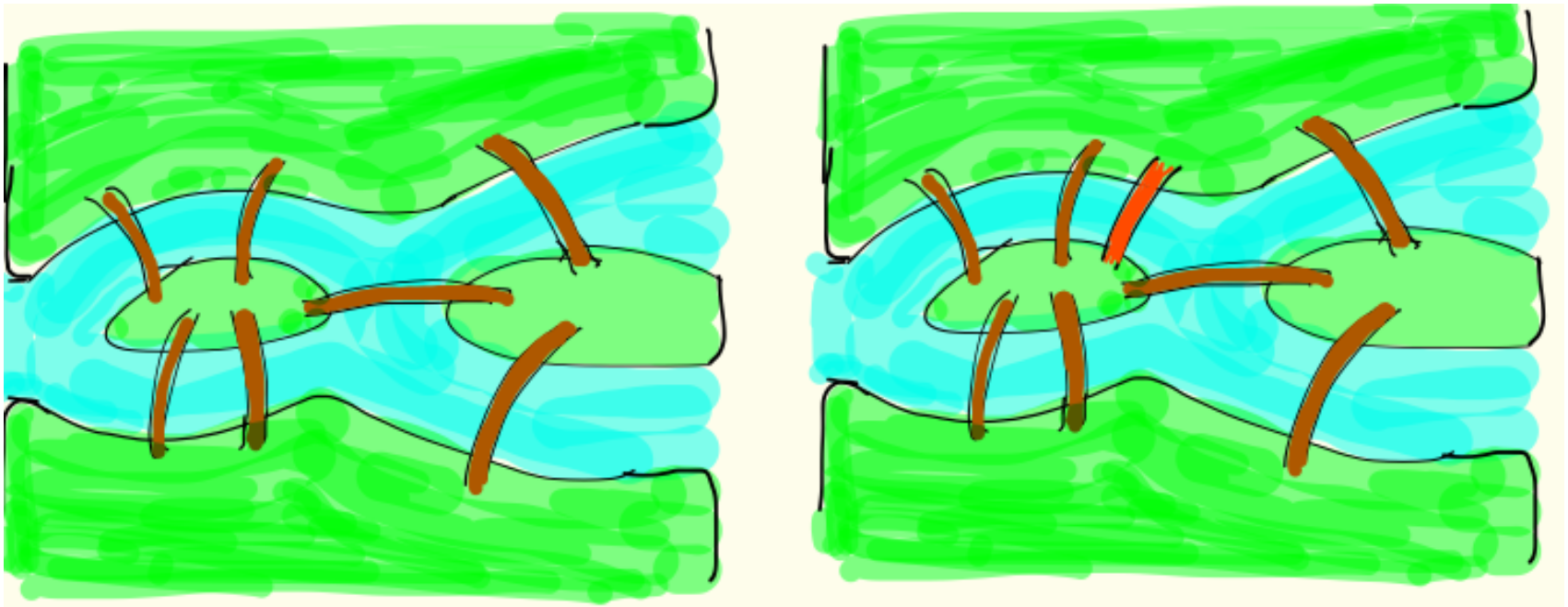
Task: In each of these two beautiful parks, try to find a path (to walk) that uses each bridge once and exactly once. Start with the park on the right first. (The blue denotes a river that contains alligators, so no swimming allowed!) Can you find a path that starts on one of the two islands?



Introduction to Topology and its Applications to Complex Data

Yogesh More

I will introduce the branch of mathematics called *Topology* through some of the puzzles that first gave rise to the subject 100 to 200 years ago. These puzzles illustrate some ideas topologists started exploring around 1900. These classical ideas have recently found applications in the analysis of complex data.

Key aspects of Topology

Three key aspects of Topology are:

1. deformation invariance
2. compressed representation (e.g. graphs or networks)
3. discrete invariants (e.g, numbers, groups, rings, ...)

What is topology?

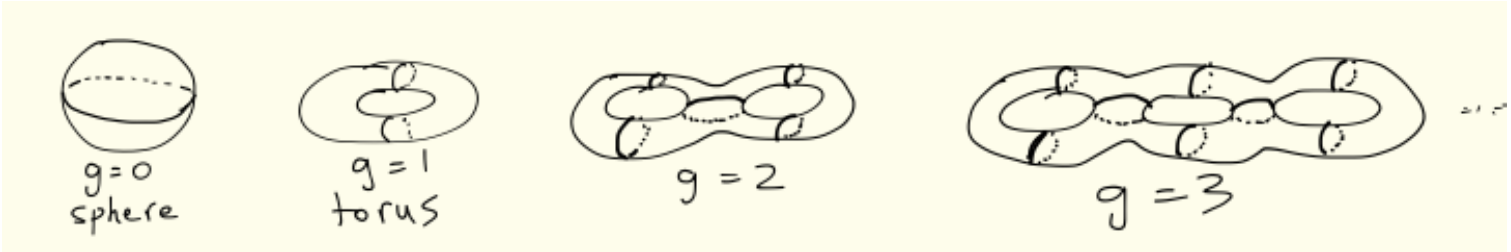
Topology is the mathematical study of the properties that are preserved through deformations, twistings, and stretchings of objects. Tearing, however, is not allowed. So topology focuses on the *qualitative* aspects of objects or shapes. (http://mathworld.wolfram.com/Topology.html)

Old joke: "A topologist is someone who cannot distinguish between a doughnut and a coffee cup"

Topology is *not* topography - the field of geoscience and planetary science comprising the study of surface shape and features of the Earth and other observable astronomical objects

History of topology and roadmap for this talk

- 1736: Leonhard Euler's paper on the Seven Bridges of Konigsberg
- 1863: Mobius gave a classification of "two-sided" surfaces: every such surface can be deformed to one of the following g holed surfaces:



- 1871: To any shape X, Betti associated *Betti numbers* $b_0(X), b_1(X), b_2(X), b_3(X), \ldots$. These numbers measure the connectedness, number of loops, holes, and higher dimensional analogs of holes in the shape X.
- Betti numbers of a (solid) 3 dimensional ball D^3 (e.g dodgeball, but not a soccer ball) are $b_0 = 1, \quad b_1 = 0, \quad b_2 = 0$.
- Betti numbers of the 2-sphere S^2 (e.g soccer ball or football, but not a dodgeball) are $b_0 = 1, \quad b_1 = 0, \quad b_2 = 1$.

History of topology and roadmap for this talk

- 1895: Henri Poincare:
  - provided new foundation to Betti numbers by introducing homology theory
  - asked whether Betti numbers are enough to specify the shape (up to homeomorphism)
  - Found the answer was no; revised question to whether contractibility of loops is enough to distinguish the 3-sphere S^3 (Poincare conjecture, took 100 years to prove the answer is yes)
- 1925: Emmy Noether shifted the focus from Betti numbers to homology *groups*
- 2000-present: Gunnar Carlsson and others are using homology theory to analyse complex data arising in all sort of situations such as cancer research, drug discovery, financial-market research, voting patterns, ...

Bridges of Konigsberg (now called Kaliningrad, in Russia)

Photo credit: Vladimir Sedach, Wikipedia

Bridges of Konigsberg

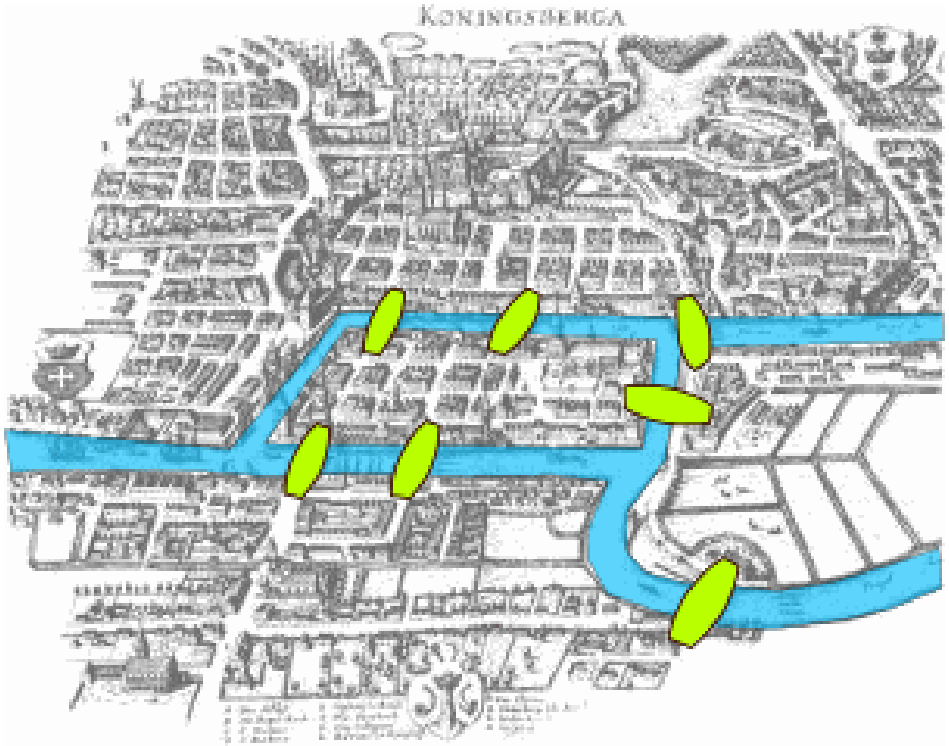
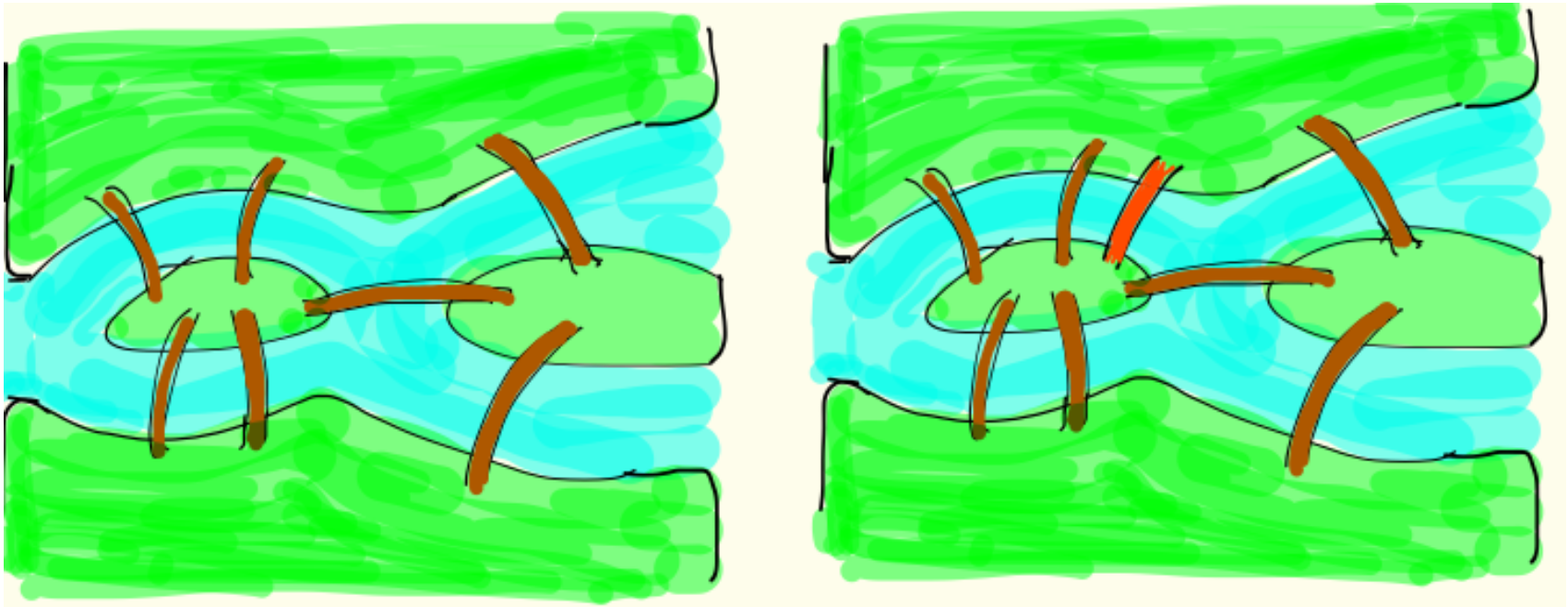


