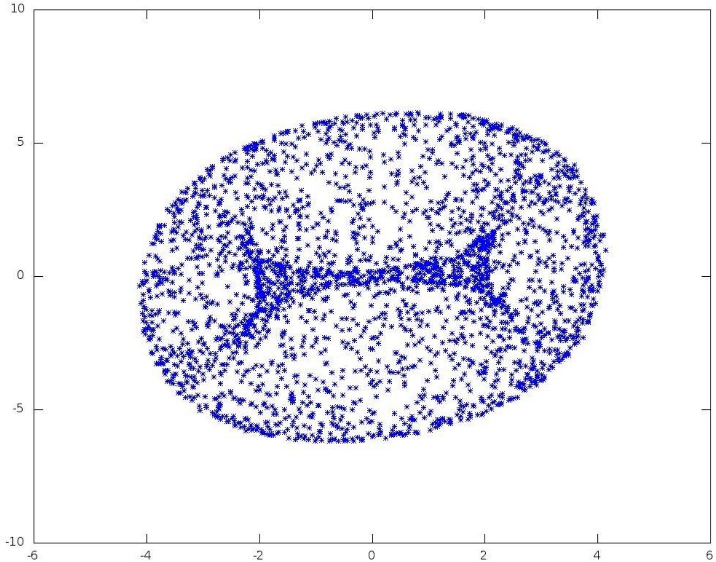
topological data analysis
and stochastic topology

yuliy **baryshnikov**
waikiki, march 2013

Promise of topological data analysis: extract the structure from the data.

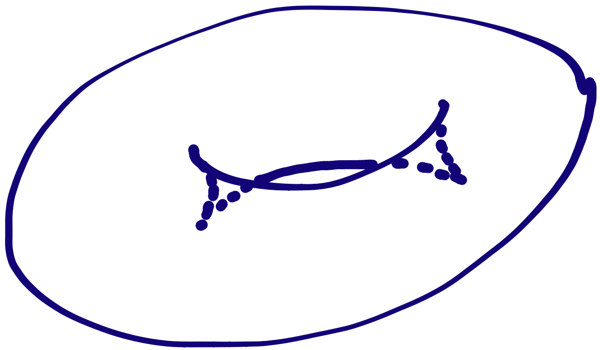
In: point clouds

Out: hidden structure



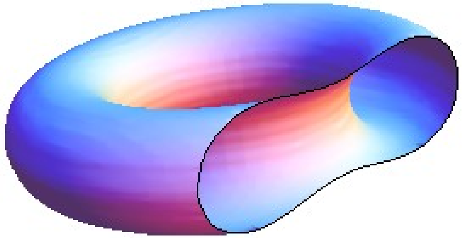
People: Edelsbrunner
Harer
De Silva
Carsson

DARPA



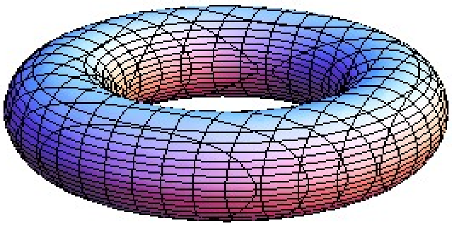
Underlying structure:
manifolds?
stratified sets?

Two ways to look at them: Poincare's and Riemann:



$$M = \bigcup_\alpha U_\alpha , \; \psi_\alpha : U_\alpha \to \mathbb{R}^n$$

$$M = \{x \in \mathbb{R}^N : f(x) = 0\}$$



Correspondingly, there are two approaches in dimensionality reduction:

- Poincare view - use ambient coordinates - PCA, Kernel PCA,...
- Riemann view - only intrinsic proximity makes sense - IsoMap, EigenMap,...

Topological Data Analysis aims at recovery of the most stable invariants of the data.

In particular, complete scale invariance is sought

This is both a curse and a blessing:

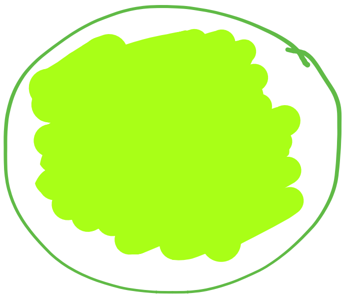
- very susceptible to noise
- only vague notion of proximity matters

Invariants of topological spaces: two spaces are the same if they are

homeomorphic?

diffeomorphic?

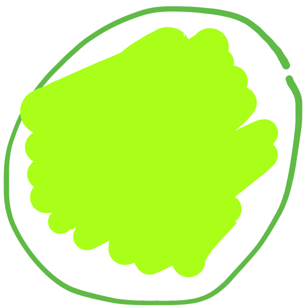
homotopic?

 $\cong_h$ 

$\not\cong_d$

$X \cong Y$: $f: X \to Y$, $g: Y \to X$

$g \circ f = id_X$, $f \circ g = id_Y$

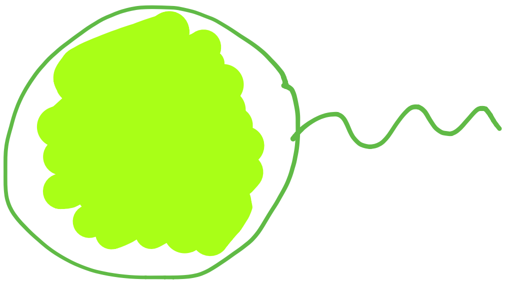
 $\sim$ 

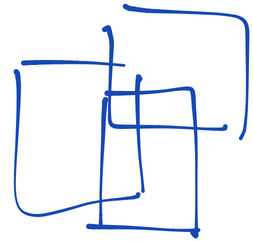
$X \sim Y$ $\quad f \circ g \sim id_Y$

$g \circ f \sim id_X$

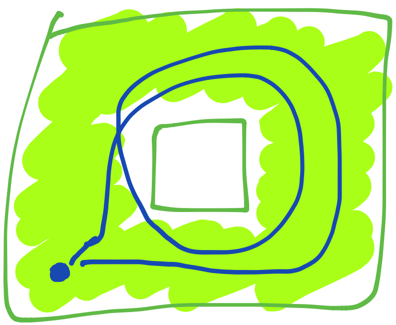
Invariants of topological spaces:

Dimension?





Homotopy groups?



$$\pi_3(S^2) =$$

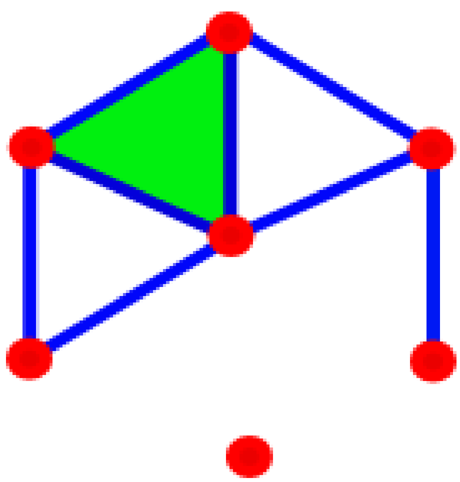
Homology and cohomology

Homology and cohomology

come in different flavors and colors

singular, de Rham, simplicial

Simplicial - easiest to deal with. Works well with meshes...



