On the "mysterious" effectiveness of mathematics in science

J Gerard Wolff*

August 25, 2017

Abstract

This paper notes first that the effectiveness of mathematics in science appears to some writers to be "mysterious" or "unreasonable". Then reasons are given for thinking that science is, at root, the search for compression in the world. At more length, several reasons are given for believing that mathematics is, fundamentally, a set of techniques for compressing information and their application. From there, it is argued that the effectiveness of mathematics in science is because it provides a means of achieving the compression of information which lies at the heart of science. The anthropic principle provides an explanation of why we find the world—aspects of it at least—to be compressible. Information compression may be seen to be important in both science and mathematics, not only as a means of representing knowledge succinctly, but as a basis for scientific and mathematical inferences—because of the intimate relation that is known to exist between information compression and concepts of prediction and probability.

Keywords: information compression, philosophy of mathematics, philosophy of science, computing, learning, perception, cognition.

1 Introduction

Although mathematics is a phenomenally successful "handmaiden" of science,[1] the reason that it is so effective in science has been described as a "mystery" that is "unreasonable". Thus:

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

[1]The slightly whimsical idea that mathematics might be some kind of servant of science, and the use of the curiously archaic word "handmaiden" seems to have originated with *The Handmaiden of the Sciences*, a book by Eric Bell [1937].

- Roger Penrose writes:

> "It is remarkable that *all* the SUPERB theories of Nature have proved to be extraordinarily fertile as sources of mathematical ideas. There is **a deep and beautiful mystery** in this fact: that these superbly accurate theories are also extraordinarily fruitful simply as *mathematics*." ([Penrose, 1989, pp. 225–226], bold face added).

- In a similar vein, John Barrow writes:

> "**For some mysterious reason** mathematics has proved itself a reliable guide to the world in which we live and of which we are a part. Mathematics works: as a result we have been tempted to equate understanding of the world with its mathematical encapsulization. ... **Why is the world found to be so unerringly mathematical?**" ([Barrow, 1992, Preface, p. vii], bold face added).

- And Eugene Wigner [1960] writes about "The unreasonable effectiveness of mathematics in the natural sciences":

> "The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which **we neither understand** nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure, even though perhaps also to **our bafflement**, to wide branches of learning." (*ibid*, p. 14, bold face added).

1.1 What about Darwinian theory, and human/social sciences?

There is no doubt that mathematics has proved to be very effective in science. But it is pertinent to note here that mathematics works better in some areas of science than in others.

For example, Charles Darwin and Alfred Wallace described their theory of evolution by natural selection with words and pictures. To this day, it is normally described in the same way. And:

> "... against Wigner's 'unreasonable effectiveness' statement (based on success in the physical sciences) one must ask why maths is often

so unreasonably ineffective in the human and social sciences of behaviour, psychology, economics, and the study of life and consciousness. These complex sciences are dominated by non-linear behaviour and only started to be explored effectively by many people (rather than only huge well-funded research groups) with the advent of small personal computers (since the late 1980s) and the availability of fast supercomputers. Some complex sciences contain unpredictabilities in principle (not just in practice): predicting the economy changes the economy whereas predicting the weather doesn't change the weather."[2]

1.2 A possible solution to the mystery

In the light of evidence and arguments described in sections that follow, there appears to be a solution to the mystery of why mathematics is so effective in (some areas of) science. That proposed solution is described in Section 7.

In brief: 1) Apart from the gathering of empirical data, science may be seen to be essentially a process of compressing those data; 2) mathematics may be seen to be a set of techniques for compression of information and their application; thus 3) For those reasons, mathematics can be a valuable aid in the process of compressing information which is a central part of good science. 4) The anthropic principle provides an explanation of why we find that much of the world is compressible (Section 7.1).

Information compression may be seen to be important in both science and mathematics, not only as a means of representing knowledge succinctly, but as a basis for scientific and mathematical inferences—because of the intimate relation that is known to exist between information compression and concepts of prediction and probability (Appendix D).

Much of the thinking in this paper derives from the development of the *SP theory of intelligence* and its realisation in the *computer model*, outlined in Appendix A.

2 SP-multiple-alignment and some basic principles and techniques for compression of information

As a preliminary to what follows, this section describes some basic principles and techniques for compression of information. All of them may be seen to be special

[2]John Barrow—personal communication, 2017-04-06.

cases of the concept of SP-multiple-alignment, a key concept in the SP system, outlined in Appendix A.

As noted there, an idea at the heart of the concept of SP-multiple-alignment is 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".

An example that illustrates the essentials of ICMUP is shown in Section 2.1, below.

This principle and its variants provide an alternative to some of the more mathematical approaches to information compression, and they are arguably more transparent and comprehensible for present purposes than those mathematically-oriented approaches. Also, any theory about the foundations of mathematics should try to reach down to a deeper level than mathematics itself, with foundation concepts that are in some sense more "primitive" than such concepts as *add*, *multiply*, *square root*, and so on.

2.1 Variants of ICMUP

There are five main variants of ICMUP, all of which are widely used in everyday life. All of these five variants of ICMUP may be modelled within the SP-multiple-alignment framework and, within that framework, they may be integrated seamlessly in any combination. The five variants are described briefly in the following five subsections.

2.1.1 Chunking-with-codes

With the first variant—a technique called *chunking-with-codes*—the unified pattern, often referred as a "chunk" of information, is given a relatively short name, identifier, or "code" which is used as a shorthand for the chunk of information wherever it occurs (except for a single 'master' copy). If, for example, the words "Treaty on the Functioning of the European Union" appear in several different places in a document, we may save space by writing the expression once, giving it a short name such as "TFEU", and then using that name as a code or shorthand for the expression wherever it occurs. Likewise for the abbreviation "ICMUP" that is used in this paper.

Figure 1 shows how this would work with the repeating pattern '`INFORMATION`' which, after unification, is assigned the relatively short identifier '`w62`'. Compression of information is achieved when the short identifier ('`w62`') replaces the longer pattern ('`INFORMATION`') that it represents.

