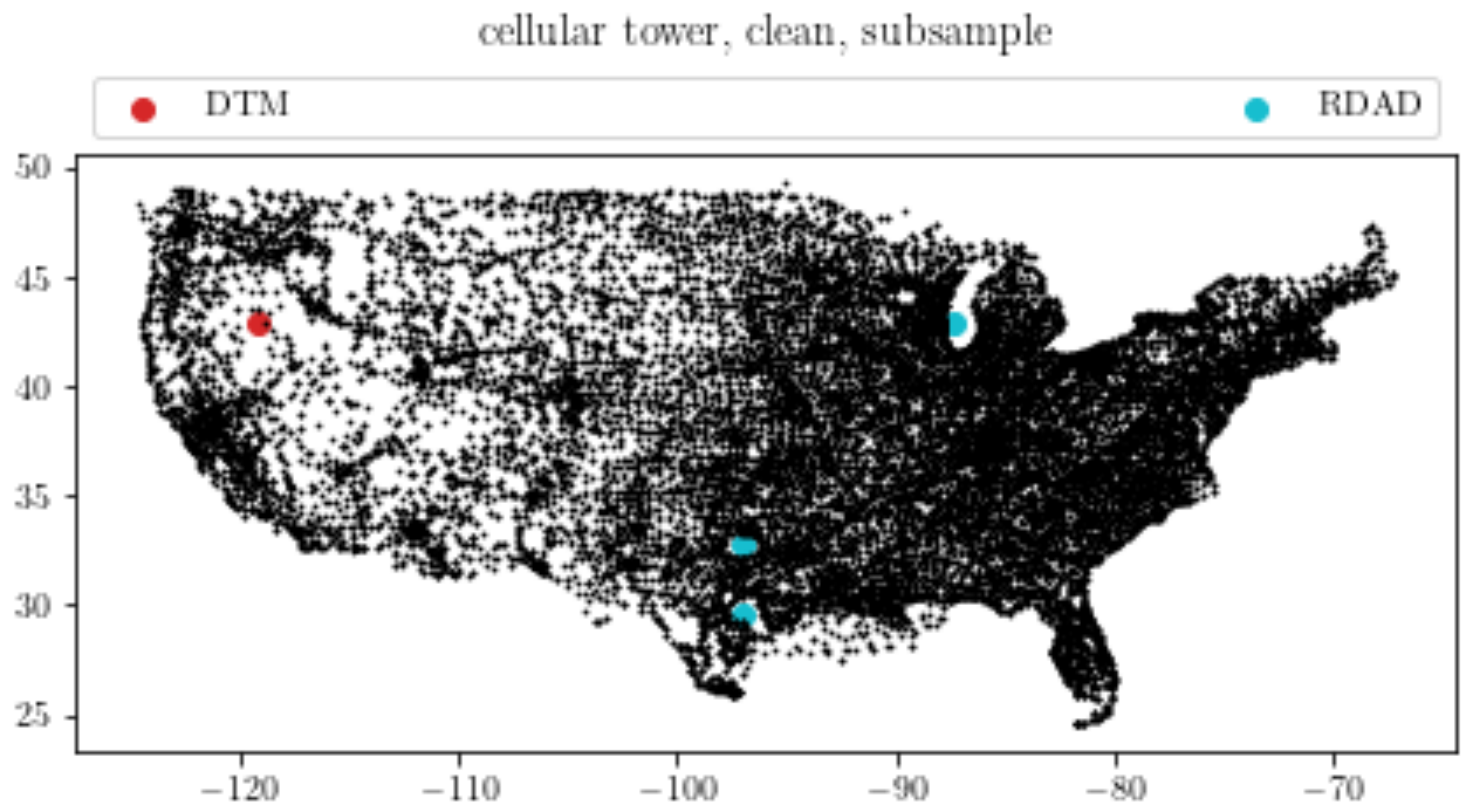
Topological Data Analysis

Detecting Weak Topological Signals in Noisy Environments



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/

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/

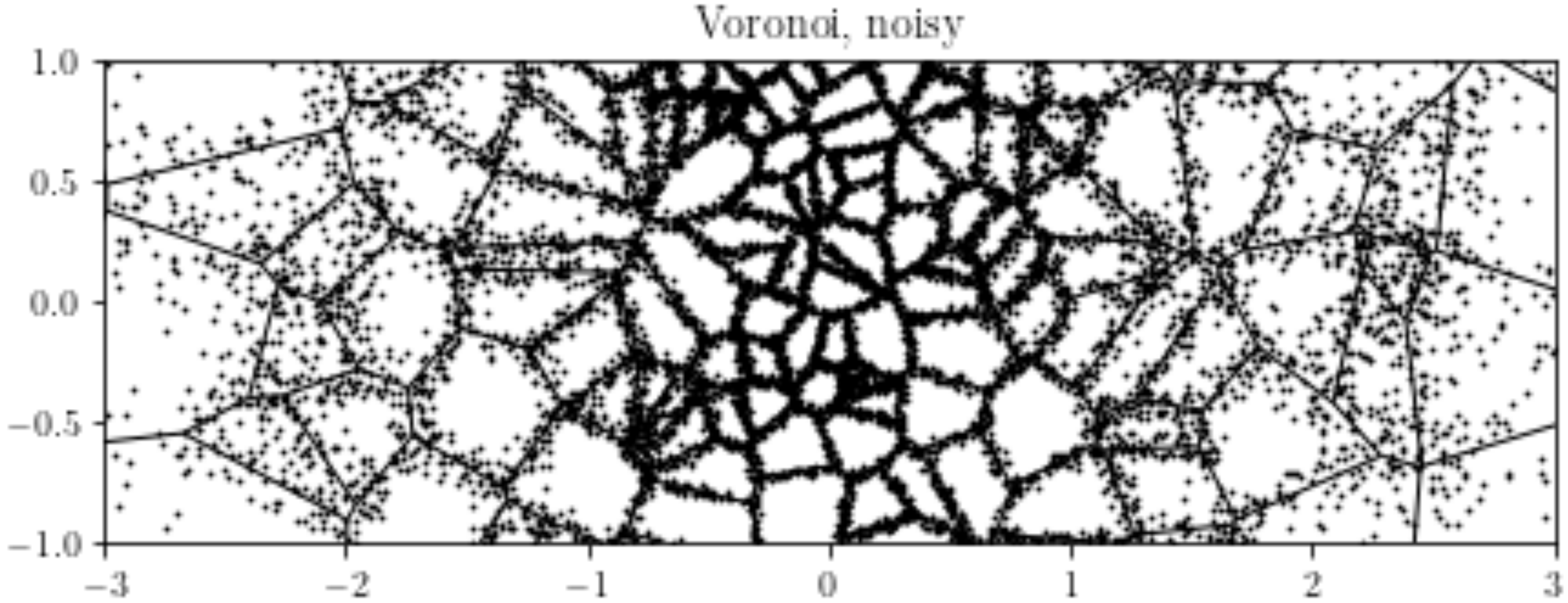
In the beginning…

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



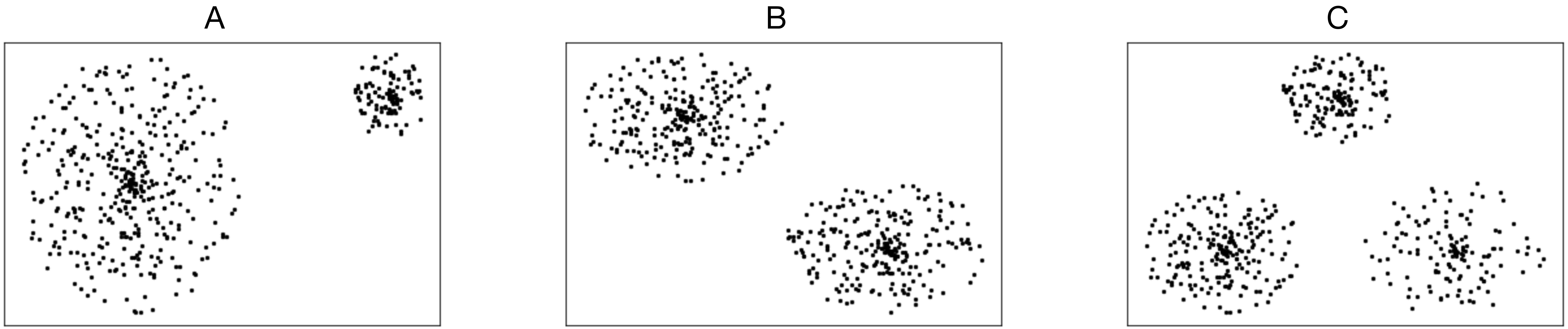
Agenda

- Topological Data Analysis: What and Why
- My Work: Weak Topological Signals amidst Noise
- Numerical Simulations

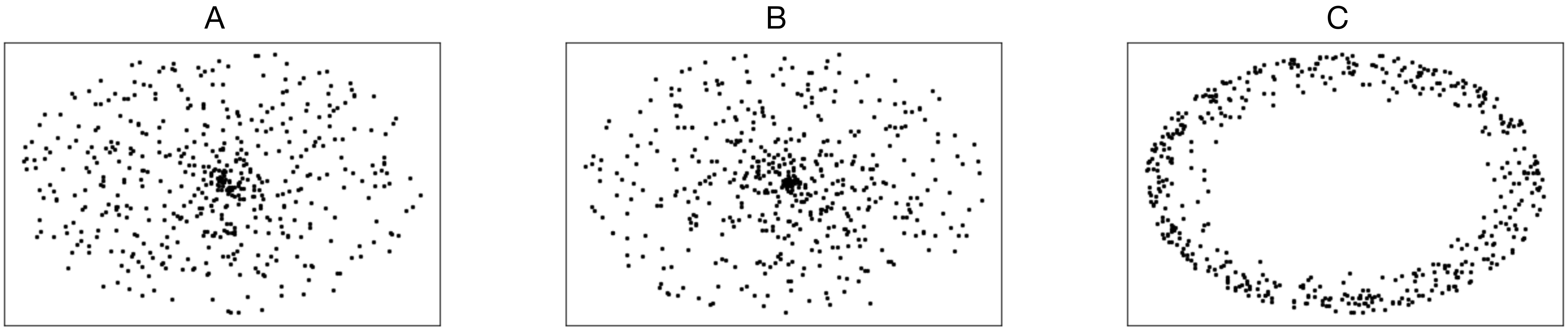
Act I

What the Fisher is Topological Data Analysis

Odd One Out



Odd One Out



Odd One Out

A

B

C

Odd One Out

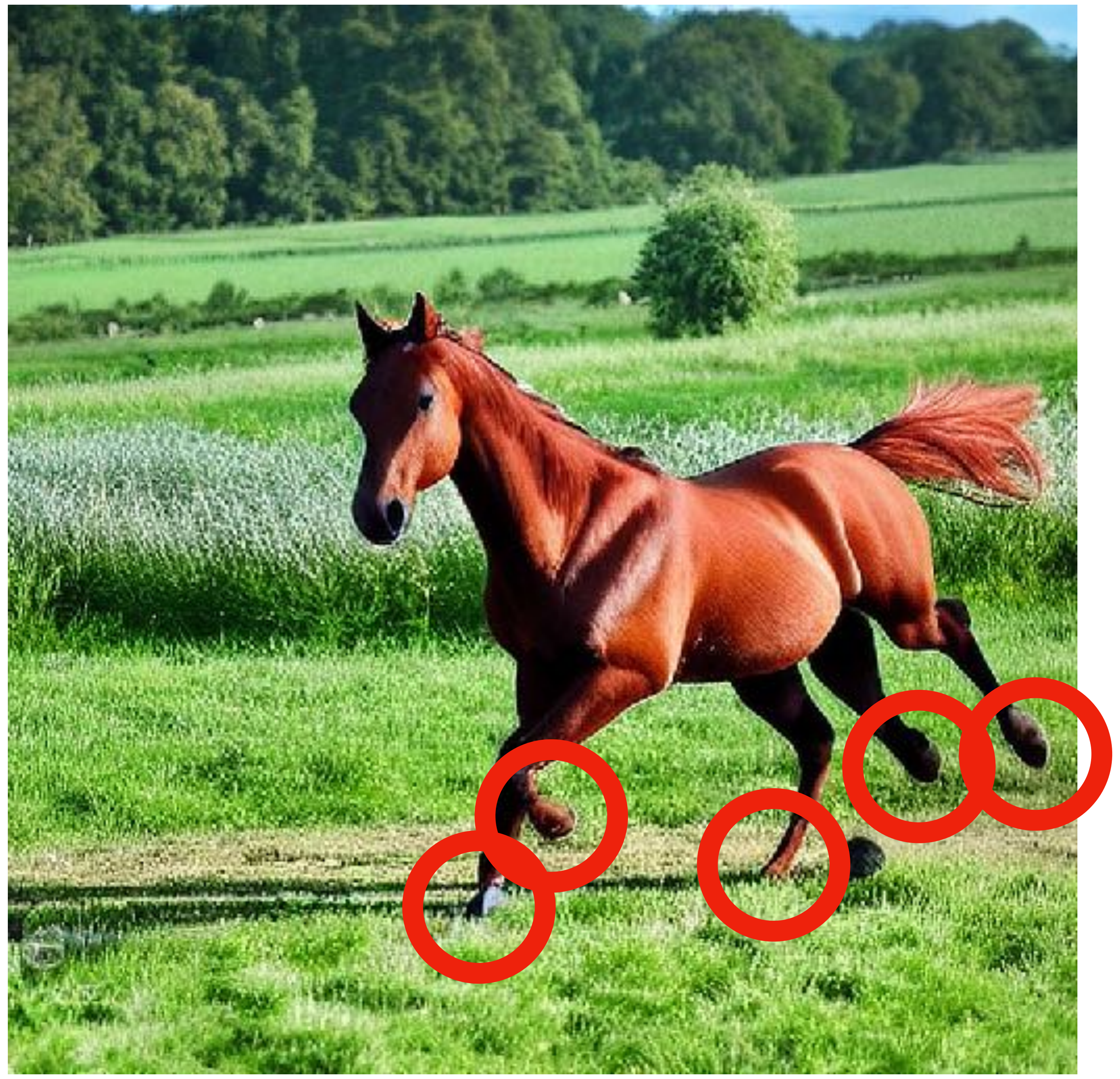
A

Photo by David Dibert from Pexels: https://www.pexels.com/photo/brown-horse-on-grass-field-635499/

B

Photo by Pixabay from Pexels: https://www.pexels.com/photo/white-horse-461717/

C



https://twitter.com/fchollet/status/1573836241875120128

Odd One Out

A

Photo by David Dibert from Pexels: https://www.pexels.com/photo/brown-horse-on-grass-field-635499/

B

Photo by Pixabay from Pexels: https://www.pexels.com/photo/white-horse-461717/

C

https://twitter.com/fchollet/status/1573836241875120128

Seriously,
you’re doing this for a PhD?

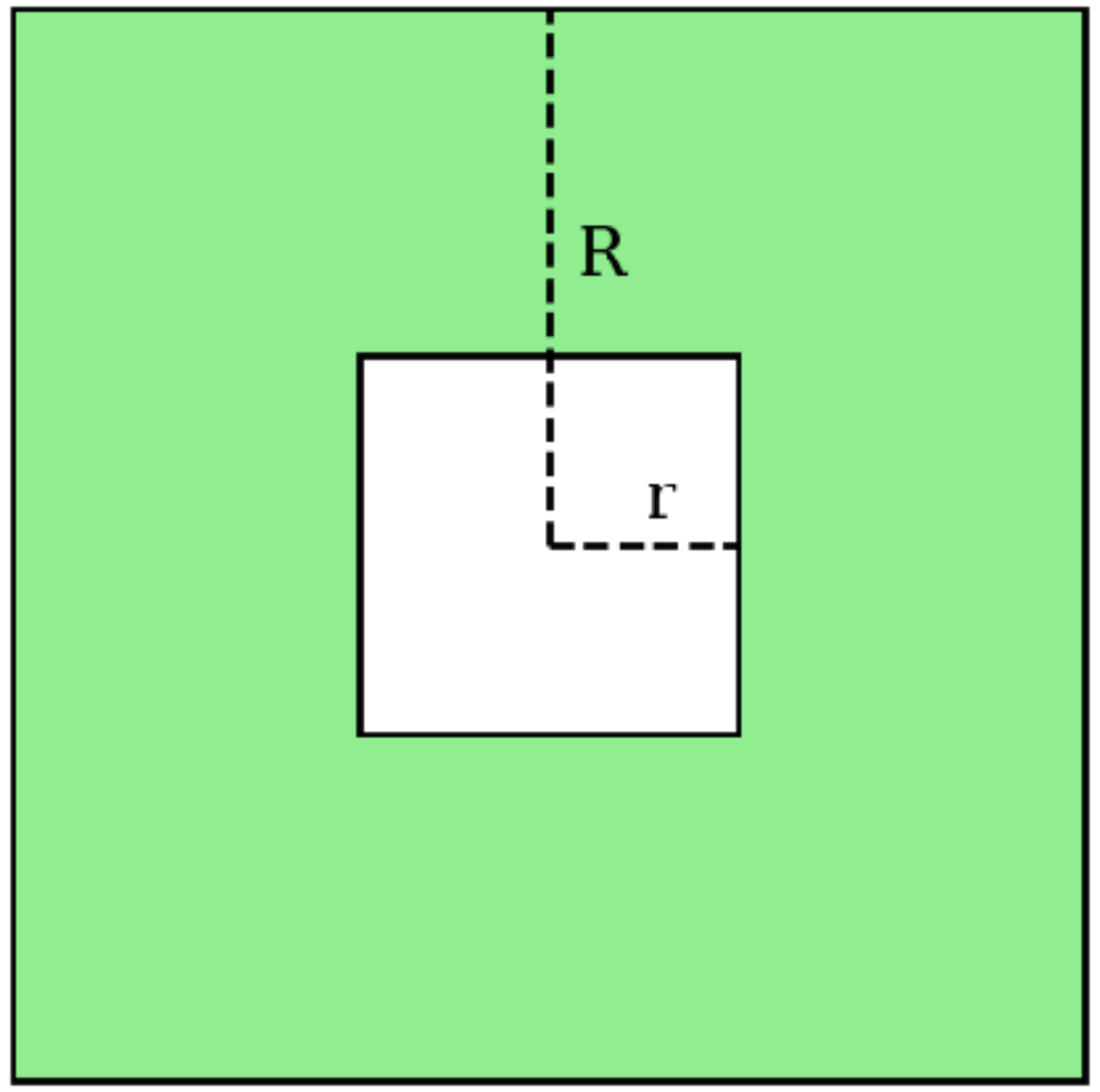
0 training data
0 parameters
100% accuracy

(for simple datasets)

