Прикладная статистика

и анализ данных

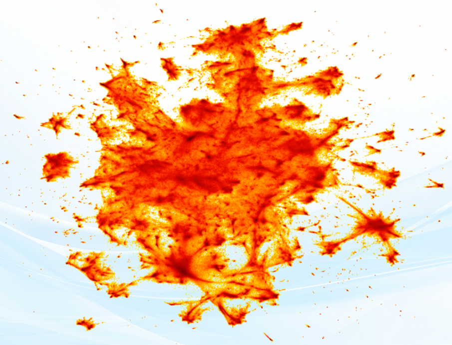
Съезд V

Метод UMAP (2018)

Uniform Approximation and Projection —
метод, выполняющий нелинейное снижение размерности.

Создан Лилендом Макиннесом + его коллегами из Таттского инст.

Цель: было получить метод, похожий на t-SNE,
но с более сильным математическим обоснованием.



Визуализация 3 млн. слов из GoogleNews dataset

Метод UMAP (2018)

Пусть $X = (X_1, ..., X_n)$ — выборка в пространстве $\mathscr{X}$.

$\rho(x_1, x_2)$ — метрика в $\mathscr{X}$

Определим **случайный ориентир. граф** на мн-ве вершин X.

Считаем, что каждое ребро появляются независимо от других.

Рассмотрим вершину X_i.
Пусть $X_{i_1}, ..., X_{i_k}$ — k ближайших соседей объекта X_i в выборке X.
$r_i = \min_s \rho(X_i, X_{i_s})$ — расстояние до ближайшего соседа.

Вероятность ребра из X_i в X_{i_s} [для остальных вер-ть = 0]:

$$\mathsf{P}(X_i \to X_{i_s}) = \exp\left(-\frac{\rho(X_i, X_{i_s}) - r_i}{\sigma_i}\right) \in [0, 1],$$

где σ_i подбирается как решение уравнения $\sum\limits_{s=1}^{k} \mathsf{P}(X_i \to X_{i_s}) = \log_2 k$,
т.е. σ_i играет роль нормировки вероятностей ребер.

Метод UMAP (2018)

На основе ориентированного графа построим **неориентированный**.

X — множество вершин

Вероятность ребра между X_i и X_j:

$$\mathsf{P}(X_i - X_j) = \mathsf{P}(\{X_i \to X_j\} \cup \{X_j \to X_i\}) =$$

$$= \mathsf{P}(X_i \to X_j) + \mathsf{P}(X_j \to X_i) - \mathsf{P}(X_i \to X_j)\, \mathsf{P}(X_j \to X_i).$$

Пусть $Y = (Y_1, ..., Y_n)$ — **вложение** X в маломерное пр-во $\mathbb{R}^d$.
Величины $Y_1, ..., Y_n$ неизвестны, их нужно подобрать.

На Y зададим случайный неор. граф с вероятностями
$\mathsf{P}(Y_i - Y_j) = \left(1 + a\|Y_i - Y_j\|_2^{2b}\right)^{-1}$, где a и b — гиперпараметры.

Минимизируем дивергенцию Кульбака-Лейблера: $KL(\mathsf{P}_X, \mathsf{P}_Y) =$

$$= \sum_{i,j} \left[\mathsf{P}(X_i - X_j) \log \frac{\mathsf{P}(X_i - X_j)}{\mathsf{P}(Y_i - Y_j)} + (1 - \mathsf{P}(X_i - X_j)) \log \frac{1 - \mathsf{P}(X_i - X_j)}{1 - \mathsf{P}(Y_i - Y_j)}\right]$$

Компоновка графа в пространстве низкой размерности

Для расположения точек в пространстве низкой размерности

используется компоновка графа **в теории**.

- Сила притяжения вдоль ребра из Y_i в Y_j

$$\frac{-2ab\|Y_i - Y_j\|_2^{2(b-1)}}{1 + \|Y_i - Y_j\|_2^2}\,\mathsf{P}(X_i \to X_j)(Y_i - Y_j),$$

  где a и b — гиперпараметры.

