ГЕОМЕТРИЧЕСКИЕ АСПЕКТЫ АЛГОРИТМОВ РАБОТЫ С ВЫСОКОРАЗМЕРНЫМИ ДАННЫМИ В МАТЕМАТИЧЕСКОМ МОДЕЛИРОВАНИИ

Александр Бернштейн
a.bernstein@skoltech.ru

Москва, ВМК МГУ им. М.В. Ломоносова, 3 декабря 2025

Skoltech
Skolkovo Institute of Science and Technology

Программирование и вычислительная математика

- решение задач математического моделирования – один из важнейших стимулов развития вычислительной математики и программирования

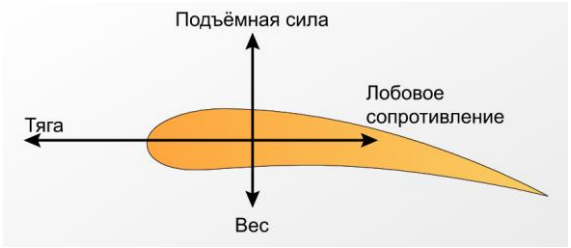
  Трифонов Н.П. Решение на ЭВМ задач структурного анализа кристаллов (кандидатская диссертация, руководитель С.Л. Соболев)

- моделирование на основе данных– один из современных трендов в математическом моделировании и предиктивной аналитике как частей искусственного интеллекта

- современные алгоритмы работы с данными являются «математикоемкими» и используют широкий спектр глубоких математических методов, включая дифференциально-геометрические и топологические методы

Математические модели на первых принципах

Аэродинамические характеристики
- подъемная сила (L),
- лобовое сопротивление (C), …



Летательный аппарат
- поверхность (S),
- тяга двигателя,
- вес, …

Режим полета
- скорость (V)
- угол атаки (α), …

Аэродинамическое проектирование: построить S:
- $L \rightarrow \max$
- $C \le C_{пред}$

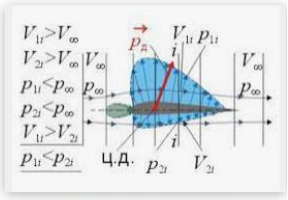
Аэродинамические расчетные модели

$L = F_L(S, V, \alpha)$

$C = F_C(S, V, \alpha)$

$$X - \frac{1}{\rho}\frac{\partial p}{\partial x} - \frac{\partial(\overline{V^2})}{\partial x} = 2(\eta w - \zeta v),$$
$$Y - \frac{1}{\rho}\frac{\partial p}{\partial y} - \frac{\partial(\overline{V^2})}{\partial y} = 2(\zeta u - \xi w),$$
$$Z - \frac{1}{\rho}\frac{\partial p}{\partial z} - \frac{\partial(\overline{V^2})}{\partial z} = 2(\xi v - \eta u),$$
$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0,$$

Аэродинамика (I)



5.4. Основные законы аэродинамики

Оптимизационная задача: построить S:

$S = \arg\max_S F_L(S, V, \alpha)$

$F_C(S, V, \alpha) \le C_{пред}$

Математические модели на первых принципах:

аэродинамические расчетные модели (CFD коды)
- пример: уравнения Эйлера с нерегулярной сеткой
- время расчёта порядка 10 ч.

Skoltech
Skolkovo Institute of Science and Technology

Математические модели на первых принципах: массовые расчеты

EU FP7 Project FFAST (2010 – 2013): Future Fast Aeroelastic Simulation Technologies

- **about 100 000 design layout releases are required for a new passenger aircraft**
- **each layout should be analyzed for 15 to 20 control and environment scenarios**

Number of conditions (cases) that are required in the Critical Loads Analysis of a large civil aircraft:

50 - flight points (altitude and speed)
100 - mass cases (loaded weight and weight distribution)
4 - control laws
10 - control surface configurations
5 - manoeuvres and gusts (gradient lengths)

$$\text{Total analysis cost/time} \approx N_{case} \times C_{sim} \times \frac{N_{sim}}{N_{span}}$$

N_{case} = Number of test cases required to define the envelope of critical loads
C_{sim} = Cost/time of each full order simulation
N_{sim} = Number of full order simulations required to construct each reduced order model
N_{span} = Number of loads cases spanned by each reduced order model

TOTAL NUMBER OF REQUIRED CASES: N_{case} = 10 000 000

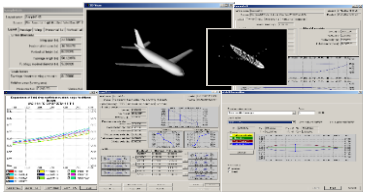
- **Boeing: a total of 800,000 cray-hours were spent for simulation of all the systems in new B-787**
- **Упрощение моделей, основанных на первых принципах** («не учет» некоторых физических феноменов, поиск приближенных решений, и т.п.), **не решает проблему кардинально**

Skoltech
Skolkovo Institute of Science and Technology

Построение аэродинамической модели по данным (1)

- **Family of Fast Aerodynamic Surrogate Models** developed in **Russian Academy of Sciences** **2004 – 2007**
- Implemented in **AIRBUS Engineering Tools for Aerodynamic design**

- **РАН:** Центр программных технологий РАН → ФИЦ ИУ РАН (ИСА), ИППИ РАН
- **ЦАГИ**



Airbus experts: application of such surrogate models "**provides the reduction of up to 10% of lead time and cost in several areas of the aircraft design process**"

1) Генерация данных $\{(S_i, V_i, \alpha_i, L_i \approx F_L(S_i, V_i, \alpha_i), C_i \approx F_C(S_i, V_i, \alpha_i)), i = 1, 2, \dots, N\}$

- построена математическая параметрическая модель 3D-поверхности самолета
- сгенерирована множество {S} поверхностей (~ 6000, включая ~ 27 000 аэродинамических профилей крыла)

технологии генеративного искусственного интеллекта

- с помощью CFD-кода (основан на решении уравнений полного потенциала) - **ЦАГИ**
  - для каждой поверхности S
  - для разных значений скорости (V) и угла атаки (α)

были вычислены подъемная сила $L \approx F_L(S, V, \alpha)$, сопротивление $C \approx F_C(S, V, \alpha)$, и другие характеристики (~ 20)

Skoltech
Skolkovo Institute of Science and Technology

Построение аэродинамической модели по данным (2)

2) Построение по данным неизвестных зависимостей (регрессионные методы/машинное обучение):

данные $\{(S, V, \alpha); (L \approx F_L(S, V, \alpha), C \approx F_C(S, V, \alpha), \ldots)\} \rightarrow$ **зависимости** $L = F_L(S, V, \alpha), C = F_C(S, V, \alpha), \ldots$

Построенная по данным аэродинамическая модель – «заменитель» математической модели «на первых принципах» (суррогатная модель, метамодель)

Построенная суррогатная модель:

- **ускорение вычислений в 360 000 раз:** время расчета **20 аэродинамических характеристик** одной компоновки для ~ **65 различных режимов полета**:
  - CFD-код (уравнение полного потенциала) ~ **2 CPU hours,**
  - построенная суррогатная модель ~ **20 CPU milliseconds**
- **относительная ошибка ~ 1%**

Reynolds number	$Re = 3\times10^6$	$Re = 3\times10^7$
Lift coefficient (**CL**)	0.848	0.796
Lateral Force coefficient (**CY**)	0.843	0.721
Profile drag coefficient (**CDFP**)	0.666	0.776
Friction drag coefficient (**CDF**)	0.319	0.234
Total drag coefficient (**CD**)	1.039	1.547
Derivatives of CL with respect to Angle of Attack	1.076	1.094

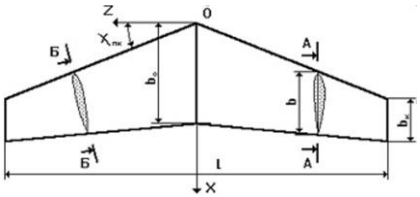
- A.V. Bernstein, A.P. Kuleshov, Yu.N. Sviridenko, V.V. Vishinsky. Fast Aerodynamic Model for Design Technology. Workbook «West-East High Speed Flow Field Conference» (November 19-22, **2007**, Moscow, Russia). 2007, 125-126
- А.В. Бернштейн, В.В. Вышинский, А.П. Кулешов, Ю.Н. Свириденко. Быстрый метод аэродинамического расчета для задач проектирования. Труды ЦАГИ “Применение искусственных нейронных сетей в задачах прикладной аэродинамики” 2678, **2008**, с. 35—45.

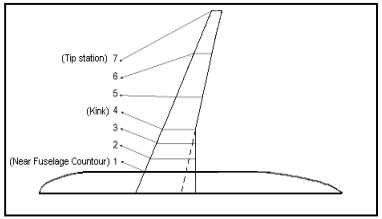
ИИ- интерпретация построенной модели

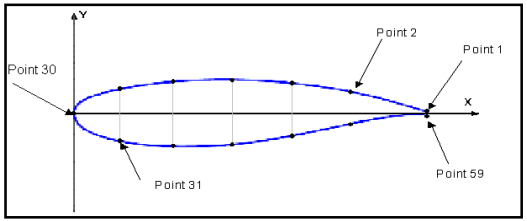
Skoltech
Skolkovo Institute of Science and Technology

Работа с многомерными данными (1)

Поверхность самолета - вектор высокой размерности







Профили крыла:

- профиль – многомерный вектор p ~ 50 ÷ 200 (**ЦАГИ: p = 59**)
- 7 профилей крыла: **59×7 = 413**

Проблемы построения математических моделей по данным

- **мир многомерен** – реальные объекты описываются векторами высокой размерности
- **многомерные данные «трудны» для анализа и обработки** (проклятие размерности, феномен «пустого пространства», …)

Skoltech
Skolkovo Institute of Science and Technology

Работа с многомерными данными (2)

Проклятие размерности (Ричард Беллман, 1961): экспоненциальный рост

- числа необходимых экспериментальных данных в зависимости от размерности пространства при решении вычислительных, оптимизационных, вероятностно-статистических задач регрессии, распознавания образов, классификации, машинного обучения, дискриминантного анализа, и др.;
- числа вариантов в комбинаторных задачах в зависимости от размерности исходных данных, что приводит к соответствующему росту сложности переборных алгоритмов

Иллюстративный пример

- $\psi(X)$: $[0, 1]^p \rightarrow R^1$ – неизвестная Липшицева функция ($\psi \in$ **Lip**) на единичном p-мерном кубе
- $\hat{\psi}(X)$ – произвольная оценка построенная по зашумленной выборке $\{(X_i, \psi(X_i)), i = 1, 2,\ldots, n\}$

Неасимптотическая нижняя граница точности (Ибрагимов, Хасьминский (1979), Stone (1982)):

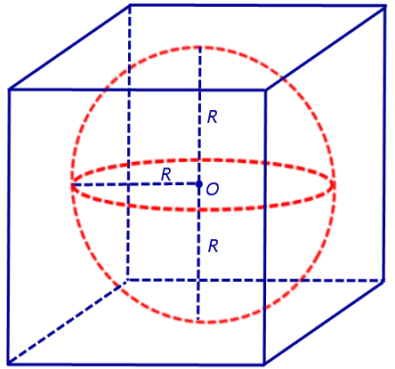
$$\sup_{\psi \in \mathbf{Lip}} E(\psi(X) - \hat{\psi}(X))^2 \geq \text{Const} \times n^{-2/(2+p)}$$ (константа не зависит от n и p)

p = 10: n = 10 000 измерений для достижения **заданной точности**

p = 20: n ~ 10 000 000 измерений для достижения **той же точности**

Работа с многомерными данными (3)

Феномен «пустого пространства

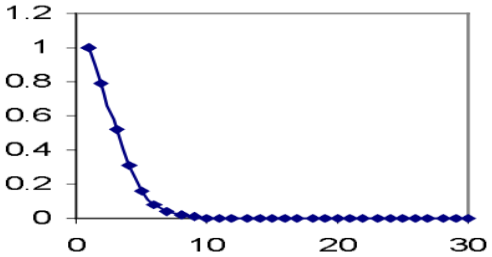


1. Объемы гиперкуба и вписанного в него шара

C(R, p) = [-R, R]^p - p-мерный куб V(C(R, p)) = $(2R)^p$

B(R, p) = { x $\in$ R^p: $|x| \le R$} - p-мерный шар $V(B(R,p)) = \frac{\pi^{p/2} \times R^p}{\Gamma\left(\frac{p}{2}+1\right)}$

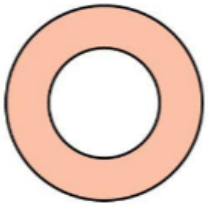
$$\lim_{p\to\infty} \frac{V(B(R,p))}{V(C(R,p))} = 0$$



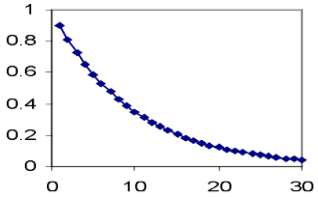
0,1 для p = 6

0,0025 для p = 10

2. Объем тонкой сферической оболочки



$$\lim_{p\to\infty} \frac{V(B(1,p)) - V(B(1-\varepsilon,p))}{V(B(1,p))} = 1$$

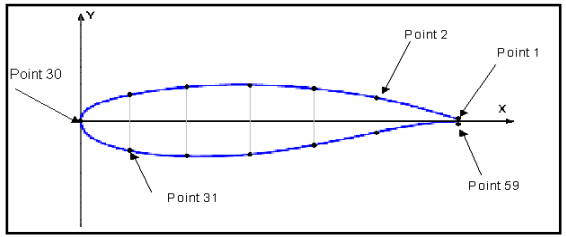


Skoltech
Skolkovo Institute of Science and Technology

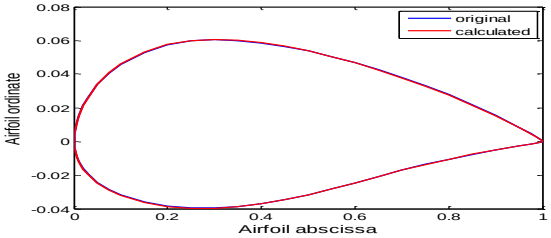
Работа с многомерными данными (4)

