Computational Optimal Transport

Gabriel Peyré **Marco Cuturi** **Nicolas Courty**

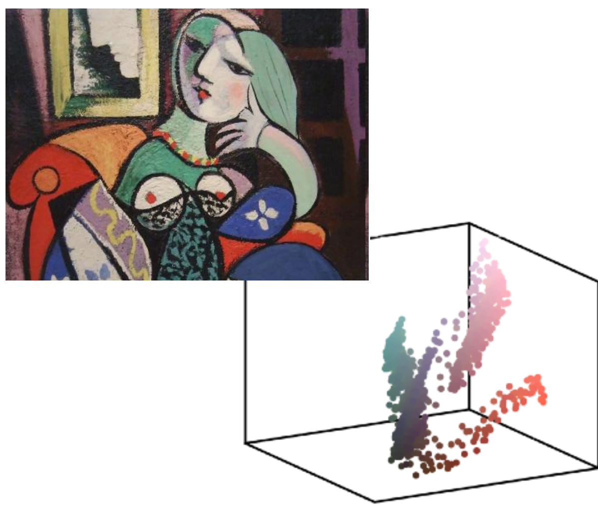
CNRS
ENS École Normale Supérieure

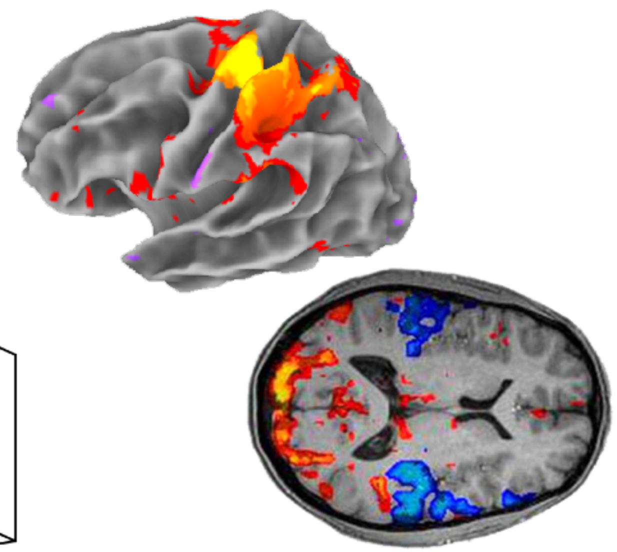
ENSAE ParisTech
École nationale de la statistique et de l'administration économique
Université Paris-Saclay

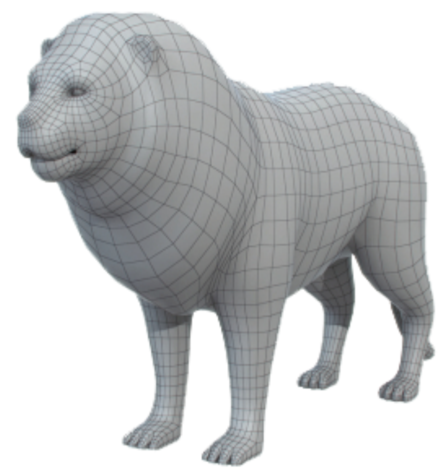
UMR IRISA

Comparing Probability Distributions

- *Probability distributions and histograms*
  $\rightarrow$ images, vision, graphics and machine learning, ...

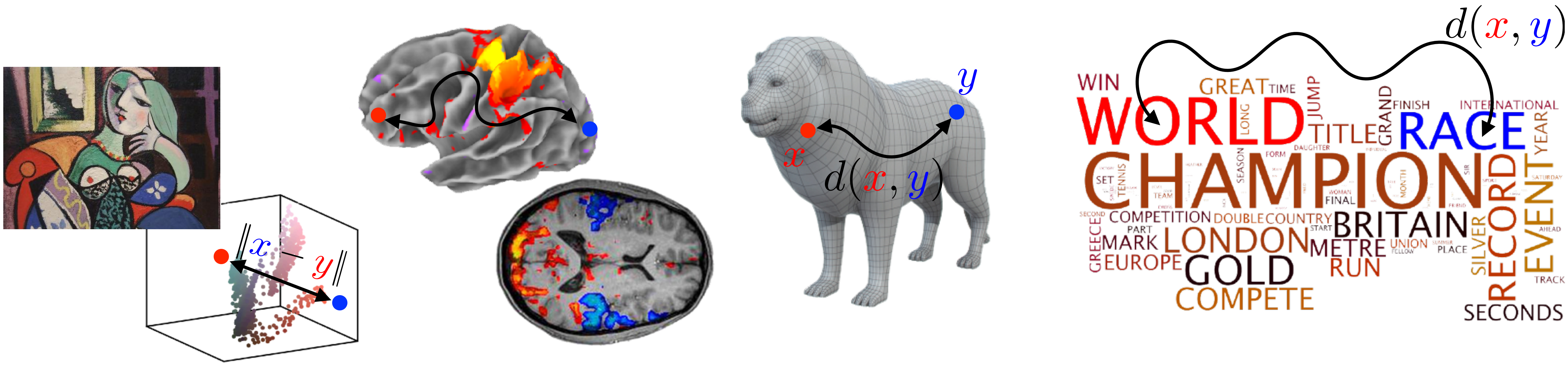






Comparing Probability Distributions

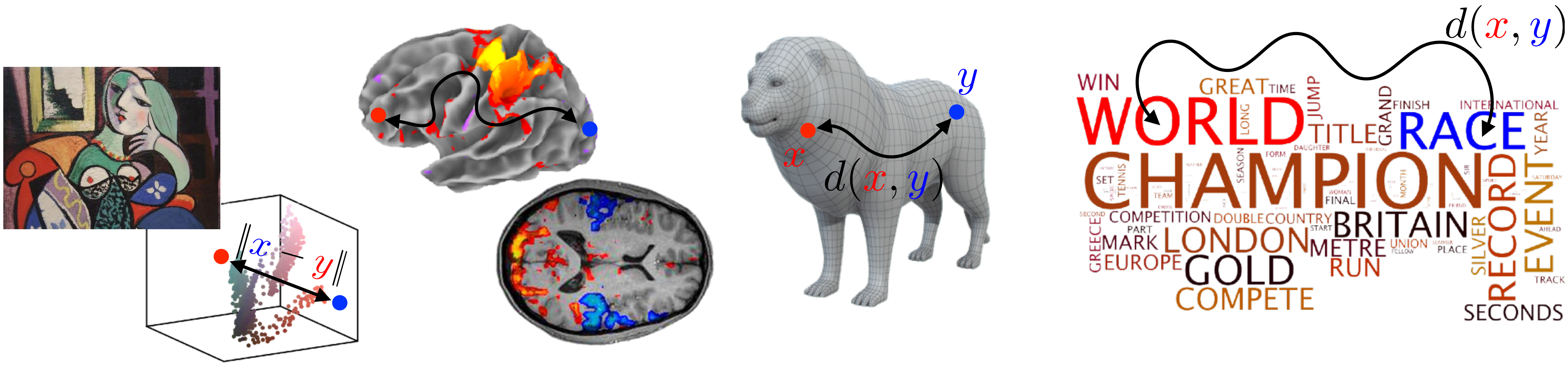
- *Probability distributions and histograms*
  $\rightarrow$ images, vision, graphics and machine learning, ...