$$H_p(X, R)$$

$$C_p = R\langle \text{simplices of dim } p \rangle$$

$$\partial\sigma = \sum_i (-)^i \sigma_i$$

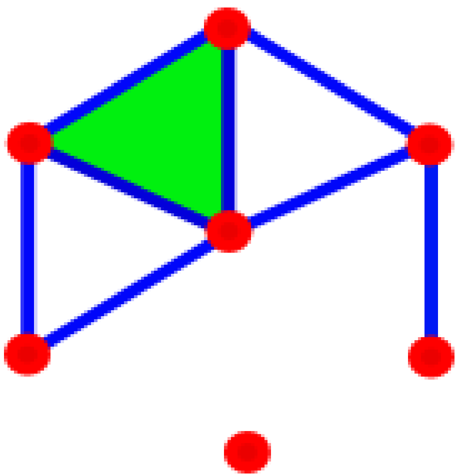
$$Z_p = \operatorname{Ker} \partial_p \qquad B_p = \operatorname{Im} \partial_{p-1}$$

$$H_p = Z_p / B_p$$

Algebraic formalism includes chains, cycles and boundaries, and the idiosyncratic relation

$$\partial^2 = 0$$

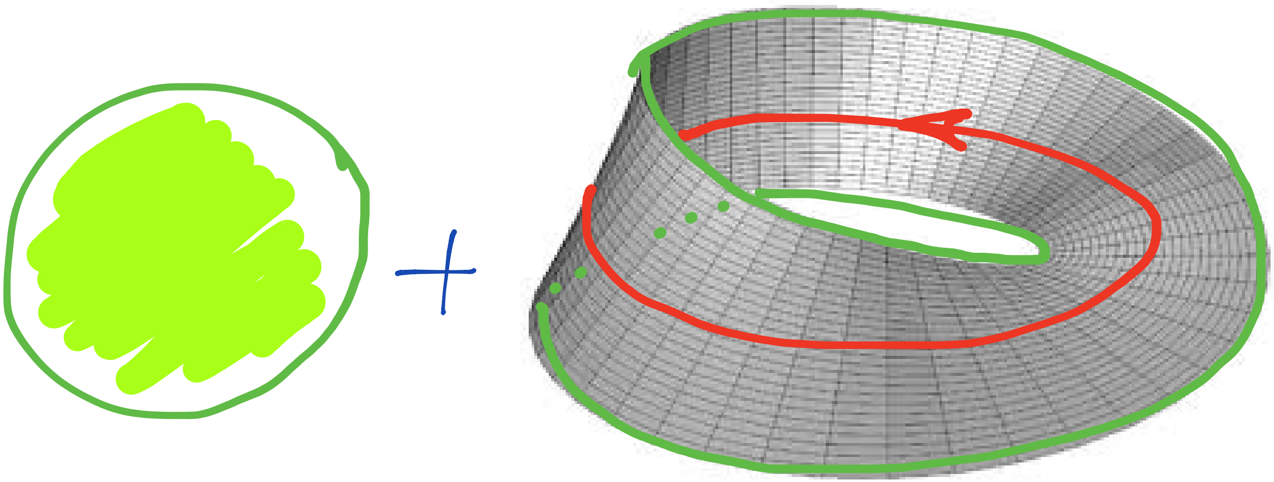
In essence, linear algebra encoding combinatorics of adjacencies.



$$\partial = \begin{array}{cc} \sigma & \\ \begin{bmatrix} & & \\ & \pm 1 & \\ & & \end{bmatrix} & \sigma' \end{array}$$

Homology groups depend on the coefficient ring: what are the entries of the boundary matrix.

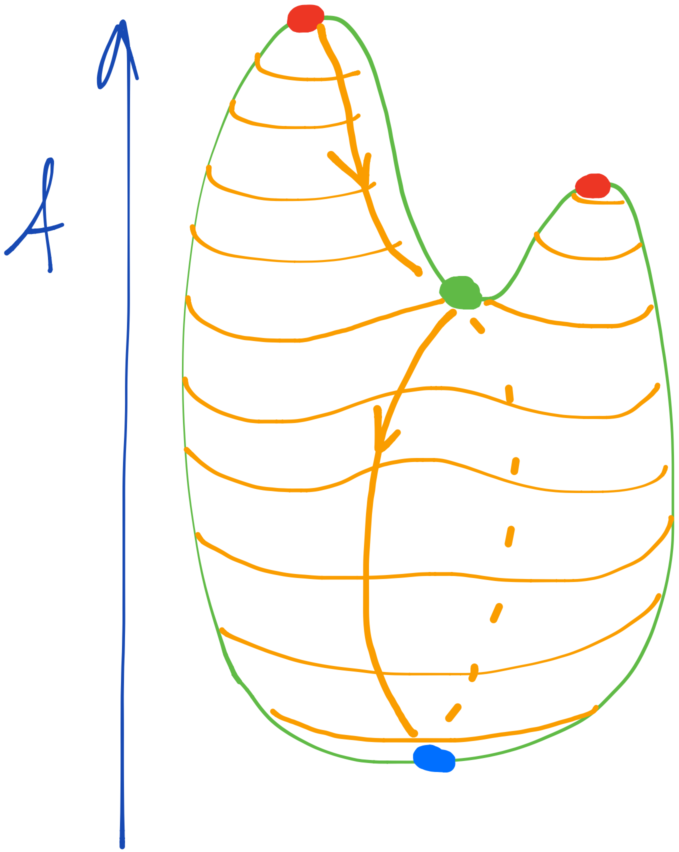
Sometimes there is torsion:



Working over a field – ranks of the homology groups are called Betti numbers: numeric invariants of a topological space

$$\beta_p = \mathrm{rk}\ H_p(X, k)$$

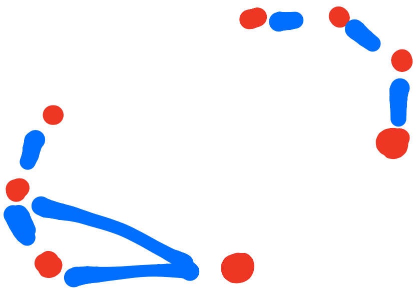
Morse theory – an important tool to generate the cellular partition of the underlying manifold, and estimate the Betti numbers

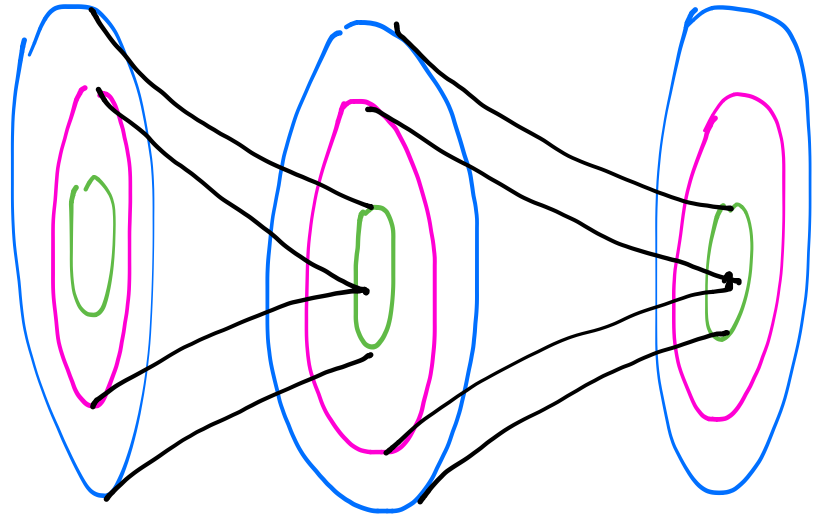


$$\beta_p \le \mu_p$$

$$\beta_p - \beta_{p-1} + \dots \pm \beta_0 \le \mu_p - \mu_{p-1} + \dots \pm \mu_0$$

What happens when one adds a simplex of dimension *d*: either a cycle in dimension *d-1* disappears, or a cycle of dimension *d* appears.



