No Coincidence, George: Capacity Limits are the Curse of Compositionality

Steven M. Frankland[1], Taylor W. Webb[2], Jonathan D. Cohen[1]
Princeton Neuroscience Institute, Princeton University, Princeton, NJ
University of California, Los Angeles, Los Angeles, CA

***preprint* - manuscript in preparation**

Archived: https://doi.org/10.31234/osf.io/cjuxb

To whom correspondence should be addressed: steven.frankland@princeton.edu

Abstract

There is striking diversity in the capacity of different cognitive processes. In some settings, humans preserve only a few bits of information over computation: for example, tasks involving working memory and attention, perceptual identification, and numerosity estimation are famously limited (Miller, 1956). Other cognitive processes seem essentially unbounded, both in what we could possibly represent (e.g., the meanings of novel sentences in natural language) as well as what we can remember, once represented (e.g., episodic memory). These strengths and apparent weaknesses are intimately related. We integrate ideas from information-theory and the cognitive sciences to argue that in order to generalize efficiently–a key cognitive strength—processing capacity will not just be finite, but profoundly limited–a famous cognitive weakness. A unified computational framework precisely predicts classic error rates and patterns of response times in working memory tasks, explains why only a few items can be enumerated with accuracy and speed ("subitizing"), and why only a few items in a set can be accurately ranked ("absolute identification"). This computational framework suggests that the human mind is optimized for a particular objective: *efficient generalization*, at the expense of processing capacity.

Keywords: capacity limits, efficient coding, information-theory, compositionality, Hopfield networks, free energy, MEME

In the title of his seminal article,"The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," George Miller (1956) refers to a striking limit in information processing, almost identical for three distinct cognitive capabilities. The most famous of these — the number of "chunks" that can be held in short-term memory — has catalyzed decades of subsequent work in the cognitive and brain sciences (Anderson, 1983; Alvarez & Cavanaugh, 2004; Baddelely, 1992; Cowan, 2001; Mathy & Feldman, 2012; Miller & Cohen, 2001). However, in less-heralded sections of the article, Miller considers the then nascent findings that other capabilities exhibit a similar capacity-limit: the precision of absolute identification (i.e., the number of perceptual stimuli that can be ordinally ranked, such as a set of tones by frequency), and the subitizing range in numerosity tasks (i.e., the number of items in a set that be enumerated with both speed and precision, without explicit counting). Although Miller wondered whether these limits owe to a common source, he ultimately concluded that the likeness was coincidental. Here, we challenge that conclusion.

Information-theory offers normative principles for optimal representation in resource-limited systems. The *capacity* of any cognitive process–whether it be perception, categorization, memory, attention–is the maximum attainable *mutual information* between the input and output of the process (Cover & Thomas, 1991). The mutual information conveys how much the uncertainty about one variable is reduced, given knowledge of another.

Rather than considering each cognitive process separately, here, we assume that stark capacity limits derive from a single representational objective: agents maximize a general mutual information between the world (*S*) (i.e., the causes of sensory data) and their internal representation ($I(S;Z)$) (Attneave, 1954; Barlow, 1961; Linsker, 1988; Friston, 2010). In this setting, profound capacity limits derive from two factors of the system that are in tension: *efficiency and flexibility.*

The first has been well characterized: it's uncontroversial that—whatever the locus of these resource limits– it's impossible to preserve all the information about the world that agents receive (Sims, 2016; Van Den Berg & Ma, 2018). Agents have finite computational resources, requiring that those resources be deployed efficiently. Efficient coding suggests allocating representational resources to states ($s_i \ni S$) in inverse proportion to their probability p(s_i). Famously, to avoid information loss, the dimensionality of $Z(S{=}s_i)$ must be at least -log $p(s_i)$. Alternatively, the system could trade off information for efficiency—this is the logic of lossy data compression (Shannon, 1948; Tishby, Pereira, & Bialek, 2000; Sims, 2016). But, in this setting, it remains unclear why agents should be willing to tolerate *so much* information loss—surely, being able to remember more than a phone-number's length of digits, for example, would be beneficial.

It is useful to consider a second feature: agents also need to represent improbable–often novel– states (requiring cognitive *flexibility)*. The flexibility and efficiency of a representation (*Z)* are related through its Shannon *entropy* ($H(Z)$). If *Z* is the set of all possible representational states [$z_1 \ldots z_n$], *H(Z)* is:

$$1. \quad H(Z) = -\sum_{z\in Z} p(z)\log p(z).$$

The more possible states–as well as the more dispersed the probability distribribution over those states is– the greater its entropy. Intuitively, the greater a system's entropy, the less statistical structure (i.e., correlations between variables) there is in the representation, and the more flexible the system is. But, the greater the entropy, the more computation that's required to resolve the uncertainty–*the less efficient it is.*

This is a version of the classic *curse of dimensionality* (Bellman, 1956) common in statistics and machine learning.

Although this tension appears inescapable, work on representation from different corners of cognitive science offers a way out: cognitive systems can be both fast and flexible by representing states *compositionally*. For example, to represent *what was where* in a visual display (Figure 1a), the agent may exploit both recurring and context-specific structure in the world. Representation depends on stable variables that describe, for example, object features and spatial locations. Then, in a particular context, the agent will re-combine those pieces to represent *a yellow square in the upper left, above a red square*.

A representational system is compositional, in this sense, if it has a vocabulary of stable representations and a way to re-combine them, conditioned on features a particular context, where context involves *observed data* (S_i) the *agent's goals* (S_k), even *situational knowledge* $(S_j)^1$. A variant of this scheme was made famous in the analysis of meaning in natural language (Frege, 1892; Partee, 1995; Szabo, 2000) and has been influential both in symbolic theories of cognition and explanations of human generalization (Fodor & Pylyshyn, 1988; Smolensky, 1990; Hummel & Holyoak, 2003; Kriete et al., 2013; Lake et al. 2017; Frankland & Greene, 2020). Here, we specify key features of a compositional representation in statistical terms, making it amenable to information-theoretic and neural network implementation.

A compositional scheme is one in which, over long time scales (ontogenetic or phylogenetic), agents acquire codes for behaviorally-relevant variables $p(Z|S;\ \phi)$ (e.g., *space, time, color, shape*). Over short time scales $(S=s)$, agents update a different set of parameters $(\Theta \mid S=s)$, conditioned on recent sensory data and context (e.g., *what is relevant?, what is where?).* These two classes of parameters differ in their rate of change—slowly (what appears "stable", ϕ) and quickly (what appears "dynamic", (Θ_z) (Hinton & Plaut, 1987; McClelland, O'Reilly, & McNaughton, 1995). Jointly, these two classes of parameters enable both efficiency and flexibility: resources can be devoted to stable features of the world in proportion to their probability $(Z|S;\ \phi)$, and the system can update Θ to represent the structure in the particular case. This enables *efficient generalization,* but, we argue, comes at the cost of severe limits in processing capacity. To see *why* this reduces information-processing capacity, consider the mutual information that agents seek to maximize.

[1] This framing departs from the classic construal in linguistics and philosophy language, where 'compositionality' often refers to the idea that the meaning of a complex expression is a function of the meaning of its parts, the way they are combined (syntax), and nothing else. Then, any additional "contextual" information that influences meaning is inconsistent with strong compositionality. Here, we allow contextual signals (e.g., goal-representations, schemas), which may be internally generated to play the same functional role as bottom-up signals. We see this as *compositional* in the relevant sense:they are shared across contexts (ϕ), and influence computation through dynamic parameterization (Θ). In considering both the principles and mechanics of the system abstractly, it is not always relevant whether the "meaning" of different representations reflect explicit features (e.g., words) or inferred neural representations ("goals"/"context").

