Regularizing Optimal Transport Using Regularity Theory

Séminaire Palaisien

November 5, 2019

François-Pierre Paty

francoispierrepaty.github.io

Based on a joint work with ***Alexandre d'Aspremont*** *and* ***Marco Cuturi***

ENSAE

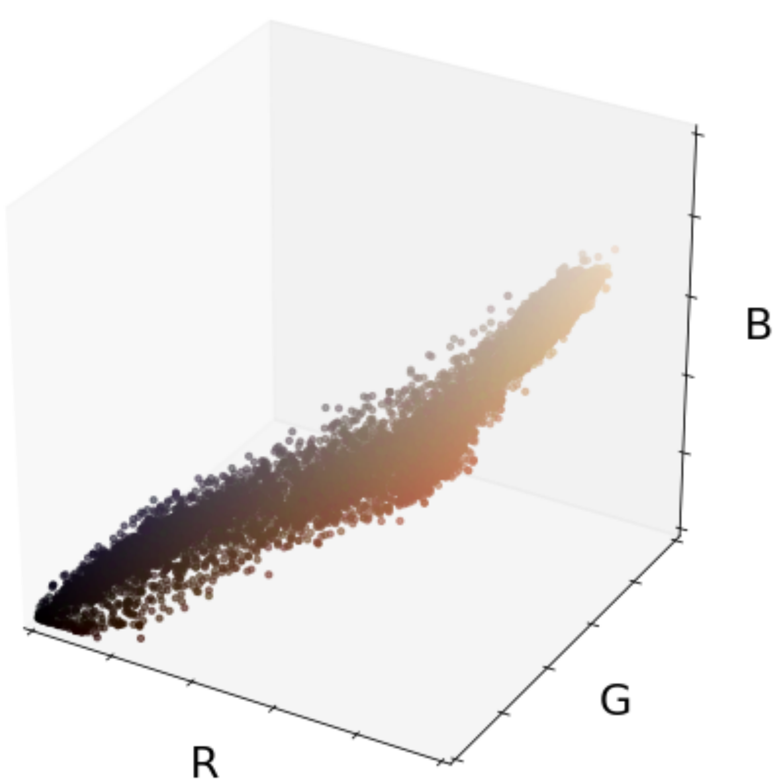
IP PARIS

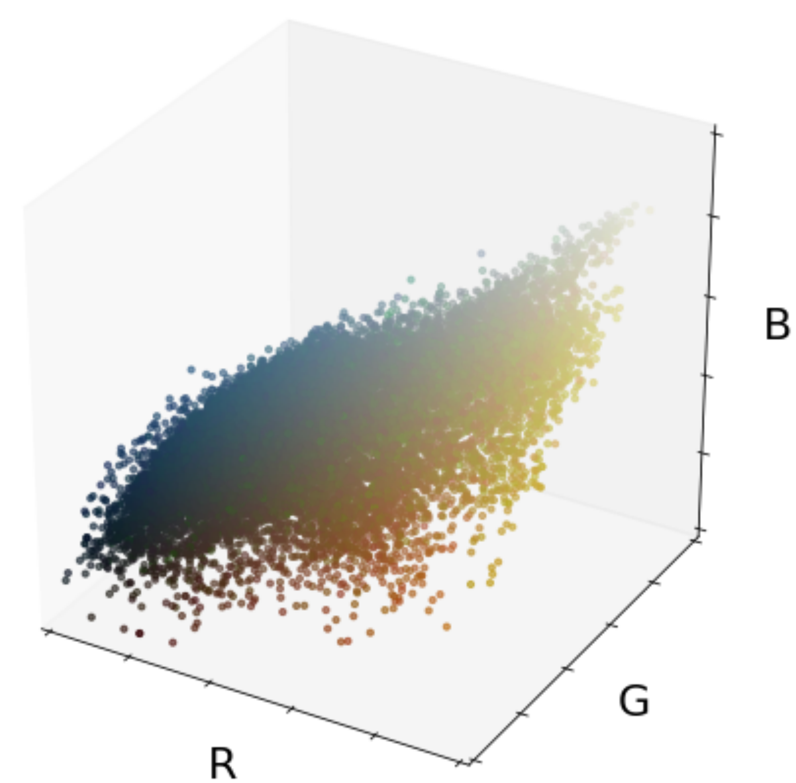
INTRODUCTION

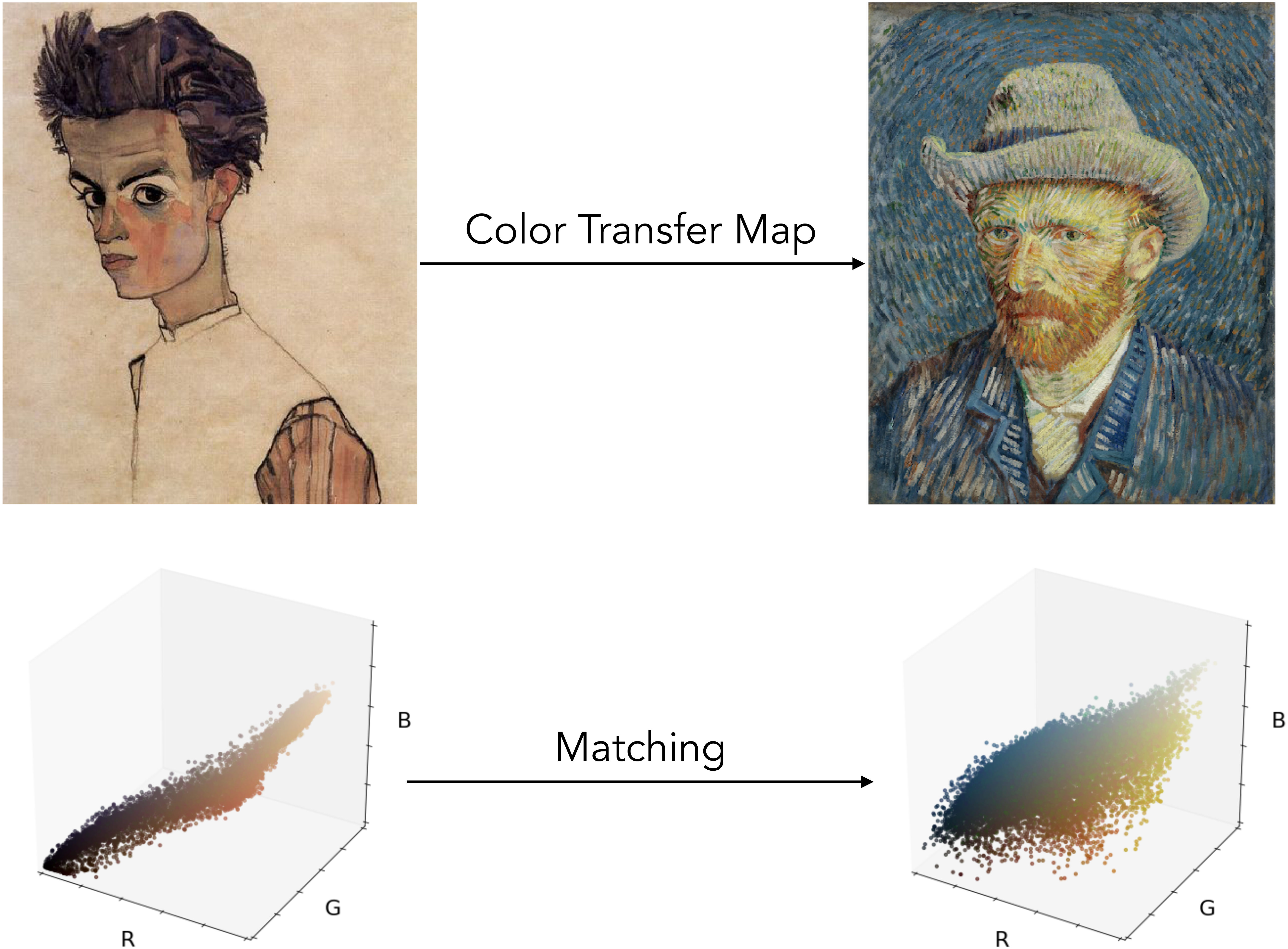
Gaspard Monge

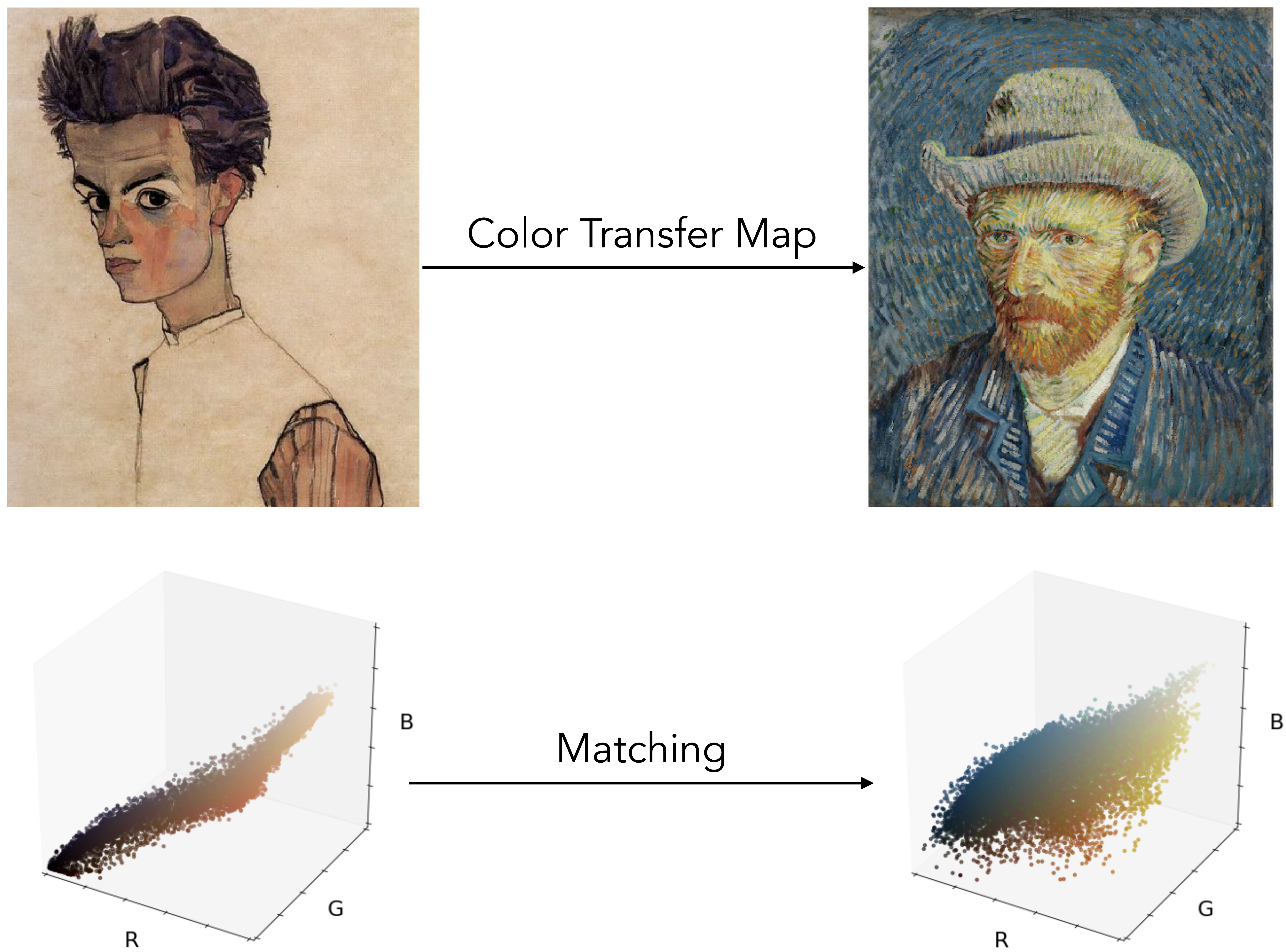
Color Transfer Map

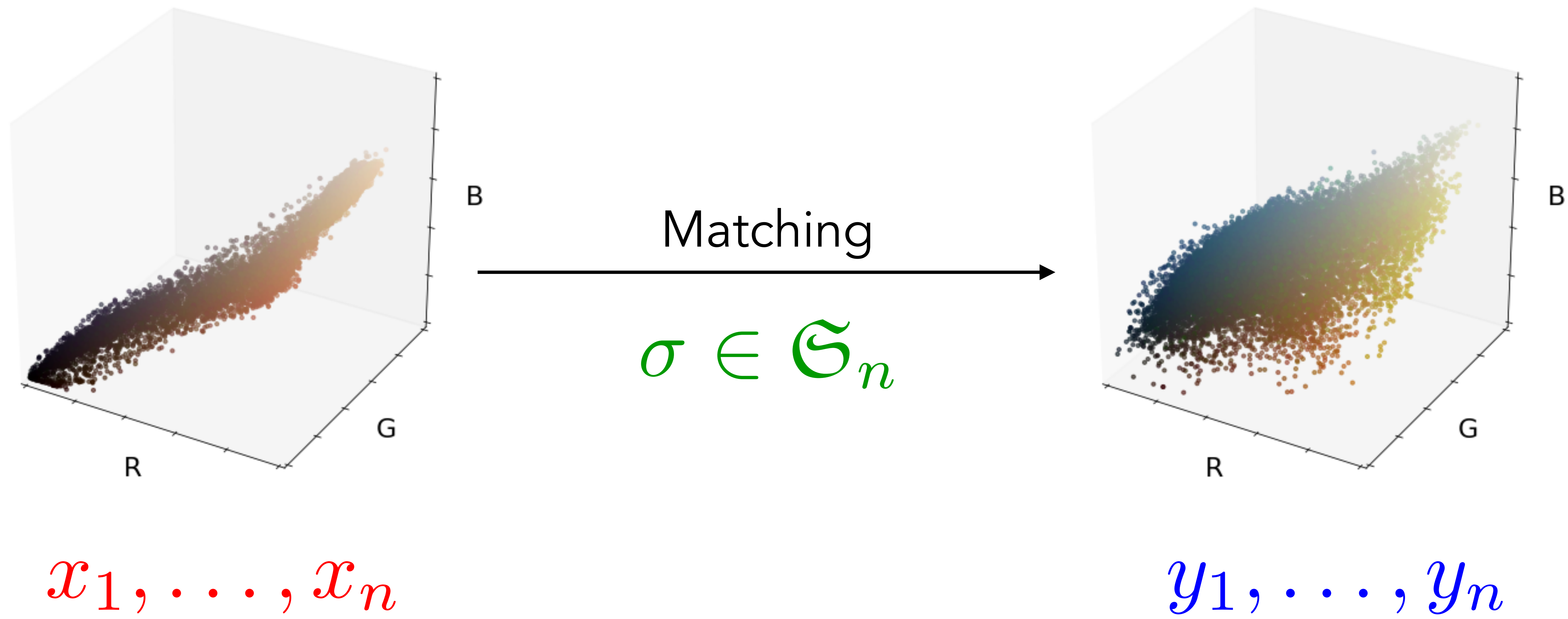
Color Transfer Map

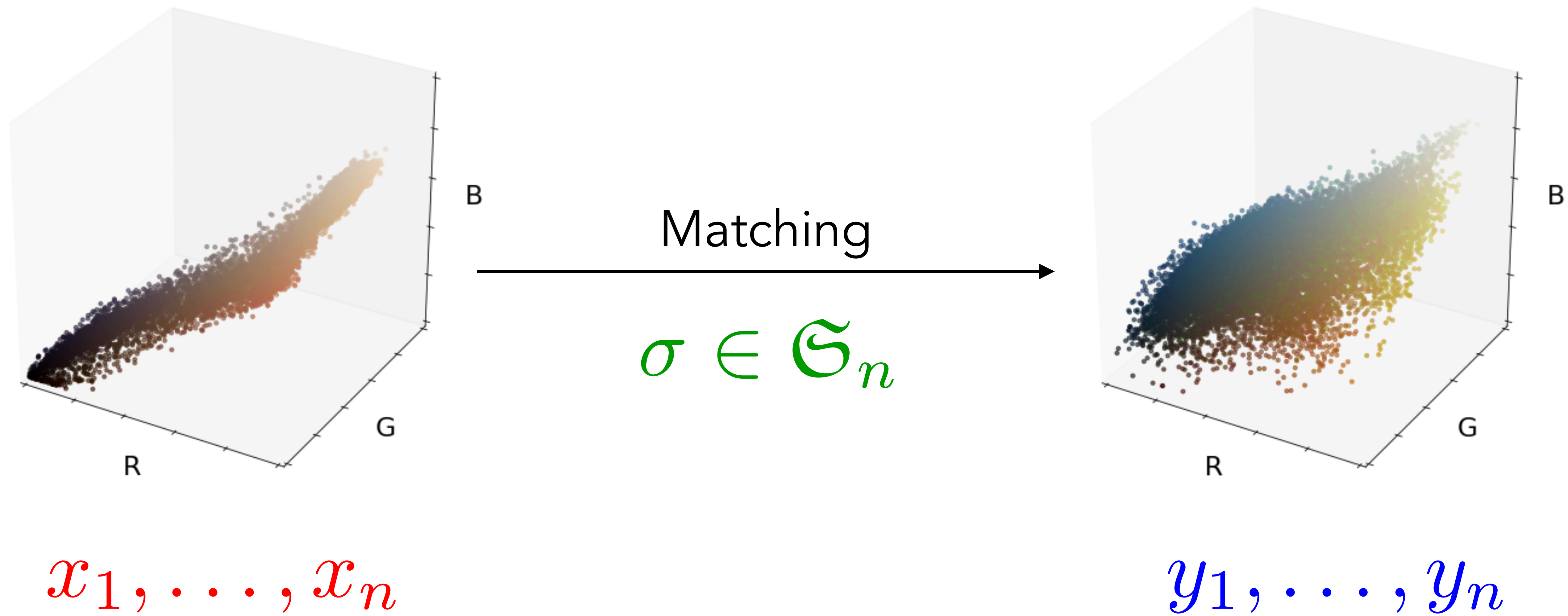


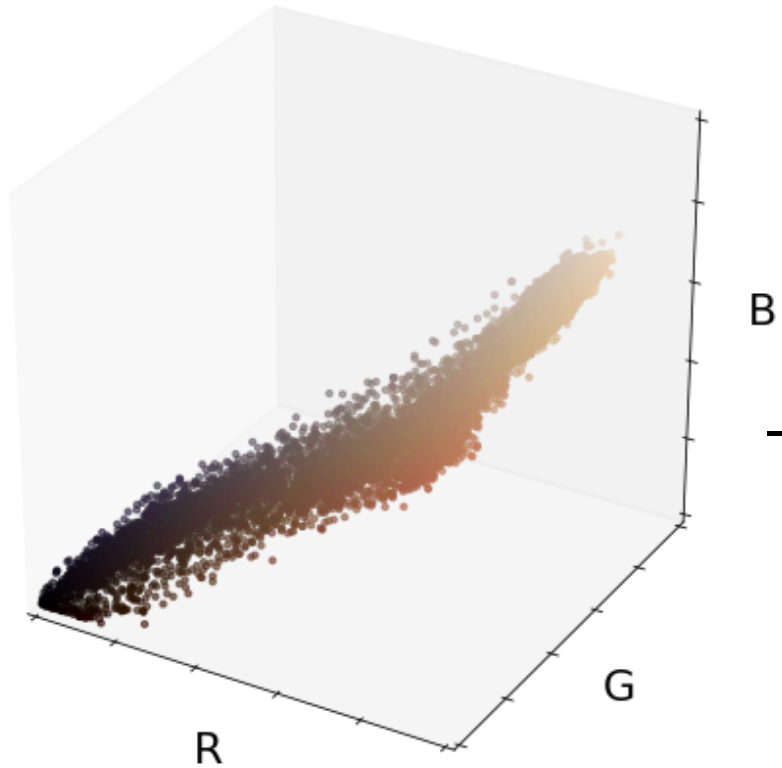



Color Transfer Map
Matching
B
G
R
B
G
R


Color Transfer Map
Matching
B
G
R
B
G
R

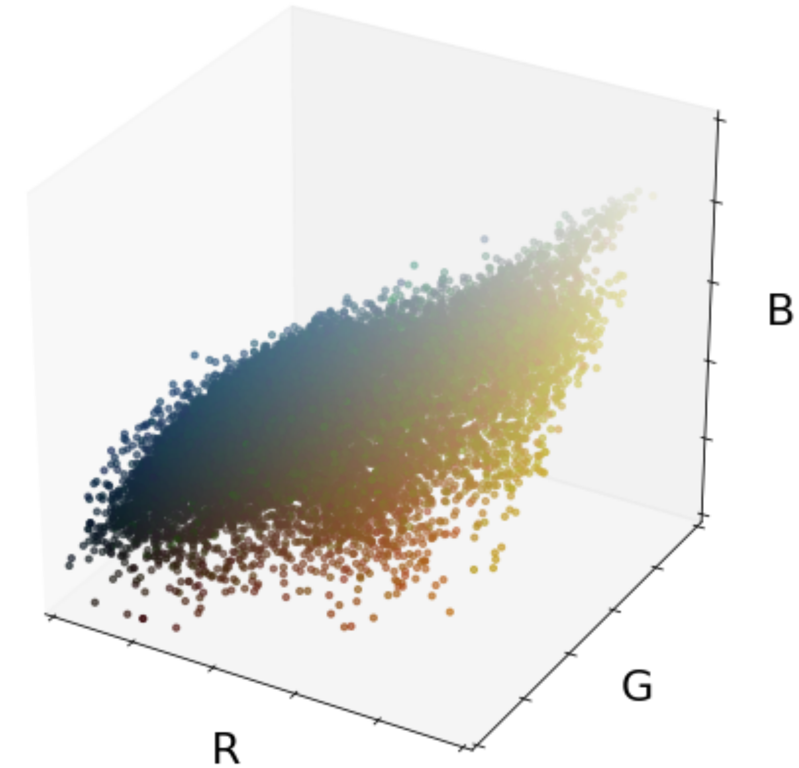

B
G
R
Matching
$\sigma \in \mathfrak{S}_n$
B
G
R
$x_1, \ldots, x_n$
$y_1, \ldots, y_n$


B
R
G
Matching
$\sigma \in \mathfrak{S}_n$
B
R
G
$x_1, \ldots, x_n$
$y_1, \ldots, y_n$