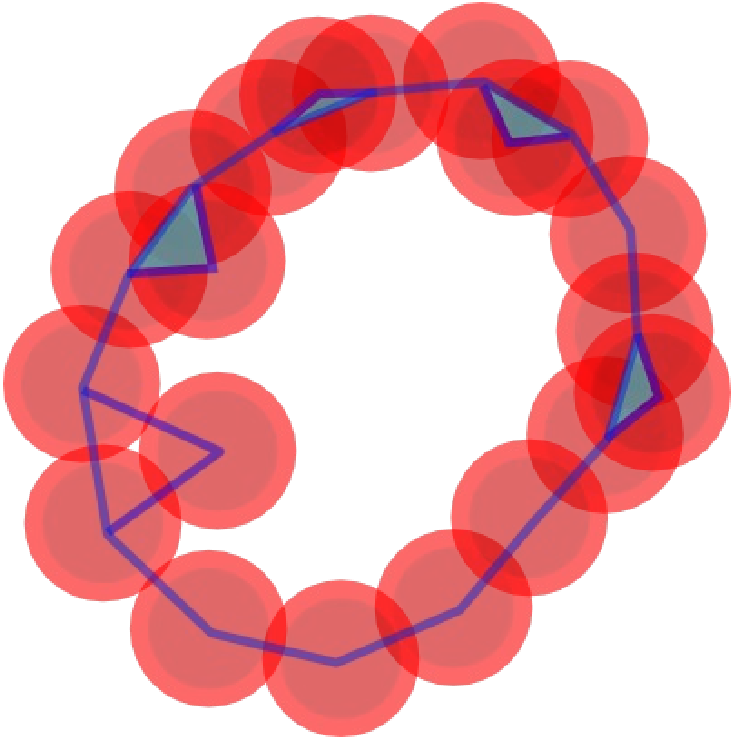


This implies existence of an aggregate invariant, the Euler characteristic.

$$\chi(X) = \sum_{1} (-1)^p \beta_p = \sum_{\sigma} (-1)^{\dim \sigma}$$

Back to data: given a point cloud, how to arrive at a topological space? Taking into account desirable scale?

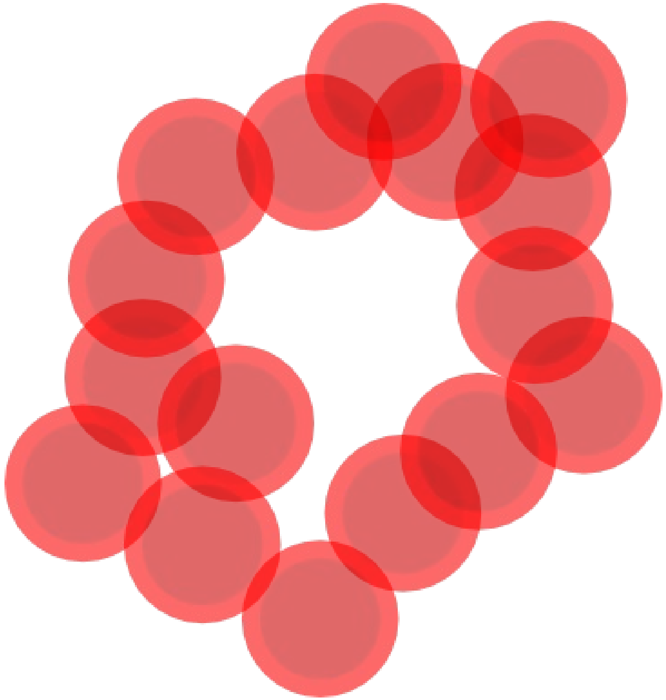
Old idea: *alpha*-shapes (Robbins, Edelsbrunner...)



Place a ball around each point in the dataset and construct the *Čech complex*: glue in a simplex for any non-empty intersection of the balls.

Nerve Theorem: If all intersections are contractible, the union of balls is homotopy equivalent to the Čech complex.

Noise is a problem, however: how to get rid of the small features?
Answer: increase the scale.
and again
and again
Until the features disappear altogether



Persistence is a way to address the onslaught of irrelevant features

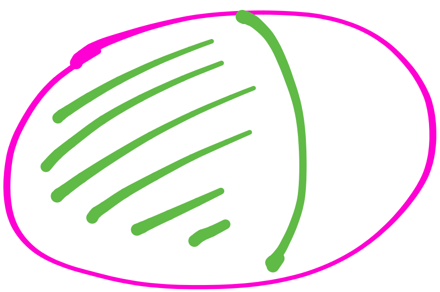
Tidal traces to be ignored: only the islands survive.

Persistence homology

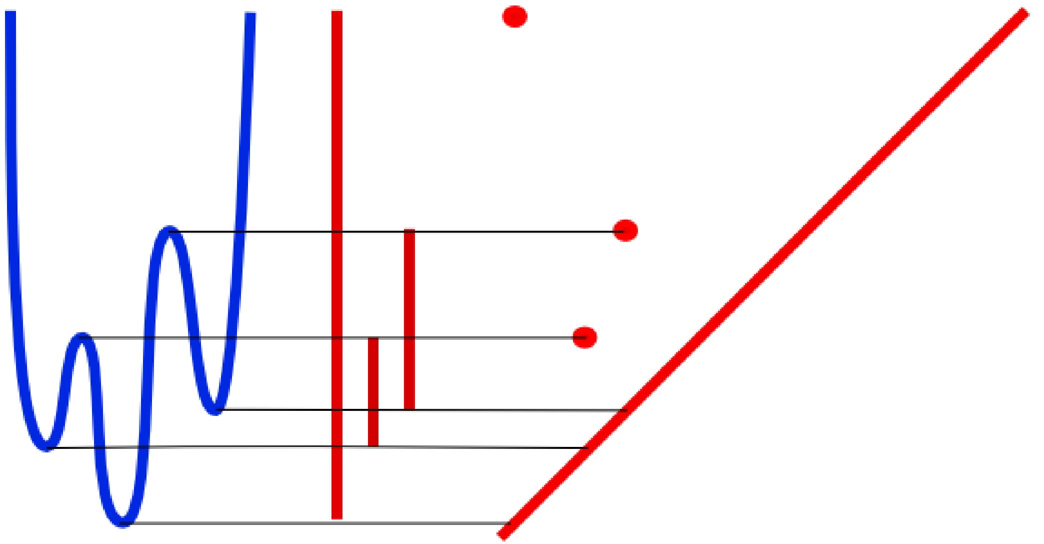
Start with a filtration, and record when a cycle appears, and when it dies.

$$\emptyset \subset K_0 \subset K_1 \subset K_2 \subset \dots \subset K_n$$

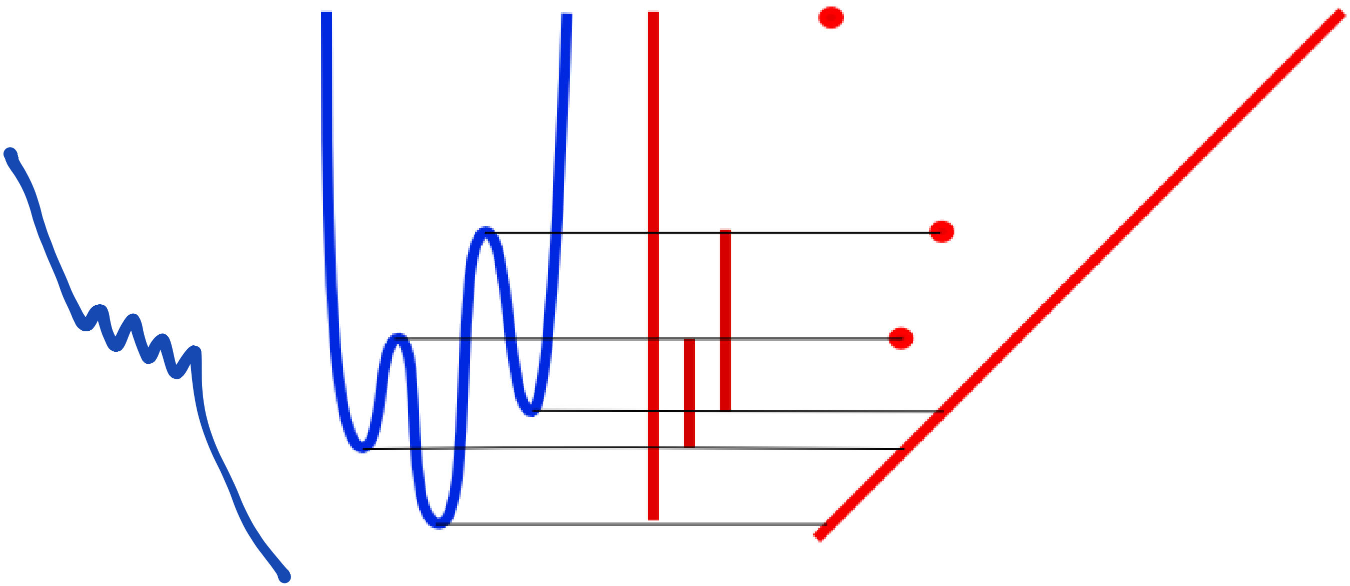
$$\begin{array}{ccc} Z_p(K_t) & \supset & B_p(K_t) \\ \cap & & \cap \\ Z_p(K_{t+1}) & \supset & B_p(K_{t+1}) \end{array}$$



