Matching Structured Data

Romain Tavenard (Université de Rennes, LETG / Irisa-Obelix)
@rtavenar

Flanders AI Research - "Time series in AI" seminar - June 2022
Companion webpage: bit.ly/fair-tav

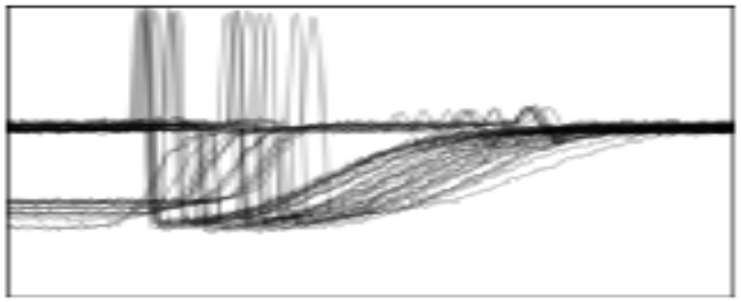
Context of my research

Machine Learning for Remote Sensing

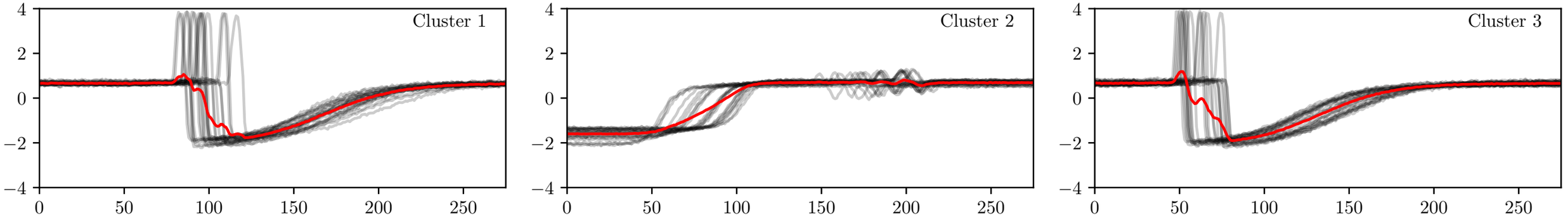
- Challenges
  - Labelled training samples are…
    - Costly to acquire
    - Noisy
  - Heterogeneous data (different sensors, …)
  - Temporal aspects (satellite constellations) and other structural information

Image from esa.int

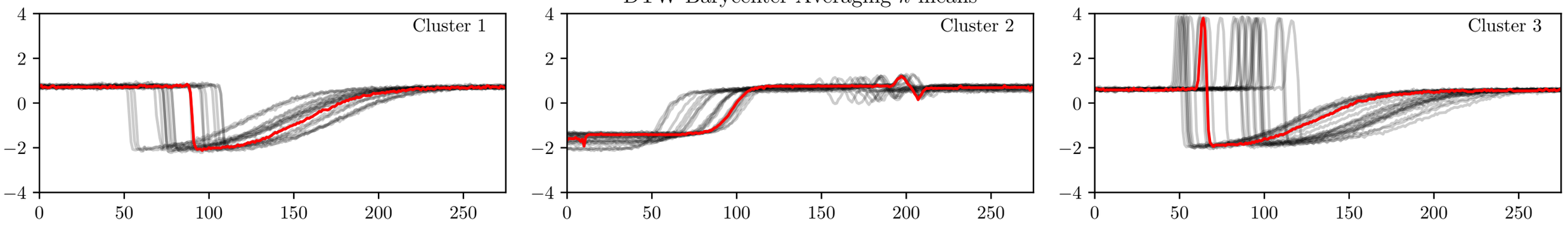
Why using dedicated matching-based metrics?



Euclidean k-means

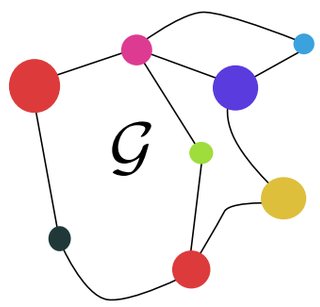


DTW Barycenter Averaging k-means



In this talk, I will…

- Introduce distances between structured objects…
  - Graphs



  - (Sets of) Time Series

$$\mathbf{x} = (\bullet, \cdots, \bullet)$$

- … to be used in machine learning models

Optimal Transport in a nutshell

Two probability distributions

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{i=j}^{m} b_j \delta_{y_j}$$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Bakeries = quantity of breads

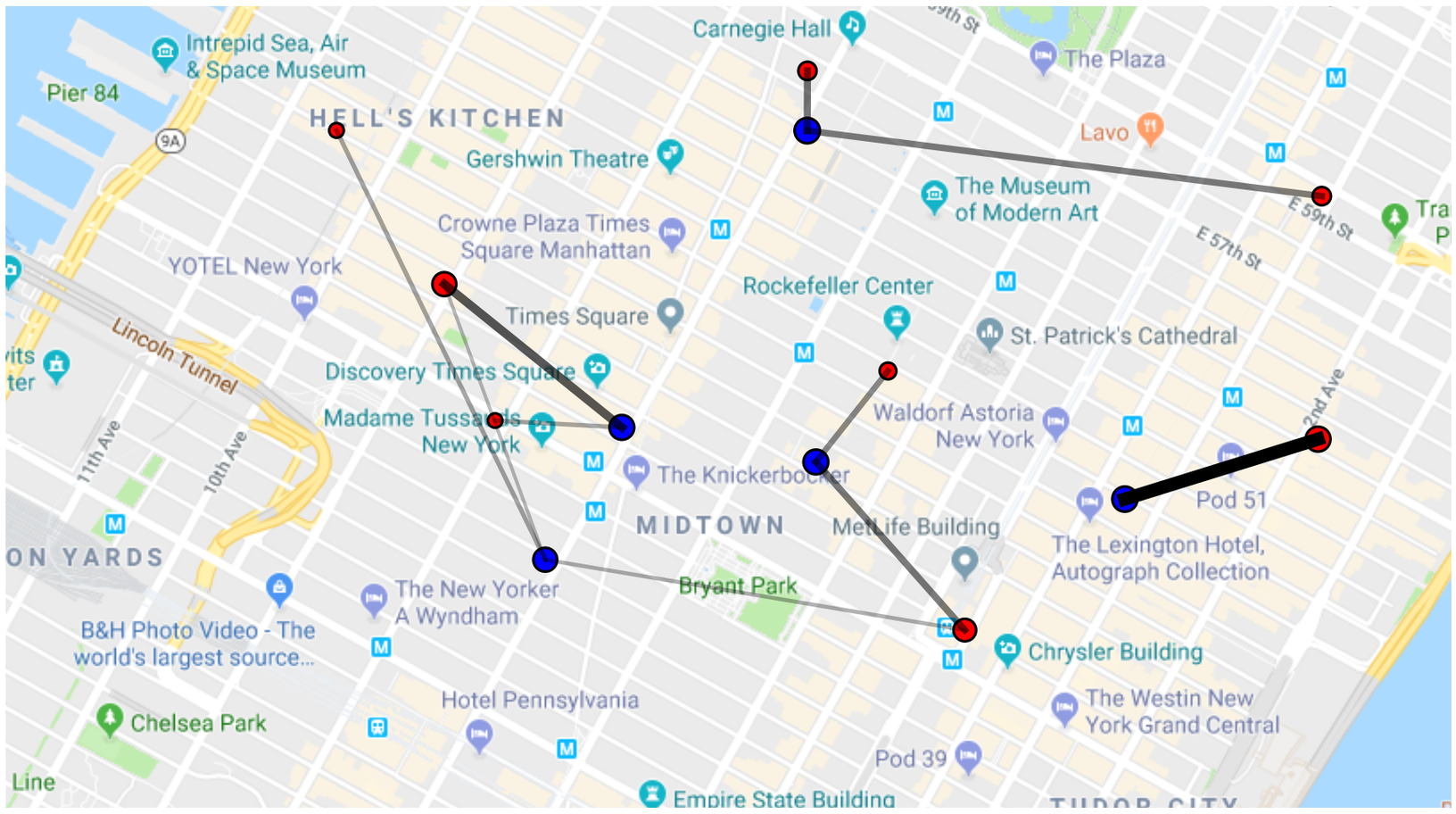
loc: x_i quantity: a_i

Cafés = demand of breads

loc: y_j demand: b_j

Distance between bakeries and cafés

$$c(x_i, y_j)$$



We want to route all the breads from bakeries to cafés the cheapest way

Optimal Transport in a nutshell

Two probability distributions

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{i=j}^{m} b_j \delta_{y_j}$$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Kantorovitch formulation

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j=1}^{m,n} c(x_i, y_j)\pi_{ij}$$

Optimal Transport in a nutshell

Two probability distributions

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{i=j}^{m} b_j \delta_{y_j}$$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Kantorovitch formulation

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j=1}^{m,n} c(x_i, y_j)\pi_{ij}$$

Set of couplings/ transport plans

$$\Pi(\mathbf{a}, \mathbf{b})$$

Optimal Transport in a nutshell

Two probability distributions

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{i=j}^{m} b_j \delta_{y_j}$$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Kantorovitch formulation

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j=1}^{m,n} c(x_i, y_j) \pi_{ij}$$

How much is shifted from x_i to y_j

Optimal Transport in a nutshell

Two probability distributions

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{i=j}^{m} b_j \delta_{y_j}$$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Kantorovitch formulation

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j=1}^{m,n} c(x_i, y_j)\pi_{ij}$$

Cost of moving masses from x_i to y_j

Optimal Transport in a nutshell

Two probability distributions

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{i=j}^{m} b_j \delta_{y_j}$$

