Information compression via the matching and unification of patterns as a unifying principle in the workings of brains and nervous systems

J Gerard Wolff[*]

July 11, 2017

Abstract

This paper presents evidence for the idea that much of the workings of brains and nervous systems may be understood as compression of information via the matching and unification of patterns. Information compression can mean selective advantage for any creature: in the efficient storage and transmission of information; and, owing to the close connection between information compression and concepts of prediction and probability, in the making of predictions about where food may be found, potential dangers, and so on. Several aspects of our everyday perceptions and thinking may be seen as information compression. For example, many words in natural languages may be seen as relatively short identifiers or "codes" for relatively complex concepts. When viewing the world with two eyes, we see one view, not two. Random-dot stereograms provide confirmation that, in binocular vision, we do indeed merge information from our two eyes and thus compress it. Information compression may be seen in the workings of sensory units in the eye of *Limulus*, the horseshoe crab. Computer models demonstrate how information compression may be a key to the unsupervised discovery of grammars for natural language, including segmental structures (words and phrases), classes of structure, and abstract patterns. Information compression may be seen in the perceptual *constancies*, including size constancy, lightness constancy, and colour constancy. Mathematics, which is a product of the human intellect, may be seen to be a set of techniques for the compression

[*] Dr Gerry Wolff BA (Cantab) PhD (Wales) CEng MIEEE MBCS; CognitionResearch.org, Menai Bridge, UK; jgw@cognitionresearch.org; +44 (0) 1248 712962; +44 (0) 7746 290775; *Skype*: gerry.wolff; *Web*: www.cognitionresearch.org.

of information, and their application. The *SP theory of intelligence* provides evidence for the importance of information compression in several aspects of human intelligence. Four objections to the main thesis of this paper are described, with answers to those objections.

Keywords: information compression, intelligence, perception, learning, cognition

1 Introduction

> "Fascinating idea! All that mental work I've done over the years, and what have I got to show for it? A goddamned zipfile! Well, why not, after all?" (John Winston Bush, 1996).

This paper describes observations and arguments in support of the idea that much of the workings of brains and nervous systems may be understood as compression of information via the matching and unification of patterns. This idea will be referred to as "BICMUP", short for "Brain: Information Compression via the Matching and Unification of Patterns" (see also Section 2.3). The aim here is to review, update, and extend the discussion in [33], itself the basis for [34, Chapter 2].

Observations and arguments in this paper provide some of the empirical support for the *SP theory of intelligence* (Section 2.2), and also *SP-neural*, a version of the SP theory expressed in terms of neurons and their interconnections. In that latter connection, this paper complements [39], a previous paper about SP-neural.

The next section describes some of the background to this research and some relevant general principles. Several sections that follow describe strands of evidence in support of BICMUP. And Section 16, with Appendix C, describes apparent contradictions of the BICMUP thesis, and the related *SP theory of intelligence*, and how they may be resolved.

2 Background and general principles

This section provides some background to this paper and summarises some general principles that have a bearing on BICMUP.

2.1 Approaches to information compression

There are many approaches to information compression, most of them with a mathematical flavour, and many of them described at length in books such as [21].

The orientation here is different. It derives largely from the SP theory of intelligence, introduced in Section 2.2. Amongst other things, the SP theory attempts to get below or behind the mathematics of other approaches, to focus on the relatively simple, 'primitive' idea that information compression may be understood in terms of the matching and unification of patterns.

A potential benefit is that, since this idea is relatively 'concrete' and less abstract than much of mathematics, it suggests avenues that may be explored in understanding possible mechanisms for information compression in artificial systems and in brains and nervous systems.

Another reason for this approach is that the SP theory aims to be, amongst other things, a theory of the foundations of mathematics [41], so it would not be appropriate for the theory to be too dependant on mathematics.

2.2 The SP theory of intelligence

The *SP theory of intelligence*, and its realisation in the *SP computer model*, is a unique attempt to simplify and integrate observations and arguments across artificial intelligence, mainstream computing, mathematics, and human learning, perception, and cognition, with information compression as a unifying theme.

There is an outline of the SP system in Appendix A with pointers to where fuller information may be found.

As noted in the Introduction, the observations and arguments in this paper provide some of the empirical support for the SP theory. More specifically they provide empirical support for the SP theory viewed as a theory of human intelligence. And the paper provides empirical underpinnings for *SP-neural*, described in [39].

There is other empirical support for the SP system, summarised in Section 15, and that evidence provides indirect support for BICMUP.

2.3 SP-multiple-alignment and information compression via the matching and unification of patterns

Here, some principles and techniques for the compression of information are outlined. Fuller descriptions may be found in [41, Section 2].

All of these principles and techniques may be seen to be special cases of the powerful concept of SP-multiple-alignment, a key part of the SP system, described briefly in Appendix A.1.

As noted there, the concept of SP-multiple-alignment is founded on the previously-mentioned concept of *information compression via the matching* *and unification of patterns* or "ICMUP". This means that a body of information may be compressed by searching for *redundancy* in the form of patterns that match each other, and then reducing that redundancy and thus compressing the information by merging or *unifying* two or more matching patterns to make one.

There are five main variants of ICMUP:

- *Chunking-with-codes.* With each unified "chunk" of information, give it a relatively short name, identifier, or "code", and use that as a shorthand for the chunk of information wherever it occurs. As mentioned in [41, Section 2.1], compression may be optimised by assigning shorter codes to more frequent chunks and longer codes to rarer chunks, in accordance with some such scheme as Shannon-Fano-Elias coding [9].

- *Schema-plus-correction.* This variant is like chunking-with-codes but the unified chunk of information may have variations or "corrections" on different occasions.

- *Run-length coding.* This variant may be used with any sequence of two or more copies of a pattern. In that case, it is only necessary to record one copy of the pattern, with the number of copies or tags to mark the start and end of the sequence.

- *Class-inclusion hierarchies with inheritance of attributes.* Here, there is a hierarchy of classes and subclasses, with "attributes" at each level. At every level except the top level, the subclass "inherits" the attributes of all higher levels.

- *Part-whole hierarchies with inheritance of contexts.* This is like class-inclusion hierarchies with inheritance of attributes except that the structure represents the parts and subparts of some entity. Each subpart may be seen to inherit its place in larger structures.

All these five variants of ICMUP may be modelled via the concept of *SP-multiple-alignment* in the SP system. Within that framework they may be integrated seamlessly in any combination.

2.4 Information compression and concepts of prediction and probability

It has been recognised for some time that there is an intimate relation between information compression and concepts of prediction and probability [29, 24, 25, 13].

In case this seems obscure, it makes sense in terms of ICMUP: a pattern that repeats is one that invites information compression via ICMUP, but it is also one that, via inductive reasoning, suggests what may happen in the future. As can be seen in the workings of the SP system, probabilities may be calculated from the frequencies with which different patterns occur ([34, Section 3,7], [36, Section 4.4]).

In more concrete terms, any repeating pattern—such as the association between black clouds and rain—provides a basis for prediction—black clouds suggest that rain may be on the way—and probabilities may be derived from the number of repetitions.

There is a little more detail in [41, Appendix D], and a lot more detail about how this works with the SP-multiple-alignment concept in [34, Section 3.7] and [36, Section 4.4].

3 Early work

Apart from the suggestion by William of Ockham in the 14th century that "Entities are not to be multiplied beyond necessity.", and remarks by prominent scientists about the importance of simplicity in science (summarised in [41, Section 3]), research with a bearing on BICMUP began in the 1950s and '60s after the publication of Claude Shannon's [23] "theory of communication" (later called "information theory"), and partly inspired by it. In what follows, there is a rough distinction between research with the main focus on human learning, perception, and cognition, and research that concentrates on issues in mathematics and computing.

3.1 Psychology-related research

In a paper called "Some informational aspects of visual perception", Fred Attneave [1] argues that we naturally compress visual information so that we can easily recognise something from an outline picture of it in which much of the repeated, redundant, information has been stripped away.

Also, "Common objects may be represented with great economy, and fairly striking fidelity, by copying the points at which their contours change direction maximally, and then connecting these points appropriately with a straight edge." (*ibid.*, p. 185), as shown in a drawing of a sleeping cat reproduced in Figure 1.