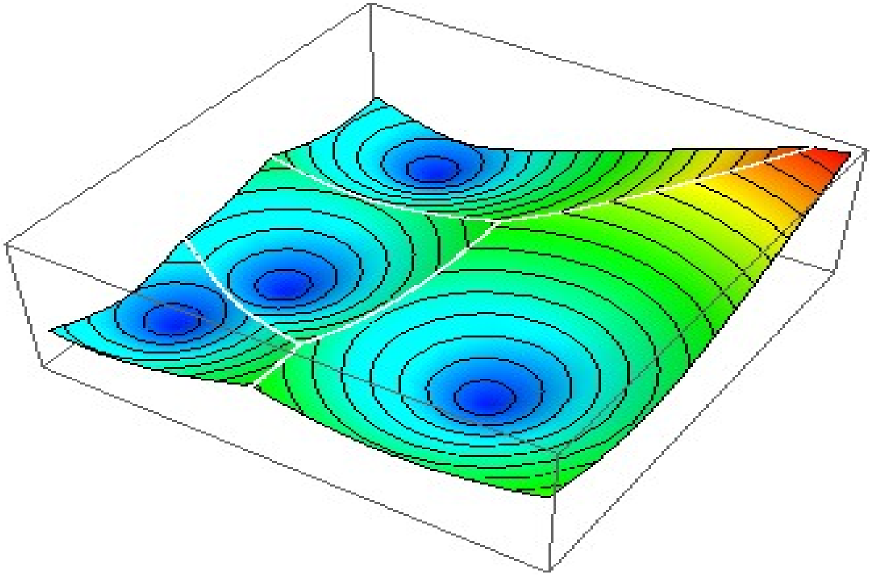
Encoding: either barcodes, or persistence diagram

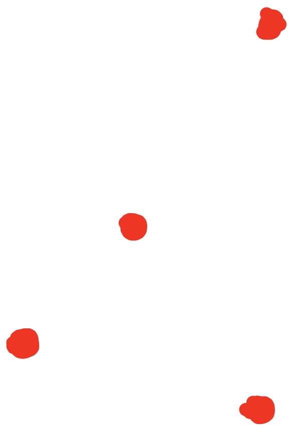


Important result: stability theorem – small perturbations of the filtration change the persistence diagram little