- Сила отталкивания между вершинами Y_i и Y_j

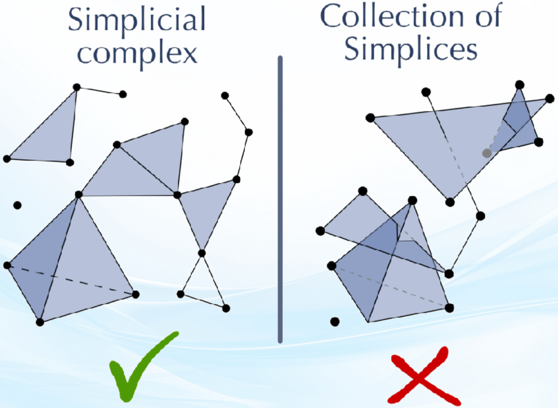
$$\frac{b}{(\varepsilon + \|Y_i - Y_j\|_2^2)(1 + \|Y_i - Y_j\|_2^2)}\,(1 - \mathsf{P}(X_i \to X_j))\,(Y_i - Y_j),$$

  где $\varepsilon = 0.001$.

- Сходимость обеспечивается за счет медленного уменьшения сил притяжения и отталкивания.
- $\mathsf{P}(Y_i - Y_j)$ — аппроксимация этих сил.

Рассмотрим теоретическое обоснование UMAP

Симплициальный комплекс — топологическое пространство с заданной на нем триангуляцией, т.е., склеенное из топологических симплексов по определенным правилам.

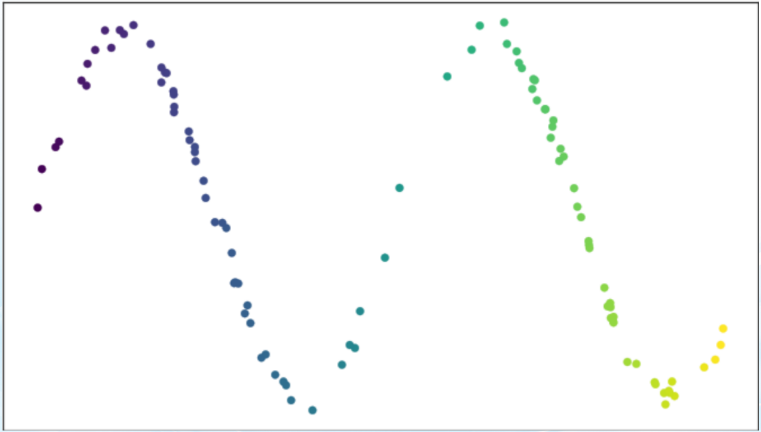


Немного топологии

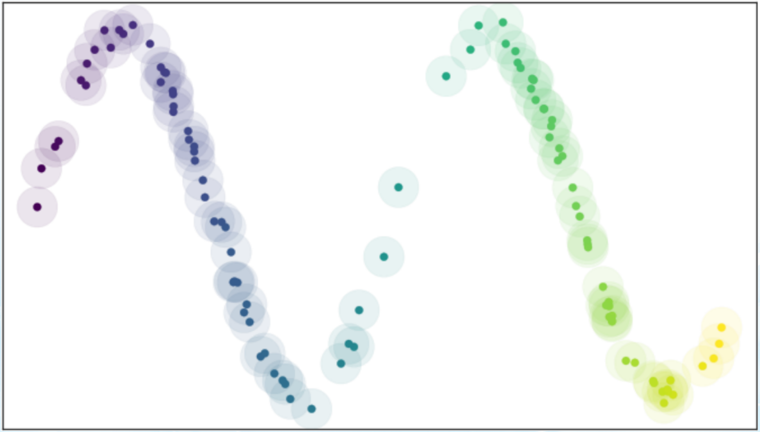
Покрытие топологического пространства — семейство множеств, таких, что их объединение содержит заданное множество.

Пусть $\{W_\alpha\}$ — конечное покрытие топологического пространства X. **Нерв покрытия** $\{W_\alpha\}$ — это абстрактный симплициальный комплекс N, множество вершин которого отождествлено с множеством индексов покрытия, при этом N содержит симплекс с вершинами $\alpha_1, \alpha_2, \ldots, \alpha_n$ тогда и только тогда, когда $\bigcap\limits_{i=1}^{n} W_{\alpha_i} \neq \varnothing$.

