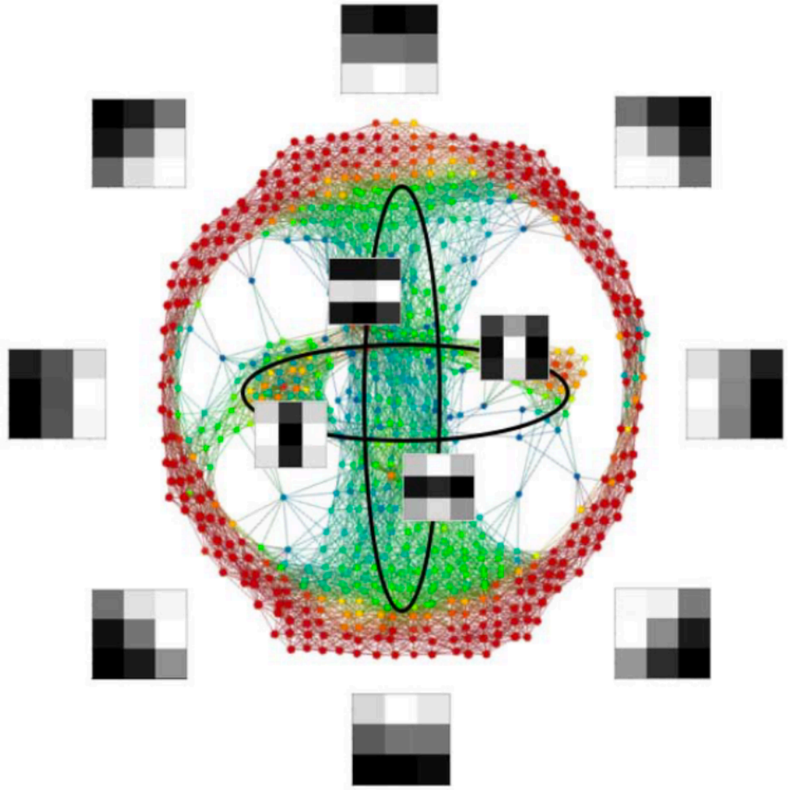
Topological Data Analysis

Klein bottles
and How Cornellians Find Them in Data



Chunyin Siu (Alex)
Center of Applied Mathematics, Cornell University
cs2323@cornell.edu

Before my talk…

Peace be with you

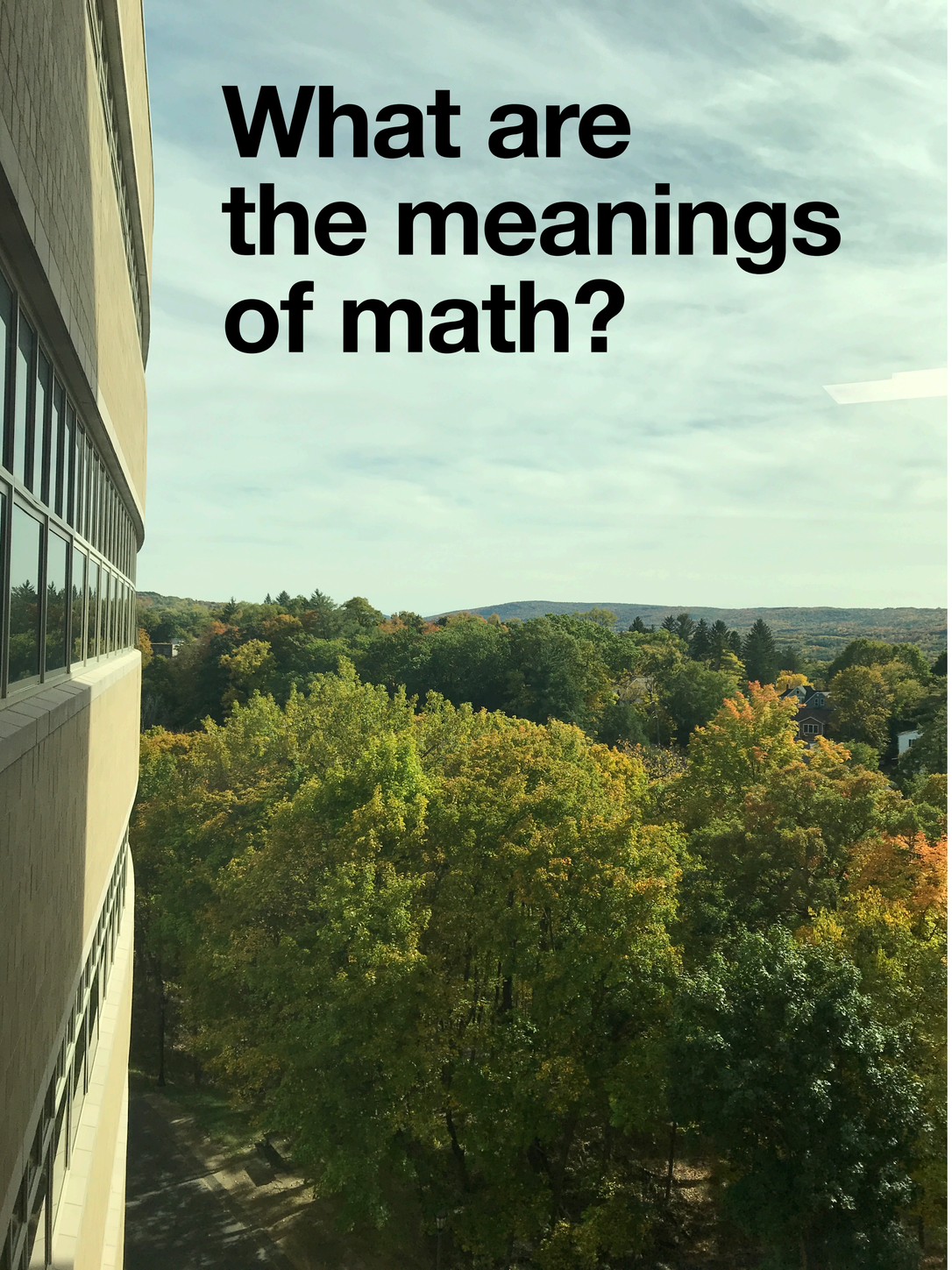
International Women's Day

My talk…

Far away by the Cascidilla Creek

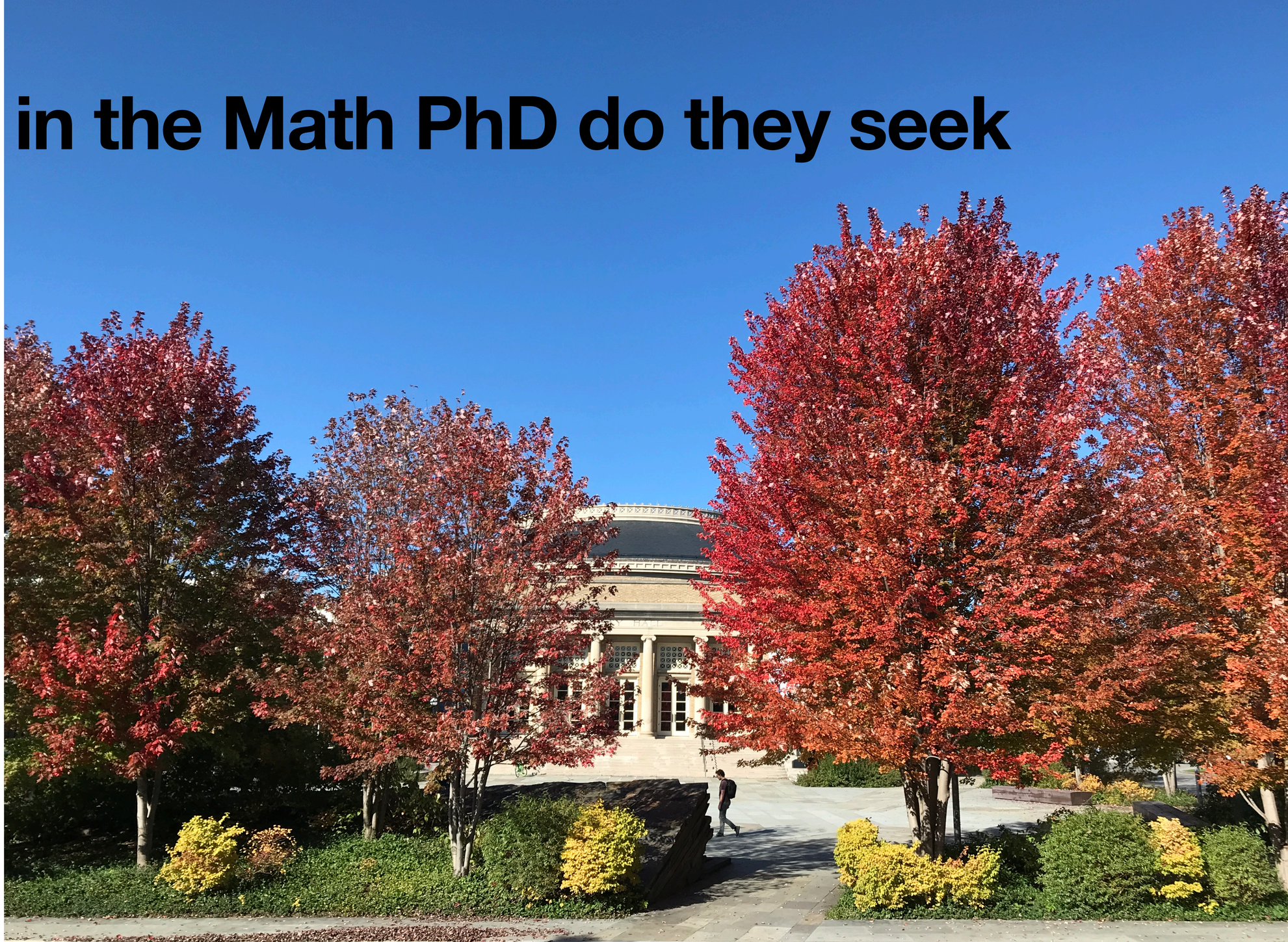
Coding all day

are the Applied Math geeks

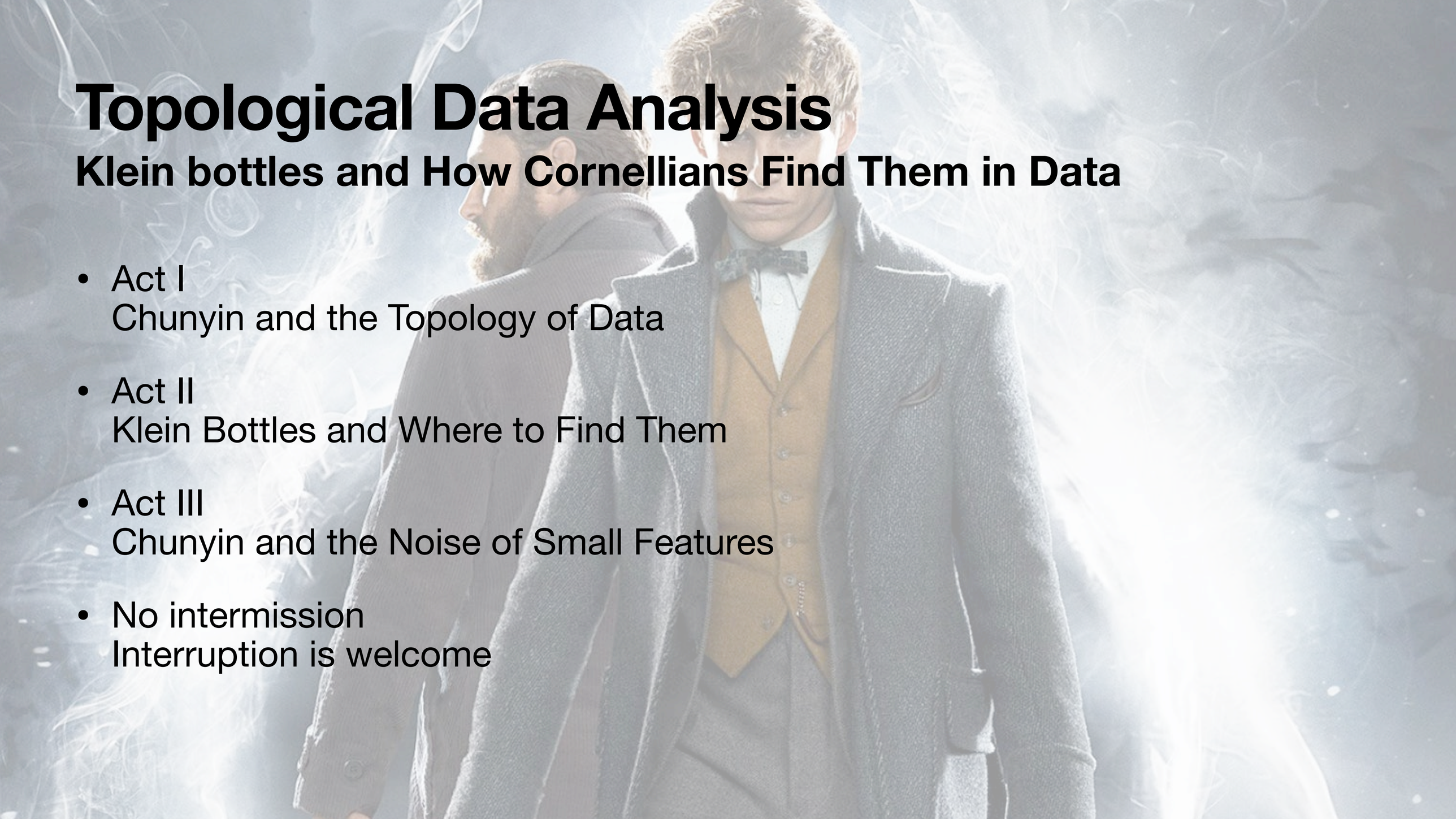


And what is the math of meanings?

Answers in the Math PhD do they seek



My talk… proper



Topological Data Analysis

Klein bottles and How Cornellians Find Them in Data

- Act I
  Chunyin and the Topology of Data
- Act II
  Klein Bottles and Where to Find Them
- Act III
  Chunyin and the Noise of Small Features
- No intermission
  Interruption is welcome

Act I

Chunyin and the Topology of Data

Topology

what popular math books tell you

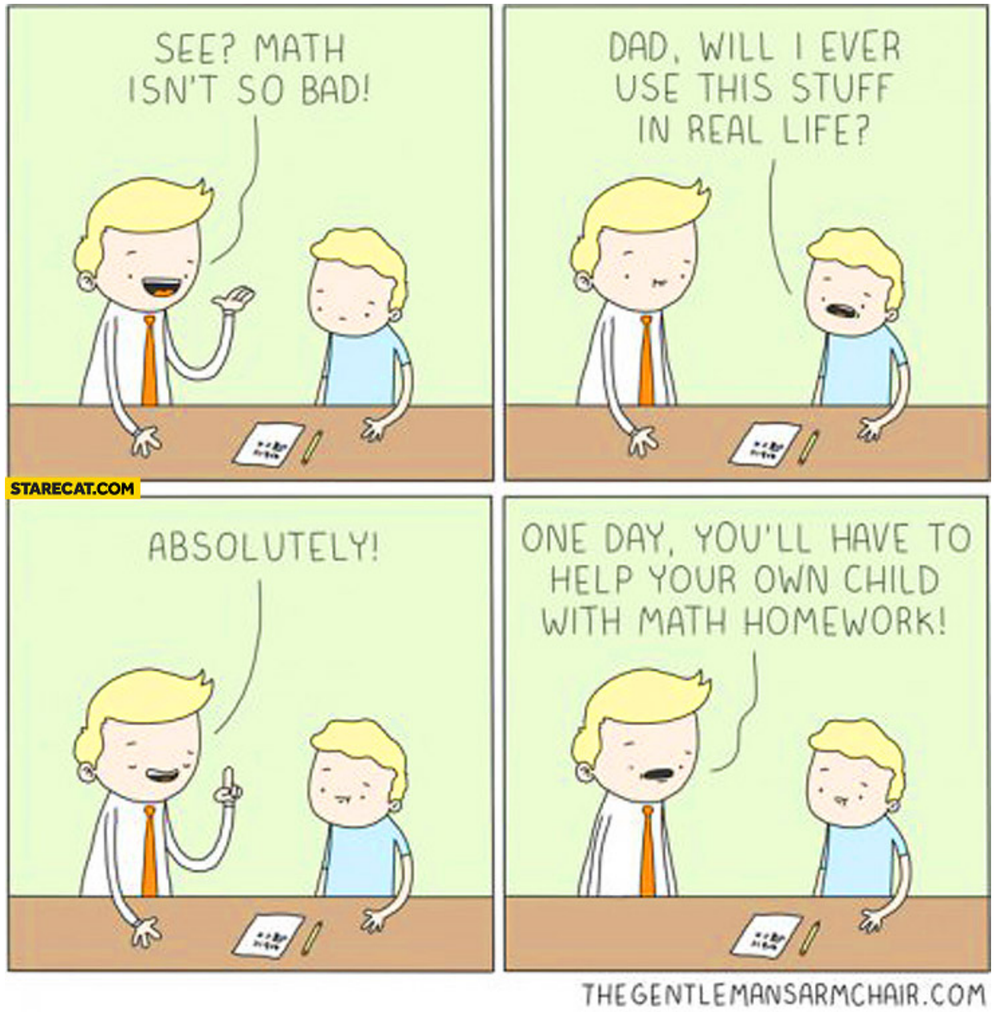
(Henry Segerman and Keenan Crane
https://www.youtube.com/watch?v=9NlqYr6-TpA)

$$
\begin{array}{ccccc}
H_n(D^n_\alpha, \partial D^n_\alpha) & \xrightarrow[\approx]{\partial} & \widetilde{H}_{n-1}(\partial D^n_\alpha) & \xrightarrow{\Delta_{\alpha\beta *}} & \widetilde{H}_{n-1}(S^{n-1}_\beta) \\
\downarrow \Phi_{\alpha *} & & \downarrow \varphi_{\alpha *} & & \uparrow q_{\beta *} \\
H_n(X^n, X^{n-1}) & \xrightarrow{\partial_n} & \widetilde{H}_{n-1}(X^{n-1}) & \xrightarrow{q_*} & \widetilde{H}_{n-1}(X^{n-1}/X^{n-2}) \\
 & \searrow^{d_n} & \downarrow j_{n-1} & & \downarrow \approx \\
 & & H_{n-1}(X^{n-1}, X^{n-2}) & \xrightarrow{\approx} & H_{n-1}(X^{n-1}/X^{n-2}, X^{n-2}/X^{n-2})
\end{array}
$$

what your see in a topology class

(Allen Hatcher
https://pi.math.cornell.edu/~hatcher/AT/AT+.pdf)

Topology

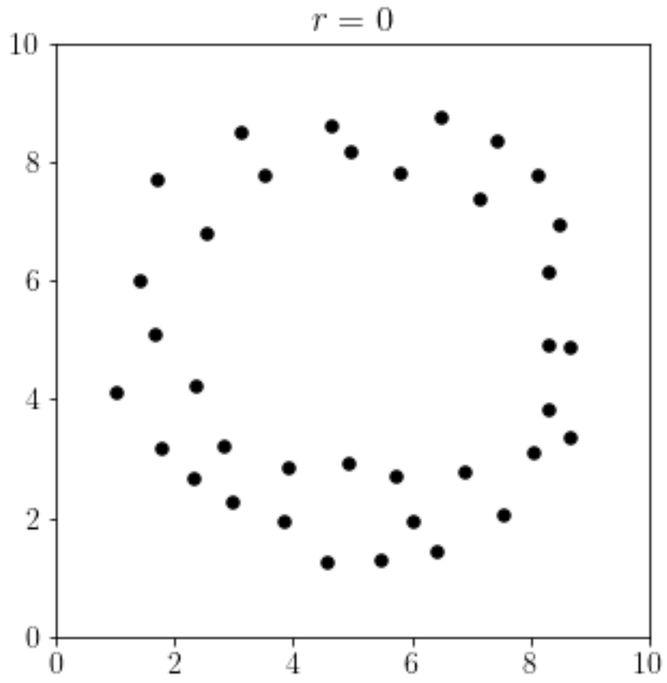

SEE? MATH ISN'T SO BAD!
DAD, WILL I EVER USE THIS STUFF IN REAL LIFE?
STARECAT.COM
ABSOLUTELY!
ONE DAY, YOU'LL HAVE TO HELP YOUR OWN CHILD WITH MATH HOMEWORK!
THEGENTLEMANSARMCHAIR.COM

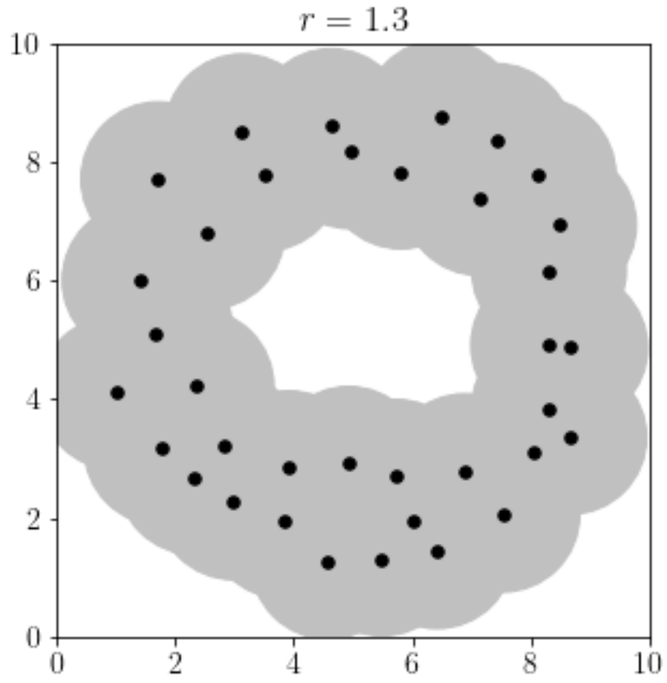
Carlsson, 2009

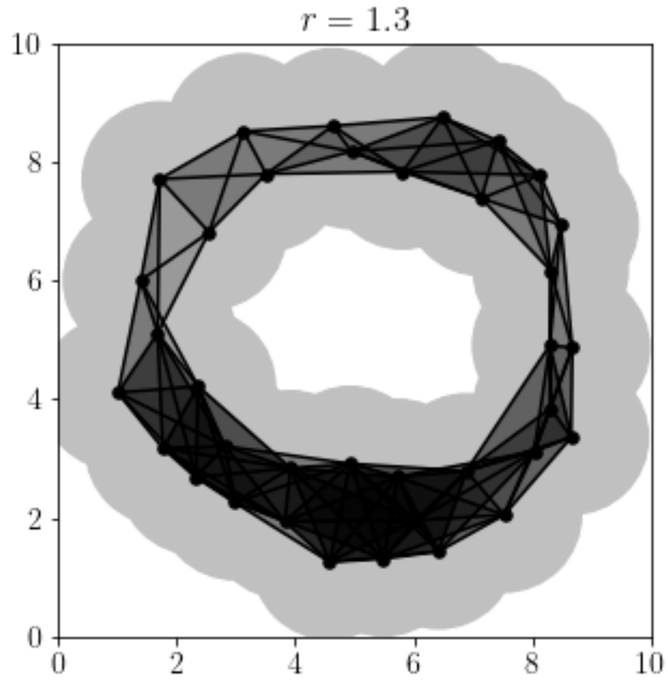
- Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308.