Matching

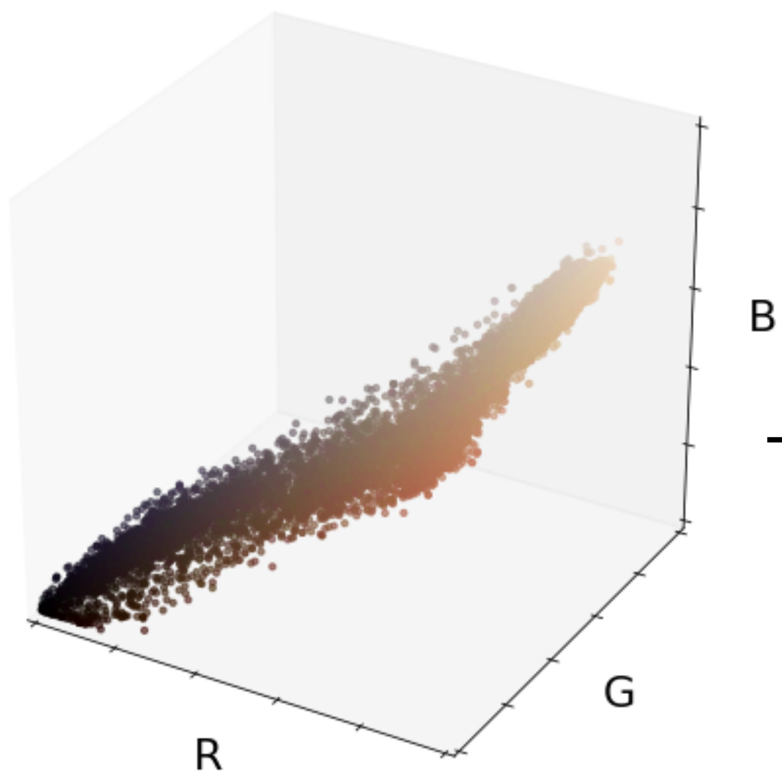
$\sigma \in \mathfrak{S}_n$



$x_1, \dots, x_n$

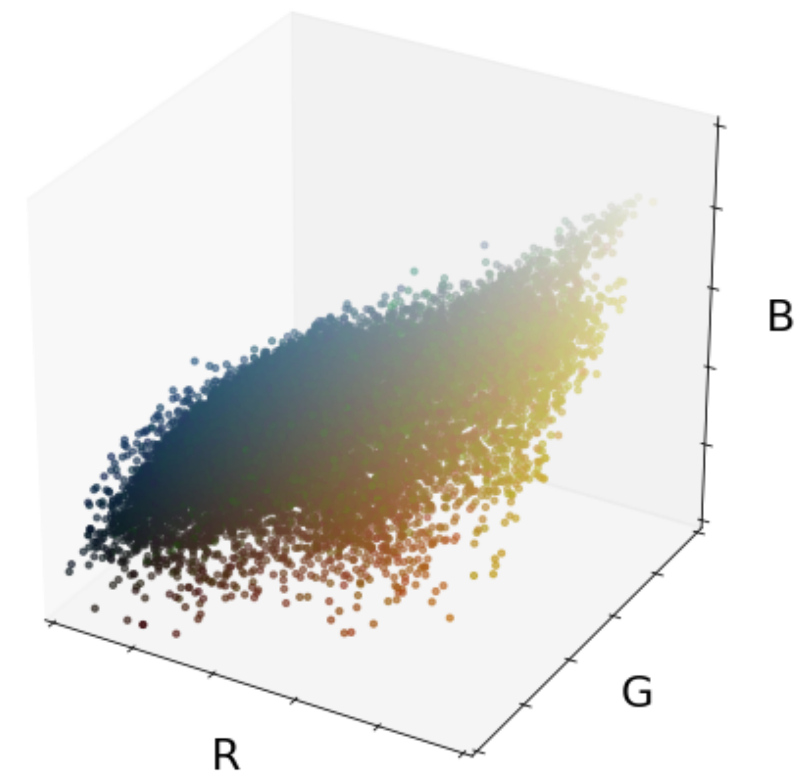
$y_1, \dots, y_n$

$$\|x_i - y_{\sigma(i)}\|^2$$



Matching

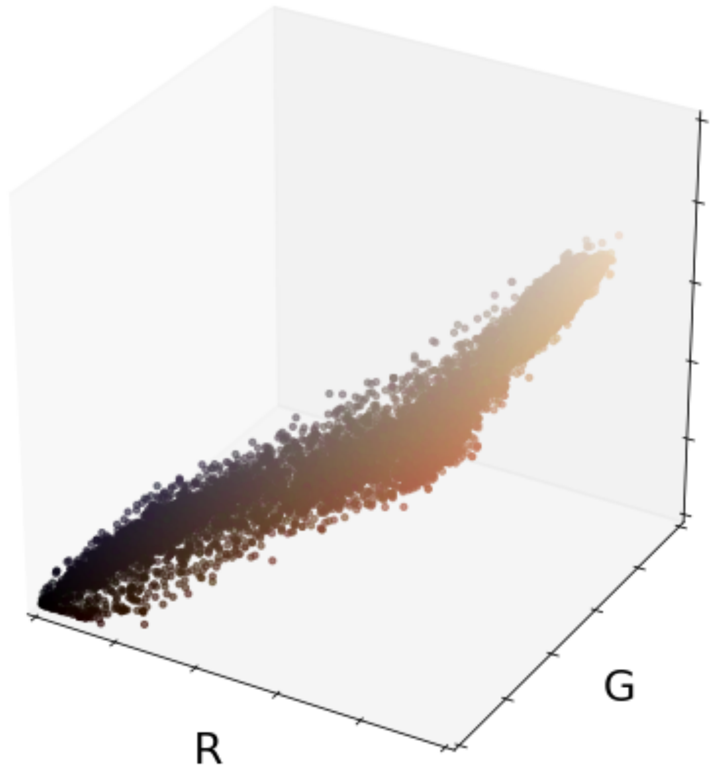
$$\sigma \in \mathfrak{S}_n$$



$$x_1, \ldots, x_n$$

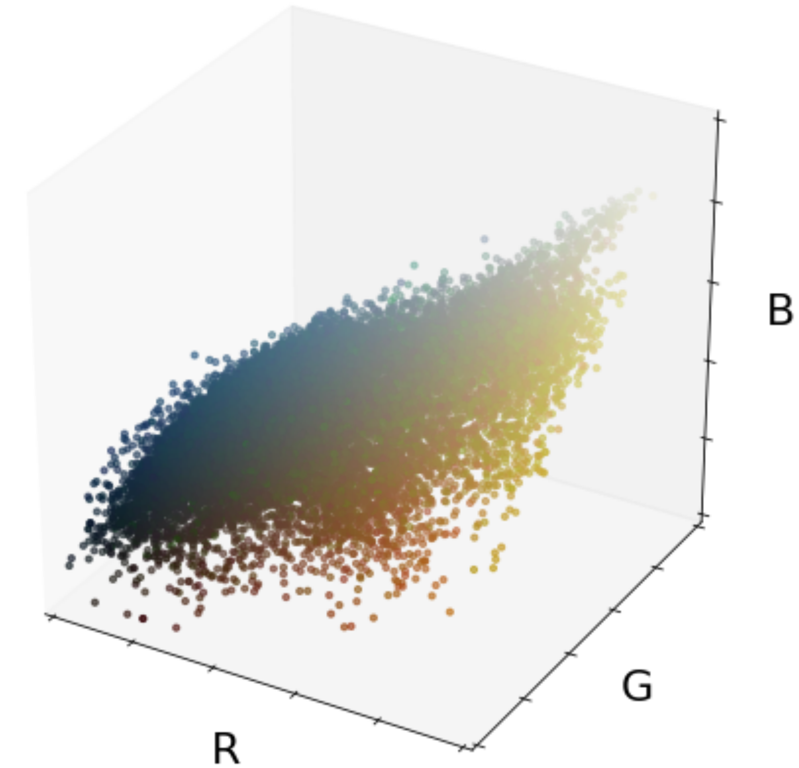
$$y_1, \ldots, y_n$$

$$\sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\|^2$$



Matching

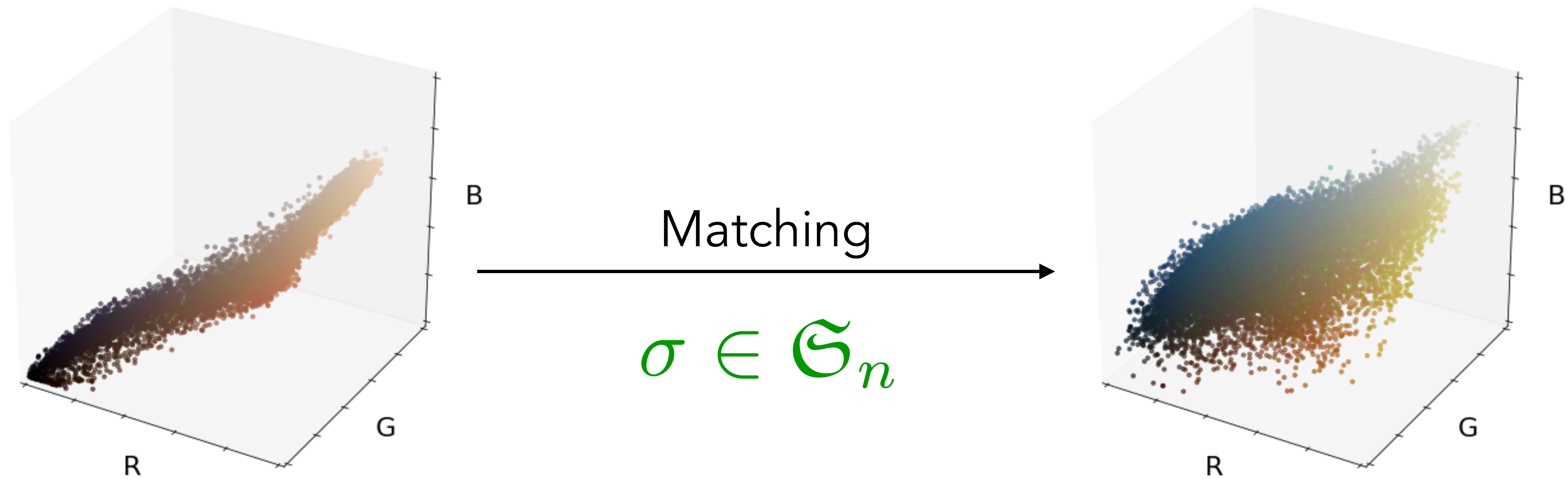
$\sigma \in \mathfrak{S}_n$



$x_1, \ldots, x_n$

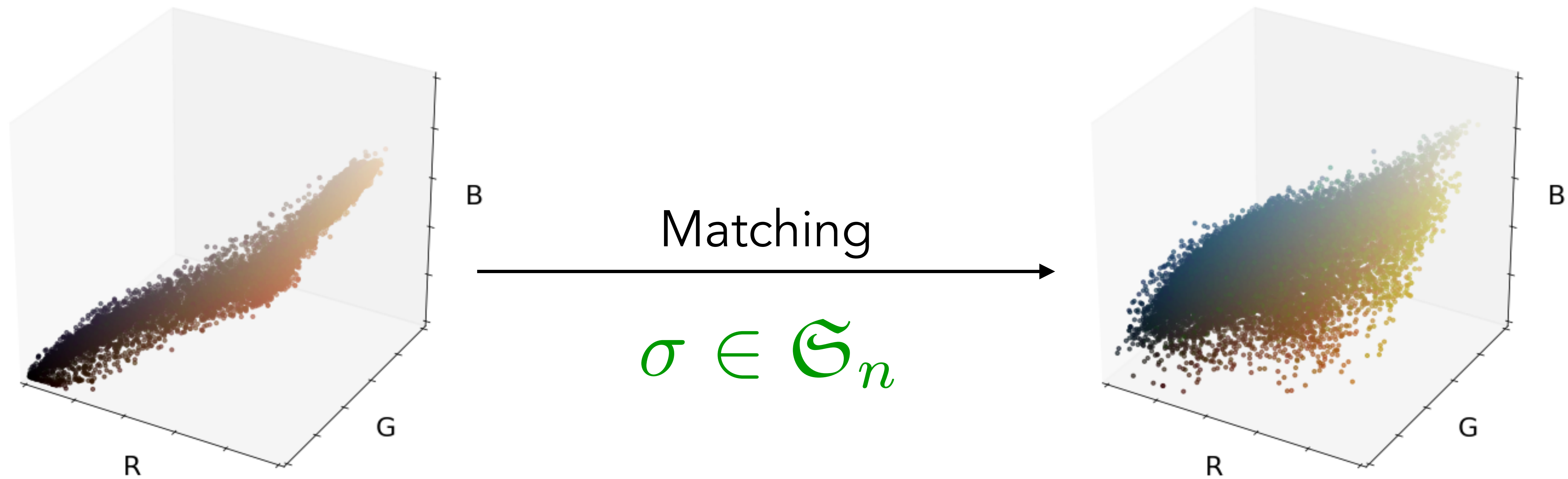
$y_1, \ldots, y_n$

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \| x_i - y_{\sigma(i)} \|^2$$



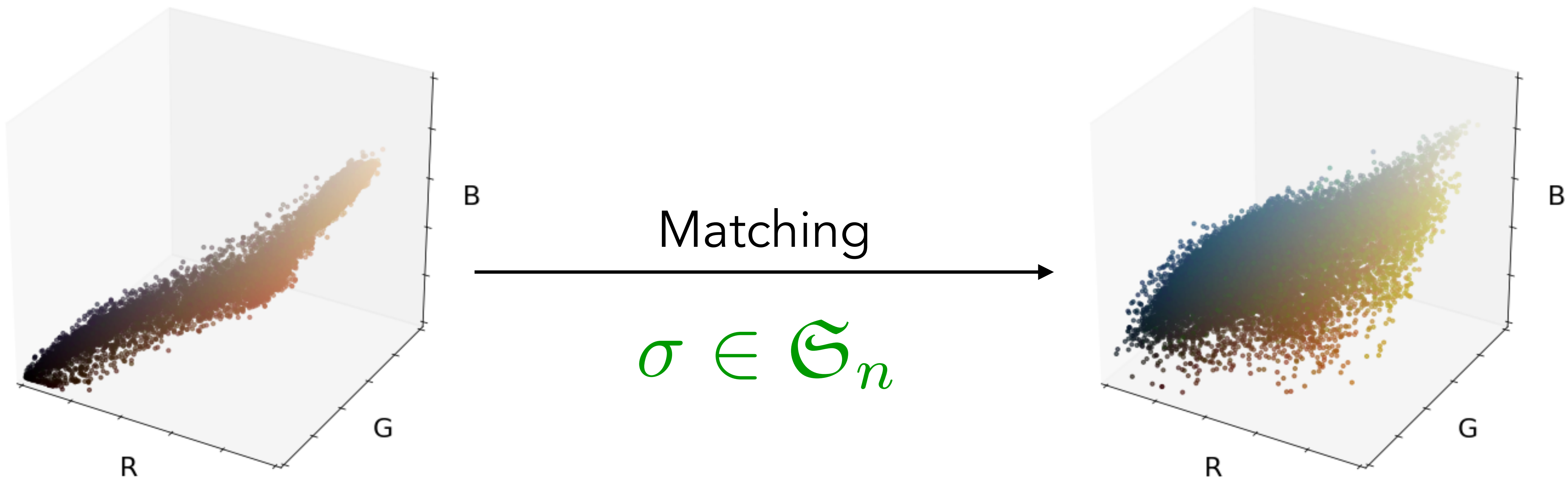
Discrete Monge Problem (1781)

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\|^2$$



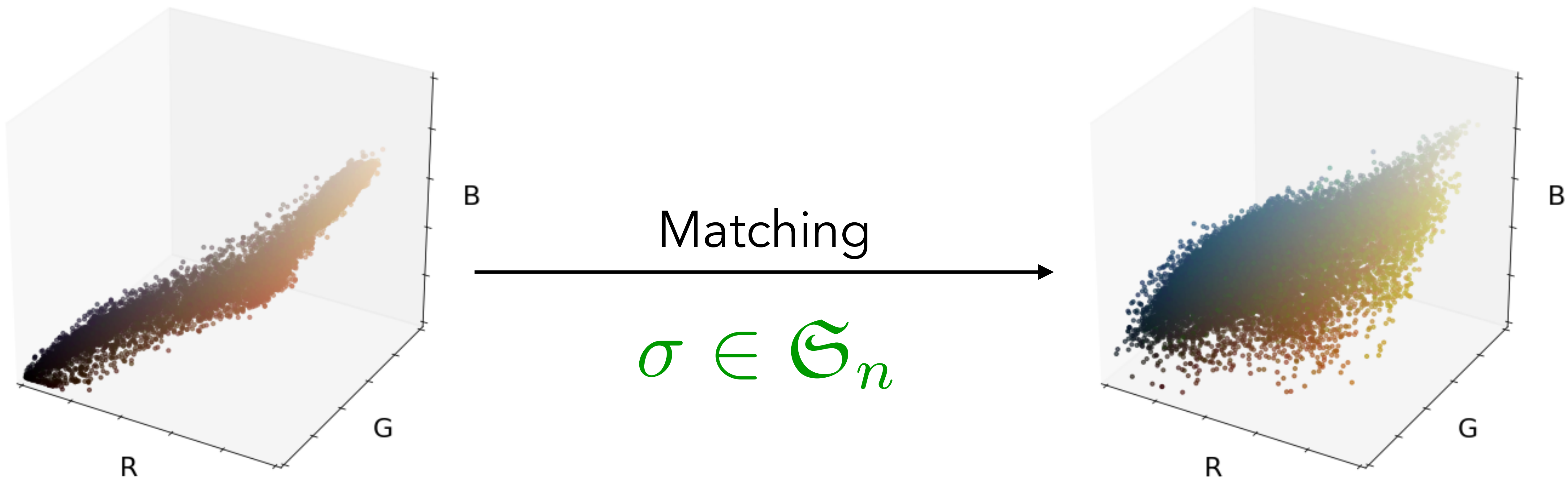
Discrete Monge Problem (1781)

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \| x_i - y_{\sigma(i)} \|$$



Discrete Monge Problem (1781)

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\|^2$$



<u>Discrete Monge Problem (1781)</u>

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \| x_i - y_{\sigma(i)} \|^2$$

Discrete Monge Problem (1781)

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\|^2$$

(i) How to handle repeated points ?

(ii) How to handle different numbers of points ?

(iii) How to compute this combinatorial problem ?

OPTIMAL TRANSPORT

Leonid Kantorovich

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \| x_i - y_{\sigma(i)} \|^2$$

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - y_j\|^2 \, \mathbb{1}_{\sigma(i)=j}$$

$$\min_{P \in \mathfrak{P}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} \| x_i - y_j \|^2 \, P_{ij}$$

$$\mathfrak{P}_n = \left\{ P \in \mathbb{R}^{n \times n} \text{ permutation matrix} \right\}$$

$$\min_{P \in \mathfrak{P}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - y_j\|^2 \, P_{ij}$$

$$\mathfrak{P}_n = \{ P \in \mathbb{R}^{n \times n} \text{ permutation matrix} \}$$

We only have to convexify and generalize $\mathfrak{P}_n$.

$$\min_{P \in \mathfrak{P}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - y_j\|^2 \, P_{ij}$$

$$\mathfrak{P}_n = \{P \in \mathbb{R}^{n \times n} \text{ permutation matrix}\}$$

We only have to convexify and generalize $\mathfrak{P}_n$.

If $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$ are probability weights, we define the associated **transportation polytope**:

$$\min_{P \in \mathfrak{P}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - y_j\|^2 P_{ij}$$

$$\mathfrak{P}_n = \{P \in \mathbb{R}^{n \times n} \text{ permutation matrix}\}$$

We only have to convexify and generalize $\mathfrak{P}_n$.

If $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$ are probability weights, we define the associated **transportation polytope**:

$$\mathfrak{U}(\mathbf{a}, \mathbf{b}) = \{P \in \mathbb{R}_+^{n \times m} \,|\, P\mathbb{1}_m = \mathbf{a}, P^\top \mathbb{1}_n = \mathbf{b}\}$$

Discrete Kantorovitch Problem

$$W_2^2(\mu,\nu) = \min_{P \in \mathfrak{U}(\mathbf{a},\mathbf{b})} \sum_{i=1}^{n} \sum_{j=1}^{m} \|x_i - y_j\|^2 P_{ij}$$

where $\mu = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$ are probability measures

2-Wasserstein distance

If $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$ are probability weights, we define the associated **transportation polytope**:

$$\mathfrak{U}(\mathbf{a},\mathbf{b}) = \{P \in \mathbb{R}_+^{n\times m} \mid P\mathbb{1}_m = \mathbf{a}, P^\top \mathbb{1}_n = \mathbf{b}\}$$

In statistics, we can interpret the data points as iid samples from two densities / probability measures:

$$x_1, \ldots, x_n \sim \mu \qquad\qquad y_1, \ldots, y_n \sim \nu$$

We can define the Monge problem and the Kantorovich problem in the general case of two probability measures.

