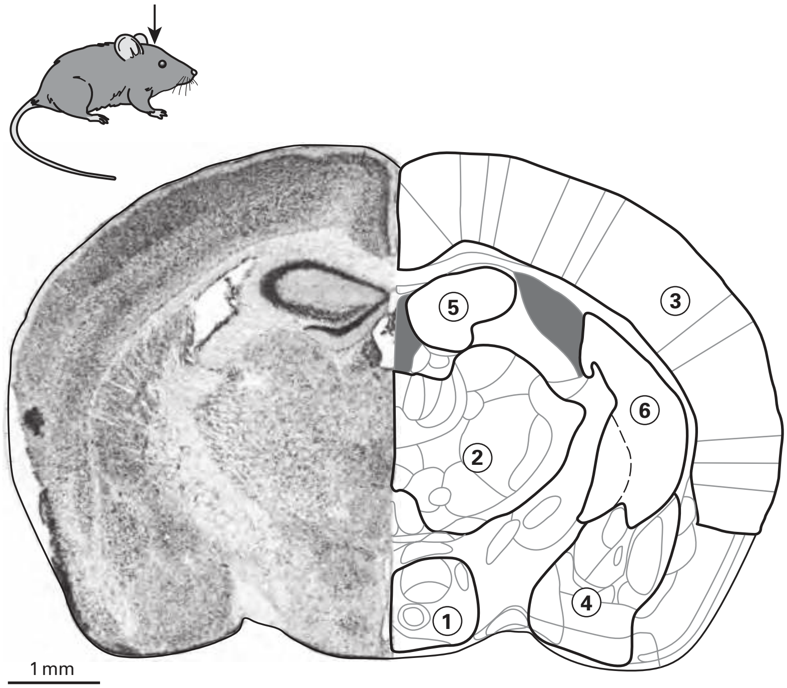
3 Why a Bigger Brain?

This chapter will explain why, despite the worm's success with 302 neurons, brains expand. The mouse cerebral cortex contains about 10^7 neurons. This seems like a lot until you consider that the cortex of the macaque monkey, a key experimental model, is larger by 100-fold, and that human cortex is 10-fold larger still (Herculano-Houzel, 2011). Despite this huge range of scales, one feels comfortable generalizing about the "mammalian brain"—because every part identified in mouse can also be identified in macaque and human (figure 3.1; Kaas, 2005).

Consider also the fly brain. It has 500-fold fewer neurons than the mouse brain, but 500-fold more neurons than the worm brain, plus a rich structure—so warranting a slot in the "large brain" category. Insect and mammal brains share many similarities. For example, both gather their neurons into clusters and their axons into cables (*tracts*). Both employ special structures to accomplish the same broad tasks: store high-level input patterns, generate low-level output patterns, and retrieve patterns using reduced instructions. Of course, there are differences, given the differences in body design and behavior. Yet, despite half a billion years of evolutionary opportunity to diverge, brain designs in insect and mammal seem to have followed the same rules.

For designs to have persisted across this immensity of time and spatial scale implies that they are neither arbitrary nor accidental. Rather, they must have emerged as responses to some broad constraint. That is what elevates the shared responses to the status of *principles*. This chapter will identify the key constraint and indicate how it leads to three principles that govern the organization of larger brains.



① Generate patterns for wireless signaling and appetitive behaviors.

② “Preprocessing” to shape signals for higher processing.

③ High-level processing: assemble larger patterns, choose behaviors.

④ “Tag” high-level patterns for emotional significance.

⑤ Store and recall.

⑥ Evaluate reward predictions.

