Topology and Data: A Tour

Tao Hou, University of Oregon

Why topological methods for data?

- In modern days, data of various kinds is being produced at an unprecedented rate, in part because of
  - new experimental methods
  - availability of highly powerful computing technology

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- In modern days, data of various kinds is being produced at an unprecedented rate, in part because of
  - new experimental methods
  - availability of highly powerful computing technology
- Also, the *nature* of the data we are obtaining is significantly different.

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- In modern days, data of various kinds is being produced at an unprecedented rate, in part because of
  - new experimental methods
  - availability of highly powerful computing technology
- Also, the *nature* of the data we are obtaining is significantly different.
- In-class task: try to Chatgpt the following:
  - “*What are the different types of data that could be produced in morden sience, engineering and everyday life?*”

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Challenges:
  - Data is often very high-dimensional, restricting our ability to understand it (e.g., visualize) and process it
  - Data obtained is also very noisy and has more missing information

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Challenges:
  - Data is often very high-dimensional, restricting our ability to understand it (e.g., visualize) and process it
  - Data obtained is also very noisy and has more missing information
- Our ability to analyze this data, both in terms of quantity and the nature of the data, is clearly not keeping pace

Why topological methods for data?

- In this course, we will see how *geometry*, but especially ***topology***, can be applied to make useful contributions to analyze various kinds of data

Based on: Gunnar Carlsson, Topology and Data

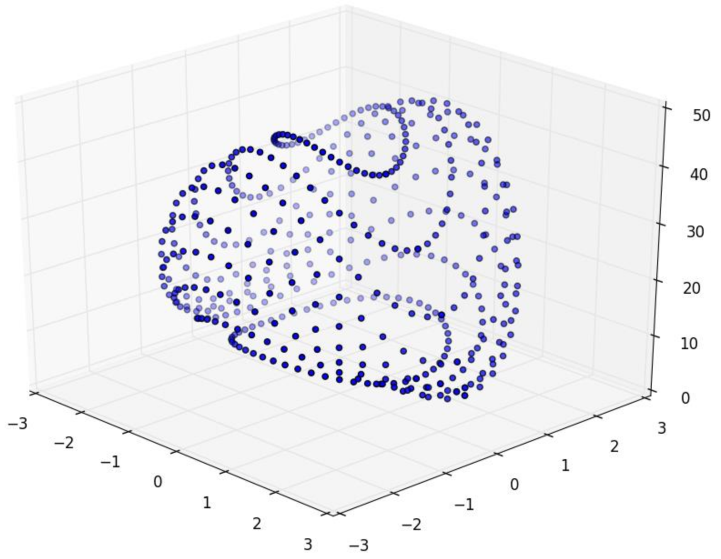
Why topological methods for data?

- In this course, we will see how *geometry*, but especially ***topology***, can be applied to make useful contributions to analyze various kinds of data
- Geometry and topology are very natural tools for this:
  - E.g., geometry can be regarded as the study of distance between points
  - We typically work with large finite sets of data with distance defined on the objects (e.g., point cloud)

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- In this course, we will see how *geometry*, but especially ***topology***, can be applied to make useful contributions to analyze various kinds of data
- Geometry and topology are very natural tools for this:
  - E.g., geometry can be regarded as the study of distance between points
  - We typically work with large finite sets of data with distance defined on the objects (e.g., point cloud)



Based on: Gunnar Carlsson, Topology and Data

Img source: https://stackoverflow.com/questions/31294355/create-surface-grid-from-point-cloud-data-in-python

Why topological methods for data?

- In this course, we will see how *geometry*, but especially ***topology***, can be applied to make useful contributions to analyze various kinds of data
- Geometry and topology are very natural tools for this:
  - E.g., geometry can be regarded as the study of distance between points
  - We typically work with large finite sets of data with distance defined on the objects (e.g., point cloud)
  - Tools from the various branches of geometry can be adapted to the study of point clouds

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- One example of data analytical methods based on geometry (and statistics) is the famous **principal component analysis** (PCA)

Why topological methods for data?

- One example of data analytical methods based on geometry (and statistics) is the famous **principal component analysis** (PCA)
- It's a dimension-reduction technique
  - projecting high-dimensional data into lower-dimensional space
  - while keeping the spread of the data in the most significant directions

Why topological methods for data?

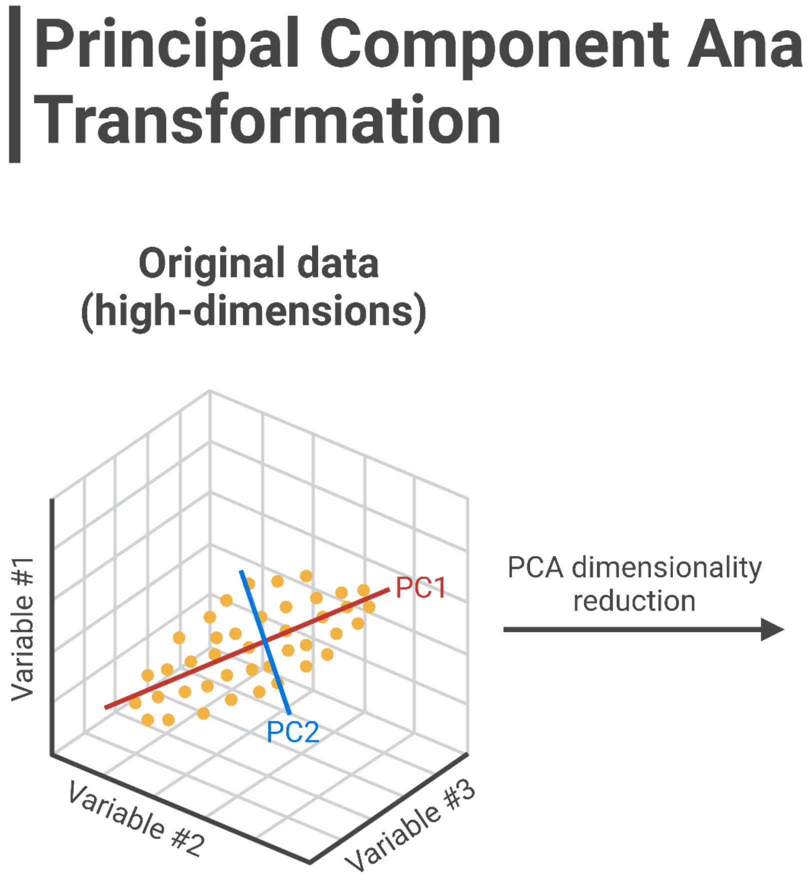


Image source: https://www.biorender.com/template/principal-component-analysis-pca-transformation

Why topological methods for data?