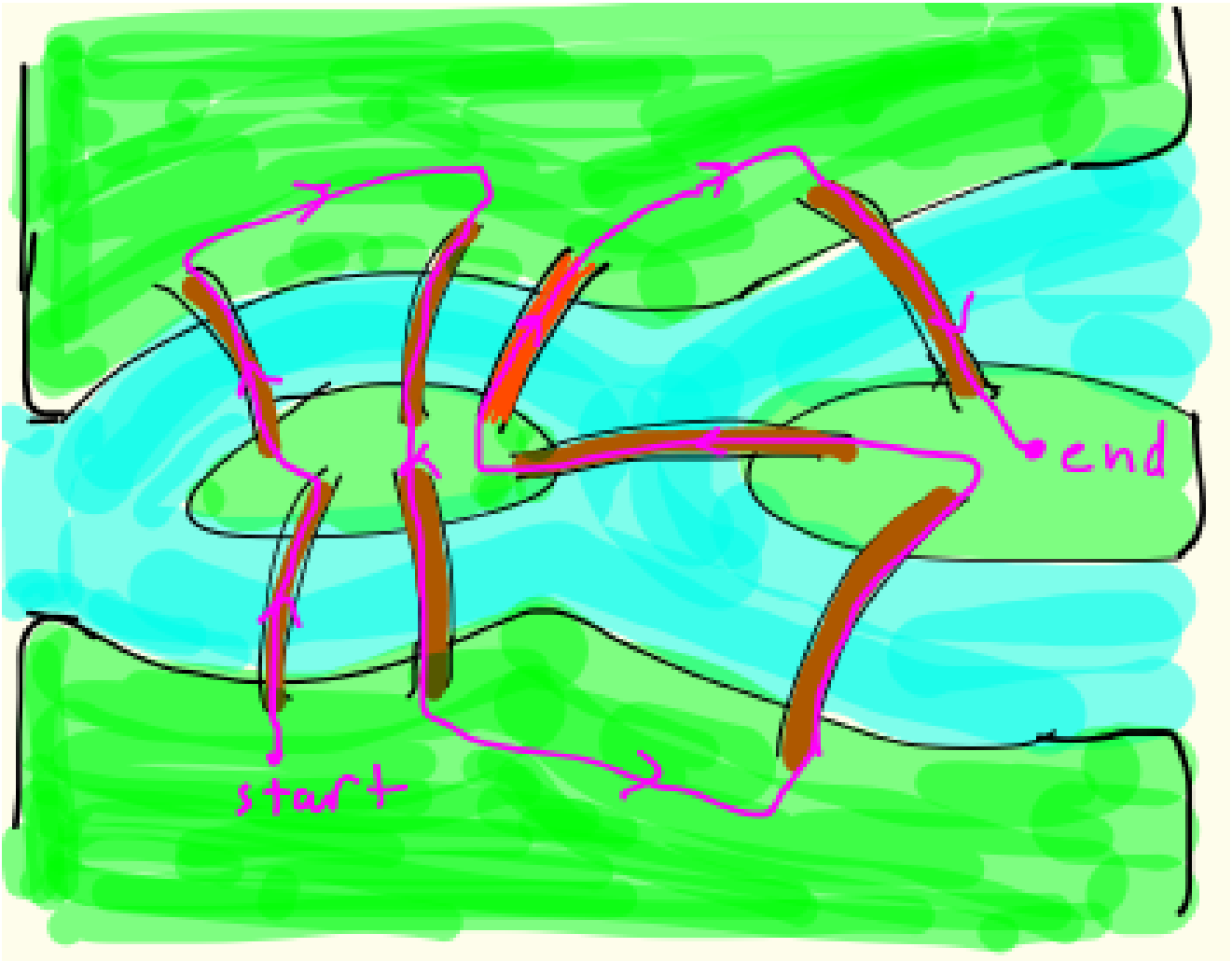
Image credit: Bogdan Giuc, Wikipedia

Konigsberg Bridge problem: Find a path that crosses each bridge exactly once. It doesn't have the start and end at the same place. Such a path is now called an *Euler path*.

Task: Try to find an Euler path in each of these two beautiful parks. Start with the park on the right first.



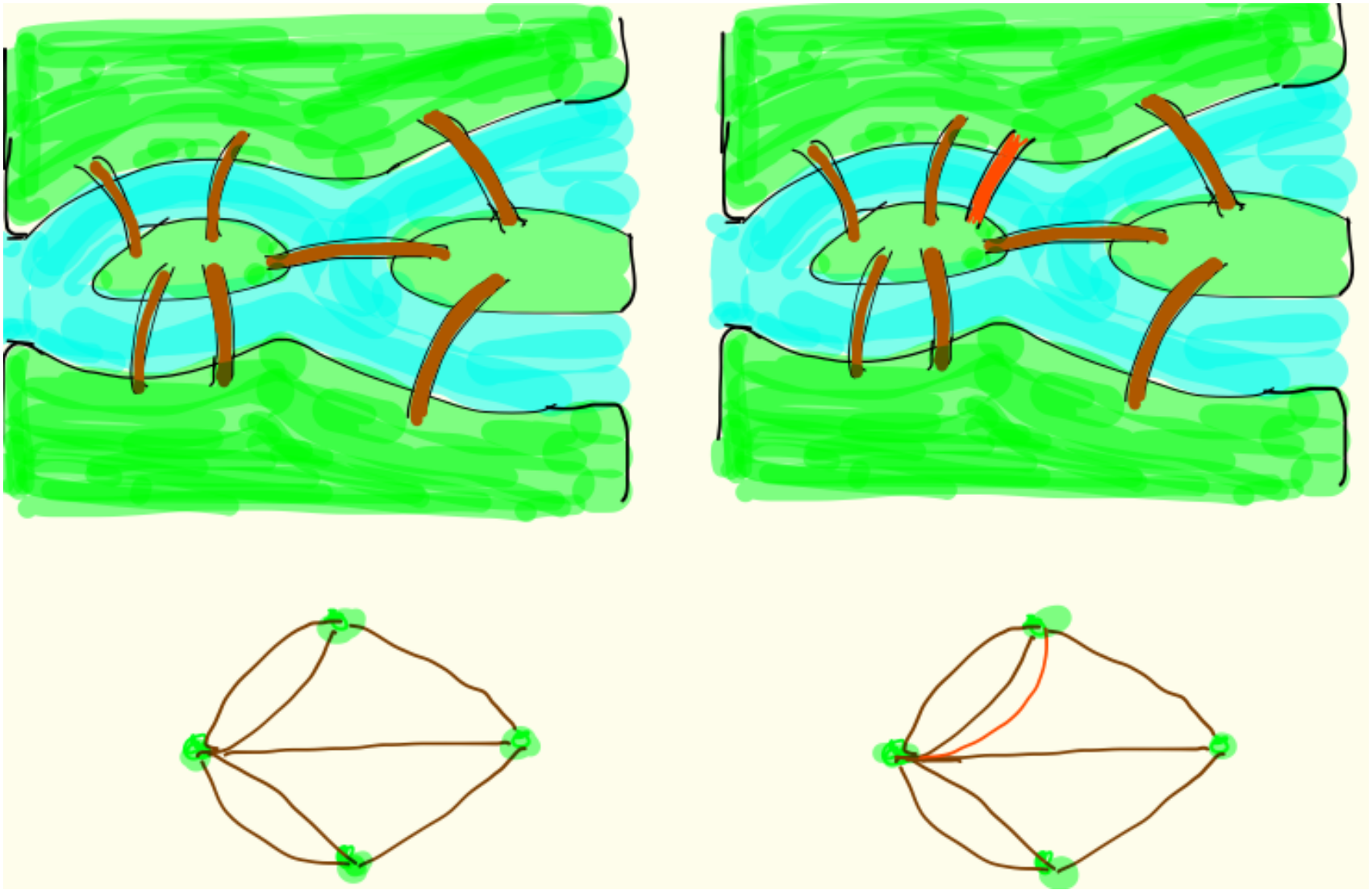
Here is one solution to the map that was on the right:



The map that was on the left has no solution. But why?

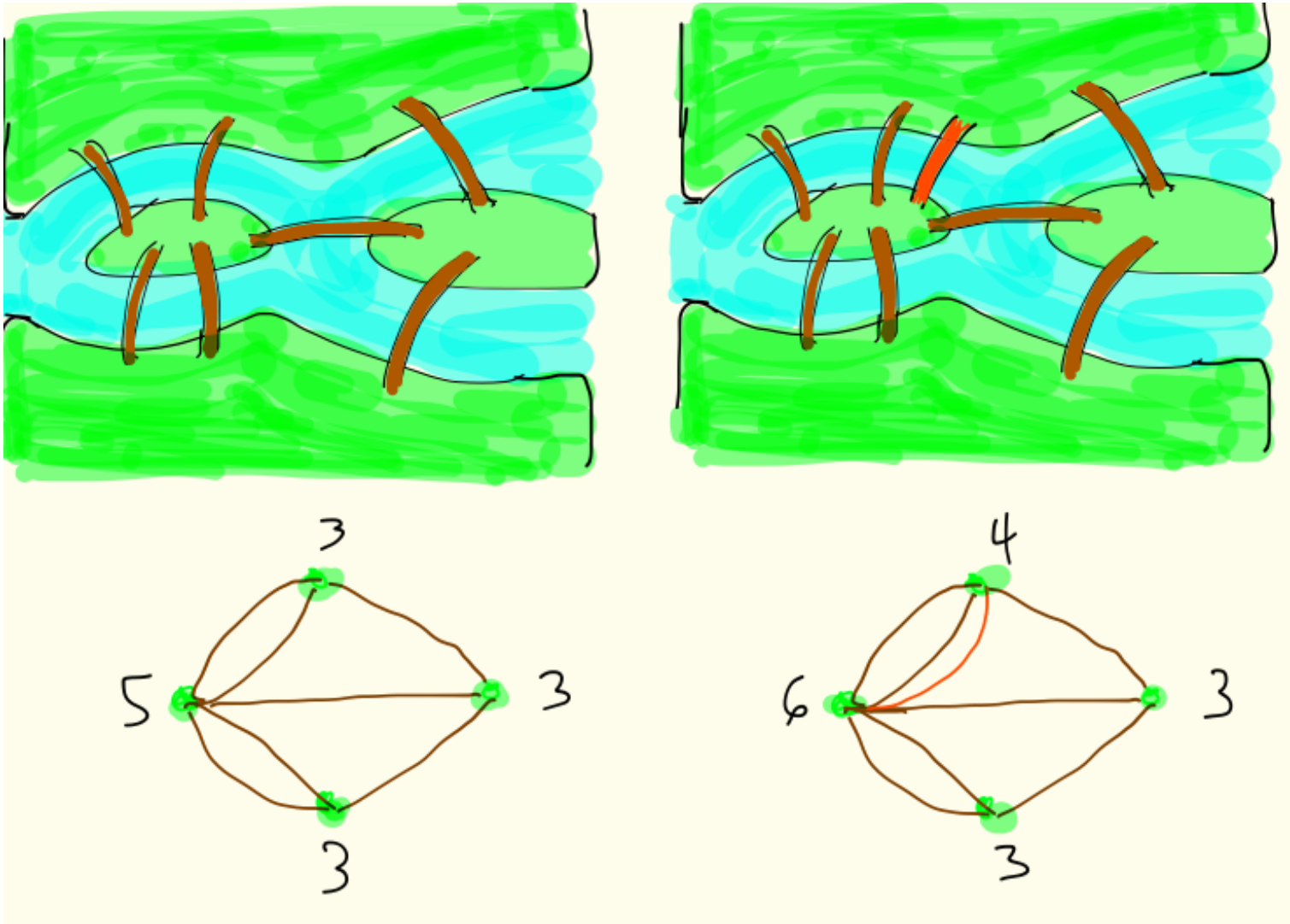
Euler's solution

Form a graph:



Euler's solution

Form a graph, label each vertex by its degree:



Euler's solution

Each time we enter and exit a vertex, we use two of the edges (basically).So for any vertex that is not the starting and ending point of an Euler path, we end up using an *even* number of edges. So since we want to use all edges, all vertices, except possibly the starting and ending points, must have even degree.

