Self-Dissimilarity: An Empirically Observable Complexity Measure

David H. Wolpert
NASA Ames Research Center
MS269-2, Moffett Field, CA, 94035
William G. Macready
Bios Group LP
317 Paseo de Peralta
Santa Fe, NM, 87501

For many systems characterized as "complex/living/intelligent" the spatio-temporal patterns exhibited on different scales differ markedly from one another. For example the biomass distribution of a human body "looks very different" depending on the spatial scale at which one examines that biomass. Conversely, the density patterns at different scales in "dead/simple" systems (e.g., gases, mountains, crystals) do not vary significantly from one another. Accordingly, we argue that the degrees of self-*dis*similarity between the various scales with which a system is examined constitute a complexity "signature" of that system. Such signatures can be empirically measured for many real-world data sets concerning spatio-temporal densities, be they mass densities, species densities, or symbol densities. This allows one to compare the complexity signatures of wholly different kinds of systems (e.g., systems involving information density in a digital computer, vs. species densities in a rain-forest, vs. capital density in an economy, *etc.*). Such signatures can also be clustered, to provide an empirically determined taxonomy of "kinds of systems" that share organizational traits. The precise measure of dissimilarity between scales that we propose is the amount of extra information on one scale beyond that which exists on a different scale. This "added information" is perhaps most naturally determined using a maximum entropy inference of the distribution of patterns at the second scale, based on the provided distribution at the first scale. We briefly discuss using our measure with other inference mechanisms (e.g., Kolmogorov complexity-based inference).

1 Introduction

Historically, the concepts of life, intelligence, culture, and complexity have resisted all attempts at formal scientific analysis. Indeed, there are not even widely agreed-upon formal definitions of those terms [6, 3]. Why is this?

We argue that the underlying problem is that many of the attempted analyses have constructed an extensive formal model before considering any experimental data. For example, some proposed definitions of complexity are founded on statistical mechanics [7], while others use computer science abstractions like finite automata [5] or universal Turing machines [4, 8, 2]. None of these models arose from consideration of any particular experimental data.

This contrasts with the more empirical approach that characterized the (astonishingly successful) growth of the natural sciences. This approach begins with the specification of readily measurable "attributes of interest" of real-world phenomena followed by observation of the inter-relationships of those attributes in real-world systems. *Then* there is an attempt to explain those inter-relationships via a theoretical model. For the most part, the natural sciences were born of raw experimental data and a need to explain it, rather than from theoretical musing.

It is not difficult to see why data-driven approaches may be more successful in general. In many respects, before a model-driven approach can be used to assign a complexity to a system, one must already fully understand that system (to the point that the system is formally encapsulated in terms of one's model class). So only once most of the work in analyzing the system has already been done can one investigate that system using these proposed measures of complexity. Another major problem with model-driven approaches is that they are prone to degeneration into theorizing and simulating, in isolation from the real world. This lack of coupling to experimental data vitiates the most important means by which theoretical models can be compared, refuted, and modified.

In this paper we follow a more data-driven approach, in which we start with an attribute of interest. Our choice for attribute of interest is based on the observation that most systems that people characterize as complex/living/intelligent have the following property: *over different space and time scales, the patterns exhibited by a complex system vary greatly, and in ways that are unexpected given the patterns on the other scales.* Accordingly, a system's self-*dis*similarity is the attribute of interest we propose be measured — completely devoid of the context of any formal model at this point. (Bar Yam also proposes a complexity profile which is based on the characteristics of a system at different scales — see [1].)

The human body is a familiar example of such self-dissimilarity; as one changes the scale of the spatio-temporal microscope with which one observes the body, the pattern one sees varies tremendously. Other examples from biology are how, as one changes the scale of observation, the internal structures of a biological cell, or of an ecosystem, differ greatly from one another. By measuring patterns in quantities other than the mass distribution (*e.g.*, in information distributions), one can also argue that the patterns in economies and other cultural institutions vary enormously with scale. It may also be that as one changes the scale of observation there are also large variations in the charge density patterns inside the human brain.

In contrast, simple systems like crystals and ideal gases may exhibit some variation in pattern over a small range of scales, but invariably when viewed over broad ranges of scales the amount of variation falls away. Similarly, viewed over a broad range of spatio-temporal scales (approximately the scales from complexes of several hundred molecules on up to microns), a mountain, or a chair, would appear to exhibit relatively little variation in mass density patterns. As an extreme example, relative to its state when alive, a creature that has died and decomposed exhibits no variation over temporal scales. Such a creature also exhibits far less variation over spatial scales than it did when alive.

Our thesis is that variation in a system's spatio-temporal patterns as one changes scales is not simply a side-effect of what is "really going on" in a complex system. Rather it is a crucial aspect of the system's complexity. We propose that it is only after we have measured such self-dissimilar aspects of real-world systems, when we have gone on to construct formal models explaining those data, that we will have models that "get at the heart" of complex systems.

There are a number of apparent contrasts between our proposed approach and much previous work on complexity. In particular, fractals have often been characterized as being incredibly complex due to their possessing nontrivial structure at all different scales; in our approach they are instead viewed as relatively simple objects since the structure found at different scales is in many respects the *same*.

Similarly, a cottage industry exists in finding self-similar degrees of freedom in all kinds of real-world systems, some of which can properly be described as complex systems. Our thesis is that independent of such self-similar degrees of freedom, it is the alternative self-dissimilar degrees of freedom which are more directly important for analyzing a system's complexity. We hypothesize that, in large measure, to concentrate on self-similar degrees of freedom of a complex system is to concentrate on the degrees of freedom that can be very compactly encoded, and therefore are not fundamental aspects of that system's complexity.

As an example, consider a successful, flexible, modern corporation, a system that is "self-similar" in certain variables ([9]). Consider such a corporation that specializes in an information processing service of some sort, so that its interaction with its environment can be characterized primarily in terms of such processing rather than in terms of gross physical manipulation of that environment. Now hypothesize that in *all* important regards that corporation is self-similar. Then the behavior of that corporation — and in particular its effective dynamic adaptation to and interaction with its environment — is specified using the extremely small amount of information determining the scaling behavior. In such a situation, one could replace that adaptive corporation with a very small computer program based on that scaling information, and the interaction with the environment would be unchanged. The patent absurdity of this claim demonstrates that *what is most important* about a corporation is not captured by those variables that are self-similar.

