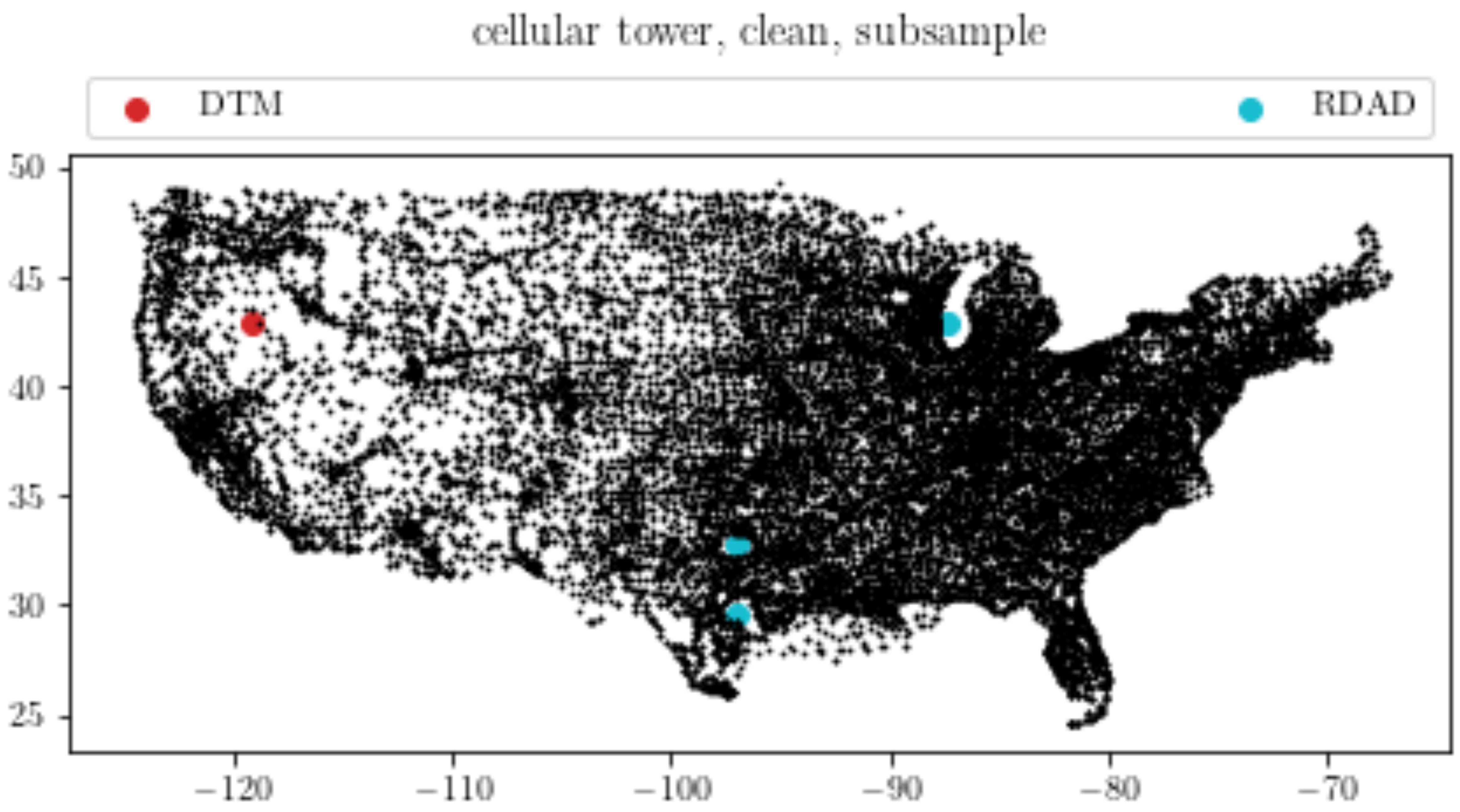
Topological Data Analysis

Small Density Vacuum and How to Find Them Robustly



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

In the beginning…

there was the data

Credit: NASA/NCSA, University of Illinois
Visualization by Frank Summers, Space Telescope Science Institute
Simulation by Martin White and Lars Hernquist, Harvard University
https://universe.nasa.gov/resources/89/cosmic-web/

In the beginning…

there was the data

and the data was non-parametric,

Credit: NASA/NCSA, University of Illinois
Visualization by Frank Summers, Space Telescope Science Institute
Simulation by Martin White and Lars Hernquist, Harvard University
https://universe.nasa.gov/resources/89/cosmic-web/

In the beginning…

there was the data

and the data was non-parametric,
and has voids,

Credit: NASA/NCSA, University of Illinois
Visualization by Frank Summers, Space Telescope Science Institute
Simulation by Martin White and Lars Hernquist, Harvard University
https://universe.nasa.gov/resources/89/cosmic-web/

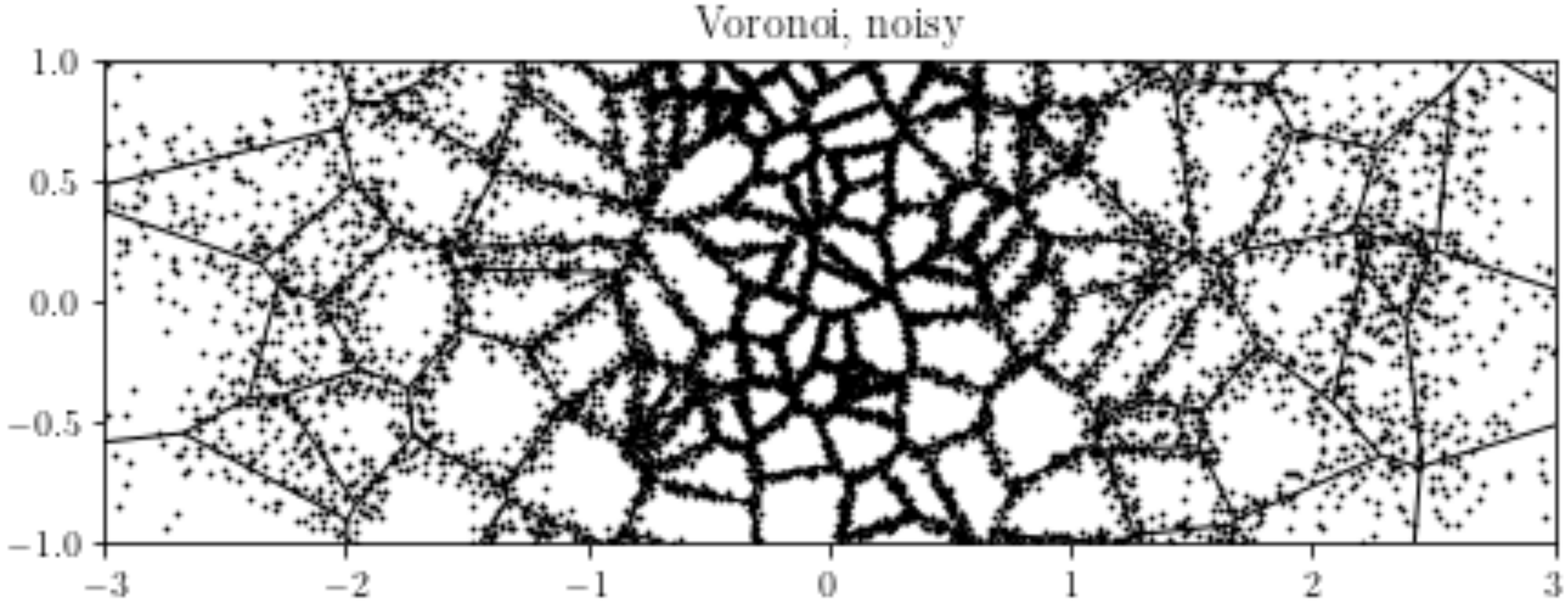
In the beginning…

there was the data

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

Credit: NASA/NCSA, University of Illinois
Visualization by Frank Summers, Space Telescope Science Institute
Simulation by Martin White and Lars Hernquist, Harvard University
https://universe.nasa.gov/resources/89/cosmic-web/

Let there be ground truth



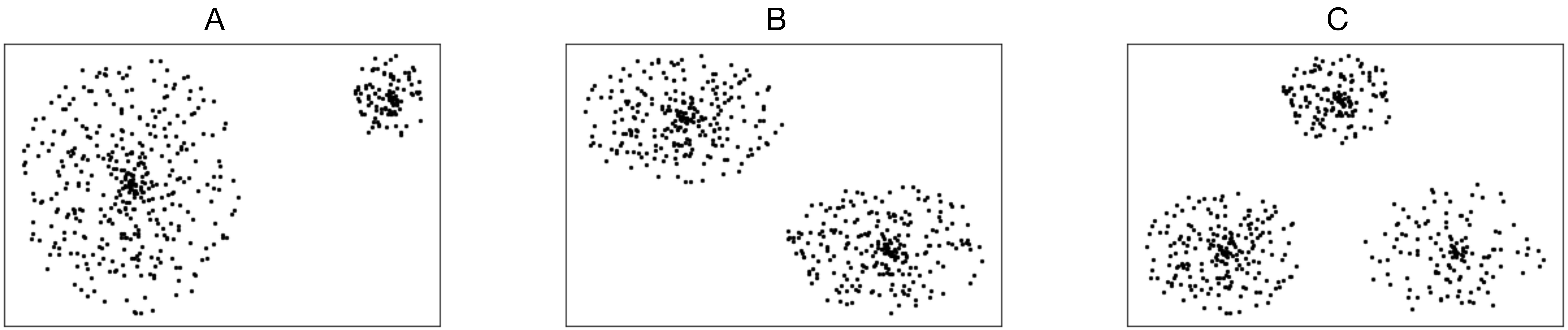
Agenda

- Topological Data Analysis: What and Why
- My Work: the Size, the Noise and the Randomness
- Numerical Simulations

Act I

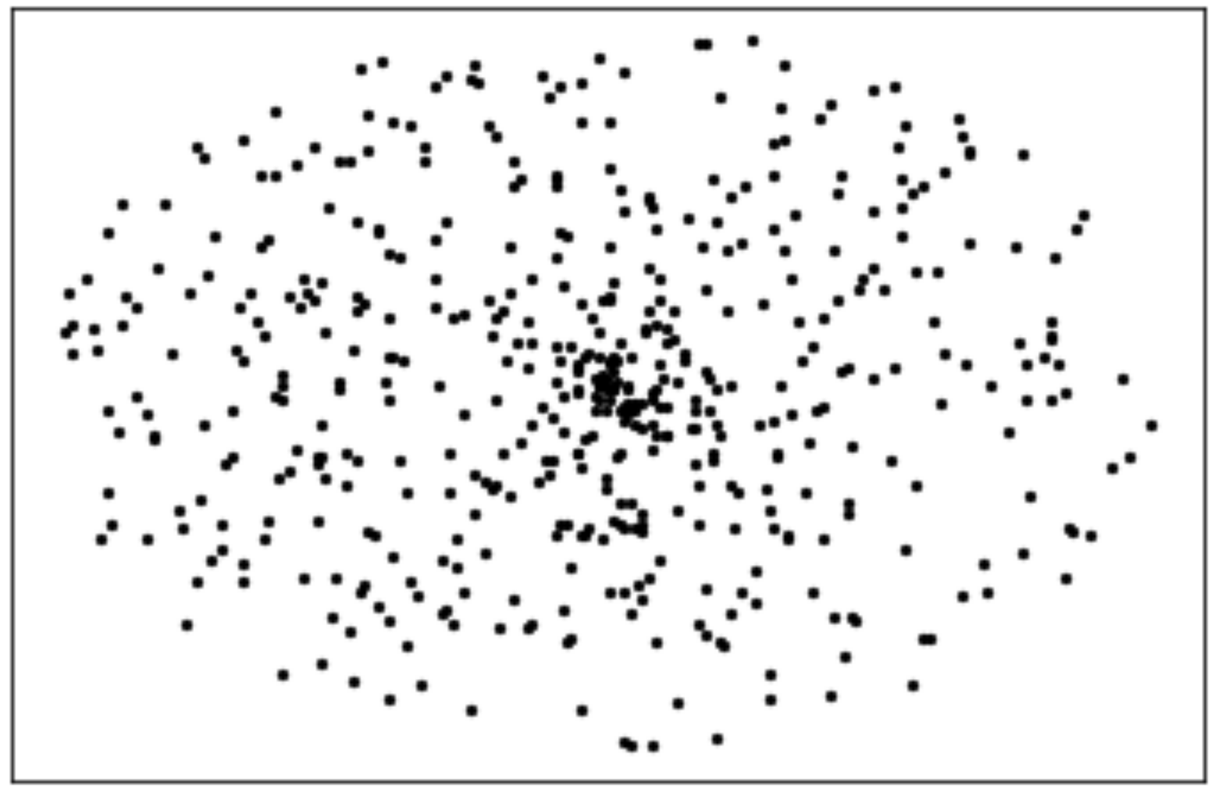
What the Fisher is Topological Data Analysis

