COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

Perceptual generalization and discrimination are fundamental cognitive abilities. For example, if a bird eats a poisonous butterfly, it will learn to avoid preying on that species again by generalizing its past experience to new perceptual stimuli. In cognitive science, the "universal law of generalization" seeks to explain this ability and states that generalization between stimuli will follow an exponential function of their distance in "psychological space." Here, I challenge existing theoretical explanations for the universal law and offer an alternative account based on the principle of efficient coding. I show that the universal law emerges inevitably from any information processing system (whether biological or artificial) that minimizes the cost of perceptual error subject to constraints on the ability to process or transmit information.

If a bird eats a poisonous or unpalatable species of butterfly, it will quickly learn to avoid preying on that species again in the future, by avoiding butterflies that look visually similar (*1*). This requires perceptual generalization, as no two butterflies look exactly alike. If generalization is too narrow—it learns to avoid one specific butterfly, but not others of the same species—the bird will continue to mistakenly consume toxic butterflies. However, if generalization is too broad—it avoids all butterflies—it will unnecessarily exclude edible food sources and consequently limit its fitness. A closely related ability is perceptual discrimination: If an edible species of butterfly closely resembles a different, toxic species (Batesian mimicry), the failure to perceptually discriminate between the two will also lead to negative consequences.

These examples demonstrate that adaptive behavior requires perceptual generalization and discrimination abilities that are finely calibrated to the costs of perceptual error. This is true not just for predator–prey relationships, but is equally important for expert-level human performance in domains such as medicine (*2*). Not surprisingly, the theoretical study of generalization is also central to progress in artificial intelligence and machine learning (*3*–*5*).

Just over 30 years ago, cognitive scientist Roger Shepard suggested that perceptual generalization was a suitable candidate for the first "universal law" in psychological science (*6*). Shepard's universal law of generalization states that the generalization between two stimuli (essentially, the probability of confusion) decreases as an exponential function of their distance within an appropriate metric "psychological space." This exponential generalization pattern has indeed proved to be near-universal, and the success of the empirical law has been impressive, accounting for data spanning a wide range of domains, sensory modalities, and across multiple species (*6*–*8*).

Shepard's explanation for this phenomenon revolves around the concept of a "consequential region" within psychological space that corresponds to a concept. For example, the concept of poisonous butterflies encompasses some set of stimuli in psychological space. Given one stimulus known to be an element of this set, the task facing the organism is to infer whether a novel stimulus will also fall in the same region; this task can be framed as one of probabilistic inference. Subsequent work (*9*, *10*) expanded on the idea of generalization as probabilistic inference, to include extrapolating from multiple exemplars and exploring alternative measures of perceptual distance or dissimilarity.

Here, I offer a qualitatively different explanation for the origins of the universal law in human perception, based on the principle of efficient coding (*11*), or the idea that biological information processing should seek to maximize performance

Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA.
***Corresponding author. Email: simsc3@rpi.edu**