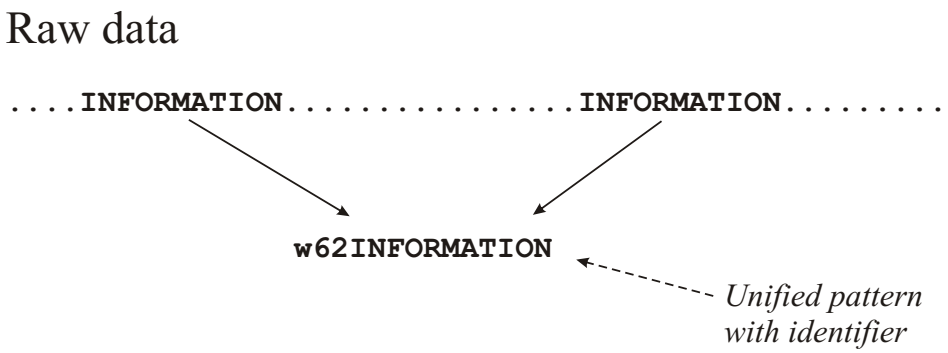

Figure 1: A schematic representation of the way two instances of the pattern 'INFORMATION' in a body of raw data may be unified to form a single 'unified pattern', with 'w62' as an identifier assigned by the system. The lower part of the figure shows how the raw data may be compressed by replacing each instance of 'INFORMATION' with a copy of the corresponding identifer. Reproduced with permission from Figure 2.3 in Wolff [2006].

An important detail here is that compression can be optimised by giving shorter codes to chunks that occur frequently and longer codes to chunks that are rare. This may be done using some such scheme as Shannon-Fano-Elias coding, described in, for example, Cover and Thomas [1991]. There is relevant discussion in [Wolff, 2006, Chapter 2].

2.1.2 Schema-plus-correction

Another variant, *schema-plus-correction*, is like chunking-with-codes but the unified chunk of information may have variations or "corrections" on different occasions. For example, a six-course menu in a restaurant may have the general form "`M1: Appetiser (S) sorbet (M) (P) coffee-and-mints`", with choices at the points marked '`S`' (starter), '`M`' (main course), and '`P`' (pudding). Then a particular meal may be encoded economically as something like '`M1:(3)(5)(1)`', where the digits determine the choices of starter, main course, and pudding.

2.1.3 Run-length coding

A third variant, *run-length coding*, may be used where there is a sequence of two or more copies of a pattern, each one except the first following immediately after its predecessor. In this case, the multiple copies may be reduced to one, as before, with something to say how many copies there are, or when the sequence begins and ends, or, more vaguely, that the pattern is repeated without anything to say when the sequence stops. For example, a sports coach might specify exercises as something like "touch toes (×15), push-ups (×10), skipping (×30), ..." or "Start running on the spot when I say 'start' and keep going until I say 'stop' ".

2.1.4 Class-inclusion hierarchies with inheritance of attributes

In this variant, there is a hierarchy of classes and subclasses, with "attributes" at each level. Each attributes may be seen as a chunk of information and the corresponding class name may be seen to be its code. At every level except the top level, the subclass "inherits" the attributes of all higher levels, thus reducing the need, on any given level, to repeat attributes from higher levels.

2.1.5 Part-whole hierarchies with inheritance of contexts

This is much the same as class-inclusion hierarchies with inheritance of attributes except that the structure represents the parts and subparts of some entity. In this case, each subpart may be seen to inherit its place in larger structures and thus the contexts of structures with which it is associated. As before, this reduces the need, at any given level, to repeat information from higher levels.

2.2 Hiding in plain sight

The five basic techniques for information compression are so familiar that they are often "hiding in plain sight": widely used because they seem like the obvious way to express things, but rarely with any recognition of their role in the compression of information. One example is the way names of things may serve as relatively short codes for relatively complex concepts, and likewise with "content" words in natural language [Wolff, 2017a, Section 5.1].

It seems that these remarks also apply to the use in mathematics of ICMUP and some of its variants, as described in Section 5.2, below.

Appendix C provides some details relating to the frequencies and sizes of repeating patterns, and their codes.

3 Human learning, perception, and thinking as information compression

A central idea in the SP system is that much of human learning, perception, and thinking may be understood as information compression, an idea for which there is now much supporting evidence. Some of this evidence is relatively direct, described in Wolff [1993], [Wolff, 2006, Chapter 2], and Wolff [2017a]. Less direct but nevertheless strong evidence is the way in which the SP computer model, which is dedicated to the compression of information, can model several different aspects of intelligence. Much of this evidence is presented in Wolff [2006, 2013, 2016a] with pointers to where further information may be found.

If it is accepted that much of human cognition may be understood as compression of information, then it should not be surprising to find that both science and mathematics, as products of the human intellect, may also be understood as compression of information.

4 Science as compression of information

Occam's Razor, the principle attributed to William of Ockham and widely seen as a key principle in science, is often expressed as "Entities are not to be multiplied beyond necessity."—meaning that, when there are two or more competing theories that explain a given set of phenomena, we should choose the simplest.

Of course, much of science is concerned with observational studies of aspects of the "world"—meaning the universe as far as we can see—or conducting experiments to obtain empirical data. But few would dispute the 'elegance' or 'beauty' of a compact expression like $E = mc^2$ compared with the relatively huge range of observations that it describes or predicts, and few would dispute the importance in science of discovering or inventing compact descriptions like that.

Respected scientists have often described the goals of science in similar terms. Isaac Newton wrote that "Nature is pleased with simplicity" [Newton, 2014, p. 320]; Ernst Mach [2004] and Karl Pearson [1892] suggested independently of each other that scientific laws promote "economy of thought"; Albert Einstein wrote that "A theory is more impressive the greater the simplicity of its premises, the more different things it relates, and the more expanded its area of application.";[3] cosmologist John Barrow has written that "Science is, at root, just the search for compression in the world" [Barrow, 1992, p. 247]; and Ming Li with Paul Vitányi have written that "Science may be regarded as the art of data compression" [Li and Vitányi, 2014, p. 585]. It is pertinent to mention that George Kingsley Zipf

[3]Quoted in Isaacson [2007, p. 512].

developed the related idea that human behaviour is governed by a "principle of least effort" [Zipf, 1949].

