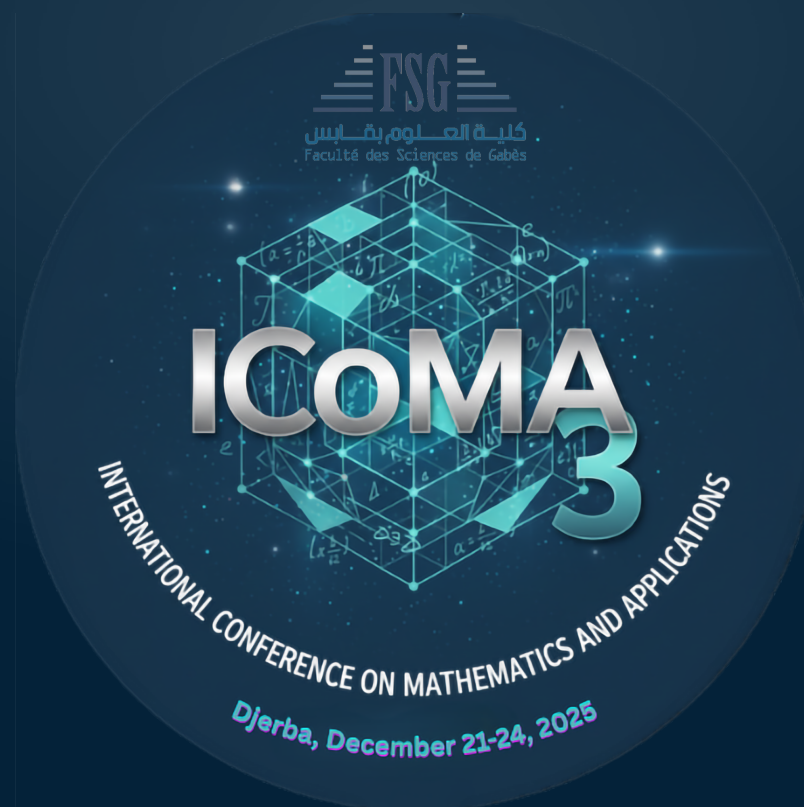


Optimal transport: A Swiss Army Knife for Machine Learning Pipelines

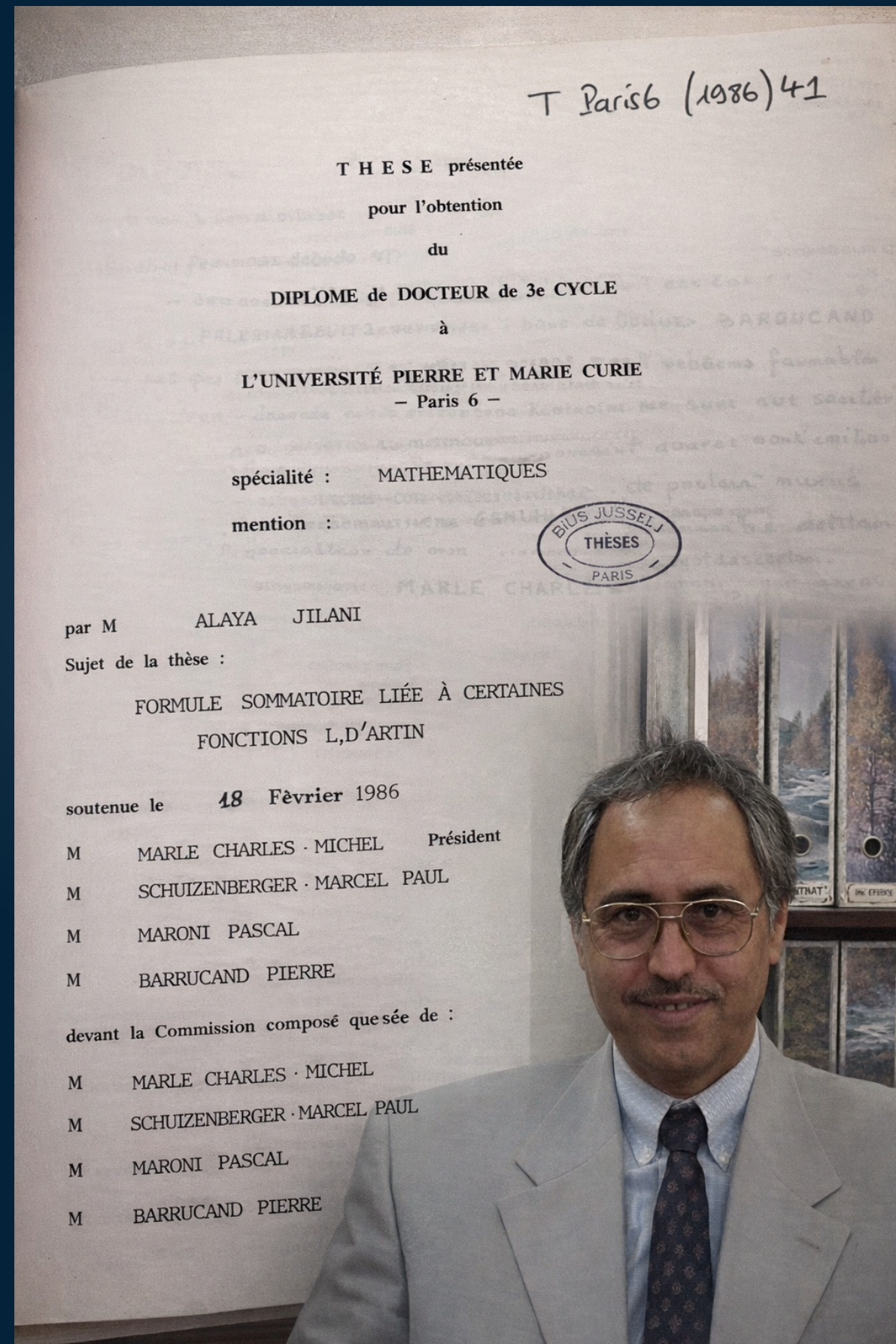
Mokhtar Z. Alaya



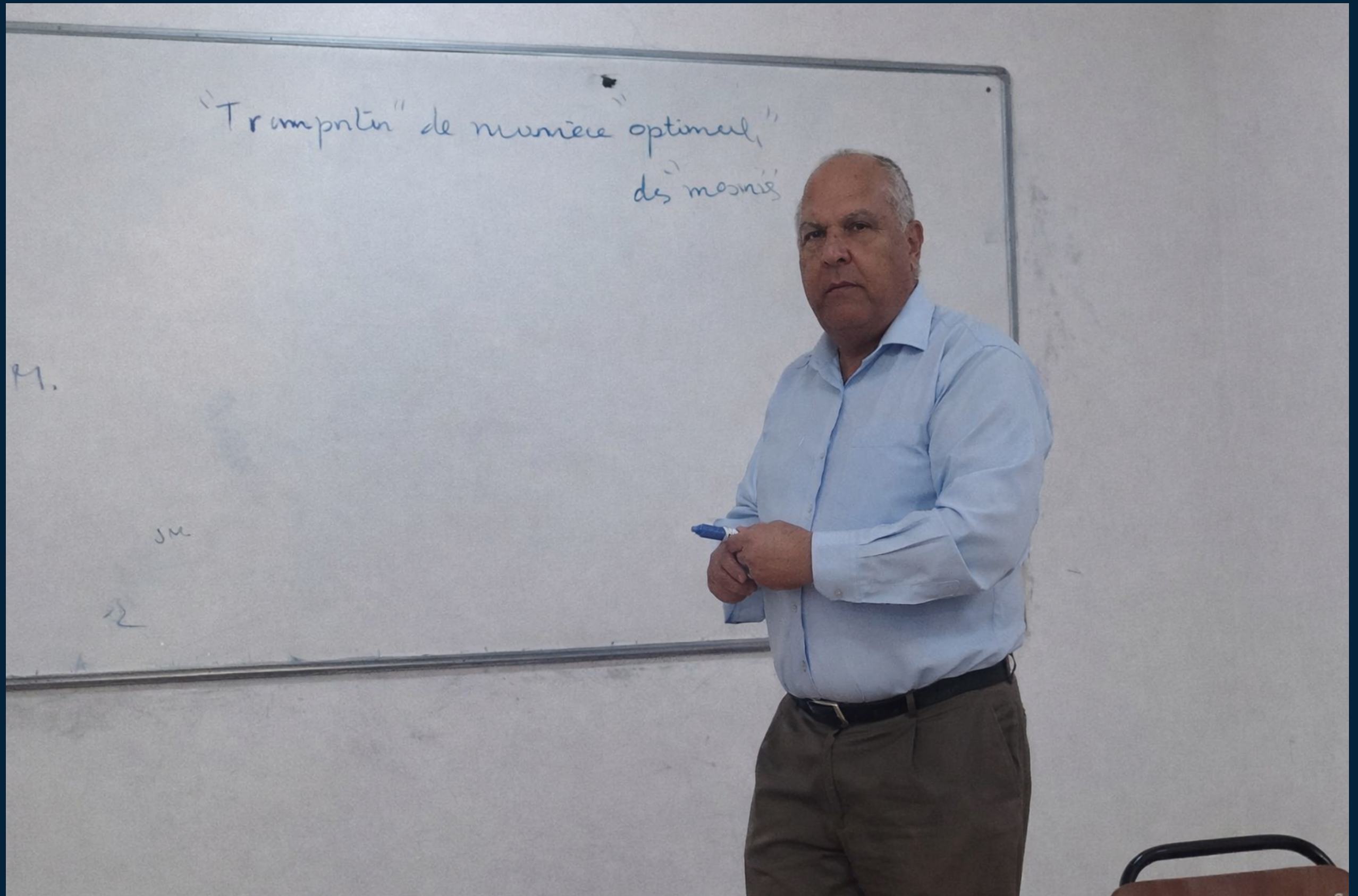
Acknowledgements



Prof. Jilani Alaya



Prof. Moncef Lahzami



Prof. Kamel Khirouni

Inorganic Chemistry Communications 179 (2025) 114753

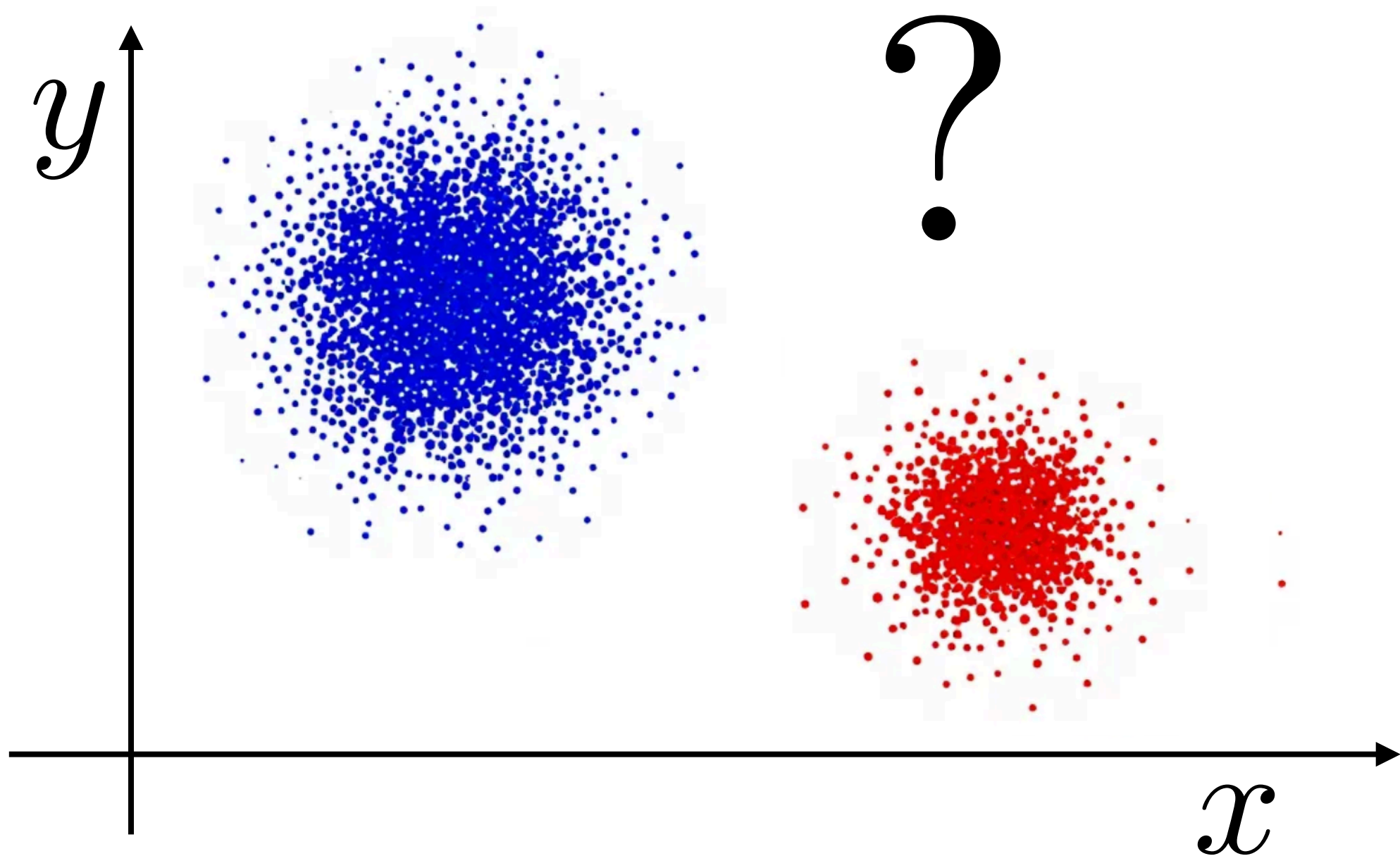


Kamel Khirouni is a full professor since 2005. He teaches various physics disciplines at the Faculty of Sciences of Gabes (University of Gabes, Tunisia). His research fields include sample preparation via solid-state reaction, plasma-enhanced chemical vapor deposition, and spray pyrolysis deposition. He conducts optical and electrical characterizations on both bulk samples and thin-film structures.

I. Motivations

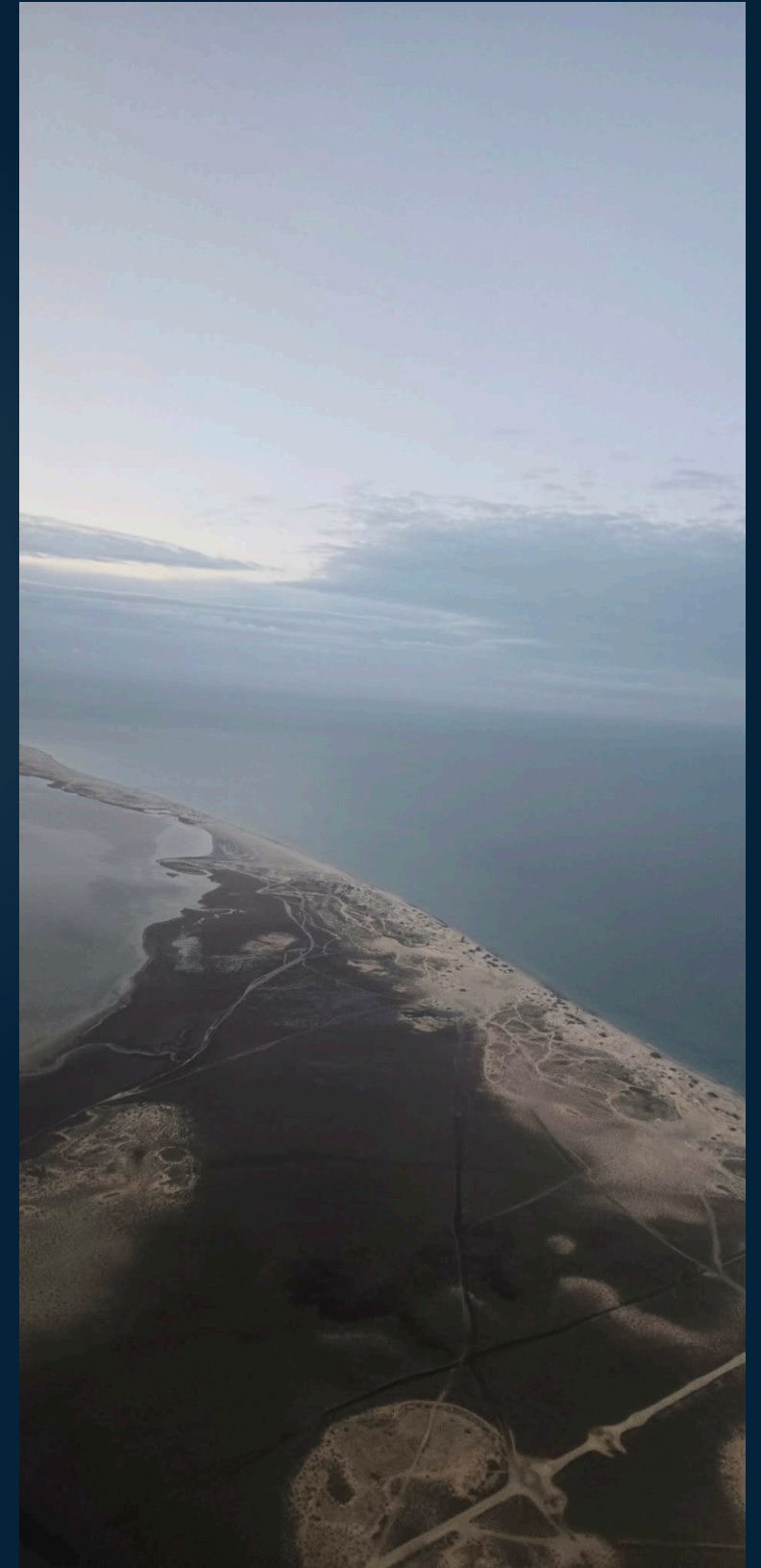
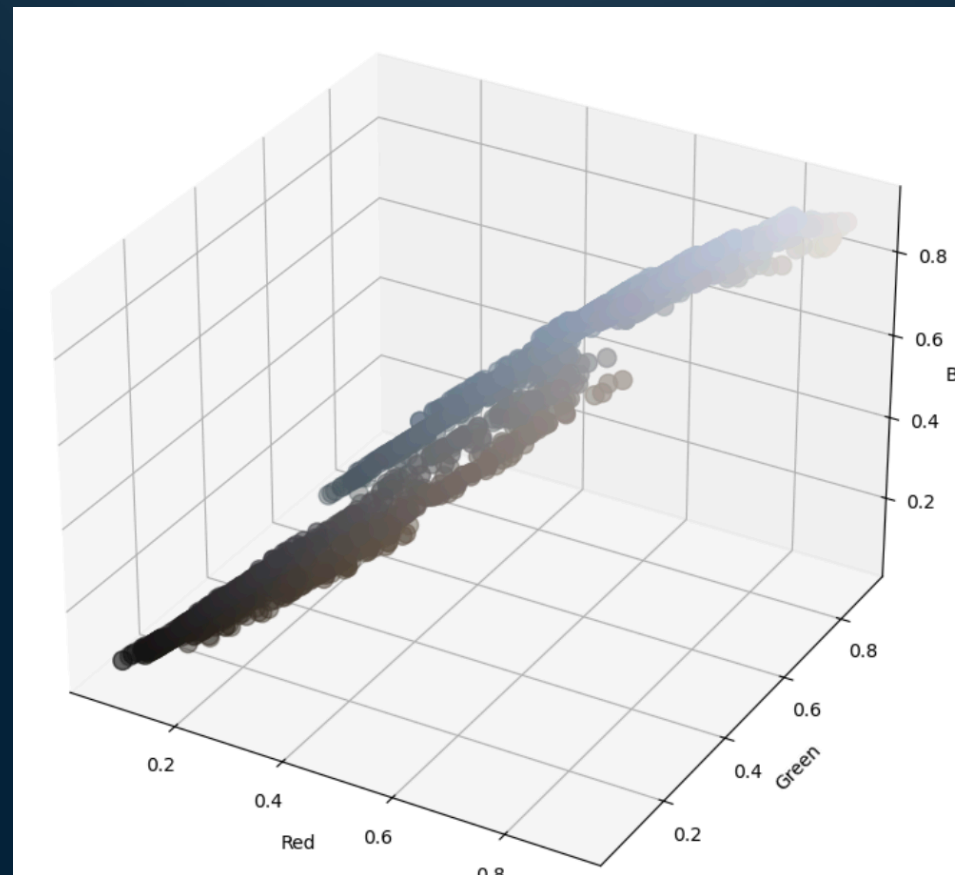
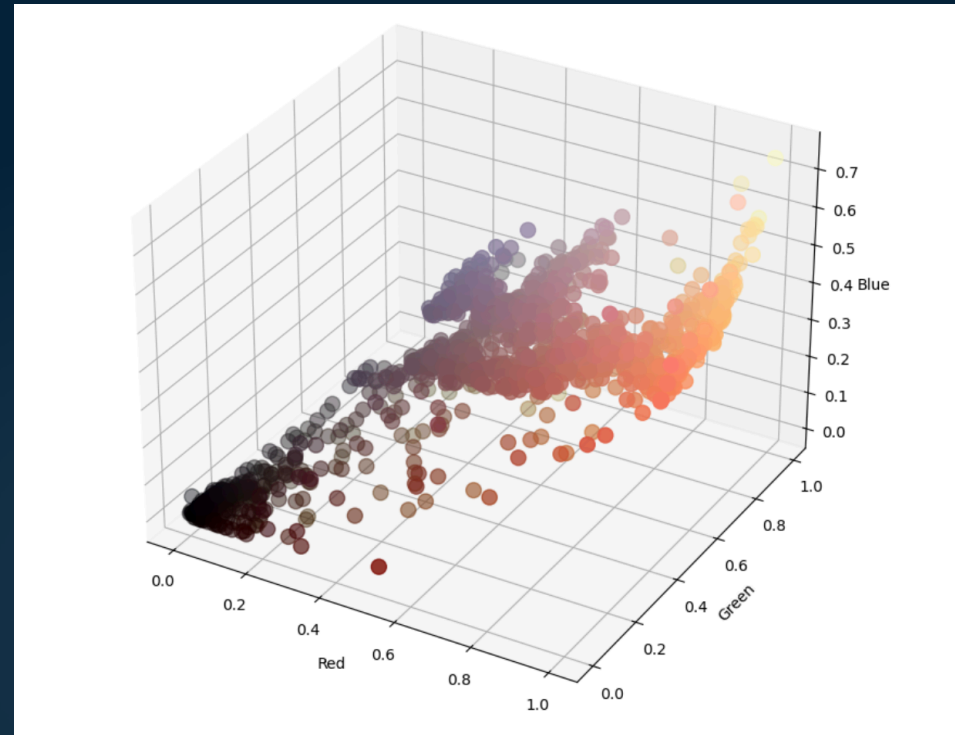
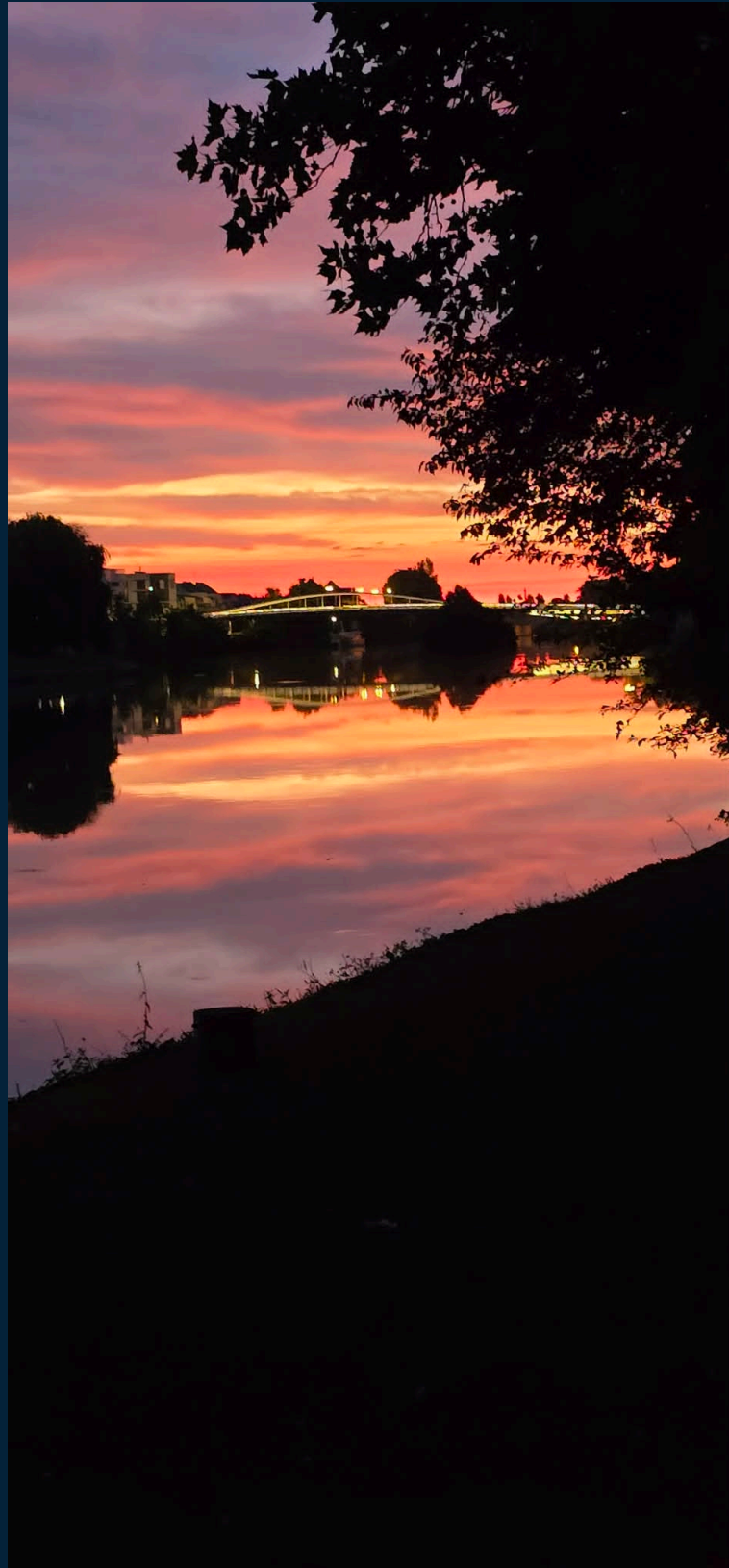
... in Cloud of points

