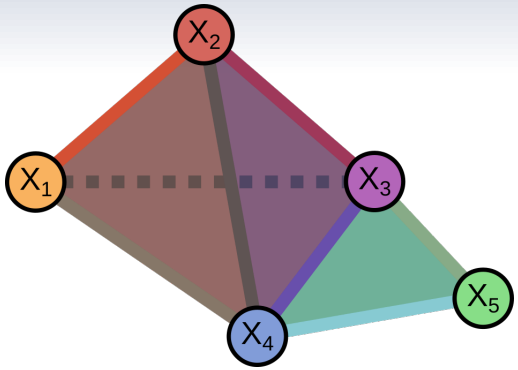


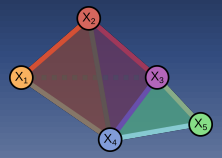
Information topological analysis, statistical structures of complex data

Pierre Baudot - Median Technologies - Inserm

in col. with Bennequin, Tapia, Goaillard

Jan 30, 2019
Data Science Meetup Nice
Learning Center Polytech

"When you use the word information, you should rather use the word form" R.Thom

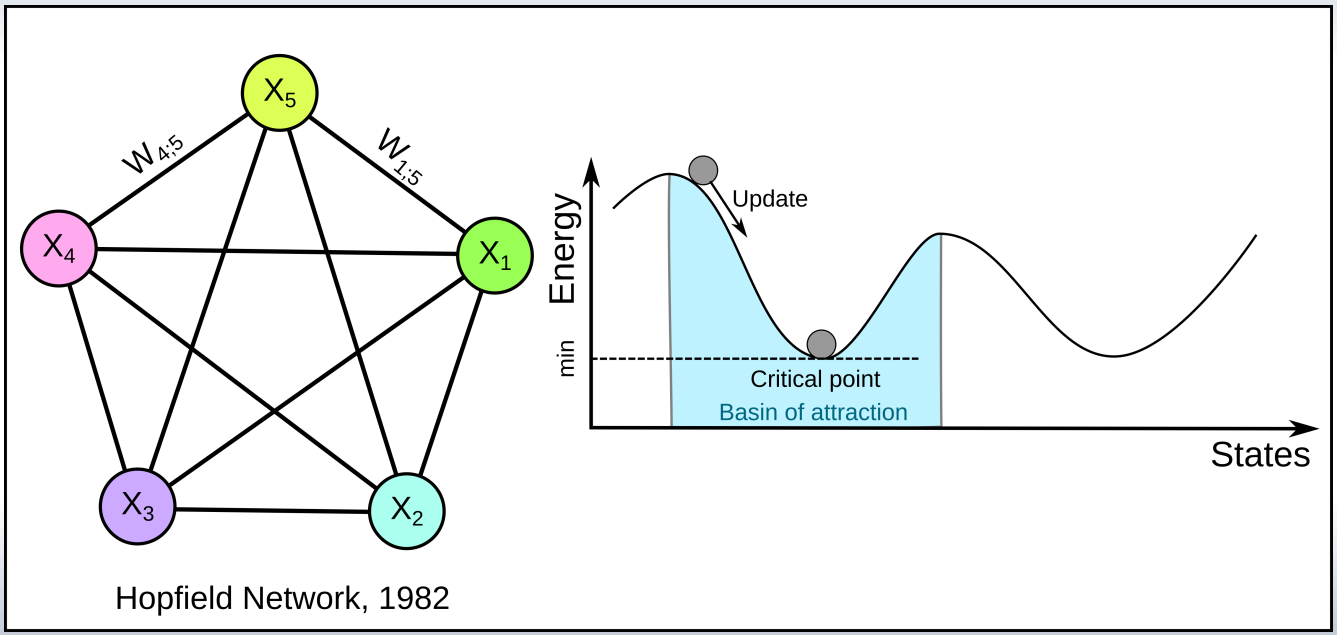


Contents

Neuroscience - Cognition

Cognition: Neural Network - Machine Learning (unsupervised):

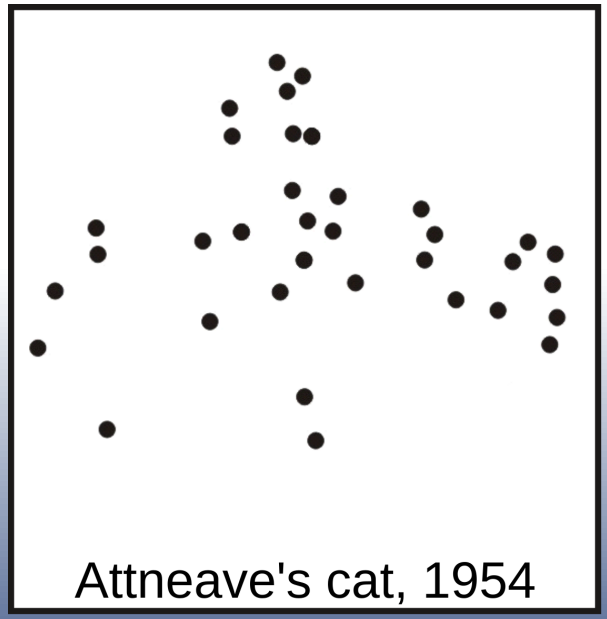
Hopfield
Hinton
Sejnowski
(Boltzmann - Helmholtz machines)
...

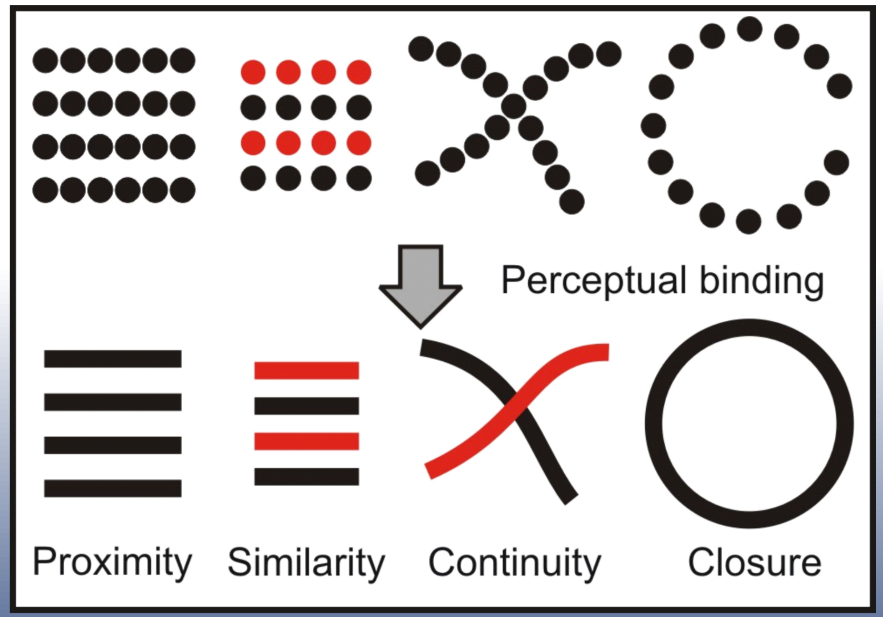


Neuroscience - Cognition

Neuroscience: Learning - Adaptation - Information sensory processing. *"Understanding is compressing"* Chaitin. **Efficient coding** (Attneave, 1952): the goal of sensory perception is to extract the redundancies and to find the most compressed representation of the environment. Any kind of symmetry and invariance are information redundancies and Gestalt principles of perception can be defined on information theoretic terms.

- Gestalt
- Barlow
- Attneave
- Laughlin
- Linsker
- Atick
- Nadal
- Sejnowski
- Bialek...

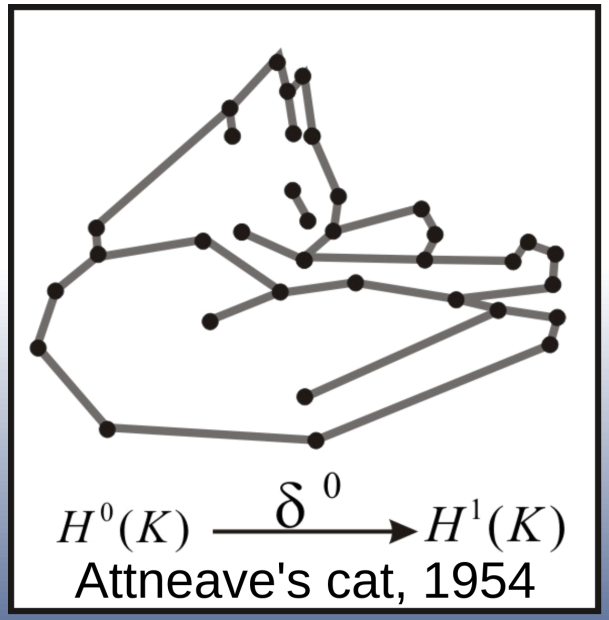




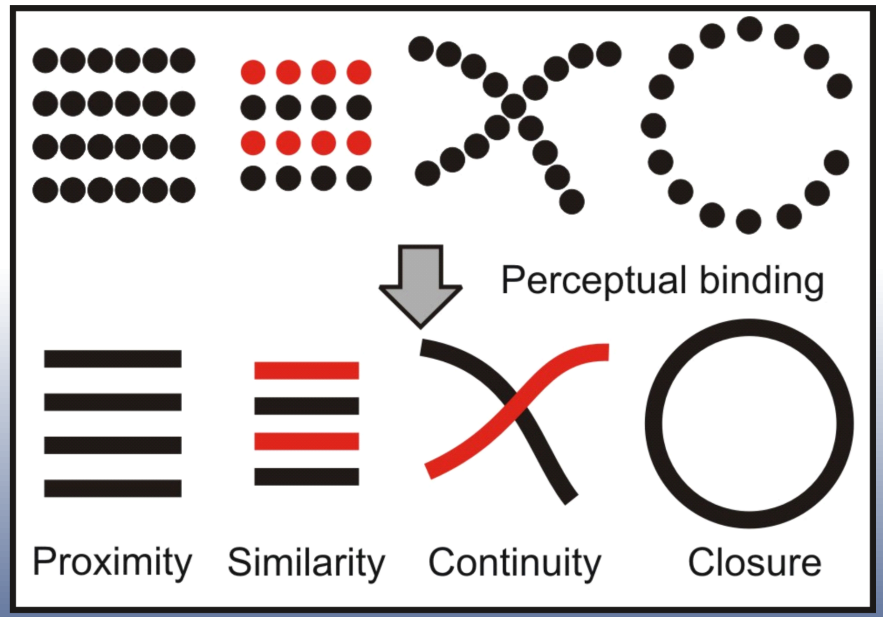
Neuroscience - Cognition

Neuroscience: Learning - Adaptation - Information sensory processing. *"Understanding is compressing"* Chaitin. **Efficient coding** (Attneave, 1952): the goal of sensory perception is to extract the redundancies and to find the most compressed representation of the environment. Any kind of symmetry and invariance are information redundancies and Gestalt principles of perception can be defined on information theoretic terms.

- Gestalt
- Barlow
- Attneave
- Laughlin
- Linsker
- Atick
- Nadal
- Sejnowski
- Bialek...



Attneave's cat, 1954



Biology

Biology: Development - Evolution - Morphogenesis:

Waddington
Thom
Wieschaus
...

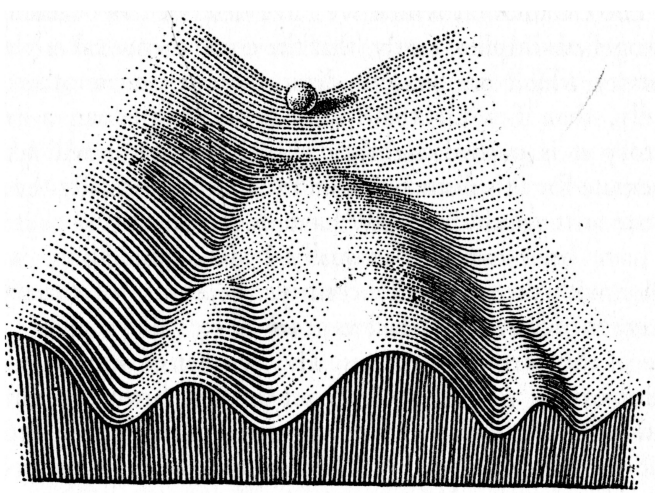


FIGURE 4
Part of an Epigenetic Landscape. The path followed by the ball, as

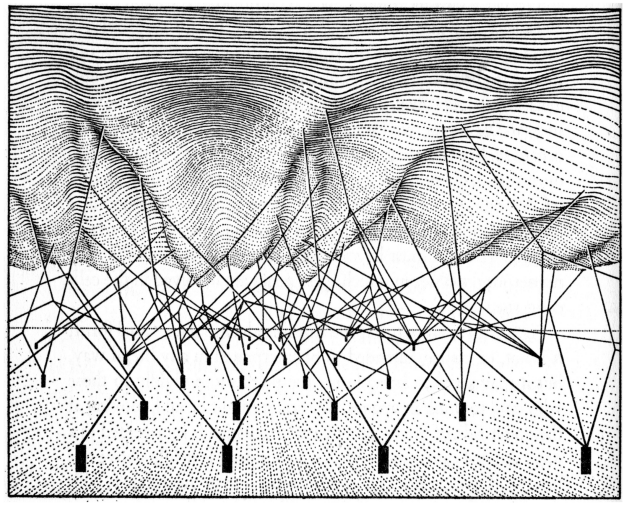
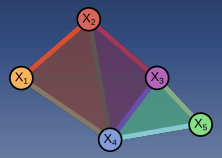


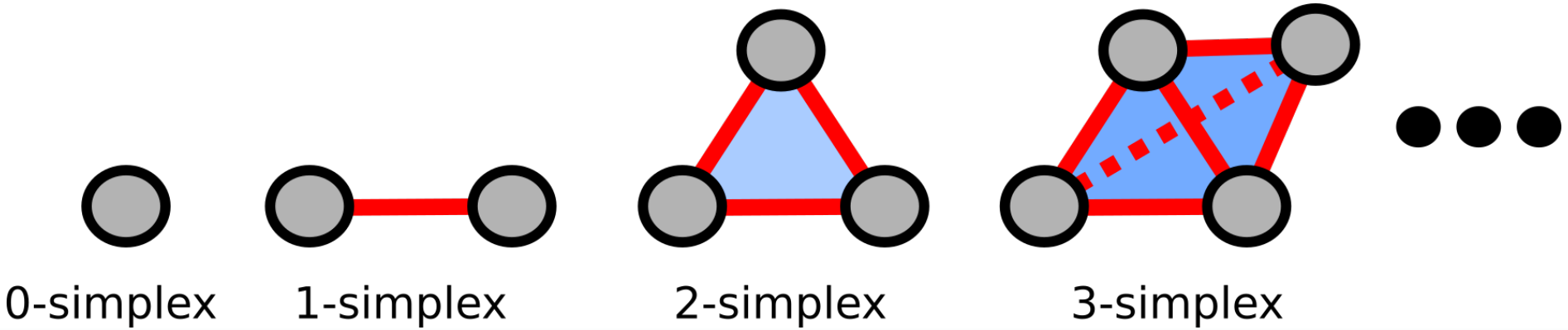
FIGURE 5
The complex system of interactions underlying the epigenetic landscape.