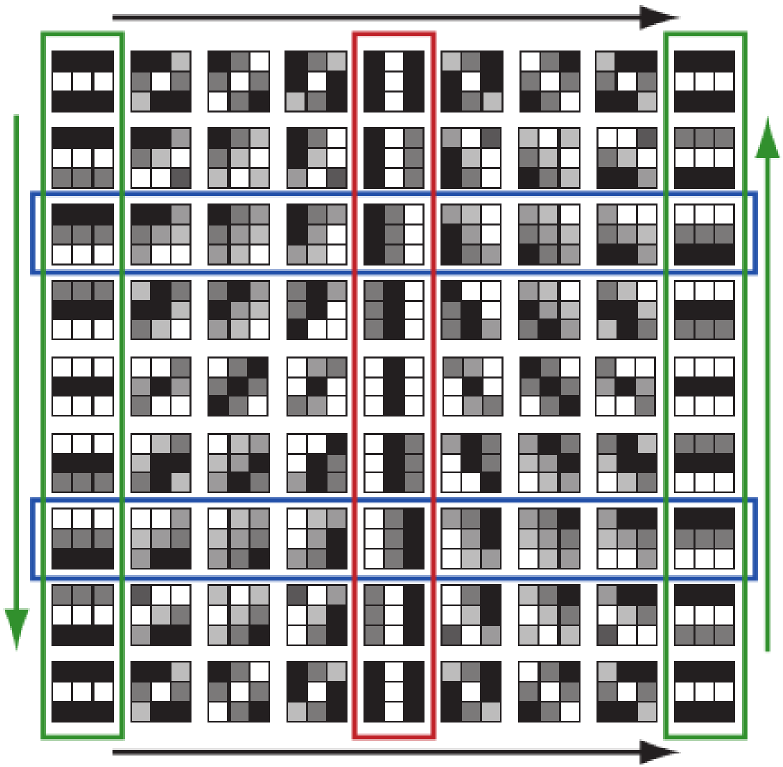
Important tool: Morse theory

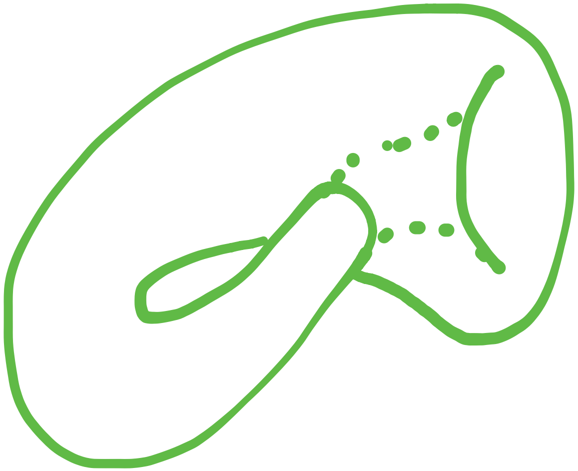




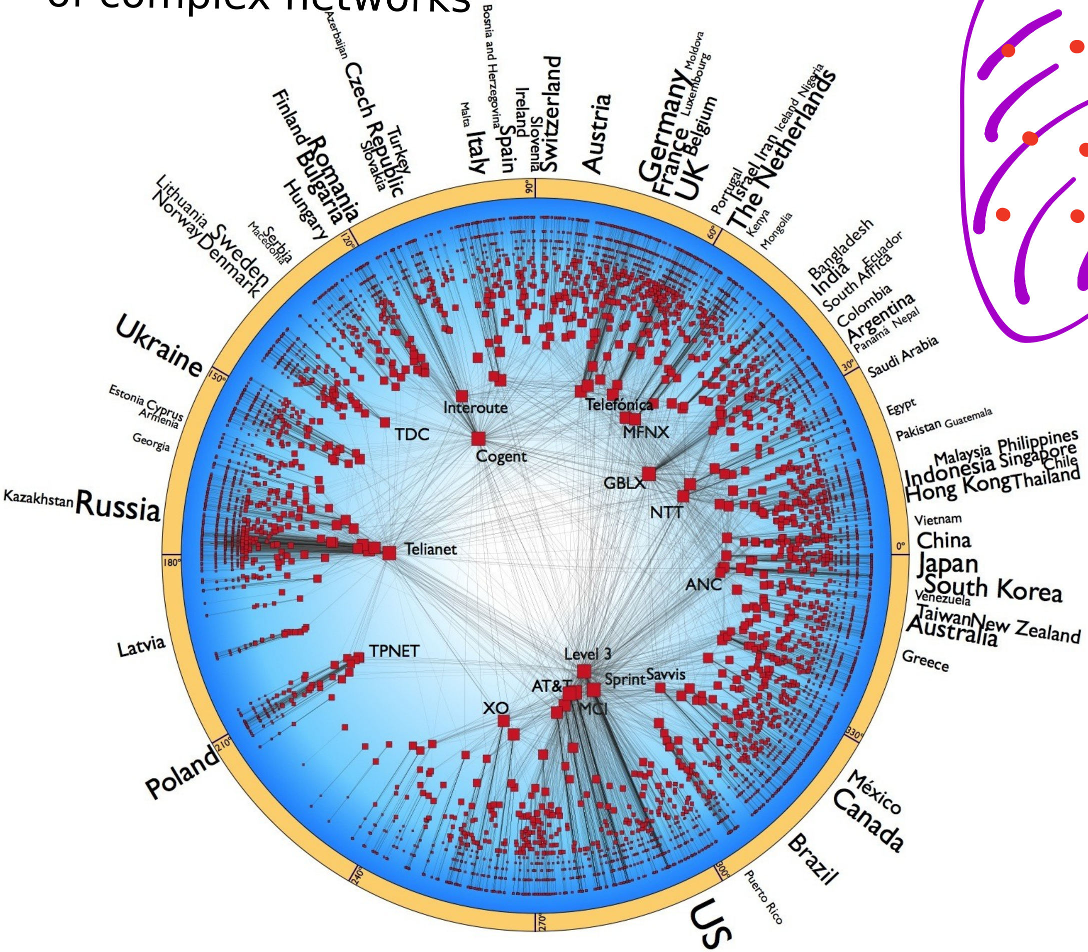
Some examples

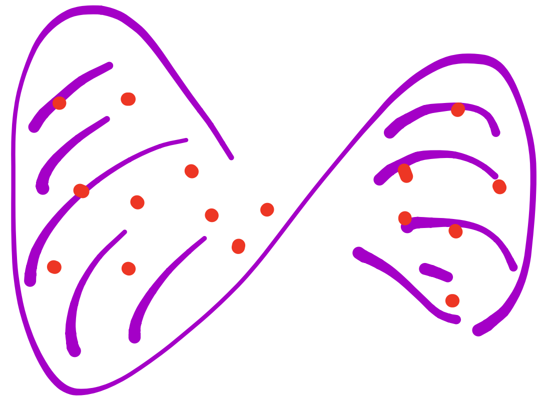
Mumford dataset and the Klein bottle





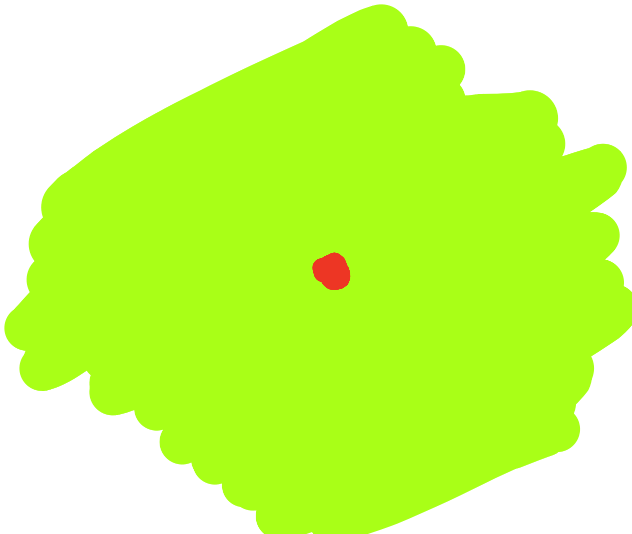
Kriookov et al. Hyperbolic mapping of complex networks





Simple test: what is the *dimension* of the internet?

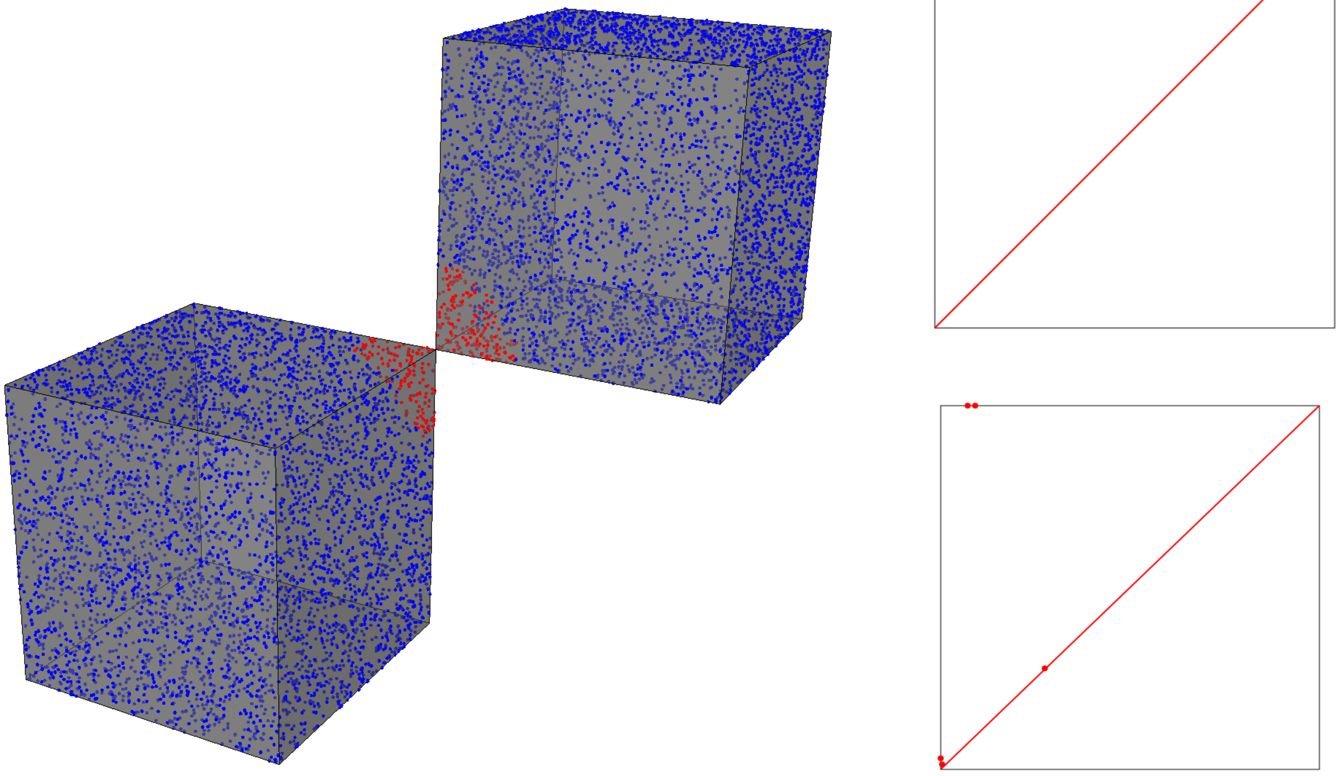
How we find the dimension knowing a sample from a manifold?