Monge and Kantorovich problems

Monge and Kantorovich problems

Kantorovich

$$\min_{P \in \mathfrak{U}(\mathbf{a},\mathbf{b})} \sum_{i=1}^{n} \sum_{j=1}^{m} \|x_i - y_j\|^2 \, P_{ij}$$

$$\min_{P \in \mathfrak{U}(\mu,\nu)} \iint \|x - y\|^2 \, dP(x, y)$$

Monge and Kantorovich problems

Monge and Kantorovich problems

Monge

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\|^2$$

$$\inf_{T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$

Monge and Kantorovich problems

Monge

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \|x_i - y_{\sigma(i)}\|^2$$

$$\inf_{T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$

$$X \sim \mu \implies T(X) \sim \nu$$

Given samples

$$x_1, \dots, x_n \sim \mu \qquad\qquad y_1, \dots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad\qquad \hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

can we reconstruct the Wasserstein distance between the generating measures ?

A natural estimator is the Wasserstein between the empirical measures:

$$|W_2(\mu, \nu) - W_2(\hat{\mu}_n, \hat{\nu}_n)| \sim \left(\frac{1}{n}\right)^{1/d}$$

Given samples

$$x_1, \ldots, x_n \sim \mu \qquad\qquad y_1, \ldots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \qquad\qquad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$$

can we reconstruct the Wasserstein distance between the generating measures ?

A natural estimator is the Wasserstein between the empirical measures:

$$|W_2(\mu, \nu) - W_2(\hat{\mu}_n, \hat{\nu}_n)| \sim \left(\frac{1}{n}\right)^{1/d}$$

Given samples

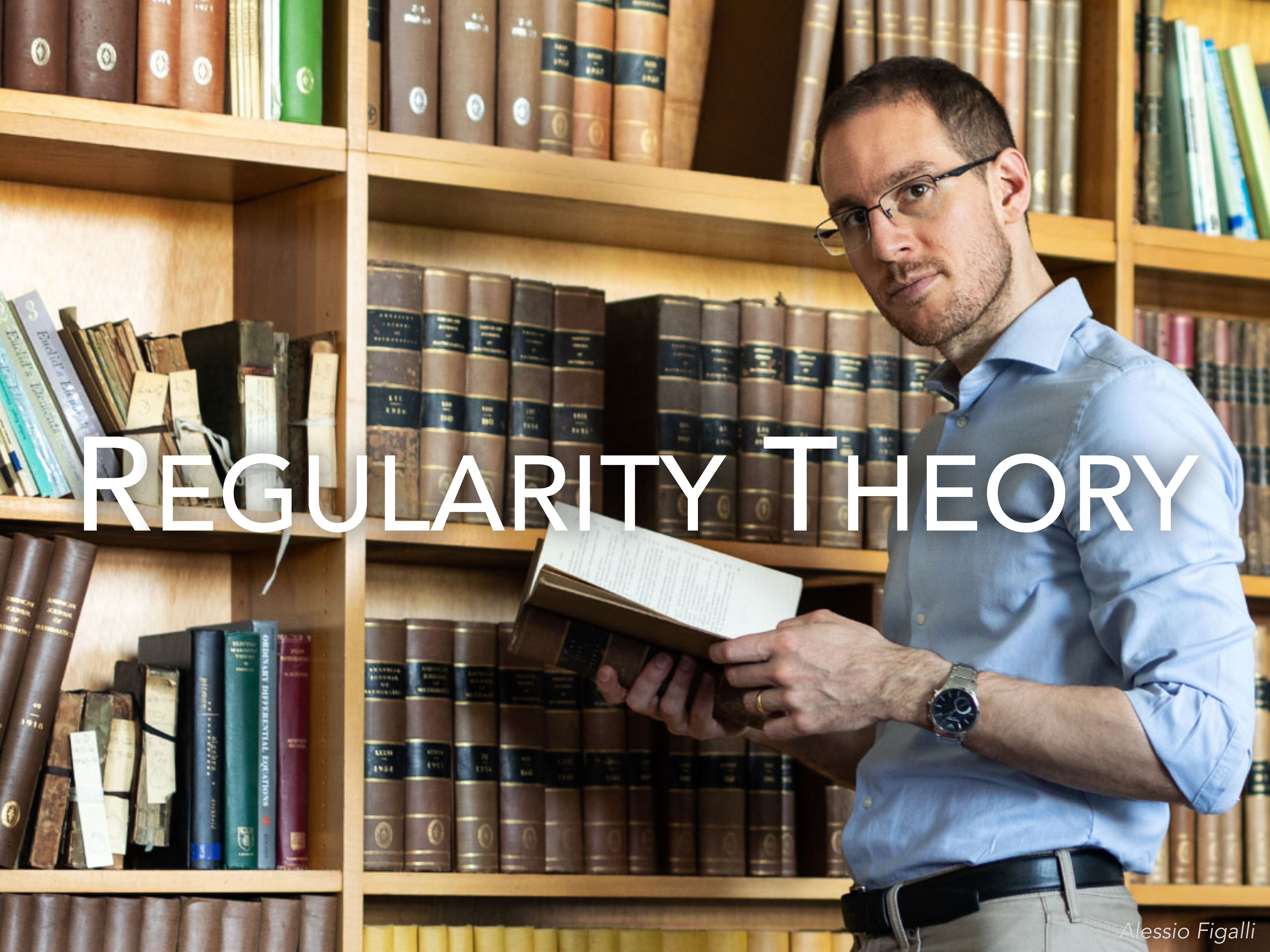
$$x_1, \ldots, x_n \sim \mu \qquad\qquad y_1, \ldots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad\qquad \hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

can we reconstruct the Wasserstein distance between the generating measures ?

Curse of Dimensionality

$$|W_2(\mu, \nu) - W_2(\hat{\mu}_n, \hat{\nu}_n)| \sim \left(\frac{1}{n}\right)^{1/d}$$


REGULARITY THEORY
Alessio Figalli

Let μ and ν be two probability measures over $\mathbb{R}^d$.

$$\inf_{T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$

When does the Monge problem admit a solution ?

What can be said about it ?

Let μ and ν be two probability measures over $\mathbb{R}^d$.

$$\inf_{T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$

Brenier Theorem

1. If μ is *absolutely continuous* with respect to the Lebesgue measure, the Monge problem admits a unique solution
2. If the Monge problem admits a solution T, then there exists a convex function f, called a **Brenier potential**, s.t.

$$T = \nabla f$$

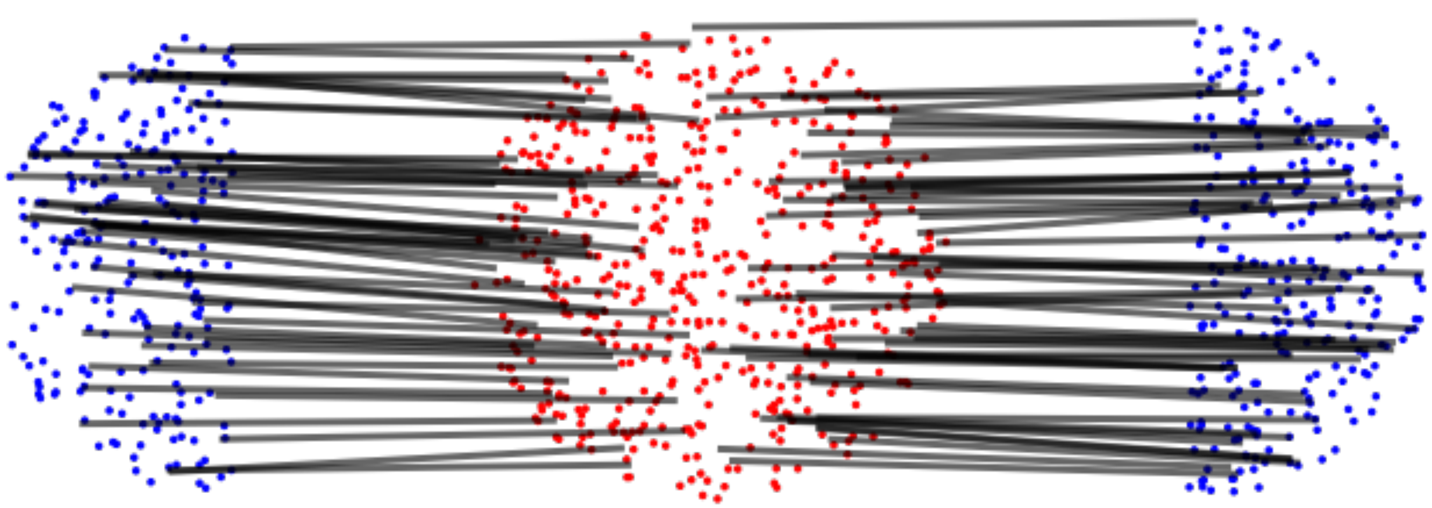
When the optimal map exists (*e.g.* when μ has a density), what kind of regularity does it exhibit ?

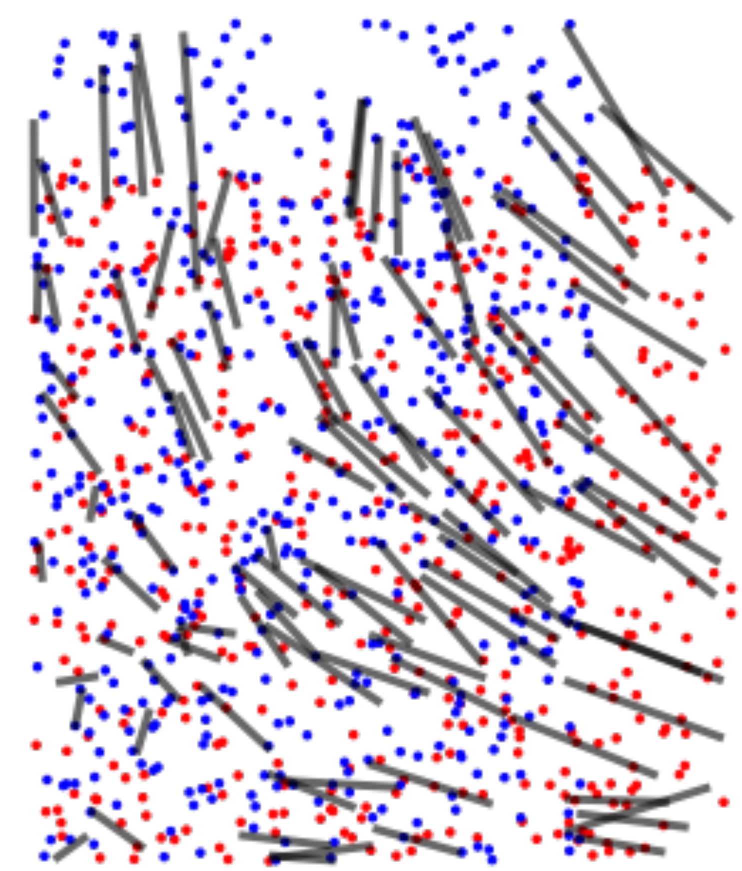
When the optimal map exists (*e.g.* when μ has a density), what kind of regularity does it exhibit ?

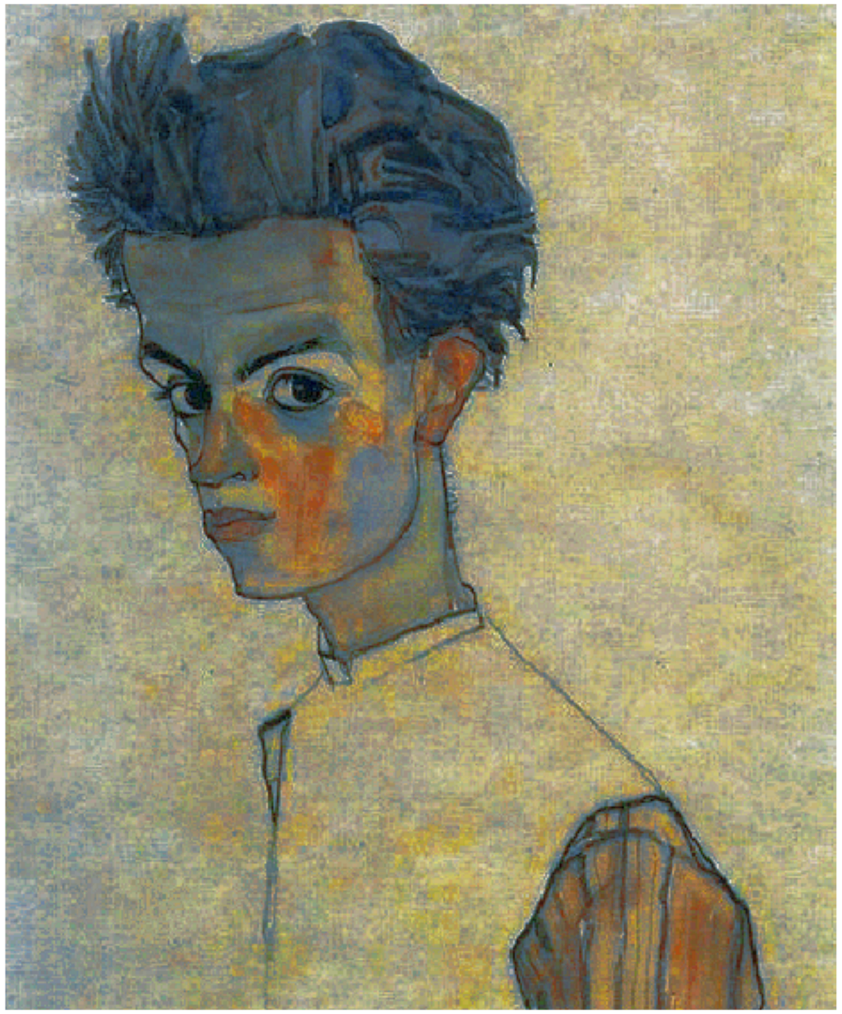
Without further assumptions on μ and ν , we cannot even hope for continuity. Many results by *Caffarelli, De Philippis, Kim, Figalli…*