Odd One Out

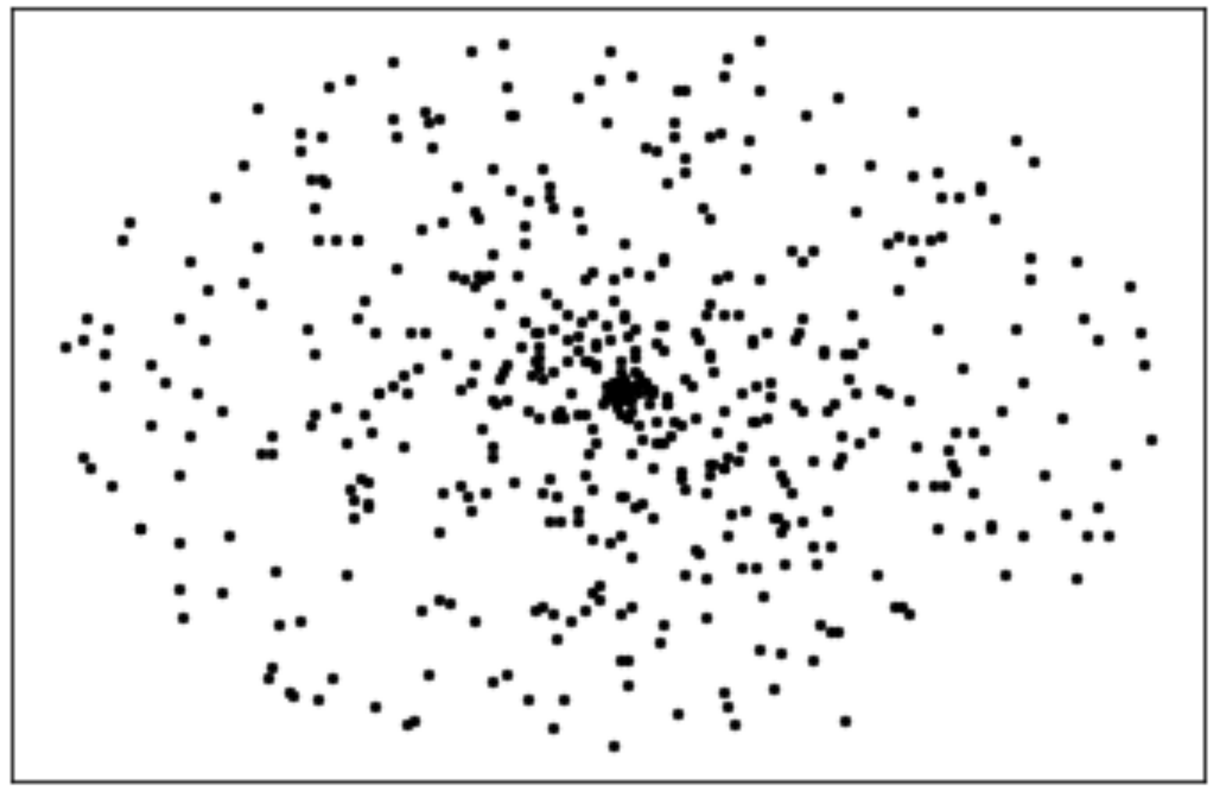


Odd One Out

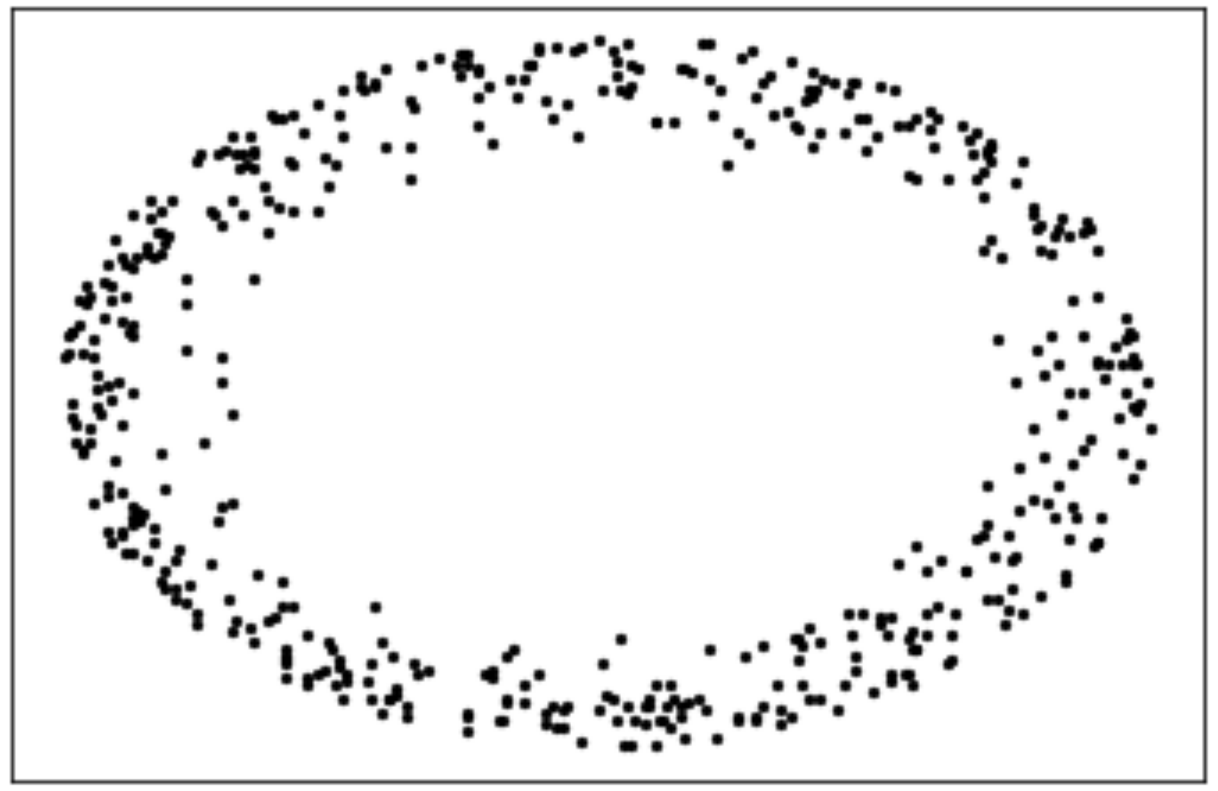
A



B



C



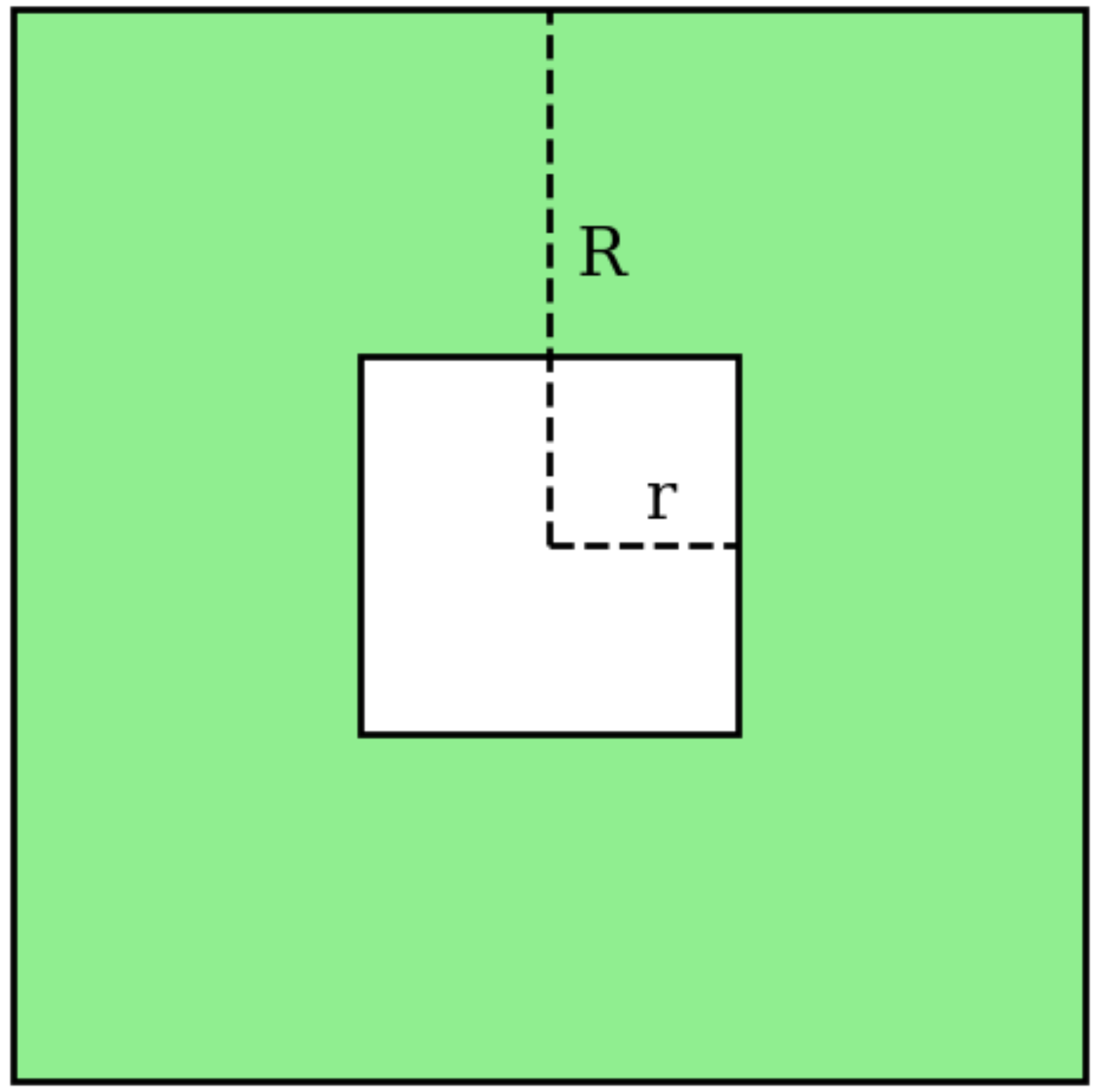
Seriously,
you’re doing this for a PhD?

0 training data
0 parameters
100% accuracy

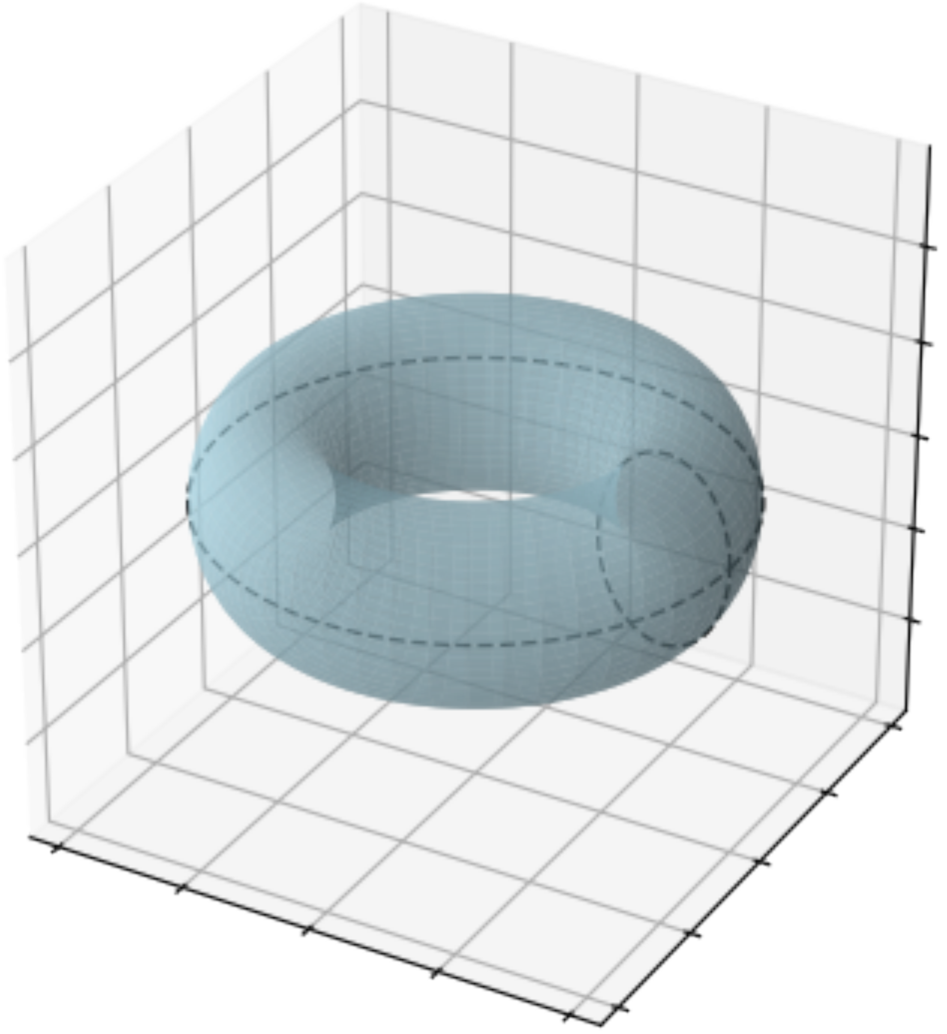
(for simple datasets)

Topological Features
of the Support of the Density

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



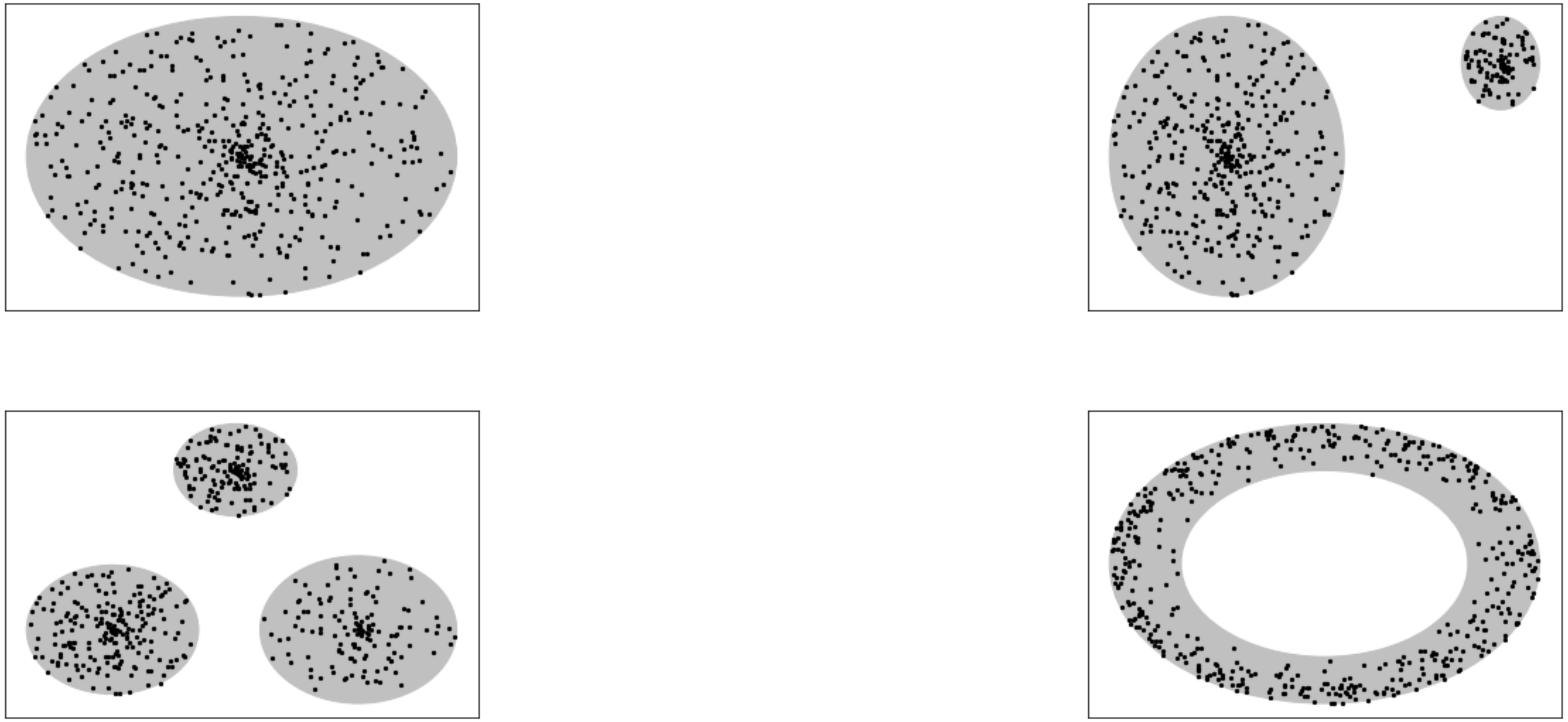
one component
one loop



one component
two loops
one cavity

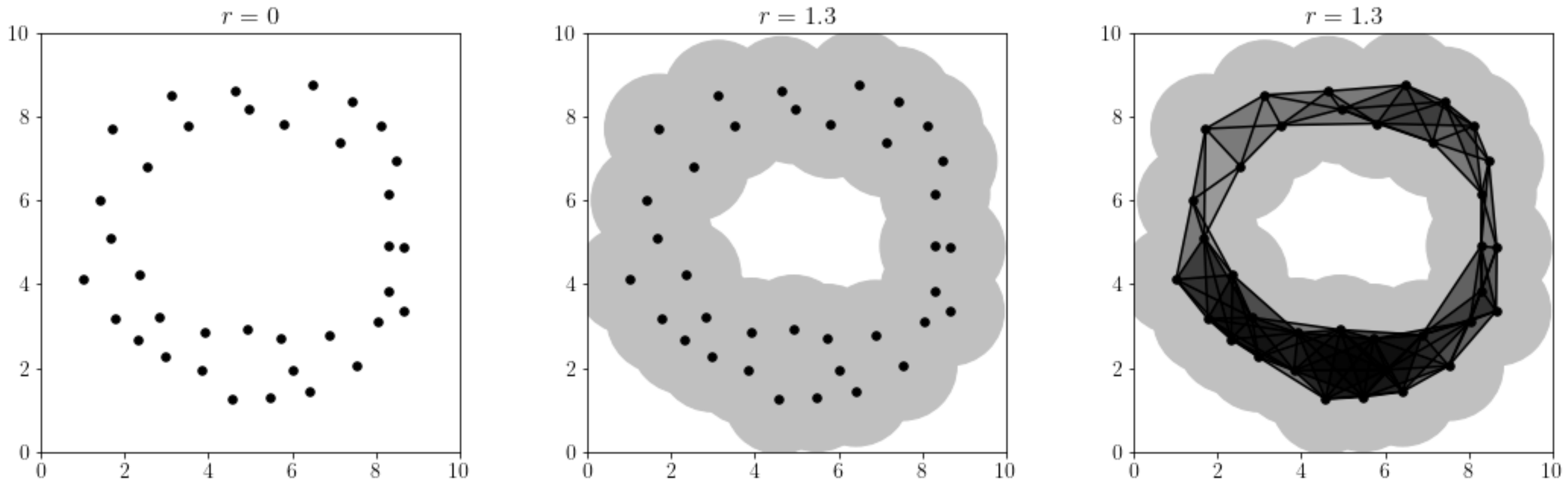
Topological Features

of the Support of the Density

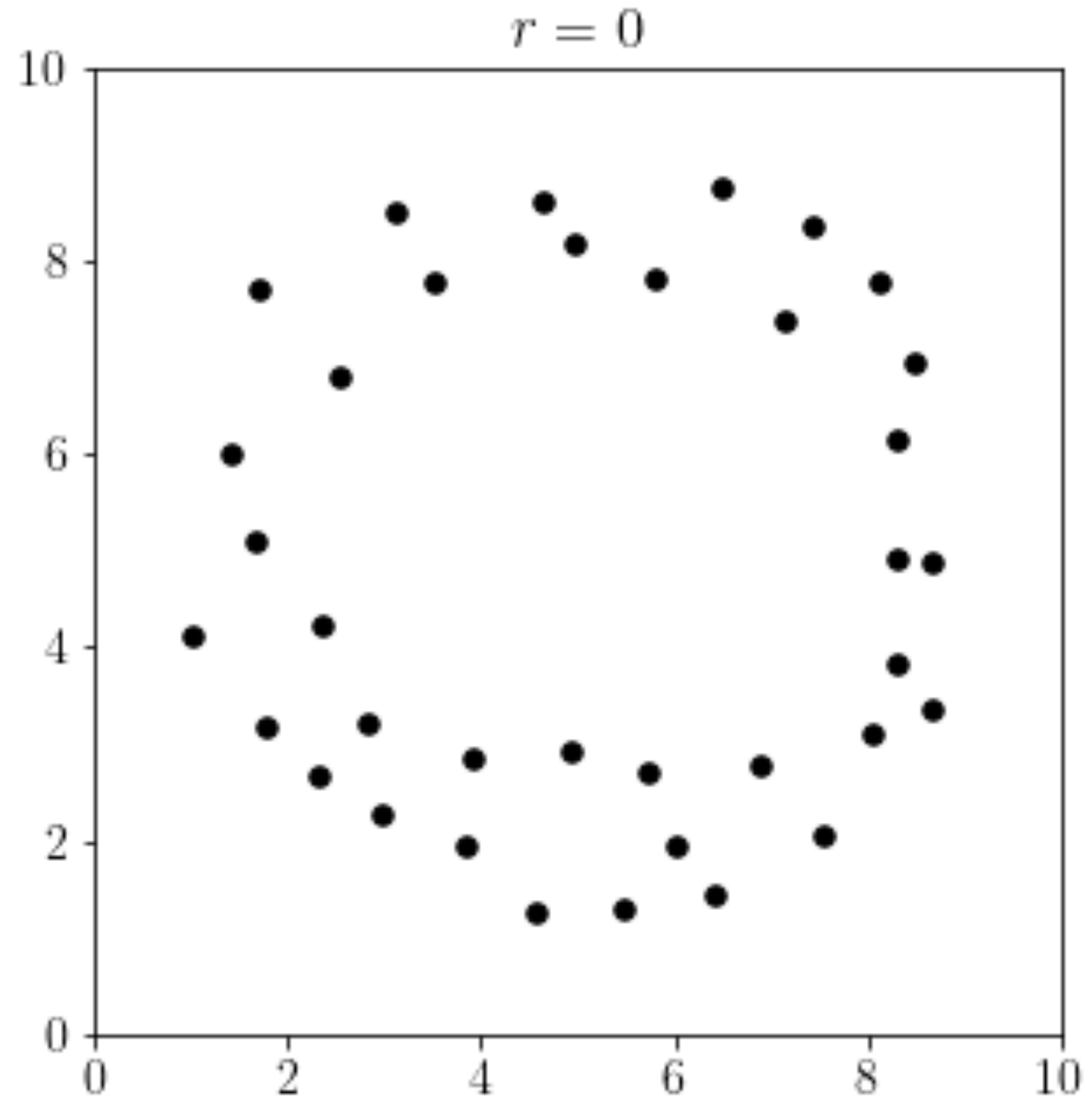


Estimator?
Mathematical Algorithm?

