Chapter Revision History

The table notes major changes between revisions. Minor changes such as small clarifications or formatting changes are not noted.

Version	Date	Changes	Principal Author(s)
0.4		Initial release	J. Hawkins
0.41	May 23, 2016	Fixed references to the time frame of human neocortex development.	
0.42	June 22, 2016	Added figure 1	J. Hawkins
0.43	Oct 13, 2016	Corrected neocortex size reference	

Hierarchical Temporal Memory: Overview

In the September 1979 issue of Scientific American, Nobel Prize-winning scientist Francis Crick wrote about the state of neuroscience. He opined that despite the great wealth of factual knowledge about the brain we had little understanding of how it actually worked. His exact words were, "What is conspicuously lacking is a broad framework of ideas within which to interpret all these different approaches" (Crick, 1979). Hierarchical Temporal Memory (HTM) is, we believe, the broad framework sought after by Dr. Crick. More specifically, HTM is a theoretical framework for how the neocortex works and how the neocortex relates to the rest of the brain to create intelligence. HTM is a theory of the neocortex and a few related brain structures; it is not an attempt to model or understand every part of the human brain. The neocortex comprises about 75% of the volume of the human brain, and it is the seat of most of what we think of as intelligence. It is what makes our species unique.

HTM is a biological theory, meaning it is derived from neuroanatomy and neurophysiology and explains how the biological neocortex works. We sometimes say HTM theory is "biologically constrained," as opposed to "biologically inspired," which is a term often used in machine learning. The biological details of the neocortex must be compatible with the HTM theory, and the theory can't rely on principles that can't possibly be implemented in biological tissue. For example, consider the pyramidal neuron, the most common type of neuron in the neocortex. Pyramidal neurons have tree-like extensions called dendrites that connect via thousands of synapses. Neuroscientists know that the dendrites are active processing units, and that communication through the synapses is a dynamic, inherently stochastic process (Poirazi and Mel, 2001). The pyramidal neuron is the core information processing element of the neocortex, and synapses are the substrate of memory. Therefore, to understand how the neocortex works we need a theory that accommodates the essential features of neurons and synapses. Artificial Neural Networks (ANNs) usually model neurons with no dendrites and few highly precise synapses, features which are incompatible with real neurons. This type of artificial neuron can't be reconciled with biological neurons and is therefore unlikely to lead to networks that work on the same principles as the brain. This observation doesn't mean ANNs aren't useful, only that they don't work on the same principles as biological neural networks. As you will see, HTM theory explains why neurons have thousands of synapses and active dendrites. We believe these and many other biological features are essential for an intelligent system and can't be ignored.