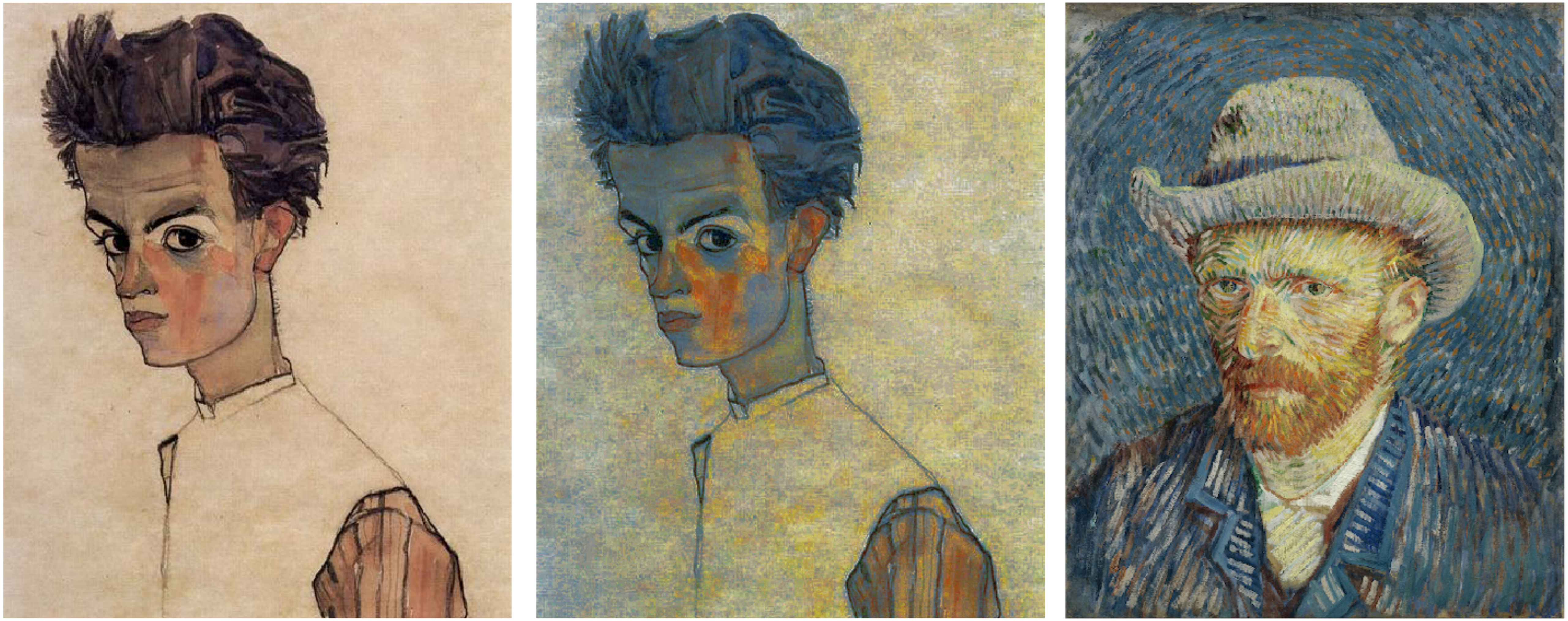
When the optimal map exists (*e.g.* when μ has a density), what kind of regularity does it exhibit ?

Without further assumptions on μ and ν , we cannot even hope for continuity. Many results by *Caffarelli, De Philippis, Kim, Figalli*…





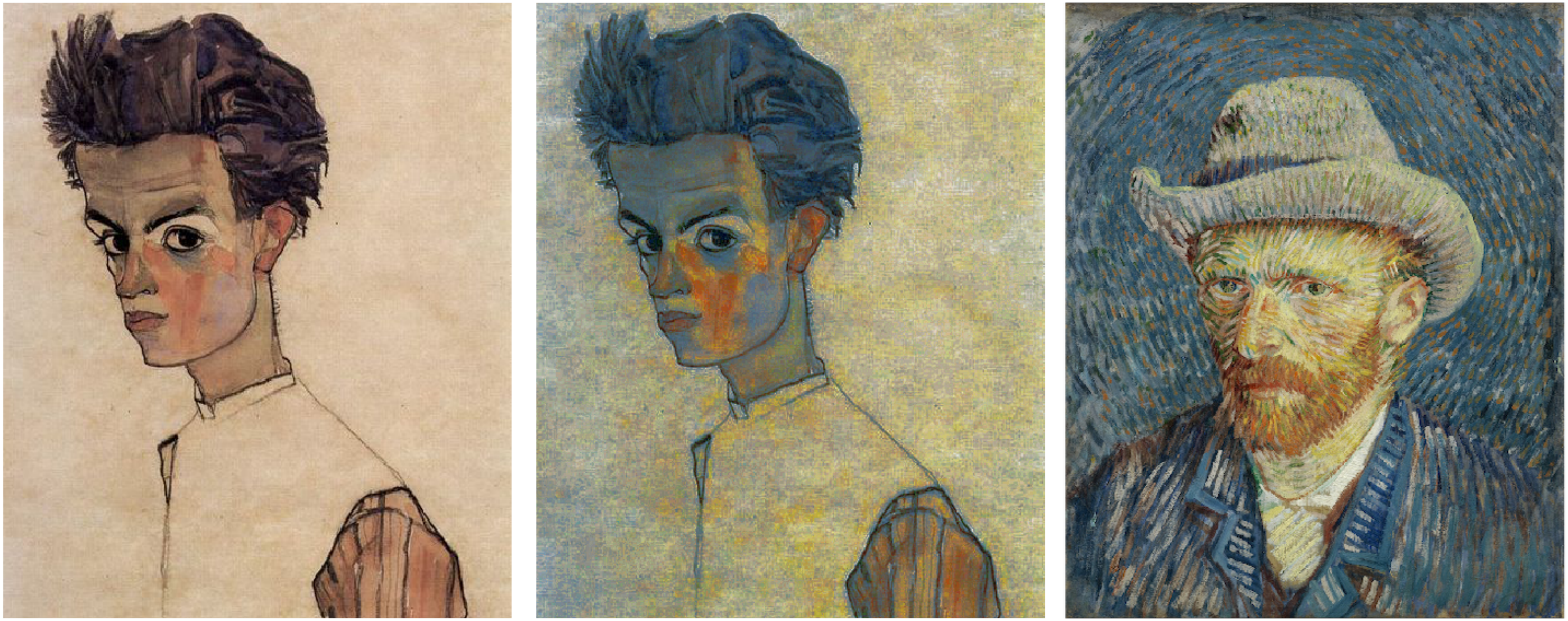


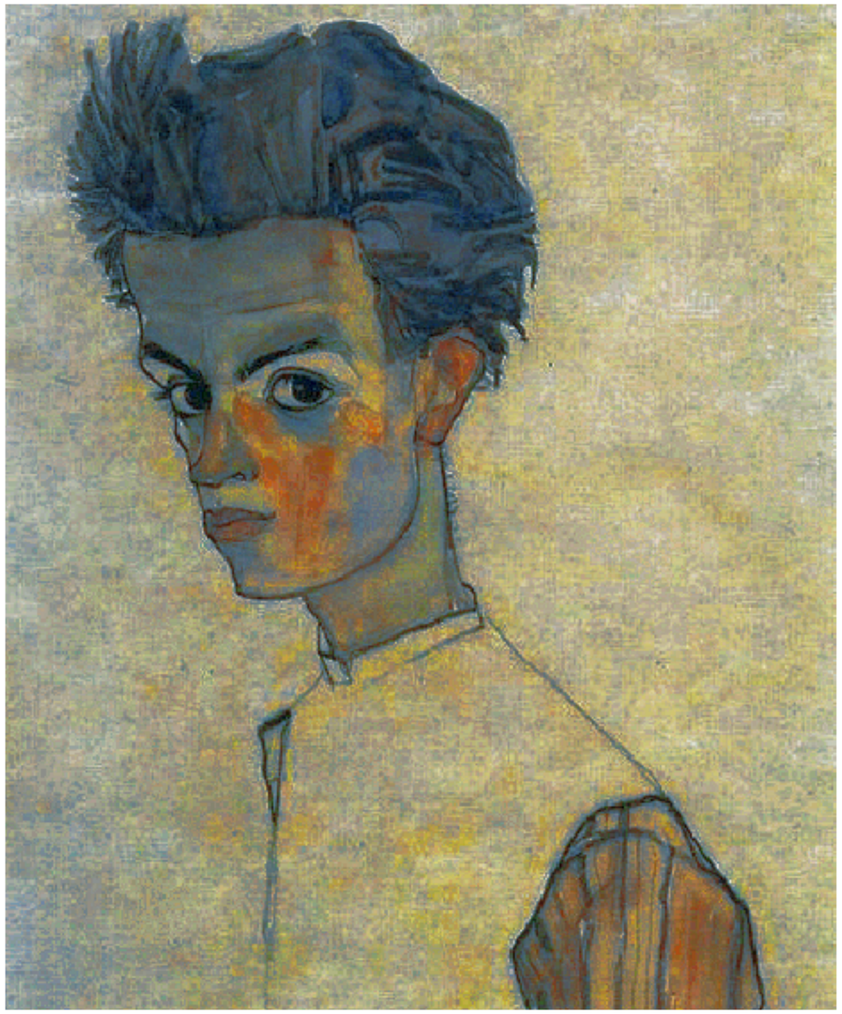


Instead of finding assumptions under which the optimal map exists and exhibits some regularity, we will enforce such regularity directly in the OT problem.

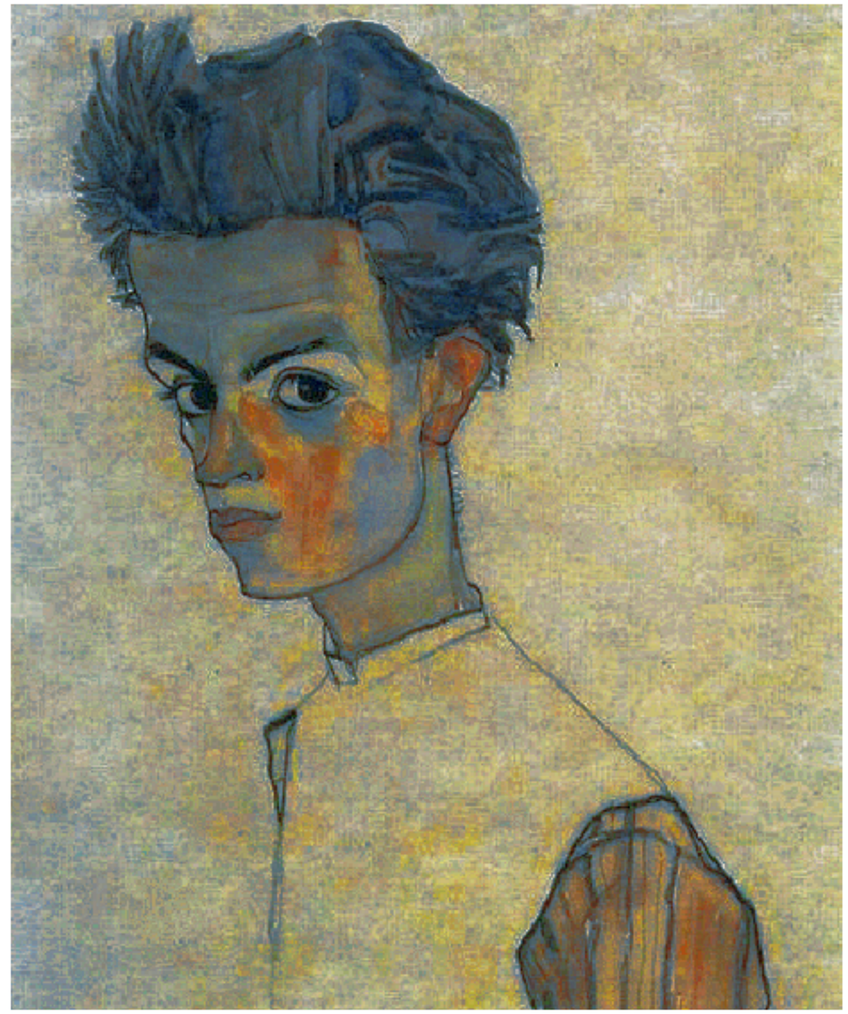
Smooth and Strongly Convex Brenier Potentials

