Theory Comput Syst (2009) 45: 650–674
DOI 10.1007/s00224-009-9173-y

Between Order and Chaos: The Quest for Meaningful Information

Pieter Adriaans

Published online: 14 February 2009

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.

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

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).

P. Adriaans (✉)
Department of Computer Science, University of Amsterdam, Kruislaan 419, 1098VA Amsterdam, The Netherlands
e-mail: pietera@science.uva.nl

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 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 [5, 17], two-part-code optimization [29], learning as data compression [30] etc. All these approaches are indebted to the formulation of an algorithmic solution to the problem of induction by Solomonoff [28], Chaitin [6] and Kolmogorov [20], 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 [18, 22]. 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 built-in 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 "Built-in" 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* [28]. This research program seems to be the driving force behind the work of researchers like Schmidhuber [26] and Hutter [19]. For a discussion of compressibility as a general cognitive principle see [7].

Occam's razor has been questioned throughout history with fierce opponents (e.g. [14]) and strong defenders (e.g. [30]). 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 [12, 23]. 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 [20, 29] and (2) the application of existing implementations of compression algo-

rithms to approximate the ideal (and uncomputable) Kolmogorov complexity as pioneered by Cilibrasi and Vitányi [8, 9].

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 [17]. Adriaans and Vitányi 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 [3]. The quality of our predictive 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 of 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* [31]. *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 [31], Schmidhuber, Lloyd [21], 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 the 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.

1.2 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:

[1] This last condition seems to rule out exactly those data sets that given the results of Adriaans and Vitanyi [3] 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. An 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).

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 [25]. Recently he introduced a notion of interestingness as *the first derivative of subjective compressibility* [27]. 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 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 built-in a priori probability distribution."

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

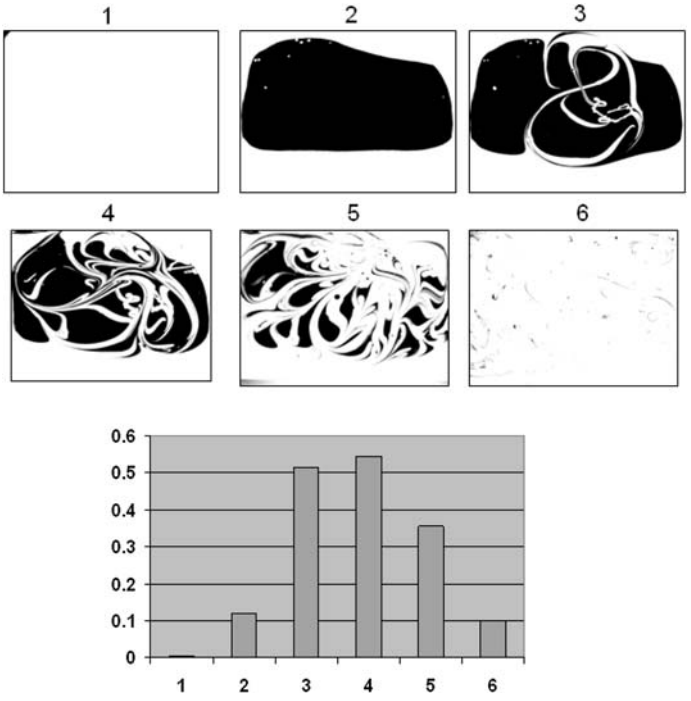

Fig. 1 Facticity scores for mixing black and white paint. For a deeper discussion see paragraph 4. The facticity of a data x is the product (times 4) of the normalized entropy $C(x)/C_{\max}(x)$ and the normalized randomness deficiency $(C_{\max}(x) - C(x))/C_{\max}(x)$. Configuration 4 has the best 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

2 Learning and Thermodynamics

Here is an experiment. Take a cup of coffee and pour some cream in it (see Fig. 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. Below I will propose a theory to make these ideas more precise.

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:

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

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:

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

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.

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

What should we conclude from these definitions 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 $x^* = \operatorname{argmin}_{M\in\mathcal{M}}(U(M) = x)$ 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

[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.

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 [20].

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 (without proof) due to Li and Vitányi [20, p. 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 $C(x|y) \geq \log m - c$.*

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.

