Topological methods for the analysis of high dimensional data sets and 3d object recognition.

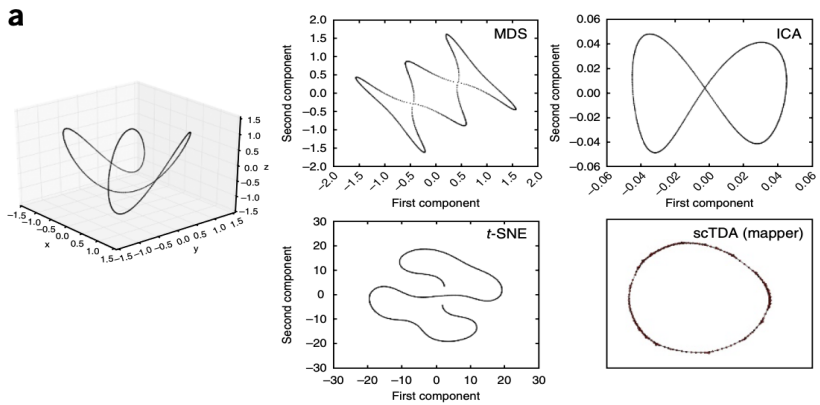
Singh, Gurjeet and Mémoli, Facundo and Carlsson, Gunnar E
Department of Mathematics, Stanford University
Institute for Computational and Mathematical Engineering, Stanford University

Journal: PBG@ Eurographics, 2007

Introduction

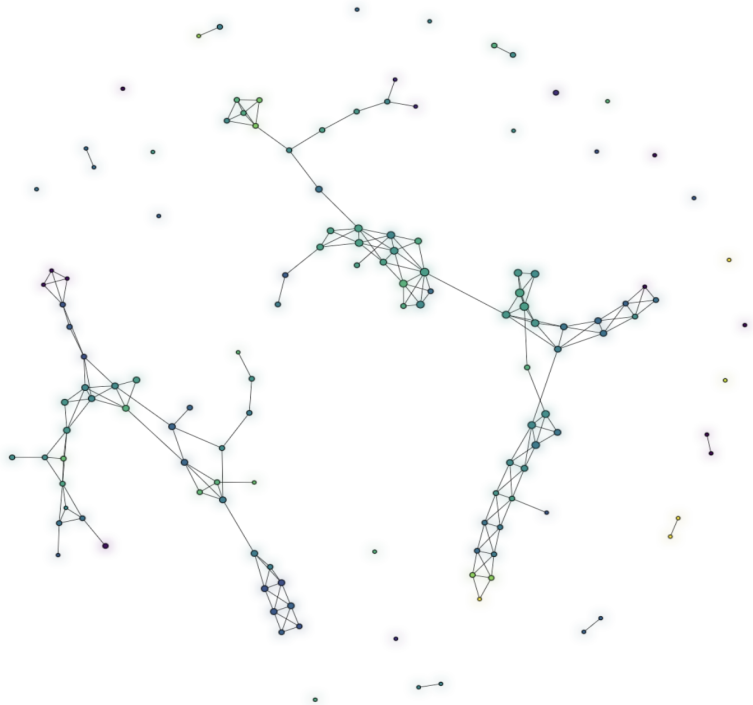
High Dimensional Data Sets

- Data coming from real applications is massive, such as big data
- Impossible to visualize and discern structure even in low dimensional projection
  - Different projection methods: PCA, tSNE, MDS, and etc.
  - Problem: Which method should we choose?



Aim

To introduce a new algorithm "Mapper" to visualize high-dimensional data sets in low-dimensional image which is easier to understand, and point to areas of interest.