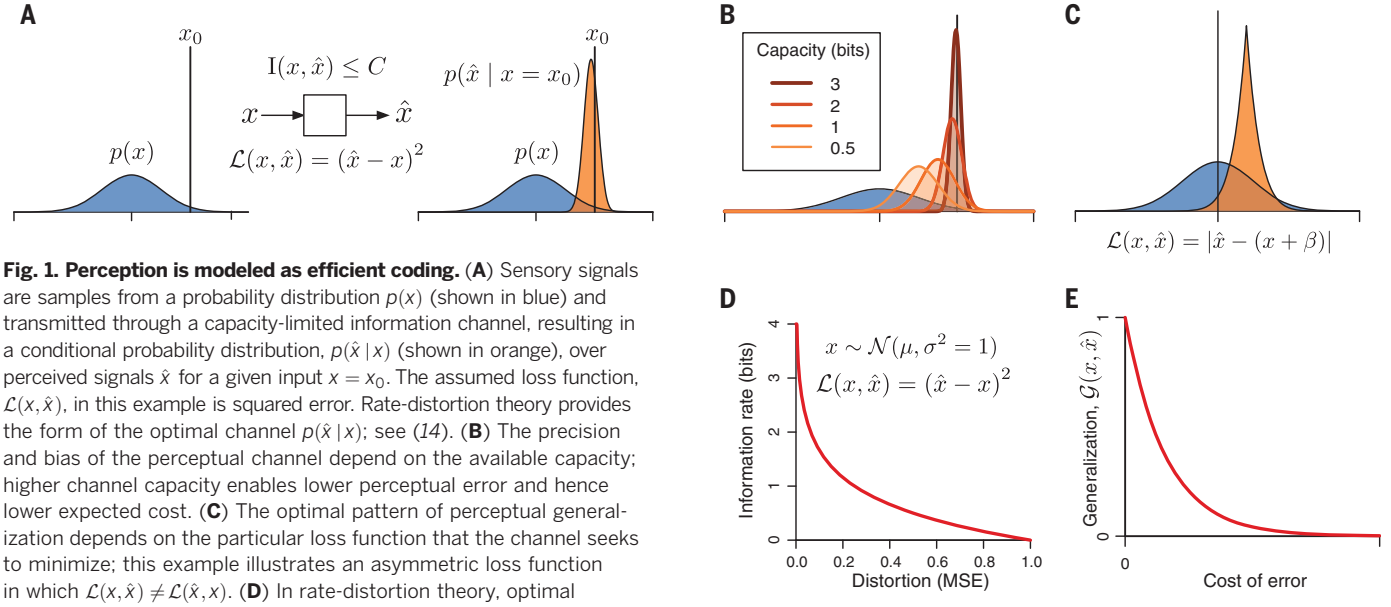

Fig. 1. Perception is modeled as efficient coding. (**A**) Sensory signals are samples from a probability distribution $p(x)$ (shown in blue) and transmitted through a capacity-limited information channel, resulting in a conditional probability distribution, $p(\hat{x} \mid x)$ (shown in orange), over perceived signals $\hat{x}$ for a given input $x = x_0$. The assumed loss function, $\mathcal{L}(x,\hat{x})$, in this example is squared error. Rate-distortion theory provides the form of the optimal channel $p(\hat{x} \mid x)$; see (*14*). (**B**) The precision and bias of the perceptual channel depend on the available capacity; higher channel capacity enables lower perceptual error and hence lower expected cost. (**C**) The optimal pattern of perceptual generalization depends on the particular loss function that the channel seeks to minimize; this example illustrates an asymmetric loss function in which $\mathcal{L}(x,\hat{x}) \neq \mathcal{L}(\hat{x},x)$. (**D**) In rate-distortion theory, optimal achievable performance is dictated by a "rate-distortion curve," which plots the minimum information rate (bits per transmission) necessary to achieve a given level of expected cost. The slope at any point along this curve is mathematically related to the steepness of the generalization gradient. This example illustrates the rate-distortion curve for a Gaussian information source with squared error cost function. (**E**) Without further assumptions, rate-distortion theory directly predicts an exponential generalization gradient as a function of the cost of perceptual error.

subject to constraints on information processing capacity.

Critically, the proposed approach also generates unique predictions that distinguish it from competing explanations for the universal law. These include predictions that relate the slope of the generalization gradient to information-theoretic quantities, asymmetric generalization gradients in situations where there are asymmetric costs for perceptual error, and the finding that artificial systems (such as the JPEG image compression algorithm) can also produce an exponential generalization gradient. The result is a revised universal law of perceptual generalization, which subsumes Shepard's statement of the law as a special case.

The approach uses results from the field of rate-distortion theory, a subdiscipline within information theory concerned with the design and analysis of optimal, but capacity-limited, information channels (*12*–*14*). Previous work has shown that rate-distortion theory offers a compelling account of human visual working memory limitations (*15*, *16*).