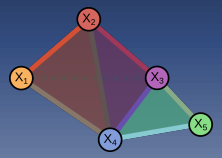
C.H. Waddington, 1957



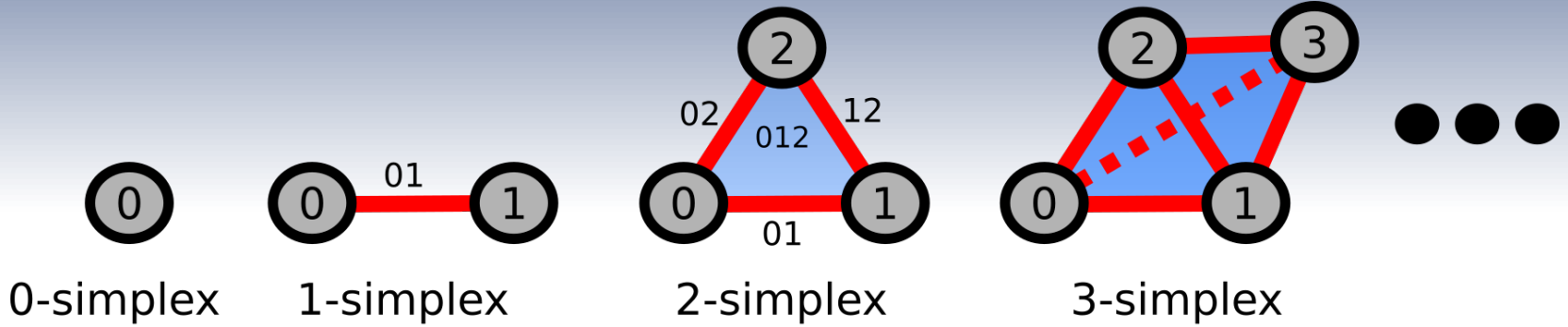
Simplex

k-simplex: k dimensional "triangle"

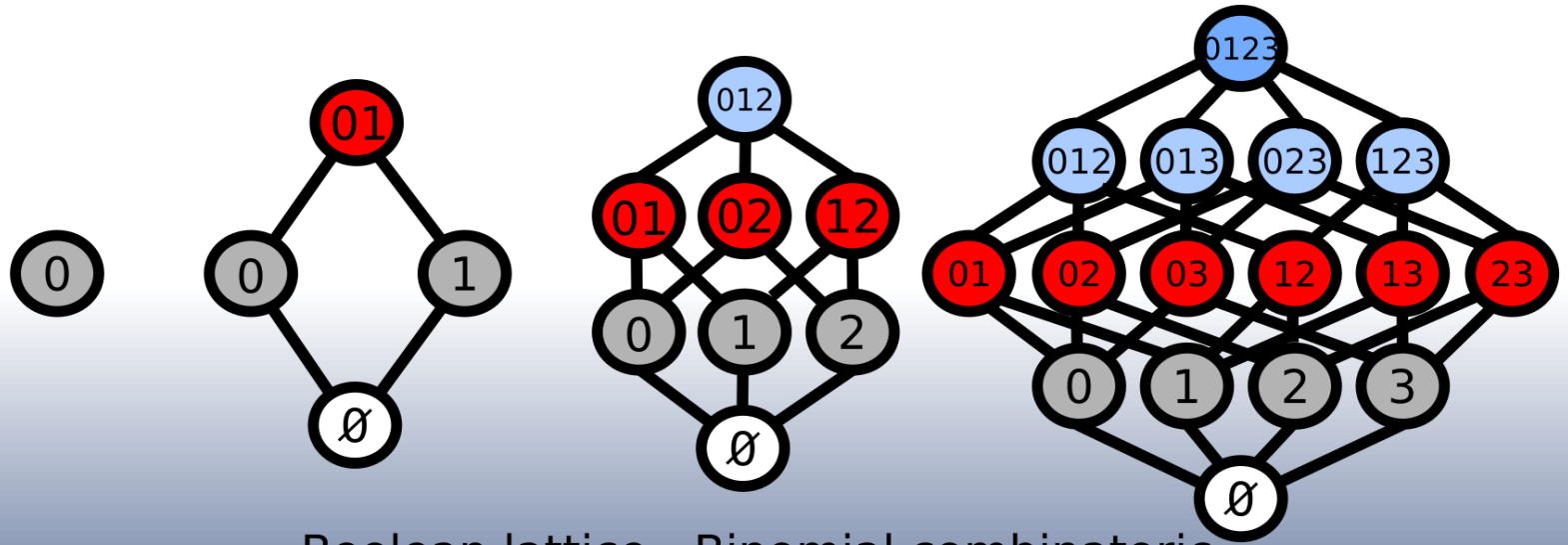




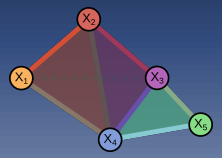
Simplex



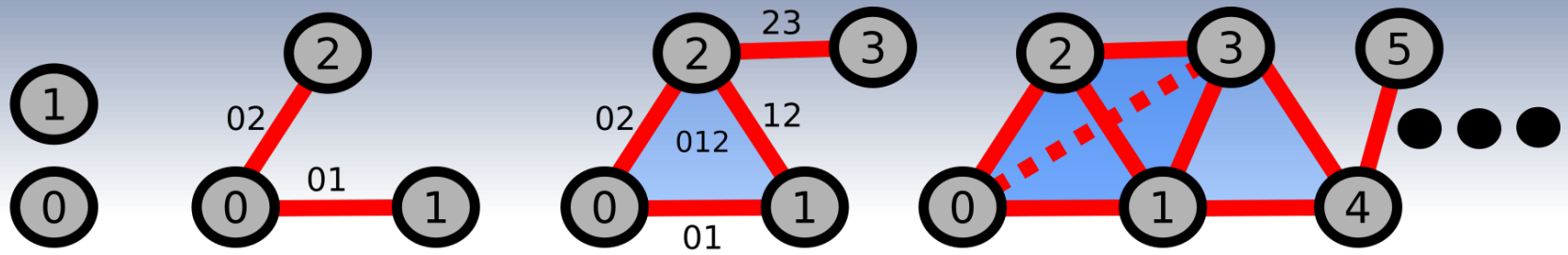
0-simplex 1-simplex 2-simplex 3-simplex



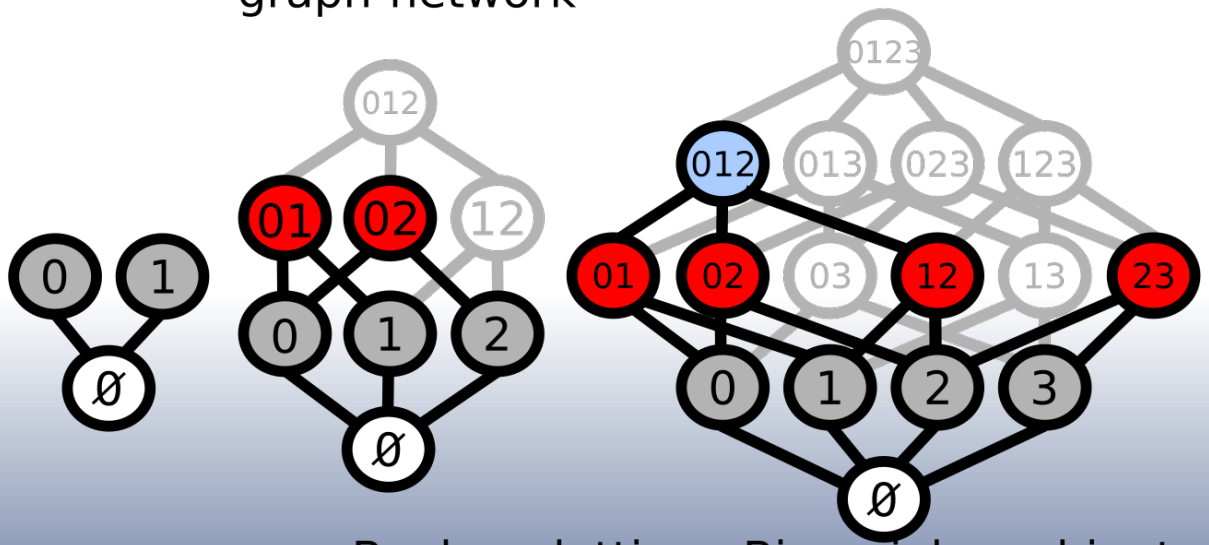
Boolean lattice - Binomial combinatoric



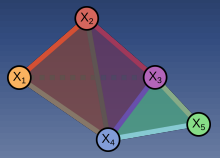
simplicial complex



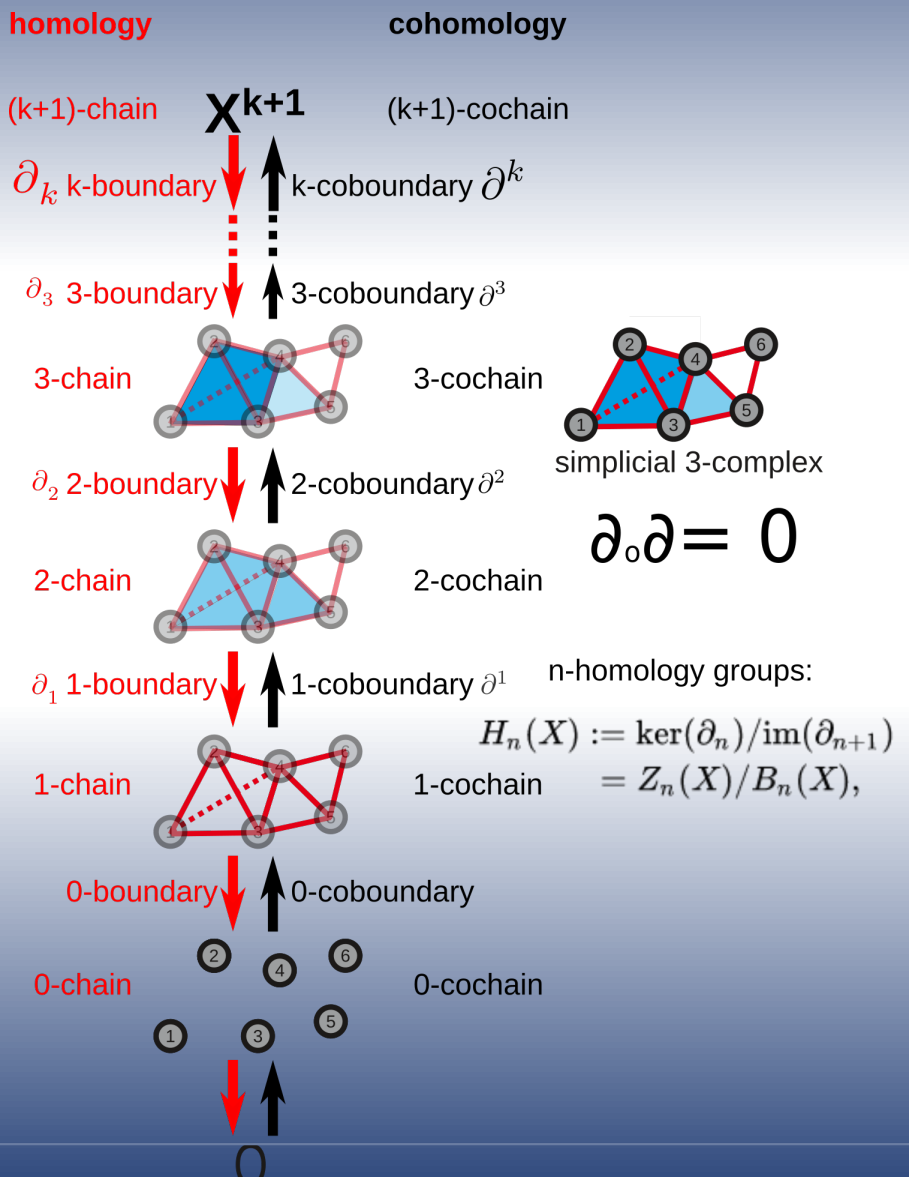
0-complex 1-complex graph-network 2-complex 3-complex