A cost function

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Kantorovitch formulation

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j=1}^{m,n} c(x_i, y_j)\pi_{ij}$$

Total cost

Optimal Transport in a nutshell

Two probability distributions

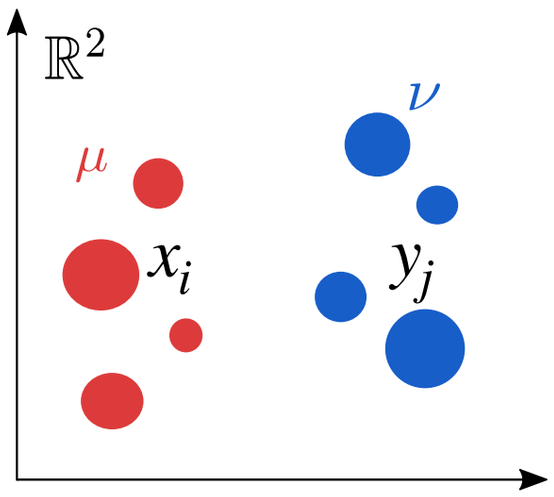
$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \quad \nu = \sum_{i=j}^{m} b_j \delta_{y_j}$$

A cost function

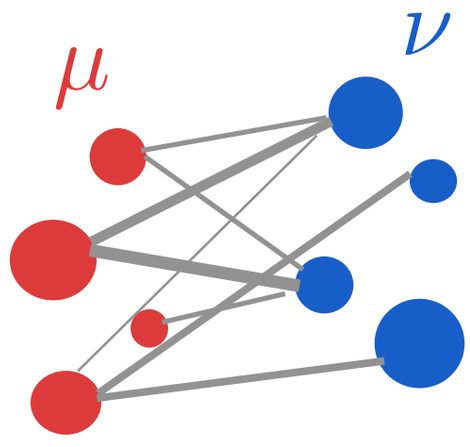
$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

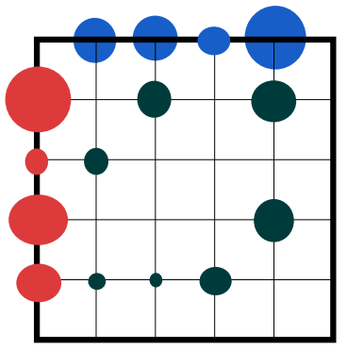
Kantorovitch formulation

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \sum_{i,j=1}^{m,n} c(x_i, y_j)\pi_{ij}$$



$$\Pi(\mathbf{a}, \mathbf{b}) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{n \times m} \mid \forall (i,j), \sum_{j=1}^{m} \pi_{ij} = a_i, \ \sum_{i=1}^{n} \pi_{ij} = b_j\}$$

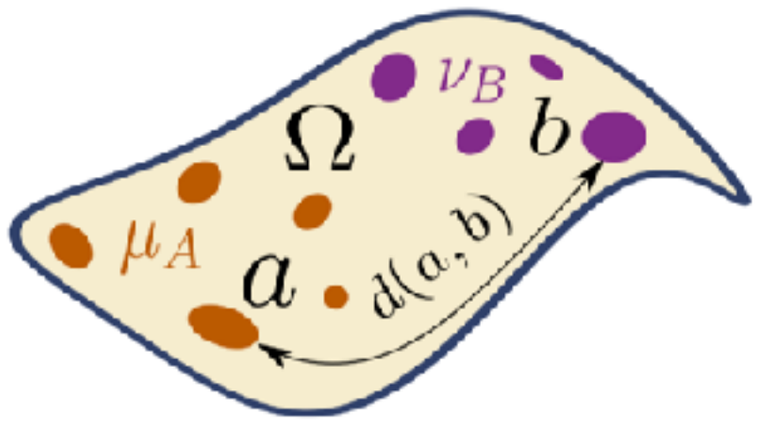


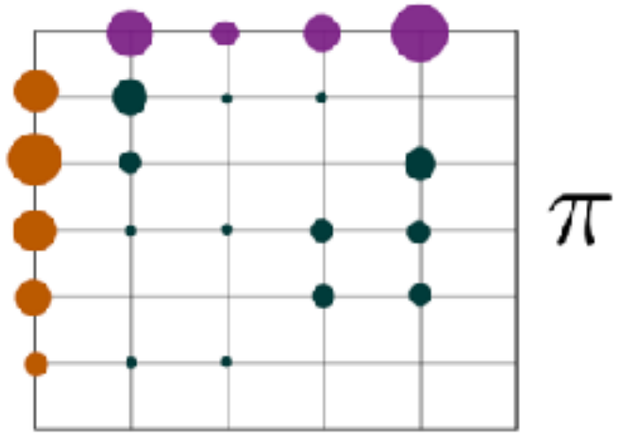


$$\pi \in \mathbb{R}_+^{n \times m}$$

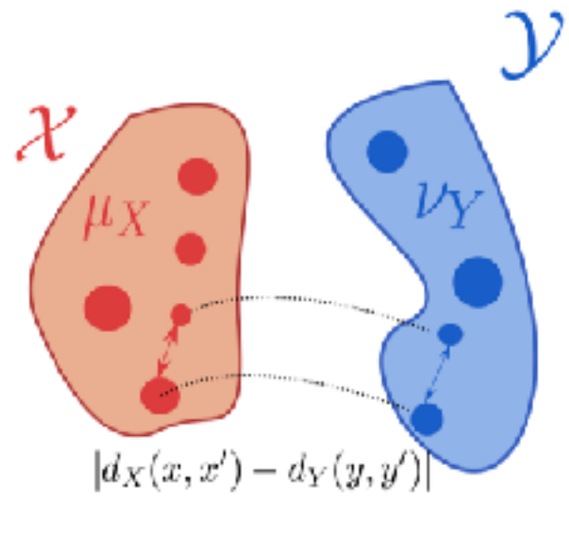
Matching Structures with Optimal Transport

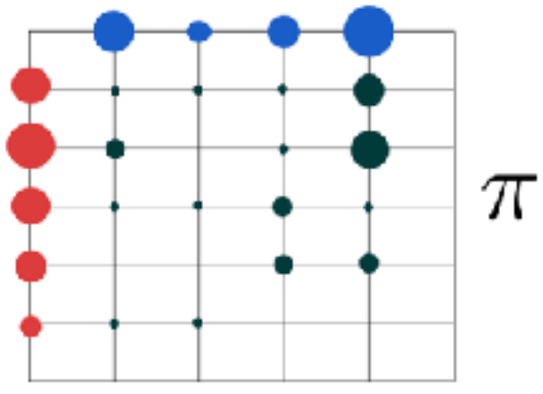
F. Memoli





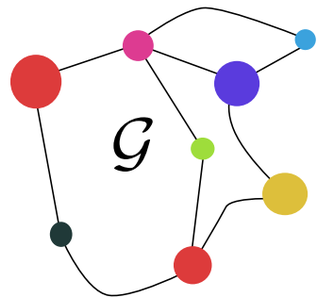
Wasserstein distance

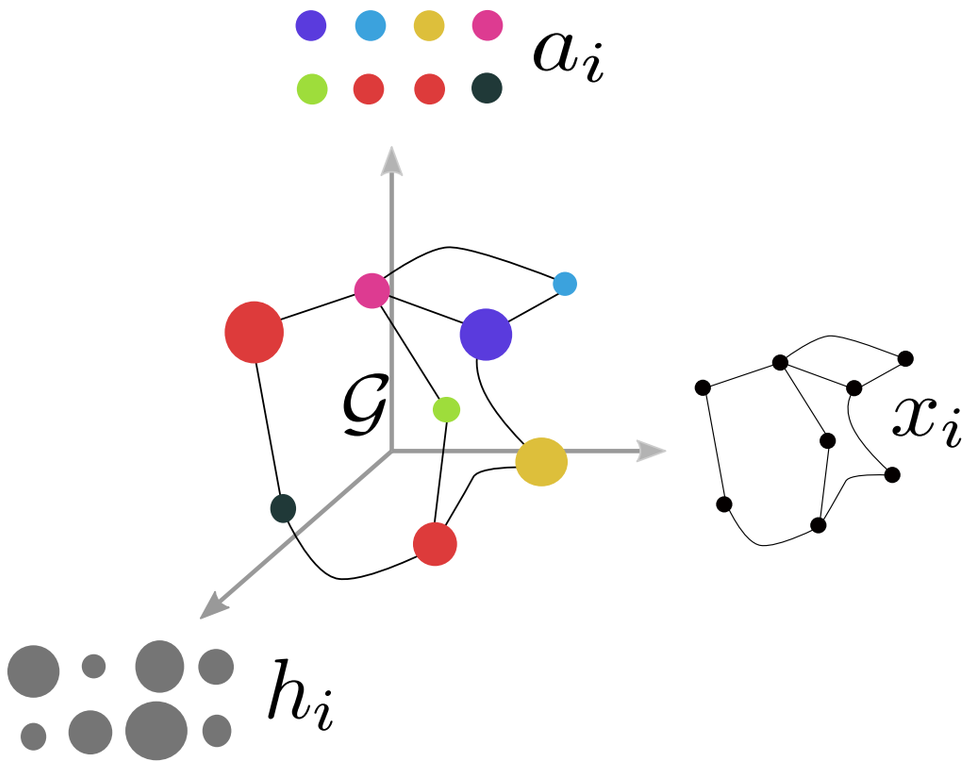




Gromov-Wasserstein distance

Matching Graphs
[Vayer *et al*, ICML'19]



