The Data Scientist's Guide to

Topological Data Analysis

Justin Skycak

Industry advisor: David Cieslak, PhD, Aunalytics
Faculty advisor: Prof. Mark Behrens, Department of Mathematics

Purpose

- TDA = Topological Data Analysis
  - Data analytic methods → useful to data scientists
  - Inspired by abstract math → esoteric terminology → hard to understand

Purpose

- TDA = Topological Data Analysis
    - Data analytic methods → useful to data scientists
    - Inspired by abstract math → esoteric terminology → hard to understand

- Goal - bridge communication gap between academia & industry
    - 1. Mapper - TDA method currently in industry
    - 2. Persistent homology - TDA method academia that may later break into industry

Purpose

- TDA = Topological Data Analysis
  - Data analytic methods → useful to data scientists
  - Inspired by abstract math → esoteric terminology → hard to understand

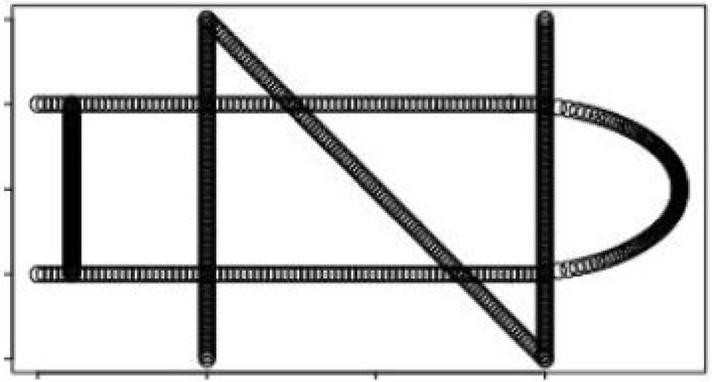
- Goal - bridge communication gap between academia & industry
  - 1. Mapper - TDA method currently in industry
  - 2. Persistent homology - TDA method academia that may later break into industry

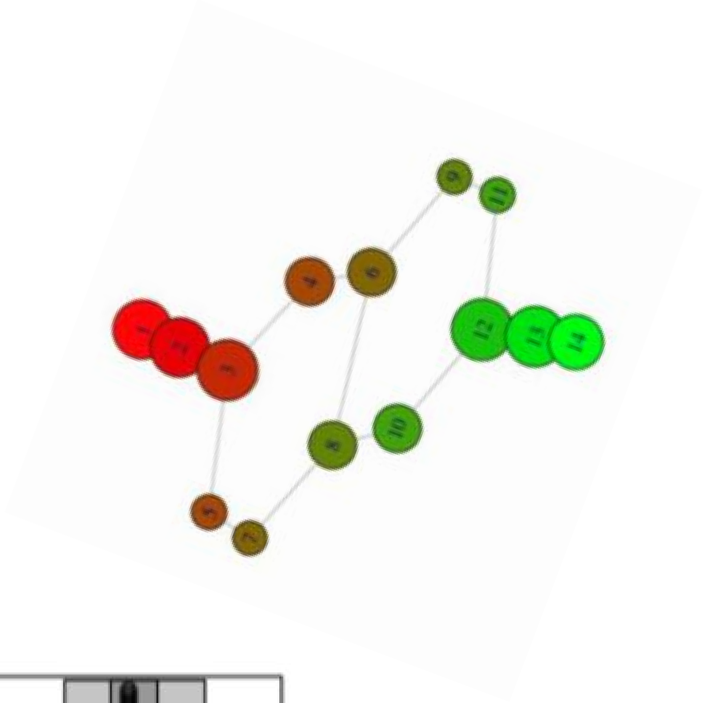
> “It is hoped that the data scientist reading this guide will be inspired to give Mapper a try in their future analytic work, and be on the lookout for future developments in persistent homology that push it from academia to industry.”