Here are some examples of simplifications in science:

> "... as space and time fuse together in a single concept of spacetime, so the electric field and the magnetic fields fuse together in the same way, merging into a single entity which today we call the electromagnetic field. The complicated equations written by Maxwell for the two fields become simple when written in this new language. ... The concepts of 'energy' and 'mass' become combined in the same way as time and space, and electric and magnetic fields, are fused together in the new mechanics. ... Einstein realizes that energy and mass are two facets of the same entity, just as the electric and magnetic fields are two facets of the same field, and as space and time are two facets of the one thing: spacetime. This implies that mass, by itself, is not conserved; and energy—as it was conceived at the time—is not independently conserved either. One may be transformed into the other: only one single law of conservation exists, not two. What is conserved is the sum of mass and energy, not each separately. Processes must exist that transform energy into mass, or mass into energy." [Rovelli, 2016, Location 812].

4.1 Simplicity and power

Since competing theories rarely address exactly the same set of phenomena, Occam's Razor may be adapted to be "In the development of a scientific theory, we should try to maximise the *simplicity* of the theory whilst retaining as much as possible of its descriptive or explanatory *power*."

There is a close connection between Occam's Razor as just described and the concept of compressing a body of information, **I**. This may be seen to be a process of maximising the *simplicity* of **I**, by extracting repeated information or *redundancy* from **I**, whilst retaining as much as possible of its non-redundant descriptive *power*.

A qualification here is that the results of information compression may be divided into two parts: a 'grammar' **G**, and an 'encoding' of **I** in terms of **G**, which we may call **E**. Here, **G** and **E** together represent lossless compression of **I**. However, **G** may be regarded as a 'theory' of **I** which is normally more 'interesting' than **E**. For reasons of that sort, **E** may sometimes be discarded (Appendix B).

4.2 Representation of knowledge and concepts of prediction and probability

There is much more to information compression than simply reducing the size of a body of information. As described in Appendix D, there is an intimate relation between information compression and concepts of prediction and probability.

Hence, compression of information is important in science, partly as a means of representing scientific knowledge in a succinct form—but at least as important is how information compression provides the key to the making of inferences [Wolff, 2006, Chapter 7] and the calculation of probabilities [Wolff, 2006, Section 3.7].

5 Mathematics as compression of information

The second step in the argument, depends on evidence that mathematics is fundamentally about the compression of information, with a set of techniques for achieving that compression. Subsections that follow present evidence in support of this idea, which we may refer to as mathematics-as-information-compression or MAIC.

5.1 An example of information compression via mathematics

It has been noted already how Einstein's equation, $E = mc^2$, may be seen to be a very compact representation of much data. Here is another example that demonstrates how ordinary mathematics—not some specialist algorithm for the compression of information—can yield high levels of information compression.

Newton's equation for his second law of motion, $s = (gt^2)/2$, is a very compact means of representing any realistically-large table showing the distance travelled by a falling object (s) in a given time since it started to fall (t), as illustrated in Table 1.[4] That small equation would represent the values in the table even if it was a 1000 times or a million times bigger, and so on. Likewise for other equations such as $a^2 + b^2 = c^2$, $PV = k$, $F = q(E + v \times B)$, and so on.

5.2 How techniques for information compression may be seen in the structure and workings of mathematics

This section describes how some of the basic principles and techniques for the compression of information that were outlined in Section 2 may be seen in the

[4] Of course, the law does not work for something like a feather falling in air. The constant, g, is the acceleration due to gravity—about $9.8m/s^2$.

Distance (m)	*Time (sec)*
0.0	0
4.9	1
19.6	2
44.1	3
78.5	4
122.6	5
176.5	6
240.3	7
313.8	8
397.2	9
490.3	10
593.3	11
706.1	12
828.7	13
961.1	14
1103.2	15
1255.3	16
Etc	*Etc*

Table 1: The distance travelled by a falling object (metres) in a given time since it started to fall (seconds).

structure and workings of mathematics.

Of course, these examples do not prove that mathematics may be understood as being entirely devoted to the compression of information. But since the techniques to be described are low-level techniques that are part of the foundations of mathematics and widely used in more complex forms of mathematics, it seems likely that mathematics may indeed be understood in its entirety to be a set of techniques for compressing information and their application.

5.3 ICMUP in mathematics and related fields

Here are some examples where ICMUP may be seen at work in mathematics and related fields:

- In mathematics, the matching and unification of patterns can be seen in the matching and unification of names or identifiers. If, for example, we want to calculate the value of z from these equations: $x = 4$; $y = 5$; $z = x + y$, we need to match the identifier x in the third equation with the identifier x in the first equation, and to unify the two so that the correct value is used for the calculation of z. Likewise for y.

- In a similar way if we wish to invoke or "call" a function such as '`sqrt(x)`' (the square root of x), there must be a match between the name of the function in the call to the function (such as '`sqrt(16)`') and the name of the function in its definition ('`sqrt(x)`'), with unification to assign the value 16 to the variable x.

- The sixth of Peano's axioms for natural numbers—for every natural number n, $S(n)$ is a natural number—provides the basis for a succession of numbers: $S(0)$, $S(S(0))$, $S(S(S(0)))$ ..., itself equivalent to unary numbers in which $1 = /$, $2 = //$, $3 = ///$, and so on. Here, S at one level in the recursive definition is repeatedly matched and unified with S at the next level.

- Emil Post's [1943] "Canonical System", which is recognised as a definition of "computing" that is equivalent to a universal Turing machine, may be seen to work largely via the matching and unification of patterns [Wolff, 2006, Chapter 4]. Much the same is true of the workings of the transition function in a universal Turing machine.

- It is true that logic gates provide the mechanism for finding an address in computer memory but, at a more abstract level, the process may be seen as one of searching for a match between the address held in the CPU and the corresponding address in computer memory. When a match has been found

between the address in the CPU and the corresponding address in memory, there is implicit unification of the two.

- Query-by-example, a popular technique for retrieving information from databases, is essentially a process of finding good matches between a query pattern and patterns in the database, with unification of the best matches.

- A system like Prolog—a computer-based version of logic—may be seen to function largely via the matching and unification of patterns.

