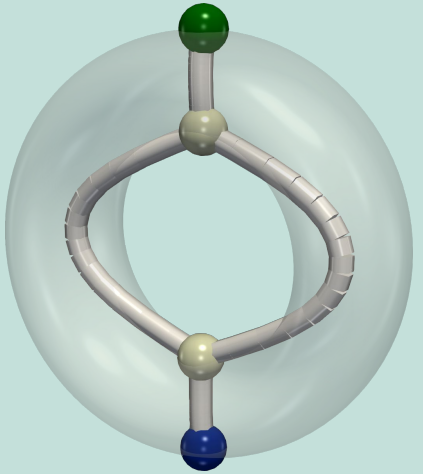
Topological Analysis of Massive Datasets



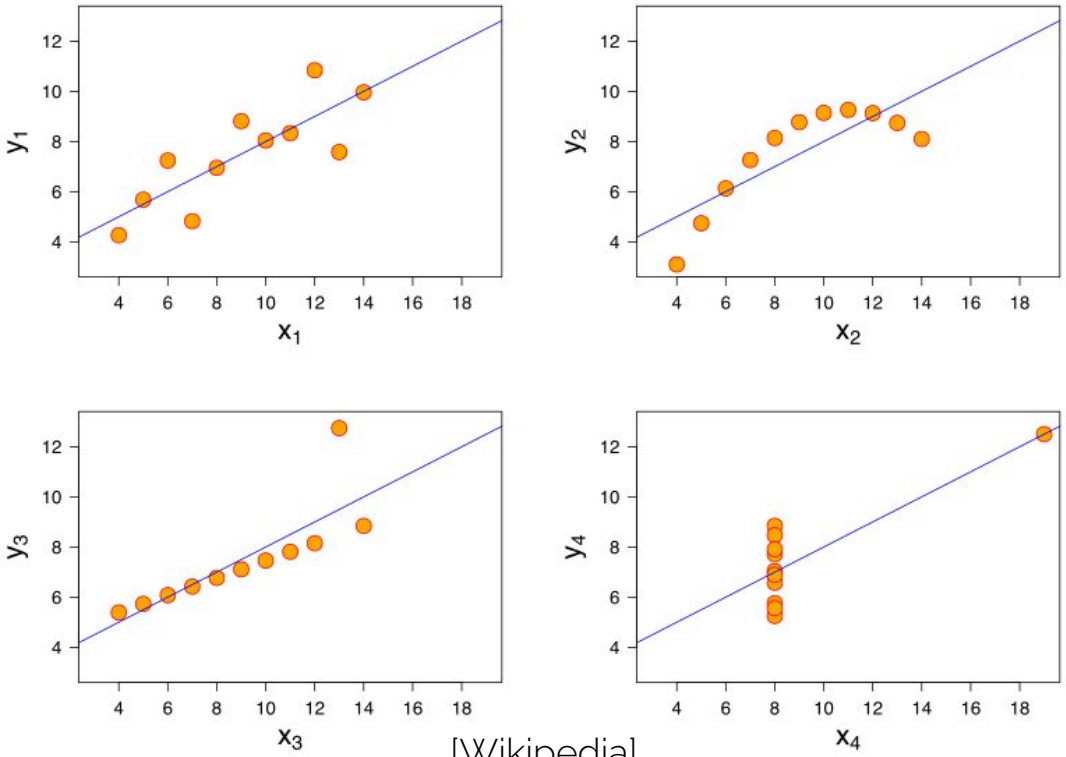
Julien Tierny

cnrs SORBONNE UNIVERSITÉ CRÉATEURS DE FUTURS DEPUIS 1257 LIP6 erc

Capturing the *“shape”* of data

Capturing the *"shape"* of data

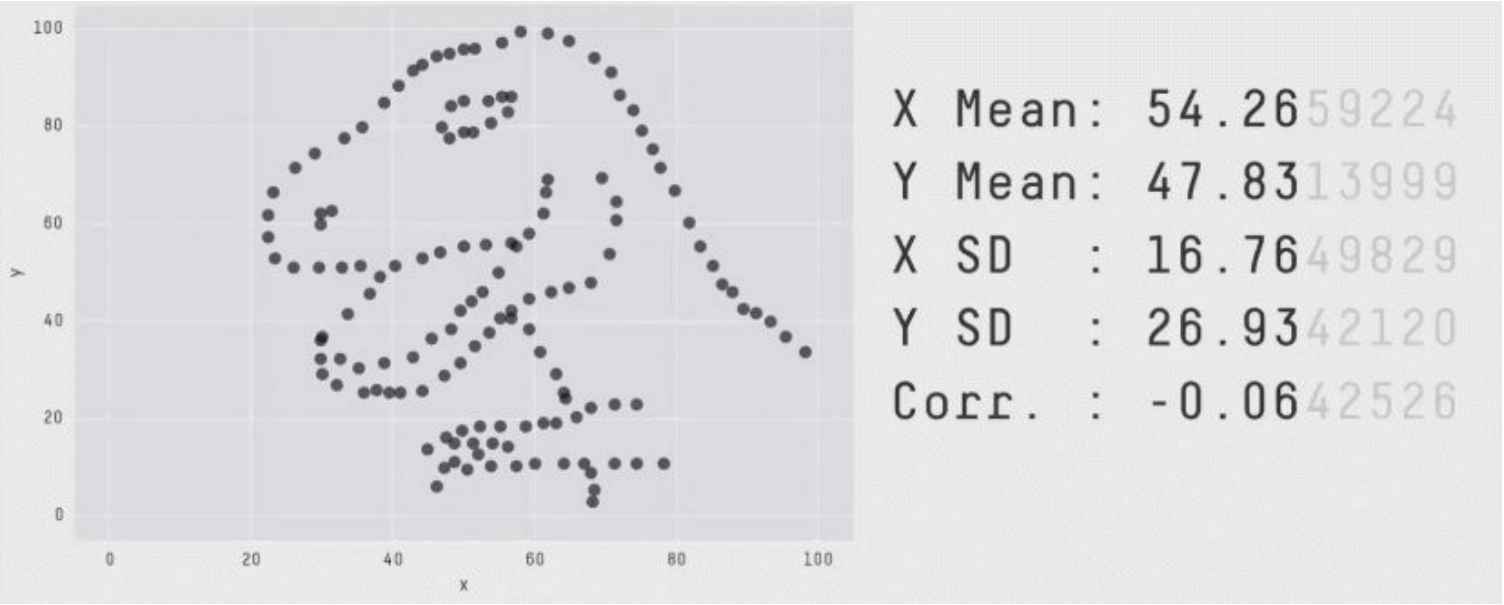
- Anscombe's quartet



[Wikipedia]

Capturing the *“shape”* of data

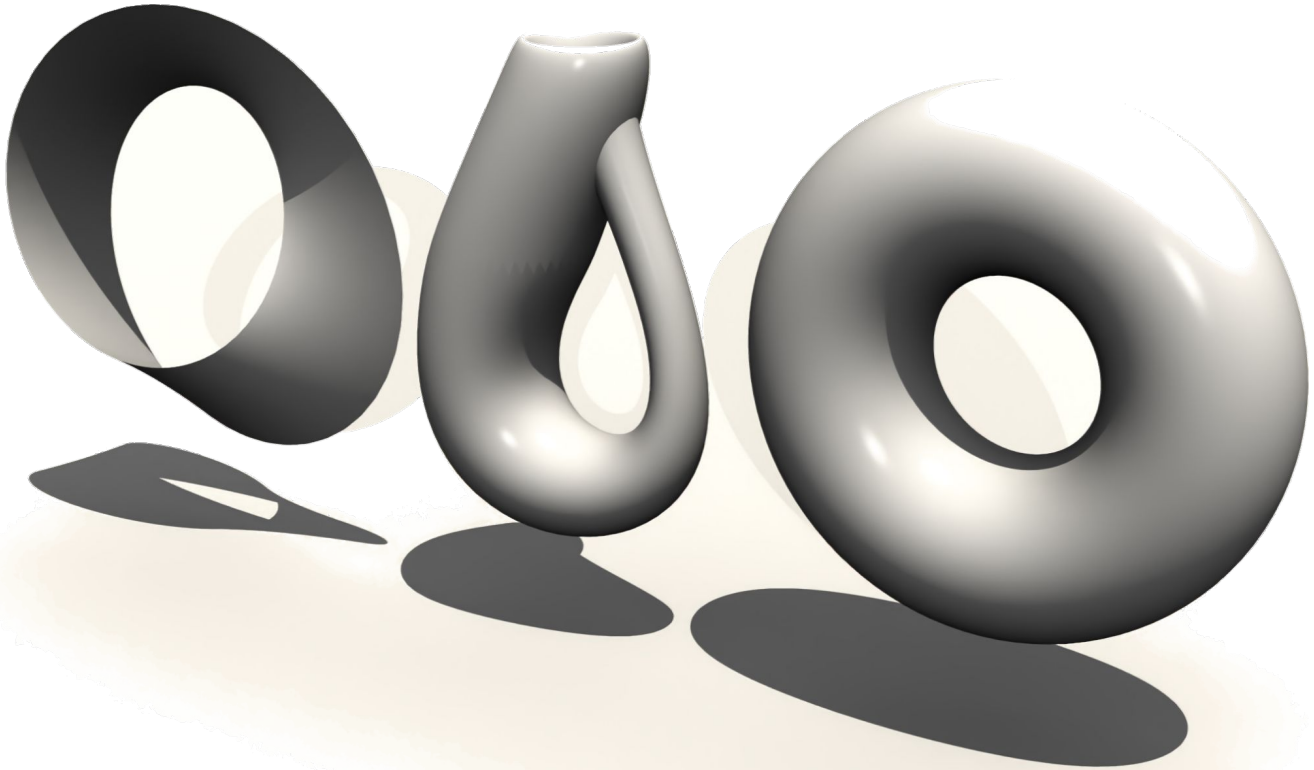
- Datasaurus Dozen



[Matejka et al. 2017]

Capturing the *"shape"* of data

- Topological invariants

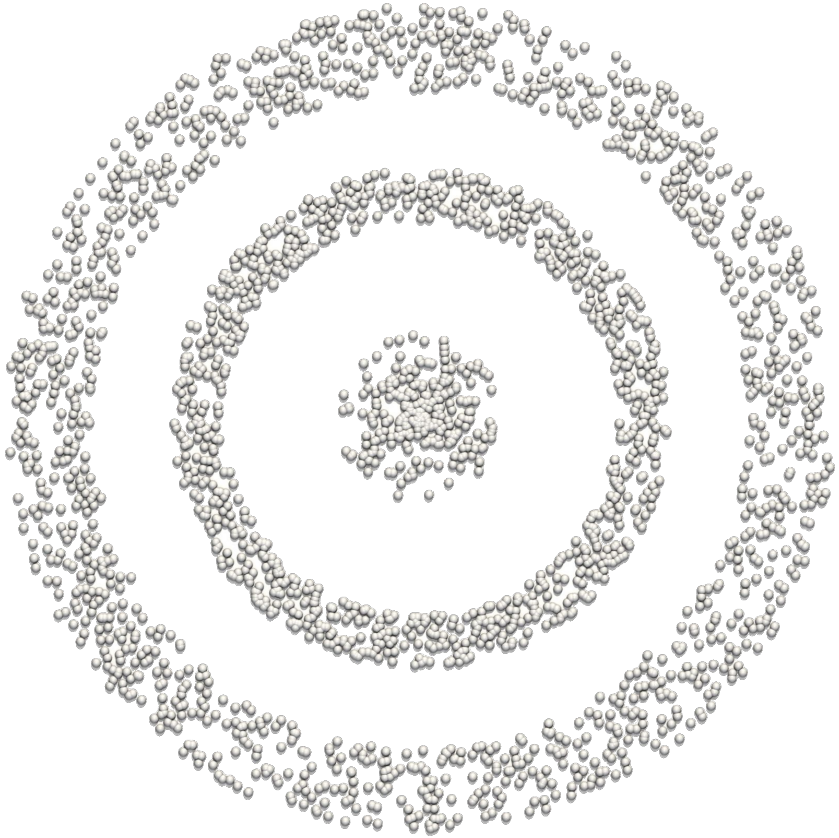


The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476

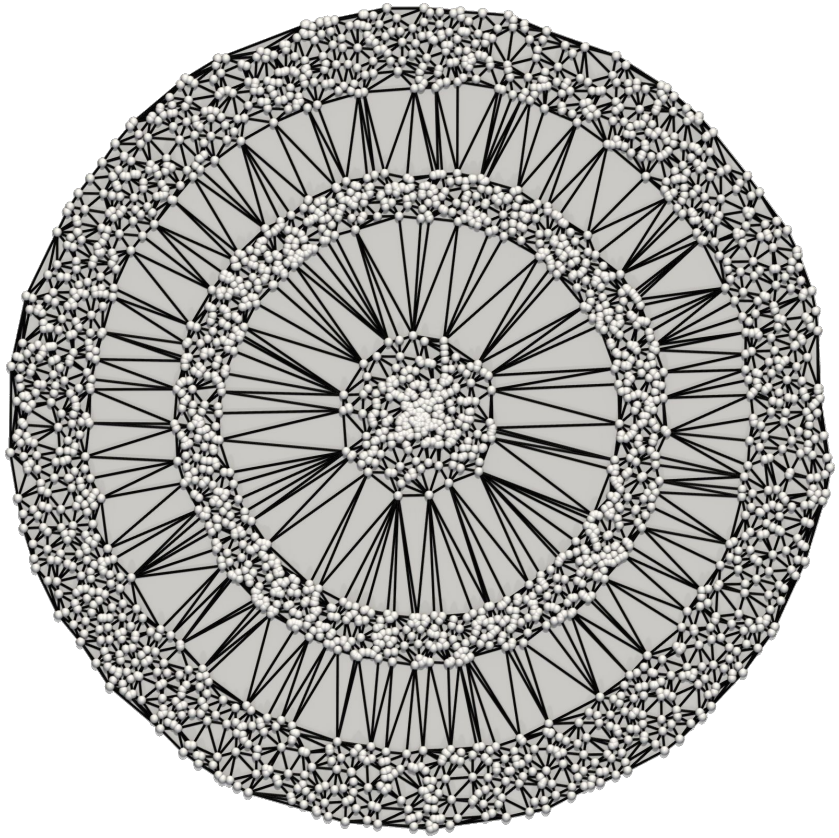
The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476