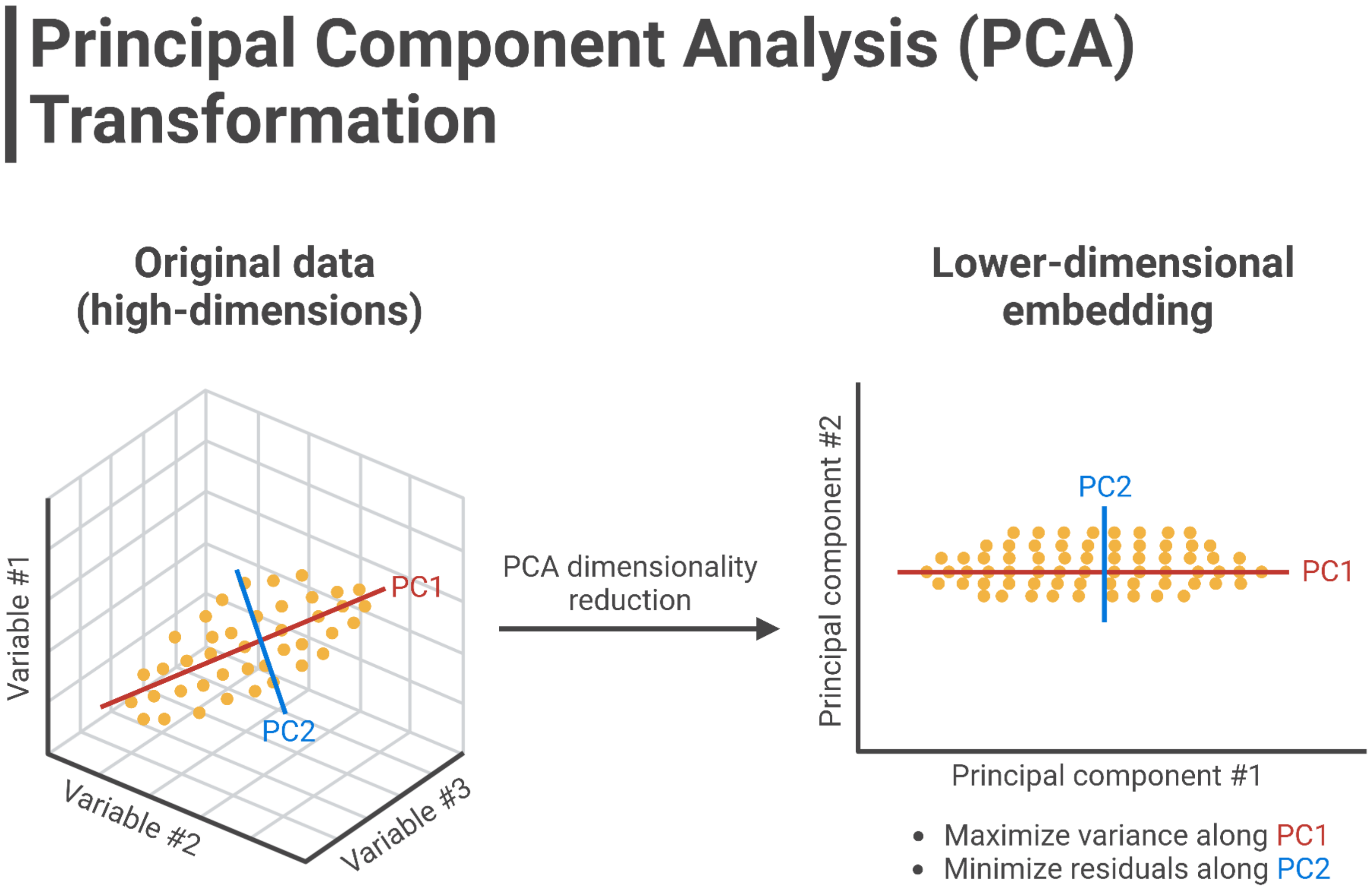


Image source: https://www.biorender.com/template/principal-component-analysis-pca-transformation

Why topological methods for data?

Some key points when applying these different methods to data analysis

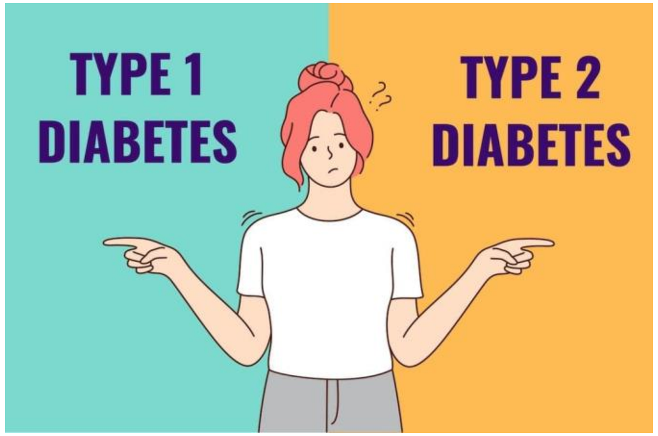
- **Qualitative** information is important:
    - An important goal of data analysis is obtain knowledge from the data (to understand how it is organized on a large scale)

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

Some key points when applying these different methods to data analysis

- **Qualitative** information is important:
  - An important goal of data analysis is obtain knowledge from the data (to understand how it is organized on a large scale)
  - E.g., when analyzing a data set for diabetes patients, it's important to understand that there are two types of the disease first, namely the juvenile and adult onset forms



Based on: Gunnar Carlsson, Topology and Data

Image source: https://www.sugarfit.com/blog/difference-between-type-1-and-type-2-diabetes

Why topological methods for data?

Some key points when applying these different methods to data analysis

- **Qualitative** information is important:
    - An important goal of data analysis is obtain knowledge from the data (to understand how it is organized on a large scale)
    - E.g., when analyzing a data set for diabetes patients, it's important to understand that there are two types of the disease first, namely the juvenile and adult onset forms
    - We could also further develop quantitative methods for distinguishing them, but the first insight about the distinct forms of the disease is key

Why topological methods for data?

- **Summaries** are more valuable than *individual parameter choices*:

Based on: Gunnar Carlsson, Topology and Data
Img: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Why topological methods for data?

- **Summaries** are more valuable than *individual parameter choices*:
  - When clustering some data points, one approach is to connect two points of distance small than a threshold ε

Based on: Gunnar Carlsson, Topology and Data
Img: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Why topological methods for data?

- **Summaries** are more valuable than *individual parameter choices*:
  - When clustering some data points, one approach is to connect two points of distance small than a threshold ε
  - The points are then clustered based on taking the connected components of the constructed graph

Based on: Gunnar Carlsson, Topology and Data
Img: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

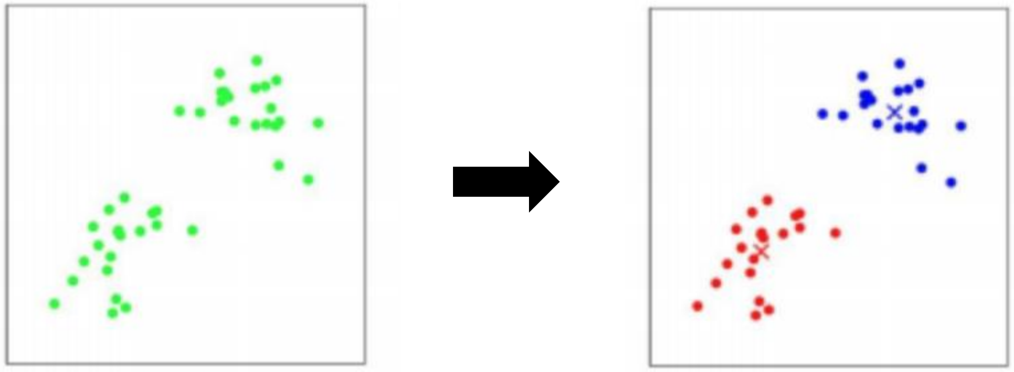
Why topological methods for data?

- **Summaries** are more valuable than *individual parameter choices*:
  - When clustering some data points, one approach is to connect two points of distance small than a threshold ε
  - The points are then clustered based on taking the connected components of the constructed graph

Based on: Gunnar Carlsson, Topology and Data
Img: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Why topological methods for data?

- **Summaries** are more valuable than *individual parameter choices*:
  - When clustering some data points, one approach is to connect two points of distance small than a threshold ε
  - The points are then clustered based on taking the connected components of the constructed graph