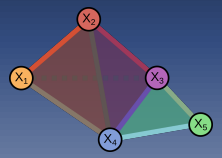


Boolean lattice - Binomial combinatoric



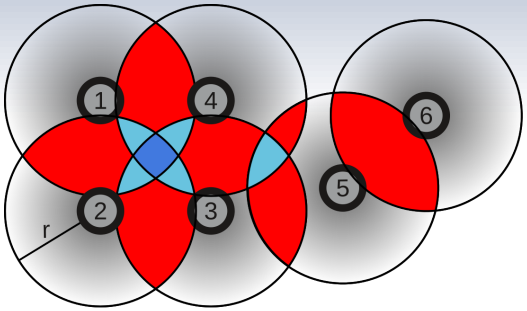
simplicial (co)-homology



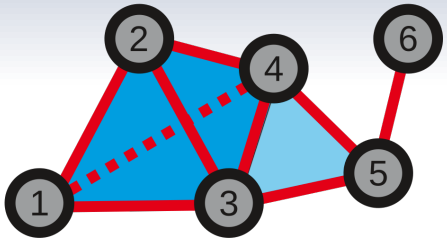


data - persistence homology

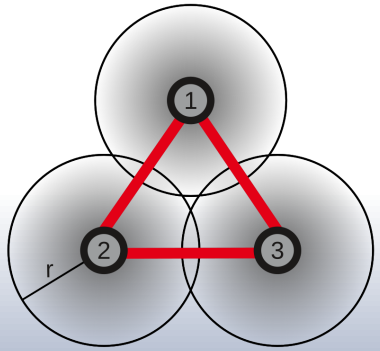
Persistence: ball for each points, make r vary, compute Vietoris complex's homology



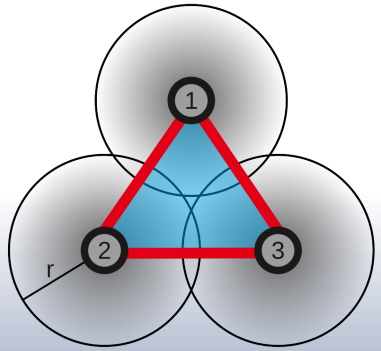
Data points with balls



Čech 3-complex



Čech complex
all intersections



Vietoris-Rips complex
only pairwise intersections

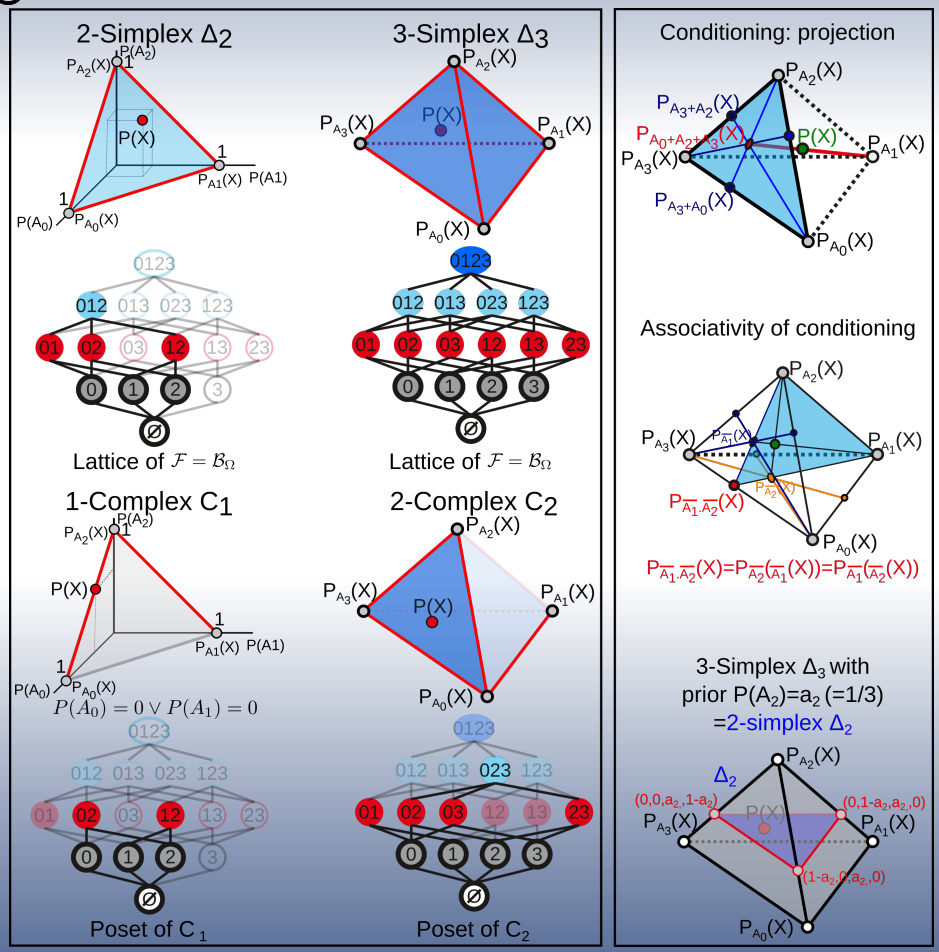
$\Rightarrow$ metric assumption (r), pairwise-graph approximation, not probabilistic

$\Rightarrow$ find a probabilist homology probabilistic without assomptions?

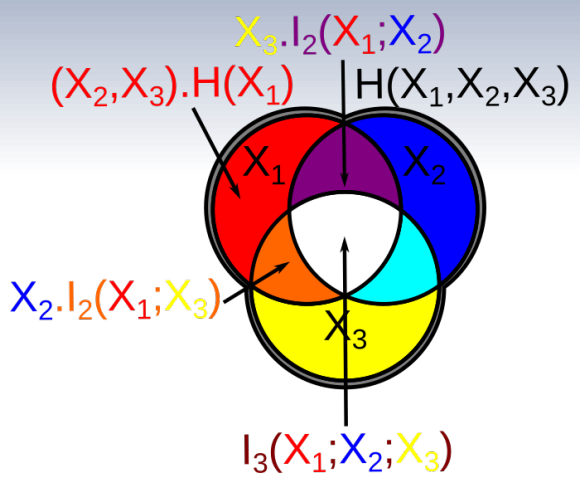
Information structures - probability simplex

The probability space $(\Omega, \mathcal{B})$, $|\Omega| = N$ is a $(N-1)$-simplex of probability, implementing geometrically Kolmogorov axiomatic:

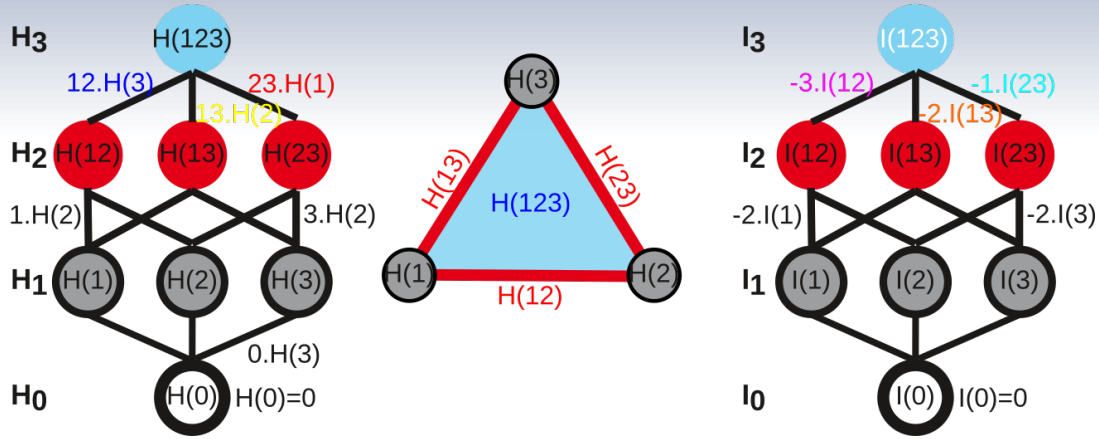
- $\sum_i P(A_i) = 1$ the geometry is affine
- $P(A_i) \geq 0$ convex
- Theorem of total probability: barycentric coordinate $P(X) = \sum_i P(A_i.X) = \sum_i P(A_i).P_{A_i}(X)$
- Conditioning is a projection on subsimplex.
- Complex of probability given by set of constraints of the form $P(A_0) = 0 \vee P(A_1) = 0$



Information functions - Entropy



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



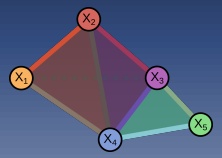
1-entropy ($k=-1/\ln 2$, bit): $H_1=H(X_j;P)=k\sum_{x\in[N_j]} p(x)\ln p(x)$

k-joint entropy: $H_k=H(X_1,...,X_k;P)=k\sum_{x_1,...,x_k\in[N_1\times...\times N_k]}^{N_1\times...\times N_k} p(x_1.....x_k)\ln p(x_1.....x_k)$

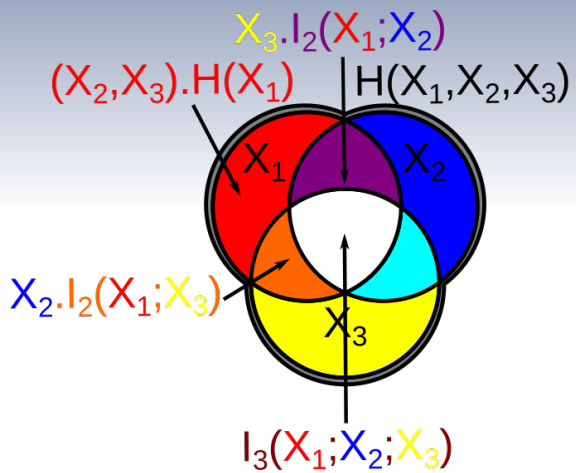
Conditional entropy: $X_2.H_1=H(X_1|X_2;P)=k\sum_{x_1,x_2\in[N_1\times N_2]}^{N_1*N_2} p(x_1.x_2)\ln p_{x_2}(x_1)$

Chain rule entropy : $H_{k+1}-H_k=(X_1,...X_k).H(X_{k+1})$

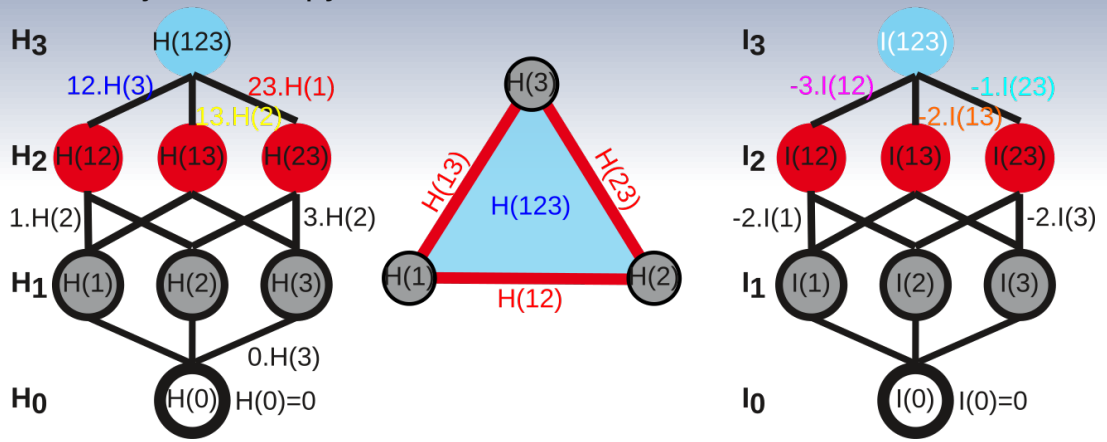
Non-negative functions (classical).



