INF556

Topological Data Analysis (TDA)

Steve Oudot

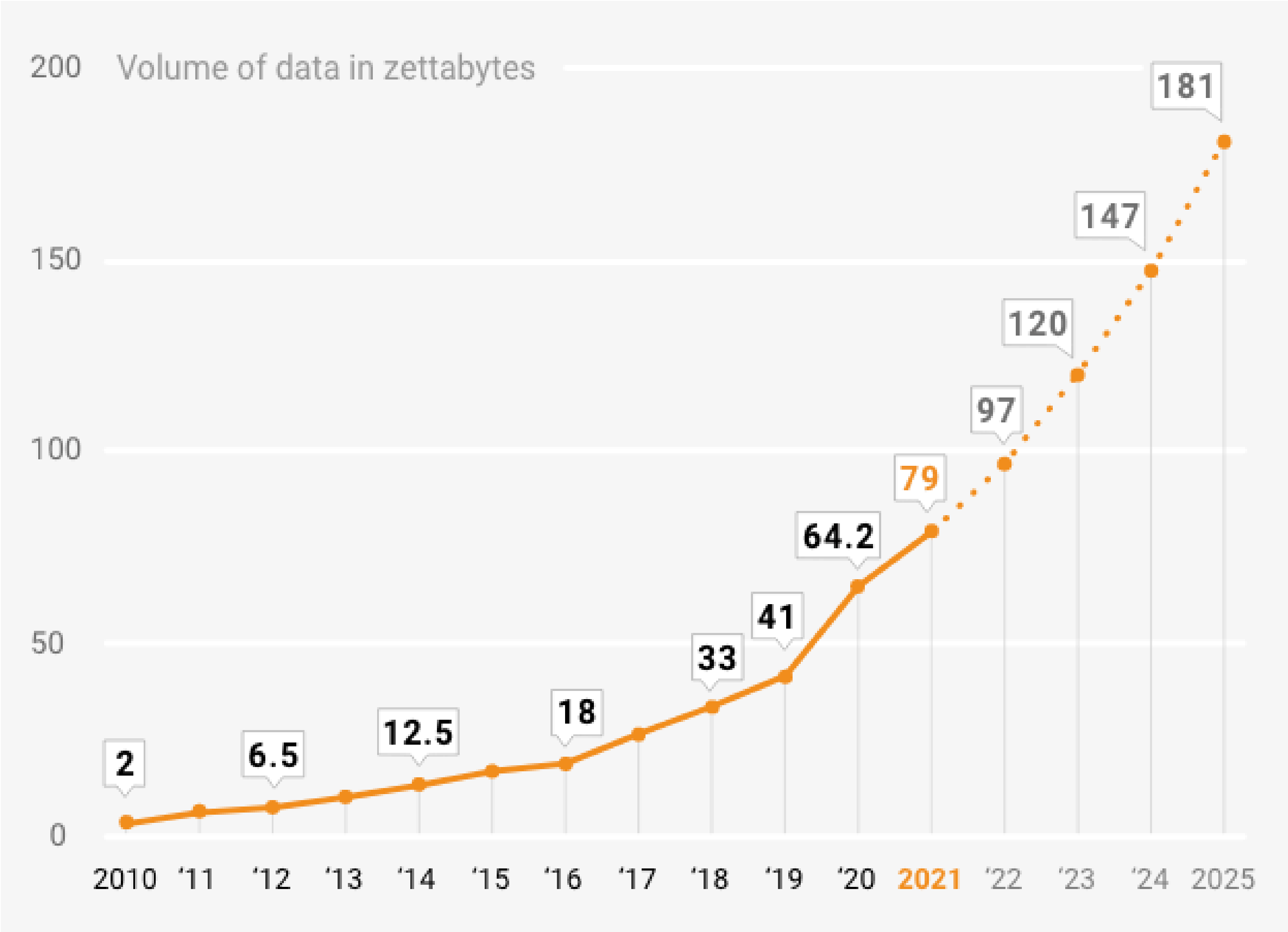
steve.oudot@inria.fr

Context: the data deluge

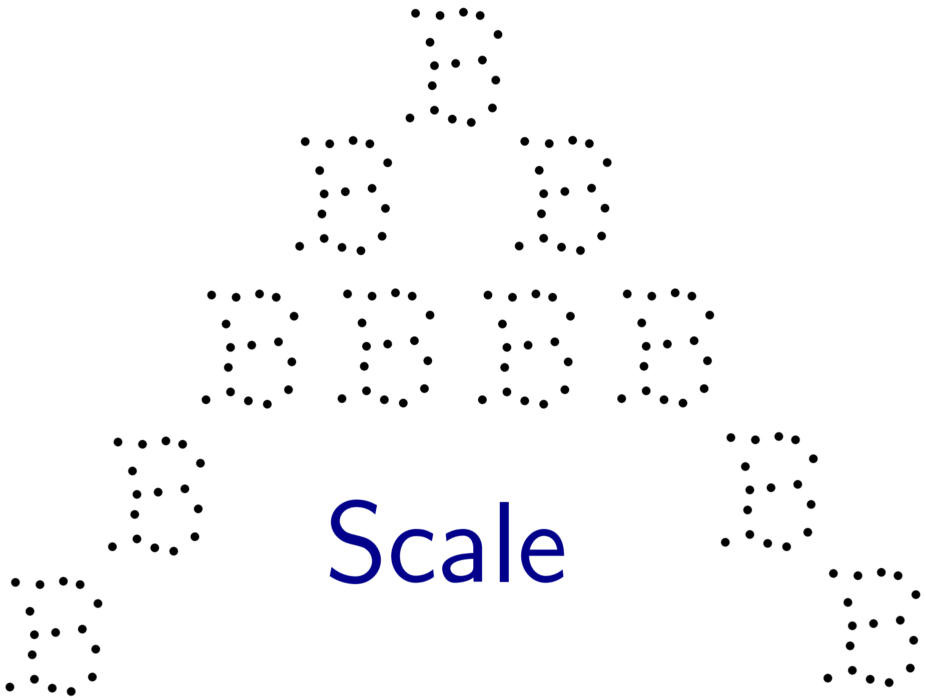
▶ **total amount of data stored in the world** (including 10% of *unique data*)**:**

2 Zo (2010) $\longrightarrow$ 79 Zo (2021) $\stackrel{\text{predict.}}{\longrightarrow}$ 181 Zo (2025) (1 Zo $= 10^{21}$ octets)

