The Information-Processing Perspective on Categorization

Manolo Martínez

Categorization behavior can be fruitfully analyzed in terms of the trade-off between as high as possible faithfulness in the transmission of information about samples of the classes to be categorized, and as low as possible transmission costs for that same information. The kinds of categorization behaviors we associate with conceptual atoms, prototypes, and exemplars emerge naturally as a result of this trade-off, in the presence of certain natural constraints on the probabilistic distribution of samples, and the ways in which we measure faithfulness.

Beyond the general structure of categorization in these circumstances, the same information-centered perspective can shed light on other, more concrete properties of human categorization performance, such as the results of experiments on supervised categorization in J. D. Smith and Minda (1998).

1 Introduction

A central debate in cognitive science concerns whether concepts are *unstructured symbols* which refer to classes of entities (this position is often called *atomism*, Fodor 1980, 2008), or instead should be identified with *bodies of information* about the class of entities targeted by the concept (henceforth, sometimes simply "the class"). I will refer to this other position as *informationism*. In the most popular development of the informationist alternative, these bodies of information are *prototypes* (Reed 1972; Rosch 1999; Hampton 2006; Minda and Smith 2011; J. D. Smith and Minda 1998): statistical summaries of the class, such as its central tendency, or the "centers of clusters of similar objects [of the class]" (Hampton 2006, 1). Another historically important way of elaborating the informationist idea is in terms of *exemplars* (Osherson et al. 1990; Nosofsky, Palmeri, and McKinley 1994; E. E. Smith and Medin 1999): individual instances of the class that the user of the concept remembers, and on which (instead of on prototypes) they rely when categorizing.

Prototypes and exemplars provide compelling explanations of important phenomena related to our use of concepts. In this paper I focus on categorization, the process through

which we determine whether some entity belongs to one class or another (Medin and Heit 1999, 100). One of the main themes of the prototype approach to categorization is that entities are classified as belonging to class A (B, C...) because they are closest to the A (B, C...) prototype, according to some abstract measure of distance, defined over some abstract space of possible entities (more on these spaces in §2.3.) Prototype theory, for example, elegantly accounts for typicality effects (e.g., that, for many classes, some instances are more quickly and reliably categorized than others, and also are perceived as being better or more paradigmatic examples of the class, Rosch 1999; Minda and Smith 2011): the typicality of an entity for a certain class can be seen as a manifestation of its distance to the prototype of that class.

Conceptual atomism and conceptual informationism are often presented as rival accounts. See, e.g., Connolly et al. (2007); Fodor and Lepore (1996); or Laurence and Margolis' introduction to their very influential (1999) edited volume. Other theorists (notably Machery 2009, chap. 2) have argued that the situation is, in fact, even worse: atomists and informationists are not even theorizing about the same phenomenon. Concepts as bodies of information are posited by psychologists as a way to model and explain our performance in, e.g., categorization tasks; while concepts as unstructured symbols are chiefly posited by philosophers, among other things, as bearers of reference, and as building blocks in a compositional language of thought. My aim in this paper is not to offer an account of human categorization performance, with all its fascinating quirks, but to show how the main gists of atom-, prototype-, and exemplar-involving categorization strategies are in fact compatible, and continuous with one another. Behavior that involves all three, in various degrees, falls out from very simple principles related to information-processing efficiency: prototypes, exemplars and atoms are, all of them, part of an efficient solution to the problem of transmitting and storing information about a class. Small wonder information categorization often relies on them. I view the analyses of categorization I will develop here as continuous with Anderson's (1990) "rationalist" strategy:[1] we start from a characterization of what cognition is supposed to do, and, relying on that, we try to recover whatever details of cognitive performance we were interested in. In a sense, the approach I sketch here goes beyond Anderson (1990, chap. 3), in that *categorization itself* can be seen as emerging from the more fundamental need to make perceptual information available downstream, in the production of behavior.[2]

In §2 I introduce and discuss the main model I explore in this piece: an agent in a toy world populated by entities with different features. Which entities will the agent encounter, and how frequently, is governed by a joint probability distribution over those

[1] See also work on rational inattention (Sims 2003) and resource rationality (Lieder and Griffiths 2020; Zaslavsky et al. 2018).

[2] One can also view the models explored here as inscribed in the tradition of idealized investigations of communication and representation pursued in, e.g., Lewis (1969–2008); Skyrms (2010); Shea, Godfrey-Smith, and Cao (2017); or Martínez (2019a). Those models do not aim to show that, e.g., human conventions, with all their quirks, just *are* Nash equilibria in signaling games, but they do show that game-theoretic coordination captures, in an economical, formally perspicuous way, a good deal of how convention comes about, and what it is. I aim at shedding a similar kind of light on categorization behavior.

features. I then consider the following problem: which coding strategy should the agent follow, if they aim to 1) transmit or store information about the entities they encounter, as faithfully as possible, while 2) keeping transmission and storage costs as low as possible. It turns out that, for worlds which present "correlational structure", in the sense Eleanor Rosch (1999) gives to this notion, and under mild assumptions, optimal codebooks are composed by atoms (that is to say, by a discrete, finite number of signals), as atomists claim; yet these atoms are produced and consumed by processes of encoding and decoding that rely on prototypes, as informationists claim. There is no conflict between atoms and prototypes; both play a necessary role in efficient information transmission and storage. The resulting agent categorizes its inputs (the entities it encounters) by instantiating a discrete number of atomic signals, each of which is decoded by relying on a prototype.

§2 can be seen as dealing with unsupervised category creation: under the principled understanding of "optimal" that I develop in that section, atomic conceptual repertoires that rely on prototypes are optimal for certain important classes of problems. In §3 I deal with *supervised* category creation: I show that efficient information transmission can explain results by J. D. Smith and Minda (1998; see also T. L. Griffiths et al. 2011) which are sometimes interpreted as showing that subjects in a categorization task shift from prototype-based to exemplar-based categorization as the task progresses. Leaving aside whether this interpretation is warranted, this change in behavior can be more parsimoniously explained in terms of changes in the make-up of the optimal categorization repertoire as its richness (technically, its rate) increases. §4 offers some concluding remarks.

2 Prototypes and Efficient Information Transmission

In §2.1 I discuss a model in which an agent encounters entities with features drawn from a continuous probability distribution. In §2.2 I discuss a model with categorical features.

2.1 The Continuous Case

We first set up a toy world. This world is populated with entities, each of which has two features, A and B. These features take (or can be represented as) real values. You can think of the value of A and B as representing, say, length and weight respectively, according to some appropriate units and scale. Figure 3a provides an example of this sort of world: samples come from an equiprobable mixture of four bivariate Gaussian distributions with means $\langle 7, 13\rangle$, $\langle 9, 3\rangle$, $\langle 14, 3\rangle$, and $\langle 14, 10\rangle$, respectively,[3] where the first

[3]There's nothing special about those values. The exercise will work in exactly the same way with a different number of Gaussians, centered at different positions. A Jupyter notebook with the code necessary to generate the results and figures in this paper can be downloaded from https://osf.io/sz49u/?view_only=264a7f3a51944142b20d87f19561b4cb I encourage the reader to try out different "toy worlds" there.

number in each of the above ordered pairs corresponds to the value of feature A, and the second to the value of feature B. These four Gaussians have the same variance and are isotropic (in particular, they all have the 2×2 identity matrix as covariance matrix.).[4]

This toy world is one in which the abstract space of possible entities (that is, the space of possible combinations of a value for feature A and a value for feature B—*feature space*, as I will be calling it, following standard usage) is occupied by four equally sized, Gaussian-shaped mounds, centered at the above four points. That is to say, as the figure shows, most samples are close to these four means, but arbitrary departures from them are possible, if increasingly unlikely the further away from the mean they are (that's why the blobs thin out towards the periphery), and the number of samples close to each of the four means is more or less equal (that's why the four blobs are more or less of equal size).

This world presents what Rosch (1999, 190) calls *perceived structure*: "[C]ombinations of what we perceive as the attributes of real objects do not occur uniformly. Some pairs, triples, etc., are quite probable, appearing in combination sometimes with one, sometimes another attribute; others are rare. . . " Our toy world is predictable in exactly these systematic ways: for example, if we know that the feature A of a certain sample has a value around 14, we can be quite confident that its feature B will be either around 3 or around 10 (and that both these options are equally likely).

The task for the agent in the model is as follows: this world produces random samples, with the probabilities dictated by the underlying probability distribution, and they are tasked with storing *as faithful* a version of the sample they encounter as possible, while using *as little resources* as possible. Alternatively (and, as far as the mathematics of the model are concerned, equivalently), you can think of the task as that of transmitting information about the sample for use downstream, say, in the production of behavior appropriate to the presence of that sample. This task is basically a redescription of what Eleanor Rosch calls *cognitive economy*, one of her two "psychological principles of categorization" (Rosch 1999, 189): "what one wishes to gain from one's categories is a great deal of information about the environment while conserving finite resources as much as possible." (Rosch 1999, 190). We have already encountered "perceived world structure", Rosch's other principle of categorization, in the description of our toy world.

Roschian cognitive economy is an optimization problem with two objectives. First, *maximizing faithfulness* in transmission or storage: the signal you send forward or store should be decodable into a set of values which are as close as possible to the values you encountered. Second, *minimizing costs* in storage and transmission[5] while doing so. One way to make this optimization problem more precise (among various other,

[4]The results I will discuss here also apply to mixtures of Gaussians with different variances. An example is worked out in the Supplementary Material, section 1.

[5]From here on out, and for the sake of brevity, I will only talk of transmission; but it should be understood that the models to be discussed in this paper apply just as well to storage. Both operations are indistinguishable from the point of view of information theory—the main formalism I will be relying on in this paper.

partially overlapping formalisms) is to cast it in the vocabulary of information theory. The main, *point-to-point* model (Shannon 1948; Cover and Thomas 2006; MacKay 2003) is well-known, and relatively straightforward (see figure 1): from left to right, we start with a source that generates samples from an underlying probability distribution, the way I described our toy world above—these samples are the M in figure 1. The *entropy* of the source, $H(M)$, gives a measure of how unpredictable this source is: e.g., if only a handful of samples have high probability, entropy will be low; if many samples are more or less equally likely, entropy will be high.[6] Entropy is systematically related to world structure in Rosch's sense: for example, when "pairs, triples, etc." of features change in tandem, the resulting source entropy is lower than if they were independent of one another. In general, structure in the relevant sense results from relation of probabilistic dependence among feature values.

In the next stage of the point-to-point model, samples coming from the source are *encoded* into a signal, X. The purpose of this encoding is to make the information in the sample able to negotiate various constraints introduced by an intervening *channel*. Here I will focus on the kind of constraint that is most relevant to the Roschian cost-faithfulness trade-off: channels cannot transmit unlimited quantities of information, but have a limited *capacity*, C. This is just the average amount of information that signals leaving the channel carry about signals entering the channel.[7] The encoder, therefore, needs to *compress* the incoming message, M, so that the resulting signal, X, can be squeezed through the channel, and decoded at the other side into a message $\hat{M}$ that recovers as much of the relevant information in the original M as possible. The entropy of the signals, $H(X)$, is also called the *rate* of the code—you can think of it as the richness, or expressiveness, of the signaling repertoire available at the encoder.[8]

We can now reformulate Rosch's cognitive economy principle as a trade-off between rate and faithfulness. Intuitively, the more compressed the encoded signal is (that is, the less expressive the signal repertoire is), the less faithful it will be—think of a high quality

[6] In this paper I focus on the qualitative aspects of information theory and the light they can shed on our theories of concepts. I will gloss over most mathematical details. For more on the formalism of information theory, the reader should consult any of a number of standard textbook treatments (e.g., Cover and Thomas 2006, chap. 1 and 2).

[7] Calculated as the mutual information between the two random variables X and $\hat{X}$, $I(X;\hat{X})$. Mutual information measures the change in the expected number of binary (yes/no) questions that one needs to ask in order to know the value of X, before and after knowing the value of $\hat{X}$—that is to say, the difference between the unconditional entropy of X and its entropy conditional on $\hat{X}$: $I(X,\hat{X}) = H(X) - H(X|\hat{X})$.