Theorem (Euler, 1736)

- *A graph has an Euler path if and only if at most two vertices have odd degree.*
- *A Bridge problem has an Euler path if and only if at most two of the land masses have an odd number of bridges.*
- *So for Konigsberg in particular, there is no such path.*

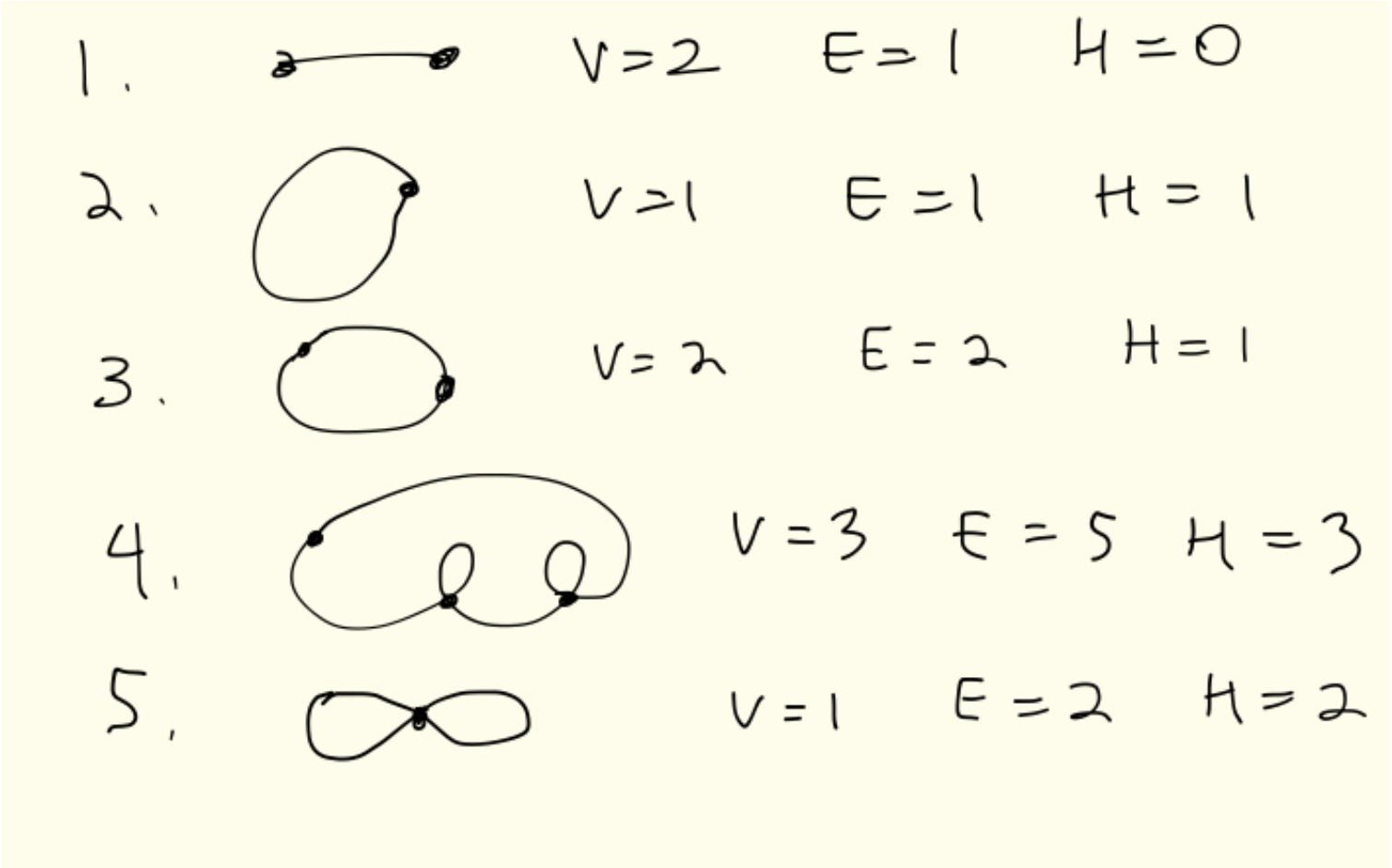
Key aspects of topology

Our three key aspects of topology, applied to Konigberg bridge problem:

1. deformation invariance: size and placement of land masses and bridges didn't matter
2. compressed representation: graphs or networks
3. discrete invariant (e.g number): degree of each vertex (number of bridges on each land mass)

Magic Trick

Draw any connected graph, tell me the number of vertices V and edges E, and I will tell you the number H of "holes" (or bounded regions) it has.



The magic formula

$$H = E - V + 1$$

This formula comes from a bigger story.

Platonic solids

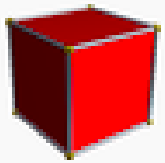
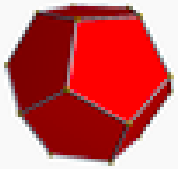
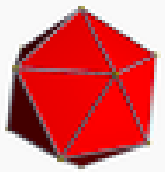
Name	Image	Vertices V	Edges E	Faces F
Tetrahedron		4	6	4
Hexahedron or cube		8	12	6
Octahedron		6	12	8
Dodecahedron		20	30	12
Icosahedron		12	30	20

Table credit: Wikipedia

Platonic solids: Euler characteristic

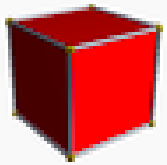
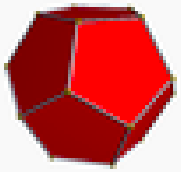
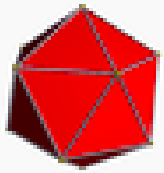
Name	Image	Vertices V	Edges E	Faces F	Euler characteristic: $V - E + F$
Tetrahedron		4	6	4	**2**
Hexahedron or cube		8	12	6	**2**
Octahedron		6	12	8	**2**
Dodecahedron		20	30	12	**2**
Icosahedron		12	30	20	**2**

Table credit: Wikipedia

Soccer ball, or truncated icosahedron

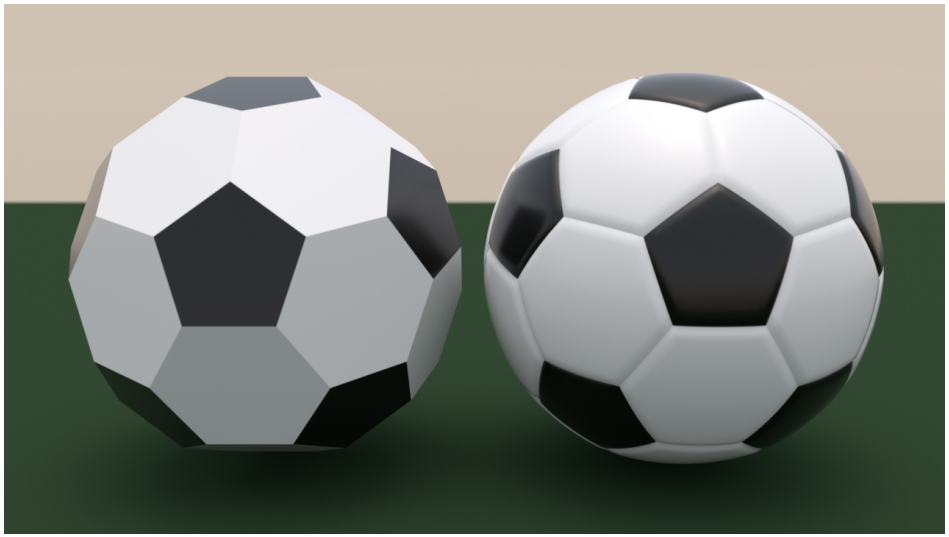


Photo credit: Aaron Rotenberg, Wikipedia

12 pentagonal faces, 20 hexagonal faces

$V = 60, E = 90, F = 12 + 20 = 32$

$V - E + F = 2$

Euler's Polyhedron Formula

Theorem (Euler's Polyhedron Formula)

For any decomposition of a spherical object into V vertices, E edges, and F faces, we have $V - E + F = 2$

Back to magic trick

How to apply Euler's polyhedron formula to our graph?
We need a sphere...
Trick: bend the paper into a sphere!
Then the unbounded region of the plane becomes a (big) face. Hence the total number of faces is one more than the number of holes in our graph:

$$\boxed{F = H + 1}$$

Plugging this into Euler's formula $V - E + F = 2$ we get

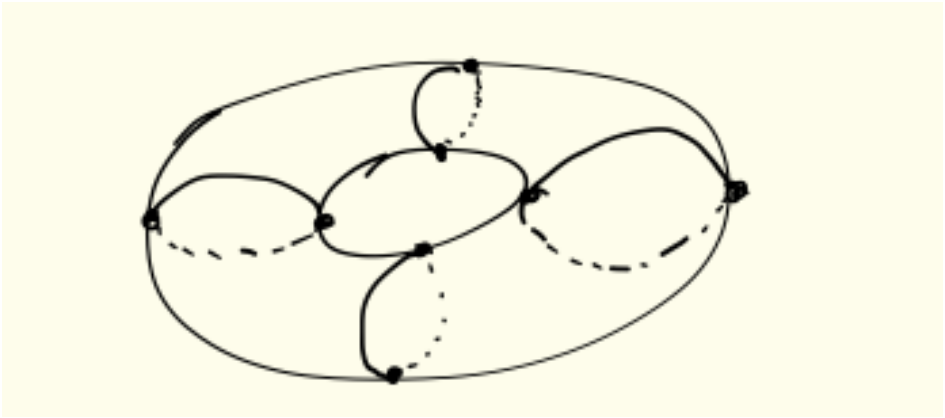
$$V - E + (H + 1) = 2$$

Solving for H gives

$$H = E - V + 1$$

Euler's formula for torus instead of a sphere

What if we replace the sphere with a torus?

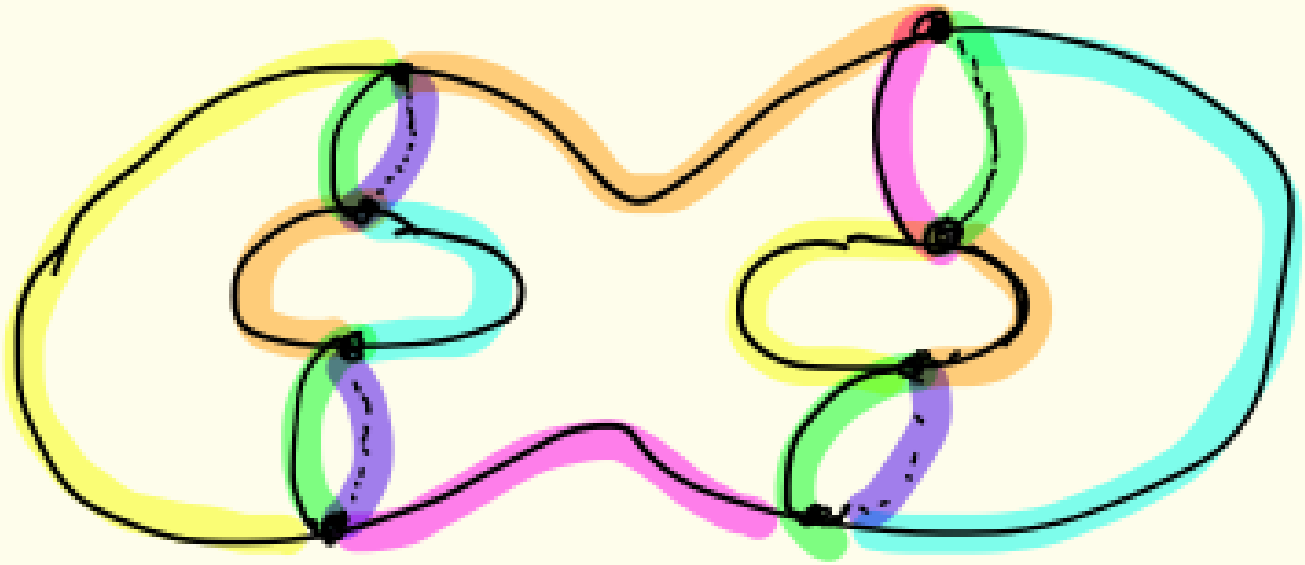


$V = 8, E = 16, F = 8$

$V - E + F = 0$

Euler's formula for surface with 2 holes

What if we replace the sphere with a surface wth 2-holes?