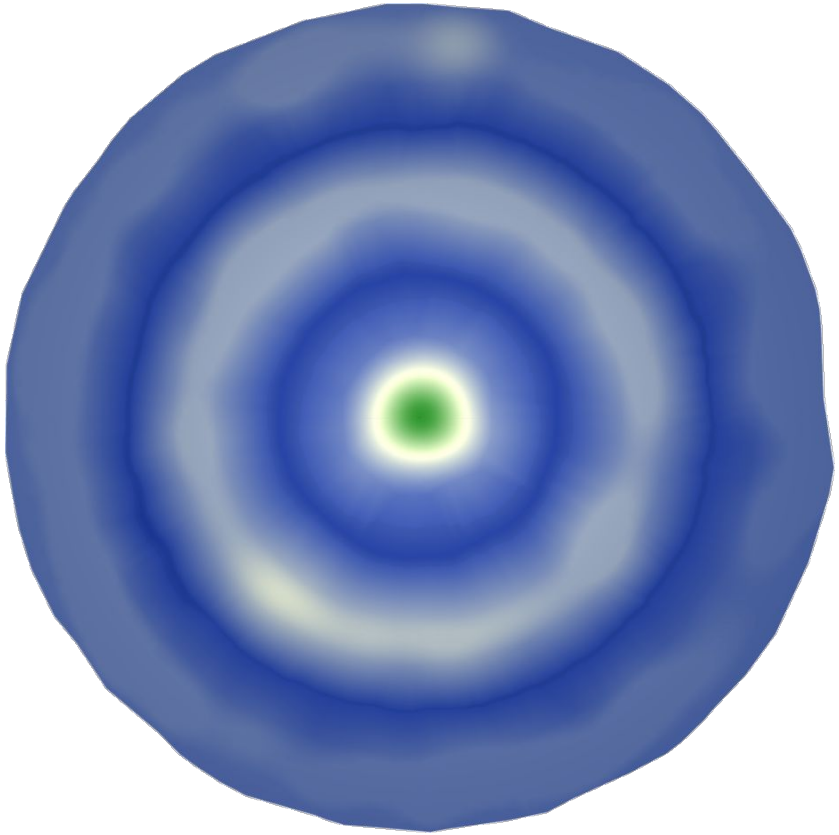
The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476



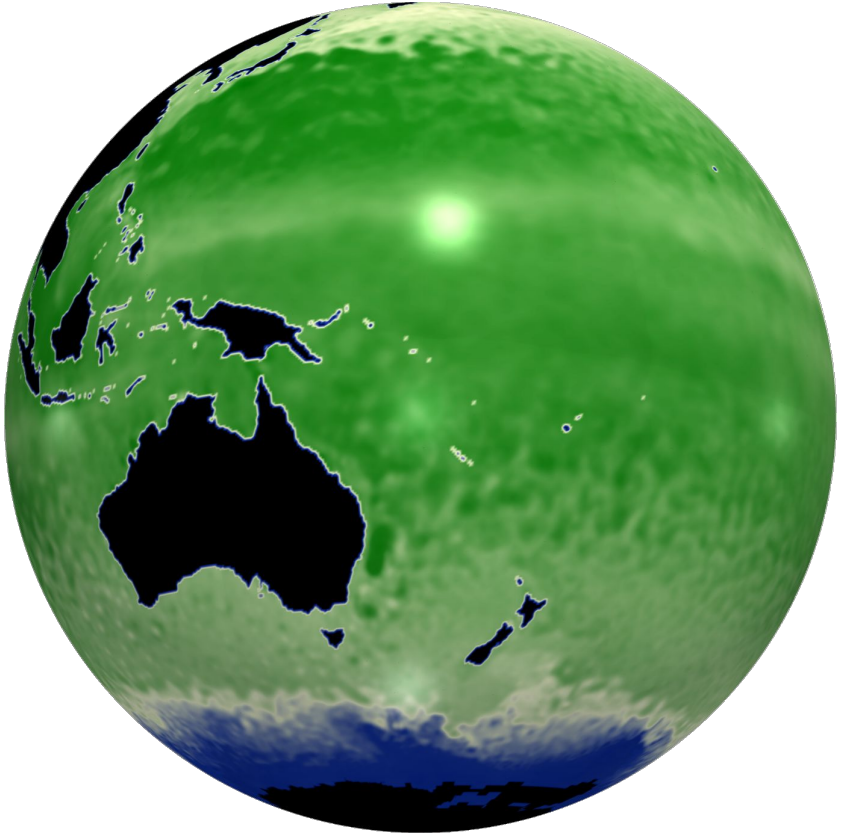
The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476