More generally, even if one could find a system commonly viewed as complex that was clearly self-similar in all important regards, it is hard to see how the same system wouldn't be considered even more "complex" if it were self-dissimilar. Indeed, it is hard to imagine a system that is highly self-dissimilar in both space and time that wouldn't be considered complex. Self-dissimilarity would appear to be a sufficient condition for a system to be complex, even if it is not a necessary condition.

In Section 2 we further motivate why self-dissimilarity is a good measure of complexity. Section 3 then takes up the challenge of formalizing some of these vague notions. The essence of our approach is the comparison of spatio-temporal structure at different scales. Since we adopt a strongly empirical perspective, how to infer structure on one scale from structure on another is a central issue. This naturally leads to the probabilistic measure we propose in this section. Finally, in Section 4 we discuss some of the general attributes of our measure and how to estimate it from data. In future work we plan to apply those estimation schemes to real-world data sets.

It is worth emphasizing that we make no claim whatsoever that self-dissimilarity captures all that is important in complex systems. Nor do we even wish to identify self-dissimilarity with complexity. We only suggest that self-dissimilarity is an important component of complexity, one with the novel advantage that it can actually be evaluating for real-world systems.

2 Self-Dissimilarity

In the real world, one analyzes a system by first being provided information (e.g., some experimental data) in one space, and then from that information making inferences about the full system living in a broader space. The essence of our approach is to characterize a system's complexity in terms of how the inferences about that broader space differ from one another as one varies the information-gathering spaces. In other words, our approach is concerned with characterizing how readily the full system can be inferred from incomplete measurements of it. Violent swings in such inferences as one changes what is measured — large self-dissimilarity — constitute complexity for us.

2.1 Why might complex systems be self-dissimilar?

Before turning to formal definitions of self-dissimilarity we speculate on why self-dissimilarity might be an important indicator of complexity. Certainly self-dissimilar systems will be *interesting*, but why should they also coincide with what are commonly considered to be complex systems?

Most systems commonly viewed as complex/interesting have been constructed by an evolutionary process (*e.g.* life, culture, intelligence). If we assume that there is some selective advantage in such systems for maximizing the amount of information processing within the system's volume, then we are led to consider systems which are able to process information in many different ways on many spatio-temporal scales, with those different processes all communicating with one another. By exploiting different scales to run different information processing, such systems are in a certain sense maximally dense with respect to how much information processing they achieve in a given volume. Systems processing information similarly on different scales, or even worse not exploiting different scales at all, are simply inefficient in their information-processing capabilities.

To make maximal use of the different information processes at different scales, presumably there must be efficient communication between those processes. Such inter-scale communication is common in systems usually viewed as complex. For example, typically the effects of large scale occurrences (like broken bones in organisms) propagate to the smallest levels (stimulating bone cell growth) in complex systems. Similarly, slight changes at small scales (the bankruptcy of a firm, or the mutation of a gene) can have marked large-scale (industry-wide, or body-wide) effects.

Despite the clear potential benefits of multi-scale information processing, explicitly constructing a system which engages in such behavior seems to be a formidable challenge. Even specifying the necessary dynamical conditions (e.g., a Hamiltonian) for a system to be able to support multi-scale information processing appears difficult. (Tellingly, it is also difficult to explicitly construct a physical system that engages in what most researchers would consider "life-like" behavior, or one that engages in "intelligent" behavior; our hypothesis is that this is not a coincidence, but reflects the fact that such systems engage in multi-scale information processing.) In this paper, rather than try to construct systems that engage in multi-scale information processing, we merely assume that nature has stumbled upon ways to do so. Our present goal is only to determine how to recognize and quantify such multi-scale information processing in the first place, and then to measure such processing in real-world systems.

This perspective of communication between scales suggests that there are upper bounds on how self-dissimilar a viable complex system can be. Since the structure at one scale must have meaning at another scale to allow communication between the two, presumably those structures cannot be *too* different. Also for a complex system to be stable it must be robust with respect to changes in its environment. This suggests that the effects of random perturbations on a particular scale should be isolated to one or a few scales lest the full system be prone to collapse. To this extent scales must be insulated from each other. Accordingly, as a function of the noise inherent in an environment, there may be very precise and constrained ways in which scales can interact in robust systems. If so it would be hoped that when applied to real-world complex systems a self-dissimilarity measure would uncover such a modularity of multi-scale information processing.

This perspective also gives rise to some interesting conjectures concerning the concept of intelligence. It is generally agreed that any "intelligent" organism has a huge amount of extra-genetic information-processing concerning the outside world, in its brain. (If all the processing could take place directly via genome-directed mechanisms, there would be no need for an adaptive structure like a brain.) In other words, the information processing in the brain of an intelligent organism is tightly and extensively coupled to the information processing of the outside world. So to an intelligent organism, the outside world — which is physically a scale up from the organism — has the same kind of information coupling with the organism that living, complex organisms have between the various scales within their own bodies.

So what is intelligence? This perspective suggests a definition. An intelligence is a system that is coupled to the broader external world exactly as though it were a subsystem of a living body consisting of that broader world. In other words, it is a system whose relationship with the outside world is similar to its relationship with its own internal subsystems. An intelligence is a system configured so that the border of what-is-living/complex extends beyond the system, to the surrounding environment.

2.2 Advantages of the approach

The reliance on self-dissimilarity as a starting point for a science of complexity has many advantages beyond its being part of a data-driven approach. For example, puzzles like how to decide whether a system "is alive" are rendered mute under such an approach. We argue that such difficulties arise from trying to squeeze physical phenomena into pre-existing theoretical models (e.g., for models concerning "life" one must identify the atomic units of the physical system, define what is meant for them to reproduce, *etc.*). Taking our purely empiricist approach, life is instead a characteristic signature of a system's self-dissimilarity over a range of spatio-temporal scales. Presumably highly complex living systems exhibit highly detailed, large self-dissimilarity signatures, while less complex, more dead systems exhibit shallower signatures with less fine detail. We argue that life is more than a yes/no bit, and even more than a real number signifying a degree—it is an entire signature. In addition to superseding sterile semantic arguments, adopting this point of view opens entirely new fields of research. For example, one can meaningfully consider questions like how the life-signature of the biosphere changes as one species (*e.g.*, humans) takes over that biosphere.