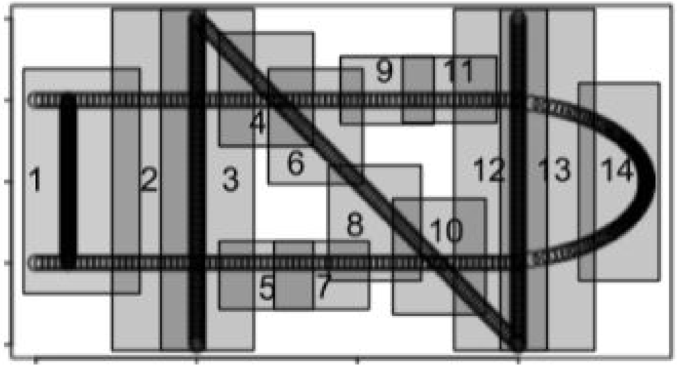
Mapper simplifies data into network

- High dimensional data → 2D network that represents overall shape of data

R package: TDAmapper





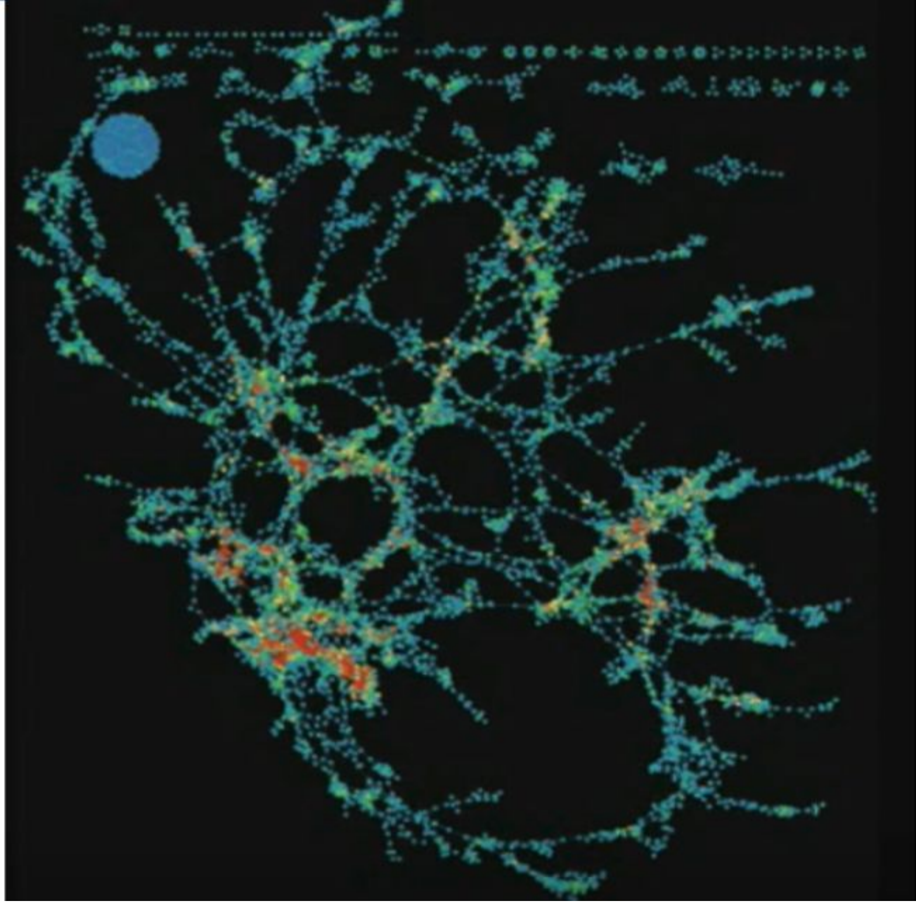


Real Use Case: Forecasting Returns

300+ market and economic variables, sampled over 25 years

Nodes colored by year

Colors are spread out → indicates repeated patterns over time



Roche, Terry, Tim Grant, Patrick Rogers, and Mukund Ramachandran. "Predicting the Future: Forecasting Returns using Machine Intelligence." *Ayasdi Resources*. 2015.

