American Psychological Association

Psychological Review

© 2020 American Psychological Association
ISSN: 0033-295X

2020, Vol. 127, No. 5, 891–917
http://dx.doi.org/10.1037/rev0000197

# Efficient Data Compression in Perception and Perceptual Memory

Christopher J. Bates and Robert A. Jacobs
University of Rochester

Efficient data compression is essential for capacity-limited systems, such as biological perception and perceptual memory. We hypothesize that the need for efficient compression shapes biological systems 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 human perception and memory. The first portion of this article discusses how three general principles for efficient data compression provide accounts for many important behavioral phenomena and experimental results. We also discuss how these principles are embodied in RDT. The second portion notes that exact RDT methods are computationally feasible only in low-dimensional stimulus 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 compression with respect to reconstruction error. Here, we introduce a new deep neural network architecture that approximately implements RDT. An important property of our architecture is that it is trained "end-to-end," operating on raw perceptual input (e.g., pixel values) rather than intermediate levels of abstraction, as is the case with most psychological models. The article's final portion conjectures on how efficient compression can occur in memory over time, thereby providing motivations for multiple memory systems operating at different time scales, and on how efficient compression may explain some attentional phenomena such as RTs in visual search.

*Keywords:* memory, perception, cognitive modeling, neural networks, information theory

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, digital resources (e.g., bandwidth, finite memory) are limited, and thus designers of these systems allocate these resources so as to maximize a system's performance, a process referred to as "bit allocation" (Gersho & Gray, 1992). 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? To address these questions, engineers have developed rate-distortion theory (RDT), a sophisticated mathematical formalism based on information theory (Berger, 1971; Cover & Thomas, 1991; MacKay, 2003).

Consider, for example, the problem of storing sound on a computer. This problem can be solved, for instance, using the MP3 file format which can store waveforms using roughly 10 times fewer bits relative to an uncompressed format. This compression is possible, in part, because most waveforms are not random, but rather possess rich statistical structure in which some frequencies are more common than others and transitions between frequencies are often predictable. In addition to exploiting statistical regularities, MP3 files compactly store waveforms because they do not attempt to encode items perfectly. To the human ear, certain frequencies are less discriminable than others, and hence it is less important to encode those frequencies exactly. These intuitions are readily formalized and quantified via RDT.

A goal of this article is to present a small set of general principles for efficient data compression that provides accounts for many behavioral phenomena (and many experimental results in the scientific literature) in multiple domains of perception and perceptual memory. Armed with these principles, we evaluate the hy-

This article was published Online First April 23, 2020.

Christopher J. Bates and Robert A. Jacobs, Department of Brain and Cognitive Sciences, University of Rochester.

We thank Chris Sims for many useful discussions. Christopher J. Bates 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). A preliminary version of a portion of this work appeared in a paper in the Proceedings of the 41st (2019) Annual Conference of the Cognitive Science Society.

Correspondence concerning this article should be addressed to Robert A. Jacobs, Department of Brain and Cognitive Sciences, University of Rochester, Meliora Hall, River Campus, Rochester, NY 14627-1268. E-mail: robbie@bcs.rochester.edu

pothesis 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 technical tools that engineers use to analyze and design systems, namely RDT, can profitably be used to understand perception and memory.

A second goal is to present a modeling framework for implementing the principles. Although exact methods exist for RDT analysis in low-dimensional stimulus spaces, approximate methods are needed for high-dimensional spaces. Previous 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. Our architecture discovers good data compressions even when the data will be used for regression, classification, recognition, or other tasks. Consequently, the model can perform the same types of tasks as participants in experimental studies (e.g., change-detection or recall tasks). A key property of the model is that it is trained “end-to-end,” operating on raw perceptual input (e.g., pixel values) rather than intermediate levels of abstraction (e.g., orientation, color, texture, or shape), as is the case with most psychological models. Our framework therefore represents an early step toward scaling up models of perception and perceptual memory toward levels of complexity faced in real-world situations.

Finally, a crucial import of our framework is that, unlike previous work that assumes representational spaces a priori, it predicts the representational spaces themselves from first principles. As we show below, this property provides new traction in understanding the relationship between different memory systems that have been identified by memory researchers (e.g., short-term vs. long-term visual memory). In addition, because it predicts representational spaces, it has the potential to elucidate important aspects of both neural representations (as revealed, for example, by representational similarity analysis applied to brain imaging [e.g., fMRI] data; Kriegeskorte, Mur, & Bandettini, 2008) and psychological representations (as revealed, for example, by multidimensional scaling, tree-fitting, and clustering applied to behavioral data; Shepard, 1980). By assuming that efficient compression plays a key role in designing neural and psychological representations, we may be able to better understand the computational underpinnings of these representations.

## Three Principles of Efficient Data Compression and Their Implications for Perception and Memory

It has recently been proposed—both by others (e.g., Brady, Konkle, & Alvarez, 2009; Mathy & Feldman, 2012; Yoo, Klyszejko, Curtis, & Ma, 2018) and ourselves (e.g., Bates, Lerch, Sims, & Jacobs, 2019; C. R. Sims, Jacobs, & Knill, 2012; Orhan & Jacobs, 2013)—that efficient data compression accounts for important aspects of visual working memory. Here, we provide a broad explication of the role of data compression in perception and perceptual memory, examining its underlying principles and motivations. This section lists three principles of efficient data compression which fall directly out of traditional RDT analyses and explores their implications for understanding a wide range of behavioral phenomena.

### Limited Capacity Principle

The limited capacity principle states that all physically realized systems are finite, and thus have finite limits on processing and storage capacities. For people, this is important because their capacities are generally less than the information content of their sensory environments. Consequently, people cannot perceive and memorize all sensory inputs in a veridical manner. Instead, 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? Within the study of visual perception, this question has been addressed in pioneering work by Horace Barlow on “efficient coding” (Barlow, 1961). The efficient coding hypothesis states that neurons represent afferent signals in the most efficient manner possible. To increase efficiency, neurons must eliminate redundancy in the information they encode. That is, the information that a neuron encodes should be as unique as possible from the information encoded by other neurons. The efficient coding paradigm has been extremely productive in understanding neural coding in early vision across many animal species (Park & Pillow, 2017; Simoncelli & Olshausen, 2001; Zhaoping, 2006). For instance, it provides a justification for divisive normalization, a ubiquitous neural mechanism in which neighboring neurons compete to be active in response to a stimulus (Carandini & Heeger, 2012). Neurons whose receptive field sensitivities most precisely overlap a stimulus’s features “win out,” suppressing the activations of their neighbors who are less well-suited to represent the stimulus. Efficient coding has also led to the idea of “sparse coding,” which predicts V1 simple-cell-like receptive fields as an optimal adaptation to natural image statistics given a constraint on how many neurons may be active at a given time (Olshausen & Field, 1996, 1997).

Although Barlow’s efficient coding hypothesis has been most applicable to early perceptual areas, a related idea—the “bounded rationality” hypothesis—has been productive in studying aspects of higher-level cognition. This idea was first championed by Herbert Simon (Simon, 1972), and views behavior as being adapted to physical constraints of the agent. For example, an agent may have limited time or energy resources to devote to a reasoning problem which prevent it from reaching the correct conclusion. Boundedly rational theories of cognition differ according to what kinds of limits they assume (Griffiths, Lieder, & Goodman, 2015; Lewis, Howes, & Singh, 2014). Our theory is most closely related to the “resource rational” framework of Griffiths et al. (2015), who propose that agents seek to optimize the “value” of a computation, defined as the expected cumulative reward of performing a particular computation minus the cost of the computation.

How should researchers quantify the costs of computation? In the resource-rational framework, capacity limits are typically quantified in ways that are specific to the choice of algorithm. For example, Vul, Goodman, Griffiths, and Tenenbaum (2014) described costs in terms of how long it takes to draw a sample from a Bayesian posterior distribution over a decision variable. If responding quickly leads to more trials and therefore more opportunities for reward, then there is an opportunity cost to drawing more samples on an individual trial, even if additional samples provide greater certainty about the correct response. As another example, Griffiths, Sanborn, Canini, Navarro, and Tenenbaum (2011) model human category learning with several different approximate algorithms, including particle filters. In particle filter algorithms, the computational cost corresponds to the number of "particles" used in the approximation, with higher numbers being more costly but more accurately approximating the distribution of interest. The effective number of particles that humans have can then be inferred from their errors.

Information theory offers a different approach to quantifying computational costs, which is more abstract and not tied to any particular algorithm. For example, Genewein, Leibfried, Grau-Moya, and Braun (2015) take a similar approach to ours, modeling an agent's actions as outputs of a limited-capacity communication channel whose inputs are world states. That is, the agent is represented by a capacity-limited stochastic function that maps world states to actions. The agent's capacity determines how its actions can be tailored to particular world states. At low capacities, an agent cannot distinguish between all world states, and therefore cannot take the optimal action for each state. Instead, the agent must abstract over multiple world states and treat those states as effectively the same. In other words, it must categorize world states. At high capacities, world states can be perfectly represented and the optimal action taken for each state.

Consistent with Genewein et al. (2015), a central point of this paper is that abstraction and categorization are key strategies for compressing memoranda (see A Corollary to the Principles of Efficient Data Compression: Categorical Bias and Abstraction section). It is not a coincidence that various forms of abstract stereotyped conceptual structures have been studied extensively in the context of memory such as schemas and scripts (Bartlett & Burt, 1933; Schank & Abelson, 1977). Psychologists and artificial intelligence researchers have proposed that people encode events relative to schemas in order to increase memory performance (R. C. Anderson, 1984), and there is evidence that people misremember these episodes (such as stories they have read) in ways that are influenced by schemas (e.g., Bartlett & Burt, 1933). These nonveridical memories based on abstract schemas suggest that long-term memory (LTM) is capacity-limited.

Is there similar evidence that short-term or working memory is capacity-limited? The answer is yes, and a large body of literature has focused on how to characterize its capacity limits especially in the domain of visual working memory (VWM; see A. Baddeley, 2003; Brady, Konkle, & Alvarez, 2011; C. R. Sims, 2016; Cowan, 2008; Luck & Vogel, 2013; Ma, Husain, & Bays, 2014). Although some researchers hypothesize that VWM is fundamentally limited by the number of discrete "items" that can be maintained, others—in the spirit of bounded rationality and efficient coding—have assumed a more generic capacity limit (Alvarez, 2011; Bates et al., 2019; Brady & Alvarez, 2015; Brady et al., 2009; Brady & Tenenbaum, 2013; C. R. Sims, 2016; Mathy & Feldman, 2012; Yoo et al., 2018). In particular, this body of work provides strong evidence that VWM makes use of a wide array of "gist" representations or summary statistics, just as we saw in the context of early visual perception, schemas, categorization, and decision-making. An interesting possibility, which we explore below, is that the abstract and categorical representations found in VWM, LTM, and other cognitive subsystems are boundedly rational adaptations to capacity limits in memory.

Another important consequence of limited capacity is that more complex stimuli, which carry more sensory information, should be more difficult to veridically remember. Supporting this prediction, there is strong evidence from the VWM literature that the complexity of to-be-remembered objects and the number of object features have important impacts on how well objects are remembered (Alvarez & Cavanagh, 2004; Brady & Alvarez, 2015; Fougnie, Asplund, & Marois, 2010; Ma et al., 2014; Oberauer & Eichenberger, 2013; Wheeler & Treisman, 2002; Xu & Chun, 2006).

## Prior Knowledge Principle

The prior knowledge principle states that accurate knowledge of stimulus statistics allows an agent to form efficient (i.e., compact) representational codes given a limited capacity. To represent a stimulus efficiently, a code must be designed using knowledge of the statistics of the to-be-coded items. As a simple example, consider Morse code, which is a method for encoding letters of the alphabet into 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. An occasional message may be assigned a long code if it happens to have many infrequent letters but, on average, messages can be assigned much shorter codes in this way. Furthermore, the more "peaky" the frequency of letters is, then the shorter the codes assigned to messages can be on average. For example, if 90% of the English language consisted of the letter "e," then messages could be coded more compactly on average than with real English in which e's are not nearly so frequent. Crucially, a peakier English language is also less informative, and therefore messages in this language are more compressible. These basic principles used in Morse code are fundamental and are universally used in digital compression algorithms such as Huffman or arithmetic coding.

There are at least four predictions suggested by the prior knowledge principle. First, neural activity should be higher (less efficient) for unnatural images than for natural images. Research on efficient neural coding provides evidence for this prediction. Investigators have found that neural codes in early visual areas are specifically adapted to natural image statistics, and therefore neural representations are less efficient (more spikes per second) in the presence of less natural images (Simoncelli & Olshausen, 2001). Note the analogy to the Morse code example in which "unnatural" (i.e., uncommon messages) are assigned longer codes.

Second, if codes are better adapted to natural image statistics, then performance should be worse in perceptual and memory tasks that use unnatural stimuli. The literatures on visual perception and VWM performance provide strong support for this prediction. A large number of studies in visual perception demonstrate performance superiority when stimuli better match natural image statistics (Bar, 2004; Boyce, Pollatsek, & Rayner, 1989; Davenport & Potter, 2004; Fang, Kersten, Schrater, & Yuille, 2004; Girshick, Landy, & Simoncelli, 2011; Greene, Botros, Beck, & Fei-Fei, 2015; Knill, Field, & Kersten, 1990; Lythgoe, 1991; Oliva, 2005; Parraga, Troscianko, & Tolhurst, 2000; Schwarzkopf & Kourtzi, 2008; Spence, Wong, Rusan, & Rastegar, 2006; Stocker & Simoncelli, 2006; Weckström & Laughlin, 1995). Moreover, although it is more popular to study VWM with simple artificial stimuli (Orhan & Jacobs, 2014), there is evidence that people perform better on memory tasks when stimuli consist of natural objects (Brady, Konkle, Oliva, & Alvarez, 2009; Brady, Störmer, & Alvarez, 2016; Melcher, 2001) or objects of expertise (Curby, Glazek, & Gauthier, 2009), or when objects are arranged in an expected manner (Kaiser, Stein, & Peelen, 2015).

Third, if people adapt their coding strategy through learning stimulus statistics, performance should be worse on training sets that are intrinsically less compressible. Experiments in VWM provide evidence for this prediction. Broadly speaking, these experiments have shown that performance limits can be predicted based on how intrinsically difficult it is to efficiently compress the stimulus set. For example, Bates et al. (2019) presented subjects in different conditions with stimuli drawn from different stimulus distributions (uniform or Gaussian). These subjects viewed images of plant-like objects that varied by the width of their leaves. They found that if leaf-widths were drawn from a Gaussian distribution, the number of correct responses was higher compared to when leaf-widths were drawn uniformly from the entire stimulus space. This result is predicted if people represent stimuli using compressed codes because uniform (flat) distributions are intrinsically less compressible than Gaussian (more peaked) distributions (as in the Morse code example above). Furthermore, these results indicate that people adapt their memory codes over time as the statistics of their environments change. By analogy to Morse code, it would be as if people learn that the frequencies of each letter in their language have shifted over time, and they modify their code assignments accordingly.

The work just mentioned is consistent with other findings. For example, Brady et al. (2009) trained subjects to memorize visual displays containing several objects, where each object consisted of two concentric, colored circles. Subjects were not told that the experimenters introduced correlations between certain color pairs such that some pairs were more likely to appear. Over several blocks of trials, subjects steadily improved performance while the correlations were present, but their performance dropped back to baseline during a small subset of trials where the correlations were removed. Thus, the improved performance must have been due to statistical learning of the color co-occurrence statistics rather than an increase in information capacity (but see Huang & Awh, 2018, regarding mechanistic ways that recall may differ in the statistical learning case).

C. R. Sims (2016) reviewed additional findings indicating that the compressibility of training stimuli influences recall performance. In "span of absolute judgment" tasks, people are trained to map a stimulus dimension (e.g., the length of a line) to a finite set of ordinal categories. For example, there may be 10 different lines increasing in length from short to long, and subjects must decide which ordinal category a line belongs to (Length 1, Length 2, etc.). Researchers have found that as the number of categories to choose from increases, performance declines, even when all stimuli are far enough apart to be fully discriminable (Rouder, Morey, Cowan, & Pealtz, 2004). This result is predicted by the requirements of efficient data compression—if people are capacity-limited, performance should decrease with category set-size because augmenting the set of values that need to be transmitted makes compression harder.