Yann Brenier



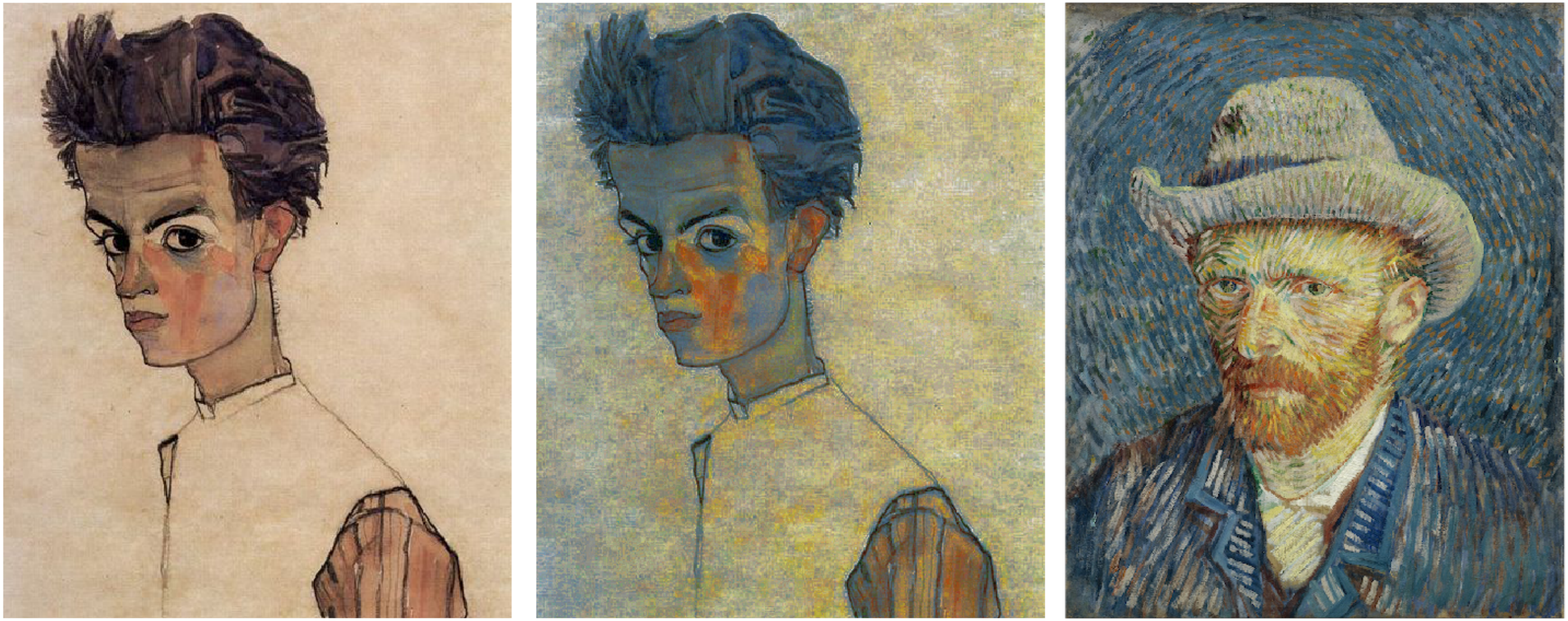


$$\ell\|x-y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$$

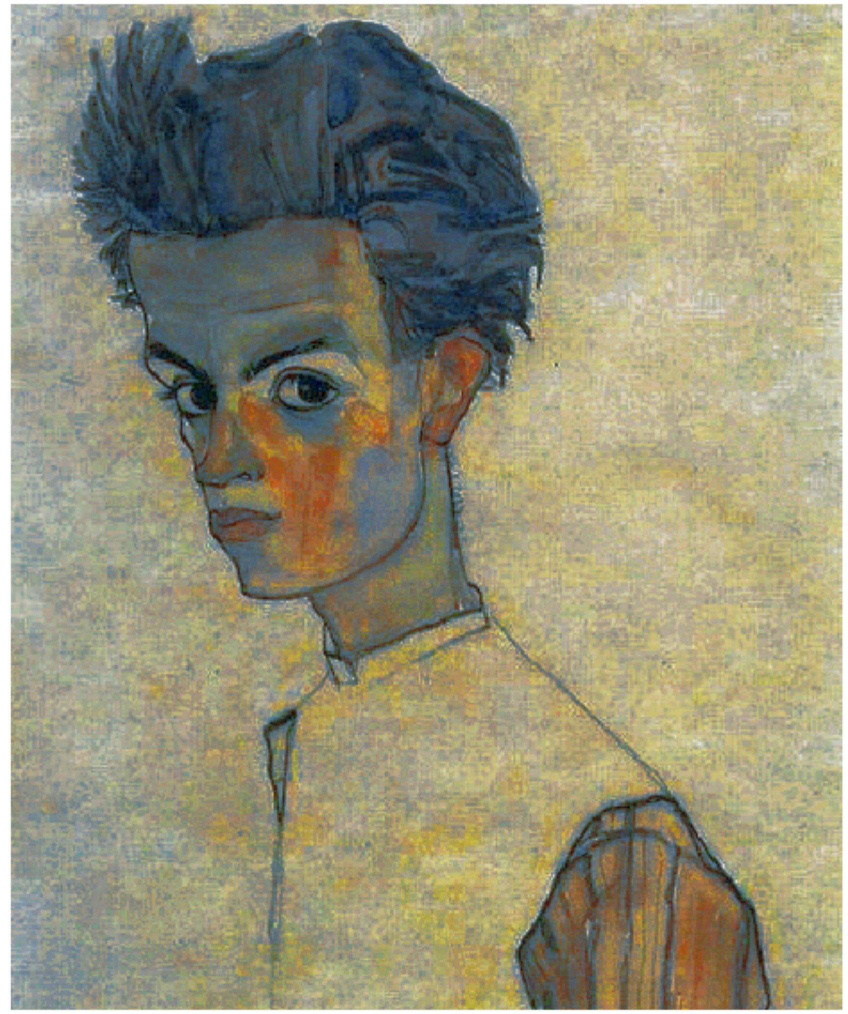
$$\ell\|x-y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$$

We ask that $T = \nabla f$ is a bi-Lipschitz map



$$\ell\|x - y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

We ask that f is **smooth** and **strongly convex**

$$\ell\|x-y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$$

We ask that f is **smooth** and **strongly convex**

$\rightarrow f \in \mathcal{F}_{\ell,L}$

But there may not even such a regular f that is admissible for the Monge problem, *i.e.* such that $(\nabla f)_\sharp \mu = \nu$.

But there may not even such a regular f that is admissible for the Monge problem, *i.e.* such that $(\nabla f)_\sharp \mu = \nu$.

Instead, we will try to best approximate ν as a push-forward of μ through a regular map:

$$\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \mu, \nu \right]$$

But there may not even such a regular f that is admissible for the Monge problem, *i.e.* such that $(\nabla f)_\sharp \mu = \nu$.

Instead, we will try to best approximate ν as a push-forward of μ through a regular map:

$$\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \mu, \nu \right]$$

<u>Smooth and Strong Convex Brenier Potentials</u>

Even when the measures are discrete, this is a *infinite dimensional* optimization problem !

$$\min_{f \in \mathcal{F}_{\ell,L}} W_2 \left[\nabla f_\sharp \mu, \nu \right]$$

Even when the measures are discrete, this is a *infinite dimensional* optimization problem !

$$\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_{\sharp} \mu, \nu \right]$$

Even when the measures are discrete, this is a *infinite dimensional* optimization problem !

$$\min_{f \in \mathcal{F}_{\ell,L}} W_2 \left[\nabla f_\sharp \mu, \nu \right]$$

$$\min_{\substack{z_1,\dots z_n \in \mathbb{R}^d \\ u \in \mathbb{R}^n}} W_2^2 \left(\sum_{i=1}^{n} a_i \delta_{z_i}, \nu \right)$$

Even when the measures are discrete, this is a *infinite dimensional* optimization problem !

$$\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \mu, \nu \right]$$

$$\min_{\substack{z_1, \ldots z_n \in \mathbb{R}^d \\ u \in \mathbb{R}^n}} W_2^2 \left(\sum_{i=1}^{n} a_i \delta_{z_i}, \nu \right)$$

$$u_i \geq u_j + \langle z_j, x_i - x_j \rangle + \frac{1}{2(1 - \ell/L)} \left(\frac{1}{L} \|z_i - z_j\|^2 + \ell \|x_i - x_j\|^2 - 2\frac{\ell}{L} \langle z_j - z_i, x_j - x_i \rangle \right)$$

$$x_1, \ldots, x_n \sim \mu \qquad y_1, \ldots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad \hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

$$x_1, \dots, x_n \sim \mu \qquad\qquad y_1, \dots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad\qquad \hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \hat{\mu}_n, \hat{\nu}_n \right]$$

$$x_1, \dots, x_n \sim \mu \qquad\qquad y_1, \dots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \qquad\qquad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}_{\ell,L}} W_2\left[\nabla f_\sharp \hat{\mu}_n, \hat{\nu}_n\right]$$

$$z_1^\star, \dots, z_n^\star, u^\star$$

$$x_1, \ldots, x_n \sim \mu \qquad y_1, \ldots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad \hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \hat{\mu}_n, \hat{\nu}_n \right]$$

$z_1^\star, \ldots, z_n^\star, u^\star$

We can easily compute the map on any new point x by solving a cheap QCQP

$$x_1, \ldots, x_n \sim \mu \qquad y_1, \ldots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \qquad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \hat{\mu}_n, \hat{\nu}_n\right]$$

$z_1^\star, \ldots, z_n^\star, u^\star$

We can easily compute the map on any new point x by solving a cheap QCQP

$$\min_{v \in \mathbb{R},\, g \in \mathbb{R}^d} v$$

$$\text{s.t.}\ \forall i,\ v \geq u_i + \langle z_i^\star, x - x_i \rangle + \frac{1}{2(1-\ell/L)}\left(\frac{1}{L}\|g - z_i^\star\|^2 + \ell\|x - x_i\|^2 - 2\frac{\ell}{L}\langle z_i^\star - g, x_i - x\rangle\right)$$

$$x_1, \dots, x_n \sim \mu \qquad\qquad y_1, \dots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \qquad\qquad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \hat{\mu}_n, \hat{\nu}_n\right]$$

$z_1^\star, \dots, z_n^\star, u^\star$

We can easily compute the map on any new point x by solving a cheap QCQP

This defines an estimator $\nabla f^\star$ of the optimal transport map sending μ to ν

$$x_1, \ldots, x_n \sim \mu \qquad y_1, \ldots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad \hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \hat{\mu}_n, \hat{\nu}_n \right]$$

$z_1^\star, \ldots, z_n^\star, u^\star$

We can easily compute the map on any new point x by solving a cheap QCQP

This defines an estimator $\nabla f^\star$ of the optimal transport map sending μ to ν

We define the *SSNB estimator* as a plug-in:

$$x_1, \ldots, x_n \sim \mu \qquad y_1, \ldots, y_n \sim \nu$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \qquad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}_{\ell, L}} W_2 \left[\nabla f_\sharp \hat{\mu}_n, \hat{\nu}_n\right]$$

$z_1^\star, \ldots, z_n^\star, u^\star$

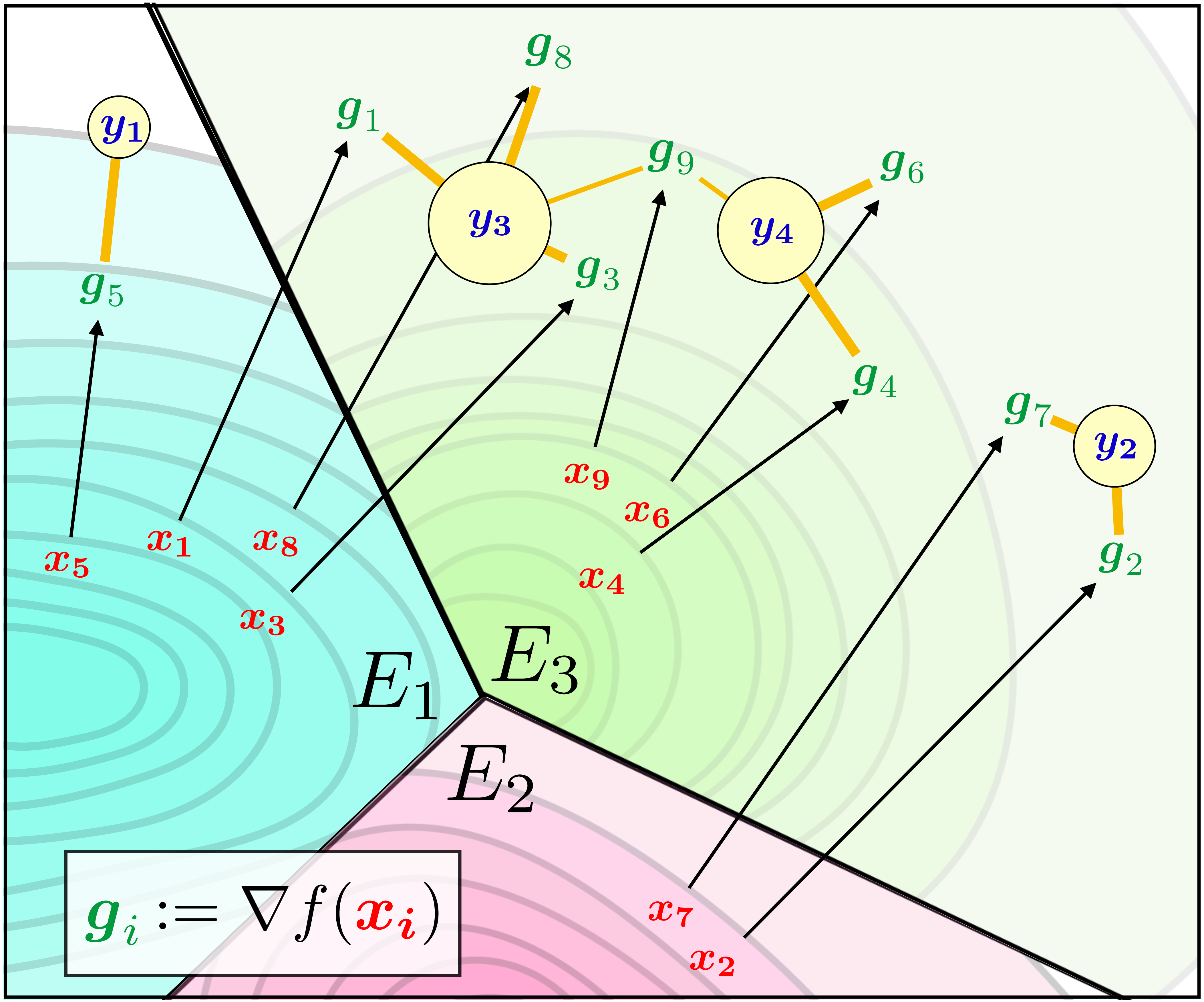
We can easily compute the map on any new point x by solving a cheap QCQP

This defines an estimator $\nabla f^\star$ of the optimal transport map sending μ to ν