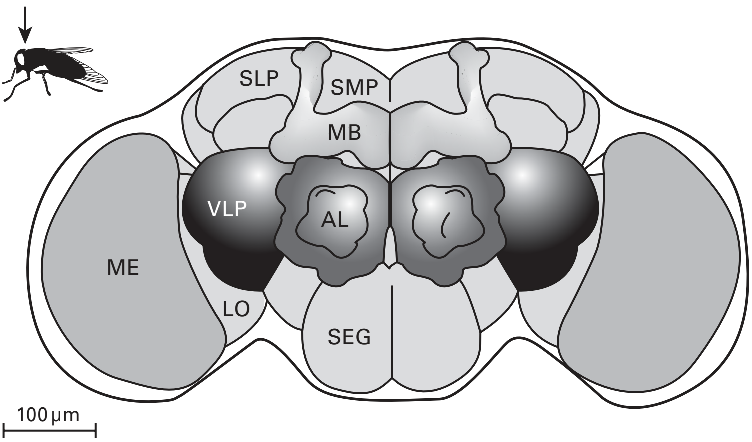


Figure 3.1
Mammalian and insect brains share many broad aspects of design. **Upper**: Cross section through mouse brain; inset indicates plane of section. **Left**: Fine dots are neurons; dark regions are neuron clusters; bright regions are myelinated tracts (chapter 4). **Right**: Numbered regions dedicated to core tasks: (1) *hypothalamus*; (2) *thalamus*; (3) *cerebral cortex*; (4) *amygdaloid complex*; (5) *hippocampus*; (6) *striatum*. Reprinted with modifications and permission from Franklin and Paxinos (1996). **Lower**: Cross section through fly brain; inset indicates plane of section. Brain is built of more than fifty clusters, each specialized for particular tasks. Depicted here are ME, medulla—detect and map local visual patterns; LO, lobula—assemble local visual patterns into larger patterns; AL, antennal lobe—preprocess olfactory signals for pattern recognition; VLP, ventrolateral protocerebrum; SLP, superior lateral protocerebrum; SMP, superior medial protocerebrum—all involved in high-level integration; MB, mushroom body—store and recall; SEG, subesophageal ganglion—integrate information for wired and wireless output to body.

A brain's core tasks

As animals emerge from the soil to a wider, less viscous world, the possibilities for foraging expand immensely. A worm explores mainly in two dimensions over an area of 0.01 m^2 whereas a honeybee typically covers an area of nearly 10^7 m^2, and a fly somewhat less. So foraging area expands by 10^9 (1 billionfold). Add the third dimension, and the volume to be explored becomes astronomical. Larger animals, such as fish, birds, and mammals, may migrate and thus forage over thousands of kilometers—thus millions of square kilometers.

Such gigantic territories contain immense resources and, of course, harbor innumerable dangers. For an animal to find the one and avoid the other requires it to rapidly gather vast amounts of information from the environment. To calibrate "vast" with one example, the eye sends the brain about 10 megabits per second, roughly the rate of an Ethernet connection (Koch et al., 2006). All sense data reach the brain in the form of tiny patterns—evanescent pieces of a dynamic jigsaw puzzle—and to be of any use, they require assembly to reveal a larger pattern. So if gathering information is to be at all rewarding, the brain must commit resources to assembling larger patterns on spatial and temporal scales that are relevant to behavior.

Yet, even a larger pattern might be useless until it is compared to a library of stored patterns where it can be identified: *edible/toxic*, *friend/foe*, or *search item not found*. Either outcome provides a basis for behavioral choice. A match allows confident choice: eat or decline, approach or flee. A nonmatch suggests caution and need to gather more data. Thus, the brain requires "pattern comparators," and these must couple to mechanisms that select behaviors: *feed*, *fight*, *copulate*, *investigate*. These, in turn, couple to mechanisms for detailed motor patterns to drive muscles for moving limbs or wings.

Any given motor behavior *might* match exactly the action that was ordered: the arrow might strike the exact point at which it was aimed. But often there are errors due to environmental or neural perturbations, and these need to be identified, so that performance can progressively improve. Thus, a brain needs mechanisms to evaluate the mismatch between the orders it gave and the actual motor performance. So, in addition to sensing and processing patterns to discover "what's important out there," the brain also devotes considerable resources to sensing and processing its own motor errors, and other errors of internal "intentional" signaling in order to improve the accuracy and efficiency of the next round. This is "motor learning."

Behaviors are subject to another important class of errors. Every action has both costs and consequences. The costs are partly energetic: how much energy was spent? But also there are "opportunity costs": could the return have been greater and the risk less for some different action? Every behavior, even when perfectly executed, needs to be evaluated from this perspective: wise or foolish? repeat or not? These evaluations of *reward prediction*, like those for motor errors, are used to update stored knowledge in order to improve the next round of predictions. The nematode worm already shows this type of evaluation to some degree, but animals in the wider world allot it major neural resources.

In sum, to succeed in the wider world, an animal must exchange larger amounts of information with its external environment and also evaluate the costs and consequences of its actions. The seven core tasks that every brain must accomplish are summarized in figure 3.2. What the brain does for the external environment it also does for the internal environment which has also expanded and complexified. Moreover, the mechanisms for managing the internal and external environments need to couple closely in order to serve each other (figure 3.2).