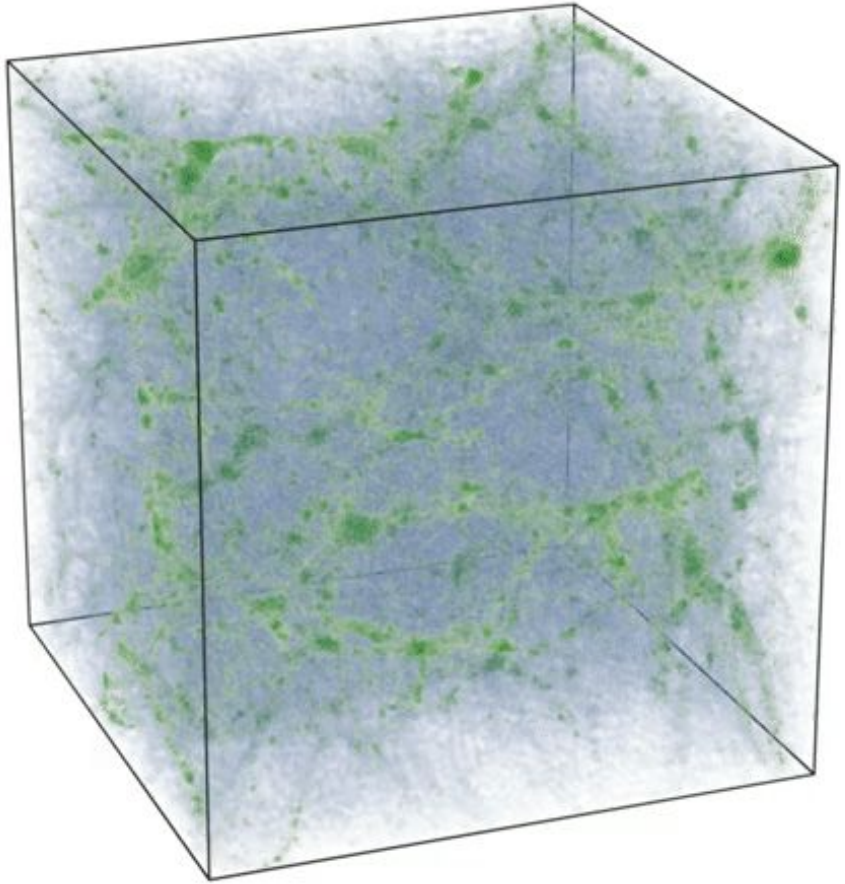


The topology of numerical data

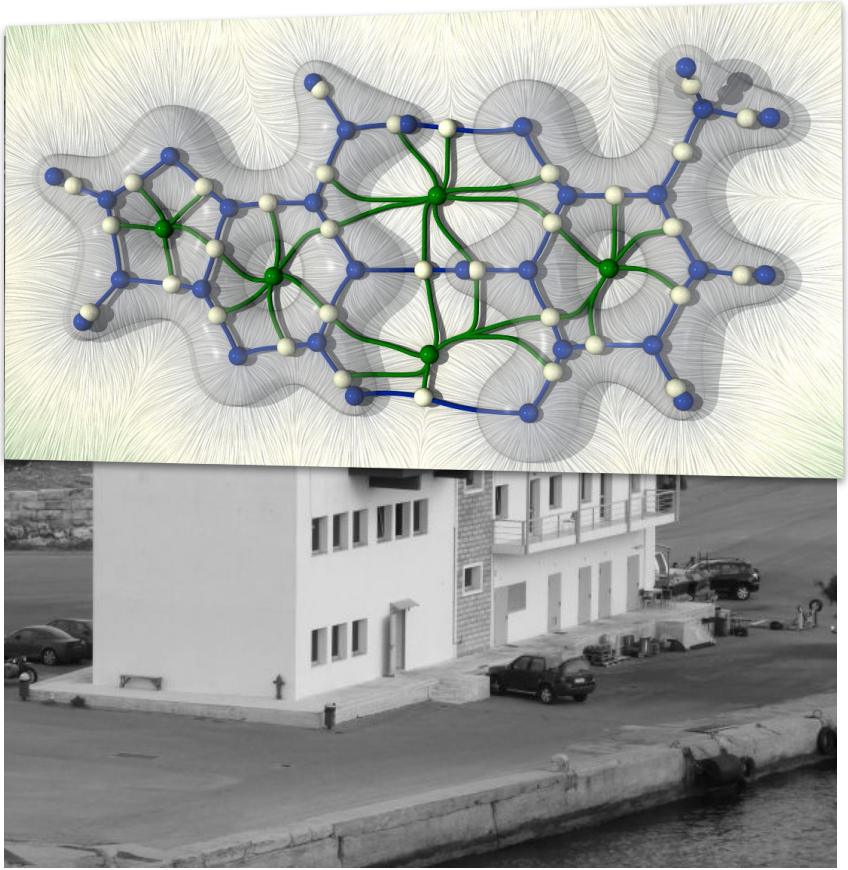
The topology of numerical data

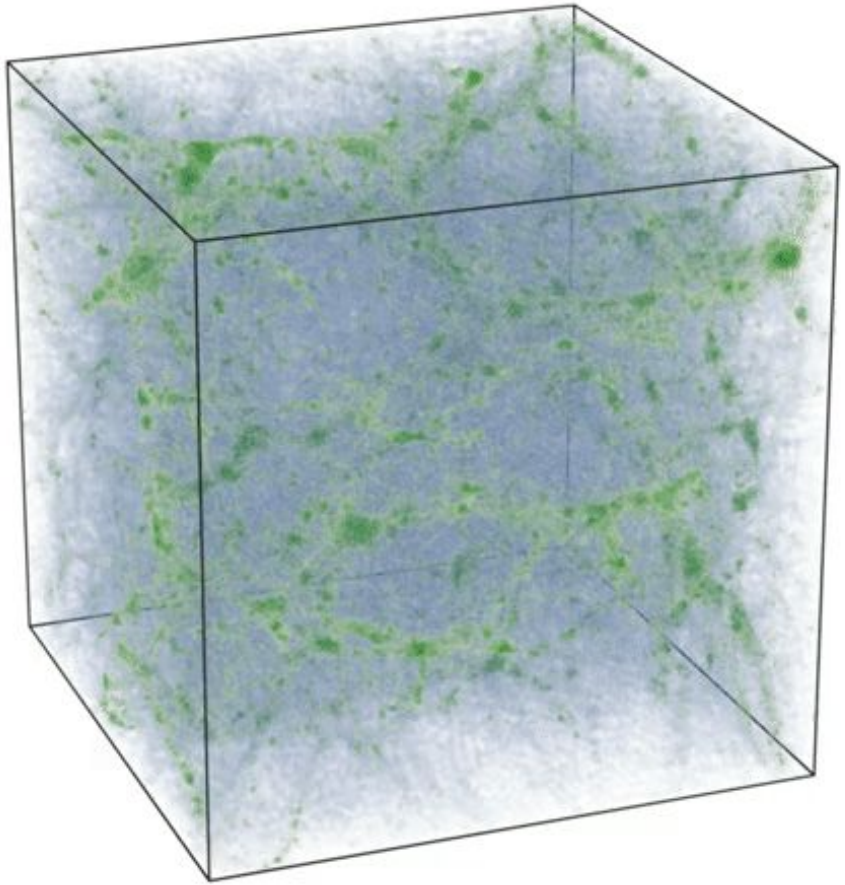


The topology of numerical data

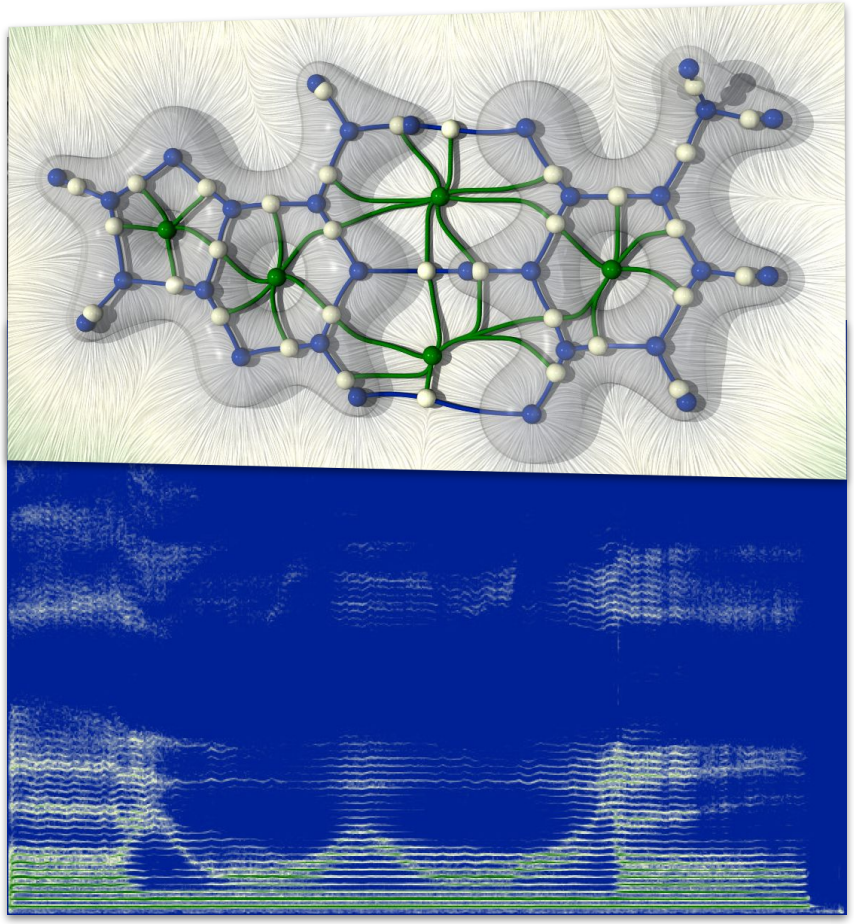


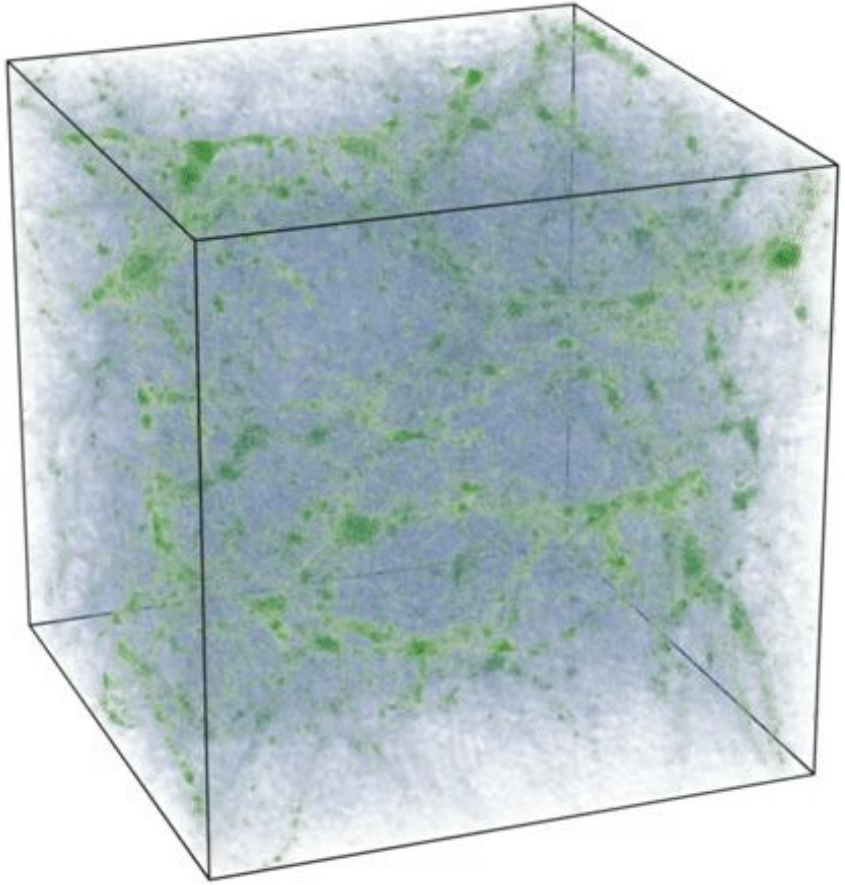
The topology of numerical data





The topology of numerical data



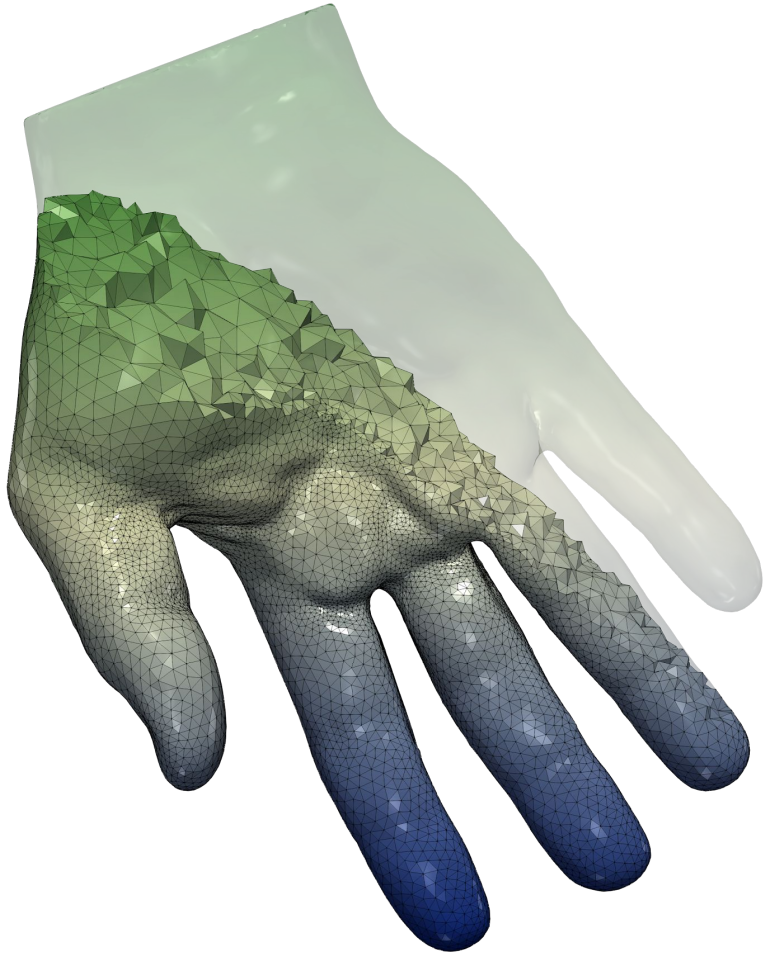


How does it work?

- Input PL scalar data
  - $f : \mathcal{M} \to \mathbb{R}$

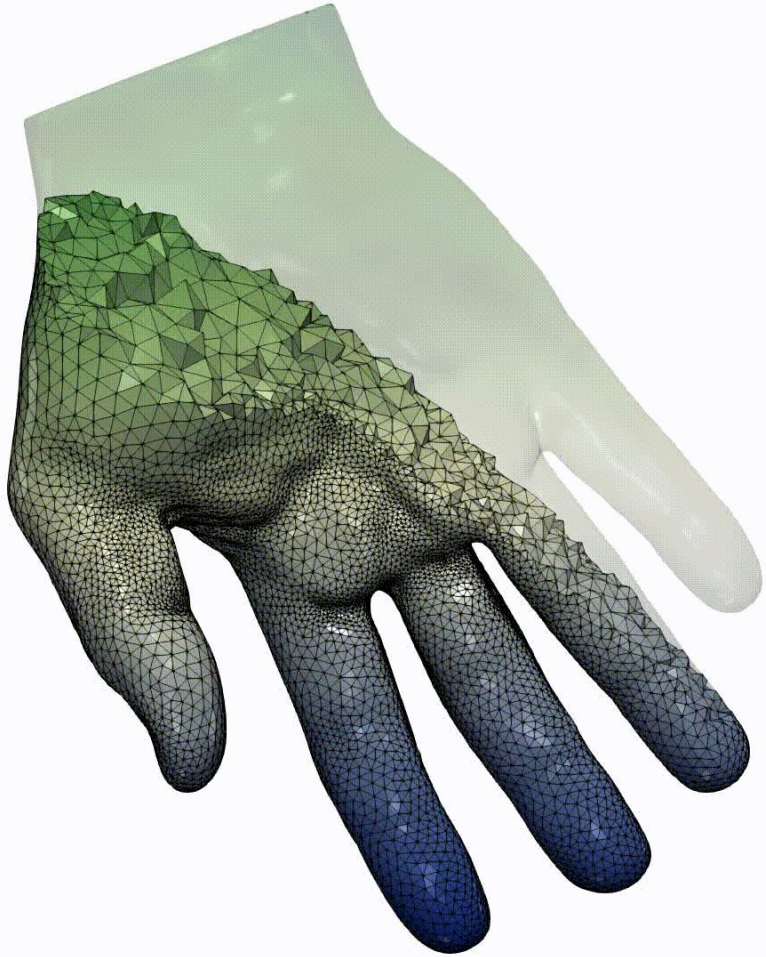
How does it work?

- **Input PL scalar data**
  - $f : \mathcal{M} \to \mathbb{R}$
  - $\mathcal{M}$: simplicial complex



How does it work?

- **Input PL scalar data**
  - $f : \mathcal{M} \to \mathbb{R}$
  - $\mathcal{M}$: simplicial complex
    - Triangulated surface
    - Tetrahedral volume
    - 2D pixel image
    - 3D voxel image
    - Point cloud data



How does it work?

- Input PL scalar data
  - $f : \mathcal{M} \to \mathbb{R}$

- Topological abstractions

How does it work?

- Input PL scalar data
  - $f : \mathcal{M} \to \mathbb{R}$

- Topological abstractions

How does it work?

- Input PL scalar data
  - $f : \mathcal{M} \to \mathbb{R}$

- Topological abstractions

How does it work?

- Input PL scalar data
  - $f : \mathcal{M} \to \mathbb{R}$

- Topological abstractions

How does it work?

- **Input PL scalar data**
  - $f : \mathcal{M} \to \mathbb{R}$

- **Topological abstractions**
  - Critical points

How does it work?

- **Input PL scalar data**
  - $f : \mathcal{M} \to \mathbb{R}$

- **Topological abstractions**
  - Critical points

How does it work?

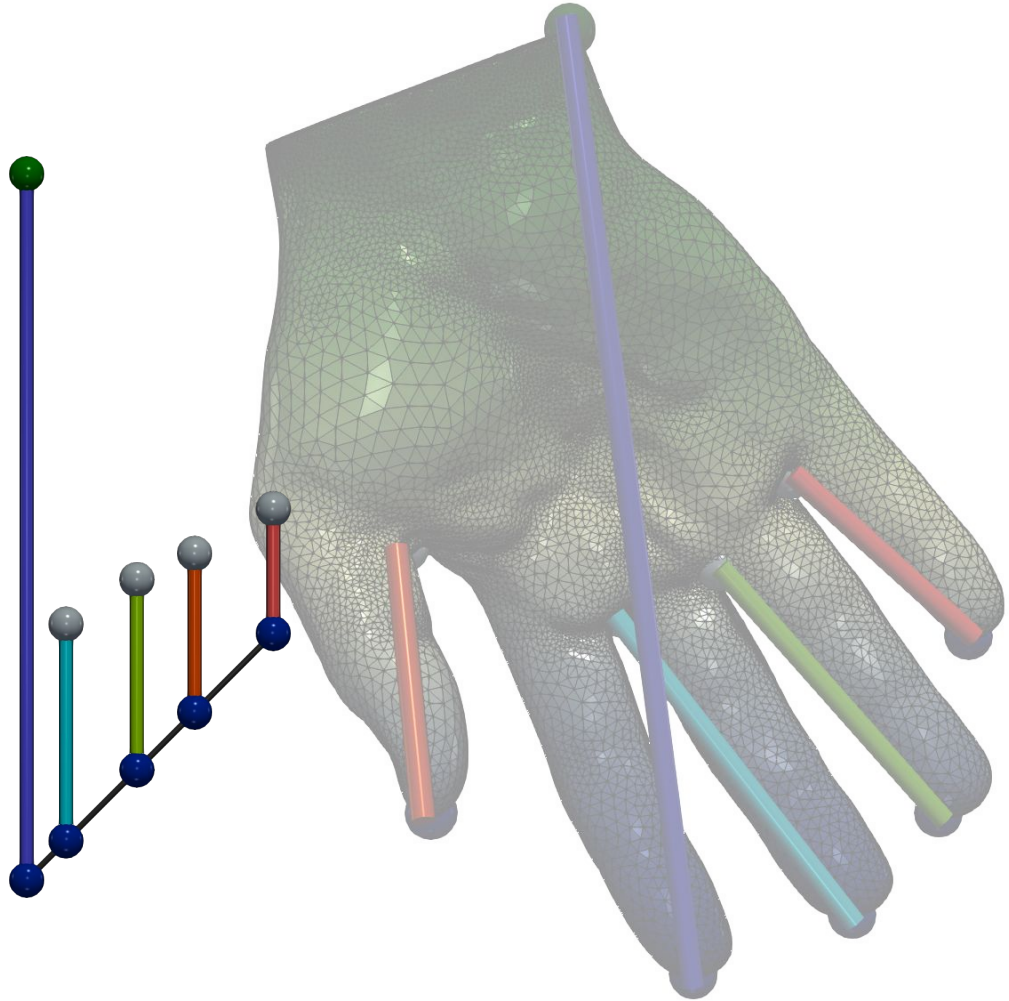
- **Input PL scalar data**
  - $f : \mathcal{M} \to \mathbb{R}$

- **Topological abstractions**
  - Critical points

How does it work?

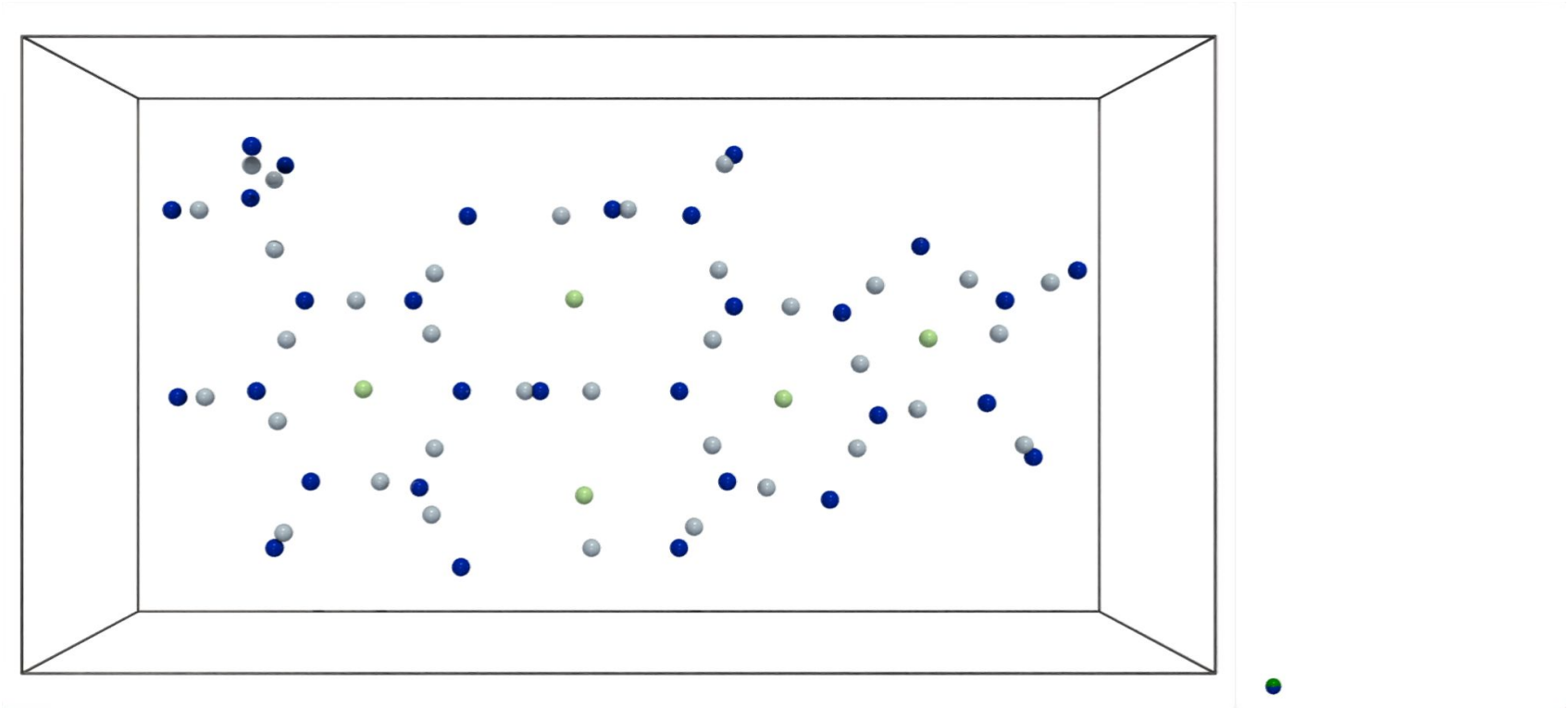
- **Input PL scalar data**
  - $f : \mathcal{M} \to \mathbb{R}$

