Information Compression, Multiple Alignment, and Intelligence

J Gerard Wolff
CognitionResearch.org
Email: jgw AT cognitionresearch.org

***Abstract*—This paper provides an overview of the *SP theory of intelligence* and its realisation in the *SP computer model*.**

The central aim in developing the SP system has been to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and human perception and cognition.

Key ideas in the theory are information compression via the matching and unification of patterns (ICMUP) and, more specifically, information compression via a concept of *multiple alignment* adapted from that concept in bioinformatics.

Multiple alignment is potentially the "double helix" of intelligence, potentially as significant for an understanding of intelligence as is DNA for biological sciences.

The background and origins of the SP theory, and the structure and workings of the system are described in outline.

The SP theory has strengths and potential in modelling several different aspects of human intelligence, as outlined in the paper.

There are many potential applications of the SP system that are summarised with references to relevant papers.

Keywords*—*Information compression; Multiple alignment; Artificial intelligence

I. Introduction

For several years, I have been developing the idea that artificial intelligence, mainstream computing, mathematics, and much of human perception and cognition, may be understood as information compression via the matching and unification of patterns (ICMUP) and, more specifically, via a concept of multiple alignment, borrowed and adapted from bioinformatics.

An early version of the idea, as applied to computing, is described in [1]. Since then, progressively more refined versions of the *SP theory of intelligence* have been described in several peer-reviewed articles and, in some detail, in a book [2].[1]

The name "SP" is short for *Simplicity* and *Power* because compression of information may be seen as a process of increasing the simplicity of information (by reducing redundancy) whilst retaining as much as possible of its non-redundant expressive power.

The main aim in this paper is to to provide an overview of the theory and its potential benefits and applications.

[1]Bibliographic details of relevant publications may be found via links from www.cognitionresearch.org/sp.htm.

II. Origins and background

The SP theory has grown out of four main strands of work:

- A body of research, pioneered by Fred Attneave [3], Horace Barlow [4], [5], and others, showing that several aspects of the workings of brains and nervous systems may be understood as compression of information.
- My own research, developing models of language learning, where the importance of information compression became increasingly clear (see, for example, [6]).[2]
- Research on principles of 'minimum length encoding', pioneered by Ray Solomonoff [7], and others.
- Several observations that suggest that information compression has a key role in computing, mathematics, and logic [2, Chapters 2 and 10].

In attempting to simplify and integrate ideas, the SP theory belongs in the same tradition as unified theories of cognition such as Soar [8] and ACT-R [9]. Although the SP programme shares some objectives with projects such as the Gödel Machine [10], and 'universal artificial intelligence' [11], the approach is very different.

III. Outline of the SP theory

The main elements of the SP theory are:

- The central aim in developing the SP system has been to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and human perception and cognition.
- The SP theory is conceived as an abstract brain-like system that receives 'New' information and stores some or all of it as 'Old' information.
- The theory is realised in the *SP computer model*, described in outline below. This may be seen as a preliminary version of the *SP machine*. It is envisaged that this will be developed as a high-parallel software virtual machine, hosted on an existing high-performance computer (Figure 4). This will be a means for researchers to see what can be done with the SP system and to create new versions of it.

[2]Details of this and other relevant publications may be found via www.cognitionresearch.org/lang_learn.html.