As indicated in Section 2.4, Satosi Watanabe picked up the baton in a paper called "Information-theoretical aspects of inductive and deductive

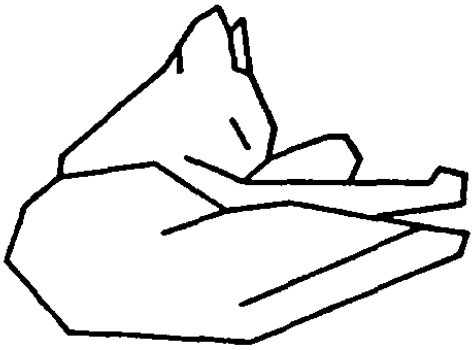

Figure 1: Drawing made by abstracting 38 points of maximum curvature from the contours of a sleeping cat, and connecting these points appropriately with a straight edge. Reproduced from Figure 3 in [1], with permission.

inference" [29]. He later wrote about the role of information compression in pattern recognition [30, 31].

Horace Barlow published a paper called "Sensory mechanisms, the reduction of redundancy, and intelligence" [2] in which he argued, on the strength of the large amounts of sensory information being fed into the human central nervous system, that "the storage and utilization of this enormous sensory inflow would be made easier if the redundancy of the incoming messages was reduced." (*ibid.* p. 537).

It is interesting to see that Barlow suggests that:

> "... the mechanism that organises [the large size of the sensory inflow] must play an important part in the production of intelligent behaviour." (*ibid.* p. 555).

and

> "... the operations required to find a less redundant code have a rather fascinating similarity to the task of answering an intelligence test, finding an appropriate scientific concept, or other exercises in the use of inductive reasoning. Thus, redundancy reduction may lead one towards understanding something about the organization of memory and intelligence, as well as pattern recognition and discrimination." [3, p. 210].

These prescient insights into the significance of information compression for the workings of human intelligence, with further discussion in [4], is a strand of thinking that has carried through into the SP theory of intelligence, with a wealth of supporting evidence.[1]

Barlow developed these and related ideas over a period of years in several papers, some of which are referenced in this paper. However, in a paper published in 2001 [5], he adopted a new position, arguing that:

> "... the [compression] idea was right in drawing attention to the importance of redundancy in sensory messages because this can often lead to crucially important knowledge of the environment, but it was wrong in emphasizing the main technical use for redundancy, which is compressive coding. The idea points to the enormous importance of estimating probabilities for almost everything the brain does, from determining what is redundant to fuelling Bayesian calculations of near optimal courses of action in a complicated world." (*ibid.* p. 242).

While there are some valid points in what Barlow says in support of his new position, his overall conclusions appear to be wrong. His main arguments are summarised in Appendix B, with what I'm sorry to say are my critical comments after each one.[2]

3.2 Mathematics- and computer-related research

Research under the general heading *minimum length encoding*, with the main emphasis on issues in mathematics and computing, is also relevant to BICMUP. This includes:

- Ray Solomonoff developed a formal theory known as *algorithmic probability* showing the intimate relation between information compression and inductive inference [24, 25].

- Chris Wallace and David Boulton explored the significance of information compression in classification and related areas in [28] and subsequent papers.

- Gregory Chaitin and Andrei Kolmogorov, working independently, built on the work of Ray Solomonoff in developing *algorithmic information*

[1] When I was an undergraduate at Cambridge University, it was fascinating lectures by Horace Barlow about the significance of information compression in the workings of brains and nervous systems, that first got me interested in those ideas.

[2] I feel apologetic about this because, as I mentioned earlier, Barlow's lectures and his earlier research relating to BICMUP have been an inspiration for me over many years.

theory. The main idea here is that the information content of a string of symbols is equivalent to the length of the shortest computer program that anyone has been able to devise that describes the string.

- Jorma Rissanen has developed related ideas in [19, 20] and other publications.

A detailed description of these and related bodies of research may be found in [13].

4 Information compression and biology

This section and those that follow describe evidence for BICMUP.

First, let's take a bird's eye view of why information compression might be important in people and other animals.

In terms of biology, information compression can confer a selective advantage to any creature:

- By allowing it to store more information in a given storage space or use less storage space for a given amount of information, and by speeding up transmission of information along nerve fibres—thus speeding up reactions—or reducing the bandwidth needed for any given volume of information.

  In connection with the last point, we have seen in Section 3.1 how Barlow [2, p. 548] draws attention to evidence that, in vertebrates at least, each optic nerve is far too small to carry reasonable amounts of the information emanating from the retina unless there is considerable compression of that information.

- Perhaps more important than the impact of information compression on the storage or transmission of information is the close connection, outlined in Section 2.4 and noted in Section 3.2, between information compression and concepts of prediction and probability. Compression of information provides a means of predicting the future from the past and estimating probabilities so that, for example, an animal may get to know where food may be found or where there may be dangers.

  As mentioned in Section 2.4, the close connection between information compression and concepts of prediction and probability makes sense in terms of ICMUP: any repeating pattern provides a basis for prediction and probabilities may be derived from the number of repetitions.

- Being able to make predictions and estimate probabilities can mean large savings in the use of energy with consequent benefits in terms of survival.

It is likely that similar principles apply in computers and other artificial systems for the processing of information.

5 Hiding in plain sight

Compression of information is so much embedded in our thinking, and seems so natural and obvious, that it is easily overlooked. Here are some examples.

5.1 Words as codes or shorthands

In the same way that “TFEU” may be a convenient code or shorthand for the rather cumbersome expression “Treaty on the Functioning of the European Union” (Appendix C.1.1), a name like “New York” is a compact way of referring to the many things of which that renowned city is composed. Likewise for the many other names that we use: “Nelson Mandela”, “George Washington”, “Mount Everest”, and so on.

More generally, most words in our everyday language stand for *classes* of things and, as such, are powerful aids to economical description. Imagine how cumbersome things would be if, on each occasion that we wanted to refer to a “table”, we had to say something like “A horizontal platform, often made of wood, used as a support for things like food, normally with four legs but sometimes three, ...”, like the slow *Entish* language of the Ents in Tolkien’s *The Lord of the Rings*.[3] Likewise for verbs like “speak” or “dance”, adjectives like “artistic” or “exuberant”, and adverbs like “quickly” or “carefully”.[4]

5.2 Merging multiple views to make one

Here is another example. If, when we are looking at something, we close our eyes for a moment and open them again, what do we see? Normally, it is the same as what we saw before. But recognising that the before and after

[3] J. R. R. Tolkien, *The Lord of the Rings*, London: HarperCollins, 2005, Kindle edition. For a description of Entish, see, for example, page 480. See also, pages 465, 468, 473, 477, 478, 486, and 565.

[4] Although natural language provides a very effective means of compressing information about the world, it is not free of redundancy. And that redundancy has a useful role to play in, for example, enabling us to understand speech in noisy conditions, and in learning the structure of language (Section C.2 and [36, Section 5.2]).

views are the same, means unifying the two patterns to make one and thus compressing the information, as shown schematically in Figure 2.

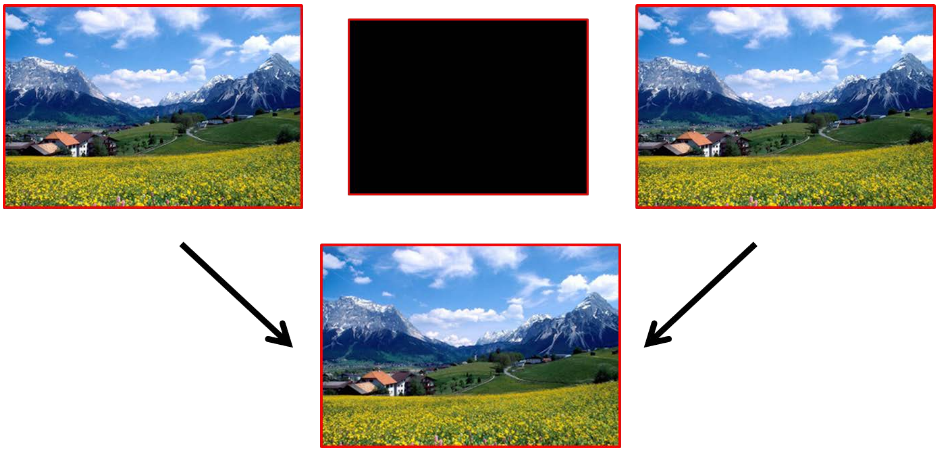

Figure 2: A schematic view of how, if we close our eyes for a moment and open them again, we normally merge the before and after views to make one. The landscape here and in Figure 3 is from Wallpapers Buzz (www.wallpapersbuzz.com), reproduced with permission.