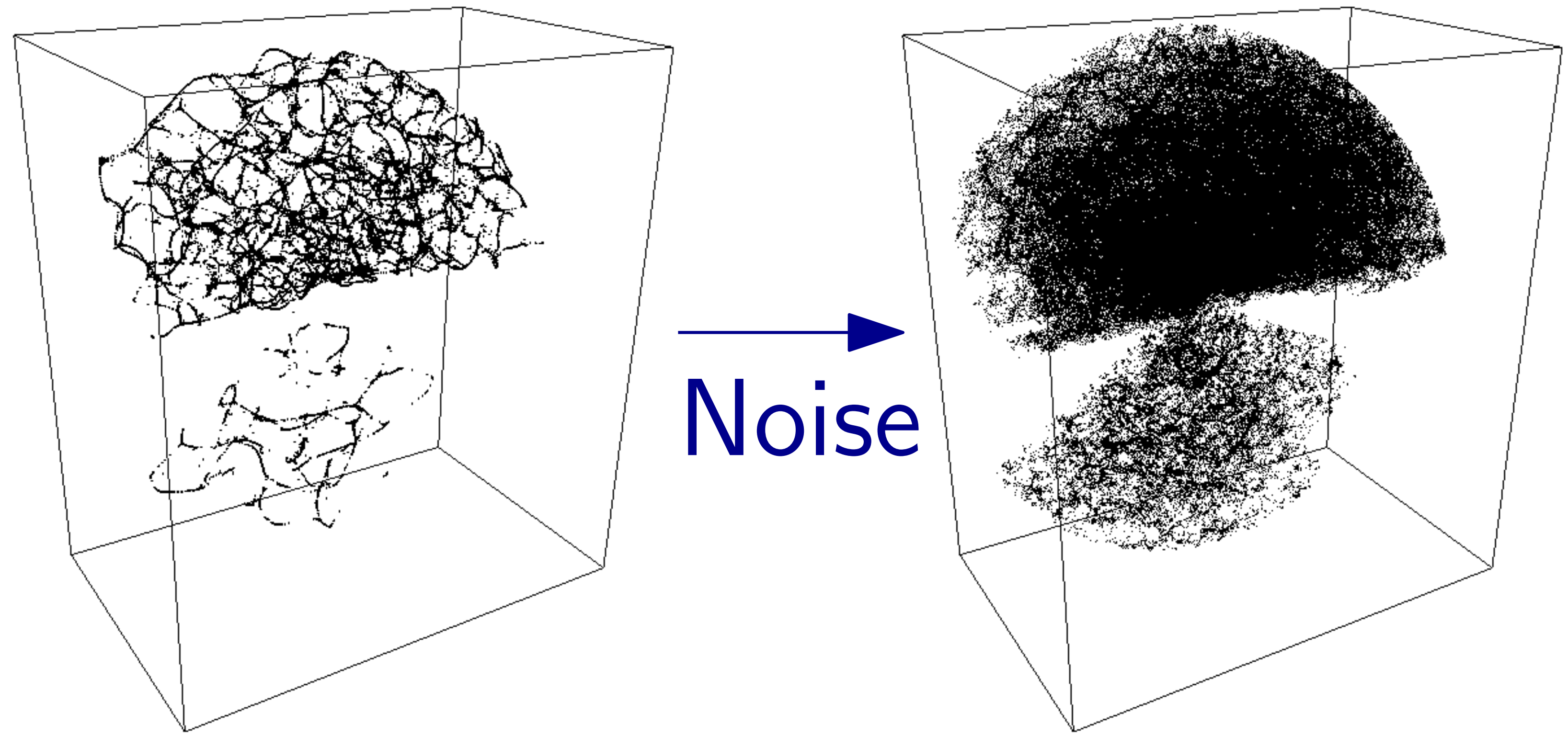
— source: International Data Corporation