Information functions - Mutual-Informations



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



2-Mutual-Information: $I_2=I(X_1;X_2;P)=k\sum_{x_1,x_2\in[N_1\times N_2]}^{N_1\times N_2} p(x_1.x_2)\ln\frac{p(x_1)p(x_2)}{p(x_1.x_2)}$

k-Mutual-Information (for $k\geq 3$, I_k can be negative, $I_1=H_1$):

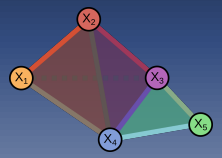
$I_n(X_1,\ldots,X_n;P)=\sum_{i=1}^{n}(-1)^{i-1}\sum_{I\subset[n];card(I)=i} H_i(X_I;P)$,

ex: $I_3=H(1)+H(2)+H(3)-H(1,2)-H(1,3)-H(2,3)+H(1,2,3)$

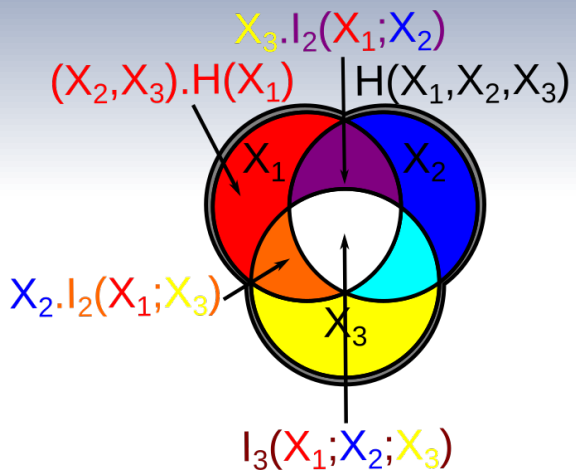
Conditional MI: $X_3.I_2=I(X_1;X_2|X_3;P)$

Chain rule MI : $I_{k-1}-I_k=X_k.I_{k-1}$

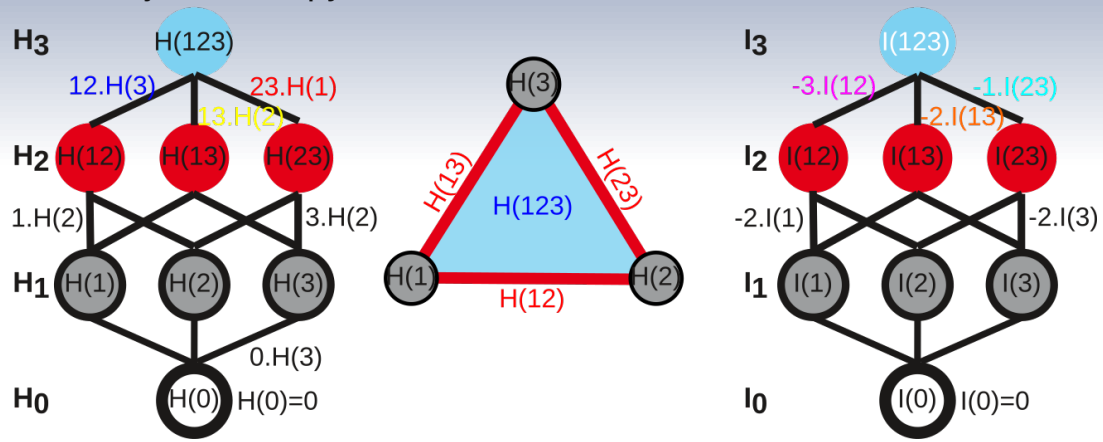
Total correlation ($C_2=I_2$ KL-div ≥ 0): $C_k=C_k(X_1;\ldots X_k;P)=\sum_{i=1}^{k} H(X_i)-H(X_1;\ldots X_k)$



Information functions - Mutual-Informations

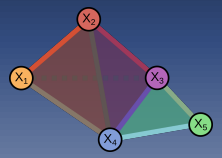


(semi) lattice and simplex of
joint-entropy and mutual-information functions



Theorem (Hu Kuo Ting)

Information functions are in bijection with finite additive (measurable) functions with operators $\cup$, $\cap$, / corresponding to Joint (;), Mutual (,) and conditional (/) information operation respectively.

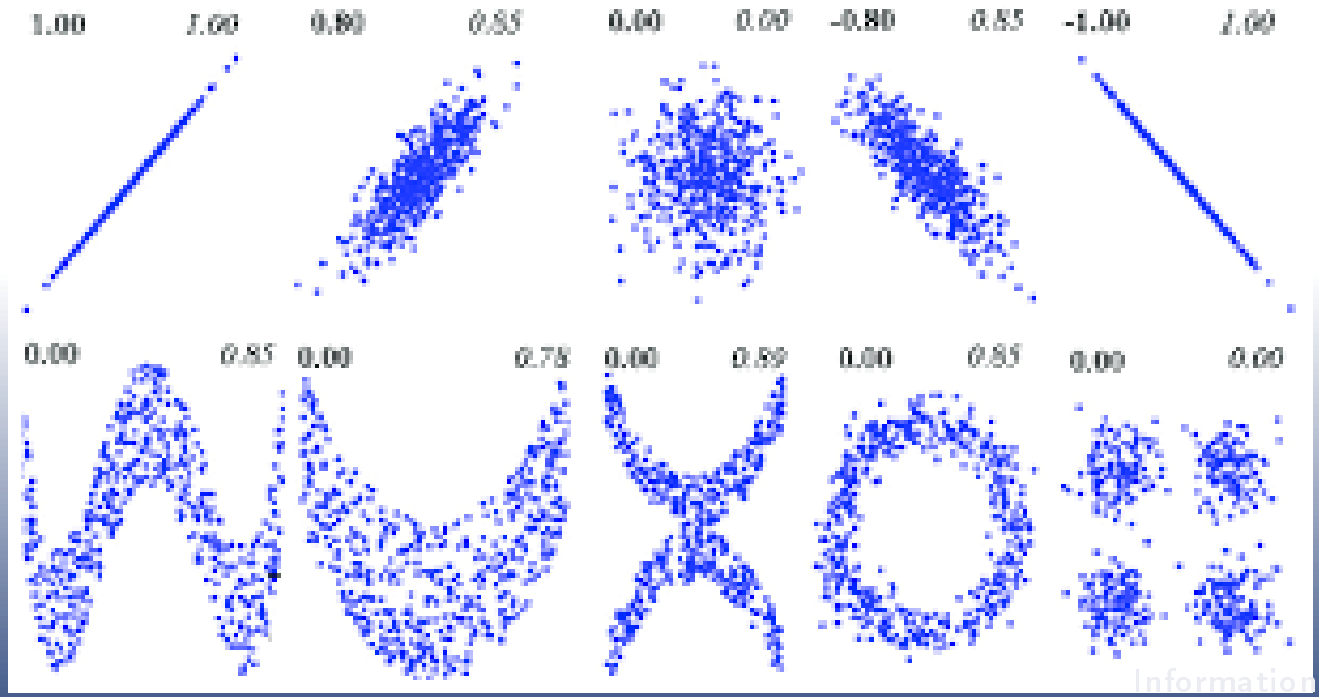


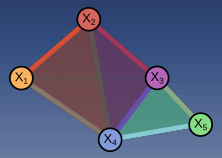
2-independence

Theorem 2-independence $\Leftrightarrow \partial^1_* = 0$ (Li, 1990)

X_1, X_2 are statistically independent if and only if $I_2 = I(X_1, X_2; P) = 0$

Moreover, $I(X_1,X_2)=0 \Rightarrow \rho_{X_1,X_2}=0$, $\rho_{X_1,X_2}=\frac{\mathrm{cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}}$ (detect non-linear)





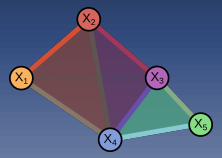
k-independence - generalization

Definition k-independence

$X_1, ..., X_k$ are k-independent if $I_k = 0$

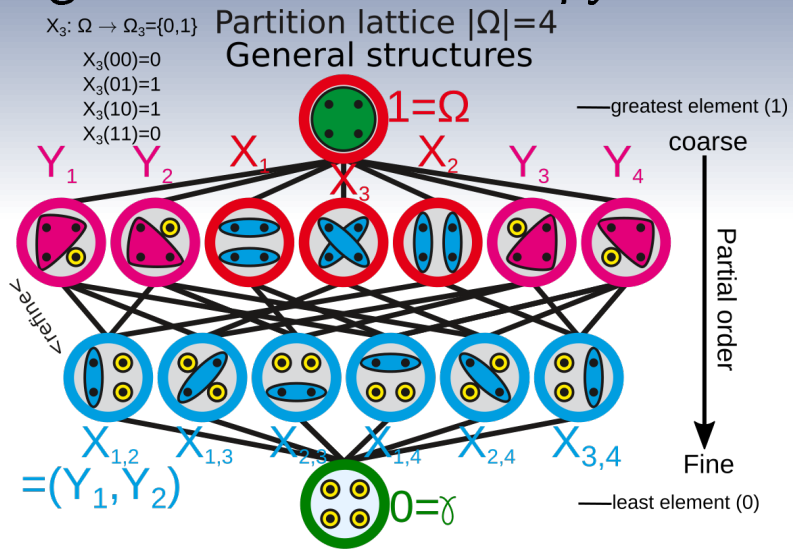
Theorem mutual-independence

$X_1, ..., X_n$ are mutually independent if and only if $\forall k \leq n$, $I_k = 0$.

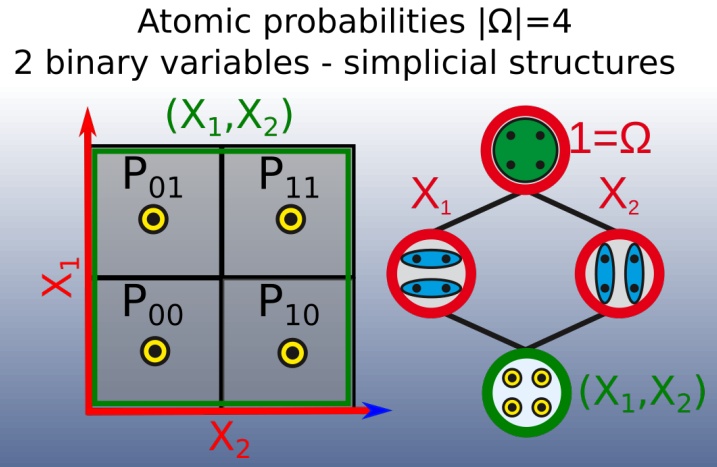


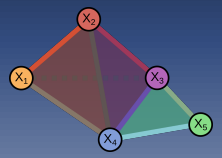
Information structures

Baudot and Bennequin, *The Homological nature of entropy*, Entropy, 2015.



- The random variables are **partitions** of the atomic probabilities of $(\Omega, \mathcal{B}, P)$ (equivalence classes).
- The **Joint-Variable** (X_1, X_2) is the less fine partition that is finer than X_1 and X_2 (gcd).





Actions and coboundaries

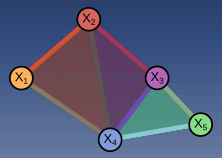
Conditioning-expectation by Y, $Y.F(X_1, ..., X_k; P)$, is the left action of Y on the functional module, $Y.F(X; P) = \sum_i P(Y = y_i)F(X; P_{Y=y_i})$. Complexes of random variables are $X^k = (X_1, ..., X_k; P)$, and we consider cochain complexes (X^k, ∂^k):

$$0 \to X^0 \xrightarrow{\partial^0} X^1 \xrightarrow{\partial^1} X^2 \xrightarrow{\partial^2} ...X^{k-1} \xrightarrow{\partial^{k-1}} X^k$$

where ∂^k is the Hochschild (or Galois) coboundary. For the first degree $k = 1$, we have the following results:

Main theorem (Baudot, Bennequin)

The information co-homology space of degree one is one-dimensional and generated by entropy.



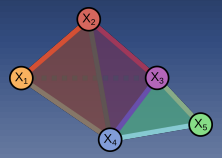
Coboundaries and higher I_k

Theorem - cocycle : independence

Let X^n be an information structure, then:

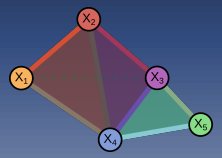
- For even degrees $2k$: $\partial^{2k} = -I_{2k+1}$ and $\partial_*^{2k} = -\partial_t^{2k} = 0$
- For odd degrees $2k+1$: $\partial^{2k-1} = 0$ and $\partial_*^{2k-1} = -\partial_t^{2k} = -I_{2k}$.

As a probabilistic interpretation, information cohomology quantifies statistical dependences at all degrees, the obstruction to factorization: k-independence coincides with cocycles.



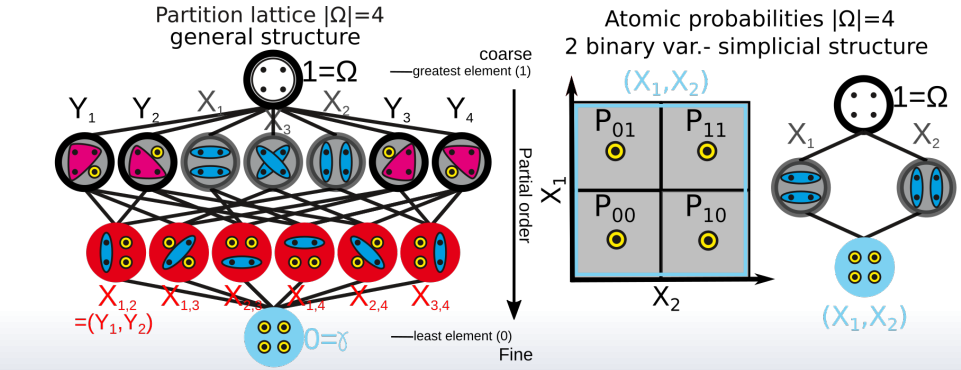
Simplicial information substructures

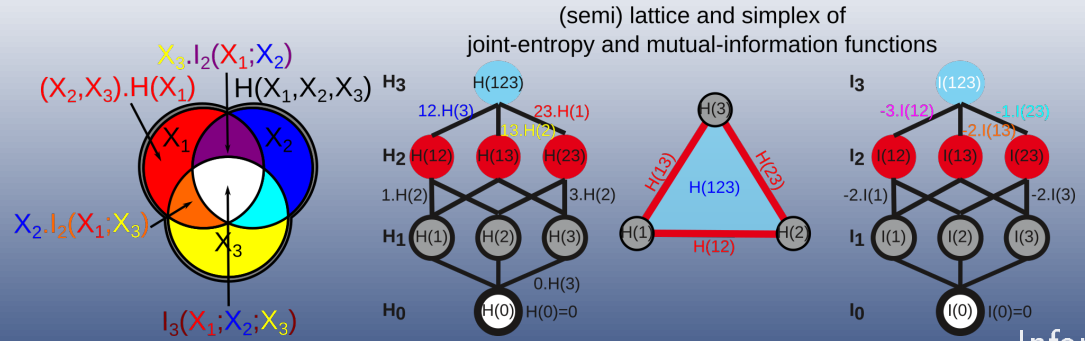
- **Computational problem:** complexity of the estimation of information functions: Bell's combinatoric $\mathcal{O}(exp(exp(N^n)))$ for n N-ary variables.
- **Computational solution**: Data analysis is developed on the simplest sub-case of the general information structure, the simplicial information structure and the simplicial information cohomology with complexity in $\mathcal{O}(2^n)$.
- **Consequence:** some possible statistical dependences cannot be detected.