Why the internal milieu needs a brain

To support richer external behaviors, an animal requires specialized internal tissues and organs. Some digest the bounty foraged from the outer

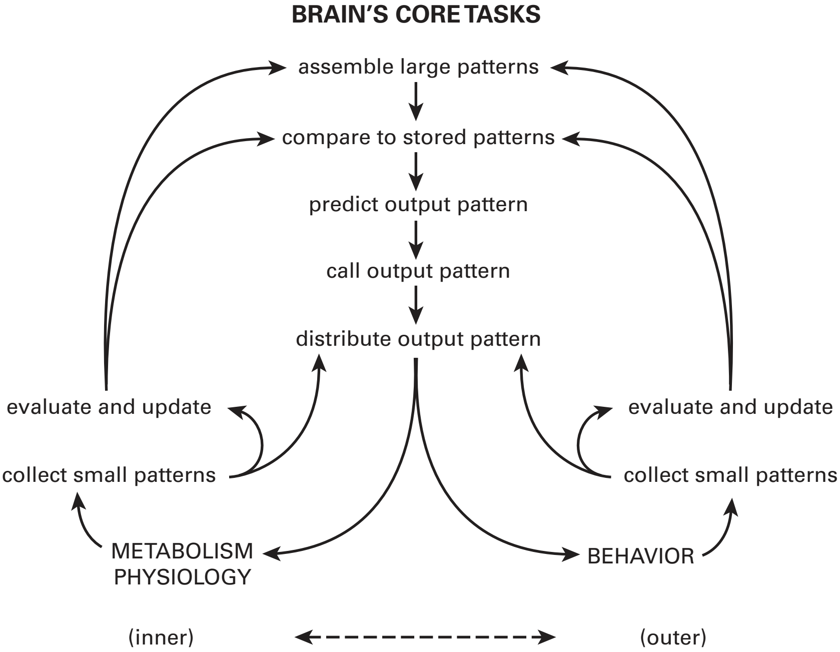


Figure 3.2
Large brains accomplish the same broad tasks. Note that inner and outer tasks couple to serve *each other* (↔).

world; others store metabolites and energy-rich compounds for release upon demand. Still others regulate ionic balance and cleanse the internal milieu, or distribute oxygen and metabolites to hungry tissues. Specialized organs of immunity protect against infectious agents and parasites. Organs couple to form systems, and systems cross-couple to optimize overall function.

The standard idea is that the internal systems more or less take care of themselves. Each parameter is supposed to have a set point, like a thermostat, from which deviations trigger feedback to correct the mismatch (*homeostasis*). Internal regulation also employs *autonomic nerves*—so termed because they are in some sense independent of voluntary control—thus, autonomous. We cannot "will" our heart to beat faster or our blood pressure to decrease. However, we can accomplish these shifts by recalling or imagining the appropriate scene. This implies the existence of neural pathways from pattern stores to pattern generators for autonomic circuits. Thus, although the autonomic nerves are generally supposed to serve emergencies (“fight or flight”), they actually serve continuous regulation—not just for panic, but for efficiency.

Efficient regulation anticipates

In fact, all internal regulation, even the mildest sort, is far from autonomous. As the external environment presents opportunity or cause for concern, internal processes must predict what the external environment is about to deliver and must prepare particular responses that will probably be needed in support. For internal processes the goal is not to correct mismatches but to prevent them.

Such predictive regulation was demonstrated for feeding and digestion by Ivan Pavlov more than a century ago: the brain processes small patterns from the outside (sight or smell of some substance) and matches them to a stored pattern that identifies a particular food. Then the brain triggers secretions all along the digestive system to prepare for what’s coming, starting in the mouth (if bread, then amylase; if fat, then lipase), then on to the stomach (if meat, then acid plus protease), the intestine (if fat, then bile), and finally the circulation (if glucose, then insulin). All of these secretions occur *before* and *during* the meal, triggered *predictively*—anticipating what will be coming down the gastrointestinal tract—thus preparing systems for absorption and uptake in order to prevent deviations that would need correction by negative feedback (Fu et al., 2011).

Modern work extends this point: as the stomach releases its contents to the next stage, it also signals the brain to prepare for the next bout of foraging. The brain responds by tuning up sensitivity of the olfactory receptors and by increasing the rate of sniffing (Julliard et al., 2007; Tong et al., 2011). Thus, the stomach warns the brain “Prepare to forage again”—well before the body has begun to deplete its reserves. Moreover, as fat reaches the small intestine, the gut can predict confidently the approach of satiety. Therefore, the gut warns the brain “cease feeding and proceed to the next activity”[1] (Fry et al., 2007).

