Презентация на тему:

Топологический анализ данных

Никита Калинин,
Доцент СПбГУ, с.н.с. ВШЭ СПб

Санкт-Петербургский
государственный
университет

Топологический анализ данных

- две группы вконтакте:
  https://vk.com/persistenthomology,
  https://vk.com/club174278716
- книги:

1) замечательные заметки от Антона Айзеберга
http://mathcenter.spb.ru/nikaan/2020/RTFM_homology_V3.pdf

2) Computational Topology: An Introduction.

3) Topological Data Analysis for Genomics and Evolution Topology in Biology

4) Elementary applied topology https://www.math.upenn.edu/~ghrist/notes.html

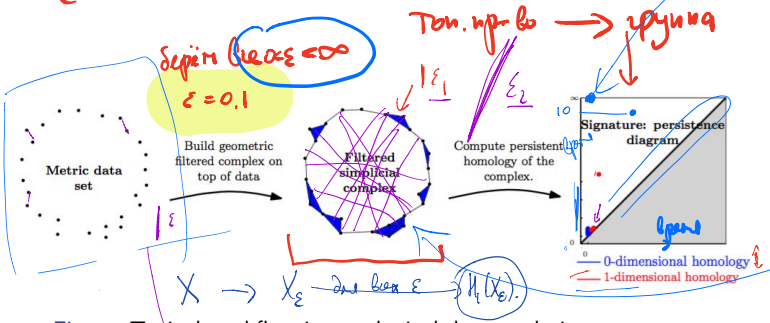
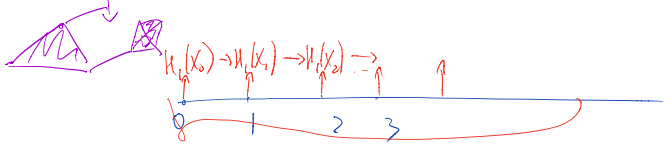


Figure: Typical workflow in topological data analysis
https://en.wikipedia.org/wiki/Topological_data_analysis



Поиск пиков вершин и компрессия сигналов

- Нульмерные гомологии кодируют компоненты связности, $H_0(X) = \mathbb{Z}^k$, где k – число компонент связности X.
- Stability of persistence diagrams https://link.springer.com/article/10.1007/s00454-006-1276-5

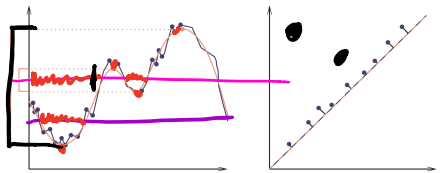


Fig. 2. (a) Two close functions, one with many and the other with just four critical values. (b) The persistence diagrams of the two functions, and the bijection between them.

- Bottleneck distance (минимум попарных расстояний по всем биекциям).
- Теорема: $d_B(D(f), D(g)) \leq |f - g|_\infty$, где $D(f)$ – диаграмма устойчивости функции f.

Мозги

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5026243/pdf/nihms777844.pdf

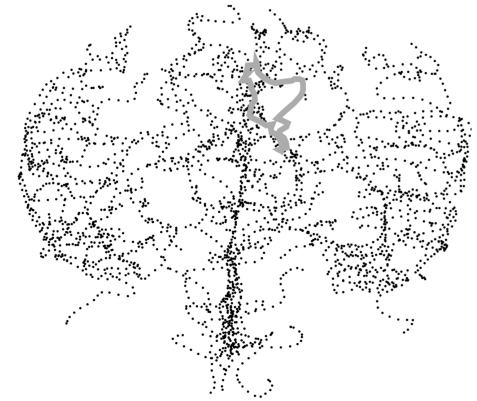
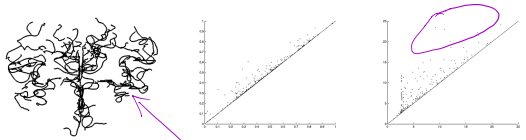
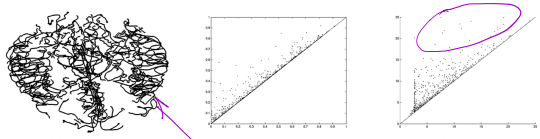


Fig. 2.
A MATLAB rendering of the brain artery tree of Patient 1. Indicated by the thick grey curve is one of the loops formed by thickening the artery tree within the brain. Also found are some of the loops and bends made by the artery tree within the 3-dimensional geometry of the brain.