Tom Hanks in Cast Away (2000), directed by Robert Zemeckis

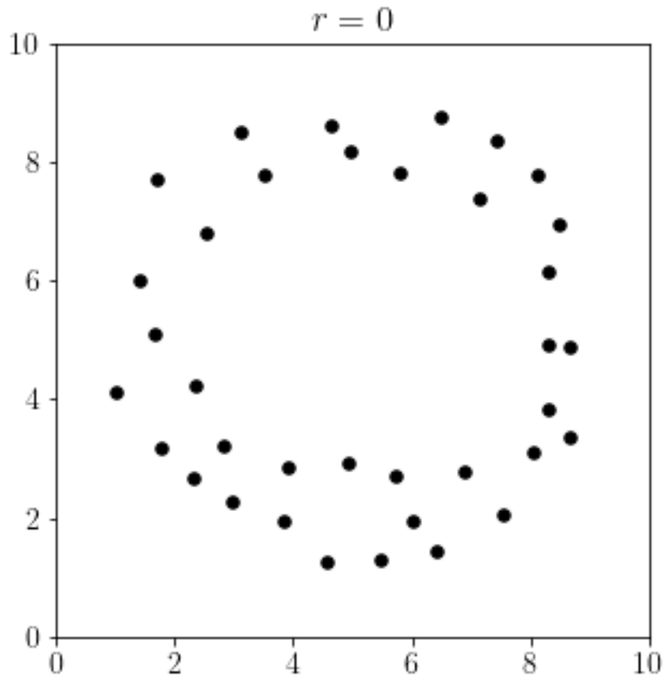
Topology of Data

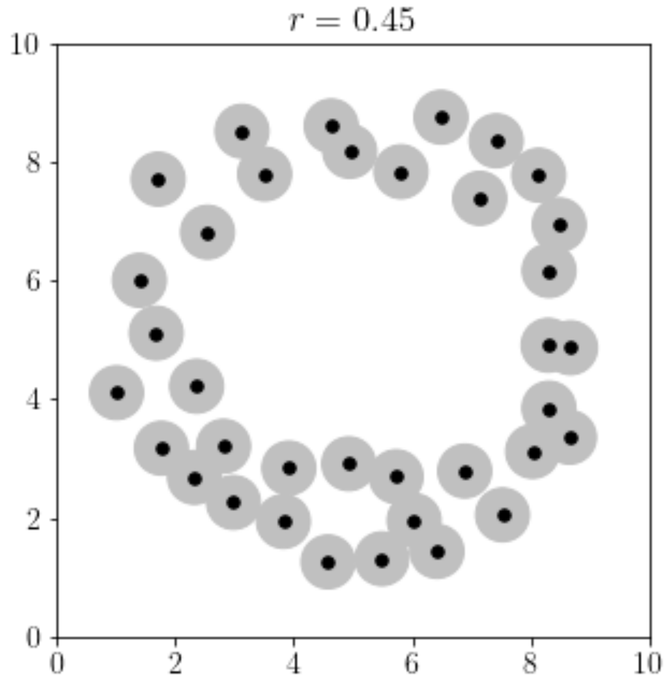


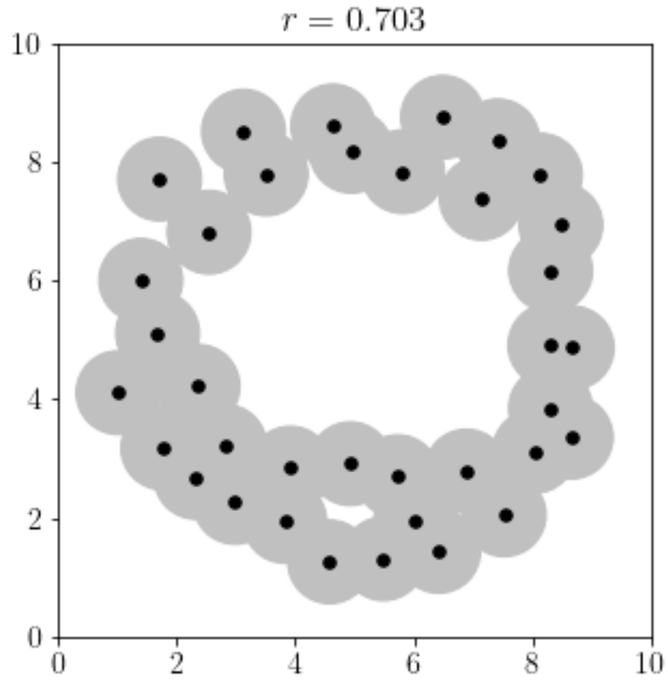


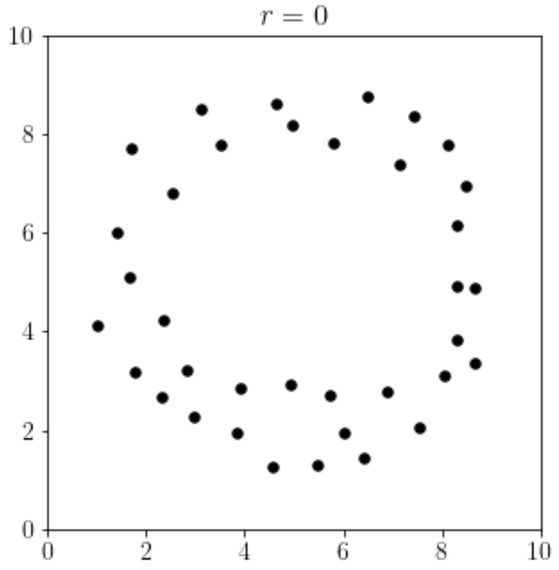


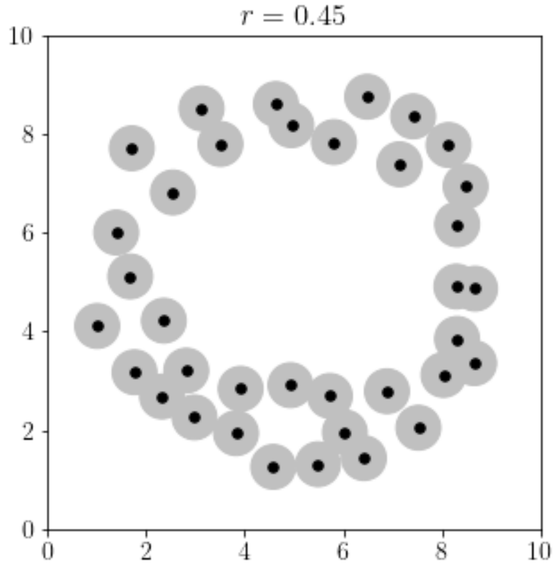
Pitfall

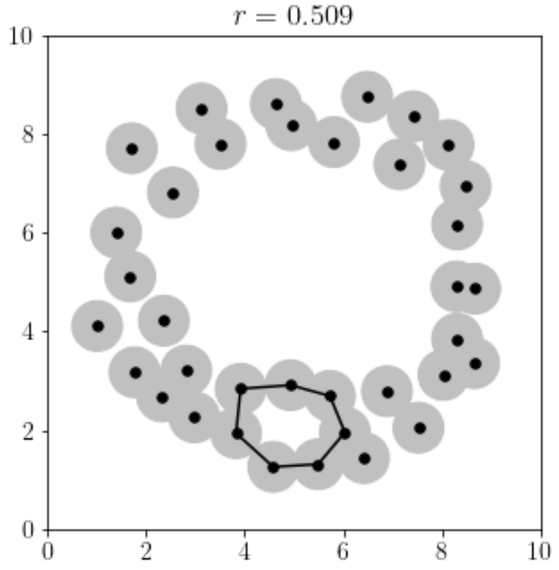


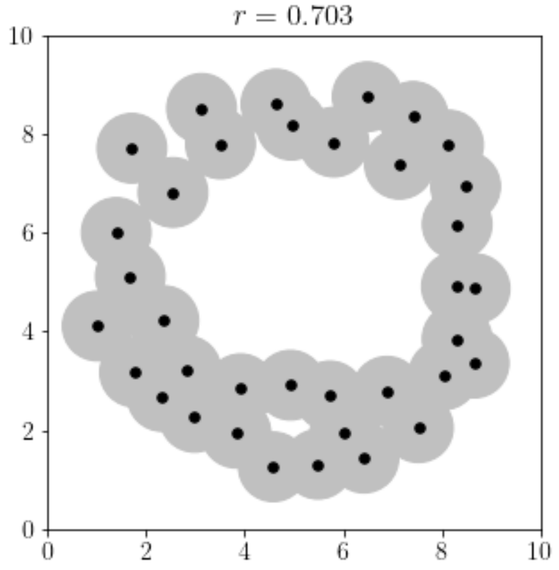


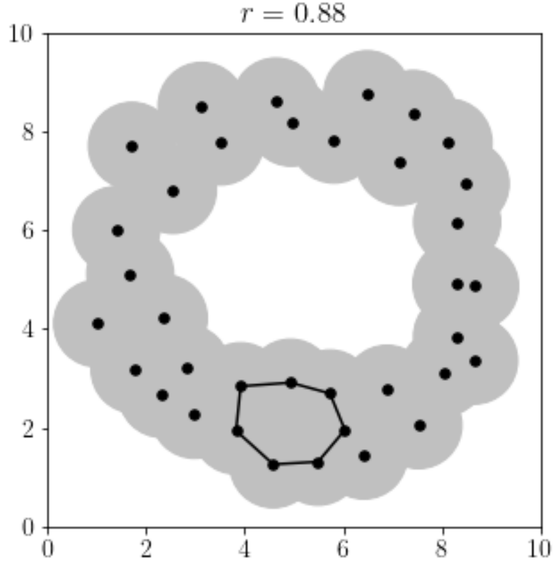


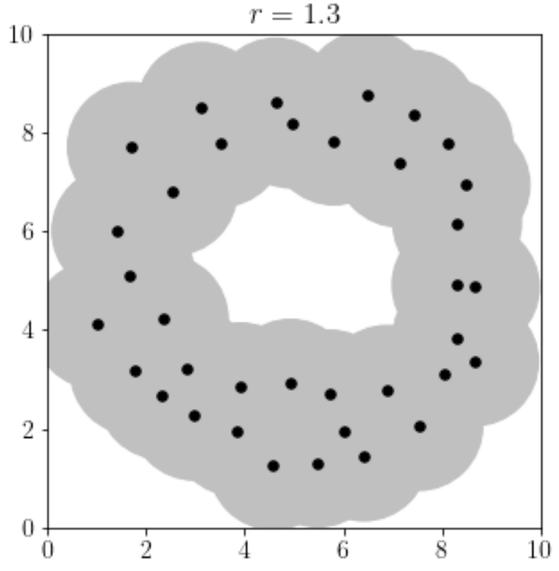

$r = 0$

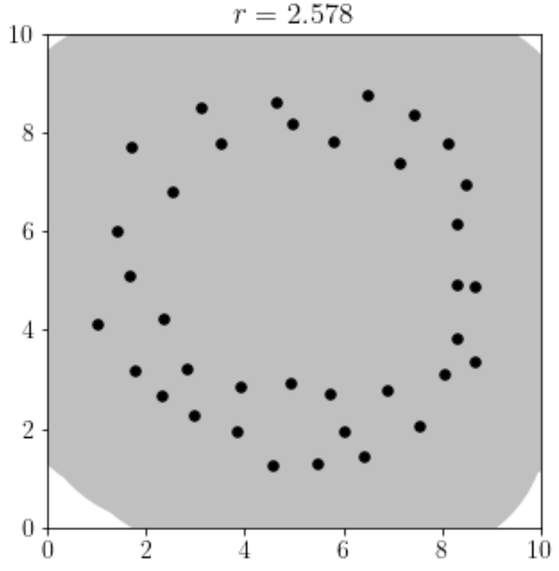

$r = 0.45$

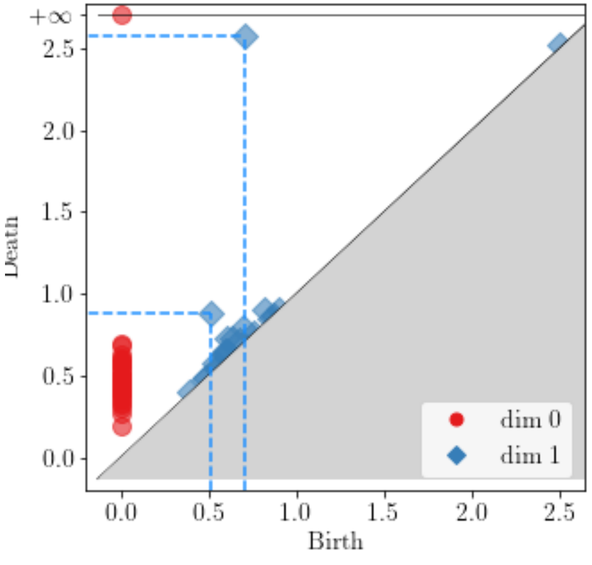

$r = 0.509$


$r = 0.703$


$r = 0.88$


$r = 1.3$


$r = 2.578$


Death
Birth
dim 0
dim 1

A Windy Road

Tom Hanks in Cast Away (2000), directed by Robert Zemeckis

Act II

Klein Bottles and Where to Find Them

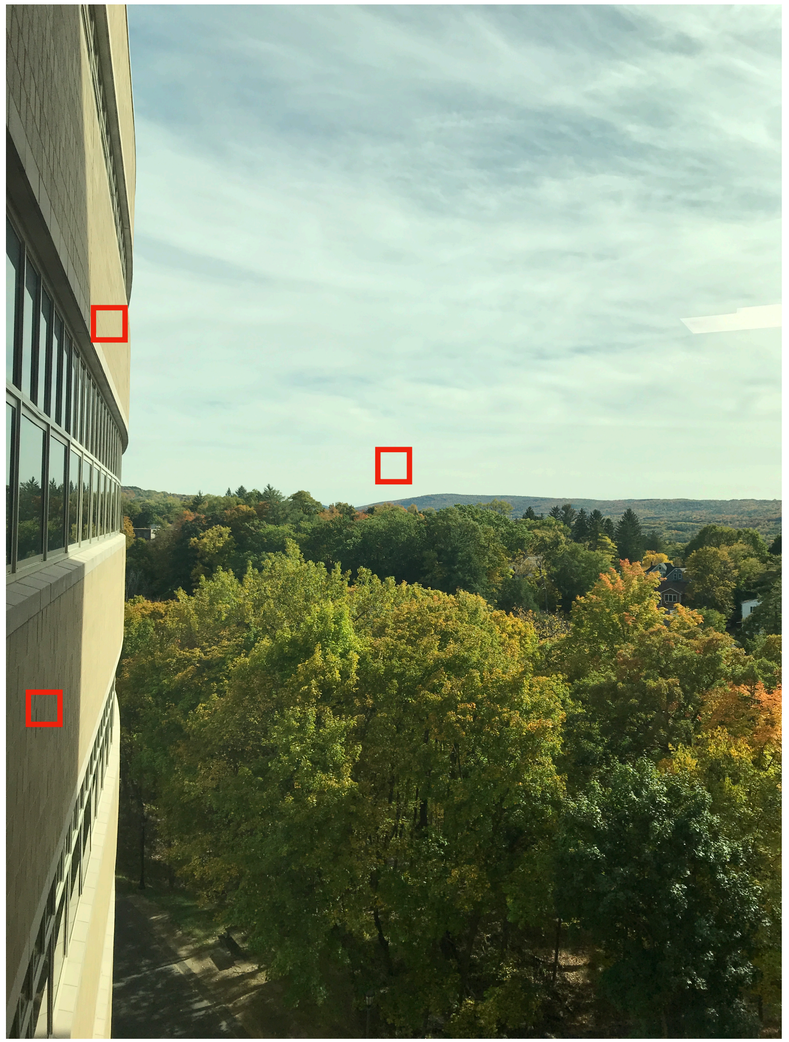
(Carlsson et al, 2008)

A dataset of natural images…

Patches in greyscale images

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{33} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \in M_{3\times 3}(\mathbb{R}) \cong \mathbb{R}^9$$

Throw away boring patches

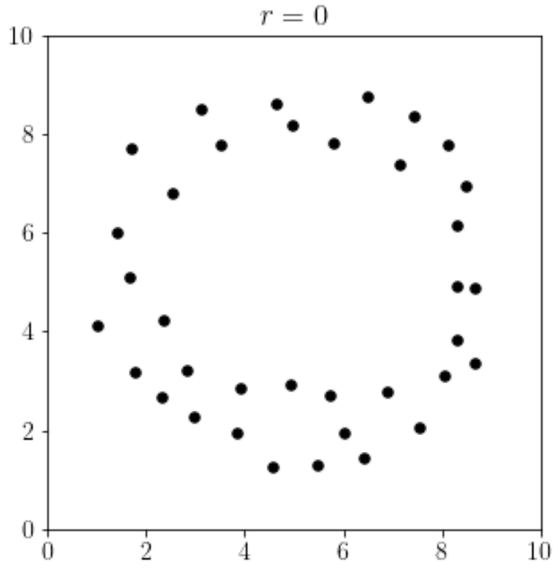


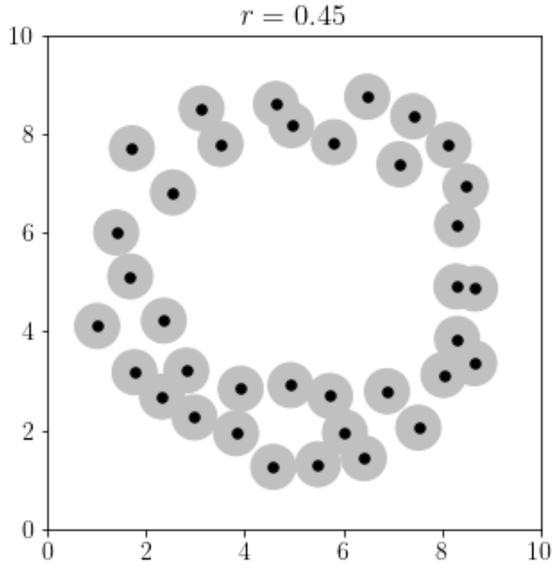
Normalize

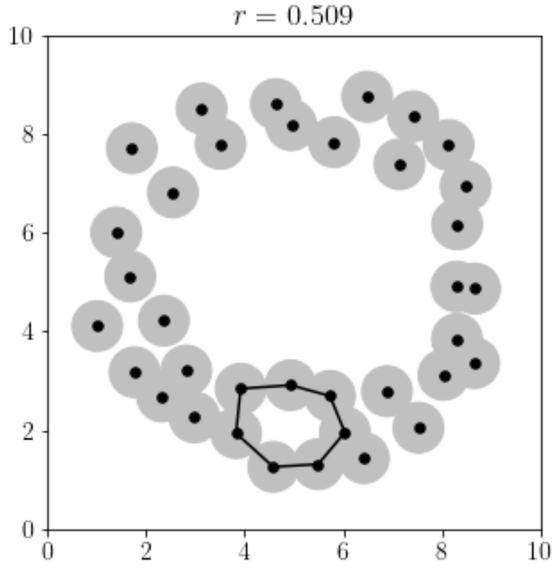
- subtract mean
- divide by norm
- $\mathbb{R}^9 \to \mathbb{S}^7$

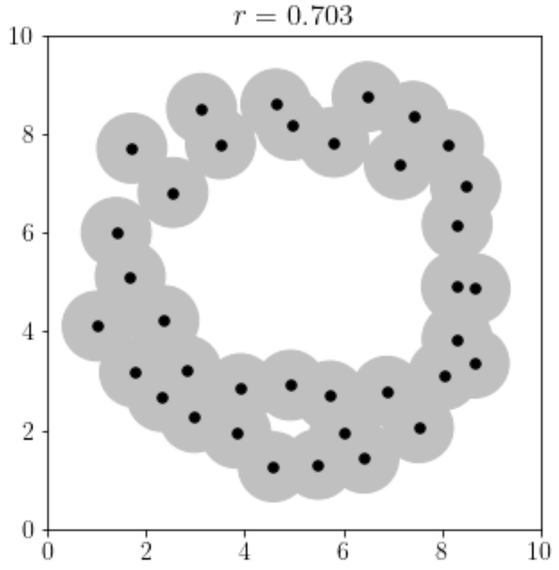
High-contrast natural image patches lie on some low-dimensional submanifold?

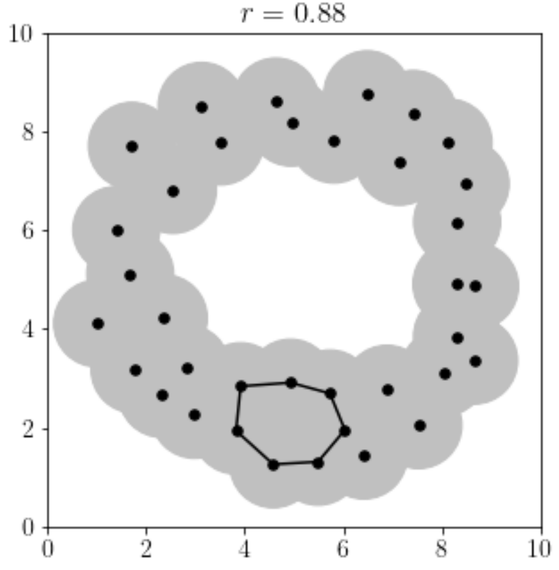
(of $\mathbb{S}^7$ upon normalization)

