User-friendly optimal transport in biology

a non-comprehensive survey in pictures and conversations

Alexandra Suvorikova [1, 2, 3]

[1]Weierstrass Institute for Applied Analysis and Stochastics, Berlin

[2]Kharkevich Institute for Information Transmisson Problems RAS, Moscow,

[3]Higher School of Economics, National Research University, Moscow

November 27, 2020

Outline

Intro to intro

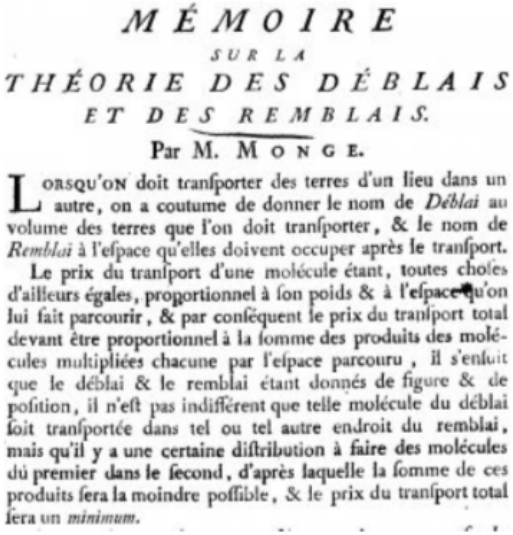
Before we begin

Sources:

- Book by M.Cuturi and G. Peyré, Peyré et al. [2019] and `https://optimaltransport.github.io`
- Twitter and tutorials of G. Peyré `http://www.gpeyre.com`
- Page of M.Cuturi `https://marcocuturi.net`
- Talk by G. Geoffrey Schiebinger at `https://sites.google.com/view/otml2019/`, see Session 3
- Works by *M.Cuturi, G.Peyré, J.Solomon, N. Lei, A.Kroshnin* and many others

Monge-Kantorovich problem

Origin of the problem

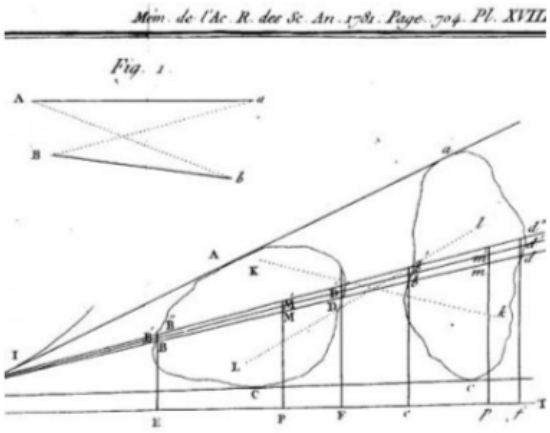


MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

Par M. MONGE.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.

Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'ensuit que le déblai & le remblai étant donnés de figure & de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, & le prix du transport total sera un *minimum*.

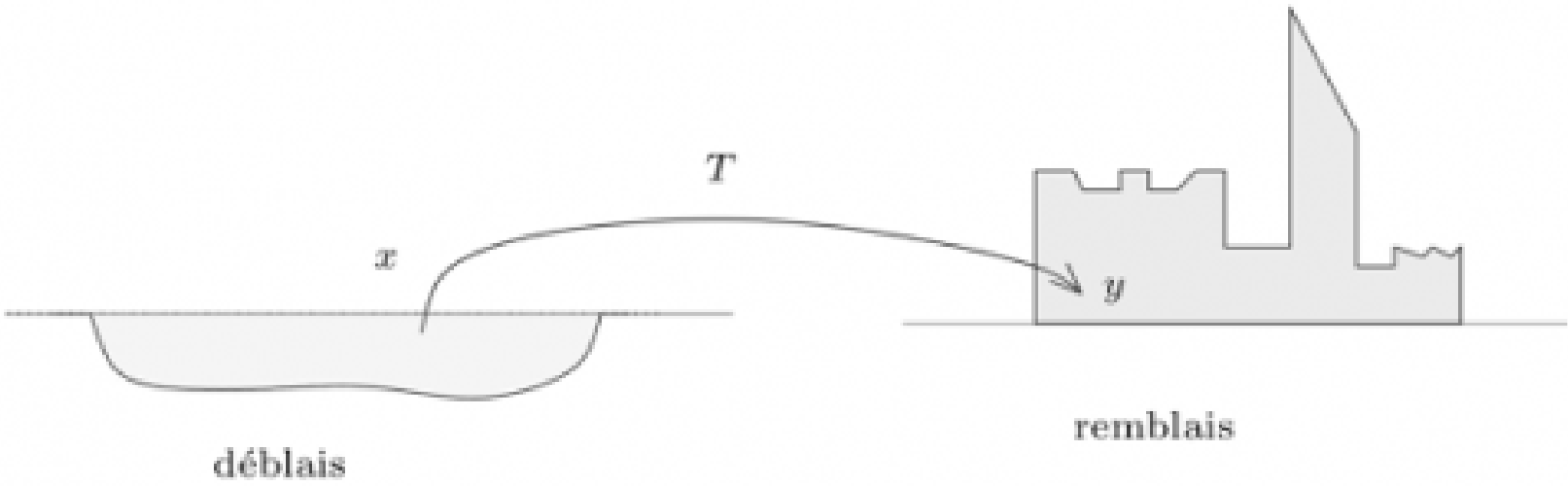


Mém. de l'Ac. R. des Sc. An. 1781. Page. 709. Pl. XVII.

Fig. 1

(1784)

Monge Problem



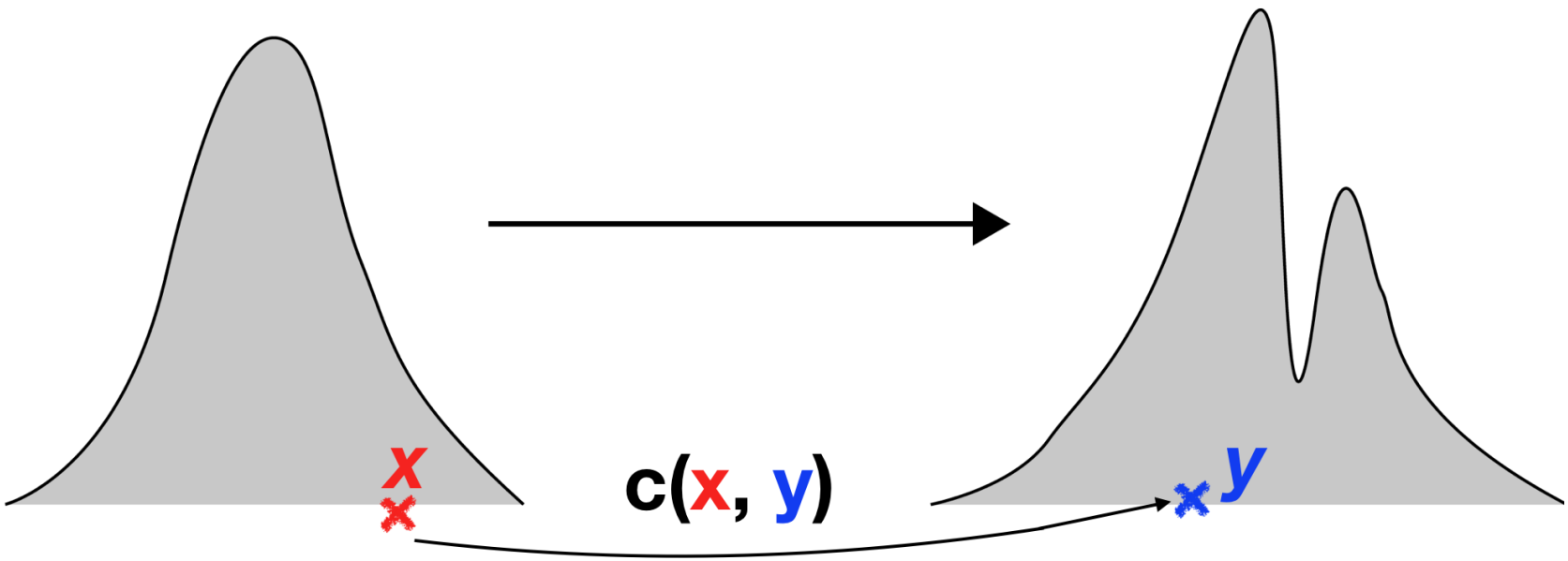
Origin of the problem

- μ , ν prob. measures supported on $\mathcal{X}$ and $\mathcal{Y}$
- cost function $c(x, y)$ $x \in \mathcal{X}$, $y \in \mathcal{Y}$