(a) Brain tree (b) Dgm_0 (c) Dgm_1

Fig. 11.
Persistent homology data objects from a 68-year old. Left: brain tree. Middle: zero-dimensional diagram. Right: one-dimensional diagram.



(a) Brain tree (b) Dgm_0 (c) Dgm_1

Fig. 10.
Persistent homology data objects from a 24-year-old. Left: brain tree. Middle: zero-dimensional diagram. Right: one-dimensional diagram.

Бар-код

- Из книги Антона Айзенберга

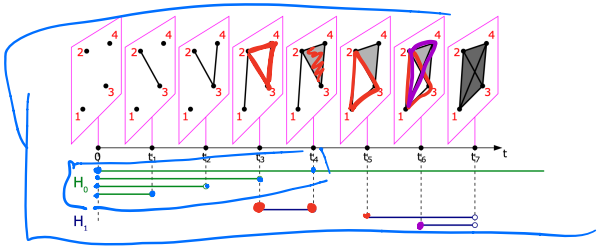


Рис. 18. Мультфильм про симплициальный комплекс: фильтрация

- бар-код и соответствующая диаграмма устойчивости

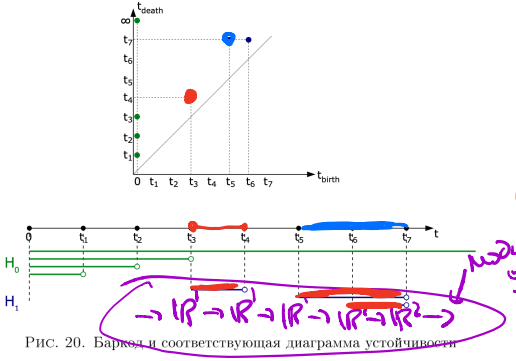
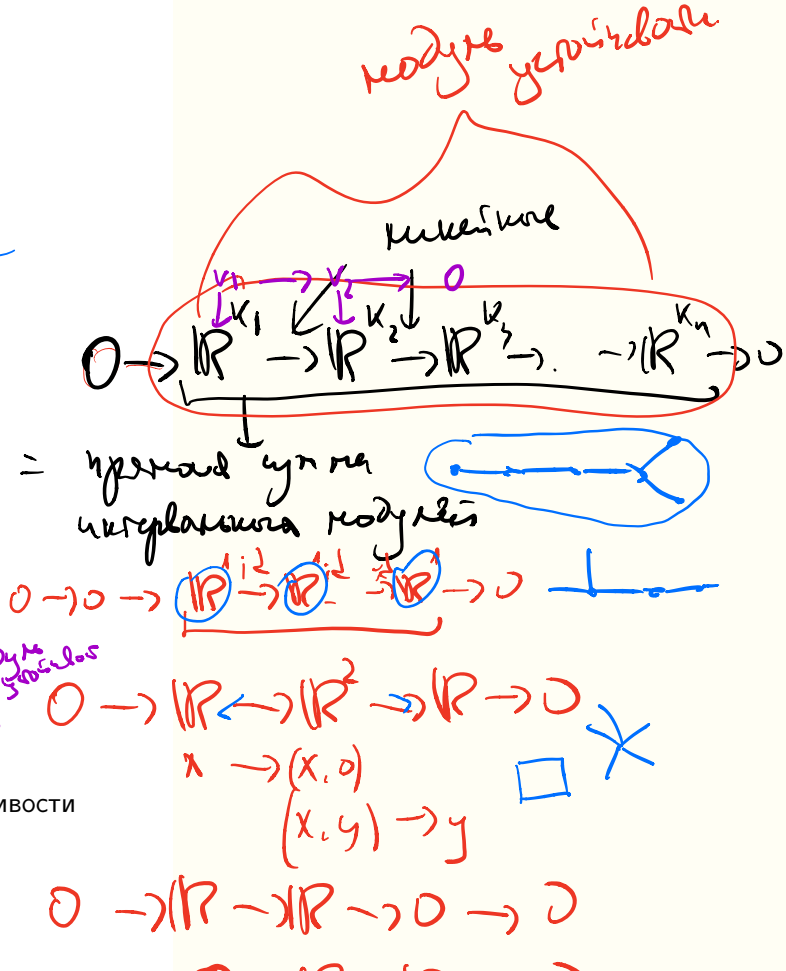


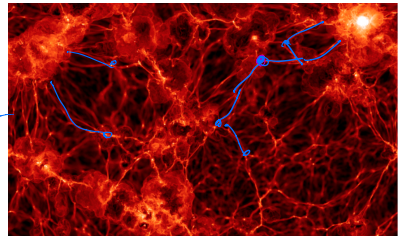
Рис. 20. Баркод и соответствующая диаграмма устойчивости

- Структурная теорема для модулей устойчивости над полем(!).