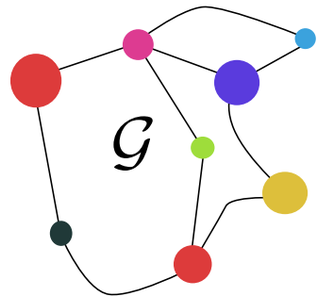


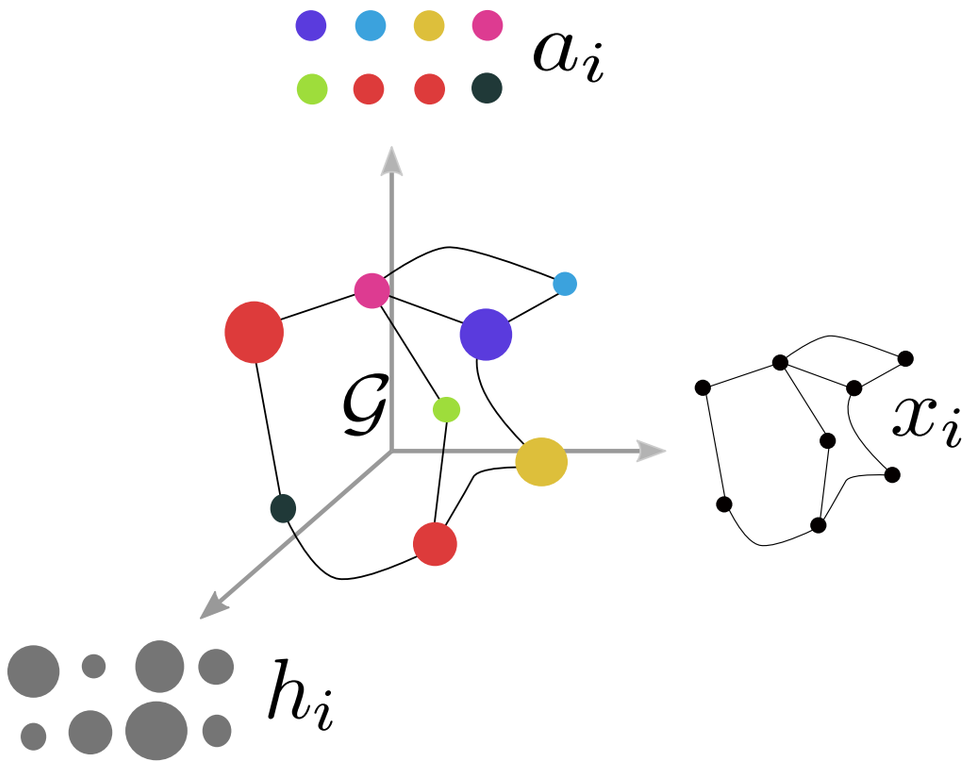
$$\left.\begin{array}{c}\\\\\\\end{array}\right\} \mu_A = \sum_i h_i \delta_{a_i}$$

Wasserstein distance

$$\min_{\pi} \sum_{i,j} d_{ij} \pi_{i,j}$$

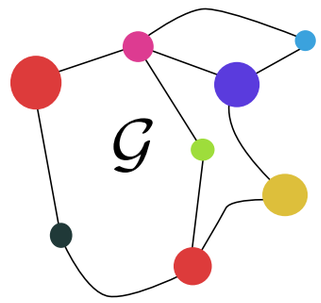
Matching Graphs
[Vayer *et al*, ICML'19]

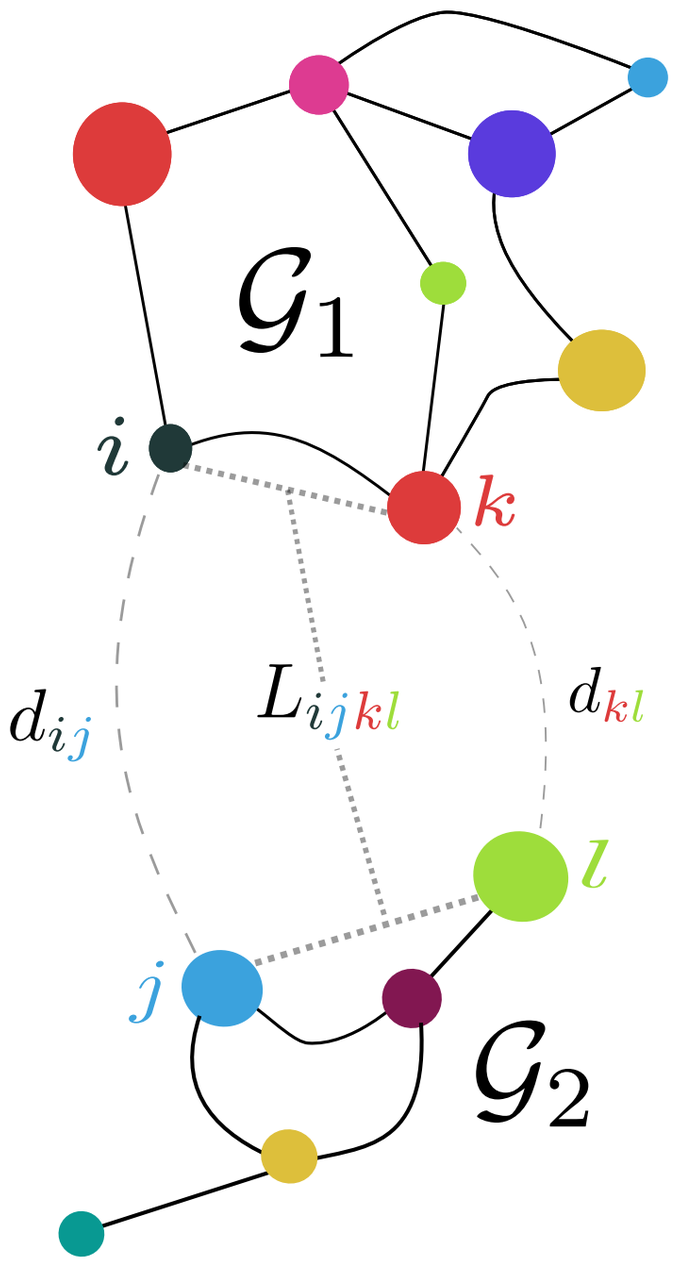




$$\left.\begin{array}{c}\\ \\ \\ \end{array}\right\} \mu = \sum_i h_i \delta_{(x_i, a_i)}$$

Matching Graphs
[Vayer *et al*, ICML'19]





Wasserstein distance

$$\min_{\pi} \sum_{i,j} d_{ij}\pi_{i,j}$$

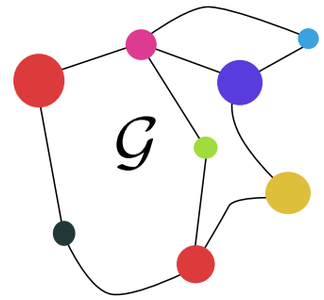
Gromov-Wasserstein distance

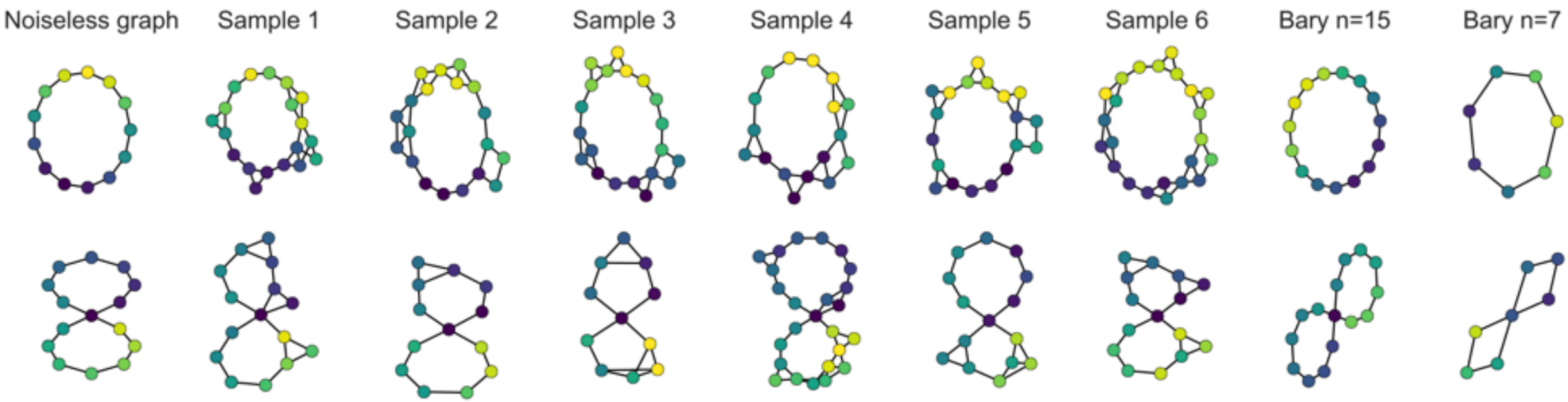
$$\min_{\pi} \sum_{i,j,k,l} L_{ijkl}\pi_{i,j}\pi_{k,l}$$

Fused-Gromov-Wasserstein (FGW) distance

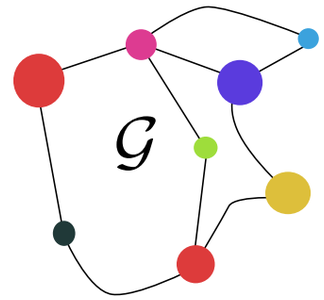
$$\min_{\pi} \sum_{i,j,k,l} \left[(1-\alpha)\, d_{ij} + \alpha L_{ijkl}\right] \pi_{i,j}\pi_{k,l}$$