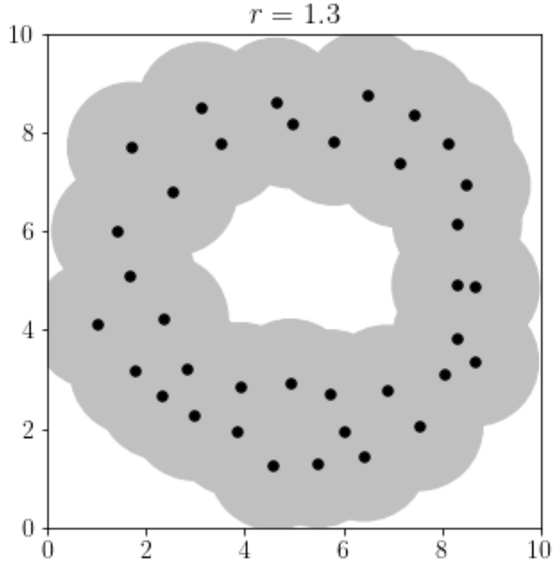

$r = 0$

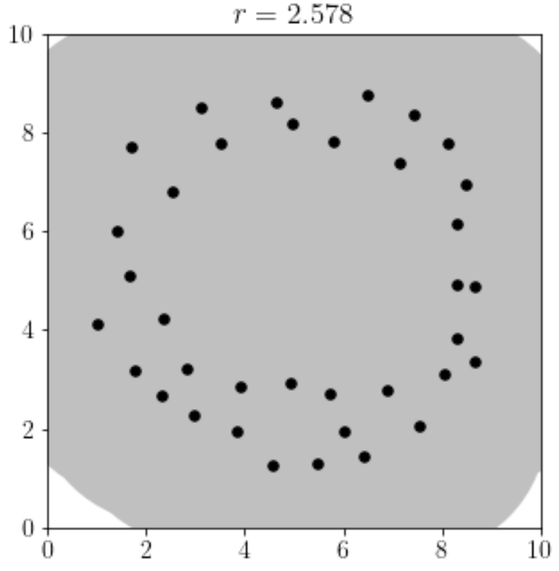

$r = 0.45$

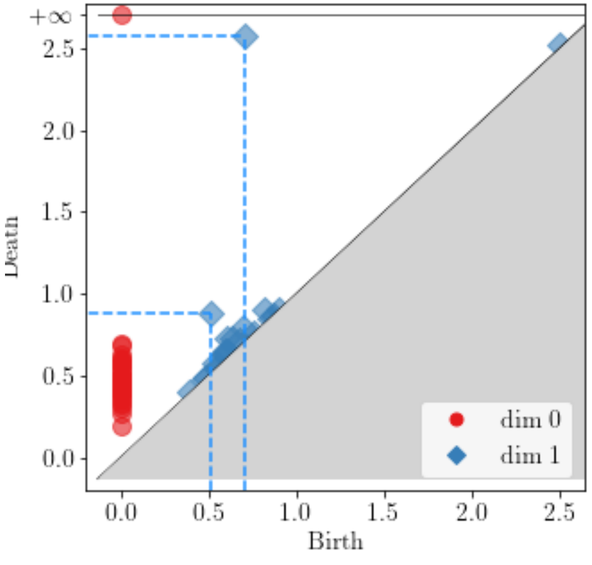

$r = 0.509$


$r = 0.703$

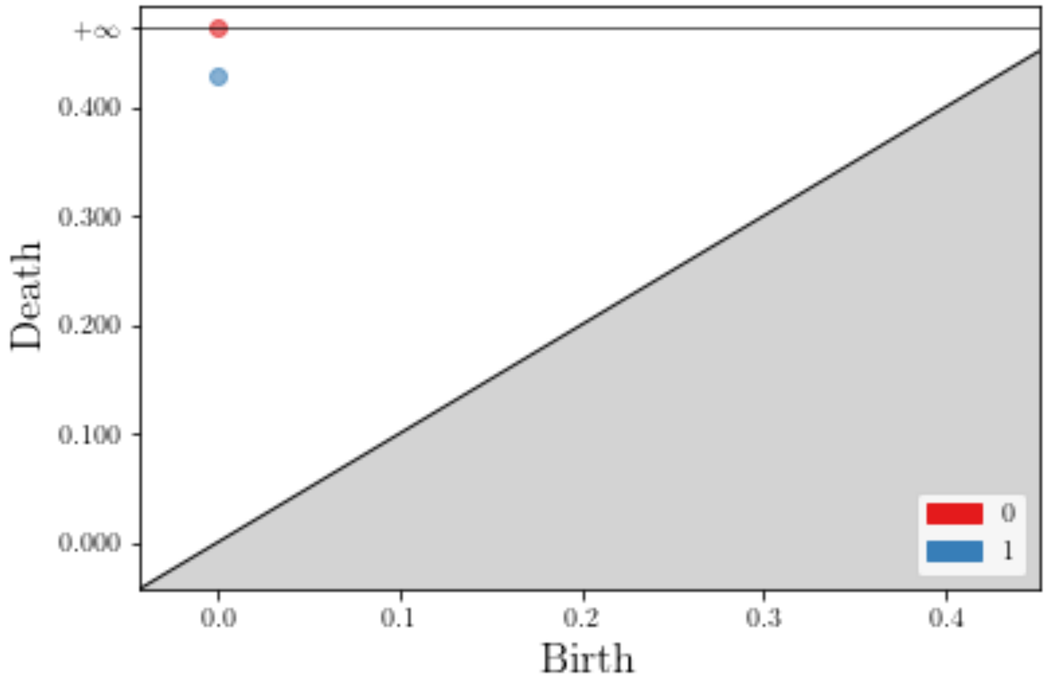

$r = 0.88$


$r = 1.3$

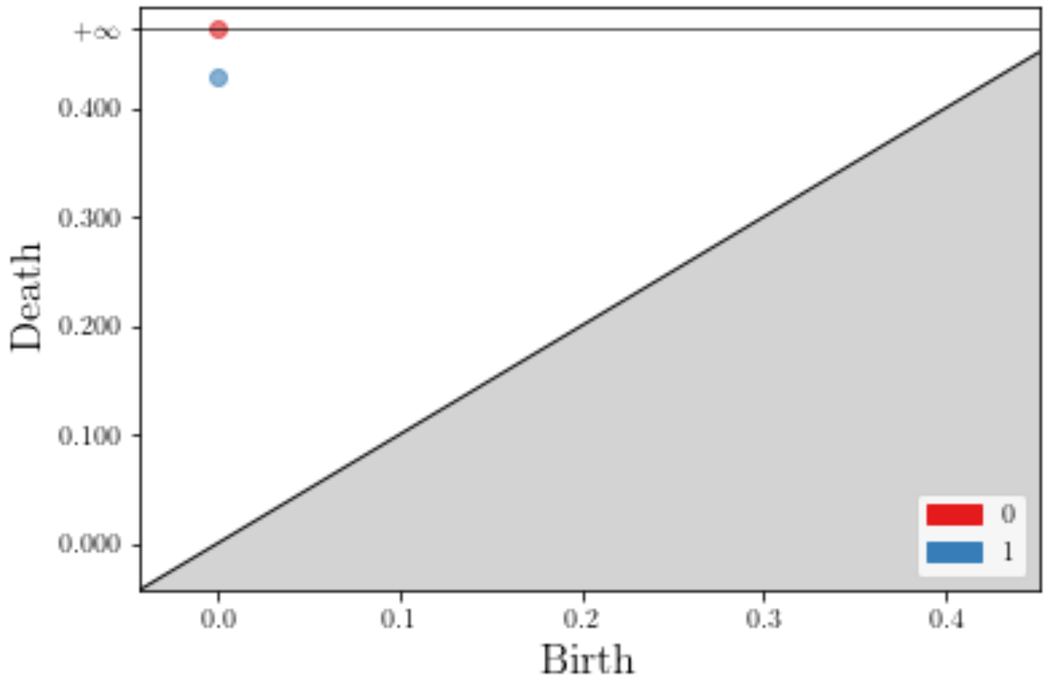

$r = 2.578$

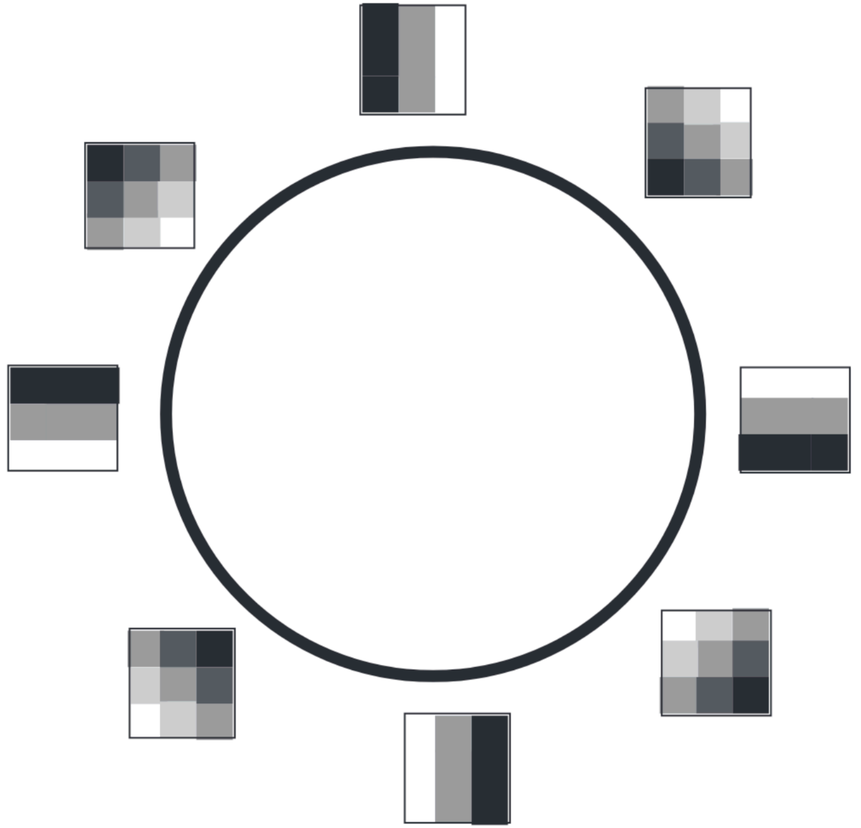

Death
Birth
dim 0
dim 1

Heavy Denoising

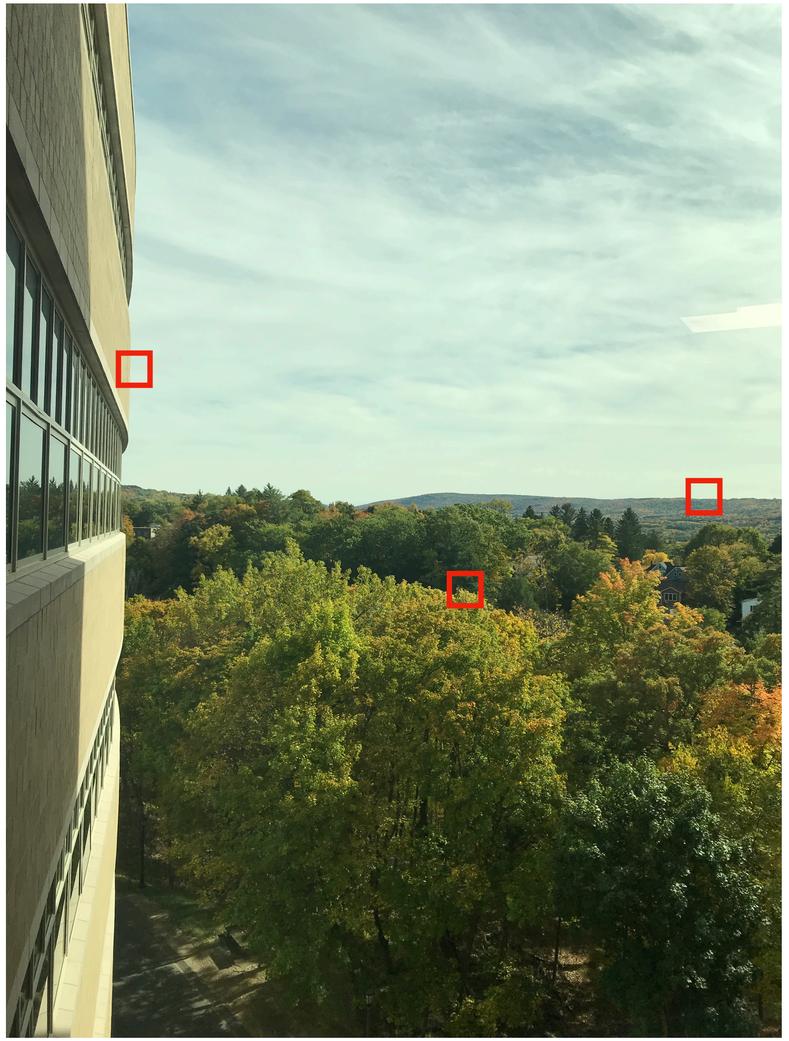


Heavy Denoising

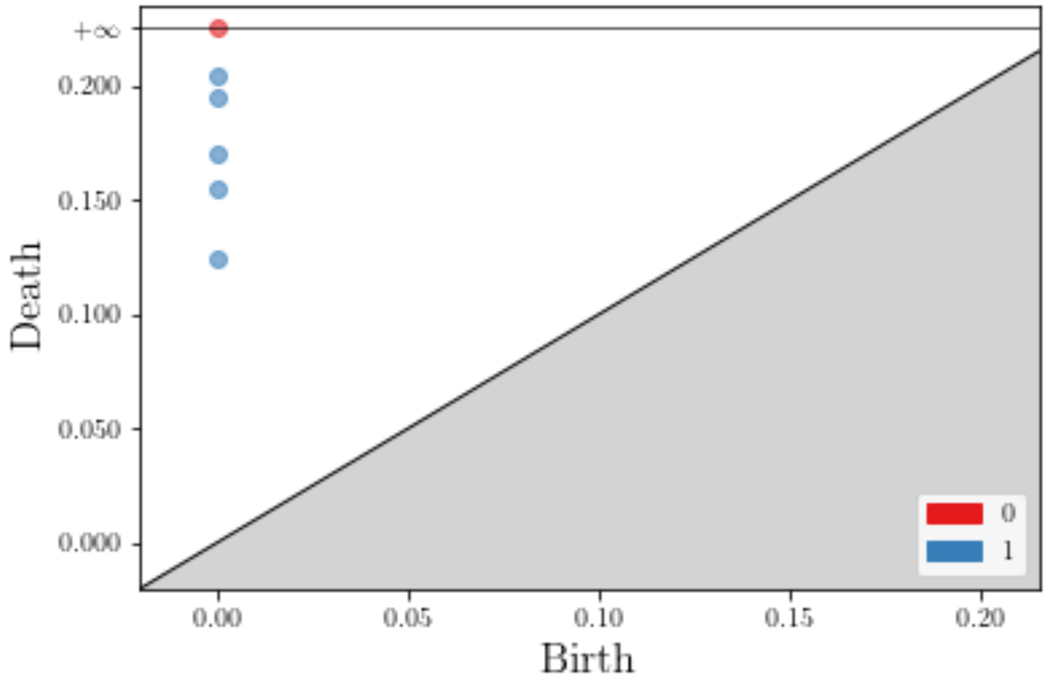




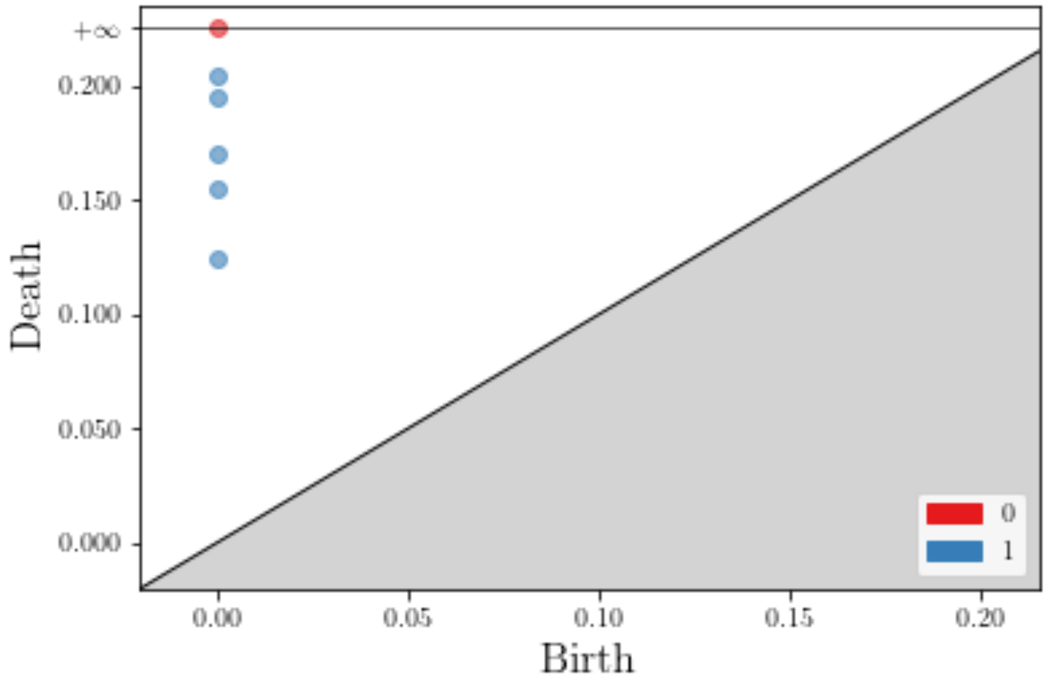
Local gradients

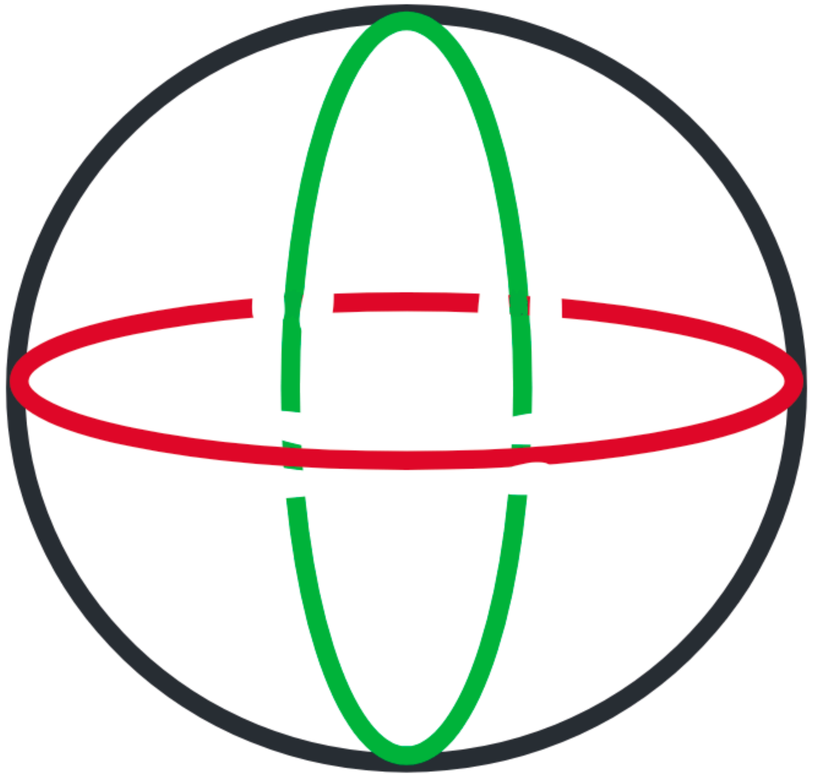


Less Aggressive Denoising

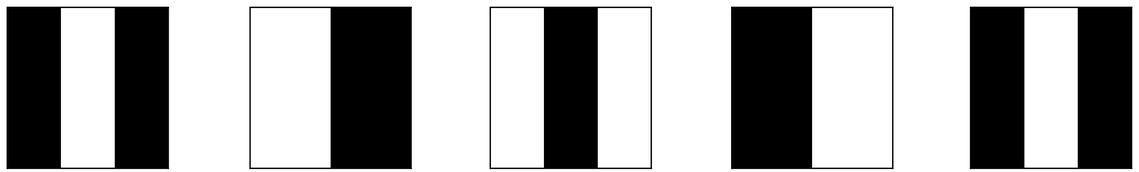


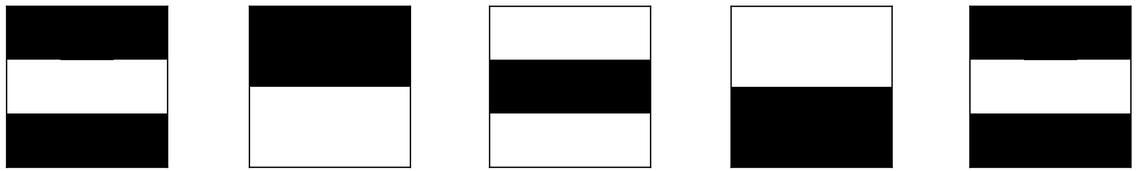
Less Aggressive Denoising

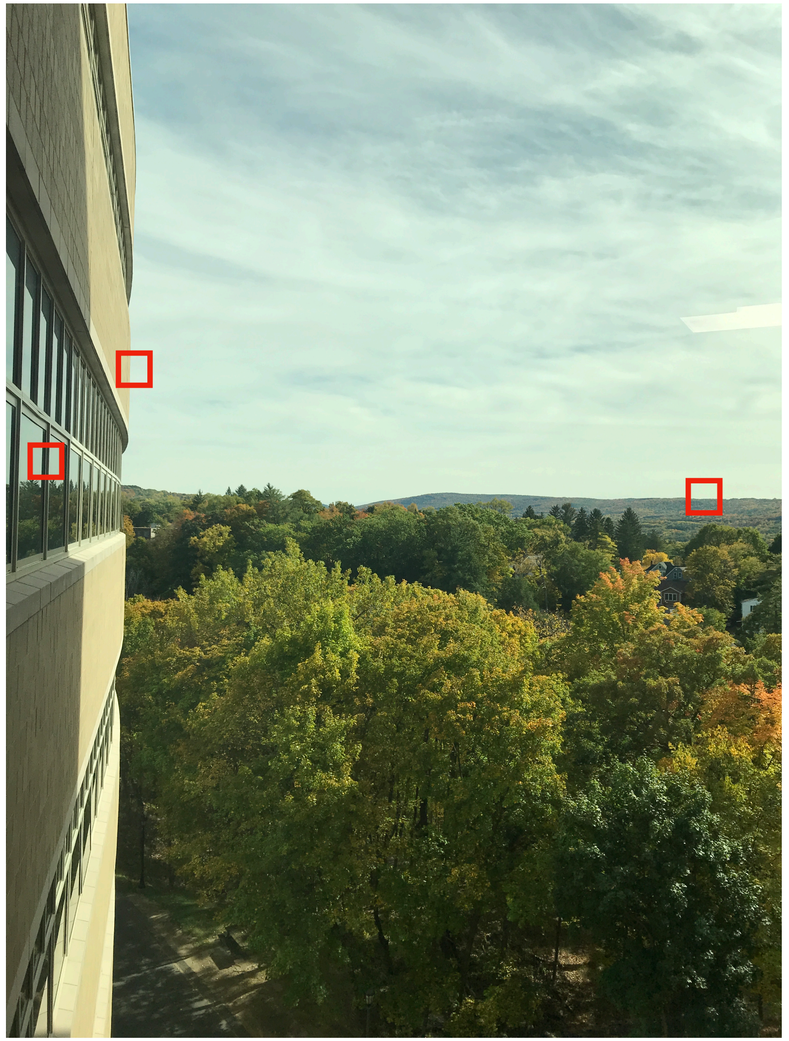




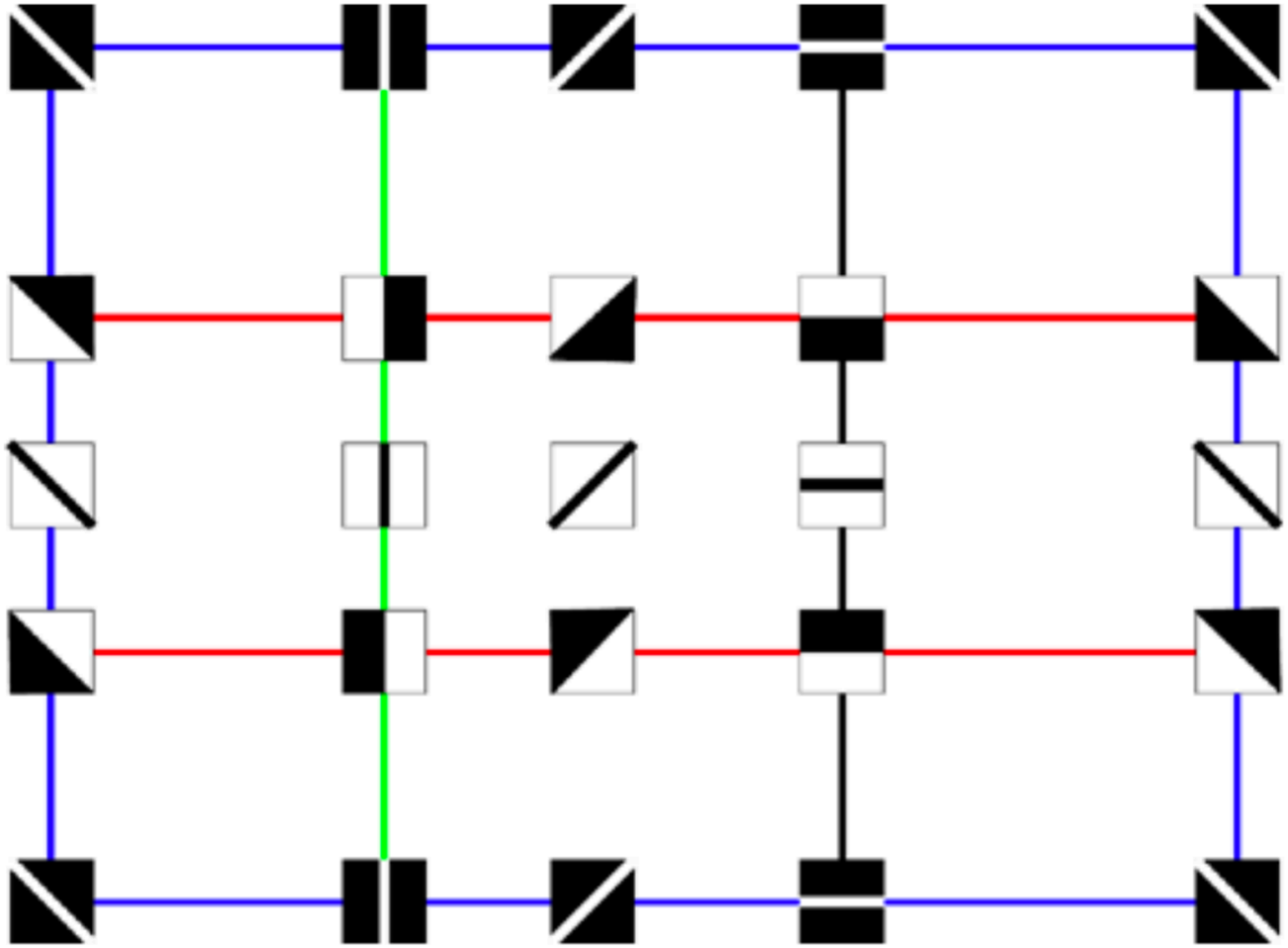
Vertical and Horizontal Patches



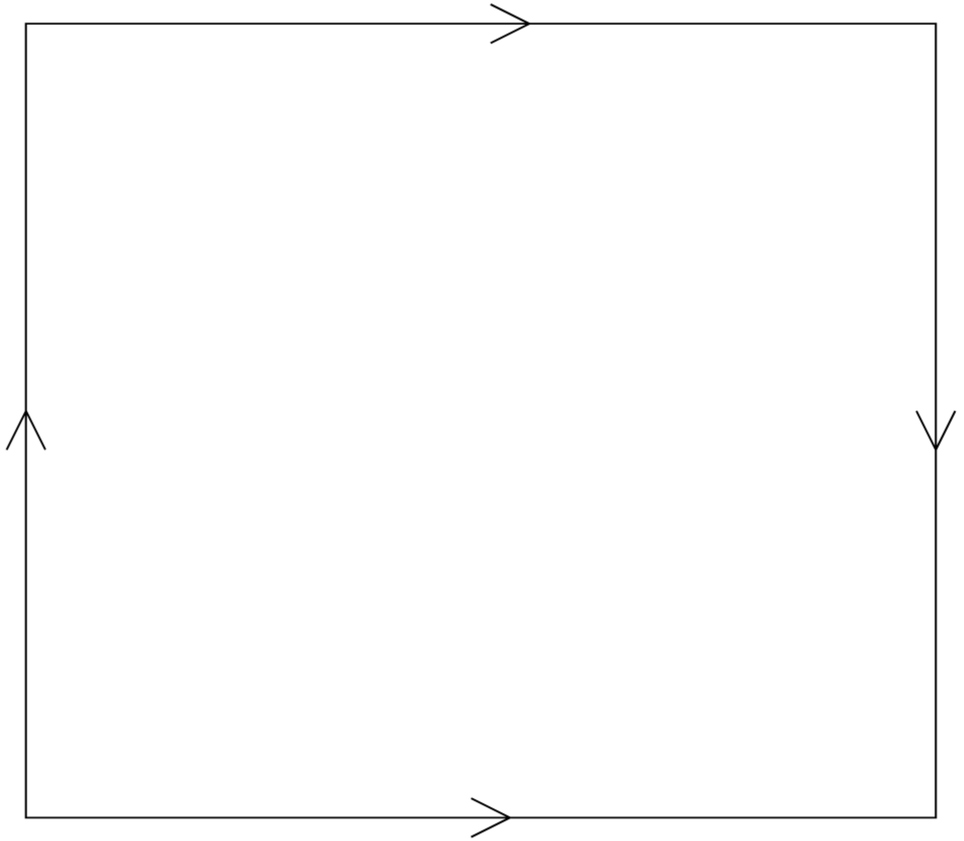


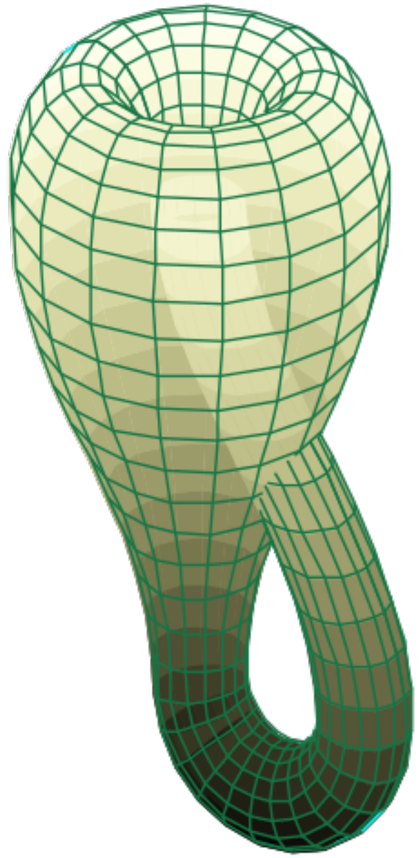


Gradients, Vertical Patches and Horizontal Patches



Klein Bottle!



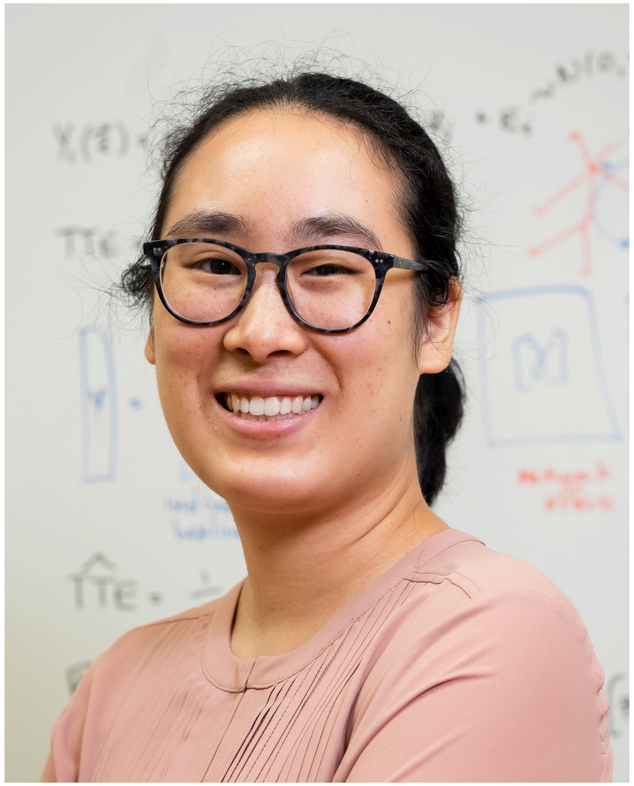


Act III

Chunyin and the Noise of Small Features

My Lovely Collaborators

Gennady Samorodnitsky

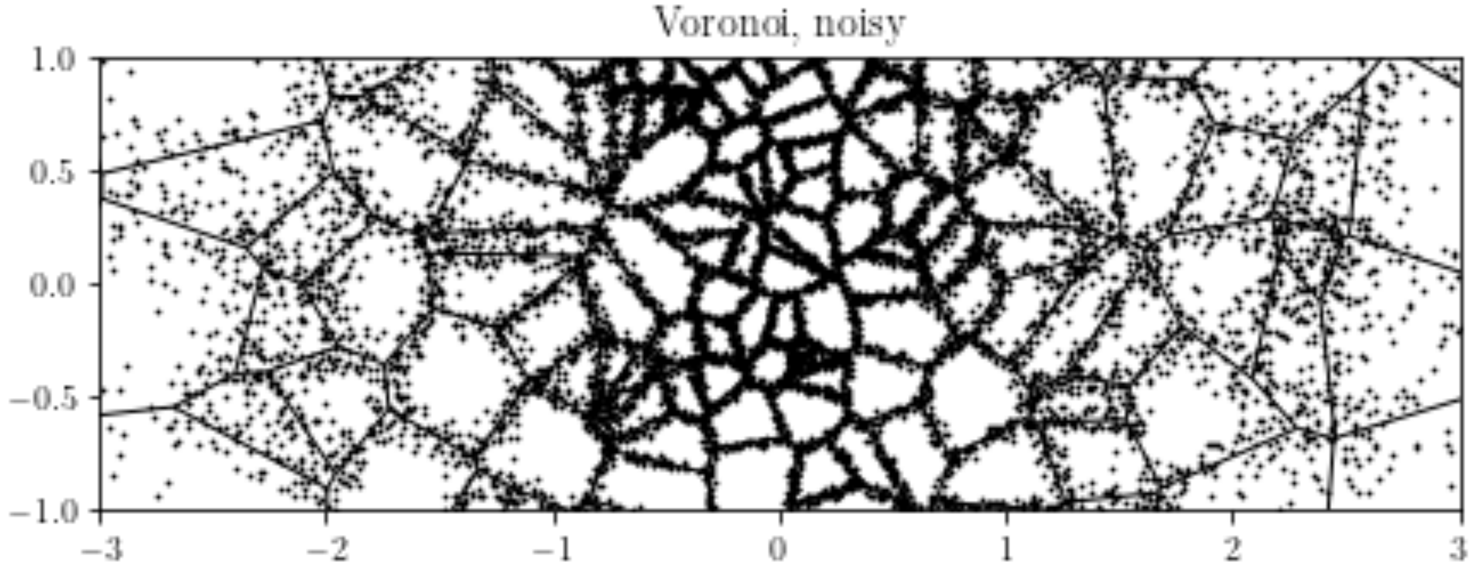


Christina Lee Yu

Andrey Yao

In the beginning…

- there was the data.



- And the data was non-parametric, and has voids, and noise is upon the face of the dataset.

Two Problems

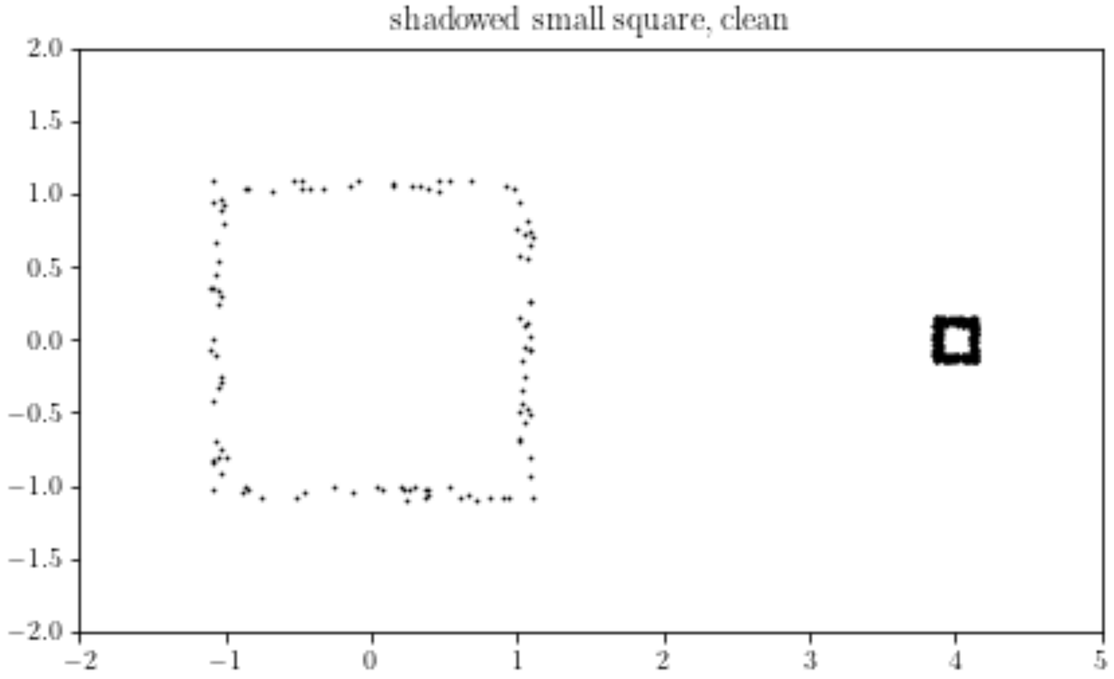
- Size
- Noise

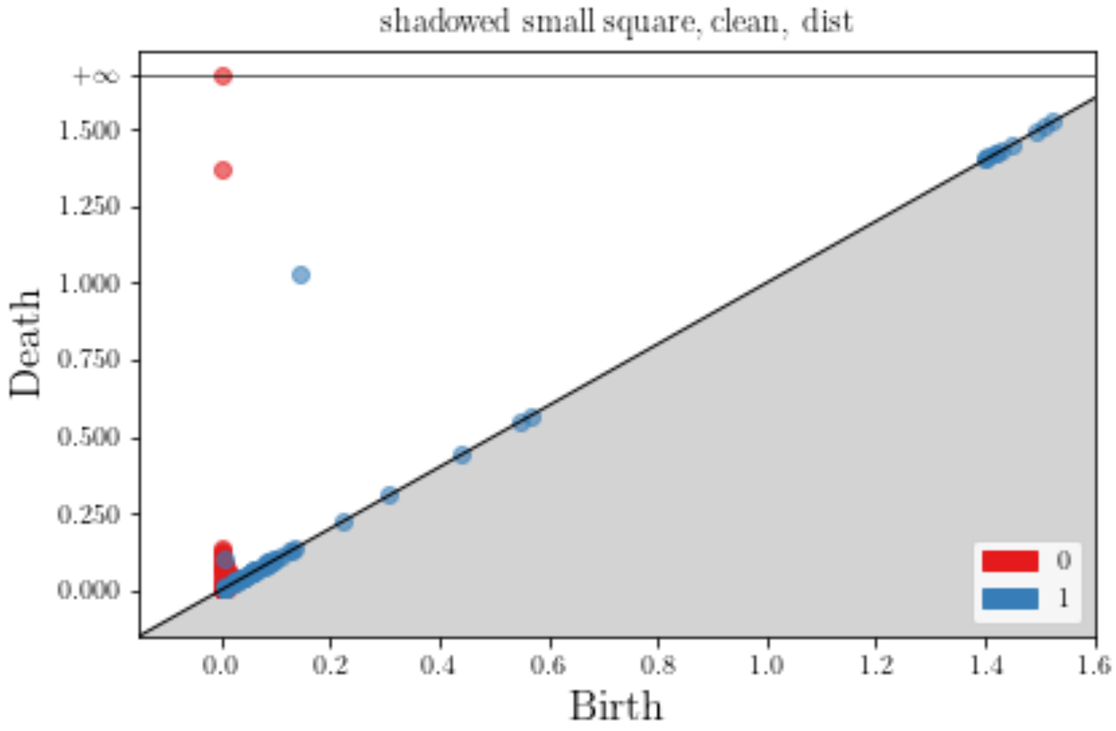
One solution

- Size
- Noise

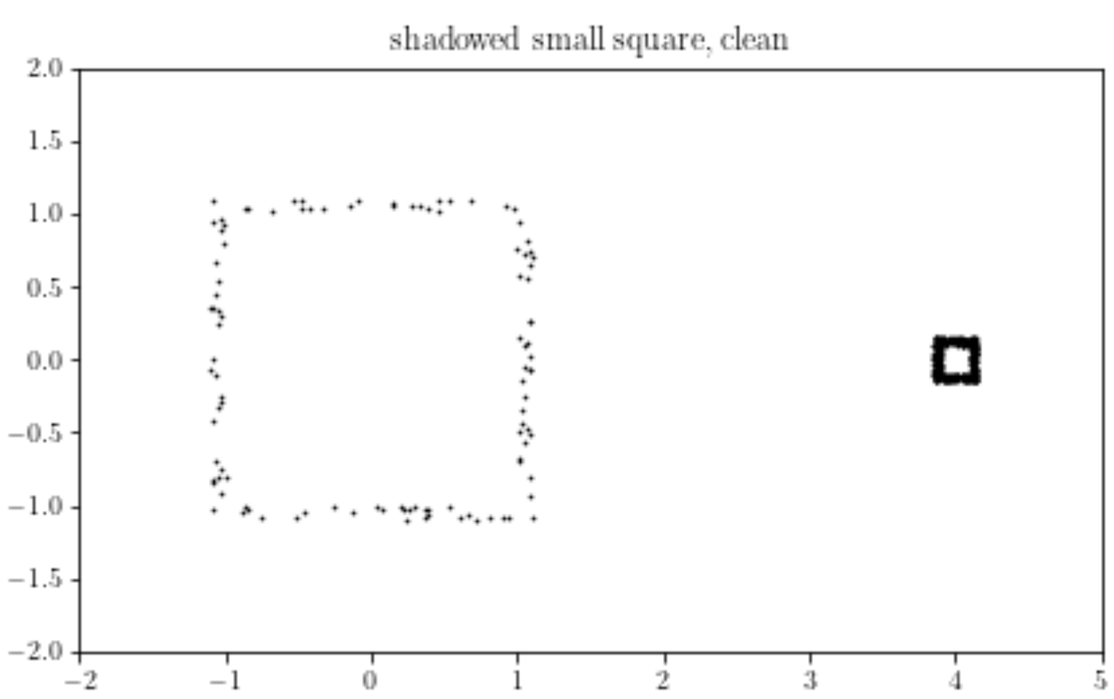
- statistical model that highlights small features
- with a robust estimator

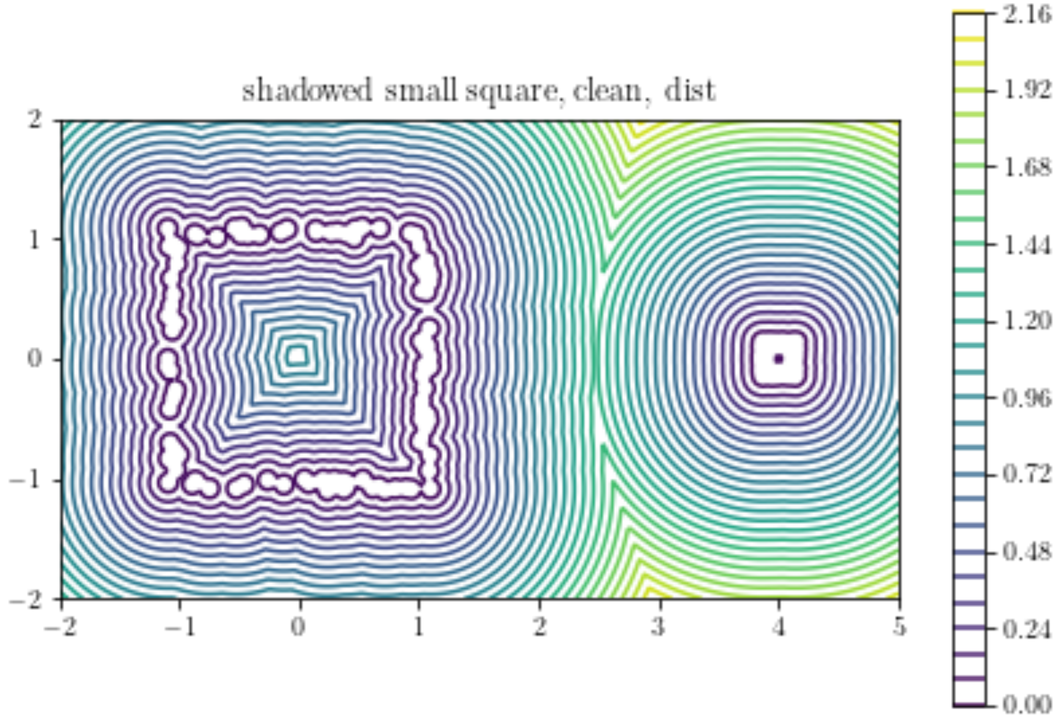
Size


shadowed small square, clean
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3
4
5


shadowed small square, clean, dist
+∞
1.500
1.250
1.000
0.750
0.500
0.250
0.000
Death
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Birth
0
1

Contour plot of min d(x, X_i)