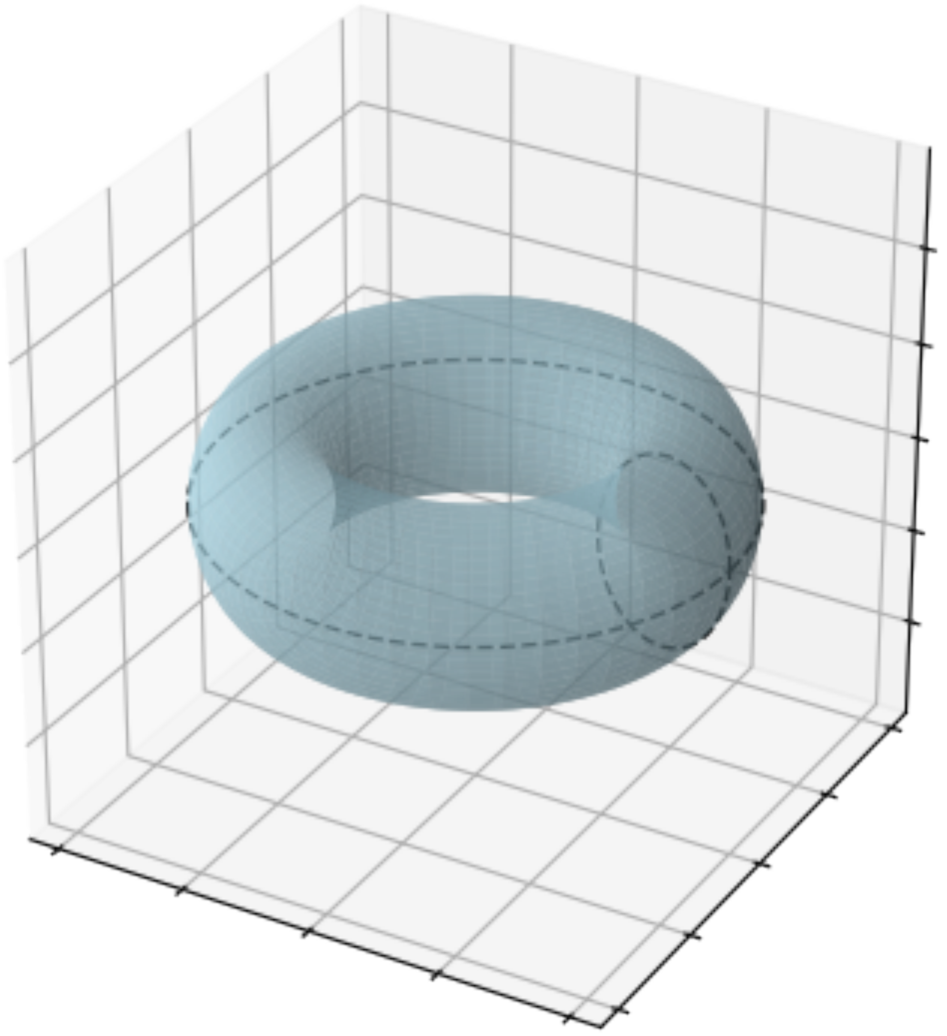
Topological Features

of the Support of the Density

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



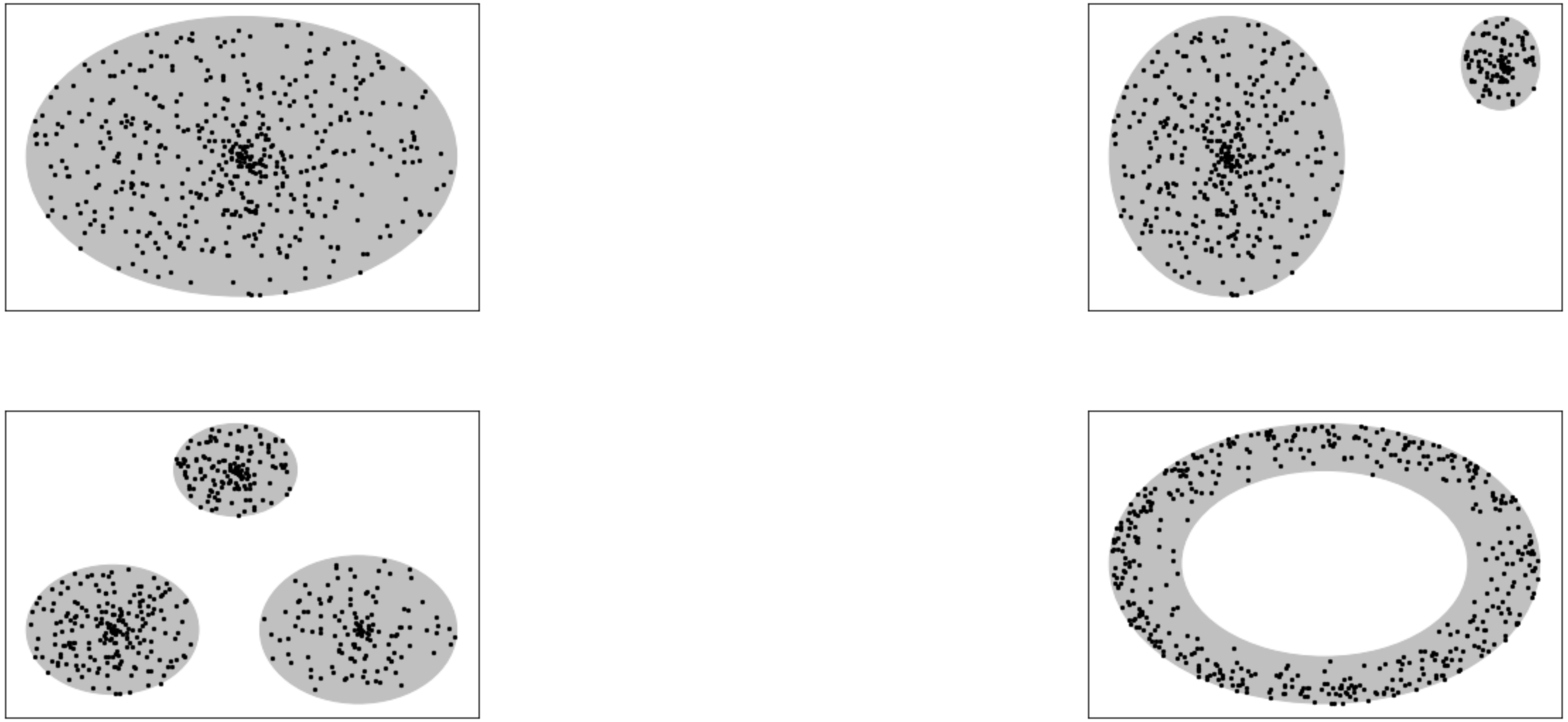
one component
one loop



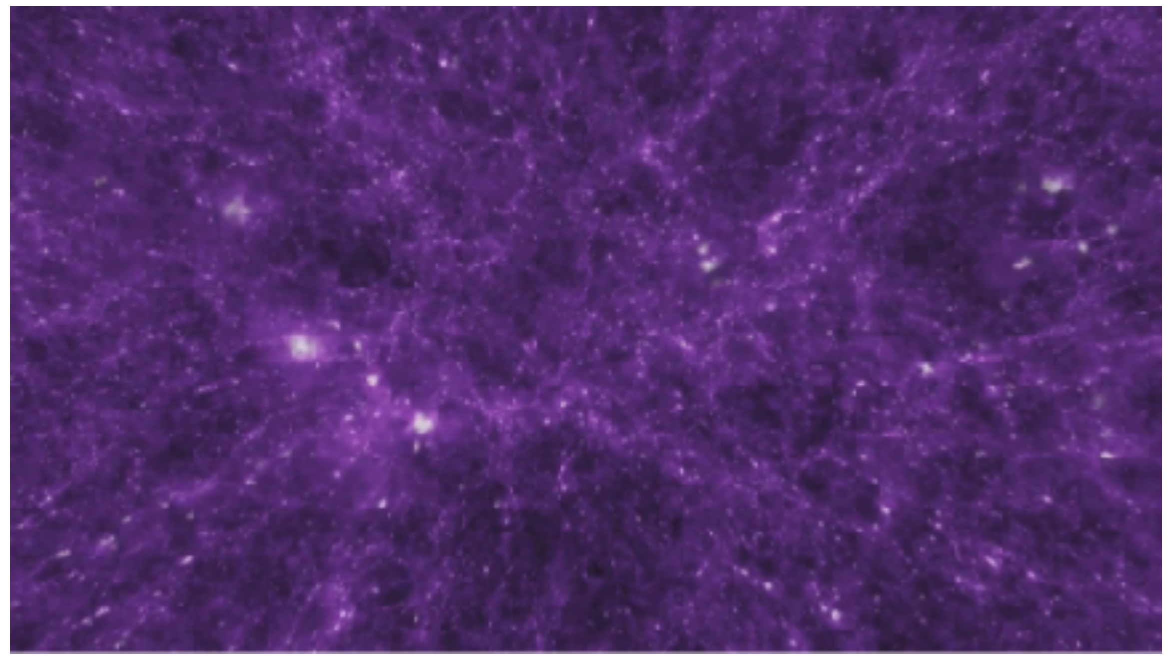
one component
two loops
one cavity

Topological Features

of the Support of the Density



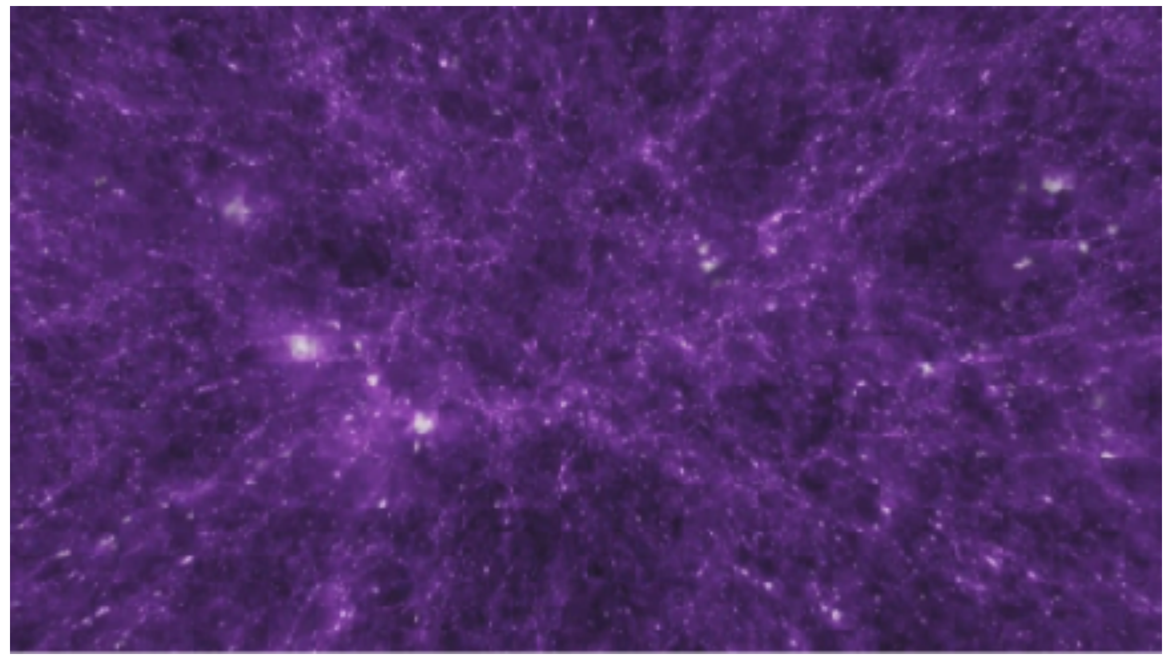
Topology in Data

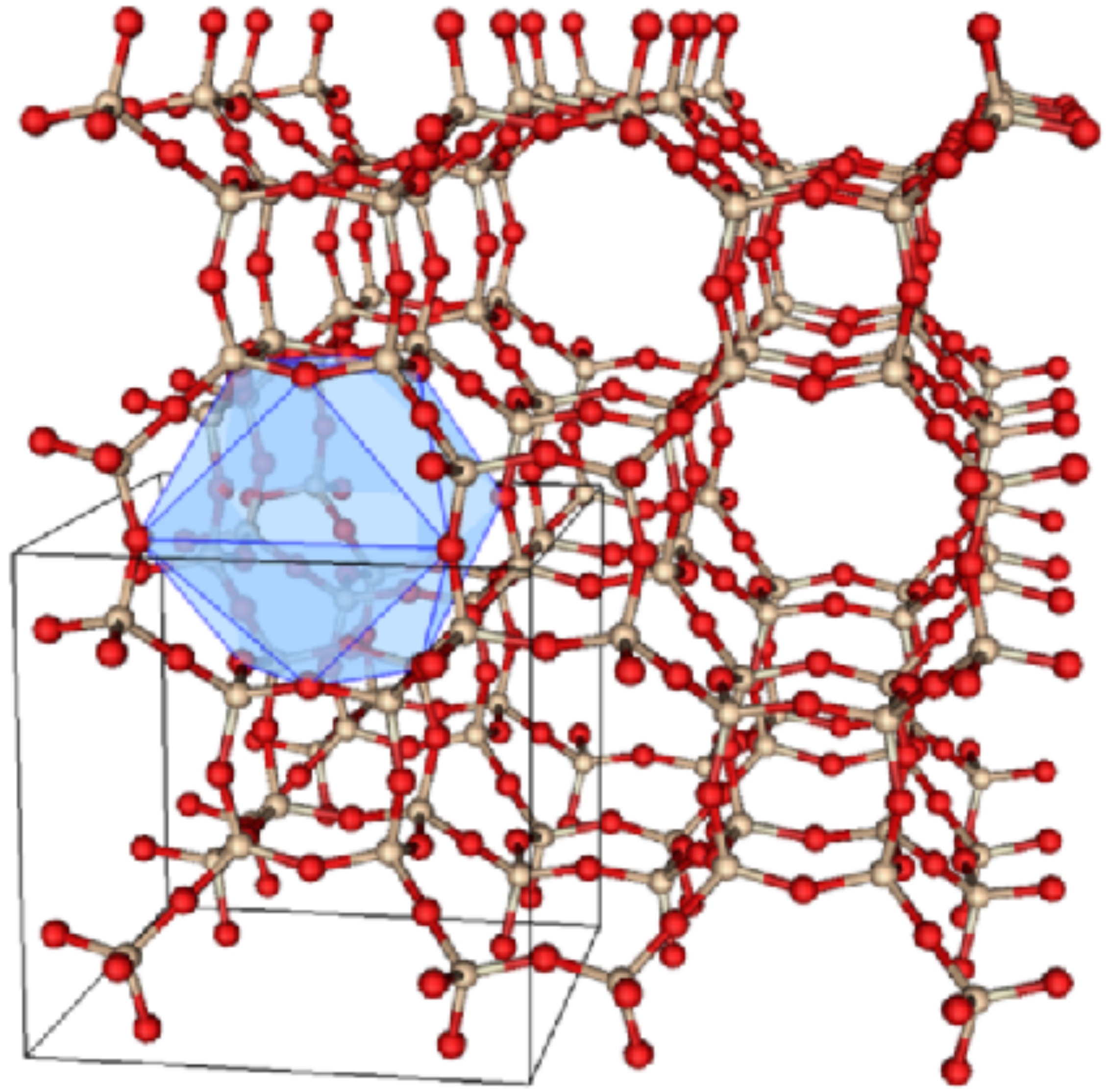


cosmology

Wilding et al 2021

Topology in Data





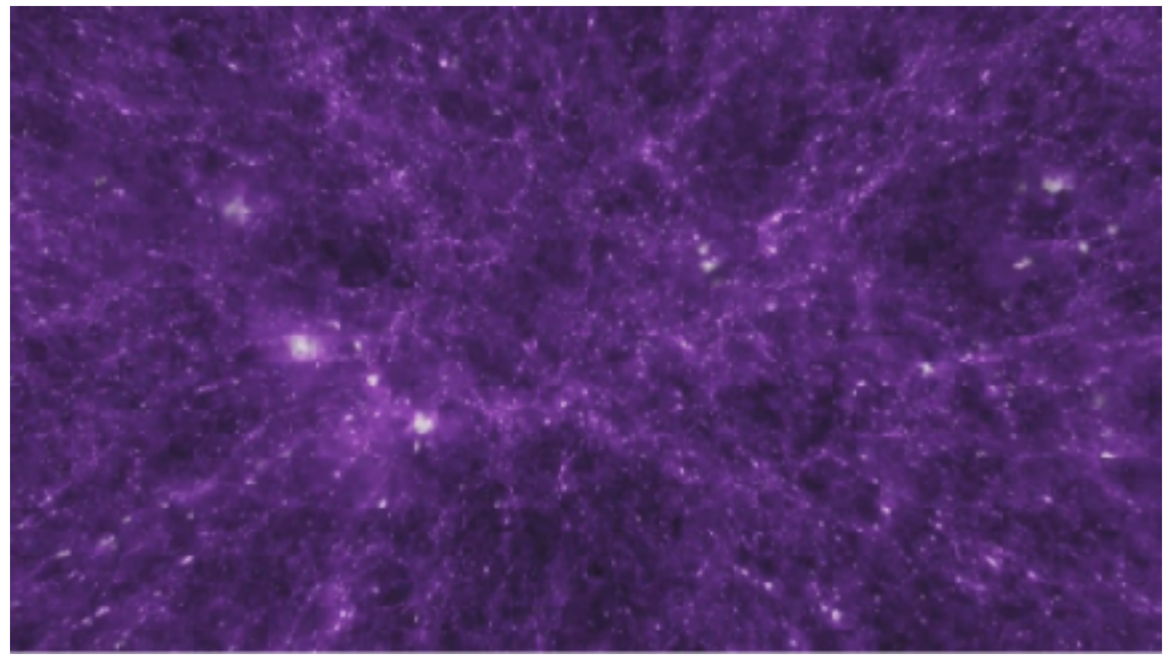
cosmology

material science

Wilding et al 2021

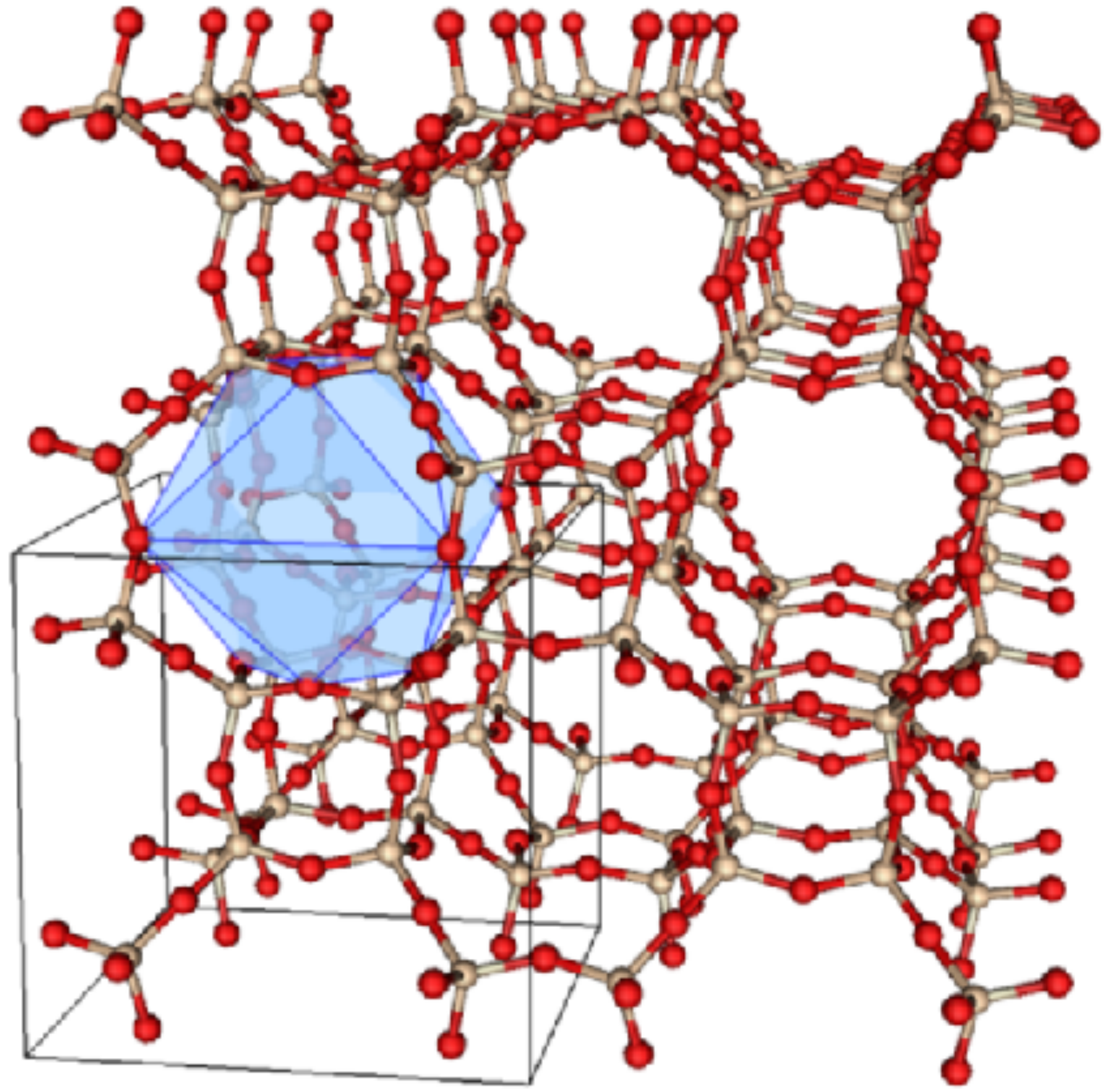
Krishnapriyan et al, 2020

Topology in Data



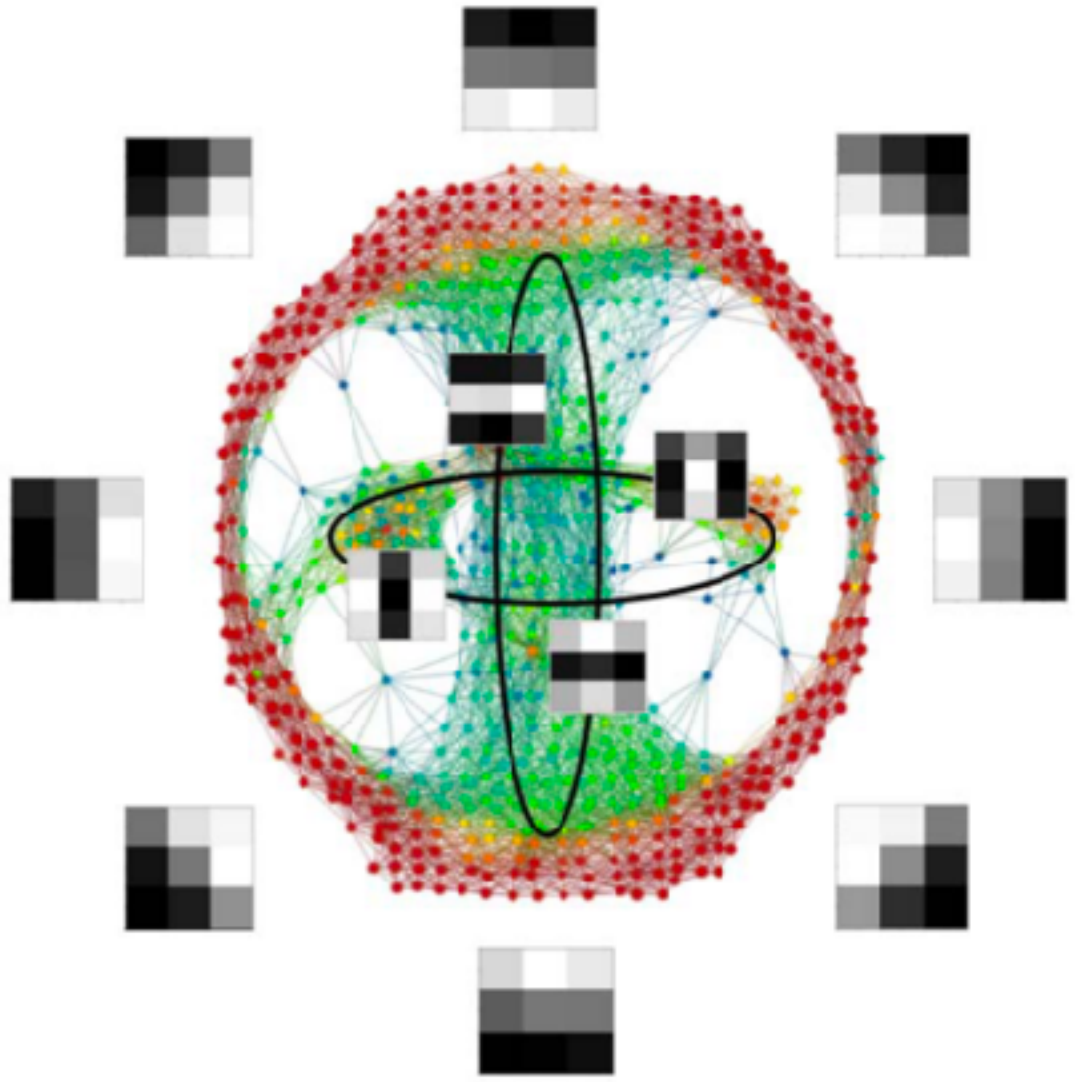
cosmology

Wilding et al 2021



material science

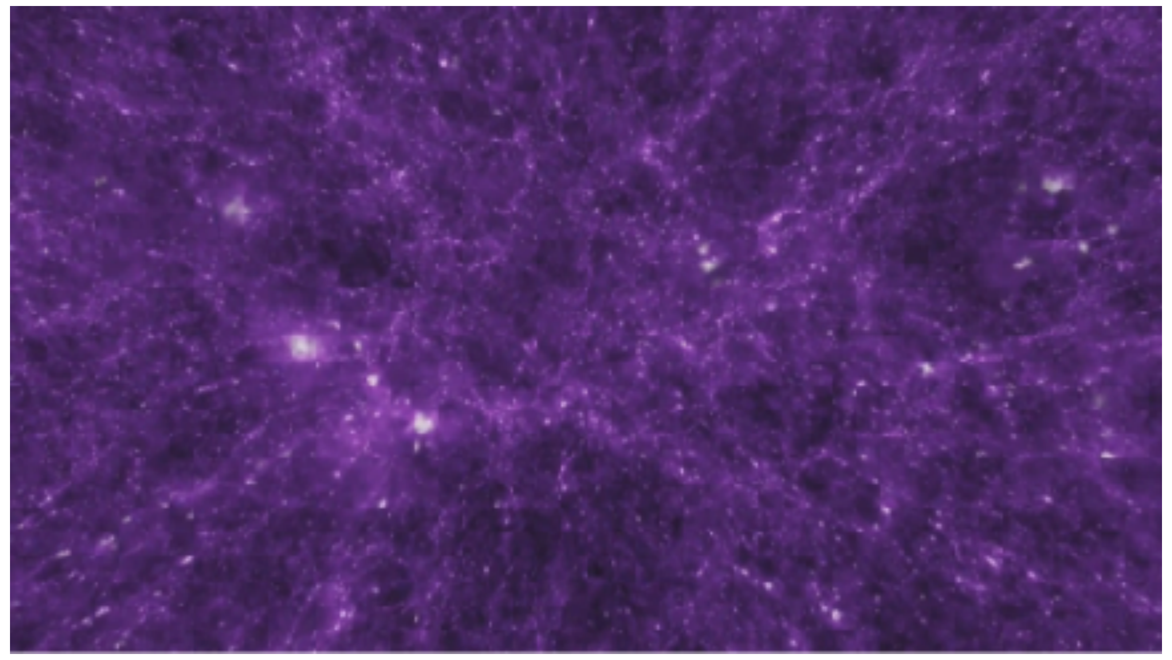
Krishnapriyan et al, 2020

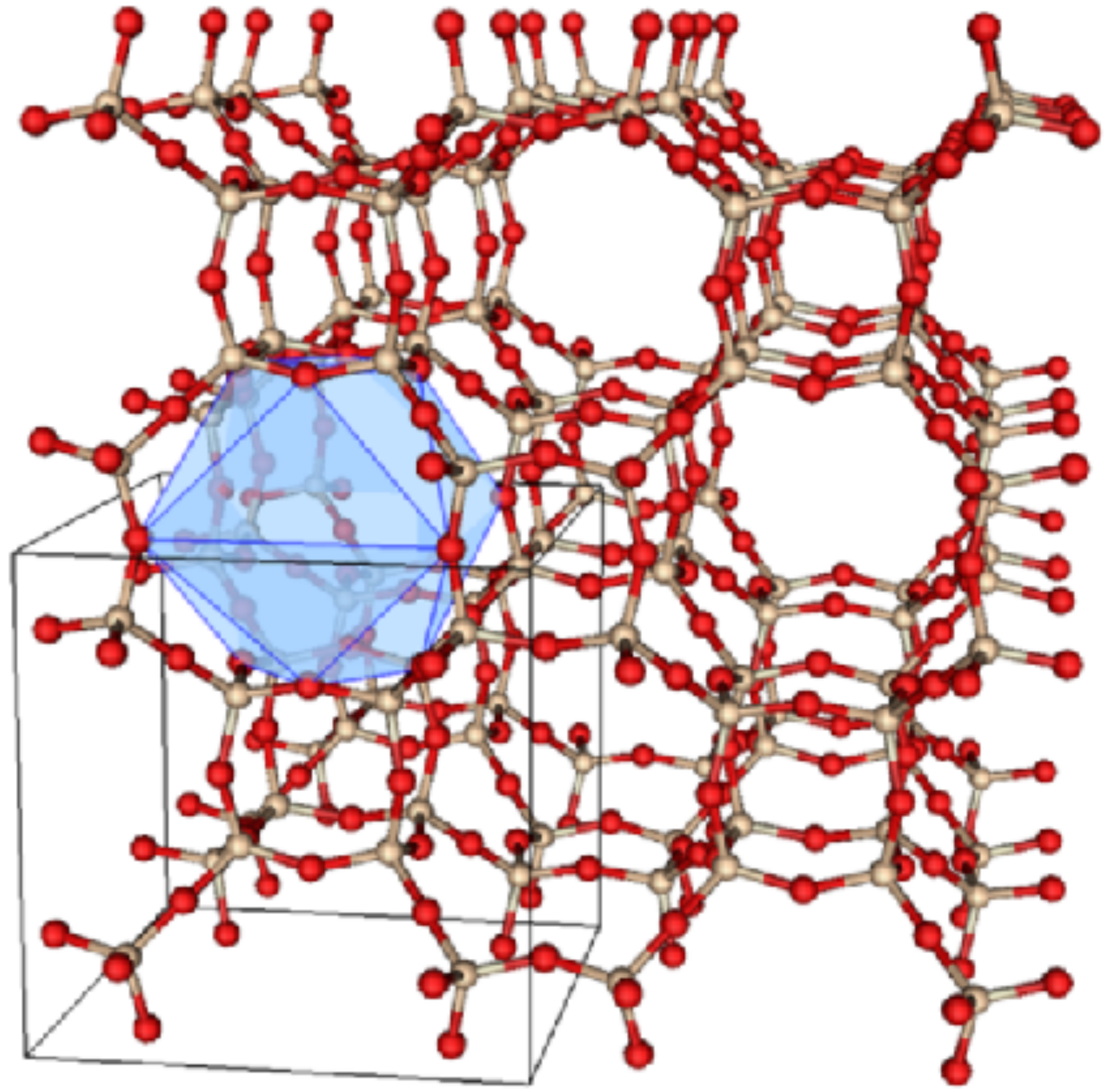


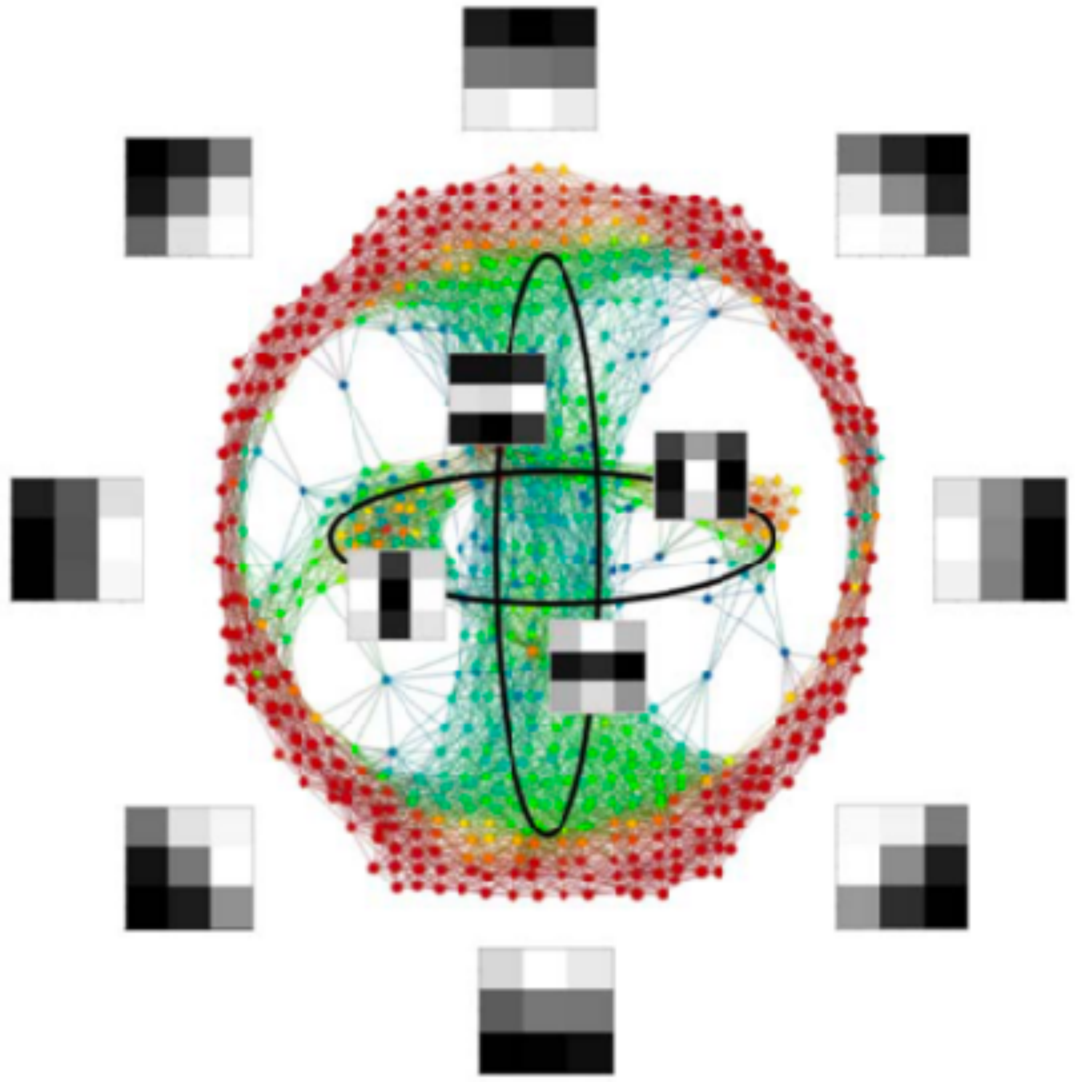
neural network

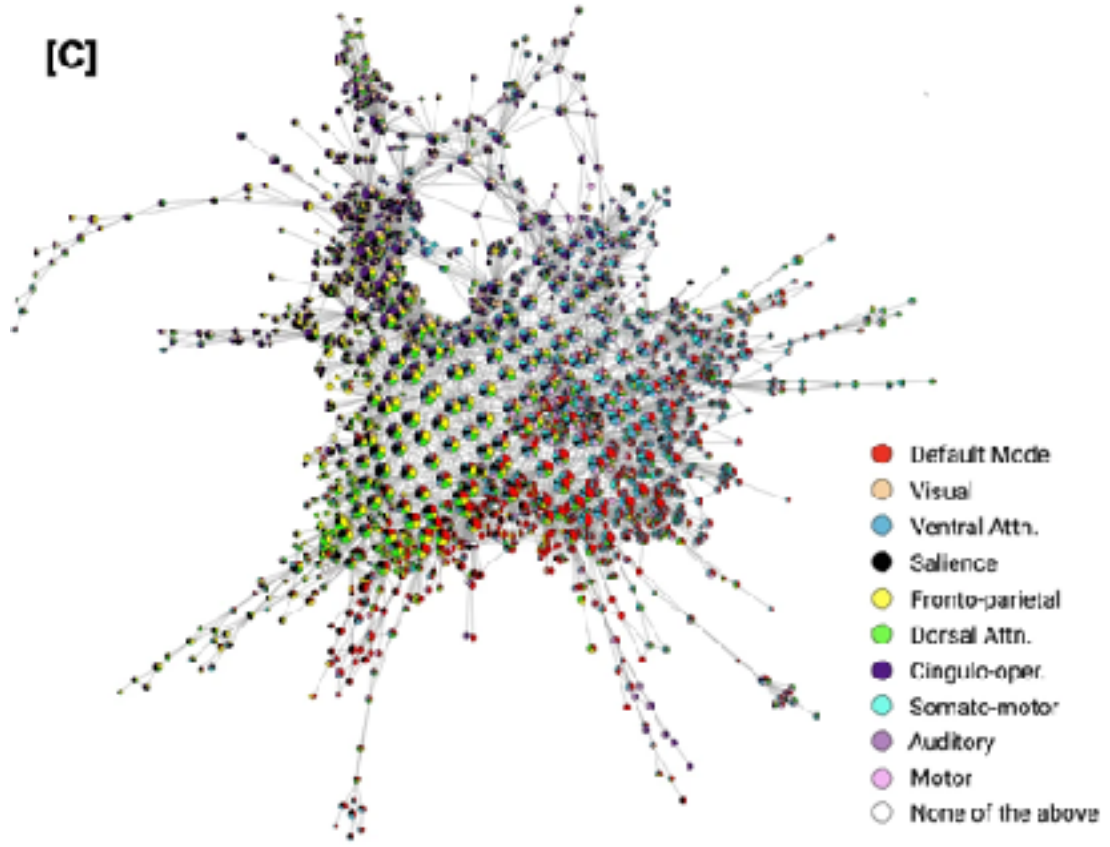
Gabrielsson and Carlsson, 2019

Topology in Data









cosmology

Wilding et al, 2021

material science

Krishnapriyan et al, 2020

neural network

Gabrielsson and Carlsson, 2019

neuroscience

Saggar et al, 2022

Interlude

Mathematics of Topological Data Analysis

Estimator?
Mathematical Algorithm?

Yes!

