INTRODUCTION TO PERSISTENT HOMOLOGY

Andreas Ott 9 May 2019 (EP structures)

Persistent homology

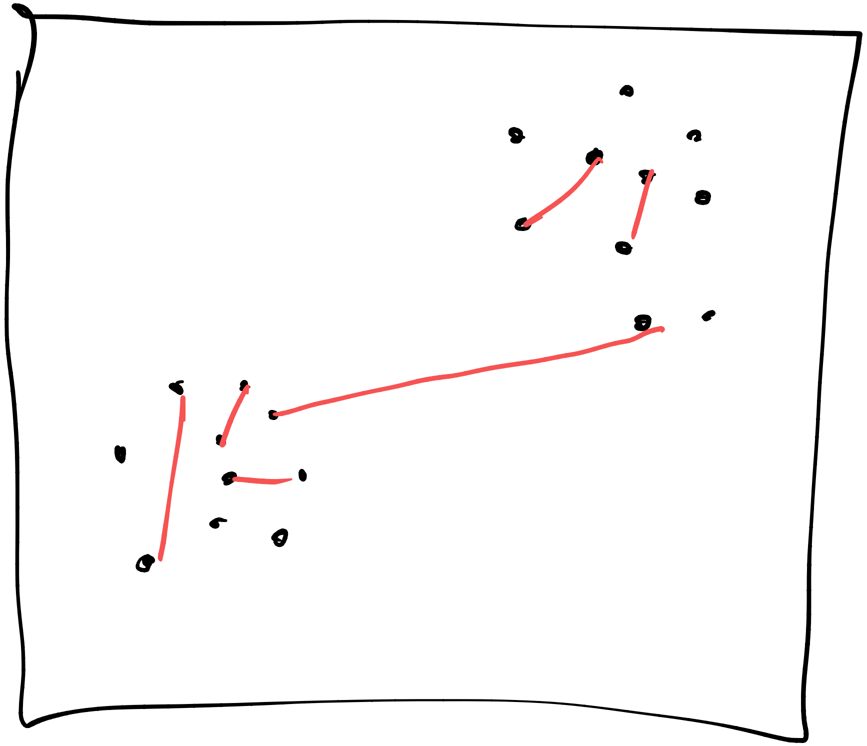
- one of the main tools in TDA
- pioneered by Frosini, Robins ('90s)
  Edelsbrunner, Letscher, Zomorodian ('00s)
  Carlsson, ...

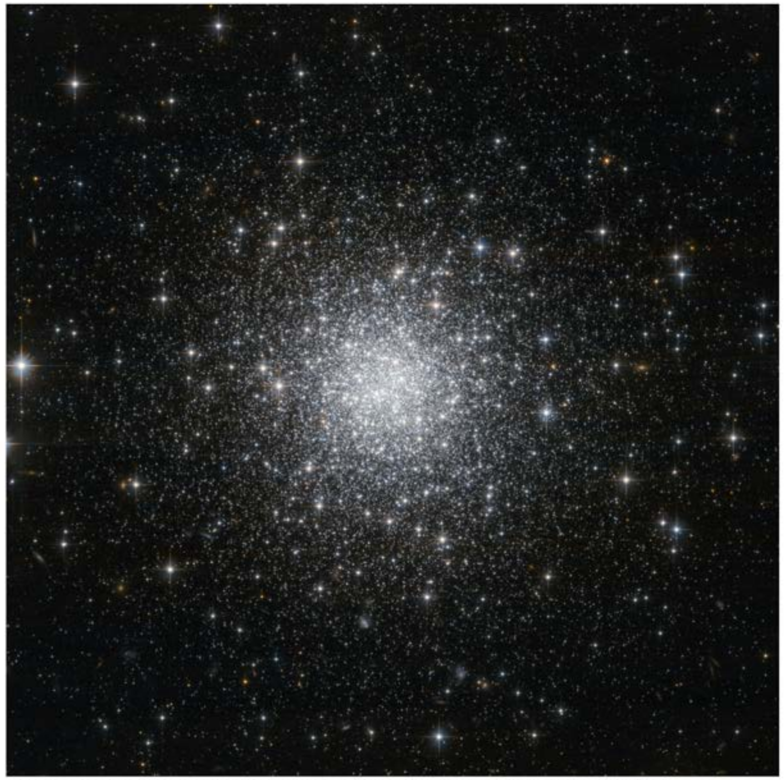
Motivation from astronomy

By ESO/D. Minniti/VVV Team - http://www.eso.org/public/images/eso1141a/, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=17060263

Observe star clusters —

think of stars as point cloud,
i.e. collection of finitely many
pts in same Eucl. vector space





To detect clusters, consider
pairwise distance between
pts.