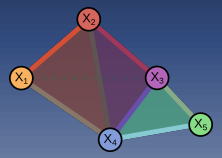


Simplicial information substructures

- **A simplicial information structure** is the triple (Ω, Δ^n, P) where Δ^n is the Boolean lattice of all subsets. A simplicial complex of random variables $X^k = (X_1, ..., X_k; P)$ is any subcomplex of the simplex Δ^n with $k \leq n$
- **Joint (X_1, X_2) and meet $(X_1; X_2)$ of variables** are the usual joint and meet of Boolean algebra and define two opposite-dual monoids.





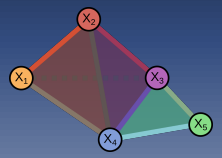


Free information energy n-body interaction

- **Internal information energy (definition):** for $k = 1$, I_1 and $\langle I_1 \rangle$ are a self-interaction $I(X_i) = H(X_i)$ that we call **internal information energy**. The total Internal energy is $E(X_1, ..., X_n; P_N) = \sum_{i=1}^{n} H(X_i)$
- **Free-information-energy (definition):** for $k > 1$ I_k quantifies the contribution of the k-body interaction, that we call the **k-free-information-energy**. The total free energy is the total correlation (Watanabe, Studeny) that quantify the total dependences $G_k = \sum_{i=2}^{k} (-1)^i \sum_{I \subset [n]; card(I)=i} I_i(X_I; P)$.

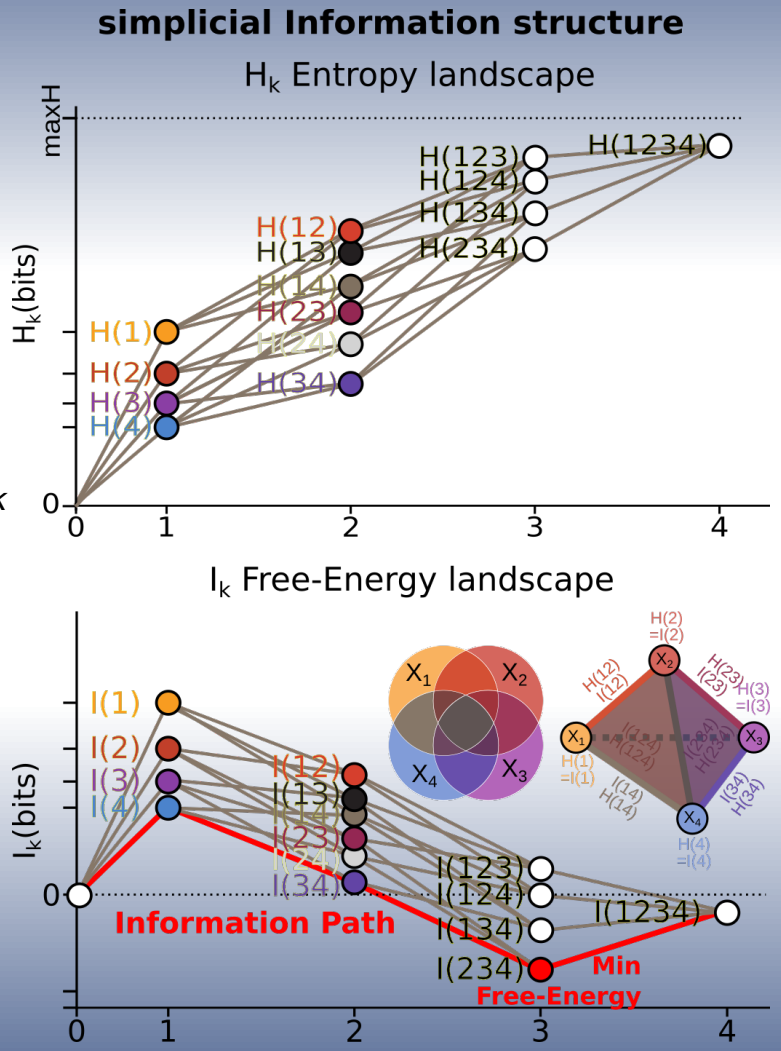
We recover the usual isotherm thermodynamic relation in the special case of Gibbs distribution $p(X_1 = x_1, ..., X_n = x_n) = p_{\underbrace{ij...n}_{\text{n indices}}} = \frac{1}{Z} e^{-E_{ij...n}/k_B T}$:

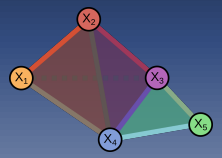
$$H_n(X_1, .., X_n; P_N) = E(X_1, .., X_n; P_N) - G(X_1, .., X_n; P_N) = E - G$$



H_k and I_k landscapes and paths

- **Information landscapes:** (semi)lattice of information as a function of the values of H_k and I_k (gives a ranking). H_k quantify variability I_k quantify statistical dependences.
- **Information path:** entropy paths HP_k and MI paths IP_k: sequence of edges, piecewise linear-functions.
- **First derivative of entropy path** is conditional entropy: $dHP_i(k)/dk = (X_1, ..., X_{k-1}).H(X_k)$, of mutual information path is minus conditional information (coface map): $dIP_i(k)/dk = -X_k.I(X_1, ..., X_{k-1})$





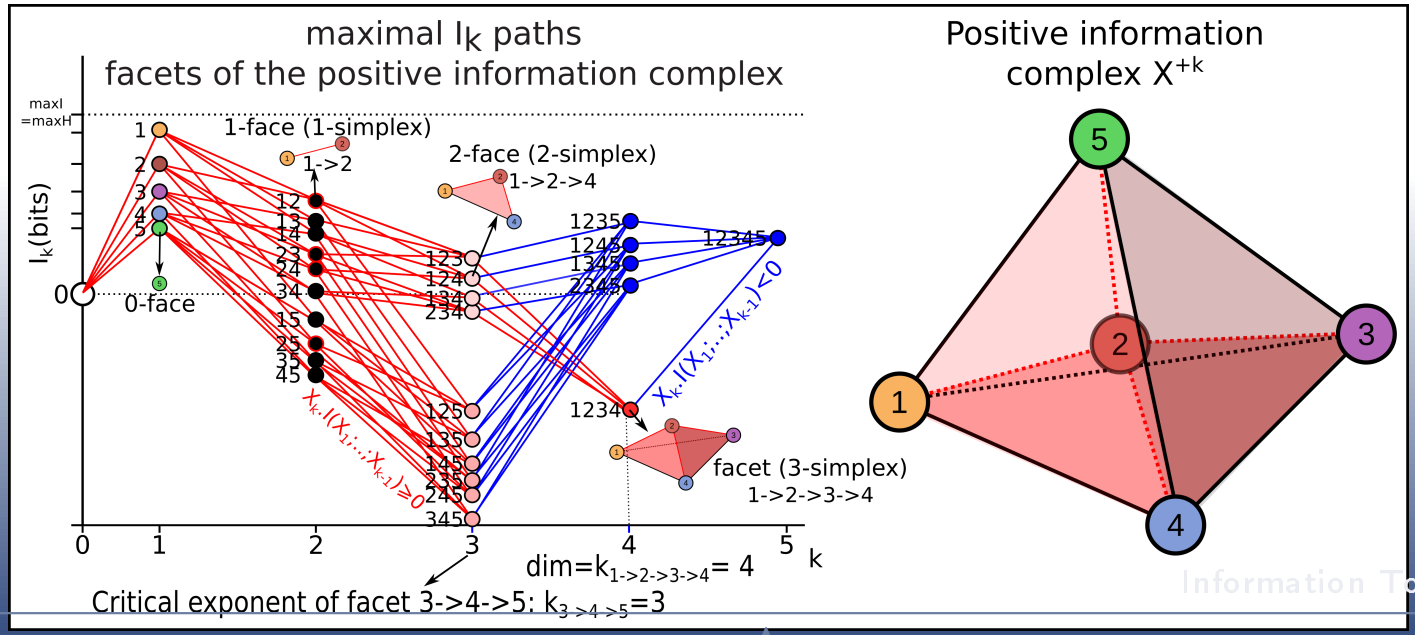
Minimum free energy complex

Positive information path: an IP_k such that $I_k < I_{k-1} < ... < I_1$.

Theorem Minimum free energy complex

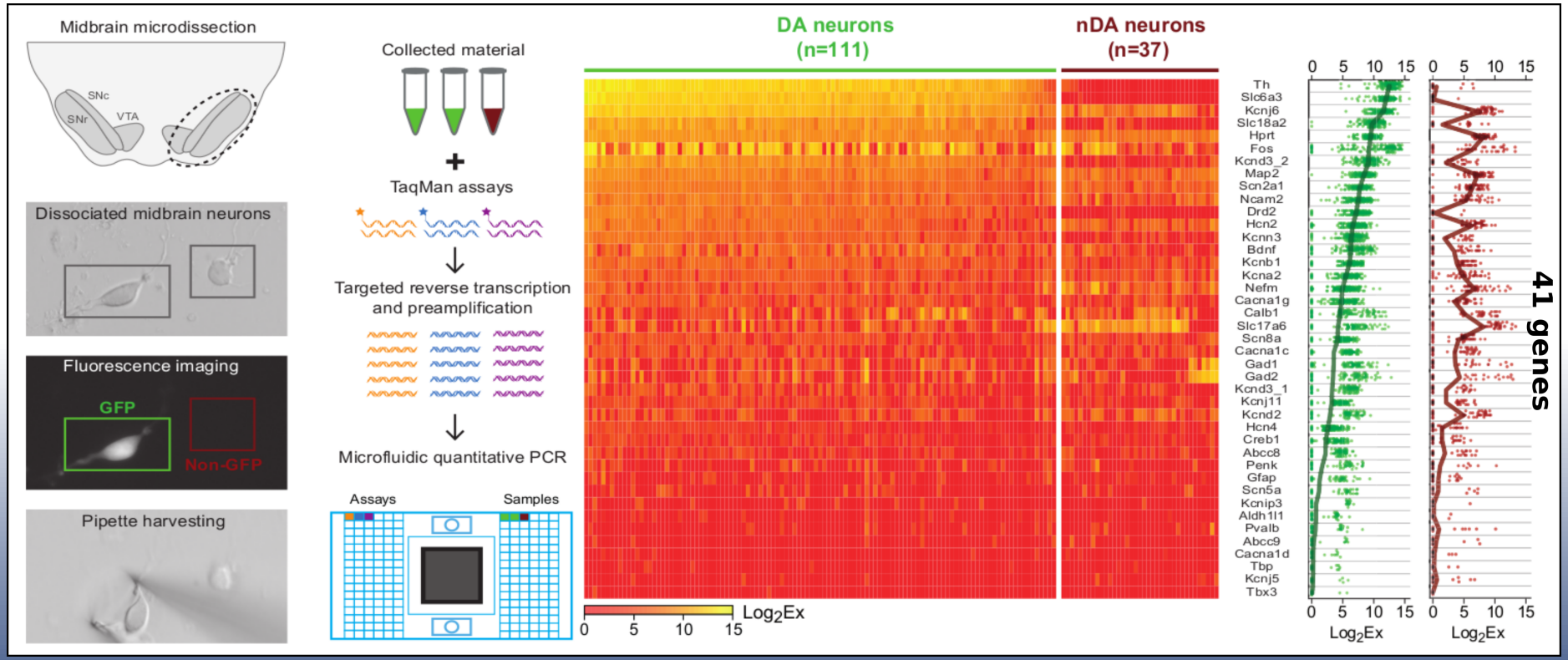
The set of all positive informations paths forms a simplicial complex. A necessary condition for this complex not to be a simplex is $d \geq 4$.

