Topological Data Analysis for Digital Rock Modeling and beyond

Evgeny Burnaev

Prof., Director of AI center
Skoltech, AIRI

Skoltech
Skolkovo Institute of Science and Technology

AIRI

ARTIFICIAL
INTELLIGENCE
JOURNEY

Research Center in Artificial Intelligence
in the direction of optimization of management decisions to reduce carbon footprint (RAIC)

Strategic Mission: develop AI-based applied SW to estimate and minimize carbon footprint and ESG risks

Core team & structure

Evgeny Burnaev
Associate Professor
Director

Core team
- 7 Profs., 1 Dr. Sci.
- 50 researchers
- 30 PhD students

Research areas
- Data Fusion and 3D Computer Vision
- Physics-Informed ML
- Efficient DL for Green AI technologies
- ML for Industrial Predictive Analytics

Recent academic achievements

- >200 papers in AI and modeling (>20 Core A/A*) in 2018-2021
- SGP Best Dataset Award 2019
- ANNPR Best Paper Award 2020
- 3 Ilya Segalovich Yandex Awards 2018, 2019, 2020
- DLGC CVPR Workshops in 2020, 2021
- Science Award of Moscow Government in 2018
- 5th IEEE Int. Conf. on Internet of People Best Paper Award in 2019
- Int. ML summer schools (MLSS, SMILES) in 2019, 2020

Partners

- **Key industry partners**: Huawei, Sber, Gazpromneft, CityAir, Yandex
- **Potential partners**: financial institutions (e.g., Sber, VEB.RF, Gazprombank), governmental bodies
- **Academic partners**: Tech. Univ. Munchen, Univ. of Hamburg, Univ. of Oxford, Institute for Applied Informatics (Germany), etc.

Innovation

- develop a software platform providing an access to frameworks for
  - ✓ Data Fusion,
  - ✓ Physics-Informed ML and
  - ✓ Green AI
- develop prototypes of AI based products for **multi-scale monitoring and control of ESG risks to optimize management decisions and reduce carbon footprint**
- deliver prototypes to the industrial companies and startups to **support Russia National Strategy in AI** and **Russia Energy Strategy**
- IP generation

Prototypes to deliver

- monitoring of a carbon footprint
- assessment of atmospheric air quality and calculations of atmospheric transport processes
- optimization of management decisions in the field of oil production
- analysis of physical and financial risks due to climate changes
- acceleration of learning and compression of large neural networks

Industrial Expertise: since 2007

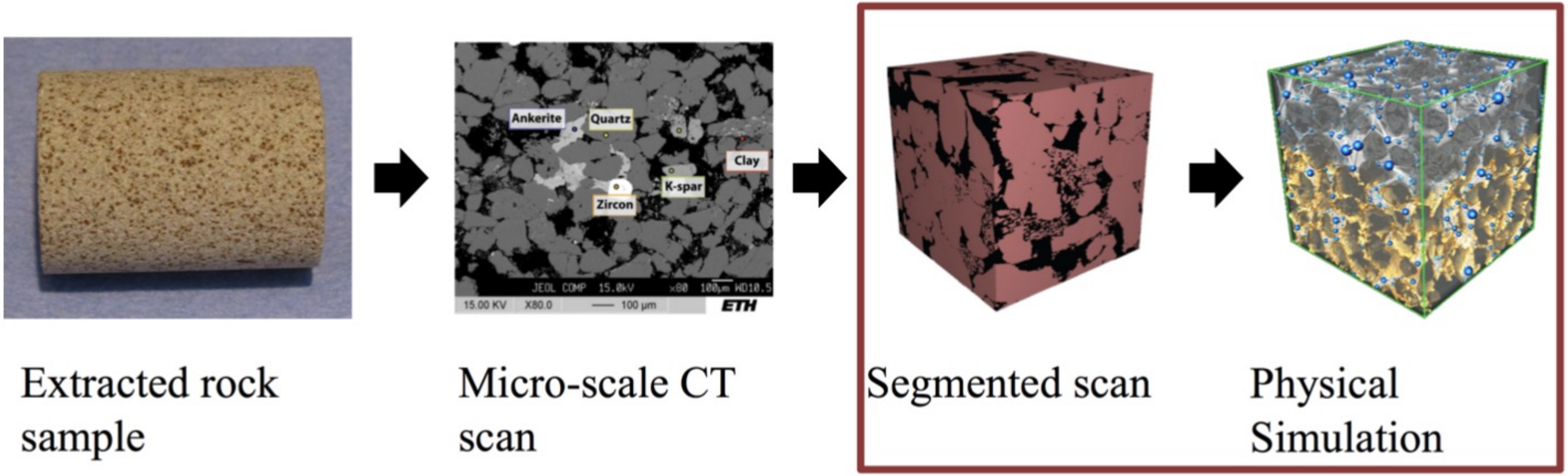
AIRBUS, LOUIS VUITTON, altran, SBERBANK, HUAWEI, AIRBUS HELICOPTERS, IPSEN Innovation for patient care, SAHARA Force India formula one team, PORSCHE, BOSCH, Alibaba Group 阿里巴巴集团

and many others...

Impact: RAIC has strong potential to become the worldwide recognizable center uniting the state-of-the-art expertise in AI and ML technologies for Industrial Analytics Applications. Such center will provide solutions to existing technological barriers in the industry based on fundamentally solid solutions, and provide elite education to future leaders in Industrially-oriented AI both in research and innovation.

Metamodeling of reservoir properties

- The paradigm of Digital Rock Physics is "Image-and-compute"
- We are working only with segmented scans and simulations



Input

→ 3D microstructural image of digitized core

Metamodeling of reservoir properties

We consider two problems

1. **Permeability prediction**
2. **Generation of artificial rock samples**

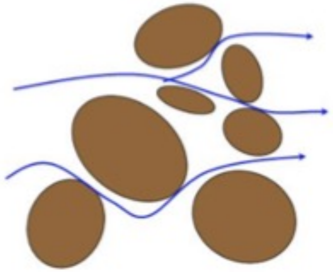
Practical advantages

→ Significant (up to 10.000 times) acceleration of permeability calculation on digitized samples

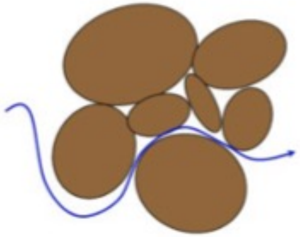
References:

- O. Sudakov, **E. Burnaev**, D. Koroteev. Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks. Computers and Geosciences, Volume 127, June 2019, Pages 91-98
- D. Volkhonskiy, E. Muravleva, O. Sudakov, D. Orlov, B. Belozerov, **E. Burnaev**, D. Koroteev. Generative Adversarial Networks for Reconstruction of 3D porous media from 2D slices. Physical Review E, 2022

- **Goal: permeability prediction with machine learning**
- Permeability (k) is a measure of the rock's ability to permit liquid to flow through its pores or voids

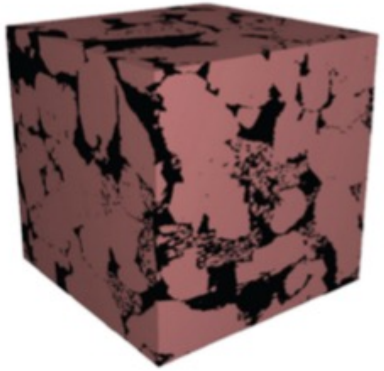


Loose structure - high κ

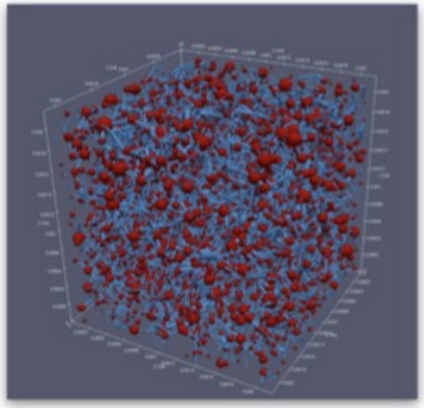


Dense structure - low κ

- OpenPNM, network model and Darcy's law and are used to compute k



Segmented scan



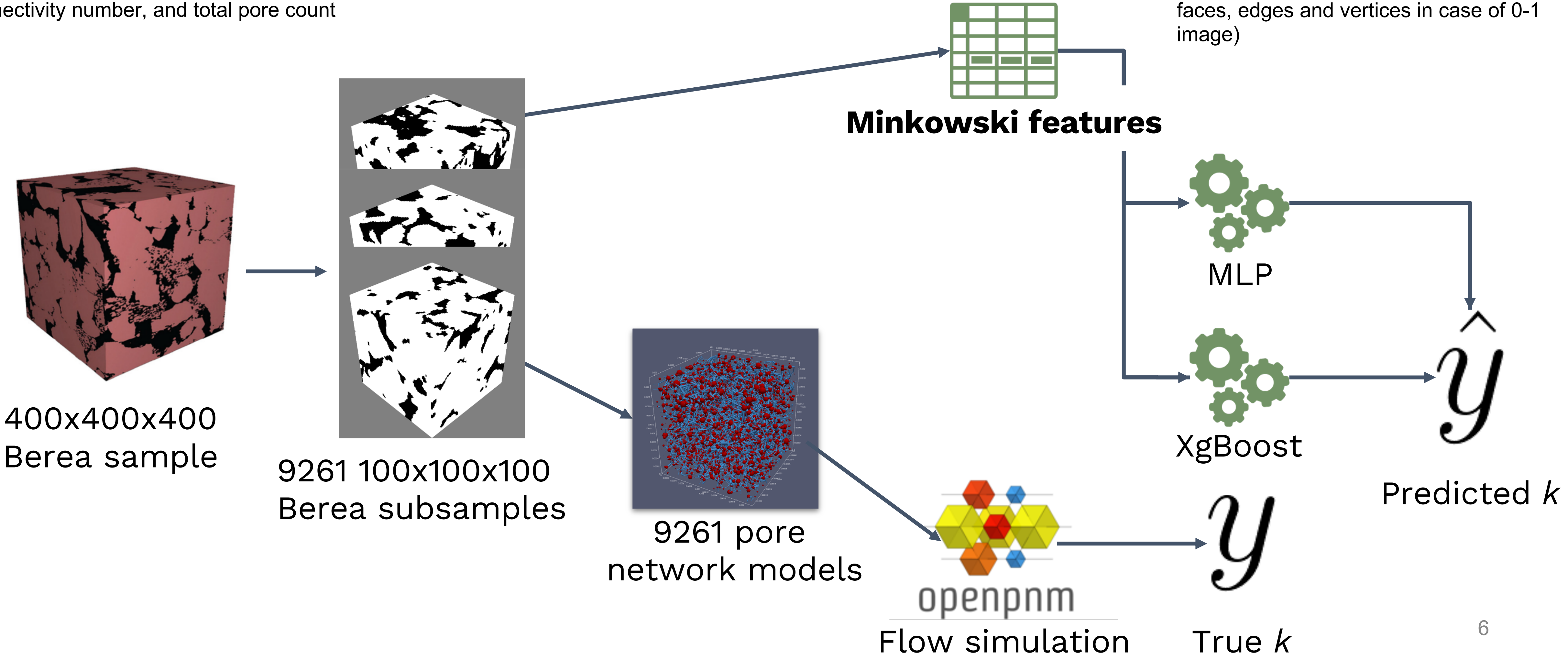
Pore-throat network model

κ value with Darcy's law

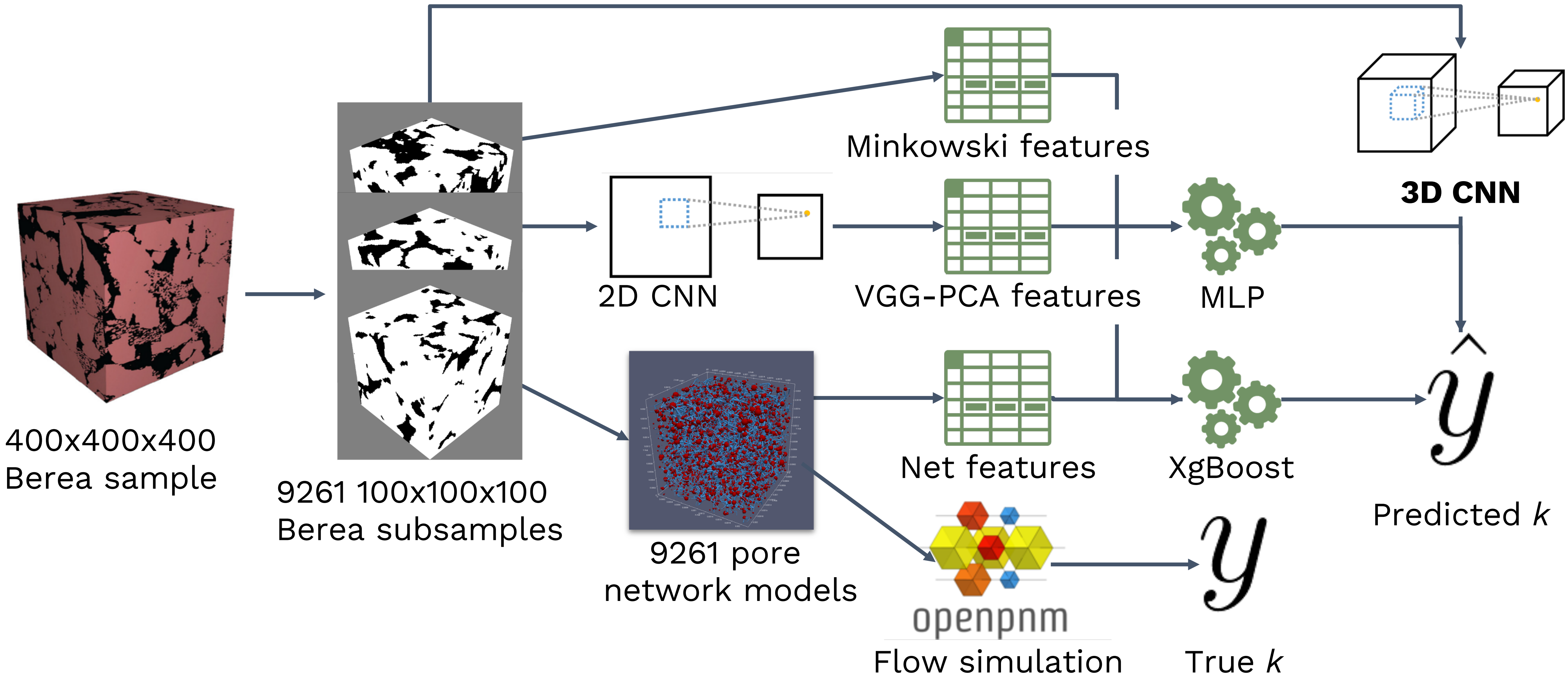
Previous permeability prediction approach

Network characteristics: median pore radius, mean pore radius, median throat radius, mean throat radius, median throat length, mean throat length, median pore connectivity number, mean pore connectivity number, and total pore count

Minkowski functionals: volume, area, mean breadth and the Euler-Poincar characteristic (enumeration of open voxels, faces, edges and vertices in case of 0-1 image)



Different feature generation pipelines

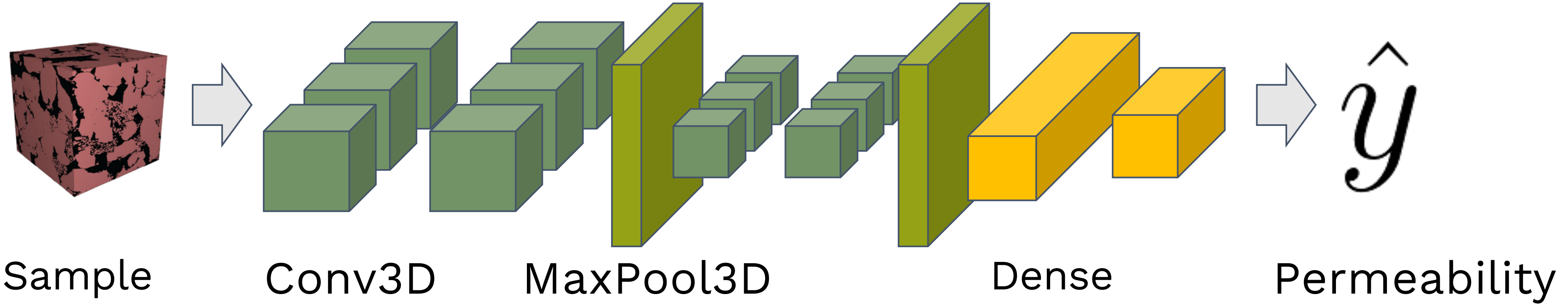


Obtained results

ABSq metric was used to provide interpretable evaluation

$$ABS_q = \frac{\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i|}{P_{99}(y) - P_1(y)}$$

Approach	ABSq
MF	*0.0396*
MF ALL	0.0370
NET	0.0372
VGG-PCA	**0.0287**
3D CNN	**0.0284**



Is this the end?