5.4 Chunking-with-codes

If a set of statements is repeated in two or more parts of a computer program then it is natural to declare them once as a 'function', 'procedure' or 'sub-routine' within the program and to replace each sequence with a "call" to the function from each part of the program where the sequence occurred. This may be seen as an example of the chunking-with-codes technique for information compression: the function may be regarded as a chunk, while the name of the function is its code or identifier.

Similar things may be done with mathematics, but most of the widely-used functions—such as '`sqrt()`', '`sin()`', or '`cosin()`'—are provided ready-made in environments like Matlab.

Number systems with bases greater than 1, like the binary, octal, decimal and hexadecimal number systems, may all be seen to illustrate the chunking-with-codes technique for compressing information. For example, with the decimal system:

- A unary number like '/////////' may be referred to more briefly as '7'. Here, '/////////' is the chunk and '7' is the code.

- A unary number like '//////////////////' may be split into two parts: '//////////' and '/////////'. Then the first part may be represented by '1' and the second part by '7', giving us the decimal number '17'. The convention is that the right-most digit represents numbers less than 10, and the next digit to the left represents the number of 10s.

- Of course, this 'positional' system can be extended so that digits in the third position from the right represent 100s, digits in the fourth position represent 1000s, and so on.

Here, we can see how the chunking-with-codes technique allows us to eliminate the repetition or redundancy that exists in all unary numbers except '/' so that large numbers, like 2035723, may be expressed in a form that is very much more compact than the equivalent unary number.

5.5 Schema-plus-correction

Most functions in mathematics and computing, like those mentioned above, are not only examples of chunking-with-codes: they are also examples of the schema-plus-correction device for compressing information. This is because they normally require input via one or more "arguments" or "parameters". For example, the square root function needs a number like 16 for it to work on. Without that number, the function is a very general "schema" for solving square root problems. With a number like 16, which may be regarded as a "correction" to the schema, the function becomes focussed much more narrowly on finding the square root of 16.

5.6 Run-length coding

Run-length coding appears in various forms in mathematics, normally combined with other things. Here are some examples:

- Multiplication (eg, 3×4) is repeated addition.
- Division of a larger number by a smaller one (eg, $12/3$) is repeated subtraction. Of course there will be a "remainder" if the larger number is not an exact multiple of the smaller number.
- The power notation (eg, 10^9) is repeated multiplication, which is itself a form of run-length coding.
- A factorial (eg, 25!) is repeated multiplication and subtraction.
- The bounded summation notation (eg, $\sum_{i=1}^{5} \frac{1}{i}$) and the bounded power notation (eg, $\prod_{n=1}^{10} \frac{n}{n-1}$) are shorthands for repeated addition and repeated multiplication, respectively. In both cases, there is normally a change in the value of a variable on each iteration, so these notations may be seen as a combination of run-length coding and schema-plus-correction.
- In matrix multiplication, AB is a shorthand for the repeated operation of multiplying each entry in matrix A with the corresponding entry in matrix B.

All of these examples may be seen as functions with one or more parameters. For example, multiplication may be written $multiply(x, y)$. As functions with parameters, the examples may be seen to illustrate the chunking-with-codes and schema-plus-correction techniques for compressing information (Section 5.5), as well as run-length coding.

5.7 SP-multiple-alignment

Preliminary work described in [Wolff, 2006, Chapter 10] shows that the SP system, with SP-multiple-alignment centre-stage, has potential to model mathematical constructs and mathematical processes. This should not be altogether surprising since, as noted in Section **?**, SP-multiple-alignments can do everything that can be done with the afore-mentioned five variants of ICMUP for compression of information, and it provides for their seamless integration.

Other reasons for believing that the SP system has potential to model many and perhaps all concepts and processes in mathematics are:

- The generality of information compression as a means of representing knowledge in a succinct manner.
- The central role of information compression in the SP-multiple-alignment framework.
- The versatility of the SP-multiple-alignment framework in the representation of knowledge (Appendix A.1).
- The close connection that is known to exist between information compression and concepts of prediction and probability (Appendix D). In case probabilities seem far removed from the non-probabilistic nature of $2 + 2 = 4$, some relevant research is noted in Section 6.5.

5.8 Well-known equations

The well-known equations that were mentioned earlier may all be interpreted in terms of the first three of our five basic techniques for compressing information, thus:

- Einstein's $E = mc^2$ illustrates run-length coding in its power notation (c^2) and in the multiplication of m with c^2.
- Newton's equation for his second law of motion, $s = (gt^2)/2$, illustrates run-length coding in its power notation (t^2), in the multiplication of g with t^2, and in the division of (gt^2) by 2.
- Pythagoras's equation, $a^2 + b^2 = c^2$, illustrates run-length coding via the power notation in a^2, b^2, and c^2.
- Boyle's law, $PV = k$, illustrates run-length coding in the multiplication of P by V.

- The charged particle equation, $F = q(E+v\times B)$, illustrates run-length coding in the multiplication of v by B and in the multiplication of $(E + v \times B)$ by q.

Since multiplication, the power notation, and division, may each be seen as an example of chunking-with-codes and schema-plus-correction (Sections 5.4 and 5.5), as well as run-length coding (Section 5.6), the same can be said about the appearance of those notations in each of the examples above.

6 Related issues

This section considers some issues related to the idea that mathematics may be seen as a set of techniques for the compression of inforamtion, and their application.

6.1 Mathematics as compression of information and the philosophy of mathematics

Amongst the several "isms" in the philosophy of mathematics—foundationism, logicism, intuitionism, formalism, Platonism, neo-Fregeanism, and more—the three which are perhaps most closely related to MAIC are *psychologism* (mathematical concepts derive from human psychology), *embodied mind theories* (mathematical thought is a natural outgrowth of human cognition), and *intuitionism* (mathematics is a creation of the human mind).

Appendix A outlines some of the evidence in support of the view that much of human learning, perception and cognition may be understood as compression of information. This is broadly consistent with the three schools of thought mentioned above.

Probably the most distinctive feature of MAIC is that it does not, to my knowledge, feature in any of psychologism, embodied mind theories, intuitionism, or any other school of thought in the philosophy of mathematics. Also, MAIC may be seen to apply not only to human thinking but also to varied kinds of artificial device for the processing of information.