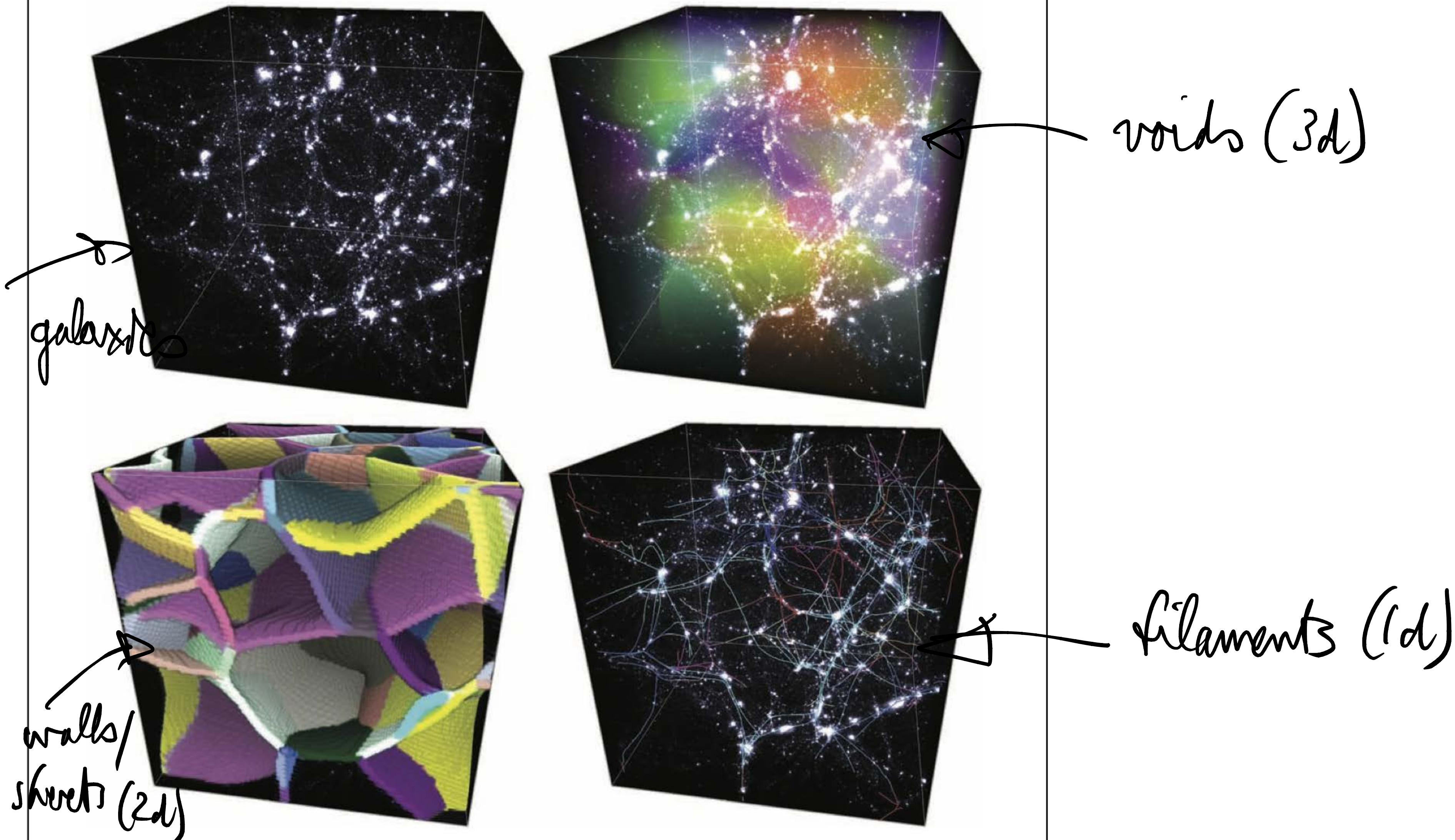


Figure 1. The dark matter density distribution in a $50\,h^{-1}$ Mpc large cosmological simulation (top left-hand panel), with its ascending 3-manifolds (i.e. the voids, top right-hand panel), ascending 2-manifolds (i.e. the walls, bottom left-hand panel) and ascending 1-manifolds (i.e. the filaments, bottom right-hand panel). The manifolds were computed using the method introduced in Sousbie et al. (2009).