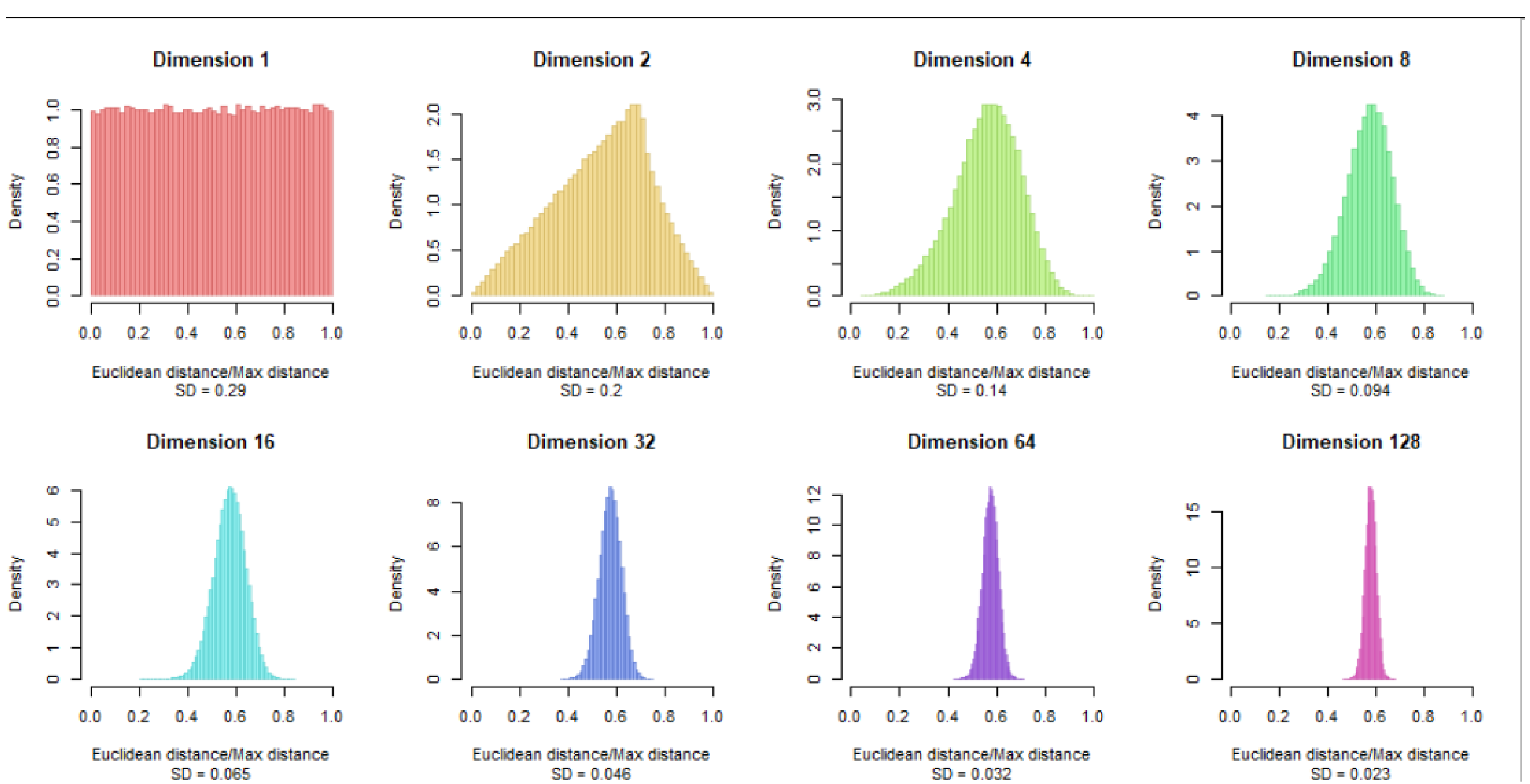
Major challenges



Scale



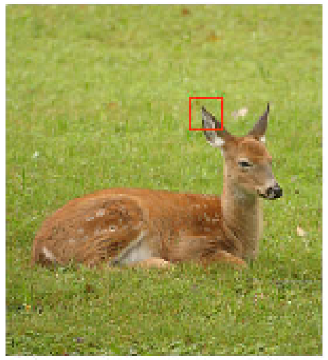
Noise

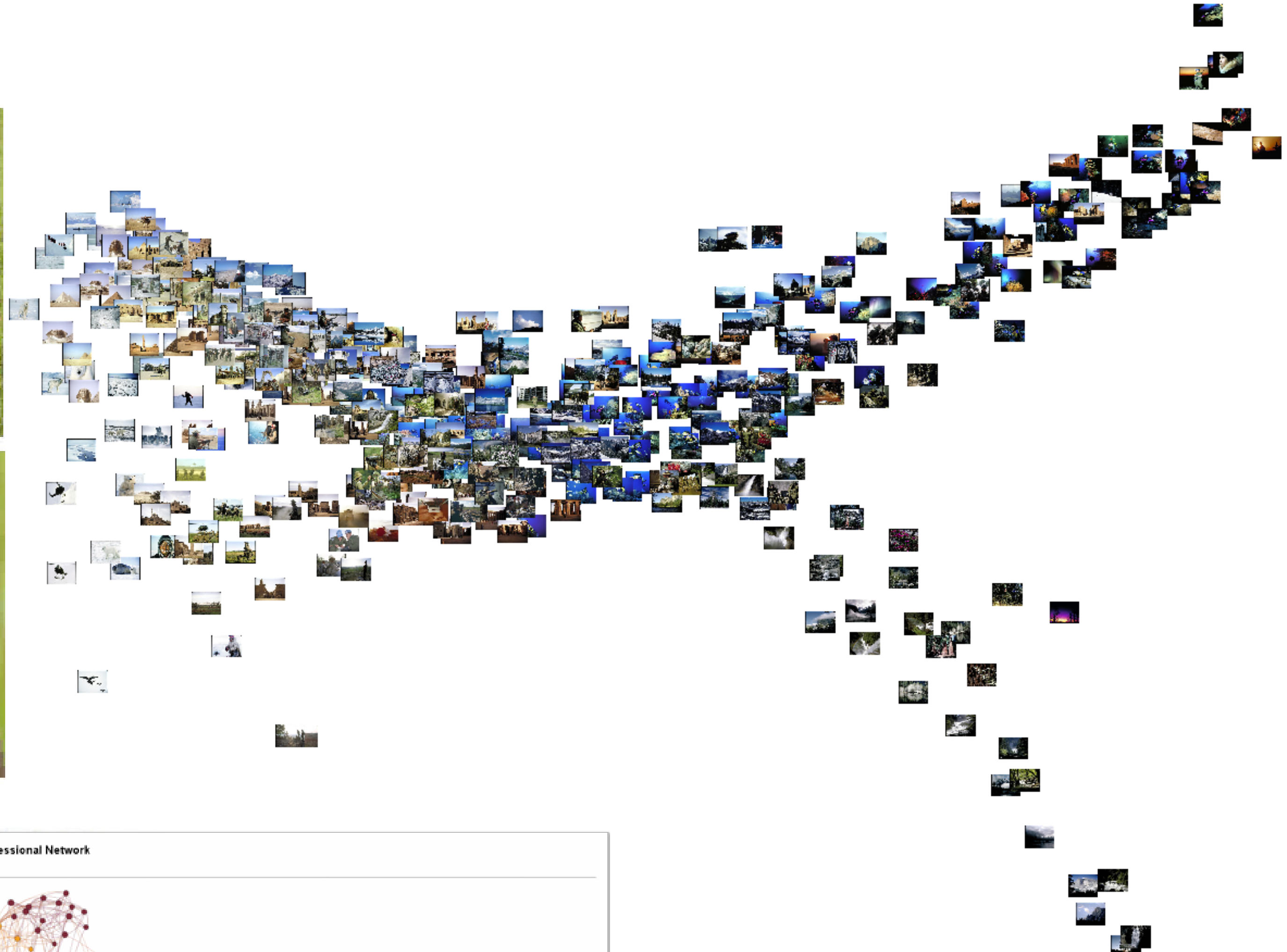


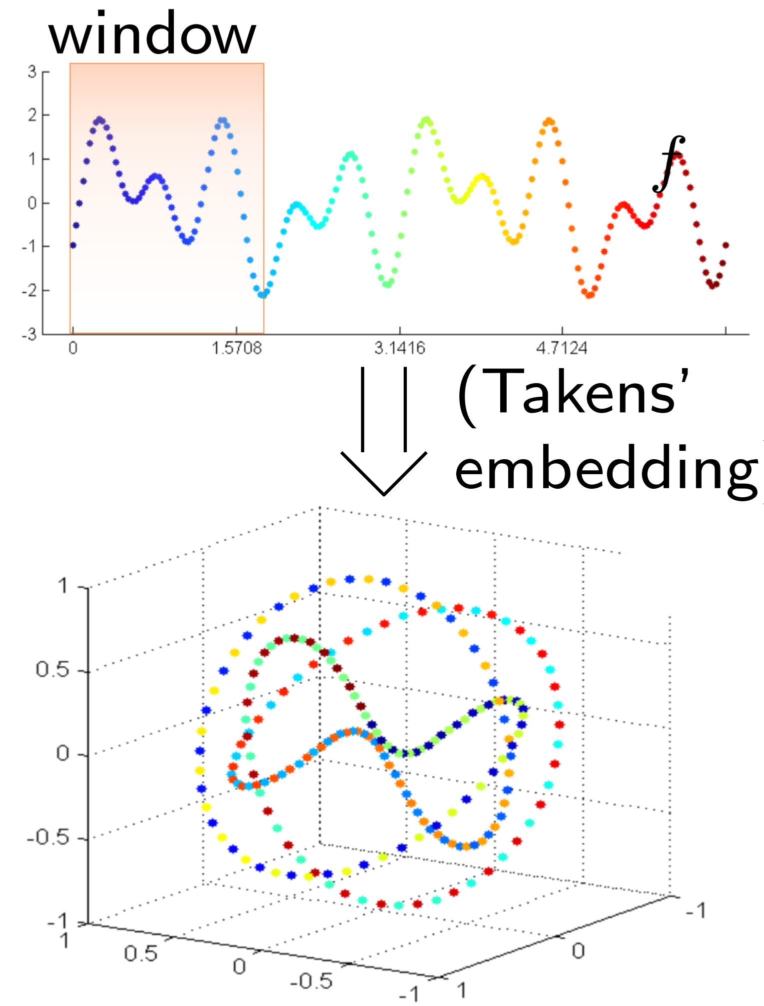