- **Topological abstractions**
  - Critical points
  - Persistence diagram



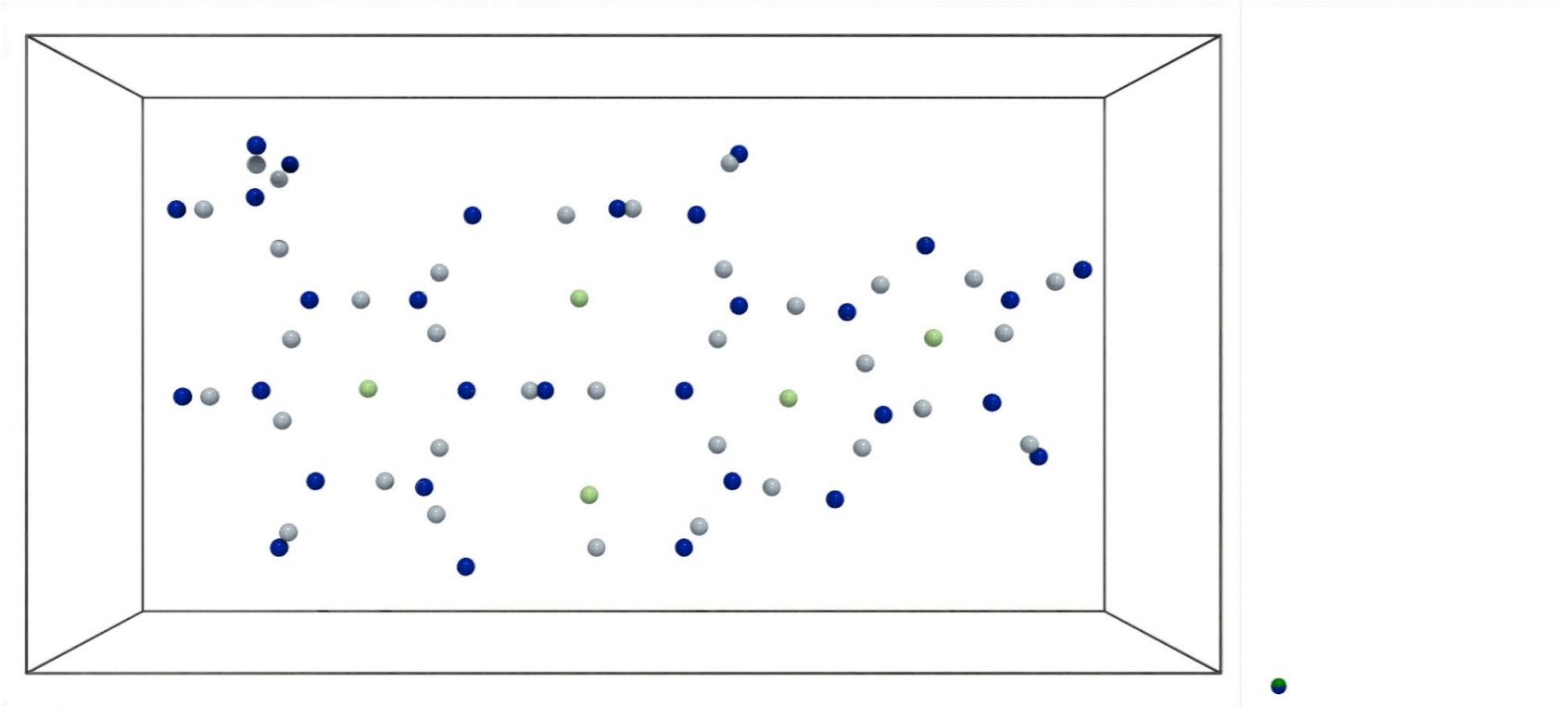
The persistence diagram

- Beyond connected components



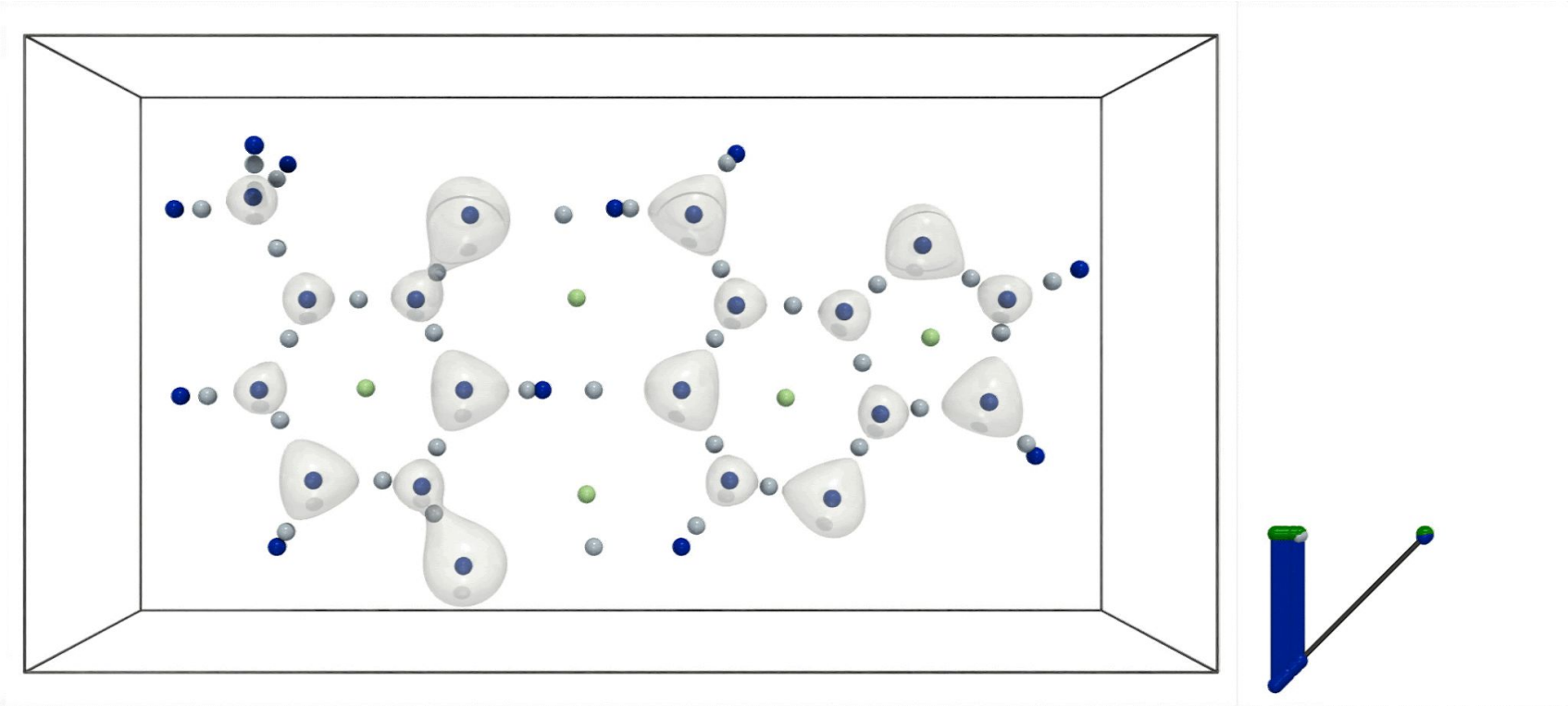
The persistence diagram

- Beyond connected components



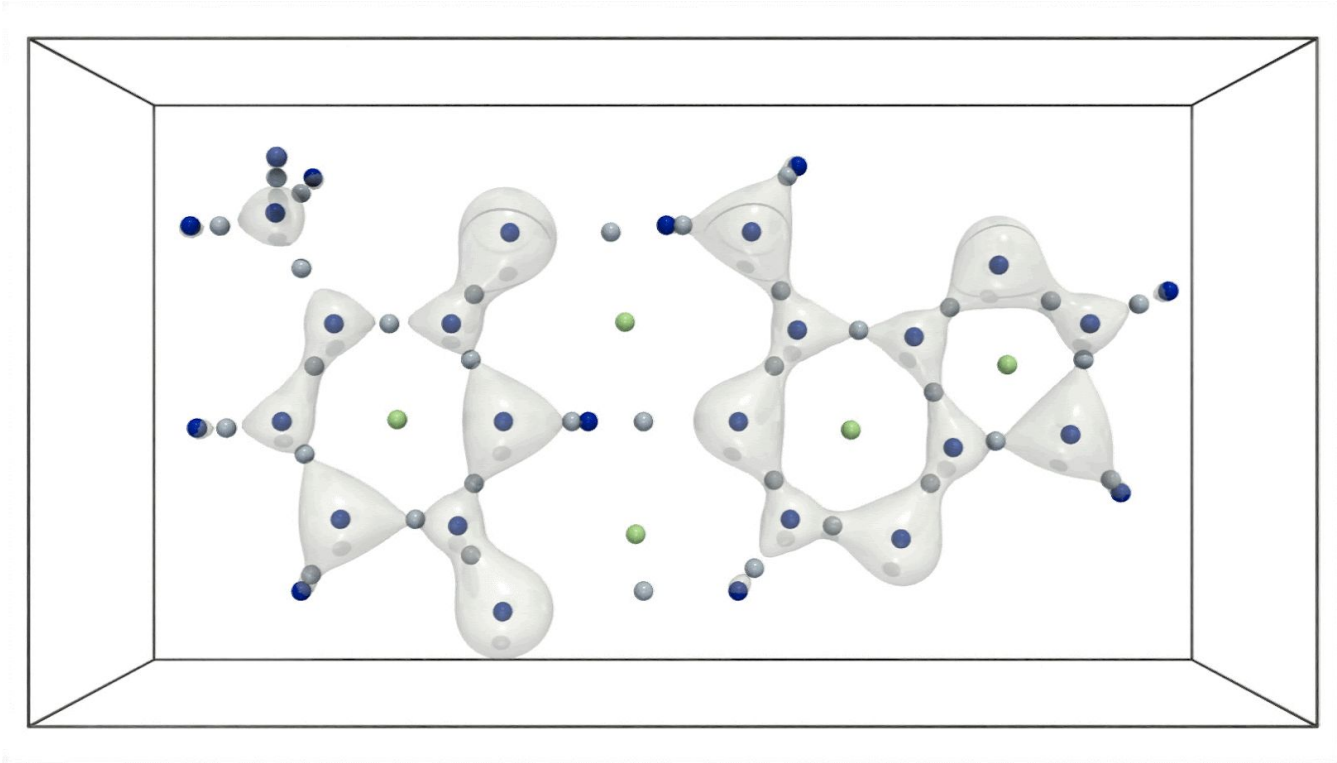
The persistence diagram

- Beyond connected components



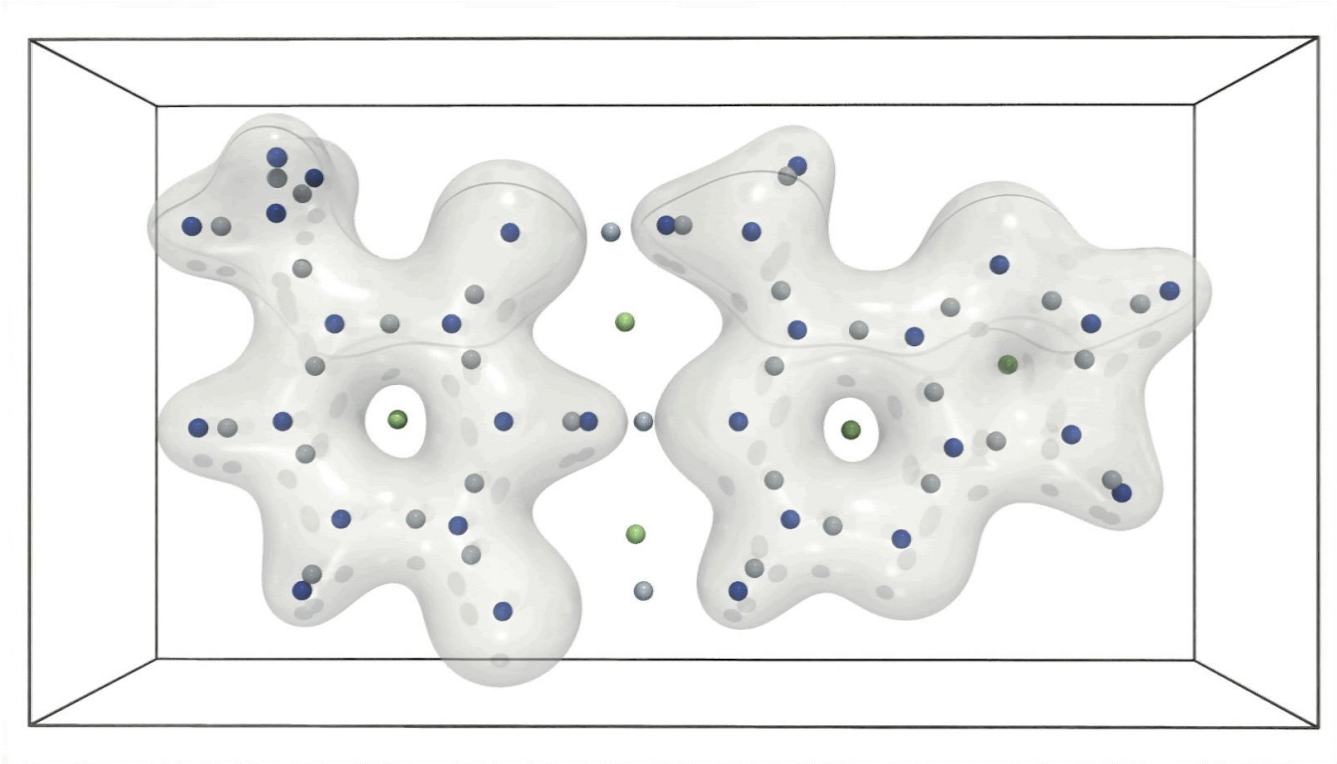
The persistence diagram

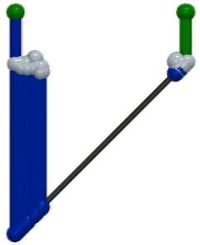
- Beyond connected components



The persistence diagram

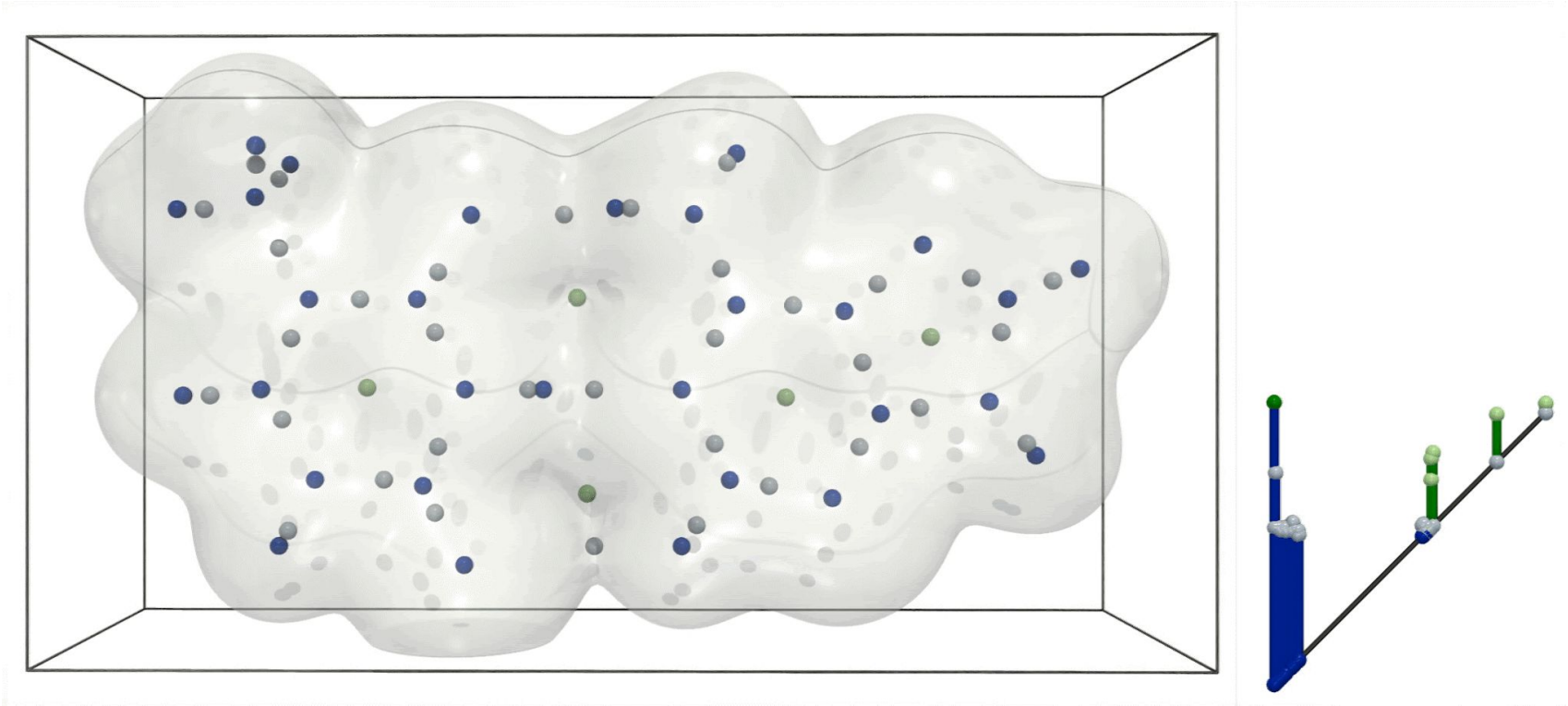
- Beyond connected components





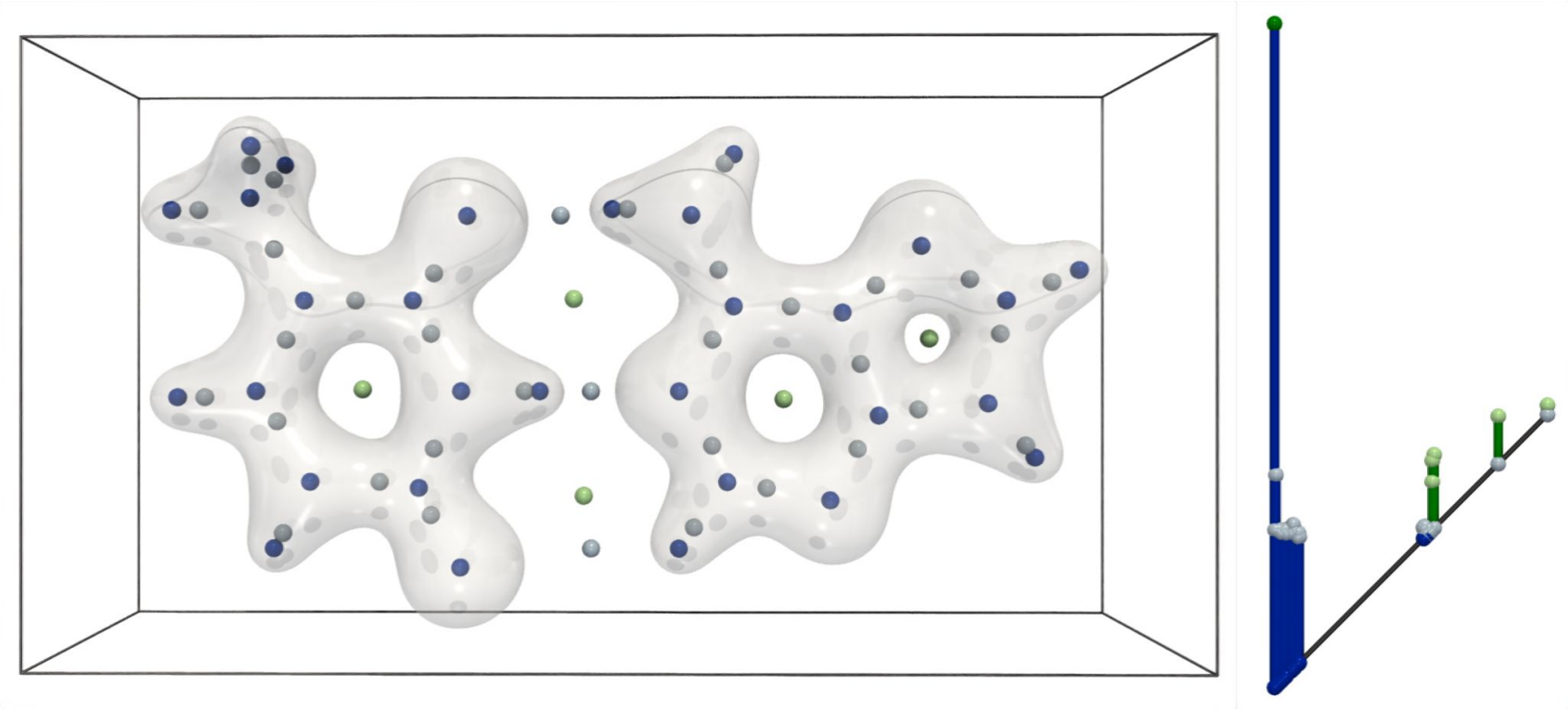
The persistence diagram

- Beyond connected components