No!

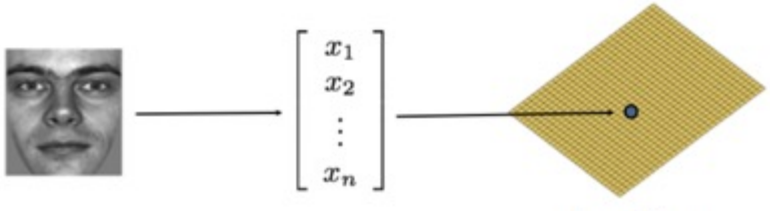
Further challenges

- **We would like to have more tractable and accurate models for prediction of transport properties of the rock**
- Validation on experimental data
- Coupling of simulations and experiments; fine-tuning of simulation models on experimental data

Topology to the rescue!

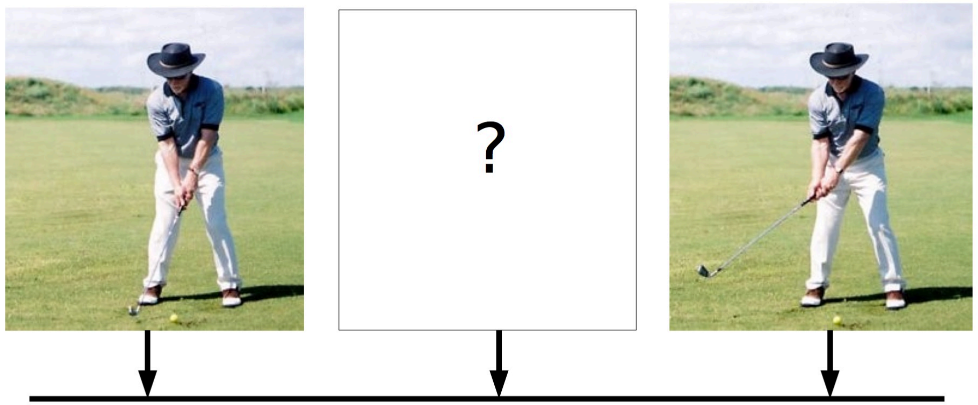
Manifold learning – Data Analysis technology based on **geometrical model** about high-dimensional data

A. **The world is multidimensional**
B. **Multidimensional data are difficult to use**
C. **Real-world data have low-dimensional structure**
D. **The world is not flat (nonlinear)**



1024×1024: $d \approx 10^6$

The world is not flat (nonlinear)



Linear interpolation

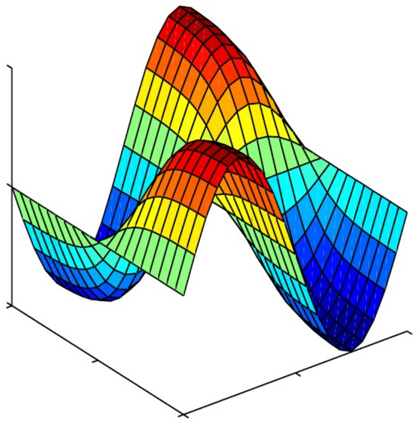
Nonlinear interpolation

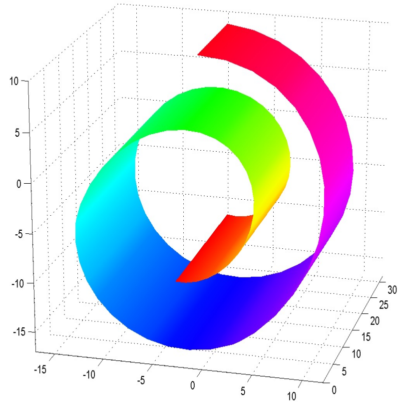
Manifold covered by a single chart (surface in $\mathbb{R}^d$)

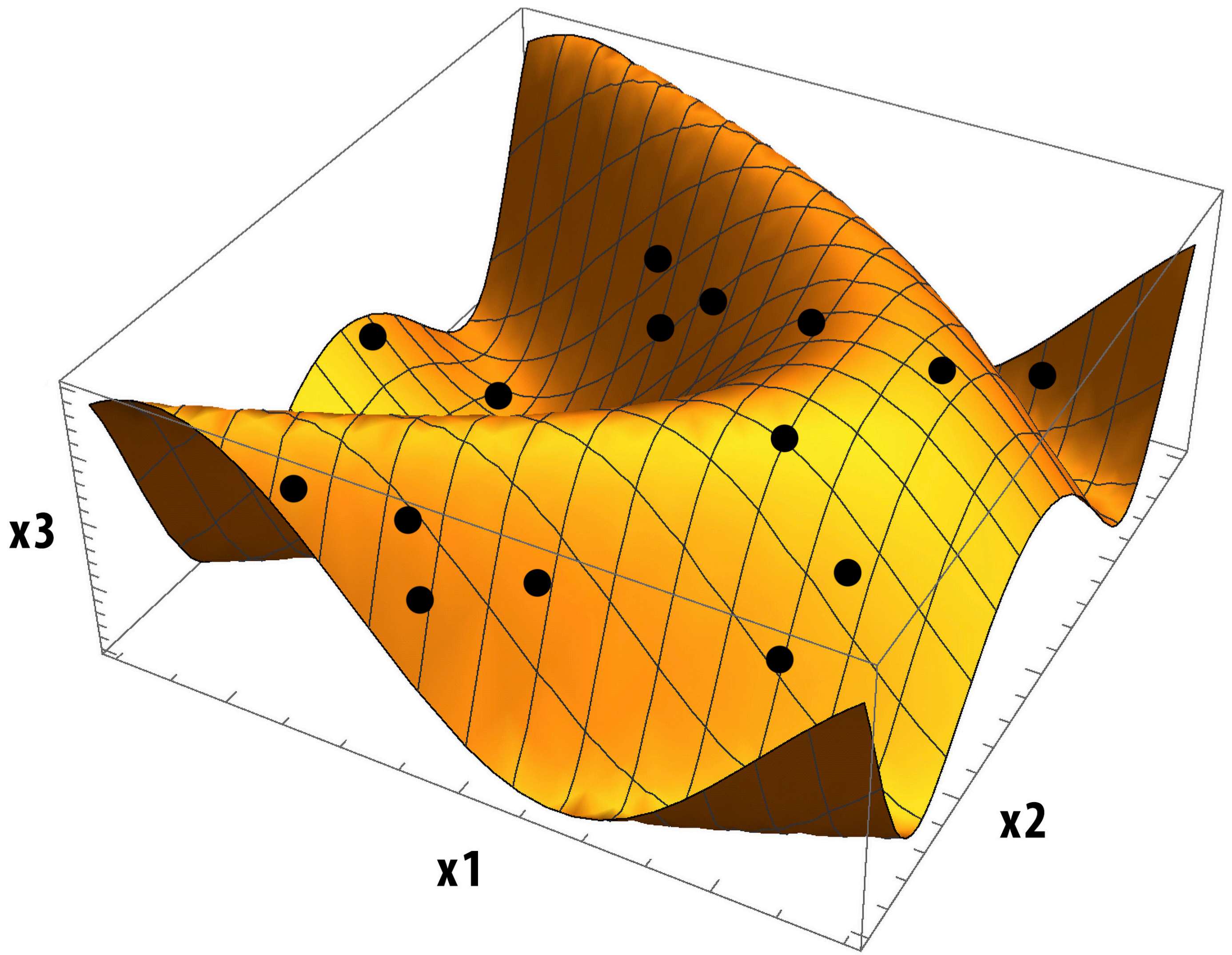
$$\mathbf{M} = \{x = g(z) \in \mathbb{R}^d : \ z \in \mathbf{Z} \subset \mathbb{R}^s\}$$

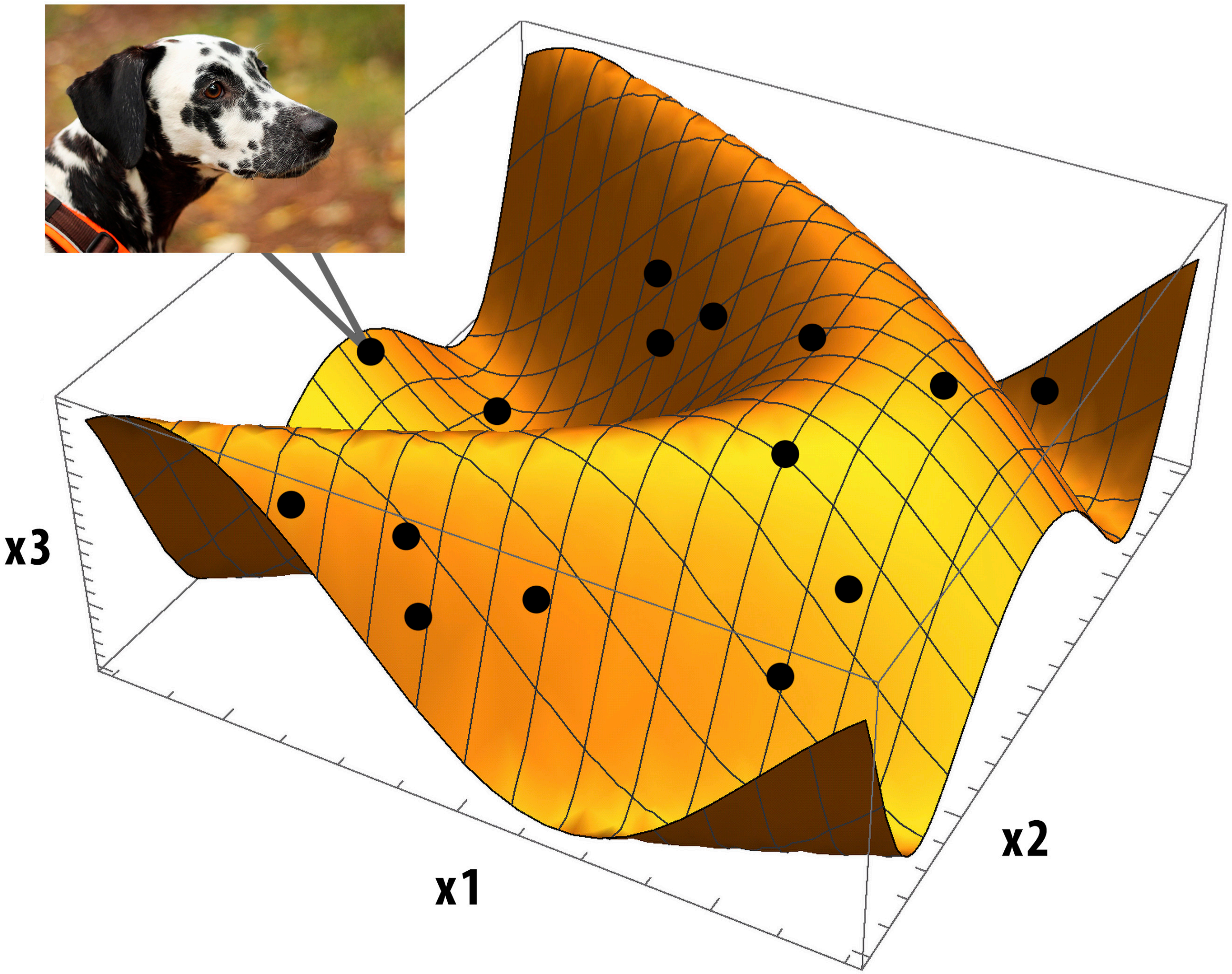
unknown s -dimensional surface – **Data manifold**

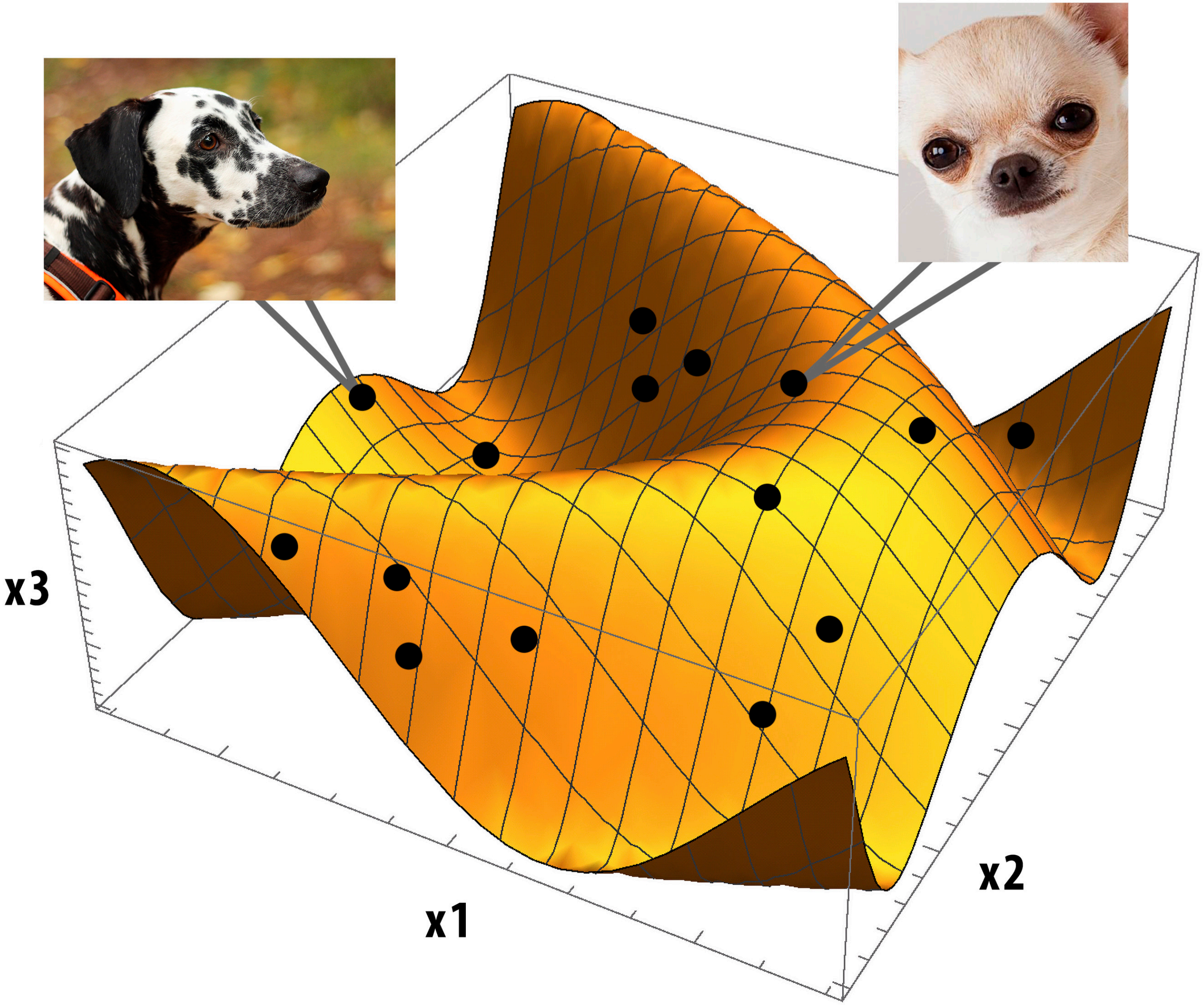
covered by **single chart** g defined on **Coordinate space** $\mathbf{Z} \subset \mathbb{R}^s$