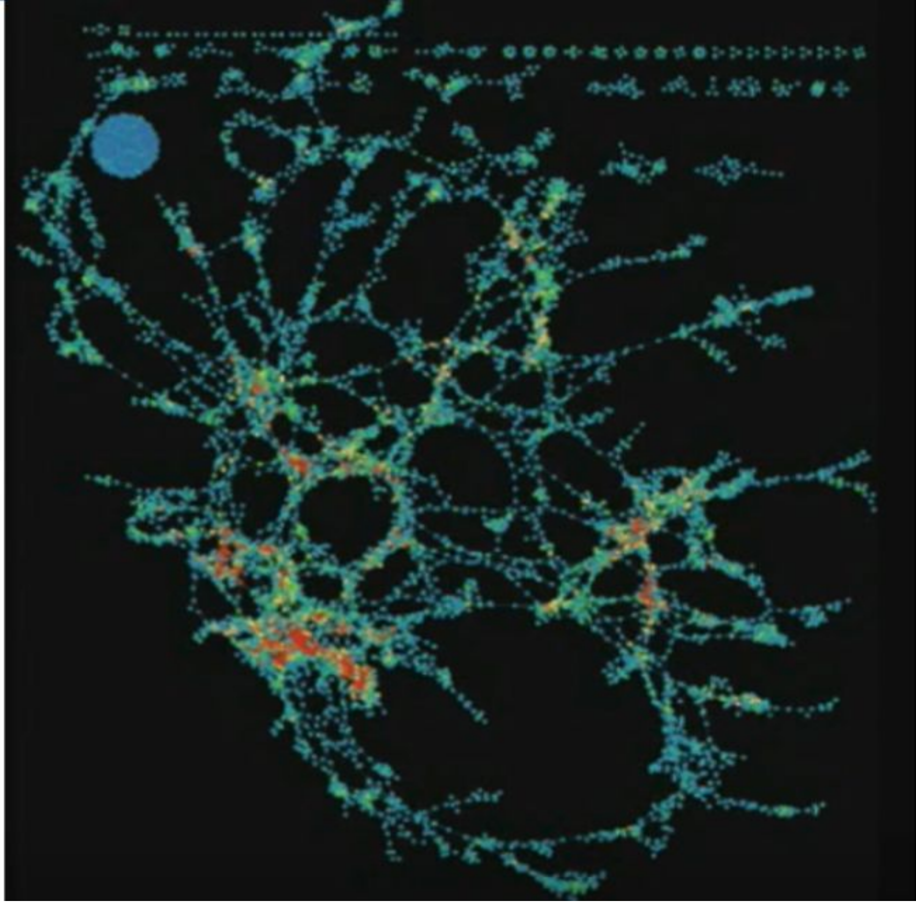
Real Use Case: Forecasting Returns

300+ market and economic variables, sampled over 25 years

Nodes colored by year

Colors are spread out → indicates repeated patterns over time

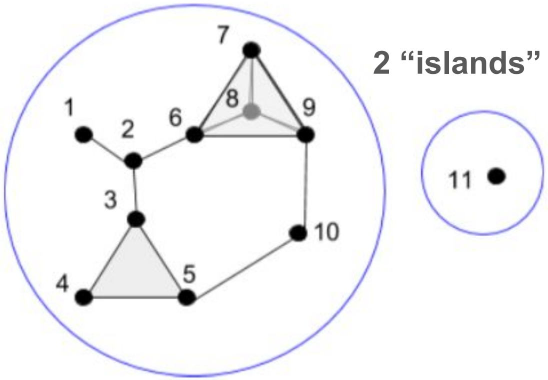
strategy to forecast from an initial date:
1. locate neighboring dates on the map
2. use their price trajectories to build a distribution of changes in price for each asset
3. use mean or median for predictions



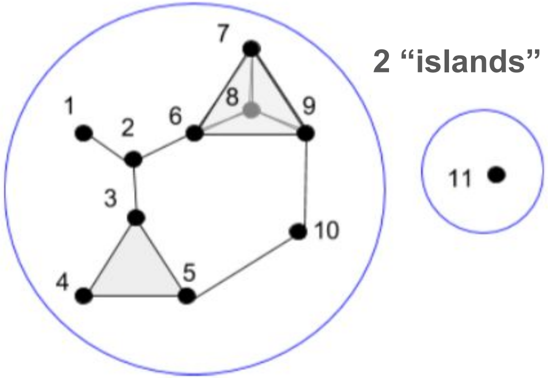
Roche, Terry, Tim Grant, Patrick Rogers, and Mukund Ramachandran. "Predicting the Future: Forecasting Returns using Machine Intelligence." *Ayasdi Resources*. 2015.

Homology counts “loops” in network

0th homology: points which cannot be shifted to each other along an edge



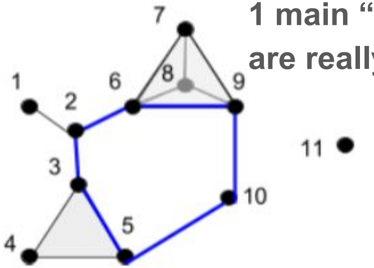
Homology counts “loops” in network

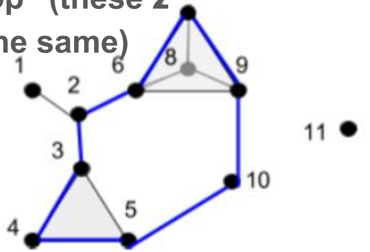


2 “islands”

0th homology: points which cannot be shifted to each other along an edge

1 main “loop” (these 2 are really the same)

