Motivation for Topological Data Analysis

Alexander C. Mueller PhD
CEO and Founder of Capnion

March 13, 2019

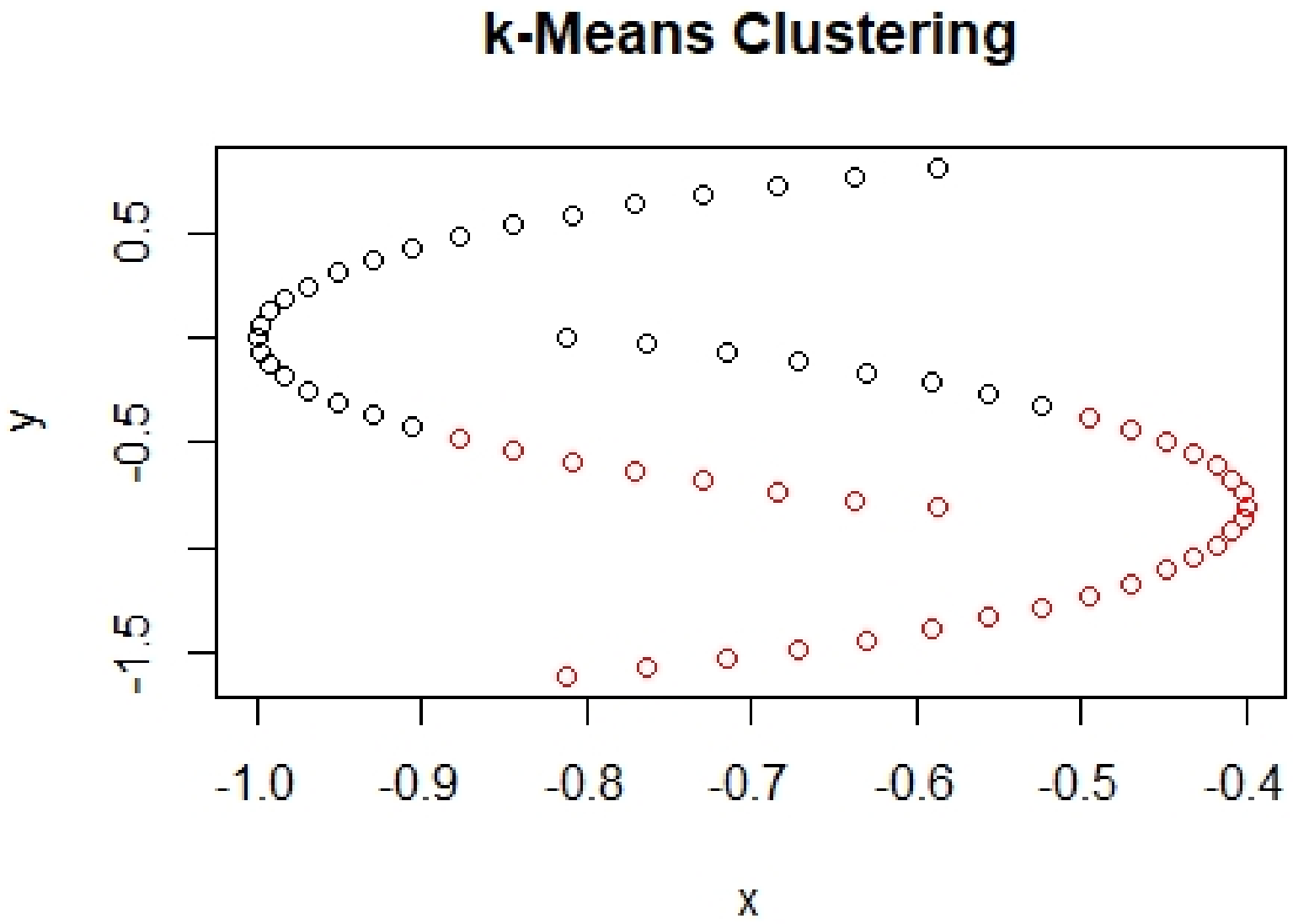
Who is the speaker? Where am I?