It seems so simple and obvious that if we are looking at a landscape like the one in the figure, there is just one landscape even though we may look at it two, three, or more times. But if we did not unify successive views we would be like an old-style cine camera that simply records a sequence of frames, without any kind of analysis or understanding that, very often, successive frames are identical or nearly so.

5.3 Recognition

Of course, we can recognise something that we have seen before even if the interval between one view and the next is hours, months, or years. In cases like that, it is more obvious that we are relying on memory, as shown schematically in Figure 3. Notwithstanding the undoubted complexities and subtleties in how we recognise things, the process may be seen in broad terms as one of matching incoming information with stored knowledge, merging or unifying patterns that are the same, and thus compressing the information. If we did not compress information in that way, our brains would quickly become cluttered with millions of copies of things that we see around us—people, furniture, cups, trees, and so on—and likewise for sounds and other sensory inputs.

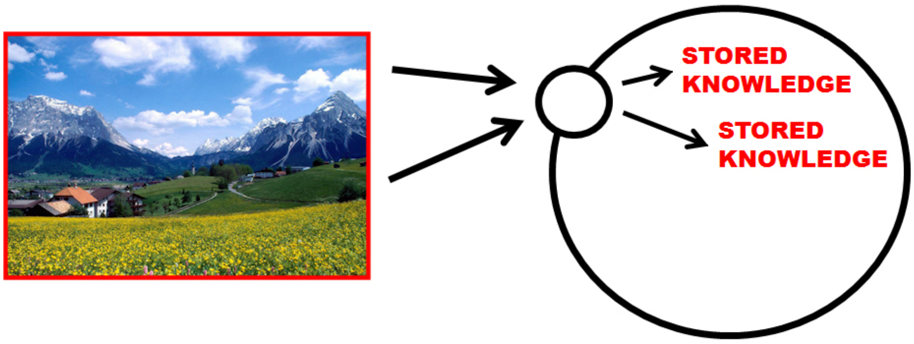

Figure 3: Schematic representation of how, in recognition, incoming visual information may be matched and unified with stored knowledge.

As mentioned earlier, Satosi Watanabe has explored the relationship between pattern recognition and information compression [30, 31].

6 Binocular vision

Information compression may also be seen at work in binocular vision:

> “In an animal in which the visual fields of the two eyes overlap extensively, as in the cat, monkey, and man, one obvious type of redundancy in the messages reaching the brain is the very nearly exact reduplication of one eye’s message by the other eye.” [3, p. 213].

In viewing a scene with two eyes, we normally see one view and not two. This suggests that there is a matching and unification of patterns, with a corresponding compression of information. A sceptic might say, somewhat implausibly, that the one view that we see comes from only one eye. But that sceptical view is undermined by the fact that, normally, the one view shows depth with a vividness that comes from merging the two slightly different views from both eyes.

Strong evidence that, in stereoscopic vision, we do indeed merge the views from both eyes, comes from a demonstration with ‘random-dot stereograms’, as described in [37, Section 5.1].

In brief, each of the two images shown in Figure 4 is a random array of black and white pixels, with no discernable structure, but they are related to each other as shown in Figure 5: both images are the same except that a square area near the middle of the left image is further to the left in the right image.

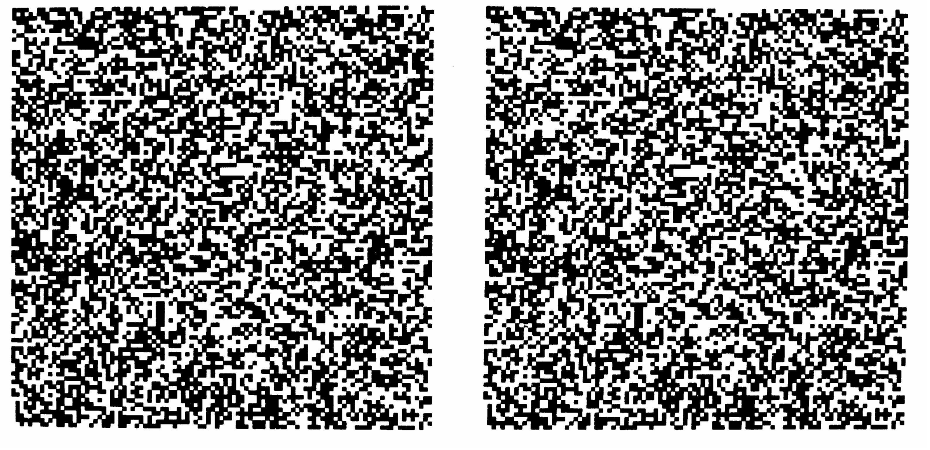

Figure 4: A random-dot stereogram from [12, Figure 2.4-1], reproduced with permission of Alcatel-Lucent/Bell Labs.

1	0	1	0	1	0	0	1	0	1
1	0	0	1	0	1	0	1	0	0
0	0	1	1	0	1	1	0	1	0
0	1	0	Y	*A*	*A*	*B*	*B*	0	0
1	1	1	X	*B*	*A*	*B*	*A*	0	1
0	0	1	X	*A*	*A*	*B*	*A*	1	0
1	1	1	Y	*B*	*B*	*A*	*B*	0	1
1	0	0	1	1	0	1	1	0	1
1	1	0	0	1	1	0	1	1	1
0	1	0	0	0	1	1	1	1	0

1	0	1	0	1	0	0	1	0	1
1	0	0	1	0	1	0	1	0	0
0	0	1	1	0	1	1	0	1	0
0	1	0	*A*	*A*	*B*	*B*	X	0	0
1	1	1	*B*	*A*	*B*	*A*	Y	0	1
0	0	1	*A*	*A*	*B*	*A*	Y	1	0
1	1	1	*B*	*B*	*A*	*B*	X	0	1
1	0	0	1	1	0	1	1	0	1
1	1	0	0	1	1	0	1	1	1
0	1	0	0	0	1	1	1	1	0

Figure 5: Diagram to show the relationship between the left and right images in Figure 4. Reproduced from [12, Figure 2.4-3], with permission of Alcatel-Lucent/Bell Labs.

When the images in Figure 4 are viewed with a stereoscope, projecting the left image to the left eye and the right image to the right eye, the central square appears gradually as a discrete object suspended above the background.

Although this illustrates depth perception in stereoscopic vision—a subject of some interest in its own right—the main interest here is on how we see the central square as a discrete object. There is no such object in either of the two images individually. It exists purely in the *relationship* between the two images, and seeing it means matching one image with the other and unifying the parts which are the same.

This example shows that, although the matching and unification of patterns is a usefully simple idea, there are interesting subtleties and complexities that arise when two patterns are similar but not identical.

Seeing the central object means finding a 'good' match between relevant pixels in the central area of the left and right images, and likewise for the background. Here, a good match is one that yields a relatively high level of information compression. Since there is normally an astronomically large number of alternative ways in which combinations of pixels in one image may be aligned with combinations of pixels in the other image, it is not normally feasible to search through all the possibilities exhaustively.

As with many such problems in artificial intelligence, the best is the enemy of the good. Instead of looking for the perfect solution, we can do better by looking for solutions that are good enough for practical purposes. With this kind of problem, acceptably good solutions can often be found in a reasonable time with heuristic search: doing the search in stages and, at each stage, concentrating the search in the most promising areas and cutting out the rest, perhaps with backtracking or something equivalent to improve the robustness of the search. One such method for the analysis of random-dot stereograms has been described by Marr and Poggio [15].

7 Abstracting object concepts via motion

It seems likely that the kinds of processes that enable us to see a hidden object in a random-dot stereogram also apply to how we see discrete objects in the world. The contrast between the relatively stable configuration of features in an object such as a car, compared with the variety of its surroundings as it travels around, seems to be an important part of what leads us to conceptualise the object as an object [37, Section 5.2]. Any creature that depends on camouflage for protection—by blending with its background—must normally stay still. As soon as it moves relative to its surroundings, it is likely to stand out as a discrete object.

The idea that information compression may provide a means of discovering 'natural' structures in the world—such as the many objects in our visual world—has been dubbed the 'DONSVIC' principle: *the discovery of natural structures via information compression* [36, Section 5.2]. Of course, the word 'natural' is not precise, but it has enough precision to be a meaningful name for the kinds of concepts which are the bread-and-butter of our everyday thinking.

Similar principles may account for how young children come to understand that their first language (or languages) is composed of words (Section 10).

8 Adaptation in the eye of *Limulus* and run-length coding

Information compression may also be seen down in the works of vision. Figure 6 shows a recording from a single sensory cell (*ommatidium*) in the eye of a horseshoe crab (*Limulus polyphemus*), first when the background illumination is low, then when a light is switched on and kept on for a while, and later switched off—shown by the step function at the bottom of the figure.