Grow Balls Slooooooooooooooooooowly

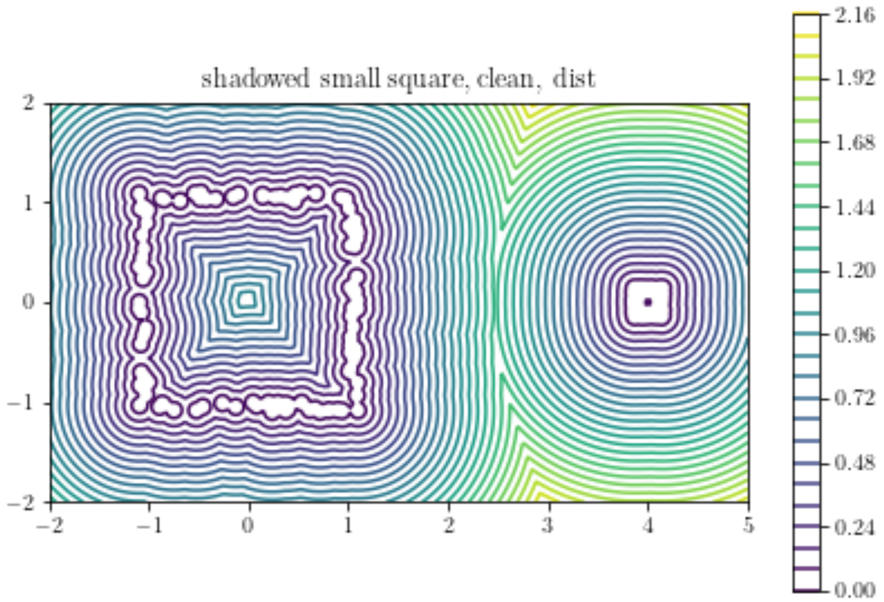
on the smaller square

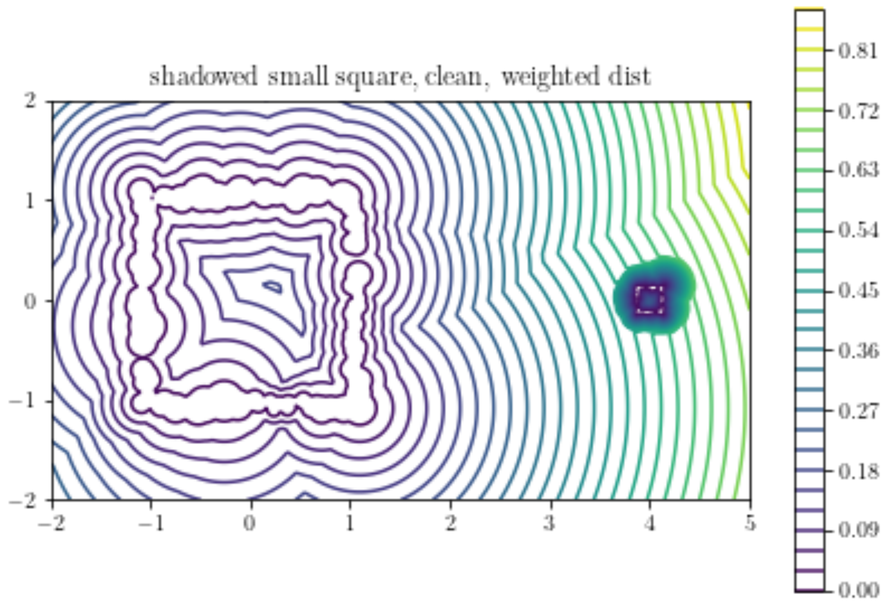
- Bell et al, 2019: growing balls at customized rates

Grow Balls Slooooooooooooooooooowly

on the smaller square

- rate = density$^{1/D}$

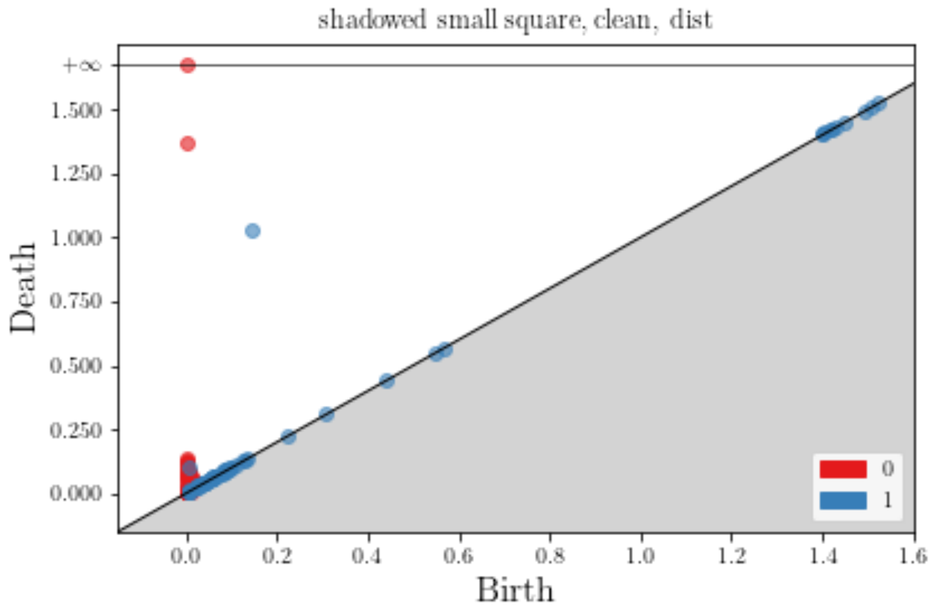


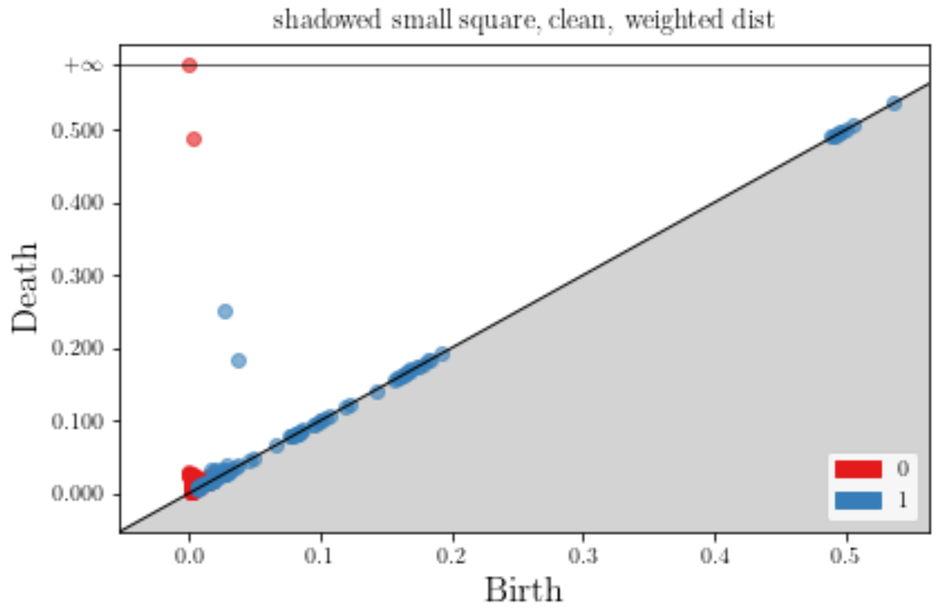


Grow Balls Slooooooooooooooooooowly

on the smaller square

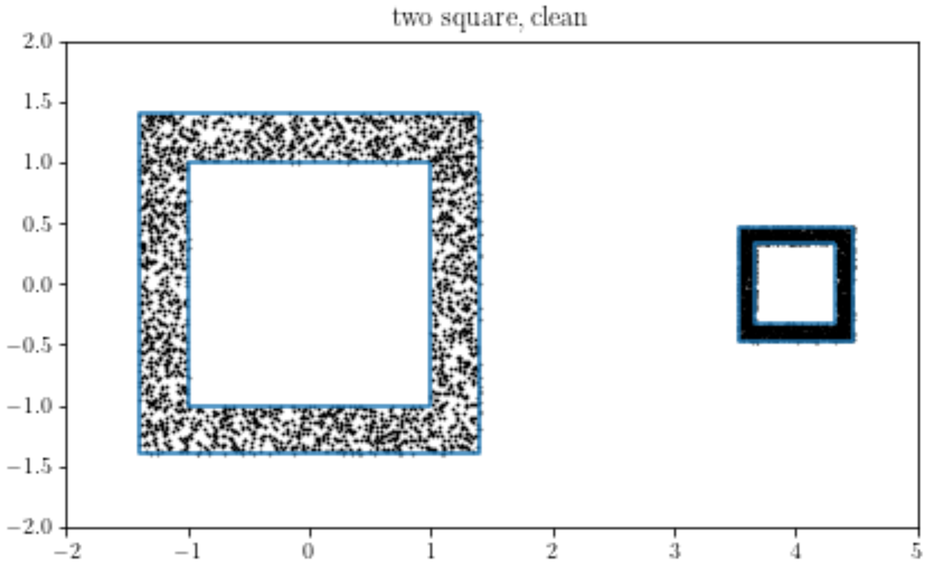
- rate = density$^{1/D}$

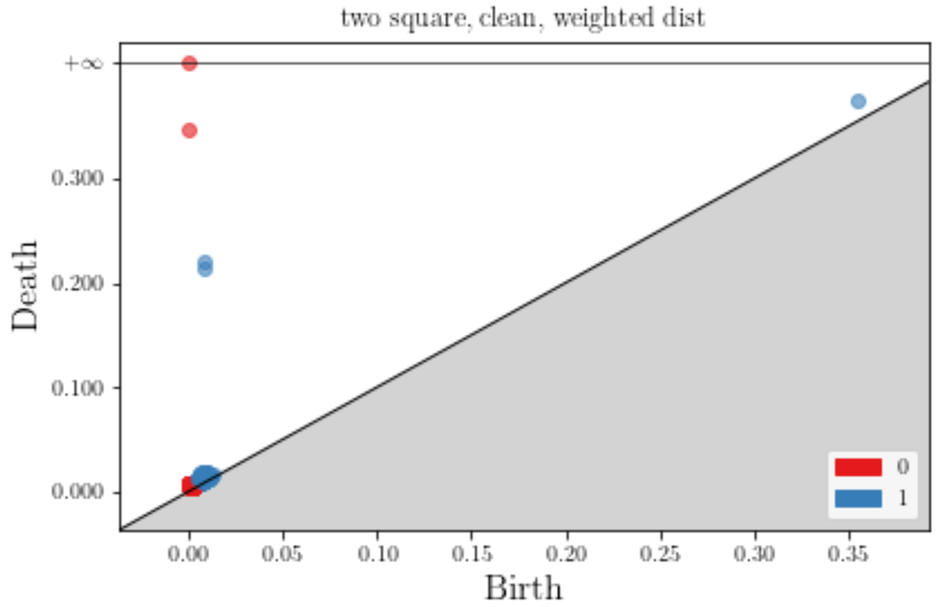




Why density$^{1/D}$?

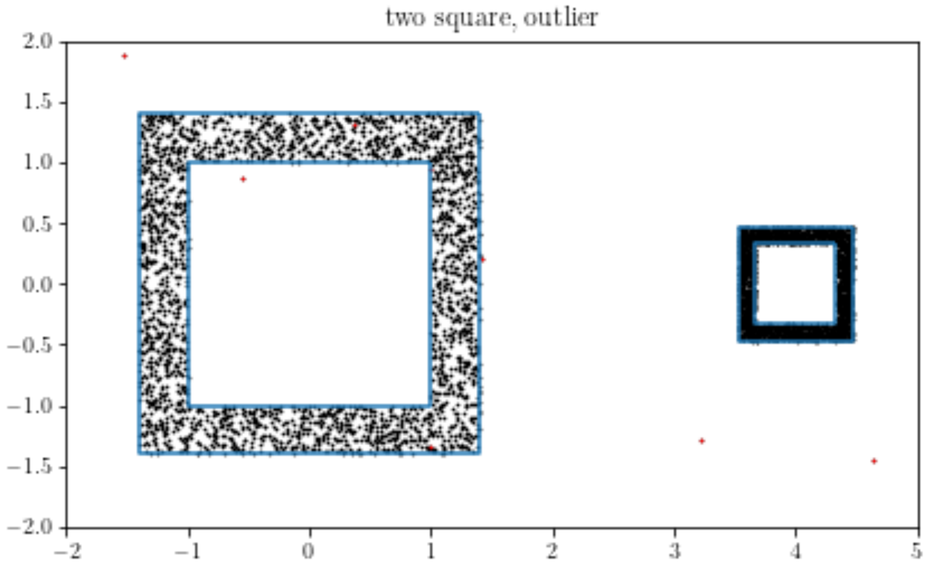
- Antman property: scaling —> same persistence diagrams





Noise

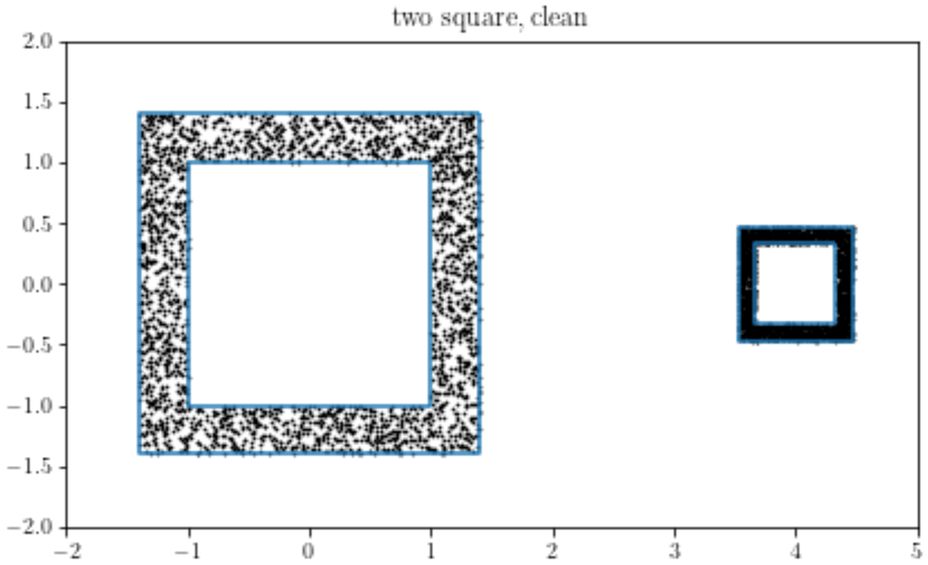
Outliers

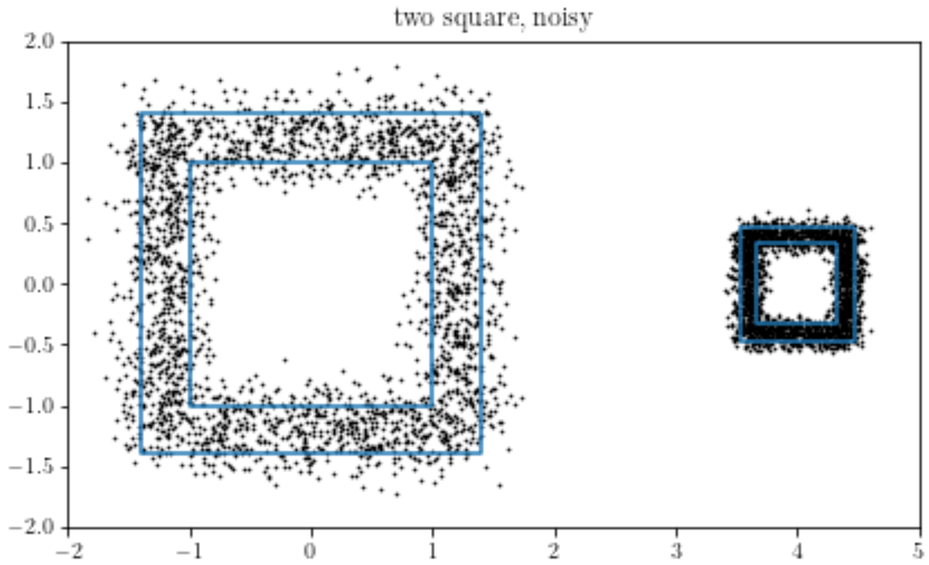

two square, outlier
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3
4
5


two square, outlier, dist
2
1
0
−1
−2
−2
−1
0
1
2
3
4
5
1.62
1.44
1.26
1.08
0.90
0.72
0.54
0.36
0.18
0.00

Additive Noise

- Gaussian noise fills the plane!





Known Problem, Known Solution

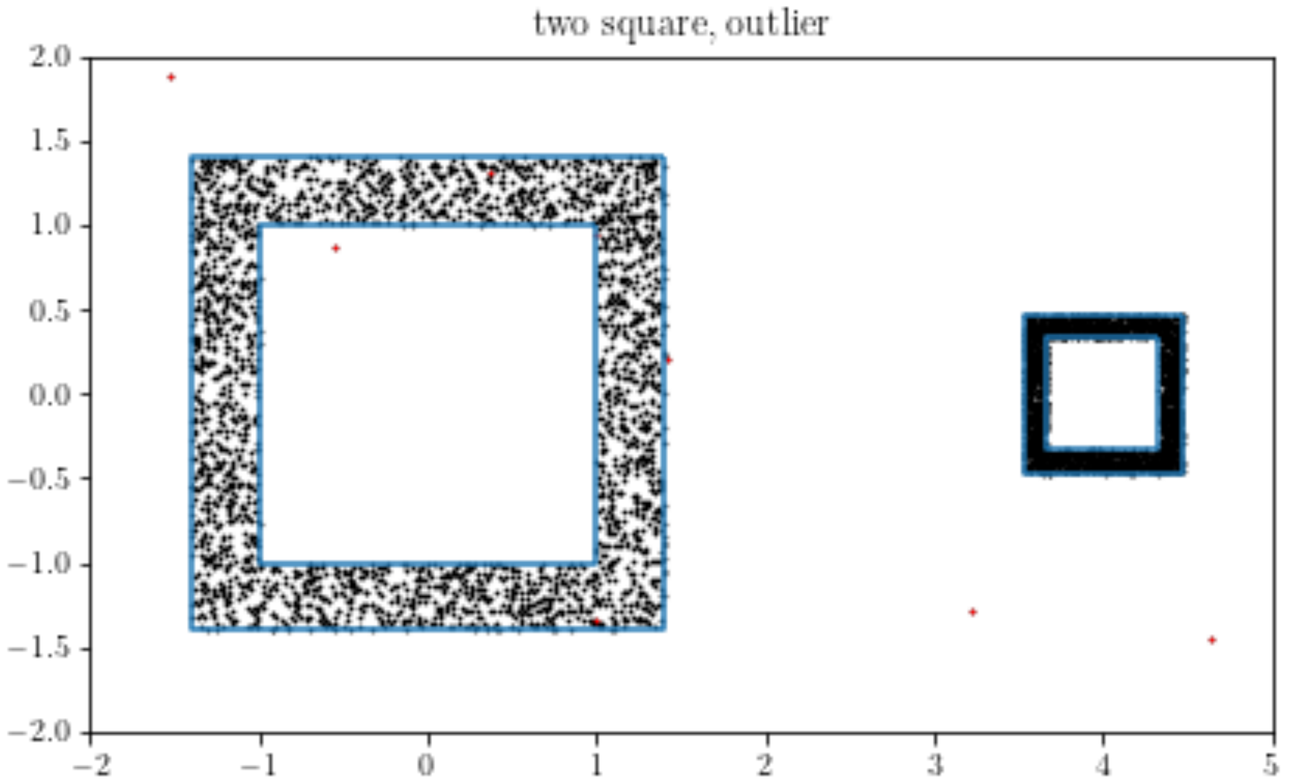
- problem: can be corrupted by 1 single data point
- solution: distance-to-measure
  - wait for more balls, and take average
  - Chazal et al (2011), Chazal et al (2018)

Distance-to-Measure (DTM)

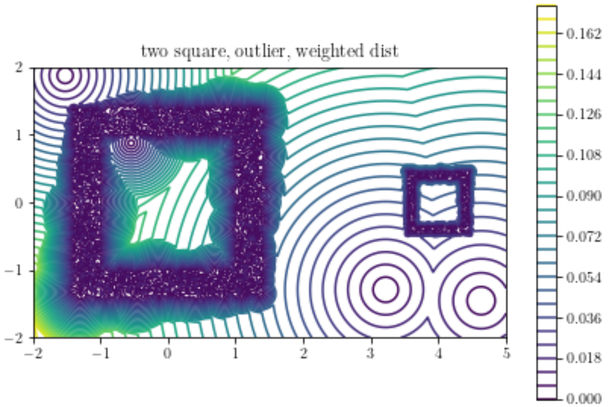
- Chazal et al (2011), Chazal et al (2018)

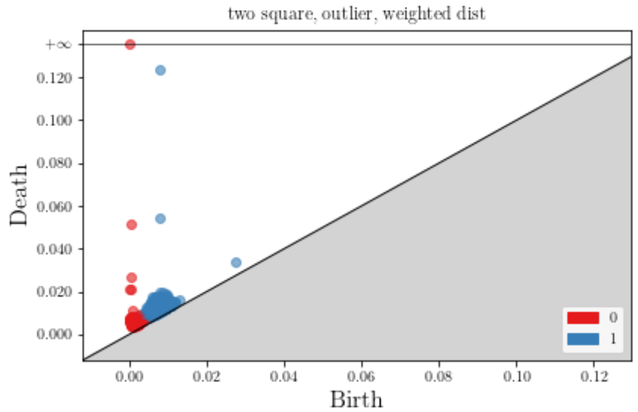
Robust Density-Aware Distance (RDAD)

