Between order and chaos: the quest for meaningful information[1⋆]

Pieter Adriaans

Department of Computer Science
University of Amsterdam,
Kruislaan 419,
1098VA Amsterdam,
The Netherlands.
pietera@science.uva.nl

Abstract. The notion of meaningful information seems to be associated with the sweet spot between order and chaos. This form of meaningfulness of information, which is primarily what science is interested in, is not captured by both Shannon information and Kolmogorov complexity. In this paper I develop a theoretical framework that can be seen as a first approximation to a study of meaningful information. In this context I introduce the notion of facticity of a data set. I discuss the relation between thermodynamics and algorithmic complexity theory in the context of this problem. I prove that under adequate measurement conditions the free energy of a system in the world is associated with the randomness deficiency of a data set with observations about this system. These insights suggest an explanation of the efficiency of human intelligence in terms of helpful distributions. Finally I give a critical discussion of Schmidhuber's views specifically his notion of low complexity art, I defend the view that artists optimize facticity instead. I suggest possibilities for empirical falsification of my views.

keywords: meaningful information, learning as compression, MDL, two-part code optimization, randomness deficiency, thermodynamics, free energy, algorithmic esthetics.

1 Introduction: Learning, compression and meaningful information

Since pre-socratic philosophy there has been a tension between a description of the world as a dynamic process (Heraclitus) or as a static structure (Parmenides. Plato's theory of ideas explains the chaotic complexity of the world around us in terms of an imperfect reflection of perfect immutable ideal forms. We can

⋆ This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

know and understand the world because our mind participates in this world of ideas. Learning to understand the world is in fact a kind of remembering what one already knows. Later philosophers like William of Occam threw the world of ideas in the dustbin ("entia non sunt multiplicands praetor necessitated", or "entities should not be multiplied beyond necessity") in favor of the nominalistic view that our descriptions of the world should be as simple as possible. This principle, often referred to as Occam's razor (to cut off Plato's beard of ideas), has had a decisive influence in the history of science. In modern methodology of science this notion is studied under various guises: Occam's razor [14], the minimal description length (MDL) principle [6; 16], two-part-code optimization [30], learning as data compression [32] etc. All these approaches are indebted to the formulation of an algorithmic solution to the problem of induction by Solomonoff [29], Chaitin [7] and Kolmogorov [31], which is one of the greater achievements of science in the 20th century.

In its modern guise this research often goes hand in hand with a computationalistic conception of the human mind as a kind of general problem solver. This conception can, via the influence of Carnap, also be traced back to the empiricist psychology of the mind of Locke and Hume [20; 18]. Solomonoff's solution to the induction problem is associated with the concept of Kolmogorov complexity as a measure of the amount of information in a binary object. Roughly the Kolmogorov complexity of a binary string is the length of the shortest prefix-free program that computes this object on a universal Turing machine. This insight allows us to formulate the notion of a universal distribution that assigns an a-priori probability to an object that is inversely logarithmic in its Kolmogorov complexity. Especially Solomonoff, who was the first to formulate the idea of a universal distribution, seems to have been driven by an ambition to solve the general problem of mathematical induction on one hand and formulate a general theory of optimal human learning based on evolution on the other: *My general conclusion was that Bayes' theorem was likely to be the key. That a person was born with a reasonably good builtin a priori probability distribution. The person would then make predictions and decisions based on this distribution. The distribution was then modified by their life experience. The initial "Builtin" distribution was obtained by organic evolution. There was a strong selection in favor of organisms that made decisions on the basis of "good" a priori probability distributions. The organisms making poor decisions would tend to have fewer descendants.* [29] This research program seems to be the driving force behind the work of researchers like Schmidhuber [27] and Hutter [19]. For a discussion of compressibility as a general cognitive principle see [8].

Occam's razor has been questioned throughout history with fierce opponents (e.g. [14]) and strong defenders (e.g. [32]). Until recently the view of learning as algorithmic data compression did not seem to have much practical value. Lots of learning algorithms in fact perform some kind of data compression, but this was not a guiding principle of their design [23; 12]. Two developments in the last five years have changed this perspective quite fundamentally : 1) a better understanding of the mathematics behind compression, specifically Kolmogorovs

structure function [30; 31] and 2) the application of existing implementations of compression algorithms to approximate the ideal (and uncomputable) Kolmogorov complexity as pioneered by Cilibrasi and Vitányi [9; 10].

1.1 A thermodynamic interpretation of Solomonoff's program

At this moment we have not only a much better understanding of the theoretical issues behind data compression. It has also become clear that MDL as a universal inductive methodology has flaws. Grünwald and Langford have identified conditions under which MDL behaves suboptimal [16]. Adriaans and Vitanyi showed that, although an optimal compression of a data set produces in a certain sense an optimal theory, this does not imply that *incremental* compression of data sets, such as most learning algorithms perform, is a generally valid strategy [4]. The quality of our predicitive models may vary indefinitely with each incremental compression step we make. Because of the uncomputatibility of the optimal compression we can never be sure to have reached a good theory in any finite time. In a purely algorithmic universe MDL actually would not be a very good strategy. The fact that bounded resource data compression 'works' in our universe has to do with its specific physical structure. Consequently there can not be a pure algorithmic explanation of the validity MDL. The extremely efficient data compression that the human mind is able to perform seems to be driven by bias that are not purely mathematical. In this context the 'built-in' a priori distribution that was referred to in the citation of Solomonoff above could be updated in our theoretical models along the following lines: "*We are intelligent agents that have evolved via a process of evolution in a universe that has the following structure*:

1. *It is spatio temporal.*
2. *It is subject to elementary physical laws. In particular it obeys the laws of thermodynamics. It has an irreversible arrow of time that is associated with a continuous increase in entropy.*
3. *It supports the spontaneous emergence of universal computational processes [33]. Since the capacity to store information presupposes the existence of reversible processes (bit-flips) and since recursive functions discard information this implies that it contains systems that can sustain thermodynamic non-equilibrium states during a certain time.*
4. *It supports various functions for the distribution of information through space: light (vision), mechanical interaction (touch, hearing) and chemical interaction (smell, taste). These information distribution functions act as 'lossy' homomorphisms that only convey partial information. In general the information decays at least polynomially with the distance in space.*

In the context of evolution we may expect our sensory organs and general problem solving capabilities to be optimized for these conditions. In particular one would expect agents emerging in these conditions to have advanced capabilities to evaluate spatial variations in entropy. Since systems increase their entropy

over time, places with low entropy are naturally 'interesting' and may create life sustaining conditions. Also the fact that such agents could emerge in an evolutionary process presupposes the environment to be benign in the following sense: the lossy information distribution functions convey enough information to survive. This implies that detection of entropy variations that are preserved under lossy compression (i.e. general detection of density variations) is sufficient for survival.[1]

This thermodynamic variant of Solomonoff's program moves us away from a more radical interpretation of his work implying a computationalistic view of the world, i.e. the metaphysical theory that the world essentially is a computational process and that the human mind is a universal computer. The connection is as follows: the application of the universal distribution to a data set seems to imply that we regard this data set as the result of a computational process. If we interpret the human mind as a general problem solving device that is the result of an evolutionary process then it is natural to suppose that it is optimized for data sets that are produced by computational processes, i.e. it evolved in a world that is itself computational. Computationalistic ideas have been defended by a variety of authors like Wolfram [33], Schmidhuber, Lloyd [22], Floridi and Zuse: "*The entire universe is being computed on a computer, possibly a cellular automaton.*"[2] It is clear that this form of computationalism is a purely metaphysical position which can not be verified at best, but which prima facie is at variance with plain observations we can make in everyday life: e.g. although the laws of gravity can be described in terms of simple mathematical regularities there is nothing that suggests that gravity is itself a computational process. Metaphysical computationalism therefore should be rejected as unscientific. Furthermore, given the flaws of MDL discussed above, it is difficult to defend the idea that the human mind evolved as a purely algorithmic compression based problem solver.

The rejection of computationalism implies a view of computational models of processes in the world as *phenomenological*: i.e. they describe processes in the world without any presupposition about their ontological status. An explanation of fact that the world at different levels of aggregation and over different phase transitions can be described by simple high level mathematical equations remains one of the great challenges of science. Assuming that the world is essentially a computational process will not bring this issue any closer to a solution.

[1] This last condition seems to rule out exactly those data sets that given the results of Adriaans and Vitanyi [4] could bring a general compression based bounded problem solver in to trouble. It is a well known principle in information theory that if a set of messages has systematic density variations it does not have maximal entropy. A environment is benign if the opposite condition also holds: If a data set is compressible it has density variations. This condition rules out the malicious demon that presents data sets that are apparently random, but in fact can be compressed substantially, e.g. decimal expansions of the number π. Such data sets indeed seem to be sufficiently rare in our universe such that a failure to recognize them in general does not create life threatening risks. Of course they still do occur in nature.

[2] Konrad Zuse, as he referred to this as "Rechnender Raum (Zuse 1967, 1982).