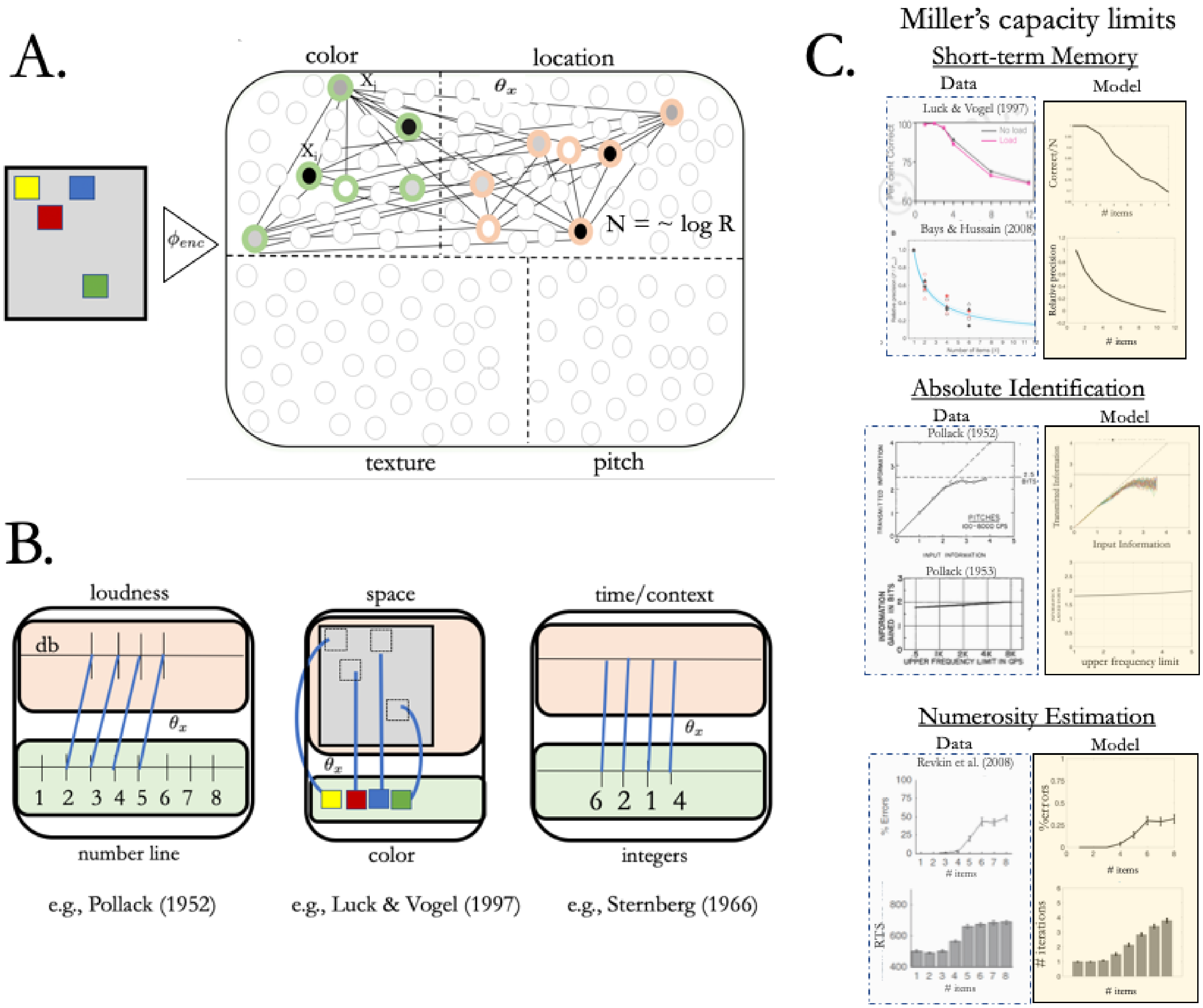


Figure 1. Overview of Computational Framework (A.) To reason quickly about unfamiliar states, (e.g., *what was where),* agents adopt a compositional coding scheme. We translate that scheme in a formalism amenable to information-theoretic and neural network analysis by defining two classes of parameters that differ in their rate of change–*slowly* (ɸ**),** conditioned on data over long time-scales, and *quickly* (*Θ*), conditioned on features of a particular context **(B.)** This scheme is general, and we argue, required to complete classic capacity-limited task paradigms in cognitive science. **(C.)** We foreshadow some of our model's results, focusing on phenomena that George Miller (1956) believed to be "coincidentally" similar. Our model precisely predicts familiar limits in (top) *short term memory*, explaining phenomena motivating both slot models, with capacity ~2 bits–and resource models, (middle) *absolute identification*, the number of items (e.g., tone frequency) that can be accurately ranked across different sampling ranges, and (bottom) *numerosity estimation*, with fast and nearly perfect estimation until ~3. This suggests these limits are no coincidence, but the consequence of a system tuned to *efficient generalization.*

Formalism. Mutual information is a function of two competing entropies

$$2. \quad I(S;Z) = H(Z) - H(Z|S).$$

The first entropy (*H(Z)*) describes the probability distribution over states of *Z*: this entropy places an upper bound on *I(S;Z)*--the larger *H*(*Z*), the greater the capacity. The conditional entropy (*H(Z*|*S)*) describes the uncertainty about *Z*, given *S:* the more predictive *S* is of *Z*, and vice versa—the lower the conditional entropy, the greater the capacity.

We assume that, to capture the variety of behavioral-relevant states*,* the agent represents *S* using a vector of variables *Z* (Z_i...Z_N)*.* Some of these variables may be devoted to, for example, representing *color (e.g.,* Z_i...$Z_{j)}$) others to *shape* (*e.g.,* Z_{j+1}...Z_k)*,*, others to *space,* yet others to *time* etc. Collectively, the joint entropy of those *N* variables is a vector-valued generalization of Eq 1:

$$3. \quad H(Z_1 \dots Z_N) = -\sum_{z_1 \in Z_1} \dots \sum_{z_n \in Z_n} p(Z_1) \dots p(Z_N) [\log p(Z_1) \dots p(Z_N)].$$

For present purposes, it is important that any statistical structure between the variables of Z decreases *H*(*Z*) and therefore also mutual information, according to Eq 2. By contrast, the conditional entropy (*H*(*Z*|*S*)) is defined as:

$$4. \quad H(Z|S) = -\sum_{s \in S, z \in Z} p(S,Z) \log \frac{p(S,Z)}{p(S)}.$$

To maximize mutual information, the system seeks variables *Z* parameterized by ϕ that co-vary with *S,* increasing the joint probability (*p(S,Z)*). But are also independent of one another (E[$Z_i^T Z_j$]= 0). Whereas statistical structure in the representation (Z) decreases mutual information, statistical structure between representation (Z) and world (S) increases mutual information.

These are inherently in tension. Over long time-scales, it's possible for the system to find configurations of *Z* and ϕ that optimize this tradeoff. However, for the agent, there is little time to update *P(S,Z)* for any particular—that is, improbable— case (*S*=*s)*.