(**Minimum free energy principle with degeneracy** = complex system)

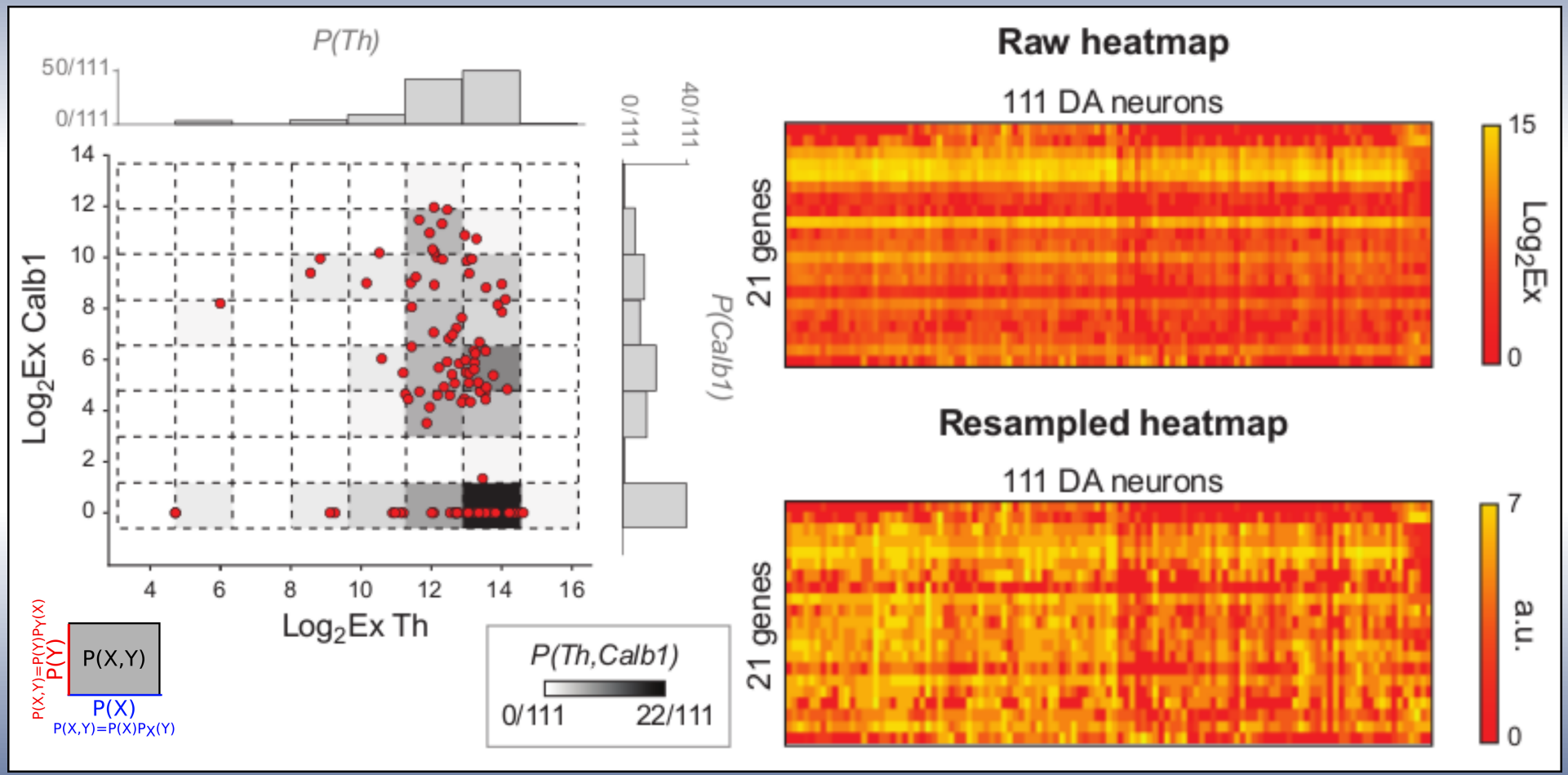


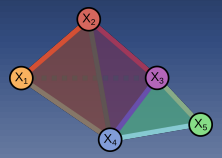
Gene expression measures

- Quantitative PCR of single neurons in SNc (dopaminergic) and other midbrain nucleus (nDA)
- mRNA expression levels for $n = 41$ genes in $m = 111$ DA and $m = 37$ nDA



Probability estimation



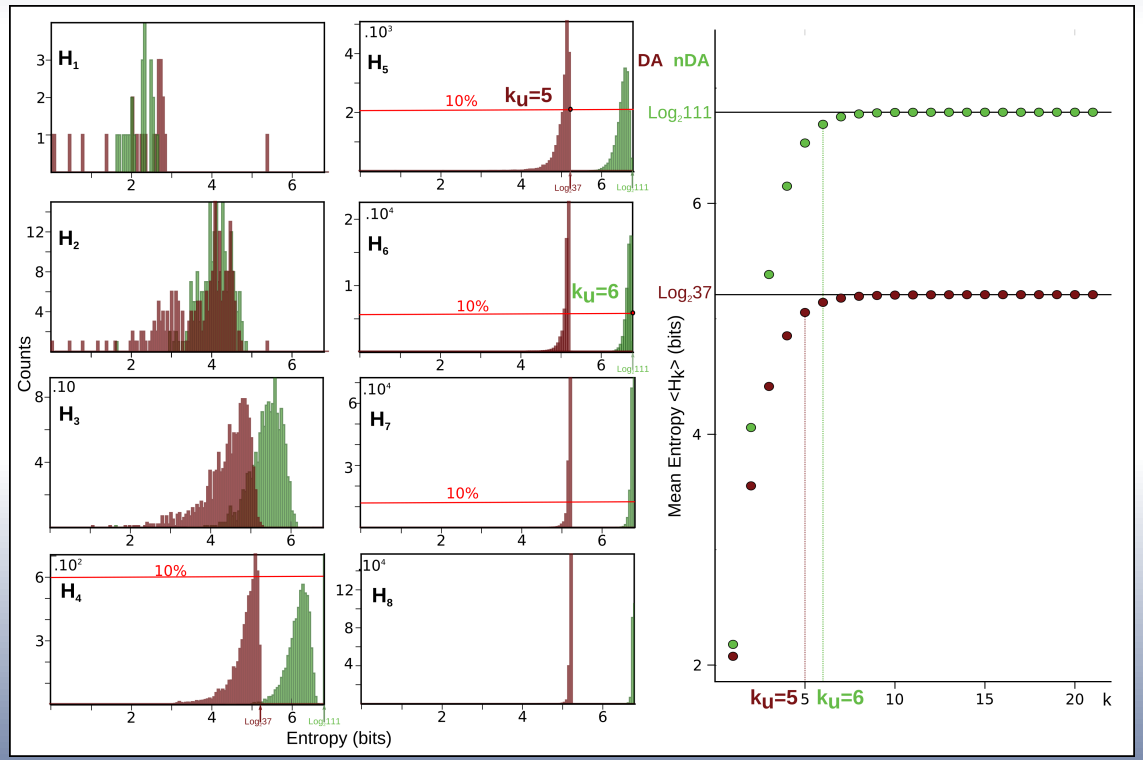


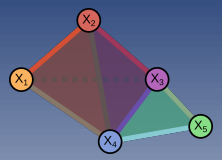
Finite size - sampling problem

Dependence on m. Undersampling (curse of dimension/sampling problem): when $N_1 \ldots N_n$ are such that only one data point falls in a box then $p = 1/m$ and $H_n = \log_2 m$.

- Degree k_u for which more than 10% of the H_k are in $\log_2 m - 0.05 \leq H_k \leq \log_2 m$.
- Analysis holds well bellow usual undersampling regime.

Computational restriction to $n = 21$ ($2^{21} \approx 2.10^6$ elements)



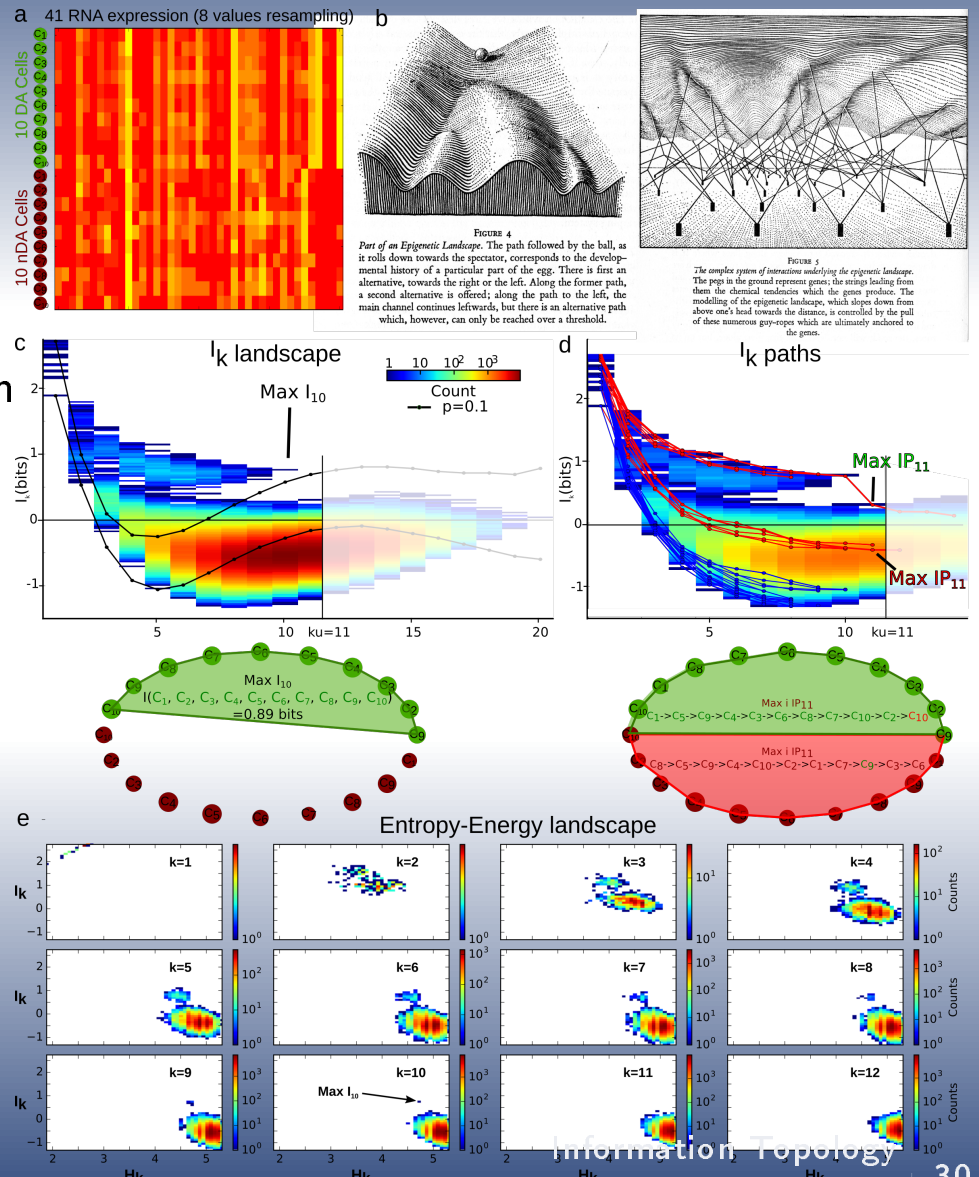


Computation of the Minimum free energy complex

- **Computational problem:** finding a global functional extrema or all the first critical dimension is NP-hard class ($\mathcal{O}(n!)$).
- **Computational solution:** At each element of the lattice, we start at one of the I_1 and at each element of the paths we explore only the two paths with lowest and highest positive values of $X_{k+1}.I(X1;..;X_k)$ (local), and iterate until it stops at the minima (whenever the conditional mutual information starts to be negative) and then rank the paths as a function of their length. It finds the maximal positive information paths that have highest and lowest I_k values at each element of a path. Computational complexity in $\mathcal{O}(n)$ but only give a **partial estimation of the minimum free energy complex** (can be richer and greater dimensionality).

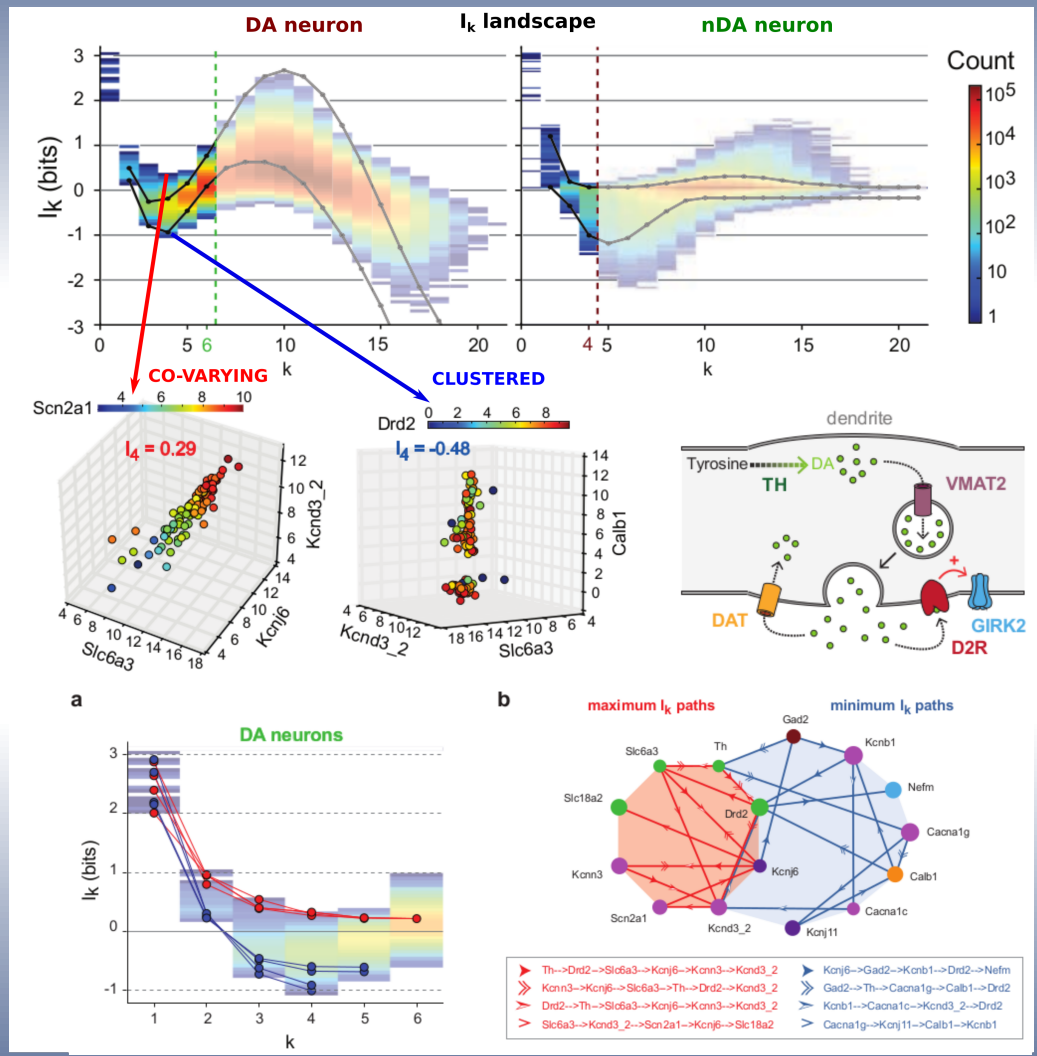
Cell type identification

- 10 DA and 10 nDA neurons preclassified by labelling. Small sample m=41 genes in dimension n=20 neurons.
- $\Rightarrow$ Beyond pairwise interactions : I_{10} identifies the DA population.
- $\Rightarrow$ Identification of the two cell types (diversity=2)
- Informational and topological formalization of epigenetic landscape à la Waddington and Thom



