 Published in *American Scientist* 77 (July-August 1989), 333-335.

SPARSE DISTRIBUTED MEMORY

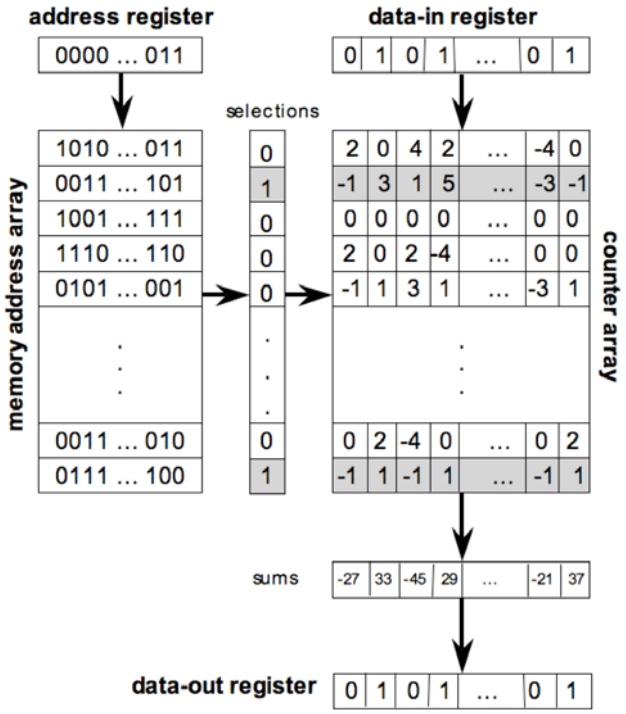
Peter J. Denning

ABSTRACT: Sparse Distributed Memory was proposed by Pentti Kanerva as a model of human long term memory. He presented it as an architecture that could store large patterns and retrieve them based on partial matches with current sensory inputs. The architecture can be realized as a neural net or as an associative memory. SDM exhibits behaviors, both in theory and in experiment, that resemble those previously unapproachable by machines -- e.g., rapid recognition of faces or odors, discovery of new connections between seemingly unrelated ideas, continuation of a sequence of events when given a cue from the middle, knowing that one doesn't know, or getting stuck with an answer on the tip of one's tongue. These behaviors are now within reach of machines that can be incorporated into the computing systems of robots capable of seeing, talking, and manipulating. Kanerva's theory is a new interpretation of learning and cognition that respects biology and the mysteries of individual human beings.

Recognizing your mother's face in a crowd. Experiencing a flood of old memories an instant after sniffing an odor you haven't smelled for years. Seeing a connection that no one ever taught you between two concepts. Discovering that an idea that seemed to have occurred to you spontaneously was actually given to you by a friend in a conversation last year. Recognizing that a particular leaf is a maple. Humming the rest of a familiar tune when given a phrase from the middle. Knowing that you don't know the answer to a question. Knowing that you do know the answer to a question, but that it is inaccessibly perched on the tip of your tongue.

These everyday phenomena illustrate capabilities of human beings that we do not know how to reproduce with a machine but that would be very useful if we could. The failure of artificial intelligence to produce machines with any of these capabilities after forty years of research is not a failure of intention. It is a failure of the rationalistic philosophy deeply rooted in Western thought *(1)*. That philosophy has produced in many disciplines a search for models that combine context-free (meaningless) elements into systems governed by formal laws. Not only have information-processing models of cognition fallen short in computer science, corresponding formal models have fallen short in anthropology, economics, linguistics, political science, psychology, and other disciplines. These shortcomings have prompted a new examination of what it means to be human, a search for a philosophy that respects the mystery of individuals and the biological roots of all learning.

Against this background, the emergence of Pentti Kanerva's theory of sparse distributed memory is refreshingly welcome *(2)*. Kanerva departs from the formalistic tradition to develop an architecture of memory, inspired by biology, in which the phenomena I mentioned in the first paragraph can arise holistically. Because his



This schematic diagram shows the relations among the components of sparse distributed memory. The memory in this example stores and retrieves 256-bit patterns across 2,000 physical locations. Each horizontal row is a location. The input pattern (cue) in the address register is compared simultaneously to all 2,000 patterns in the memory address array; each line in the array holds the address of one location. The distances from each address pattern are compared with the memory's built-in threshold radius and a subset of the locations is selected (*shaded areas*). The 256-bit data pattern is stored at the selected locations by adding 1 to each counter in the counter array corresponding to each 1 in the pattern and subtracting 1 from each counter corresponding to a 0 in the pattern. A 256-bit pattern is retrieved by forming 256 sums from the corresponding counters in each selected location and then forming a 1 output bit in the data-out register for each sum that is nonnegative and a 0 for each sum that is negative. The retrieved pattern is a statistical reconstruction determined from the contents of all selected locations. All selections can be done in parallel, and all data bits can be handled in parallel, giving the memory great speed over a wide range of pattern widths and physical locations.