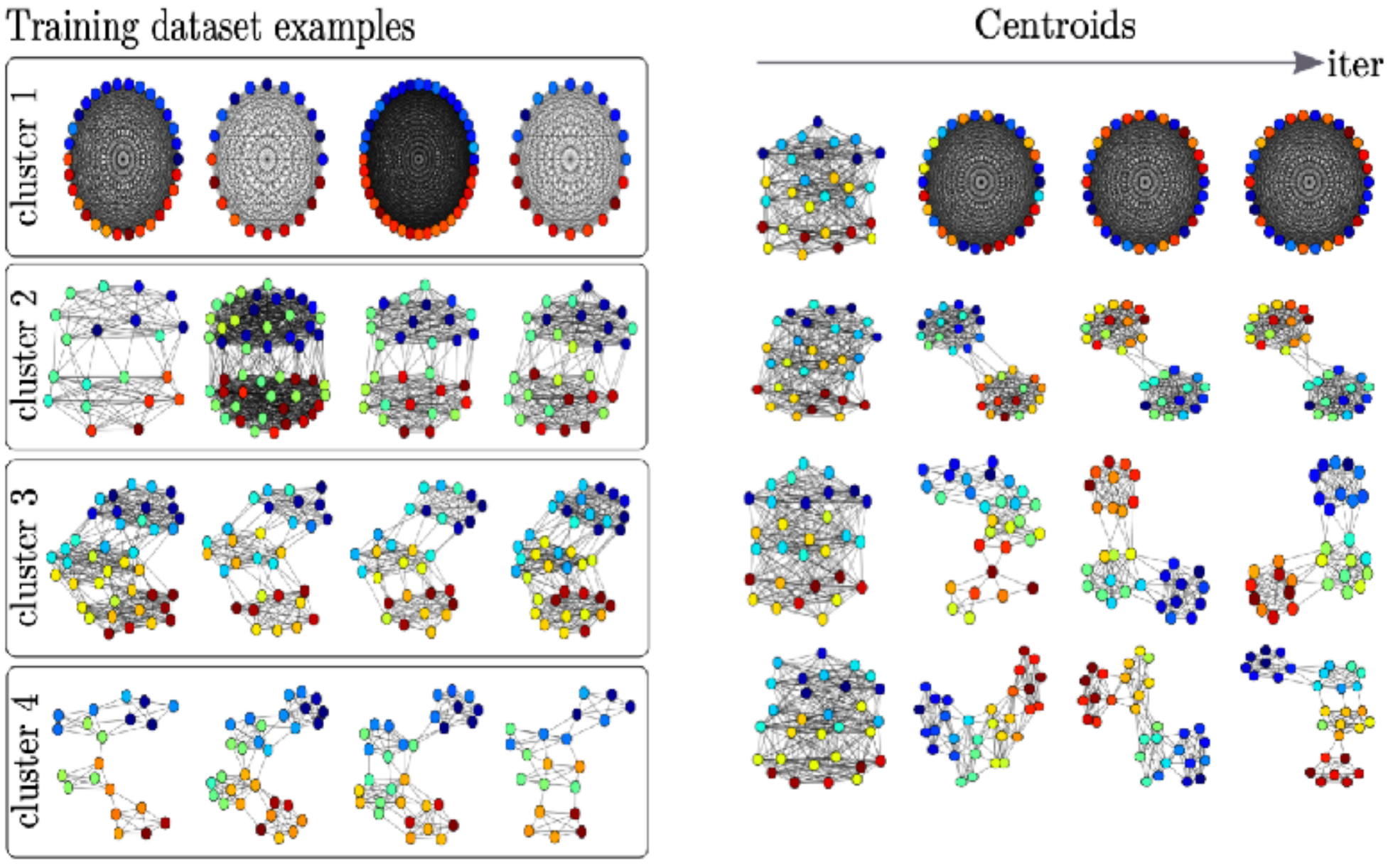
Matching Graphs
[Vayer *et al*, ICML'19]





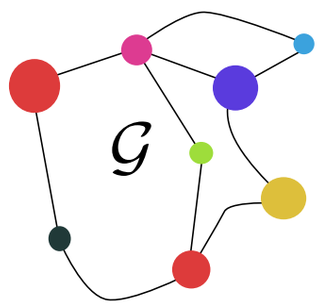
Matching Graphs
[Vayer *et al*, ICML'19]





In this talk, I will…

- Introduce distances between structured objects…
  - Graphs



  - (Sets of) Time Series

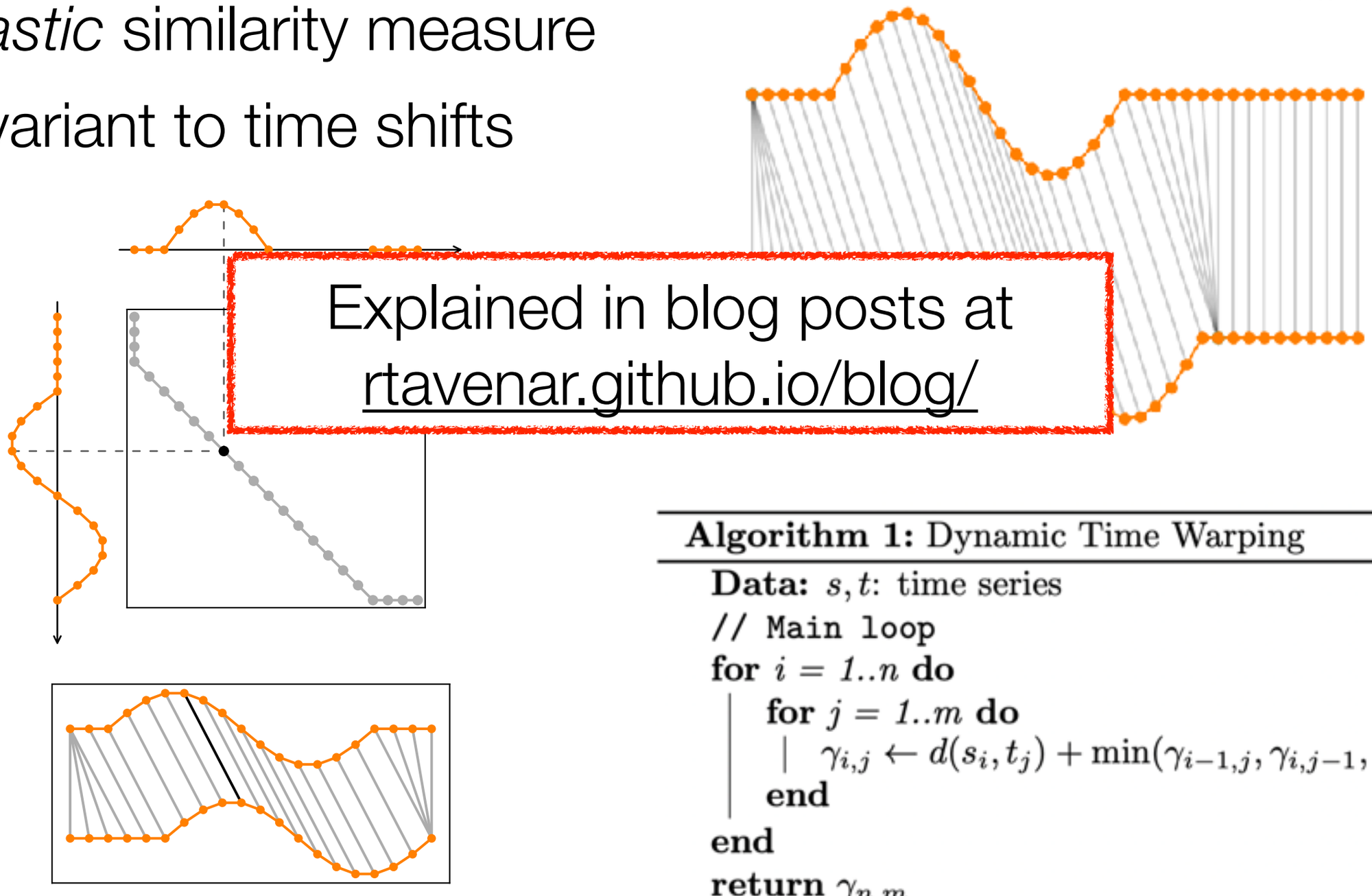
$$\mathbf{x} = (\bullet, \cdots, \bullet)$$

- … to be used in machine learning models

Matching Time Series $\mathbf{x} = (\bullet, \cdots, \bullet)$
Dynamic Time Warping (DTW)

H. Sakoe

- *Elastic* similarity measure
- Invariant to time shifts



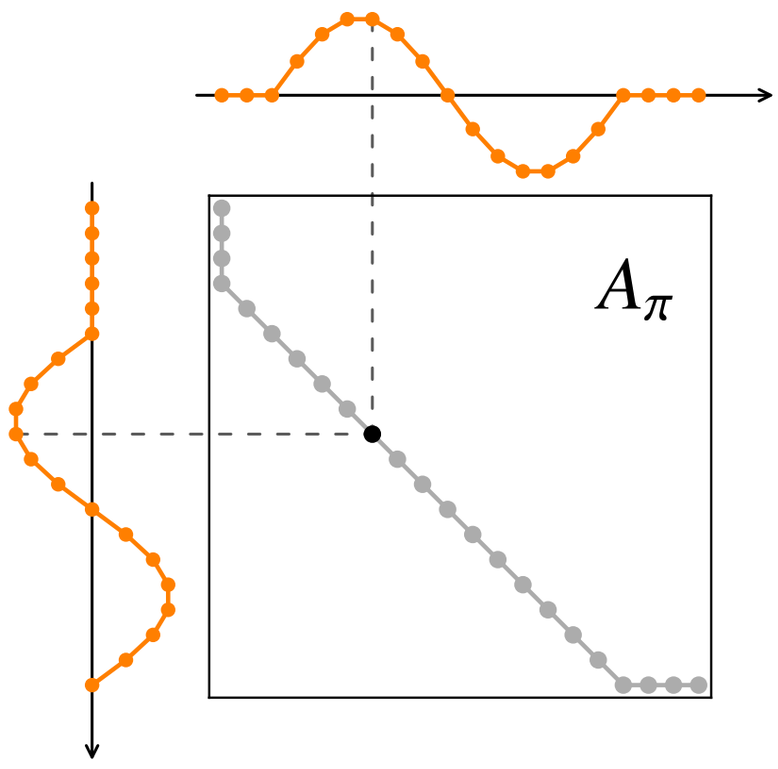
Explained in blog posts at
rtavenar.github.io/blog/