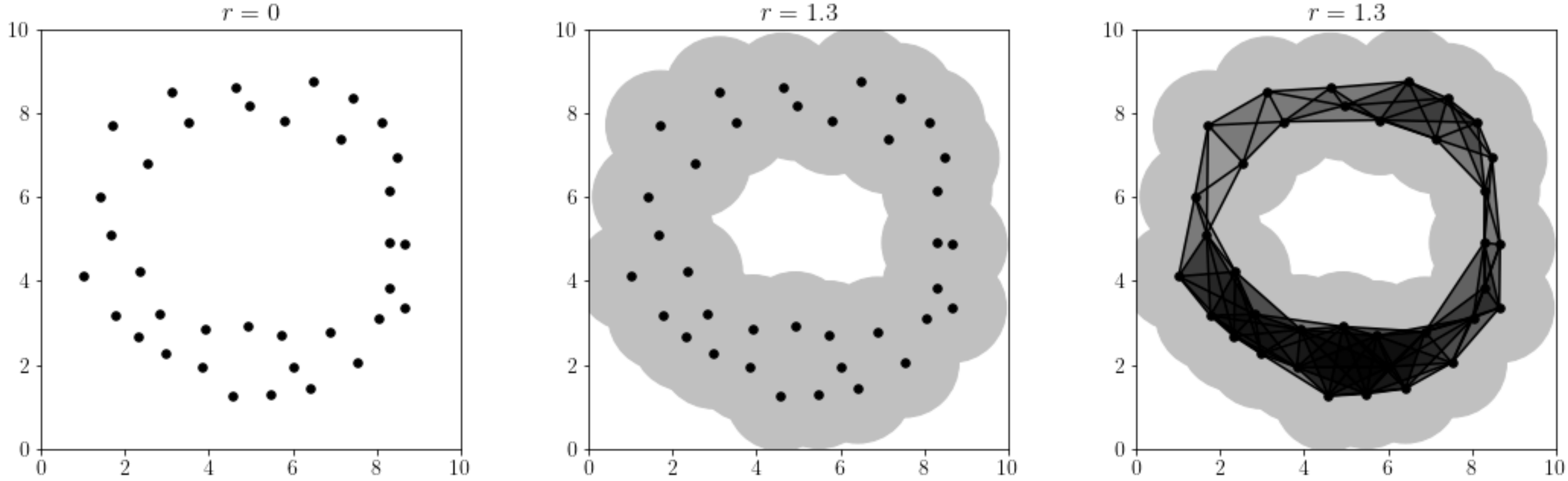


diagram credit: Andrey Yao

Pitfall

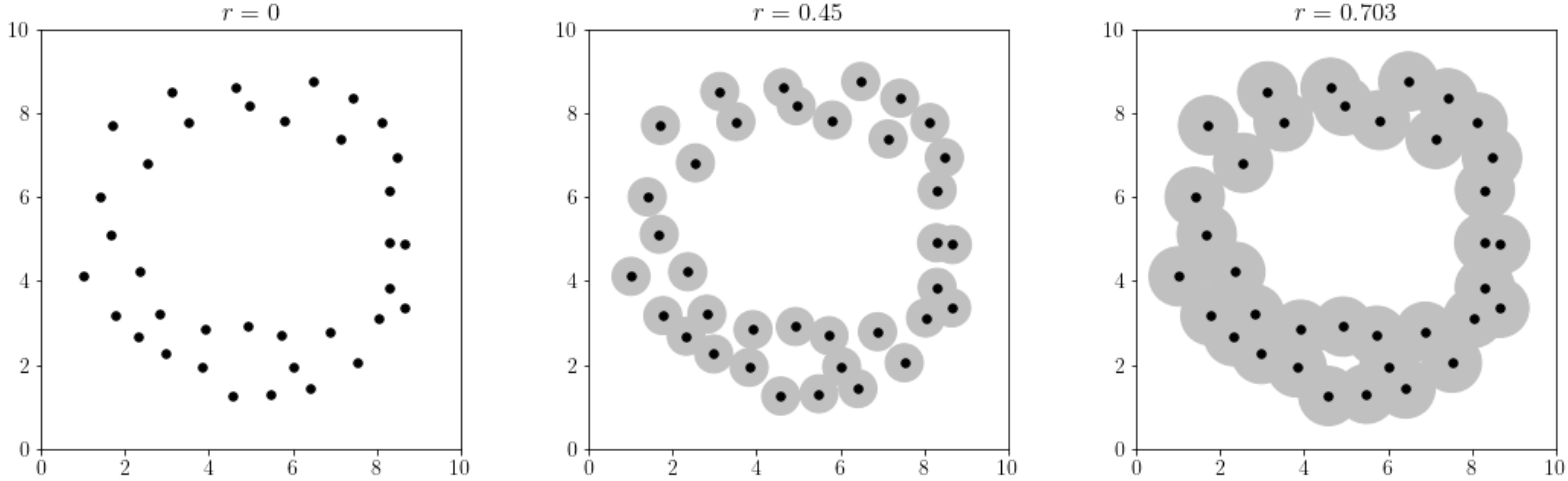


diagram credit: Andrey Yao

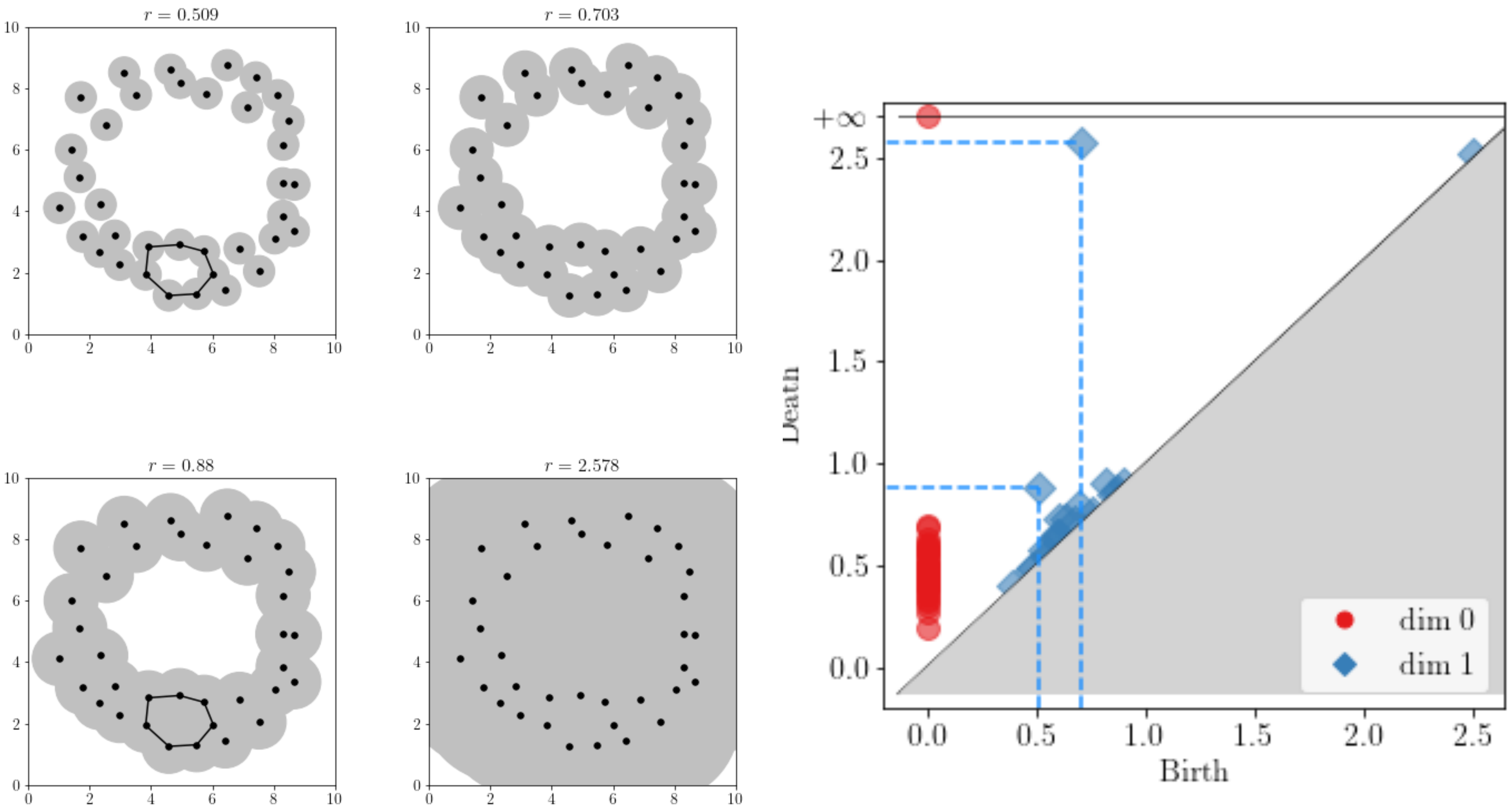
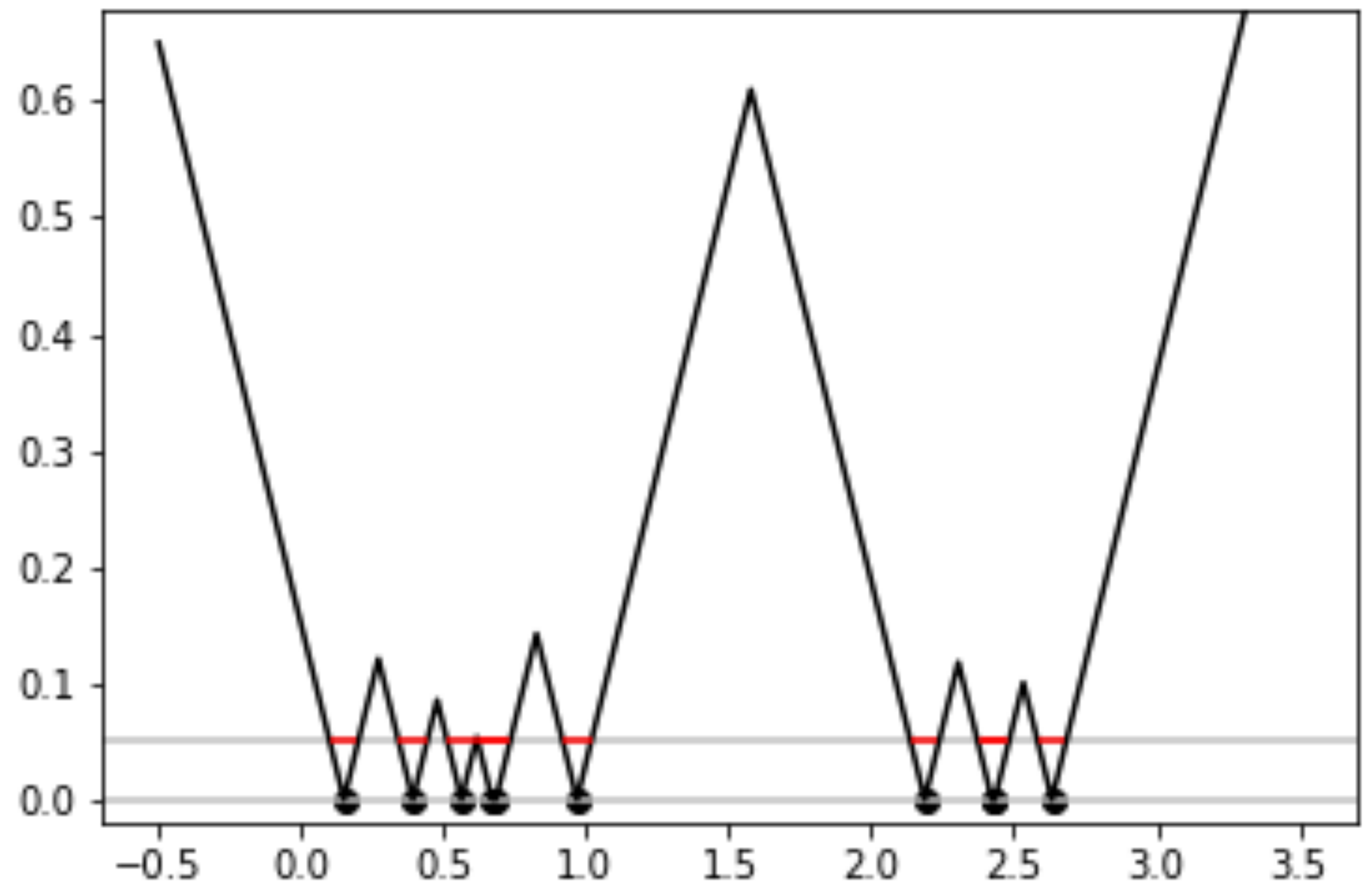


diagram credit: Andrey Yao

The Ground Truth Persistence Diagram?

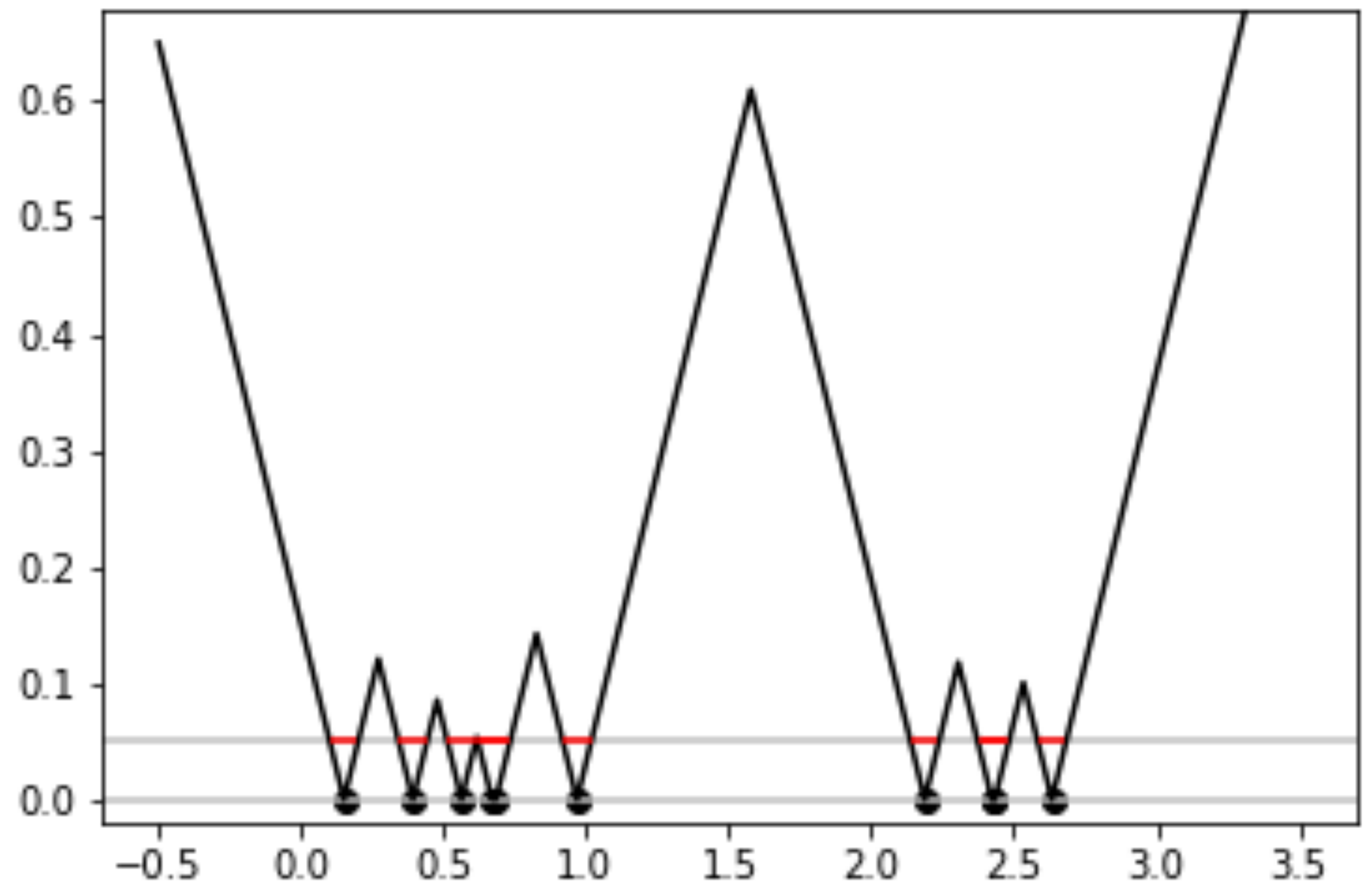
Higher-Dimensional Perspective

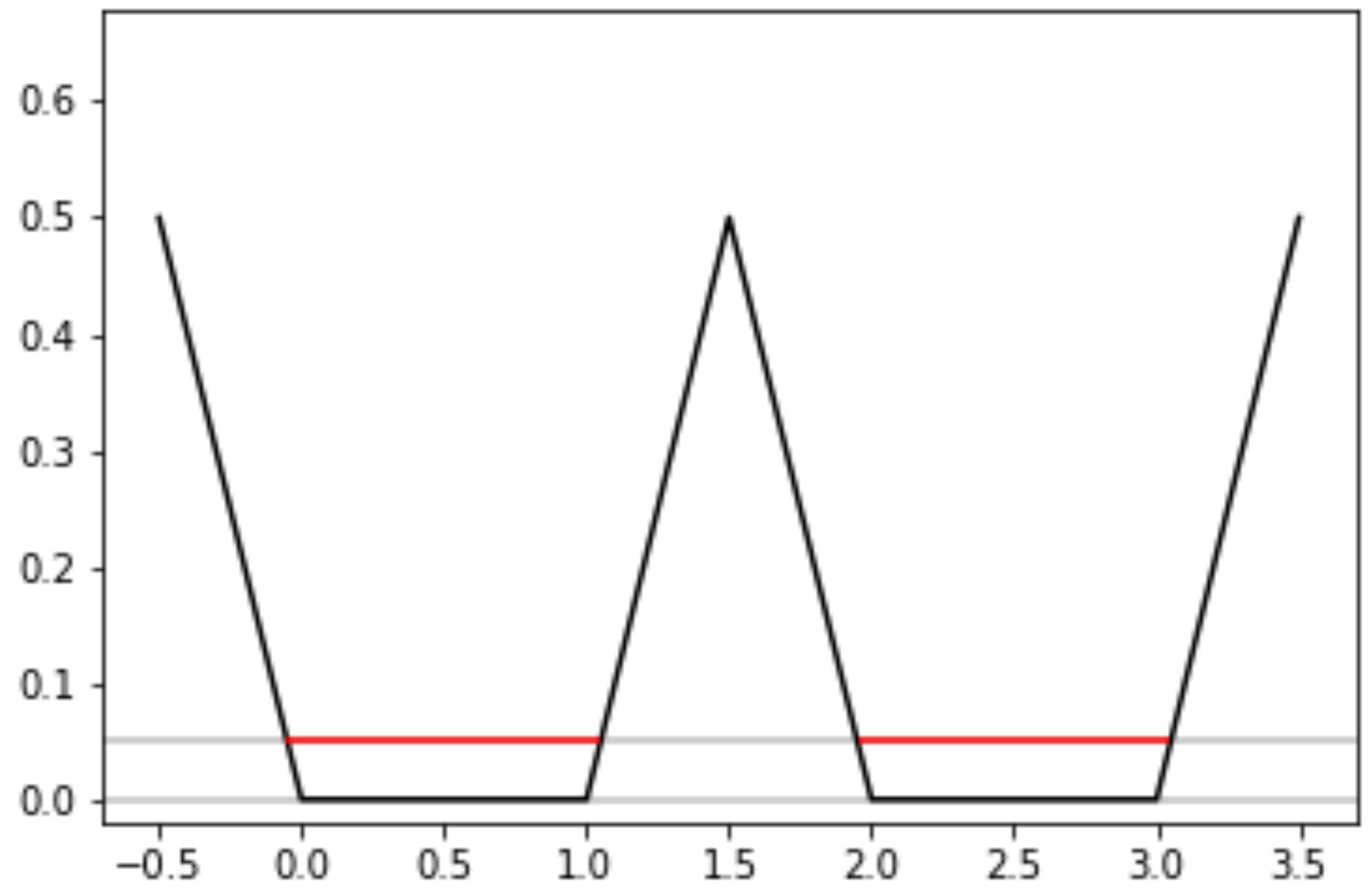
- Balls are lower-level sets of distance function
- $d_{emp}(x) = \min d(x, X_i)$



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

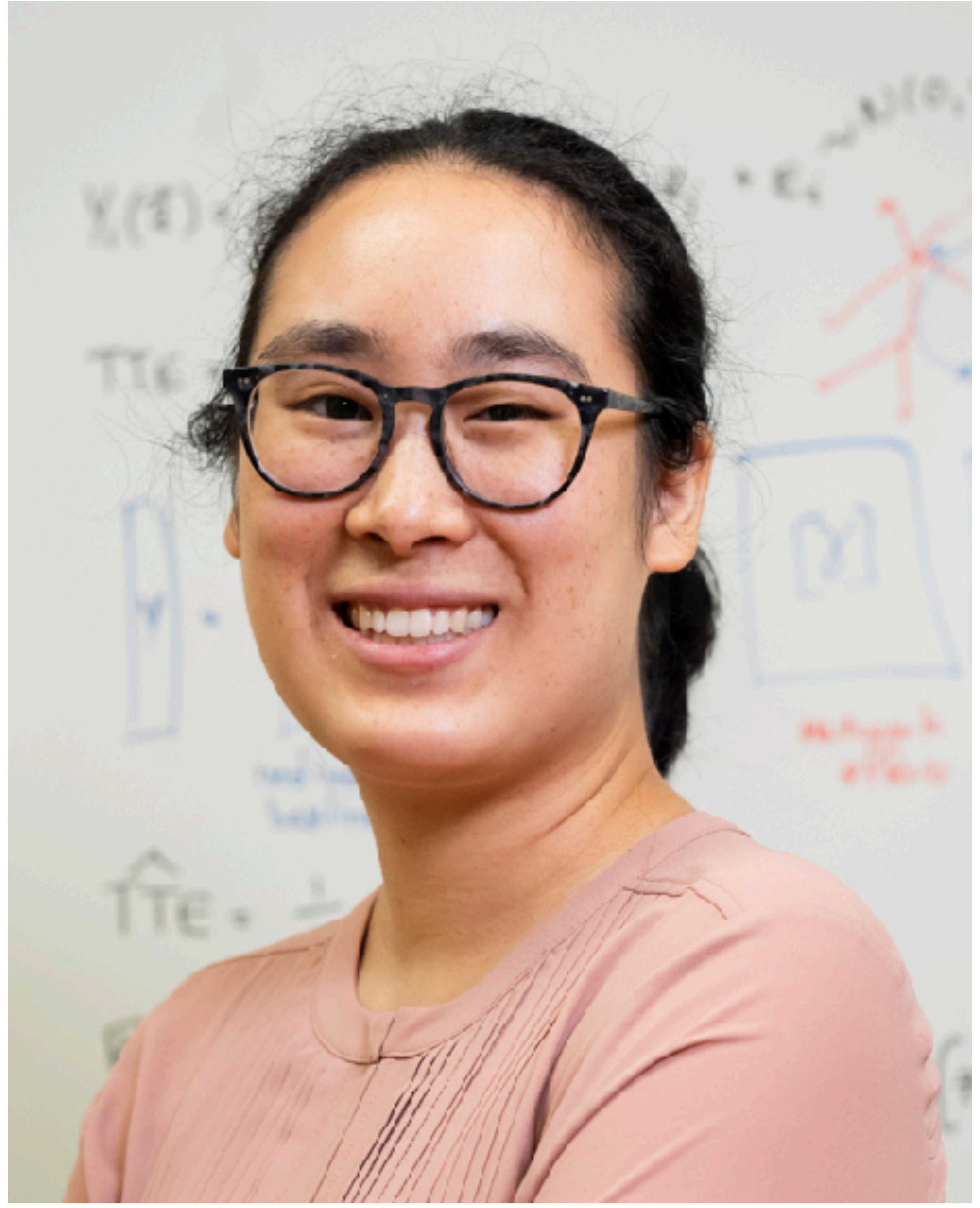
- useful when the dataset has global structures like loops and holes
- these structures can be estimated
- their information can be summarized in persistence diagrams

Act II

Weak Topological Signals Amidst Noise

My Lovely Collaborators

Gennady Samorodnitsky

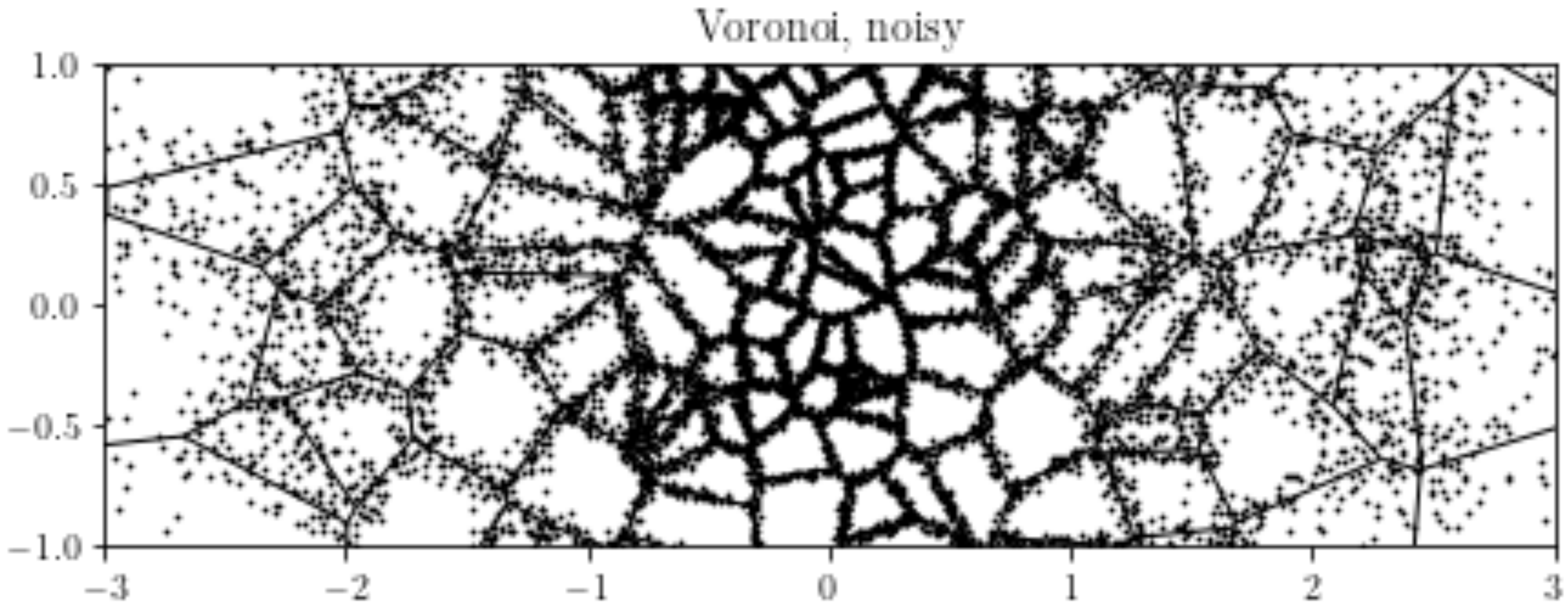


Christina Lee Yu

Andrey Yao

Two problems

- Size
- Noise



Two Problems

- Size
- Noise

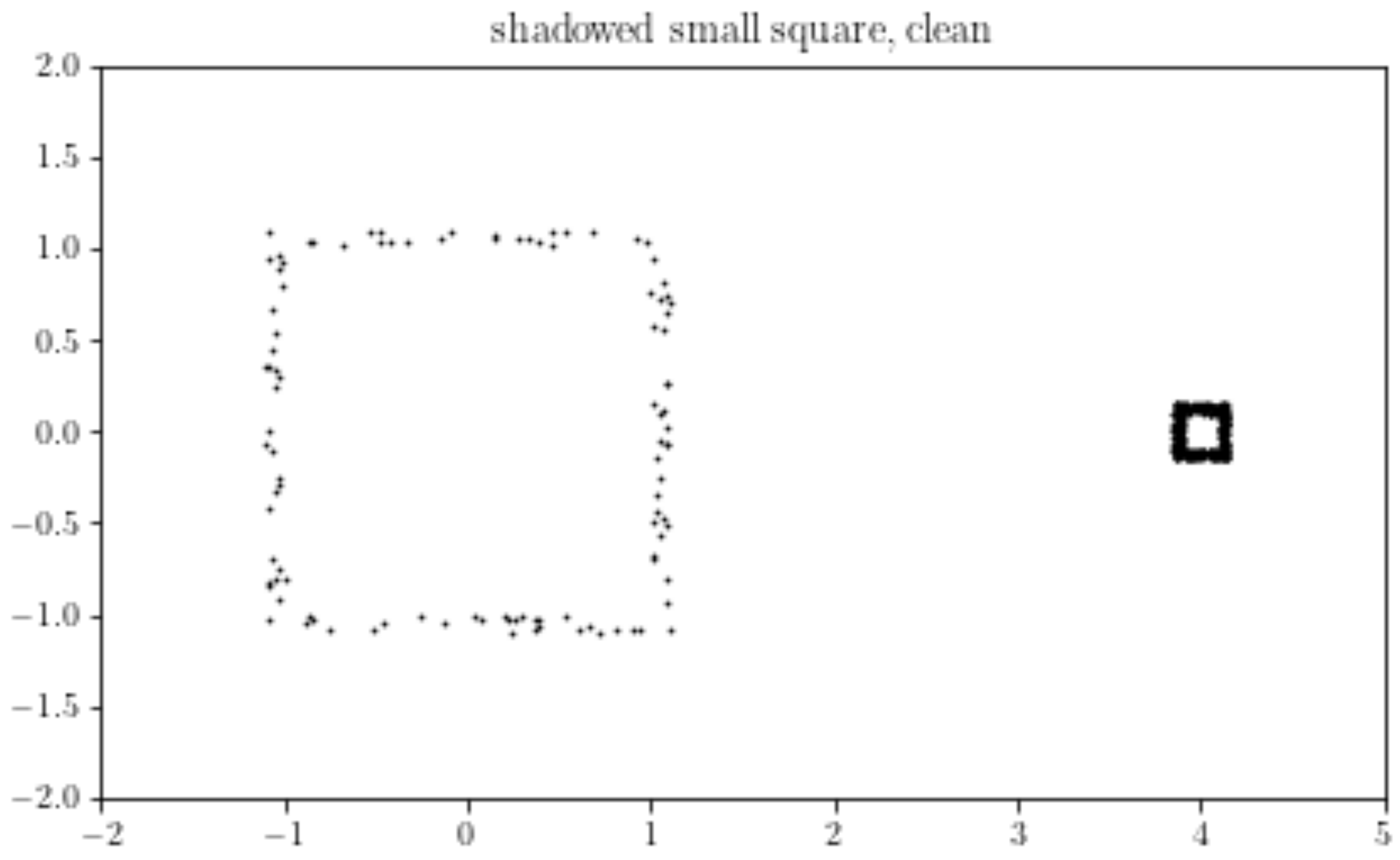
- Related works
  - Hickok (2022)
  - Berry and Sauer (2019)
  - Moon et al (2018)
  - Carlsson and Zomorodian (2009)
  - etc…