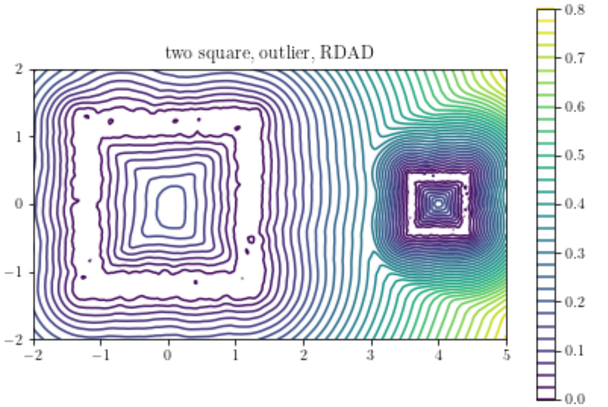
Outlier

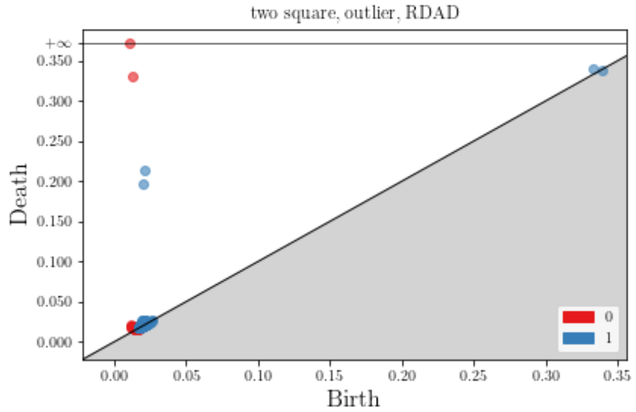


Weighted distance v.s. RDAD

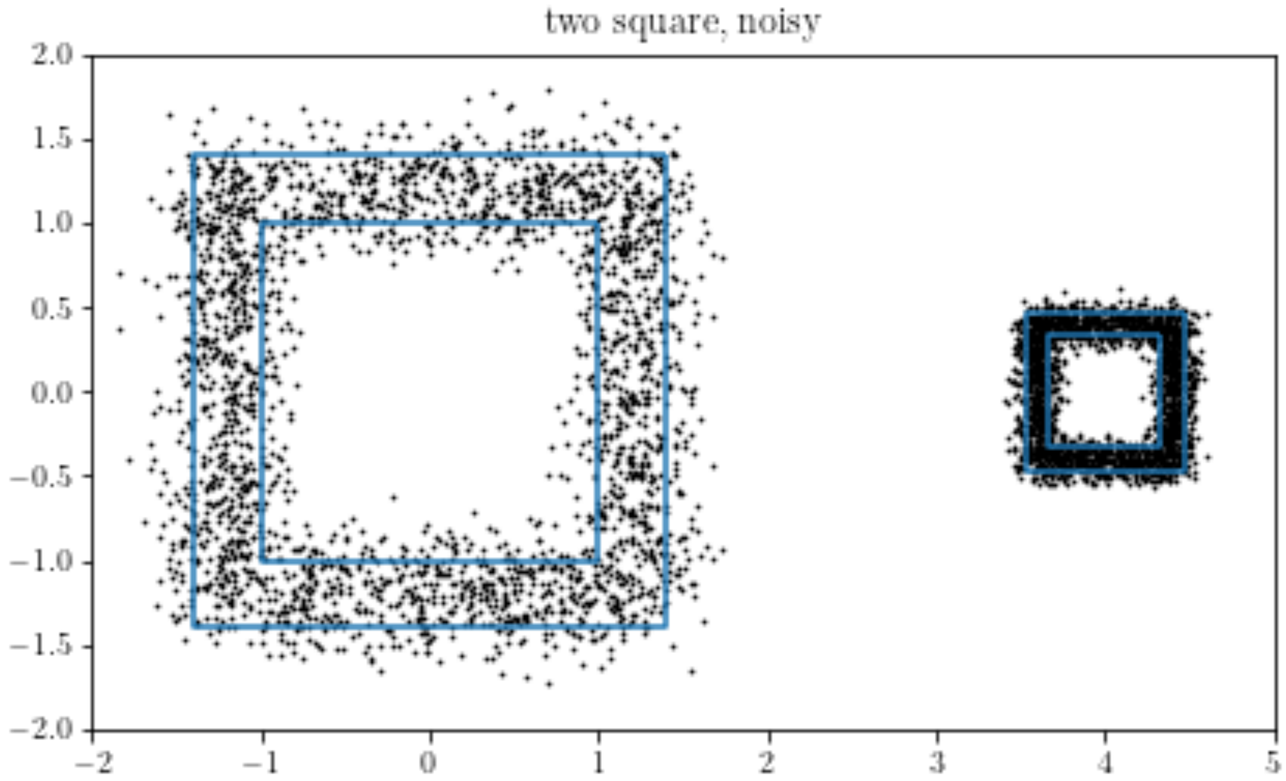




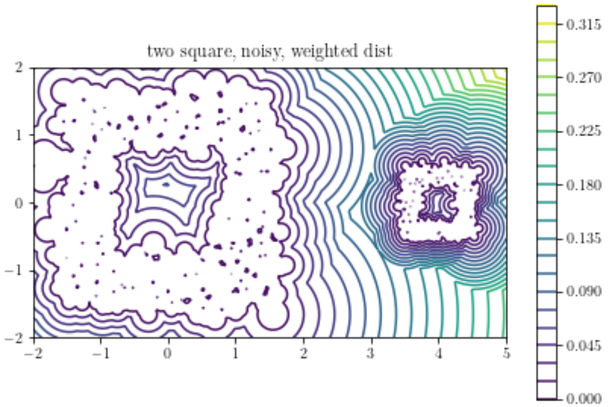


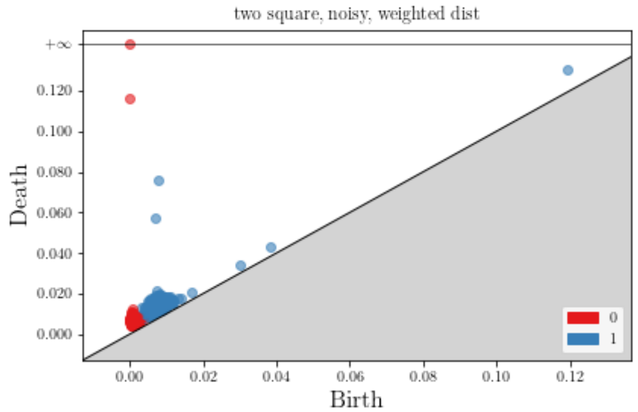


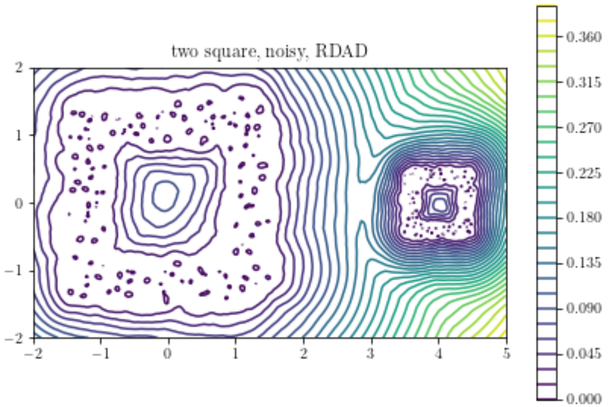
Additive noise

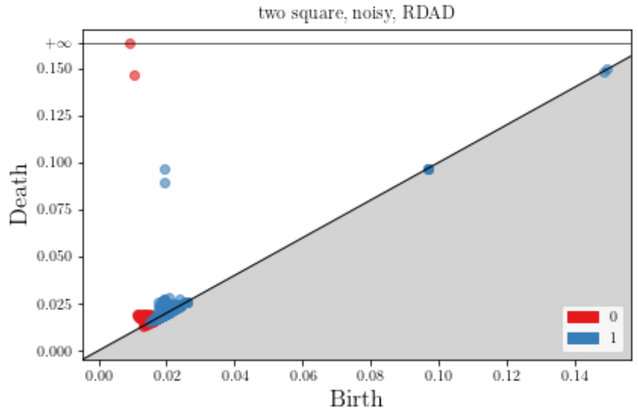


Weighted distance v.s. RDAD

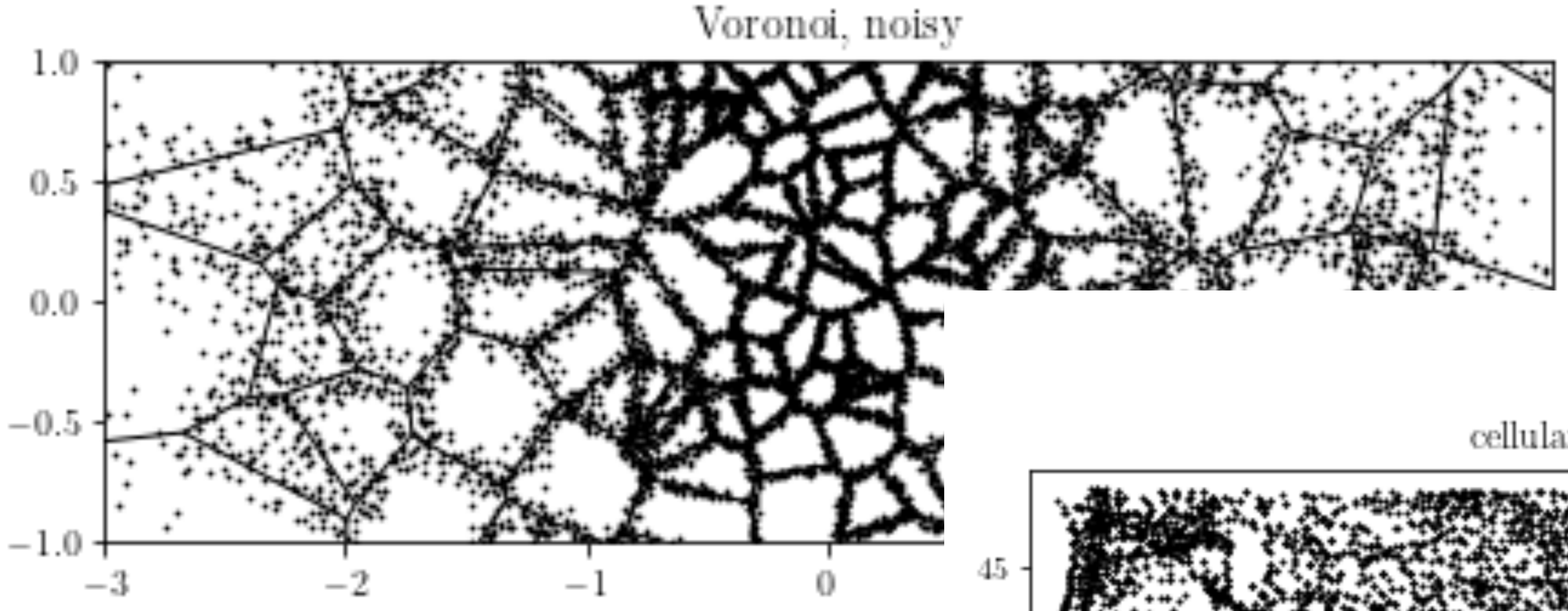


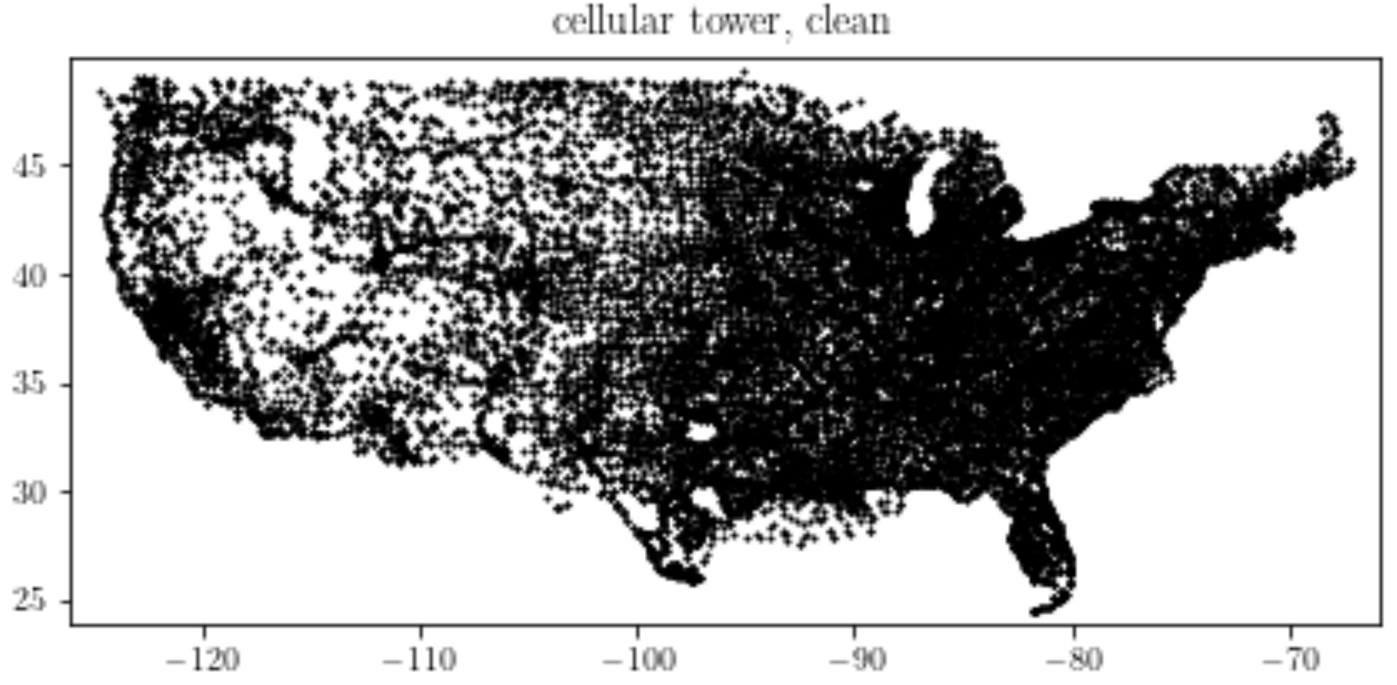






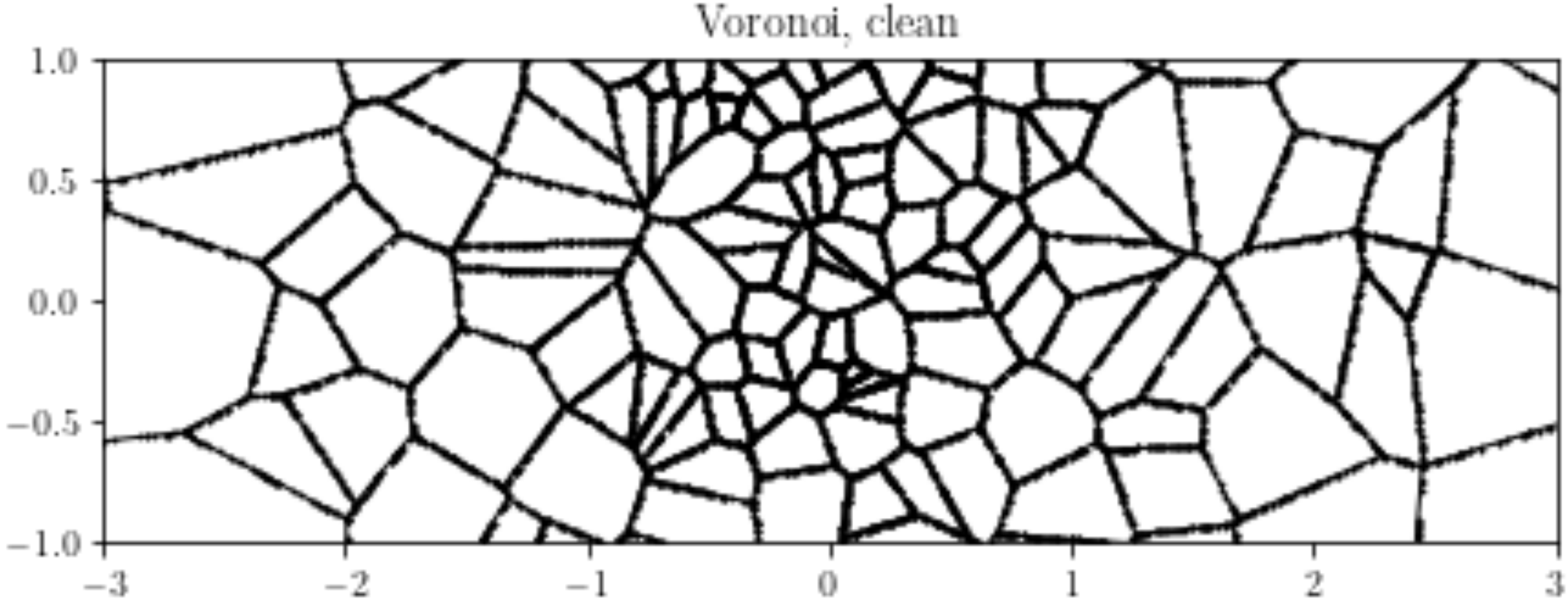
Simulations


Voronoi, noisy
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0

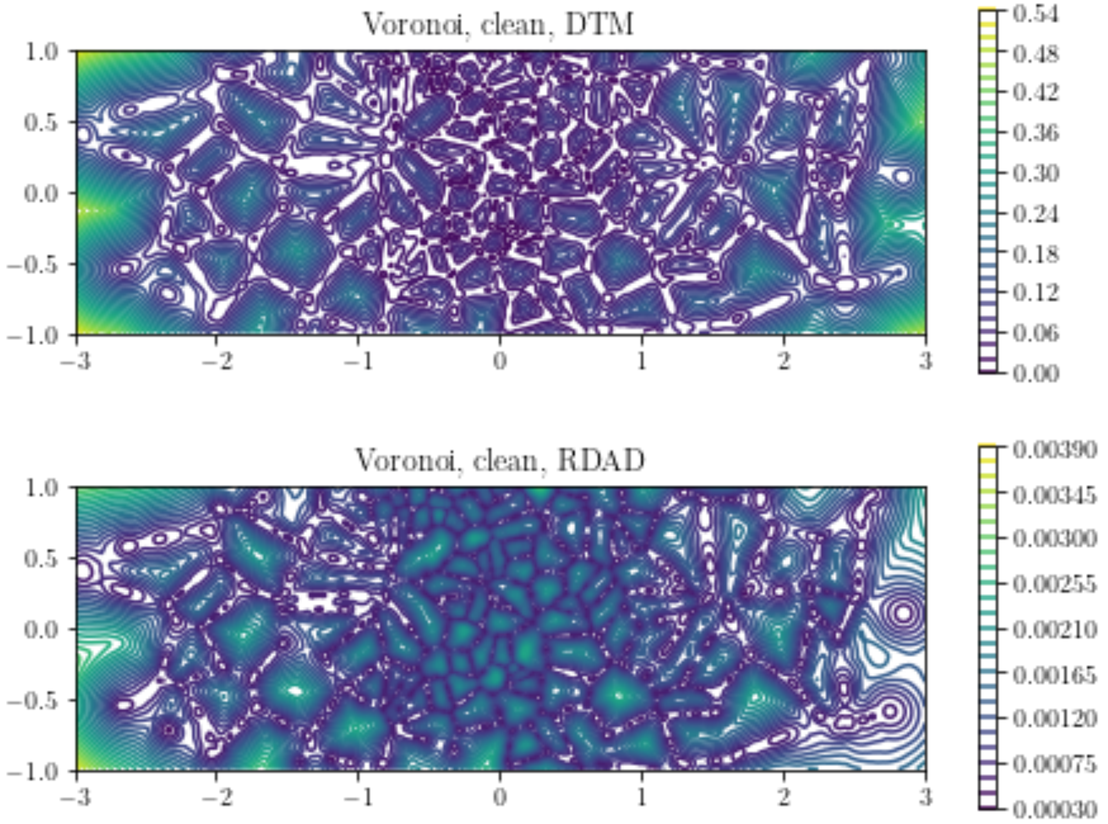

cellular tower, clean
45
40
35
30
25
−120
−110
−100
−90
−80
−70

clean Voronoi

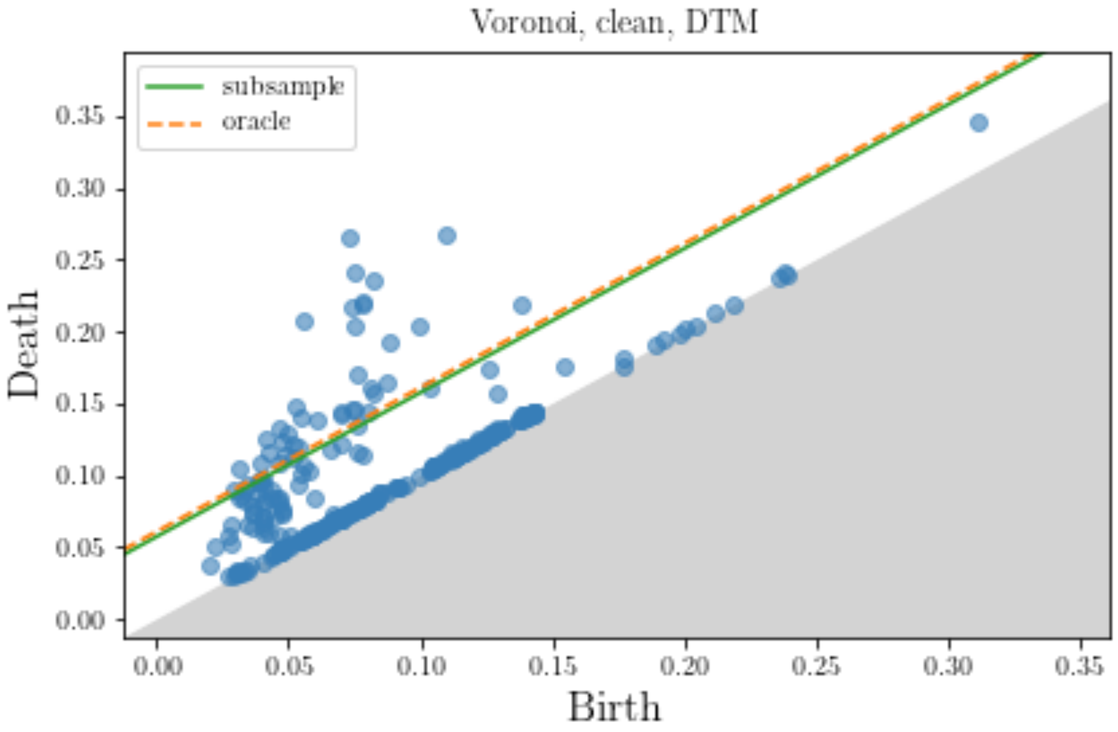
each cell has the same number of points

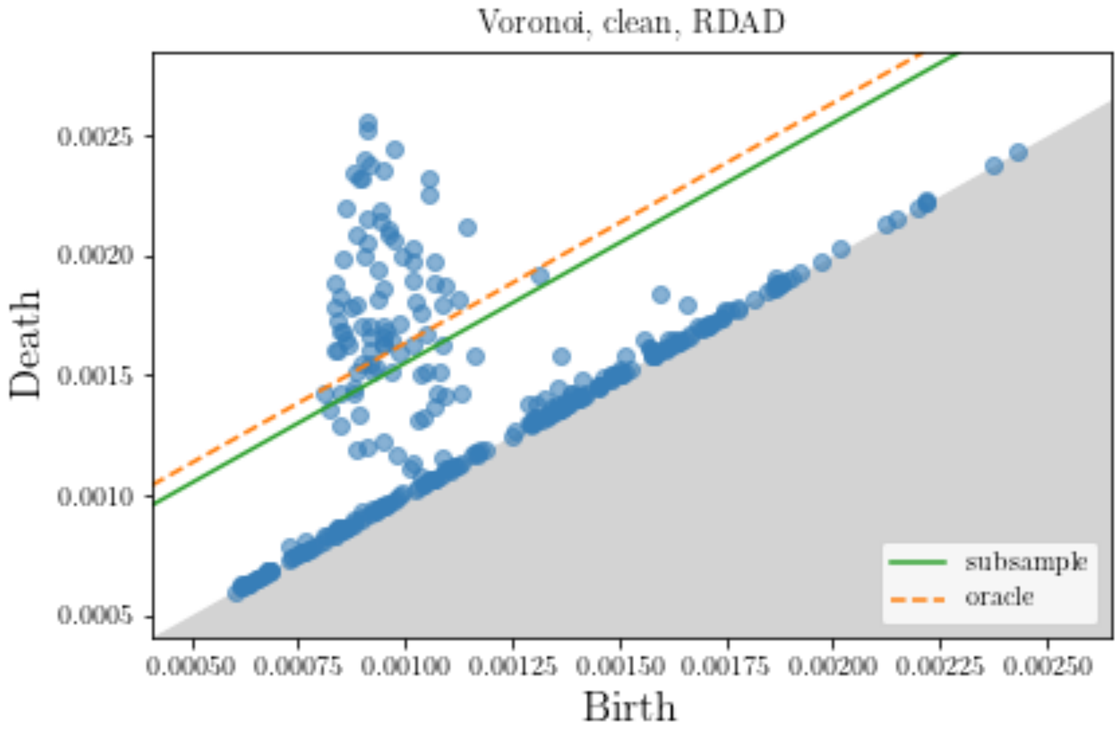


DTM and RDAD

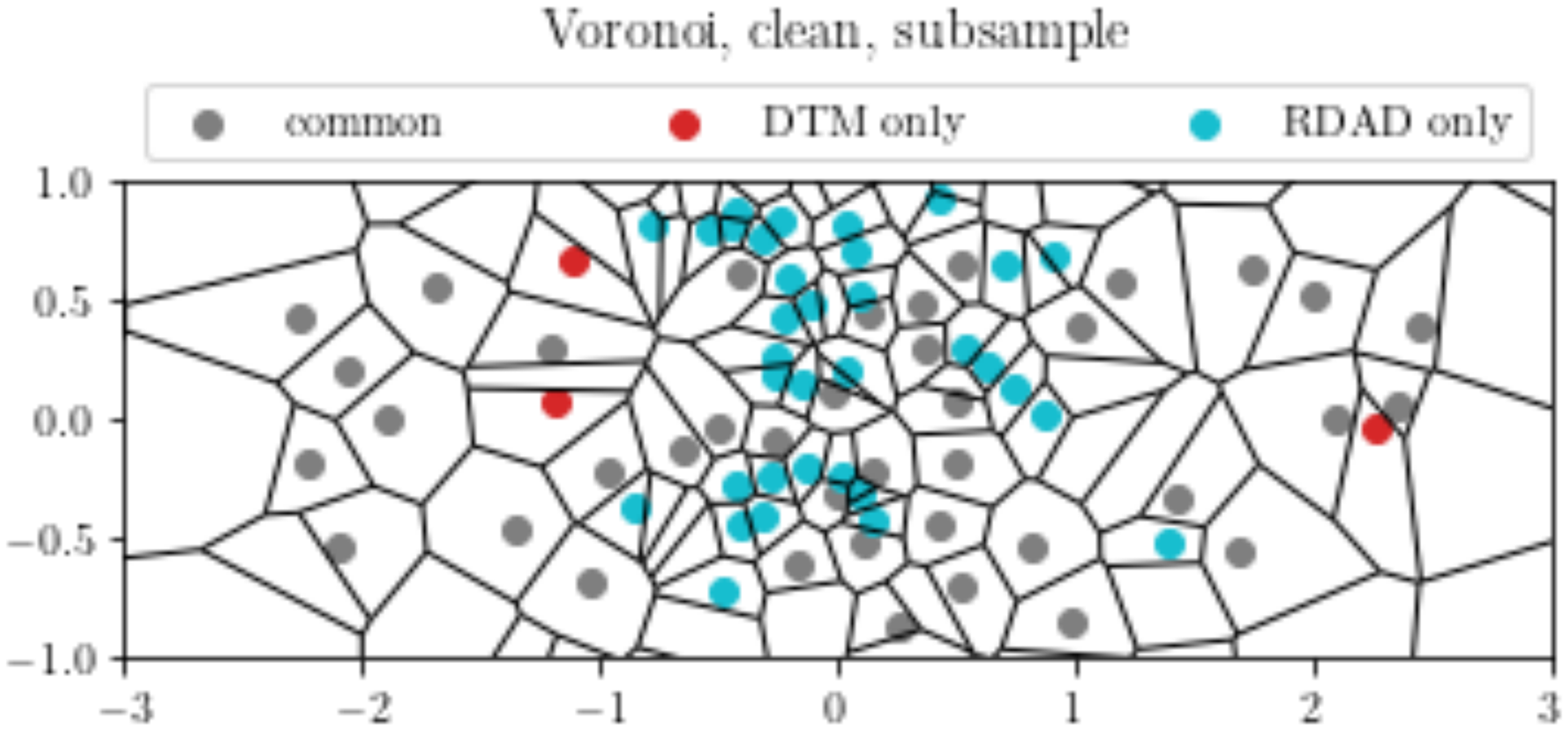

Voronoi, clean, DTM
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
0.54
0.48
0.42
0.36
0.30
0.24
0.18
0.12
0.06
0.00
Voronoi, clean, RDAD
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
0.00390
0.00345
0.00300
0.00255
0.00210
0.00165
0.00120
0.00075
0.00030

DTM and RDAD

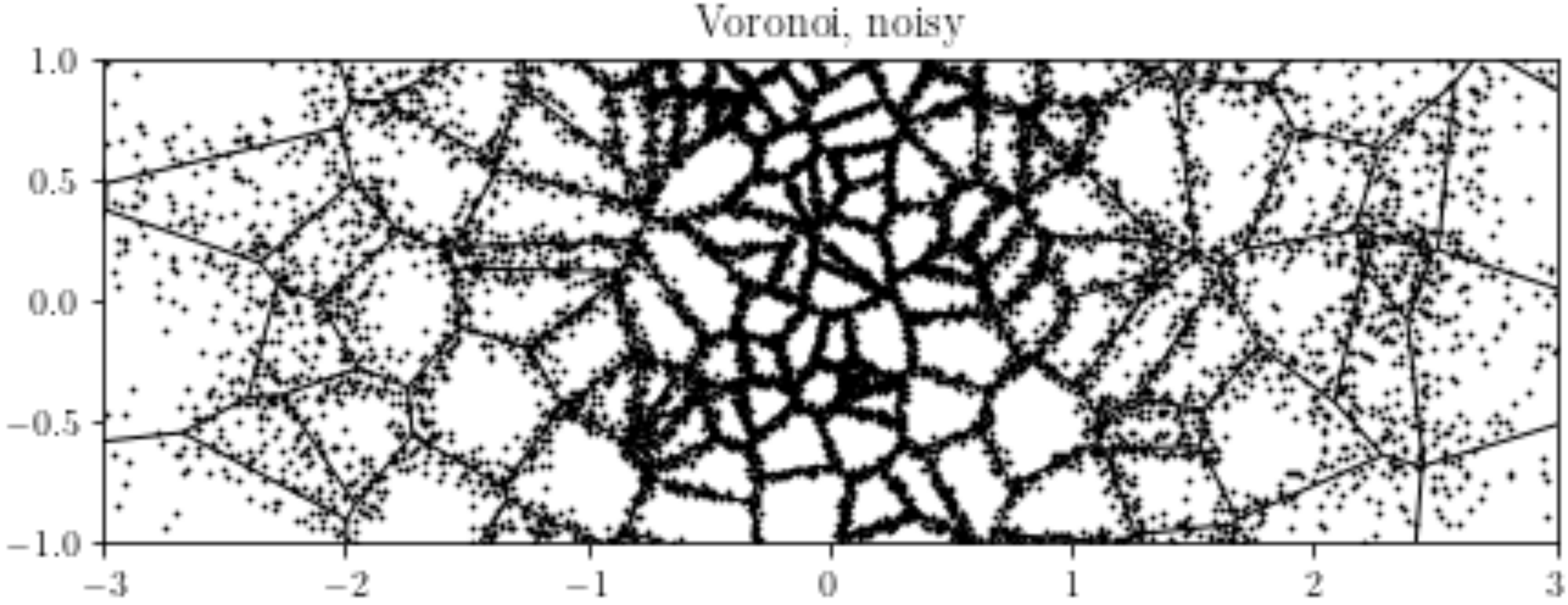

Voronoi, clean, DTM
subsample
oracle
Death
Birth
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35


Voronoi, clean, RDAD
subsample
oracle
Death
Birth
0.0005
0.0010
0.0015
0.0020
0.0025
0.00050
0.00075
0.00100
0.00125
0.00150
0.00175
0.00200
0.00225
0.00250

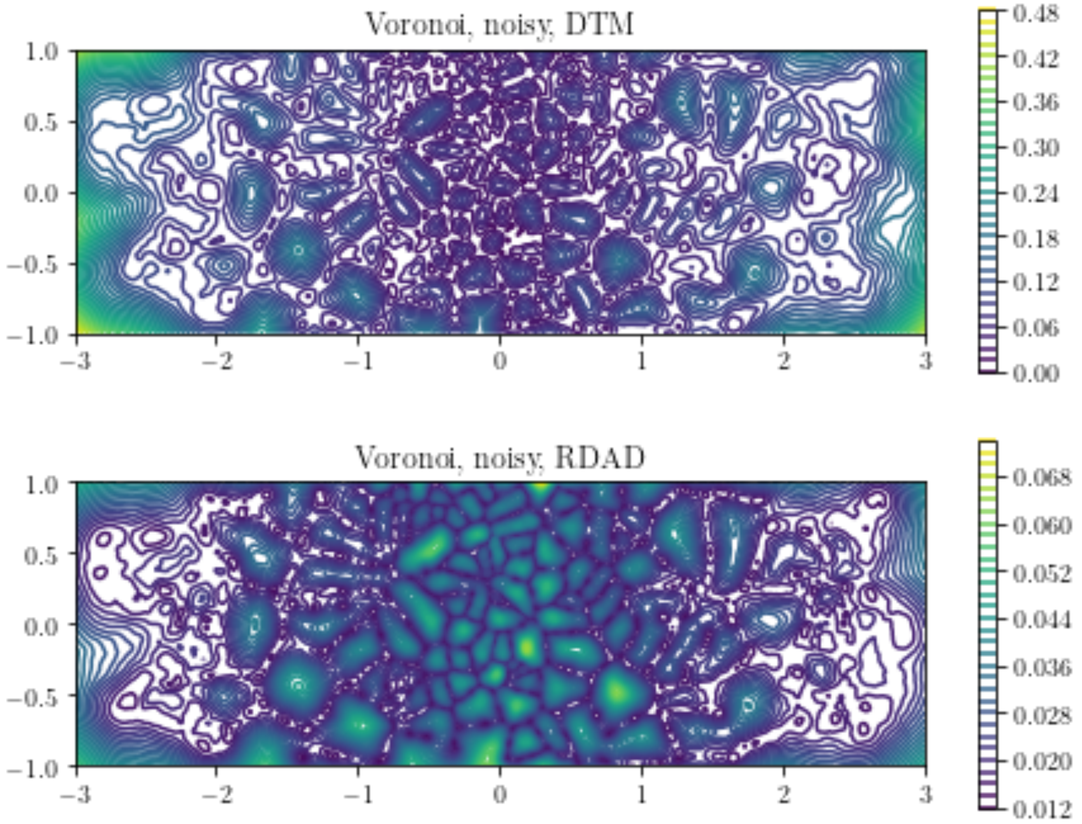
DTM and RDAD


Voronoi, clean, subsample
common
DTM only
RDAD only
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3