As one might expect, the ommatidium fires at a relatively low rate of about 20 impulses per second even when the illumination is relatively low (shown at the left of the figure). When the light is switched on, the rate of firing increases sharply but instead of staying high while the light is on (as one might expect), it drops back almost immediately to the background rate. The rate of firing remains at that level until the light is switched off, at which point it drops sharply and then returns to the background level, a mirror image of what happened when the light was switched on.

For the main theme of this paper, a point of interest is that the positive spike when the light is switched on, and the negative spike when the light is switched off, have the effect of marking boundaries, first between dark and light, and later between light and dark. In effect, this is a form of run-length coding (Section 2.3). At the first boundary, the positive spike marks the fact of the light coming on. As long as the light stays on, there is no need for that information to be constantly repeated, so there is no need for the rate of firing to remain at a high level. Likewise, when the light is switched off, the negative spike marks the transition to darkness and, as before, there is no need for constant repetition of information about the new low level of

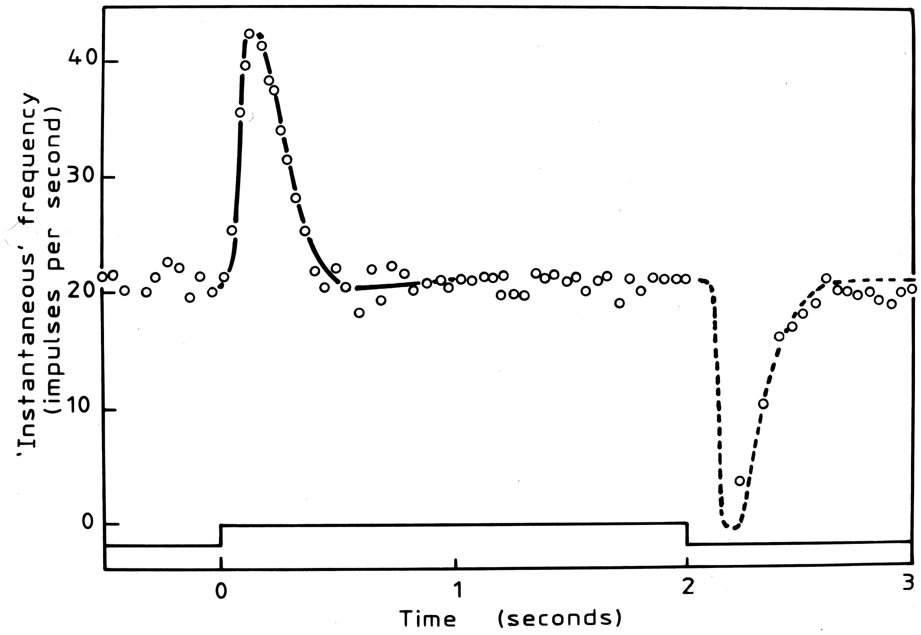

Figure 6: Variation in the rate of firing of a single ommatidium of the eye of a horseshoe crab in response to changing levels of illumination. Reproduced from [18, p. 118.], with permission from the Optical Society of America.

illumination.[5]

Another point of interest is that this pattern of responding—adaptation to constant stimulation—can be explained via the action of inhibitory nerve fibres that bring the rate of firing back to the background rate when there is little or no variation in the sensory input [27]. Inhibitory mechanisms are widespread in the brain [26, p. 45] and it appears that, in general, their role is to reduce or eliminate redundancy in information, in keeping with the main theme of this paper.

9 Other examples of adaptation

Adaptation is also evident at the level of conscious awareness. If, for example, a fan starts working nearby, we may notice the hum at first but then adapt to the sound and cease to be aware of it. But when the fan stops, we are likely to notice the new quietness at first but adapt again and stop noticing it.

Another example is the contrast between how we become aware if something or someone touches us but we are mostly unaware of how our clothes touch us in many places all day long. We are sensitive to something new and different and we are relatively insensitive to things that are repeated.

As with adaptation in the eye of *Limulus*, these other kinds of adaptation may be seen as examples of the run-length coding technique for compression of information.

10 Discovering the segmental structure of language

There is evidence that much of the segmental structure of language—words and phrases—may be discovered via information compression, as described in the following two subsections.

[5] It is recognised that this kind of adaptation in eyes is a likely reason for small eye movements when we are looking at something, including sudden small shifts in position ('microsaccades'), drift in the direction of gaze, and tremor [16]. Without those movements, there would be an unvarying image on the retina so that, via adaptation, what we are looking at would soon disappear.

10.1 The word structure of natural language

As can be seen in Figure 7, people normally speak in ‘ribbons’ of sound, without gaps between words or other consistent markers of the boundaries between words. In the figure—the waveform for a recording of the spoken phrase “on our website”—it is not obvious where the word “on” ends and the word “our” begins, and likewise for the words “our” and “website”. Just to confuse matters, there are three places within the word “website” that look as if they might be word boundaries.

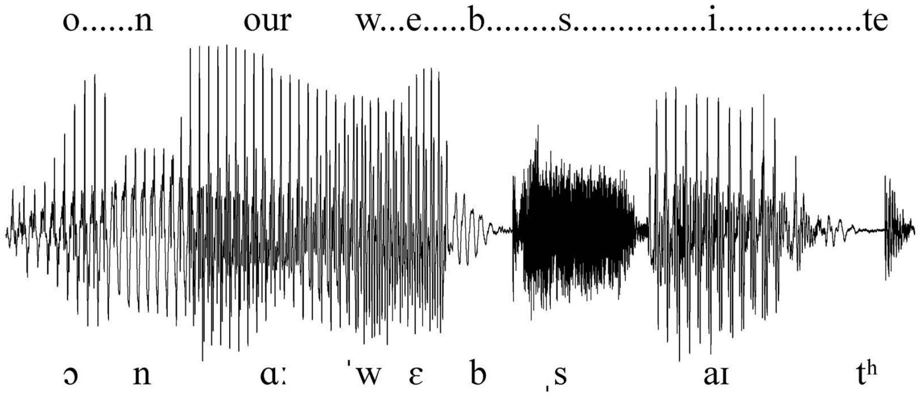

Figure 7: Waveform for the spoken phrase “On our website” with an alphabetic transcription above the waveform and a phonetic transcription below it. With thanks to Sidney Wood of SWPhonetics (swphonetics.com) for the figure and for permission to reproduce it.

Given that words are not clearly marked in the speech that young children hear, how do they get to know that language is composed of words? Learning to read could provide an answer but it appears that young children develop an understanding that language is composed of words well before the age when, normally, they are introduced to reading. Perhaps more to the point is that there are still, regrettably, many children throughout the world that are never introduced to reading but, in learning to talk and to understand speech, they inevitably develop a knowledge of the structure of language, including words.[6]

[6] It has been recognised for some time that skilled speakers of any language have an ability to create or recognise sentences that are grammatical but new to the world. Chomsky’s well-known example of such a sentence is *Colorless green ideas sleep furiously.* [8, p. 15], which, when it was first published, was undoubtedly novel. This ability to create or recognise grammatical but novel sentences implies that knowledge of a language means knowledge of words as discrete entities that can form novel combinations.

In keeping with the main theme of this paper, information compression appears to provide an answer. With computer model MK10 (described in [32] and earlier publications referenced therein), it has been shown that heuristic search for high levels of information compression can reveal much of the word structure in an English-language text from which all spaces and punctuation has been removed [36, Section 5.2]. This discovery of word structure is achieved without the aid of any kind of dictionary or other information about the structure of English. It is also achieved in "unsupervised" mode, without the assistance of any kind of "teacher", or data that is marked as "wrong", or the grading of samples from simple to complex (*cf.* [11]).

It true that there are added complications with speech but it seems likely that similar principles apply.

Discovering the word structure of language via information compression is another example of the DONSVIC principle, mentioned in Section 7—because words are the kinds of 'natural' structure which are the subject of the DONSVIC principle, and because information compression provides a key to how they may be discovered.

10.2 The phrase structure of natural language

Program MK10, mentioned in Section 10.1, does quite a good job at discovering the phrase structure of unsegmented text in which each word has been replaced by a symbol representing the grammatical class of the word [32].[7] As before, it works without any prior knowledge of the structure of English and it works in usupervised mode without the assistance of any kind of "teacher", or anything equivalent.

This result suggests that information compression may have a role to play, not merely in discovering the word structure of language, but more generally in discovering the grammatical structure of language (next).