3.1 Randomness Deficiency and Minimum Description Length

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 [20]. 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 set M then

$$C(x|M) \leq \log |M| + O(1) \tag{6}$$

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:

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

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:

$$h_x(\alpha) = \min_S \{\log |S| : x \in S, C(S) \leq \alpha\}. \tag{8}$$

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:

$$C(D|M,d) \leq \log \binom{m}{d} + O(1). \tag{9}$$

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 (6) and (9) become 'close' to equalities when respectively x and D are *typical* for M, i.e. when they are random in M. 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:

$$\delta(D|M,d) = \log \binom{m}{d} - C(D|M,d), \tag{10}$$

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

$$\beta_x(\alpha) = \beta_D(\alpha) = \min_M \{\delta(D|M) : M \supseteq D, C(M) \leq \alpha\}. \tag{11}$$

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 [29].

Kolmogorov complexity thus is useful in the context of the socalled Minimum Description Length Principle (MDL). We give the traditional formulation of MDL [5, 23]:

Definition 2 *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.

If D is a data set then the 'best' model M_{MDL} to explain D is given by:

$$\operatorname*{argmin}_{M \in \mathcal{M}} -\log P(M) - \log P(D|M) = \operatorname*{argmin}_{M \in \mathcal{M}} C(M) + C(D|M) = M_{MDL}. \tag{12}$$

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

Under this interpretation of $\mathcal{M}$, the length of the optimal code for an object is equivalent to its Kolmogorov complexity. This specific formulation is also known as two-part code optimization. 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 [16]. Rather than two-part codes, one uses general universal codes for individual sequences; two-part codes are only a special case.

The formula $\operatorname{argmin}_{M\in\mathcal{M}} -\log P(M) - \log P(D|M)$ 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 [29]. 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. Equation (12) 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:

$$\Lambda(M_i, d) = \log \binom{m_i}{d} + C(M_i) \geq C(D) - O(1). \tag{13}$$

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 by means of 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 [3] 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 the 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" [29].[7]
- 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.

3.2 Kolmogorov Complexity Meets Thermodynamics

The mathematical relation between thermodynamic entropy and Kolmogorov complexity is rather straightforward while the philosophical implications are quite formi-

[7] 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 [17].

dable. The expression for the Gibbs entropy in thermodynamics is:

$$S = -\sum_i p_i \ln p_i .$$

The corresponding definition for Shannon entropy is:

$$H \equiv -\sum_i p_i \log_2 p_i .$$

According to Bais and Farmer: *"... this exact quantitative definition of information and its applications transcend the limited origin and scope in conventional thermodynamics and statistical mechanics"* [4]. They consider information theory to be more fundamental then thermodynamics.

The close connection between Shannon entropy and Kolmogorov complexity is observed by, amongst others, Cover and Thomas: *"Gratifyingly, the Kolmogorov complexity K is approximately equal to the Shannon entropy H if the sequence is drawn at random from a distribution that has entropy H. So the tie-in between information theory and Kolmogorov complexity is perfect"* [10, p. 3].

The two observations together i.e. the mathematical equivalence of Shannon entropy and Gibbs entropy and the approximate equivalence between Shannon entropy and Kolmogorov complexity suggest a deep connection between physics and complexity theory. A similar (but much stronger view) is expressed by Li and Vitányi in their standard textbook. On the basis of a somewhat different analysis they conclude: *"... it seems reasonable to assign to each string x an effective thermodynamic entropy equal to its complexity $K(x)$"* [20, p. 551]. They also discuss the relation between Shannon entropy and Gibbs entropy (p. 564).

So let's take the suggestion of Li and Vitányi seriously. What happens when we observe a dynamic system at a certain point in time and store the results in a binary string? One would expect that there is a relation between the thermodynamic qualities of the system and the mathematical qualities of the string. In this paragraph I present a theorem that stipulates a possible interpretation of this connection. For this purpose I will assume that it makes sense to talk about the temperature of a string:

Conjecture 1 *We can assign a temperature to strings.*

For the moment the reader might interpret this as either a very deep insight or a rather surrealistic artefact of the theory. Fact is that in the proof of the central theorem below temperature will be cancelled out against other variables. This is what one would expect, because in our day to day experience the temperature of data sets is irrelevant. My paradigmatic example will be that of a digital camera, but the theorem in principle holds for a range of physical systems for which we store observations in data sets. First let's assume that we can observe a system by means of a *canonical measurement function* h.