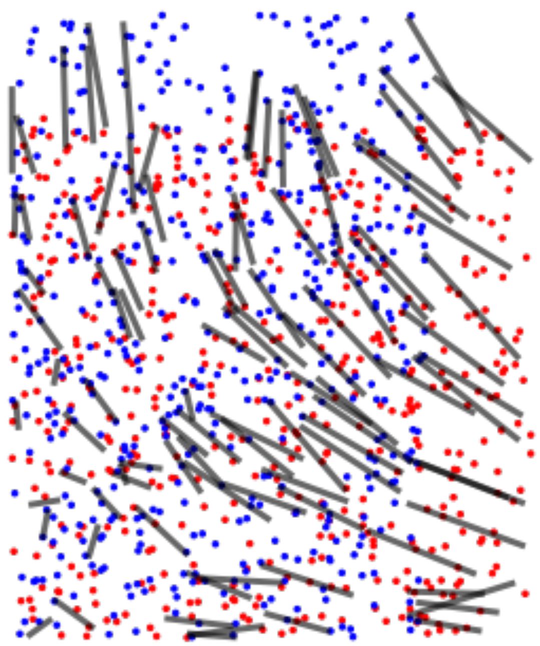
We define the *SSNB estimator* as a plug-in:

$$\widehat{W}_2^2 = \int \|x - \nabla f^\star(x)\|^2 d\mu(x)$$

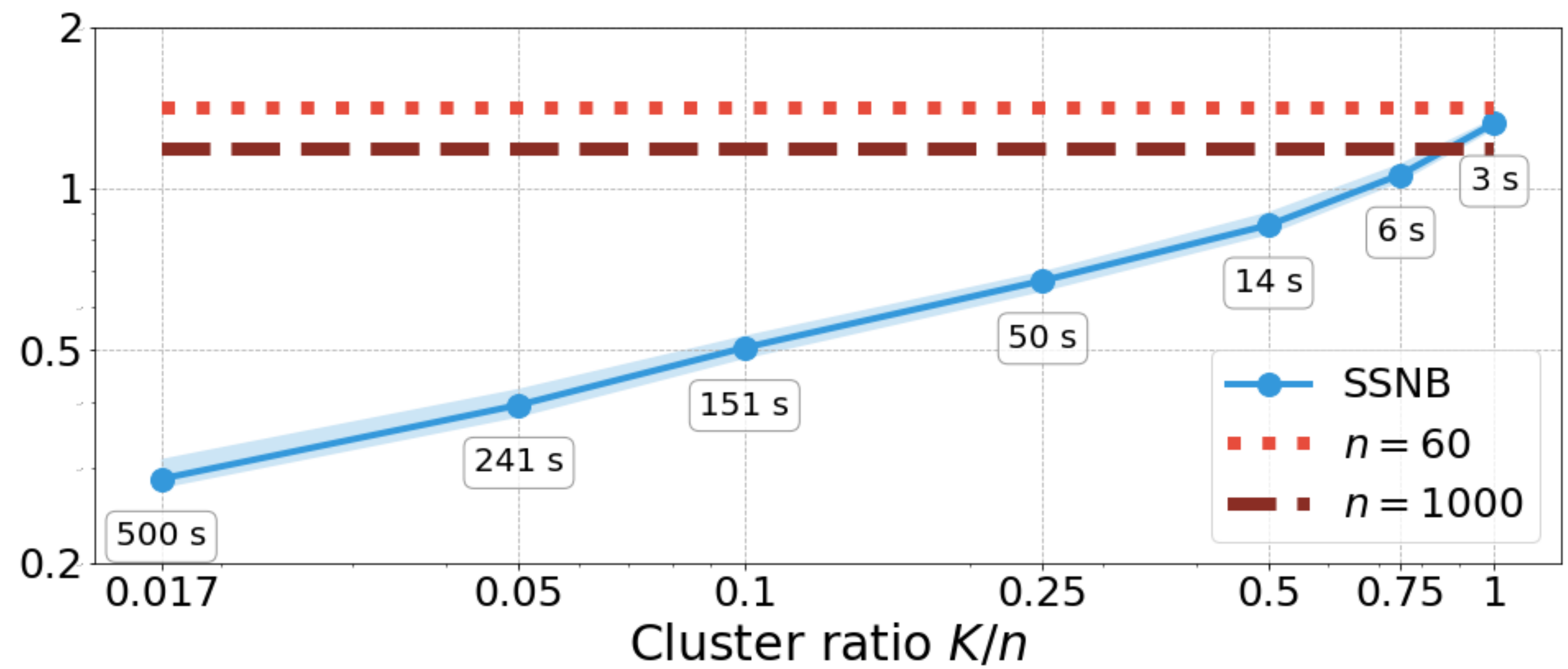
Regularity "by part"