[8] In this paragraph I have made liberal use of "conduit metaphors" (Reddy 1979; Eubanks 2001) according to which information about samples is encoded, transmitted, and then decoded for its use downstream. It is important to note, though, that fully explicit, non-metaphorical readings of the relevant notions are available: for example, "coding" M into X just means implementing a function that takes M as input and produces X as output. No more needs to be read into it, and, in particular, it is not necessary to think of coding as translation, in a semantically charged sense. The quality of the coding scheme in question, which is one of the main topics of what follows, will also be formalized in a way that does not depend (or not more than pretty much everything else, anyway) on covert, semantically-charged metaphors. I would like to thank an anonymous reviewer for pressing me here.

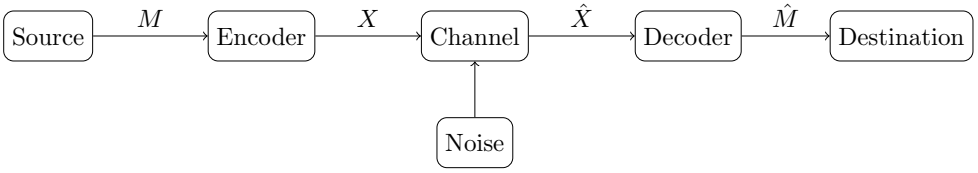


Figure 1: The main, *point-to-point*, Shannon model of information transmission.

CD track versus a compressed `mp3` version thereof, or Goya's *Caprichos*, as seen in the original printings versus low-resolution `jpeg` versions thereof. In order to quantify this trade-off we need a measure of faithfulness, or (more common in information theory) its converse, *distortion*, or *loss*—a function, d, which gives a score (say, a positive real number) to each pair of an incoming and a decoded message: $d\colon M \times \hat{M} \to \mathbb{R}^+$, higher scores meaning that the reconstruction is of worse quality, for whatever purposes the decoded message is to be put to at its destination. One widely used distortion measure when dealing with continuous data is the *mean squared Euclidean distance*, or *mean squared error* [MSE]:[9]

$$d(M, \hat{M}) = \frac{1}{n} \sum_{1 \leq i \leq n} (M_i - \hat{M}_i)^2$$

One of the foundational results in information theory, Shannon's so-called *lossy source coding theorem* (Shannon 1948; Berger 1971, chaps. 2–3; Cover and Thomas 2006, chap. 10,) formalizes the intuitive idea of a trade-off between expressiveness and faithfulness. Suppose that we wish to keep the average distortion of our signals below a certain value D. This theorem states that there is a specific minimum rate R, such that only signaling repertoires with a rate bigger than R can achieve an average distortion of D. Conversely, suppose that we can only afford to spend a rate R' in our signaling repertoire. Then the theory states that there is a certain average distortion D' which is the minimum we can achieve with that rate.

In general, there exists a monotonically increasing function $R(D)$ that gives the minimum

[9] For illustration, if you are presented with a sample with values $\langle 9.1, 2.8 \rangle$ for features A and B respectively, and you decode it as $\langle 9, 3 \rangle$, the distortion you are incurring in, according to the MSE measure, is:

$$\frac{(9.1 - 9)^2 + (2.8 - 3)^2}{2} = .025$$

If, on the other hand, you decode it as $\langle 9.5, 2.5 \rangle$, which is intuitively further away from the original message, you end up with a higher distortion:

$$\frac{(9.1 - 9.5)^2 + (2.8 - 2.5)^2}{2} = .125$$

rate, R, at which a certain target distortion D (the expected value of d) is achievable,[10] and a function $D(R)$ that gives the minimum distortion D which can be achieved with a rate of R. Furthermore, there are algorithms that can calculate $R(D)$ efficiently, at least for relatively simple, low-dimensional sources.[11]

To gain some initial intuition about how this rate-distortion function works, consider a very simple source: a fair coin that is repeatedly tossed. This source has two possible values, heads and tails, with probabilities $P(\text{HEADS}) = P(\text{TAILS}) = .5$. Suppose that we want to communicate the value of one of these tosses downstream. If we wish to communicate it in full (with no distortion) then we need 1 bit: e.g., we send a `1` if heads, and a `0` if tails. That is to say, $R(0) = 1$ (in words: the minimum rate for zero distortion is one bit.) Suppose on the other hand that we want to spend *no* rate at all. That is to say, we don't want to send anything. Then, the best that the decoder can do is to guess, say, heads every time, and be right half of the time. So, $R(.5) = 0$. We may also decide to use only .5 bits to encode each toss: this corresponds to an optimal distortion of 0.11.[12] And so on. Figure 2 is the full rate-distortion curve for this source.

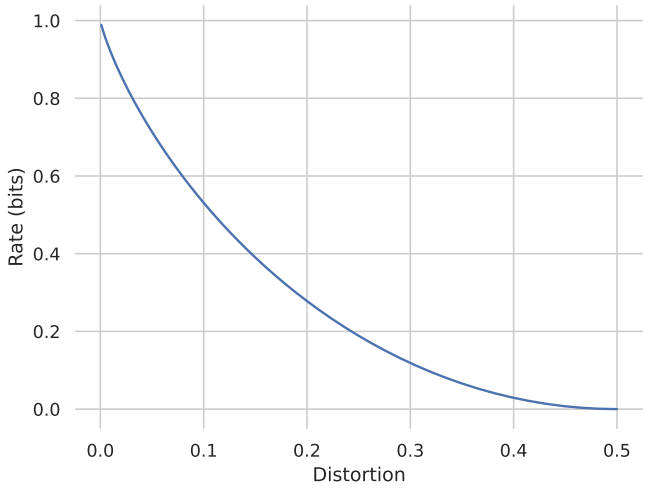


Figure 2: The rate-distortion curve for a source consisting of tosses of a fair coin

[10] $R(D)$ happens to be the minimal mutual information, $I(M; \hat{M})$, at which the target distortion D can be achieved (Cover and Thomas 2006, theorem 10.25).

[11] For the analyses in this paper I have used *deterministic annealing* for continuous data (Rose 1994, 1998) and the *Blahut-Arimoto* algorithm for discrete data (Blahut 1972; Arimoto 1972). General-purpose optimization algorithms can also be used.

[12] One way to achieve this rate-distortion pair (that is to say, <0.5 bits, 0.11 distortion>) is to use a probabilistic coder that encodes "heads" as `1` with probability .89 and as `0` otherwise; and vice versa for "tails".

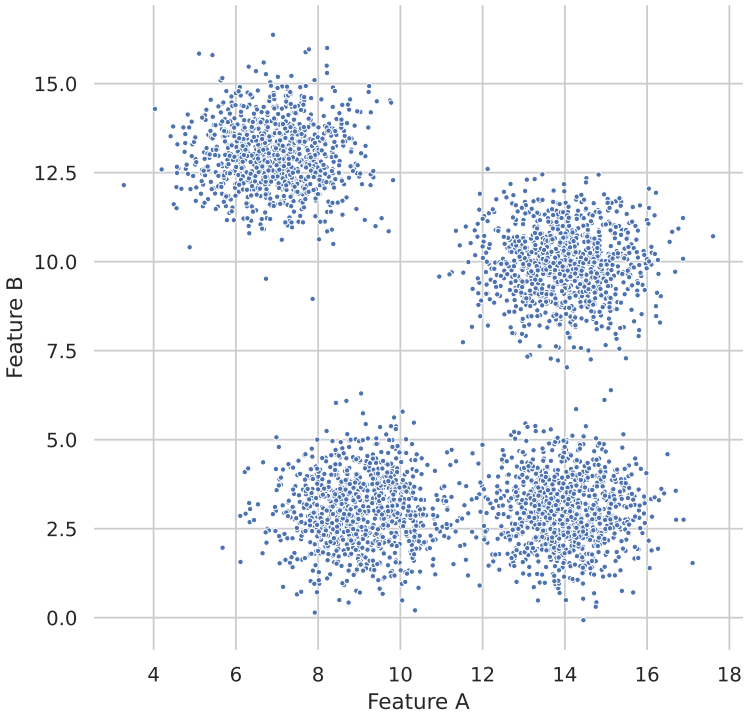
One can think of this as meaning that we can accept that level of unreliability, or noise, in our encoder if we are prepared to put up with .11 distortion.

Having operationalized richness-faithfulness trade-offs as rate-distortion trade-offs, we can now calculate the $R(D)$ curve for the source in figure 3a, using MSE as our distortion measure. The result is in figure 3b. This is what's going on in that curve: each point corresponds to a different *source codec*—that is to say, a pair of an encoder that takes every incoming source message M to a signal X, and a decoder that takes this signal to a decoded message, $\hat{M}$. The rate corresponds to the mutual information between incoming and decoded messages, $I(M;\hat{M})$, and the distortion (in this example) is the mean squared error between incoming and decoded messages. Distortion diminishes monotonically as rate grows, but the slope of the curve picks up the pace somewhat when the rate hits 2 bits—that is to say, once the encoder can use *four* different signals; this is the cross on the curve. Figure 3c summarizes what the codec is doing at that point: the encoder has a repertoire of four different signals, and, for example, signal `0` is sent whenever a sample corresponding to a point in the blue cluster is received. Signal `0`, in its turn, is decoded as the centroid of the blue cluster. Analogously with the other three signals and the other three clusters.

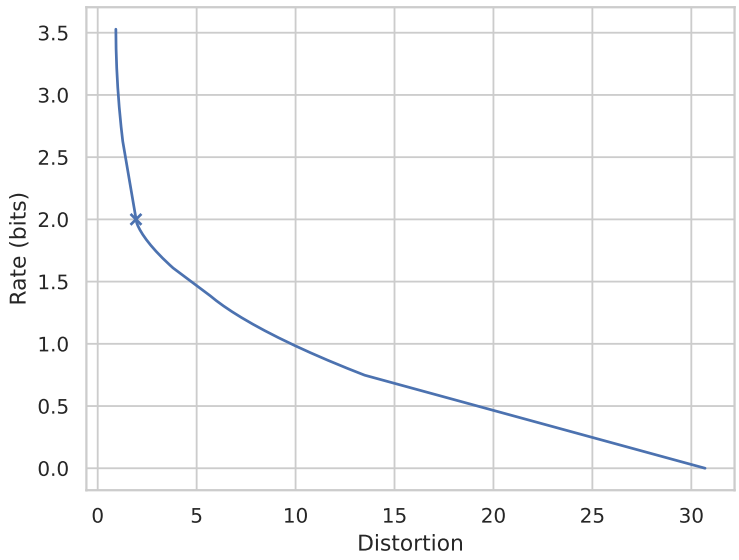
I claimed in the introduction that conceptual atoms and prototypes are not incompatible, and in fact participate jointly in efficient strategies of information transmission. The behavior of the codec in figure 3c provides a concrete illustration of this. First, it is a paradigmatic example of prototype-based categorization: each incoming sample, s, is encoded to a signal that, in turn, is decoded as s's closest prototype (one of the four cluster centroids). The rule the encoder is using can be summarized as follows: *encode the incoming sample using the signal corresponding to its closest prototype*. The encoder is effectively classifying (encoding) s under a concept (a signal) that will subsequently be decoded as its closest prototype. Among the samples that are closest to prototype p than to any other prototype, some are closer to p than others (that is to say, some are closer to the centroid of their cluster than others): explanations of typicality effects can rely on this fact just as much as they do in traditional prototype theory.

Second, optimal information transmission at this particular point in the rate-distortion curve is achieved with just four *atomic* signals. Note that this is not merely a consequence of the constraint that the rate at this point has to be 2 bits. There are indefinitely many ways to achieve a rate of 2 bits with more than four signals (although not with less than four): they involve probabilistically encoding individual samples to two or more signals (say, "toss a fair coin; if heads, encode this sample as signal 1, if tails, encode it as signal 2"). It might have seemed plausible that having more available signals, perhaps even a continuum of them (while keeping rate fixed) might help reducing distortion: say, having forty signals to play with, even if we have to restrict ourselves to 2 bits in total, would seem to put us in an advantageous position compared to someone who has to restrict themselves to four signals. Somewhat surprisingly, that's not how things turn out. Four atomic signals are enough for optimality.

