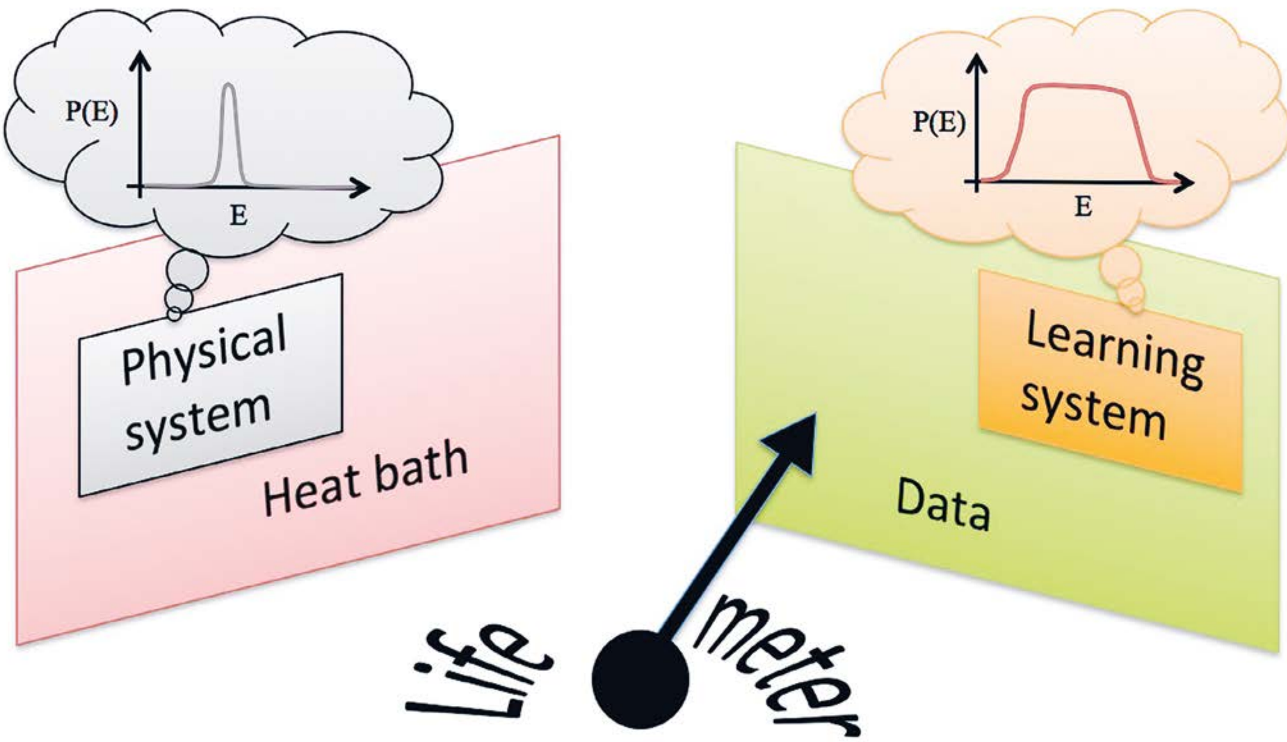


ON THE IMPORTANCE OF BEING CRITICAL

Matteo Marsili – DOI: https://doi.org/10.1051/epn/2020508
[1]The Abdus Salam International Centre for Theoretical Physics – 34151 Trieste, Italy

Being critical, *i.e.* able to process and distill relevant information, is crucial for living systems. Learning distinguishes living from inanimate matter. Quantifying this distinction may provide a "life meter" [1] that, for example, can allow us to detect alien life forms in astrobiology. Living systems also respond in an anomalous manner to perturbations, as compared to inanimate matter, unless the latter is poised at a critical state (in the statistical physics sense). I argue below that these two notions of criticality are only apparently different, because a system that learns is inherently critical, also in the statistical physics sense.

Artificial Intelligence provides an ideal framework to investigate learning in quantitative terms. A learning machine differs in two important respects from a physical system. First, the latter is ruled by the maximum entropy principle, that constrains the distribution to the Gibbs-Boltzmann form and fixes its parameter, the temperature, to that of the environment. Therefore, all a physical system knows about its environment is just one number. Instead, learning machines, such as deep neural networks, can adjust billions of parameters to shape their internal representations. Second, the "environment" in which learning machines are immersed is the data they learn. This "environment" (*e.g.* digital pictures) generally has a very rich statistical structure of hidden features. After all, there should be something to learn [2]!

Taking this comparison further, one [3,4,5] realises that learning is a dual problem with respect to statistical mechanics. In the latter, the Hamiltonian is given and the distribution over internal states is dictated by the maximum entropy principle. In learning, the Hamiltonian is adjusted in the training process. The energy takes the natural meaning of a coding cost and its average fixes the **resolution** of the representation, *i.e.* the number of bits spent to represent one data point. Physical systems have a very narrow distribution of energies. A learning machine, instead, will make its spectrum of energies (*i.e.* coding costs) as broad as possible to distinguish significantly different structures in the data [3,5]. This is because the ***relevance***, which is an entropic measure of the width of the energy range, lower bounds the amount of information that the representation extracts about the hidden features of the data [5]. Therefore, machines that achieve a maximal relevance at a given resolution (*i.e.* those with a broad energy distribution) are those that are maximally informative about the hidden features.