Dimensionality

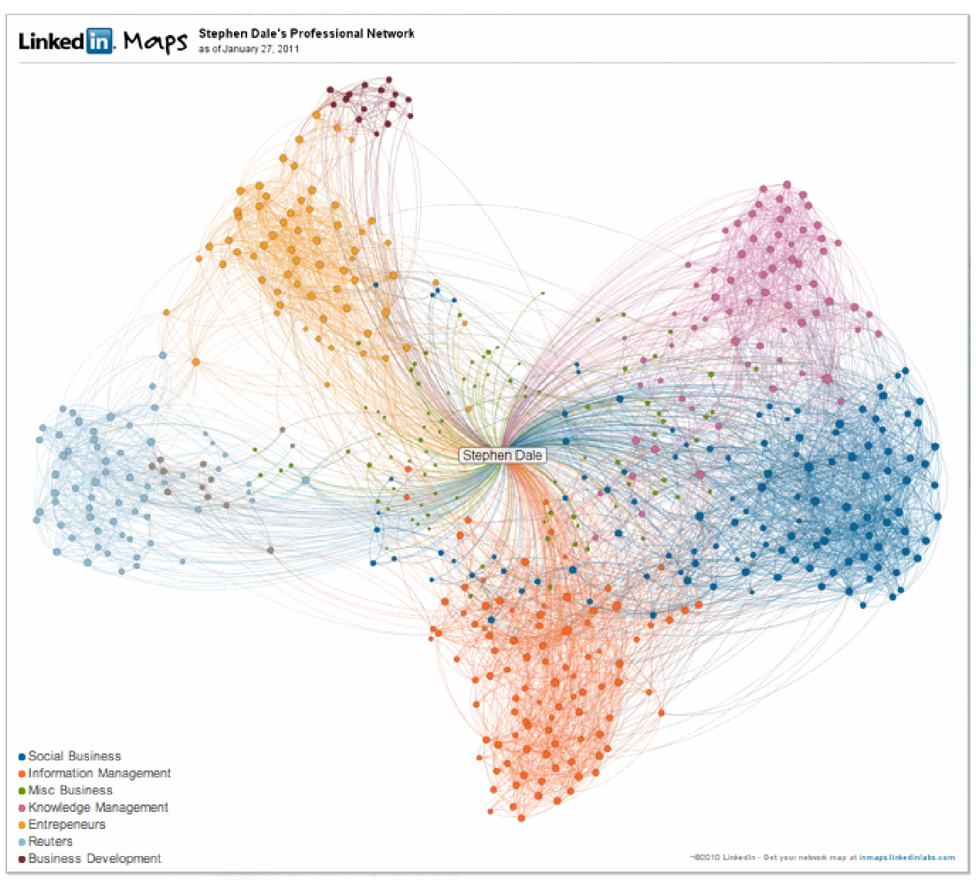
source de l'image : `https://stats.stackexchange.com`

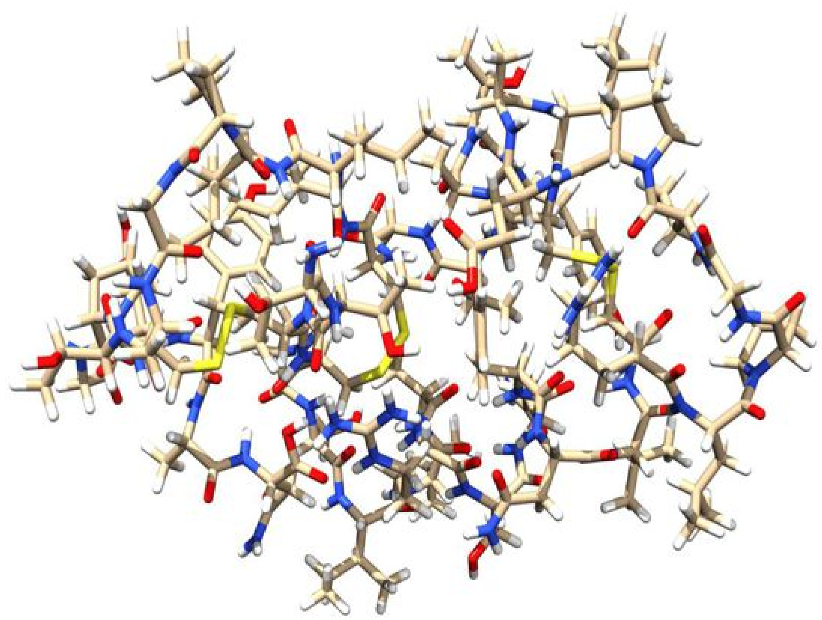
Remedy : reveal & exploit underlying structures