11 Grammatical inference

Picking up the last point from the previous subsection, it seems likely that learning the grammar of a language may also be understood in terms of information compression. Evidence in support of that expectation comes from research with two programs designed for grammatical inference:

- Program SNPR, which was developed from program MK10, can discover plausible grammars from samples of English-like artificial lan-

[7] This was done by Dr. Isabel Forbes, a person qualified in theoretical linguistics.

guages [32]. This includes the discovery of segmental structures, classes of structure, and abstract patterns. Information compression is central in how the program works.

- Program SP71, one of the main products of the SP programme of research, achieves results at a similar level to that of SNPR. As before, information compression is central in how the program works. With the solution of some residual problems, outlined in [36, Section 3.3], there seems to be a real possibility that the SP system will be able to discover plausible grammars from samples of natural language. Also, it is anticipated that, with further development, the program may be applied to the learning of non-syntactic "semantic" knowledge, and the learning of grammars in which syntax and semantics are integrated.

What was the point of developing SP71 when it does no better at grammatical inference than program SNPR? The reason is that the SNPR program, which was designed to build structures hierarchically, was not compatible with the new goal of the SP programme of research: to simplify and integrate observations and concepts across a broad canvass. What was needed was a new organising principle that would accommodate hierarchical structures and several other kinds of structure as well. It turned out that the SP-multiple-alignment concept, borrowed and adapted from bioinformatics, was much more promising than the organising principle in the SNPR program.

12 Generalisation, the correction of over- and under-generalisations, and "dirty data"

In connection with the learning of language and other kinds of knowledge, it appears that information compression provides an elegant solution to two problems:

- *Generalisation.* How can we generalise our knowledge without over-generalising or under-generalising;

- *Dirty data.* How can we learn correct knowledge despite errors in the examples we hear;

There is evidence that both these things can be achieved with unsupervised learning via information compression, without the correction of errors by parents or teachers or anything equivalent. In brief, a grammar that is good in terms of information compression is one that generalises without over-generalising or under-generalising; and such a grammar is also one that weeds out errors in the data. These things are described more fully in [34, Section 9.5.3] and [36, Section 5.3]. There is also relevant discussion in [40, Section V-H and XI-C].

This problem of generalising our learning without over- or under-generalisation applies to the learning of a natural language and also to the learning of such things as visual images. It appears that the solution outlined here has distinct advantages compared with, for example, what appear to be largely *ad hoc* solutions that have been proposed for deep learning in artificial neural networks [40, Section V-H].

13 Perceptual constancies

It has long been recognised that our perceptions are governed by *constancies*:

- *Size constancy.* To a large extent, we judge the size of an object to be constant despite wide variations in the size of its image on the retina [10, pp. 40-41].

- *Lightness constancy.* We judge the lightness of an object to be constant despite wide variations in the intensity of its illumination [10, p. 376].

- *Colour constancy.* We judge the colour of an object to be constant despite wide variations in the colour of its illumination [10, p. 402].

These kinds of constancy, and others such as shape constancy and location constancy, may each be seen as a means of encoding information economically. It is simpler to remember that a particular person is "about my height" than many different judgements of size, depending on how far away that person is. In a similar way, it is simpler to remember that a particular object is "black" or "red" than all the complexity of how its lightness or its colour changes in different lighting conditions.

14 Mathematics

A discussion of mathematics may seem out of place in a paper about BICMUP but maths and computing are both products of the human intellect so, for that reason, a consideration of their organisation and workings is relevant to the matter in hand.

In [41] it has been argued that mathematics may be seen as a set of techniques for the compression of information, and their application. In case this seems implausible:

- An equation like Albert Einstein's $E = mc^2$ may be seen as a very compressed representation of what may be a very large set of data points relating energy (E) and mass (m), with the speed of light (c) as a constant. Similar things may be said about such well-known equations as $s = (gt^2)/2$ (Newton's second law of motion), $a^2 + b^2 = c^2$ (Pythagoras's equation), $PV = k$ (Boyle's law), and $F = q(E + v \times B)$ (the charged-particle equation).

- The first three of the basic techniques for information compression outlined in Section 2.3 may be seen at work in mathematical notations. For example: multiplication as repeated edition may be seen as an example of run-length coding;

Owing to the close connections between logic and mathematics, and between computing and mathematics, it seems likely that similar principles apply in logic and in computing.

15 The SP system and its empirical support as evidence for BICMUP

Empirical support for the SP system, viewed as a theory of human intelligence, is largely independent of the kind of direct empirical support for BICMUP described in the body of this paper. Thus empirical support for the SP theory provides additional empirical support for BICMUP.

As outlined in Appendix A, the SP system has strengths in three main aspects of human intelligence: versatility in the representation of diverse kinds of knowledge, versatility in diverse aspects of intelligence, and the kind of seamless integration of diverse kinds of knowledge and diverse aspects of intelligence, in any combination, which appears to be necessary to model the fluidity, versatility, and adaptability of human intelligence.

This versatility in aspects of human intelligence provides empirical support for the SP system as a theory of human intelligence, including its central organising principles: SP-multiple-alignment and information compression. Thus the SP system and its empirical support provides indirect but powerful empirical support for BICMUP.

16 Some apparent contradictions and how they may be resolved

The idea that information compression is fundamental in human learning, perception and cognition (BICMUP), and also in AI, mainstream computing, and mathematics (aspects of the SP theory besides human cognition), seems to be contradicted by:

- The productivity of the human brain and the ways in which computers and mathematics may be used to create redundant copies of information as well as to compress information;
- The fact that redundancy in information is often useful in both the storage and processing of information;
- A less direct challenge to BICMUP and the SP theory is persuasive evidence, described by Gary Marcus [14], that in many respects, the human mind is a kluge, meaning "a clumsy or inelegant—yet surprisingly effective—solution to a problem" (*ibid.*, p 2).
- The fact that certain kinds of redundancy are difficult or impossible for people to detect and exploit.

These apparent contradictions and how they may be resolved are discussed in Appendix C.

17 Conclusion

This paper presents evidence for the idea, referred to as "BICMUP", that much of human learning, perception, and thinking, and the workings of nervous systems, may be understood as compression of information.

As background to the main body of the paper: the *SP theory of intelligence* is introduced, with information compression as its unifying theme; it is often useful to view information compression as a process of finding patterns that match each other and merging or "unifying" multiple instances of a pattern to make one, and there are five main variants of that principle; it has been recognised for some time that there is an intimate connection between information compression and concepts of prediction and probability.

Research related to BICMUP was developed by such people as Fred Attneave and Horace Barlow in the 1950s and later. At about the same time and largely independently, related research, with an emphasis on issues in mathematics and computing, was developed by such people as Ray Solomonoff, Gregory Chaitin, and Andrei Kolmogorov.

As with artificial systems, information compression can mean selective advantages for animals: in the efficient storage and transmission of information; in being able to make predictions about sources of food, where there may be dangers, and so on; and in corresponding savings in energy.

Some aspects of information compression and its benefits are so much embedded in our everyday thinking that they are easily overlooked: most nouns, verbs and adjectives may be seen as short codes for relatively complex concepts; we frequently create shorthands for relatively long expressions; if we blink or otherwise close our eyes for a moment, we normally compress the before and after views by merging them into a single percept; In recognising something after a longer period, we are, in effect, compressing the sensory information by merging it with information that we have stored.

Quite independently of interesting issues related to depth perception, the fact that, normally, we see a single image when we are viewing something with two eyes, suggests that we are compressing sensory data by merging the two retinal images to make one. The fact that people see the central figure in random-dot stereograms proves that they are indeed merging the two random-dot images. This is because the central figure is not present in either of the random-dot images individually—it can only be seen by unifying them.

Similar principles apply in the way the boundaries of a visual object can be seen most clearly when the object moves in relation to its background.

Information compression may be seen at a “low” level in the workings of a sensory unit or *ommatidium* in the eye of *Limulus*, the horseshoe crab. Here, there is a medium-level rate of firing when illumination is constant, with a sharp upswing when a light is switched on, and a downswing when the light is switched off. These swings have the effect of marking the boundaries of a period of uniform illumination, in accordance with the principle of run-length coding. Similar principles of adaptation apply at the level of consciousness in one’s responses to sounds or things that may touch one’s skin.

A computer model of the unsupervised learning of segmental structure in language, with a central role for information compression, produces results that suggest that information compression is a key principle in the unsupervised discovery of segmental structures, both at the level of words and of phrases.