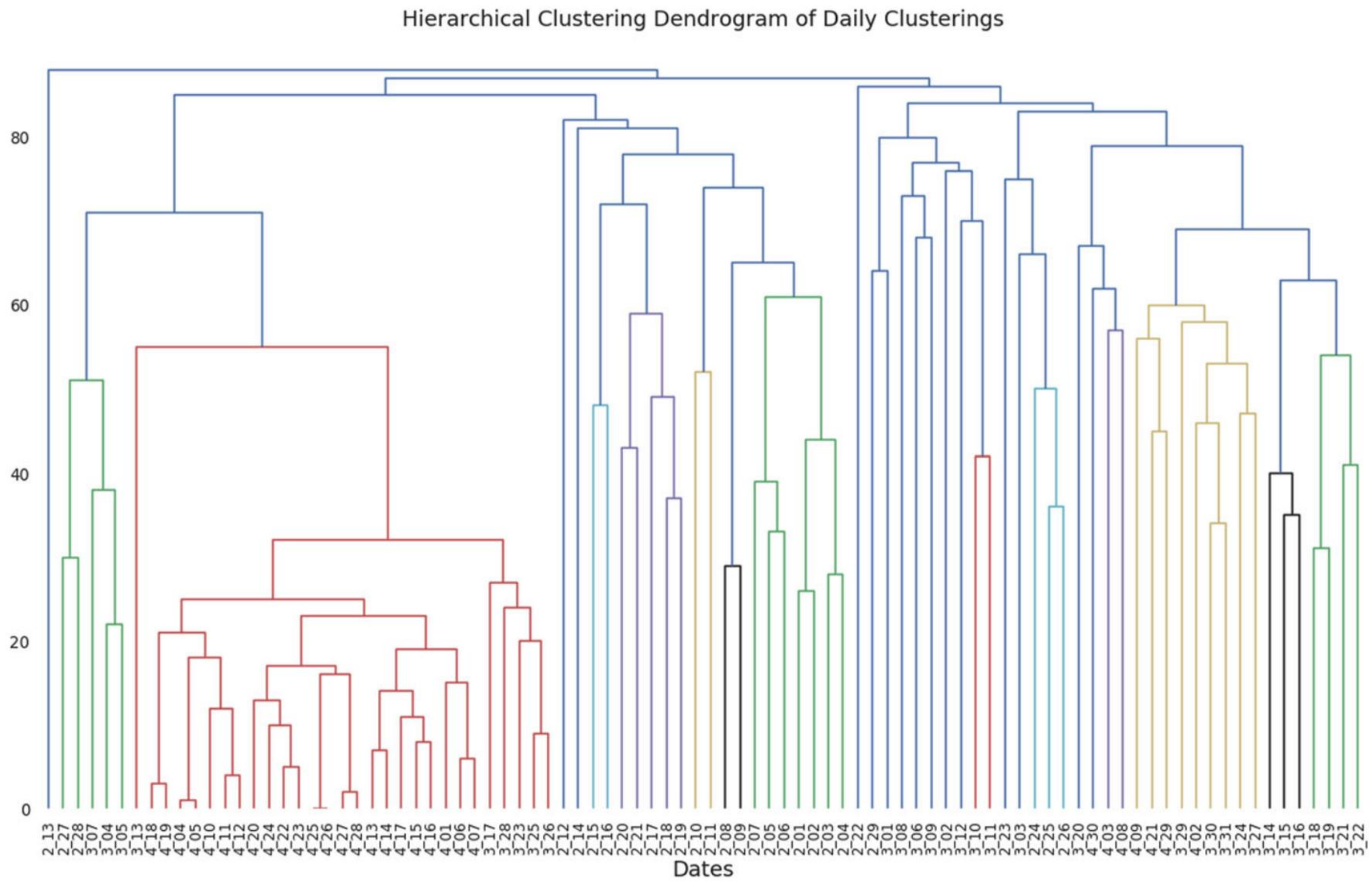
Based on: Gunnar Carlsson, Topology and Data
Img: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Why topological methods for data?

- **Summaries** are more valuable than *individual parameter choices*:
  - When clustering some data points, one approach is to connect two points of distance small than a threshold ε
  - The points are then clustered based on taking the connected components of the constructed graph
  - Some clustering theory has been trying to determine the optimal choice of ε (the parameter)
  - But it is now well understood that maintaining the summary of the entire behavior of clustering under all possible parameter ε at once (called **dendrogram**) is more helpful

Based on: Gunnar Carlsson, Topology and Data

- Example of dendrogram (also called **Hierarchical Clustering**) hashtag usage on twitter during COVID-19:



Img source: Cruickshank and Carley. Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering

Why topological methods for data?

- **Summaries** are more valuable than *individual parameter choices*:
  - When clustering some data points, one approach is to connect two points of distance small than a threshold ε
  - The points are then clustered based on taking the connected components of the constructed graph
  - Some clustering theory has been trying to determine the optimal choice of ε (the parameter)
  - But it is now well understood that maintaining the summary of the entire behavior of clustering under all possible parameter ε at once (called **dendrogram**) is more helpful
  - We will learn topological methods that helps summarize **invariants** of data under a change of parameters

Based on: Gunnar Carlsson, Topology and Data

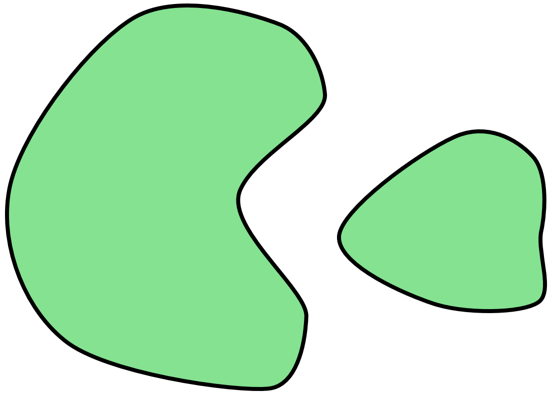
Why topological methods for data?

- Topology is exactly the branch of mathematics dealing with qualitative geometric information:
  - what the connected components of a space are
  - and more generally the connectivity information: the classification of loops and higher dimensional holes within the space

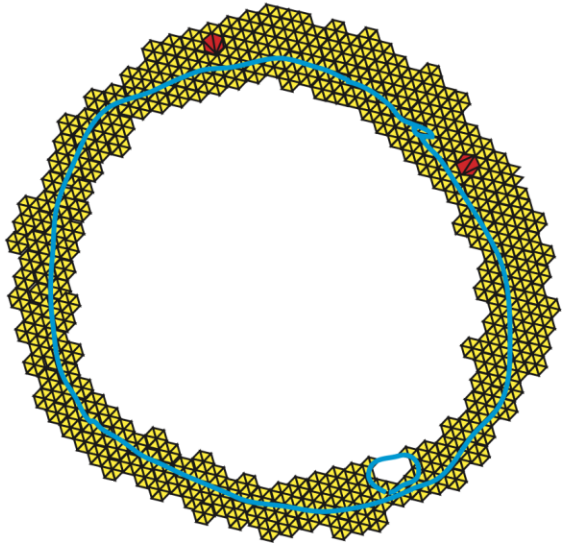
Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Topology is exactly the branch of mathematics dealing with qualitative geometric information:
  - what the connected components of a space are
  - and more generally the connectivity information: the classification of loops and higher dimensional holes within the space



0-dimensional hole (gaps between different components)

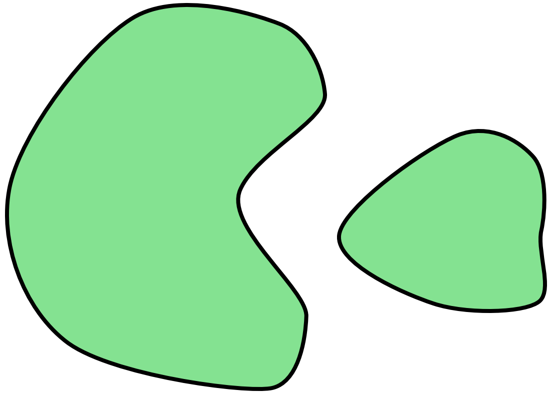


1-dimensional hole

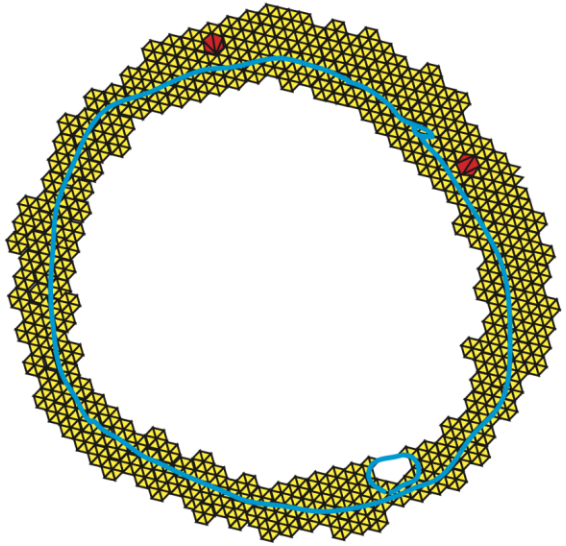
Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Topology is exactly the branch of mathematics dealing with qualitative geometric information:
  - what the connected components of a space are
  - and more generally the connectivity information: the classification of loops and higher dimensional holes within the space
- This suggests that topological methodologies for data should be helpful in studying data qualitatively



0-dimensional hole (gaps between different components)



1-dimensional hole

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Compared to straightforward geometric methods, topology studies properties of data in a way
  - much less sensitive to a choice of metrics (distance)
  - e.g., not sensitive geometric properties such as curvature