Formally, we can follow Cheyette & Piantadosi (2020) and quantify the amount of information processing that is required to update the agent's prior knowledge *p*(Z;ϕ) to her posterior representation *p*(Z|S=s; ϕ) using the Kullback-Leibler-divergence (D_{KL})---a measure of the dissimilarity of probability distributions– between the two.

$$5. \quad D_{kl}(P(Z|S=s) \;||\; |P(Z)) = \sum_{z \in Z} p(Z|S=s) \log \frac{p(Z|S=s)}{p(Z)}.$$

It is standard in information-theory to treat D_{KL} as a measure of *how much computation* is required to transform a representation optimized for $p(Z)$ to a representation optimized for $p(Z|S$=s). Environmental demands naturally place limits on how "far" agents' prior distribution (p(Z)) can be from its posterior p(Z|S=s). For an agent, the time-bound on information processing is often on the order of milliseconds to seconds to represent sensory data, such that it can guide action.

For a single set of parameters, this is in conflict with the need to represent many possible states. The more possibilities in Z, the greater the entropy, $H(Z)$, the greater that distance $D_{KL}(p(Z|S=s) \| p(Z))$. According to Equation 1, tuning a system to maximize mutual information over long time-periods increases $H(Z)$. When the prior probability of any particular state is *low*—which occurs when the dimensionality is high, that distance ($\mathrm{D}_{\mathrm{KL}}(p(Z|S=s) \| p(Z))$) will be large.

In this context, a system that had only slow (ɸ) parameters would be *efficient*, but *inflexible.* It could only represent those states that occurred with enough frequency to assign (estimate?) $P(S=s,Z=z)$. By contrast, a system that had *only (Θ)* parameters would be *flexible*, but if those were not defined over abstract, recurring features (*color, space, time*, etc.) and their expected relevance, it would be *inefficient.* To meet the joint demands of flexibility and efficiency, agents adopt multiple *classes* of parameters for Z, defined over different time-windows. The *rate at which Θ changes* must be less than the time bound so that any subsequent information processing that is required is also less that bound $((D_{KL}(\mathrm{p}(Z|S=s \| \mathrm{p}(Z)) < b)$.

A variety of cognitive processes–for example, *attention and variable-binding*– can be seen as ways to update fast parameters to deal with this problem. Attention, for example, reduces entropy by associating top-down ("goals") and bottom-up ("salience") signals with a subset of representations (Z_{i}), restricting the space of behaviorally-relevant possibilities (e.g., is the task about *color* or *shape*?), and thereby decreasing entropy.

Then, to represent particular states (e.g., *what was where?)* agents must re-combine these task-relevant variables (Von der Malsburg; 1999; Smolensky, 1990; Roskies, 1999; Marcus, 2003). Although how neural circuits rapidly perform variable-binding (Von der Malsburg, 1994; Doumas et al. 2008; Hayworth, 2012; Kriete, 2013; Maas et al. 2019) remains an outstanding question, here, we focus on the *statistical* properties. Whatever the mechanism, binding results in a transient statistical dependency between Z_i and Z_{j} (e.g., *what was where*?).

This is the situation the agent in classic capacity-limited task paradigms finds herself in: a clear set of task-relevant variables specified by the experimental context, with no long-term relationship between semantic domains (See Figure 1B). To represent *what was where,* agents update *Θ* to speed information processing, but limiting processing capacity. The slow parameters ɸ remain undisturbed, allowing efficient representation of future states. But the ability to represent novel states quickly using *Θ* reduces processing capacity–that is, within-context entropy–profoundly. This is the *curse of compositionality.* It motivates a particular process model that can explain classic error rates and response times in cognitive science.

Computational Model. Consider a neural network of N nodes that jointly represent a collection of behaviorally-relevant variables ($Z_i \ldots Z_N$; ɸ, Θ_z). A node in this framework is a cognitive variable, and as such, can be considered a statistical summary of the information-content of many individually noisy neurons. To model cognitive phenomena, we can assume that these variables reflect features like the *spatial* or *temporal* structure of a stimulus or set of stimuli, the *color* of an object, the *intensity* of a tone etc[2]. Efficient coding dictates that representation (*Z*) minimizes dimensionality ("code-length"), while preserving behaviorally-relevant information about the variable.

Shannon's classic work showed that *N* is bounded by the entropy of the represented variable *x* *(*-log[p(*x*)]). Although the number of distinctions that an agent might need to make for one variable (for example, *auditory pitch*) might be quite different than the number of relevant distinctions for another (for example, *space*), given that the efficient coding scheme is a logarithmic function of the number of distinctions, *N* is more similar across variables than might be expected (See Figure 2). Intuitively, for a binary coding scheme, with *N* variables we have a maximum of ~2^N distinct states in *x*.

If the agent knew the possible patterns, it might be possible to define a decoder capable of "looking-up" the most likely *S*=*s*, given some partial information (*S*'). However, the agent's situation is different: she may need to decode novel combinations of variables: for example, *a green square in the upper left*, given only partial information (*where was the green square*?).

We consider computation in terms inspired by models from statistical mechanics as systems that minimize *energies* (Hopfield, 1982; Hinton & Sejnjowski, 1983). The "energy" of a neural network is a global measure of the agreement between the knowledge stored in its parameters and current activity values. This mathematical framework has been influential in computational neuroscience (Hopfield, 1982; Friston et al. 2010; Gottlaub & Braun, 2021) as well as machine learning (Zemel et al. 1995; Hinton & Salaktudinov, 2007; Salakhutdinov, 2018; LeCun et al. 2006).

Given a population of binary variables [Z_i...Z_N]---here, representations— that have discrete-time (*t)* varying values (*a*) , computed as

$$6. \quad a_i = \sum_{i \neq j} \theta_{ij} x_j + \theta.$$

update these individual variables according to

$$7. \;\; a(t{+}1) {=} \begin{cases} 1, & \text{if}\, a_i(t) > 0 \\ -1, & otherwise. \end{cases}$$

That is, "asynchronously", until the energy function *(E), (*the negative of its probability)

$$8. \quad E[S] = -\frac{1}{2} \sum_{i \neq j} \theta_{ij} s_i sj + \beta.$$

reaches a local minimum. In the absence of additional noise, dynamics will drive the network toward local energy minima. If this process is repeated over many different populations, it can be seen as maximizing *p(S,Z)* and minimizing the conditional entropy *H(Z=z |S=s)*, therein increasing *I(S;Z).*

Given no previous knowledge of *S*=*s*', the system can update (*Θ*|S=s, ф) based on the observed data and goals. We assume that this update of *Θ* is performed according to Hebb's rule—simply a

[2] At the neural level, each dimension can be represented by a large number of noisy-neurons that, in aggregate, are essentially noiseless, by the law of large numbers. Here, the model is a *cognitive* model.

description of what varied (what's *relevant*) and what co-varied (what is bound to what) in the observed data.

If Θ reflects structure observed in data over a short time-period, under certain conditions, the system can use the available information about *Z (Z')* in concert with the knowledge in Θ*,* to recover missing information. This has allowed Hopfield's algorithm to serve as a potential memory for biological systems.