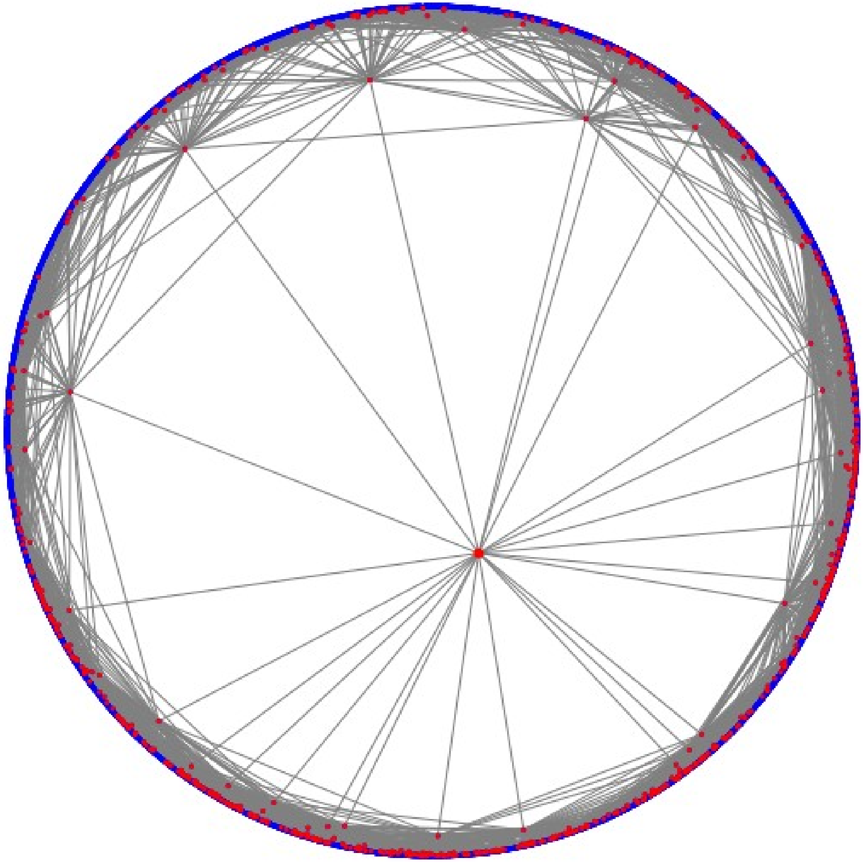
$\cong (S^2, pt)$

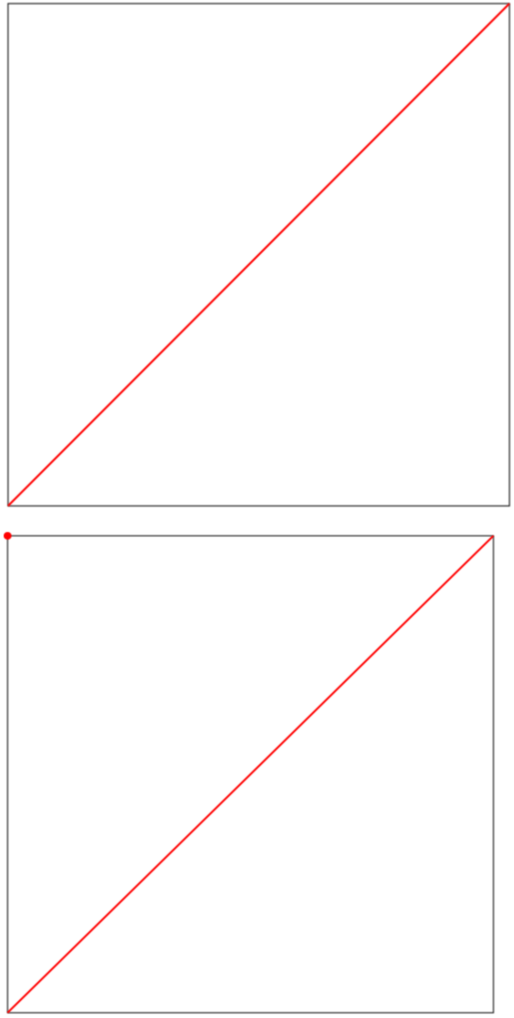
$(M, M-\{x\}) \cong (S^n, pt)$

Testing on a special
(stratified) set

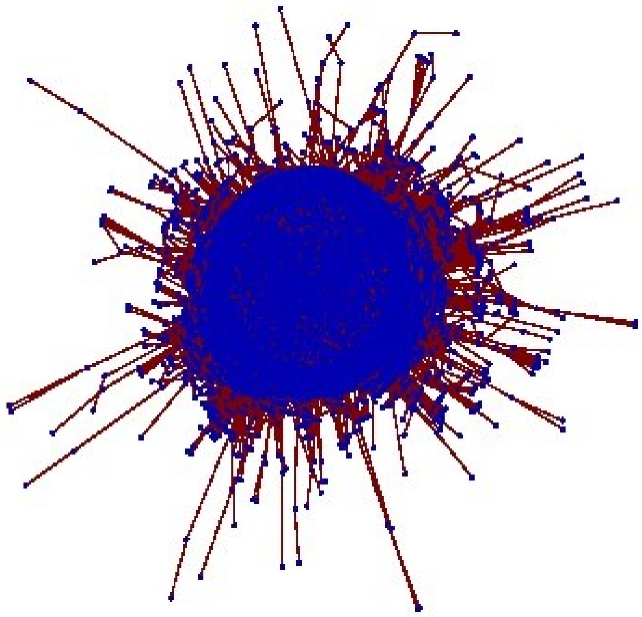


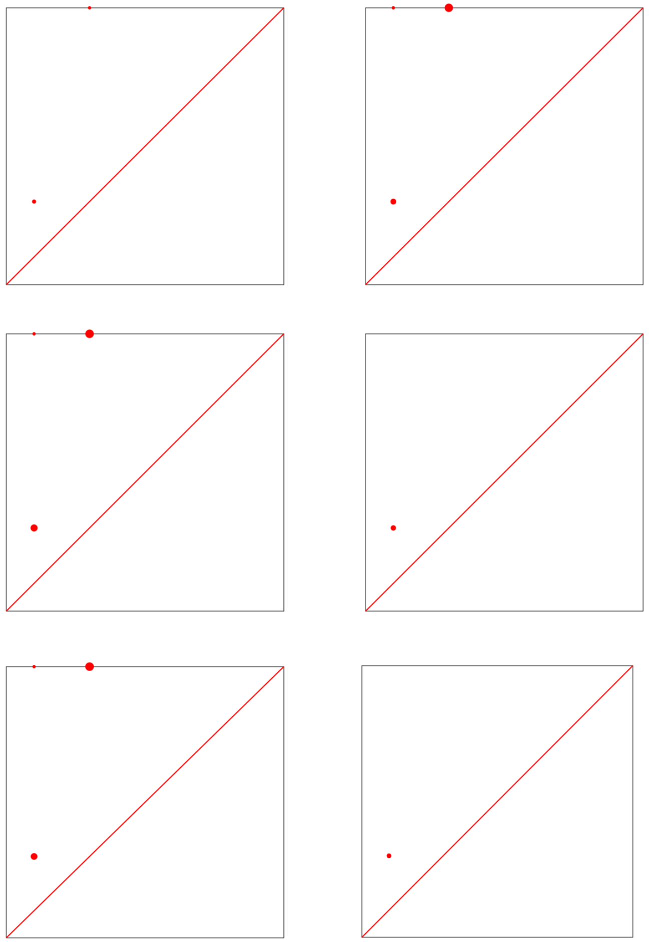
Testing on a special
(stratified) set



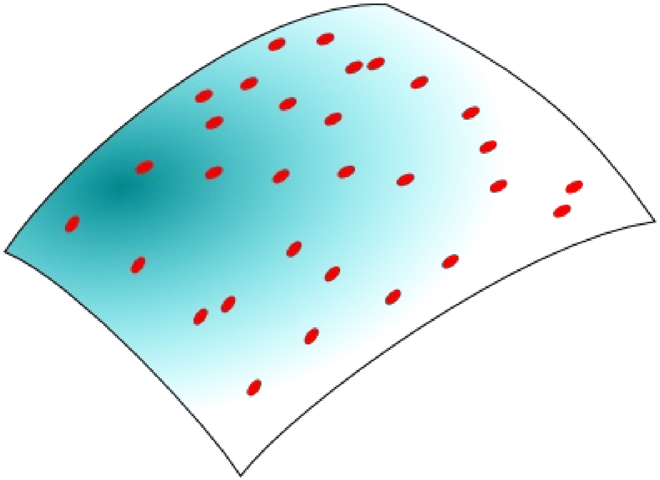


Autonomous Systems
("*Internet*")



