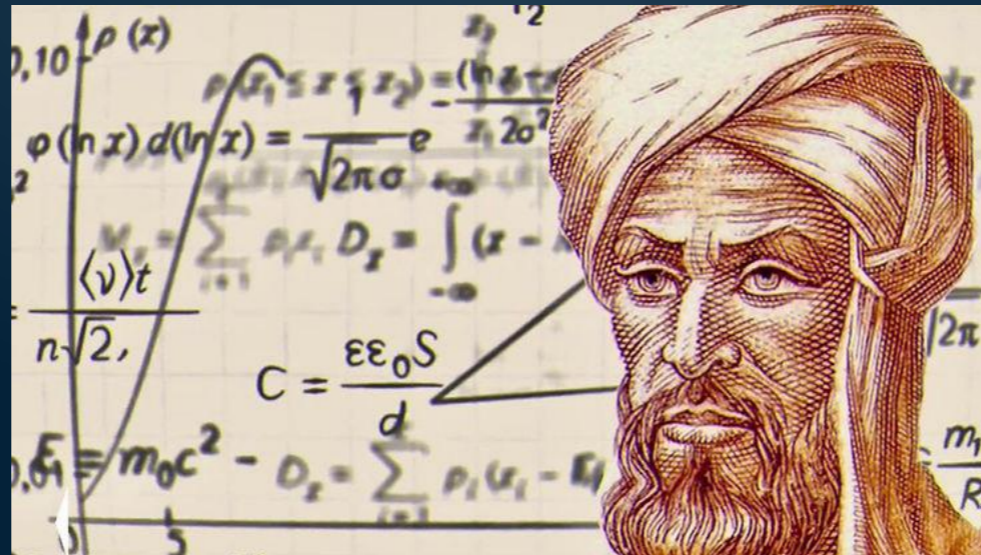


# Optimal Transport Meets Privacy: Gaussian-Smoothed Divergences for Private Distribution Learning and Domain Adaptation

Mokhtar Z. Alaya

Al-Khwarizmi

Applied Mathematics Webinar

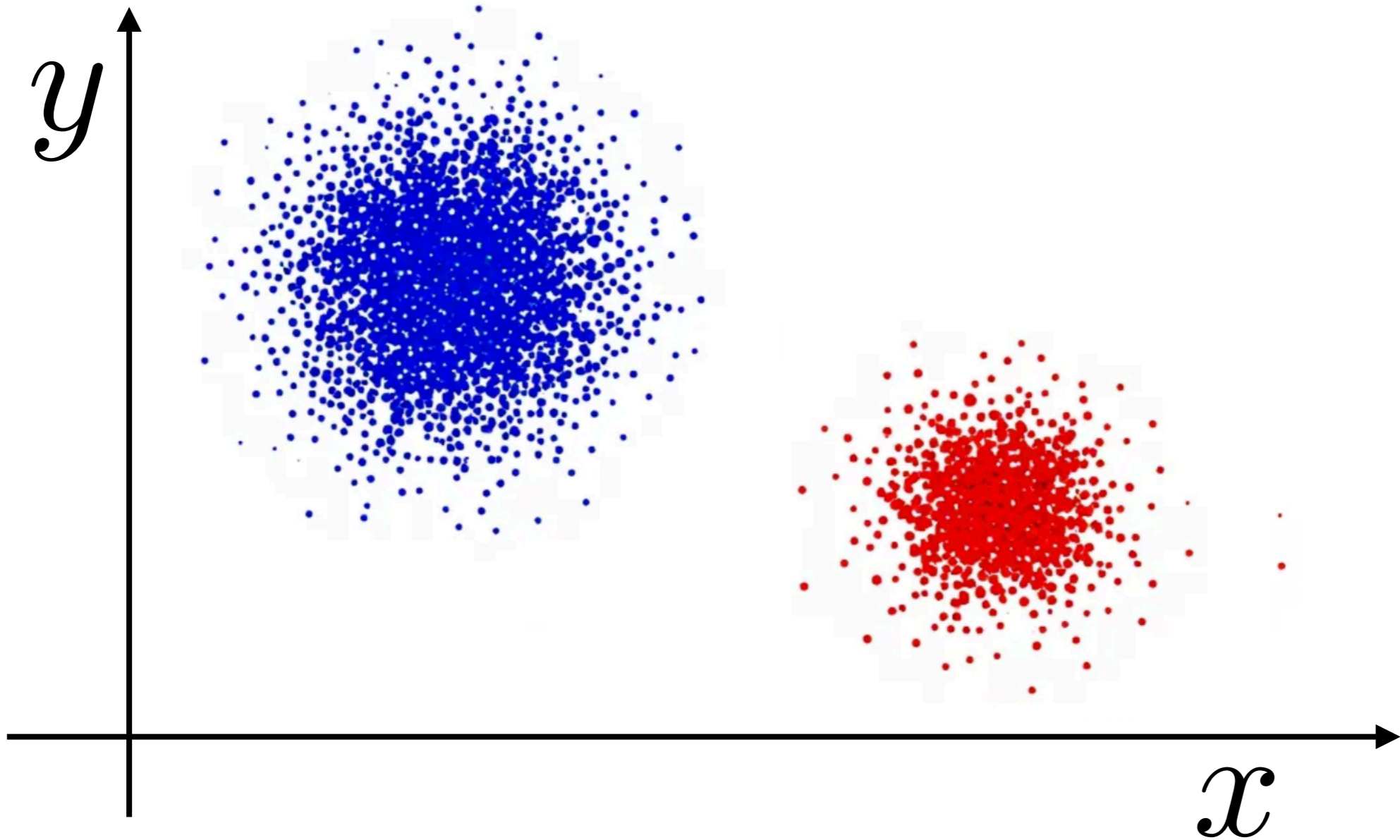


May 12, 2026

# I. Motivations

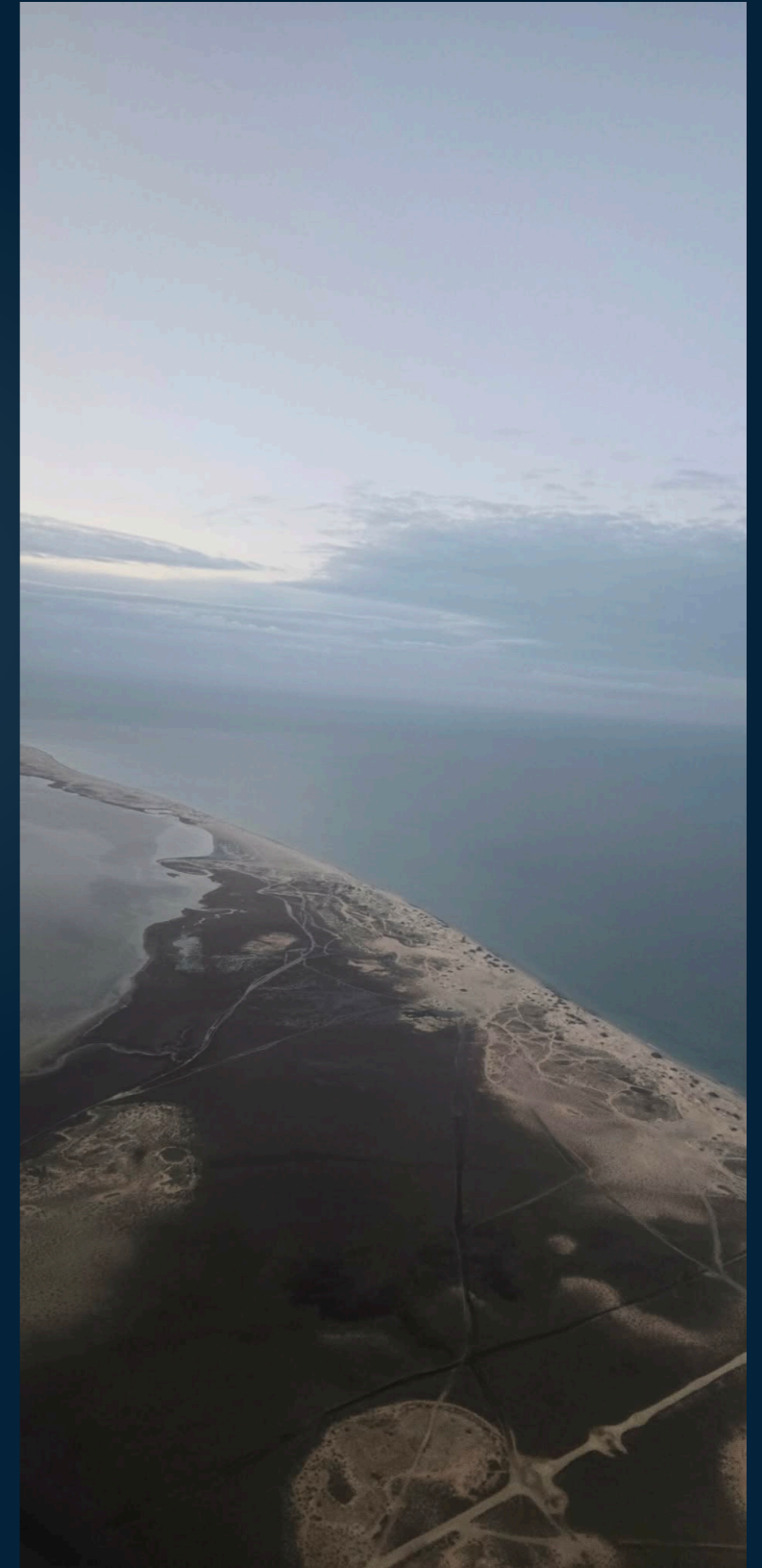
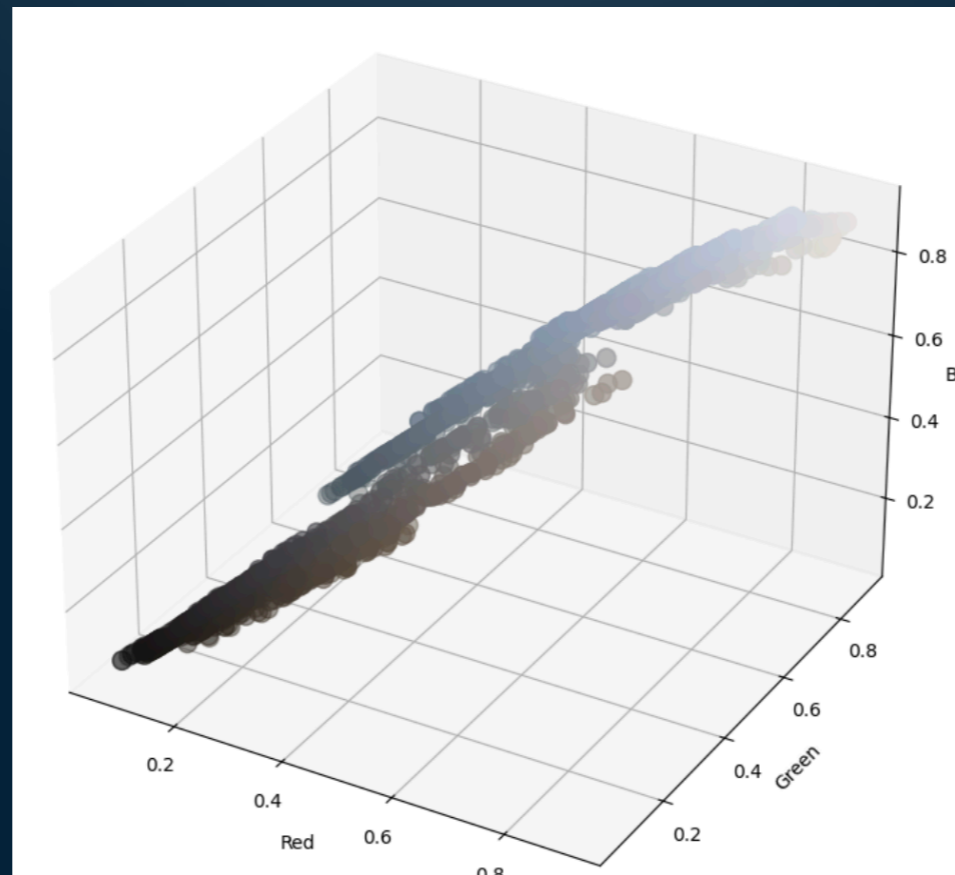
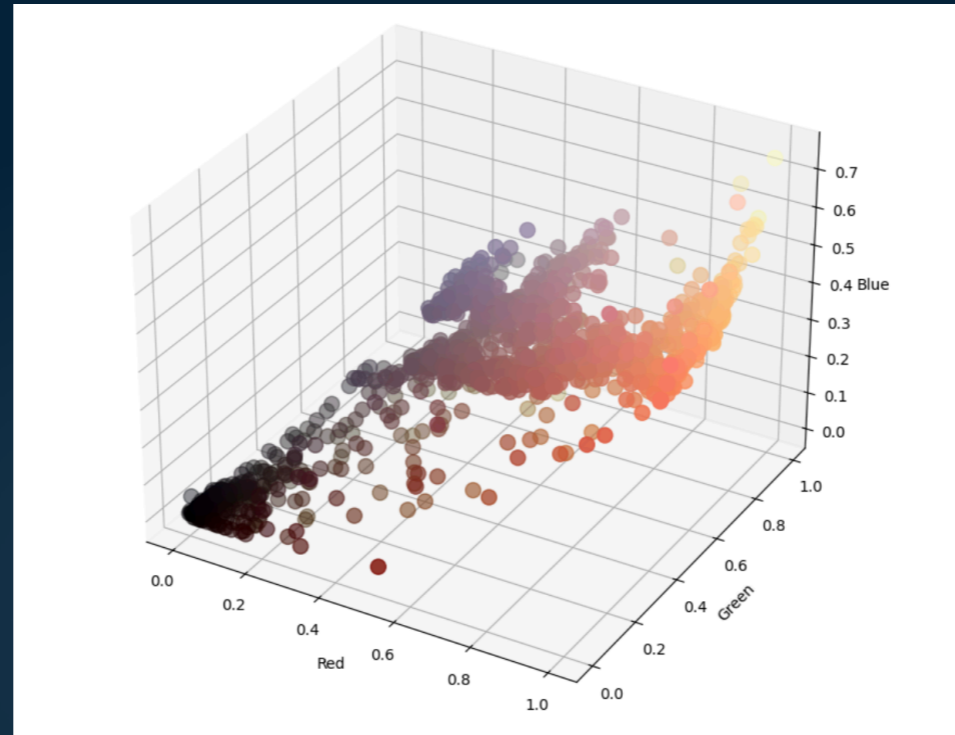
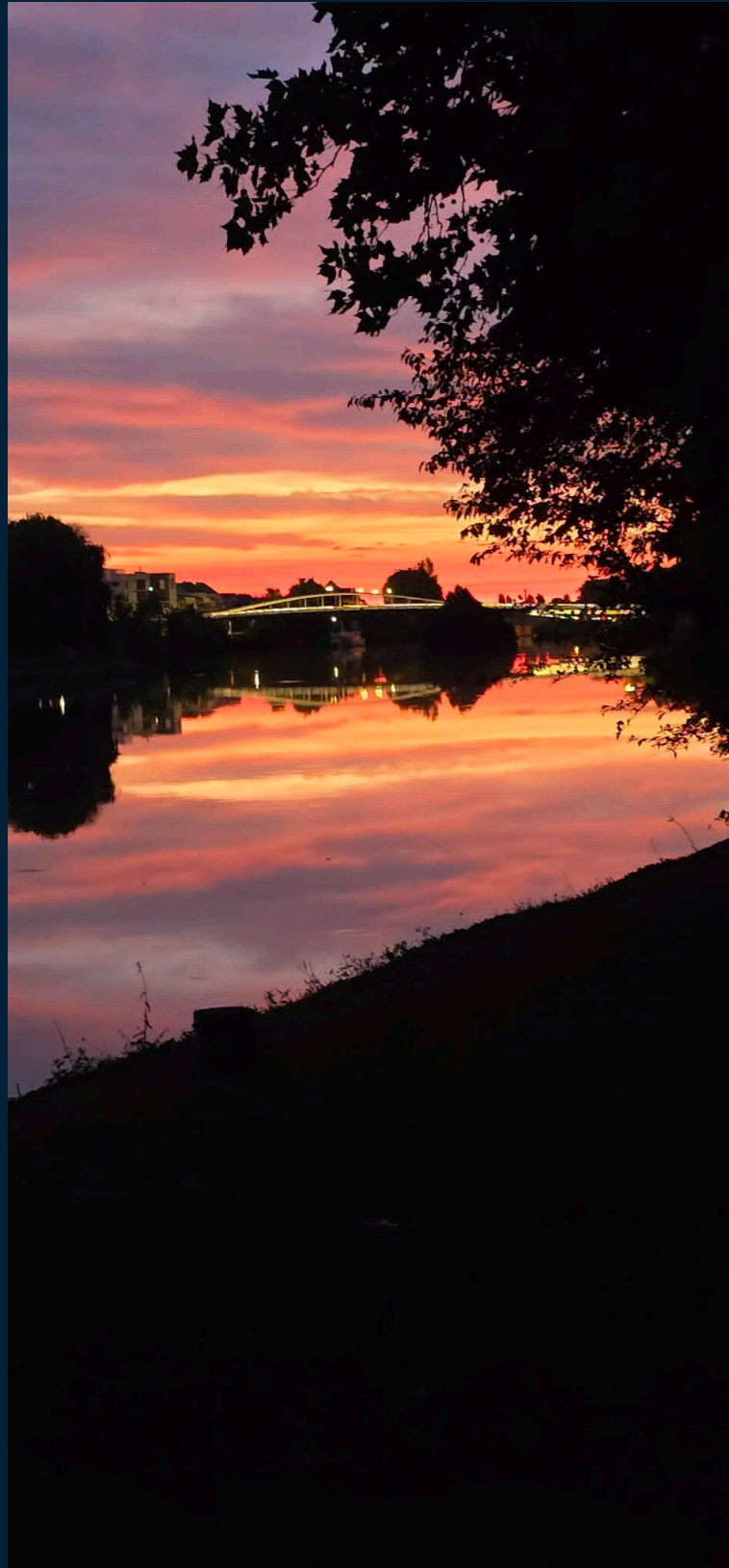
# ... in Cloud of points