- *Optimal transport*
  $\rightarrow$ takes into account a metric d.

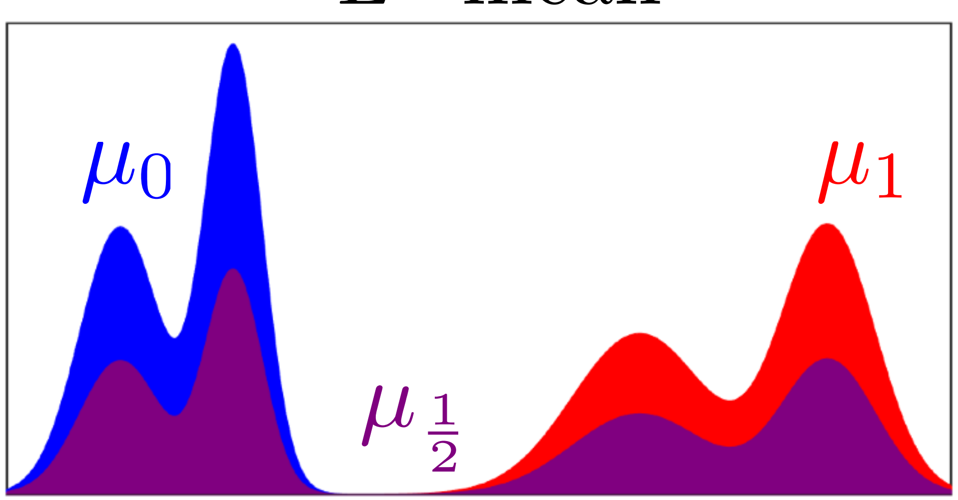
Comparing Probability Distributions

- *Probability distributions and histograms*
  → images, vision, graphics and machine learning, ...

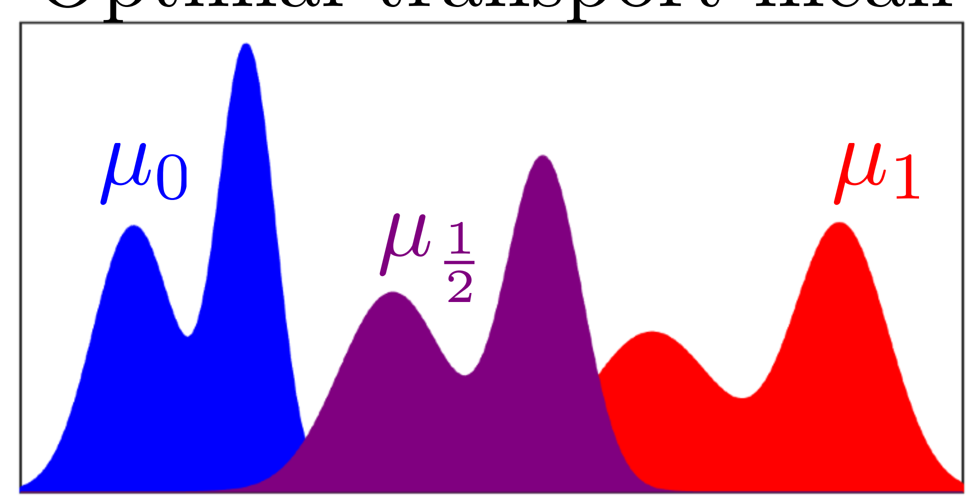


- *Optimal transport*
  → takes into account a metric d.

L^2 mean



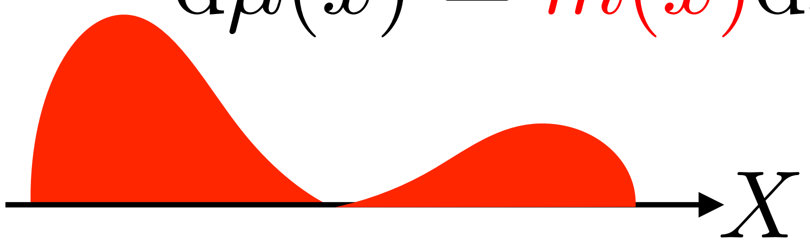
Optimal transport mean



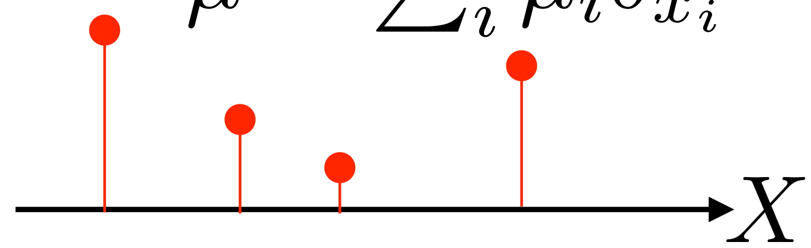
Probability Measures

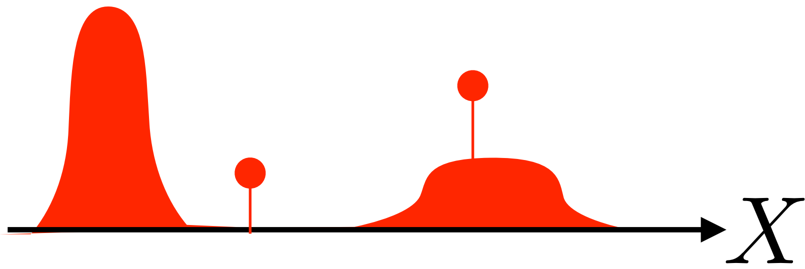
Positive Radon measure μ *on a set* X.

$$\mathrm{d}\mu(x) = m(x)\mathrm{d}x$$



$$\mu = \sum_i \mu_i \delta_{x_i}$$



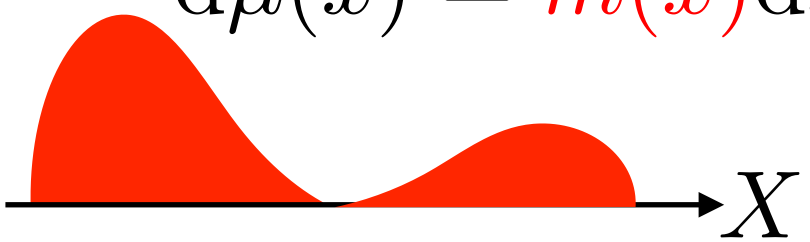


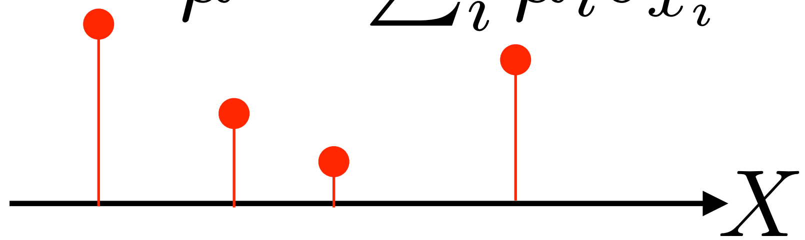
Measure of sets $A \subset X$: $\mu(A) = \int_A \mathrm{d}\mu(x) \geqslant 0$