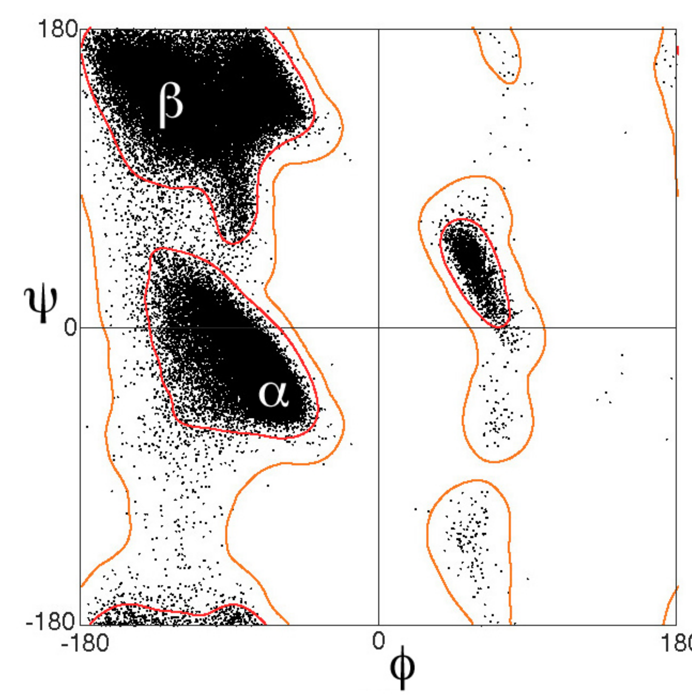




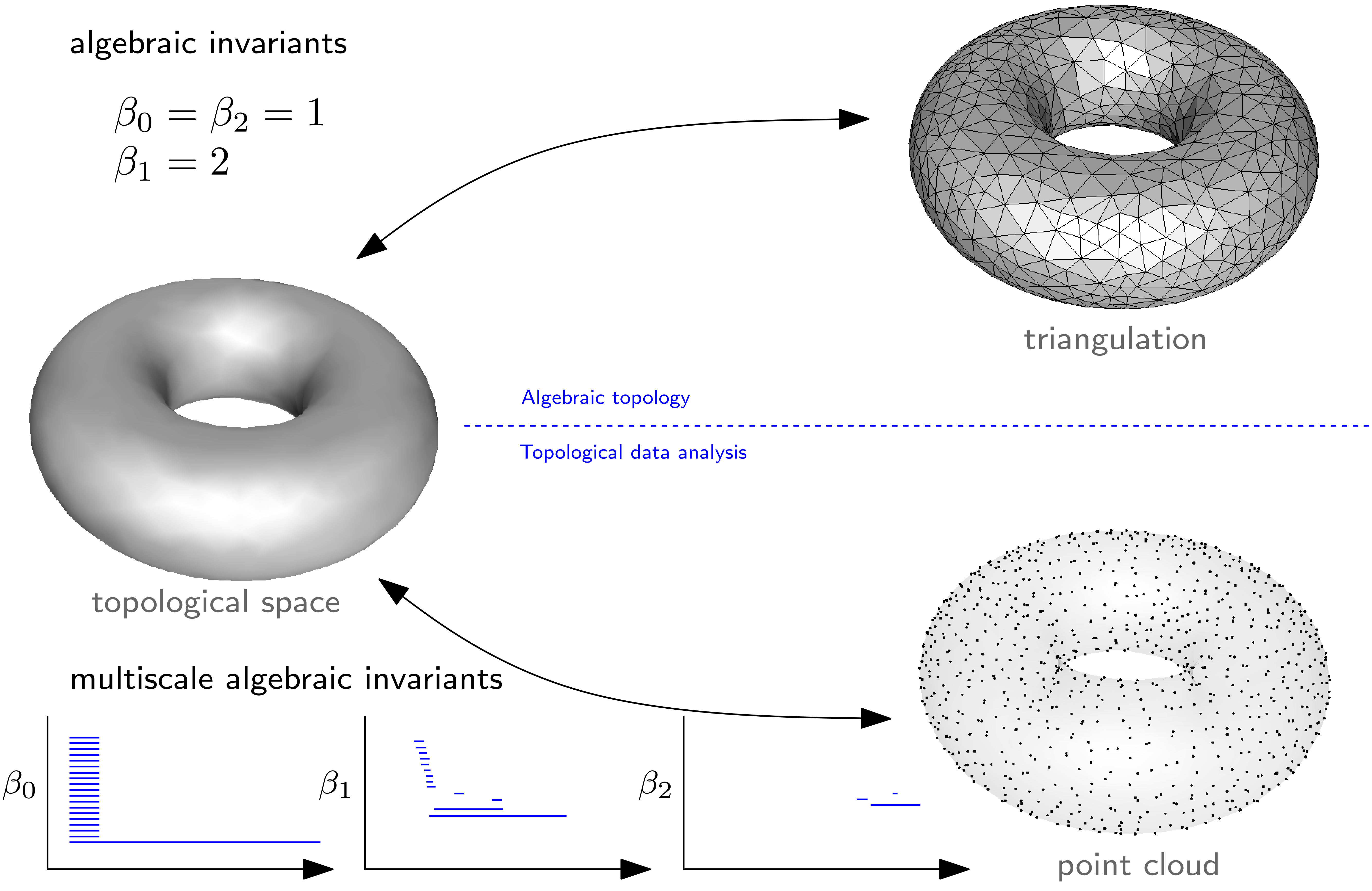






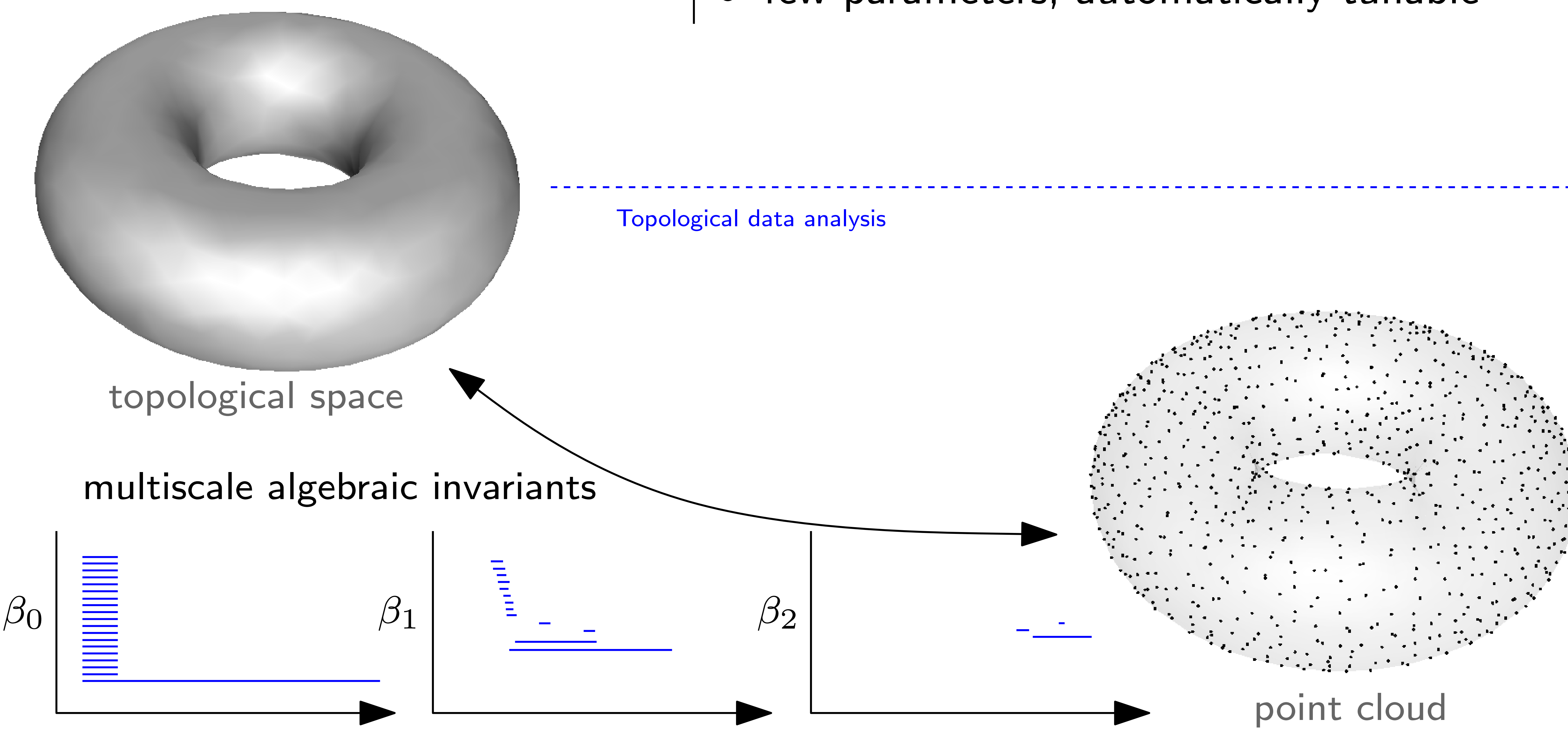


Enter topological data analysis (TDA)



Enter topological data analysis (TDA)

- invariance properties
- provable stability guarantees
- complementary to other descriptors
- few parameters, automatically tunable

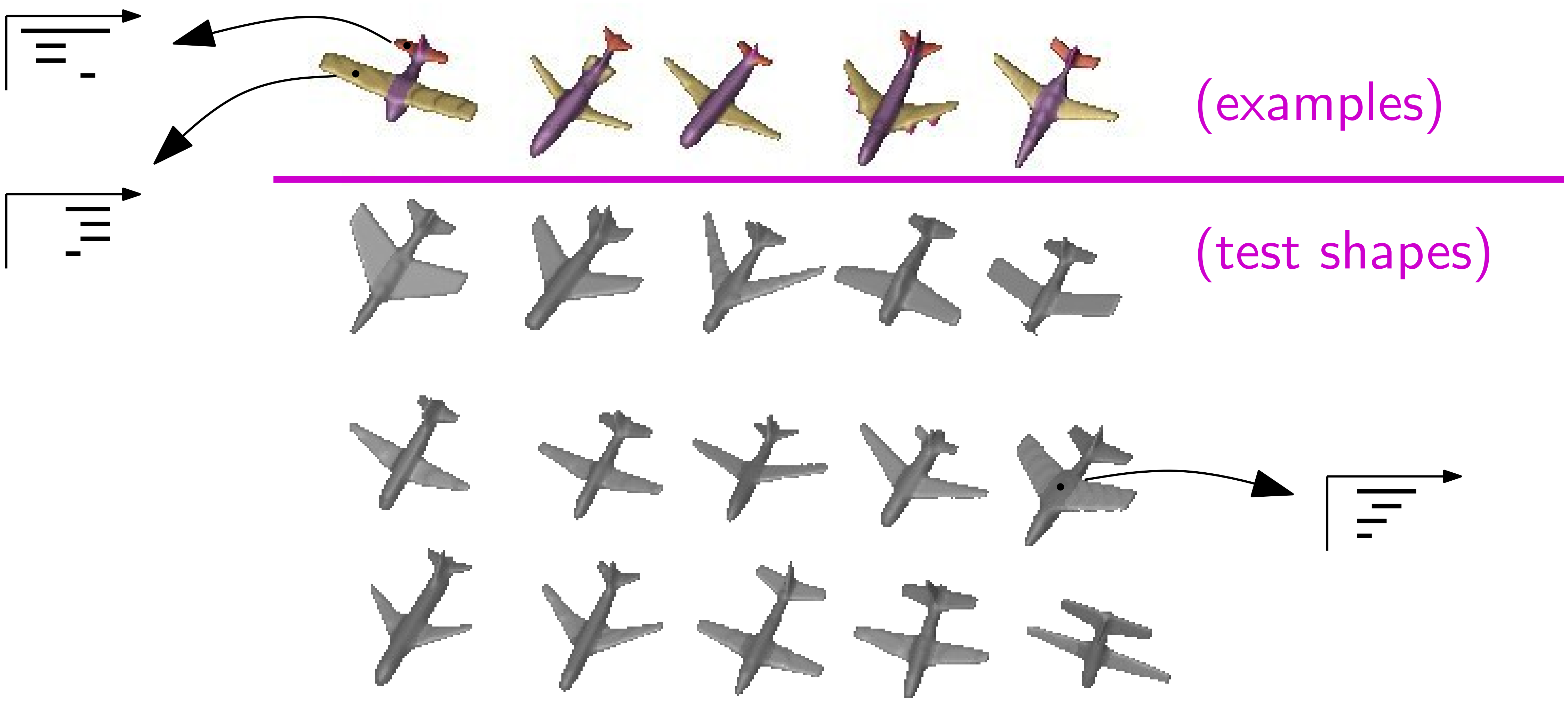


Example of application: shape segmentation

Task : segment 3d shapes from examples

Supervised learning approach:

- train a predictor on descriptors extracted from the examples
- apply the predictor to the descriptors extracted from the test shapes



[M. Carrière, S. Oudot, M. Ovsjanikov: Stable Topological Signatures for Points on 3D Shapes, *Proc. SGP*, 2015]

Example of application: shape segmentation

Task : segment 3d shapes from examples

Supervised learning approach:

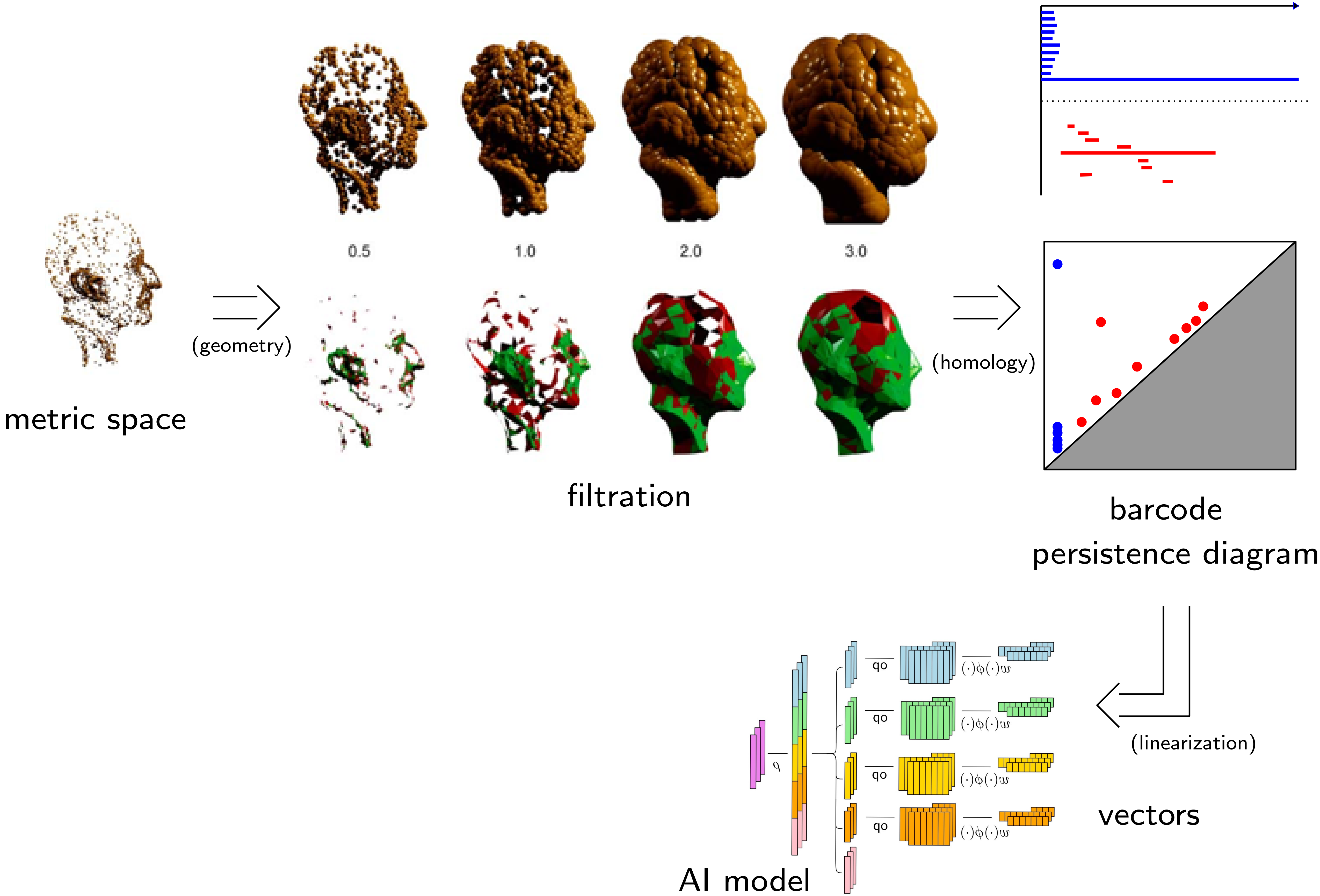
- train a predictor on descriptors extracted from the examples
- apply the predictor to the descriptors extracted from the test shapes

Error rates (%):

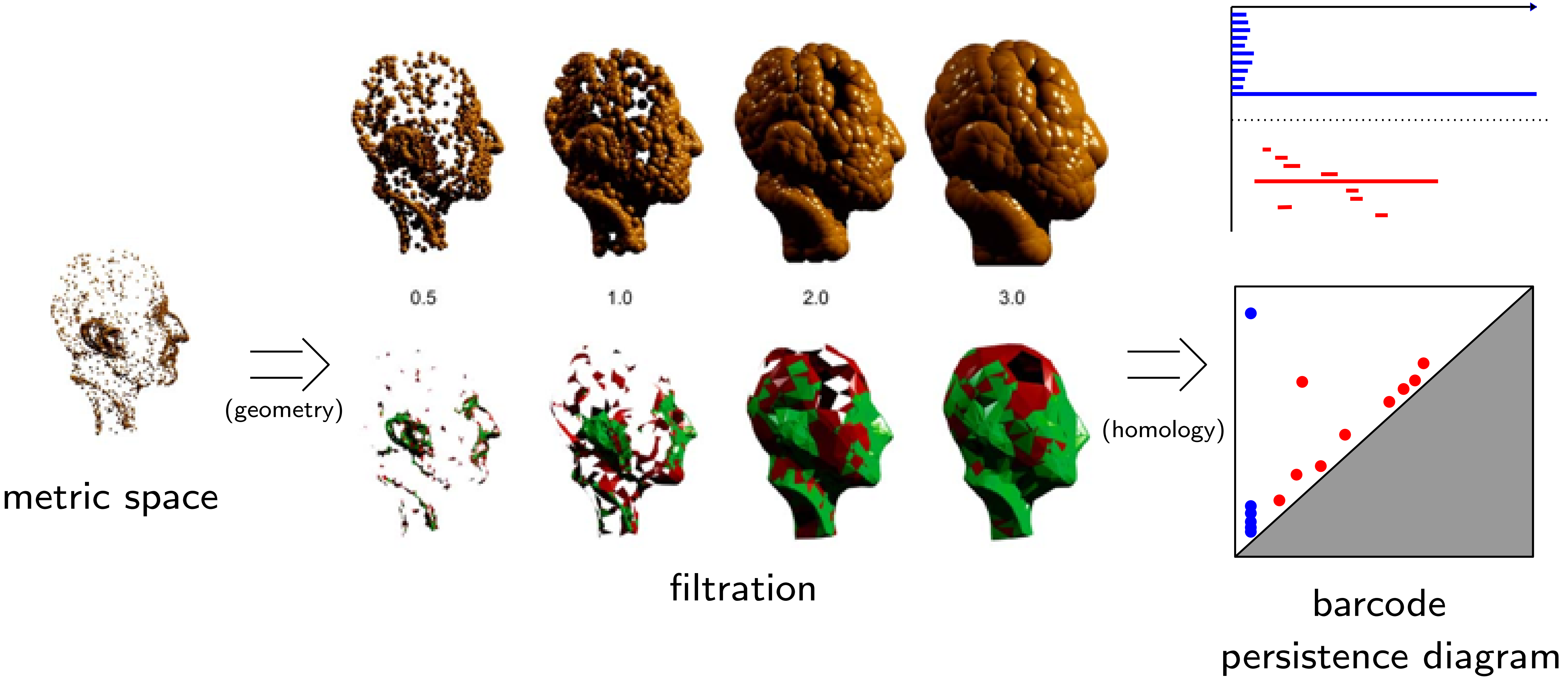
	geom/stat	TDA	geom/stat + TDA
Humanoids	21.3	26.0	**11.3**
Airplanes	18.7	27.4	**9.3**
Ants	9.7	7.7	**1.5**
Four legs	25.6	27.0	**15.8**
Octopuses	5.5	14.8	**3.4**
Birds	24.8	28.0	**13.5**
Fish	20.9	20.4	**7.7**