- All New and Old information is expressed as arrays of atomic symbols called *patterns* in one or two dimensions. So far, the SP computer model works only with one-dimensional patterns. But it envisaged that the model will be generalised to work with 2D patterns.
- Despite its relative simplicity, the SP system has strengths and potential in several aspects of intelligence as described in Section IV and it has several potential benefits and applications described in Section V.
- The system works by the economical encoding of New patterns in terms of Old patterns, including Old patterns that are newly created by the system and those that have been created previously. This may be seen to achieve such things as unsupervised learning, pattern recognition, parsing or understanding of natural language, and the other aspects of intelligence mentioned above.
- Compression of information is achieved via the matching and unification (merging) of patterns (ICMUP). An important part of this process is an improved version of dynamic programming [2, Appendix A] providing flexibility in matching, and with key roles for the frequency of occurrence of patterns, and their sizes.
- More specifically, information compression is achieved via a concept of *multiple alignment* with ICMUP centre stage. Multiple alignment, which is outlined in Section III-B, is a powerful central idea, similar to the concept of multiple alignment in bioinformatics but with important differences. It has the potential to be the "double helix" of intelligence—as significant for the understanding of intelligence as is DNA in biological sciences.
- Owing to the intimate connection between information compression and concepts of probability (Section **??**), it is relatively straightforward for the SP system to calculate probabilities for inferences made by the system, and probabilities for parsings, recognition of patterns, and so on.
- A 'neural' version of the SP theory—*SP-neural*—describes how abstract concepts in the SP theory may be realised with neurons and their interconnections and intercommunication ([2, Chapter 11], [12]).
- The SP system has several distinctive features and advantages compared with alternative approaches to AI [13]. Section V of that paper describes several problems with 'deep learning in artificial neural networks' and how, in the SP system, these problems may be overcome.

A. The SP computer model

The SP theory is currently expressed in the form of a computer model (SP71). This has two main components: procedures for building multiple alignments and procedures for unsupervised learning, as outlined in the following two subsections.

B. Multiple alignments

The multiple alignment concept in the SP theory has been adapted from a similar concept in bioinformatics, where it means a process of arranging, in rows or columns, two or more DNA sequences or amino-acid sequences so that matching symbols—as many as possible—are aligned orthogonally in columns or rows.

The main difference between the two concepts is that, in bioinformatics, all sequences have the same status, whereas in the SP theory, the system attempts to create a multiple alignment which enables one New pattern (sometimes more) to be encoded economically in terms of one or more Old patterns.

As an illustration of the concept, Figure 1 shows two multiple alignments which are, in effect, two alternative parsings of the ambiguous sentence 'Fruit flies like a banana'.[3]

Each of the two multiple alignments is the best one found by the SP computer model, with a set of Old patterns representing grammatical rules (including words and their grammatical categories) and a New pattern representing the sentence to be parsed. Here, 'best' means that each multiple alignment achieves the greatest degree of compression of the New patterns via its encoding in terms the Old patterns. More detail may be found in [2, Section 3.5].

Although this example does not illustrate the point, it is pertinent to mention that the process of forming multiple alignments is robust in the face of errors. An example is shown in Figure 2 in Section IV-C.

Multiple alignments are built in stages, with pairwise matching and merging of patterns, with merged patterns from any stage being carried forward to later stages, and with a weeding out, at all stages, of low-scoring multiple alignments. This is broadly similar to programs for the creation of multiple alignments in bioinformatics.

At the heart of the process for building multiple alignments is a process for finding good full or partial matches between patterns ([2, Appendix A], [14, Section 4]), somewhat like the WinMerge utility for finding similarities and differences between files, or standard 'dynamic programming' methods for the alignment of sequences. The main difference between the SP process and others, is that the former can deliver several alternative matches between patterns, while WinMerge and standard methods deliver one 'best' result.

C. Unsupervised learning

For the unsupervised learning of Old patterns, there are additional processes of deriving Old patterns from multiple alignments, evaluating newly-created *grammars* (sets of Old patterns) in terms of their effectiveness for the economical encoding of the New information, and the weeding out low-scoring grammars ([2, Sections 3.9, 3.10, and 9.2], [14, Section 5]).

Because of the way the model searches for a global optimum, it does not depend on the presence or absence of any particular feature or combination of features. In both the