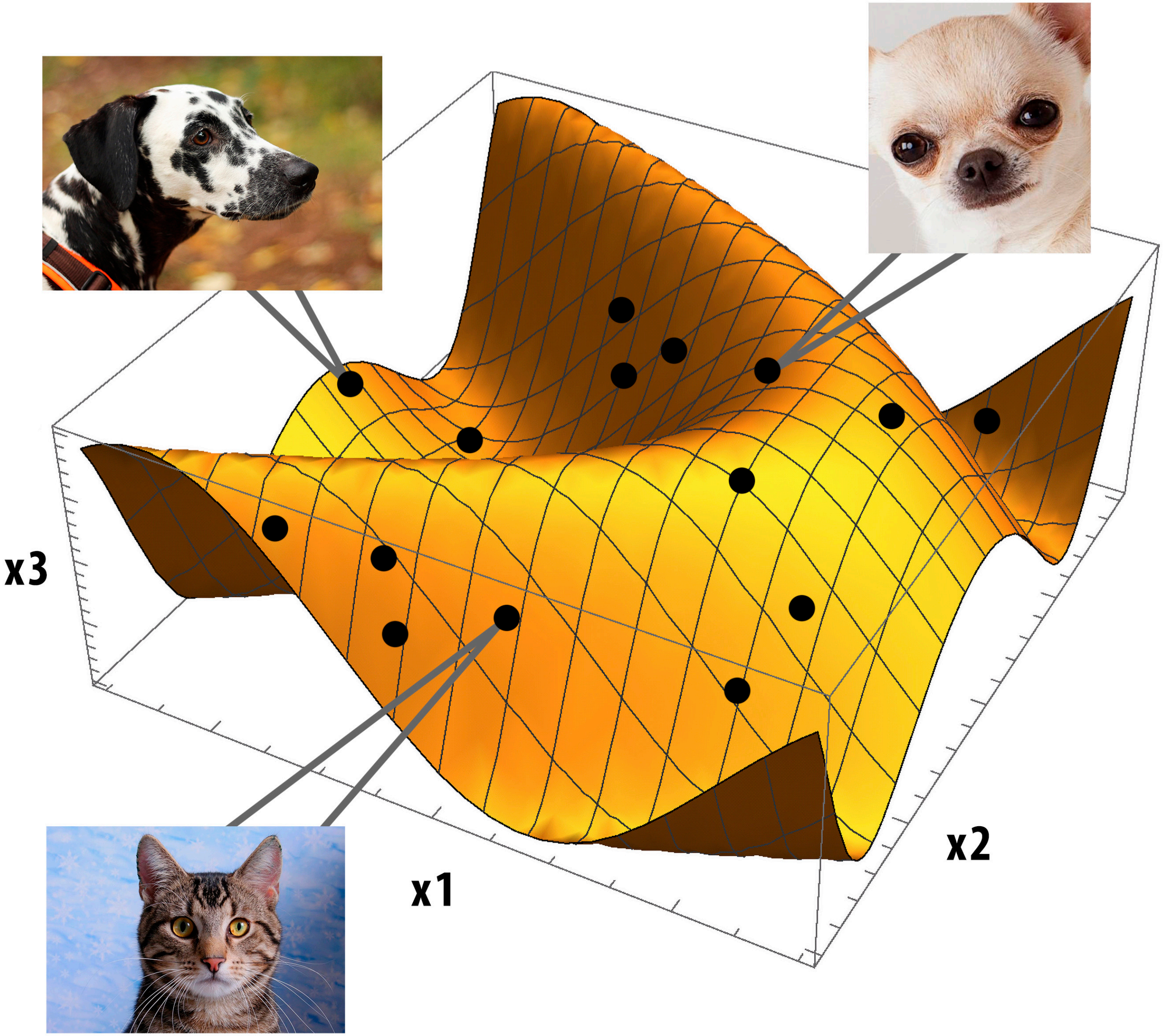


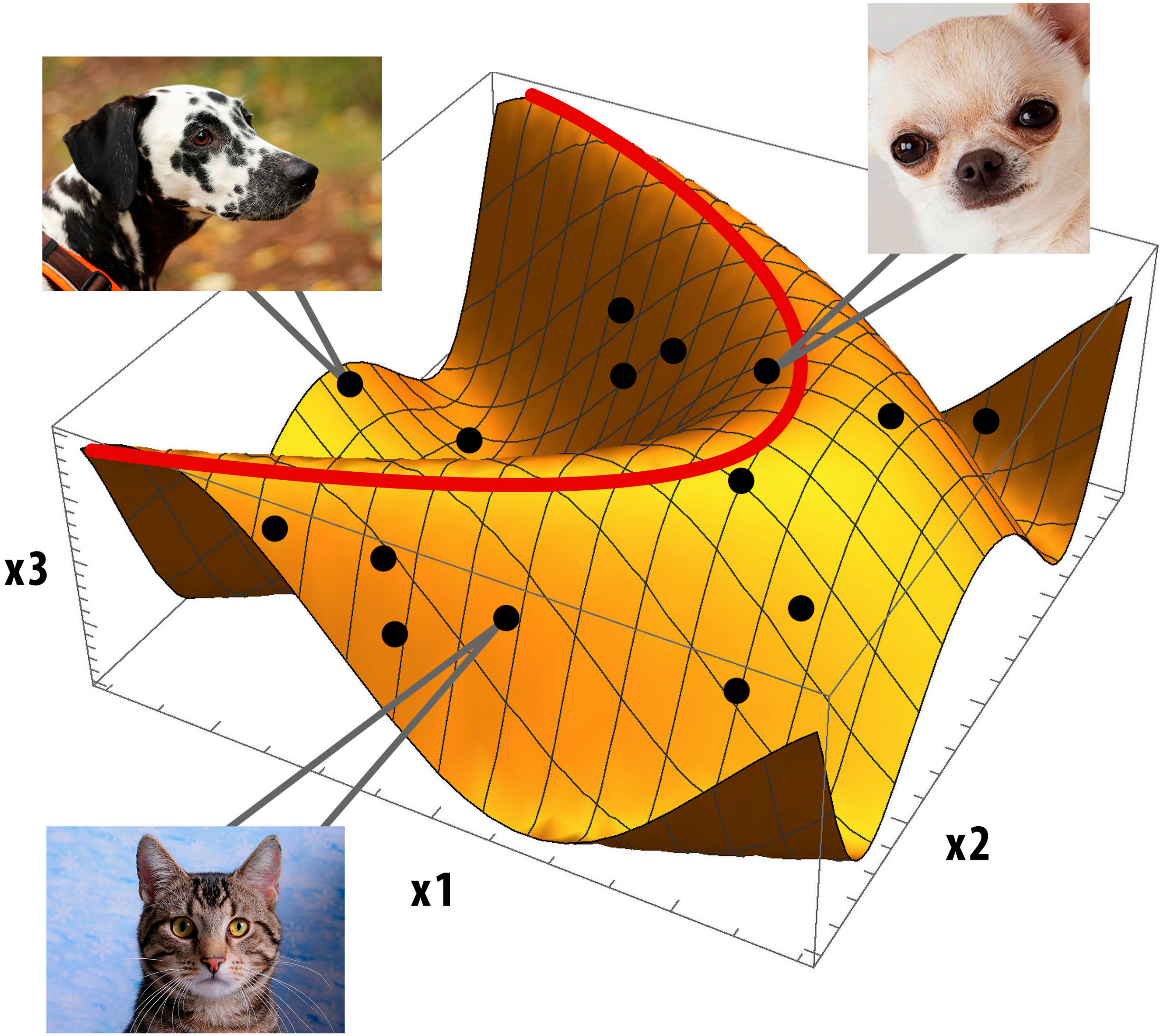

x3
x1
x2

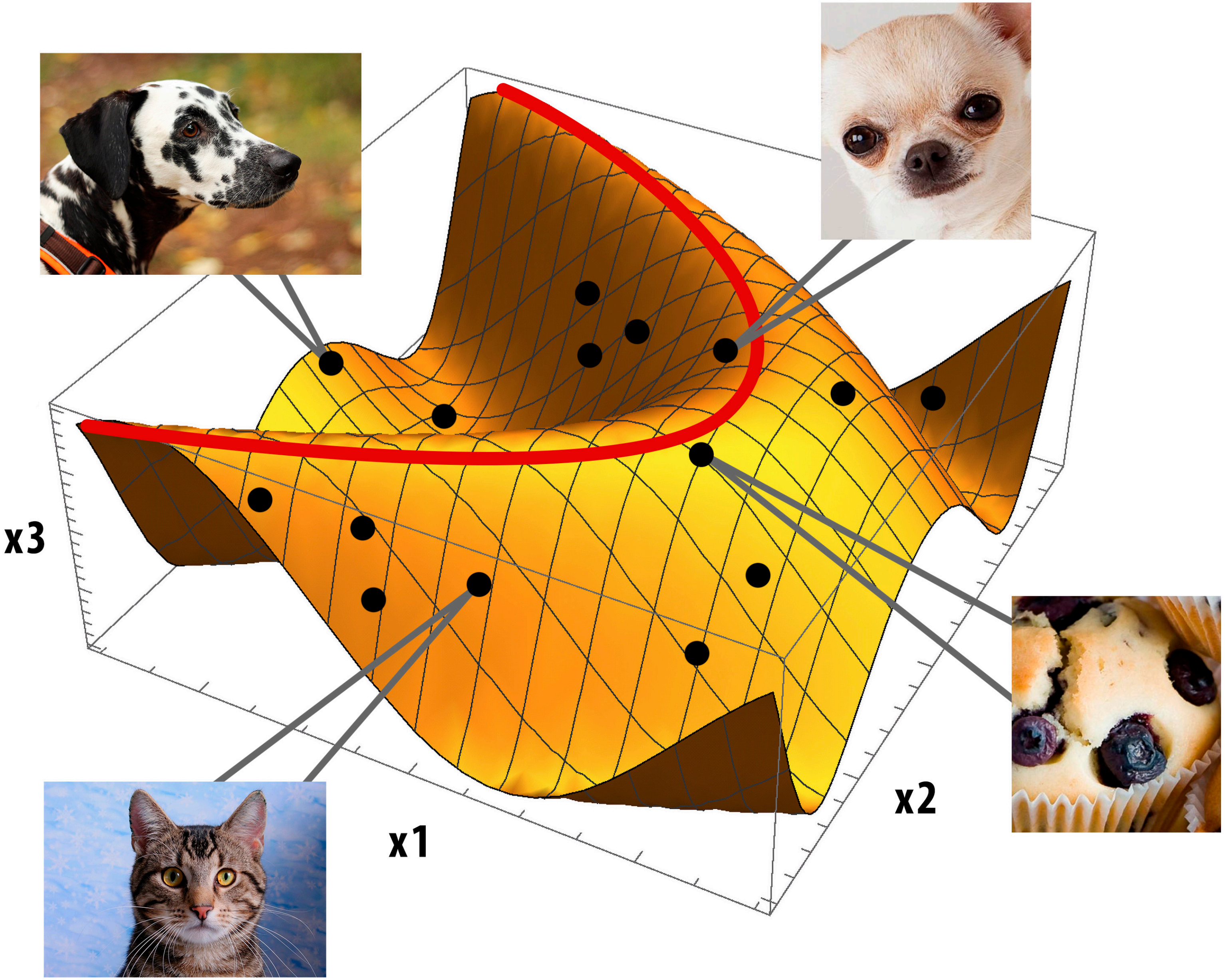

x3
x1
x2



x3
x1
x2


x3
x1
x2

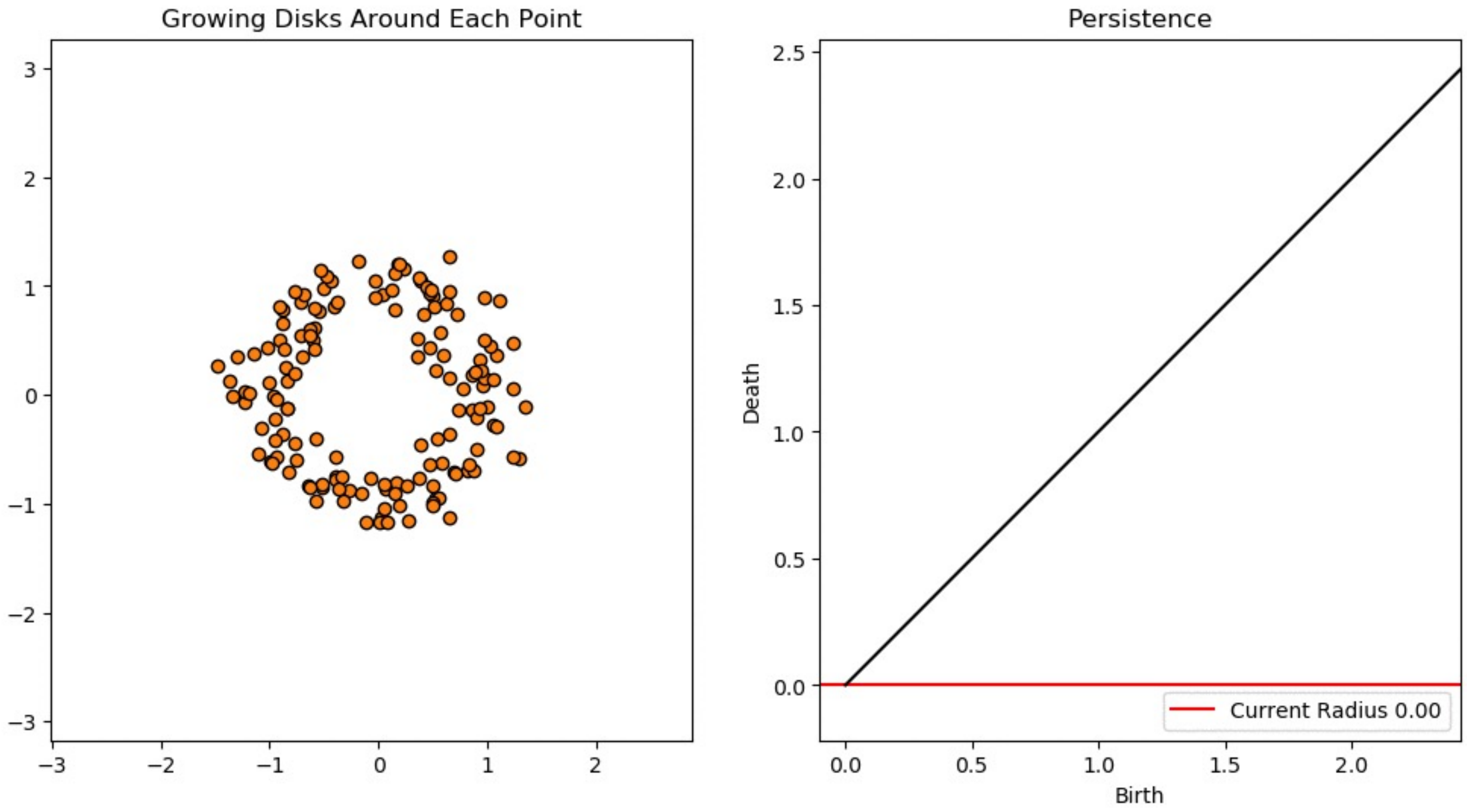

x3
x1
x2


x3
x1
x2

Topological Data Analysis

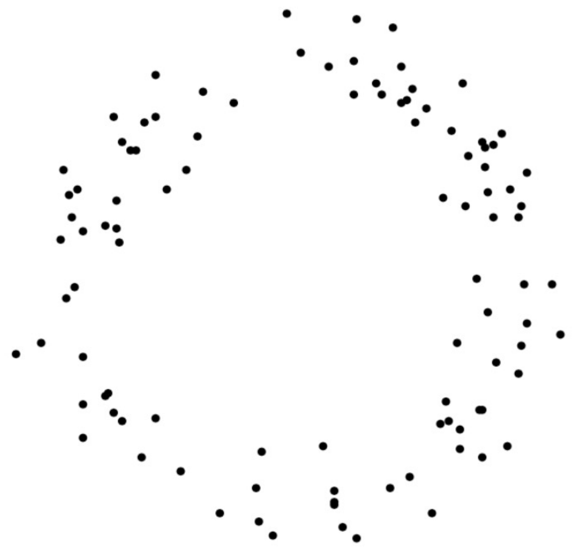
Source: Ulrich Bauer. Topological Data Analysis: An Introduction to Persistent Homology. MLSS, 2019

Persistent Diagram (in 1d)



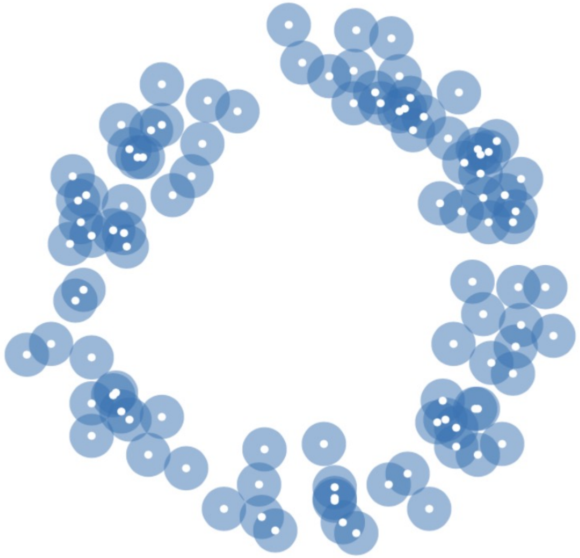
Source: https://towardsdatascience.com/persistent-homology-with-examples-1974d4b9c3d0

Persistent Barcodes



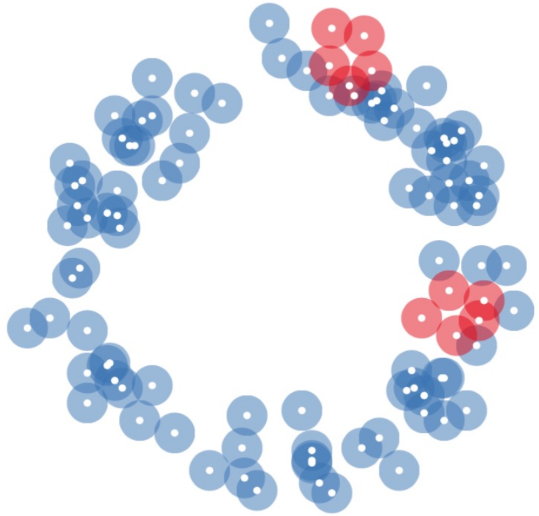