Yes!



Pitfall



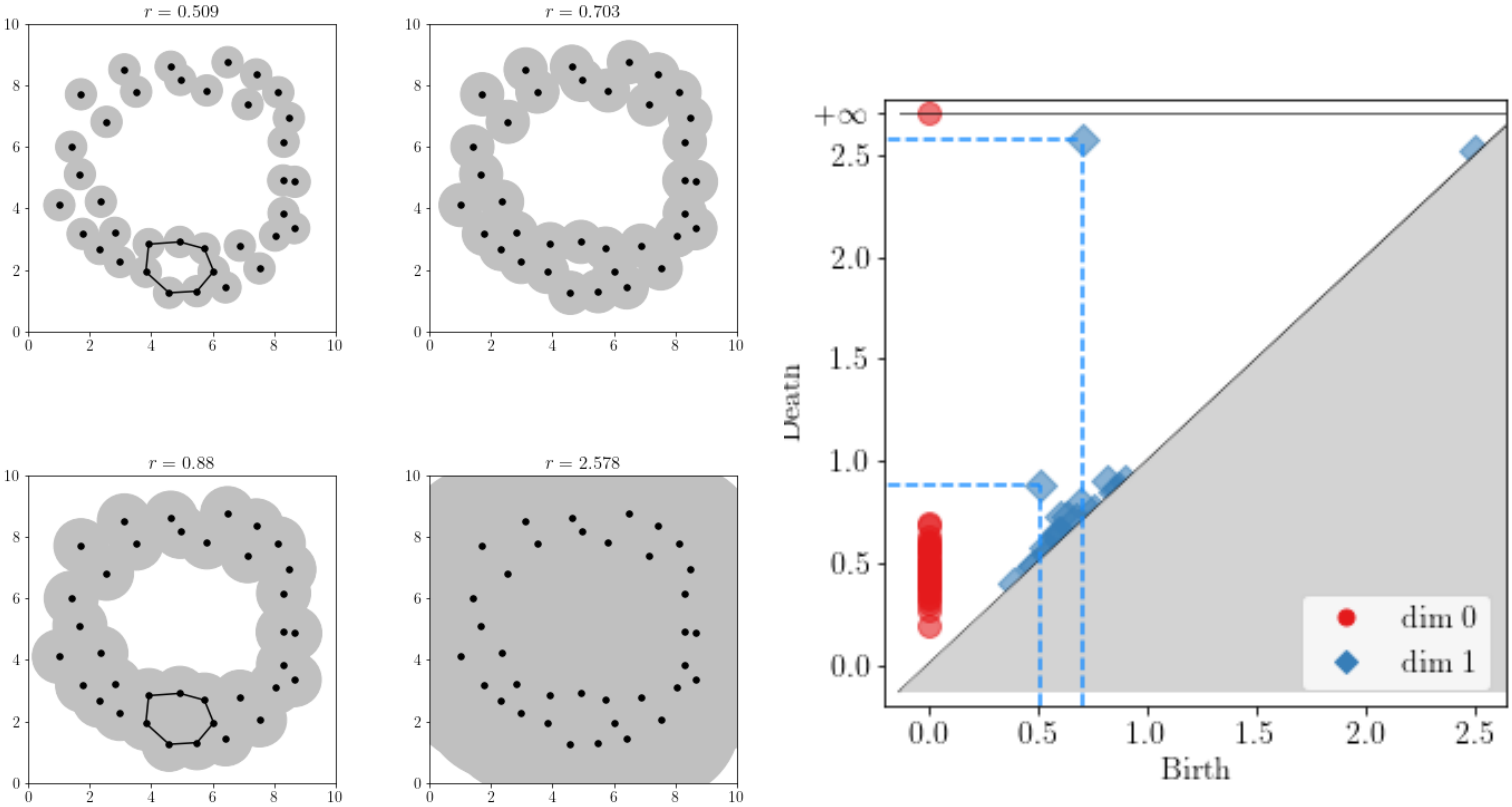


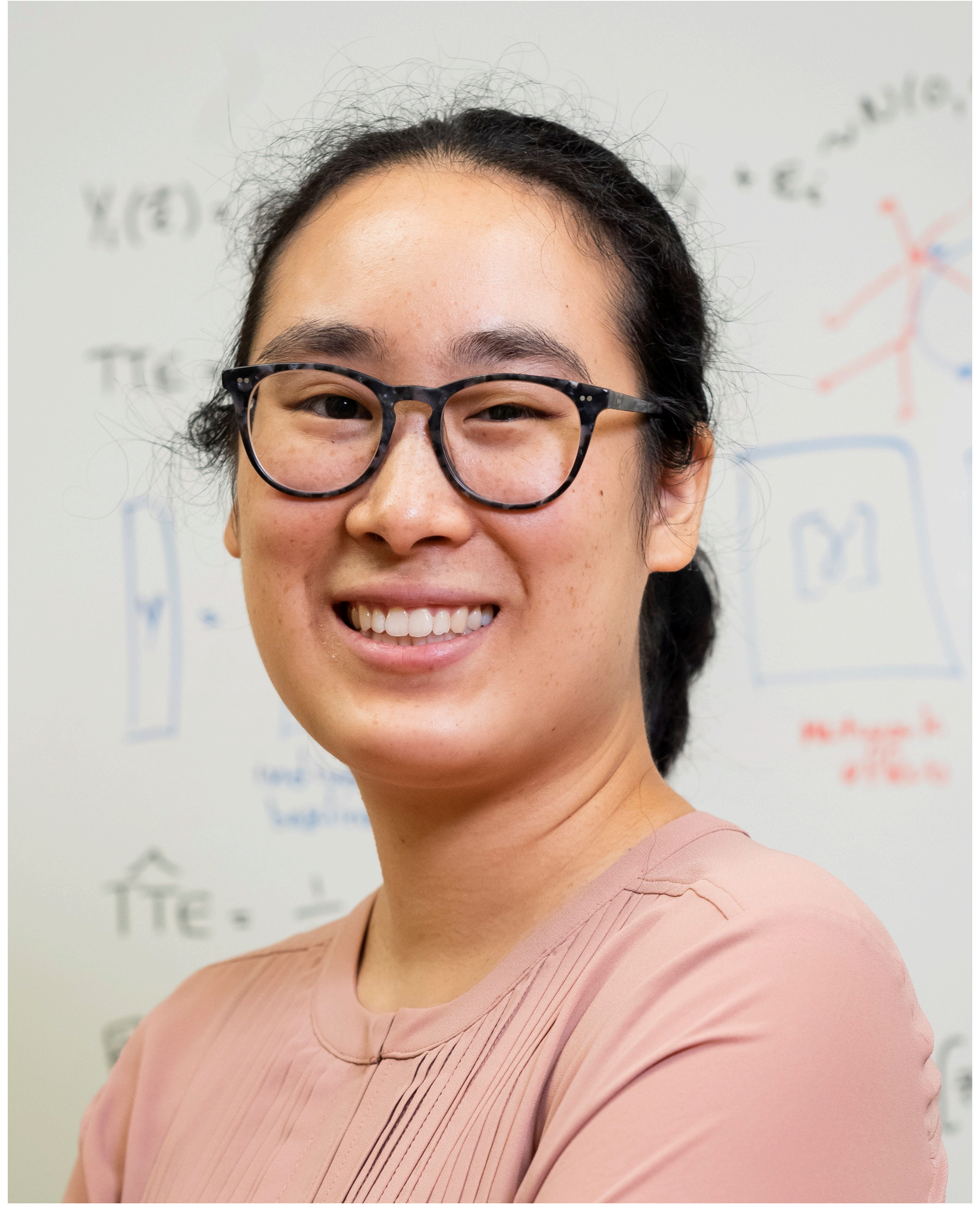
diagram credit: Andrey Yao

Act II

Small Density Vacuum and How to Find Them Robustly

My Lovely Collaborators

Gennady Samorodnitsky

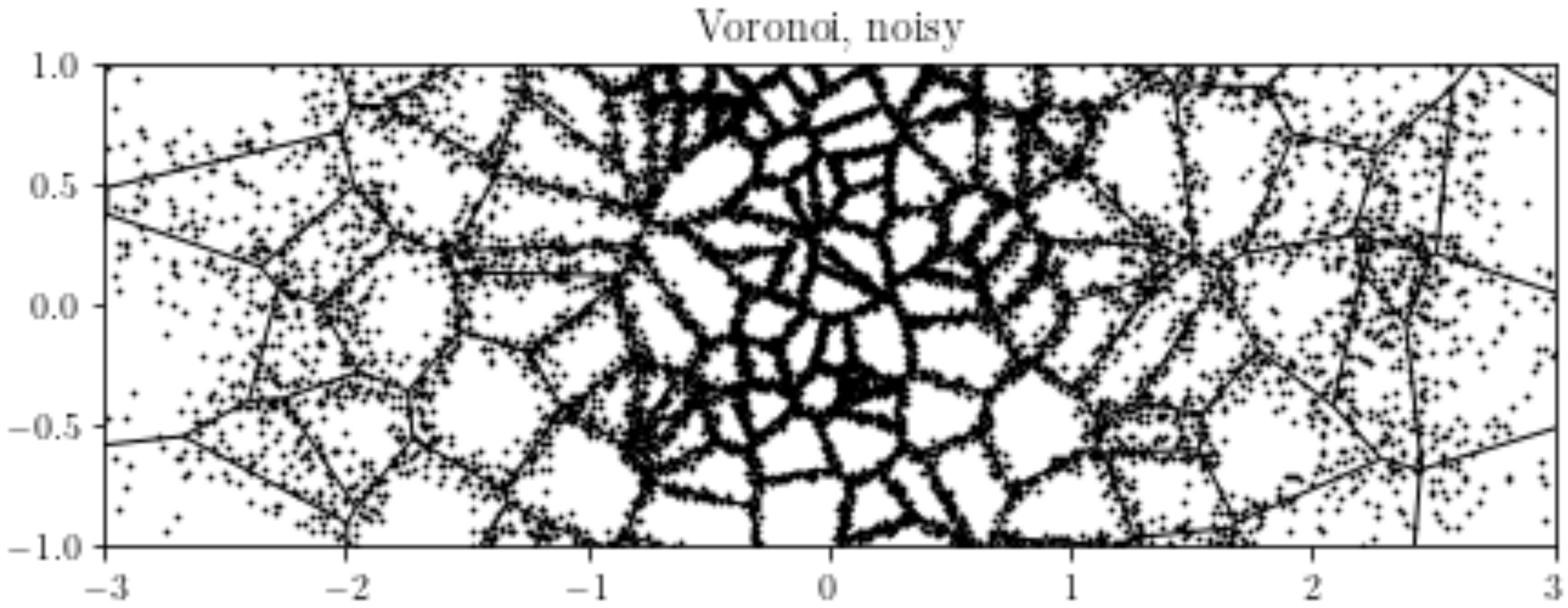


Christina Lee Yu

Andrey Yao

Two problems

- Size
- Noise



Two Problems

- Size
- Noise

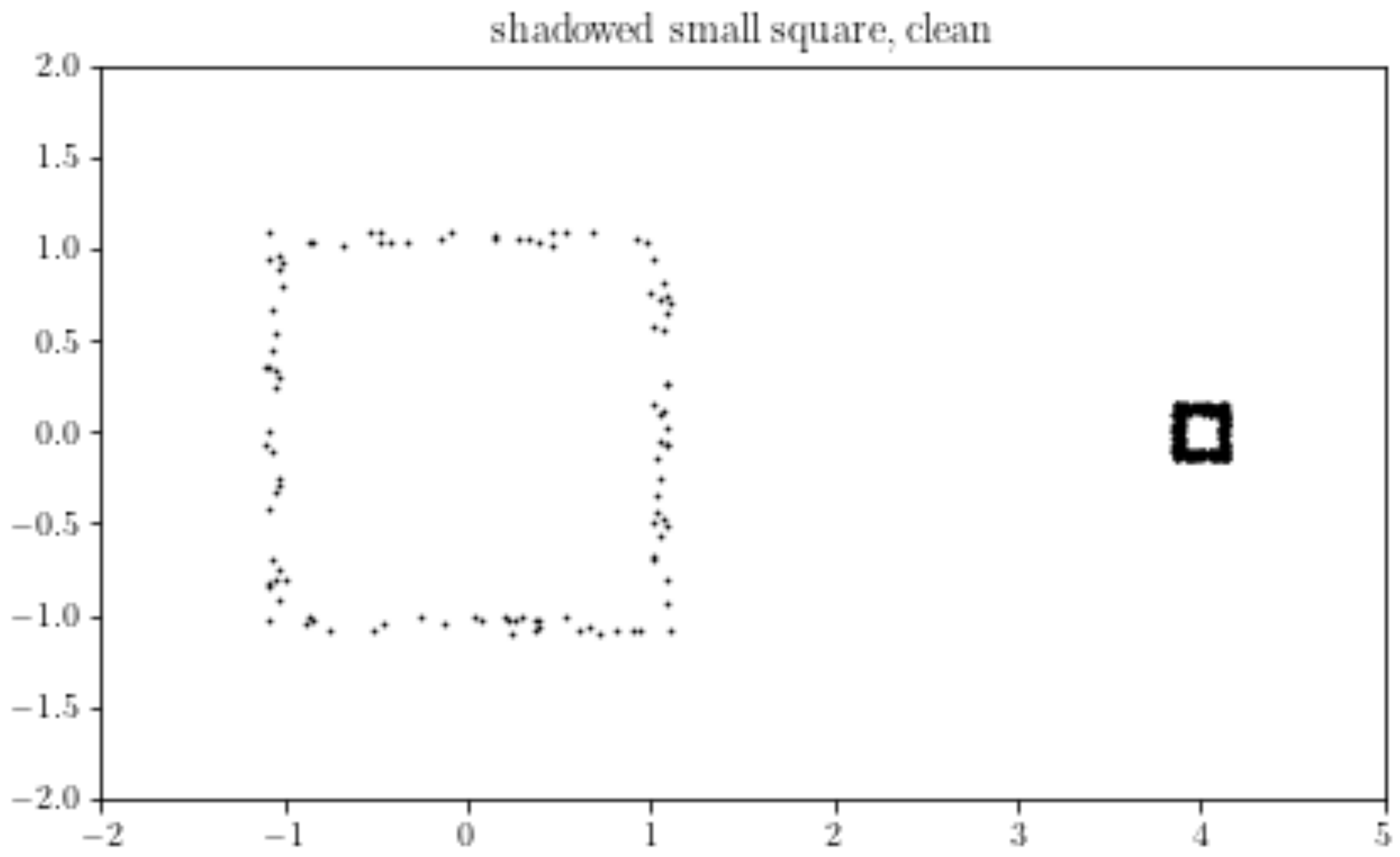
- Related works
    - Hickok (2022)
    - Berry and Sauer (2019)
    - Moon et al (2018)
    - Carlsson and Zomorodian (2009)
    - etc…