[3]This sentence is the second part of *Time flies like an arrow. Fruit flies like a banana.*, attributed to Groucho Marx.

```
0           fruit         flies                 like                 a         banana       0
              |             |                     |                  |           |
1             |             |                     |                  |   < N 1 banana >     1
              |             |                     |                  |   | |          |
2             |             |                     |     < NP 1 < D   | > < N          > >   2
              |             |                     |     | |    | |   | |                |
3             |             |                     |     | |    < D 1 a >                |   3
              |             |                     |     | |                             |
4             |             |           < ADV 0 like >  | |                             |   4
              |             |           |  |         |  | |                             |
5             |             |     < ADP < ADV        >  < NP                            > > 5
              |             |     |  |                                                    |
6     < N 2 fruit >         |     |  |                                                    | 6
      | |         |         |     |  |                                                    |
7 S 0 < N         > < V     |   > < ADP                                                   > 7
                    | |     |   |
8                   < V 0 flies >                                                           8

(a)
0                  fruit         flies           like                 a         banana     0
                     |             |               |                  |           |
1                    |             |               |                  |   < N 1 banana >   1
                     |             |               |                  |   | |          |
2                    |             |               |     < NP 1 < D   | > < N          > > 2
                     |             |               |     | |    | |   | |                |
3                    |             |               |     | |    < D 1 a >                | 3
                     |             |               |     | |                             |
4                    |             |       < V 1 like >  | |                             | 4
                     |             |       | |        |  | |                             |
5 S 1 < NP           |             |     > < V        >  < NP                            > 5
      | |            |             |     |
6     | |    < A 0 fruit >         |     |                                                 6
      | |    | |         |         |     |
7     < NP 0 < A         > < N     |   > >                                                 7
                           | |     |   |
8                          < N 0 flies >                                                   8

(b)
```

Fig. 1: The two best multiple alignments found by a near-identical precursor of the SP computer model with Old patterns representing grammatical rules (rows 1 to 8 in (a) and (b)) and the ambiguous sentence fruit flies like a banana as a 'New' pattern (row 0 in each multiple alignment). Reproduced from Figure 5.1 in [2], with permission.

building of multiple alignments and unsupervised learning, plausible results may be obtained in the face of errors of omission, commission and substitution in the data.

IV. Strengths and potential of the SP system

Despite the relative simplicity of the SP system, it has strengths and potential in several areas, outlined in the following subsections.

A. *Unsupervised learning*

From a set of appropriate sentences, the SP computer model can derive a plausible generative grammar for the syntax of an English-like artificial language, including the learning of segmental structures, classes of structure, and abstract patterns [2, Chapter 9]. This learning is achieved without supervision or error correction by a 'teacher', without the provision of 'negative' samples, and without the grading of samples from simple to complex (*cf.* [15]). It thus overcomes restrictions on what can be achieved with some other models of learning and reflects more accurately what is known about how children learn their first language or languages.

The model draws on earlier research showing that inductive learning via principles of 'minimum length encoding' can lead to the discovery of entities that are psychologically natural—such as words in natural languages [6].

As it stands now, the model is not able to derive intermediate levels of abstraction or discontinuous dependencies in data but those problems appear to be soluble [14, Section 3.3]. The model also needs to be generalised to work with 2D patterns.

B. *Representation of knowledge*

Despite the simplicity of representing knowledge with patterns, the way they are processed within the multiple alignment framework gives them the versatility to represent several kinds of knowledge, including the syntax of natural language (Section IV-C), class hierarchies (with cross-classification), part-whole hierarchies (and their integration with class hierarchies), decision networks and trees, relational tuples, and if-then rules [2, Chapters 5 to 8]. The system may also represent causal relations [2, Section 7.9] and concepts in mathematics, logic, and computing, such as 'function', 'variable', 'value', 'set', and 'type definition' ([2, Chapter 10], [16, Section 6.6]).

The versatility of the SP system in the representation of knowledge suggests that it has potential as a *Universal Framework for the Representation and Processing of Diverse Kinds of Knowledge* (UFK) [17, Section III]. There are considerable potential benefits in the simplification of computing systems.

One universal format for knowledge and one universal framework for processing means that different kinds of knowledge may be combined flexibly and seamlessly according to need. This kind of flexibility is likely to be needed in any system that aspires to human-level intelligence.

C. Natural language processing

Grammatical rules, including words and their grammatical categories, may be represented with SP patterns. As we have seen (Figure **??**) the parsing of natural language may be modelled via the building of multiple alignments [2, Chapter 5]. The same is true of the production of natural language [2, Section 3.8]. The framework provides an elegant means of representing discontinuous dependencies in syntax, including overlapping dependencies such as number dependencies and gender dependencies in languages like French [2, Section 5.4].