The current results can be concisely stated as follows: Perceptual generalization in any efficient communication system will necessarily follow an exponential function of the cost of perceptual error. In this framework, the emergence of the universal law is the signature of an organism that seeks to perceive the world as best as possible, according to some utility measure, subject to available resource limitations.

Figure 1 shows the theoretical framework and its properties. Perception is modeled as a capacity-limited information channel in which afferent sensory signals (x) are distributed according to the distribution $p(x)$. The perceived signal ($\hat{x}$) is related to its veridical value by a conditional probability distribution $p(\hat{x} \mid x)$. Capacity limits in the channel prevent transmitting sensory signals with perfect fidelity, and hence in general, $\hat{x} \neq x$. Instead, the goal of the channel is to minimize a given loss function, specified by $\mathcal{L}(x, \hat{x})$, subject to the constraint that the amount of information transmitted by the channel, measured by the mutual information $I(x, \hat{x})$, is at or below a capacity limit C. Rate-distortion theory provides analytical and numerical tools for solving such constrained optimization problems (*12*–*14*).

Notably, several of the properties illustrated in Fig. 1 (such as a "bias to the mean effect," Fig. 1A) are also predicted by Bayesian models of perception. As both are rational or optimal models of cognition, this is not surprising. Whereas Bayesian models of perception often make atheoretic assumptions about the nature of "internal noise" within a perceptual channel [e.g., (*17*)], rate-distortion theory instead gives sensory processing limitations a strong theoretical interpretation in terms of constructs from information theory. Hence, rate-distortion theory can be viewed as a special case of the more general class of Bayesian models of perception. As will be shown presently, this also allows the framework to generate unique predictions.

To connect rate-distortion theory to perceptual generalization, one needs a measure of the strength of generalization from one stimulus to another. Shepard (*6*) defined the following measure:

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}} \right)^{\frac{1}{2}} \tag{1}$$

where $p_{x\hat{x}}$ indicates the probability that a response associated with stimulus $\hat{x}$ is made to stimulus x. According to Shepard's universal law, generalization will follow an exponential function of the distance between x and $\hat{x}$ in an appropriate psychological space, where the distance is assumed to obey the basic metric axioms. Rate-distortion theory suggests a more general formulation for this law. Using Shepard's measure of generalization, rate-distortion theory directly predicts that generalization should follow

$$\mathcal{G}_{x\hat{x}} = \exp\left[s \frac{1}{2} \Big(\mathcal{L}(x, \hat{x}) + \mathcal{L}(\hat{x}, x) - \mathcal{L}(x, x) - \mathcal{L}(\hat{x}, \hat{x}) \Big) \right] \tag{2}$$

where the constant parameter $s < 0$ is monotonically related to the capacity of the channel. Note that this includes Shepard's original universal law as a special case. If the loss function satisfies two of the axioms of distance metrics, namely symmetry $[\mathcal{L}(x, \hat{x}) = \mathcal{L}(\hat{x}, x)]$ and identity $[\mathcal{L}(x, x) = \mathcal{L}(\hat{x}, \hat{x}) = 0]$, then one can easily verify that the generalization function reduces to

$$\mathcal{G}_{x\hat{x}} = \exp[s\ \mathcal{L}(x, \hat{x})] \tag{3}$$

Consequently, when the loss function is taken to be distance in a psychological space, Shepard's original universal law emerges from rate-distortion theory exactly. However, the result in Eq. 2 holds true under very general conditions, even when the psychological representation does not correspond to a metric space. As one example, if the mental representation of complex stimuli consists of a taxonomy of nested categories (*18*), the loss function may be defined in terms of tree distance between exemplars.

