Motivation for Topological Data Analysis

Alexander C. Mueller PhD
CEO and Founder of Capnion

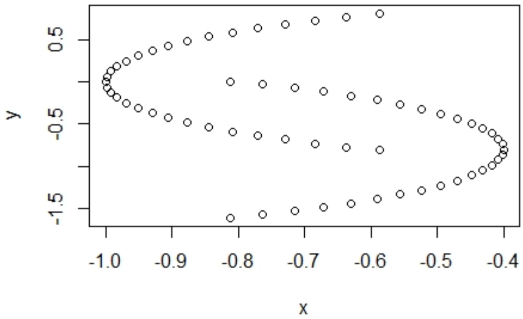
February 13, 2019

Who is the speaker?

30 second resume:

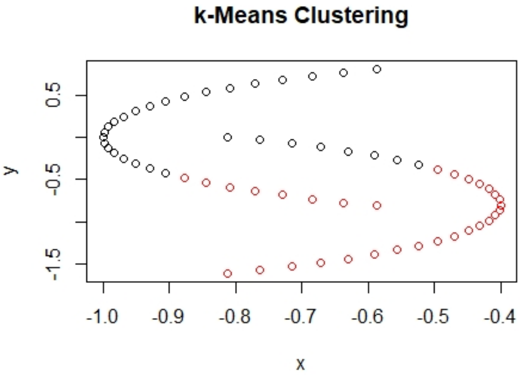
- an ancient metro Saint Louis townie
- grew up in University City
- University City High School
- B.A. Washington University (econ and math)
- Ph.D. University of Michigan (math)
- private sector data science
- founded data privacy company Capnion

A Clustering Challenge



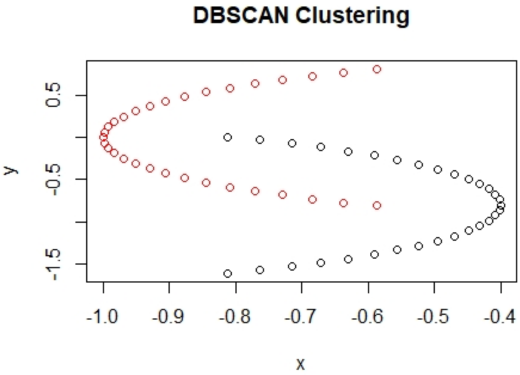
Does anyone see any difficulty applying common clustering algorithms to this dataset? What is the right answer?

Failure of *k*-Means



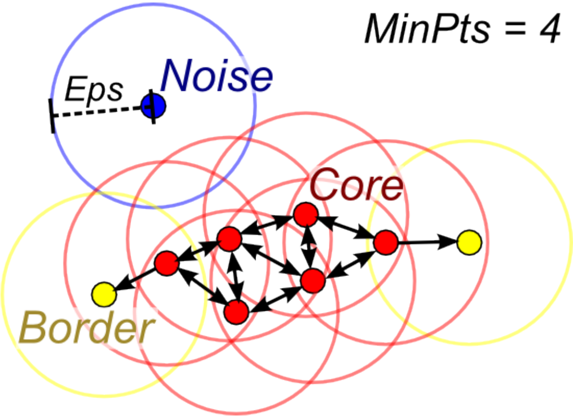
Why is *k*-means giving this obvious bad pair of clusters?

Success of DBSCAN



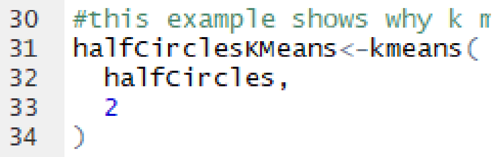
DBSCAN is successful because it is is a local method.

How DBSCAN Works

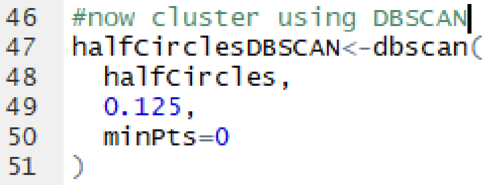


Roughly: points are in the same cluster if the distance between them is ϵ or less. This allows clusters that are not convex.