More generally, self-dissimilarity signatures can be used to compare entirely different kinds of systems (*e.g.*, information densities in human organizations versus mass distributions in galaxies). With this complexity measure we can, in theory at least, meaningfully address questions like the following: How does a modern economy's complexity signature compare to that of the organelles inside a prokaryotic cell? What naturally occurring ecology is most like that of a modern city? Most like that of the charge densities moving across the internet? Can cultures be distinguished according to their self-dissimilarity measure? Can one reliably distinguish between different kinds of text streams, like poetry and prose, in terms of their complexity?

By concentrating on self-dissimilarity signatures we can compare systems over different regions of scales, thereby investigating how the complexity character itself changes as one varies the scale. This allows us to address questions like: For what range of scales is the associated self-dissimilarity signature of a transportation system most like the signature of the current densities inside a computer? How much is the self-dissimilarity signature of the mass density of the astronomy-scale universe like that of an ideal gas when examined on mesoscopic scales, *etc.*?

In fact, by applying the statistical technique of clustering to self-dissimilarity signatures, we should be able to create empirically-defined taxonomies ranging over broad classes of real-world systems. For example, self-dissimilarity signatures certainly will separate marine environments (where the mass density within organisms is similar to the mass density of the environment) from terrestrial environments (where the mass densities within organisms is quite different from that of their environment). One might also hope that such signatures would divide marine creatures from terrestrial ones, since the bodily processes of marine creatures observe broad commonalities not present in terrestrial creatures (and vice-versa). Certainly one would expect that such signatures could separate prokaryotes from eukaryotes, plants from animals, *etc.* In short, statistical clustering of self-dissimilarity signatures may provide a purely data-driven (rather than model-driven or — worse still — subjective) means of generating a biological taxonomy. Moreover, we can extend the set of signatures being clustered far beyond biological systems, thereby creating, in theory at least, a taxonomy of all natural phenomena. For example, not only could we cluster cultural institutions (do Eastern and Western socio-economic institutions break up into distinct clusters?); we could also cluster the signatures of such institutions together with those of insect colonies (do hives fall in the same cluster as human feudal societies, or are they more like democracies?).

The self-dissimilarity concept also leads to many interesting conjectures. For example, in the spirit of the Church-Turing thesis, one might posit that any naturally-occurring system with sufficiently complex yet non-random behavior at some scale s must have a relatively large and detailed self-dissimilarity signature at scales finer than s. If this hypothesis holds, then (for example) due to the fact that its large-scale physical behavior (i.e., the dynamics of its intelligent actions) is complex, the human mind *necessarily* has a large and detailed self-dissimilarity signature at scales smaller than that of the brain. Such a scenario suggests that the different dynamical patterns on different scales within the human brain is not some side-effect of how nature happened to solve the question of how to build an intelligence, given its constraints of noisy carbon-based life. Rather it is fundamental, being required for any (naturally occurring) intelligence. This would in turn suggest that (for example) work on artificial neural nets will have difficulty creating convincing mimics of human beings until those nets are built on several different scales at once.

3 Probabilistic Measures of Self-Dissimilarity

We begin by noting that any physical system is a realization of a stochastic process, and it is the properties of that underlying process that are fundamentally important. This leads us to consider an explicitly probabilistic setting for measuring self-dissimilarity, in which we are comparing the probability distributions over the various scale s patterns that the process can generate.

By incorporating probabilistic concerns into its foundations in this way, the proposed measure explicitly reflects the fundamental role that statistical inference (for example of patterns at one scale from patterns at another scale) plays in complexity. It also means that the framework will involve the quantities that are of direct interest physically. In addition, via information theory, it provides us with some very natural candidate measures for the amount of dissimilarity between structures at two different scales (*e.g.*, the Kullback-Leibler [10] distance between those structures). The implicit viewpoint of such measures is that "how dissimilar" two structures at different scales are is how much information is provided in the larger-scale structure that is absent in the smaller-scale structure. (The exploration of other, non-information-theoretic measures of self-dissimilarity is the subject of future research.)

To formalize the proposed measure of self-dissimilarity, we begin with a definition of a scale's "stochastic structure". Then we specify how to convert structures on different scales to the same scale by using statistical inference. As the final step, we specify how to quantify the difference between two structures on the same scale. Applied on a scale s_c to a pair of structures converted from scales s_1 and s_2, this quantity will be our measure of the self-dissimilarity exhibited by scales s_1 and s_2.

3.1 Defining the structure at a scale

Assume an integer-indexed set of spaces, Ω_s. The indices on the spaces are called *scales*. For any two scales s_1 and $s_2 > s_1$, assume also that we have a set of mappings $\{\rho^{(i)}_{s_1 \leftarrow s_2}\}$ labeled by i, each taking elements of Ω_{s_2} to elements of the smaller scale space Ω_{s_1}.

In this paper, "scales" will be akin to the widths of the translatable masking windows with which a system is examined, rather than to different levels of precision with which it is examined. The index i labeling the mapping set specifies the location of the masking window through which the system is examined (colloquially, i tells us where we are pointing our microscope). The fact that we have a full mapping *set* simply reflects the multitude of such locations.

Two elaborations of window-based scales are provided by the following two examples. Both examples involve one-dimensional sequences of characters as the objects under study

Example 1: The members of Ω_{s_2} are the sequences of s_2 successive characters. Indicate such a sequence as $\omega_{s_2}(k)$, with $1 \leq k \leq s_2$ indexing the characters. $\rho^{(i)}_{s_1 \leftarrow s_2}$ is the projective mapping taking any ω_{s_2} to the sequence of s_1 characters ω_{s_1} where $\omega_{s_1}(j) = \omega_{s_2}(j+i)$ for $1 \leq j \leq s_1$, and $0 \leq i \leq s_2 - s_1$. So the $\rho^{(i)}_{s_1 \leftarrow s_2}$ are translations of a simple masking operation creating a subsequence of s_1 characters, with i indicating the translation.

Example 2: This is a modification of example 1 so that the mapping sets and spaces Ω_s are as scale-invariant as possible, and therefore introduce minimal *a priori* bias into the self-dissimilarity measure. We require that s_1 must equal $a2^{k_1}$ and s_2 must equal $b2^{k_2}$, for some integer constants $b > a > 0$ and $k_2 \geq k_1 \geq 0$. We then have $\omega_{s_1}(j) = \omega_{s_2}(j + i2^{k_1})$, where $1 \leq j \leq a2^{k_1}$, and $0 \leq i \leq b2^{k_2 - k_1} - a$. So for example if b and a are fixed and $k_2 = k_1$, then for all (k_1-indexed) pairs of a small scale and a large scale, the kinds of of overlaps among the small scale windows appear the same, "from the perspective" of the large scale.