Rate-distortion theory was applied to the results of several published perceptual identification

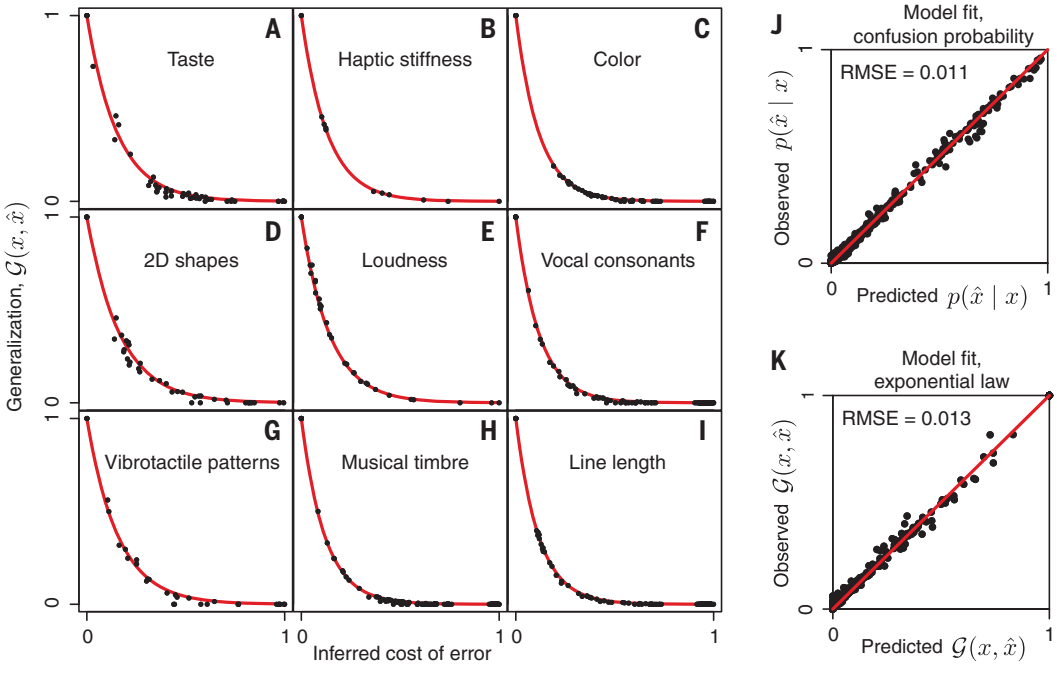

Fig. 2. Rate-distortion theory applied to a range of perceptual identification experiments. (**A** to **I**) Each panel plots the strength of generalization against the inferred cost of perceptual error. Cost functions were estimated by using Bayesian inference. Sources of empirical data are as follows: (A) (*27*); (B) (*28*) veterinarian data; (C) (*29*) data from block 3; (D) (*30*) experiment 1; (E) (*23*) unbiased condition; (F) (*21*) signal-to-noise = 0 dB condition; (G) (*31*) combined data from finger and forearm; (H) (*32*); (I) (*33*) subject MP, set size = 13 condition. (**J**) Goodness of fit to the empirical confusion matrices across all experiments, in terms of root mean squared error (RMSE). (**K**) Deviation between empirical generalization and the exponential generalization gradient predicted by rate-distortion theory across all experiments.