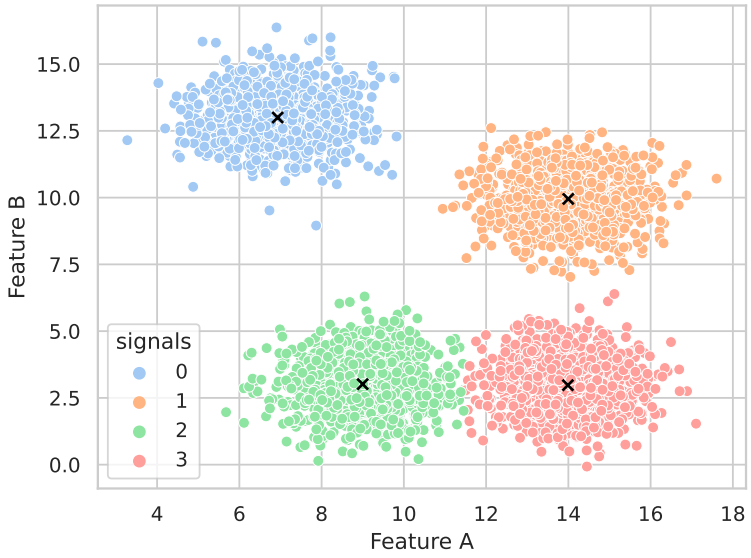
One way to see how and why this works is to focus on how different signals contribute to the $R(D)$ function. Figure 3d shows how optimal groups of 1, 2, 3, ..., up to 12 signals can be used to categorize our toy world. The figure shows the *slope* of the $R(D)$ curve


15.0
12.5
10.0
7.5
5.0
2.5
0.0
Feature B
4
6
8
10
12
14
16
18
Feature A

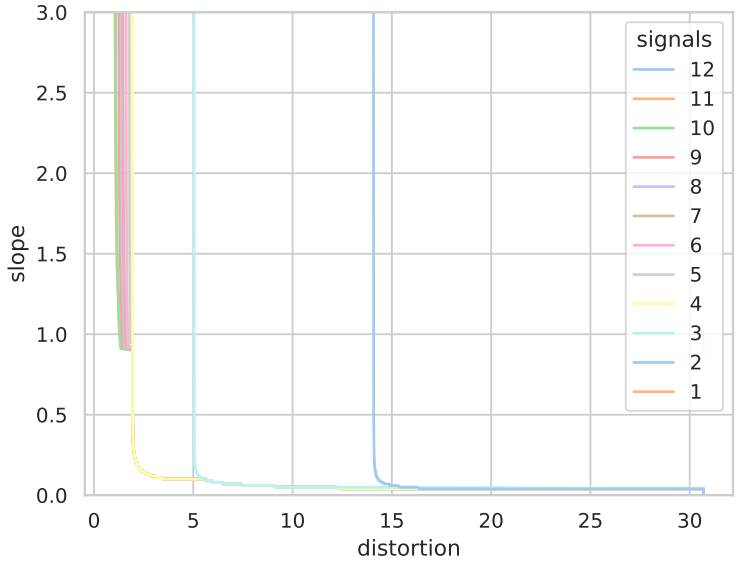
(a)


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rate (bits)
0
5
10
15
20
25
30
Distortion

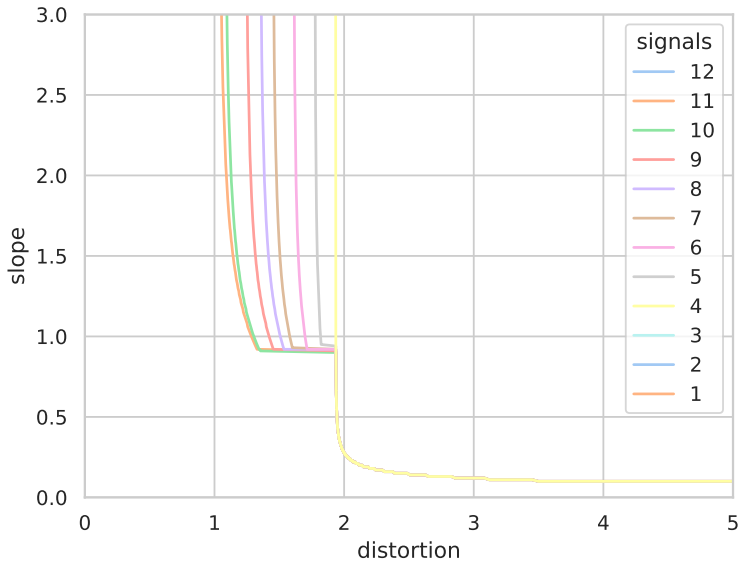
(b)


Feature B
Feature A
signals
0
1
2
3
15.0
12.5
10.0
7.5
5.0
2.5
0.0
4
6
8
10
12
14
16
18

(c)


slope
distortion
signals
12
11
10
9
8
7
6
5
4
3
2
1
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20
25
30

(d)



(e)

Figure 3: Categorizing a mixture of Gaussians with atoms and prototypes. 3a: The stimulus set is 4000 points coming from an equiprobable mixture of four bivariate Gaussian distributions. Each point corresponds to a combination of two feature values (the x and y coordinates). 3b: The rate-distortion curve for the source in figure 3a and a mean square error distortion measure. The cross marks the rate-distortion of the optimal 2-bit codec (four signals), which coincides with a certain change of slope. 3c: This 2-bit codec is shown here: each color represents points sent to the same signal. That signal, in turn, is decoded as the cross at the centroid of each group of points. 3d: The contribution of each new signal to the $R(D)$ curve. Each new signal takes the curve a bit further. The contribution made by larger groups overlaps that made by smaller groups. This happens until there are four signals, at which point no discrete group of signals is optimal. 3e: A close-up of the 'explosion' after four signals.

plotted against distortion. Reading the figure from right to left, we start with just one signal. This we cannot really see, as it corresponds to the rightmost point on the curve: with just one signal there can be no information transfer, and the rate is strictly zero.[13] With two signals (the first blue stretch, from right to left) we can account for a reduction of distortion from just above 30 to just below 15. After that, two signals exhaust their categorizing potential (that's why the slope shoots to infinity), and we need three signals to continue reducing distortion.

The interesting thing to note here is that the three-signal curve (and in fact all n-signal curves for $n > 2$) perfectly overlap the two-signal curve. As I said above, while we might have expected that a codec that utilizes three signals to communicate two bits (by being slightly inefficient with each signal) would be better than a two-bit two-signal codec, it is not. As long as the rate is below 1 bit, two signals are optimal. The same thing happens with four versus three signals. But beyond that point things change: once we have more than four signals, there are no longer groups of signals that are both rate-distortion optimal *and* discrete.[14] Four is the biggest such group. Figure 3e is a close-up of this transition from discrete to continuous.

The fact that one can minimize distortion at a certain rate with atomic (discrete) signals is not a peculiarity of this example. In general, if the distortion measure is the MSE, it can be shown that, unless we are working in high rate / low distortion regimes, atomic signals are enough to meet the rate-distortion optimum (Rose 1994, sec. III).[15] In particular, for mixtures of Gaussians such as our toy world, the rate-distortion optimum can be achieved with atomic signals up until all sources of variation (all different Gaussians) have been accounted for. This is, precisely, the point marked with a cross in figure 3b which I have been discussing.[16]

What I take to be the most important lesson of the example is this: I have not had to *posit* atoms and prototypes. They have emerged naturally as a solution to the problem of transmitting information about samples, under two mild constraints: MSE as a distortion measure, and a regime of relatively high distortion (Rose 1994). The results linking atomicity to regimes where information transmission happens at very low rates (see *ibid.*) suggest that concepts can afford to be atomic at least partly because they are, precisely, signals that convey the gist of a class, while aggressively disregarding finer details.[17]

[13] The way we count stuff in information theory, one signal and its absence would be two signals.

[14] This is related to the fact that, in the rate-distortion-optimal way of clustering, cluster-splitting happens "along the principal axis of the cluster" (Rose 1998, 2216). Once we have accounted for all isotropic Gaussians there are no principal components left, and all directions are equal.

[15] More precisely, if the so-called *Shannon lower bound* [SLB] on $R(D)$ is not tight, then the lowest achievable distortion at any given rate can be achieved with atomic signals. For an introduction to the SLB, see Gray (1990, chap. 4). Shannon introduced this notion in his (1959). For more on the conditions under which the SLB is tight, see Linder and Zamir (1994), Koch (2016).

[16] Section 3 of the Supplementary Material presents a case in which there are no limits to the rate-distortion optimality of discrete sets of signals, precisely because the sources of variation are not Gaussian (but rather depend on a uniform probability distribution.)

[17] It is suggestive to think of the codec in figure 3c as a *prototype denoiser*: we can see the four clusters in figure 3a as composed of noisy versions of the four centroids, which the four signals (concepts) clean

2.2 The Discrete Case

In fact, the MSE-distortion constraint can often be relaxed as well: consider now a different "stimulus set", this time constructed out of a set of nine categorical features $F_1, \ldots, F_9$. In our stimulus set, they will be binary features, that can be simply "on" or "off", present or absent. So as to have a concrete example in mind, we could think of these features as being of the kind birds may or may not have, such as, e.g., HAS WINGS, which could be present (+HAS WINGS) or absent (-HAS WINGS). Other such features are FLIES, HAS FEATHERS, or HATCHES EGGS (Hampton 2006). The instantiation of each of these nine features replicates, noisily, the state of a central, hidden node which is instantiated at random. You can think of it as some kind of probabilistic essence, perhaps, as in Boyd's homeostatic property clusters (1999). Figure 4a shows the very simple structure of this class as a graph. Specifically, the probabilities of instantiations of nodes in the graph in figure 4a are:

- $\Pr(+\text{HIDDEN}) = \Pr(\text{-HIDDEN}) = .5$

And, for all i,

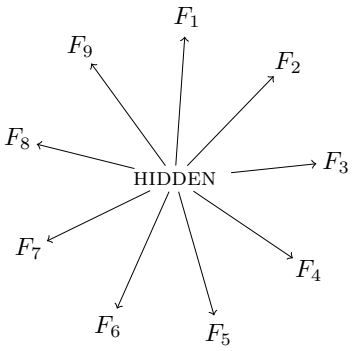
- $\Pr(+F_i|+\text{HIDDEN}) = \Pr(-F_i|\text{-HIDDEN}) = .95$
- $\Pr(+F_i|\text{-HIDDEN}) = \Pr(-F_i|+\text{HIDDEN}) = .05$

Here, each sample can be thought of as a binary vector with nine entries, such as, e.g., [0, 0, 1, 1, 0, 1, 0, 1, 1]. For each entry, 1 means that the corresponding feature is present, and 0 that it is not. The naïve method of storing or transmitting this information requires, therefore, 9 bits. The entropy of this source is, in fact, not 9 but ~3.6 bits, though, because features are far from independent from one another. But we can compress further than this, if we are ready to accept some distortion. Because we are dealing with categorical data, we cannot use MSEs to measure our distortion. One common alternative for discrete sources is the so-called *Hamming distortion*, which simply counts the number of differences between original and decoded vectors, and then normalizes.[18] Figure 4b shows the $R(D)$ curve for this stimulus set under the Hamming distortion. Here, too, there is a comparatively sudden change of slope—at 1 bit this time. The explanation is entirely analogous to the previous example: 1 bit is all you need to account for the main source of variation (the hidden node, in this case), and the rest, literally, is noise.

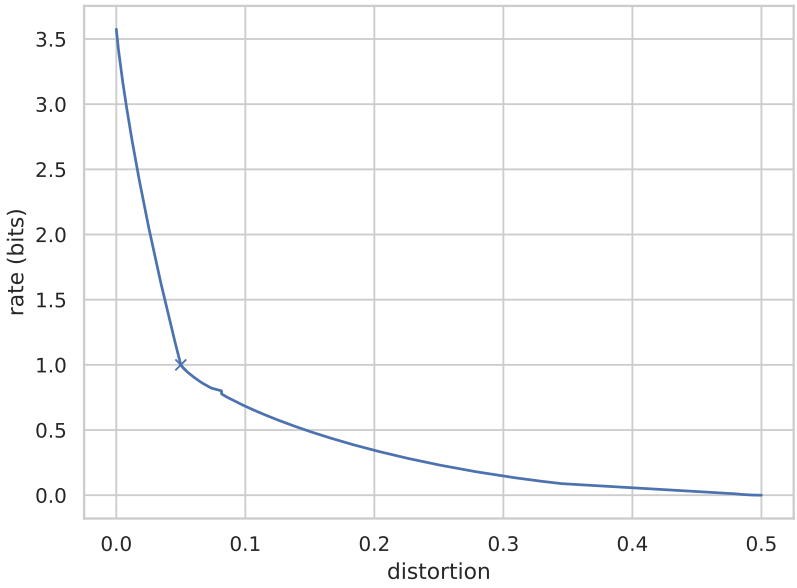
Encoder and decoder at the cross in figure 4b are, again, relying on two prototypes: on the one hand, the all-ones vector (you can think of this as the most typical member of the class (the prototypical bird, with all of its usual birdy features); on the other, the all-zeros vector (something like the "prototypical absence" of a class member: no birdy features at all). The encoder sends a different signal depending on which of these two

and recover. This perhaps partly explains why thinking of concepts as ideal versions of real-world samples, from Plato's *Phaedo* to Barsalou (1985), has often seemed attractive.

