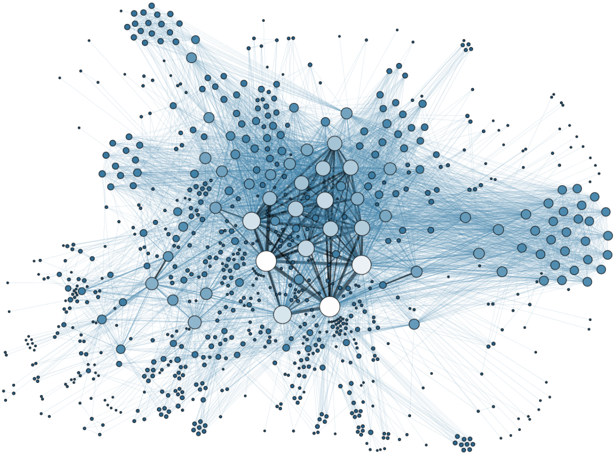
A Short Introduction to TDA (Topological Data Analysis)

Sara Kališnik

Introduction

Data is often given in the form of point clouds in $\mathbb{R}^n$.



Introduction

We have problems analyzing this data because it is often

Introduction

We have problems analyzing this data because it is often

- given in the form of very long vectors, where not all coordinates are relevant,

Introduction

We have problems analyzing this data because it is often

- given in the form of very long vectors, where not all coordinates are relevant,
- very high-dimensional,

Introduction

We have problems analyzing this data because it is often

- given in the form of very long vectors, where not all coordinates are relevant,
- very high-dimensional,
- noisy.

Introduction

We have problems analyzing this data because it is often

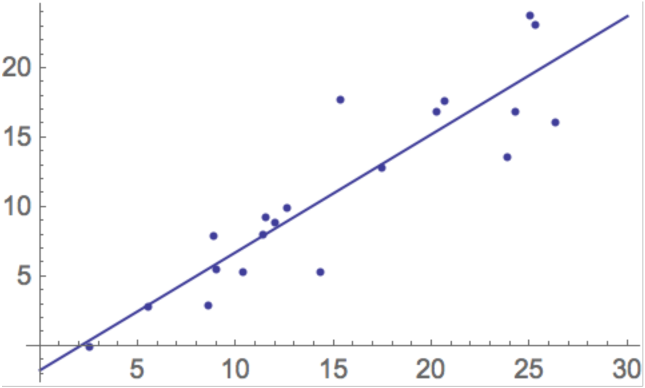
- given in the form of very long vectors, where not all coordinates are relevant,
- very high-dimensional,
- noisy.

Goal of topological data analysis:

Leverage machinery of algebraic topology to develop tools for studying 'qualitative' features of data.

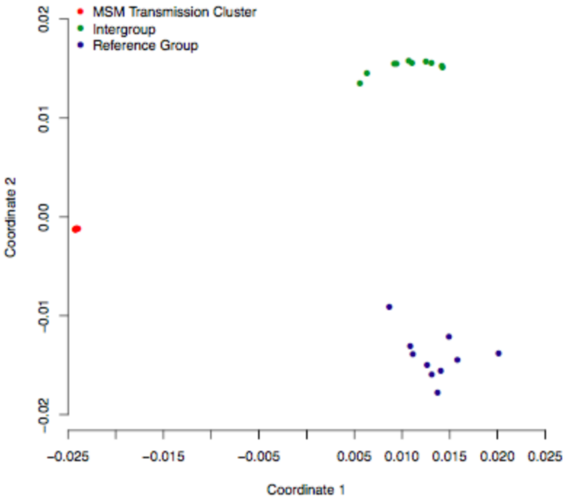
Shape of Data

Linear Regression



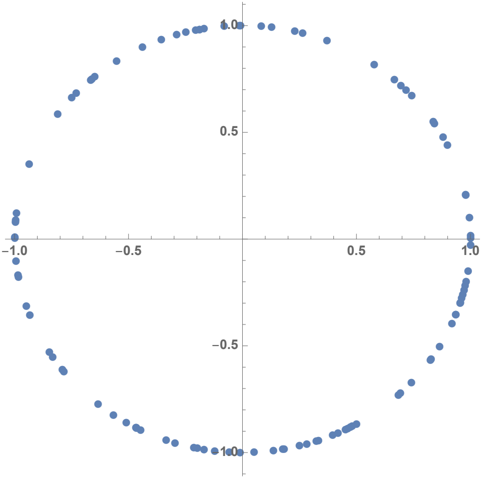
Shape of Data

Clusters



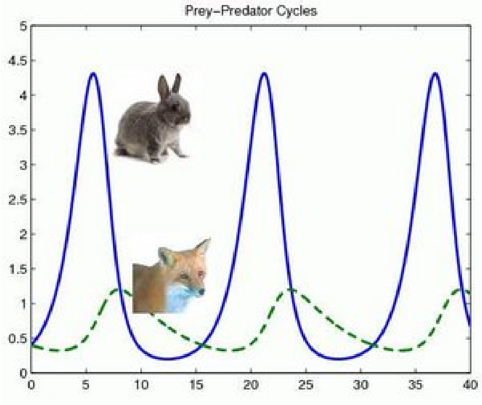
Shape of Data

Loops



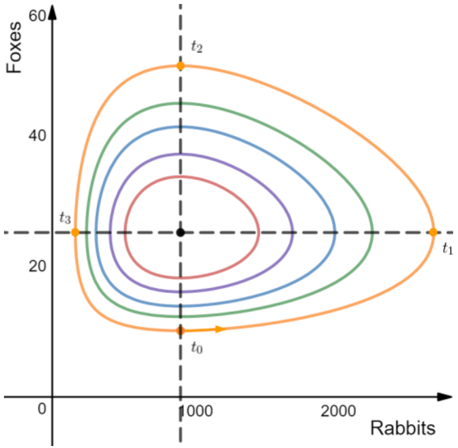
Shape of Data

Holes/Cycles/Loops



Shape of Data

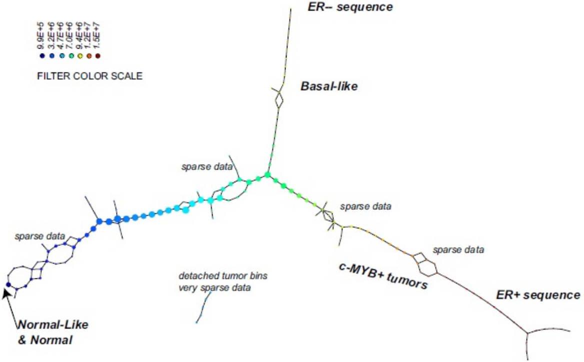
Holes/Cycles/Loops



Shape of Data

Tendrils/Flares

Breast Cancer Study [Nicolau, Levine, Carlsson 2011]



Why Topology?

Three key ideas:

- Invariance under deformation
- Coordinate freeness
- Compressed representations

Why Topology?