1.2 Meaningful information

There is a connection with the notion of meaningful information. Formal definitions of information like those of Shannon and Kolmogorov do measure information in data sets but they do not capture the notion of meaningful information. This is immediately clear when we note that the most information rich radio transmission we could send is pure noise. Any station following this strategy would soon loose its audience. Data sets with maximum entropy are not considered to be interesting by human beings: such sets are rich in information but they contain no meaningful information. On the other hand a transmission of pure silence would also not be considered to be very informative. They contain no information at all. Meaningful information seems to exist in the 'sweet spot' between order and chaos.

In this paper I associate meaningfulness with facticity, but this is no doubt only a crude approximation. In general science, in the study of human cognition and even in art we seem to have an interest in systems that have a complexity between order and chaos, between boredom and noise. The 'interestingness' of these data sets is related to compressibility ([11], [13]). The thermodynamic explanation for this seems to be the fact that, in a universe in which entropy naturally increases over time, systems that maintain a low entropy over a period of time are 'by definition' interesting. Compressibility is associated with structure, with self-organization and with the principles of life itself.

It is important to distinguish this question from the related ambition of researchers that are interested in formulating a theory of optimal learners based on Kolmogorov complexity. Schmidhuber even has formulated a theory of algorithmic aesthetics and low complexity art along these lines [26]. Recently he introduced a notion of interestingness as *the first derivative of subjective compressibility* [28]. This theory deals with a subjective notion of interestingness *at a certain time for a certain agent.* Facticity on the other hand is an a priori quality of data sets, i.e. products of the human mind. As such it leads to predictions that can in principle be verified empirically given the present state of technology. Since I am also interested in a theory of algorithmic esthetics I will present a critical discussion of the ideas of Schmidhuber in a separate paragraph at the end of this paper.[3]

In the context of this paper I am not so much interested in the definition of *an optimal problem solver* but in the question *why the universe produces data sets from which anything can be learned at all.* Why does the universe act as a cooperative teacher? Why do we live in a universe in which MDL is a valuable methodological principle? The reason for this shift in direction is the insight that the study of algorithmic strategies for problem solving, as such, do not explain the efficiency with which we solve problems. Theories about algorithmically optimal problems solvers give an interesting framework for the transcendental analysis of learning but in order to explain the efficiency of learning an analysis of additional bias is necessary. This paper does a first step in this direction by

[3] The ideas on a dialectics of facticity and art were presented in my Paradiso lecture at the beginning of 2007.

analyzing bias that stem from thermodynamics. This shift is not in conflict with Solomonoff's research program but more or less orthogonal to it. Surprisingly, from a philosophical point of view, this change of direction is associated with a shift from an empiricist tabula rasa position to a more Cartesian/Kantian view in which a learning agent shares bias with the world in which it is embedded. This should be interpreted not so much as innate ideas, but as the theory that an agent inherits distributions from the world from which it originates. This is fully compatible with the observation cited above of Solomonoff that human beings are: "born with a reasonably good builtin a priori probability distribution."

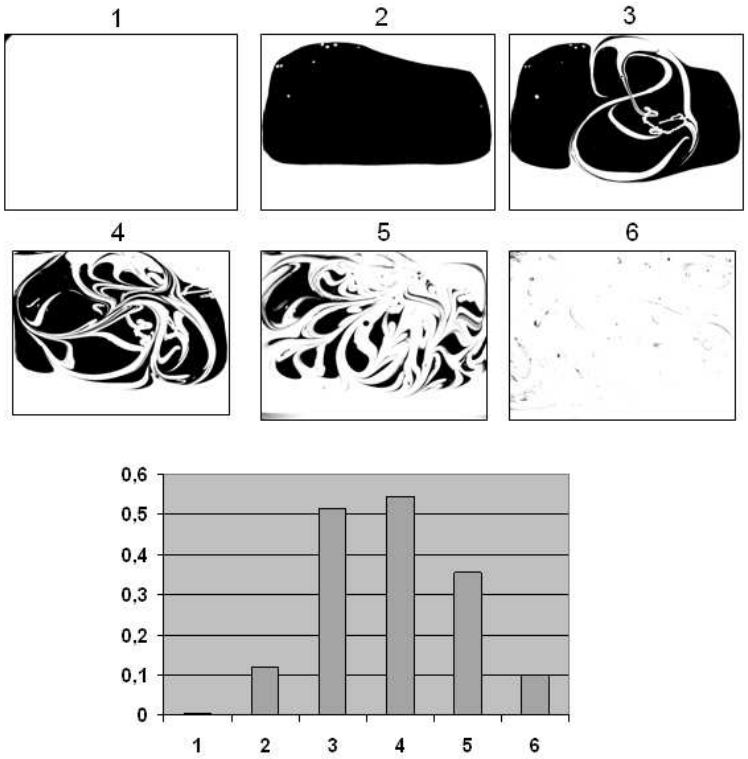

Fig. 1. Facticity scores for mixing black and white paint. The facticity of a data x is the product (times 4) of the normalized entropy $C(x)/C_{max}(x)$ and the normalized randomness deficiency $(C_{max}(x)-C(x))/C_{max}(x)$. Configuration 4 has the best balance between order and chaos and thus would be the most 'interesting' one. The scores have been calculated using JPEG, followed by RAR compression. Maximal entropy $C_{max}(x)$ has been approximated by adding 400 % noise to the images. The standard entropy $C(x)$ is approximated by the file size after compression. Note that the resolution of the camera influences the measurements. The addition of hard pixel noise creates a random image that the camera never could capture. This is the reason that none of the pictures reach the maximal facticity of 1.

1.3 An experiment

Here is an experiment. Take a cup of coffee and pour some cream in it (See Figure 1). Take a picture of it with your digital camera. In the beginning the cream will be just an uninteresting blob. Stir slowly and make pictures of various stages that have nice patterns. Continue until the cream has dissolved and your cup has an even brown color. Drink the coffee, then look at the file size of the different pictures.

If your camera uses an adequate compression algorithm you will find that the file size has increased up to a certain point and then decreases. The compression algorithm of your camera reflects the complexity of the data set until the moment that the complexity has reached a global equilibrium and is beyond its resolution. In this experiment we have a system that evolves in time, the cup of coffee, and a data set of observations, the pictures. The crux of this experiment is that the size of the individual pictures somehow reflects the 'interestingness' of the system. In the beginning there is a lot of order in the system. This is not very interesting. In the end there is an equilibrium that also has little cognitive appeal.

In the following it will prove useful to describe these compressions in terms of a so-called two-part-code: a description of a general class of sets, the *model code* and an element or a set of elements of this set, the *data-to-model-code* [30; 31].

Let me give some examples:

- **Symmetry**. This is one of the most fundamental ordering principles in nature. Most living creatures have symmetry: plants, trees, predator, prey. If a data set has symmetry it means that we only have to describe half of it (the data-to-model-code) plus some information about the nature of the symmetry of constant length (the model-code). In the limit such a data set can be compressed to at least half its size. In terms of generating languages symmetry is context free: a symmetric data set can be produced by a simple memoryless central process. Discovering symmetry in a data set can be seen as a very simple learning problem. It can easily be discovered in linear time. For some thoughts on symmetry and entropy see [21].
- **Repetition**. In order to describe a repeating pattern I only have to give a description of the generating pattern (the data-to-model-code) and some information about the way the pattern repeats itself (the model-code). Repetition is more complex than symmetry in the sense that it presupposes a generating process with a memory. In terms of languages: repetition is context sensitive. Finding repeating patterns in a data set is also a basic learning problem that can be solved in time $n \log n$ [2].
- **Grammar**. A corpus of a language could be described in terms of the grammar G (the model-code) of the language and a set of indexes corresponding to an enumeration of the sentences in the corpus (the data-to-model-code). If the size of the corpus is large enough in relation to the size of the grammar G then this description in terms of two will be shorter than an extensional description of the sentences in the corpus. Finding this description is a well studied learning problem. If the language is regular, then the task of ap-

proximating the smallest DFA consistent with a set of sentences is NP-hard [24; 1].

- **Program**. We could ask ourselves, given a certain data set: what would be the shortest program generating this data set in a certain programming language, or, even more general, we could try to find the shortest combination of a Turing machine T_i (the model-code) and a program P (the data-to-model-code). In a sense this would be, from a computational point of view, the ultimate compression possible and the Turing machine T_i would be the ultimate 'explanation' of the data set. Needless to say that because of the Halting problem there is no algorithm that will construct this ultimate compression for us. The problem is undecidable. Still, conditional to the programming language we choose, the notion of the shortest program generating a certain data set is well defined. Kolmogorov complexity studies these optimal compressions from the perspective of universal Turing machines [31].

Here I have described four classes of learning problems (varying from very easy, via NP-hard, to undecidable) as compression problems where the task is to find a two-part code compression for a data set. Apparently there is a deep connection between data compression and learning.

