the right of the boundary are also nearly identical. However, reconstructions on one side of the boundary are quite different from those on the other. In other words, there is a strong bias in the reconstructions to the appropriate category means, and thus a

sharp discontinuity at the category boundary. These results imply that at low capacity, the memory representation is a code that simply indicates which category the input belonged to. The best the autoecoder can do in this case is to produce the mean or prototype of that category. At higher capacities, the memory code contains more perceptual details beyond the category membership.

**Experiment 2: Natural Images** Experiment 2 used the Fruits-360 dataset to show that our approach scales to natural images. Again, we show that our models have increasing categorical bias as capacity decreases. However, our analyses in this experiment differ in a few ways. First, because natural image datasets do not contain a clear set of dimensions along which stimuli vary (like leaf width and leaf angle in Experiment 1), we indirectly measure the categorical bias in the trained models using autoencoder reconstructions and principle components analysis (PCA). An additional difference is that the decision module was trained to categorize each image, rather than to detect a change between target and probe.

Figure 4 (top panel) shows image reconstructions from the autoencoder at high, medium, and low capacity. These images demonstrate that the amount of detail that is retained in memory decreases as capacity decreases. At low capacity, the reconstructions are clearly categorical: each type of fruit corresponds to a unique output, which is the average of all images in that category. At medium capacity, different varieties within each species of fruit can begin to be distinguished. The figure's bottom panel demonstrates that the model's memory codes become more categorical at lower capacities. We performed PCA on memory vector activations and plotted stimuli in the space defined by the first two principle components. At medium or low capacity, memory codes for stimuli that belong to the same class are very similar, whereas at high capacity, memories of stimuli within a category are quite distinguishable from each other, and thus more perceptual details may be recovered[2].

## Conclusion

We have argued, from both theoretical and empirical standpoints, that efficient data compression may be a central goal of perceptual and memory subsystems. In future work, we will discuss the extensive empirical evidence that efficient data compression is implemented in biological perception and memory, beyond the limited examples given here. In the current work, we highlighted one interesting piece of evidence that neural systems follow these principles, specifically that categorical representations are prevalent in memory. In simulation, we showed how categorical representations can be a natural outgrowth of efficient compression. These mechanisms for categorical bias generate hypotheses that can be tested in future empirical work. Because our modeling framework operates in an end-to-end and task-general manner, we believe that it shows promise for being scalable in ways that most psychological models are not.

[2]Note that even though the principle-components space appears to scale with capacity, this does not imply that the degree of categorical bias stays constant. For example, if the magnitude of noise that is added to the latent activations is fixed, more separation between two points in principle-components space implies that the decoder can more easily distinguish between them despite the noisiness. In fact, as network capacity is increased, the magnitude of noise added to the latents tends to *decrease* (because this allow more information to be stored), and thus two points that are a distance $d$ apart in principle-components space are at least as distinguishable at high capacity compared to low capacity.

## Acknowledgments

We thank Chris Sims for many useful discussions. The first author was supported by an NSF NRT graduate training grant (NRT-1449828) and an NSF Graduate Research Fellowship (DGE-1419118). This work was also supported by an NSF research grant (DRL-1561335).

## References

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2018). Fixing a broken elbo. *arXiv preprint arXiv:1711.00464v3*.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–485.

Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages.

Bartlett, F. C., & Burt, C. (1933). Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology*, *3*(2), 187–192.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, *120*(1), 85.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Norwell, MA: Kluwer Academic Publishers.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718.

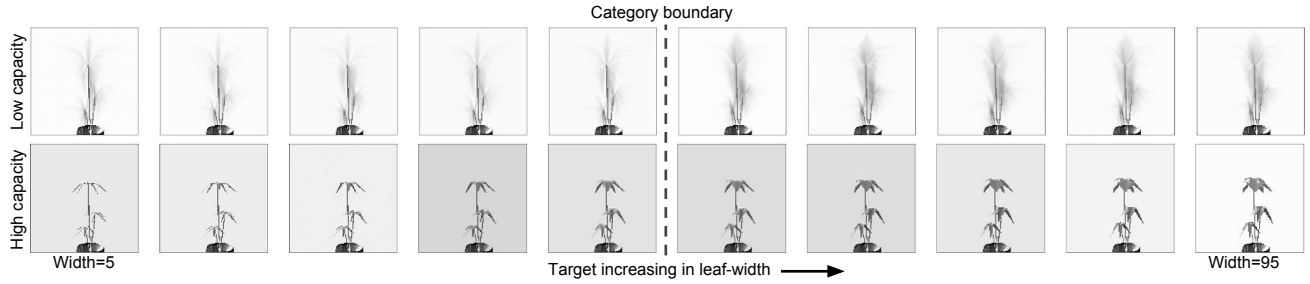


Figure 3: Visualizing categorical bias. At low capacity, all images on either side of the category boundary are highly similar to each other. Training-set stimuli included all combinations of leaf width and leaf angle. The category boundary divided skinny leaves from wide leaves, but was agnostic to leaf angle.

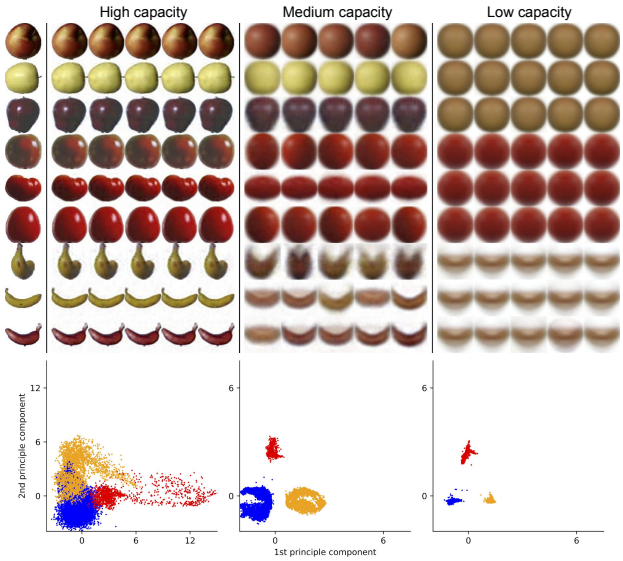


Figure 4: Top: Reconstructions from models trained on Fruits-360 dataset at three different capacities. Bottom: Corresponding PCA analysis of latent (memory) unit activations. At high capacity, latent activations have high variability within a class, whereas at lower capacities, same-class memories are highly similar.


Hemmer, P., & Steyvers, M. (2009). A bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Huttenlocher, J., Hedges, L. V., Corrigan, B., & Crawford, L. E. (2004). Spatial categories and the estimation of location. *Cognition*, *93*(2), 75–97.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological review*, *98*(3), 352.

Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive psychology*, *27*(2), 115–147.

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive psychology*, *23*(3), 420–456.

Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 307.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558.

Luck, S. J. (2008). Visual short-term memory. In *Visual memory* (pp. 43–85). New York: Oxford University Press.

Mathy, F., & Feldman, J. (2012). Whats magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362.

Oliva, A. (2005). Gist of the scene. In *Neurobiology of attention* (pp. 251–256). Elsevier.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193–1216.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*, 1–29.