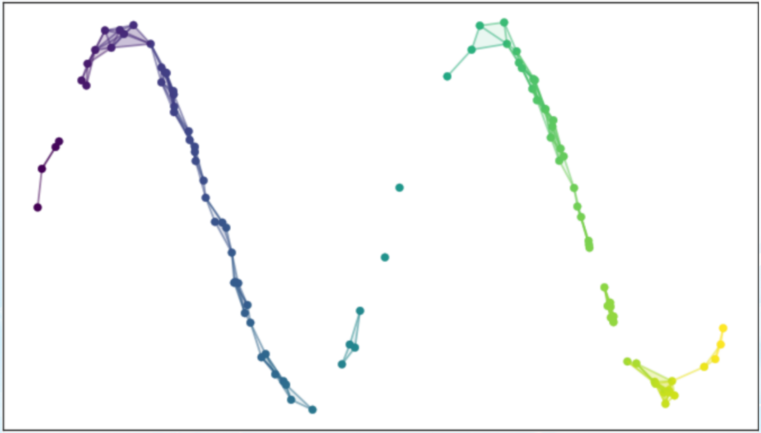
Топологическое множество



Покрытие топологического множества



Нерв покрытия



Теорема о нерве

Если пространство X триангулируемо и $\{W_\alpha\}$ — конечное покрытие замкнутыми множествами, причём все непустые пересечения стягиваемы, то нерв покрытия гомотопически эквивалентен X.

Гомотопия из X в Y — непрерывное отображение $f : [0,1] \times X \to Y$.

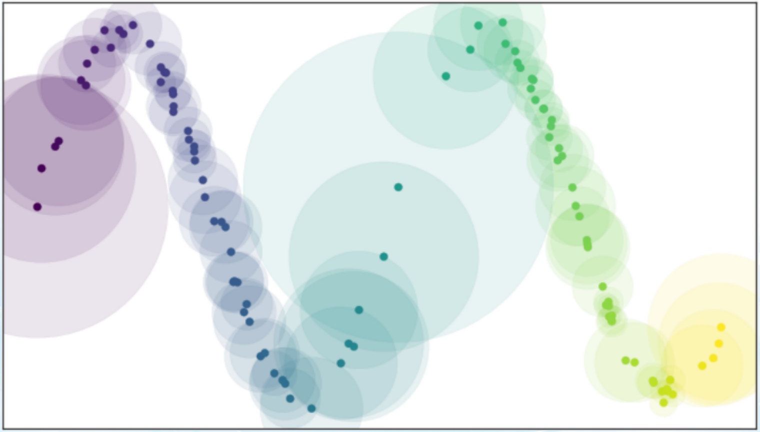
Если данные равномерно распределены по многообразию,
то покрытие будет “хорошим”.

Когда данные так хорошо себя ведут?

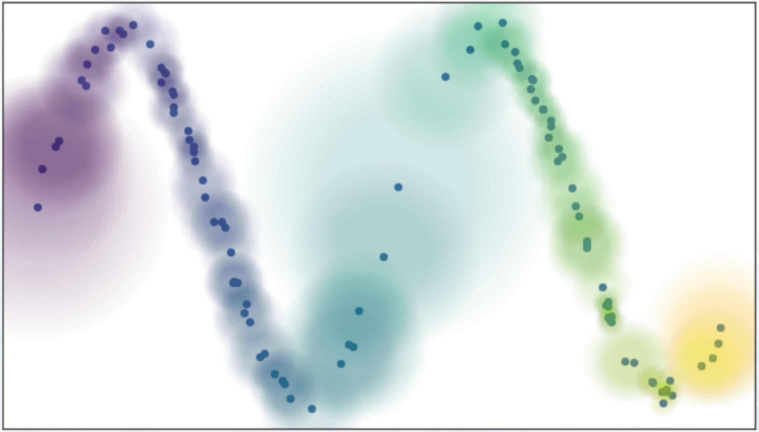
Предположение: данные равномерно распределены на многообразии!

Определим Риманову метрику на многообразии,
чтобы сделать это предположение истинным.