$V = 8, \quad E = 16, \quad F = 6$

$$V - E + F = -2$$

Euler's formula for a surface with g holes

Theorem

For a surface with g-holes,

$$V - E + F = 2 - 2g$$

The quantity $V - E + F$ is called the *Euler characteristic*. More generally, the Euler characteristic $\chi(X)$ of any shape X is defined to be the alternating sum of the Betti numbers

$$\chi(X) = b_0 - b_1 + b_2 - b_3 + \cdots$$

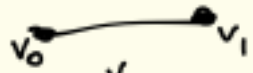
Recall our three key aspects of topology:

1. deformation invariance: the value of $V - E + F$ is the same for tetrahedron, soccer ball, or any triangulation of a spherical surface
2. compressed representation: we can represent a sphere by (the exterior of) a tetrahedron
3. discrete invariants: $V - E + F, g$

Simplices: A topologist's building blocks

0-simplex = point

1-simplex = edge



2-simplex = triangle

3-simplex = tetrahedron

n-simplex = n-dimensional tetrahedron

e.g $\{(x_0, \ldots, x_n) \in \mathbb{R}^{n+1}_{\geq 0} \mid x_0 + \ldots + x_n = 1\}$

Betti numbers

Any shape has Betti numbers, $b_0, b_1, b_2, \ldots,$. The number b_k is roughly a measure of the number of $k+1$-dimensional holes. The first few Betti numbers have the following definitions for 0-dimensional, 1-dimensional, and 2-dimensional simplicial complexes:

- b_0 is the number of connected components
- b_1 is the number of "circular" holes
- b_2 is the number of two-dimensional "voids" or "cavities"
- In higher dimension we lose the ability to visualize geometry, but $b_3, b_4, \cdots$ can be defined algebraically.

Example of Betti numbers

Betti numbers of a (solid) 3 dimensional ball D^3 (e.g dodgeball, but not a soccer ball) are $b_0 = 1, \quad b_1 = 0, \quad b_2 = 0.$

Betti numbers of the 2-sphere S^2 (e.g soccer ball or football, but not a dodgeball) are $b_0 = 1, \quad b_1 = 0, \quad b_2 = 1.$

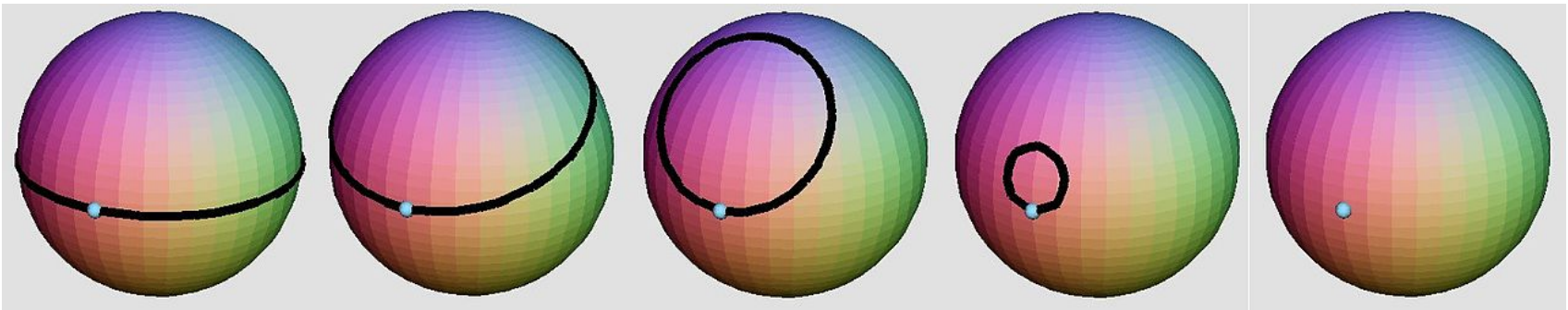


Image credit: Salix alba, Wikipedia

Example of Betti numbers

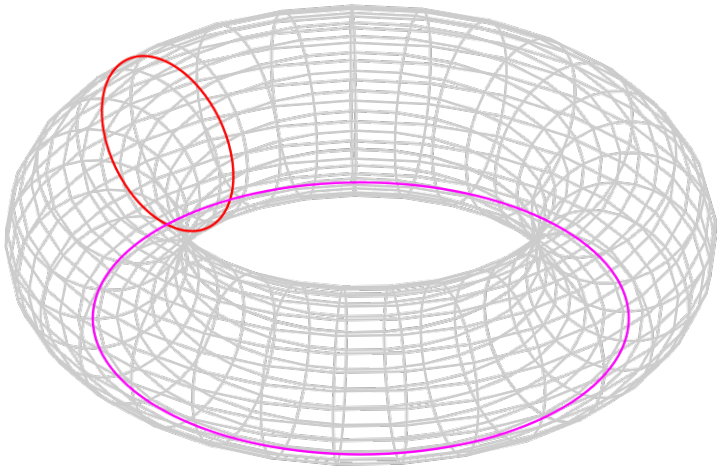


Image credit: Krishnavedala, Wikipedia

Betti numbers of a (hollow) torus are $b_0 = 1, b_1 = 2, b_2 = 1$.

Betti numbers of a solid torus (donut!) are $b_0 = 1, b_1 = 1, b_2 = 0$.

Betti numbers of the surface X_g with g-holes are
$b_0(X_g) = 1, b_1(X_g) = 2g, b_2(X_g) = 1$.

Emmy Noether (1882-1935)

Photo credit: Bryn Mawr College Archives, Wikipedia

- Emmy Noether was one of the leading mathematicians of her time. She developed the theories of rings, fields, and algebras (MA 5120 Abstract Algebra).
- One of her insights was to shift everyone's focus from Betti numbers $b_k(X)$ of a shape X to more complicated algebraic objects called the homology *groups* $H_k(X)$
- Betti numbers can be recovered as the size the homology groups:

$$b_k(X) = \text{rank } H_k(X)$$

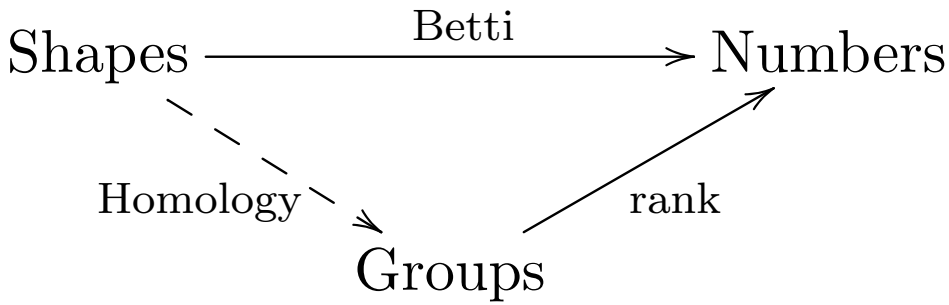
- Recall we said discrete invariants can be numbers, **groups, rings,** ...

Emmy Noether (1882-1935)

- In the spring of 1915, Noether was invited to the University of Gottingen by David Hilbert and Felix Klein.
- Their effort to recruit her, however, was blocked by some of the faculty. One faculty member protested: "What will our soldiers think when they return [from WW I] to the university and find that they are required to learn at the feet of a woman?"
- Hilbert responded with indignation, stating, "I do not see that the sex of the candidate is an argument against her admission as privatdozent [subordinate teaching duties]. After all, we are a university, not a bath house."
- During her first years teaching at Gottingen she did not have an official position and was not paid; her family paid for her room and board and supported her academic work. Her lectures often were advertised under Hilbert's name.

Emmy Noether (1882-1935)

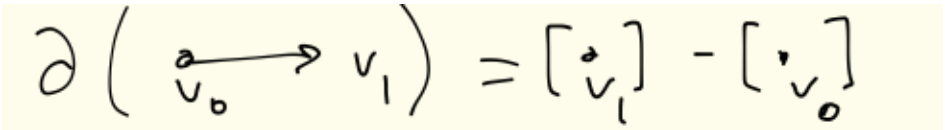
- 1933 Hitler, Nazi party, came to power. Noether, who was Jewish, was fired from her position.
- 1933 Noether moved to Bryn Mawr College (a women's liberal arts college near Philadelphia)
- Noether died in 1935 of cancer
- Noether provided invaluable methods of abstract conceptualization. Van der Waerden said that Noether's originality was"absolute beyond comparison."

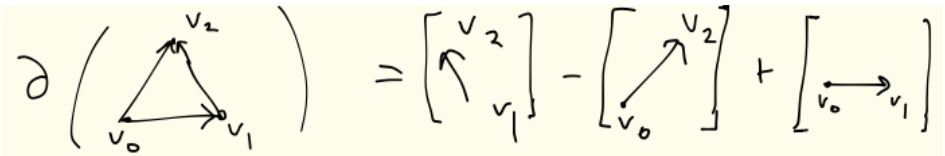


Simplicial Homology (1900s)

Let X be our shape, e.g a sphere. Divide it up into simplices in any way you like (deformation invariance, compressed representation).

- Let $C_k(X)$ denote the (free abelian group generated by) k-simplices.
- Define the boundary map (or function) $\partial_k : C_k(X) \to C_{k-1}(X)$ to be the map taking a k-simplex σ as input and giving the alternating sum of its boundary as output





Simplicial homology

- Putting these groups and maps together we get a chain complex

$$\cdots \to C_3(X) \xrightarrow{\partial_3} C_2(X) \xrightarrow{\partial_2} C_1(X) \xrightarrow{\partial_1} C_0(X) \to 0$$

- The simplicial homology group $H_k(X)$ are the "cycles modulo boundaries":
  - **cycles** are elements of $C_k(X)$ that ∂_k maps to 0
  - **boundaries** are elements of $C_k(X)$ that equal $\partial_{k+1}\sigma$ for some $\sigma \in C_{k+1}$

Simplicial homology

- Computing the simplicial homology groups $H_k(X)$ "by hand" can get complicated quickly, so mathematicians have found other equivalent definitions (cellular homology, singular homology).
- Mathematicians developed these other equivalent theories from 1900-1930
- But computing simplicial homology is relatively easy for a computer (foreshadowing).

Little blurb about my research interests

- Simplicial/singular/cellular homology has been extensively studied. But it turns out to be just one way of going from

$$\text{Shapes} \to \text{Groups}$$

- 1950s, 1960s, 1970s mathematicians started to find other sorts of (co)-homology theories:
  - K-theory
  - Cobordism theories
  - Morava K-theories $K_{n,p}$, one for every nonnegative integer n, and and prime number p

And now for something completely different

This talk is getting too technical, so let's change directions...
from the Math circa 1900 to Art circa 1900!

Sunday Afternoon on the Island of La Grande Jatte

Image credit: Georges Seurat - Art Institute of Chicago, Wikipedia

"Some say they see poetry in my painting, I see only science."