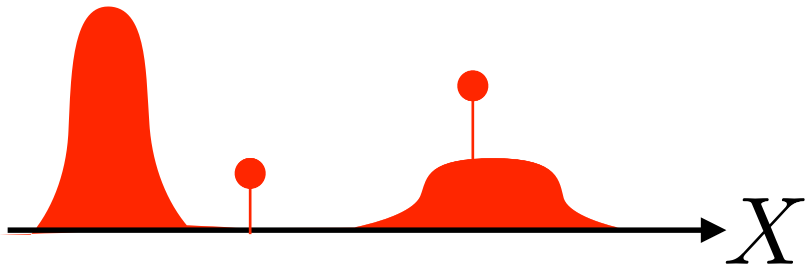
Probability Measures

Positive Radon measure μ *on a set* X*.*

$$\mathrm{d}\mu(x) = {\color{red} m(x)}\mathrm{d}x \qquad \mu = \sum_i \mu_i \delta_{x_i}$$







Measure of sets $A \subset X$: $\mu(A) = \int_A \mathrm{d}\mu(x) \geqslant 0$

Integration against continuous functions: $\int_X g(x)\mathrm{d}\mu(x) \geqslant 0$

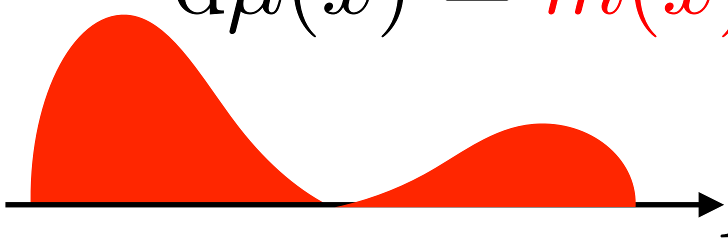
$$\mathrm{d}\mu(x) = {\color{red} m(x)}\mathrm{d}x \longrightarrow \int_X g\mathrm{d}\mu = \int_X m(x)\mathrm{d}x$$

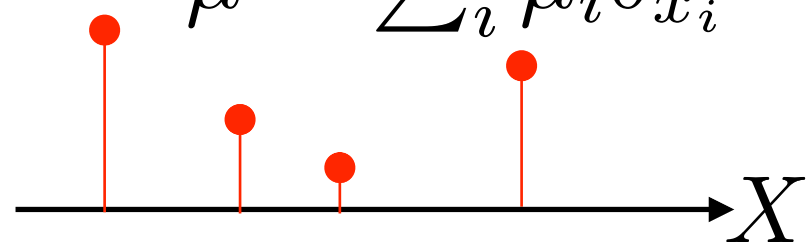
$$\mu = \sum_i \mu_i \delta_{x_i} \longrightarrow \int_X g\mathrm{d}\mu = \sum_i \mu_i g(x_i)$$

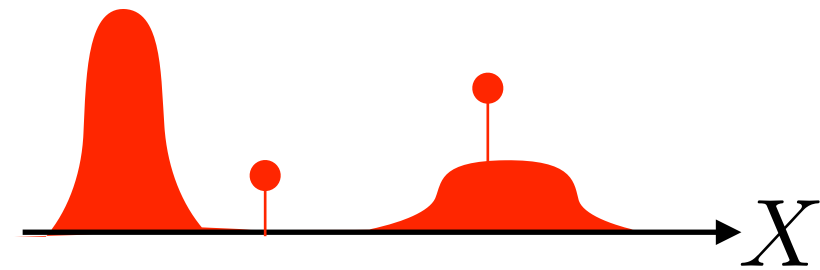
Probability Measures

Positive Radon measure μ *on a set* X.

$$\mathrm{d}\mu(x) = {\color{red}m(x)}\mathrm{d}x \qquad \mu = \sum_i \mu_i \delta_{x_i}$$







Measure of sets $A \subset X$: $\mu(A) = \int_A \mathrm{d}\mu(x) \geqslant 0$

Integration against continuous functions: $\int_X g(x)\mathrm{d}\mu(x) \geqslant 0$

$$\mathrm{d}\mu(x) = {\color{red}m(x)}\mathrm{d}x \longrightarrow \int_X g\mathrm{d}\mu = \int_X m(x)\mathrm{d}x$$

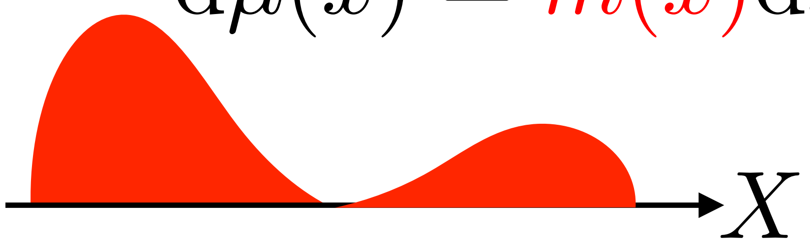
$$\mu = \sum_i \mu_i \delta_{x_i} \longrightarrow \int_X g\mathrm{d}\mu = \sum_i \mu_i g(x_i)$$

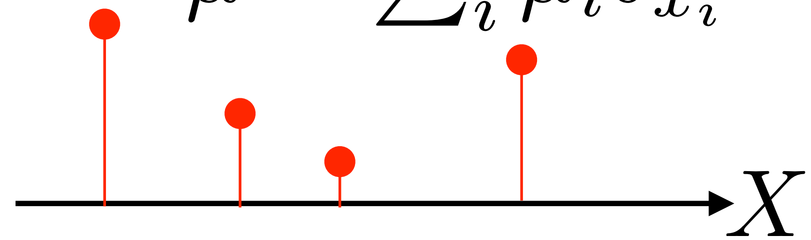
Probability (normalized) measure: $\mu(X) = \int_X \mathrm{d}\mu(x) = 1$

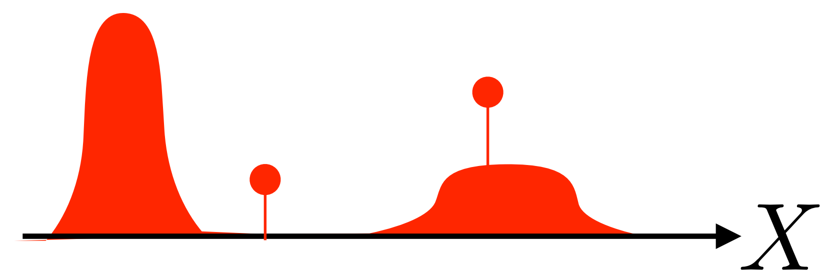
Probability Measures

Positive Radon measure μ *on a set* X.