As already noted in Section III-B, the SP system is robust in the face of errors in data, as shown in Figure 2. Here, a sentence to be parsed, shown in row 0, contains an error of addition (an '`x`' in '`t h e`'), an error of omission (the '`l`' in '`a p p l e s`'), and an error of substitution (the letter '`k`' replacing '`w`' in '`s w e e t`'). Despite these errors, the SP computer model creates a multiple alignment representing a plausible parsing of the sentence.

As indicated in Section IV-B, the system may also model non-syntactic 'semantic' structures. For reasons discussed in Section IV-I, there is potential for the seamless integration of syntax with semantics—with a consequent potential for the understanding of natural languages. Preliminary trials show how the system may be used in the translation of surface forms into meanings, and in the use of meanings to produce surface forms [2, Section 5.7]. There is also potential for translations amongst natural languages, with meanings functioning as an interlingua amongst different surface forms.

The importance of context in the processing of language is accommodated in the way the system searches for a global best match for patterns: any pattern or partial pattern may be a context for any other [2, Section 5.2.2].

D. Pattern recognition

. Thanks largely to the versatility of the multiple alignment concept, the SP system provides a powerful framework for pattern recognition. It can model pattern recognition at multiple levels of abstraction as shown in Figure 3.[4] Here, an unknown plant with features shown in column 0 is recognised as a Meadow Buttercup (column 1), which is in the genus *Ranunculus* (column 6), which is in the family Ranunculaceae (column 5), and so on.

[4]Compared with the two multiple alignments shown in Figure 1, this multiple alignment has been rotated by 90°, with the New pattern in column 0 and Old patterns in columns 1 to 6. The choice between these two ways of displaying multiple alignments depends purely on what fits best on the page.

A probability may be calculated for any given classification or any associated inference. As in the processing of natural language, the system copes with errors in data, and it provides a role for context in recognition. These ideas appear to have potential in the field of computer vision (Section IV-E).

E. Computer vision

When the SP computer model has been generalised work with patterns in two dimensions, it is likely to have applications in computer vision as discussed in [18].

F. Information retrieval

The SP system provides for the retrieval of information from a knowledge base in the manner of query-by-example, and has potential to support the development of query languages, if required. The system may serve as an intelligent database that also supports the use of traditional data models—but with advantages compared with existing systems ([19], [2, Chapter 6]).

G. Reasoning

The system has strengths with several kinds of reasoning including: one-step 'deductive' reasoning; chains of reasoning; abductive reasoning; reasoning with probabilistic networks and trees; reasoning with 'rules'; nonmonotonic reasoning and reasoning with default values; Bayesian reasoning with "explaining away"; causal reasoning; reasoning that is not supported by evidence; and inference via inheritance of attributes [2, Chapter 7]. The system also has potential for spatial reasoning [20, Section IV-F.1], for what-if reasoning [20, Section IV-F.2], and for commonsense reasoning [21].

H. Planning and problem solving

The SP framework provides a means of planning a route between two places, and, with the translation of geometric patterns into textual form, it can solve the kind of geometric analogy problem that may be seen in some puzzle books and IQ tests [2, Chapter 8].

I. Seamless integration of diverse kinds of knowledge and diverse aspects of intelligence

The provision of one simple format for different kinds of knowledge and one relatively simple framework (multiple alignment) for processing diverse kinds of knowledge is likely to facilitate the seamless integration of diverse kinds of knowledge and diverse aspects of intelligence in any combination.

This kind of integration appears to be essential if we are ever to achieve human-like AI. It appears that most of the problems that we encounter in everyday situations, and which we solve with commonsense knowledge and commonsense reasoning, require seamless integration of relevant knowledge and seamless integration of diverse aspects of intelligence [21].