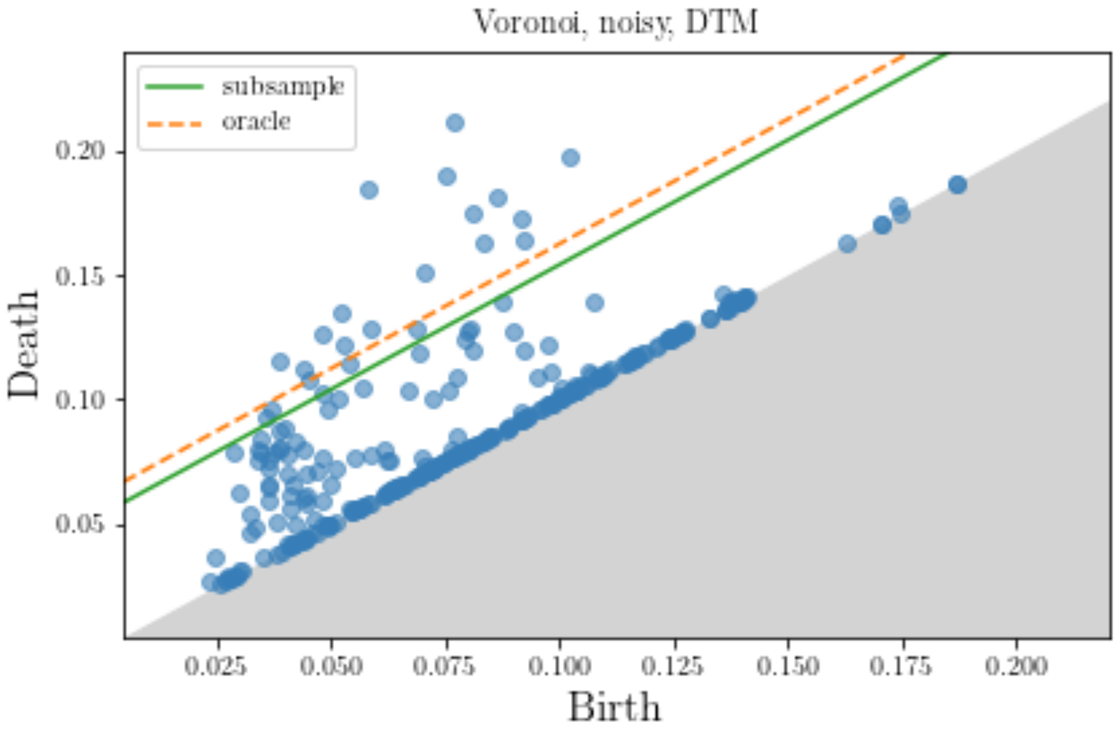
Noisy Voronoi

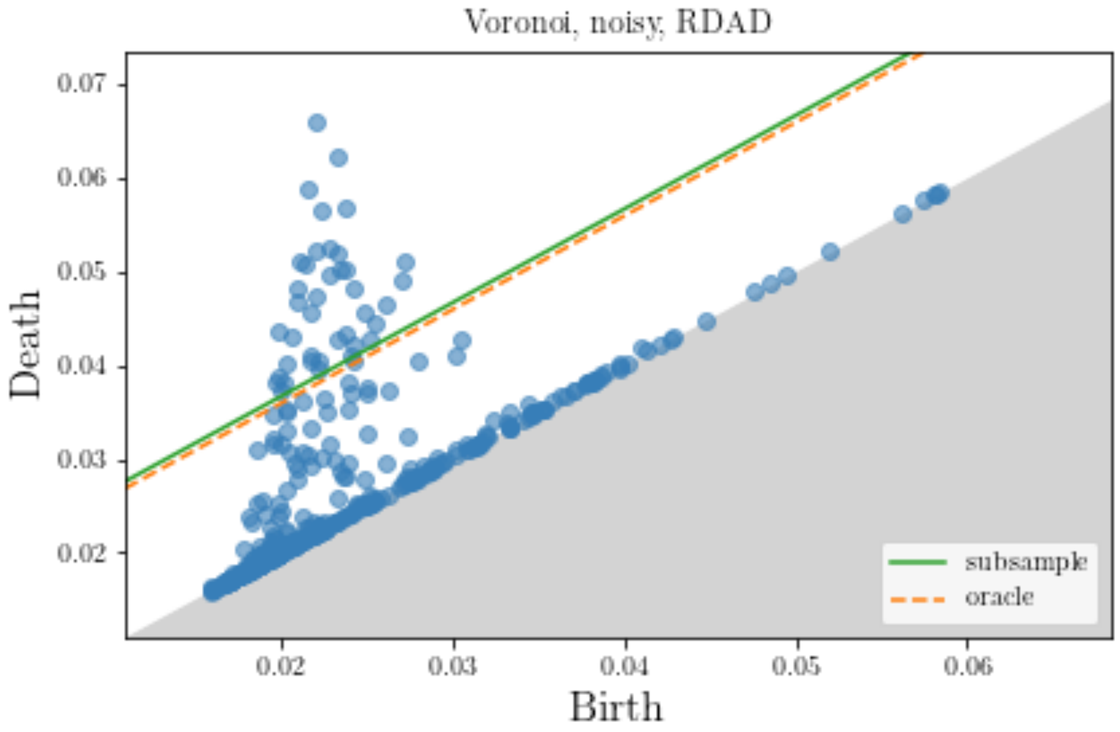


DTM and RDAD

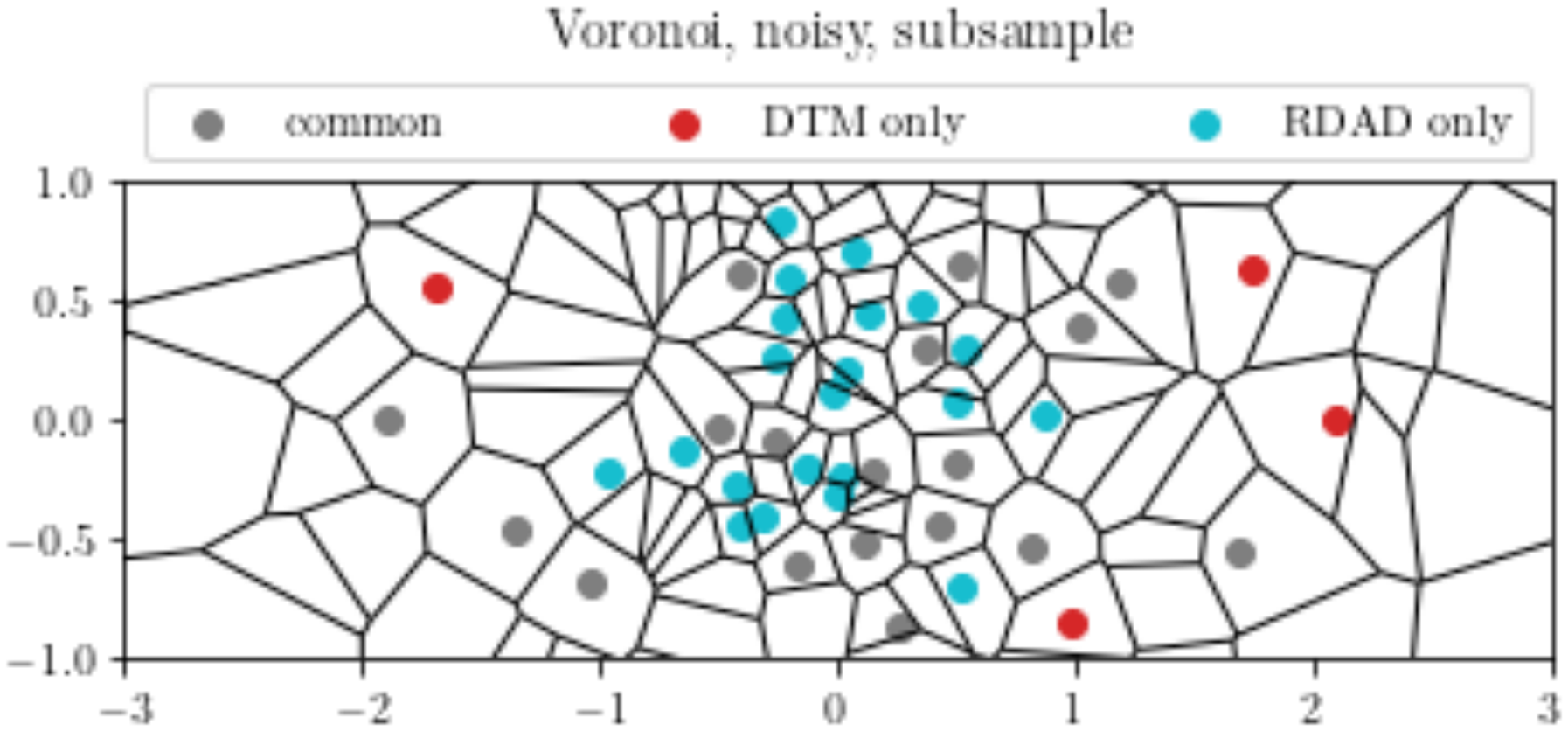

Voronoi, noisy, DTM
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
0.48
0.42
0.36
0.30
0.24
0.18
0.12
0.06
0.00
Voronoi, noisy, RDAD
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3
0.068
0.060
0.052
0.044
0.036
0.028
0.020
0.012

DTM and RDAD


Voronoi, noisy, DTM
subsample
oracle
Death
Birth
0.05
0.10
0.15
0.20
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200


Voronoi, noisy, RDAD
subsample
oracle
Death
Birth
0.02
0.03
0.04
0.05
0.06
0.07

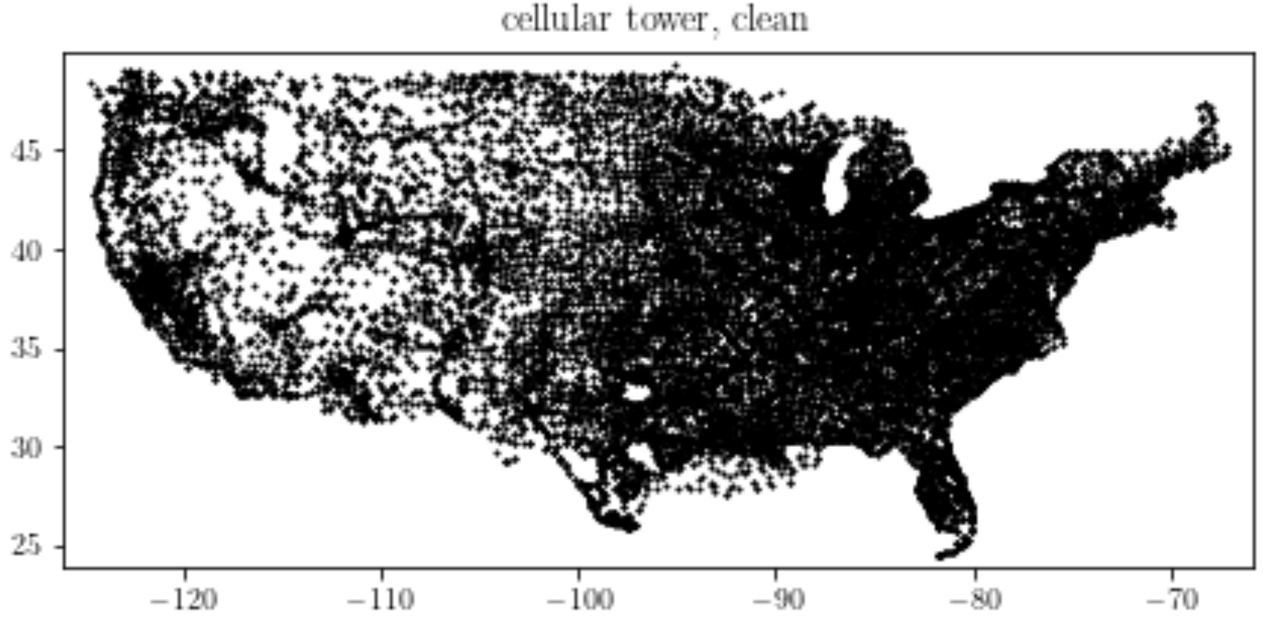
DTM and RDAD


Voronoi, noisy, subsample
common
DTM only
RDAD only
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3

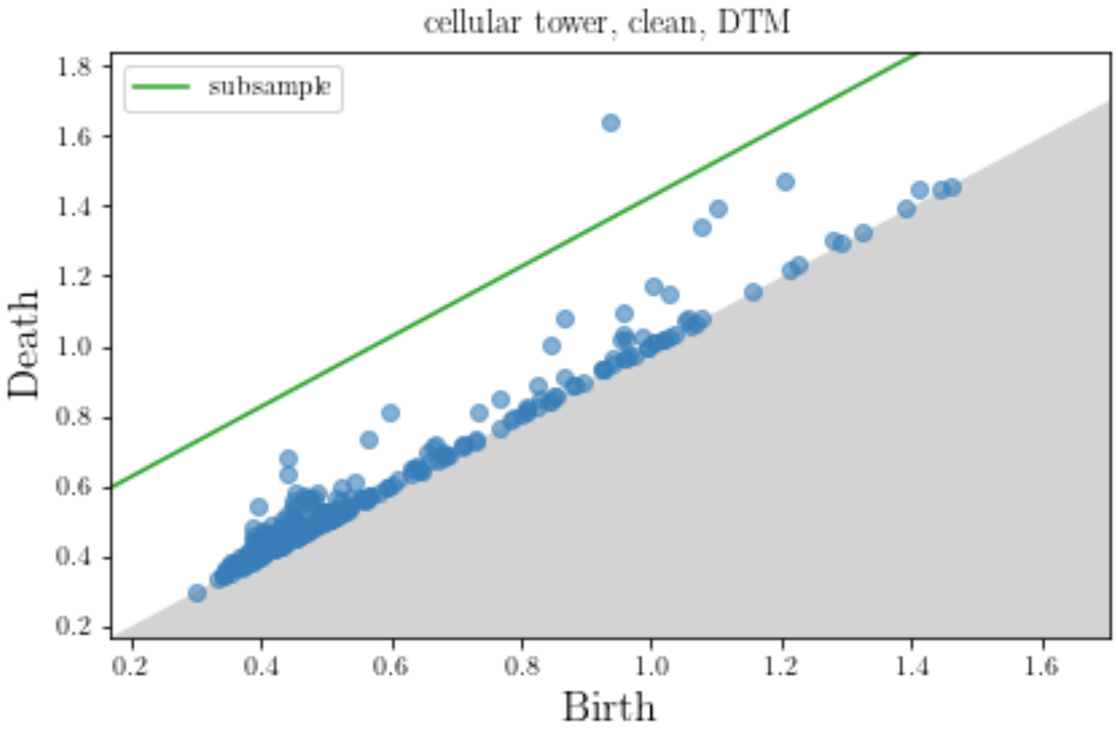
Cellular Towers

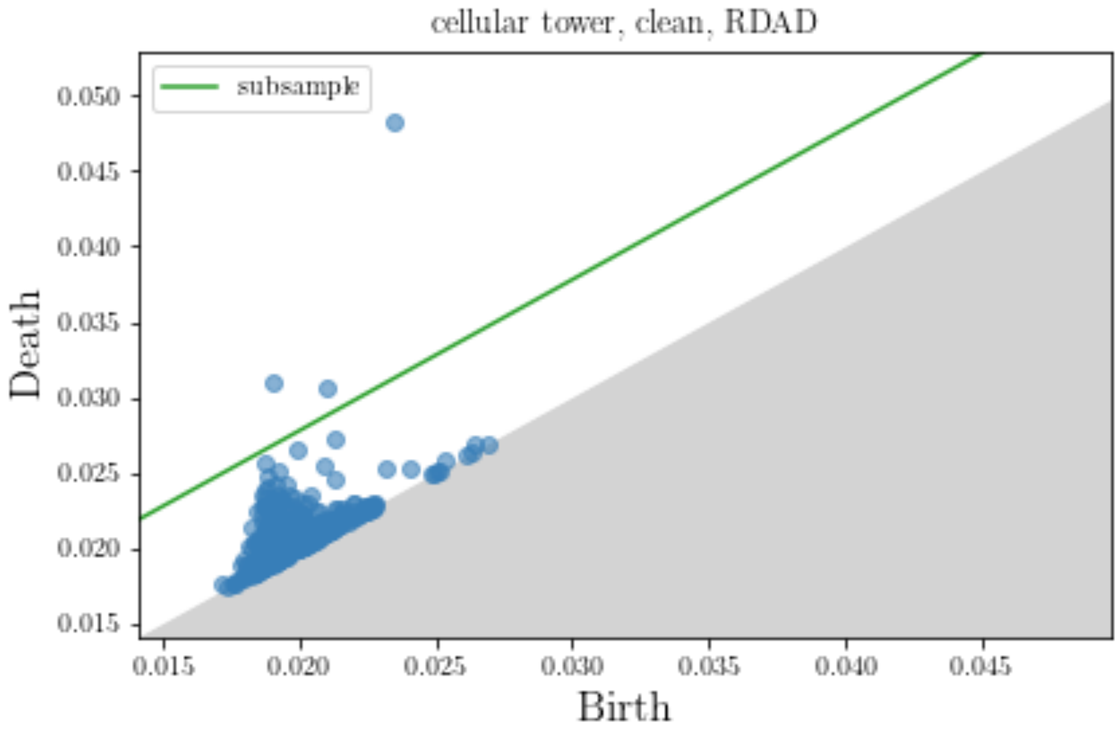
Cellular Towers

(HIFLD, 2021)

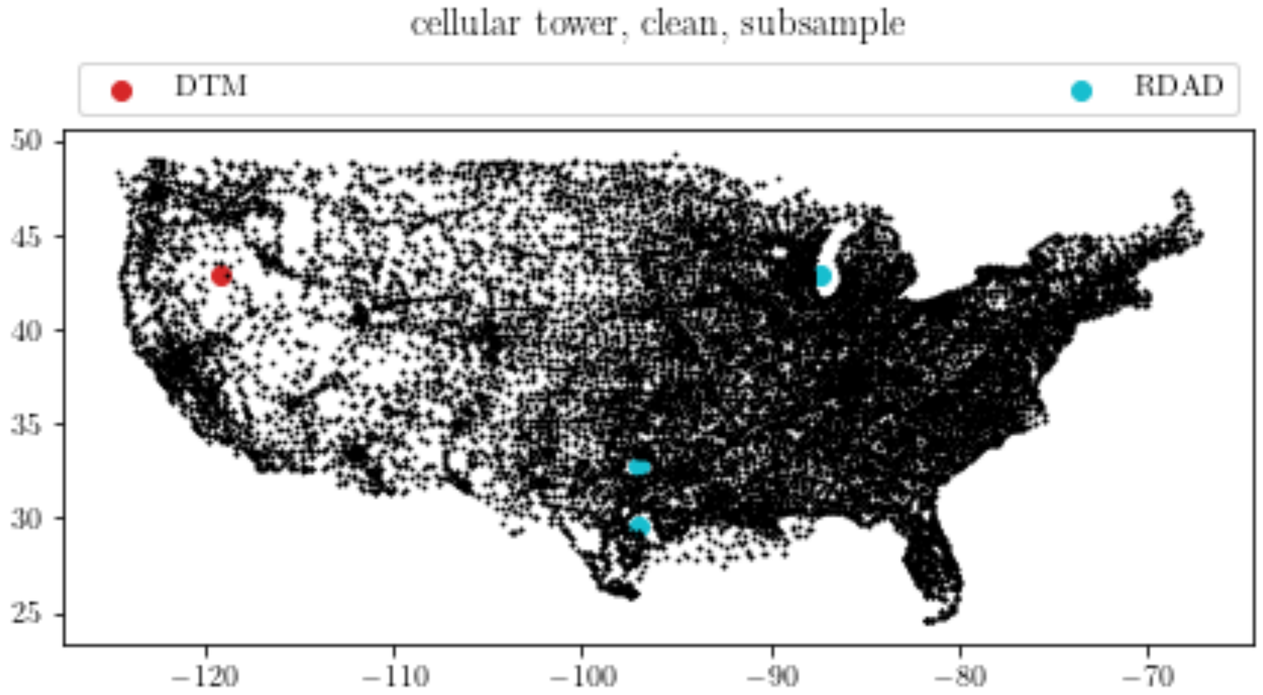


DTM and RDAD


cellular tower, clean, DTM
subsample
Death
Birth
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6


cellular tower, clean, RDAD
subsample
Death
Birth
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.015
0.020
0.025
0.030
0.035
0.040
0.045

Cellular Towers



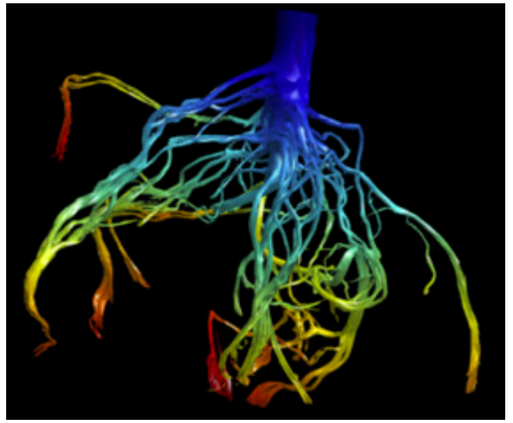
Epilogue:
The Beginning of the End

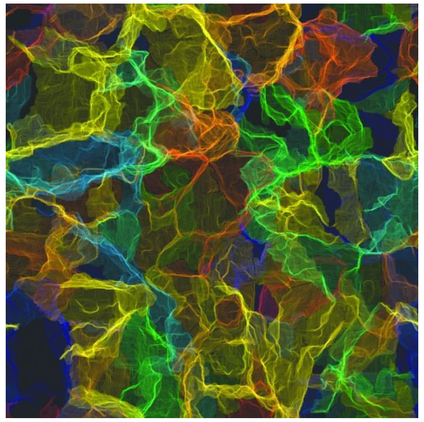
The End of the Beginning

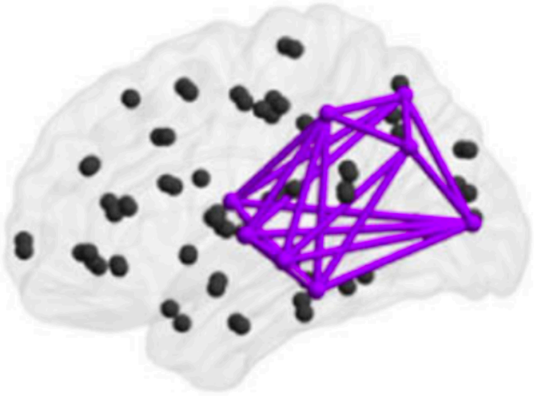
What Research Opportunities are Out There?

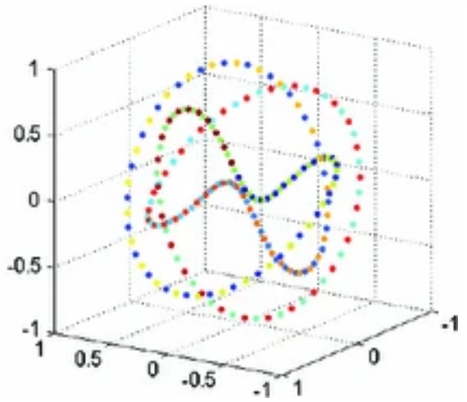
Applications

- image analysis (Gabrielsson and Carlsson, 2019)
- cosmology (Aragon-Calvo and M. A. and Szalay, 2012)
- neuroscience (Sizemore et al, 2018)
- material science (Ichinomiya et al, 2017)
- botany (Li et al, 2017)
- time series (Perea, 2015)
- and many others…



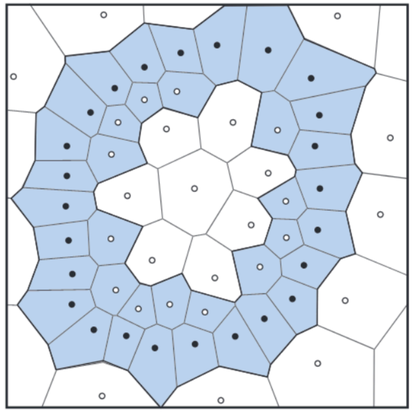


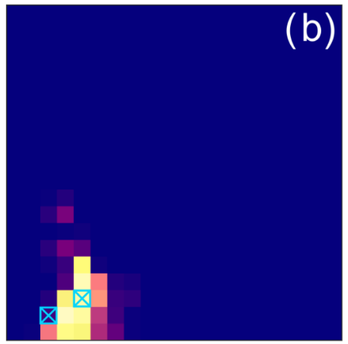


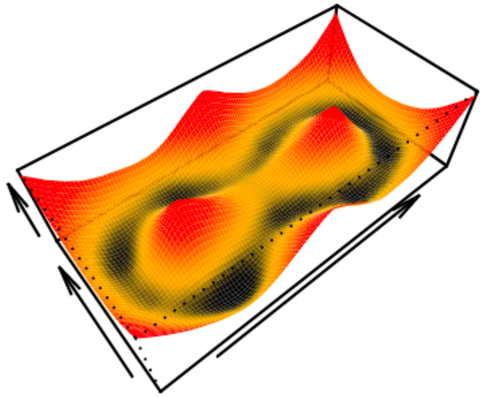


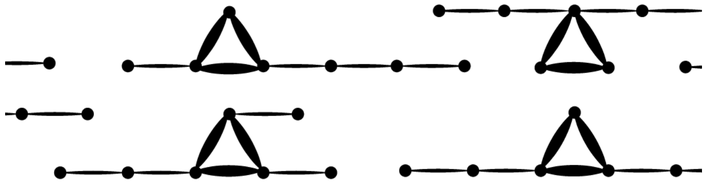
Research Directions

- Numerical analysis (Hudson et al, 2010)
- Probability (Kahle, 2011)
- Statistics (Chazal et al, 2018)
- Machine learning (Adams et al, 2017)
- and many others…









Ongoing / Future Works

- Speeding up Topological Computation (joint work with Yao)
- Stability of Homology Groups of Scale-Free Simplicial Complexes (joint work with Samorodnitsky and Yu)
- Random A_n-type Quiver Representation (?)
- Stability of Simplicial Set Representation of Sample Points Near a Manifold (joint work with Hoderlein)

Conclusion

- Topology is useful, when combined with data science, statistics, numerical analysis and computer science.
- Topological Data Analysis is happening right now, right here at Cornell.
- The potential is unlimited.

Chunyin Siu (Alex)
Center of Applied Mathematics, Cornell University
cs2323@cornell.edu

Thank you!

Chunyin Siu (Alex)
Center of Applied Mathematics, Cornell University
cs2323@cornell.edu

References

- Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017). Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.*, 18(1):218–252.
- Aragon-Calvo, M. A. and Szalay, A. S. (2012). The hierarchical structure and dynamics of voids. *Monthly Notices of the Royal Astronomical Society*, 428(4):3409–3424.
- Bell, G., Lawson, A., Martin, J., Rudzinski, J., and Smyth, C. (2019). Weighted persistent homology. *Involve*, 12(5):823–837.
- Bru¨el Gabrielsson, R. and Carlsson, G. (2019). Exposition and interpretation of the topology of neural networks. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1069–1076.
- Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308.
- Carlsson, G., Ishkhanov, T., de Silva, V., and Zomorodian, A. (2008). On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76:1–12.
- Chazal, F., Cohen-Steiner, D., and M´erigot, Q. (2011). Geometric inference for probability measures. *Found Comput Math*, 11:733–751.
- Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2018). Robust topological inference: Distance to a measure and kernel distance. *Journal of Machine Learning Research*, 18:1 – 40.
- Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

- HIFLD (2021). Cellular towers.
- Hudson, B., Miller, G. L., Oudot, S. Y., and Sheehy, D. R. (2010). Topological inference via meshing. In *Proceedings of the Twenty-Sixth Annual Symposium on Computational Geometry*, SoCG '10, pages 277–286, New York, NY, USA. Association for Computing Machinery.
- Kahle, M. (2011). Random geometric complexes. *Discrete & Computational Geometry*, 45(3):553–573.
- Li, M., Duncan, K., Topp, C. N., and Chitwood, D. H. (2017). Persistent homology and the branching topologies of plants. *American Journal of Botany*, 104(3):349–353.
- Perea, J. A. and Harer, J. (2015). Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838.
- Sizemore, A. E., Giusti, C., Kahn, A., Vettel, J. M., Betzel, R. F., and Bassett, D. S. (2018). Cliques and cavities in the human connectome. *Journal of Computational Neuroscience*, 44(1):115—145.

Diamonds in the slides

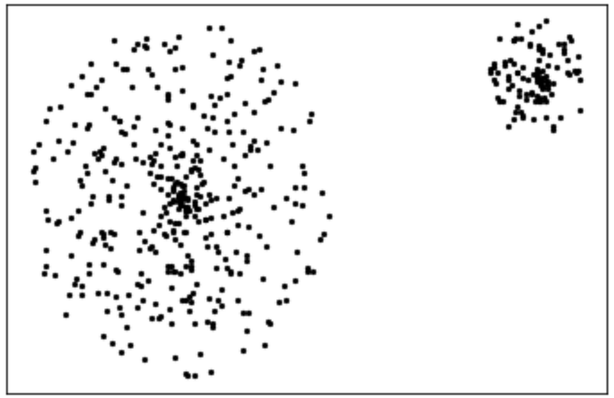
reserve slides in case I need them

TDA overview

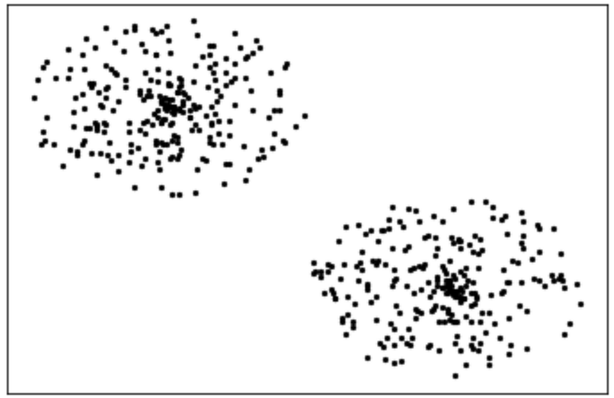
But what is TDA, really?

Odd One Out

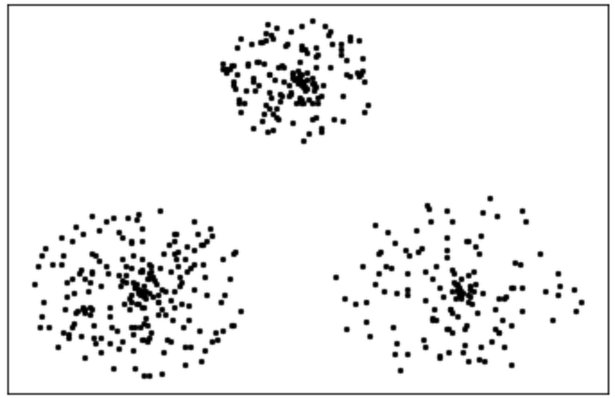
A



B

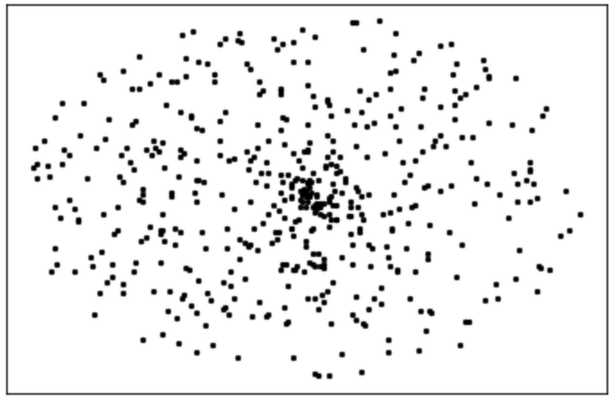


C

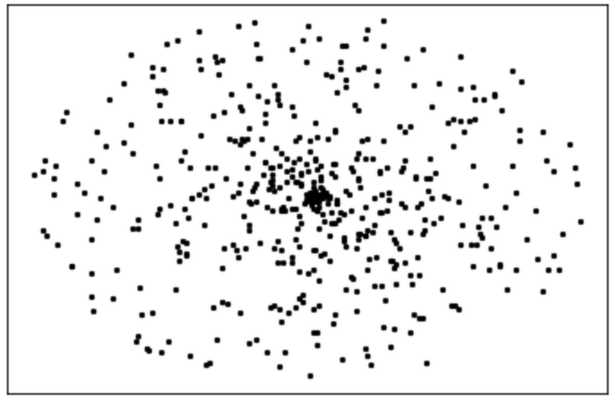


Odd One Out

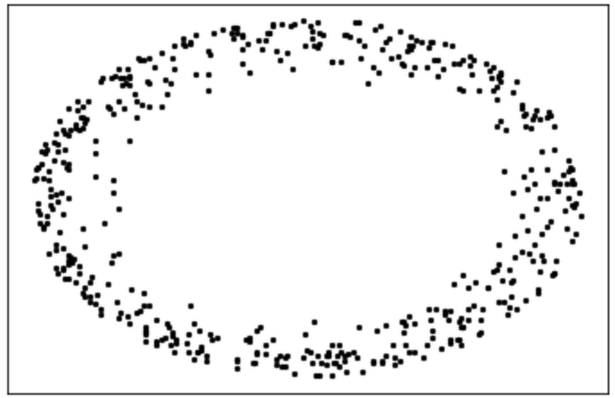
A



B



C



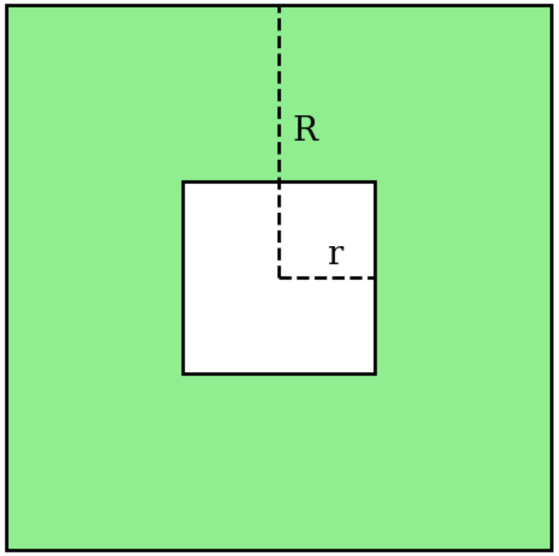
Seriously,
you’re doing this for a PhD?