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Compared to straightforward geometric methods, topology studies properties of data in a way
  - much less sensitive to a choice of metrics (distance)
  - e.g., not sensitive geometric properties such as curvature
- Technically speaking, topology ignores the quantitative values of distances and replaces them with the notion “nearness” of points without relying on distance

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Compared to straightforward geometric methods, topology studies properties of data in a way
  - much less sensitive to a choice of metrics (distance)
  - e.g., not sensitive geometric properties such as curvature
- Technically speaking, topology ignores the quantitative values of distances and replaces them with the notion “nearness” of points without relying on distance
- This insensitivity to the metric is useful in studying situations
  - where we can only understand data (and its metric) in a coarse way
  - where we want a “global” view of data instead its local geometric details

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- Compared to straightforward geometric methods, topology studies properties of data in a way
  - much less sensitive to a choice of metrics (distance)
  - e.g., not sensitive geometric properties such as curvature
- Technically speaking, topology ignores the quantitative values of distances and replaces them with the notion “nearness” of points without relying on distance
- This insensitivity to the metric is useful in studying situations
  - where we can only understand data (and its metric) in a coarse way
  - where we want a “global” view of data instead its local geometric details
- We will look at more concrete examples of what topology can do later on

Based on: Gunnar Carlsson, Topology and Data

Why topological methods for data?

- In summary

“Data has Shape, Shape has Meaning”

Shape of Data

- “Data has Shape, Shape has Meaning ”
- Ex1:

Image source: https://web.stanford.edu/class/archive/ee/ee392n/ee392n.1146/lecture/may13/EE392n_TDA_online.pdf

Shape of Data

- “Data has Shape, Shape has Meaning ”
- Ex1:

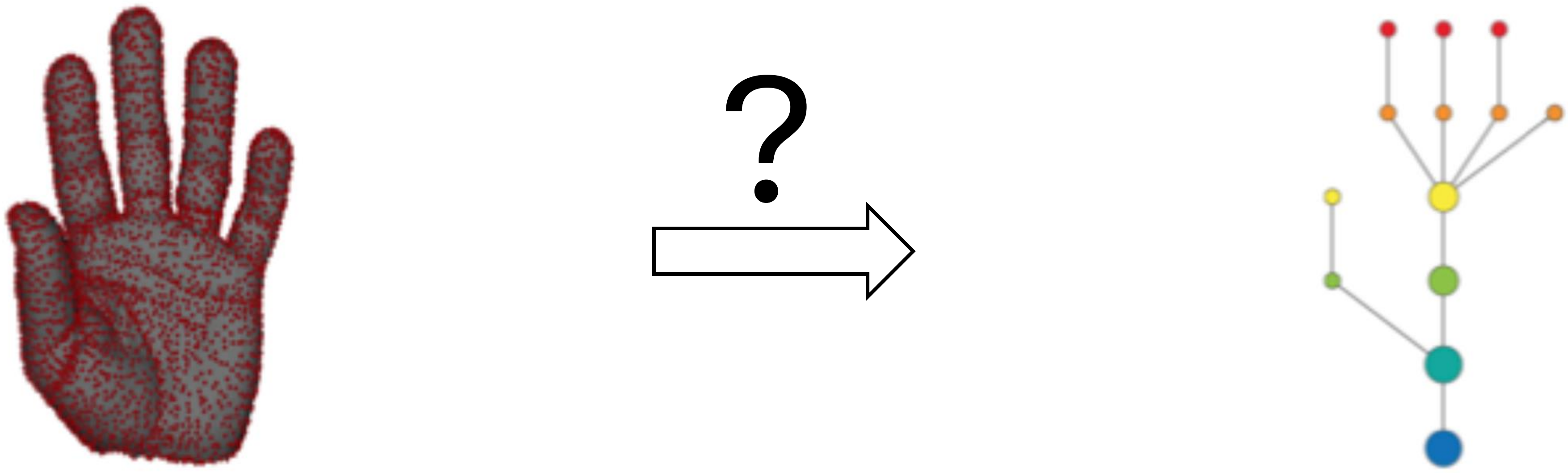


Image source: https://web.stanford.edu/class/archive/ee/ee392n/ee392n.1146/lecture/may13/EE392n_TDA_online.pdf

Shape of Data

- “Data has Shape, Shape has Meaning ”
- Ex1: Using **Mapper** graph

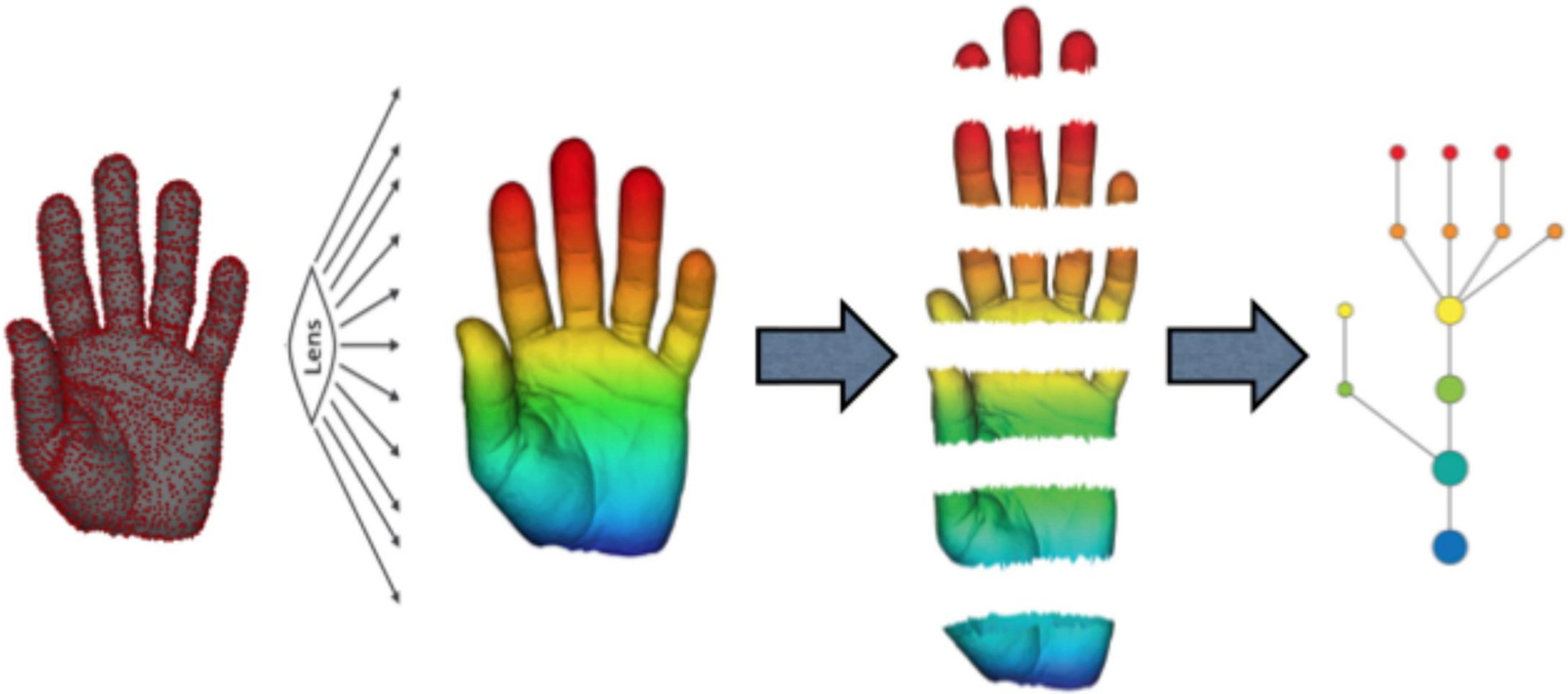


Image source: https://web.stanford.edu/class/archive/ee/ee392n/ee392n.1146/lecture/may13/EE392n_TDA_online.pdf

Shape of Data

- “Data has Shape, Shape has Meaning ”
- Ex2: How do we describe the shapes of the following three sets of points (for pokemons) and characterize their differences?