$\mathrm{d}\mu(x) = m(x)\mathrm{d}x$ $\qquad$ $\mu = \sum_i \mu_i \delta_{x_i}$







Measure of sets $A \subset X$: $\mu(A) = \int_A \mathrm{d}\mu(x) \geqslant 0$

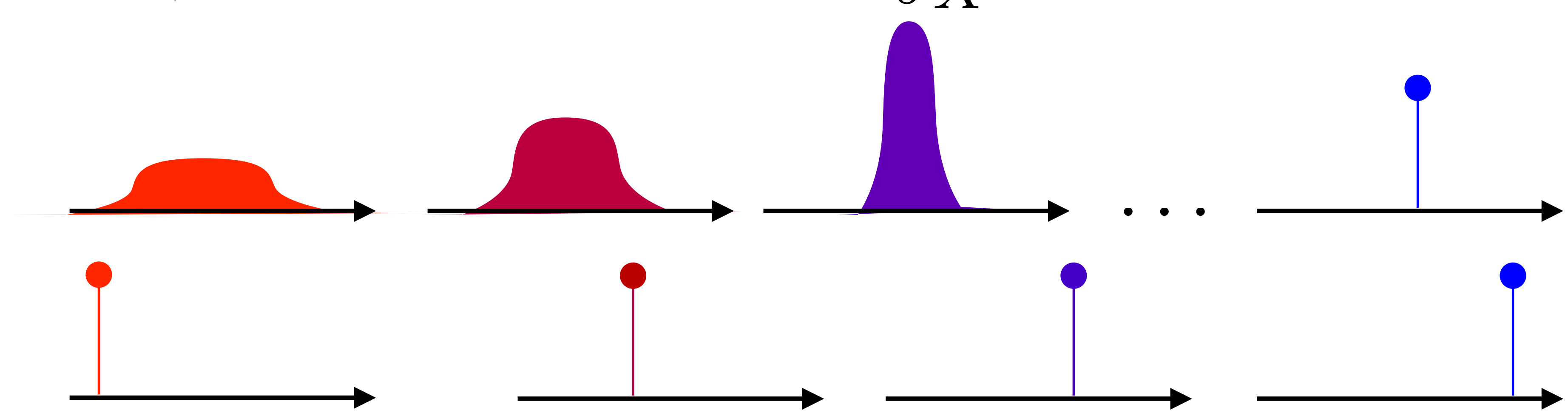
Integration against continuous functions: $\int_X g(x)\mathrm{d}\mu(x) \geqslant 0$

$$\mathrm{d}\mu(x) = m(x)\mathrm{d}x \longrightarrow \int_X g\mathrm{d}\mu = \int_X m(x)\mathrm{d}x$$

$$\mu = \sum_i \mu_i \delta_{x_i} \longrightarrow \int_X g\mathrm{d}\mu = \sum_i \mu_i g(x_i)$$

Probability (normalized) measure: $\mu(X) = \int_X \mathrm{d}\mu(x) = 1$

Weak convergence:

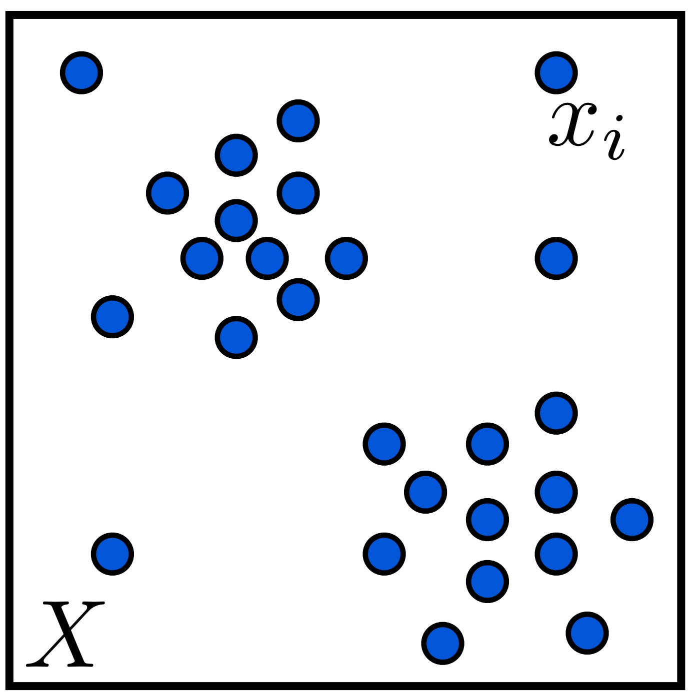


Discretization: Histogram vs. Empirical

Discrete measure: $$\mu = \sum_{i=1}^{N} \mu_i \delta_{x_i} \qquad x_i \in X, \quad \sum_i \mu_i = 1$$

Lagrangian (point clouds)

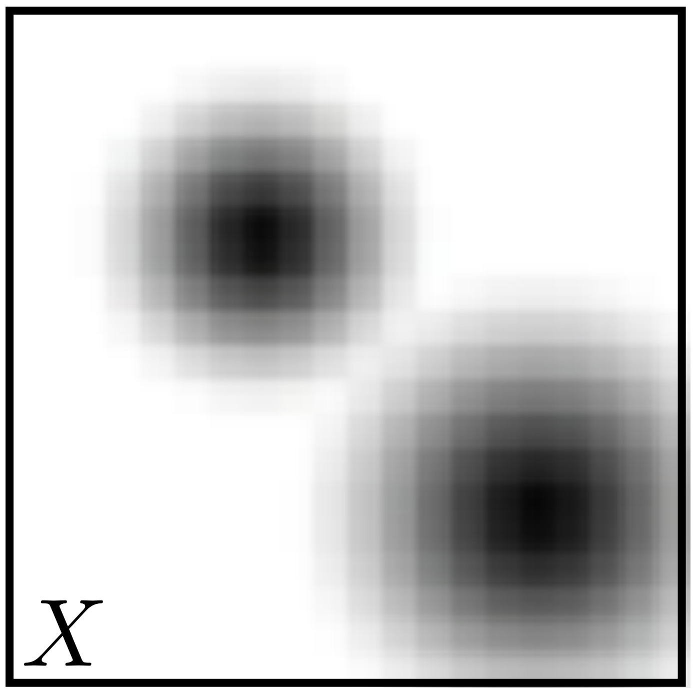
Constant weights $\mu_i = \frac{1}{N}$



Quotient space:
X^N / Σ_N

Eulerian (histograms)

Fixed positions x_i (e.g. grid)