0 training data
0 parameters
100% accuracy

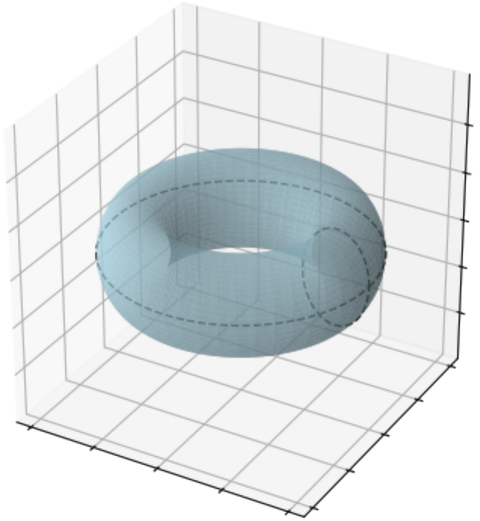
(for simple datasets)

Topological Features
of the Support of the Density

- i.e. components, loops, cavities and higher-dimensional holes



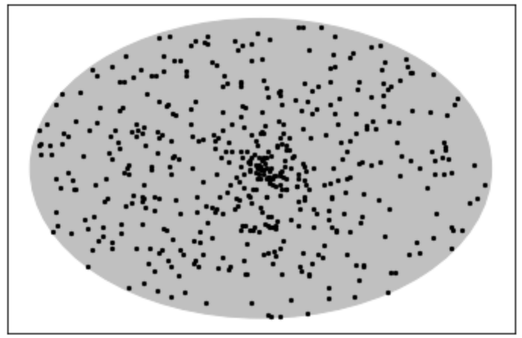
one component
one loop

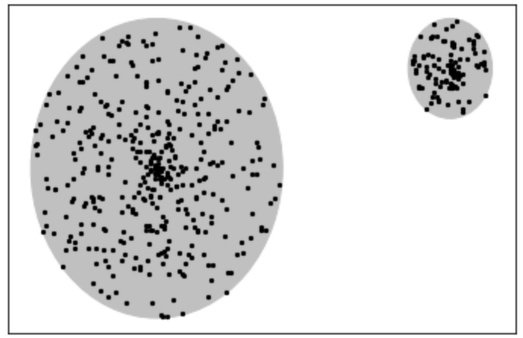


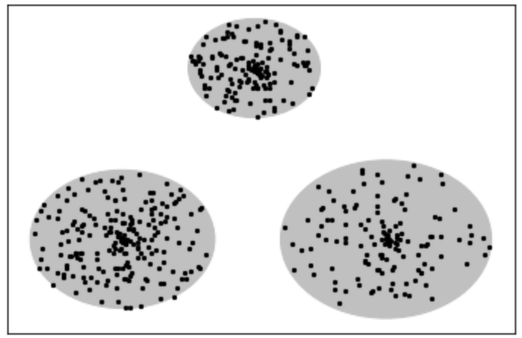
one component
two loops
one cavity

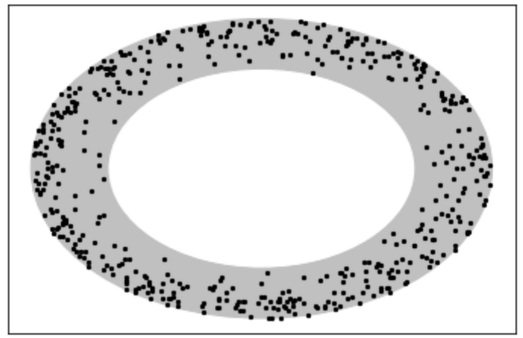
Topological Features

of the Support of the Density



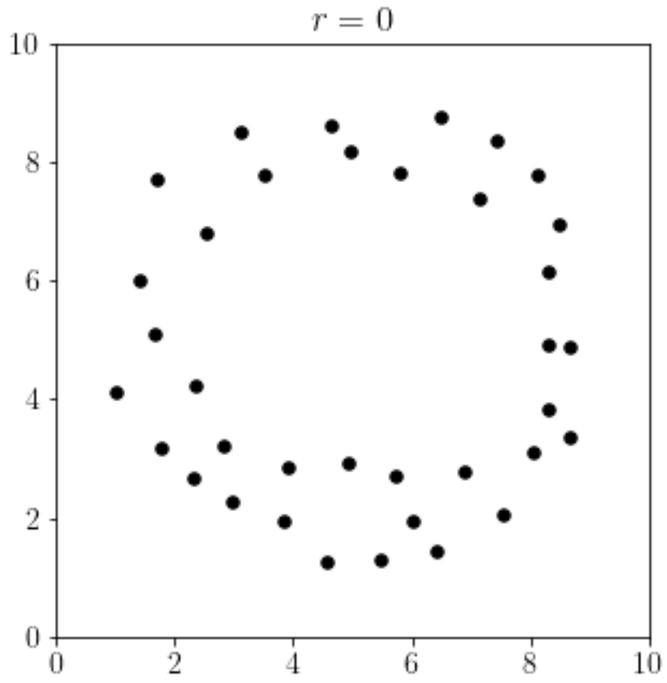


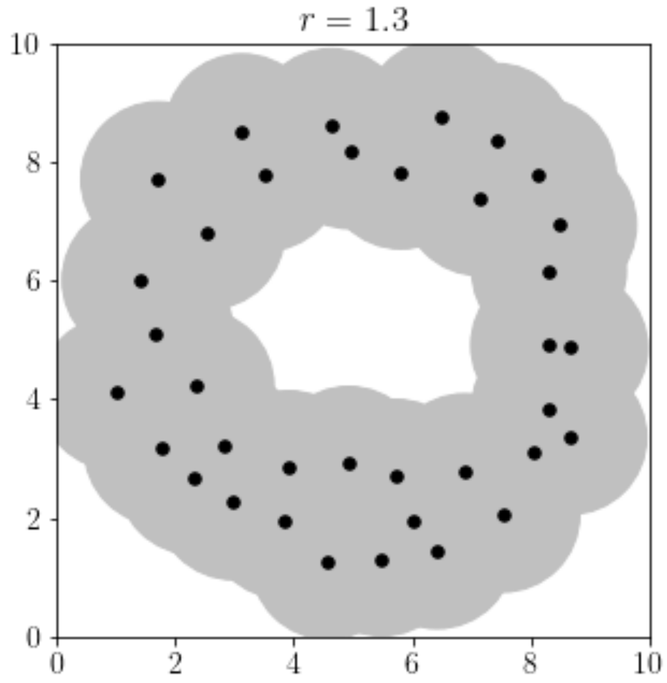


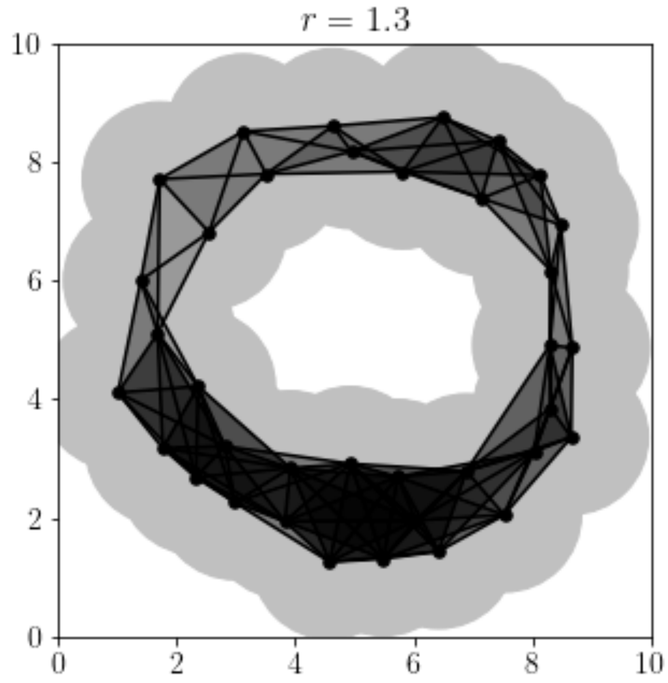


Estimator?
Mathematical Algorithm?

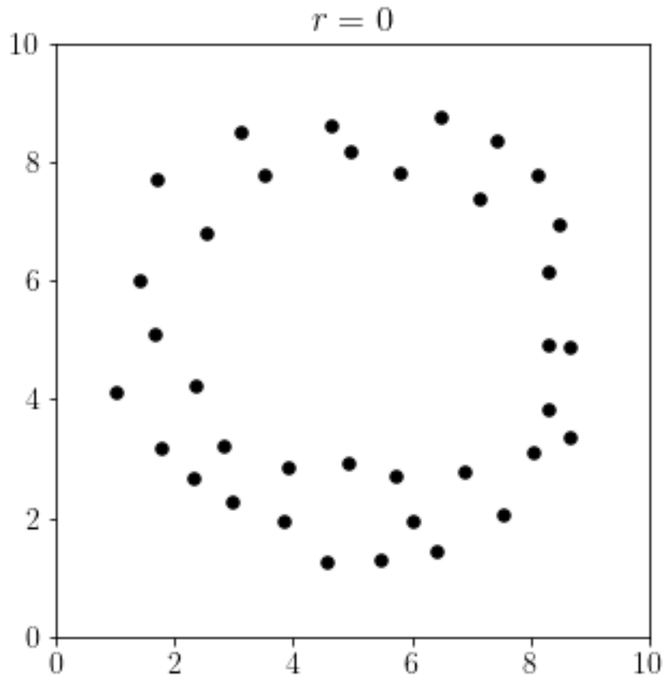
Yes!

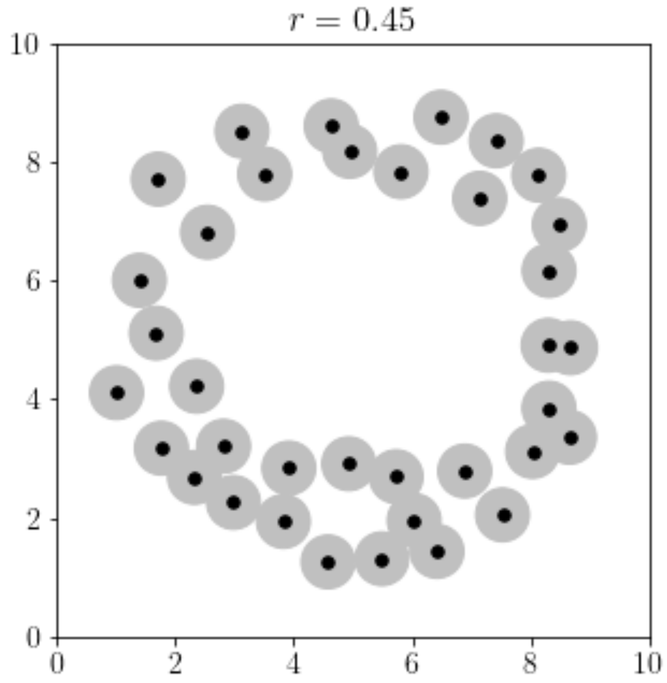


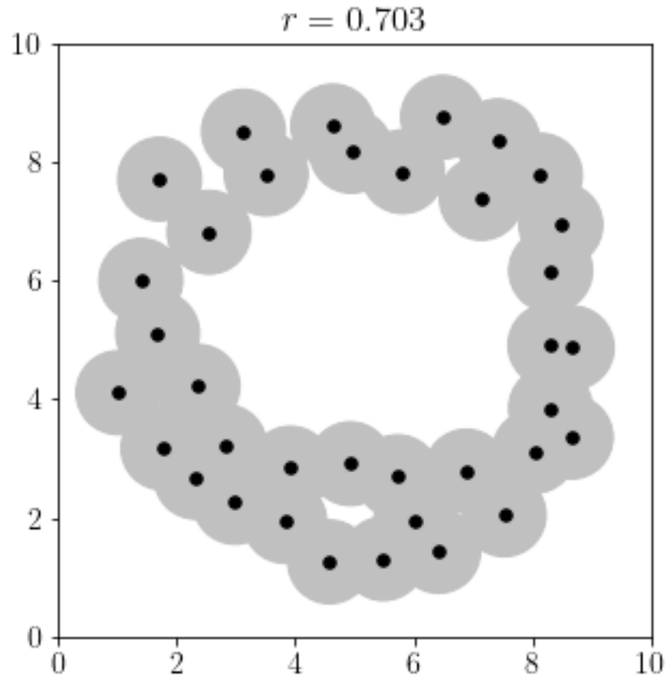


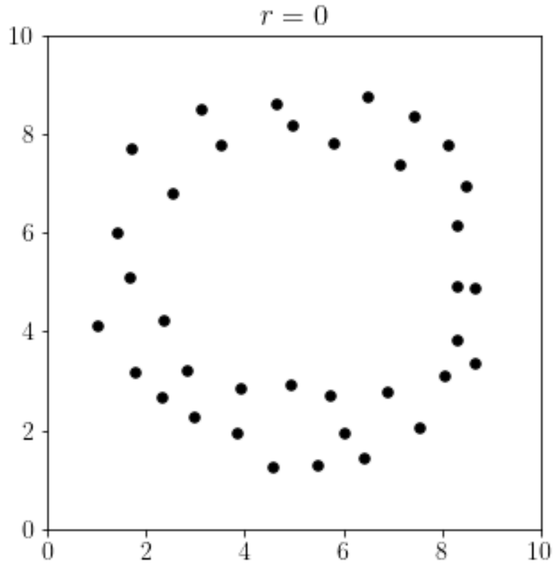


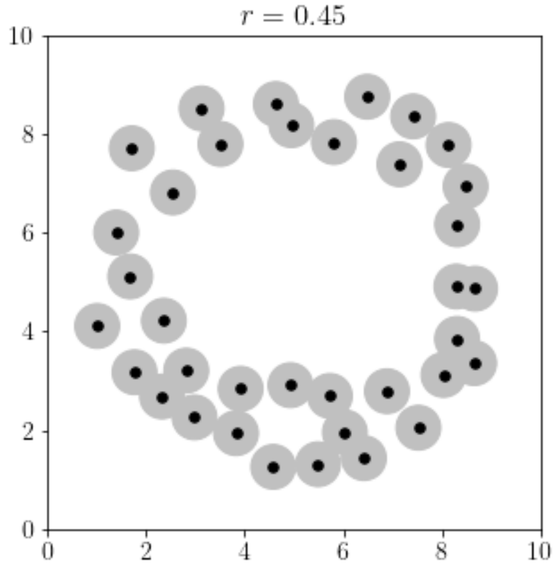
Pitfall

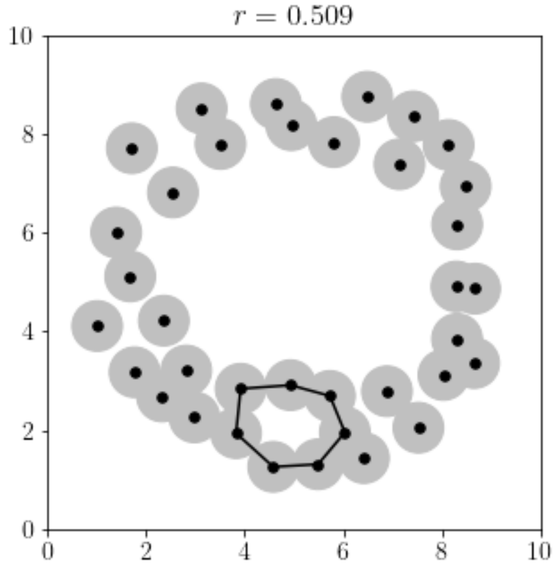


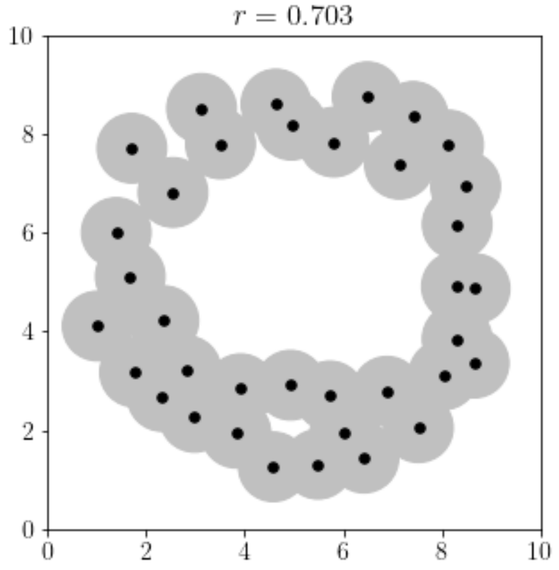


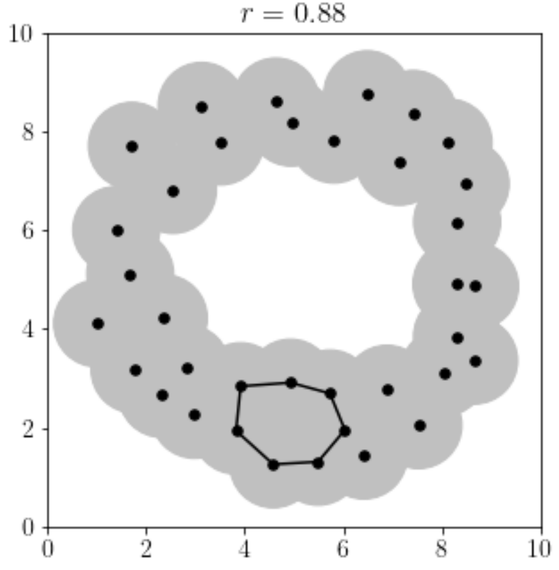


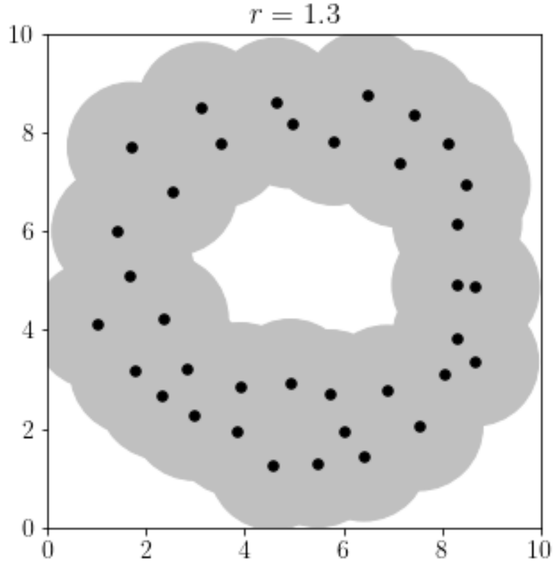

$r = 0$

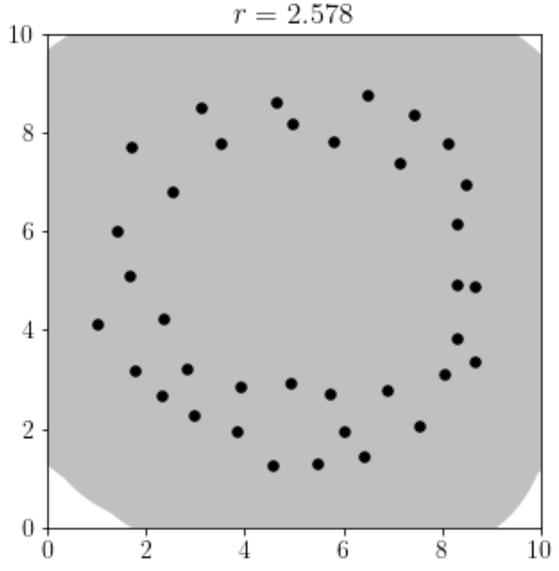

$r = 0.45$

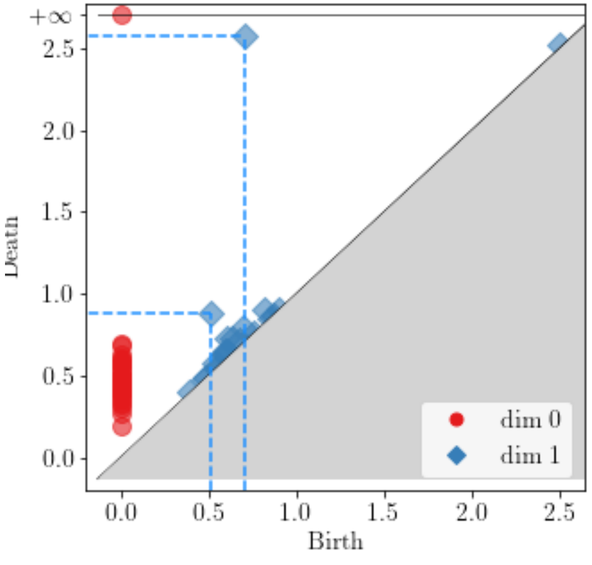

$r = 0.509$


$r = 0.703$


$r = 0.88$


$r = 1.3$

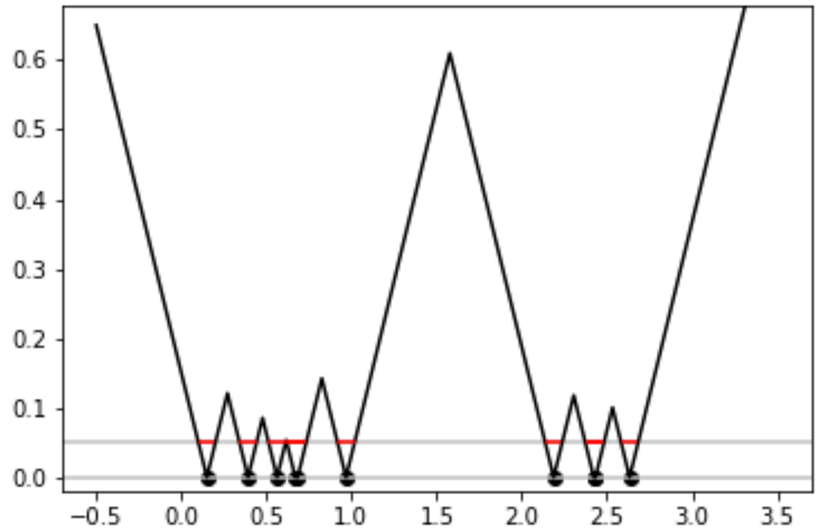

$r = 2.578$


Death
Birth
dim 0
dim 1

The Ground Truth Persistence Diagram?

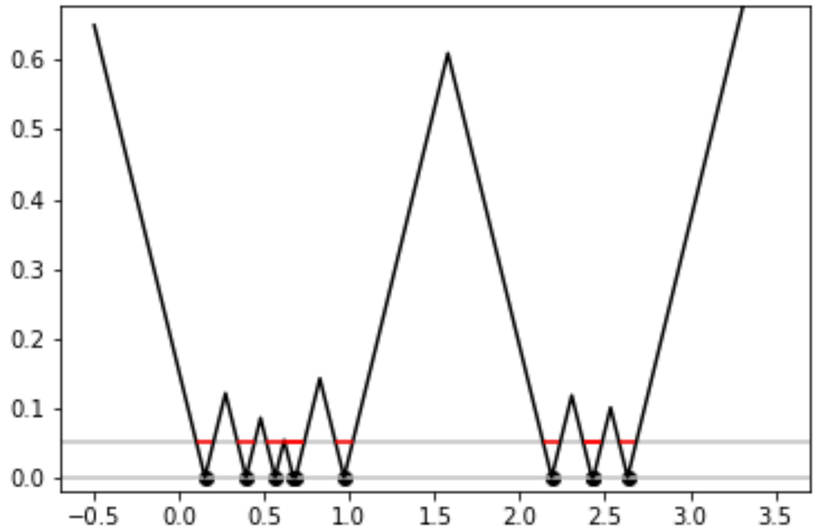
Higher-Dimensional Perspective

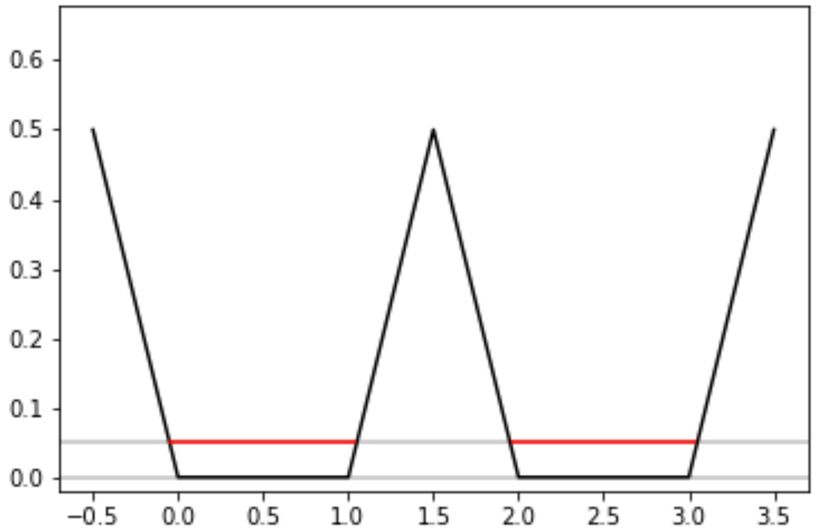
- Balls are lower-level sets of distance function
- $d_{emp}(x) = \min d(x, X_i)$



Higher-Dimensional Perspective

- Estimator of (lower-level sets of) the distance function of the support
- $d_{gt}(x) = \inf d(x, y)$; y ranges over the support of the density

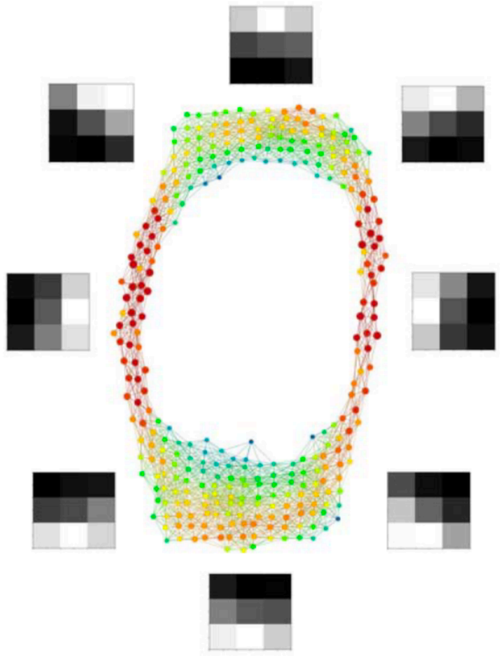


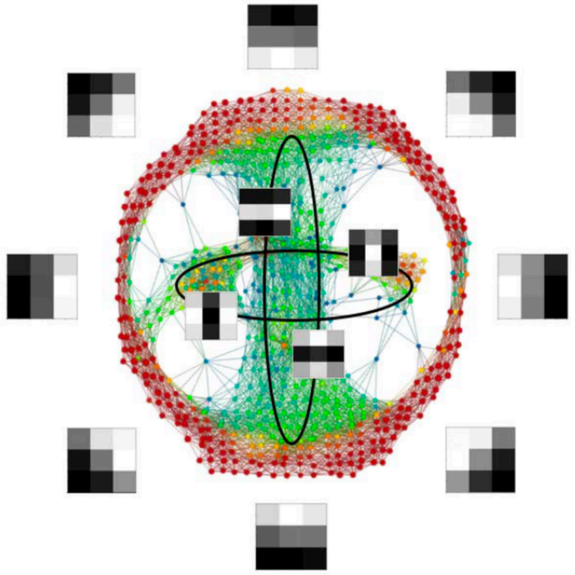


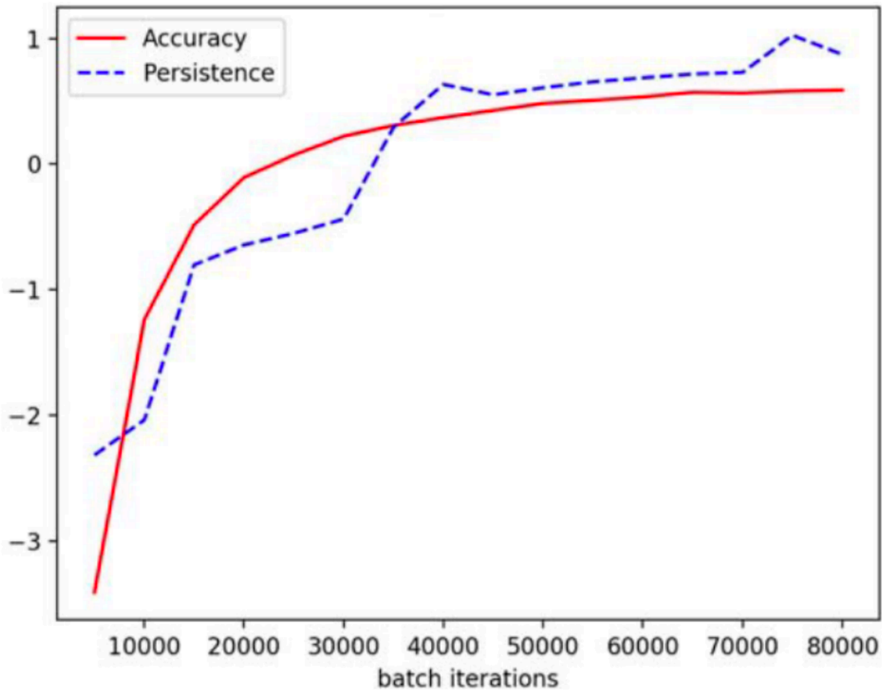
Take-Home Messages

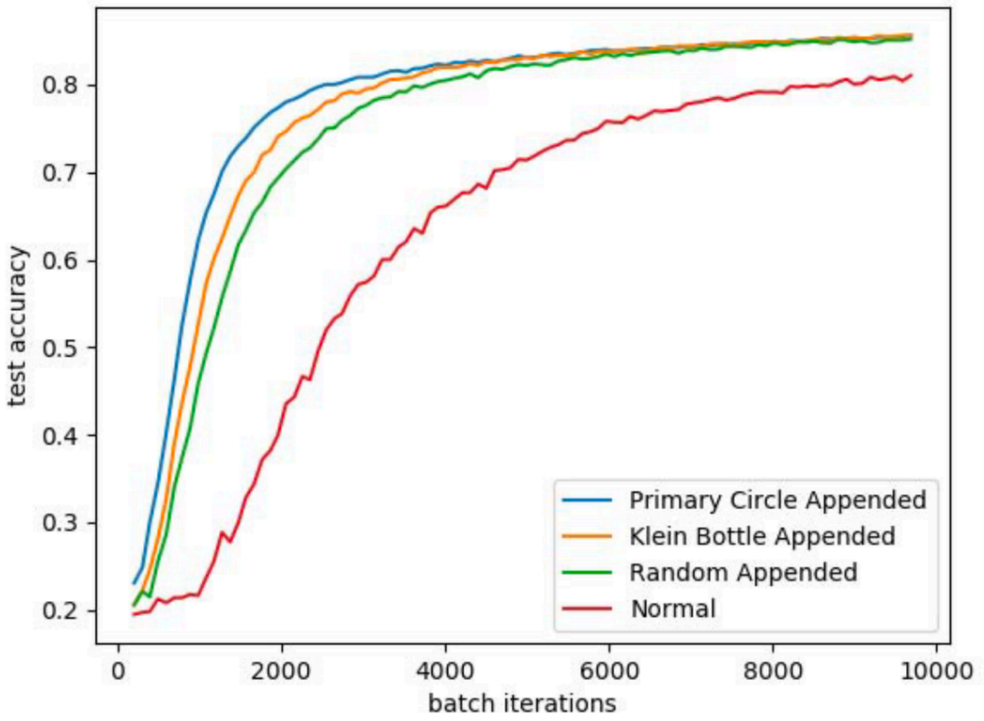
- useful when the dataset has global structures like loops and holes
- these structures can be estimated
- their information can be summarized in persistence diagrams

TDA and Machine Learning






Accuracy
Persistence
1
0
−1
−2
−3
10000
20000
30000
40000
50000
60000
70000
80000
batch iterations


test accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
2000
4000
6000
8000
10000
batch iterations
Primary Circle Appended
Klein Bottle Appended
Random Appended
Normal

What have we got so far?

