Introduction to Optimal Transport

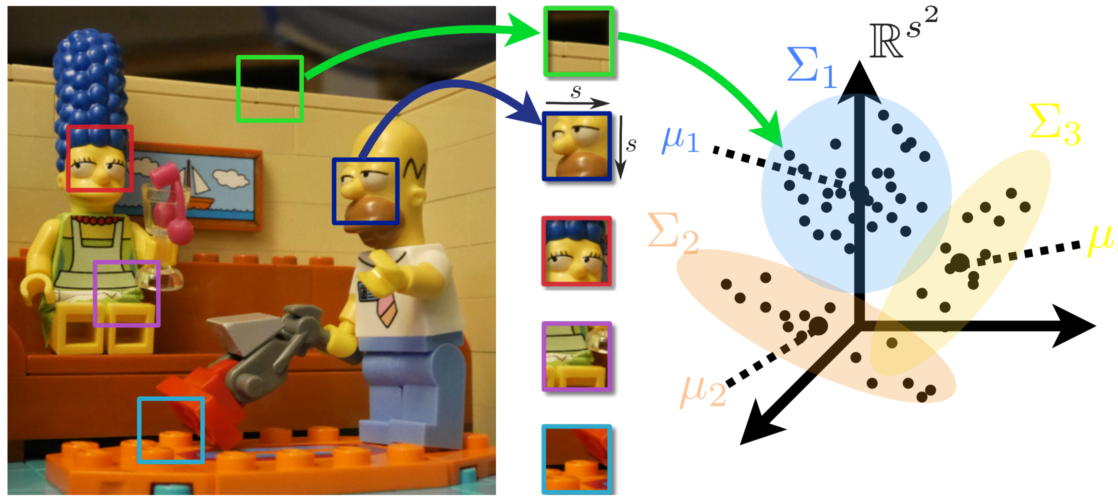
Julie Delon

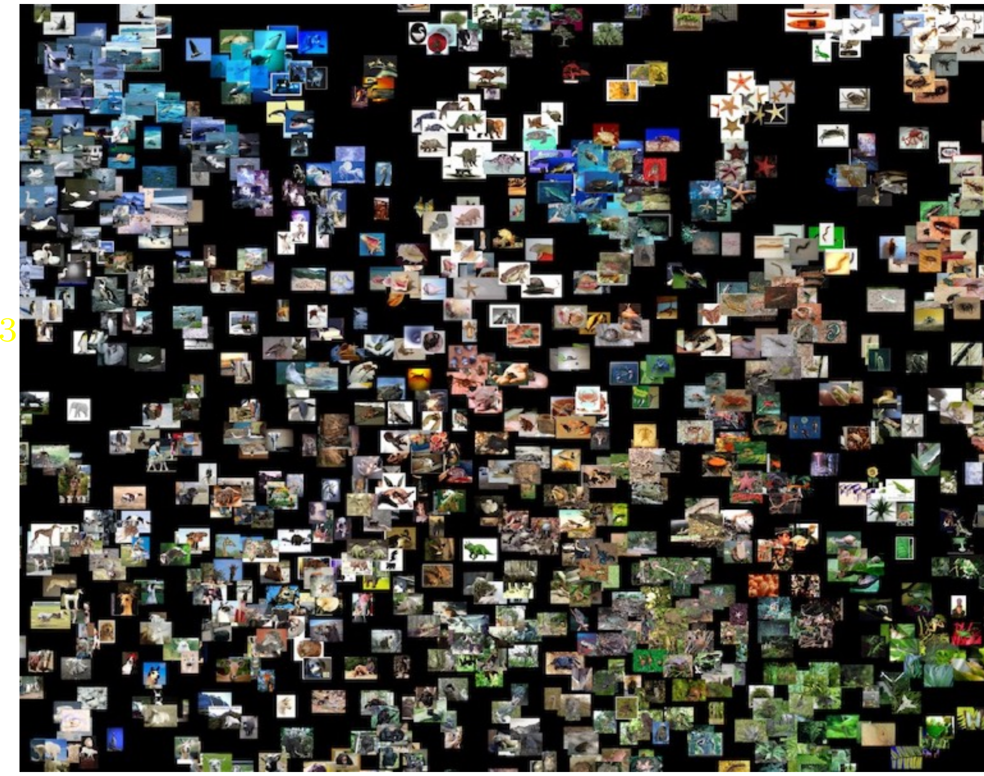
Ecole de recherche GDR IG-RV, 02/11/2020

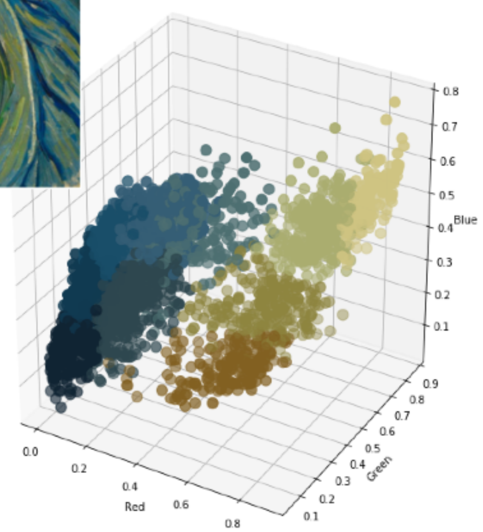
LABORATOIRE MAP5

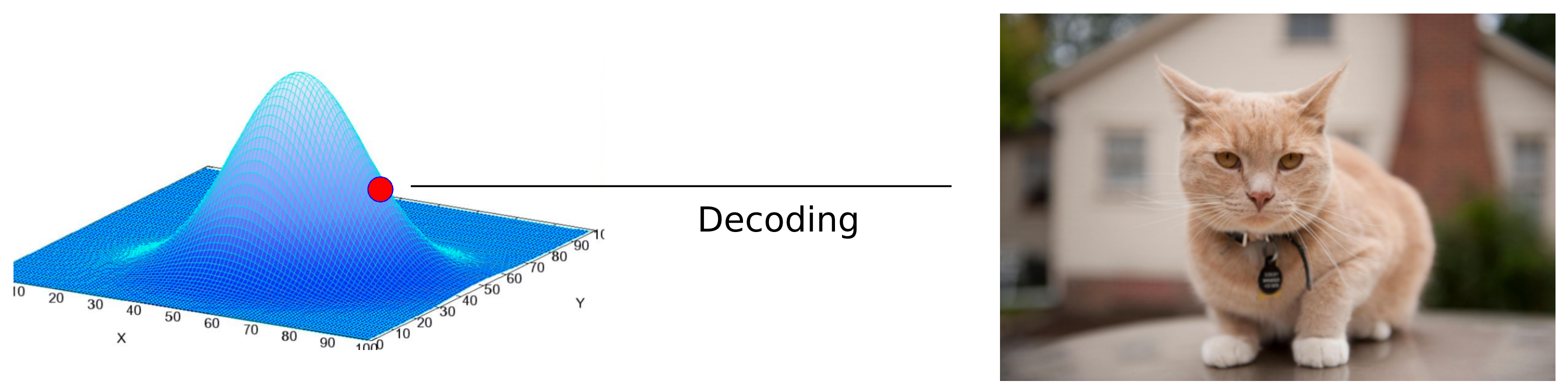
Université de Paris

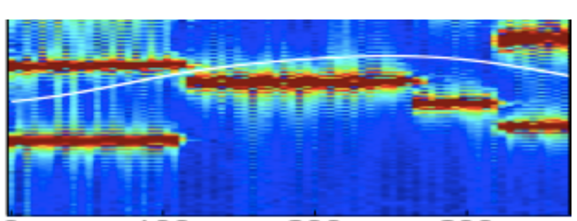
Optimal Transport in data science

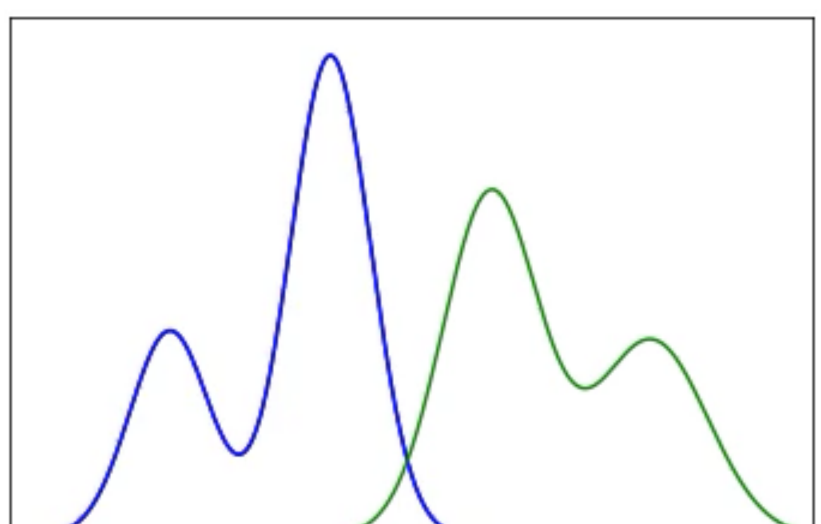


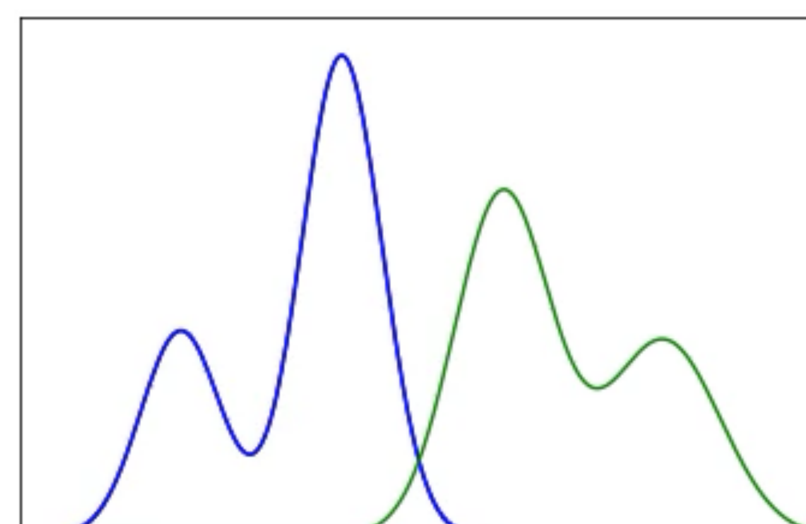




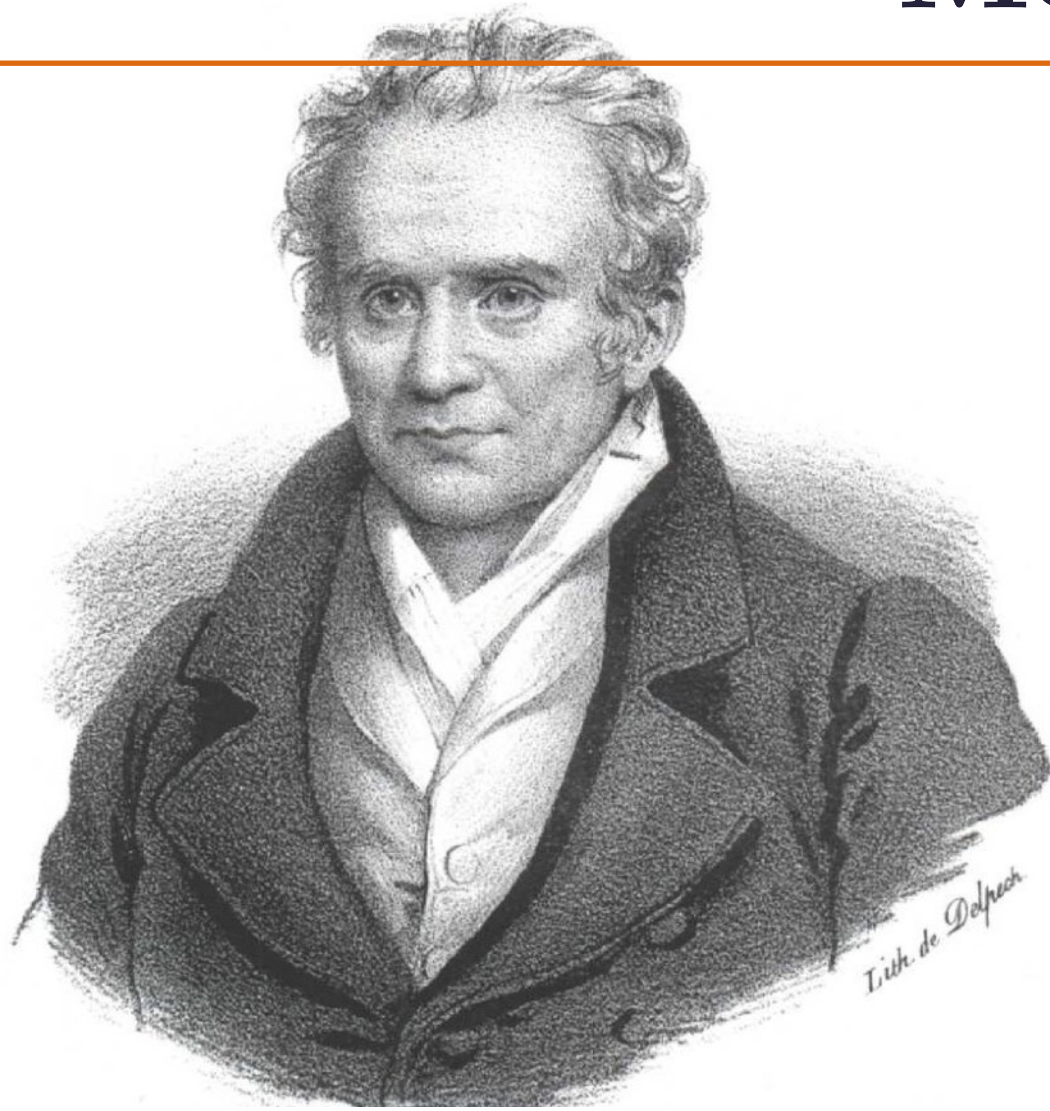


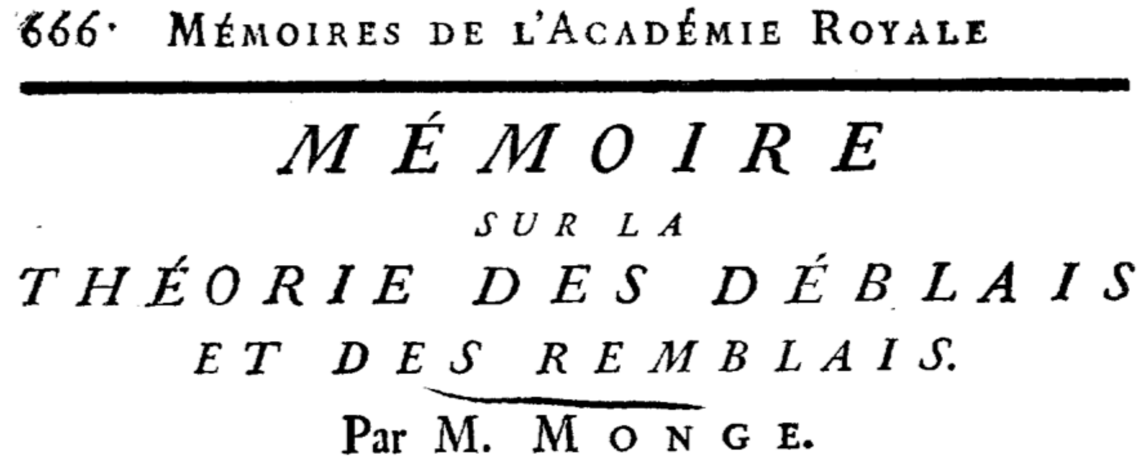






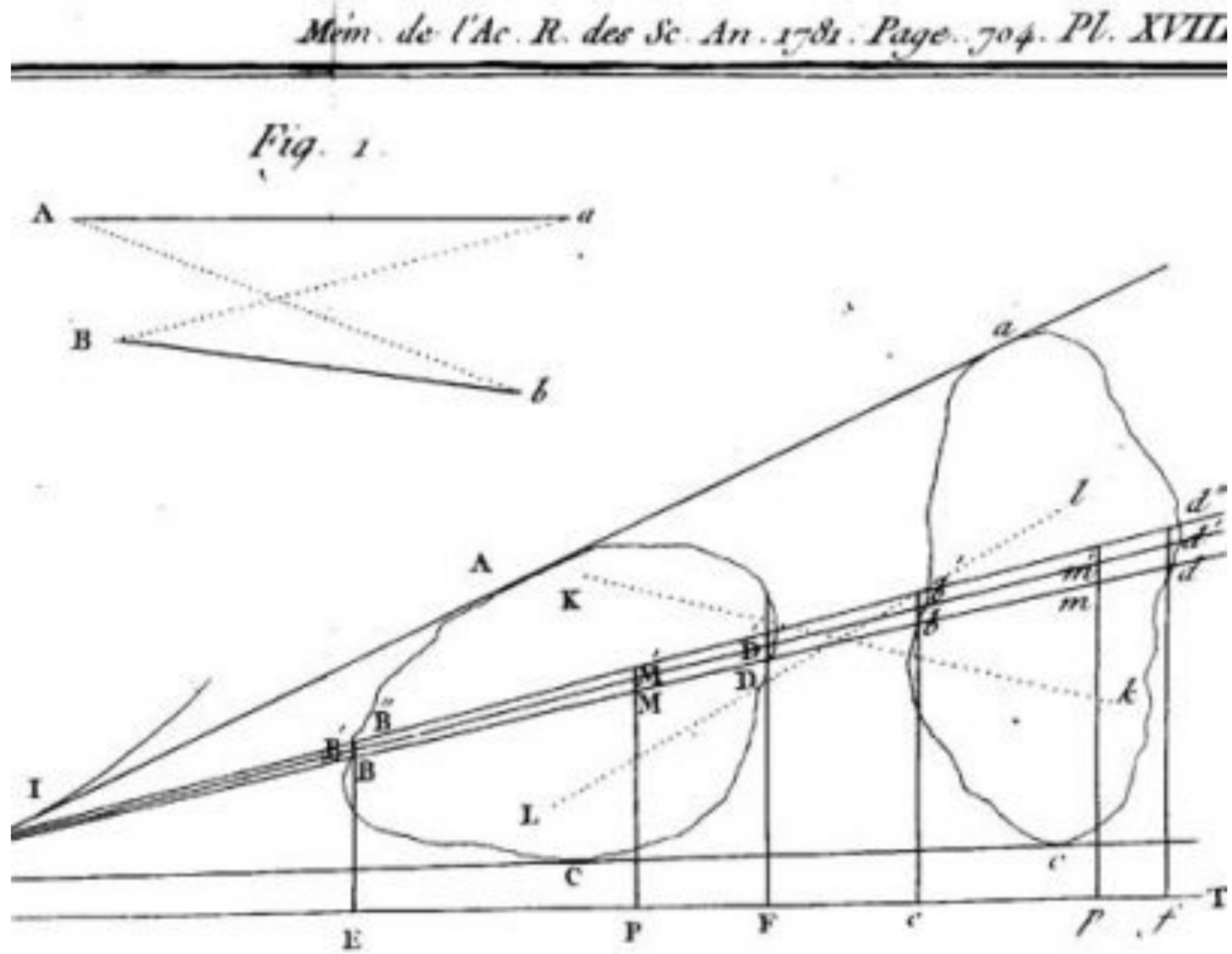
Monge, 1781

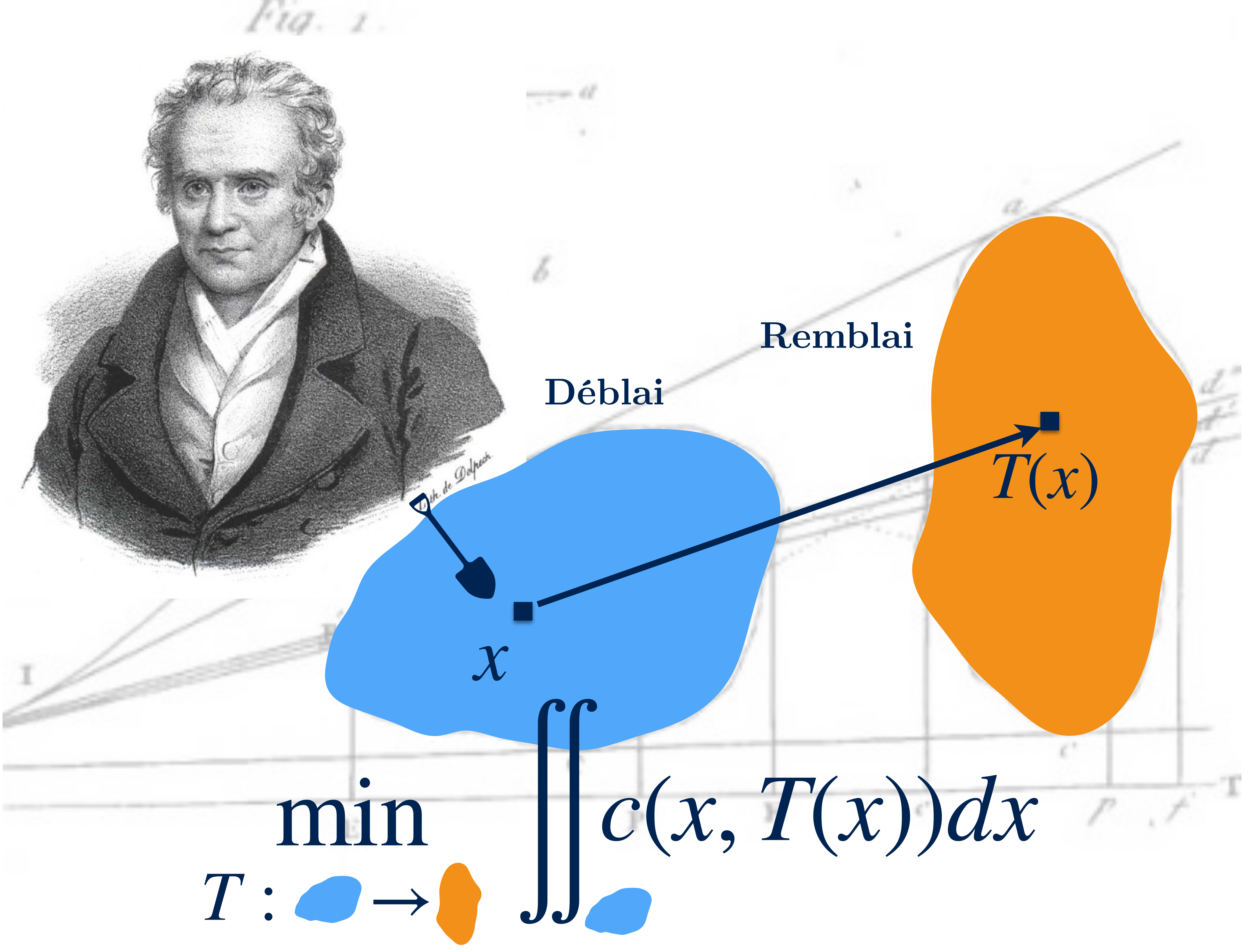




666 · Mémoires de l'Académie Royale

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. Monge.




Remblai
Déblai
$T(x)$
x
$\min_{T:} \iint c(x, T(x))dx$

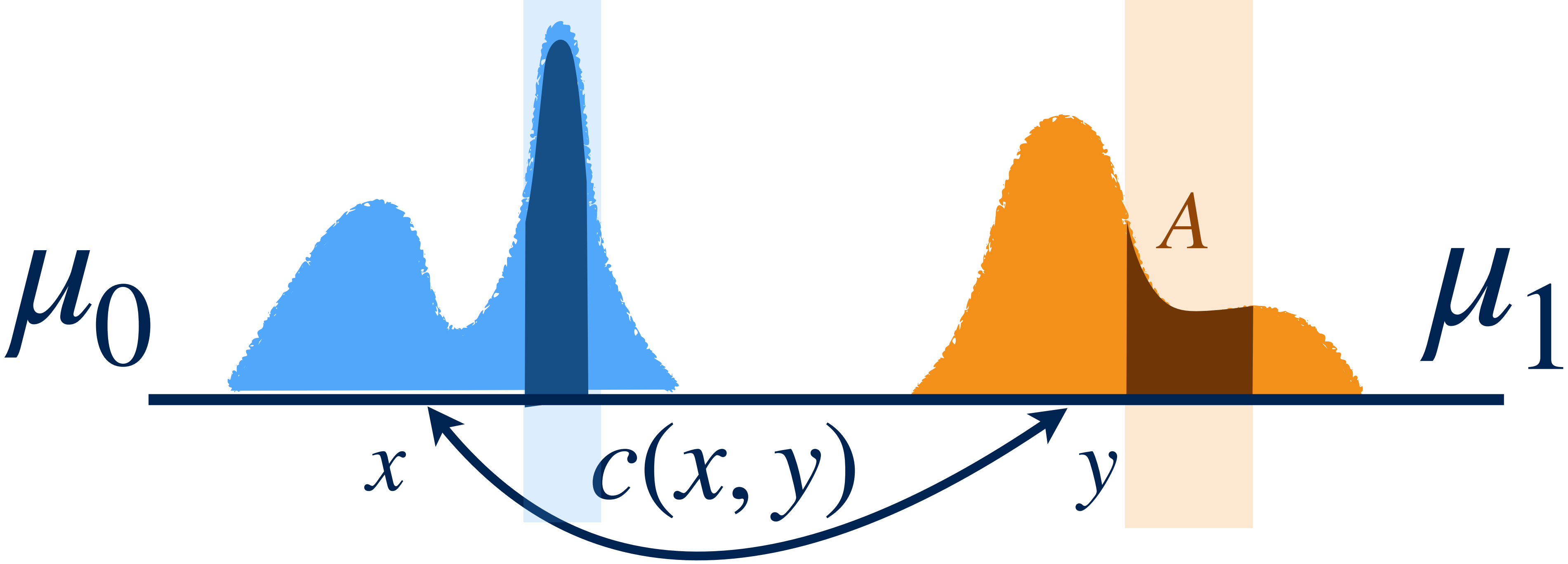
Monge optimal transport

How to transfer the mass from μ_0 to μ_1 at minimal cost?

Push forward $\mu_1 = T\#\mu_0$

$$\forall A, \quad \mu_1(A) = \mu_0(T^{-1}(A))$$

$$T^{-1}(A) = \{x, T(x) \in A\}$$

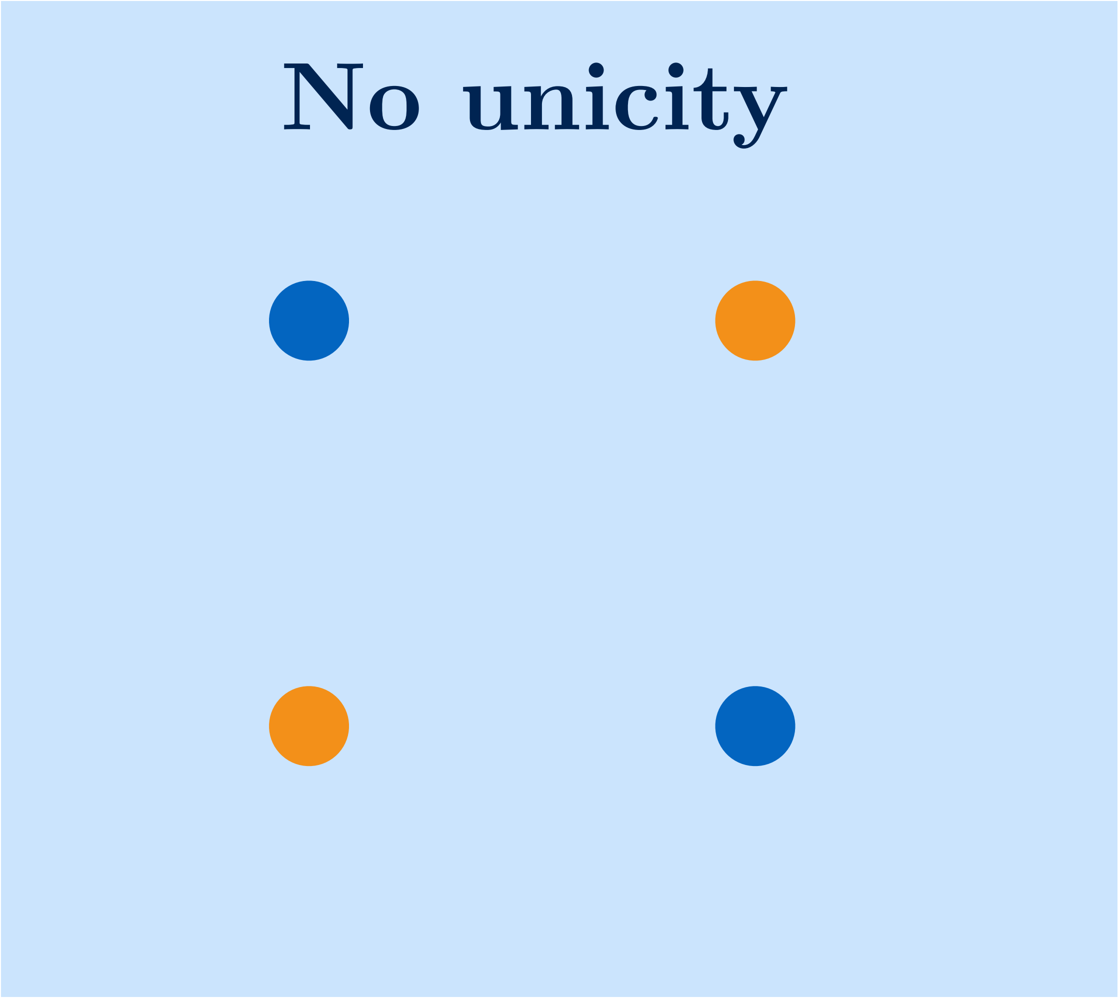


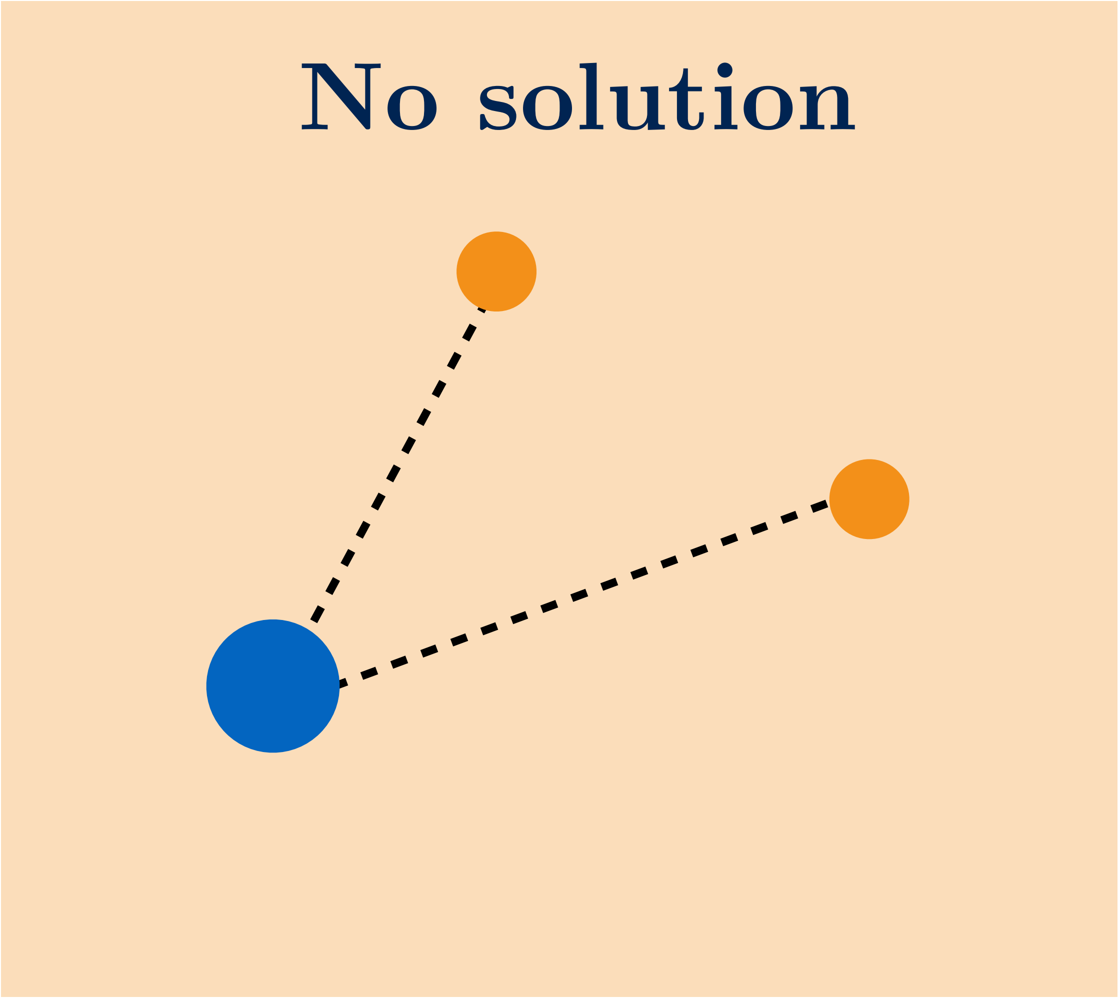
Total cost of the mass transfer = sum of costs of displacements of elementary masses.

$$\inf_{T\#\mu_0=\mu_1} \int c(x, T(x)) d\mu_0(x)$$

Finding T???

<u>**Difficult Problem, lack of symmetry, not convex.**</u>





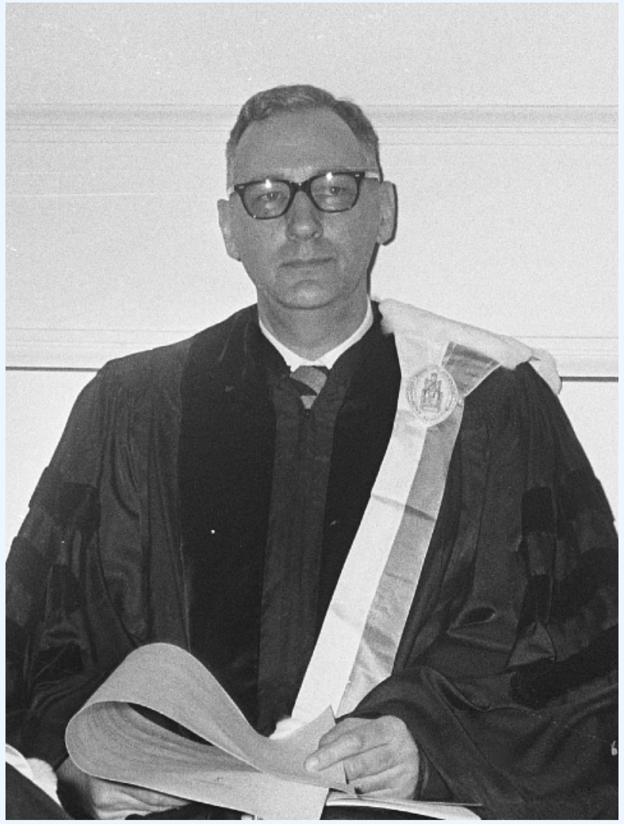
Linear programming

A.N.Tostoi, 1930

L. Kantorovich, 1939

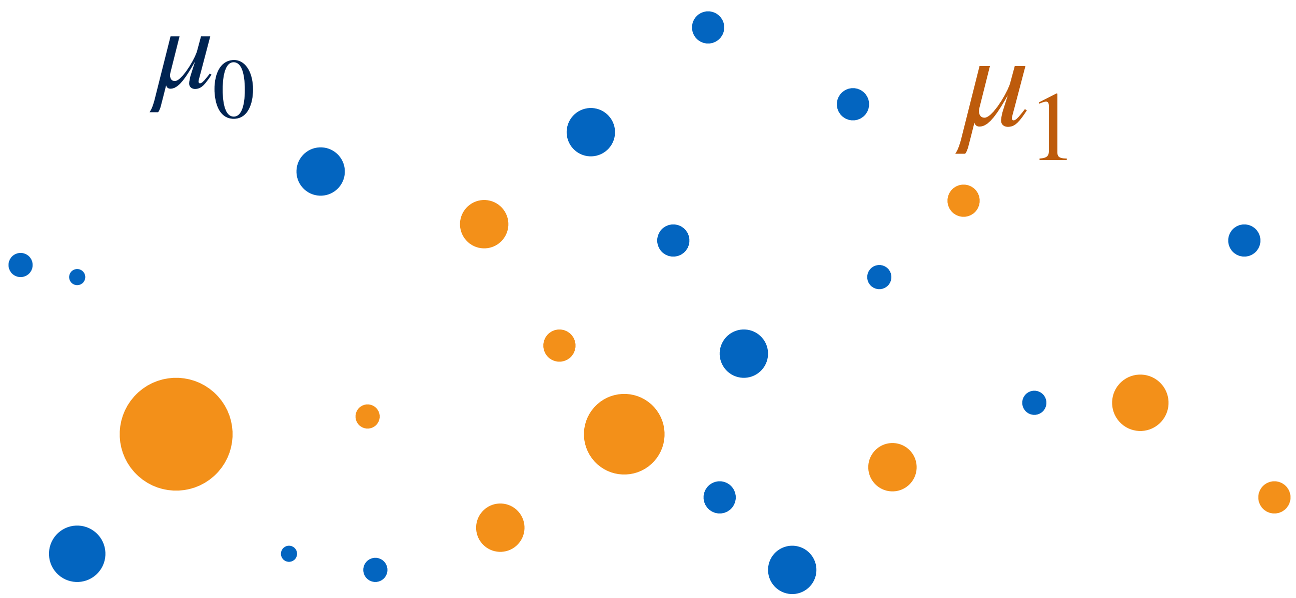
F.L. Hitchcock, 1941

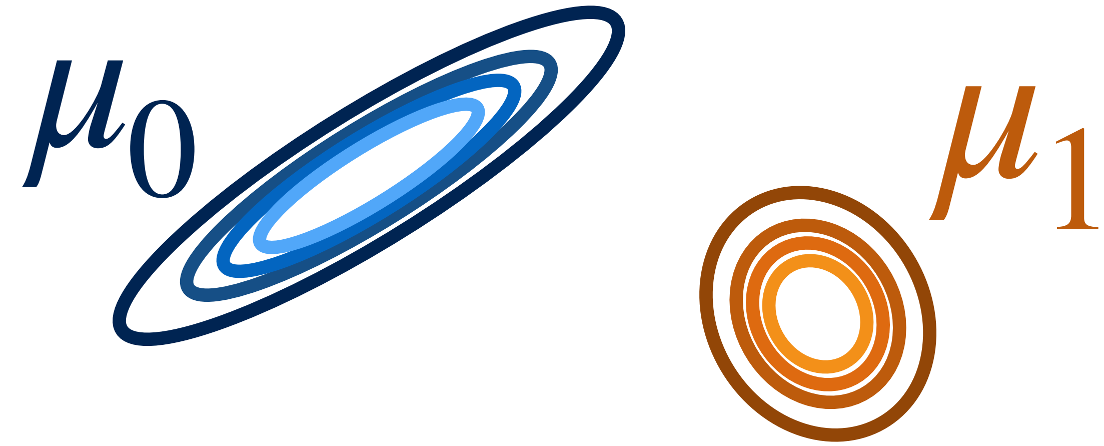
T.C. Koopmans, 1942



G. Dantzig, J. Von Neumann, 40’s, 50’s

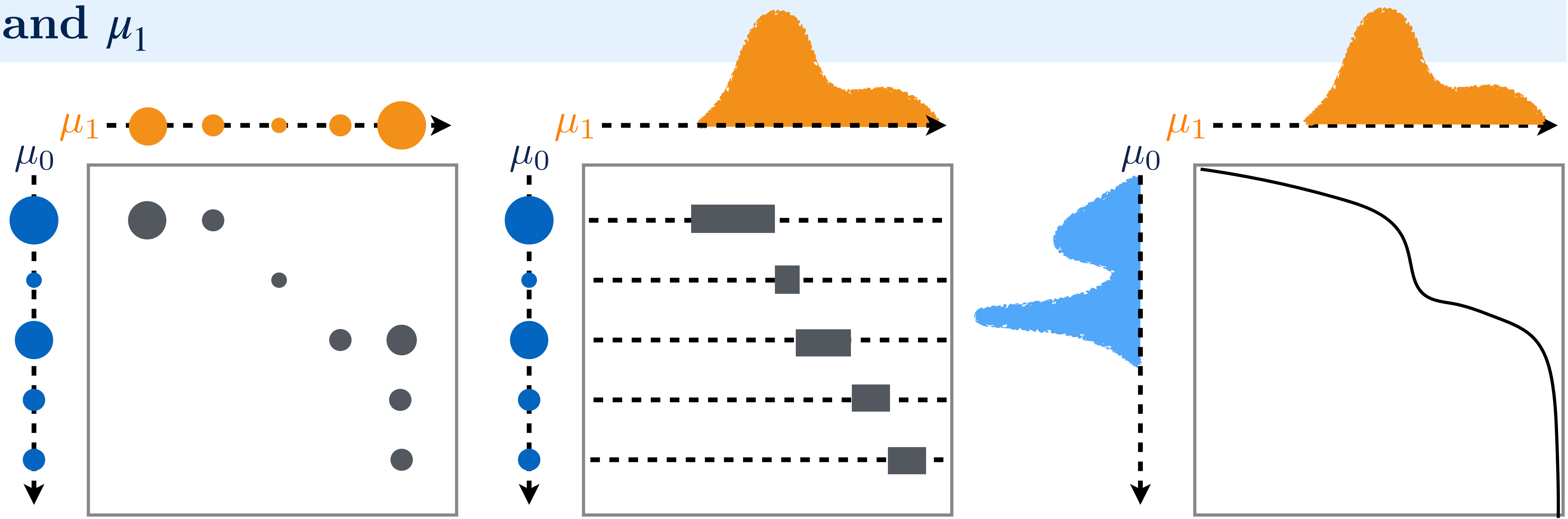
Couplings





<u>Couplings</u>

$\Pi(\mu_0, \mu_1)$ = **probability distributions on** $X \times X$ **with marginals** μ_0 **and** μ_1

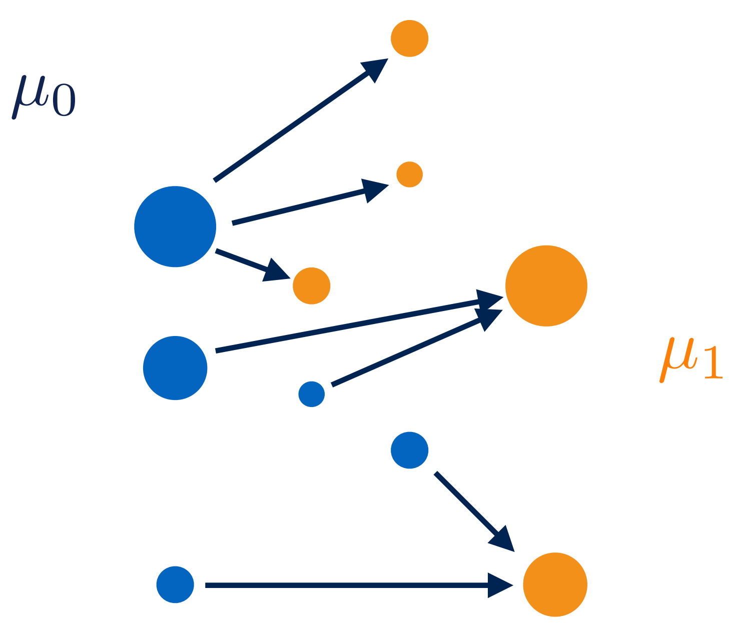


General formulation

[Kantorovich, *On the transfer of masses*, 1942]

$$W_c(\mu_0,\mu_1) = \inf_{\gamma\in\Pi(\mu_0,\mu_1)} \int_{X\times X} c(x,y)d\gamma(x,y).$$

Discrete case: $\mu_0 = \sum_i s_i\, \delta_{x_i},\ \mu_1 = \sum_j t_j\, \delta_{y_j}$ **with** $\sum_i s_i = \sum_j t_j$



$$\boxed{W_c(\mu_0,\mu_1) = \min_{\gamma\in\Pi(\mu_0,\mu_1)} \sum_i \sum_j c(x_i,y_j)\gamma_{ij}}$$

$$\Pi(\mu_0,\mu_1) = \left\{ \textbf{matrices } \gamma \textbf{ s.t. } \gamma_{i,j} \geq 0, \sum_i \gamma_{i,j} = t_j,\ \sum_j \gamma_{i,j} = s_i \right\}$$

Monge-Kantorovich

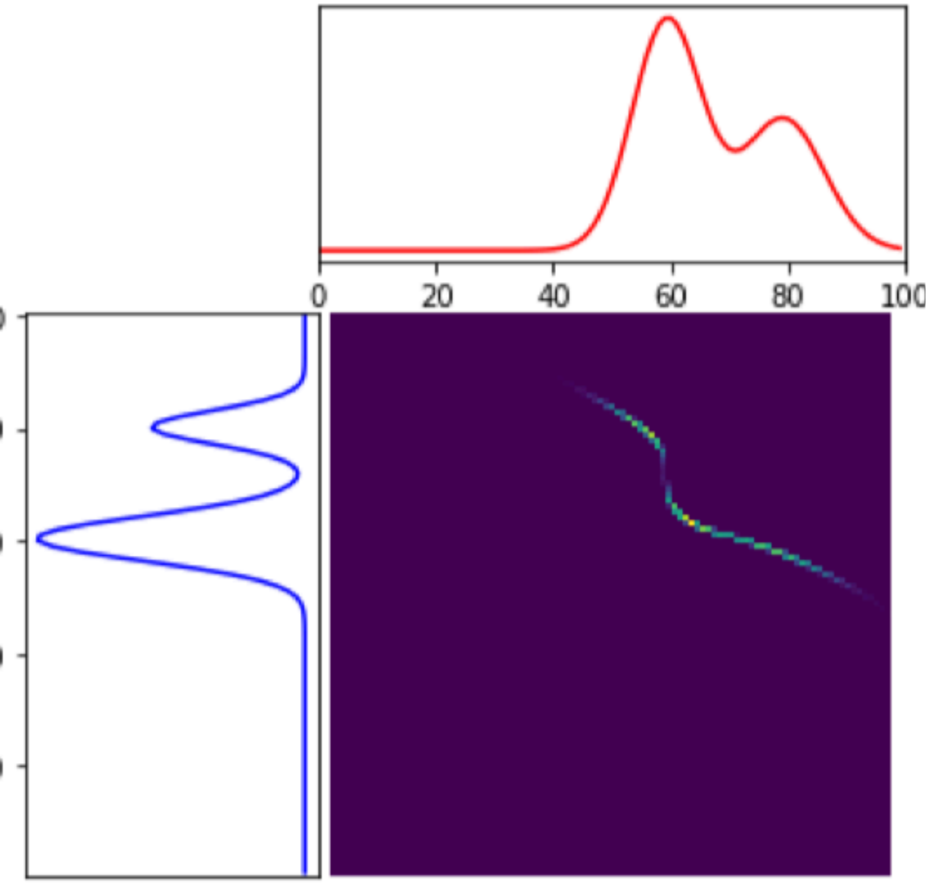
Monge, 1781

$$\inf_{T\#\mu_0=\mu_1} \int c(x,T(x))d\mu_0(x)$$

Kantorovich, 1939

$$\inf_{\gamma\in\Pi(\mu_0,\mu_1)} \int c(x,y)d\gamma(x,y)$$

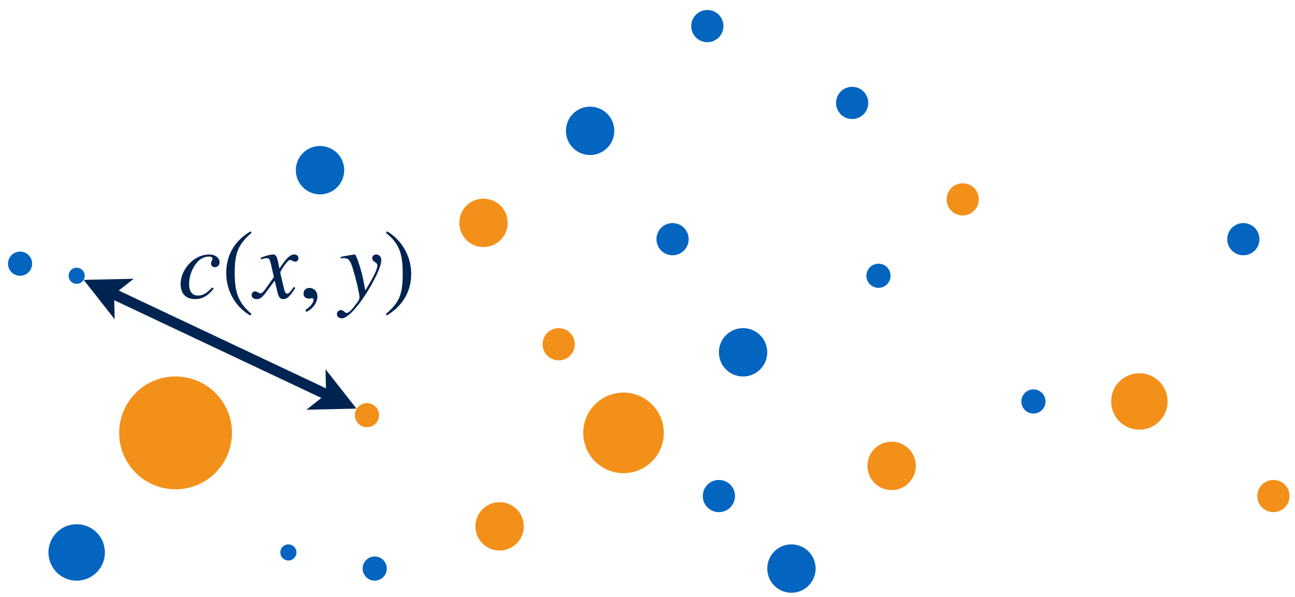
Brenier, 1991 If $c(x,y) = \|x-y\|^2$, if μ_0 has a density, Monge problem has a solution $T = \nabla\psi$ where ψ unique convex function s.t. $\nabla\psi\#\mu_0 = \mu_1$. The plan $\gamma = (Id, T)\#\mu_0$ is solution of Kantorovich pb.

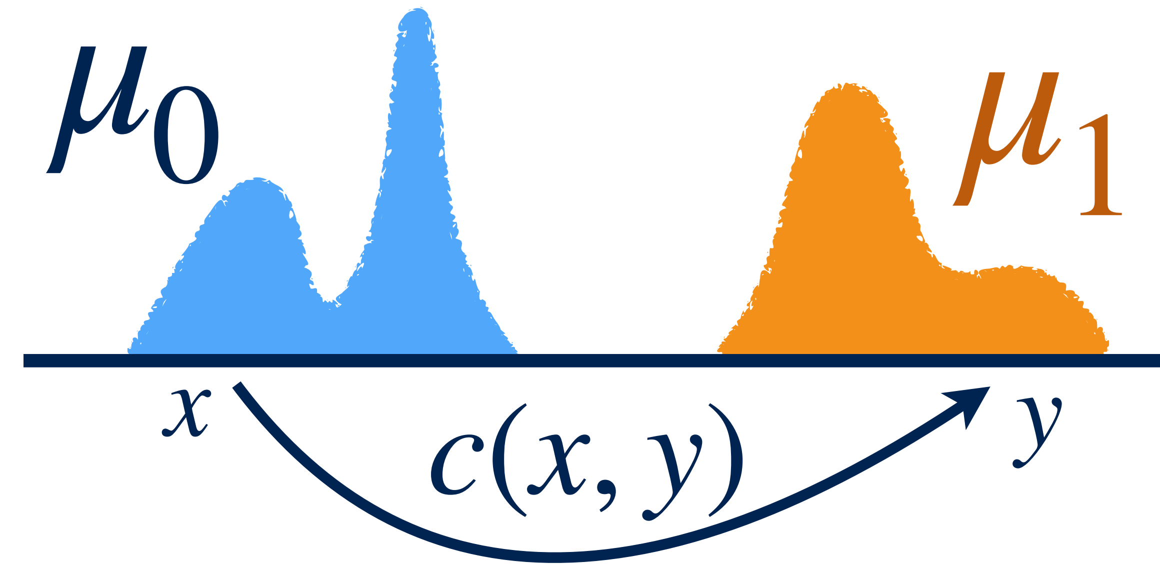


Displacement interpolation:

$$\mu_t = ((1-t)Id + tT)\#\mu_0, \;\; t \in [0,1]$$

Wasserstein distances





If $c(x, y) = d(x, y)^p$ with $p \geq 1$ and d a distance,

$$W_p(\mu_0, \mu_1) = \left(\inf_{\gamma \in \Pi(\mu_0, \mu_1)} \iint c(x, y) d\gamma(x, y) \right)^{\frac{1}{p}}$$

defines a **distance between probability measures**.

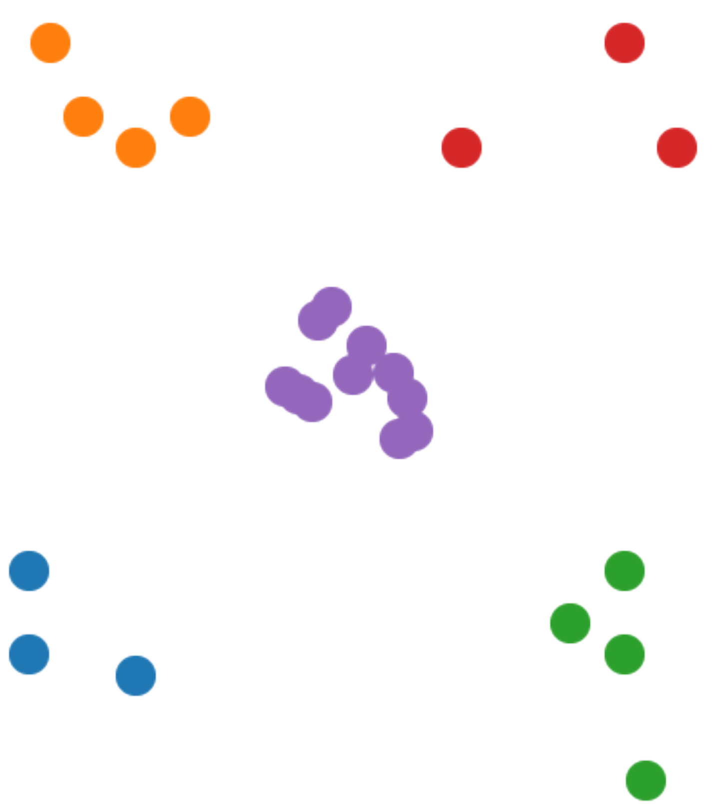
p=2 or 1 used in numerous applications

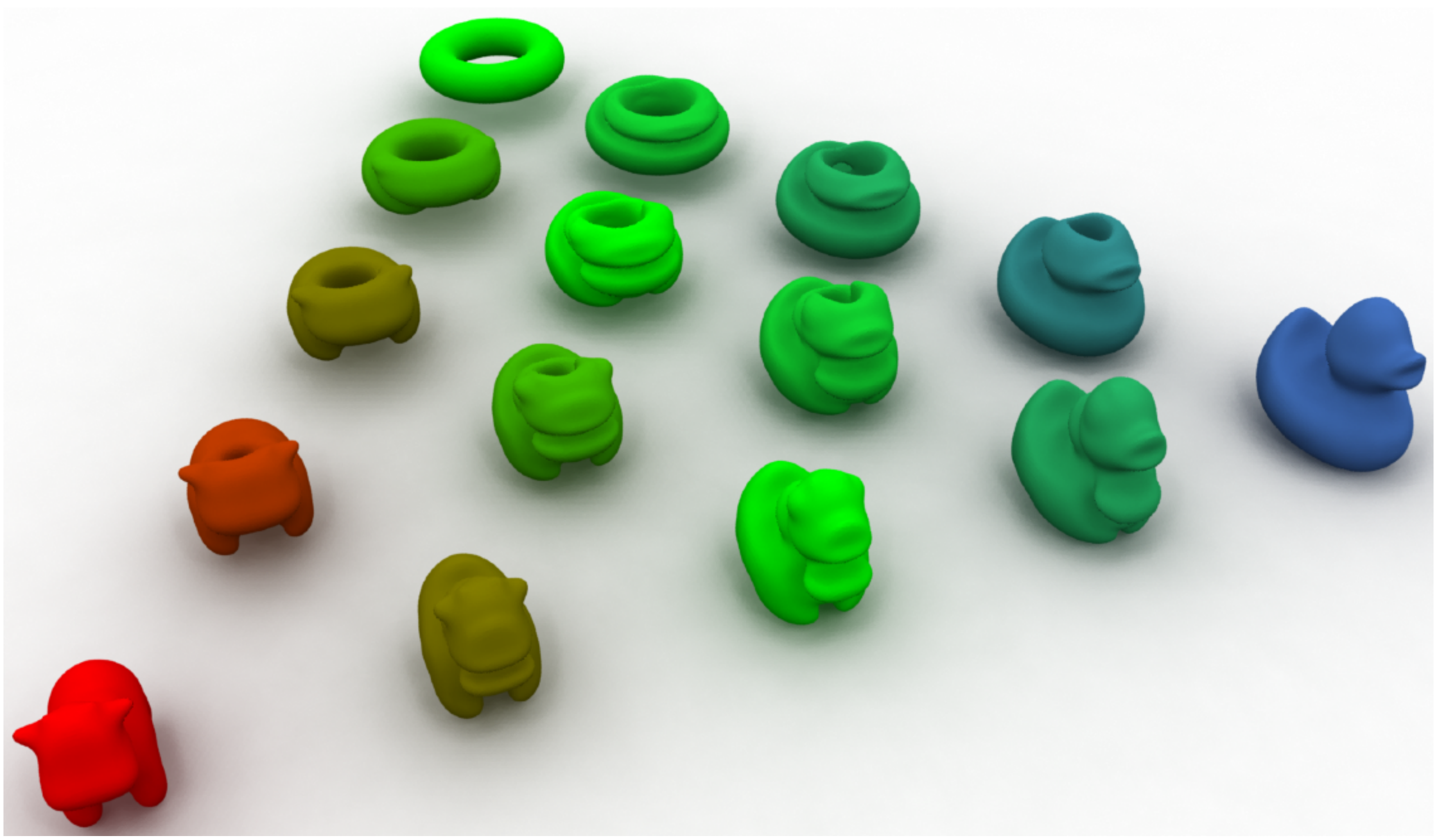
Wasserstein barycenters

Barycenter of $(\mu_i)_{i\in\{0,\dots,I-1\}}$, weights $\sum_i \lambda_i = 1$

$$\mu_{bary} \in \operatorname*{argmin}_{\rho} \sum_i \lambda_i W_2^2(\mu_i, \rho)$$

Prop. [Agueh, Carlier 2011]: existence and unicity of the barycenter for $c(x,y) = ||x - y||^2$ if the μ_i vanish on small sets.





[Solomon et al. 2015]

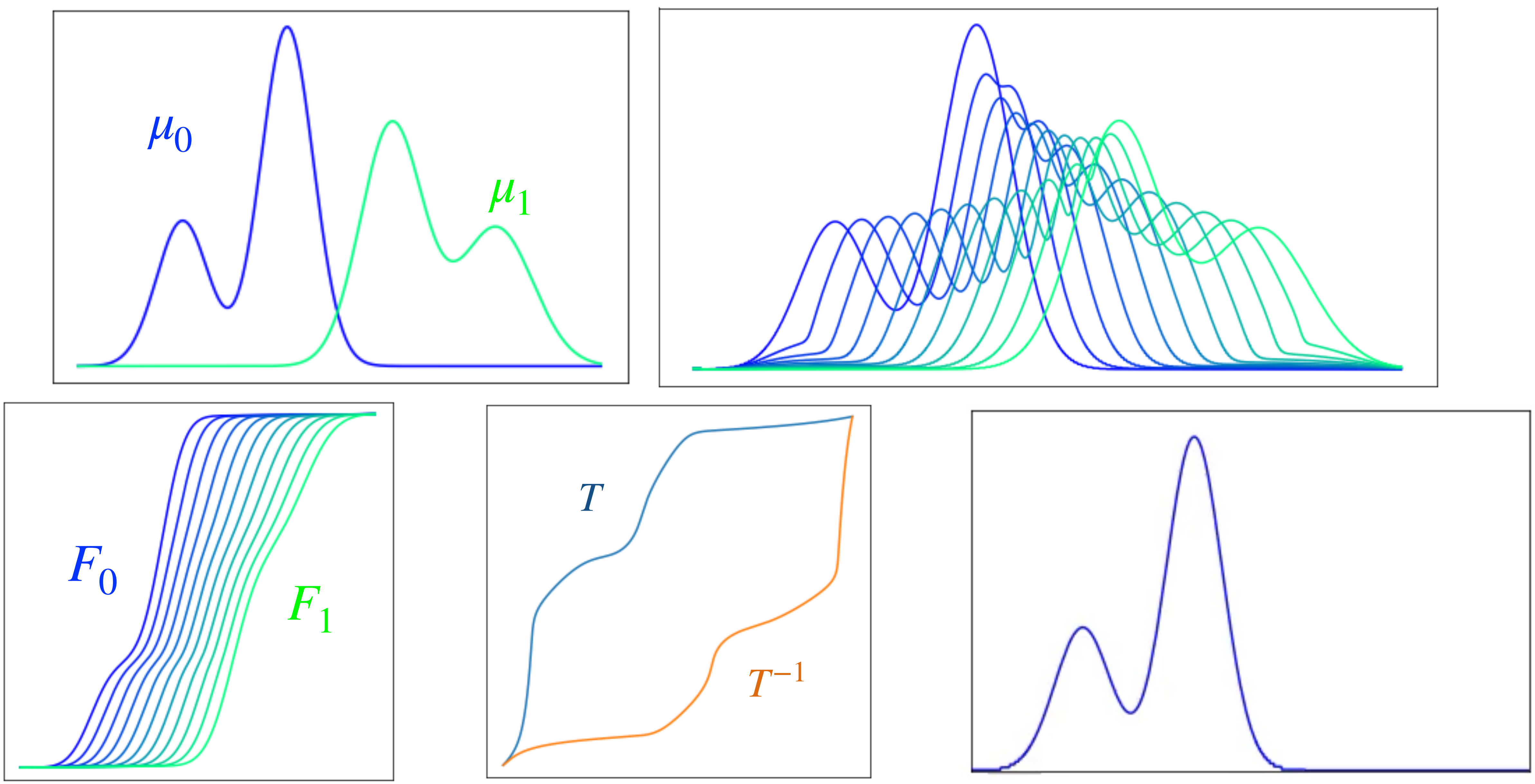
Optimal transport in one dimension

Optimal Transport in 1D

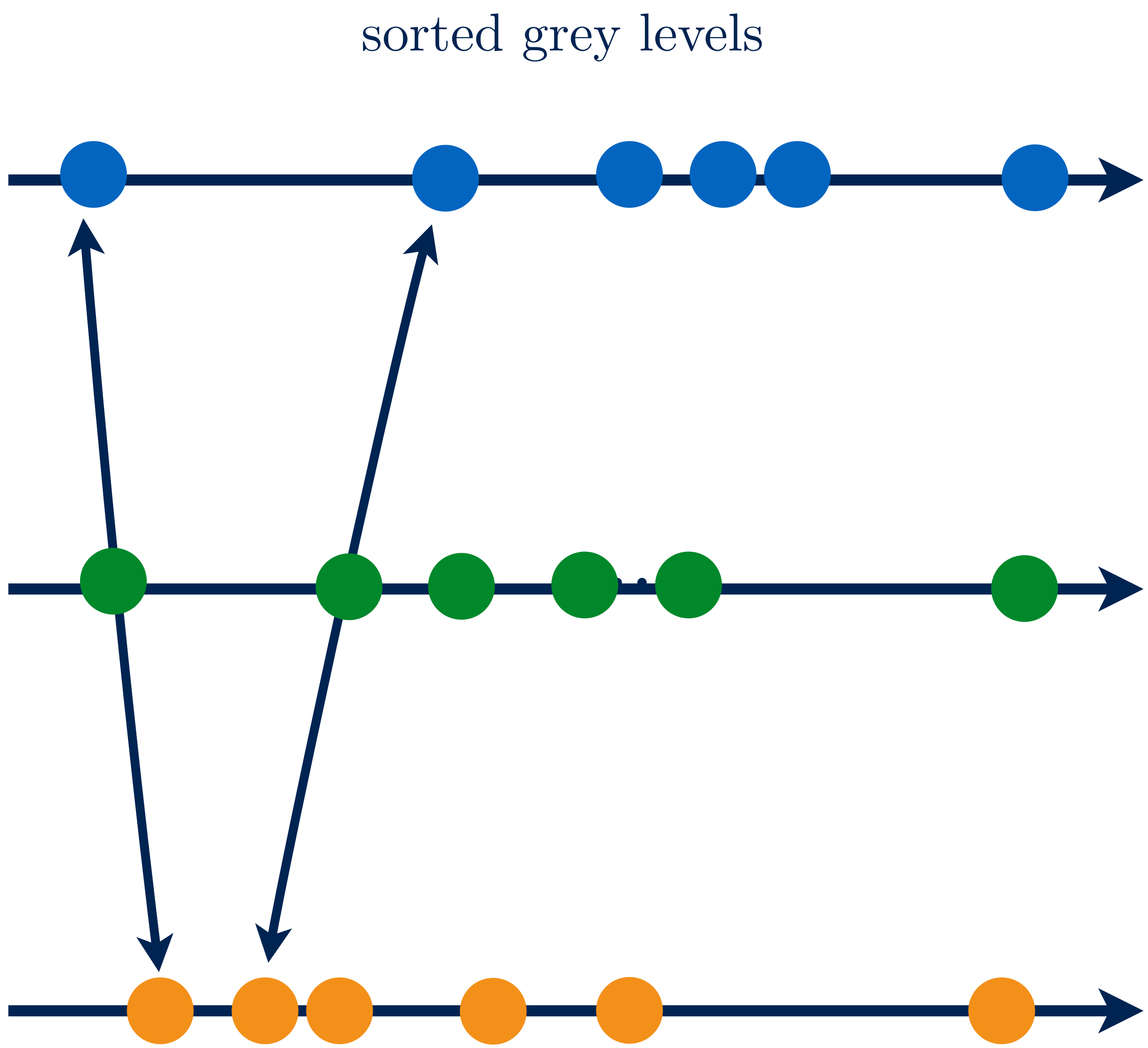
On $\mathbb{R}$, if $c(x,y) = f(|x-y|)$ with f convex,

$$W_c(\mu_0, \mu_1) = \int_0^1 f(|F_0^{-1}(t) - F_1^{-1}(t)|)dt,$$

with F_0 and F_1 the distribution functions of μ_0 and μ_1. Moreover, if μ_0 has no atoms, $T = F_1^{-1} \circ F_0$ is solution of the Monge problem.



Midway histogram



OT between Gaussians

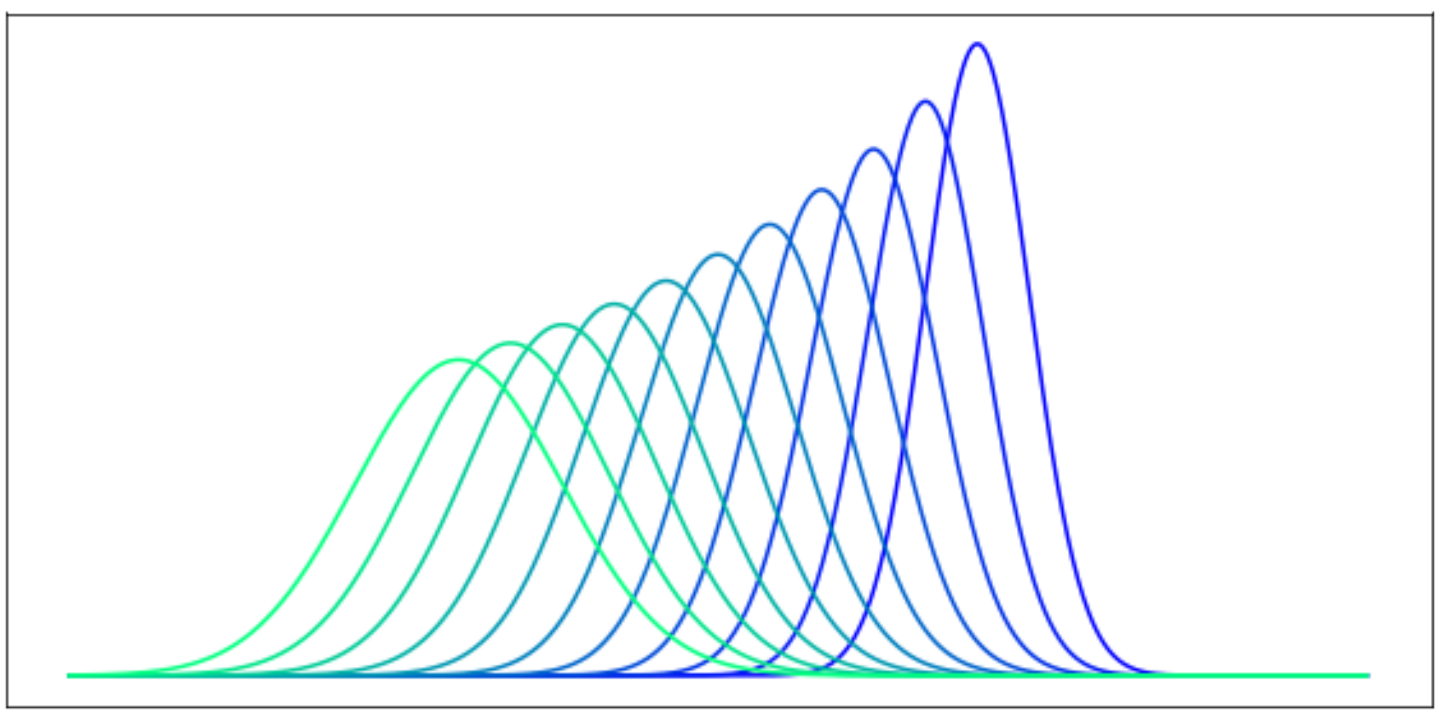
Optimal transport between Gaussians

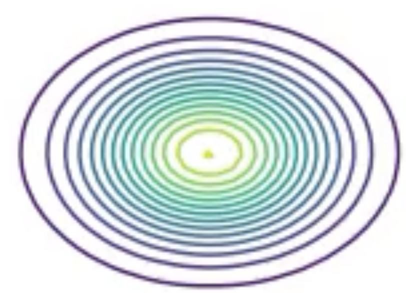
$\mu_i = \mathcal{N}(m_i, \Sigma_i), i \in \{0,1\}$ two Gaussian distributions on $\mathbb{R}^d$

$$W_2^2(\mu_0,\mu_1) = \|m_0 - m_1\|^2 + \underbrace{\mathrm{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)}_{B^2(\Sigma_0,\Sigma_1)}$$

If Σ_0 non-singular, affine optimal map

$$T(x) = m_1 + \Sigma_0^{-\frac{1}{2}}\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}(x - m_0)$$





Barycenters between Gaussians

$\mu_i = \mathcal{N}(m_i, \Sigma_i), i \in \{0, \dots, I-1\}$ Gaussian distributions on $\mathbb{R}^d$

Barycenter [Agueh, Carlier 2011]:

$$\operatorname{argmin}_\mu \sum_{i=0}^{I-1} \lambda_i W_2^2(\mu_i, \mu) = \mathcal{N}(m^*, \Sigma^*)$$

$$m^* = \sum \lambda_i m_i \qquad \Sigma^* = \min_{\Sigma} \sum_i \lambda_i \mathrm{B}(\Sigma, \Sigma_i)^2$$

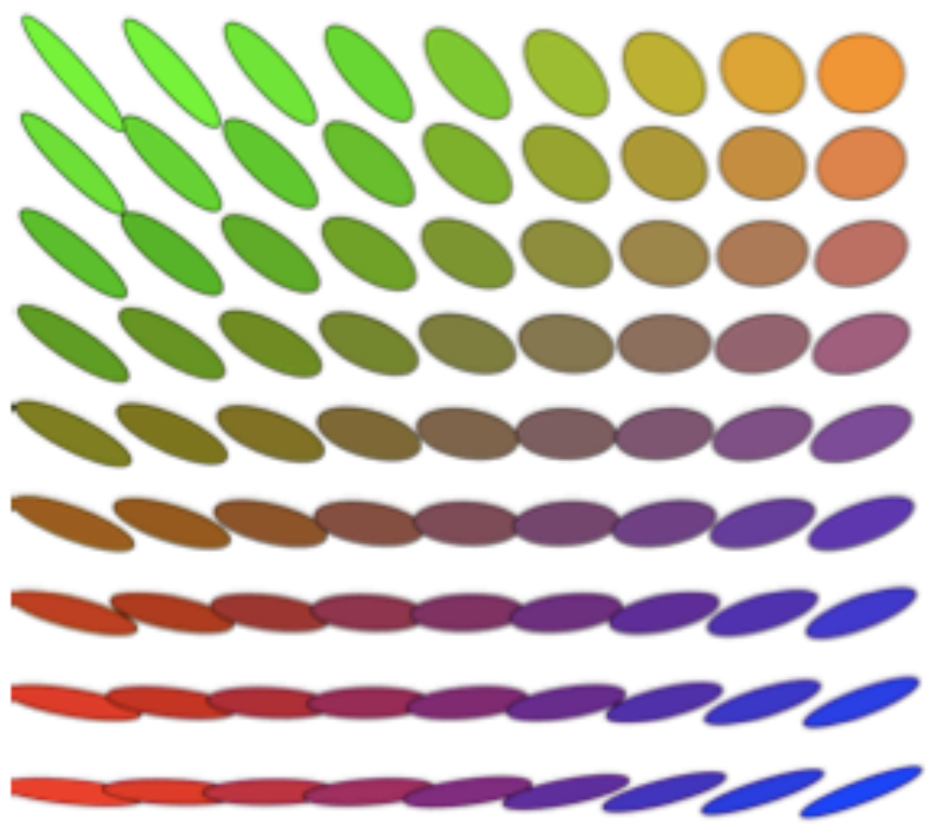
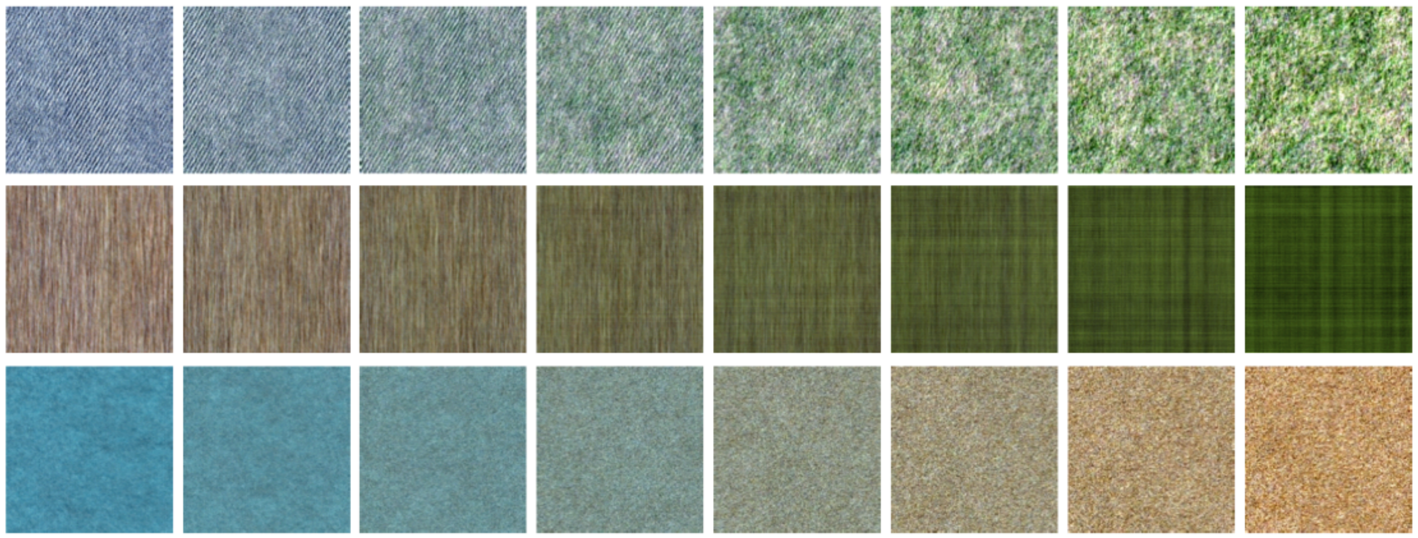


illustration: Cuturi, Peyré, OT book



Texture mixing [Xia et al, 2014]

Numerical approaches

Linear programming

Input $\mu_0 = \sum_{i=1}^{K_0} s_i\, \delta_{x_i}$, $\mu_1 = \sum_{j=1}^{K_1} t_j\, \delta_{y_j}$ **with** $\sum_i s_i = \sum_j t_j = 1$

(LP) $\operatorname{argmin}_{\gamma \in \Pi(\mu_0,\mu_1)} \sum_{i,j} c_{i,j}\gamma_{i,j}$ **with**

$$\Pi(\mu_0,\mu_1) = \left\{ \textbf{matrices } \gamma \textbf{ s.t. } \gamma_{i,j} \geq 0, \sum_i \gamma_{i,j} = t_j, \sum_j \gamma_{i,j} = s_i \right\}$$

One solution has less than $K_0 + K_1 - 1$ values $\neq 0$

Assignment: Hungarian algo. [Kuhn 1995] $O(N^3)$, Auction [Bertsekas 1992]

LP: Network Simplex [Cunningham 1976]$O(N^3)$

Dynamic formulation [Brenier, Benamou 2000]

Semi-discrete OT [Mérigot 11, Levy 15]

Sliced OT [Rabin et al. 11, Rabin et al.15]

Entropic OT [Cuturi 13,...]

Sliced optimal transport

Replace classical OT by

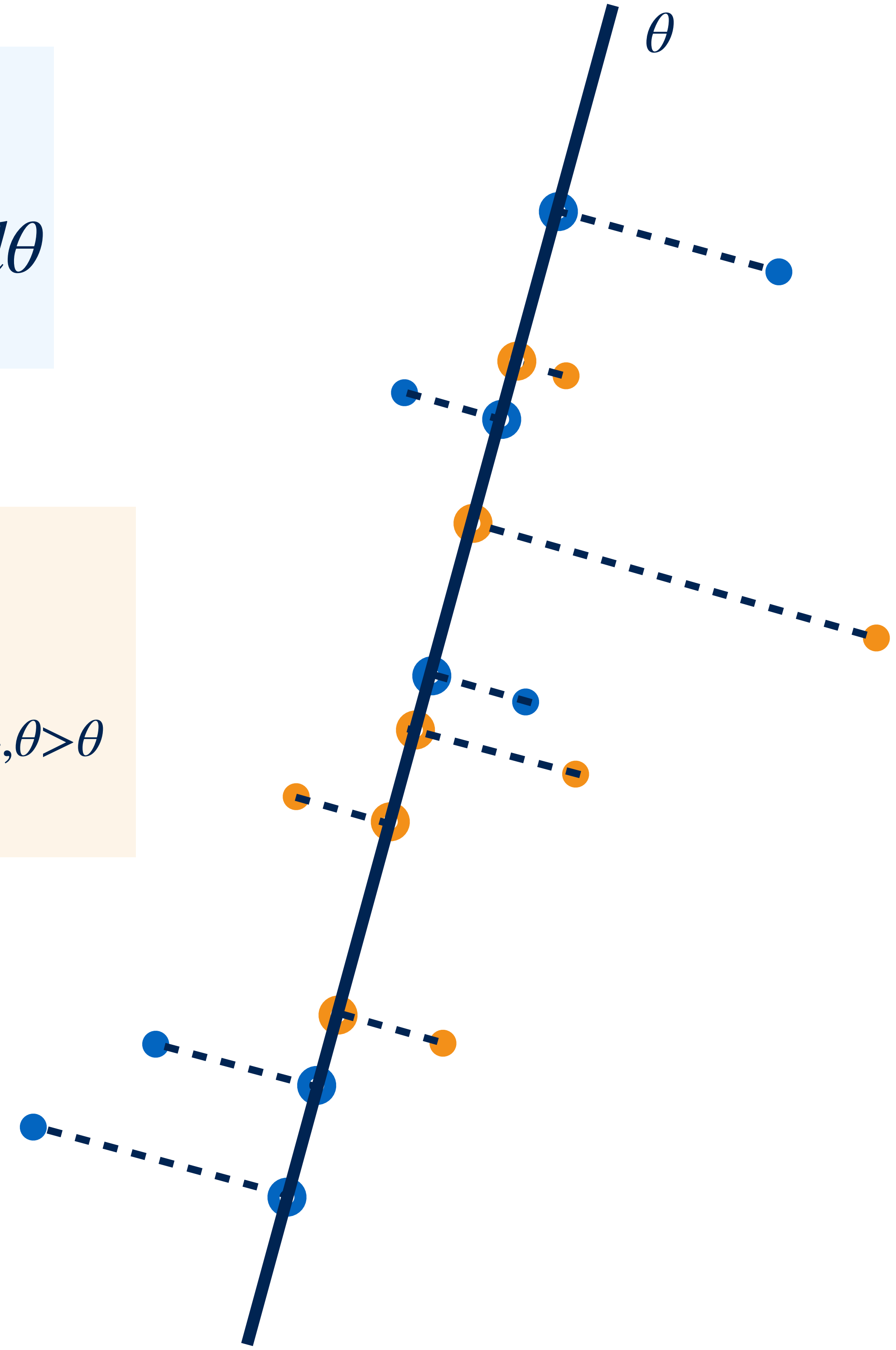
$$SW_2^2(\mu_0, \mu_1) = \int_{\mathbb{S}^{d-1}} W_2^2(p_\theta \# \mu_0, p_\theta \# \mu_1) d\theta$$

<u>**Discrete measures**</u>

$$\mu_0 = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \ , \ \mu_1 = \frac{1}{n} \sum_{j=1}^{n} \delta_{y_j}, \ p_\theta \# \mu_0 = \frac{1}{n} \sum_{i} \delta_{<x_i, \theta> \theta}$$

$$SW_2^2(\mu_0, \mu_1) = \int_{\mathbb{S}^{d-1}} \sum_i |\langle x_i - y_{\sigma_\theta(i)}, \theta \rangle|^2 d\theta$$

with σ_θ monotone rearrangement between $< \mu_0, \theta >$ and $< \mu_1, \theta >$.

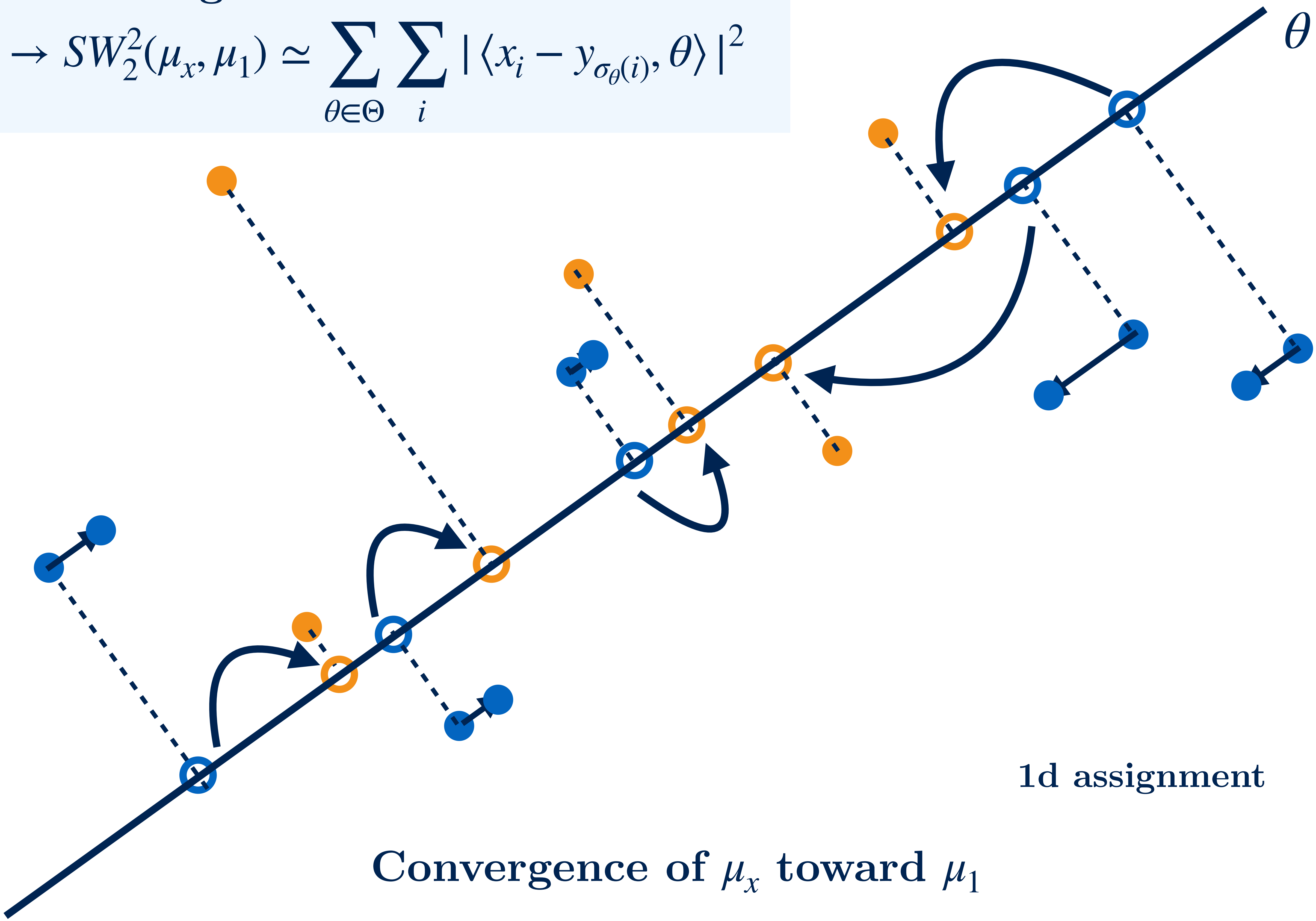


[Rabin et al. 11, Rabin et al. 2015]

Assignment with Sliced OT

Stochastic gradient descent on

$$x \to SW_2^2(\mu_x, \mu_1) \simeq \sum_{\theta \in \Theta} \sum_i |\langle x_i - y_{\sigma_\theta(i)}, \theta \rangle|^2$$



Convergence of μ_x toward μ_1

Entropic OT

Entropy of the matrix γ $\quad H(\gamma) = \sum_{i,j} \gamma_{i,j}(\log(\gamma_{i,j}) - 1)$

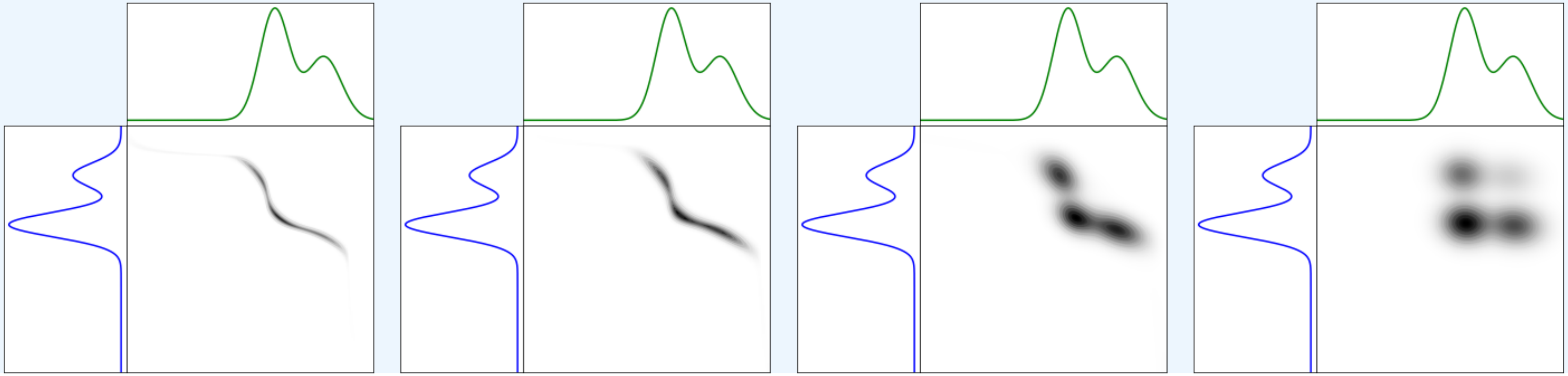
Entropic OT [Cuturi '13]

$$\underset{\gamma \in \Pi(\mu_0,\mu_1)}{\operatorname{argmin}} \sum_{i,j} c(x_i, y_j)\gamma_{i,j} - \varepsilon H(\gamma)$$

With $K_{i,j} = e^{-\frac{1}{\varepsilon}c(x_i,y_j)}$ the pb becomes

$$\underset{\gamma \in \Pi(\mu_0,\mu_1)}{\operatorname{argmin}} \sum_{i,j} \gamma_{i,j} \log\left(\frac{\gamma_{i,j}}{K_{i,j}}\right) = \underset{\gamma \in \Pi(\mu_0,\mu_1)}{\operatorname{argmin}} \operatorname{KL}(\gamma || K)$$

Sinkhorn algorithm = alternate projections of K on $\Pi(\mu_0, \mu_1)$



$\varepsilon = 3 \times 10^{-4}$ $\quad\varepsilon = 10^{-3}$ $\quad\varepsilon = 10^{-2}$ $\quad\varepsilon = 10^{-1}$

Sinkhorn algorithm

Prop: solution γ of $\underset{\gamma\in\Pi(\mu_0,\mu_1)}{\operatorname{argmin}} \mathrm{KL}(\gamma||K)$ satisfies $\gamma = \mathrm{diag}(a) K \,\mathrm{diag}(b)$

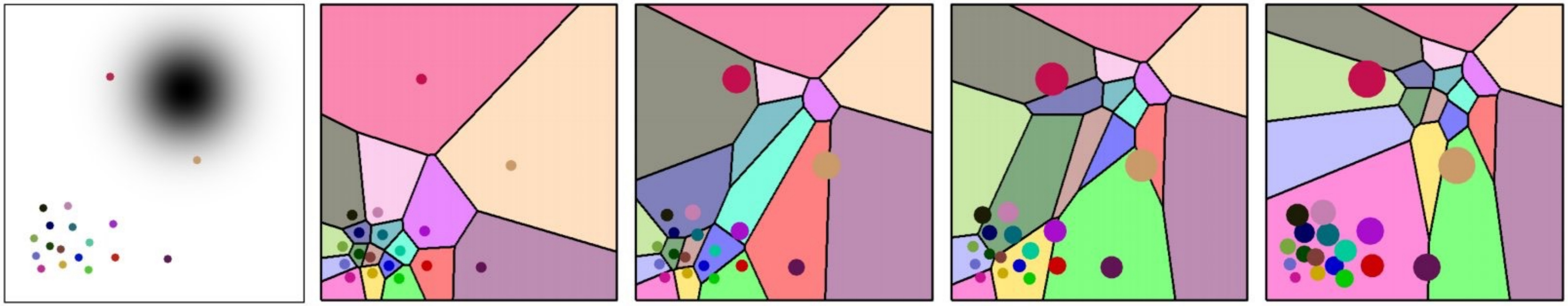
Since $\gamma \in \Pi(\mu_0, \mu_1)$, it implies that
$$\begin{aligned} a \odot Kb &= \mu_0 \\ b \odot K^T a &= \mu_1 \end{aligned}$$

Iterations: $a \leftarrow \dfrac{\mu_0}{Kb} \quad b \leftarrow \dfrac{\mu_1}{K^T a}$

- Iterative projections on the constraints.
- Simple extension to compute barycenters of more than 2 measures
- Matrix-vector multiplications
- For regular grids, products Kx can be written as convolutions.
- Numerical pb when $\varepsilon \to 0$.

Barycenters between superheroes!

Semi-discrete OT ?



α and β $\quad$ $\ell = 1$ $\quad$ $\ell = 20$ $\quad$ $\ell = 40$ $\quad$ $\ell = 100$

Duality

Duality

Under mild conditions on the cost c (l.s.c.),

$$W_c(\mu_0,\mu_1) = \sup_{\phi,\psi\in\Phi_c(\mu_0,\mu_1)} \int \phi d\mu_0 + \int \psi d\mu_1, \text{ where}$$

$$\Phi_c(\mu_0,\mu_1) = \{\phi,\psi \in L^1 \text{ s.t. } \forall x,y,\ \phi(x)+\psi(y) \leq c(x,y)\}.$$

ϕ, ψ are called Kantorovich potentials

c-transform

$$\phi^c(y) = \inf_x c(x,y) - \phi(y)$$

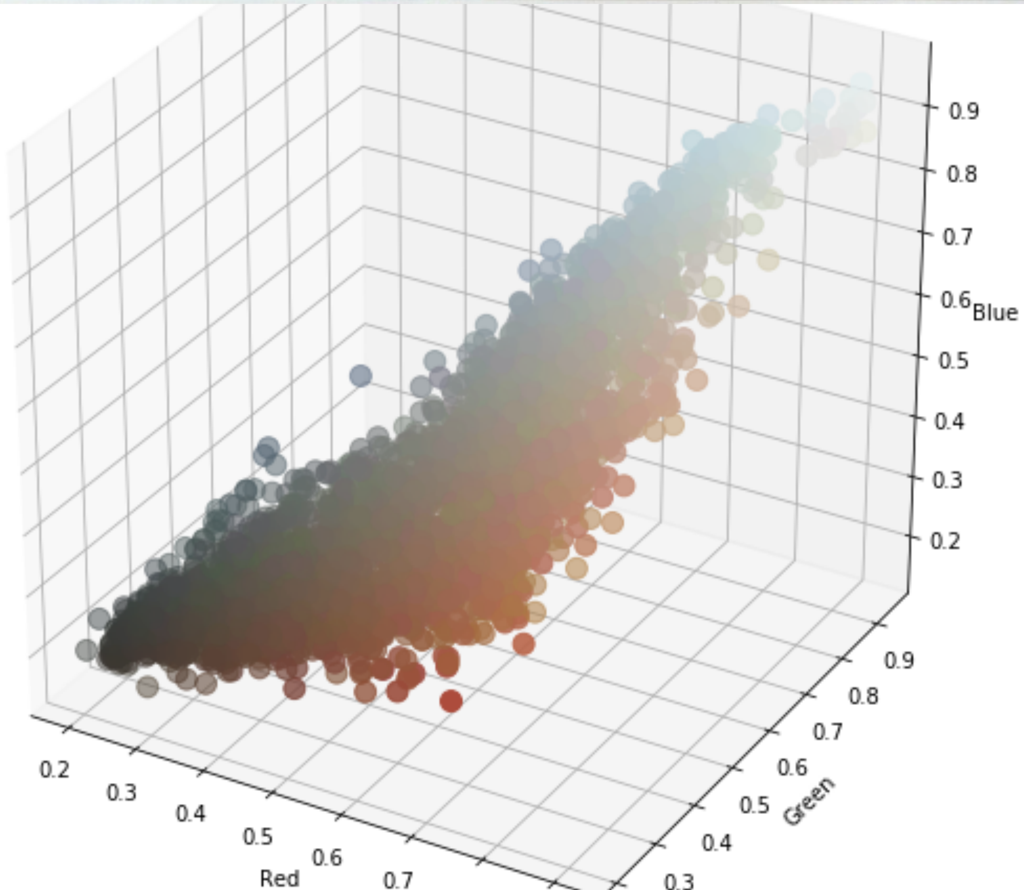
$$W_c(\mu_0,\mu_1) = \sup_{\phi \text{ c-concave}} \int \phi d\mu_0 + \int \phi^c d\mu_1$$

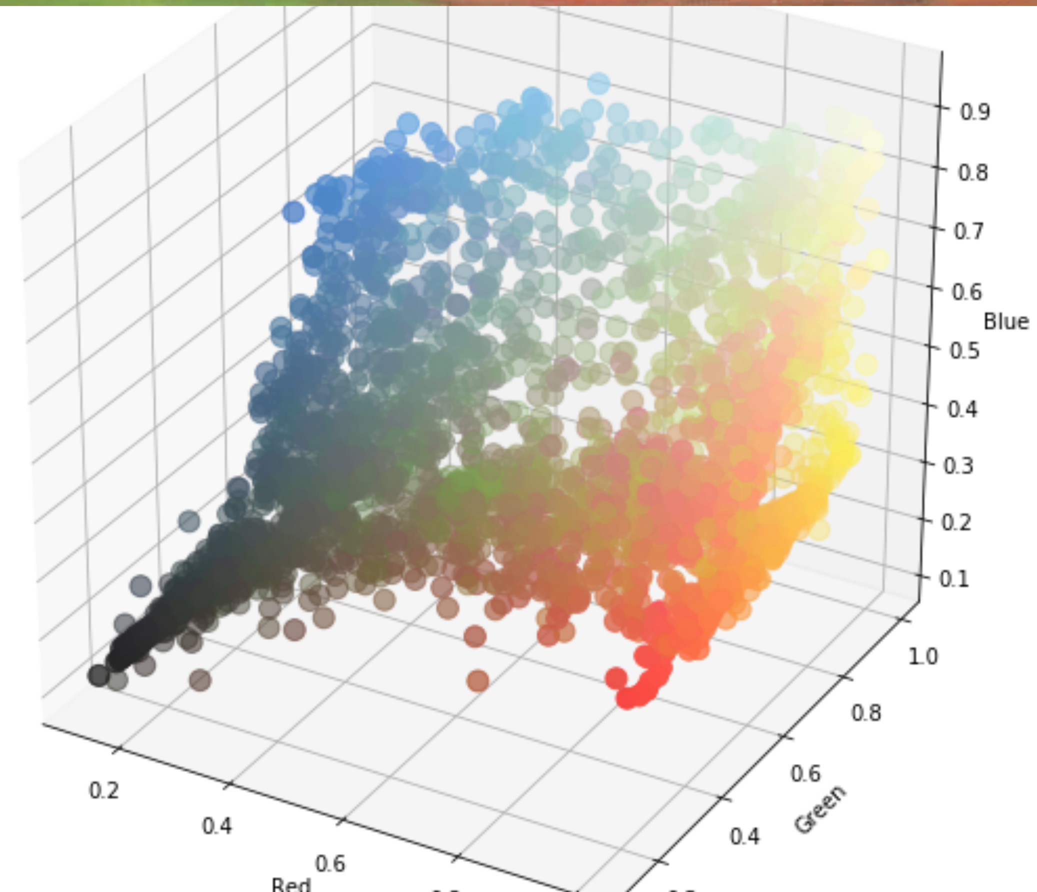
If c is a distance, then

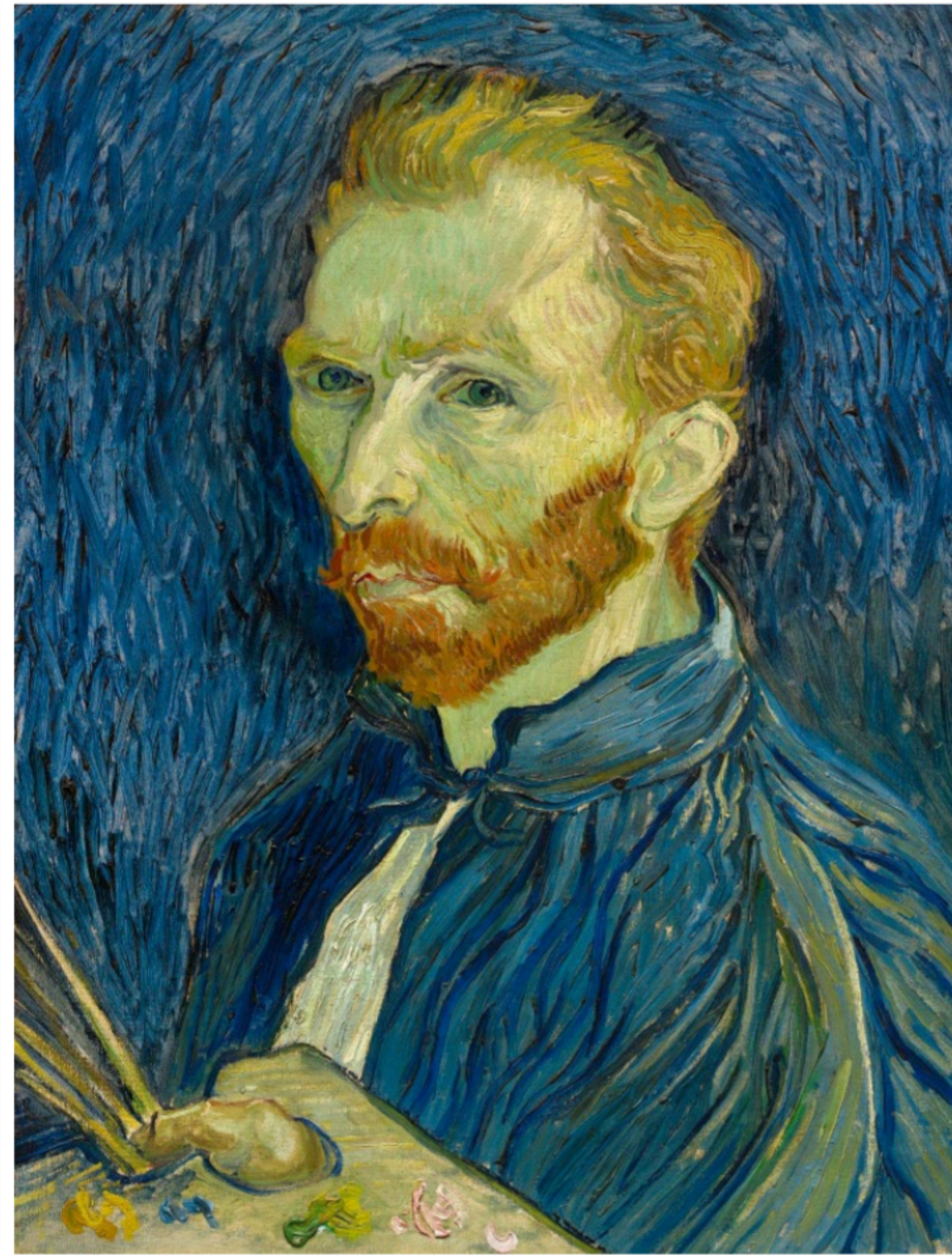
$$W_c(\mu_0,\mu_1) = \sup_{\phi Lip_1} \int \phi d\mu_0 - \int \phi d\mu_1$$ $\rightarrow$Wasserstein GANs !!!

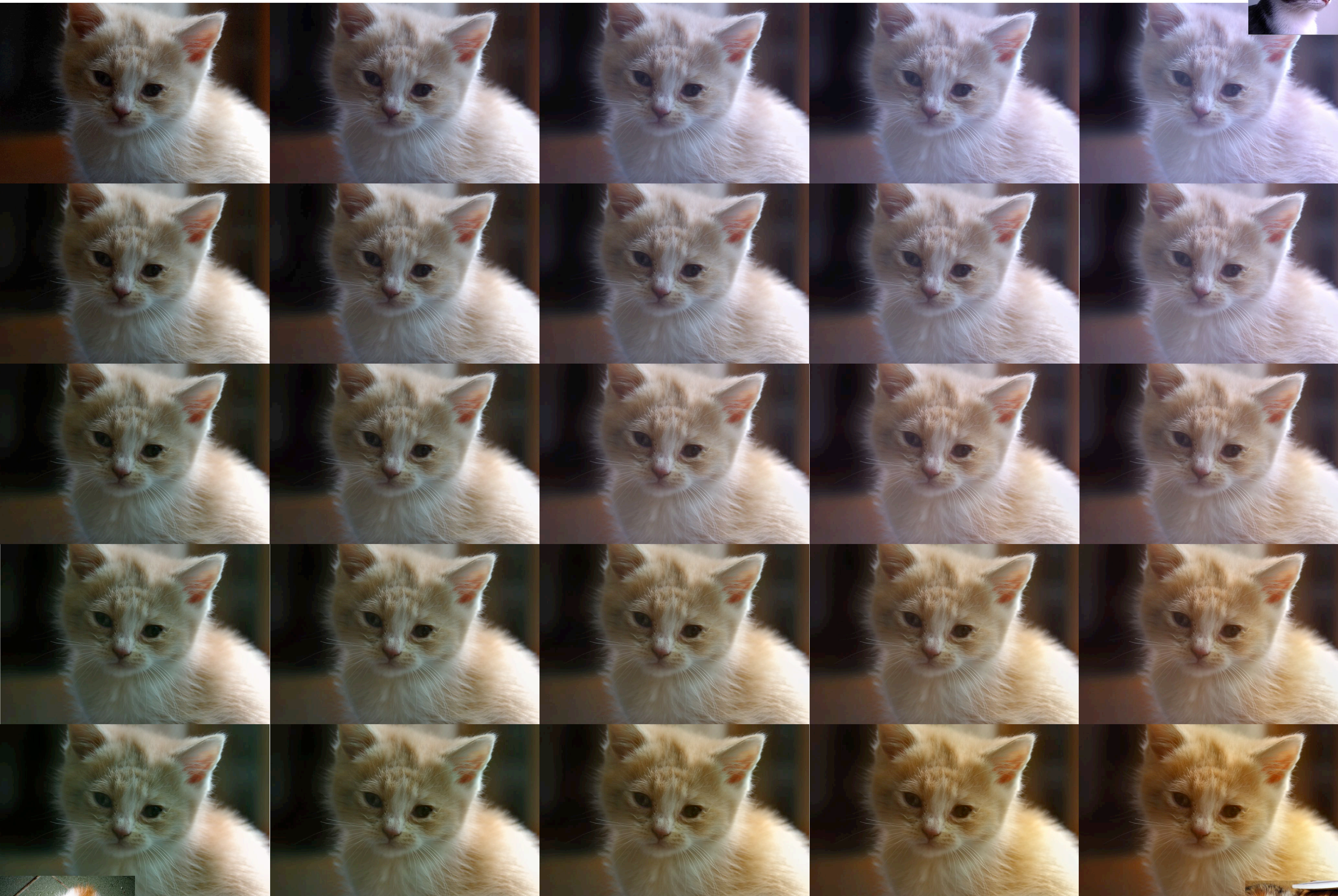
Some applications

Color transfer

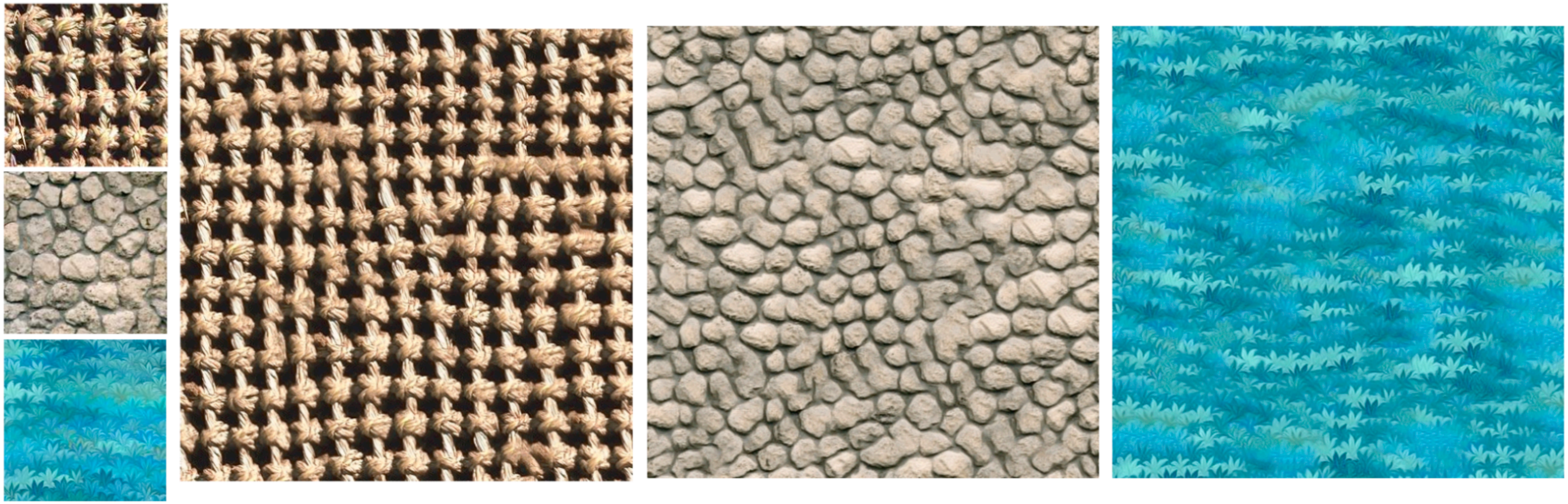






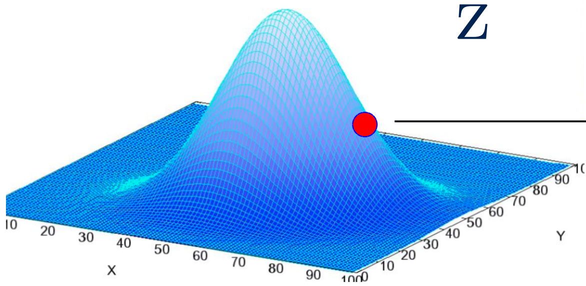


Texture synthesis and style transfer



[Leclaire, Rabin 2019]

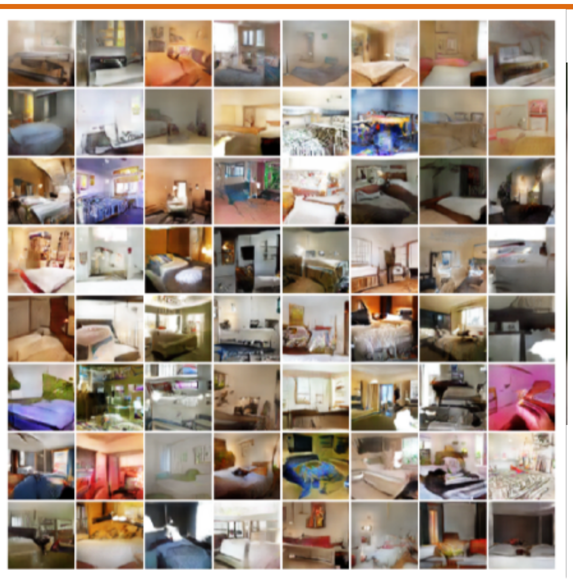
Generative networks



Z

Decoding

Probabilistic model in latent space



Generative Adversarial Networks (GANs) [Goodfellow et al. 2014]

WGAN [Arjowsky et al. 2017]

$$\min_G W_1^1(G(z), \mu_1) \textbf{ with } z \sim \mathcal{N}(0, I)$$

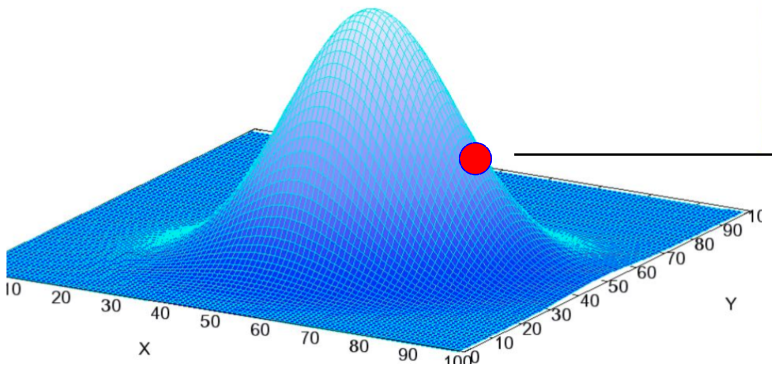
- minimizes the W_1^1 distance between the generated data and the database distribution μ_1
- Wasserstein distance dual computation

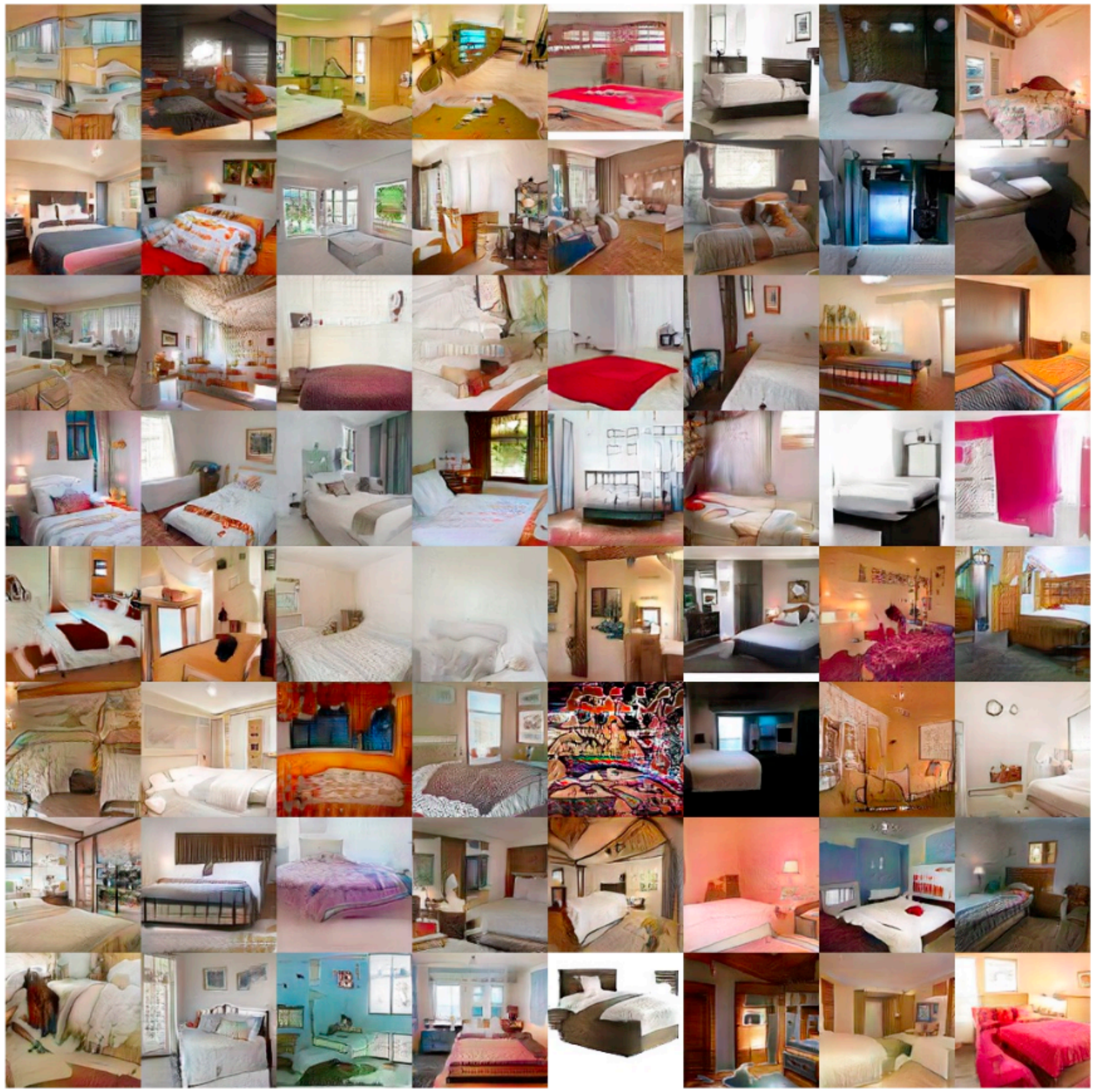
$$\min_G \sup_{\phi \in Lip_1} \mathbb{E}_{\mu_1}[\phi(X)] - \mathbb{E}_{Z \sim \mathcal{N}(0,I)}[\phi(G(Z))]$$

- avoid vanishing gradients of [Goodfellow et al. 2014]

Wasserstein GANs

Generative models

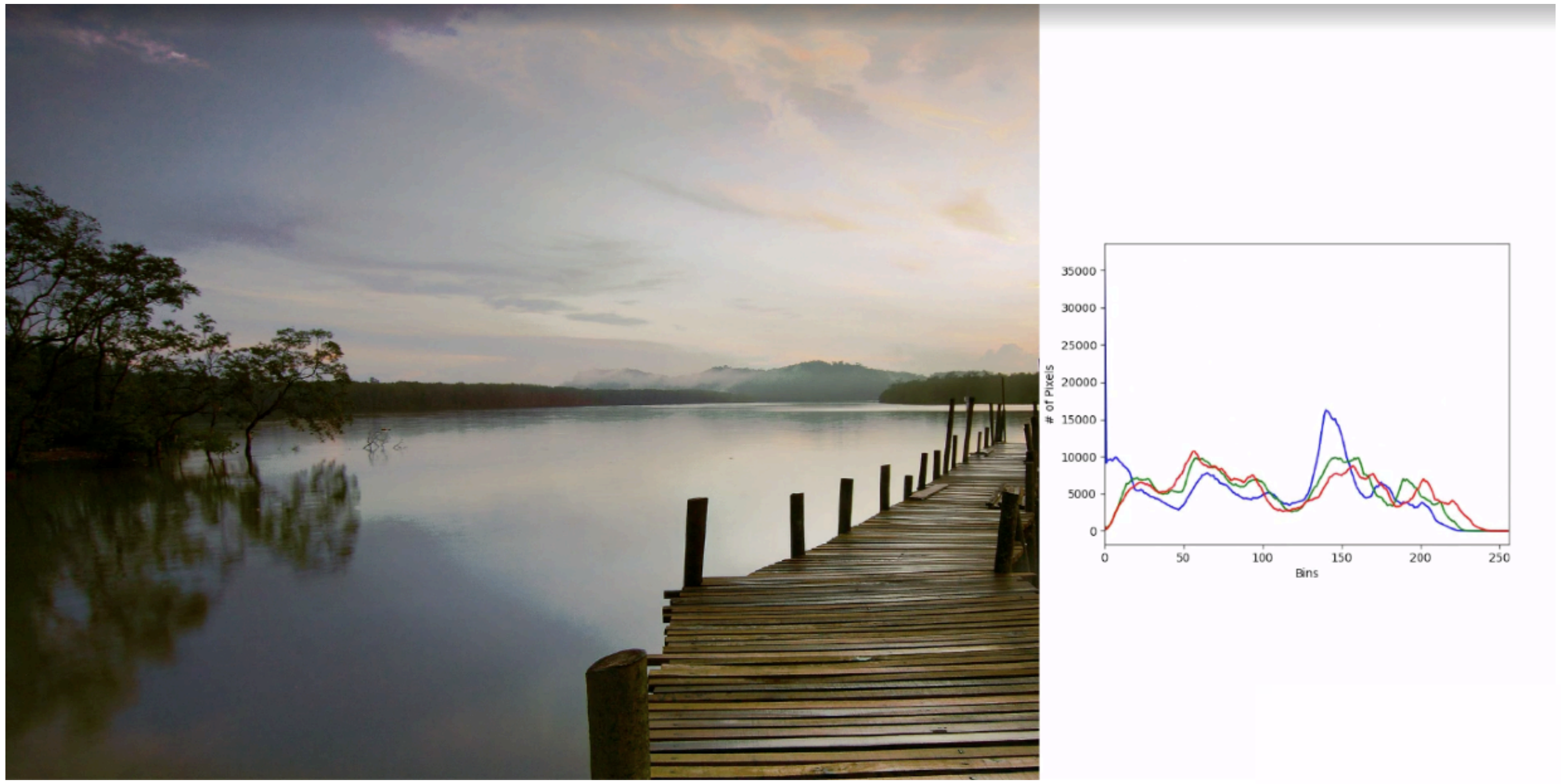


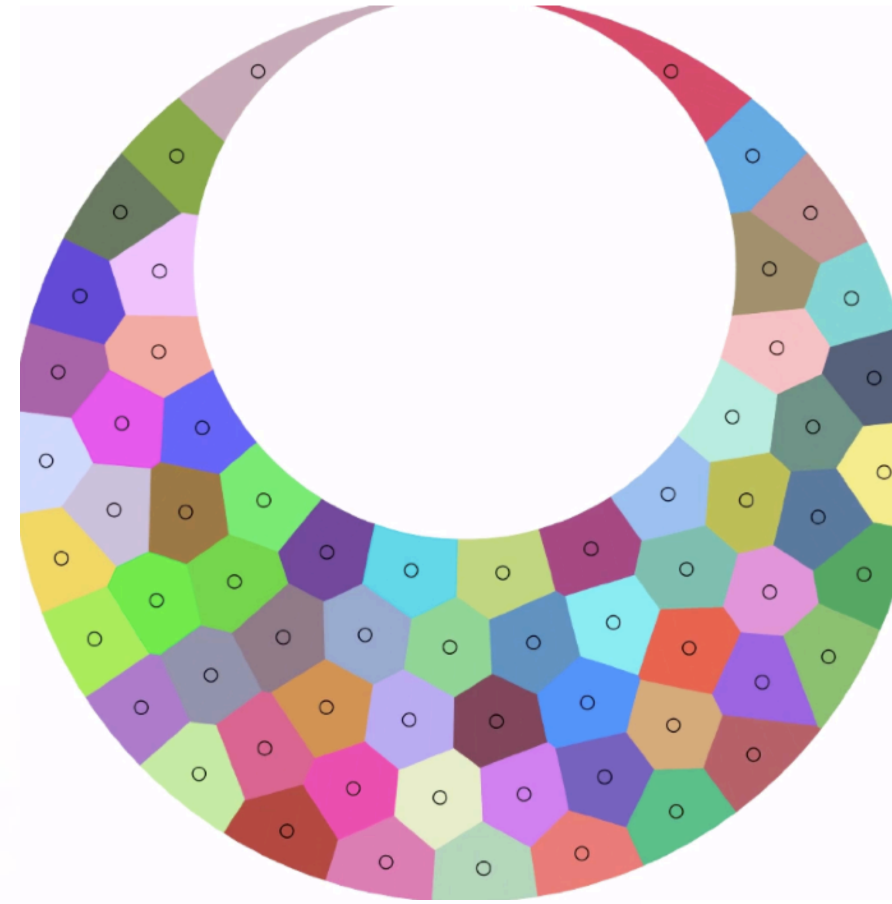


[Gulrajani et al., 2017]

[Karras et al. 2018]

Conclusion


35000
30000
25000
20000
15000
10000
5000
0
# of Pixels
0
50
100
150
200
250
Bins




GIF