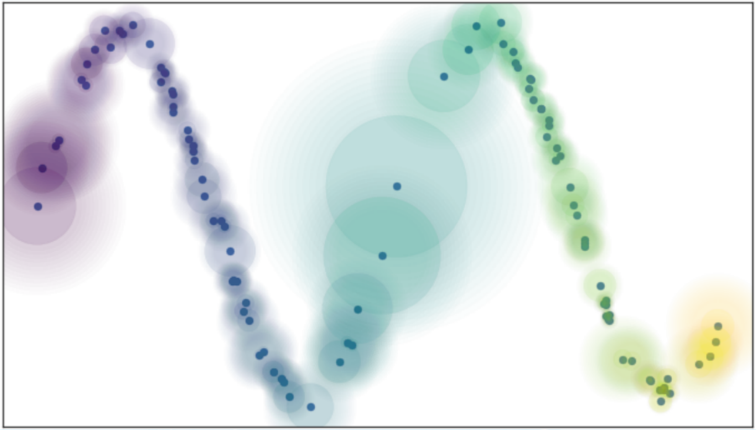
Покрытие с разными радиусами



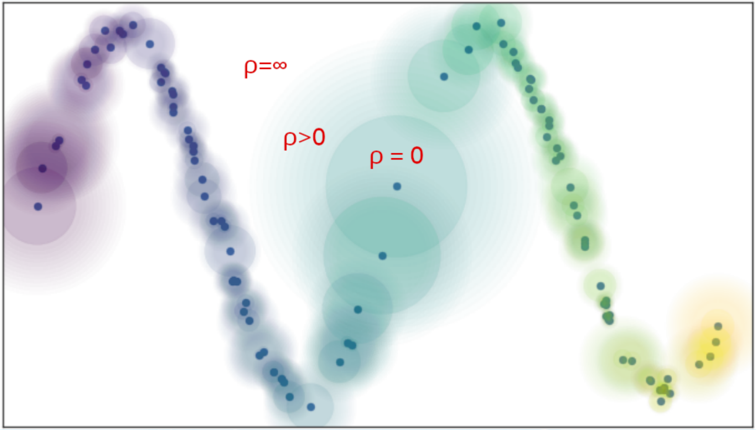
Но почему радиус фиксирован? Берем случайный:



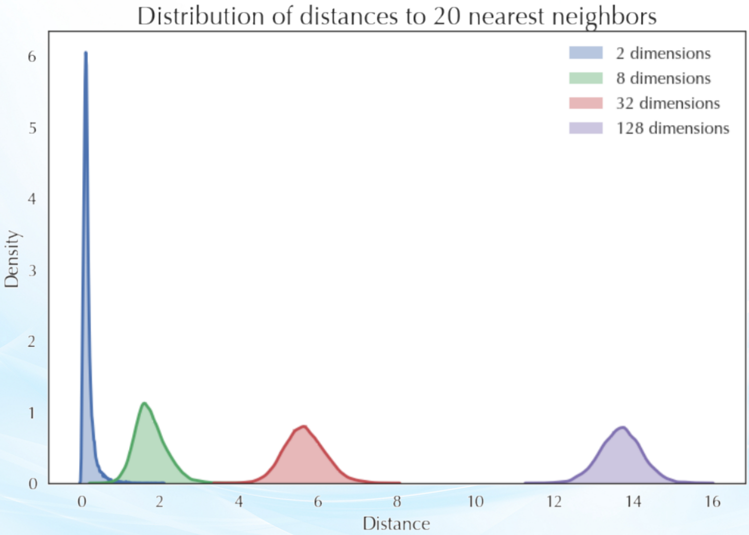
Предположение: многообразие локально связно