V. POTENTIAL BENEFITS AND APPLICATIONS

Potential benefits and applications of the SP system are summarised in the following subsections.

```
0                      t h x e               a p p   e    s                a r e        s   k e e t       0
                       | |   |               | | |   |    |                | | |        |     | | |
1                      | |   |    N     Nr 6 a p p l e #N |                | | |        |     | | |       1
                       | |   |    |     |              |  |                | | |        |     | | |
2                      | |   |    N Np N Nr            #N s #N             | | |        |     | | |       2
                       | |   |    | |                       |              | | |        |     | | |
3                 D 17 t h   e #D | |                       |              | | |        |     | | |       3
                  |            |  | |                       |              | | |        |     | | |
4           NP 0a D            #D N |                       #N #NP         | | |        |     | | |       4
            |                       |                           |          | | |        |     | | |
5           |                       |                           |  V Vp 11 a r e #V     |     | | |       5
            |                       |                           |  | |           |      |     | | |
6 S Num   ; NP                      |                          #NP V |           #V A   |     | | | #A #S 6
     |    |                         |                                |              |   |     | | | |
7    |    |                         |                                |             A 21 s w   e e t #A    7
     |    |                         |                                |
8   Num PL ;                        Np                               Vp                                   8
```

Fig. 2: A parsing via multiple alignment created by the SP computer model with errors of omission, commission, and substitution as described in the text. Reproduced from Figure 4 in [17], with permission.

A. Artificial intelligence

Perhaps the main strength and potential of the SP system is in its capabilities in diverse aspects of intelligence (Section IV). As noted in Section IV-I, a valuable feature of the system is its potential for the seamless integration of diverse kinds of knowledge and diverse aspects of intelligence in any combination.

B. Simplification of computing systems, including software

Because the SP system has potential as a UFK (Section IV-B), and because of the central role of information compression in the system, it has potential to yield an overall simplification of computing systems, including software [16, Section 5]. This has potential benefits in minimising faults in computing systems (both hardware and software) and in the time, effort, and cost of creating computer programs (see also Section V-H).

C. Unsupervised learning

The SP system has potential for the unsupervised learning of various kinds of knowledge including grammars for the syntax of natural language, non-syntactic semantic knowledge, and associations between syntax and semantics [16, Section 6.1].

D. Applications with natural language

As we have seen (Section IV-C), the SP system has strengths in the representation of syntactic knowledge and in the parsing and production of natural language.

If the syntax and semantics of natural languages can be learned, and their inter-connections (Section V-C), this would open up applications in the processing of natural language [16, Section 6.2] with potential for human-level understanding and production of natural language and for translations between languages at the level of good human translators.

E. The development of intelligence in autonomous robots

The SP system has potential for the development of human-like versatility and adaptability in autonomous robots and for reductions in the energy required to support that versatility and adaptability ([20], [16, Section 6.3]). Matters discussed include how the SP system may process parallel streams of information; generalisation of knowledge, correction of over-generalisations, and learning from dirty data; how to cut the cost of learning; and reinforcements and motivations [20].

F. Computer vision and the understanding of natural vision

When the SP computer model has been generalised to work with patterns in two dimensions, there is potential for it to be applied in computer vision and in modelling aspects of natural vision. This has been discussed at some length in [18] and more briefly in [16, Section 6.4]. The SP system has potential for the unsupervised learning of visual objects and classes of objects, and suggests how coherent concepts may be derived from fragments. It may be applied in the recognition of objects and also in scene analysis. The theory suggests how, via vision, we may piece together a knowledge of the three-dimensional structure of objects and of our environment, it provides an account of how we may see things that are not objectively present in an image, how we may recognise something despite variations in the size of its retinal image, and how raster graphics and vector graphics may be unified. And it has things to say about the phenomena of lightness constancy and colour constancy, the role of context in recognition, ambiguities in visual perception, and the integration of vision with other senses and other aspects of intelligence.