If we are given a probability distribution π_{s_2} over Ω_{s_2} and any single member of the mapping set $\{\rho^{(i)}_{s_1 \leftarrow s_2}\}$, we obtain an induced probability distribution over Ω_{s_1} in the usual way. Call that distribution $\rho^{(i)}_{s_1 \leftarrow s_2}(\pi_{s_2})$, or just $\pi^{(i)}_{s_1 \leftarrow s_2}$ for short. It will often be convenient to construct a quantitative synopsis of the set of all of these scale s_1 distributions. If that synopsis is a single probability distribution, then forming this synopsis puts Ω_{s_1} and Ω_{s_2} on equal footing, in that they are both associated with a single distribution. In this paper, we use the average $\rho_{s_1 \leftarrow s_2}(\pi_{s_2}) \equiv \pi_{s_1 \leftarrow s_2} \equiv \sum_i \pi^{(i)}_{s_1 \leftarrow s_2} / \sum_i 1$ as the synopsis of $\{\pi^{(i)}_{s_1 \leftarrow s_2}\}$.

We would like to be able to talk about the probabilistic structure at scale s (i.e. a distribution describing the kinds of patterns seen at scale s). This structure may characterize the statistical regularities of a single object or the regularities of an ensemble of the objects. Either way though, we would like this distribution to be independent of quantities at scales other than s.

Accordingly, we restrict attention to mapping sets such that for some fixed generating scale s_g, for any $s_1 < s_2 < s_g$, the set $\{\rho^{(k)}_{s_1 \leftarrow s_g}\}$ is the set of all compositions $\rho^{(i)}_{s_1 \leftarrow s_2}\rho^{(j)}_{s_2 \leftarrow s_g}$. We call this restriction *composability of mapping sets*. By itself, composability of mapping sets does not quite force $\rho_{s_1 \leftarrow s_g}(\pi_{s_g})$ to equal $\rho_{s_1 \leftarrow s_2}(\rho_{s_2 \leftarrow s_g}(\pi_{s_g}))$.[1] In this paper though we focus on mapping sets such that for the scales of interest $\pi_{s_1 \leftarrow s_g} \approx \rho_{s_1 \leftarrow s_2}(\rho_{s_2 \leftarrow s_g}(\pi_{s_g}))$. Under this restriction we can, with small error, just write π_s for any scale of interest s, without specifying how it is generated from π_{s_g}. For situations where this restriction holds we will say that we have (approximate) *composability of distributions*. Given such composability, we adopt the sitribution π_s as our definition of the *stochastic structure at scale s*.

Example 1 continued: Here $\pi^{(i)}_{s_1 \leftarrow s_2}(\omega_{s_1})$ is the probability that a sequence randomly sampled from Ω_{s_2} (according to π_{s_2}) will have the subsequence ω_{s_1} starting at its i'th character. So $\pi_{s_1 \leftarrow s_2}(\omega_{s_1})$ is the probability that a sequence randomly sampled from Ω_{s_2} will, when sampled starting at a random character i, have the sequence ω_{s_1}.

In this example, although we have composability of mapping sets, in general we do not have composability of distributions unless s_g/s_2 is quite large. The problem arises from edge effects due to the finite extent of Ω_{s_g}. Say $\pi_{s_g}(\omega_{s_g}) = 1$ for some particular ω_{s_g}; all other elements of Ω_{s_g} are disallowed. Then a subsequence of s_1 characters occurring only once in ω_{s_g} will occur just once in $\{\rho^{(k)}_{s_1 \leftarrow s_g}(\omega_{s_g})\}$, and accordingly is assigned the value $1/(s_g - s_1)$ by $\pi_{s_1 \leftarrow s_g}$, regardless of where it occurs in ω_{s_g}. If that subsequence arises at the end of ω_{s_g} and nowhere else it will also occur just once in the set $\{\rho^{(i)}_{s_1 \leftarrow s_2}\rho^{(j)}_{s_2 \leftarrow s_g}(\omega_{s_g})\}$. However if it occurs just once in ω_{s_g}, but away from the ends of ω_{s_g}, it will occur more than once in the set $\{\rho^{(i)}_{s_1 \leftarrow s_2}\rho^{(j)}_{s_2 \leftarrow s_g}(\omega_{s_g})\}$. Accordingly, its value under $\rho_{s_1 \leftarrow s_2}(\rho_{s_2 \leftarrow s_g}(\pi_{s_g}))$ is dependent on its position in ω_{s_g}, in contrast to its value under $\rho_{s_1 \leftarrow s_g}(\pi_{s_g})$.

Fortunately, so long as s_g/s_2 is large, we would expect that any sequence of s_1 characters in ω_{s_g} that has a significantly non-zero probability will occur many times in ω_{s_g}, and in particular will occur many times in regions far enough away from the edges of ω_{s_g} so that the edges are effectively invisible. Accordingly, we would expect that the edge effects are negligible under those conditions, and therefore that we have approximate composability of distributions.

The fact that they are generated via mappings $\rho_{s_1 \leftarrow s_g}$ and $\rho_{s_2 \leftarrow s_g}$ imposes some restrictions relating the stochastic structures π_{s_1} and π_{s_2}. Firstly, note that the mapping from the space of possible π_{s_2} to the space of possible π_{s_1} given by a particular $\rho_{s_1 \leftarrow s_2}(\cdot)$ usually will not be one-to-one. In addition, it need not be onto, i.e. there may be π_{s_1}'s that do not live in the space of possible $\pi_{s_1 \leftarrow s_2}$. In particular, consider example 1 above, where the character set is binary. Say that $s_1 = 2$. Then $\pi_{s_1}(\omega_{s_1}) = \delta_{\omega_{s_1},(0,1)}$ is not an allowed $\pi_{s_1 \leftarrow s_2}$. For such a distribution to exist in the set of possible $\pi_{s_1 \leftarrow s_2}$ would require that there be sequences ω_{s_2} for which any successive pair of bits is the sequence (0, 1). Clearly this is impossible for there must necessarily be successive pairs of bits in ω_{s_2} consisting of (1,0).

Accordingly, for any $s < s_g$, in general not all π_s are possible, due solely to the mapping set $\rho_{s \leftarrow s_g}$. Therefore for any $s_1 < s_2$, the posterior probability[2] $P(\pi_{s_2}|\pi_{s_1})$ must reflect a mapping set concerned with a scale other than s_1 or s_2, namely $\rho_{s_2 \leftarrow s_g}$. This is in addition to reflecting $\rho_{s_1 \leftarrow s_2}$, and holds even for composable distributions.