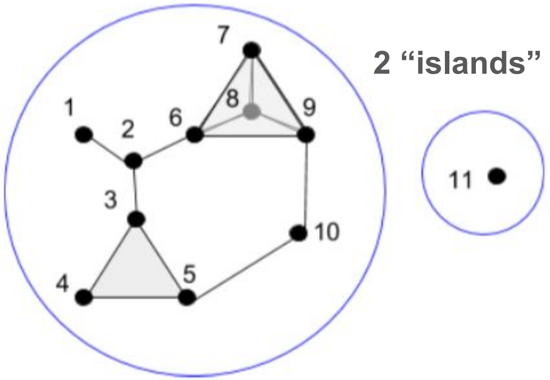




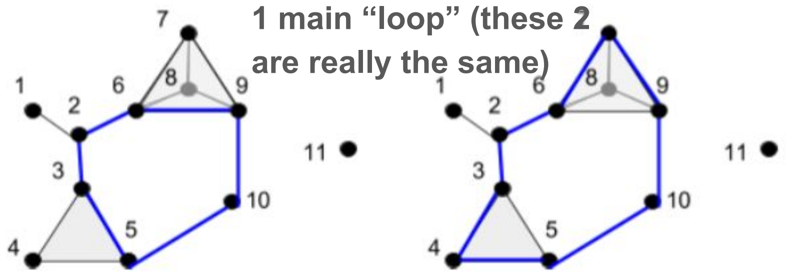
1st homology: edge loops which cannot be shifted to each other along a surface

Homology counts “loops” in network

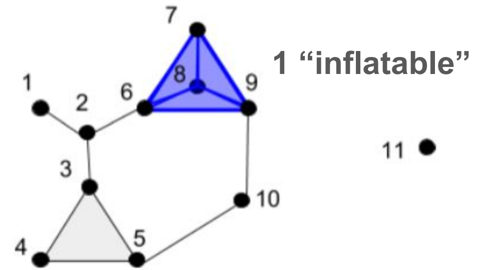
0th homology: points which cannot be shifted to each other along an edge



1st homology: edge loops which cannot be shifted to each other along a surface



2nd homology: closed surfaces which cannot be stretched into one another along solid tetrahedrons



Persistent Homology counts “loops” across scales

- To convert cloud of data points to network, you connect points that are “close enough”
  - Scale = choice of “close enough”
  - Depending on choice scale, network can be densely connected or sparsely connected (or in between)

Persistent Homology counts “loops” across scales

- To convert cloud of data points to network, you connect points that are “close enough”
  - Scale = choice of “close enough”
  - Depending on choice scale, network can be densely connected or sparsely connected (or in between)

- Q: What scale should you use when you compute homology of data?
  - A: Look at all scales
  - Make “barcode graph”

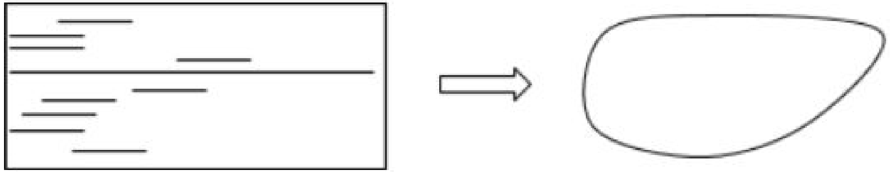
Persistent Homology counts “loops” across scales

- To convert cloud of data points to network, you connect points that are “close enough”
  - Scale = choice of “close enough”
  - Depending on choice scale, network can be densely connected or sparsely connected (or in between)

- Q: What scale should you use when you compute homology of data?
  - A: Look at all scales
  - Make “barcode graph”

1 component in first homology
means data has a “main loop”

First homology
components



Small scale . . . large scale

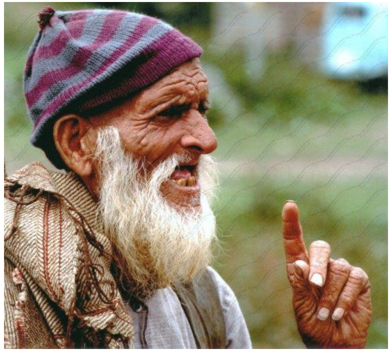
Betti numbers

- We can represent any space as a point, where nth component counts number of components in nth homology (aka nth Betti number)
  - E.g. 2 components in 0th homology, 1 component in 1st homology, 1 component in 2nd homology, 0 components in all following homologies → (2, 1, 1, 0, 0, …)

Betti numbers

- We can represent any space as a point, where nth component counts number of components in nth homology (aka nth Betti number)
  - E.g. 2 components in 0th homology, 1 component in 1st homology, 1 component in 2nd homology, 0 components in all following homologies → (2, 1, 1, 0, 0, …)

Math rambles...



We have lots of mathematical machinery to operate on transformations between points, e.g. probability and calculus.

Up until topology, we were limited to using these tools within a particular space at a given time.

Topology gives us a way to talk about entire spaces as points.

We can now use distance, probability, and calculus to study transformations between entire spaces! (in theory)

Thanks for your time.

Questions/comments?