2 Learning and Thermodynamics

Let us redefine the problem of learning as a general problem of induction. Suppose we study some universe Υ that contains a certain system Σ. In principle Σ could be anything: the human brain, the living cell, a black hole, the weather. For the moment we will suppose that Σ is an isolated physical system that exists in space and time. The problem of induction now takes the following form: can we develop a description of Σ that: 1) *explains its structure* 2) *predicts its behavior*? Behind these issues there is still a deeper problem. Note that by denoting S as a system we have already made a hermeneutic jump. By considering Σ as a system we have decided that it is interesting. The question is: can we give a formal description of this notion of interestingness. This last question cannot be answered by means of an analysis of the formal complexity of Σ alone. In order to understand these questions we must look at the physical background and specifically at the theory of thermodynamics [4]. The first law of thermodynamics describes the change of internal energy U of a system in terms of the difference between the amount of heat Q absorbed by the system and the amount of work W done by the system:

$$dU = \bar{d}Q - \bar{d}W \tag{1}$$

The second law of thermodynamics states that a change of entropy of any system is directly related to a change in the amount of heat absorbed by the system, and inversely related to the absolute temperature T. Moreover the entropy never decreases in time:

$$dS = \frac{\bar{d}Q}{T}, \qquad \frac{dS}{dt} \geq 0. \tag{2}$$

[4] For a discussion of the relation between physics and information see [5]

An important notion for our research is that of *free energy*:

$$F \equiv U - TS \tag{3}$$

The free energy is associated with the amount of energy in the system that is free to do work. If a system is in a state of thermal equilibrium then the free energy is minimal and the entropy is maximal. In a gas the total entropy in equilibrium is given by:

$$S = -\sum_i p_i \log p_i \tag{4}$$

where p_i are the individual probabilities of the velocities of the particles. In the limiting case where all probabilities are equal $p_i = p = 1/w$ we get:

$$S = \ln w. \tag{5}$$

This is the formula that Boltzmann had engraved on his tombstone. It tells us that in a state of maximal equilibrium the entropy is the log of the number of accessible states.

What should we conclude from this analysis in the context of learning? Note that for a closed system in thermodynamic equilibrium macroscopically measurable quantities do not vary over time. This means that there is very little that we can learn about a system in thermodynamic equilibrium. Such systems do not have an internal structure and they do not have an interesting history. Consequently learnability is associated with non-equilibrium states of systems. Here is one possible objective answer to the question what distinguishes a system from its environment. Separate systems are those parts of the world that maintain en entropy that is different from their environment during a certain period of time. Consequently learnable systems are associated with variation in entropy. This implies no maximal entropy and thus an amount of free energy larger than zero. Self-organization is typically associated with systems that maintain an entropy that is different from the environment for a certain period of time. A world that is in a state of thermal equilibrium does not contain any meaningful information, has no structure, no interesting development and no free energy.

3 Kolmogorov complexity

Now we turn our attention to Kolmogorov complexity as a theory about optimal complexity of data sets. Let $x, y, z \in \mathcal{N}$, where $\mathcal{N}$ denotes the natural numbers and we identify $\mathcal{N}$ and $\{0,1\}^*$ according to the correspondence

$$(0, \epsilon), (1, 0), (2, 1), (3, 00), (4, 01), \ldots$$

Here ϵ denotes the *empty word*. The *length* $|x|$ of x is the number of bits in the binary string x, not to be confused with the *cardinality* $|S|$ of a finite set S. For example, $|010| = 3$ and $|\epsilon| = 0$, while $|\{0,1\}^n| = 2^n$ and $|\emptyset| = 0$. The emphasis is on binary sequences only for convenience; observations in any alphabet can be

encoded in a 'theory neutral' way. Below we will use the natural numbers and the binary strings interchangeably. In the rest of the paper we will interpret the set of models $\mathcal{M}$ in the following way:

Definition 1. *Given the correspondence between natural numbers and binary strings, $\mathcal{M}$ consists of an enumeration of all possible self-delimiting programs for a preselected arbitrary universal Turing machine U.*[5] *Let x be an arbitrary bit string. The shortest program that produces x on U is $x^* = argmin_{M \in \mathcal{M}}(U(M) = x)$ and the Kolmogorov complexity of x is $C(x) = |x^*|$. The conditional Kolmogorov complexity of a string x given a string y is $C(x|y)$, this can be interpreted as the length of a program for x given input y. A string is defined to be random if $C(x) \geq |x|$.*

This makes $\mathcal{M}$ one of the most general model classes with a number of very desirable properties: it is universal since all possible programs are enumerated, because the programs are self-delimiting we can concatenate programs at will, in order to create complex objects out of simple ones we can define an a-priori complexity and probability for binary strings. There are also some less desirable properties: $C(x)$ cannot be computed (but it can be approximated) and $C(x)$ is asymptotic, i.e. since it is defined relative to an arbitrary Turing machine U it makes less sense for objects of a size that is close to the size of the definition of U. Details can be checked in [31]. We have:

$$argmin_{M \in \mathcal{M}} - \log P(M) - \log P(D|M) =$$

$$argmin_{M \in \mathcal{M}} C(M) + C(D|M) = M_{MDL} \tag{6}$$

Under this interpretation of $\mathcal{M}$, the length of the optimal code for an object is equivalent to its Kolmogorov complexity.

In this paper I will often use the notions of *typicality* and *incompressibility* of elements of a set, e.g. in those cases where I state that the vast majority of elements of a set have a certain quality. This might at first sight sound a bit inaccurate. To show that this notion actually has an exact definition I give the following theorem due to Li and Vitányi [31] pg. 109):

Theorem 1. *Let c be a positive integer. For each fixed y, every finite set A of cardinality m has at least $m(1 - 2^{-c}) + 1$ elements x with $C(x|y) \geq \log m - c$.*

Proof: The number of programs of length less than $\log m - c$ is

$$\sum_{i=0}^{\log m - c - 1} 2^i = 2^{log m - c} - 1$$

Hence, there are at least $m - m2^{-c} + 1$ elements in A that have no program of length less than $\log m - c$.

[5] Here the notational conventions of two disciplines clash. U is the internal energy of a system $U(x)$ is the Universal Turing machine with input x. Which interpretation is meant should be clear from the context.

This shows that in the limit the number of elements of a set that have low Kolmogorov complexity is a vanishing fraction. In the limit a typical element of a set is a random element. In general the vast majority of elements of a set is not compressible. One of the problems with Kolmogorov complexity is that it specifies the length of a program but tells us nothing about the time complexity of the computation involved. Therefore Kolmogorov complexity can not be used directly to prove lower bounds for the time complexity of problems.

3.1 Randomness deficiency

It is important to note that objects that are non-random are very rare. To make this more specific: in the limit the density of compressible strings x in the set $\{0,1\}^{\leq k}$ for which we have $C(x) < |x|$ is zero [31]. The overwhelming majority of strings is random. In different words: an element is *typical* for a data set if and only if it is *random* in this data set. In yet different words: if it has maximal entropy in the data set. This insight allows us to formulate a theory independent measure for the quality of models: *randomness deficiency*.

We start by giving some estimates for upper-bounds of conditional complexity. Let $x \in M$ be a string in a finite model M then

$$C(x|M) \leq \log |M| + O(1) \tag{7}$$

i.e. if we know the set M then we only have to specify an index of size $\log |M|$ to identify x in M. Consequently:

$$C(x) \leq C(M) + \log |M| + O(1) \tag{8}$$

The factor $O(1)$ is needed for additional information to reconstruct x from M and the index. Its importance is thus limited for larger data sets. These definitions motivate the famous Kolmogorov structure function:

$$h_x(\alpha) = \min_S \{\log |S| : x \in S, C(S) \leq \alpha\} \tag{9}$$

Here α limits the complexity of the model class S that we construct in order to 'explain' an object x that is identified by an index in S. [6] Let $D \subseteq M$ be a subset of a finite model M. We specify $d = |D|$ and $m = |M|$. Now we have:

$$C(D|M,d) \leq \log \binom{m}{d} + O(1) \tag{10}$$

Here the term $\binom{m}{d}$ specifies the size of the class of possible selections of d elements out of a set of m elements. The term $\log \binom{m}{d}$ gives the length of an index for this set. If we know M and d then this index allows us to reconstruct D.

A crucial insight is that the inequalities 7 and 10 become 'close' to equalities when respectively x and D are *typical* for M, i.e. when they are random in M.

[6] This α could be seen as a factor that limits the resolution of the camera in figure 1.

This typicality can be interpreted as a measure for the goodness of fit of the model M. A model M for a data set D is optimal if D is random in M, i.e. the randomness deficiency of D in M is minimal. The following definitions formulate this intuition. The *randomness deficiency* of D in M is defined by:

$$\delta(D|M,d) = \log \binom{m}{d} - C(D|M,d), \tag{11}$$

for $D \subseteq M$, and ∞ otherwise. If the randomness deficiency is close to 0, then there are no simple special properties that single D out from the majority of data samples to be drawn from M.

