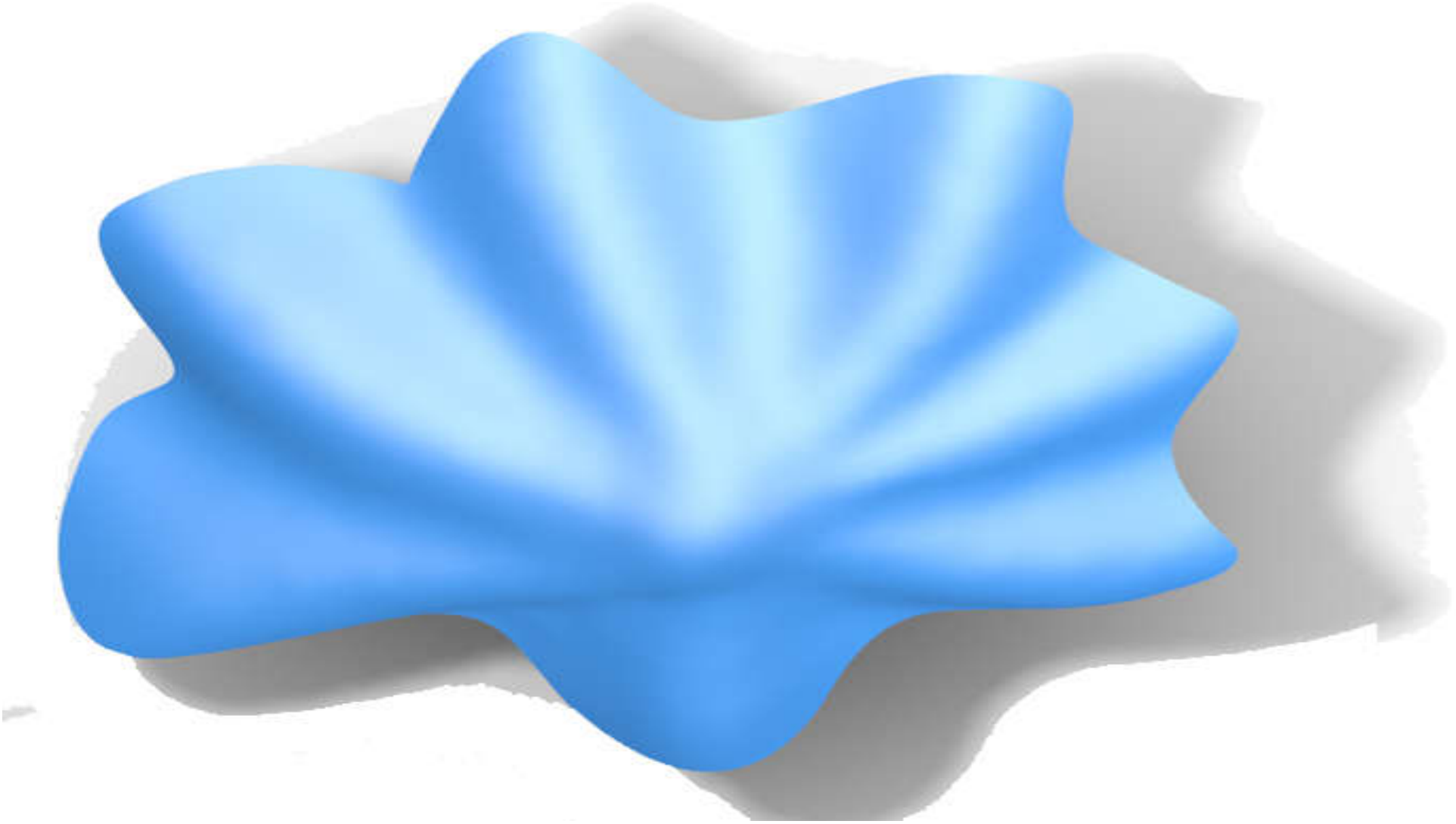
Discovering low-dimensional manifolds in high-dimensional data

Ingrid Daubechies
Duke University

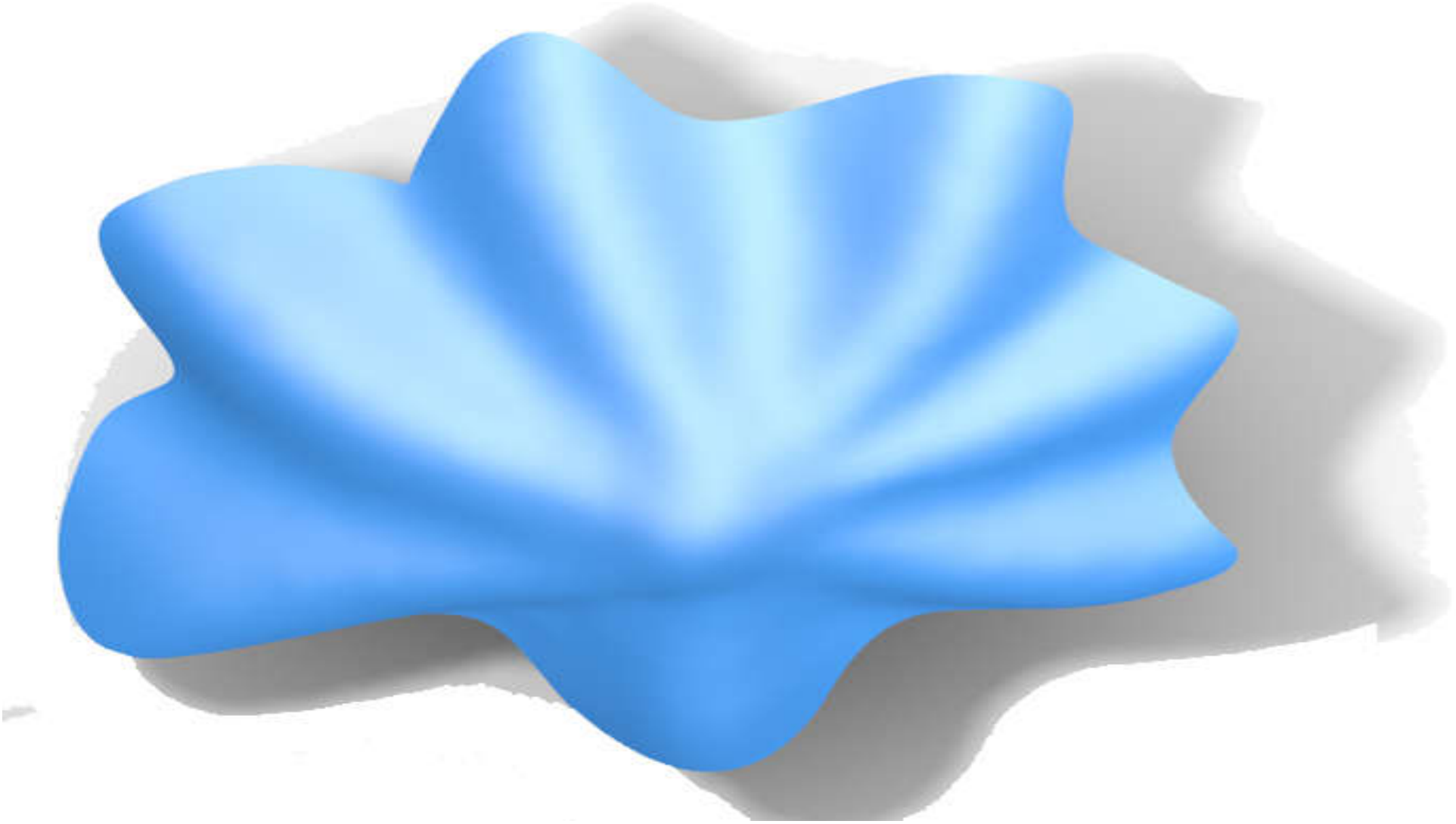
Paris, July 4, 2025

A manifold or variety may have a smaller dimension than how it is presented at first

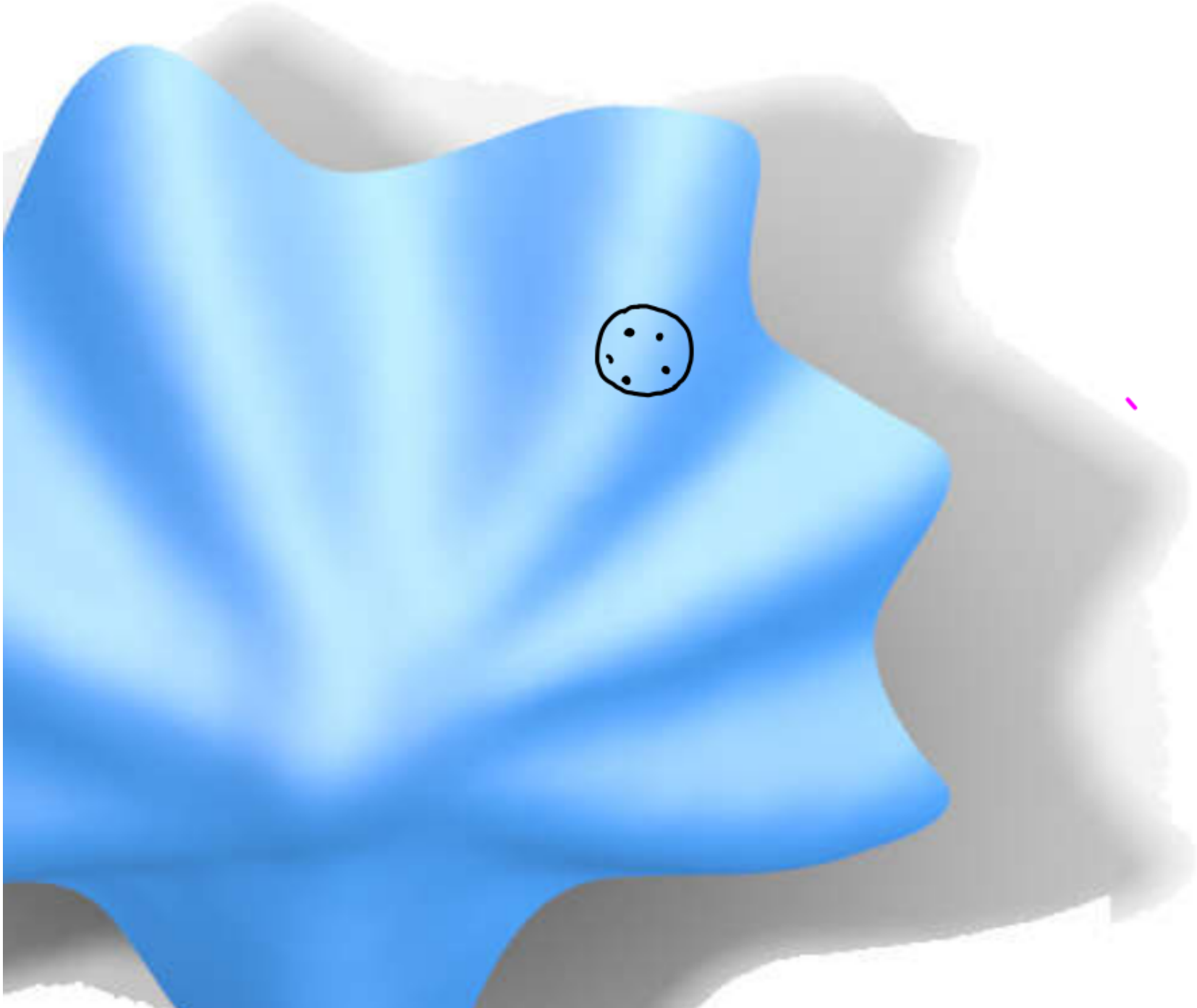


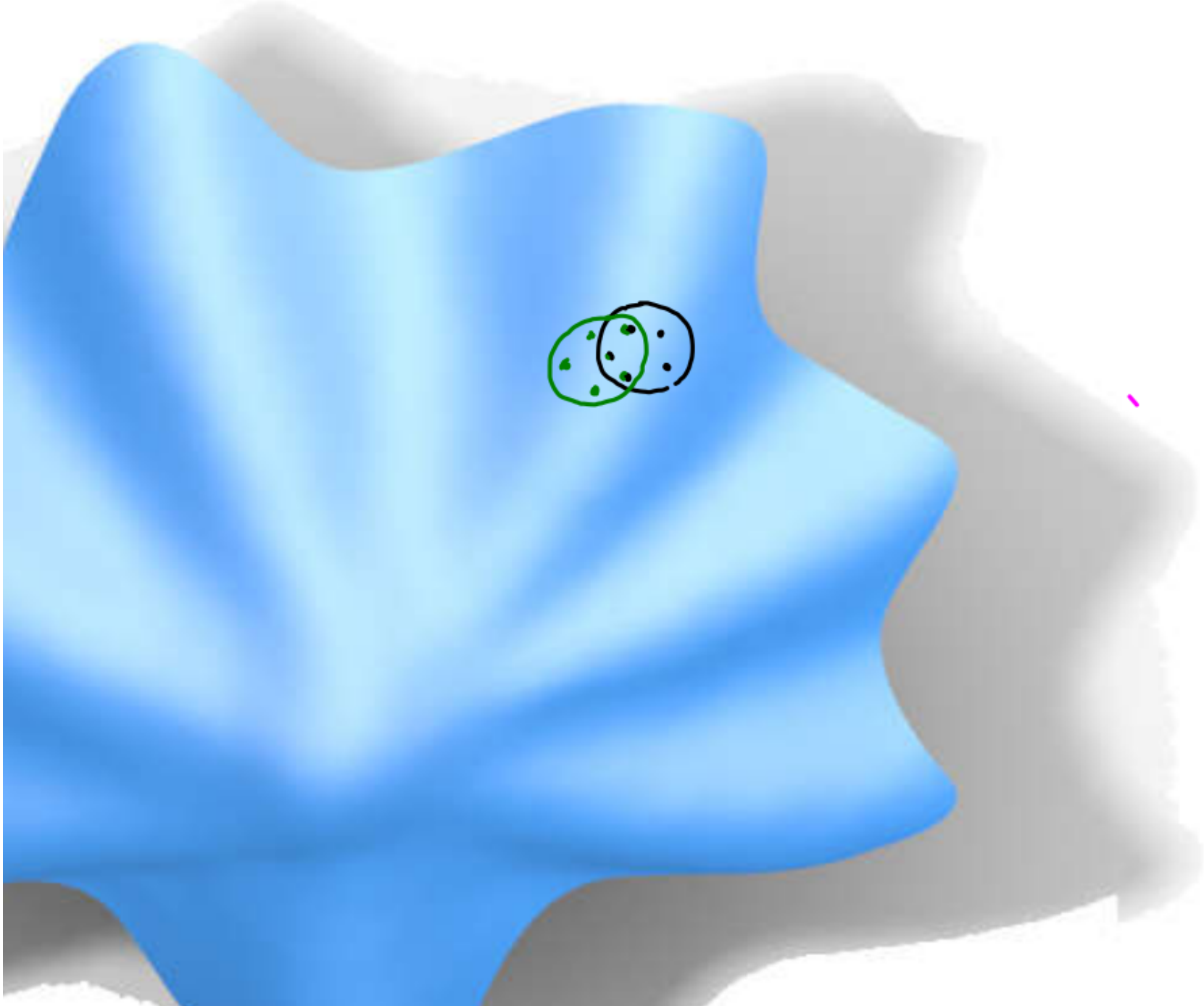
Sample points: given by **3** coordinates *(x,y,z)*

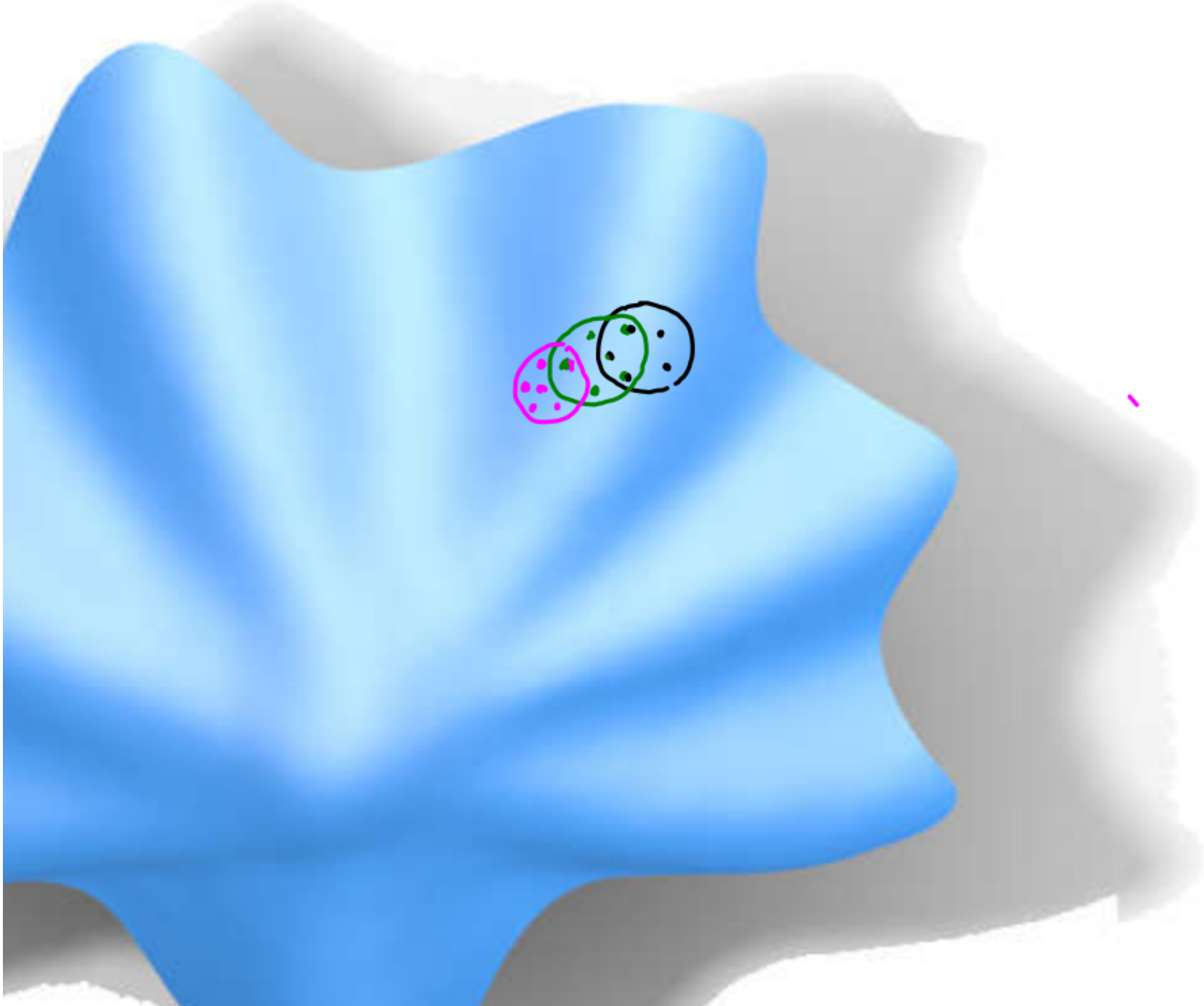
but the blue surface is only **2**-dimensional

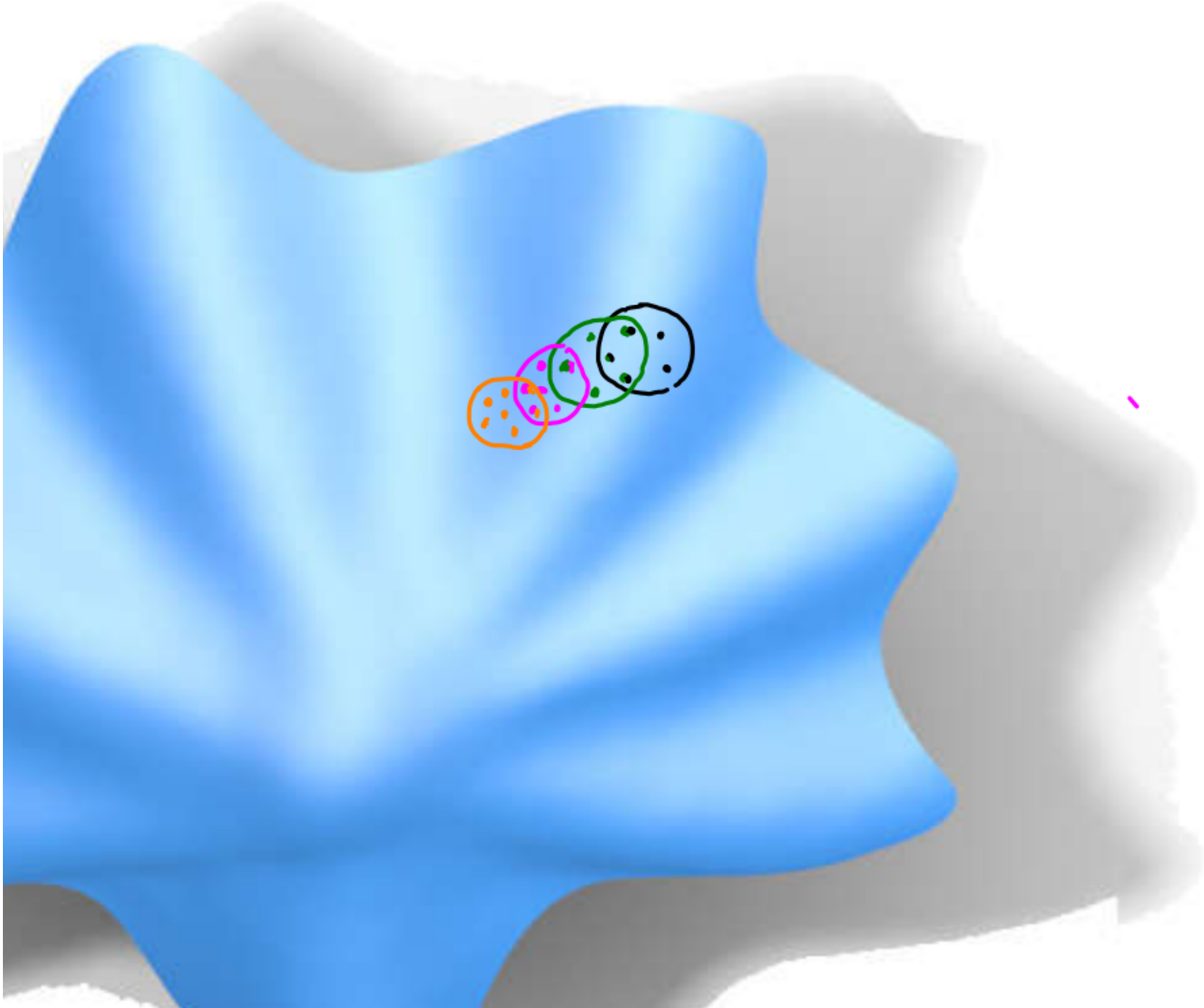


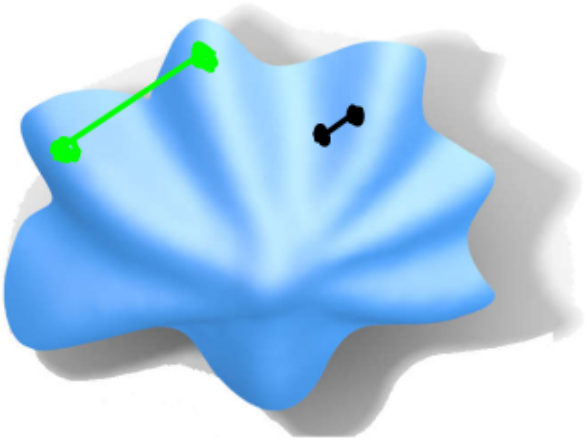
A collection of small tangent patches does give a good first approximation of a surface



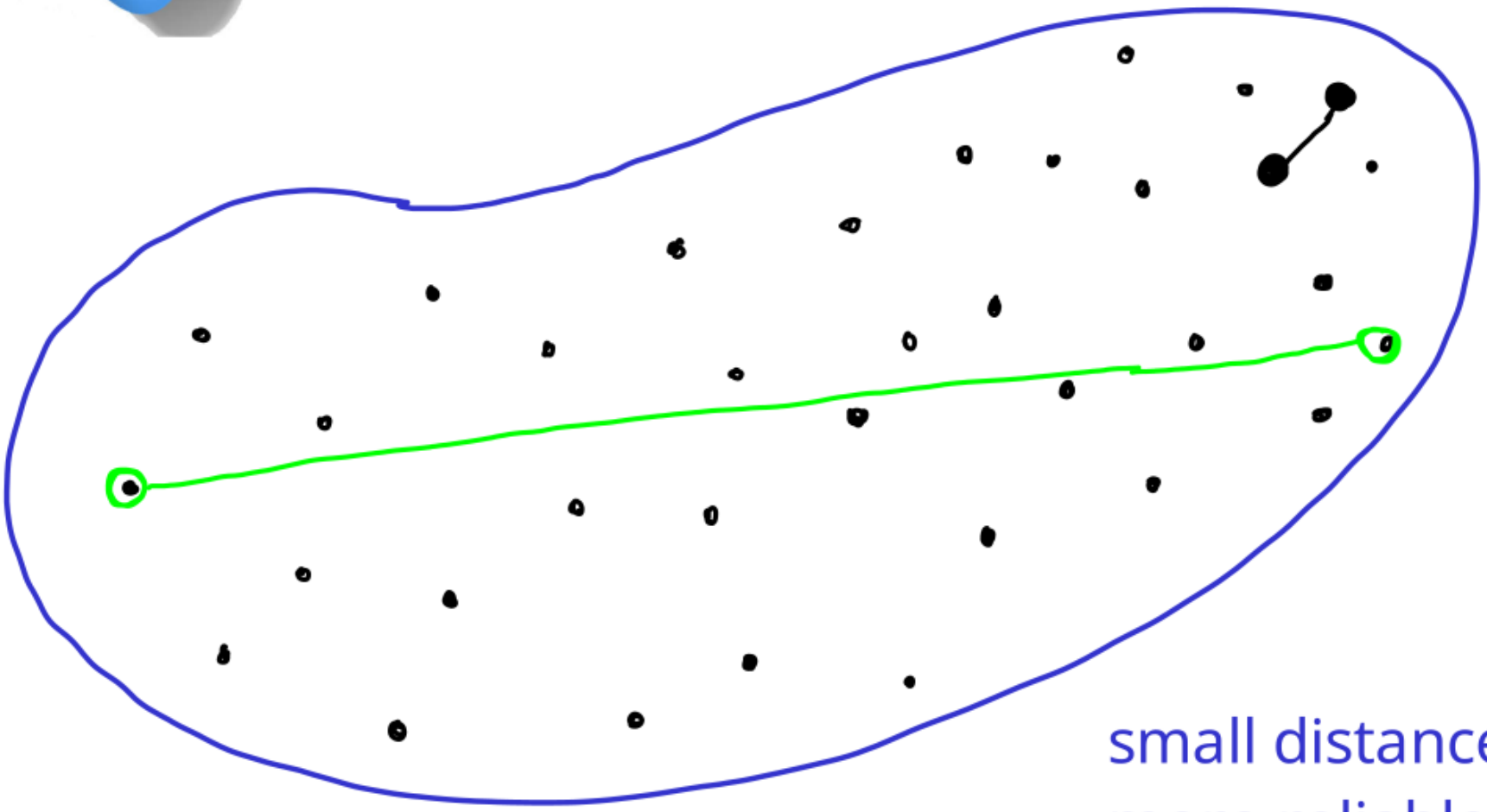




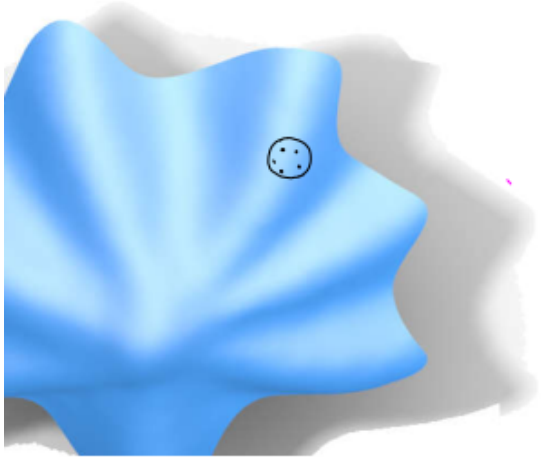




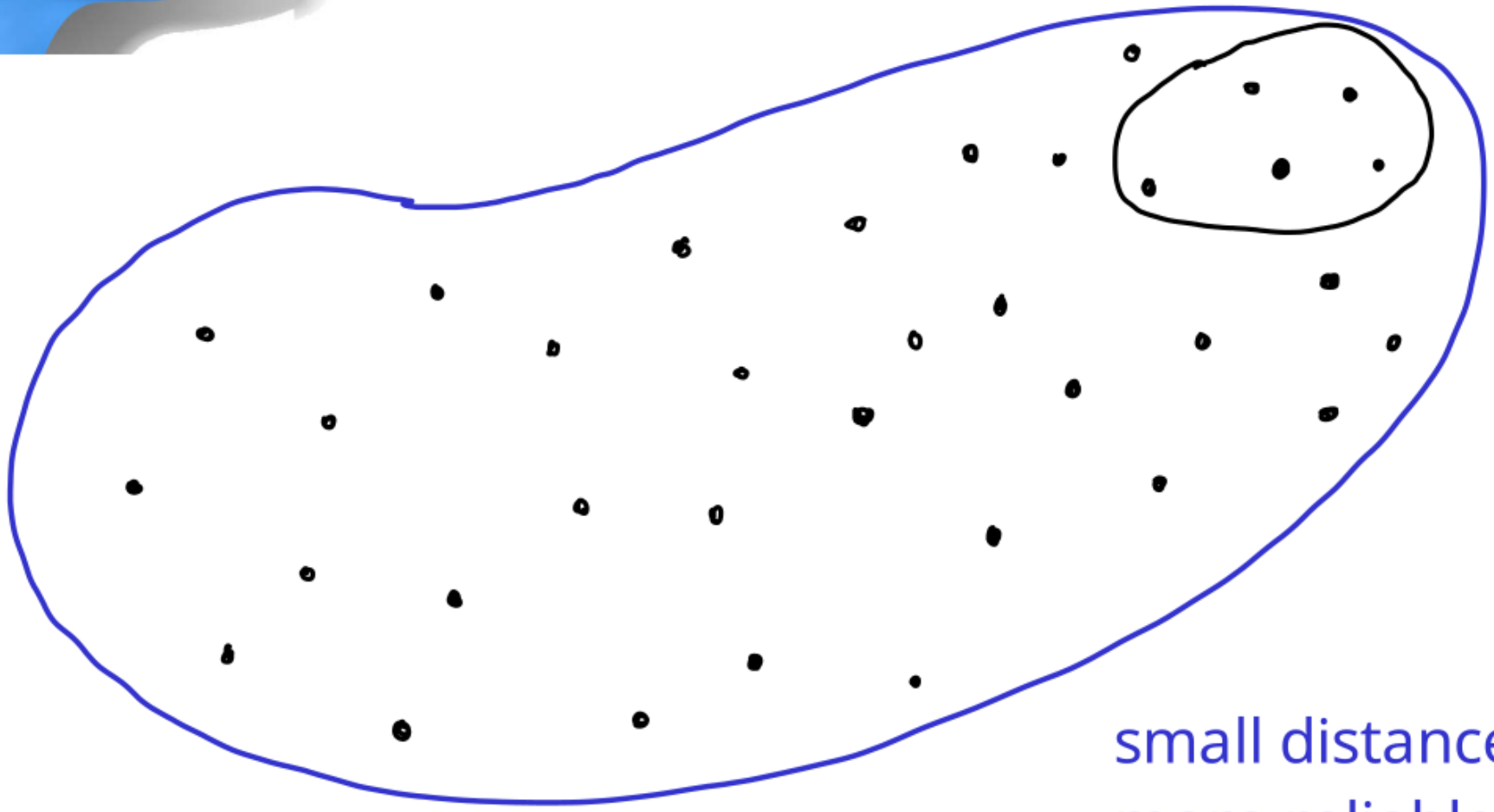
Data set



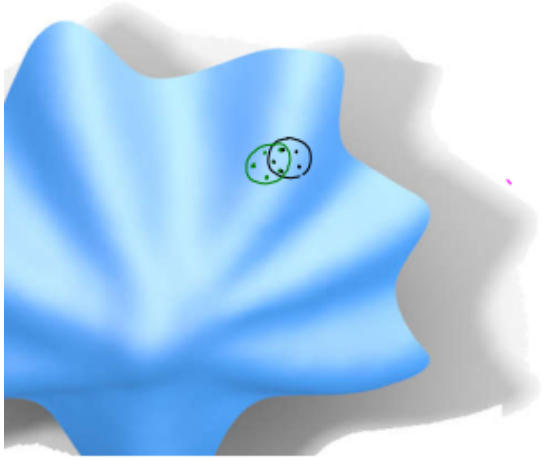
small distances are more reliable!