Convex polytope (simplex):
$\{(\mu_i)_i \geqslant 0 \; ; \; \sum_i \mu_i = 1\}$

Push Forward

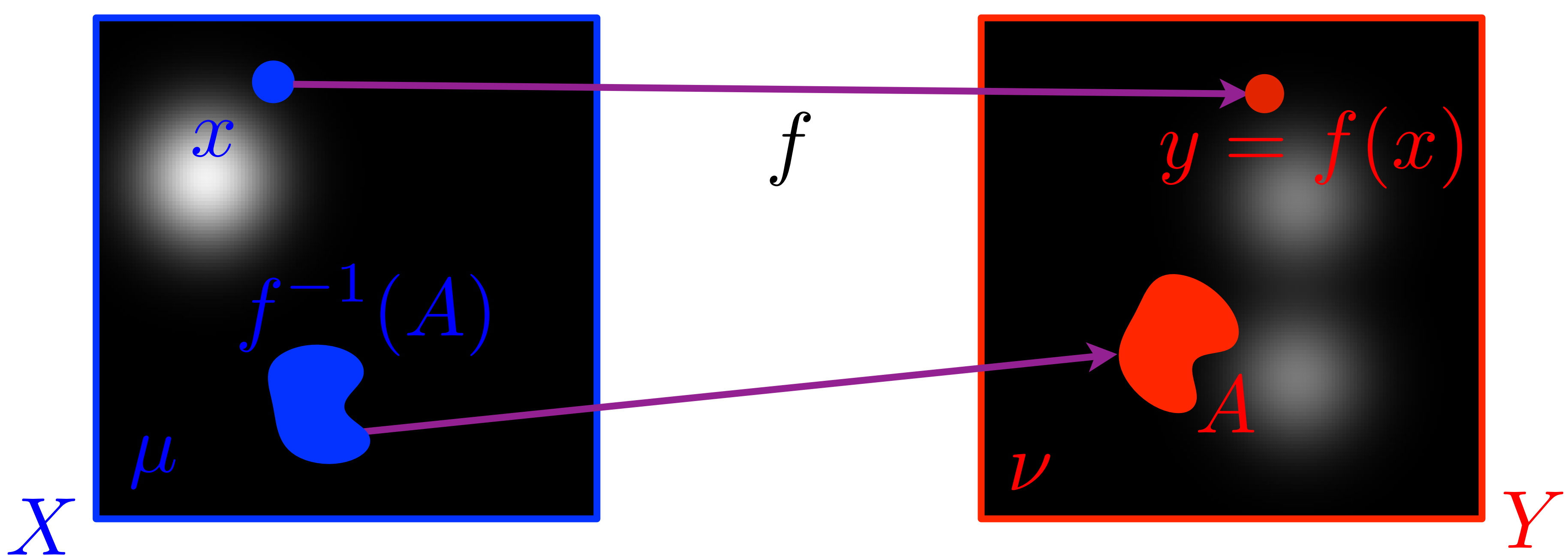
Radon measures (μ, ν) on (X, Y).

Transfer of measure by $f : X \to Y$: *push forward.*

$\nu = f_\sharp \mu$ defined by:

$$\nu(A) \overset{\text{def.}}{=} \mu(f^{-1}(A))$$

$$\iff \int_Y g(y)\mathrm{d}\nu(y) \overset{\text{def.}}{=} \int_X g(f(x))\mathrm{d}\mu(x)$$



Push Forward

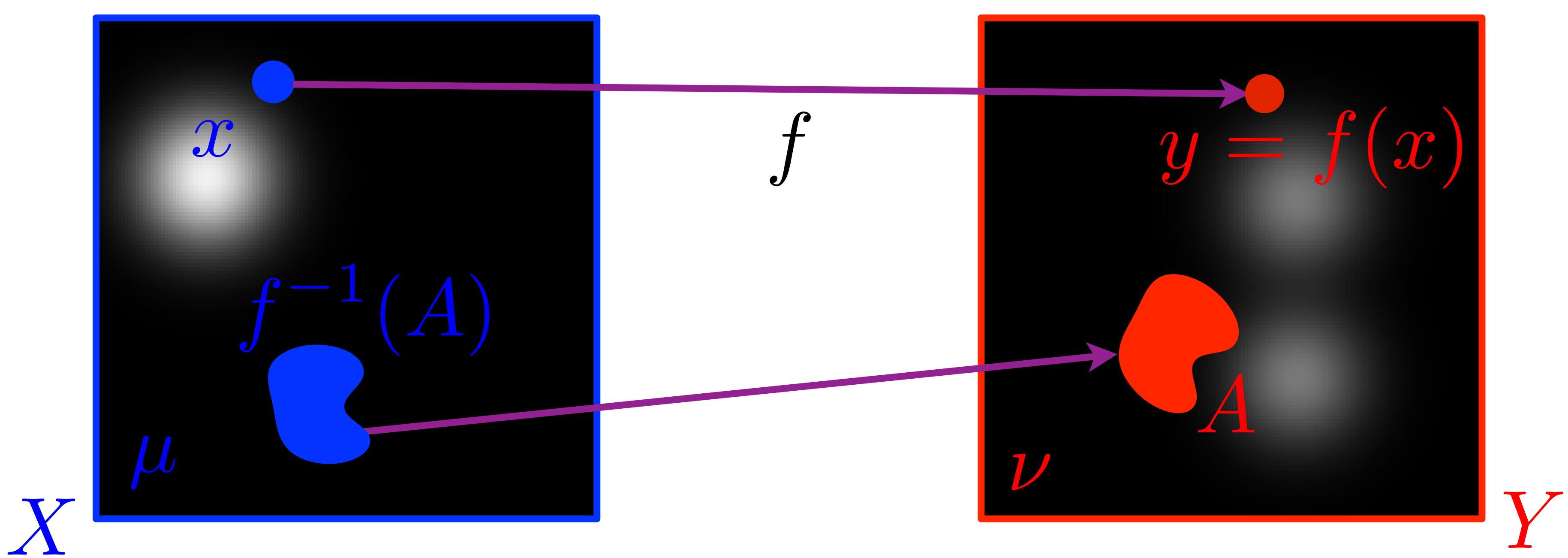
Radon measures (μ, ν) on (X, Y).

Transfer of measure by $f : X \to Y$: *push forward.*

$\nu = f_\sharp \mu$ defined by:

$$\nu(A) \overset{\text{def.}}{=} \mu(f^{-1}(A))$$

$$\iff \int_Y g(y)\mathrm{d}\nu(y) \overset{\text{def.}}{=} \int_X g(f(x))\mathrm{d}\mu(x)$$

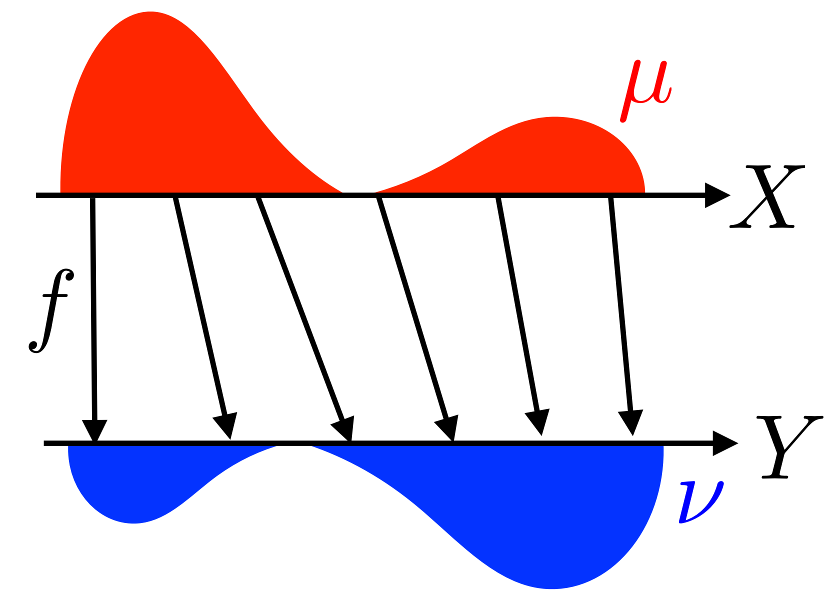


Smooth densities: $\mathrm{d}\mu = \rho(x)\mathrm{d}x$, $\mathrm{d}\nu = \xi(x)\mathrm{d}x$