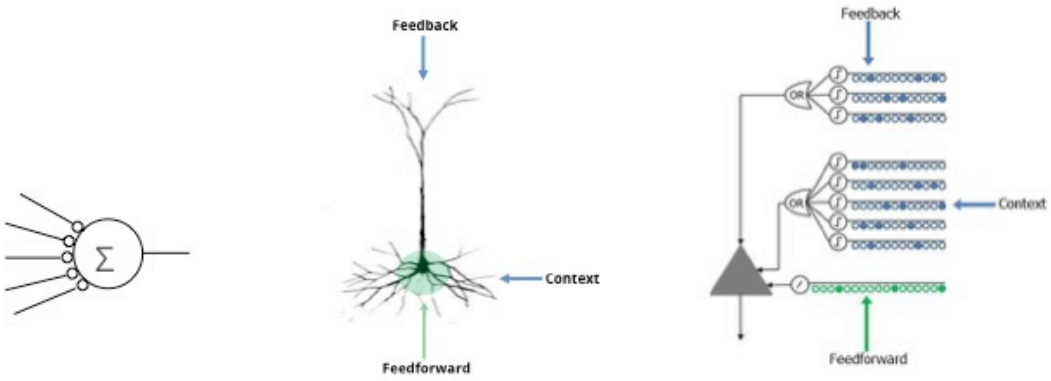

Figure 1a Figure 1b Figure 1c

Figure 1 Biological and artificial neurons. *Figure 1a* shows an artificial neuron typically used in machine learning and artificial neural networks. Often called a "point neuron" this form of artificial neuron has relatively few synapses and no dendrites. Learning in a point neuron occurs by changing the "weight" of the synapses which are represented by a scalar value that can take a positive or negative value. A point neuron calculates a weighted sum of its inputs which is applied to a non-linear function to determine the output value of the neuron. *Figure 1b* shows a pyramidal neuron which is the most common type of neuron in the neocortex. Biological neurons have thousands of synapses arranged along dendrites. Dendrites are active processing elements allowing the neuron to recognize hundreds of unique patterns. Biological synapses are partly stochastic and therefore are low precision. Learning in a biological neuron is mostly due to the formation of new synapses and the removal of unused synapses. Pyramidal neurons have multiple synaptic integration zones that receive input from different sources and have differing effects on the cell. *Figure 1c* shows an HTM artificial neuron. Similar to a pyramidal neuron it has thousands of synapses arranged on active dendrites. It recognizes hundreds of patterns in multiple integration zones. The HTM neuron uses binary synapses and learns by modeling the growth of new synapses and the decay of unused synapses. HTM neurons don't attempt to model all aspects of biological neurons, only those that are essential for information theoretic aspects of the neocortex.

Although we want to understand how biological brains work, we don't have to adhere to all the biological details. Once we understand how real neurons work and how biological networks of neurons lead to memory and behavior, we might decide to implement them in software or in hardware in a way that differs from the biology in detail, but not in principle. But we shouldn't do that before we understand how biological neural systems work. People often ask, "How do you know which biological details matter and which don't?" The answer is: once you know how the biology works, you can decide which biological details to include in your models and which to leave out, but you will know what you are giving up, if anything, if your software model leaves out a particular biological feature. Human brains are just one implementation of intelligence; yet today, humans are the only things everyone agrees are intelligent. Our challenge is to separate aspects of brains and neurons that are essential for intelligence from those aspects that are artifacts of the brain's particular implementation of intelligence principles. Our goal isn't to recreate a brain, but to understand how brains work in sufficient detail so that we can test the theory biologically and also build systems that, although not identical to brains, work on the same principles.

Sometime in the future designers of intelligent machines may not care about brains and the details of how brains implement the principles of intelligence. The field of machine intelligence may by then be so advanced that it has departed from its biological origin. But we aren't there yet. Today we still have much to learn from biological brains and therefore to understand HTM principles, to advance HTM theory, and to build intelligent machines, it is necessary to know neuroscience terms and the basics of the brain's design.

Bear in mind that HTM is an evolving theory. We are not done with a complete theory of neocortex, as will become obvious in the remainder of this chapter and the rest of the book. There are entire sections yet to be written, and some of what is written will be modified as new knowledge is acquired and integrated into the theory. The good news is that although we have a long way to go for a complete theory of neocortex and intelligence, we have made significant progress on the fundamental aspects of the theory. The theory includes the representation format used by the neocortex ("sparse distributed representations" or SDRs), the mathematical and semantic operations that are enabled by this representation format, and how neurons throughout the neocortex learn sequences and make predictions, which is the core component of all inference and behavior. We also understand how knowledge is stored by the formation of sets of new synapses on the

dendrites of neurons. These are basic elements of biological intelligence analogous to how random access memory, busses, and instruction sets are basic elements of computers. Once you understand these basic elements, you can combine them in different ways to create full systems.

The remainder of this chapter introduces the key concepts of Hierarchical Temporal Memory. We will describe an aspect of the neocortex and then relate that biological feature to one or more principles of HTM theory. In-depth descriptions and technical details of the HTM principles are provided in subsequent chapters.

Biological Observation: The Structure of the Neocortex

The human brain comprises several components such as the brain stem, the basal ganglia, and the cerebellum. These organs are loosely stacked on top of the spinal cord. The neocortex is just one more component of the brain, but it dominates the human brain, occupying about 75% of its volume. The evolutionary history of the brain is reflected in its overall design. Simple animals, such as worms, have the equivalent of the spinal cord and nothing else. The spinal cord of a worm and a human receives sensory input and generates useful, albeit simple, behaviors. Over evolutionary time scales, new brain structures were added such as the brainstem and basal ganglia. Each addition did not replace what was there before. Typically the new brain structure received input from the older parts of the brain, and the new structure's output led to new behaviors by controlling the older brain regions. The addition of each new brain structure had to incrementally improve the animal's behaviors. Because of this evolutionary path, the entire brain is physically and logically a hierarchy of brain regions.

