Information Cohomology methods for learning the statistical structures of data

Pierre Baudot*, PhD[1], Mathieu Bernardi, Mscl[1]
[1]Median Technologies, France
in col. with Daniel Bennequin, Monica Tapia, Jean-Marc Goaillard.
*corresponding author e-mail: pierre.baudot@mediantechnologies.com

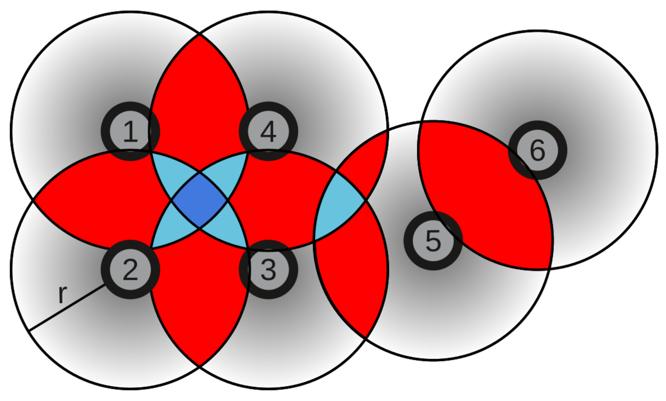
Abstract

Information cohomology is a branch of topological data analysis that allows us to quantify directly statistical dependencies and independences in a given dataset. Theorems establish Shannon entropy as a first cohomology class and mutual information as coboundaries on finite probability space endowed with a random variable chain complex structure. We present some simplicial subcase applications to data analysis in different contexts: transcriptomic and digits and medical CT image classification.

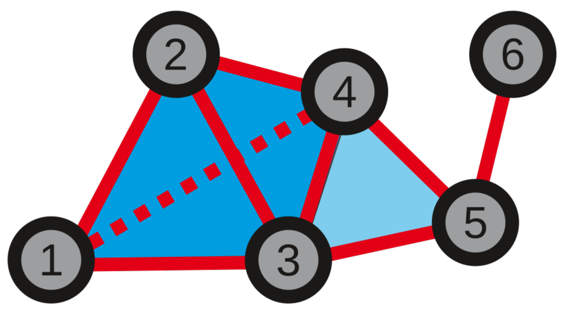
Principles and Theory

We consider random variables as partitions of atomic probabilities and the associated poset given by their lattice. The basic cohomology is settled by the Hochschild coboundary, with a left action corresponding to information conditioning. The first degree cocycle is the entropy chain rule, allowing to derive the functional equation of information and hence to characterize **entropy uniquely as the first group of the cohomology**. (minus) Odd multivariate mutual informations (MI, I_{2k+1}) appears as even degrees coboundary, and the introduction of a second trivial or symmetric action coboundary gives even MI (I_{2k}) in the odd degrees. **Mutual statistical independence is equivalent to the vanishing of all k-MI (I_k=0)**, leading to the conclusion that the Ik define refined measures of statistical dependencies and that the cohomology quantifies the obstruction to statistical factorization. We develop the computationally tractable subcase of on the simplicial (Boolean) sub-lattice, represented by **entropy H_k and information I_k landscapes.**

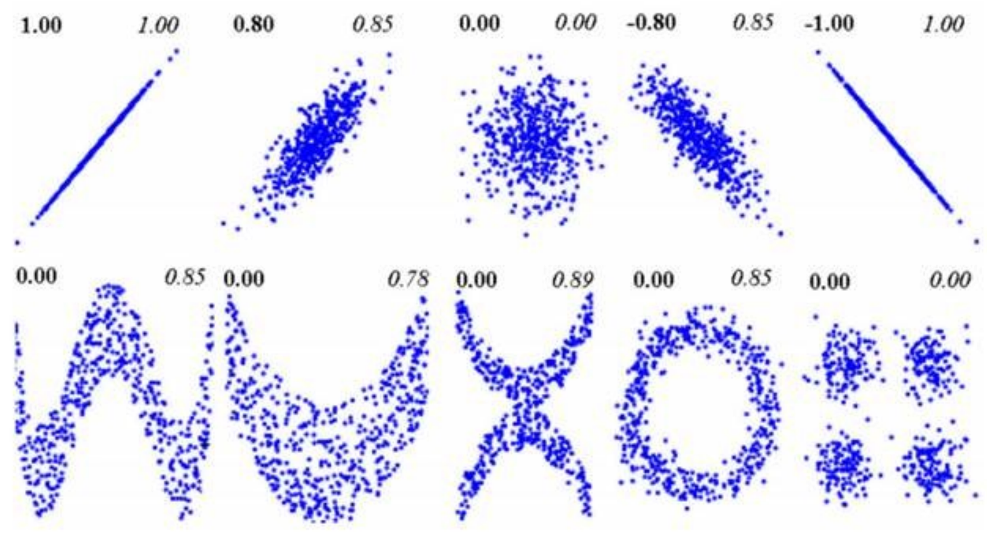
Persistence analysis principle (for comparison)