- How to compare two sets of clouds points?



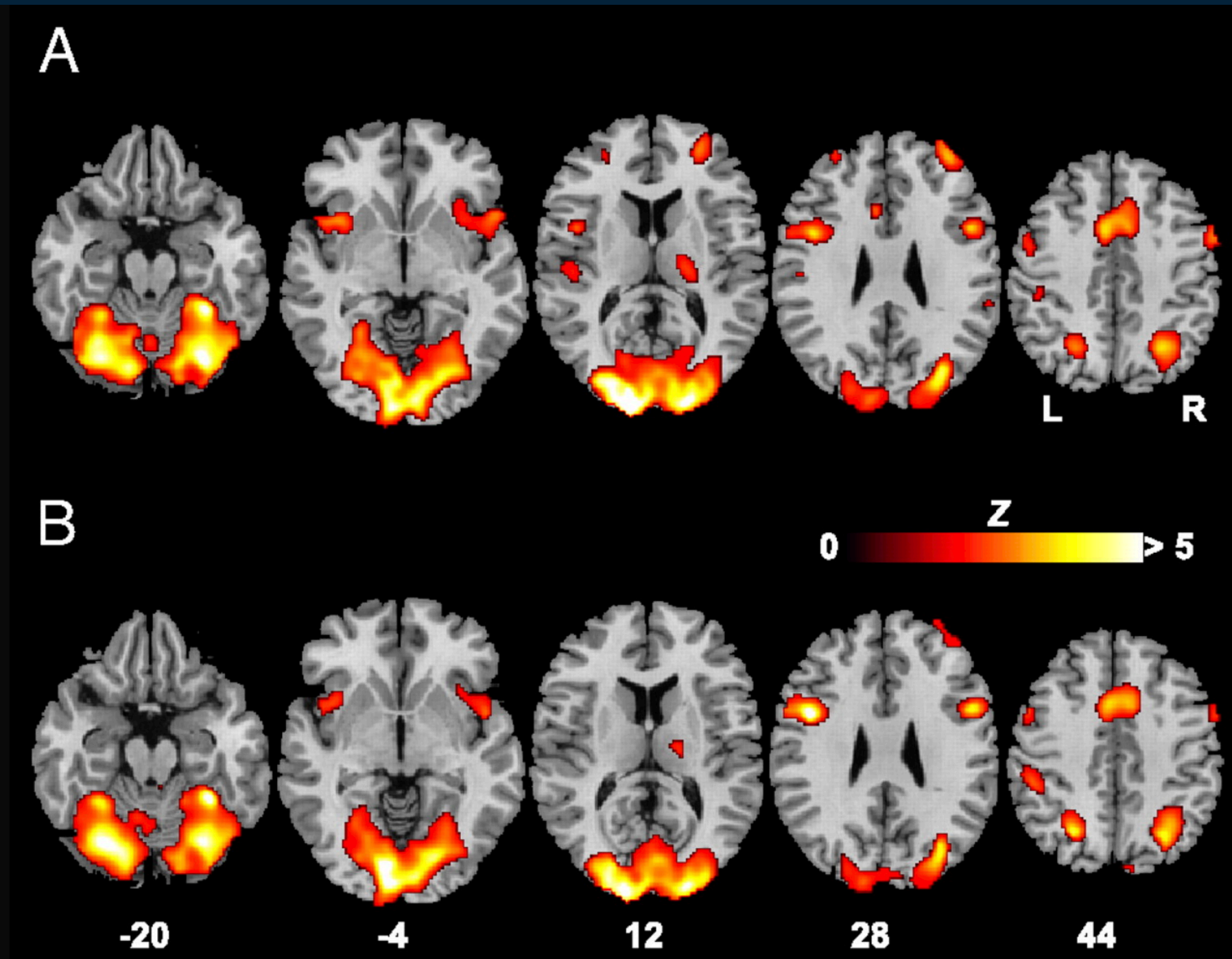
... in Image Processing

- How to measure the similarity between two images?



... in Neuroscience

- How to compare two brain activation maps?



... in Classification: CheXpert data



No Finding



Cardiomegaly



Lung Opacity



Pneumonia



Pneumothorax



Pleural Effusion



Lung Lesion



Edema



Consolidation



Atelectasis



Pleural Other



Fracture

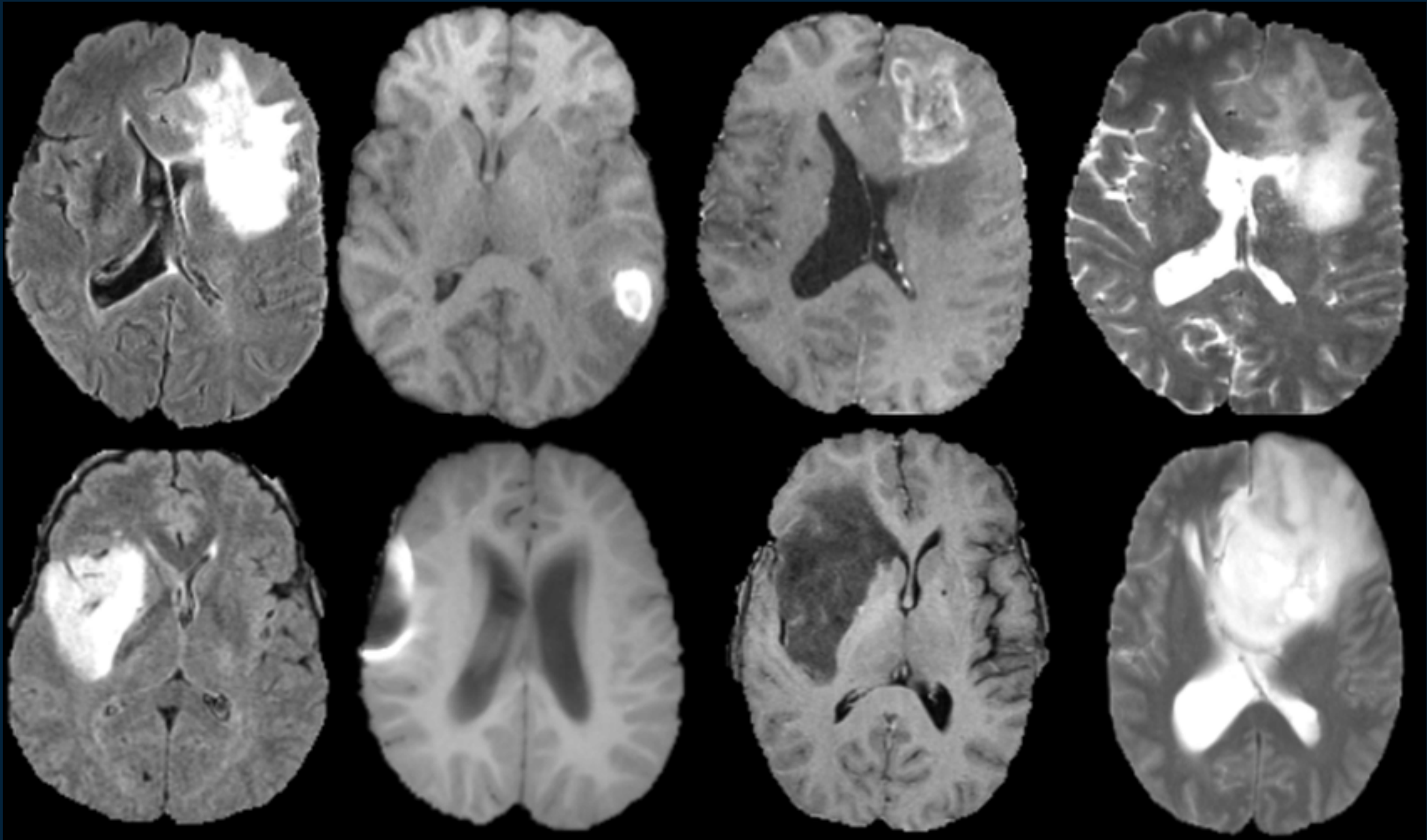


Support Devices



Enlarged
Cardiomeastinum

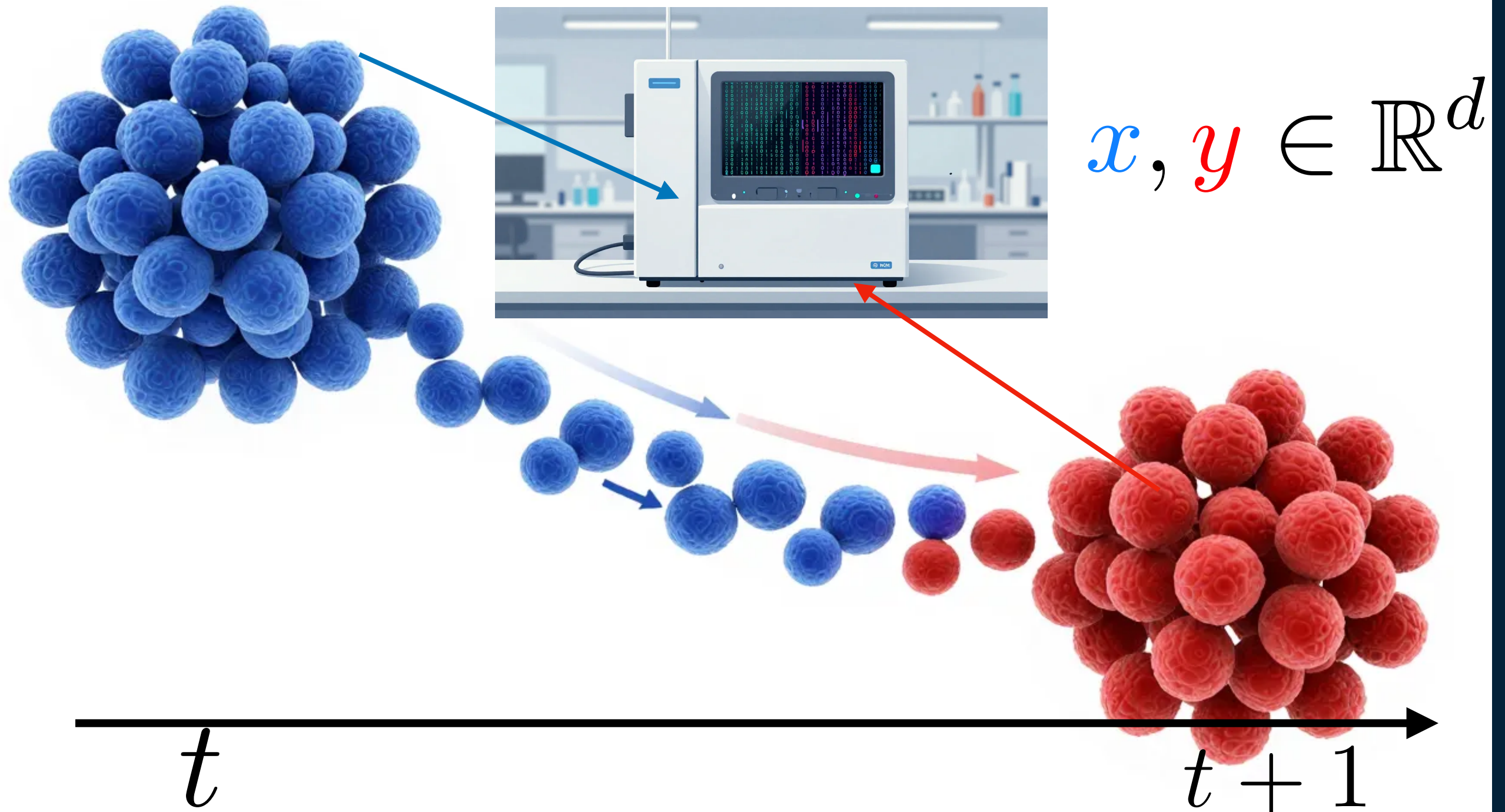
... in Anomaly detection: BraTS data



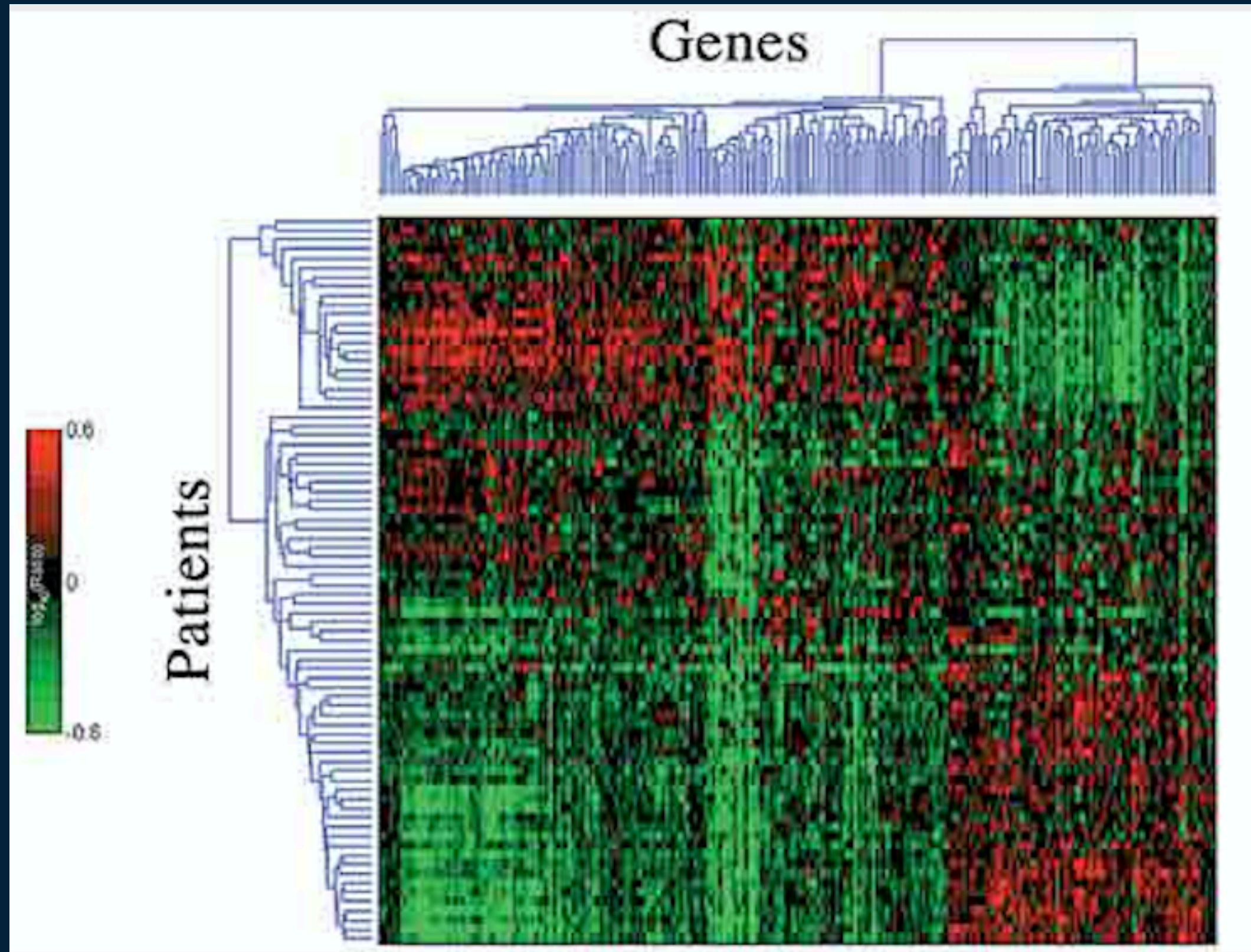
Source: <https://www.med.upenn.edu/cbica/brats2020/registration.html>

... in Genomics (I)

- Understanding dynamics at individual cell level.



... in Genomics (II)



Source: Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002

2. Machine Learning: Predictive Modeling

Training Set

 $\mathcal{X}$

Feature Space

 $\mathcal{Y}$

Label Space

$$(\vec{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$$

$$\vec{x}_i = (x_{i1}, \dots, x_{id})^\top \in \mathbb{R}^d$$

Sampling Data

$$\mathcal{D}_n = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\}$$

Learning Task

- The learning task consists of assuming that the labels have been calculated using a function.

$$h^* : \mathcal{X} \rightarrow \mathcal{Y}$$

- Find a hypothesis that better approximates the target function.
Build a predictor model, classifier, or regression function.

$$\hat{h} \equiv h_{\hat{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$$

Learning Task: loss function

- A cost function, also known as a loss function, is a function

$$\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$$

used to quantify prediction quality

$$\ell(y, h(\vec{x}))$$

Real Risk

- Risk is defined as the expectation of a cost function (or expectation of prediction error), i.e.,

$$\mathcal{R}(h) = \mathbb{E}_{(\vec{x}, y) \sim \mathbb{P}} [\ell(h(\vec{x}), y)]$$

- Risk minimization is impossible in practice because the joint law $\mathbb{P}$ of observations is unknown.
- Idea: replace the theoretical probability measure $\mathbb{P}$ with the empirical probability measure $\mathbb{P}_n$:

$$d\mathbb{P}_n(x, y) = \frac{1}{n} \sum_{i=1}^n \delta_{(\vec{x}_i, y_i)}(\vec{x}, y)$$

Empirical Risk Minimisation

- Empirical risk assesses the cumulative effect of errors

$$\mathcal{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, h(\vec{x}_i))$$

- Predictor by empirical risk minimization:

$$h_{\hat{\theta}} = \arg \min_{h \in \mathcal{H}} \mathcal{R}_n(h)$$

3. Optimal Transport

OT is ...

A method for comparing probability distributions with the ability to incorporate spatial information.



