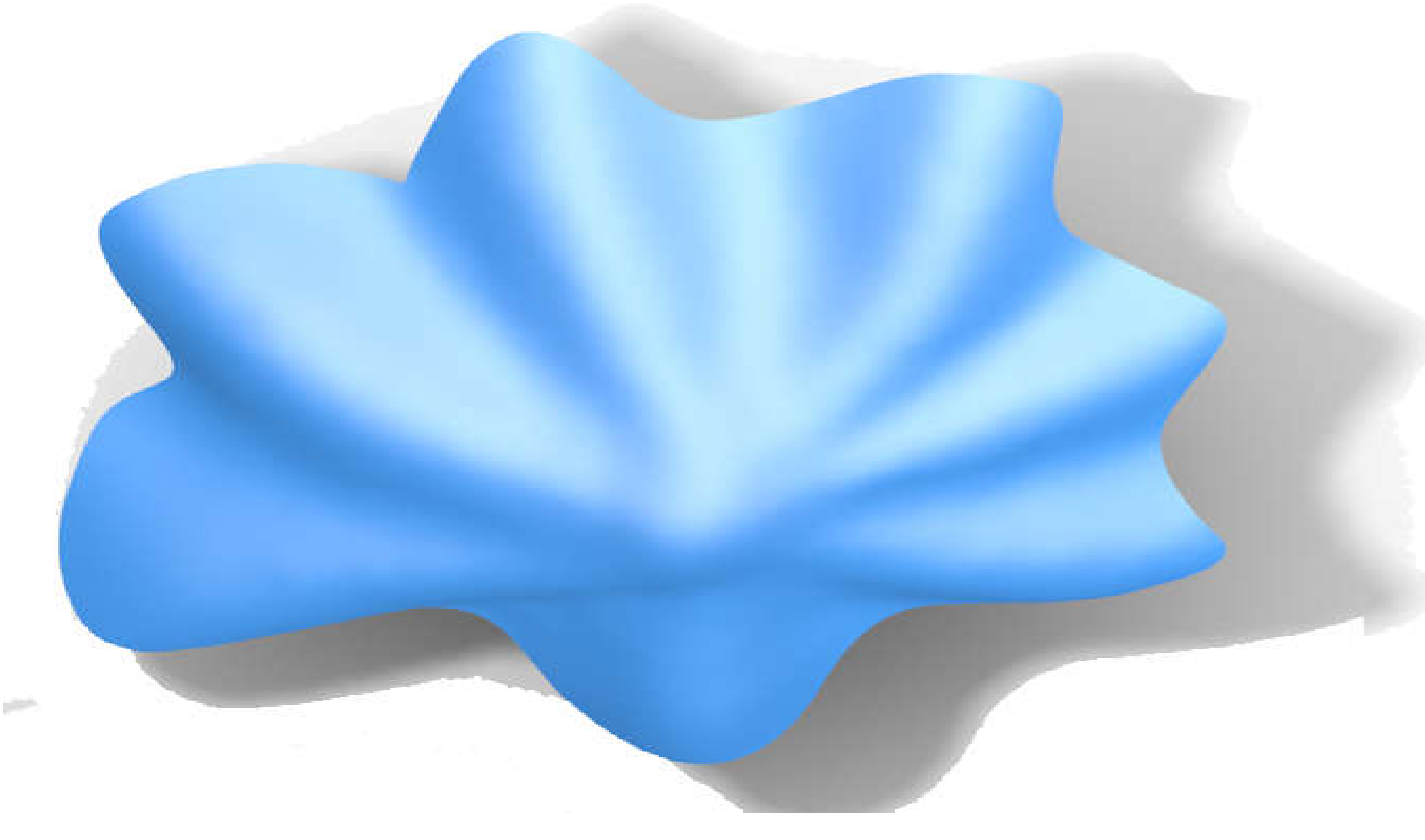
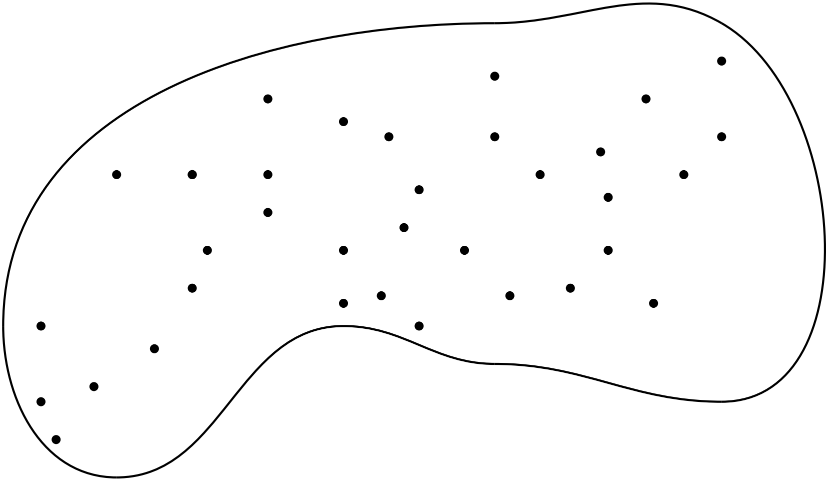
Discovering low-dimensional manifolds in high-dimensional data sets

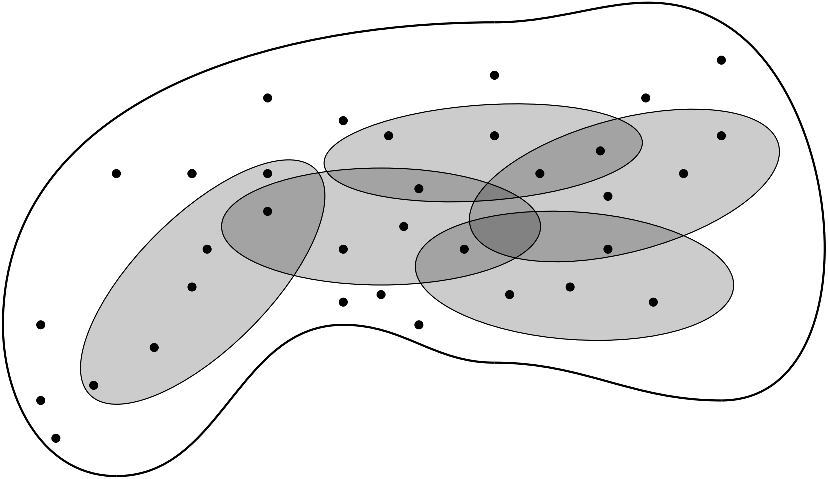


Diffusion Maps: "Knit together" local geometry to get "better" distances



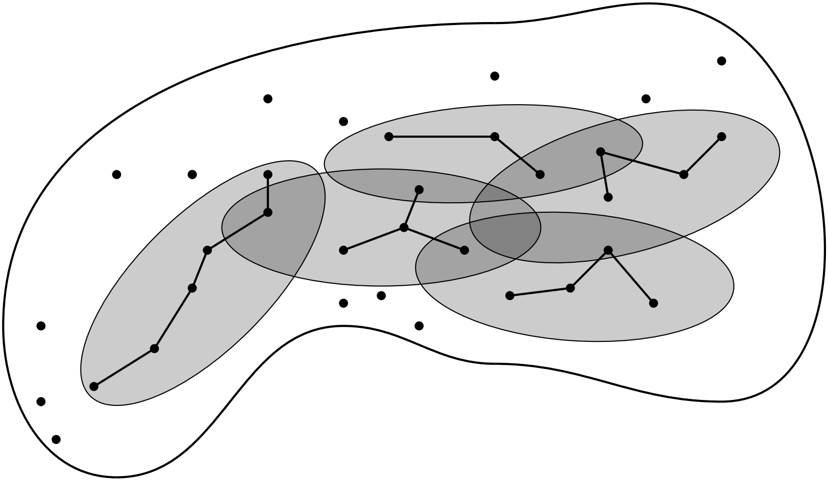
Small distances are much more reliable!

Diffusion Maps: “knitting together” local geometry

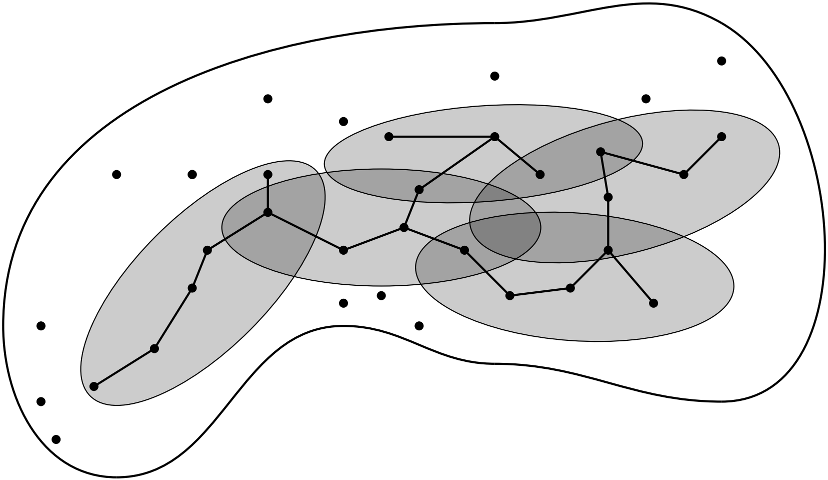


A collection of small tangent patches does give a good first approximation of a surface

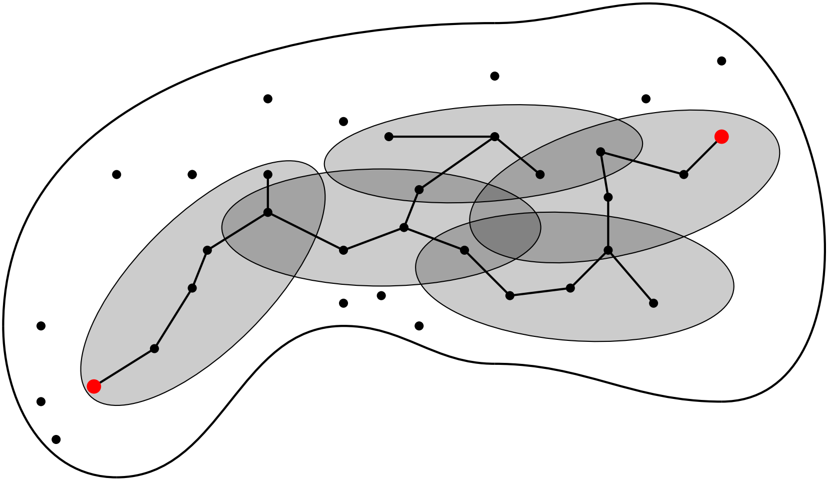
Diffusion Maps: “knitting together” local geometry



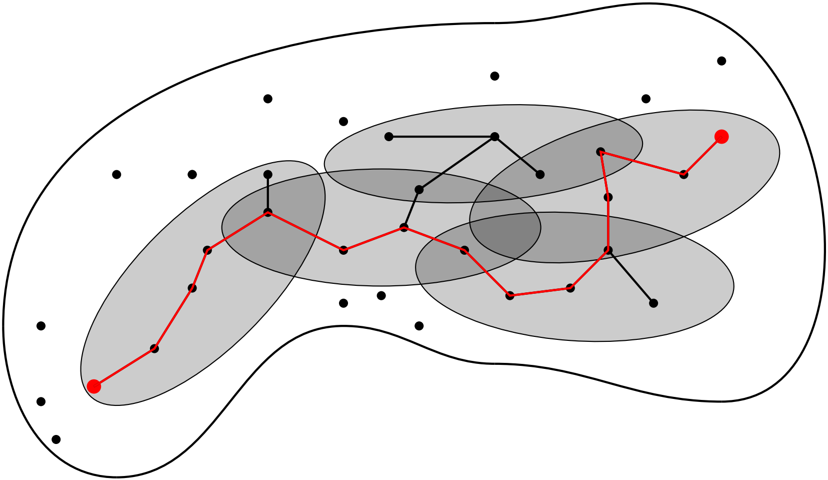
Diffusion Maps: "knitting together" local geometry

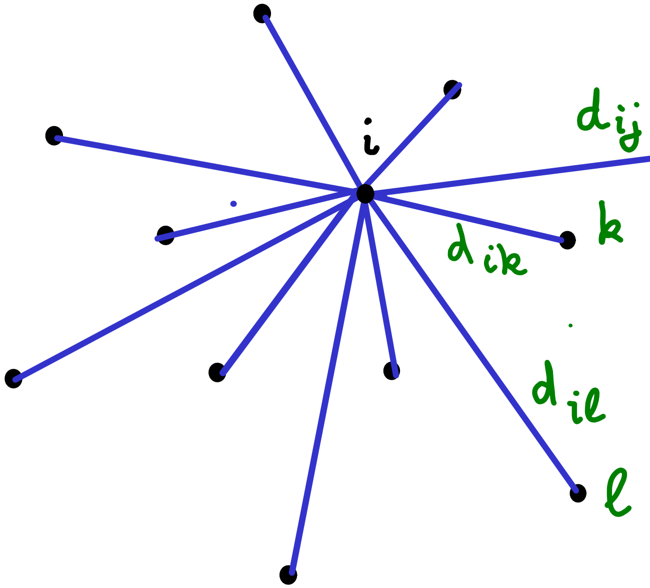


Diffusion Maps: “knitting together” local geometry



Diffusion Maps: "knitting together" local geometry





$$e^{-d_{ij}^2/2\tau} = W_{i,j;\tau}$$

$$D_{i,i;\tau} = \sum_{j \neq i} W_{i,j;\tau}$$

$D_\tau^{-1} W_\tau$: defines a random walk on graph.