Figure 2 a) real brain b) logical hierarchy *(placeholder)*

The neocortex is the most recent addition to our brain. All mammals, and only mammals, have a neocortex. The neocortex first appeared about 200 million years ago in the early small mammals that emerged from their reptilian ancestors during the transition of the Triassic/Jurassic periods. The modern human neocortex separated from those of monkeys in terms of size and complexity about 25 million years ago (Rakic, 2009). The human neocortex continued to evolve to be bigger and bigger, reaching its current size in humans between 800,000 and 200,000 years ago[1]. In humans, the neocortex is a sheet of neural tissue about the size of a large dinner napkin (2,500 square centimeters) in area and 2.5mm thick. It lies just under the skull and wraps around the other parts of the brain. (From here on, the word "neocortex" will refer to the human neocortex. References to the neocortex of other mammals will be explicit.) The neocortex is heavily folded to fit in the skull but this isn't important to how it works, so we will always refer to it and illustrate it as a flat sheet. The human neocortex is large both in absolute terms and also relative to the size of our body compared to other mammals. We are an intelligent species mostly because of the size of our neocortex.

The most remarkable aspect of the neocortex is its homogeneity. The types of cells and their patterns of connectivity are nearly identical no matter what part of the neocortex you look at. This fact is largely true across species as well. Sections of human, rat, and monkey neocortex look remarkably the same. The primary difference between the neocortex of different animals is the size of the neocortical sheet. Many pieces of evidence suggest that the human neocortex got large by replicating a basic element over and over. This observation led to the 1978 conjecture by Vernon Mountcastle that every part of the neocortex must be doing the same thing. So even though some parts of the neocortex process vision, some process hearing, and other parts create language, at a fundamental level these are all variations of the same problem, and are solved by the same neural algorithms. Mountcastle argued that the vision regions of the neocortex are vision regions because they receive input from the eyes and not because they have special vision neurons or vision algorithms (Mountcastle, 1978). This discovery is incredibly important and is supported by multiple lines of evidence.

Even though the neocortex is largely homogenous, some neuroscientists are quick to point out the differences between neocortical regions. One region may have more of a certain cell type, another region has extra layers, and other regions may exhibit variations in connectivity patterns. But there is no question that neocortical regions are remarkably similar and that the variations are relatively minor. The debate is only about how critical the variations are in terms of functionality.

The neocortical sheet is divided into dozens of regions situated next to each other. Looking at a neocortex you would not see any regions or demarcations. The regions are defined by connectivity. Regions pass information to each other by sending bundles of nerve fibers into the white matter just below the neocortex. The nerve

[1] http://humanorigins.si.edu/human-characteristics/brains

fibers reenter at another neocortical region. The connections between regions define a logical hierarchy. Information from a sensory organ is processed by one region, which passes its output to another region, which passes its output to yet another region, etc. The number of regions and their connectivity is determined by our genes and is the same for all members of a species. So, as far as we know, the hierarchical organization of each human's neocortex is the same, but our hierarchy differs from the hierarchy of a dog or a whale. The actual hierarchy for some species has been mapped in detail (Zingg, 2014). They are complicated, not like a simple flow chart. There are parallel paths up the hierarchy and information often skips levels and goes sideways between parallel paths. Despite this complexity the hierarchical structure of the neocortex is well established.

We can now see the big picture of how the brain is organized. The entire brain is a hierarchy of brain regions, where each region interacts with a fully functional stack of evolutionarily older regions below it. For most of evolutionary history new brain regions, such as the spinal cord, brain stem, and basal ganglia, were heterogeneous, adding capabilities that were specific to particular senses and behaviors. This evolutionary process was slow. Starting with mammals, evolution discovered a way to extend the brain's hierarchy using new brain regions with a homogenous design, an algorithm that works with any type of sensor data and any type of behavior. This replication is the beauty of the neocortex. Once the universal neocortical algorithms were established, evolution could extend the brain's hierarchy rapidly because it only had to replicate an existing structure. This explains how human brains evolved to become large so quickly.