Likewise, computer models for learning the syntax of English-like artificial languages, with central roles for information compression, produce results that suggest that information compression is a key principle in the unsupervised learning of syntax, including the learning of segmental structures, classes of structure, and abstract patterns.

It seems likely that these principles will generalise to natural languages.

Information compression may be seen in the perceptual *constancies*, including size constancy, lightness constancy, and colour constancy.

In another paper, it has been argued that mathematics may be seen to be a set of techniques for the compression of information, and their application. Since mathematics is a product of the human intellect, the evidence and conclusions of this paper provides further support for BICMUP.

Since information compression is central in the workings of the SP system, and since it is successful in modelling several aspects of human intelligence, the system provides indirect but quite strong support for BICMUP.

Four possible objections to BICMUP and the SP theory are outlined in Section 16 but it appears that there are good answers to all three of these objections, as described in Appendix C.

A Outline of the SP system

As noted in Section 2.2, the *SP theory of intelligence*, and its realisation in the *SP computer model*, is a unique attempt to simplify and integrate observations and concepts across a broad canvass, with information compression as a unifying theme.

The name "SP" is short for *Simplicity* and *Power*. This is because (lossless) information compression is central in the workings of the SP system, and information compression may be seen to be a process of maximising the *simplicity* of a body of information, **I**, by extracting redundancy from **I**, whilst retaining as much as possible of its non-redundant descriptive *power*.

The SP theory is described most fully in [34] and more briefly in [36]. Details of other publications in this programme of research may found, most with download links, on www.cognitionresearch.org/sp.htm. Some of them are referenced elsewhere in this paper.

The SP theory is conceived as a brain-like system that receives *New* information via its senses and stores some or all of it, in compressed form, as *Old* information.

In the SP system, all kinds of knowledge are stored as arrays of atomic symbols called *patterns*. At present, the SP computer model works only with one-dimensional patterns but it is envisaged that it will be generalised to work with two-dimensional patterns, in addition to 1D patterns.

A.1 SP-multiple-alignment

A key idea in the SP system is the concept of *SP-multiple-alignment* borrowed and adapted from the concept of multiple alignment in bioinformatics. An example of a multiple alignment from bioinformatics is shown in Figure 8. Here, five DNA sequences have been arranged in rows and, by judicious "stretching" of sequences in a computer, matching symbols have brought into line. A "good" multiple alignment is one with a relatively large number of matching symbols.

```
  G G A     G     C A G G G A G G A     T G     G   G G A
  | | |     |     | | | | | | | | |     | |     |   | | |
  G G | G   G C C C A G G G A G G A     | G G C G   G G A
  | | |     | | | | | | | | | | | |     | |     |   | | |
A | G A C T G C C C A G G G | G G | G C T G     G A | G A
  | | |           | | | | | | | | |   |   |     |   | | |
  G G A A         | A G G G A G G A   | A G     G   G G A
  | |   |         | | | | | | | |     |   |     |   | | |
  G G C A         C A G G G A G G     C   G     G   G G A
```

Figure 8: A 'good' multiple alignment amongst five DNA sequences. Reproduced with permission from Figure 3.1 in [34].

It turns out that, in most cases, there is an astronomically large number of possible multiple alignments. This means that, in creating good multiple alignments, heuristic techniques must be used, building each multiple alignment in stages and, at each stage, selecting only the best multiple alignments for further processing.

In the SP system, the SP-multiple-alignment concept has been adapted so that one or more of the patterns (normally only one) is a New pattern and the rest are Old patterns, and the system is designed to create SP-multiple-alignments that enable the New pattern (or patterns) to be encoded economically in terms of the Old patterns.

At the heart of the concept of SP-multiple-alignment is the idea that we may identify repetition or *redundancy* in information by searching for patterns that match each other, and that we may reduce that redundancy and thus compress information by merging or *unifying* two or more matching patterns to make one. This idea—*information compression via the matching and unification of patterns*—may be referred to in brief as "ICMUP". Variants of ICMUP are described in Section 2.3.

As with the creation of multiple alignments in bioinformatics, heuristic techniques are needed for the creation of SP-multiple-alignment. With this kind of technique, it is normally not possible guarantee that the best possible multiple alignment has been found. We must be content with the creation of SP-multiple-alignments that are “reasonably good”.

A.2 Strengths of the SP system

It turns out that the SP system, with SP-multiple-alignment at centre-stage, is a versatile means of representing diverse kinds of knowledge—such as the syntax of natural languages, class-inclusion hierarchies, part-whole hierarchies, and more. And the SP system is a versatile means of modelling diverse aspects of intelligence—such as unsupervised learning, the processing natural language, pattern recognition, several kinds of reasoning, and more. With tasks like pattern recognition, it has human-like abilities to recognise patterns despite errors of omission, commission, and substitution.

Because these things all flow from one relatively simple framework—SP-multiple-alignment—there is potential for the seamless integration of diverse kinds of knowledge and diverse aspects of intelligence, in any combination. That kind of seamless integration appears to be *essential* for modelling the fluidity, versatility, and adaptability of human intelligence. The SP-multiple-alignment concept has potential to be as significant for understanding human-level intelligence as is DNA for biological sciences. SP-multiple-alignment could be the “double helix” of intelligence.

A.3 SP-neural

The abstract concepts in the SP system may be mapped into groupings of neurons and their interconnections in a version of the SP theory called *SP-neural* [39].

A.4 Potential benefits and applications of the SP system

Potential benefits and applications of the SP system include: helping to solve nine problems with big data; helping to develop human-like intelligence in autonomous robots; understanding natural vision and the development of computer vision; medical diagnosis; and the SP system as an intelligent database. Papers about these areas of application may be found on www.cognitionresearch.org/sp.htm. Other potential benefits and applications are described in [38].

B Barlow's change of view about the significance of information compression, with comments

As noted in Section 3, Horace Barlow, in a paper published in 2001 [5], argued that "... the [compression] idea was right in drawing attention to the importance of redundancy in sensory messages ... but it was wrong in emphasizing the main technical use for redundancy, which is compressive coding." (*ibid.* p. 242).

As mentioned before, it seems to me that, while there are some valid points in what Barlow says in support of his new position, his overall conclusions are wrong. His main arguments follow, with my comments after each one, flagged with "JGW".

1. "It is important to realize that redundancy is not something useless that can be stripped off and ignored. An animal must identify what is redundant in its sensory messages, for this can tell it about structure and statistical regularity in its environment that are important for its survival." [5, p. 243], and "It is ... knowledge and recognition of ... redundancy, not its reduction, that matters." [5, p. 244].

   JGW: It seems to me that the error here is to assume that compression of information means the elimination of redundant patterns. On the contrary, lossless compression of something like '`tabletabletabletabletable`' means retaining one instance of '`table`' with a record of the number of occurrence, or something equivalent. Knowledge of the frequency of occurrence of any pattern may serve in the calculation of probabilities, something that has been worked out in detail in the SP system ([34, Section 3.7], [36, Section 4.4]).

   In general, compression of information is entirely compatible with a knowledge of redundant patterns and what they can say about statistical regularity in a creature's environment that is important for its survival.

2. "Redundancy is mainly useful for error avoidance and correction" [5, p. 244]. This heading in [5] appears to be a relatively strong point in support of Barlow's new position, but he writes: "Since it is certainly true that sensory transducers and neural communication channels introduce noise, this is likely to be important in the brain, but the correction of such internally generated errors is a separate problem, and it will not be considered further here." [5, p. 244].

JGW: Redundancy can certainly be useful in the avoidance or correction of errors. But that does not invalidate BICMUP. As noted in Appendix C.2, the SP system, which is dedicated to the compression of information, will not work properly in such tasks as parsing, pattern recognition and grammatical inference, unless there are redundancies in its raw data. For that reason, it needs those redundancies in order to correct errors of omission, commission, and substitution, as described in [34, Section 6.2], [35, Section 2.2.2], and [36, Section 6.2].

3. Following the remark that "This is the point on which my own opinion has changed most, partly in response to criticism, partly in response to new facts that have emerged." [5, p. 244], Barlow writes:

   > "Originally both Attneave and I strongly emphasized the economy that could be achieved by recoding sensory messages to take advantage of their redundancy, but two points have become clear since those early days. First, anatomical evidence shows that there are very many more neurons at higher levels in the brain, suggesting that redundancy does not decrease, but actually increases. Second, the obvious forms of compressed, non-redundant, representation would not be at all suitable for the kinds of task that brains have to perform with the information represented; ..." [5, pp. 244–245].

   and

   > "I think one has to recognize that the information capacity of the higher representations is likely to be greater than that of the representation in the retina or optic nerve. If this is so, redundancy must increase, not decrease, because information cannot be created." [5, p. 245].