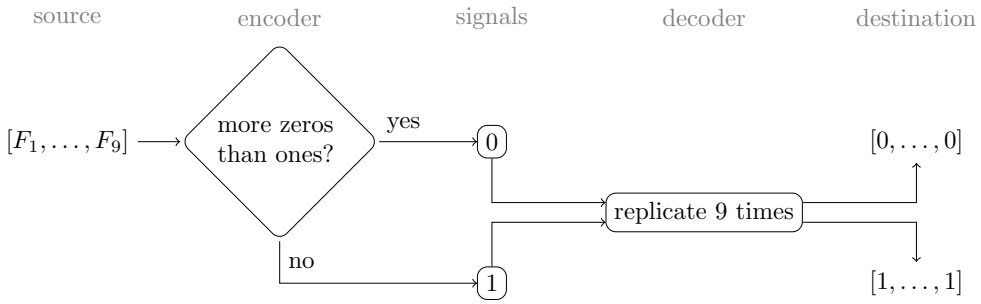
[18] For illustration, if [0, 0, 1, 1, 0, 1, 0, 1, 1] were to be decoded as [1, 1, 1, 1, 1, 1, 1, 1, 1], the Hamming distortion would be $\frac{4}{9}$: 4 mistakes in 9 entries.


F_1
F_9
F_2
F_8
F_3
HIDDEN
F_7
F_4
F_6
F_5

(a)


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
rate (bits)
0.0
0.1
0.2
0.3
0.4
0.5
distortion

(b)



(c)

Figure 4: Categorizing a cluster of categorical features. 4a: A model of a class with nine categorical features that noisily replicate the state of a hidden node. 4b: The $R(D)$ curve for the class in figure 4a and Hamming distortion. The cross marks a comparatively sudden change of slope. (The small hump to the right of the cross is noise in the numerical approximation.) 4c: 1-bit codec that attains the rate-distortion pair at the cross of the $R(D)$ curve in figure 4b.

vectors is closest to the sample received. This signal is in turn decoded as its associated prototype. This is, again, an example of cooperation of atomic signals (two of them, in this case) with prototypes in providing efficient solutions to information-transmission problems.[19]

As we have seen, Rosch's cognitive-economy principle presents a multiobjective optimization problem (optimize *both* information about the environment *and* resource expenditure) which is, therefore, underdetermined: multiobjective optimization problems are "solved" by providing a *Pareto frontier*—the set of solutions such that you cannot improve one of the objectives (say, information about the environment) without worsening the other (say, resource expenditure). The discussion so far in this section can be read as an argument that the $R(D)$ curve is a compelling formalization of at least an important aspect of the cognitive-economy Pareto frontier.[20] Furthermore, as we have also seen, not all points

[19] It is also interesting to note that, while the encoder only sees the surface features F_i, the signal most closely correlates with none of them, but with the hidden node. The codec is recovering the causal structure of its class by compressing it.

[20] In this paper I am not distinguishing between memory and channel capacity on the one hand (these are the kinds of resources that information theory concerns itself with), and computational complexity (Rooij et al. 2019; Arora and Barak 2009; Li and Vitányi 2008) on the other hand. Complexity is as central a "resource", in Rosch's sense, as memory or capacity. In particular, the main reasons to prefer atomic signals to, say, a probability distribution of continuous signals, all else being equal, are complexity-related ones: a repertoire of (say) four signals is computationally a much simpler object than a probability distribution over a space of signals. In this paper I am focusing on information-theoretic constraints, but a full evaluation of how cognitive-economy-related considerations should inform our theories of concepts will need to treat computational complexity independently as a third optimization objective, alongside rate and distortion.

in the $R(D)$ curve are equal. In both the examples discussed so far in this section (the Gaussian mixture in figure 3a and the cluster of categorical features in figure 4a), there is a change of slope, an elbow, that corresponds to the point at which all of the main sources of variations have been accounted for (each of the Gaussians in the first example, the hidden node in the second), and the remaining distortion corresponds to noise (cf. Martínez 2019b). In these two examples, this elbow was also the most informative point for which atomic signals are optimal.[21]

In the optimization problems in particular that I have examined here, the elbow of the $R(D)$ curve (the points marked with a cross in figures 3b and 4b) offer excellent cognitive-economic compromises. For the system portrayed in figure 4b, as we saw, zero distortion can only be achieved with ~3.6 bits (this is the entropy of the stimulus set), and the maximum distortion (at zero rate) is 0.5 (this is the best expected distortion you can get when you are simply guessing the sample). Yet the distortion at rate 1 bit (i.e., at the cross) is .05. That is to say, with only $\frac{1}{3.6} = 28\%$ of all the rate you can throw at this problem, you get from 50% distortion to 5% distortion—a 90% improvement. For the system portrayed in figure 3b, the least expected distortion you can get at rate 0 is around 30.6: this is the distortion when you have to guess the sample without any information, and corresponds to the expected squared distance to the centroid of the whole stimulus set. With the codec in figure 3c, on the other hand, we attain a distortion of ~1.93 with 2 bits. That is a reduction of distortion of 96%. In this example, furthermore, getting the distortion all the way to zero essentially requires as much entropy as there are data points; in our case, approximately 12 bits for 4000 samples.

2.3 Conceptual Spaces

A word on how the above continuous and discrete toy models relate to work on "conceptual spaces", as developed by Peter Gärdenfors (2000; see also Chella, Frixione, and Gaglio 2001; Millikan 2017, among many others). The main asumptions embodied in the above models are that

- Samples to be classified are points in an abstract n-dimensional feature space; each point corresponding to a different combination of values of n different features.
- Treating a certain point p in feature space as if it was a different point p' instead incurs in a penalty (a "distortion") that, in the models above, is cashed out in terms of a *distance* between p and p': Euclidean for continuous feature values, Hamming for discrete ones.

[21] The elbow in the slope of the $R(D)$ curve need not always coincide with the minimum distortion achieved by discrete signals: they will not coincide, for example, if various Gaussians are close enough as to be unimodal. The fact remains, though, that in those cases the largest optimal, discrete set of signals has as many signals as there are independent Gaussians in the mixture.

An example of unimodality is presented in the Supplementary Material. I would like to thank an anonymous referee for prompting me to discuss this kind of case.

- Feature space is not uniformly occupied. There are regions that concentrate most of the probability of instantiation of samples to be classified, and regions that are mostly empty.

Seeing categorization behavior as relying on some pre-existing "psychological distance" among samples (and therefore, implicitly, seeing those samples as embedded in an abstract feature space) is a widespread modeling decision since at least Shepard (1957). Furthermore, the notion that feature space is not uniformly occupied, as I briefly discussed above, can be seen as a more general, more formally perspicuous way of cashing out what Eleanor Rosch, and many others following her, call "correlational world structure".

There are at least two ways in which the above discussion treats feature spaces in a way that is different from, and could fruitfully inform, work on the conceptual-spaces tradition. First, for Gärdenfors (and many other cognitive psychologists before and after him, including Rosch), feature space does not model how physical samples are, but how they are represented. That is to say, the space in question is a psychological entity—hence the talk of "conceptual" or "cognitive" (Bellmund et al. 2018) spaces.[22] In the above models no such assumption is made: they are agnostic as to whether feature space models the actual distribution of features of physical objects in a certain relevant domain and context; or instead models some internal representation thereof. In fact, much of the appeal of information-theoretic analyses comes from noting that resource-efficient representation for categorization does not need a psychological space, fully populated with samples; but that a handful of prototypes is often enough.

A second important way in which the above models differ from conceptual-spaces developments of the idea of a feature space is that, e.g., Gärdenfors (2000) makes several assumptions as to what conditions a region of feature space needs to meet in order to fall under a single concept. Importantly, he claims that such regions need to be convex (Gärdenfors 2000, chap. 3). I, on the other hand, have not made any such assumptions: regions of space mapped to each prototype, indeed, come out convex for the Euclidean and Hamming distances I have utilized here—but this, and the very presence of prototypes, are side effects of the process of optimizing a rate-distortion trade-off, not put in by hand.[23]

In fact, it is entirely possible to devise ecologically plausible distortion measures such that the related rate-distortion-optimal concepts are not convex. For example, if the distortion in question is relative to the distance to a single designated focal point, then

[22] Gärdenfors (2000, sec. 1.4) distinguishes between "phenomenal" and "scientific" spaces, where the latter are best conceived as objective, non-mental entities (such as, e.g., literal Newtonian space.) In any event, in his discussions of categorization he always takes the relevant spaces to be psychological.

[23] The existence of an optimal and discrete set of signals (a set of atoms) does depend on feature space being "clumpy" (Millikan 2017, chap. 1), but the optimality of convex regions around propototypes does not: rate-distortion-optimal categorization of any feature space under an MSE distortion measure will result in a Voronoi tessellation (cf. Jäger and Van Rooij 2007). See the Supplementary Material for an example of this in a dataset sampled from an uniform probability distribution.

the optimal categories are (non-convex) concentrical bands around that focal point.[24] Investigating categorization from the point of view of information-processing efficiency reveals possibilities that other treatments of conceptual spaces may be prone to overlook.

In this section I have shown how atoms and prototypes, two standard components of the psychologist's categorization toolbox, emerge naturally from the trade-off between faithfulness and resource expenditure that Eleanor Rosch called "cognitive economy". In the following section I show how, beyond shedding light on the phenomenon of categorization in general, rate-distortion analyses can also illuminate other features of our categorization behavior; in particular, some aspects of supervised categorization that are sometimes interpreted as demonstrating a shift from reliance on prototypes to reliance on exemplars.

3 Supervised Categorization

Exemplars are actual instances of a class—actual birds, cats or chairs. In exemplar-based models of categorization, the class to which a certain sample belongs is decided by calculating its distance to those exemplars, not to a prototype (E. E. Smith and Medin 1999; T. Griffiths et al. 2007). I should first note that the difference between exemplar- and prototype-based models is often not as momentous as one might initially think, and as the literature sometimes makes it out to be. Many of the classes that psychologists focus on (because they appear to be the kinds of classes we care most about) are highly correlational in Rosch's sense: instances of the class do not uniformly occupy feature space, but are confined, with high probability, to small regions, or low-dimensional manifolds, of feature space. That is to say, often, randomly picking an exemplar will land you close to a typical member of the class; consequently, categorization based on a random exemplar will typically be close to categorization based on a prototype. For example, if the probability distribution of a one-dimensional stimulus set is Gaussian, ~68% of exemplars are less than one standard deviation away from the mean (the prototype), and ~99.7% less than three standard deviations away. If our exemplar-based categorization is based on the expected distance to n exemplars, exemplar- and prototype-based categorization become more and more similar the larger n is, and indistinguishable in the limit.

I will not develop these observations here. In any event, leaving aside their behavior in the limit, categorization models relying on exemplars and prototypes can make very different predictions when the classes they are dealing with are small, or when they do not closely align with the structure of feature space. The two models discussed in section 2 can be seen as instances of *unsupervised categorization*: I only fixed the probabilistic

[24] A distortion like this might plausibly be relevant, e.g., to sports such as golf (where the focal point would be the hole) or basketball (where it would be the basket). I present a model of this kind of situation in the Supplementary Material. It should be possible to investigate empirically whether enforcing this kind of distortion measure in a laboratory task results in the emergence of non-convex categorization behavior.

structure of the stimulus set (the source) and what counts as more or less faithful decoding (the distortion measure), and categorization took care of itself. The resulting categories are comparatively natural, in that they are exploiting source structure to find efficient solutions to the rate-distortion trade-off. But much of the debate on the relation between prototype- and exemplar-based categorization depends on classes being antecedently defined by the researcher, in ways which do not necessarily exploit this structure, or that go against its grain.