The *minimal randomness deficiency* function is

$$\beta_x(\alpha) = \beta_D(\alpha) = \min_M \{\delta(D|M) : M \supseteq D,\ C(M) \leq \alpha\}, \tag{12}$$

If the randomness deficiency is minimal then the data set is typical for the theory and, with high probability, future data sets will share the same characteristics, i.e. minimal randomness deficiency is also a good measure for the future performance of models. For a formal proof of this intuition, see [30].

3.2 Kolmogorov complexity meets thermodynamics

In this paragraph we analyze the following central theorem that relates the free energy of a system with the randomness deficiency of the data set resulting from observations of the system:

Theorem 2. *For a static measurement $h_t : \Sigma \to \{0,1\}^*$ at moment t of a dynamic system Σ with free energy F we have:*

$$h_t(F) = \delta(x).$$

Here $\delta(x)$ is the randomness deficiency of the data set x represented as a string.

Proof: Consider the following thought experiment. We collect a number of observations (measurements) of a system Σ at a certain time t. A paradigmatic example could be a single photo of a dynamic system that evolves in time, say a satellite picture of a hurricane. Such a picture could be seen as a homomorphism from a system Σ to a data set x. Call the system Σ and the data set that results from the measurements D represented as a binary string x. Since the information in the measurements is supposed to be representative, the information in the string x reflects the thermodynamic state of the system *frozen in time*. Suppose we want to define such a homomorphism h_t that depicts the structure of Σ on x at moment t. Since we are abstracting from the dimension of time we can interpret definition 3 as time neutral, much in the same way that we can not deduce the speed of a car from its picture. This means that we can interpret T to be the unit temperature of the new system. The definition 3 for free energy would be transformed in the following way

$$h_t(F) = h_t(U - TS) =$$

$$h_t(U) - h_t(T)h_t(S) =$$
$$h_t(U) - h_t(S)$$

The other units get an interpretation along the following lines. If Σ is in equilibrium we expect x to be random. According to 5 the maximal entropy of x would be the set of all accessible states. This is for a binary string $2^{|x|}$, which gives $\log 2^{|x|} = |x|$ as its maximal Boltzmann entropy. This is in line with the Kolmogorov estimate of the maximum complexity of a random string where $C(x) \geq |x|$. Conversely if x is compressible and there are no measurement distortions we may (because of h_t) conclude with high probability that Σ is not in thermal equilibrium. We may see the Kolmogorov complexity as an a priori entropy measure of a binary string, i.e. $h_t(S) = C(x)$. Clearly the maximal entropy of a string x is reached when $C(x) \geq |x|$, i.e. $h_t(U) = |x|$:

$$h_t(F) = |x| - C(x)$$

Note that $|x| - C(x)$ is the randomness deficiency of the string x. Consequently the Helmholz free energy $U - TS$ of the system Σ is under the homomorphism h_t transformed in to the randomness deficiency $\delta(x)$ of x:

$$h_t(F) = \delta(x)$$

This concludes the proof of the theorem. If we collect a set of adequate measurements of a system at time t we may say that the compressibility or randomness deficiency of the resulting data set reflects the free energy of the system. If the data set is compressible then the system contains free energy. In that case it is not in thermodynamic equilibrium and capable of performing work. One might call theorem 2 the fundamental learnability theorem for physical systems. It shows how learning as data compression and thermodynamics interact. Data compression identifies systems that are not in thermal equilibrium: i.e. systems with structure, systems with self organization, living systems etc.

Consider the following simple example. Below we have 4 data sets. Data-set-1 is visually associated with a non-eqilibrium state, Data-set-2 is equivalent to an equilibrium state. The last two sets are the result of applying a 3 to 1 bit homomorfism to these data sets.

Data-set-1 Non-equilibrium: A 57x8 binary space with 40 bits densely packed
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000
10101101101111000000000000000000000000000000000000000000
11110111011101000000000000000000000000000000000000000000
11011010111101000000000000000000000000000000000000000000
10011101101110000000000000000000000000000000000000000000

Data-set-2 Equilibrium: A 57x8 binary space with 40 bits randomly distributed

00100001001000000000100000000000010000000000100000000000
00001010000000000000010000000100000001000000000000010000000
01000000000000000000000010000000000000100000000010100000
00000010000000001000000000100000000000000100000000100000
00100000000100000000000100000000001000000010000000001000
00000000000000000000010000000000000100010000000000000100
00001000000000000000000000010000000000000000000000000000000
00000010000000000001000001000000000010000000010010000000

Data-set-1' the result of a 3 to 1 density homomorfism

0000000000000000000
0000000000000000000
0000000000000000000
0000000000000000000
1111100000000000000
1111000000000000000
1111000000000000000
0111100000000000000

Data-set-2' the result of a 3 to 1 density homomorfism

0000000000000000000
0000000000000000000
0000000000000000000
0000000000000000000
0000000000000000000
0000000000000000000
0000000000000000000
0000000000000000000

Figure 2 shows the effects of applying Rar compression to these sets. Direct Rar compression gives an increase in size for Data-set-2 versus Data-set-1, however, after the application of a lossy density homomorfism (3 bits to 1) the image Data-set-2' is smaller then Data-set-1', i.e. the complexity of Data-set-2 is not preserved under simple density homomorfisms. This illustrates the fact that phase transitions that are normally associated with high complexity (in the order of Avogadro's number) in thermodynamic systems already occur under simple conditions in discrete systems and that they can be measured using state of the art commercial data compression routines. Given the fact that information distribution functions in our universe are of this lossy type it is to be expected that agents that evolve in such an environment ignore the complexity in equilibrium data sets.

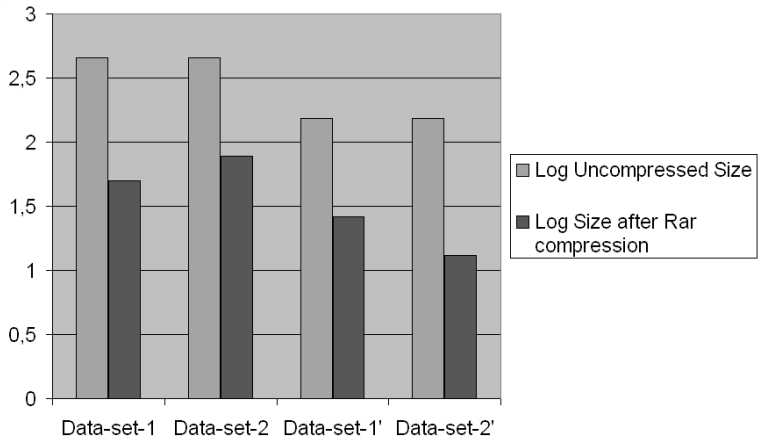

Fig. 2. The effect of data compression under density homomorfisms for very simple equilibrium and non-equilibrium data sets (456 bytes).

4 Facticity

This analysis shows that entropy and Kolmogorov complexity not necessarily measure the *interestingness* of a system of a data set. All systems in the universe will eventually reach a state of maximal entropy. A system in maximal entropy has played its part and has no interesting structure. Likewise, although a random string x contains in a way the maximum amount of information possible for a string of length $|x|$, it contains without any context no meaningful information. We can not expect to learn very much about a system that is in a state of thermodynamic equilibrium. On the other hand a string with low complexity does not contain very much information and thus by definition it does not contain much meaningful information. Interestingness or meaningfulness of a data set seems to be lying in a tension between chaos and structure. As a first approximation of this notion I will define the idea of the facticity of a data set. The facticity of a binary string will be maximal if $C(x) = 1/2|x|$. The maximum amount of meaningful information can be measured in terms of the what I call the normalized facticity of a string. It is the product of the normalized entropy $C(x)/C_{max}(x)$ and the normalized randomness deficiency $(C_{max}(x) - C(x))/C_{max}(x)$. For strings this is:

$$\varphi(x) = 4\frac{C(x)}{|x|} \times \frac{|x| - C(x)}{|x|} \tag{13}$$

The factor 4 serves to secure a maximum facticity of 1. Facticity can be seen as a normalized information density measure. For thermodynamic systems this equation is transformed in to:

$$\varphi(\Sigma) = 4\frac{S}{S_{max}} \times \frac{S_{max} - S}{S_{max}} \tag{14}$$

This is the rationale behind the experiment represented in figure 1. Here I have taken pictures of the process of mixing black and white paint. I use the facticity score to select the most *interesting* picture.

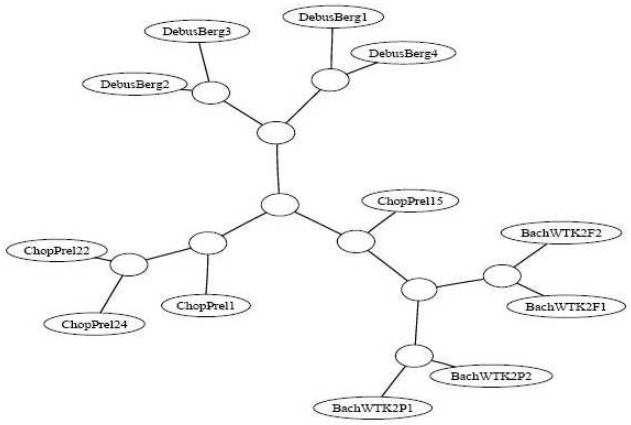

Fig. 3. A tree representation based on the normalized compression distance between 12 Piano pieces.