Концентрация мер, разреженная структура, **низкоразмерная внутренняя структура «реальных» данных:**

- носители **реальных** данных лежат на **малой** части «объемлющего» многомерного пространства,
- могут быть описаны/параметризованы **небольшим числом параметров**



59 → 6



A. Bernstein, E. Burnaev, S. Chernova, F. Zhu, N. Qin. **Comparison of Three Geometric Parameterization methods and Their Effect on Aerodynamic Optimization.** Eurogen-2011, Sira, Italy, September 14-16, 2011, pp. 758–772

- **вычислительно эффективные алгоритмы** работы с многомерными данными (предиктивная аналитика, машинное обучение, Искусственный интеллект, …) в явном или неявном виде **используют низкоразмерную внутреннюю структуру носителей данных**
- **нахождение** по данным **внутренней структуры носителей дан**ных и их использование в алгоритмах – отдельное научное **направление исследований**

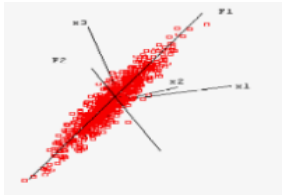
‘Machine learning is about the shape of data’ (ICML’2014)

Skoltech
Skolkovo Institute of Science and Technology

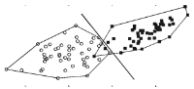
Методы работы с многомерными данными (20 век)

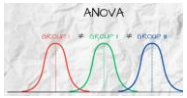
Многомерный статистический анализ

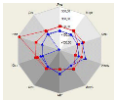
Снижение размерности
PCA - метод главных компонент (1903)



Регрессионный анализ (1795/1805/…)
Дискриминантный анализ (1936)
Дисперсионный анализ (1918)
Факторный анализ (1904), **…**

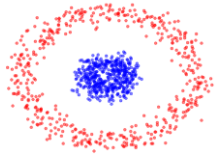






Линейные модели данных
Гауссовские распределения

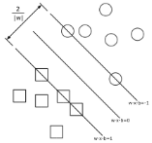
PCA → Спектральное разложение матриц/тензоров → «тензорные поезда» →



Эвристические алгоритмы работы с «нелинейными» многомерными данными

Метод потенциальных функций (1964) → **Kernel PCA** (1998)
(методы гильбертова пространства с воспроизводящим ядром)

Метод опорных векторов (1963)



Искусственные нейронные сети/автоэнкодеры (1974 → 1991 → …)

Отсутствие математических моделей многомерных данных

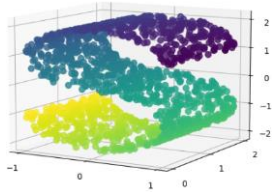
Skoltech
Skolkovo Institute of Science and Technology

Новые «математикоемкие» методы работы с многомерными данными

Моделирование многообразий (Manifold Learning) (2000)

Топологический анализ данных (Topological Data Analysis) (2009), **…**

Нелинейная модель многомерных данных: многомерные данные лежат на (или вблизи) **нелинейного многообразия невысокой размерности** (**Многообразия данных**), вложенного в **высокоразмерное пространство наблюдений** Seung, Lee: The Manifold Ways of Perception. Science, 2000



Гипотеза многообразия: выборка p-мерных данных $\mathbf{X}_n = \{X_1, X_2, \dots, X_n\} \subset R^p$:

- лежит на **неизвестном** Многообразии данных **M** вложенном в p-мерное пространство ($\mathbf{M} \subset R^p$)
- многообразие имеет **неизвестную** внутреннюю размерность q = Dim **M**
- получена в соответствии с **неизвестным** вероятностным распределением μ на **M**

Моделирование многообразий: задачи

По заданной выборке решаются различные задачи:

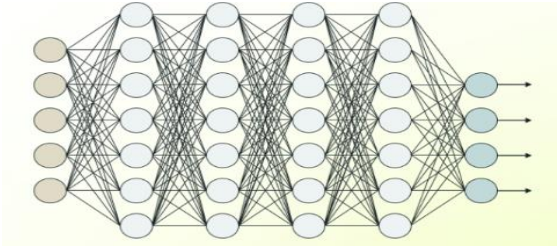
- **снижение размерности:** при выбранном q, построить низкоразмерные описания данных

  $\mathbf{X}_n = \{X_1, X_2, \dots, X_n\} \subset R^p \to \mathbf{Y}_n = \{y_1, y_2, \dots, y_n\} \subset R^q$:

  (построить отображение вложения $h: X \in \mathbf{M} \subset R^p \to y = h(X) \in R^q$) с заданными свойствами)

- оценить внутреннюю размерность q = Dim **M**
- построить оценку $\widehat{\mathbf{M}}$ многообразия данных **M**
- построить оценки различных элементов многообразия (касательных пространств $T_{\mathbf{M}}(X)$, риманова тензора $Q_{\mathbf{M}}(X)$, …, в различных точках многообразия $X \in \mathbf{M}$)
- оценить вероятностное распределение μ на **M**
- решение различных статистических задач (регрессии, классификации, …), носители данных которых лежат на неизвестном Многообразии данных

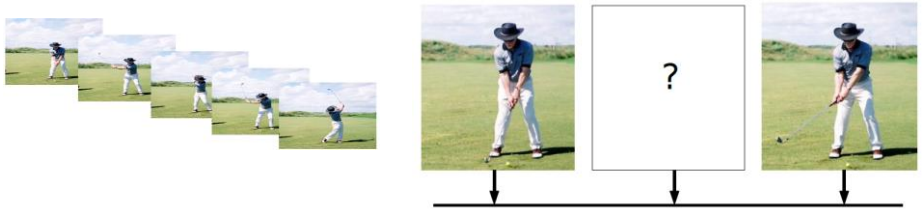
Алгоритмы снижения размерности применяются не только к исходным данным, но и к данным, возникающим на промежуточных слоях глубоких нейронных сетей



Skoltech
Skolkovo Institute of Science and Technology

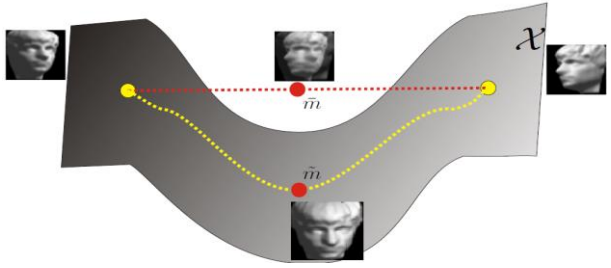
Моделирование многообразий: пример 1 (геодезические линии)