The example I will discuss here (J. D. Smith and Minda 1998; but I learned about it from T. L. Griffiths et al. 2011) relies on the artificial classes given in Table 1. We can think of these classes as emerging from adding a small amount of noise to `000000` for class A and `111111` for class B, and *then* swapping one of the members of each class with one another (those would be the last members in each class enumeration). The resulting classes have an odd member out each. When human subjects try to learn these categories, they follow the pattern in figure 5a: the odd ones out are incorrectly classified with what would be their "natural" classes, and only after some learning do they start moving to the correct ones.

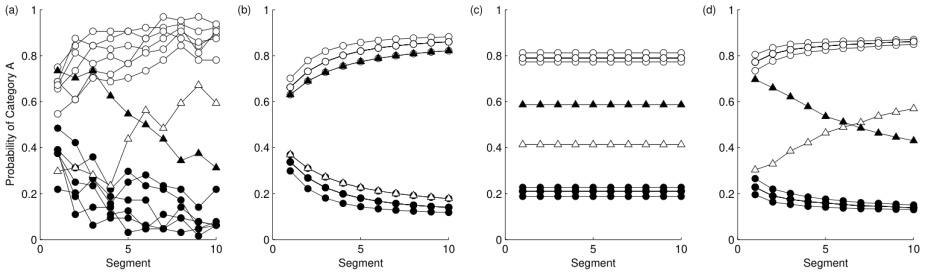


Figure 5: Learning non-linearly separable categories. Reproduced from T. L. Griffiths et al. (2011)

Table 1: The two linearly non-separable classes in J. D. Smith and Minda (1998). The "odd ones out" are the last elements in each column.

A	B
000000	111111
100000	011111
010000	101111
001000	110111
000010	111011
000001	111110
111101	000100

5b and 5c show the behavior of prototype- and exemplar-based categorizers respectively.

None of them adequately captures the gist of human categorization: the prototype model always categorizes the odd ones out with their natural classes, and the exemplar model never does, and hence fails to cluster them with their natural classes during the early training segments. The way these results were interpreted in J. D. Smith and Minda (1998), subjects can be seen as first employing a prototype-based strategy and, after some learning, switching to an exemplar-based strategy.[25]

As it happens, just like in the previous section, the behavior of human categorizers can be explained directly as the result of rate-distortion optimization. I first turn the two classes in Table 1 into a single source by adding the class each sample belongs to as an extra feature (following Anderson 1990, 99, and many others). See Table 2. I will also assume that all stimuli are equiprobable, as each was presented an equal number of times in the original experiment, but of course this could be modified as needed.

Table 2: Representing the two classes in Table 1 as a single source. The class each sample belongs to is represented as an extra feature (`0` for class A and `1` for class B, in red).

0000000
1000000
0100000
0010000
0000100
0000010
1111010
1111111
0111111
1011111
1101111
1110111
1111101
0001001

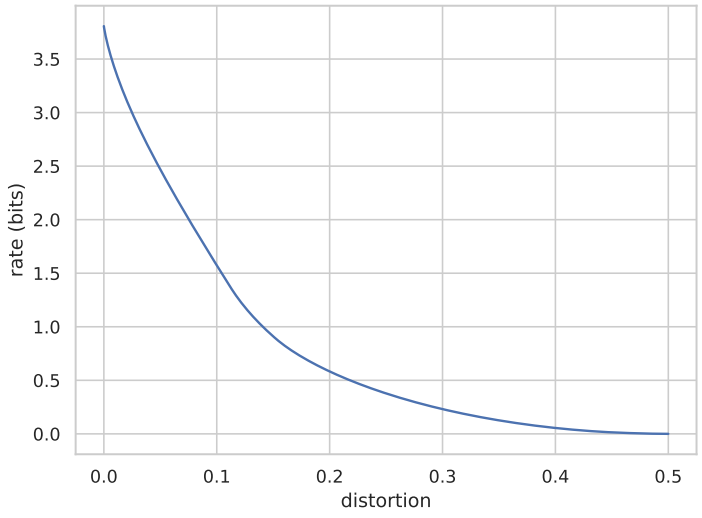
[25] 5d records the behavior of a Dirichlet-process mixture model (details in T. L. Griffiths et al. 2011), which is able to capture the crisscrossing pattern typical of human data. The rate-distortion approach I am exploring in this paper comes to this problem from a very different, perhaps ecologically more basic perspective: not (as in the work by Griffiths and colleagues) by trying to model a probabilistic source, but by trying to transmit information from perception to behavior.

Dasgupta and Griffiths (2022) is a recent introduction to non-parametric Bayesian approaches to categorization. Other approaches that, like mine, view prototype-exemplar transitions as gradual, and not a sharp substitution of one categorizing strategy by another are the SUSTAIN model (Love, Medin, and Gureckis 2004) and the varying abstraction model (Vanpaemel and Storms 2010). None of these models gets to categorization behavior purely from information-theoretic considerations, but comparing them in detail with the rate-distortion approach is matter for future research.

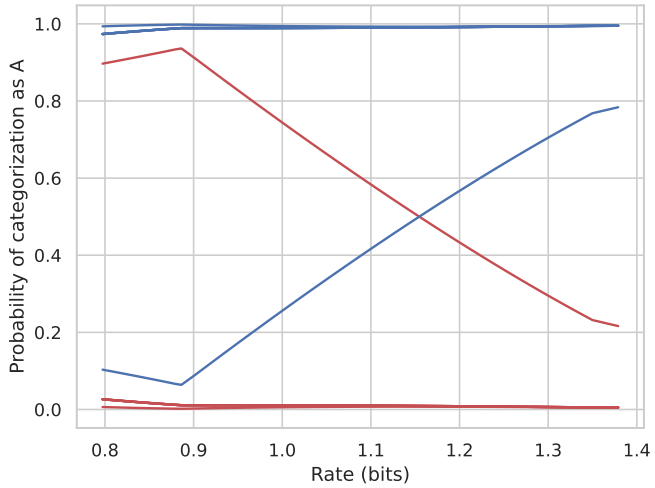
We now have a source of binary strings. As we did in §2.2, we can explore what happens as we try to transmit as much information about stimuli as possible (including the last bit with the class they belong to), at different rates, quantifying faithfulness with Hamming distortion. The rate-distortion curve for this exercise is in figure 6a. Here, in particular, we are interested in how different samples are classified. This can be calculated by focusing on the last bit of the stimuli (which corresponds to the preassigned class; see Table 2), and keeping track of whether, and how, it changes after passing through the codec. So, for example, if 0111111 is decoded as 0111110 with a probability of .6, as 1110111 with a probability of .2, and as 0111111 with a probability of .2, we'll say that the original sample is categorized correctly with a probability of .4 (corresponding to the sum of the probabilities of the two ways in which the last bit is decoded correctly). Figure 6b shows what happens when we do this exercise for all samples, using the optimal codec at each rate from .8 to 1.5 bits.

Here too we find the familiar pattern in which all samples are consistently categorized into the correct classes, except for the two odd ones out, which are initially categorized with the classes that would correspond to them as if a prototype was governing the process, and only later are assigned to their correct class, as in exemplar-based categorization. Why does this behavior emerge? Recall that the x-axis measures the rate at which information is transmitted. That is to say, it measures the amount of information about samples that can be used in the categorization decision. At low rates (i.e., around 1 bit at the left end of the plot) there is barely enough information to losslessly transmit the value of a single binary feature, let alone seven of them (the six original ones plus the category feature). The codec therefore has to find a way to provide a gist of the stimulus in (less than) 1 bit, or a single binary feature, and rely on the statistics of the source to "puff up" this single feature into the seven features of the reconstructed stimulus. The result is not unlike the majority rule that the optimal codec in figure 4c relies on: send a `1` if the majority of features are `1`s, send a `0` otherwise—then copy the received signal seven times at the receiver side. This effectively sees the source as a collection of noisy departures from the all `1`s and all `0`s vectors, which aligns with the externally enforced classes (the seventh feature) very well, except, of course, for the two odd ones out. This is why the probability of their being misclassified is very high. As we increase rate (as we move to the right along the x-axis) we gain expressive power and can start accommodating the odd ones out with their own signal, at least probabilistically. That's how the probability of correct classification grows, until we hit the entropy of the source and all samples are classified correctly (at ~2.8 bits, well to the right of the region shown in figure 6b.)

In their description of the Smith and Minda experiment, Griffiths and colleagues claim that "a prototype model was found to provide a better explanation for human performance on a categorization task during the early stages of learning, while an exemplar model was found to be a better fit to the later stages" (T. L. Griffiths et al. 2011, 190f). We have seen that, in fact, capturing the gist of human performance in the experiment just requires a system that aims at minimizing distortion at different rates. Such a system may know nothing of prototypes or exemplars, but will display qualitatively equivalent behavior.



(a)



(b)

Figure 6: A rate-distortion analysis of the Smith-Minda experiment: 6a: Rate-distortion curve for the source in Table 2 and Hamming distortion. 6b: Classification of samples in classes by the optimal codec at each rate.

Many critiques of the intended interpretation of the Smith and Minda experiment, according to which it provides evidence of a "representational shift" (Johansen and Palmeri 2002) from prototypes to exemplars as category learning progresses, point out that prototype-like behavior in the early segments of training can be just as well explained by subjects focusing their attention on one, or a few, of the more highly predictive features of the stimuli (Nosofsky and Zaki 2002, 938; Johansen and Palmeri 2002, 531; see also Nosofsky 1986 for more on this "attention-optimization" idea; Nosofsky cites Reed 1972; and Shepard, Hovland, and Jenkins 1961 as early suggestions along similar lines.) But, if the values of different features are highly correlated (as they are in the Smith and Minda experiment, and as demanded by Rosch's principle of perceived world structure,) this is equivalent to saying that in the early stages of category learning subjects are using low-rate coders to categorize stimuli.[26] In §2 I showed that categorization against a discrete number of prototypes can be seen as emergent behavior that results from rate-distortion-efficient coding at low rates. Under that perspective, protesting that, instead of prototype-based categorization, what we have is categorization based on one or a few features is somewhat arbitrary: both are, under the relevant circumstances, largely equivalent ways of describing rate-distortion-efficient behavior at low rates.

This does not mean that prototype- and exemplar-based models are somehow irrelevant or misguided, of course. For one thing, they aim at capturing not just the gist, but the actual numerical detail of human performance, which is why they have various tunable parameters while the rate-distortion analysis I have presented here has none.[27] For another, they can be seen as the way cognitive systems approximate rate-distortion optima: they provide much needed implementational detail to the purely abstract "solution" offered by rate-distortion analysis. The same can be said about the more sophisticated Dirichlet-process approach to categorization in T. L. Griffiths et al. (2011).

My point is, rather, that a picture of human categorization performance in which categorizers come to the task with a repertoire of tools (prototypes and exemplars, among others) and then, somewhat fancifully, switch from one to another as the task progresses risks missing the forest for the trees. What happens is that the coding strategy that optimally minimizes distortion evolves as rate increases. It is fine to interpret this evolution as a switch from prototypes to exemplars, if one remembers that what is

[26] More precisely, focusing one's attention on one or a few dimensions is a sufficient, but not necessary, condition for implementing a low-rate coder: it is theoretically possible, for example, that the rate-distortion optimal 1-bit coder need to be calculated by taking into account two or more stimulus dimensions. This will happen if each such dimension is not very predictive of the class the stimulus belongs to, but the two of them together are. That is, if they carry information about their class *synergistically* (Williams and Beer 2010; Wibral et al. 2017; Martínez 2020). It should be possible to test empirically whether these considerations of informational efficiency make a contribution to explaining categorization behavior, over and above the purported broadening of attentional scope from one to more dimensions. Developing these ideas is matter for another paper.

[27] For example, in J. D. Smith and Minda (1998, 1414), there is an "attentional weight", w_k, for each of the k features that to-be-classified items and exemplars share, and a "sensitivity parameter", c, that governs the whole process of categorization. These $k + 1$ parameters are set so as to maximize fit with human performance.

driving the process, at a higher level of abstraction, is a uniform process of optimizing rate-distortion trade-offs.

One important assumption I have made in this section is that learning, of the sort subjects undergo in the Smith-Minda experiment as they go through task segments, results in higher rate in the flow of information between input samples and their decoded reconstructions (again see figure 1). That is to say, *learning to do a task, among other things, consists in a widening of the informational bottleneck between the random variables that describe inputs to the task, sensory or otherwise, and the random variables that describe task-related behavior.* This seems to capture an important aspect of what learning consists in.