How pick τ? want: approximation to diffusion on manifold

"true" diffusion: semi-group property

$$e^{-tL} e^{-sL} = e^{-(t+s)L}$$

Shan Shan

$\Rightarrow$ pick τ so that $(D_\tau^{-1} W_\tau)^2 \simeq D_{2\tau}^{-1} W_{2\tau}$

spectral decomposition of $D_\tau^{-1} W_\tau$

$\Rightarrow$ eigenvectors ψ_ℓ with

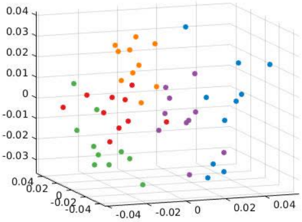
eigenvalues $\lambda_{\ell;\tau}$

$$\left(e^{-\gamma\tau L}\right)_{ij} \simeq \sum_{\ell=1}^{N} (\lambda_{\ell;\tau})^{\gamma} \psi_\ell(i)\, \psi_\ell(j)$$

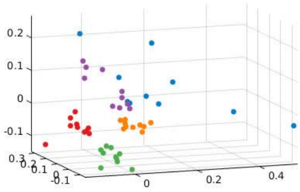
each data point j is mapped to the feature vector $\left((\lambda_{\ell,\tau})^{\gamma/2} \psi_\ell(j)\right)_{\ell=1}^{L}$

$$D^2_{ij;t} := \sum_{\ell=1}^{N} (\lambda_{\ell;t})^{t} \left|\psi_\ell(i) - \psi_\ell(j)\right|^2$$

MDS for cPD & DD



cPD

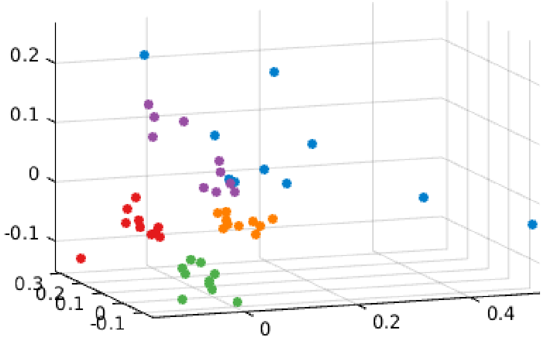


DD

Diffusion Distance (DD)

Fix $1 \leq m \leq N$, $t \geq 0$,

$$D_m^t\left(S_i, S_j\right) = \left(\sum_{k=1}^{m} \lambda_k^t \left(u_k\left(i\right) - u_k\left(j\right)\right)^2\right)^{\frac{1}{2}}$$



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

Doug Boyer

Jukka Jernvall

More Precisely: biological morphologists

↓

Study Teeth & Bones of

extant & extinct animals

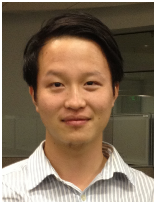
still live today   fossils

Collaborators

Rima Alaifari
ETH Zürich

Doug Boyer
Duke

Ingrid Daubechies
Duke



Tingran Gao
Duke

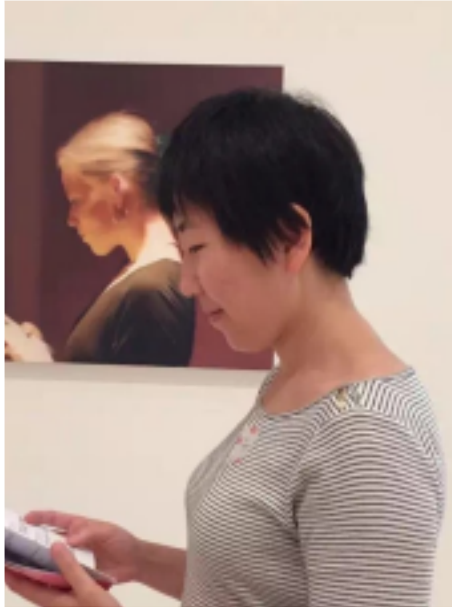
Yaron Lipman
Weizmann

Roi Poranne
ETH Zürich

Jesús Puente
J.P. Morgan

Robert Ravier
Duke

Shahar Kovalsky



Shan Shan

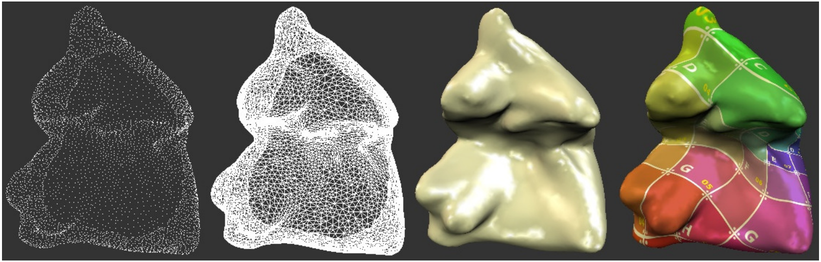
Nadav Dym

Chen-Yun Lin

First: project on “complexity” of teeth

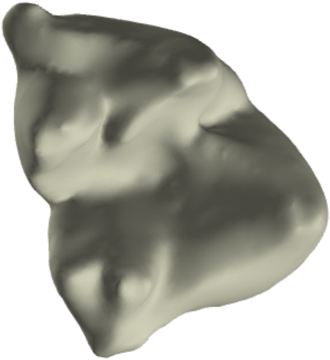
First: project on “complexity” of teeth
Then: find automatic way to compute Procrustes distances between surfaces — without landmarks

Data Acquisition



Surface reconstructed from μCT-scanned voxel data

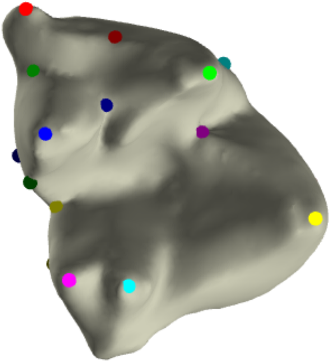
Geometric Morphometrics



second mandibular molar of a Philippine flying lemur

- Manually put k landmarks

Geometric Morphometrics

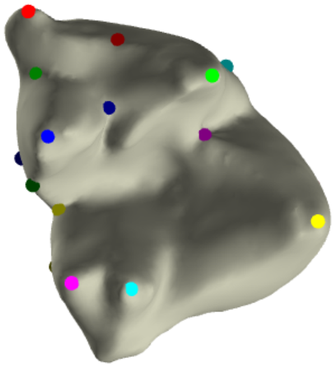


second mandibular molar of a Philippine flying lemur

- Manually put k landmarks

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

Geometric Morphometrics



second mandibular molar of a Philippine flying lemur

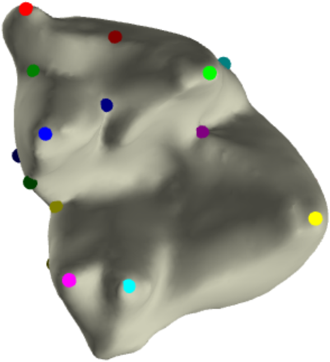
- Manually put k landmarks

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

- 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

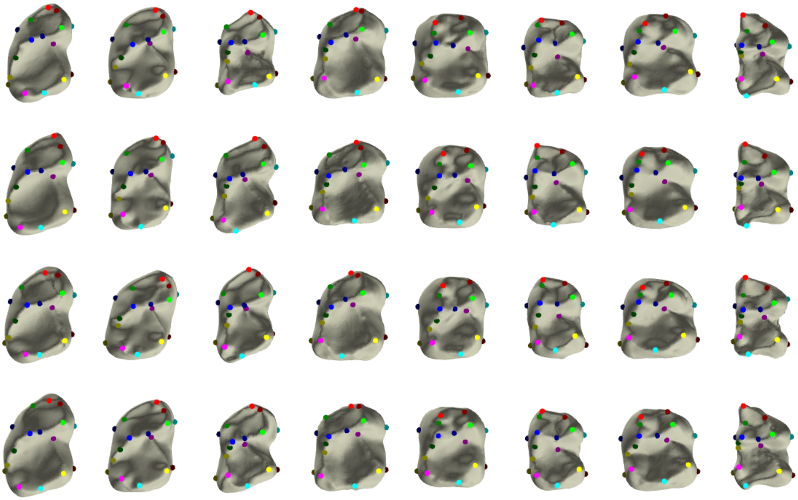
$$p_1, p_2, \cdots, p_k$$

- Use spatial coordinates of the landmarks as features

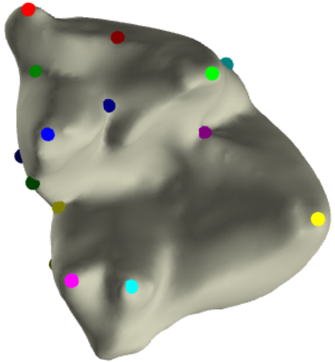
$$p_j = (x_j, y_j, z_j), \; j = 1, \cdots, k$$

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

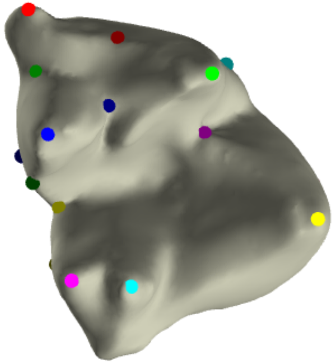
The *Shape Space* of k landmarks in $\mathbb{R}^3$



Geometric Morphometrics: Limitation of Landmarks

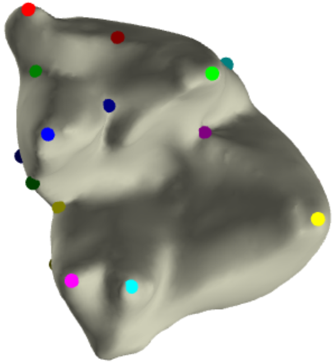


Geometric Morphometrics: Limitation of Landmarks



- Landmark Placement: tedious and time-consuming

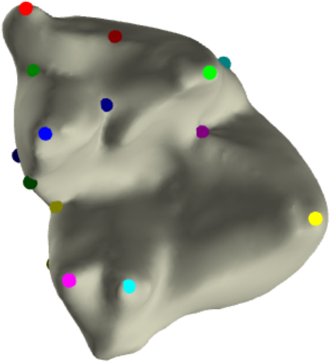
Geometric Morphometrics: Limitation of Landmarks