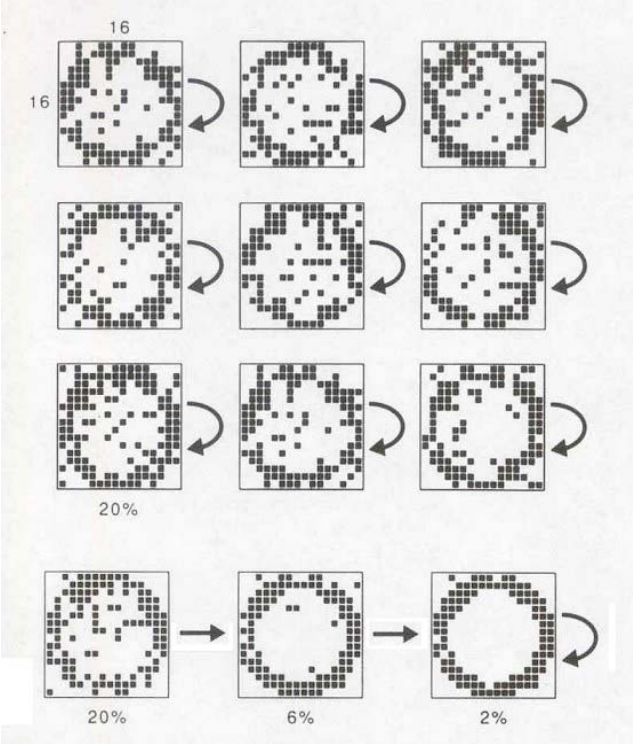
theory deals with patterns recalled statistically from patterns previously stored across large regions of the memory space, he does not insist that anyone can ever know precisely how the phenomena arise. In what follows, I will describe the central ideas of sparse distributed memory; I would encourage you to read the details in Kanerva's book.

The theory begins with an interpretation of human long-term memory as a storage system that associates sensory input patterns quickly with actions that are appropriate for the situation. In Kanerva's model, sensory input is represented in the form of very long bit vectors containing thousands or tens of thousands of bits. Because no two external situations are identical, the memory must respond to partial matches between the current sensory pattern and previously stored patterns. The measure of dissimilarity between patterns is the number of bits in which they differ, a metric known as the Hamming distance. For example, the distance between 01101 and 10111 is 3 bits.

Kanerva illustrates his design with an example of 1,000-bit patterns, giving rise to a space of 2^{1000} possible patterns. In this space, 1/1000 of the patterns are within 451 bits of any given pattern, and all but 1/1000 of the patterns are within 549 bits. The extremely large number of patterns that are so close (±49 bits) to the mean distance of 500 bits between two random patterns is crucial to the memory's ability to make connections between patterns that seemingly have little to do with each other.

Ordinary (random-access) computer memories are designed around a simple idea: Within a few nanoseconds after a memory cue (address) is presented for a read operation, the memory responds with an output pattern (data). High speed is achieved by associating one physical location with each possible address. Current technology limits the designs to about 25 address bits and 64 data bits, nowhere near the pattern lengths needed for simulation of human long-term memory.

Kanerva proposes an architecture that encompasses an affordable number of physical locations (say 1,000,000) and a large pattern size (say 1,000 bits). Each location is assigned an address (1,000-bit pattern) at random, and the set of location addresses constitutes a sparse subset of the memory space. The memory has an input register for the cue (address) pattern and an input register for the data pattern, and it has a register for an output pattern (these registers each hold 1,000 bits). Each location has



Each of the nine patterns at the top of the figure was stored in a simulated sparse distributed memory by addressing the memory with the pattern itself. Each pattern is a 16x16 array of bits that transforms into a 256-bit vector. The three figures at the bottom show the result of an iterative search in which the result of the first retrieval was used as the input cue for the second retrieval. The final output pattern was none of the patterns stored. Because each of the nine stored patterns was constructed from an O with 20% of the bits randomly reversed, this behavior may be interpreted as the memory's ability to extract a signal from noise. Another interpretation is that the memory formed a statistical interpolation among the stored patterns; the new pattern is stable (it will retrieve itself) and thus serves as a conceptualization of the data.

an address decoder that compares its own address with the input cue, selecting that location as a participant in the next storage or retrieval operation whenever the cue is within distance *d* of the location's address. Kanerva demonstrates that the address decoders can be built of linear threshold circuits -- gates that produce a 1 at their output whenever the number of 1s among their many inputs is at least 1000-*d* -- and notes a similarity of operation between these circuits and neurons in the nervous systems of many animals.