4 Conclusion

In this paper I have, first, intervened in the debate between atomists and informationists about concepts. I have argued that, far from being alternative hypotheses as to the nature of concepts (and *a fortiori* far from being incompatible) both atoms and bodies of information are jointly useful for efficient transmission or storage of information about a class.

For prototypes to emerge in efficient transmission, though, one needs the world to be relevantly similar to the mixture of Gaussians in Fig. 3: that is to say, the world needs to be sufficiently "clumpy", in Millikan's (2017, chap. 1) sense; or, more or less equivalently, show correlational world structure in Rosch's (1999) sense; or, also more or less equivalently, be composed out of property clusters in Boyd's sense (1999; see also Slater 2015; Martínez 2015, among many others). One can see all of these attempts at characterizing the metaphysics of knowable worlds as ways of ensuring that those worlds are *compressible*—and that, furthermore, their associated rate-distortion function shows the kind of elbow we see in Figs. 3b and 4b.

For atoms to emerge, we also need to be working at relatively low rates: in particular, in the case of Gaussian mixtures, we can optimally transmit information with atoms insofar as we are content with the level of distortion that comes from simply ignoring Gaussian dispersion around its mean: that is to say, a maximum n atoms for a mixture of n Gaussians. This fact (proven by Rose 1994) sheds light on two intuitive properties of conceptual repertoires: first, conceptual repertoires are *comparatively small*, and certainly smaller than what we take to be the repertoire of possible ("nonconceptual") perceptual contents. Second, concepts are sometimes taken to be closely related to idealized versions of samples in their target class plausibly because of their being tightly related to the centroids of more or less Gaussian regions in feature space.

I have also shown how thinking of concepts in the context of processes of information transmission helps explain apparently unrelated data about human categorization performance: the claimed substitution of prototypes by exemplars in J. D. Smith and Minda

(1998). This result further showcases the explanatory usefulness of information-processing (and in particular rate-distortion) models and analyses of concepts and categorization.

Obviously, none of the above entails that other theoretical approaches to concepts, and in particular classical prototype- and exemplar-based theories, are without merit. There is a lot that the analyses in this paper do not explain, from the actual detail of human categorization performance, to the actual detail of how concepts are learned. For these other ends, a parametrized theory, which can be fit to numerical data, is needed. My aim has been, rather, to show that a good deal of what would perhaps appear to be surprising features of concepts in fact fall right off the way efficient transmission of information needs to behave.

Acknowledgements

I would like to thank Nick Shea, James Hampton, the Buenas Migas work in progress group, three reviewers for this journal, and audiences in Barcelona, Durham, Düsseldorf, and New York for very helpful feedback.

This work has been funded by the Spanish Ministry of Science and Innovation, through grants PID2021-127046NA-I00 and CEX2021-001169-M (MCIN/AEI/10.13039/501100011033); and by the Generalitat de Catalunya, through grant 2017-SGR-63.

References

Anderson, John R. 1990. *The Adaptive Character of Thought*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Arimoto, Suguru. 1972. "An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels." *IEEE Transactions on Information Theory* 18 (1): 14–20.

Arora, Sanjeev, and Boaz Barak. 2009. *Computational Complexity*. Cambridge University Press.

Barsalou, Lawrence W. 1985. "Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure in Categories." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11 (4): 629.

Bellmund, Jacob L. S., Peter Gärdenfors, Edvard I. Moser, and Christian F. Doeller. 2018. "Navigating Cognition: Spatial Codes for Human Thinking." *Science* 362 (6415): eaat6766. https://doi.org/10.1126/science.aat6766.

Berger, Toby. 1971. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall Series in Information and System Sciences. Inglewood Cliffs, New Jersey: Prentice-Hall.

Blahut, Richard. 1972. "Computation of Channel Capacity and Rate-Distortion Functions." *IEEE Transactions on Information Theory* 18 (4): 460–73.

Boyd, Richard. 1999. "Homeostasis, Species, and Higher Taxa." In *Species: New Interdisciplinary Essays*, edited by R A Wilson, 141–85. Mit Press.

Chella, Antonio, Marcello Frixione, and Salvatore Gaglio. 2001. "Conceptual Spaces for Computer Vision Representations." *Artificial Intelligence Review* 16 (2): 137–52. https://doi.org/10.1023/A:1011658027344.

Connolly, Andrew C., Jerry A. Fodor, Lila R. Gleitman, and Henry Gleitman. 2007. "Why Stereotypes Don't Even Make Good Defaults." *Cognition* 103 (1): 1–22. https://doi.org/10.1016/j.cognition.2006.02.005.

Cover, T. M., and J. A. Thomas. 2006. *Elements of Information Theory*. New York: Wiley.

Dasgupta, Ishita, and Thomas L. Griffiths. 2022. "Clustering and the Efficient Use of Cognitive Resources." *Journal of Mathematical Psychology* 109 (August): 102675. https://doi.org/10.1016/j.jmp.2022.102675.

Eubanks, Philip. 2001. "Understanding Metaphors for Writing: In Defense of the Conduit Metaphor." *College Composition and Communication* 53 (1): 92. https://doi.org/10.2307/359064.

Fodor, Jerry A. 1980. *The Language of Thought*. 1 edition. Cambridge, Mass: Harvard University Press.

———. 2008. *LOT 2: The Language of Thought Revisited*. Oxford Clarendon Press.

Fodor, Jerry A., and Ernest Lepore. 1996. "The Red Herring and the Pet Fish: Why Concepts Still Can't Be Prototypes." *Cognition* 58 (2): 253–70. https://doi.org/10.1016/0010-0277(95)00694-X.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. MIT press.

Gray, Robert M. 1990. *Source Coding Theory*. The Springer International Series in Engineering and Computer Science. Springer US. https://doi.org/10.1007/978-1-4613-1643-5.

Griffiths, Thomas L., Adam Sanborn, K R Canini, D J Navarro, and J B Tenenbaum. 2011. "Nonparametric Bayesian Models of Categorization." In *Formal Approaches in Categorization*, edited by Emmanuel M. Pothos and Andy J. Wills, 173–98.

Griffiths, Tom, Kevin Canini, Adam Sanborn, and Danielle Navarro. 2007. "Unifying Rational Models of Categorization via the Hierarchical Dirichlet Process."

Hampton, James A. 2006. "Concepts as Prototypes." *Psychology of Learning and Motivation* 46 (January): 79–113. https://doi.org/10.1016/S0079-7421(06)46003-5.

Jäger, Gerhard, and Robert Van Rooij. 2007. "Language Structure: Psychological and Social Constraints." *Synthese* 159 (1): 99. https://doi.org/10.1007/s11229-006-9073-5.

Johansen, Mark J, and Thomas J. Palmeri. 2002. "Are There Representational Shifts During Category Learning?" *Cognitive Psychology* 45 (4): 482–553. https://doi.org/10.1016/S0010-0285(02)00505-4.

Koch, Tobias. 2016. "The Shannon Lower Bound Is Asymptotically Tight." *IEEE Transactions on Information Theory* 62 (11): 6155–61.

Laurence, Stephen, and Eric Margolis. 1999. "Concepts and Cognitive Science." In *Concepts: Core Readings*, edited by Eric Margolis and Stephen Laurence, 3–81. Cambridge, MA.

Lewis, David. 1969–2008. *Convention: A Philosophical Study*. John Wiley & Sons.
Li, Ming, and Paul Vitányi. 2008. *An Introduction to Kolmogorov Complexity and Its Applications. Texts in Computer Science*. Vol. 9. Springer, New York,.
Lieder, Falk, and Thomas L. Griffiths. 2020. "Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences* 43: e1. https://doi.org/10.1017/S0140525X1900061X.
Linder, Tamas, and Ram Zamir. 1994. "On the Asymptotic Tightness of the Shannon Lower Bound." *IEEE Transactions on Information Theory* 40 (6): 2026–31.
Love, Bradley C., Douglas L. Medin, and Todd M. Gureckis. 2004. "SUSTAIN: A Network Model of Category Learning." *Psychological Review* 111 (2): 309–32. https://doi.org/10.1037/0033-295X.111.2.309.
Machery, Edouard. 2009. *Doing Without Concepts*. Oxford University Press.
MacKay, David JC. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.
Martínez, Manolo. 2015. "Informationally-Connected Property Clusters, and Polymorphism." *Biology and Philosophy* 30 (1): 99–117.
———. 2019a. "Deception as Cooperation." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 77 (October): 101184. https://doi.org/10.1016/j.shpsc.2019.101184.
———. 2019b. "Representations Are Rate-Distortion Sweet Spots." *Philosophy of Science* 86 (5): 1214–26. https://doi.org/10.1086/705493.
———. 2020. "Synergic Kinds." *Synthese* 197 (5): 1931–46. https://doi.org/10.1007/s11229-017-1480-2.
Medin, Douglas L., and Evan Heit. 1999. "Categorization." In *Cognitive Science*, edited by Benjamin Martin Bly and David E. Rumelhart, 99–143. San Diego: Academic Press. https://doi.org/10.1016/B978-012601730-4/50005-6.
Millikan, Ruth Garrett. 2017. *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford University Press.
Minda, John Paul, and J. David Smith. 2011. "Prototype Models of Categorization: Basic Formulation, Predictions, and Limitations." *Formal Approaches in Categorization*, 40–64.
Nosofsky, Robert M. 1986. "Attention, Similarity, and the Identification–Categorization Relationship." *Journal of Experimental Psychology: General* 115 (1): 39. https://doi.org/10.1037/0096-3445.115.1.39.
Nosofsky, Robert M., Thomas J. Palmeri, and Stephen C. McKinley. 1994. "Rule-Plus-Exception Model of Classification Learning." *Psychological Review* 101 (1): 53.
Nosofsky, Robert M., and Safa R. Zaki. 2002. "Exemplar and Prototype Models Revisited: Response Strategies, Selective Attention, and Stimulus Generalization." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28 (5): 924–40. https://doi.org/10.1037/0278-7393.28.5.924.
Osherson, Daniel N., Edward E. Smith, Ormond Wilkie, Alejandro Lopez, and Eldar Shafir. 1990. "Category-Based Induction." *Psychological Review* 97 (2): 185.
Reddy, Michael. 1979. "The Conduit Metaphor." *Metaphor and Thought* 2: 285–324.
Reed, Stephen K. 1972. "Pattern Recognition and Categorization." *Cognitive Psychology*

3 (3): 382–407. https://doi.org/10.1016/0010-0285(72)90014-X.

Rooij, Iris van, Mark Blokpoel, Johan Kwisthout, and Todd Wareham. 2019. *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis*. Cambridge ; New York, NY: Cambridge University Press.

Rosch, Eleanor. 1999. "Principles of Categorization." In *Concepts: Core Readings*, edited by Eric Margolis and Stephen Laurence, 189–206. The MIT Press.

Rose, Kenneth. 1994. "A Mapping Approach to Rate-Distortion Computation and Analysis." *IEEE Transactions on Information Theory* 40 (6): 1939–52.

———. 1998. "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems." *Proceedings of the IEEE* 86 (11): 2210–39. https://doi.org/10.1109/5.726788.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423.

———. 1959. "Coding Theorems for a Discrete Source with a Fidelity Criterion." *IRE Nat. Conv. Rec* 4 (142-163): 1.

Shea, Nicholas, Peter Godfrey-Smith, and Rosa Cao. 2017. "Content in Simple Signalling Systems." *The British Journal for the Philosophy of Science*.

Shepard, Roger N. 1957. "Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space." *Psychometrika* 22 (4): 325–45. https://doi.org/10.1007/BF02288967.

Shepard, Roger N., Carl I. Hovland, and Herbert M. Jenkins. 1961. "Learning and Memorization of Classifications." *Psychological Monographs: General and Applied* 75 (13): 1–42. https://doi.org/10.1037/h0093825.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics*, Swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information, 50 (3): 665–90. https://doi.org/10.1016/S0304-3932(03)00029-1.

Skyrms, Brian. 2010. *Signals: Evolution, Learning & Information*. New York: Oxford University Press.

Slater, Matthew H. 2015. "Natural Kindness." *British Journal for the Philosophy of Science* 66: 374–411.

Smith, Edward E., and Douglas L. Medin. 1999. "The Exemplar View." In *Concepts: Core Readings*, edited by Eric Margolis and Stephen Laurence, 207–22. The MIT Press. Bradford Books.