Topological Background and Motivation

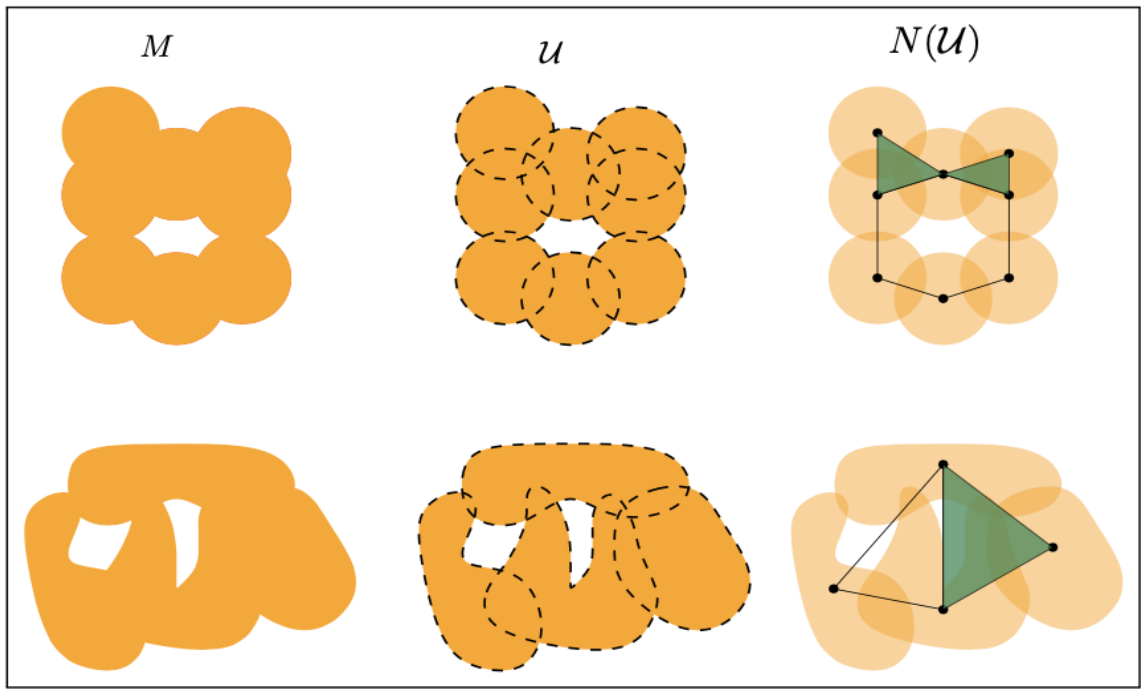
Definition

Nerve: Given a finite collection of sets $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, the *nerve* of $\mathcal{U}$ is defined to be the simplicial complex N($\mathcal{U}$) whose vertex is the index set A, and where a subset $\{\alpha_0, \alpha_1, ...\alpha_k\}$ of A spans a k-simplex in N($\mathcal{U}$) if and only if $U_{\alpha_0} \cap U_{\alpha_1} \cap ... \cap U_{\alpha_k} \neq \phi$

Topological Background and Motivation

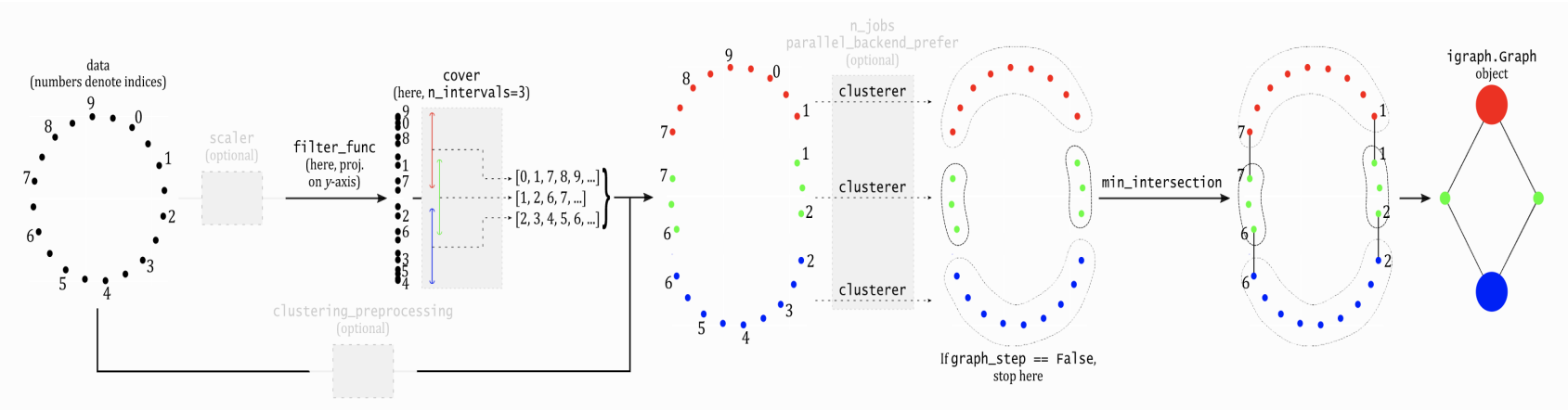
Definition

Nerve: Given a finite collection of sets $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, the *nerve* of $\mathcal{U}$ is defined to be the simplicial complex N($\mathcal{U}$) whose vertex is the index set A, and where a subset $\{\alpha_0, \alpha_1, ...\alpha_k\}$ of A spans a k-simplex in N($\mathcal{U}$) if and only if $U_{\alpha_0} \cap U_{\alpha_1} \cap ... \cap U_{\alpha_k} \neq \phi$