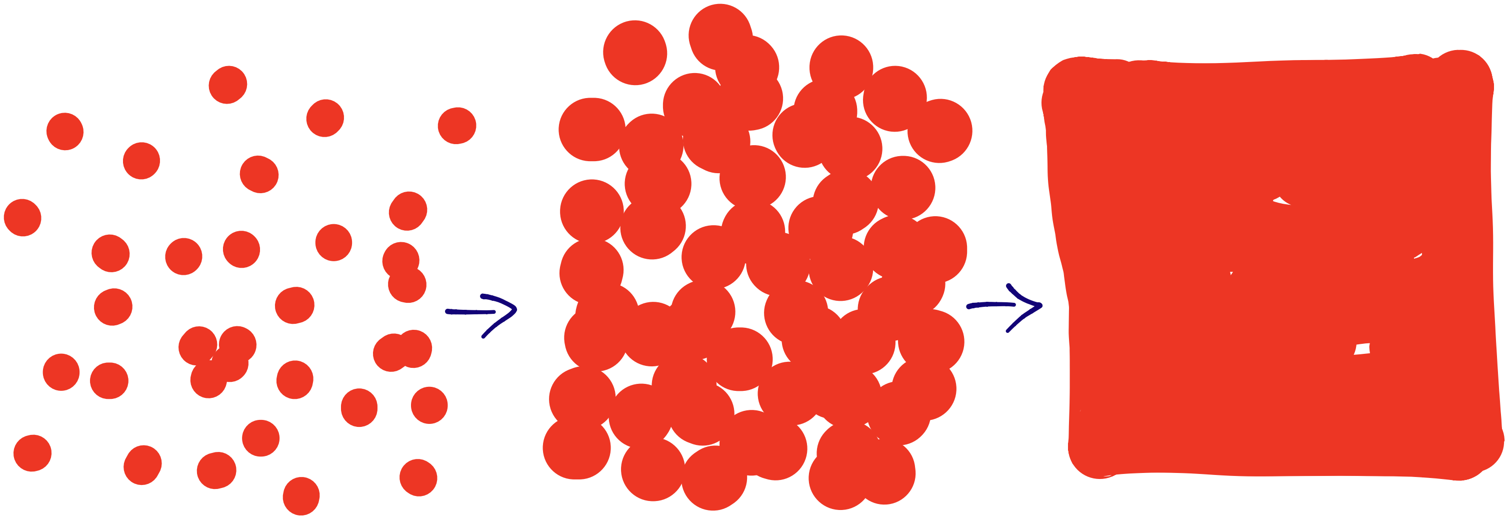


Stochastic Topology: dealing with the baseline case



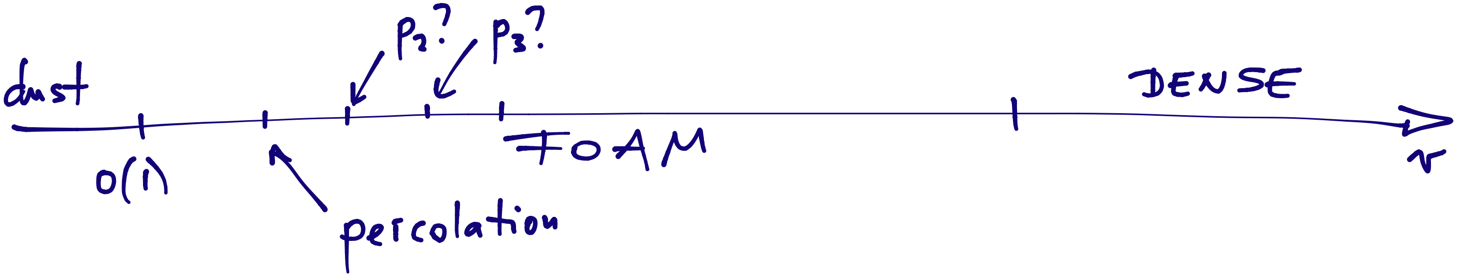
Niyogi, Smale, Weinberger: Cech complex for small enough *r* and dense enough sample with has the homotopy type of the underlying manifold