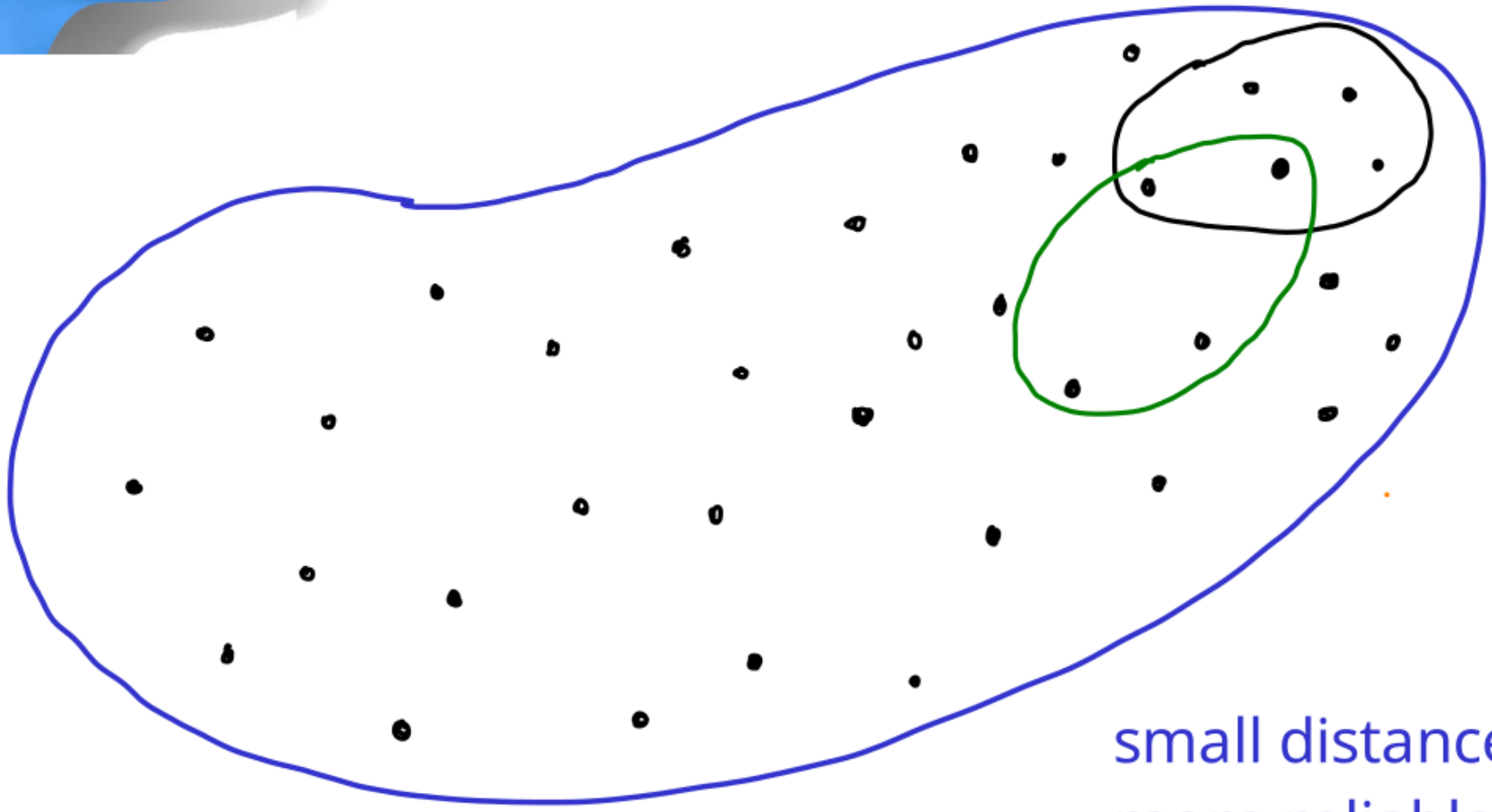
Data set



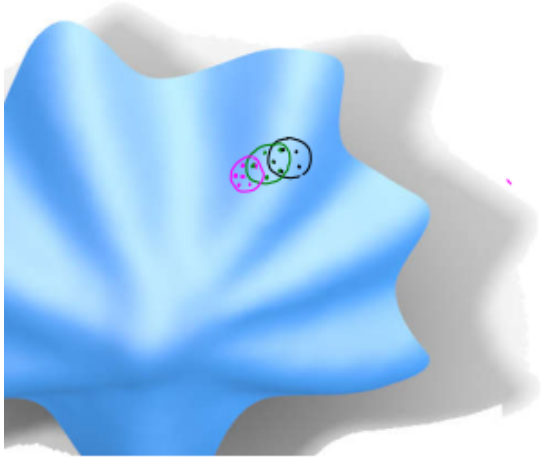
small distances are more reliable!



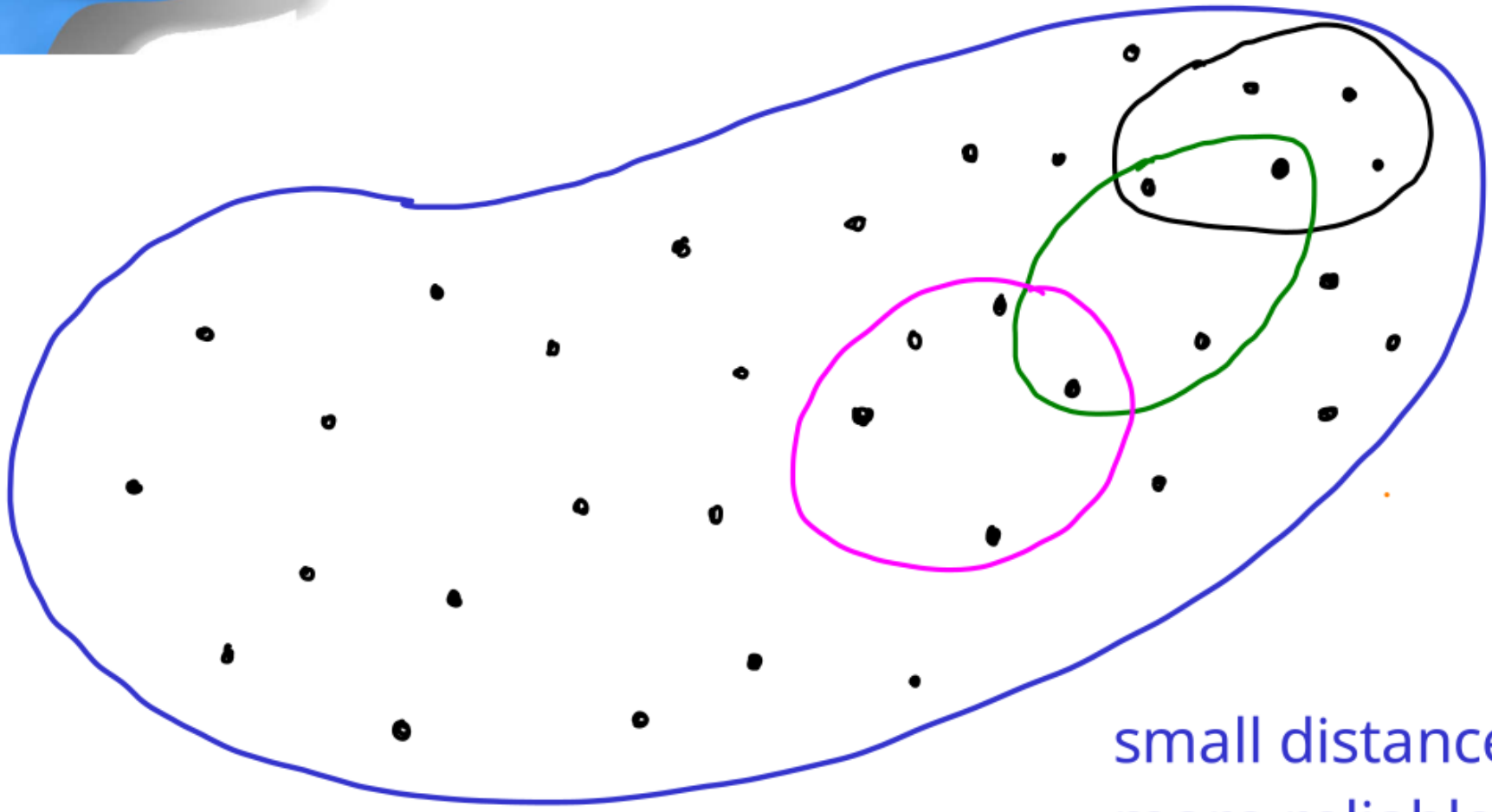
Data set



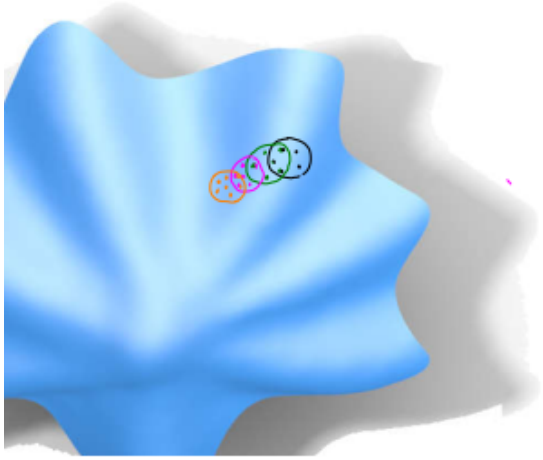
small distances are more reliable!



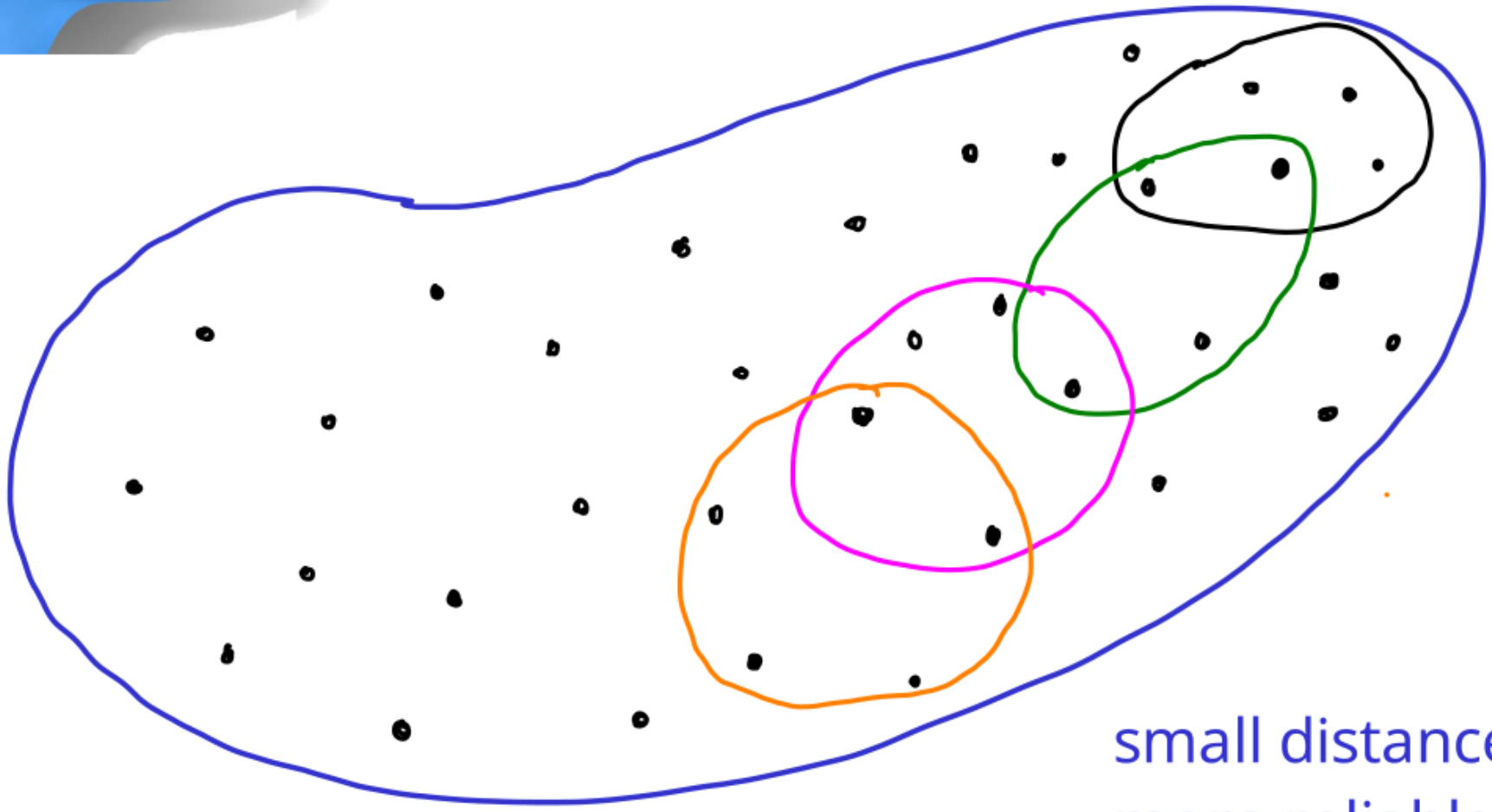
Data set



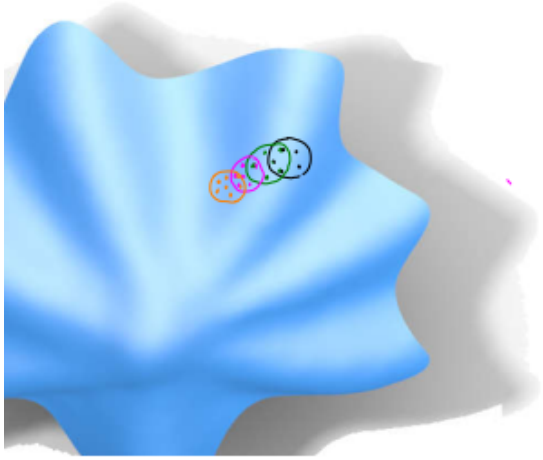
small distances are more reliable!



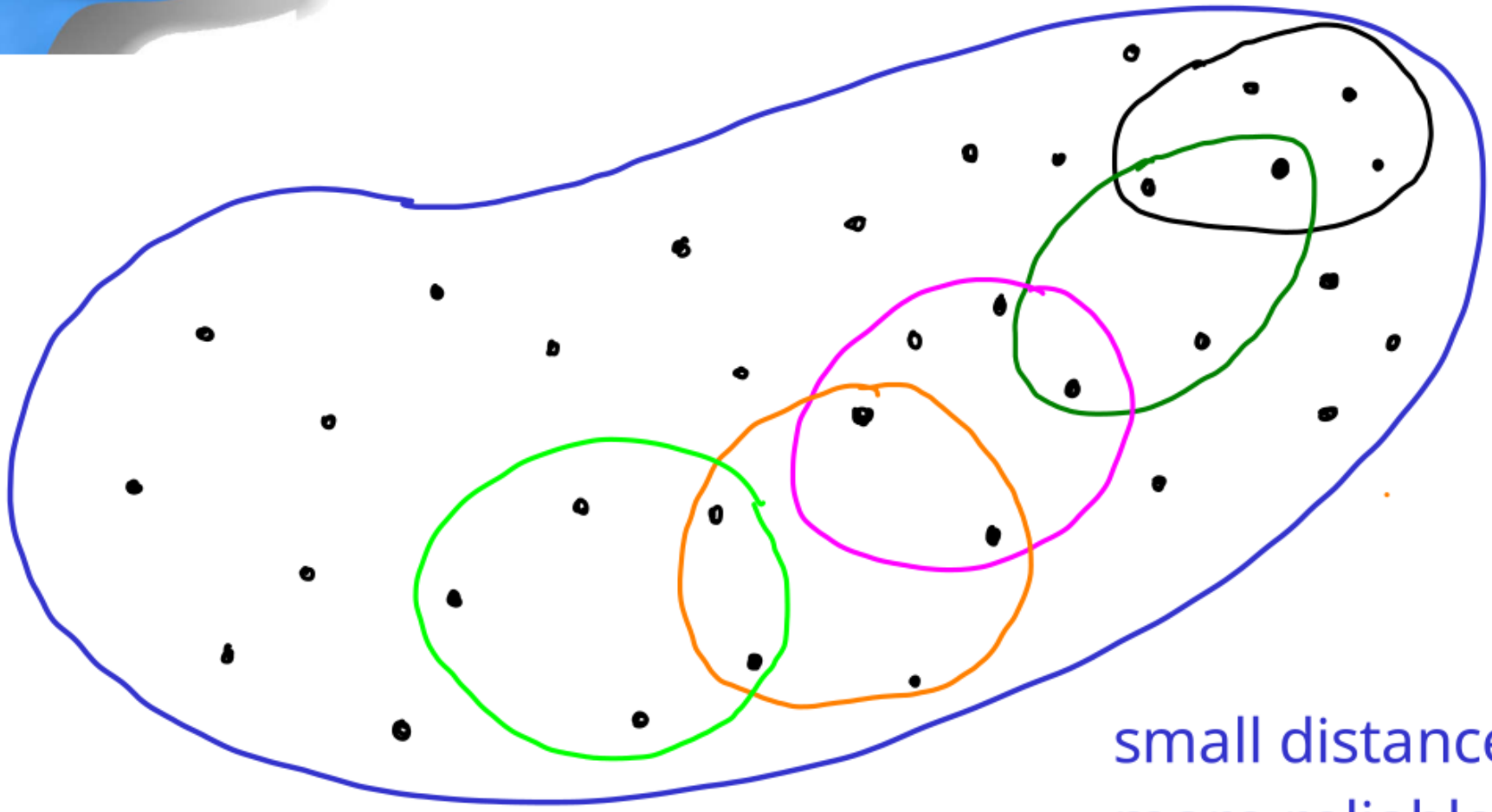
Data set



small distances are more reliable!



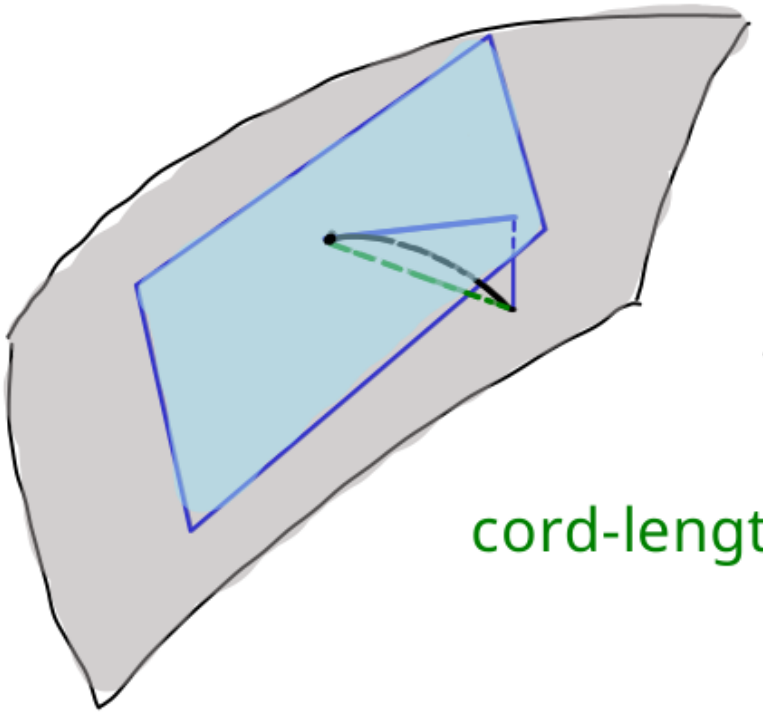
Data set



small distances are more reliable!

How do this computationally?

build diffusion operator (matrix)



in first approximation, and
very locally,

cord-length $\simeq$ dist. on manifold

$\simeq$ proj. dist. on tangent plane

$\Rightarrow$ use standard expression for heat diffusion kernel

use standard expression for heat kernel

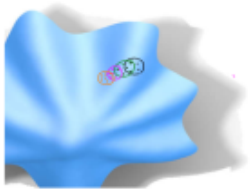
data points $P_i, P_j, \ldots$

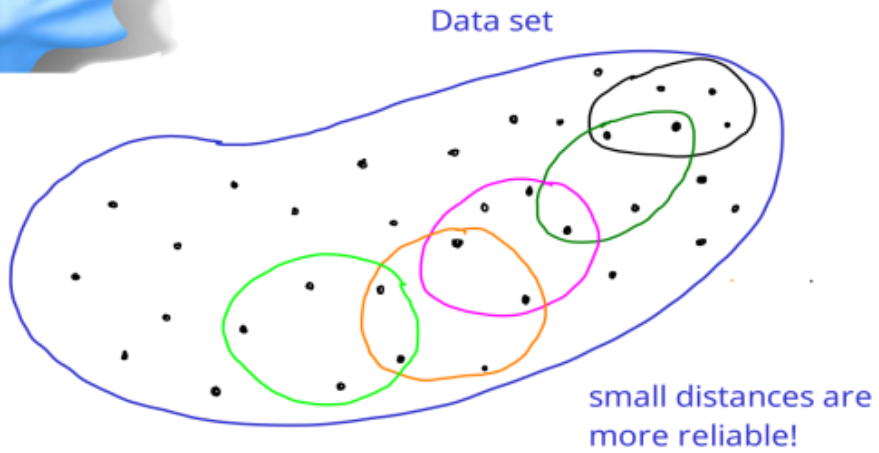
"dissimilarity" distance $\text{dist}(P_i, P_j) = d_{i,j}$

diffusion kernel in Euclidean space:

$$D_{t;\,i,j} = N_t \, e^{-d_{ij}^2/2t}$$

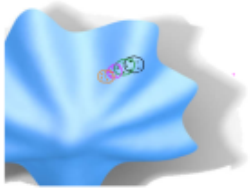
size of t measures extent of the diffusion

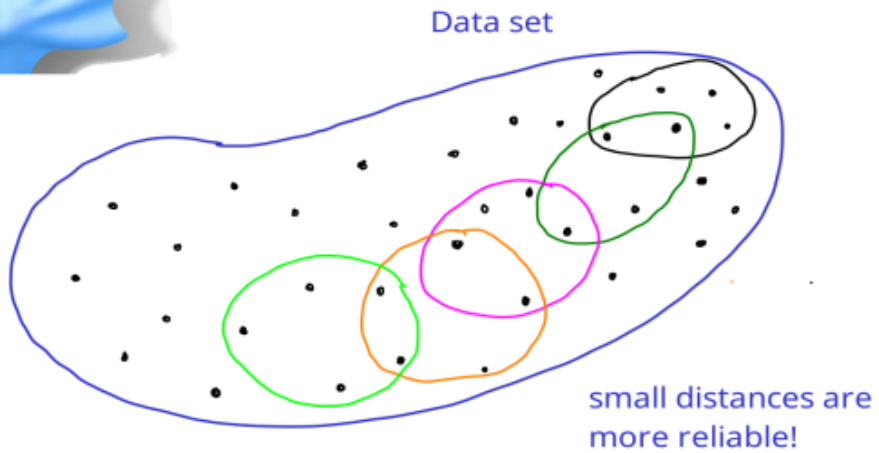




use standard expression for heat kernel

$$D_{t;i,j} = N_t \, e^{-d_{ij}^2/2t}$$





size of t indicates extent of the diffusion

t needs to be sufficiently small

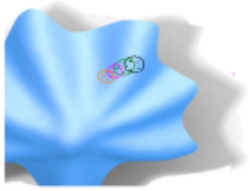
To "diffuse" further: consider powers of the diffusion matrix!

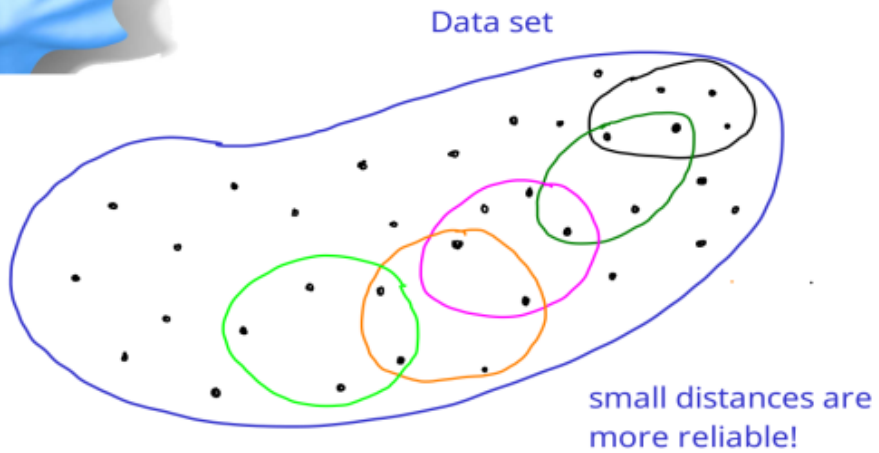
Normalization?

use standard expression for heat kernel

$$\mathcal{D}_{t;\,i,j} = N_t \, e^{-\,d_{ij}^2 / 2t}$$

Normalization?



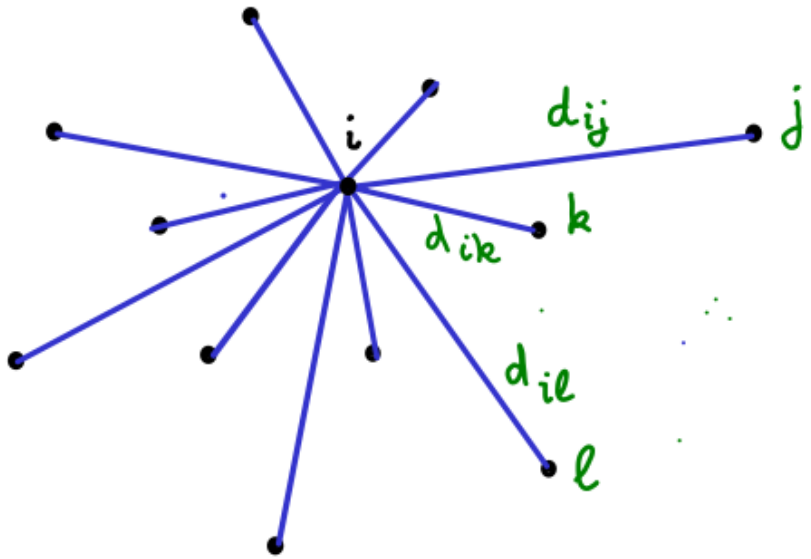


Heat diffusion on a manifold defines a random walk.

The approximation built from the data should do likewise:

$$\sum_j \mathcal{D}_{t;\,i,j} = 1.$$

$$\Rightarrow \quad N_t = \left(\sum_j e^{-\,d_{ij}^2 / 2t} \right)^{-1}$$



The data points and their distances define a weighted graph

$$e^{-d_{ij}^2/2t} = \mathcal{D}_{t;ij}$$

$$\widetilde{\mathcal{D}}_{t;i,i} = \sum_j \mathcal{D}_{t;i,j}$$

$$W_t = \widetilde{\mathcal{D}}_t^{-1} \mathcal{D}_t$$: defines a random walk on this graph

$\Longrightarrow$ candidate for a discrete approximation of the diffusion on the manifold

How should one pick t ?