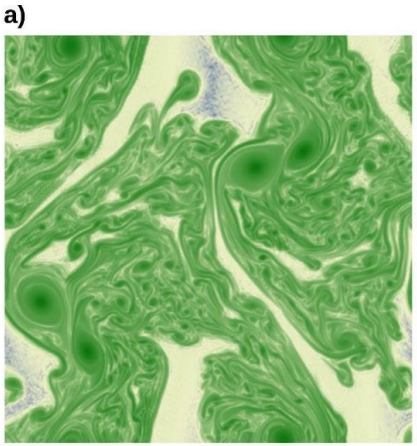
The persistence diagram

- Beyond connected components



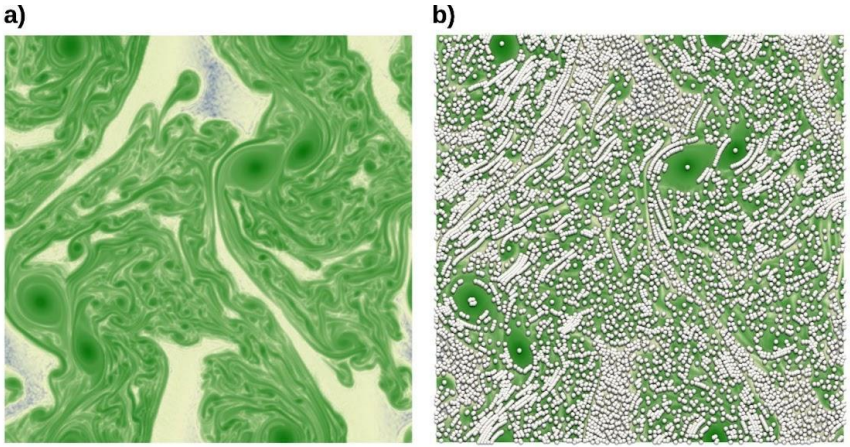
The persistence diagram

- **Structural features**
  - Vortex centers
    - Enstrophy maxima



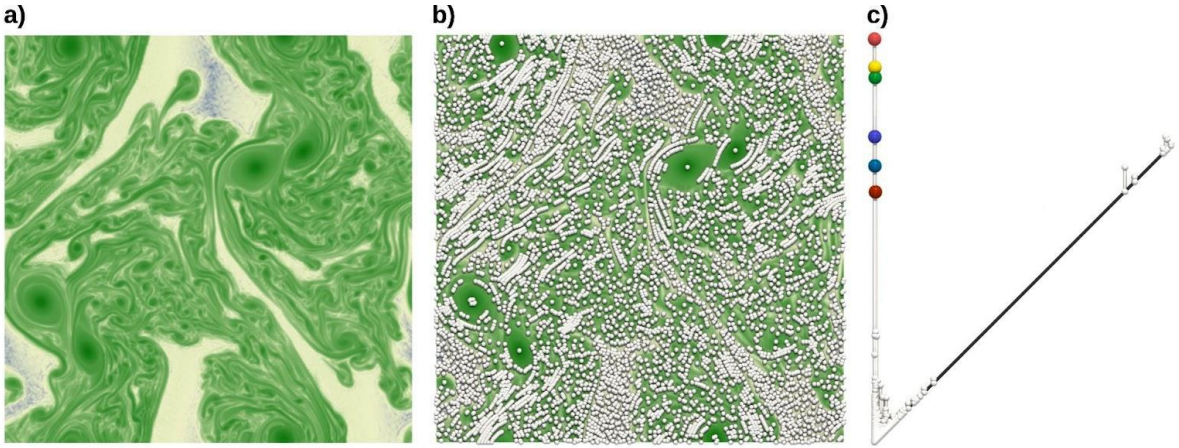
The persistence diagram

- **Structural features**
  - Vortex centers
    - Enstrophy maxima



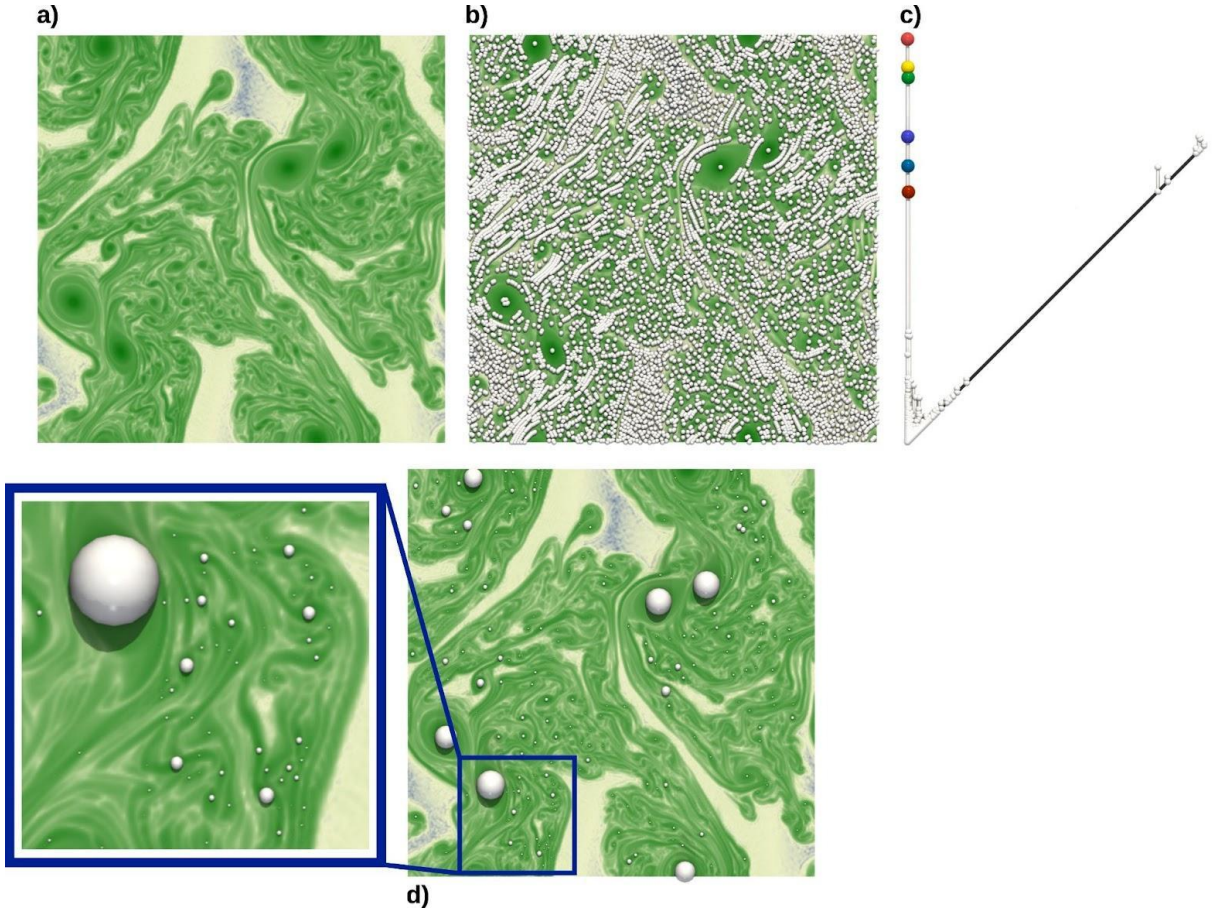
The persistence diagram

- **Structural features**
  - Vortex centers
    - Enstrophy maxima
  - Large-scale vortices
    - Persistent maxima



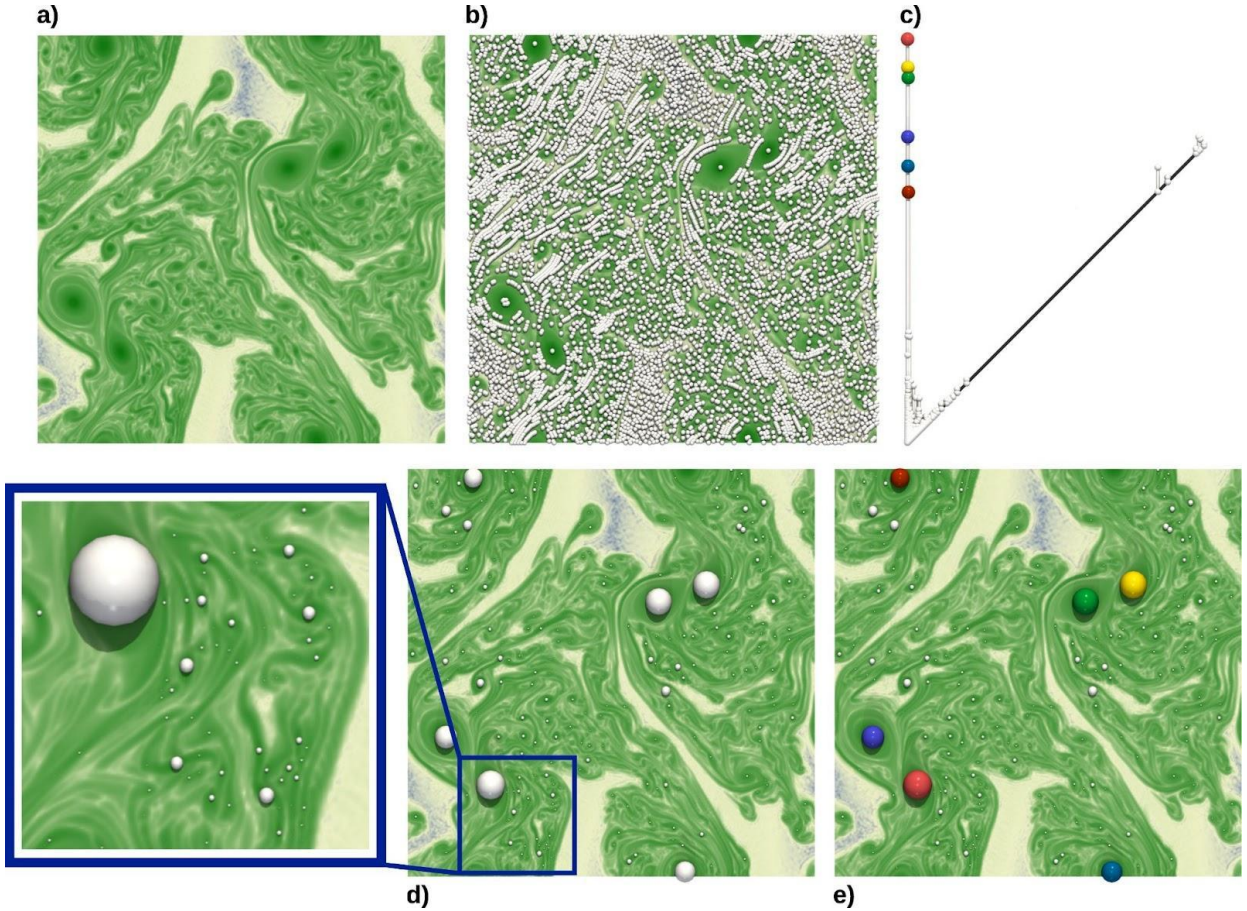
The persistence diagram

- **Structural features**
  - Vortex centers
    - Enstrophy maxima
  - Large-scale vortices
    - Persistent maxima