   JGW: It seems to me that there are two errors here:

   - The likelihood that there are "very many more neurons at higher levels in the brain [than at the sensory levels]" and that "the information capacity of the higher representations is likely to be greater than that of the representation in the retina or optic nerve" does not in any way invalidate BICMUP.

     It seems likely that many of the neurons at higher levels are concerned with the storage of one's accumulated knowledge over the period from one's birth to one's current age ([34, Chapter 11], [39, Section 4]). By contrast, neurons at the sensory level would be concerned only with the processing of sensory information at any one time.

Although knowledge in one's long-term memory stores is likely to be highly compressed and only a partial record of one's experiences, it is likely, for most of one's life except early childhood, to be very much larger than the sensory information one is processing at any one time. Hence, it should be no surprise to find many more neurons at higher levels than at the sensory level.

- For reasons given under point 4, there seem to be errors in the proposition that "the obvious forms of compressed, non-redundant, representation would not be at all suitable for the kinds of task that brains have to perform with the information represented."

4. Under the heading "Compressed representations are unsuitable for the brain", Barlow writes:

> "The typical result of a redundancy-reducing code would be to produce a distributed representation of the sensory input with a high activity ratio, in which many neurons are active simultaneously, and with high and nearly equal frequencies. It can be shown that, for one of the operations that is most essential in order to perform brain-like tasks, such high activity-ratio distributed representations are not only inconvenient, but also grossly inefficient from a statistical viewpoint ..." [5, p. 245].

JGW: It is not clear why Barlow should assume that we store knowledge in a distributed representation or why such a representation should necessarily be inefficient:

- With regard to the second point, it is true that deep learning in artificial neural networks [22], with their distributed representations, are often hungry for computing resources. But otherwise they are quite successful with certain kinds of task, and there appears to be scope for increasing their efficiencies [7].
- But the SP system demonstrates that localist representations with such things as class-inclusion hierarchies and part-whole hierarchies are efficient and effective in a variety of kinds of task (see, for example, [34] and [36]), and they are biologically plausible [39].

C Some apparent contradictions of BICMUP and the SP theory, and how they may be resolved

The three subsections here discuss the apparent contradictions to BICMUP and the SP theory mentioned in Section 16, and how they may be resolved.

C.1 The creation of redundancy via information compression: "decompression by compression"

The idea that information may be decompressed by compressing information—"decompression by compression"—seems paradoxical at first sight. Examples described here may help to show why the paradox is more apparent than real.

C.1.1 A simple example of "decompression by compression"

The chunking-with-codes idea mentioned in Section 2.3 provides a simple example of decompression by compression. If, for example, a document contains many instances of the expression "Treaty on the Functioning of the European Union" we may shorten it by giving that expression a relatively short name or code like "TFEU" and then replacing most instances of the long expression with its short code.

This achieves compression of information because, in effect, multiple instances of "Treaty on the Functioning of the European Union" have been matched with each other and unified.

We can reverse the process and thus decompress the document by searching for instances of "TFEU" and replacing each one with "Treaty on the Functioning of the European Union". But to achieve that result, the search pattern, "TFEU", needs to be matched and unified with each instance of "TFEU" in the document. And that process of matching and unification is itself, in effect, a process of compressing information. Hence, decompression of information has been achieved via information compression.

C.1.2 Creating redundancy via information compression

With a computer, it is very easy to create information containing large amounts of redundancy and to do it by a process which may itself be seen to entail the compression of information.

We can, for example, make a 'call' to the function defined in Figure 9, using the pattern '`oranges_and_lemons(100)`'. The effect of that call is to print out a highly redundant sequence containing 100 copies of the expression "Oranges and lemons, Say the bells of St. Clement's; ".

```
void oranges_and_lemons(int x)
{
     printf("Oranges and lemons, Say the bells of St. Clement's; ");
     if (x > 1) oranges_and_lemons(x - 1) ;
}.
```

Figure 9: A simple recursive function showing how, via computing, it is possible to create repeated (redundant) copies of "Oranges and lemons, Say the bells of St. Clement's; ".

Taking things step by step, this works as follows:

1. The pattern '`oranges_and_lemons(100)`' is matched with the pattern '`void oranges_and_lemons(int x)`' in the first line of the function.

2. The two instances of '`oranges_and_lemons`' are unified and the value 100 is assigned to the variable x. The assignment may also be understood in terms of the matching and unification of patterns but the details would be a distraction from the main point here.

3. The instruction '`printf("Oranges and lemons, Say the bells of St. Clement's; ");`' in the function has the effect of printing out '`Oranges and lemons, Say the bells of St. Clement's; `'.

4. Then if $x > 1$, the instruction '`oranges_and_lemons(x - 1)`' has the effect of calling the function again but this time with 99 as the value of x (because of the instruction $x - 1$ in the pattern '`oranges_and_lemons(x - 1)`', meaning that 1 is to be subtracted from the current value of x).

5. Much as with the first call to the function, the pattern '`oranges_and_lemons(99)`' is matched with the pattern '`void oranges_and_lemons(int x)`' in the first line of the function.

6. Much as before, the two instances of '`oranges_and_lemons`' are unified and the value 99 is assigned to the variable x.

7. This cycle continues until the value of x is 0.

Where does compression of information come in? It happens mainly when one copy of '`oranges_and_lemons`' is matched and unified with another copy so that, in effect, two copies are reduced to one.

There is more about recursion at the end of Appendix C.1.3, next.

C.1.3 Decompression by compression in the SP system

How the SP system may achieve decompression by compression is described in [34, Section 3.8] and [36, Section 4.5].

There are two important points here:

- Decompression of a body of information **I**, may be achieved by a process which is *exactly* the same as the process that achieved the original compression of **I**: there is no modification to the program of any kind.

  All that is needed to achieve decompression is to ensure that there is some residual redundancy in the compressed version of **I**, so that the program has something to work on, as noted in Appendix B.

- The SP computer model is entirely devoted to compression of information, without any special provision for decompression of information.

Those two things establish that it is indeed possible to achieve decompression by compression, meaning that, in that idea, there is really no paradox or contradiction.

With regard to the example with recursion discussed in Appendix C.1.2, readers may find it useful to examine examples of recursion with the SP system, described in [34, Sections 4.3.2.1 and 5.3], [37, Section 3.3], and [39, Section 7]. In all these examples, recursion is driven by a process which is unambiguously devoted to the compression of information.

C.2 Redundancy is often useful in the storage and processing of information

The fact that redundancy—repetition of information—is often useful in both the storage and processing of information is the second apparent contradiction to BICMUP and the SP theory. Here are some examples:

- With any kind of database, it is normal practice to maintain one or more backup copies as a safeguard against catastrophic loss of the data. Each backup copy represents redundancy in the system.

- With information on the internet, it is common practice to maintain two or more ‘mirror’ copies in different places to minimise transmission times and to spread processing loads across two or more sites, thus reducing the chance of overload at any one site. Again, each mirror copy represents redundancy in the system.

- Redundancies in natural language can be a very useful aid to the comprehension of speech in noisy conditions.

- It is normal practice to add redundancies to electronic messages, in the form of additional bits of information together with checksums, and also by repeating the transmission of any part of a message that has become corrupted. These things help to safeguard messages against accidental errors caused by such things as birds flying across transmission beams, or electronic noise in the system, and so on.

These kinds of uses of redundancy may seem to conflict with the idea that information compression—which means reducing redundancy—is fundamental in computing and cognition. However, the two things are largely independent. For example: “... it is entirely possible for a database to be designed to minimise internal redundancies and, at the same time, for redundancies to be used in backup copies or mirror copies of the database ... Paradoxical as it may sound, knowledge can be compressed and redundant at the same time.” [34, Section 2.3.7].

Also, the SP system, which is dedicated to the compression of information, will not work properly with totally random information containing no redundancy. It needs redundancy in its raw data in order to achieve such things as the parsing of natural language, pattern recognition, and grammatical inference, and, in those and other areas, it needs redundancy in its data for the correction of errors of omission, commission, and substitution.