Finally, if both perception and memory obey the prior knowledge principle, then there should be overlap in the biases that are observed in these subsystems. Although it is not typical to directly compare response biases in perception and memory, there is at least some evidence that memory inherits biases from perception (e.g., Montaser-Kouhsari & Carrasco, 2009). Indirect evidence comes from the fact that there is a large degree of neural and representational overlap between perception and memory. For example, which stimulus a subject is retaining in working memory can often be decoded from the responses of early visual cortex (Christophel, Hebart, & Haynes, 2012; Emrich, Riggall, LaRocque, & Postle, 2013; Harrison & Tong, 2009; Kang, Hong, Blake, & Woodman, 2011; Serences, Ester, Vogel, & Awh, 2009; Wolff, Jochim, Akyürek, & Stokes, 2017), and the amount of decodable information decreases with memory load (D'Esposito & Postle, 2015; Emrich et al., 2013; Wolff et al., 2017). It has also been found that memory maintenance impacts concurrent (Konstantinou, Bahrami, Rees, & Lavie, 2012) as well as subsequent perceptual processing (Saad & Silvanto, 2013a, 2013b) in ways suggesting memory and perception share common neural substrates. In addition, perceptual stimuli are found to interfere with memory maintenance, but only when they are specifically targeted to the values being maintained in memory (Pasternak & Greenlee, 2005).

## Task-Dependency Principle

For a code to be optimal, it must take into account the behavioral goals of an agent. The task-dependency principle states that codes should allocate resources according to how an agent will use that 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 stimulus values. This principle has two entailments: (a) stimulus dimensions that are task-irrelevant should be allocated minimal resources, and (b) particular stimulus values within a given dimension that are task-irrelevant should be allocated minimal resources.

Experiments on VWM provide strong evidence in support of the first entailment. For example, subjects in an experiment reported by Bates et al. (2019) learned arbitrary category boundaries in which one stimulus dimension (e.g., leaf width), but not another dimension (leaf angle), was category-relevant. It was found that their errors in a subsequent memory task increased along the irrelevant dimension but decreased along the relevant dimension, as would be expected if they were strategically reallocating resources based on task demands. Yoo et al. (2018) reported a VWM experiment in which subjects were precued on each trial as to how likely it was that each dot in a multidot display would need to be remembered. They showed that memory errors for location decreased monotonically with probability that a dot would be probed. Fougnie, Cormiea, Kanabar, and Alvarez (2016) similarly manipulated the information that subjects maintained in VWM in a task-dependent manner. In one condition, people remembered the colors of all items in a display and were asked to recall one of the colors selected at random. In the other condition, subjects were queried about all items' colors but a response was only considered correct if all recall judgments were on the correct half of the color wheel. Results indicate that both guessing rates and precision fell in the latter condition, consistent with the prediction that subjects spread their limited memory resources more evenly across the items in a display. Swan, Collins, and Wyble (2016) used a "surprise" trial in a VWM experiment to show that people allocated more capacity to a task-relevant dimension. They blocked trials, such that the first half of the experiment only queried about one stimulus dimension (color), and the second block, which began without warning, queried about a different dimension (orientation). On the first trial of the second block (the "surprise" trial), subjects had lower precision on orientation and higher precision on color compared to the remaining trials. The above examples indicate that people can rationally allocate their capacity to different stimulus dimensions of an image to improve task performance.

The second entailment mentioned above states that to allocate maximal resources to the most task-relevant information, it is important to consider how confusable different stimulus values along a single dimension are to prevent costly confusions. According to C. R. Sims (2018), evidence for this entailment is provided by classic experiments on perceptual discrimination and generalization. In generalization experiments, people are required to discriminate between pairs of stimuli. If two stimuli are more similar to each other, they will be more likely to be confused, meaning that a subject will be more likely to generalize from one to the other. However, the task-dependency principle predicts that the probability of generalizing from stimulus $x$ to stimulus $y$ should reflect how behaviorally costly it is to confuse them. Moreover, if people can adapt their codes to changing demands (e.g., rewards from the experimenter), their pattern of errors should be predicted by efficient compression. An experiment by Kornbrot (1978) provides such an example. On each trial, a subject heard a tone and had to categorize the intensity of the tone as belonging to one of $N$ ordinal categories. Points were awarded for correct answers and subtracted for incorrect answers, but overestimates were penalized more than underestimates, making penalties asymmetric. In the analysis by C. R. Sims (2018), subjects were found to have efficiently adapted to this particular asymmetric cost function imposed by the experimenter. Specifically, it was found that subjects made fewer overestimation errors but more underestimation errors compared to a control condition with symmetric penalties, suggesting a strategic tradeoff.

## RDT

In the previous section, we presented several core principles of efficient data compression, discussed their implications for perception and perceptual memory, and reviewed experimental data indicating that the principles provide accounts of a wide range of perceptual and mnemonic phenomena. In this section, we present a brief introduction to concepts in information theory and RDT, which form the mathematical basis for those principles.

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. RDT focuses on the case when the capacity (or rate of transmission) 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).[1] 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: (a) the prior distribution over inputs to the channel ("prior knowledge principle"), and (b) how the transmitted signal will be used after transmission ("task-dependency principle"). The first factor is important because inputs that are uncommon do not need to be transmitted as accurately as inputs that are more common. The second factor is important because, depending on the application, some kinds of errors may be more costly than others.

Recall the example introduced toward the beginning of this paper in which audio signals are coded using the MP3 file format. This format is able to compress files to a fraction of their original size by taking advantage of the two factors just mentioned. First, audio waves are not usually completely random, and therefore their regularities can be taken advantage of by replacing parts of the original waveform with educated "guesses" based on context. Second, people are unable to perceive extreme frequencies, so this information can be discarded without perceptible differences. In theory, these same engineering strategies for signal transmission are also applicable to biological information processing including human perception and memory.

Whereas much of the scientific literature uses 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 arrays) can be regarded as inputs to an information channel, and neural codes are the channel's outputs. Stimuli observed in the world follow some distribution, and the neural memory codes follow a distribution conditioned on the observed stimuli. The capacity of memory is the mutual information between the stimulus and the neural code. If the mutual information is high, then memory has high capacity and neural decoding (i.e., predicting sensory stimuli from neural activity) can be detailed and accurate. In contrast, if it is low, then memory has low capacity and neural decoding cannot be as precise.

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

[1] Although we will use the terms interchangeably in this article, technically speaking, *rate* and *capacity* are not completely synonymous. Both rate and capacity are measured in units of bits per item (or "symbol"). For example, a biological memory system would be able to transmit an average of $x$ bits per image studied. But in RDT, *capacity* refers more specifically to the maximum achievable rate for a given channel design, whereas *rate* is the average information per symbol given the choice of a particular prior. Thus, capacity is expressed as the supremum of rate over all possible prior input distributions.

$d(x, \hat{x})$ without exceeding an upper limit $C$ on mutual information. Mathematically, this minimization is the following constrained optimization problem:

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

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

As illustrated in Figure 1, codes that are optimized according to RDT have several intuitive properties. For explanatory purposes, this figure assumes a one-dimensional stimulus space in which the distribution of stimulus values, often referred to as the prior or input distribution and denoted $p(x)$ (blue distribution in this figure), is either Gaussian or uniform. As illustrated in Figure 1a, as capacity decreases, the optimal channel distribution $p(\hat{x}\,|\,x)$ where $x = x_0$ (orange distribution) gets less precise or flatter (compare the channel distribution when the capacity is three bits [left graph] vs. when it is one bit [right graph]). Figure 1b illustrates that $p(\hat{x}\,|\,x)$ also becomes less precise when the entropy of the input distribution $p(x)$ increases.[2] With a narrow input distribution (low entropy; right graph), many stimulus values can be largely ignored based on how infrequently they are encountered, meaning that the channel can allocate more resources or bits to more frequent values. But with a more dispersed input distribution (high entropy; left graph), many more stimulus values are potentially important, and thus the optimal channel must allocate bits more evenly across the stimulus space. In addition, this panel demonstrates that the output of the channel will be biased toward the mean of the input distribution, with the bias increasing as this distribution's entropy decreases (Figure 1b, left graph: large input distribution entropy; right graph: small entropy). Intuitively, biasing toward the mean reduces average channel error because it ensures that more outputs will fall in regions of higher input probability. Figure 1c shows that if a visual display has multiple objects and capacity is allocated equally across them, then channel precision will decrease with number of items. Finally, the choice of distortion function can have a profound effect on the channel distribution (not illustrated here).

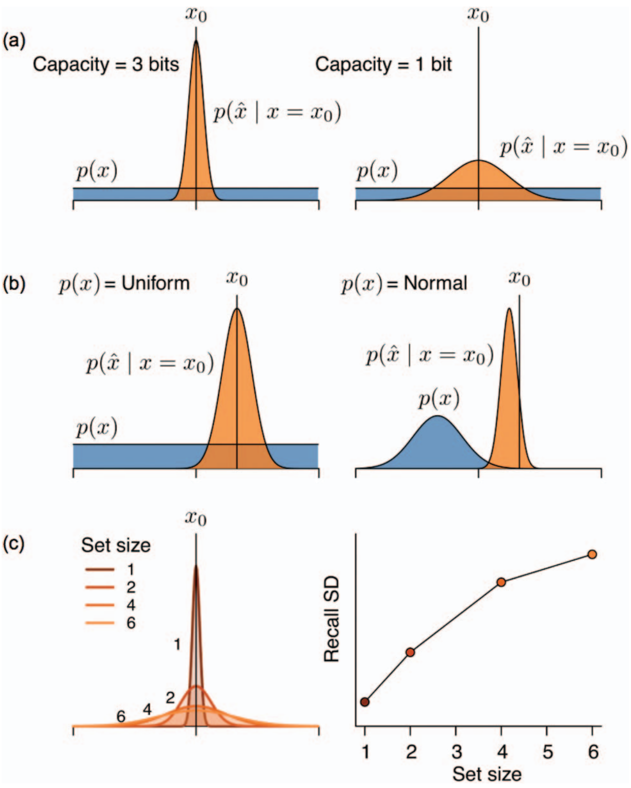

*Figure 1.* Predictions of rate-distortion theory for unidimensional sources. See text for details. *SD* = standard deviation of the response error. From "Adaptive Allocation of Human Visual Working Memory Capacity During Statistical and Categorical Learning," by C. J. Bates, R. A. Lerch, C. R. Sims, and Robert A. Jacobs, 2019, *Journal of Vision, 19*(2), p. 11. Adapted with permission. See the online article for the color version of this figure.

## A Corollary to the Principles of Efficient Data Compression: Categorical Bias and Abstraction

The study of categorization has played a prominent role in the field of psychology. Here, we explore an important link between categorical representations and efficient data compression. As noted briefly above in the context of the limited capacity principle, optimal data compressions often produce abstract or categorical representations. For instance, representations in VWM are often biased toward category means and representations in LTM appear to be more sensitive to categorical than perceptual features (see below). This section demonstrates that these properties can arise from efficient compression, making certain assumptions about either the stimulus prior $p(x)$ or the distortion function $d$.

Consider the case where the input distribution over stimulus features has multiple modes. For example, stimuli may be drawn from a mixture of Gaussian distributions, where each mixture component can be interpreted as a category. In this scenario, optimal compressions should appear more categorical as the available capacity decreases because a channel can reduce expected error by producing outputs that are biased toward modes of the input distribution (i.e., regions of the stimulus space from which stimuli are very common). Roughly speaking, as the capacity of a channel decreases, categorical bias increases. At very low capacities, the best a channel can do is to output category prototypes (i.e., mean values of mixture components).

This idea is illustrated in Figure 2 for a one-dimensional stimulus space in which there are two categories of stimuli. In each graph of this figure, the horizontal axis plots the stimulus space, the vertical axis plots probability, the dotted vertical line is the category boundary, and the solid vertical line plots the true stimulus value ($x = x_0$). The input distribution $p(x)$ is colored blue and the channel or output distribution $p(\hat{x}\,|\,x)$ is colored orange. The top and bottom rows show results for low and high capacity channels, respectively, assuming a square-error loss function. When the channel has low capacity, the output distribution is always centered near the category prototype (i.e., the mean of the component Gaussian distribution) that is closest to stimulus $x_0$.

[2] Entropy is a measure of how predictable a random variable is, and therefore how much information it carries. On one extreme, if a variable is uniformly random over some domain, it is hard to predict and therefore one gains a lot of information by observing its value. On the other extreme, if a variable is a delta function (i.e., it always takes on the same value), then no information is gained by observing its value.

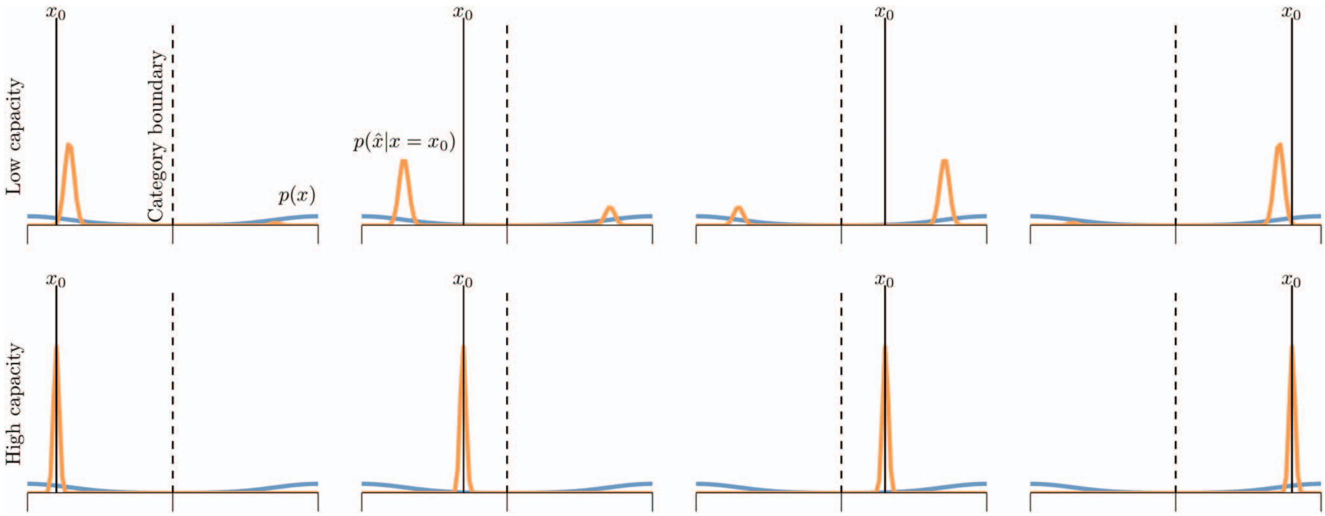

*Figure 2.* Illustration of how categorical bias can be explained via the input distribution $p(x)$. See text for details. The means of the two mixture components of the input distribution were 0 and 100 (coinciding with the axis extremes). Distortion is square-error. See the online article for the color version of this figure.

That is, the channel shows a large categorical bias. In contrast, when the channel has high capacity, the output distribution is centered near stimulus $x_0$. It transmits much more fine-scale perceptual detail about the stimulus in this case.

Alternatively, categorical bias can arise through the distortion function *d*. Consider a channel with a categorical distortion function and a uniform input distribution. According to the distortion function, 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 a high cost, whereas misremembering an edible plant as a different edible plant has low cost. As illustrated in Figure 3, this scenario, like the scenario above in which the input distribution was a mixture distribution, will also yield categorical bias at low capacity. In the top row of this figure, there is a sharp jump in outputs when the true stimulus value crosses the category boundary. This categorical bias arises because it minimizes the possibility of making a costly miscategorization at low capacity.