Big Picture: from dust to foam to monolith

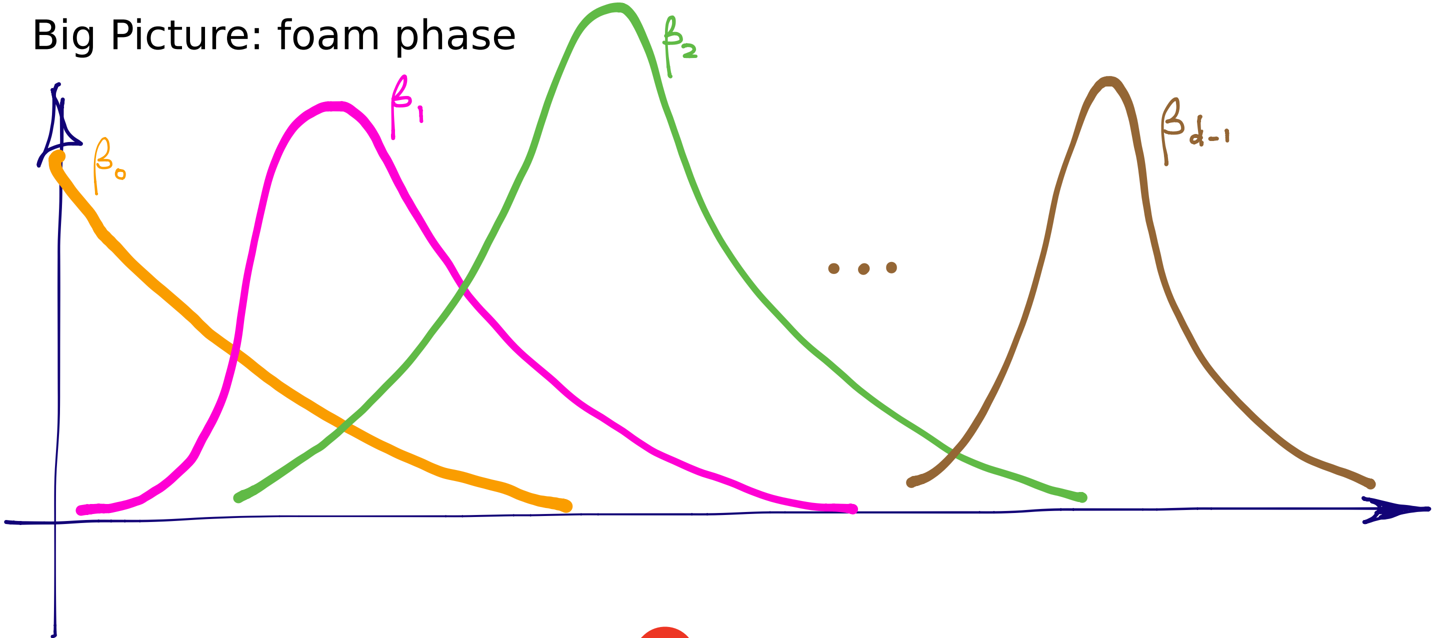


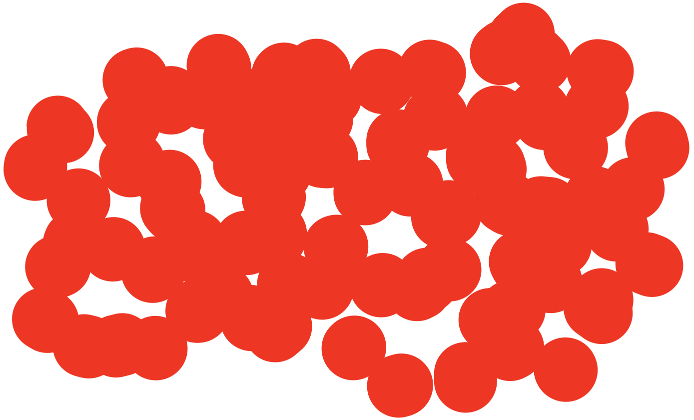
Big Picture: timescale

$$\nu = n \cdot r^d$$



Big Picture: foam phase





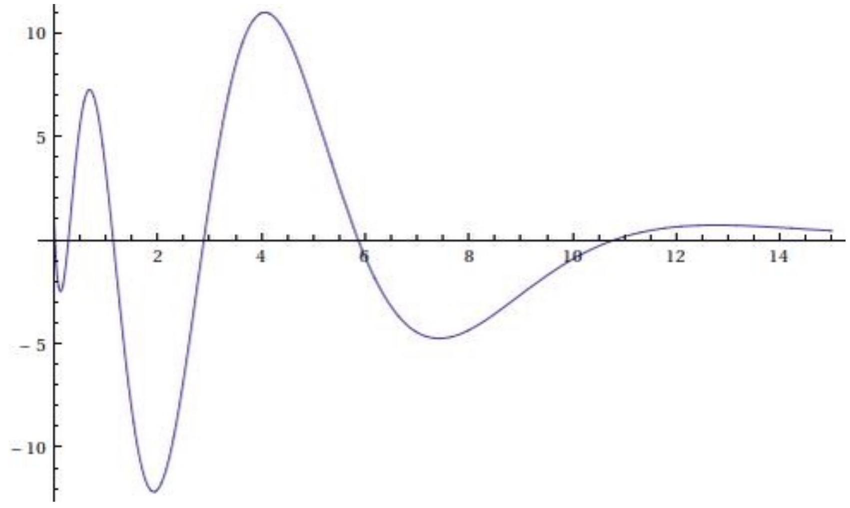
Foam phase: oscillation of Euler characteristic

$$Y = X + C, \quad |C| = \nu$$

Thm (Okun)

$$\chi_d(\nu) = \frac{d}{d\nu}\left(\nu\, \chi_{d-1}(\nu)\right)$$

$$\chi_1 = e^{-\nu}$$



Euler characteristic in dim=7

Foam phase: concentrations of Betti numbers

$X \subset \mathbb{R}^d$ - Poisson, $\rho \equiv 1$

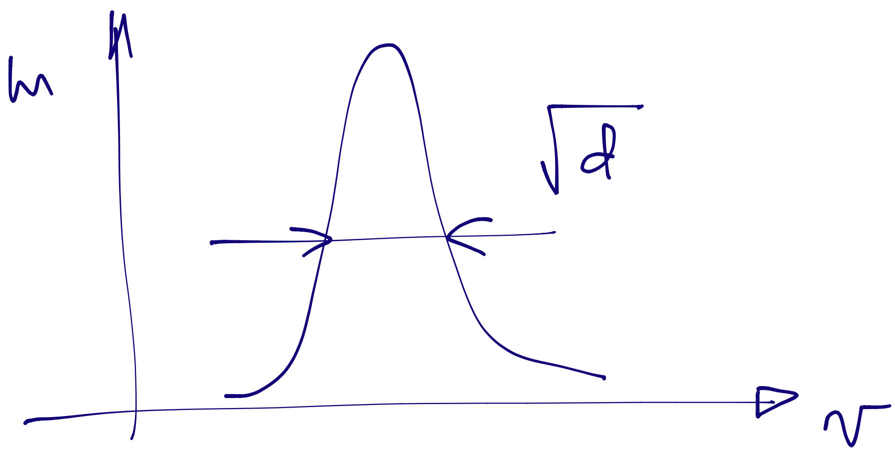
$X_r = X + B$, $|B| = r$

$$\beta_p(r) := \lim_{\Omega \to \mathbb{R}^d} \frac{\beta_p(X_r \cap \Omega)}{|\Omega|}$$

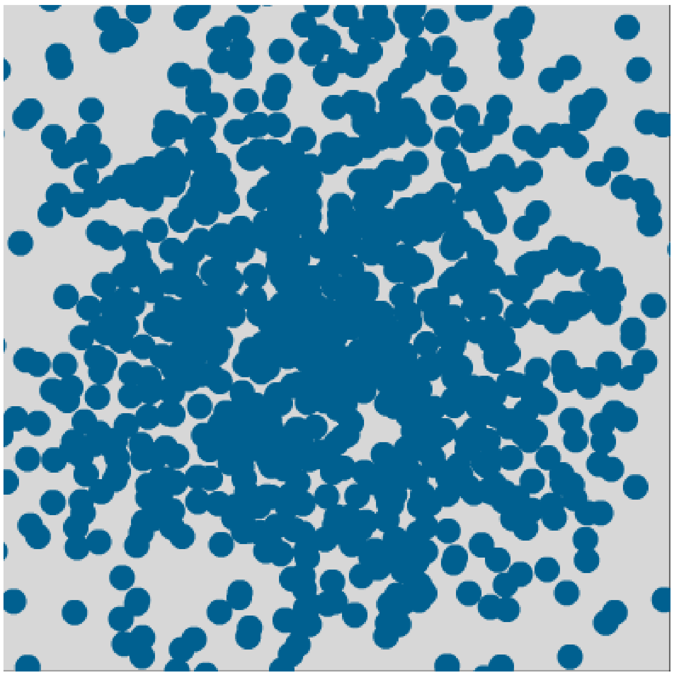
$$m_p(r) = \lim_{\Omega \to \mathbb{R}^d} \frac{\#(Cr(p) \cap \Omega)}{|\Omega|}$$

Thm: m_p is concentrated around $r_p = C_* + p$

[width $\sim \sqrt{d}$]



Life in the ambient space: crackle phenomenon

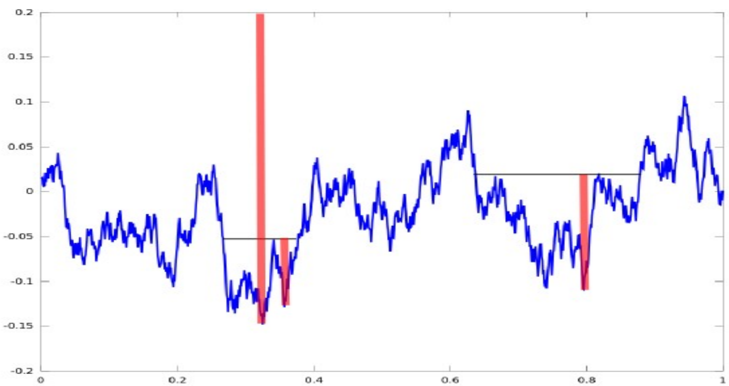


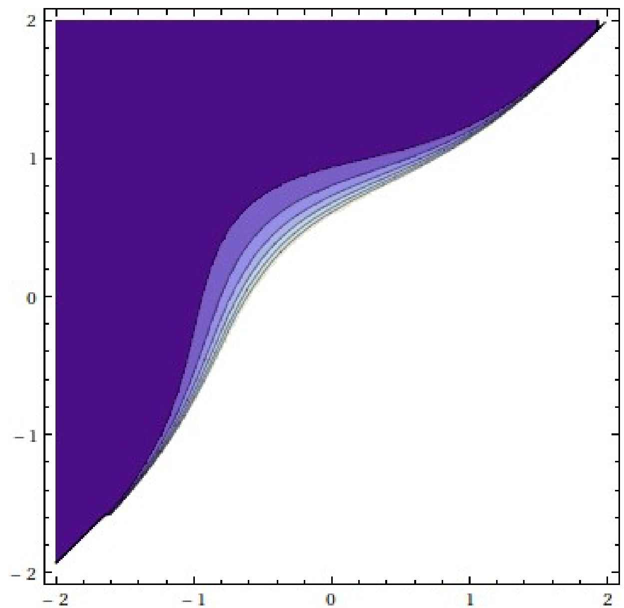
If the noise is Gaussian, the balls form a solid ball. In exponential and subexponential cases, the homologies form in layers.

Plenty of *further models and questions* in stochastic topology

$$\#[x,y] \sim (y-x)^{-1} \int_x^\infty e^{-2\xi^2} d\xi$$

Persistence diagrams for Brownian bridge





Plenty of *further models and questions* in stochastic topology

Topology of the level sets of random functions

Random simplicial complexes

Random configuration spaces

...