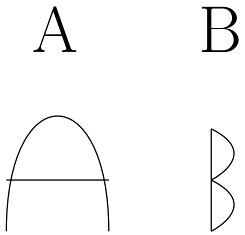
Three key ideas:

- Invariance under deformation

Why Topology?

Three key ideas:

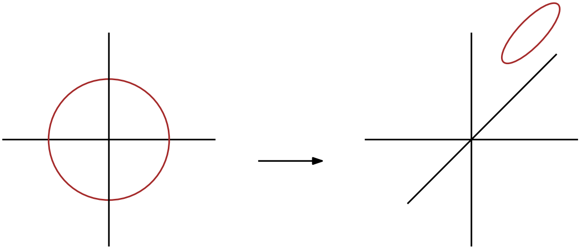
- Invariance under deformation



Why Topology?

Three key ideas:

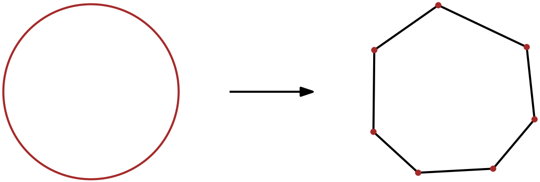
- Coordinate Freeness



Why Topology?

Three key ideas:

- Compressed representations



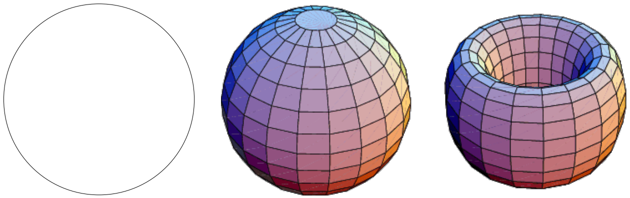
How to deal with shape?

Two tasks:

- Measure Shape
- Represent Shape

Measuring Shape

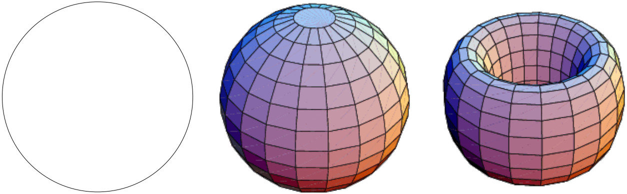
Homology is a formalism for measuring shape...



$b_1 = 1$	$b_1 = 0$	$b_1 = 2$
$b_2 = 0$	$b_2 = 1$	$b_2 = 1$

Measuring Shape

Homology is a formalism for measuring shape...



$b_1 = 1$	$b_1 = 0$	$b_1 = 2$
$b_2 = 0$	$b_2 = 1$	$b_2 = 1$

The extension of homology to more general setting including point clouds is called persistent homology.

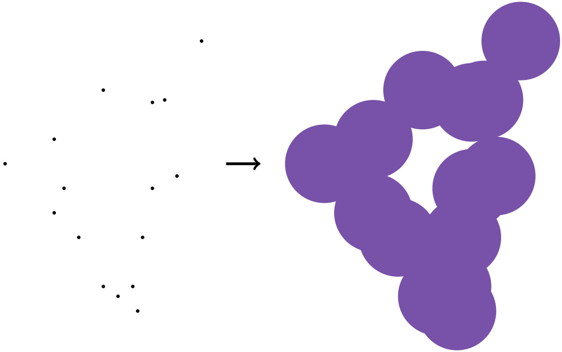
The concept emerged independently in the work of Frosini, Ferri, and collaborators in Bologna, Italy, of Robins at Boulder, Colorado, and of Edelsbrunner, Letscher and Zomorodian at Duke, North Carolina.

Persistent Homology

A finite metric space $\mathbb{X}$ has no interesting topology.

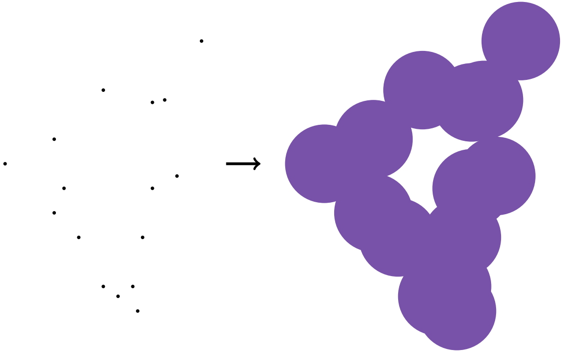
Persistent Homology

A finite metric space $\mathbb{X}$ has no interesting topology.



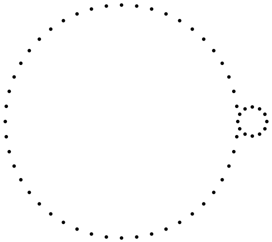
Persistent Homology

A finite metric space $\mathbb{X}$ has no interesting topology.

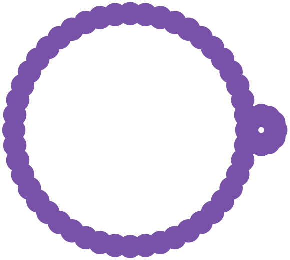


Let $U(\mathbb{X}, R)$ be the union of balls of radius R centered at the points of $\mathbb{X}$. For any $R > 0$ and $i \geq 0$, i-th Betti number of $U(\mathbb{X}, R)$ gives us a qualitative descriptor of $\mathbb{X}$.

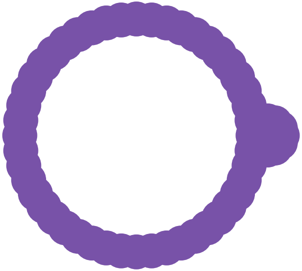
Persistent Homology



Persistent Homology



$b_0 = 1$
$b_1 = 2$



$b_0 = 1$
$b_1 = 1$

Persistent Homology

Problems with this descriptor

- No canonical choice of R.
- Invariant is unstable with respect to perturbation of data or small changes in R.
- Does not distinguish 'small' holes from 'big' ones.

Persistent Homology

Persistent Homology

- Consider not only single reconstruction $U(\mathbb{X}, R)$ of $\mathbb{X}$, but a 1-parameter family of reconstructions

$$F(\mathbb{X}) = \{U(\mathbb{X}, r)\}_{r \in [0,\infty)}$$

and inclusion maps $U(\mathbb{X}, r) \hookrightarrow U(\mathbb{X}, r')$ whenever $r \leq r'$.