One solution

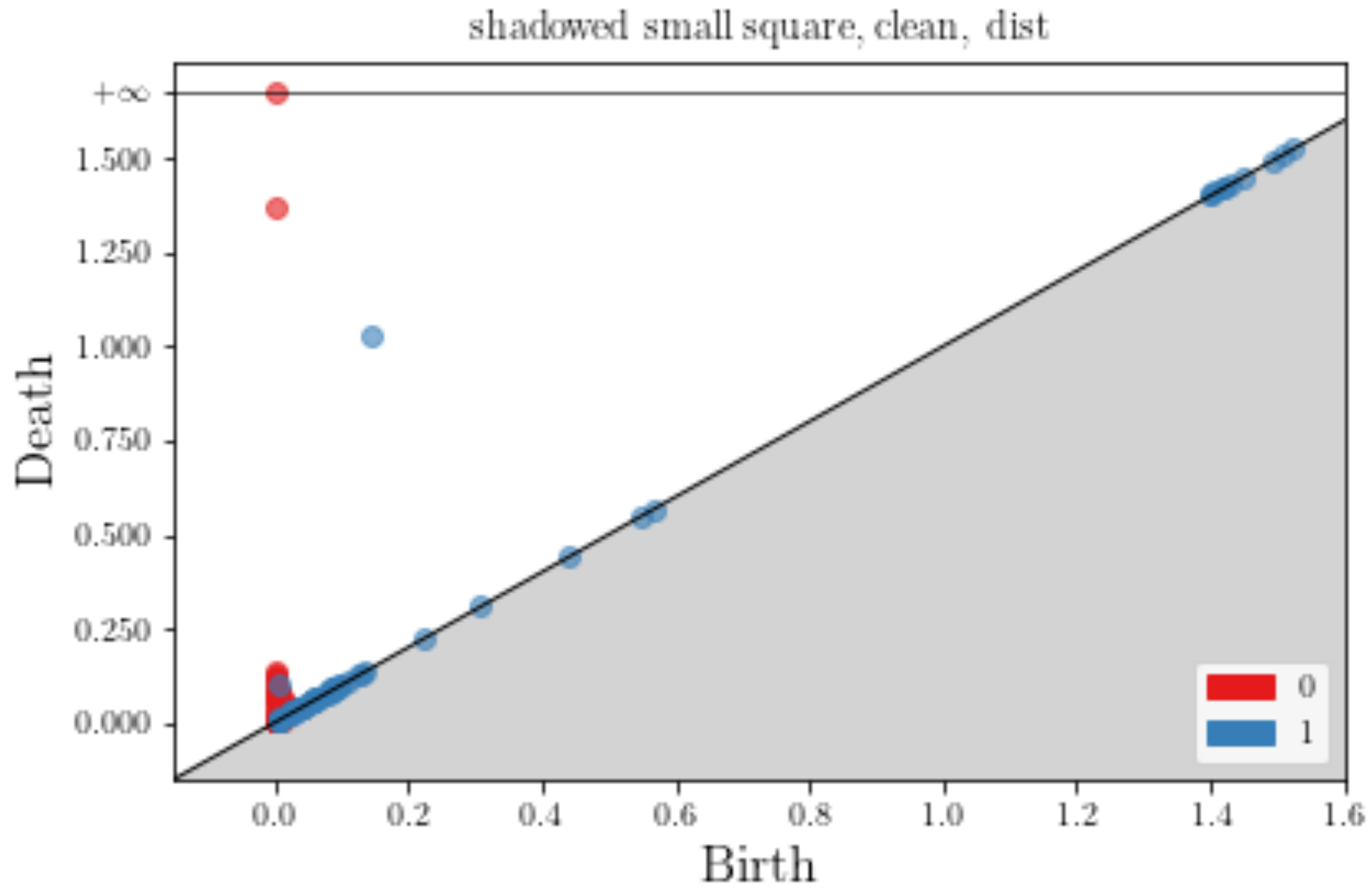
- Size
- Noise

- statistical model that highlights small features
- with a provably 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

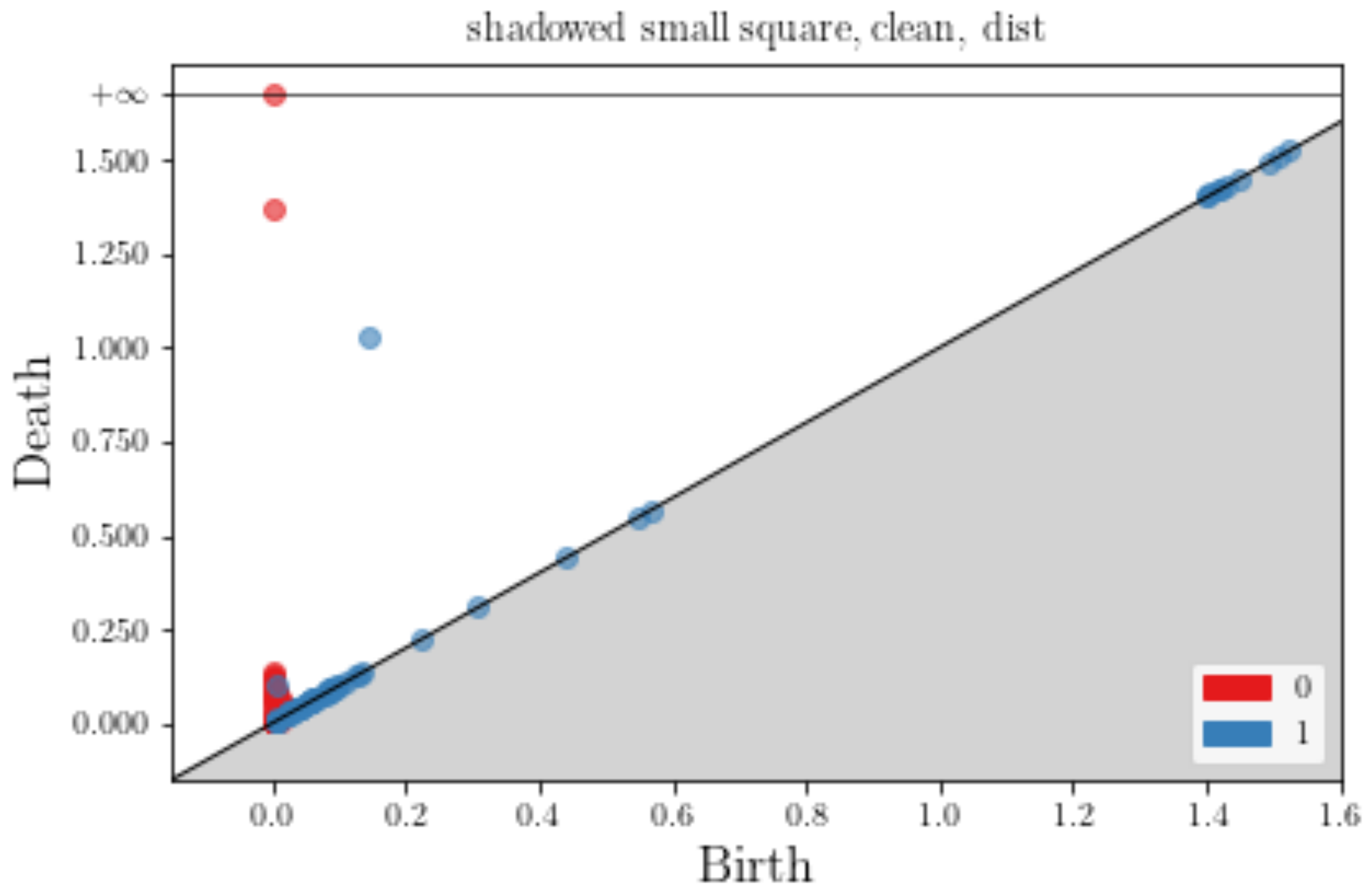
Grow Balls Sloooooooooooooooooooowly
on the smaller square

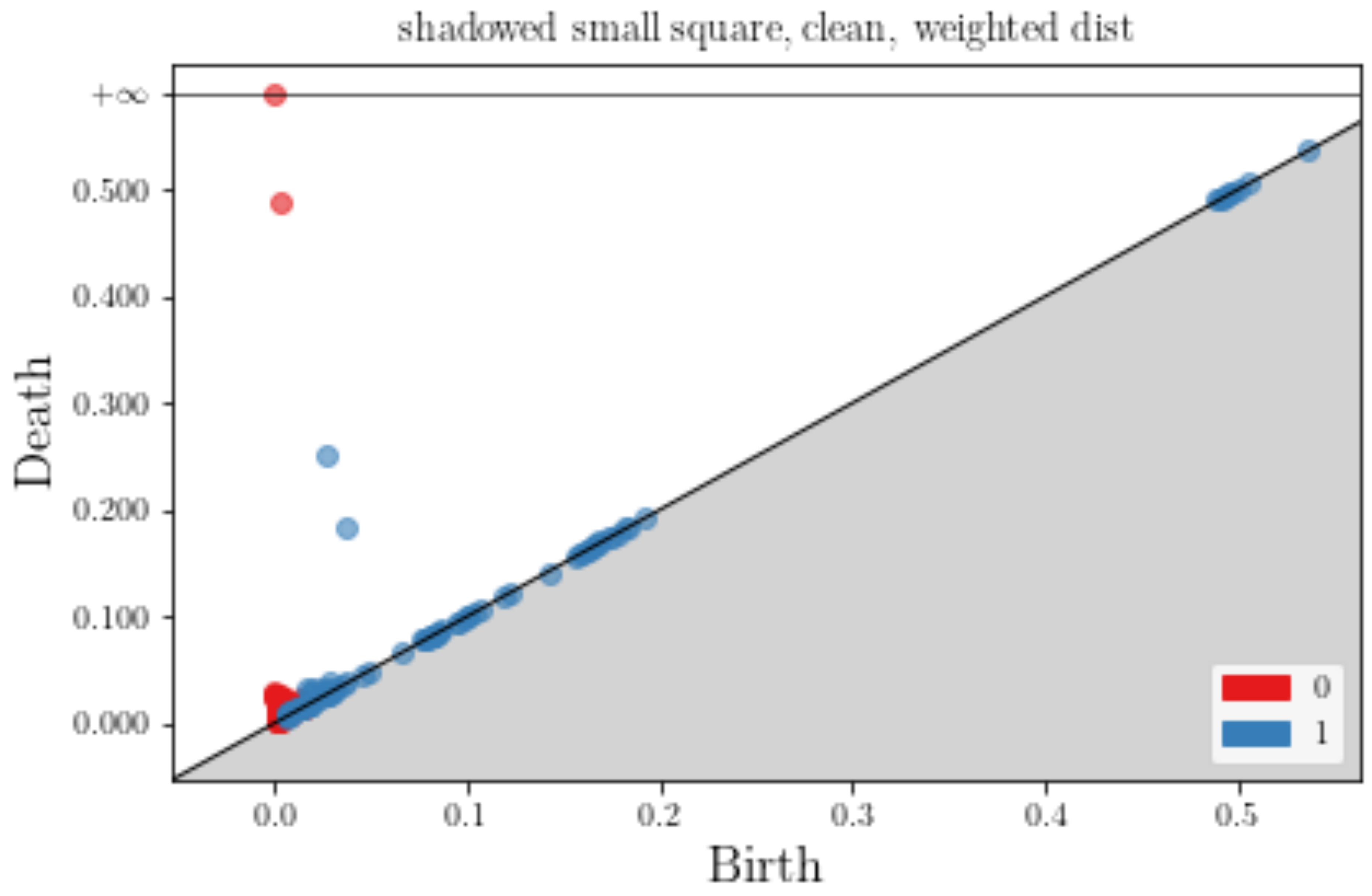
- Bell et al, 2019: growing balls at customized rates