Задачи повышения четкости изображений



Линейная интерполяция

Линейные методы не учитывают, что реальные изображения «живут» на **нелинейных** низкоразмерных структурах

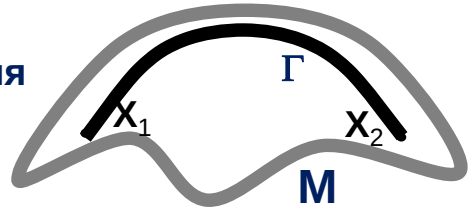


нелинейные методы

Skoltech
Skolkovo Institute of Science and Technology

Моделирование многообразий: примеры (1)

Геодезические линии на многообразии – кратчайшие пути → расстояния



ISOmetric MAPing (ISOMAP) **Tehenbaum, de Silva, Langford: A global geometric framework for nonlinear dimensionality reduction, 2000**

Многомерное шкалирование: сохранение расстояний в низкоразмерных данных

$$\Delta_{\text{MetricMDS}} = \sum_{i,j=1}^{n} \left(\|X_i - X_j\|^2 - \|y_i - y_j\|^2 \right)^2$$

ISOMAP: евклидовы расстояния заменяются длинами геодезических $D(X_i, X_j)$ (**оцениваются**)

$$\Delta_{\text{ISOMAP}} = \sum_{i,j=1}^{n} \left(\left(D(X_i, X_j) \right)^2 - \|y_i - y_j\|^2 \right)^2$$

ISOMAP-решение: **гильбертово пространство** со специфическим **воспроизводящим ядром**

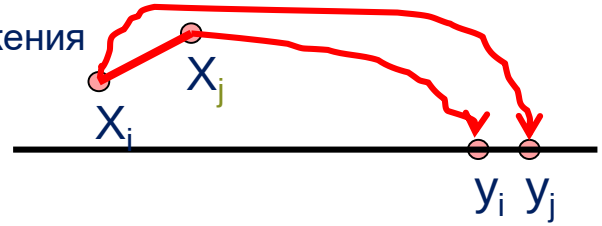
$$K_{\text{ISOMAP}}(X_i, X_j) = - D^2(X_i, X_j) - \sum_{k=1}^{n} D^2(X_k, X_i) - \sum_{k=1}^{n} D^2(X_j, X_k) + \sum_{k,s=1}^{n} D^2(X_s, X_k)$$

LogMap (2005), Riemannian manifold learning (2006, 2008), …

Skoltech
Skolkovo Institute of Science and Technology

Моделирование многообразий: пример 2 (уравнения на многообразиях)

$\mathbf{X}_n$ – обучающая выборка, h: $\mathbf{M} \subset R^p \to R^q$ – желаемое отображение вложения



$$|h(X') - h(X)| \le |\nabla_{\mathbf{M}} h(X)| \times |X' - X| + o(|X' - X|) \qquad \nabla h(X) = \begin{pmatrix} \frac{\partial h}{\partial x_1} \\ \cdots \\ \frac{\partial h}{\partial x_p} \end{pmatrix}$$

$$F(h) = \int_{\mathbf{M}} |\nabla_{\mathbf{M}} h(X)|^2 mes(dX) \to \min$$

Теорема Стокса: $F(h) = \int_{\mathbf{M}} (h \times \Delta_{\mathbf{M}} h)(X) mes(dX)$

$\Delta_{\mathbf{M}}(h)$: $h(X) \to -\operatorname{div}(\nabla_{\mathbf{M}} h(X)) = -\sum_{k=1}^{p} \frac{\partial^2 h(X)}{\partial x_k^2}$

- оператор Лапласа-Бельтрами

Компоненты оптимального вложения $h(X) = \begin{pmatrix} h_1(X) \\ \cdots \\ h_q(X) \end{pmatrix}$ состоят из q **собственных функций оператора Лапласа-Бельтрами**, отвечающих q минимальным собственным числам

Laplacian Eigenmaps (Belkin, Niyogi: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, **2003**)

- строится n×n **матрица Лапласиана** $L_{LE}(\mathbf{X}_n)$ – выборочный аналог оператора Лапласа-Бельтрами
- n-мерные векторы $(h_k(X_1), h_k(X_2), \dots, h_k(X_n))^T$, состоящие из k-ых компонент векторов $\{h(X_i)\}$, k = 1, 2, … , q, являются собственными векторами матрицы Лапласиана, отвечающих q минимальным собственным числам

Моделирование многообразий: пример 3 (касательные расслоения)

Оценка обобщающей способности методов снижения размерности в произвольной точке $X \in \mathbf{M}$

$$\gamma_- \times d_{P,2}(T_M(X), T_{\widehat{M}}(\widehat{X})) \le \delta^*(X) \le \gamma_+ \times d_{P,2}(T_M(X), T_{\widehat{M}}(\widehat{X}))$$

- $\delta^*(X)$ – некоторая характеристика обобщающей способности
- $T_M(X)$ - касательное пространство к многообразию M данных в точке X
- $T_{\widehat{M}}(\widehat{X})$ - касательное пространство к построенной по данным оценке $\widehat{M}$ многообразия данных в
- «восстановленной» точке $\widehat{X}$
- $d_{P,2}(L, \hat{L})$ – расстояние между q-мерными линейными пространствами в R^p (проекционная 2-норма на многообразия Грассмана)
- γ_- и γ_+, $0 \le \gamma_- \le \gamma_+ \le 1$, - константы

Bernstein, Kuleshov. **Manifold Learning: generalizing ability and tangent proximity, 2013**

Задача оценивания касательного расслоения многообразия данных - построение **алгоритма снижения размерности (Grassmann&Stiefel Eigenmaps)**, обеспечивающий близости между

- $M \approx \widehat{M}$ - между точками $X \in \mathbf{M}$ и $\widehat{X} \in \widehat{M}$ исходного и восстановленного по данным многообразий M и $\widehat{M}$
- $T_M(X) \approx T_{\widehat{M}}(\widehat{X})$ - между касательными пространствами $T_M(X)$ и $T_{\widehat{M}}(\widehat{X})$ к многообразиям M и $\widehat{M}$

$TB(\mathbf{M}) = \{(X, T_M(X), X \in \mathbf{M}\}$ - касательное расслоение многообразия данных **M**

Bernstein, Kuleshov. **Tangent Bundle Manifold Learning via Grassmann & Stiefel Eigenmaps, 2012**

Моделирование многообразий: геометрия и стохастика

$$\mathbf{X}_n = \{X_1, X_2, \dots, X_n\} \subset R^p \to \mathbf{Y}_n = \{y_1, y_2, \dots, y_n\} \subset R^q$$