- statistical model that highlights small features
- with a robust estimator
- and a bootstrapping method

What have we got so far?

- ✅ statistical model that highlights small features
- ✅ with a robust estimator
- and a bootstrapping method

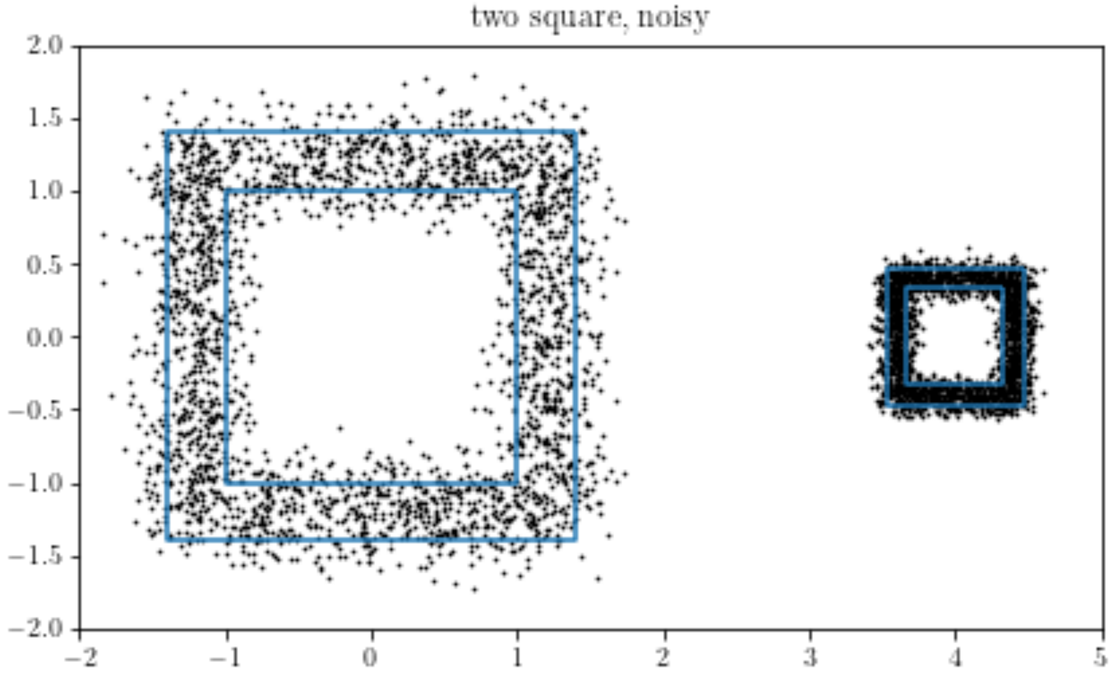
What have we got so far?

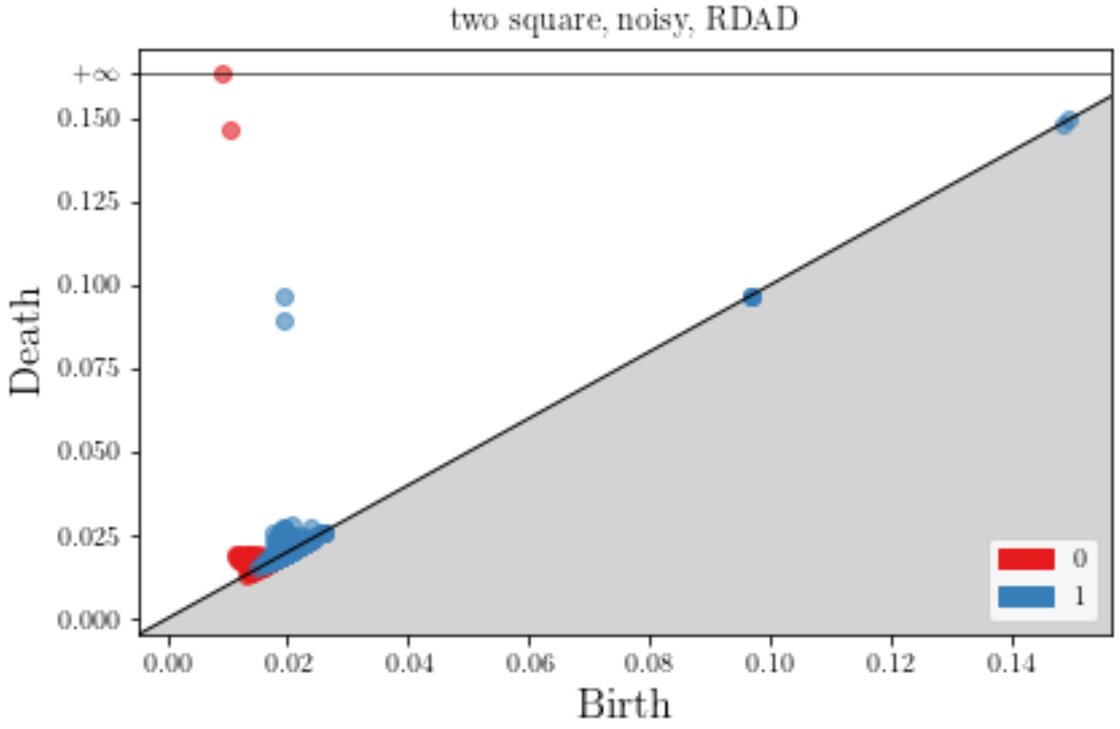
- ✅ statistical model that highlights small features— weighted distance and RDAD
- ✅ with a robust estimator— empirical version of RDAD
- and a bootstrapping method

Bootstrapping

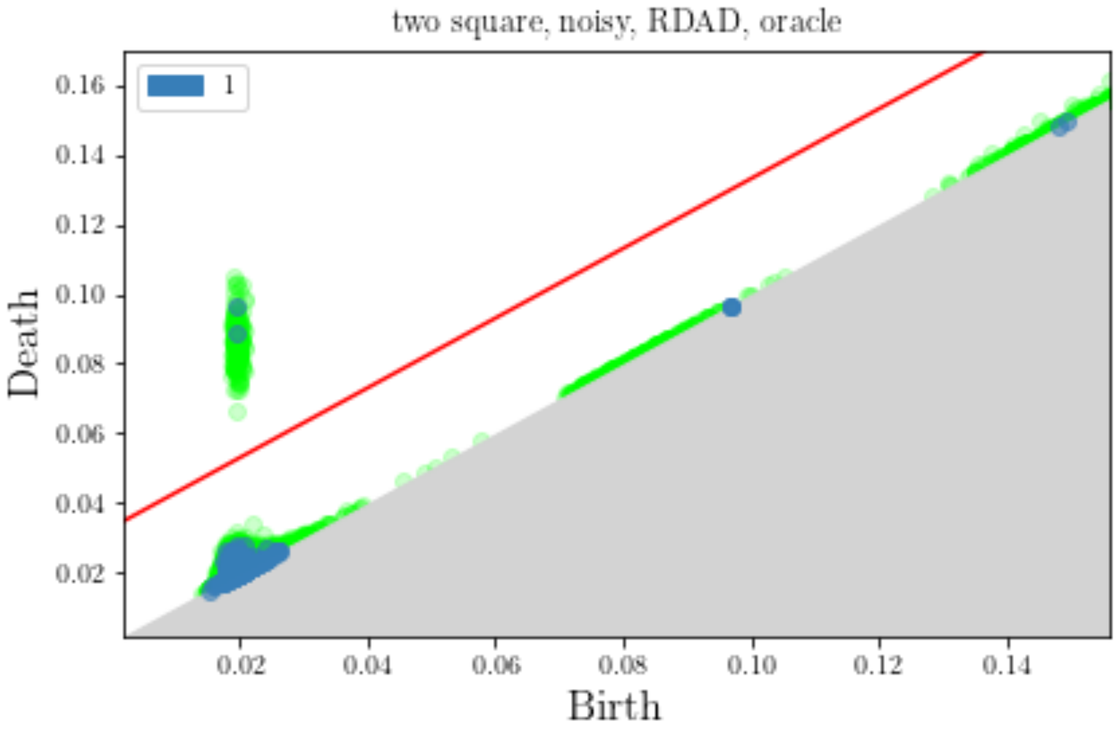
the Randomness

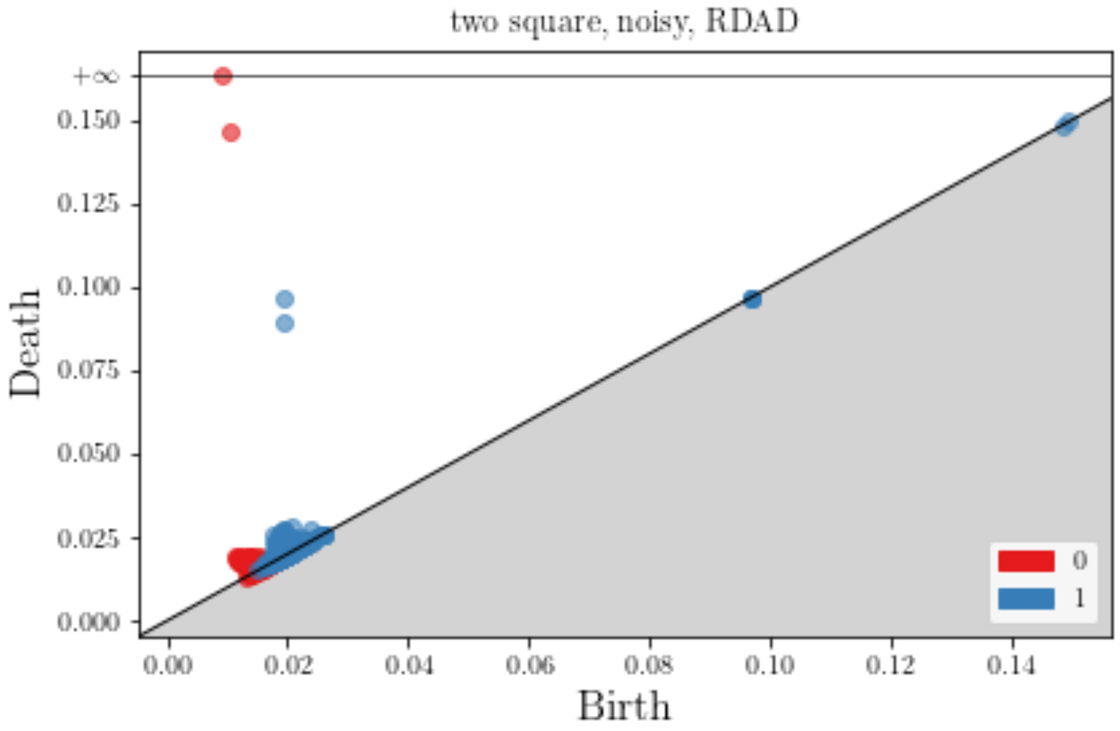
Persistence diagrams in parallel universes?





Persistence diagrams in parallel universes?



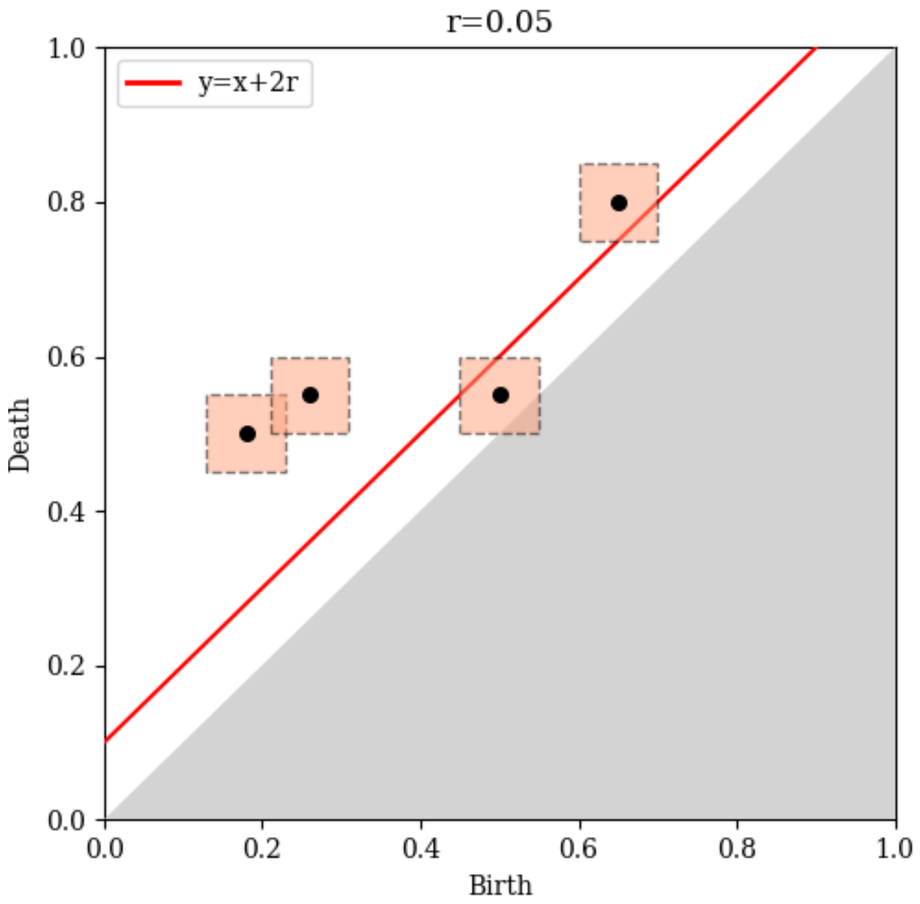


Known problem

- Fasy et al (2014) — distance function
- Chazal et al (2018) — distance-to-measure
  - proven conservative method
  - unproven tighter method

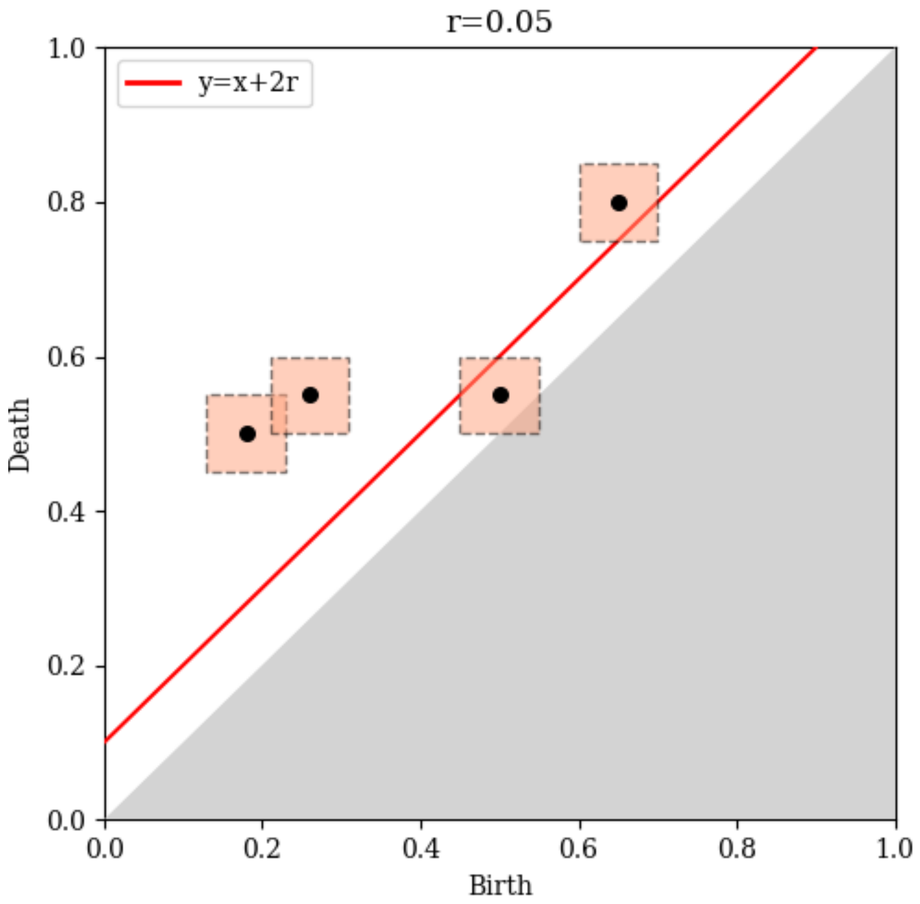
We need a notion of dissimilarity…

- bottleneck distance:
  metric between two persistence diagrams
- bottleneck distance metric ball
  - 1 point from each square
  - any number of points beneath the red line



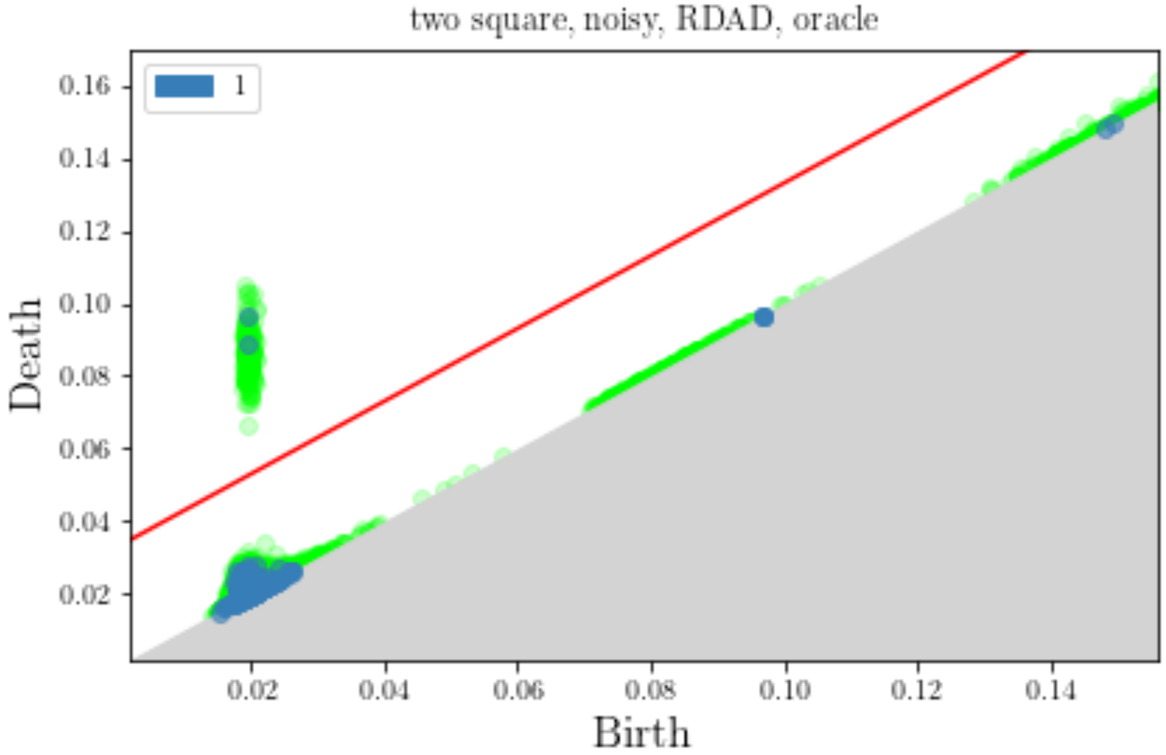
Meta-Bootstrapping

- Find a suitable radius for the bottleneck distance metric ball
- Draw the red line
- Declare points above red line significant



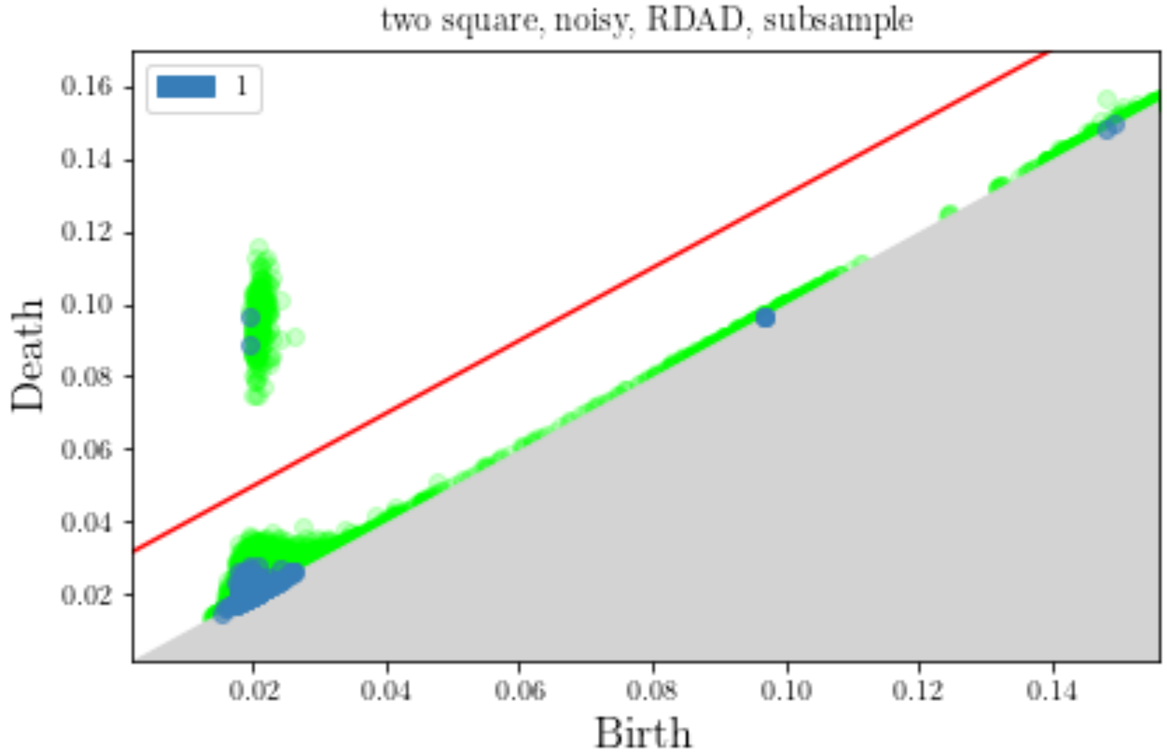
What is a suitable radius?

- if we have the oracle:
  - 95-percentile of bottleneck distances from the persistence diagrams of other samples with the same number of points

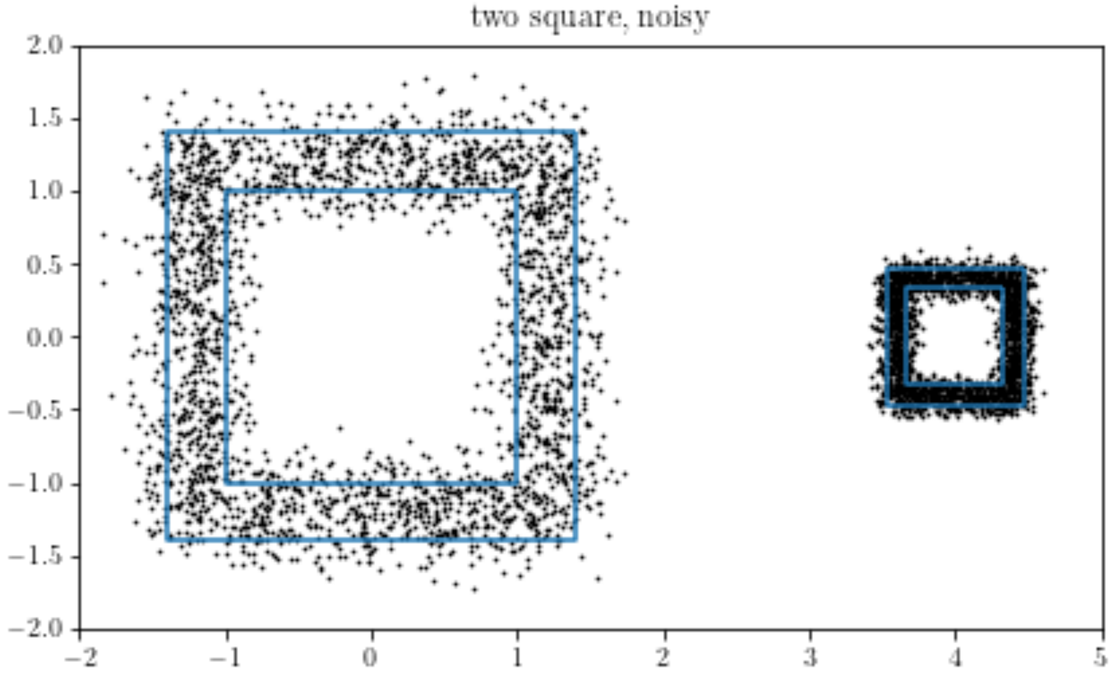


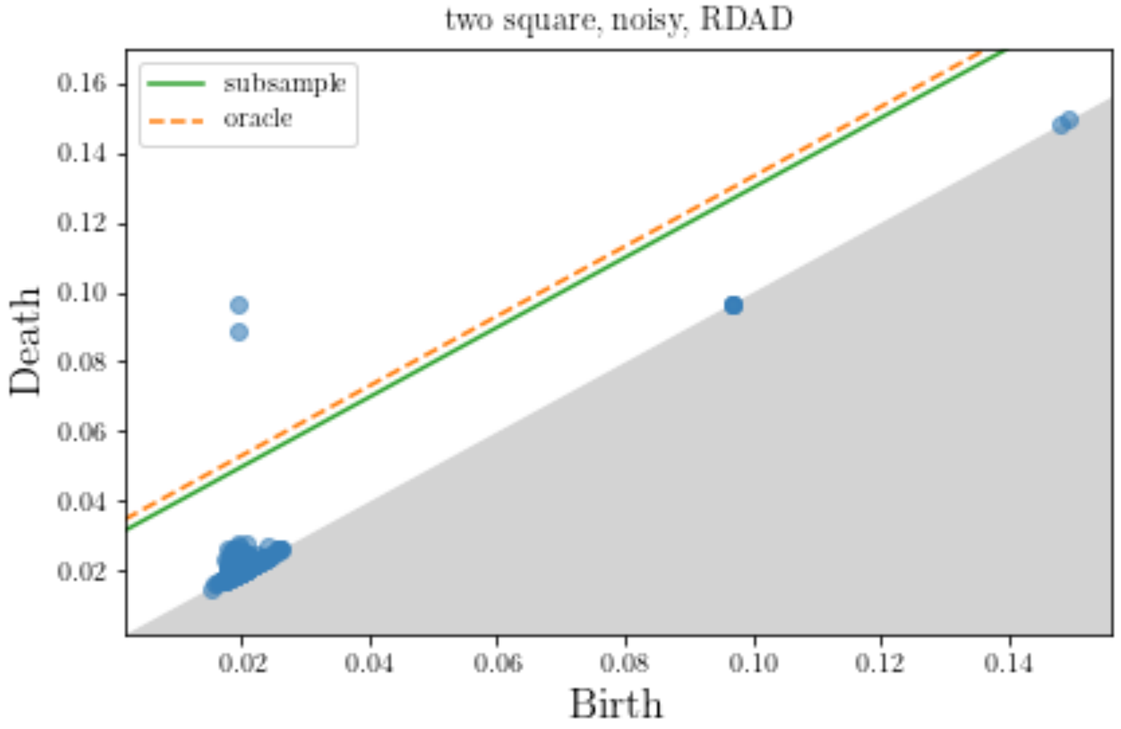
What if we don’t have the oracle?

- Generate proxy same-size samples by drawing points from the empirical sample with replacement



How do the two compare?





Just a snap to go…

- statistical model that highlights small features— weighted distance and RDAD
- with a robust estimator— empirical version of RDAD
- and a bootstrapping method— subsample bootstrapping