A B C

Image source: https://kids.frontiersin.org/articles/10.3389/frym.2021.551557

Shape of Data

- “Data has Shape, Shape has Meaning ”
- Ex2: How do we describe the shapes of the following three sets of points (for pokemons) and characterize their differences?
  - It’s easy for human beings to ‘delineate’ the ‘shapes’ of these points form and to understand their difference (we naturally have intuitions about shapes)
  - But how to make computer understand shape?

A B C

Image source: https://kids.frontiersin.org/articles/10.3389/frym.2021.551557

Topology, a branch of mathematics

- Topology: A branch of mathematics that people use to study the shapes of objects

Topology, a branch of mathematics

- Topology: A branch of mathematics that people use to study the shapes of objects
- From Wikipedia: A branch of mathematics studying the properties of a object that are *preserved under continuous deformations*, such as
  - stretching, twisting, crumpling, and bending;

  that is, without
  - tearing, gluing, closing holes, opening holes, or passing through itself.

Topology, a branch of mathematics

- Topology: A branch of mathematics that people use to study the shapes of objects
- From Wikipedia: A branch of mathematics studying the properties of a object that are *preserved under continuous deformations*, such as
  - stretching, twisting, crumpling, and bending;

  that is, without
  - tearing, gluing, closing holes, opening holes, or passing through itself.
- Another interpretation: "*rubber-sheet geometry*": concerned with the essence of shapes rather than their rigid measurements.

Topology, a branch of mathematics

- Topology: A branch of mathematics that people use to study the shapes of objects
- From Wikipedia: A branch of mathematics studying the properties of a object that are *preserved under continuous deformations*, such as
  - stretching, twisting, crumpling, and bending;

  that is, without
  - tearing, gluing, closing holes, opening holes, or passing through itself.
- Another interpretation: "*rubber-sheet geometry*": concerned with the essence of shapes rather than their rigid measurements.
- In my own words: Topology studies *how points in a space connect to each other* within the space

Topology (informally speaking)

- Suppose we have a circular rubber band.

Topology (informally speaking)

- Suppose we have a circular rubber band.
- Topology describes the properties of it that stay the same if we *stretch* it or *shrink* it or *bend* it, but without *gluing* things together or *breaking* it.

Topology (informally speaking)

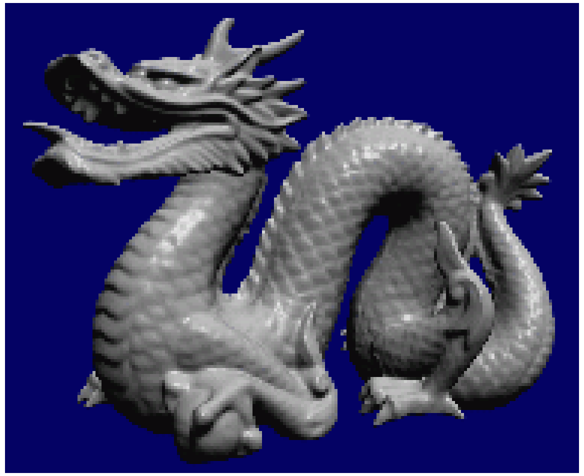
- Suppose we have a circular rubber band.
- Topology describes the properties of it that stay the same if we *stretch* it or *shrink* it or *bend* it, but without *gluing* things together or *breaking* it.
- From a topological viewpoint, the following three are equivalent

=

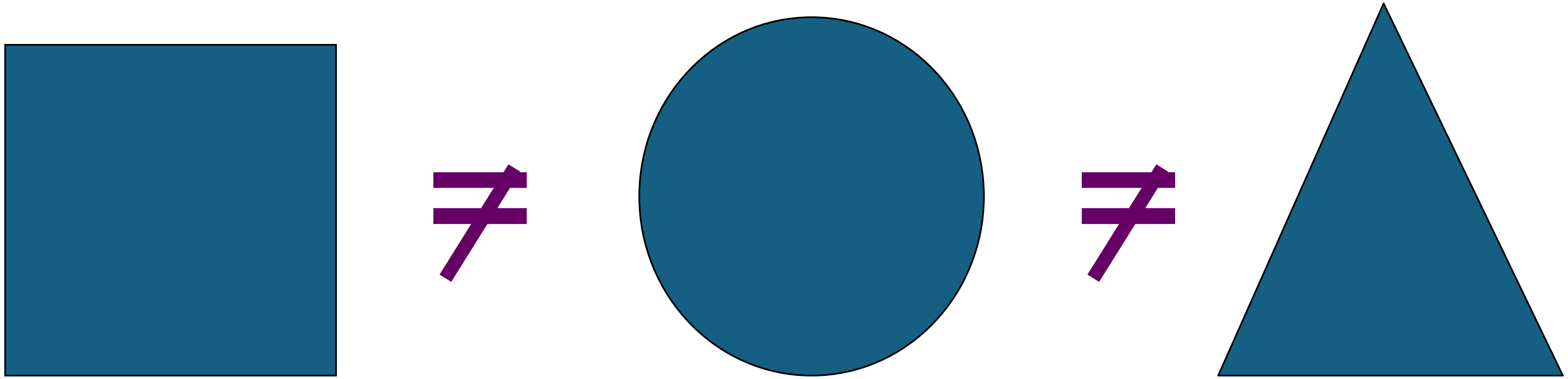
=

Topology (informally speaking)

- Suppose we have a circular rubber band.
- Topology describes the properties of it that stay the same if we *stretch* it or *shrink* it or *bend* it, but without *gluing* things together or *breaking* it.
- While the following are **not** equivalent

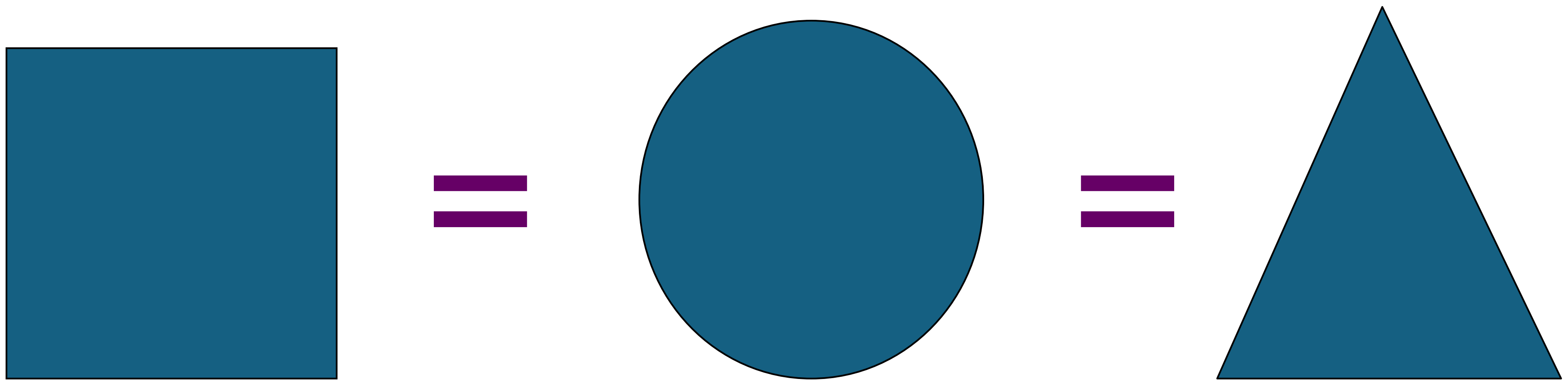
 $\neq$  $\neq$ 

Geometry



Courtesy of Isabel K. Darcy: http://www.math.uiowa.edu/~idarcy/AppliedTopology.html

Topology

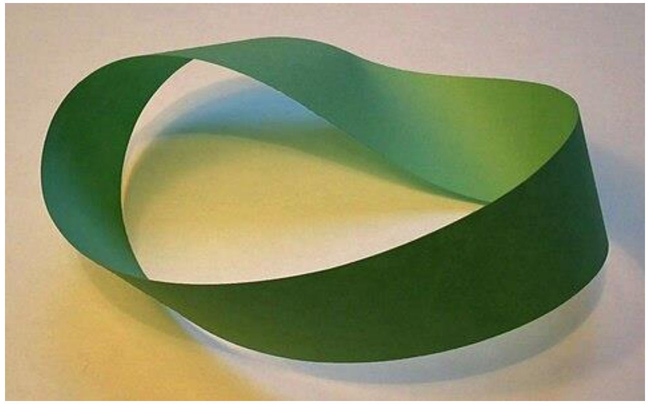


Courtesy of Isabel K. Darcy: http://www.math.uiowa.edu/~idarcy/AppliedTopology.html