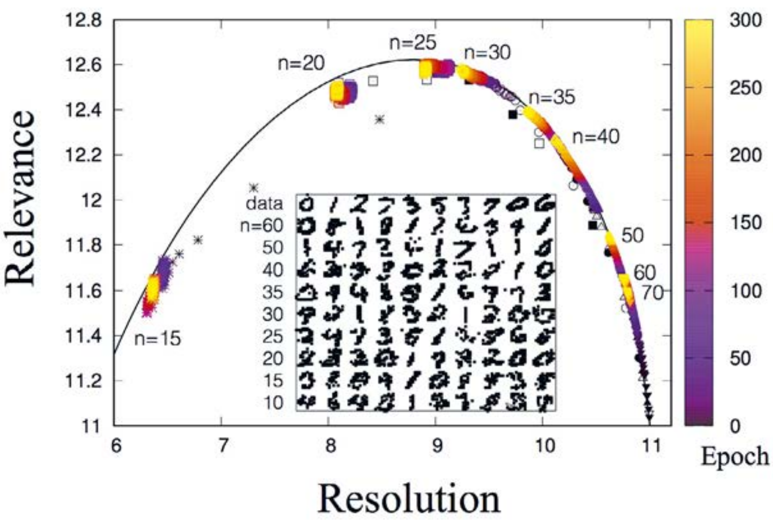
The ***principle of maximal relevance*** has been verified in learning machines such as Restricted Boltzmann Machines (see Fig. 1) or Deep Belief Network [5,6] and efficient coding [7]. This principle predicts distributions of energy levels that are consistent with the suggested hyperbolic geometric nature of sensory representations [8], and with the (expected) invariance of representations under coarse graining of the data (*i.e.* a picture of a dog should be classified in the same way, both at high and low resolution). This principle also predicts a linear relation between the (micro canonical) entropy and the energy, in the statistical mechanics analogy, which is a signature of criticality [7]. A sample of states from this distribution exhibits statistical criticality, *i.e.* the frequency with which a given state is observed scales with an inverse power law with its frequency rank. The exponent of this law is related to the trade-off between resolution and relevance (see Fig. 1): at high-resolution, the representation is noisy. Further compression extracts more relevant information (*i.e.* the relevance increases). The relevance reaches a maximum at an intermediate resolution and then it decreases.

At the maximum, the distribution of energies is flat, which corresponds to the occurrence of Zipf's law [10]. Hence Zipf's law appears as a signature of maximally informative representations. In this perspective, the occurrence of Zipf's law in the frequency of words in most languages is not surprising. More generally, biological systems are critical not because they are "poised" at a critical point, but rather because their internal states (*e.g.* the immune system) are efficient representations of their environments (*e.g.* the space of pathogens). This makes criticality a candidate "life meter" [1], because it occurs generically in learning systems, irrespective of what is learned.

“ The relevance lower bounds the amount of information that the representation extracts about the hidden features of the data ”

The abundance of experimental data in life sciences presages remarkable advances in unveiling Nature's code. This requires statistical inference to venture in scarcely explored domains. First, our ignorance of the underlying laws that govern living system requires model-free methods. Second, inference needs to operate in the deep under-sampling regime, where the number of samples is very small compared to the dimensionality of the data. For example, the number of sequences of a protein domain that evolution dispensed us with may be much less than those necessary to estimate parametric models. Life itself operates on very sparse data. A bacterium cannot afford the luxury of estimating changes in gradient concentrations to an arbitrary precision, before mounting the appropriate response. Inference methods need to take into account this trade-off between "chance" and "necessity", to extract information on biological function from data. The concept of relevance provides a guide to explore the under-sampling

▾ **FIG. 1: Learning machines converge to representations of maximal relevance at a given resolution. The values of the resolution and of the relevance are reported at different stages of training (from black to yellow) for Restricted Boltzmann Machines (RBM) trained on a dataset of had-written digits (see Ref. [5] for more details). Different symbols correspond to different number of hidden units, from 15 (left) to 70 (right). The full black curve corresponds to the theoretical limit of maximal relevance. Inset: sample digits in the data (top row) are compared with those generated by RBM's with varying number of hidden units.**