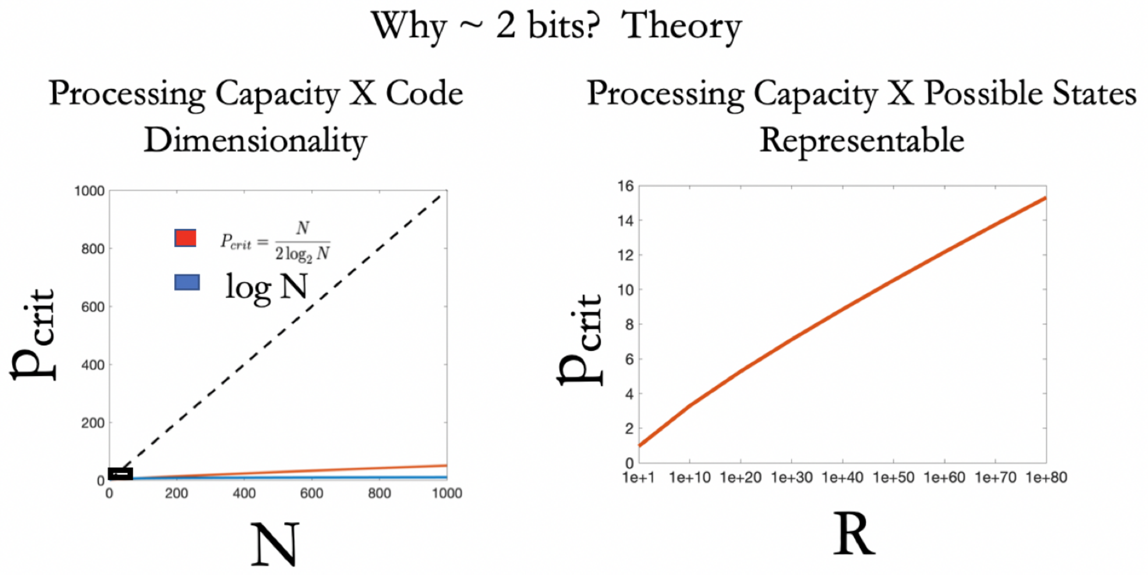
$$9. \quad \theta_{i,j} =< x_i x_j > - < x_i >< x_j > .$$

Classic work has found Hopfield networks are profoundly capacity-limited (Hopfield, 1982; McEliece et al. 1987; Sompolinsky, Amit, Guttfreund, 1985; Amit, 1989). The function that relates memory retrieval to set size has a critical point (p_{crit}), at which a memory is dramatically less likely to be retrieved. When the set size (p) $<p_{crit}$, the equilibrium states of the network are likely to correspond to previously experienced patterns. When $p/n > p_{crit}$, they may reflect statistical summaries of those patterns.

This limit is often seen as a disadvantage for Hopfield networks and has motivated considerable bodies of work in theoretical physics and machine learning aimed to increase their capacity (Krotov & Hopfield, 2016). However, from another perspective (Mackay, 1991), classic Hopfield networks aren't solving a memory retrieval problem, but a representational problem: what is the least-biased (most flexible) representation of the agent's knowledge, given the statistics of the data. This is Jaynes' (1957) principle of *maximum entropy*: choose the probability distribution that maximizes the Shannon entropy, consistent with the statistical information given. In the particular case, agent's assume the least–minimal update to prior knowledge– and perform inference through energy minimizing processes. An objective function that maximizes entropy and minimizes energy is equivalent to minimizing *free energy* (Amit, 1989; Friston, 2010; Gottlaub & Braun, 2020). Here, we call the computational model we use to explain psychological data "MEME" (Maximum Entropy/Minimum Energy)[3].

These two properties—efficient slow-learned representation and maximum flexibility in dynamic re-combination— limit processing capacity, not just in principle, but profoundly (See Figure 3). As we will see in the next section, given minimal assumptions, this framework precisely predicts the error rates and qualitative response time curves for visual short term memory, absolute identification, and numerosity estimation. The mind is not optimized for working memory tasks, nor for perceptual discrimination, nor numerosity estimation, but for *efficient generalization*—a fundamentally different objective.

[3] While the current colloquial use of this term is not lost on us, its use here is intended to directly reflect its original use and meaning (Dawkins, 1976), which was a reference to configuration of knowledge that, by analogy to genes in the evolutionary process, satisfy stability conditions defined by an interaction between internal and external sources of selective pressure. Here, we use it as an acronym, to describe the nature of those selective forces — Maximization of Entropy and Minimization of Energy — that lead to the formation of stable knowledge representations within a system.


Why ~ 2 bits? Theory

Processing Capacity X Code Dimensionality

Processing Capacity X Possible States Representable

Figure 2. Two-part explanation for low processing capacity. LEFT. Processing capacity (P_{crit}) of a maximally flexible (fully-connected) system, as a function of representation dimensionality (*N*) (RED). Human agents occupy a small, low-capacity part of the space (BLACK BOX). RIGHT. However, assuming efficient coding, only a relatively small *N* is needed to represent a vast number of possible states (R.) (N = ~log *R*). As a result, to distinguish a space of individuals equal in size to the number of atoms in the universe (~10^{80}), we would only need 256 binary variables, resulting in an expected processing capacity of ~16.

George Miller's Cases. To model cognition, we consider how a system would represent task-relevant dimensions efficiently. All tasks can be considered to have a "cue"--what is perceptually available to the participant as a stimulus– and a "target"---what must be inferred to complete the task. This information is presented in the form of *N*-dimensional patterns (*Z*) presented to the Hopfield network. We call the d_x dimensional representation for the cue '*X'* and the d_y dimensional code for the target '*Y'*. *X* and *Y* are slow-learned variables (e.g., X(s,ϕ_x)). Given efficiency constraints, we can make informed guesses about *N (*and d_x and d_y*)* in a particular task-domain. For a continuous input domain——-locations in space, points in time, values along a unidimensional continuum— an efficient re-coding of a 1D and 2D environments bears a striking resemblance to the grid-like codes found in medial entorhinal cortex (Hafting et al. 2005; Dordek, 2016; Stachenfeld et al. 2017; Wei et al. 2015). This is an increasingly well-studied case of a more general set of desiderata: efficient low-dimensional representations of the structure of a domain (Fiete, Burak, & Brookings, 2008; Wei et al., 2015; Stachenfeld, 2017). Grid-codes have exponential representational capacity (Fiete, Burak, & Brookings, 2008) and the particular scaling of frequencies in mammalian entorhinal cortex—$\sqrt{e}$— minimize the number of variables necessary to cover the space (Wei et al., 2015). Mathematical analyses have shown that grid-like codes can be derived from simple techniques like PCA (Dordek et al. 2016; Stachenfeld et al. 2017). Empirically, grid-like codes have been found to represent non-spatial dimensions, such as sound (Aronov et al. 2017), olfactory stimuli (Horner et al. 2018), 2D visual arrays (Bicanski & Burgess, 2019), and abstract conceptual dimensions (Constantinescu et al. 2016).

Then, on a shorter time-scale, the experimental context determines what is relevant, and what covaries (*Θ*). We update these statistics according to Hebb's rule. Given the cue (*X),* the model uses the statistical knowledge in its slow learned representation X(s,ϕ_X), Y(s,ϕ_y) and its rapidly-updated parameters (*Θ*) to infer the most likely "missing information" using biologically-plausible learning and updating rules.