Grow Balls Sloooooooooooooooooowly

on the smaller square

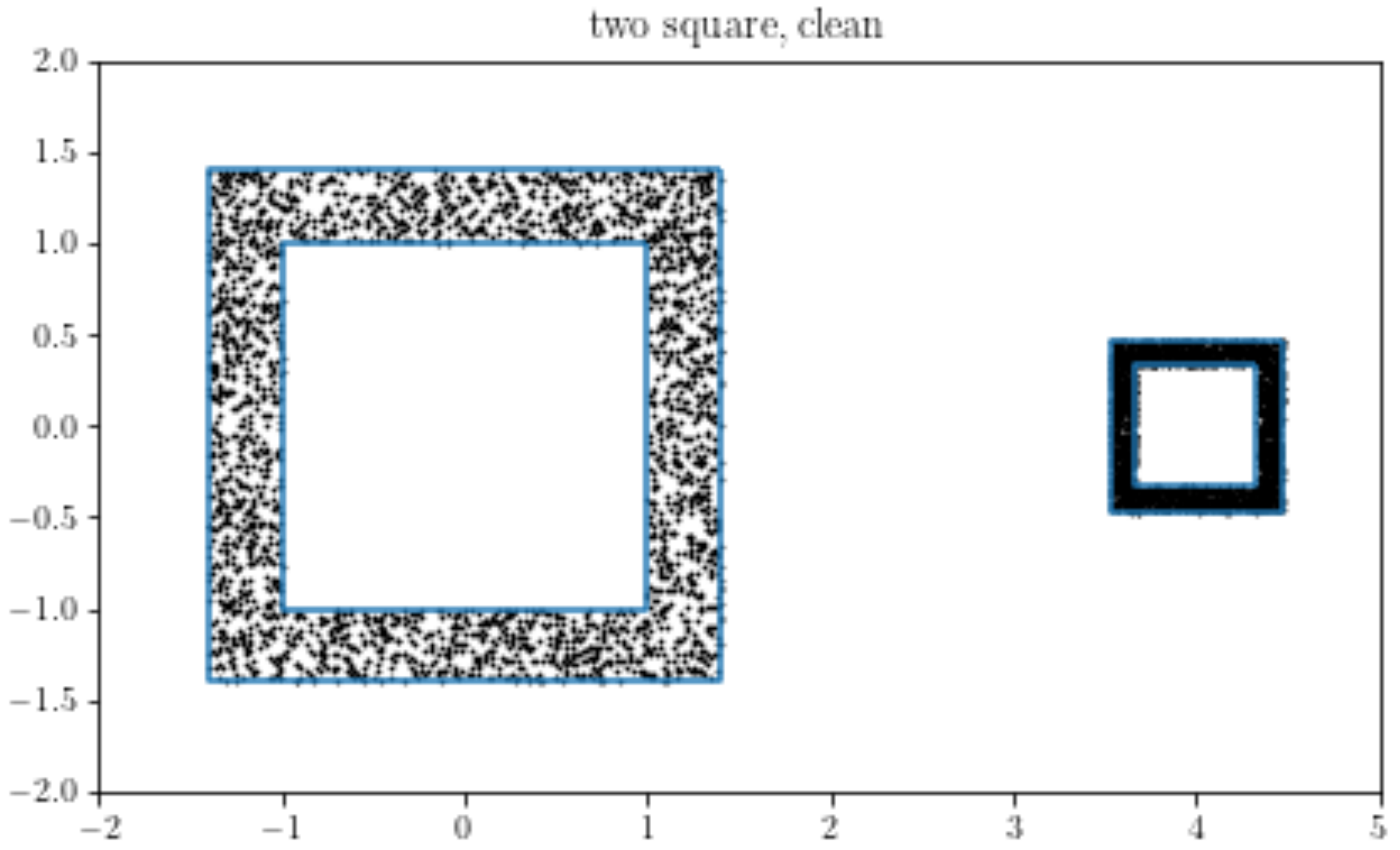
- rate = 1/density$^{1/D}$

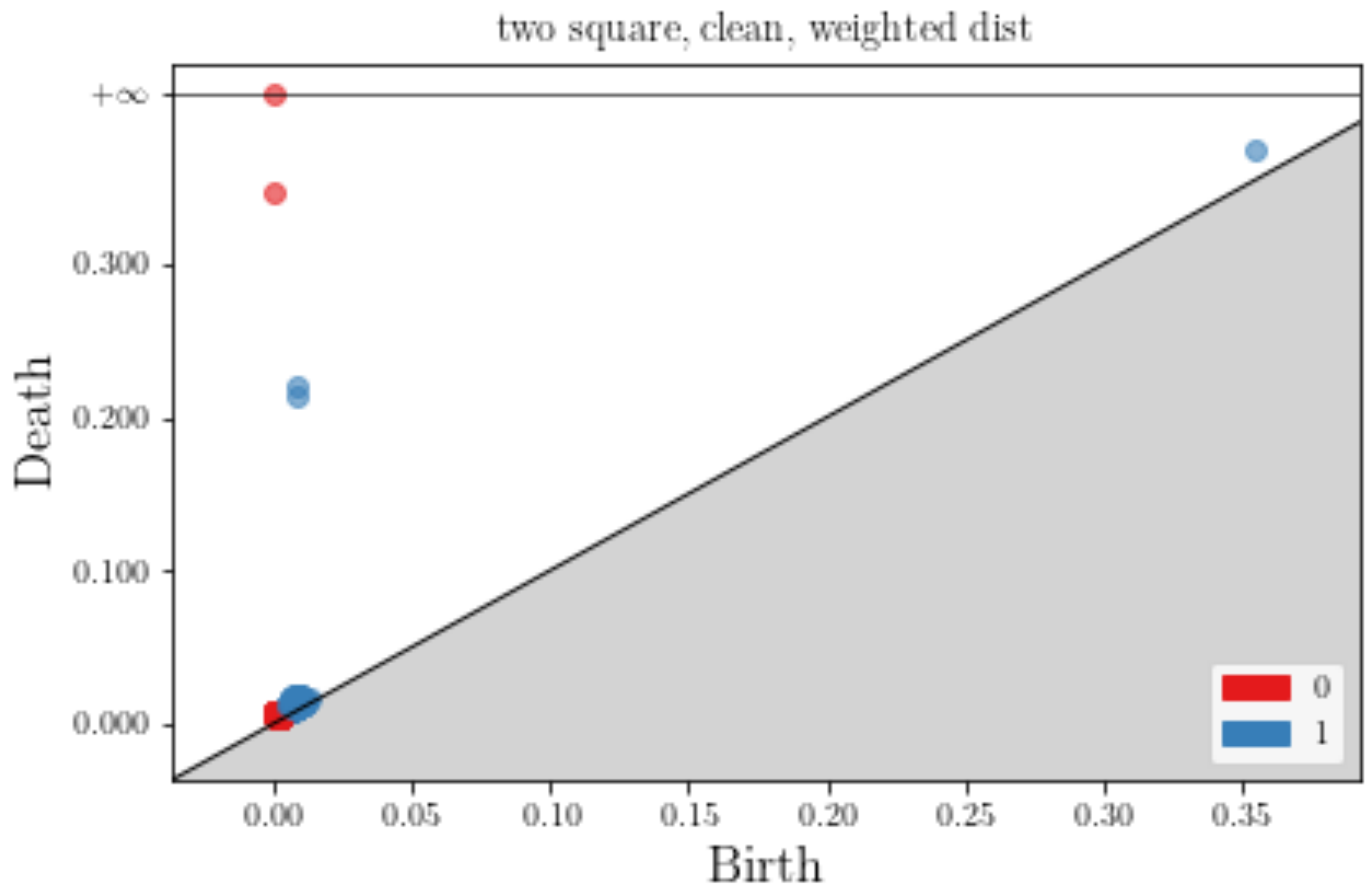




Scale invariance

- uniform scaling —> same persistence diagrams



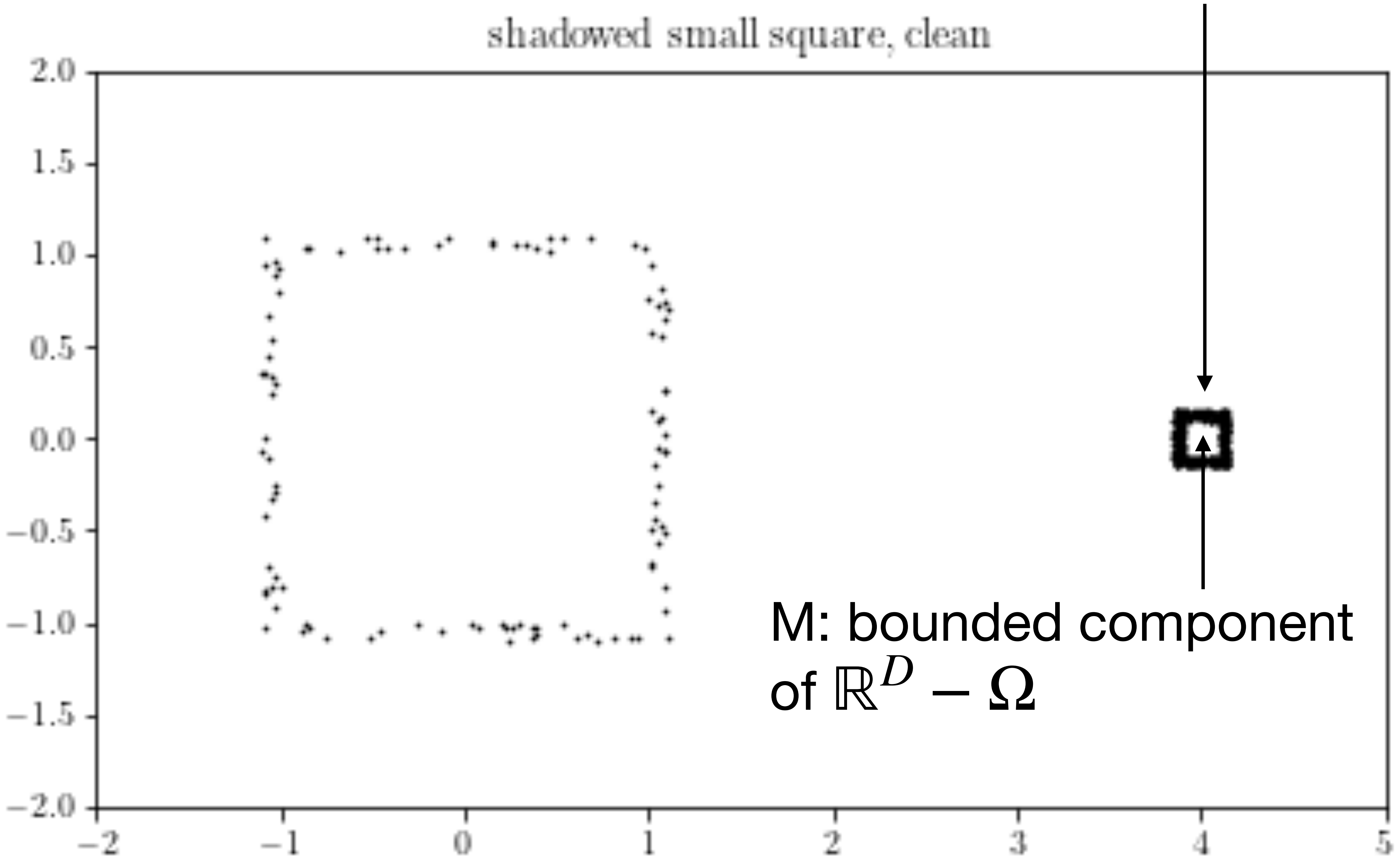


Theorem

TLDR: Small holes of high-density regions are far from diagonal.

- Let t be a density threshold.
- As in the figure, let M be a "hole" of a high-density region Ω with size $r = \max_{x \in M} d(x, \partial M)$.
- Under nice assumptions,
  M induces a (D − 1)-dimensional homology class
  with persistence at least $\frac{1}{\sqrt{2}} t^{1/D} r - O(m^{1/D})$

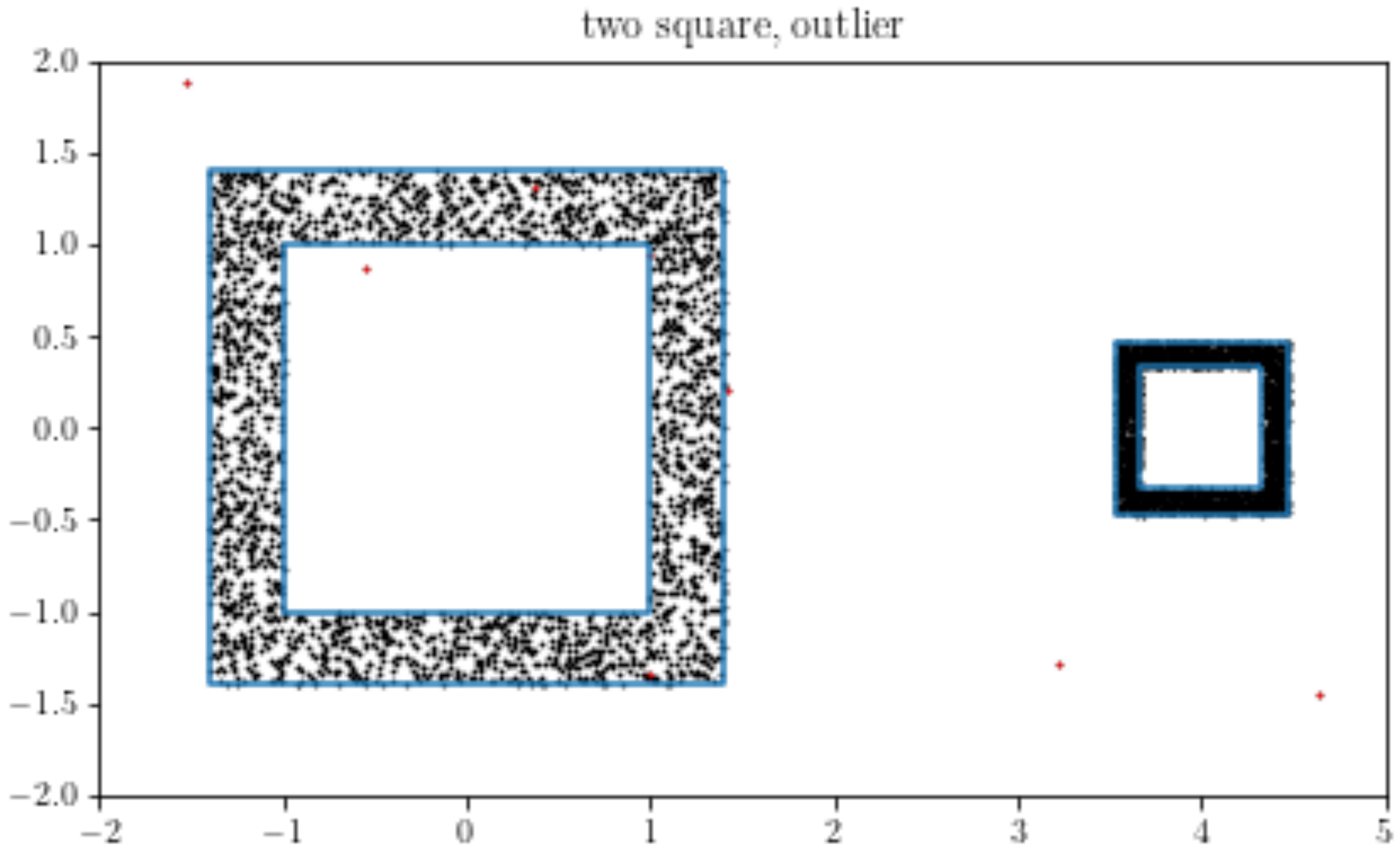
Ω: component of the the high-density region $\{\xi : f(\xi) \geq t\}$

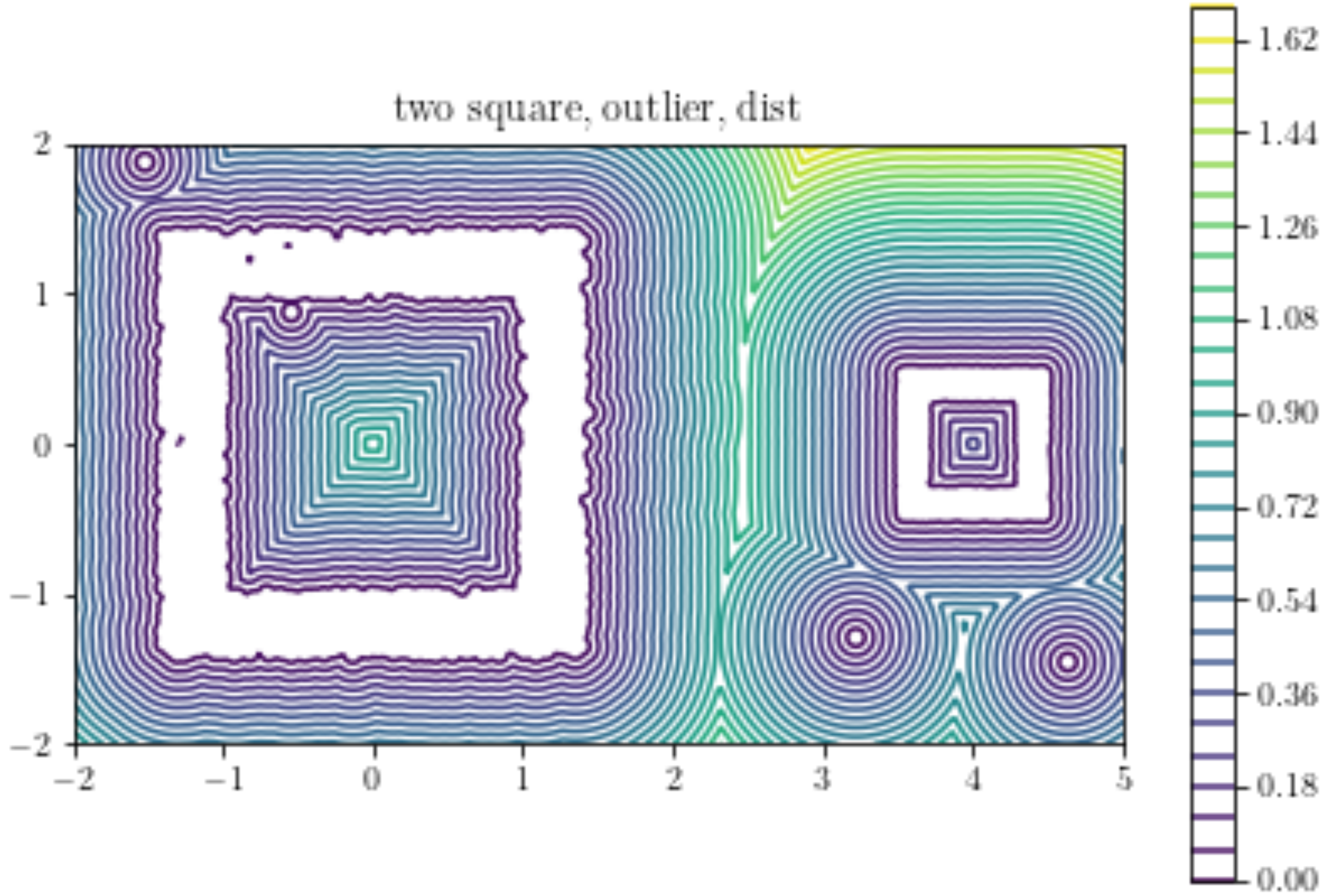


M: bounded component of $\mathbb{R}^D - \Omega$

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

Known Problem, Known Solution

- solution: distance-to-measure
  - wait for more balls, and take average
  - can leverage empirical process theory
  - Chazal et al (2011), Chazal et al (2018)

Robust Density-Aware Distance
(RDAD)