Each “next activity” requires the brain to predict continuously, and in timely fashion, the need for a particular blood pressure. Consider the record of mean arterial pressure over 24 hours (figure 3.3). In early afternoon, as the subject attends a lecture, his brain anticipates reduced demand and allows him to doze: pressure falls. Startled awake by the jab of a pin, the brain predicts danger: pressure spikes; then, identifying a prank, the brain directs the nap to resume: pressure falls. At midnight the subject has sexual intercourse: pressure spikes, but then falls profoundly and stays low during sleep. Come morning, the brain predicting a busy day, restores the pressure.

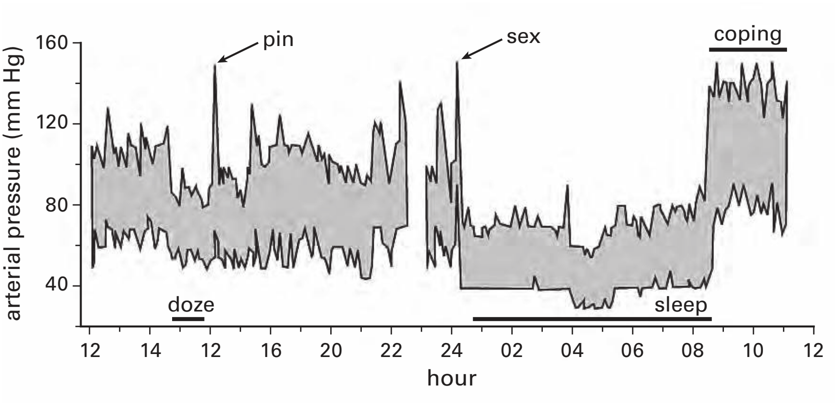


Figure 3.3
Internal systems match behavior. Arterial pressure fluctuates with demand. Each shift in pressure is accompanied by parallel shifts in hormonal and neural signaling that follow the broad catabolic and anabolic patterns. Redrawn from Bevan et al. (1969) and reprinted from Sterling (2004b).

Such anticipatory tuning requires coordinated action of multiple organs and organ systems. To raise pressure, the heart accelerates and vessels constrict. Also the kidney expands blood volume by pumping more salt water into the circulation. The kidney also signals the brain that the body will soon need more supplies of salt and water. Thus, like the gastrointestinal tract, the kidney alerts the brain well in advance of an upcoming need to resupply. Each contribution operates on a different timescale: faster for heart and vessels, slower for kidney's pumping, and still slower for the brain's rise of salt appetite and thirst. These contributions to internal regulation are all initiated simultaneously—and largely by the same signals.

In short, every move we make is matched by a corresponding cardiovascular and renal pattern. Of this we are generally unaware. Yet if the motor command ("Arise!") slightly precedes the internal command ("Tighten vessels!"), blood flow to the head drops, and we faint. That this experience, *postural hypotension*, occurs rarely attests to the rigorous coupling between the cardiovascular pattern and muscular patterns on a 100-ms timescale. On a slower timescale "Arise!" increases by eightfold a signal to the kidney to save water.[2]

Note that matching blood pressure to environmental context requires all of the brain's broad tasks as diagramed in figure 3.2—the collecting and assembling of patterns, the comparison to stores, and so forth. How else to decide if the jab is from a friend or enemy? Moreover, every high-level call to external action is delivered simultaneously to multiple internal organs. Thus, collecting patterns and distributing patterns are both thoroughly coupled between inner and outer worlds. Where and how the brain effects this coupling will be treated in chapter 4.