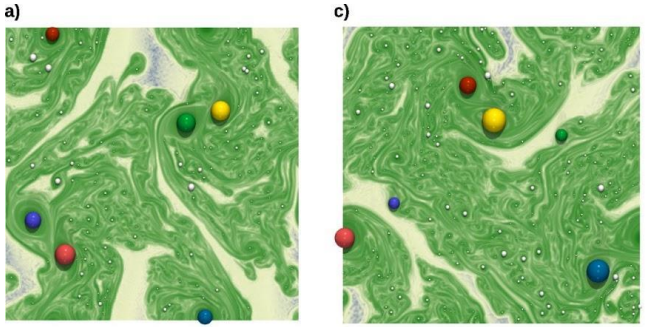
The persistence diagram

- **Structural features**
  - Vortex centers
    - Enstrophy maxima
  - Large-scale vortices
    - Persistent maxima



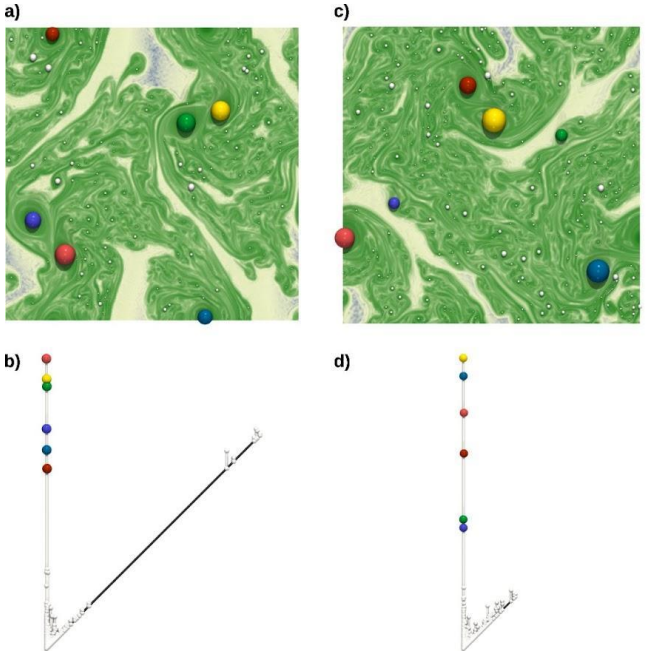
Comparing topologies

- Data representation



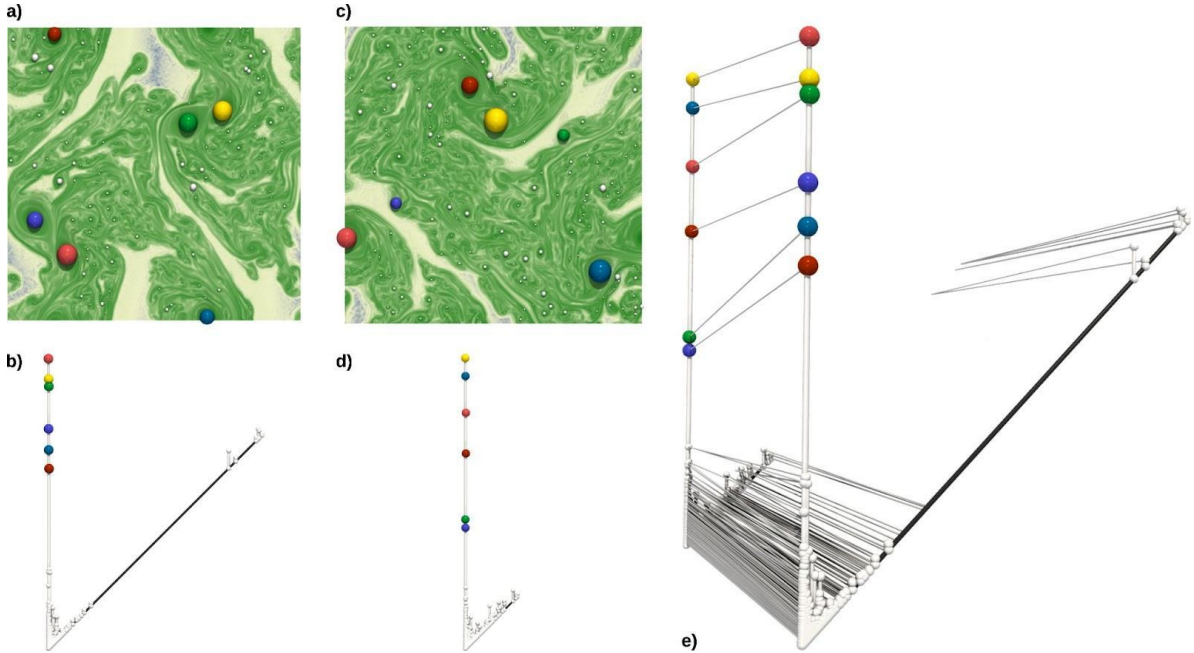
Comparing topologies

- **Data representation**
  - Topological signature
    - Persistence diagram



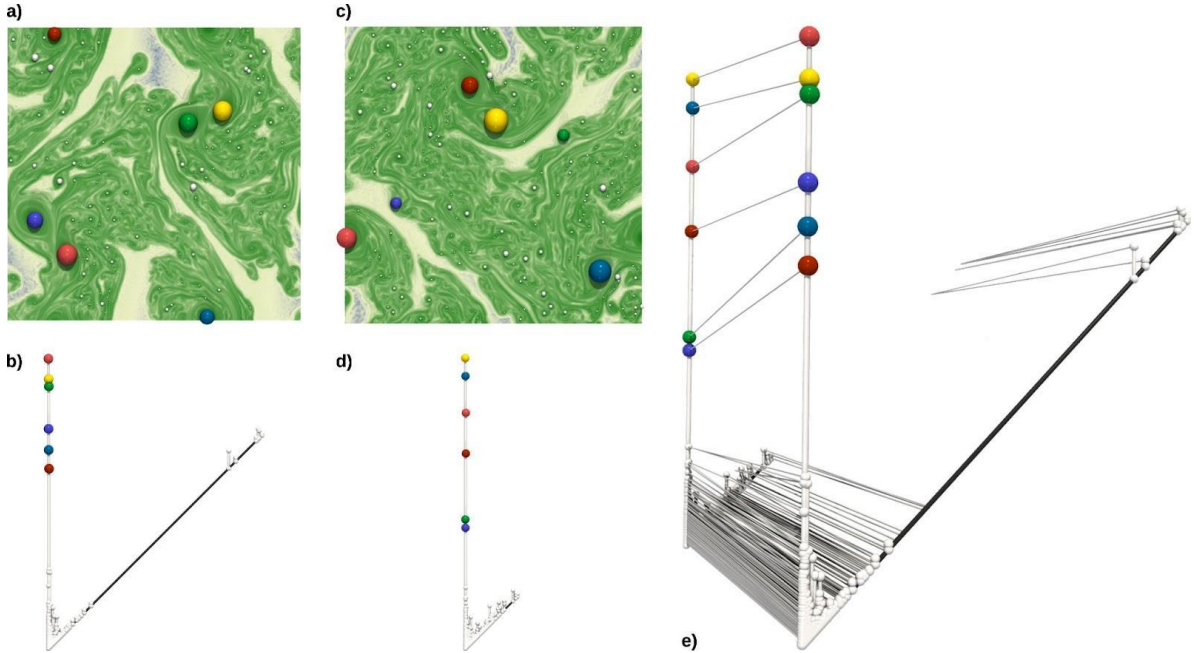
Comparing topologies

- **Data representation**
  - Topological signature
    - Persistence diagram
- **Data comparison**
  - Wasserstein distance
    - Optimal assignment
    - Computable, stable

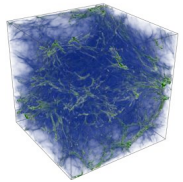


Comparing topologies

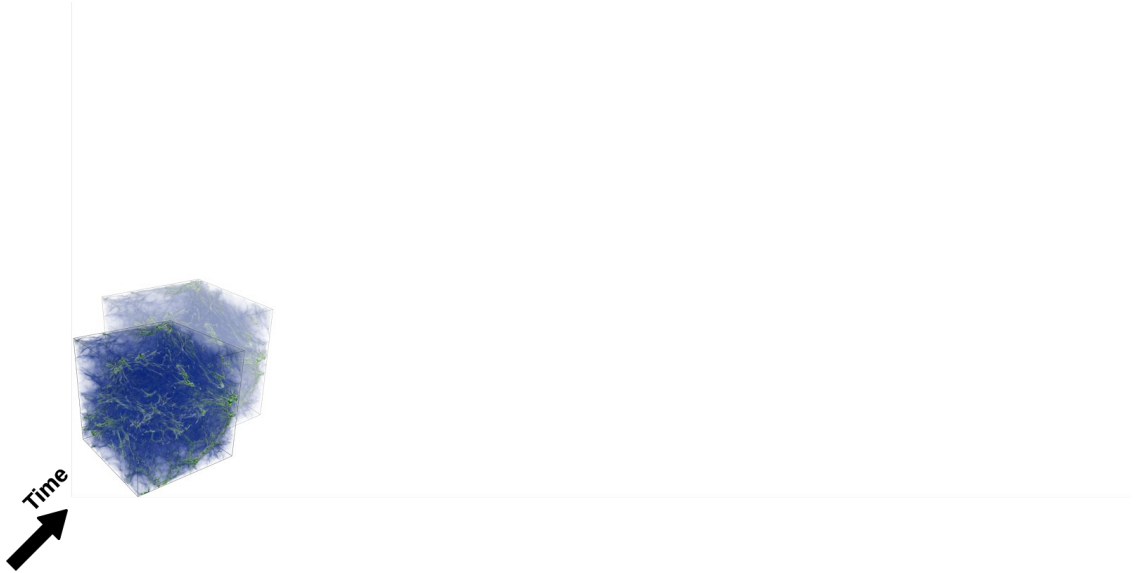
- **Data representation**
  - Topological signature
    - Persistence diagram
- **Data comparison**
  - Wasserstein distance
    - Optimal assignment
    - Computable, stable
- ***Topological statistics***
  - Geodesics, means, …



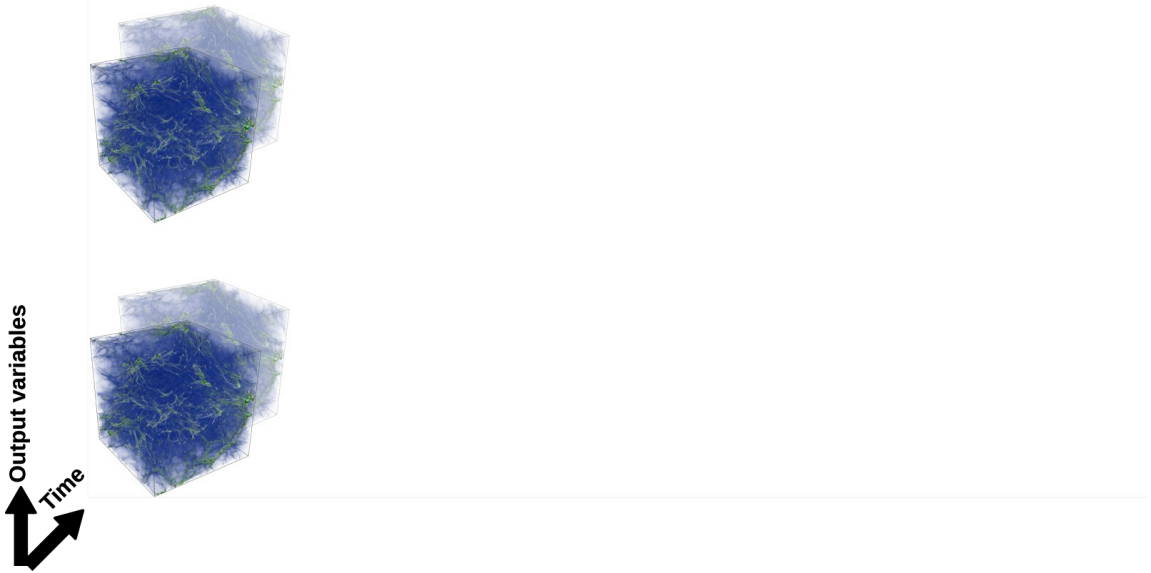
The ERC project “TORI”