6.2 The apparent paradox of creating redundancy via information compression

The idea that mathematics or computing is largely, perhaps entirely, about compression of information may seem to conflict with the undoubted fact that, with some simple mathematics or a simple computer program, it is possible to create data containing large amounts of repetition or redundancy.

This issue and how it may be resolved is discussed in [Wolff, 2017a, Appendix C.1].

6.3 Redundancy is often useful in the storage and processing of information

There is no doubt that informational redundancy—repetition of information—is often useful. For example, it is standard practice in computing to maintain two or more copies of important data, and redundancy in messages can provide a useful means of correcting errors. These kinds of uses of redundancy may seem to conflict with the idea that information compression—which means reducing redundancy—is fundamental in mathematics, computing and cognition.

This issue and how it may be resolved is discussed in [Wolff, 2017a, Appendix C.2].

6.4 Mathematical "beauty" and information compression

In a paper with the title "Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes" [Schmidhuber, 2009], and in several earlier papers, Schmidhuber describes how mathematical "beauty", amongst other things, may be understood in terms of the compression of information. His analysis, which is largely in terms of algorithmic information theory and related concepts [Li and Vitányi, 2014], is somewhat different from what has been described in Sections 5.1 to 6.2, above, which attempts to reach down to relatively "primitive" concepts like the discovery of matches between patterns and the merging or "unification" of patterns that are the same.

6.5 Probabilities

If it is accepted that information compression is central in the structure and workings of mathematics, then in view of the intimate connection between information compression and concepts of prediction and probability (Section 4.2 and Appendix D), there are reasons to think that mathematics is fundamentally probabilistic.

Although this seems to conflict with the apparent certainty of equations like $2+2=4$, a probabilistic foundation for mathematics is consistent with the discovery of randomness in number theory:

> "I have recently been able to take a further step along the path laid out by Gödel and Turing. By translating a particular computer program into an algebraic equation of a type that was familiar even to the

> ancient Greeks, I have shown that there is randomness in the branch of pure mathematics known as number theory. My work indicates that—to borrow Einstein's metaphor—God sometimes plays dice with whole numbers." [Chaitin, 1988, p. 80].

As indicated in this quotation, randomness in number theory is closely related to Gödel's incompleteness theorems. These are themselves closely related to the phenomenon of recursion, a feature of many formal systems, several of Escher's pictures, and much of Bach's music, as described in some detail by Douglas Hofstadter in *Gödel, Escher, Bach: An Eternal Golden Braid* [Hofstadter, 1980].

7 An apparent solution to the mystery of why mathematics is so effective in science

In view of evidence that: 1) science is fundamentally a search for compression in the world (Section 4); and evidence that 2) mathematics may be seen to be largely a set of techniques for compressing information and their application (Section 5); and bearing in mind 3) the afore-mentioned intimate relation between information compression and concepts of prediction and probability (Appendix D); it seems reasonable to conclude that those three things may explain why mathematics is so effective as a means of representing scientific knowledge and in the making of scientific inferences.

There is relevant discussion in Appendix B.

7.1 The anthropic principle

An objection to the arguments above is that, while $E = mc^2$ is undoubtedly a compressed representation of the data that it describes, that observation does not explain why nature can be so compressible. But a possible answer is that:

> "... maths is best thought of as the catalogue of all possible patterns and so it is inevitable that mathematics is effective in describing the world—it could not be otherwise because the world must have pattern to allow life to exist."[5]

And this appeal to the anthropic principle[6] may be adapted for information compression as something like: "the world must be compressible because otherwise everything, including ourselves, would be a soup of randomness."

[5] John Barrow—personal communication, 2017-04-06.

[6] "Anthropic principle", *Wikipedia*, bit.ly/2pVf1W8, retrieved 2017-04-24.

8 Computing

Most of this paper is about mathematics but of course there is a close relation between mathematics and computing in its modern sense of computing by machine, witness the several examples in the paper drawn from the field of computing. It seems likely that most of what has been said about mathematics in the paper will apply also to computing. That said, software systems often employ two of the previously-mentioned devices for compression of information (Section 2.1)—class-inclusion hierarchies and part-whole hierarchies—which are rarely if ever employed in mathematics.

It is perhaps relevant to mention here that the concept of a universal Turing machine is often seen as point of reference in theories of computation[7] and related concepts such as 'computability',[8] and 'algorithmic information theory'.[9] And that, by contrast, the SP theory of intelligence (Appendix A) is itself a theory of computation, based on the conjecture, supported by evidence, that much of artificial intelligence, mainstream computing, mathematics, and human learning, perception, and cognition, may be understood as information compression.

With the second perspective, the SP theory is perhaps better seen as providing a framework for understanding the concept of a universal Turing machine, rather than the other way round. In that connection, ICMUP may be seen in the workings of the 'transition function' within the universal Turing machine.

The key differences between the SP system as a theory of computing and earlier models such as the concept of a universal Turing machine, a Post canonical system, and lambda calculus is that:

- The SP system is founded on the concept of SP-multiple-alignment which is itself an expression of ICMUP.
- It is designed to achieve relatively high levels of information compression via SP-multiple-alignment with the application of heuristic methods in the building of SP-multiple-alignments and in the creation of grammars that score well in terms of their capacity for the economical encoding of a target body of information.
- Via the concept of SP-multiple-alignment, it incorporates all five of the variants of ICMUP described in Section 2.1.

[7] See, for example, "Theory of computation", *Wikipedia*, bit.ly/2sQxseE, retrieved 2017-07-06.

[8] See, for example, 'computability', *Wikipedia*, bit.ly/2tmrAM5, retrieved 2017-07-04.

[9] See, for example, "Algorithmic information theory", *Wikipedia*, bit.ly/2tMZ1KB, retrieved 2017-07-06.

- It has strengths in artificial intelligence which are missing from the earlier models.
- It has clear potential to achieve an overall simplification of computing systems, including software, as described in [Wolff, 2014, Section 5].

Reasons for thinking that the SP system is or has the potential to be a 'universal' system for performing any kind of computation is discussed in [Wolff, 2006, Chapter 4] and [Wolff, 2017c, Appendix B].