Definition 3 Suppose that Σ is a dynamical physical system that evolves over time. A canonical measurement function $h : \Sigma \to \{01\}^{c_e}$ has the following properties:

- Every string produced by h has the same length c_e, which is called the *equilibrium complexity* associated with h and
- h actually measures the entropy S of Σ at time t in terms of the Kolmogorov complexity of its output: $h_t(S) = c_m(C(h_t(\Sigma)))$, where c_m is a constant.
- Specifically $C(h_t(\Sigma)) = c_e$ if Σ is in equilibrium, i.e. in that case the output of h is a random string.

A canonical measurement function brings us from the dynamic world of systems to the static world of binary data sets. Note that it is quite possible that h is a lossy function that gives only a partial model of Σ. A digital camera that always makes pictures with equal file size is an approximation of a canonical measurement function. The length c_e of the binary string that is the output of h is a measure of the maximal amount of information that can be produced. This amount of information will, by definition, only be reached if Σ is in equilibrium, hence the name equilibrium complexity. Note that in an equilibrium state the system has no free energy. All internal energy is converted to work. This is associated with a random string as output of the measurement. This motivates:

Lemma 1 $h_t(U) = c_e$: *the internal energy U of the system is associated with the maximal Kolmogorov complexity c_e of the output of h.*

The following theorem relates the free energy of a system with the randomness deficiency of the data set resulting from observations of the system:

Theorem 2 *Given Conjecture 1, Lemma 1 and a set of canonical measurements $h : \Sigma \to \{0,1\}^{c_e}$ of a dynamic system Σ with free energy F and constant temperature we have*:

$$h_t(F) \equiv \delta(h_t(\Sigma))$$

i.e. the free energy of the system is linear in the randomness deficiency of the data set containing the measurement.

Proof Note that h is a function from a system Σ to a set of binary strings. For Σ by Definition 3 we have $F \equiv U - TS$ which, under the homomorphism h gives:

$$h_t(F) \equiv h_t(U) - h_t(T)h_t(S).$$

By Conjecture 1 we stipulate that $h_t(T) = c_t$. By Lemma 1 we have that $h_t(U) = c_e$. Definition 3 gives: $h_t(S) = c_m(C(h_t(\Sigma)))$

$$h_t(F) \equiv c_e - c_t c_m(C(h_t(\Sigma))).$$

If Σ is in equilibrium we have zero free energy. This gives:

$$c_e = c_t c_m(C(h_t(\Sigma))).$$

At the same time by Definition 3 we have $C(h_t(\Sigma)) = c_e$ for equilibrium situations. So we have $c_e - c_t c_m c_e = 0$, which gives:

$$c_t c_m = 1.$$

Since the temperature is constant and c_m is only dependent on h the corrections for the homomorphism and the temperature cancel each other out. Consequently:

$$h_t(F) \equiv c_e - C(h_t(\Sigma)).$$

Here c_e gives the maximal complexity of the output of h and $C(h_t(\Sigma))$ the actual complexity at time t, this amounts to:

$$h_t(F) \equiv \delta(h_t(\Sigma)).$$ □

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. In real life perfect canonical measurement systems do not exist, if only for loss of energy because of the system being observed. Canonical measurement systems allow us to ignore temperature in our data sets *because* they deliver a perfect image of the entropy of the original system. Of course this theoretical exercise is far from completed, but I hope that it offers a first sketch of the complex interaction between thermodynamics and complexity theory.

4 Joule's Free Expansion Experiment: An Example of Theorem 2

In Joule's free expansion experiment, which is a standard textbook example, a high pressure ideal gas streams in to an isolated vacuum chamber. This is a adiabatic non-equilibrium process for which most of the approximations of thermodynamics do not hold. One would expect the gas to cool down in this process, because the temperature of vacuum is zero. Experiments show that this is not the case: *the temperature remains constant*. The results from the previous paragraph can help us to understand this. This result is in line with the predictions of Theorem 2 in the sense that the only relevant variable fluctuation in this process is the descriptive entropy.