G. The SP system as an intelligent database

The SP system may serve as a database system with advantages compared with a conventional database system ([19], [16, Section 6.5]). It provides a simple but versatile format for different kinds of knowledge (including those mentioned in Section IV-B). In particular, it provides for the integration of object-oriented representations, relational tuples, and network or tree models. The system has capabilities in artificial intelligence (as described in Section IV), it can function effectively in the face of errors in its input data, and

```
0                 1                2                  3              4              5                  6

                  <species>
                  acris
                  <genus> ---------------------------------------------------------------------------- <genus>
                  Ranunculus ------------------------------------------------------------------------- Ranunculus
                                                                                    <family> --------- <family>
                                                                                    Ranunculaceae ---- Ranunculaceae
                                                                     <order> ------ <order>
                                                                     Ranunculales - Ranunculales
                                                      <class> ------ <class>
                                                      Angiospermae - Angiospermae
                                   <phylum> --------- <phylum>
                                   Plants ----------- Plants
                                   <feeding>
has_chlorophyll ------------------ has_chlorophyll
                                   photosynthesises
                                   <feeding>
                                   <structure> ------ <structure>
                                                      <shoot>
<stem> ---------- <stem> --------------------------- <stem>
hairy ----------- hairy
</stem> --------- </stem> -------------------------- </stem>
                  <leaves> -------------------------- <leaves>
                  compound
                  palmately_cut
                  </leaves> ------------------------- </leaves>
                                                      <flowers> ------------------- <flowers>
                                                                                    <arrangement>
                                                                                    regular
                                                                                    all_parts_free
                                                                                    </arrangement>
                  <sepals> -------------------------------------------------------- <sepals>
                  not_reflexed
                  </sepals> ------------------------------------------------------- </sepals>
<petals> -------- <petals> -------------------------------------------------------- <petals> --------- <petals>
                                                                                    <number> --------- <number>
                                                                                                       five
                                                                                    </number> -------- </number>
                  <colour> -------------------------------------------------------- <colour>
yellow ---------- yellow
                  </colour> ------------------------------------------------------- </colour>
</petals> ------- </petals> ------------------------------------------------------- </petals> -------- </petals>
                                                                                    <hermaphrodite>
<stamens> ------------------------------------------------------------------------- <stamens>
numerous -------------------------------------------------------------------------- numerous
</stamens> ------------------------------------------------------------------------ </stamens>
                                                                                    <pistil>
                                                                                    ovary
                                                                                    style
                                                                                    stigma
                                                                                    </pistil>
                                                                                    </hermaphrodite>
                                                      </flowers> ------------------ </flowers>
                                                      </shoot>
                                                      <root>
                                                      </root>
                                   </structure> ----- </structure>
<habitat> ------- <habitat> ------ <habitat>
meadows --------- meadows
</habitat> ------ </habitat> ----- </habitat>
                  <common_name> -- <common_name>
                  Meadow
                  Buttercup
                  </common_name> - </common_name>
                                   <food_value> ----------------------------------- <food_value>
                                                                                    poisonous
                                   </food_value> ---------------------------------- </food_value>
                                   </phylum> -------- </phylum>
                                                      </class> ----- </class>
                                                                     </order> ----- </order>
                                                                                    </family> -------- </family>
                  </genus> --------------------------------------------------------------------------- </genus>
                  </species>

0                 1                2                  3              4              5                  6
```

Fig. 3: The best multiple alignment created by the SP model, with a set of New patterns (in column 0) that describe some features of an unknown plant, and a set of Old patterns, including those shown in columns 1 to 6, that describe different categories of plant, with their parts and sub-parts, and other attributes. Reproduced from Figure 16 in [14], with permission.

it can function like established database models when that is required.

H. Applicztions in software engineering

The SP system has considerable potential for increasing the efficiency of software engineering [16, Section 6.6]. This may be achieved via two routes that may complement each other:

- *Unsupervised learning*. With appropriate data, there is potential for the automatic or semi-automatic production of software. Here, "appropriate data" means data that completely defines the entities and processes to be modelled. This is most likely to be available for applications in science and probably less likely to be available for applications for business or administration.
- *Systems analysis*. Where the development of software depends on verbal descriptions of how, for example, a business or industrial facility functions, entities and processes may be represented using SP patterns. There is potential for the representation of processing occur in parallel [20, Sections V-G, V-H, and V-I, and Appendix C.].
  In this approach to software engineering, there is no "programming" of the traditional kind. The high-level analysis of the system is also the program. There is no need for instructions to take account of how the underlying hardware works. This is all handled by the SP machine.

Potential advantages of this approach to software engineering are reductions in human time and effort with corresponding benefits in terms of costs and reductions in errors.

I. Information compression

Because the SP machine works by compression of information, it is likely to prove useful where compression of information is the primary requirement [16, Section 6.7].