Topology (informally speaking)

- Suppose we have a circular rubber band.
- Topology describes the properties of it that stay the same if we *stretch* it or *shrink* it or *bend* it, but without *gluing* things together or *breaking* it.
- Trick question: is the rubber band equivalent to a mobius strip (formed by inversely glueing the two ends of a paper tape)

$\stackrel{?}{=}$



Topology (informally speaking)

- Suppose we have a circular rubber band.
- Topology describes the properties of it that stay the same if we *stretch* it or *shrink* it or *bend* it, but without *gluing* things together or *breaking* it.
- Answer: no. Several reasons:
    1. To form a mobius band from the rubber band, you have to *break* the rubber band and *re-glue* the two ends (inversely), and these operations are not allowed (they are *not continuous*)

Topology (informally speaking)

- Suppose we have a circular rubber band.
- Topology describes the properties of it that stay the same if we *stretch* it or *shrink* it or *bend* it, but without *gluing* things together or *breaking* it.
- Answer: no. Several reasons:
  1. To form a mobius band from the rubber band, you have to *break* the rubber band and *re-glue* the two ends (inversely), and these operations are not allowed (they are *not continuous*)
  2. The rubber band is *orientable* (has two sides) while the mobius band is not (cannot differentiate the two sides): Orientability is an *invariant* that should be preserved if two spaces are equivalent. (TDA heavily draw upon other invariants such as *homology*, which we will look at later)

Topology (more formally)

- From the book *Topology* by James Munkres

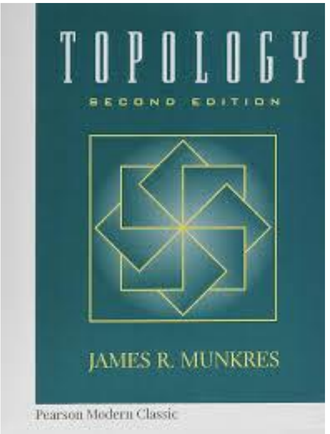
Definition. A ***topology*** on a set X is a collection $\mathcal{T}$ of subsets of X having the following properties:

(1) $\varnothing$ and X are in $\mathcal{T}$.

(2) The union of the elements of any subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

(3) The intersection of the elements of any finite subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

A set X for which a topology $\mathcal{T}$ has been specified is called a ***topological space***.



Topology (more formally)

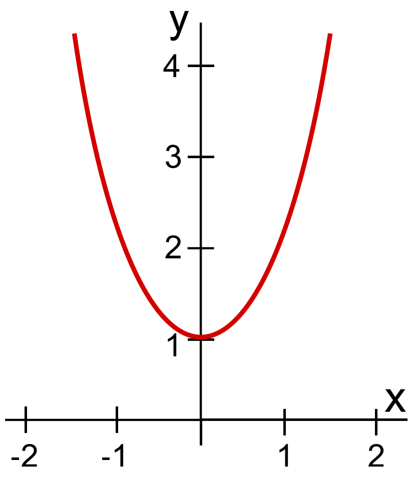
- In the previous definition, each set $S \in \mathcal{T}$ (notice that $S \subseteq X$) is called an *open set*.
- The open sets are usually chosen to provide a notion of "*nearness*" without having a notion of distance defined.
- A topology allows defining properties such as
  - Continuity
  - Connectedness
  - Compactness

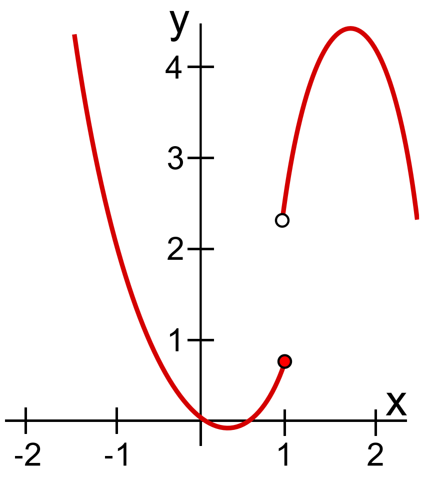
  without defining a distance.

Topology (more formally)

- An example of continuity:

Continuous function





Non-continous function

image src: https://www.sfu.ca/math-coursenotes/Math%20157%20Course%20Notes/sec_ContinuityIVT.html

Continuity and topological equivalence (more formally)

Let X and Y be topological spaces. A function $f : X \to Y$ is said to be ***continuous*** if for each open subset V of Y, the set $f^{-1}(V)$ is an open subset of X.

Continuity and topological equivalence (more formally)

Let X and Y be topological spaces. A function $f : X \to Y$ is said to be ***continuous*** if for each open subset V of Y, the set $f^{-1}(V)$ is an open subset of X.

Let X and Y be topological spaces; let $f : X \to Y$ be a bijection. If both the function f and the inverse function

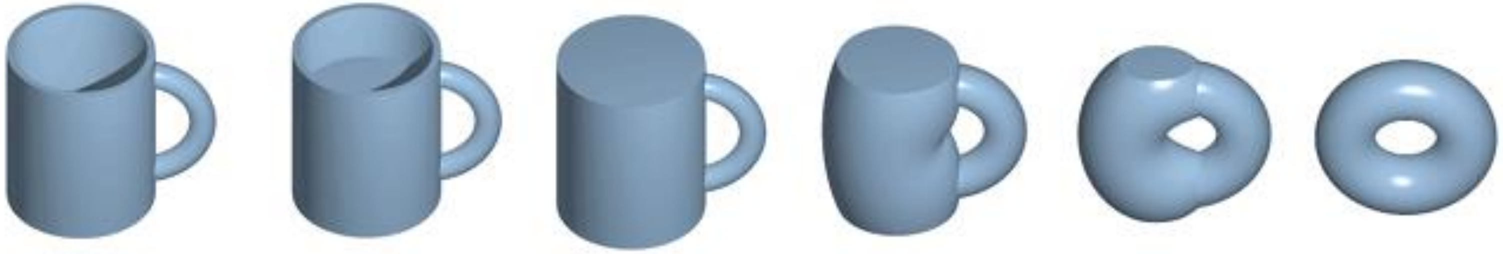
$$f^{-1} : Y \to X$$

are continuous, then f is called a ***homeomorphism***.

"Homeomorphism" is the formal terminology for "topological equivalence" that we have been talking about

Examples of homeomorphic spaces

- On Wikipedia: https://en.wikipedia.org/wiki/Topology



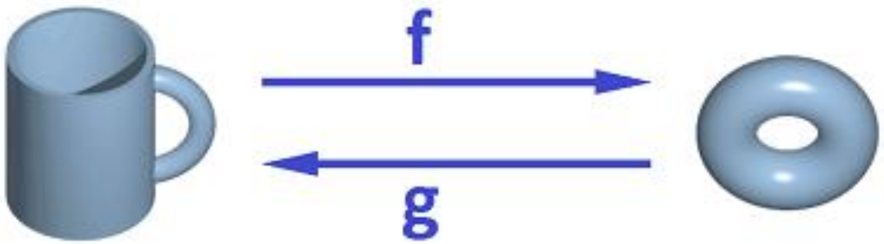


Image source: https://www.facebook.com/photo/?fbid=113751385892899&set=a.113751795892858

Problem for our practical purpose

- Now we have (very roughly) defined a “topological structure” on our data, which can be used to describe the “shape” for the data
- But to be honest, this “topology” (a set of subset of the dataset X) is too abstract, there is almost nothing we can do about it

Problem for our practical purpose

- Now we have (very roughly) defined a "topological structure" on our data, which can be used to describe the "shape" for the data
- But to be honest, this "topology" (a set of subset of the dataset X) is too abstract, there is almost nothing we can do about it
- We need to "encode" the topology of your data into some format processible by the computer
- For this, we utilize some "numeric invariants" for the topological spaces
  - *Invariant*: something that does not change between spaces that are topologically equivalent (homeomorphic)