Source: Ulrich Bauer. Topological Data Analysis: An Introduction to Persistent Homology. MLSS, 2019

Persistent Barcodes



Source: Ulrich Bauer. Topological Data Analysis: An Introduction to Persistent Homology. MLSS, 2019

Persistent Barcodes

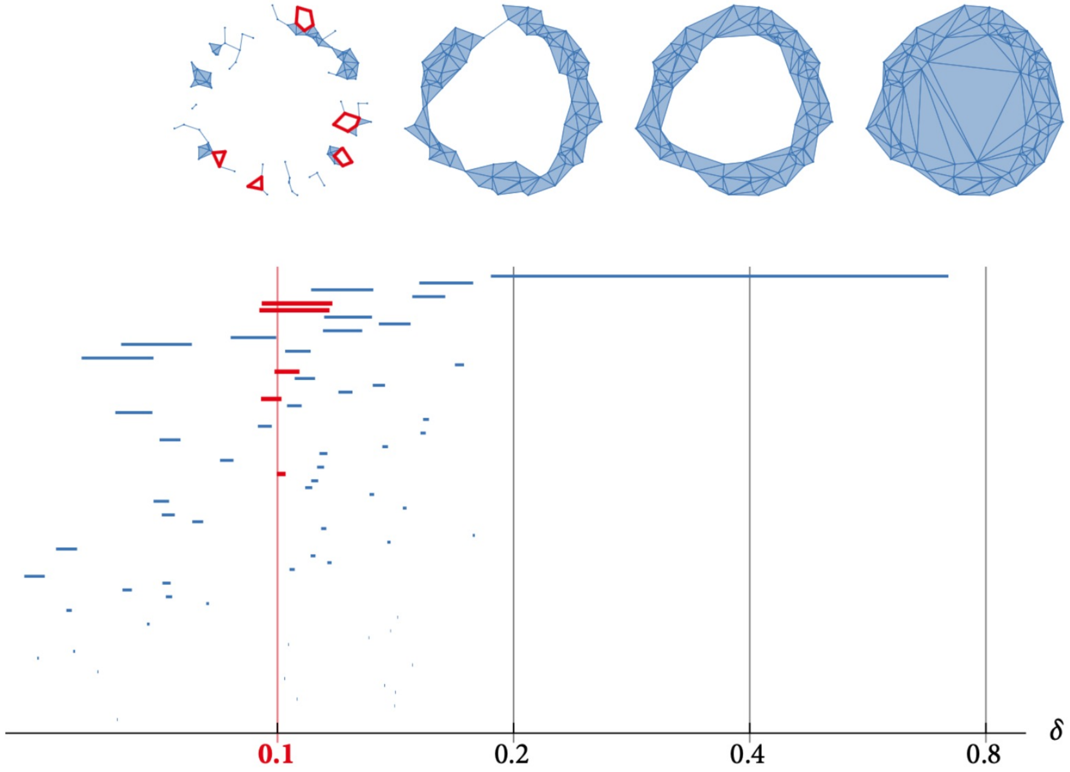


Source: Ulrich Bauer. Topological Data Analysis: An Introduction to Persistent Homology. MLSS, 2019

Persistent Barcodes

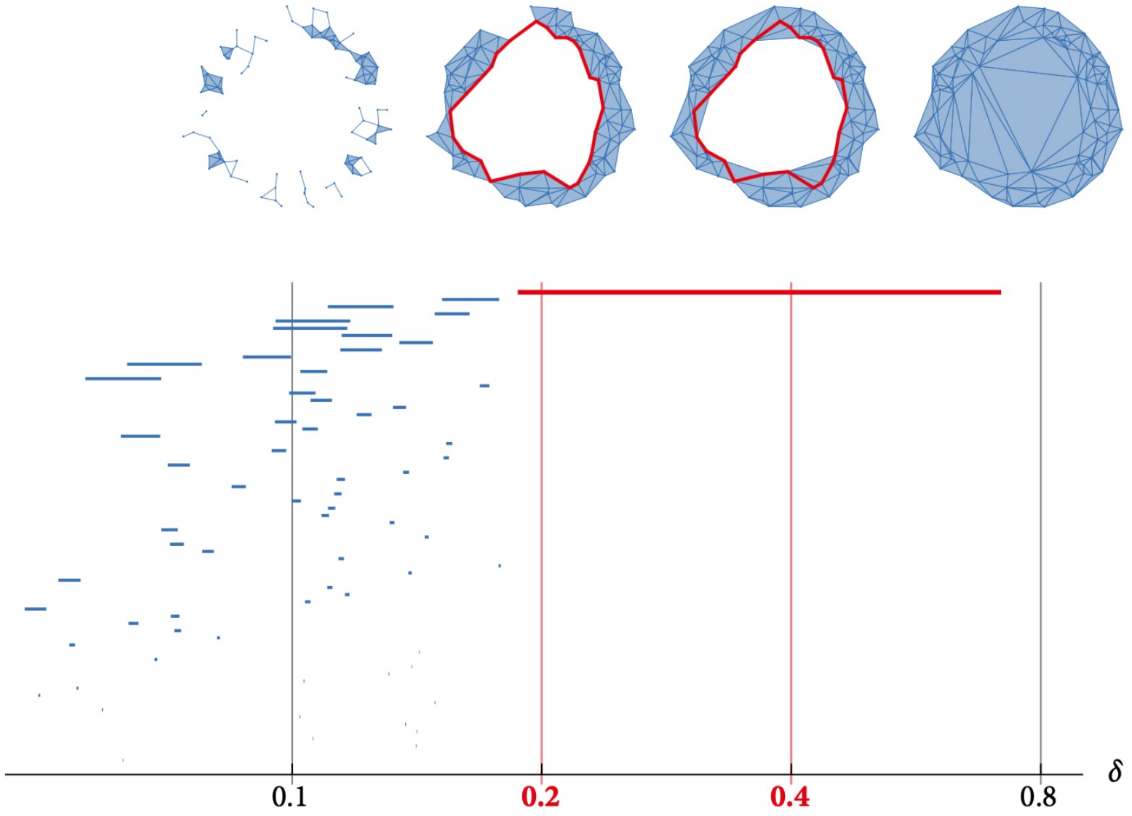
Source: Ulrich Bauer. Topological Data Analysis: An Introduction to Persistent Homology. MLSS, 2019

Persistent Barcodes



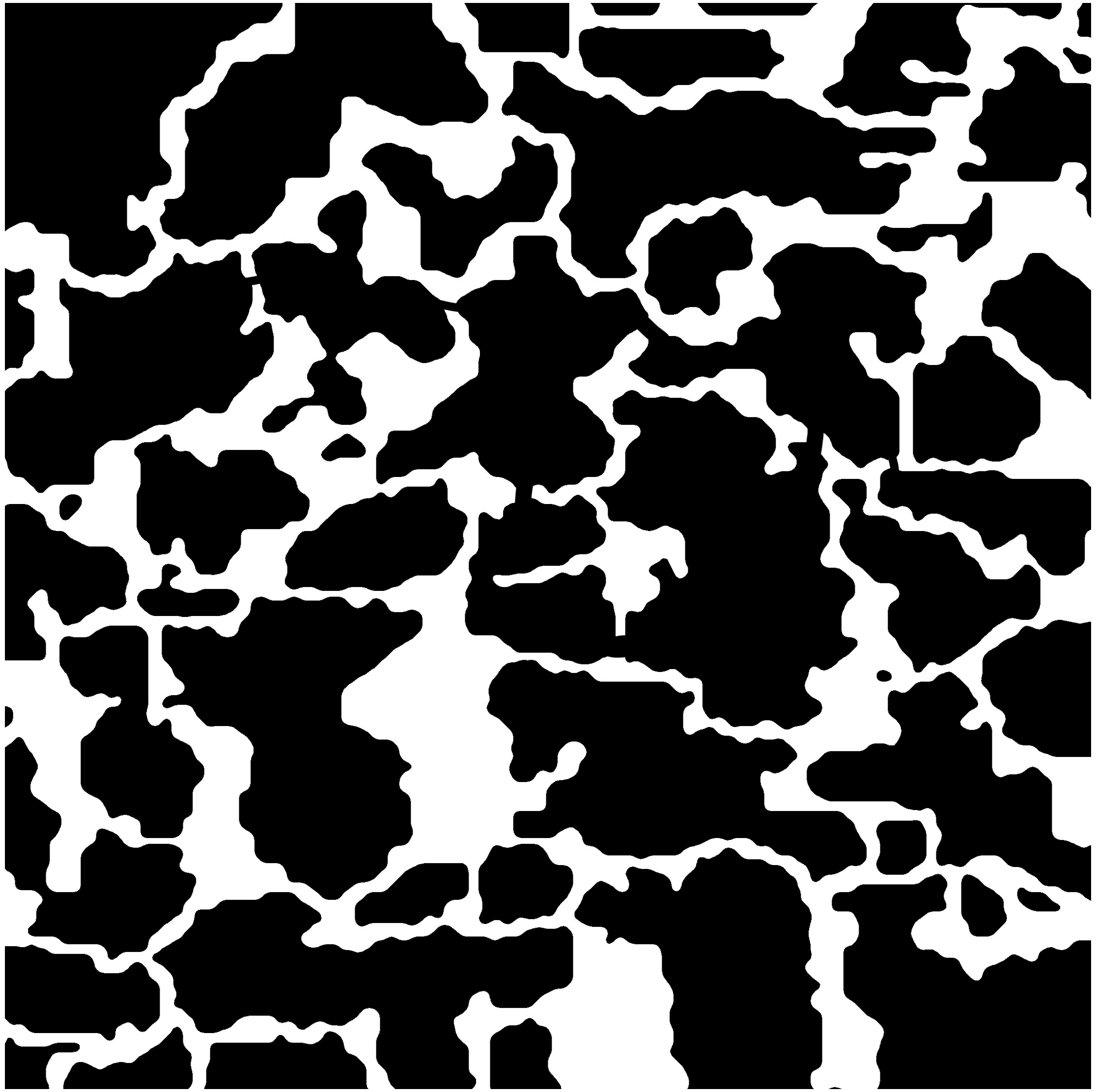
Source: Ulrich Bauer. Topological Data Analysis: An Introduction to Persistent Homology. MLSS, 2019