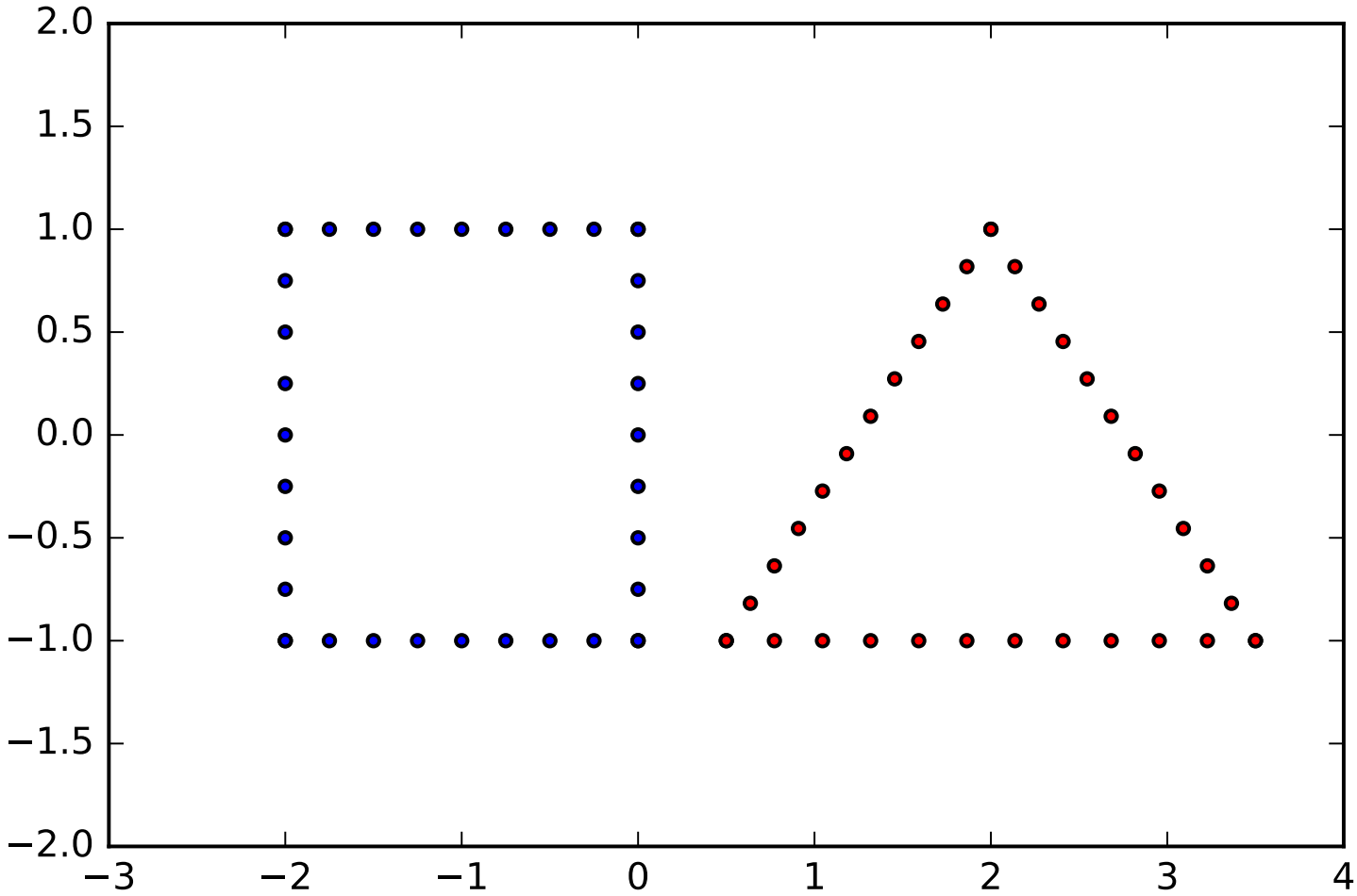
$$d_M(\mu, \nu) = \min_{T : T_{\#}\mu = \nu} \int_{\mathcal{X}} c(x, T(x)) d\mu(x)$$

Measure-preserving maps:

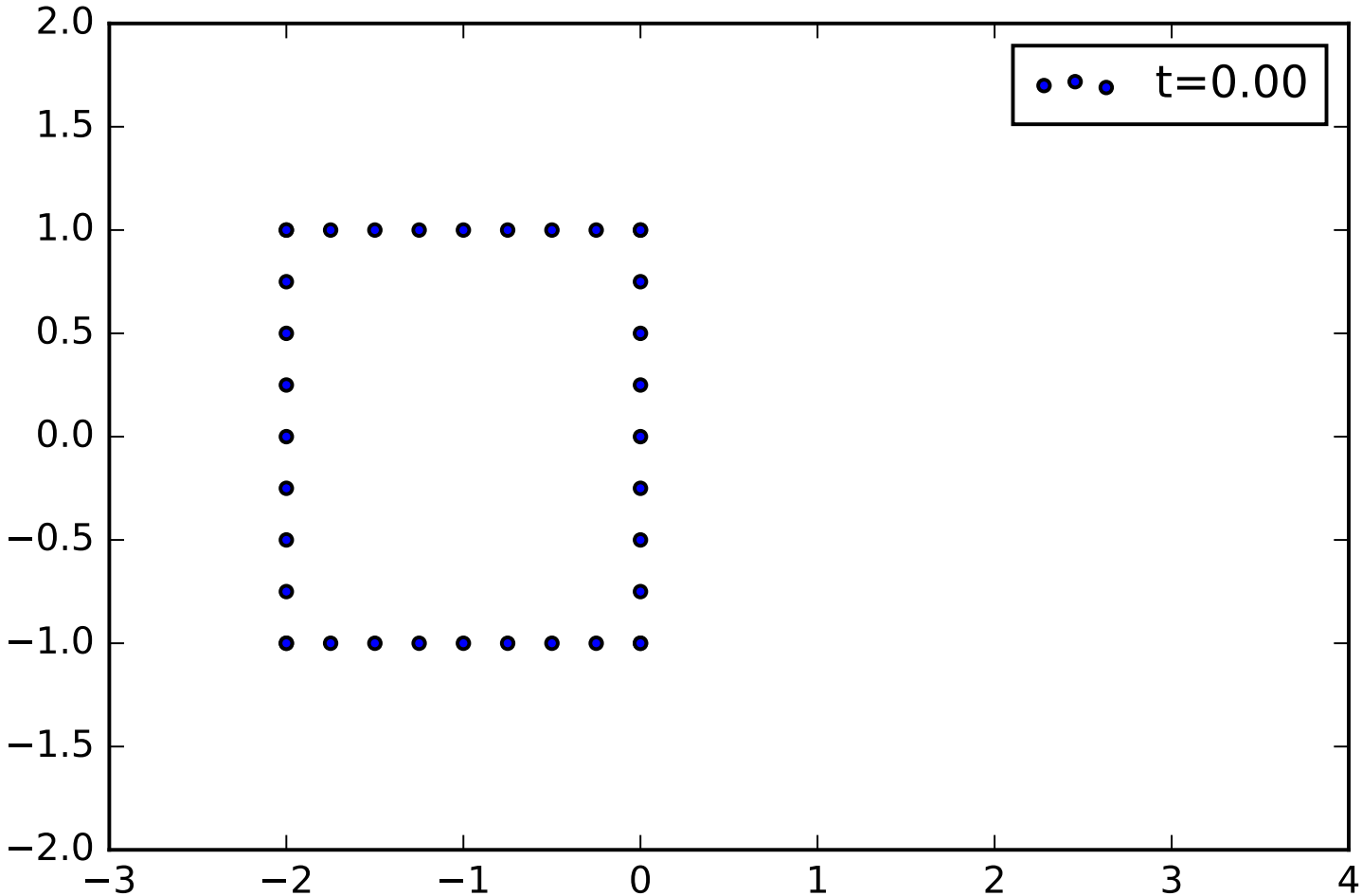
$$\forall \mathcal{A} \in \mathcal{Y}, \quad b(T^{-1}(\mathcal{A})) = \nu(\mathcal{A})$$