-Georges Seurat

Sunday Afternoon on the Island of La Grande Jatte

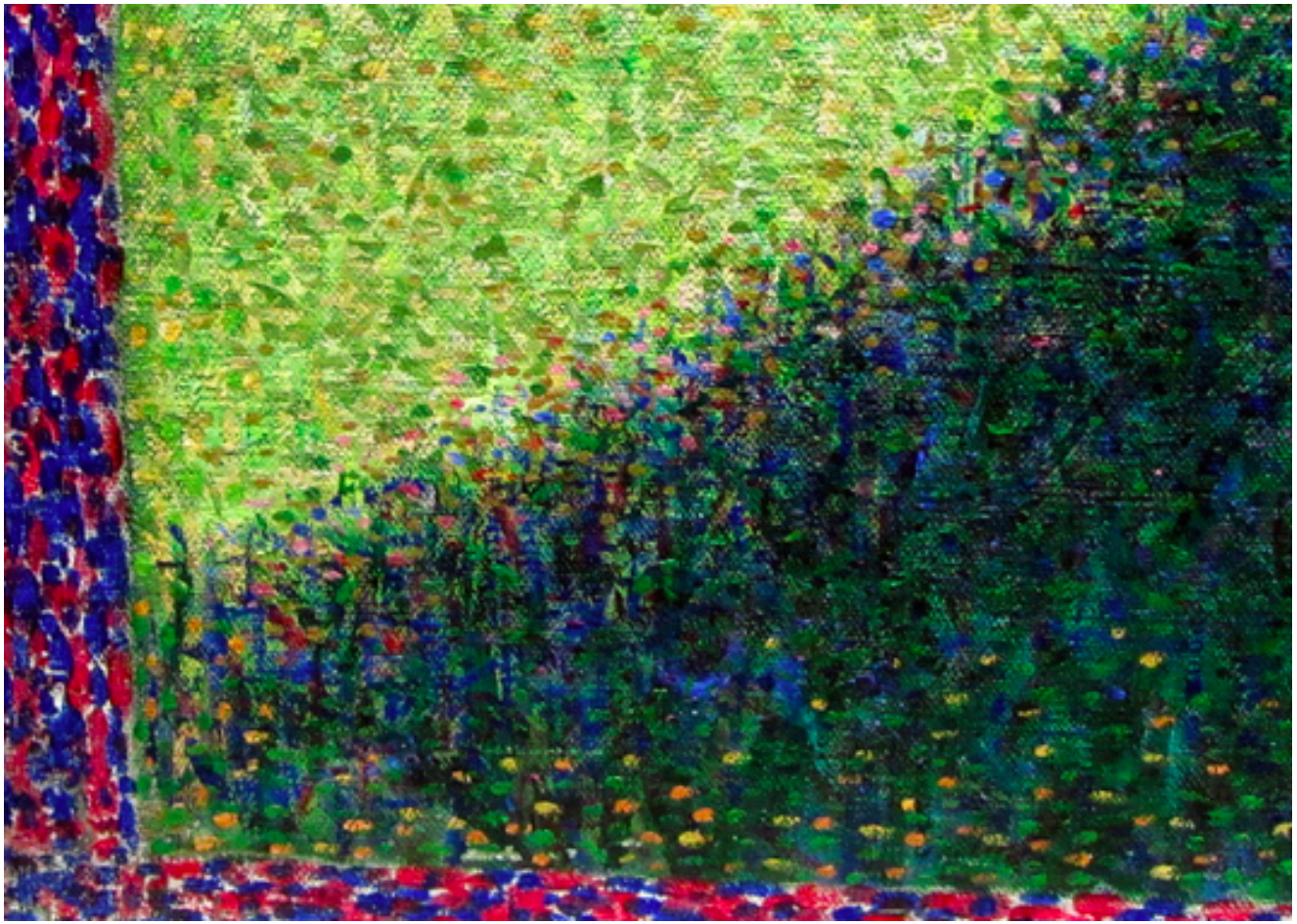


Photo credit: Jennifer Tharp, Flickr

Sunday Afternoon on the Island of La Grande Jatte

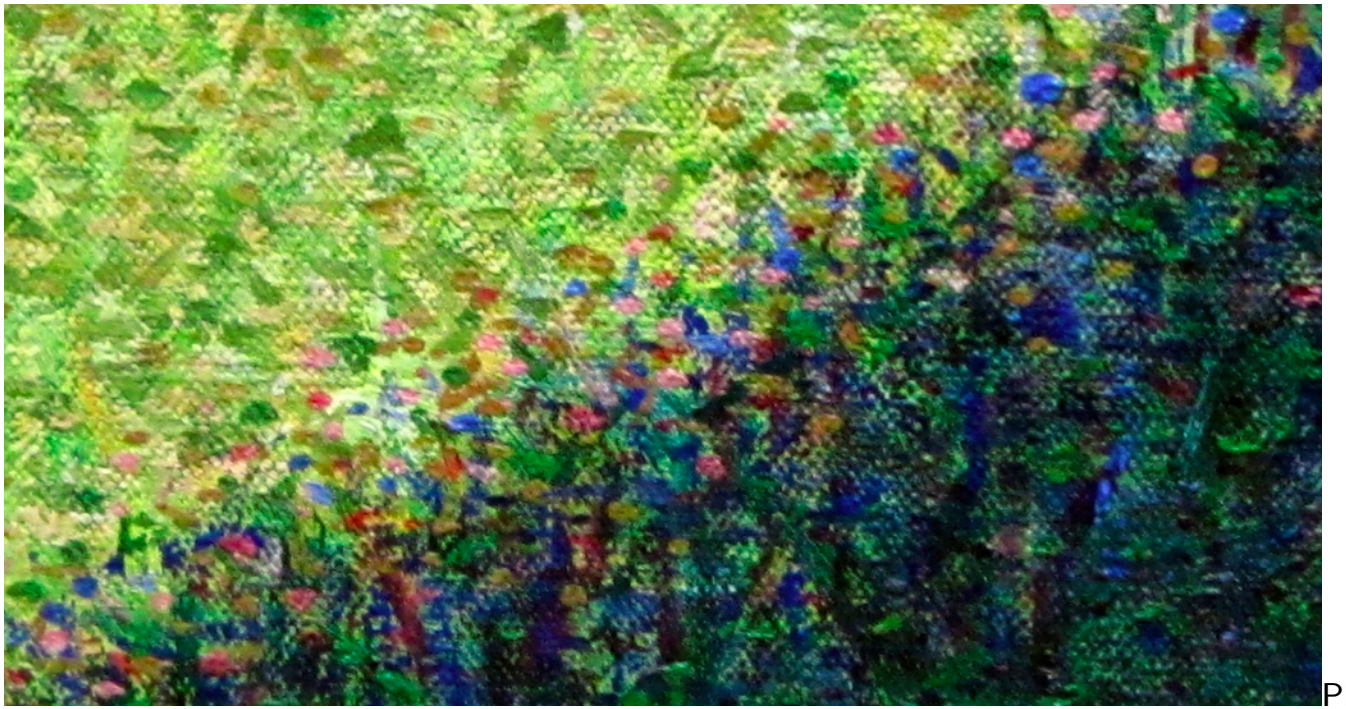


Photo credit: Jennifer Tharp, Flickr

Sunday Afternoon on the Island of La Grande Jatte

Photo credit: Tom S., Wikipedia

- Georges Seurat's most famous work, and is an example of pointillism.
- 7×10 feet in size, now exhibited in the Art Institute of Chicago
- Took two years to paint (1884-1886)
- Estimated to consist of approximately 3,456,000 dots (or so says a page on internet)

Your brain is a topological data analysis machine

Image credit: Georges Seurat - Art Institute of Chicago, Wikipedia

Your brain can:

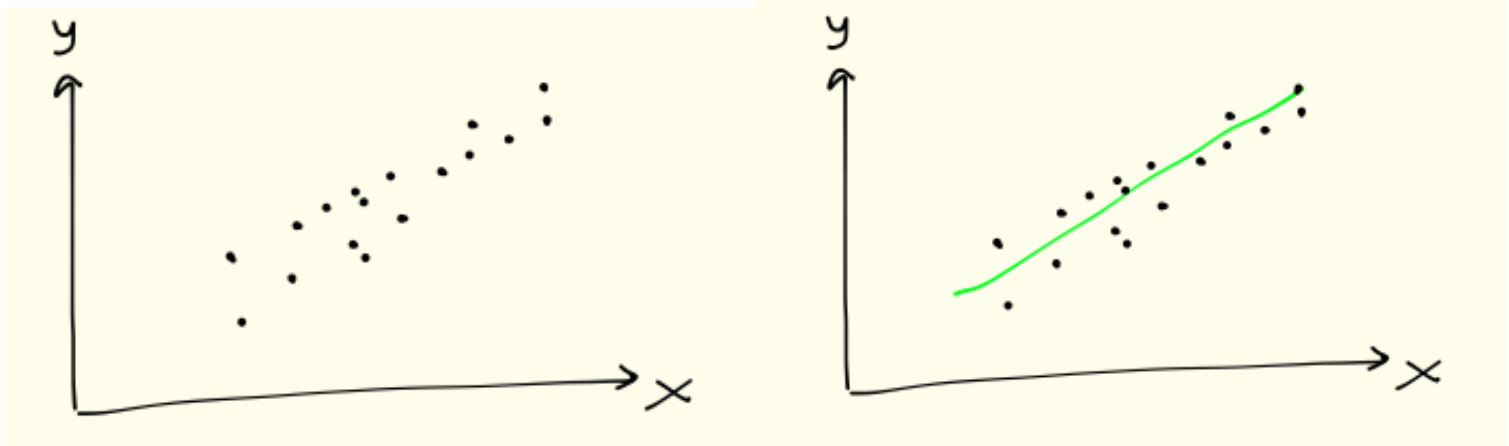
- extract shape from millions of dots (or much fewer than a million dots)
- extract meaning from the shape

What's the difference between the two?

Very big difference. The same as the difference between taking a reservation and *holding* a reservation. ▸ Seinfeld Clip

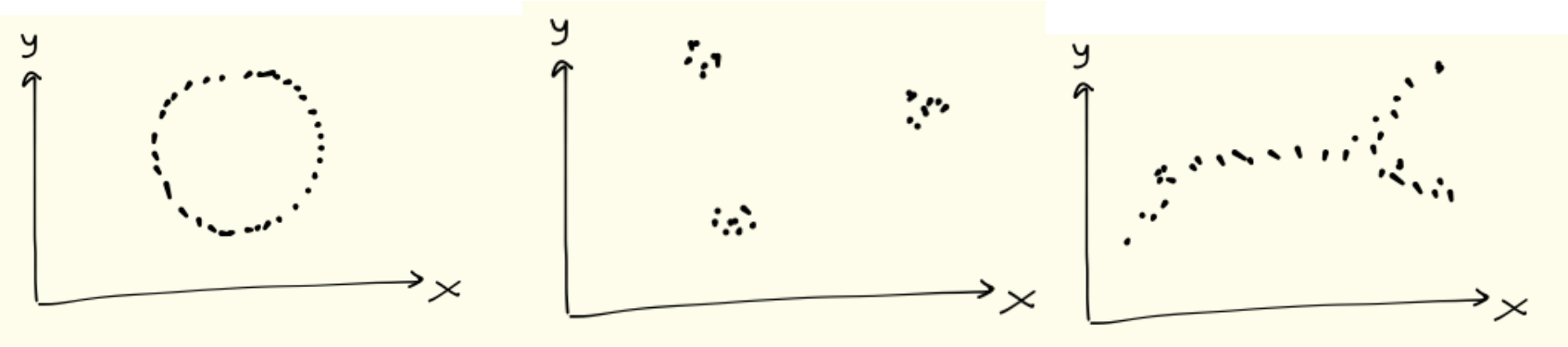
Data

- Studying data has become an extremely hot topic within the past decade:
- Data science, machine learning, data analytics, etc.
- Data can often be represented as points (in the xy-plane, or in $\mathbb{R}^n$ for some n)
- There are many tools to study data: Average, Linear regression, Principal Component Analysis (PCA), Clustering ...



Data is messy

- Data doesn't always form a straight line. It can take various shapes:



- One could try to develop or find a regression model for each type of shape.
- But that's not practical since there are an immense variety of possible shapes
- Instead, let's find a flexible way of dealing with all shapes. That's where topology comes in.

Topology is back

- New idea: use topology to construct a set of tools to find *shape* in data.
- Topological Data Analysis, used in the work Gunnar Carlsson (Stanford) and others 2000-present
- Persistent homology, popularized by Robert Ghrist (U Penn) 2000-present
- Carlsson co-founded a company, Ayasdi that applies these techniques to various business sectors: drug discovery, oil and gas exploration, and financial-market research
- **Caution:** Find meaning (if any) in the shape is a separate task, requiring domain expertise.

Data and Topology

Why study data using topology?