Also due to this fact that (depending on the mapping

[1] The problem is that the ratio of the number of times a particular mapping $\rho^{(k^*)}_{s_1 \leftarrow s_g}$ occurs in the set $\{\rho^{(k)}_{s_1 \leftarrow s_g}\}$, divided by the number of times it can be created by compositions $\rho^{(i)}_{s_1 \leftarrow s_2}\rho^{(j)}_{s_2 \leftarrow s_g}$, may not be the same for all k^*.

[2] $P(\pi_{s_2}|\pi_{s_1})$ is the probability of stochastic structure π_{s_2} at scale s_2 given a stochastic structure π_{s_1} at scale s_1.

set) not all π_s are possible in general, the functional form of any $P(\pi_{s_g})$ will often not be "consistent" with the associated induced functional form of $P(\pi_s) = \int d\pi_s P(\pi_{s_g})\delta(\pi_s - \rho_{s\leftarrow s_g}(\pi_{s_g}))$ (the integral is implicitly restricted to the unit s-dimensional simplex). When this happens, we cannot employ first-principles arguments to set a functional form for a prior probability distribution over structures π_s and then apply that prior to all scales s simultaneously. In particular, a $P(\pi_{s_g})$ that assigns non-zero weight to all possible π_{s_g} will not assign non-zero weight to all possible π_s in general, and in this sense the functional forms on the two scales are not consistent.

3.2 Comparison to traditional methods of scaling

It is worth taking a brief aside to discuss the numerous alternative ways one might define the structure at a particular scale. In particular, one could imagine modifying any of the several different methods that have been used for studying self-similarity. Although we plan to investigate those methods in future work, it is important to note that they often have aspects that make them appear problematic for the study of self-dissimilarity. For example, one potential approach would start by decomposing the full pattern at the largest scale into a linear combination of patterns over smaller scales, as in wavelet analysis for example. One could then measure the "weight" of the combining coefficients for each scale, to ascertain how much the various scales contribute to the full pattern. However such an approach has the difficulty that comparing the weight associated with the patterns at a pair of scales in no sense directly compares the patterns at those scales. At best, it reflects — in a non-information-theoretic sense — how much is "left over" and still needs to be explained in the small scale pattern, once the full scale pattern is taken into account.

Many of the other traditional methods for studying self-similarity rely on scale-indexed blurring functions (*e.g.* convolution functions, or even scaled and translated mother wavelets) B_s that wash out detail at scales finer than s (for example by forming convolutions of the distribution with such blurring functions). With all such approaches one compares some aspect of the pattern one gets after applying B_s to one's underlying distribution, to the pattern one gets after applying $B_{s'\neq s}$. If after appropriate rescaling those patterns are the same for all s and s' then the underlying system is self-similar.

There are certain respects shared by our approach and these alternatives. For example, usually a set of spaces $\{\rho^{(i)}_{s_1\leftarrow s_2}\Omega_{s_2}\}$ are used by those alternative approaches in defining the structure at a particular scale. (Often those spaces are translations of one another, corresponding to translations of the blurring function.)

However unlike these traditional approaches our approach makes no use of a blurring function. This is important since there are a number of difficulties with using a blurring function to characterize self-dissimilarity. One obvious problem is how to choose the blurring function, a problem that is especially vexing if one wishes to apply the same (or at least closely-related) self-dissimilarity measure to a broad range of systems, including both systems made up of symbols and systems that are numeric. Indeed, for symbolic spaces how even to define blurring functions in general is problematic. This is because the essence of a blurring function B_s is that for any point x, applying B_s reduces the pattern over a neighborhood of width s about x to a single value. There is some form of average or integration involving that blurring function that produces the pattern at the new scale — this is how information on smaller scales than s is washed out. But what general rule should one use to reduce a symbol sequence of width s to a single symbol?

More generally, even for numeric spaces, how should one deal with the statistical artifacts that arise from the fact that the probability distribution of possible values at a point x will differ before and after application of blurring at x? In traditional approaches, for numeric spaces, this issue is addressed by dividing by the variance of the distribution. But that leaves higher order moments unaccounted for, an oversight that can be crucial if one is quantifying how patterns at two different scales differ from one another.

Such artifacts reflect two dangers that should be avoided by any candidate self-dissimilarity measure:

1. The possibility of changes in the underlying statistical process that don't affect how we view the process's self-dissimilarity, but that do modify the value the candidate self-dissimilarity measure assigns to that process.

2. The possibility of changes in the underlying process that modify how we view the self-dissimilarity of the process but not the value assigned to that process by our candidate measure.

In general, unless the measure is derived in a first principles fashion directly from the concept of self-dissimilarity, we can never be sure that the measure is free of such artifacts.

Our current focus is on approaches that are based on mapping sets, and in which rather than directly compare two scale-indexed structures that live in different spaces (as in the traditional approaches), one first performs statistical inference to map the structures to the same space. There will always be the possibility of artifacts when making comparisons between systems that are different in kind (e.g., that live in non-isomorphic spaces). However properly done, an inference-based approach should at least avoid hidden statistical artifacts in comparisons between scales within a single system since the statistical aspects are explicit.[3] In particular, with such an inference-based approach there is no need for a blurring function, and the problems inherent in careless use of such functions can be avoided. Intuitively, the inference-based approach achieves this by having the information at scale s_2 be a superset of the information at any scale $s_1 < s_2$. This is clarified in the following discussion.

[3]Indeed, it may even prove possible to combine such inference-based mappings — and the associated lack of unforeseen statistical artifacts — with the structures used in the traditional approaches (e.g., blurring-based structures). This is the subject of future research.

3.3 Converting structures on different scales to the same scale

It will be convenient to introduce yet a fourth "comparison scale", s_c, at which to compare our (inferences based on) our structures. Often s_c is set in some manner by the problem at hand, and in particular, we can have $s_c = s_2$, and/or $s_c = s_g$. But this is not required by the general formulation. For the rest of this paper, we will always take $s_c \geq \max[s_1, s_2]$, where s_1 and s_2 are the two scales whose structures are being compared.