9 Conclusion

This paper notes first that the effectiveness of mathematics in science appears to some writers to be "mysterious" or "unreasonable". Then reasons are given for thinking that science is fundamentally a search for compression of empirical data. At more length, several reasons are given for believing that mathematics is, fundamentally, a set of techniques for compressing information—including the matching and unification of patterns, chunking-with-codes, schema-plus-correction, and run-length coding—and their application. From there, it is argued that mathematics has proved to be effective in science because it provides a means of achieving the compression of information which lies at the heart of science.

The anthropic principle provides an explanation of why we find the world—aspects of it at least—to be compressible.

Information compression may be seen to be important in both science and mathematics, not only as a means of representing knowledge succinctly, but as a basis for scientific and mathematical inferences—because of the intimate relation that is known to exist between information compression and concepts of prediction and probability.

That mathematics may be seen to be a set of techniques for compressing information and their application, is in keeping with the view that much of human learning, perception, and cognition may understood as the compression of information.

In accordance with Occam's Razor and the best traditions in science, these ideas may help to unify thinking across the philosophy of mathematics and the philosophy of science (Sections 2 to 7), concepts of 'computing' (Section 8), artificial intelligence (Appendix A.1), human learning, perception, and thinking (Section 3), and neuroscience (Section A.2).

These ideas may also have a bearing on how mathematics may be developed in the future Wolff [2017b].

Acknowledgements

I'm grateful to Roger Penrose and John Barrow for helpful comments on an earlier version of this paper. For helpful comments on drafts of a related paper, I'm grateful to Robert Thomas, Michele Friend, and Alex Paseau. I'm also grateful for discussion from time to time of some of the ideas in this paper with Tim Porter and Chris Wensley.

A The SP theory of intelligence and the SP computer model

As noted in the Introduction, much of the thinking in this paper derives from the *SP theory of intelligence* and its realisation in the *SP computer model*. This theory, which is described quite fully in Wolff [2006] and more briefly in Wolff [2013], aims to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and human learning, perception, and cognition.

Several other papers in the SP programme of research, most with download links, may be found via www.cognitionresearch.org/sp.htm.

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 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 2. 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.

The key difference between the concept of multiple alignment in bioinformatics and the concept of SP-multiple-alignment is that, in the latter, a 'good' SP-multiple-alignment is one that allows one New pattern (sometimes more than one) to be encoded economically in terms of one or more 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.

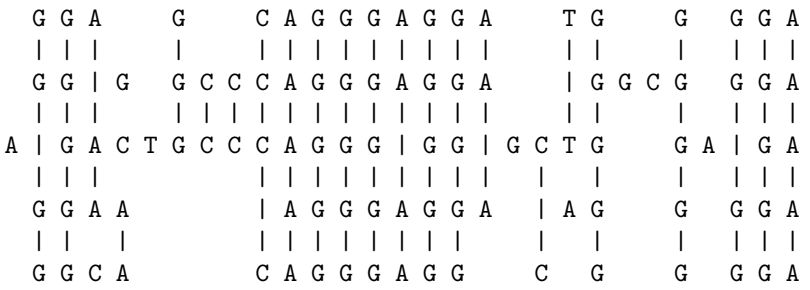

Figure 2: A 'good' multiple alignment amongst five DNA sequences. Reproduced with permission from Figure 3.1 in Wolff [2006].

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 2.1.

A.1 Strengths of the SP system

Distinctive features and advantages of the SP system compared with other AI-related systems are described in Wolff [2016a].

Key strengths of the SP system, due mainly to the powerful concept of SP-multiple-alignment, are: versatility in the representation of knowledge; versatility in aspects of intelligence; seamless integration of diverse kinds of knowledge and diverse aspects of intelligence, in any combination. More detail may be found in [Wolff, 2017c, Appendix B].

It appears that the concept of SP-multiple-alignment has the potential to be as significant for an understanding of intelligence, broadly construed, as has DNA for an understanding of many phenomena in biological sciences. SP-multiple-alignment may prove to be the "double helix" of intelligence.

A.2 SP-neural

Key concepts in the SP theory may be mapped on to structures of neurons and their interconnections in a version of the SP theory called SP-neural Wolff [2016b].

B Generalisation in science

Science is not merely about describing things in an economical manner, it is about making predictions or inferences that go beyond what has actually been observed. As described in Appendix D, there is an intimate connection between information compression and concepts of prediction and probability.

This appendix summarises some ideas discussed elsewhere (Wolff [2006, Section 2.2.12], Wolff [2013, Section 5.3]) about how we can or should generalise our concepts without over-generalisation (sometimes called under-fitting) or under-generalisation (sometimes called over-fitting).

This issue is important in understanding how a child learns his or her first language or languages. The learning of a given language, **L**, is based on a finite sample of **L** that has been heard, normally quite large. This is represented by the smallest envelope in Figure 3, marked as "A sample of utterances".[10] From that finite sample, we learn to understand and to produce a range of possible utterances that is much larger than the finite sample we have heard. This is represented by the envelope marked "All utterances in language L". But although that range of utterances is large, it is smaller than the range of all possible utterances represented by the envelope marked "All possible utterances". Notice that the smallest envelope—the basis for learning—is partly inside the envelope for "All utterances in language L" and partly outside it: children learn partly from good examples of **L** and partly from corrupted examples of **L**, which are marked in the figure as "dirty data".

In summary, learning **L** means generalising beyond the finite sample of utterances that we have heard but without either over-generalisation or under-generalisation, and it means somehow correcting for dirty data. Although young children say things like "gooses" and "sheeps"—apparently overgeneralising from what they have heard—they grow out of these errors. The weight of evidence is that children can learn a first language without the need for explicit correction of errors by adults or older children.[11]

There is evidence that learning of **L** can be achieved, without over-generalisation or under-generalisation, correcting for dirty data, and without the need for explicit correction of errors. Here's how:

1. Start with a finite sample of **L** which we may call **I**. The sample may contain dirty data as described above.

[10] The weight of evidence is overwhelmingly against the nativist, Chomskian view that children are born with substantial knowledge of the structure of language. Some of the evidence is described in Wolff [1988, pp. 208–209 and pp. 210-211]. Perhaps the strongest argument, not made in that publication, is that, to explain why a newborn baby can learn any natural language, the nativist view depends on the existence of some kind of *universal grammar* that describes the structure of every one of the thousands of natural languages and is in every infant's head at the time of his or her birth. Despite decades of research, no such universal grammar has been found.