Problem for our practical purpose

- Now we have (very roughly) defined a "topological structure" on our data, which can be used to describe the "shape" for the data
- But to be honest, this "topology" (a set of subset of the dataset X) is too abstract, there is almost nothing we can do about it
- We need to "encode" the topology of your data into some format processible by the computer
- For this, we utilize some "numeric invariants" for the topological spaces
  - *Invariant*: something that does not change between spaces that are topologically equivalent (homeomorphic)
- `Remark: most formally, the numeric invariants are indeed called algebraic invariants`

A toy version of algebraic invariant

- Counting the number of pieces and number of holes in an object

A toy version of algebraic invariant

- Counting the number of pieces and number of holes in an object

Number of pieces: 5
Number of holes: 0

Number of pieces: 4
Number of holes: 2

Number of pieces: 1
Number of holes: 3

- We will study this algebraic invariant more extensively later

Image source: https://kids.frontiersin.org/articles/10.3389/frym.2021.551557

Discovering the shape of data by connecting the dots

- Let's try to discover the shape of the pokemon below

- A problem with the data is that, it's just a discrete set of points —— it doesn't form any "meaningful shapes" that we could count (and further make computer to process)

Image source: https://kids.frontiersin.org/articles/10.3389/frym.2021.551557

Discovering the shape of data by connecting the dots

- Let’s try to discover the shape of the pokemon below

- A problem with the data is that, it’s just a discrete set of points —— it doesn’t form any “meaningful shapes” that we could count (and further make computer to process)
- In order to form some meaningful shape, we need to find a way to “connect the dots”

Image source: https://kids.frontiersin.org/articles/10.3389/frym.2021.551557

Connecting the dots

- We connect the dots by increasing their size.
  - As we make the dots larger, gaps between the dots become smaller, and eventually the dots overlap

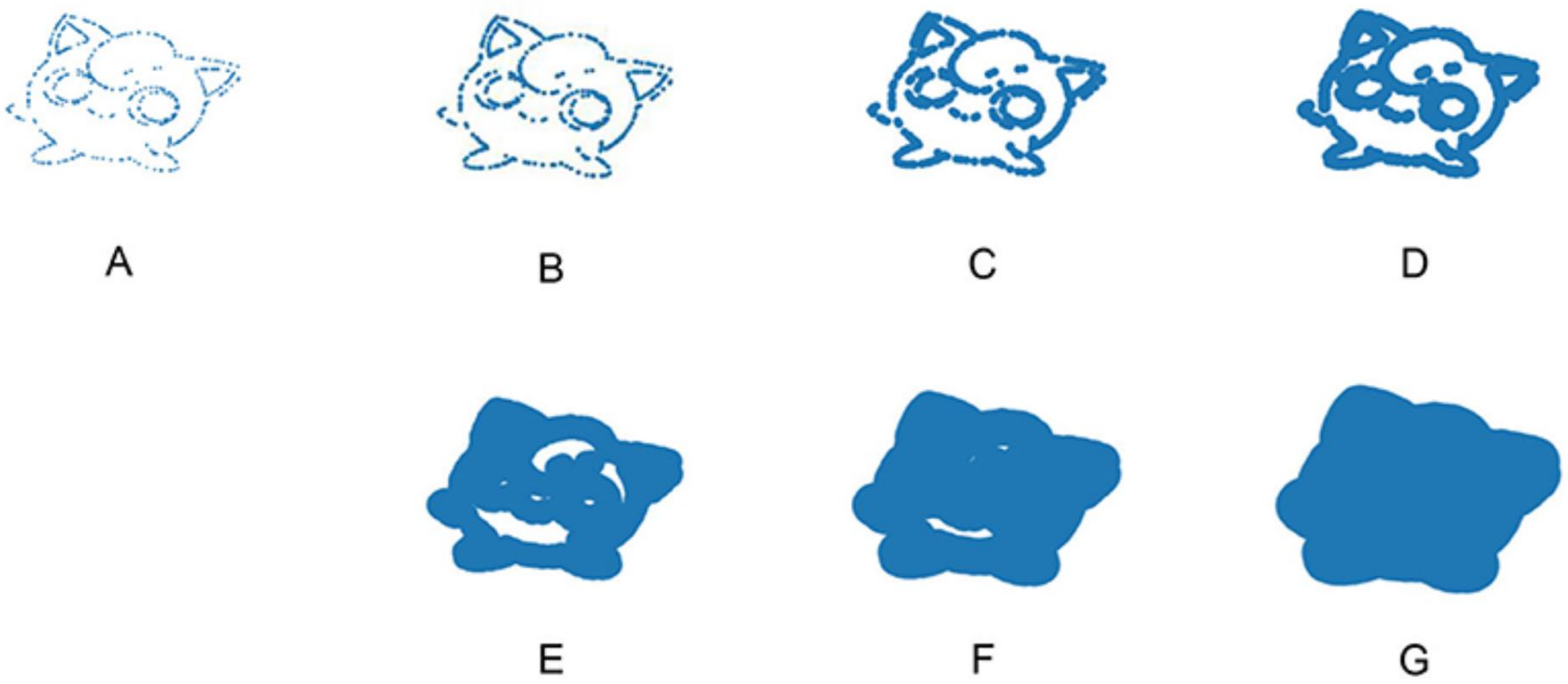


Image source: https://kids.frontiersin.org/articles/10.3389/frym.2021.551557

Connecting the dots

- Remark: a more natural way of connecting the dots by drawing lines between the dots. We will do that more formally and extensively later

	Number of pieces	Number of holes
	224	0
	101	0
	17	2
	1	6
	1	6
	1	3
	1	0

Image source:
https://kids.frontiersin.org/articles/10
.3389/frym.2021.551557

Connecting the dots

- But another question arises: which size do we choose?
- A person *may* be able to detect the “right size” for the previous example. But what about more involved shapes?
- Furthermore, how do we let computer choose such a size?

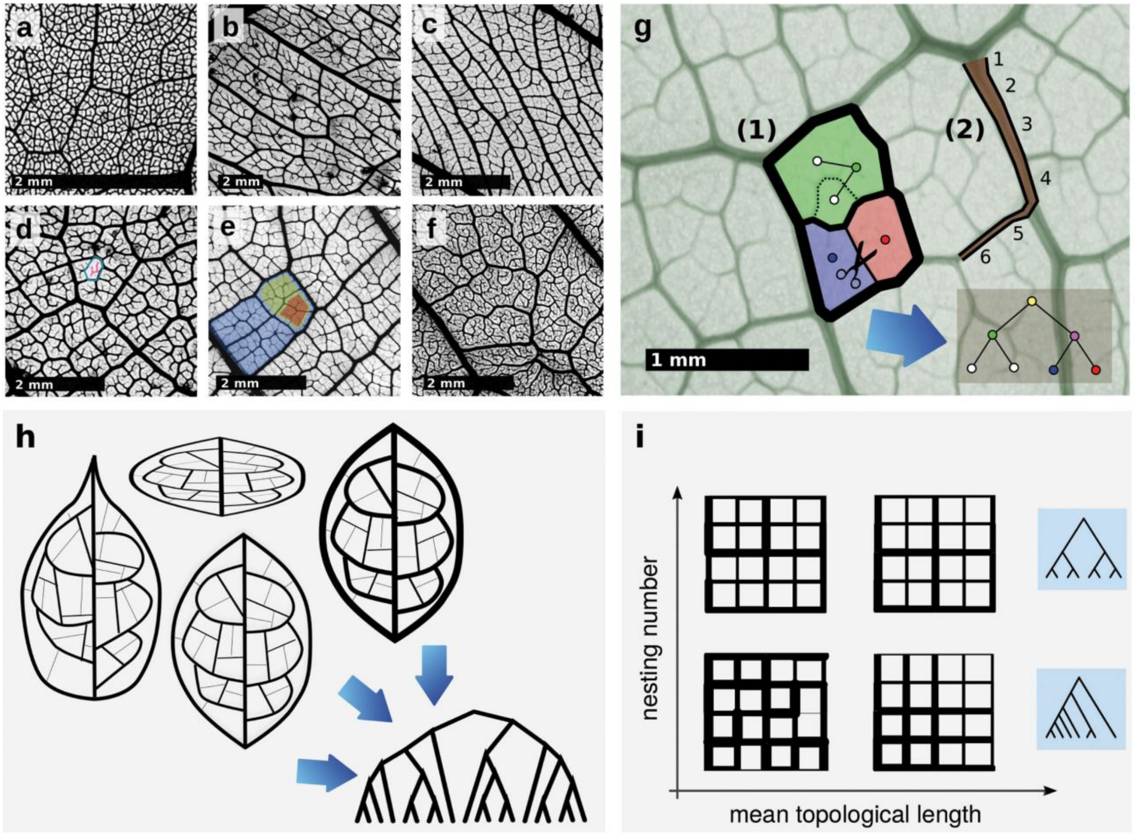
Connecting the dots

- But another question arises: which size do we choose?
- A person *may* be able to detect the "right size" for the previous example. But what about more involved shapes?
- Furthermore, how do we let computer choose such a size?
- We will study a major tool in topological data analysis (TDA) called **Persistent Homology**, which provides a solution
- Hint: Persistent Homology does not try to find such a size, but rather it *considers all the sizes* and *tracks the changes of the topological invariants, by tracking the how the pieces and holes persist,* across all the sizes

How are topological methods useful?