Short Term Memory. We model a related body of work on short-term memory for visual displays (Visual Short Term Memory: VSTM) central to the study of short-term memory in recent decades (Pashler, 1988; Luck & Vogel, 1997; Wilken & Ma, 2004; Alvarez & Cavanagh, 2004; Bays & Husain, 2008; Sims, Knill, & Jacobs, 2012; Luck & Vogel, 2013; Bays, 2015). In its simplest form, the change detection paradigm involves a brief presentation of a visual stimulus. After a momentary delay (~1000 ms), subjects determine whether a change occurred with respect to the previous display. The number of items in the display is varied over trials, and in canonical cases, only one item has the potential to change. We assume that *Θ* reflects the statistical structure of the image on a particular experimental trial—the binding of location-codes (*X*) and color codes (*Y*) as in Eq. 4. That is, *what colors* were *located where* on the screen.

On this account, There is no dedicated working memory system: only a procedure for representation and inference. We query the network with the *location* code of one object in the original display, and allow the computational dynamics to evolve until equilibrium to identify the most likely features of an item at that location. The model thus retrieves the memories in *Θ*— a representation of the structure of the stimulus*,* given a location code (*X*). We compute the Hamming distance between the network's settled state and the set of possible targets. This distance can be converted into a simple decision rule for change detection. We use *d*<0.05, standard in work on Hopfield Networks (McEliece et al. 1987). This algorithm naturally transitions from near precise retrieval to high-probability of error, as has motivated classic slot models (Miller, 1956; Luck & Vogel, 1997).

Although the form of error rates will match the observed data, the exact quantities depend on d_x,d_y. Here, we use a grid-like representations of a 2D visuo-spatial array (Bicanski & Burgess, 2019) using the size and distance of their visual display and known exponential scaling properties of grid modules (Stensola et al. 2012; Wei et al., 2015). Colors were treated as abstract locations in a circular "environment", and projected into a repeating grid-like code of a single dimension. The size of this dimension was obtained based on knowledge of the range and discriminability of perceivable wavelengths (e.g., Long et al. 2006). The frequencies were scaled according to empirically observed scaling of grid-cells, known to minimize representation-length for a given number of distinctions (Stensola et al., 2012; Wei et al., 2015). Remarkably, this simple algorithm captures not just the form, but the quantitative limits of human change-detection error rates as a function of set size (See Figure 3) (Luck & Vogel, 1997). When the particular colors and locations are uniformly sampled, this coding scheme gives rise to a qualitative decrease in accuracy between 3 and 4, as if the visual system contained a finite number (~3 or 4) of slots.

Other empirical work has shown that representational precision is not fixed, but instead varies as a function of stimulus properties encoded in Θ (e.g., Wilken & Ma, 2004; Alavarez & Cavanagh, 2004; Bays & Husain, 2008; Sims et al., 2012). For example, Bays & Husain (2008) showed that as set size increases, the encoded stimulus-precision declines as a power-law (See Figure 2H). The more items in a display, the less precisely we encode the spatial locations or orientations of individual items. We measure the precision as the Hamming distance of the equilibrium state from the target (*inverse precision*). Figure 2 plots inverse precision as a function of set size. Inverse precision tracks the empirically observed decline (Bays & Husain (2008)): the more items present in the image, the further the equilibrium state is from the target, and the more difficult it becomes to differentiate similar queries.

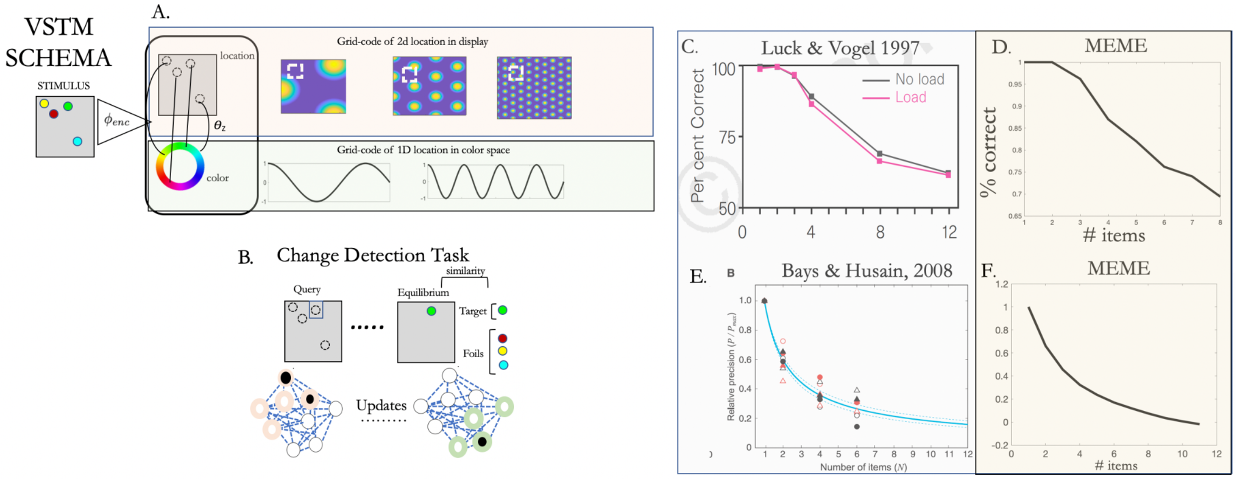


Figure 3. Classic Visual Short Term Memory (VSTM) phenomena. **(A)** We consider a visual stimulus of colored shapes at particular locations (Luck & Vogel, 1997), and assume features are factored into separate representational streams and re-combined. Locations are represented by grid-like codes of the 2D array and colors are grid-like codes of the 1D color space. The stimulus-specific weights (Θ) reflect the correlational structure of those variables observed in a particular image (*what color was where*?). **(B)** Change detection involves presenting the network with a location code and allowing it to evolve until equilibrium. Changes are reported on 50% of change trials in which the correct color is not within a Hamming Distance of 0.05 from the target. (**C & D).** The model's performance closely tracks Luck & Vogel's empirical observation of qualitative change in performance at ~3 items (**E,F**). In this framework, representational precision also decreases as an approximate power law, as observed in Bays & Husain (2008). The model thus captures phenomena central to both slot and resource models.

The iterative nature of the inference algorithm generates predictions about the functional form of human response times, as well as error rates. We find that a surprising number of previously observed human response times—logarithmic, linear, "elbow"— can be explained as different parameterization of the Hopfield model, where these parameterizations are determined by efficient representation of task-relevant domains (*X*,*Y*) and their statistical dependencies (*Θ,* "bindings") (See Figure 4).

We consider, as an illustrative case, a famous memory experiment owing to Sternberg (1966). There, participants were presented with a random sequence of integers (e.g., 8, 4, 2, 5). Seconds later, participants were queried with targets or foils to give a *yes/no* response regarding whether the query was in the stimulus set (e.g., 2). Sternberg (1966) famously found strikingly linear response times as a function of set size, with nearly equal response times for yes/no responses, suggesting exhaustive serial search of a memory system.

In the current framework, the settling time function depends on d_x/d_y and *p*. It's thus necessary to consider the representation of *X* (the integers), as well as the episodic context *Y* (context) associated with it. (task-relevant subset of number-line). An efficient representation of the task-relevant integers {0,9}requires only ~3.32 bits. The representation of the spatio-temporal (episodic) context is higher-dimensional. In the regime in which $d_x = 3.32$ and $d_x<d_y$, settling times are approximately linear as a function *p,* with a maximum r^2 on a linear model at d_y =~10*d_x. ~0.99. See Figure 4. Notably, ~10*3.32 is in the range of what would be expected for a grid-like representation of the recent past—i.e., time-code that is bound to each successive integer. Sternberg's recognition memory task can therefore be seen as a case in which an efficient low-dimensional integer code (*X*) is used to query the structured knowledge in *W* to retrieve a higher-dimensional context code (*Y*), rather than an exhaustive serial search of slots.