Figure 3 a) brain with information flowing posterior to anterior b) logical hierarchical stack showing old brain regions and neocortical regions *(placeholder)*

Sensory information enters the human neocortex in regions that are in the rear and side of the head. As information moves up the hierarchy it eventually passes into regions in the front half of the neocortex. Some of the regions at the very top of the neocortical hierarchy, in the frontal lobes and also the hippocampus, have unique properties such as the ability for short term memory, which allows you to keep a phone number in your head for a few minutes. These regions also exhibit more heterogeneity, and some of them are older than the neocortex. The neocortex in some sense was inserted near the top of the old brain's hierarchical stack. Therefore as we develop HTM theory, we first try to understand the homogenous regions that are near the bottom of the neocortical hierarchy. In other words, we first need to understand how the neocortex builds a basic model of the world from sensory data and how it generates basic behaviors.

HTM Principle: Common Algorithms

HTM theory focuses on the common properties across the neocortex. We strive not to understand vision or hearing or robotics as separate problems, but to understand how these capabilities are fundamentally all the same, and what set of algorithms can see AND hear AND generate behavior. Initially, this general approach makes our task seem harder, but ultimately it is liberating. When we describe or study a particular HTM learning algorithm we often will start with a particular problem, such as vision, to understand or test the algorithm. But we then ask how the exact same algorithm would work for a different problem such as understanding language. This process leads to realizations that might not at first be obvious, such as vision being a primarily temporal inference problem, meaning the temporal order of patterns coming from the retina is as important in vision as is the temporal order of words in language. Once we understand the common algorithms of the neocortex, we can ask how evolution might have tweaked these algorithms to achieve even better performance on a particular problem. But our focus is to first understand the common algorithms that are manifest in all neocortical regions.

HTM Principle: Hierarchy

Every neocortex, from a mouse to a human, has a hierarchy of regions, although the number of levels and number of regions in the hierarchy varies. It is clear that hierarchy is essential to form high-level percepts of the world from low-level sensory arrays such as the retina or cochlea. As its name implies, HTM theory incorporates the concept of hierarchy. Because each region is performing the same set of memory and algorithmic functions, the capabilities exhibited by the entire neocortical hierarchy have to be present in each region. Thus if we can understand how a single region works and how that region interacts with its hierarchical neighbors, then we can build hierarchical models of indefinite complexity and apply them to any type of sensory/motor system. Consequently most of current HTM theory focuses on how a single neocortical region works and how two regions work together.

Biological Observation: Neurons are Sparsely Activated

The neocortex is made up of neurons. No one knows exactly how many neurons are in a human neocortex, but recent "primate scale up" methods put the estimate at 86 billion (Herculano-Houzel, 2012). The moment-to-moment state of the neocortex, some of which defines our perceptions and thoughts, is determined by which neurons are active at any point in time. An active neuron is one that is generating spikes, or action potentials. One of the most remarkable observations about the neocortex is that no matter where you look, the activity of neurons is sparse, meaning only a small percentage of them are rapidly spiking at any point in time. The sparsity might vary from less than one percent to several percent, but it is always sparse.

HTM Principle: Sparse Distributed Representations

The representations used in HTM theory are called Sparse Distributed Representations, or SDRs. SDRs are vectors with thousands of bits. At any point in time a small percentage of the bits are 1's and the rest are 0's. HTM theory explains why it is important that there are always a minimum number of 1's distributed in the SDR, and also why the percentage of 1's must be low, typically less than 2%. The bits in an SDR correspond to the neurons in the neocortex.

SDRs have some essential and surprising properties. For comparison, consider the representations used in programmable computers. The meaning of a word in a computer is not inherent in the word itself. If you were shown 64 bits from a location in a computer's memory you couldn't say anything about what it represents. At one moment in the execution of the program the bits could represent one thing and at another moment they might represent something else, and in either case the meaning of the 64 bits can only be known by relying on knowledge not contained in the physical location of the bits themselves. With SDRs, the bits of the representation encode the semantic properties of the representation; the representation and its meaning are one and the same. Two SDRs that have 1 bits in the same location share a semantic property. The more 1 bits two SDRs share, the more semantically similar are the two representations. The SDR explains how brains make semantic generalizations; it is an inherent property of the representation method. Another example of a unique capability of sparse representations is that a set of neurons can simultaneously activate multiple representations without confusion. It is as if a location in computer memory could hold not just one value but twenty simultaneous values and not get confused! We call this unique characteristic the "union property" and it is used throughout HTM theory for such things as making multiple predictions at the same time.