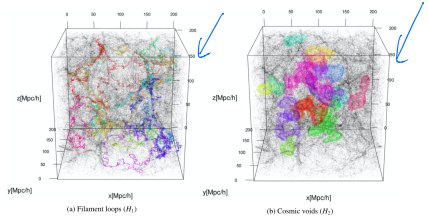


TDA в астрономии

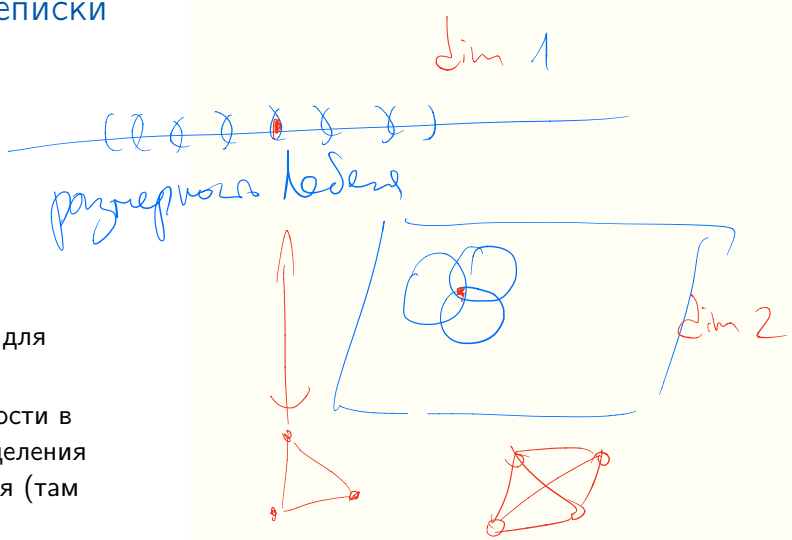
- Что было: считать сколько пар объектов на заданном расстоянии (two-point correlation function). Брать три точки?
- Имеются filament loops и cosmic voids (без чётких определений). Можно анализировать имеющиеся данные и сравнивать их с моделями именно с точки зрения устойчивых гомологий – а как ещё подбирать параметры для модели?



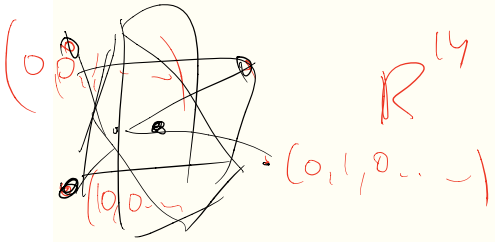
https://www.wired.co.uk/article/dipole-repeller-milky-way



TDA в биологии (из личной переписки с Е. Андроновым)



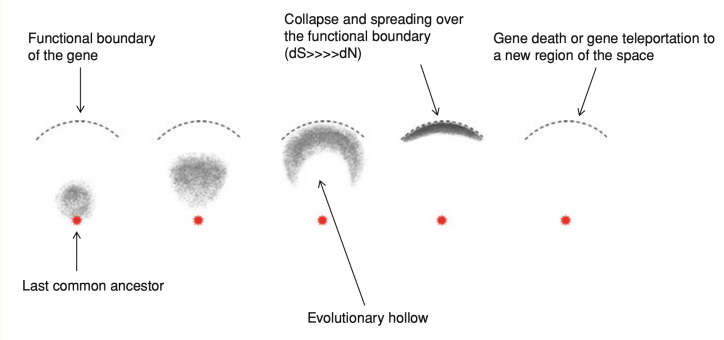
- алгоритм размещения множества точек, для которых есть матрица:
- поиск симплекса максимальной размерности в матрице попарных расстояний для определения минимальной размерности представления (там было 13D);
- использование его вершин как системы многомерной "GPS" и определение метрических координат каждой точки в этой системе;
- построение серий срезов полученного множества для визуализации паттернов (искали прежде всего дырки).