- Landmark Placement: tedious and time-consuming

- Fixed Number of Landmarks: lack of flexibility

Geometric Morphometrics: Limitation of Landmarks

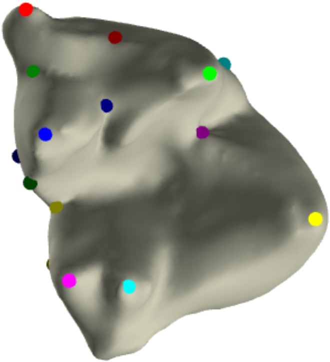


- Landmark Placement: tedious and time-consuming

- 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 $\longrightarrow$

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

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




Landmarked Teeth $\longrightarrow$

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

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

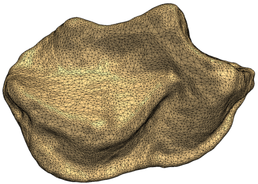
First: project on "complexity" of teeth
Then: find automatic way to compute Procrustes distances between surfaces — without landmarks

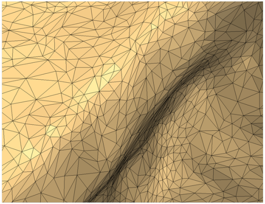
Landmarked Teeth $\longrightarrow$

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

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

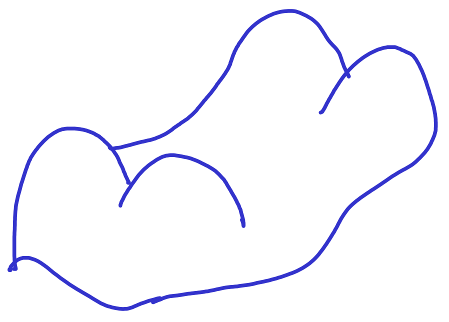
examples: finely discretized triangulated surfaces

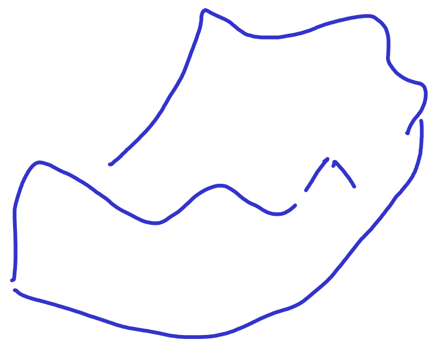


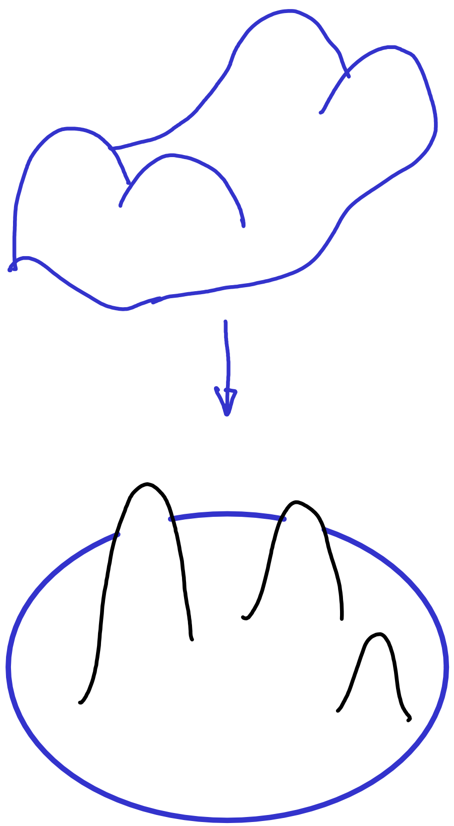


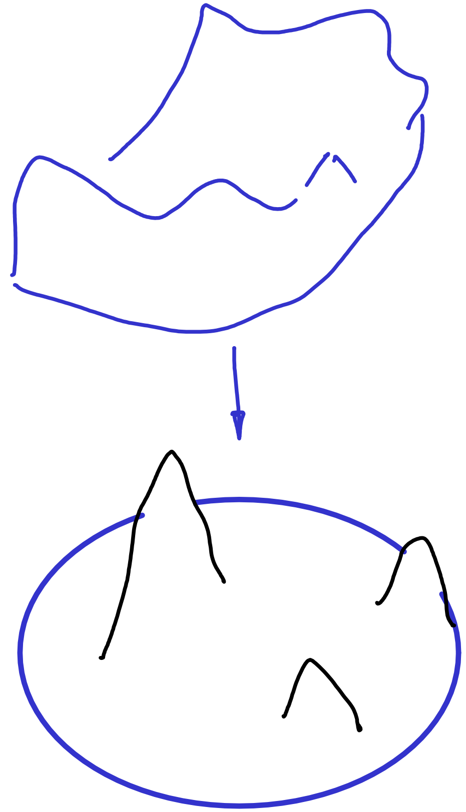
We defined 2 different distances

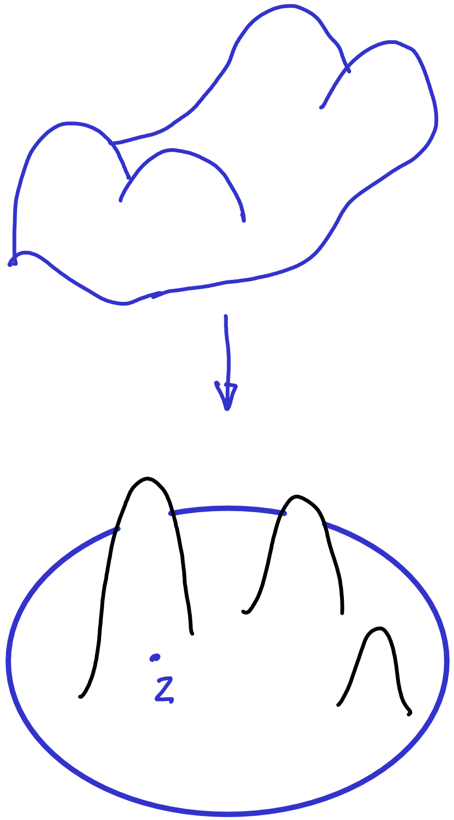
$d_{\mathrm{cWn}}(S_1, S_2)$:	conformal flattening comparison of neighborhood geometry optimal mass transport
$d_{\mathrm{cP}}(S_1, S_2)$:	continuous Procrustes distance