- $\mathbf{X}_n \to \Gamma(\mathbf{X}_n)$ – граф с вершинами в точках выборки, ребра связывают каждую вершину с k ближайшими соседями, веса ребер определяются нормированными ядрами «теплопроводности»:

  $(i, j) \to P_{ij} = \text{const} \times \exp\left\{-\frac{1}{t}\left|X_i - X_j\right|^2\right\} \to$ вероятности появления ребер ($\sum_{i,j} P_{ij} = 1$, $t = 2\sigma^2$)
- $\mathbf{Y}_n \to \Gamma(\mathbf{Y}_n)$ – граф строится подобным образом по «искомым» низкоразмерным представлениям с вероятностями появления ребер $\{Q_{ij}\}$

- **Stochastic Neighbor Embedding (SNE):** G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In: Advances in Neural Information Processing Systems, vol. 15, pp. 833–840, Cambridge, MA, USA, 2002
- **Symmetric SNE (SSNE):** J.A. Cook, I. Sutskever, A. Mnih, and G.E. Hinton. Visualizing similarity data with a mixture of maps. In: Proc. of the 11th International Conference on Artificial Intelligence and Statistics, vol. 2, pp. 67–74, 2007

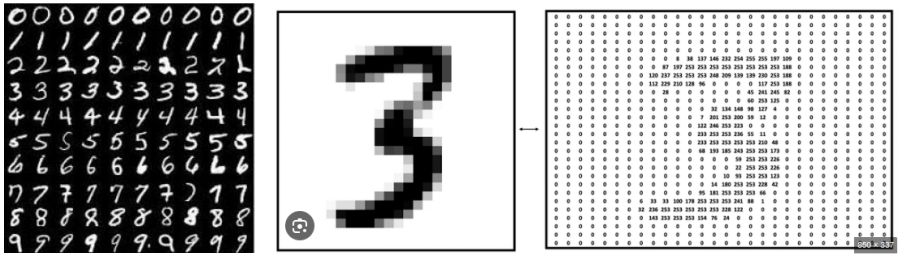
  Минимизация расстояния Кульбака-Лейблера: $L_{SSNE}(\mathbf{Y}_n|\mathbf{X}_n) = KL(P \| Q) = \sum_i \sum_j P_{ij} \log_2 \frac{P_{ij}}{Q_{ij}} \to \min$

- **t-distributed SNE (t-SNE):** L.J.P. van der Maaten, G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 9 (Nov), pp. 2579-2605, 2008

  $L_{t\text{-}SNE}(\mathbf{Y}_n|\mathbf{X}_n) = KL(P \| Q)$, распределение Q имеет t-распределение Стьюдента
- **Uniform Manifold Approximation and Projection (UMAP):** L. McInnes, et. al. UMAP: Uniform Manifold Approximation and Projection. The Journal of Open Source Software, 3(29), p. 861, 2018

Моделирование многообразий: геометрия и стохастика

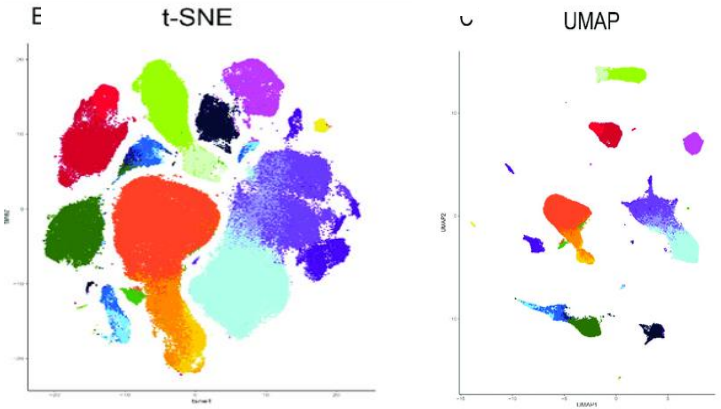
MNIST



28x28 = 756 пикселей

Рукописная цифра – 756-мерный вектор

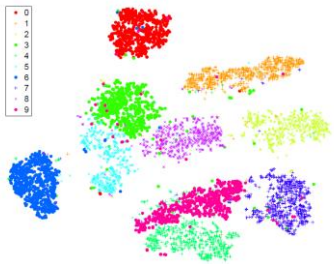
Построение **двумерных представлений** **756-мерных векторов** по 6 000 цифрам



ISOMAP

Laplacian Eigenmaps

t-SNE



Skoltech
Skolkovo Institute of Science and Technology

Моделирование многообразий: исследования

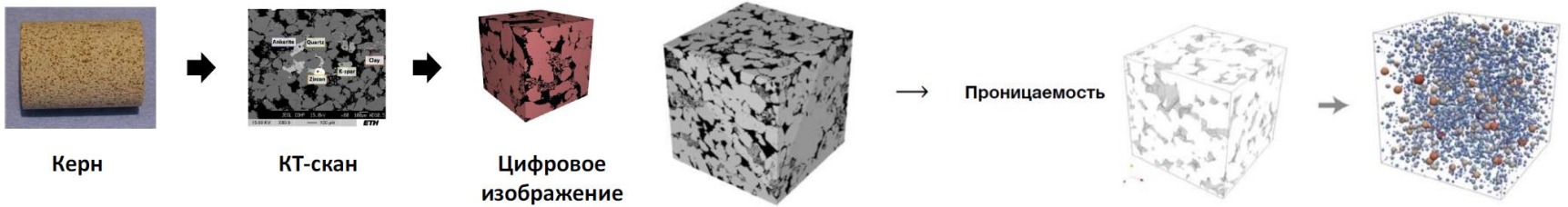
- Bernstein A.V., Kuleshov A.P. **Low-Dimensional Data Representation in Data Analysis**. **Lecture Notes in Artificial Intelligence**, vol. 8774 “**Artificial Neural Networks in Pattern Recognition**”, Springer International Publishing, Switzerland, pp. 47-58, 2014
- A.P. Kuleshov, A.V. Bernstein. **Statistical Learning on Manifold-valued Data**. **Lecture Notes in Artificial Intelligence** Series, vol. 9729 ‘**Machine Learning and Data Mining in Pattern Recognition**,’ Switzerland, Springer International Publishing, pp. 311–325, 2016.
- Kuleshov, A., Bernstein, A. **Nonlinear multi-output regression on unknown input manifold**. **Annals of Mathematics and Artificial Intelligence**, vol. 81, №1-2, pp. 209-240, 2017
- Kuleshov, A., Bernstein, A., Burnaev, E.,Yanovich, Yu. **Machine Learning in Appearance-based Robot Self-localization**. Proceedings of the 16th International **IEEE Conference on Machine Learning and Applications** (ICMLA-2017), IEEE Computer Society, Mexico, December 18-21, pp. 106-112, 2017