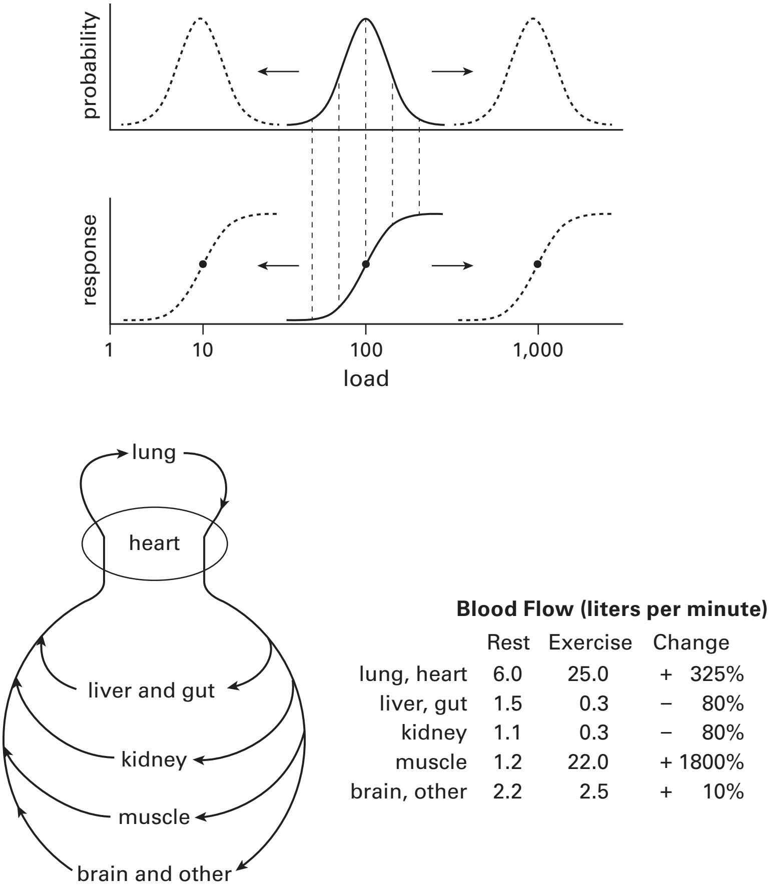
Adapt, match, trade

Although this book concerns efficient neural design, we must keep in mind that the brain comprises only 2% of the body's mass and 20% of its energy. So the body also needs to operate efficiently. Each organ should match its capacity to the anticipated need of the organ downstream. Too little and the system will fail; too much and capacity is wasted. So each organ needs constant tuning to anticipate the next demand (figure 3.4). But what happens when a need exceeds the capacity to supply? This problem is solved by arranging various short-term "trade-offs." Such cooperation enhances the range of performance while greatly reducing average excess capacity (figure 3.4).

For example, the "resting" heart pumps 6 L of blood per minute through the respiratory system and then out to the general circulation. Resting skeletal muscle uses about 20% of the oxygenated blood—matched to its modest need for maintaining posture. During peak exercise, muscle must increase its supply by nearly 20-fold, but the pulmonary and systemic circulation can increase their outputs only fourfold. Therefore, the body must either reduce its peak capacity for exercise or increase its peak pulmonary and cardiovascular capacity by fivefold—imagine the chest! Or it can borrow.

Indeed, during peak exercise the splanchnic circulation (gut and liver) and the renal circulation (kidney) both reduce their shares by four- to fivefold, enough to pay part of muscle's bill for exercise. During digestion, when the splanchnic circulation needs more blood, it borrows from muscle and skin—unless skin needs blood for cooling. The brain neither makes loans nor allows overdrafts that might cause it to overheat. Anyone who has eaten and then exercised in the sun will recall how these conflicting demands from muscle, gut, and skin are resolved: by corrective motor commands to internal systems ("Vomit!") and to external systems ("Lie down!"). Moreover, the experience receives a strongly negative evaluation that updates the knowledge store ("Do not repeat!").

This example illustrates three key rules for efficient regulation: (1) adapt response capacity to changes in input level, (2) match response capacities across the system, and (3) trade between systems. Regulatory responses begin promptly—as soon as there is sufficient statistical evidence to predict



Blood Flow (liters per minute)

	Rest	Exercise	Change
lung, heart	6.0	25.0	+ 325%
liver, gut	1.5	0.3	− 80%
kidney	1.1	0.3	− 80%
muscle	1.2	22.0	+ 1800%
brain, other	2.2	2.5	+ 10%

Figure 3.4

Adapt, match, trade. Upper: Adapt response capacity to load. Every system confronts some distribution of probable loads (bold). As conditions shift, so does the distribution (dashed). The response curve (bold) is typically sigmoid with its most sensitive region (steep part) matched to the most probable loads. As a sensor detects a statistically reliable change in the distribution, it prepares the effectors by shifting their response curves to match the new distribution (dashed). Each sensor also adapts its own sensitivity. Reprinted from Sterling (2004b). **Lower:** Organs and organ systems couple efficiently by matching loads to capacities. Trade-offs allow better performance while reducing unused capacity and enhancing "portability." Blood flow pattern changes with exercise: total flow quadruples, but that is insufficient for muscle. To meet the full need, blood is routed from liver, gut, and kidney, temporarily reducing their performance but eventually benefiting from what the muscular effort has accomplished. Data from Weibel (2000).

a new target level. By comparison, self-regulation by feedback to a set point would be hopelessly inefficient. But to execute these principles of predictive regulation requires an organ with knowledge of the outside, knowledge of the inside, and knowledge of the past to anticipate what the whole animal will need over various timescales—the whole brain (Sterling, 2012).

