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*.

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 …". 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 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 like this:

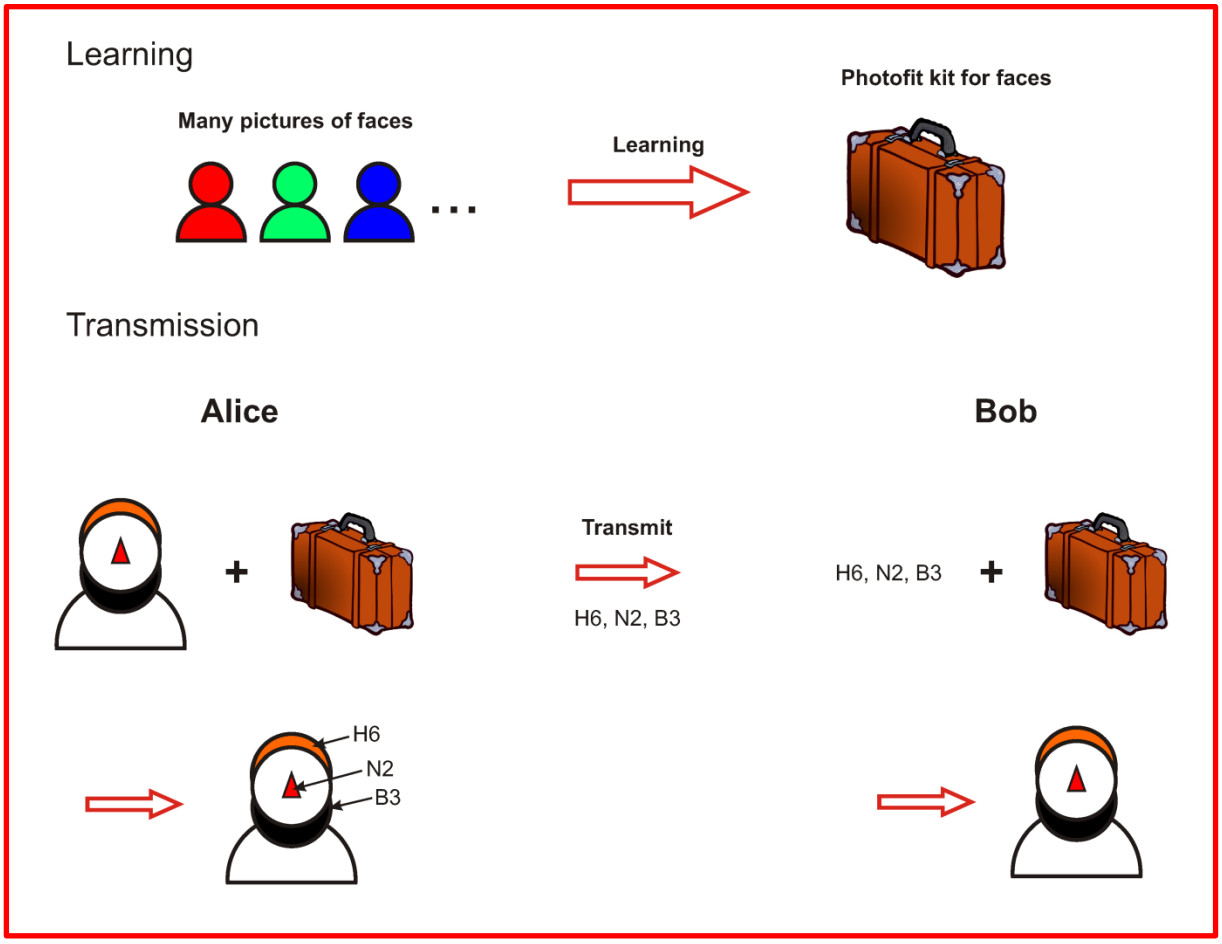
> "Imagine that we had at the receiver a sort of rubbery model of a human face. Or we might have a description of such a model stored in the memory of a huge electronic computer. First, the transmitter would have to look at the face to be transmitted and 'make up' the model at the receiver in shape and 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 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."

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**.

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.”.

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 a bit like the workings of an old-style Photofit system for constructing pictures of faces, as shown in the figure.



In the first “learning” phase, many pictures of faces would be fed into a learning system which would use those data to construct something like a Photofit kit for faces containing pictures of many kinds of hair, many kinds of nose, many kinds of beard, and so on.

In the second “transmission” phase, which may be repeated many times, individual faces would be transmitted from A to B (from “Alice” to “Bob”). For any one face, Alice would use the Photofit kit to identify “codes” for individual features, such as “H6” for red hair, “N2” for a big red nose, and “B3” for a black beard. Then “H6, N2, B3” would be transmitted to Bob who would use that information, with the Photofit kit, to recreate the original face.

Since the codes would normally be very small compared with the information in the corresponding picture of a face, there would be a big saving in the amount of information to be transmitted—much larger than if ordinary compression methods are used—and big savings in the amount of energy needed for transmission.

What has been described is just a simple example to explain the general idea. Something more sophisticated would be needed to transmit the sounds and moving pictures of something like a TV programme.

To develop transmission of information via model-based coding as it has been described, a promising way forward is via the *SP theory of intelligence*, the product of a long-term programme of research. It has clear potential to provide the main functions that are needed.

Without going into details, the SP theory is based on the idea that much of intelligence is about searching for patterns that match each other and then merging such patterns to reduce two or more instances to one.

With learning, the SP computer model has already demonstrated an ability to learn the kinds of things that would be needed in a "kit" like the Photofit kit described above. It has also demonstrated how Alice could use such a kit to create codes from raw information, and how Bob could use the kit to recreate the original information from the codes.

As with the simplified example described earlier, there is clear potential for the SP system to make big cuts in the amount of information to be transmitted in moving big data from one place to another, and big cuts in the amount of energy required. 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.

Details of the main publications in the SP programme of research are given, in many cases with download links, on [*www.cognitionresearch.org/sp.htm*](http://www.cognitionresearch.org/sp.htm)*.*