[11] Christy Brown was a cerebral-palsied child who not only lacked any ability to speak but whose bodily handicap was so severe that for much of his childhood he was unable to demonstrate that he had normal comprehension of speech and non-verbal forms of communication [Brown, 2014]. Hence, his learning of language must have been achieved without the possibility that anyone might correct errors in his spoken language.

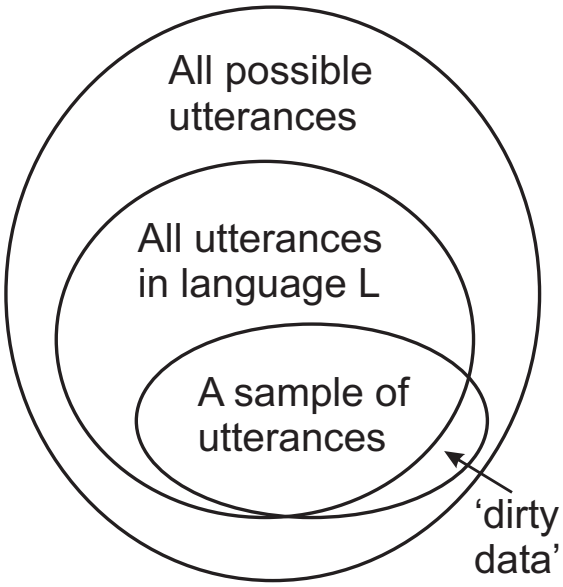

Figure 3: Categories of utterances involved in the learning of a first language, **L**. In ascending order size, they are: the finite sample of utterances from which a child learns; the (infinite) set of utterances in **L**; and the (infinite) set of all possible utterances. Adapted from Figure 7.1 in Wolff [1988], with permission.

2. Compress **I** using something like the SP system, designed to achieve high levels of lossless information compression.[12]

3. The result of compressing **I**, which we may call **IC**, may be divided into two parts: a 'grammar' **G**, and an 'encoding' of **I** in terms of **G**, which we may call **E**.

4. Of **G** and **E**, the most interesting seems normally to be **G**, which may be regarded as a 'theory' of **I**. It appears that, normally, **G** represents a distillation of the 'essence' of **I**, weeding out any dirty data in **I**, and generalising from **I** without overgeneralising. **E** is a description of **I** in terms of the theory, **G**.

This solution for language learning appears to be general and applicable to the learning of any kind of kind of knowledge. It seems likely that, for example, it will provide a solution to the problem of how, via unsupervised learning, a concept like 'horse' may be learned without under-generalisation (meaning that, for example,

[12]Ordinary compression algorithms, like the popular 'ZIP' algorithms, are not really suitable because they are designed to work fast with low-powered computers and may thus miss relatively large amounts of redundancy.

the system would only recognise horses that are very similar to, or identical with, the examples of horses in **I**), and without over-generalisation (meaning that, for example, the system would regard cows, sheep, or dogs, as horses).

It appears that this solution is altogether simpler and more comprehensive than several alternatives, as discussed in Wolff [2016a, Section V-H].

Without such generalisation, any learning system would be severely handicapped: only able to recognise or understand things that were exactly the same as it had seen before.

C Frequency of occurrence, sizes of patterns, and ICMUP

A point to notice about ICMUP is that, to achieve lossless compression of information, it is necessary to use some kind of "code" to mark the positions of redundant copies of any pattern that have been unified, as outlined under the heading "Chunking-with-codes" in Section 2.1. And to ensure that there is an overall compression of a given body of information, **I**, it is necessary to ensure that:

- The given pattern must repeat more often in **I** than we would expect by chance *for patterns of that size*. In general, large patterns yield more compression than small ones, and the minimum frequency needed to achieve information compression is smaller for large patterns than it is for small patterns.
- The code should not be too large. Normally, its size should be at or near the theoretical minimum needed to ensure an overall compression of **I**.

These points relate to the close connection between information compression and concepts of prediction and probability, discussed in Appendix D.

D Information compression and concepts of prediction and probability

It has been recognised for some time that there is an intimate connection between information compression and concepts of prediction and probability, as described in Ray Solomonoff's Algorithmic Probability Theory [Solomonoff, 1964, 1997], and in the closely-related Kolmogorov Complexity Theory [Li and Vitányi, 2014]. Information compression and concepts of prediction and probability may be seen as two sides of the same coin.

The close connection between those two things makes sense in terms of ICMUP (Section 2):

- A pattern that repeats is one that, via inductive reasoning, we naturally regard as a guide to what may happen in the future (more in Appendix D.2, below).
- A pattern that repeats is one that, via unification, is likely to yield compression of information.
- A partial match between one pattern and another can be the basis for predicting the occurrence of the unmatched parts, a form of inference that is sometimes called *prediction by partial matching*.[13]

D.1 Frequencies of occurrence, sizes of patterns, and probabilities

A point of interest is that, in the same way that information compression depends partly on the frequency of occurrence of a pattern that repeats, and also on its size (Appendix C), the probabilities of inferences that may be drawn from any repeating pattern depend partly on the frequency of occurrence of the given pattern and partly on its size. How relevant calculations are made in the SP system is described in Wolff [2006, Sections and 3.7 7.2].

More specifically, a repeating pattern of size **S** can only yield inferences with probabilities greater than chance if its frequency of occurrence within a given body of information, **I**, is greater than the 'threshold' frequency of occurrence that would be expected by chance *for a pattern of size* **S**; and that threshold frequency is greater for small patterns than it is for large patterns.