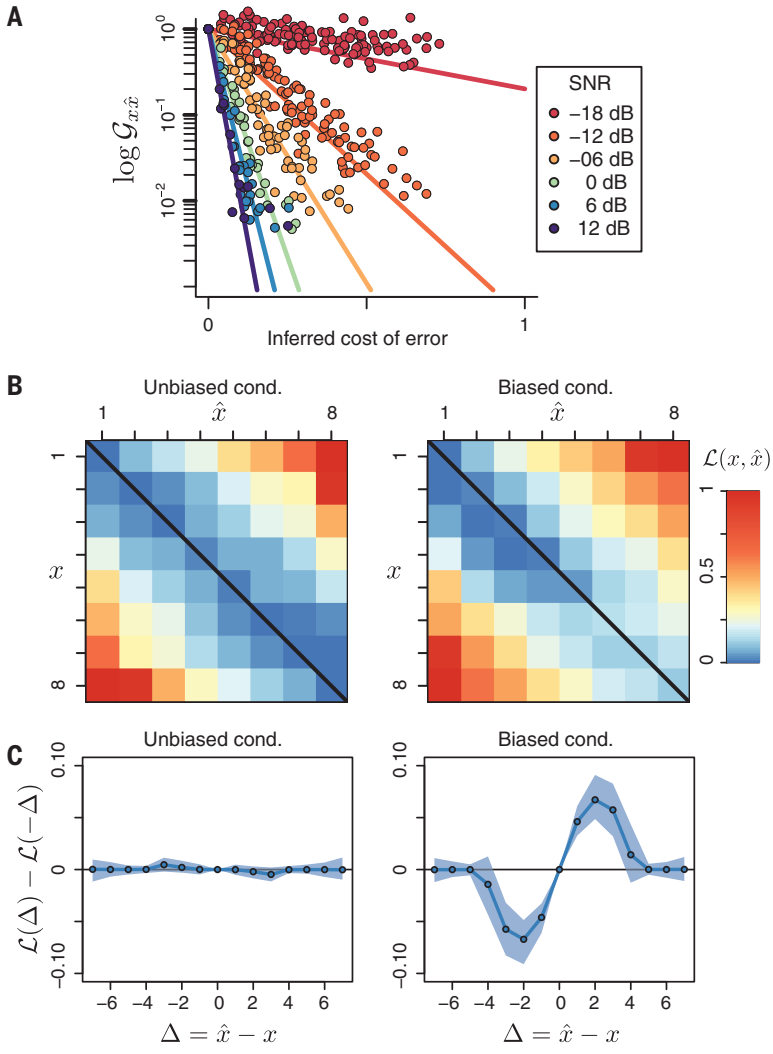

Fig. 3. Unique predictions of rate-distortion theory. (**A**) Rate-distortion theory predicts that the slope of the generalization gradient depends on the information rate of the perceptual channel. In this experiment (*21*), observers must identify vocal consonants embedded in varying levels of white noise. When plotted on a logarithmic axis, an exponential generalization gradient appears as a straight line. Rate-distortion theory is used to predict the slope of the generalization gradient for each condition. (**B**) Left and right panels show the inferred loss functions for the unbiased payoff condition, and biased payoff condition of the experiment reported in (*23*). (**C**) Asymmetry in the loss function is revealed by plotting the average cost for an overestimation error relative to the cost for an underestimation error of the same magnitude. The solid line indicates the maximum a posteriori estimate. The shaded region indicates the estimated 95% highest-density Bayesian credible interval. The figure demonstrates that asymmetry in the cost function appears only when there are task-defined asymmetries in the cost of perceptual error.

experiments (Fig. 2) that use a range of perceptual modalities (visual, haptic, auditory, gustatory). Archives of these data, along with model code, are provided online (*19*). On each trial of an identification experiment, a stimulus is randomly selected from a set, and the observer must identify it with a unique response. The resulting data consist of a perceptual confusion matrix, which gives the empirical frequency that stimulus x produced response $\hat{x}$. The perceptual loss function, $\mathcal{L}(x,\hat{x})$, is estimated from this confusion matrix by means of Bayesian inference.

As shown in Fig. 2, the observed relationship between the inferred cost of perceptual error (the estimated loss function $\mathcal{L}$) and the empirical generalization strength ($\mathcal{G}_{x\hat{x}}$, given by Eq. 1) follows an exponential gradient nearly exactly. Notably, this is not a consequence of a model that fits the data poorly but forces an exponential gradient. Rather, as shown in Fig. 2, J and K, rate-distortion theory simultaneously produces a precise model of the full probability distribution over perceptual confusion, as well as accurately predicts the exponential form of the generalization gradient. The supplementary materials also include a comparison of rate-distortion theory to an alternative existing model of perceptual identification, known as the Luce–Shepard choice model (*20*).