Persistent Homology

Persistent Homology

- Consider not only single reconstruction $U(\mathbb{X}, R)$ of $\mathbb{X}$, but a 1-parameter family of reconstructions

$$F(\mathbb{X}) = \{U(\mathbb{X}, r)\}_{r \in [0,\infty)}$$

  and inclusion maps $U(\mathbb{X}, r) \hookrightarrow U(\mathbb{X}, r')$ whenever $r \leq r'$.
- Apply i-dimensional homology functor H_i with field coefficients

Persistent Homology

Persistent Homology

- Consider not only single reconstruction $U(\mathbb{X}, R)$ of $\mathbb{X}$, but a 1-parameter family of reconstructions

$$F(\mathbb{X}) = \{U(\mathbb{X}, r)\}_{r \in [0,\infty)}$$

and inclusion maps $U(\mathbb{X}, r) \hookrightarrow U(\mathbb{X}, r')$ whenever $r \leq r'$.
- Apply i-dimensional homology functor H_i with field coefficients
- Obtain a family of vector spaces $\{V_r\}_r$ and linear maps between them. Call such algebraic structures persistence vector spaces.

Persistent Homology

Persistent Homology

- Consider not only single reconstruction $U(\mathbb{X}, R)$ of $\mathbb{X}$, but a 1-parameter family of reconstructions

$$F(\mathbb{X}) = \{U(\mathbb{X}, r)\}_{r \in [0,\infty)}$$

  and inclusion maps $U(\mathbb{X}, r) \hookrightarrow U(\mathbb{X}, r')$ whenever $r \leq r'$.
- Apply i-dimensional homology functor H_i with field coefficients
- Obtain a family of vector spaces $\{V_r\}_r$ and linear maps between them. Call such algebraic structures persistence vector spaces.

Can we classify persistence vector spaces that arise from filtrations up to isomorphism?

Persistent Homology

Persistent Homology

- Consider not only single reconstruction $U(\mathbb{X}, R)$ of $\mathbb{X}$, but a 1-parameter family of reconstructions

$$F(\mathbb{X}) = \{U(\mathbb{X}, r)\}_{r \in [0,\infty)}$$

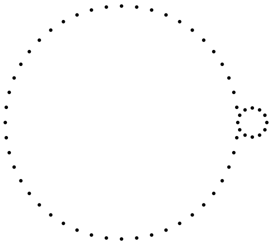
  and inclusion maps $U(\mathbb{X}, r) \hookrightarrow U(\mathbb{X}, r')$ whenever $r \leq r'$.
- Apply i-dimensional homology functor H_i with field coefficients
- Obtain a family of vector spaces $\{V_r\}_r$ and linear maps between them. Call such algebraic structures persistence vector spaces.

Can we classify persistence vector spaces that arise from filtrations up to isomorphism?

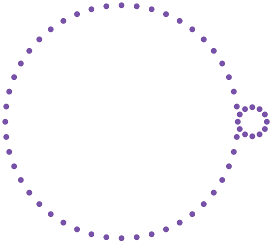
Yes, by barcodes.

(Computing Persistent Homology, Carlsson and Zomorodian)

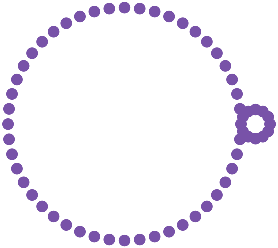
Persistent Homology



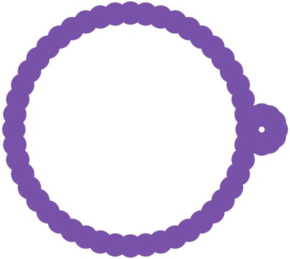
Persistent Homology



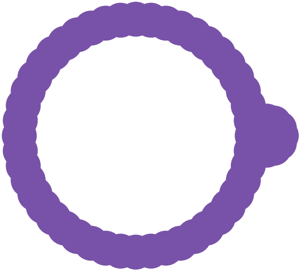
Persistent Homology



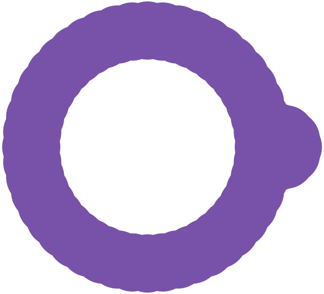
Persistent Homology



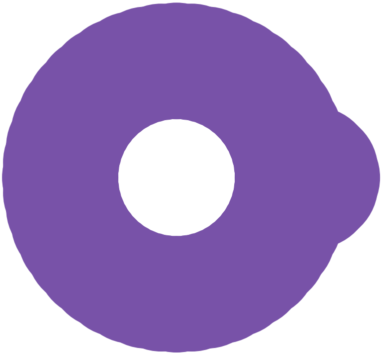
Persistent Homology



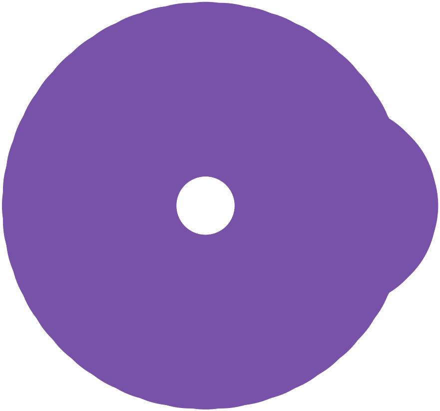
Persistent Homology



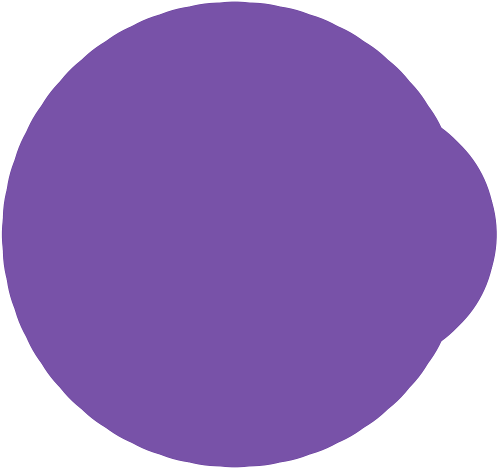
Persistent Homology



Persistent Homology

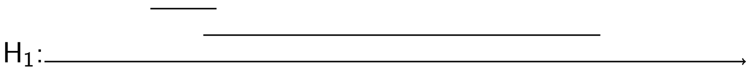


Persistent Homology



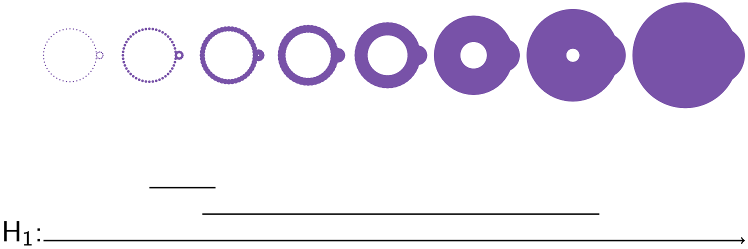
Persistent Homology

Barcode for



Persistent Homology

Barcode for



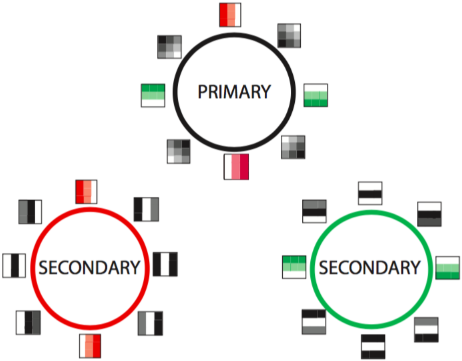
For each interval:

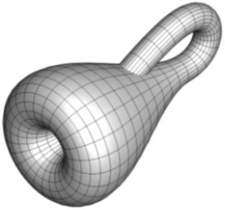
- Left endpoint is the index at which the hole is born
- Right endpoint is index at which hole dies
- Length of interval is the lifetime of a hole in filtration

Applications of Persistent Homology

Natural Scene Statistics/Image Processing

(Local structure of spaces of natural images by G. Carlsson, Vin de Silva, T. Ishkanov and A. Zomorodian)





Natural Scene Statistics/Image Processing

A long time ago in a country far far away (the Netherlands) J. van Hateren and A. van der Schaaf were taking photos in a town called Groningen and in the surrounding countryside.

Natural Scene Statistics/Image Processing

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Natural Scene Statistics/Image Processing

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Typical camera uses tens of thousands of pixels, so images lie in a very high dimensional pixel space, $\mathbb{R}^P$.

Natural Scene Statistics/Image Processing

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Typical camera uses tens of thousands of pixels, so images lie in a very high dimensional pixel space, $\mathbb{R}^P$.

David Mumford: What can be said about the set of images $\mathcal{I} \subseteq \mathcal{P}$ lying within $\mathbb{R}^P$? Can it be modeled as a submanifold or a subspace of $\mathbb{R}^P$?

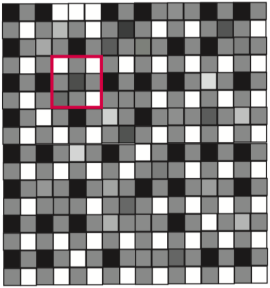
Natural Scene Statistics/Image Processing

The whole manifold of images is not accessible in a useful way, a space of small image patches might in fact contain quite useful information.

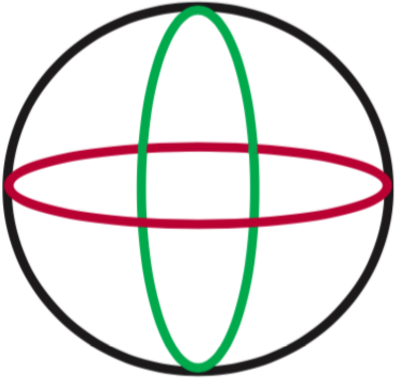
Natural Scene Statistics/Image Processing

The whole manifold of images is not accessible in a useful way, a space of small image patches might in fact contain quite useful information.

Solution: observe 3×3 patches.

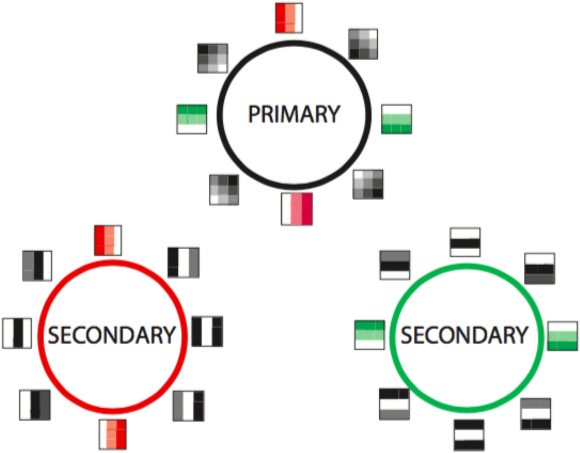


Natural Scene Statistics/Image Processing



Three circle model

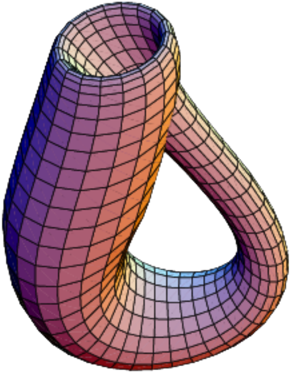
Natural Scene Statistics/Image Processing



Three circle model in the data

Natural Scene Statistics/Image Processing

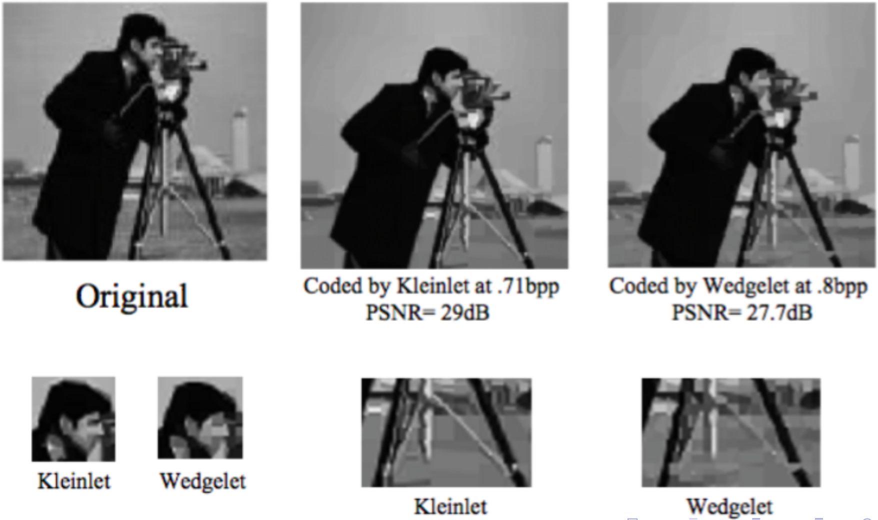
Klein Bottle



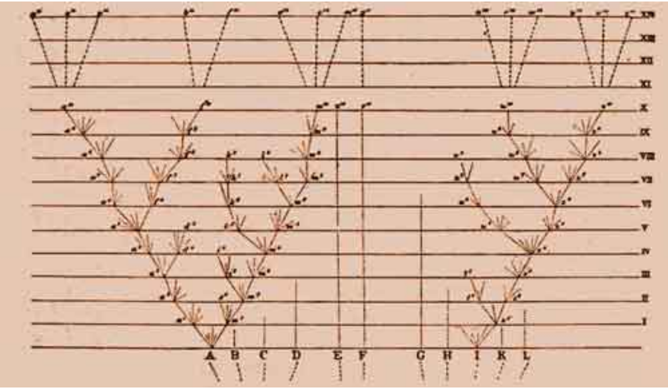
A Picture is worth 1,000 words

J. Perea, G. Carlsson: Compression based on the Klein bottle mode (Kleinlets).

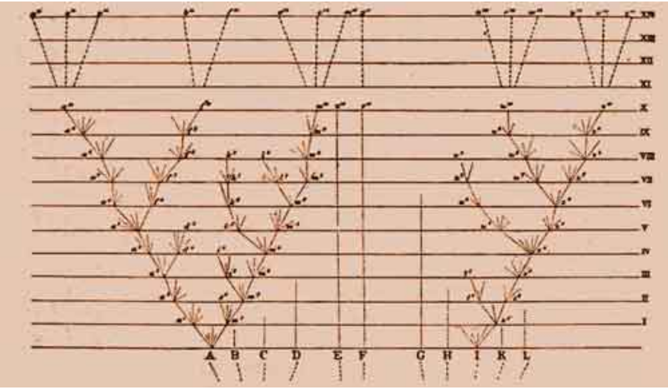
The evidence for Kleinlets over Wedglets