Consider, for example, what inferences one might make from the occurrence, in an English text, of the neighbouring letters, '`th`'. Given only those two letters, one may guess that they may be part of several different words such as '`the`', '`this`', '`that`', '`those`', and so on, each one with a probability that is substantially less than 1. But although, notwithstanding its fame, the pattern of words, '`Let me not to the marriage of true minds`', is much rarer in English than the pattern '`th`', we infer with near certainty that it will be followed by the words '`Admit impediments`'.

With regard to frequencies and sizes of patterns in the calculation of probabilities:

[13] See "Prediction by partial matching", *Wikipedia*, bit.ly/1BUtAYo, retrieved 2017-03-01.

- There is a sharp contrast between the SP system, which takes account of both the frequencies and the sizes of patterns in calculating probabilities, and frequentist approaches to statistics which emphasise the frequencies of occurrence of entities, without taking account of their sizes.

- "Hebbian" learning, first proposed by neuroscientist Donald Hebb [1949, p. 62], with a central role in most versions of "deep learning" [Schmidhuber, 2015], is focussed entirely on the frequency with which one neuron fires another, without any role for the sizes of neural structures involved in learning, perception and cognition.

D.2 Inductive reasoning

With regard to inductive reasoning mentioned in the first bullet point in Appendix D:

> "We can, of course, ... ask, as philosophers have done for many years: 'What is the rational basis for inductive reasoning?' Why do most people have this strong intuition that because the sun has always risen every morning it will do it again tomorrow, or because every paving stone in a path has held our weight so far, the next one will too? None of these conclusions can be proved logically.
>
> "It is no good arguing that inductive reasoning is rational because it has always worked in the past. This argument eats its own tail. Here is an argument why inductive reasoning is rational which does not depend on the principle which it is trying to justify:
>
> > "If we assume that the world, in the future, will contain redundancy in the form of recurring patterns of events, then brains and computers which store information and make inductive inferences will be useful in enabling us to anticipate events. If it turns out that the world, in the future, does indeed contain redundancy then our investment in the means of storing and processing information will pay off. If it turns out that the world, in the future, does not contain redundancy then we are dead anyway—reduced to a pulp of total chaos!
>
> "This kind of reasoning made fortunes for speculators after World War II: it was rational to buy up London bomb sites during the war because, if the war were won, they would become valuable. If the war

were to be lost, the money saved by not making the investment would, in an uncomfortable and uncertain future, probably not be much use anyway." [Wolff, 1991, pp. 28–29].

References

E. C. Banks. The philosophical roots of Ernst Mach's economy of thought. *Synthese*, 139:25–53, 2004.

J. D. Barrow. *Pi in the Sky*. Penguin Books, Harmondsworth, 1992.

E. T. Bell. *The Handmaiden of the Sciences*. The Williams & Wilkins Company, Baltimore, 1937.

C. Brown. *My Left Foot*. Vintage Digital, London, Kindle edition, 2014. First published in 1954.

G. J. Chaitin. Randomness in arithmetic. *Scientific American*, 259(1):80–85, 1988.

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

D. O. Hebb. *The Organization of Behaviour*. John Wiley & Sons, New York, 1949.

D. Hofstadter. *Gödel, Escher, Bach: an Eternal Golden Braid*. Penguin Books, Harmondsworth, 1980.

W. Isaacson. *Einstein: His Life and Universe*. Pocket Books, London, Kindle edition, 2007.

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

I. Newton. *The Mathematical Principles of Natural Philosophy*. The Perfect Library, Kindle edition, 2014. First published 1687. Illustrated and bundled with *Life of Sir Isaac Newton*.

K. Pearson. *The Grammar of Science*. Walter Scott, London, 1892. Republished by Dover Publications, 2004, ISBN 0-486-49581-7. Internet archive: bit.ly/1g2gNfk.

R. Penrose. *The Emperor's New Mind*. Oxford University Press, Oxford, 1989.

E. L. Post. Formal reductions of the general combinatorial decision problem. *American Journal of Mathematics*, 65:197–268, 1943.

C. Rovelli. *Reality Is Not What It Seems: The Journey to Quantum Gravity*. Penguin Books, London, kindle edition, 2016.

J. Schmidhuber. Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In G. Pezzulo, M. V. Butz, O. Sigaud, and G. Baldassarre, editors, *Anticipatory Behavior in Adaptive Learning Systems, from Sensorimotor to Higher-level Cognitive Capabilities*, Lecture Notes in Artificial Intelligence. Springer, Berlin, 2009.

J. Schmidhuber. Deep learning in neural networks: an overview. *Neural Networks*, 61:85–117, 2015. doi: 10.1016/j.neunet.2014.09.003.

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

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

E. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics*, 13:1–14, 1960.

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.

J. G. Wolff. *Towards a Theory of Cognition and Computing*. Ellis Horwood, Chichester, 1991.

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

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.

J. G. Wolff. The SP theory of intelligence: an overview. *Information*, 4(3):283–341, 2013. doi: 10.3390/info4030283. bit.ly/1NOMJ6l.

J. G. Wolff. The SP theory of intelligence: benefits and applications. *Information*, 5(1):1–27, 2014. doi: 10.3390/info5010001. bit.ly/1FRYwew.

J. G. Wolff. The SP theory of intelligence: its distinctive features and advantages. *IEEE Access*, 4:216–246, 2016a. doi: 10.1109/ACCESS.2015.2513822. bit.ly/2qgq5QF.

J. G. Wolff. Information compression, multiple alignment, and the representation and processing of knowledge in the brain. *Frontiers in Psychology*, 7:1584, 2016b. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.01584. bit.ly/2esmYyt.

J. G. Wolff. Evidence that much of the workings of brains and nervous systems may be understood as compression of information via the matching and unification of patterns. Technical report, CognitionResearch.org, 2017a. Submitted for publication. See bit.ly/2ruLnrV.

J. G. Wolff. Towards a new mathematics for science. Technical report, CognitionResearch.org, 2017b. bit.ly/2o1pr8p. This report is also achieved in vixra.org/ and hal.archives-ouvertes.fr/hal-01534619.

J. G. Wolff. Software engineering and the sp theory of intelligence. Technical report, CognitionResearch.org, 2017c. Submitted for publication. bit.ly/2w99Wzq.

G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Hafner, New York, 1949. Republished by Martino Publishing, Mansfield Centre, CT, 2012.