Bigger brains

We seem to have answered "Why a bigger brain?" In a wider world, a more effective brain expands the possibilities for behavior. Control of appendages such as fins, wings, and legs lends speed and scope to exploration, so that vastly more small patterns are encountered which then require selection and assembly. More large patterns require more comparisons, requiring a larger library; more comparisons also require more decisions, and these require more evaluation. Naturally, more neurons are needed, and since neuronal components are irreducibly small (chapter 7), a brain must enlarge.[3]

The larger brain, to be effective, must operate in real time. One need not watch a sloth for very long to realize the limits to life in slow motion. The larger, faster brain must still remain portable and also metabolically affordable. So a brain needs to be both functionally effective and cost-effective. These demands for speed, portability, and affordability all interact; therefore, individually and together they raise questions of brain design. We turn now to the fundamental constraint on any brain design that leads to the first three design principles. Then, in the context of these few principles, we discuss some actual designs (mammal and insect).

Design constraints

The fundamental constraint on brain design emerges from a law of physics. This law governs the costs of capturing, sending, and storing *information*. This law, embodied in a family of equations developed by Claude Shannon, applies equally to a telephone line and a neural cable, equally to a silicon circuit and a neural circuit. This law constrains neural design at all scales and cannot be avoided any more than a B-29 bomber can avoid the law of gravity. But, though the brain is fundamentally an organ that manipulates information, few neuroscientists are familiar with this law or aware of its value for understanding brain organization. We explain it briefly here and give more detail in chapters 5 and 6.

What *is* "information"?

Information is *the reduction of uncertainty about some situation X associated with observing any variable Y that is causally correlated with X*. Uncertainty defines the standard measure: one *bit* is the information needed to decide between two equally likely alternatives. Information depends on causality because, to reduce uncertainty, a message must be reliably relatable to its source, the event that caused it. Any factor that reduces the reliability of this connection, such as noise, increases uncertainty and destroys information.

Reduction of uncertainty succinctly describes the brain's purpose. A spike in an ON ganglion cell reduces the brain's uncertainty that a brighter than average object is located in a particular region of the visual field (chapter 11). And when the brain matches the sensory pattern coded by a patch of ganglion cells to a stored pattern, it reduces a key uncertainty: "Friend or foe?" The answer helps to select the next behavior and implement it. To this end, a motor neuron spike decreases the uncertainty that its target muscle fibers will contract and help the animal move in the appropriate direction. In short, to achieve its core purpose, the brain uses physical devices (neurons and circuits) that represent and manipulate information. So now we must ask: how much information can a neuron represent, and what constrains its capacity?

A neuron's information capacity

To convey information, a neuron must represent the state of its input as a distinct output (input and output must be causally related). It follows that a neuron's capacity to convey information is limited by the number of distinctly different outputs that it can generate. The number of different outputs a spiking neuron can generate in a given time is the number of distinctly different spike trains that it can produce in that time. This depends on two factors, mean firing rate (*R* spikes per second) and the precision of spike timing (Δt seconds). The upper bound on firing rate is set by spike duration plus the period following a spike when a neuron is refractory (cannot spike). Certain neurons reach this limit during brief bursts, but most neurons operate far below this limit. Precision is limited by channel noise and membrane time constant. Here biophysics limits information capacity.

What is the relation between spike rate, timing precision, and the number of different spike trains a neuron can produce? When a neuron transmits for 1 s, it produces *R* spikes with a timing precision of Δt (Rieke et al., 1997). The number of different spike trains, *M*, is the number of ways the neuron can place its *R* spikes in $T = 1/\Delta t$ intervals (figure 3.5). Deriving *M* is a standard exercise in calculating combinations that is often set to students in quaint terms, such as placing peas in pots. The solution is

$$M = T!/(R!(T - R)!), \tag{3.1}$$

where ! denotes factorial and $(T - R)$ is the number of empty (spikeless) intervals.

The number of different messages, *M*, that a neuron can generate in 1 s converts to information rate. According to Shannon, the information, *H*, is given by

$$H = \log_2(M). \tag{3.2}$$

Substituting for *M* using (3.1) gives

$$H = \log_2(T!/(R!(T - R(!)) = \log_2(T!) — \log_2(R!) — \log_2((T - R)!). \tag{3.3}$$

Because Shannon used a logarithmic scale, a message lasting twice as long conveys twice as much information. And, because he used log base 2, information is in bits. Thus, *H*, the information that a neuron can transmit with messages 1 s long, is its information capacity in bits per second (figure 3.5).