$$f_\sharp \mu = \nu \iff \rho(f(x))|\det(\partial f(x))| = \xi(x)$$

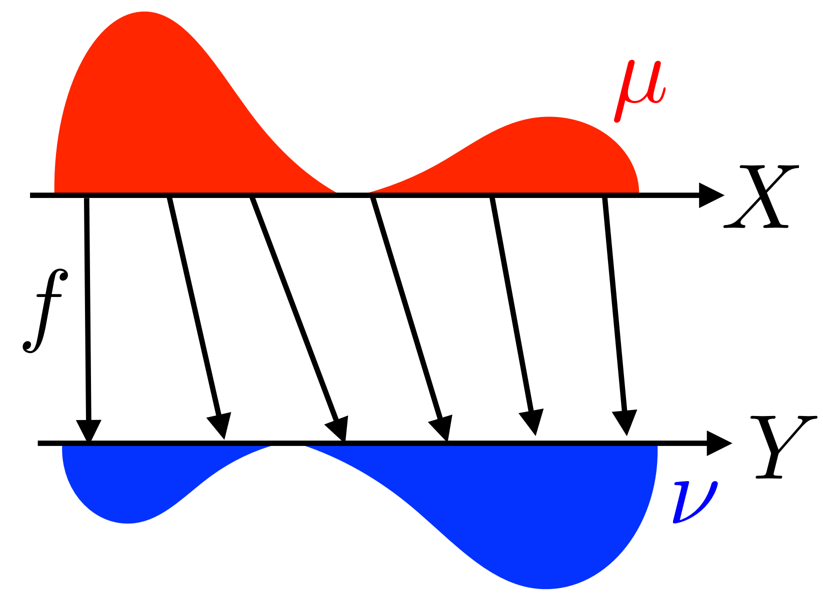
Monge Transport

$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$



Monge Transport

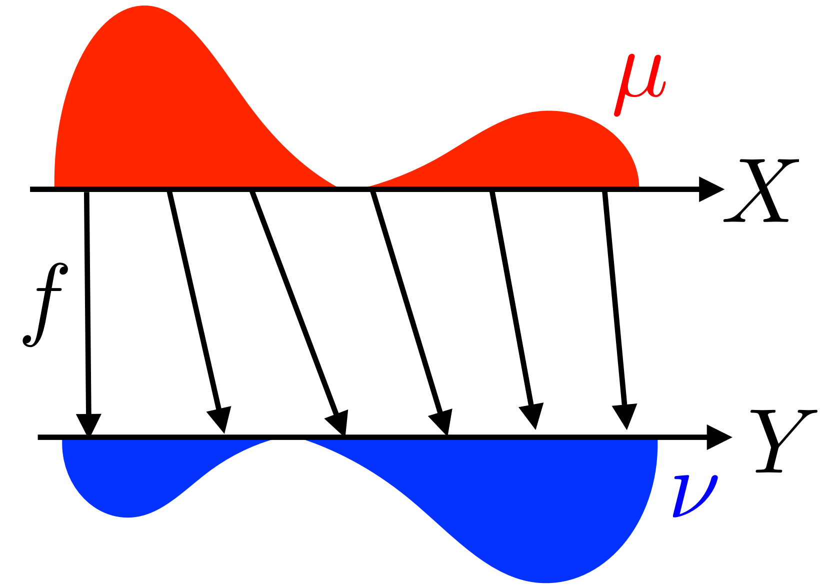
$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$



Theorem: [Brenier] for $c(x,y) = \|x-y\|^2$, (μ, ν) with density, there exists a unique optimal f. One has $f = \nabla\psi$ where ψ is the unique convex function such that $(\nabla\psi)_\sharp \mu = \nu$

Monge Transport

$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$

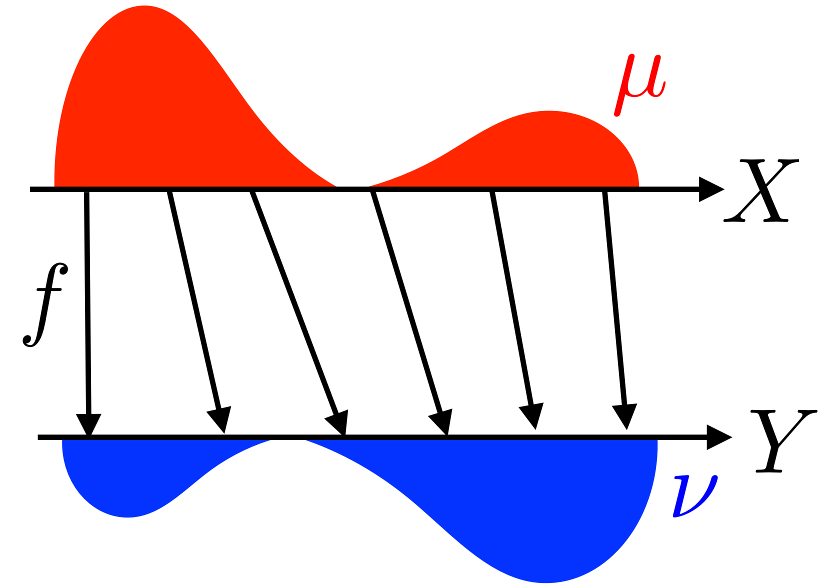


Theorem: [Brenier] for $c(x, y) = \|x - y\|^2$, (μ, ν) with density, there exists a unique optimal f. One has $f = \nabla\psi$ where ψ is the unique convex function such that $(\nabla\psi)_\sharp \mu = \nu$

Monge-Ampère equation: $\rho(\nabla\psi) \det(\partial^2 \psi) = \xi$

Monge Transport

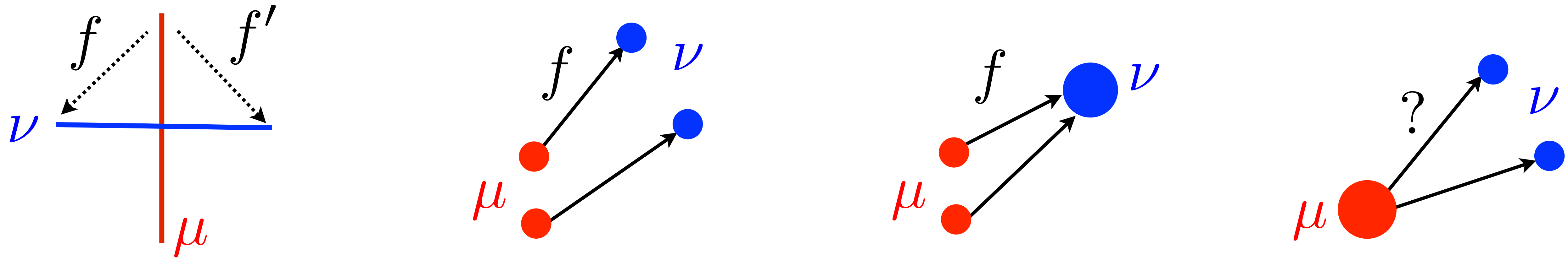
$$\min_{\nu = f_\sharp \mu} \int_X c(x, f(x)) \mathrm{d}\mu(x)$$