Not too large , not too small

$$e^{-d_{ij}^2/2t} = D_{t;ij}$$

$$\tilde{D}_{t;i,i} = \sum_j D_{t;i,j}$$

$$W_t = \tilde{D}_t^{-1} D_t$$

candidate for a discrete approximation of the diffusion on the manifold

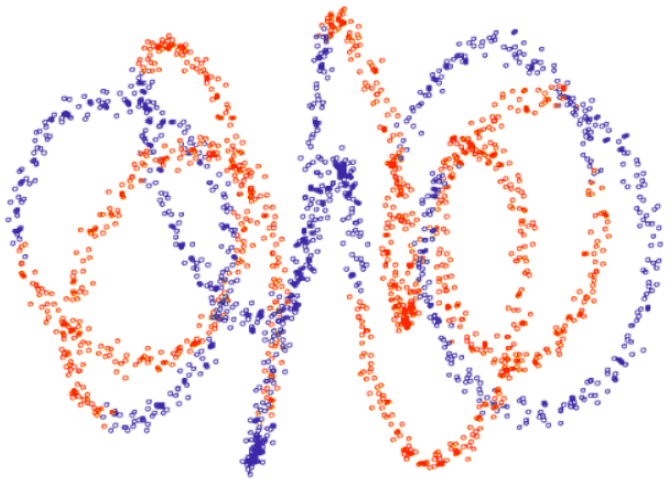
How should one pick t ?

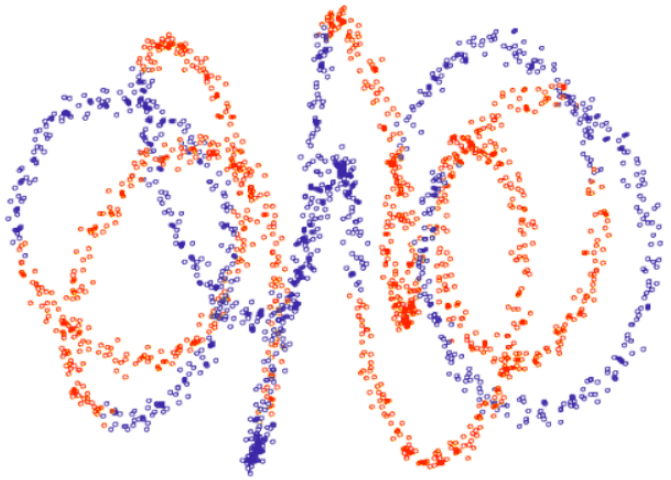
Not too large , not too small

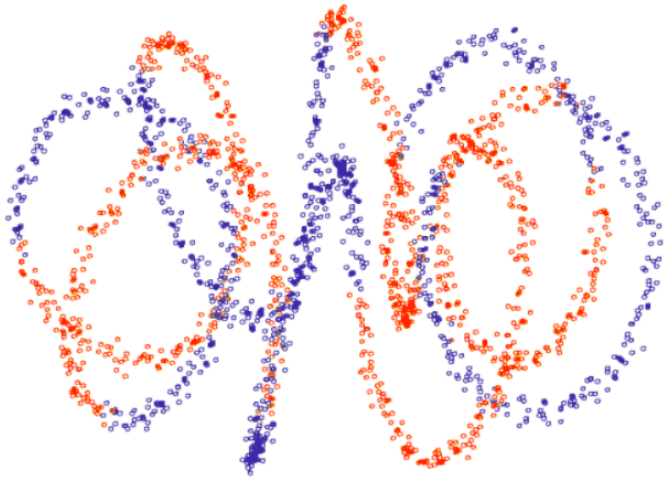
$W_t \; W_s$ should be close to W_{t+s}

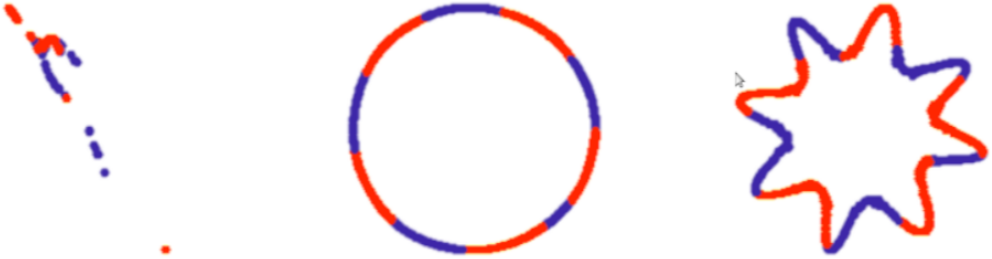
$\| (W_t)^2 - W_{2t} \| \rightarrow$ should be small

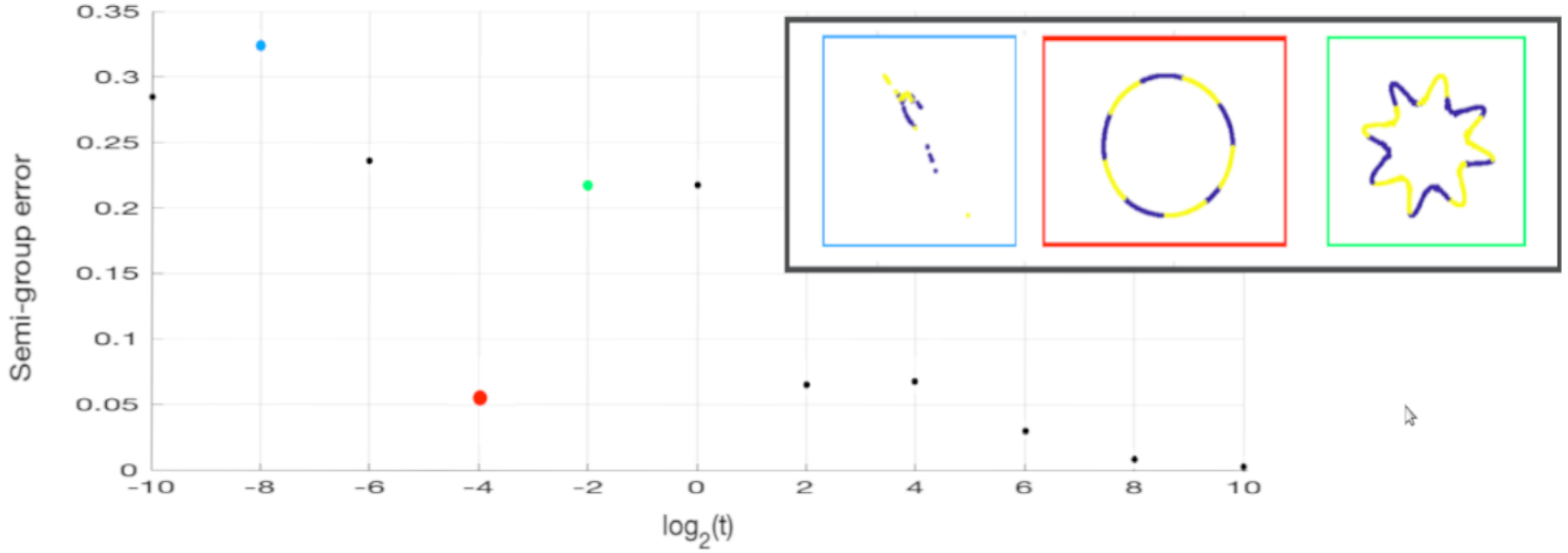
Shan Shan & ID

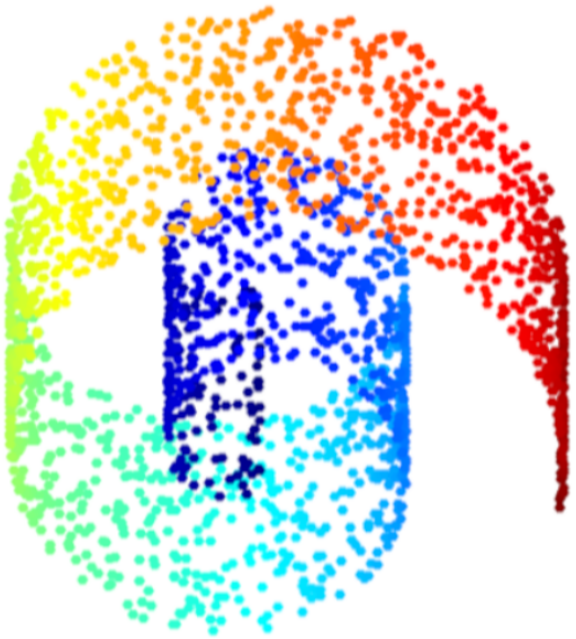


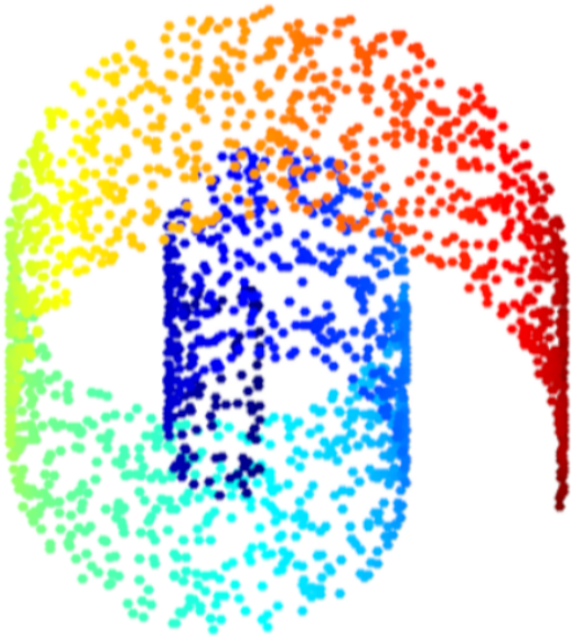


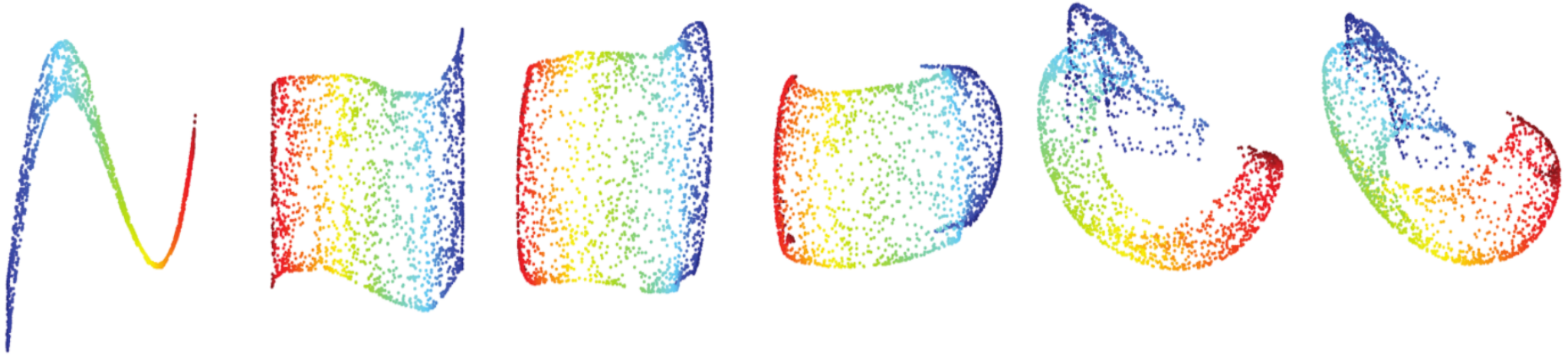


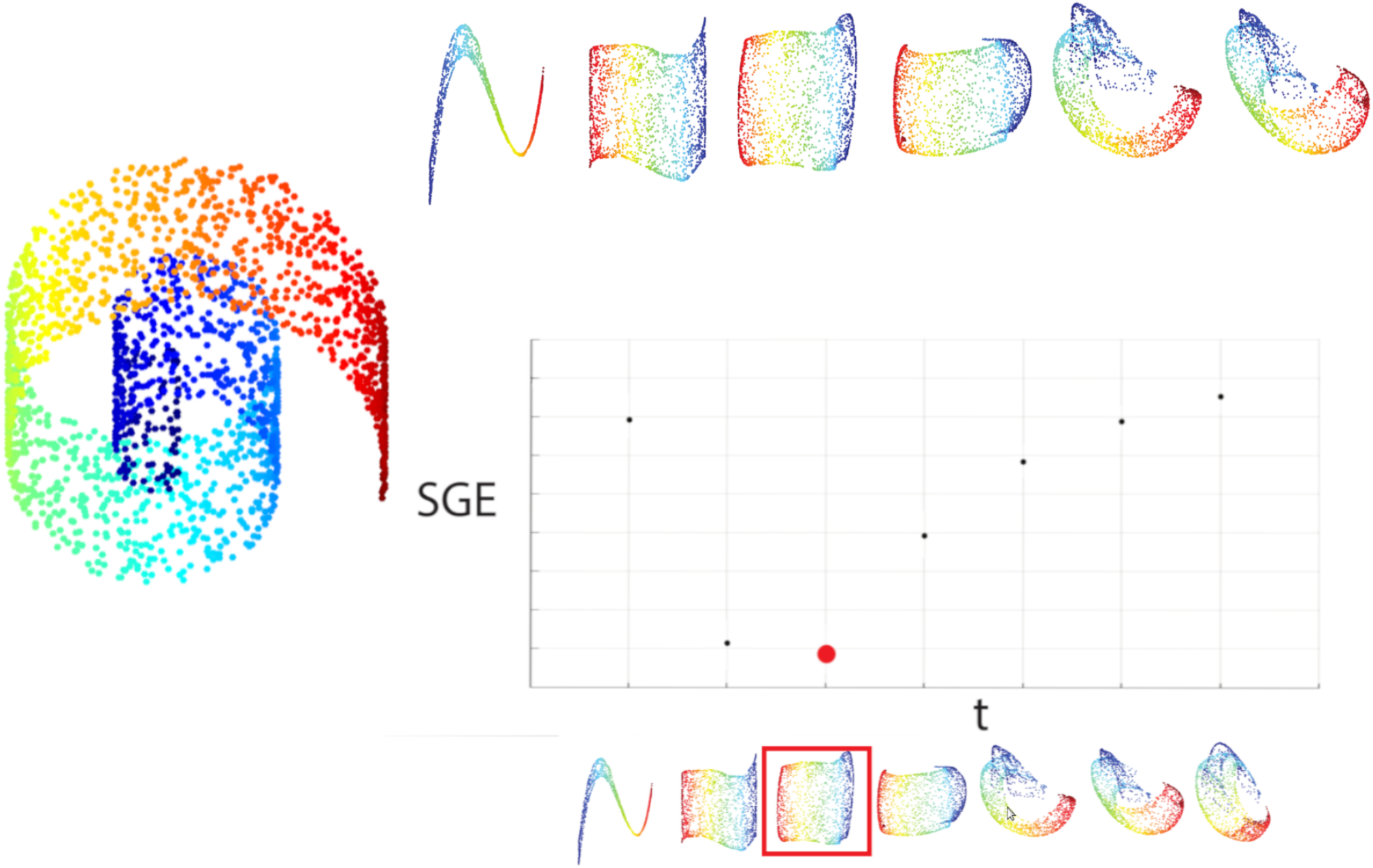



0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Semi-group error
-10
-8
-6
-4
-2
0
2
4
6
8
10
log2(t)








SGE
t

$$e^{-d_{ij}^2/2t} = \mathcal{D}_{t;ij} \qquad \tilde{\mathcal{D}}_{t;i,i} = \sum_j \mathcal{D}_{t;i,j}$$

$$W_t = \tilde{\mathcal{D}}_t^{-1} \mathcal{D}_t$$

candidate for a discrete approximation of the diffusion on the manifold

Next: spectral decomposition:

$$W_{t;i,j} \cong \sum_\ell w_t \, \psi_{\ell;i} \, \psi_{\ell;j}$$

diffusion over time T : $$\cong \sum_\ell (w_t)^{T/t} \, \psi_{\ell;i} \, \psi_{\ell;j}$$

new parametrization for data points: $$\left(w_t^\lambda \, \psi_{\ell;i} \right)_{\ell=1}^{L}$$

$$e^{-d_{ij}^2/2t} = \mathcal{D}_{t;ij} \qquad \tilde{\mathcal{D}}_{t;i,i} = \sum_j \mathcal{D}_{t;i,j}$$

$$W_t = \tilde{\mathcal{D}}_t^{-1} \mathcal{D}_t$$

candidate for a discrete approximation of the diffusion on the manifold

Next: spectral decomposition:

$$W_{t;i,j} \cong \sum_\ell w_t \, \psi_{\ell;i} \, \psi_{\ell;j}$$

new parametrization for data points: $\left(w_t^\lambda \, \psi_{\ell;i} \right)_{\ell=1}^{L}$

$\Rightarrow$ "diffusion distance" $d_{diff}^2(i,j) = \sum_{\ell=1}^{L} w_t^{2\lambda} \left| \psi_{\ell;i} - \psi_{\ell;j} \right|^2$

.

So far: transition from a "dissimilarity distance" to a distance more governed by the entirety of the dataset.

ref: Belkin & Nyogi
Coifman, Lafon, Maggioni
Shan Shan

In this talk: application to data set of biological shapes
additional use of diffusion methods to provide even better data-adapted distances.

It all started with a conversation with biologists....

Doug Boyer

Jukka Jernvall

More Precisely: biological morphologists

↓

Study Teeth & Bones of

extant & extinct animals

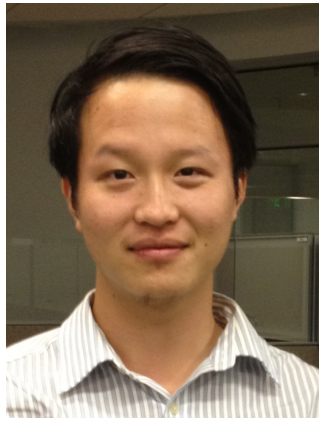
still live today — fossils

Collaborators

Rima Alaifari
ETH Zürich

Doug Boyer
Duke

Ingrid Daubechies
Duke



Tingran Gao
Duke

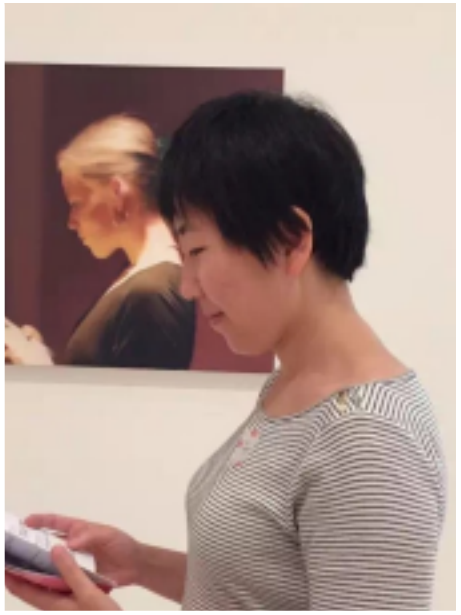
Yaron Lipman
Weizmann

Roi Poranne
ETH Zürich

Jesús Puente
J.P. Morgan

Robert Ravier
Duke

Shahar Kovalsky



Shan Shan

Nadav Dym

Chen-Yun Lin

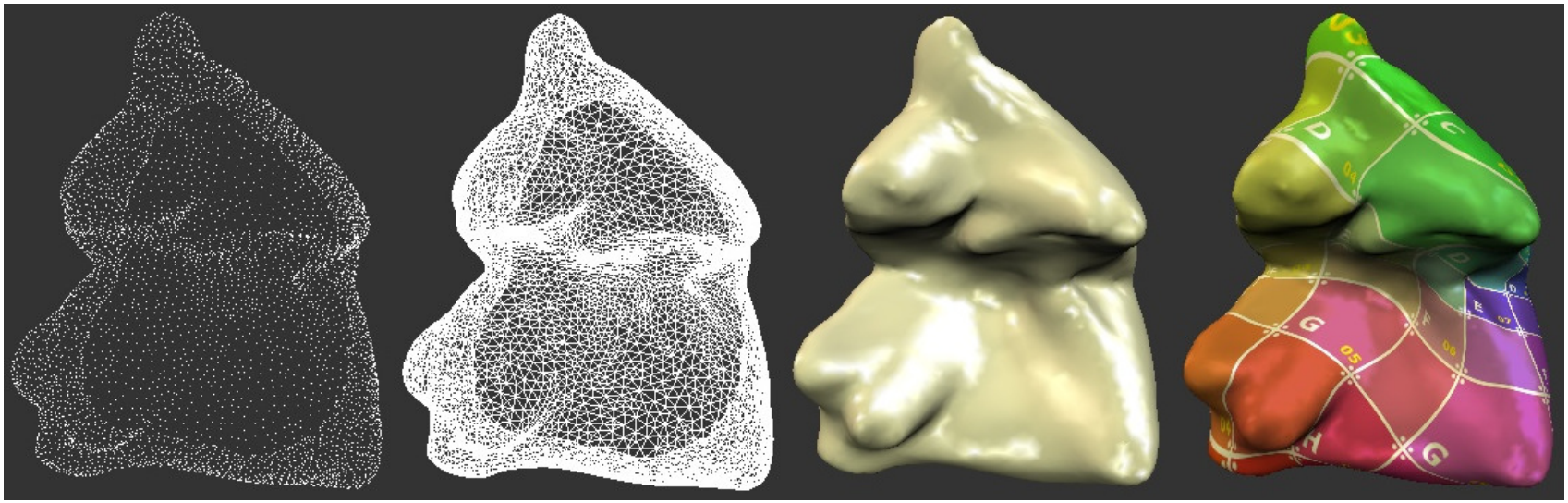
Shira Faigenbaum

Alex Winn

First: project on “complexity” of teeth

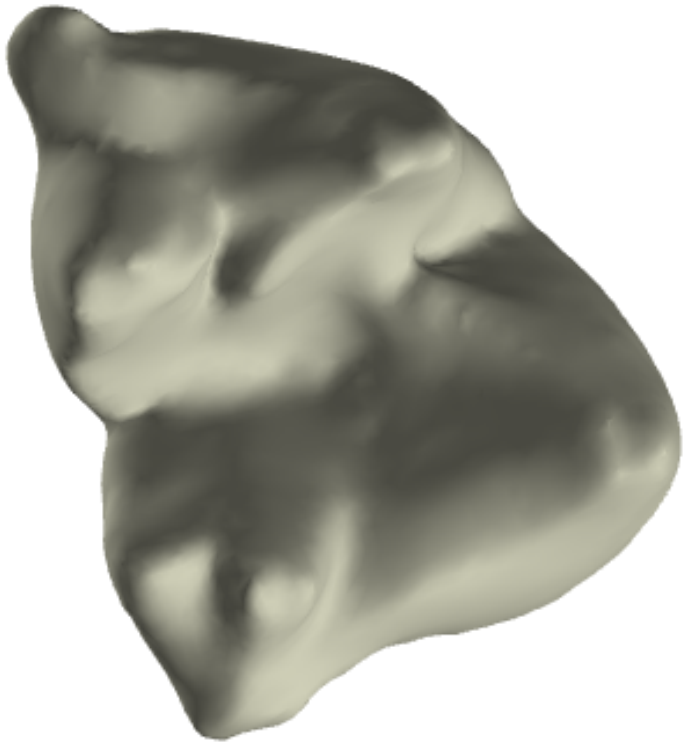
First: project on "complexity" of teeth
Then: find automatic way to compute Procrustes distances between surfaces — without landmarks