## Rate-Distortion and Bayesian Approaches

RDT provides a framework for finding optimal codes or representations, and thus it can be used to model aspects of optimal

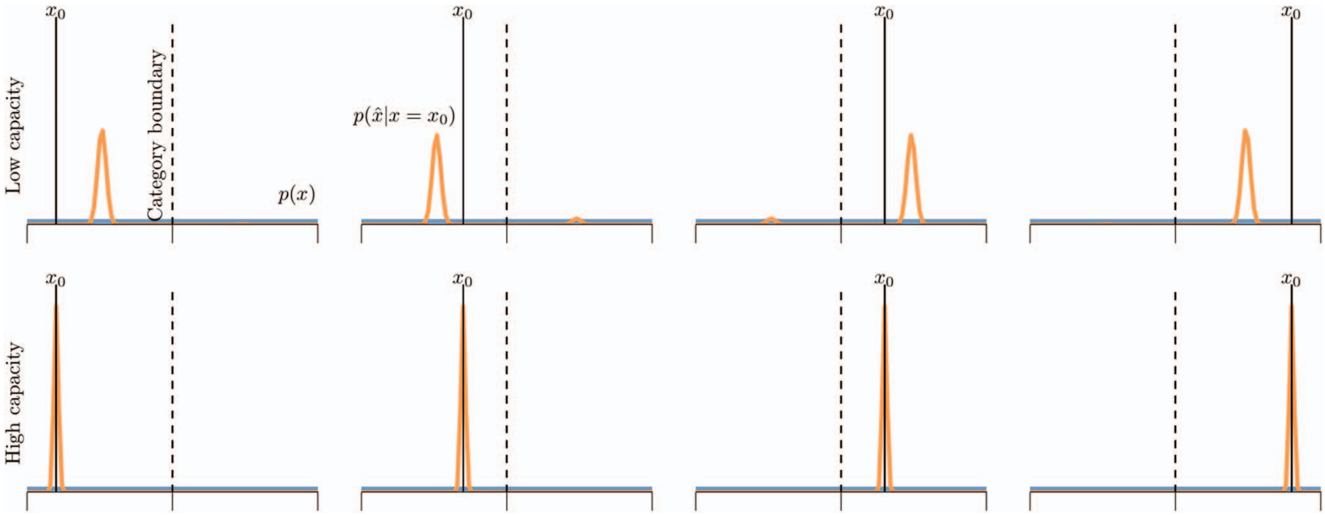

*Figure 3.* Illustration of how categorical bias can be explained via the distortion function *d*. See text for details. The distortion function was a weighted sum between a pure categorical loss and a square-error loss with weights of 1 and 0.001, respectively. See the online article for the color version of this figure.

information processing. In the field of psychology, however, it is more common for researchers to model optimal information processing using a Bayesian approach. What is the relationship between RDT and Bayesian approaches?

RDT and Bayesian approaches often make similar predictions. This is expected because RDT makes extensive use of Bayesian statistics. However, these predictions are not always identical. Differences in their predictions stem from the fact that Bayesian approaches do not make assumptions about capacity limits, whereas RDT assumes that the processes under study are capacity-limited. For Bayesian approaches, performance is limited solely by "noise." In different contexts, this noise is referred to as sensory noise, perceptual noise, memory noise, decision-making noise, or motor noise. For RDT, in contrast, performance is limited by both noise and limits on capacity.

For example, consider a VWM experiment in which subjects must remember and later recall the visual features of objects in displays. When relatively few objects appear in displays, recall performance is often good, but it degrades rapidly when displays contain more objects, a phenomenon known as the "set size" effect. Researchers have modeled set-size effects using Bayesian approaches by assuming that the variance of perceptual noise increases as the number of objects in displays increases (e.g., van de Berg, Awh, & Ma, 2014). Unfortunately, although this assumption is needed to account for the empirical data, it is lacking an independent theoretical justification. In contrast, set-size effects can be modeled using an RDT approach by assuming that the variance of perceptual noise and the capacity of VWM are fixed constants (see Figure 1, Panel c). Intuitively, RDT accounts for set-size effects because VWM's limited capacity is "spread thinner" across objects when displays contain more objects.

As a second example, consider how optimal memory performance should change with changes in the stimulus or input distribution $p(x)$. As discussed in Bates et al. (2019), RDT predicts that performance should steadily decline with increases in the standard deviation of this distribution, whereas a conventional Bayesian approach predicts that this performance will degrade slowly. At an intuitive level, these differences in predictions are expected. RDT assumes that an optimal system allocates its limited capacity to cover the entire stimulus range, and thus this capacity is "spread thinner" across the stimulus space as the size of the range increases. In contrast, a conventional Bayesian approach assumes that an optimal system does not have a capacity limit, and thus optimal memory performance can be robust to increases in the size of the stimulus range. As reported in Bates et al. (2019), experimental data are qualitatively consistent with the predictions of RDT.

## Computer Simulations: Preliminaries

Exact methods exist to find optimal channels based on RDT (Blahut, 1972), and past work has used these methods to develop RDT accounts of perception and memory (Bates et al., 2019; C. R. Sims et al., 2012; C. R. Sims, 2016, 2018; Lerch, Cui, Patwardhan, Visell, & Sims, 2016). Above, we used these methods to find optimal channels exhibiting different degrees of categorical bias.

Unfortunately, exact methods are computationally feasible only with low-dimensional stimulus spaces, and thus cannot be used in more realistic situations. Researchers have therefore considered approximate methods such as the use of deep neural networks to approximately implement RDT in high-dimensional spaces. To date, however, these implementations have been limited to tasks in which the sole goal is data compression with respect to reconstruction error (e.g., Ballé et al., 2016; see also Alemi et al., 2018; Han, Lombardo, Schroers, & Mandt, 2018; Santurkar, Budden, & Shavit, 2018). An innovation of the research presented here is that we introduce a new deep neural network architecture that approximately implements RDT. Our architecture discovers good data compressions even when 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., pixel values) rather than intermediate levels of abstraction (e.g., orientation, color, texture, or 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.

There are three key motivations for generalizing RDT models of biological systems to high-dimensional spaces. First, if biological systems have discovered sophisticated schemes to compress high-dimensional data, it should be productive for psychologists to create models that solve the same problem. Biological and model solutions can be compared, and the latter evaluated in terms of their predictive power. Second, models that take as input raw perceptual data (e.g., pixel values) are more general than models that assume a stimulus-specific, intermediate set of features. Many models in the psychology literature assume that images are (magically) parsed into objects, and that objects are (magically) encoded as finite sets of simple features (e.g., orientation, color, etc.). However, most natural images cannot be unambiguously parsed in these ways (Orhan & Jacobs, 2014). As an extreme example, if subjects view images with no semantic content (e.g., white noise), it is not clear which features they should use to encode the image information. After all, these images contain no objects, parts, surfaces, contours, and so forth. However, a model that takes raw pixel values directly as input can still make predictions about behavioral performance in this case. Third, compressing raw perceptual data results in a representational code, which can be compared to neural codes. Thus, aspects of neural activity can in principle be predicted using the representational spaces that result from efficient compression of raw inputs.

Furthermore, although many engineering algorithms already exist for compressing digital images, video, and audio, these methods are not suitable as cognitive models for several reasons. First, such algorithms do not typically include ways to vary either the loss function or the capacity arbitrarily—manipulations that are key for the purposes of modeling human behavior. Second, they are not designed to adapt over time to new input distributions, contrary to the adaptive nature of biological systems (Bates et al., 2019). Finally, popular compression algorithms do not include a hierarchy of abstractions, from highly perceptual to highly categorical, as seen in perception and memory.

## Rate-Distortion Autoencoders

A key component of our models is the "autoencoder." In general, autoencoders are parameterized models (e.g., neural networks) that are trained in an unsupervised manner to map input data items to (approximate) copies or reconstructions of these items subject to an "information bottleneck." Our simulations used a specific variant of autoencoders known as variational autoencoders (VAEs; Kingma & Welling, 2013; Rezende, Mohamed, & Wierstra, 2014).

A VAE consists of three parts or stages. The goal of the first two stages, referred to as the *encoder* and the *sampler*, is to map each data item to a probability distribution over hidden or latent variables. As explained below, a latent representation is not a fixed value such as a fixed vector of neural network activation values. Instead, a latent representation is a probability distribution over vectors of activation values. For instance, if $\mathbf{x}_i$ denotes the $i$th data item and $\mathbf{z}_i$ denotes its corresponding latent variable, then the goal of the first two stages is to estimate the distribution $p(\mathbf{z}_i \mid \mathbf{x}_i)$ and to sample $\mathbf{z}_i$ from this distribution.

Assume, for example, that latent variable $\mathbf{z}_i$ is a vector with $J$ components (i.e., $\mathbf{z}_i = [z_{i1}, \ldots, z_{iJ}]^T$) where each individual component, labeled $z_{ij}$, has a Gaussian distribution with mean $\mu_{ij}$ and variance $\sigma^2_{ij}$. Then $p(\mathbf{z}_i \mid \mathbf{x}_i)$ is a Gaussian distribution defined by $2J$ parameters (denoted $\{\mu_{ij}, \sigma^2_{ij}\}_{j=1}^{J}$; there is one mean and one variance for each component of $\mathbf{z}_i$).

The first stage of a VAE is an encoder that maps each data item $\mathbf{x}_i$ to a set of latent mean and variance parameters. This encoder is typically implemented as a conventional neural network. For instance, it may have one or more layers of hidden units, where each hidden unit first computes a weighted sum of its inputs and then computes its activation value using a nonlinear activation function. An encoder has $2J$ output units whose activation values correspond to the values of the latent mean $\{\mu_{ij}\}_{j=1}^{J}$ and variance $\{\sigma^2_{ij}\}_{j=1}^{J}$ parameters.

The output of the encoder is the input to the second stage of a VAE, the sampler. The sampler takes the mean and variance parameter values produced by the encoder and samples from a Gaussian distribution defined by these values to produce a value for latent variable $\mathbf{z}_i$ (i.e., $z_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2_{ij})$). In some applications, it may be desirable to sample multiple times in order to obtain a histogram estimating the latent distribution $p(\mathbf{z}_i \mid \mathbf{x}_i)$.

Ideally, one would infer the exact latent distribution $p(\mathbf{z}_i \mid \mathbf{x}_i) = p(\mathbf{x}_i \mid \mathbf{z}_i)p(\mathbf{z}_i)/p(\mathbf{x}_i)$. However, this inference problem is computationally intractable in general. In practice, encoders of VAEs are therefore trained to infer a particular approximate distribution, denoted $q_\phi(\mathbf{z}_i \mid \mathbf{x}_i)$ (where $\phi$ denotes encoder network weights), known as a variational distribution. This training attempts to find values of $\phi$ that make $q_\phi(\mathbf{z}_i \mid \mathbf{x}_i)$ as close as possible to the true posterior $p(\mathbf{z}_i \mid \mathbf{x}_i)$. Through our choice of the approximating distribution $q$, we can introduce an "information bottleneck" into the encoder, such that latent representations fail to represent at least some aspects of input data items. This loss of information is a desirable outcome, because the latent variables are then forced to find meaningful abstractions in the data. Below, we will see there is a precise way of controlling the amount of information lost by introducing a scalar parameter (denoted $\beta$) inside the training objective.

The final stage of a VAE is a decoder. Like the encoder, the decoder is typically implemented as a conventional neural network. Its input is a sample $\mathbf{z}_i$ produced by the sampler, and its output is an (approximate) reconstruction of data item $\mathbf{x}_i$ based on the sample. Because latent variables have probability distributions, reconstructions also have distributions. If the sampler produces multiple samples of latent variable $\mathbf{z}_i$, then the decoder can decode each sample, thereby producing multiple data item reconstructions. That is, samplers and decoders can be used to obtain a histogram estimating distribution $p_\theta(\mathbf{x}_i \mid \mathbf{z}_i)$ (where $\theta$ denotes decoder network weights). Reconstructions tend to be imperfect due to the loss of information in the latent variable (and because $\mathbf{z}_i$ is a noisy sample from approximate distribution $q_\phi(\mathbf{z}_i \mid \mathbf{x}_i)$).

Intuitively, VAEs are closely related to RDT (these intuitions can be made mathematically precise; see Alemi et al., 2017, 2018; Ballé et al., 2016; Burgess et al., 2018). Channels in RDT correspond to VAEs. Messages or signals (e.g., sensory stimuli) in RDT correspond to input data items in VAEs. Codes (e.g., memories) in RDT correspond to latent representations in VAEs, and capacity limits in RDT correspond to constraints on latent representations. Distortion functions (penalizing differences between input signals and their reconstructions) in RDT correspond to (at least one term in) objective functions in VAEs.

During training, VAEs adjust their weight values to minimize the following loss or objective function:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = -\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] + \beta\, D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})).$$

A mathematical derivation and explanation of this equation can be found in Kingma and Welling (2013). Intuitively, the equation can be understood as follows.

The right side of the equation has two terms. The first term is the (negative of the) expected log probability of input $\mathbf{x}$ given latent variable $\mathbf{z}$. It is often referred to as the "reconstruction error" because, in practice, it often ends up being some measure of error between the input and its approximate reconstruction. In VAEs, it plays a role analogous to the distortion function in RDT. If, for example, each input feature is assumed to have a Gaussian distribution given a value for $\mathbf{z}$, then $\log p_\theta(\mathbf{x} \mid \mathbf{z})$ is proportional to the sum of square-error between the true input feature values and their estimated or reconstructed values.

The second term is the Kullback-Leibler divergence (a measure of the difference between two probability distributions) between the posterior distribution of latent representation $\mathbf{z}$ (after observing $\mathbf{x}$), denoted $q_\phi(\mathbf{z} \mid \mathbf{x})$, and its prior distribution, denoted $p(\mathbf{z})$ (chosen by the experimenter; in our simulations we set $p(z_{ij})$ to be a Gaussian distribution with a mean of zero and variance of one). This term acts as a regularizer that constrains the latent representation acquired during training by biasing the posterior distribution of this representation toward its prior distribution. In VAEs, this term plays a role analogous to the capacity constraint in RDT because it limits or constrains the latent representations (or codes) acquired by a VAE.

For VAEs, the coefficient $\beta$ in the objective function is set to one. $\beta$-VAEs are a variant of VAEs in which $\beta$ can be set to any nonnegative value (Higgins et al., 2017). Both VAEs and $\beta$-VAEs have previously been characterized using RDT. As mentioned above, this is accomplished by treating the first term of the objective function as a measure of distortion and the second term as related to capacity (e.g., Alemi et al., 2017, 2018; Ballé et al., 2016; Burgess et al., 2018). The capacity of $\beta$-VAEs can be indirectly influenced through the choice of $\beta$. When $\beta$ is set to a small value, $\beta$-VAEs are relatively unconstrained by the regularizer and capacity tends to be large. In this case, $\beta$-VAEs acquire latent representations that codify the fine details of data items. In contrast, when β is large, β-VAEs are more constrained by the regularizer and capacity tends to be small. These β-VAEs acquire latent representations that are relatively constrained, compressed, or abstract. After a β-VAE is trained, it is useful to quantify its actual capacity (using, e.g., the formula in Alemi et al. (2018)).

## Extended Model Architecture

Some of the simulations discussed below use an extended model architecture containing two modules: a β-VAE and a decision network (see Figure 4). As described above, the β-VAE assigns a value to latent variable $\mathbf{z}_i$ based on stimulus $\mathbf{x}_i$ which can be regarded as a memory code. The decision network takes as input this memory code and, optionally, a task-related probe image. It outputs a decision variable. In a change-detection task, for example, the input to the β-VAE is a target image, the input to the decision network is the β-VAE's latent code of the target image and a probe image, and the decision network's output is the probability that the target and probe images depict different objects.

A critical aspect of the extended architecture is the link between the two modules. Because the β-VAE's memory code is an input to the decision network, error information flows during training through the decision network into the β-VAE (via the backpropagation process used to train neural networks; Rumelhart, Hinton, & Williams, 1986). This allows task-based decision errors to influence the acquisition of β-VAE memory codes. The objective function minimized by the extended model during training has three terms which can be weighted differently to achieve different trade-offs, corresponding to (a) the distortion of the β-VAE's image reconstruction, (b) the information capacity of the β-VAE's memory code, and (c) the decision error. See the Appendix for additional implementation details about the model used for each task.