Theorem: [Brenier] for $c(x, y) = \|x - y\|^2$, (μ, ν) with density, there exists a unique optimal f. One has $f = \nabla\psi$ where ψ is the unique convex function such that $(\nabla\psi)_\sharp \mu = \nu$

Monge-Ampère equation: $\rho(\nabla\psi) \det(\partial^2 \psi) = \xi$

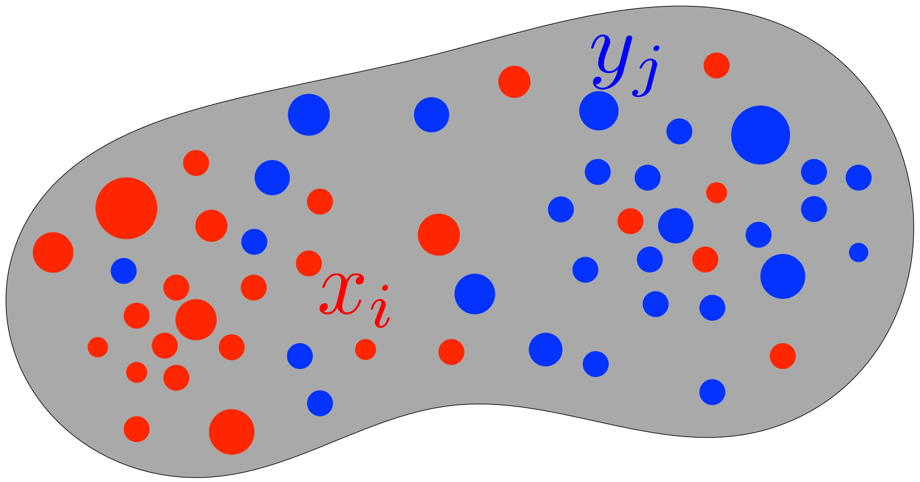
Non-uniqueness / non-existence:



Kantorovitch's Formulation

Input distributions $\mu = \sum_i \mu_i \delta_{x_i}$

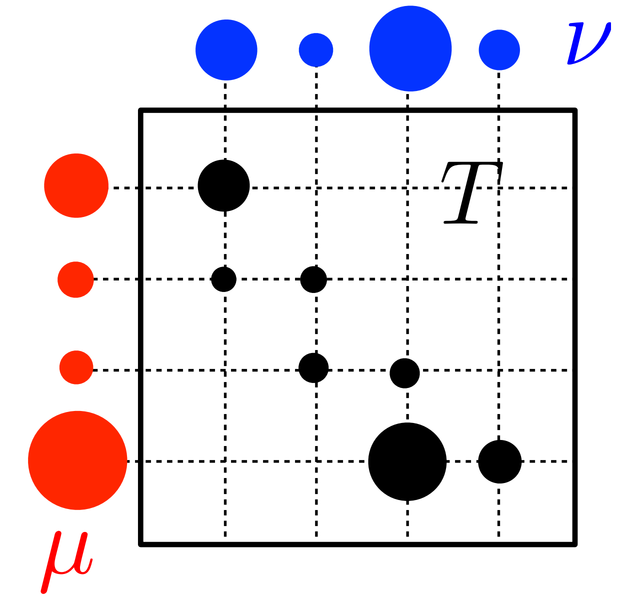
$\nu = \sum_j \nu_j \delta_{y_j}$



Points $(x_i)_i$, $(y_j)_j$

Weights $\mu_i \geqslant 0$, $\nu_j \geqslant 0$.

$\sum_{i=1}^{N_1} \mu_i = \sum_{j=1}^{N_2} \nu_j = 1$ $\qquad d_{i,j} = d(x_i, y_j)$



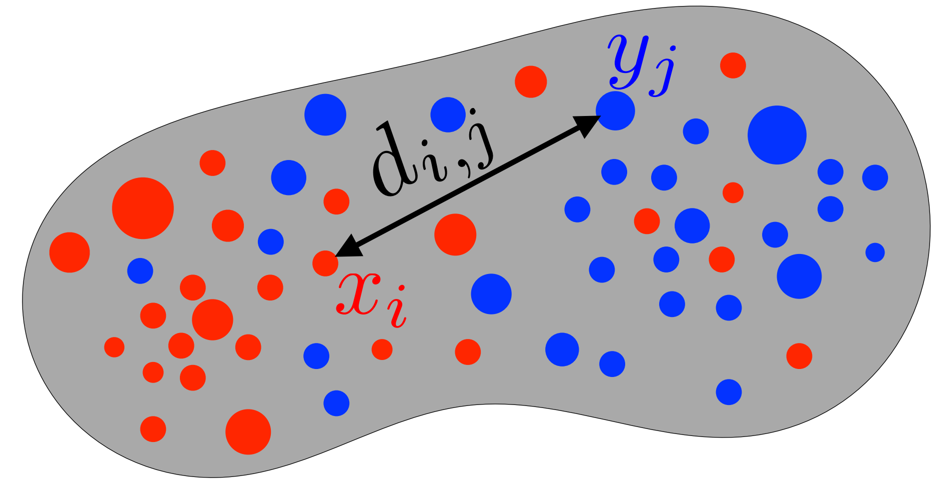
Def. *Couplings*

$$\mathcal{C}_{\mu,\nu} \stackrel{\text{def.}}{=} \left\{ T \in \mathbb{R}_+^{N_1 \times N_2} \; ; \; T\mathbb{1}_{N_1} = \mu, T^\top \mathbb{1}_{N_2} = \nu \right\}$$

Kantorovitch's Formulation

Input distributions $\mu = \sum_i \mu_i \delta_{x_i}$

$\nu = \sum_j \nu_j \delta_{y_j}$

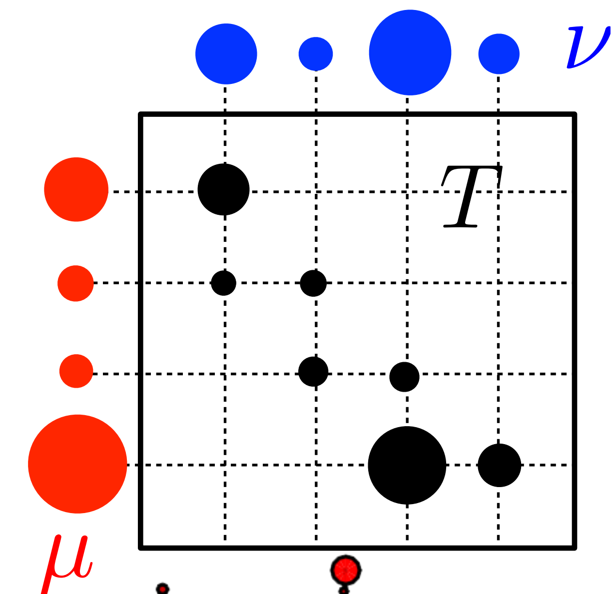


Points $(x_i)_i$, $(y_j)_j$

Weights $\mu_i \geqslant 0$, $\nu_j \geqslant 0$.