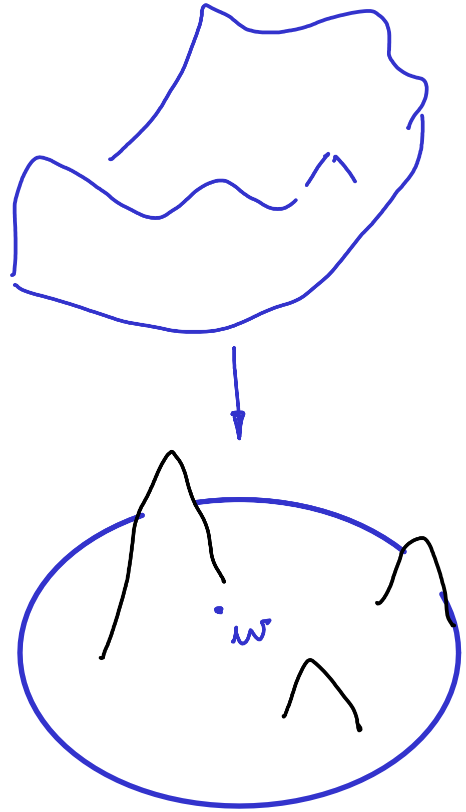


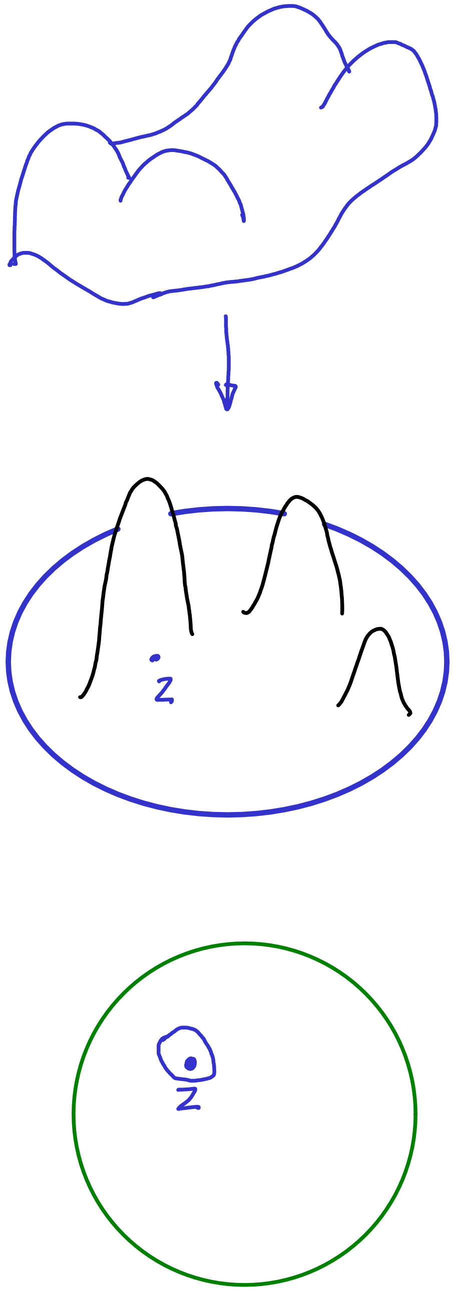




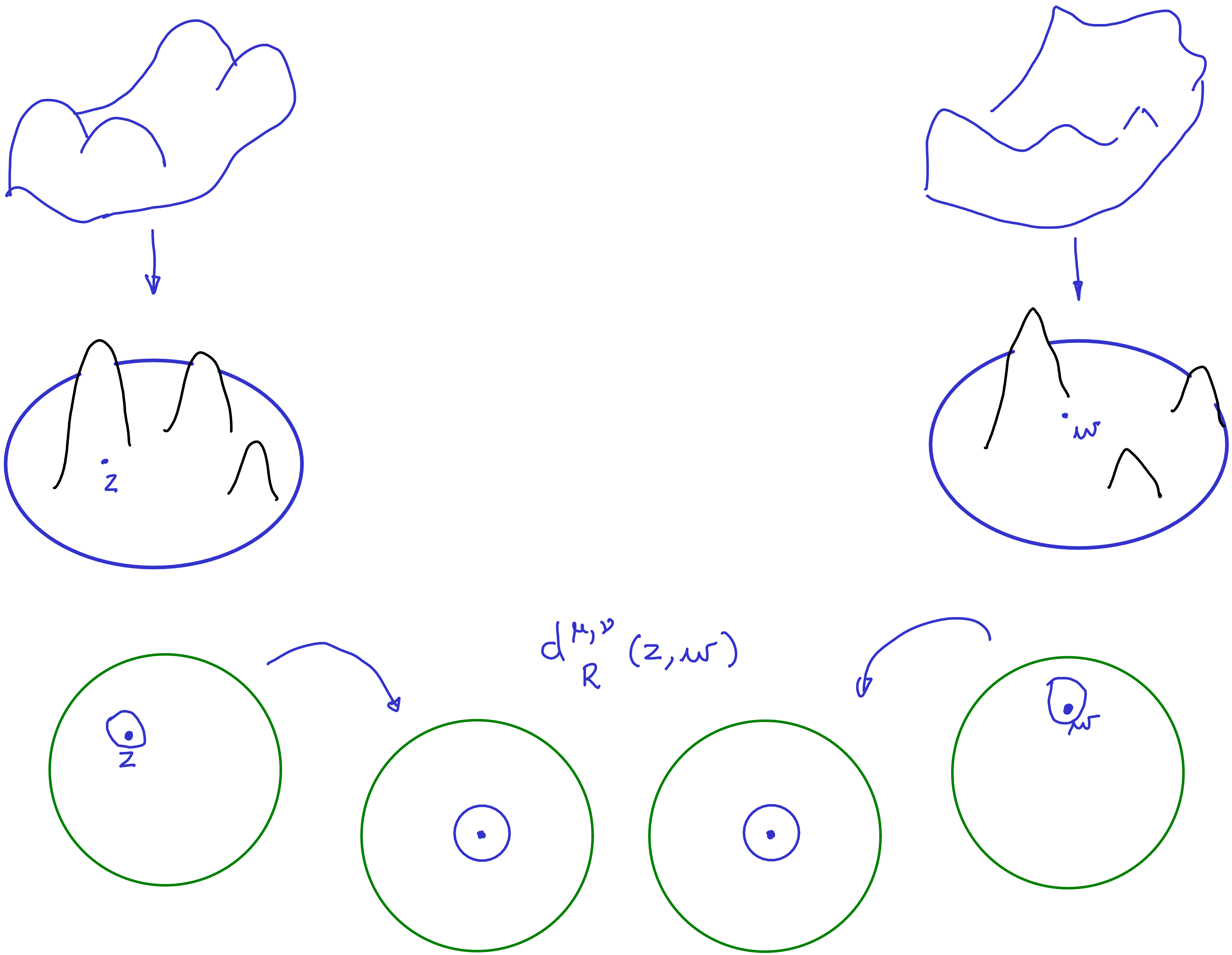



z

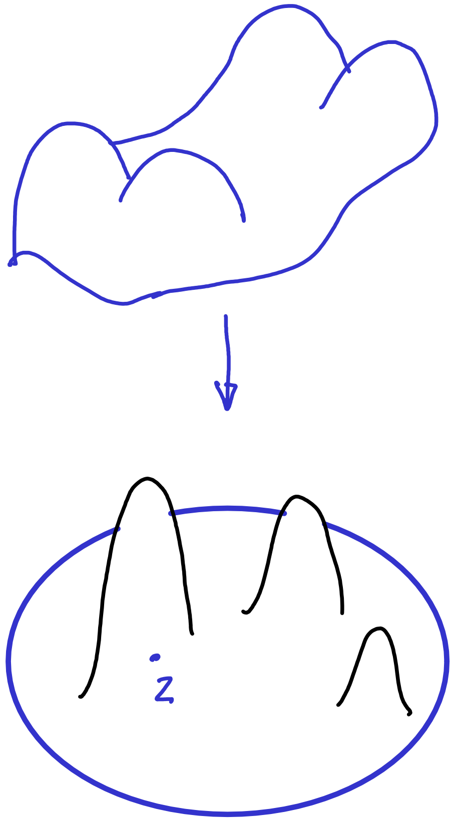

w

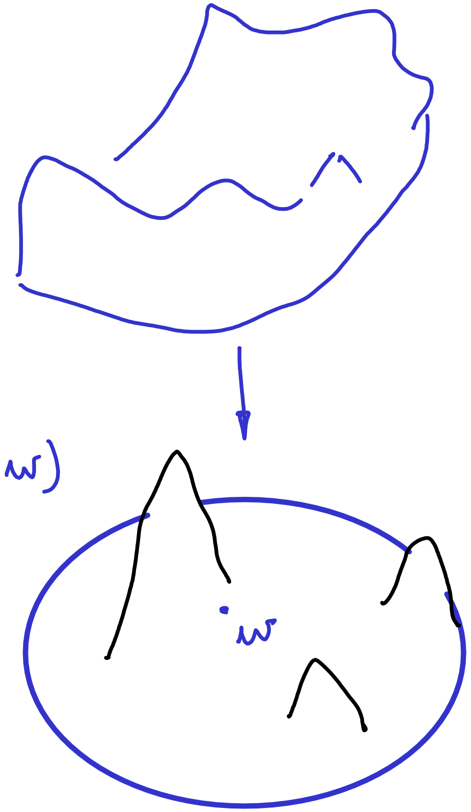

z
z


w
w

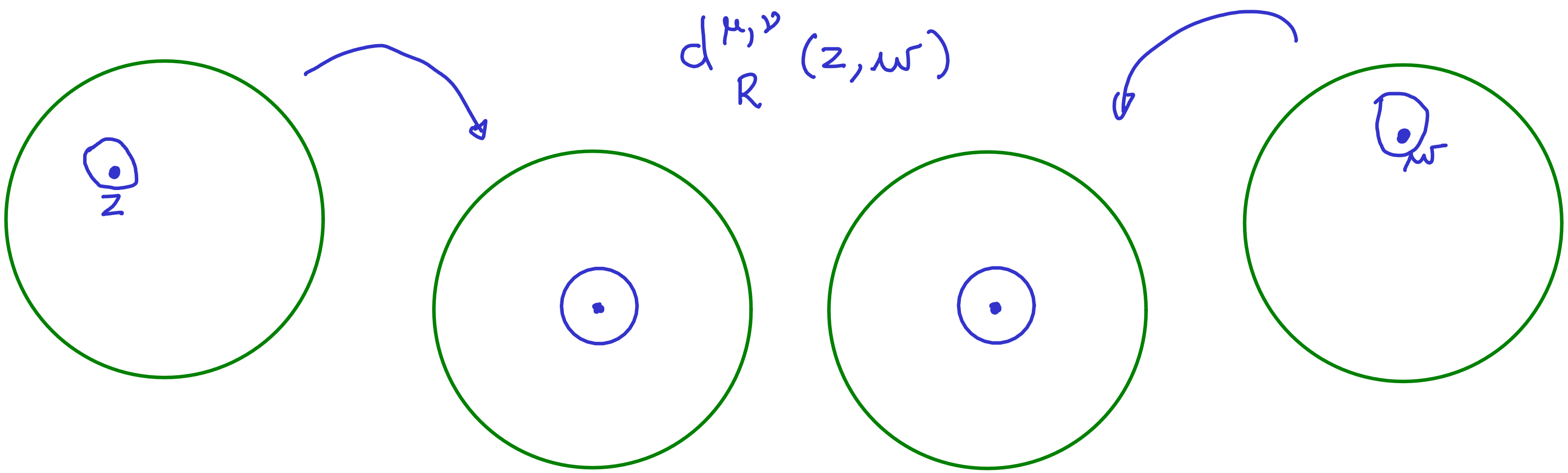


z
w
$d_R^{\mu,\nu}(z,w)$
z
w

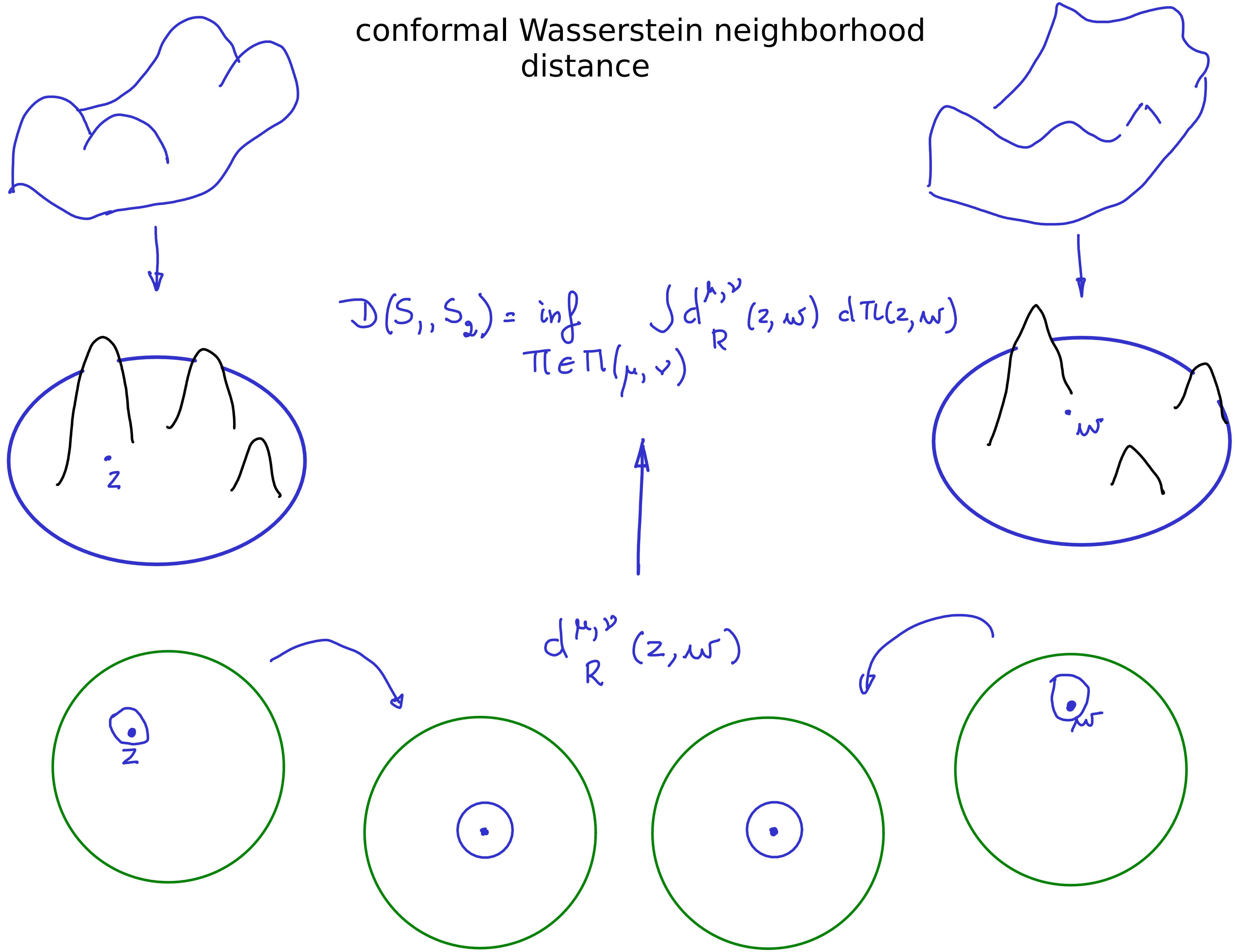




$$D(S_1, S_2) = \inf_{\pi \in \Pi(\mu,\nu)} \int d_R^{\mu,\nu}(z,w)\, d\pi(z,w)$$



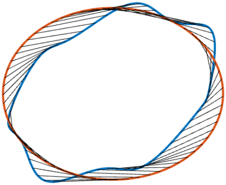
conformal Wasserstein neighborhood distance



Continuous Procrustes Distance (cPD)

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

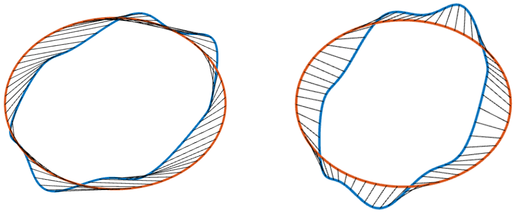
where $\mathcal{C} : S_1 \to S_2$ is an area-preserving diffeomorphism.



Continuous Procrustes Distance (cPD)

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

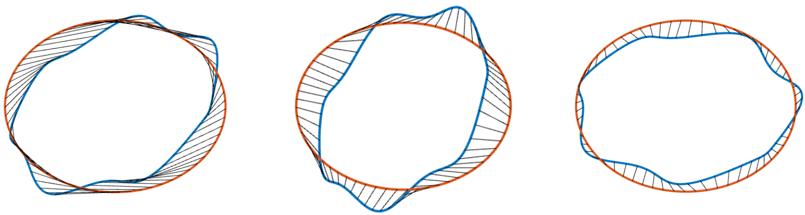
where $\mathcal{C} : S_1 \to S_2$ is an area-preserving diffeomorphism, and $\mathbb{E}_3$ is the Euclidean group on $\mathbb{R}^3$.