Robust Density-Aware Distance function

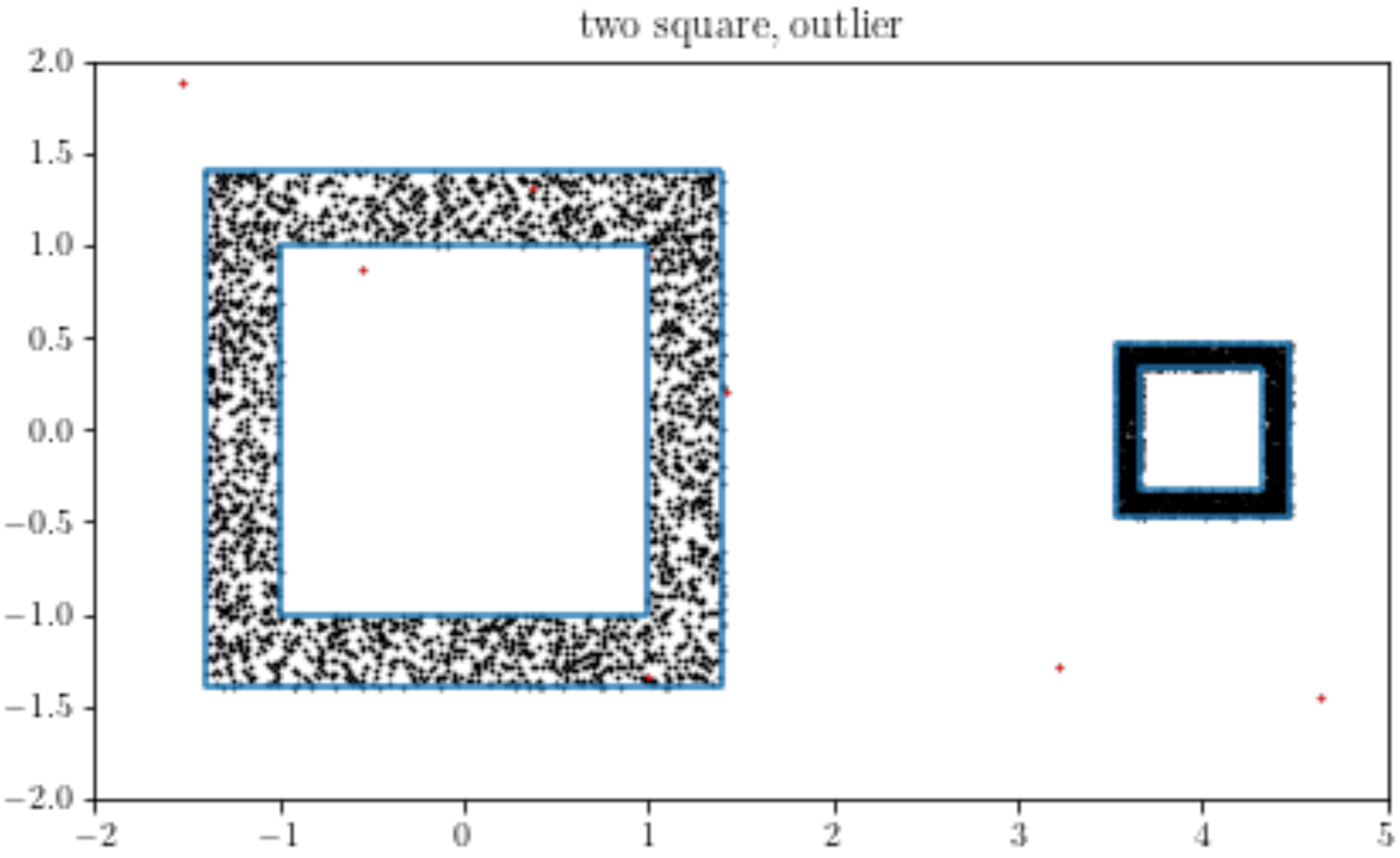
$$DTM(x) = \sqrt{\frac{1}{m}\int_0^m G_x^{-1}(q)^2 dq}$$

$$G_x(r) = P\{d(x, X) \leq r\}$$

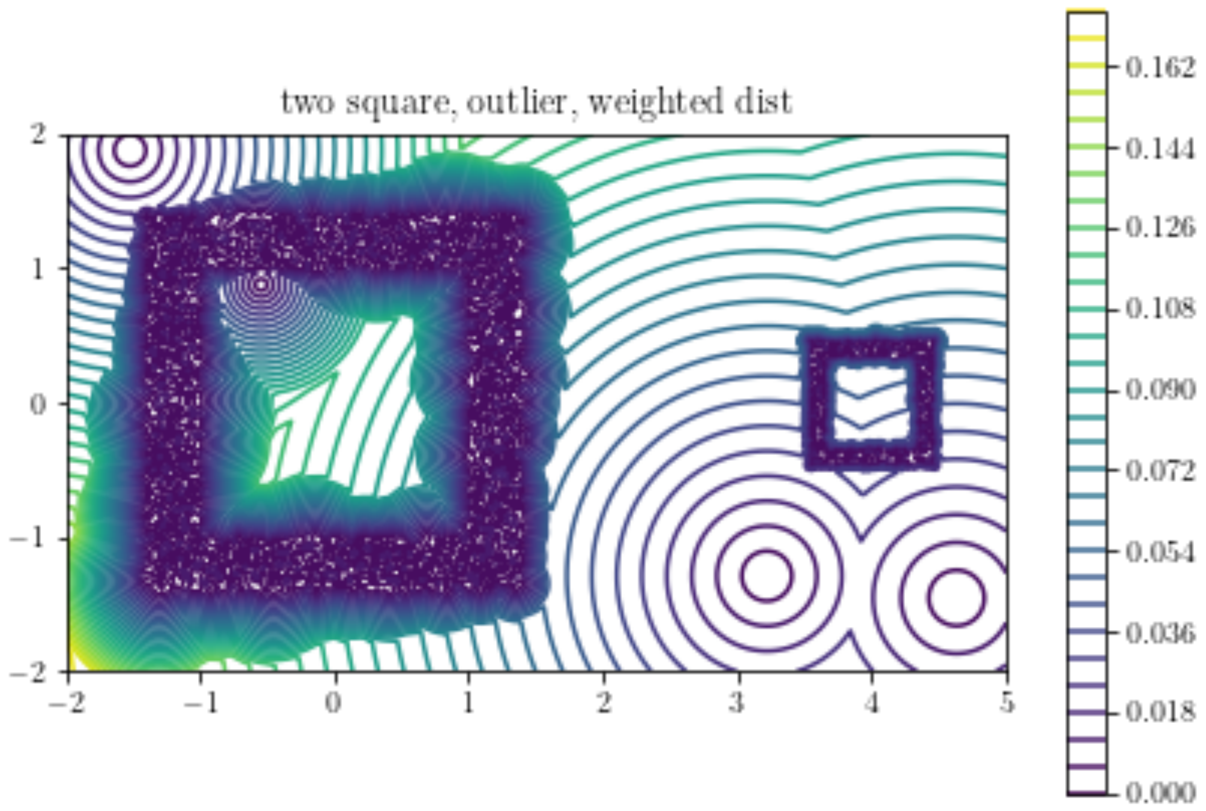
$$RDAD(x) = \sqrt{\frac{1}{m}\int_0^m F_x^{-1}(q)^2 dq}$$

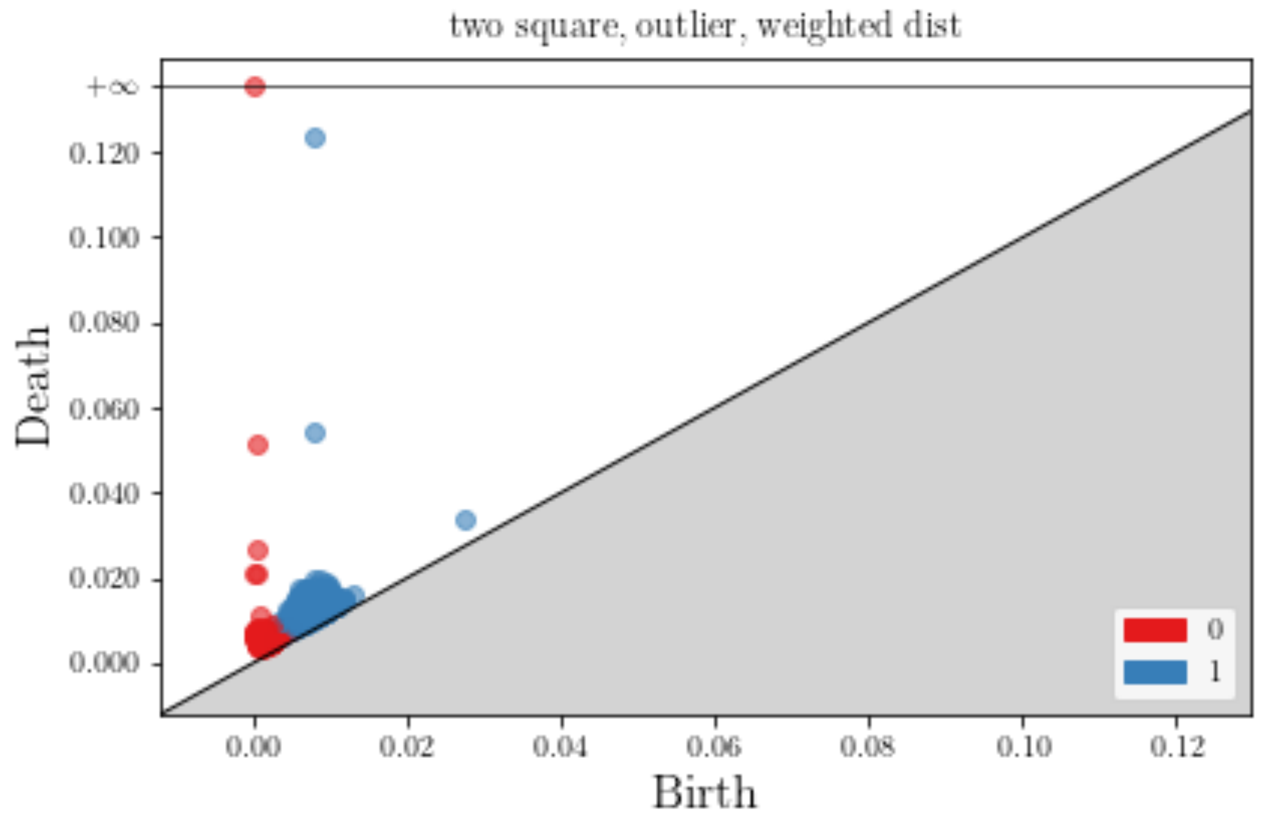
$$F_x(r) = P\{d(x, X)f(X)^{1/D} \leq r\}$$

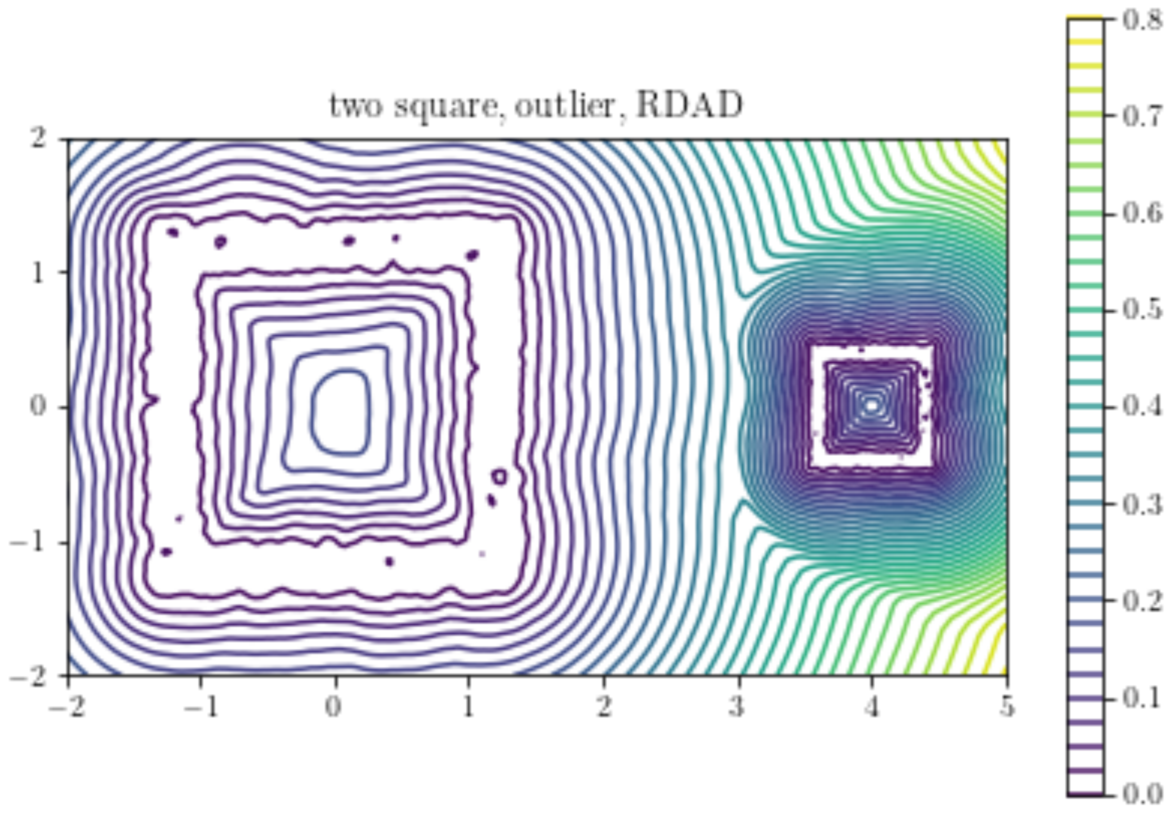
Outlier

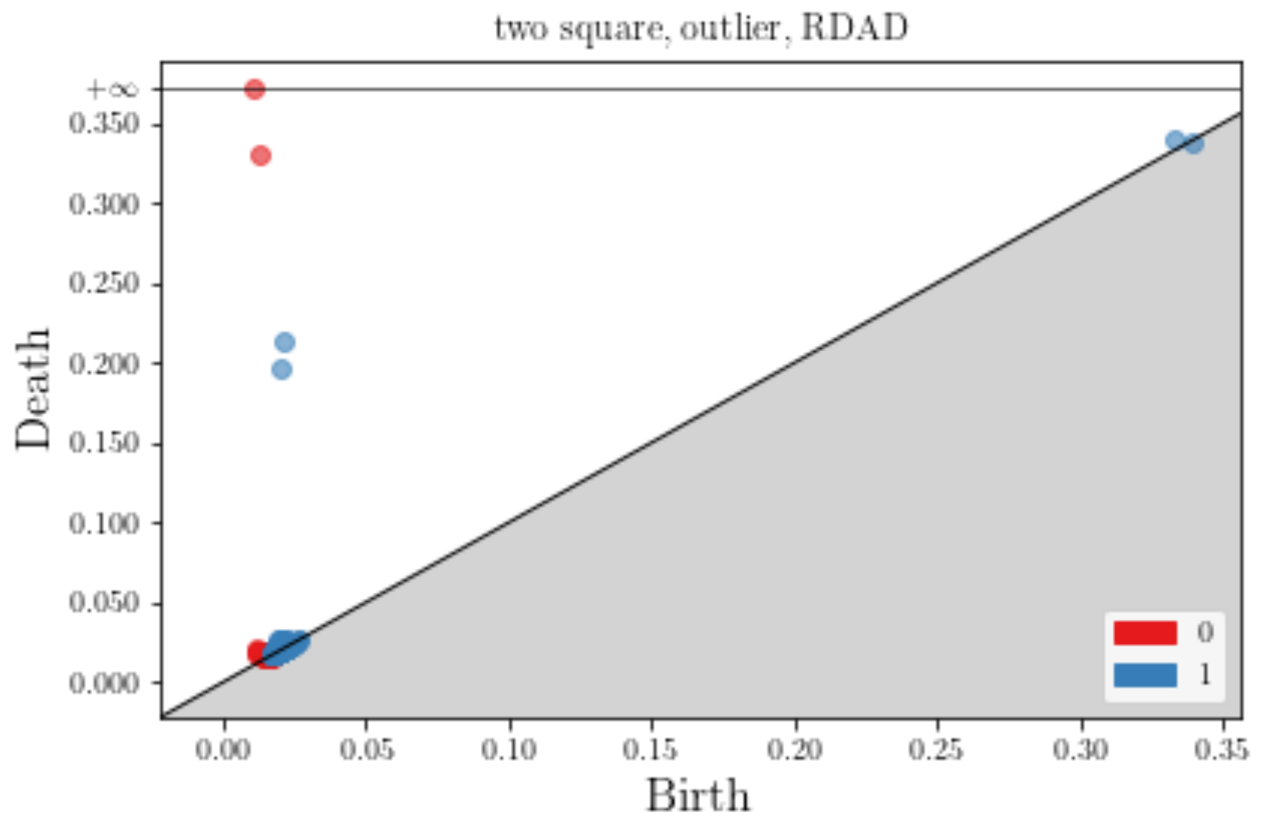


Weighted distance v.s. RDAD









Theorem

- Let f and $\tilde{f}$ be two densities.
- Under nice condition, the persistence diagrams of $RDAD_f$ and $RDAD_{\tilde{f}}$ on a compact set *K* have bottleneck distance bounded by

$$O(W_p(f, \tilde{f}) + \|f - \tilde{f}\|_\infty)$$

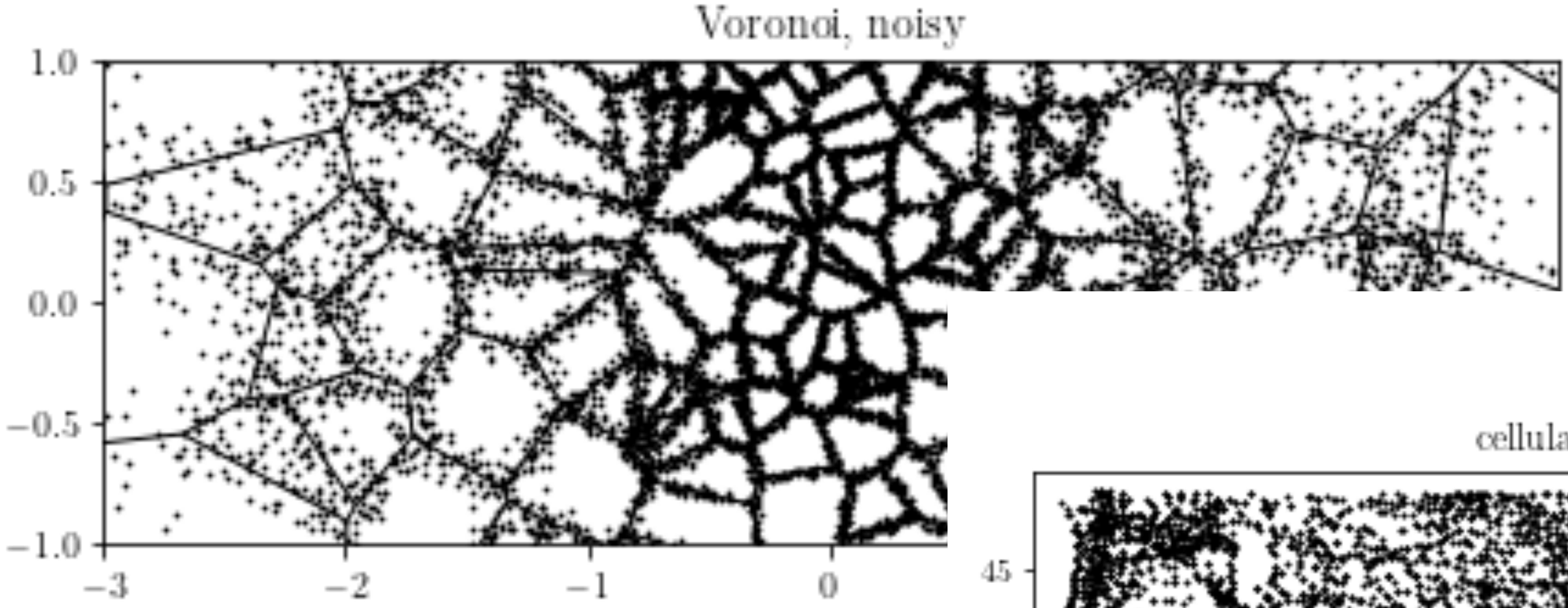
Statistical Convergence?