If the settling process is interrupted at random points, as in Reed (1973)'s stop-signal behavioral paradigm, the model displays the same functional form as human speed/accuracy tradeoff—exponential increases in accuracy before asymptote. This supports the contention that the algorithms for learning and inference are compelling models of human cognitive phenomena, when specified at the right level of description. Although how these forms map onto the absolute values of RTs will depend on the additive contribution of upstream perceptual processes and downstream motor output, we suggest the functional form of many response times can be characterized by *(a)* an efficient representations of *X (the cue)* and *Y (the target), (b)* their rapid association according to the Hebb rule, and *(c)* the use of a generalized form of Hopfield's algorithm to find local energy minima.

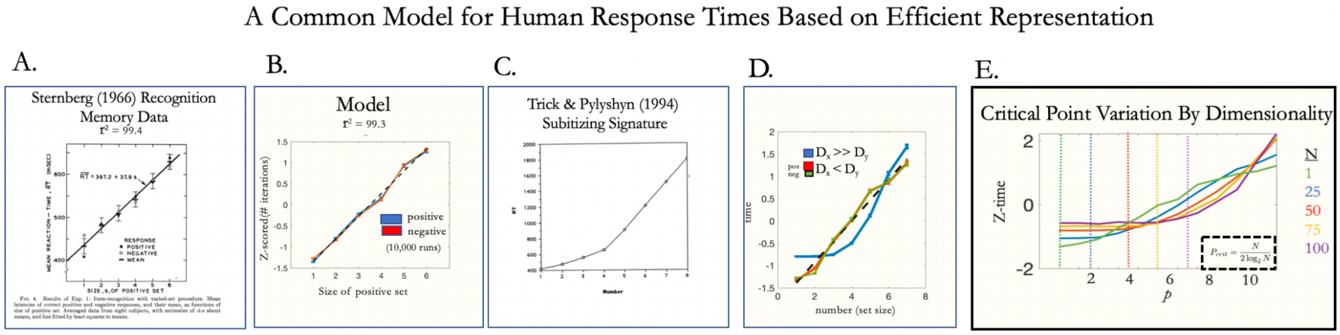


Figure 4. Our model explains a variety of qualitatively different response time functions in cognitive tasks, as a function of efficient task-relevant representation. **(A)** Display of Sternberg's (1966) classic recognition memory results: linear response times as a function of set size, indicative of serial exhaustive search through the system's memory. **(B)** However, a Hopfield network with asynchronous updating and efficient representations of task-relevant domains predicts the same result. **(C.)** In addition it captures the character of other paradigms—for example, for numerosity estimation as a function of set size—, where classic behavioral results show an "elbow" response time curve, often taken as evidence for separate systems. **(D)** Our model predicts both phenomena due to differences in the dimensionality (efficient representation) of the cue (d_x) and target (d_y) domains. The elbow arises when $d_x >> d_y$ (here, a spatial array (which requires many bits) vs. the integers 1-8 (requires only 3 bits). **(E)**.This suggests a general function for defining the critical points (vertical lines) in RTs as a function of representation dimensionality (N, 1-100, colors). We find the function to be $\sim \frac{N}{2\ log\ N}$, an equation previously described in theoretical work on memory capacity in Hopfield Networks. (McEliece et al. 1987).

Absolute Identification. We can extend the model to account for a collection of phenomena in absolute identification that were of particular interest to Miller. In an absolute identification task, subjects are presented with a set of stimuli along with ranks to be associated with those stimuli: for example, a set of nine evenly spaced tones between 1000 hz and 5000 hz, with a rank of one assigned to 1000 hz, a rank of two assigned to 1500 hz, and so on. This occurs at the beginning of the experiment ("demonstration period"). Then, when presented with each stimulus, subjects must report its rank: for example, in a set of pitches, a *2000 hz* tone might have been the tone with the *third* lowest frequency. There, psychologists identified severe limits in the number of unidimensional perceptual stimuli that humans can reliably identify (~2.5 bits) — strikingly similar to capacity limits in short-term memory. Moreover, these effects are famously invariant to the sampling range. Pollack (1953) found that AI limits are nearly identical regardless of whether tones are sampled across the entire range of perceivable pitches, (e.g., from 100 to 10100 hz in intervals of 1000 hz) or from a narrow range (e.g. only from 4000 to 5000 hz in intervals of 100 hz). Despite the tenfold increase in absolute difference in hZ subjects show little improvement in performance.

We assume that absolute identification involves slow-learned descriptions of the relevant stimulus dimension (pitch in the above example) and a mental number line for ranking them. The particular mapping in Θ between codes for the stimulus dimension and codes for the number line are learned for that particular experimental context.We derive expected grid codes for the Pollack's pitch dimension based on empirical JND of the modality-specific perceptual range (20hz to 20khz with a JND of 0.5%), as well as general grid-cell scaling properties (Stensola et al., 2012; Wei et al., 2015). Pollack (1952) employed equally spaced stimuli in the range of 100 hz to 10000 hz.

To perform the identification task, we map each sampled pitch from its 1D location to a multivariate space of frequencies and phases. The network is presented with the grid-code corresponding to a particular stimulus and the activity states evolve until equilibrium, in order to identify the most likely number (i.e., the number whose code has the shortest Hamming distance to the settled state). Figure 3e shows model results over 1000 trials for Pollack's pitch limits plotted in bits, capturing the observed capacity limit of <2.5 bits. Here, the within-context correlations determine Θ, and as a result, variables that are *stable within-context exert no effect on the computational dynamics*: no variance in Eq. 4. means no opportunity to influence the local updates. When the stimuli are concentrated at narrow ranges (e.g., 4000-5000 hz), low-frequency grid-cells thus exert no effect. However, we call this "approximate" invariance, given that, as we gradually increase the range, lower frequencies vary across stimuli with greater probability. This introduction of more variables leads to a small increase in capacity with increased sample range. But given the geometric progression of the frequencies in the code, the increase is small. See Figure 3D-F.

The finding of approximate scale-invariance depends on the experimental design decision to keep the sampling ratios equivalent across ranges. Gravetter and Lockhead (1973) found that when the same stimuli are flanked by nearby neighbors, precision is greater than when they are flanked by the same number of peripheral stimuli ("broad") (See Figure 3G-I) (See also Braida & Durlach, 1972, Rouder, 2001). Gravetter & Lockhead's (1973) "broad sampling" condition introduces within-context variance in low-frequency grid-cells. These low-frequency variables carry no information about neighboring items, but still influence the computational dynamics in Eqs. 5 & 6, given their variance.

Broad sampling thus decreases precision for the same reason that increasing set size decreases precision in the VSTM work outlined in the previous section (Wilken & Ma, 2004; Bays & Husain, 2008): both introduce redundancies in the code through correlations, decreasing achievable information rates for certain discriminations. For a cognitive system, these within-context redundancies are the statistics induced by compositionality in the code; mechanisms capable of dramatic increases in the number of possible states we could encode quickly. This positions us well to deal with unfamiliar futures, but at the cost of capacity limits in information processing when we actually encounter them.