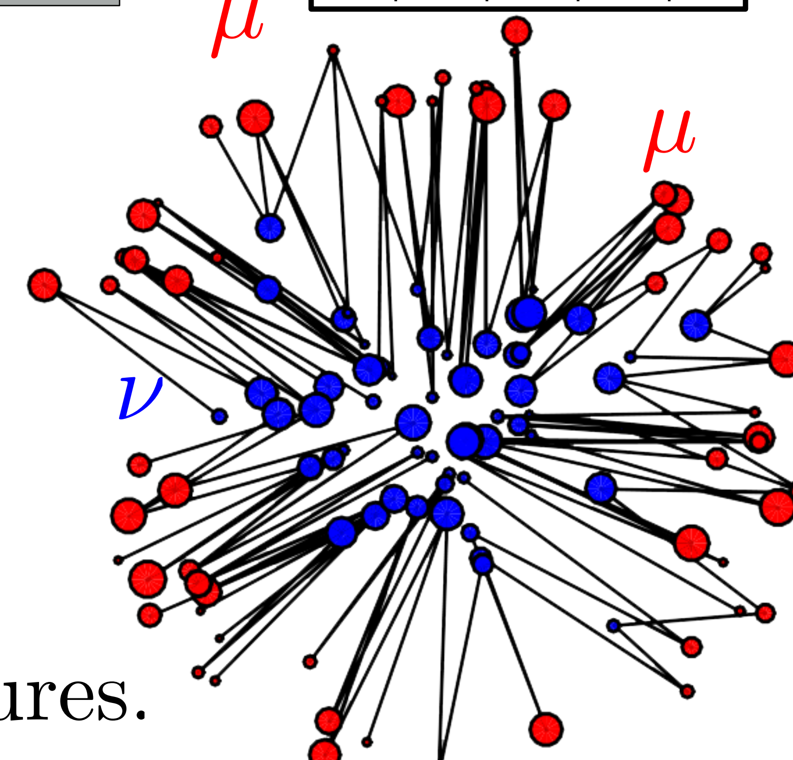
$\sum_{i=1}^{N_1} \mu_i = \sum_{j=1}^{N_2} \nu_j = 1$

$d_{i,j} = d(x_i, y_j)$



Def. *Couplings*

$$\mathcal{C}_{\mu,\nu} \overset{\text{def.}}{=} \left\{ T \in \mathbb{R}_+^{N_1 \times N_2} \; ; \; T \mathbb{1}_{N_1} = \mu, T^\top \mathbb{1}_{N_2} = \nu \right\}$$



Def. *Wasserstein Distance / EMD*

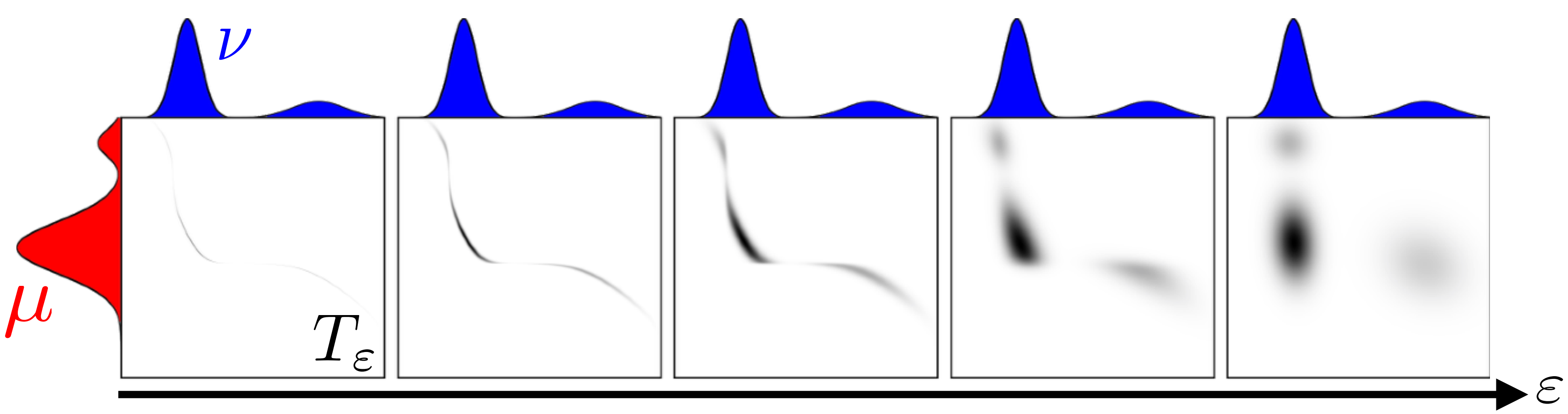
$$W_p^p(\mu, \nu) \overset{\text{def.}}{=} \min \left\{ \sum_{i,j} T_{i,j} d_{i,j}^p \; ; \; T \in \mathcal{C}_{\mu,\nu} \right\}$$

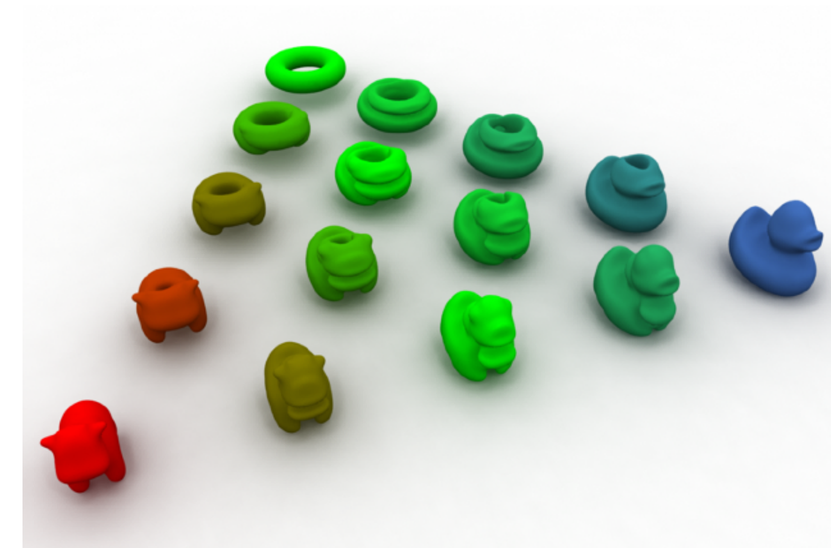
[Kantorovich 1942]

$\to W_p$ is a distance over Radon probability measures.

What's next

Marco Cuturi: fast entropic numerical solvers, applications.





Nicolas Courty: Optimal Transport for machine learning.