The key test, however, is whether rate-distortion theory generates predictions that distinguish it from competing explanations. The remainder of the paper focuses on three such predictions. The first is that the steepness of the generalization gradient should be monotonically related to the information rate of the perceptual channel. Specifically, when plotted on a logarithmic axis, exponential curves such as those shown in Fig. 2 will appear as straight lines with slope s. Whereas prior work has treated the slope of generalization as a free parameter, rate-distortion theory uniquely provides a strong theoretical interpretation for this quantity. In particular, for an optimal communication channel, the slope satisfies

$$s = \frac{\mathrm{d}R}{\mathrm{d}D} \tag{4}$$

where the term on the right-hand side of this equation is the slope of the rate-distortion curve for the channel (*12*), as illustrated in Fig. 1D. Consequently, experimental manipulations designed to influence the information rate of the perceptual channel (the numerator of this equation) should have a direct and predictable impact on the slope of the generalization gradient.

A test of this prediction is provided by the classic experiments reported in (*21*). In these experiments, subjects were asked to identify vocal consonants embedded in six different levels of white noise (signal-to-noise ratio ranging from 12 to −18 dB). Intuitively, increasing the amount of noise will decrease the amount of information about the signal that the observer can process. Under the assumption that the stimulus noise influences the information rate of the channel (the numerator of Eq. 4), but not the cost function for perceptual error (the denominator), it is possible to predict the slope of the generalization gradient in a parameter-free manner. The results are shown in Fig. 3A. In this plot, the generalization curves are shown on a logarithmic

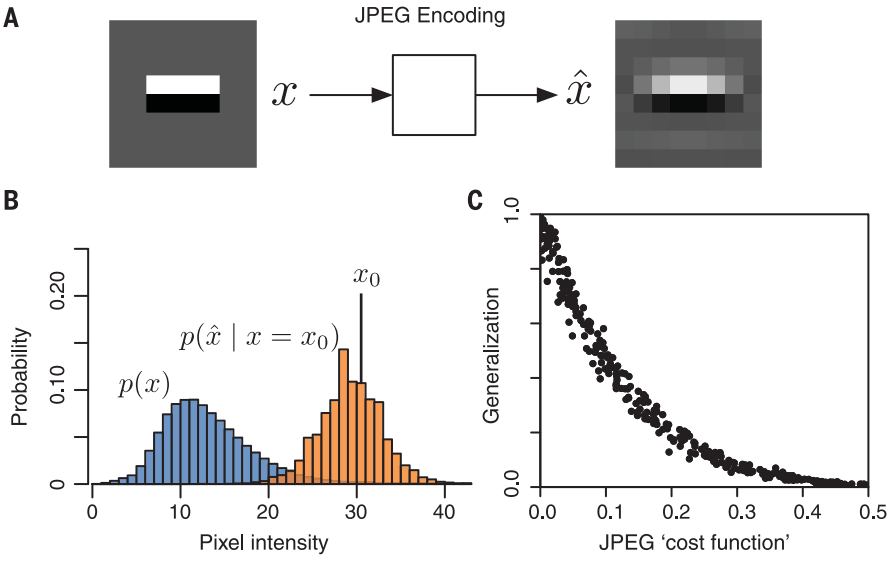

Fig. 4. Investigating perceptual generalization in the JPEG image compression algorithm. (**A**) An image patch encoded by JPEG will typically introduce pixel-level deviations. In particular, JPEG optimizes the transmission of low–spatial frequency components at the expense of introducing larger errors in coding higher spatial frequencies. Squared error within this scaled frequency domain closely approximates the "cost function" for JPEG image coding. (**B**) Conditional probability distribution over JPEG encoding of a particular pixel intensity, averaged over a large number of image patches drawn from a natural scene image. The behavior of the channel closely mirrors the predictions of rate-distortion theory (compare to Fig. 1A). (**C**) Deviations between an input and JPEG-encoded grayscale image of a natural scene conform to the universal law of generalization.

axis to illustrate the change in slope across stimulus noise conditions. The empirical slope of the generalization gradient closely follows the predictions of rate-distortion theory.