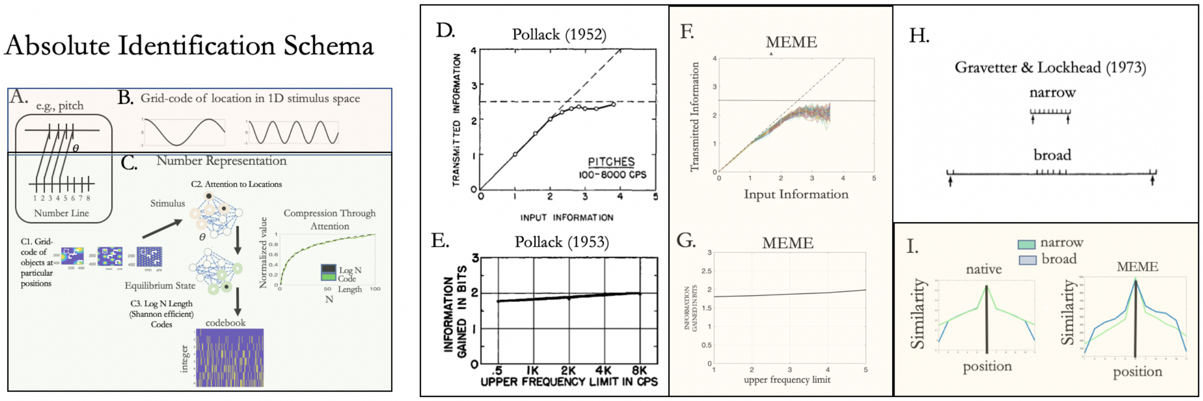


Figure 5. (A) Schema of Absolute Identification task. Subjects are presented with a set of stimuli along a target dimension (e.g. pitch) together with associated ranks. The task is to later report the corresponding rank when queried with a particular stimulus. To model these phenomena, we assume (**B**) 1D grid-like code for perceptual dimensions, and a multi-step program for deriving a representation of the mental number line from first principles and data **(C).** We derive a number line representation by assuming (C1) variable number of objects are presented at random locations in a 2D array. In each case, agents (C2) update θ to attend to those locations. We find (C3) that the equilibrium states are perfectly efficient (log *N*), matching well-known psychophysical phenomena indicating a “compressed” mental number line (Dehaene, 2003). We use this compressed number line for modeling. **(D)** Pollack found a limit of ~2.5 bits of information for absolute identification of pitch, highlighted in Miller (1956). This limit is approximately scale invariant **(E)**, as increasing the absolute difference between pitches has little effect on discriminability. Our model reproduces the limit **(F)**, and the approximate scale-invariance effect **(G),** to the point of predicting the slight linear increase (E vs.G). **(H)** However, this is not due to a fixed item capacity, but instead depends on the relationship between sampling distribution and precision. For example, “broad sampling” to include stimuli near the min and max of the perceivable range (Gravetter & Lockhead, 1973) reduces precision on the middle items, relative to narrow sampling. (**I)** Our model predicts this pattern, as broad sampling introduces low-frequency redundancies into the codes, causing increased errors.

Numerosity Estimation. Finally, Miller addresses a quantitatively similar limit in the number of items in a visual display that can be rapidly and accurately enumerated while the stimulus remains on the screen. There, subjects show fast and nearly perfect report of small numbers (the "subitizing" range), but increasingly slow and error-prone reports thereafter. Given standard experimental conditions, the limit with visual stimuli is believed to be between 3 and 4 (Mandler & Shebo, 1982; Trick & Pylyshyn, 1993; Revkin et al., 2008; Cheyette & Piantadosi, 2020, smaller than the original ~6 from Kaufman et al., 1948), and is often taken to evidence a separate "small number" system (Revkin et al., 2008). As with short term memory, we suggest a single system that exhibits qualitatively different behavior as a function of the observed statistics.

Here, the slow-learned representation of 2D visual array is as described in the VSTM section, and the slow-learned 1D number line is as described in the section on absolute identification. Subjects' task is to infer the latter, given the former. The fast-weights reflect the objects presented at particular locations, which will vary from 1 to N_{task}. In this context, the Hopfield network can be seen as a model of *what captures attention* with limits imposed by the rapid association of spatial variables in Θ.

We assume that Θ is updated on each trial, and further that it contains *perfect* knowledge of the covariance between the 2D grid code for a stimulus display of *N* items, and the 1D codes from 1 to *N,* "as if" the agent were sequentially counting them in a random order. Of course, subjects are not mentally counting in this task, but this forms an idealized model. The model is thus "told" the correct answer by associating the complete image with correct integer code. But it must maintain the set of associations from 1 to *N* in a common representational space (Θ*).* Any capacity limits the model exhibits therefore stem from the representation and computational dynamics.

Figure 6 shows that this model exhibits phase transitions in error rates at ~3 or 4, like human subjects. In addition to the error curves, we compute the mean number of passes through the variables until equilibrium. These functions track the RT curves commonly observed: if p<~3, RTs are almost flat, but if p>3, reaction times increase monotonically (Kaufman, 1948; Mandler & Shebo, 1983; Trick & Pylyshyn, 1993; Revkin et al., 2008; Cheyette & Piantadosi, 2020). Iterations until equilibrium in the Hopfield network follow the same structure.As with VSTM, we expect differences in numerosity estimation to be influenced by recent updates to Θ*,* contingent on the experimental context. These drive differences in the computational dynamics.

Revkin et al. (2008) observe fast RTs, low error rates and minimal response variance when the possible responses are 1-8, but not when the possible responses are 10-80 in intervals of 10 (See Figure 6). This failure of universal Weber-like precision is taken as evidence for a special system for small numbers. In our view, this can be explained by the nature of the representation and the computational dynamics. Here, we assume Θ reflects the range of integers under consideration (1-8 vs. 10-80). As above, the model is "told" the correct answer by associating the complete image with correct integer code, and limits are imposed by the nature of the representation inherited from the slow-codes for space and number, and the computational dynamics. Under this scheme, the Hopfield network tracks empirical results in both conditions (See Figure 6).

There is no need to invoke separate fixed-capacity systems: Representations optimized for efficient generalization are sufficient to explain these phenomena. Although the current framework agrees with the long-standing observation that small numbers are *cognitively* special (Revkin et al. 2008), this doesn’t require them to be executed by separate systems. Like Gallistel & Gelman, (1992) and Cheyenne & Piantadosi (2020), our model suggests that the specialness is in the output, not the mechanism. We further suggest that representations optimized for efficient generalization are sufficient to explain not only the qualitative forms of these results, but also the particular capacity-limits (~3-4). As with the other phenomena we model, we do so without fitting any task-specific empirical data, or making assumptions about the amount of internal “noise”. Instead, profound capacity-limits are predictable using general estimates of the JND of task-relevant variables, and critically, an abstract description of a computational system optimized for efficient generalization.

On our view, qualitative changes in error rates and response times are expected at critical points, the particular location of these points will depend on the nature of the representation, and the nature of the representation itself has been determined by (a) efficient coding of behaviorally-relevant distinctions and (b) the need to generalize quickly. Jointly, this offers a framework for understanding classic capacity-limited phenomena. Profound capacity-limits do not owe to any biophysical property, nor to exhaustion of a finite number of slots. Instead, the main objective is simply to configure an agent to stand ready to represent novel states when they appear, given limited time. On this view, capacity limits— one of humans most striking deficits—can be seen as a consequence of one of our greatest strengths: *efficient generalization*. This is the curse of compositionality.