With this expression we can "follow the money." That is, using a standard currency (bits) we can ask like good engineers: how fast does a neuron send information (bits per second) and how efficiently (bits per spike)? And at what cost in space (bits per cubic millimeter) and energy (bits per molecule of adenosine tri-phosphate)? This molecule, abbreviated *ATP*, is the standard intracellular molecule for transferring energy.

Information costs energy and space

Information rate increases with spike rate and with spike timing precision, that is, reduction in Δt. However, for any given precision, information rate increases sublinearly with spike rate (figure 3.5). Consequently, as spike rate rises, bits per spike should fall, and this theoretical decline in bits per spike is observed experimentally (figure 3.5).

There is another way to explain why more frequent spikes carry less information. A symbol that occurs less frequently is more surprising and so more informative (chapter 4, equation 4.2). This effect, which Shannon called *surprisal*, makes a code with fewer spikes more efficient. For example, a code that distributes spikes sparsely among a population of neurons conveys more bits per spike (chapter 12; Levy & Baxter, 1996).

This simple law—infrequent spikes carry more bits—profoundly influences neural design because, following the money, one finds that spikes are

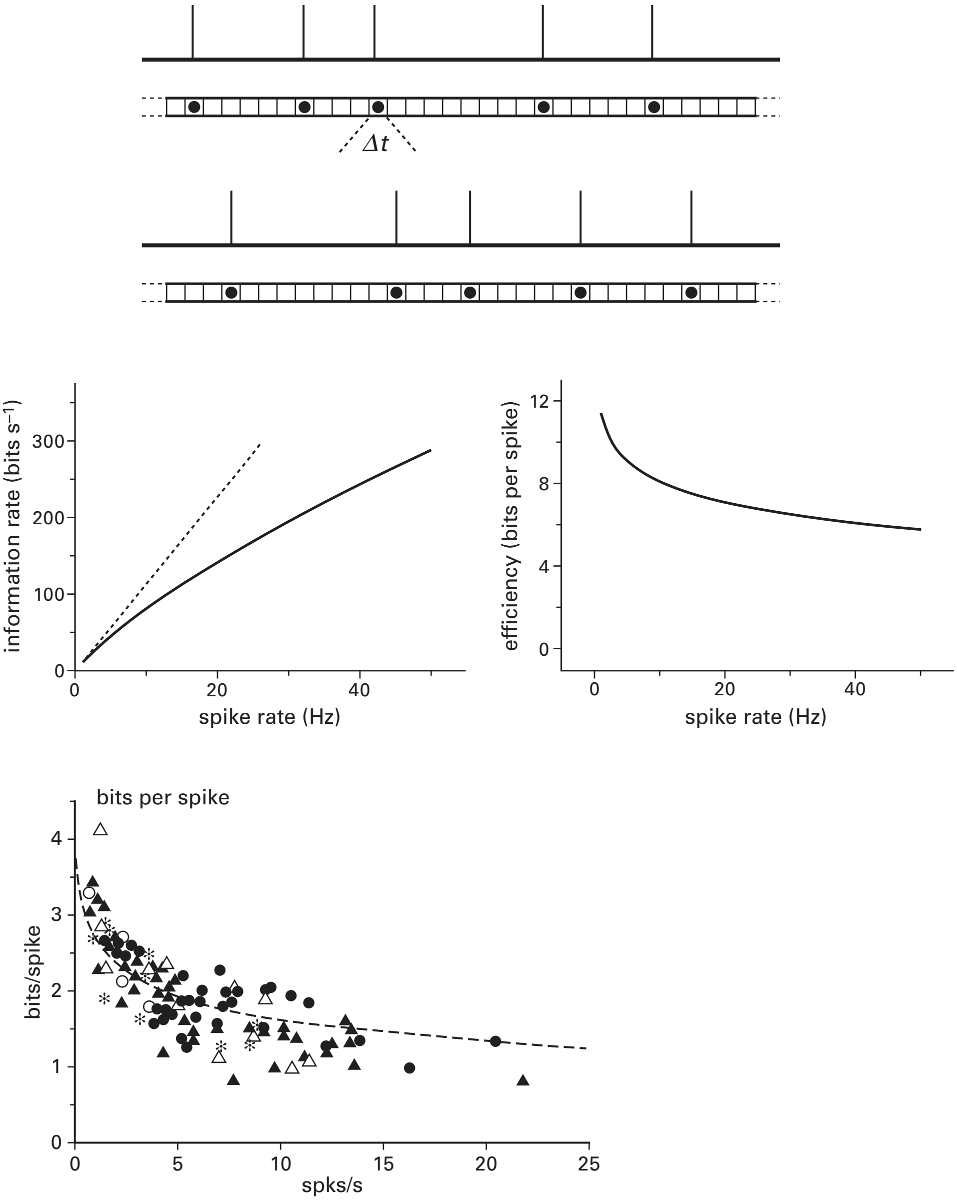


Figure 3.5

Mathematics and biophysics govern the representational capacity of signal trains. Upper: Distinct sequences of spikes in time intervals Δt represent different inputs. **Middle left**: Theory predicts information rate to increase sublinearly with spike rate, with the consequence shown at **middle right**: Increasing spike rate reduces the information transmitted per spike. These theoretical curves were calculated using the standard approximation for signal entropy at low spike rates (Rieke et al., 1997, equation 3.22). In general neurons do not achieve their theoretical capacity because of noise and redundancy; consequently, measured values of bits/spike are lower (figure 11.25). **Lower**: Measured bits per spike falls as mean spike rate increases. Data pooled from several classes of guinea pig retinal ganglion cell. Reprinted with permission from Balasubramanian & Sterling (2009).

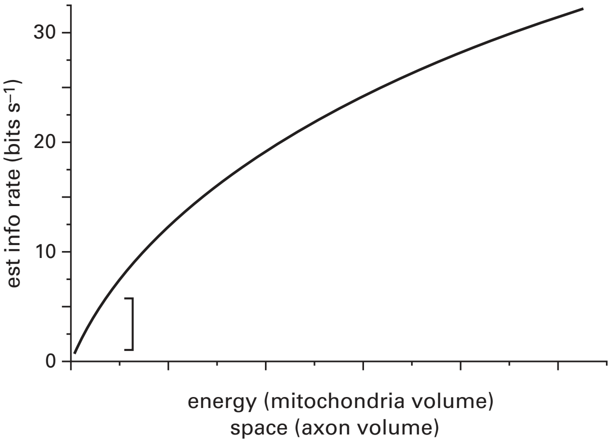


Figure 3.6

Law of diminishing returns. Doubling information rate of retinal ganglion cells more than doubles space and energy costs. Consequently, neural designs try to stay on the steep region of this empirically measured curve. Modified from Balasubramanian & Sterling (2009) and reprinted with permission.