The use of sparse distributed representations is a key component of HTM theory. We believe that all truly intelligent systems must use sparse distributed representations. To become facile with HTM theory, you will need to develop an intuitive sense for the mathematical and representational properties of SDRs.

Biological Observation: The Inputs and Outputs of the Neocortex

As mentioned earlier, the neocortex appeared recently in evolutionary time. The other parts of the brain existed before the neocortex appeared. You can think of a human brain as consisting of a reptile brain (the old stuff) with a neocortex (literally "new layer") attached on top of it. The older parts of the brain still have the ability to sense the environment and to act. We humans still have a reptile inside of us. The neocortex is not a stand-alone system, it learns how to interact with and control the older brain areas to create novel and improved behaviors.

There are two basic inputs to the neocortex. One is data from the senses. As a general rule, sensory data is processed first in the sensory organs such as the retina, cochlea, and sensory cells in the skin and joints. It then goes to older brain regions that further process it and control basic behaviors. Somewhere along this path the neurons split their axons in two and send one branch to the neocortex. The sensory input to the neocortex is literally a copy of the sensory data that the old brain is getting.

The second input to the neocortex is a copy of motor commands being executed by the old parts of the brain. For example, walking is partially controlled by neurons in the brain stem and spinal cord. These neurons also split their axons in two, one branch generates behavior in the old brain and the other goes to the neocortex. Another example is eye movements, which are controlled by an older brain structure called the superior colliculus. The axons of superior colliculus neurons send a copy of their activity to the neocortex, letting the neocortex know what movement is about to happen. This motor integration is a nearly universal property of the brain. The neocortex is told what behaviors the rest of the brain is generating as well as what the sensors are sensing. Imagine what would happen if the neocortex wasn't informed that the body was moving in some

way. If the neocortex didn't know the eyes were about to move, and how, then a change of pattern on the optic nerve would be perceived as the world moving. The fact that our perception is stable while the eyes move tells us the neocortex is relying on knowledge of eye movements. When you touch, hear, or see something, the neocortex needs to distinguish changes in sensation caused by your own movement from changes caused by movements in the world. The majority of changes on your sensors are the result of your own movements. This "sensory-motor" integration is the foundation of how most learning occurs in the neocortex. The neocortex uses behavior to learn the structure of the world.

Figure 4 showing sensory & motor command inputs to the neocortex (block diagram) *(placeholder)*

No matter what the sensory data represents - light, sound, touch or behavior - the patterns sent to the neocortex are constantly changing. The flowing nature of sensory data is perhaps most obvious with sound, but the eyes move several times a second, and to feel something we must move our fingers over objects and surfaces. Irrespective of sensory modality, input to the neocortex is like a movie, not a still image. The input patterns completely change typically several times a second. The changes in input are not something the neocortex has to work around, or ignore; instead, they are essential to how the neocortex works. The neocortex is memory of time-based patterns.

The primary outputs of the neocortex come from neurons that generate behavior. However, the neocortex never controls muscles directly; instead the neocortex sends its axons to the old brain regions that actually generate behavior. Thus the neocortex tries to control the old brain regions that in turn control muscles. For example, consider the simple behavior of breathing. Most of the time breathing is controlled completely by the brain stem, but the neocortex can learn to control the brain stem and therefore exhibit some control of breathing when desired.

A region of neocortex doesn't "know" what its inputs represent or what its output might do. A region doesn't even "know" where it is in the hierarchy of neocortical regions. A region accepts a stream of sensory data plus a stream of motor commands. From these inputs it learns of the changes in the inputs. The region will output a stream of motor commands, but it only knows how its output changes its input. The outputs of the neocortex are not pre-wired to do anything. The neocortex has to learn how to control behavior via associative linking.

HTM Principle: Sensory Encoders

Every HTM system needs the equivalent of sensory organs. We call these "encoders." An encoder takes some type of data–it could be a number, time, temperature, image, or GPS location–and turns it into a sparse distributed representation that can be digested by the HTM learning algorithms. Encoders are designed for specific data types, and often there are multiple ways an encoder can convert an input to an SDR, in the same way that there are variations of retinas in mammals. The HTM learning algorithms will work with any kind of sensory data as long as it is encoded into proper SDRs.