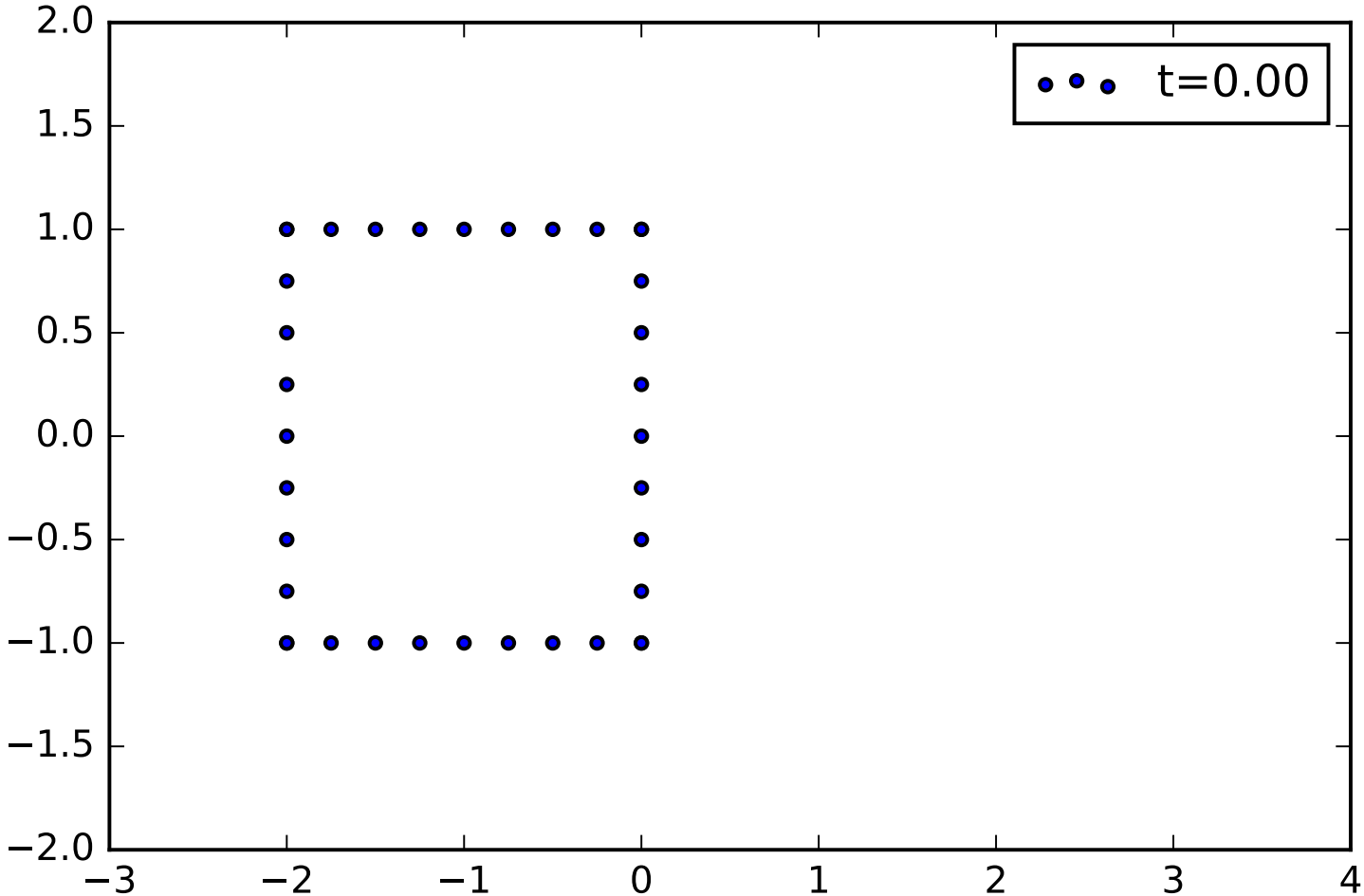
Optimal mass transportation



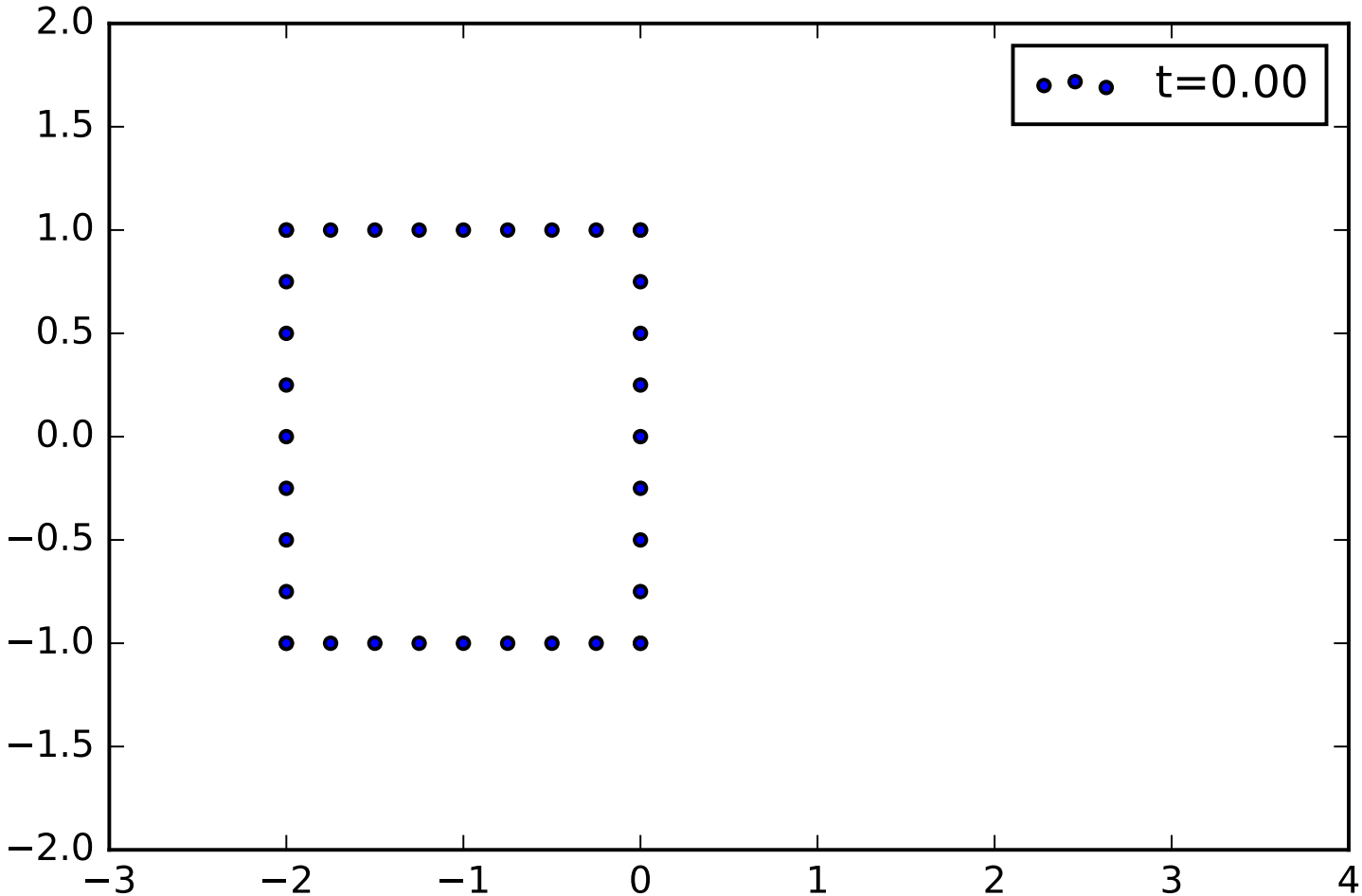
Optimal mass transportation - I



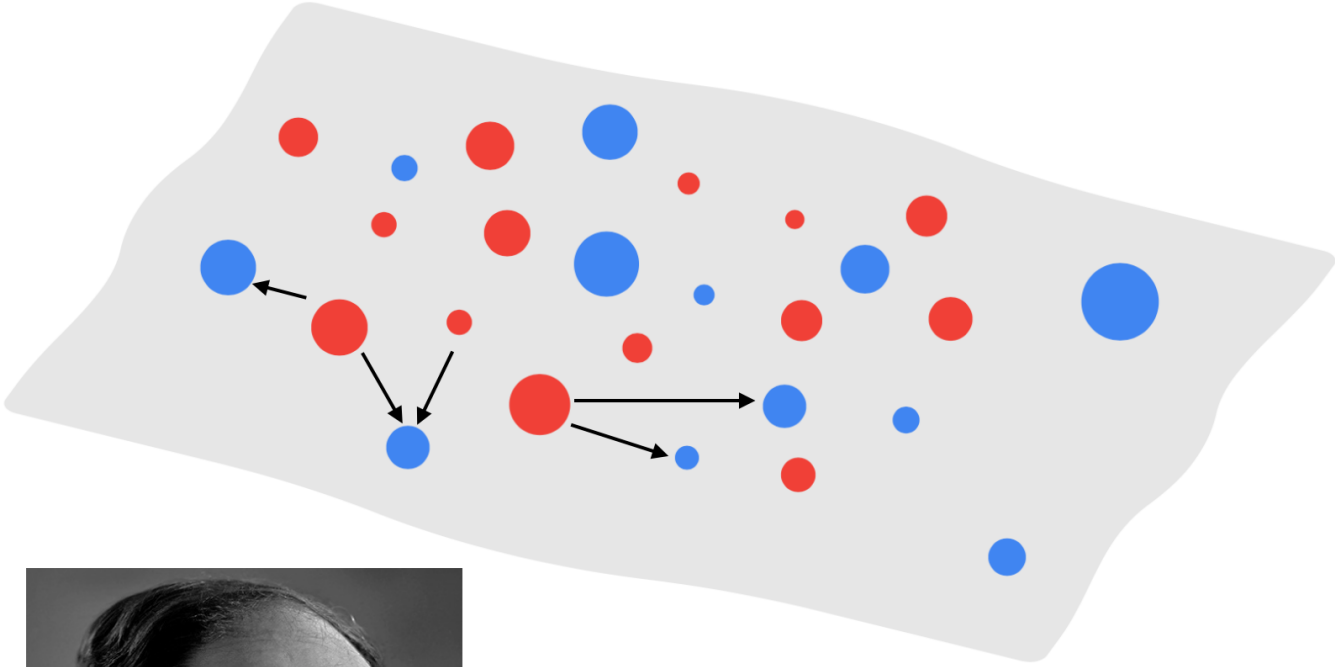
Optimal mass transportation - II

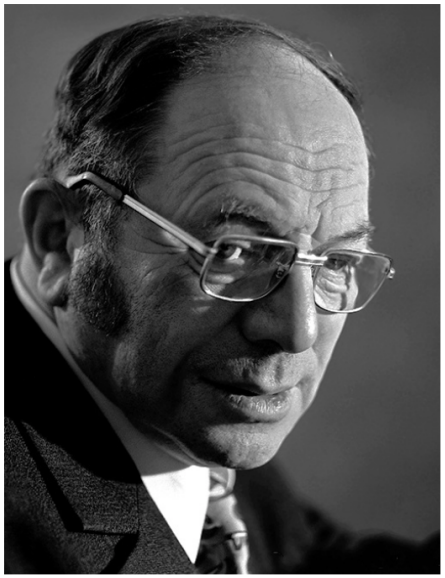


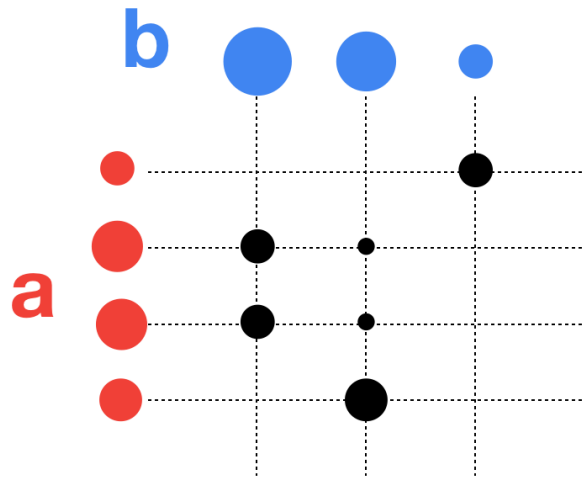
Optimal mass transportation - III



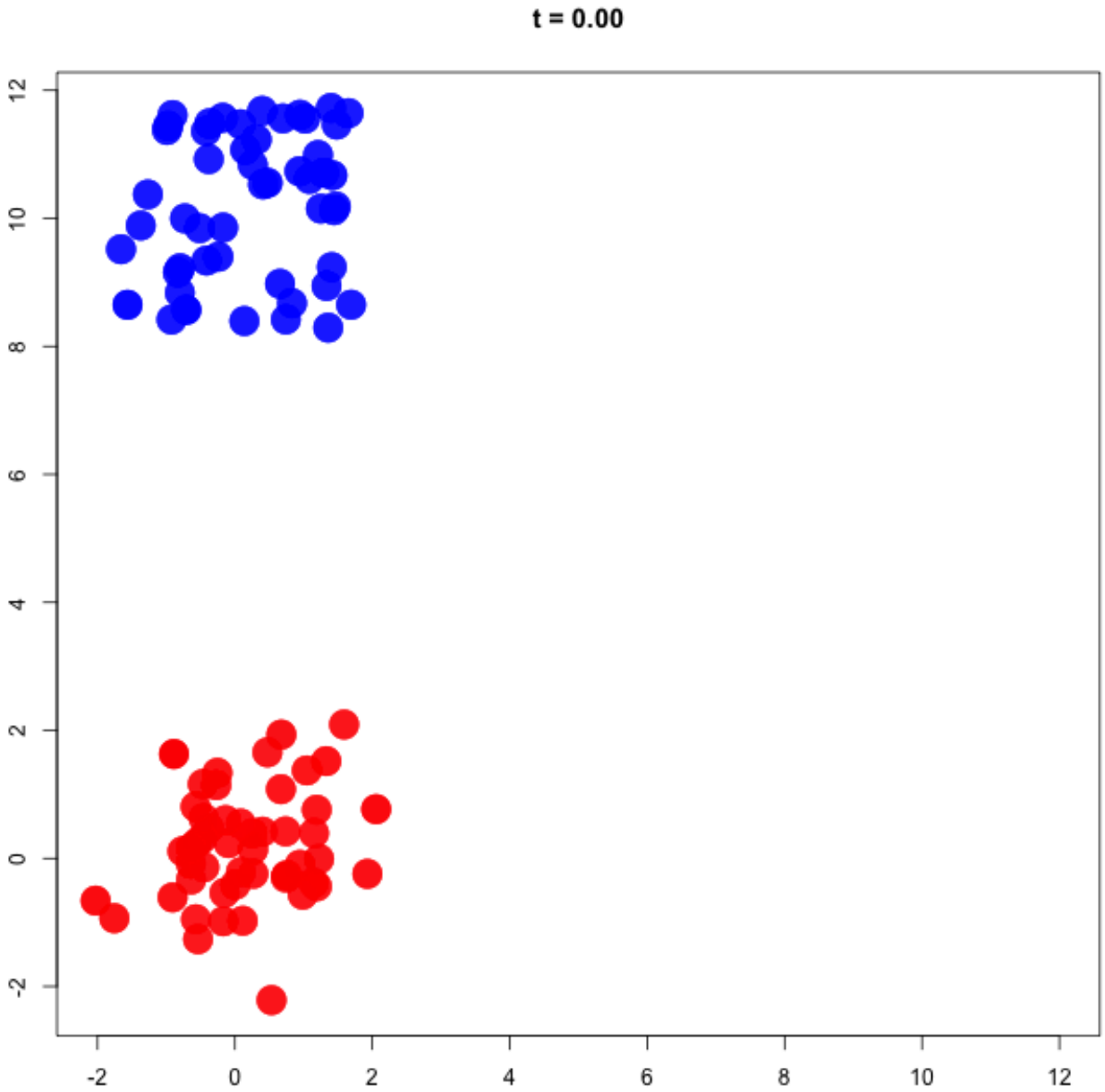
Kantorovich: mass split



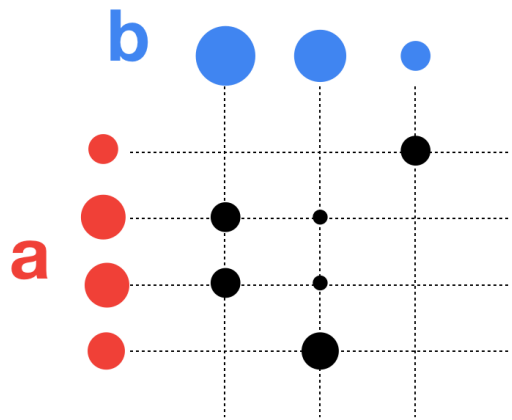




Mass split



Monge-Kantorovich-Problem



$$a = \sum_i a_i \delta_{X_i}, \quad b = \sum_j b_j \delta_{Y_j}, \quad \sum_{i=1}^{N} a_i = \sum_{j=1}^{M} b_j = 1$$

we choose cost function $c(X_i, Y_j) := c_{ij}$

$$d_{MK}(a, b) := \min_{P \in \Pi_{a,b}} \langle P, C \rangle = \min_{P \in \Pi_{a,b}} \sum_{i,j} P_{ij} c_{ij}$$

$$\Pi_{a,b} = \left\{ P \in \mathbb{R}^{N \times M} : P \mathbb{I}_N = a, P^T \mathbb{I}_M = b \right\}$$