Skoltech
Skolkovo Institute of Science and Technology

Топологический Анализ Данных: мотивация

- в задачах ИИ (предиктивной аналитики, машинного обучения, суррогатного моделирования, …)
  - обучающая информация представлена в виде наборов **дискретных данных/облаков точек**
  - информация в компьютере представлена **в дискретной форме**
- **топология** - часть математики/геометрии, изучающая в самом общем виде **явление непрерывности**, а также свойства обобщённых геометрических объектов, не меняющиеся при малых деформациях и не зависящие от способа их задания

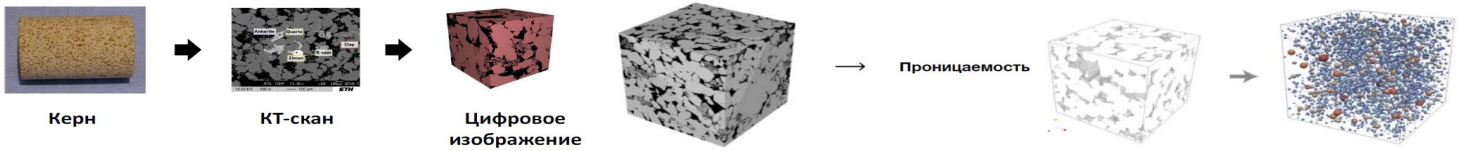
Предсказания проницаемости пористых и гранулярных сред



Skoltech
Skolkovo Institute of Science and Technology

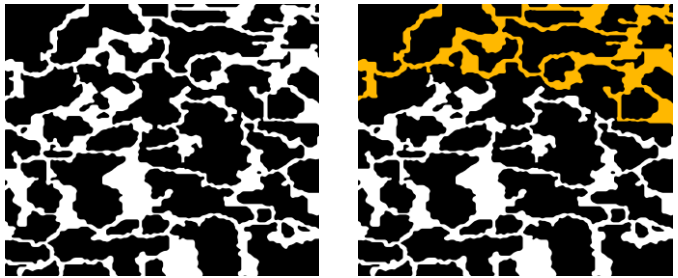
Топологический Анализ Данных: примеры

Предсказания проницаемости пористых и гранулярных сред



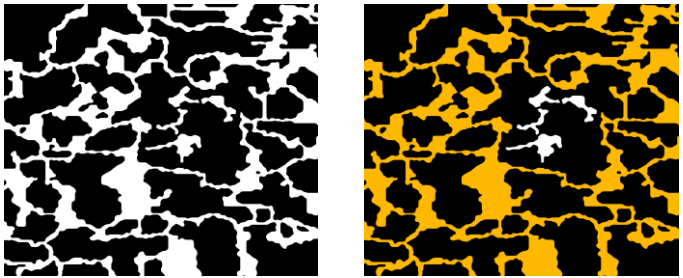
- Геометрические характеристики (пористость, извилистость, …)
- Функционалы Минковского (площади, периметры, эйлеровы характеристики, …)

Образец А



Нулевая проницаемость (в вертикальном направлении)

Образец Б



Ненулевая проницаемость

Значения функционалов Минковского

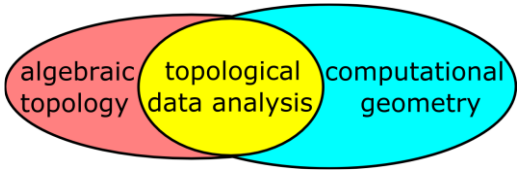
	Образец А	Образец Б	Разница, %
Площадь А	1593603	1593394	0.01
Периметр S	58087	58112	0.04
Эйлерова характеристика χ	-32	-32	–

Skoltech
Skolkovo Institute of Science and Technology

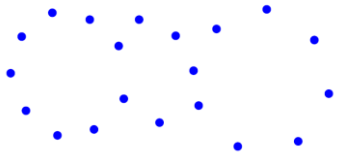
Топологический Анализ Данных: задачи

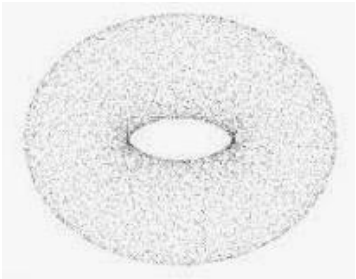
Топологический анализ данных: объединение дискретных фрагментов в непрерывные образы (глобальные структуры) и **вычисление** их различных **характеристик**

- замена набора элементов данных некоторым семейством **симплициальных комплексов** в соответствии с параметром близости.
- анализ этих топологических комплексов с помощью алгебраической топологии (новой теорией **персистентных гомологий)**
- перекодировка устойчивой гомологии набора данных в параметризованную версию **чисел Бетти** (баркоды, персистентные диаграммы, …)



Облако точек (пиксели/воксели изображения)
критерий близости





- выявление **скрытых структур**
- построение новых **признаков** (дескрипторов), называемых **топологическими инвариантами**, …

Skoltech
Skolkovo Institute of Science and Technology

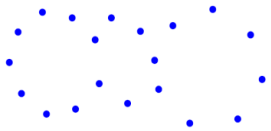
Топологический Анализ Данных: пример (фильтрация)

Числа Бетти:

- β_0 – число связных компонент (0-мерные гомологии)
- β_1 – число «щелей» (1-мерные гомологии)