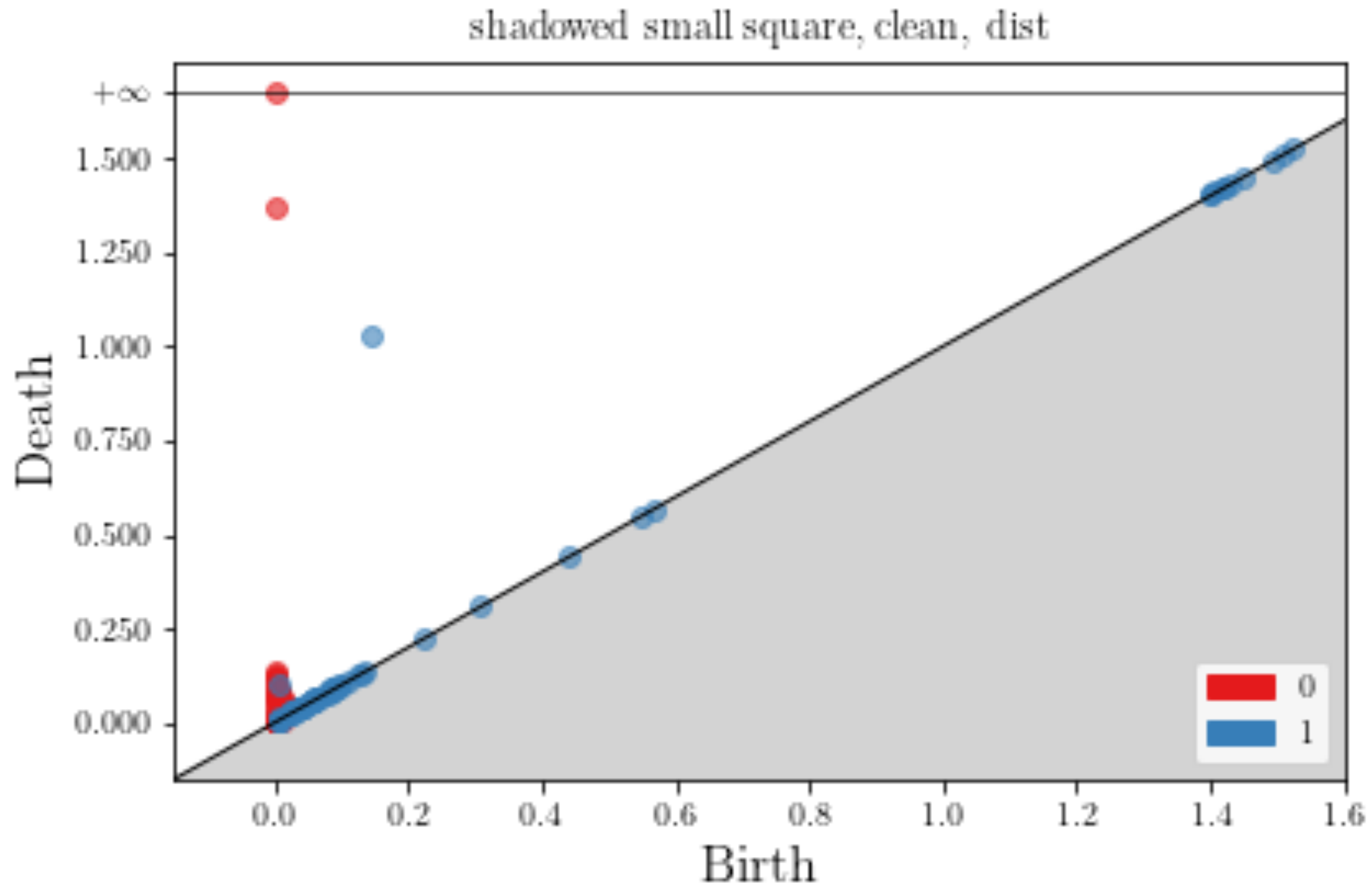
One solution

- Size
- Noise

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

Size


shadowed small square, clean


shadowed small square, clean, dist
Death
Birth
0
1

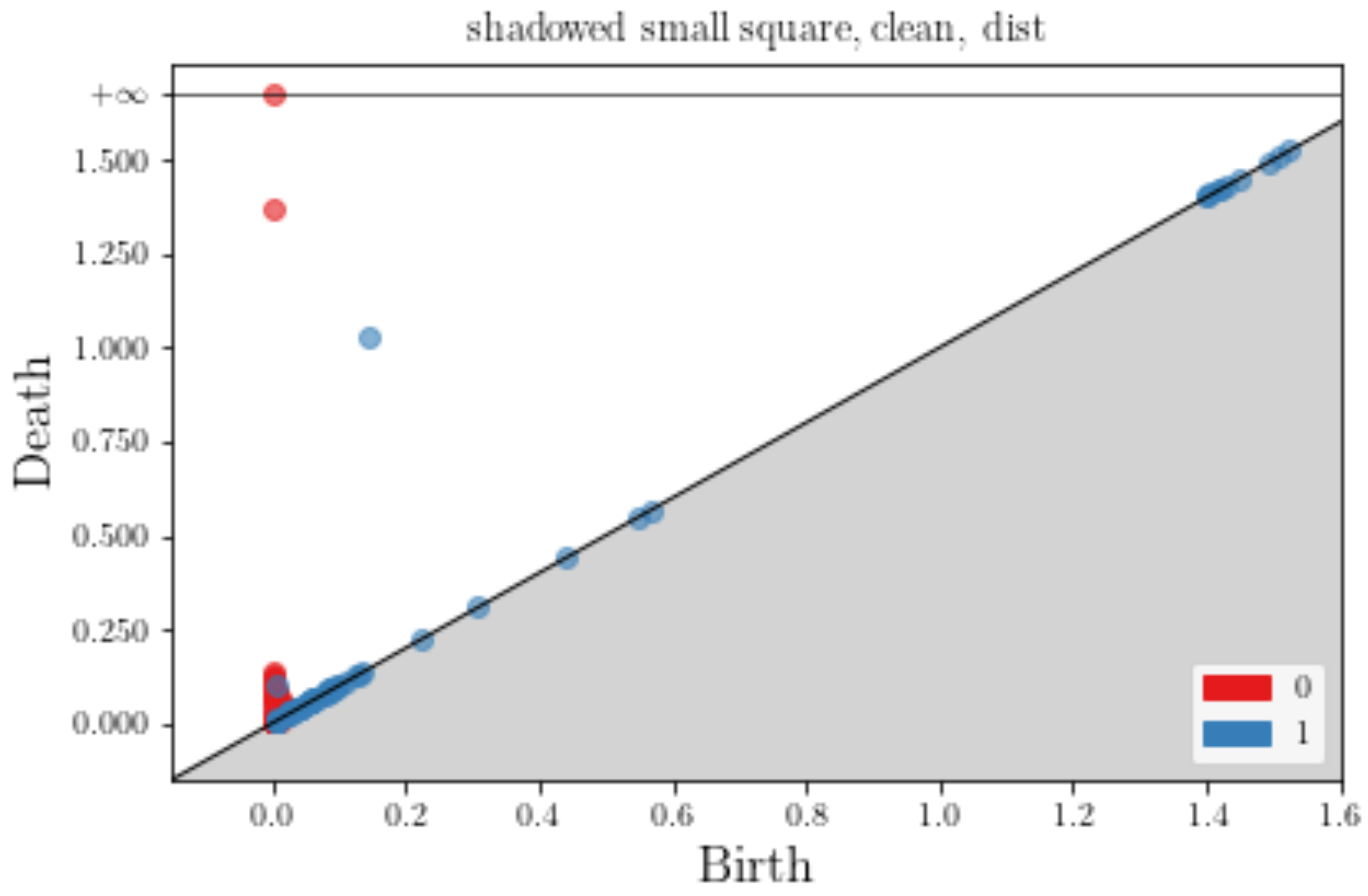
Grow Balls Sloooooooooooooooooooowly

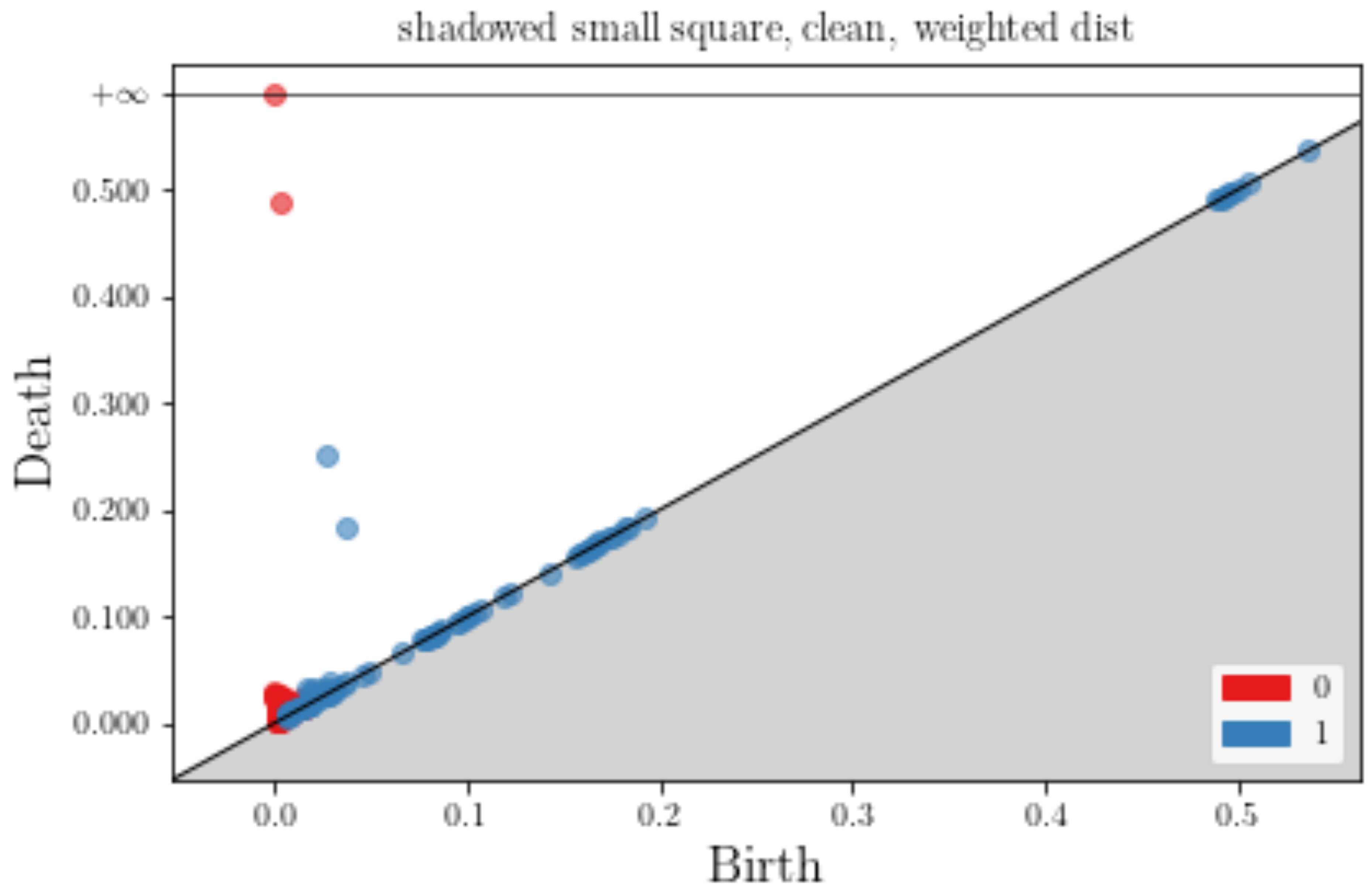
on the smaller square

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

Grow Balls Sloooooooooooooooooooowly
on the smaller square

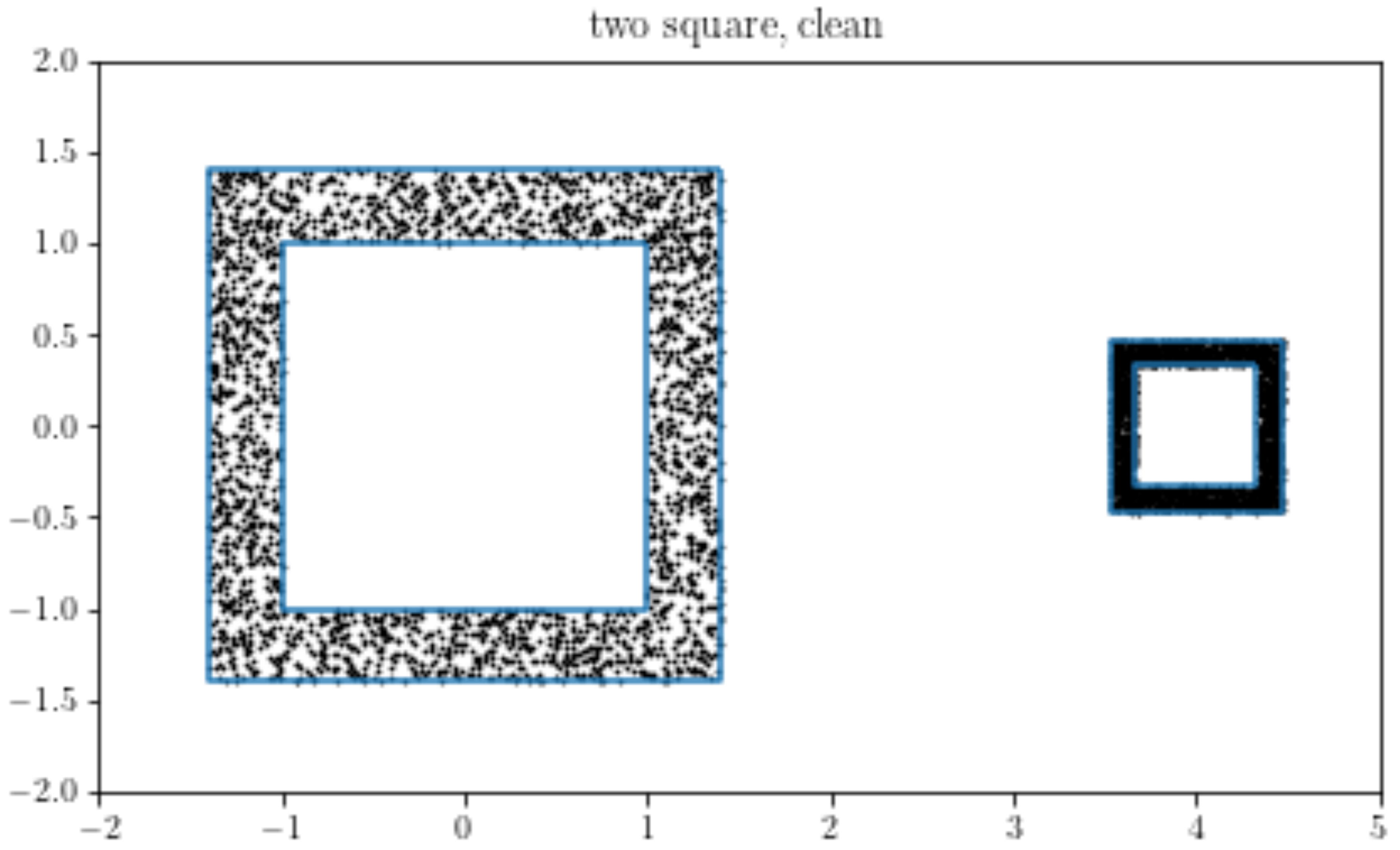
- rate = 1/density$^{1/D}$

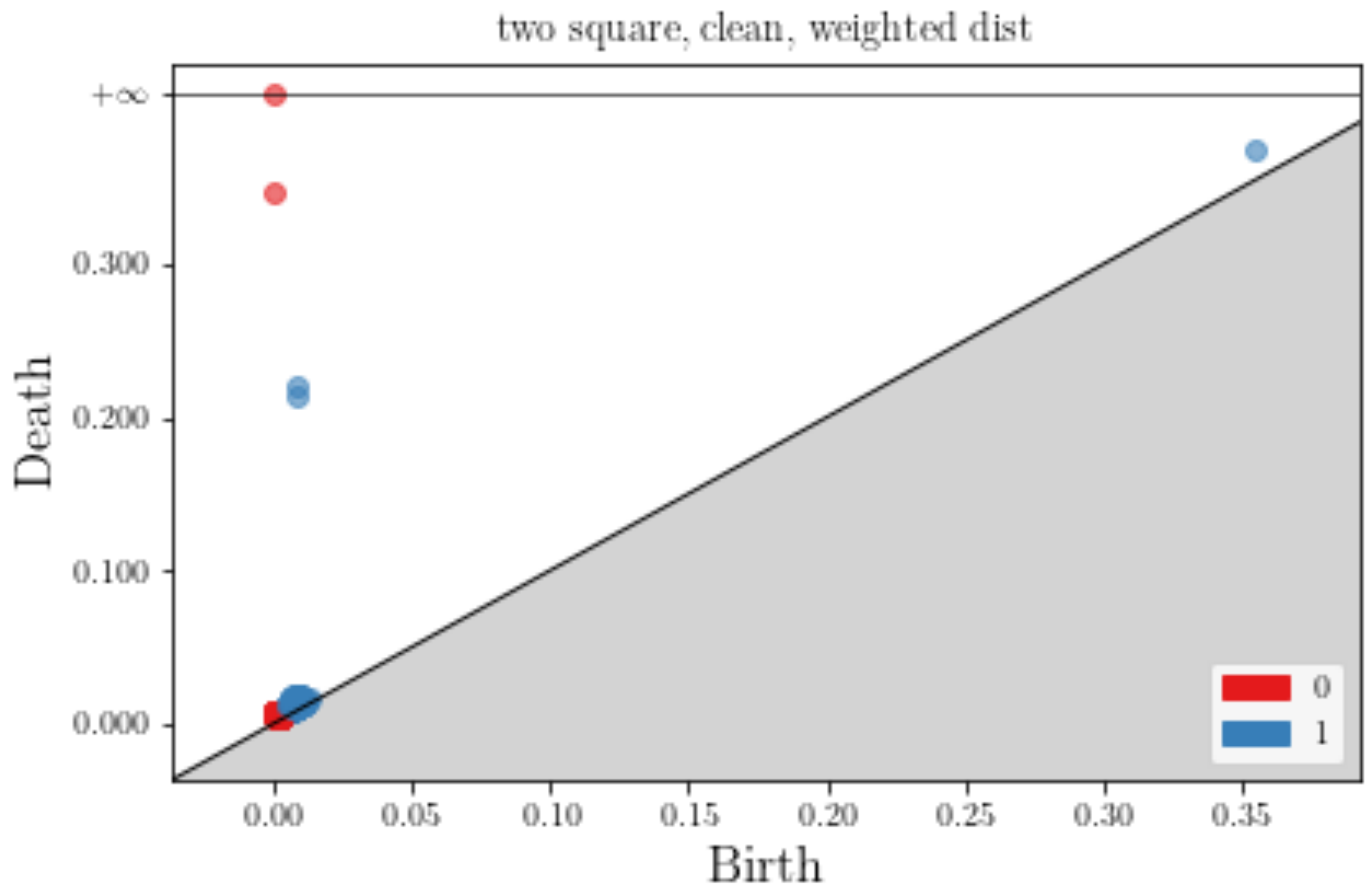




Scale invariance

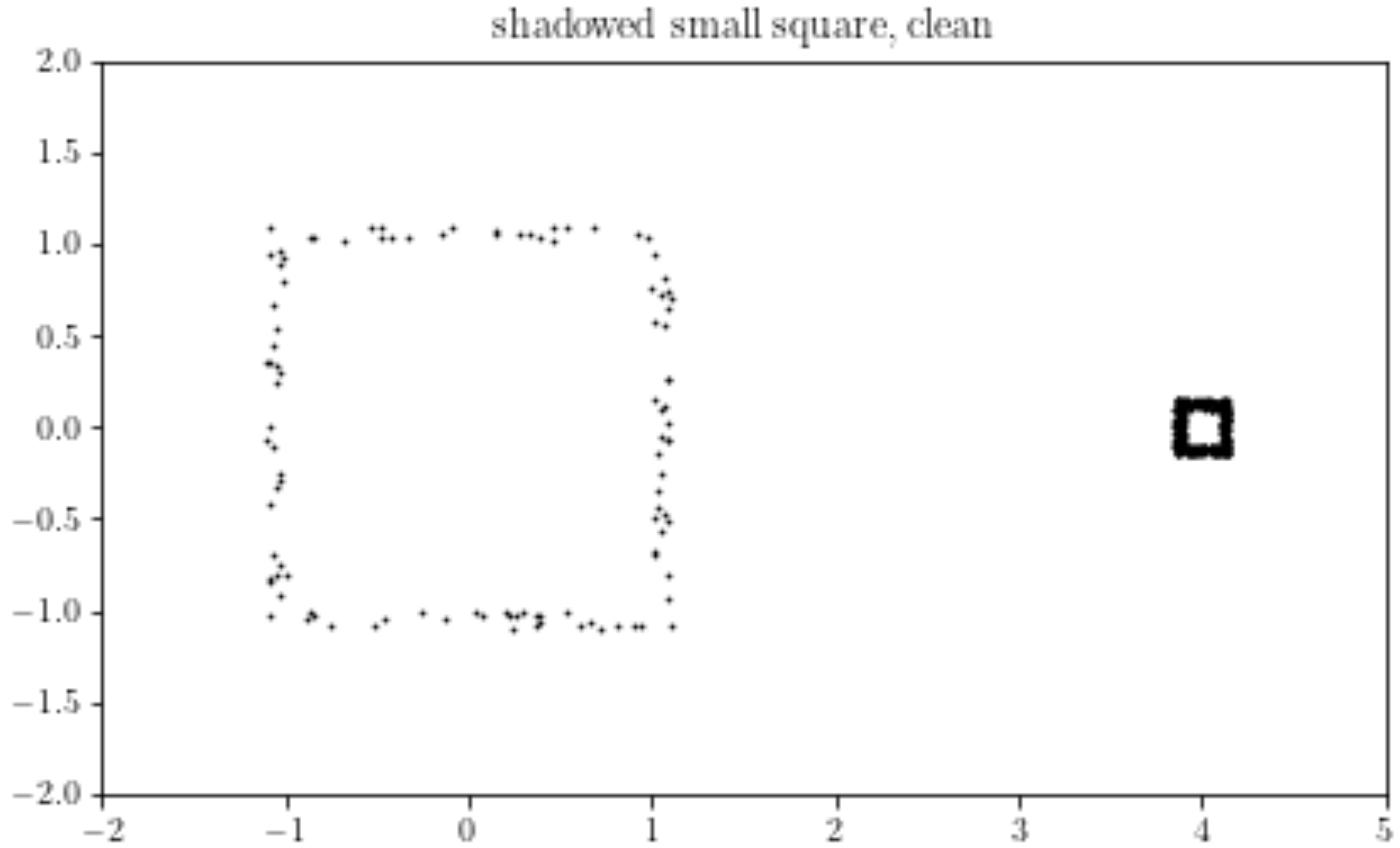
- uniform scaling —> same persistence diagrams





Theorem

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

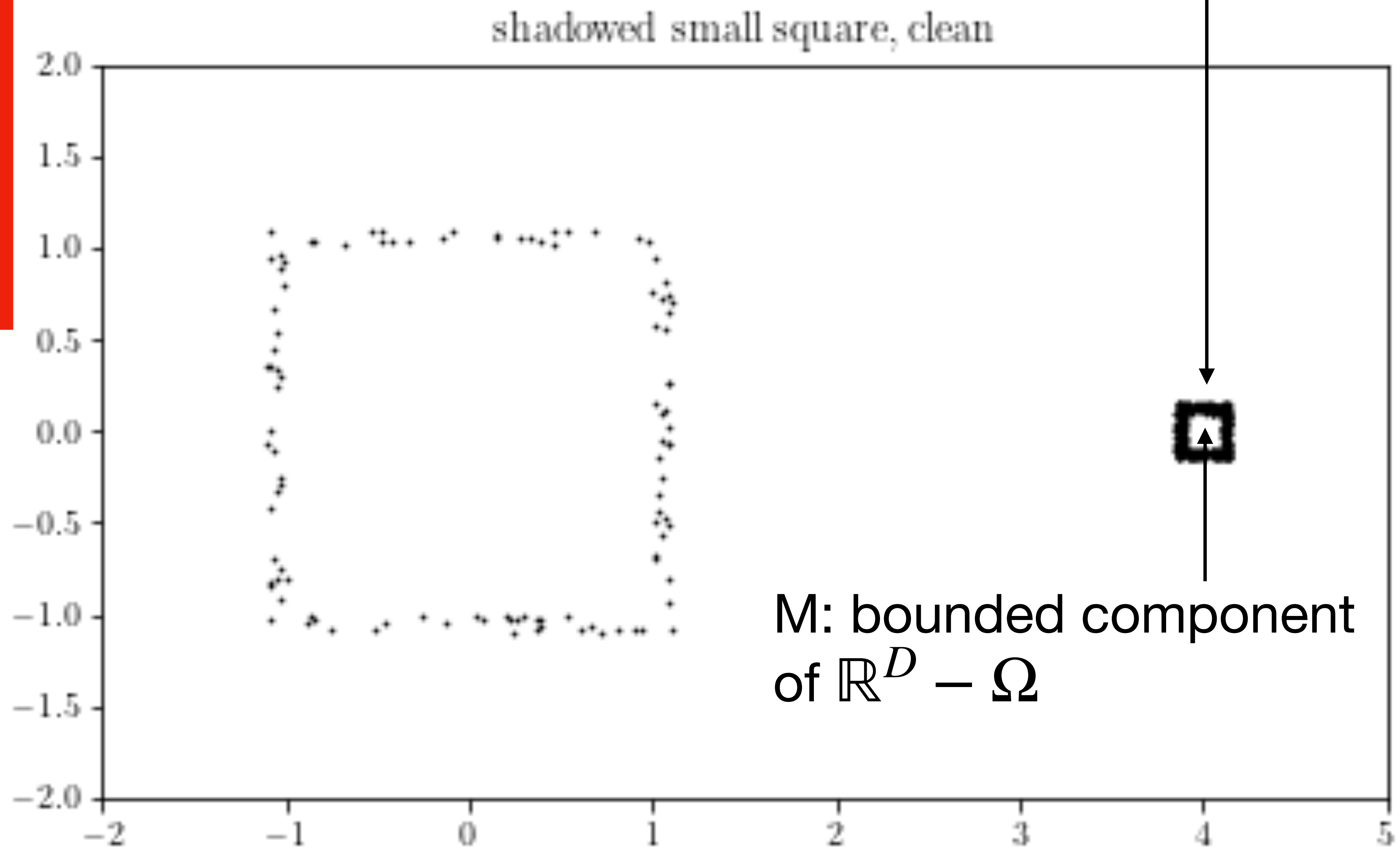


Theorem

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 $r = \max_{x \in M} d(x, \partial M)$.