Suppose we have an ideal gas in continuous space, basically a set of n identical perfectly elastic snooker balls in an isolated vacuum cylinder with no gravity. Suppose that at t_0 the particles are all in one half of the cylinder with random positions and velocities. This means that the system has not reached an equilibrium at time t_0. With high probability, after a certain period of time the particles will be evenly distributed over the cylinder. Now the system Σ is given by the following description:

– The exact position and velocity of every particle given in real numbers at time zero.
– A description of Newton's laws that regulates how the system evolves over time.

Note that the descriptive complexity of this system Σ is in principle *infinite*. A randomly selected real from any non-empty interval contains infinite information with probability 1. Now consider a homomorphism p_t that takes the exact position of each particle at time t and sends it to an integer $1 \leq i \leq k$ associated with a grid of k cells defined by a certain discrete coordinate system for the cylinder. $p_t : \Sigma \to P(\mathbf{N})$ is a function from the system Σ to a set of integers that is associated with the position of the balls in the cylinder at time t. Apart from the infinite size of the input there is nothing tricky about this function. Any student could write the program on the basis of a sufficiently rich approximation of the real values in the input. To make the example complete, suppose a second function $q : P(\mathbf{N}) \to \{01\}^{c_e}$ that takes a set of integers S to a binary string s of length c_e that describes this set. Again there is nothing tricky about this function. Any student could implement it. Finally let $h_t \equiv qp_t : \Sigma \to \{01\}^{c_e}$, i.e. the composition of p_t and q. Thus h_t approximates a canonical measurement function that takes the system Σ and produces a file with an approximate description of the position of the particles at time t.

First I analyze this situation from the perspective of information theory. An observer that analyzes the history of Σ will see an increase of the Kolmogorov complexity of output of h to a certain level, after which it stabilizes. After this point in time Σ has reached a thermodynamic equilibrium. Note that the equilibrium complexity is dependent on the granularity of the grid used in h, i.e. we never measure the entropy of the original system directly. It can in this case be defined as: $c_e = C(x) = \log k + \log n + \log \binom{k}{n} + O(1)$. This is the equilibrium complexity of Σ with respect to h. Here the terms $\log k$ and $\log n$ are needed to code the number of cells in the grid and the number of particles in the system and the term $\log \binom{k}{n}$ is the size of an index of the selection of n out of k cells.

If one takes the granularity to be sufficiently high one can prove the following lemma:

Lemma 2 *For all moments in time t_i in which Σ is in equilibrium and each discrete cell contains at most one particle the complexity of the output x of h_{t_i} will be roughly the same with $C(x) = \log k + \log n + \log \binom{k}{n} + O(1)$, i.e. the equilibrium complexity.*

Proof Observe that since the particles are randomly distributed over the space the string x describes a random selection of n cells out of k possibilities, i.e. a random selection of n integers $\leq k$. This gives the desired estimate.

Lemma 2 allows us to make the following observation: if complexity of the output h_{t_i} is smaller than the equilibrium complexity then Σ is not in a state of equilibrium at time t_i. Specifically, when all the particles will be in one half of the cylinder, the upperbound for the complexity of the output x will be: $C(x) \leq \log(k/2) + \log n + \log \binom{(k/2)}{n} + O(1)$, which, for large enough n, is much smaller than the equilibrium bound.

Note that the opposite situation is possible: there are low entropy states that are not 'sensed' by h_t e.g. the situation in which the particles are randomly distributed over the cells, but each particle is *exactly* in the middle of a cell. These states, however, are extremely improbable. So we have gained the following insight: if our data set is compressible below the standard equilibrium description complexity, then the system is not in equilibrium and will have free energy. The converse is not true. Theorem 2 gives the exact connection.