Applications of Persistent Homology



Applications of Persistent Homology



Tree of Life

Applications of Persistent Homology

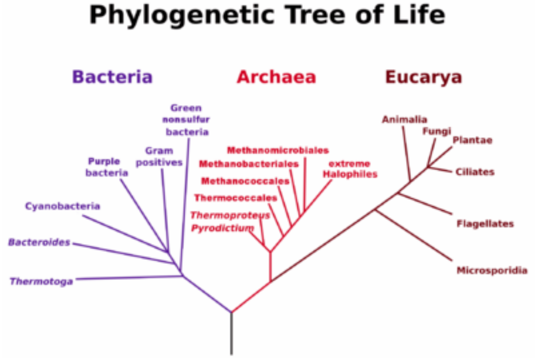
- 1970s molecular phylogenetic analysis based on nucleotide and protein sequences

Applications of Persistent Homology

- 1970s molecular phylogenetic analysis based on nucleotide and protein sequences
- 1977 Carl Woese identifies archaea as new domain in life

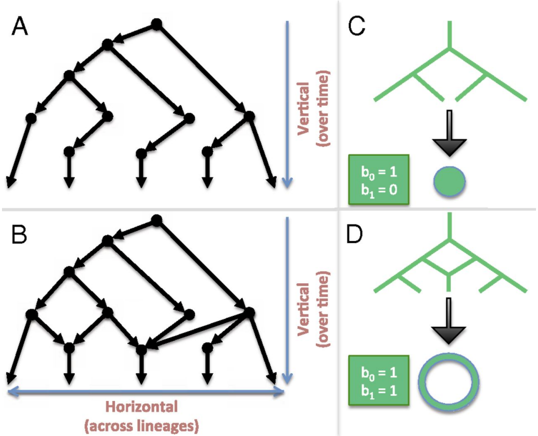
Applications of Persistent Homology

- 1970s molecular phylogenetic analysis based on nucleotide and protein sequences
- 1977 Carl Woese identifies archaea as new domain in life
- since 1990s a true revolution in genomic sequencing techniques providing hard data for evolutionary biology



Applications of Persistent Homology

Viral Evolution (Topology of viral evolution by J.M. Chan, G. Carlsson, and R. Rabadan)

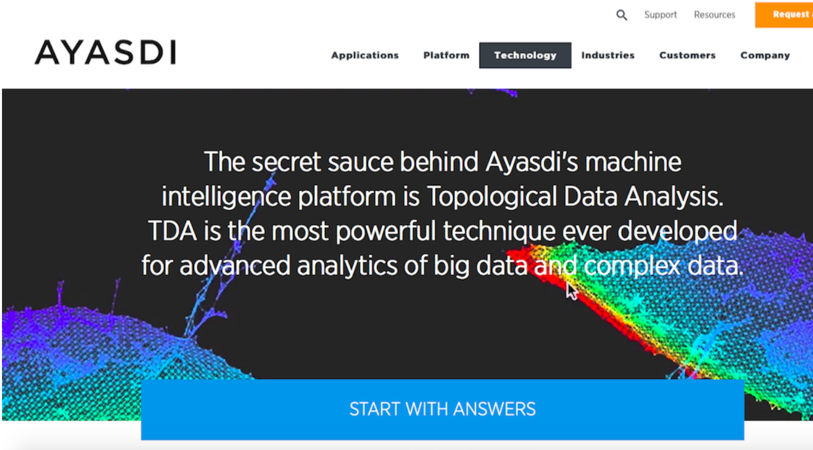


Representing Shape

A very popular TDA method for representing shape is called mapper and was developed by G. Singh, F. Memoli and G. Carlsson.

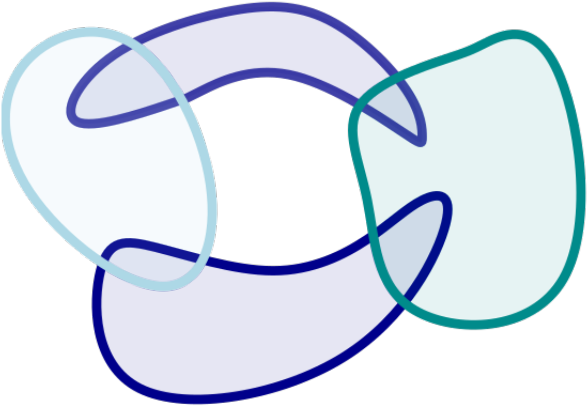
Representing Shape

A very popular TDA method for representing shape is called mapper and was developed by G. Singh, F. Memoli and G. Carlsson.



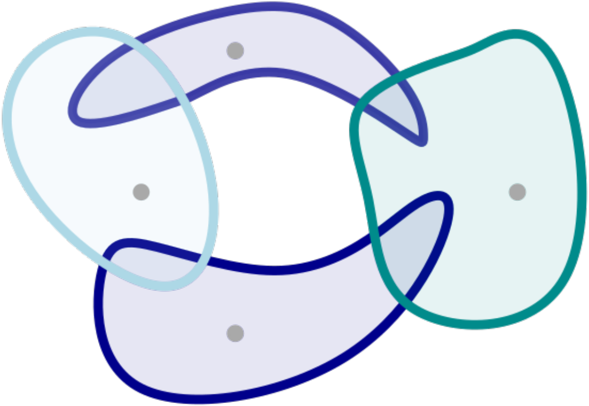
Representing Shape

Suppose we have a covering of a circle:



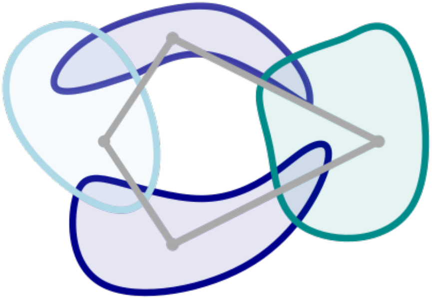
Representing Shape

We assign a vertex to each connected component of this covering



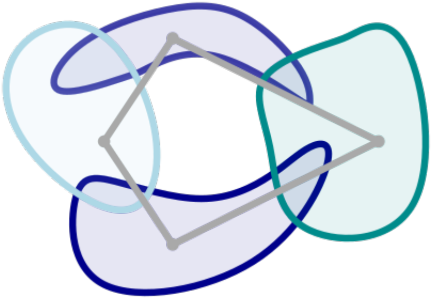
Representing Shape

When precisely two connected components intersect, we connect the corresponding vertices with an edge.



