Oiling the wheels under big data

How big gains in efficiency may be achieved in transmitting big data from one place to another

Gerry Wolff

"The Square Kilometre Array is one of the most ambitious scientific projects ever undertaken. Its organizers plan on setting up a massive radio telescope made up of more than half a million antennas spread out across vast swaths of Australia and South Africa." So say John Kelly and Steve Hamm, both of IBM, in their book *Smart Machines* [1, p. 62].

Their reason for writing about the SKA is that it will create huge problems for even the smartest or most powerful of smart machines. "The SKA is the ultimate big data challenge." say Kelly and Hamm. "The telescope will collect a veritable deluge of radio signals from outer space—amounting to fourteen exabytes of digital data per day …" (*ibid*., p. 63). Of the several problems arising from quantities of data like that, one that may seem surprising is that the amount of energy required merely to move the data from one place to another is proving to be a significant headache for the SKA project and other projects of that kind.

This problem may be solved or at least reduced via a new approach to old ideas: "analysis/synthesis" and, more specifically, the relatively challenging idea of "model-based coding".

Analysis/synthesis has been described by Khalid Sayood [3, p. 592] like this:

> "Consider an image transmission system that works like this. At the transmitter, we have a person who examines the image to be transmitted and comes up with a description of the image. At the receiver, we have another person who then proceeds to create that image. For example, suppose the image we wish to transmit is a picture of a field of sunflowers. Instead of trying to send the picture, we simply send the words 'field of sunflowers'. The person at the receiver paints a picture of a field of sunflowers on a piece of paper and gives it to the user. Thus an image of an object is transmitted from the transmitter to the receiver in a highly compressed form."

This approach works best with the transmission of speech, probably because the physical structure and properties of the vocal cords, tongue, teeth, and so on, help in the process of creating an analysis of any given sample of speech and in any synthesis of speech that may be derived from that analysis. But things are more difficult with images, especially if they are moving.

The concept of model-based coding was described by John Pierce in 1961 [2, pp. 139-140] like this:

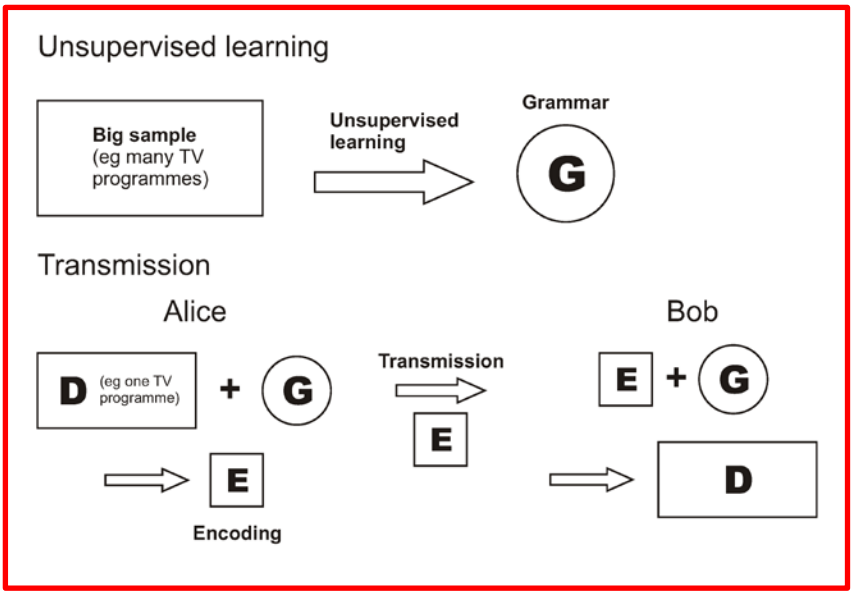
> "Imagine that we had at the receiver a sort of rubbery model of a human face. Or we might have a description of such a model stored in the memory of a huge electronic computer. First, the transmitter would have to look at the face to be transmitted and 'make up' the model at the receiver in shape and tint. The transmitter would also have to note the sources of light and reproduce these in intensity and direction at the receiver. Then, as the person before the transmitter talked, the transmitter would

Contact: Dr Gerry Wolff PhD CEng MBCS MIEEE, CognitionResearch.org, Menai Bridge, UK; *Phone:* +44 (0) 1248 712962; +44 (0) 7746 290775; *Skype:* gerry.wolff; *E-mail:* **jgw@cognitionresearch.org**; *Web:* **www.cognitionresearch.org**.

have to follow the movements of his eyes, lips and jaws, and other muscular movements and transmit these so that the model at the receiver could do likewise."