- How to compare two sets of clouds points?



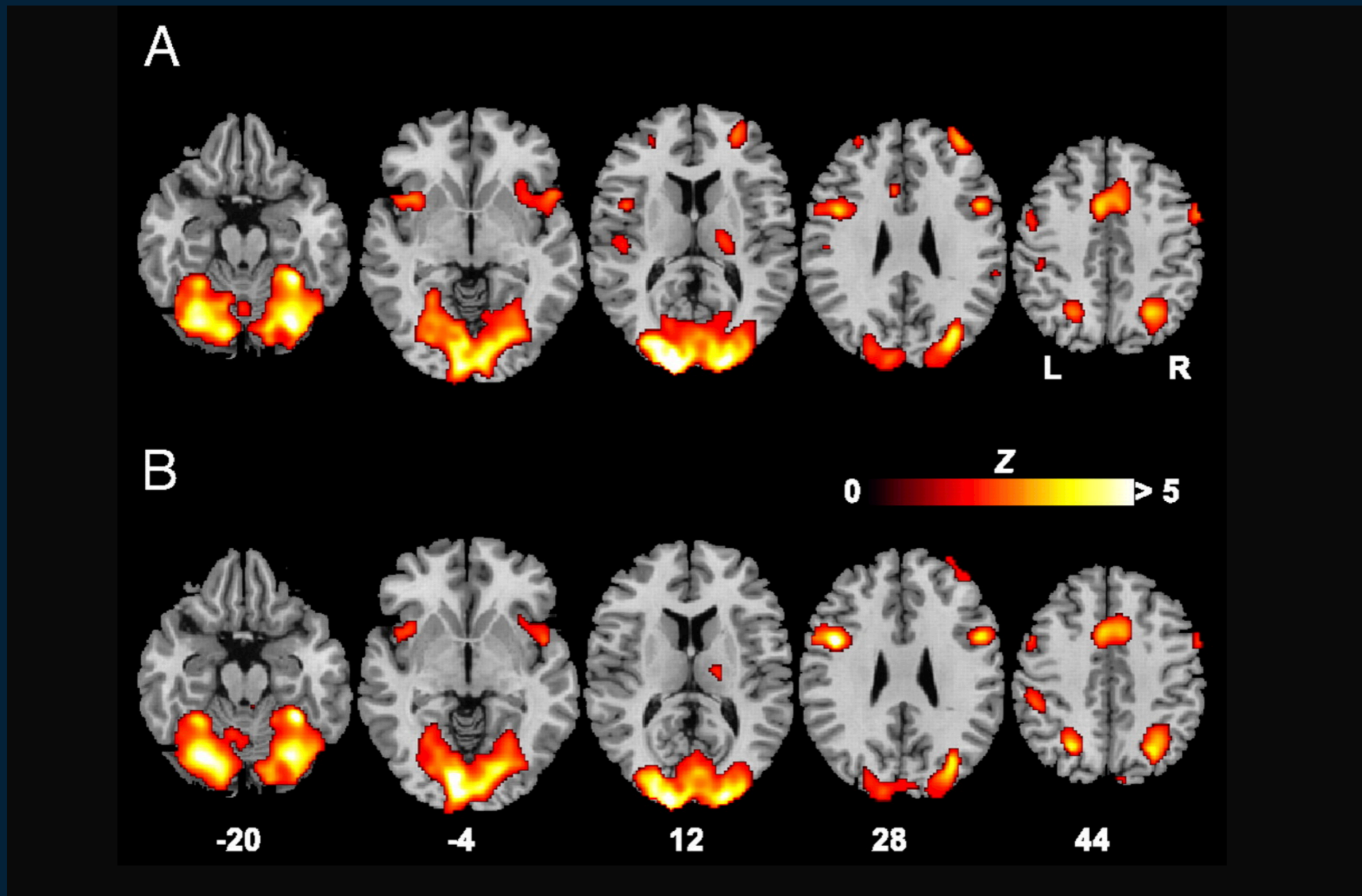
# ... in Image Processing

- 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

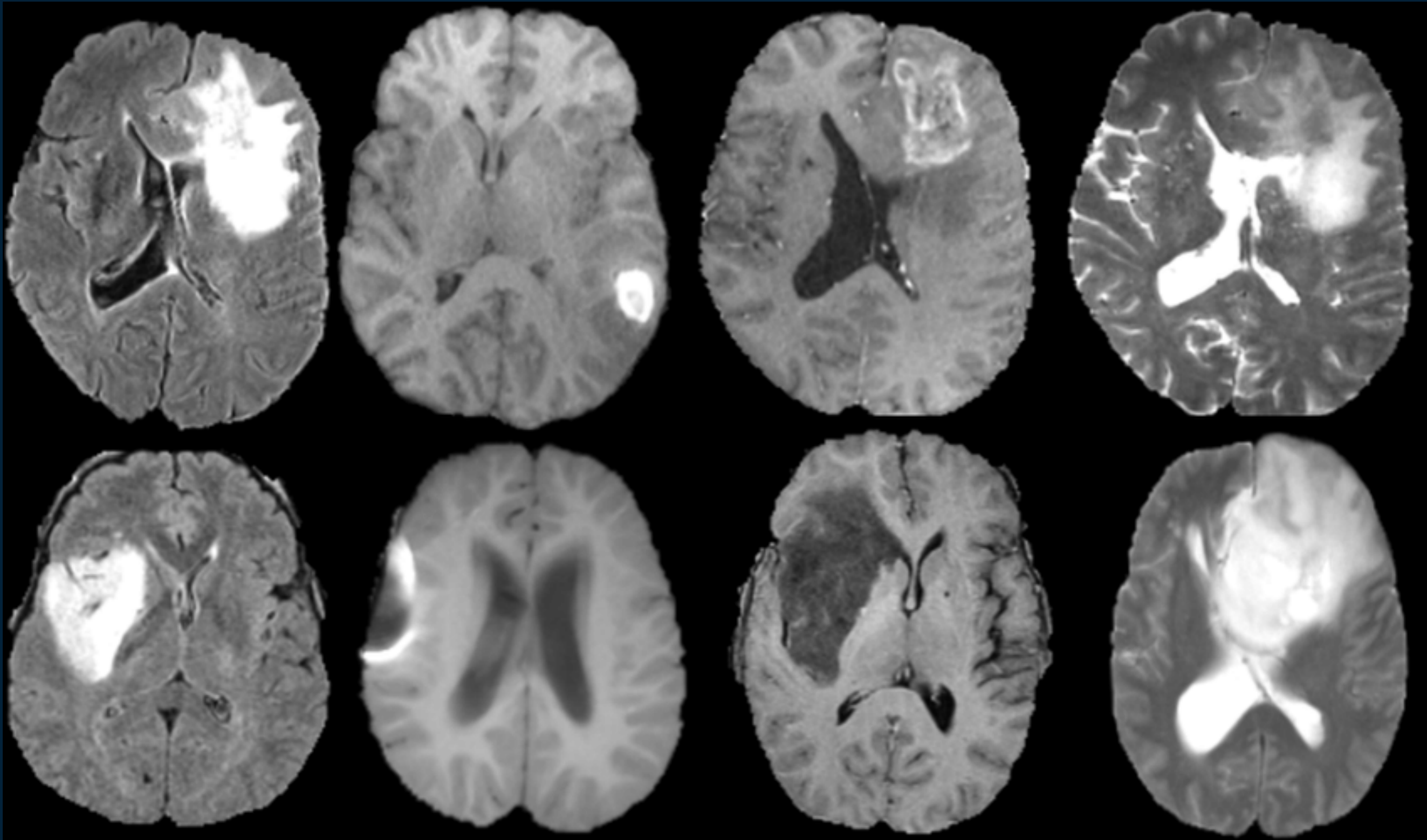


Support Devices



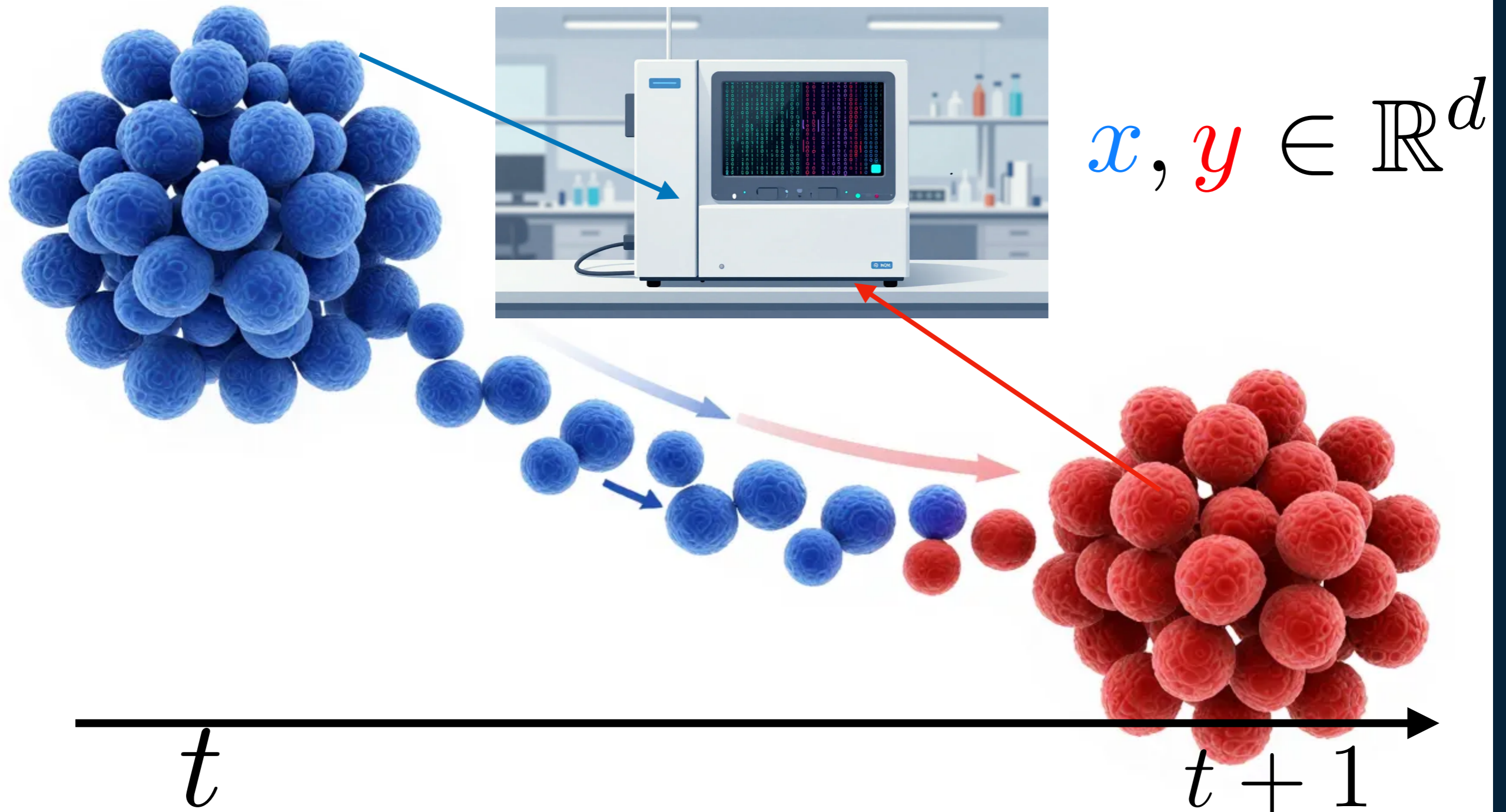
Enlarged  
Cardiomeastinum

# ... in Anomaly detection: BraTS data



# ... in Genomics

- Understanding dynamics at individual cell level.



# 2. Machine Learning: Predictive Modeling

# Training Set

 $\mathcal{X}$ 

Feature Space

 $\mathcal{Y}$ 

Label Space

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

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

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

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

used to quantify prediction quality

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

# 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

# Origin: Monge Problem (1781)



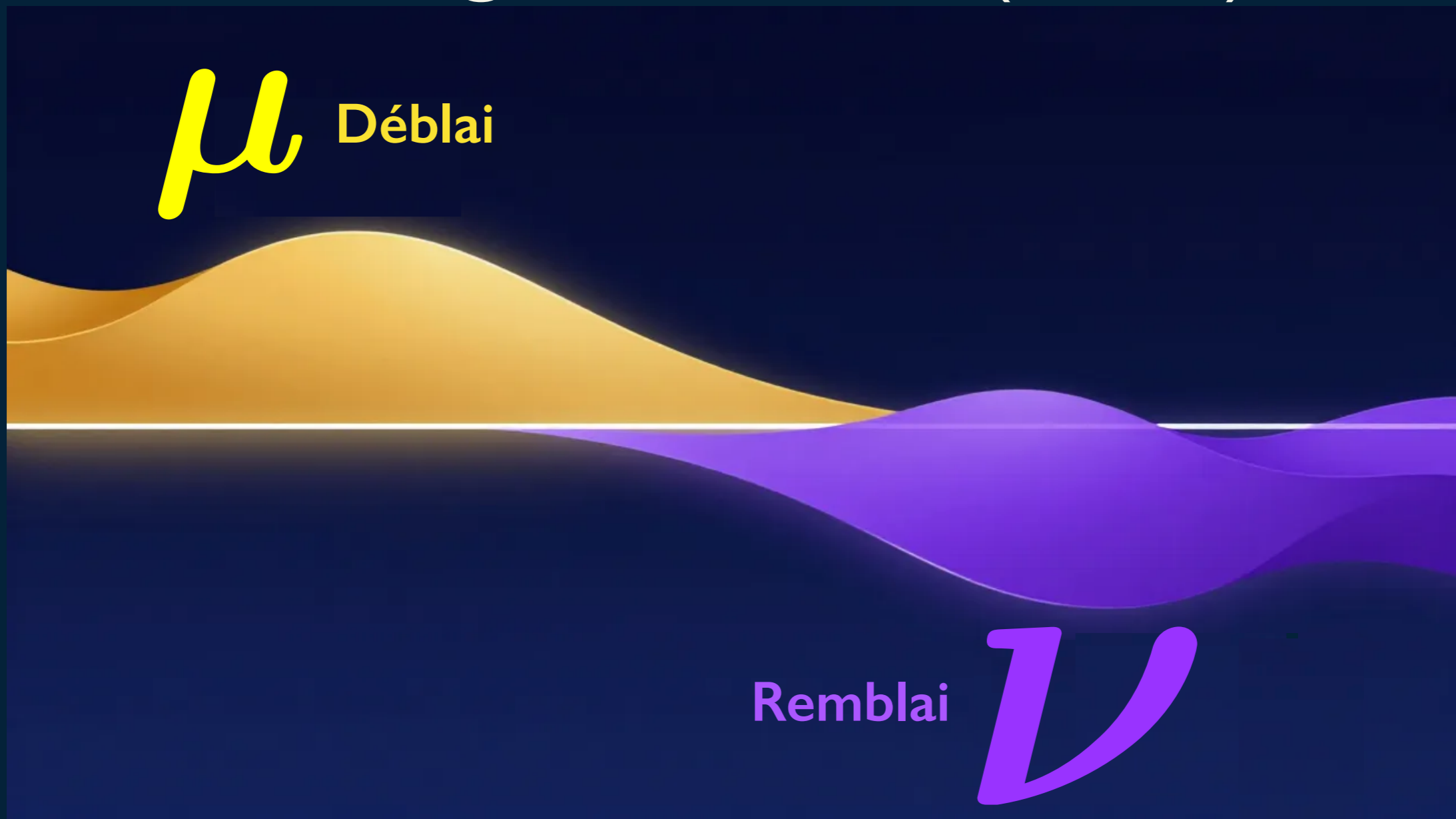
*M É M O I R E*  
*S U R L A*  
*T H É O R I E D E S D É B L A I S*  
*E T D E S R E M B L A I S.*

Par M. M O N G E.

**L**ORSQU'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'en suit 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*.

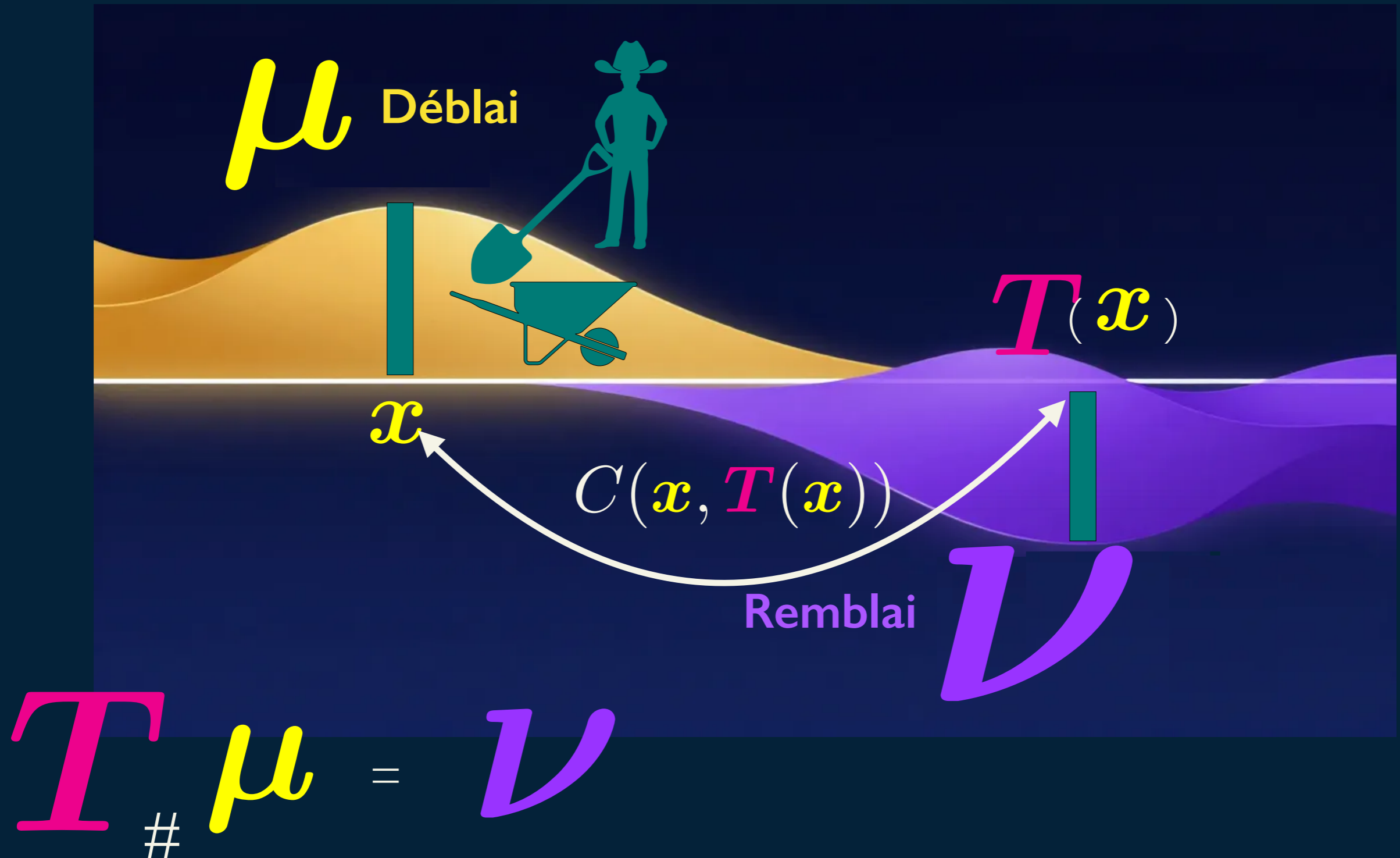
# Monge Problem (1781)



- 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**”.



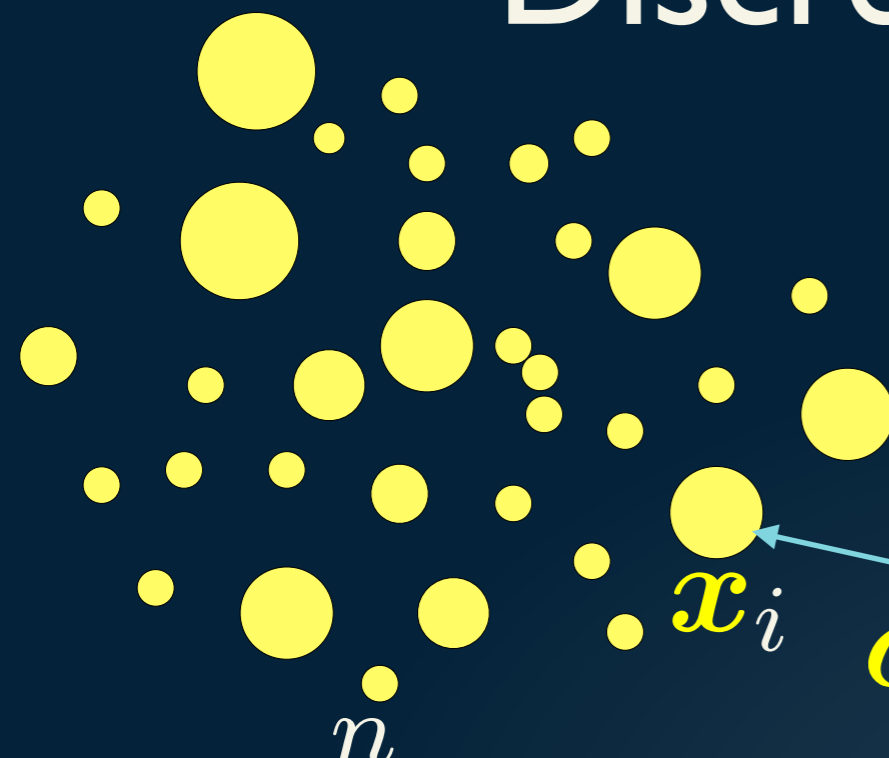
# Monge Problem (1781)

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

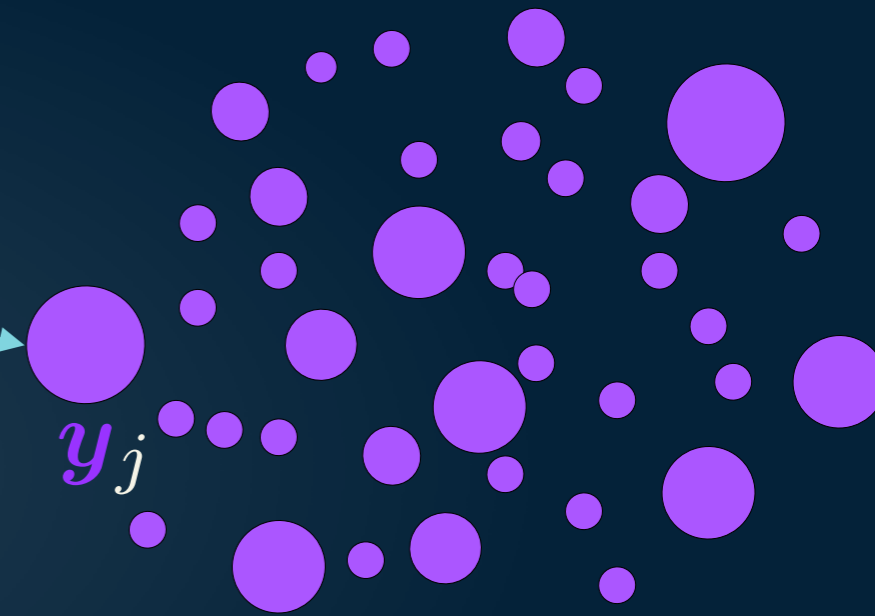
$$\inf_{T \# \mu = \nu} \int C(x, T(x)) \mu(x) dx$$

- 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



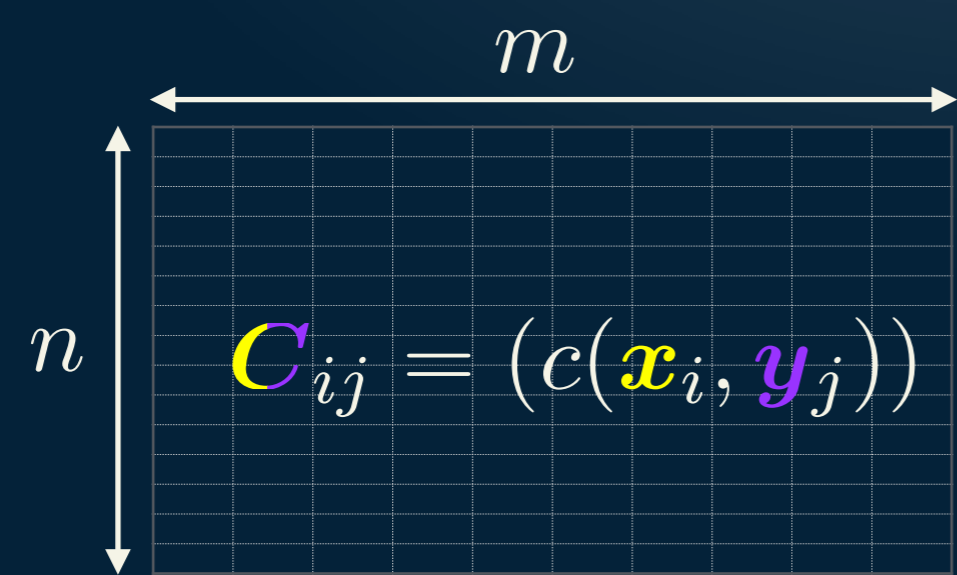
$$\nu = \sum_{j=1}^m \nu_j \delta_{y_j}$$



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

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

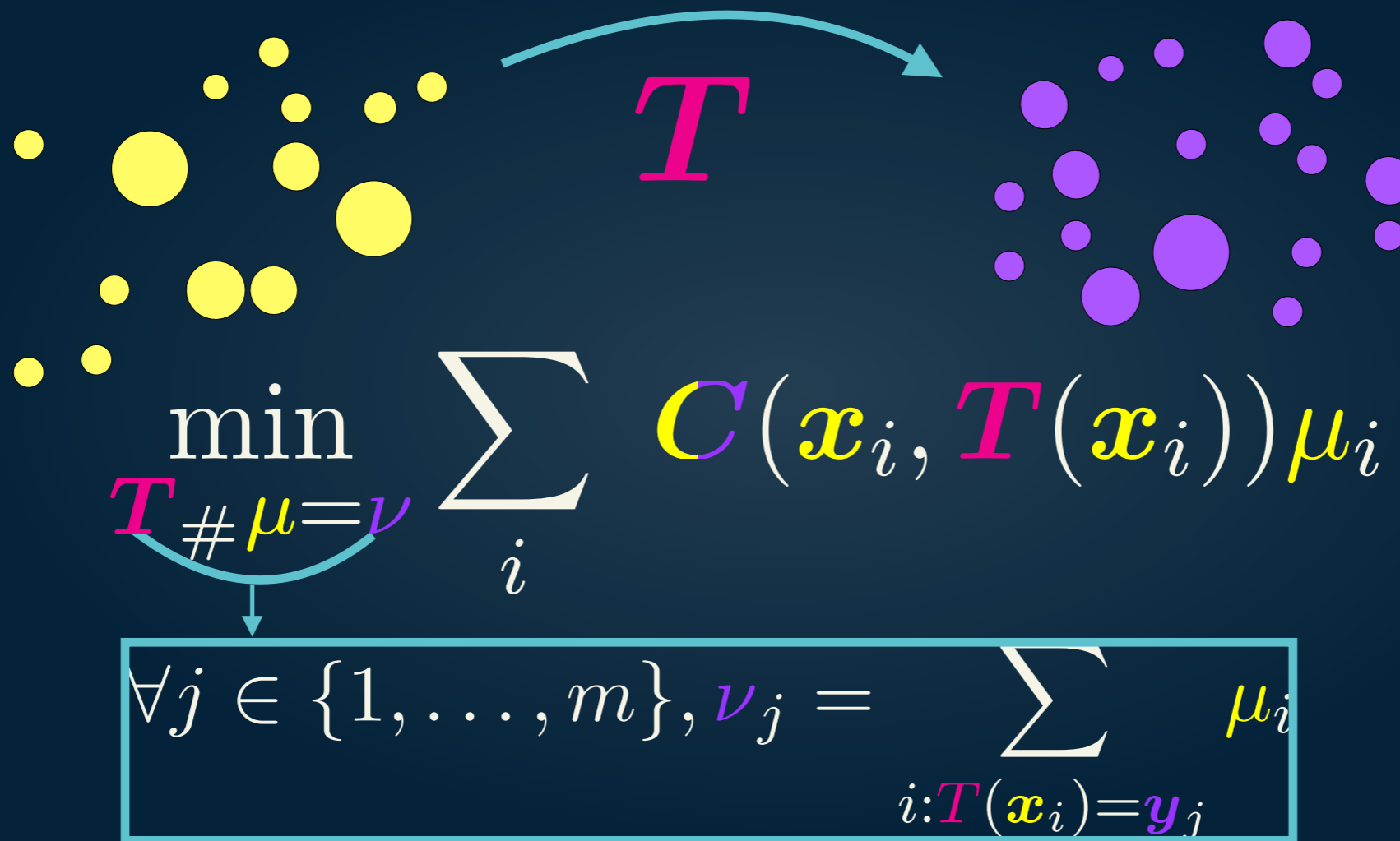
**C**  
Cost Matrix



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

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

# Discrete OT Framework: Monge's Formula



Strict: Deterministic Assignments

# Discrete OT Framework: Monge's Formula

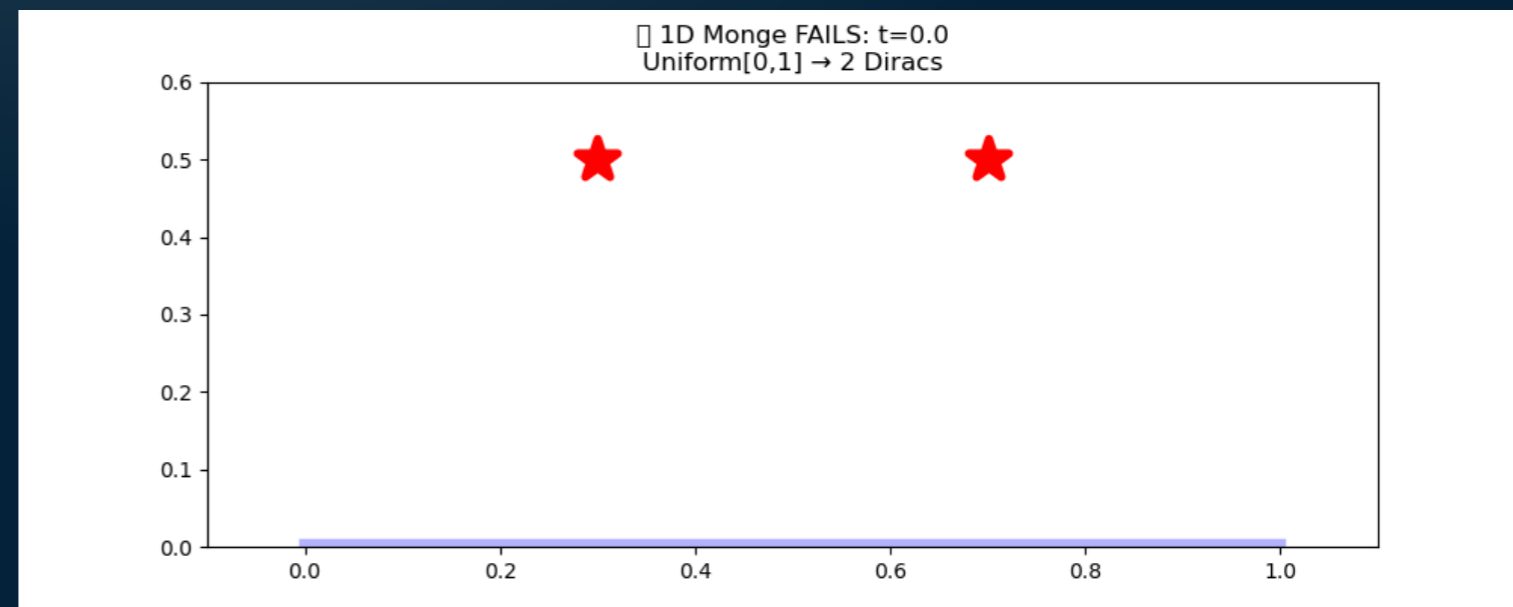
non-convex

combinatorial

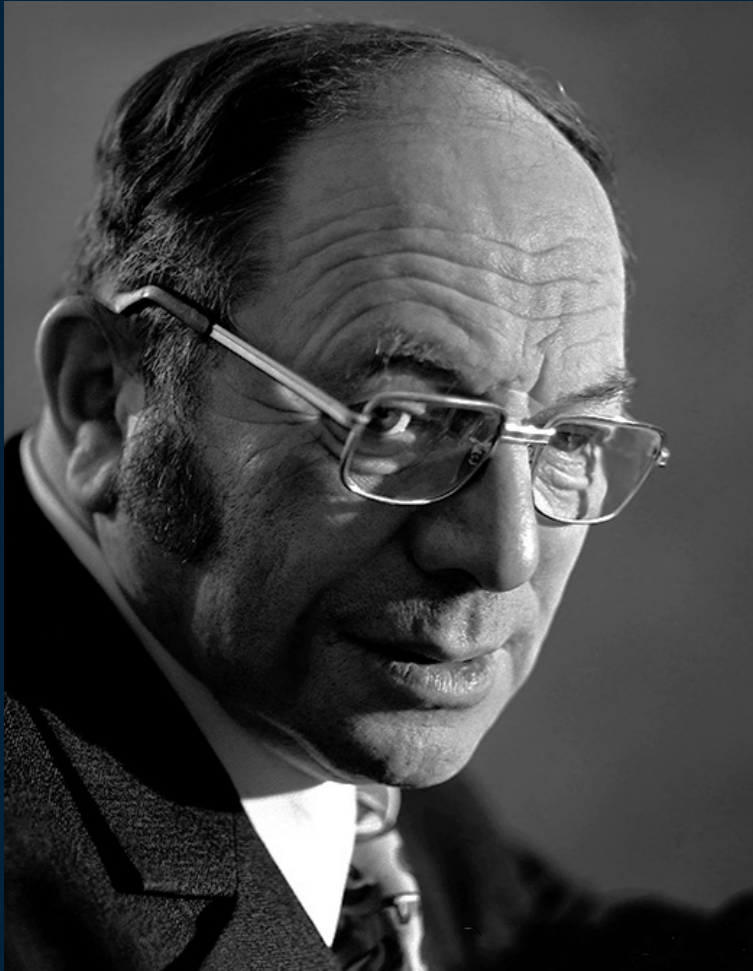
non-existent

Uniform weights

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

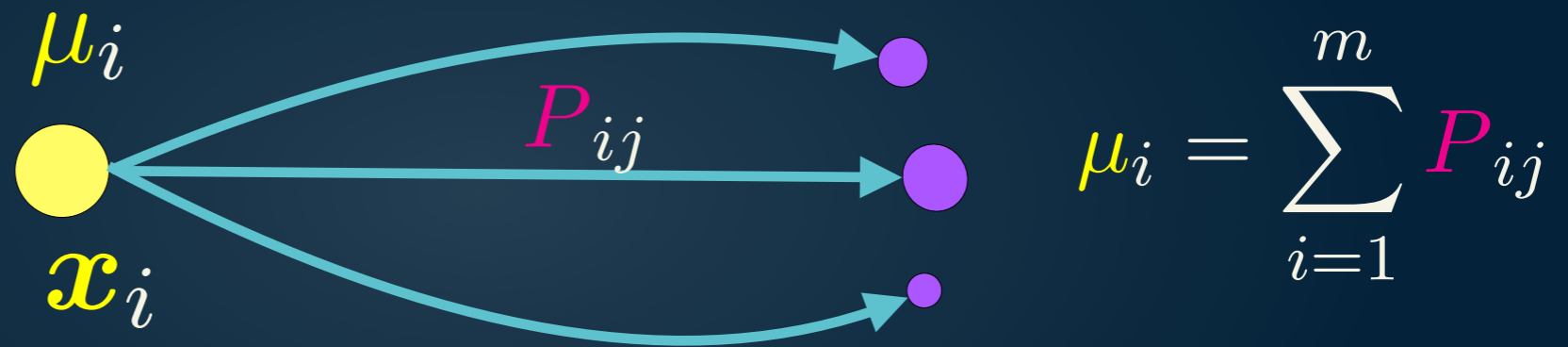


# Discrete OT Framework: Kantorovich's Formula



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  $\mu$  and  $\nu$  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

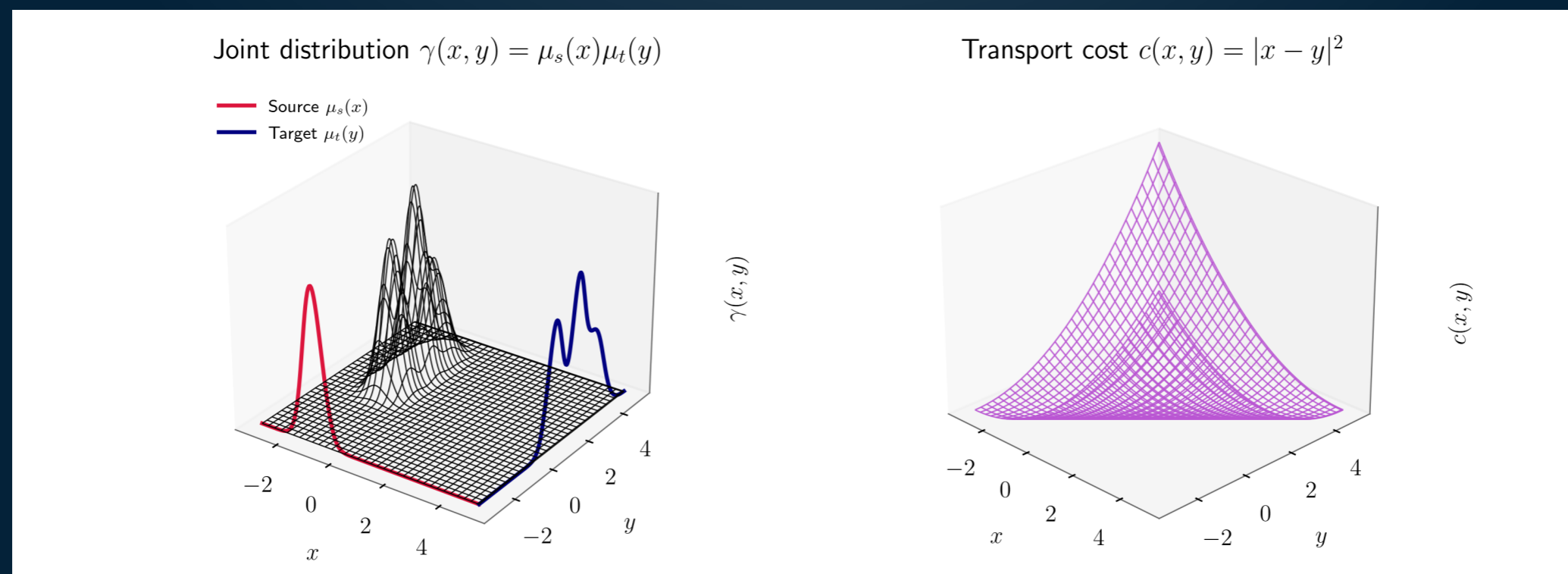
Monge-Kantorovich /  
Wasserstein Distance

# Continuous OT Framework: Kantorovich's Formula

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

Probabilistic couplings set (Transport Polytope)

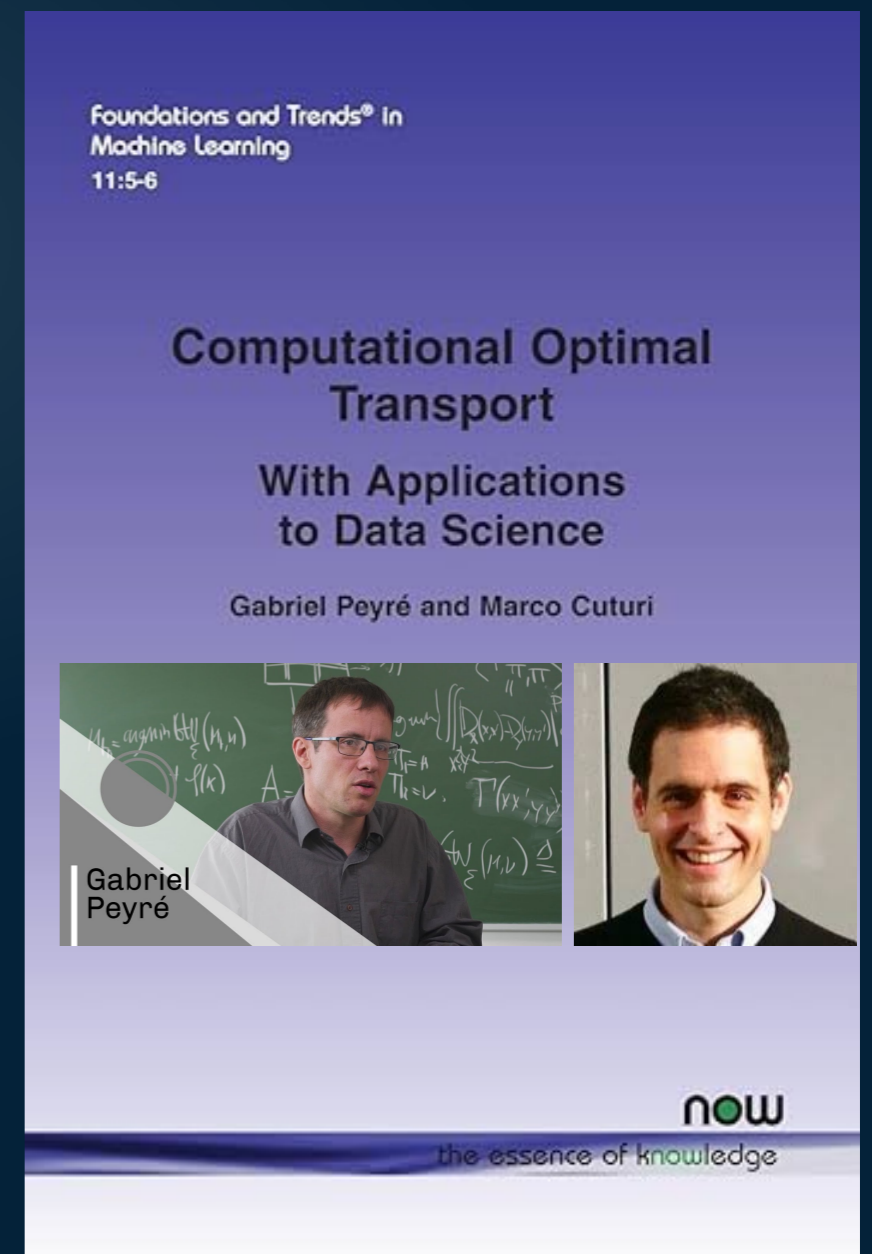
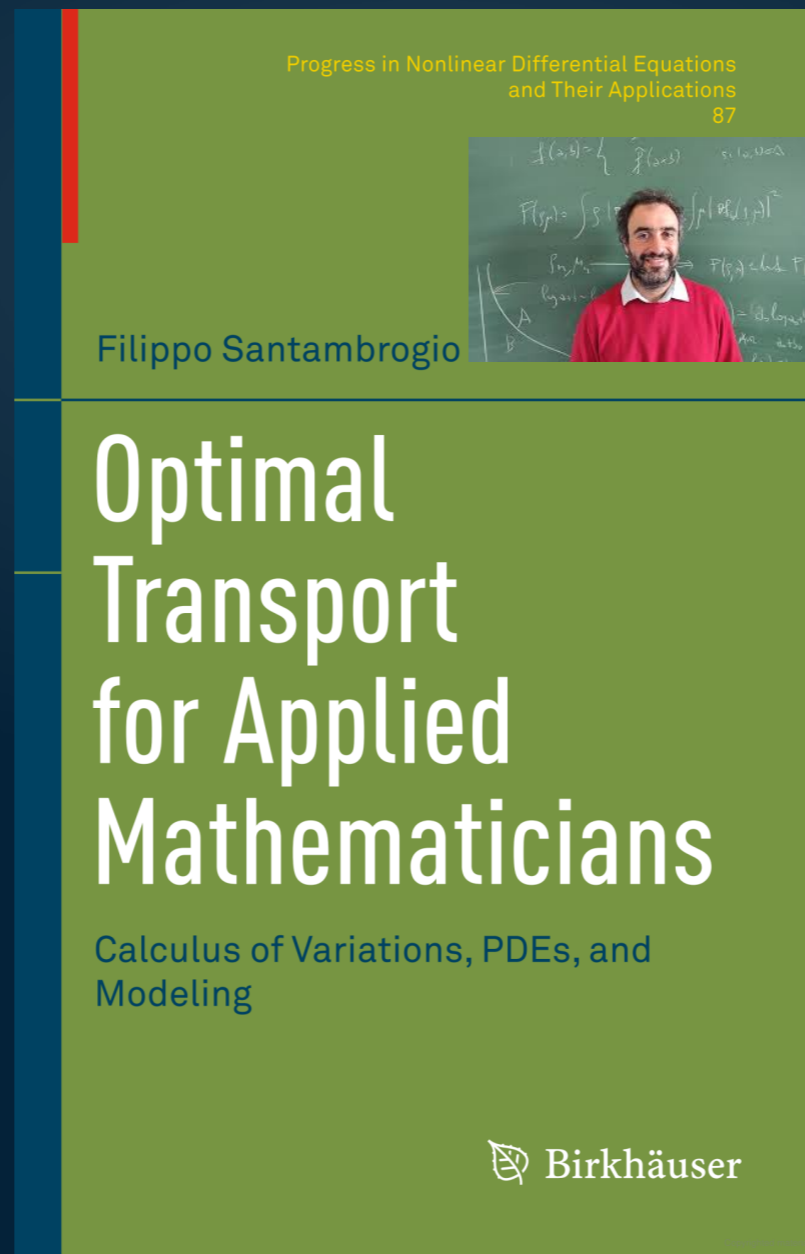
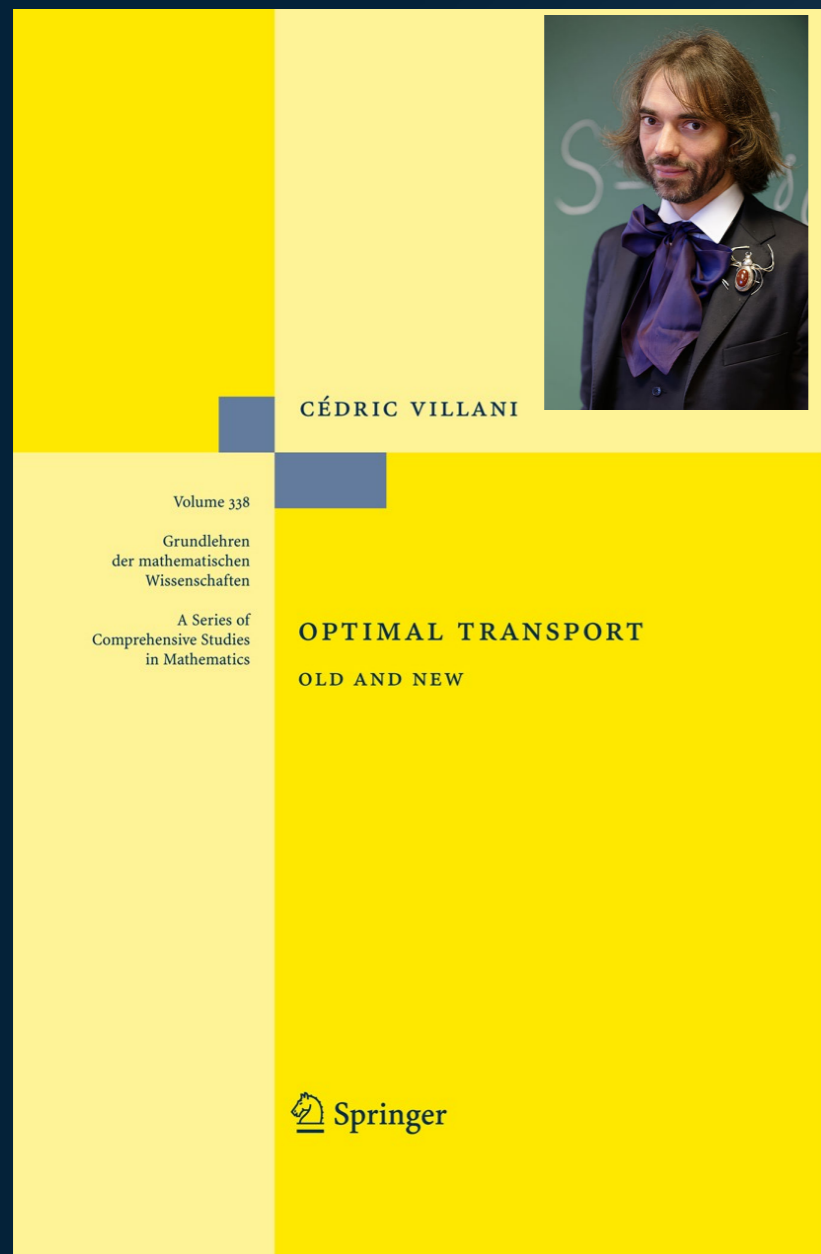
$$\Pi_{\text{con}}(\mu, \nu) = \left\{ \gamma \geq 0, \int_{\mathbb{R}^d} \gamma(x, y) dy = \mu, \int_{\mathbb{R}^d} \gamma(x, y) dx = \nu \right\}$$



# Continuous OT Framework: Wasserstein Distance

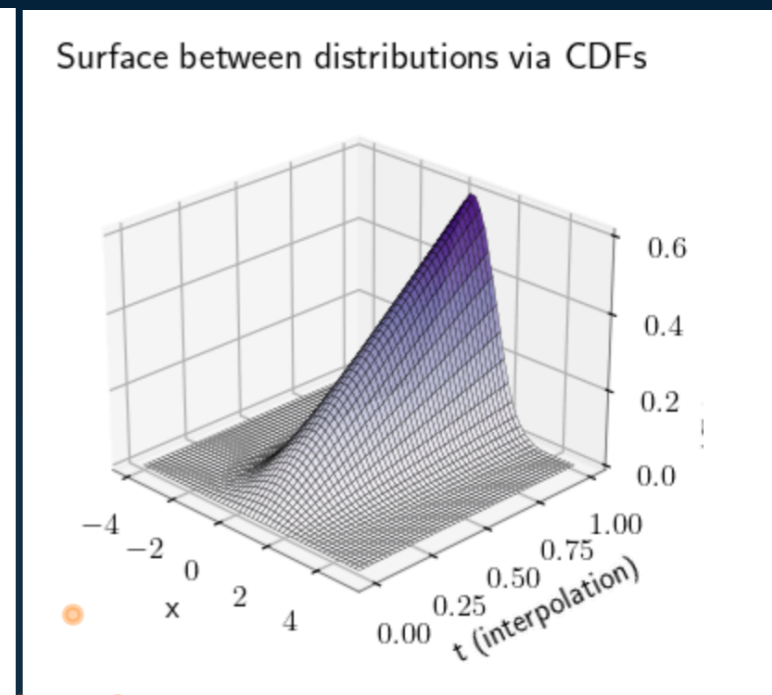
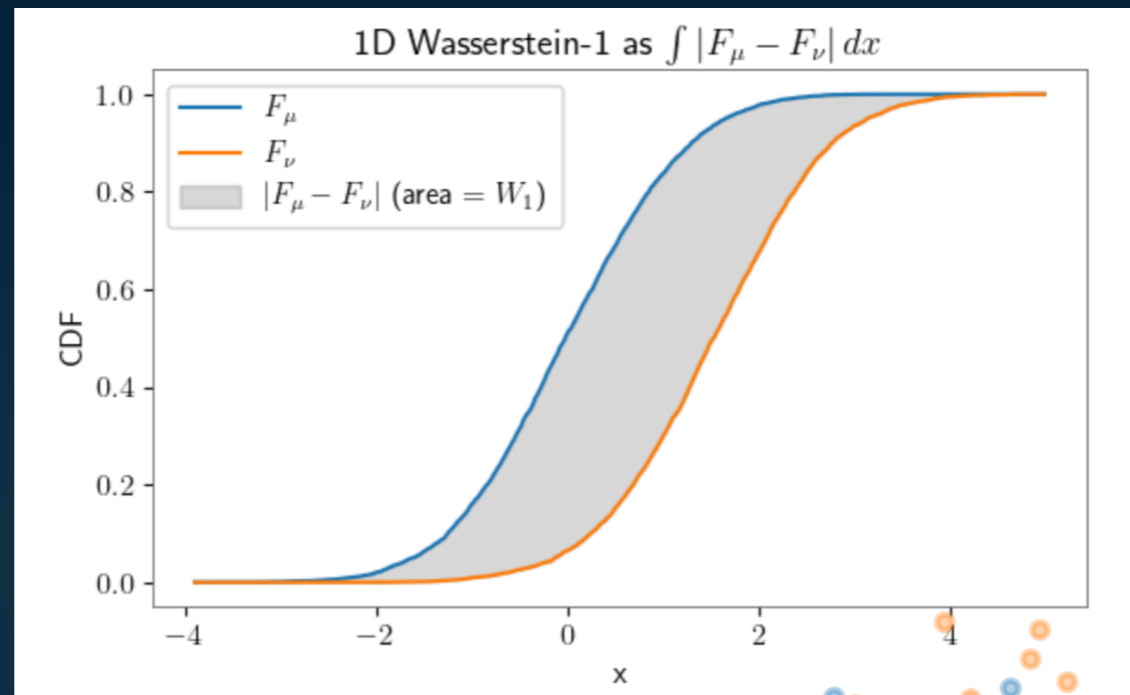
## Wasserstein Distance

$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left( \inf_{\gamma \in \Pi_{\text{con}}(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^d} D^p(\boldsymbol{x}, \boldsymbol{y}) \gamma(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \right)^{1/p} = \left( \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \gamma} [D^p(\boldsymbol{x}, \boldsymbol{y})] \right)^{1/p}$$



# Continuous OT Framework: Wasserstein Distance

$$W_1(\mu, \nu)$$



$$\mathcal{N}(m_0, \Sigma_0)$$

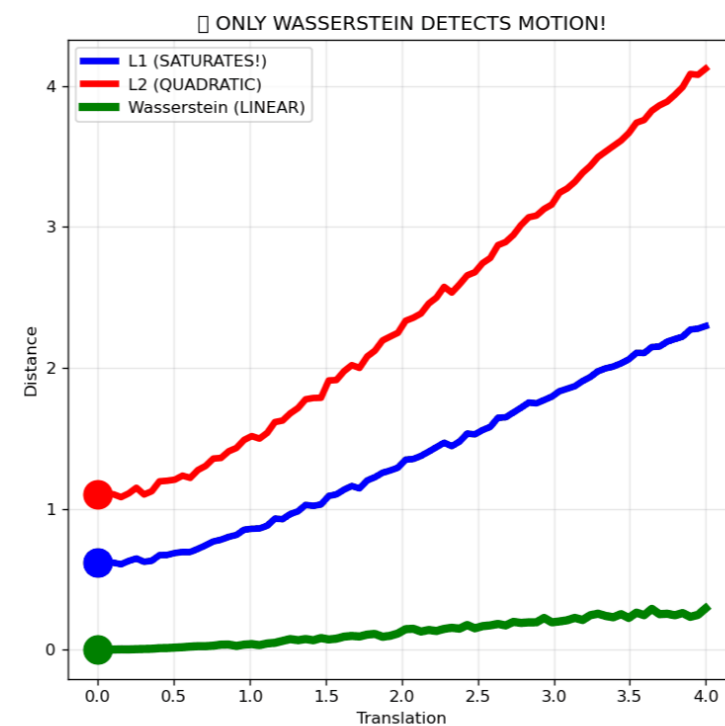
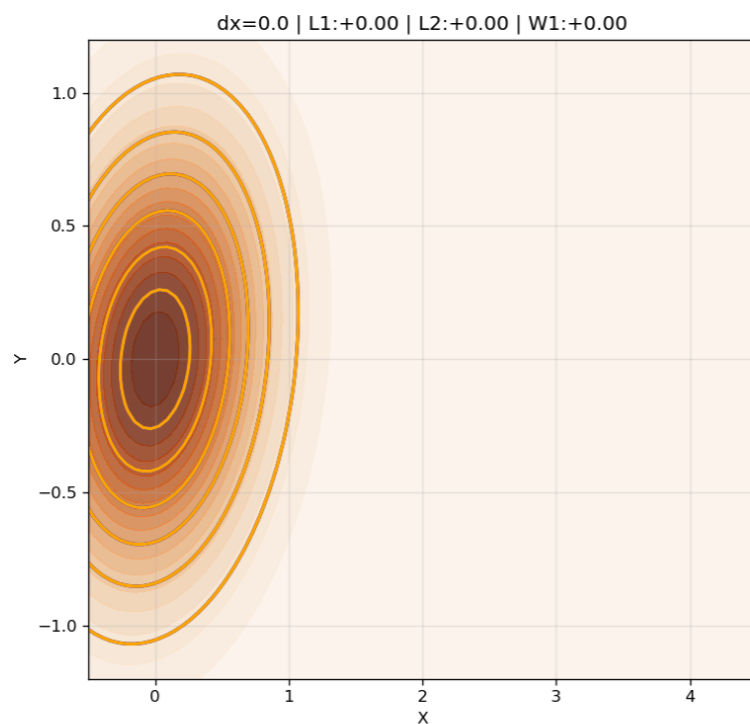
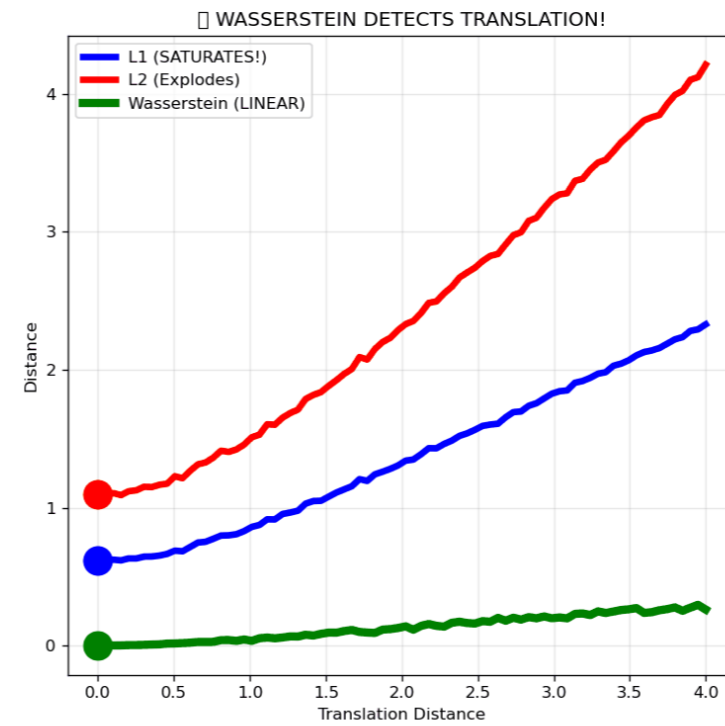
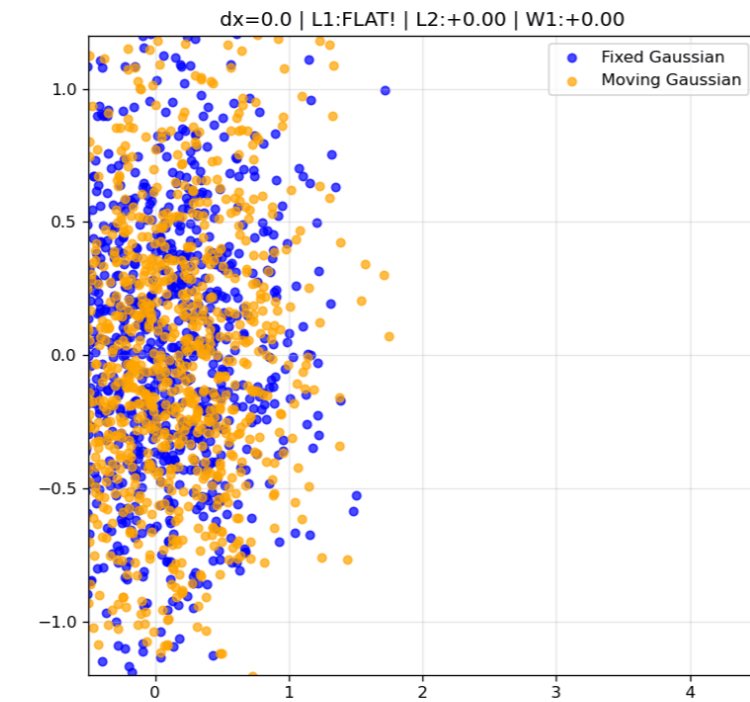
$$\mathcal{N}(m_1, \Sigma_1)$$

Bures Distance

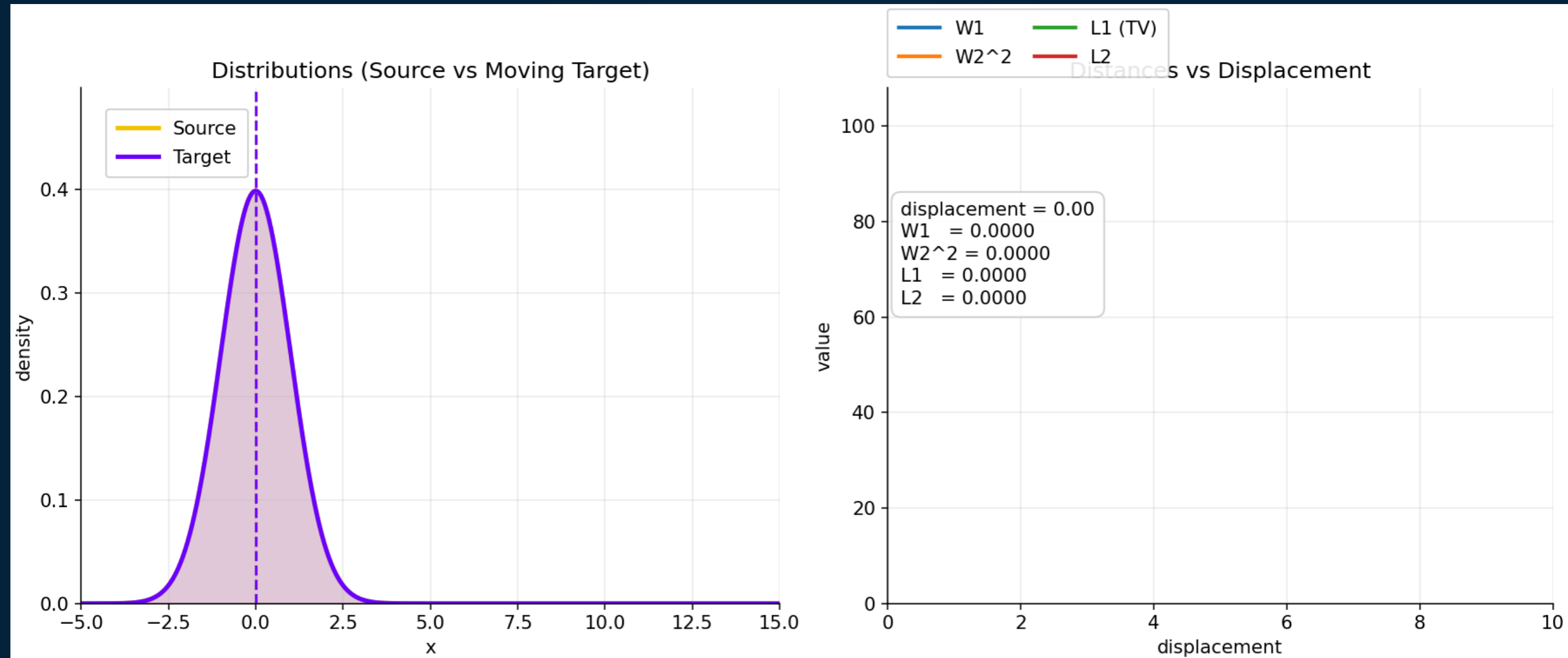
$$W_2^2(\mu, \nu)$$

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

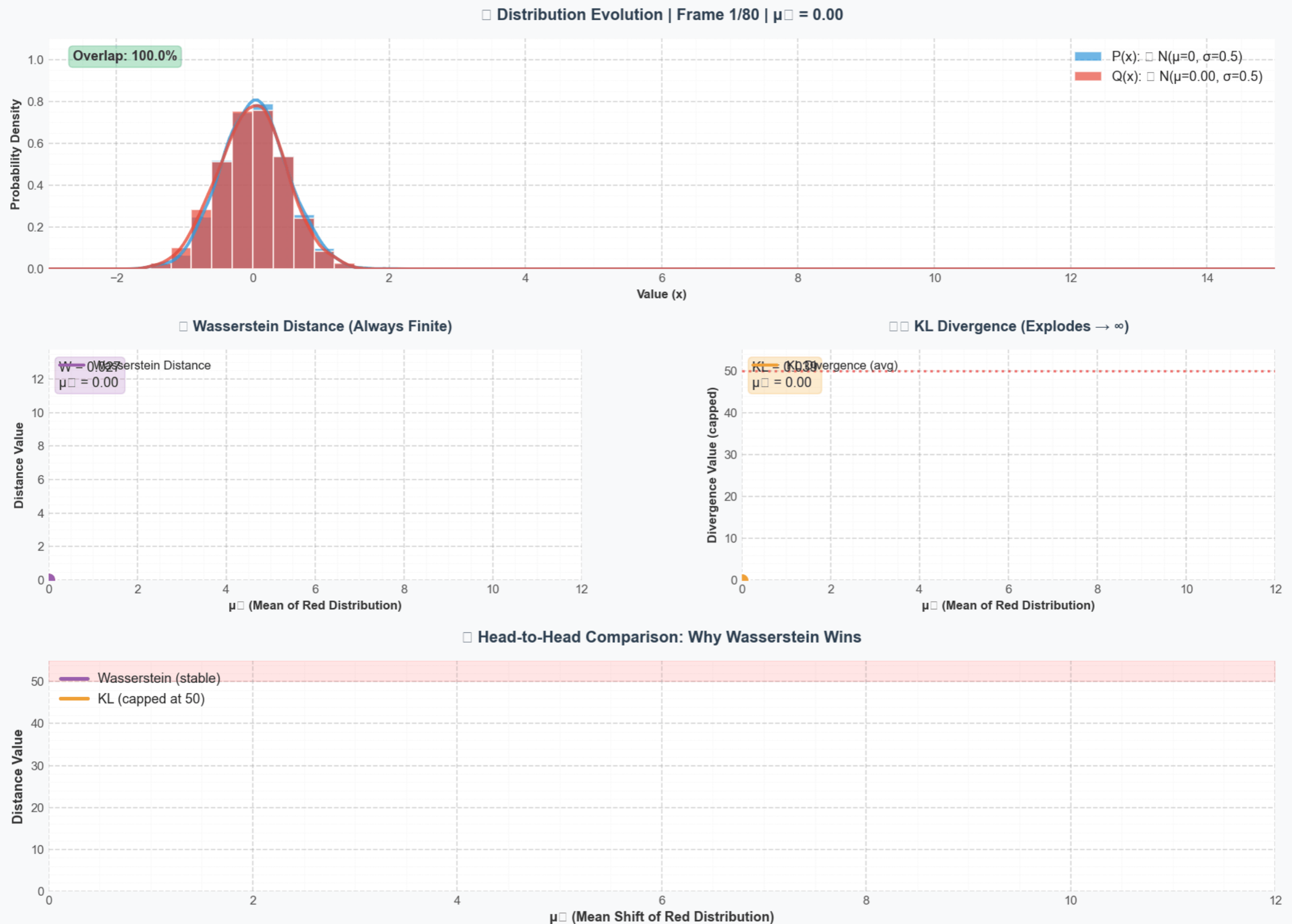
# Benefit of Wasserstein Distance (I)