Image taken from: T. Sousbie, The persistent cosmic web and its filamentary structure – I. Theory and implementation, Monthly Notices of the Royal Astronomical Society, 2011, doi: 10.1111/j.1365-2966.2011.18394.x

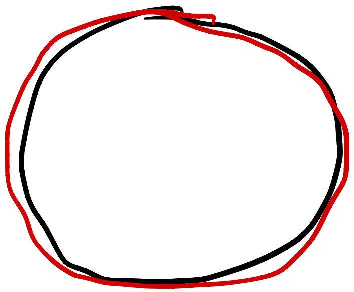
Observe 'holes' i.e. higher dim'l structure.

<u>Problem.</u> How detect 'holes' with a computer?

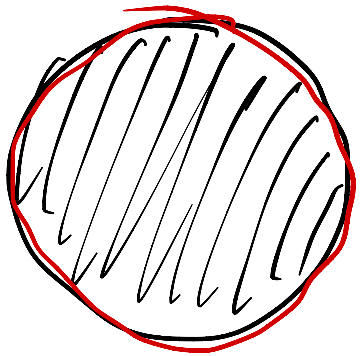
Nontrivial, can be solved with a tool from alg. top. called 'homology'.

Persistent homology

<u>Basic idea:</u> "A hole in some top. space is something that is not the boundary of something else."[9]



hole



no hole

Step 1 How to get a top. space from point cloud?

Here we construct a Vietoris–Rips cplx:

If for any $\begin{Bmatrix} 2 \\ 3 \\ \vdots \\ n \end{Bmatrix}$ pts out of the given pts the pairwise distance is $\leq \varepsilon$ then we include these pts as an $\begin{Bmatrix} \text{edge} \\ \text{triangle} \\ \vdots \\ n\text{-simplex} \end{Bmatrix}$.

Step 2 Compute homology of VR cplx (depends on ε) = combinatorics + linear algebra

Step 3 Display dependence on ε in a compressed, easy-to-read way in terms of a 'barcode'.

The barcode displays the (persistence) of homology classes ⇝ persistent homology (PH).

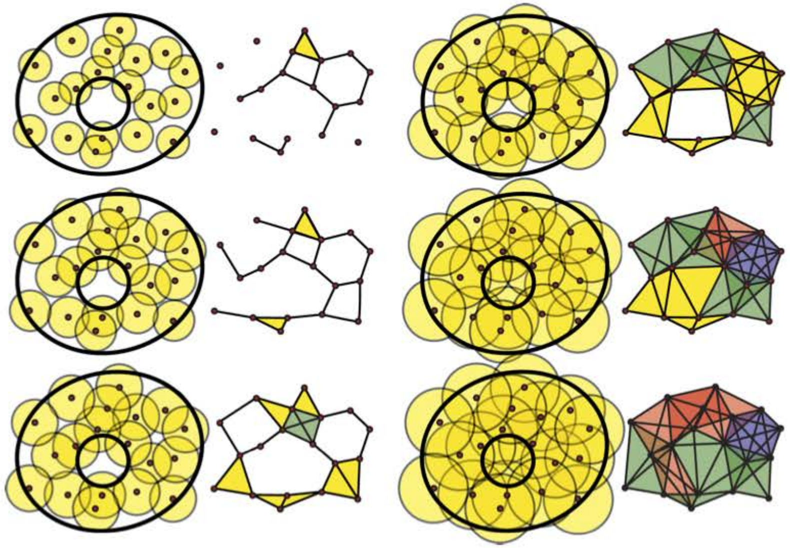
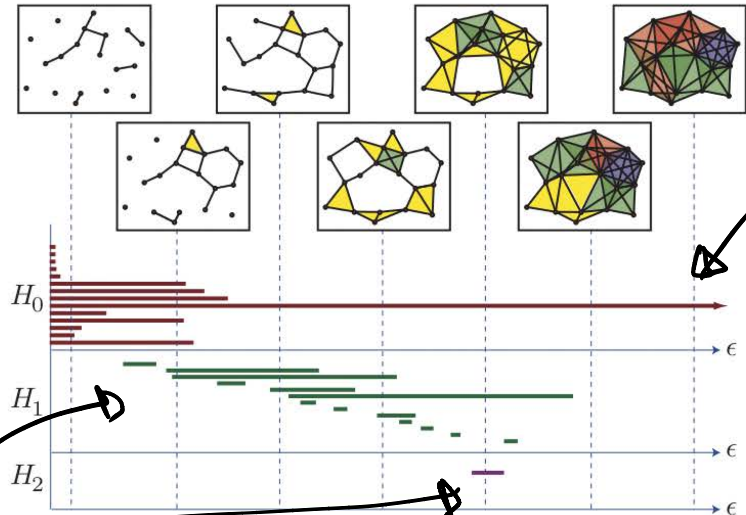