Persistent Barcodes

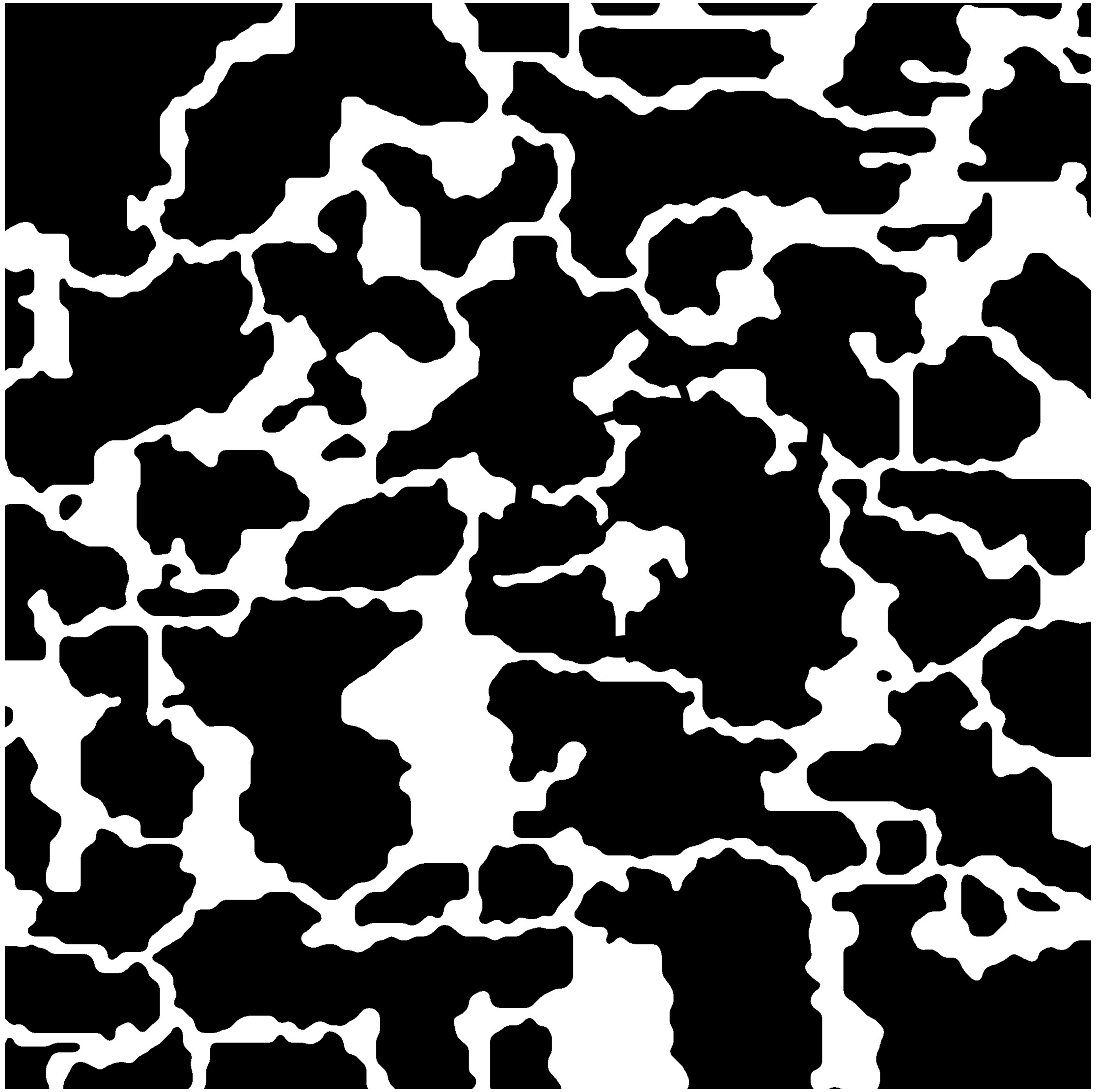


Source: Ulrich Bauer. Topological Data Analysis: An Introduction to Persistent Homology. MLSS, 2019

Example of core samples

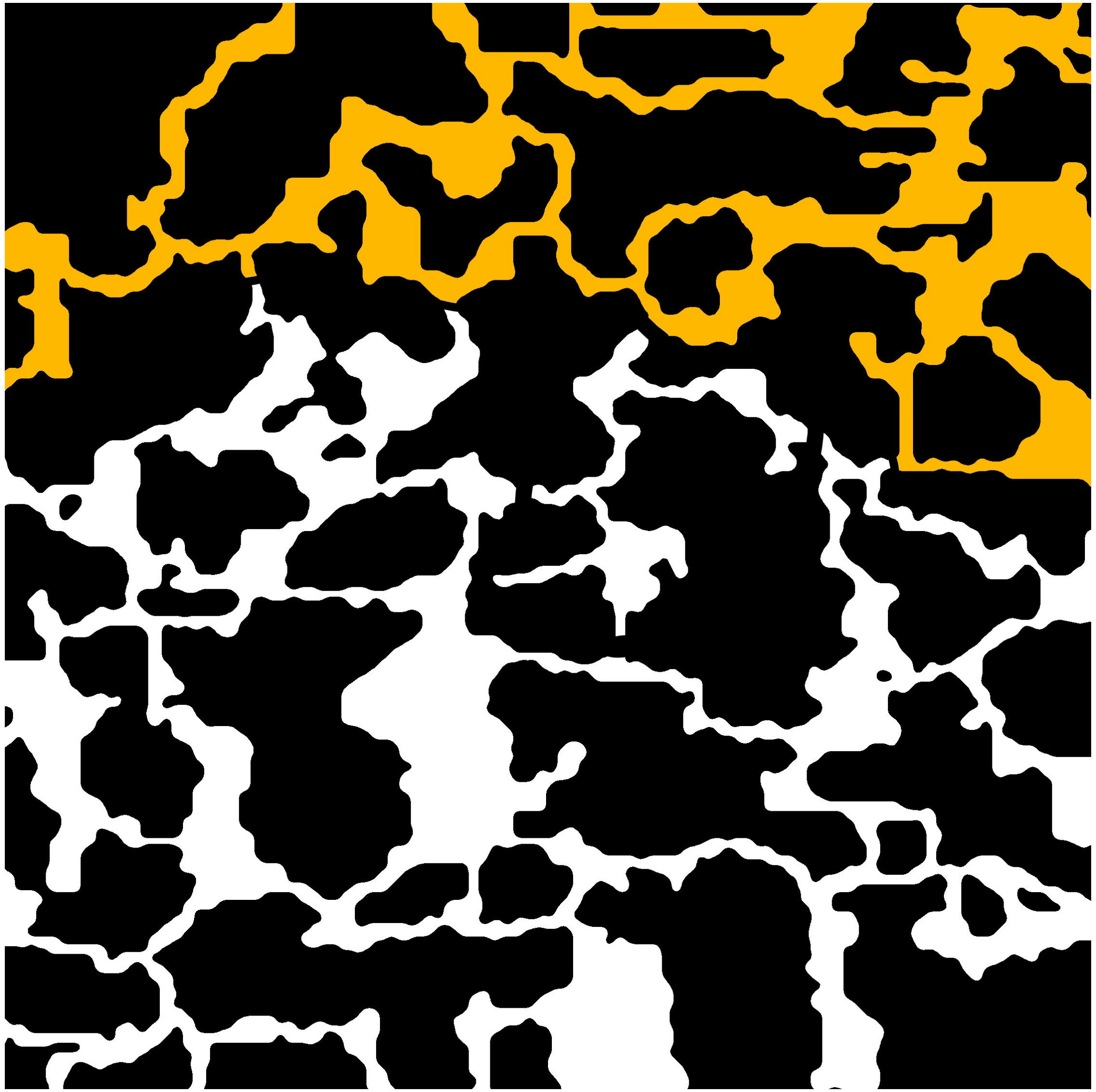


Zero permeability (in vertical direction)

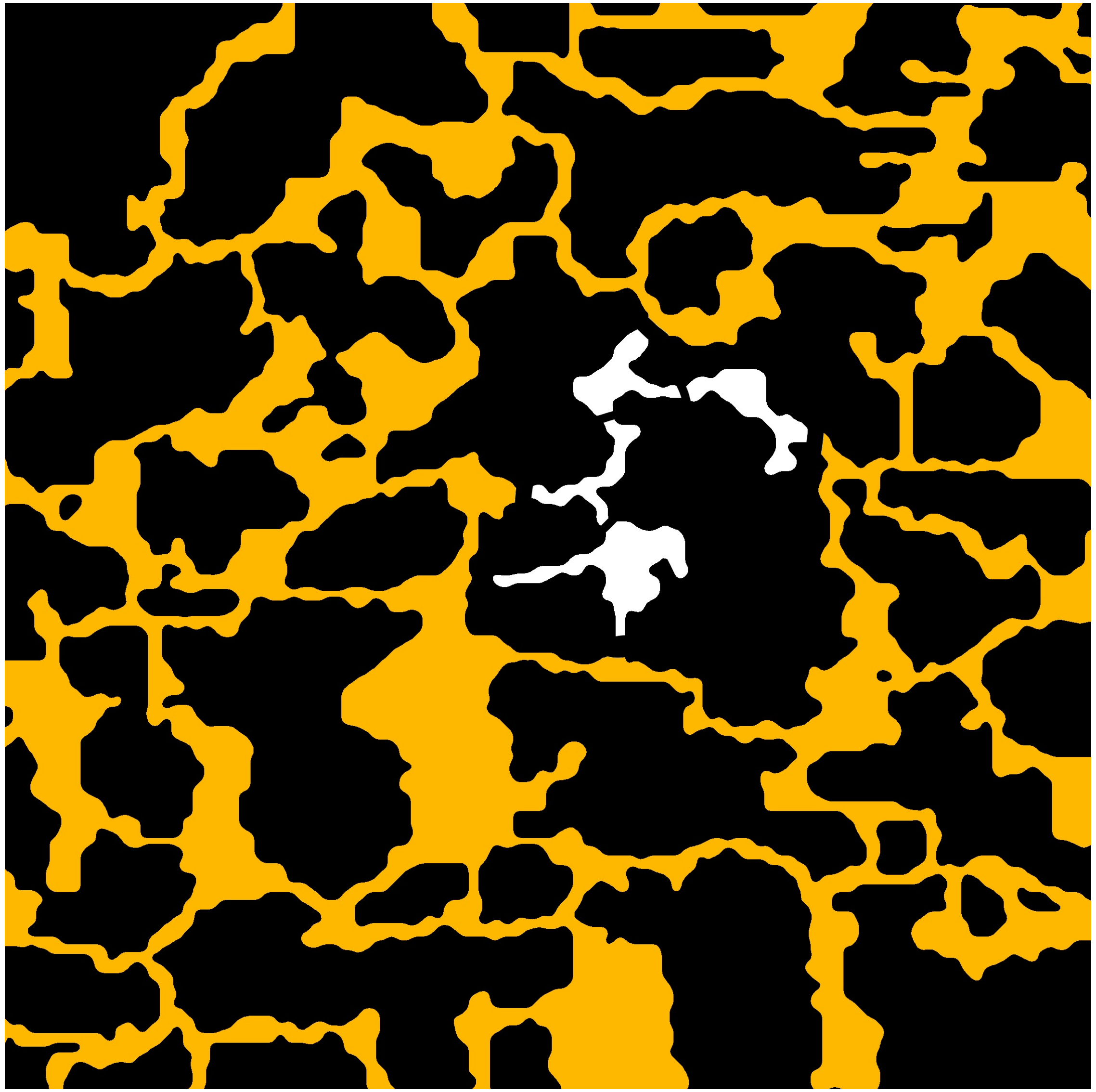


Non-zero permeability

Example of core samples



Zero permeability (in vertical direction)

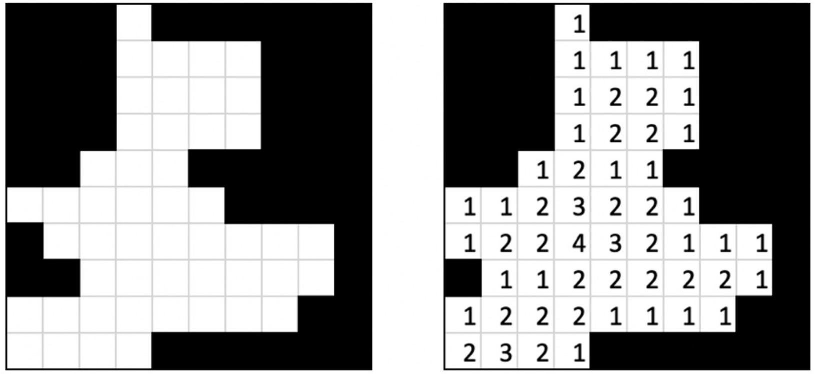


Non-zero permeability

Euclidean Distance Transform

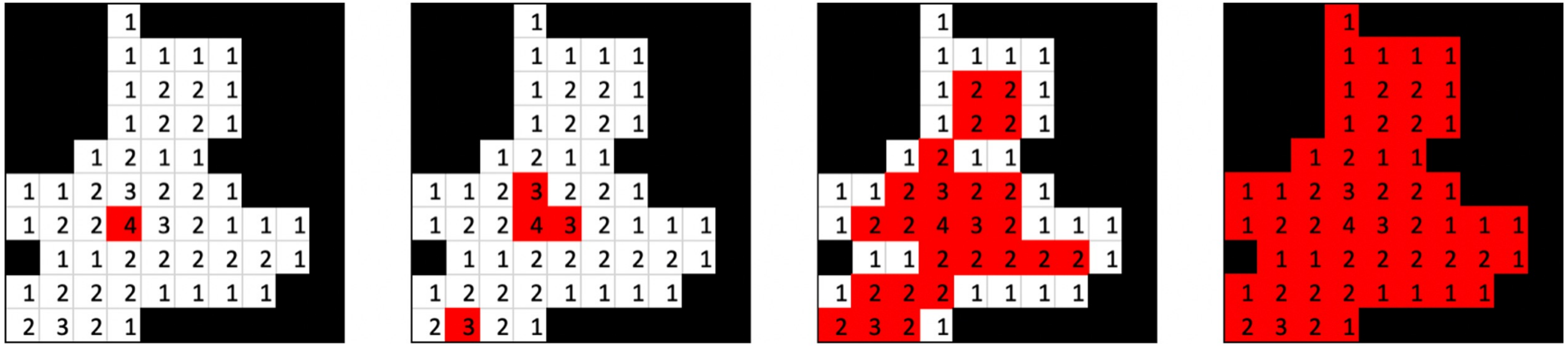
Let $A \in X$ then the Euclidean distance from the boundary of A to all points $x \in X$

$$f(x) = d(x, \partial A) := \inf_{y \in \partial A} \|x - y\|_2$$



Example for the norm $\|\cdot\|_\infty$

Filtration



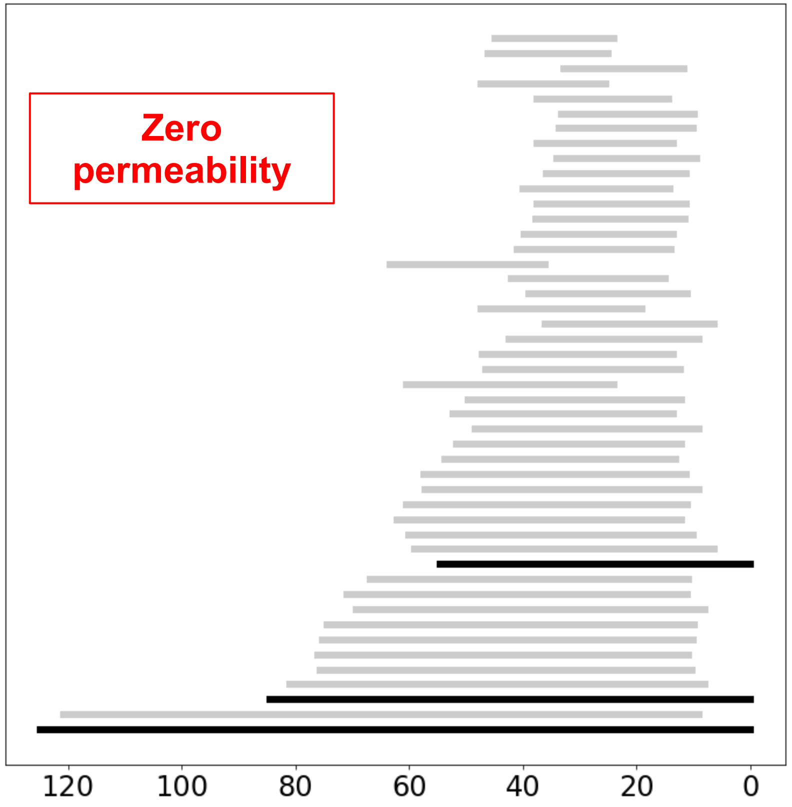
$$X^t = f^{-1}[t, +\infty) = \{x \in X \mid f(x) \geq t\}$$