30 second resume:

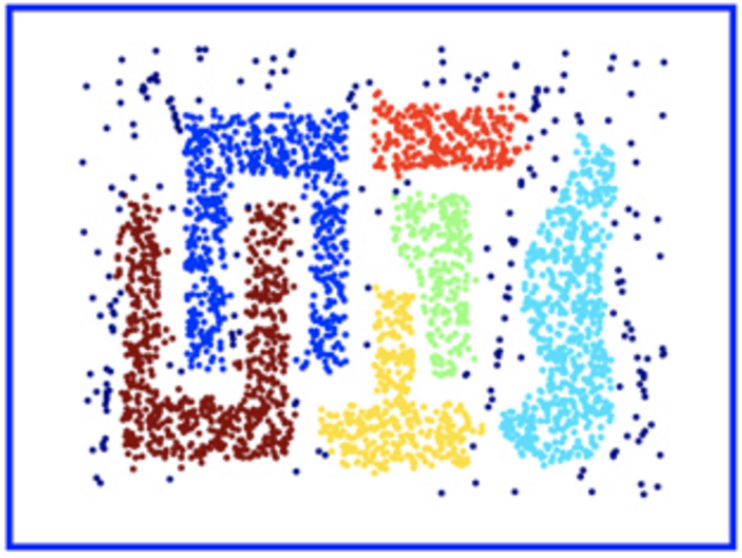
- an ancient metro Saint Louis townie
- grew up in University City
- University City High School
- B.A. Washington University (econ and math)
- Ph.D. University of Michigan (math)
- private sector data science
- founded data privacy company Capnion

The goal of this event is to provide a space to talk about emerging data technologies in detail.

Failure of *k*-Means and "Holes"