FIGURE 3. A sequence of Rips complexes for a point cloud data set representing an annulus. Upon increasing ϵ, holes appear and disappear. Which holes are real and which are noise?



clusters

barcode

holes

FIGURE 4. [bottom] An example of the barcodes for $H_*(\mathsf{R})$ in the example of Figure 3. [top] The rank of $H_k(\mathcal{R}_{\epsilon_i})$ equals the number of intervals in the barcode for $H_k(\mathsf{R})$ intersecting the (dashed) line $\epsilon = \epsilon_i$.

Images taken from: Robert Ghrist, Barcodes: the persistent topology of data
Bulletin of the AMS, Vol 45, No 1, 2008, 61-75

Moral: The barcode describes the 'shape' of the point cloud data.

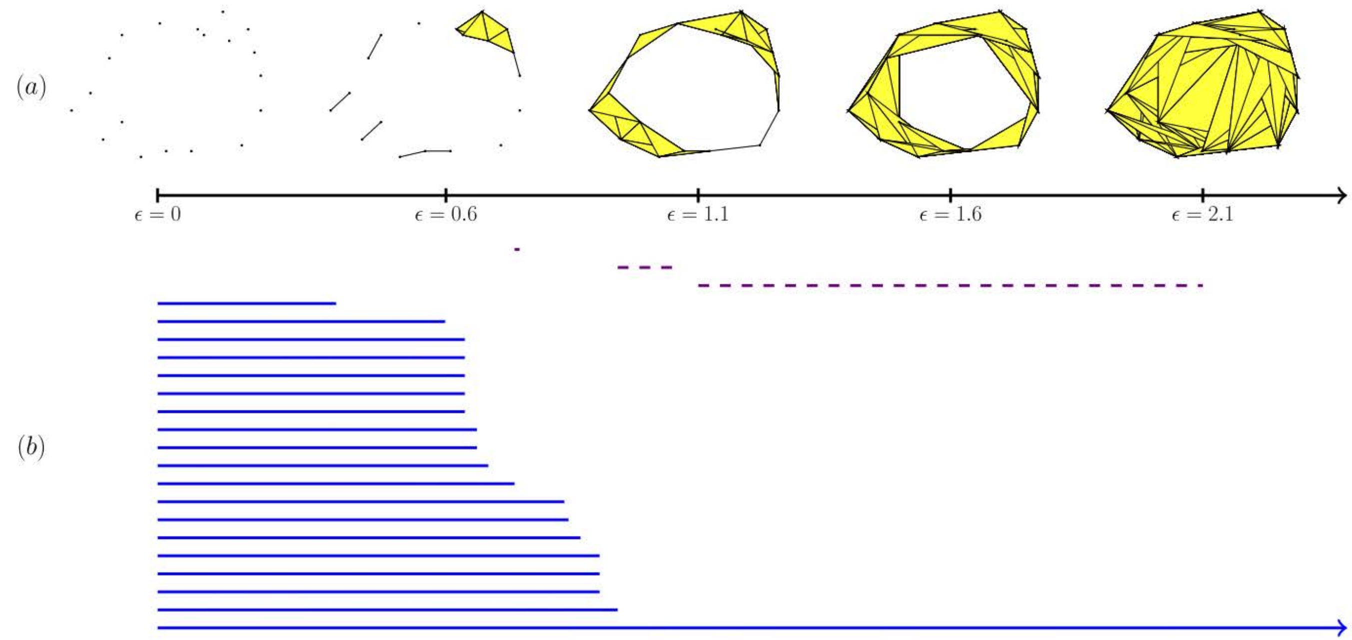


Figure 1 Example of persistent homology for a point cloud. (a) A finite set of points in $\mathbb{R}^2$ (for $\epsilon = 0$) and a nested sequence of spaces obtained from it (from $\epsilon = 0$ to $\epsilon = 2.1$). **(b)** Barcode for the nested sequence of spaces illustrated in (a). Solid lines represent the lifetime of components, and dashed lines represent the lifetime of holes.