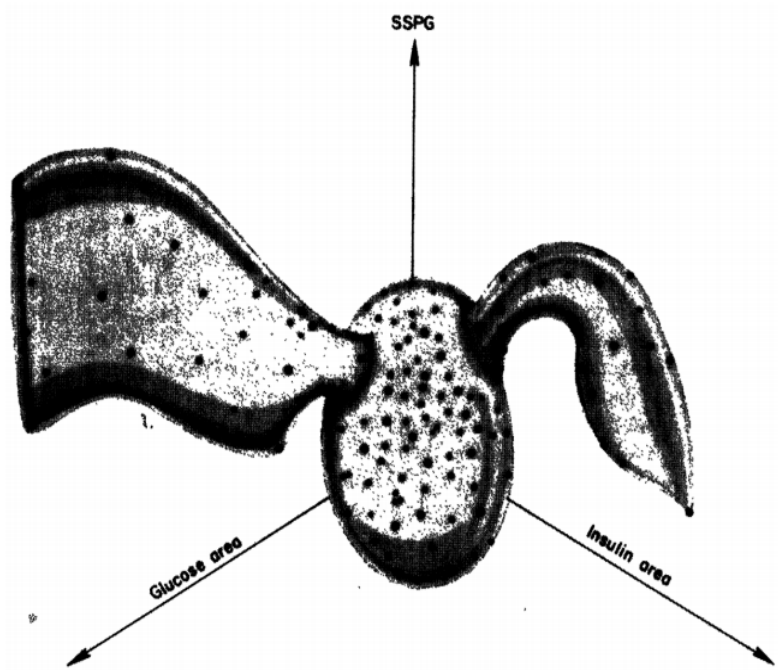
Topological Background and Motivation

1. If we can not imagine what our data looks like, how is it possible to select a cover $\mathcal{U}$?
2. Solution: We can project the data into a space which is easier for us to imagine, such as the Euclidean space or sphere S_1.