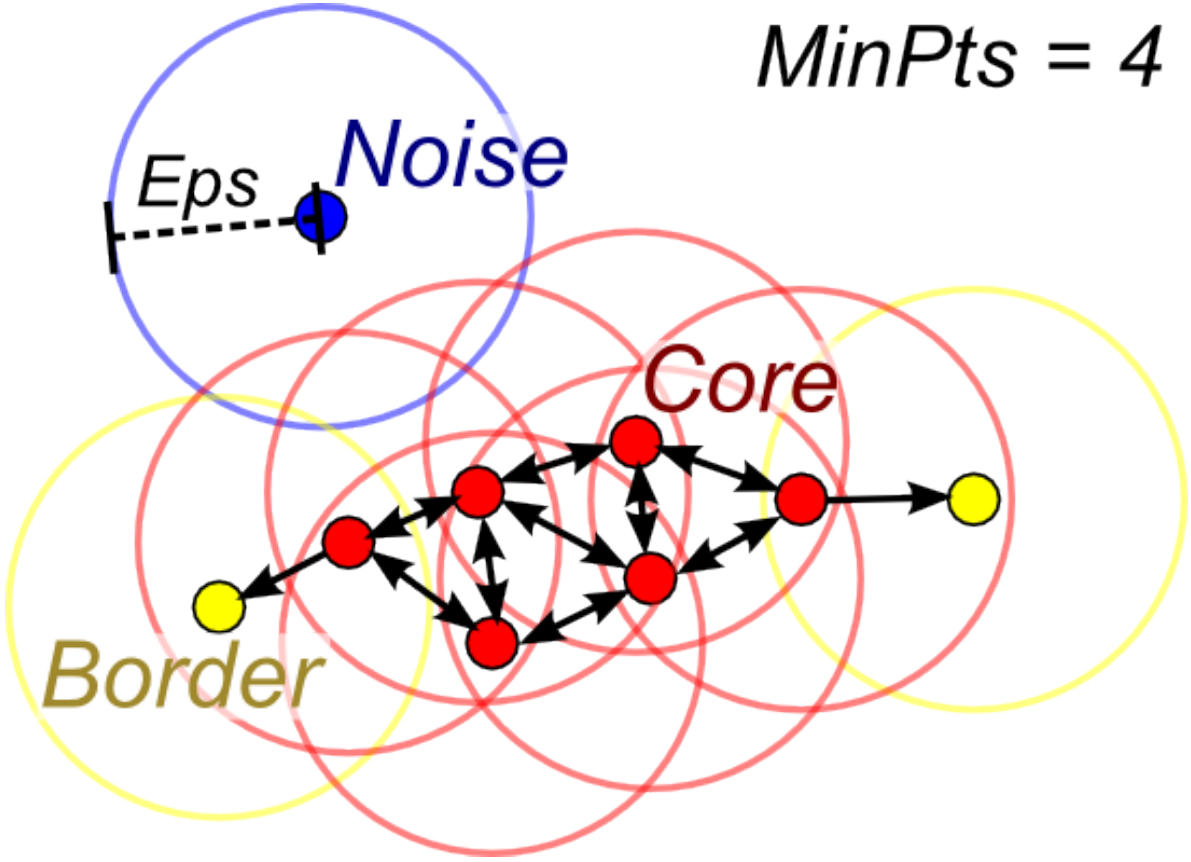


Datasets with Interesting Shape



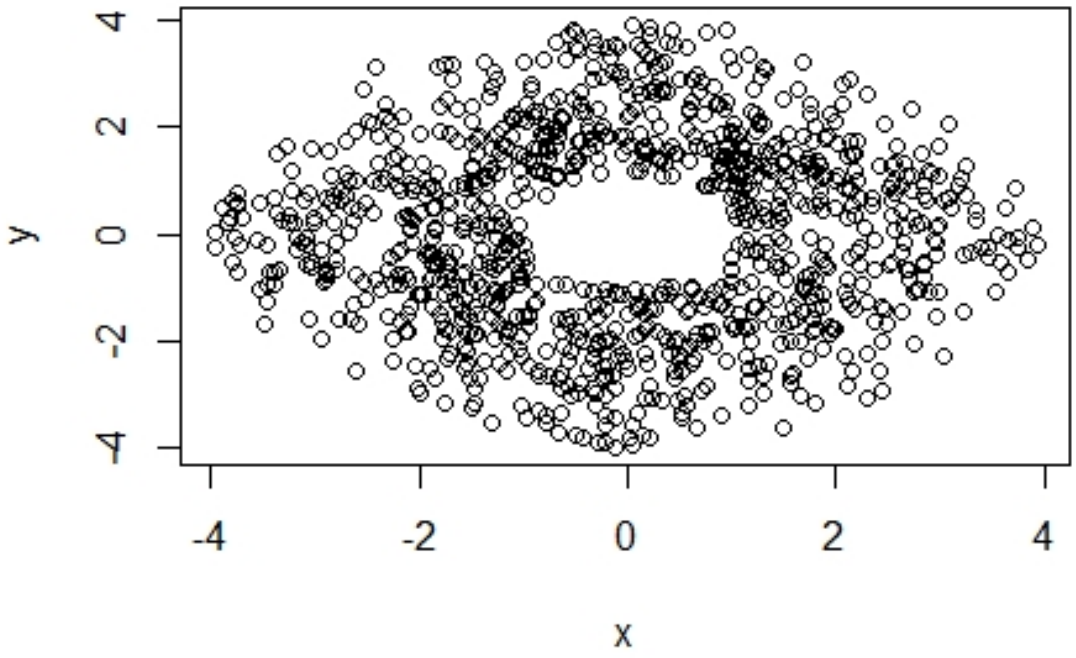
In higher dimension, how might we even know we were working with a funny shaped dataset like this? One could easily waste a lot of time getting mediocre results from k-means.

“Connected” Point Clouds



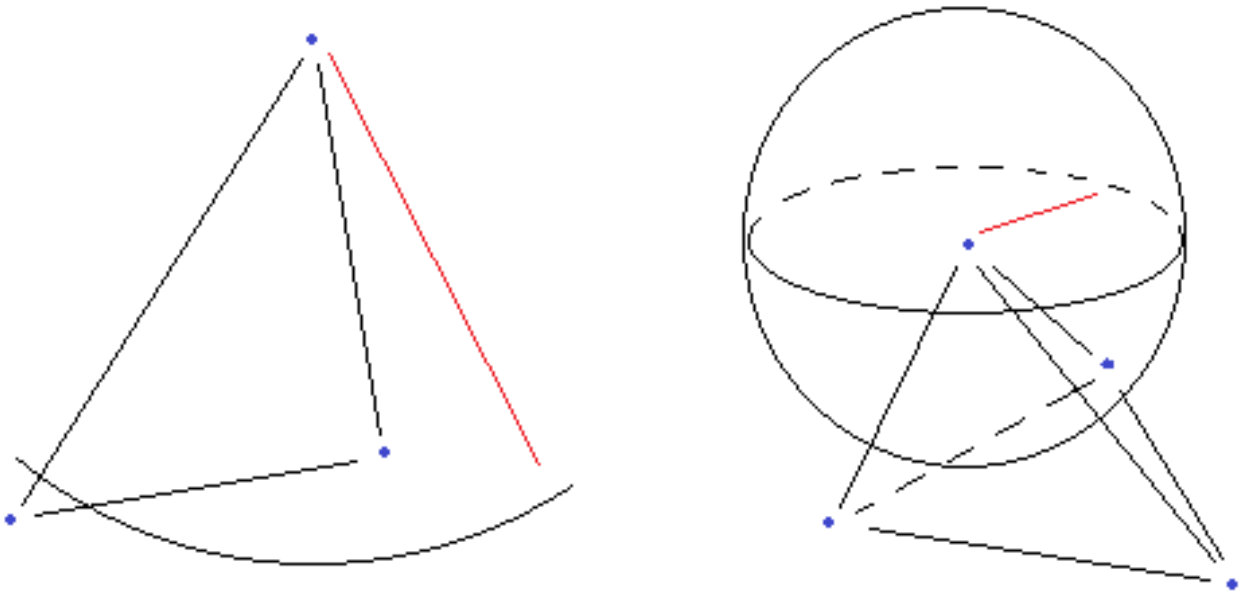
Fixed *Eps* gives you the DBSCAN clustering algorithm, to do TDA we try a range of *Eps* and see how the clustering changes.