Let us analyze this example again in terms of classical thermodynamics. This is not unproblematic because thermodynamical derivations only work under strict equilibrium conditions that are not always met. Note, that also in the derivation of Gibbs entropy a partition function Z is introduced to renormalize the classes of velocities in to a sound probability distribution. Gibbs entropy is only defined for canonical ensembles. The number of particles and the volume are constant so two conditions for canonical ensembles are met. Others vary over time in the example. In principle there are three different phases:

– At time t_0 all the particles are in one half of the cylinder with random velocities and spatial distribution. For this part of the cylinder we could calculate the standard macroscopic variables, temperature, pressure and entropy by considering it (somewhat erroneously) as a microcanonical ensemble with fixed volume, number of particles and energy. The other half of the cylinder is empty and thus has a vacuum: the pressure, the temperature and the entropy are all zero. One can use this separation of temperature to run a heat engine by allowing the heat to flow from the hot side to the cold side. The Gibbs entropy for the total cylinder is not defined because it is not in equilibrium.
– In the second phase the atoms distribute themselves over the total space, but no equilibrium is reached yet. In this phase the standard macroscopic variables like temperature and pressure are *not defined*. The same holds for the Gibbs entropy.
– In the last phase a state of equilibrium is reached. The Gibbs entropy as well as temperature and pressure are well defined.

There is no exchange of energy with the environment so we have rapid adiabatic expansion. We cannot use the standard definition to estimate the work done by the system, $dU = TdS - PdV$, since P, V and T are not well defined for the whole system during all the three phases. Still, internally we have heat flow and this must be associated with a potential amount of work done by the gas. Since the velocities of the particles do not change during the expansion, the temperature will remain the same. The change in free energy can be explained completely in terms of a change of entropy. Since the gas does not do any work during the expansion the temperature remains the same. This would be different if the gas has to push away a piston during the expansion, then the temperature would also drop. The original free energy of the gas is completely transformed in to entropy:

$$F = TdS. \tag{14}$$

The homomorphism h allows us to estimate the relative change of entropy. This is associated with the relative difference between the equilibrium complexity of the gas

distributed over the whole cylinder minus the initial complexity of the gas distributed over half of the cylinder:

$$h(dS) = c_e^{-1}\left(\log\binom{k}{n} - \log\binom{k/2}{n} + O(1)\right)$$
$$\approx c_e^{-1}\left(\int_{k-n}^{k} \log x dx - \int_{k/2-n}^{k/2} \log x dx + O(1)\right). \tag{15}$$

The integrals in the last part of this equation nicely show that the notion of 'volume change' is also transferred to the information theoretical part of the theory. This expression has to be corrected for the length of the output with a factor c_e^{-1} where c_e is the equilibrium complexity. This is associated with the granularity of the homomorphism h. Also h must be fine grained to such a degree that it reflects the change in entropy of the system. In order for Theorem 2 to do its work it is not necessary to use a grid as fine as in this example (i.e. one particle per cell). This was only introduced to make the mathematics easier. Note that, since q in h is a canonical measurement function we can estimate the randomness deficiency of the output of h at t_0:

$$\delta(h_{t_0}(\Sigma)) = \log\binom{k}{n} - \log\binom{k/2}{n} + O(1). \tag{16}$$

Since the temperature does not change we can consider its image to be constant $c_t = h(T)$. Combining this with (14) and (15) we get:

$$h_{t_0}(F) = h(T)h(dS) = c_t/c_e\left(\log\binom{k}{n} - \log\binom{k/2}{n} + O(1)\right). \tag{17}$$

Insertion of (16) gives:

$$h_{t_0}(F) = c_t/c_e(\delta(h_{t_0}(\Sigma)). \tag{18}$$

This is the desired result: for canonical measurements of adiabatic processes, with constant temperature, the free energy of the system is proportional to the randomness deficiency of the measurements with corrections for temperature and the granularity of the measurement. The factors c_t/c_e remain in the final result because in this case our homomorphism h does not obey the strict conditions of Theorem 2. Information theory can help us to quantify thermodynamic variables in situations in which some of the units are ill defined. □

5 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:

$$\varphi(x) = 4\frac{C(x)}{|x|} \times \frac{|x| - C(x)}{|x|}. \tag{19}$$

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:

$$\varphi(\Sigma) = 4\frac{S}{S_{\max}} \times \frac{S_{\max} - S}{S_{\max}}. \tag{20}$$

This is the rationale behind the experiment represented in Fig. 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.