At the time this was written, it would have been impossibly difficult to make things work as described. Pierce says: "Such a scheme might be very effective, and it could become an important invention if anyone could specify a useful way of carrying out the operations I have described. Alas, how much easier it is to say what one would like to do (whether it be making such an invention, composing Beethoven's tenth symphony, or painting a masterpiece on an assigned subject) than it is to do it." (*ibid.*, p. 140).

Even today, Piece's vision is a major challenge. But there appears to be a way forward, described in the rest of this article. If it can be made to work, it would indeed be very effective, oiling the wheels under big data by providing a means of moving it from one place to another with relatively small expenditures of energy.



In outline, model-based coding may be made to work as shown in the figure. There would be two main elements to the scheme: learning of an abstract description or 'grammar' ('G') of the kind of information to be transmitted, such as TV programmes or indeed information received via the SKA; and transmission of a specific body of data ('D'), such as a single TV programme or a body of data from the SKA, from the transmitter ('Alice') to the receiver ('Bob').

The learning would be "unsupervised", meaning learning directly from data without assistance of any kind of "teacher", or the labelling of examples, or rewards or punishments, or anything equivalent. In this scheme, learning would normally be done independently of any specific transmission, it would be done by a relatively powerful computer and with a relatively large sample of the kind of data that is to be transmitted. Alice and Bob would each receive a copy of G.

In transmission, D would first be processed by Alice in conjunction with G to create an 'encoding' ('E'), which would describe D in terms of the entities and abstract concepts in G. The encoding, E, would then be transmitted to Bob who would use it, in conjunction with his own copy of G, to reconstruct D.

Since E would normally be very small compared with D, there would, with one qualification, normally be a large saving in the amount of information to be transmitted compared with the transmission of raw data, or indeed, transmission of information that has been compressed in the normal way, without the benefit of model-based coding. The one qualification is that any given G would be used for the transmission of many different Ds. If

it is only used once or twice, any saving is likely to be relatively small—because it would be reduced by the cost of transmitting G to Alice and Bob.

The main differences between what is described here and alternative schemes without model-based coding is that, in the latter type of scheme: any "learning" is part of the encoding stage, not an independent process; any such learning is normally relatively unsophisticated and designed to favour speed of processing on low-powered computers rather than high levels of information compression; where there has been some learning, Alice normally transmits both G and E together, not E by itself, meaning much smaller savings than if E is transmitted alone; and while Alice and Bob may be provided with some elements of G—such as the structure of human faces or bodies—this is normally hard coded and not learned, and it is normally restricted to very few kinds of things.

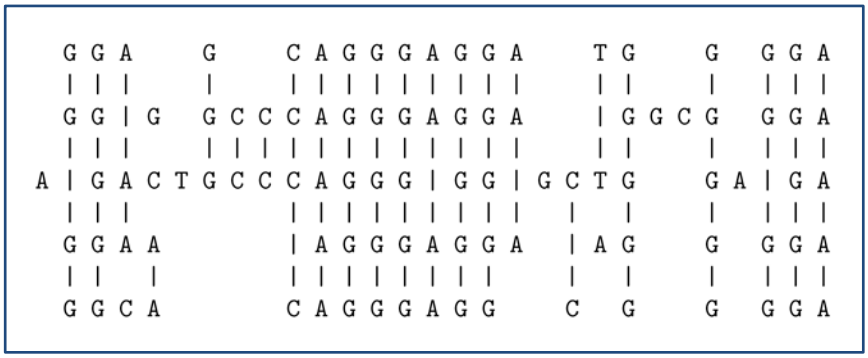
To develop transmission of information via model-based coding as outlined above, a promising way forward is via the *SP theory of intelligence*, outlined in the box. This system, the product of a long-term programme of research, has clear potential to provide the main functions that are needed: unsupervised learning of G; encoding of D in terms of G; and recreation of D from E and G [6, Section VIII].

Regarding the first of these functions, the SP computer model has already demonstrated unsupervised learning of plausible generative grammars for the syntax of English-like artificial languages, including the learning of segmental structures, classes of structure, and abstract patterns. With both the surface forms of language and non-linguistic or "semantic" forms of knowledge, it has clear potential to learn such things as class hierarchies, class heterarchies (meaning class hierarchies with cross classification), part-whole hierarchies, and other forms of knowledge.