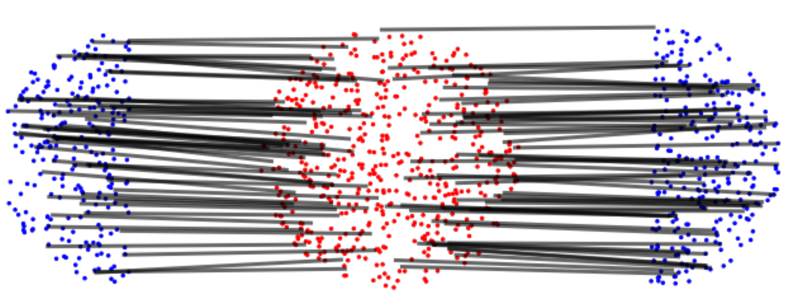


Estimation Error depending on K

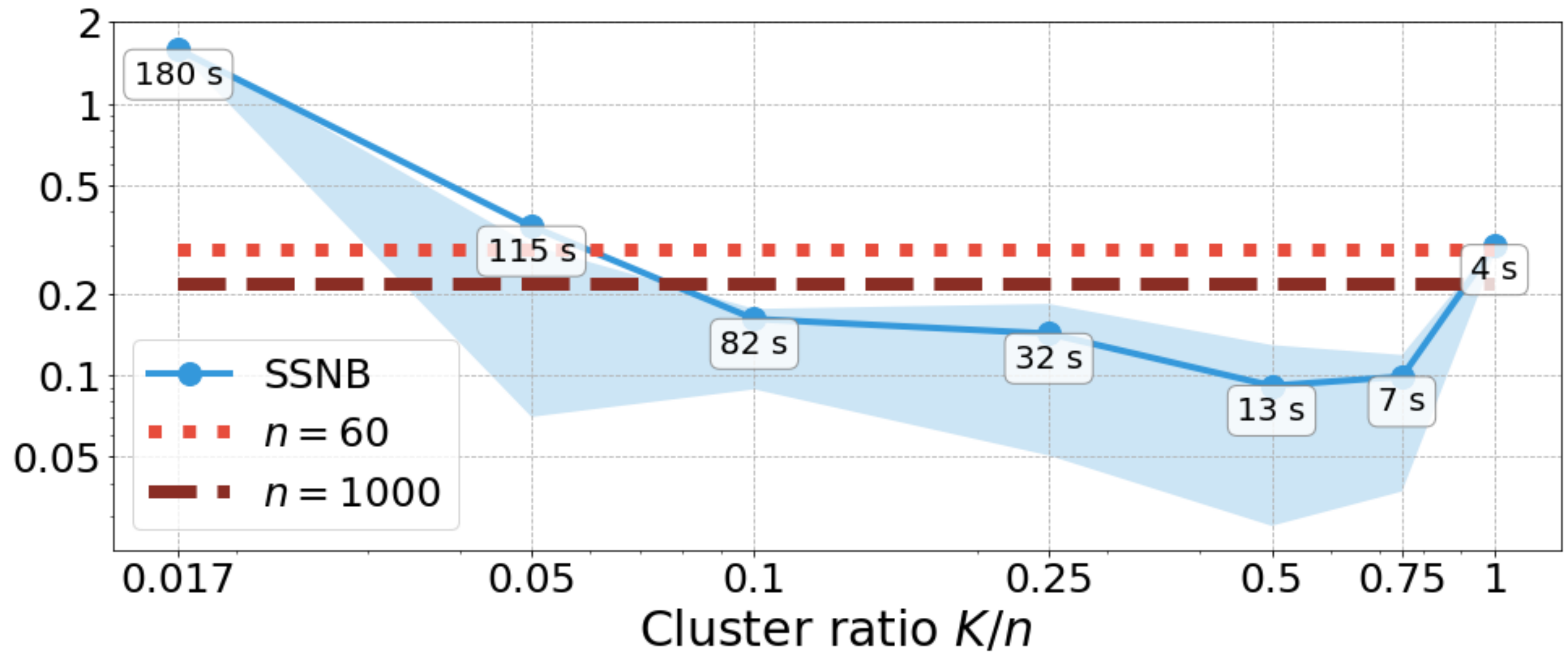


Global Regularity

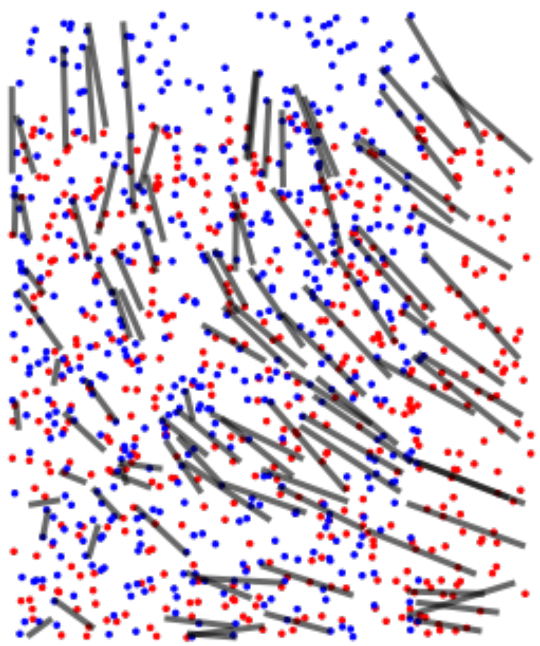




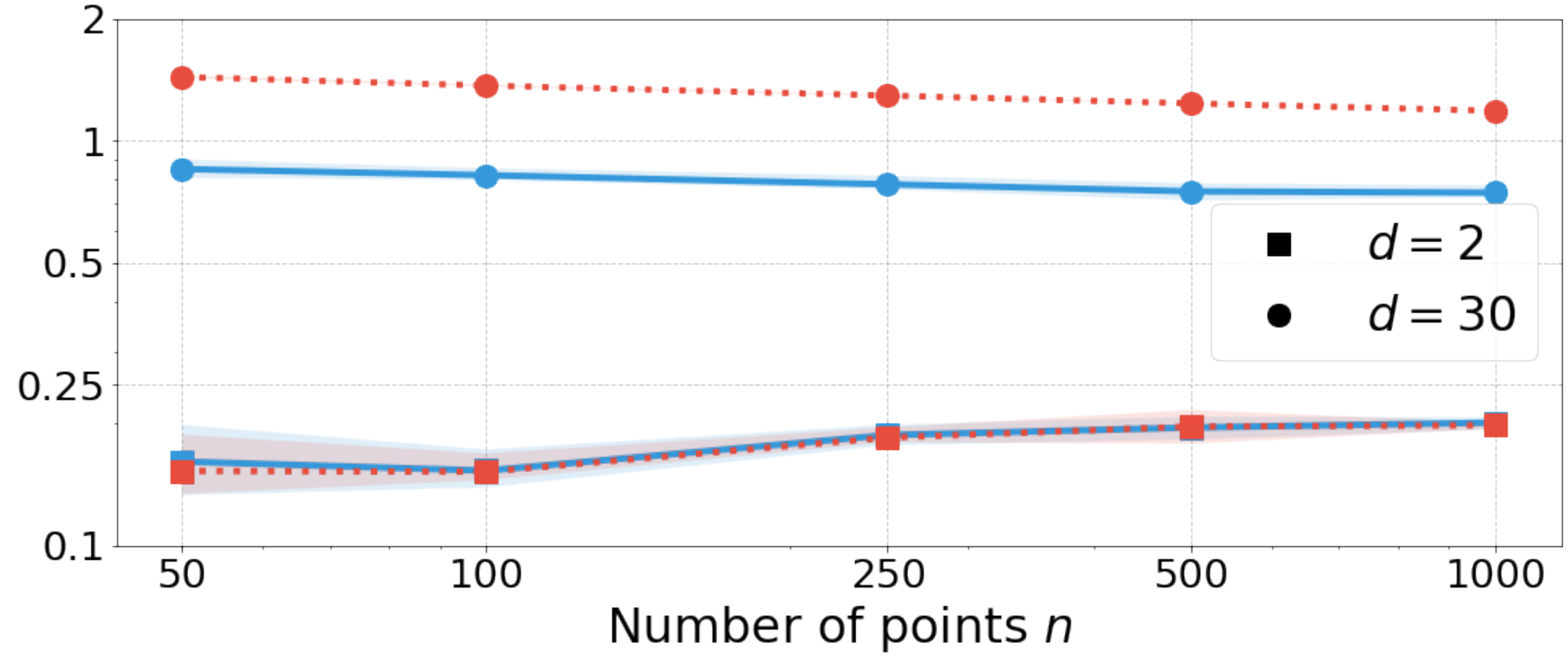
Local Regularity

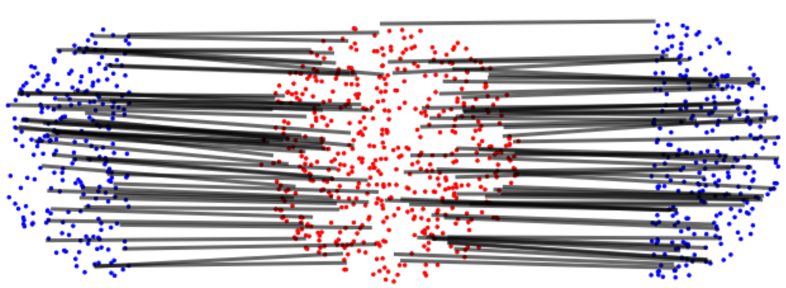


Estimation Error depending on n

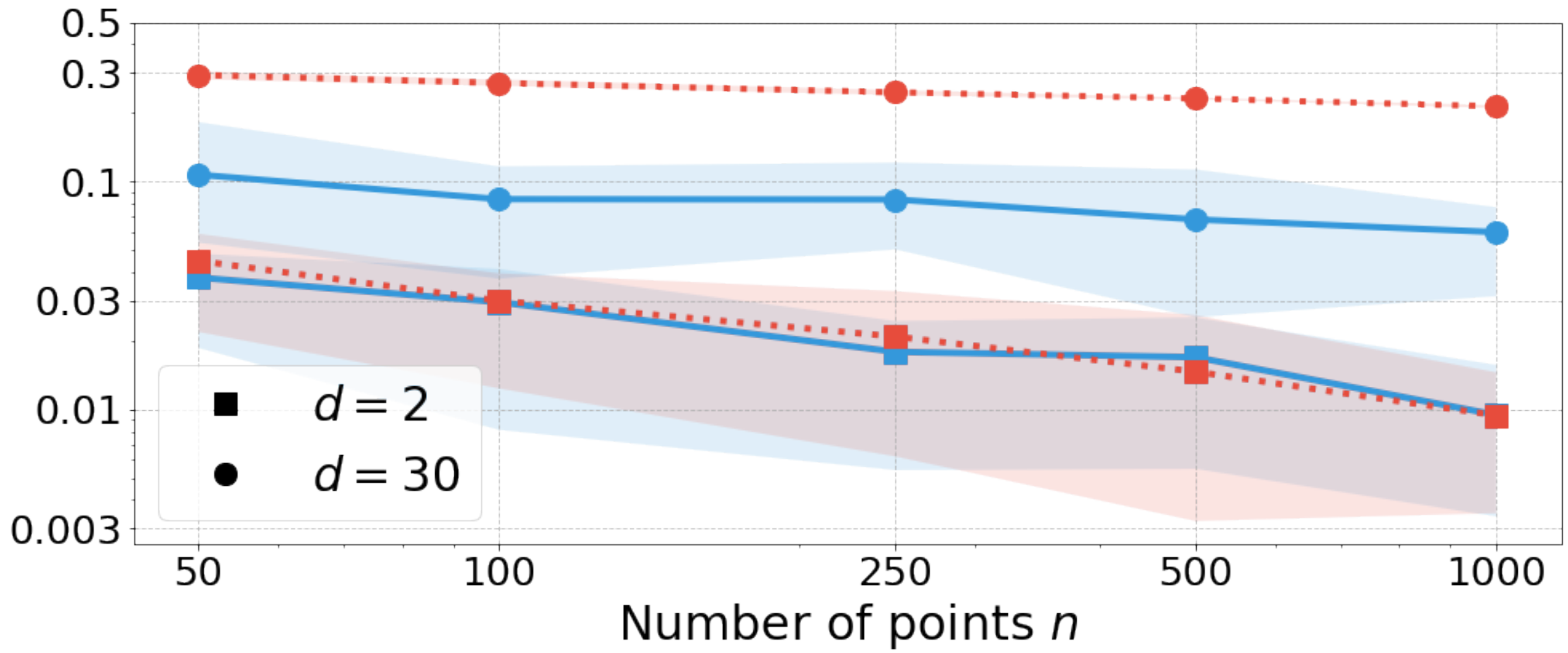


Global Regularity

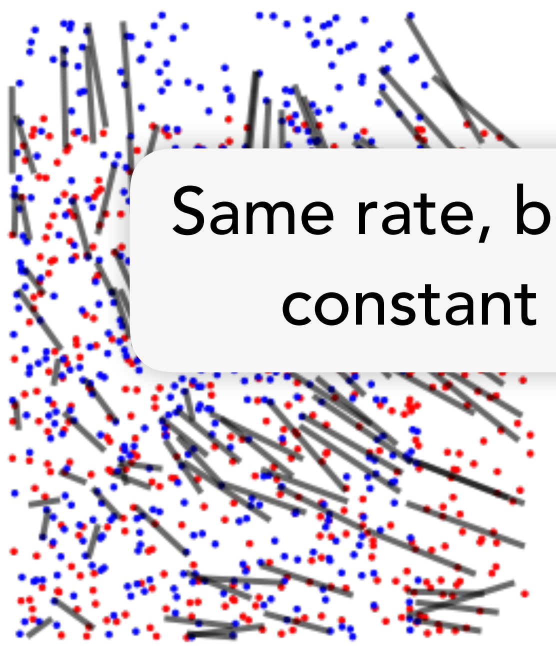