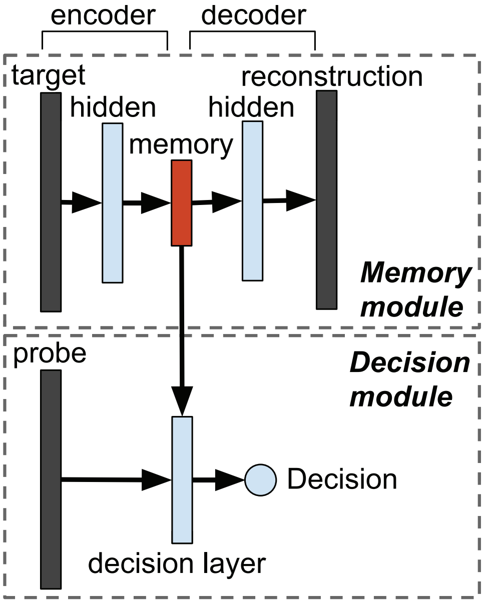

*Figure 4.* Schematic of the extended model architecture. Dark gray boxes represent a vector of pixel values, while other boxes represent layers (or a set of layers) in a network. The layer representing the memory code is in red. See the online article for the color version of this figure.

As discussed above, the information that gets stored or transmitted in RDT models of behavior can vary widely depending on an organism's prior knowledge and goals. Accordingly, researchers can manipulate during training what kind of information is encoded in memory in several ways. For example, a researcher may seek a model that learns a memory code that "cares" a lot about making accurate decisions but is less concerned about remembering all image pixels accurately. Technically, this can be achieved by weighting the distortion of the β-VAE's reconstruction in a model's objective function by an especially small value. Another way to manipulate the contents of memory is to vary the decision task. For example, consider a case where a model is trained only to minimize decision loss. If the error for the decision network depends solely on stimulus dimension one, and is independent of dimension two, then the β-VAE will learn a memory code that contains information about the first stimulus dimension only. Finally, memory contents should change as the prior distribution over stimulus inputs changes. For example, if some stimuli appear more frequently during training, then the acquired memory code will tend to store more information about these stimuli at the expense of others.

## Data Sets

The simulations reported below used three sets of stimuli: artificial potted plants, multiplant images, and natural images of fruits.

**Artificial potted plants.** This data set consisted of the same stimuli used in Bates et al. (2019). The set consists of images of an artificial plant-like object (rendered via Blender, a 3D graphics program), which we varied along two dimensions: leaf width and leaf angle (see Figure 5). Images were converted to gray scale, and down-sampled and cropped to a size of 120 × 120 pixels. The stimulus space was discretized to 100 values along each dimension for a total of 10,000 unique stimuli.

**Plants set-size stimuli.** For the set-size task discussed below, we created a data set where each image had between one and six potted plant objects taken from the artificial potted plants data set (see Figure 8 for examples). All images were gray-scale and 300 × 300 pixels. For experiments below, we kept leaf width fixed and varied leaf angle.

**Natural images of fruits.** For our simulations with natural images, we used the Fruits-360 database (Mureşan & Oltean, 2018) which is comprised of photographs of individual fruits on a white background. We chose a subset of the image classes to train on, specifically apples, tomatoes, and bananas. We augmented the data set by randomly zooming and cropping images, as well as randomly flipping images.

# Computer Simulations: Results

In this section, we present the results of a set of computer simulations demonstrating qualitative agreement between our neural network models, optimal data compression as defined by RDT, and people's behaviors on perceptual and memory tasks. To test

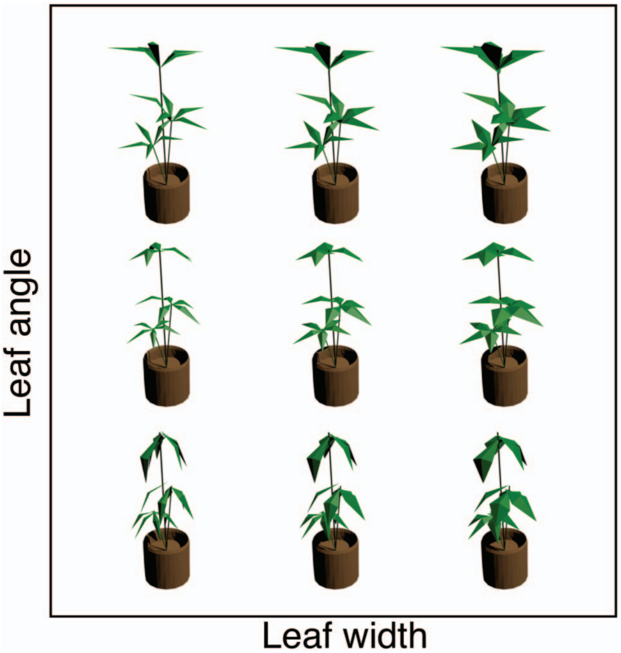

*Figure 5.* Two-dimensional stimulus space of artificial potted plants. Leaf width varies along the horizontal axis, while leaf angle varies along the vertical axis. From "Adaptive Allocation of Human Visual Working Memory Capacity During Statistical and Categorical Learning," by C. J. Bates, R. A. Lerch, C. R. Sims, and Robert A. Jacobs, 2019, *Journal of Vision, 19*(2), p. 11. Adapted with permission. See the online article for the color version of this figure.

ideas about the role of efficient data compression in perception and memory, it is not sufficient to use simple models that make strong, a priori assumptions about the stimulus dimensions that people represent. Rather, we need to build models that match the level of algorithmic sophistication of people who, for example, perceive and remember aspects of their visual environments starting from photoreceptor activities (akin to pixel values).

## Fundamental Phenomena of Efficient Data Compression

This subsection focuses on the fundamental phenomena of efficient data compression that were described above. The simulations reported here are presented in an order roughly following the panels of Figure 1.

**Varying capacity.** As will be discussed below (see the Efficient Compression Over Time and Reasons for Multiple Memory Systems section), experimental evidence indicates that the average information content of perceptual traces is large and thus these traces can represent the fine details of sensory stimuli, whereas the information content of mnemonic traces is smaller meaning that these traces tend to represent coarse-scale abstractions of stimuli. To demonstrate that our models are consistent with this property, we trained β-VAEs on the artificial potted plants data set at different capacities. The top, middle, and bottom rows of Figure 6 show the results for networks with low, medium, and high capacities, respectively. The leftmost image in each row is an instance of a data item, and the remaining images in a row are samples of a β-VAE's reconstructions. Networks with low capacity tended to produce blurry, averaged reconstructions. In fact, the reconstructions of the network with the lowest (near zero) capacity were nearly identical to the average image over the entire data set. This result was expected because the average of the data set is the value that minimizes the expected reconstruction error. In contrast, networks with high capacity tended to produce crisp, accurate reconstructions. The reconstructions of the network with the highest capacity were nearly veridical copies of each data item.

**Varying the stimulus distribution.** If people are efficient but capacity-limited, their performance in memory tasks should depend on the prior or stimulus distribution (see the discussion of the prior knowledge principle). In the simple case of a Gaussian versus a uniform distribution over stimuli, RDT predicts that (a) performance should be worse for the uniform distribution (since all stimuli are equally important and thus capacity will be spread more thinly across the stimulus space), (b) memory representations should be systematically biased toward the mean of a nonuniform stimulus distribution, with bias increasing for stimuli farther from the mean, and (c) memory representations should become less precise and more biased toward the mean as capacity decreases. Together, Bates et al. (2019; discussed above) and Huttenlocher, Hedges, and Vevea (2000) provided empirical support for these predictions.

As above, we trained β-VAEs on the plants data set at different capacities. In the uniform condition, training data items were drawn from a uniform distribution across both leaf-width and leaf-angle stimulus dimensions. In the Gaussian condition, items were drawn from a uniform distribution over leaf angle and a Gaussian distribution ($M = 50$, $SD = 10$) over leaf width. Because leaf width is the dimension on which the two conditions use different distributions, it is the "relevant" dimension for comparing performances across the two conditions.

Because the model's memory does not explicitly encode leaf width or leaf angle, we use an indirect measure to assess what the model has stored about these dimensions and evaluate our prediction that its representations should be systematically biased. Specifically, we use the decoder of the β-VAE to produce reconstructions from the memories, and then compare these reconstructed images to other images in the data set using pixel-by-pixel correlations as a distance metric. When a reconstructed image is very accurate, it should be highly correlated with the image it was reconstructing, and it should be less highly correlated with other images in the data set. On the other hand, if memory codes are biased toward the mean, the reconstructed image may be most similar to, and thus most highly correlated with, a different image, one which is closer to the mean of the prior. For instance, if the model observes a plant with leaf-width equal to 30, and the prior is Gaussian with mean 50, the image reconstructed from memory may be most similar to, say, images of plants with leaf-width 35, which are closer to the mean of the prior.

Results of this analysis are shown in Figure 7. In each graph of this figure, the horizontal axis plots the value along the leaf-width dimension (i.e., the relevant dimension). Each vertical colored dashed line demarcates a value of leaf width that was observed by the model. The solid curve of matching color represents the correlation values between the resulting reconstructions and other plant images in the data set. For example, consider the purple lines. To produce the purple curve in each panel, we first input plant

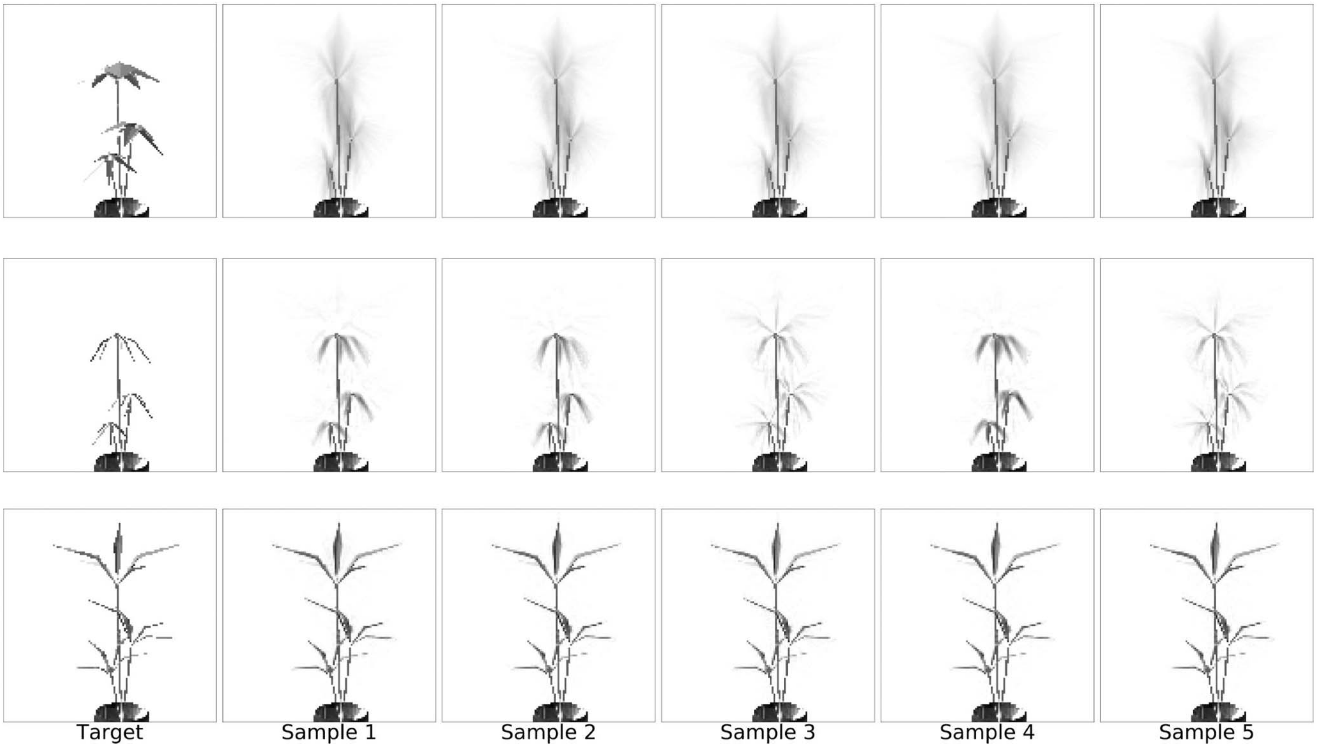

*Figure 6.* β-VAE reconstructions at three different capacities: $\beta = 10$ (top row; network capacity = 0.01 nats), $\beta = 1$ (middle row; network capacity = 2.6 nats), and $\beta = 0.1$ (bottom row; network capacity = 31.6 nats). Networks were trained using a uniform prior or stimulus distribution over both leaf-width and leaf-angle stimulus dimensions.

images of leaf-width 40 (and each of the possible leaf angles) to the β-VAE. Then, we produced reconstructions for each of those images. Next, we compared each reconstructed image to all 10,000 images in the data set to get 10,000 correlation values. Finally, we marginalized (averaged) over leaf-angle (i.e., the irrelevant dimension), resulting in a single, averaged correlation value for each value of leaf-width that the reconstruction was compared to.[3]

The procedure just described was carried out with four different models, corresponding to the four panels of Figure 7. The graphs in the top and bottom rows correspond to β-VAEs trained on Gaussian and uniform prior distributions, respectively, and the graphs in the left and right columns correspond to β-VAEs with high- and low-capacity, respectively.

As expected, β-VAE image reconstructions exhibited significant bias toward the overall leaf-width mean (leaf-width $M = 50$) in the Gaussian condition, especially when using low-capacity networks. In contrast, no bias was evident in the uniform condition with networks of either low or high capacity. Moreover, reconstructions were more biased for leaf-width values farther from the leaf-width mean in the Gaussian condition. We also examined the reconstruction errors and found errors to be higher in the uniform condition than in the Gaussian condition with both low- and high-capacity networks (when equated for measured capacity), as predicted by RDT. Finally, the relatively flatter curves in the right-most panels indicate that reconstructions were less precise with low-capacity networks.

**Set-size effects.** Set-size effects—decreases in memory performance with increases in the number of to-be-remembered items—have been reported in numerous VWM experiments and are predicted by RDT (see Figure 1c). To test if our models can account for these effects, we created a data set consisting of images with varying numbers of plants (see leftmost column of Figure 8). For this set of simulations, we used three versions of the extended model architecture described above, where versions differed in the capacities of their β-VAEs ranging from low to high (β values of 0.1, 0.01, and 0.003). Each model was trained separately on each set-size, which varied from one to six. The decision network of each model was trained to predict the leaf angle (as above), which varied from 0 to 99. To measure precision, we calculated the mean squared error between the prediction and true value.

Examples of models' performances are shown in Figure 8. The leftmost column shows displays with set sizes of two and six. The remaining columns show sample image reconstructions. To produce the top row, we trained a model on images belonging to set-size 2 with $\beta = 0.003$. To produce the bottom row, we trained a model on images belonging to set-size 6 with $\beta = 0.01$. The capacities of the two networks were measured to be approximately equal. As expected, reconstructions of set-size 2 are crisp and accurate because all the capacity can be allocated to just two objects, whereas reconstructions of the large set-size are blurrier and less accurate because the same capacity is "spread" thinly across multiple plants.

[3] Since reconstructions from the β-VAE are noisy samples, we repeated the analysis multiple times and averaged the results.