Algorithm 1: Dynamic Time Warping

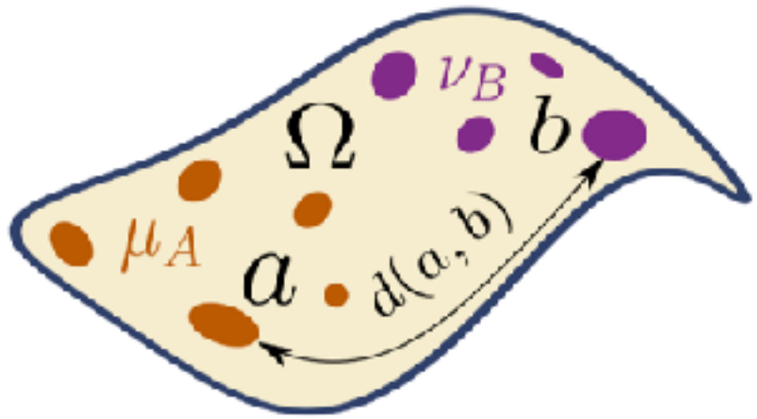
Data: s, t: time series
`// Main loop`
for $i = 1..n$ **do**
  for $j = 1..m$ **do**
    $\gamma_{i,j} \leftarrow d(s_i, t_j) + \min(\gamma_{i-1,j}, \gamma_{i,j-1}, \gamma_{i-1,j-1})$
  end
end
return $\gamma_{n,m}$

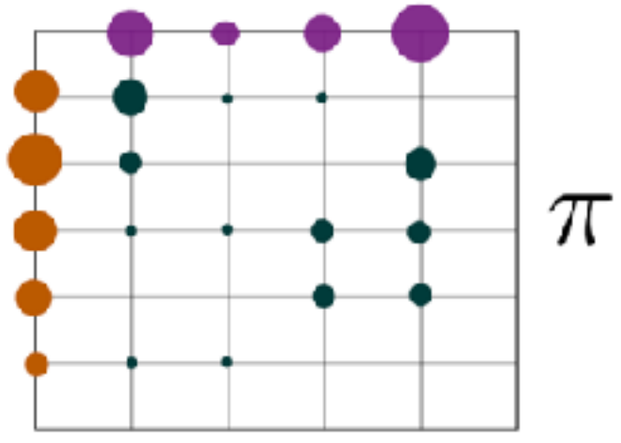
Matching-based metrics
Time Series *vs* distributions

$$\min_{\pi} \sum_{i,j} d_{i,j} \pi_{i,j}$$



Dynamic Time Warping
π is a monotonically increasing connected path





Wasserstein distance
π is a map w. fixed marginals

Matching Time Series

$\mathbf{x} = (\bullet, \cdots, \bullet)$

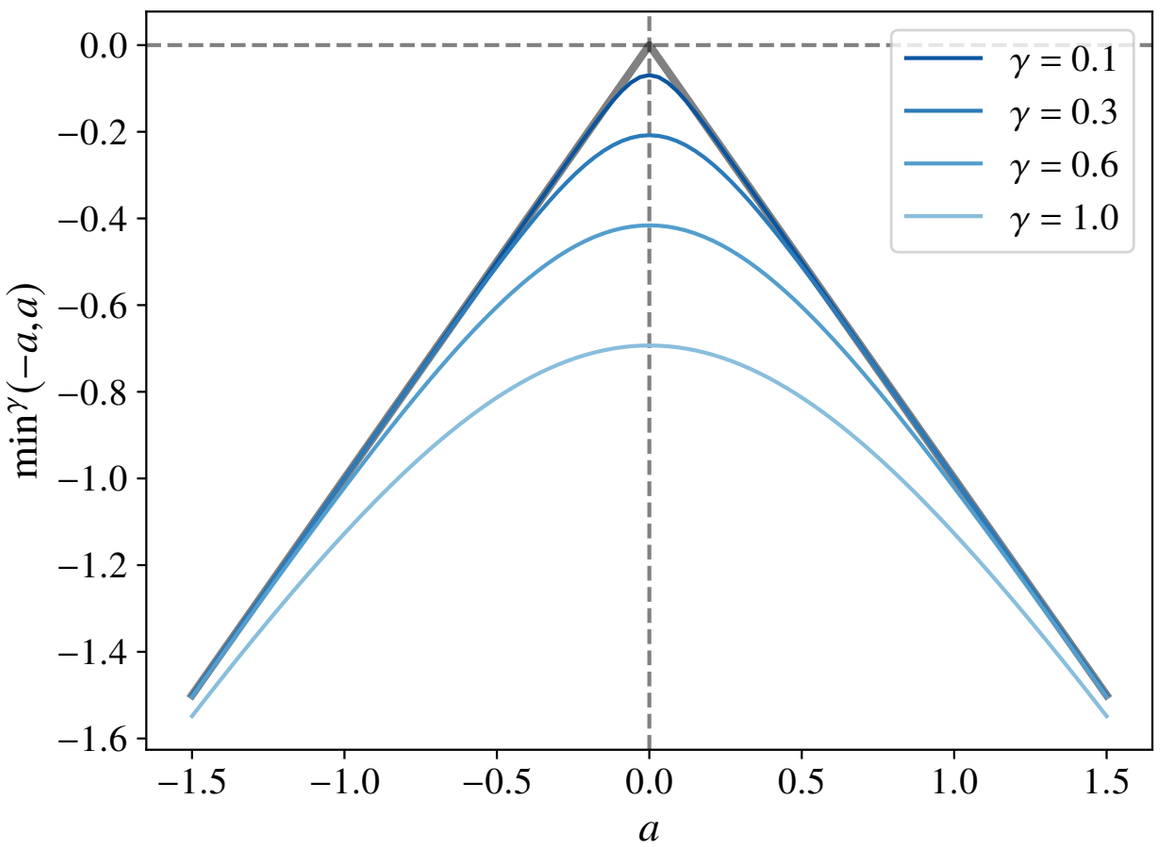
softDTW: a differentiable surrogate for DTW

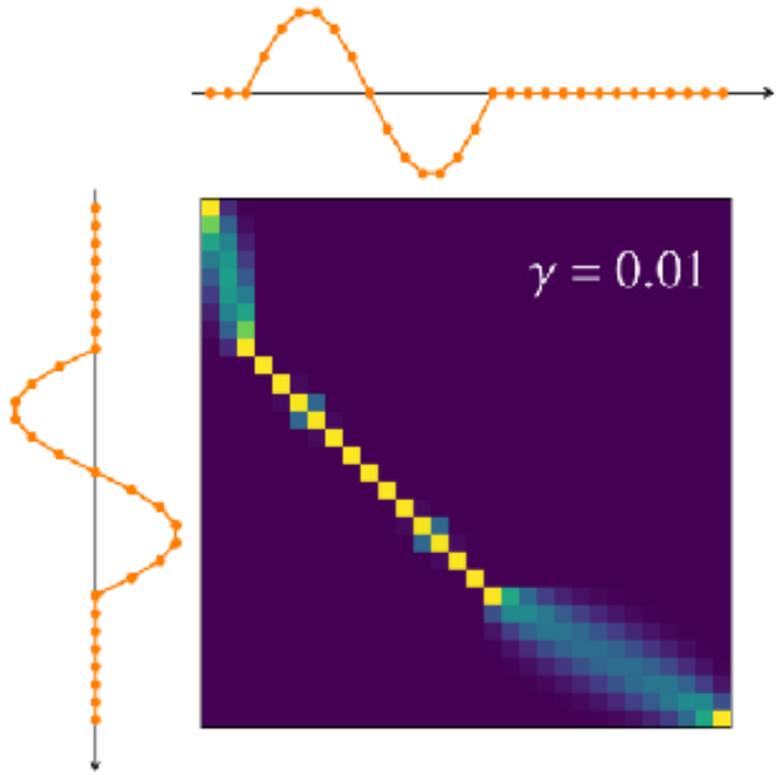
M. Cuturi

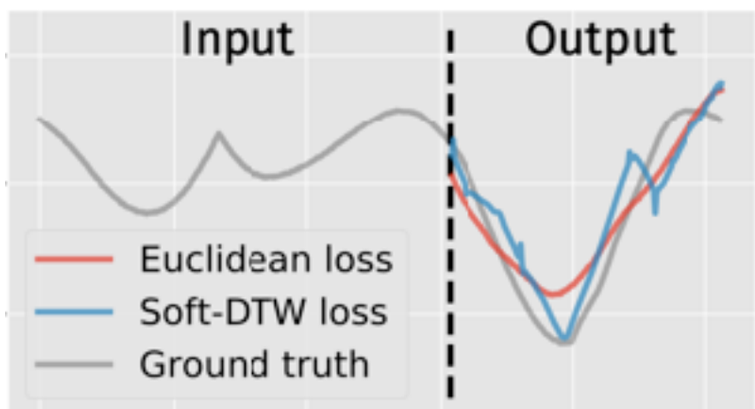
M. Blondel

- Soft-min instead of min in DTW formulation:

$$\min_{\pi}{}^{\gamma} \sum_{i,j} d_{i,j}\pi_{i,j}$$

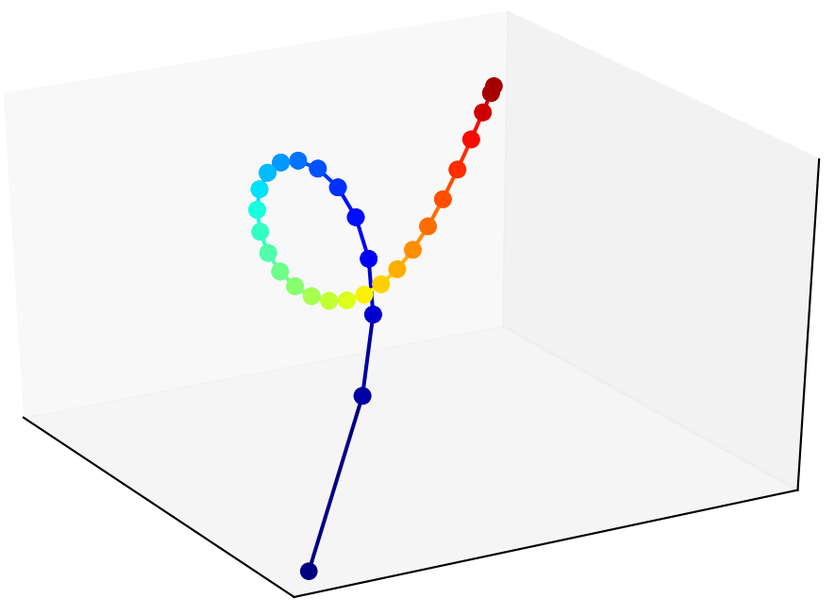


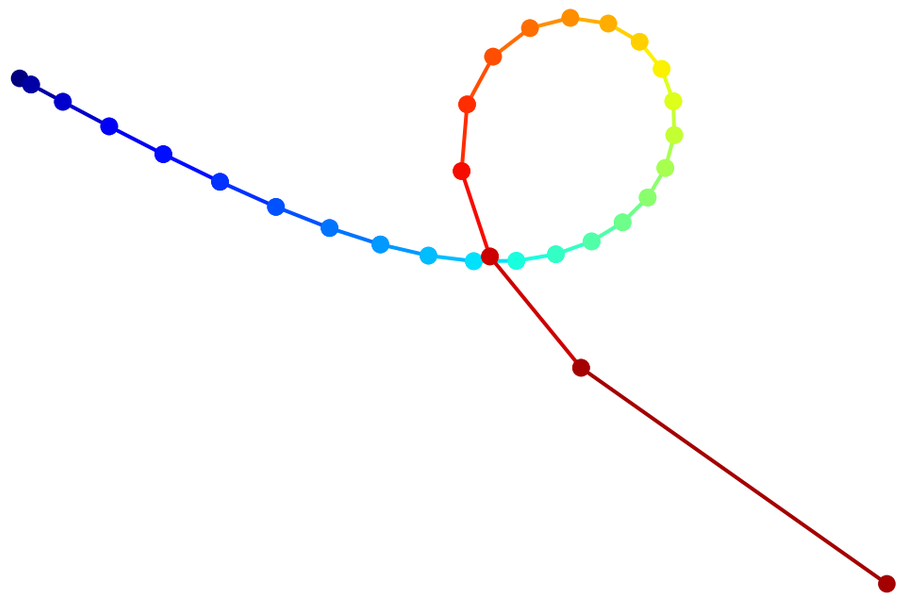




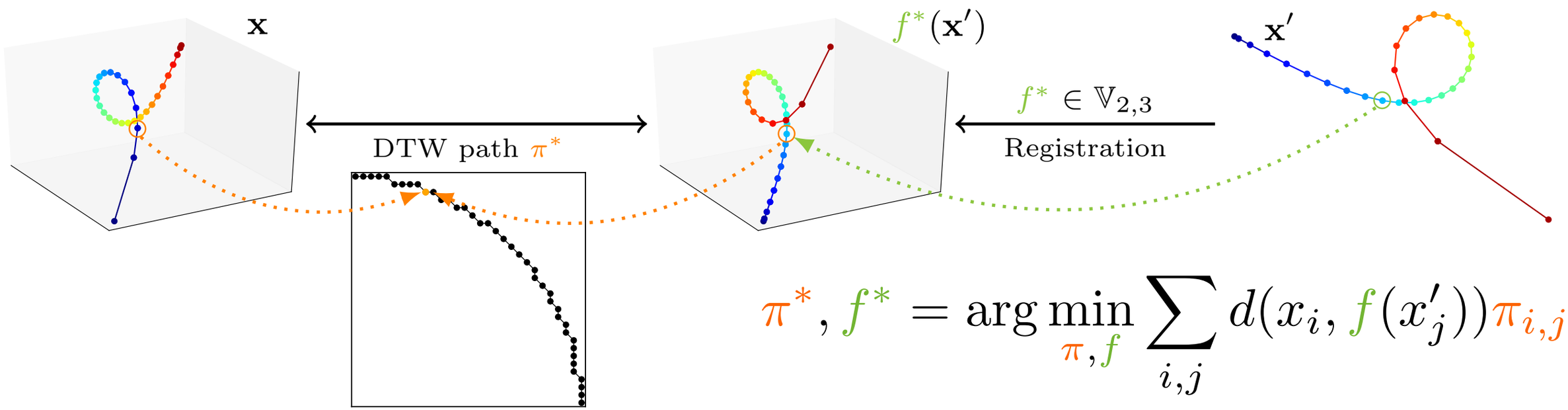
Matching Time Series $\mathbf{x} = (\bullet, \cdots, \bullet)$

DTW with Global Invariances [Vayer *et al*, ArXiv'20]





Matching Time Series $\mathbf{x} = (\bullet, \cdots, \bullet)$
DTW with Global Invariances [Vayer *et al*, ArXiv'20]

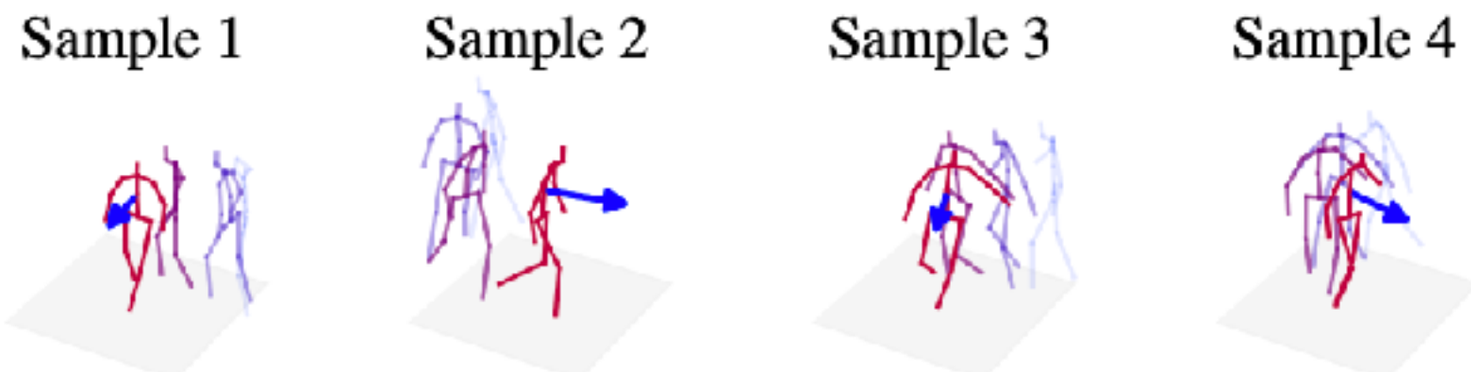


$$\pi^*, f^* = \arg\min_{\pi, f} \sum_{i,j} d(x_i, f(x'_j))\pi_{i,j}$$

- DTW-GI
  - Alternate between
    - DTW (for a fixed registration)
    - Procrustes (for a fixed temporal alignment)
- softDTW-GI
  - Gradient-descent on manifold (Stiefel constraints)
  - Parametrized family of registrations (eg. a neural network)

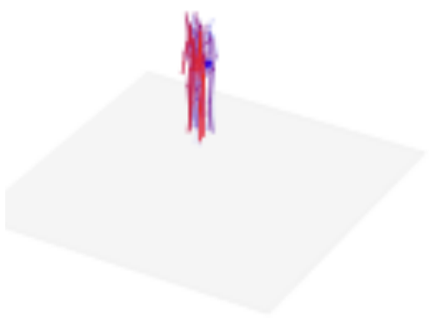
Matching Time Series $\mathbf{x} = (\bullet, \cdots, \bullet)$
DTW with Global Invariances [Vayer *et al*, ArXiv'20]