Image taken from: Nina Otter, Mason Porter, Ulrike Tillmann, Peter Grindrod, Heather Harrington, A roadmap for the computation of persistent homology
EPJ Data Science, 6:17 2017, doi: 10.1140/epjds/s13688-017-0109-5

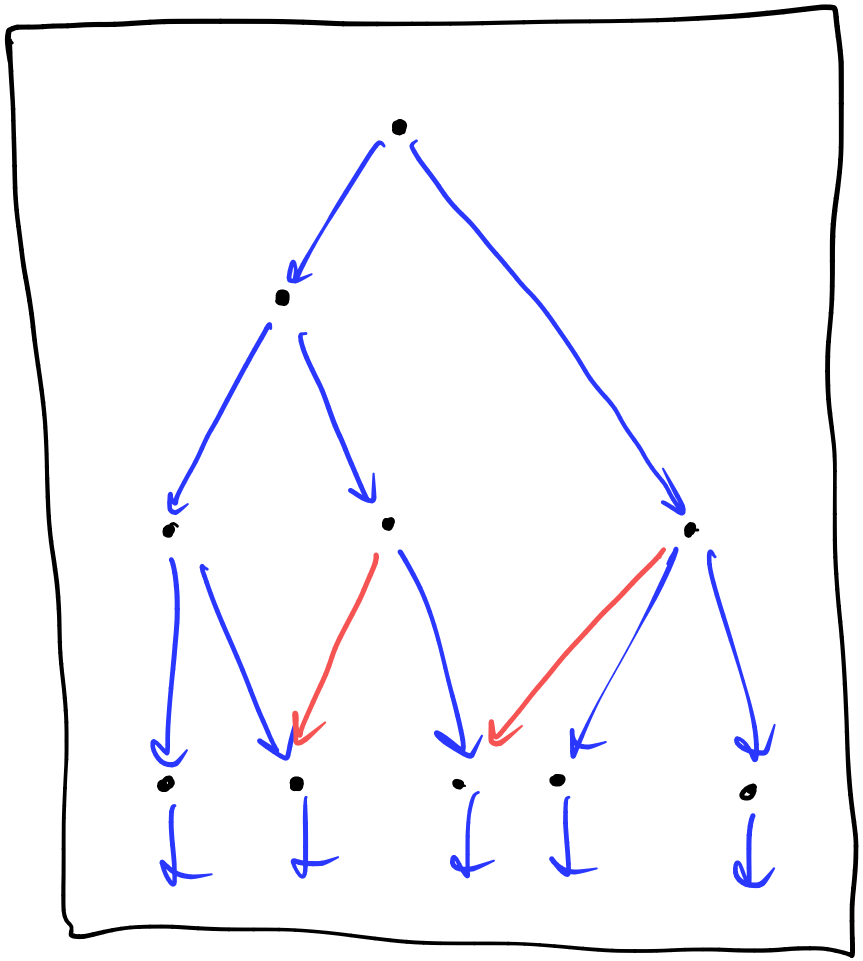
Features of PH

- stable under small perturbations of the point cloud
  ⇝ can handle 'noisy' data

- works more generally for any point cloud that has the structure of a finite metric space

<u>Example</u> Phylogenetics and viral evolution (Chan, Carlsson, Rabadan '13)

A phylogenetic tree ('tree of life') represents the evolutionary relationship between organisms — here we consider certain influenza viruses.



vertical evolution: random mutations over a number of generations

horizontal evolution: mixture of genetic material between individuals of different lineages

Run PH on this dataset:

points (finite) genetic sequences from the alphabet $\{A, C, G, T\}$ of nucleotides