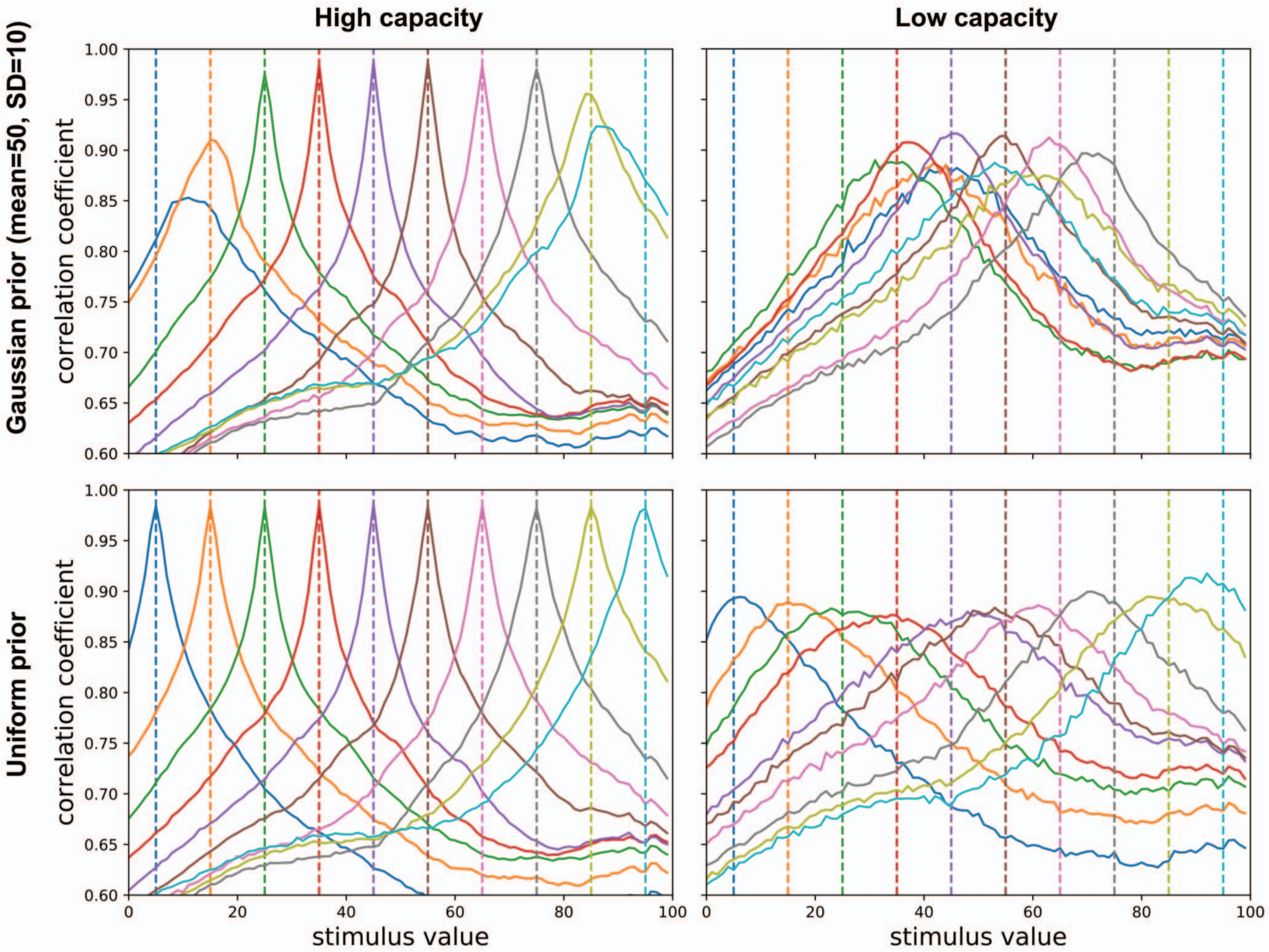

*Figure 7.* Effect of the prior or stimulus distribution on β-VAE reconstructions. See text for details. See the online article for the color version of this figure.

The key results are shown in Figure 9 which shows estimated "rate-distortion" curves for each set-size.[4] The horizontal axis plots calculated rates or capacities for the β-VAE portions of models. The three points defining each curve correspond to the three versions of the β-VAEs ranging from low to high capacity. The vertical axis plots distortion or mean squared error in responses. Colors of curves indicate the set-size condition.

Consistent with set-size effects, mean squared error increased monotonically as a function of set-size. This is evidenced by the fact that curves shift upward with set-size, implying that response precision decreases with set-size for a fixed rate (as in Figure 1c). Critically, the decrease in precision naturally emerges from the introduction of capacity limits, and requires no additional assumptions. When there is less capacity available, a model is forced to "spread" its resources across the to-be-remembered plants in a display to minimize overall error.

**Varying the objective function.** RDT predicts that the contents of memory vary based on the nature of a task (task-dependency principle). Here, we demonstrate that our model shares this property. We trained an extended model architecture to detect changes between a target image and a randomly drawn probe image in two conditions. In one condition, the model was penalized solely for errors in leaf width, and in the other condition it was penalized solely for errors in leaf angle. For example, if the target image had leaf-width = 50 and leaf-angle = 50, but the probe had leaf-width = 60 and leaf-angle = 50, then the model would be penalized in the former condition, but not the latter.[5]

As expected, latent representations of models' β-VAEs contained little information about the unpenalized stimulus dimension in each condition. This result is illustrated in Figure 10. The left and right columns show results for the leaf width-relevant and leaf angle-relevant conditions, respectively. Within a column, the left image of a pair is the target image and the right image is a β-VAE's reconstruction. One can see that the image reconstruc-

[4] Rate-distortion curves are commonplace in the engineering literature.

[5] In these simulations, objective functions were set so that reconstruction error was ignored and so that decision error depended only on leaf width (first condition) or leaf angle (second condition).

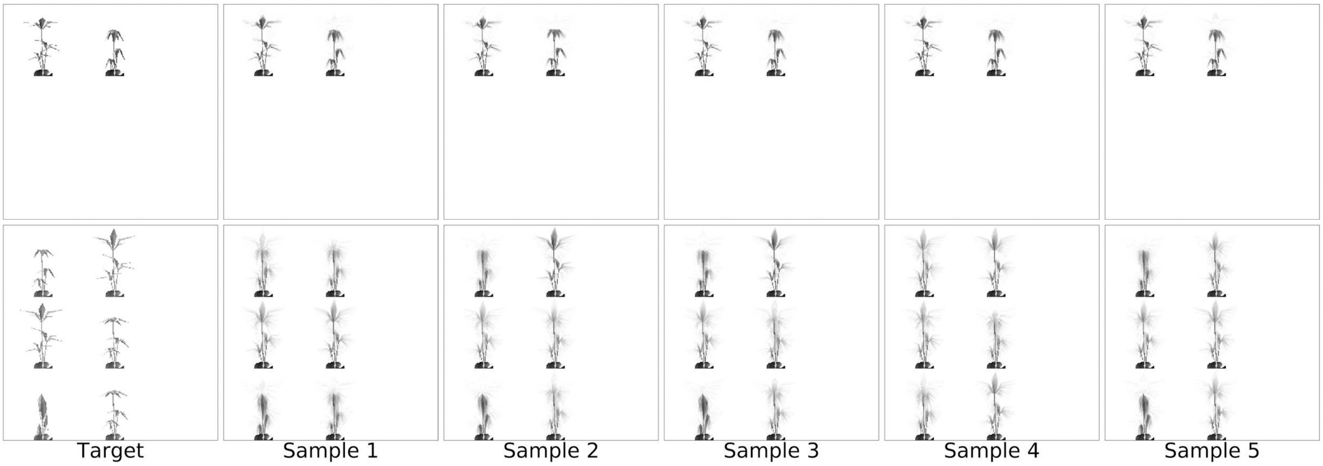

*Figure 8.* Sample reconstructions for plants set-size data set. The top row is an example of set-size 2, whereas the bottom row is set-size 6. Each row shows the results from a separately trained β-VAE, but the two networks had approximately the same capacity. As a result, reconstructions are less accurate with the larger set size.

tions most closely match the target images along the leaf-width dimension in the left column, and most closely match the target images along the leaf-angle dimension in the right column.

## Modeling Categorical Bias in High-Dimensional Spaces

In this subsection, we demonstrate that the categorical bias found with exact RDT methods in low-dimensional stimulus spaces can also be found with our models in high-dimensional spaces. The two sets of simulations discussed here used the artificial potted plants and the natural images of fruits data sets, respectively.

**Artificial potted plants.** Extended model architectures were trained in two conditions. In the categorical condition, images were sampled from a uniform distribution across the stimulus space (both leaf width and angle). A model was trained using an objective function that combined categorical and pixel-reconstruction errors. In regard to the categorical error, the decision network of a model was trained to detect a change between target and probe plants, but errors were only nonzero when the network responded "same" but the target and probe belonged to different categories or when the network responded "different" but the target and probe belonged to the same category. We set the category boundary along a single dimension—leaf width—at the value of 50. If the reconstruction error was always set exactly to zero and the β-VAE portion of a model was trained only with respect to the categorical error, then the model would be incapable of learning perfect reconstructions. Therefore, we set the weight on the reconstruction error term to a small positive value so that the categorical error would dominate for low-capacity models, but high capacity models could still learn excellent reconstructions.

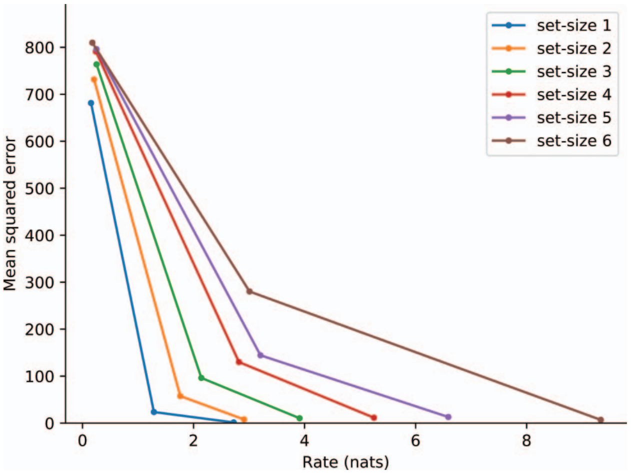

*Figure 9.* Estimated rate-distortion curves for each stimulus set-size. Error is based on Euclidean distance between actual and predicted leaf angles. See the online article for the color version of this figure.

In the modal condition, a model was trained only with respect to reconstruction error, but we manipulated the prior or stimulus distribution such that there were two modes corresponding to separate categories. Stimuli were restricted to vary along a single dimension, leaf width, keeping the other dimension fixed. The stimulus distribution was a mixture of two 1-D Gaussian distributions with means set to 0 and 100 and with standard deviations set to 15.

In both categorical and modal conditions, we predicted that the latent representations of low-capacity models would store little more than the category of target images. As illustrated in Figure 11, the results confirmed this prediction. In this figure, the leaf-width of a target plant increases from left to right. Critically, image reconstructions are similar when leaf widths are to the left of the category boundary, resembling one of the category prototypes, or to the right of the category boundary, resembling the other category prototype. However, they change dramatically at the category boundary.

We also verified that the extent of this categorical bias increased gradually as the capacity of a model decreased. This is illustrated in Figure 12. The graphs in this figure were plotted using the same method as those in Figure 7. For example, the red solid line plots

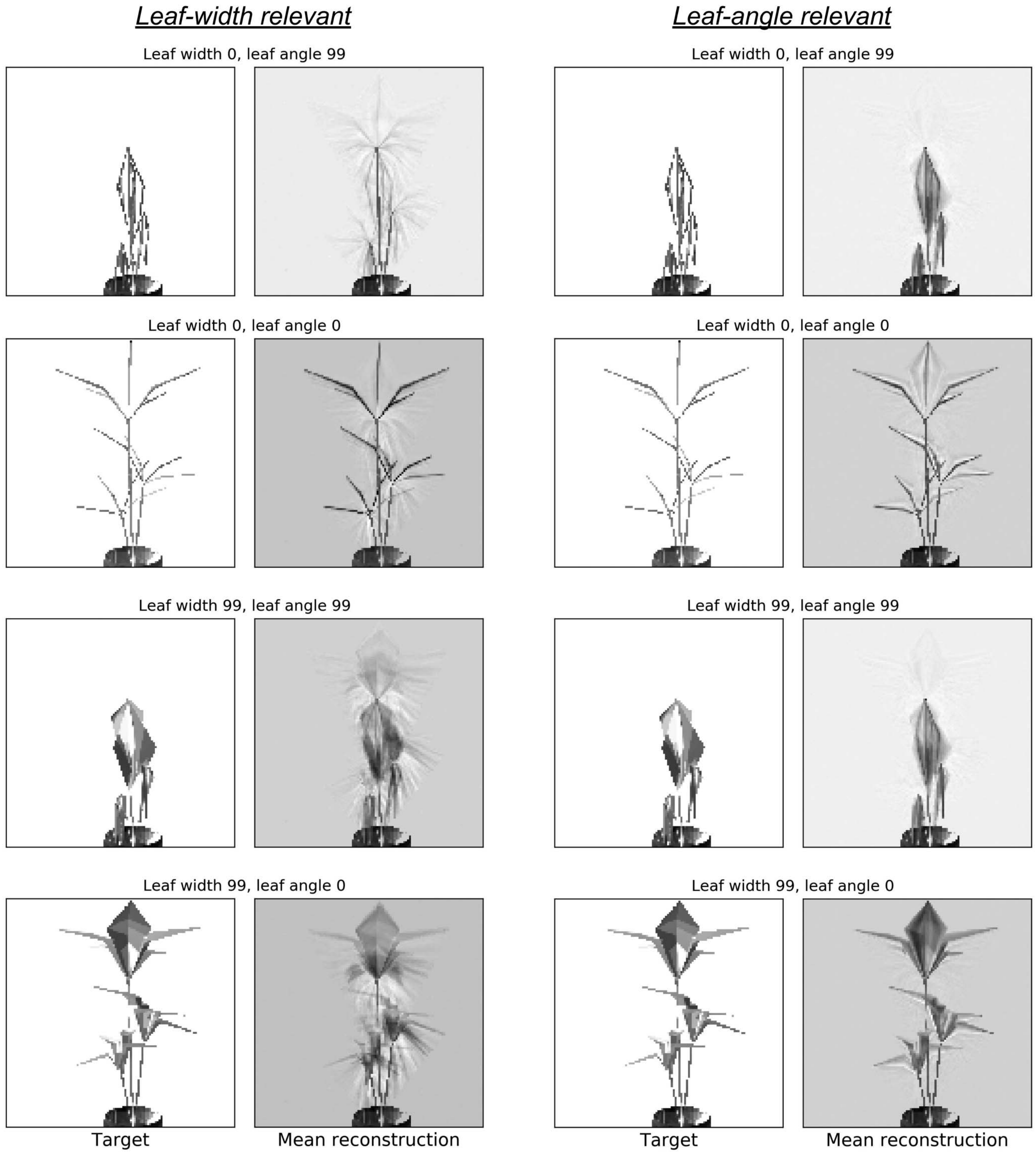
Leaf-width relevant
Leaf-angle relevant
Leaf width 0, leaf angle 99
Leaf width 0, leaf angle 99
Leaf width 0, leaf angle 0
Leaf width 0, leaf angle 0
Leaf width 99, leaf angle 99
Leaf width 99, leaf angle 99
Leaf width 99, leaf angle 0
Leaf width 99, leaf angle 0
Target
Mean reconstruction
Target
Mean reconstruction

*Figure 10.* Image reconstructions of target images when models were penalized for error in leaf-width (left column) or for error in leaf-angle (right column).

the correlation curve when a target plant had a leaf width of 35 and comparison plants had leaf widths indicated by the horizontal axis. The left, middle, and right graphs are for high-, medium- and low-capacity models, respectively. Clearly, target reconstructions of high-capacity models are most similar to true target leaf-width values, but as capacity decreases, correlation curves corresponding to all target values within a category collapse onto each other. In other words, for low-capacity models, there are essentially two

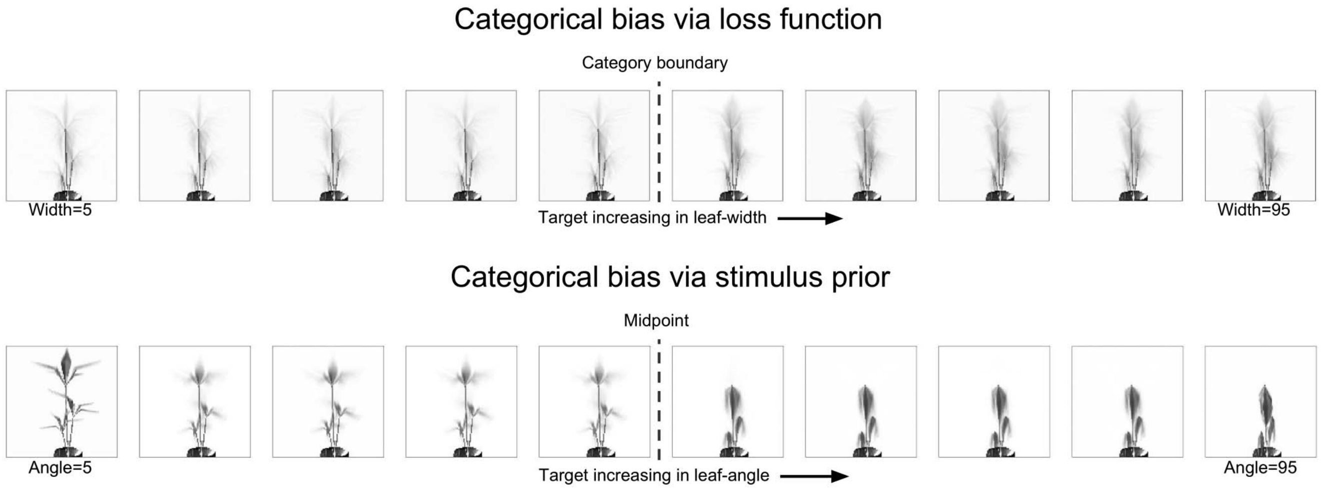
Categorical bias via loss function
Category boundary
Width=5
Target increasing in leaf-width
Width=95
Categorical bias via stimulus prior
Midpoint
Angle=5
Target increasing in leaf-angle
Angle=95

*Figure 11.* Target image reconstructions from low-capacity models trained in the categorical (top) and modal (bottom) conditions. Leaf widths (top) and leaf angles (bottom) of target plants increase from left to right. In both conditions, similar results were found by varying the other leaf dimension.

possible reconstructions corresponding to the two category prototypes.