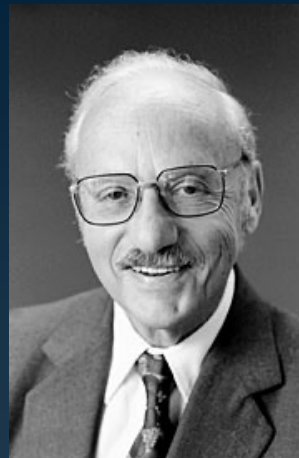
G. Monge
(1746 - 1818)



L. Kantorovich
(1912 - 1986)



T. Koopmans
(1910 - 1985)



G. Dantzig
(1914 - 2005)



Y. Brenier



F. Otto



R. McCann

Nobel Prize '75



C. Villani

Fields'10



A. Figalli

Fields'18

Origin: Monge Problem (1781)



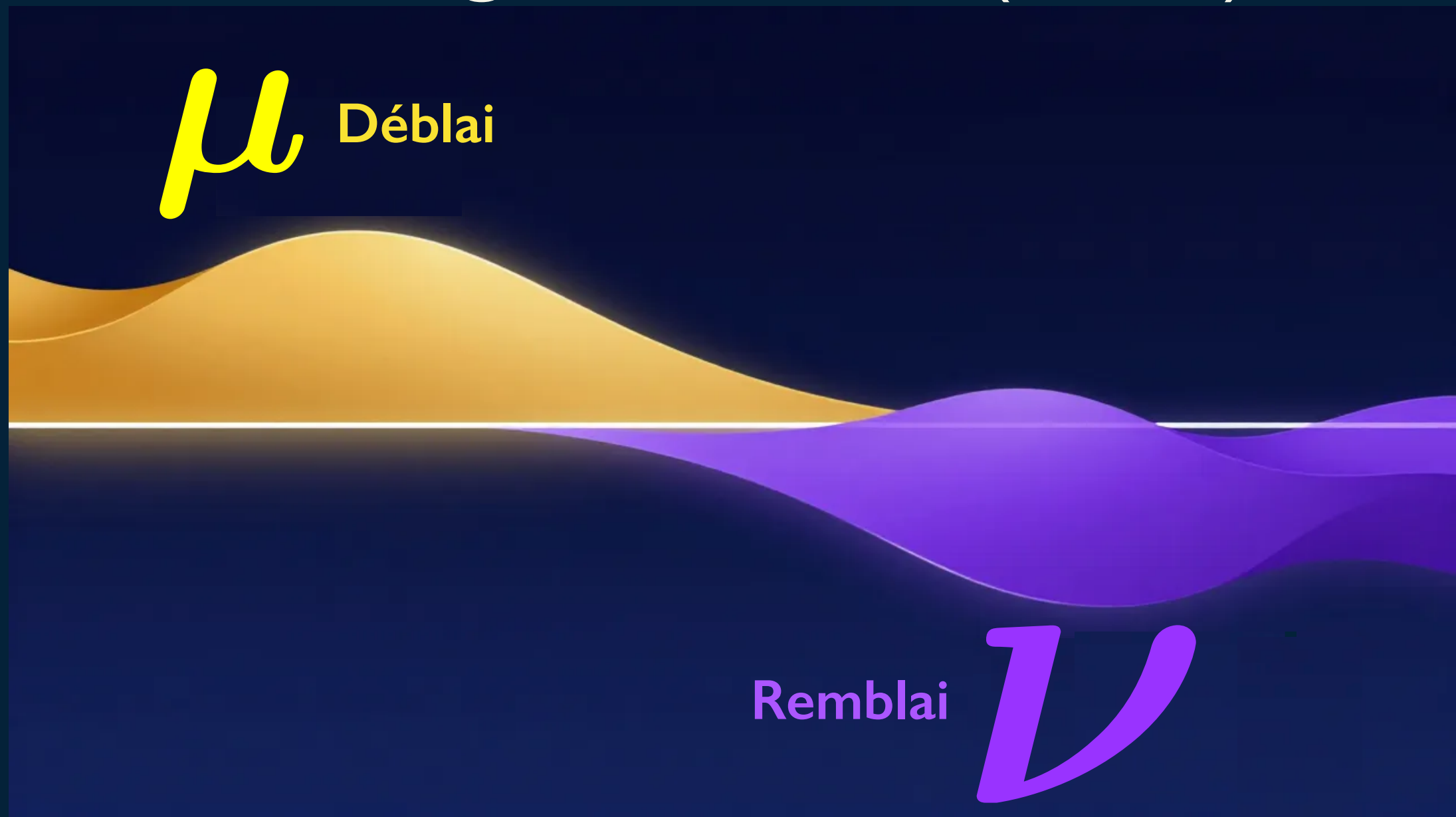
M É M O I R E *S U R L A* *T H É O R I E D E S D É B L A I S* *E T D E S R E M B L A I S.*

Par M. M O N G E.

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 fera un *minimum*.

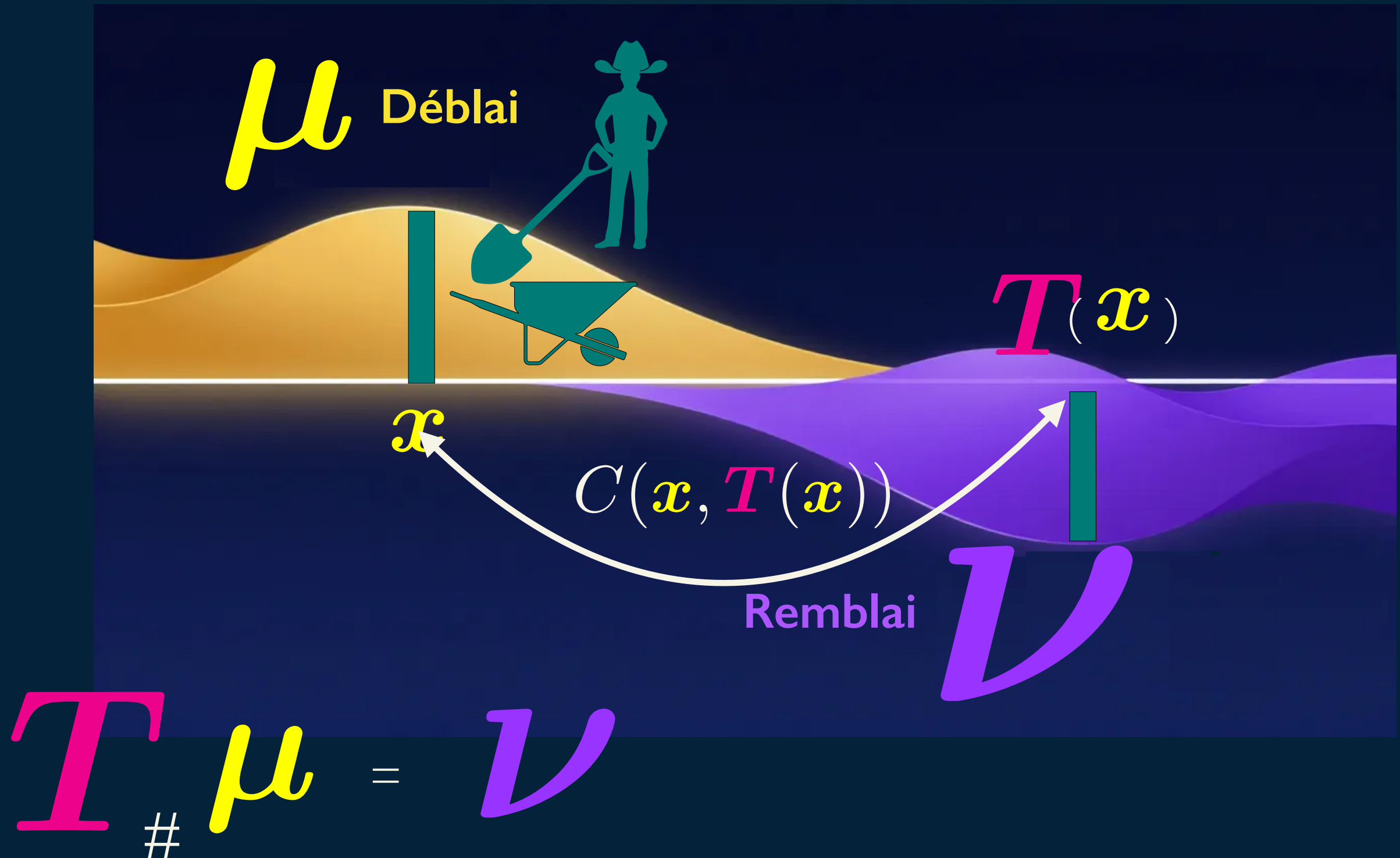
Monge Problem (1781)



- How to move dirt from one place (**déblai**) to another (**remblai**) while minimizing the effort?
- Find a mapping T between the two distributions of mass (**transport**).
- Optimize with respect to a displacement cost (**optimal**).

Monge Problem (1781)

- The mapping T must **push-forward** the “**déblai**” measure towards the “**remblai**”.



Monge Problem (1781)

- Monge formulation aim at finding a mapping T such that:

$$\inf_{T_{\#}\mu=\nu} \int C(x, T(x)) \mu(x) dx$$

- Mapping T does not exist in the general case.
- Brenier, 1991 proved existence and unicity of the Monge map for Euclidean cost and distributions with densities.

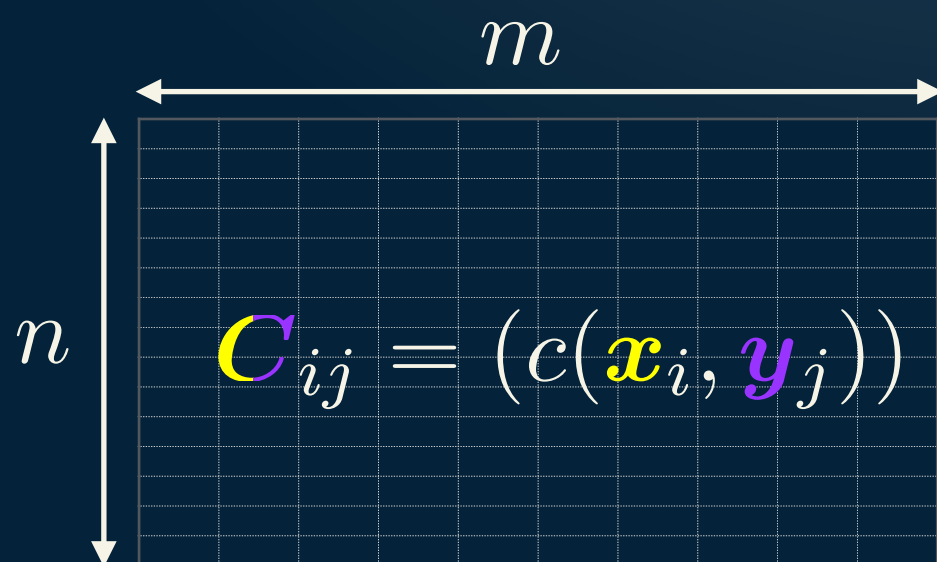
Discrete OT Framework

$$\nu = \sum_{j=1}^m \nu_j \delta_{y_j}$$

$$\mu = \sum_{i=1}^n \mu_i \delta_{x_i}$$

$$C_{ij} = (c(x_i, y_j))$$

C
Cost Matrix



$$\min_{T_{\#} \mu = \nu} \sum_i C(x_i, T(x_i)) \mu_i$$

$$\forall j \in \{1, \dots, m\}, \nu_j = \sum_{i: T(x_i) = y_j} \mu_i$$

Discrete OT Framework: Monge's Formula

$$\min_{T_{\# \mu = \nu}} \sum_i C(x_i, T(x_i)) \mu_i$$

$$\forall j \in \{1, \dots, m\}, \nu_j = \sum_{i: T(x_i) = y_j} \mu_i$$

Strict: Deterministic Assignments

Discrete OT Framework: Monge's Formula

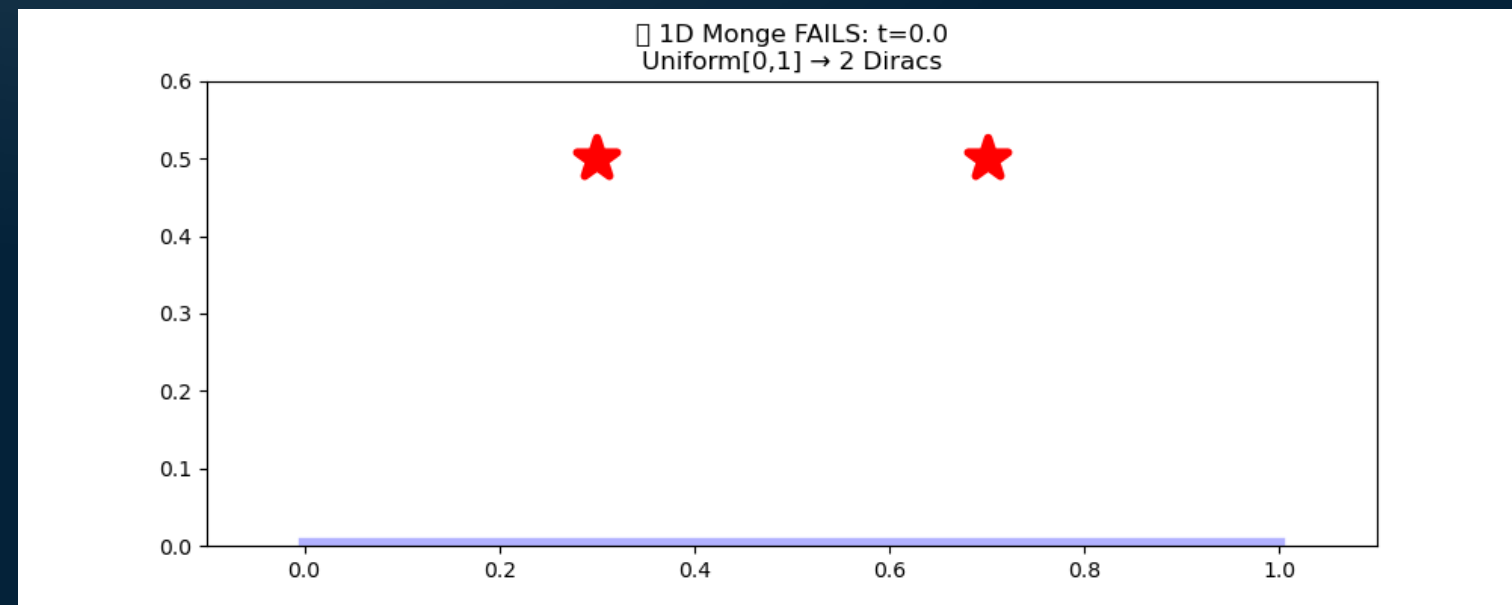
non-convex

combinatorial

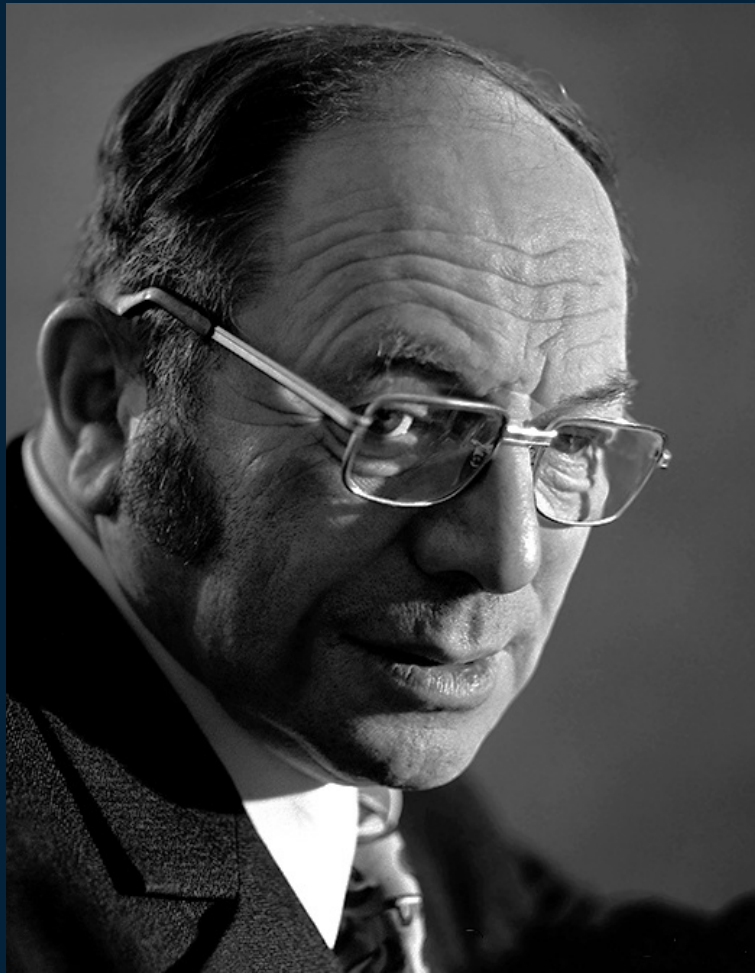
non-existent

Uniform weights

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

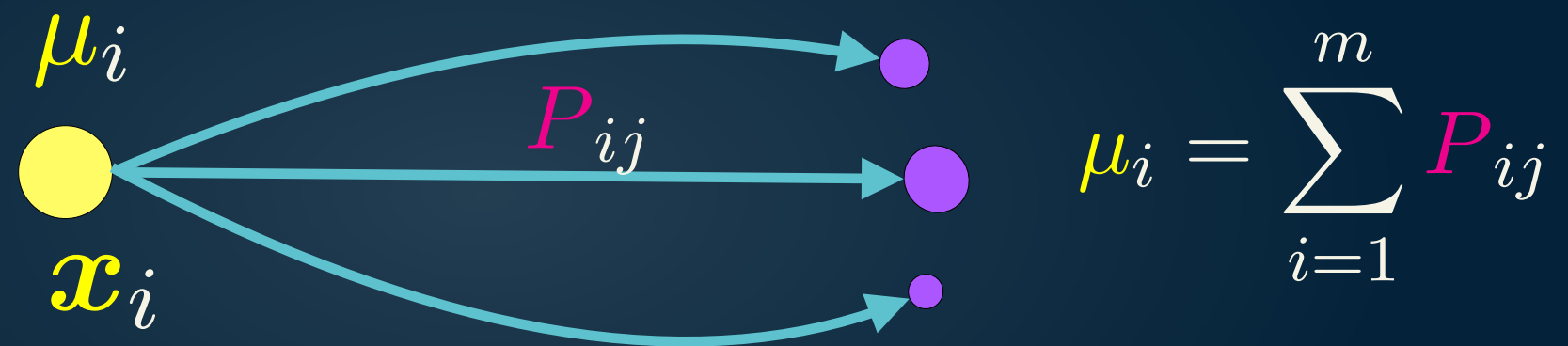


Discrete OT Framework: Kantorovich's Formula



Leonid Kantorovich
(1912-1986)

- Focus on where the mass goes, allow splitting.
- Applications mainly for resource allocation problems.



Relaxed: Fractional Assignments

Probabilistic couplings set (Transport Polytope)

$$\Pi(\mu, \nu) = \{P \in \mathbb{R}_+^{n \times m}, P \mathbf{1}_m = \mu, P^\top \mathbf{1}_n = \nu\}$$

Mass conservation constraints

Discrete OT Framework: Kantorovich's Formula

- Computing OT between μ and ν amounts to solving a linear problem:

Kantorovich 1942

$$\mathcal{S}(\mu, \nu) = \min_{P \in \Pi(\mu, \nu)} \left\{ \langle C, P \rangle = \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij} \right\}$$



Distance

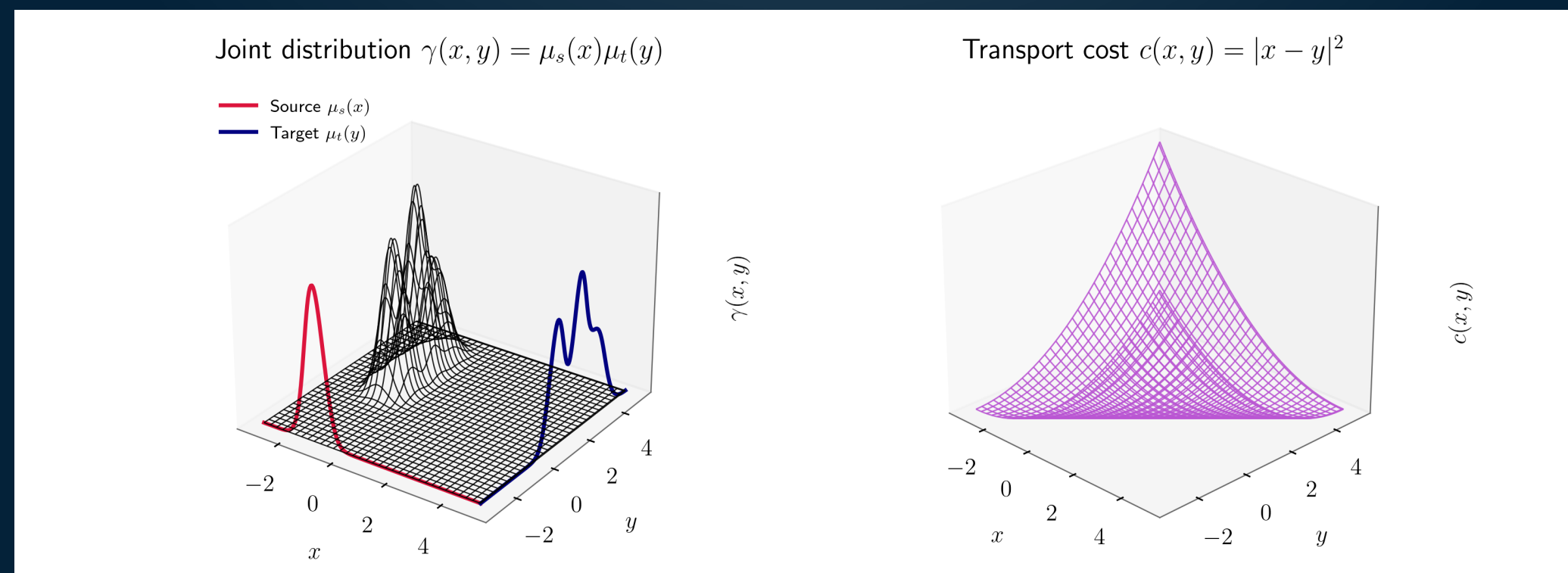
Monge-Kantorovich /
Wasserstein Distance

Continuous OT Framework: Kantorovich's Formula

$$\min_{\gamma \in \Pi_{\text{con}}(\mu, \nu)} \int_{\mathbb{R}^n \times \mathbb{R}^m} C(x, y) \gamma(x, y) dx dy$$

Probabilistic couplings set (Transport Polytope)