Examining *patterns of veins in leaves*: studied structure of >=100 leaves and found different patterns—like human fingerprints—in them

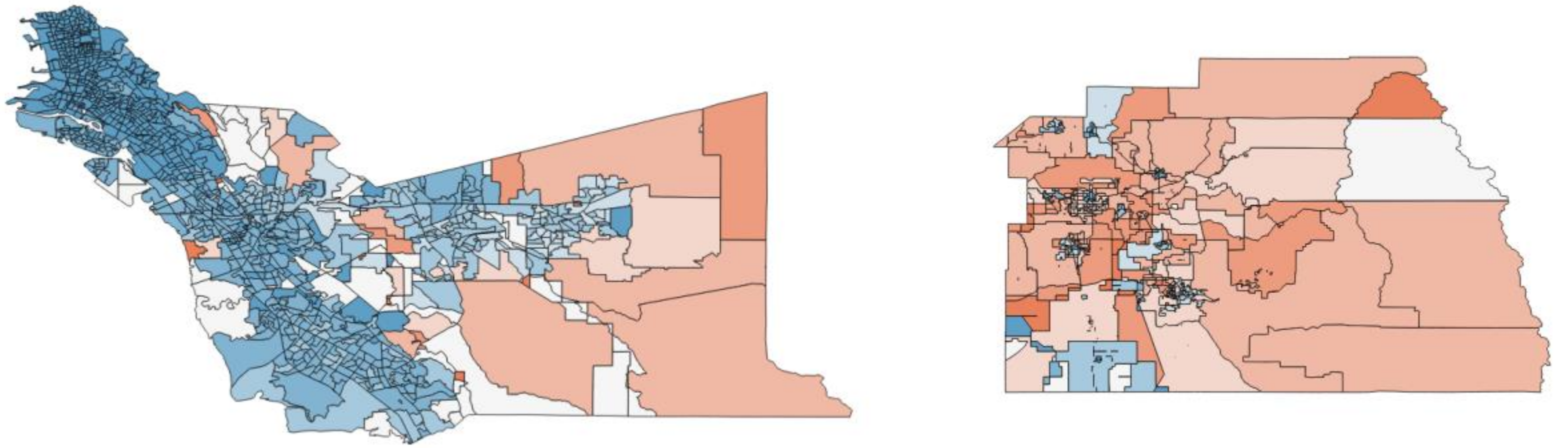
- These fingerprints help scientists identify leaves from small leaf fragments, and may also be helpful for improving our understanding of how leaves grow.

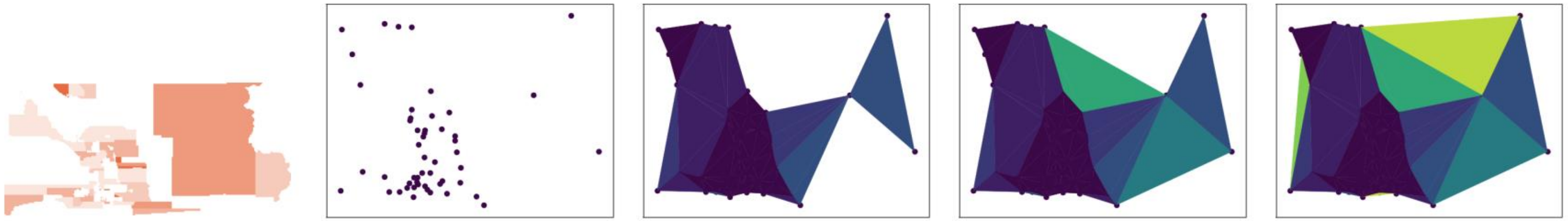


Ronellenfitsch, H., Lasser, J., Daly, D. C., and Katifori, E. 2015. Topological phenotypes constitute a new dimension in the phenotypic space of leaf venation networks.

How are topological methods useful?

Studying the voting patterns in California





Feng, M., and Porter, M. A. 2021. Persistent homology of geospatial data: A case study with voting.

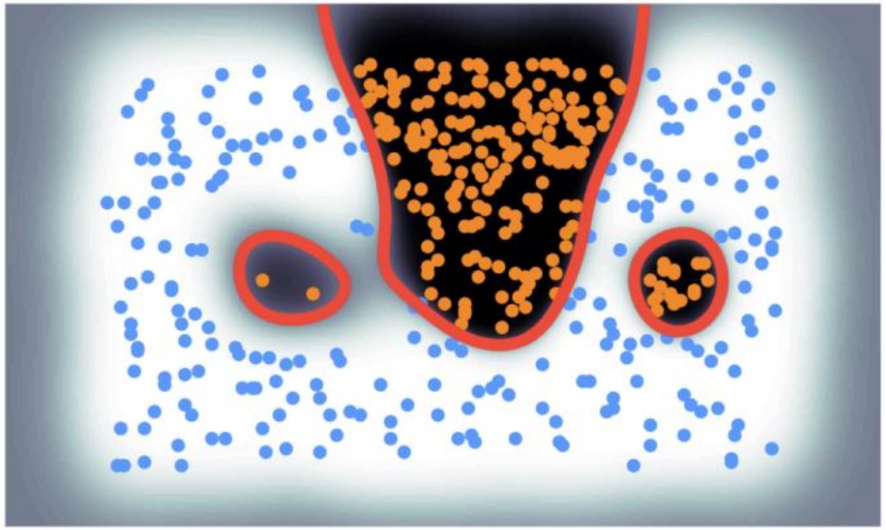
How are topological methods useful?

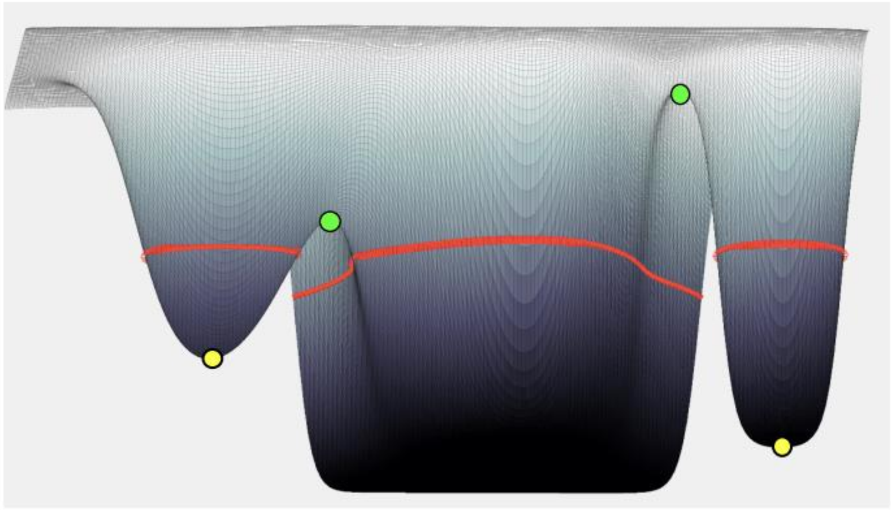
Utilizes *topological regularization* losses to rectify topological artifacts (broken legs, unrealistic thin structures, and small holes) in generating synthetic 3D models

Chen et al., Enhancing Implicit Shape Generators Using Topological Regularizations.

How are topological methods useful?

Utilizes *topological regularization* losses to reduce the topological complexity of the classification boundary of a binary classification task





Chen, Chao, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology.