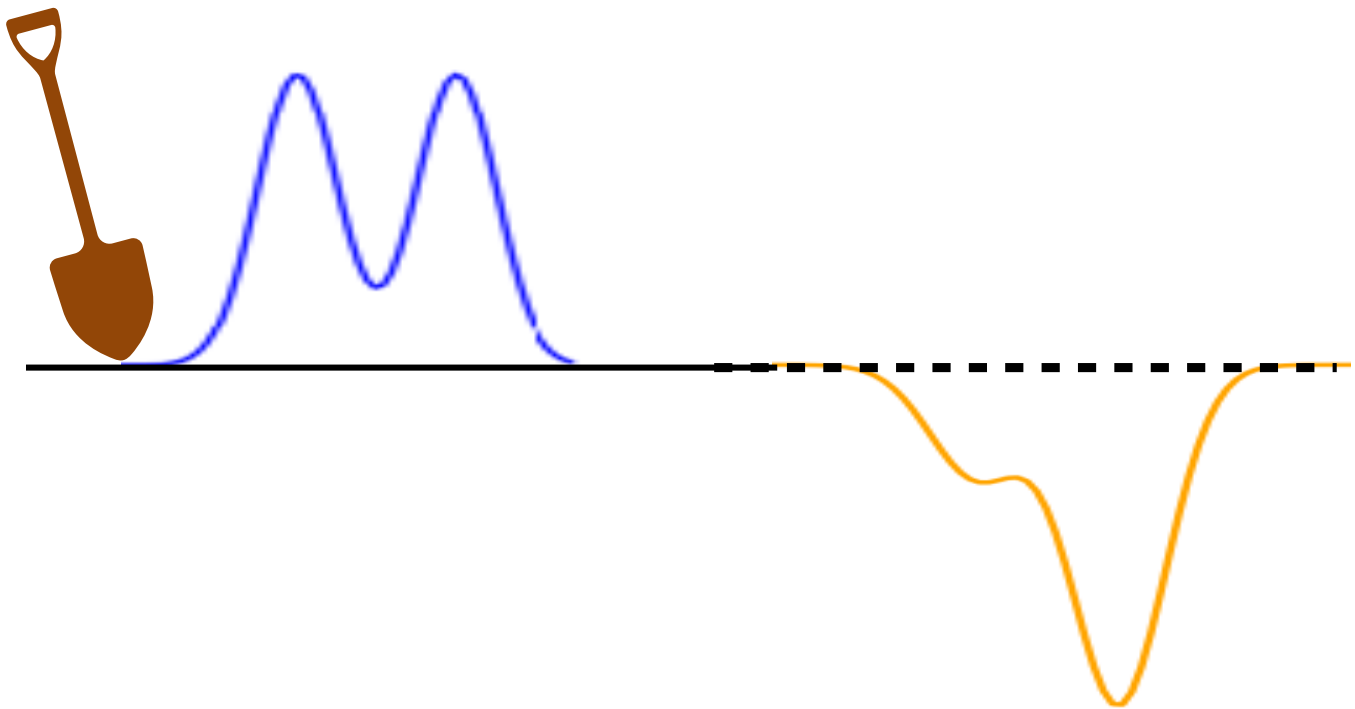
Introduction to Optimal Transport for Machine Learning

Aude Genevay

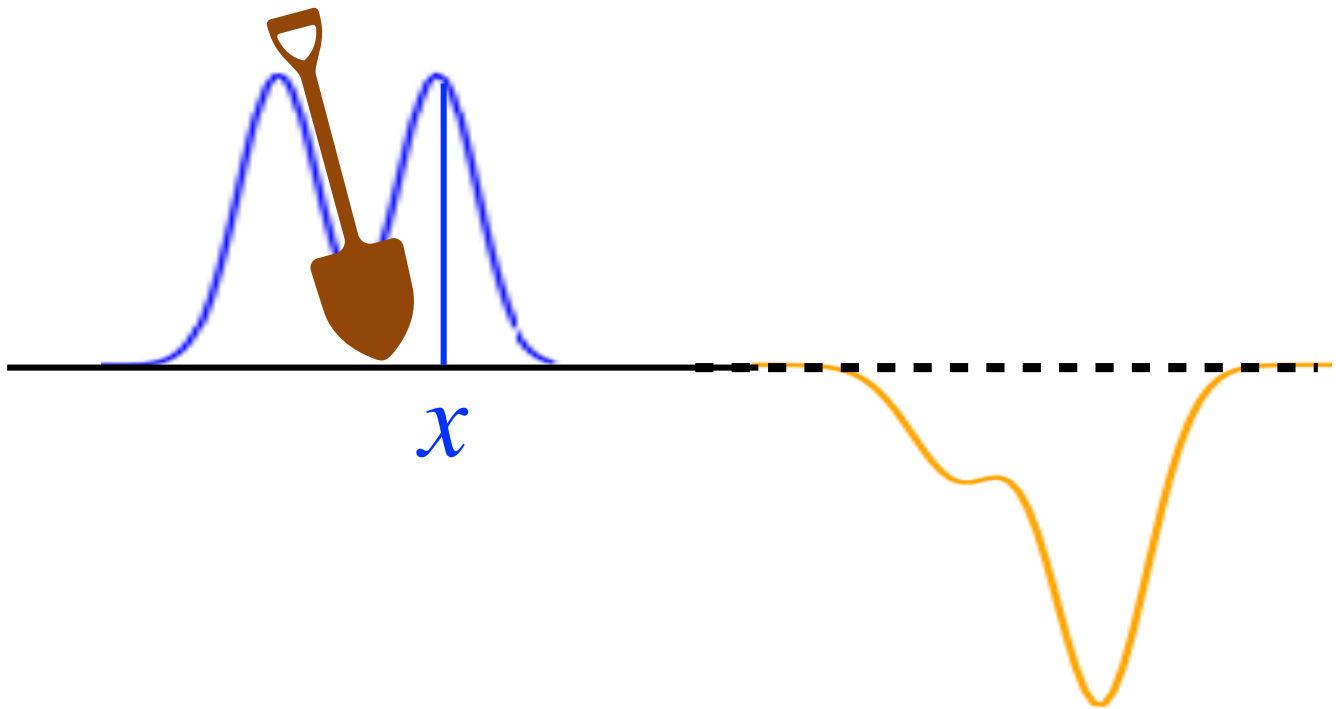
MIT CSAIL

Einstein ML Reading Group - Feb. 2020

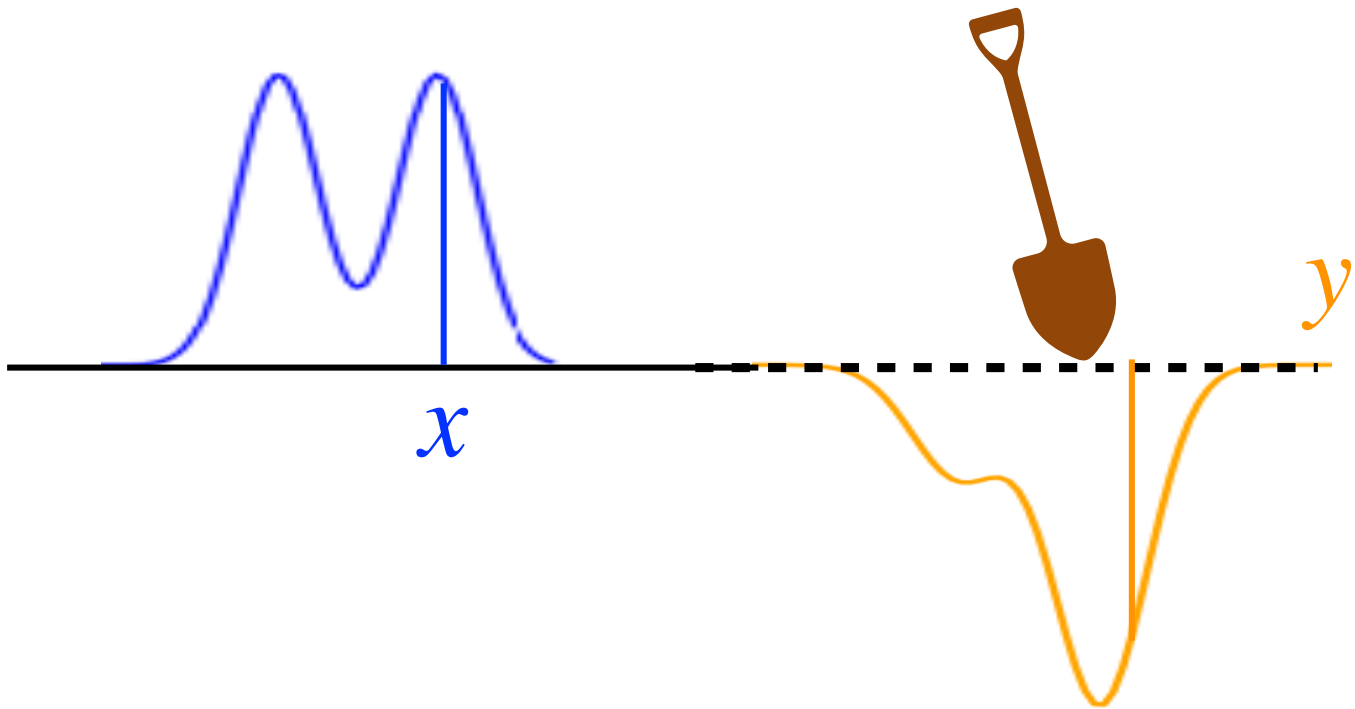
Optimal Transport (Monge 1781)



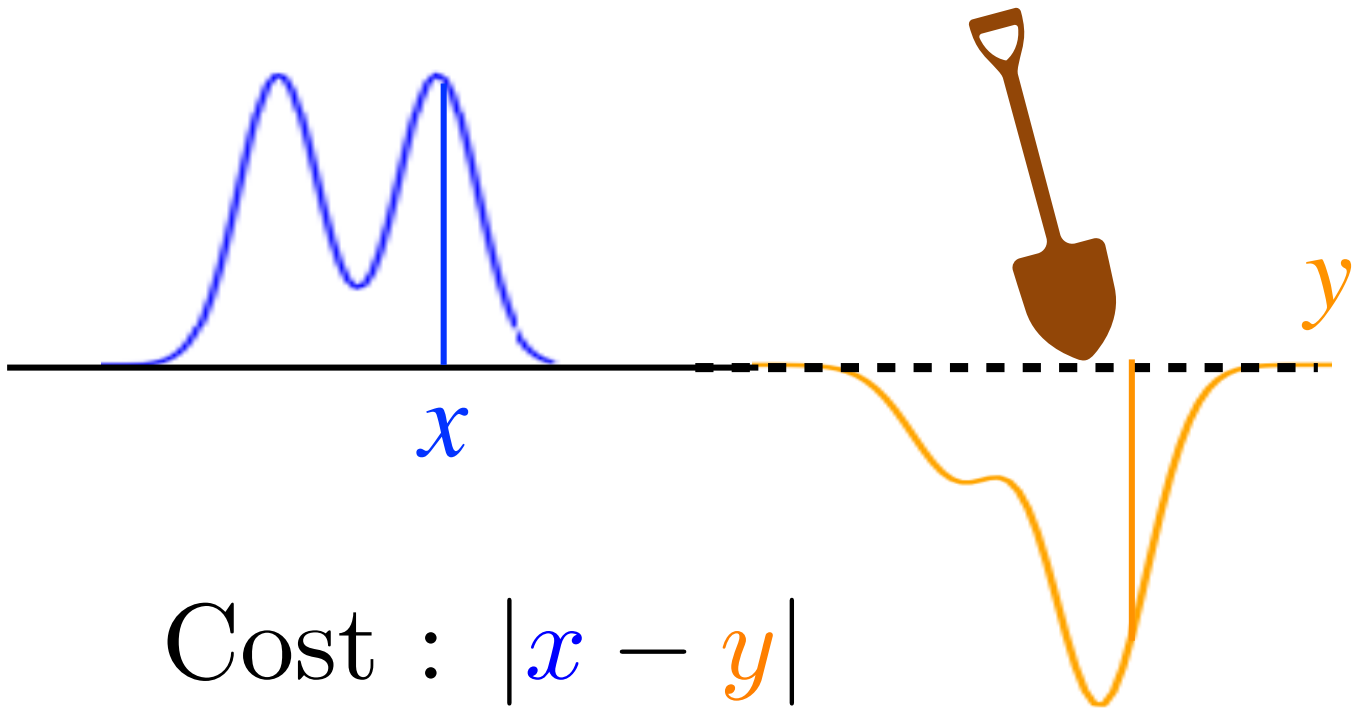
Optimal Transport (Monge 1781)



Optimal Transport (Monge 1781)

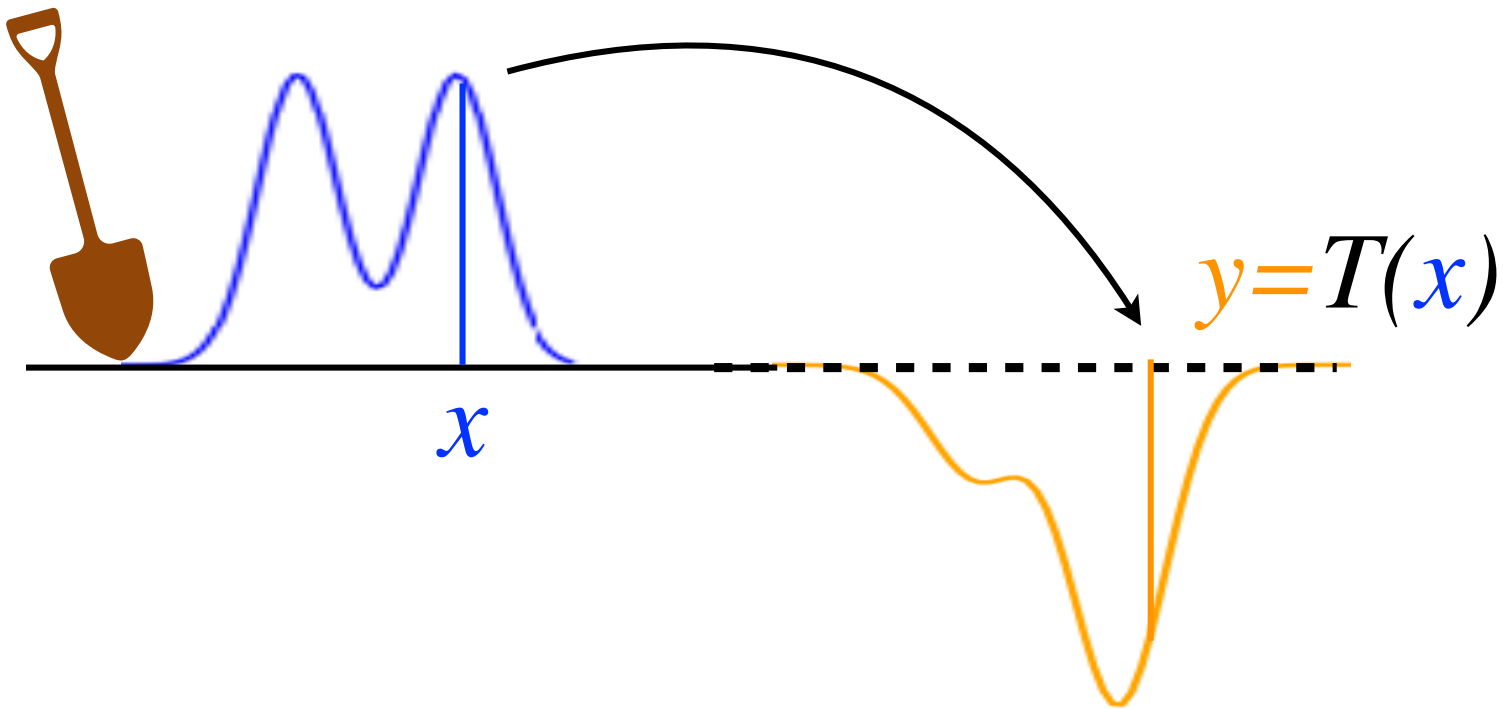


Optimal Transport (Monge 1781)

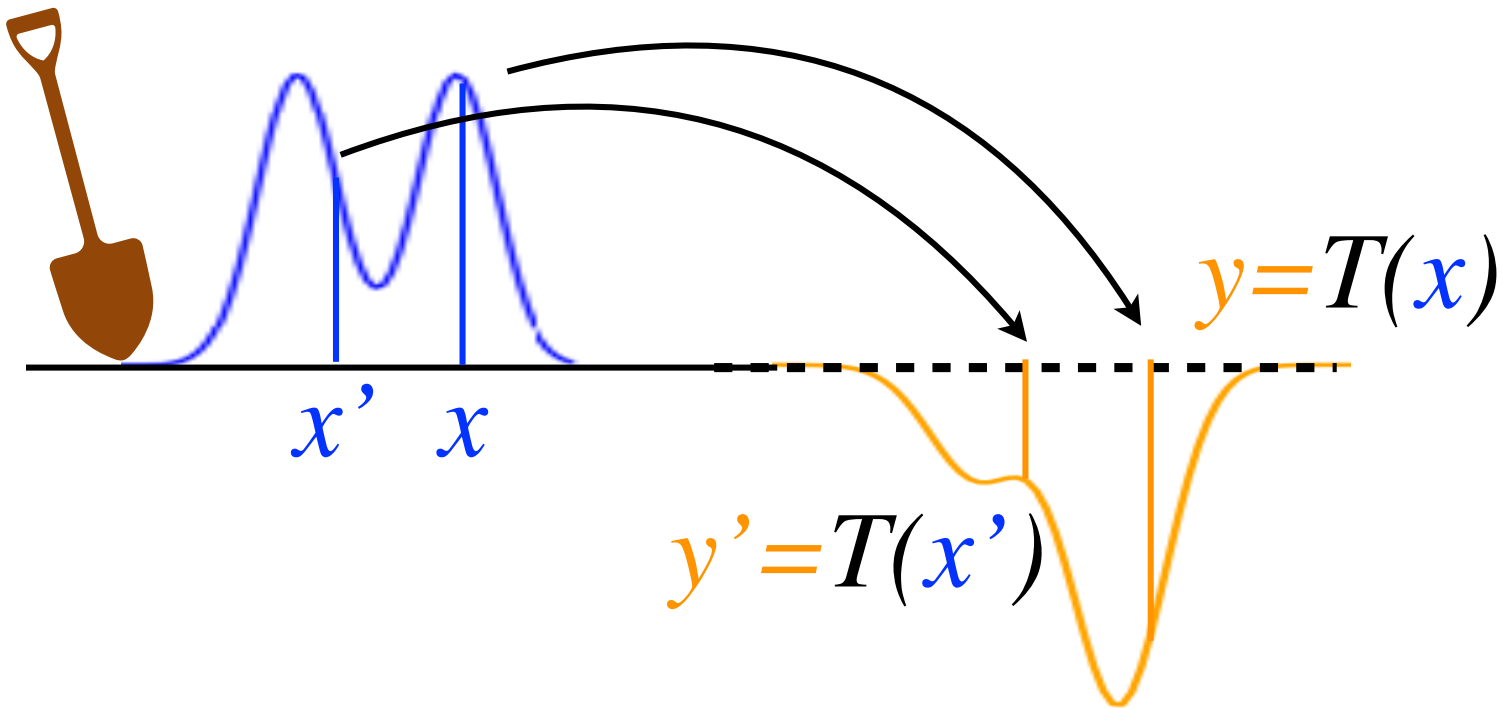


$$\text{Cost} : |x - y|$$

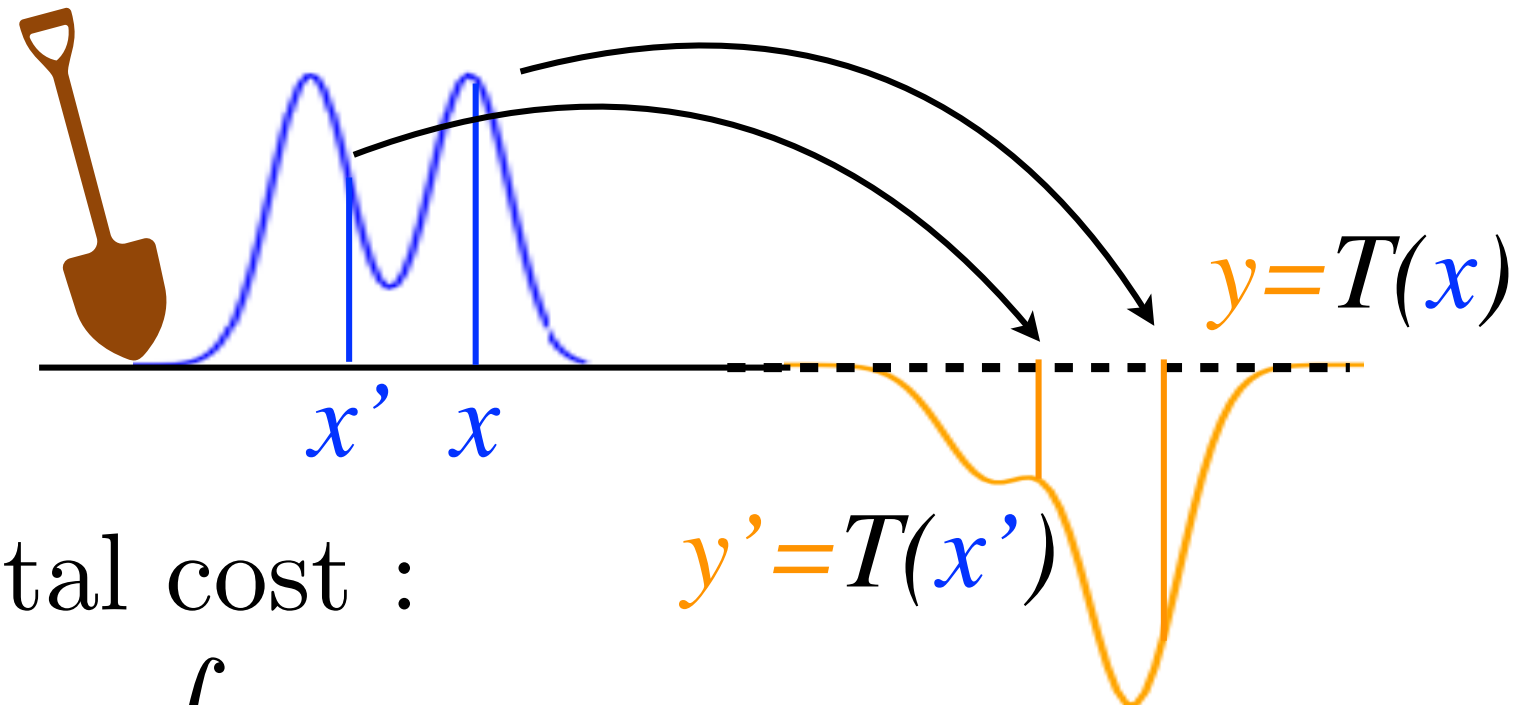
Optimal Transport (Monge 1781)



Optimal Transport (Monge 1781)



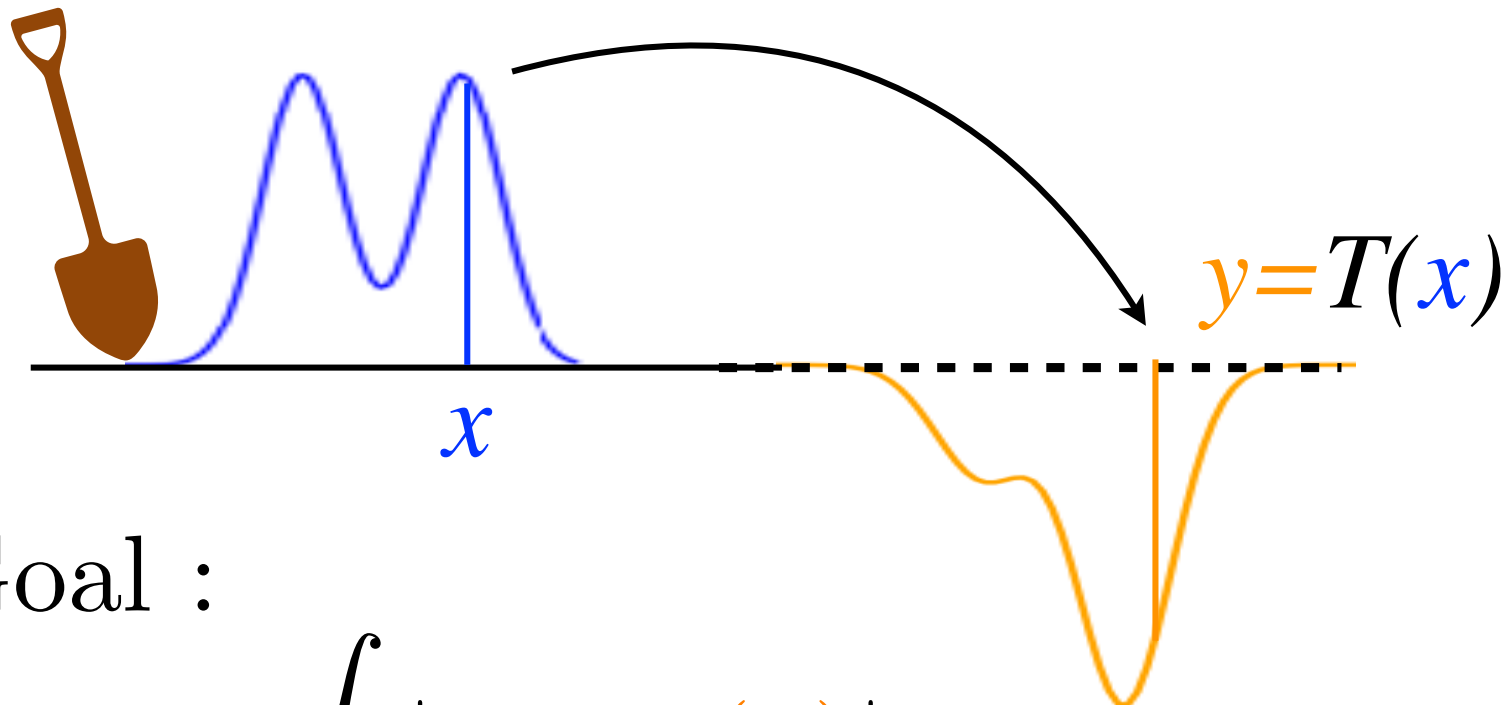
Optimal Transport (Monge 1781)



Total cost :

$$\int |x - T(x)| dx$$

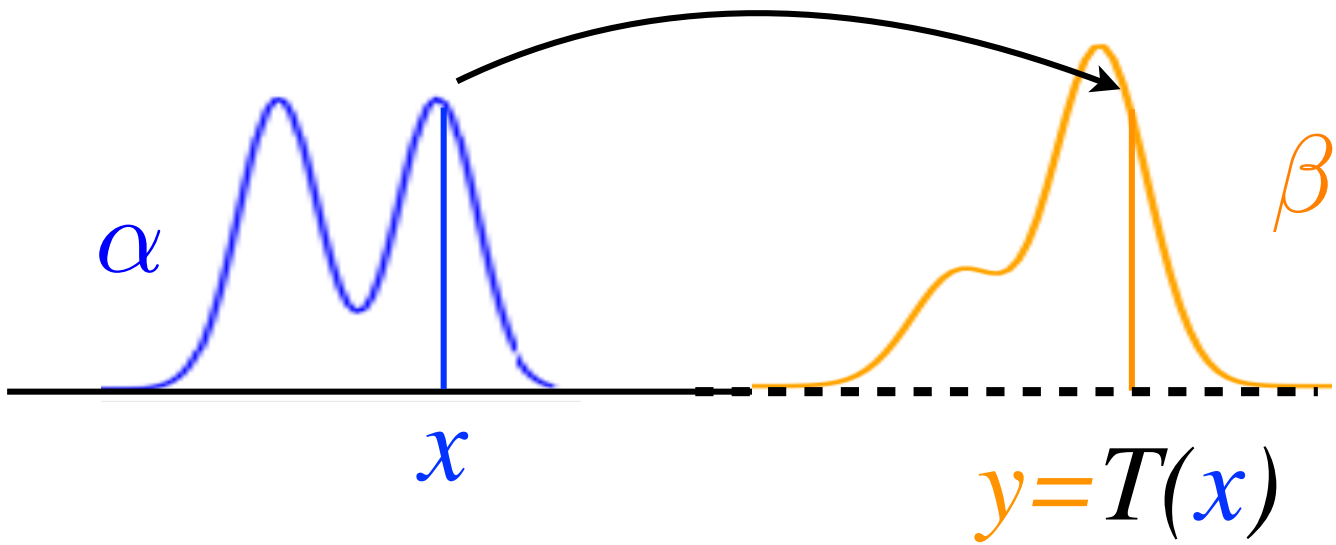
Optimal Transport (Monge 1781)



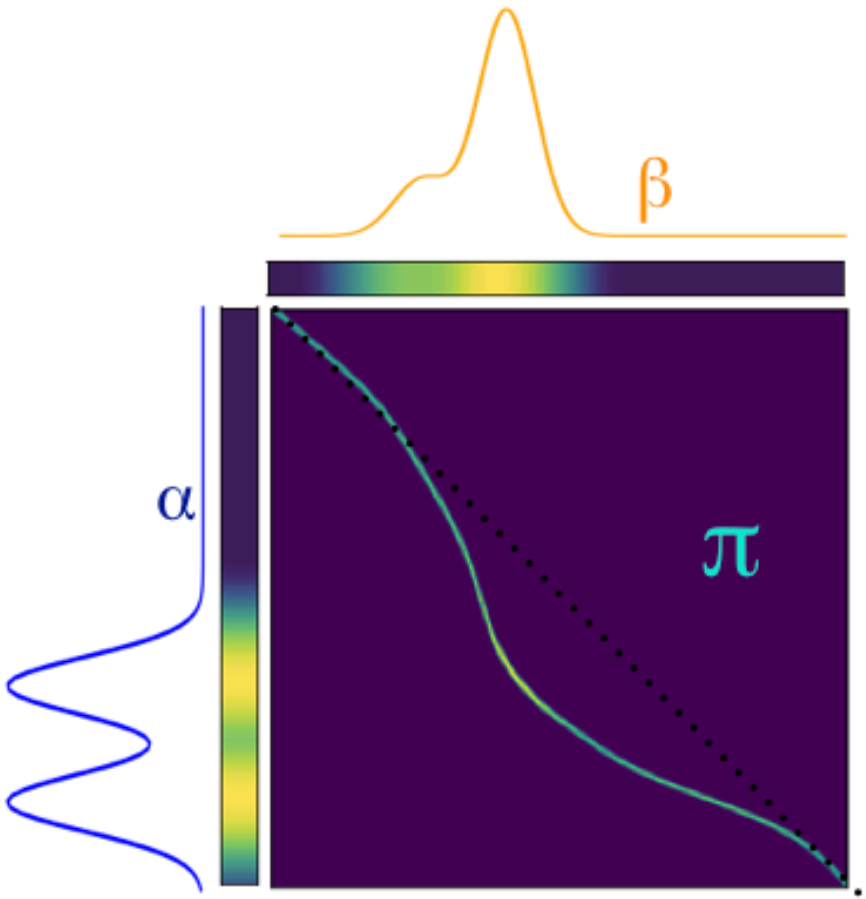
Goal :

$$\min_T \int |x - T(x)| dx$$

Optimal Transport (Monge 1781



Kantorovich Formulation ('42)



Kantorovich Formulation ('42)

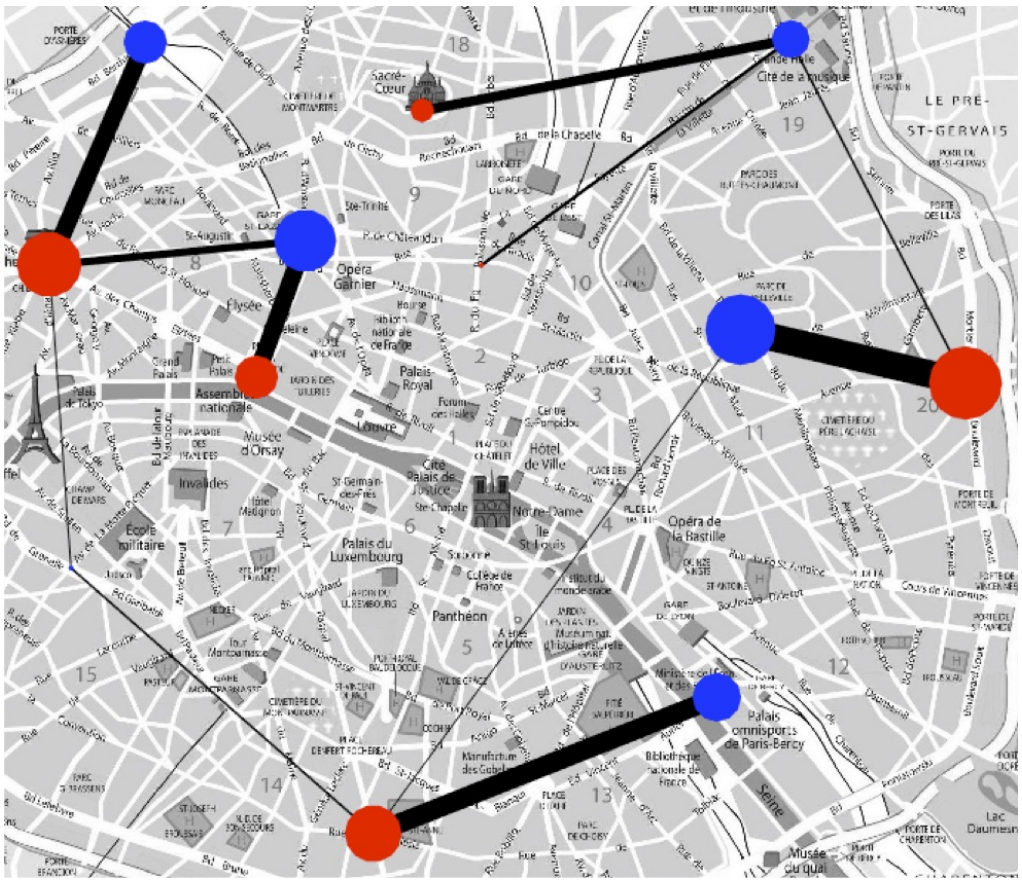


Image from G. Peyré : the bakery supply problem in Paris

Kantorovich Formulation ('42)

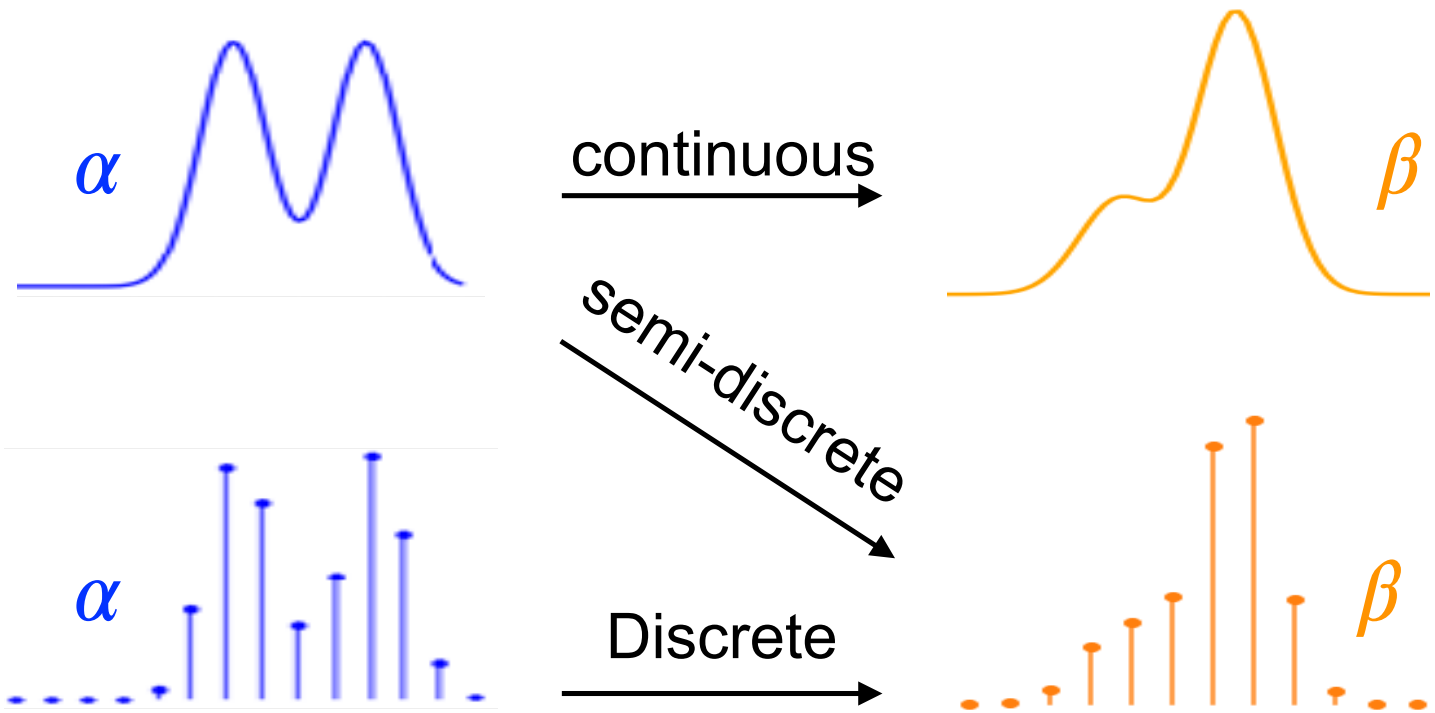
Given c cost of moving one unit of mass from x to y,
Minimal cost of moving ALL the mass from α to β ?

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

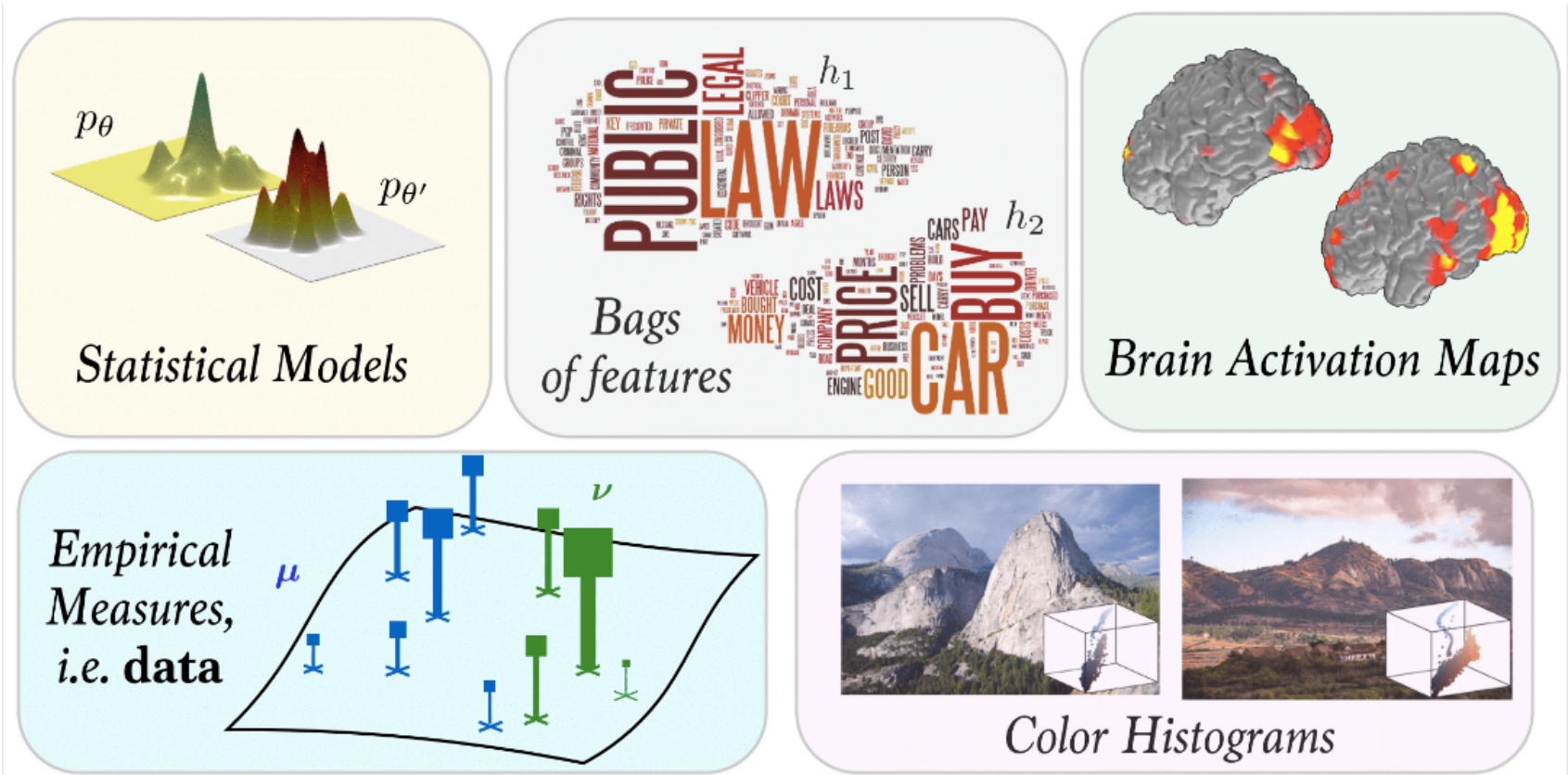
$$W_c(\alpha, \beta) = \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) \qquad (\mathcal{P})$$

For $c(x, y) = \|x - y\|_2^p$, $W_c(\alpha, \beta)^{1/p}$ is the **p-Wasserstein distance**.

Comparing Probability Measures

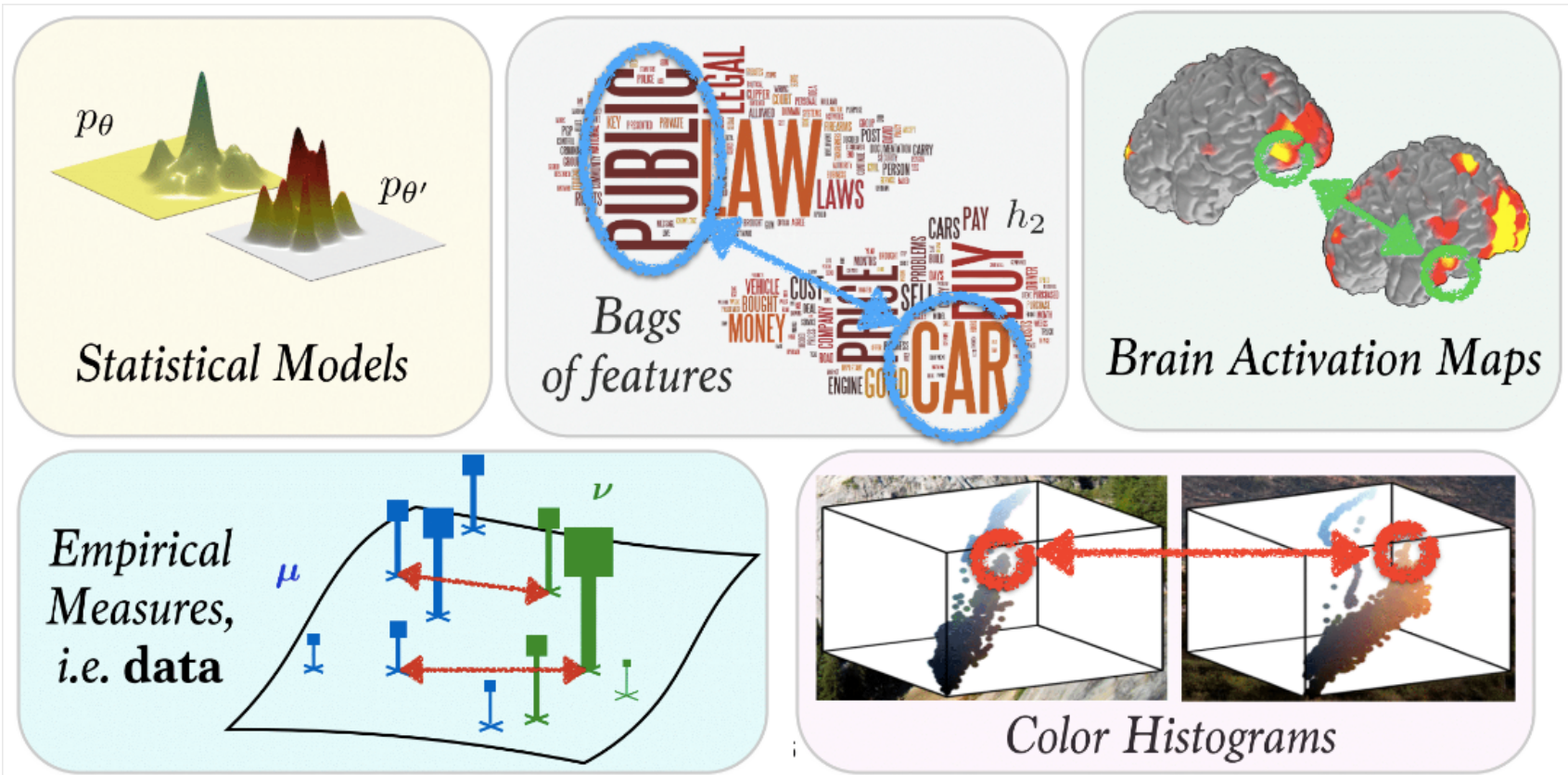


Probability Measures in Data Science



courtesy of Marco Cuturi

Probability Measures in Data Science



courtesy of Marco Cuturi

Entropic Regularization (Cuturi '13)

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$W_c\ (\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \qquad (\mathcal{P})$$

Entropic Regularization (Cuturi '13)

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$W_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon H(\pi|\alpha\otimes\beta), \quad (\mathcal{P}_\varepsilon)$$

where

$$H(\pi|\alpha\otimes\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\pi(x,y).$$

relative entropy of the transport plan π with respect to the product measure $\alpha\otimes\beta$.

Entropic Regularization

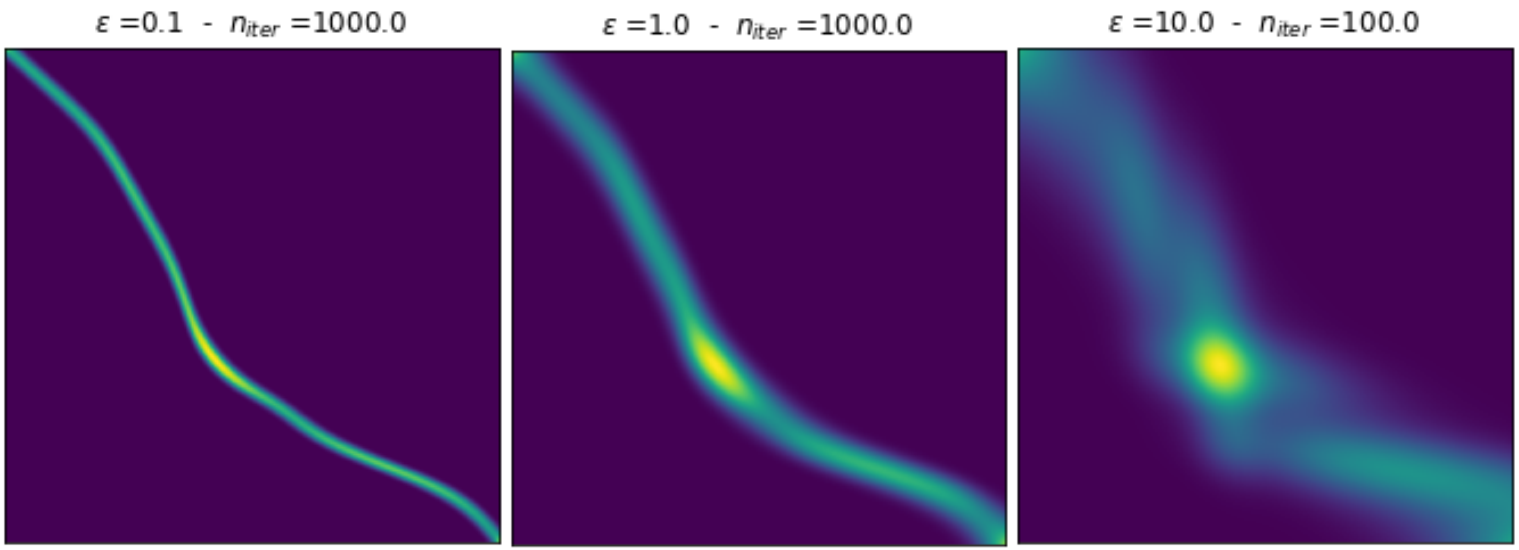


Figure 1 – Influence of the regularization parameter ε on the transport plan π.

Intuition : the entropic penalty 'smoothes' the problem and avoids over fitting (think of ridge regression for least squares)

Dual Formulation

Contrary to standard OT, no constraint on the dual problem :

$$W_c\ (\alpha, \beta) = \max_{\substack{u \in \mathcal{C}(\mathcal{X}) \\ v \in \mathcal{C}(\mathcal{Y})}} \int_{\mathcal{X}} u(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y) \qquad (\mathcal{D})$$

$$\text{such that} \quad \{u(x) + v(y) \leqslant c(x, y)\ \forall\ (x, y) \in \mathcal{X} \times \mathcal{Y}\}$$

Dual Formulation

Contrary to standard OT, no constraint on the dual problem :

$$W_{c,\varepsilon}(\alpha,\beta) = \max_{\substack{u\in\mathcal{C}(\mathcal{X})\\ v\in\mathcal{C}(\mathcal{Y})}} \int_{\mathcal{X}} u(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon.$$

$$= \max_{\substack{u\in\mathcal{C}(\mathcal{X})\\ v\in\mathcal{C}(\mathcal{Y})}} \mathbb{E}_{\alpha\otimes\beta}\left[f_\varepsilon^{XY}(u,v)\right] + \varepsilon, \qquad (\mathcal{D}_\varepsilon)$$

with $f_\varepsilon^{xy}(u,v) \overset{\text{def.}}{=} u(x) + v(y) - \varepsilon e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}$

Sinkhorn's Algorithm

Iterative algorithm : alternate between optimizing over u with fixed v and optimizing over v with fixed u.

Sinkhorn's Algorithm

Iterative algorithm : alternate between optimizing over u with fixed v and optimizing over v with fixed u.

Sinkhorn's Algorithm

Let $\mathbf{K}_{ij} = e^{-\frac{c(x_i,y_j)}{\varepsilon}}, \mathbf{a} = e^{\frac{\mathbf{u}}{\varepsilon}}, \mathbf{b} = e^{\frac{\mathbf{v}}{\varepsilon}}$.

$$\mathbf{a}^{(\ell+1)} = \frac{1}{\mathbf{K}(\mathbf{b}^{(\ell)} \odot \boldsymbol{\beta})} \quad ; \quad \mathbf{b}^{(\ell+1)} = \frac{1}{\mathbf{K}^T(\mathbf{a}^{(\ell+1)} \odot \boldsymbol{\alpha})}$$

Complexity of each iteration : $O(n^2)$,
Linear convergence, constant degrades when $\varepsilon \to 0$.

Differentiable approximation of OT

Bonus : Sinkhorn procedure is fully differentiable with auto-diff tools (e.g TensorFlow) $\Rightarrow$ yields a differentiable approximation of OT !

Some applications :

- Differentiable sorting (Cuturi et al '19)
- Differentiable (or 'soft') assignments
- Differentiable clustering (G. et al '19)
- Learning with a regularized Wasserstein loss ($\rightarrow$ more on that later...)

Minimizing W_ε, $\varepsilon = 1$

$$\min_{(x_1,\dots,x_k)} W_\varepsilon\left(\frac{1}{k}\sum_{i=1}^{k}\delta x_i, \frac{1}{n}\sum_{i=1}^{n}\delta y_j\right)$$

Sinkhorn Divergences

Issue of regularized Wass. Distance : $W_{c,\varepsilon}(\alpha, \alpha) \neq 0$
Proposed Solution : introduce corrective terms to 'debias' regularized Wasserstein distance

Definition (Sinkhorn Divergences)

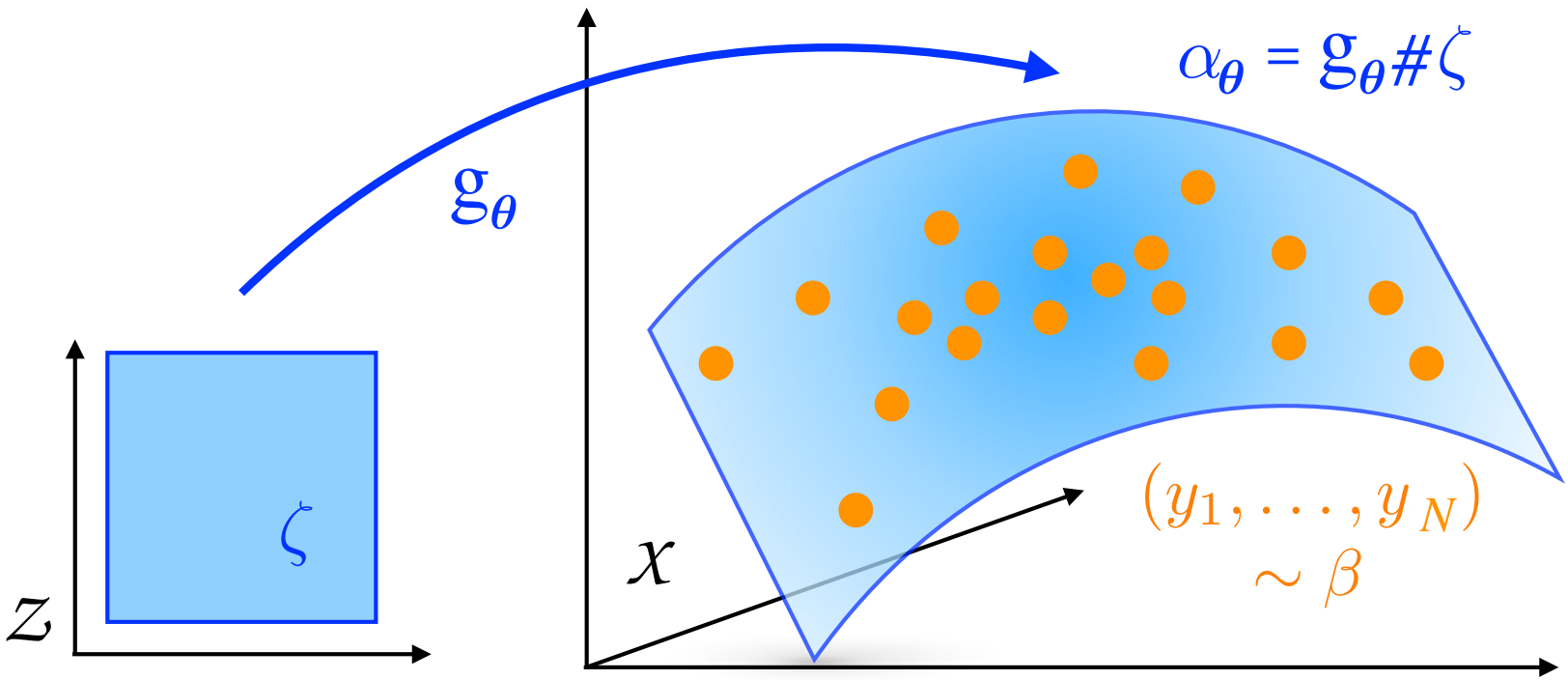
Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$SD_{c,\varepsilon}(\alpha, \beta) \stackrel{\text{def.}}{=} W_{c,\varepsilon}(\alpha, \beta) - \frac{1}{2} W_{c,\varepsilon}(\alpha, \alpha) - \frac{1}{2} W_{c,\varepsilon}(\beta, \beta),$$

Minimizing SD_ε

$$\min_{(x_1,\dots,x_k)} SD_\varepsilon\left(\frac{1}{k}\sum_{i=1}^{k}\delta x_i, \frac{1}{n}\sum_{i=1}^{n}\delta y_j\right)$$

Generative Models



Problem Formulation

- β the **unknown** measure of the data :
  finite number of samples $(y_1, \dots, y_N) \sim \beta$
- α_θ the parametric model of the form $\alpha_\theta \overset{\text{def.}}{=} g_{\theta\#}\zeta$:
  to sample $x \sim \alpha_\theta$, draw $z \sim \zeta$ and take $x = g_\theta(z)$.

We want to solve by gradient descent

$$\min_\theta SD_{c,\varepsilon}(\alpha_\theta, \beta)$$

NB : α_θ and β are only known via their samples.

Computing the Gradient in Practice

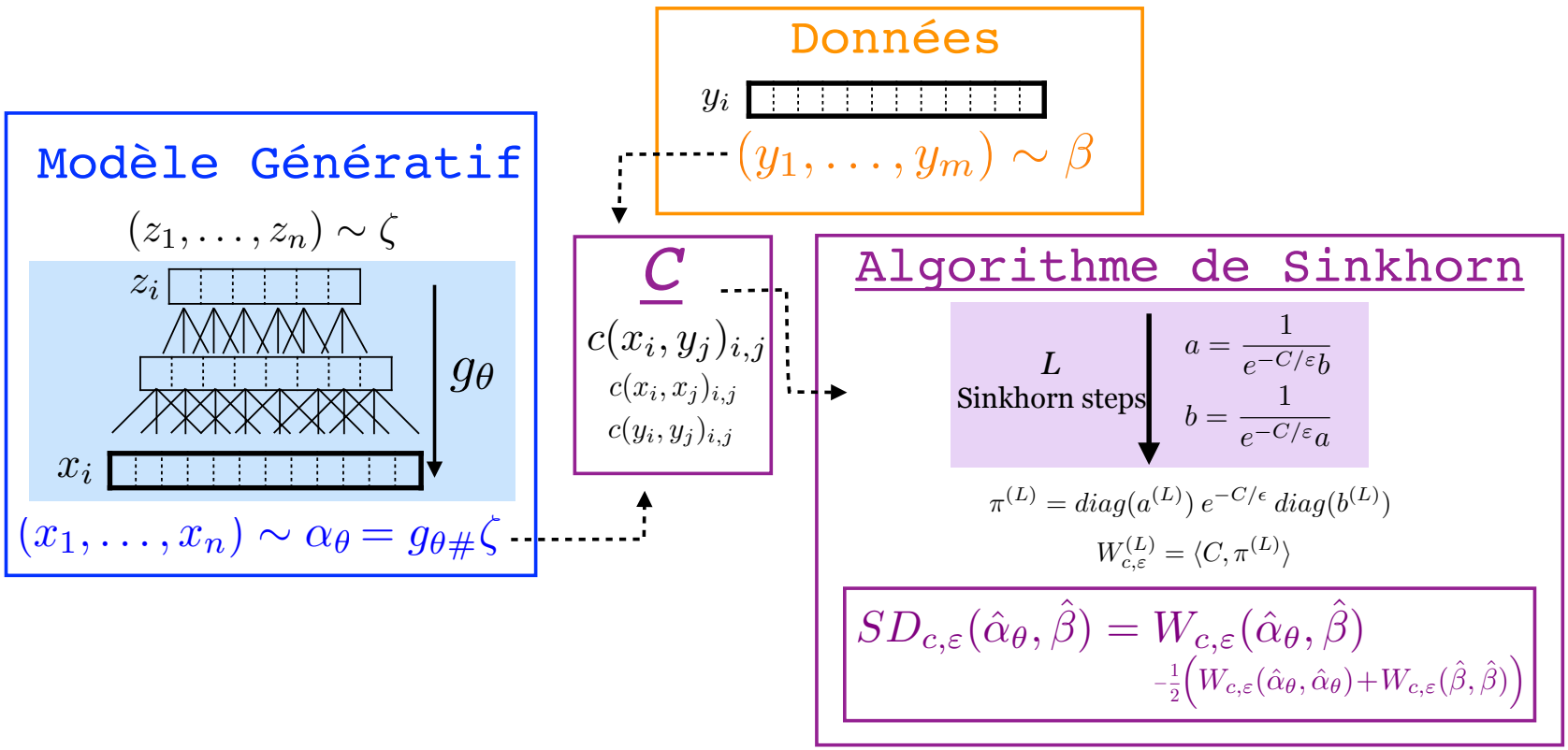
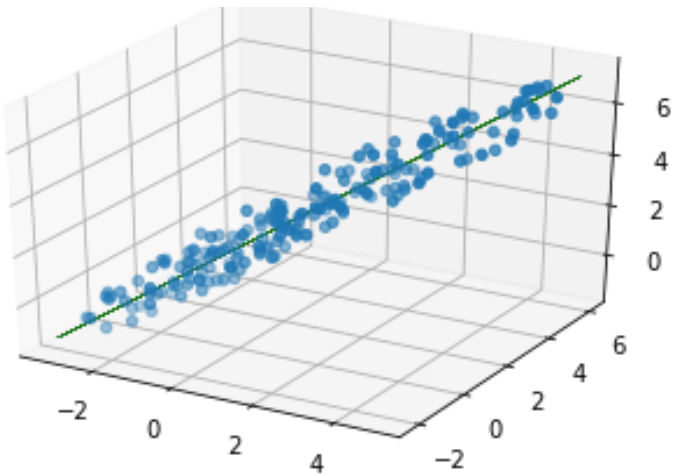


Figure 2 – Scheme of the approximation of the Sinkhorn Divergence from samples (here, $g_\theta : z \mapsto x$ is represented as a 2-layer NN).

Empirical Results

$W_{c,\varepsilon} - \varepsilon = 1, c = ||\cdot||_2^2$



$SD_{c,\varepsilon} - \varepsilon = 1, c = ||\cdot||_2^2$

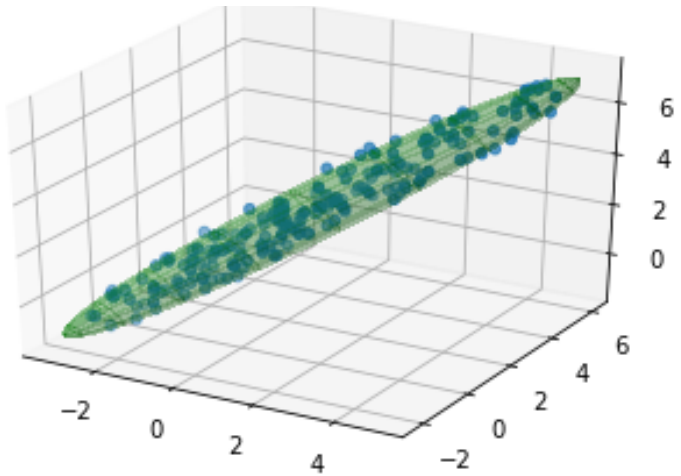


Figure 3 – Influence of the 'debiasing' of the Sinkhorn Divergence (SD_ε) compared to regularized OT (W_ε). Data are generated uniformly inside an ellipse, we want to infer the parameters A, ω (covariance and center).

Shape Registration

Goal : align 2 images of a same object from different patients, or between patient and atlas

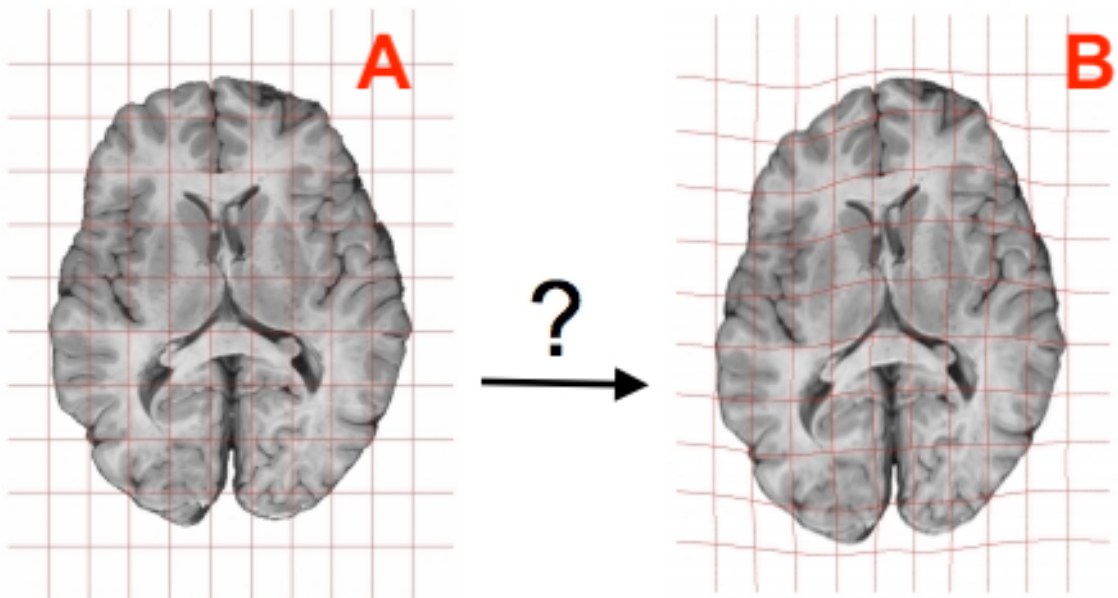


Image from Marc Niethammer

Mathematical Formulation

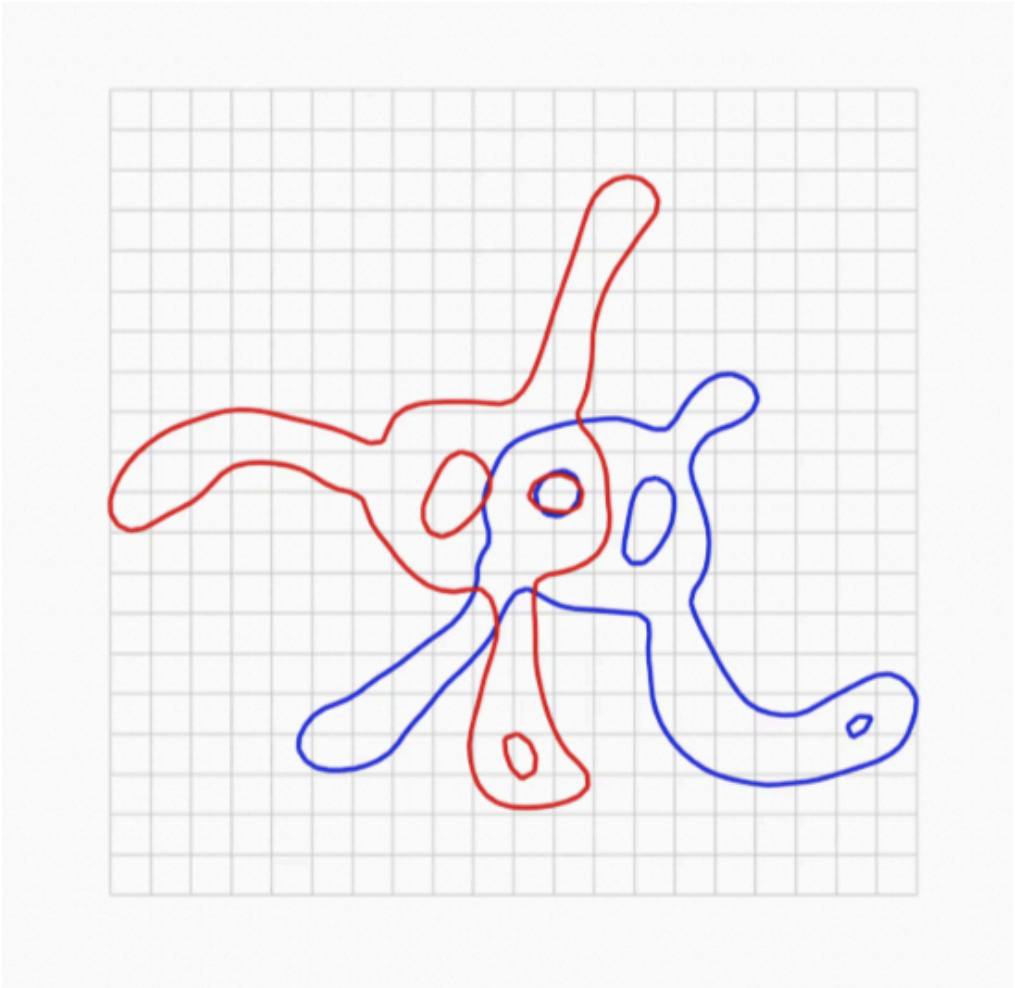
We are looking for φ diffeomorphism (smooth function with smooth inverse) that solves

$$\min_{\varphi} D(\varphi(A) - B) + R(\varphi)$$

where D is a distance on point clouds and R is a regularizer.

$\rightarrow$ we can use SD_ε and minimize over φ using backpropagation ! **(Feydy et. al. '18)**

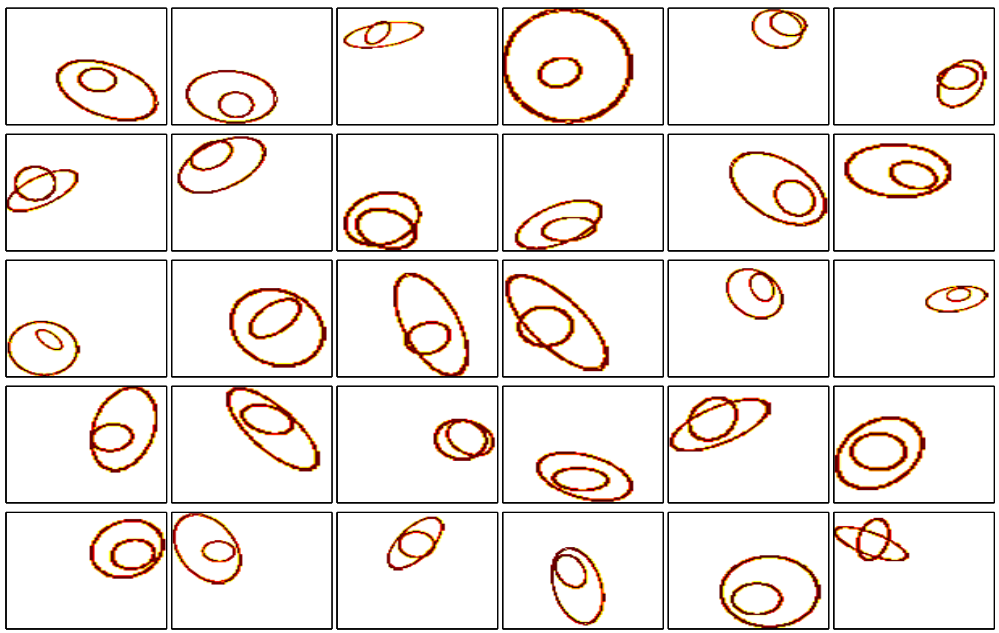
Registration using Sinkhorn Divergences



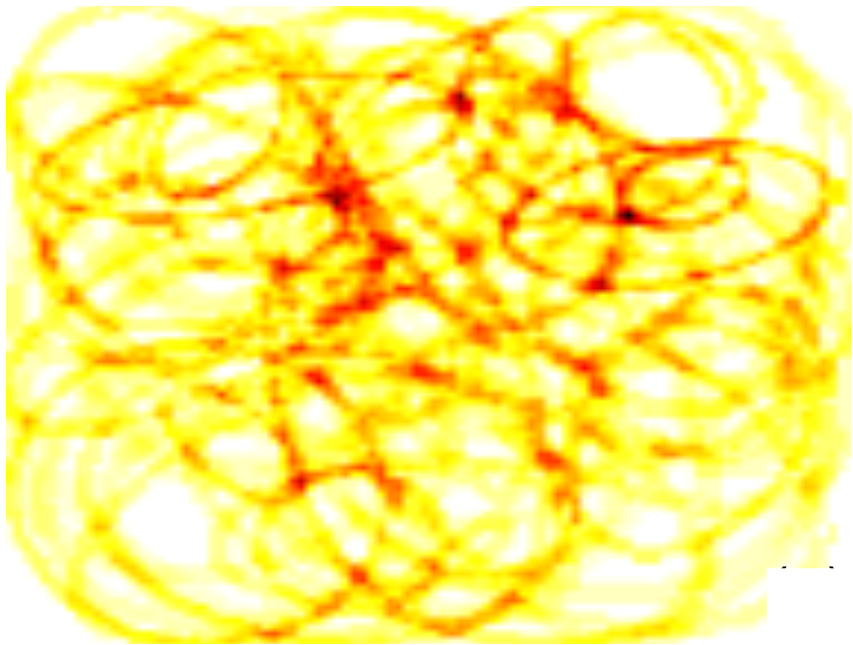
Images from Jean Feydy

Averaging point clouds / images

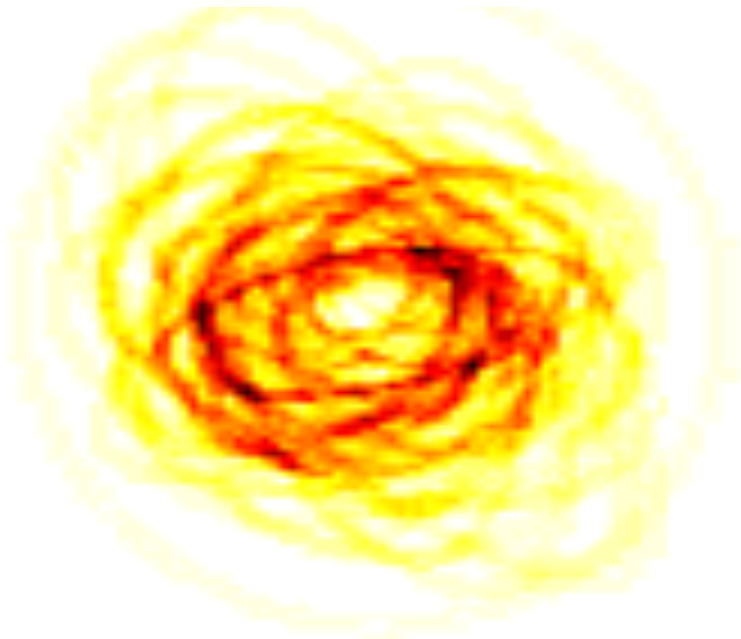
Goal : Given images of a same object from different sources, get and 'average' representation of that object.



The Euclidean average fails



The centered Euclidean average (also) fails

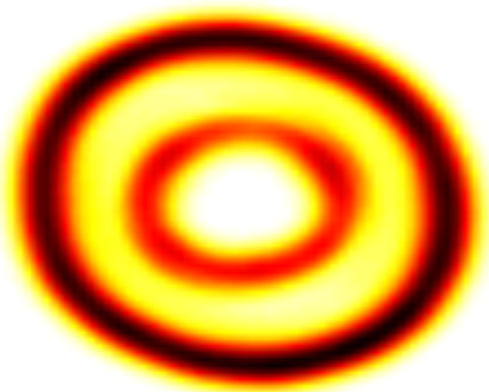


Introducing : the Wasserstein barycenter

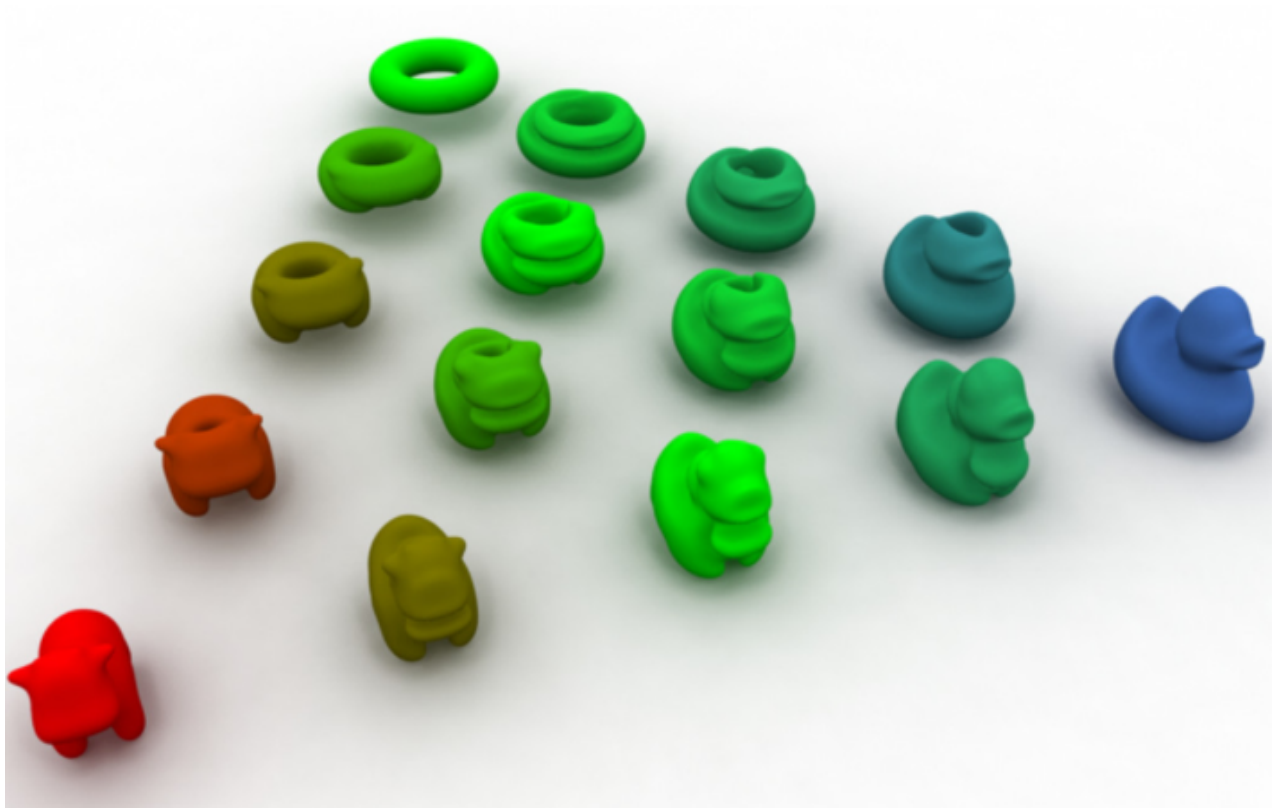
$$\min_{\alpha} \sum_{i=1}^{n} W(\alpha, \beta_i)$$

Introducing : the Wasserstein barycenter

$$\min_{\alpha} \sum_{i=1}^{n} W(\alpha, \beta_i)$$

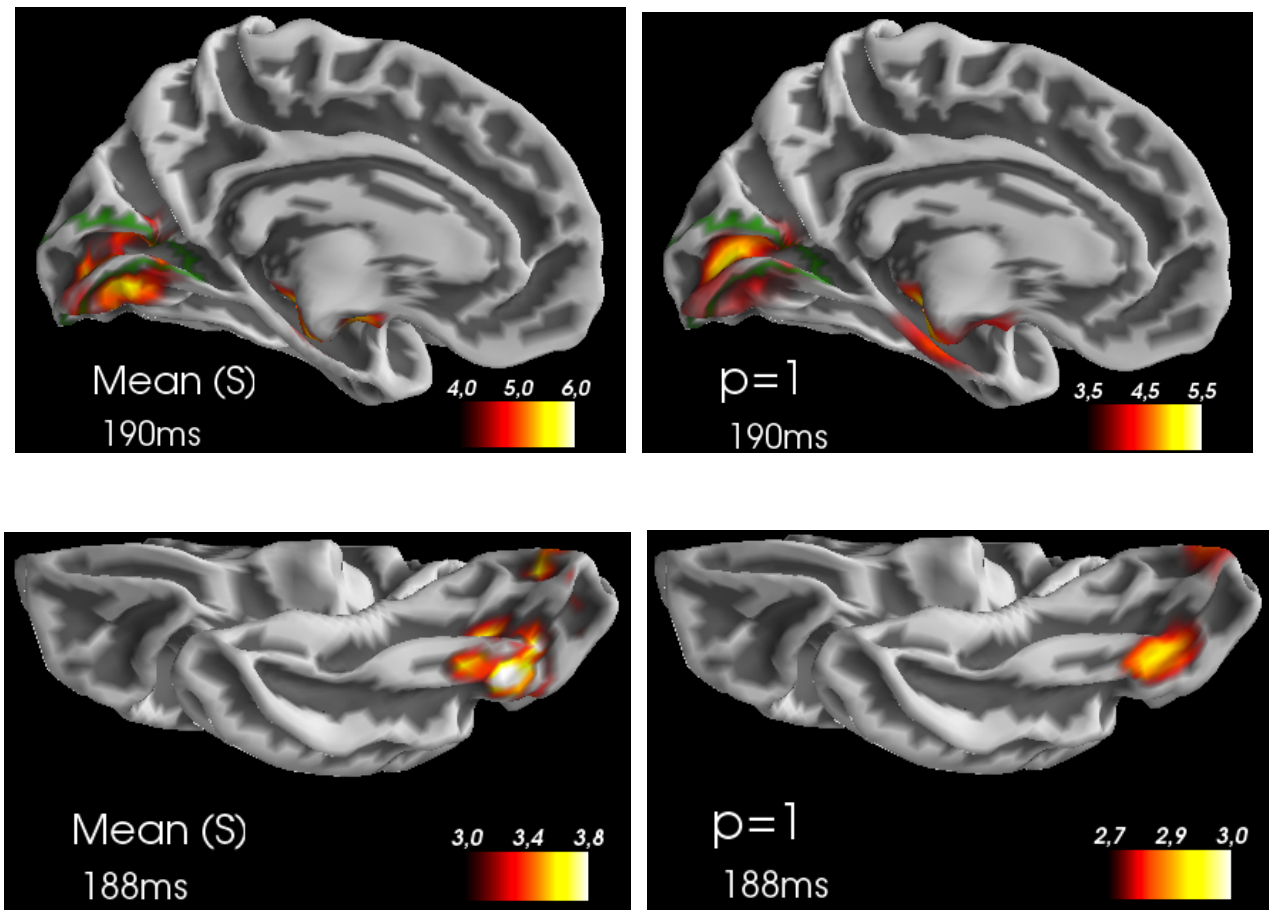


Interpolation between shapes



From Solomon et. al '15

Application to brain imaging



Euclidean Wasserstein

Waddington's Landscape Metaphor for Cell Differentiation

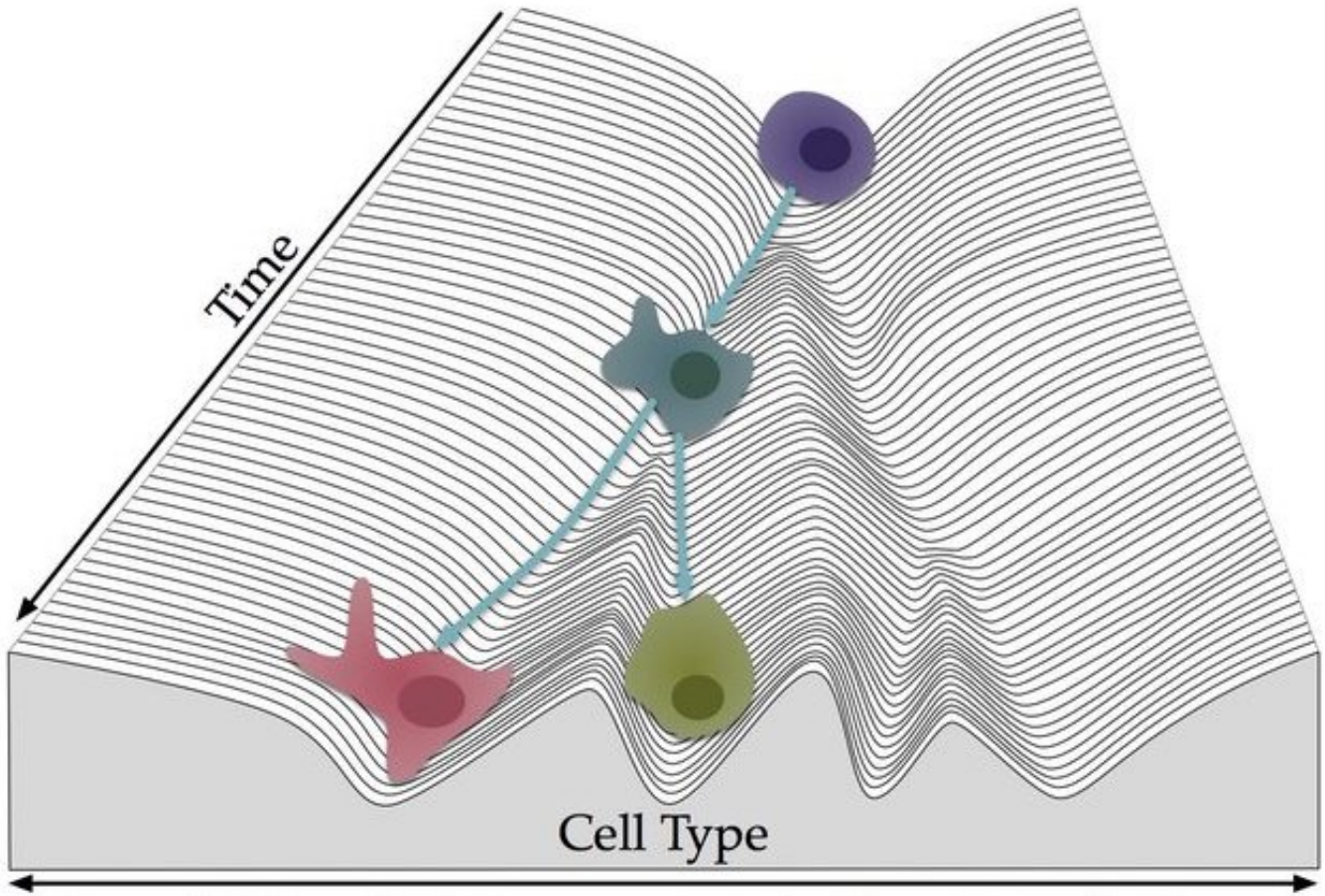


Figure 4 – Illustration taken from Zwiessele and Lawrence (2017)

Tracking single-cell development with OT

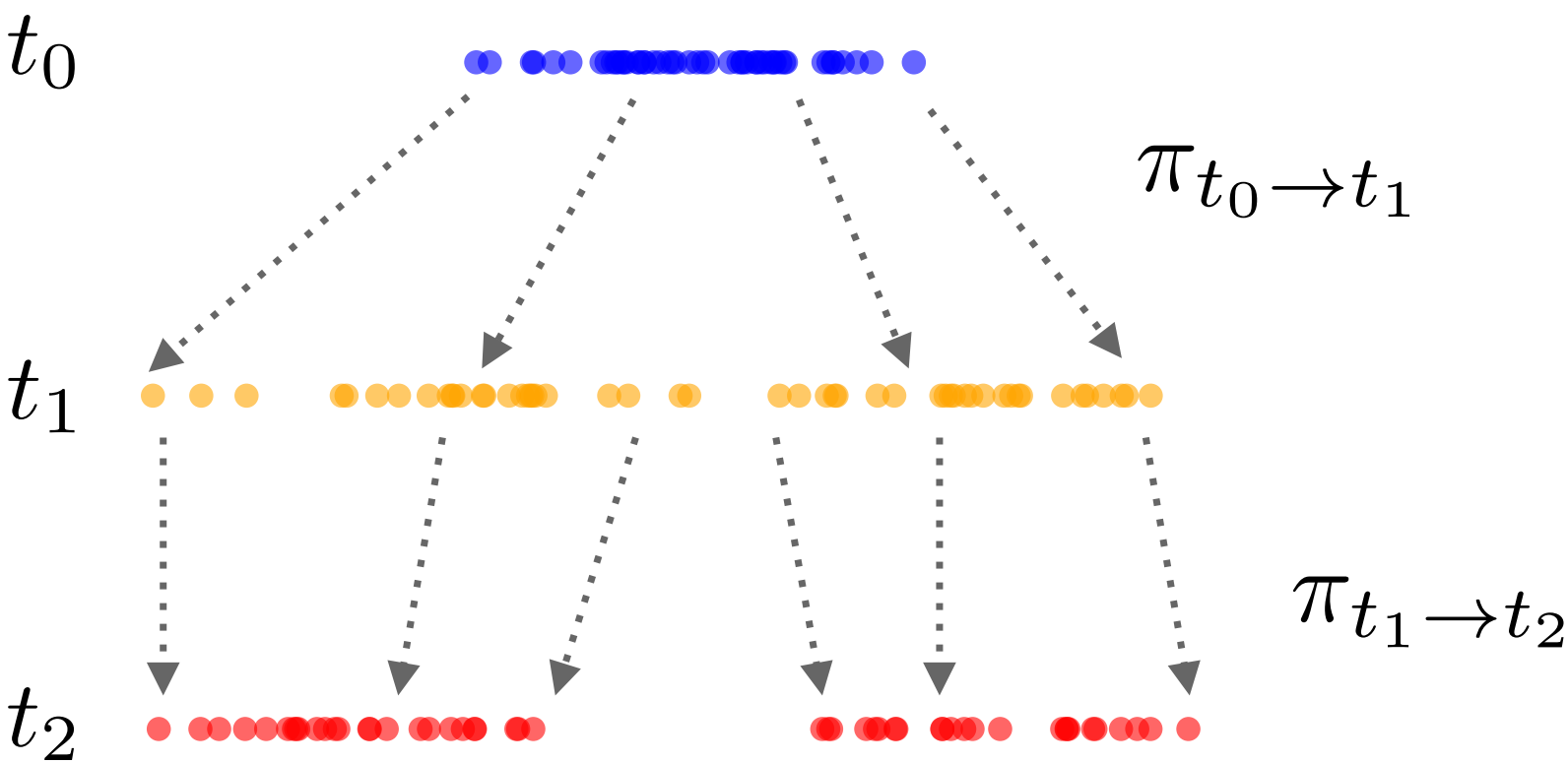
Goal : Use single cell RNA-sequencing data to recover developmental landscape of cells

Challenge : Single cell is destroyed by sequencing, so can't be followed over time

Idea (Schiebinger et. al. '19) :

- at time t, represent single cells as data points in gene expression space $\rightarrow$ point cloud in $\mathbb{R}^d$
- use optimal transport to infer trajectory of cells between time t and $t+1$.

Tracking single-cell development with OT



Take home message

- Optimal Transport gives a powerful tool to compare probability distribution
- Also induces a coupling between points in both measures
- Can be applied to a variety of problems where geometry matters :
  - fitting generative models and shape registration
  - averaging images or shapes
  - tracking evolution phenomenon (e.g. single cell)