Regularization

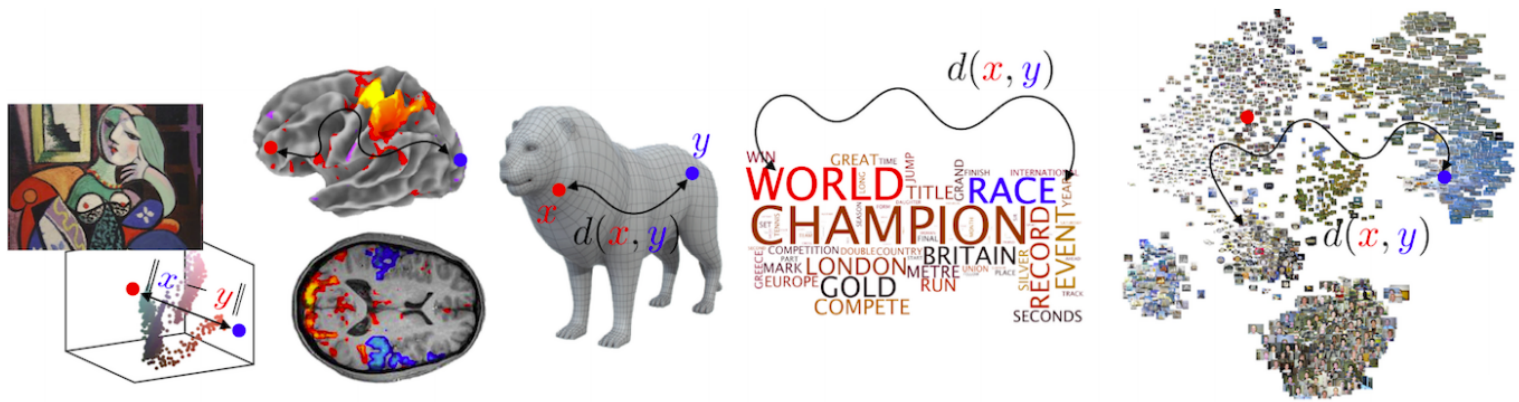
- Computation tractability: *regularization* Peyré et al. [2019]

$$d^2_{MK,\gamma}(a, b) := \inf_{P \in \Pi_{a,b}} \left[\sum_{i,j} c(X_i, Y_j) P_{ij} + \gamma H(P) \right]$$

Probability distributions in data science

Statistical inference: samples (point clouds) can be interpreted as probability distributions

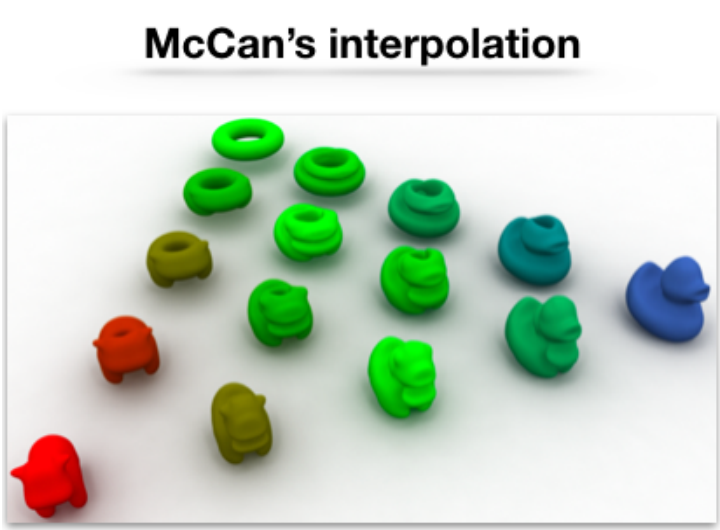
- non-linear spaces: medical images, 3D shapes, phylogenetic trees, word emeddings
- low-dimensional data in high dimensions: GANs, autoencoders

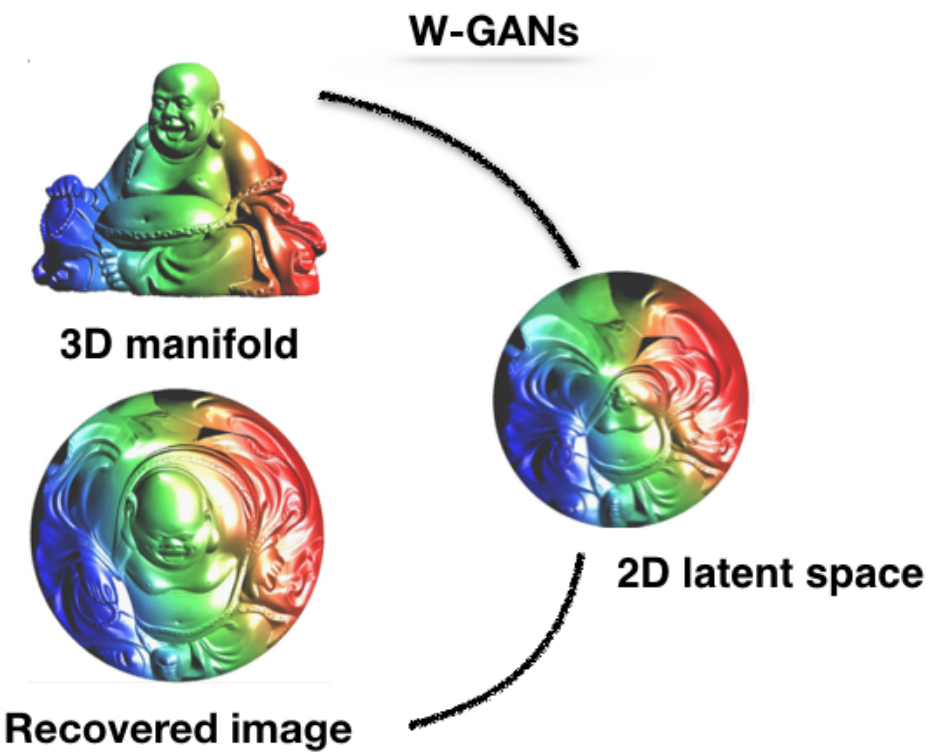


"Wassersteinization"

OT distance is a suitable replacement for... anything?

- **As a distance:** barycenters, Mc Cann's displacement interpolation; Wasserstei PCA
- **As a loss function:** GANs; transfer learning, recovery of group-related dynamics





Survey

Enormous amount of literature:

- **NeurIPS'19: over 100** submissions containig "Wasserstein" or "Optimal Transport" in in the title
- **ICML'19: over 50** submissions

A comprehensive survey on theoretical backgrounds of OT: Villani [2009], Santambrogio [2010]

Computation aspects: Peyré et al. [2019]