C.3 The human mind as a kluge

As mentioned in Section 16, Gary Marcus has described persuasive evidence that, in many respects, the human mind is a kluge. To illustrate the point, here is a sample of what Marcus says:

> “Our memory is both spectacular and a constant source of disappointment: we recognize photos from our high school year-books decades later—yet find it impossible to remember what we had for breakfast yesterday. Our memory is also prone to distortion, conflation, and simple failure. We can know a word but not be

> able to remember it when we need it ... or we can learn something valuable ... and promptly forget it. The average high school student spends four years memorising dates, names, and places, drill after drill, and yet a significant number of teenagers can't even identify the *century* in which World War I took place." [14, p. 18], emphasis as in the original.

Clearly, human memory is, in some respects, much less effective than a computer disk drive or even a book. And it seems likely that at least part of the reason for this and other shortcomings of the human mind is that "Evolution [by natural selection] tends to work with what is already in place, making modifications rather than starting from scratch." and "piling new systems on top of old ones" [14, p. 12].

Superficially, this and other evidence in [14] seems to undermine the idea that there is some grand unifying principle—such as information compression via SP-multiple-alignment—that governs the organisation and workings of the human mind.

Perhaps, as Marvin Minsky suggested, "each [human] mind is made of many smaller processes" called *agents* each one of which "can only do some simple thing that needs no mind or thought at all. Yet when we join these agents in societies—in certain very special ways—this leads to true intelligence." [17, p. 17].

In answer to these points:

- The evidence that Marcus presents is persuasive: it is difficult to deny that, in certain respects, the human mind is a kluge. And evolution by natural selection provides a plausible explanation for anomalies and inconsistencies in the workings of the human mind.

- But those conclusions are entirely compatible with BICMUP (supported by evidence presented in this paper), and the SP theory as a theory of mind (supported by evidence presented in [34, 36, 39] and elsewhere). As Marcus says:

  > "I don't mean to chuck the baby along with its bath—or even to suggest that kluges outnumber more beneficial adaptations. The biologist Leslie Orgel once wrote that 'Mother Nature is smarter than you are,' and most of the time it is." [14, p. 16].

  although Marcus warns that in comparisons between artificial systems and natural ones, nature does not always come out on top.

In general it seems that, despite the evidence for kluges in the human mind, there can be powerful organising principles too. Since BICMUP and the SP theory are well supported by evidence, they are likely to provide useful insights into the nature of human intelligence, alongside an understanding that there are likely to be kluge-related anomalies and inconsistencies too.

Minsky's counsel of despair—"The power of intelligence stems from our vast diversity, not from any single, perfect principle." [17, p. 308]—is probably too strong. It is likely that there is at least one unifying principle for human-level intelligence, and there may be more. And it is likely that, with people, any such principle or principles operates alongside the somewhat haphazard influences of evolution by natural selection.

C.4 Some kinds of redundancy are difficult or impossible for people to detect and exploit

There is no doubt that people are imperfect in their abilities to detect and exploit redundancy. For example:

> "... a grid in which pixels encoded the binary expansion of π would, of course, have a very simple description, but this structure would not be identified by the perceptual system; the grid would, instead, appear completely unstructured." [6, p. 578].

At first sight, this shortfall in our abilities seems to undermine the idea that information compression is a unifying principle in the workings of brains and nervous systems. But that idea does not imply that brains and nervous systems are perfect compressors of information. Indeed, it appears that with all but the smallest or most regular bodies of information, it is necessary to use heuristic techniques for compression of the information and that these cannot guarantee that the best possible result has been found (Section A.1).

References

[1] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[2] H. B. Barlow. Sensory mechanisms, the reduction of redundancy, and intelligence. In HMSO, editor, *The Mechanisation of Thought Processes*, pages 535–559. Her Majesty's Stationery Office, London, 1959.

[3] H. B. Barlow. Trigger features, adaptation and economy of impulses. In K. N. Leibovic, editor, *Information Processes in the Nervous System*, pages 209–230. Springer, New York, 1969.

[4] H. B. Barlow. Intelligence, guesswork, language. *Nature*, 304:207–209, 1983.

[5] H. B. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12:241–253, 2001.

[6] N. Chater. Reconciling simplicity and likelihood principles in perceptual organisation. *Psychological Review*, 103(3):566–581, 1996.

[7] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman. Project adam: building an efficient and scalable deep learning training system. In *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 571–582. USENIX Association, 2014.

[8] N. Chomsky. *Syntactic Structures*. Mouton, The Hague, 1957.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1991.

[10] J. P. Frisby and J. V. Stone. *Seeing: The Computational Approach to Biological Vision*. The MIT Press, London, England, 2010.

[11] M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[12] B. Julesz. *Foundations of Cyclopean Perception*. Chicago University Press, Chicago, 1971.

[13] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 3rd edition, 2014.

[14] G. Marcus. *Kluge: the Hapharzard Construction of the Human Mind*. Faber and Faber, London, paperback edition, 2008. ISBN: 978-0-571-23652-7.

[15] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B*, 204(1156):301–328, 1979.

[16] S. Martinez-Conde, J. Otero-Millan, and S. L. Macknik. The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nature Reviews Neuroscience*, 14:83–96, 2013.

[17] M. Minsky, editor. *The Society of Mind*. Simon & Schuster, New York, 1986.

[18] F. Ratliff, H. K. Hartline, and W. H. Miller. Spatial and temporal aspects of retinal inhibitory interaction. *Journal of the Optical Society of America*, 53:110–120, 1963.

[19] J. Rissanen. Modelling by the shortest data description. *Automatica*, 14(5):465–471, 1978.

[20] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society B*, 49(3):223–239, 1987.

[21] K. Sayood. *Introduction to data compression*. Morgan Kaufmann, Amsterdam, 2012.

[22] J. Schmidhuber. Deep learning in neural networks: an overview. *Neural Networks*, 61:85–117, 2015.

[23] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[24] R. J. Solomonoff. A formal theory of inductive inference. Parts I and II. *Information and Control*, 7:1–22 and 224–254, 1964.

[25] R. J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997.

[26] L. R. Squire, D. Berg, F. E. Bloom, S. du Lac, A. Ghosh, and N. C. Spitzer, editors. *Fundamental neuroscience*. Elsevier, Amsterdam, fourth edition, 2013.

[27] G. von Békésy. *Sensory Inhibition*. Princeton University Press, Princeton, NJ, 1967.

[28] C. S. Wallace and D. M. Boulton.

[29] S. Watamabe. Information-theoretical aspects of inductive and deductive inference. *IBM Journal of Research and Development*, 4:208–231, 1960.

[30] S. Watanabe, editor. *Frontiers of Pattern Recognition*. Academic Press, New York, 1972.

[31] S. Watanabe. Pattern recognition as information compression. In *Frontiers of Pattern Recognition* [30].

[32] J. G. Wolff. Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I. M. Schlesinger, and M. D. S. Braine, editors, *Categories and Processes in Language Acquisition*, pages 179–215. Lawrence Erlbaum, Hillsdale, NJ, 1988. bit.ly/ZIGjyc.

[33] J. G. Wolff. Computing, cognition and information compression. *AI Communications*, 6(2):107–127, 1993. bit.ly/XL359b.

[34] J. G. Wolff. *Unifying Computing and Cognition: the SP Theory and Its Applications*. CognitionResearch.org, Menai Bridge, 2006. ISBNs: 0-9550726-0-3 (ebook edition), 0-9550726-1-1 (print edition). Distributors, including Amazon.com, are detailed on bit.ly/WmB1rs.

[35] J. G. Wolff. Towards an intelligent database system founded on the SP theory of computing and cognition. *Data & Knowledge Engineering*, 60:596–624, 2007. bit.ly/1CUldR6.

[36] J. G. Wolff. The SP theory of intelligence: an overview. *Information*, 4(3):283–341, 2013. bit.ly/1NOMJ6l.

[37] J. G. Wolff. Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision. *SpringerPlus*, 3(1):552–570, 2014. bit.ly/2oIpZB6.

[38] J. G. Wolff. The SP theory of intelligence: benefits and applications. *Information*, 5(1):1–27, 2014. bit.ly/1FRYwew.

[39] J. G. Wolff. Information compression, multiple alignment, and the representation and processing of knowledge in the brain. *Frontiers in Psychology*, 7:1584, 2016. bit.ly/2esmYyt.

[40] J. G. Wolff. The SP theory of intelligence: its distinctive features and advantages. *IEEE Access*, 4:216–246, 2016. bit.ly/2qgq5QF.

[41] J. G. Wolff. On the "mysterious" effectiveness of mathematics in science. Technical report, CognitionResearch.org, 2017. Submitted for publication. bit.ly/2otrHD0. This report is also archived in vixra.org/ and hal.archives-ouvertes.fr/hal-01534622.