Data Acquisition



Surface reconstructed from μCT-scanned voxel data

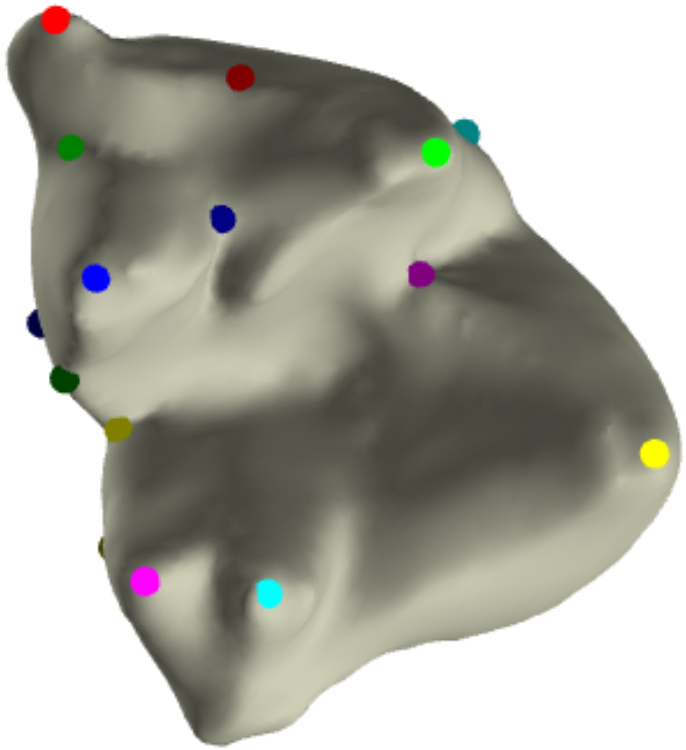
Geometric Morphometrics



second mandibular molar of a Philippine flying lemur

- Manually put k landmarks

Geometric Morphometrics

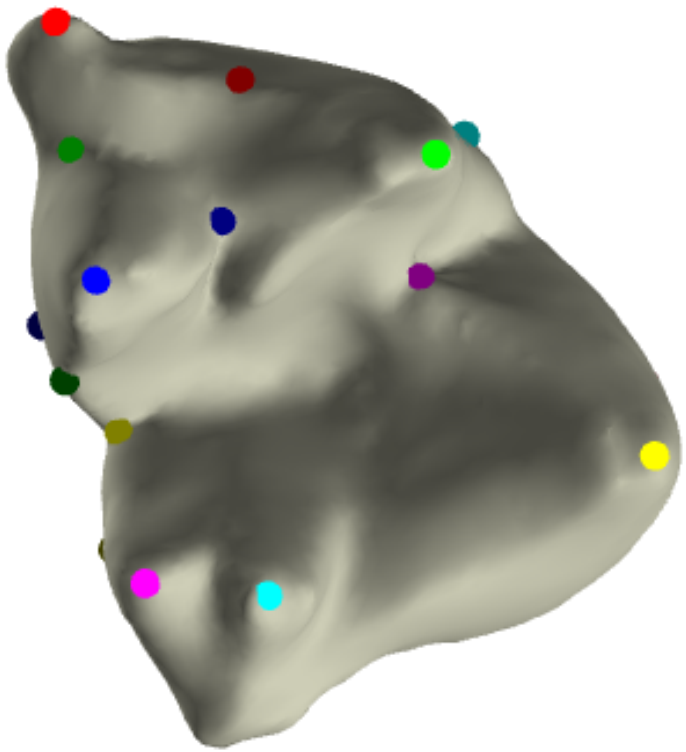


second mandibular molar of a Philippine flying lemur

- Manually put k landmarks

$$p_1, p_2, \cdots, p_k$$

Geometric Morphometrics



second mandibular molar of a Philippine flying lemur

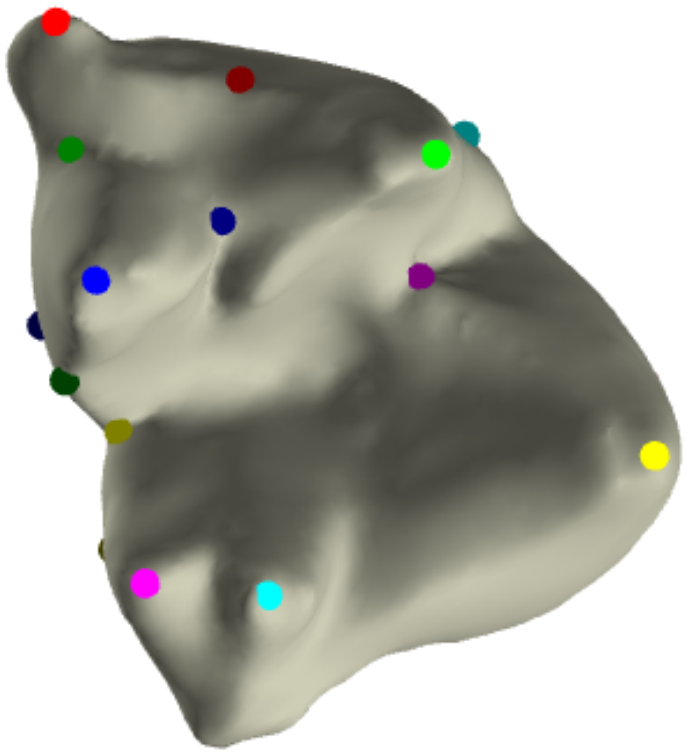
- Manually put k landmarks

$$p_1, p_2, \cdots, p_k$$

- Use spatial coordinates of the landmarks as features

$$p_j = (x_j, y_j, z_j), \ j = 1, \cdots, k$$

Geometric Morphometrics



second mandibular molar of a Philippine flying lemur

- Manually put k landmarks

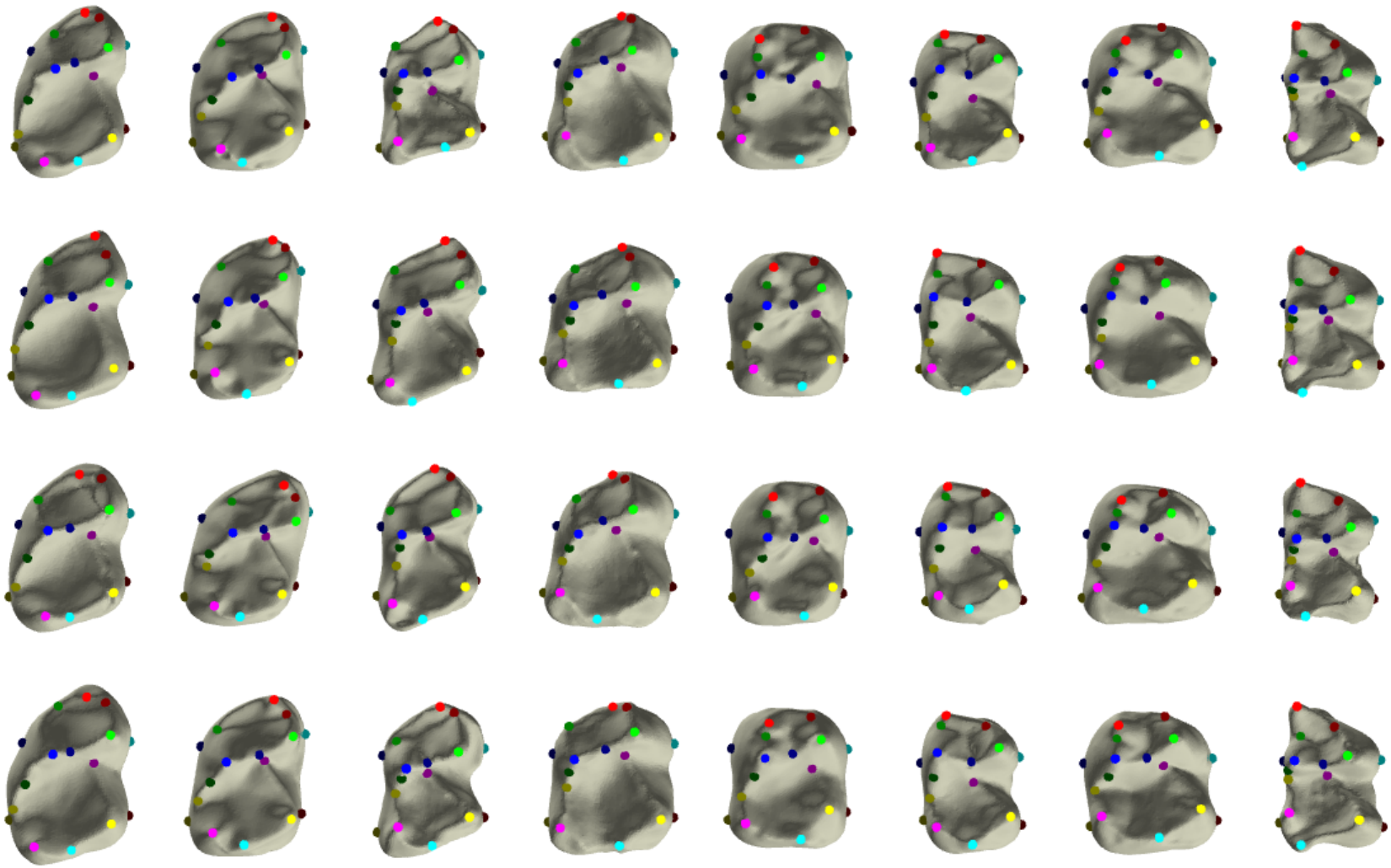
$$p_1, p_2, \cdots, p_k$$

- Use spatial coordinates of the landmarks as features

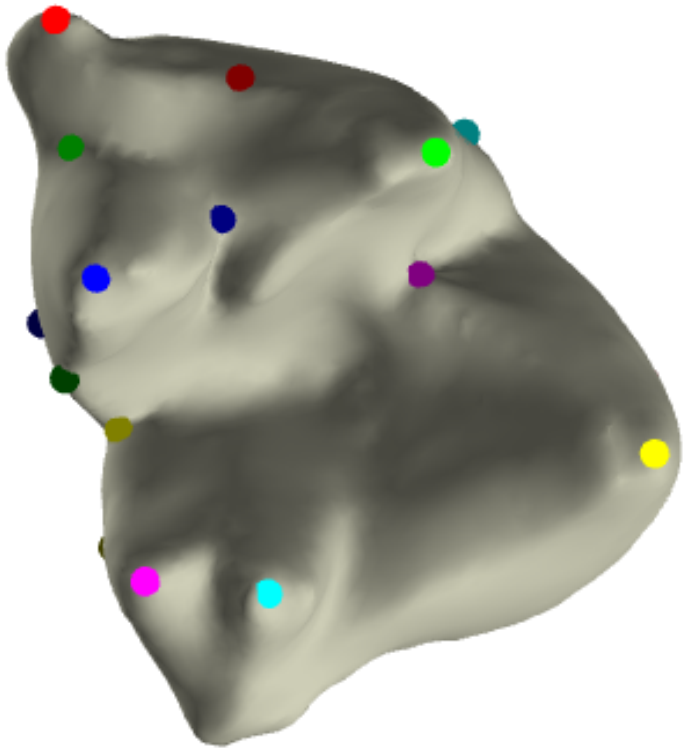
$$p_j = (x_j, y_j, z_j), \ j = 1, \cdots, k$$

- Represent a shape in $\mathbb{R}^{3\times k}$

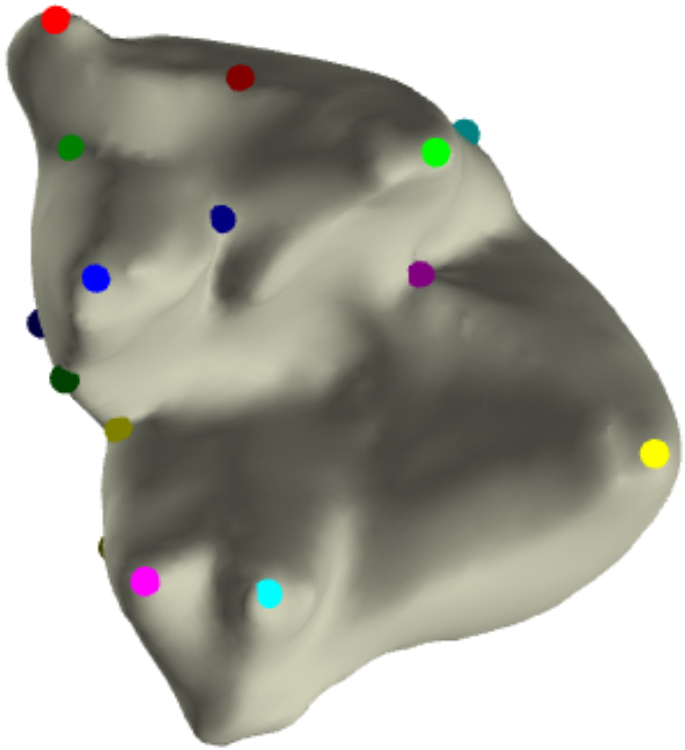
The *Shape Space* of k landmarks in $\mathbb{R}^3$



Geometric Morphometrics: Limitation of Landmarks

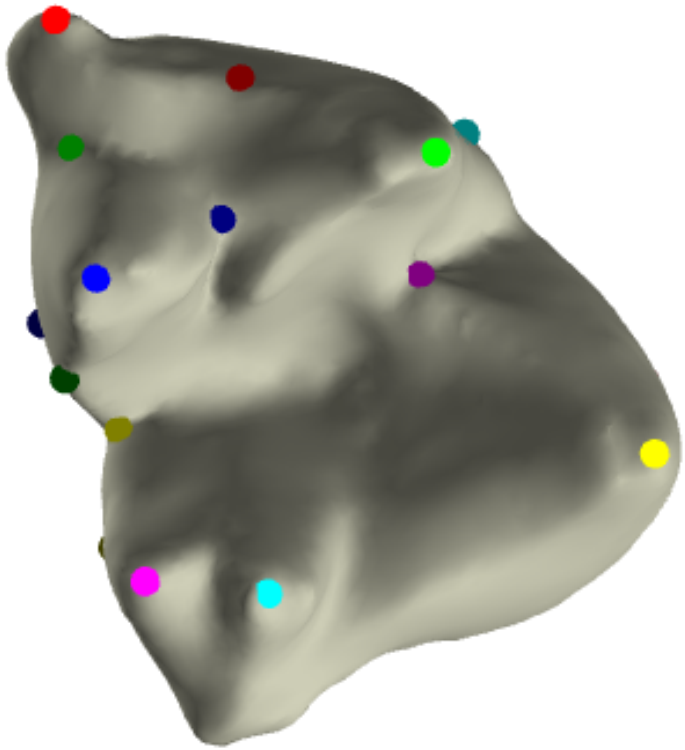


Geometric Morphometrics: Limitation of Landmarks



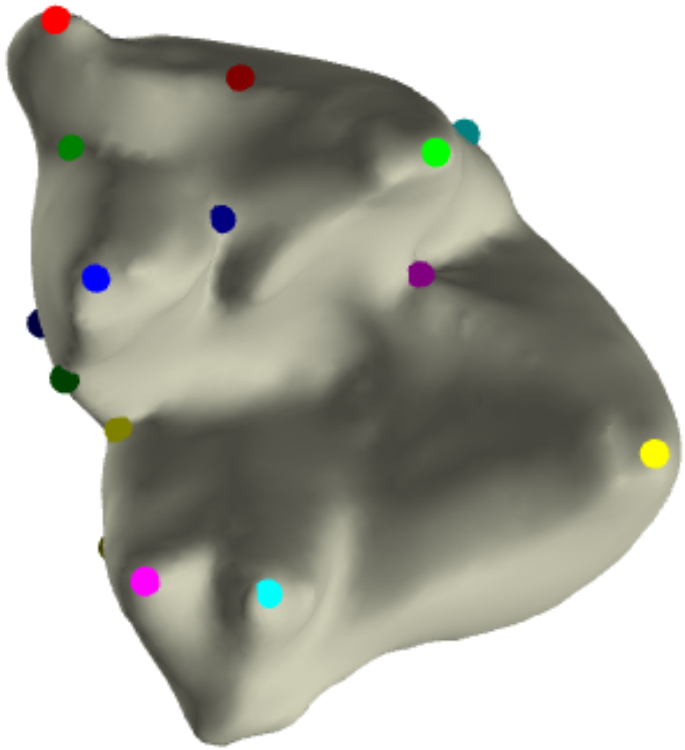
- Landmark Placement: tedious and time-consuming

Geometric Morphometrics: Limitation of Landmarks



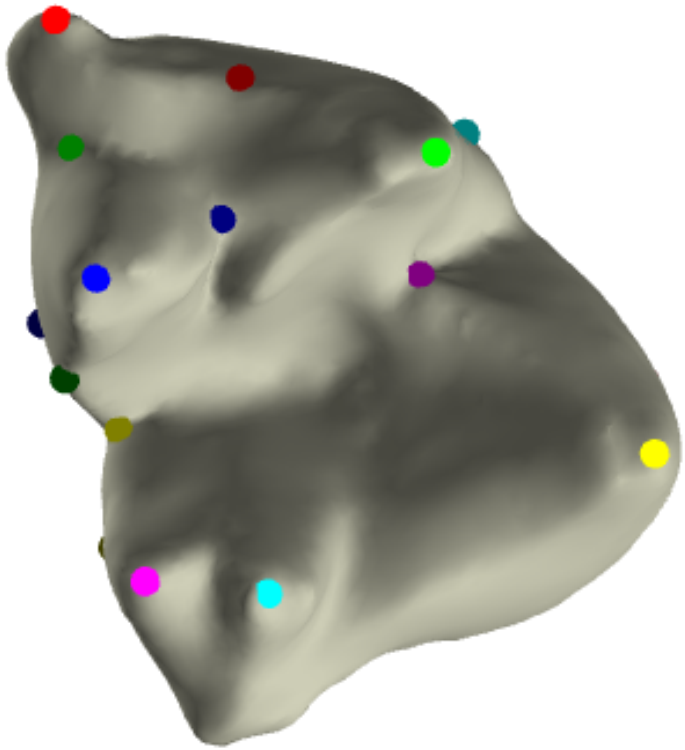
- Landmark Placement: tedious and time-consuming
- Fixed Number of Landmarks: lack of flexibility

Geometric Morphometrics: Limitation of Landmarks



- Landmark Placement: tedious and time-consuming
- Fixed Number of Landmarks: lack of flexibility
- Domain Knowledge: high degree of expertise needed, not easily accessible

Geometric Morphometrics: Limitation of Landmarks



- Landmark Placement: tedious and time-consuming
- Fixed Number of Landmarks: lack of flexibility
- Domain Knowledge: high degree of expertise needed, not easily accessible
- Subjectivity: debates exist even among experts

First: project on "complexity" of teeth
Then: find automatic way to compute Procrustes distances between surfaces — without landmarks

Landmarked Teeth $\longrightarrow$

$$d^2_{Procrustes}(S_1, S_2) = \min_{R \text{ rigid tr.}} \sum_{j=1}^{J} \|R(x_j) - y_j\|^2$$

First: project on "complexity" of teeth
Then: find automatic way to compute Procrustes distances between surfaces — without landmarks

Landmarked Teeth $\longrightarrow$

$$d^2_{Procrustes}(S_1, S_2) = \min_{R \text{ rigid tr.}} \sum_{j=1}^{J} \|R(x_j) - y_j\|^2$$

Find way to compute a distance that does as well, for biological purposes, as Procrustes distance, based on expert-placed landmarks, automatically?

First: project on "complexity" of teeth
Then: find automatic way to compute Procrustes distances between surfaces — without landmarks

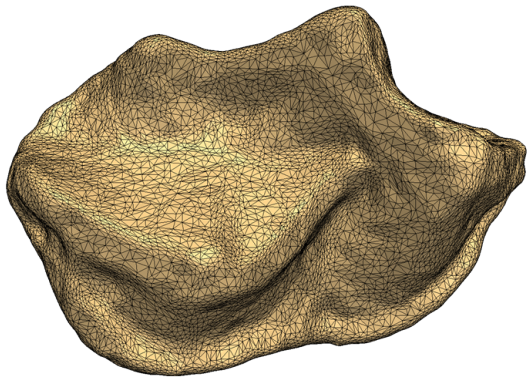
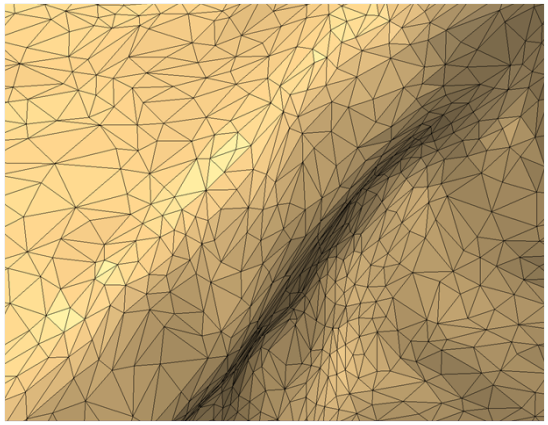



Landmarked Teeth $\longrightarrow$

$$d^2_{Procrustes}(S_1, S_2) = \min_{R \text{ rigid tr.}} \sum_{j=1}^{J} \|R(x_j) - y_j\|^2$$

Find way to compute a distance that does as well, for biological purposes, as Procrustes distance, based on expert-placed landmarks, automatically?

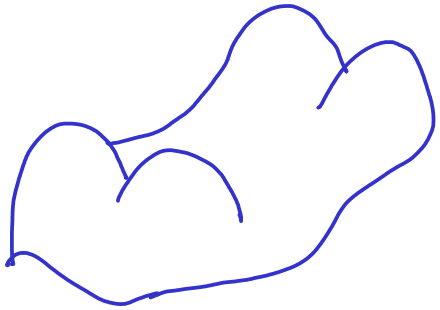
examples: finely discretized triangulated surfaces

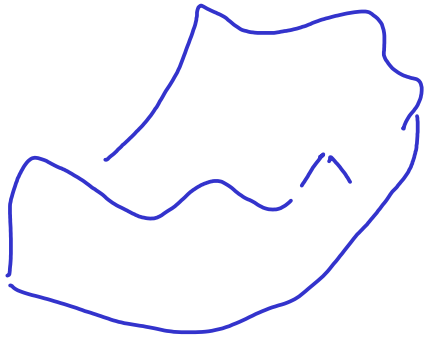



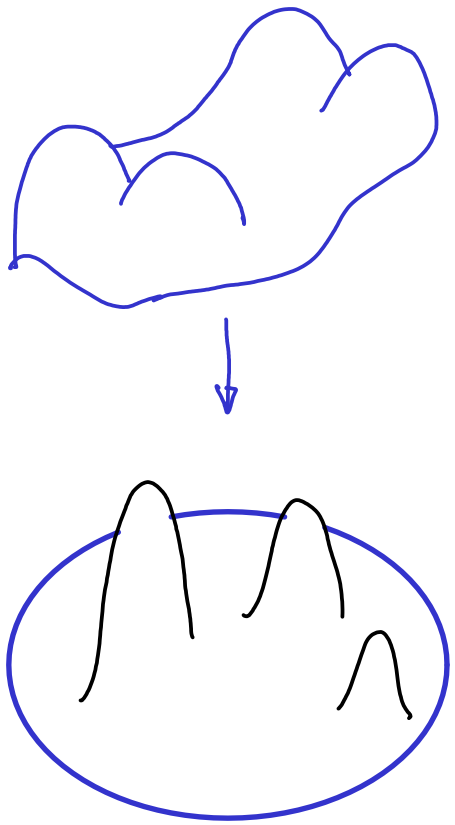
We defined 2 different distances

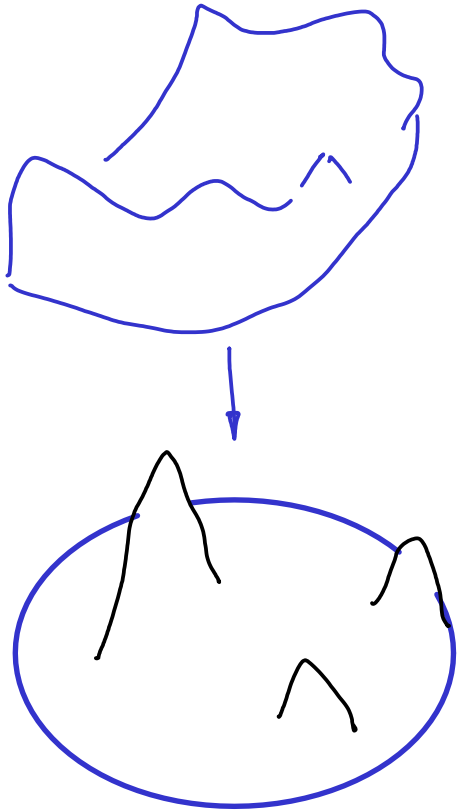
$d_{\mathrm{cWn}}(S_1, S_2)$: conformal flattening
comparison of neighborhood geometry
optimal mass transport

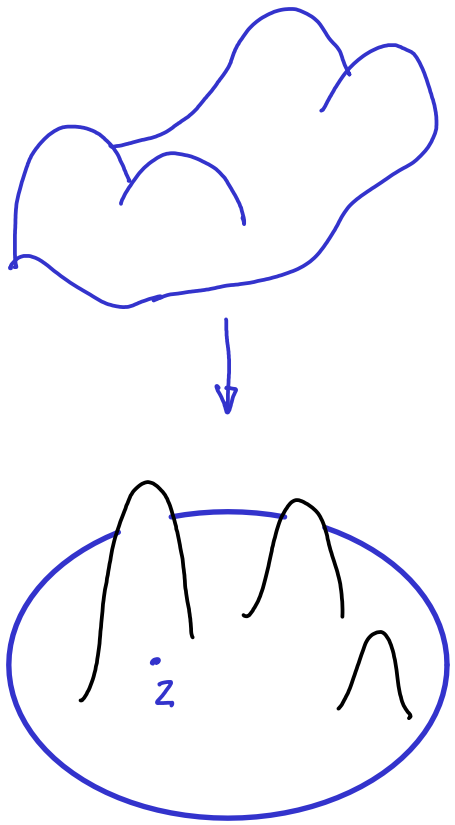
$d_{\mathrm{cP}}(S_1, S_2)$: continuous Procrustes distance