The fact that state of the art data compression routines can be used to make predictions about data sets that seem to have cognitive relevance was recently discovered by Vitányi and Cilibrasi [9]. Suppose that x and y are data sets and that we have a concatenation operation on these sets that allows us to form xy. Let C be a general compression routine such that $C(x)$ is the length in bits of data set x when compressed by C. We can now define the related *Normalized Compression Distance* (NCD):

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}} \tag{15}$$

Figure 3 shows that NCD seems to be able to identify style connections between different piano pieces. NCD seems to work well for data sets that have a natural linear representation such as music and language. For images it seems to work less well due to the fact that we do not have good general purpose compression algorithms for higher dimensional data sets.

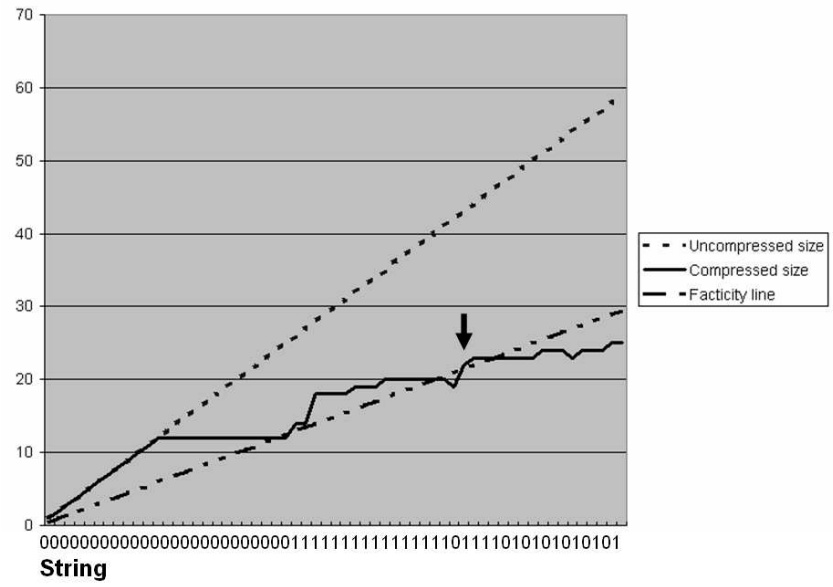

Fig. 4. Example of a simple rule breaking process that is not factic, i.e. it is not complex enough to sustain facticity. We start with a zero. At each move the rule breaking routine checks the facticity of the string using Rar compression. If the string is too complex that last symbol is repeated. If the complexity is too low the symbol of the sequence is changed from 0 to 1 or vice versa. At the arrow this rule breaking routine ends in a loop of 01 sequences that can easily be compressed by Rar and thus continues indefinitely. This is the fate of any recursive rule breaking routine that is not 'clever' enough to outsmart the compressor. Recursion is not creative.

4.1 Factic processes and factic data sets

The facticity is optimal if the balance between order and chaos is optimal. Facticity is partly motivated by insights from thermodynamics, but also be introduced via other constructions. Facticity can be seen as a rule breaking concept. Functions that follow and break rules with some regularity create data sets with high facticity. Suppose we want to construct a binary string of k bits with maximum facticity, i.e. $C(x) = k/2$. For any k of sufficient size, strings with near optimal facticity exist in abundancy: just concatenate a low complexity string of length of ca. $k/2 + \log k/2 + O(1)$ to a random string of length ca. $k/2 - \log k/2 - O(1)$, where the term $\log k/2$ serves to code the length and $O(1)$ serves to concatenate the first part to the second part. This gives at least $2^{k/2-\log k/2-O(1)}$ strings with basic near optimal facticity and there are many more. We are interested in processes that create facticity. The following definition is useful:

Definition 2. *An incremental information creation process is called* ***factic*** *if it maintains constant facticity of the total generated data set.*

We call data sets with high facticity also *factic*. Note that in order for a process to be factic it must have access to an unlimited source of new information during its execution. In general, factic processes are the result of two conflicting functions: one *generating function* that is an unlimited source of new information and a *constraining function* that regulates the production of information. Note that although factic data sets exist in abundance there is no recursive routine that can construct them since the Kolmogorov complexity needed to judge the faciticity score can not be computed (See figure 4). Another way to say the same thing is that recursive routines can not create new information fast enough to sustain facticity: recursion is not factic. Data sets that are factic with high probability can easiliy be approximated by computational routines that use a random generator as generating function and a standard data compression function as constraining function. (See figure 5).

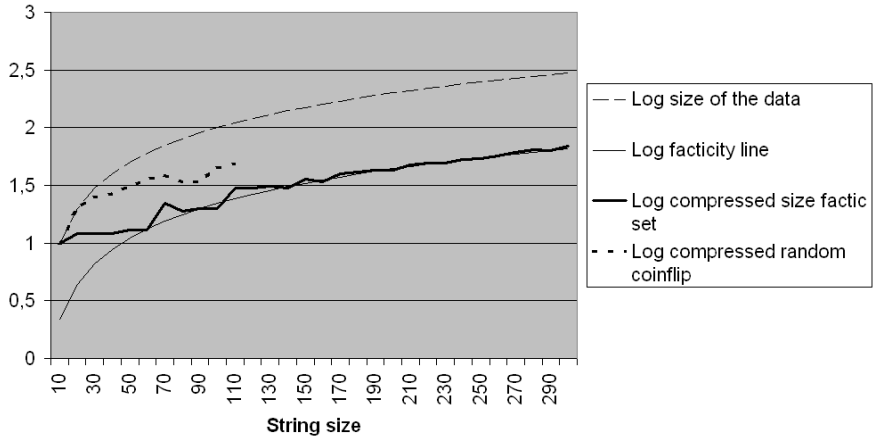

Fig. 5. Example of a factic process based on a random generator. This set was created in the following way. First a test run with a data set of random coinflips was used to estimate C_{max}, since Rar is not an optimal compressor for binary strings represented as bytes. As a result a base compression factor 0.22 was chosen as a facticity target. The string starts with 10 zero's. If the Rar complexity is too high, the sequence is continued with 10 versions of the last symbol. If the Rar complexity is too low, the sequence is continued with 10 random selections from $\{1, 0\}$. The resulting string is:
00000000000000000000000000000000000000000000000000000000000111000000000000
00000000000000000000000000011000101000000000000000000000000000000000110100001
11111111111100100000111111111111111111111111111111101011110111111111111111
11111111101100011000000000001110011000011110001000000000000000000000010100000011

There is an abundance of examples of factic processes:

- Evolutionary processes are in general factic. Here mutation is the information generating function and the environment that regulates survival serves as a constraining function.
- A cooperative teacher (See [3]. If we have a learning agent with limited computational resources (the constraining function) a cooperative teacher (the generation function) would follow a strategy of selecting simple examples that allow the 'pupil' to compress the examples in to rules with relative ease. When the pupil has digested the simple examples the teacher can shift to more complex ones. Thus the complexity of the examples increases monotonically. The teacher will select his examples in a narrow band between what the pupil already knows (order) and what is too complex to process (subjective chaos).
- Curiosity driven 'creative' agents as proposed by Schmidhuber (See [28]). Under assumption that the general capacity to learn gives an evolutionary benefit we expect learning agents that are the product of evolution to have some mechanism that drives them to select new examples that are optimal given their current theories about the structure of their environment. Such an explanation of the evolutionary benefits of curiosity seems plausible. By the same token such a curiosity driven agent should be inclined to ignore any low-complexity examples that are already processed as boring and search examples that 'satisfy' its curiosity. These are the examples that the agent will find 'interesting' in this stage of the learning process. Here the search process of the agent of the generating function and the subjective compression routine of the agent is the constraining function. One might even interpret curiosity driven scientific heuristics as an advanced variant of such an evolutionary survival strategy for the human race.