Suppose we are interested in the scale s_c structure, π_{s_c}, and are given the structure on scale s. Then via Bayes' theorem, that scale s structure fixes a posterior distribution over the elements of $\omega_{s_c} \in \Omega_{s_c}$, i.e., it fixes an estimate of the scale s_c structure:

$$\begin{aligned} P(\omega_{s_c} \mid \pi_s) &= \int d\pi_{s_c}\; P(\omega_{s_c} \mid \pi_{s_c})\, P(\pi_{s_c} \mid \pi_s) \\ &= \int d\pi_{s_c}\; \pi_{s_c}(\omega_{s_c})\, P(\pi_{s_c} \mid \pi_s) \\ &= \frac{\int d\pi_{s_c}\; \pi_{s_c}(\omega_{s_c})\, P(\pi_s \mid \pi_{s_c}) P(\pi_{s_c})}{\int d\pi'_{s_c}\; P(\pi_s \mid \pi'_{s_c})\, P(\pi'_{s_c})} \end{aligned} \tag{1}$$

where π'_{s_c} is a dummy argument π_{s_c}, and in the usual Bayesian way

$$\begin{aligned} P(\pi_{s_c}) &= \int d\pi_{s_g} P(\pi_{s_c} \mid \pi_{s_g}) P(\pi_{s_g}) \\ &= \int d\pi_{s_g} \delta(\pi_{s_c} - \rho_{s_c \leftarrow s_g} \pi_{s_g}) P(\pi_{s_g}), \end{aligned}$$

where $P(\pi_{s_g})$ is a prior over the real-valued multi-dimensional vector π_{s_g}.

The implicit model here is that π_{s_c} is formed by first sampling $P(\pi_{s_g})$ to get a π_{s_g}, and then having the mapping set $\rho_{s_c \leftarrow s_g}$ generate π_{s_c} from that π_{s_g}. Then ω_{s_c} is formed by sampling that π_{s_c}. To generate π_s, one applies the mapping $\rho_{s \leftarrow s_g}$ to π_{s_g} directly.

As an example, by composability $\pi_s = \rho_{s \leftarrow s_c} \pi_{s_c}$, and therefore

$$P(\omega_{s_c} \mid \pi_s) = \frac{\int d\pi_{s_c}\; \pi_{s_c}(\omega_{s_c})\, \delta\big(\pi_s - \rho_{s \leftarrow s_c}(\pi_{s_c})\big)\, P(\pi_{s_c})}{\int d\pi'_{s_c}\; \delta\big(\pi_s - \rho_{s \leftarrow s_c}(\pi'_{s_c})\big)\, P(\pi'_{s_c})}$$

(As always, sums replace integrals if appropriate.) In this situation, $P(\omega_{s_c} \mid \pi_s)$ may not even be an allowed distribution, in the sense that $P(\pi_{s_c})$ assigns zero probability to the distribution whose ω_{s_c}-dependence is given by $P(\omega_{s_c} \mid \pi_s)$. As an alternative decomposition, we can write

$$\begin{aligned} P(\omega_{s_c} \mid \pi_s) &= \frac{P(\pi_s \mid \omega_{s_c}) P(\omega_{s_c})}{P(\pi_s)} \\ &= \frac{P(\pi_s \mid \omega_{s_c}) \int d\pi_{s_c} \pi_{s_c}(\omega_{s_c}) P(\pi_{s_c})}{\sum_{\omega_{s_c}} \text{numerator}}\;. \end{aligned} \tag{2}$$

In practice, rather than set the prior $P(\pi_{s_c})$ and try to evaluate the integrals in equations (1) and (2), one might approximate the fully Bayesian approach of equations (1) and (2), for example via MAXENT [11], MDL [12], or by minimizing algorithmic complexity [14]. Indeed, even if we were to restrict ourselves to analyses relying on Bayes' theorem and even if $s_g \neq s_c$, we might (for example) wish to "pretend" that s_c is our generating scale, and therefore measure the dissimilarity between $P(\omega_{s_g} \mid \pi_{s_1})|_{s_g = s_c}$ and $P(\omega_{s_g} \mid \pi_{s_2})|_{s_g = s_c}$, rather than the dissimilarity between $P(\omega_{s_c} \mid \pi_{s_1})$ and $P(\omega_{s_c} \mid \pi_{s_2})$.

To allow full generality then, for each pair of scales s_2 and $s_1 < s_2$, introduce the random variable $\pi^{s_1}_{s_2}$ to indicate a distribution over Ω_{s_2} that is inferred from the structure at scale s_1. Indicate an element sampled from $\pi^{s_1}_{s_2}$ by $\omega^{s_1}_{s_2}$. Given a structure at scale s_1, π_{s_1}, we call the rule taking π_{s_1} to a distribution $\pi^{s_1}_{s_2}$ the *inference mechanism* for going from that scale-s_1 structure to a guess for the distribution at scale s_2, and indicate the action of the inference mechanism by writing $\pi^{s_1}_{s_2} = \pi^{s_1}_{s_2}(\pi_{s_1})$. As examples, equations (1) and (2) provide two formulations of a Bayesian inference mechanism.

Once we have calculated both $\pi^{s_1}_{s_c}(\pi_{s_1})$, the scale-$s_1$-inferred distribution over Ω_{s_c}, and $\pi^{s_2}_{s_c}(\pi_{s_2})$, the scale-$s_2$-inferred distribution over Ω_{s_c}, we have translated both our structures at scale s_1 and s_2 into two new structures, both of which are in the same space, Ω_{s_c}. We can now directly compare the two new structures that were generated by the structures at scales s_1 and s_2. In this way we can quantify how dissimilar the structures over s_1 and s_2 are. In this paper, we will concentrate on quantifications that can be viewed as the amount of information (concerning scale s_c) inferable from the structure at scale s_2 that goes beyond what is inferrable from the structure at scale s_1.

3.4 Comparing structures on the same scale

To define a complexity measure we must next choose a scalar-valued function Δ_{s_c} that measures a distance between probability distributions over Ω_{s_c}.[4] Intuitively, $\Delta_s(Q_s, Q'_s)$ should reflect the information-theoretic similarity between the two distributions over Ω_s given by Q_s and Q'_s. Accordingly Δ_{s_c} should satisfy some simple requirements. It is reasonable to require that for a fixed π_s, $\Delta_s(\pi_s, Q_s)$ is minimized by setting Q_s to equal π_s. Also, in some circumstances it might be appropriate to require that for any s_2, $s_1 < s_2$, π_{s_2}, and Q_{s_2}, $\Delta_{s_2}(\pi_{s_2}, Q_{s_2}) \geq \Delta_{s_1}(\rho_{s_1 \leftarrow s_2}(\pi_{s_2}), \rho_{s_1 \leftarrow s_2}(Q_{s_2}))$. In this paper we will not impose a rigid set of requirements on Δ_s, but rather as we discuss various candidate Δ_s we will note how they are related to such desiderata.