[M. Carrière, S. Oudot, M. Ovsjanikov: Stable Topological Signatures for Points on 3D Shapes, *Proc. SGP*, 2015]

The TDA pipeline in a nutshell



The TDA pipeline in a nutshell



4 foundational pillars (persistence theory):

- decomposition theorems ($\exists$ barcodes)
- algorithms (barcodes calculation)
- stability theorems (barcodes as descriptors)
- connections to other domains (geometry / analysis / statistics / AI)

Course outline

- Session 1: introduction to persistence theory + lab
- Sessions 2-3: homology theory + exercises
- Session 4-5: persistence theory + lab (**graded**)
- Session 5: topological inference + lab
- Session 6: topological descriptors + lab
- Session 7: learning with topological descriptors + exercises
- Session 8: statistics with topological descriptors + exercises
- Session 9: Reeb graphs and Mapper + lab
- **Final written exam**

Historical view

- origins : Morse theory (1930's-1940's)
  - *rank* $\equiv$ interval
  - *span* $\equiv$ length / persistence

Annals of Mathematics
Vol. 41, No. 2, April, 1940

RANK AND SPAN IN FUNCTIONAL TOPOLOGY

By Marston Morse

(Received August 9, 1939)

1. Introduction.

The analysis of functions F on metric spaces M of the type which appear in variational theories is made difficult by the fact that the critical limits, such as absolute minima, relative minima, minimax values etc., are in general infinite in number. These limits are associated with relative k-cycles of various dimensions and are classified as 0-limits, 1-limits etc. The number of k-limits suitably counted is called the k^{th} type number m_k of F. The theory seeks to establish relations between the numbers m_k and the connectivities p_k of M. The numbers p_k are finite in the most important applications. It is otherwise with the numbers m_k .

Historical view

- origins : Morse theory (1930's-1940's)
  - *rank* ≡ interval
  - *span* ≡ length / persistence

- subsequent echoes (1990's)
  - size theory [Frosini et al.]
  - canonical forms [Barannikov]
  - persistent Betti numbers [Robins]

ANNALS OF MATHEMATICS
Vol. 41, No. 2, April, 1940

RANK AND SPAN IN FUNCTIONAL TOPOLOGY

BY MARSTON MORSE

(Received August 9, 1939)

1. Introduction.

The analysis of functions F on metric spaces M of the type which appear in variational theories is made difficult by the fact that the critical limits, such as absolute minima, relative minima, minimax values etc., are in general infinite in number. These limits are associated with relative k-cycles of various dimensions and are classified as 0-limits, 1-limits etc. The number of k-limits suitably counted is called the k^{th} type number m_k of F. The theory seeks to establish relations between the numbers m_k and the connectivities p_k of M. The numbers p_k are finite in the most important applications. It is otherwise with the numbers m_k .

Historical view

- origins : Morse theory (1930's-1940's)
  - *rank* $\equiv$ interval
  - *span* $\equiv$ length / persistence

- subsequent echoes (1990's)
  - size theory [Frosini et al.]
  - canonical forms [Barannikov]
  - persistent Betti numbers [Robins]

- persistence theory (2000's-)
  - persistence algorithm [Edelsbrunner, Letscher, Zomorodian]
  - persistence modules [Carlsson, Zomorodian]
  - interleaving distance and stability [Cohen-Steiner, Edelsbrunner, Harer] [Chazal, Cohen-Steiner, Glisse, Guibas, Oudot] [Lesnick]

ANNALS OF MATHEMATICS
Vol. 41, No. 2, April, 1940

RANK AND SPAN IN FUNCTIONAL TOPOLOGY

BY MARSTON MORSE

(Received August 9, 1939)

1. Introduction.

The analysis of functions F on metric spaces M of the type which appear in variational theories is made difficult by the fact that the critical limits, such as absolute minima, relative minima, minimax values etc., are in general infinite in number. These limits are associated with relative k-cycles of various dimensions and are classified as 0-limits, 1-limits etc. The number of k-limits suitably counted is called the k^{th} type number m_k of F. The theory seeks to establish relations between the numbers m_k and the connectivities p_k of M. The numbers p_k are finite in the most important applications. It is otherwise with the numbers m_k .

Historical view

- origins : Morse theory (1930's-1940's)
  - *rank* $\equiv$ interval
  - *span* $\equiv$ length / persistence

ANNALS OF MATHEMATICS
Vol. 41, No. 2, April, 1940

RANK AND SPAN IN FUNCTIONAL TOPOLOGY

BY MARSTON MORSE

(Received August 9, 1939)

1. Introduction.

The analysis of functions F on metric spaces M of the type which appear in variational theories is made difficult by the fact that the critical limits, such as absolute minima, relative minima, minimax values etc., are in general infinite in number. These limits are associated with relative k-cycles of various dimensions and are classified as 0-limits, 1-limits etc. The number of k-limits suitably counted is called the k^{th} type number m_k of F. The theory seeks to establish relations between the numbers m_k and the connectivities p_k of M. The numbers p_k are finite in the most important applications. It is otherwise with the numbers m_k.

- subsequent echoes (1990's)
  - size theory [Frosini et al.]
  - canonical forms [Barannikov]
  - persistent Betti numbers [Robins]

- persistence theory (2000's-)
  - persistence algorithm [Edelsbrunner, Letscher, Zomorodian]
  - persistence modules [Carlsson, Zomorodian]
  - interleaving distance and stability [Cohen-Steiner, Edelsbrunner, Harer] [Chazal, Cohen-Steiner, Glisse, Guibas, Oudot] [Lesnick]
  - connection to representation theory [Carlsson, de Silva] [Crawley-Boevey]
  - connection to sheaf theory [Curry] [Kashiwara, Shapira]
  - connection to symplectic geometry [Polterovich, Shelukhin et al.]
  - connection to statistical inference [Wasserman et al.] [Chazal et al.]
  - connection to optimal transport [Carrière, Cuturi, Oudot] [Divol, Lacombe]