Representing Shape

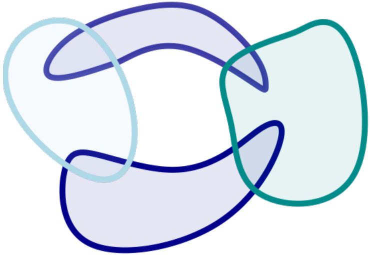
When precisely two connected components intersect, we connect the corresponding vertices with an edge.

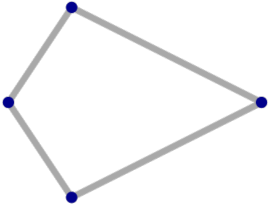


When more than two, add a face of appropriate dimension.

Representing Shape

Voila!





Representing Shape

Topological version of Mapper

Setting:

We are given a space X equipped with a continuous map $f\colon X \to Z$ to a parameter space Z, and that the space Z is equipped with a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ for some finite indexing set A.

Representing Shape

Topological version of Mapper

Setting:
We are given a space X equipped with a continuous map $f\colon X \to Z$ to a parameter space Z, and that the space Z is equipped with a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ for some finite indexing set A.

- Since f is continuous, the sets $f^{-1}(U_\alpha)$ form an open covering of X.

Representing Shape

Topological version of Mapper

Setting:
We are given a space X equipped with a continuous map $f\colon X \to Z$ to a parameter space Z, and that the space Z is equipped with a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ for some finite indexing set A.

- Since f is continuous, the sets $f^{-1}(U_\alpha)$ form an open covering of X.
- We write $f^{-1}(U_\alpha) = \cup_{j=1}^{j_\alpha} V(\alpha, i)$ where j_α is the number of connected components of $f^{-1}(U_\alpha)$. We write $\overline{\mathcal{U}}$ for the covering of X obtained by taking these connected components.

Representing Shape

Topological version of Mapper

Setting:
We are given a space X equipped with a continuous map $f\colon X \to Z$ to a parameter space Z, and that the space Z is equipped with a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ for some finite indexing set A.

- Since f is continuous, the sets $f^{-1}(U_\alpha)$ form an open covering of X.
- We write $f^{-1}(U_\alpha) = \cup_{j=1}^{j_\alpha} V(\alpha, i)$ where j_α is the number of connected components of $f^{-1}(U_\alpha)$. We write $\overline{\mathcal{U}}$ for the covering of X obtained by taking these connected components.
- Represent the topological space by a nerve of $\overline{\mathcal{U}}$.

Representing Shape

The Statistical version of Mapper

- Define a reference map $f\colon X \to Z$, where X is the given a point cloud and Z is the reference metric space.

Representing Shape

The Statistical version of Mapper

- Define a reference map $f\colon X \to Z$, where X is the given a point cloud and Z is the reference metric space.
- Select a covering $\mathcal{U}$ of Z.

Representing Shape

The Statistical version of Mapper

- Define a reference map $f\colon X \to Z$, where X is the given a point cloud and Z is the reference metric space.
- Select a covering $\mathcal{U}$ of Z.
- If $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, then construct the subsets $X_\alpha = f^{-1}(U_\alpha)$.

Representing Shape

The Statistical version of Mapper

- Define a reference map $f\colon X \to Z$, where X is the given a point cloud and Z is the reference metric space.
- Select a covering $\mathcal{U}$ of Z.
- If $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, then construct the subsets $X_\alpha = f^{-1}(U_\alpha)$.
- The analog of taking connected components in the point cloud world is clustering. Clusters form a covering of X parametrized by pairs (α, c), where $\alpha \in A$ and c is one of the clusters of X_α.

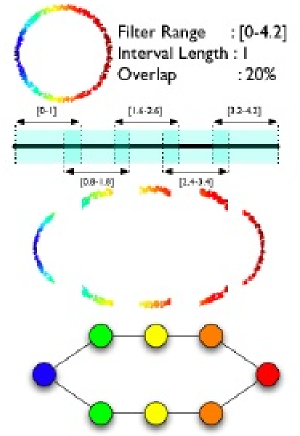
Representing Shape

The Statistical version of Mapper

- Define a reference map $f: X \to Z$, where X is the given a point cloud and Z is the reference metric space.
- Select a covering $\mathcal{U}$ of Z.
- If $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, then construct the subsets $X_\alpha = f^{-1}(U_\alpha)$.
- The analog of taking connected components in the point cloud world is clustering. Clusters form a covering of X parametrized by pairs (α, c), where $\alpha \in A$ and c is one of the clusters of X_α.
- Construct a graph whose vertex set is the set of all possible such pairs (α, c), and where an edge connects (α_1, c_1) and (α_2, c_2) if and only if the corresponding clusters have a point in common.

Representing Shape

The Statistical version of Mapper



Example:
Consider point cloud data which is sampled from a noisy circle in $\mathbb{R}^2$, and the filter $f(x) = ||x - p||^2$, where p is the left most point in the data.

Vertices are colored by the average filter value.

Mapper

The Miller-Reaven diabetes study

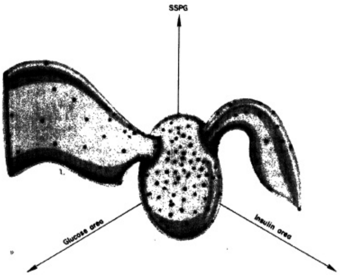
G.M. Reaven and R.G. Miller conducted a diabetes study at Stanford in the 1970'.

Mapper

The Miller-Reaven diabetes study

G.M. Reaven and R.G. Miller conducted a diabetes study at Stanford in the 1970'.

145 patients were included and six quantities were measured: age, relative weight, fasting plasma glucose, area under the plasma glucose curve for the three hour glucose tolerance test(OGTT), area under the plasma insulin curve for OGTT, steady state plasma glucose response.

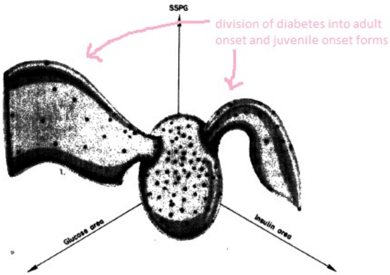


Mapper

The Miller-Reaven diabetes study

G.M. Reaven and R.G. Miller conducted a diabetes study at Stanford in the 1970'.