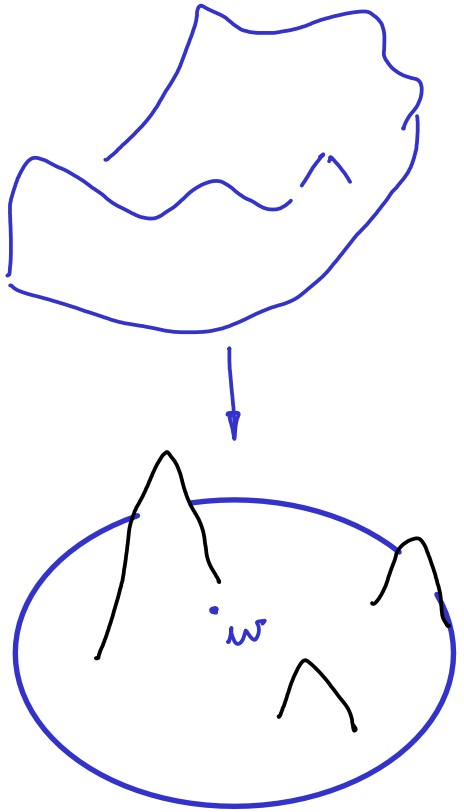


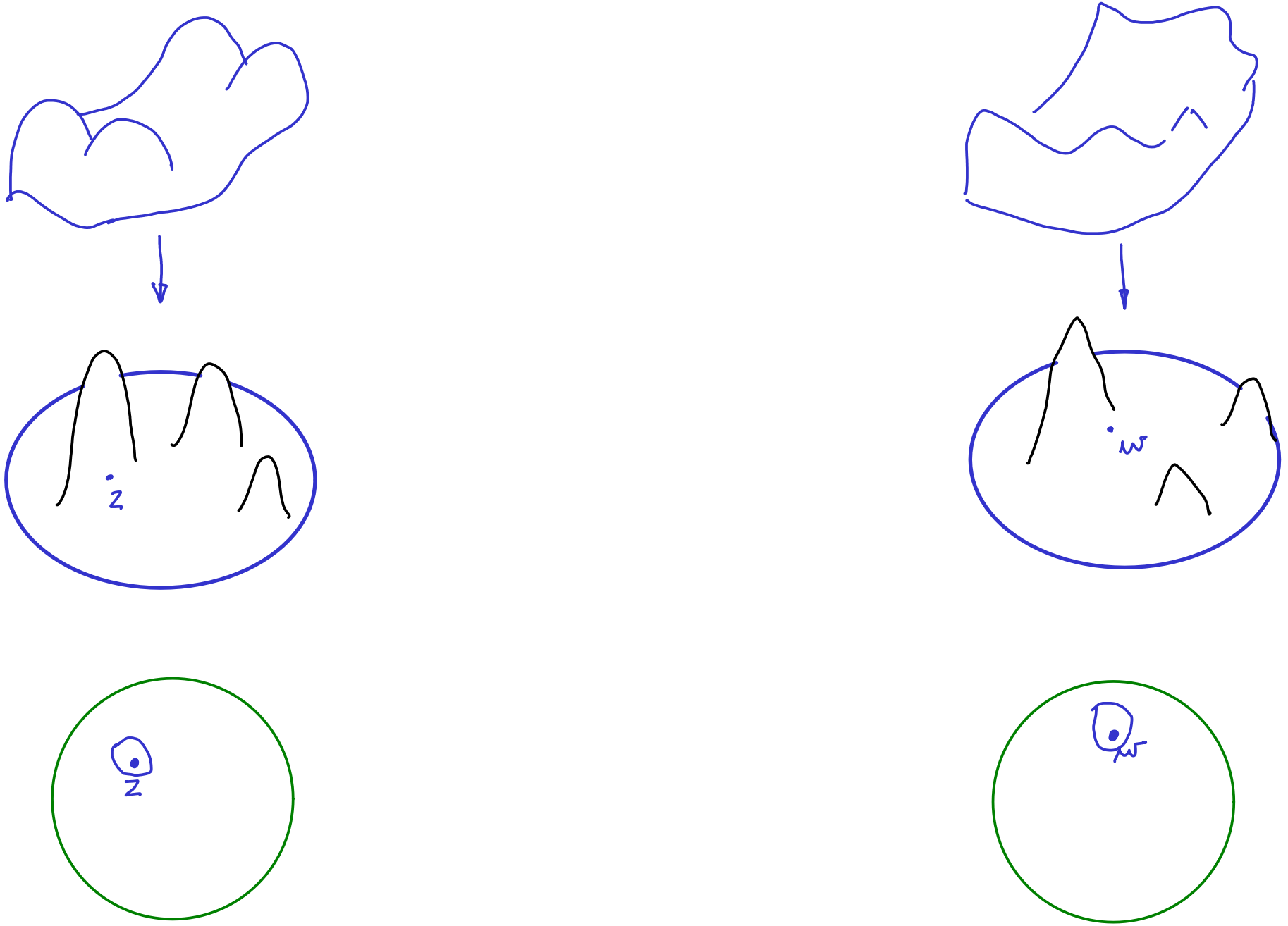


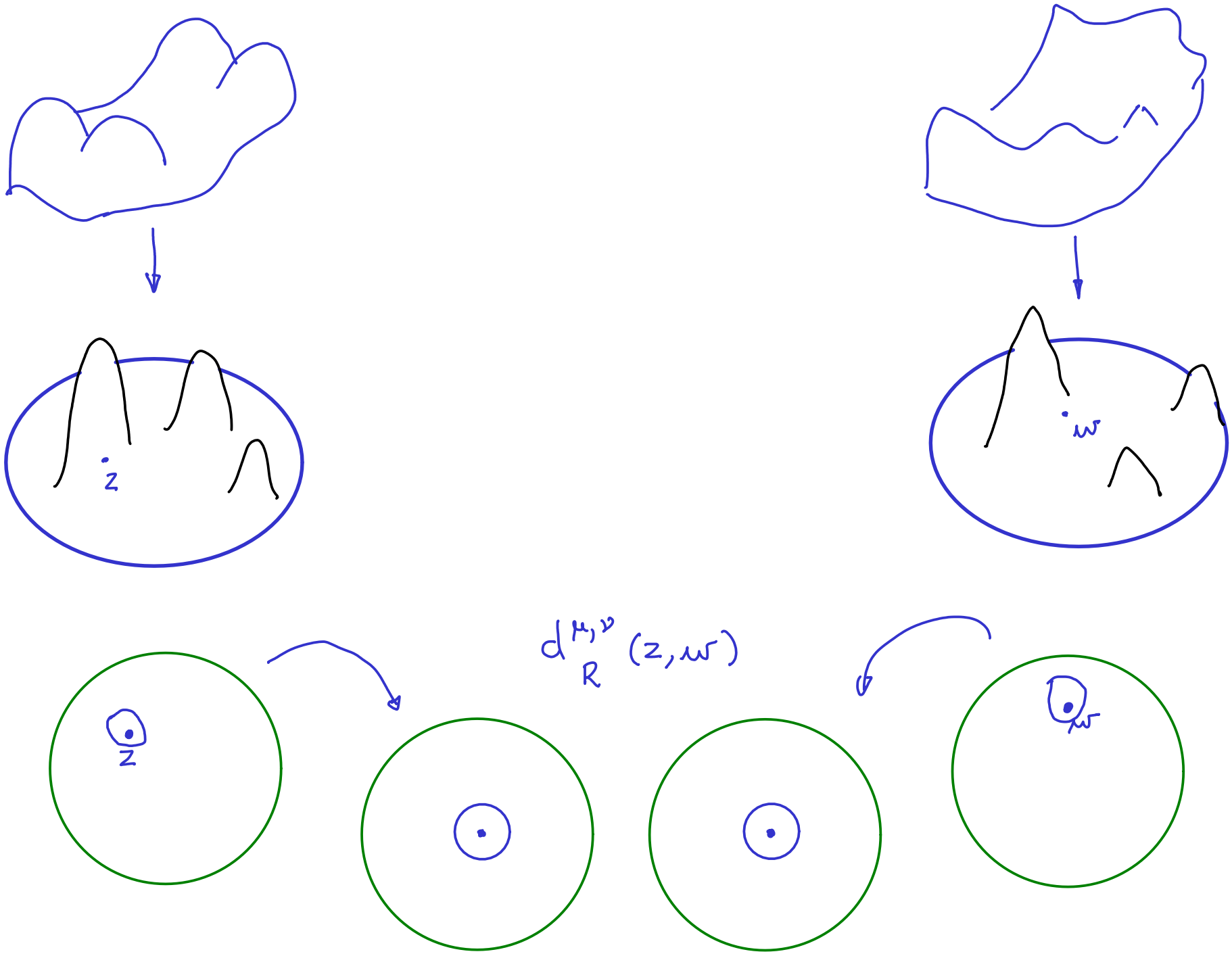


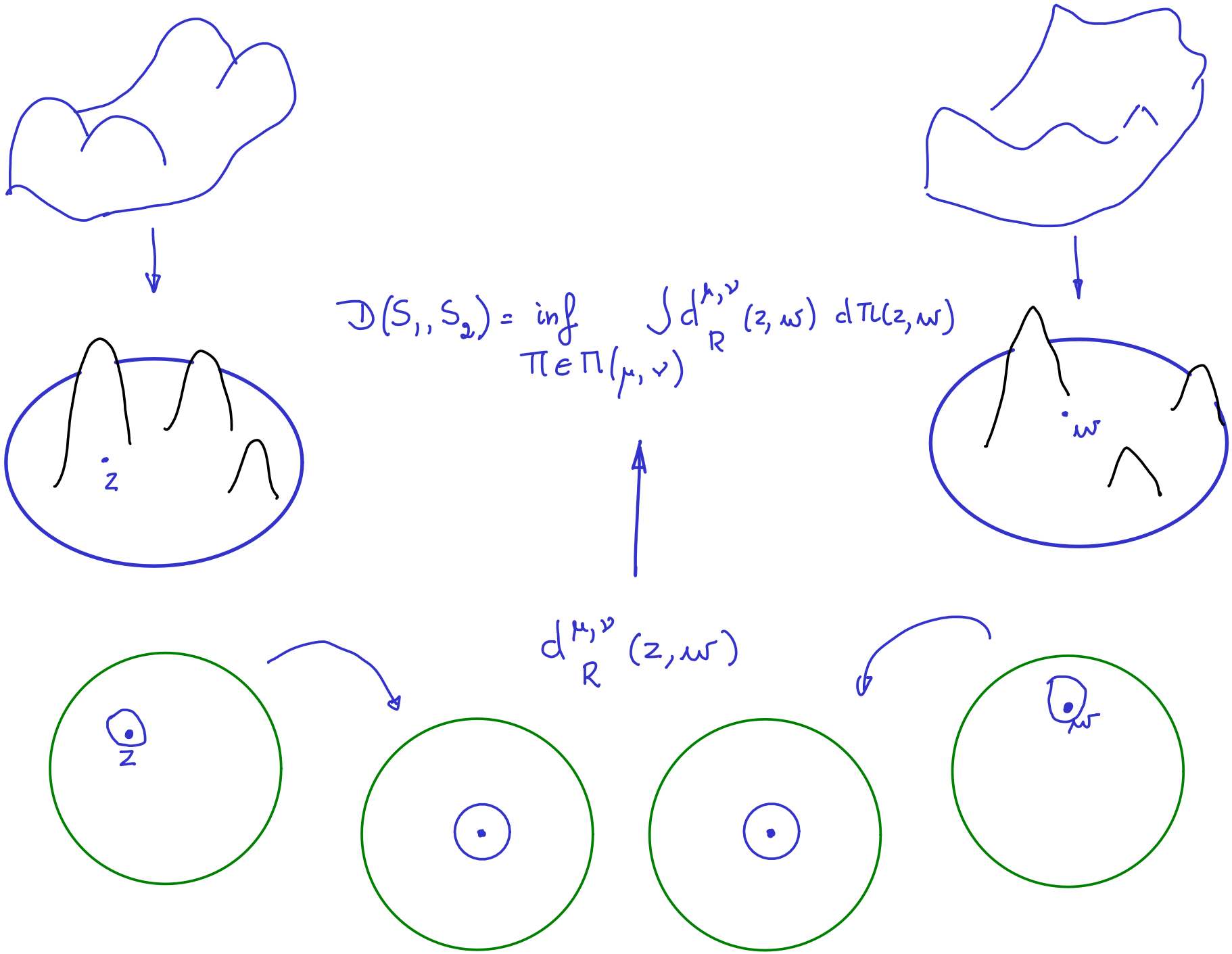



z

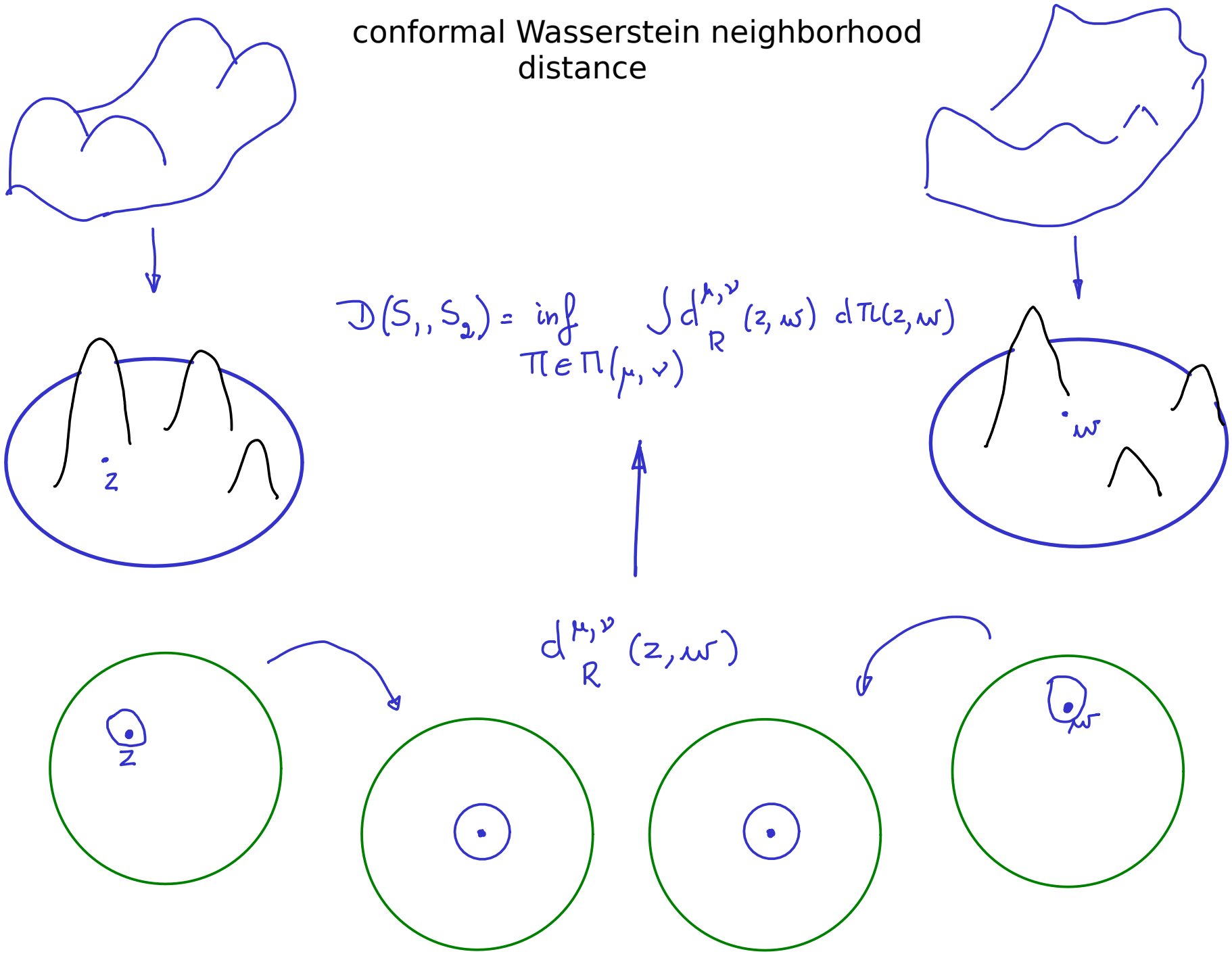

w


z
w
z
w


z
w
$d_R^{\mu,\nu}(z,w)$
z
w


$\mathcal{D}(S_1, S_2) = \inf_{\pi \in \Pi(\mu,\nu)} \int d_R^{\mu,\nu}(z,w)\, d\pi(z,w)$
z
w
$d_R^{\mu,\nu}(z,w)$
z
w

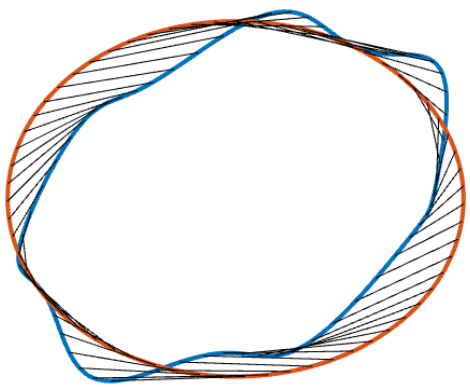
conformal Wasserstein neighborhood distance



Continuous Procrustes Distance (cPD)

$$D_{\mathrm{cP}}(S_1, S_2) = \left(\int_{S_1} \| \quad x \quad - \mathcal{C}(x) \|^2 \, d\mathrm{vol}_{S_1}(x) \right)^{\frac{1}{2}},$$

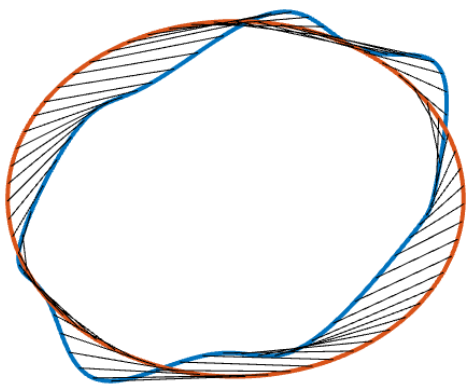
where $\mathcal{C} : S_1 \to S_2$ is an area-preserving diffeomorphism.

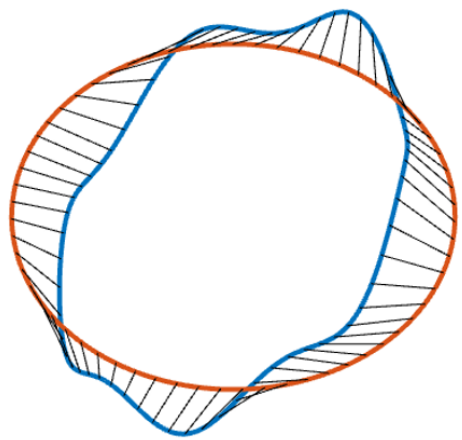


Continuous Procrustes Distance (cPD)

$$D_{\mathrm{cP}}(S_1, S_2) = \left(\inf_{R \in \mathbb{E}(3)} \int_{S_1} \| R(x) - \mathcal{C}(x) \|^2 \, d\mathrm{vol}_{S_1}(x) \right)^{\frac{1}{2}},$$

where $\mathcal{C} : S_1 \to S_2$ is an area-preserving diffeomorphism, and $\mathbb{E}_3$ is the Euclidean group on $\mathbb{R}^3$.

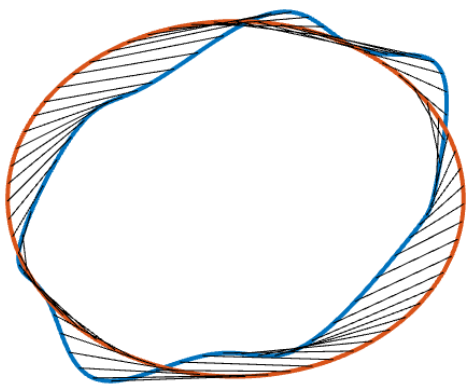
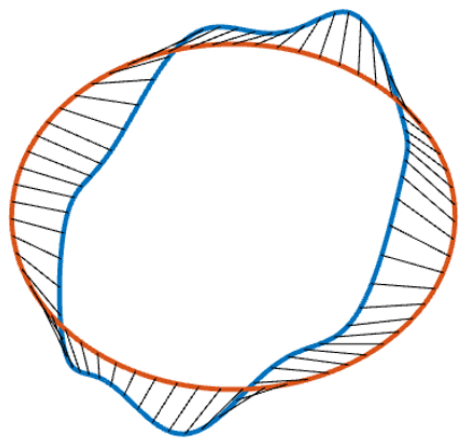
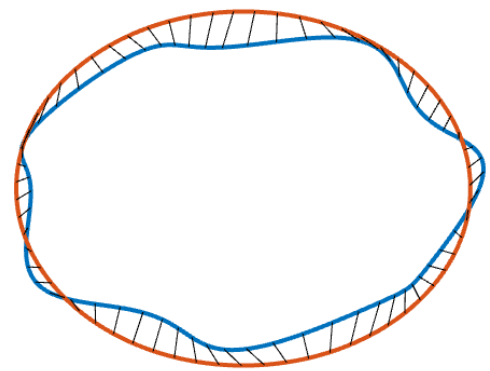




Continuous Procrustes Distance (cPD)

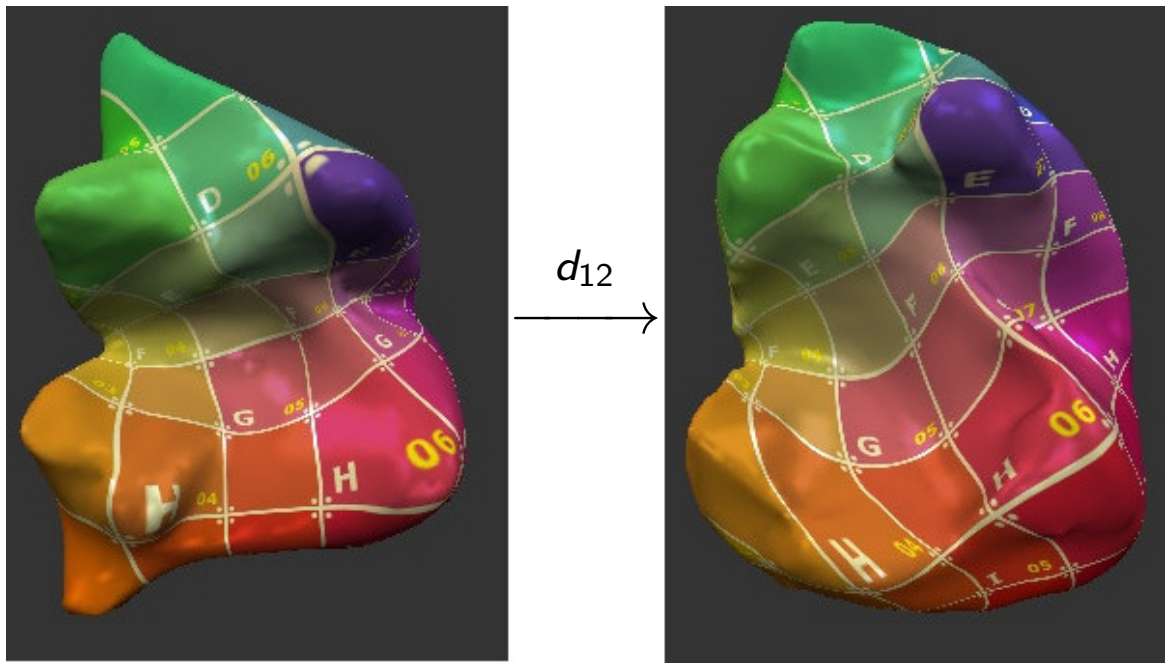
$$D_{\mathrm{cP}}(S_1, S_2) = \left(\inf_{\mathcal{C} \in \mathcal{A}(S_1, S_2)} \inf_{R \in \mathbb{E}(3)} \int_{S_1} \| R(x) - \mathcal{C}(x) \|^2 \, d\mathrm{vol}_{S_1}(x) \right)^{\frac{1}{2}},$$

where $\mathcal{A}(S_1, S_2)$ is the set of area-preserving diffeomorphisms between S_1 and S_2, and $\mathbb{E}_3$ is the Euclidean group on $\mathbb{R}^3$.

Continuous Procrustes Distance (cPD)

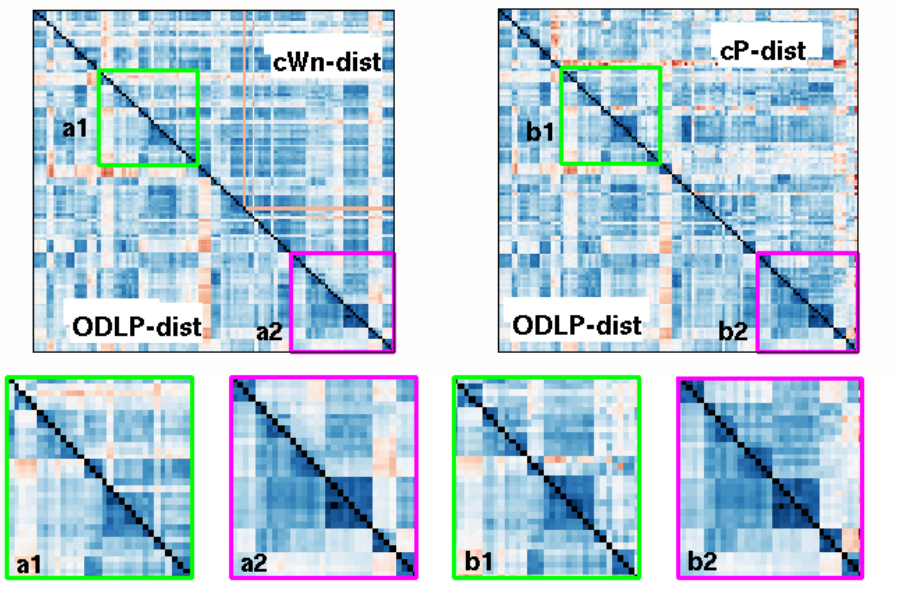
$$d_{cP}(S_1, S_2) = \inf_{\mathcal{C} \in \mathscr{A}} \inf_{R \in \mathbb{E}_3} \left(\int_{S_1} \| R(x) - \mathcal{C}(x) \|^2 \, d\mathrm{vol}_{S_1}(x) \right)^{1/2}$$

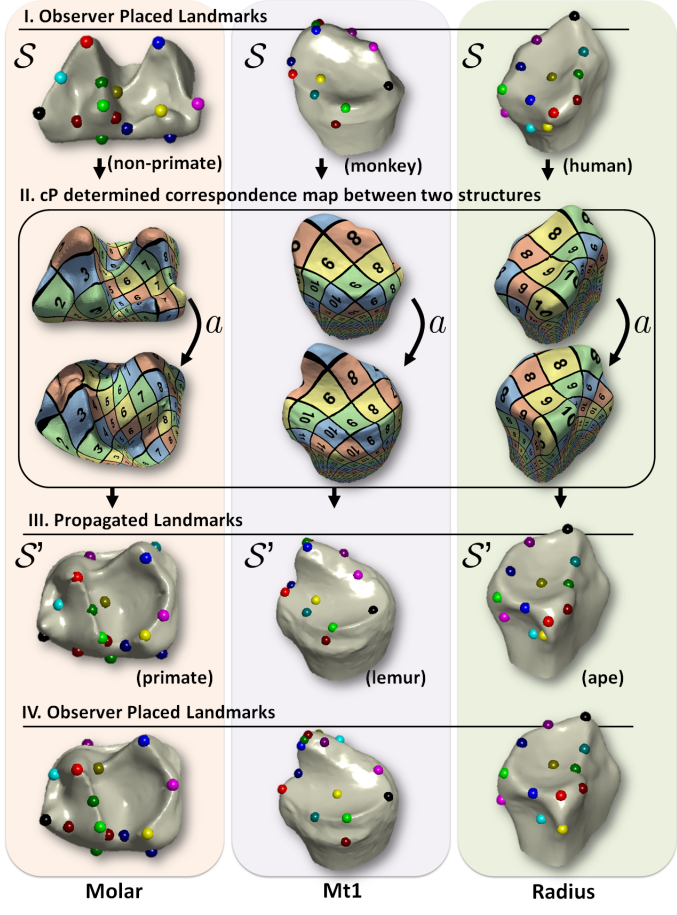


We defined 2 different distances

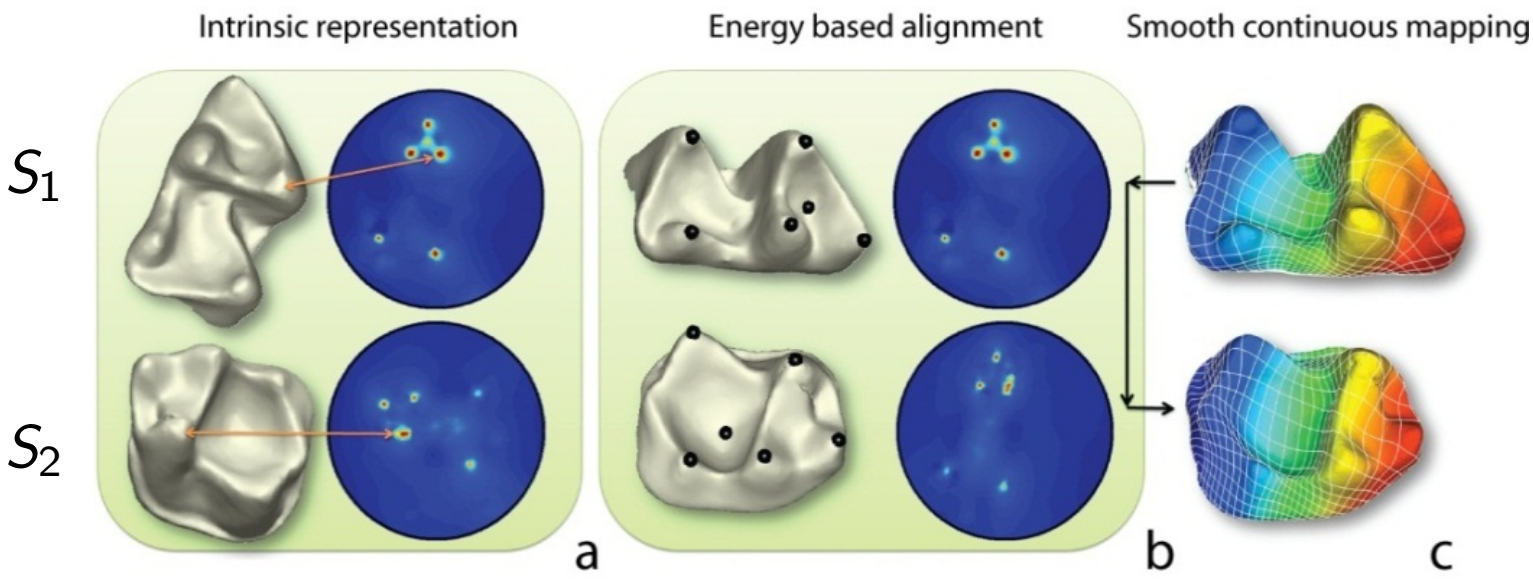
$d_{\mathrm{cWn}}(S_1, S_2)$: conformal flattening
comparison of neighborhood geometry
optimal mass transport

$d_{\mathrm{cP}}(S_1, S_2)$: continuous Procrustes distance

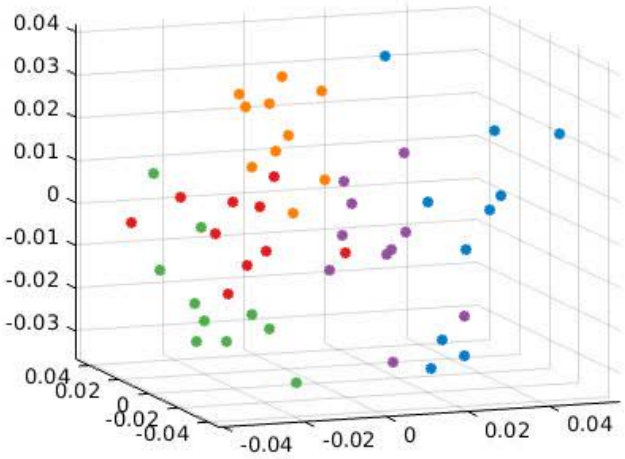




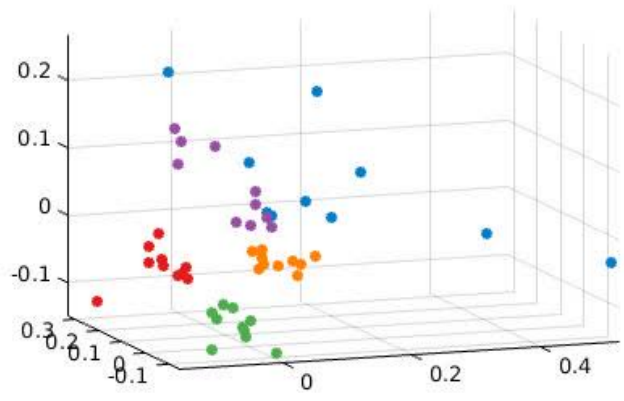
Bypass Explicit Feature Extraction



MDS for cPD & DD

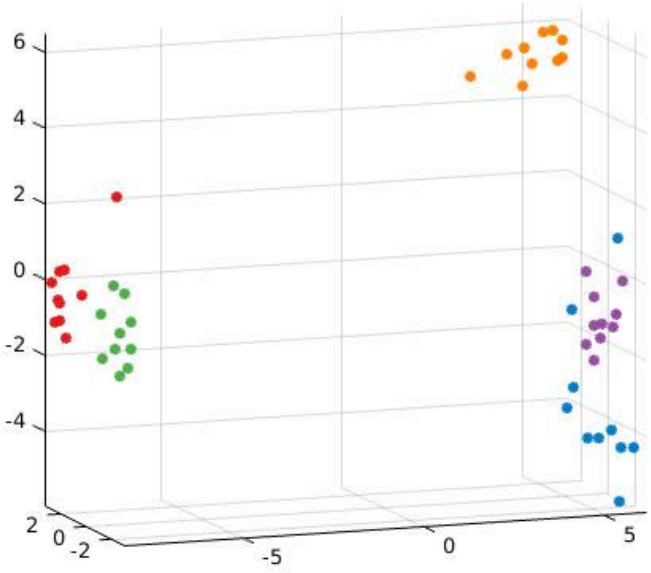


cPD

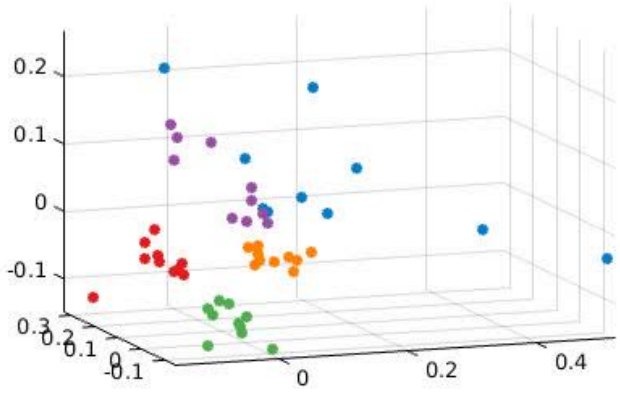


DD

Even better can be obtained!

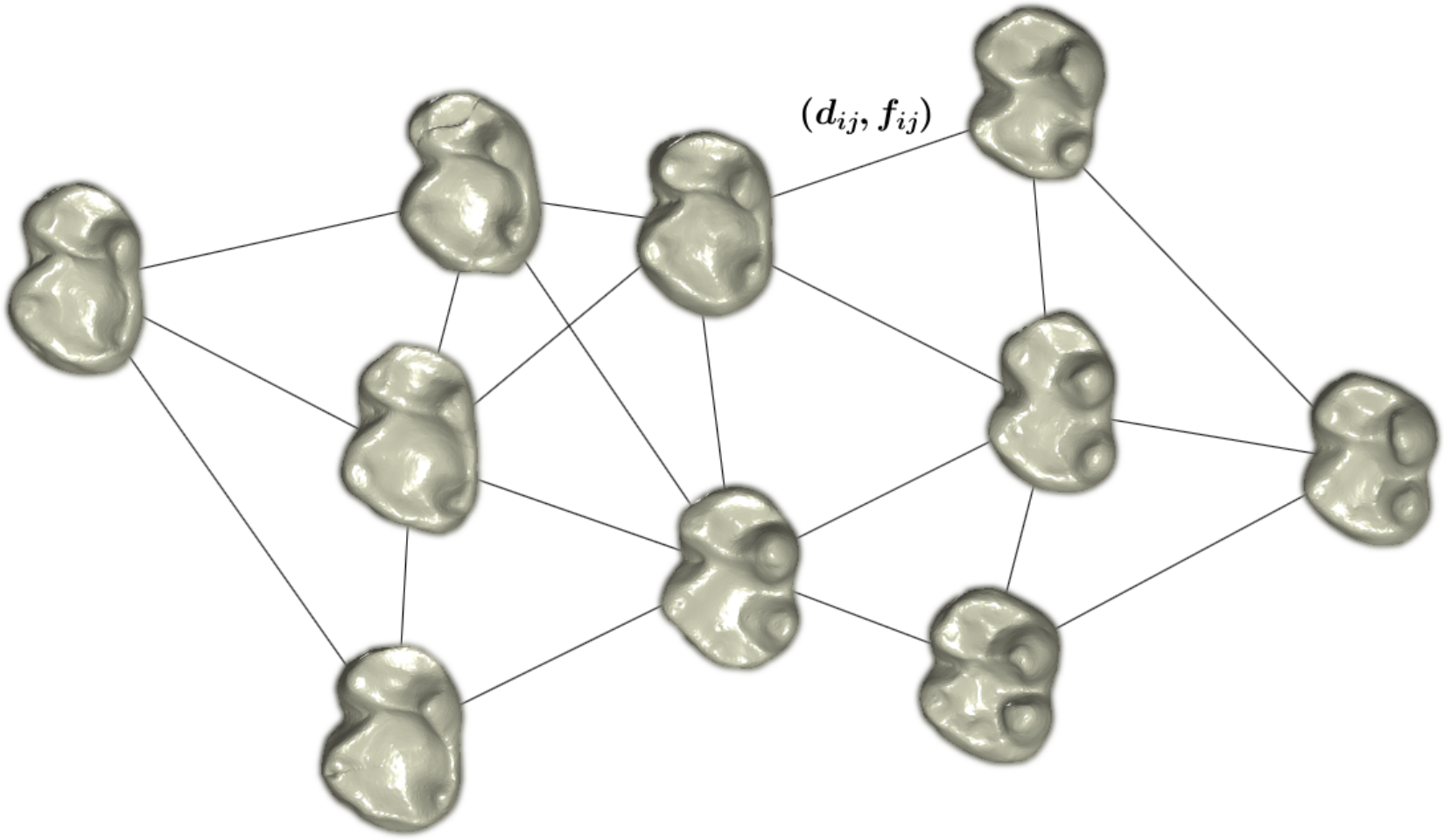


HBDD



DD

to get Diffusion Distance : used local distances
knitted together
→ spectral parametrization
→ distance.


(d_{ij}, f_{ij})

to get Diffusion Distance : used local distances
knitted together
→ spectral parametrization
→ distance.

mappings were used only to obtain numerical values for local distances.

to get Diffusion Distance : used local distances
knitted together
→ spectral parametrization
→ distance.

mappings were used only to obtain numerical values for local distances.

but they can do much more for us!