- Application to MoCap data forecasting

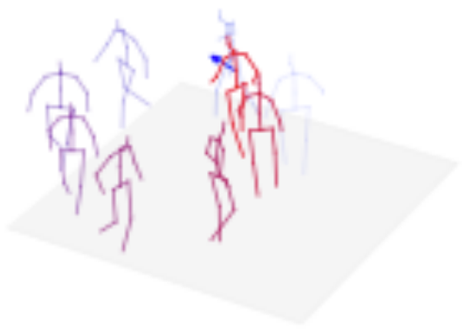


- Results

Ours



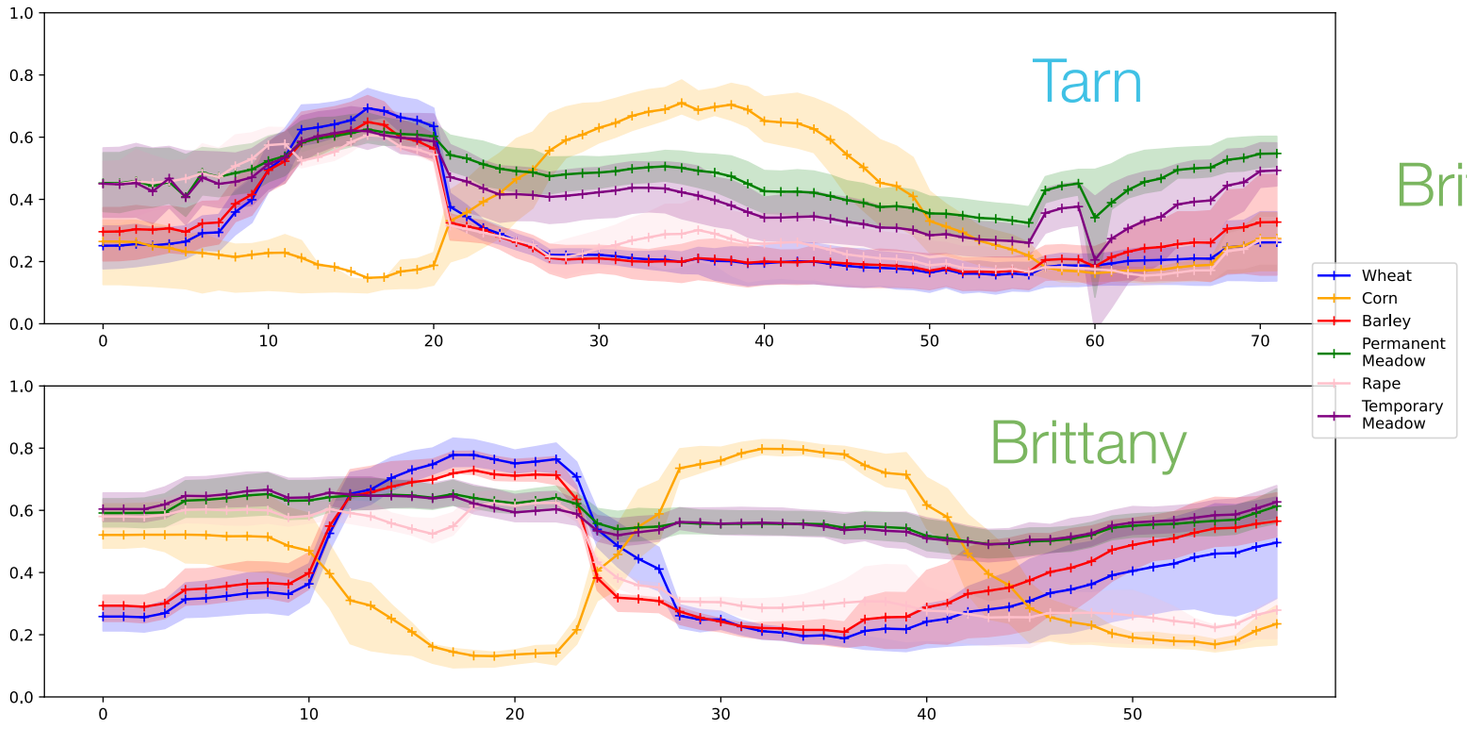
Baseline without explicit map (CTW)

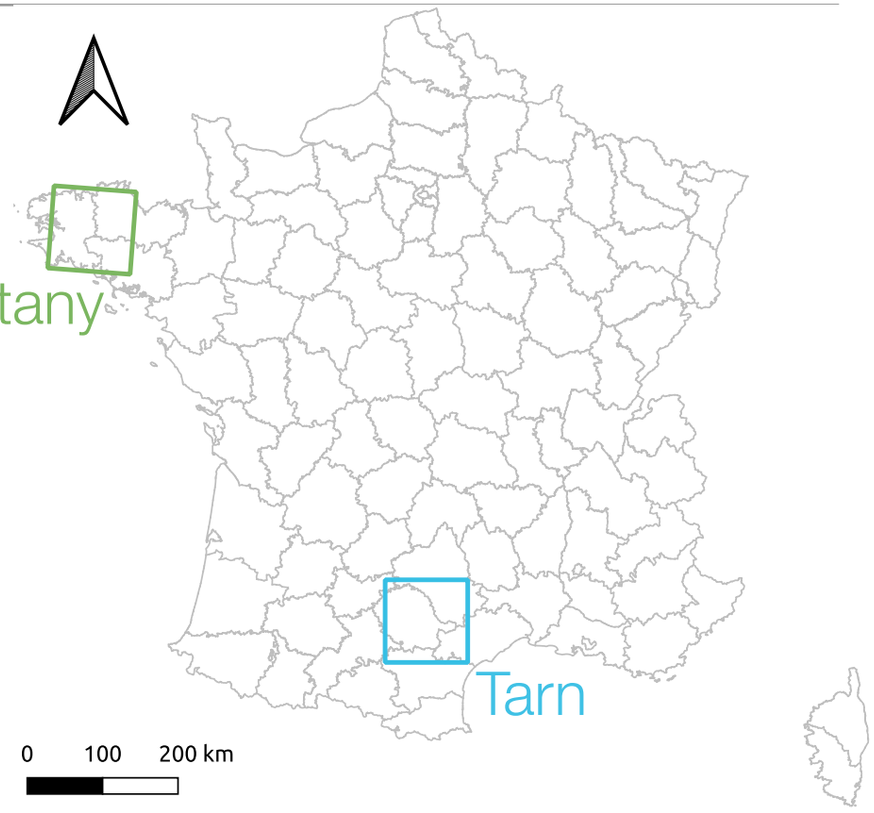


Baseline without time warping (L2+Proc.)

Method	Average test error
L2	183.11 +/- 3.90
softDTW [24]	183.12 +/- 3.90
CTW [16]	183.11 +/- 3.90
GromovDTW [20]	181.28 +/- 3.71
L2+Procrustes	46.33 +/- 0.21
softDTW+Procrustes	43.16 +/- 0.06
softDTW-GI (ours)	**39.58 +/- 0.34**

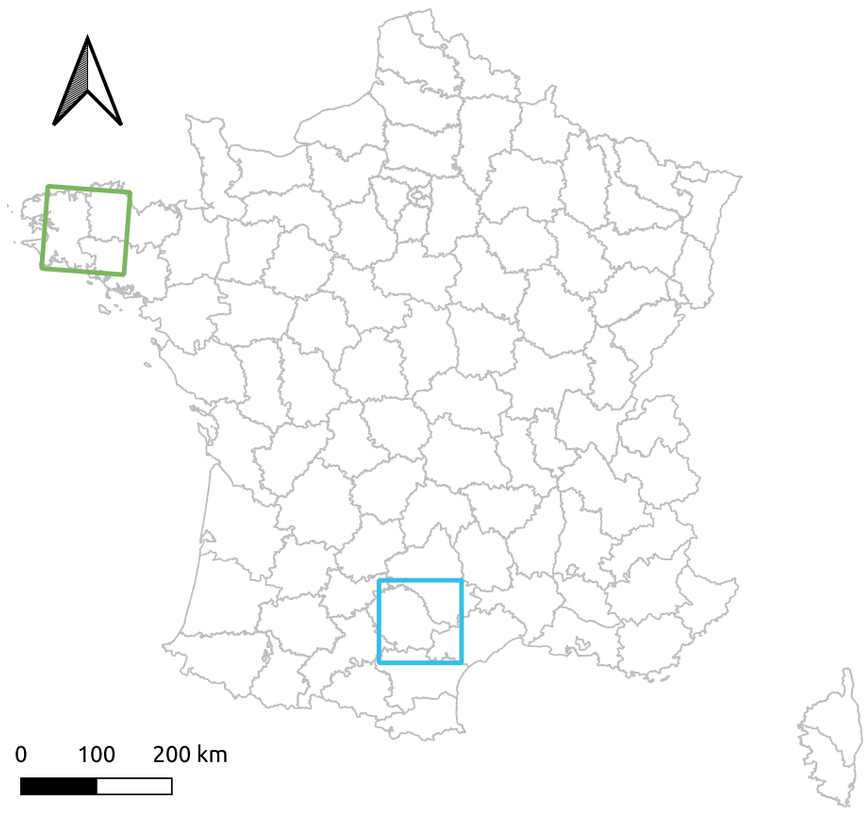
Matching sets of Time Series
Domain Adaptation in Time & Space (ongoing)

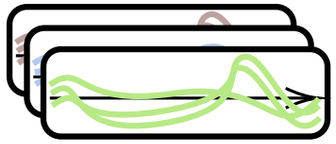


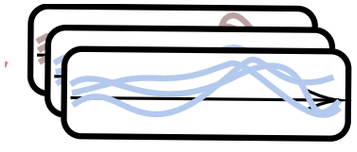


- Remote Sensing context
  - Lots of data but annotations are costly
  - Domain Adaptation (DA) is key

Matching sets of Time Series
Domain Adaptation in Time & Space (ongoing)