Code Compared



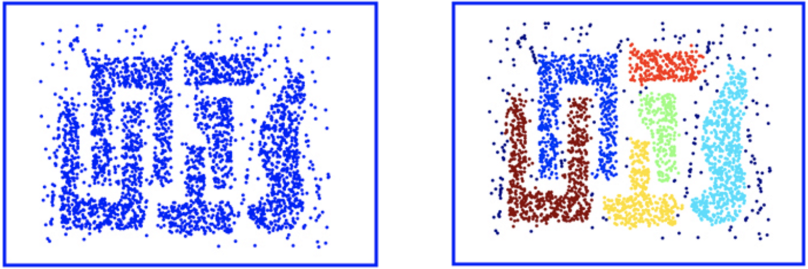
```
#this example shows why k m
halfcirclesKMeans<-kmeans(
  halfcircles,
  2
)
```



```
#now cluster using DBSCAN
halfcirclesDBSCAN<-dbscan(
  halfcircles,
  0.125,
  minPts=0
)
```

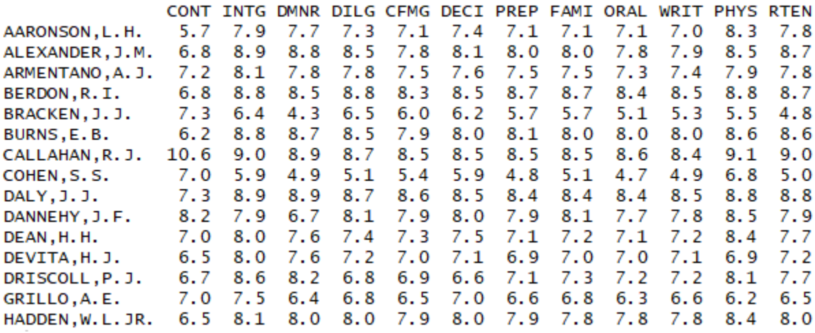
Rather than setting the number of clusters, we tell DBSCAN how to decide when things are “very close” to one another via a parameter ϵ (here it is 0.125).

Suitable Datasets



DBSCAN will consistently outperform centroid-oriented clustering methods like *k*-means for datasets that with clusters that have “holes” in them, are not convex, or are generally weird shapes.

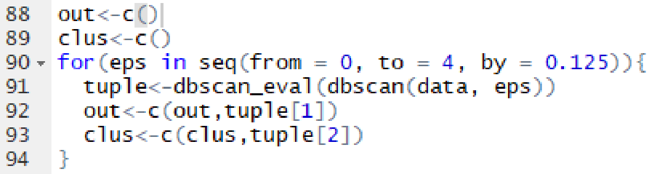
The U.S. Judges Dataset



```
                  CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
AARONSON,L.H.      5.7  7.9  7.7  7.3  7.1  7.4  7.1  7.1  7.1  7.0  8.3  7.8
ALEXANDER,J.M.     6.8  8.9  8.8  8.5  7.8  8.1  8.0  8.0  7.8  7.9  8.5  8.7
ARMENTANO,A.J.     7.2  8.1  7.8  7.8  7.5  7.6  7.5  7.5  7.3  7.4  7.9  7.8
BERDON,R.I.        6.8  8.8  8.5  8.8  8.3  8.5  8.7  8.7  8.4  8.5  8.8  8.7
BRACKEN,J.J.       7.3  6.4  4.3  6.5  6.0  6.2  5.7  5.7  5.1  5.3  5.5  4.8
BURNS,E.B.         6.2  8.8  8.7  8.5  7.9  8.0  8.1  8.0  8.0  8.0  8.6  8.6
CALLAHAN,R.J.     10.6  9.0  8.9  8.7  8.5  8.5  8.5  8.5  8.6  8.4  9.1  9.0
COHEN,S.S.         7.0  5.9  4.9  5.1  5.4  5.9  4.8  5.1  4.7  4.9  6.8  5.0
DALY,J.J.          7.3  8.9  8.9  8.7  8.6  8.5  8.4  8.4  8.4  8.5  8.8  8.8
DANNEHY,J.F.       8.2  7.9  6.7  8.1  7.9  8.0  7.9  8.1  7.7  7.8  8.5  7.9
DEAN,H.H.          7.0  8.0  7.6  7.4  7.3  7.5  7.1  7.2  7.1  7.2  8.4  7.7
DEVITA,H.J.        6.5  8.0  7.6  7.2  7.0  7.1  6.9  7.0  7.0  7.1  6.9  7.2
DRISCOLL,P.J.      6.7  8.6  8.2  6.8  6.9  6.6  7.1  7.3  7.2  7.2  8.1  7.7
GRILLO,A.E.        7.0  7.5  6.4  6.8  6.5  7.0  6.6  6.8  6.3  6.6  6.2  6.5
HADDEN,W.L.JR.     6.5  8.1  8.0  8.0  7.9  8.0  7.9  7.8  7.8  7.8  8.4  8.0
```

We are now working in twelve dimensions and it will be much harder to eyeball a good choice of ϵ and it is easy to get silly results if ϵ is way off.