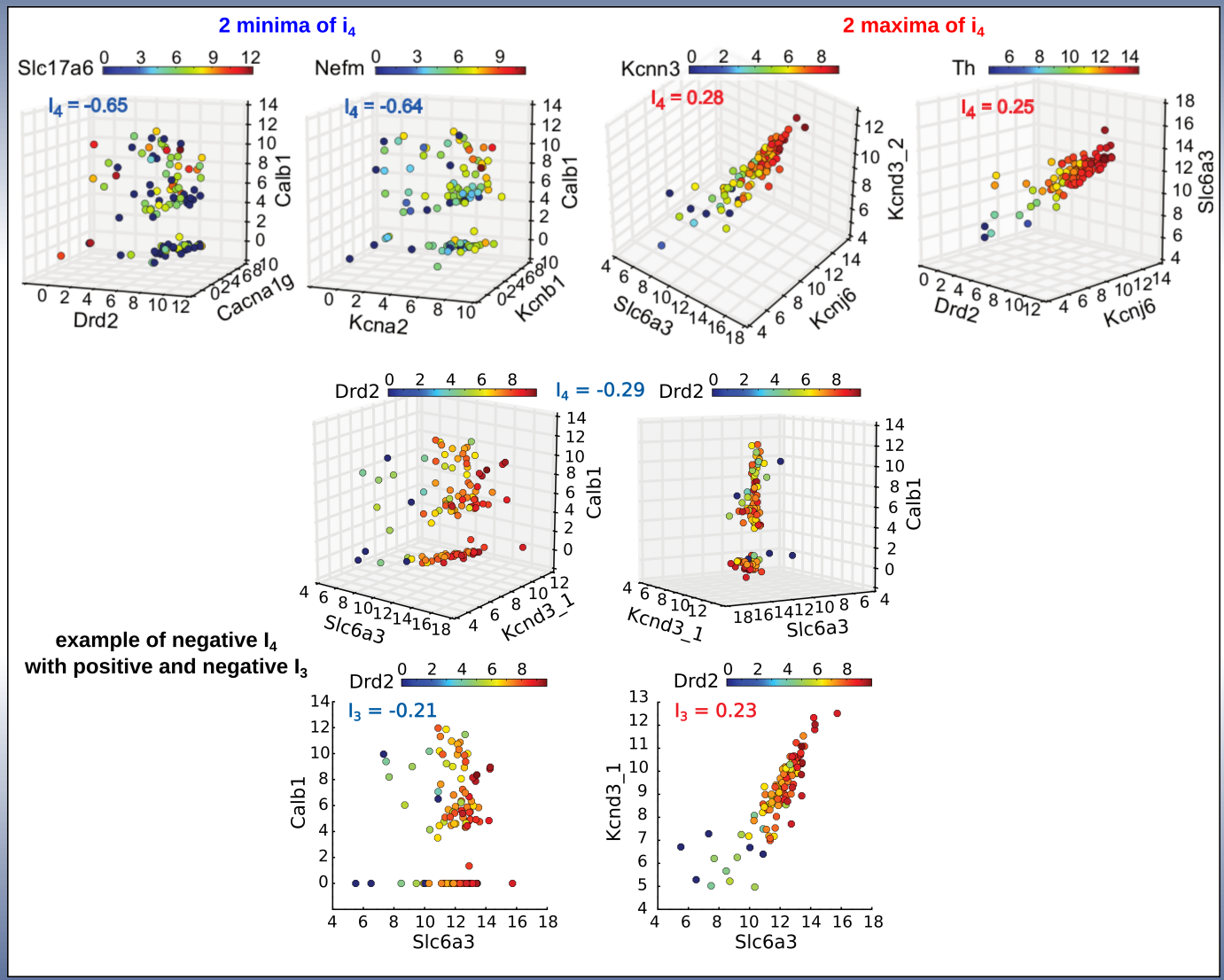
Gene modules identification

- Combinatorial explosion of interactions (analog to Van der Walls), dependences and independences in high dimension k. Modules with high I_k identify the metabolic chain of dopamine and reveals new modules of coregulation, biologically relevant (neuromodulator-electrical coupling).
- I_k positivity detects covarying variables (co-expression, common transcription factor). I_k negativity (synergie) detects **clusters**, differential expression (spatial) and known cell sub-types.



Maximum and minimum I_k "modules"

- Negative I_k detects clusters
- Positive I_k detects "covariations" even non-linear (Reshef,2011)

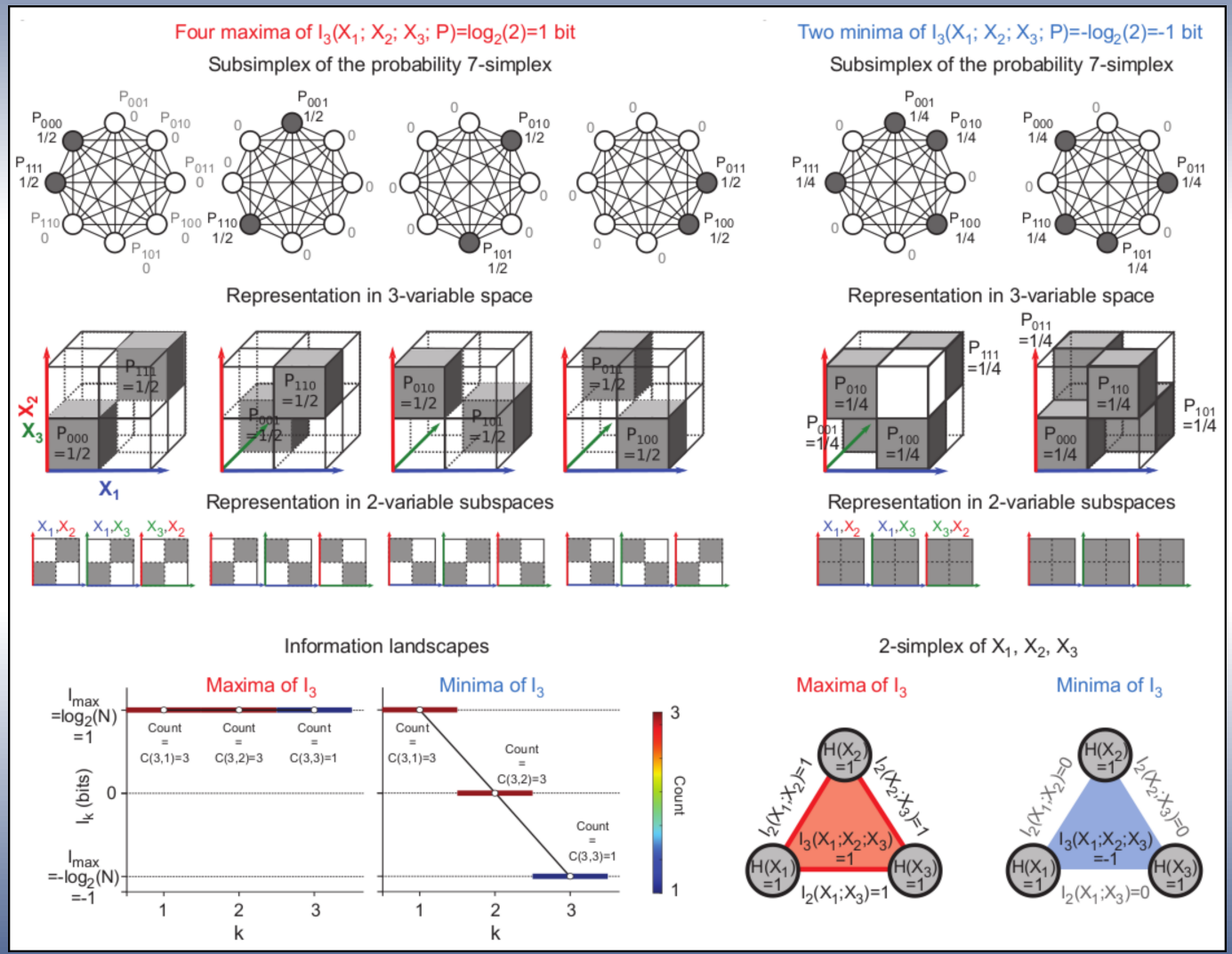


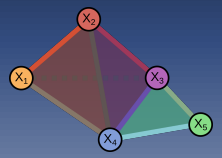
I_k extrema and negativity - Special cases

Theorem (Hu Kuo Ting, 1962)

For $k \geq 3$ I_k can be negative.

- Schrödinger "what is life?": living system feed upon negentropy (free-energy)
- Synergy (Brenner et al.)
- Frustrated spin glasses (Matsuda)



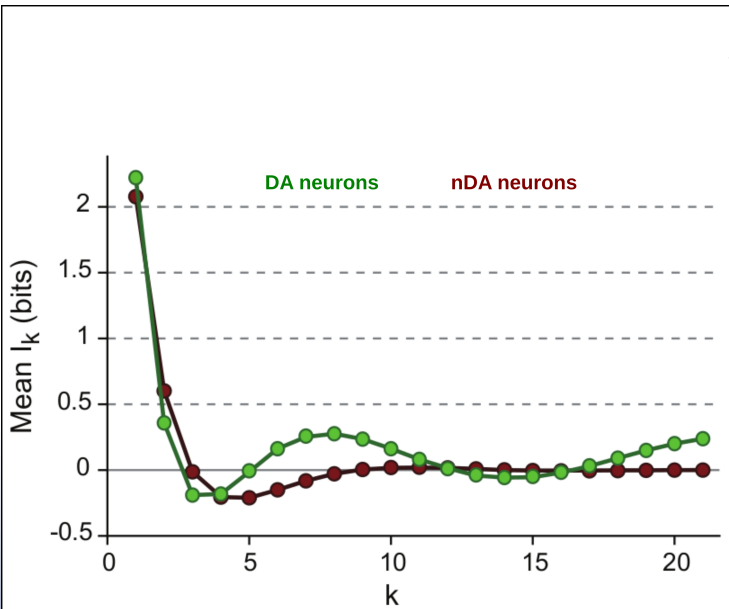


Mean H_k and I_k - Homogeneous system

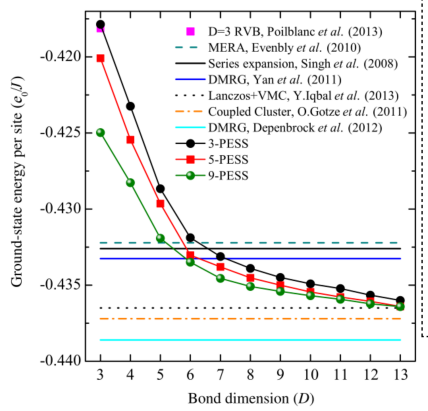
Mean behavior of the information structure defined the mean H_k and I_k:

$$\langle H_k \rangle = \frac{\sum_{T \subset [n]; card(T)=i} H_k(X_T; P)}{\binom{n}{k}}, \quad \langle I_k \rangle = \frac{\sum_{T \subset [n]; card(T)=i} I_k(X_T; P)}{\binom{n}{k}}$$

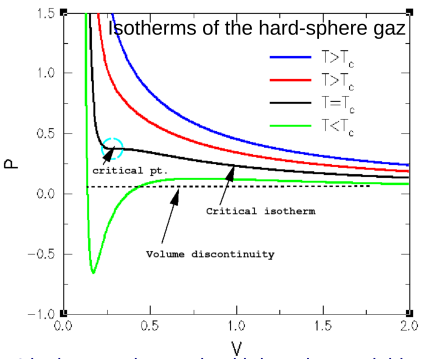
Mean information correspond to ideal homogeneous structure X^n_{hom} with homogeneous k-body interactions.



Xi et al, Tensor Renormalization of Quantum Many-Body Systems Using Projected Entangled Simplex States, 2014.