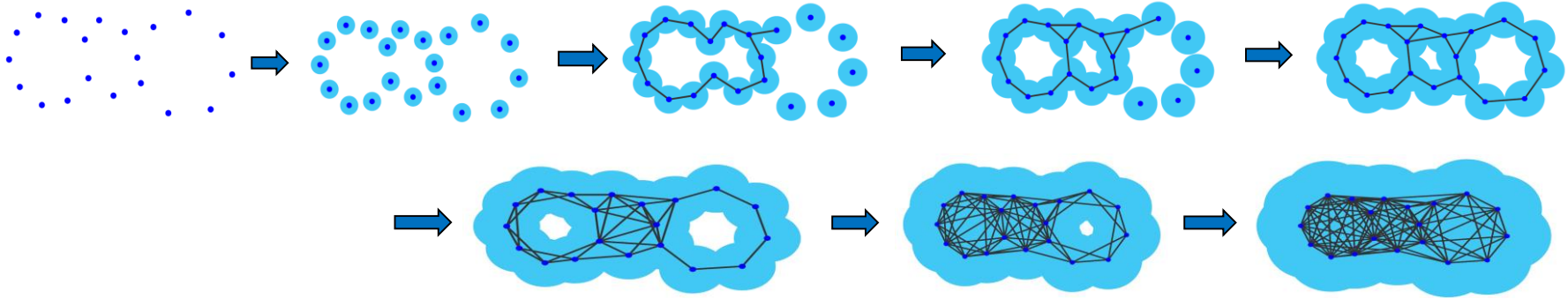
$\beta_0 = 1, \beta_1 = 2$

$\beta_0 = 2, \beta_1 = 1$



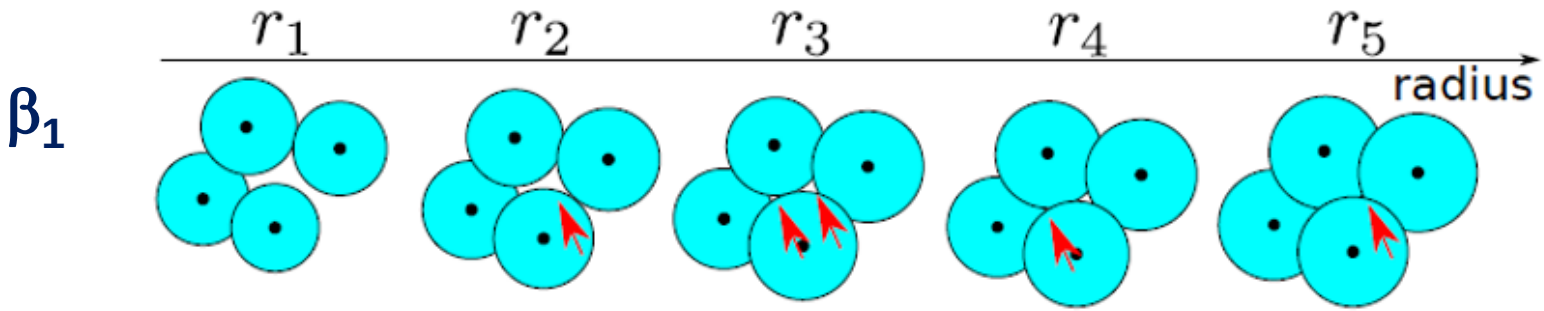
фильтрация: точка → шар радиуса r → увеличивая r, «следим» за:

- количеством связных компонент (β_0)
- количеством щелей (β_1)
- «временами жизни» связных компонент и щелей



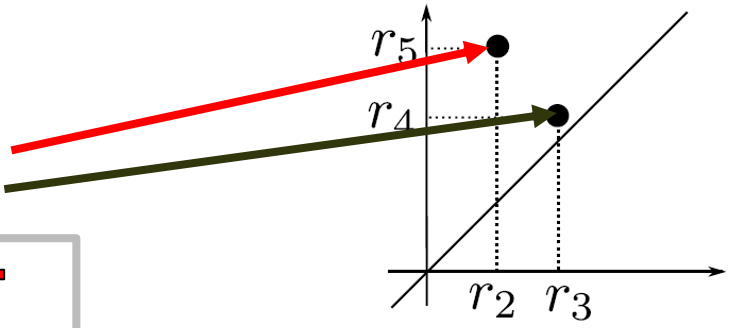
Skoltech
Skolkovo Institute of Science and Technology

Топологический Анализ Данных: пример (процессы «рождения и смерти»)

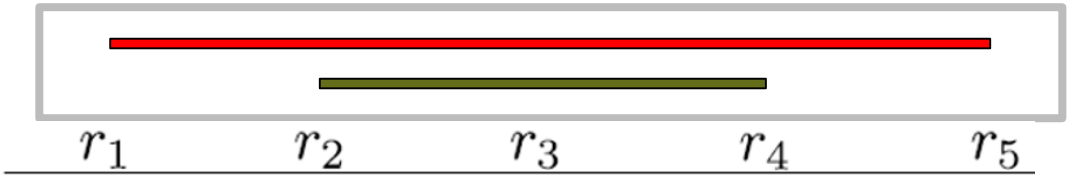


- в момент r_2 **«родилась» Щель 1**
- в момент r_3 **«родилась» Щель 2**
- в момент r_4 **«умерла» Щель 2**
- в момент r_5 **«умерла» Щель 1**

Персистентная диаграмма
(persistence diagram)



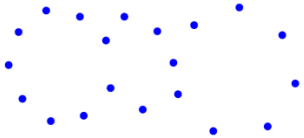
- **Щель 1: родилась в момент** r_2 и умерла **в момент** r_5
- **Щель 2: родилась в момент** r_3 и умерла **в момент** r_4



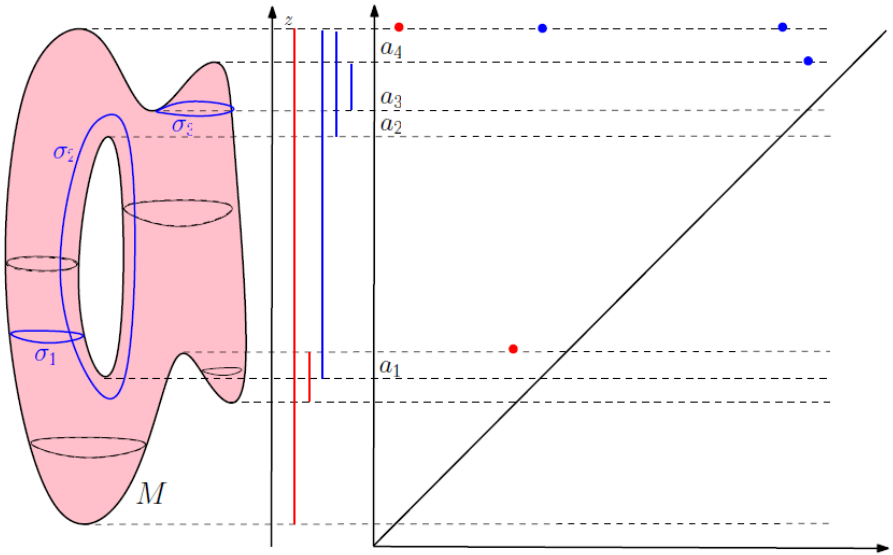
Баркод (barcode)

Skoltech
Skolkovo Institute of Science and Technology

Топологический Анализ Данных: примеры



Фильтрация по «линиям уровня»

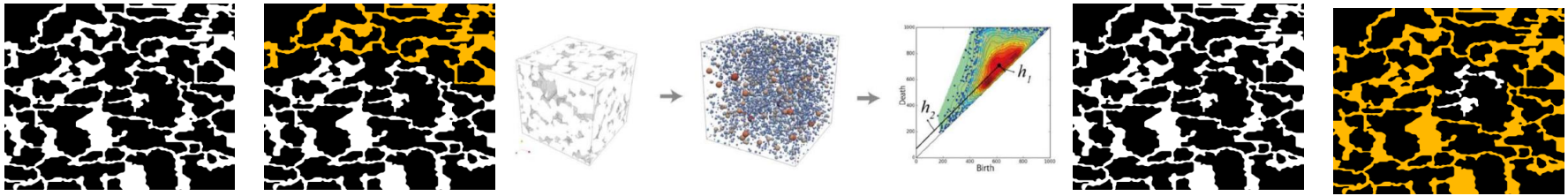


Баркод и персистентная диаграмма:

- **Связные компоненты**
- **цели**

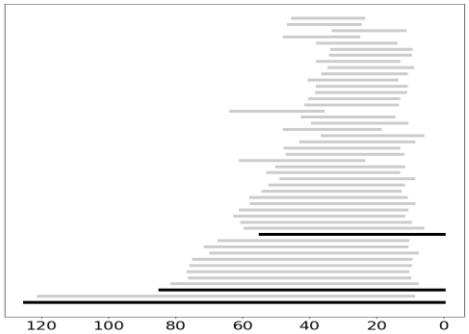
Skoltech
Skolkovo Institute of Science and Technology

ТАД: предсказания проницаемости пористых и гранулярных сред



Образец А (нулевая проницаемость)

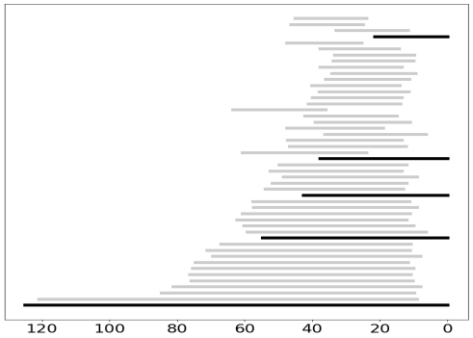
Образец Б (ненулевая проницаемость)



Баркоды и персистентные диаграммы (размерности 0)

в зависимости от характерного размера

- **Черные полосы: топологические характеристики, соответствующие компонентам связности множества пор**
- Для непроницаемой/проницаемой пород персистентные диаграммы существенно различаются



Топологические признаки для предсказания проницаемости пористых и гранулярных сред

Topological Characteristics for prediction of permeability of porous and granular media

Бернштейн А., Бурнаев Е., Осипцов А., Баранников С., Качан О.
Сколковский институт науки и технологий

ГАЗПРОМ НЕФТЬ

5 марта 2020
Санкт-Петербург

Skoltech
Skolkovo Institute of Science and Technology

Сравнение стандартных характеристик (функционалов Минковского) с современными подходами на основе топологического анализа данных

Бернштейн А., Бурнаев Е., Осипцов А., Баранников С., Качан О.
Сколтех

ГАЗПРОМ НЕФТЬ

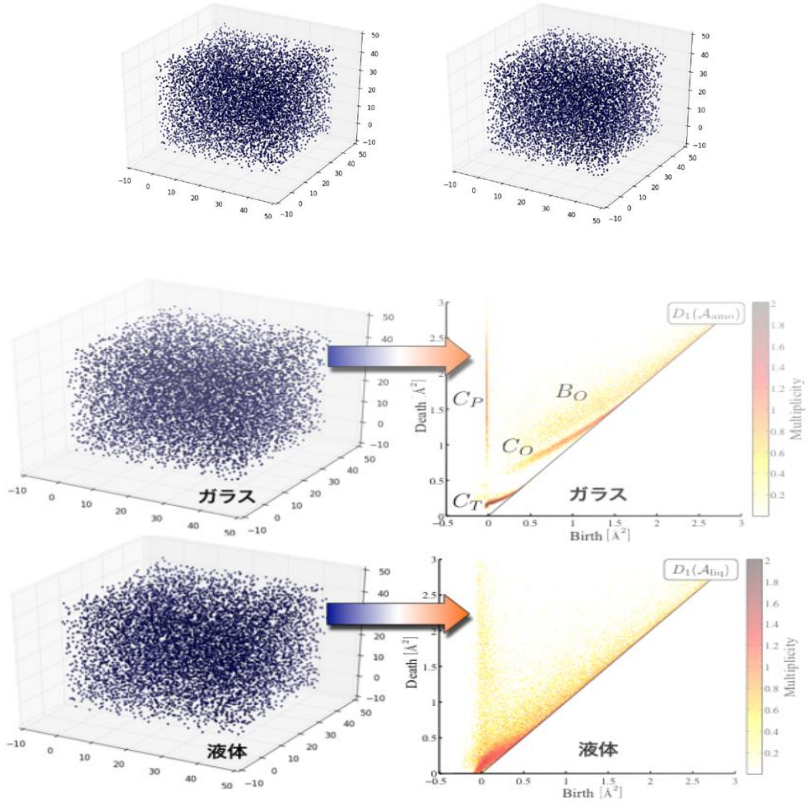
Skoltech
Skolkovo Institute of Science and Technology

Skoltech
Skolkovo Institute of Science and Technology

Топологический Анализ Данных: машинное зрение

Атомные конфигурации различных химических соединений

(аморфный и жидкий кремнеземы)



Topological Data Analysis in Machine Vision

Alexander Bernstein

Joint work with E. Burnaev, M. Sharaev, E. Kondrateva, O. Kachan

Skolkovo Institute of Science and Technology, Moscow, Russia

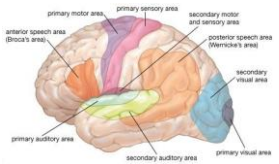
SPIE. DIGITAL LIBRARY

Amsterdam, The Netherlands, November 16, 2019

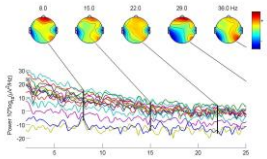
Skoltech
Skolkovo Institute of Science and Technology

Топологический Анализ Данных: медицинские приложения

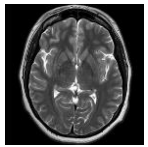
Нейрональные зоны мозга

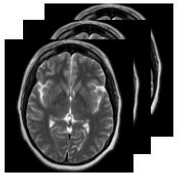


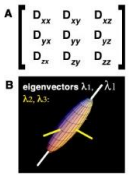
ЭЭГ

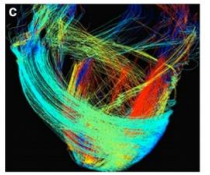


МРТ/фМРТ/диффузионно-тензорная МРТ

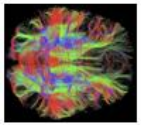




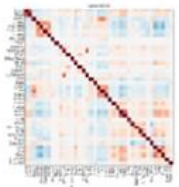




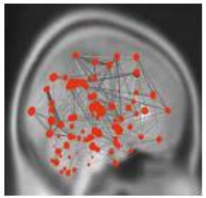
матрицы связей (коннективности)



Матрица коннективности



Граф коннективности



Здоровый

Больной

Skoltech
Skolkovo Institute of Science and Technology

Топологический Анализ Данных: медицинские приложения

Skoltech
Skolkovo Institute of Science and Technology

Диагностика депрессии по ЭЭГ в состоянии покоя
(19 каналов) (Resting-state)

Точность:

по 8 стандартным характеристикам	**0,79**
+ 2 топологических характеристики (Entropy Complex on Envelopes, 0,64)	**0.85**
+ еще 2 топологических характеристики (Persistent Diagram Bins, 0.63)	**0.88**

Skoltech
Skolkovo Institute of Science and Technology

СПАСИБО ЗА ВНИМАНИЕ