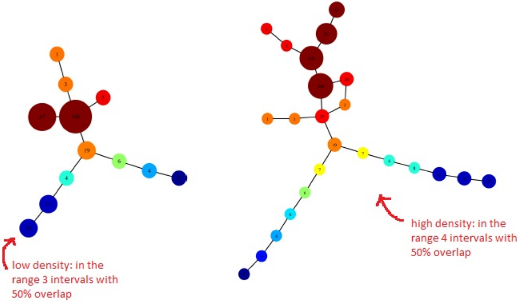
145 patients were included and six quantities were measured: age, relative weight, fasting plasma glucose, area under the plasma glucose curve for the three hour glucose tolerance test(OGTT), area under the plasma insulin curve for OGTT, steady state plasma glucose response.



Mapper

The Miller-Reaven diabetes study

If we take the filter to be a density estimator, we get the following representations for two different resolutions:

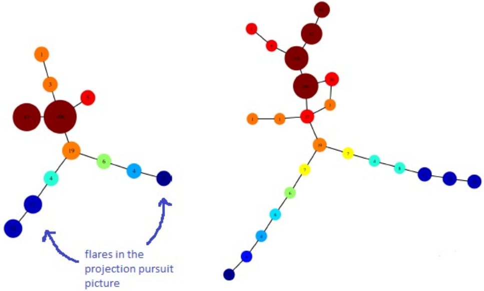


Red is indicative of high density, and blue of low. The size of the node and the number indicate the size of the cluster.

Mapper

The Miller-Reaven diabetes study

If we take the filter to be a density estimator, we get the following representations for two different resolutions:



Red is indicative of high density, and blue of low. The size of the node and the number indicate the size of the cluster.

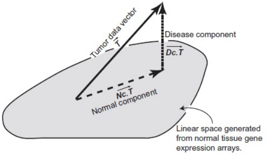
Mapper

Breast cancer data

What should the filter be?

Mapper

Breast cancer data
What should the filter be?



- Take linear combinations of normal expression data and denote the subspace they span by $\mathcal{N}$.
- Decompose the original data - vector $\vec{T}$ into normal-like expression, $\vec{Nc.T}$, which is the projection onto $\mathcal{N}$.
- The disease, deviation $\vec{Dc.T}$ from normal-like expression, is defined to be the difference between diseased tissue expression and normal-like expression.

Mapper

Breast cancer data

The family of functions we take as filters is

$$f_{p,k}(\vec{V}) = [\sum |g_r|^p]^{\frac{k}{p}}$$

where $\vec{V} = \langle g_1, g_2, \ldots, g_s \rangle$ and coordinates g_i are individual genes.

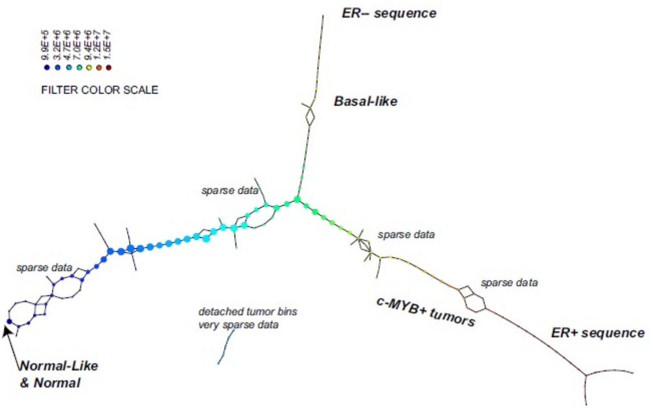
If $k = 1, p = 2$, the function computes standard (Euclidean) norm of a vector.

Essentially, all these different filter functions, $f_{p,k}$, measure the overall amount of deviation from the normal state.

The effect of the different choices of p determining the choice of L^p norm is that, for larger values of p the weight of genes with larger expression levels is greater.