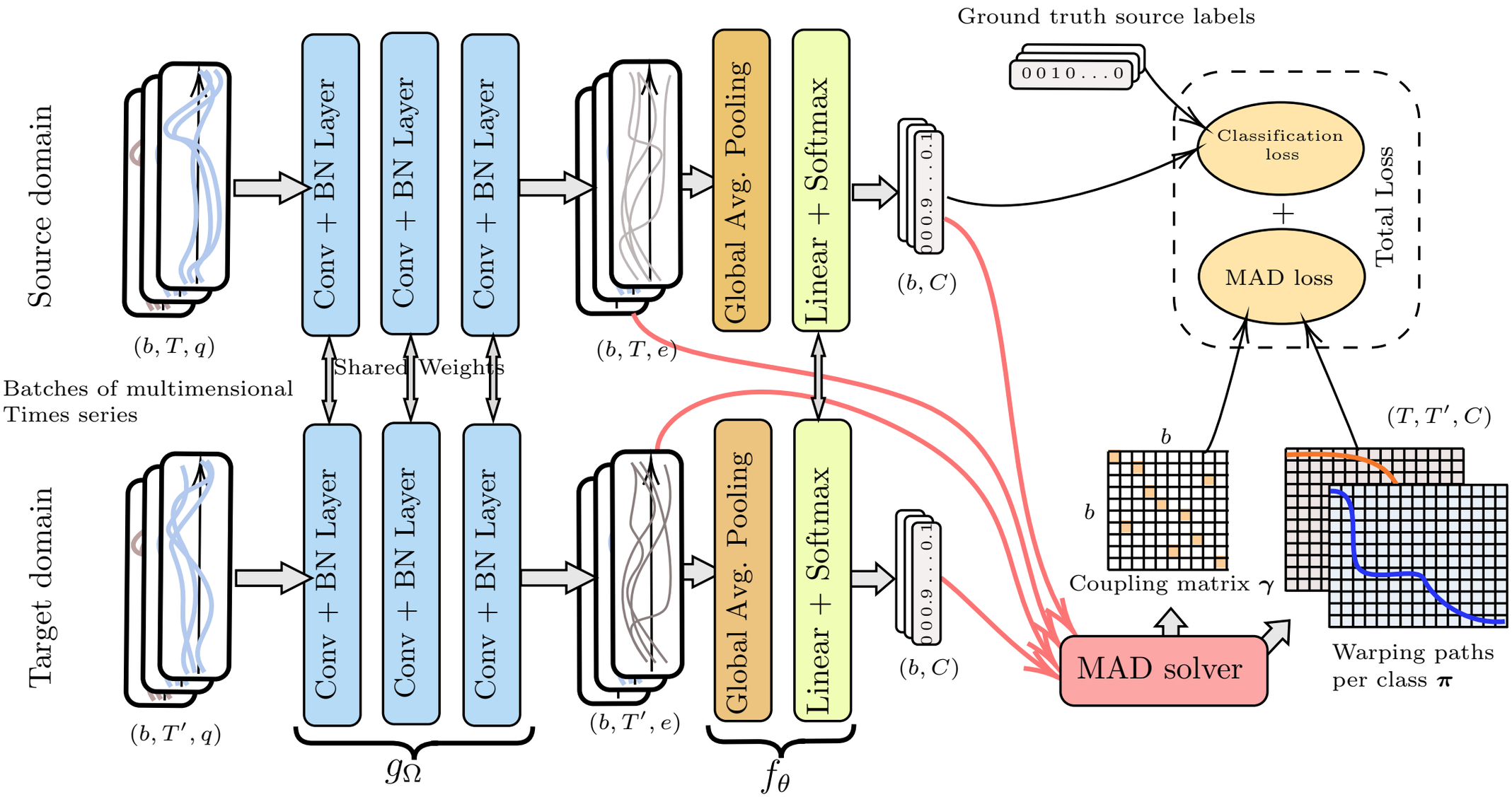


Dynamic Time Warping path

$$\min_{\pi^{(s)},\pi^{(t)}} \sum_{i,i'} \sum_{t,t'} d(X_{i,t}, Y_{i',t'}) \pi^{(s)}_{i,i'} \pi^{(t)}_{t,t'}$$

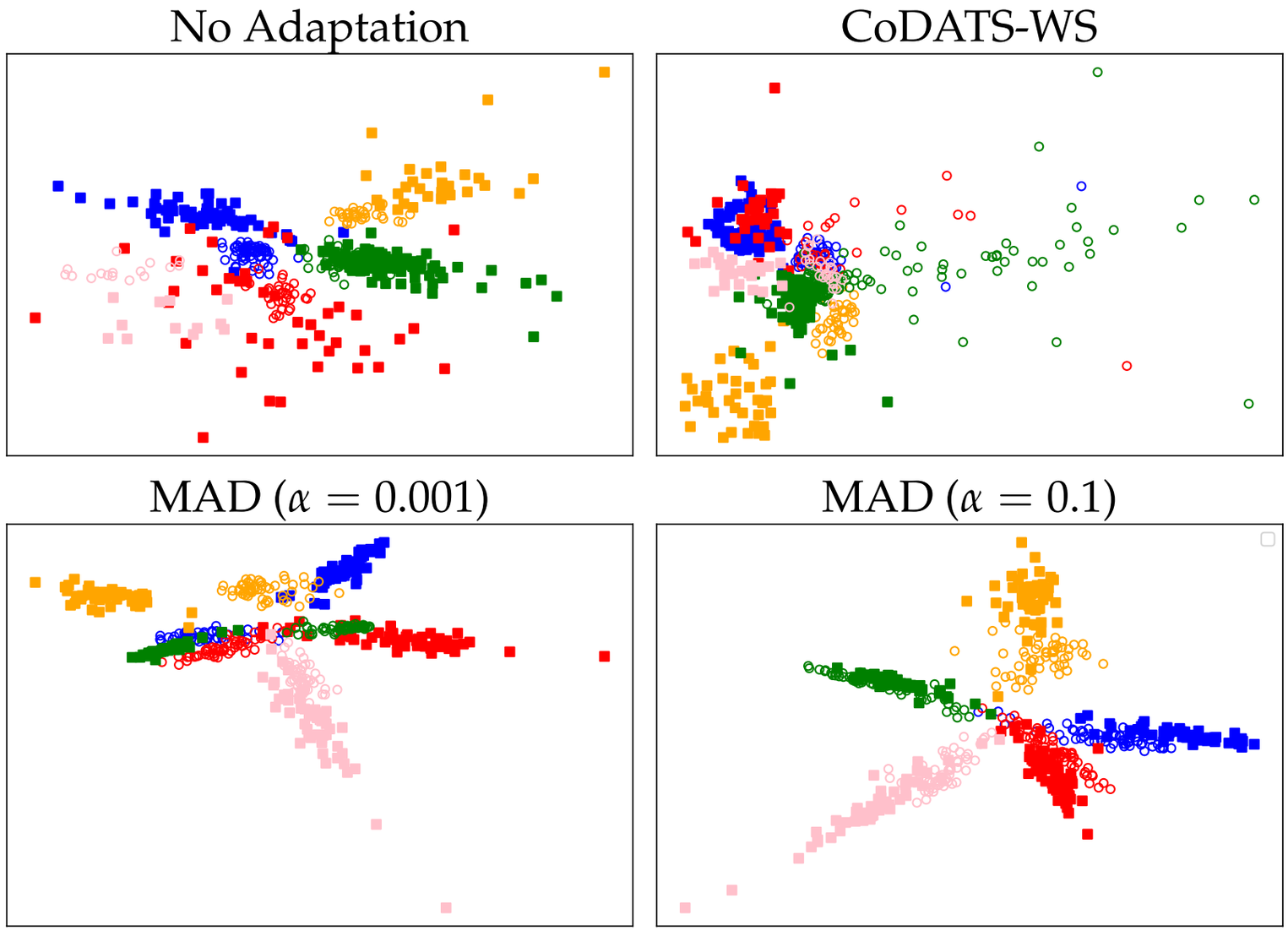
Optimal Transport plan

Matching sets of Time Series
Domain Adaptation in Time & Space (ongoing)



Overall architecture

Matching sets of Time Series
Domain Adaptation in Time & Space (ongoing)



Latent space visualization
(MDS)

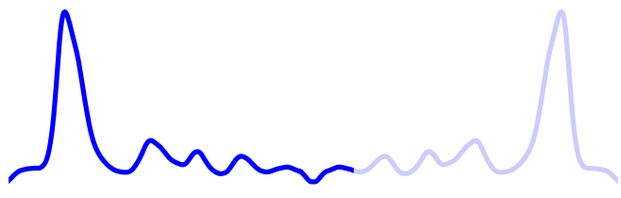
Matching sets of Time Series
Domain Adaptation in Time & Space (ongoing)

Problem	No adaptation	CoDATS-WS	MAD	Target only
DK1 → FR1	65.2 ± 2.1	66.2 ± 6.8	$\mathbf{75.7 \pm 1.6}$	93.3 ± 0.2
DK1 → FR2	65.4 ± 1.4	70.1 ± 2.0	$\mathbf{71.0 \pm 0.7}$	93.2 ± 0.2
DK1 → AT1	76.5 ± 2.7	78.1 ± 4.4	$\mathbf{81.0 \pm 3.1}$	96.4 ± 0.2
FR1 → DK1	56.0 ± 1.3	70.5 ± 9.3	$\mathbf{71.7 \pm 2.8}$	98.4 ± 0.1
FR1 → FR2	57.8 ± 8.5	75.7 ± 4.0	$\mathbf{79.1 \pm 2.0}$	93.2 ± 0.2
Tarn → Brittany	88.45 ± 3.65	96.0 ± 1.6	$\mathbf{98.87 \pm 0.44}$	99.68 ± 0.05
Brittany → Tarn	48.91 ± 0.63	$\mathbf{93.6 \pm 0.1}$	90.63 ± 1.11	98.64 ± 0.18

Table 3. Classification accuracies on miniTimeMatch and TarnBZH datasets.

Conclusion

- Structured data is a great playground!
  - Matching-based methods have a role to play
- Links between OT and DTW
  - Still a lot to be done



- Software (Python libraries)
  - tslearn for time series
  - POT for optimal transport
- Find more (incl. code) at
  - rtavenar.github.io/hdr/ (my work)
  - rtavenar.github.io/blog/ (intro. to DTW and variants)