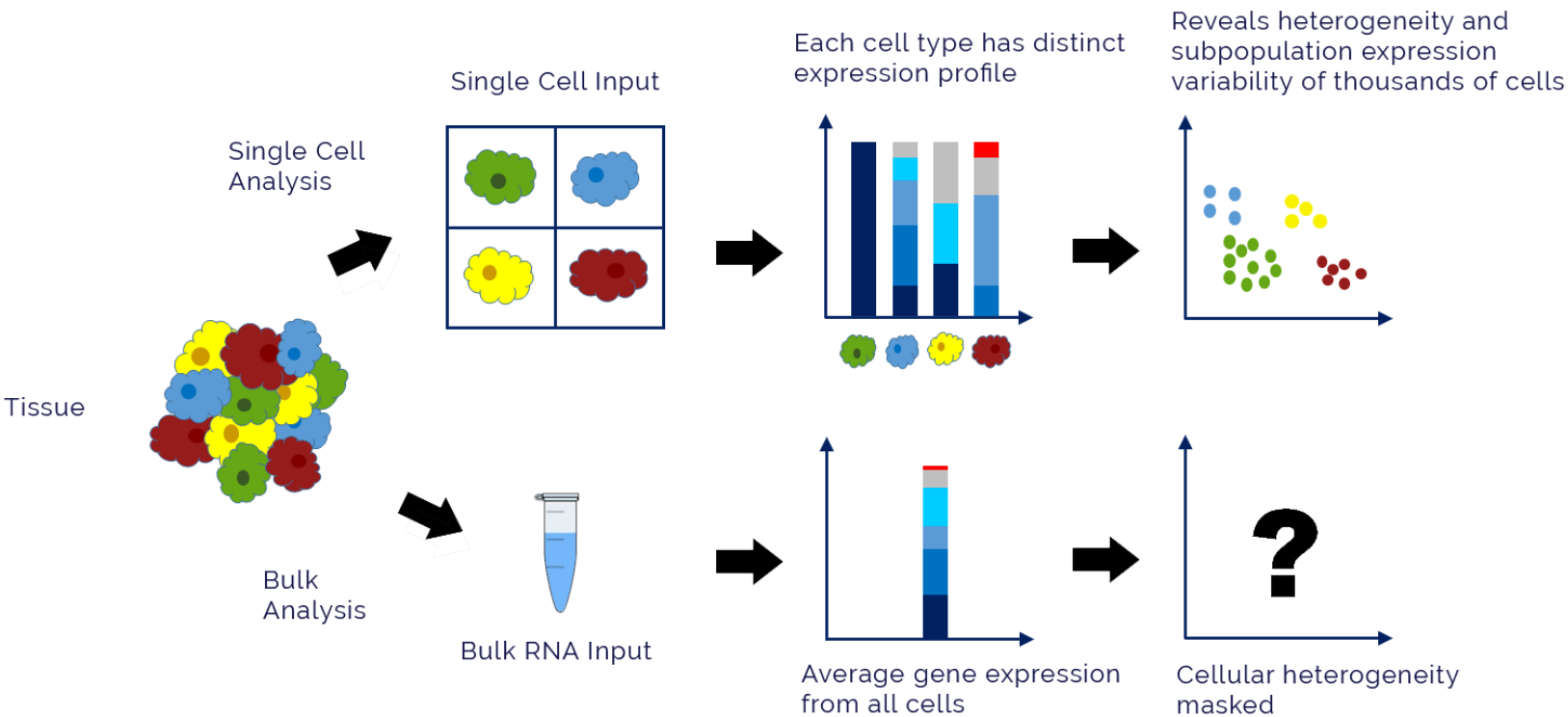
OT in ML

- `marcocuturi.net`
- `www.gpeyre.com`; twitter @gabrielpeyre
- `otml2019.github.io`

Single cell genome analysis

Single-cell data

- Single-cell Hi-C
- Single-cell gene expression based on Schiebinger et al. [2019]



Topologically associated domains in DNA ?

Spatial geometry is important for gene expression regulation

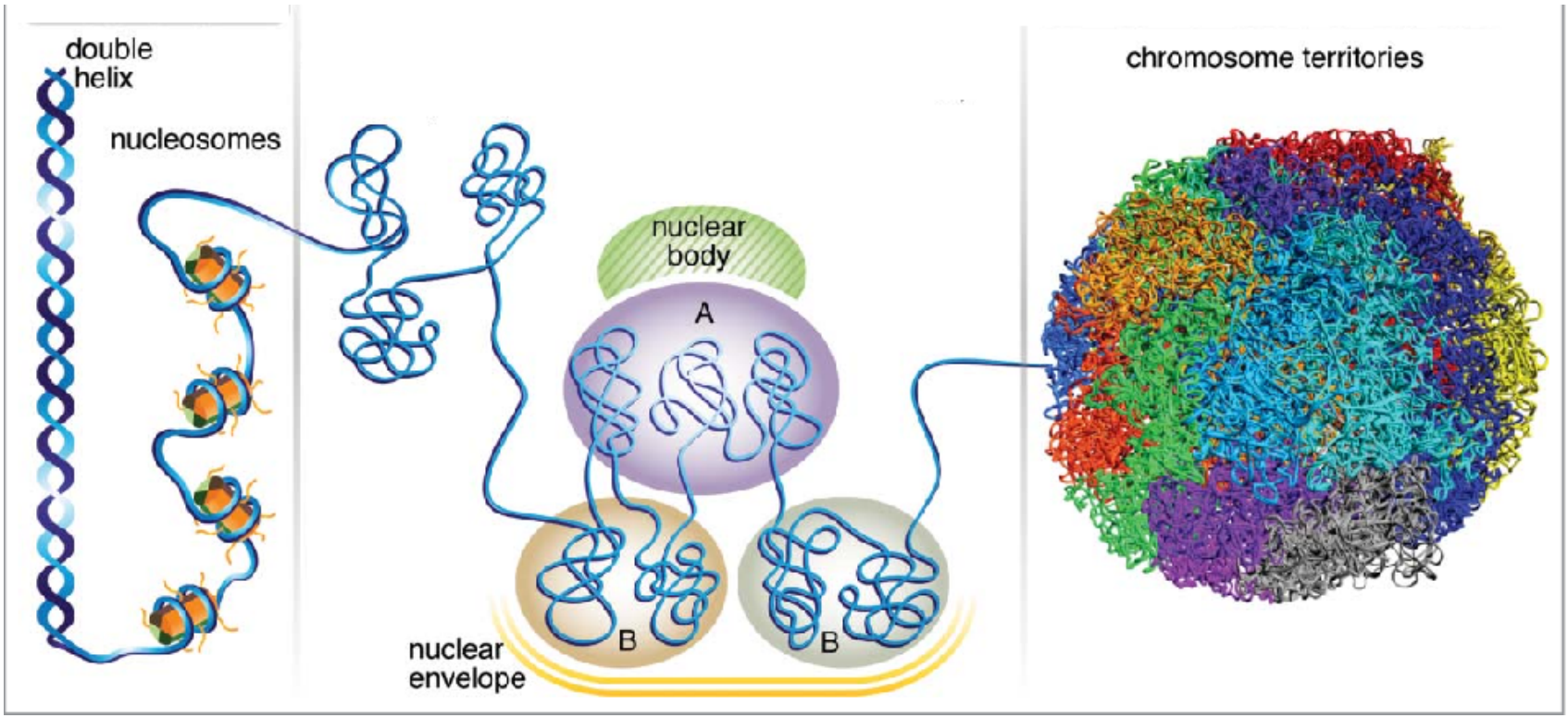


Figure: ?