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



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

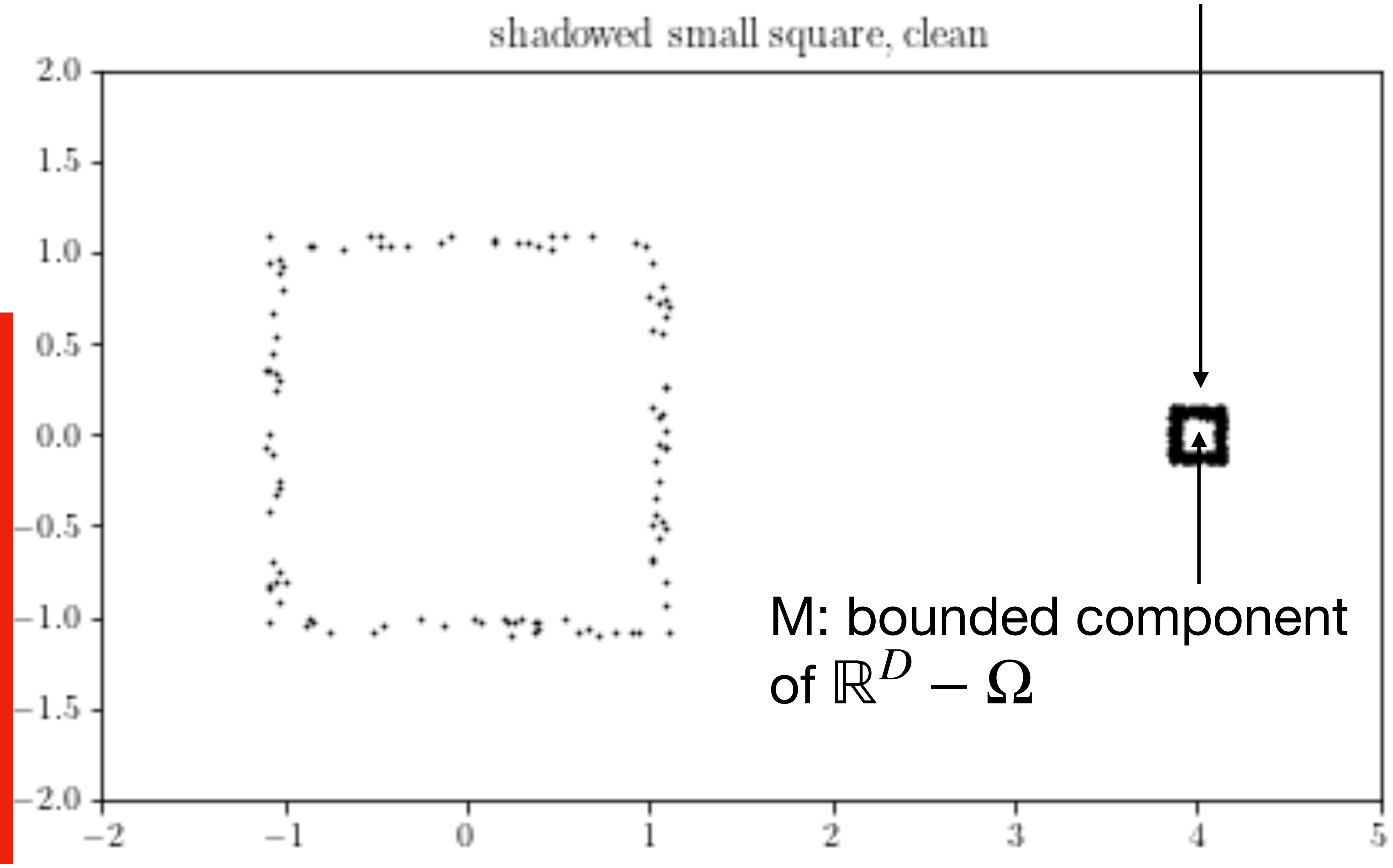
Theorem

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

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

- 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

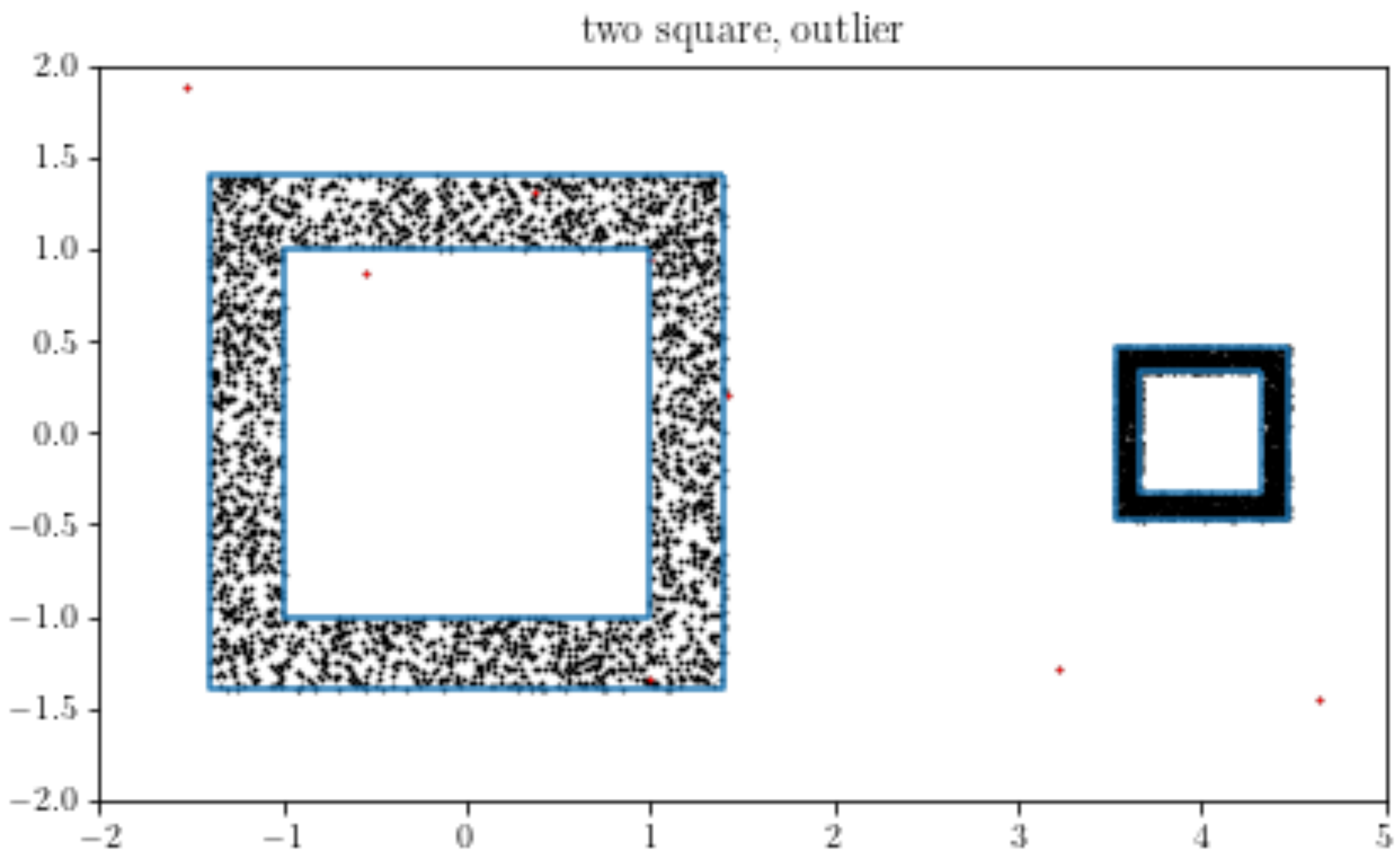
  with persistence at least $\frac{1}{\sqrt{2}} t^{1/D} r - O(m^{1/D})$



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

Noise

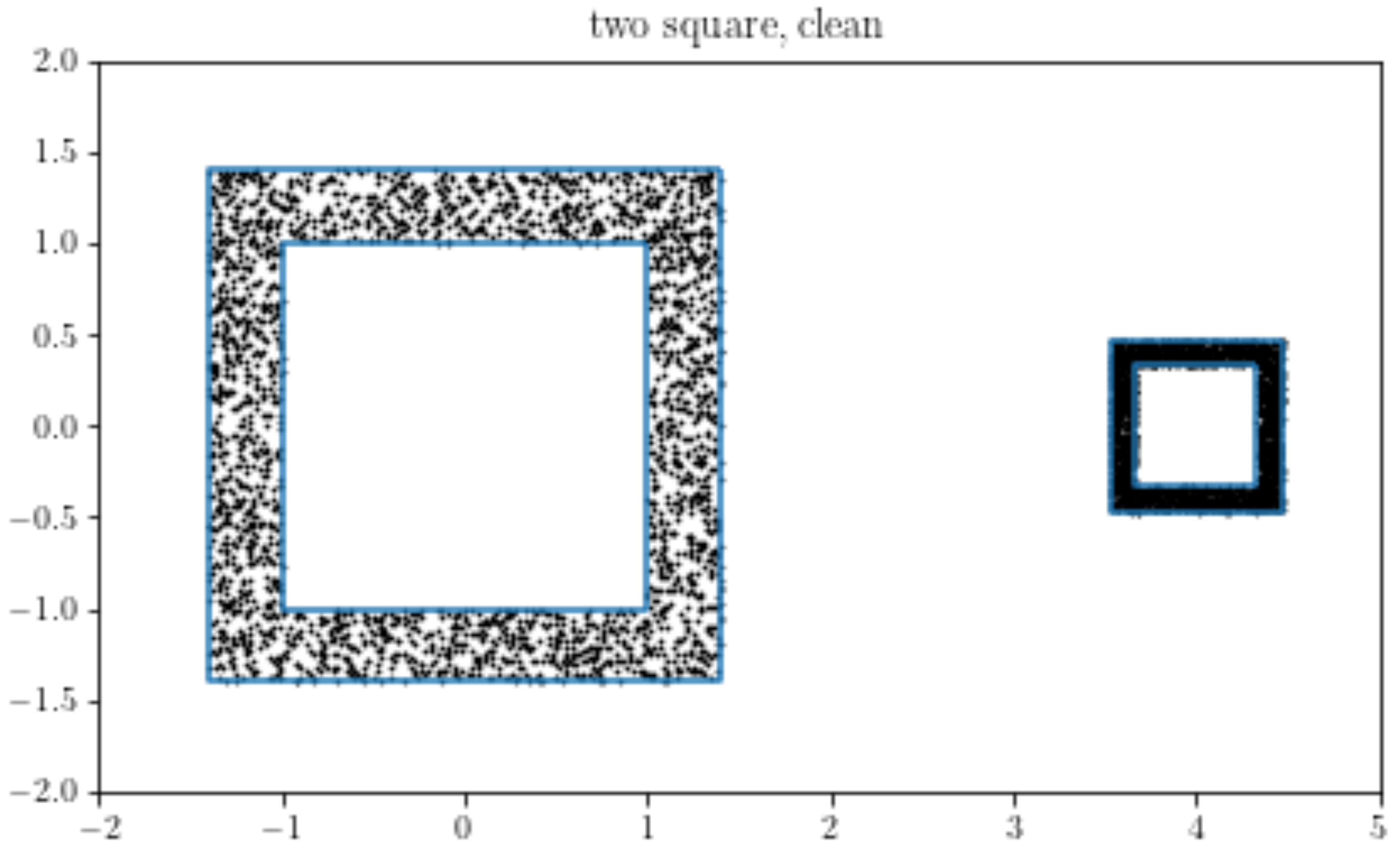
Outliers

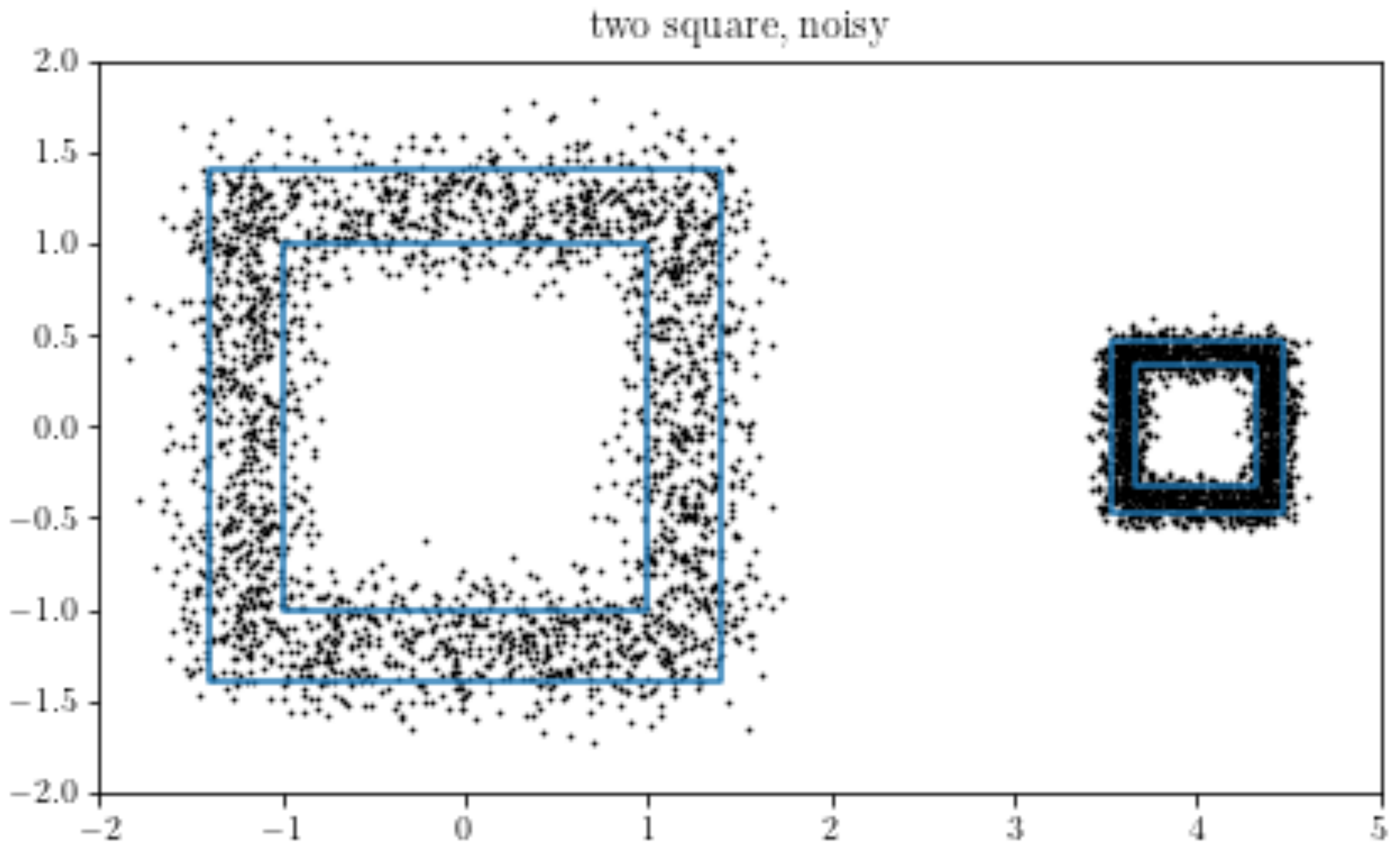

two square, outlier
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3
4
5


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

Additive Noise

- $d_{emp}(x) = \min d(x, X_i)$
- $d(x) = \inf d(x, y)$, where y ranges over the support





Distance-to-measure

Known Problem, Known Solution [Chazal et al (2011), Chazal et al (2018)]

Distance-to-measure

Known Problem, Known Solution [Chazal et al (2011), Chazal et al (2018)]

- d(x) = inf d(x, y) = 0-th quantile of d(x, .)

Distance-to-measure

Known Problem, Known Solution [Chazal et al (2011), Chazal et al (2018)]

- d(x) = inf d(x, y) = 0-th quantile of d(x, .)
- DTM(x) = average of the first m-th quantiles of d(x, .)

Distance-to-measure

Known Problem, Known Solution [Chazal et al (2011), Chazal et al (2018)]

- d(x) = inf d(x, y) = 0-th quantile of d(x, .)
- DTM(x) = average of the first m-th quantiles of d(x, .)
- can leverage empirical process theory

Robust Density-Aware Distance (RDAD)

Robust Density-Aware Distance function

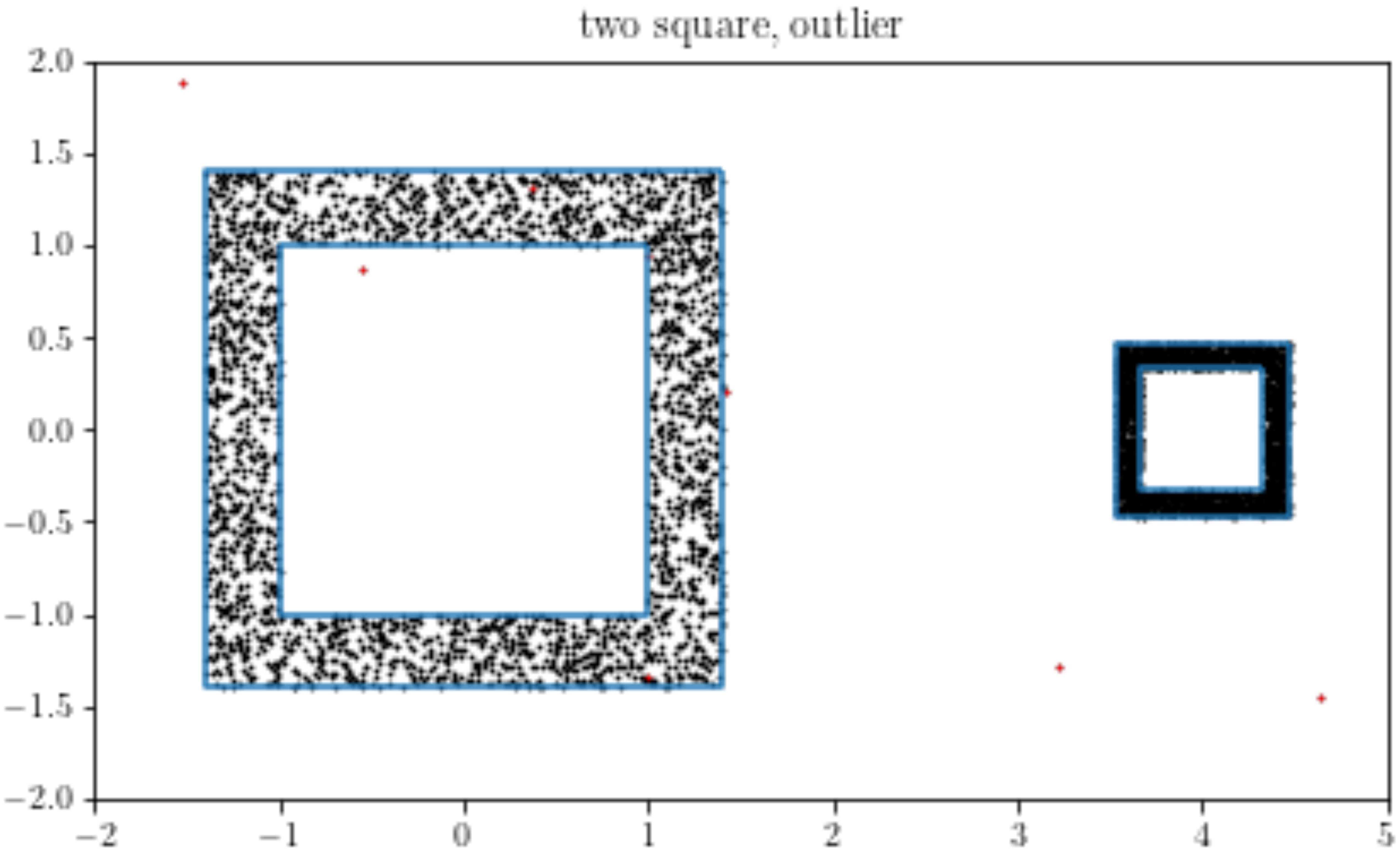
$$DTM(x) = \sqrt{\frac{1}{m}\int_0^m G_x^{-1}(q)^2 dq}$$

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

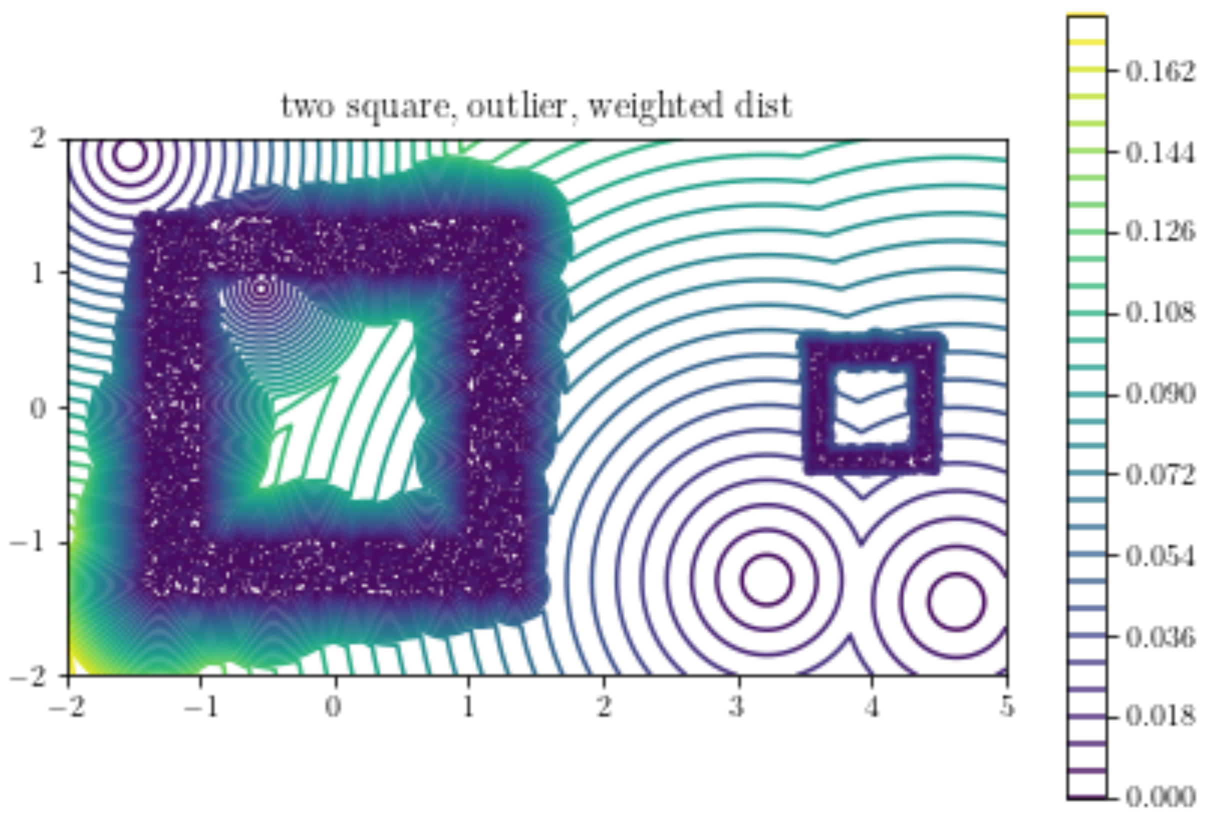
$$RDAD(x) = \sqrt{\frac{1}{m}\int_0^m F_x^{-1}(q)^2 dq}$$

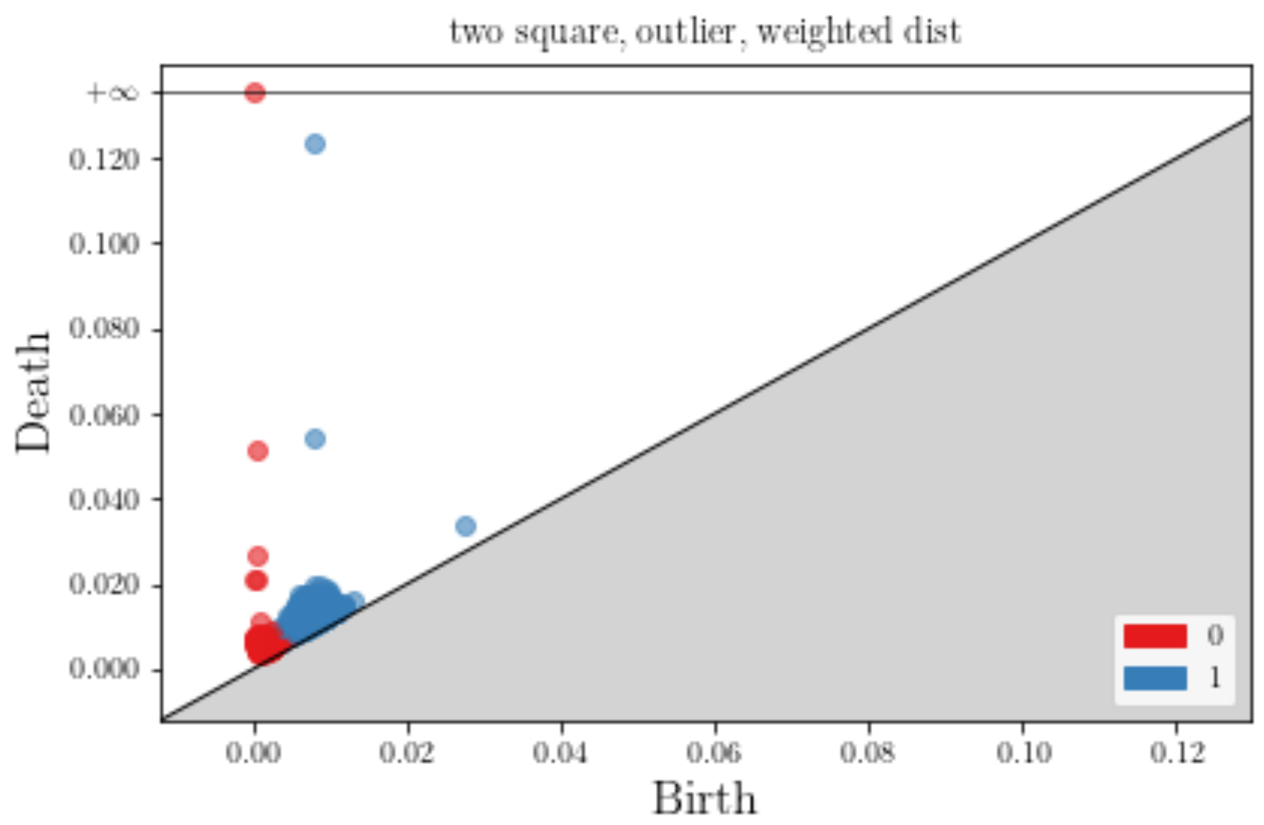
$$F_x(r) = P\{d(x, X) f(X)^{1/D} \leq r\}$$