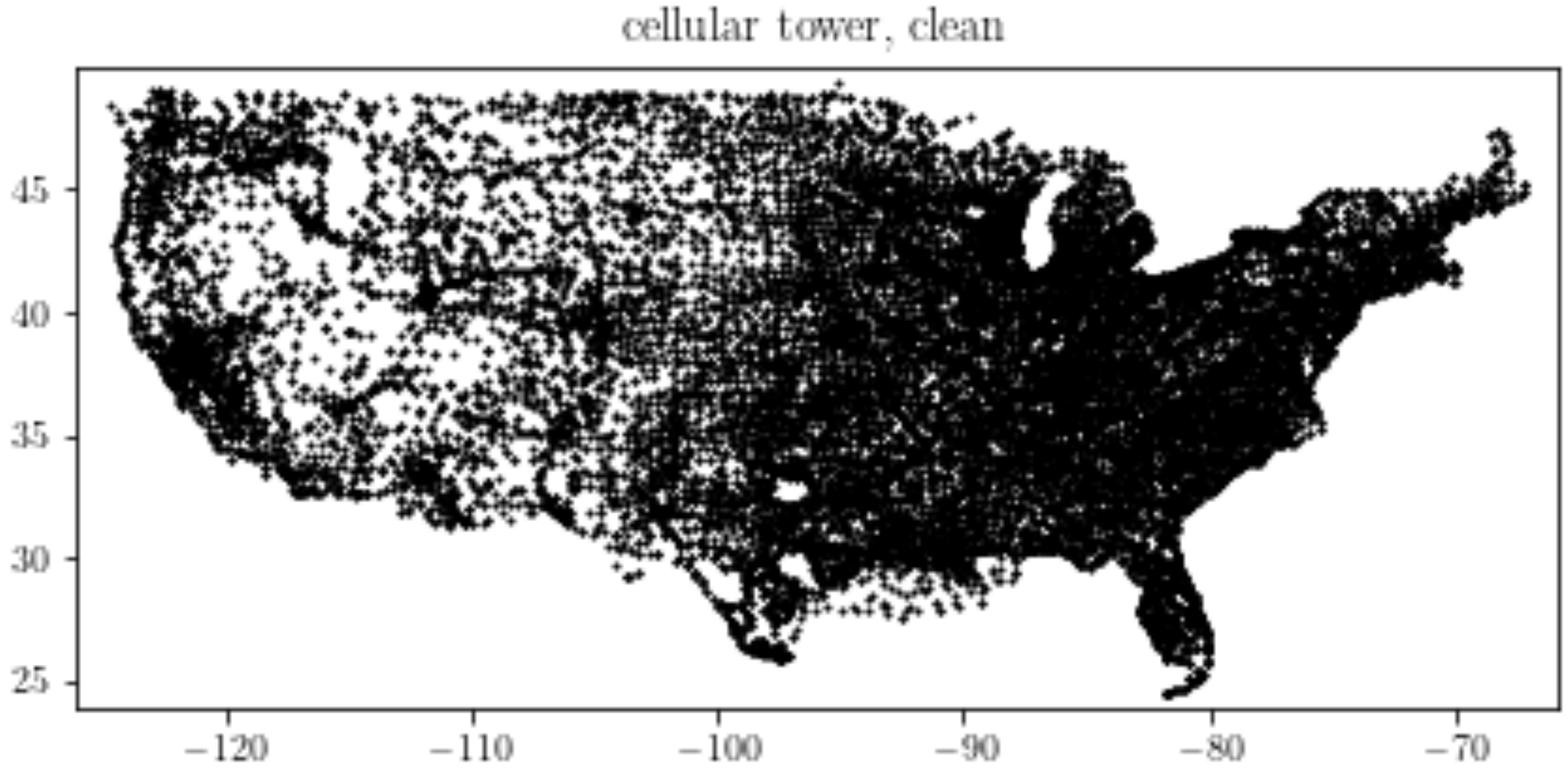
Theorem

- Let $X_1, \ldots, X_N$ be iid points sampled from a nice density.
- Then on every compact set *K*,

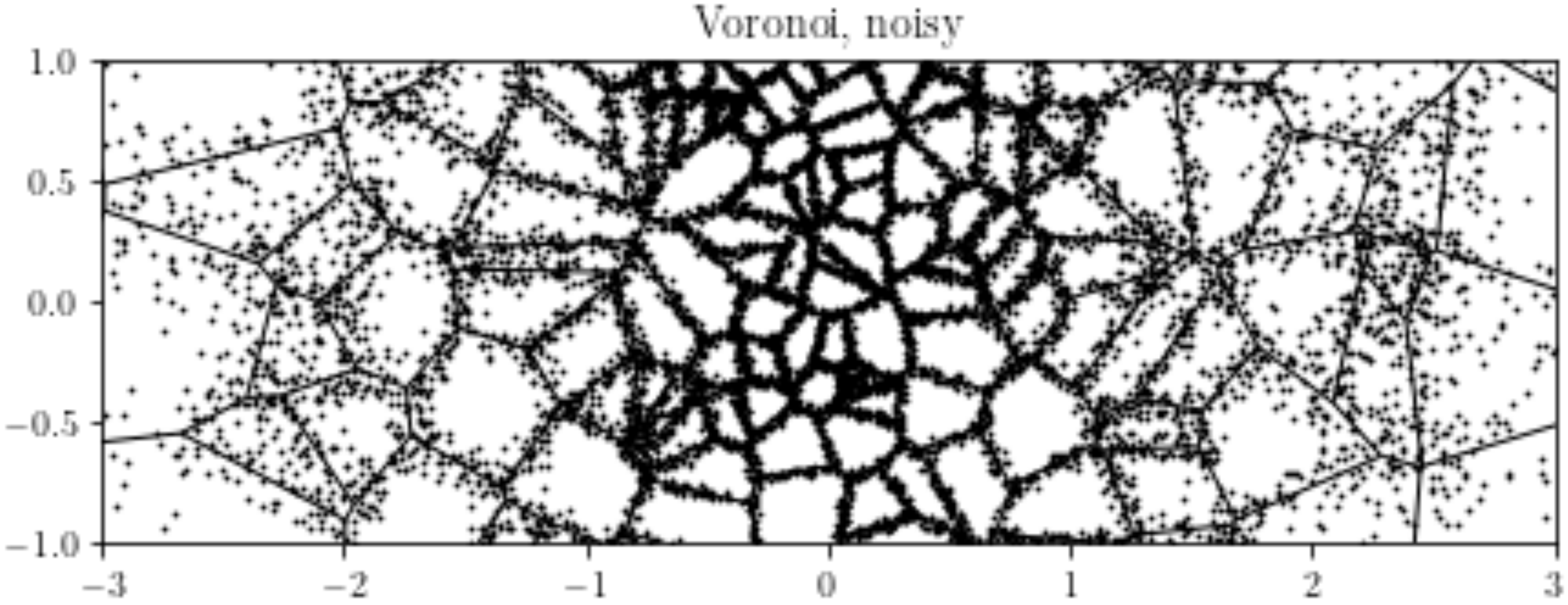
$$\sqrt{N}(\widehat{RDAD}^2 - RDAD^2) \xrightarrow{\text{weakly in } L^\infty(K)} \text{a centered Gaussian process}$$