Hi-C matrix for a fixed cell-type

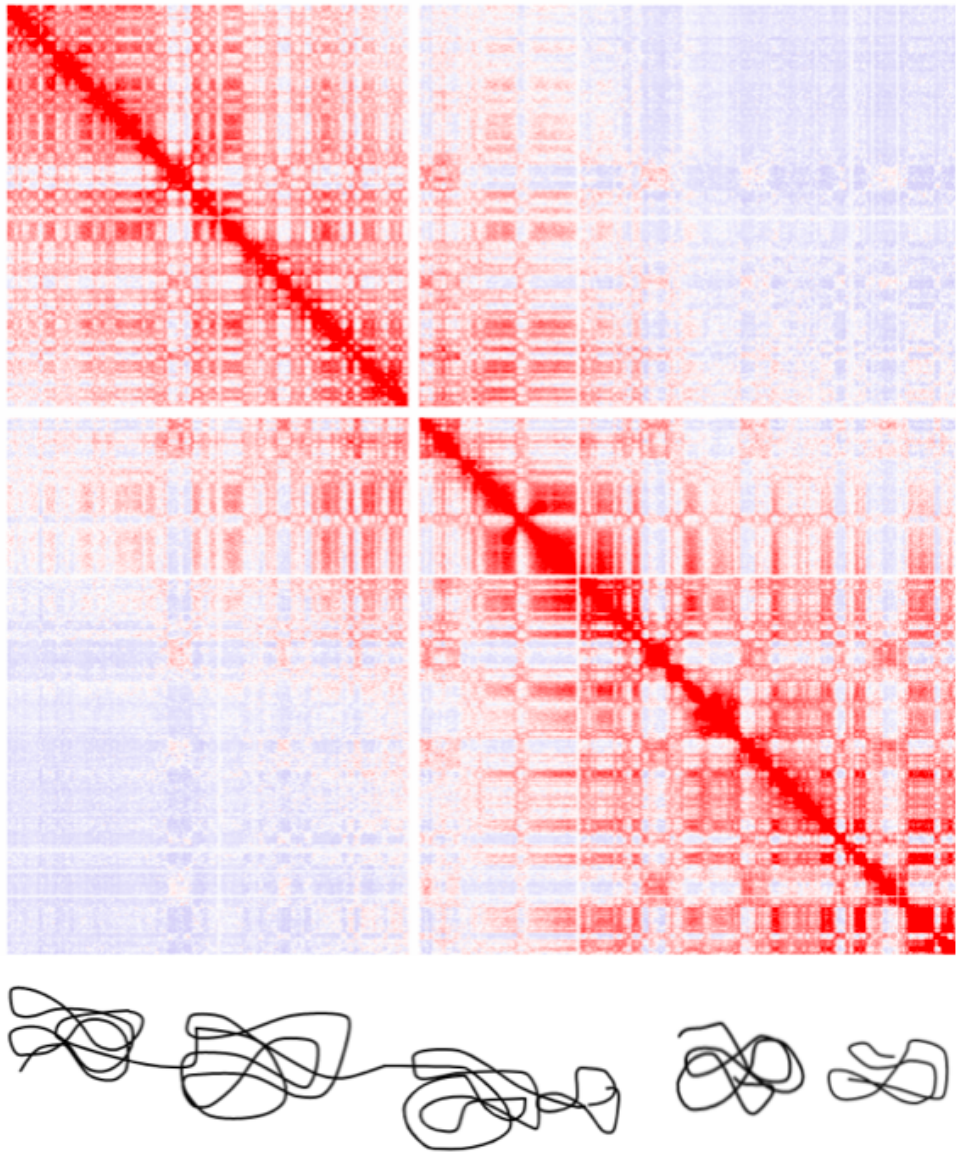


Figure: Hi-C matrices vs. chromosome alignment

Avergaing

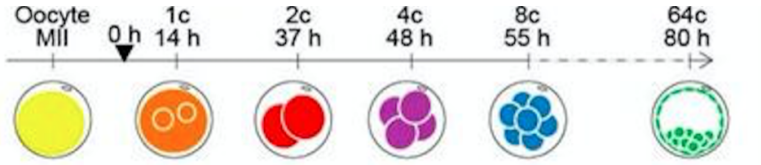
Data $S_1, \dots S_n$, $S_i \overset{iid}{\sim} P$, P on $\mathcal{X}$
Goal: compute mean w.r.t. d_{MK} with some cost function

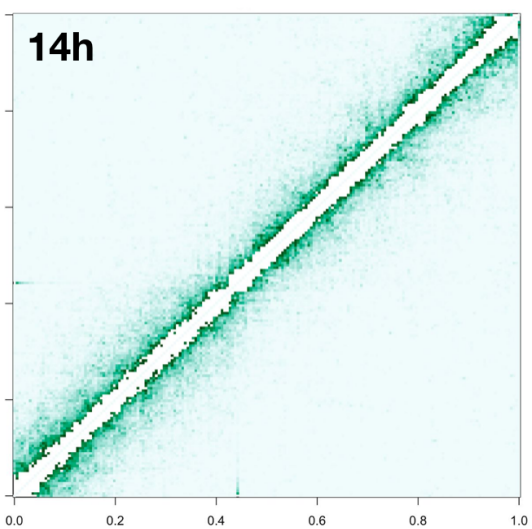
$$Q_* \overset{\text{def}}{=} \operatorname*{argmin}_{Q \in \mathcal{X}} \int d^2_{MK}(S, Q) dP(S), \quad Q_n \overset{\text{def}}{=} \operatorname*{argmin}_{Q \in \mathcal{X}} \sum_i d^2_{MK}(S_i, Q)$$

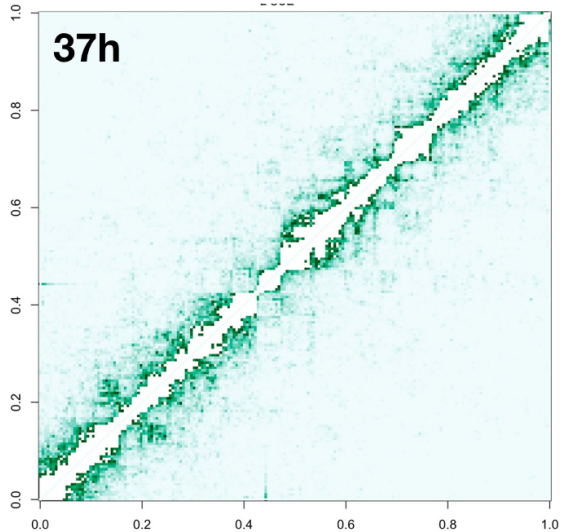
Questions to Q_n

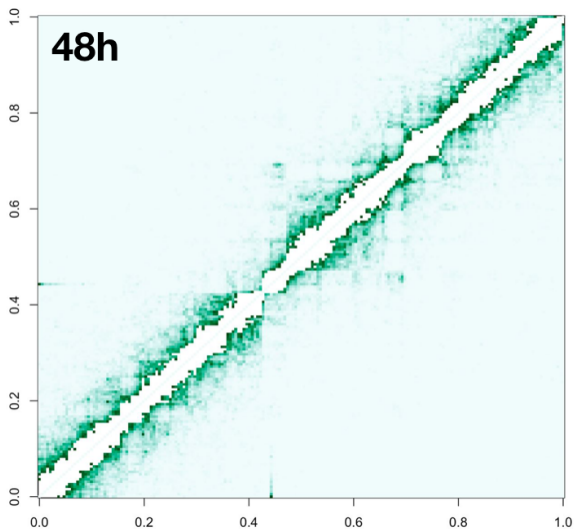
- consistency
- concentration
- CLT
- confidence regions

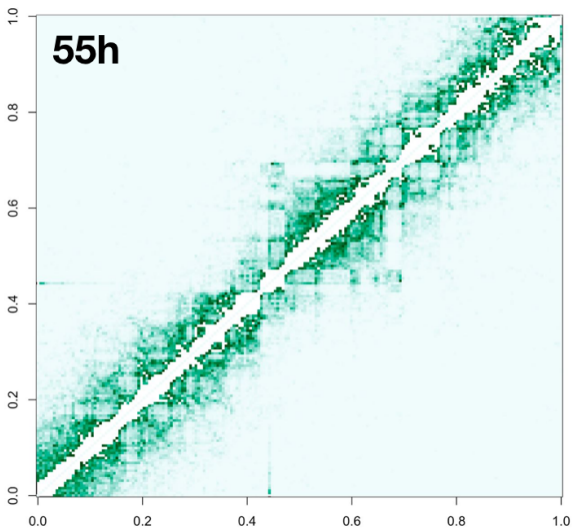
Embryogenesis in mice [work in progress]