The SP theory and the SP computer model

In outline, the SP theory, and its realisation in the SP computer model, has been designed to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and human perception and cognition [4].

The system comprises these main features: 1) All kinds of knowledge are represented with arrays of atomic symbols in one or two dimensions called "patterns"; 2) All kinds of processing are done via a process of searching for patterns or parts of patterns that match each other and via the merging or "unification" of patterns, or parts of patterns, that are the same—with a consequent compression of information; 3) More specifically, all kinds of processing are done via the building and manipulation of *multiple alignments* like the one shown in the figure, but adapted for the SP system; 4) The whole system is inherently probabilistic because of the very close connection that is known to exist between information compression and concepts of prediction and probability.



The powerful concept of multiple alignment, as it has been developed in the SP programme of research, may provide the long-sought-after key to general AI, meaning AI with the versatility and adaptability of human intelligence. I believe it is fair to say that it could be the "double helix" of intelligence—as significant for an understanding of "intelligence" in a broad sense as is DNA for the biological sciences.

In keeping with the quest for simplification and integration across a broad canvass, the SP system has strengths in several different areas including: unsupervised learning, the representation and processing of diverse kinds of knowledge; the processing of natural language, fuzzy pattern recognition, recognition at multiple levels of abstraction, best-match and semantic forms of information retrieval, several kinds of reasoning, planning, and problem solving.

The SP system also has several potential benefits and applications described in peer-reviewed papers that may be downloaded via links from www.cognitionresearch.org/sp.htm.

A key idea in the SP framework is that the entities and abstract concepts discovered by the system would be "natural" in the sense that they would be the kinds of things that people recognise, including specific things like "my cat" and more general concepts like "animal". Evidence to date suggests that the SP system conforms to this principle—*the discovery of natural structures via information compression*, or "DONSVIC" for short [4, Section 5.2]. It appears that unsupervised learning in accordance with the DONSVIC principle requires

relatively high levels of information compression and that the SP system can meet that requirement.

When the SP system has been generalised to process 2D patterns, it is anticipated that unsupervised learning in the SP system may be extended to the learning of 3D digital models of objects, in much the same way that some existing applications can build such models, each one from overlapping digital photographs of an object taken from different angles [5, Section 6.1]. The SP system should also be able to build 3D digital models of environments from overlapping images, much as Google Streetview builds what are essentially 3D models of streets, using overlapping photographs (*ibid.*, Section 6.2).

The second of the functions mentioned above is accommodated in the way the system builds multiple alignments from "New" information (received from the system's environment) and "Old" knowledge (that is derived via earlier learning and is stored for current and future use). A key part of that process is the creation of a relatively compact encoding of the New information in terms of the Old knowledge.

If the SP system is being used by Alice as a means of transmitting big data economically to Bob, then, with a previously-learned G playing the part of Old knowledge and a given body of big data playing the part of New information, the encoding created by the system may play the part of E in the transmission of big data, as described above.

Regarding the third of the functions mentioned above, a neat feature of the SP system is that decoding of information is done in exactly the same way as the encoding of information, with E playing the part of New information and, as before, G playing the part of Old knowledge. So it is a straightforward matter for Bob to use the SP system to decode any E received from Alice, using his own copy of G.

With this kind of oiling of its wheels, big data may glide quickly and efficiently from one place to another, without the need for massive bandwidth, and without needing the output of a small power station to haul it on its way.

References

[1] Kelly, J. E. and Hamm, S., *Smart machines: IBM's Watson and the era of cognitive computing*, New York: Columbia University Press, 2013.

[2] Pierce, J., *Symbols, Signals & Noise*, New York: Harper & Brothers, 1961.

[3] Sayood, K., *Introduction to Data Compression*, Waltham MA: Morgan Kaufmann, 2012, ISBN 978-0-12-415796-5.

[4] Wolff, J. G., "The SP theory of intelligence: an overview", *Information*, 4(3), 283-341, 2013.

[5] Wolff, J. G., "Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision", *SpringerPlus*, 3, 552, 2014.

[6] Wolff, J. G., "Big data and the SP theory of intelligence", *IEEE Access*, 2, 301-315, 2014.

Details of the main publications in the SP programme of research are given, in many cases with download links, near the top of www.cognitionresearch.org/sp.htm.