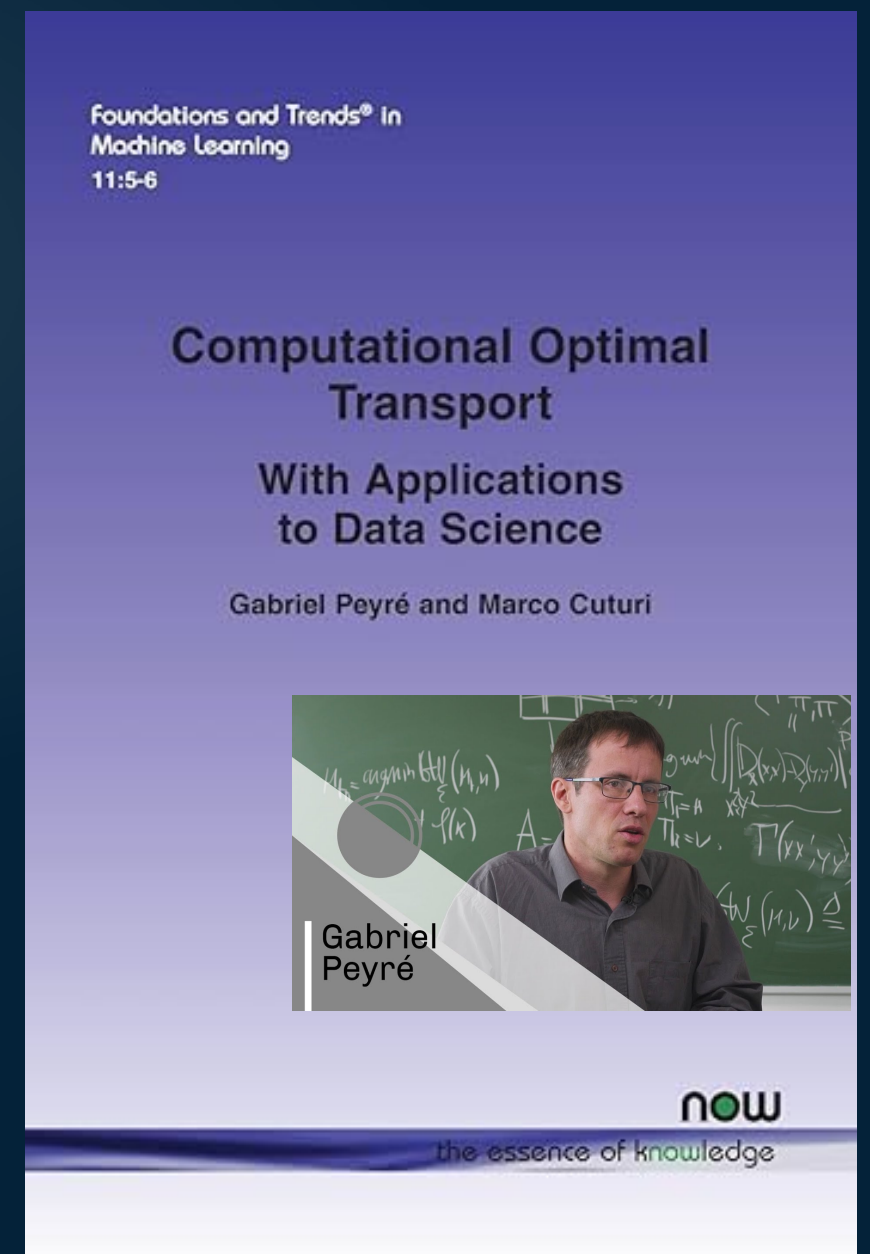
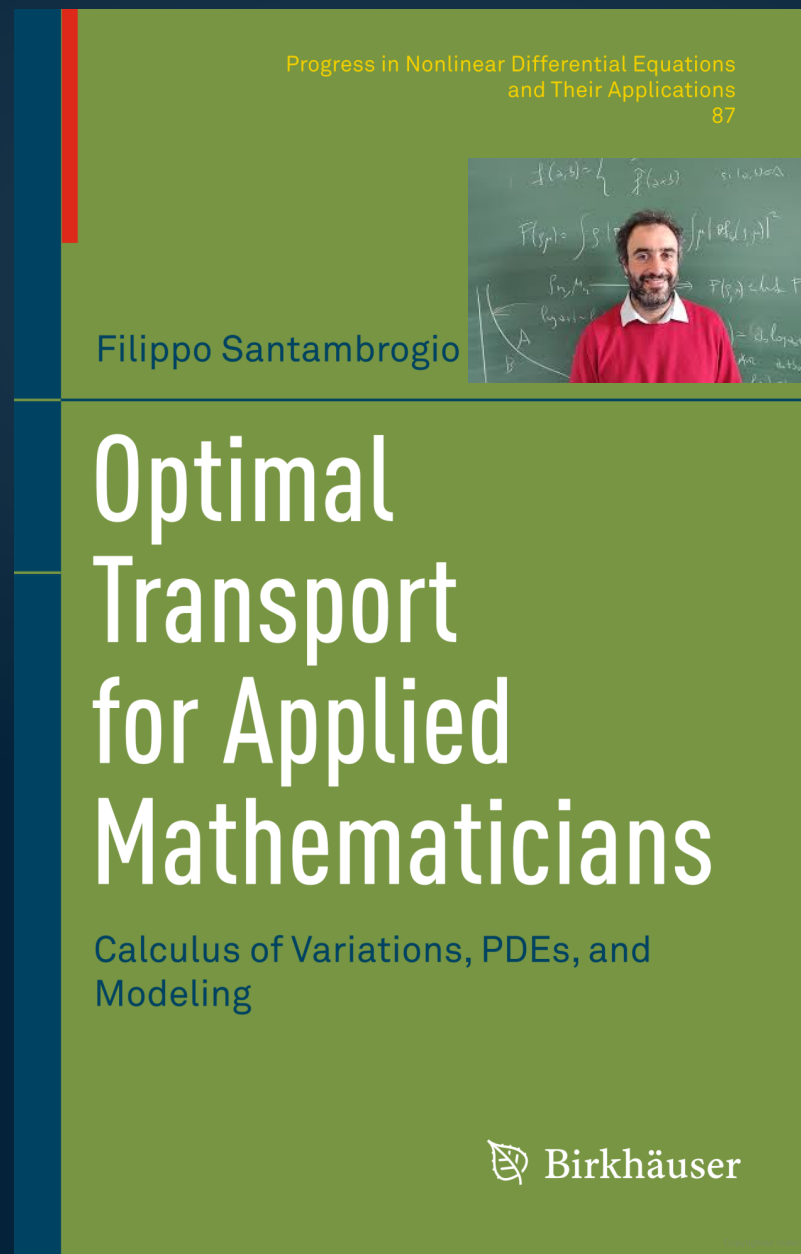
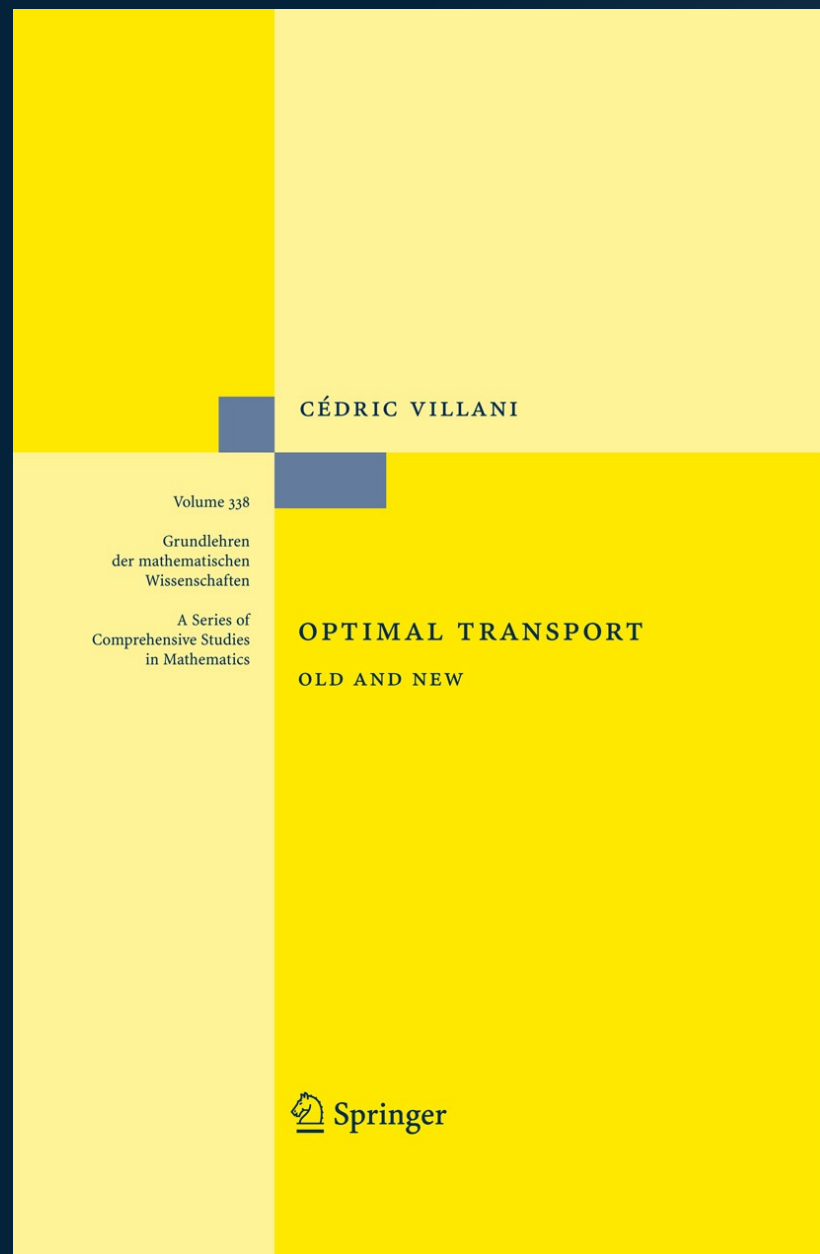
$$\Pi_{\text{con}}(\mu, \nu) = \left\{ \gamma \geq 0, \int_{\mathbb{R}^m} \gamma(x, y) dy = \mu, \int_{\mathbb{R}^n} \gamma(x, y) dx = \nu \right\}$$



Continuous OT Framework: Wasserstein Distance

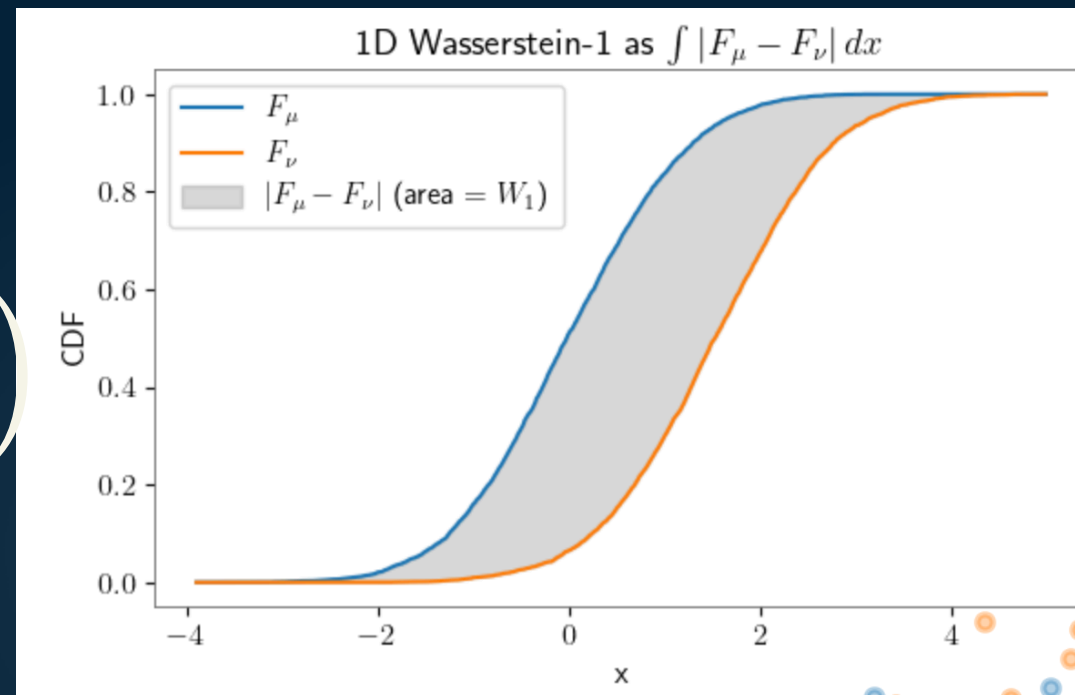
Wasserstein Distance

$$\mathcal{W}_p^p(\mu, \nu) = \min_{\gamma \in \Pi_{\text{con}}(\mu, \nu)} \int_{\mathbb{R}^n \times \mathbb{R}^m} D^p(x, y) \gamma(x, y) dx dy = \mathbb{E}_{(x, y) \sim \gamma} [D^p(x, y)]$$

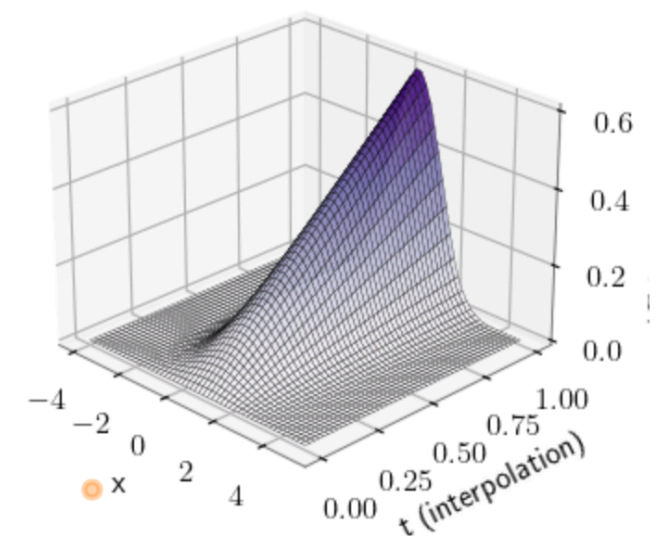


Continuous OT Framework: Wasserstein Distance,

$$\mathcal{W}_1(\mu, \nu)$$



Surface between distributions via CDFs



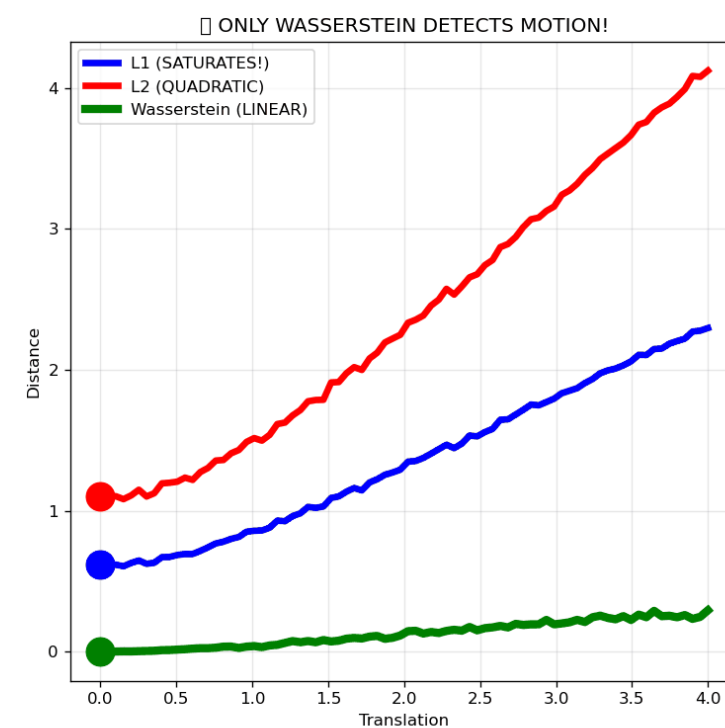
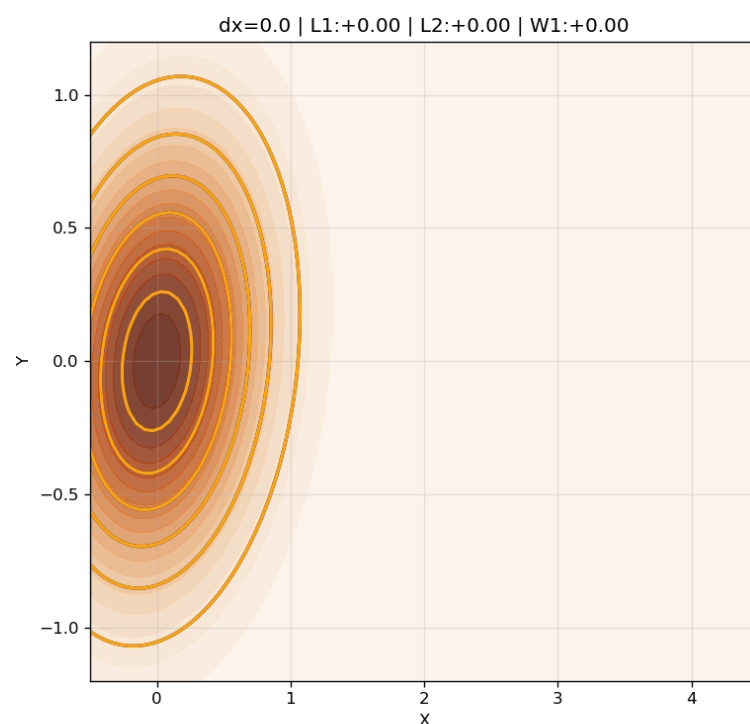
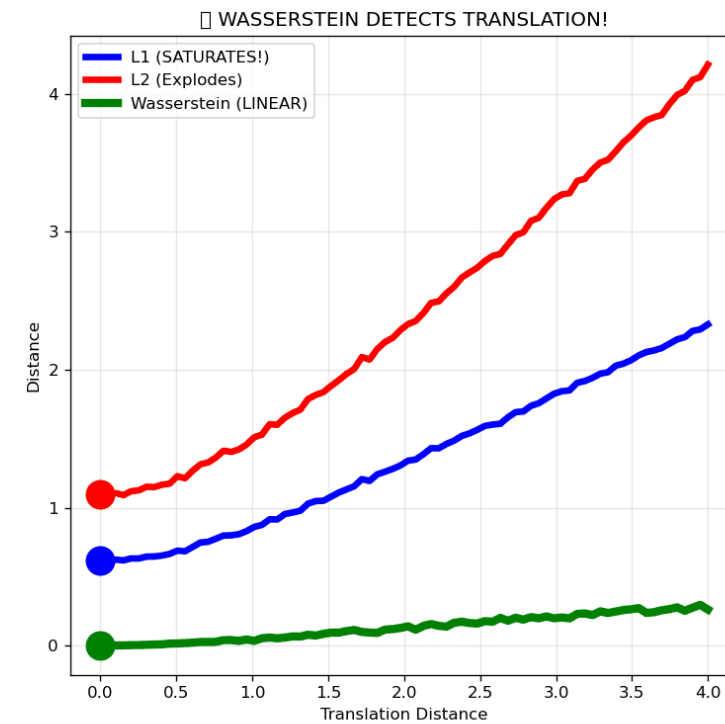
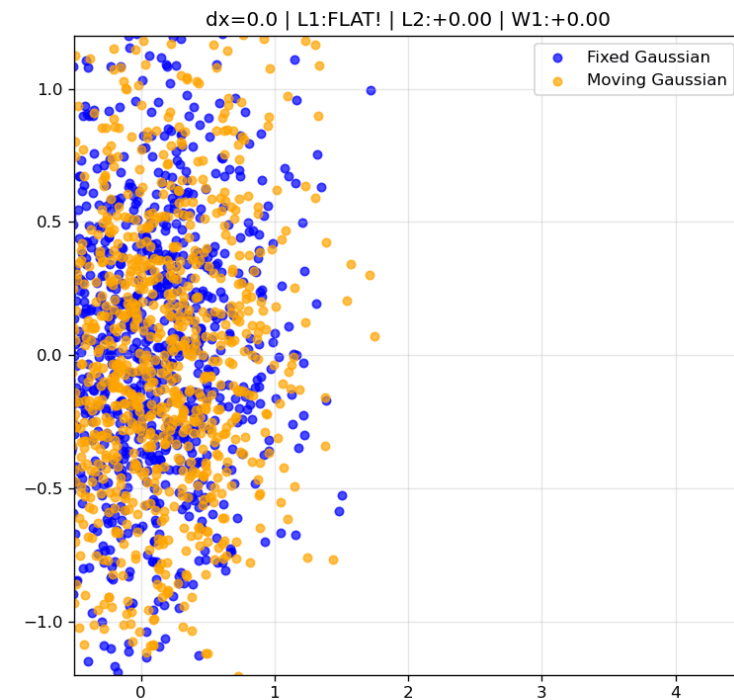
Bures Distance

1

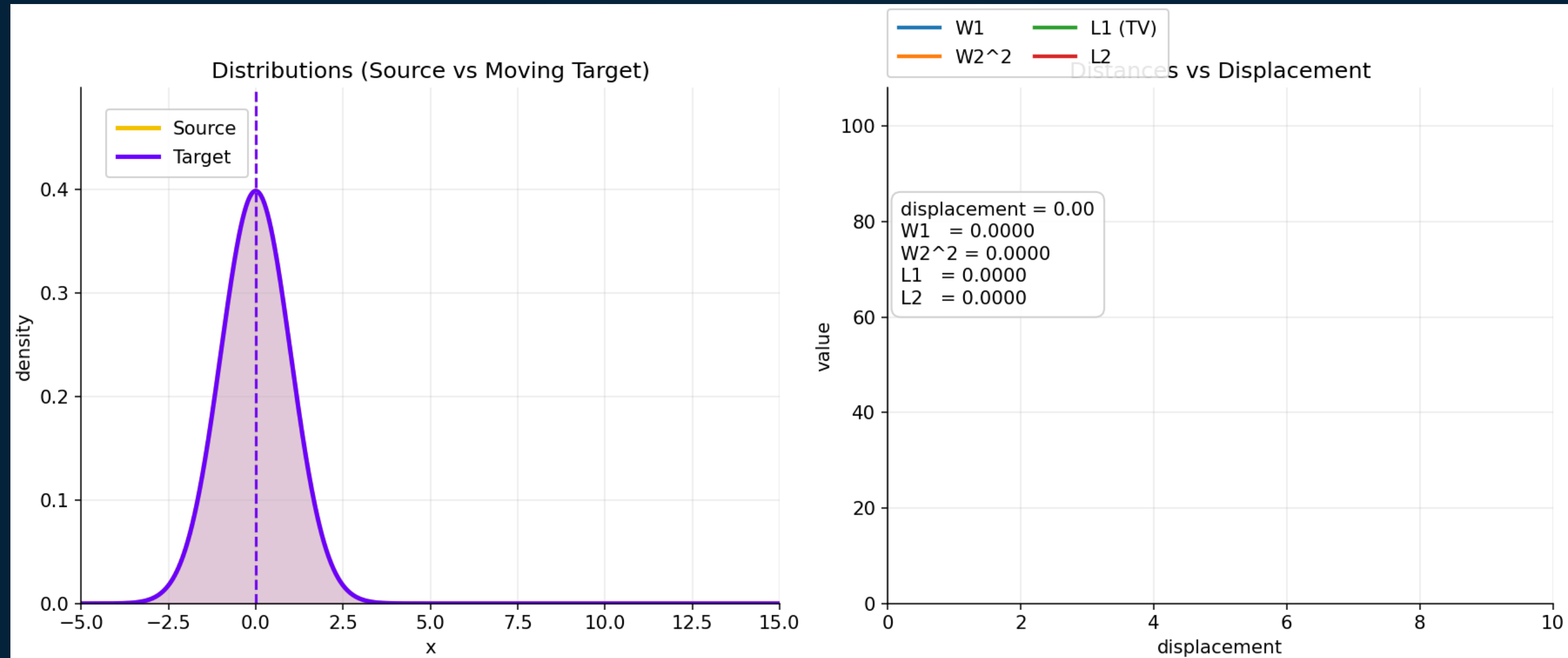
$$\mathcal{W}_2^2(\mu, \nu)$$

$$= \|m_0 - m_1\|^2 + \text{Tr} \left(\Sigma_0 + \Sigma_1 - 2(\Sigma_1^{1/2} \Sigma_0 \Sigma_1^{1/2})^{1/2} \right)$$