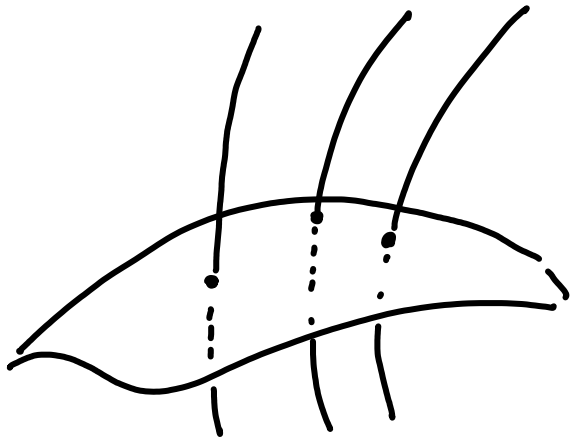
in fact: we have a fiber bundle.
(because of the mappings)

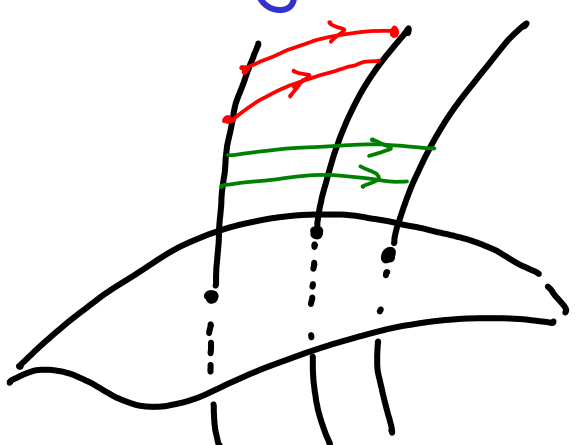
to get Diffusion Distance : used local distances
knitted together
→ spectral parametrization
→ distance.

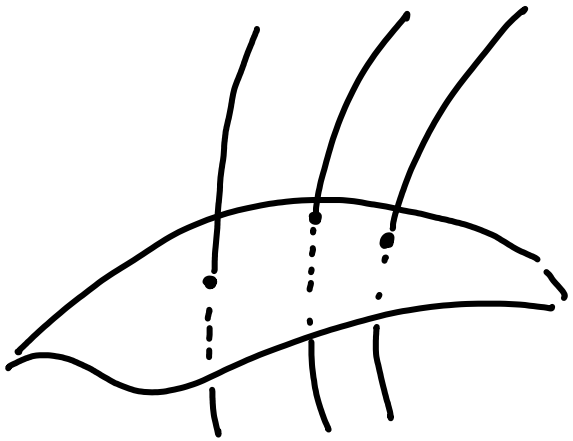
mappings were used only to obtain numerical values for local distances.

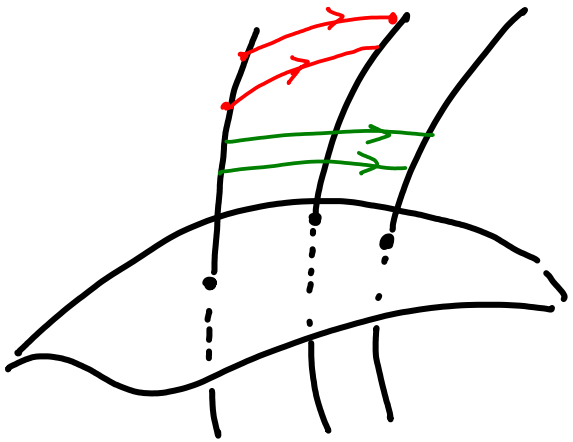
but they can do much more for us!

in fact: we have a fiber bundle.
(because of the mappings)





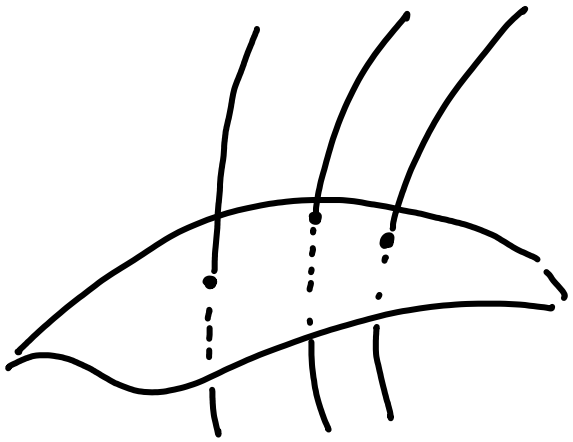


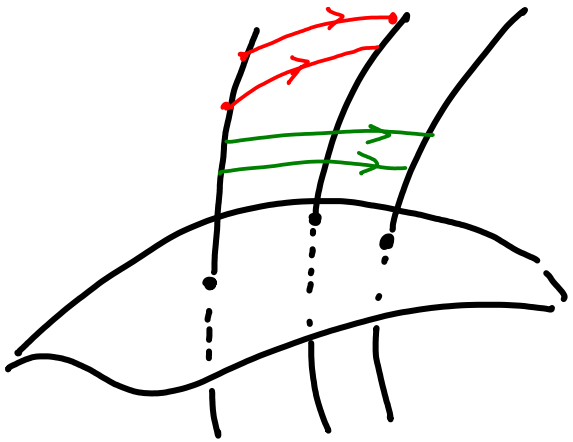


Connection.

↓

family of mappings between fibers





Connection.

↓

family of mappings between fibers

Tingran Gao: use these to define a much more detailed diffusion structure on the higher-dimensional object

→ "project" at a later stage to obtain "horizontal" part of diffusion.

Horizontal Random Walk on a *Fibre Bundle*

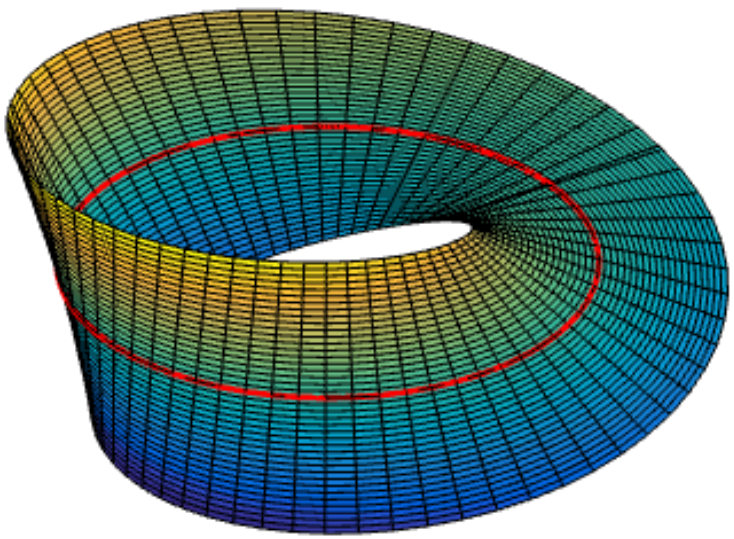
Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

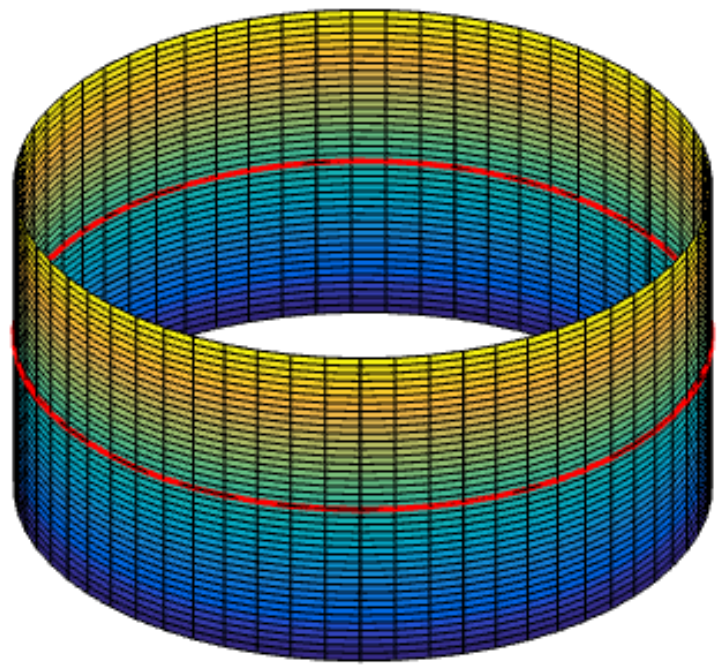
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold

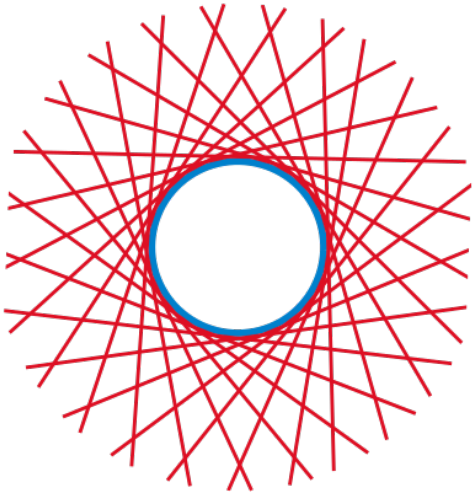
Horizontal Random Walk on a *Fibre Bundle*

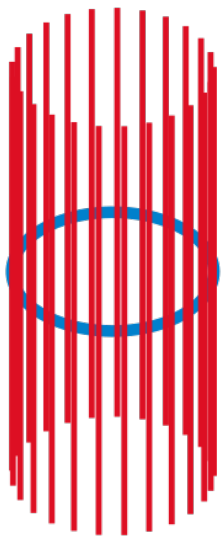
Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

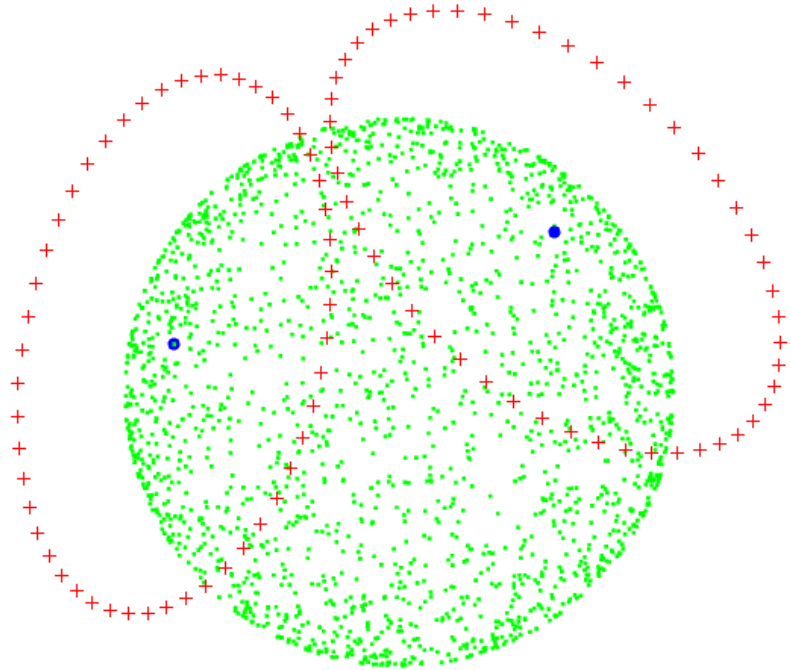
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$







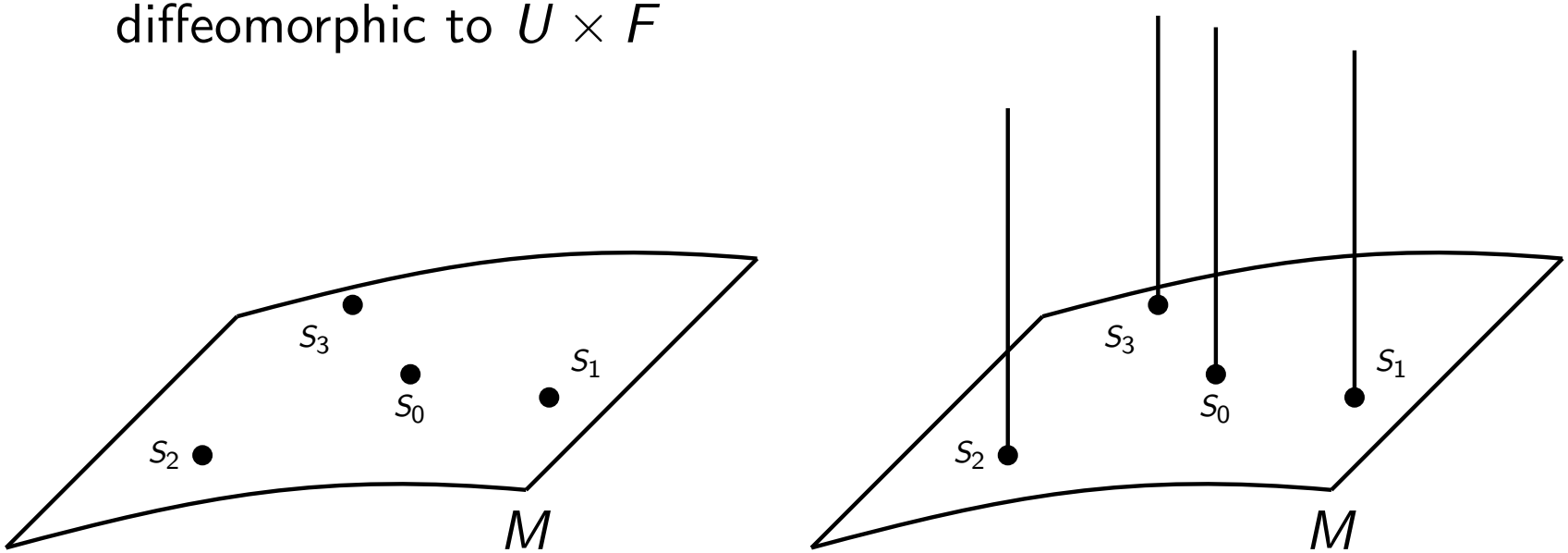




Horizontal Random Walk on a *Fibre Bundle*

Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

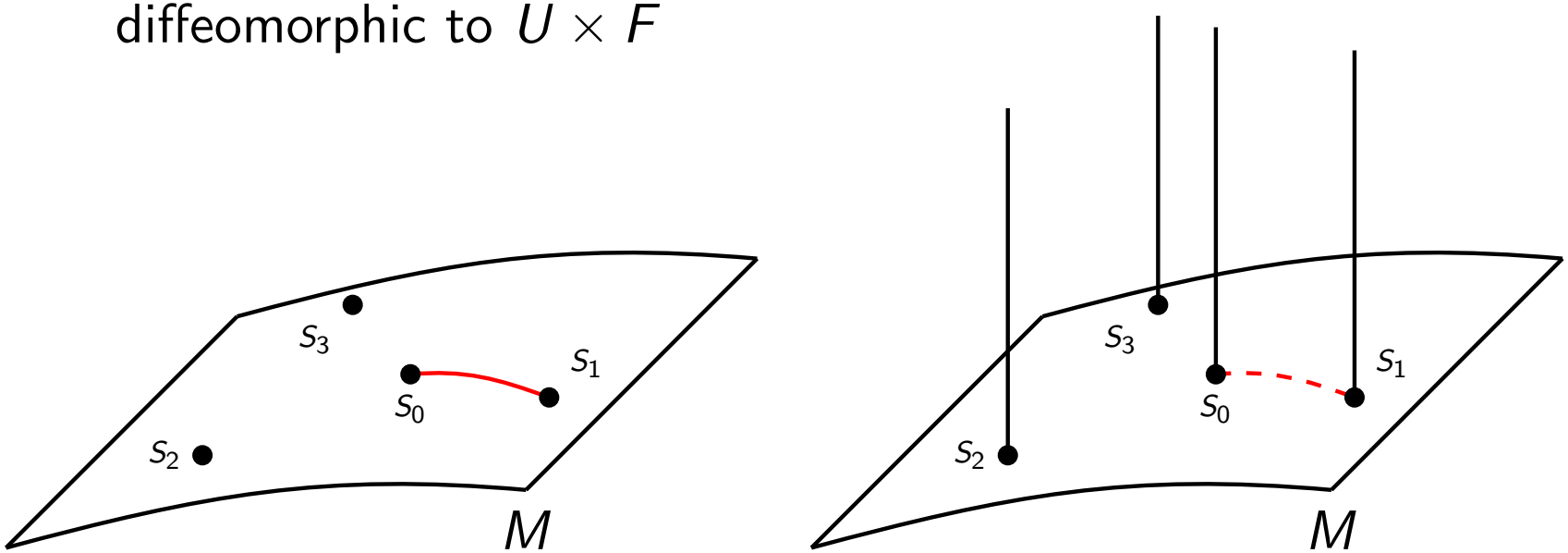
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$



Horizontal Random Walk on a *Fibre Bundle*

Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

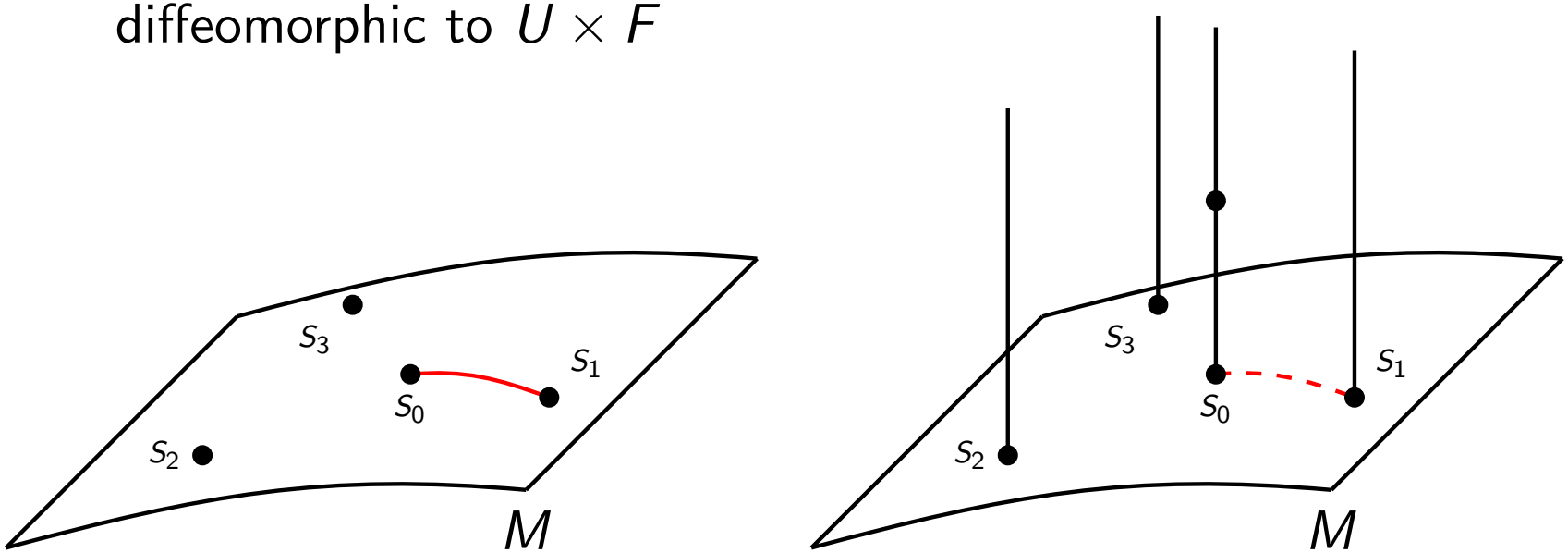
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$



Horizontal Random Walk on a *Fibre Bundle*

Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

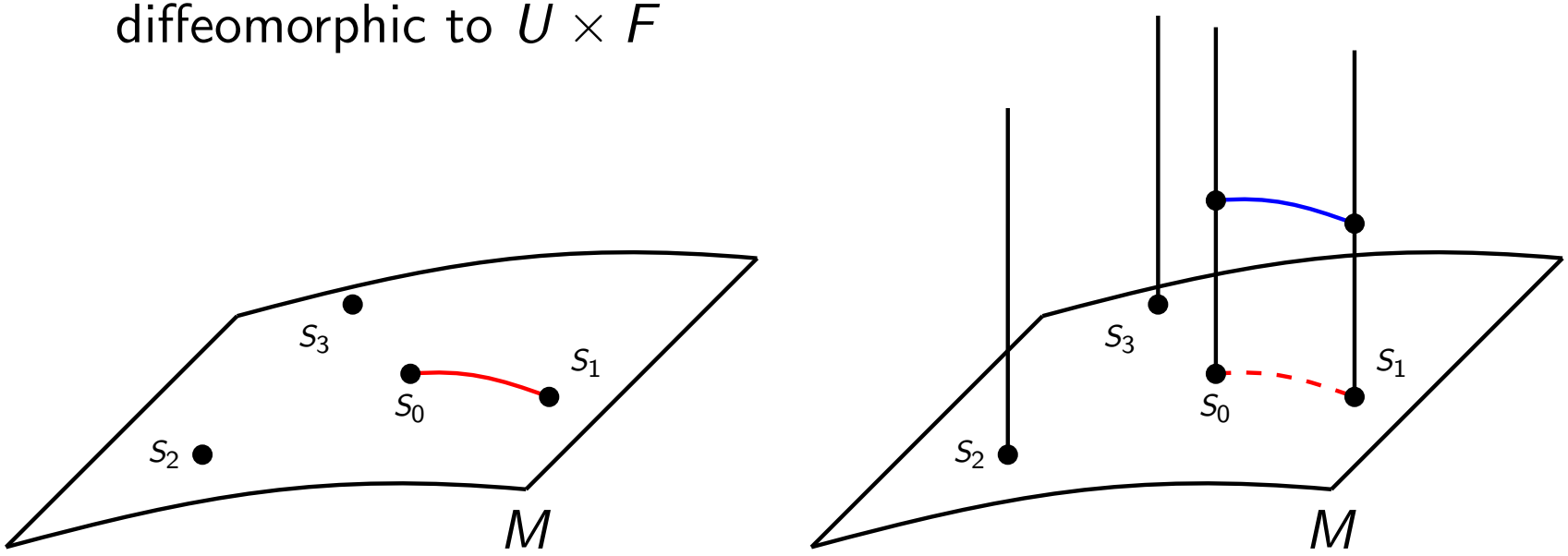
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$



Horizontal Random Walk on a *Fibre Bundle*

Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

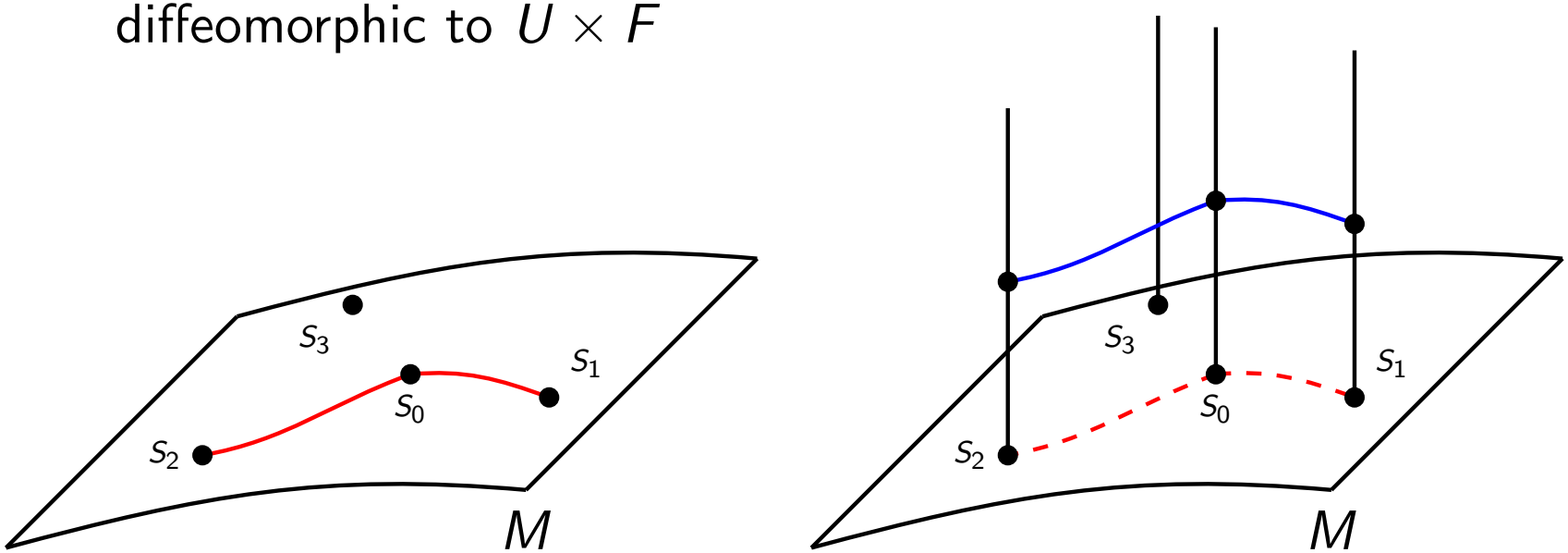
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$



Horizontal Random Walk on a *Fibre Bundle*

Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

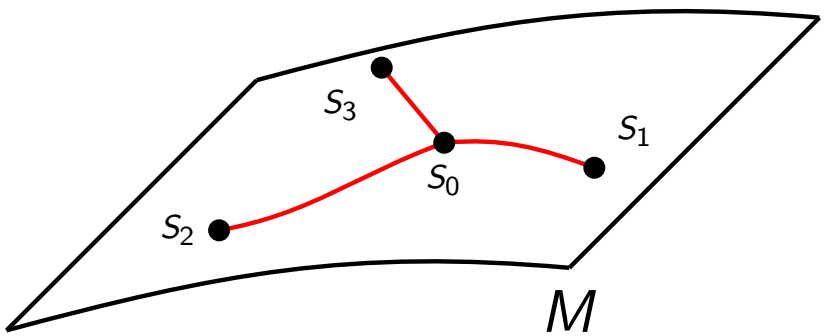
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$

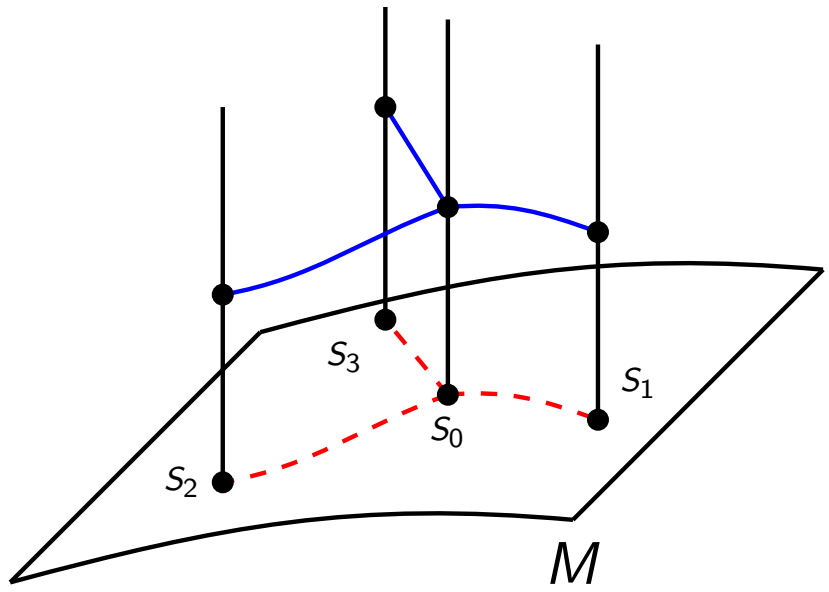


Horizontal Random Walk on a *Fibre Bundle*

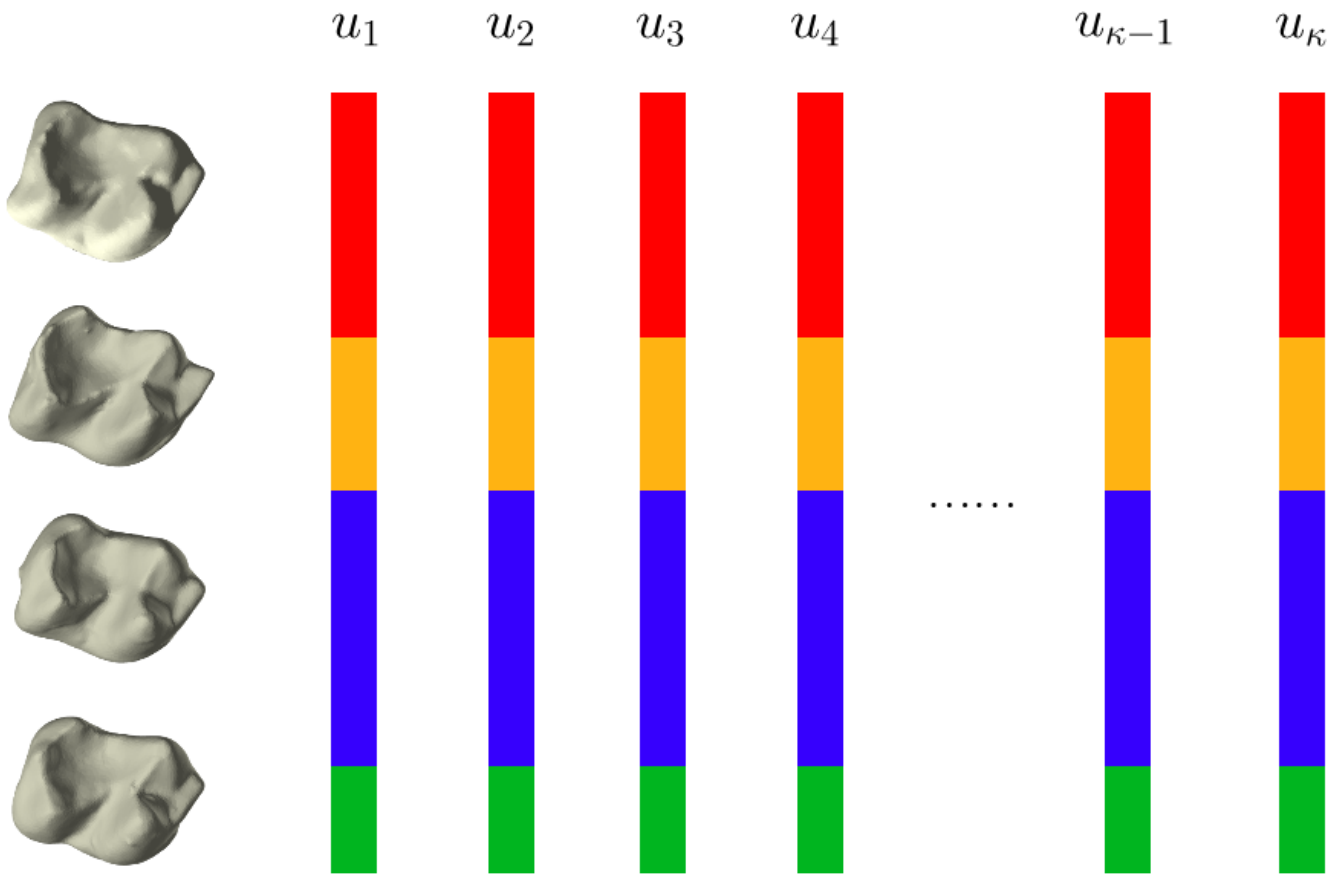
Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$