As an example $\Delta_s(Q_s, Q'_s)$ might be $|KL(\pi_s, Q_s) - KL(\pi_s, Q'_s)|$, where $KL(\cdot, \cdot)$ is the Kullback-Leibler (KL) distance [10] and π_{s_c} is the implicit true distribution over Ω_{s_c}.[5] One nice aspect of $\Delta^{KL}_{s_c}$ is that it can be viewed as a quantification of the amount of extra information concerning Ω_{s_c} that exists in Q_s but not in $Q_{s'}$. I.e., it is the amount of extra information in Q_s beyond that in $Q_{s'}$.

[4] We use the word "distance" advisedly, since we do not require that Δ_{s_c} obey the properties of a metric in general.

[5] When, as in this case, specification of π_{s_c} is needed, we should properly write $\Delta^{KL}_s(Q_s, Q_{s'}; \pi_s)$.

Consider $s_c = s_2$. In this case $\pi^{s_2}_{s_c}(\cdot)$ is the identity function and $\Delta^{KL}_{s_c}(\pi^{s_2}_{s_c}(\pi_{s_2}), \pi^{s_1}_{s_c}(\pi_{s_1}); \pi_{s_c}) = KL(\pi_{s_c}, \pi^{s_1}_{s_c}(\pi_{s_1}))$. I.e., in this scenario, $\Delta^{KL}_{s_c}$ is the KL distance between π_{s_c} and the inference for π_{s_c} based on π_{s_1}. This suggests another natural choice for $\Delta_s(Q_s, Q'_s)$, which is to set it to $KL(Q_s, Q'_s)$ always, regardless of the scale s_c distribution or of whether $s_c = s_2$. However this choice for Δ_s could be misleading if neither Q_s nor Q'_s is "well-aligned" with the true π_s; in such a case the two distributions may appear very similar according to Δ_s, but that similarity is specious. In contrast, Δ^{KL}_s forces the inference mechanisms to be "honest", as far as the resultant value of dissimilarity is concerned. In addition, $\Delta^{KL}_s(Q_s, Q'_s)$ obeys the triangle inequality, and unlike $KL(Q_s, Q'_s)$, $\Delta^{KL}_s(Q_s, Q'_s)$ is symmetric in its arguments. Unfortunately though, $\Delta^{KL}_s(Q_s, Q'_s) = 0$ does not imply that $Q_s = Q'_s$. So Δ^{KL}_s is not ideal, and there may be situations where $KL(\cdot, \cdot)$ is preferable.

4 Discussion

In this section we discuss how to estimate our self-dissimilarity measure from finite data and discuss some of the broad features of our measure.

4.1 Comparing Structures when Information is Limited

In the previous section we saw that to measure how dissimilar two structures π_{s_1} and π_{s_2} are we translate both to a distribution over the common space Ω_{s_c} and then measure how dissimilar those two distributions are. Unless we know the structures π_{s_1}, π_{s_2}, and π_{s_c} though, rather than evaluate Δ_{s_c}, we have to be content with the expected value of Δ_{s_c} conditioned on our provided information, $\mathcal{I}$. We indicate such an expectation in its full generality as follows:

$$I_{s_1,s_2;s_c}(\mathcal{I}) \equiv \int d\pi_{s_1} d\pi_{s_2} d\pi_{s_c} \Delta_{s_c}(\pi^{s_1}_{s_c}(\pi_{s_1}), \pi^{s_2}_{s_c}(\pi_{s_2}); \pi_{s_c}) \times P(\pi_{s_1}, \pi_{s_2}, \pi_{s_c} \mid \mathcal{I}) \ , \qquad (3)$$

where in turn

$$P(\pi_{s_1}, \pi_{s_2}, \pi_{s_c} \mid \mathcal{I}) = P(\pi_{s_c} \mid \pi_{s_1}, \pi_{s_2}, \mathcal{I}) \times P(\pi_{s_1}, \pi_{s_2} \mid \pi_{s_c}, \mathcal{I}) \ .$$

In this last equation, the last term on the right-hand side is the likelihood function for generating the structures at scales s_1 and s_2.

As an example, if the provided information is π_{s_1} and π_{s_2}, then we can write the expected distance as

$$I_{s_1,s_2;s_c}(\pi_{s_1}, \pi_{s_2}) = \int d\pi_{s_c} \ \Delta_{s_c}(\pi^{s_1}_{s_c}(\pi_{s_1}), \pi^{s_2}_{s_c}(\pi_{s_2}); \pi_{s_c}) \times P(\pi_{s_c} \mid \pi_{s_1}, \pi_{s_2}) \ , \qquad (4)$$

where by Bayes' theorem

$$P(\pi_{s_c} \mid \pi_{s_1}, \pi_{s_2}) \propto \delta[\pi_{s_1} - \rho_{s_1 \leftarrow s_c}(\pi_{s_c})] \delta[\pi_{s_2} - \rho_{s_2 \leftarrow s_c}(\pi_{s_c})] P(\pi_{s_c}) \ , \qquad (5)$$

with the proportionality constant set by normalization.

$I_{s_1,s_2;s_c}$ is a quantification of how dissimilar the structures at scales s_1 and s_2 are. The dissimilarity signature of a system is the upper-triangular matrix $\Delta_{s_1,s_2} = I_{s_1,s_2;s_c}$. Large matrix elements correspond to unanticipated new structure between scales.

In light of the foregoing, there are a number of restrictions we might impose on our inference mechanism, in addition to the possible restrictions on the distance measure. For example, it is reasonable to expect that for scales $i < j < k$ that $I_{i,k;s_c} \geq I_{i,j;s_c}$. Plugging in equation (4) with $\rho_{i \leftarrow k}$ set equal to $\rho_{i \leftarrow j} \rho_{j \leftarrow k}$ translates this inequality into a restriction on allowed inference mechanisms π^i_k and π^j_k. As with a full investigation of restrictions on distance measures, an investigation of restrictions on inference mechanisms is the subject of future research.