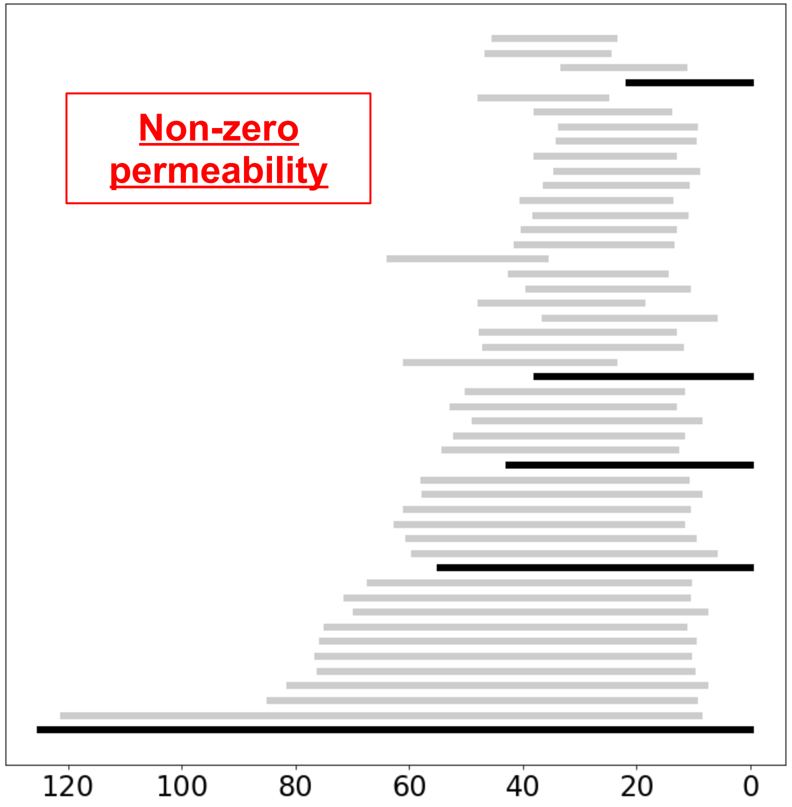
Minkowski functionals

	Zero permeability	Non-zero permeability	
	Core A	Core B	Diff. in %
Square	1593603	1593394	0.01
Perimeter	58087	58112	0.04
Euler characteristics	-32	-32	–

- **Conclusion**: Minkowski functionals can not detect permeability of core samples reliably

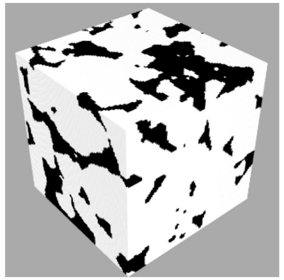
Persistent barcodes (dimension 0)





- **Persistent barcodes:** behaviour of topological characteristics of dimension 0 depending on characteristic size
- **Black lines** denotes those topological characteristics which corresponds to components of the connectivity of a set of pores
- We can calculate features from persistence barcodes to differentiate between zero/non-zero permeability

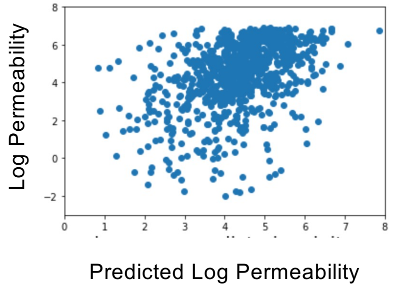
Some results



1. Linear model

$$\boxed{\text{Log Permeability} \sim a \cdot \text{Log Porosity} + b}$$

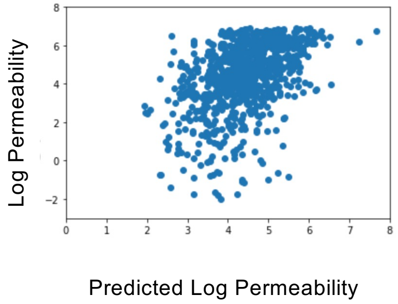
MAE(Permeability) ~ 157 mD



2. Linear model + correcting term

$$\boxed{\text{Log Permeability} \sim a \cdot \text{Log Porosity} + b + \Delta(x)}$$

- $\Delta(x)$ – Random Forest
- x = ("max", "mean", "std", "count", "entropy", "median", "sum", "kurtosis", "skewness") – features calculated from the persistence diagram
- MAE(Permeability) ~ 121 mD



Comparing Data Manifolds via Manifold Topology Divergence

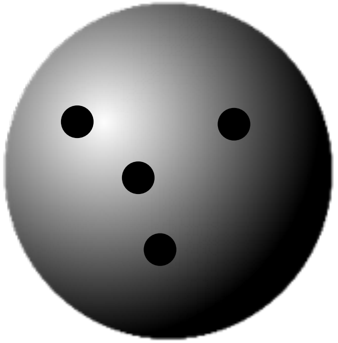
joint with S.Barannikov, I.Trofimov, G.Sotnikov, E.Trimbach,
A.Korotin, A.Filippov

Manifold Topology Divergence: a Framework for Comparing Data Manifolds. NeurIPS, 2021

Latent Generative Model

$$x \sim p(x|g_\theta(z)) \cdot p(z)$$

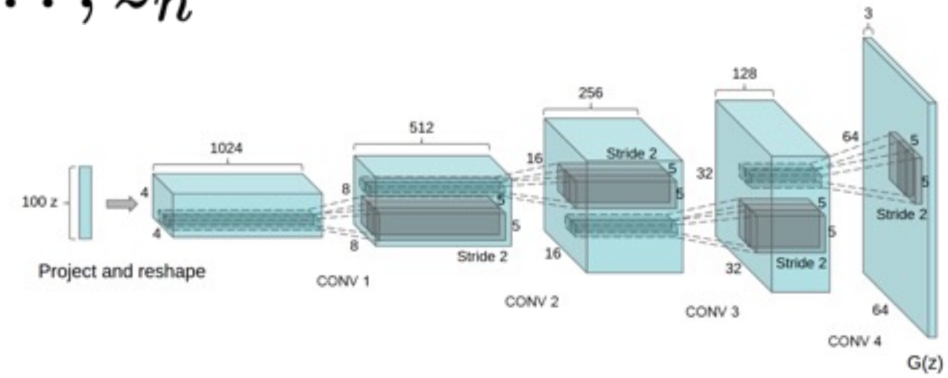
$$z \sim p(z)$$



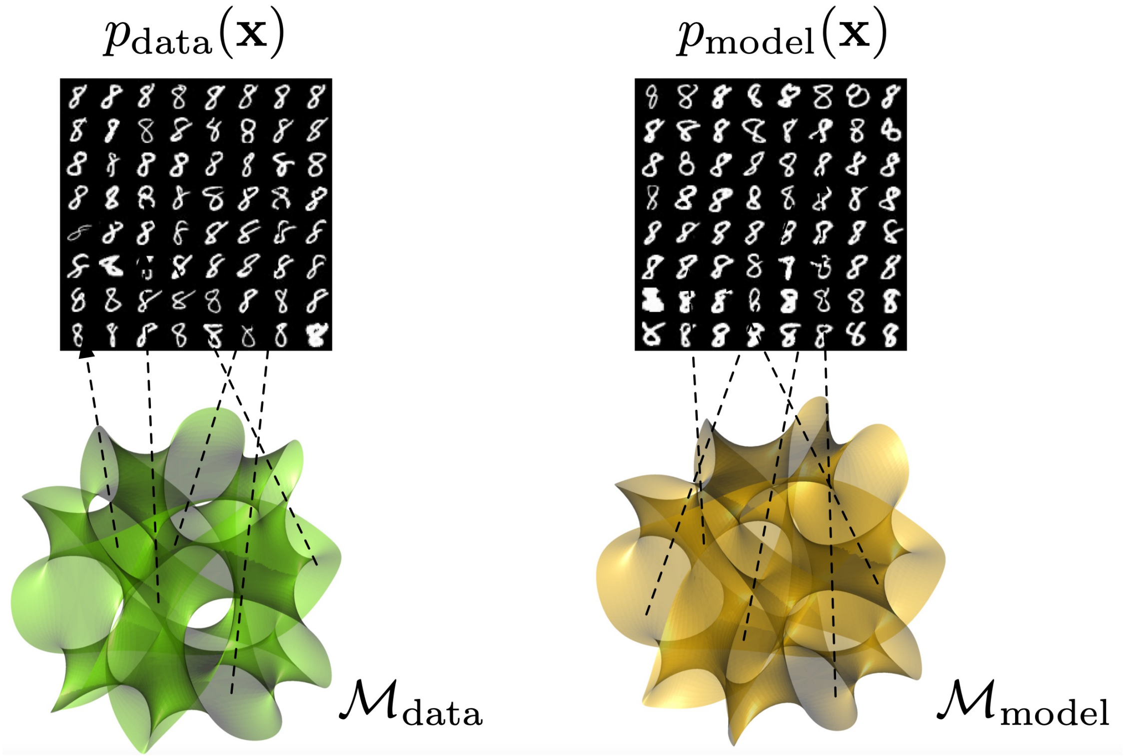
$$g_\theta$$

$$z_1, z_2, \ldots, z_n$$

$$x_1, x_2, \ldots, x_n$$



Evaluation of GANs



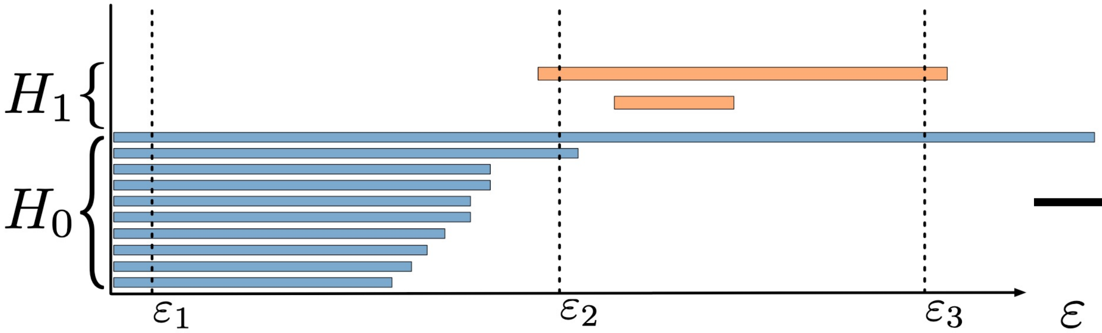
The manifold of **data** (real objects)

The manifold of a **model** (generated objects)

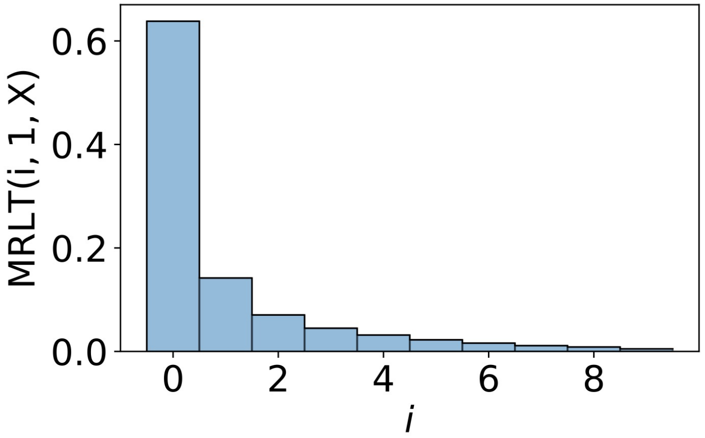
IDEA: evaluate GAN by comparing ***manifolds*** of real and generated objects

Source: Geometry score: A method for comparing generative adversarial networks. ICLR, 2018

Geometry Score



Persistent Barcode



Mean Relative Living Times (MRLT)

$$\mathrm{GeomScore}(X_1, X_2) \triangleq \sum_{i=0}^{i_{\max}-1} \left(\mathrm{MRLT}(i, 1, X_1) - \mathrm{MRLT}(i, 1, X_2)\right)^2$$

Khrulkov, V., & Oseledets, I. Geometry score: A method for comparing generative adversarial networks. ICML, 2018

Manifold Topology Divergence [NeurIPS, 2021]