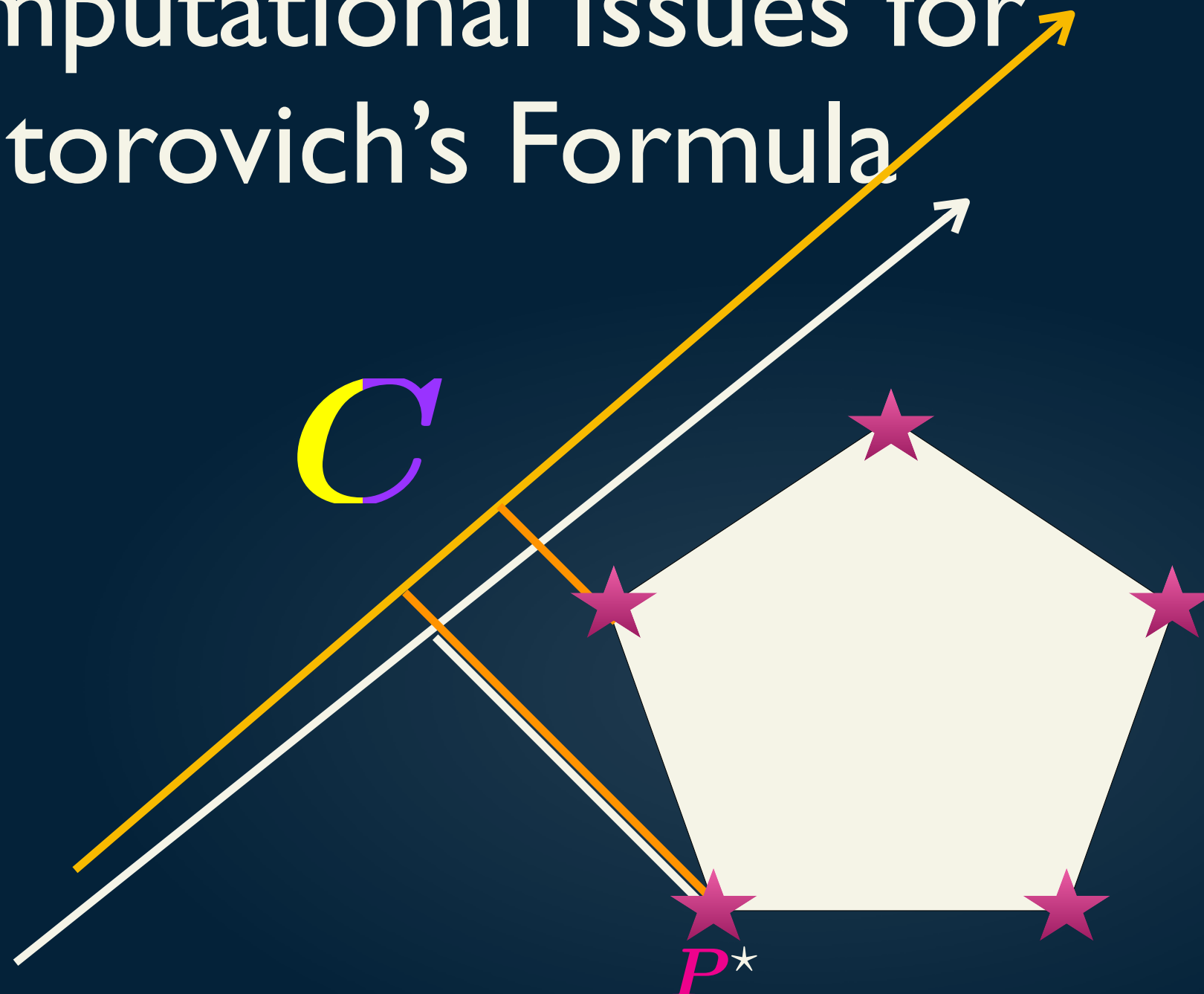
Benefit of Wasserstein Distance



Benefit of Wasserstein Distance



Computational Issues for Kantorovich's Formula



$\Pi(\mu, \nu)$ is the Birkhoff polytope

- No unique solution in some cases, numerical instabilities.
- Linear programming problem that requires generally $\mathcal{O}(n^3 \log(n)^2)$ arithmetic operations.

Regularized Discrete OT Framework: Sinkhorn Divergence

- Entropic regularization of OT distances relies on the addition of a penalty term as follows:

Regularisation parameter

$$\mathcal{S}_\eta(\mu, \nu) = \min_{P \in \Pi(\mu, \nu)} \{ \langle C, P \rangle - \eta H(P) \}$$

Sinkhorn divergence

Negative entropy

$$H(P) = - \sum_{i,j} P_{ij} \log(P_{ij})$$

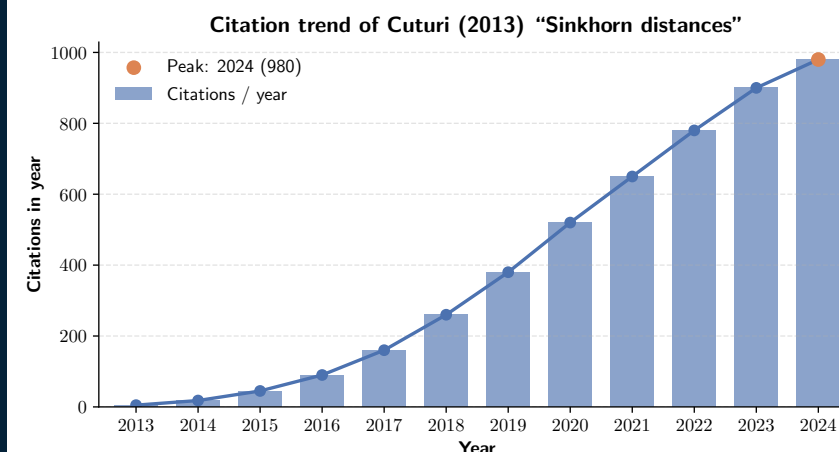


Marco Cuturi

Conference NeurIPS, 2013

**Sinkhorn Distances:
Lightspeed Computation of Optimal Transport**

Marco Cuturi
Graduate School of Informatics, Kyoto University
mcuturi@i.kyoto-u.ac.jp



Regularized Discrete OT Framework:

Dual of $\mathcal{S}_\eta(\boldsymbol{\mu}, \boldsymbol{\nu})$

Dual of Sinkhorn divergence

$$\mathcal{S}_\eta^d(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{v} \in \mathbb{R}^m} \left\{ \Psi(\boldsymbol{u}, \boldsymbol{v}) := \mathbf{1}_n^\top B(\boldsymbol{u}, \boldsymbol{v}) \mathbf{1}_m - \langle \boldsymbol{u}, \boldsymbol{\mu} \rangle - \langle \boldsymbol{v}, \boldsymbol{\nu} \rangle \right\}$$

where

$$B(\boldsymbol{u}, \boldsymbol{v}) := \text{diag}(e^{\boldsymbol{u}}) \boldsymbol{K} \text{diag}(e^{\boldsymbol{v}}) \quad \boldsymbol{K} = e^{-\boldsymbol{C}/\eta}$$

Gibbs Kernel

- The primal optimal solution $\boldsymbol{P}^*$ takes the form:

$$\boldsymbol{P}^* = \text{diag}(e^{\boldsymbol{u}^*}) \boldsymbol{K} \text{diag}(e^{\boldsymbol{v}^*})$$

Optimal Transportation Plan

$$\text{with } (\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{v} \in \mathbb{R}^m}{\operatorname{argmin}} \left\{ \Psi(\boldsymbol{u}, \boldsymbol{v}) \right\}$$

Dual Optimal Variables

Regularized Discrete OT Framework: Sinkhorn Algorithm

- P^* can be solved efficiently by Sinkhorn iterations (near- $\mathcal{O}(n^2)$ complexity (Altschuler et al., 2017))

SINKHORN(C, μ, ν) Matrix-Scaling Problem

1. $a^{(0)} \leftarrow \mathbf{1}_n/n, b^{(0)} \leftarrow \mathbf{1}_m/m;$

2. $K \leftarrow e^{-C/\eta};$

3. For $k = 1, 2, 3, \dots$

$a^{(k)} \leftarrow \mu \oslash K b^{(k-1)};$

$b^{(k)} \leftarrow \nu \oslash K^\top a^{(k-1)};$

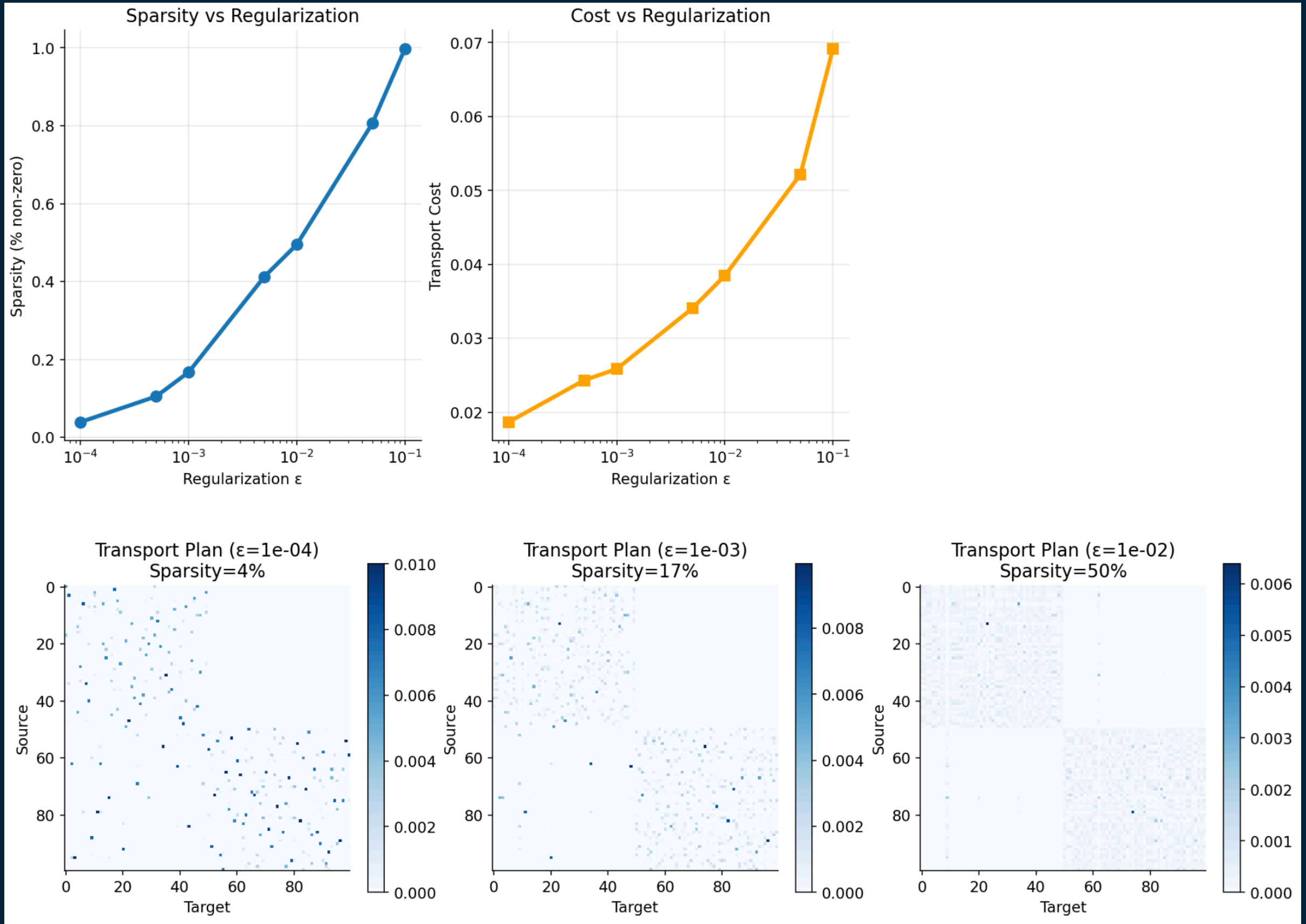
4. Return $\text{diag}(a^{(k)}) K \text{diag}(b^{(k)})$



(Flamary et al. 2017)

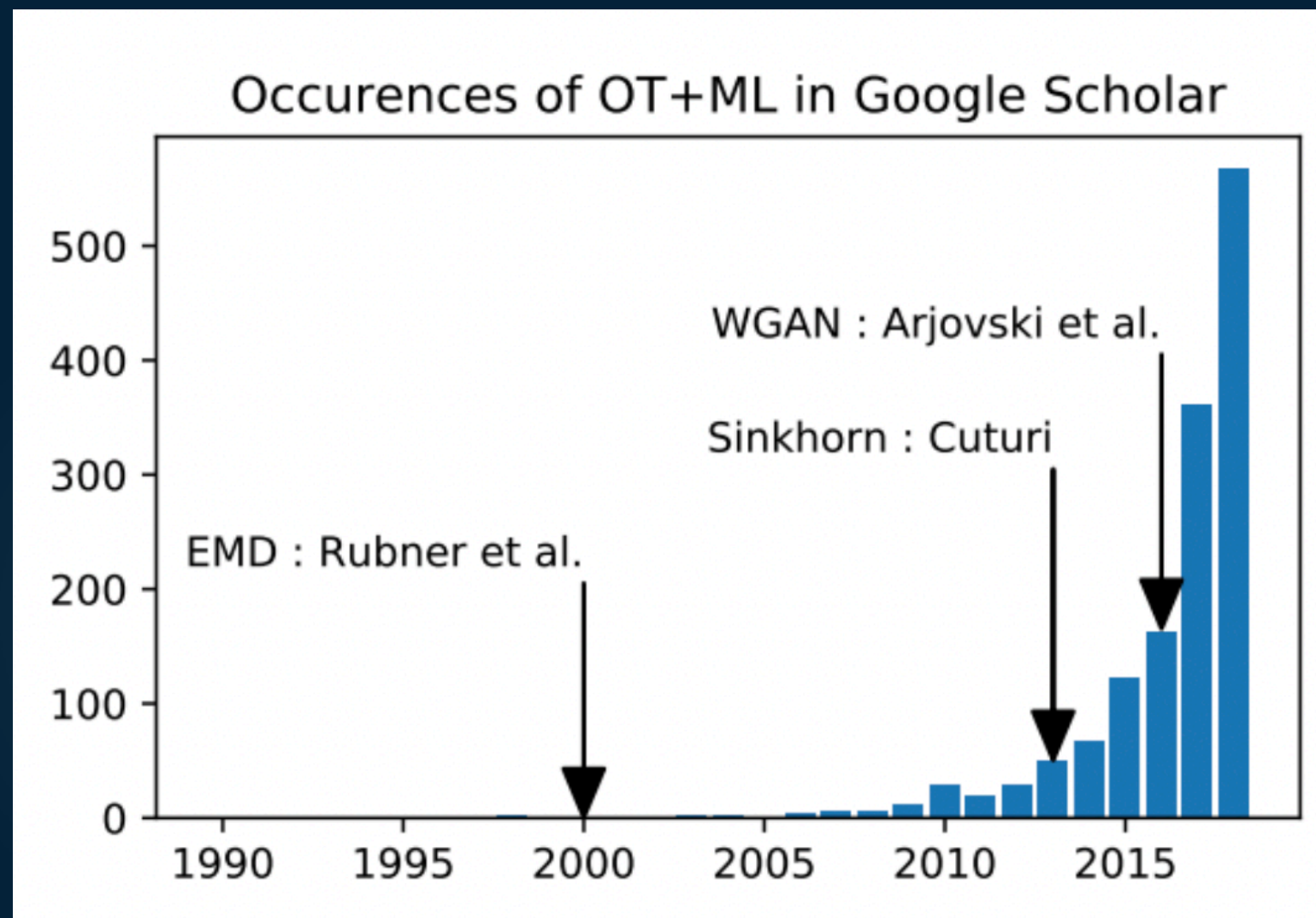
```
from ot import sinkhorn  
P_star = sinkhorn(mu, nu, C, eta)
```

Sparsity of Transport Plan



4. How can it be used in data science?

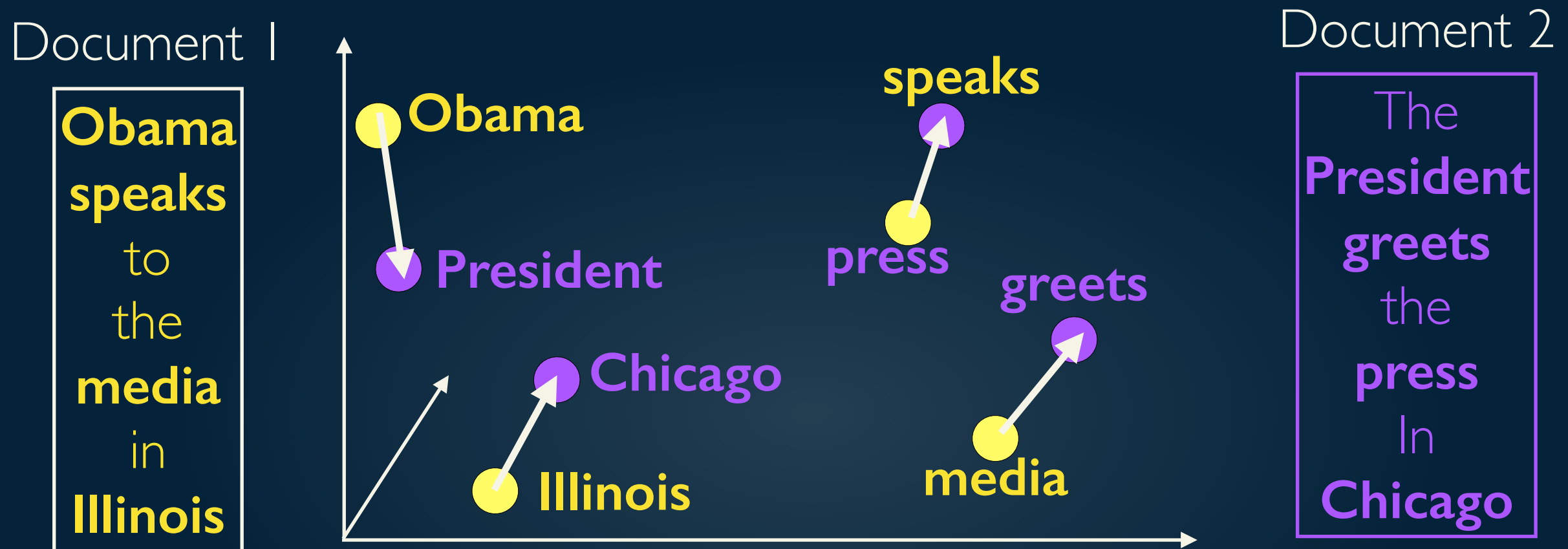
History of OT for machine learning



[R. Flamary, 2019 (HDR)]

- Recently introduced to ML (well known in image processing since 2000).
- Computational OT allows numerous applications (regularization).
- Deep learning boost (numerical optimisation and GAN).