4.2 Features of the Measure

Although we are primarily interested in cases where the indices s correspond to physical scales and the Ω_s to versions of physical spaces observed on those scales, our proposed self-dissimilarity measure does not require this, especially if one allows for non-composable mapping sets. Rather our measure simply acknowledges that in the real world information is gathered in one space, and from that information inferences are made about the full system. The essence of our measure is to characterize a system's complexity in terms of how those inferences change as one varies the information-gathering spaces.

Accordingly, there are three elements involved in specifying $I_{s_1,s_2;s_c}(\pi_{s_1}, \pi_{s_2})$:

1. A set of mapping sets $\{\rho^{(i)}_{s \leftarrow s';i}\}$ relating various scales s and s', to define the "structure" at a particular scale;
2. An inference mechanism to estimate structure on one scale from a structure on another scale;
3. A measure of how alike two same-scale structures are (potentially based on a third structure on that scale).

The choice of these elements can often be made in an axiomatic manner. First, the measure in (3) can often be uniquely determined based on information theory and the issues under investigation. Next, assuming one has a prior probability distribution over the possible states of the system, then for any provided mapping set, one can combine that prior with the measure of (3) to fix the unique "Bayes-optimal" inference mechanism: The optimal inference mechanism is the one that produces the minimal expected value of the measure in (3) given the information provided by application of the mapping set. For $s_2 = s_c$, $I_{s_1,s_2;s_c}(\pi_{s_1}, \pi_{s_2}) = \Delta_{s_2}(\pi_{s_2}, \pi^{s_1}_{s_2}(\pi_{s_1}))$, and for example for the Kullback-Leibler Δ, the Bayes-optimal $\pi^{s_1}_{s_2}(\pi_{s_1})$ is $P(\omega_{s_2} \mid \pi_{s_1})$, as given in equation 1. (This solution for the Bayes-optimal inference mechanism holds for many natural choices of Δ; see the discussion on scoring and density estimation in ([13]).)

Finally, given the mapping-set-indexed Bayes-optimal inference mechanisms, and given the measure of (3), one can axiomatically choose the mapping set itself: The optimal mapping set of size K from Ω_s to $\Omega_{s' \neq s}$ is the set of K mappings that *minimizes* the expected value of the self-dissimilarity of the system. In other words, one can choose the mapping set so that the expected result of applying it to a particular Ω_s results in a distribution over $\Omega_{s'}$ that is maximally informative concerning the distribution over Ω_s, in the sense of inducing a small expected value of the measure in (3). At this point all three components of I are specified. The only input from the researcher was what issues they wish to investigate concerning the system, and their prior knowledge concerning the system.

In practice, one might not wish to pursue such a full axiomatization of (1,2,3). We view the ease with which our measure allows one to slot in portions of such an alternative non-axiomatic approach to be one of the measure's strengths. For example, one could fix (1) and (3), perhaps without much concern for *a priori* justifiability, and then choose the inference mechanism in a more axiomatic manner. In particular, if we know that the system has certain symmetries (e.g., translational invariance), then those symmetries can be made part of the inference mechanism. This would allow us to incorporate our prior knowledge concerning the system directly into our analysis of its complexity without following the fully axiomatic approach.

Another advantage of allowing various inference mechanisms is that it allows us to create more refined versions of some of the traditional measures of complexity. For example, consider a real-world scheme for estimating the algorithmic information complexity of a particular infinite real-world system. Such a scheme would involve gathering a finite amount of data about the system (e.g., data from a finite window), and then finding small Turing machines that can account for that data [14]. The size of the smallest such machine is an upper bound on the algorithmic complexity of the data. In addition, the appropriately weighted distribution of the full patterns these Turing machines would produce if allowed to run forever can be taken as a probabilistic inference for the full underlying system. Self-dissimilarity then measures how this inference for the full system varies as one gathers data in more and more refined spaces. Systems with small algorithmic complexity should be quite self-similar according to such a measure, since once a certain quality of data has been gathered, refining the data further (*i.e.*, increasing the window size) will not affect the set of minimal Turing machines that could have produced that data. Accordingly, such refining will not significantly affect the inference for the full underlying system, and therefore will result in low dissimilarity values. Conversely, algorithmically complex systems should possess large amounts of self-dissimilarity. Note also that rather than characterize a system with just a single number, as the traditional use of algorithmic complexity does, this proposed variant yields a more nuanced signature (the set $\{I_{s_i,s_j}\}$).

The self-dissimilarity measure can even be made to closely approximate traditional, blurring-function-based measures of similarity by an appropriate choice of the inference mechanism. This would be the case if for example the inference mechanism worked by estimating the fractal character of the pattern at scale s_1, and extrapolated that character upward to scales $s_2 > s_1$.

Acknowledgements We would like to thank Tony Begg, Liane Gabora, Isaac Saias, and Kevin Wheeler for insightful discussions.

References

[1] Bar-Yam, Y. *Dynamics of Complex Systems*, Addison-Wesley (1997).

[2] Bennett, C. H. *Found. Phys.*, **16**, (1986), 585.

[3] Casti, J. L., "What if", *New Scientist* **151** (1996), 36–40.

[4] Chaitin, G. *Algorithmic Information Theory*, Cambridge University Press, 1987

[5] Crutchfield, J. P., "The calculi of emergence", *Physica D*, **75** (1994), 11–54.

[6] Lloyd, S, "Physical Measures of Complexity", *1989 Lectures in Complex Systems*, (E. Jen ed), Addison-Wesley, 1990.

[7] Lloyd, S. and H. Pagels, "Complexity as thermodynamic depth", *Annals of Physics*, (1988), 186–213.

[8] Solomonoff, R. J., *Inform. Control*, **7**, (1964), 1.

[9] Stanley, M. "Scaling Behaviour in the Growth of Companies", *Nature*, **379**, (1996), 804–806.

[10] Cover, T. M., and J. A. Thomas, *Elements of information theory*, John Wiley & Sons (1991).

[11] Jaynes E. T., *Probability theory: the logic of science*, fragmentary edition available at ftp://bayes.wustl.edu/pub/Jaynes/book.probability.theory

[12] Buntine, W. "Bayesian back-propagation", *Complex Systems*, **5**, (1991), 603–643.

[13] Bernardo, and Smith, *Bayesian Theory*, John Wiley & Sons (1995).

[14] Schmidhuber, J. "Discovering solutions with low Kolmogorov complexity and high generalization ability", *The Twelfth International Conference on Machine Learning*, (Prieditis and Russel Eds.), Morgan Kauffman, 1995.