IDEA: compare manifolds

$$M_{\text{data}},\ \ M_{\text{model}}$$

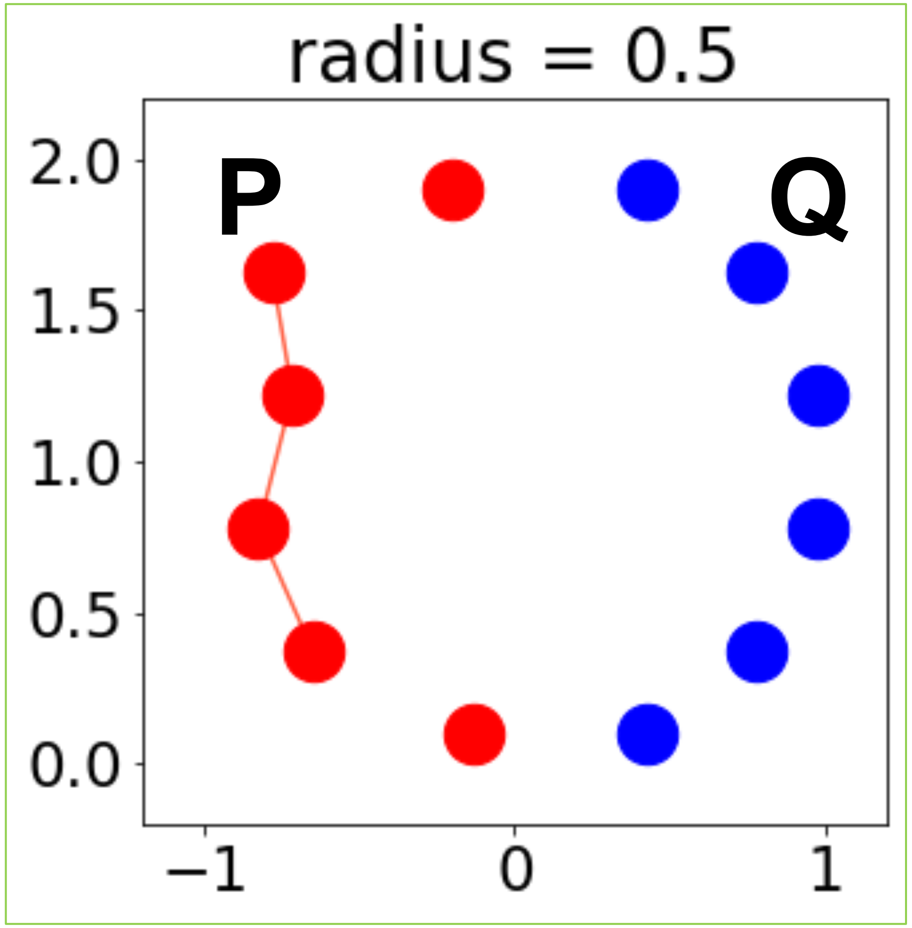
via calculating topological features of

$$\boxed{(M_{\text{data}} \cup M_{\text{model}})/M_{\text{model}}}$$

and $\boxed{(M_{\text{data}} \cup M_{\text{model}})/M_{\text{data}}}$

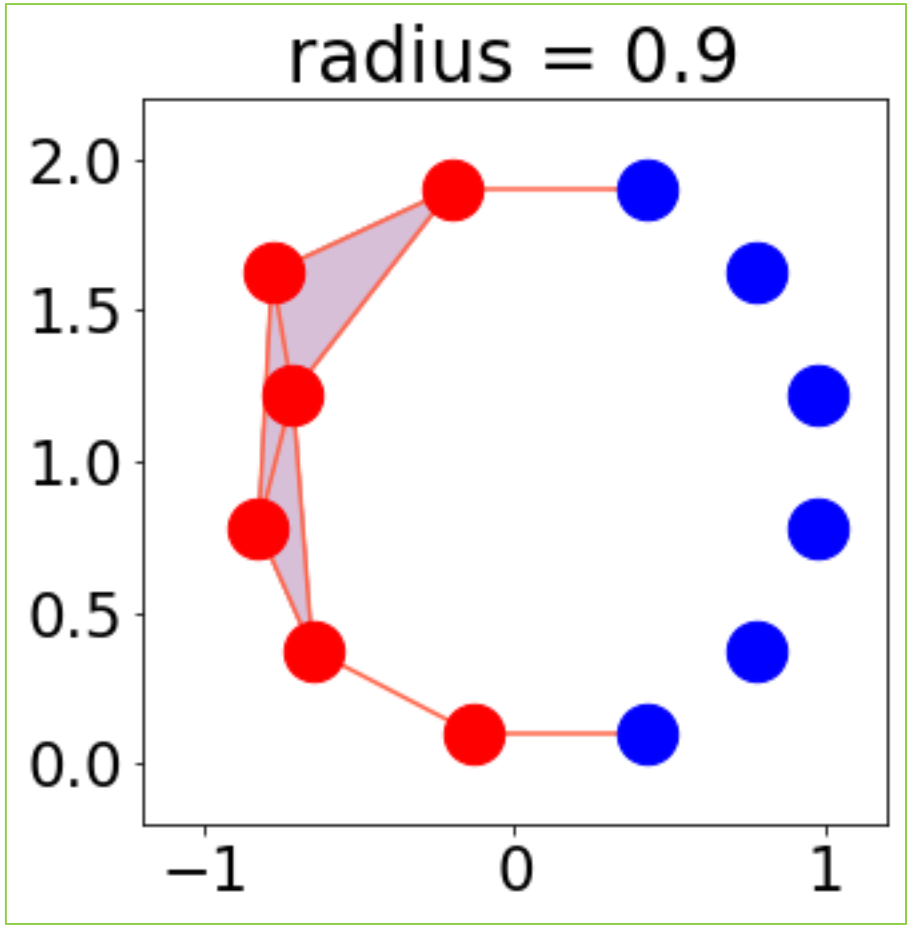
Manifold Topology Divergence: a Framework for Comparing Data Manifolds. NeurIPS, 2021. S.Barannikov, I.Trofimov, G.Sotnikov, E.Trimbach, A.Korotin, A.Filippov, E.Burnaev.

Manifold Topology Divergence [NeurIPS, 2021]



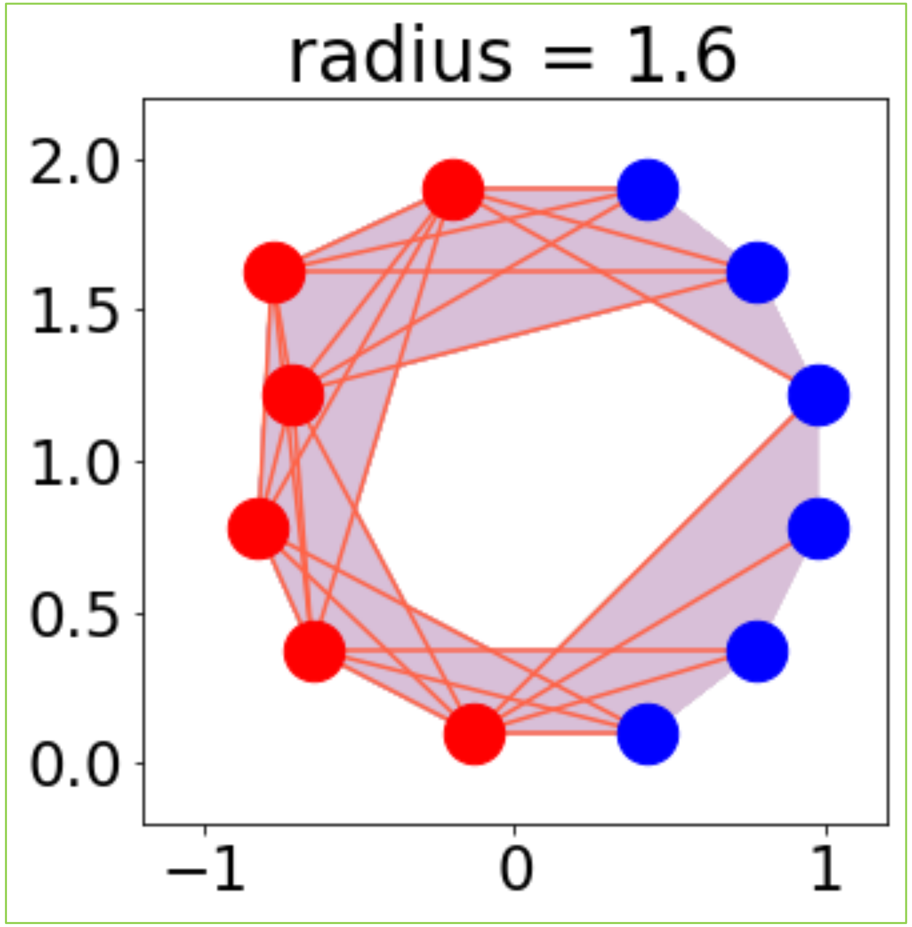
Manifold Topology Divergence: a Framework for Comparing Data Manifolds. NeurIPS, 2021. S.Barannikov, I.Trofimov, G.Sotnikov, E.Trimbach, A.Korotin, A.Filippov, E.Burnaev.

Manifold Topology Divergence [NeurIPS, 2021]



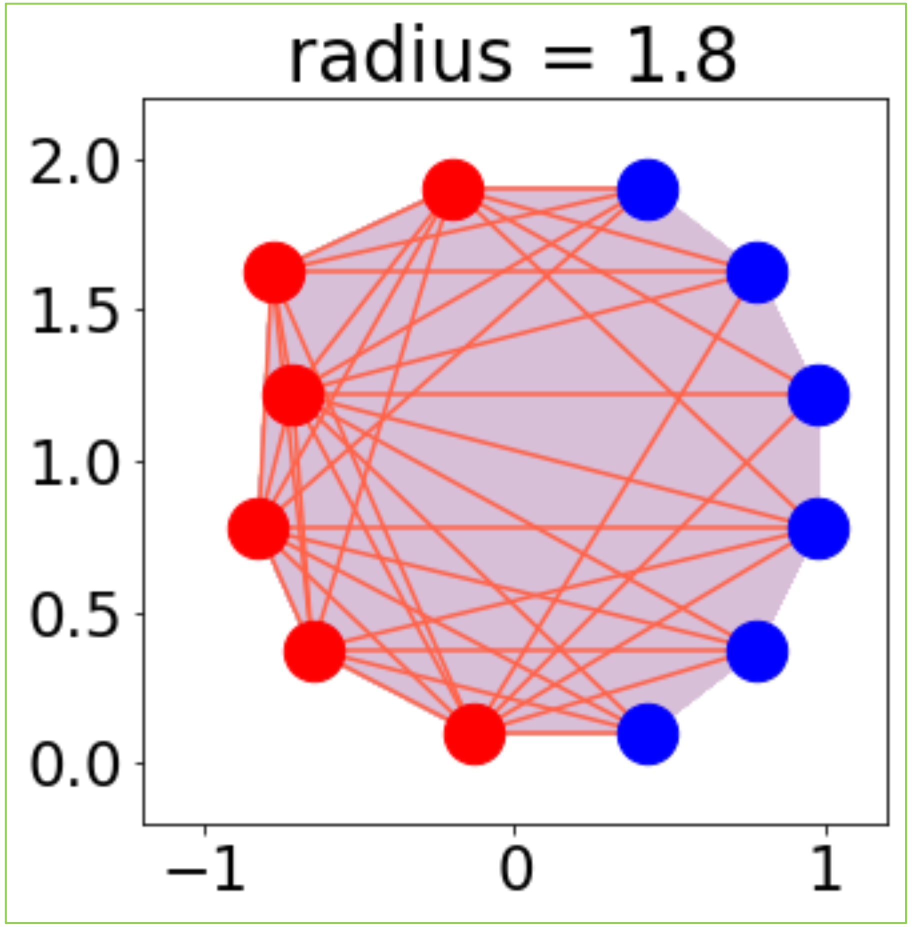
Manifold Topology Divergence: a Framework for Comparing Data Manifolds. NeurIPS, 2021. S.Barannikov, I.Trofimov, G.Sotnikov, E.Trimbach, A.Korotin, A.Filippov, E.Burnaev.

Manifold Topology Divergence [NeurIPS, 2021]



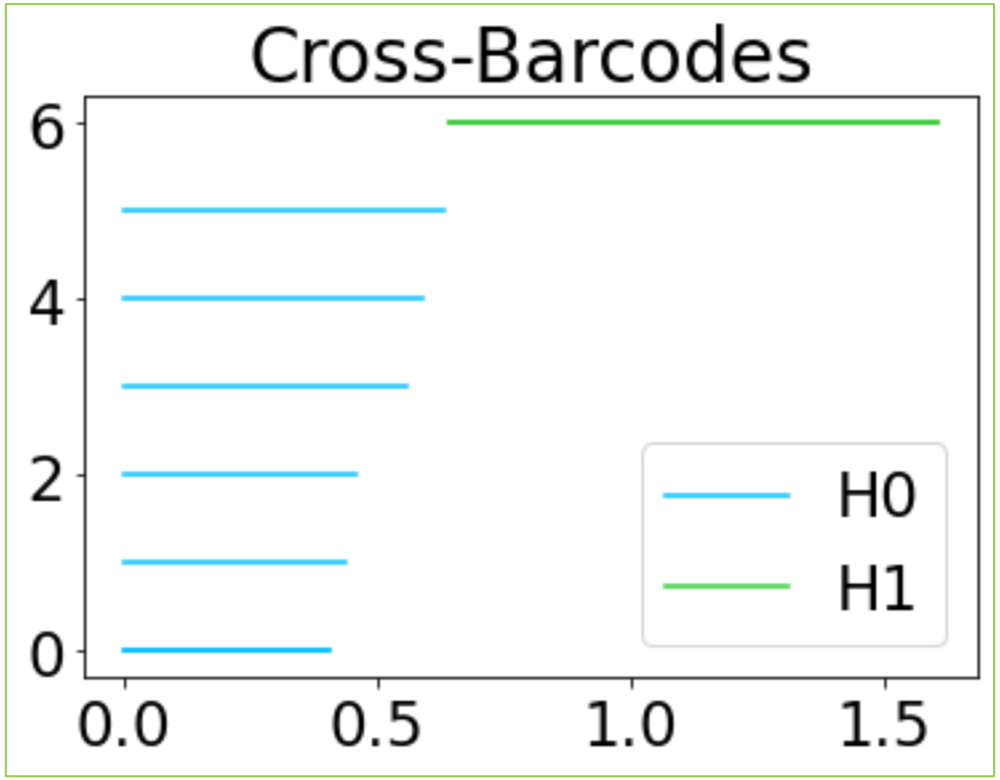
Manifold Topology Divergence: a Framework for Comparing Data Manifolds. NeurIPS, 2021. S.Barannikov, I.Trofimov, G.Sotnikov, E.Trimbach, A.Korotin, A.Filippov, E.Burnaev.

Manifold Topology Divergence [NeurIPS, 2021]



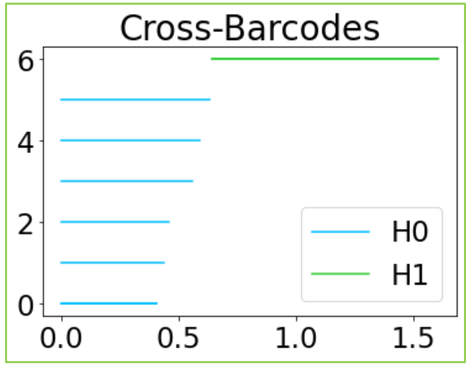
Manifold Topology Divergence: a Framework for Comparing Data Manifolds. NeurIPS, 2021. S.Barannikov, I.Trofimov, G.Sotnikov, E.Trimbach, A.Korotin, A.Filippov, E.Burnaev.

Manifold Topology Divergence [NeurIPS, 2021]



Manifold Topology Divergence: a Framework for Comparing Data Manifolds. NeurIPS, 2021. S.Barannikov, I.Trofimov, G.Sotnikov, E.Trimbach, A.Korotin, A.Filippov, E.Burnaev.