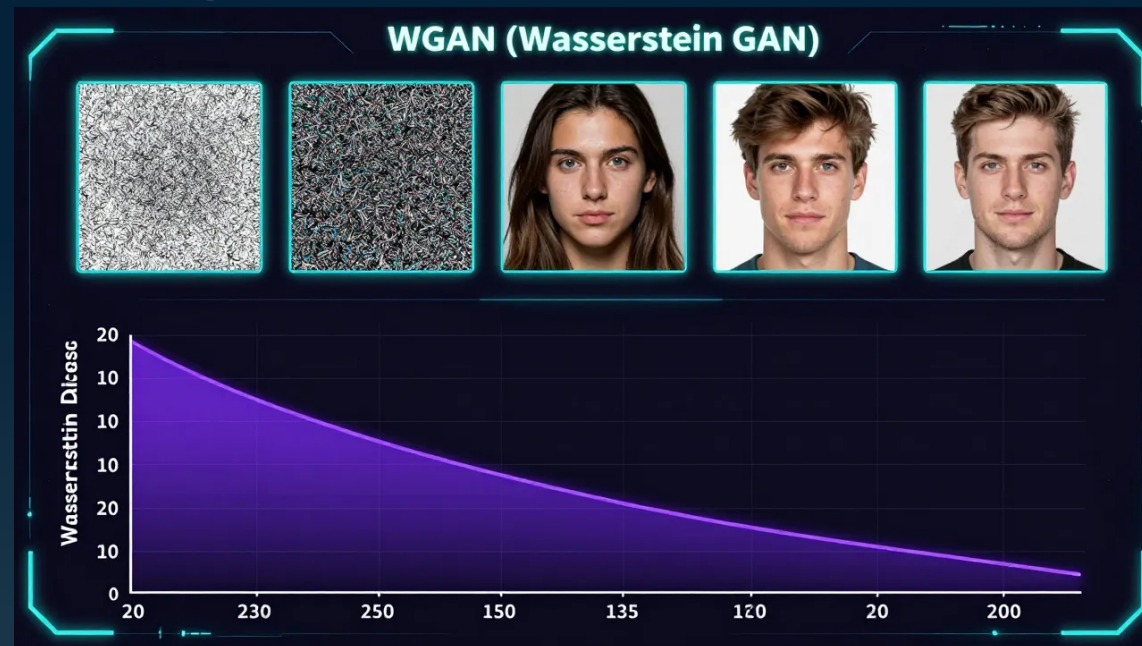
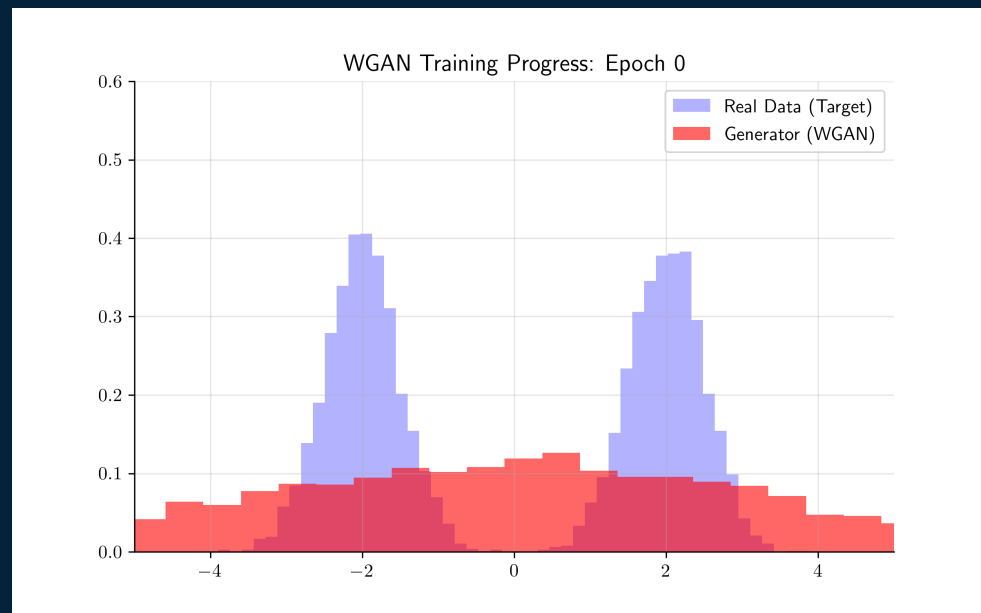
Matching words embeddings



Word Mover's Distance avec Word2vec embeddings
[Kusner et al, 2015 (ICML)]

- Words are embedded in a high-dimensional space with deep neural networks.
- Matching two documents in an OT problem, with the Euclidean distance in the embedded space.

Wasserstein loss for generative modelling



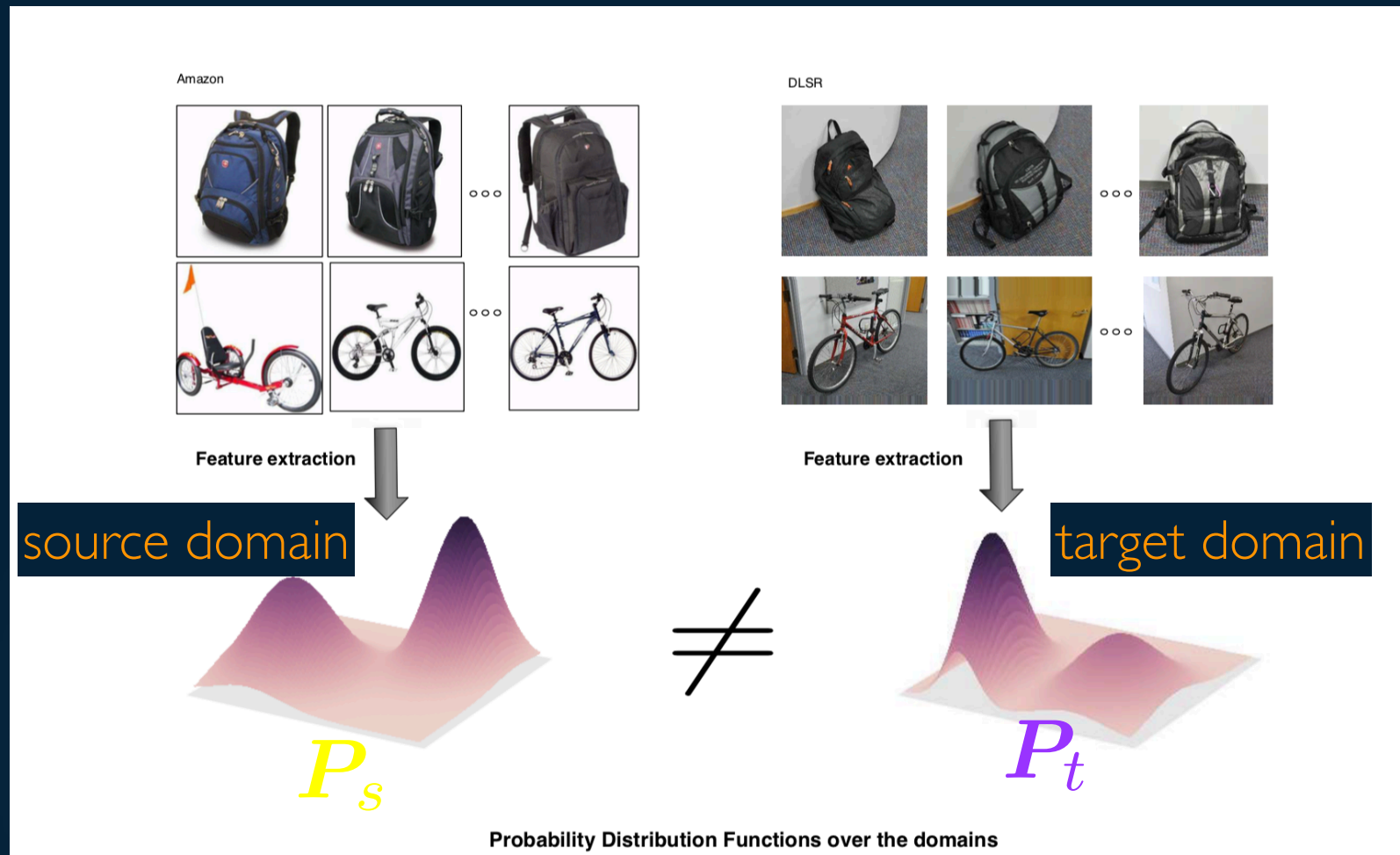
Generative modelling as a matching distribution problem

- Learn a model that maps random vector to target space.
- Distribution of the model is targeted to be similar to the learning samples.
- Similarity as Wasserstein sense [Arjovsky et al. 2017, Deshpande et al. 2018, Nguyen et al. 2020).

$$\min_{f_{\theta}} W_p^p \left(\left\{ f_{\theta}(z_i) \right\}_{i=1}^K, \left\{ x_j \right\}_{j=1}^K \right)$$

$\{z_i\}$ some random vectors, $\{x_j\}$ some samples from the target distribution.

Unsupervised Domain Adaptation



- Traditional machine learning hypothesis:
 - We have access to training data. Probability distribution of the training set and the testing are the same.
 - We want to learn a classifier that generalizes to new data.

Unsupervised Domain Adaptation



- Domain adaptation: classification problem with data coming from different sources (domains).
- Labels only available in the source domain, and classification is conducted in the target domain.
- Classifier trained on the source domain data performs badly in the target domain.

OTDA: Optimal Transport Domain Adaptation

Assumptions

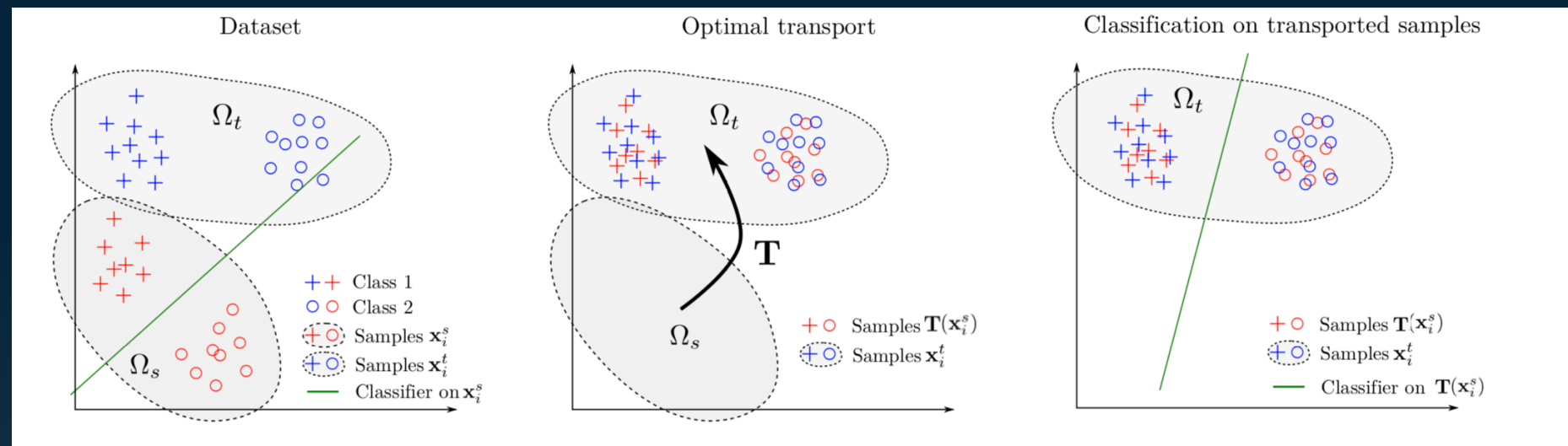
- There exist a transport T in the feature space between the two domains.
- The transport preserves the conditional distributions:

$$P_s[y|\mathbf{x}_s] = P_t[y|T(\mathbf{x}_s)]$$

3-step strategy

1. Estimate optimal transport between distributions.
2. Transport the training samples onto the target distribution using barycentric mapping (Ferradans et al., 2013)
3. Learn a classifier on the transported training samples.

(Source image: Courty et al., 2017)



5. Screened Dual of Sinkhorn Divergence

Screening Sinkhorn Algorithm for Regularized Optimal Transport

Mokhtar Z. Alaya

LITIS EA4108

University of Rouen Normandy

mokhtarzahdi.alaya@gmail.com

Maxime Bérar

LITIS EA4108

University of Rouen Normandy

maxime.berar@univ-rouen.fr



Gilles Gasso

LITIS EA4108

INSA, University of Rouen Normandy

gilles.gasso@insa-rouen.fr



Alain Rakotomamonjy

LITIS EA4108

University of Rouen Normandy

and Criteo AI Lab, Criteo Paris

alain.rakoto@insa-rouen.fr

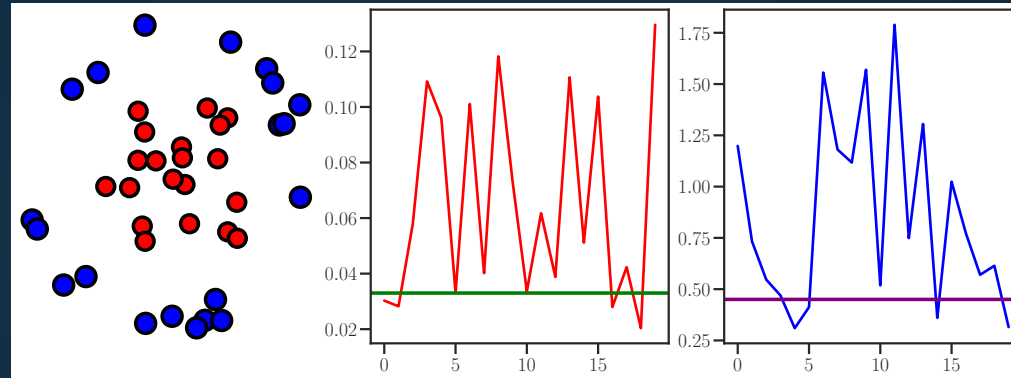


33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

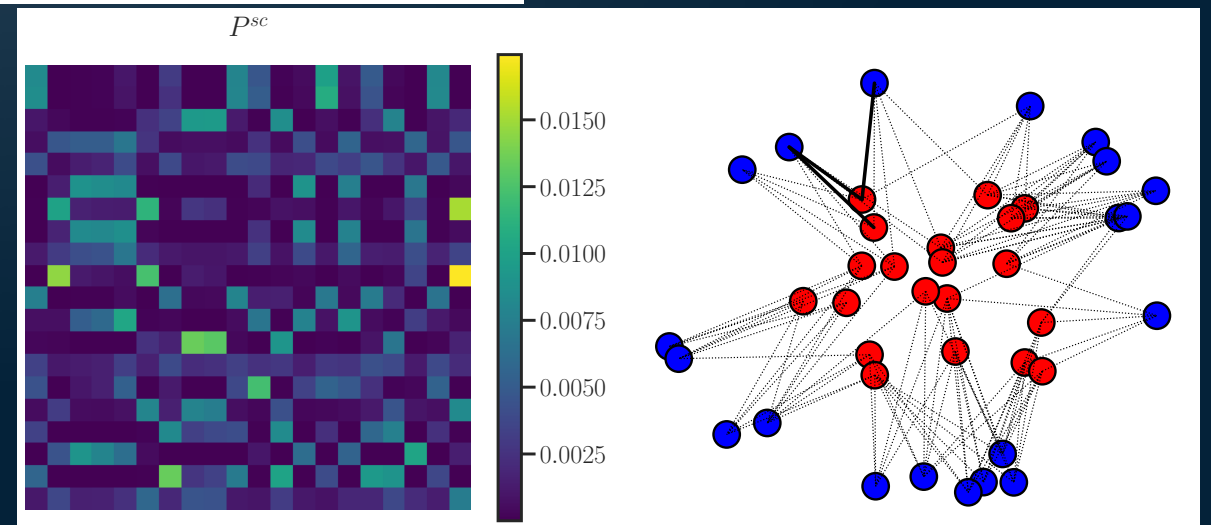
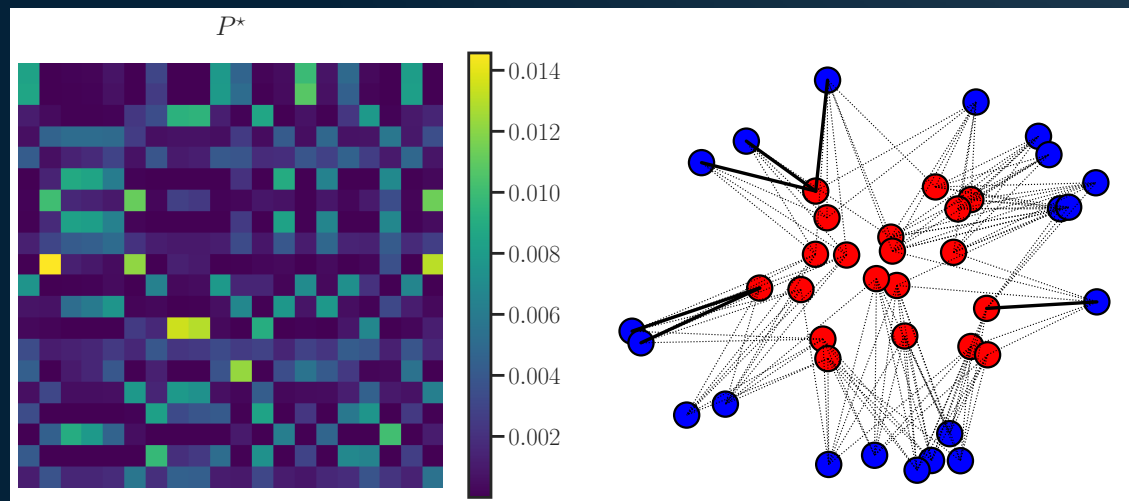
Screened Dual of Sinkhorn Divergence: Motivations

- OT plan presents a large number of neglectable values (Blondel et al., 2018)
- *Static screening test* in Lasso (Ghaoui et al., 2010)
- We define the convex set $\mathcal{C}_\alpha^r = \{w \in \mathbb{R}^r : e^{w_i} \geq \alpha\}$, for $\alpha > 0$

$P^\star$



P^{sc}



- Identify these indices and fixed at the thresholds before solving the problem.
- ➡ Reduce the scale of the optimization procedure.

Static Screening Test:

Approximate Dual of $\mathcal{S}_\eta(\mu, \nu)$

- Based on this idea, we define a so-called *approximate dual of Sinkhorn divergence*.

Approximate dual of Sinkhorn divergence

$$\mathcal{S}_\eta^{\text{ad}}(\mu, \nu) = \min_{\mathbf{u} \in \mathcal{C}_{\frac{\varepsilon}{\kappa}}^n, \mathbf{v} \in \mathcal{C}_{\varepsilon\kappa}^m} \left\{ \Psi_{\kappa}(\mathbf{u}, \mathbf{v}) := \mathbf{1}_n^\top B(\mathbf{u}, \mathbf{v}) \mathbf{1}_m - \langle \kappa \mathbf{u}, \mu \rangle - \langle \frac{\mathbf{v}}{\kappa}, \nu \rangle \right\}$$

- This is a simply dual Sinkhorn with lower-bounded variables, where the bounds are

$$\alpha_{\mathbf{u}} = \frac{\varepsilon}{\kappa} \quad \text{and} \quad \alpha_{\mathbf{v}} = \varepsilon \kappa$$

where $\varepsilon > 0$ and $\kappa > 0$ being fixed numeric constants

Static Screening Test: Definition

- κ -parameter plays a role of *scaling factor*

→ closed order of the potential components e^u and e^v

- Without κ the components e^u and e^v can have inversely related scale:

→ e^u being too large and e^v being too small

- The static screening test aims at locating two subsets of indices

$$(I, J) \in \{1, \dots, n\} \times \{1, \dots, m\}$$

Static screening test

$$\mathcal{T}(I, J)$$

$$(u, v) \in \mathcal{C}_{\alpha_u}^n \times \mathcal{C}_{\alpha_v}^m \equiv \begin{cases} e^{u_i} > \alpha_u \text{ and } e^{v_j} > \alpha_v, \forall (i, j) \in I \times J, \\ e^{u_{i'}} = \alpha_u \text{ and } e^{v_{j'}} = \alpha_v, \forall (i', j') \in I^c \times J^c. \end{cases}$$

Static Screening Test: $\mathcal{T}(I_{\varepsilon, \kappa}, J_{\varepsilon, \kappa})$

$$r_i(\mathbf{K}) = \sum_{j=1}^m \mathbf{K}_{ij} \quad \text{and} \quad c_j(\mathbf{K}) = \sum_{i=1}^n \mathbf{K}_{ij}$$

i-th row sum of $\mathbf{K}$ j-th row sum of $\mathbf{K}$

Proposition

Let $(\mathbf{u}^*, \mathbf{v}^*)$ be optimal solution of $\mathcal{S}_\eta^{\text{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Define

$$I_{\varepsilon, \kappa} = \left\{ i = 1, \dots, n : \boldsymbol{\mu}_i \geq \frac{\varepsilon^2}{\kappa} r_i(\mathbf{K}) \right\}$$

$$\text{and } J_{\varepsilon, \kappa} = \left\{ j = 1, \dots, m : \boldsymbol{\nu}_j \geq \kappa \varepsilon^2 c_j(\mathbf{K}) \right\}.$$

Then,

$$e^{\mathbf{u}_i^*} = \frac{\varepsilon}{\kappa} \text{ and } e^{\mathbf{v}_j^*} = \varepsilon \kappa \text{ for all } i \in I_{\varepsilon, \kappa}^{\mathbb{C}} \text{ and } j \in J_{\varepsilon, \kappa}^{\mathbb{C}}$$

- The parameters ε and κ are difficult to interpret, we exhibit their relations with a fixed number budget of points from the supports of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Screening with a Fixed Number Budget of Points

- We denote by $n_b \in \{1, \dots, n\}$ and $m_b \in \{1, \dots, m\}$ the number of points that are going to be optimized in $\mathcal{S}_\eta^{\text{ad}}(\mu, \nu)$.
- Let $\xi \in \mathbb{R}^n$ and $\zeta \in \mathbb{R}^m$ to be the ordered decreasing vectors of $\mu \oslash r(K)$ and $\nu \oslash c(K)$
- To keep only n_b -budget and m_b -budget of points, the parameters

$$\varepsilon^2 \kappa = \zeta_{m_b} \quad \text{and} \quad \frac{\varepsilon^2}{\kappa} = \xi_{n_b}$$

Then

$$\varepsilon = (\xi_{n_b} \zeta_{m_b})^{1/4} \quad \text{and} \quad \kappa = \sqrt{\frac{\zeta_{m_b}}{\xi_{n_b}}}$$

- This guarantees

$$|I_{\varepsilon, \kappa}| = n_b \quad \text{and} \quad |J_{\varepsilon, \kappa}| = m_b$$

Screening with a Fixed Number Budget of Points

- If $(\mathbf{u}^*, \mathbf{v}^*)$ solution of $\mathcal{S}_\eta^{\text{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ satisfies $\mathcal{T}(I_{\varepsilon, \kappa}, J_{\varepsilon, \kappa})$ with $\alpha_{\mathbf{u}^*} = \frac{\varepsilon}{\kappa}$ and $\alpha_{\mathbf{v}^*} = \varepsilon \kappa$
- We can restrict the variables in $\mathcal{S}_\eta^{\text{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ to variables in $I_{\varepsilon, \kappa}$ and $J_{\varepsilon, \kappa}$
- This boils down to restricting the constraints feasibility $\mathcal{C}_{\frac{\varepsilon}{\kappa}}^n \cap \mathcal{C}_{\varepsilon \kappa}^m$ to the screened domain defined by $\mathcal{U}^{\text{sc}} \cap \mathcal{V}^{\text{sc}}$

Screened Feasibility Domain

$$\mathcal{U}^{\text{sc}} = \{\mathbf{u} \in \mathbb{R}^{n_b} : e^{\mathbf{u}_i} \geq \frac{\varepsilon}{\kappa}\} \text{ and } \mathcal{V}^{\text{sc}} = \{\mathbf{v} \in \mathbb{R}^{m_b} : e^{\mathbf{v}_j} \geq \varepsilon \kappa\}$$

Screening with a Fixed Number Budget of Points

- By replacing in $\mathcal{S}_\eta^{\text{ad}}(\mu, \nu)$, the variables belonging to $(I_{\varepsilon, \kappa}^{\mathbb{C}} \times J_{\varepsilon, \kappa}^{\mathbb{C}})$ by $\frac{\varepsilon}{\kappa}$ and $\varepsilon\kappa$, we derive the screened dual of Sinkhorn divergence as