A second prediction of rate-distortion stems from the fact that unlike in Shepard's theoretical account, there is no requirement that perceptual generalization must be symmetric. Empirical asymmetries in generalization have previously been raised as an argument against a metric representation of perceptual similarity (*22*). In the present case, a different theoretical origin for asymmetry is predicted in terms of asymmetric costs of perceptual error. An empirical test of this prediction is found in an experiment reported in (*23*). In this experiment, subjects were tasked with identifying pure tones of varying loudness. Subjects were motivated to perform accurately by awarding points for correct responses and deducting points for errors; points were exchanged for a monetary bonus at the end of the experiment. Each subject completed two experimental conditions. In the neutral condition, payoffs were symmetric for all types of errors, whereas in the biased condition, overestimate errors were more costly than underestimate errors.

The inferred loss functions are illustrated in Fig. 3B for both the neutral and biased condition. Inferred costs for perceptual error are symmetric in the unbiased penalty condition, but substantially asymmetric in the biased penalty condition. Formal model comparisons (reported in the supplementary materials) reveal that the data are better explained by rate-distortion theory with an asymmetric cost function, compared to an alternative model that assumes symmetric perceptual distance.

Lastly, rate-distortion theory predicts that exponential generalization gradients should not be limited to biological information processing, but rather should be exhibited by any communication system that operates efficiently in the rate-distortion sense, whether natural or artificial. Figure 4 illustrates an identification "experiment" conducted on the JPEG image compression algorithm. The experiment was performed by taking grayscale photographs from a natural scene database (*24*) and encoding them using the JPEG algorithm. As JPEG is a form of lossy compression, the encoded images will almost certainly introduce perceptual "confusions"—an input pixel replaced by a somewhat different pixel at the output stage (Fig. 4A). A confusion matrix is obtained by collecting the joint statistics of input and JPEG-encoded pixels. Compared to human participants, JPEG has the useful feature that the objective for perceptual coding is obtainable by inspection of its algorithm. In brief, JPEG performs a discrete cosine transform (DCT) on an input image and scales the coefficients by a weight matrix that emphasizes coding accuracy for low spatial frequencies. This weighted DCT representation is essentially the "psychological space" for JPEG encoding. Figure 4C plots the strength of generalization between pixel values against the average squared error distance in quantized DCT space. The results illustrate that JPEG image coding also conforms to the universal law of generalization. Although this finding is consistent with rate-distortion theory, it is difficult to reconcile with alternative explanations for the universal law.

The current work is only part of a growing body of literature showing the broad applicability of efficient coding as a means of understanding biological information processing (*25*, *26*). As a theoretical framework, efficient coding is not an alternative to the popular Bayesian perception framework, but rather is an extension in which sensory limitations are attributed to information processing capacity limitations. As perception exists to maximize the utility of behavior, it is a compelling idea that evolution drives perceptual systems toward the regime of rate-distortion efficiency: optimizing performance subject to information processing constraints.