5 MDL and code optimization

Let us return to our original ambition. Given a system Σ we collect a set of measurements D and represent them in a string x. We are interested in an explanation of the structure of Σ and a prediction of its behavior. What do these ambitions mean in the context of the framework that I have described? We select a sufficiently small universal Turing machine U. [7]This would be a general universal machine that does not contain any information about x. The shortest program that produces x on U is $x^* = argmin_{M \in \mathcal{M}}(U(M) = x)$ and the Kolmogorov complexity of x relative to U is $C(x) = |x^*|$. Note that in this context x^* is a random string, but it is certainly not meaningless, since we have $U(x^*) = x$. In this sense x^* 'explains' all of the structure of x. x^* cannot be computed, but it can be found in finite time by means of dovetailing an enumeration of all possible computations on U. However x^* does not help us much in terms of understanding the structure of x. This changes if we try to compress x in to

[7] Recently Alex Smith proved an intuition of Wolfram about the existence of a very simple universal Turing machine with 3 symbols and 2 internal states [33].

a so-called two part code. It is important to note that two part code optimization is a specific application of MDL. The majority of work on MDL is closer in spirit to the statistical than to the Kolmogorov complexity world ([17]). Rather than two-part codes, one uses general universal codes for individual sequences; two-part codes are only a special case. We give the traditional formulation of MDL [23; 6]:

Definition 3. The Minimum Description Length principle*: The best theory to explain a set of data is the one which minimizes the sum of*

- *the length, in bits, of the description of the theory and*
- *the length, in bits, of the data when encoded with the help of the theory*

Let $M \in \mathcal{M}$ be a model in a class of models $\mathcal{M}$, and let D be a data set. The **prior probability** of a hypothesis or model M is $P(M)$. Probability of the data D is $P(D)$. **Posterior probability** of the model given the data is:

$$P(M|D) = \frac{P(M)P(D|M)}{P(D)}$$

The following derivation [23] illustrates the well known equivalence between MDL and the selection of the Maximum A posteriori hypothesis in the context of Shannon's information theory. Selecting the **Maximum A Posteriori hypothesis(MAP)**:

$$M_{MAP} \equiv argmax_{M \in \mathcal{M}}\ P(M|D)$$

$$= argmax_{M \in \mathcal{M}}\ (P(M)P(D|M))/P(D)$$

(since D is constant)

$$\equiv argmax_{M \in \mathcal{M}}\ (P(M)P(D|M))$$

$$\equiv argmax_{M \in \mathcal{M}} \log P(M) + \log P(D|M)$$

$$\equiv argmin_{M \in \mathcal{M}} - \log P(M) - \log P(D|M)$$

where according to Shannon $-\log P(M)$ is the length of the optimal *model-code* in bits and $-\log P(D|M)$ is the length of the optimal *data-to-mode-code* in bits. This implies that the model that is chosen with Bayes' rule is equal to the model that MDL would select:

$$M_{MAP} \equiv M_{MDL}$$

The formula $argmin_{M \in \mathcal{M}} - \log P(M) - \log P(D|M)$ indicates that a model that generates an optimal data compression (i.e. the shortest code) is also the best model. This is true even if $\mathcal{M}$ does not contain the original intended model as was proved by [30]. It also suggests that compression algorithms can be used to approximate an optimal solution in terms of successive steps of incremental

compression of the data set D. This is *not* true as was shown by Adriaans and Vitányi[4]. Yet this illicit use of the principle of MDL is common practice.

We now turn our attention to incremental compression. Equation 6 gives the length of the optimal *two-part-code*. The length of the two-part-code of an intermediate model M_i is given by:

$$\Lambda(M_i, d) = \log \binom{m_i}{d} + C(M_i) \geq C(D) - O(1) \tag{16}$$

This equation suggests that the optimal solution for a learning problem can be approximated using an incremental compression approach. This is indeed what a lot of learning algorithms seem to be doing: find a lossy compression of the data set finding regularities. This holds for such diverse approaches as nearest neighbor search, decision tree induction, induction of association rules and neural networks. There is a caveat however; Adriaans and Vitányi [4] have shown that the randomness deficiency not necessarily decreases with the length of the MDL code, i.e. shorter code does not always give smaller randomness deficiency, e.g. a better theory. This leads to the following observations:

- The optimal compression of a data set in terms of model and a data-to-model code always gives the best model approximation "irrespective of whether the 'true' model is in the model class considered or not" [30][8].
- This optimal compression cannot be computed.
- Shorter code does not necessarily mean a better model.

These observations show that the naive use of the MDL principle is quite risky. Learning by means of incremental compression might lead to a model that is worse then the one we started with. These observations should make us cautious about the use of incremental compression algorithms. Yet in the real world data compression seems to be a reasonable inductive strategy. This amounts to the following:

Claim. The distributions we find in the world are generally benign in the sense that time and memory bounded tests with reasonable limits for Kolmogorov complexity are sufficient for an adequate complexity estimate.

What the memory and processing time limits would be is a problem for an other paper, but a reasonable intuition would be that the limits lie well within the processing capacity of the human brain. Another way of formulating the same principle is: if a system looks like it is in thermodynamic equilibrium, with high probability it is. This implies that data sets that look random but in fact are highly structured, like the decimal expansion of the number π are highly rare

[8] This is true only in this specific computational framework of reference. In a probabilistic context, both for Bayesian and MDL inference, the assumption that the true model is in the model class considered can sometimes be crucial - this also explains why in Vapnik-Chervonenkis type approaches, complexity is penalized much more heavily than in MDL [16]).

in nature. Why (and if) these data sets do not occur is not completely clear, but a natural assumption would be that natural systems that are capable of calculating such rich data sets are by nature instable and therefore do not exist long enough in time.

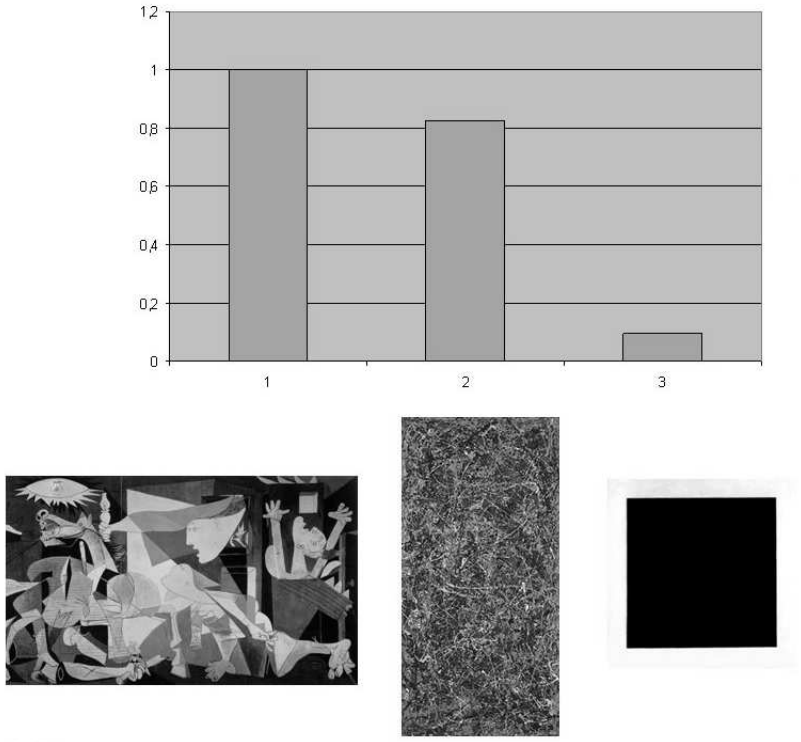

Fig. 6. Facticity scores for three well known works of art. Picasso's Guernica scores a maximal 1. It contains optimal meaningful information. As was to be expected, the black square of Malewich has a low score on the interestingness scale. It contains little information. But also Pollocks composition No. 5 has a lower score. In a way, it contains 'too much' information to be interesting. Note that people always speak about 'the drippings' of Pollock. Apparently it is difficult to keep these high entropy images apart. The facticity scores were calculated in the same way as in figure 1. These works of art typically represent the period of crises in painting in the 20th century in which painters were trying to redefine the conceptual space of their art.

6 Algorithmic esthetics: the dialectic of facticity

Recently Schmidhuber defined a notion of 'interestingness' in a paper with the rather ambitious title "Simple Algorithmic Principles of Discovery, Subjective beauty, Selective Attention, Curiosity & Creativity" [28]. Since there is a relation with the notion of facticity it is useful to present a critical discussion of these ideas. Although I am critical of Schidhuber's theories, at least we seem to agree on one point: algorithmic information theory is a useful formalism to evaluate esthetic theories. Indeed, as we saw in the previous paragraphs, curiosity driven agents tend to produce factic data sets. But it seems not right to equate the notion of 'interestingness' that can be defined for these agents with beauty. As an algorithmic esthetics Schmidhubers conception is not satisfactory. In the following I will argue that the notion of subjective compressibility in art is much more complex than Schmidhuber assumes. In particular great works of art seem to be a rich source of meaning because of the fact that they transcend our rationality (i.e. they have high facticity in themselves and can not be compressed) and not because they have low complexity. *Beauty is not an evolutionary concept.* Artist do not try to construct simple didactic objects, they try to construct objects that are as rich in meaning as possible, i.e. they try to optimize facticity.