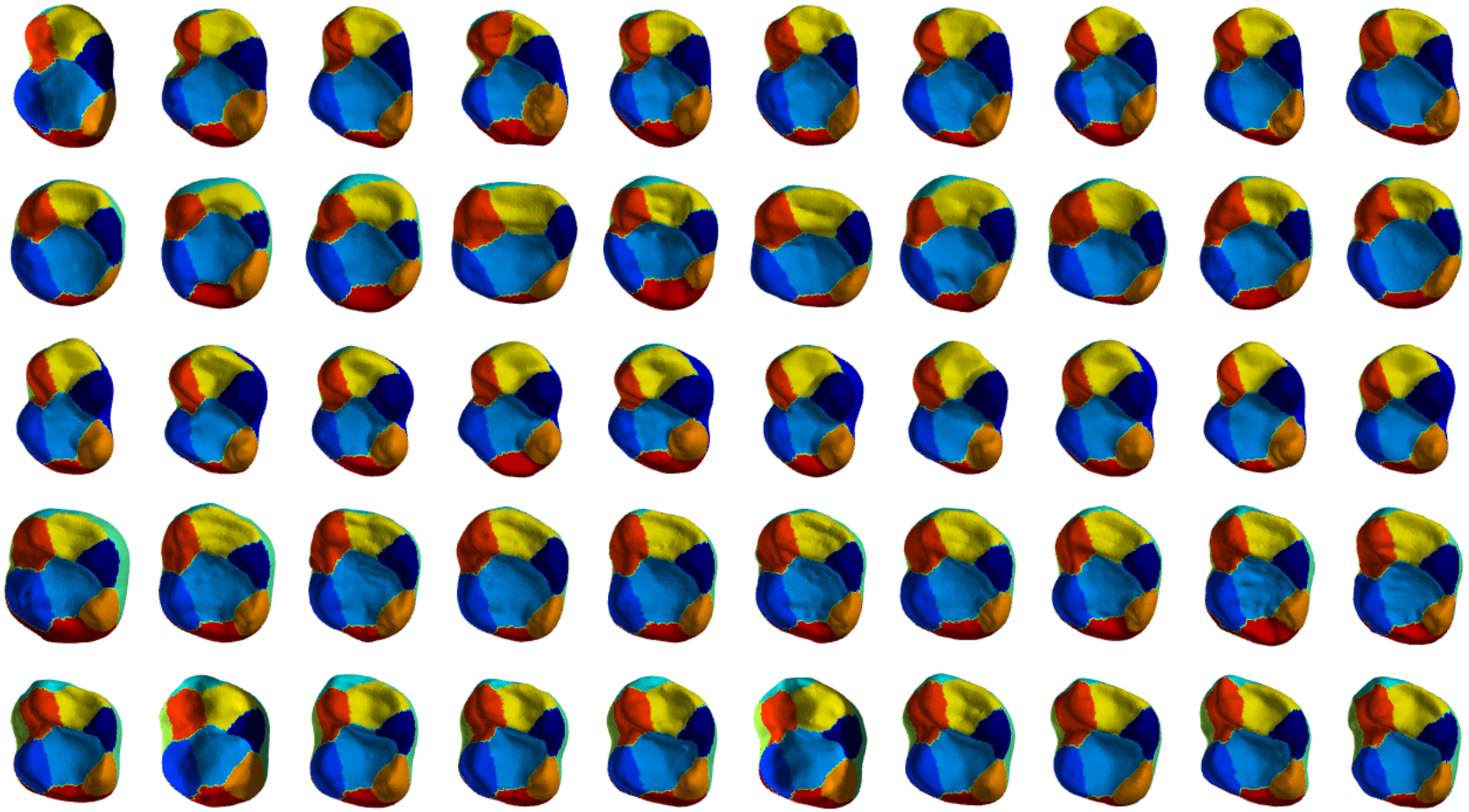




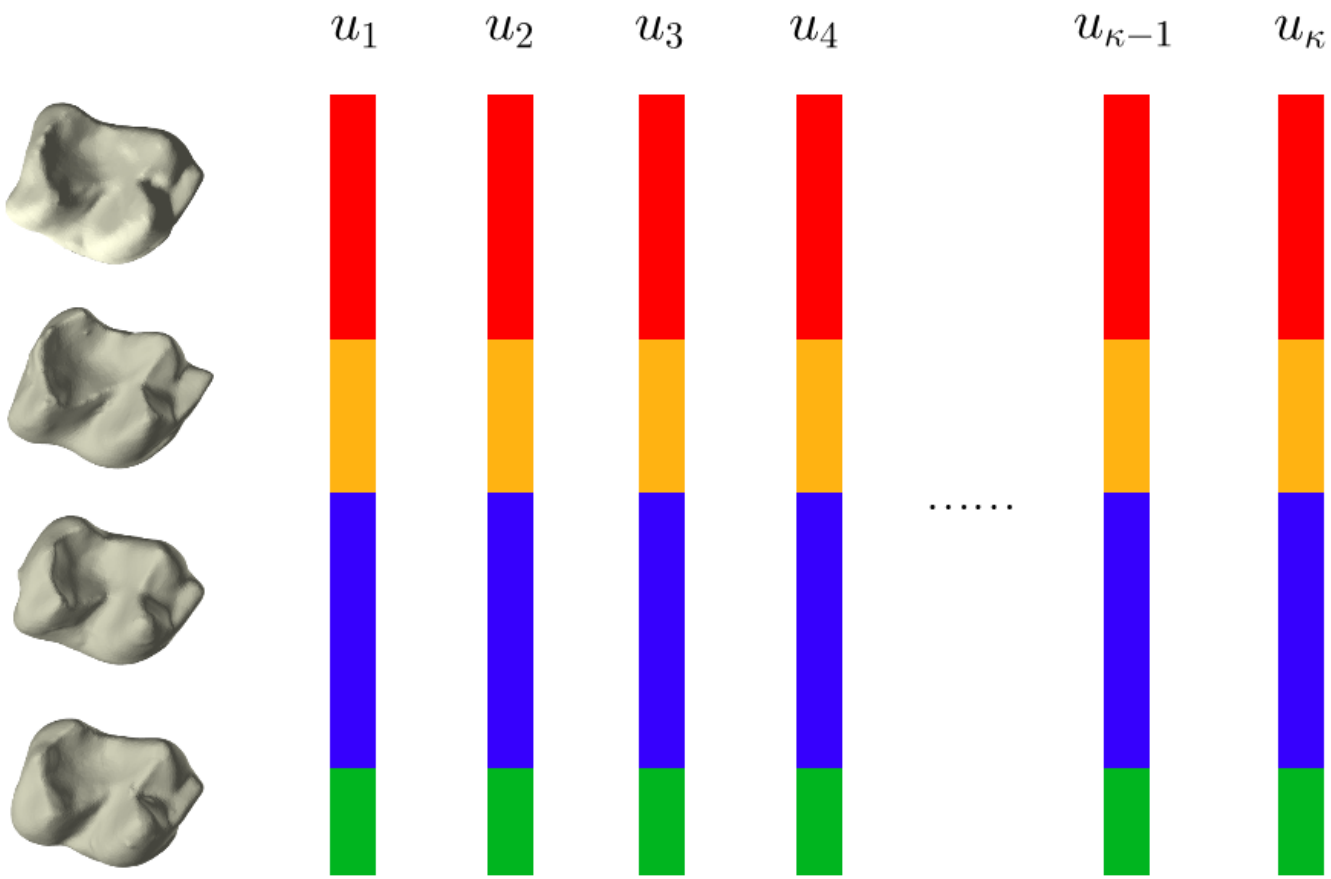
Horizontal Diffusion Maps: Embedding the Entire Bundle



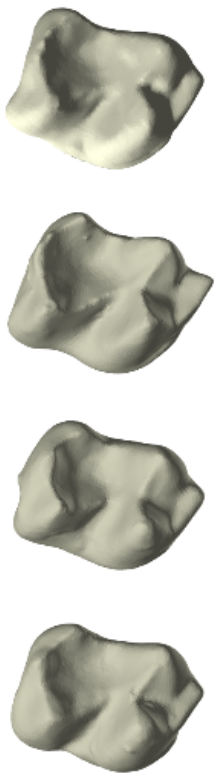
Automatic Landmarking — Interpretability

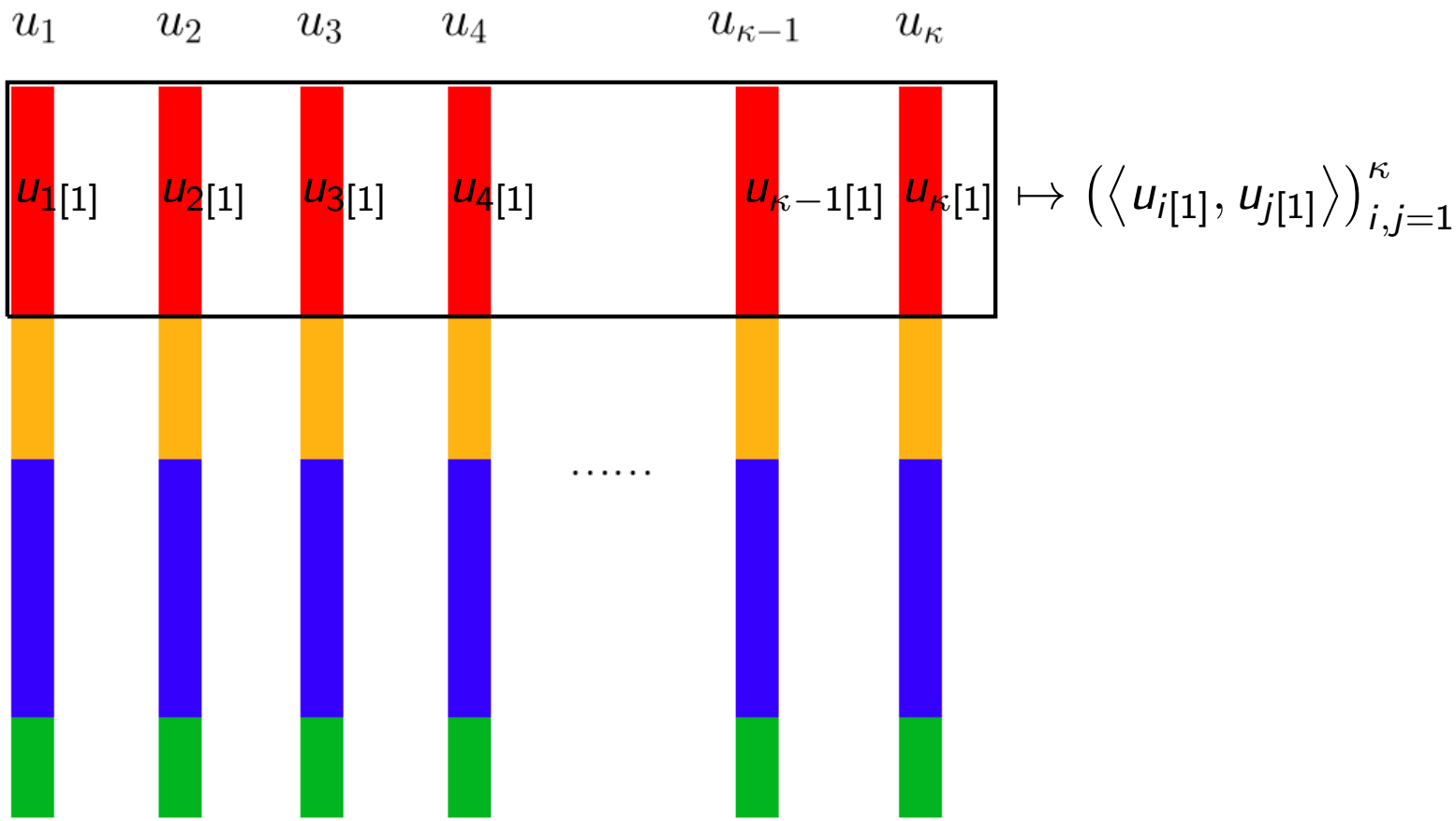


Horizontal Diffusion Maps: Embedding the Base Manifold

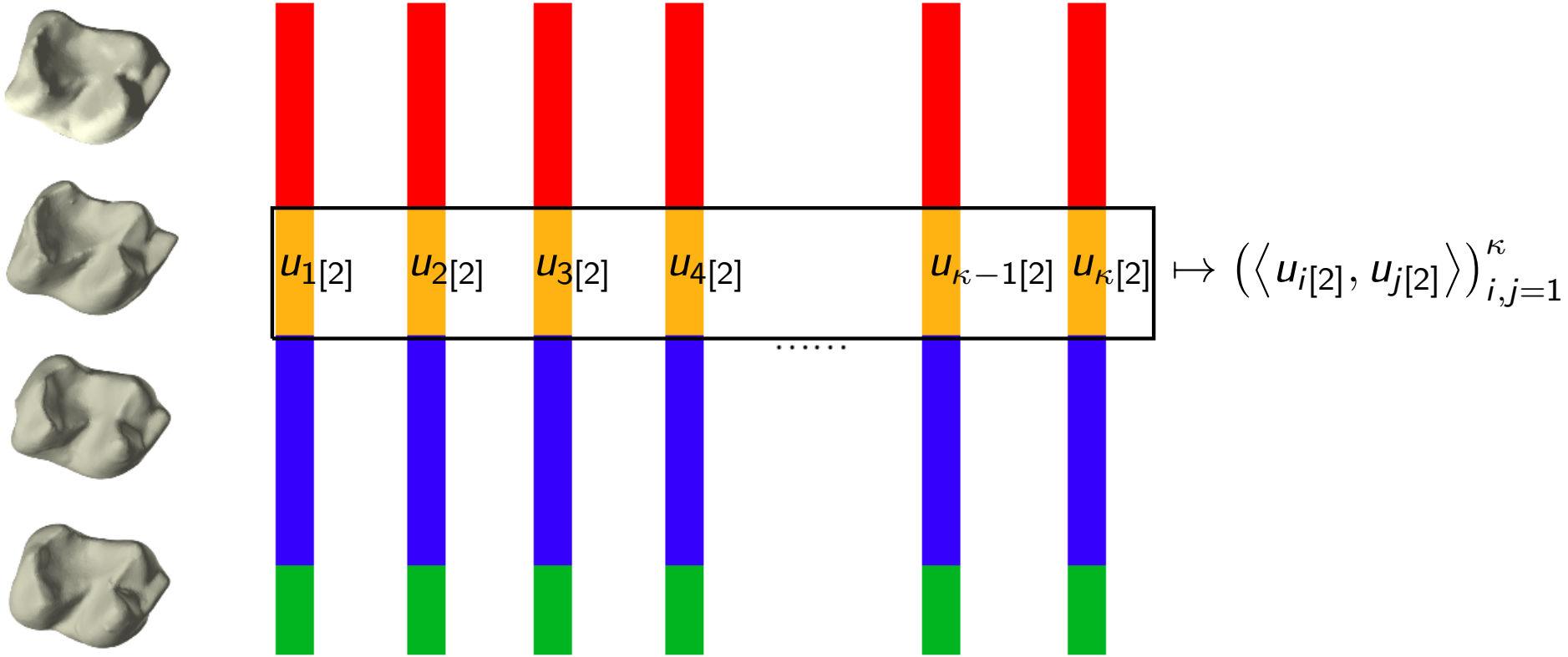


Horizontal Diffusion Maps: Embedding the Base Manifold

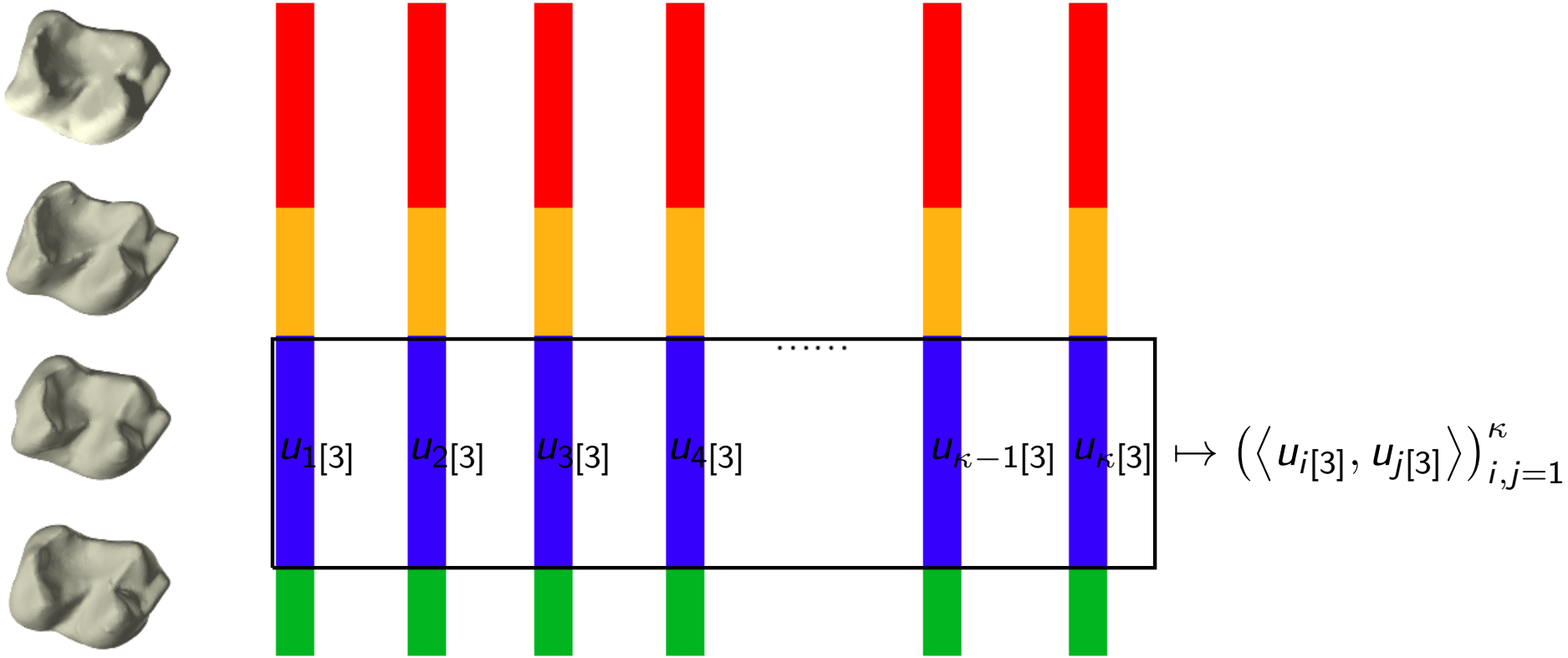




Horizontal Diffusion Maps: Embedding the Base Manifold



Horizontal Diffusion Maps: Embedding the Base Manifold



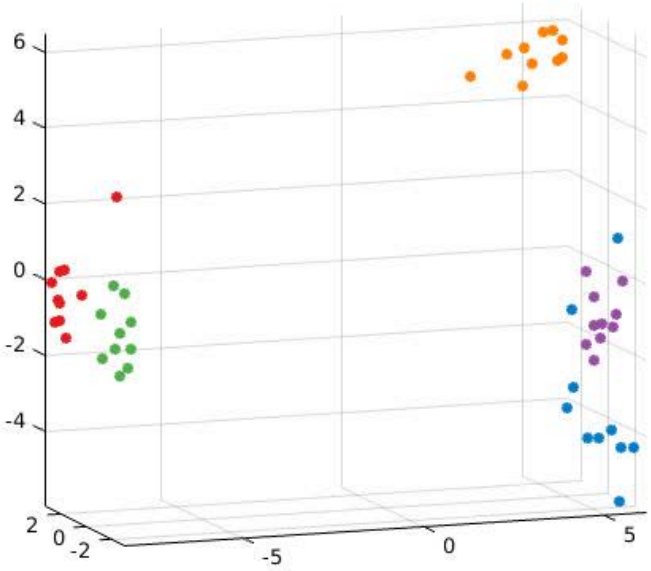
$(j, p) \longrightarrow (u_k(j,p))_{k=1}^{K}$

j: surface S_j; p: point p on S_j

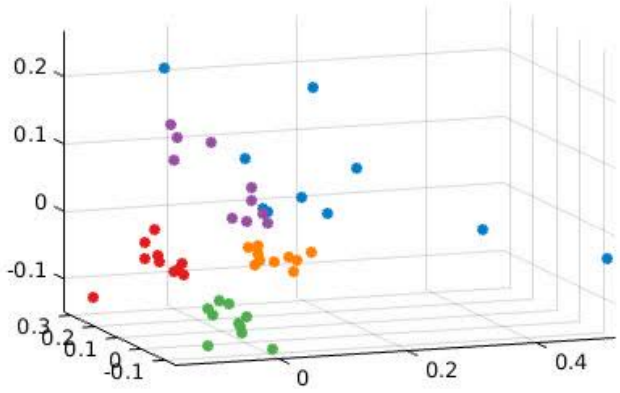
$$M_{j;k,\ell} = \sum_p w^{\lambda}_{t,k}\, u_k(j,p)\, \overline{u_\ell(j,p)}$$

$$\mathrm{dist}(S_j, S_{j'}) = \| M_j - M_{j'} \|.$$

Species Clustering

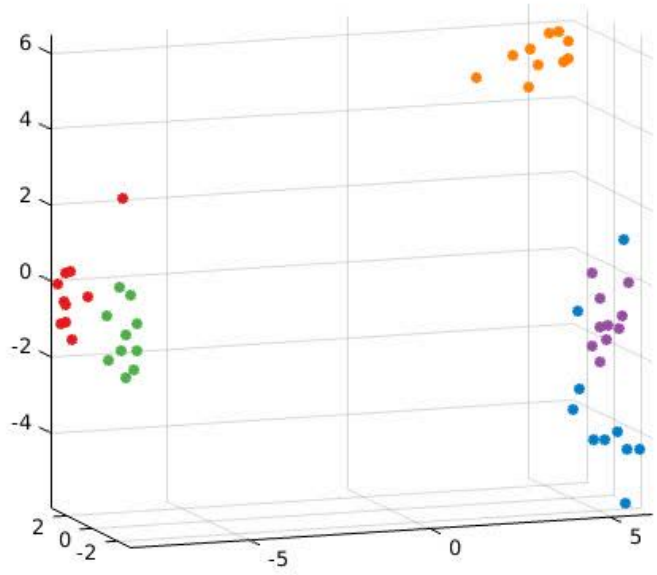


Horizontal Base Diffusion Distance (with Maps)

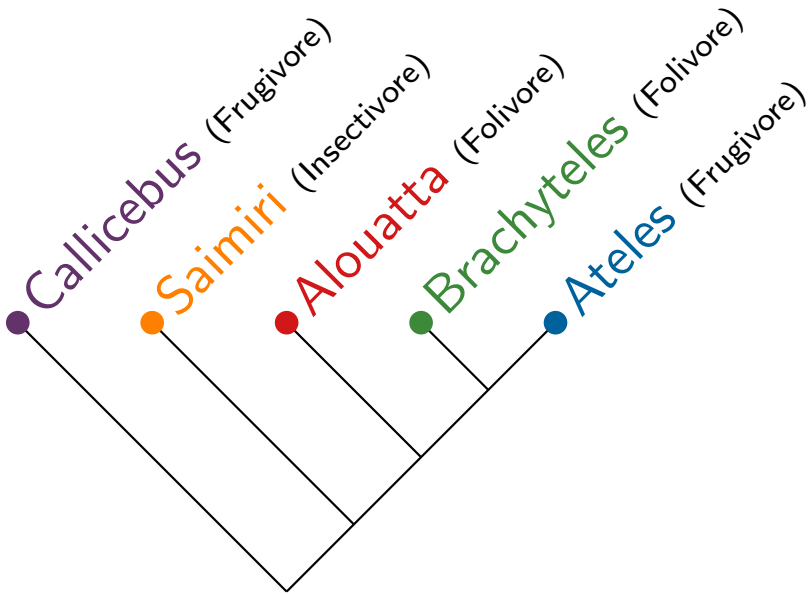


Diffusion Distance (without Maps)

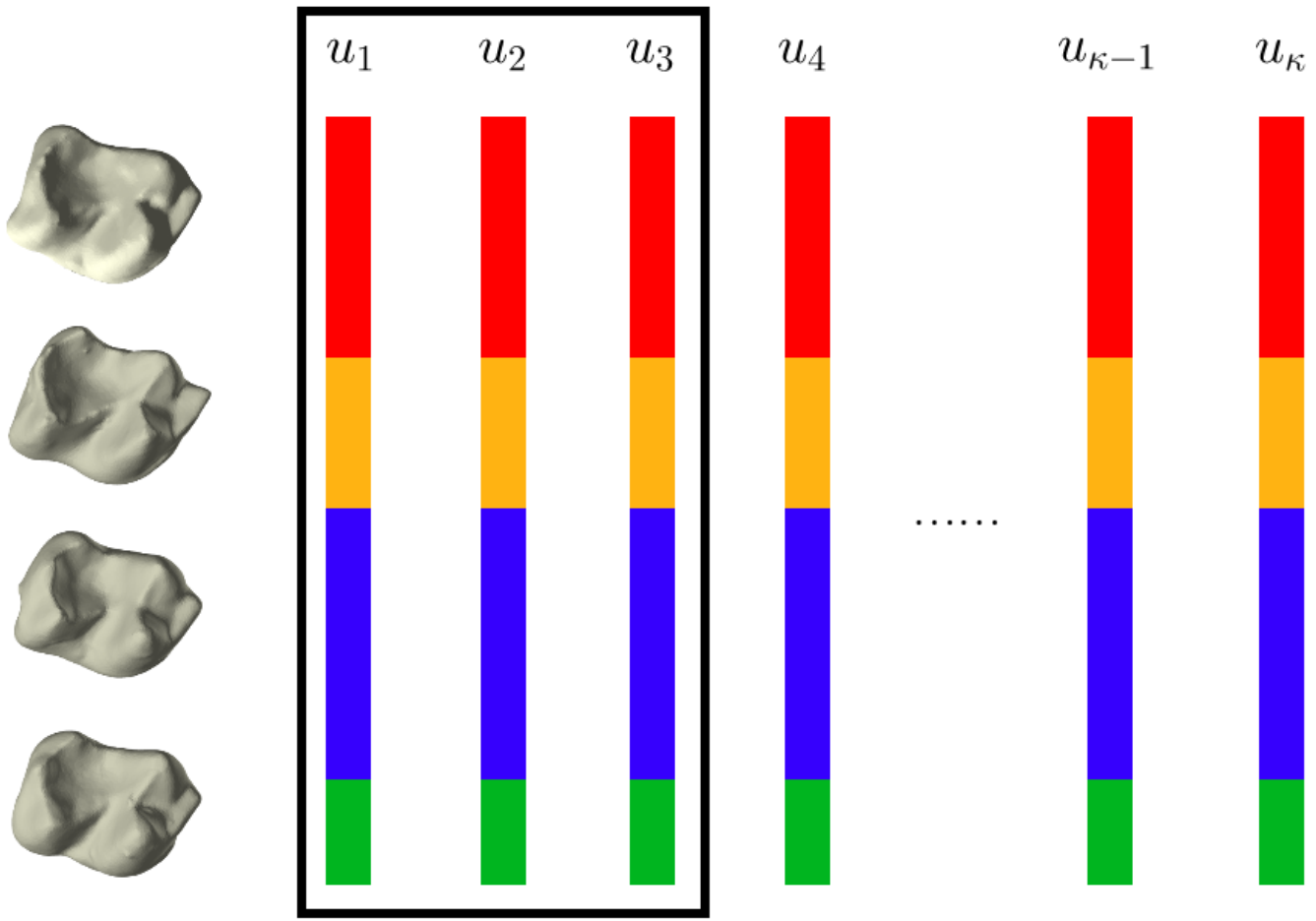
Species Clustering



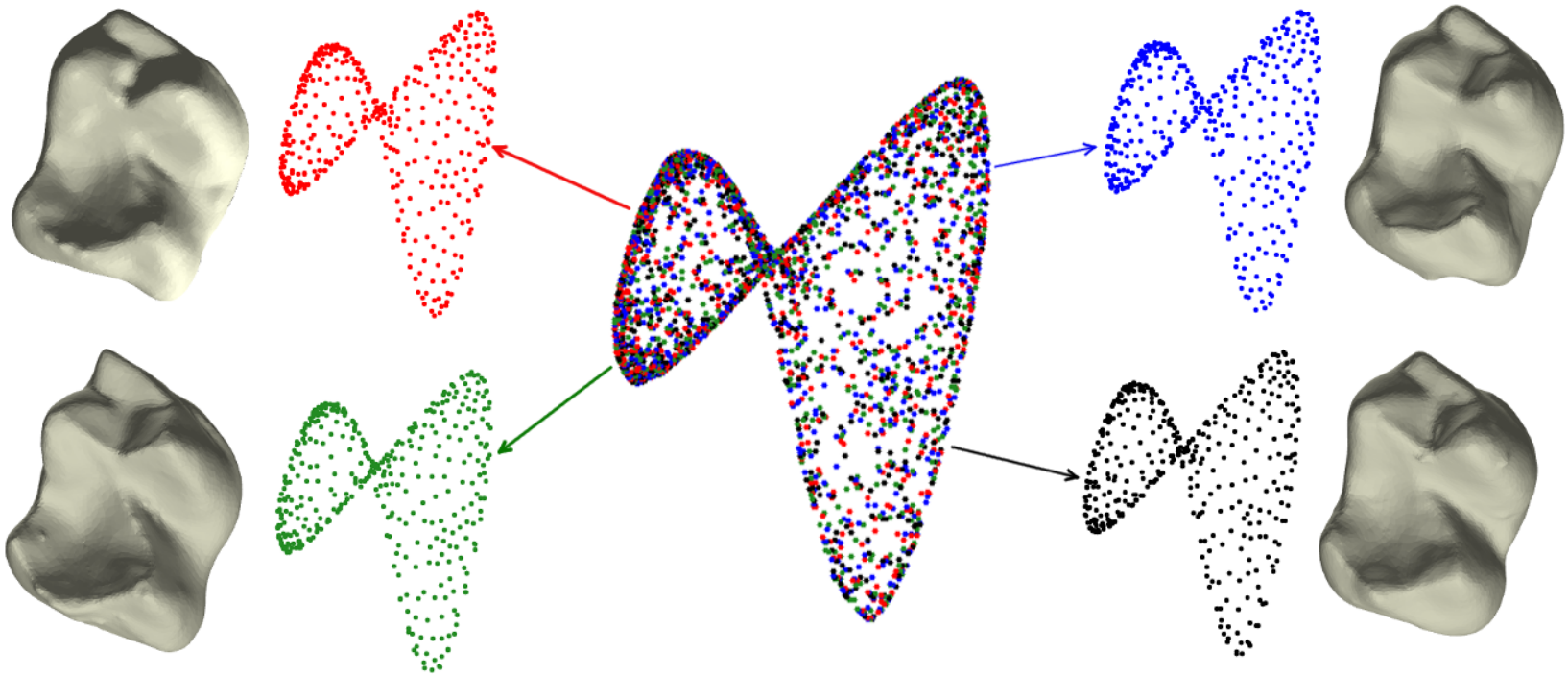
Horizontal Base Diffusion Distance (with Maps)