REFERENCES AND NOTES

1. J. V. Z. Brower, *Evolution* **12**, 32–47 (1958).
2. M. Myles-Worsley, W. A. Johnston, M. A. Simons, *J. Exp. Psychol. Learn. Mem. Cogn.* **14**, 553–557 (1988).
3. T. M. Mitchell, *Artif. Intell.* **18**, 203–226 (1982).
4. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, *Science* **350**, 1332–1338 (2015).
5. A. Esteva *et al*., *Nature* **542**, 115–118 (2017).
6. R. N. Shepard, *Science* **237**, 1317–1323 (1987).
7. K. Cheng, *Psychol. Sci.* **11**, 403–408 (2000).
8. S. Ghirlanda, M. Enquist, *Anim. Behav.* **66**, 15–36 (2003).
9. J. B. Tenenbaum, T. L. Griffiths, *Behav. Brain Sci.* **24**, 629–640 (2001).
10. N. Chater, N. P. M. Vitányi, *J. Math. Psychol.* **47**, 346–369 (2003).
11. H. Barlow, in *Sensory Communication* (MIT Press, Cambridge, MA, 1961), pp. 217–234.
12. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression* (Prentice Hall, Englewood Cliffs, NJ, 1971).
13. R. Blahut, *IEEE Trans. Inf. Theory* **18**, 460–473 (1972).
14. C. R. Sims, *Cognition* **152**, 181–198 (2016).
15. C. R. Sims, R. A. Jacobs, D. C. Knill, *Psychol. Rev.* **119**, 807–830 (2012).
16. C. R. Sims, *J. Vis.* **15**, 2 (2015).
17. A. A. Stocker, E. P. Simoncelli, *Nat. Neurosci.* **9**, 578–585 (2006).
18. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, *Science* **331**, 1279–1285 (2011).
19. For an online archive of data sets and model code, see https://osf.io/x5ckn. Additional details of the model and parameter estimation are provided in the supplementary materials.
20. R. N. Shepard, *Psychometrika* **22**, 325–345 (1957).
21. G. A. Miller, P. E. Nicely, *J. Acoust. Soc. Am.* **27**, 338–352 (1955).
22. A. Tversky, *Psychol. Rev.* **84**, 327–352 (1977).
23. D. E. Kornbrot, *Atten. Percept. Psychophys.* **24**, 193–208 (1978).
24. W. S. Geisler, J. S. Perry, *J. Vis.* **11**, 14 (2011).
25. X. X. Wei, A. A. Stocker, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10244–10249 (2017).
26. S. E. Marzen, S. DeDeo, *J. R. Soc. Interface* **14**, 20170166 (2017).
27. T. P. Hettinger, J. F. Gent, L. E. Marks, M. E. Frank, *Percept. Psychophys.* **61**, 1510–1521 (1999).
28. N. Forrest, S. Baillie, H. Z. Tan, in *EuroHaptics 2009*, IEEE, pp. 646–651 (2009).
29. R. M. Nosofsky, *J. Exp. Psychol. Learn. Mem. Cogn.* **13**, 87–108 (1987).

30. F. G. Ashby, W. W. Lee, *J. Exp. Psychol. Gen.* **120**, 150–172 (1991).
31. M. Azadi, L. Jones, in *World Haptics Conference (WHC)*, IEEE, pp. 347–352 (2013).
32. J. M. Grey, *J. Acoust. Soc. Am.* **61**, 1270–1277 (1977).
33. J. N. Rouder, R. D. Morey, N. Cowan, M. Pfaltz, *Psychon. Bull. Rev.* **11**, 938–944 (2004).

ACKNOWLEDGMENTS

Funding: This research was supported by NSF grant DRL-1560829. **Author contributions:** C.R.S. conducted the research and wrote the manuscript. **Competing interests:** None declared. **Data and materials availability:** Online data archives associated with this paper are provided via the Open Science Framework, at https://osf.io/x5ckn/.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6389/652/suppl/DC1
Materials and Methods
Fig. S1

2 October 2017; accepted 3 April 2018
10.1126/science.aaq1118

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims

Science, **360** (6389), .
DOI: 10.1126/science.aaq1118

Balancing costs and performance

Deciding whether a novel object is another instance of something already known or an example of something different is an easily solved problem. Empirical mapping of human performance across a wide range of domains has established an exponential relationship between the generalization gradient and interstimuli distance. Sims now shows that this relationship can be derived from a consideration of the costs of optimal information coding.

Science, this issue p. 652