**Natural images.** In this experiment, we only examine categorical bias driven by the loss function, since we are unable to precisely manipulate the modality of the prior distribution for a compilation of natural images, like we did with the artificial images. Accordingly, we weighted the decision loss heavily compared to the reconstruction loss in order to ensure that the target object category was accurately remembered as opposed to remembering each pixel value equally well.

To measure categorical bias on the natural images data set, we examined image reconstructions (see Figure 13) and performed principal components analysis (PCA) on the memory representations (see Figure 14). We can infer from the reconstructions that the memories became less diagnostic as to, for example, which variety of apple was seen or what angle it was seen from. 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 PCA analysis demonstrates that at low capacity, all memory vectors within a particular class were highly similar to each other, and clearly distinct from all other classes; whereas at high capacity, memories from the same class were more distinct from each other and had more overlap with other classes.[6]

## Interim Summary

To this point in the article, we have argued that efficient data compression shapes biological perception and perceptual memory in many of the same ways that it shapes engineered systems. We have stated three principles—the capacity-limited, prior knowledge, and task dependency principles—that follow directly from RDT, a primary tool that engineers use to design and analyze capacity-limited machines, and shown how these principles provide accounts for many important behavioral phenomena and experimental results. We also presented an extended deep neural network architecture that approximately implements RDT in high-dimensional spaces. Importantly, the architecture is trained end-to-end, operating on raw perceptual input as opposed to features selected by an investigator. We demonstrated that the architecture accounts for several errors and biases in human perception and perceptual memory, including categorical biases.

We believe that this work establishes a firm foundation for the hypothesis that principles of efficient data compression can serve as unifying principles accounting for many aspects of perception and perceptual memory. In the remainder of this article, we attempt to expand this foundation in new directions by offering conjectures or speculations about how efficient compression may play a role in other areas. The next section examines efficient compression over time. An important aspect of this work is that it motivates the need for multiple memory systems operating at multiple time scales. The following section applies efficient compression to the study of perceptual attention. Here, it is argued that efficient compression can account for several attentional phenomena including “pop out” effects.

## Efficient Compression Over Time and Reasons for Multiple Memory Systems

In sections above, we demonstrated cases in which categorical bias increases as compression increases. For instance, in the case of an autoencoder trained with photographs of fruits, we assumed

[6] Although the principle-component 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-component 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 latent variables tends to decrease (because this allows more information to be stored), and thus two points that are a distance $d$ apart in principle-component space are at least as distinguishable at high capacity compared to low capacity.

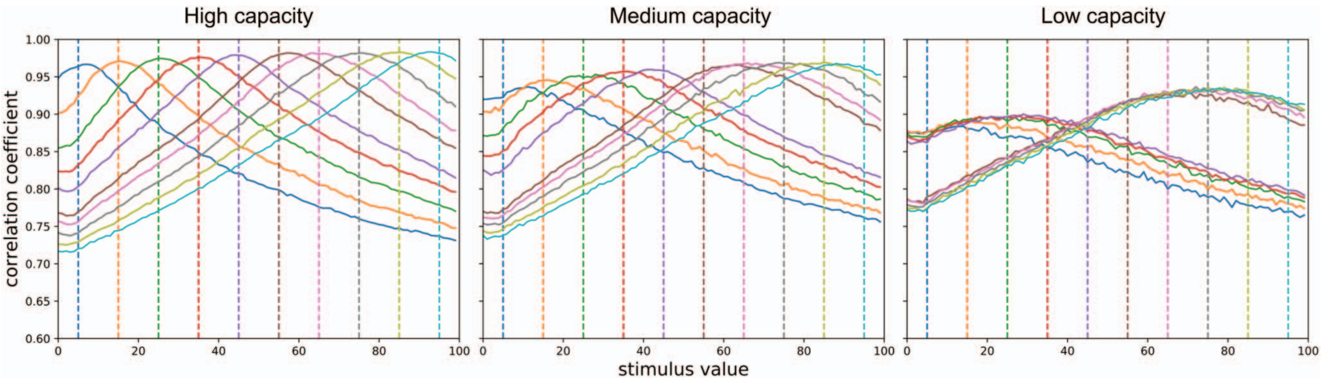

*Figure 12.* Pixel-wise correlations between reconstructions of target and comparison images as a function of a model's capacity. See the online article for the color version of this figure.

a loss function *d* that combined reconstruction and categorical errors and showed that the network's memories contained high visual detail at high capacity but only retained category information (apple vs. banana vs. tomato) at low capacity. Here we demonstrate how these results can further our understanding of memory systems in people. In brief, we conjecture that people's longer-term memory representations (older traces) correspond to highly compressed codes, whereas their shorter-term memory representations (younger traces) correspond to less-compressed codes. We argue that this conjecture is both consistent with a large body of empirical evidence and predicted by an extension of the principles of efficient compression over the dimension of time.

Conventional RDT, presented above, is atemporal in nature. It studies the problem of transmitting information from point A to point B without consideration of what happens to that information after it is transmitted. Practical applications, however, must consider the problem of storage. For example, people cannot simultaneously store all of their life experiences in their full perceptual detail. They therefore must choose which information is maintained and for how long. This problem motivates an extension to our study of efficient compression to the case in which representations are optimized over time. In this section, we sketch out this extension in broad theoretical terms but leave to future work many of its details. We start by presenting an additional principle that pertains to the problem of deciding what transmitted information to keep in storage over time, along with empirical evidence for its implementation in human perceptual memory systems. We then provide an extension to RDT which formalizes this principle.

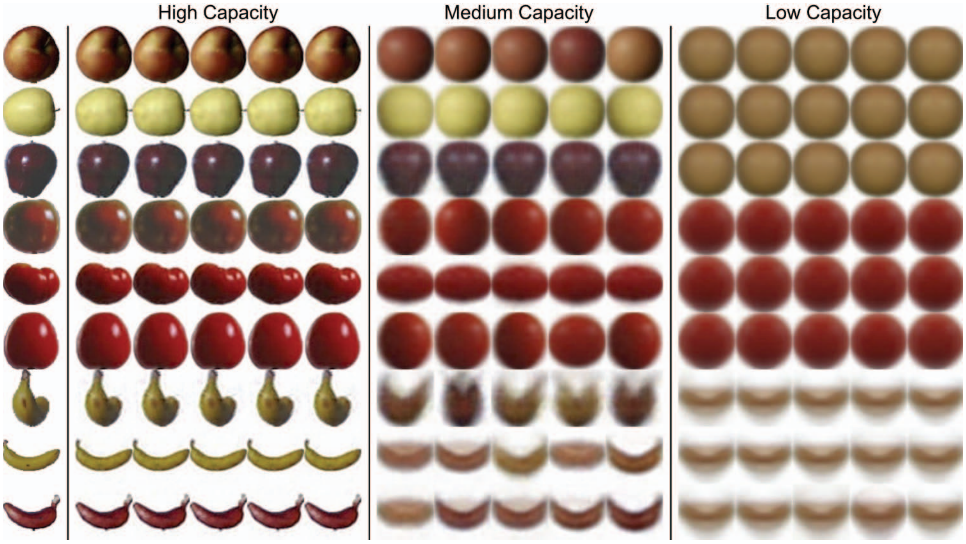

*Figure 13.* Reconstructions at different capacities. The left-most column contains the original images, while the remaining columns (separated by vertical dark lines) contain reconstructions at three different capacities. Five random samples are drawn for each combination of capacity level and original image. A high-capacity network exhibits high-fidelity memory for the inputs, while at lower capacities, network memories become more categorical and less certain of specific visual details, although they can still reliably distinguish the three chosen classes (apples, tomatoes, bananas). See the online article for the color version of this figure.

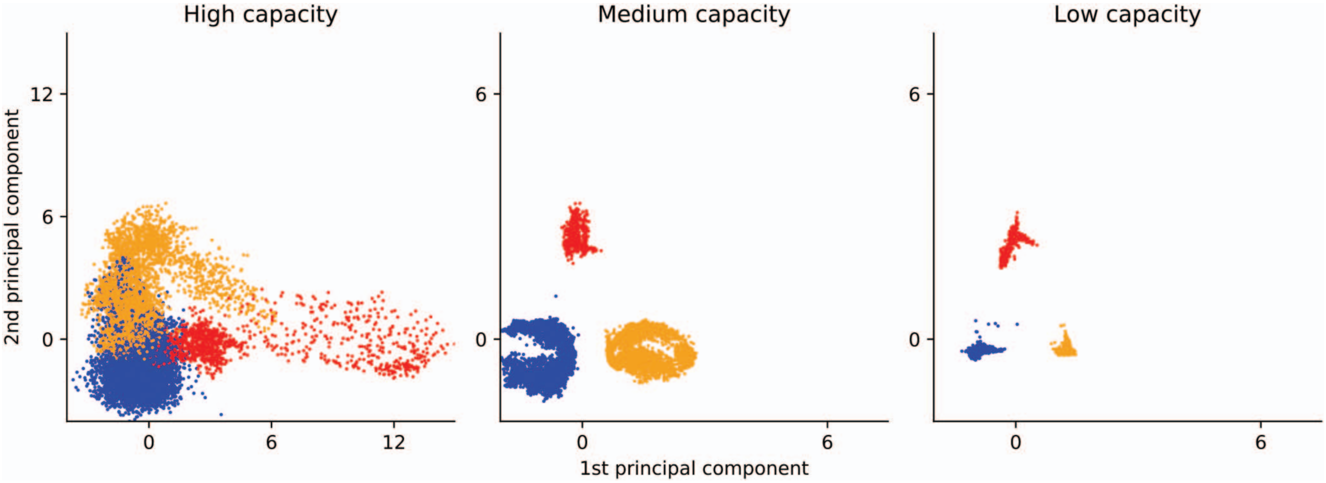

*Figure 14.* Principal components analysis of network memories. A low capacity network remembers little more than the category of the input, while a higher-capacity network remembers more visual detail, as indicated by the larger spread of points. The colors blue, red, and orange correspond to the classes "apple," "banana," and "tomato," respectively. Note the difference in axis limits between plots. See the online article for the color version of this figure.

## Information Decay Continuum Principle

The information decay continuum principle states that the average information content of individual memory traces tends to decline over time, and the rate of this decay is roughly monotonic in (e.g., proportional to) its current information content. Thus, the timeline of a hypothetical memory trace would look something like the following. 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 sensory detail about the stimulus, are even more categorical and abstract, and decay more slowly than those of sensory memory. Finally, LTM representations contain the least amount of fine detail about the originally observed stimulus, are the most categorical and abstract, and decay slowest.

We account for the decay of individual traces by hypothesizing that memory is biased toward representing recent information because recent information tends to be more task-relevant (J. R. Anderson, 1991; J. R. Anderson & Schooler, 1991). Consequently, memory adaptively reallocates bits over time such that fewer resources are devoted to older memory 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 memory resources that can then be used to encode new information. An exception occurs when information in a memory trace is actively rehearsed or refreshed (A. Baddeley, 2003; Ricker & Cowan, 2010). Because rehearsal of older traces indicates that the information in those traces remains task-relevant, those traces are not recoded using fewer resources. For further intuition, consider the analogy of trying to make room on a full computer hard drive. It would be efficient to first remove large video files before worrying about much smaller text files. Moreover, one could "recode" a large video file by replacing it with a text file containing a summary of its contents. Because LTM traces are highly abstract and summary in nature (similar to text files), they can be retained cheaply and many of these traces can be accrued over time. By contrast, sensory and working memory traces are more detailed (more similar to video files), and therefore not as many traces can be kept concurrently.

The optimal recoding behavior of a system will depend on the precise form of the recency bias. At one extreme, a trace may be equally likely to be useful over a large range of delays (i.e., no recency bias), in which case all memories will be highly compressed, even soon after stimulus offset. At another extreme, the average usefulness of a memory could drop precipitously over time. In this case, all memories would be high resolution, but have a short "shelf-life", being wiped away quickly with no recoding over time. Biological memory systems are likely to fall between these two extremes, with perceptual memory traces that decay linearly, exponentially, or according to a power law. Although we leave further details to future work, we suspect that for a wide range of candidate functional forms, the same qualitative results will obtain: Older traces will tend to be more compressed and abstract and will be maintained longer, whereas newer traces will tend to be less compressed and more detailed and will be maintained for less time.

The psychology literature has attributed memory phenomena to multiple distinct systems (e.g., sensory memory, working or short-term perceptual memory, long-term perceptual memory, etc.). The information continuum decay principle does not directly predict separate systems, and so it is important to address how this division could be explained. We speculate on two possible explanations. First, to be optimal in its bit reallocation over time, a compressed memory code would potentially need to change at every moment in time, and this would demand an infinitely large set of potential codes. Because neural substrates are finite, it is unlikely that they could support an unbounded set of codes. Therefore, the (relatively small) number of memory systems that exist may strike a balance between implementational costs of neural hardware and optimal bit reallocation over time. Such a "bag" of representations could provide a basis for an approximate solution to the "recoding" problem in which the optimal code at each point in time is approximated by a mixture or combination of codes. If traces in each system degrade according to their mean information contents (e.g., sensory memory degrades quickly, short-term memory [STM] degrades more slowly, etc.), then the overall abstractness of a memory will increase over time. Thus, a discrete mixture of representations could approximately implement the information decay continuum principle. Alternatively, it may be possible to recode information over time within the same neural circuits.

However, there is a second motivation for multiple memory systems to exist. As perceptual memory tends to share neural substrates with other perceptual functions, a division into multiple systems may be in part due to the demands placed on perception more broadly. For example, it is likely advantageous to possess a hierarchy of representations extending from perceptually detailed to more abstract, because agents then possess a wide range of representations suitable for a wide range of tasks. For instance, we know that animals possess both perceptually detailed representations of recent stimuli needed, for example, for planning eye movements, as well as more abstract or conceptual representations needed to, say, decide what kind of fruit one is looking at. Because the "best" representation is highly task-dependent, meaning that a wide variety of representations must be maintained, the best overall choice of representational architecture may be complex, depending on at least the two partially competing demands just presented.

The questions raised here, and the speculative answers that we have provided, deserve extensive future investigation. Although this investigation is in its early stages, we believe that there is substantial empirical evidence for the information decay continuum principle, including trace decay both within and across memory systems.

Evidence for decay within systems can be found in many experiments. Sperling (1960) reported that visual representations of individual displays decayed in iconic memory within a few hundred milliseconds after stimulus offset. Luck (2008) reviewed evidence that VWM representations of individual displays decay or drift over a period of seconds. Konkle, Brady, Alvarez, and Oliva (2010) reported that power law fits to their experimental data suggest that $d'$ (a measure of discriminability) on a visual LTM recognition task would be above 1.0 after a day following study, would fall below 1.0 after a month, and would be below 0.6 after a year.

There is also evidence for decay across memory systems. Experimental findings indicate that nearly all systems are influenced by a mix of perceptual and more categorical factors, though the representations of some systems tend to be relatively more perceptual in nature, whereas the representations of other systems are more categorical. For example, in the auditory domain, A. D. Baddeley (1966b) reported that subjects' performance on a verbal STM task was worse when words were acoustically similar than when they were semantically similar. A. D. Baddeley (1966a) found the opposite results on a verbal LTM task. In this case, performance was worse when words were semantically similar, and performance did not decline when words were acoustically similar. Baddeley's results suggest that STM representations are more perceptual, whereas LTM representations are more categorical or conceptual.