Data points with balls

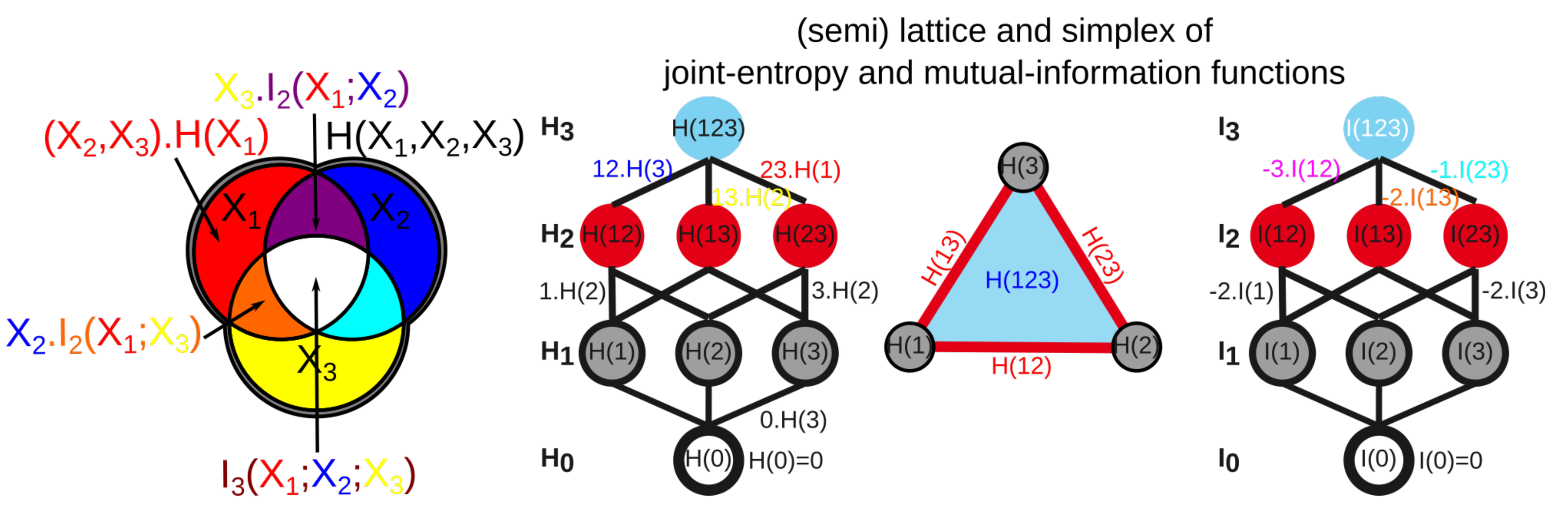


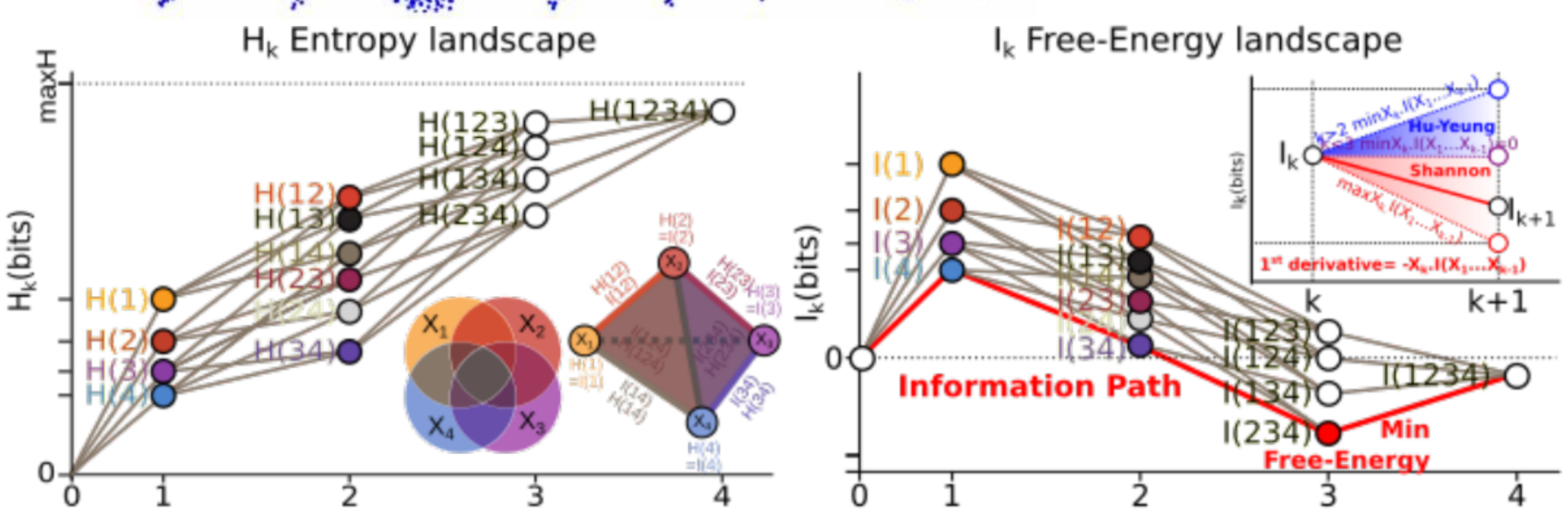
Čech 3-complex



Mutual Informations