One of the exciting aspects of machine intelligence based on HTM theory is that we can create encoders for data types that have no biological counterpart. For example, we have created an encoder that accepts GPS coordinates and converts them to SDRs. This encoder allows an HTM system to directly sense movement through space. The HTM system can then classify movements, make predictions of future locations, and detect anomalies in movements. The ability to use non-human senses offers a hint of where intelligent machines might go. Instead of intelligent machines just being better at what humans do, they will be applied to problems where humans are poorly equipped to sense and to act.

HTM Principle: HTM Systems are Embedded Within Sensory-motor Systems

To create an intelligent system, the HTM learning algorithms need both sensory encoders and some type of behavioral framework. You might say that the HTM learning algorithms need a body. But the behaviors of the system do not need to be anything like the behaviors of a human or robot. Fundamentally, behavior is a means of moving a sensor to sample a different part of the world. For example, the behavior of an HTM system could be traversing links on the world-wide web or exploring files on a server.

It is possible to create HTM systems without behavior. If the sensory data naturally changes over time, then an HTM system can learn the patterns in the data, classify the patterns, detect anomalies, and make predictions of future values. The early work on HTM theory focuses on these kinds of problems, without a behavioral component. Ultimately, to realize the full potential of the HTM theory, behavior needs to be incorporated fully.

HTM Principle: HTM Relies On Streaming Data and Sequence Memory

The HTM learning algorithms are designed to work with sensor and motor data that is constantly changing. Sensor input data may be changing naturally such as metrics from a server or the sounds of someone speaking. Alternately the input data may be changing because the sensor itself is moving such as moving the eyes while looking at a still picture. At the heart of HTM theory is a learning algorithm called Temporal Memory, or TM. As its name implies, Temporal Memory is a memory of sequences, it is a memory of transitions in a data stream. TM is used in both sensory inference and motor generation. HTM theory postulates that every excitatory neuron in the neocortex is learning transitions of patterns and that the majority of synapses on every neuron are dedicated to learning these transitions. Temporal Memory is therefore the substrate upon which all neocortical functions are built. TM is probably the biggest difference between HTM theory and most other artificial neural network theories. HTM starts with the assumption that everything the neocortex does is based on memory and recall of sequences of patterns.

HTM Principle: On-line Learning

HTM systems learn continuously, which is often referred to as "on-line learning". With each change in the inputs the memory of the HTM system is updated. There are no batch learning data sets and no batch testing sets as is the norm for most machine learning algorithms. Sometimes people ask, "If there are no labeled training and test sets, how does the system know if it is working correctly and how can it correct its behavior?" HTM builds a predictive model of the world, which means that at every point in time the HTM-based system is predicting what it expects will happen next. The prediction is compared to what actually happens and forms the basis of learning. HTM systems try to minimize the error of their predictions.

Another advantage of continuous learning is that the system will constantly adapt if the patterns in the world change. For a biological organism this is essential to survival. HTM theory is built on the assumption that intelligent machines need to continuously learn as well. However, there will likely be applications where we don't want a system to learn continuously, but these are the exceptions, not the norm.

Conclusion

The biology of the neocortex informs HTM theory. In the following chapters we discuss details of HTM theory and continue to draw parallels between HTM and the neocortex. Like HTM theory, this book will evolve over time. At first release there are only a few chapters describing some of the HTM Principles in detail. With the addition of this documentation, we hope to inspire others to understand and use HTM theory now and in the future.

References

Crick, F. H.C. (1979) Thinking About the Brain. Scientific American September 1979, pp. 229, 230. Ch. 4 27

Poirazi, P. & Mel, B. W. (2001) Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. Neuron, 2001 doi:10.1016/S0896-6273(01)00252-5

Rakic, P. (2009). Evolution of the neocortex: Perspective from developmental biology. Nature Reviews Neuroscience. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2913577/

Mountcastle, V. B. (1978) An Organizing Principle for Cerebral Function: The Unit Model and he Distributed System, in Gerald M. Edelman & Vernon V. Mountcastle, ed., 'The Mindful Brain' , MIT Press, Cambridge, MA , pp. 7-50

Zingg, B. (2014) Neural networks of the mouse neocortex. Cell, 2014 Feb 27;156(5):1096-111. doi: 10.1016/j.cell.2014.02.023

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proceedings of the National Academy of Sciences of the United States of America, 109 (Suppl 1), 10661–10668. http://doi.org/10.1073/pnas.1201895109