Aspects of Data:

- subject to noise (e.g experimental error)
- can be large (e.g millions of data points)
- we want to extract some information from the data

Aspects of Topology

- deformation invariance: not sensitive to minor variations
- compressed representation
- discrete invariants: "topological statistics"

Topological Data Analysis applied to breast cancer research

- In 2010, Gunnar Carlsson, et. al. [Lum] applied a Topological Data Analysis to two old breast cancer databases and within minutes discovered something new:
- "We identified a *previously unknown* subgroup of oncology *survivors* who exhibited genetic indicators of *poor survivors* [low ESR1 levels]. This will allow us to better understand this group and potentially help improve survival rates for this disease, which might potentially help us find a cure."

  -Devi Ramanan, Ayasdi head of collaboration [Ram]
- "These insights had eluded more traditional study for more than a decade. Using Topological Data Analysis, Ayasdi was to discover new insights within minutes." [Sym]

Breast cancer research results [Lum]

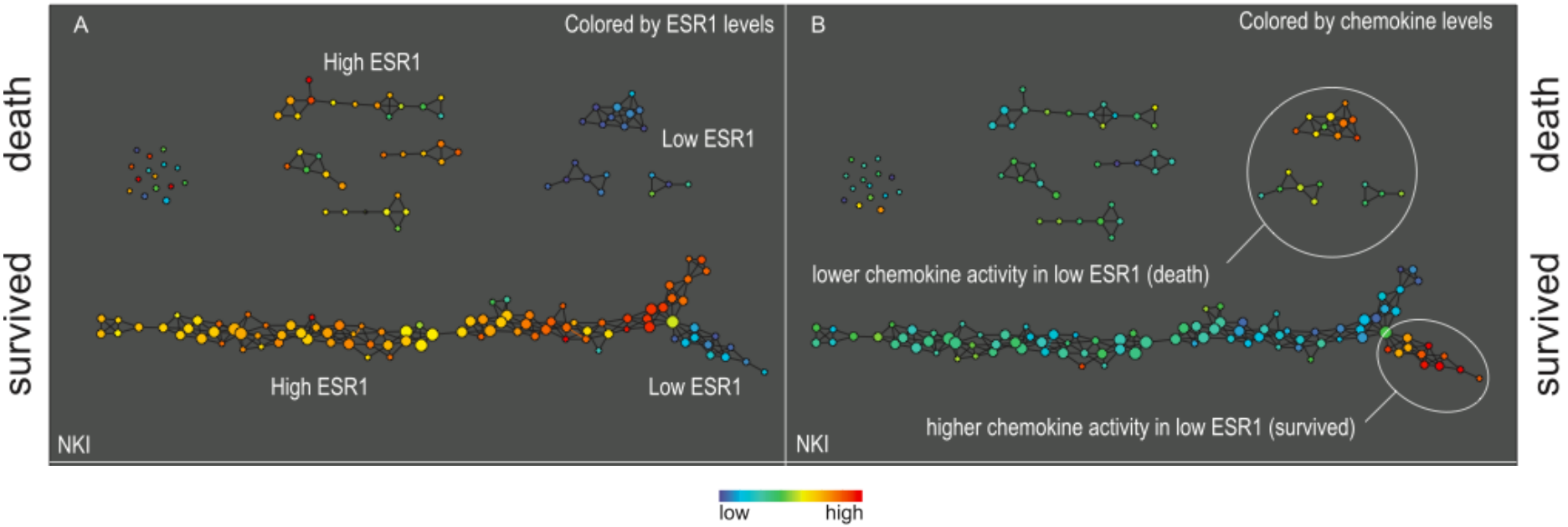


Image credit: [Lum]

Breast cancer research results [Lum]

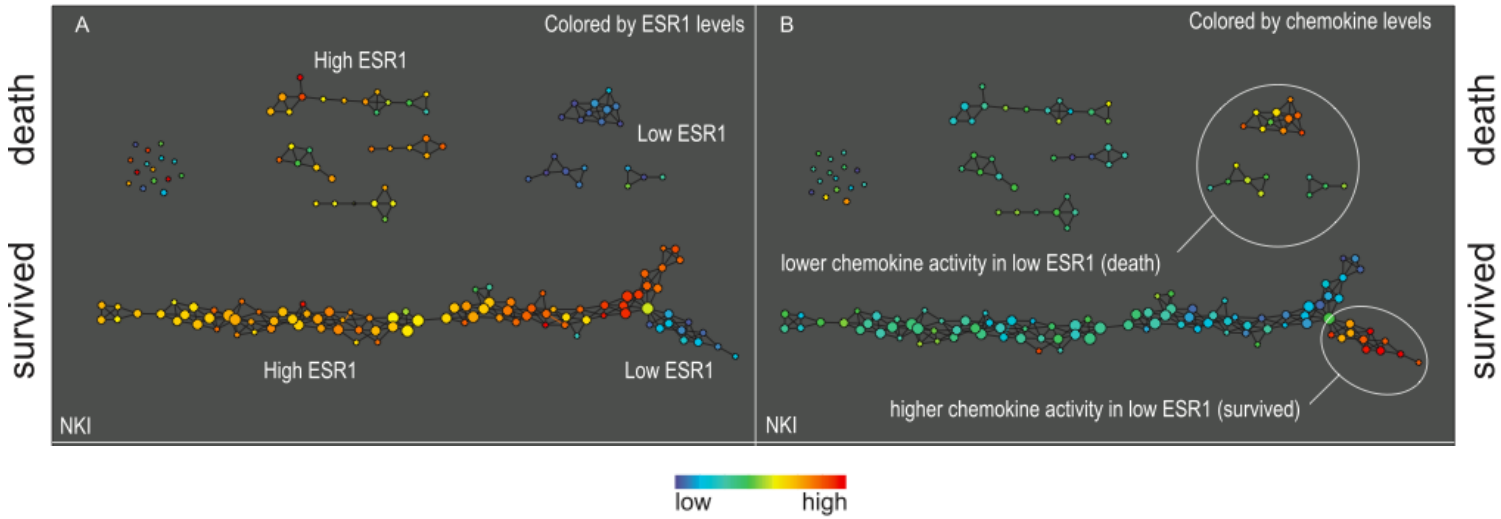
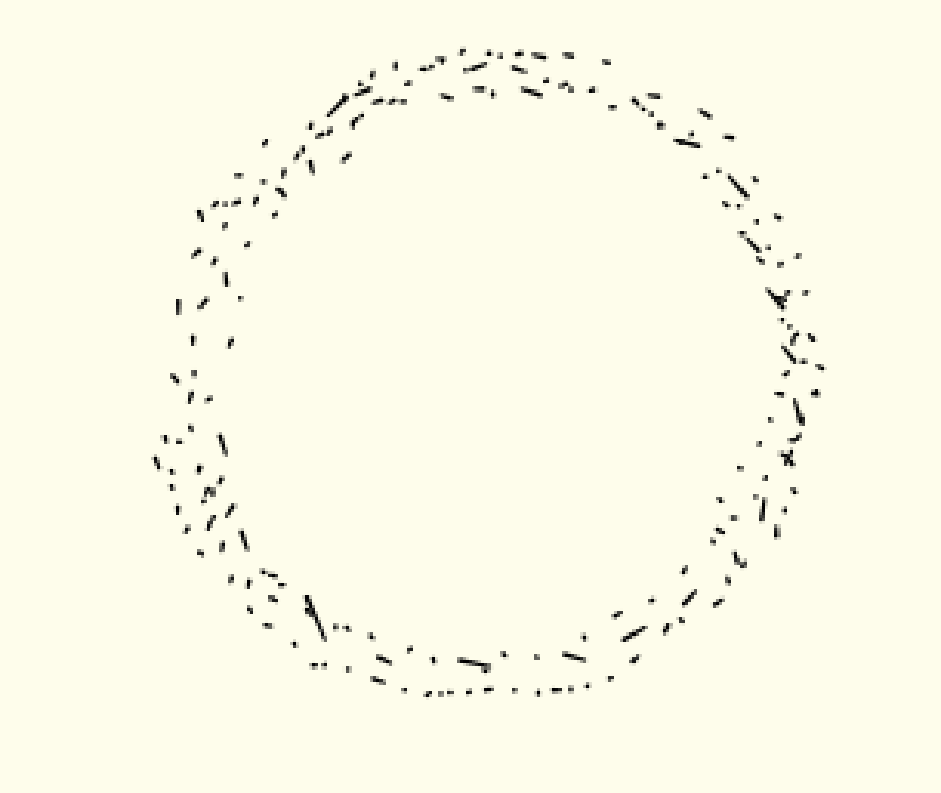


Image credit: [Lum]

- Raw data was gene expression levels of 1500 genes, in 272 tumors. So 272 points in $\mathbb{R}^{1500}$. Define a notion of distance between two points in $\mathbb{R}^{1500}$
- Use a filter function f to put tumors into *overlapping* bins/boxes/groups $f^{-1}(U_i)$
- Each node represents a cluster of tumors in a bin
- Nodes are connected if and only if they have at least one tumor in common

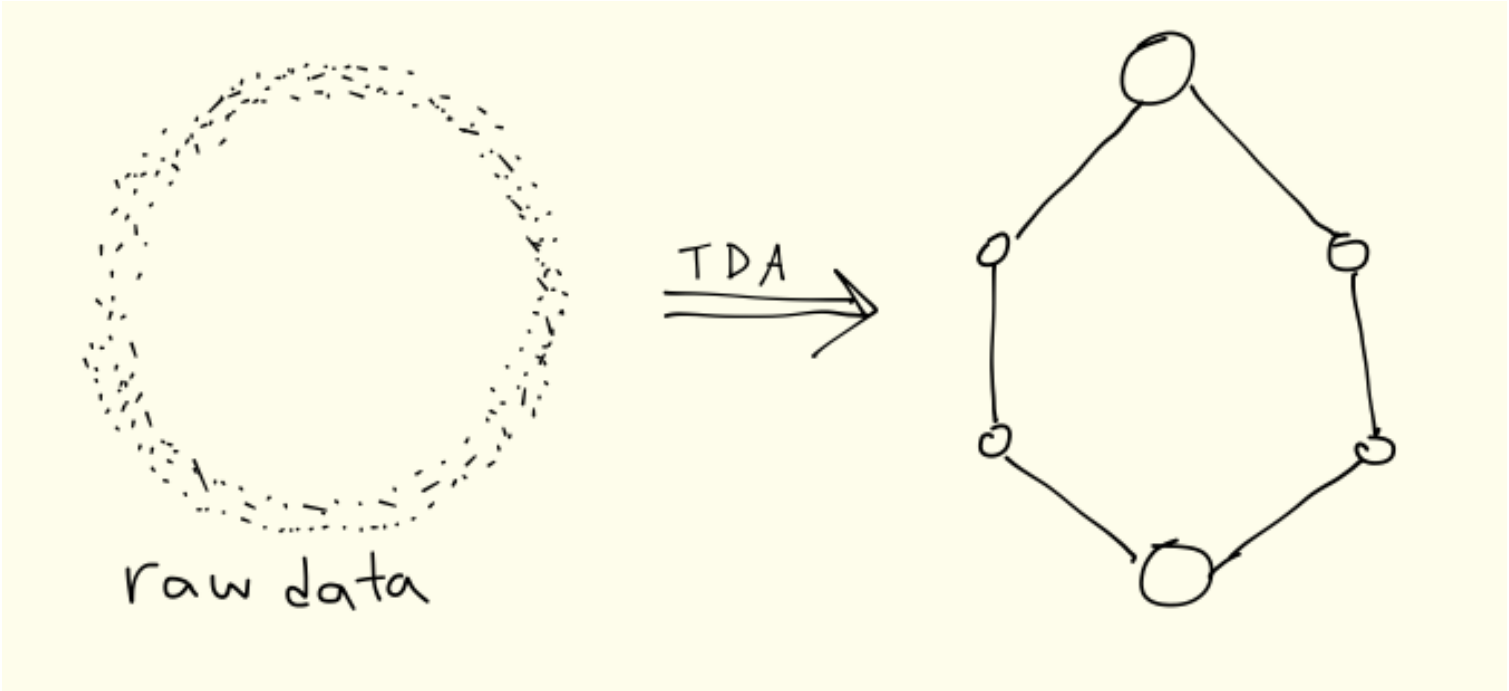
Topological Data Analysis Recipe, by toy example