Van Der Walls n-body interactions

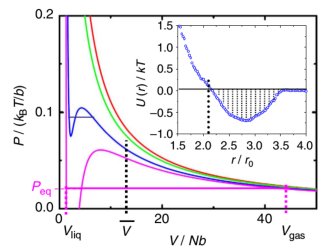


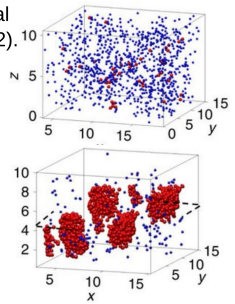
Ideal gaz no interaction i.i.d random variables
PV=nRT

Gaz with attractive n-body interaction
$(P+an^2/V^2)$ $(V-nb)=nRT$

A lot of small interactions makes important interaction

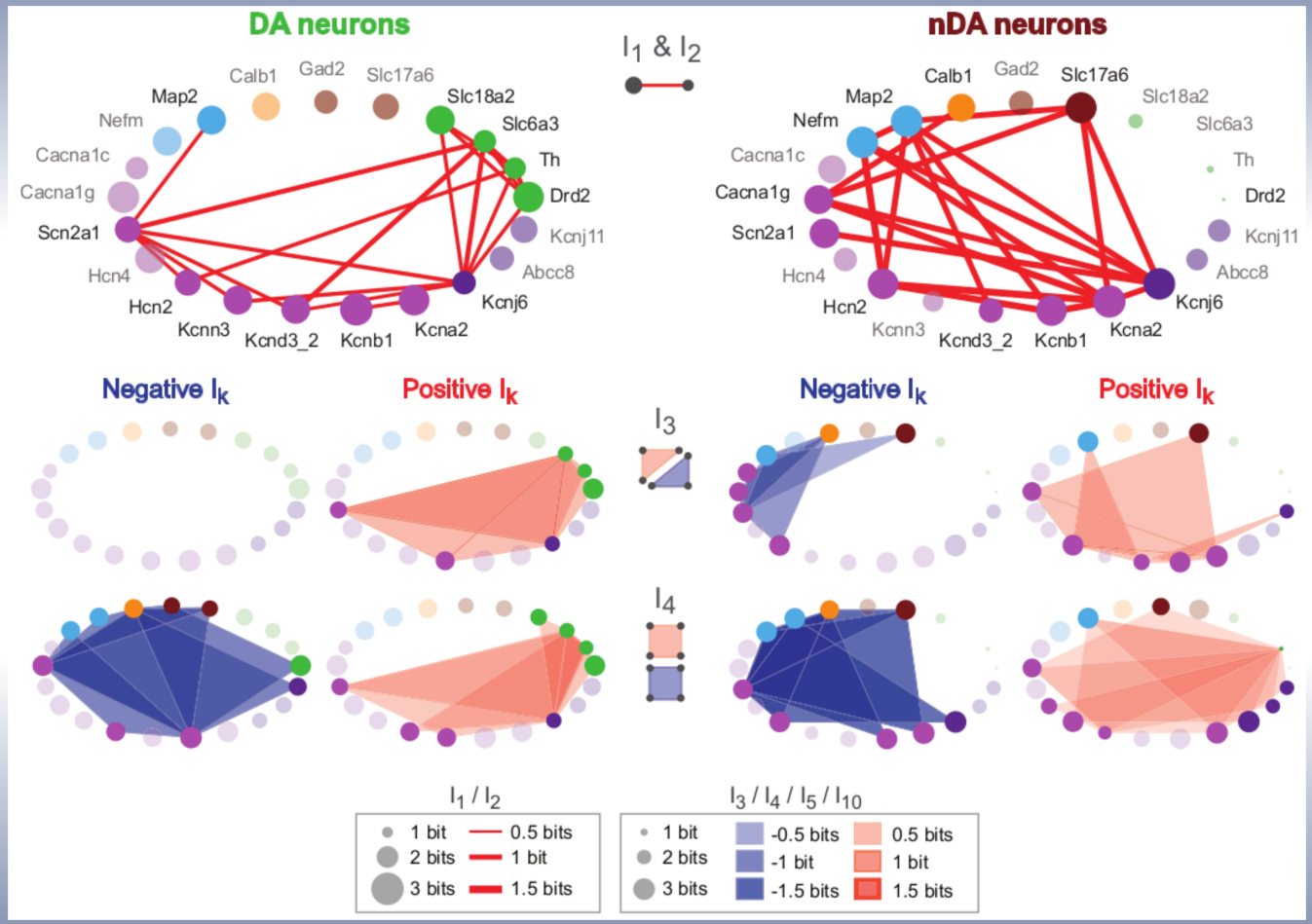
Van Der Walls isotherms of the colloidal Gaz-liquid transition (Nguyen et al, 2012).





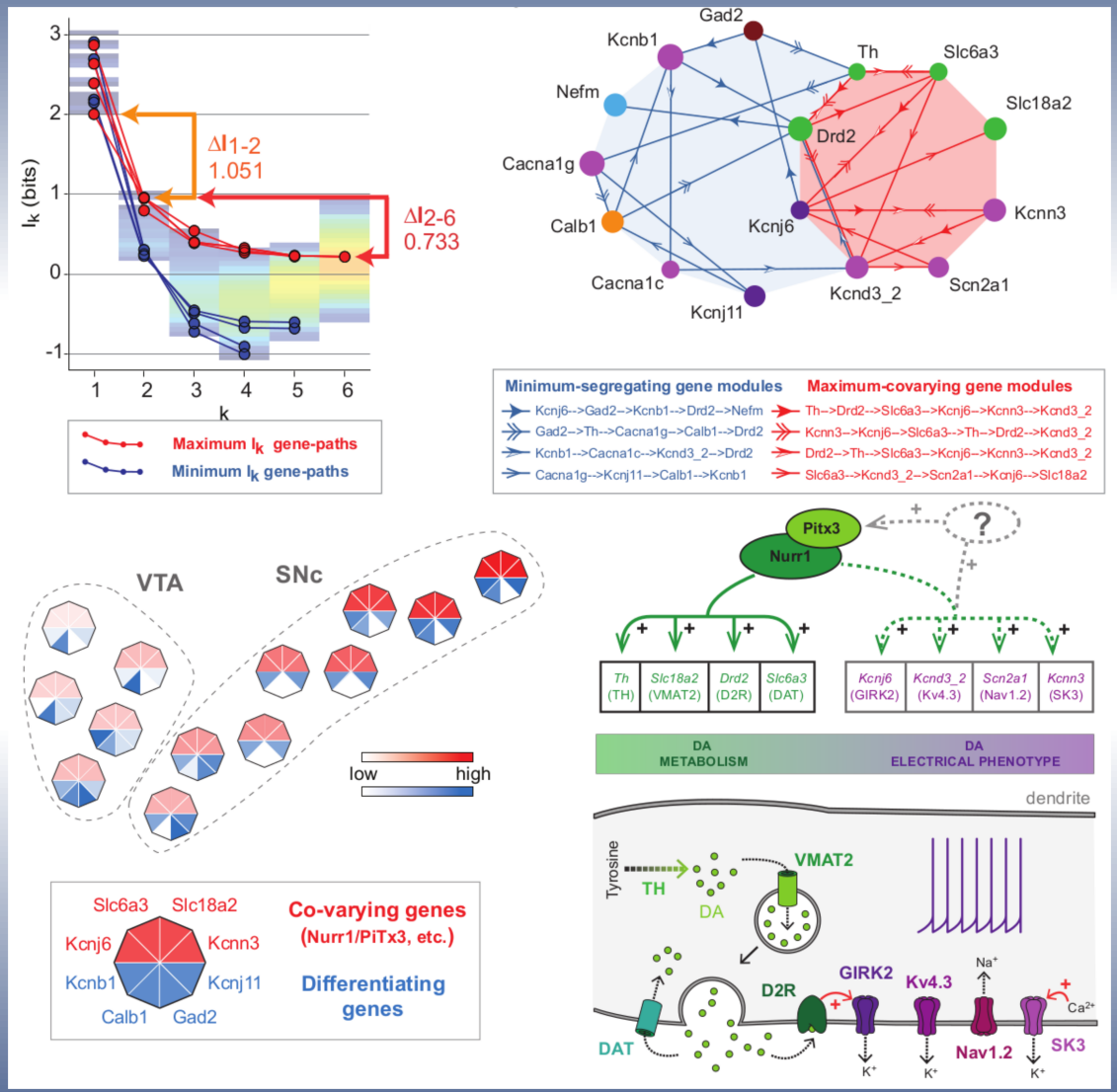
Maximum and minimum I_k "modules"

- I_2 qualitatively similar to $\rho_{X,Y}$ (Reshef, 2011).
- I_k are nontheless specific to a given cell type: **cell identity signature**.



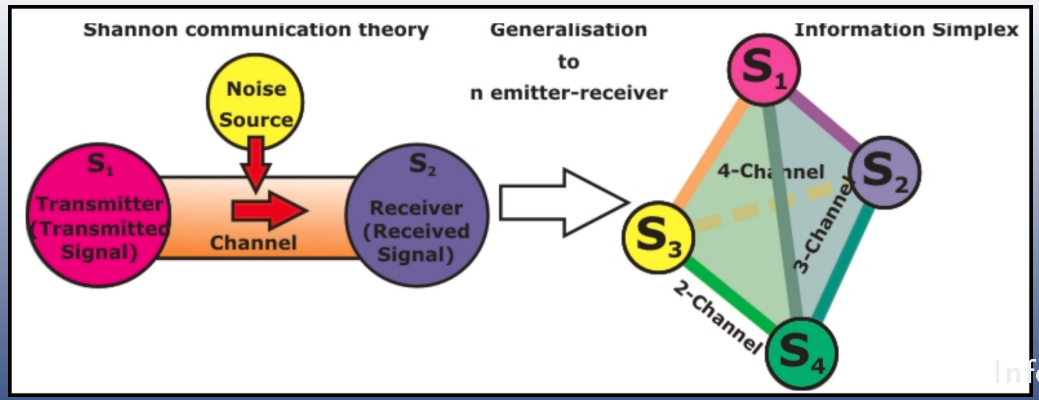
DA Minimum free energy complex

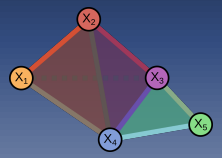
- Identifies functional module up to $k_i = 6$.
- Maximum path detect the metabolic chain of Dopamine, genes having common transcription regulators and unravel electrophysiological and neuromediator identity coupling.
- Minimum path detect heterogenity, subclasses and spatial differential expressions.



Conclusion

- New methods of **Machine learning and topological data analysis, Shannon-Poincaré Machine**: Identifies relevant modules up to $k_i = 10$ $k_i = 6$ with sample $m = 41$ and $m = 111$, resp. Available opensource-python program INFOTOPO.
- Information theory and data analysis tools without metric, Markovian, iid, Gibbs distribution, phase space or symplectic assumptions.
- Common framework for epigenetic learning
- Beyond pairwise statistics: from complex networks to complex $\Rightarrow$ Topological neural complexes (binary variable).

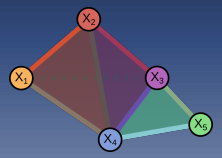




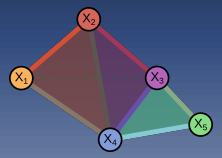
Thank you!

Thank Median, Jean-marc, Monica, Daniel ... UNIS1072 inserm, ERC Chanelomics..

Questions?

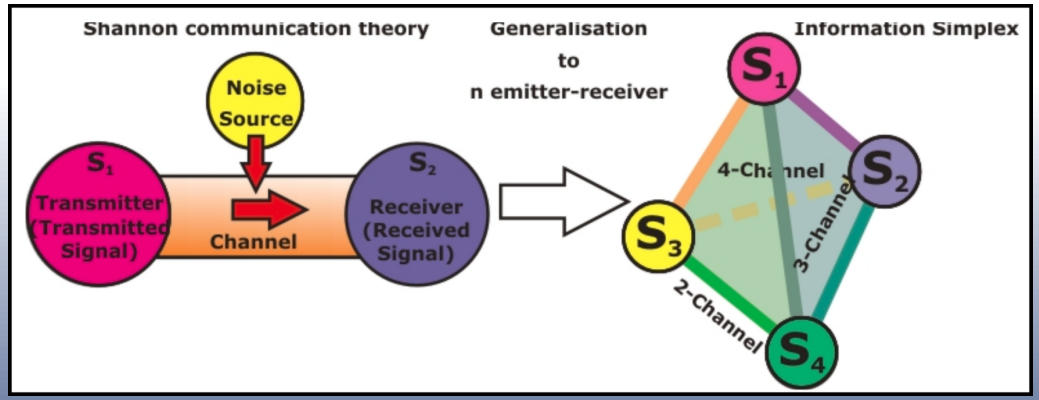


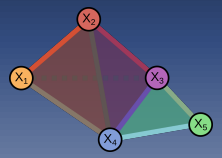
Questions



Information theory

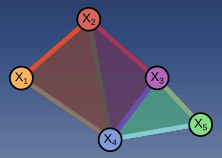
- The global picture: information communication is only partially accounted by pairwise exchange of information, formalized by a communication channel, that is a 1-simplex between two variables, the emitter and the receiver. By considering n emitters/receivers and defining k-communication channels as the k-face of a simplicial structure, with respective capacity $\max(I_k)$, the present topological formalism gives very preliminary basement for such a generalized communication theory. Moreover, it suggests refined data compression algorithm.





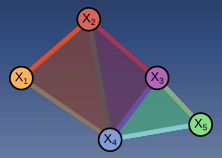
Statistical physic

- At least in genetic expression, but we propose that it is a generic feature of biological structures, high order than pairwise statistical interaction exist, can be non negligible, and moreover can be combinatorially numerous.
- Clustering of data points analog to matter condensation, a simple picture.
- Topological and informational formalization of the Potts model, negativity signature of frustration, multiplicity of local minima.
- mean information path is analog to DFT treatment of the n-body problem, but the formalism here is different, it is finite and discrete, it computes the cohomology group of measurable function, do not assume any metric (like an interaction distance r), nor Hamiltonian or Lagrangian structure, symplectic or contact structure, configuration or phase space (etc.). The main difference with classical statistical physic determinations of free energy and entropy is the absence of predefined metric and the finiteness-discreteness of the formalism (no assymptotic limit, no Stirling approximation).



Statistical physic

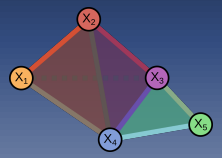
- Our theorem applied to 3n dimensions of a configuration space (like in DFT) implies that whereas the minimum free information energy complex of an elementary body can only be a simplex, the configuration space of n elementary body can be a complex with quite arbitrary topology (possible heterogeneity at large "scales").
- What should be done next: discrete analog of Noether theorem.



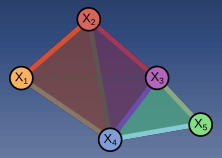
Ecology - Biology - Complex systems

- **Ecology** is the scientific analysis and study of interactions among organisms and their environment...
- Biology and ecology: the main interest of the present formalism is to capture and identify diversity, while yet allowing selectivity. It gives a quantitative framework the cellular identity and its differentiation.
- From complex network to ... complex.

Appendices



First appendix



Second appendix