Continuous Procrustes Distance (cPD)

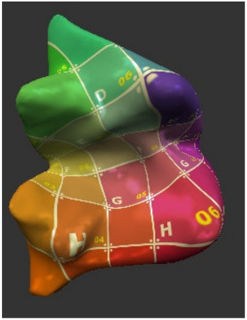
$$D_{\mathrm{cP}}\left(S_1, S_2\right) = \left(\inf_{\mathcal{C}\in\mathcal{A}(S_1,S_2)} \inf_{R\in\mathbb{E}(3)} \int_{S_1} \left\|R\left(x\right) - \mathcal{C}\left(x\right)\right\|^2 d\mathrm{vol}_{S_1}\left(x\right)\right)^{\frac{1}{2}},$$

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

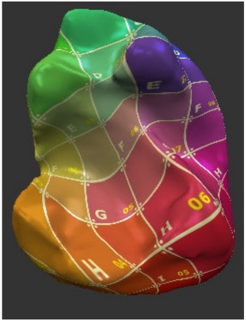


Continuous Procrustes Distance (cPD)

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



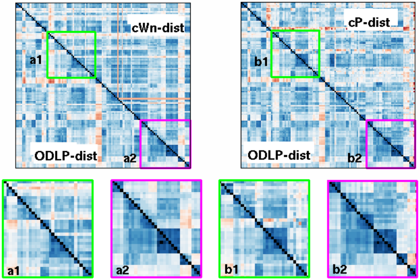
$\xrightarrow{d_{12}}$

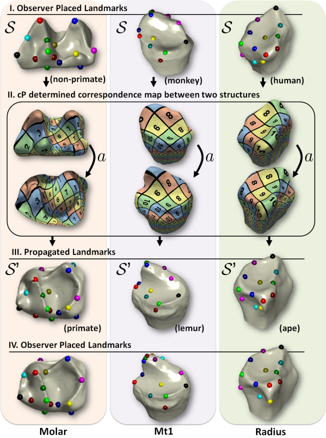


We defined 2 different distances

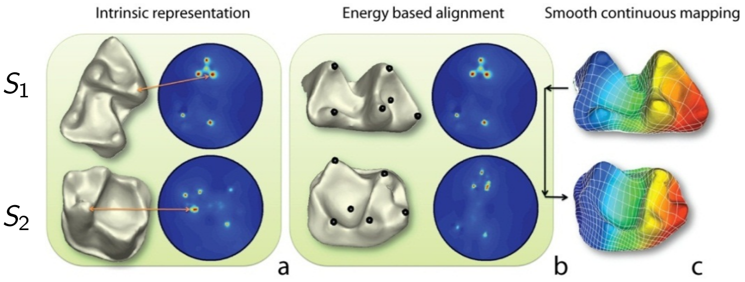
$d_{\mathrm{cWn}}(S_1, S_2)$: conformal flattening
comparison of neighborhood geometry
optimal mass transport

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

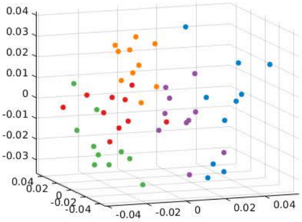




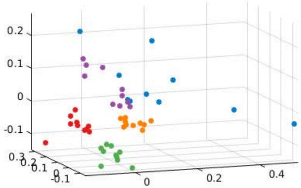
Bypass Explicit Feature Extraction



MDS for cPD & DD

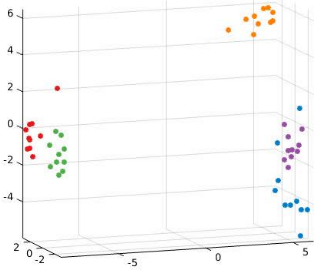


cPD

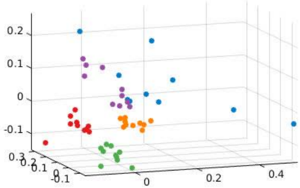


DD

Even better can be obtained!

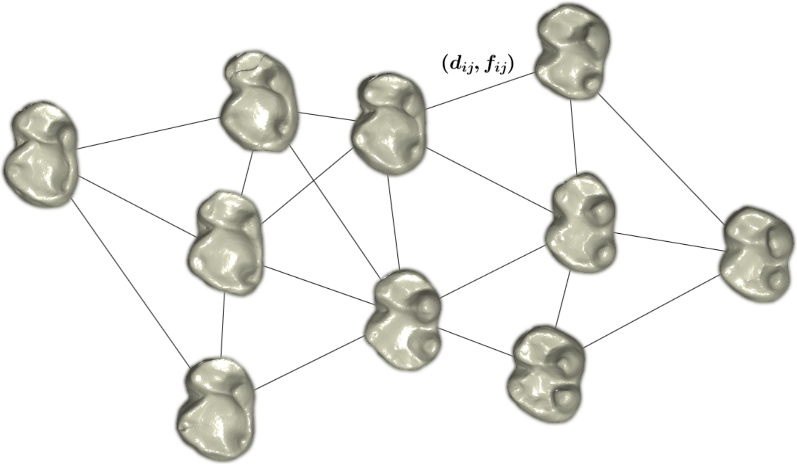


HBDD



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!

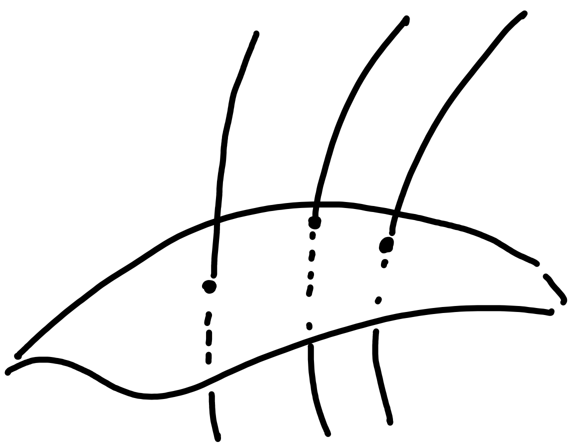
in fact: we have a fiber bundle.
(because of the mappings)

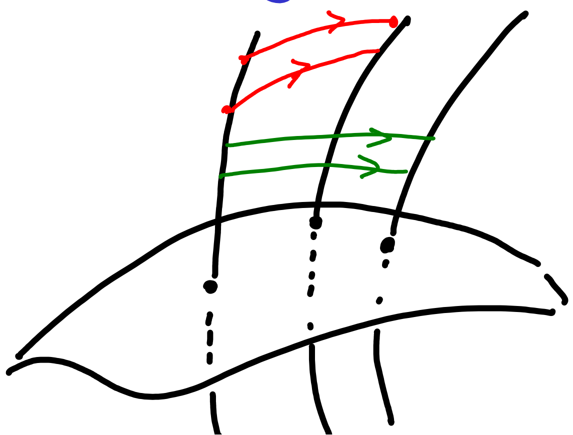
to get Diffusion Distance : used local distances
knitted together
→ spectral parametrization
→ distance.

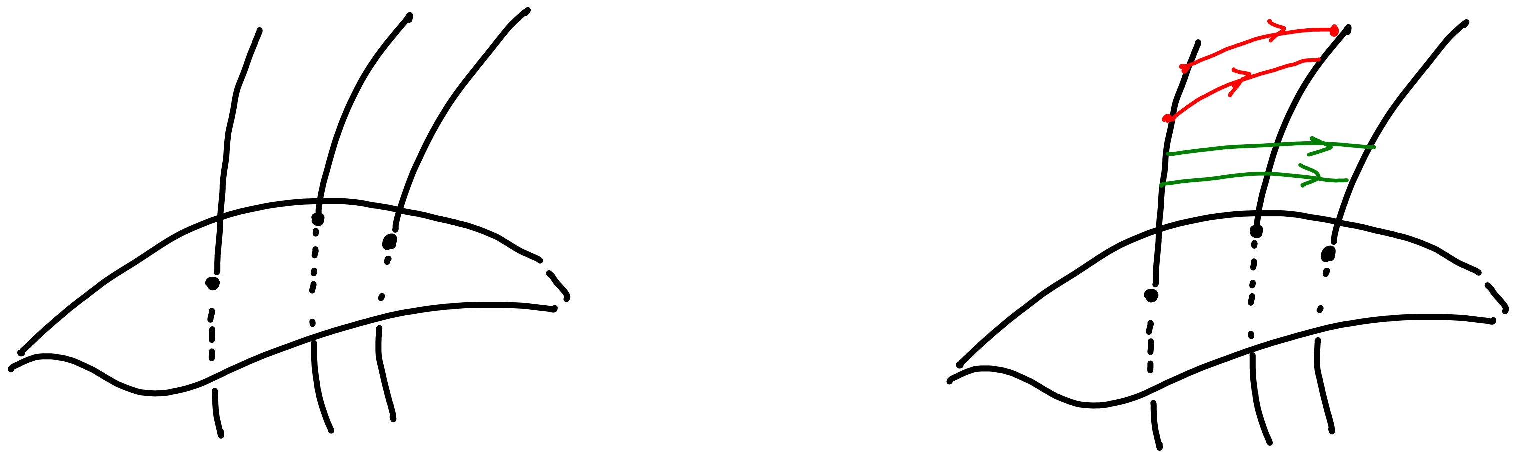
mappings were used only to obtain numerical values for local distances.

but they can do much more for us!

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



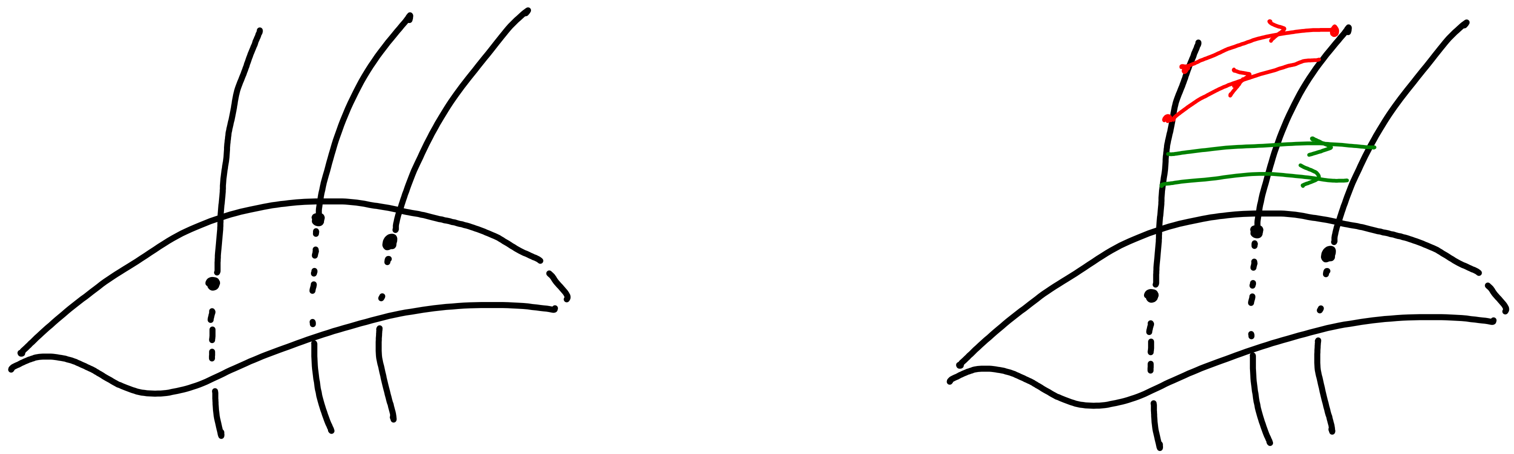




Connection.

↓

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*

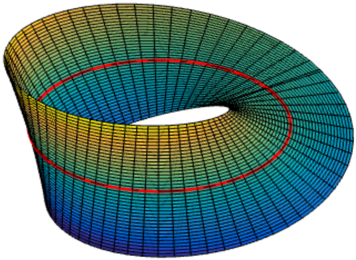
Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

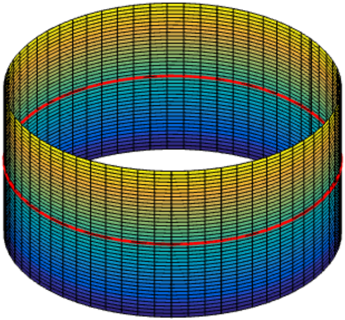
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold

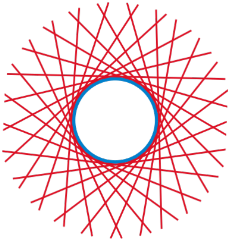
Horizontal Random Walk on a *Fibre Bundle*

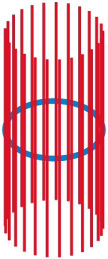
Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

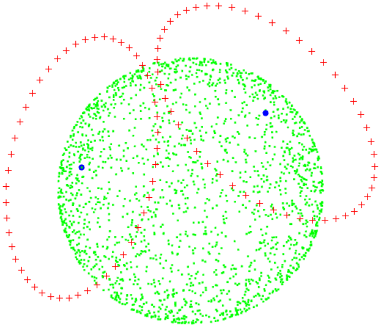
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$