generalize correlations and Cov.

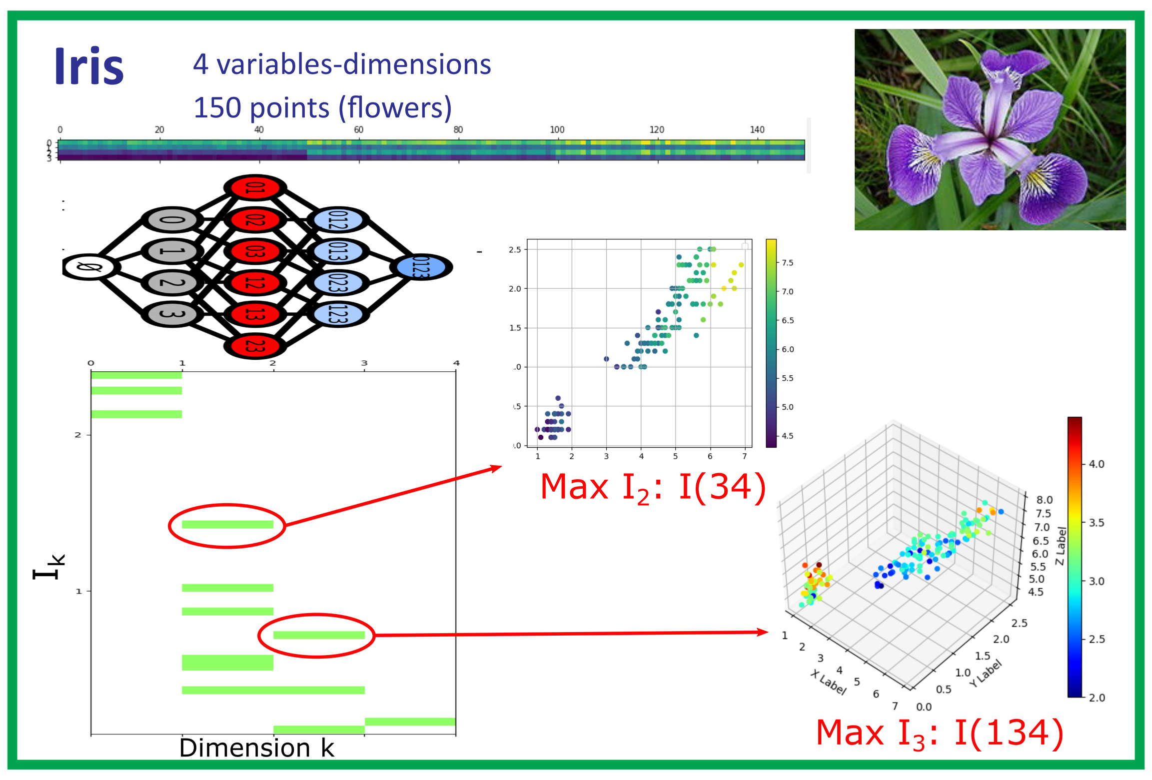
(semi) lattice and simplex of joint-entropy and mutual-information functions