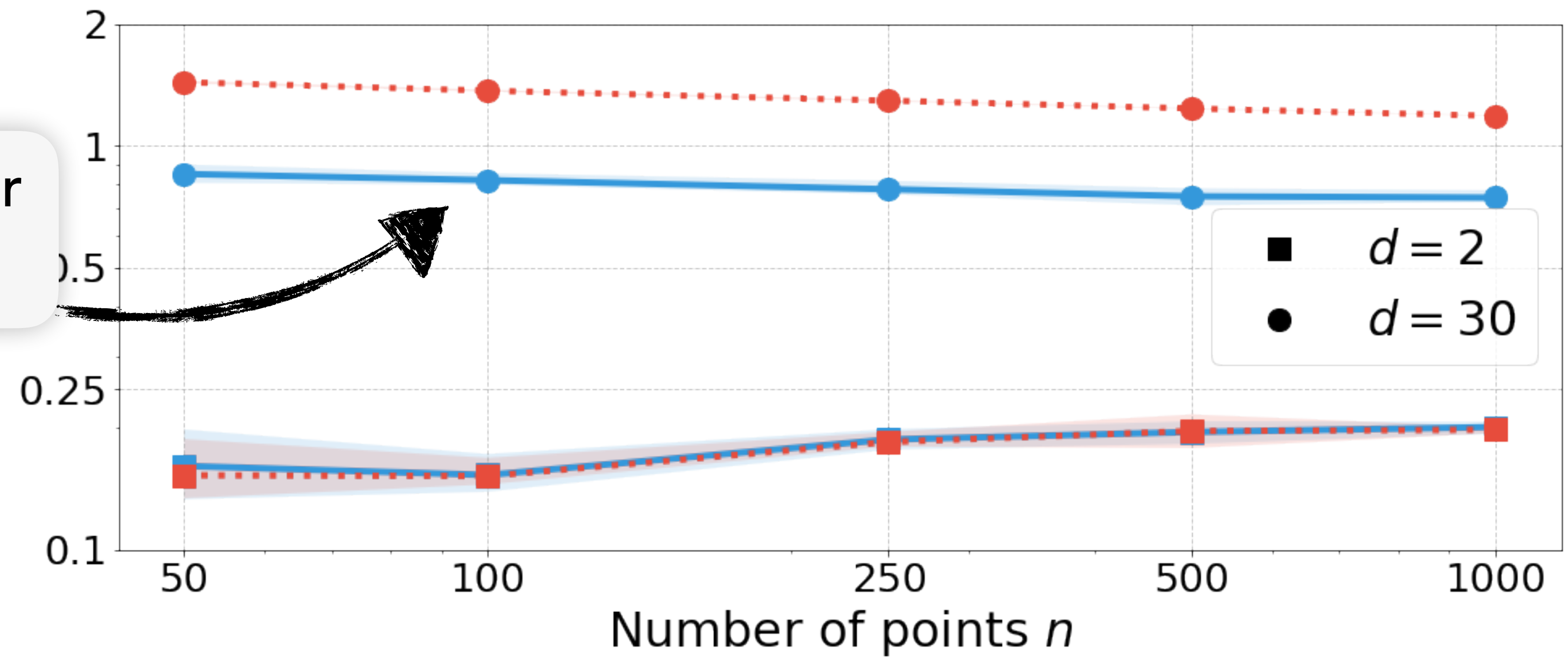
Local Regularity



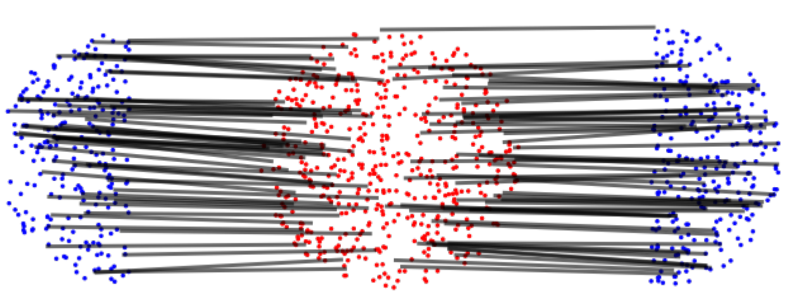
Estimation Error depending on n



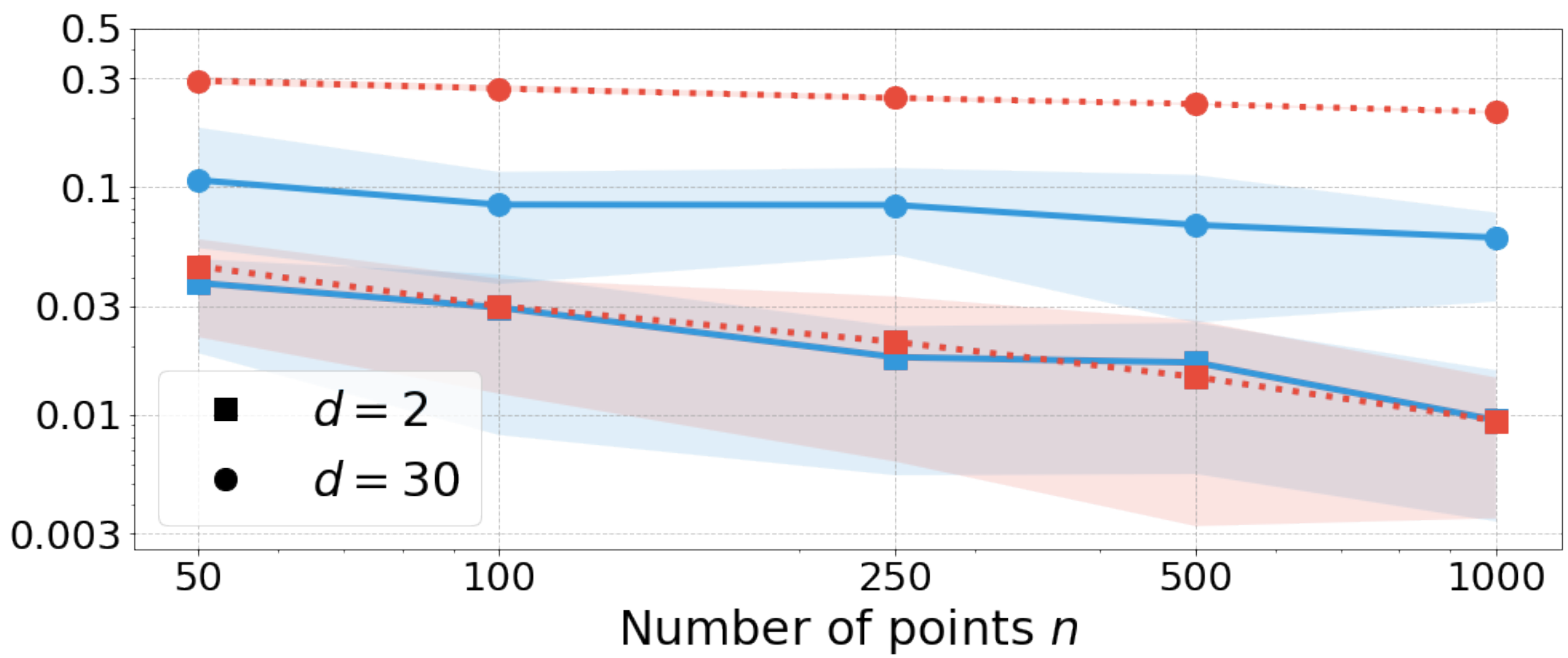
Same rate, better constant ?

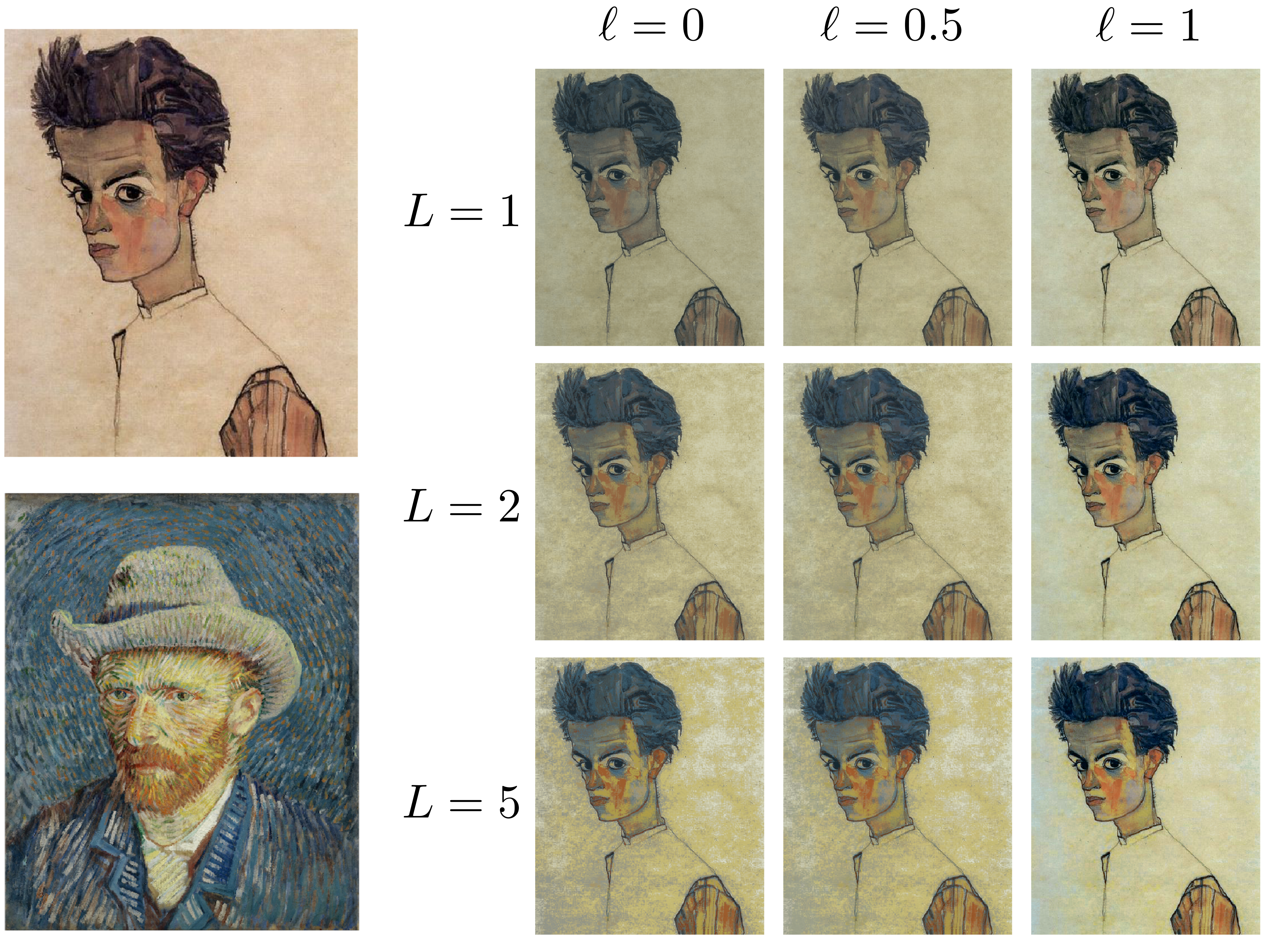


Global Regularity



Local Regularity





$\ell = 0$
$\ell = 0.5$
$\ell = 1$
$L = 1$
$L = 2$
$L = 5$


QUESTIONS ?
Cédric Villani