One might object that my definition of facticity is arbitrary. Why select the maximum on the balance between order and chaos? Why not 1/3 or 1/8? The motivation lies in Theorem 2 in the previous paragraph. If the data set is produced by a canonical measurement function then we have maximal facticity in the exact spot were the product of the free energy stored in the system and the amount of information stored in the system is maximal. *Facticity faithfully measures the amount of useful information in a system*: if facticity is high then there is a lot of information in the system and the system has a lot of free energy to do something with this information. Of course there is a certain arbitrariness and one could choose another optimum. This is a form of arbitrariness that is very common in science. We can measure temperature in degrees Celsius, Fahrenheit or Kelvin. This is OK as long as there are clear conversions and all units of measurement refer to the same underlying concept, in this case temperature. Here I present facticity as an abstract formal concept with a well founded stimulative definition.

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 [8]. 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):

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}. \tag{21}$$

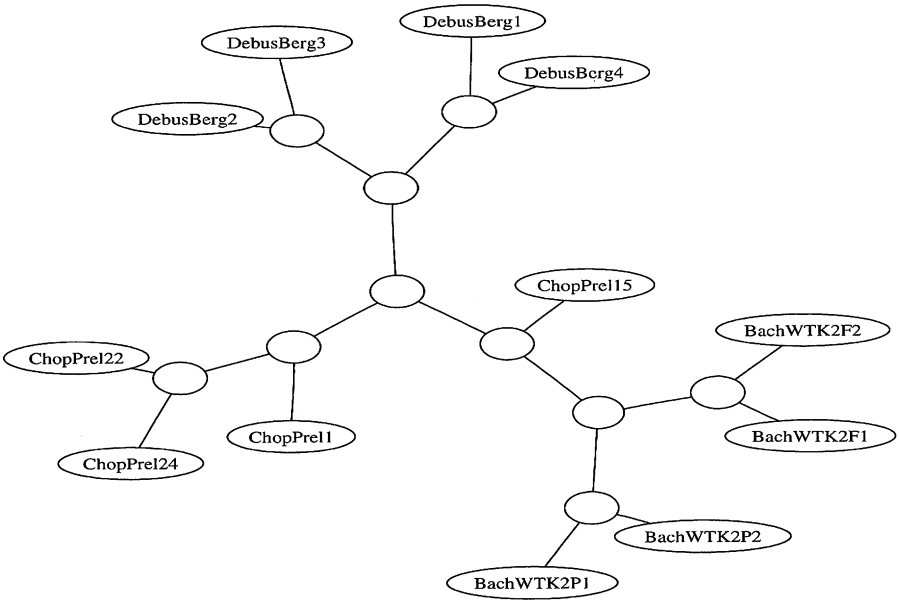

Fig. 2 A tree representation based on the normalized compression distance between 12 Piano pieces

Figure 2 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.

5.1 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 can also be introduced via other constructions. Facticity in a dynamic setting 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 4 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 seem to be 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 facticity score can not be computed. 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 easily be approximated by computational routines that use a random generator as generating function and a standard data compression function as constraining function. 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 [2]). 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 [27]). 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.

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 have seen that we should be 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 intuitive:

Claim 1 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 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.

6 Algorithmic Esthetics

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" [27]. 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 (see Fig. 3).

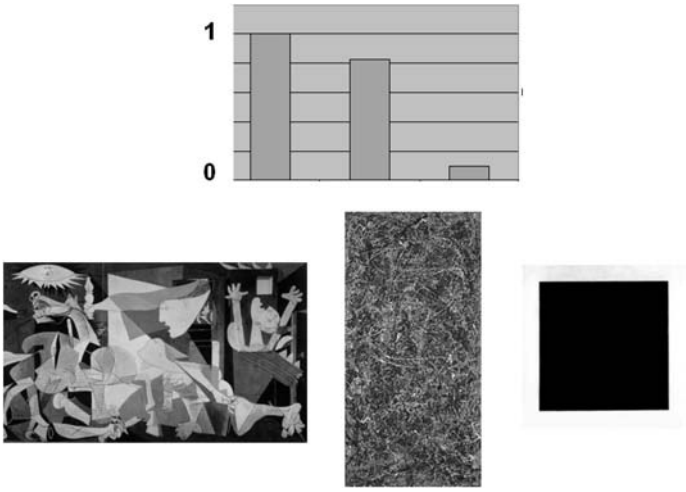

Fig. 3 Facticity scores for three well known works of art. Picasso's Guernica scores a maximal 1. It contains optimal meaningful information. As was to be expected, the black square of Malewich has a low score on the interestingness scale. It contains little information. But also Pollocks composition 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 Fig. 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