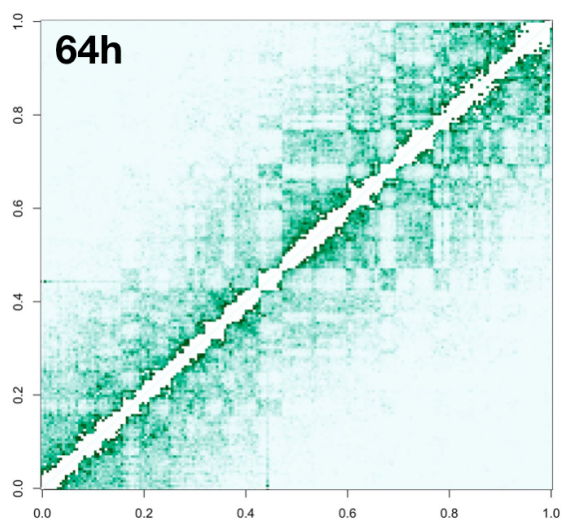












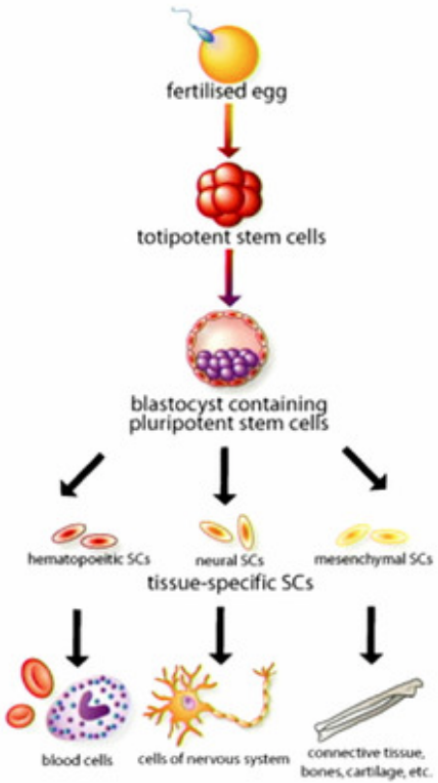
Talk by G.Schiebeinger at NeurIPS'19

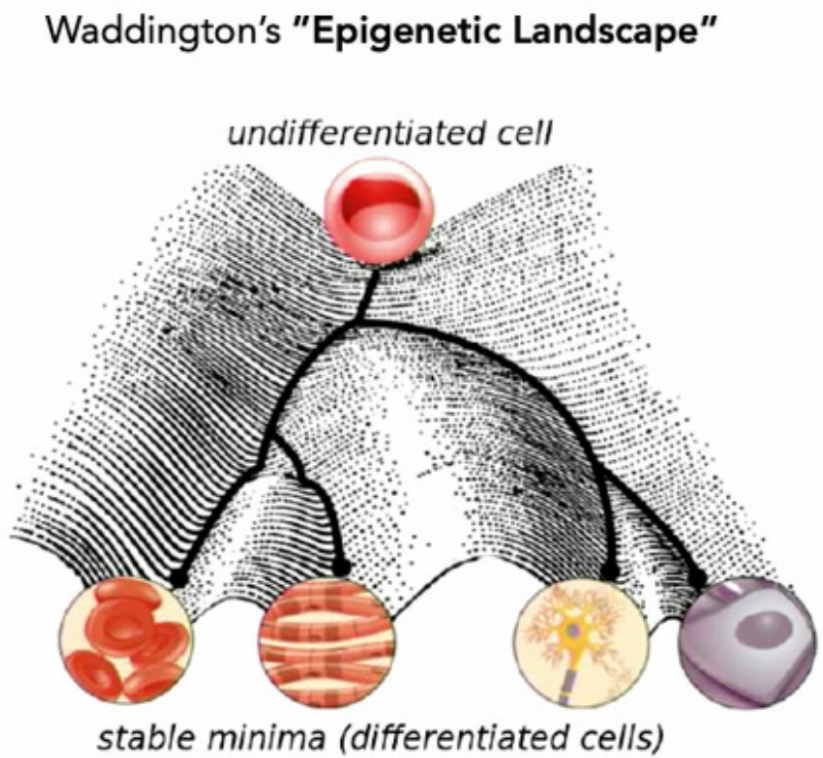
- Paper: Schiebinger et al. [2019]
- Talk by Geoffrey Schiebinger at NeurIPS'19, see `https://sites.google.com/view/otml2019/`, Session 3
- the bioinformatics chat `https://bioinformatics.chat/optimal-transport`

Model

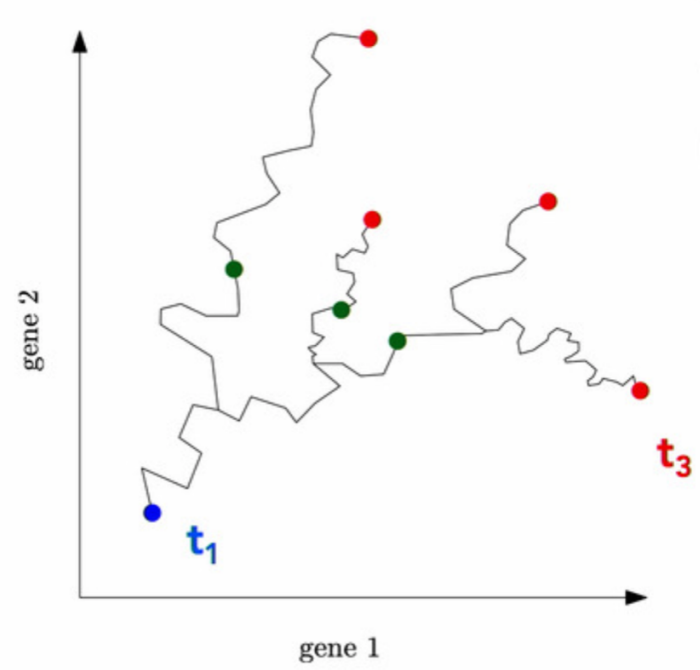
- Each cell: $x \in \mathbb{R}^{\text{num of genes}}$
- Each dim: how many RNA molecules for this gene

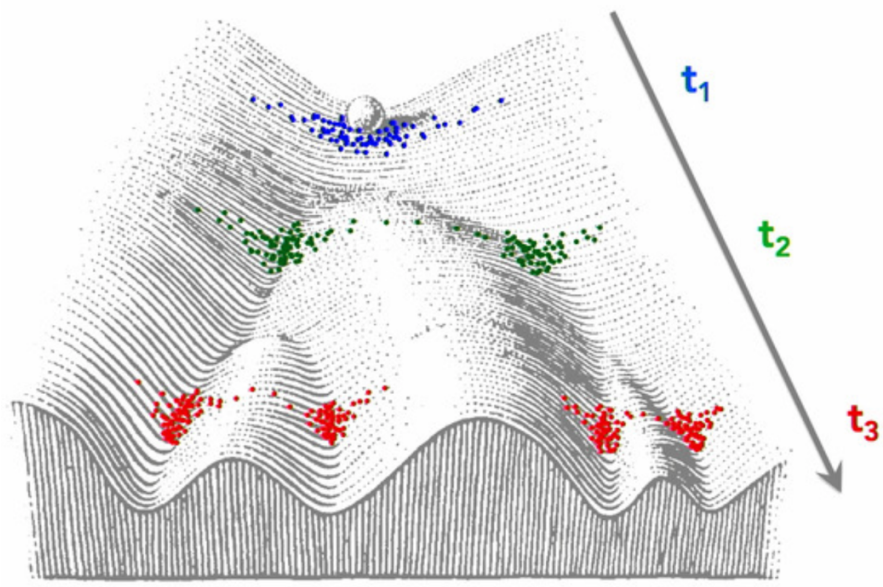
Goal: describe developmental process of stem cells





Stem cell evolution





Cells change gene expression over time.

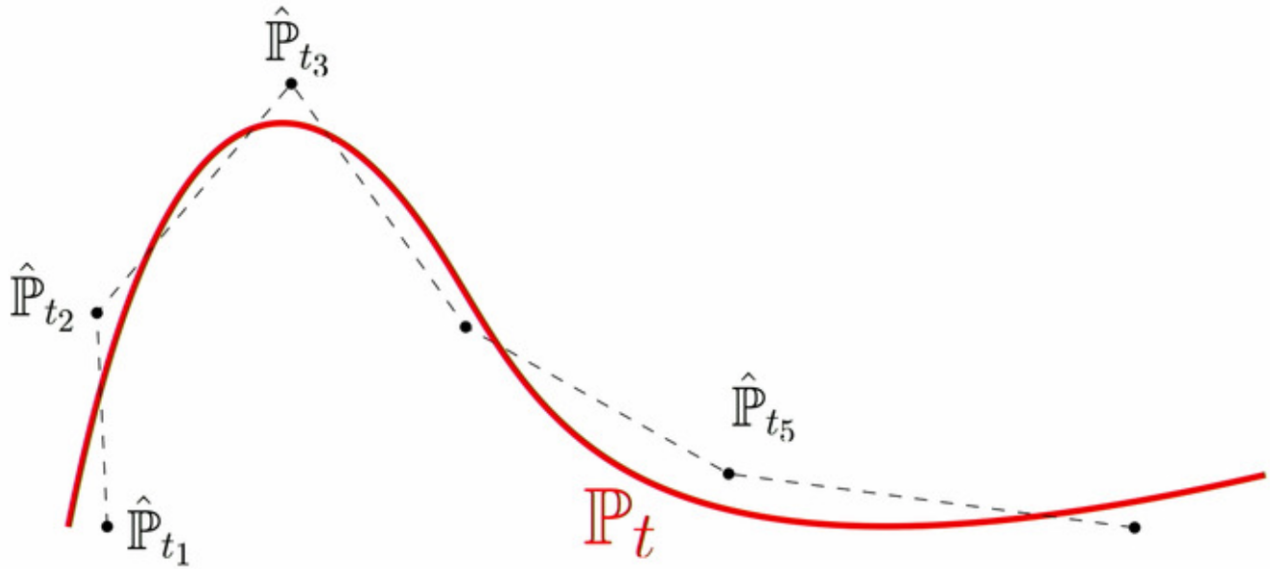
Cell division creates branching paths.

Measurement kills cells so we cannot observe paths!

Intuition behind

Intuition

The process is **locally linear** in Wasserstein space:



Potential problems

$\mathcal{Q}$: How to model cell proliferation? (Increase of mass!)

$\mathcal{A}$: Non-balanced optimal transport

$\mathcal{Q}$: How to model cell differentiation?

$\mathcal{A}$: Entropic regularization

References I

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. 2013.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Filippo Santambrogio. Introduction to optimal transport theory. 2010.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4): 928–943, 2019.

References II

Cédric Villani. *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-71049-3 978-3-540-71050-9.