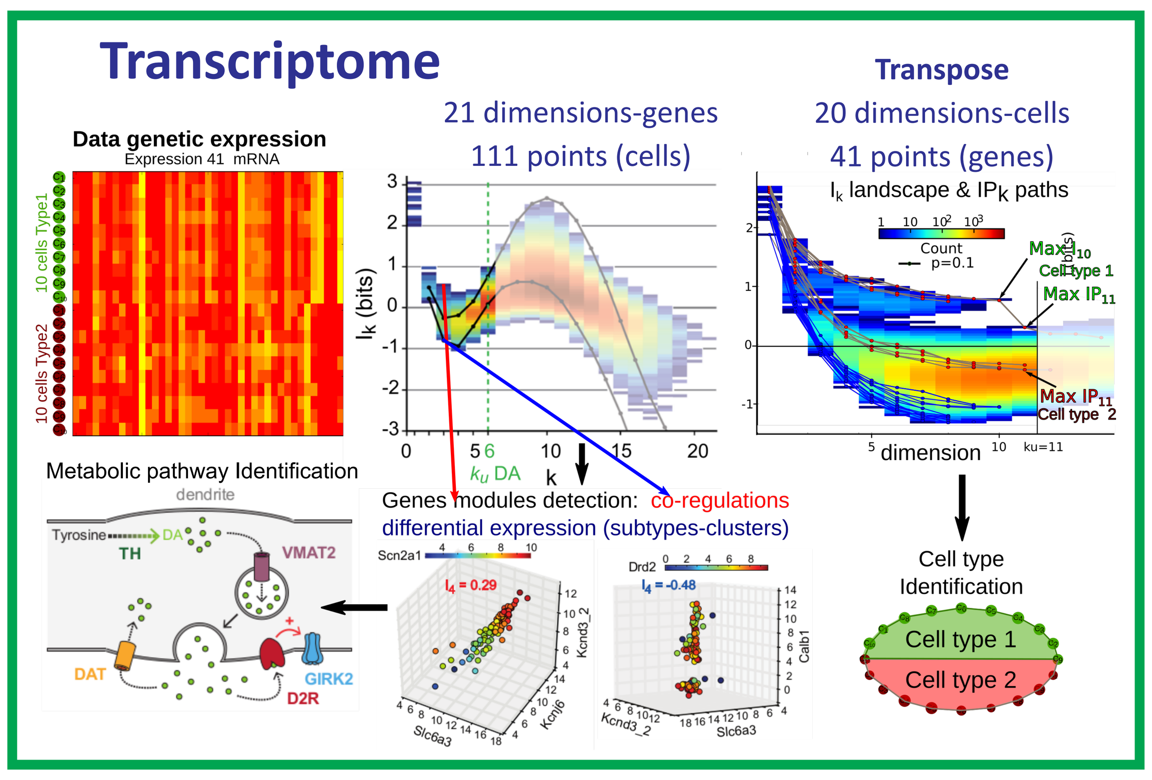


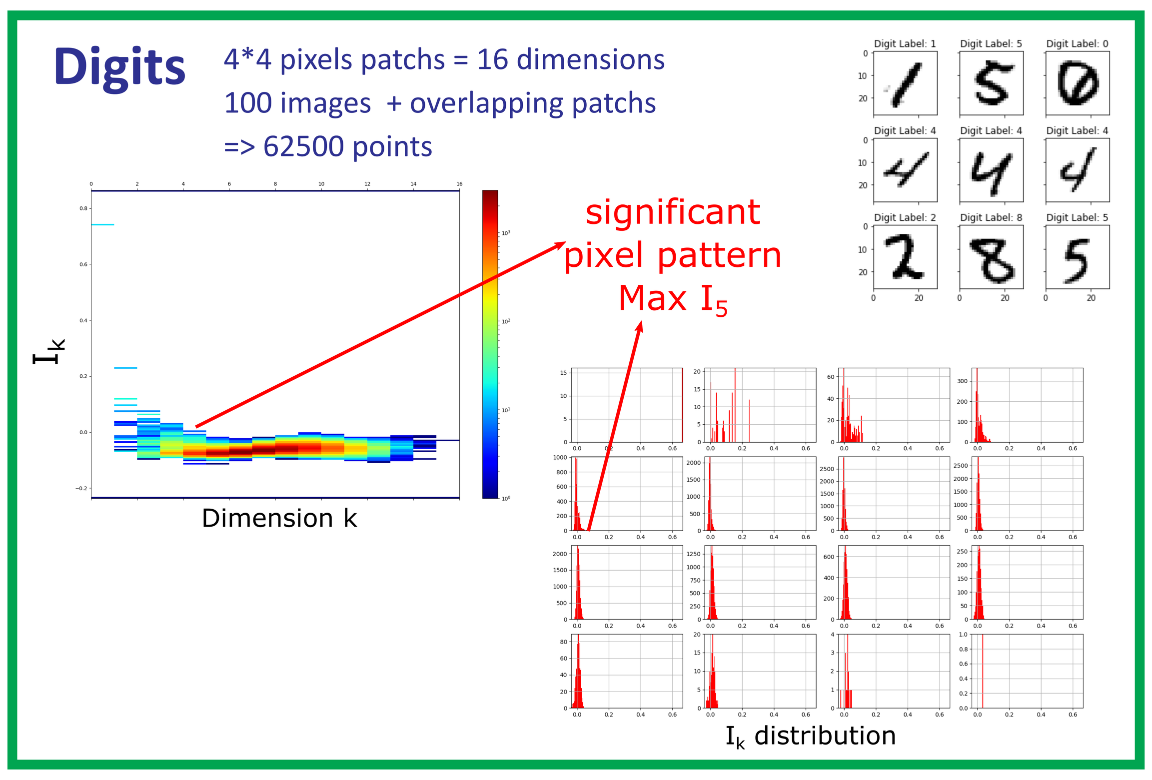


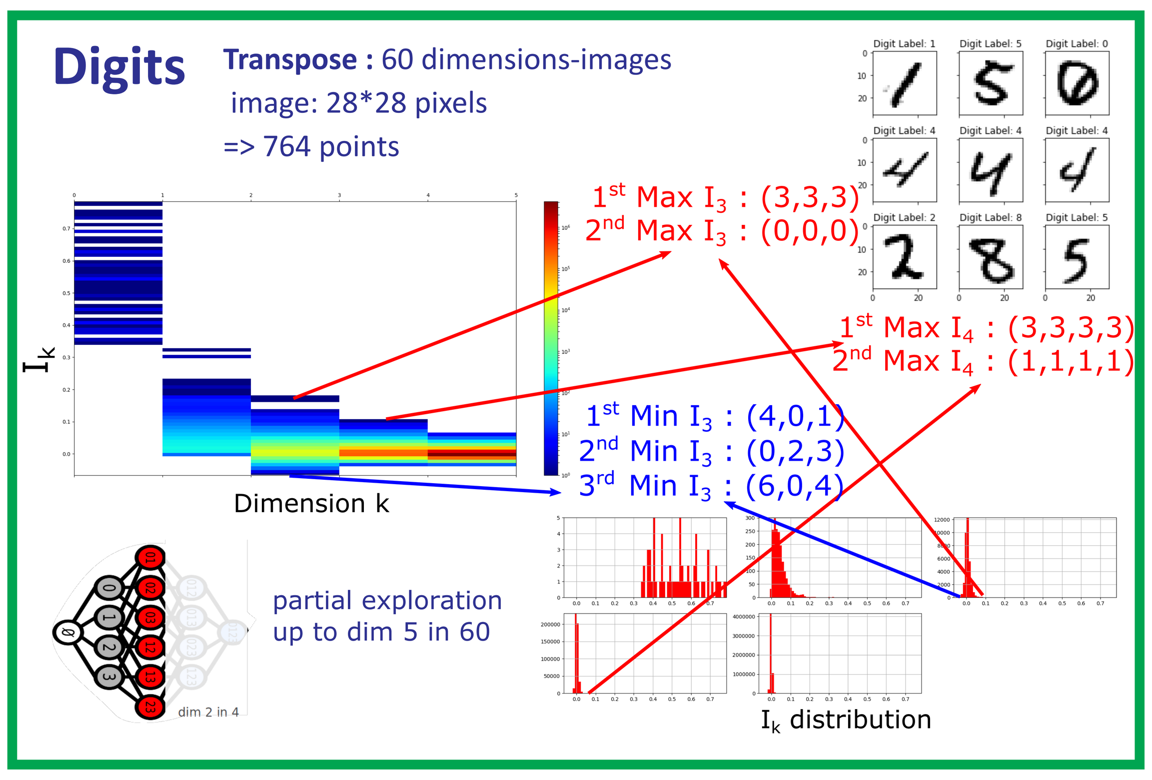
The marginal I_1 component defines a self-internal energy functional U_k, and **$I_{k,k>1}$** define the contribution of the **k-body interactions to the free energy functional G_k** given by the KL divergence between marginals and the joined variable (the "total correlation"). The set of information paths in simplicial structure is in bijection with the symmetric group.