Similar results have been found in the visual domain. 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. Although VWM maintains representations that are somewhat detailed, recent research reveals that these representations are also surprisingly abstract. 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).

As abstract as VWM representations seem to be, visual LTM representations appear to be even more so. Konkle et al. (2010) performed a visual LTM experiment in which subjects viewed 2,800 color images of real-world objects for 3 s each during a study session. Objects belonged to categories, and subjects studied between one and 16 exemplars per category. Following study, subjects 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).

Taken as a whole, the pattern of results described here (and many more results from the scientific literature) strongly support the hypothesized continuum of systems ranging from those with fast decay rates and high perceptual detail to slow decay rates and small amounts of perceptual detail. The time-scales attributed to different systems also support the hypothesized continuum of decay rates. For example, iconic visual memories decay within roughly 100 ms, visual short-term memories decay over a span of seconds, and visual long-term memories decay over much longer time spans.

## Extending RDT

Here, we consider an extension of RDT to the "online" case in which an agent both accrues new information and maintains old information at each moment in time subject to an overall limit on storage capacity. As we demonstrate, a consequence of a limited storage capacity is that less abstract (e.g., perceptual) representations "decay" more rapidly than more abstract (e.g., mnemonic) ones.

Because past experiences may be useful to future behavior, it is desirable to store as much information about them as possible. However, if the brain has limited capacity, then not everything can be permanently maintained. Because recent memories are more likely to be task-relevant (J. R. Anderson, 1991; J. R. Anderson &

Schooler, 1991), the brain should delete information about past experiences to "make room" for new ones. But memories do not have to be completely deleted. Rather, they can be replaced with compact summaries of their contents which may still convey useful information. As a given memory recedes into the past, it may be successively recoded with increasingly abstract summaries, thereby allowing "freed up" memory resources to be used to code new experiences.

This more general problem can be formalized as follows:

$$\begin{aligned} Q^{*} = \underset{\{p(\hat{x}_{0:t} \mid x_{0:t})\}_0^T}{\arg\min} \int_0^T \mathbb{E}_{p(x_{0:t},\, \hat{x}_{0:t})} \; d(x_{0:t}, \hat{x}_{0:t}) \; dt, \\ \text{subject to } I(x_{0:t}; \hat{x}_{0:t}) \leq C \qquad \forall \quad 0 \leq t \leq T. \end{aligned} \tag{3}$$

In this RDT formulation, we have introduced subscripts denoting time because this optimization considers all inputs from the start of time (all past experiences starting from time step zero) to the current time (time step *t*). The random variable $\hat{x}_{0:t}$ represents the brain's entire memory contents at time *t*. As the brain's contents may shift over time (e.g., to recode old memories), we allow a different random variable $\hat{x}_{0:t}$ for each moment in time. The distortion function *d* then measures the error between all the current memory contents $\hat{x}_{0:t}$ and all the corresponding observations up to that point $x_{0:t}$.

Because this optimization problem is computationally intractable, we do not attempt to solve it explicitly. However, if we assume that the distortion function *d* includes a recency bias, as motivated above, then there should be an intuitive result: old memory traces should tend to resemble outputs of an optimal low-capacity channel (i.e., they should tend to be compact summaries), whereas newer traces should tend to resemble outputs of a higher-capacity channel (e.g., they should tend to contain fine-scale perceptual detail). That is, using optimal data compression, it is possible for a capacity-limited system to strike a balance in which the system maintains both fine-scale details about a relatively small number of recent experiences (these experiences are highly likely to be task-relevant) and compact summaries of a larger number of older experiences (which might also prove to be task-relevant).

# Perceptual Attention

## Empirical Evidence for Attention as Data Compression

In the preceding sections we used the perspective of "efficient data compression" to study important aspects of perception and perceptual memory. In this section, we claim that this perspective can also yield insights into perceptual attention. Indeed, researchers have argued that attentional shifts are an adaptation to capacity limits—since people cannot perceive everything of interest in a scene at once, people serially scan a scene over time to extract needed information (Pashler, Johnston, & Ruthruff, 2001). In other words, perception is capacity-limited, and attentional shifts are a strategy for managing this limit. Here, we do not consider all aspects of attention—attention is a notoriously sprawling and unwieldy research domain—but rather focus on two common experimental paradigms.

First, using the "multiple object tracking" (MOT) paradigm, the works of Alvarez and Oliva (2008, 2009) are particularly relevant. In Alvarez and Oliva (2009), subjects were given a primary task of tracking moving objects, and a secondary task of remembering the background texture. Although tracking the moving objects, texture elements of the background changed on some trials. When there was a change, texture elements were rotated by the same number of degrees of angle. When rotated in one direction, they created a highly noticeable change in global texture. However, when rotated in the other direction, they maintained a constant global pattern. When performing the concurrent tracking task, subjects were able to detect changes to global texture but had much more difficulty detecting the nonglobal texture changes. Without the concurrent tracking task, subjects were better able to detect the nonglobal texture changes. Alvarez and Oliva (2008) used a variant of the concurrent tracking task to also show that people extract summary statistics when attentional load is increased. Together, these data suggest that people's attentional filters are not all-or-none. For example, when attentional load is high, the abstract "gist" of seemingly unattended aspects of a scene can still be perceived and remembered. From the standpoint of efficient data compression, it seems that subjects in the Alvarez and Oliva experiments formed a compressed code of a scene based on a high cost to misrepresenting the moving objects in a scene, and a low cost to misrepresenting the texture background. As we explain in more detail below, the fact that subjects were able to detect a categorical change to the background texture better than a noncategorical change is well-explained by the properties of efficient data compression.

Second, within the "visual search" experimental paradigm, we take the well-known "pop-out" phenomenon as an illustrative example. Experimentalists have found that certain targets, when surrounded by certain kinds of distractors, are so easy to find that they immediately "pop out". More precisely, the amount of time required to find the target increases very little with number of distractors. Research has identified four important and complementary factors determining search times and pop-out: (a) target-distractor similarity (Avraham, Yeshurun, & Lindenbaum, 2008; Duncan & Humphreys, 1989); (b) distractor-distractor similarity (Avraham et al., 2008; Duncan & Humphreys, 1989); (c) how much "scrutiny" (i.e., representational detail) is required to distinguish targets and distractors (Ahissar & Hochstein, 1993; Hochstein & Ahissar, 2002); and (d) familiarity with a display (Chun & Jiang, 1998; Corneille, Goldstone, Queller, & Potter, 2006; Eckstein, 2011; Wang, Cavanagh, & Green, 1994).

Efficient data compression provides a coherent and theoretically grounded explanation for all four of these factors. Our key assumption is that when a search array is easy to compress, enough perceptual detail is represented so that the target can be distinguished from distractors at first glance (i.e., pop-out occurs; Hochstein & Ahissar, 2002). If, however, a search array is hard to compress, this will not be possible, and thus it will be necessary to shift attention to different areas in the display to perceive finer details. The effects of target-distractor and distractor-distractor similarity can be explained via the prior-knowledge principle discussed above because more homogeneous regions of a display contain less information and may be represented with shorter codes, whereas more heterogeneous regions contain more information and thus require longer codes. Indeed, this strategy is used in popular image compression algorithms (Deutsch, 1996). The task-dependency principle plays a key role in defining similarity because the nature of a task defines which objects or conjunctions of features the visual system should consider to be similar or dissimilar. The amount of scrutiny required to distinguish target and distractor is a result of the limited-capacity principle because a limited capacity makes representation of fine details more difficult. Lastly, the effects of stimulus familiarity on search times is another obvious outgrowth of the prior-knowledge principle.

The theory of visual search we present here is broadly consistent with another theory, the texture tiling model (TTM; Chang & Rosenholtz, 2016). TTM posits that peripheral vision is well described as a more compressed version of foveal vision, and provides a particular algorithmic implementation of that compression. This theory accounts for many effects in visual attention (including search performance) by considering the amount of information that can be extracted from the periphery in visual attention tasks (Rosenholtz, 2017). Our theory can be seen as related to TTM but more abstracted, in that it does not explicitly take into account differences between foveal and peripheral vision and does not commit to any particular compression algorithm.

## Modeling Attention as Data Compression

Here, we present simulations demonstrating how attentional allocation of perceptual resources can be modeled in an end-to-end manner as lossy compression. We consider simple visual search tasks in which the target is defined by a single stimulus feature (e.g., shape), and complex tasks in which the target is defined by a conjunction of features (e.g., shape and color). Our hypothesis is that people's search speeds are related to the compressibility of search arrays. In particular, we conjecture that single-feature search tasks tend to use displays that are highly compressible (compact codes can still represent fine-scale sensory information for these displays), and thus search speeds are high, whereas conjunction tasks use displays that are less compressible (codes can only represent coarser-scale information for these displays) leading to slower search speeds.

To begin to explore this hypothesis, we created a single-feature search task in which distractors were red objects composed of horizontal and vertical edges, and targets were red squares. In this case, targets could be located based on a single stimulus feature, namely shape. In the conjunction search task, distractors were either red or blue objects composed of horizontal and vertical edges or were blue squares, and targets were red squares. A conjunction of shape and color features was required to locate targets in this type of task.

Extended model architectures were trained to perform the two tasks. The autoencoder of an architecture was trained with respect to both image reconstruction and decision errors, with a higher weight on reconstruction error. The decision module was trained to output the spatial coordinates of the target. The weight on the reconstruction error was higher than the decision error because we assume that people's visual systems are primarily optimized for more general visual features rather than ones that are specific to this particular visual search task.

Figure 15 illustrates autoencoder image reconstructions corresponding to single-feature (top row) and conjunction (bottom row) search displays. Reconstructions for each type of display were produced by autoencoders with roughly equal measured capacity. Because conjunction search displays contain more stimulus information on average, reconstructions of these displays are lower-fidelity than those of single-feature search displays.

To compare the compressibility of single-feature and conjunction search displays more carefully, we produced approximate rate-distortion curves (as was done with the set-size experiment above). If one type of search array is less compressible than another, then its rate-distortion curve should be higher (larger distortion or error values for the same rates). The results, shown in Figure 16, confirmed our intuition that conjunction search displays contain more visual information, as the rate-distortion curve for the conjunction displays was clearly higher than that of the single-feature displays. Taken as a whole, these results are consistent with our hypothesis that single-feature search tasks use displays that are highly compressible, and thus people's search speeds are high, whereas conjunction tasks use displays that are less compressible leading to slower search speeds.

## General Discussion

Engineers must take capacity limits into consideration when designing artificial systems. A common way of ameliorating the negative effects of capacity limits is through data compression. When designing efficient data compression algorithms, RDT provides engineers with a principled mathematical framework for quantifying the trade-offs between rate (or capacity) and distortion

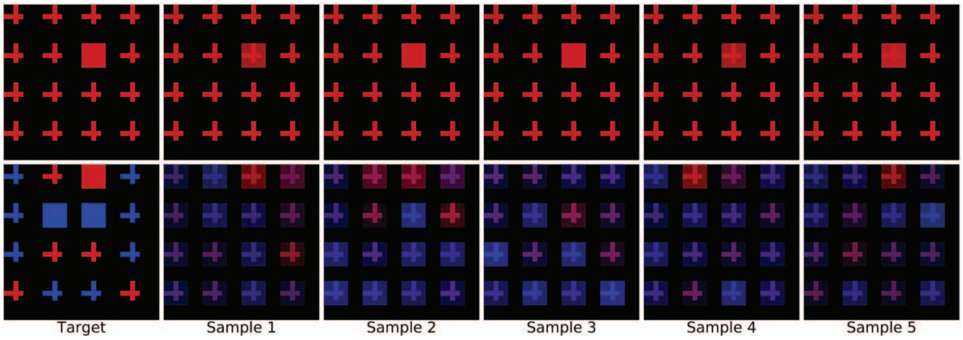

*Figure 15.* Autoencoder image reconstructions for single-feature (top) and conjunction (bottom) search displays. The leftmost image in a row shows the original display and the remaining images are image reconstruction samples. See the online article for the color version of this figure.

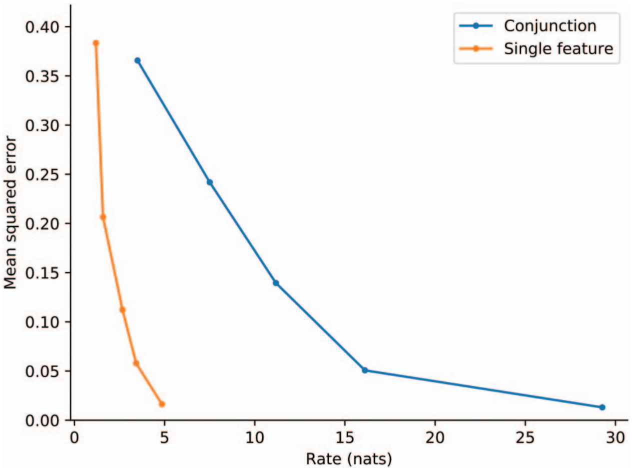

*Figure 16.* Approximate rate-distortion curves for single-feature and conjunction search displays. Error is based on Euclidean distance between actual and predicted target coordinates. See the online article for the color version of this figure.

(or error). In this article, we have argued that a similar situation holds with respect to people and other biological organisms. Because people's perceptual and perceptual memory subsystems are physically realized, they are necessarily capacity-limited. Biology (or evolution/development) has discovered that data compression can help people make the most of their limited capacities. Consequently, RDT can provide psychologists with a rigorous framework for understanding important aspects of perception and perceptual memory.

A goal of this article has been to describe a small set of general principles for efficient data compression that provides accounts for many behavioral phenomena (and many experimental results in the scientific literature) in multiple domains of perception and perceptual memory. These principles follow directly from RDT. This aspect of our work can be regarded as a "computational theory." According to Marr (1982), a computational theory of an information processing system analyzes the system's goals to determine what the system *should* do, such as the optimal computations a system should perform in order to achieve its goals.

A second goal of this paper has been to present a modeling framework for implementing the principles. We noted that exact methods exist for RDT analysis in low-dimensional stimulus spaces, but that approximate methods are needed for high-dimensional spaces. Although previous researchers have used deep neural networks to approximately implement RDT in high-dimensional spaces, these implementations have been limited to tasks in which the sole goal is data compression with respect to reconstruction error (e.g., Ballé et al., 2016). An important contribution of the research presented here is that we introduced a new deep neural network architecture that approximately implements RDT. Our architecture discovers good data compressions even when the data will be used for regression, classification, recognition, or other tasks. Consequently, the model can perform the same types of tasks as participants in experimental studies (e.g., change-detection or recall tasks). A key property of our model is that it is trained end-to-end, operating on raw perceptual input (e.g., pixel values) rather than intermediate levels of abstraction, as is the case with most psychological models. Our framework therefore represents an early step toward scaling up models of perception and perceptual memory toward levels of complexity faced in real-world situations. The neural network implementation of our theory can be regarded as providing a "process" model of aspects of perception and perceptual memory, focusing on the mechanisms giving rise to perceptual and mnemonic phenomena (McClelland et al., 2010).

A final goal of this article has been to offer conjectures about possible implications of efficient compression for memory organization and attention. We discussed how efficient compression can occur in memory over time, thereby providing motivations for multiple memory systems operating at multiple time scales. We also discussed how efficient compression may explain some attentional phenomena such as RTs in visual search.

The work presented here is a broad explication, touching on many different areas of perception and memory. Future research will need to conduct detailed studies of the implications of this work within each specific area. We close this article by describing several directions for future research that we believe might be particularly productive.

As mentioned above, our framework has the potential to shed light on patterns of neural activity in biological organisms. For instance, although some researchers have argued that the primary goal of the visual ventral stream is categorization (Yamins & DiCarlo, 2016), one could test the alternative hypothesis that the primary goal is efficient data compression. If one assumes a categorical loss function (as in some of our simulations), then the former hypothesis can be recovered as a special case when the weight on the pixel-reconstruction term in the loss function is zero. Importantly, categorization and compression are not mutually exclusive goals but, to the contrary, may be viewed as complementary. As demonstrated above, compression can lead to abstract codes with a categorical bias.