Simulations

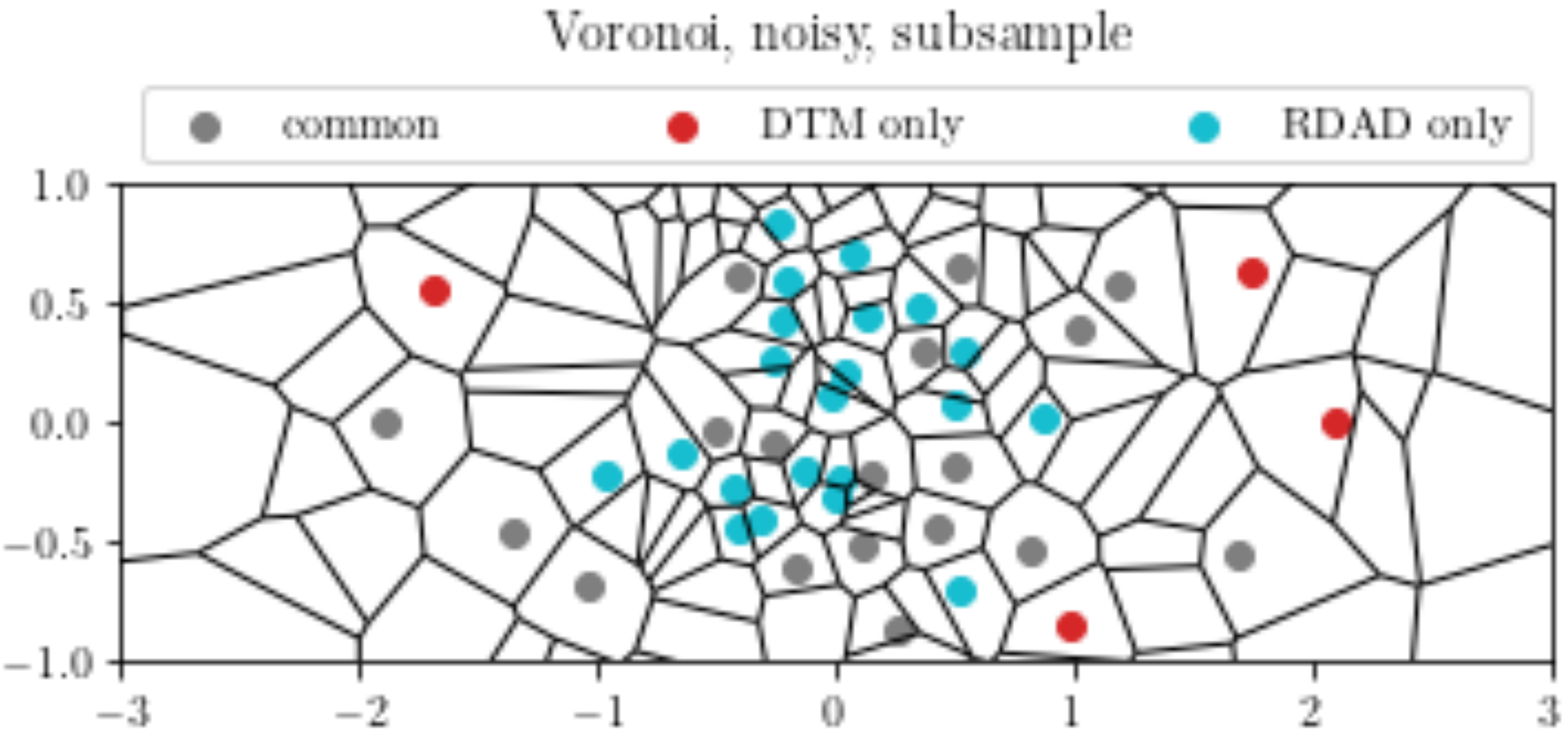

Voronoi, noisy
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0


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

Noisy Voronoi

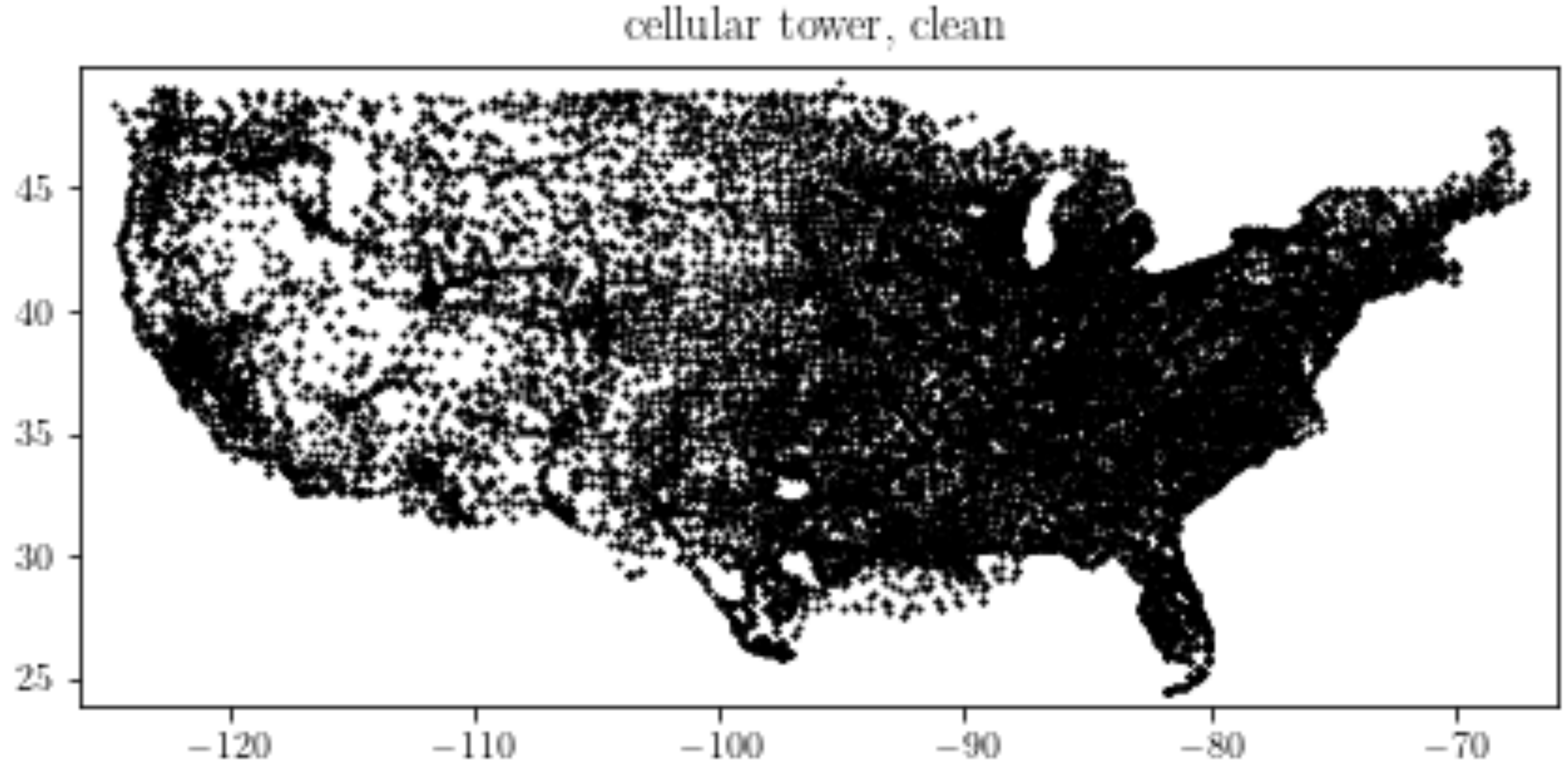


DTM and RDAD

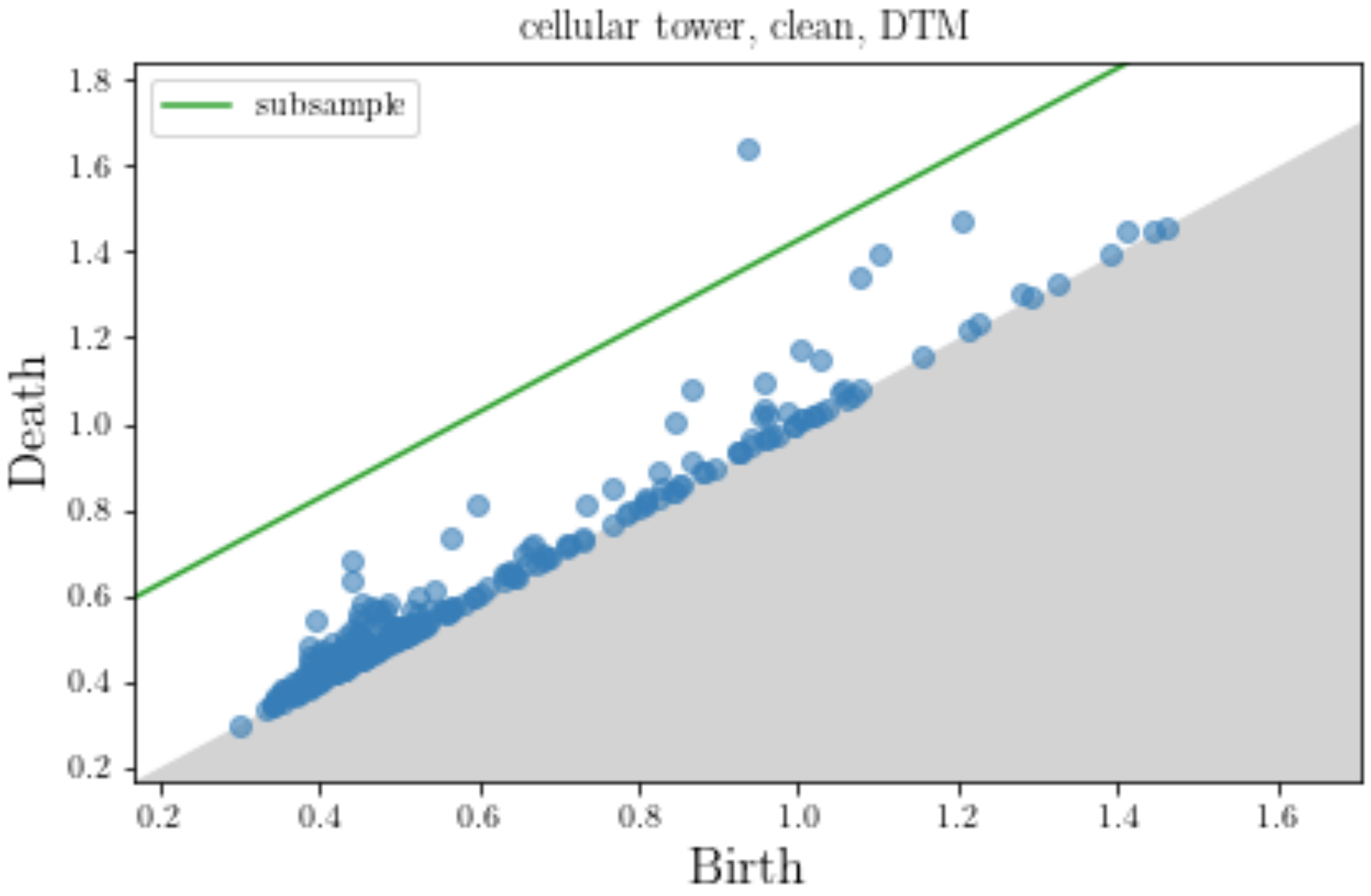

Voronoi, noisy, subsample
common
DTM only
RDAD only
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3

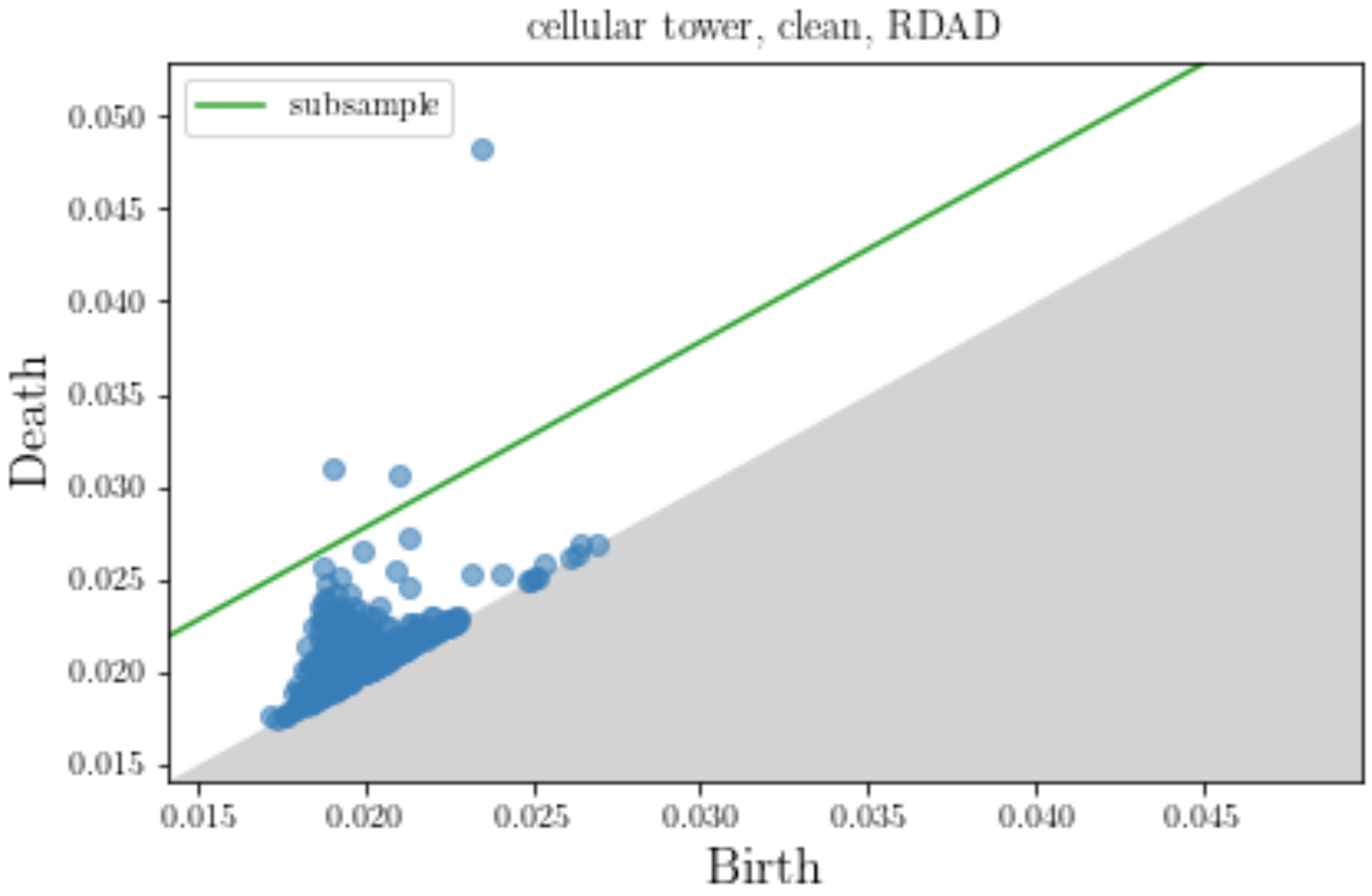
Cellular Towers

Cellular Towers
(HIFLD, 2021)

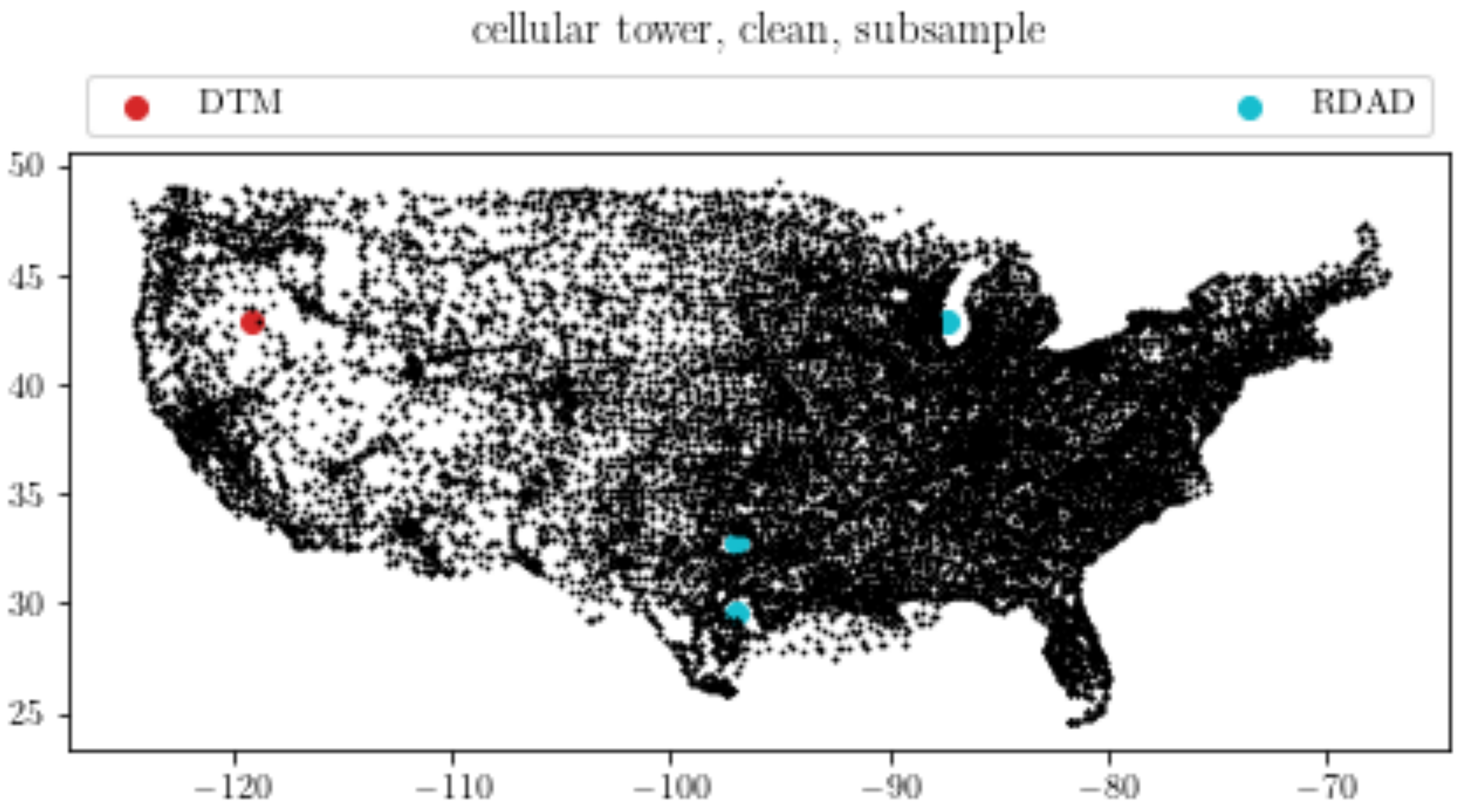


DTM and RDAD


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


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

Cellular Towers



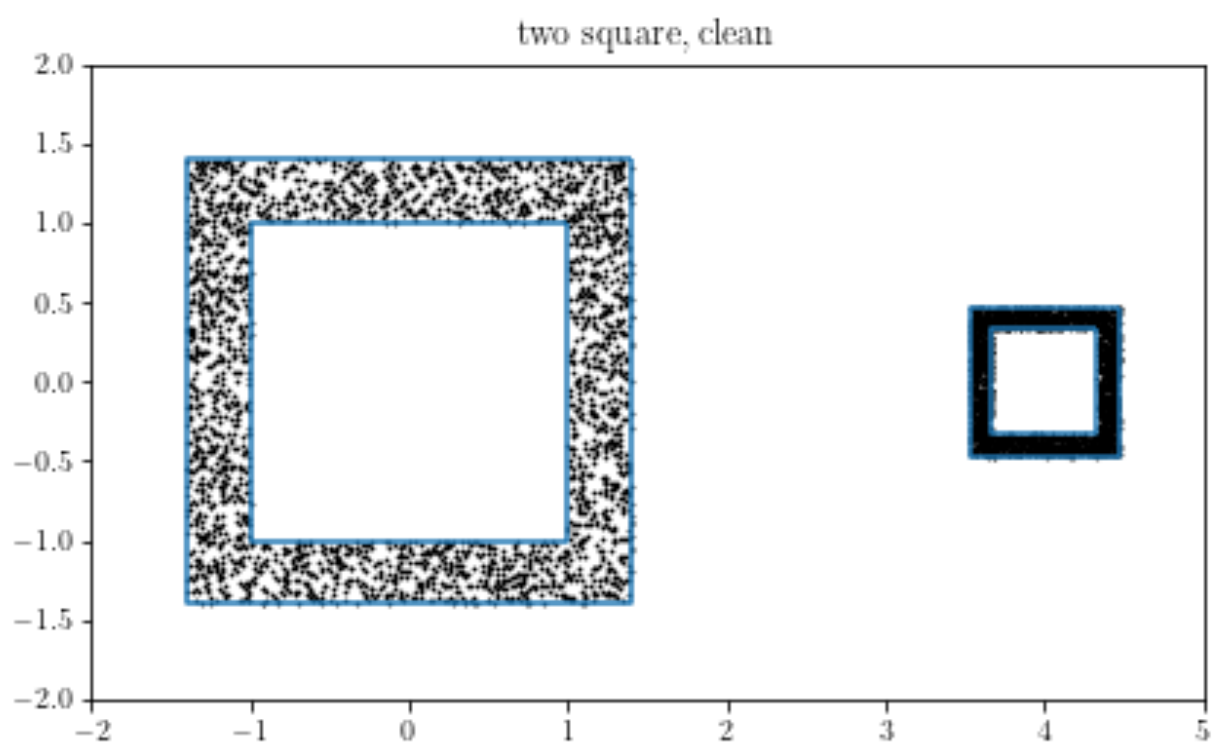
Epilogue:
The End of the Beginning

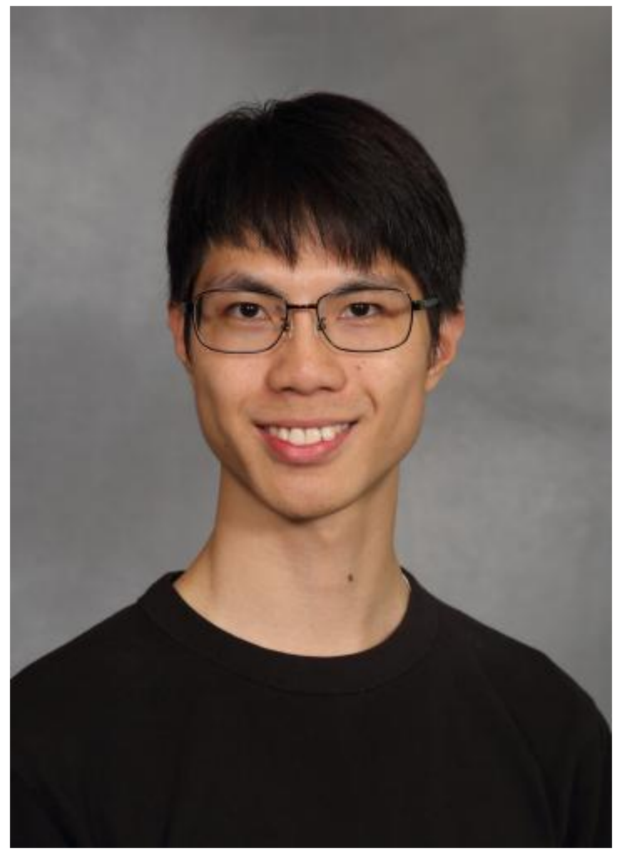
Ongoing / Future Works

- Bootstrapping properties of RDAD?
- Inference of Cosmological Parameters?
- Organic combination of topology and statistics???

Thank you!

- Chunyin Siu (Alex)
- 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.
- Berry, T., and Sauer, T. (2019). Consistent manifold representation for topological data analysis. *Foundations of Data Science* 1(1): 1–38
- 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.
- Carlsson, G., Zomorodian, A. (2009). The theory of multidimensional persistence. *Discrete Comput Geom*, 71–93
- 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.

- Hickok, A. (2022). A Family of Density-Scaled Filtered Complexes
- 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.