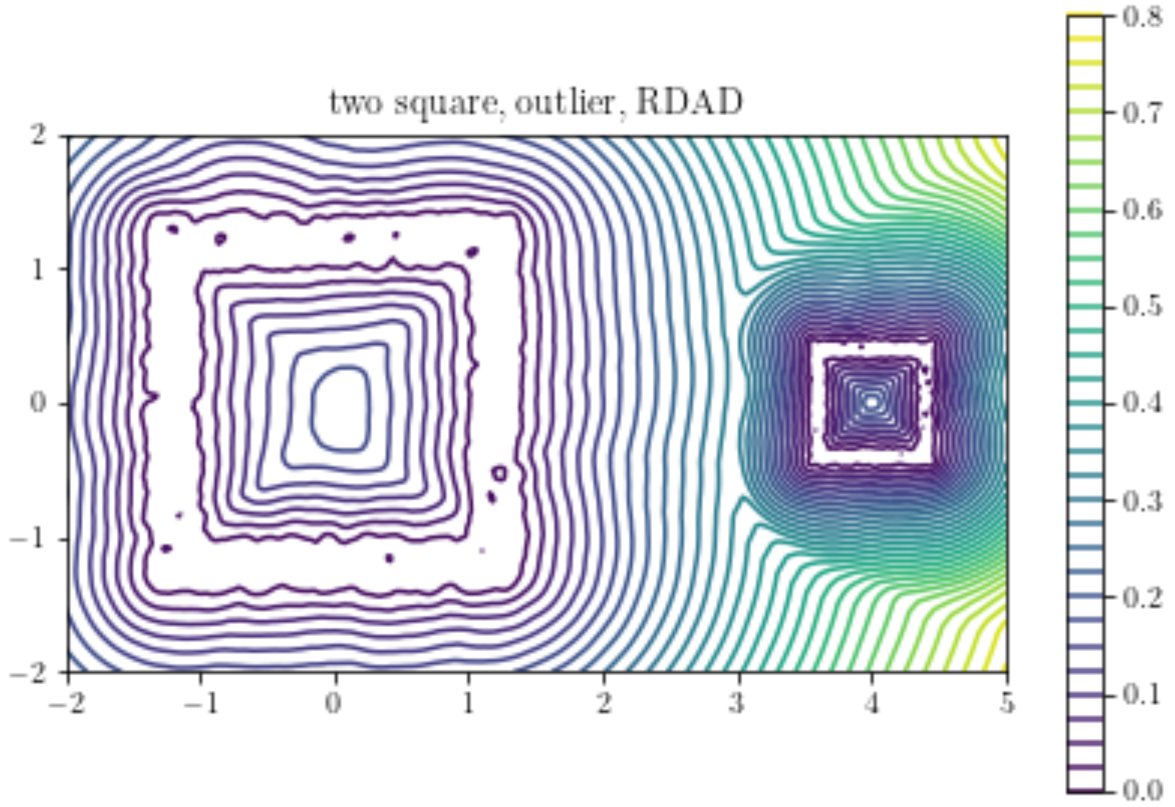
Outlier

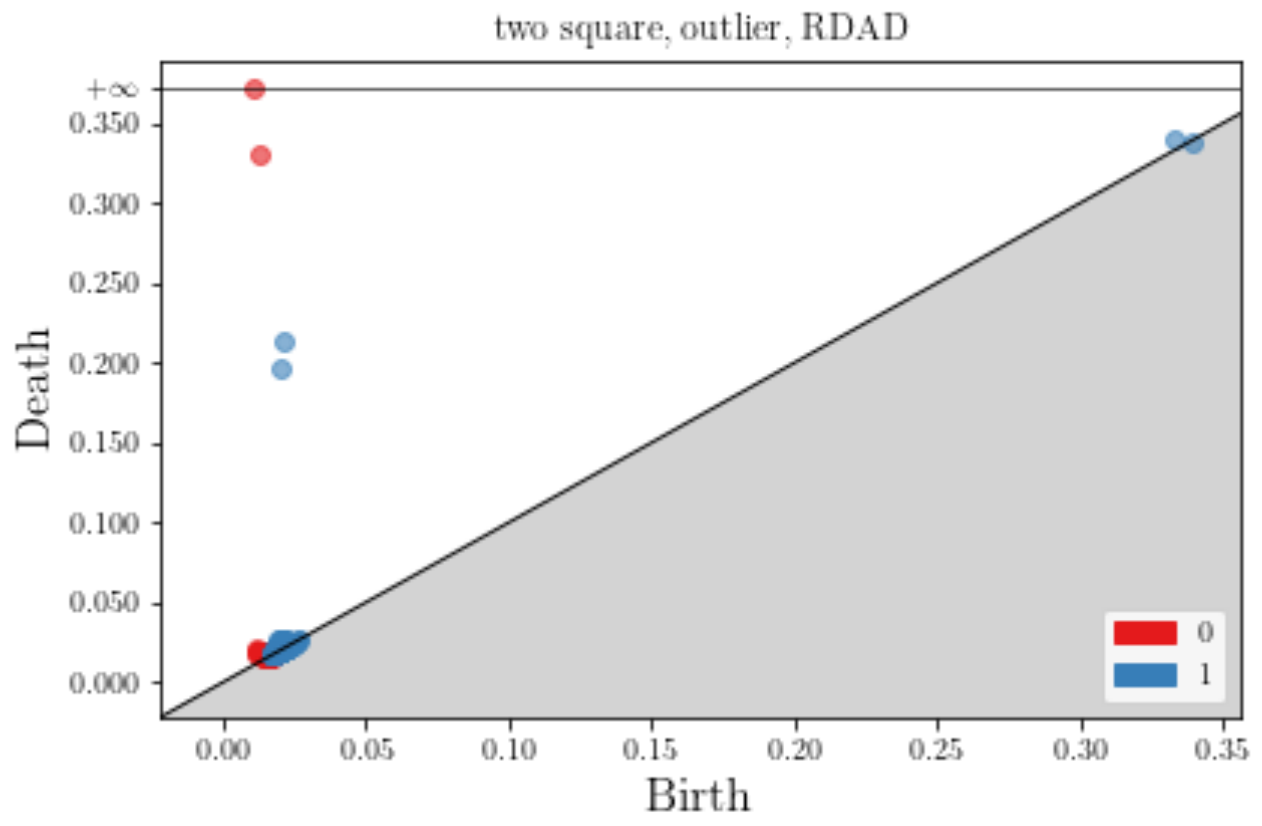


Weighted distance v.s. RDAD

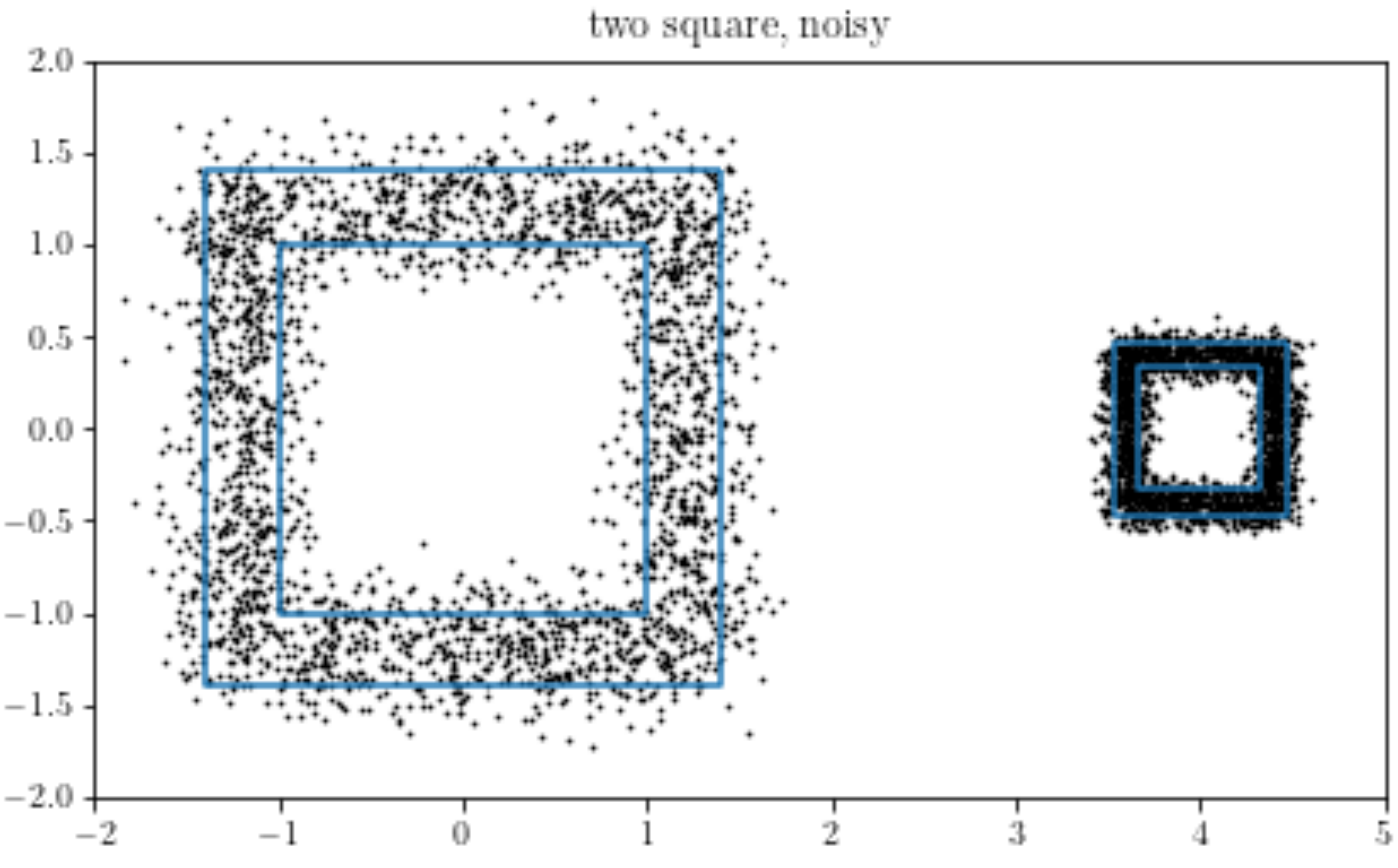




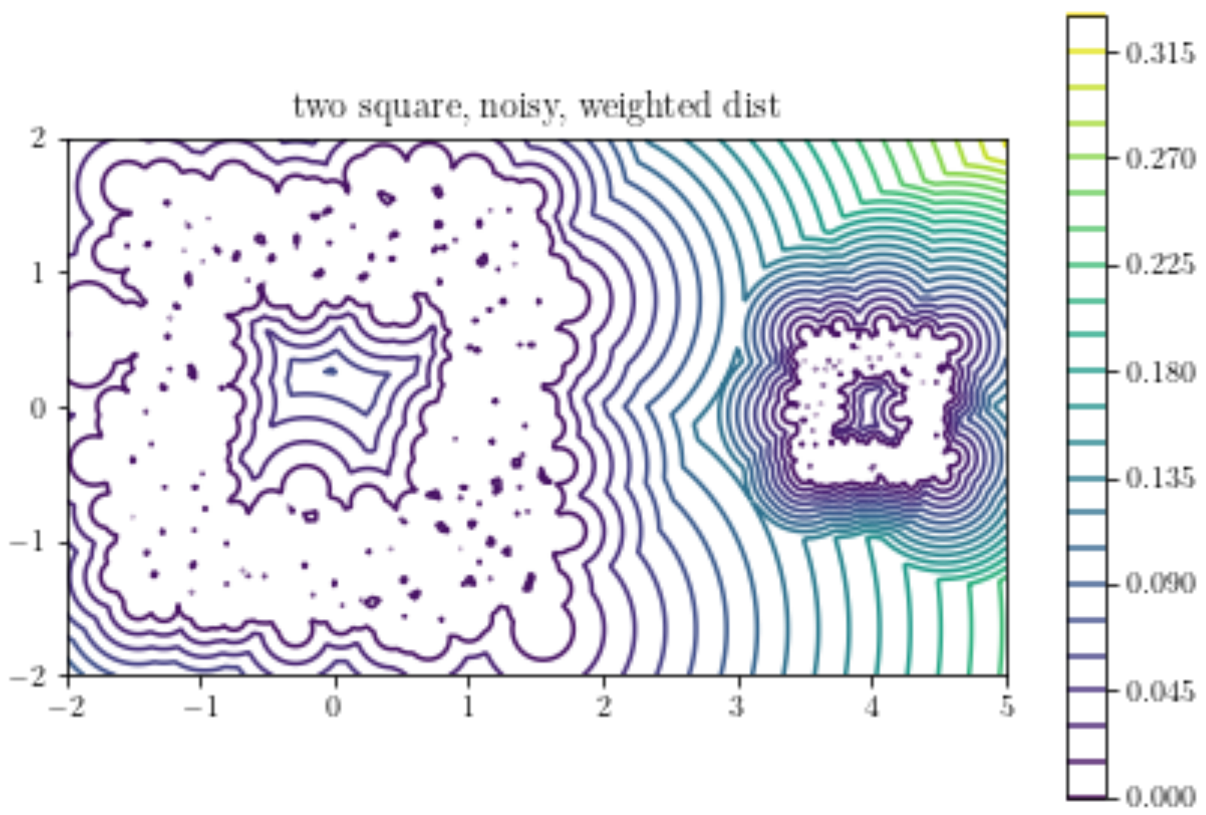


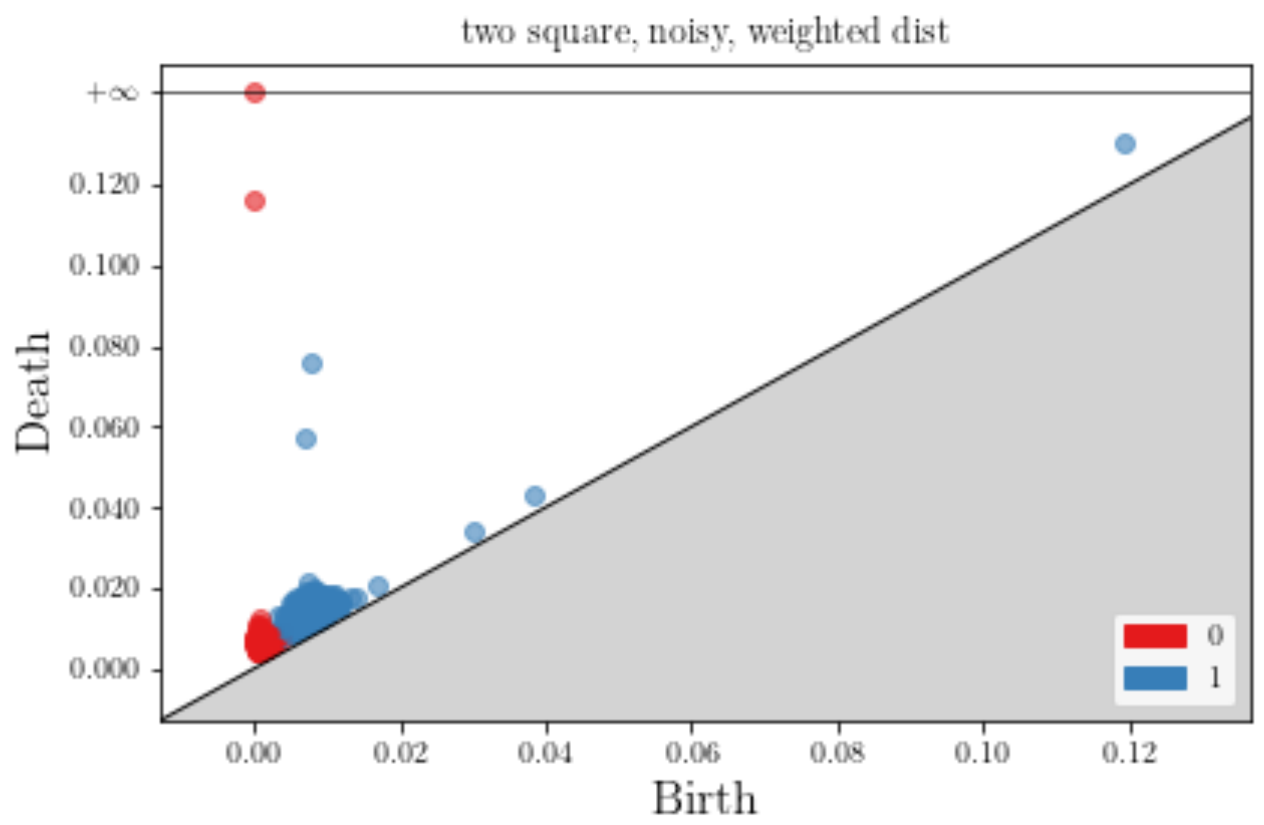


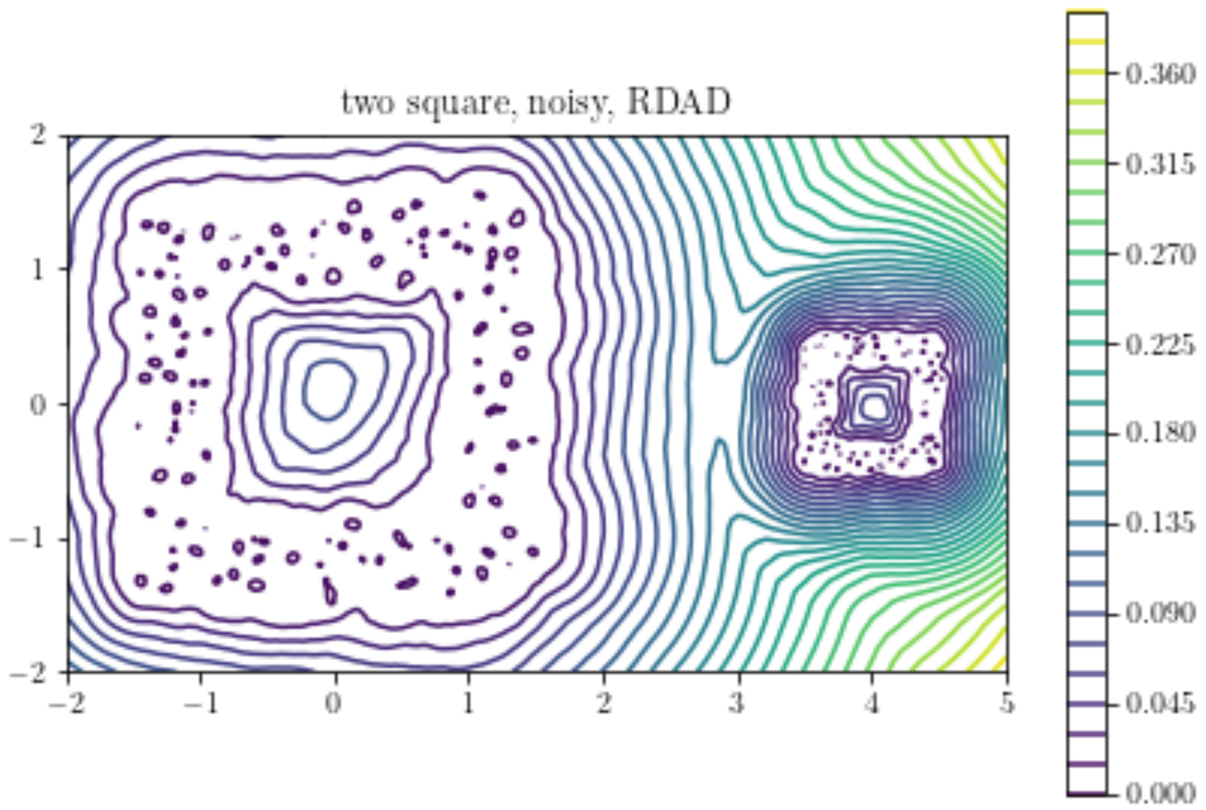
Additive noise

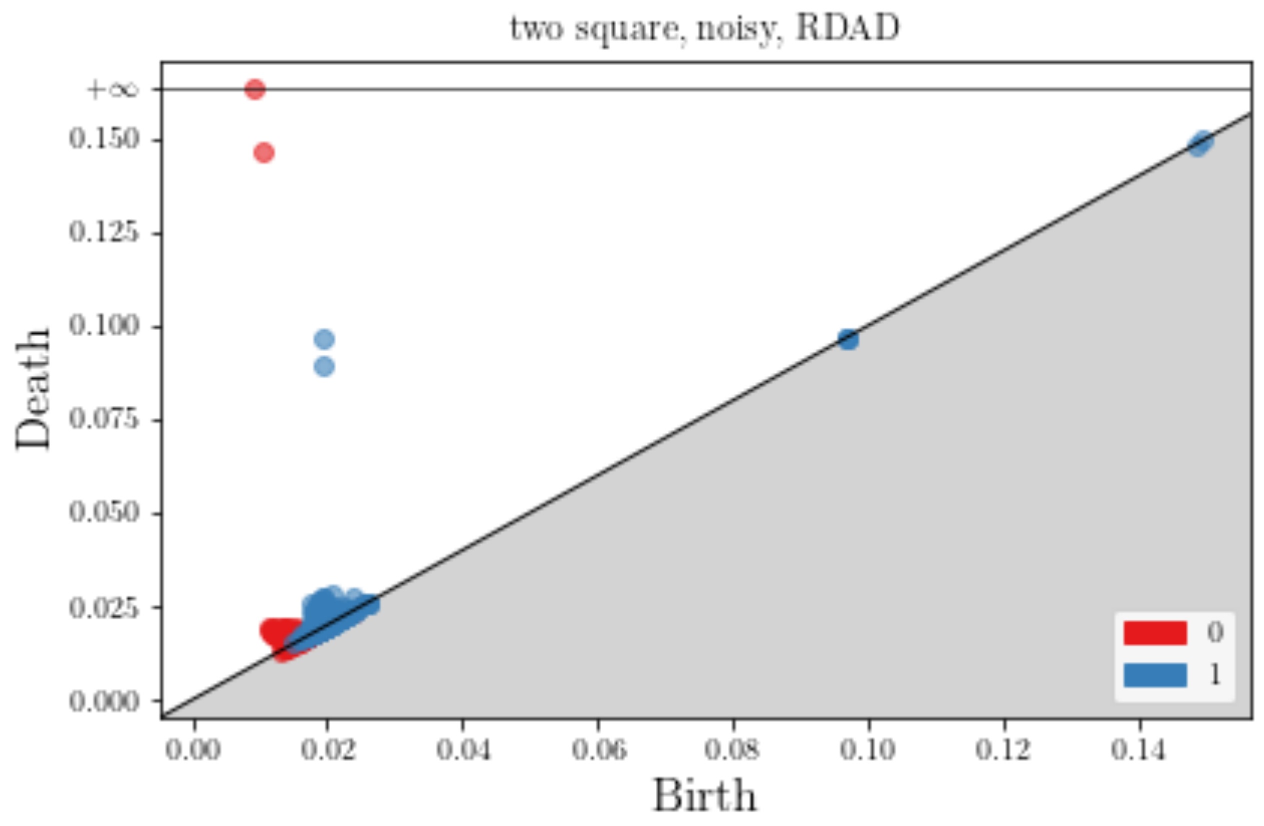


Weighted distance v.s. RDAD









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

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

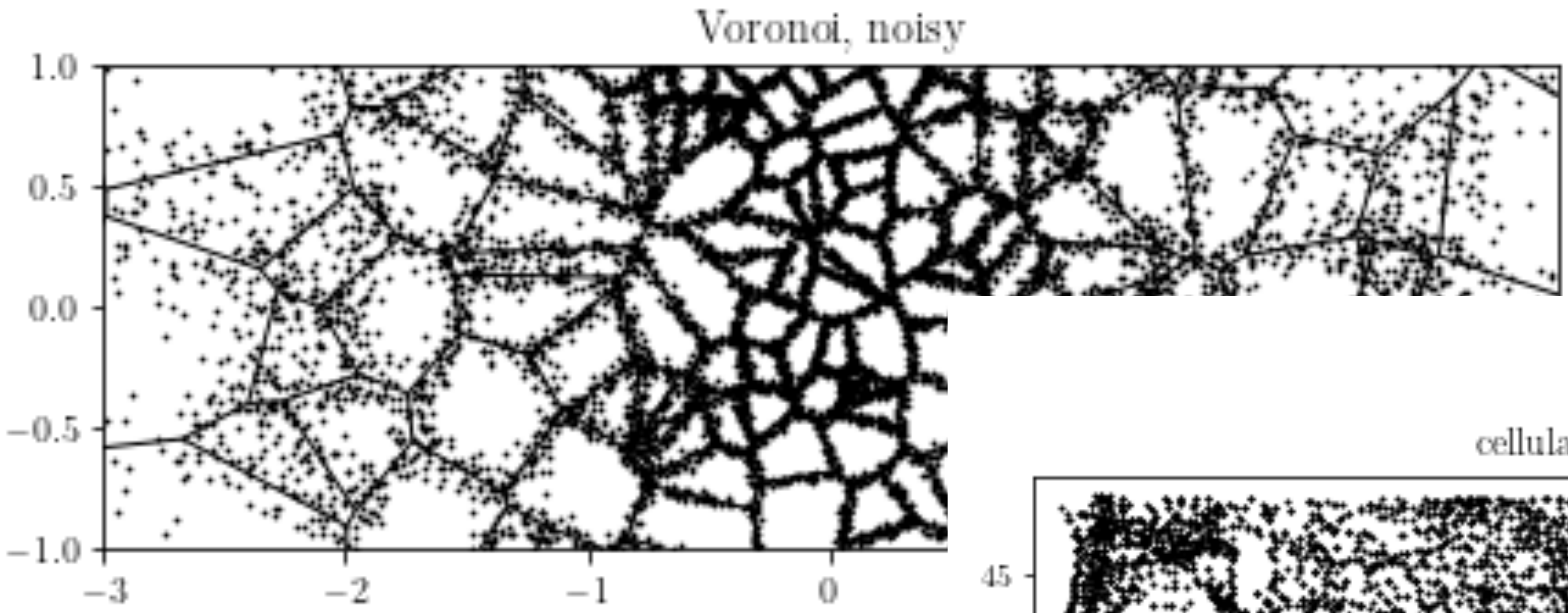
Statistical Convergence?

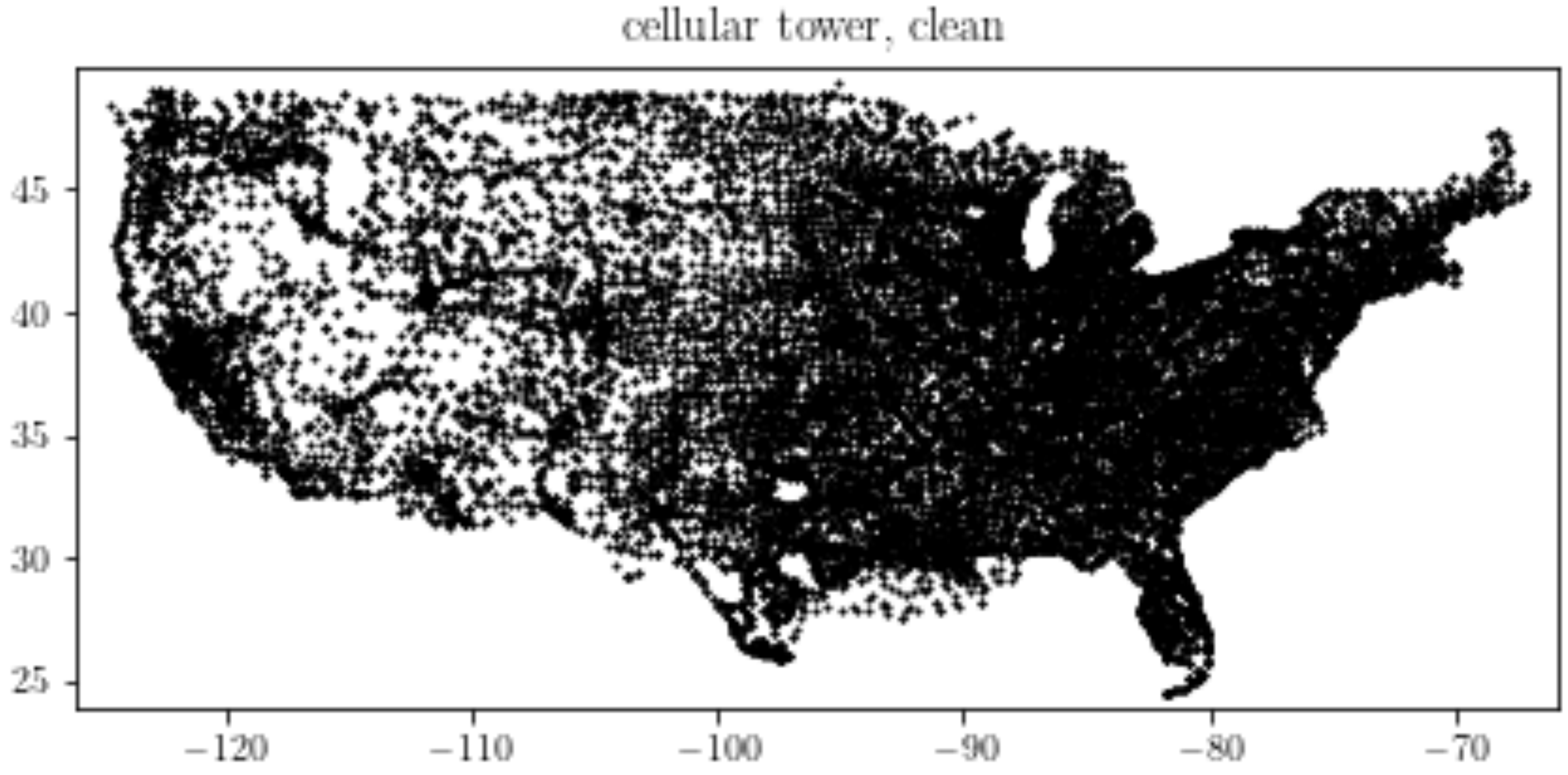
Theorem

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

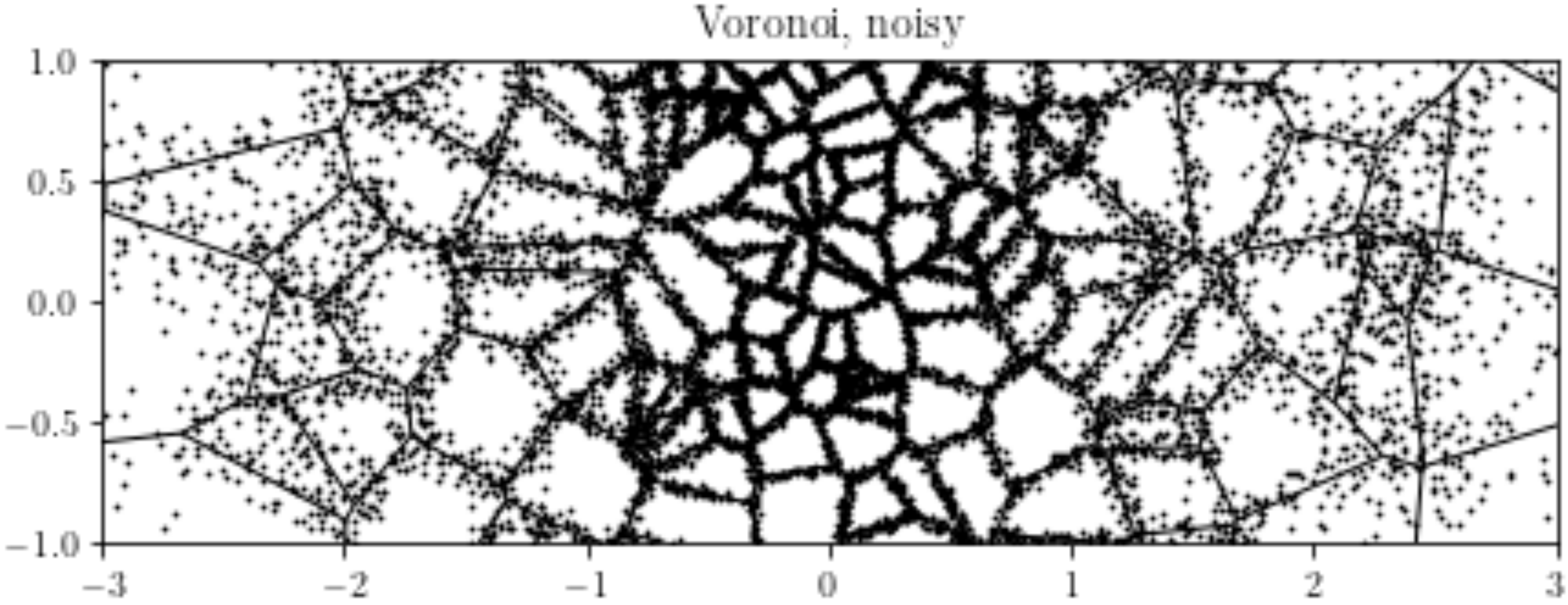
$$\sqrt{N}(\widehat{RDAD}^2 - RDAD^2) \xrightarrow{\text{weakly in } L^\infty(K)} \text{a centered Gaussian process}$$

Simulations

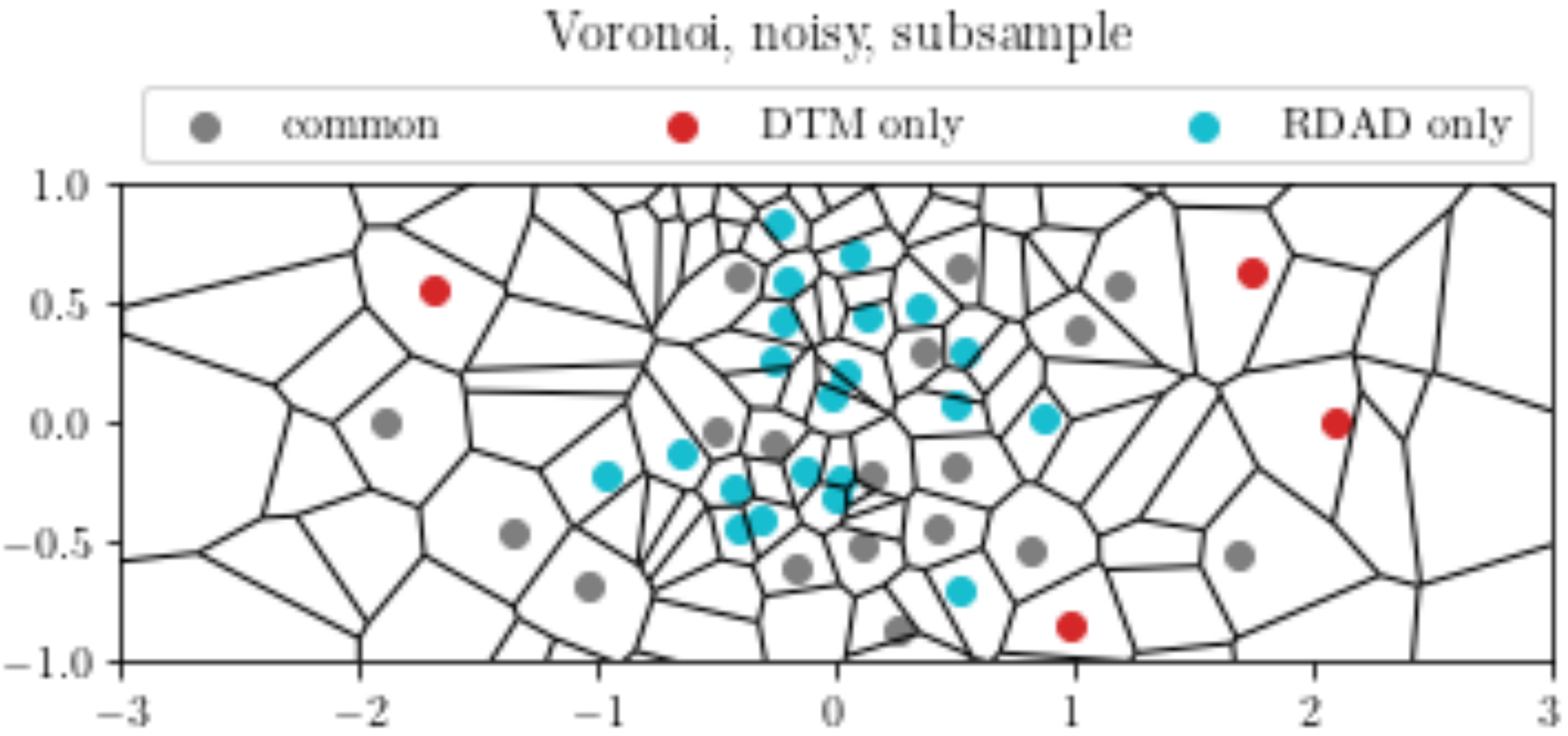

Voronoi, noisy
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0


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

Noisy Voronoi



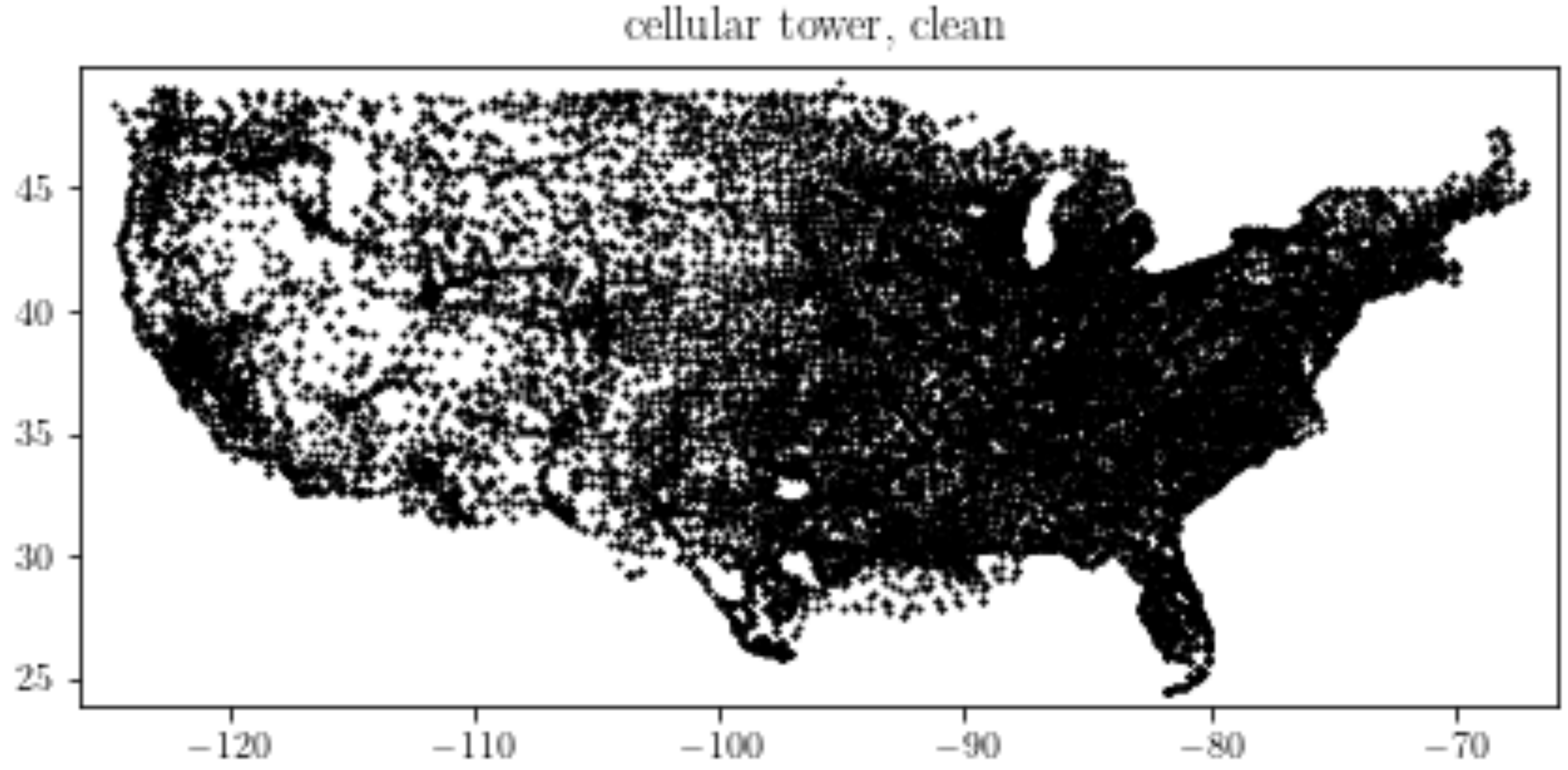
DTM and RDAD



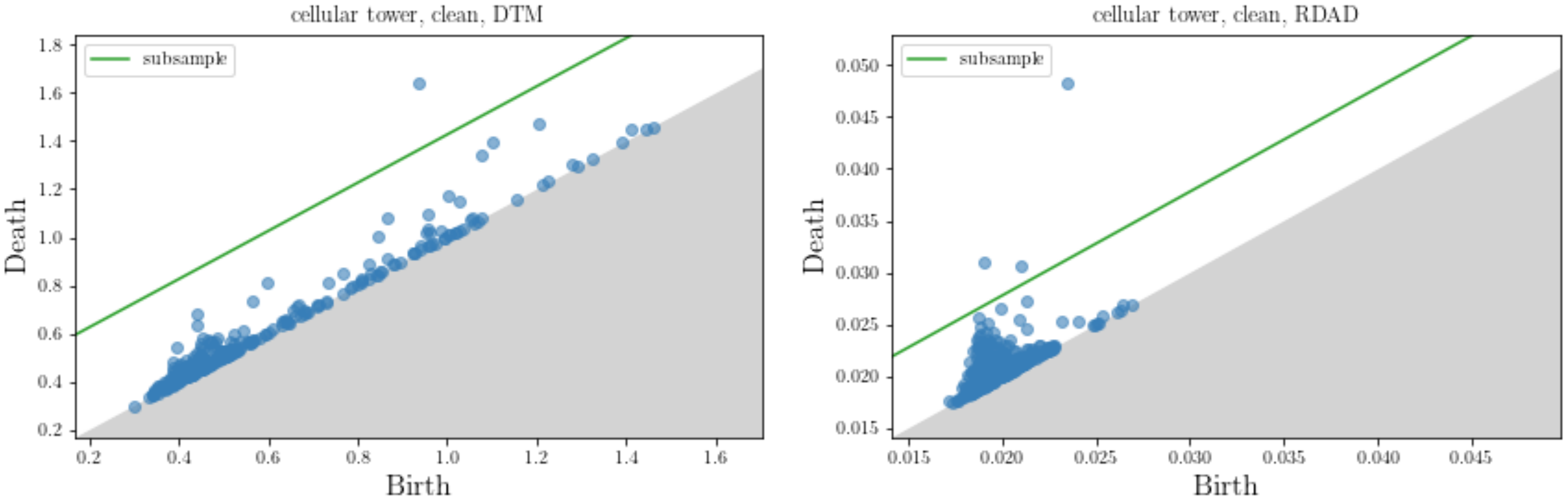
Voronoi, noisy, subsample
common
DTM only
RDAD only
1.0
0.5
0.0
−0.5
−1.0
−3
−2
−1
0
1
2
3

Cellular Towers

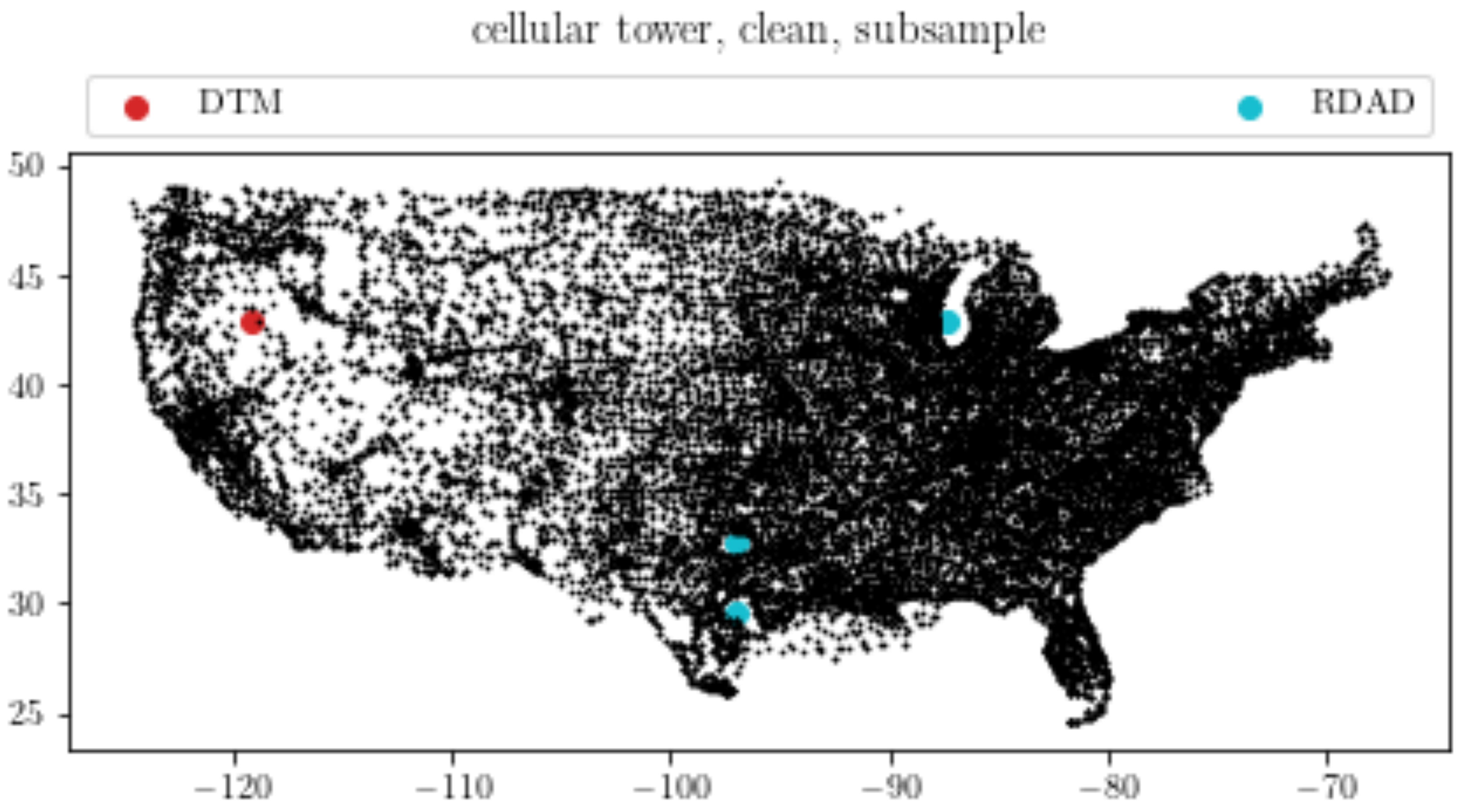
Cellular Towers
(HIFLD, 2021)



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



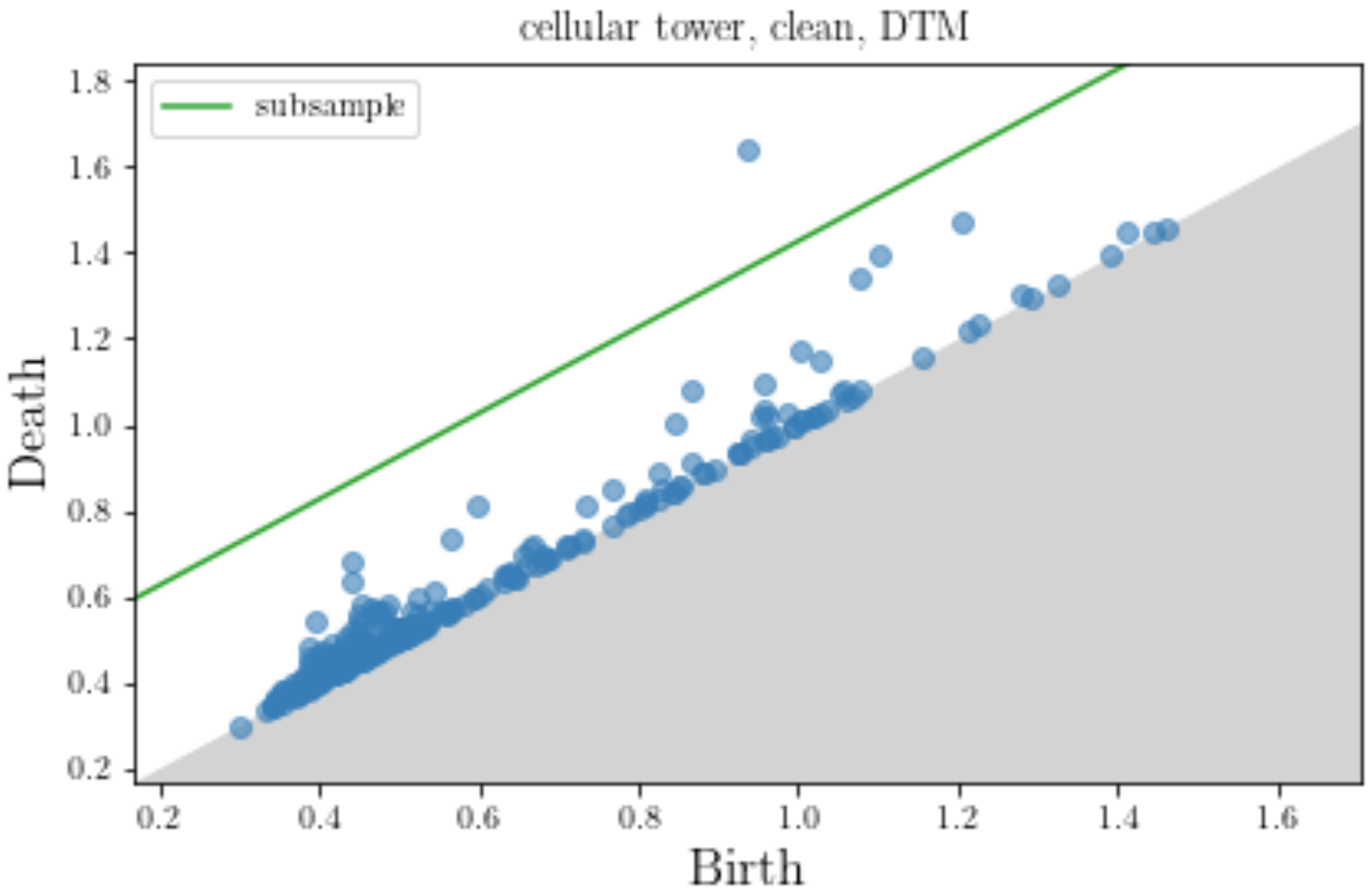
Looking Forward

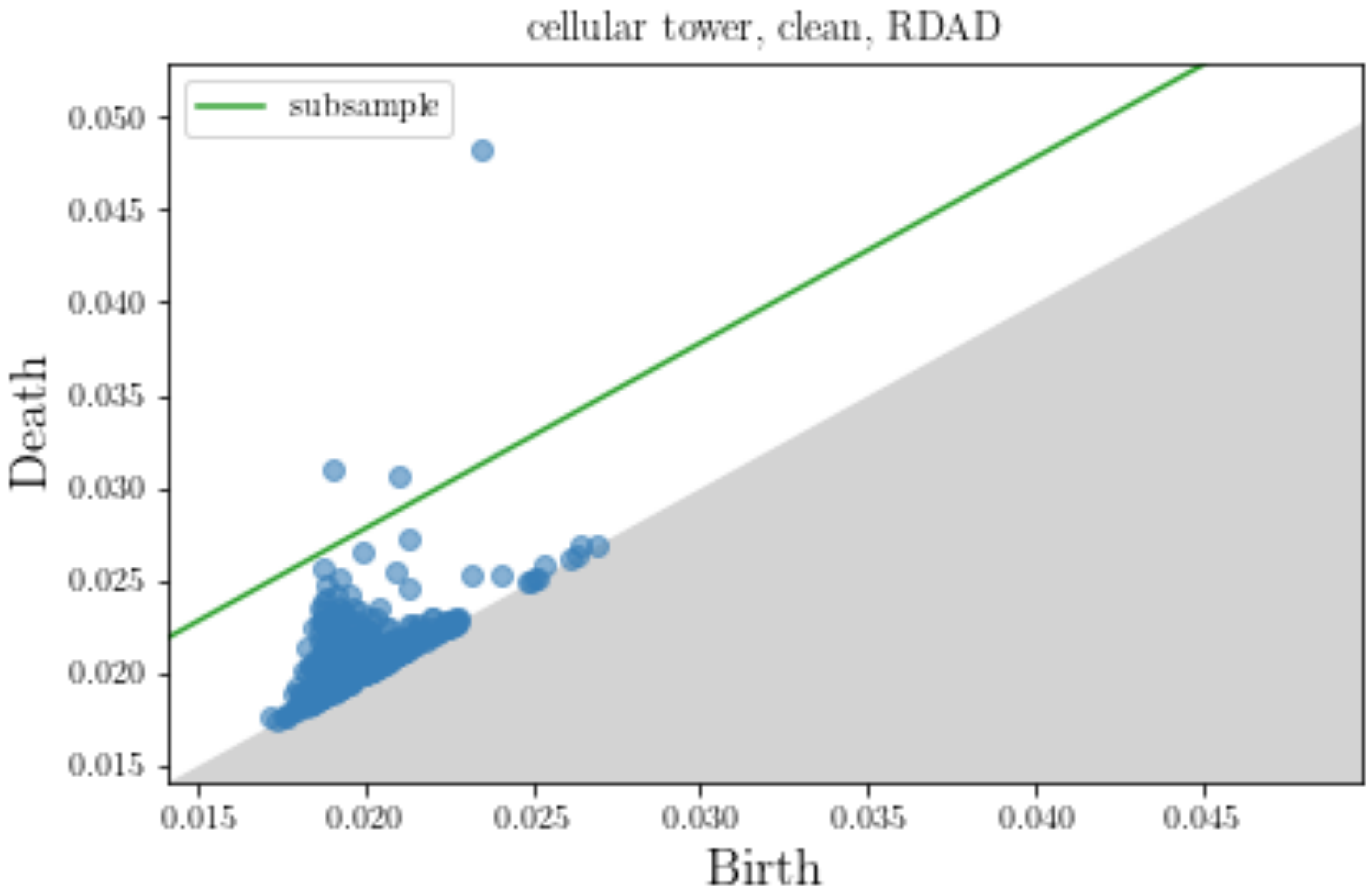
Ongoing / Future Works

- Bootstrapping properties of RDAD?

Ongoing / Future Works

- Bootstrapping properties of RDAD?



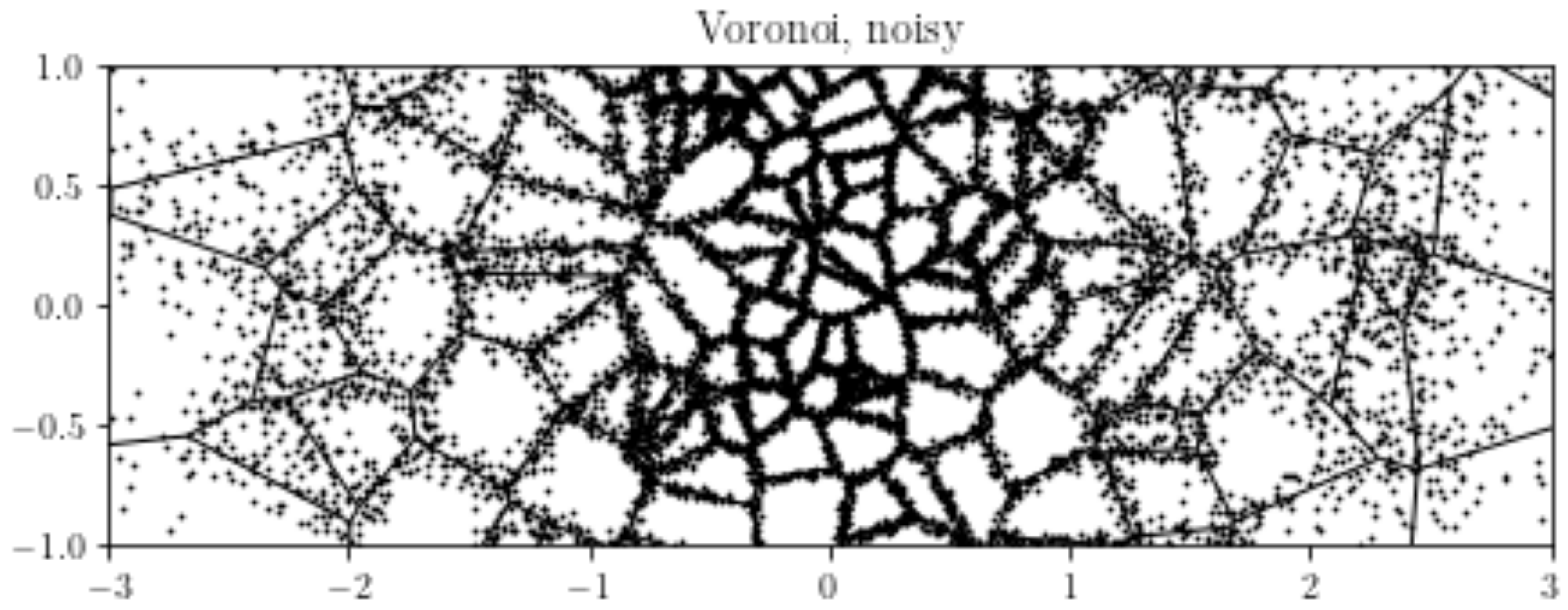


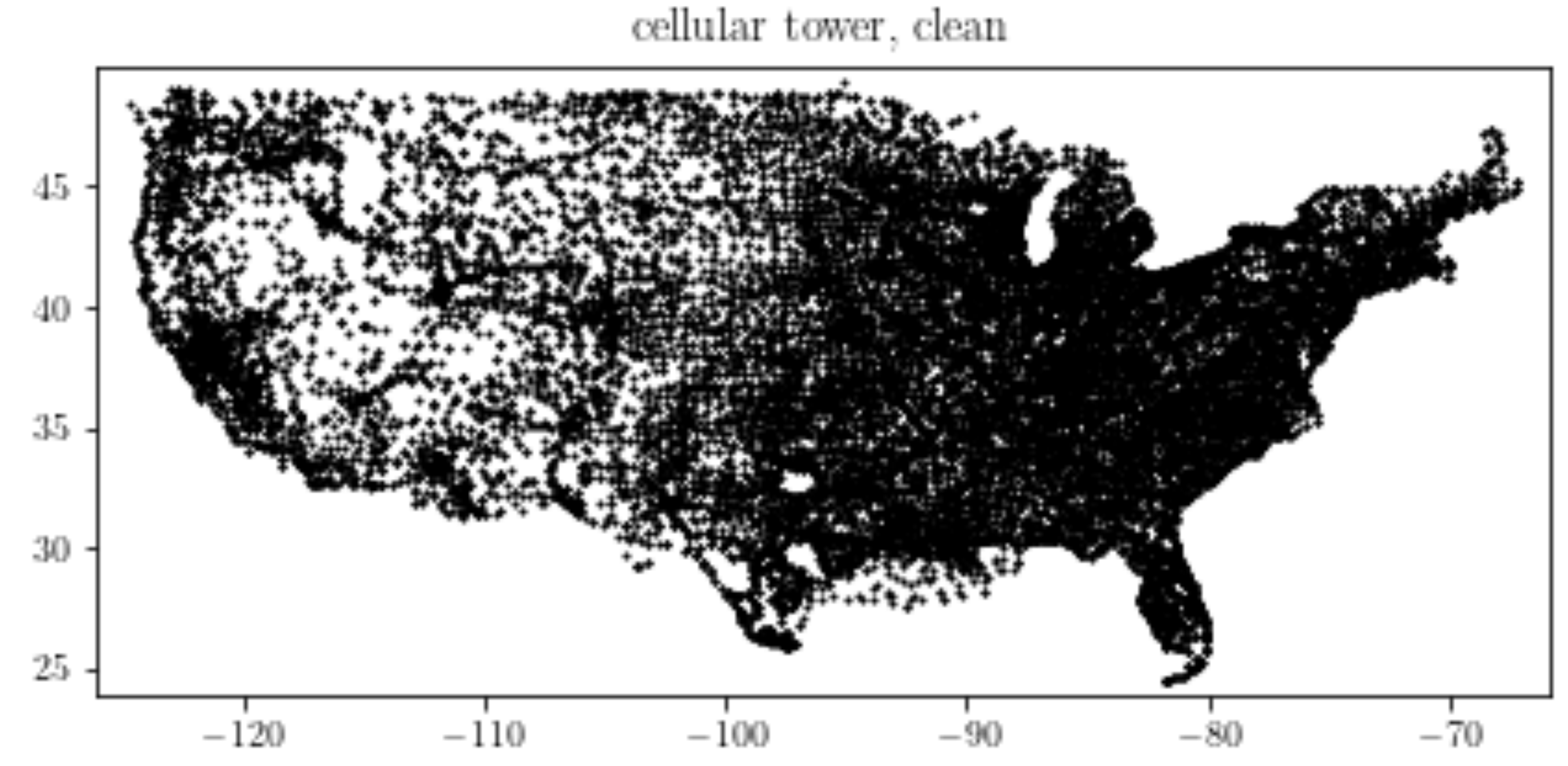
Ongoing / Future Works

- Bootstrapping properties of RDAD?
- Efficient implementation?

Ongoing / Future Works

- Bootstrapping properties of RDAD?
- Efficient implementation?



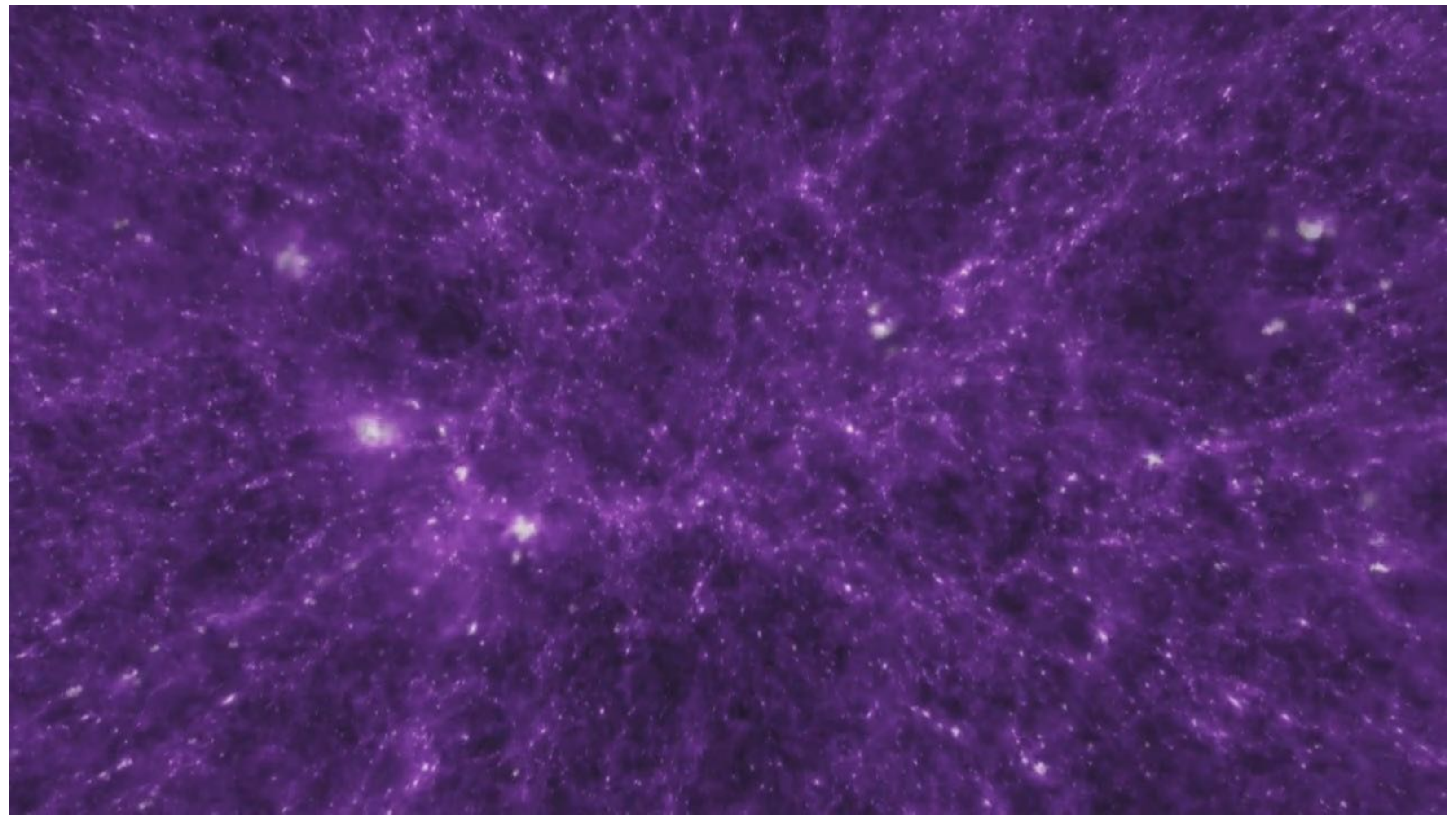


Ongoing / Future Works

- Bootstrapping properties of RDAD?
- Efficient implementation?
- Inference of Cosmological Parameters?