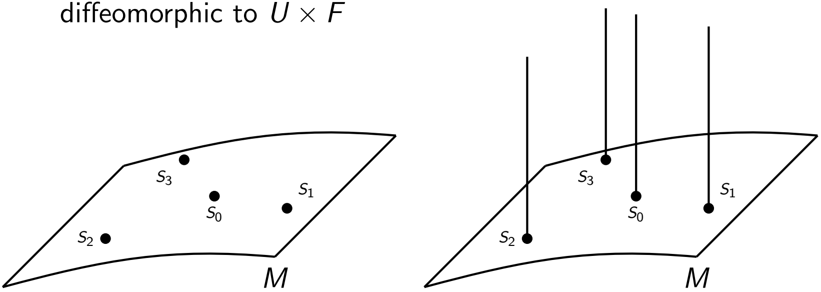




Horizontal Random Walk on a *Fibre Bundle*

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

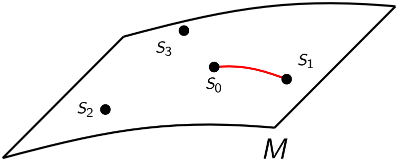
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$

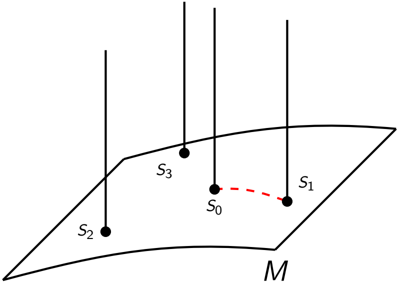


Horizontal Random Walk on a *Fibre Bundle*

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 $U \times F$

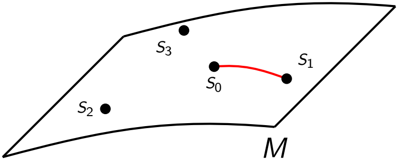


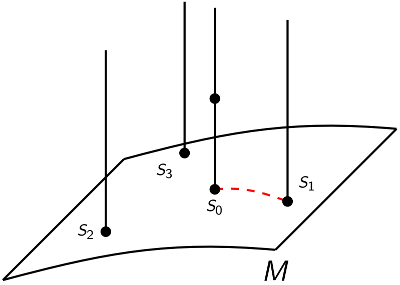


Horizontal Random Walk on a *Fibre Bundle*

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 $U \times F$

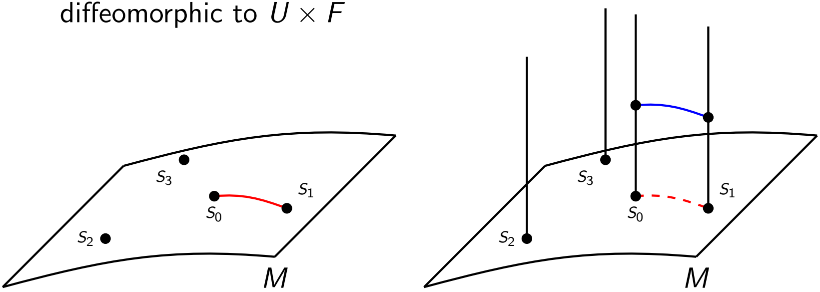




Horizontal Random Walk on a Fibre Bundle

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

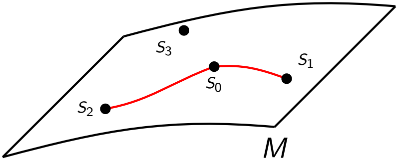
- E: *total* manifold
- M: *base* manifold
- $\pi : E \to M$: smooth surjective map (*bundle projection*)
- F: *fibre* manifold
- *local triviality*: for "small" open set $U \subset M$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$

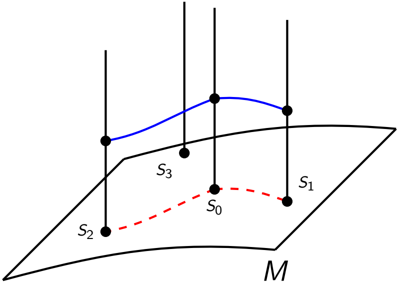


Horizontal Random Walk on a *Fibre Bundle*

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 $U \times F$

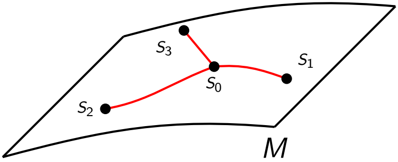


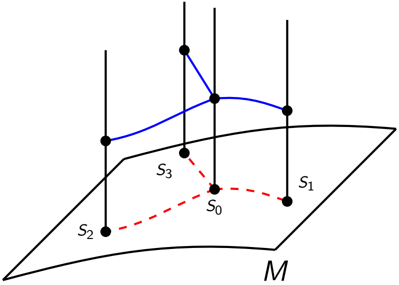


Horizontal Random Walk on a *Fibre Bundle*

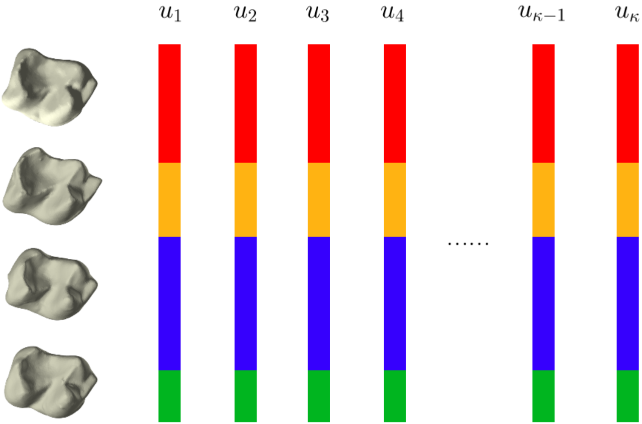
Fibre Bundle $\mathscr{E} = (E, M, F, \pi)$

- 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 $U \times F$

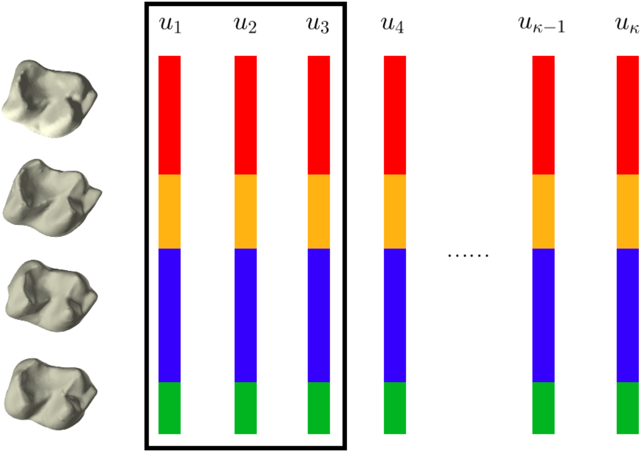




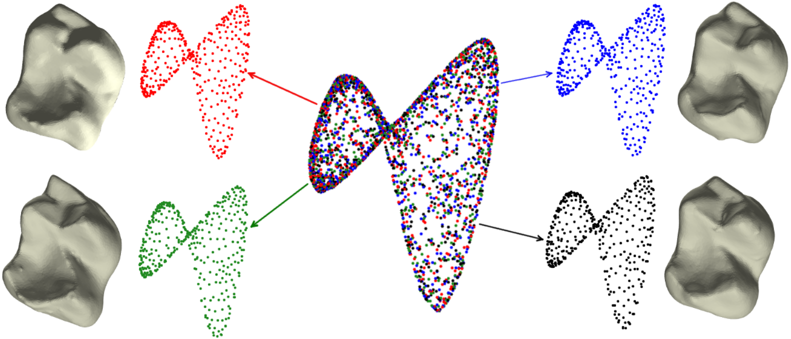
Horizontal Diffusion Maps: Embedding the Entire Bundle



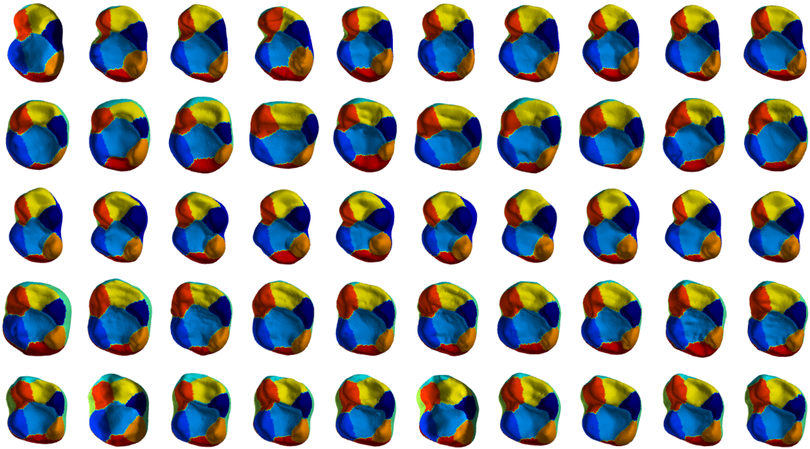
Horizontal Diffusion Maps: Embedding the Entire Bundle



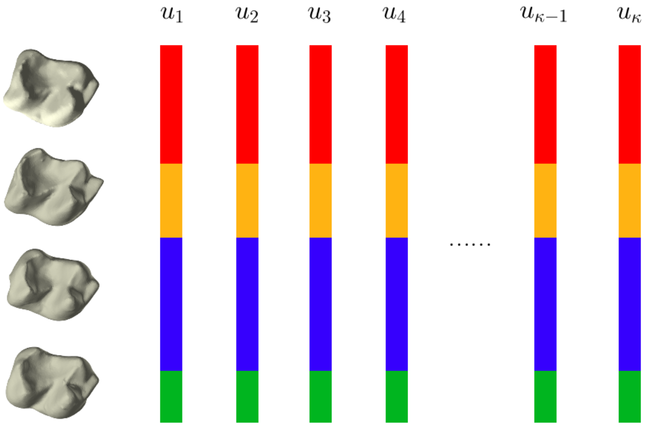
Horizontal Diffusion Maps



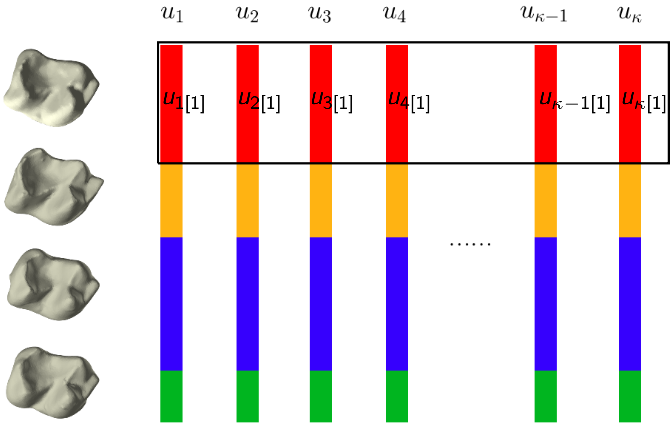
Automatic Landmarking — Interpretability



Horizontal Diffusion Maps: Embedding the Base Manifold

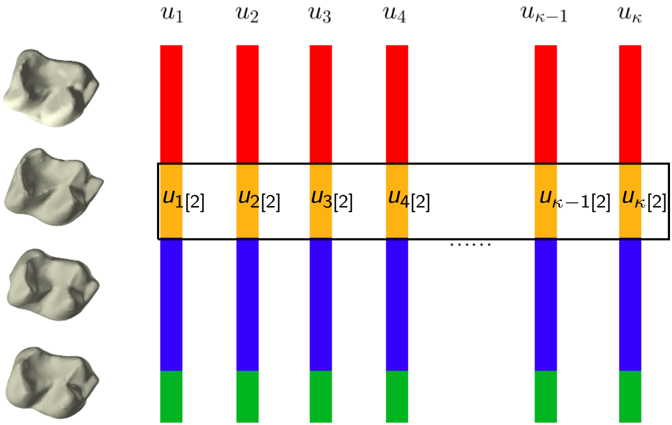


Horizontal Diffusion Maps: Embedding the Base Manifold



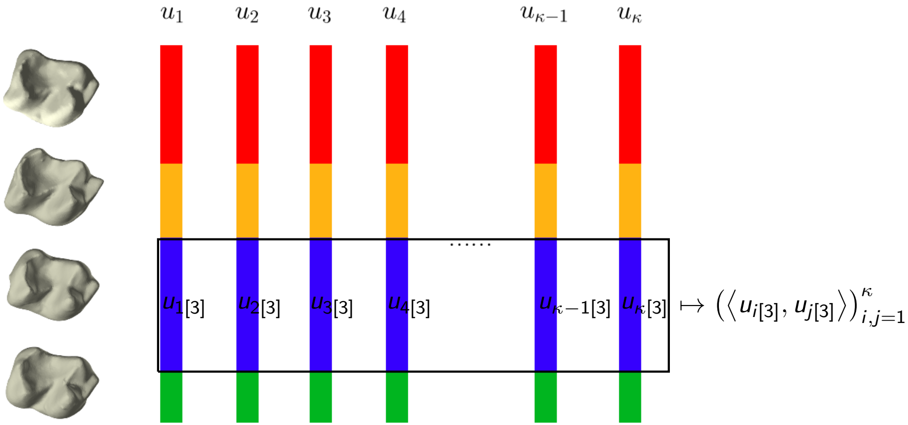
$$\mapsto \left(\left\langle u_{i[1]}, u_{j[1]}\right\rangle\right)_{i,j=1}^{\kappa}$$