Kanerva recommends *d*=451 for the 1,000,000-location memory of 1,000-bit patterns. With these parameters, approximately 1/1000 of the physical locations will be selected by any given input cue. How are storage and retrieval carried out with this arrangement?

To store a 1,000-bit data pattern at address *A*, the memory works as follows. The input cue pattern *A* is presented to the memory, and all locations within 451 bits of *A* select themselves. This set of selected locations is called the sphere selected by *A*. A copy of the input data pattern, which is to be associated with *A*, is then entered into each of the selected locations. Because any given location is within the spheres of selection of many distinct cue patterns, entering a new value must not obliterate the previous contents of the location. This is accomplished by implementing each location as a set of 1,000 counters, one for each bit position of the data. Data are entered by adding 1 to each counter for which the corresponding data bit is 1, and subtracting 1 from each counter for which the corresponding data bit is 0. Kanerva calculates that 8-bit counters are adequate for most applications.

To retrieve a 1,000-bit pattern corresponding to input cue *A*, the memory works as follows. The sphere of selected locations is formed as described above. A set of 1,000 output counter values is constructed from all the selected locations by summing all the corresponding selected counters; for example, the counter in output bit position 2 is the sum of the bit-2 counters of each selected physical location. The 1,000-bit output pattern is constructed from the 1,000 output counters by a threshold method: if an output counter is nonnegative, that output bit is 1, otherwise it is 0.

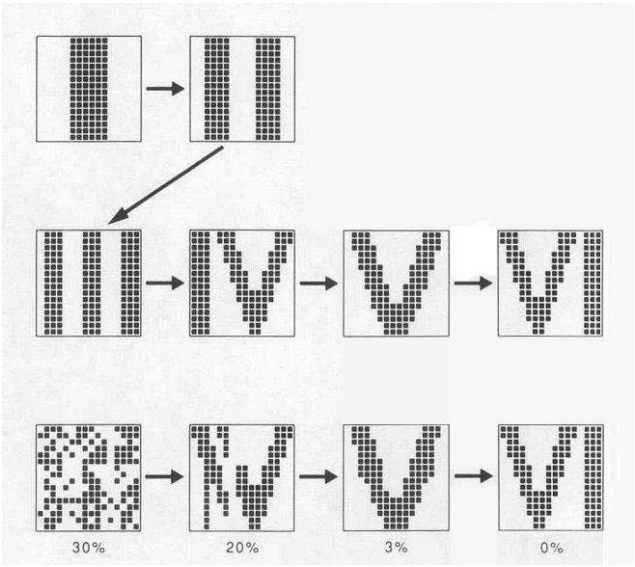
The rationale for the name is now obvious: the memory is *sparse* because the physical locations are a vanishingly small subset of the memory space; it is *distributed* because a pattern is stored in many locations and retrieved by statistical reconstruction from many locations. Distribution enables the memory to retrieve a stored pattern when the input cue only partially matches any stored pattern, an ability that arises from the large overlap between the spheres of selected locations of two similar cues. It also renders the memory robust in case of failure of portions of the addressing or storage hardware.

Each storage and retrieval can be carried out with massively parallel operations among the address decoders and counters, allowing the memory to respond rapidly. At the NASA Ames Research Center, David Rogers has built a simulator of the sparse distributed memory running on a 32,768-node Connection Machine 2 of the Thinking Machines Corporation; it simulates 250,000 locations with 256-bit patterns, with cycle time of about 1/2 of a second.