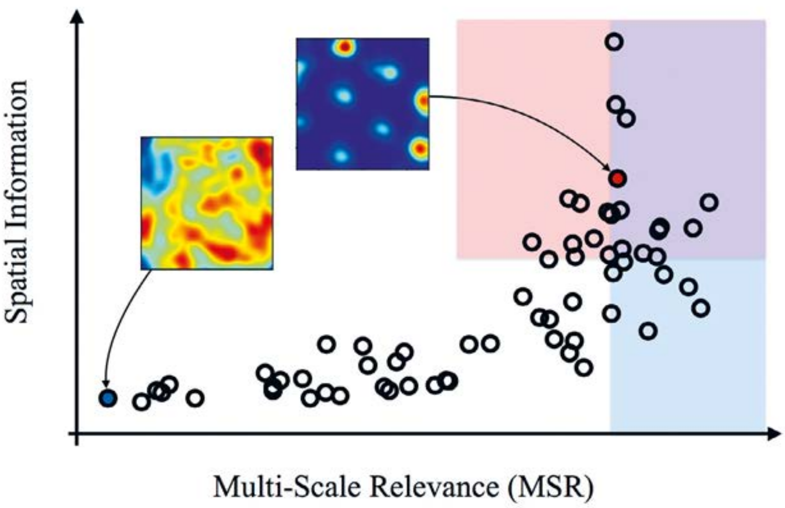
> **Biological systems are critical not because they are "poised" at a critical point, but rather because their internal states are efficient representations of their environments**

regime in a model-free manner, and to detect the signature that "necessity" leaves in biological data. If the internal state of a living system is a maximally informative representation, searching for critical subsets of variables (those with high relevance) in the data, may shed light on how that living system copes with its environment.

The neural code responsible for spatial cognition is a good example of how relevance-based inference differs from standard methods. The Moser's lab [11] deciphered the neural code responsible for spatial cognition in rats, by identifying specialised neurons that fire when the rat visits the vertices of a triangular lattice (see Fig. 2). They did this by correlating neural activity with the rat's position. Evidently, upstream neurons in the cortex don't need to know the rat's position to identify neurons responsible for spatial cognition. This means that one can find these neurons on the basis of their spiking time series alone. Cubero *et al.* [12] have shown that it is possible to identify relevant neurons, without a clue of what they are coding for, using only data on neural activity. They propose an indicator, called Multi-Scale Relevance (MSR), which characterises the response of a neuron on all time-scales. Neurons with a low MSR are uninformative about navigation covariates (position or direction), whereas all informative neurons have a large value of the MSR (see Fig. 2). The group of neurons with largest values of the MSR allows one to decode the position of the animal as precisely as the group of neurons with the largest spatial information content.

The concept of relevance has also been used to identify biologically relevant sites in protein sequences [13], but it can also be applied to the study of single cell expression data or of DNA methylation, to mention just two examples. In each case, the concepts needs to be declined in appropriate ways, always keeping in mind Doctor Zhivago's warning that "[...] life is never a material, [...] it is infinitely beyond your or my obtuse theories about it".

▾ **FIG. 2: Multi-Scale relevance and Spatial Information for 65 neurons recorded in the medial Enthorinal Cortex of navigating rats, by Stensola *et al.* [11] (data from Ref. [12]). The spatial information is the mutual information between the rat's position and the neural activity, during the experiment. The insets show the firing rate maps for two representative neurons, that reports the probability of spiking at a given location. The 20 most relevant neurons (in the shaded blue region) can predict the position of the rat at least as well as the 20 most informative ones (red shaded region). Note that identifying the most relevant neurons only requires data on the neural activity, whereas the most (spatially) informative neurons can only be found if also the spatial covariate is known.**



About the Author

Matteo Marsili is a theoretical physicist, leading the Quantitative Life Sciences Section of the Abdus Salam ICTP. He's interested in understanding collective phenomena in different disciplines (physics, biology, economics and finance, statistical inference and learning) with statistical physics methods.

References

[1] P. Davies, *Physics Today* **73(8)**, 34 (2020)

[2] M. Mézard, *Phys. Rev. E* **95**, 022117 (2017)

[3] R J Cubero, et al, *Journal of Statistical Mechanics: Theory and Experiment* **6**, 063402 (2019)

[4] M. Marsili, *Journal of Statistical Mechanics: Theory and Experiment* **10**, 103401 (2019)

[5] O. Duranthon, M. Marsili and R. Xie, *Maximal Relevance and Optimal Learning Machines*, arXiv:1909.12792 (2019)

[6] J. Song, M. Marsili and J. Jo, *Journal of Statistical Mechanics: Theory and Experiment* **12**, 123406 (2018)

[7] R. J. Cubero, M. Marsili and Y. Roudi, *Entropy* **20(10)**, 755 (2018)

[8] T. O. Sharpee, *Current Opinion in Neurobiology* **58**, 101 (2019)

[9] T. Mora and W. Bialek, *J. Stat. Phys.* **144.2**, 268 (2011).

[10] L. Aitchison, N. Corradi and P. E. Latham, *PLoS computational biology* **12(12)**, e1005110 (2016)

[11] H. Stensola, *et al.*, *Nature* **492** (7427), 72

[12] R. J. Cubero, M. Marsili and Y. Roudi, *J. Comput. Neurosci.* **48(1)**, 85 (2020)

[13] S. Grigolon, S. Franz and M. Marsili, *Molecular BioSystems* **12(7)**, 2147 (2016)