J. Medical diagnosis

The SP system may be used to assist physicians in medical diagnosis ([22], [16, Section 6.8]). The main attractions are: a format for representing diseases that is simple and intuitive; an ability to cope with errors and uncertainties in diagnostic information; the simplicity of storing statistical information as frequencies of occurrence of diseases; a method for evaluating alternative diagnostic hypotheses that yields true probabilities; and a framework that should facilitate unsupervised learning of medical knowledge and the integration of medical diagnosis with other AI applications.

K. Helping to solve problems with big data

Somewhat unexpectedly, the SP system may help to solve nine problems associated with big data ([17], [16, Section 6.9]):

- *Overcoming the problem of variety in big data*. In computing, the great diversity of different formalisms and formats for knowledge makes it difficult to analyse big data. The potential of the SP system to be a UFK has potential to help solve this problem.
- *Learning and discovery*. The SP system has clear potential for the unsupervised learning or discovery of structures and associations in big data. In this it may be assisted by reductions in the variety of formalisms and formats for knowledge in big data (previous bullet point).
- *Interpretation of data*. The strengths of SP system in such areas as pattern recognition, parsing and production of natural language, several kinds of reasoning, and more (Section IV), may collectively be seen as "interpretation" of data and are clearly relevant to the analysis of big data.
- *Velocity: analysis of streaming data*. The SP system lends itself to an incremental style, assimilating information as it is received, much as people do.
- *Volume: making big data smaller*. Reducing the size of big data via lossless compression can yield direct benefits in the storage, management, and transmission of data, and indirect benefits in areas such as the interpretation of data.
- *Additional economies in the transmission of data*. There is potential for additional economies in the transmission of data, which are potentially very substantial, by judicious separation of 'grammar' and 'encoding'.
- *Energy, speed, and bulk*. There is potential for big cuts in the use of energy in computing, for greater speed of processing with a given computational resource, and for corresponding reductions in the size and weight of computers.
- *Veracity: managing errors and uncertainties in data*. The SP system can identify possible errors or uncertainties in data, suggest possible corrections or interpolations, and calculate associated probabilities.
- *Visualisation*. Knowledge structures created by the system, and inferential processes in the system, are all transparent and open to inspection.

L. Commonsense reasoning

The SP system has potential to solve several problems in the modelling of commonsense reasoning and commonsense knowledge, as described in [23]. A draft paper is in [21].

M. Other areas of potential application of the SP system

Some other potential areas of application for the SP system include: reasoning and the representation of knowledge in the semantic web; bioinformatics; maintaining multiple versions and parts of a document or web page; detection of computer viruses; data fusion; the creation of new kinds of computer; and the development of scientific theories [16, Section 6.10].

VI. Conclusion

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.

Key ideas in the theory are information compression via the matching and unification of patterns and, more specifically, information compression via a concept of multiple alignment, borrowed and adapted from bioinformatics. Multiple alignment has proved to be a powerful idea, potentially the "double helix" of intelligence, potentially as significant for an understanding of intelligence as is DNA for biological sciences.

The SP system has strengths and potential in several different aspects of intelligence, which have been described.

The SP system also has potential in several areas of application, also described.

As mentioned in Section III, a useful step forward in the development of these ideas would be the creation of a version of the SP machine as a high-parallel, open-source, software virtual machine, with a good user interface, hosted on an existing high-performance computer. This would be a means for researchers everywhere to explore what can be done with the system and to create new versions of it. How things may develop is shown schematically in Figure 4.

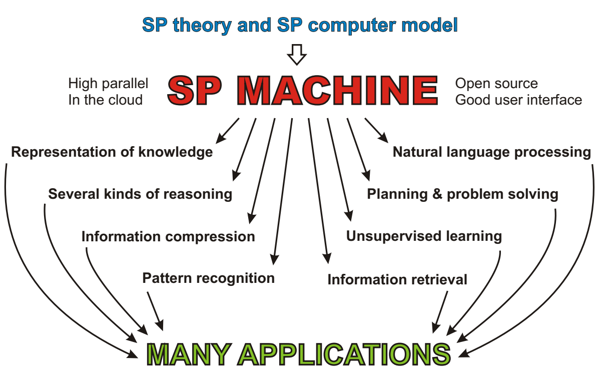

Fig. 4: Schematic representation of the development and application of the SP machine. Reproduced from Figure 2 in [14], with permission.