Ongoing / Future Works

- Bootstrapping properties of RDAD?
- Efficient implementation?
- Inference of Cosmological Parameters?



Topological Data Analysis Resources

AATRN

Topological Data Analysis Resources

AATRN (Henry Adams)

Gudhi

Scikit-TDA

Topological Data Analysis Resources

AATRN (Henry Adams)

गुढी

Gudhi

Scikit-TDA

Otter et al. *EPJ Data Science* (2017) 6:17
DOI 10.1140/epjds/s13688-017-0109-5

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE — Open Access

CrossMark

A roadmap for the computation of persistent homology

Nina Otter[1,3], Mason A Porter[4,1,2*], Ulrike Tillmann[1,3], Peter Grindrod[1] and Heather A Harrington[1]

*Correspondence: mason@math.ucla.edu
[4]Department of Mathematics, UCLA, Los Angeles, CA 90095, USA
Full list of author information is available at the end of the article

Abstract
Persistent homology (PH) is a method used in topological data analysis (TDA) to study qualitative features of data that persist across multiple scales. It is robust to perturbations of input data, independent of dimensions and coordinates, and

survey [Nina et al, 2017]

Topological Data Analysis Resources

AATRN (Henry Adams)

Gudhi

Scikit-TDA


Otter et al. EPJ Data Science (2017) 6:17
DOI 10.1140/epjds/s13688-017-0109-5

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

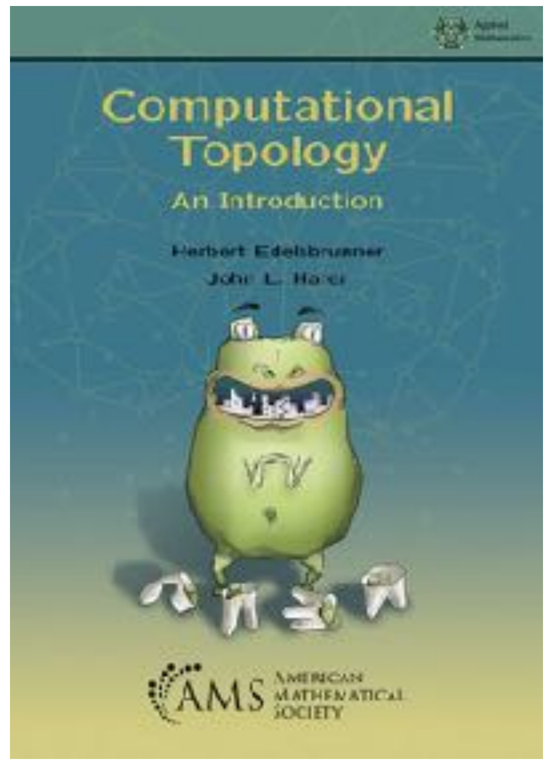
CrossMark

A roadmap for the computation of persistent homology

Nina Otter[1,3], Mason A Porter[4,1,2*], Ulrike Tillmann[1,3], Peter Grindrod[1] and Heather A Harrington[1]

*Correspondence: mason@math.ucla.edu
[4]Department of Mathematics, UCLA, Los Angeles, CA 90095, USA
Full list of author information is available at the end of the article

Abstract
Persistent homology (PH) is a method used in topological data analysis (TDA) to study qualitative features of data that persist across multiple scales. It is robust to perturbations of input data, independent of dimensions and coordinates, and



TDA textbook [Edelsbrunner and Harer, 2010]

survey [Nina et al, 2017]

Topological Data Analysis Resources

AATRN (Henry Adams)

Gudhi

Scikit-TDA


Otter et al. EPJ Data Science (2017) 6:17
DOI 10.1140/epjds/s13688-017-0109-5

EPJ Data Science
a SpringerOpen Journal

EPJ.org

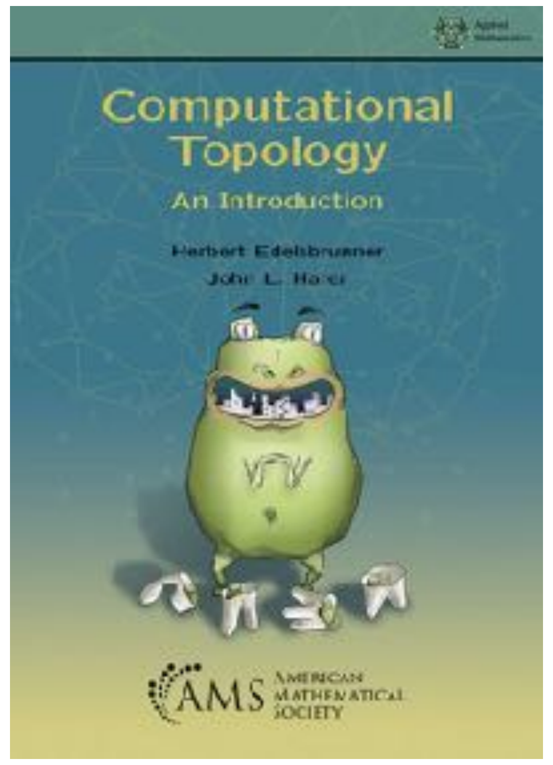
REGULAR ARTICLE Open Access

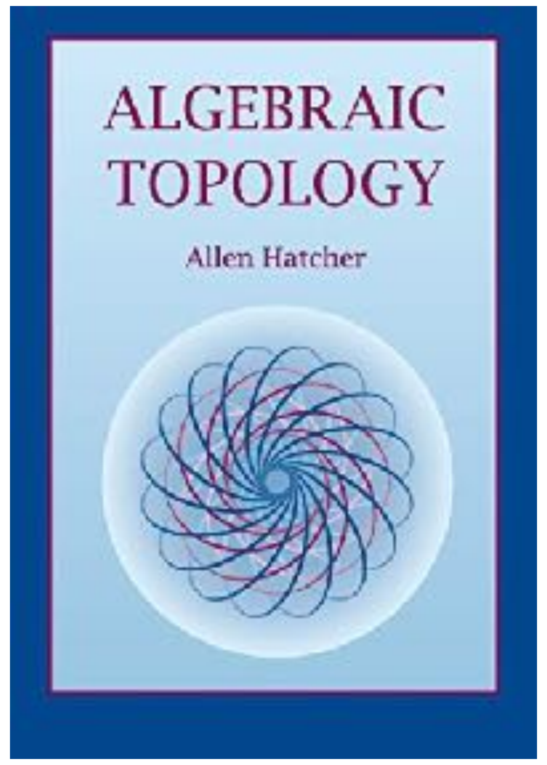
A roadmap for the computation of persistent homology

Nina Otter[1,3], Mason A Porter[4,1,2*], Ulrike Tillmann[1,3], Peter Grindrod[1] and Heather A Harrington[1]

*Correspondence: mason@math.ucla.edu
[4]Department of Mathematics, UCLA, Los Angeles, CA 90095, USA
Full list of author information is available at the end of the article

Abstract
Persistent homology (PH) is a method used in topological data analysis (TDA) to study qualitative features of data that persist across multiple scales. It is robust to perturbations of input data, independent of dimensions and coordinates, and





survey [Nina et al, 2017]

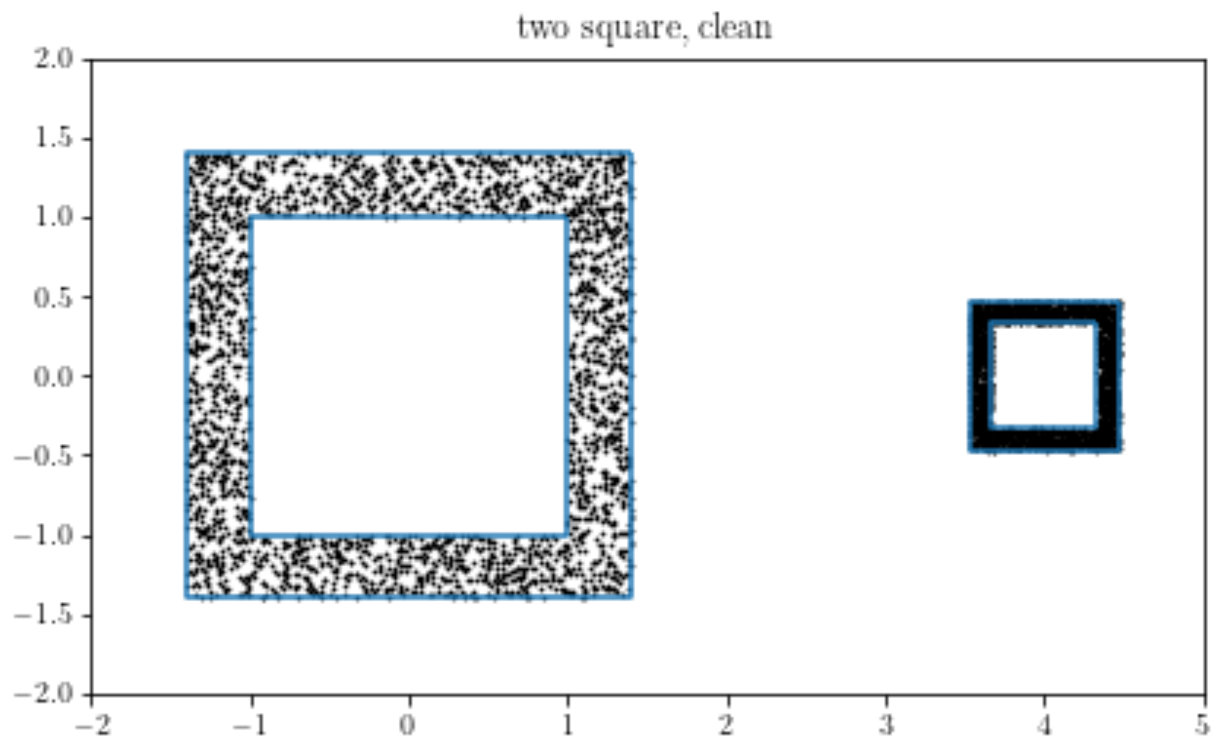
TDA textbook [Edelsbrunner and Harer, 2010]

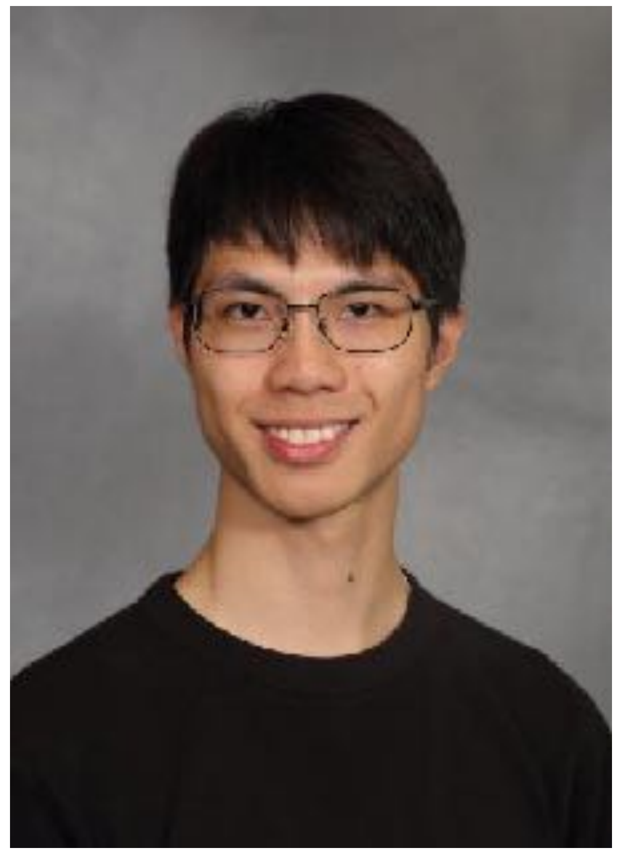
Topology textbook [Hatcher, 2002]

Take-Home Messages

- Topology is useful for understanding nonlinear geometric structures.
- Topological features in low signal-to-noise environment is hard, but doable.

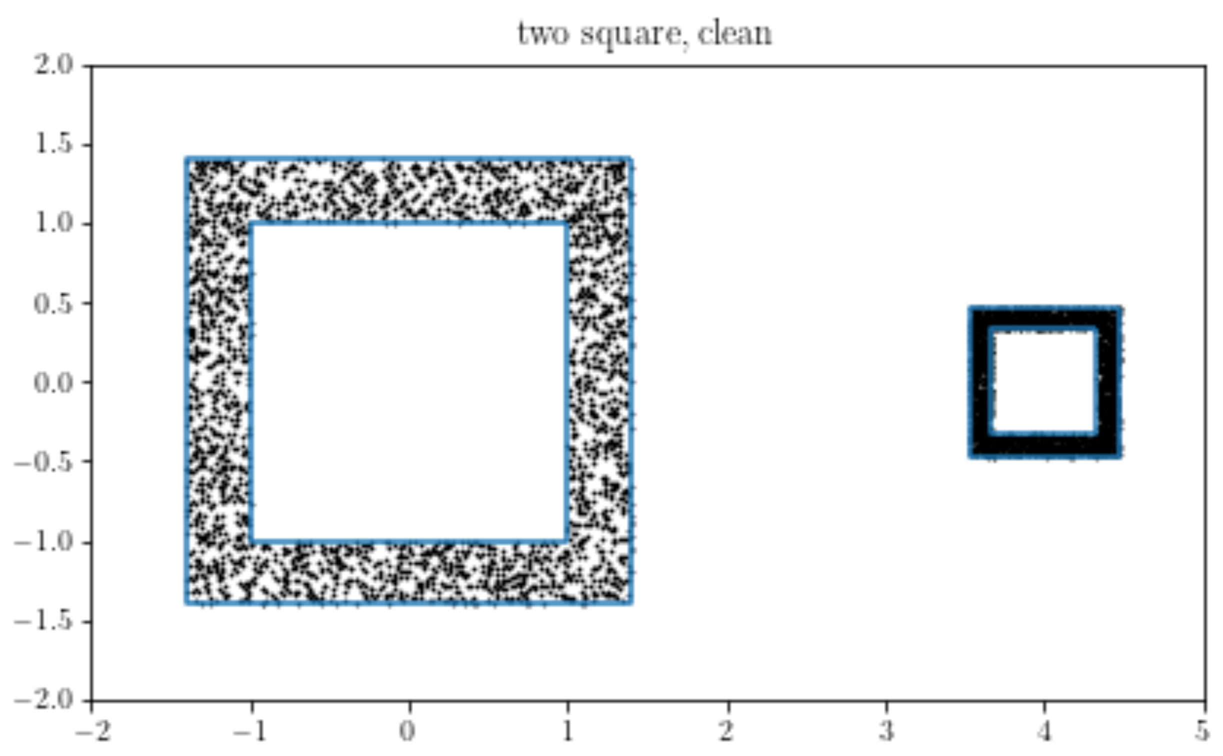
- Chunyin Siu (Alex)
- Cornell University
- cs2323@cornell.edu

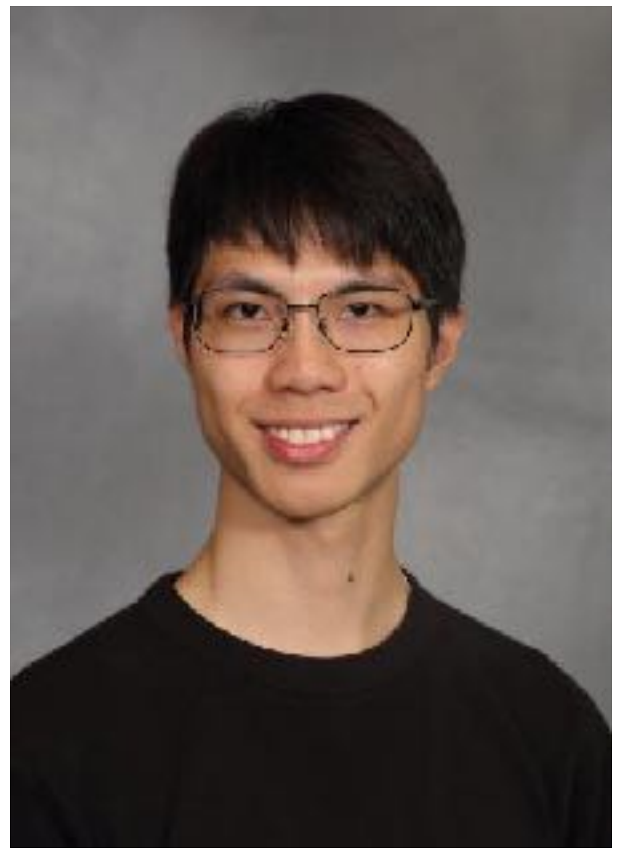




Thank you!

- Chunyin Siu (Alex)
- Cornell University
- cs2323@cornell.edu

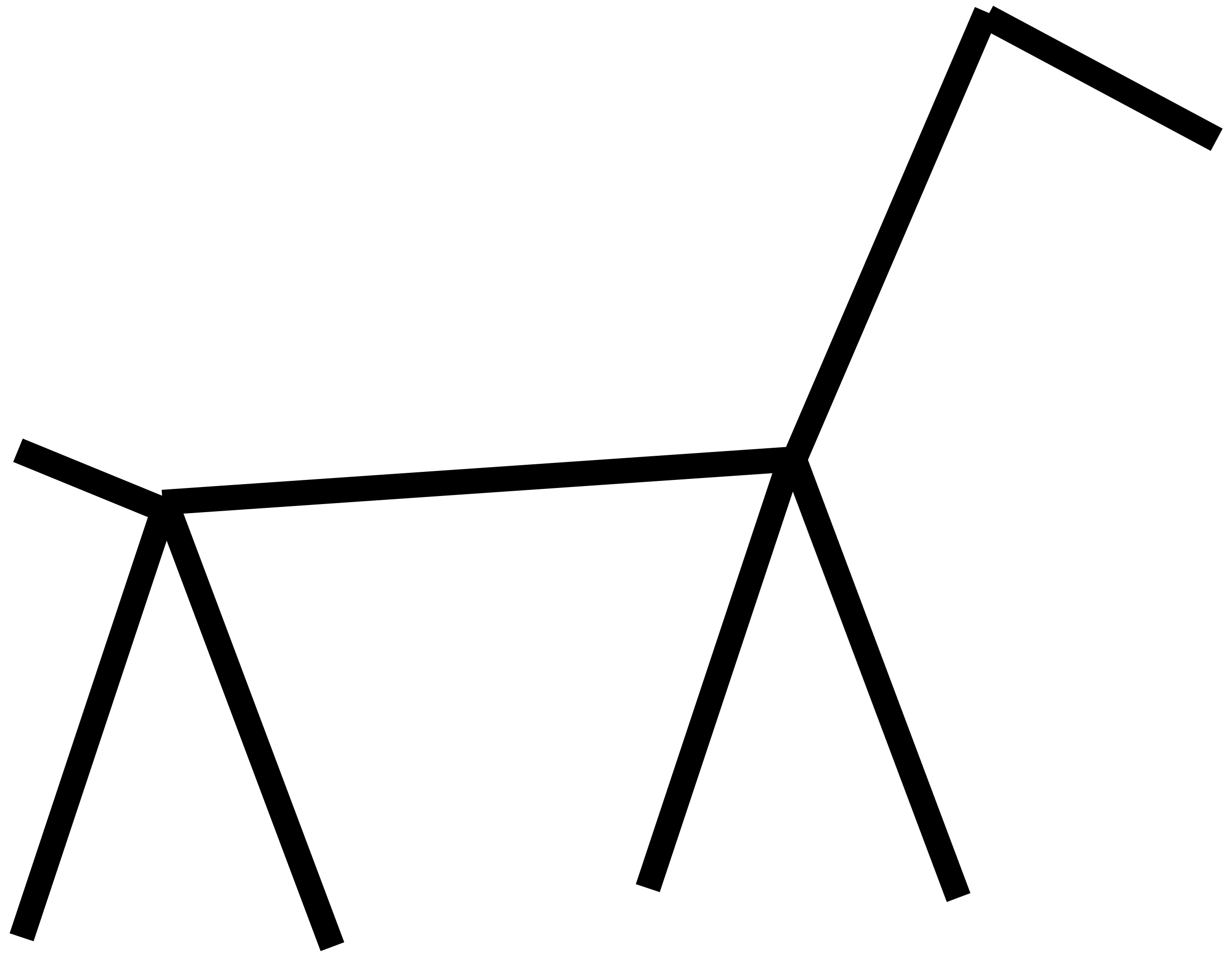




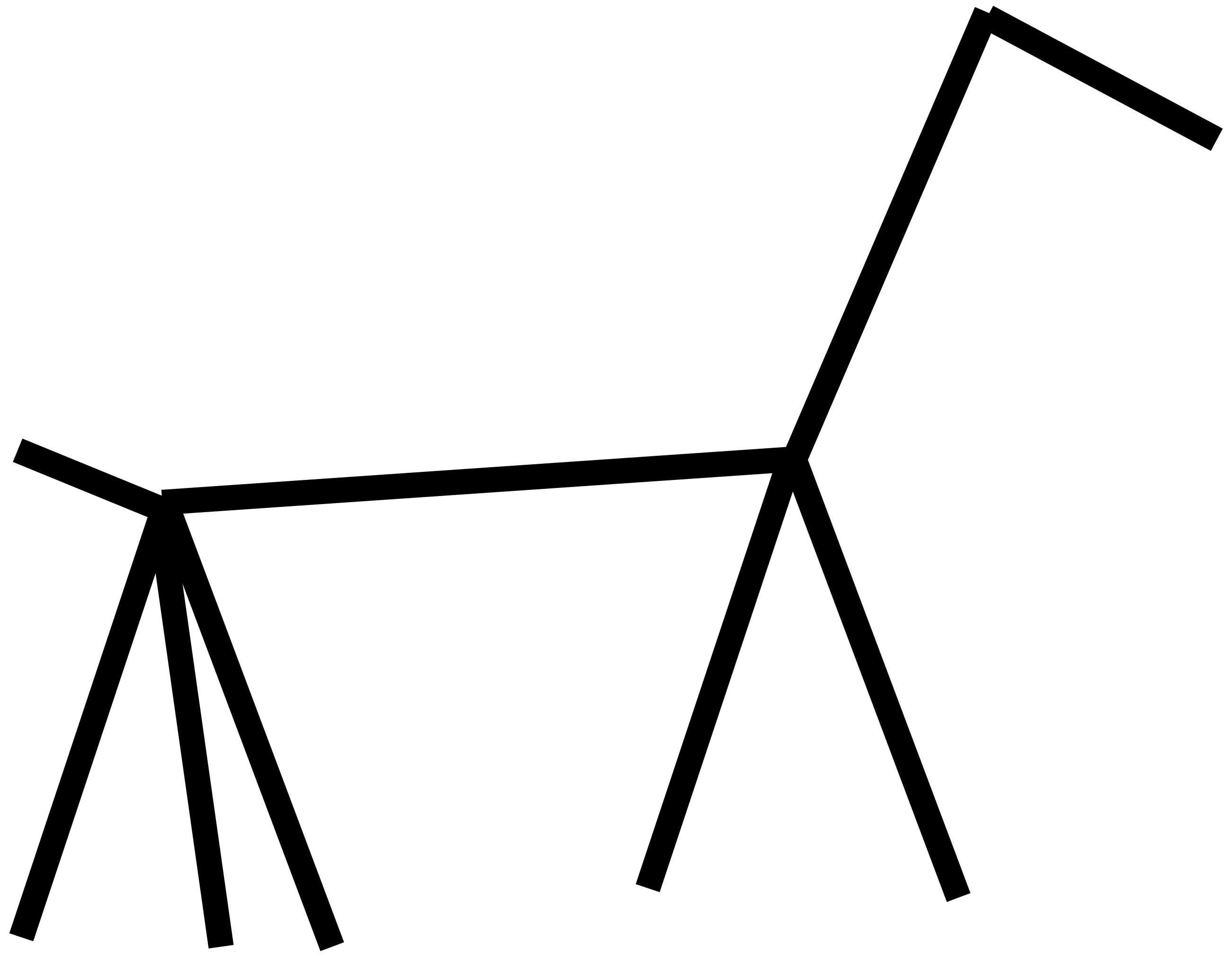
References

- 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
- Bruel 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., and Zomorodian, A. (2009). The theory of multidimensional persistence. *Discrete Comput Geom*, 71–93
- Chazal, F., Cohen-Steiner, D., and Merigot, 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.

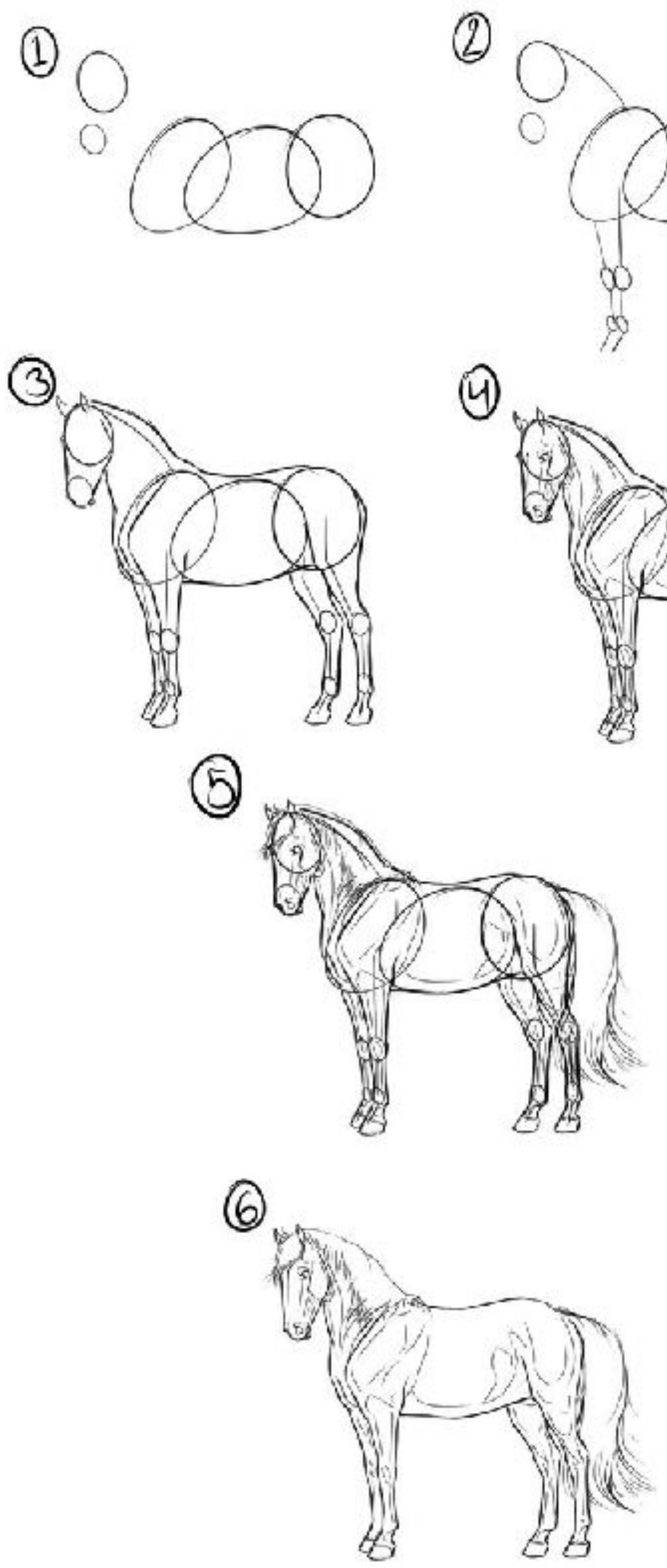
- Moon, C., Giansiracusa, N., and Lazar, N.A. (2018). Persistence terrace for topological inference of point cloud data. *Journal of Computational and Graphical Statistics,* 27(3), 576–586.
- Hickok, A. (2022). A Family of Density-Scaled Filtered Complexes
- HIFLD (2021). Cellular towers.
- Krishnapriyan A.S., Haranczyk M., and Morozov D. (2020). Topological descriptors help predict guest adsorption in nanoporous materials. *The Journal of Physical Chemistry C*, 124(17): 9360–9368.
- Saggar M., Shine J.M., Liegeois R., Dosenbach N.U.F., Damien Fair D. (2022). Precision dynamical mapping using topological data analysis reveals a hub-like transition state at rest. *Nature Communications*, 13(1): 4791.
- Wilding, G., Nevenzeel K., van de Weygaert R., Vegter G., Pranav P., Jones B.J.T., Efstathiou K., Feldbrugge J. (2021) Persistent homology of the cosmic web – I. Hierarchical topology in ΛCDM cosmologies. *Monthly Notices of the Royal Astronomical Society*, 507 (2): 2968–2990.



horse



non-horse



TinyGlitch on DeviantArt, from https://www.pinterest.cl/pin/151644712438831173/