Toy example presented in [Aya].
Step 0: Raw Data



Goal: Get a compressed representation (i.e. graph or network) that captures the shape

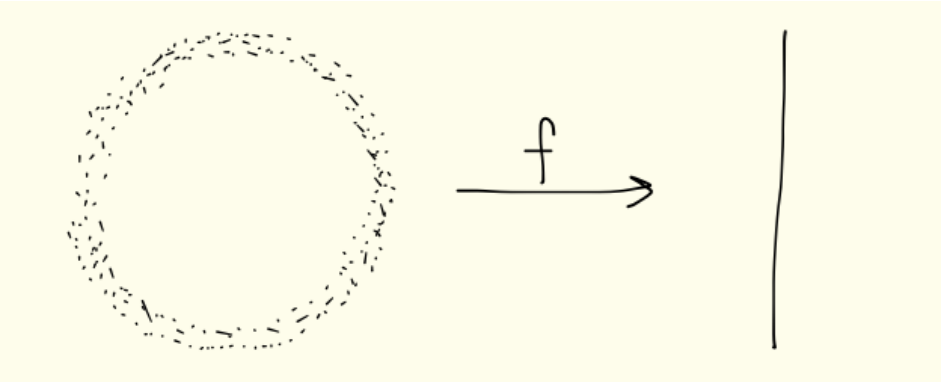
Topological Data Analysis Recipe, by toy example



Goal: Get a compressed representation (i.e. graph or network) that captures the shape

Topological Data Analysis Recipe

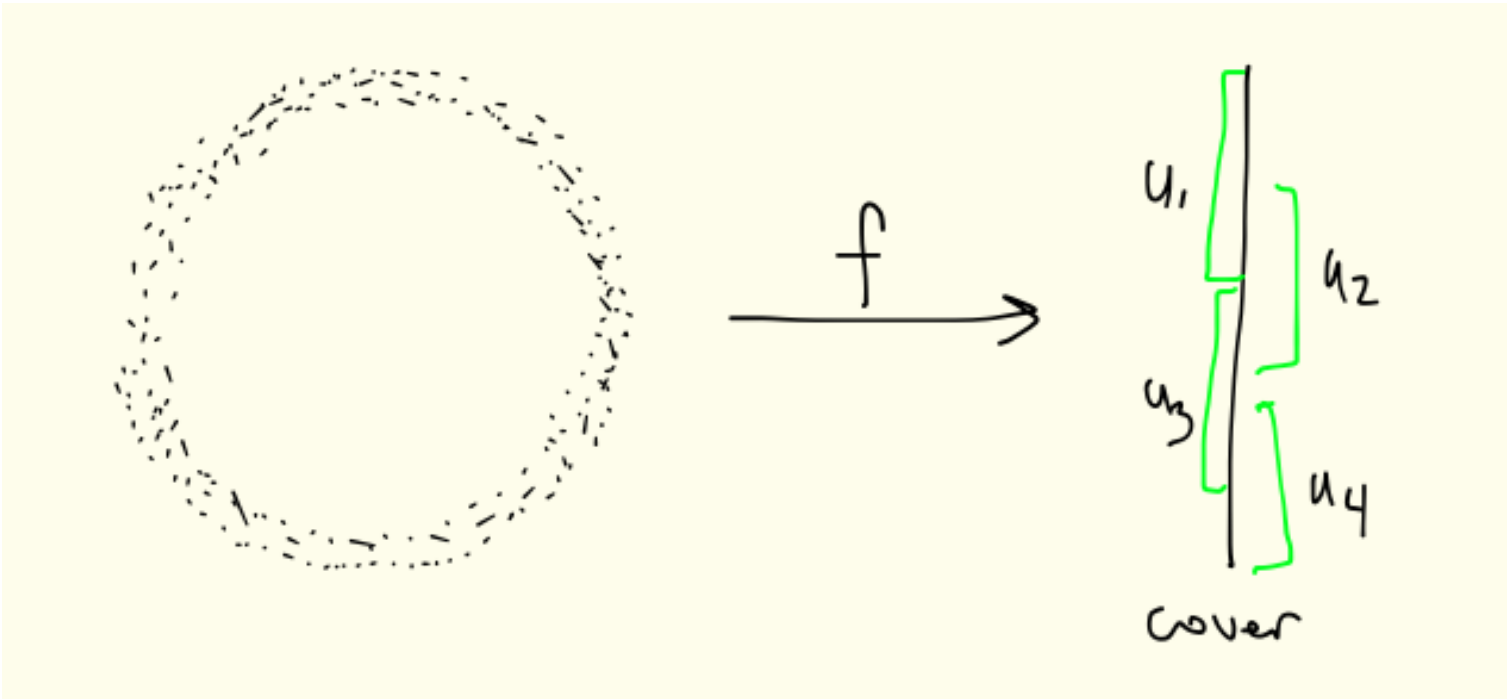
Step 1: Apply filter function f



In our toy example, f takes each data point and returns its y-coordinate. (In the breast cancer study, the filter functor took each data point and returned the distance to the furthest data point. This is called the L_∞-centrality function. The choice of filter function affects the output significantly.)

Topological Data Analysis Recipe

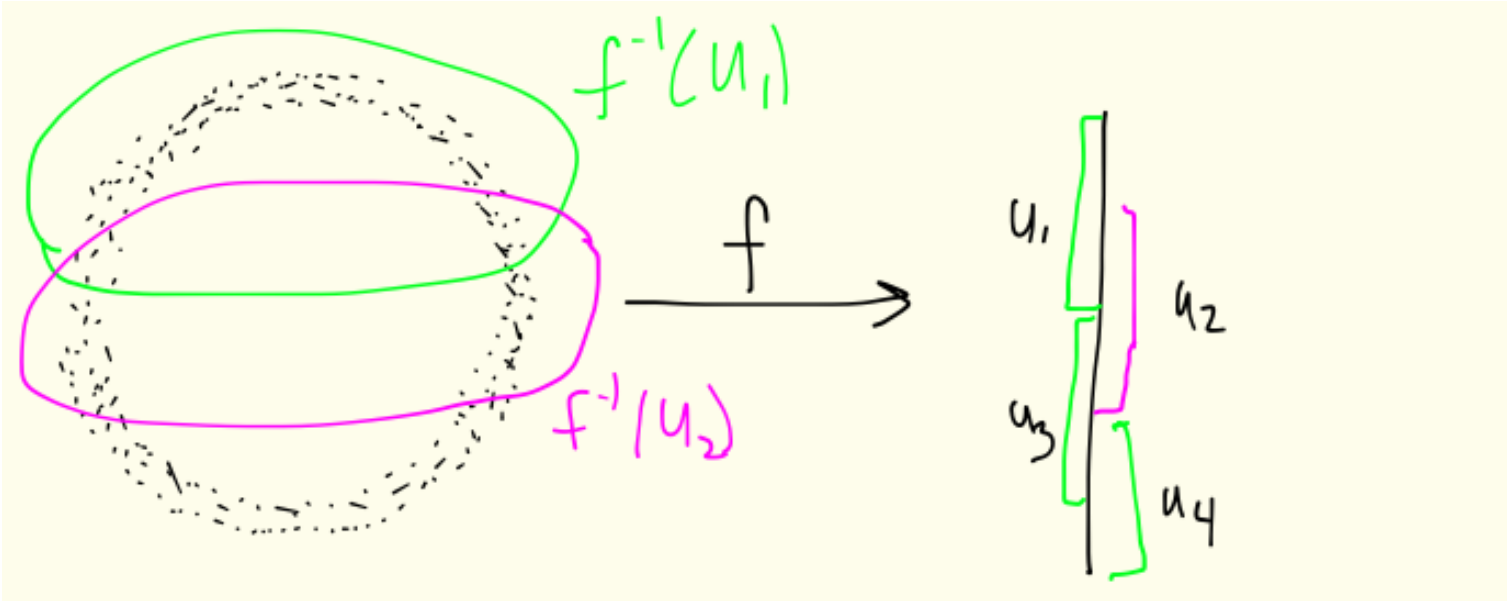
Step 2: Cover the target of the filter function by intervals.



The size and number of the intervals can be varied.

Topological Data Analysis Recipe

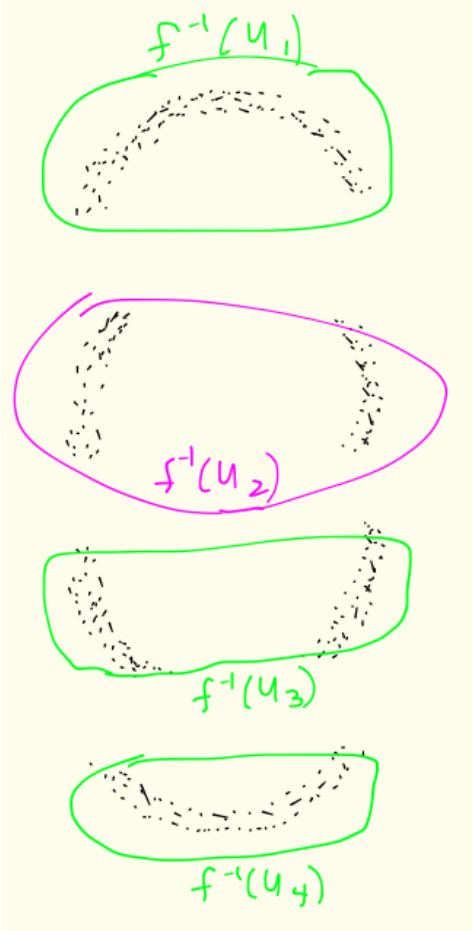
Step 3: Take the inverse image under f of each interval U_i, to create bins of data points.



There are four bins of data, only two are shown on this slide.

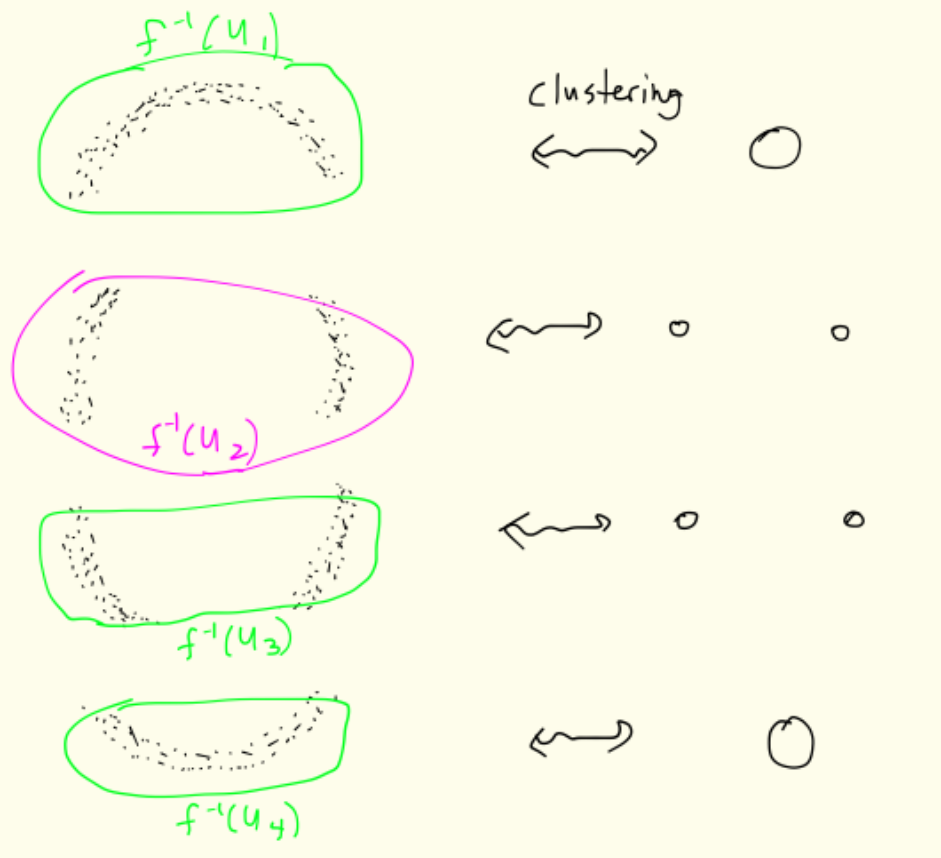
Topological Data Analysis Recipe

Step 3: Take the inverse image under f of each interval U_i, to create bins of data points. Here are all four bins.