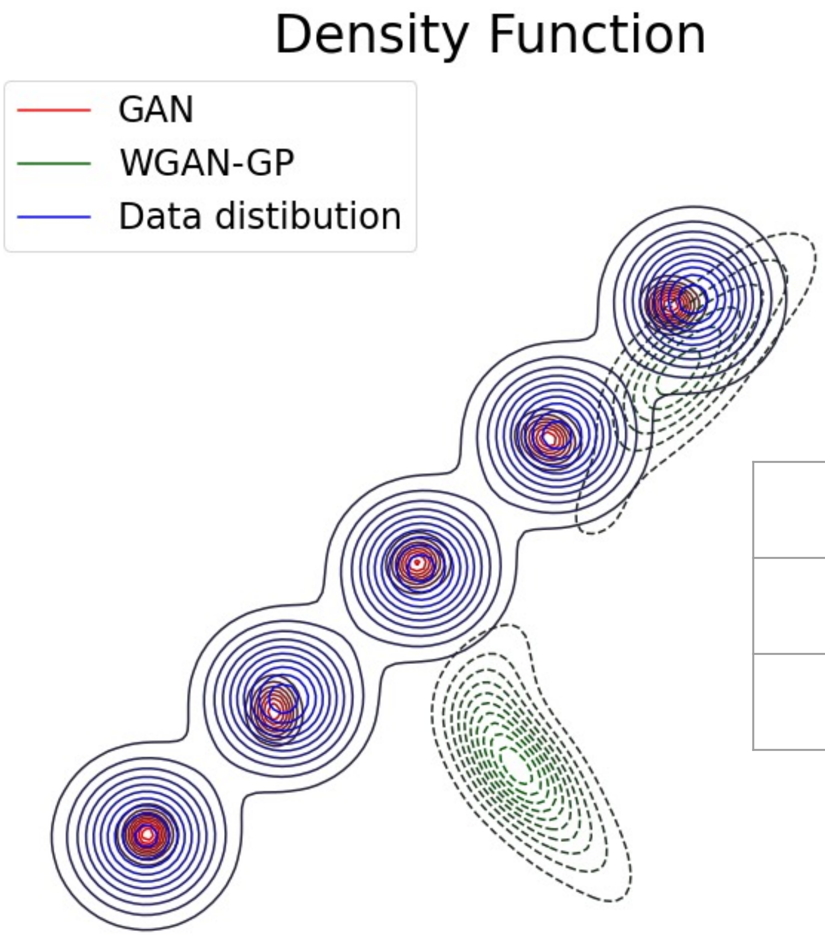
Cross-Barcode(P,Q)



$||$Cross-Barcode(P, Q)$||_B$ is bounded from above by the **Hausdorff distance** between **P** and **Q,** where $||\ .\ ||_B$ is the **bottleneck distance**

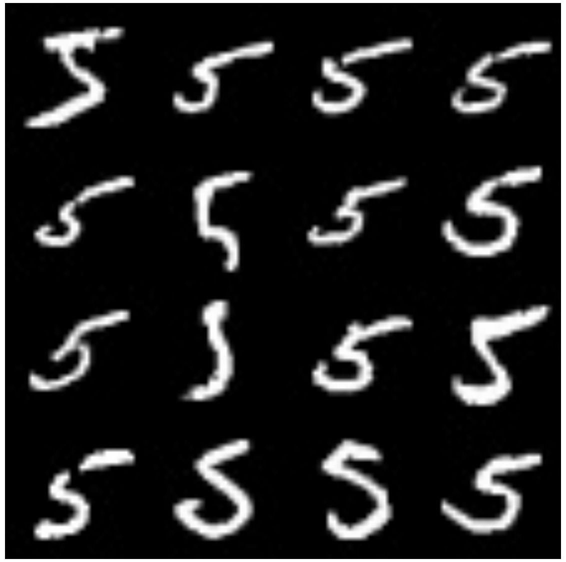
MTop-Div(P, Q) by definition equals the **sum of lengths of segments** in **Cross-Barcode$_1$(P, Q)**

Mode dropping detection



Gen. Model	G. SCORE	MTop-Div
WGAN-GP	**1.083**	0.562
orig. GAN	1.087	**0.081**

‘5’s vs. flipped ‘5’s

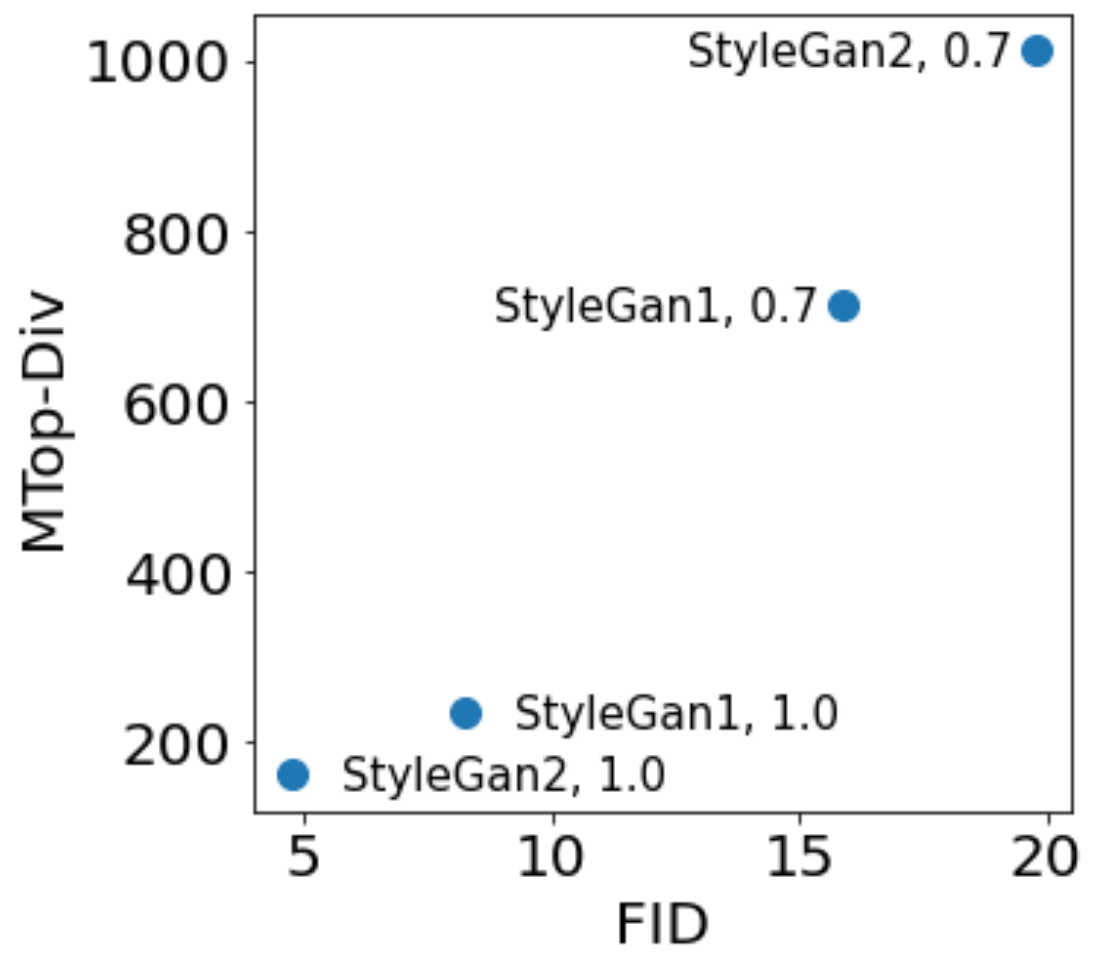


vs

Geometry Score = 0.0

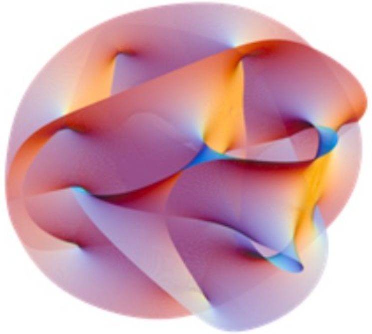
MTop-Div = 6154.0

FID vs. **MTop-Div** for **StyleGAN**, **StyleGAN2** on **FFHQ**



MTop-Div is monotonically increasing in good correlation with **FID**

Conclusions



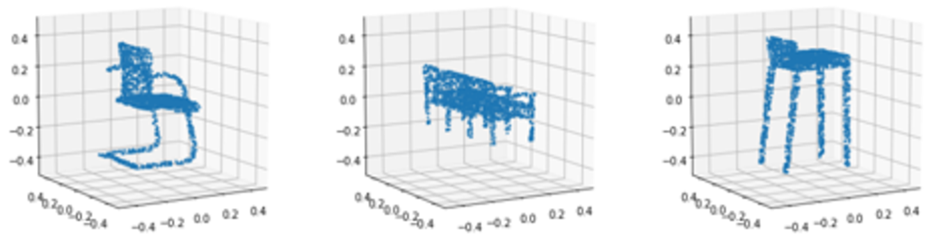
- **Machine Learning is about Shape of Data**

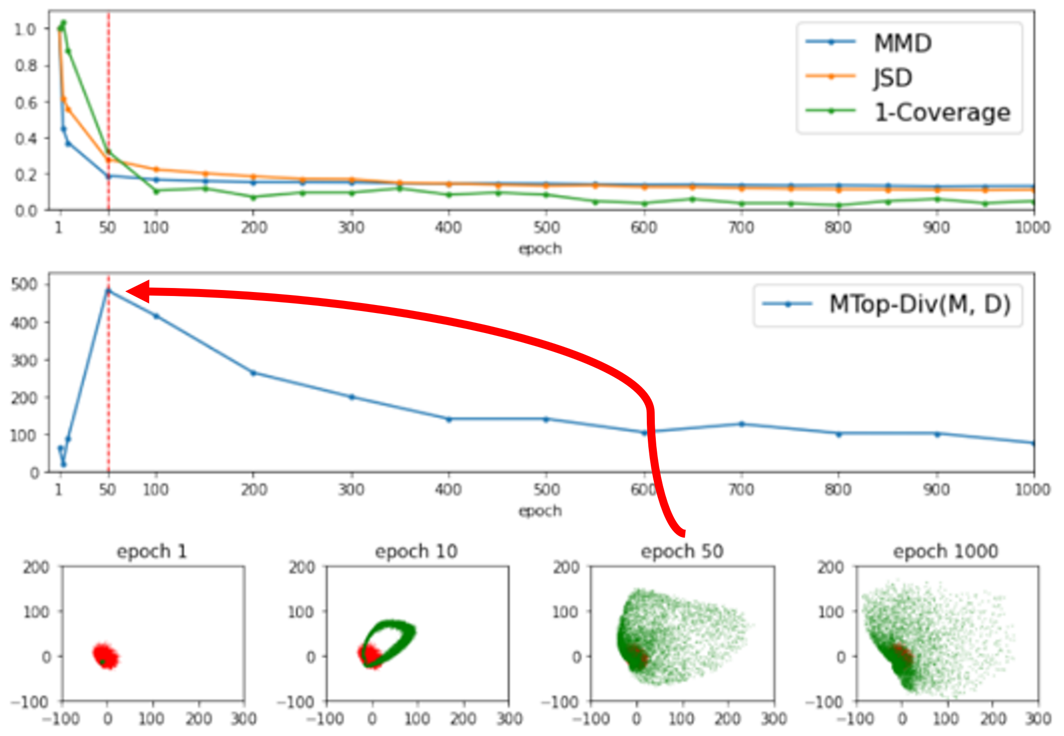
- TDA-based methods for feature generation and rock properties assessment

- New MTop-Div divergence, compared against 6 established evaluation methods: FID, discriminative score, MMD, JSD, 1-coverage, and Geometry score. MTop-Div is able to capture subtle differences in data geometry

- We overcame the known TDA scalability issues and in particular have carried out the MTop-Div calculations on most recent datasets such as FFHQ, with dimensionality $d \sim 10^7$

Experiments. 3D GAN.



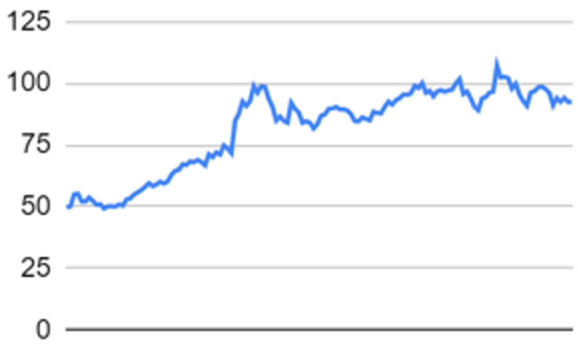


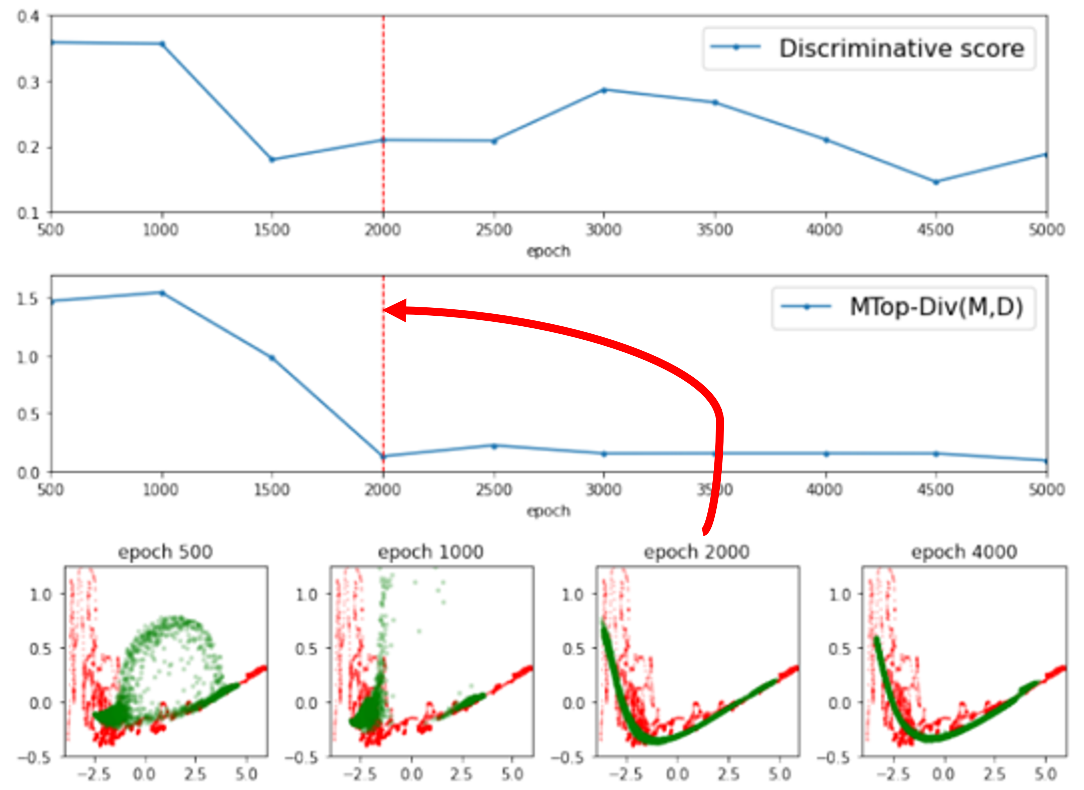
Training process of GAN applied to 3D shapes. Normalized quality measures MMD, JSD, 1-Coverage, MTop-Div vs. epoch. Lower is better.

MTop-Div is more sensitive than standard quality measures.

PCA projection of real objects (red) and generated objects (green). Vertical red line (epoch 50) depicts the moment, when the manifold of generated objects “explodes” and becomes much more diverse.

Experiments. TimeGAN.





Training dynamics of TimeGAN applied to market stock data. Discriminative score vs. epoch, MTop-Div vs. epoch. Lower is better.

MTop-Div agrees with discriminative score.

PCA projection of real time-series (red) and generated time-series (green). Vertical red line (epoch 2000) depicts the moment when manifolds of real and generated objects become close.

Conclusions

1. We introduced a new tool: Cross-Barcode(P, Q). For a pair of point clouds P and Q, the Cross-Barcode(P, Q) records the differences in multiscale topology between two manifolds approximated by the point clouds;

2. We proposed a new measure for comparing two data manifolds approximated by point clouds: Manifold Topology Divergence (MTop-Div);

3. We applied the MTop-Div to evaluate performance of GANs in various domains: 2D images, 3D shapes, time-series. We show that the MTop-Div correlates well with domain-specific measures and can be used for model selection. Also it provides insights about evolution of generated data manifold during training;