Smith, J. David, and John Paul Minda. 1998. "Prototypes in the Mist: The Early Epochs of Category Learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24 (6): 1411–36. https://doi.org/10.1037/0278-7393.24.6.1411.

Vanpaemel, Wolf, and Gert Storms. 2010. "Abstraction and Model Evaluation in Category Learning." *Behavior Research Methods* 42 (2): 421–37. https://doi.org/10.3758/BRM.42.2.421.

Wibral, Michael, Viola Priesemann, Jim W. Kay, Joseph T. Lizier, and William A. Phillips. 2017. "Partial Information Decomposition as a Unified Approach to the Specification of Neural Goal Functions." *Brain and Cognition*, Perspectives on Human Probabilistic Inferences and the 'Bayesian Brain', 112 (March): 25–38. https://doi.or

g/10.1016/j.bandc.2015.09.004.
Williams, Paul L., and Randall D. Beer. 2010. "Nonnegative Decomposition of Multivariate Information." http://arxiv.org/abs/1004.2515.
Zaslavsky, Noga, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. "Efficient Compression in Color Naming and Its Evolution." *Proceedings of the National Academy of Sciences*, July, 201800521. https://doi.org/10.1073/pnas.1800521115.

The Information-Processing Perspective on Categorization

Supplementary Material

Manolo Martínez

1 Different Variances

The example discussed in section §2 of the main document is a dataset sampled from a mixture of four Gaussians with the same variance. While the mathematical results on which the discussion in that section relies do require that Gaussians be isotropic (i.e., that they have a covariance matrix proportional to the identity matrix,) they do not require those variances to be equal (i.e., the proportionality constant may change from Gaussian to Gaussian.) I present here an example, fully analogous to the one discussed in §2 of the main paper, in which the equal-variance condition is relaxed.

The dataset is in fig. S1: a mixture of 5 Gaussians with different variances. The rate-distortion curve and the best codec with 5 signals (in figs. S2 and S3) are entirely analogous to those calculated in section §2.

2 Unimodal Gaussians

In the example discussed in §2.1 of the paper, the maximum number of discrete signals which still can be rate-distortion optimal is 4. This optimal codec also corresponds to a change of tendency (an "elbow") in the rate-distortion function. In general, the following two quantities need not coincide:

1. On the one hand, the maximum number of discrete signals that can be rate-distortion-optimal.
2. On the other, the number of signals at which the optimal codec corresponds to an elbow in the rate-distortion curve.

The first quantity still corresponds to the number of Gaussians in the mixture. This is just the straightforward consequence of the results proven in (Rose 1994, sec. III): accounting for all Gaussian sources of variation (i.e., for our current purposes, placing

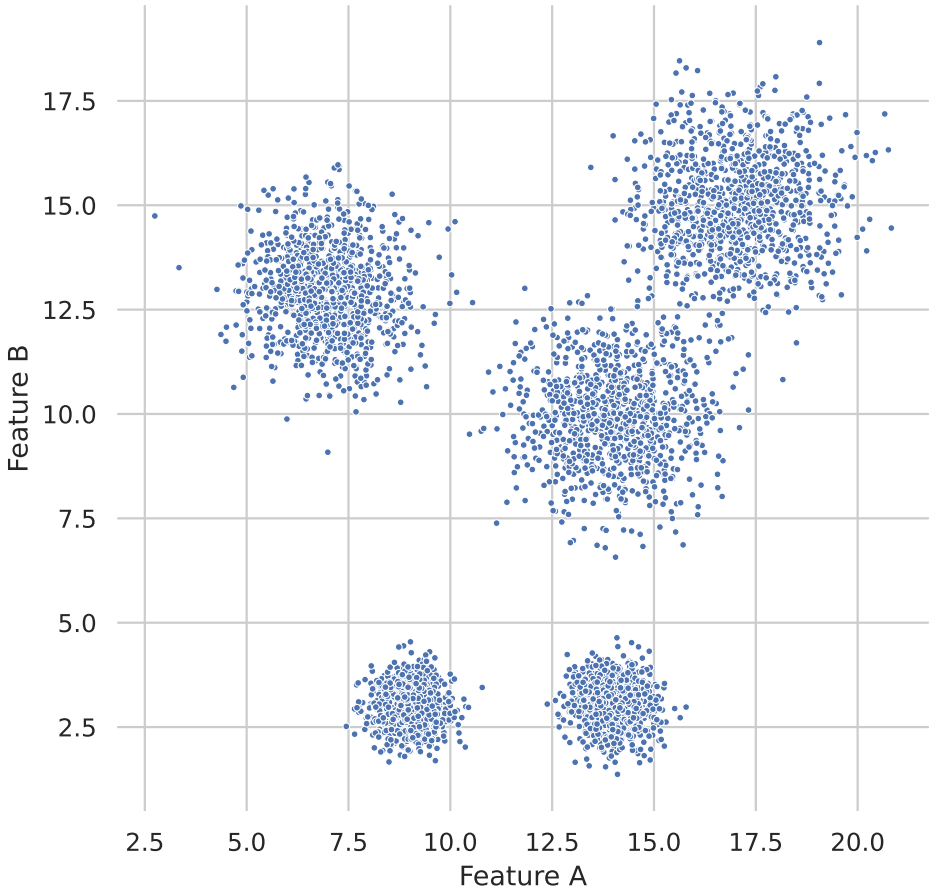


Figure S1: A toy world sampled from five Gaussians with unequal variances.

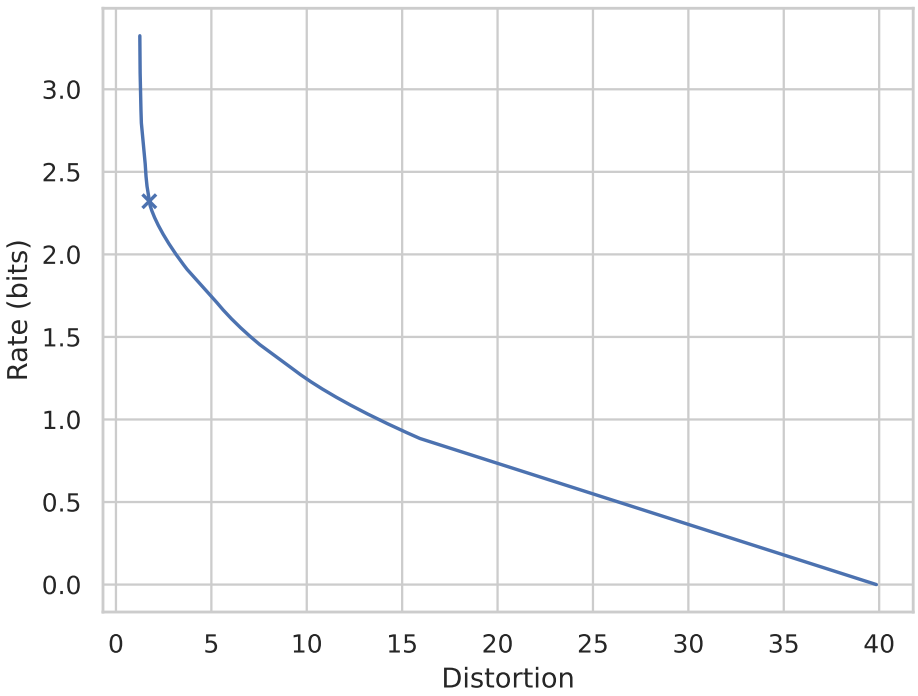


Figure S2: The rate-distortion curve for the unequal-variances dataset in fig. S1.

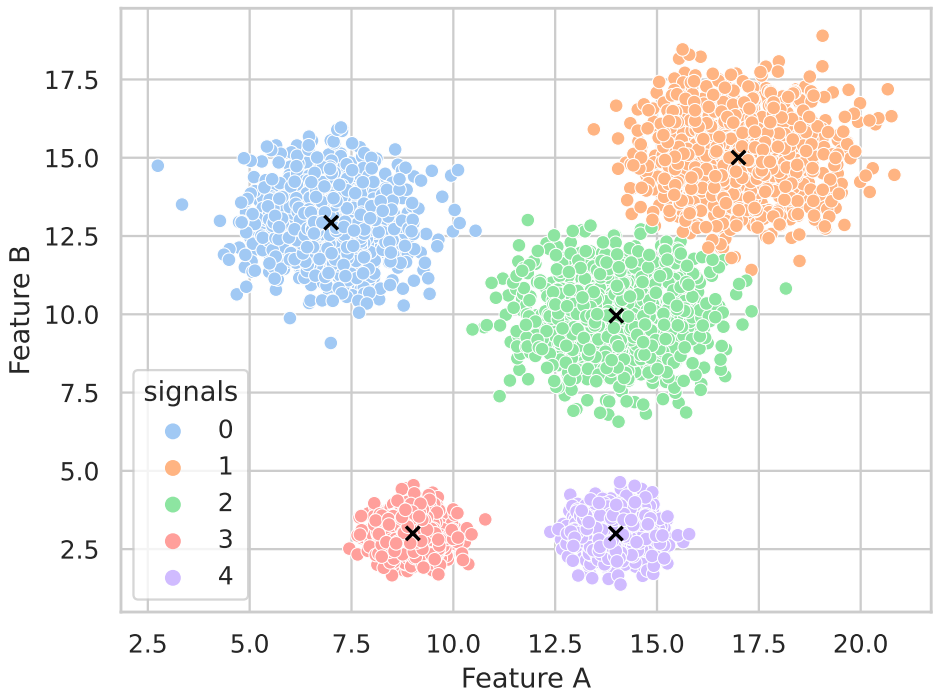


Figure S3: Optimal 5-signal codec at the cross of fig. S2.

prototypes on their means) always corresponds to a rate-distortion-optimal codec. But the second quantity roughly follows the number of *modes* in the mixture, which might be smaller than the number of Gaussians itself. Here I present example of such a mixture (fig. S4).

The 1-bit (that is, two-signal) codec in fig. S6 is rate-distortion-optimal and sits on an "elbow", as shown by fig. S5. Still, there are rate-distortion-optimal codecs with 3 and 4 signals—just not with more.

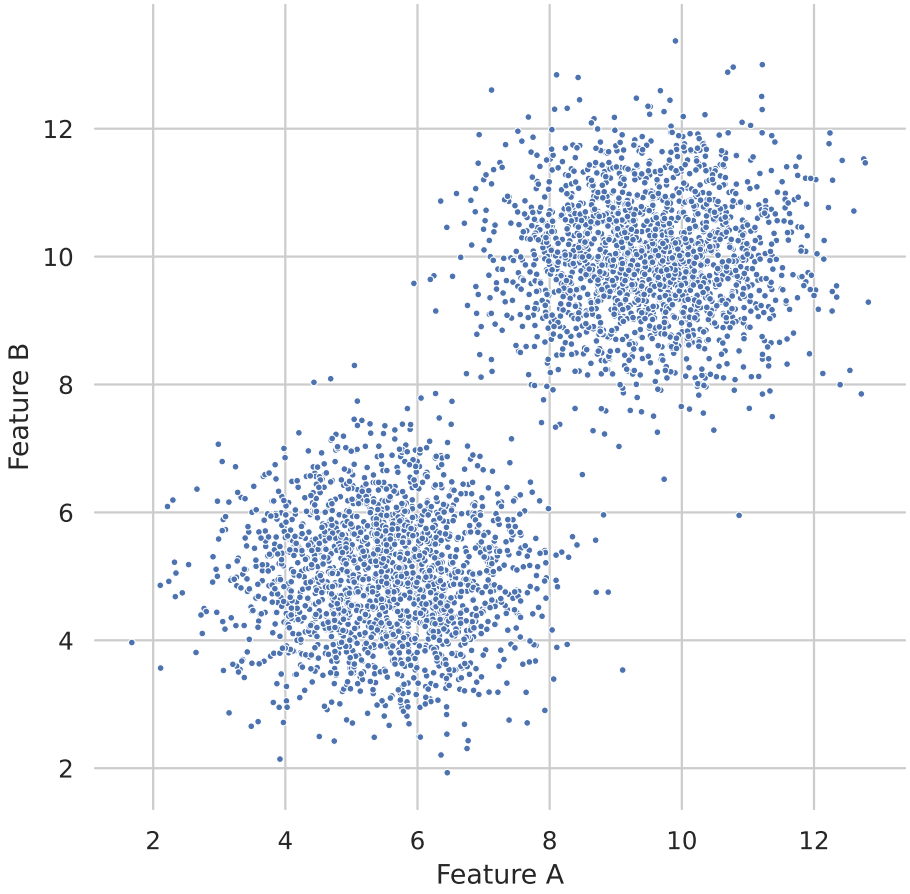


Figure S4: A toy world sampled from four Gaussians which result in two modes.