Screened Dual of Sinkhorn Divergence

$$\mathcal{S}_\eta^{\text{scd}}(\mu, \nu) = \min_{u \in \mathcal{U}_{\text{sc}}, v \in \mathcal{V}_{\text{sc}}} \{ \Psi_{\varepsilon, \kappa}(u, v) \}$$

where

$$\Psi_{\varepsilon, \kappa}(u, v) := (e^{u_{I_{\varepsilon, \kappa}}})^\top K_{(I_{\varepsilon, \kappa}, J_{\varepsilon, \kappa})} e^{v_{J_{\varepsilon, \kappa}}} + \varepsilon\kappa (e^{u_{I_{\varepsilon, \kappa}}})^\top K_{(I_{\varepsilon, \kappa}, J_{\varepsilon, \kappa}^{\mathbb{C}})} \mathbf{1}_{m_b} + \frac{\varepsilon}{\kappa} \mathbf{1}_{n_b}^\top K_{(I_{\varepsilon, \kappa}^{\mathbb{C}}, J_{\varepsilon, \kappa})} e^{v_{J_{\varepsilon, \kappa}}} \\ - \kappa \mu_{I_{\varepsilon, \kappa}}^\top u_{I_{\varepsilon, \kappa}} - \kappa^{-1} \nu_{J_{\varepsilon, \kappa}}^\top v_{J_{\varepsilon, \kappa}} + \Xi$$

with

$$\Xi = \varepsilon^2 \sum_{i \in I_{\varepsilon, \kappa}^{\mathbb{C}}, j \in J_{\varepsilon, \kappa}^{\mathbb{C}}} K_{ij} - \kappa \log(\varepsilon\kappa^{-1}) \sum_{i \in I_{\varepsilon, \kappa}^{\mathbb{C}}} \mu_i - \kappa^{-1} \log(\varepsilon\kappa) \sum_{j \in J_{\varepsilon, \kappa}^{\mathbb{C}}} \nu_j$$

L-BFGS-B: Box Constraints on $(\mathbf{u}^{\text{sc}}, \mathbf{v}^{\text{sc}})$

For calculating objective function of $\mathcal{S}_\eta^{\text{scd}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ uses only the restricted parts

$$\mathbf{K} \begin{array}{|c|c|} \hline \mathbf{K}_{(I_{\varepsilon, \kappa}, J_{\varepsilon, \kappa})} & \mathbf{K}_{(I_{\varepsilon, \kappa}, J_{\varepsilon, \kappa}^{\mathcal{C}})} \\ \hline \mathbf{K}_{(I_{\varepsilon, \kappa}^{\mathcal{C}}, J_{\varepsilon, \kappa})} & \text{Diagonal lines} \\ \hline \end{array}$$

Proposition

Let $(\mathbf{u}^{\text{sc}}, \mathbf{v}^{\text{sc}})$ be an optimal pair solution of the screened dual $\mathcal{S}_\eta^{\text{scd}}(\boldsymbol{\mu}, \boldsymbol{\nu})$

Define $\mathbf{K}_{\min} = \min_{i \in I_{\varepsilon, \kappa}, j \in J_{\varepsilon, \kappa}} \mathbf{K}_{ij}$ Then,

$$\frac{\varepsilon}{\kappa} \vee \frac{\min_{i \in I_{\varepsilon, \kappa}} \mu_i}{\varepsilon(m - m_b) + \varepsilon \vee \frac{\max_{j \in J_{\varepsilon, \kappa}} \nu_j}{n \varepsilon \mathbf{K}_{\min}} m_b} \leq e^{\mathbf{u}_i^{\text{sc}}} \leq \frac{\varepsilon}{\kappa} \vee \frac{\max_{i \in I_{\varepsilon, \kappa}} \mu_i}{m \varepsilon \mathbf{K}_{\min}}$$

and

$$\varepsilon \kappa \vee \frac{\min_{j \in J_{\varepsilon, \kappa}} \nu_j}{\varepsilon(n - n_b) + \varepsilon \vee \frac{\kappa \max_{i \in I_{\varepsilon, \kappa}} \mu_i}{m \varepsilon \mathbf{K}_{\min}} n_b} \leq e^{\mathbf{v}_j^{\text{sc}}} \leq \varepsilon \kappa \vee \frac{\max_{j \in J_{\varepsilon, \kappa}} \nu_j}{n \varepsilon \mathbf{K}_{\min}}$$

Screenkhorn

SCREENKHORN($\mathbf{C}, \eta, \boldsymbol{\mu}, \boldsymbol{\nu}, n_b, m_b$)

1. $\mathbf{K} \leftarrow e^{-\mathbf{C}/\eta}$;
2. $\boldsymbol{\xi} \leftarrow \text{sort}(\boldsymbol{\mu} \odot r(\mathbf{K})), \boldsymbol{\zeta} \leftarrow \text{sort}(\boldsymbol{\nu} \odot c(\mathbf{K}))$; (decreasing order)
3. $\varepsilon \leftarrow (\boldsymbol{\xi}_{n_b} \boldsymbol{\zeta}_{m_b})^{1/4}, \kappa \leftarrow \sqrt{\boldsymbol{\zeta}_{m_b} / \boldsymbol{\xi}_{n_b}}$
4. $I_{\varepsilon, \kappa} \leftarrow \{i = 1, \dots, n : \mu_i \geq \varepsilon^2 \kappa^{-1} r_i(\mathbf{K})\}$;
5. $J_{\varepsilon, \kappa} \leftarrow \{j = 1, \dots, m : \nu_j \geq \varepsilon^2 \kappa c_j(\mathbf{K})\}$;
6. $\mathbf{K}_{\min} = \min_{I_{\varepsilon, \kappa}, J_{\varepsilon, \kappa}} \mathbf{K}_{ij}$
7. $\underline{\boldsymbol{\mu}} \leftarrow \min_{i \in I_{\varepsilon, \kappa}} \mu_i, \bar{\boldsymbol{\mu}} \leftarrow \max_{i \in I_{\varepsilon, \kappa}} \mu_i; \quad \underline{\boldsymbol{\nu}} \leftarrow \min_{j \in J_{\varepsilon, \kappa}} \nu_j, \bar{\boldsymbol{\nu}} \leftarrow \max_{j \in J_{\varepsilon, \kappa}} \nu_j$;
8. $\underline{\mathbf{u}} \leftarrow \log\left(\frac{\varepsilon}{\kappa} \vee \frac{\underline{\boldsymbol{\mu}}}{\varepsilon(m - m_b) + \varepsilon \vee \frac{\bar{\boldsymbol{\nu}}}{n\varepsilon\mathbf{K}_{\min}} m_b}\right), \bar{\mathbf{u}} \leftarrow \log\left(\frac{\varepsilon}{\kappa} \vee \frac{\bar{\boldsymbol{\mu}}}{m\varepsilon\mathbf{K}_{\min}}\right)$;
9. $\underline{\mathbf{v}} \leftarrow \log\left(\varepsilon\kappa \vee \frac{\underline{\boldsymbol{\nu}}}{\varepsilon(n - n_b) + \varepsilon \vee \frac{\kappa\bar{\boldsymbol{\mu}}}{m\varepsilon\mathbf{K}_{\min}} n_b}\right), \bar{\mathbf{v}} \leftarrow \log\left(\varepsilon\kappa \vee \frac{\bar{\boldsymbol{\nu}}}{n\varepsilon\mathbf{K}_{\min}}\right)$;
10. $\bar{\boldsymbol{\theta}} \leftarrow \text{stack}(\bar{\mathbf{u}}\mathbf{1}_{n_b}, \bar{\mathbf{v}}\mathbf{1}_{m_b}), \underline{\boldsymbol{\theta}} \leftarrow \text{stack}(\underline{\mathbf{u}}\mathbf{1}_{n_b}, \underline{\mathbf{v}}\mathbf{1}_{m_b})$;

STEP 1: SCREENING

11. $\mathbf{u}^{(0)} \leftarrow \log(\varepsilon\kappa^{-1})\mathbf{1}_{n_b}, \mathbf{v}^{(0)} \leftarrow \log(\varepsilon\kappa)\mathbf{1}_{m_b}$;
12. $\boldsymbol{\theta}^{(0)} \leftarrow \text{stack}(\mathbf{u}^{(0)}, \mathbf{v}^{(0)})$;
13. $\boldsymbol{\theta} \leftarrow \text{L-BFGS-B}(\boldsymbol{\theta}^{(0)}, \underline{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})$;
14. $\boldsymbol{\theta}_u \leftarrow (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n_b})^\top$;
15. $\boldsymbol{\theta}_v \leftarrow (\boldsymbol{\theta}_{n_b+1}, \dots, \boldsymbol{\theta}_{n_b+m_b})^\top$;
16. $\mathbf{u}_i^{\text{sc}} \leftarrow (\boldsymbol{\theta}_u)_i$ if $i \in I_{\varepsilon, \kappa}$ and $\mathbf{u}_i^{\text{sc}} \leftarrow \log(\varepsilon\kappa^{-1})$ if $i \in I_{\varepsilon, \kappa}^c$;
17. $\mathbf{v}_j^{\text{sc}} \leftarrow (\boldsymbol{\theta}_v)_j$ if $j \in J_{\varepsilon, \kappa}$ and $\mathbf{v}_j^{\text{sc}} \leftarrow \log(\varepsilon\kappa)$ if $j \in J_{\varepsilon, \kappa}^c$;
18. Return $B(\mathbf{u}^{\text{sc}}, \mathbf{v}^{\text{sc}})$;

STEP 2: L-BFGS-B

Theoretical Analysis and Guarantees

Screened optimal transport plan $P^{\text{sc}} = \text{diag}(e^{\mathbf{u}^{\text{sc}}}) \mathbf{K} \text{diag}(e^{\mathbf{v}^{\text{sc}}})$

Screened marginals $\mu^{\text{sc}} = P^{\text{sc}} \mathbf{1}_m$ and $\nu^{\text{sc}} = (P^{\text{sc}})^{\top} \mathbf{1}_n$

Proposition

$$\|\mu - \mu^{\text{sc}}\|_1^2 = \mathcal{O}\left(n_b c_{\kappa} + (n - n_b) \left(\frac{\|C\|_{\infty}}{\eta} + \frac{m_b}{\sqrt{nm} c_{\mu\nu}} \mathbf{K}_{\min}^{3/2} + \frac{m - m_b}{\sqrt{nm} \mathbf{K}_{\min}} + \log \left(\frac{\sqrt{nm}}{m_b c_{\mu\nu}^{5/2}} \right) \right) \right)$$

$$\|\nu - \nu^{\text{sc}}\|_1^2 = \mathcal{O}\left(m_b c_{\frac{1}{\kappa}} + (m - m_b) \left(\frac{\|C\|_{\infty}}{\eta} + \frac{n_b}{\sqrt{nm} c_{\mu\nu}} \mathbf{K}_{\min}^{3/2} + \frac{n - n_b}{\sqrt{nm} \mathbf{K}_{\min}} + \log \left(\frac{\sqrt{nm}}{n_b c_{\mu\nu}^{5/2}} \right) \right) \right)$$

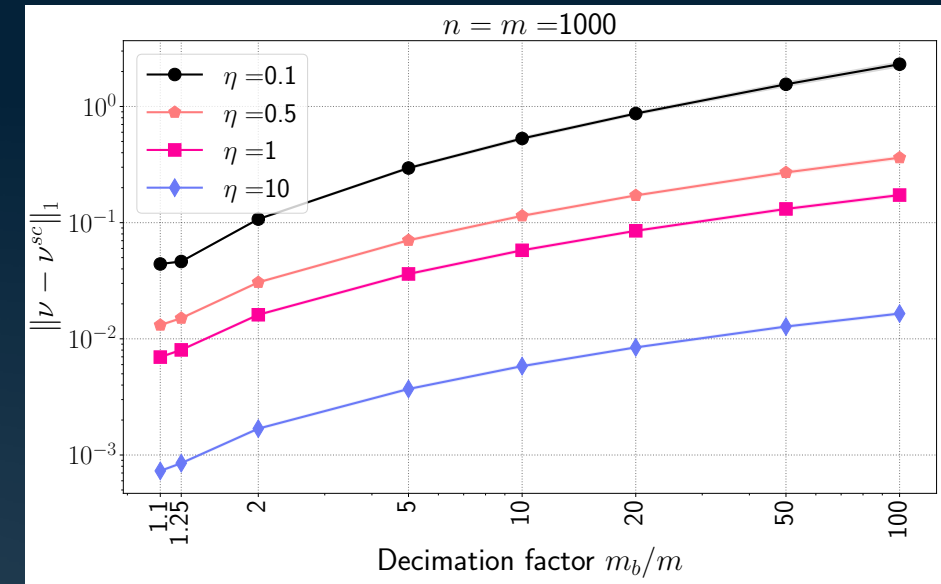
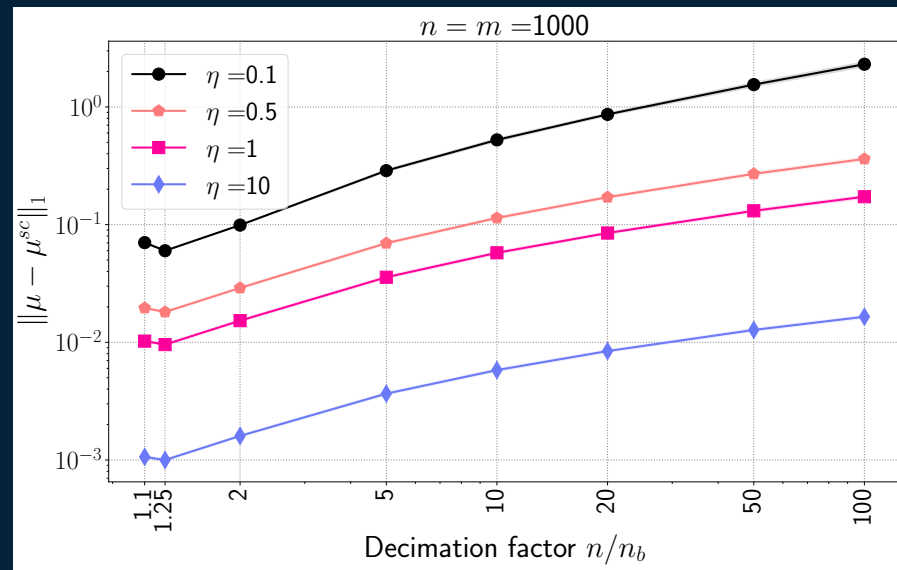
$$c_z = z - \log z - 1 \text{ for } z > 0, c_{\mu\nu} = \underline{\mu} \wedge \underline{\nu} \text{ with } \underline{\mu} = \min_{i \in I_{\varepsilon, \kappa}} \mu_i, \quad \underline{\nu} = \min_{j \in J_{\varepsilon, \kappa}} \nu_j.$$

Proposition

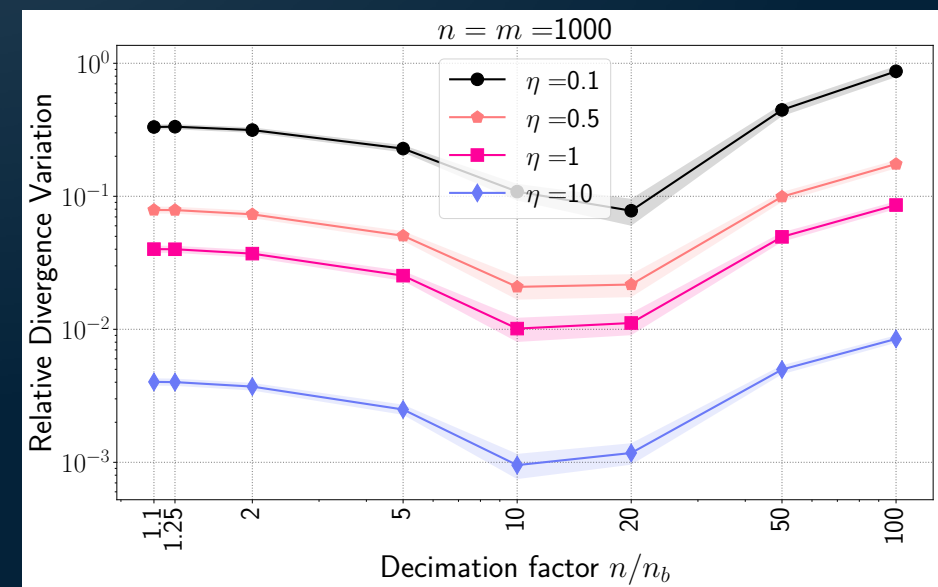
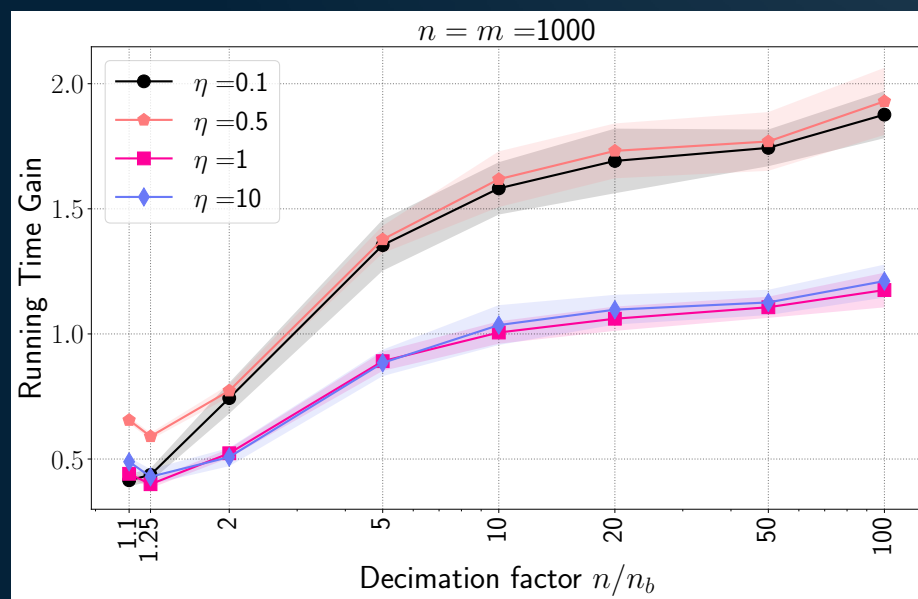
$$\Psi_{\varepsilon, \kappa}(\mathbf{u}^{\text{sc}}, \mathbf{v}^{\text{sc}}) - \Psi(\mathbf{u}^{\star}, \mathbf{v}^{\star}) = \mathcal{O}\left(R(\|\mu - \mu^{\text{sc}}\|_1 + \|\nu - \nu^{\text{sc}}\|_1 + \omega_{\kappa})\right)$$

$$R = \frac{\|C\|_{\infty}}{\eta} + \log \left(\frac{(n \vee m)^2}{nm c_{\mu\nu}^{7/2}} \right) \text{ and } \omega_{\kappa} = |1 - \kappa| \|\mu^{\text{sc}}\|_1 + |1 - \kappa^{-1}| \|\nu^{\text{sc}}\|_1 + |1 - \kappa| + |1 - \kappa^{-1}|$$

Theoretical Analysis and Guarantees: (Simulation on Toy Data)



Marginal violations in relation with the budget of points



$$\frac{T_{\text{SINKHORN}}}{T_{\text{SCREENKHORN}}}$$

$$\frac{|\langle \mathbf{C}, \mathbf{P}^* \rangle - \langle \mathbf{C}, \mathbf{P}^{sc} \rangle|}{\langle \mathbf{C}, \mathbf{P}^* \rangle}$$

Real Dataset for OTDA: MNIST(source) to USPS (target)