At first sight the idea of low complexity art seems to fit nicely with some predominant themes of western philosophy dating back to ancient Greek thought: 1) the Platonic identification of beauty and truth and 2) the identification of truth with simplicity. In various sources from antiquity we find the notion that truth and beauty can be reached through a process of 'idealization' removing all the errors and faults from a collection of similar objects.[9] The fact that there are philosophers that defend those ideas does not imply that they describe what artists actually do. Figure 7 shows that the reality is much more complex. Artists certainly use compression, but not in such a way that beauty can in general be identified with low-complexity. The following variants seem to occur:

- Realism: the representation is isomorphic to the data.
- Idealization: ideal schemas optimally compress the description of a set of examples with errors.
- Schematization: optimal compression under bounded complexity.
- Characterization: optimal bounded compression of an individual example conditional to the optimal general theory.

What is more, all these variants occur side by side throughout history. There is no development from simple to more complex art as would be predicted by Schmidhuber's theory. Especially Plato's identification of truth and beauty that fits so nicely with the concept of a curiosity driven notion of evolutionary beauty should be regarded with suspicion. In the end artists were banned from Plato's ideal state. Artists do not follow rules, they break them.

[9] See e.g. Xenophon, Memorabilia III. This actually shows that the notion of data compression as a process of idealization that approximates some form of truth is much older than Occam. MDL as a scientific methodology has its roots in Greek thought.

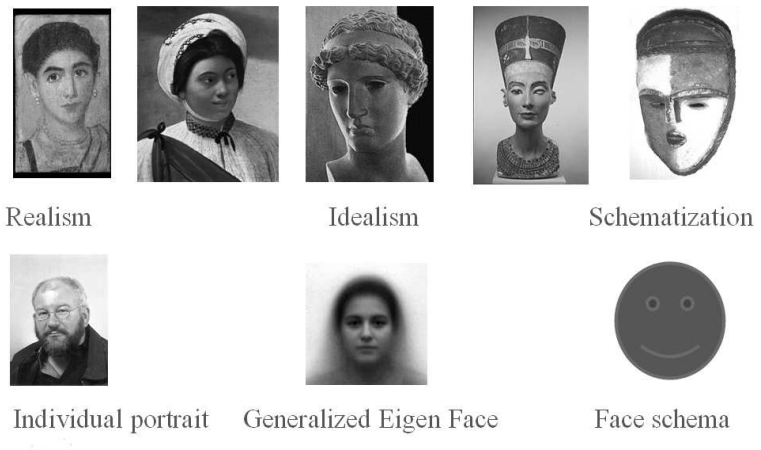

Fig. 7. An illustration of the complex relation between data compression and idealization in art. The eigenface shows that a process of data compression in to a general ideal form is an element of a certain artistic tradition. At the same time extreme realism (very little compression) and schematization (extreme compression) exist. Note that the portrait in the upper left is from Fayoum. It shows that individual portraits already occured in antiquity, illustrating the a-historical character of this form of realism. The idea that beauty has a relation with low-complexity and that the history of art shows an evolution to objects of increasing complexity is simply denied by the facts. The automatically constructed eigen face is due to Luis Jañez Escalada and Miguel Angel Castellanos of the University of Madrid.

The world of art and science have different rhetorical models. An artist communicates directly with his audience through his products. De gustibus non est disputandum. A work of art either fascinates and moves us, or it does not. No scientific argumentation or theoretical explanation can change this, although of course one can train ones sensitivity for the quality of art. There is a consensus amongst most art critics and artists that esthetic judgements are not verbal. If the essential quality of a work of art could be described adequately in language then the work of art would be nothing but an illustration of the text, and thus stop to be an independent work of art. One can say that a work of art is 'good' or 'beautiful' but this verbal judgement only serves as a recommendation and not as an explanation. The beauty of a work of art 'shows' itself in the sense of Wittgenstein's Tractatus. A beautiful object is a constant source of pleasure that defies explanantion. From this perspective any attempt to formulate a scientific theory explaining what beauty is or prescribing what human beings should or would find beautiful is doomed to fail. Books and theories by authors like Hofstadter, Boden, Scha, Ramachandran [25] and Schmidhuber [28] all present us with hypothetical models of the human mind and then try to define beauty or

creativity in terms of these models. Such an excercise may give us deep insights, it does not change the fact that beauty transcends the tools of science.

First of all I observe that we as agents live in a world that is not completely transparent for us. We know that we can influence our environment, but this does not imply that we wholly understand the processes that are involved. We can prepare food without a full understanding of the underlying biological and chemical processes. We can make children without a grasp of the miracle of life. By the same token artists create art objects without an intellectual understanding of the creative process involved. An intellectual conscious decision to make art seldom leads to anything of interest. This is in conflict with the view of the artist as a low-complexity teacher that is one step ahead of his audience and in line with the view that a real object of art transcends verbal analysis.

Secondly, art does not seem to evolve from lower to higher forms of complexity in the way Schmidhuber's theory predicts. On the contrary, as soon as certain conceptual spaces are opening up, we see that artists immediately explore the maximal extension of the artistic possibilities. Homer's Iliad and Odyssey are not boring low complexity books from the beginning of literature, but epic stories that have fascinated people for centuries and that function as inspirational examples for contemporary authors. At the same time the existence of the Homer's works does not keep writers from coming up with new ideas. There is space enough for new stories and plots. Painting in antiquity was certainly not of lower quality than anything that has been done since the Renaissance. The same holds for poetry: Sappho is not a low complexity precursor of Shakespeare's sonnets. Objects in history that were created in the early history of art frequently seem to have a deeper meaning then anything that follows.

On the other hand, it cannot be denied that some forms of art know a historical development. Abstract painting seems partly to be a reaction to the invention of photography forcing painters to redefine the conceptual space of their art (See figure 6). The development of western music is one of increasingly rich harmonic possibilities. But then again, the invention of the 12-tone technique by Schönberg does not make the works of Mozart less beautiful, just as the emergence of abstract painting takes anything away from the fascination of a painting by Raphael. A theory that explains these phenomena is the one that states that artists try to *maximize meaning in a historical context*. A good work of art immediately conquers the full possibilities of the conceptual space in which it emerges (See figure 8). As such its full importance can not be appreciated completely by both the artist and the audience at the time of its conception. This theory implies that artists work in a conceptual space in which they try to optimize meaning (i.e. facticity) right from the start. There is no development form simple to more complex. However dynamic development of art is driven by the fact that any finite conceptual space can be exhausted. If this happens, the possibilities for artistic development are gone, and if the conceptual space is not redefined the art form dies.

Claim. Good art maximizes conditional facticity in a conceptual space that develops historically.

Here is an abstract description of such a process. Consider a community of agents with a general bounded algorithmic compression routine C and some factic creation routine R. The conceptual space for 'works of art' is the set of binary strings of length k. The history h_t at time t of this community consists of sequential individual proposals of works of art $r_1, r_2..., r_t$. A work of art r_{t+1} is considered to be 'good' if:

- It is *meaningful*: It has optimal facticity according to equation 13 using C as compression routine.
- It is *original*: The normalized compression distance NCD of r_{t+1} to any of the elements of the history $r_1, r_2..., r_t$ using C as compression routine is at least c where $0 \ll c \leq 1$.

It is clear that such a combination of demands leads to an interesting non-linear dynamics within this history. Consider the beginning of the process. The first agent makes a proposal r_1 and succeeds in maximizing the facticity of this object. This first object of art r_t consequently structures the rest of the development. Any new object has to stay away from r_1. As soon as a proposed string has a low normalized compression distance to r_1 it is considered to be plagiarism. A second object r_2 again conquers part of the conceptual space away from the region taken by r_1. If the process continues after some time it become impossible to select completely original objects with high facticity. At this moment the production of original meaningful objects of art becomes impossible. Supposing that the agents still have an interest in the creation of art, a possible solution is the expansion of the conceptual space, i.e. increase k to k'. Note however that if this happens it remains still impossible to define new meaningful objects of a length $< k$.

7 Conclusions and further work

In this paper I studied the notion of meaningful information. I showed that this notion is intricately connected with the idea of learning by compression. I introduced the concept of facticity as a first approximation of meaningful information. I studied data compression in the context of thermodynamics and I showed that under adequate measurement conditions the randomness deficiency of a data set is associated with the free energy in the data set.

I also studied two-part code optimization. Here I analyzed compression algorithms that separate a data set in a structural and an ad hoc part. In this way the system is reduced to a typical element of a model and thus the model description has a high probability to produce an adequate prediction of any future development of the system.

Note that systems in thermodynamical equilibrium have no significant development in time. Reducing the description of these systems to random two part-codes compresses the description of the system to those elements that are *time invariant*. That is why such descriptions can be used to predict the future of the system.

