# On Autistic Interpretations of Occam's Razor

**Jose Hernández-Orallo**[1] and **Ismael García-Varea**[2]

**Abstract.** Recently, an overhype about the MDL principle is surrounding some fields of Artificial Intelligence, especially machine learning, neural networks, ILP and many others, supported by recent proofs that the shorter the more likely. In this paper we discuss in a critical and sometimes informal way these justifications because they are proved for very ideal, infinite and, in our view, artificial situations. Using the same information-theoretic approach, we study the case for finite and short data and we arrive to a slightly different result: MDL is a good principle but not the best one for finite strings and perfect hypotheses. The argument is based upon recently introduced variants and definitions around the idea of Intensional Complexity, which penalise or 'simply' do not allow exceptions, seen these as extensional descriptions. Intensional considerations change the statement that *"optimal compression (Minimal Description Length (MDL)) gives you the best hypothesis provided the data are random with respect to the hypothesis, the data are not completely perfect and the data grow to infinity"* into the following one *"the intensionality criterion gives you a more feasible hypothesis when the data are perfect ensuring and not supposing that the data are random to the hypothesis."* Also it does not require that the data grow to infinity, so it can be used to "understand" finite real problems. More importantly, our definitions are free from the "MDL's principle" paradox, since the shortest hypothesis is never random to the data. In one of the two formalisations we present, the connection with learning and Levin's "Universal Search Problems" is made explicitly. In the end, and very far from the classical notion of 'identification', we propose a different notion of learning: *the more a system learns the more intensional the description is respect to the data.* Consequently the blurry notions of underfitting and overfitting may be better understood.

## 1 Introduction

After the initial relevance of deduction in AI, nowadays its counterpart, inductive learning, and all its varieties (abductive reasoning, reasoning by analogy, connectionism, ILP, grammatical inference, HMM, EBR, etc.), are beginning to play the more central and still agglutinative role in the proper subset of AI devoted to make intelligent machines.

This has motivated new unified frameworks for understanding reasoning. In particular, a radical approach has been undertaken by Wolff, with the claim that *"all kinds of computing and formal reasoning may usefully be understood as information compression by pattern matching, unification and search"* [40].

In this paper, we will discuss critically a central and now classical issue in this line, which is based on the view of unsupervised learning as compression [31], its famous operative principle MDL [27] and, mainly, its formal justification. The discussion is motivated from the apparent contradiction between the so-called "no-free-lunch" theorems about induction [41] [29] stating that one learner cannot be better than another when performance is averaged uniformly over all possible problems. These results only allow that a learner could be better than another for a particular distribution of problems. [36] show that the MDL principle is almost optimal for the universal distribution $2^{-K(x)}$. Of course, universal distribution is an objective choice when you have not any information at all about the real origin of the information, but is this the case in real applications of machine learning? Are we always so autistic about the source of the information that we pretend to learn?

After this discussion under the view of an intensional variant of Kolmogorov Complexity, we will introduce some considerations on knowledge construction, meaning and ontology, which are also capital questions in AI.

## 2 Learning by Data Compression

Recently, there is some kind of overhype [26] [13] [16] [15] [23] [42] [25] [14] about unsupervised learning and pattern recognition as data compression [39] and Rissanen's Minimum Description Length Principle (MDL) [27]. This is a variant of the Minimum Message Length (MML) principle [38] but the idea is *formally* older [31]. They all are "fresh interpretations" [14] under Algorithmic Information Theory of a much older idea attributed to William of Ockham 1290?-1349?:

*Occam's Razor Principle: "If there are alternative explanations for a phenomenon, then,* **all other things being equal**, *we should select the simplest one."*

This principle was rejected by Popper because he said that it had no sense because there is no objective criterion for simplicity. But Algorithmic Complexity $C(x)$ or Kolmogorov Complexity $K(x)$ is an objective criterion for simplicity. This is precisely what R.J.Solomonoff proposed as a perfect theory of induction, in [36] words. Algorithmic Complexity inspired J. Rissanen in 1978 to use it as a general modelling method, giving the popular MDL principle [27]:

**Minimum Description Length (MDL) principle:** *The best theory to explain a set of data is the one which minimises the sum of: the length, in bits, of the description of the theory; and, the length, in bits, of data when encoded with the help of the theory.*
*Then, we enclose the exceptions, if any.*

This principle is not computable in general and it can only be used if we restrict the descriptive mechanism. The first justified reason to use the MDL principle is to avoid overgeneralisation.

But MDL is used mainly for another reason, *"the shorter the hypothesis the more predictable it is"*. This has been proved for statistical prediction. In [36] and [37] it is shown that in many cases Bayesian reasoning "is prone to overfitting" while MDL is not. [35] discuss an "ideal MDL principle" where *"with a more complex description of the hypothesis H, it may fit the data better and therefore decreases the misclassified data. If H describes all the data, then it does not allow for measuring errors. A simpler description of H may be penalized by increasing the number of misclassified data. If H is a trivial hy-*

[1] Universitat Politècnica de València, Departament de Sistemes Informàtics i Computació. Camí de Vera 14, Aptat. 22.012 E-46071, València, Spain. E-mail:jorallo@dsic.upv.es

[2] Universitat Politècnica de València, Institut Tecnològic d'Informàtica. Camí de Vera 14, Aptat. 22.012 E-46071, València, Spain. E-mail:ivarea@iti.upv.es


Submitted to ECAI 98
January 20, 1998