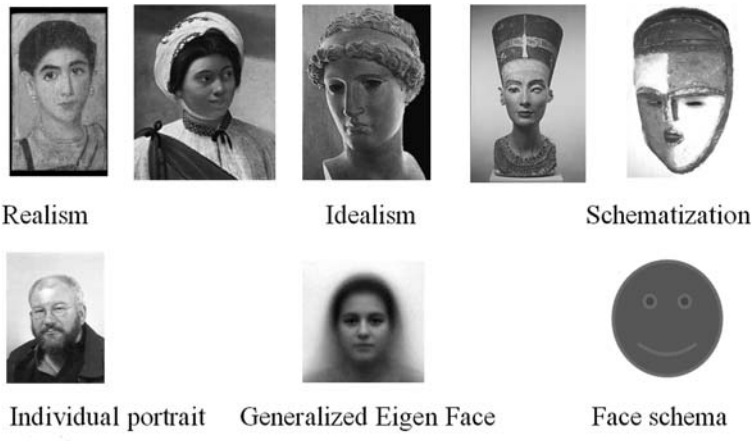

Fig. 4 An illustration of the complex relation between data compression and idealization in art. The eigenface shows that a process of data compression in to a general ideal form is an element of a certain artistic tradition. At the same time extreme realism (very little compression) and schematization (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

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 (see Fig. 5). 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*. Artists 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.[8] The fact that there are philosophers that defend those ideas does not imply that they describe what artists actually do. Figure 4 shows that the reality is much more complex. Artists certainly use compression, but not in such

[8]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.

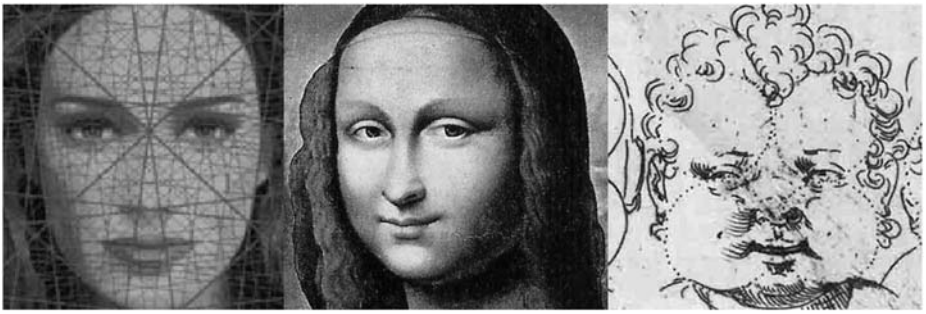

Fig. 5 *Left*, a picture of a regular schematic feminine face due to Schmidhuber [25]. In the *middle*, a detail of a copy of the Mona Lisa by Leonardo's untalented protegé Salai. On the *right* a scheme for a childs head based on an arrangement of four circles in a square due to Fioletti (1608). The 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

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.

The world of art and science have different rhetorical models. An artist communicates directly with his audience through his products. 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. 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 Ramachandran [24] and Schmidhuber [27] present us with hypothetical models of the human mind and then try to define beauty or creativity in terms of these models. Such an exercises may give us deep insights, it does not change the fact that beauty transcends the tools of science.

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.

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.

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 stretch 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 variations 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.

References

1. Adriaans, P.W.: Using MDL for grammar induction. In: Sakakibara, Y., Kobayashi, S., Sato, K., Tomita, T.N.E. (eds.) Grammatical Inference: Algorithms and Applications, 8th International Colloquium, ICGI 2006, Tokyo, Japan. Springer, Berlin (2006)
2. Adriaans, P.W.: The philosophy of learning. In: Adriaans, P.W., van Benthem, J. (eds.) Handbook of the Philosophy of Information. Handbook of the Philosophy of Science, series edited by D.M. Gabbay, P. Thagard and J. Woods (2008)