Choosing ϵ Part 1



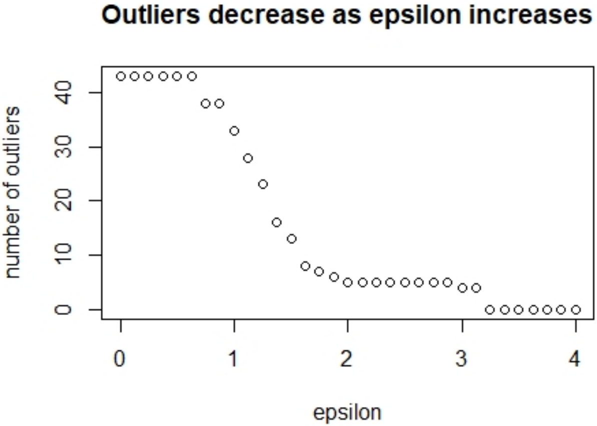
```
out<-c()
clus<-c()
for(eps in seq(from = 0, to = 4, by = 0.125)){
  tuple<-dbscan_eval(dbscan(data, eps))
  out<-c(out,tuple[1])
  clus<-c(clus,tuple[2])
}
```

We are going to iterate through a range of values for ϵ and record some crude metrics...

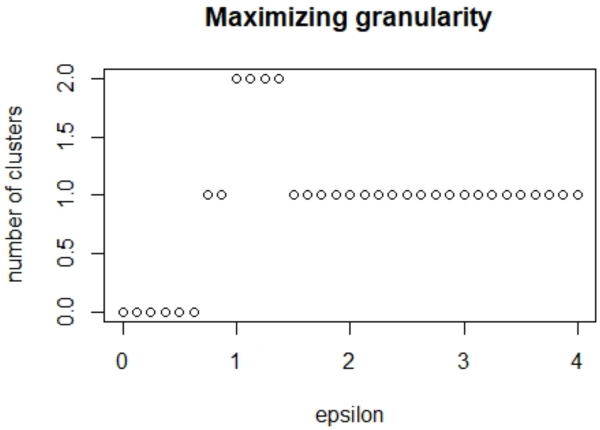
- number of clusters
- number of outliers

... of whether the results are good or bad.

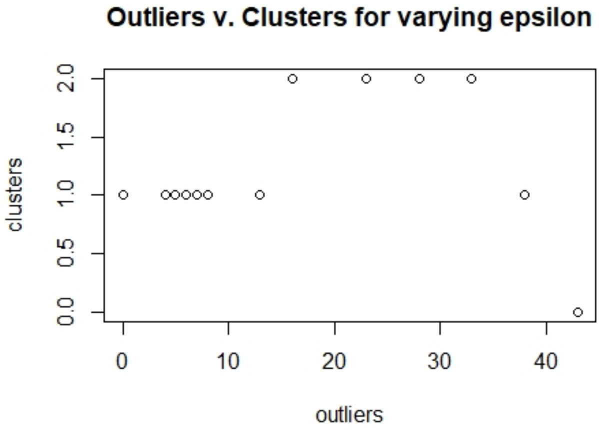
Choosing ϵ Part 2



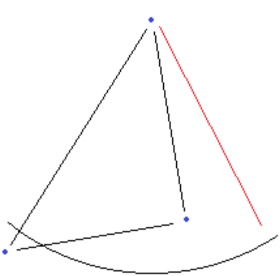
Choosing ϵ Part 3

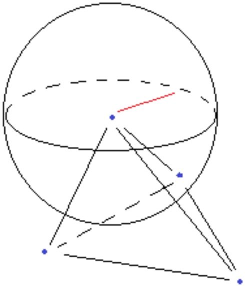


Choosing ϵ Part 4



triangles, simplices, and little balls





Instead of working with balls of radius ϵ we could look at triangles, or simplices in higher dimension, with sides less than ϵ in length. Is there anything intelligent we can do with the collection of such things? (Yes, algebraic topology)

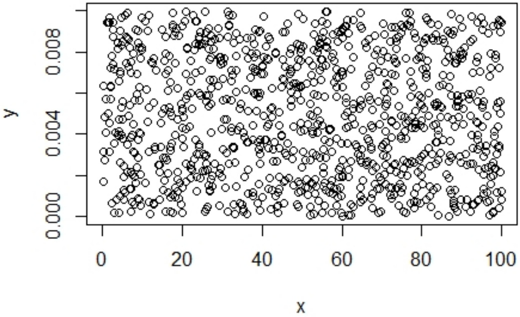
Algebraic Topology

Algebraic topology seeks to answer questions 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?