Horizontal Diffusion Maps: Embedding the Base Manifold

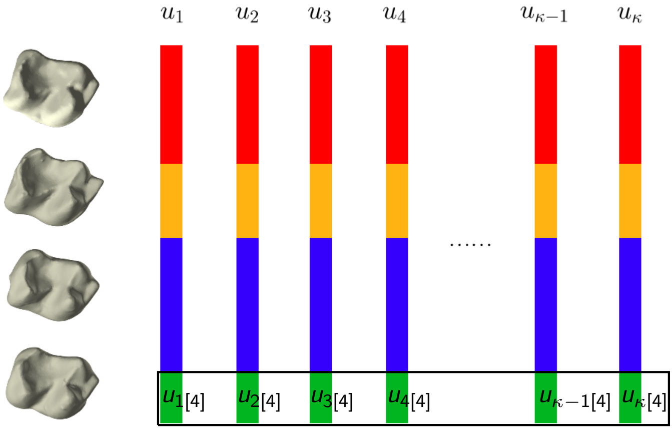


$$\mapsto \left(\left\langle u_{i[2]}, u_{j[2]}\right\rangle\right)_{i,j=1}^{\kappa}$$

Horizontal Diffusion Maps: Embedding the Base Manifold



Horizontal Diffusion Maps: Embedding the Base Manifold



$\mapsto \left(\left\langle u_{i[4]}, u_{j[4]} \right\rangle\right)_{i,j=1}^{\kappa}$

spectral coordinates for points in fiber bundle:

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

S_j

pt p on S_j

spectral coordinates for points in fiber bundle:

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

j: S_j

p: pt p on S_j

↓ "project" to geometry on base manifold

Spectral coordinates for points in fiber bundle:

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

S_j pt p on S_j

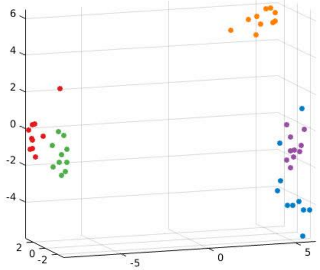
"project" to geometry on base manifold

$$\text{hor. dist}(S_i, S_j)$$

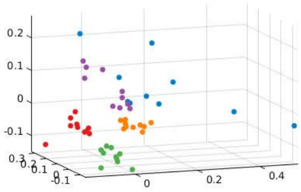
= dist. between corresponding point clouds in K-dim space.

$$= \Big[\sum_{p,q} \lambda_k^{\varepsilon,\delta}(p,q) \, |u_k(i,p) - u_k(j,q)|^2 \Big]^{1/2}$$

Species Clustering

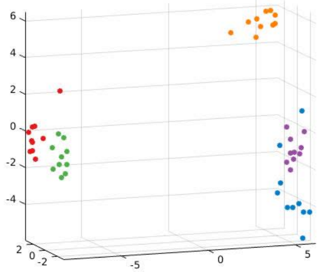


Horizontal Base Diffusion Distance (with Maps)

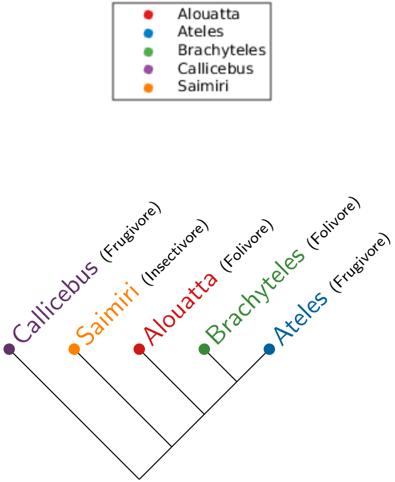


Diffusion Distance (without Maps)

Species Clustering



Horizontal Base Diffusion Distance (with Maps)

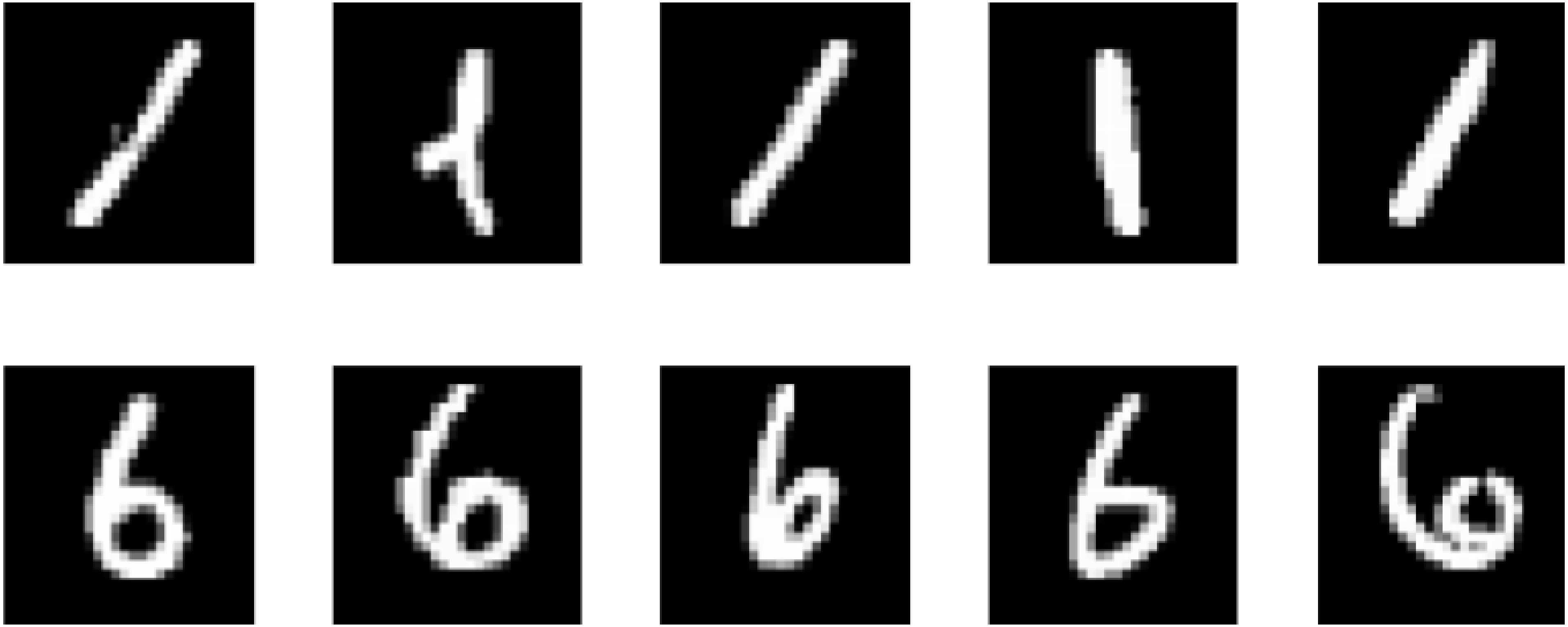


Learning the base manifold in a fibre bundle

This fibre bundle learning method applies much more generally!

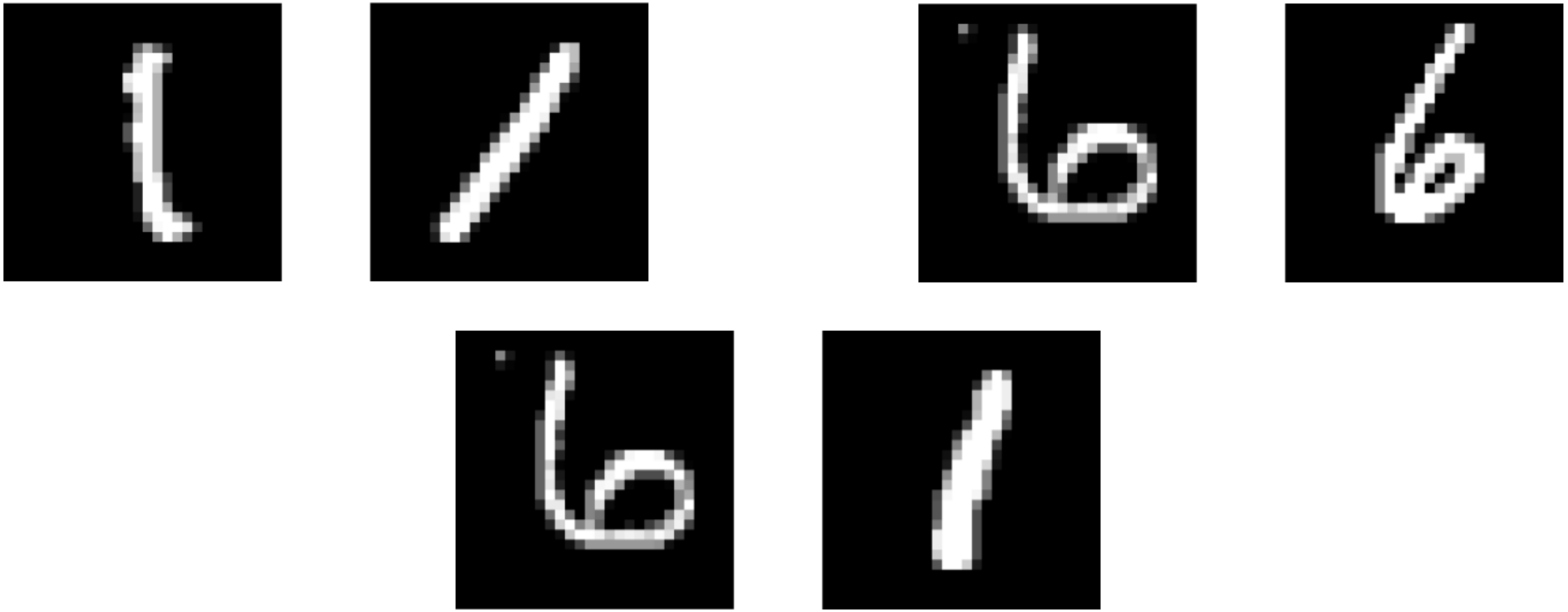
Learning the base manifold in a fibre bundle

This fibre bundle learning method applies much more generally!