Let us consider again the phenomena mentioned at the start of this essay. The memory's ability to retrieve patterns associated with sensory input quickly could allow it to recognize instantly your mother's face or a long-forgotten odor. The memory can form associations between patterns without ever being explicitly taught those associations because the distance between two patterns is sufficiently small that the one pattern retrieves the other. Similarly the memory can retrieve a forgotten pattern from some cue that seemingly had nothing to do with it, giving the impression of generating a new pattern. It can retrieve the pattern corresponding to "maple leaves" that was formed internally after storing many patterns encoding specific maple leaves. It can store patterns in lists representing their temporal order, and begin an iterative retrieval from anywhere in the list. Fast convergence of an iterative search can be interpreted as "knowing that you know" and nonconvergence as "knowing that you don't know;"' the tip-of-the-tongue phenomenon would occur somewhere between these two cases.

It is important to remember that the theory predicts that these phenomena will occur in sparse distributed memory, but it cannot predict the details. It cannot predict which connections you might see between ideas, which concepts you will form, or what will be on the tip of your tongue.

Kanerva began to develop his theory in the early 1970s. He did so independently of James Albus and David Marr, who developed similar theories from observation of the structure of the human nervous system and the cerebellum *(3,4)*. These theories have the distinguishing feature that they can be readily tested; they have thus inspired much work with simulators that verify their mathematical properties and predictions.



The six patterns at the top were stored as a list in a simulated sparse distributed memory by storing each pattern as the data associated with the previous pattern in the sequence. The four patterns at the bottom resulted from an iterative search, beginning with a noisy version of the third pattern and culminating with a clean version of the sixth. This behavior may be interpreted as the memory's ability to locate the remainder of a temporal sequence given a pattern that is similar to one of the members. This behavior will occur even when the sequence stored in the memory is noisy, suggesting that the memory can generate an abstract form of a sequence.

Albus's theory also emphasizes the hierarchical organization of the nervous system and suggests that associative memory and sensory encoding may be organized into levels. All these theories are consistent with the biological theory of learning proposed by Maturana and Varela *(5)*.

The sparse distributed memory is intended as an integral component of a larger system that includes sensory apparatus and a scheme for encoding sensory input into binary patterns. Such a system also includes motors that act when driven with stimulus patterns. Kanerva calls this an autonomous learning system. It includes a component called the focus that contains a pattern updated constantly from both sensory input and the contents of the sparse distributed memory and that generates the patterns used to drive motors. The focus represents the current moment of consciousness, which continuously changes as the sensory input and the context retrieved from memory change.

A major research area is the design of sensory encoders. How does visual input get encoded so that the patterns stored in memory are relatively insensitive to small rotations, translations, zooms, and pans of the visual field? Or so that certain shapes are easily detectable within any visual field? How does speech input get encoded so that the same word produces similar patterns independently of the speaker? How does tactile input get encoded so that different surface textures are distinguishable? These and similar questions are occupying Kanerva and his colleagues, who seek to build prototypes of devices that recognize visual shapes, continuous speech, and fine textures. A theory of Robert Erickson about how the power of visual systems arises from large numbers of simple components illustrates a possible sensory-encoding system that might mesh well with sparse distributed memory *(6)*.

The theory cannot predict which connections you might see between ideas, which concepts you will form, or what will be on the tip of your tongue.

David Rogers has been studying the sparse distributed memory as a statistical inference machine. In one experiment, he fed in a stream of patterns, each derived from a vector of measurements of 15 weather-related factors from a four-hour interval at a weather station in Darwin, Australia. There were 50,000 vectors covering about 23 years of observations. The 15 components of each vector were encoded as a 256-bit pattern that was the storage address of the single bit indicating rain in the subsequent four-hour period. Rogers modified the operation of the memory so that the address array was dynamically altered to add addresses similar to those associated with rain, and delete addresses not associated with rain. At the end of the experiment, the address array identified the combinations of bits that were the most reliable predictors of rain in the data.

However promising his theory is, Pentti Kanerva advises that it is not a final answer. It is only a step in a line of investigation whose final outcomes cannot be predicted. His theory opens the possibility that machines can perform some of the actions of which we are capable, while leaving plenty of room for the biological roots of intelligence and the mysteries of each human being.

References

1. T. Winograd and F. Flores. 1986. *Understanding Computers and Cognition.* Addison-Wesley.
2. P. Kanerva. 1988. *Sparse Distributed Memory.* Bradford Books of MIT Press.
3. J. Albus. 1981. *Brains, Behavior, and Robotics.* Byte Books of McGraw-Hill.
4. D. Marr. 1971. Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London 262*. 23-81.
5. H. Maturana and F. Varela. 1987. *The Tree of Knowledge.* Shambhala New Science Library.
6. R. Erickson. 1984. On the neural bases of behavior. *American Scientist 72*, 3 (May-June). 233-241.