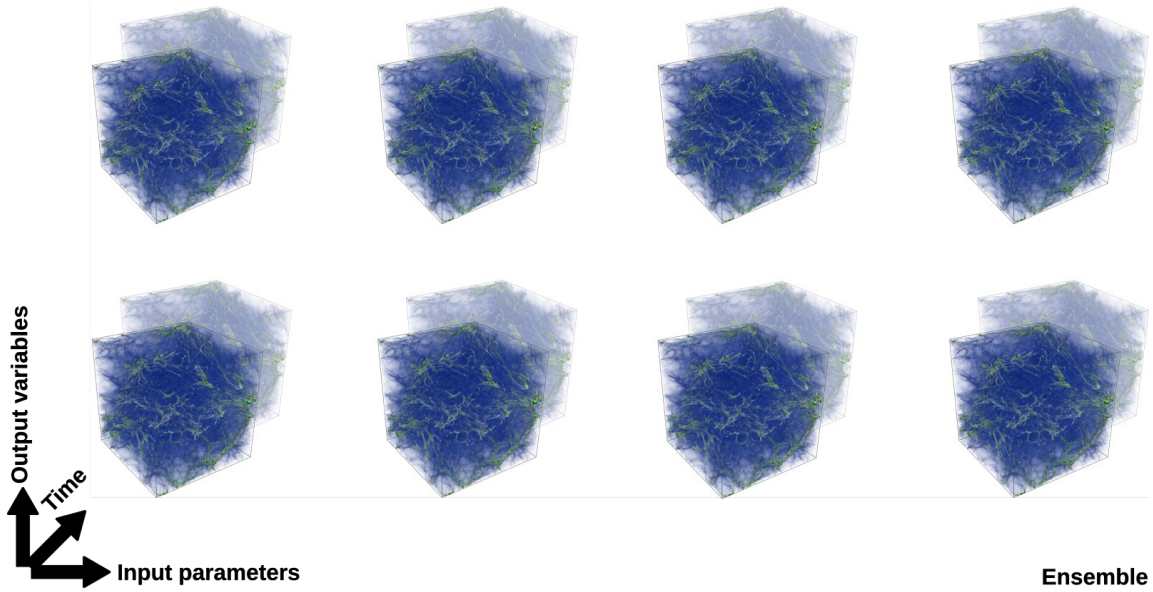
The ERC project “TORI”



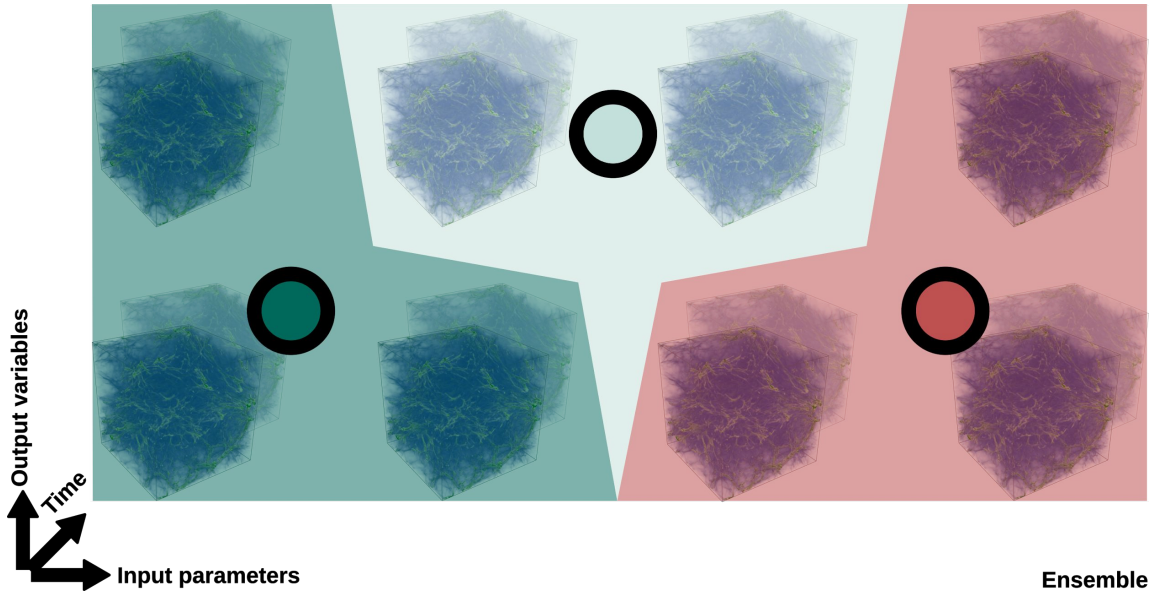
The ERC project “TORI”



The ERC project “TORI”

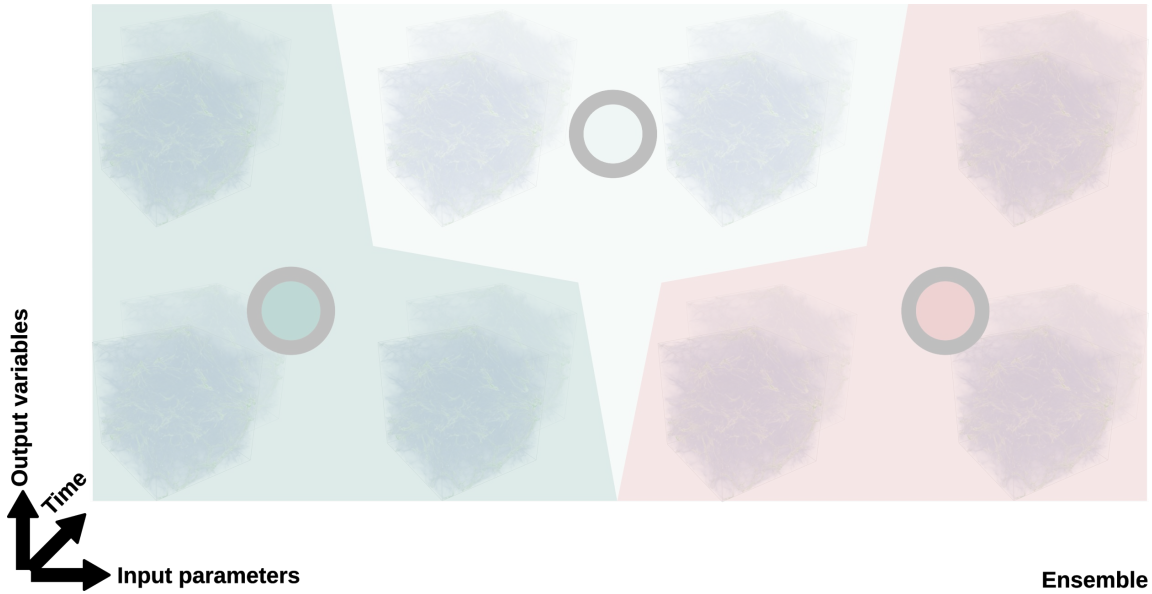


The ERC project “TORI”



The ERC project “TORI”

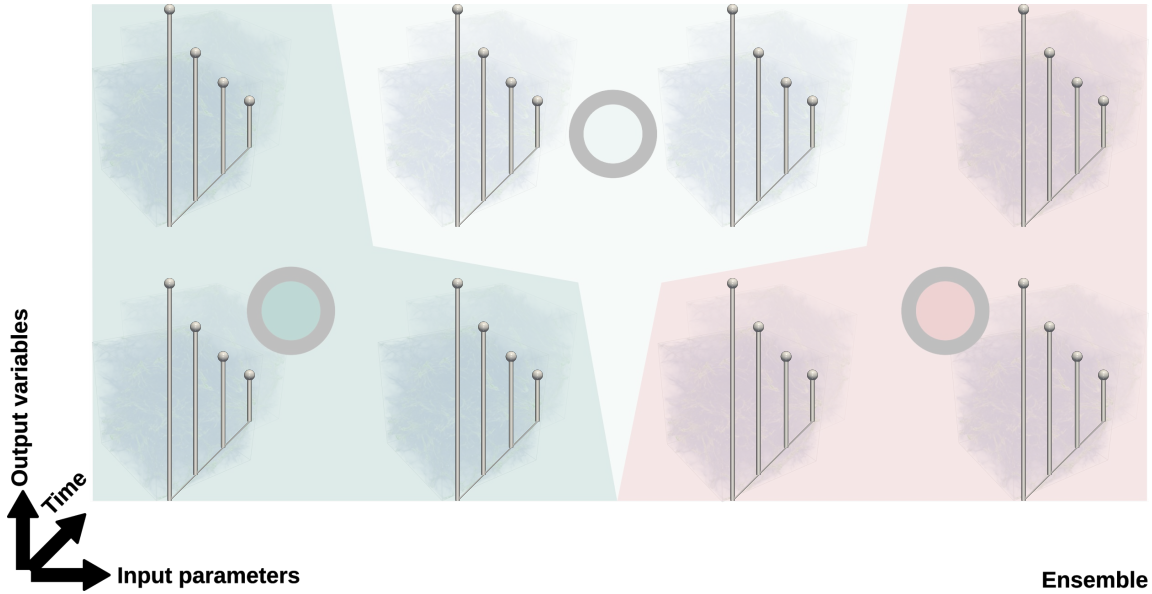
- **Challenges**
  - Significance
  - Data size



The ERC project “TORI”

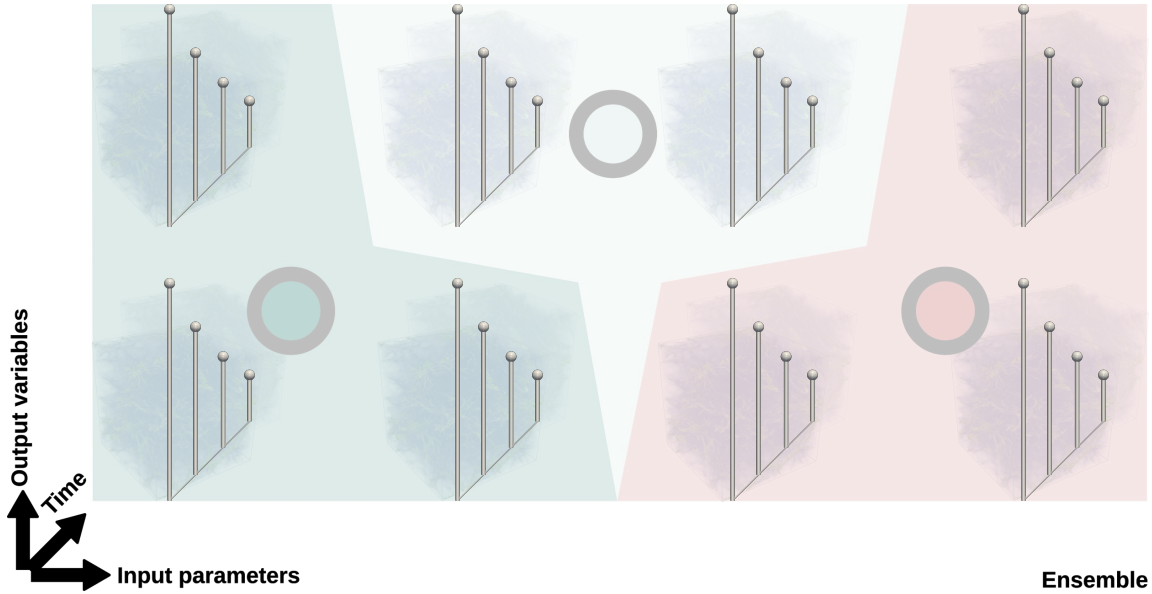
- **Challenges**
  - Significance
  - Data size

- **Strategy**
  - Topological reduction



The ERC project “TORI”

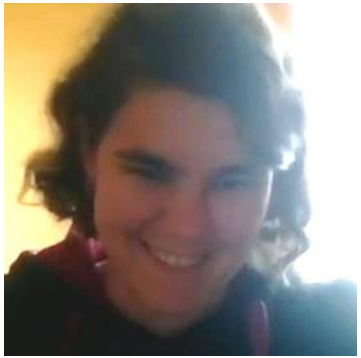
- **Challenges**
  - Significance
  - Data size
- **Strategy**
  - Topological reduction