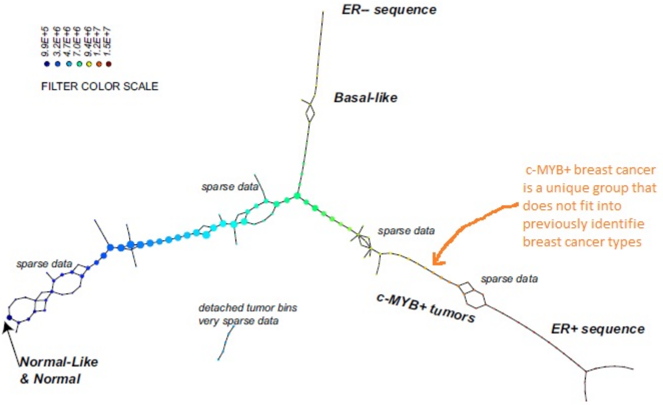
Mapper

Breast cancer data



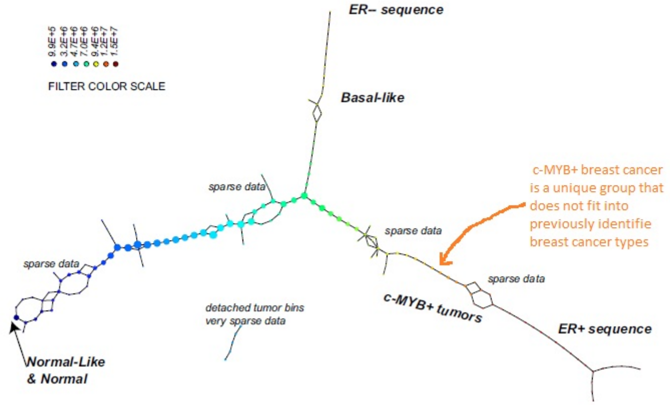
Mapper

Breast cancer data



Mapper

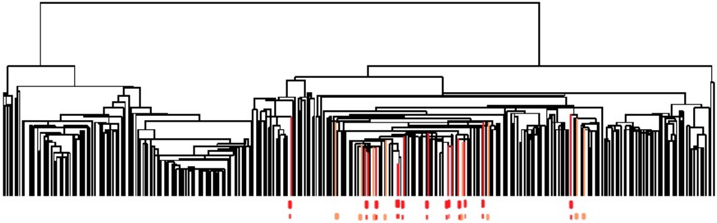
Breast cancer data



Both ER+ tumors (Estrogen Receptor positive) showed a 100%

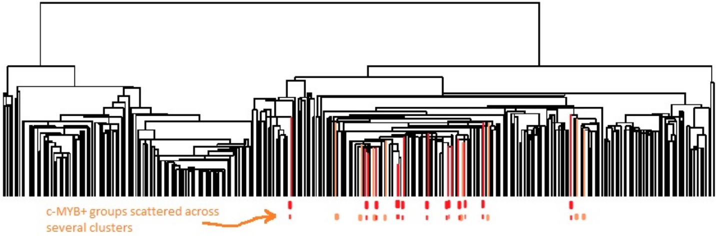
Mapper

Clustering versus Mapper



Mapper

Clustering versus Mapper



Representing Shape

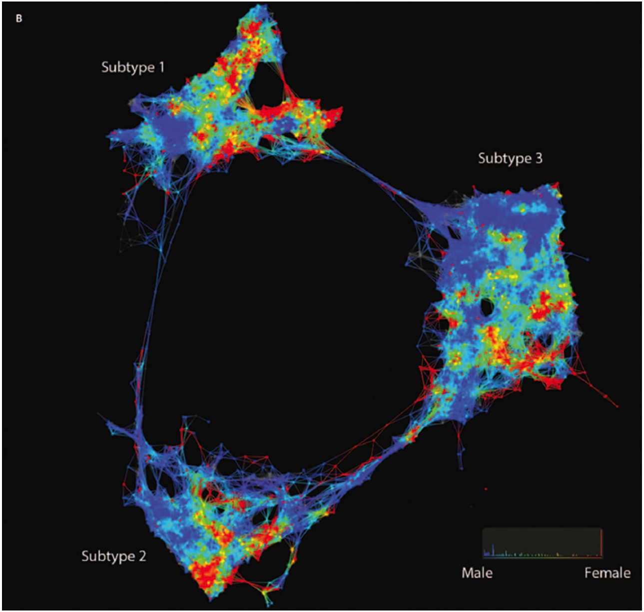
Type 2 Diabetes

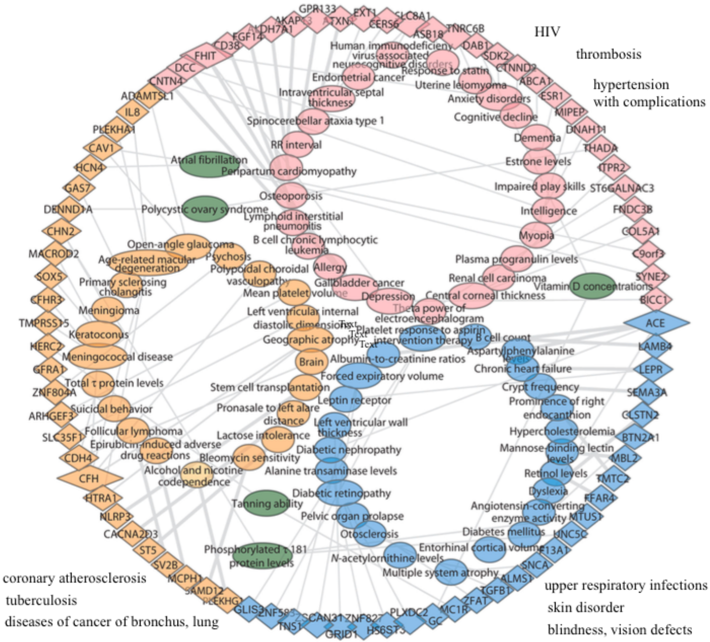
Current clinical definitions classify diabetes into three major subtypes: type 1 diabetes (T1D), T2D, and maturity-onset diabetes of the young.

Differences among T2D patients suggest several T2D subtypes.

Li Li, Wei-Yi Cheng, Benjamin S. Glicksberg, Omri Gottesman, Ronald Tamler, Rong Chen, Erwin P. Bottinger, and Joel T. Dudley (Icahn School of Medicine at Mount Sinai) use a topology-based approach.


Science
Translational
Medicine
28 October 2015
AAAS


B
Subtype 1
Subtype 3
Subtype 2
Male
Female


HIV
thrombosis
hypertension with complications
coronary atherosclerosis
tuberculosis
diseases of cancer of bronchus, lung
upper respiratory infections
skin disorder
blindness, vision defects
GPR133
AKAP13
ATXN
EXT1
SLC8A1
CERS6
ASB18
TNRC6B
ALDH7A1
FGF14
CD38
FHIT
DCC
CNTN4
ADAMTSL1
IL8
PLEKHA1
CAV1
HCN4
GAS7
DENND1A
CHN2
MACROD2
SOX5
CFHR3
TMPRSS15
HERC2
GFRA1
ZNF804A
ARHGEF3
SLC35F1
CDH4
CFH
HTRA1
NLRP3
CACNA2D3
ST5
SV2B
MCPH1
SAMD12
PLEKHG1
GLIS3
ZNF583
TNS1
SCAN31
GRID1
ZNF827
HS6ST3
PLXDC2
GC
MC1R
ZFAT
TGFB1
ALMS1
SNCA
F13A1
UNC5C
MTUS1
FFAR4
TMTC2
MBL2
BTN2A1
CLSTN2
SEMA3A
LEPR
LAMB4
ACE
BICC1
SYNE2
C9orf3
COL5A1
FNDC3B
ST6GALNAC3
ITPR2
THADA
DNAH11
MIPEP
ESR1
ABCA1
CTNND2
SDK2
DAB1
Human immunodeficiency virus-associated neurocognitive disorders
Response to statin
Endometrial cancer
Uterine leiomyoma
Intraventricular septal thickness
Anxiety disorders
Spinocerebellar ataxia type 1
Cognitive decline
Dementia
RR interval
Atrial fibrillation
Peripartum cardiomyopathy
Estrone levels
Impaired play skills
Osteoporosis
Polycystic ovary syndrome
Lymphoid interstitial pneumonitis
Intelligence
Myopia
B cell chronic lymphocytic leukemia
Open-angle glaucoma
Age-related macular degeneration
Psychosis
Polypoidal choroidal vasculopathy
Allergy
Plasma progranulin levels
Primary sclerosing cholangitis
Gallbladder cancer
Renal cell carcinoma
Vitamin D concentrations
Mean platelet volume
Depression
Central corneal thickness
Meningioma
Theta power of electroencephalogram
Left ventricular internal diastolic dimension
Keratoconus
Platelet response to aspirin intervention therapy
B cell count
Geographic atrophy
Aspartylphenylalanine levels
Meningococcal disease
Brain
Albumin-to-creatinine ratios
Chronic heart failure
Forced expiratory volume
Total τ protein levels
Stem cell transplantation
Crypt frequency
Prominence of right endocanthion
Leptin receptor
Suicidal behavior
Pronasale to left alare distance
Left ventricular wall thickness
Hypercholesterolemia
Follicular lymphoma
Lactose intolerance
Epirubicin-induced adverse drug reactions
Mannose-binding lectin levels
Diabetic nephropathy
Bleomycin sensitivity
Retinol levels
Alanine transaminase levels
Alcohol and nicotine codependence
Dyslexia
Diabetic retinopathy
Angiotensin-converting enzyme activity
Tanning ability
Pelvic organ prolapse
Diabetes mellitus
Otosclerosis
Phosphorylated τ181 protein levels
Entorhinal cortical volume
N-acetylornithine levels
Multiple system atrophy

Applied Algebraic Topology Research Network

I am one of the co-director of the Applied Algebraic Topology Research Network, which hosts a weekly Online Seminar. Recordings of our seminar are available at our YouTube Channel, which has over 6000 YouTube subscribers.