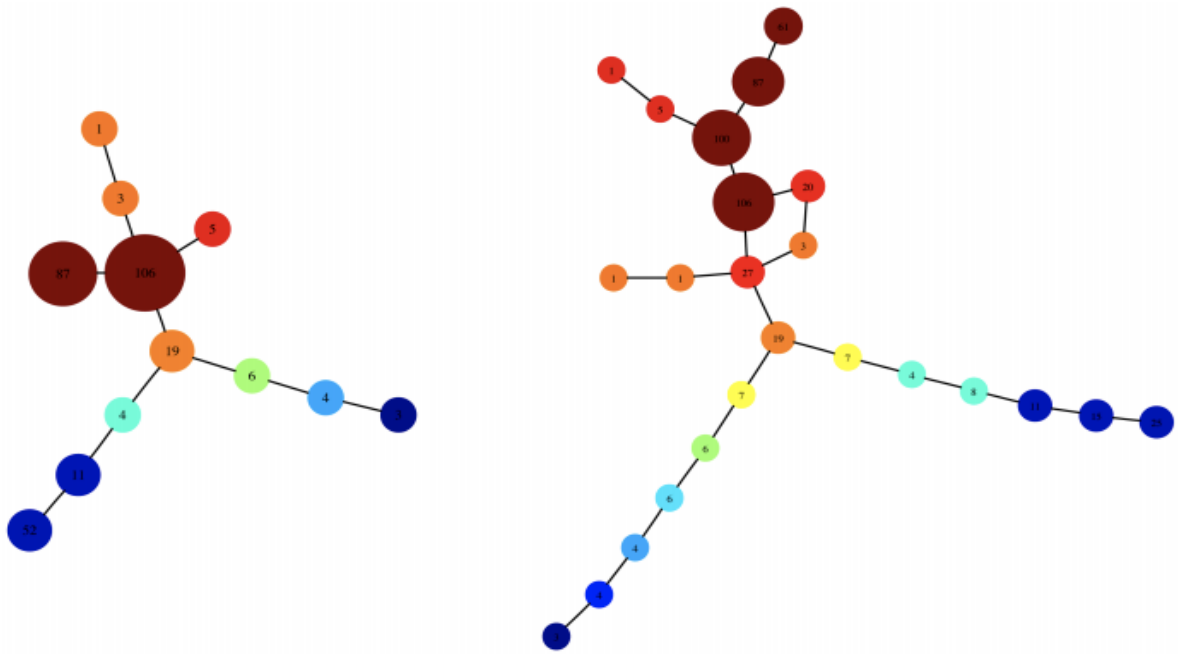
Applications

- The Miller-Reaven diabetes study
- six dimensional data:
  1. age,
  2. relative weight,
  3. fasting plasma glucose,
  4. area under the plasma glucose curve for the three hour glucose tolerance test (OGTT),
  5. area under the plasma insulin curve for the (OGTT)



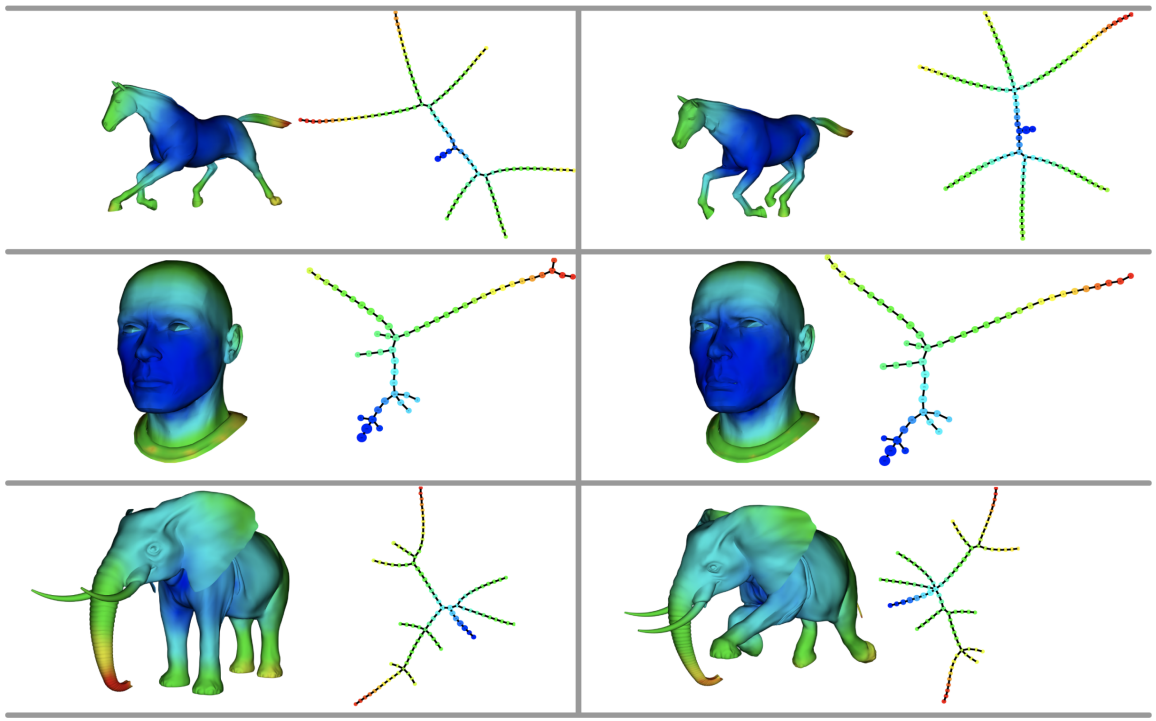
Applications

- The Miller-Reaven diabetes study
- six dimensional data:
  1. age,
  2. relative weight,
  3. fasting plasma glucose,
  4. area under the plasma glucose curve for the three hour glucose tolerance test (OGTT),
  5. area under the plasma insulin curve for the (OGTT)



Applications

- For AI



Reference

- Computational Topology for Data Analysis
- https://docs-tda.giotto.ai/0.2.0/modules/mapper.html
- van Veen et al., (2019). Kepler Mapper: A flexible Python implementation of the Mapper algorithm. Journal of Open Source Software, 4(42), 1315
- Hendrik Jacob van Veen, Nathaniel Saul, David Eargle, and Sam W. Mangham. (2019, October 14). Kepler Mapper: A flexible Python implementation of the Mapper algorithm (Version 1.4.1).
- Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development