Unsupervised

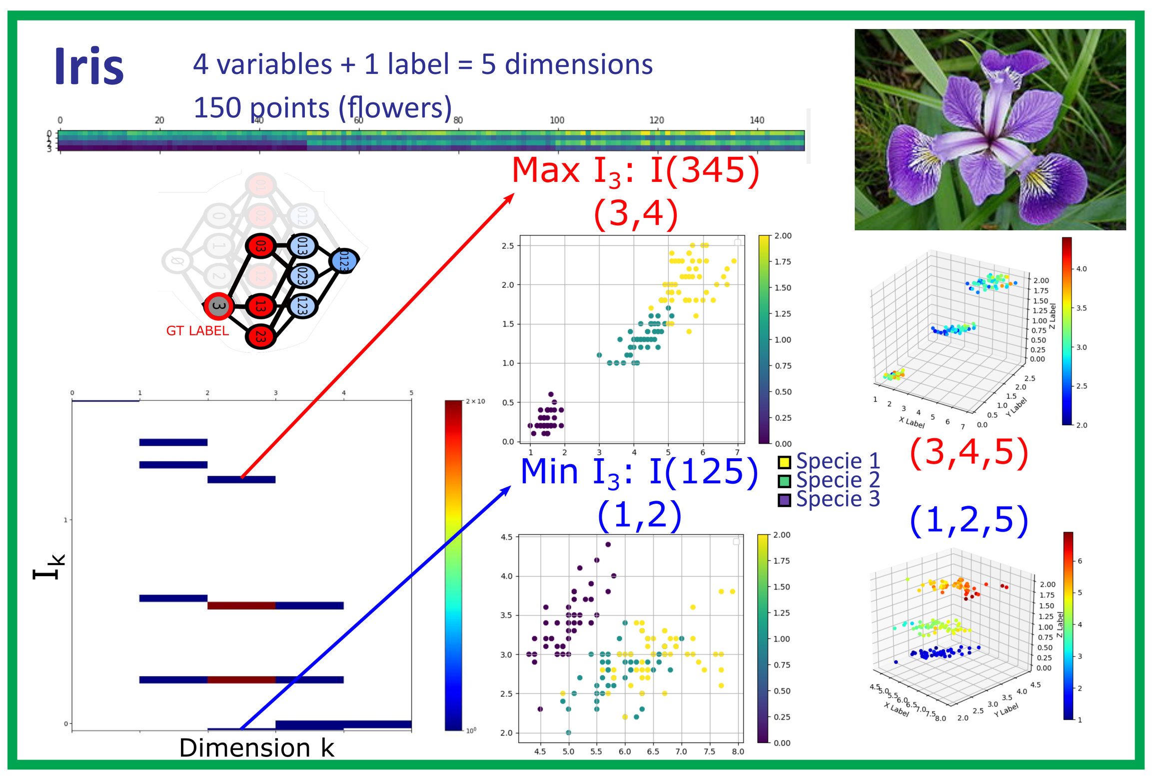


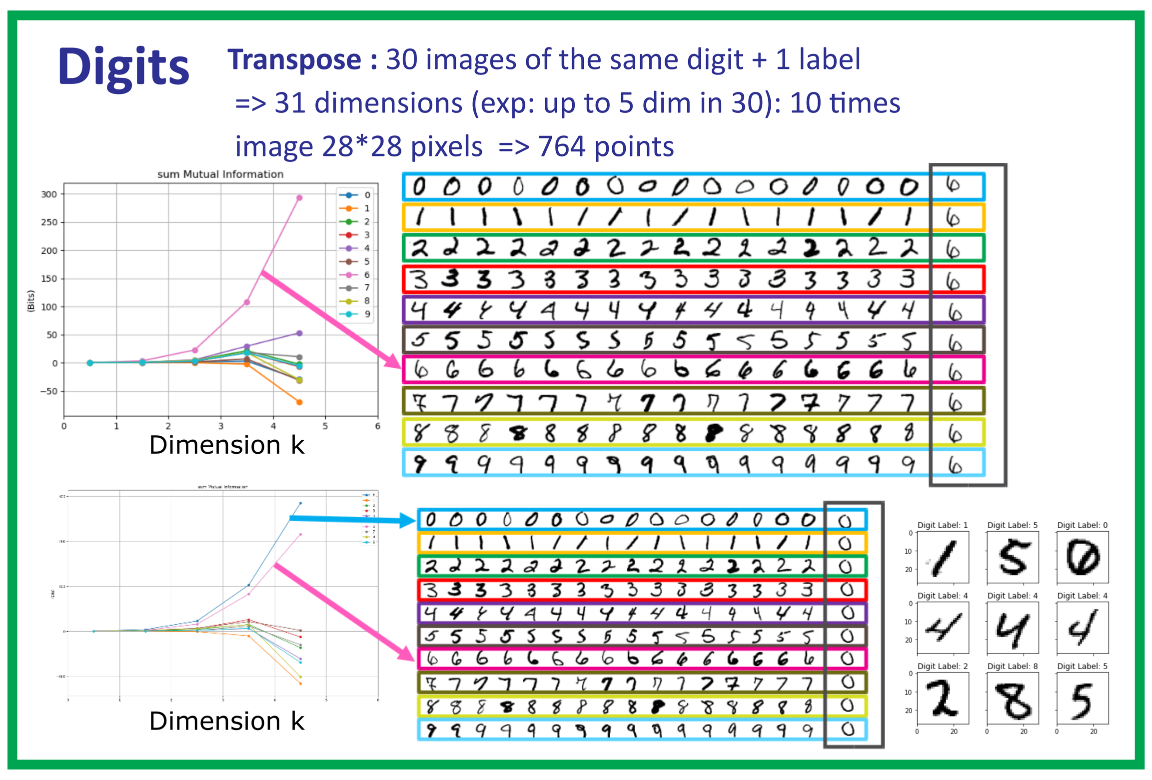






Supervised





Conclusions

1. New methods for topological data analysis intrinsicaly based on statistics. Encouraging results on data to be confirmed on larger training set and comparted to deep networks.
2. Generalization and formalisation of Neural Networks (binary variables) with algebraic topology. New formalization of supervised learning as a subcase of supervised learning.
3. Computationaly expensive O(2^n) or C(k,n) in partial exploration: current development of parrallell and GPU processing of the programs.

Bibliography

[1] P. Baudot and D. Bennequin. The homological nature of entropy. Entropy, 17(5):3253–3318, 2015.
[2] J.P. Vigneaux. Topology of statistical systems: a cohomological approach to information theory. PhD Thesis, 2019.
[3] M. Tapia & al; Neurotransmitter identity and electrophysiological phenotype are genetically coupled in midbrain dopaminergic neurons. Sci; Reports. 2018.
[4] P. Baudot, Tapia M., Goaillard J.M., Topological Information Data Analysis: Poincare-Shannon Machine and Statistical Physic of Finite Heterogeneous Systems. 2017 preprints201804.0157.v1