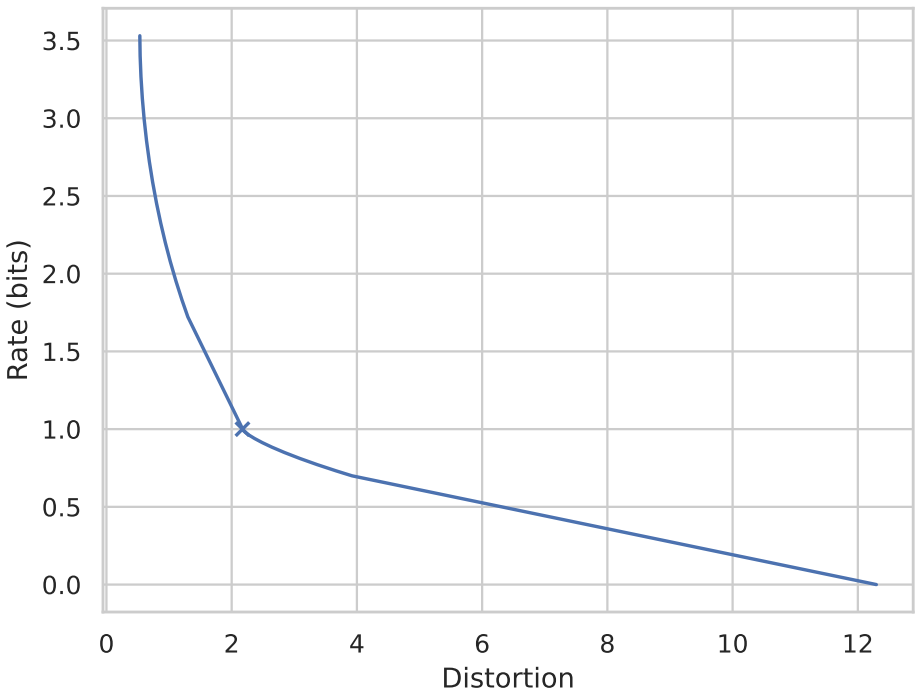


Figure S5: The rate-distortion curve for the unimodal-variances dataset.

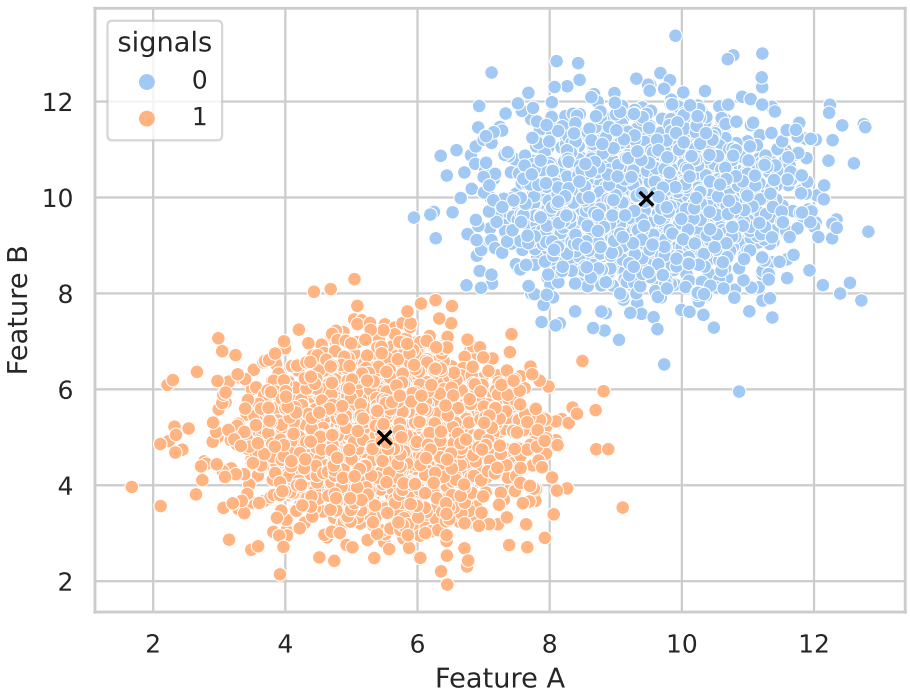


Figure S6: The 1-bit codec sitting at the elbow of fig. S5.

3 Uniform Source Distributions and Voronoi Tessellations

In this paper I aim at showing, among other things, that discrete repertoires of signals (what Fodor 1998 calls "conceptual atoms") and prototypes can be both part of an optimal information-processing strategy. I have focused on Gaussian mixtures because, for them, there is a maximum optimal number of discrete signals. This, I suggest, partly explains why conceptual repertoires typically provide a crude gist of a situation, while entirely disregarding finer detail.

For other kinds of source (e.g., uniform distributions), discrete repertoires of arbitrarily large numbers of signals can still be optimal. In such cases, optimal repertoires are reminiscent of the evolutionarily stable repertoires of color terms studied in Jäger and Van Rooij (2007): if the target distortion measure is an Euclidean distance, the resulting optimal partition, for any arbitrary numbers of signals, is a Voronoi tessellation, with prototypes in the centroid of each Voronoi cell. I show this in fig. S7, for an uniform source, mean squared error as the target distortion error, and various numbers of signals.

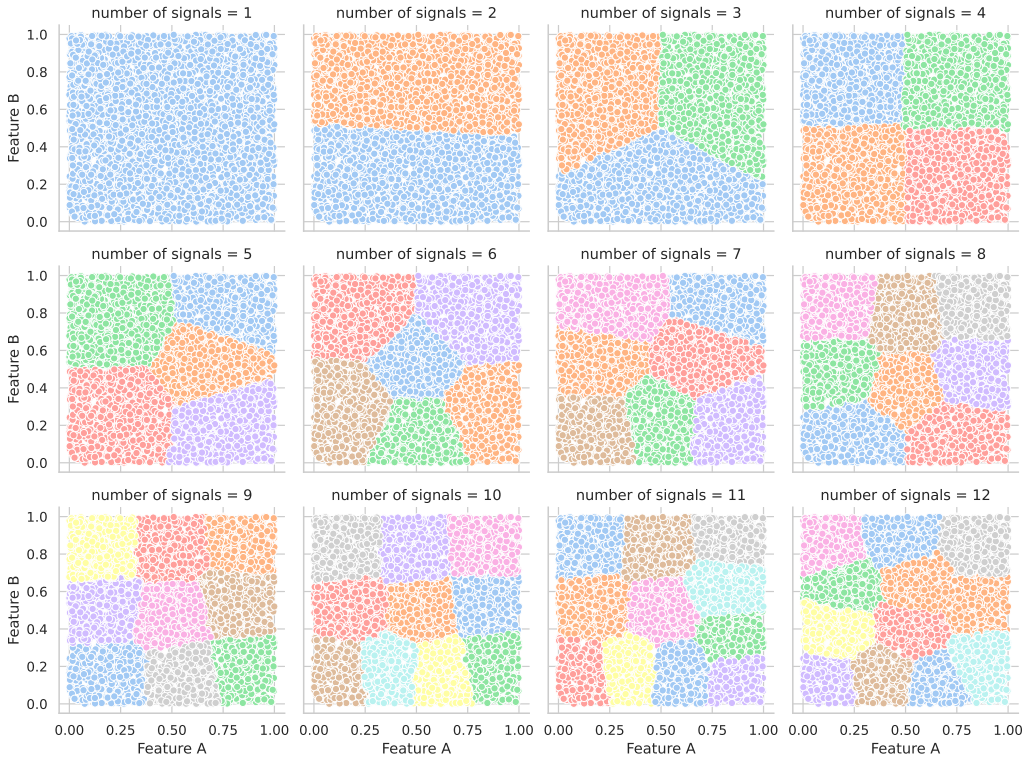


Figure S7: Optimal Voronoi tessellations for 1 to 12 signals, for a uniform toy world.

4 Non-Euclidean Distances

All analyses in this paper have used an Euclidean distance (or its discrete counterpart, the Hamming distortion) as distortion measure. This is why the rate-distortion-optimal partitions have always been Voronoi tessellations.

In fact, rate-distortion analyses can rely on arbitrary distortion measures. For example, here, I run a similar analysis to §2.1 in the main paper, but now using a distortion measure according to which $\hat{m}$ is a good decoding of m to the extent that both are equally close to some antecedently designated point, p. That is to say, if we want to encode and subsequently decode a certain point, m, in our abstract feature space, the decoded counterpart, $\hat{m}$ will have no distortion iff m and $\hat{m}$ are at the same distance to p. The more dissimilar these two distances are, the more distortion.

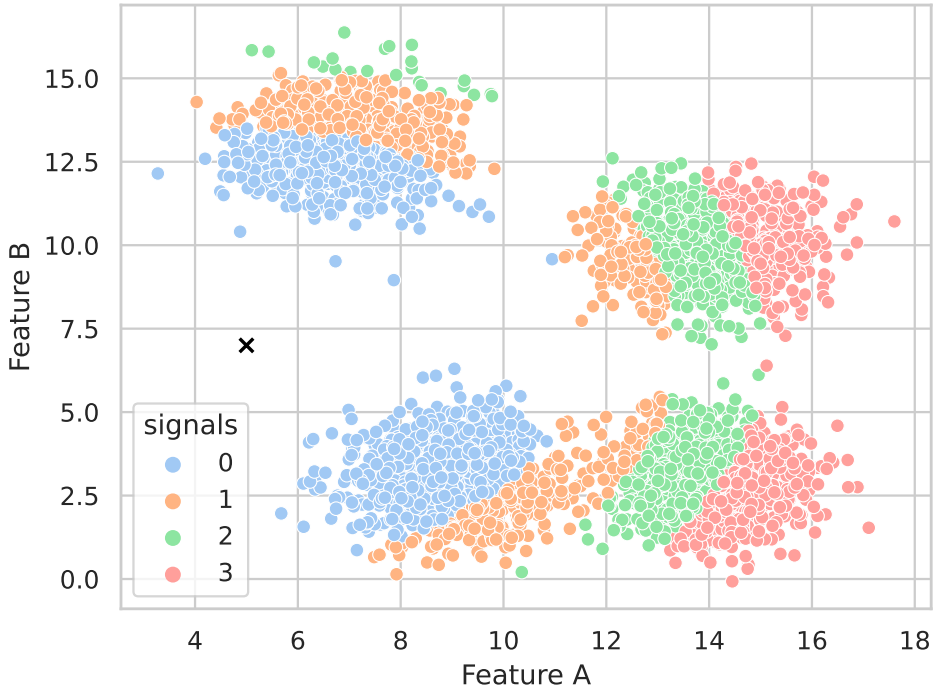


Figure S8: The rate-distortion-optimal 2-bit codec for the dataset in Fig. 3a of the main paper, using a "distance to a designated point" distortion measure. The black cross marks the position of the designated point.

For this distortion measure, the optimal categories are concentrical bands around the designated point (marked with a black cross in Fig. S8.) Two things to note about this.

First, these categories no longer rely on the "natural" structure imposed by the mixture of Gaussians. This structure is only relevant for Euclidian-distance-based distortion measures. Second, the resulting categories are no longer convex (i.e., straight lines connecting two points in a category may pass through other categories.) Gärdenfors (2000), among many others, have argued that "natural properties" are convex. Fig. 8 shows that convexity depends on an Euclidean-distance-minimizing goal. The implicit assumption that all natural categorization systems are of this sort needs to be explicitly tested and validated.

References

Fodor, Jerry A. 1998. *Concepts: Where Cognitive Science Went Wrong.* Oxford University Press.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought.* MIT press.

Jäger, Gerhard, and Robert Van Rooij. 2007. "Language Structure: Psychological and Social Constraints." *Synthese* 159 (1): 99. https://doi.org/10.1007/s11229-006-9073-5.

Rose, Kenneth. 1994. "A Mapping Approach to Rate-Distortion Computation and Analysis." *IEEE Transactions on Information Theory* 40 (6): 1939–52.