Same but different:

using distance correlation to compare different presentations of persistent homology

Katharine Turner (ANU). Joint work with Gard Spreemann (EPFL).

Statistical Challenges for Topological Data Analysis

Also known as “how STAT101 is no help”

STAT101

- Bell Curve (i.e. normal distribution)
- Z-scores
- t-tests (p-values for null hypothesis testing)
- Correlation
- Regression
- Confidence intervals for means
- Chi-Square Tests

Treasure Chest: Non-parametric Statistics

"it is difficult to give a precise definition of nonparametric inference"

-Larry Wasserman, All of Nonparametric Statistics, Springer. (2007)

- distribution-free
- can sometimes be applied for data in general metric spaces
- no free lunch: *when a parametric test would be appropriate*, non-parametric tests usually have less power

In TDA we don't have parametric models and distributions

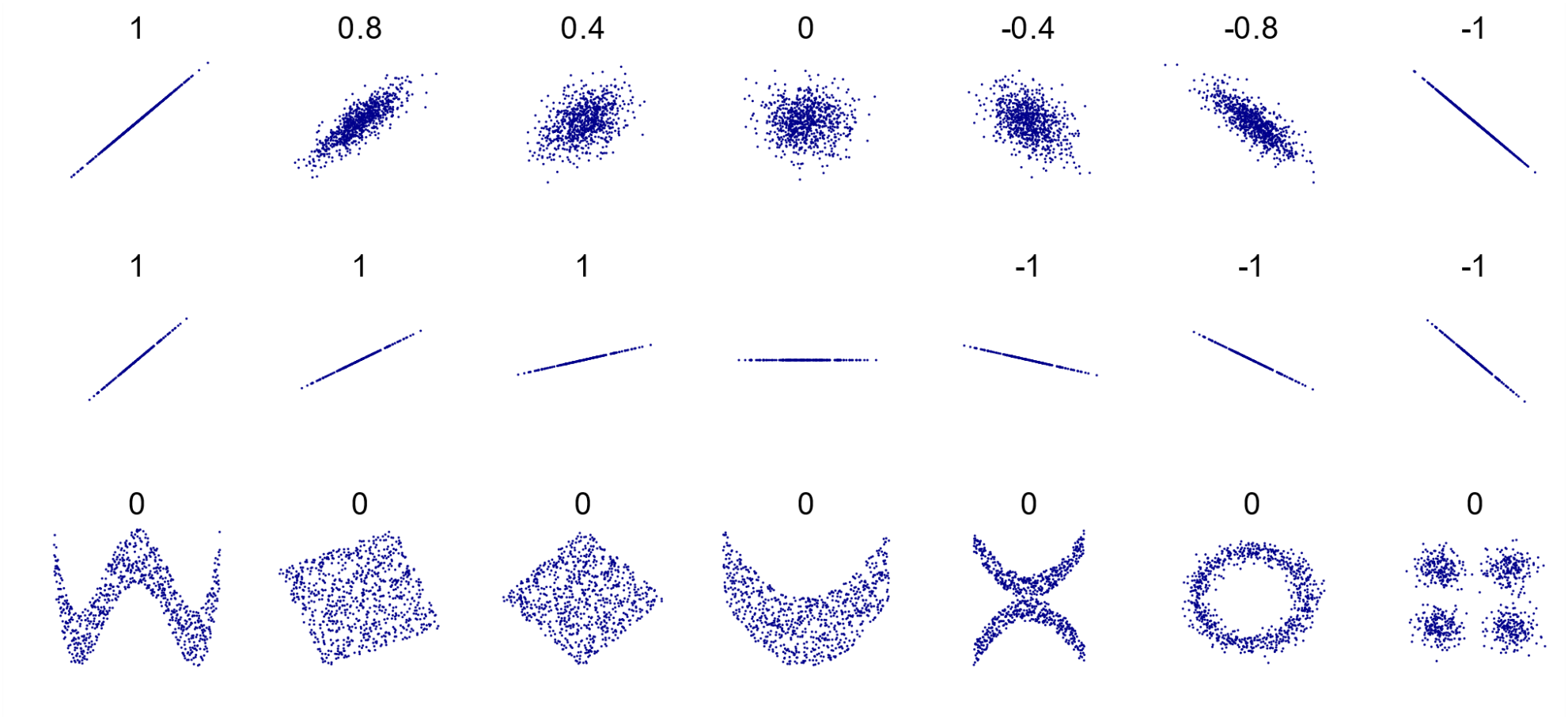
cor·re·la·tion

/ˌkôrəˈlāSH(ə)n/

noun

- a mutual relationship or connection between two or more things.
- interdependence of variable quantities.
- a quantity measuring the extent of interdependence of variable quantities.
- the process of establishing a relationship or connection between two or more measures.

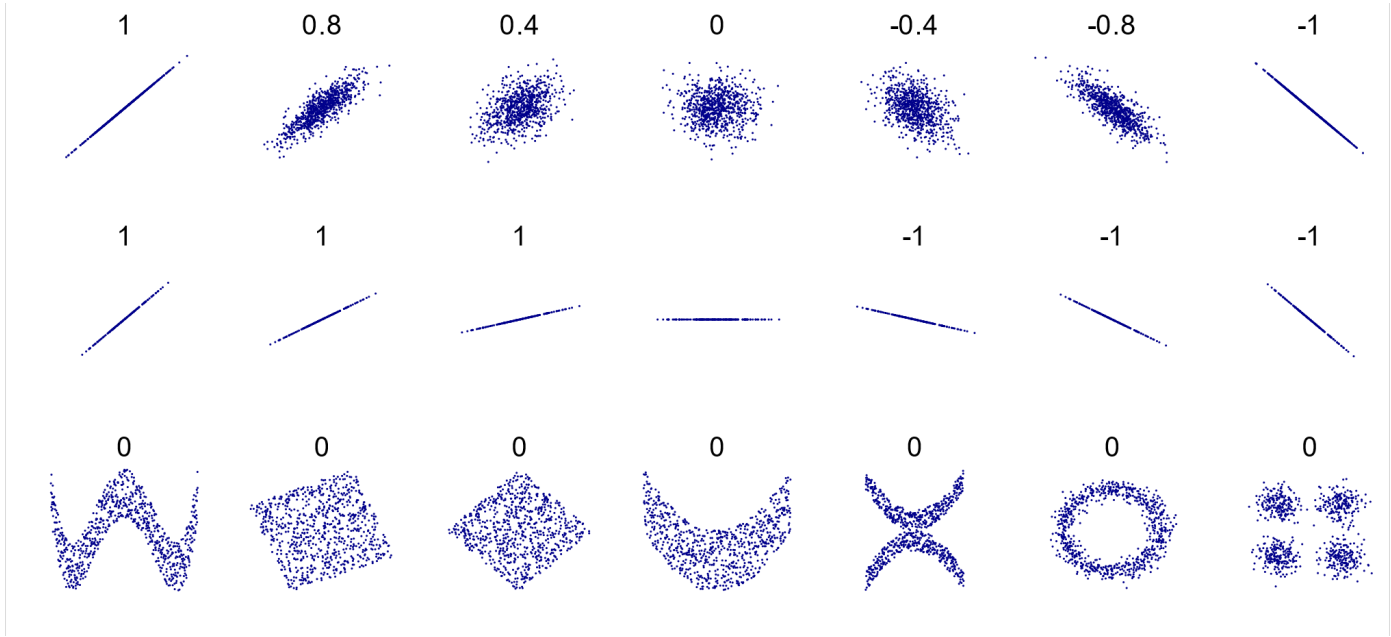
Pearson correlation coefficient



Several sets of (*x*, *y*) points, with the Pearson correlation coefficient

Source: Wikipeadia

Pearson correlation coefficient



Several sets of (*x*, *y*) points, with the Pearson correlation coefficient

Source: Wikipeadia

It reflects the noisiness and direction of a linear relationship

It does not measure the slope of that relationship

It does not capture many aspects of nonlinear relationships

Pearson correlation coefficient

- “Default” measure of correlation
- Only defined for real variables
- Parametric method – best suited for comparing Gaussians

$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

OFTEN NOT DEFINED/SUITABLE FOR TOPOLOGICAL SUMMARIES

Introducing Distance Correlation

The non-parametric answer to correlation that can be used for samples in any metric space

Distance Correlation

Given paired samples (X, Y) that can lie in DIFFERENT metric spaces, we can ask what the joint variability is of the pairwise distances

That is how related is $d_Y(y_i, y_j)$ to $d_X(x_i, x_j)$

We need to doubly recentre the pairwise distances before taking the normal correlation.

Doubly recentred distance function

X a random element with values in a connected metric space $(\mathcal{X}, d_X)$ with distribution μ.

$$a_\mu(x) = \int d_X(x, x')d\mu(x') \qquad D_\mu = \int d_X(x, x')d\mu^2(x)(x')$$

When these are finite we say that X has finite first moment

Definition: the doubly recentred distance function is defined as

$$d_\mu(x,y) := d_X(x,y) - a_\mu(x) - a_\mu(y) + D(\mu).$$

Defining Distance Correlation

Let X and Y be metric spaces. Let θ = (X, Y) have marginals μ and ν. We define the distance covariance of θ as

$$dCov(X,Y)^2 = \int d_\mu(x,x')d_\nu(y,y')d\theta^2((x,y)(x'y'))$$

We can define

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}$$

Strong Negative Type

- A metric space is of strong negative type if for all $x_1, x_2, \ldots x_n$

$$\sum_{i=1}^{n} \alpha_i = 0 \implies \sum_{i,j=1}^{n} \alpha_i \alpha_j d(x_i, x_j) \leq 0$$

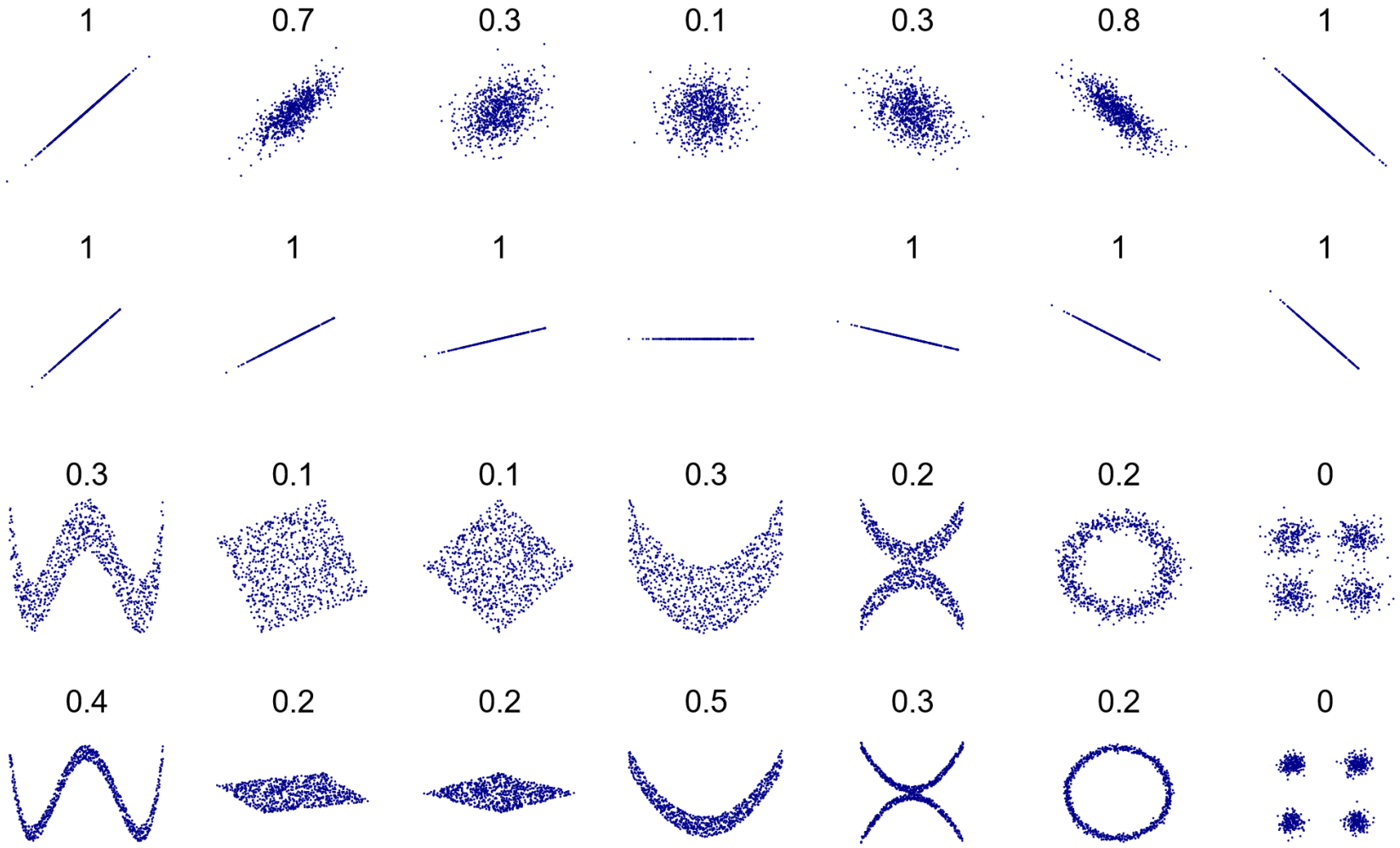
- Equivalence to existence of a Hilbert space embedding

  $\phi : X \to H$ with $||\phi(x) - \phi(y)||^2 = d(x, y)$.

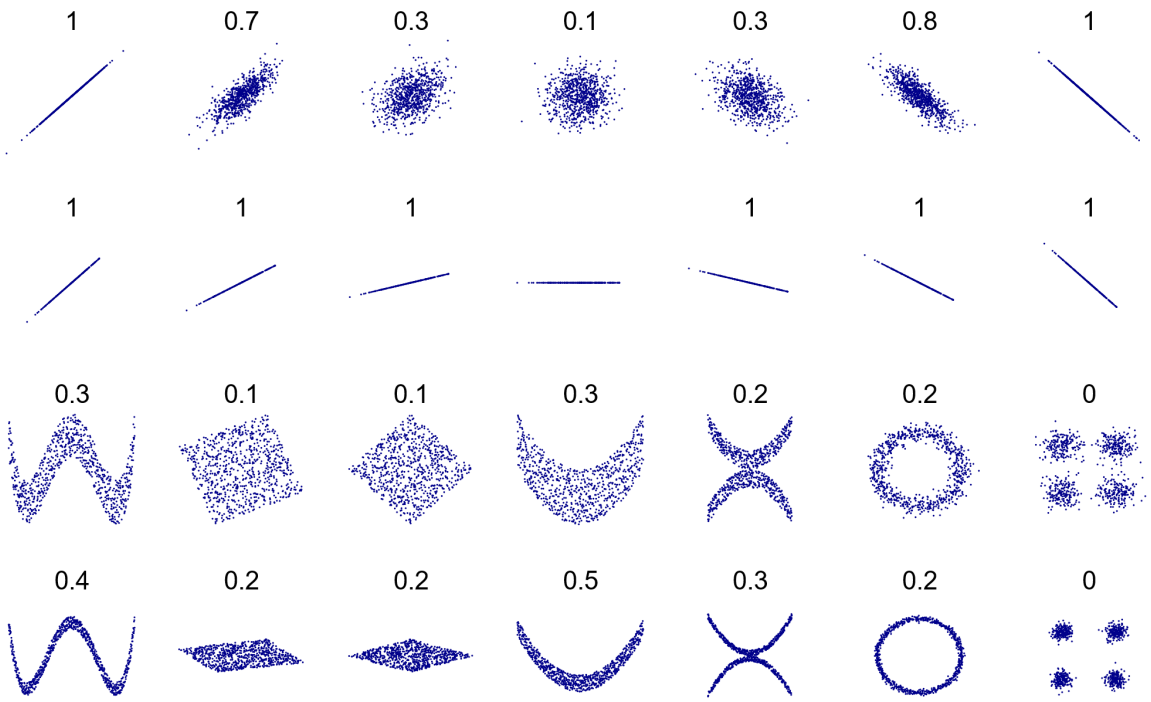
- Every separable Hilbert space is of strong negative type

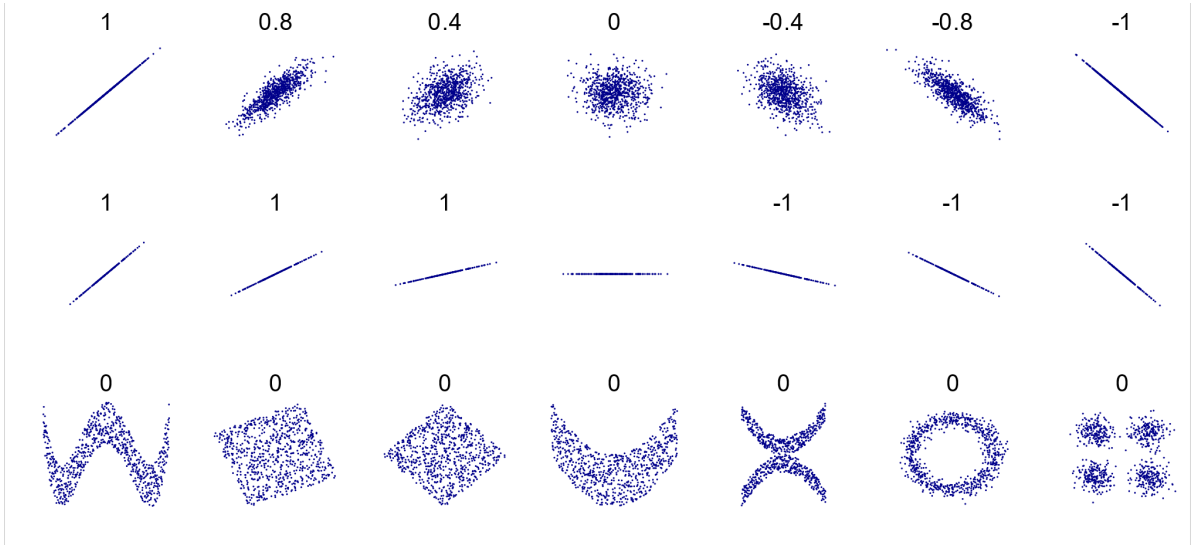
  In particular, Euclidean spaces are of strong negative type

Example Distance Correlations



Source:
Wikipedia





Distance Correlation VS Pearson Correlation

Which TDA summaries are of Strong Negative Type?

Yes

L^1 and L^2 versions of metric spaces for

- landscapes,
- rank functions,
- kernel maps, etc
- Betti and Euler curves

No

- $p = \infty$ functional topological summaries
- persistence diagrams with bottleneck
- persistence diagrams with any p-Wasserstein metric (shown by counterexamples)

NOTE: Strictly speaking L^1 versions are only of negative type rather than of strong negative type. But if (X, d) is a metric space of negative type, then (X, d^r) has strong negative type for all $r \in (0, 1)$;

Getting Meta

Using the SAME persistent homology but in DIFFERENT metric spaces

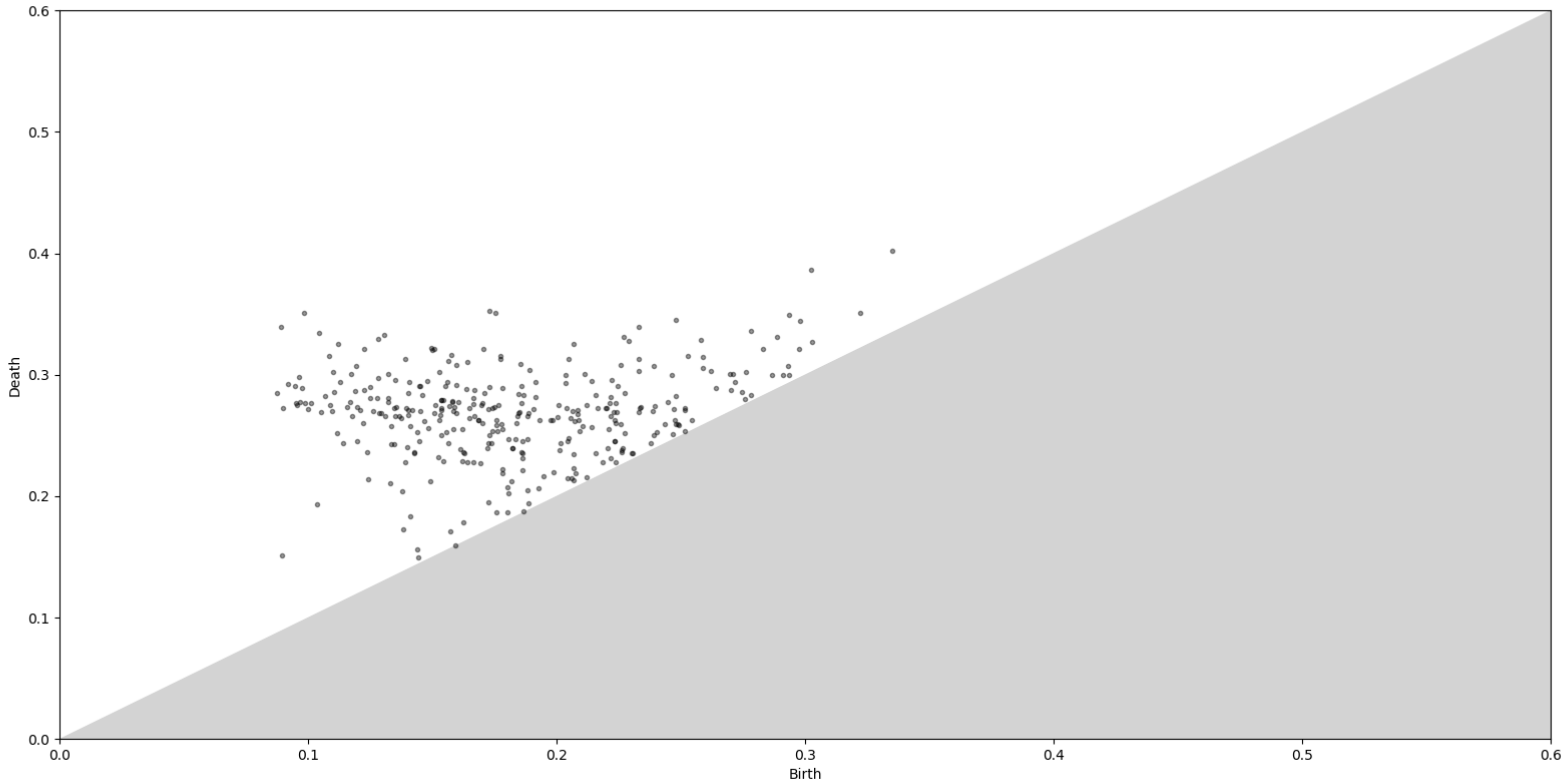
A veritable zoo of topological summaries

- Persistence diagrams
  - -Bottleneck, p-Wasserstein
- Heat kernel maps
  - - L^p functional distances
  - - Different bandwidths
- Persistence landscape distance
  - – L^p metrics
- sliced Wasserstein kernel
  - -different bandwidths

- L^p distance of the Betti curves
- L^p distance of the Euler curves
- L^p distance of the number of k-simplices

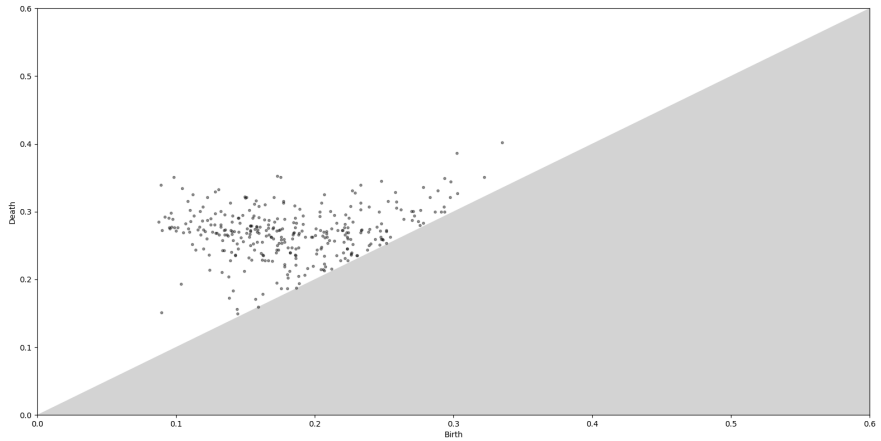
Erdos-Renyi

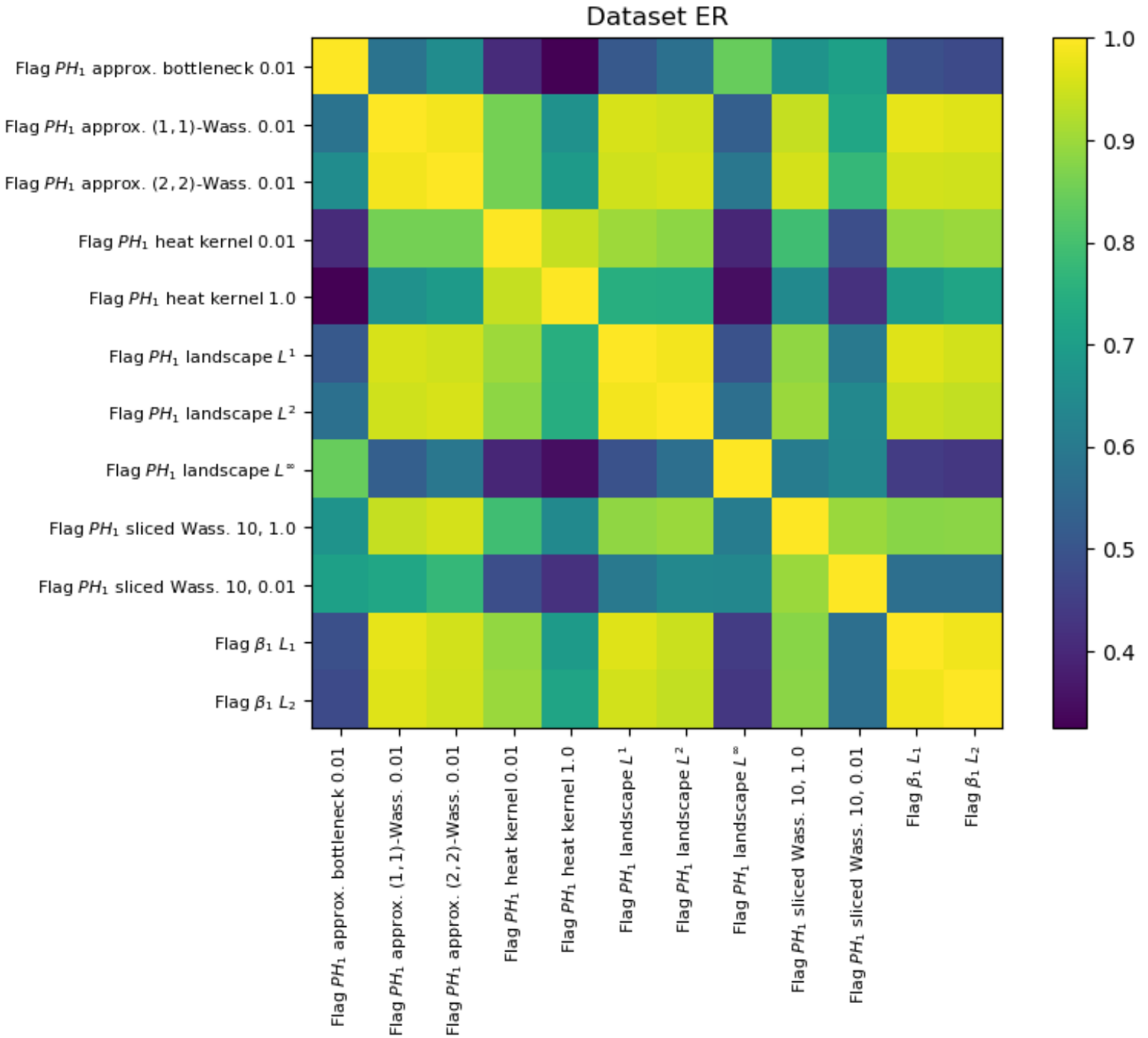
- density-0.5 random graphs on 100 vertices.
- Complex built: flag.
- 100 datasets.



Erdos-Renyi

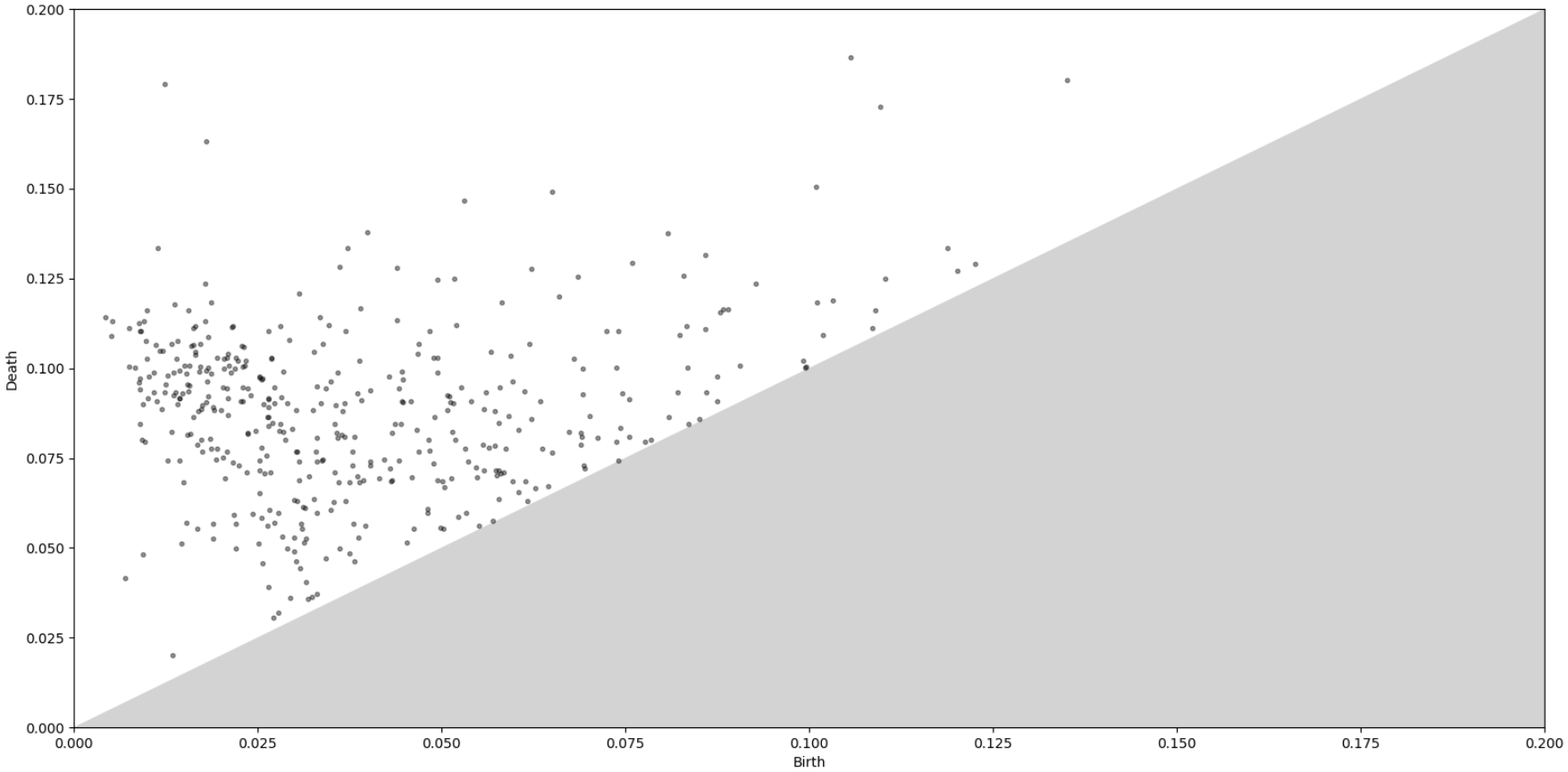
- density-0.5 random graphs on 100 vertices.
- Complex built: flag.
- 100 datasets.





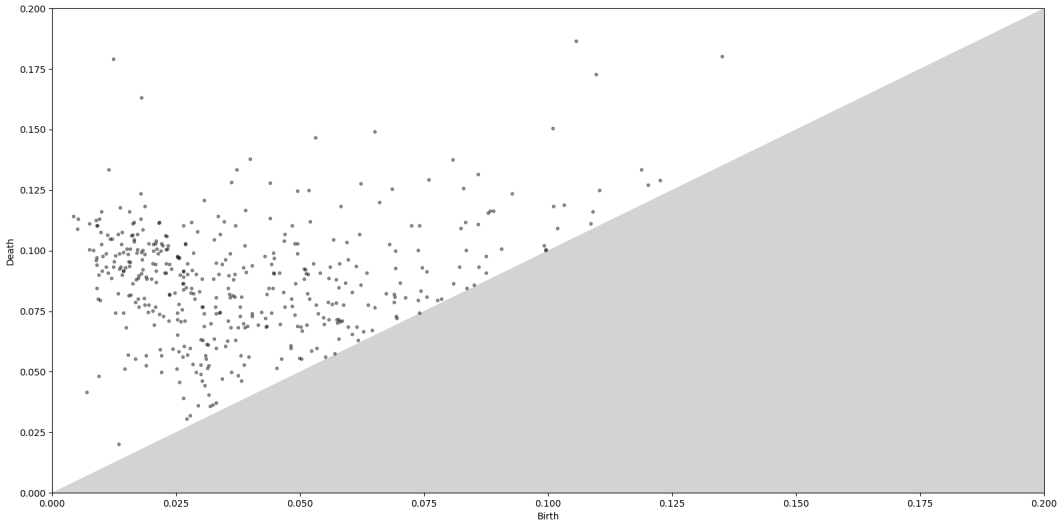
Directed Erdos-Renyi

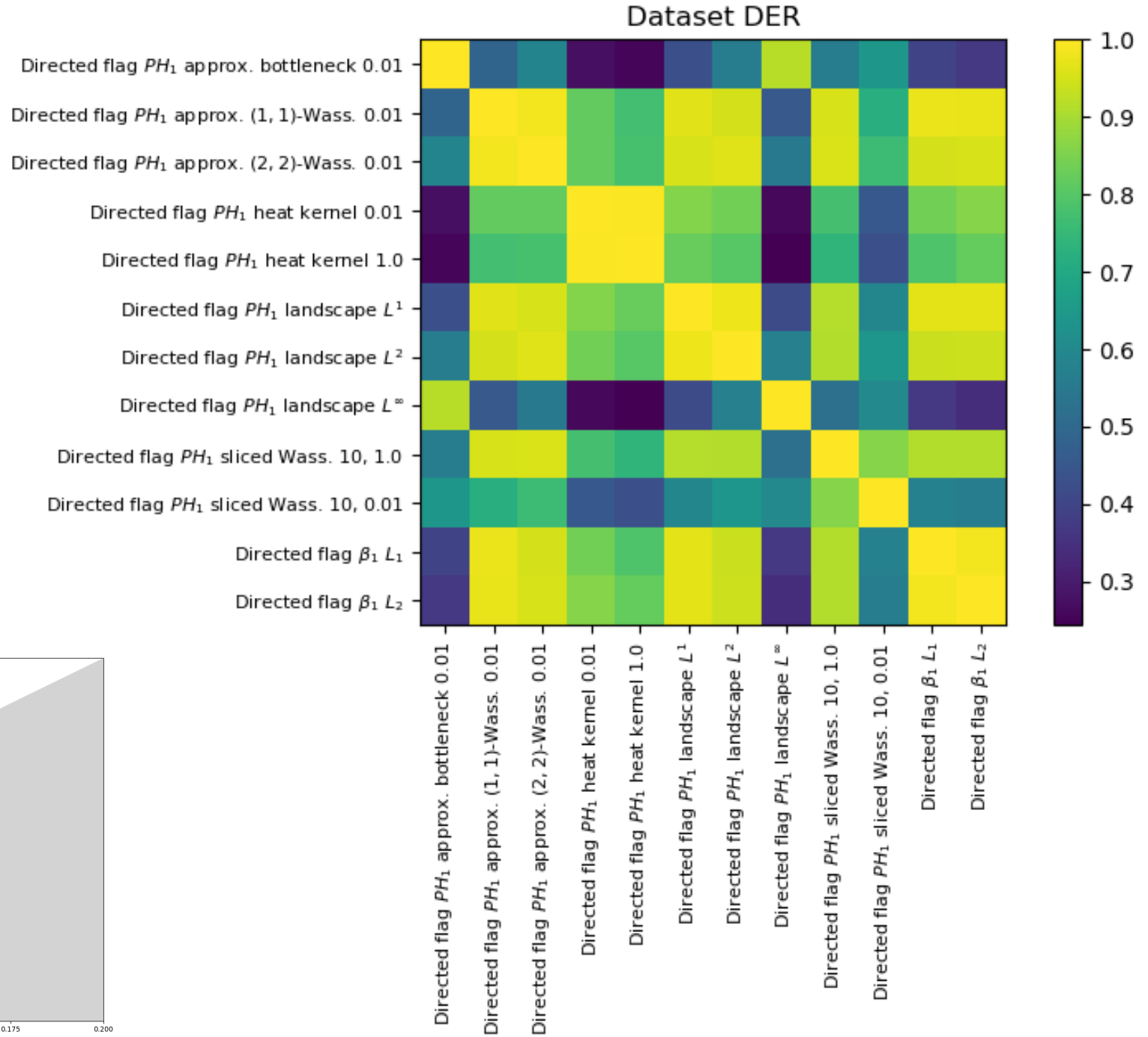
- density-0.5 random digraphs on 100 vertices.
- Complex built: dflag.
- 100 datasets.



Directed Erdos-Renyi

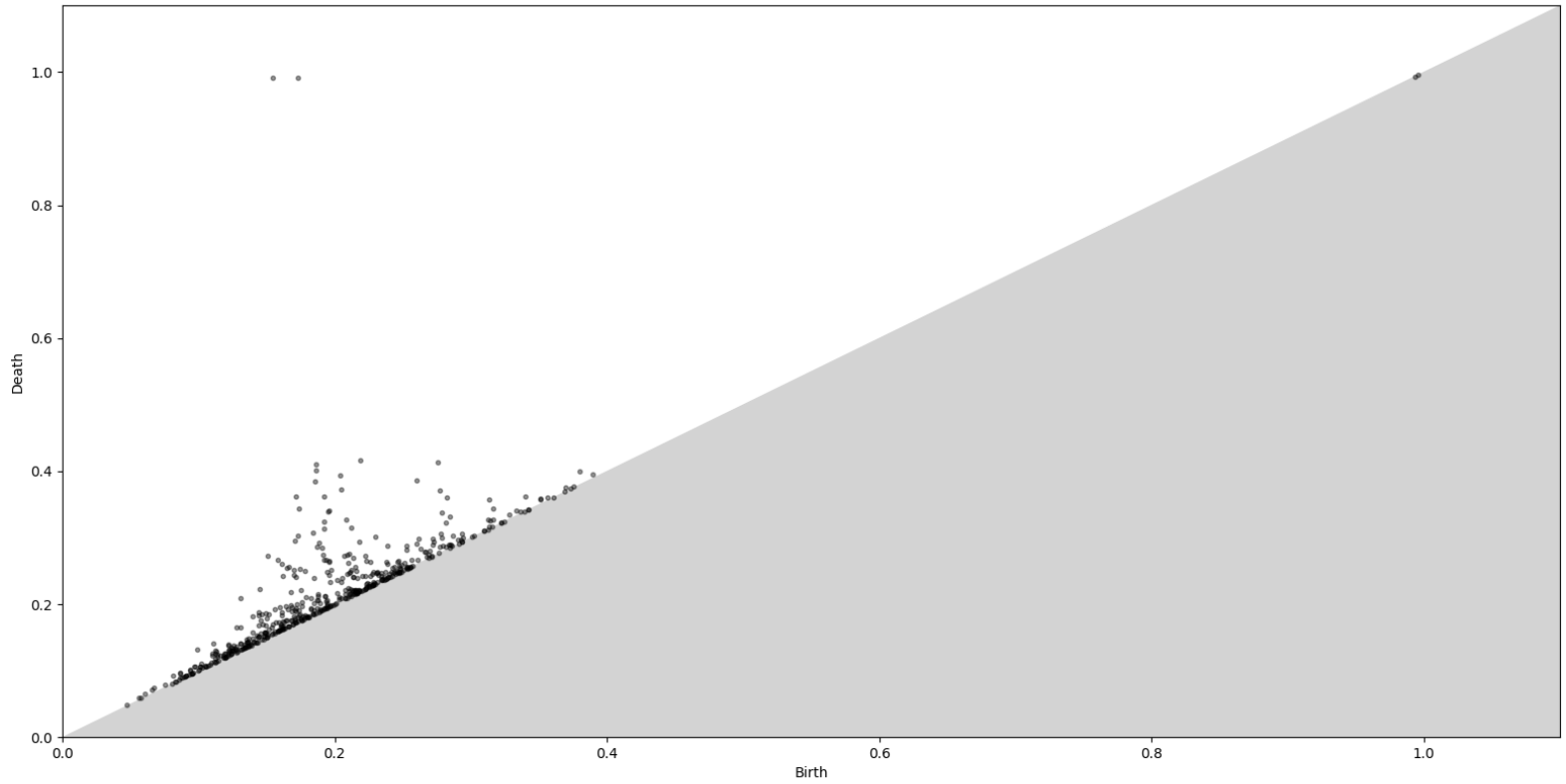
- density-0.5 random digraphs on 100 vertices.
- Complex built: dflag.
- 100 datasets.





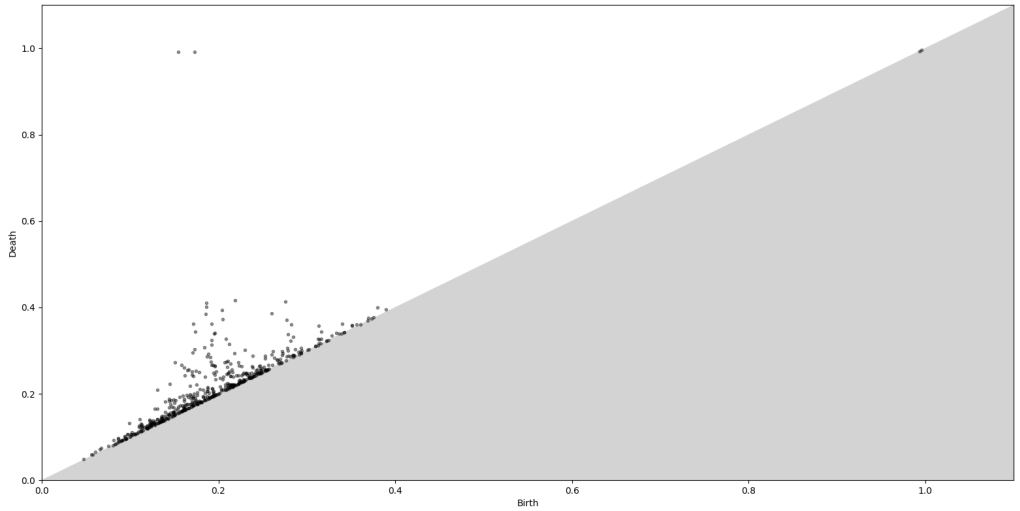
Geometric on a Torus

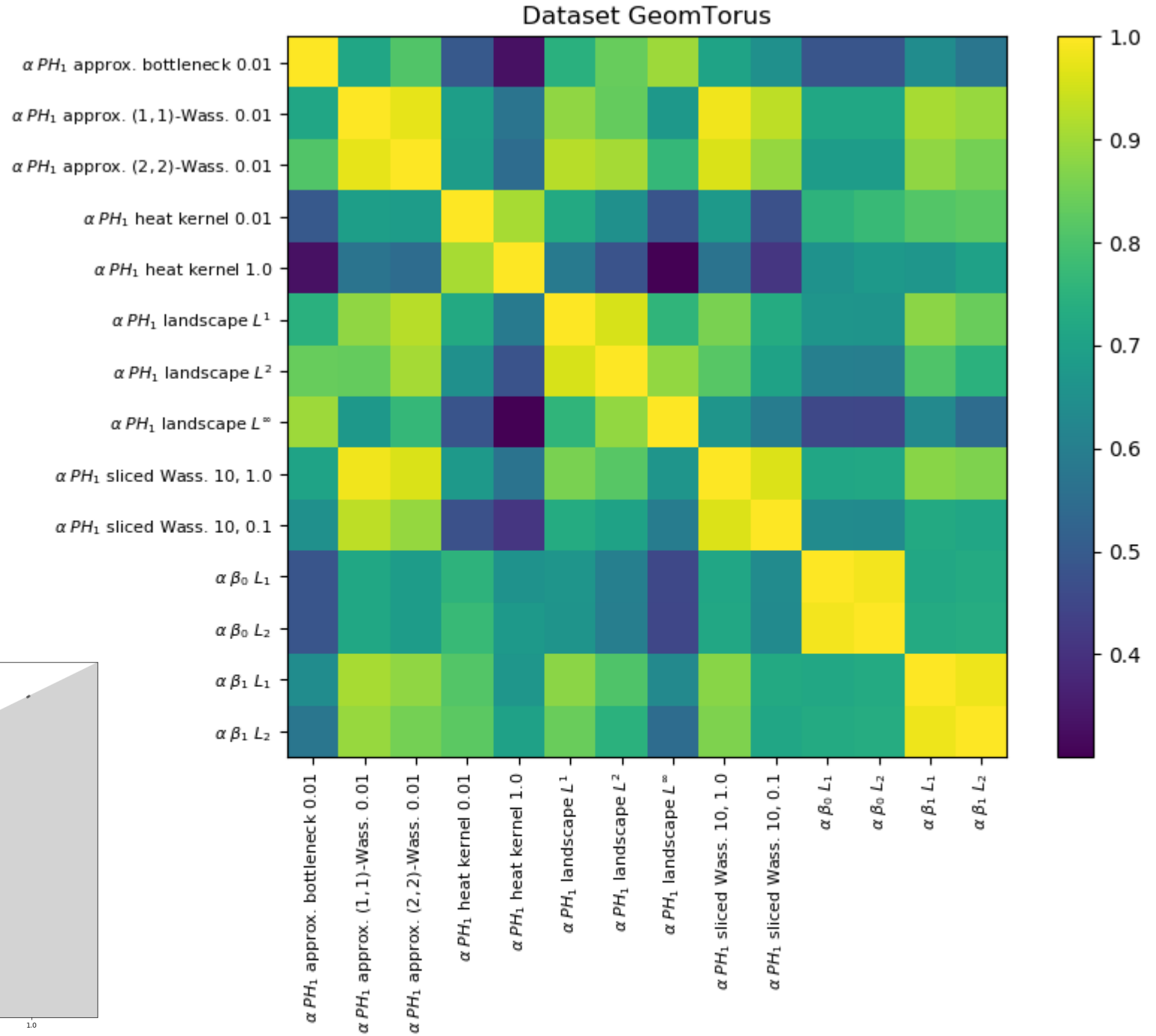
- Uniformly randomly sampled 100 points from a noisy torus in R^4.
- Complex built: alpha
- 100 datasets.



Geometric on a Torus

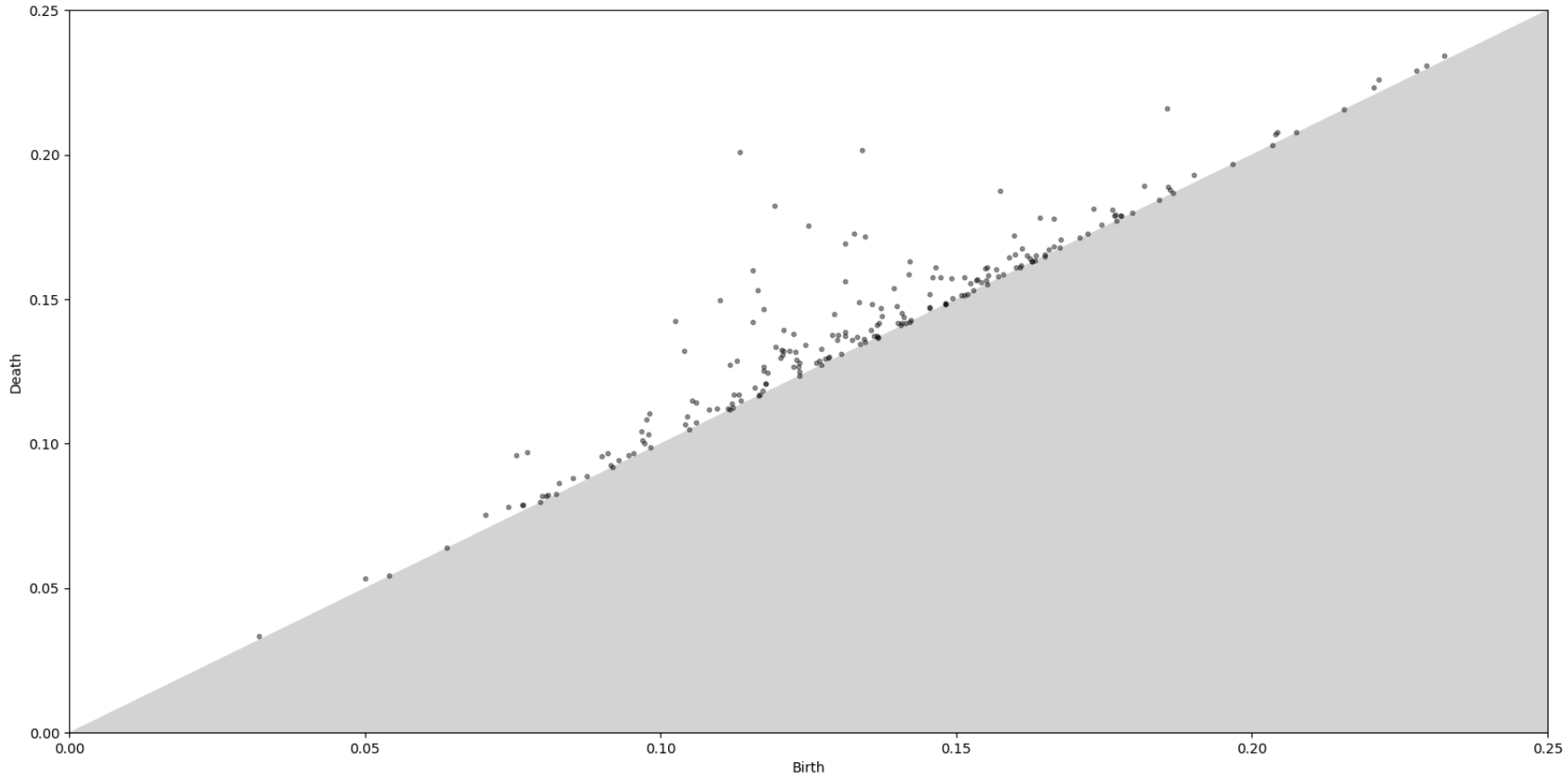
- Uniformly randomly sampled 100 points from a noisy torus in R^4.
- Complex built: alpha
- 100 datasets.





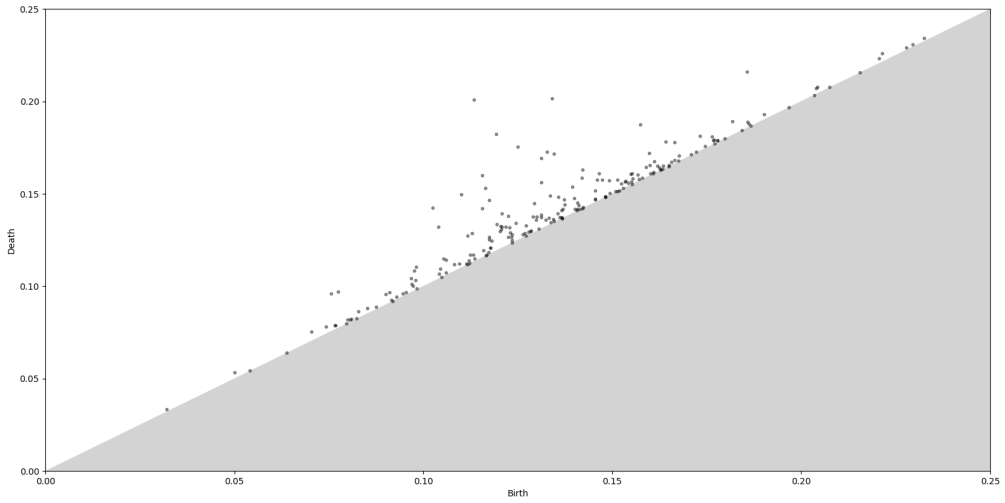
Random Geometric

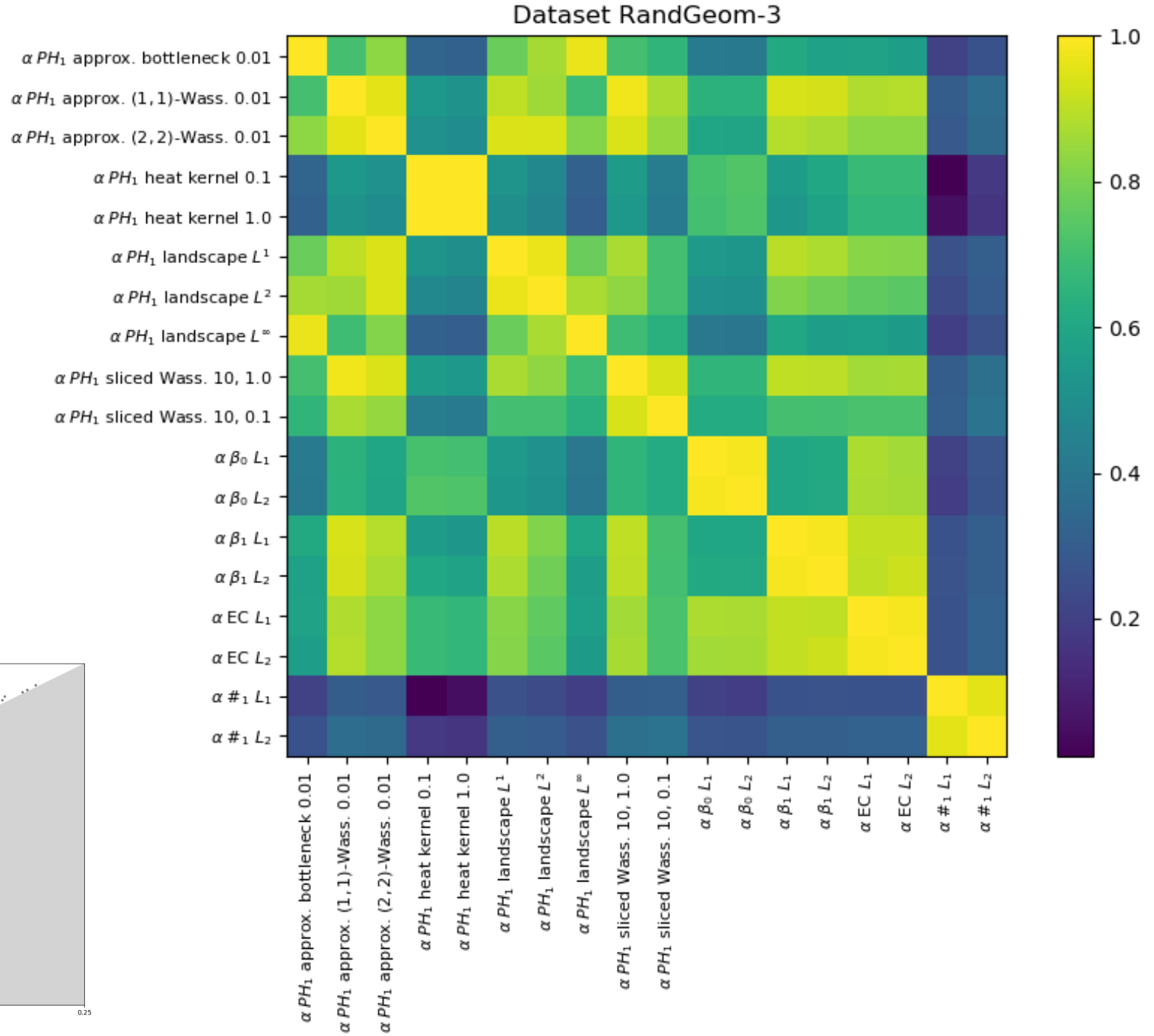
- Uniformly randomly sampled 100 points unit cube in R^3
- Complex built: alpha
- 100 datasets.



Random Geometric

- Uniformly randomly sampled 100 points unit cube in R^3
- Complex built: alpha
- 100 datasets.





Erdos-Renyi to Geometric

Fix a parameter p in [0,1].

Draw 100 random points uniformly from the 3D unit cube.

Let X be complete weighted graph with weights the Euclidean distances.

Let Y be the complete random graph with random weights.

Construct complete weighted graph Z by:

For each pair (i,j),

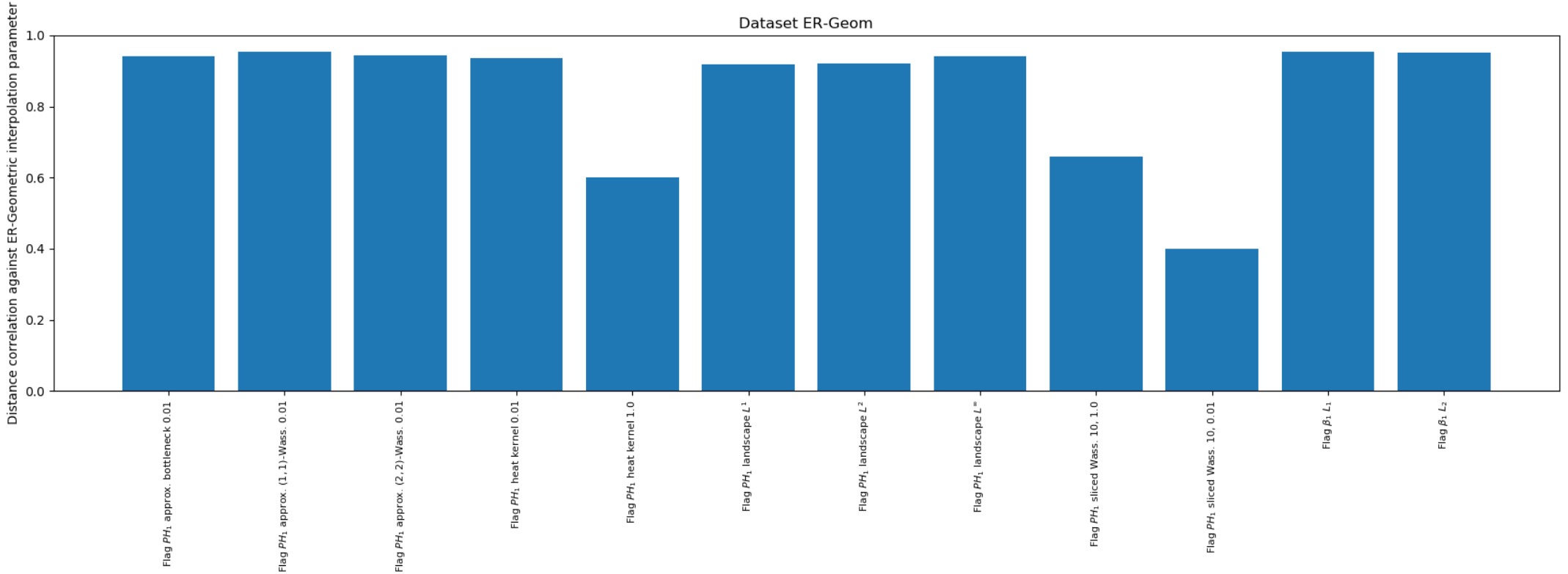
$$Z_{i,j} = \begin{cases} X_{i,j} \text{ with probability } p \\ Y_{i,j} \text{ with probability } 1-p \end{cases}$$

The model ranges from “fully ER” (p=0) to “fully geometric” (p=1).

Geometric to Erdos-Renyi

Simulated varying p from 0 to 1 (inclusive) in 100 steps

Distance correlation between the topological summaries and the value of interpolation parameter



Maps of Norway

Correlations between PH of topographic data with other variables

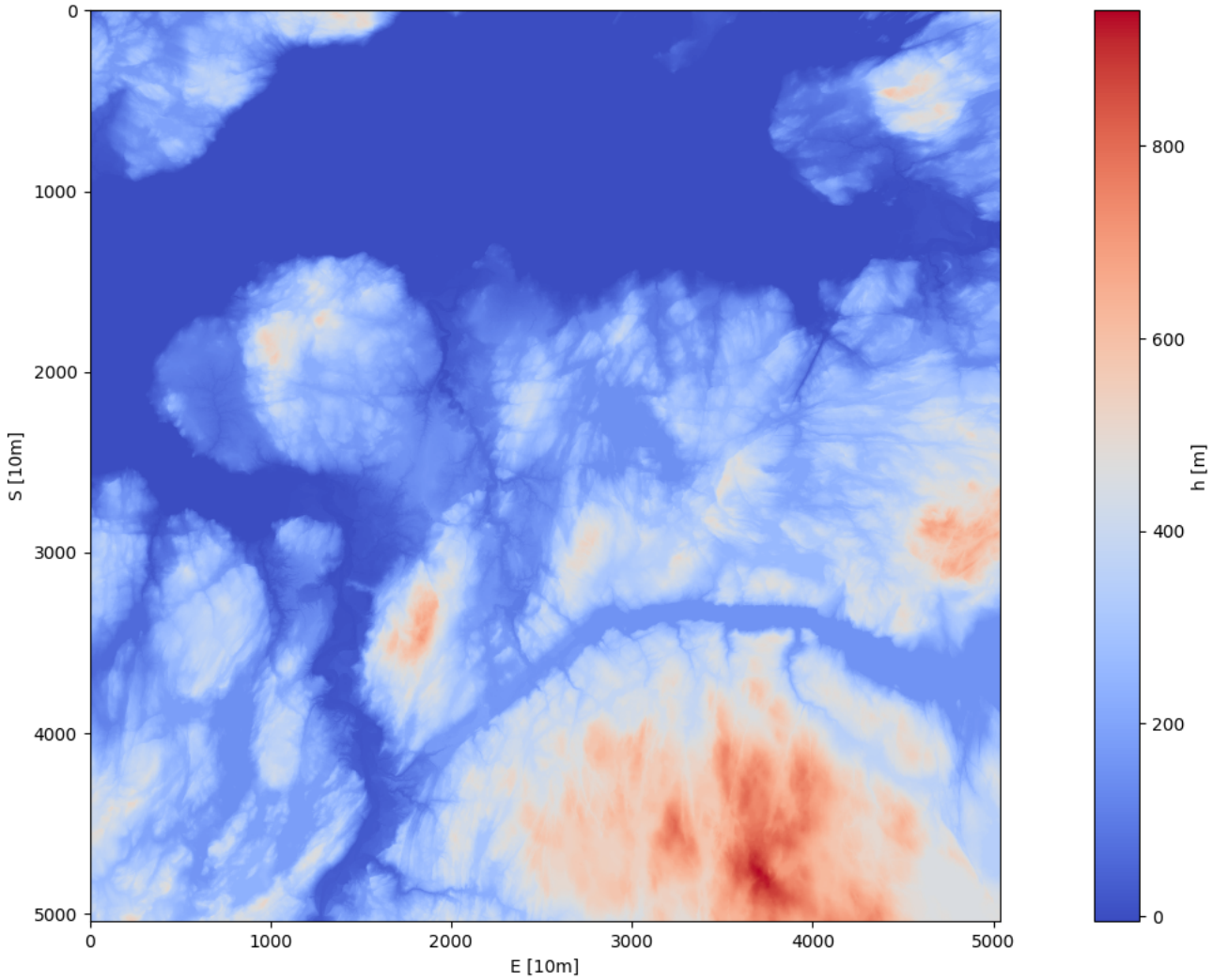
Digital Elevation Model

These are 50 km by 50 km patches of elevation data with a horizontal resolution of 10 m, vertical resolution of about 1 m.

Mapping data courtesy

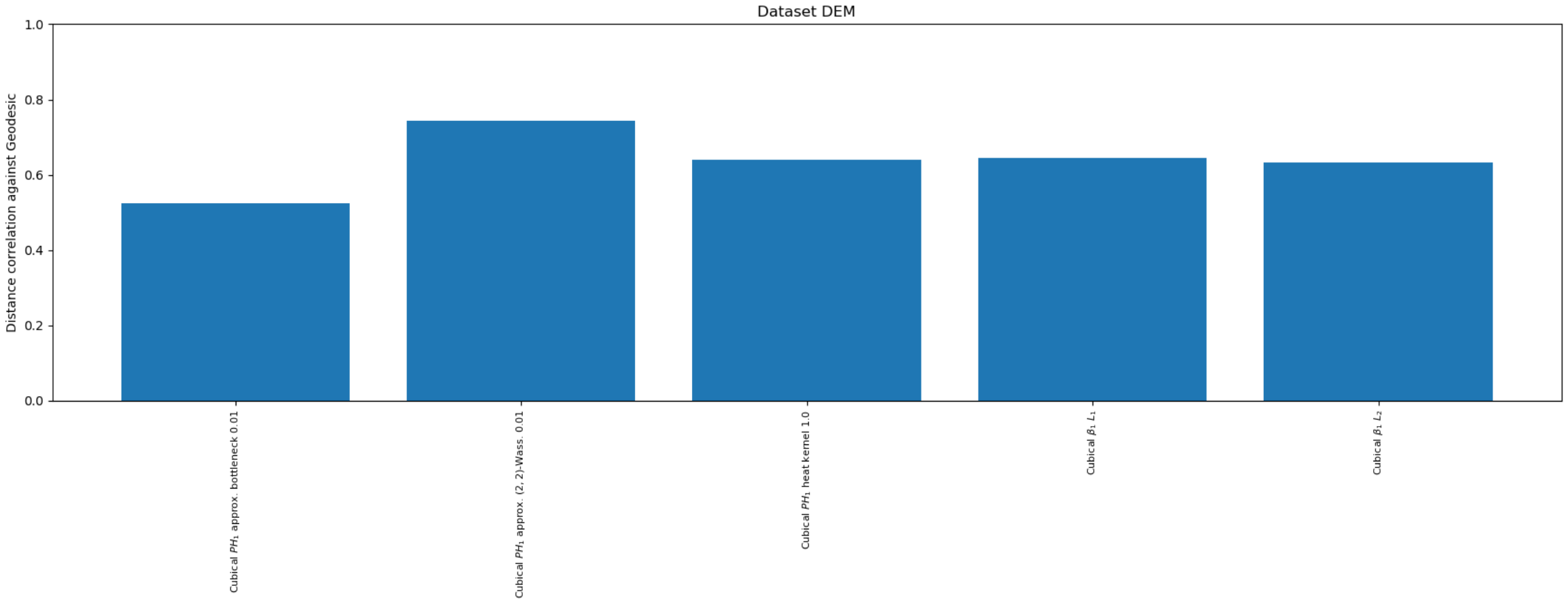
of the Norwegian Mapping Authority.

Compute persistent homology with respect to the height function



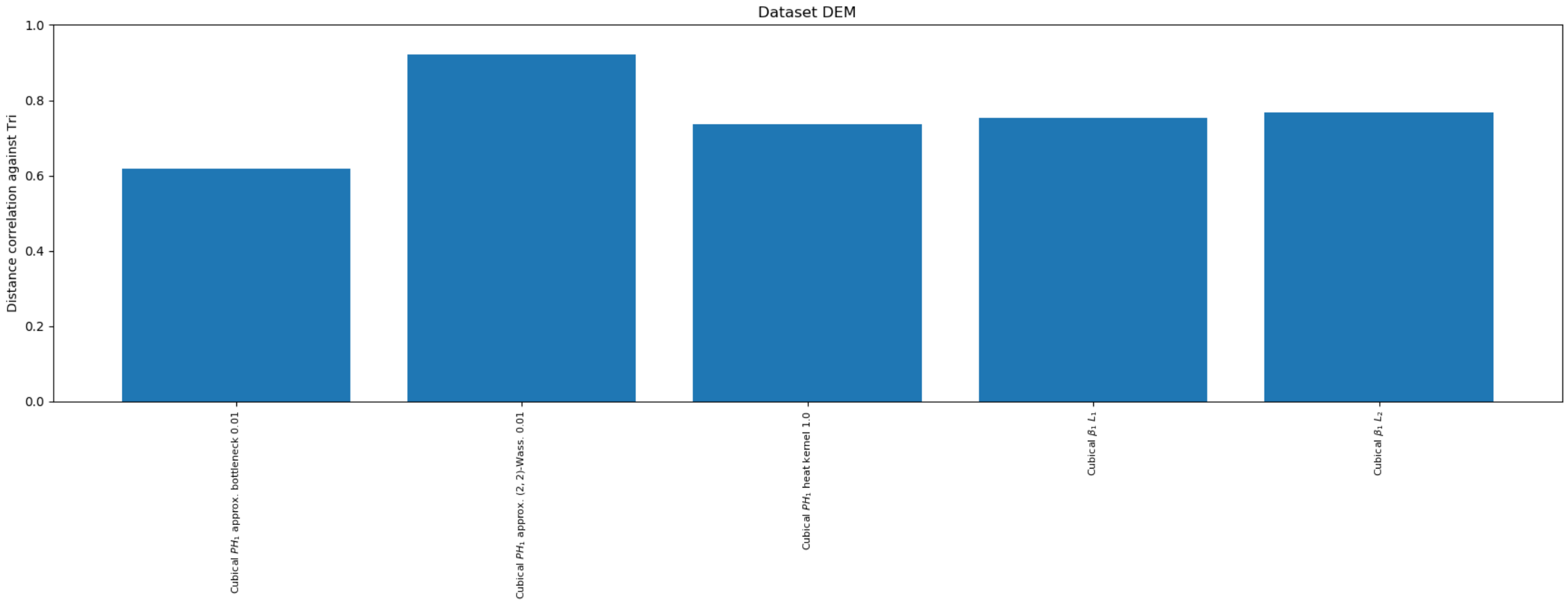
Digital Elevation Model: distance correlation with geodesic distance

Other variable is the geodesic distance between the centres of the patches.



Digital Elevation Model: distance correlation with Terrain Roughness Indicator

TRI = Terrain Roughness Indicator, a very simple real-valued model (one of many) geographers and geologists use for terrain roughness.



Software/data attributions:

- Directed flag complex PH: Flagser by Daniel Lütgehetmann.
- Flag complex PH: Ripser by Uli Bauer.
- Cubical complexes + their PH: GUDHI by INRIA
- Alpha complexes + their PH: GUDHI by INRIA
- Persistence landscapes: Persistence Landscape Toolbox by Paweł Dłotko
- Approximate Wasserstein distances: Hera by Arnur Nigmetov et al.
- Elevation data by Norwegian Mapping Authority.

Thanks for listening.