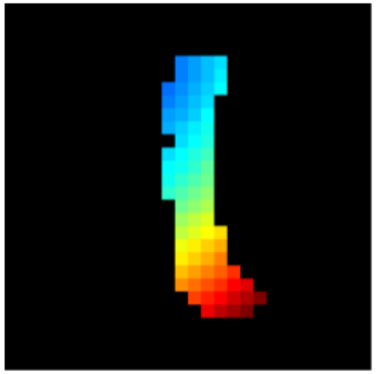
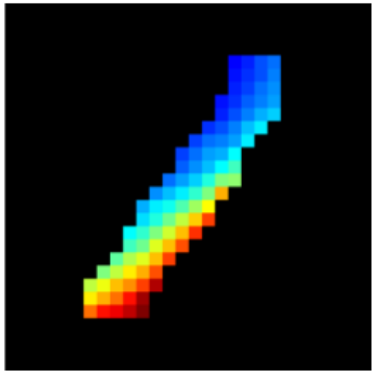
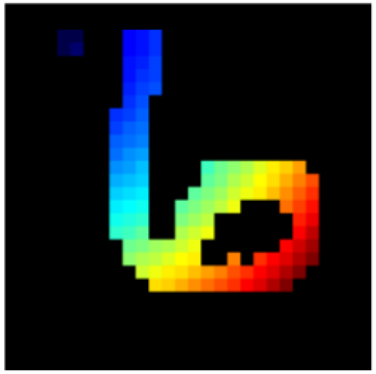
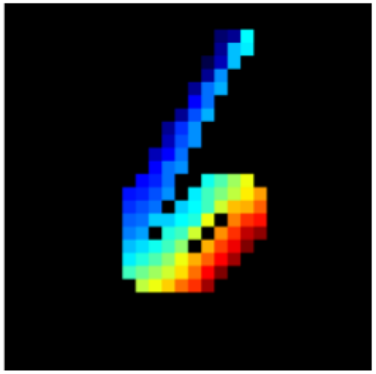
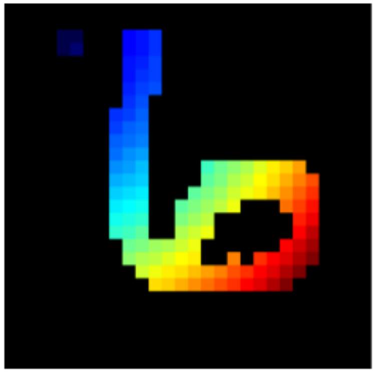
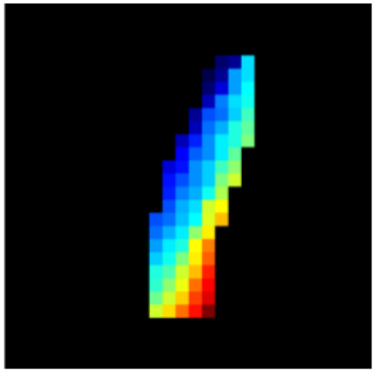
Learning the base manifold in a fibre bundle

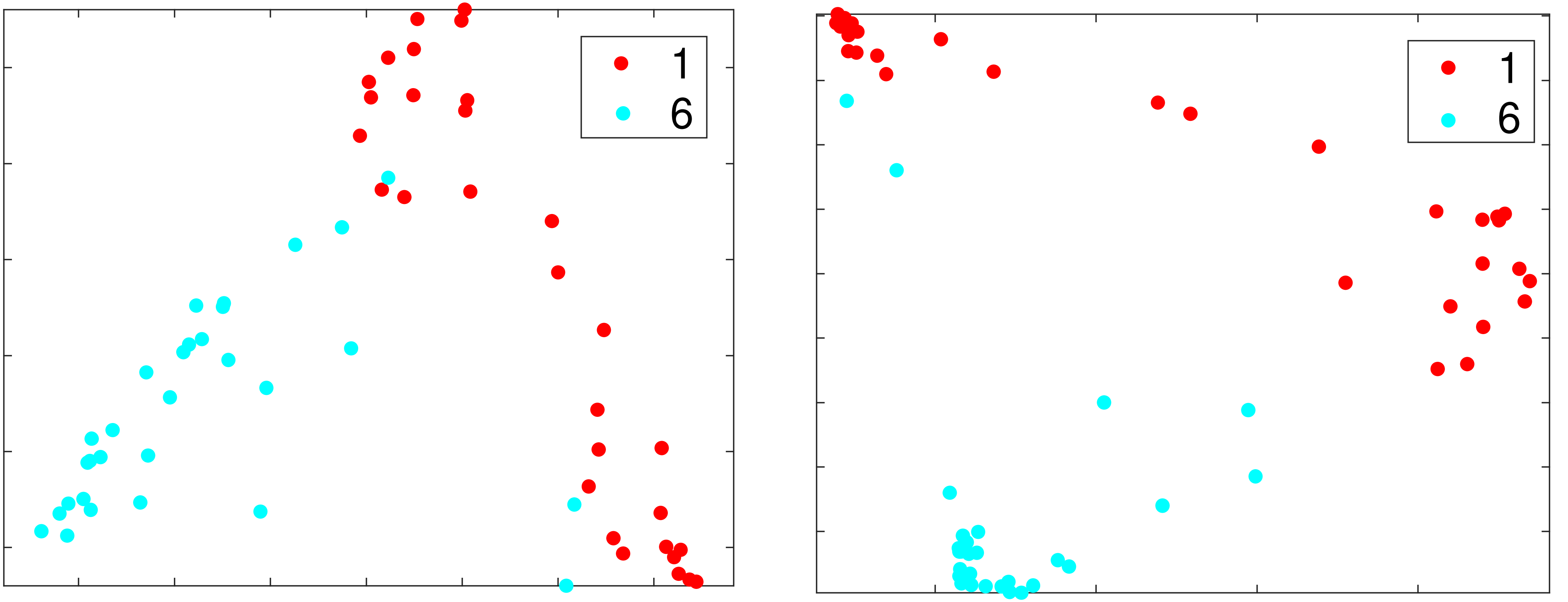
This fibre bundle learning method applies much more generally!

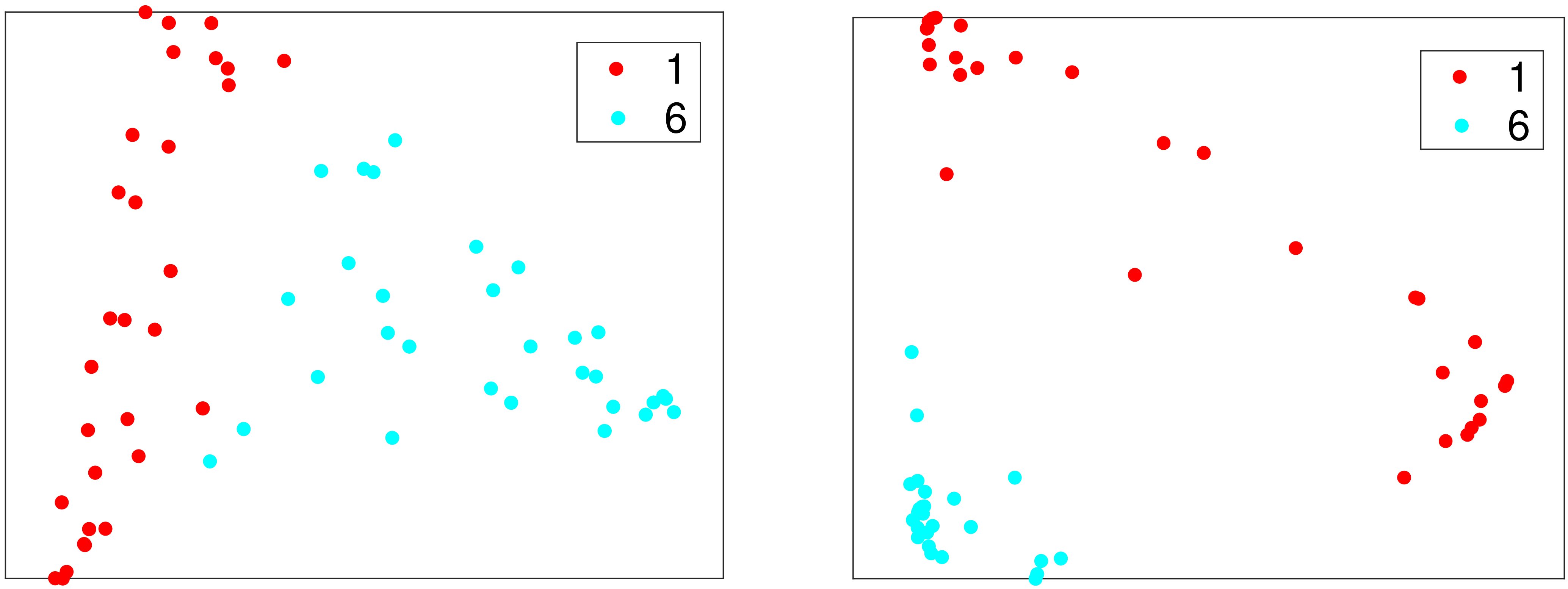


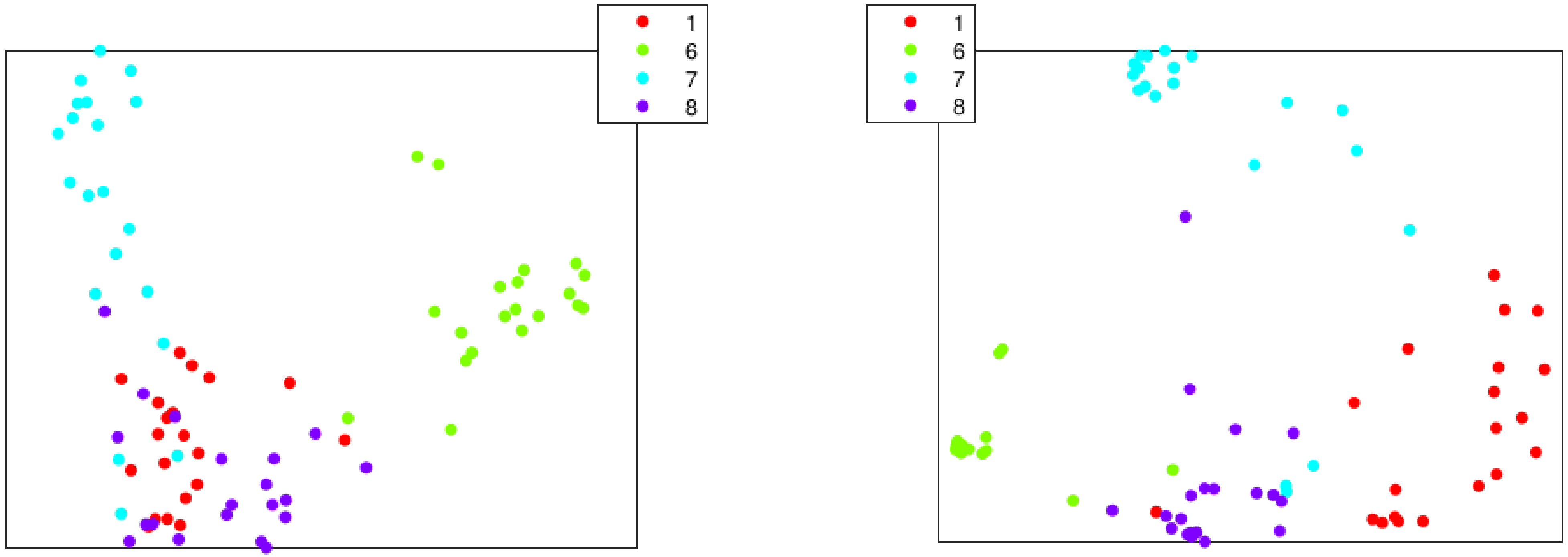
Learning the base manifold in a fibre bundle

This fibre bundle learning method applies much more generally!

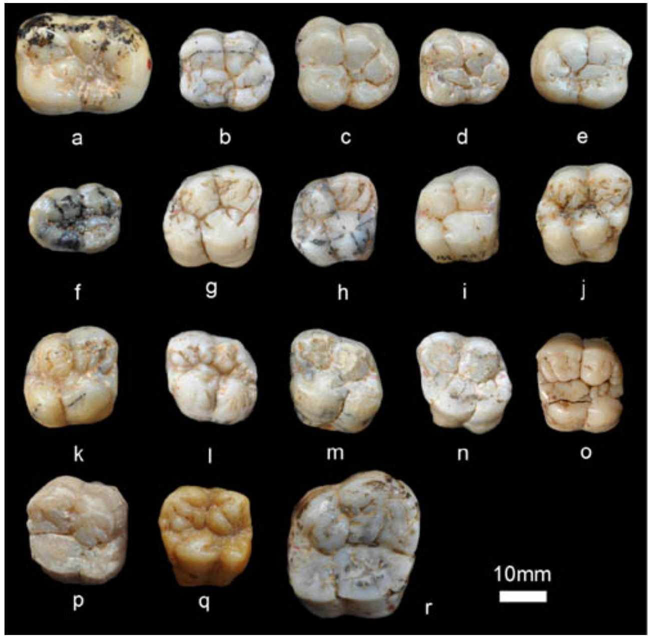

1
6
1
6


1
6
1
6


1
6
7
8
1
6
7
8

- multi-resolution ; coarse- & fine-graining.

Connection is reasonable for bones/teeth of closely related species.



primate molars

crabeater seal molars