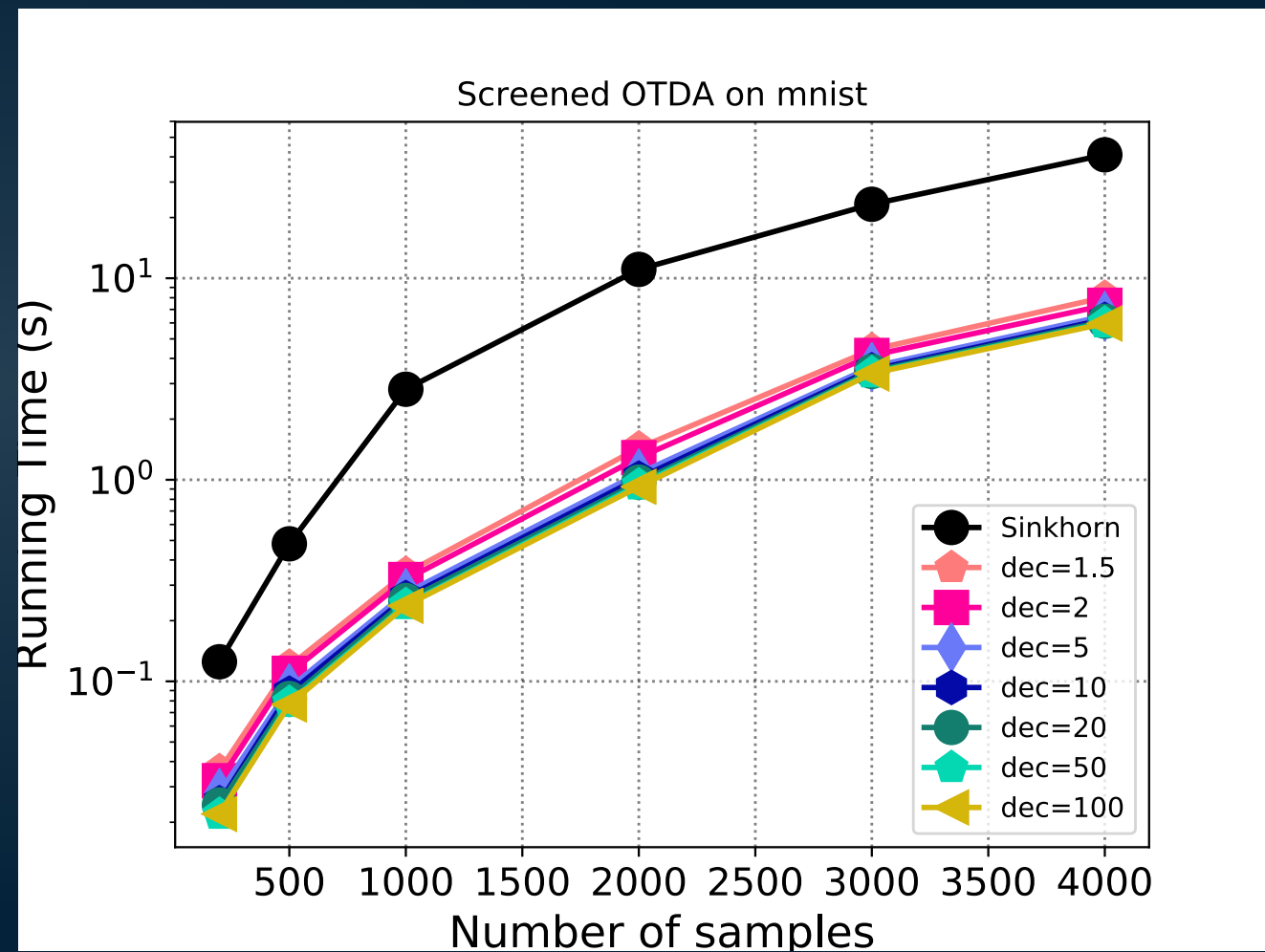
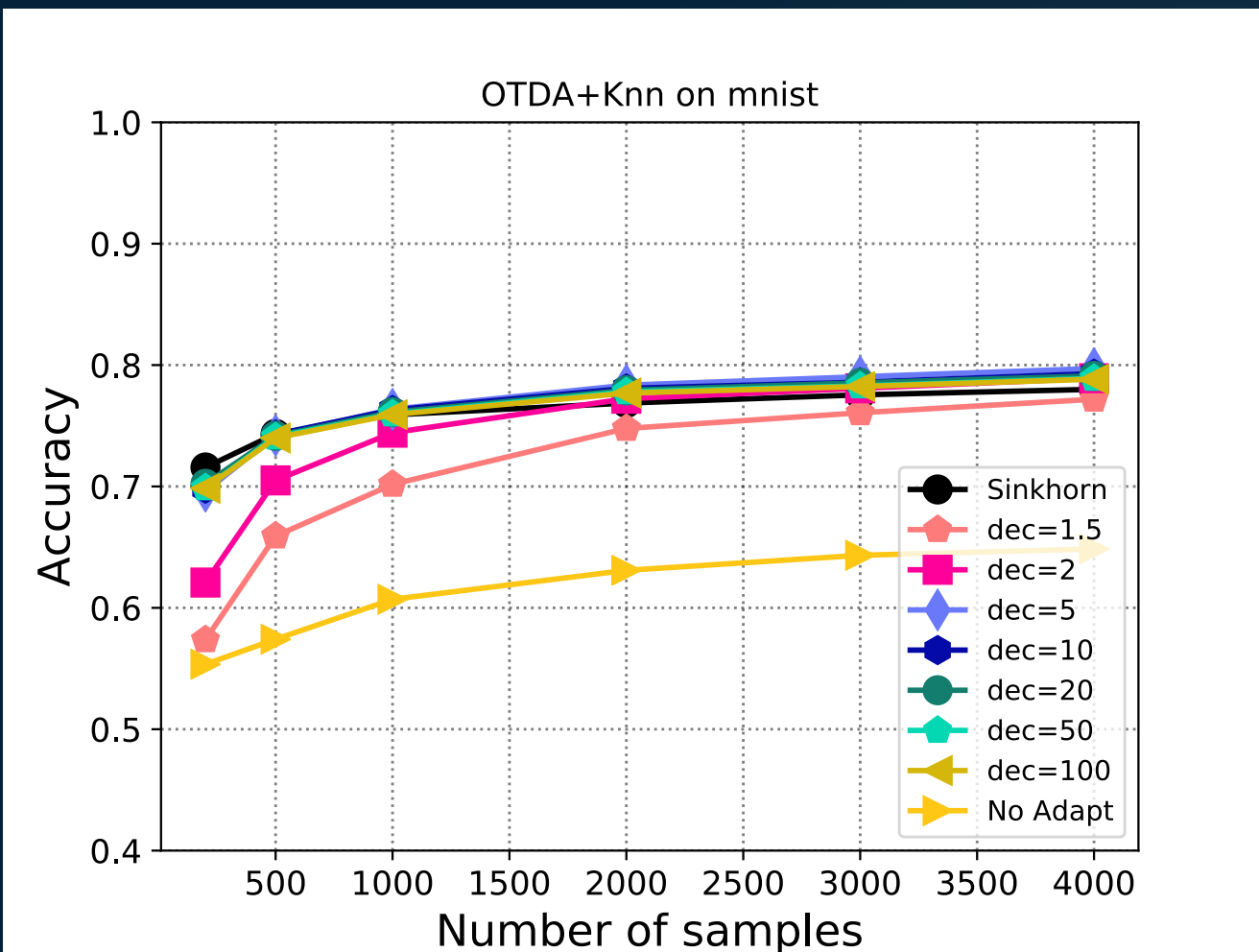


MNIST



USPS

Real Dataset for OTDA: MNIST(source) to USPS (target)



2. Unblanced OT: Partial Wasserstein Distance

Partial Optimal Transport with Applications on Positive-Unlabeled Learning



Laetitia Chapel
Univ. Bretagne-Sud, CNRS, IRISA
F-56000 Vannes
laetitia.chapel@irisa.fr

Mokhtar Z. Alaya
LITIS EA4108

University of Rouen Normandy
mokhtarzahdi.alaya@gmail.com



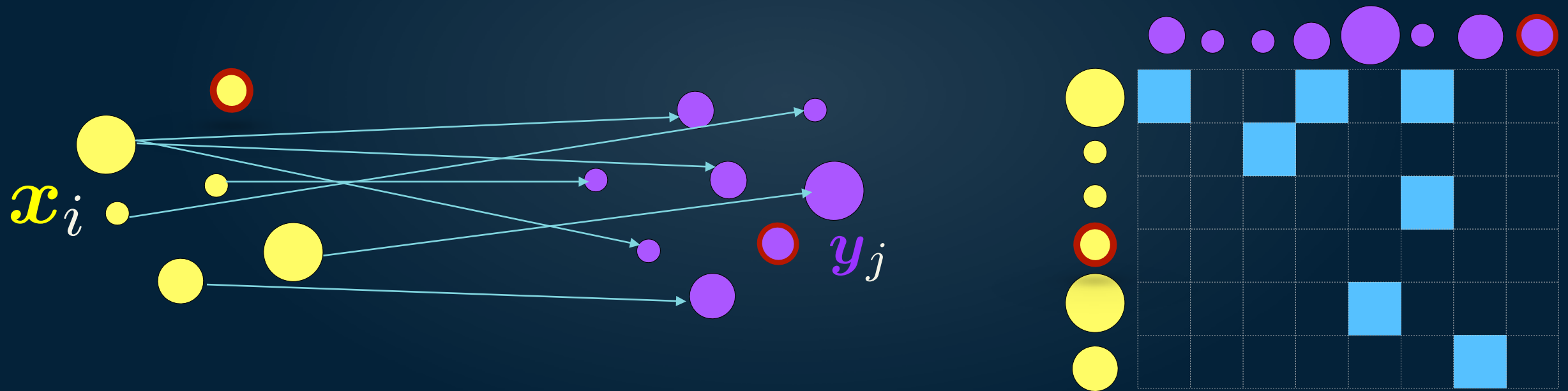
Gilles Gasso
LITIS EA4108
INSA, University of Rouen Normandy
gilles.gasso@insa-rouen.fr

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Partial OT: Partial Wasserstein Distance

- Partial OT problem focuses on transporting only a fraction

$$0 \leq s \leq \min(\|\mathbf{p}\|_1, \|\mathbf{q}\|_1)$$



- The set of admissible coupling becomes

$$\Pi^u(\mathbf{p}, \mathbf{q}) = \{\mathbf{T} \in \mathbb{R}_+^{n \times m}, \mathbf{T}\mathbf{1}_m \leq \mathbf{p}, \mathbf{T}^\top \mathbf{1}_n \leq \mathbf{q}, \mathbf{1}_n^\top \mathbf{T} \mathbf{1}_m = s\}$$

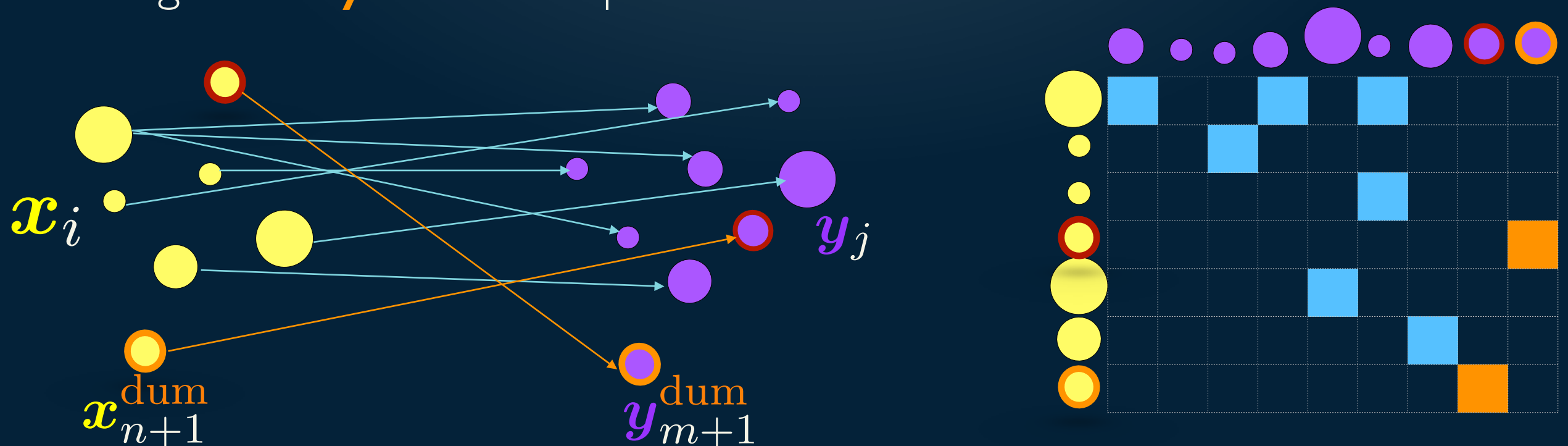
Partial OT: Partial Wasserstein Distance

- The partial-Wasserstein distance reads as:

$$\mathcal{PW}_p^p(\mathbf{p}, \mathbf{q}) = \min_{\mathbf{T} \in \Pi^u(\mathbf{p}, \mathbf{q})} \left\{ \langle \mathbf{C}, \mathbf{T} \rangle = \sum_{i=1}^n \sum_{j=1}^m \mathbf{C}_{ij} \mathbf{T}_{ij} \right\}$$

[Caffarelli and McCann, '10; Figalli, '10, Benamou et al., '15; Chizat et al., '18]

- Solution:** we propose to directly solve the exact partial Wasserstein by adding **dummy** or **virtual** points $\mathbf{x}_{n+1}^{\text{dum}}$ and $\mathbf{y}_{m+1}^{\text{dum}}$:



Partial OT: Partial Wasserstein Distance

- We extend the cost matrix as follows:

$$\bar{C} = \begin{bmatrix} C & \xi \mathbf{1}_m \\ \xi \mathbf{1}_n^\top & 2\xi + A \end{bmatrix} \quad \text{for some } A > 0 \text{ and } \xi \geq 0,$$

and the mass probability vectors as:

$$\bar{p} = [p, \|q\|_1 - s] \quad \bar{q} = [\bar{q}, \|p\|_1 - s]$$

- Hence,

$$\bar{p} = \sum_{i=1}^n p_i \delta_{x_i} + (\|q\|_1 - s) \delta_{x_{n+1}^{\text{dum}}}$$

$$\bar{q} = \sum_{j=1}^m q_j \delta_{y_j} + (\|p\|_1 - s) \delta_{y_{m+1}^{\text{dum}}}$$

Exact Partial Wasserstein Distance

- Let us define $\bar{T}^*$ the solution of the extended problem with $(\bar{C}, \bar{p}, \bar{q})$.
Namely:

$$\bar{T}^* \in \mathcal{W}_p^p(\bar{p}, \bar{q}) := \min_{\bar{T} \in \Pi(\bar{p}, \bar{q})} \langle \bar{C}, \bar{T} \rangle.$$

Proposition

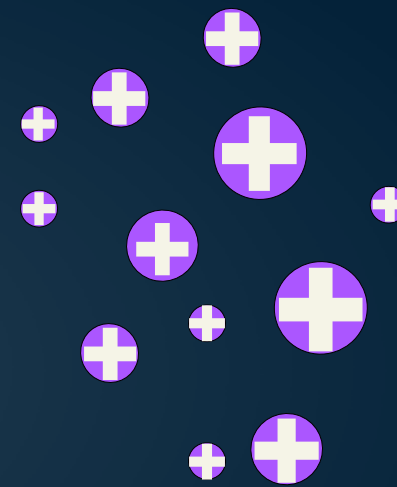
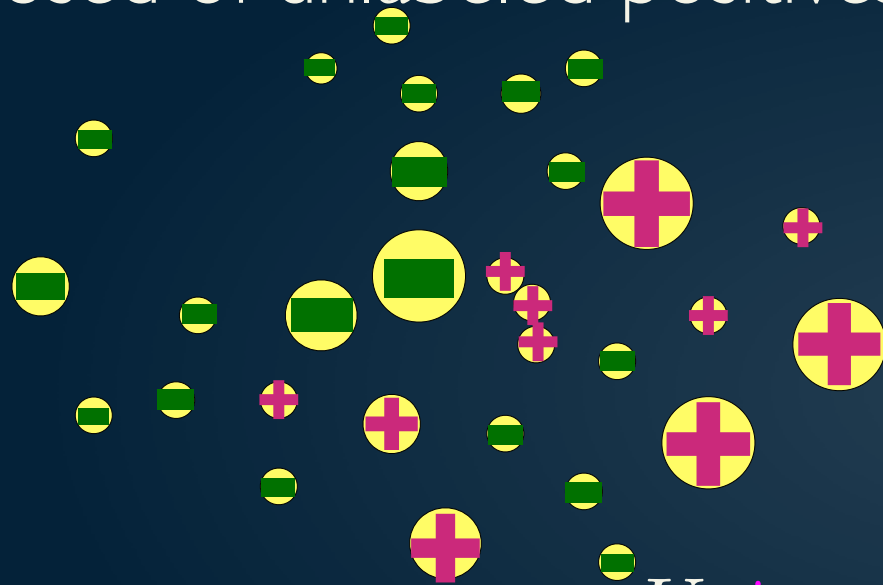
Assume that $A > \max(C_{ij})$ and ξ is bounded, one has:

$$\mathcal{W}_p^p(\bar{p}, \bar{q}) - \mathcal{PW}_p^p(p, q) = \xi(\|p\|_1 + \|q\|_1 - 2s).$$

The partial optimum transport plan T^* of the partial Wasserstein problem is the optimum transport plan $\bar{T}^*$ deprived from its last row and column.

Overview of PU Learning

- PU learning is a variant of classical binary classification problem.
- Training data consists of only positive points **Pos** and testing data is composed of unlabeled positives and negatives **Unl**.



$$\mathbf{Unl} = \{x_i^U\}_{i=1}^{n_U} = \{x_i^{U,+}\} \cup \{x_i^{U,-}\} \quad \mathbf{Pos} = \{x_i^P\}_{i=1}^{n_P}$$

[Liu et al., '03; Denis et al., '05; Elkan and Noto, '08; Du Pelessis, '15; Bekker and Davis, '20]

- The true proportion of positives within **Unl**, called class prior, is assumed to be known and given by:

$$\pi = \mathbb{P}(y = +1 | o = 0)$$

unobserved positive sample

Take Home Message

- A powerful tool, well theoretically grounded, for manipulating distributions in machine learning.
- Despite its initial computational complexity, a lot of applications, even in large scale/deep learning settings.
- Others OT aspects (out the scope of the presentation): **Gromov-Wasserstein distance** (working with structured data), **Sliced Wasserstein**, **Multimarginal Optimal Transport (MOT)** and many more !

Some References

- G. Peyré and M. Cuturi,
Computational Optimal Transport with Applications to Data Sciences, 2019
- N. country, R. Flamary, D. Tuia and A. Rakotomamonjy.
Optimal Transport for Domain Adaptation, *PAMI* 2017
- R. Flamary et al. POT: Python Optimal Transport Library, 2017.

POT: Python Optimal Transport

Contents

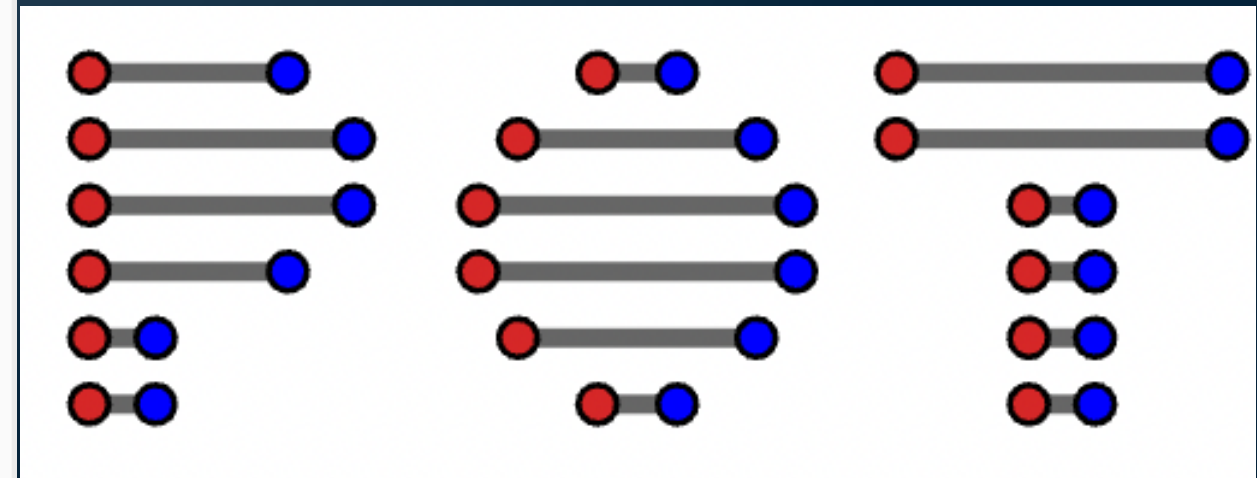
- [POT: Python Optimal Transport](#)
- [Quick start guide](#)
- [API and modules](#)
- [Examples gallery](#)
- [Releases](#)

pypi package **0.7.0** Anaconda Cloud **0.7.0** build **passing** codecov **92%** downloads **177k**
downloads **86k total** license **MIT**

This open source Python library provide several solvers for optimization problems related to Optimal Transport for signal, image processing and machine learning.

Website and documentation: <https://PythonOT.github.io/>

Source Code (MIT): <https://github.com/PythonOT/POT>

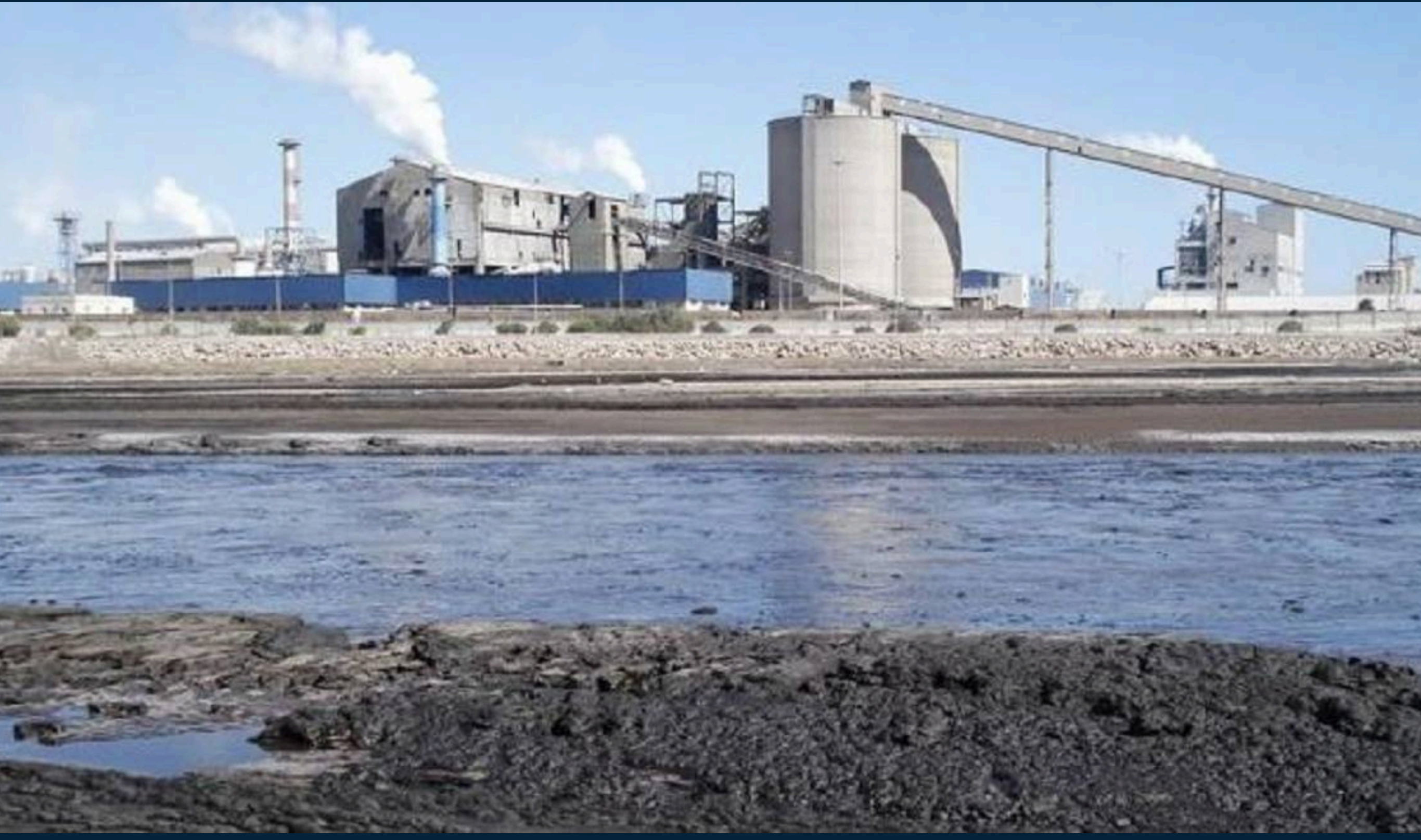


Color Transfer: Sensibility to Pollution Problem in Gabes

Color Transfer: Oasis Chenini



Color Transfer: GCT



Color Transfer: CGT -> Oasis



Thank You for your attention!