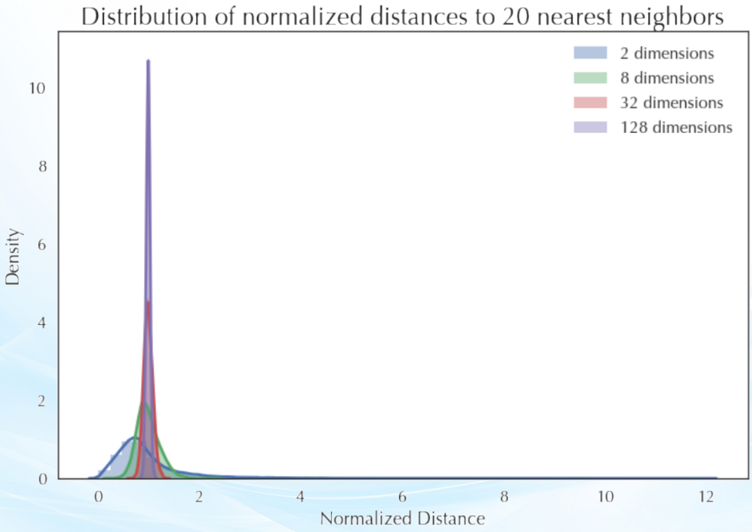
Предположение: многообразие локально связно



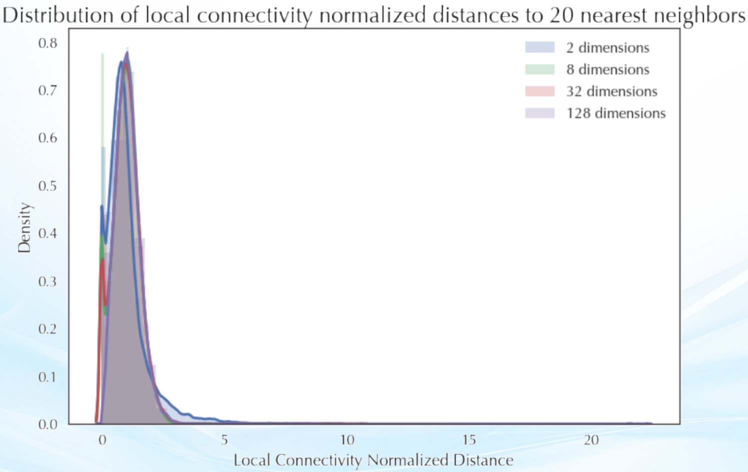
Распределение расстояний



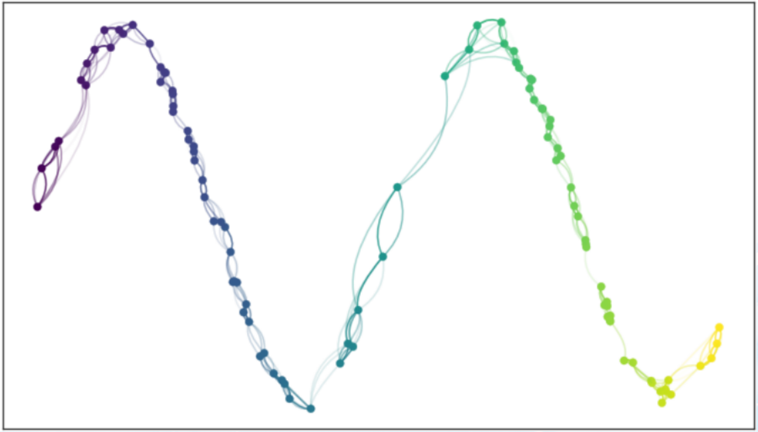
Распределение нормированных расстояний



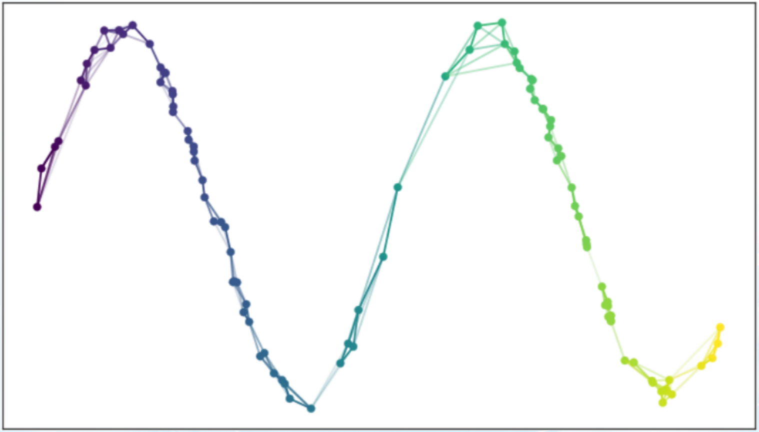
Распр. норм. расстояний на основе лок. связности



Но наша локальная метрика не симметрична!