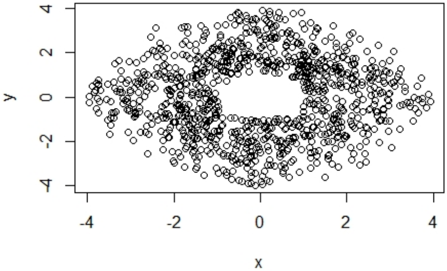
The “algebraic” part refers to the mathematics that powers these tools, which is technically demanding. However, from a datascience standpoint it is better (or at least easier) to start with the sorts of questions the tools answer, and just what sort of answers it gives. We’ll do this in a point cloud context.

Example



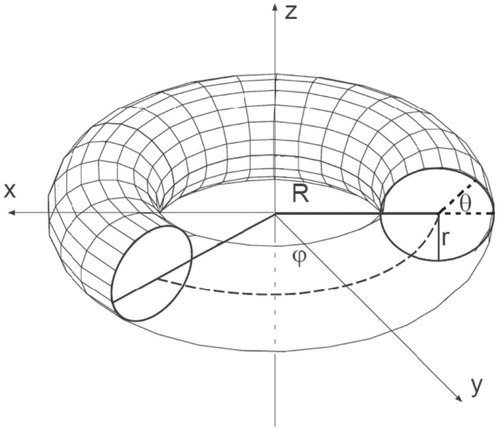
What is the “dimension” of this collection of points?

The Annulus



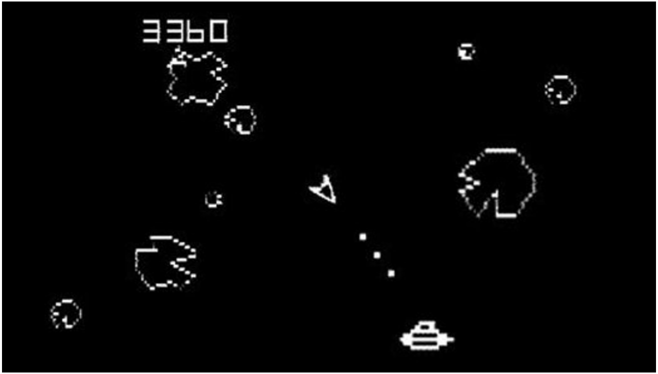
Is this just a collection of points, or somehow a solid ring with a hole in it? Does the empty space in the middle have any privilege over the empty space elsewhere? Why or why not?

The Torus



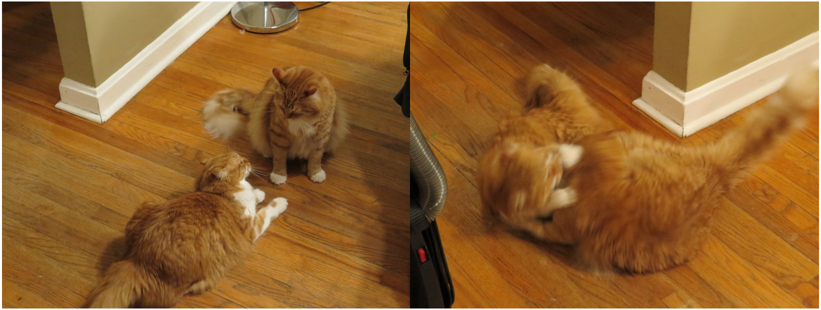
The term “torus” refers to the two-dimensional *surface formed by the outside of the donut* and not the donut itself.

The Asteroids Torus



Look familiar? Why bring up an old arcade game?

Application: Quantifying Blurriness of Images



How precise can we be about what it means to say an image is clear or blurry? How can we quantify it?

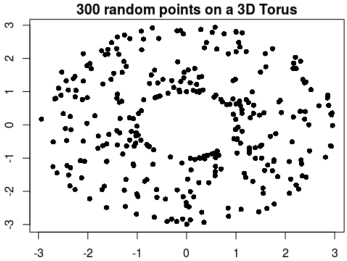
Clear Image Zoom

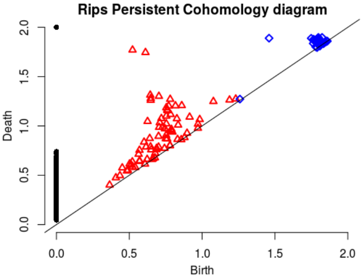
This image has relatively good separation of light and dark pixels. *The region of dark points has few holes in it*

Fuzzy Image Zoom

This image has relatively bad separation of light and dark pixels. *The region of dark points has many holes in it*

Next Time: Persistent (Co)Homology





Agenda and Homework

- Look for a heuristic description of what it does
- Ignore all the math defining it
- Look at pictures and work on interpreting them
- Play with the R TDA library on a toy dataset

Questions

Any questions?

Feel free to contact me at acmueller@capnion.com

Slides are available at
https://github.com/capnion/ ...
random/blob/master/acm_feb19stlskunkworks.pdf