expensive. They use about 20% of the brain's energy (Attwell & Laughlin, 2001; Sengupta et al., 2010). A spike charges a neuron's membrane capacitance by about 100 mV, and the membrane area is substantial due to a neuron's local branching. Higher mean spike rates require a larger cell body with greater membrane area; this increases energy cost per spike and adds to the cost of transmitting bits at high rates. Consequently, where spikes are sent sporadically and at low mean rates, more information can be sent for the same energy—more bits per ATP. This saving in energy by low rates is compounded by a saving in space.

Higher spike rates also require thicker axons.[4] Because axon diameter, *d*, increases directly with firing rate, axon volume rises as d^2; therefore, doubling the firing rate quadruples axon volume. The concentration of mitochondria, an indicator of energy cost, tends to be constant with axon diameter; therefore, as volume quadruples, so does the energy supply (Perge et al., 2009, 2012). In summary, there is a *law of diminishing returns*: cost per bit, both in energy and space, rises steeply with bit rate (figure 3.6).

Three principles of neural design

The inescapable cost of sending any information and the disproportionate cost of sending at higher rates lead to three design principles: *send only what is needed*; *send at the lowest acceptable rate*; *minimize wire, that is, length and* *diameter of all neural processes*. This last principle seems obvious, but it actually reflects a subtle point that arises from the constraint on rate.

Designs should reduce wire, of course, because wire uses space and energy. But wires also use *time* for transmission, and that is time lost to processing and action (Howarth et al., 2012). The constraint is particularly onerous for neural wires because they transmit more slowly than copper wire. Neural conduction velocity is 100 millionfold lower and, for biophysical reasons, faster conduction requires thicker wires (chapter 7). Thus saving time by sending at higher information rates (bits per second) and higher conduction velocities (meters per second) requires thicker axons, which, as noted, involves disproportionate costs in energy and space (Wen & Chklovskii, 2005). Thus, the only economical way to save time is to rigorously shorten wires. This principle shapes brain design across all scales, from an axon's branching and the microscopic design of local circuits, to the overall layout (chapter 13).

With these few principles we can now consider how the mammalian and fly brains are organized on a scale of about 1 mm and why. This macro-organization cannot explain the actual computations because those occur mostly on a finer scale. Nor do we claim that every feature represents the best of all possible designs. Others might work just as well—but they have not been tested. All we can say is that these three principles illuminate the layout of real brains—across a millionfold range of scale and half a billion years of evolution.