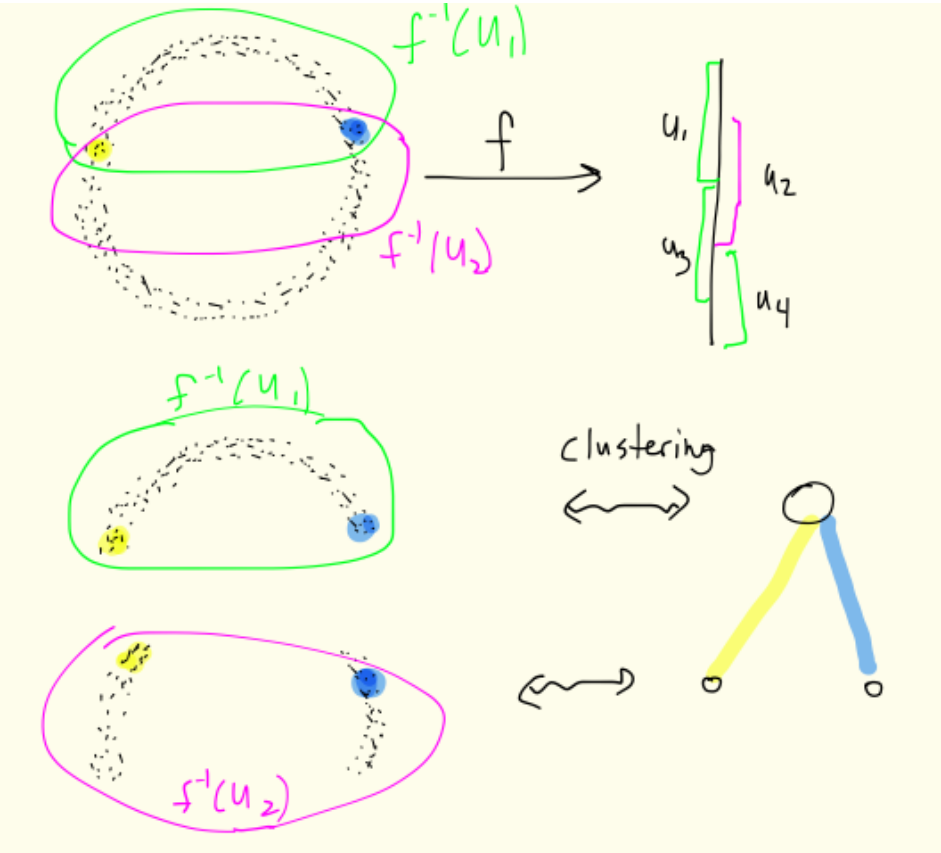
Topological Data Analysis Recipe

Step 4: Apply a clustering algorithm to each bin



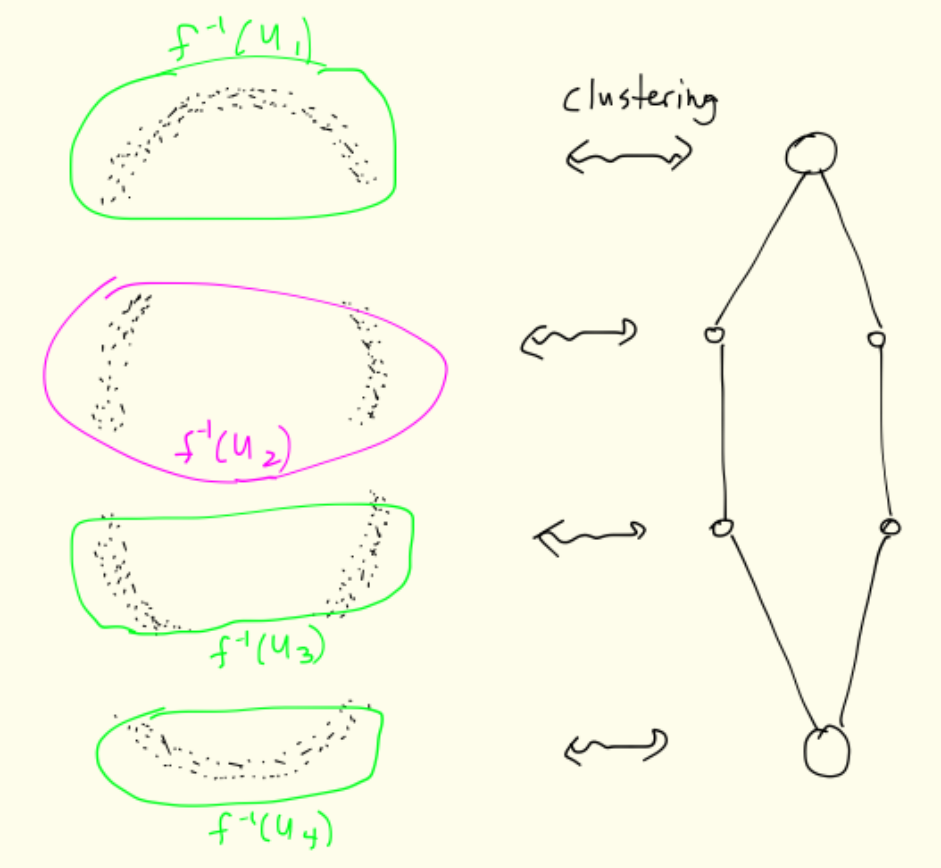
Topological Data Analysis Recipe

Step 5: Connect two nodes if they both represent a common data point



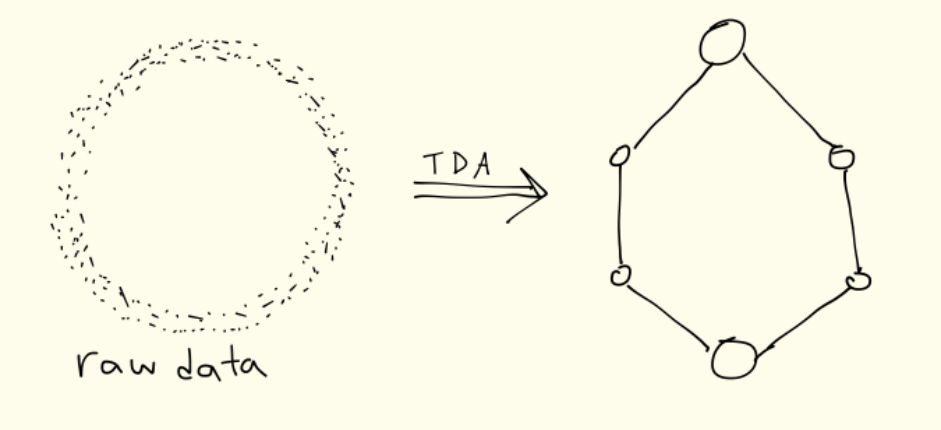
Topological Data Analysis Recipe

Step 5: Connect two nodes if they both represent a common data point



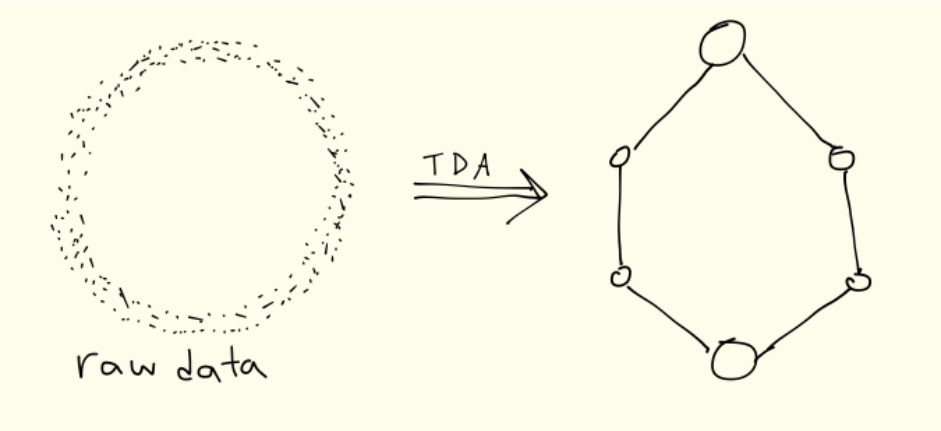
Topological Data Analysis Recipe

Step 6: Color the nodes via functions of interest



Topological Data Analysis Recipe

Step 7 (The hardest and most important): Find meaning!



- Check if your findings are consistent with the experience of domain experts
- Be willing to consider the possibility that TDA might not be give anything useful, in which case use other tools (persistent homology, other tools from statistics)

References

Aya Ayasdi. TDA and Machine Learning: Better Together.

GC1 Gunnar Carlsson. Why topological data analysis works http://www.ayasdi.com/blog/bigdata/why-topological-data-analysis-works/

Lum P.Y. Lum et al. Extracting insights from the shape of complex data using topology, Sci Rep 3:1236, (2013)

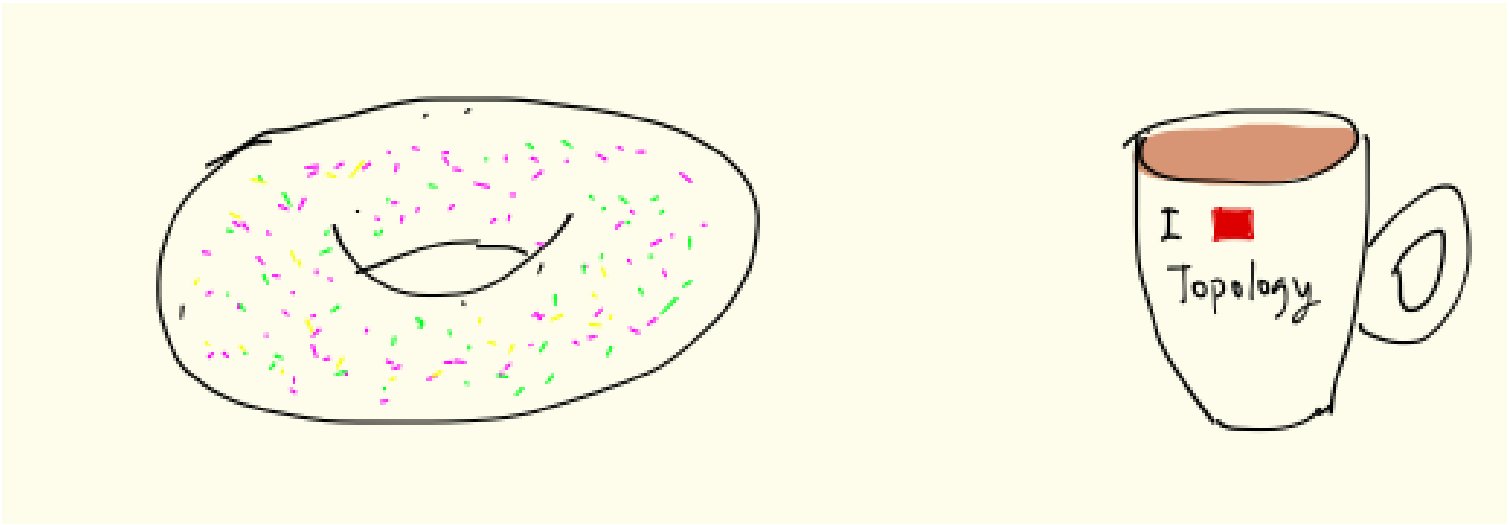
NYT New York Times, Jauary 16, 2013.

Ram http://www.ourdigitalmags.com/publication/?i=252002p=49

Sym Jonathan Symonds. http://www.ayasdi.com/blog/bigdata/big-data-brings-big-benefits-to-drug-discovery/

Wiki Wikipedia All photos

The End



Don't forget to hold the reservation!