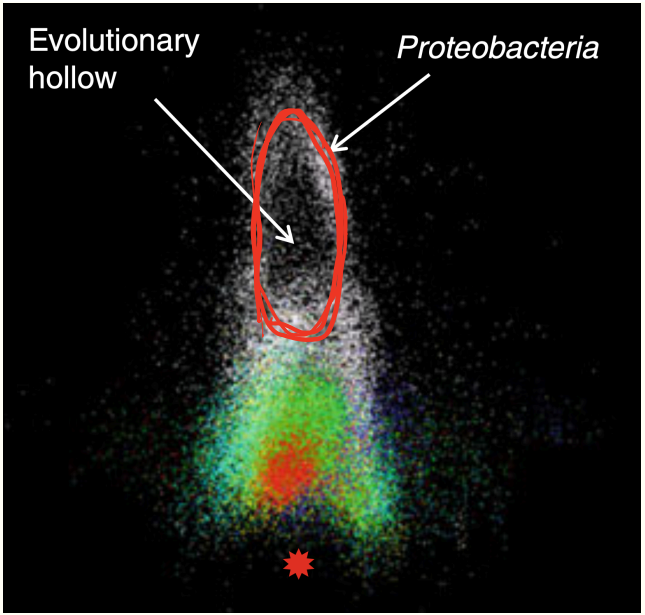


TDA в биологии (из личной переписки с Е. Андроновым)

- Ген 16S rRNA часто переносился горизонтально.
- Интересно с точки зрения реконструкции эволюционного процесса в этом пространстве, который должен быть примерно такой:
- начинаться с одной точки (мы попробовали найти место, где она была, сейчас оно пустое);
- представлять собой необратимое радиальное расширение (хотя все не так просто);
- в ходе этого расширения предковые формы должны необратимо вымываться, искали пустые места, где они когда-то были. Это довольно интересно и непросто, так как есть пустые места, где никогда ничего не было (или пока не было, или уже никогда не будет, или в принципе невозможно).
  Типа Большого взрыва, только в эволюции прокариотического разнообразия. Проблема только в том, что все это можно увидеть только в "достаточной" размерности.

The Evolutionary Space Model to be Used for the Metagenomic Analysis of Molecular and Adaptive Evolution in the Bacterial Communities, Pershina et al





Распространение ZIKV вируса в Бразилии

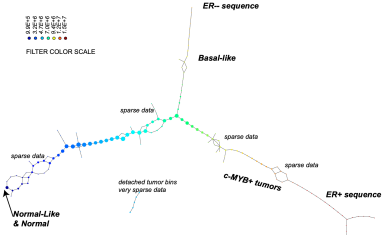
- Modeling the spread of the Zika virus using topological data analysis,

  https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0192120
- можно строить диффуры (но там мало параметров и не учитывают географию), можно пытаться из текущей карты заболевания, погоды и количества комаров вывести, но тоже так себе (потому что данные приходят в разнобой и с большой задержкой, а про комаров так совсем плохие).
- отметили райцентры, вставили туда количество заболевших, комаров и пр, и потом показали, что простая топология приводит к большей скорости заболеваний, сложная – к медленной (добавили топологические параметры в список фич по которым регрессия строится)
- комары и вирусы не летают через гомологические дырки.

Отступление: TDA (topological data analysis) для рака груди

https://www.pnas.org/content/108/17/7265

- Устойчивые гомологии выявляют дырки, а бывают щупальца.
- рак очень разный, нужно много данных чтобы классифицировать, данные не жёсткие
- полезная тулза: Mapper
- $f : X \to [0, 1], \bigcup U_i = [0, 1]$ кластеризуем в $f^{-1}(U_i)$ и соединяем два кластера ребром, если они пересекаются.
- препроцессинг данных: строчки — гены и протеины, столбцы – пациенты, выделение "болезнетворной" компоненты, из неё же функция для Mapper.
- Результат: выделен тип рака, который всегда излечивается, клетки не похожие на здоровые в этом случае (излечиваемый рак с клетками, похожими на здоровые, был известен)



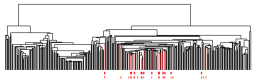


Fig. 4. Clustering vs. *PAD*. Can *Mapper* extract something new from the data that clustering does not? We compare the outputs of clustering (average linkage) vs. *Mapper* as applied to the same exact data matrix (*DSGA*-transformed *NKI*) to show that these two procedures are different. The bins defining the *c-MYB*$^+$ group were marked on the cluster dendrogram (red for the tighter—no outliers—group, and orange for the larger *c-MYB*$^+$ group containing outliers) The *c-MYB*$^+$ tumors are scattered among different clusters, but *PAD* has been able to extract this group that turns out to be both statistically and biologically/clinically coherent.

Место для заметок

Место для заметок

Место для заметок