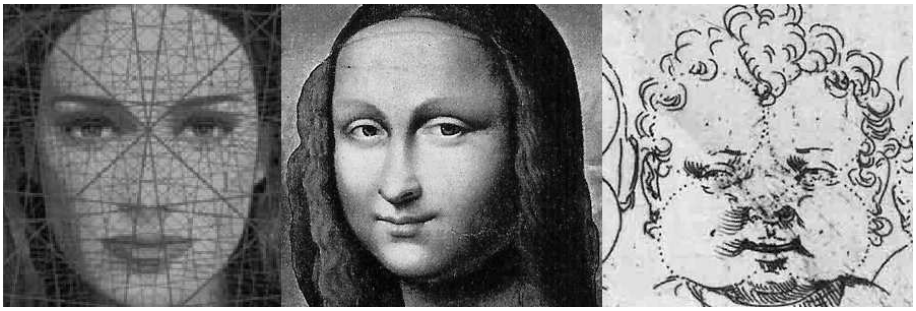

Fig. 8. Left, a picture of a regular schematic feminine face due to Schmidhuber [26]. In the middle, a detail of a copy of the Mona Lisa by Leonardo's untalented protegé Salai. On the right a scheme for a childs head based on an arrangement of four circles in a square due to Fioletti (1608). The last image shows that construction of faces according to simple geometrical schemes was an element of artistical training in the Renaissance. It is clear from the plain look of Salai's painting, which conveys nothing of the fascination of the original, that great works of art are difficult to copy, i.e. they have a meaning that can not be captured by simple geometrical schemas. This supports the view that great works of art optimize facticity and can not be compressed in to low-complexity data sets.

There are a number of ways in which this research could be expanded. Firstly there is the issue of developing good complexity estimates for specific problem classes, so that MDL approaches can be used. I have given initial reports for DFA induction but much improvement is possible [1]. Another direction of research is a deeper analysis of the distributions that I suppose are essential for our capabilities to analyze the world around us. Another interesting exercise could be a further embedding of these insights in the history of philosophy.

Interestingly the claims of the role of facticity in art I have defended here seem to be open for empirical testing (and thus to plain Popperian falsification). This is due to the fact that Cilibrasi's Normal Compression Distance seems to measure cognitive relevant aspects of music represented as midi files. The need felt by composers to strech the limits of consonancy and counterpoint at a certain point in history, should be measurable as an impossibility to come up with interesting original melodies given enough Midi representations of melodies up to that moment. Secondly, given the current status of fMRI technology it is possible to present melodies with various variantions in complexity and facticity and to study invariants in representation in the brain. Normal compression distance seems not to be able to measure cognitive relevant aspect of images but at this moment comparable fMRI and PET-scan studies are done measuring the brain's reaction to images with various Weibull and non-Weibull distributions that have a relation with facticity.[15] Even if the creation of real art will remain a miracle for ever we are bound to get a much deeper insight in the 'innate' probability distributions that our brain uses to analyse and predict the world around us.

Bibliography

[1] Pieter W. Adriaans, Using MDL for Grammar Induction, in Grammatical Inference: Algorithms and Applications, Yasubumi Sakakibara, Satoshi Kobayashi, Kengo Sato, Tetsuro Nishino Etsuji Tomita (Eds.), 8th International Colloquium, ICGI 2006, Tokyo, Japan, September 2006, Proceedings.

[2] Pieter W. Adriaans, Learning Deterministic DEC Grammars is Learning Rational Numbers, in Grammatical Inference: Algorithms and Applications, Yasubumi Sakakibara, Satoshi Kobayashi, Kengo Sato, Tetsuro Nishino Etsuji Tomita (Eds.), 8th International Colloquium, ICGI 2006, Tokyo, Japan, September 20-22, 2006. Proceedings

[3] P.W.Adriaans, The philosophy of learning, in Handbook of the philosophy of information, P.W.Adriaans, J. van Benthem eds. in Handbook of the philosophy of science, Series edited by Dov M. Gabbay , Paul Thagard and John Woods, to appear.

[4] P. Adriaans, P.M.B. Vitányi, The power and perils of MDL, Proc. IEEE International Symposium on Information Theory (ISIT), Nice, France, 24-29 June, 2007, 2216-2220.

[5] F.A. Bais and J.D. Farmer, The physics of information, Handbook of the philosophy of information, P.W.Adriaans, J. van Benthem eds. in Handbook of the philosophy of science, Series edited by Dov M. Gabbay , Paul Thagard and John Woods, to appear.

[6] A. Barron, J. Rissanen, and B. Yu, The minimum description length principle in coding and modeling. IEEE Trans. Information Theory, vol. 44 (1998), no. 6, pp. 2743-2760.

[7] Gregory J. Chaitin, Algorithmic information theory, Cambridge University Press, New York, NY, USA, 1987.

[8] N. Chater and P. Vitányi, Simplicity: A unifying principle in cognitive science? Trends in Cognitive Sciences, 7:1(2003), 19–22.

[9] R. Cilibrasi and P. Vitányi, Clustering by compression, IEEE Trans. Infomat. Th., Submitted. See http://arxiv.org/abs/cs.CV/0312044.

[10] Rudi Cilibrasi and Paul M. B. Vitányi, Automatic Meaning Discovery Using Google, http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0412098, 2004.

[11] James P. Crutchfield and Karl Young, Inferring Statistical Complexity. Physical Review Letters, Vol 63, no. 2, pp. 105-108, July 1989.

[12] Curnéjols A., Miclet L., Apprentissage artificiel, concepts et algorithmes, Eyrolles, (2003)

[13] Dalkilic, M. M., W. T. Clark, J. C. Costello and P. Radiovojac (2006), "Using Compression to Identify Classes of Inauthenic Texts," Proceedings of the 2006 SIAM Conference on Data Mining, ¡http://www.siam.org/meetings/sdm06/proceedings.htm¿, accessed July 23, 2007.

[14] Pedro Domingos, The Role of Occam's Razor in Knowledge Discovery, Data Mining and Knowledge Discovery, volume 3, no. 4, pages 409–425, 1999.

[15] J. M. Geusebroek and A. W. M. Smeulders. A six-stimulus theory for stochastic texture. Int. J. Comput. Vision, 62(1/2):7-16, 2005.
[16] P.D. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification . Machine Learning 66(2-3), DOI 10.1007/s10994-007-0716-7, pages 119-149, 2007.
[17] P.D. Grünwald, The Minimum Description Length Principle. MIT Press, June 2007. 570 pages.
[18] David Hume, An Enquiry Concerning Human Understanding, The Harvard Classics, P.F. Collier & Son, Vol. XXXVII, Part 3, 1909, 1914.
[19] Marcus Hutter, Universal Algorithmic Intelligence: A Mathematical Top→Down Approach, in Artificial General Intelligence, B. Goertzel and C. Pennachin, Springer, Berlin, Cognitive Technologies, pp. 227–290, 2007.
[20] John Locke, An Essay Concerning Human Understanding, London : Dent ; New York : Dutton, 1961.
[21] Shu-Kun Lin, The Nature of the Chemical Process. 1. Symmetry Evolution - Revised Information Theory, Similarity Principle and Ugly Symmetry. Int. J. Mol. Sci. 2, pp. 10-39, 2001.
[22] Lloyd, S. (2000-08-31). "Ultimate physical limits to computation". Nature 406: 10471054.
[23] Mitchell T. M., , Machine Learning, McGraw-Hill, New York, (1997)
[24] Pitt L., Warmuth M. K., The Minimum Consistent DFA Problem Cannot be Approximated within any Polynomial, Journal of the ACM, vol. 40, nr. 1, 95–142, (1993)
[25] V.S. Ramachandran and William Hirstein. The science of art, a neurological theory of aesthetic experience. Journal of Consciousness Studies, 6(6-7):15-51, 1999.
[26] J. Schmidhuber, Low-complexity art, Leonardo, Journal of the international Society for the Arts, Sciences and Technology, vol. 30:2, p. 97-103, MIT Press, 1997.
[27] J. Schmidhuber, Completely Self-Referential Optimal Reinforcement Learners. In W. Duch et al. (Eds.): Proc. Intl. Conf. on Artificial Neural Networks ICANN'05, LNCS 3697, pp. 223-233, Springer-Verlag Berlin Heidelberg, 2005.
[28] J. Schmidhuber. Simple Algorithmic Principles of Discovery, Subjective beauty, Selective Attention, Curiosity & Creativity, to appear in: V. Corruble, M. Takeda, and E. Suzuki (Eds.): DS 2007, LNAI 4755, pp. 26-38, Springer-Verlag, Berlin Heidelberg, 2007.
[29] Ray J. Solomonoff, The Discovery of Algorithmic Probability, Journal of Computer and System Sciences", vol. 55, no 1, pp. 73-88, 1997.
[30] Vereshchagin N.K., Vitányi P.M.B., Kolmogorov's structure functions and model selection, IEEE Trans. Information Theory, vol. 50, nr. 12, 3265–3290, (2004)
[31] Li M., Vitányi P.M.B. An Introduction to Kolmogorov Complexity and Its Applications, 2nd ed., Springer-Verlag, New York, (1997)
[32] J. Gerard Wolff, Unifying Computing And Cognition, The SP Theory and its Applications, CognitionResearch.org.uk, 2006.
[33] Stephen Wolfram, A new kind of science, Wolfram Media Inc., 2002.