References

[1] J. G. Wolff, "Simplicity and power—some unifying ideas in computing," *Computer Journal*, vol. 33, no. 6, pp. 518–534, 1990, see bit.ly/R3ga9n.

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

[3] F. Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, pp. 183–193, 1954.

[4] H. B. Barlow, "Sensory mechanisms, the reduction of redundancy, and intelligence," in *The Mechanisation of Thought Processes*, HMSO, Ed. London: Her Majesty's Stationery Office, 1959, pp. 535–559.

[5] ——, "Trigger features, adaptation and economy of impulses," in *Information Processes in the Nervous System*, K. N. Leibovic, Ed. New York: Springer, 1969, pp. 209–230.

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

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

[8] J. E. Laird, *The Soar Cognitive Architecture*. Cambridge, Mass.: The MIT Press, 2012, ISBN-13: 978-0-262-12296-2.

[9] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," *Psychological Review*, vol. 111, no. 4, pp. 1036–1060, 2004.

[10] B. Steunebrink and J. Schmidhuber, "A family of Gödel machine implementations," in *Proceedings of the Fouth International Conference on Artificial General Intelligence (AGI 2011)*, ser. Lecture Notes in Artificial Intelligence, J. Schmidhuber, K. R. Thórisson, and M. Looks, Eds., vol. 6830. Springer, 2011, ISBN 978-3-642-22886-5.

[11] M. Hutter, *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer, 2005, ISBN 3-540-22139-5, www.hutter1.net/ai/uaibook.htm.

[12] J. G. Wolff, "Information compression, multiple alignment, and the representation and processing of knowledge in the brain," *Frontiers in Psychology*, vol. 7, 2016, accepted for publication.

[13] ——, "The SP theory of intelligence: its distinctive features and advantages," *IEEE Access*, vol. 4, pp. 216–246, 2016, bit.ly/21gv2jT.

[14] ——, "The SP theory of intelligence: an overview," *Information*, vol. 4, no. 3, pp. 283–341, 2013, bit.ly/1hz0lFE.

[15] M. Gold, "Language identification in the limit," *Information and Control*, vol. 10, pp. 447–474, 1967.

[16] J. G. Wolff, "The SP theory of intelligence: benefits and applications," *Information*, vol. 5, no. 1, pp. 1–27, 2014, bit.ly/1lcquWF.

[17] ——, "Big data and the SP theory of intelligence," *IEEE Access*, vol. 2, pp. 301–315, 2014, bit.ly/1jGWXDH. This article, with minor revisions, is reproduced in Fei Hu (Ed.), *Big Data: Storage, Sharing, and Security (3S)*, Taylor & Francis LLC, CRC Press, 2016, pp. 143–170.

[18] ——, "Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision," *SpringerPlus*, vol. 3, no. 1, pp. 552–570, 2014, bit.ly/1scmpV9.

[19] ——, "Towards an intelligent database system founded on the SP theory of computing and cognition," *Data & Knowledge Engineering*, vol. 60, pp. 596–624, 2007, bit.ly/Yg2onp, arXiv:cs/0311031 [cs.DB].

[20] ——, "Autonomous robots and the SP theory of intelligence," *IEEE Access*, vol. 2, pp. 1629–1651, 2014, bit.ly/1zrSemu.

[21] ——, "Commonsense reasoning, commonsense knowledge, and the sp theory of intelligence," 2016, in preparation. The current draft may be seen via bit.ly/2eBoE9E.

[22] ——, "Medical diagnosis as pattern recognition in a framework of information compression by multiple alignment, unification and search," *Decision Support Systems*, vol. 42, pp. 608–625, 2006, bit.ly/XE7pRG, arXiv:1409.8053 [cs.AI].

[23] E. Davis and G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence," *Communications of the ACM*, vol. 58, no. 9, pp. 92–103, 2015.