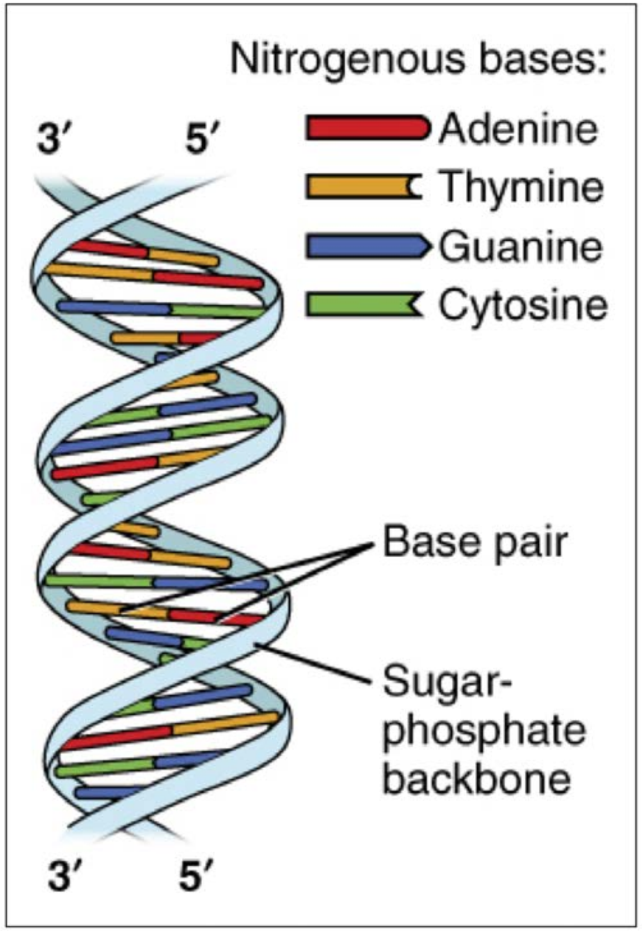
e.g. AATCGGTA

metric Hamming distance

A A T C G G T A

A C T G G G C A

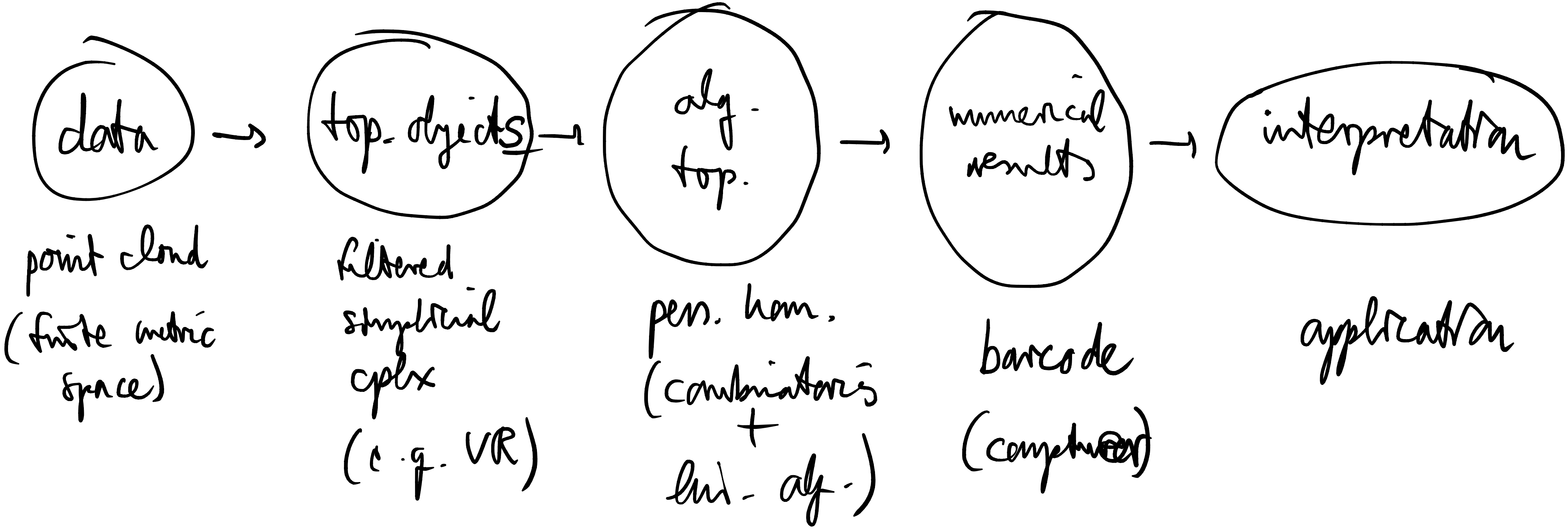
distance := # mismatches (here it's 3)



Upshot: PH detects nontrivial homology

→ supposed to be responsible for epidemic outbreaks of influenza in the past.

The PH workflow



References

- Carlsson: Top. pattern recognition for point cloud data
- OPTGH (Oxford): A roadmap for the computation of PH

Computing

several packages (e.g. JavaPlex, Ripser, ...)

Proposal for EP

- understand app'n (e.g. genetics) in full detail
  - → get data set
  - → run PH
  - → discuss the results with biologists
- find a problem that
  - — biologists want to solve
  - — PH cannot do (yet)
- improve theory.