3. Adriaans, P.W., Vitányi, P.M.B.: The power and perils of MDL. In: Proc. IEEE International Symposium on Information Theory (ISIT), Nice, France, 24–29 June, pp. 2216–2220 (2007)
4. Bais, F.A., Farmer, J.D.: The physics of information. In: Adriaans, P.W., van Benthem, J. (eds.) Handbook of the Philosophy of Information. Handbook of the Philosophy of Science, series edited by D.M. Gabbay, P. Thagard and J. Woods (2008)
5. Barron, A., Rissanen, J., Yu, B.: The minimum description length principle in coding and modeling. IEEE Trans. Inf. Theory **44**(6), 2743–2760 (1998)
6. Chaitin, G.J.: Algorithmic Information Theory. Cambridge University Press, Cambridge (1987)
7. Chater, N., Vitányi, P.: Simplicity: a unifying principle in cognitive science? Trends Cogn. Sci. **7**(1), 19–22 (2003)
8. Cilibrasi, R., Vitányi, P.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005)
9. Cilibrasi, R., Vitányi, P.M.B.: Automatic meaning discovery using Google. http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0412098 (2004)
10. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2006)
11. Crutchfield, J.P., Young, K.: Inferring statistical complexity. Phys. Rev. Lett. **63**(2), 105–108 (1989)
12. Curnéjols, A., Miclet, L.: Apprentissage Artificiel, Concepts et Algorithmes. Eyrolles, Paris (2003)
13. Dalkilic, M.M., Clark, W.T., Costello, J.C., Radiovojac, P.: Using compression to identify classes of inauthenic texts. In: Proceedings of the 2006 SIAM Conference on Data Mining (2007). http://www.siam.org/meetings/sdm06/proceedings.htm
14. Domingos, P.: The role of Occam's Razor in knowledge discovery. Data Min. Knowl. Discov. **3**(4), 409–425 (1999)
15. Geusebroek, J.M., Smeulders, A.W.M.: A six-stimulus theory for stochastic texture. Int. J. Comput. Vis. **62**(1/2), 7–16 (2005)
16. Grünwald, P.D.: The Minimum Description Length Principle. MIT Press, Cambridge (2007)
17. Grünwald, P.D., Langford, J.: Suboptimal behavior of Bayes and MDL in classification under misspecification. Mach. Learn. **66**(2–3), 119–149 (2007). doi:10.1007/s10994-007-0716-7
18. Hume, D.: An Enquiry Concerning Human Understanding, The Harvard Classics, vol. XXXVII, Part 3. PF Collier, Toronto (1909–1914)
19. Hutter, M.: Universal algorithmic intelligence: a mathematical top→down approach. In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence. Cognitive Technologies, pp. 227–290. Springer, Berlin (2007)
20. Li, M., Vitányi, P.M.B.: An Introduction to Kolmogorov Complexity and Its Applications, 2nd edn. Springer, New York (1997)
21. Lloyd, S.: Ultimate physical limits to computation. Nature **406**, 1047–1054 (2000)
22. Locke, J.: An Essay Concerning Human Understanding. Dent/Dutton, London/New York (1961)
23. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
24. Ramachandran, V.S., Hirstein, W.: The science of art, a neurological theory of aesthetic experience. J. Conscious. Stud. **6**(6–7), 15–51 (1999)
25. Schmidhuber, J.: Low-complexity art, Leonardo. J. Int. Soc. Arts Sci. Technol. **30**(2), 97–103 (1997)
26. Schmidhuber, J.: Completely self-referential optimal reinforcement learners. In: Duch, W., et al. (eds.) Proceedings of the International Conference on Artificial Neural Networks ICANN'05. LNCS, vol. 3697, pp. 223–233. Springer, Berlin (2005)
27. Schmidhuber, J.: Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity and creativity. In: V. Corruble, M. Takeda, E. Suzuki (eds.) DS 2007. LNAI, vol. 4755, pp. 26–38 (2007)
28. Solomonoff, R.J.: The discovery of algorithmic probability. J. Comput. Syst. Sci. **55**(1), 73–88 (1997)
29. Vereshchagin, N.K., Vitányi, P.M.B.: Kolmogorov's structure functions and model selection. IEEE Trans. Inf. Theory **50**(12), 3265–3290 (2004)
30. Wolff, J.G.: Unifying Computing and Cognition, the SP Theory and Its Applications. CognitionResearch.org.uk, Menai Bridge (2006)
31. Wolfram, S.: A New Kind of Science. Wolfram Media, Champaign (2002)