Three of the principles for efficient data compression—limited capacity, prior knowledge, and task dependency principles—follow directly from RDT. The remaining principle—information decay continuum—was introduced to address important aspects of biological systems that are not addressed by traditional engineering work. The information decay continuum principle states that there is a decline over time in the information content of individual perceptual memory traces. We hypothesized that this decline is useful to capacity-limited agents because it allows agents to devote fewer resources to older traces over time, thereby freeing up resources that can be used to encode new information (J. R. Anderson, 1991; J. R. Anderson & Schooler, 1991).

The information decay continuum principle raises many challenging questions. For instance: How might "recoding" of memory traces over time be implemented in biological systems? Can the existence of multiple memory systems be explained as optimal under certain assumptions, and do these systems correspond to the ones hypothesized in the scientific literature? Given that there is neural overlap between perception and perceptual memory, to what extent do perceptual demands constrain memory performance (and to what extent do memory demands constrain perceptual performance)?

An implication of the information decay continuum principle is that perception and perceptual memory systems have substantial commonalities, operating in similar ways because they use similar organizing principles (albeit at different time scales, and thus with different parameter settings). If so, this implication is inconsistent with theories arguing that the mind consists of multiple modules (e.g., modules for perception, language, motor action, etc.), each operating by its own set of principles (e.g., Fodor, 1983). As discussed above, several studies have found that perceptual and memory systems often function similarly, with similar representational biases and overlapping neural substrates. Future research will need to explore the commonalities, both cognitive and neural, among perception and memory in more detailed ways.

Another implication of the information-content continuum principle is that systems have a continuum of representations ranging from perceptually detailed representations of recent stimuli at one end of the continuum to more categorical and abstract representations of recent and older stimuli at the other end. We hypothesized that this diversity of representations is useful because different tasks require different types of information. This hypothesis has important ramifications for the study of decision making. For example, Lennie (1998) argued that the visual system is composed of multiple hierarchical levels and that task-relevant information can be recovered at every level. If so, then decision making in perceptual tasks requiring relatively low-level visual details (e.g., planning eye movements) is subserved by one set of levels, whereas decision making in tasks requiring higher-level information (e.g., distinguishing facial emotional expressions under multiple viewpoint and lighting conditions) is subserved by another set. The existence of a continuum of perceptual and perceptual memory systems with a continuum of representations motivates the need for future work performing careful task analyses on a range of behavioral tasks to uncover what information is needed for each task, and what types of mental and neural representations are most suitable for each task.

Above, we described our early steps toward thinking about relationships between data compression and perceptual attention. We described simulation results indicating that search times in visual search tasks may be explained by capacity limits and lossy compression. Although this is highly preliminary work, it suggests the potential merits of a new conceptualization of attention based on efficient data compression. Future work could explore this research direction, further studying RDT accounts of search times with different configurations of distractor and target objects, as well as accounts of other phenomena studied in the domain of perceptual attention.

Efficient data compression also has implications for understanding learning and expertise. How should we understand the differences between domain experts and novices? RDT suggests three important factors. It may be that experts outperform novices in a domain because they have higher capacities when processing stimuli in the domain. It may be that experts have greater prior knowledge about the stimuli in the domain. Or it may be that experts have loss functions that are more finely tuned to the tasks in the domain. Future research can use RDT to quantitatively and rigorously study the differences in these factors between experts and novices.

The work presented here has been restricted to episodic perceptual memory, but the field of psychology has identified several different types of memory systems. Do the principles underlying perceptual memory also play a role in, say, semantic or procedural memory? Our simulation results demonstrate that low-capacity systems tend to develop categorical or abstract compressions of stimuli. Do principles of efficient data compression also underlie abstraction in semantic and procedural memory? For example, can semantic and procedural memory also be shown to demonstrate categorical bias and sensitivity to task demands and prior distributions?

Lastly, we are encouraged by the successes of recent work studying the implications of principles of efficient data compression in nonperceptual and nonmeneumonic domains. For instance, Zaslavsky, Kemp, Regier, and Tishby (2018) argued that languages efficiently compress ideas into words. These authors found that color-naming systems across languages achieve near-optimal compression, studying this problem using an "information bottleneck" framework closely related to RDT. Similarly, C. A. Sims (2003, 2006) developed a theory, called "rational inattention," explaining why it is optimal for capacity-limited economic agents to attend to some decision variables while ignoring others. This theory, which uses an optimization framework that is highly similar to RDT, has important implications for macroeconomic, finance, behavioral economic, labor, trade, and political economy issues. We conjecture that principles of efficient data compression may serve as a unifying framework (among others) for understanding aspects of human behavior in a wide range of domains.

## References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America, 90,* 5718–5722.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). An information-theoretic analysis of deep latent-variable models. *arXiv preprint arXiv:1711.00464*. Retrieved from https://arxiv.org/pdf/1711.00464v1.pdf

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*. Retrieved from https://arxiv.org/pdf/1711.00464v3.pdf

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15,* 122–131.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15,* 106–111.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19,* 392–398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 7345–7350.

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

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Anderson, R. C. (1984). Role of the reader's schema in comprehension, learning, and memory. *Learning to Read in American Schools: Basal Readers and Content Texts, 29,* 243–257.

Avraham, T., Yeshurun, Y., & Lindenbaum, M. (2008). Predicting visual search performance by quantifying stimuli similarities. *Journal of Vision, 8*(4), 9.

Baddeley, A. D. (1966a). The influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology, 18,* 302–309.

Baddeley, A. D. (1966b). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology, 18,* 362–365.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4,* 829–839.

Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*. Retrieved from https://arxiv.org/pdf/1611.01704.pdf

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5,* 617–629.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 782–790). Cambridge, MA: MIT Press.

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

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision, 19*(2), 11.

Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. Cliffs, NJ: Prentice Hall.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory, 18,* 460–473.

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 556–566.

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

Brady, T. F., & Alvarez, G. A. (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 921–929.

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,* 487–502.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision, 11*(5), 4.

Brady, T. F., Konkle, T., Oliva, A., & Alvarez, G. A. (2009). Detecting changes in real-world objects: The relationship between visual long-term memory and change blindness. *Communicative & Integrative Biology, 2,* 1–3.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 7459–7464.

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,* 85–109.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., . . . Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*. Retrieved from https://arxiv.org/pdf/1804.03599.pdf

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13,* 51–62.

Chang, H., & Rosenholtz, R. (2016). Search performance is better predicted by tileability than presence of a unique basic feature. *Journal of Vision, 16*(10), 13.

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience, 32,* 12983–12989.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36,* 28–71.

Corneille, O., Goldstone, R. L., Queller, S., & Potter, T. (2006). Asymmetries in categorization, perceptual discrimination, and visual search for reference and nonreference exemplars. *Memory & Cognition, 34,* 556–567.

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

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169,* 323–338.

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 94–107.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15,* 559–564.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology, 66,* 115–142.

Deutsch, P. (1996). *DEFLATE compressed data format specification version 1.3* (Tech. Rep.). San Francisco Peninsula, CA: Aladdin Enterprises. http://dx.doi.org/10.17487/RFC1951

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5), 14.

Emrich, S. M., Riggall, A. C., LaRocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *Journal of Neuroscience, 33,* 6516–6523.

Fang, F., Kersten, D., Schrater, P. R., & Yuille, A. L. (2004). Human and ideal observers for detecting image curves. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems* (pp. 1459–1466). Cambridge, MA: MIT Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision, 10*(12), 27.

Fougnie, D., Cormiea, S. M., Kanabar, A., & Alvarez, G. A. (2016). Strategic trade-offs between quantity and quality in working memory. *Journal of Experimental Psychology: Human Perception and Performance, 42,* 1231–1240.

Genewein, T., Leibfried, F., Grau-Moya, J., & Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI, 2, 27,* 1–24.

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

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14,* 926–932.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics, 77,* 1239–1251.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7,* 217–229.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J., & Tenenbaum, J. B. (2011). Nonparametric Bayesian models of categorization. In E. M. Pothos & A. J. Wills (Eds.) *Formal approaches in categorization* (pp. 173–198). Cambridge, UK: Cambridge University Press.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 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,* 718–734.

Han, J., Lombardo, S., Schroers, C., & Mandt, S. (2018). Deep probabilistic video compression. *arXiv preprint arXiv:1810.02845*. Retrieved from https://arxiv.org/pdf/1810.02845.pdf

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458,* 632–635.

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

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., . . . Lerchner, A. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework. *Proceedings of the 2017 International Conference on Learning Representations, 2,* 6.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36,* 791–804.

Huang, L., & Awh, E. (2018). Chunking in working memory via content-free labels. *Scientific Reports, 8*(23), 1–10.

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

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98,* 352–376.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129,* 220–241.

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

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

Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 307–317.

Kaiser, D., Stein, T., & Peelen, M. V. (2015). Real-world spatial regularities affect visual working memory for objects. *Psychonomic Bulletin & Review, 22,* 1784–1790.

Kang, M.-S., Hong, S. W., Blake, R., & Woodman, G. F. (2011). Visual working memory contaminates perception. *Psychonomic Bulletin & Review, 18,* 860–869.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*. Retrieved from https://arxiv.org/pdf/1312.6114.pdf

Knill, D. C., Field, D., & Kersten, D. (1990). Human discrimination of fractal images. *Journal of the Optical Society of America A, 7,* 1113–1123.

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,* 558–578.

Konstantinou, N., Bahrami, B., Rees, G., & Lavie, N. (2012). Visual short-term memory load reduces retinotopic cortex response to contrast. *Journal of Cognitive Neuroscience, 24,* 2199–2210.

Kornbrot, D. E. (1978). Theoretical and empirical comparison of luce's choice model and logistic Thurstone model of categorical judgment. *Perception & Psychophysics, 24,* 193–208.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(4), 1–28.

Lennie, P. (1998). Single units and visual cortical organization. *Perception, 27,* 889–935.

Lerch, R. A., Cui, H., Patwardhan, S., Visell, Y., & Sims, C. R. (2016). Exploring haptic working memory as a capacity-limited information channel. *Proceedings of IEEE haptics symposium* (pp. 113–118). Philadelphia, PA: IEEE.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science, 6,* 279–311.

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

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17,* 391–400.

Lythgoe, J. N. (1991). Evolution of visual behaviour. In J. R. Cronley-Dillon & R. L. Gregory (Eds.), *Evolution of the eye and visual system* (pp. 3–14). London, UK: Macmillan Press.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17,* 347–356.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

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

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., . . . Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14,* 348–356.

Melcher, D. (2001). Persistence of visual memory for scenes. *Nature, 412,* 401.

Montaser-Kouhsari, L., & Carrasco, M. (2009). Perceptual asymmetries are preserved in short-term memory tasks. *Attention, Perception, & Psychophysics, 71,* 1782–1792.

Mureşan, H., & Oltean, M. (2018). Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica, 10,* 26–42.

Oberauer, K., & Eichenberger, S. (2013). Visual working memory declines when more features must be remembered for each object. *Memory & Cognition, 41,* 1212–1227.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). Burlington, VT: Elsevier Academic Press.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37,* 3311–3325.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review, 120,* 297–328.

Orhan, A. E., & Jacobs, R. A. (2014). Toward ecologically realistic theories in visual short-term memory research. *Attention, Perception, & Psychophysics, 76,* 2158–2170.

Park, I. M., & Pillow, J. W. (2017). Bayesian efficient coding. *bioRxiv preprint 178418.*

Parraga, C. A., Troscianko, T., & Tolhurst, D. J. (2000). The human visual system is optimised for processing the spatial information in natural visual images. *Current Biology, 10,* 35–38.

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology, 52,* 629–651.

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience, 6,* 97–107.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Ricker, T. J., & Cowan, N. (2010). Loss of visual working memory within seconds: The combined use of refreshable and non-refreshable features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1355–1368.

Rosenholtz, R. (2017). Capacity limits and how the visual system copes with them. *Electronic Imaging, 2017,* 8–23.

Rouder, J. N., Morey, R. D., Cowan, N., & Pealtz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review, 11,* 938–944.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533–536.

Saad, E., & Silvanto, J. (2013a). How visual short-term memory maintenance modulates the encoding of external input: Evidence from concurrent visual adaptation and TMS. *Neuroimage, 72,* 243–251.

Saad, E., & Silvanto, J. (2013b). How visual short-term memory maintenance modulates subsequent visual aftereffects. *Psychological Science, 24,* 803–808.

Santurkar, S., Budden, D., & Shavit, N. (2018). Generative compression. *Picture coding symposium* (pp. 258–262). San Francisco, CA: IEEE.

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

Schwarzkopf, D. S., & Kourtzi, Z. (2008). Experience shapes the utility of natural statistics for perceptual contour integration. *Current Biology, 18,* 1162–1167.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science, 20,* 207–214.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210,* 390–398.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization, 1,* 161–176.

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

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics, 50,* 665–690.

Sims, C. A. (2006). Rational inattention: Beyond the linear-quadratic case. *The American Economic Review, 96,* 158–163.

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

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science, 360,* 652–656.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review, 119,* 807–830.

Spence, I., Wong, P., Rusan, M., & Rastegar, N. (2006). How color enhances visual memory for natural scenes. *Psychological Science, 17,* 1–6.

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

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9,* 578–585.

Swan, G., Collins, J., & Wyble, B. (2016). Memory for a single object has differently variable precisions for relevant and irrelevant features. *Journal of Vision, 16*(3), 32.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121,* 124–149.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38,* 599–637.

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception & Psychophysics, 56,* 495–500.

Weckström, M., & Laughlin, S. B. (1995). Visual ecology and voltage-gated ion channels in insect photoreceptors. *Trends in Neurosciences, 18,* 17–21.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General, 131,* 48–64.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience, 20,* 864–871.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature, 440,* 91–95.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19,* 356.

Yoo, A. H., Klyszejko, Z., Curtis, C. E., & Ma, W. J. (2018). Strategic allocation of working memory resource. *Scientific Reports, 8,* 16162.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences of the United States of America, 115,* 7937–7942.

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Network: Computation in Neural Systems, 17,* 301–334.

(*Appendix follows*)

# Appendix

## Model Details

All simulations used layers that are standard within the neural network literature. No specific fine-tuning was required to produce our results, and results were relatively insensitive to the choices of number of hidden units and layers, as long as the number of units was large. In simulations involving the plants datasets, we chose standard fully-connected layers with "tanh" activation functions. In experiments involving Fruits-360 and visual-search datasets, all hidden units used rectified-linear activations (ReLU). All convolutional layers used 3 × 3 kernels, with a stride of two. In experiments that used a change-detection task, the decision module output was a single sigmoidal unit, and was trained with cross-entropy loss. For experiments in which the decision module was trained to recall specific feature values, the decision loss was squared error. Finally, in the natural-images experiment, the decision module output was a softmax layer with one output unit for each of the three categories. All networks were trained with the "Adam" optimization algorithm. Specific information about the number of layers and number of units per layer is detailed in Table A1. Code to run the experiments can be downloaded from https://github.com/Rick-C-137/efficientPerceptualDataCompression.

Table A1
*Details of the Network Architectures for Each Experiment*

| Experiment | Encoder hidden layers | Latent units | Decoder hidden layers | Decision hidden layers |
|---|---|---|---|---|
| Plants | MLP (500)<br>MLP (500) | 500 | MLP (500)<br>MLP (500) | MLP (100) |
| Plants set-size | MLP (500)<br>MLP (500) | 500 | MLP (500)<br>MLP (500) | MLP (100) |
| Fruits-360 | Conv 3 × 3 (32)<br>Conv 3 × 3 (64)<br>Conv 3 × 3 (64)<br>Conv 3 × 3 (64)<br>MLP (1,000) | 1000 | MLP (3,136)<br>Conv 3 × 3 (64)<br>Conv 3 × 3 (32)<br>Conv 3 × 3 (32) | None |
| Visual search | Conv 3 × 3 (32)<br>Conv 3 × 3 (64)<br>MLP (2,000) | 500 | MLP (4,096)<br>Conv 3 × 3 (64) | None |

*Note*. MLP = standard, fully-connected perceptron layers; Conv = standard 2D convolutional layers. Convolutional layers for the decoders were standard "convolution-transpose" layers. Numbers in parentheses indicate the number of hidden units for MLP layers, or the number of filters for convolutional layers. Each table entry for encoder, decoder, and decision layers is a comma-separated list of hidden layers. For example, in the plants experiments, the encoder of the β-VAE had two fully-connected hidden layers, each with 500 units.

Received February 20, 2019
Revision received February 11, 2020
Accepted February 11, 2020 ■