Сделаем ее симметричной



Далее

Топология, теория представлений, нечеткие множества...

Definition 7. *Define the functor* FinReal : **Fin-sFuzz** $\to$ **FinEPMet** *by setting*

$$\mathsf{FinReal}(\Delta^n_{<a}) \triangleq (\{x_1, x_2, \dots, x_n\}, d_a),$$

where

$$d_a(x_i, x_j) = \begin{cases} -\log(a) & \text{if } i \neq j, \\ 0 & \text{otherwise} \end{cases}.$$

and then defining

$$\mathsf{FinReal}(X) \triangleq \operatorname*{colim}_{\Delta^n_{<a} \to X} \mathsf{FinReal}(\Delta^n_{<a}).$$

Similar to Spivak's construction, the action of FinReal on a map $\Delta^n_{<a} \to \Delta^m_{<b}$, where $a \leq b$ defined by $\sigma : \Delta^n \to \Delta^m$, is given by

$$(\{x_1, x_2, \dots, x_n\}, d_a) \mapsto (\{x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}\}, d_b),$$

which is a non-expansive map since $a \leq b$ implies $d_a \geq d_b$.

Since FinReal preserves colimits it admits a right adjoint, the fuzzy singular set functor FinSing. We can then define the (finite) fuzzy singular set functor in terms of the action of its image on $\mathbf{\Delta} \times I$, analogously to Sing.

Definition 8. *Define the functor* FinSing : **FinEPMet** $\to$ **Fin-sFuzz** *by*

$$\mathsf{FinSing}(Y) : ([n], [0, a)) \mapsto \hom_{\mathbf{FinEPMet}}(\mathsf{FinReal}(\Delta^n_{<a}), Y).$$

We then have the following theorem.

Theorem 1. *The functors* FinReal : **Fin-sFuzz** $\to$ **FinEPMet** *and* FinSing : **FinEPMet** $\to$ **Fin-sFuzz** *form an adjunction with* FinReal *the left adjoint and* FinSing *the right adjoint.*

Proof. The adjunction is evident by construction, but can be made more explicit as follows. Define a functor $F : \mathbf{\Delta} \times I \to \mathbf{FinEPMet}$ by

$$F([n], [0, a)) = (\{x_1, x_2, \dots, x_n\}, d_a),$$

where

$$d_a(x_i, x_j) = \begin{cases} -\log(a) & \text{if } i \neq j, \\ 0 & \text{otherwise} \end{cases}.$$

Now FinSing can be defined in terms of F as

$$\mathsf{FinSing}(Y) : ([n], [0, a)) \mapsto \hom_{\mathbf{FinEPMet}}(F([n], [0, a)), Y).$$

where the face maps d_i are given by pre-composition with Fd^i, and similarly for degeneracy maps, at any given value of a. Furthermore post-composition with F level-wise for each a defines maps of fuzzy simplicial sets making FinSing a functor.

We now construct FinReal as the left Kan extension of F along the Yoneda embedding:

$$\begin{tikzcd} & \mathbf{Fin\text{-}sFuzz} \arrow[dr, dashed, "\mathsf{FinReal}"] & \\ \mathbf{\Delta} \times I \arrow[ur, hook, "y"] \arrow[rr, "F"'] & & \mathbf{FinEPMet} \end{tikzcd}$$

Explicitly this results in a definition of FinReal at a fuzzy simplicial set X as a colimit:

$$\mathsf{FinReal}(X) = \operatorname*{colim}_{y([n],[0,a)) \to X} F([n]).$$

Further, it follows from the Yoneda lemma that $\mathsf{FinReal}(\Delta^n_{<a}) \cong F([n], [0, a))$, and hence this definition as a left Kan extension agrees with Definition 7, and the definition of FinSing above agrees with that of Definition 8. To see that FinReal and FinSing are adjoint we note that


ВСЁ!