An Annular Point Cloud



Is it reasonable to call this a circle with a hole in it? Or is a cloud of points just a cloud of points? Why or why not?

triangles, simplices, and little balls



Simplices have well-defined boundaries, and we'll detect holes by finding groups of points that define simplices with their boundaries, but not their interiors, "inside" the dataset.

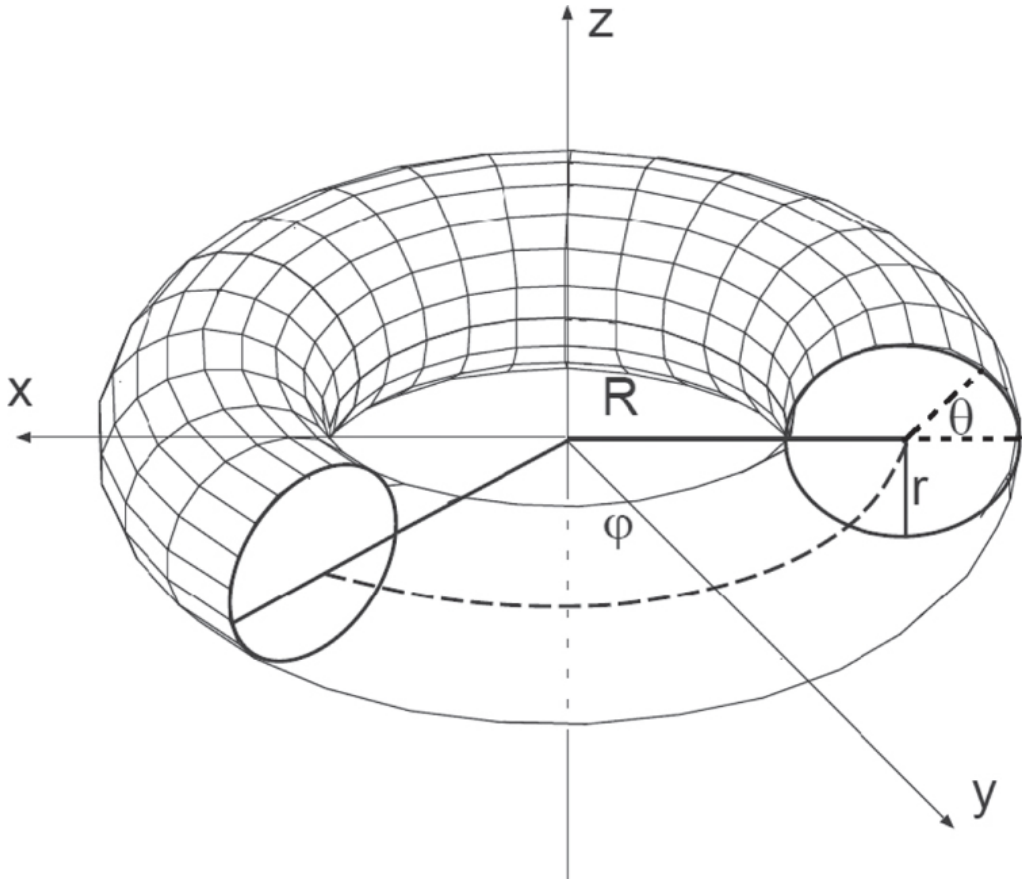
Applying Algebraic Topology

Algebraic topology is a field that contains tools for sifting all our simplices to find the most “interesting” ones. This will answer questions about the dataset like...

- What dimension is it really?
- Is it all the way solid?
- Or does it have holes in it?
- How many holes?
- What dimension are the holes really?