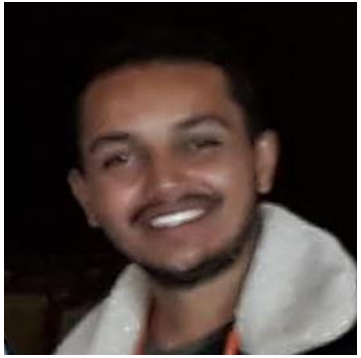
erc http://erc-tori.github.io/

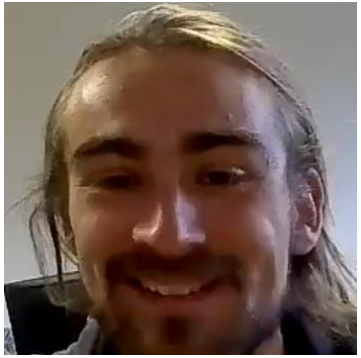
The TORI team





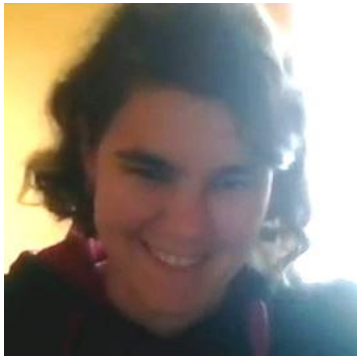




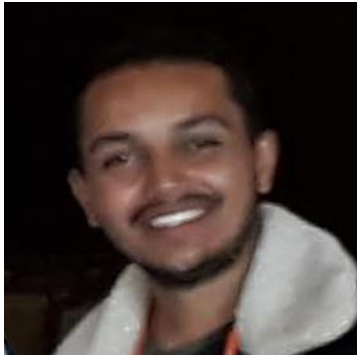
The TORI team

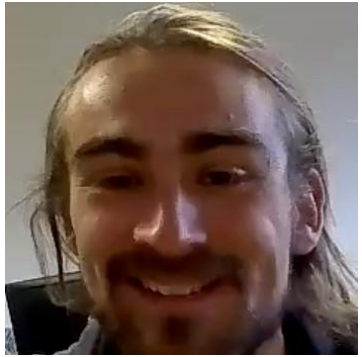
Topology at scale





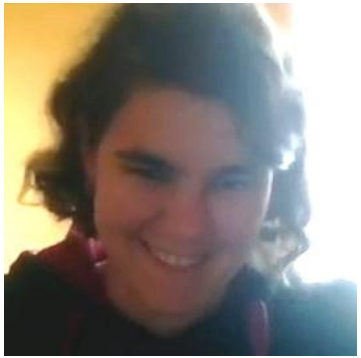




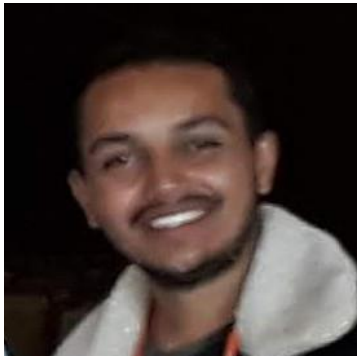
The TORI team

Topology at scale





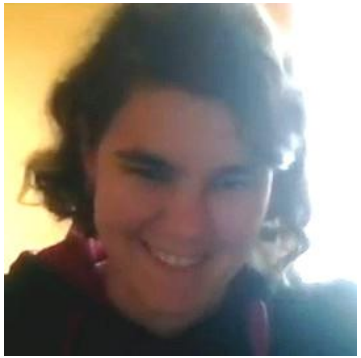



Topological statistics

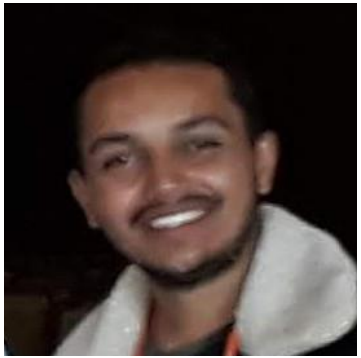
The TORI team

Topology at scale





Topological statistics

http://erc-tori.github.io/publications

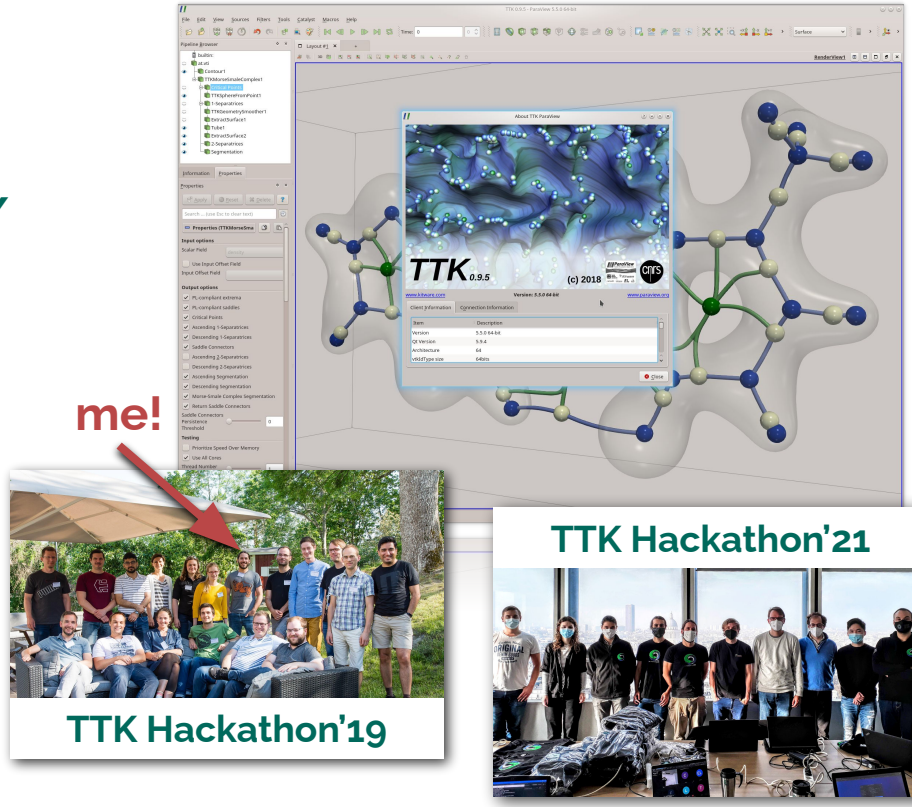
Reproducible research

- Replicable papers
  - https://www.replicabilitystamp.org/

Reproducible research

- **Replicable papers**
  - **https://www.replicabilitystamp.org/**

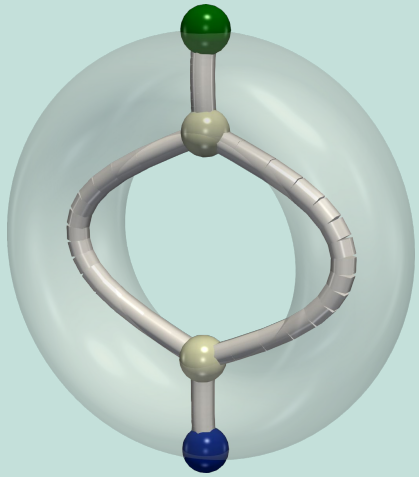
- **The Topology ToolKit (TTK)**
  - Open-source library C++ (2017)
    - **http://topology-tool-kit.github.io**
  - 17 contributing institutions
  - Online example database



Take-home messages

- **Numerical data?**
  - Recover/analyze/compare ‘structural’ information

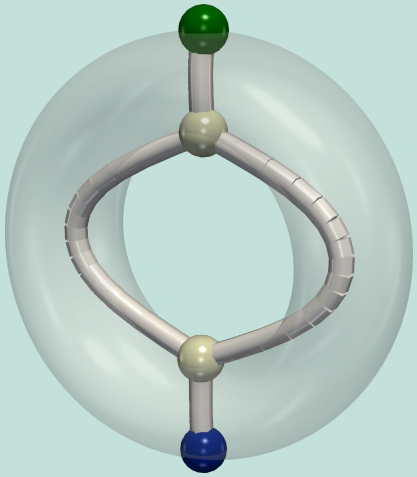
  ⇨ Topological Data Analysis
    - Robust, multiscale
    - Successful in applications



Take-home messages

- **Numerical data?**
  - Recover/analyze/compare ‘structural’ information
  - ⇨ Topological Data Analysis
    - Robust, multiscale
    - Successful in applications

- **The TORI project**
  - Large scale data analysis
  - Available in open-source! (TTK)
  - Many perspectives



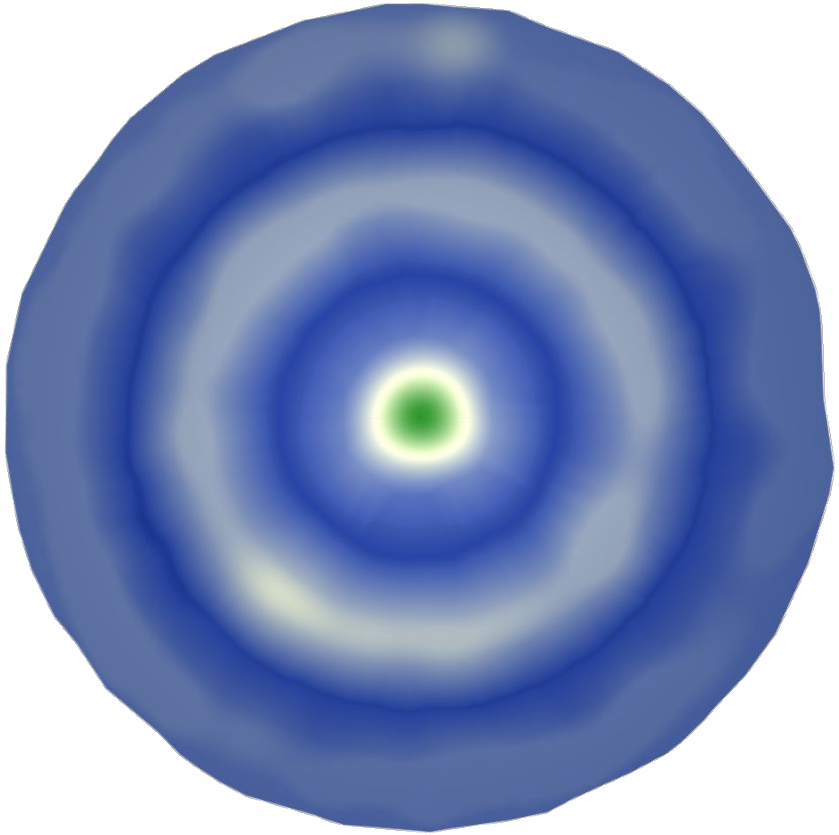
Thanks!

- **TTK**
  - Homepage: **https://topology-tool-kit.github.io/**
  - Examples: **https://topology-tool-kit.github.io/examples/**
  - Code: **https://github.com/topology-tool-kit/ttk/**

- **ERC Project TORI**
  - **https://erc-tori.github.io/**

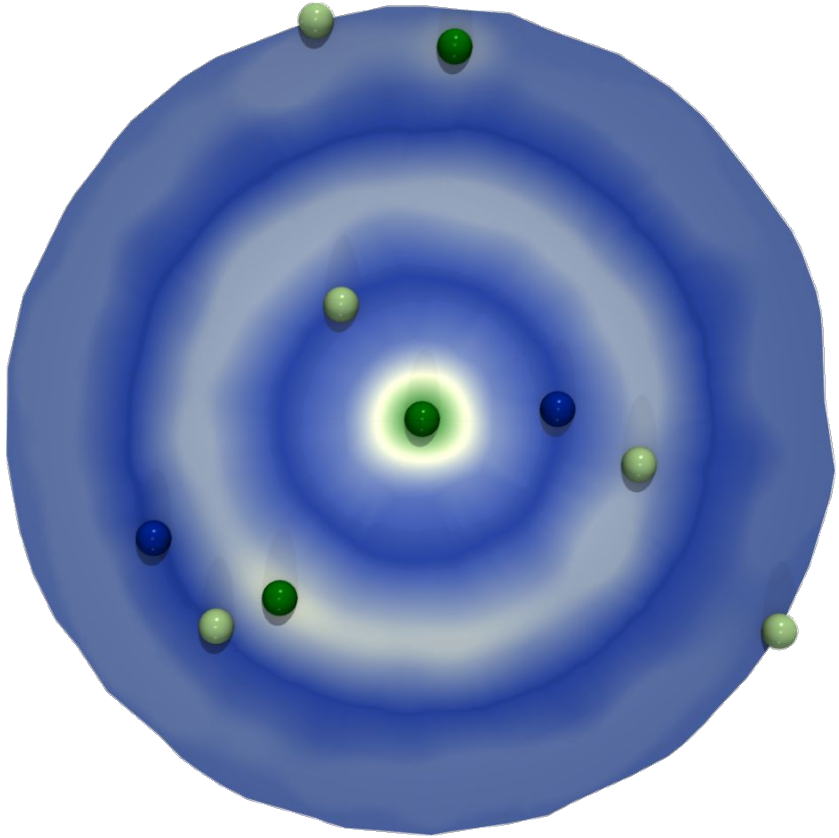
The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476



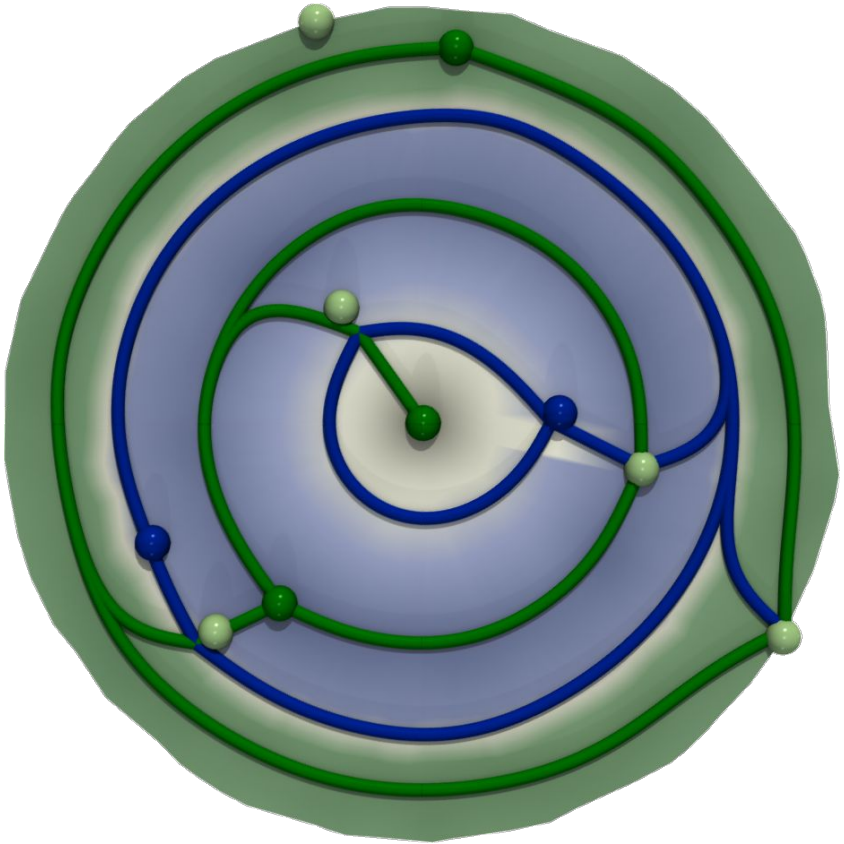
The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476



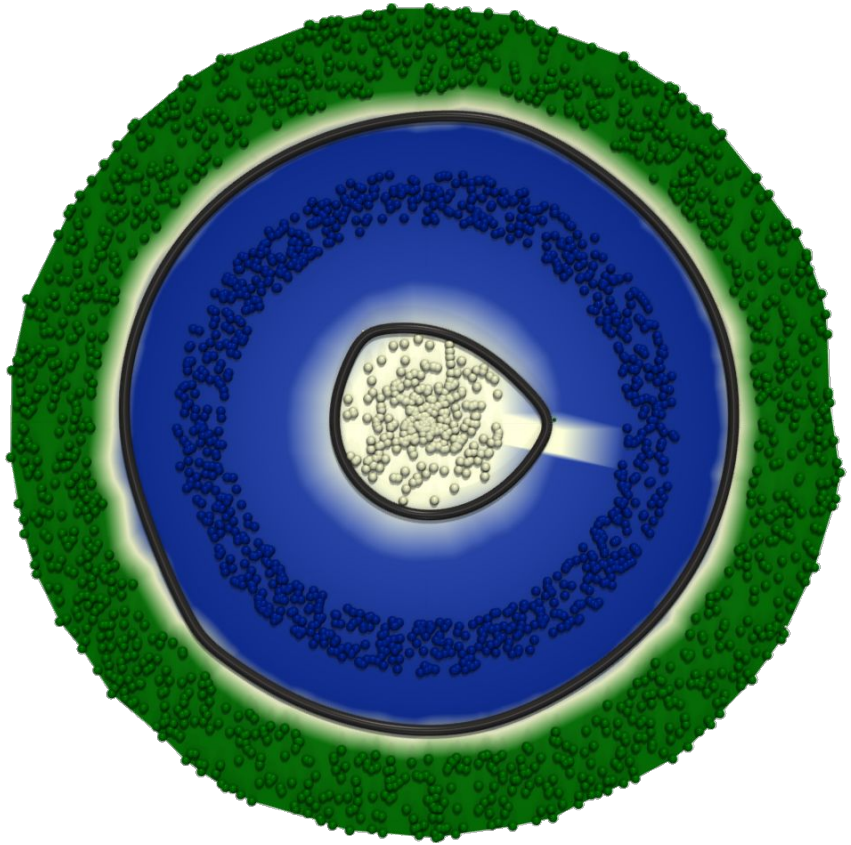
The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476



The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476



The topology of numerical data

	Row ID	x	y
0	0	-0.565872	0.433185
1	1	-0.852855	-0.0414391
2	2	-0.74526	-0.0704809
3	3	-0.283754	0.732374
4	4	-0.348494	0.708251
5	5	-0.141337	-0.791217
6	6	-0.590461	0.429422
7	7	-0.611657	-0.551685
8	8	-0.601515	-0.45884
9	9	0.240081	0.697199
10	10	0.101971	0.836754
11	11	-0.706277	-0.36127
12	12	0.301446	-0.731123
13	13	-0.759347	0.297477
14	14	-0.685959	0.140645
15	15	-0.220021	-0.828495
16	16	0.794982	0.352219
17	17	0.61109	0.561273
18	18	-0.642173	0.313909
19	19	-0.769766	-0.0934766
20	20	-0.0163218	-0.845476