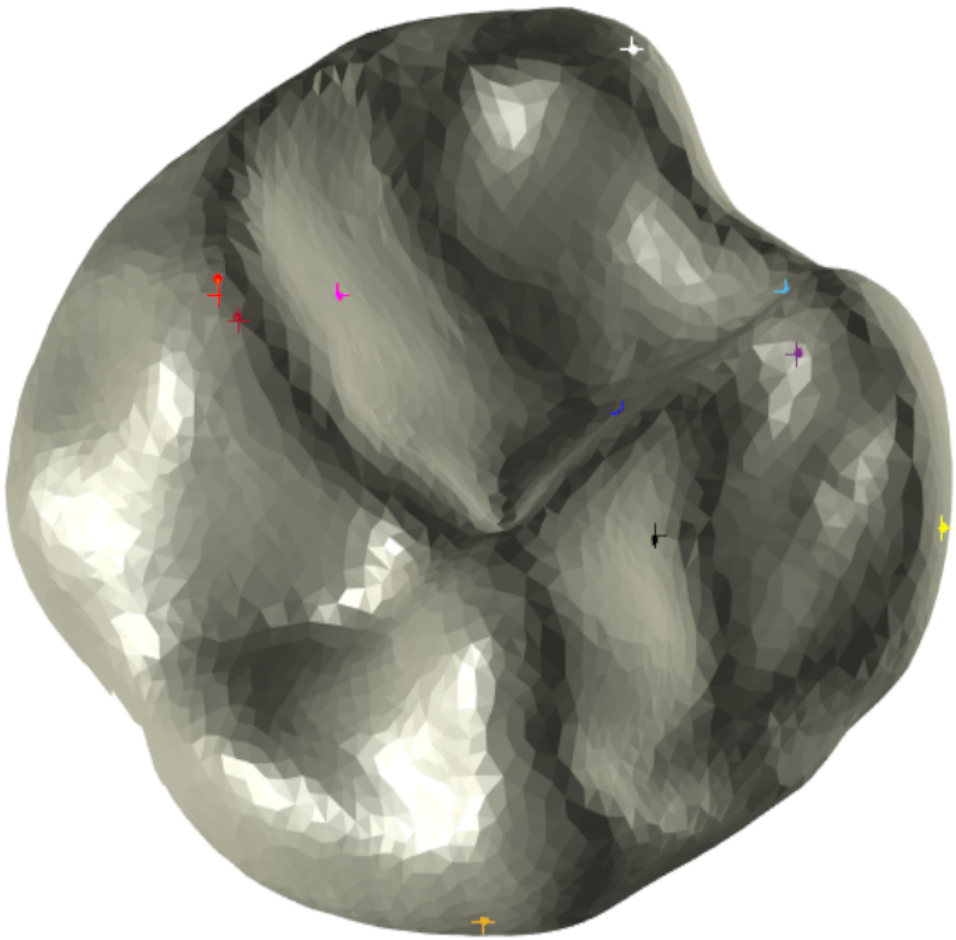


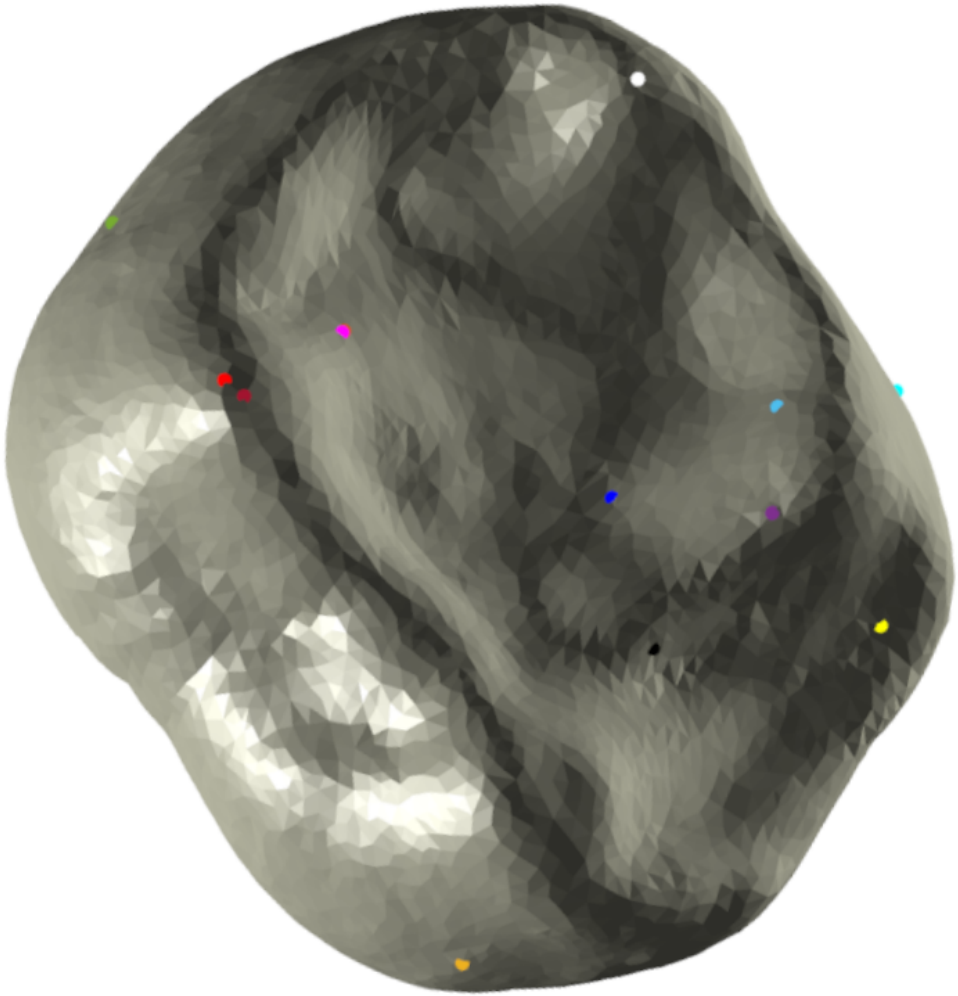
Horizontal Diffusion Maps: Embedding the Entire Bundle

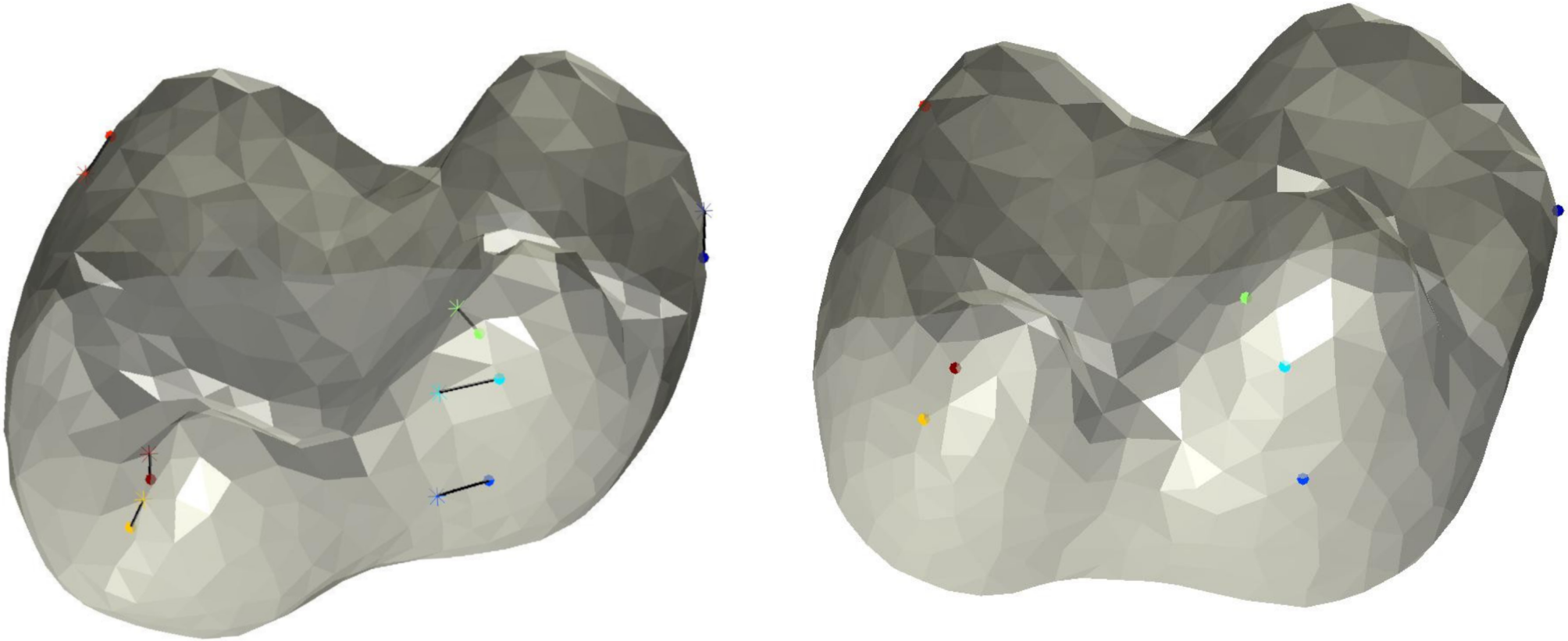


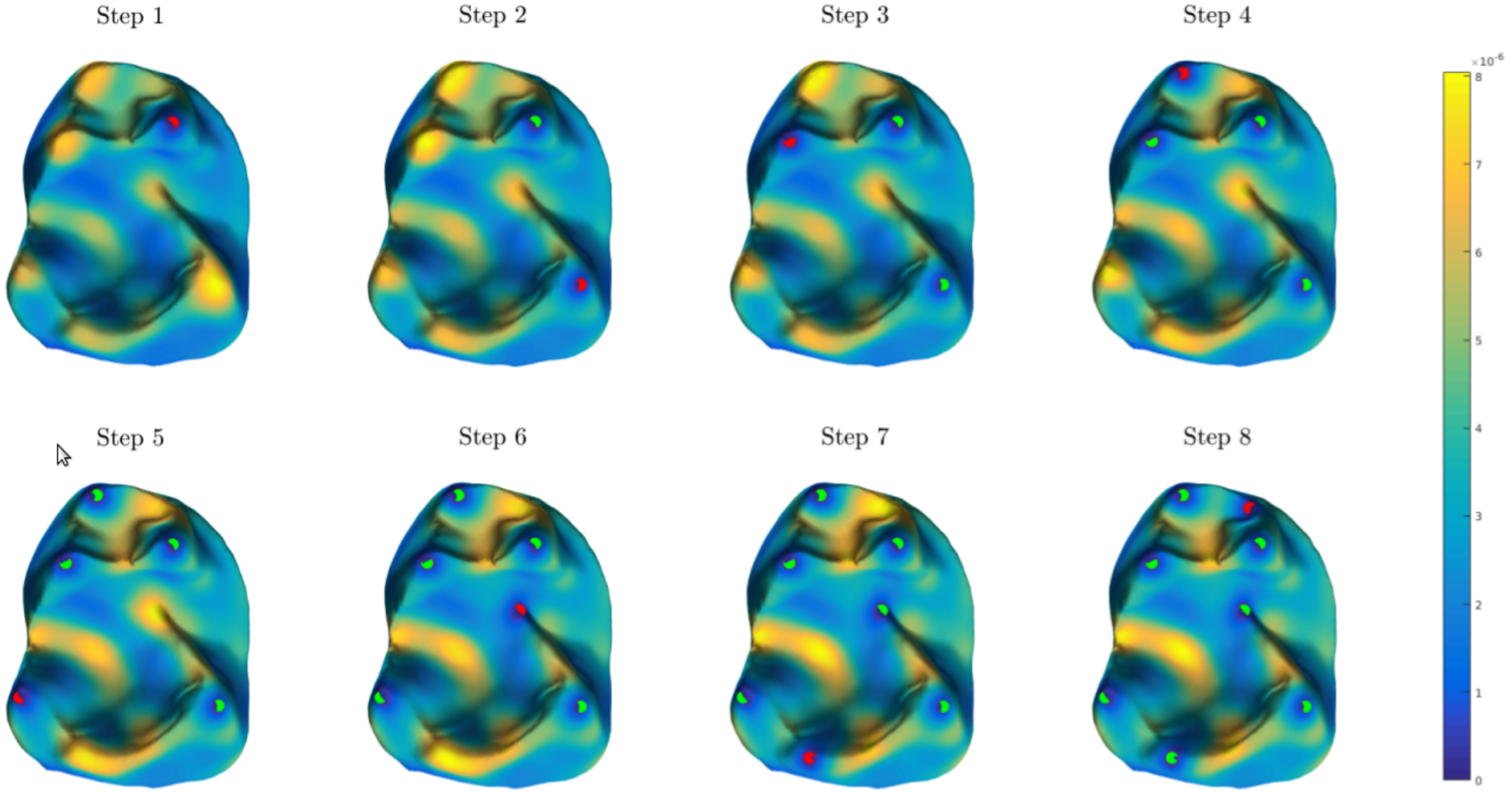
Horizontal Diffusion Maps





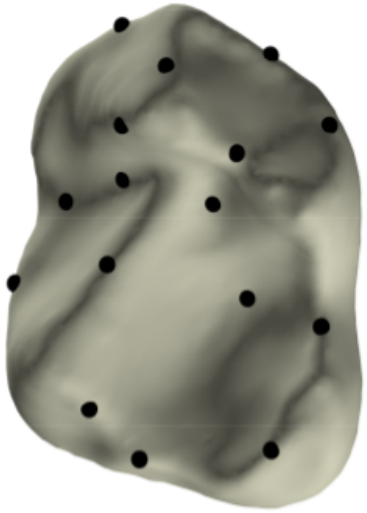





Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Step 8
×10^-6
8
7
6
5
4
3
2
1
0

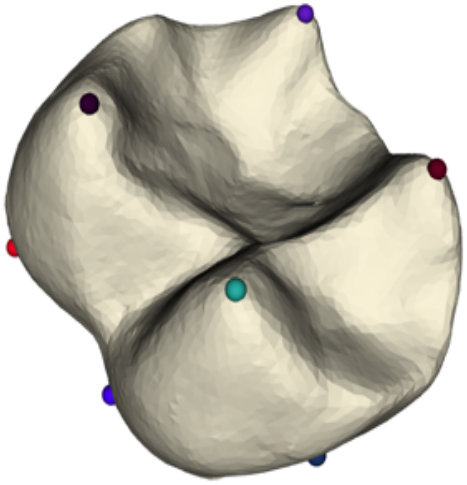
Gaussian Process Landmarking

Local Weight Maximum

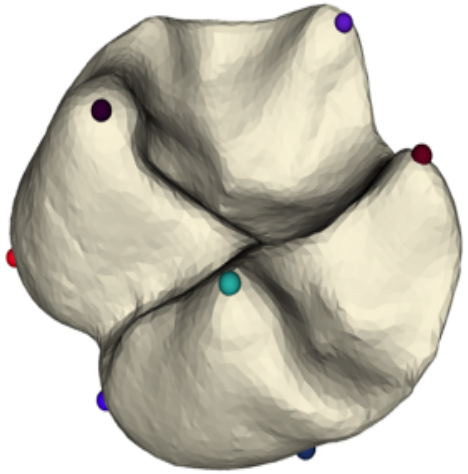


Geodesic Farthest Point Sampling

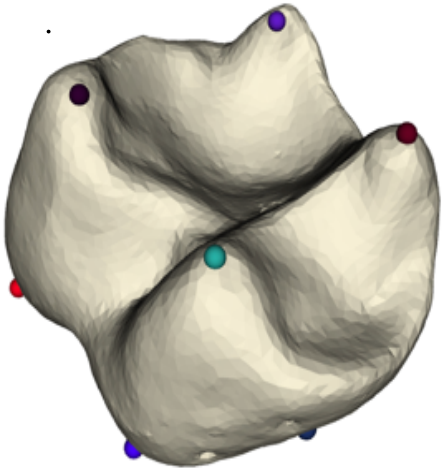
USNM-281751_M759



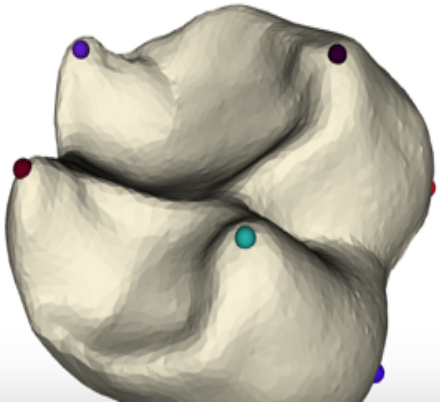
USNM-281758_M761



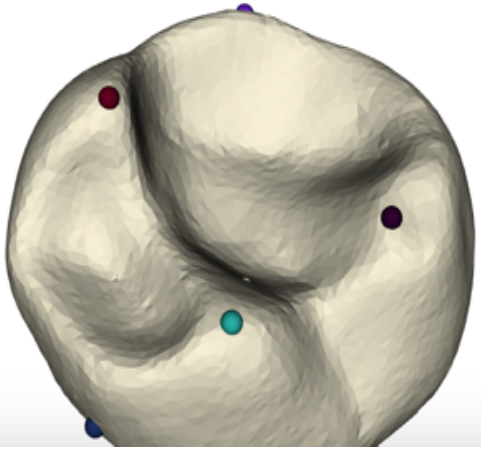
USNM-284782_M1081



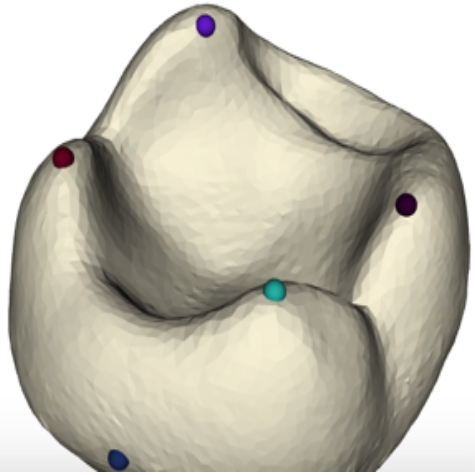
USNM-290601_M1083



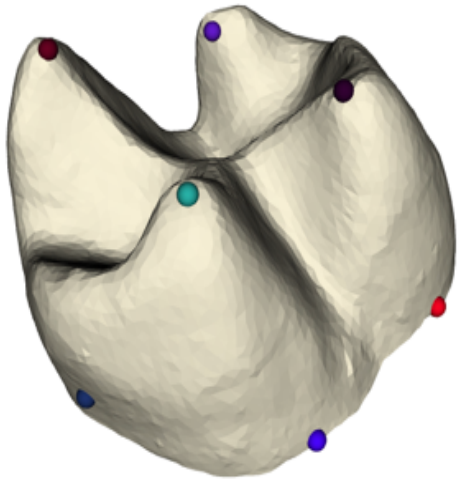
AMNH-
M-67102_M1099



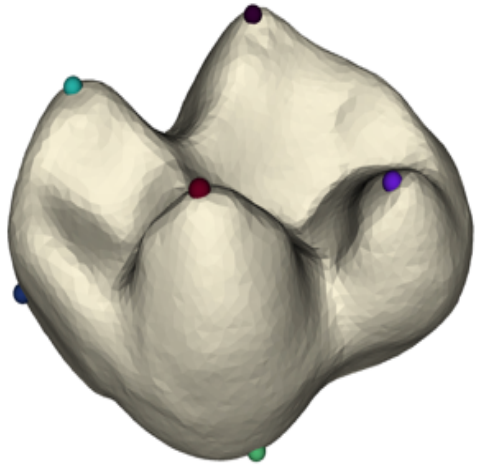
AMNH-
M-71787_M784



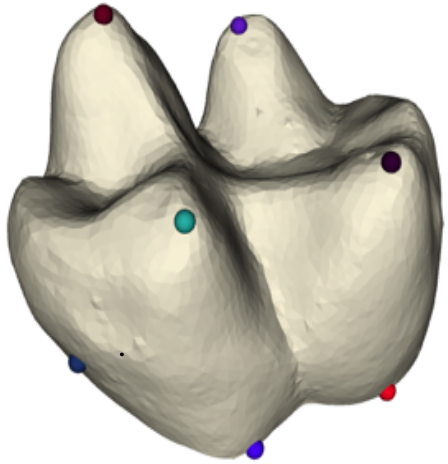
MN-Rio-106_M1214



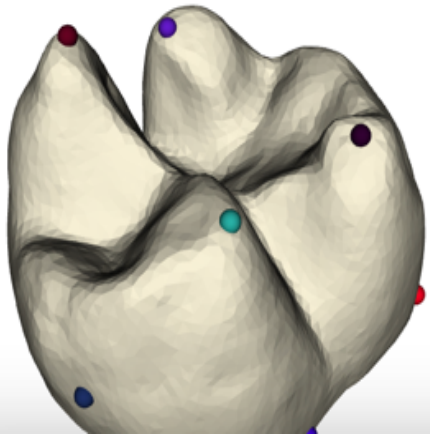
MN-Rio-526_M1205



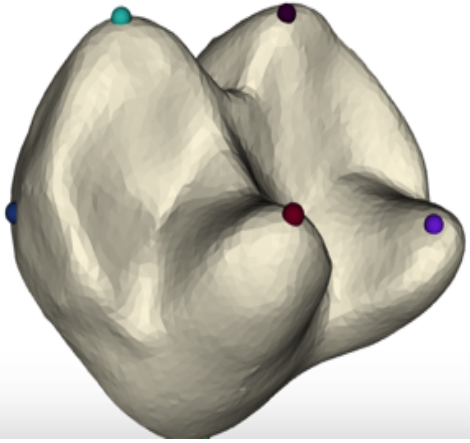
MN-Rio-2718_M1198



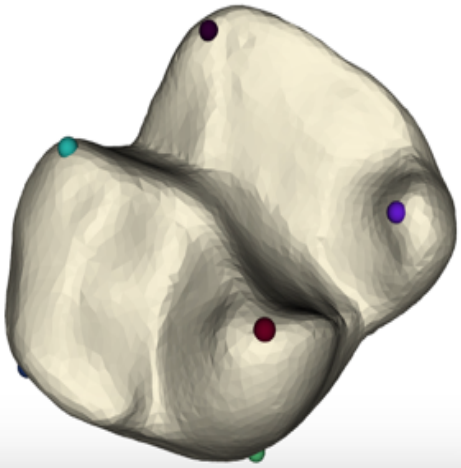
MN-Rio-6699_M1211



MN-
Rio-7724_M1223



MN-
Rio-8513_M1220



Other generalizations in progress:

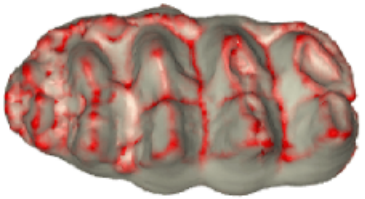
use a local geometry-adapted version of Procrustes distance.

DNE

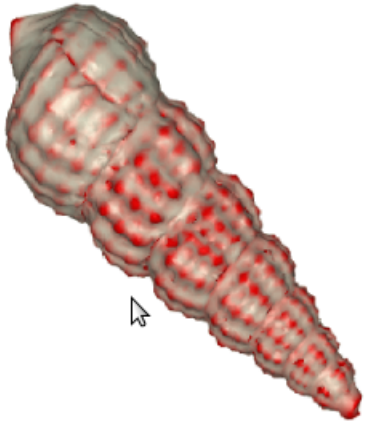
(AriaDNE)
Shan Shan

Oryctolagus
0.027

Mammut
0.040



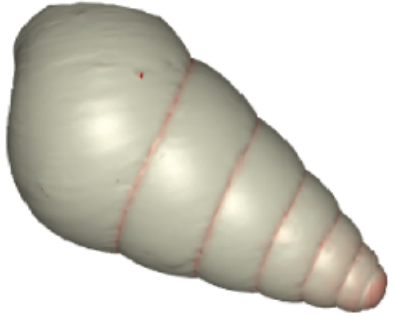
Lirobittium rugatum
0.058



Hemicentetes
0.014

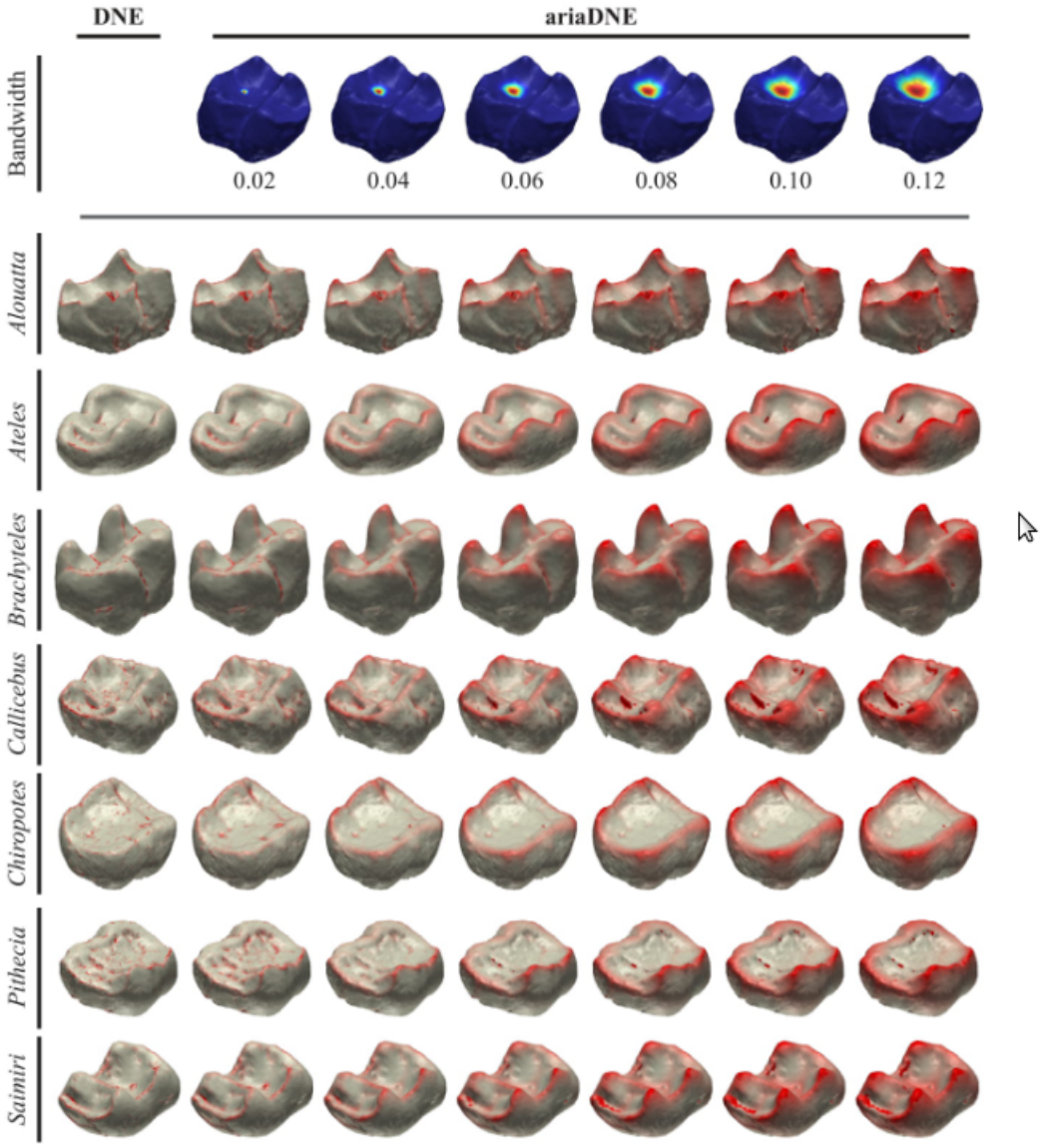
Mammuthus
0.018

Tornatellaria adelinae
0.009



DNE

(AriaDNE)
Shan Shan



Other generalizations in progress:

use a local geometry-adapted version of Procrustes distance.

varying ratios of diffusion times on fiber and base: first optimize registration by emphasizing fiber diffusion; then increasingly emphasize base diffusion to obtain base geometry insights

Happy Birthday, Albert!