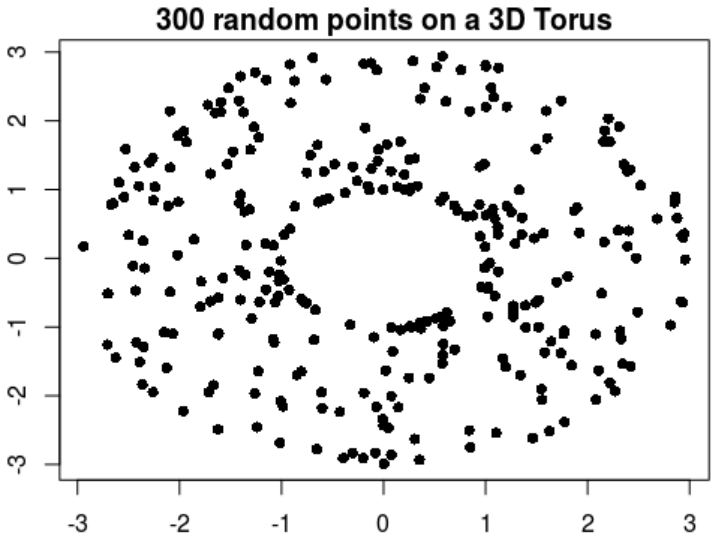
In practice, we’ll use a tool in R that does this sifting for us to find and categorize holes in the data.

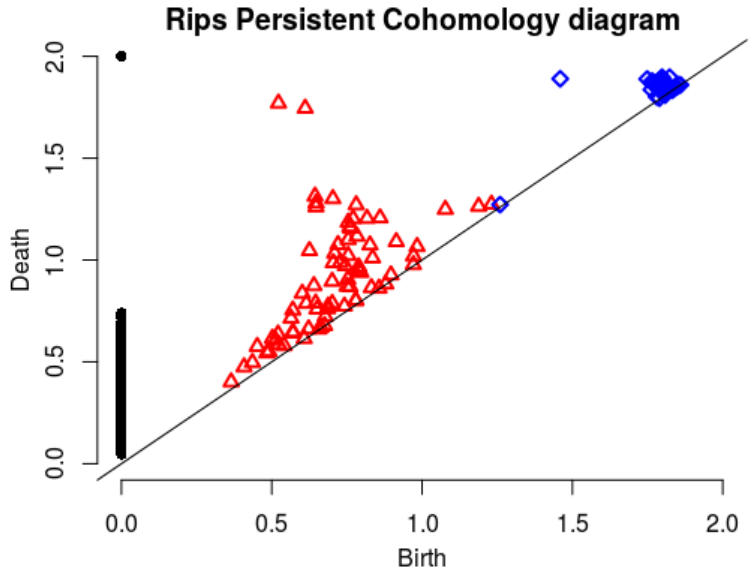
The Torus



The term “torus” refers to the two-dimensional *surface formed by the outside of the donut* and not the donut itself.

Rips Diagrams





The torus two “holes“ which show up as 1 dimensional cycles in our Rips diagram. We’ll cover some simpler examples in R and come back to this one.

Under the Hood

TDA is computationally demanding in general. We'll talk about relative efficiency when we go through our examples.

R's TDA library uses Rcpp and is really a wrapper for three common C++ libraries built for optimized TDA computations...

- GUDHI
- Dionysus
- PHAT

Persistent homology is a key mathematical concept underlying our Rips diagrams, and other tools in the library, but we won't do anything with it directly.

Interpreting Rips Diagrams

What to look for:

- Points far from the diagonal are more likely to represent "real" properties.
- High dimension cycles (points represented by shapes with more sides) tend to be rarer and more interesting.

What it means for the point cloud:

- High dimension cycles are empty spaces.

What it means in practical applications:

- A high-dimension cycle (roughly) represents a hidden tradeoff.
- A high-dimension cycle is represented by a group of observations that (roughly) with whose "average" is not similar to any observation in the dataset.

R Example Agenda

We'll consider a few examples, first handmade random datasets with obvious holes and then some familiar canned R datasets. We'll see that the Rips diagram finds some unusual features in the real-world datasets that we might speculate describe hidden tradeoffs.

- hollow circle
- fuzzily solid circle
- hollow sphere
- U.S. Judge Ratings
- Motor Trend Car Road Tests

Questions

Any questions?

Feel free to contact me at acmueller@capnion.com

Slides and code are available at
https://github.com/capnion/random