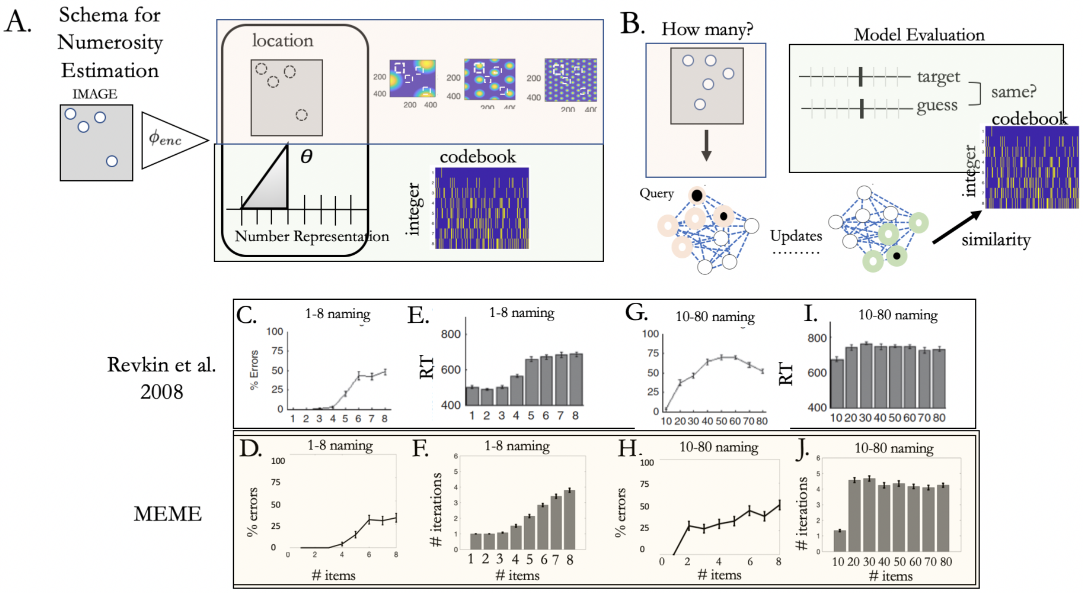


Figure 4. Numerosity estimation. **(A)** Stimuli are visual displays consisting of sets of objects. Set size (*N*) varies across trials. Subjects' task is to report the cardinality of the set, and error rates and response times are collected. To model this task, we assume that agents factor the stimulus into a spatial representation (grid code of 2D space) and a number line representation that must then be related ("how many?"). Here, we allow the network to "know" the true cardinality on each trial by associating the code for the complete spatial layout with a representation of *N* in θ. However, we assume this requires that cardinalities<*N* also be in θ. The capacity-limits then stem from (1) the correlations in the spatial structure as *N* increases and (2) the decreasing precision in the number code with increasing cardinality (itself determined by correlational structure on a slower time scale, as in the absolute identification task). **(B).** To evaluate the model, the network is queried with the spatial representation of the image, and allowed to evolve until equilibrium to identify the most similar number-code. **(C).** Example human behavioral data from Revkin et al. (2008) showing standard subitizing effect in error rates. **(D)** Like humans, the model shows nearly perfect performance when *N*<=3, but declining performance thereafter. (**E&F)** Likewise, the canonical human RT signature is qualitatively predicted by the number of iterations (settling time) through the Hopfield network. **(G,H,I,J).** Both human and model performance depends on the experimental context shaping the statistics in The weights (θ) encode the network's knowledge of this relationship. (1-8 naming left vs. 10-80 naming right).

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, *9*(1), 147-169.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, *15*(2), 106-111.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, *55*(14), 1530.

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge university press.

Anderson, J. R. (1983). *The architecture of cognition*. Psychology Press.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, *61*(3), 183.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01).

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851-854.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, *7*(6), 1129-1159.

Bellman, R. (1957). *Dynamic Programming.* Princeton University Press, Princeton, NJ.

Bicanski, A., & Burgess, N. (2019). A computational model of visual recognition memory via grid cells. *Current Biology*, *29*(6), 979-990.

Braida, L. D., & Durlach, N. I. (1972). Intensity Perception. II. Resolution in One‐Interval Paradigms. *The Journal of the Acoustical Society of America*, *51*(2B), 483-502.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature human behaviour*, *4*(12), 1265-1272.

Cover, T. M., & Thomas, J.A. (1991). *Elements of information theory*. John Wiley & Sons.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Dordek, Y., Soudry, D., Meir, R., & Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *Elife*, *5*, e10094.

Fiete, I. R., Burak, Y., & Brookings, T. (2008). What grid cells convey about rat location. *Journal of Neuroscience*, *28*(27), 6858-6871.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.

Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, *71*, 273-303.

Frege, G. (1892). Über begriff und gegenstand. *Vierteljahrsschrift für wissenschaftliche Philosophie*, *16*(2).

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1-2), 43-74.

Gravetter, F., & Lockhead, G. R. (1973). Criterial range as a frame of reference for stimulus judgment. *Psychological Review*, *80*(3), 203.

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801-806.

Hayworth, K. J. (2012). Dynamically partitionable autoassociative networks as a solution to the neural binding problem. *Frontiers in computational neuroscience*, *6*, 73.

Hinton, G. E., & Plaut, D. C. (1987, July). Using fast weights to deblur old memories. In *Proceedings of the 9th annual conference of the cognitive science society* (pp. 177-186).

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498-525.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), 16390-16395.

Krotov, D., & Hopfield, J. J. (2016). Dense associative memory for pattern recognition. *Advances in neural information processing systems*, *29*.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, *1*(0).

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences*, *83*(19), 7508-7512.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*(3), 105-117.

Lisman, J. E., & Idiart, M. A. (1995). Storage of 7+/-2 short-term memories in oscillatory subcycles. *Science*, *267*(5203), 1512-1515.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279-281.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347-356.

MacKay, D. J. C. (1991). Maximum entropy connections: Neural networks. In *Maximum entropy and Bayesian methods* (pp. 237-244). Springer, Dordrecht.

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of experimental psychology: general*, *111*(1), 1.

Marcus, G. F. (2003). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, *88*(5), 375.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, *15*(3), 267-273.

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science*, *1*, 311-360.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, *24*(6), 745-749.

Pollack, I. (1953). The information of elementary auditory displays. II. *The Journal of the Acoustical Society of America*, *25*(4), 765-769.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation?. *Psychological science*, *19*(6), 607-614

Roskies, A. L. (1999). The binding problem. *Neuron*, *24*(1), 7-9.

Rouder, J. N. (2001). Absolute identification with simple and complex stimuli. *Psychological Science*, *12*(4), 318-322.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, *2*(6), 459-473.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

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

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, *119*(4), 807.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, *46*(1-2), 159-216.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, *20*(11), 1643-1653.

Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M. B., & Moser, E. I. (2012). The entorhinal grid map is discretized. *Nature*, *492*(7427), 72-78

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652-654.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 331.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological review*, *101*(1), 80.

Van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*, e34963.

Von Der Malsburg, C. (1994). The correlation theory of brain function. In *Models of neural networks* (pp. 95-119). Springer, New York, NY.

Von der Malsburg, C. (1999). The what and why of binding: the modeler's perspective. *Neuron*, *24*(1), 95-104.

Wei, X. X., Prentice, J., & Balasubramanian, V. (2015). A principle of economy predicts the functional architecture of grid cells. *Elife*, *4*, e08362.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of vision*, *4*(12), 11-11.

Szabó, Z. G. (2000). *Problems of compositionality*. Routledge.