# Benefit of Wasserstein Distance (2)

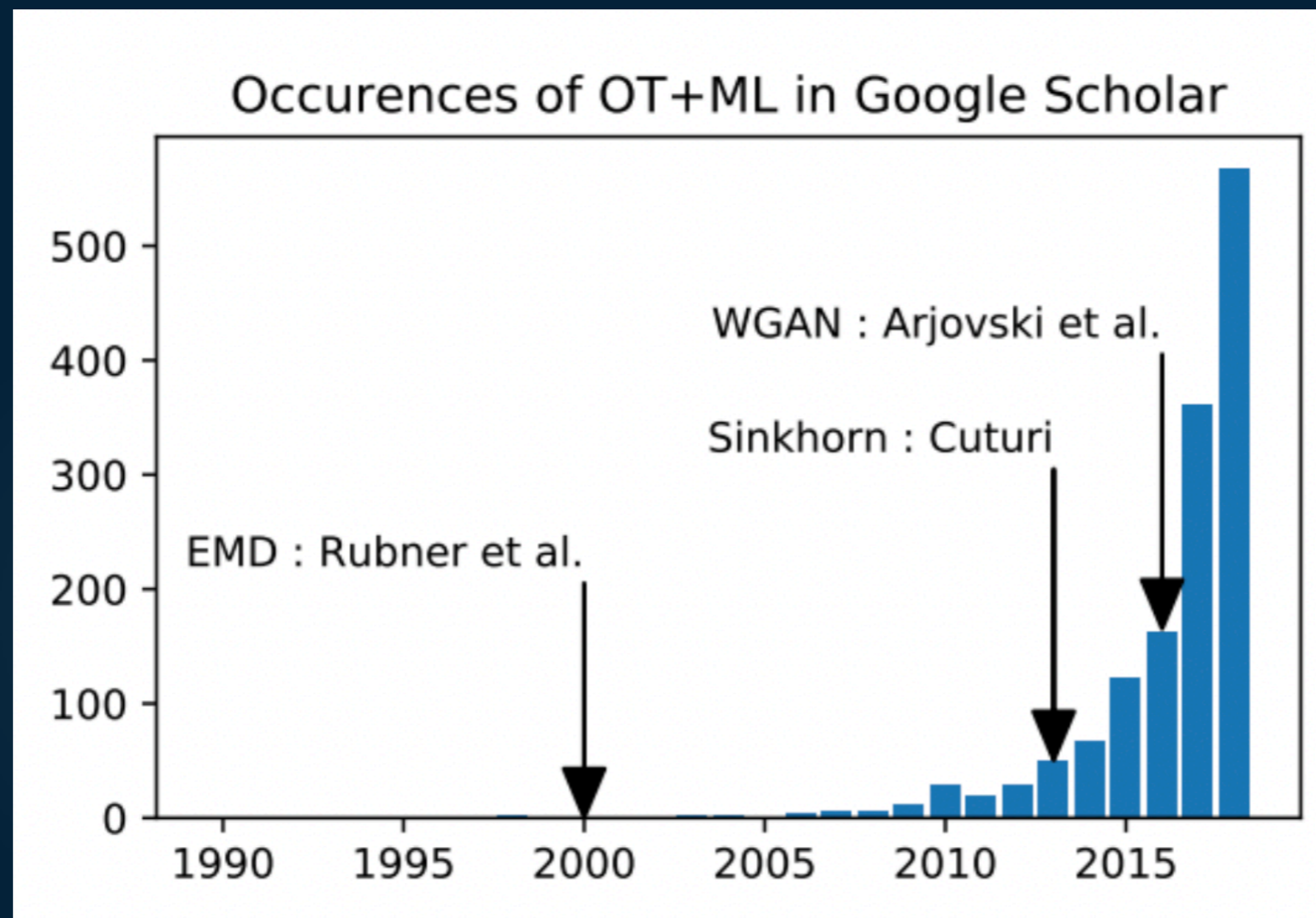


# Benefit of Wasserstein Distance (3)



4. How can it be used in data science?

# History of OT for machine learning



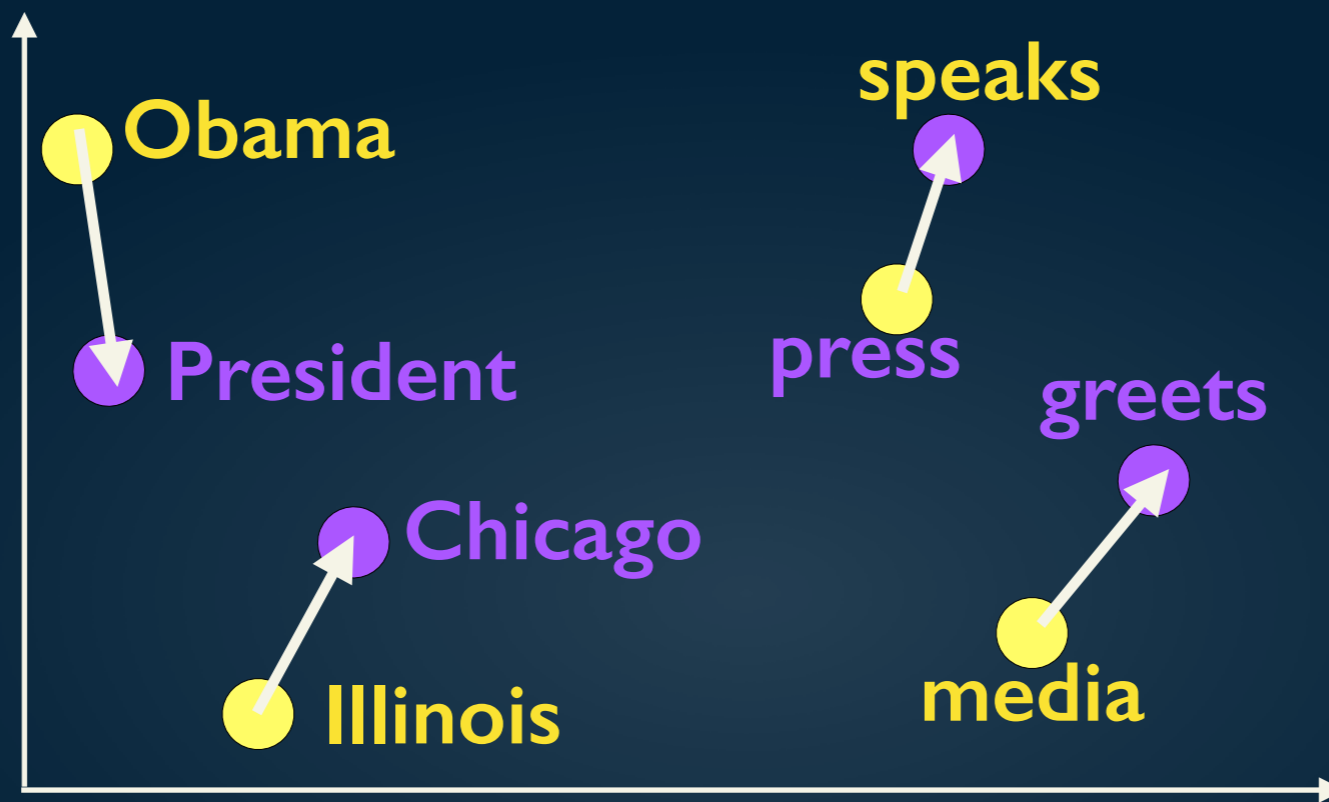
[R. Flamary, 2019 (HDR)]

- 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

Document 1

**Obama**  
**speaks**  
to  
the  
**media**  
in  
**Illinois**



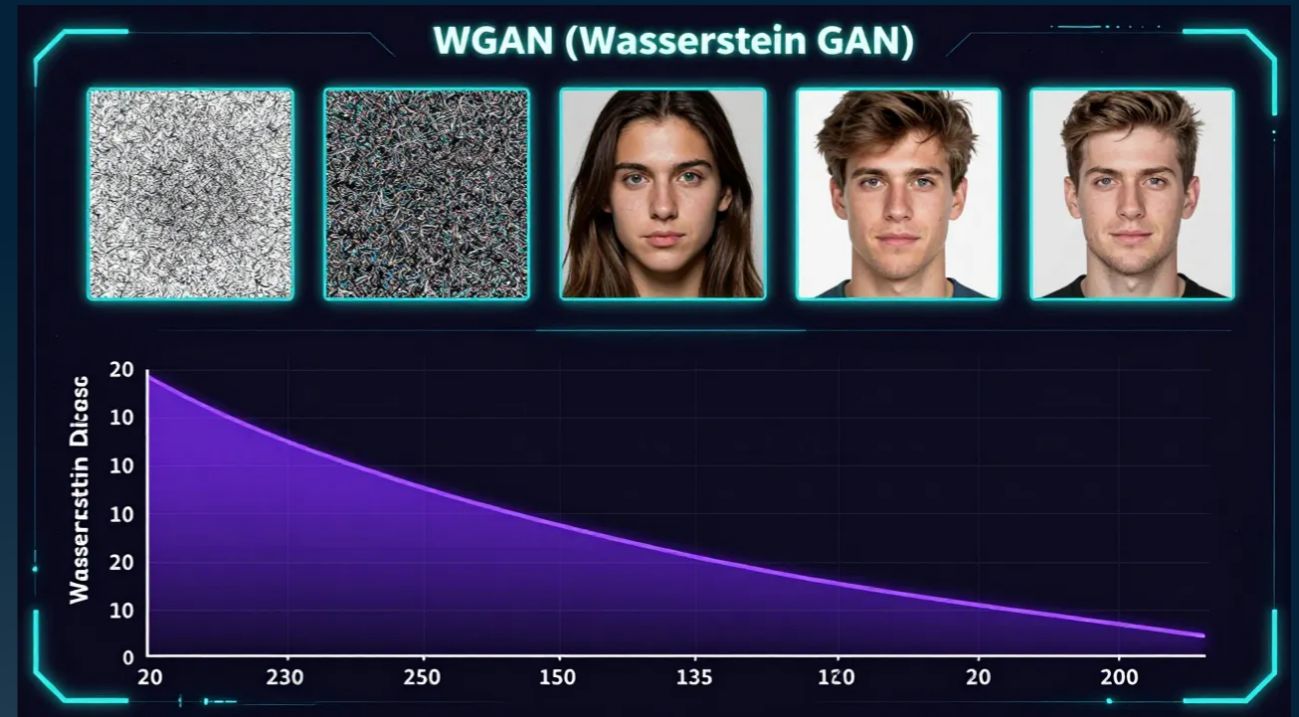
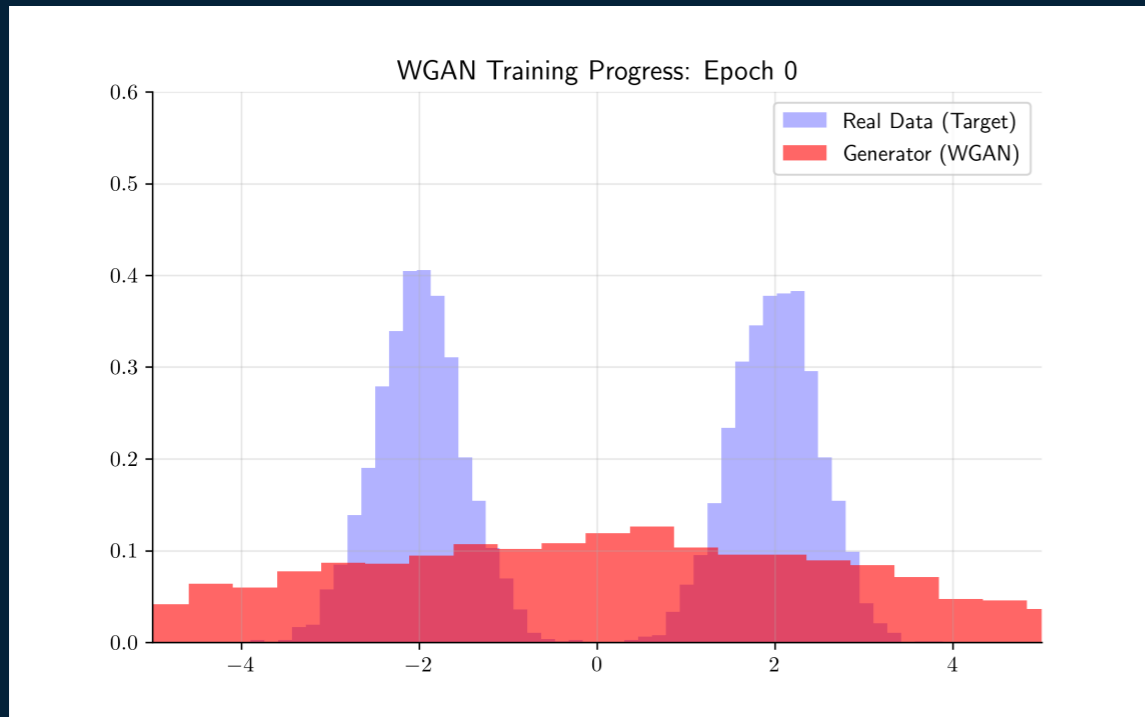
Document 2

The  
**President**  
**greet**s  
the  
**press**  
In  
**Chicago**

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



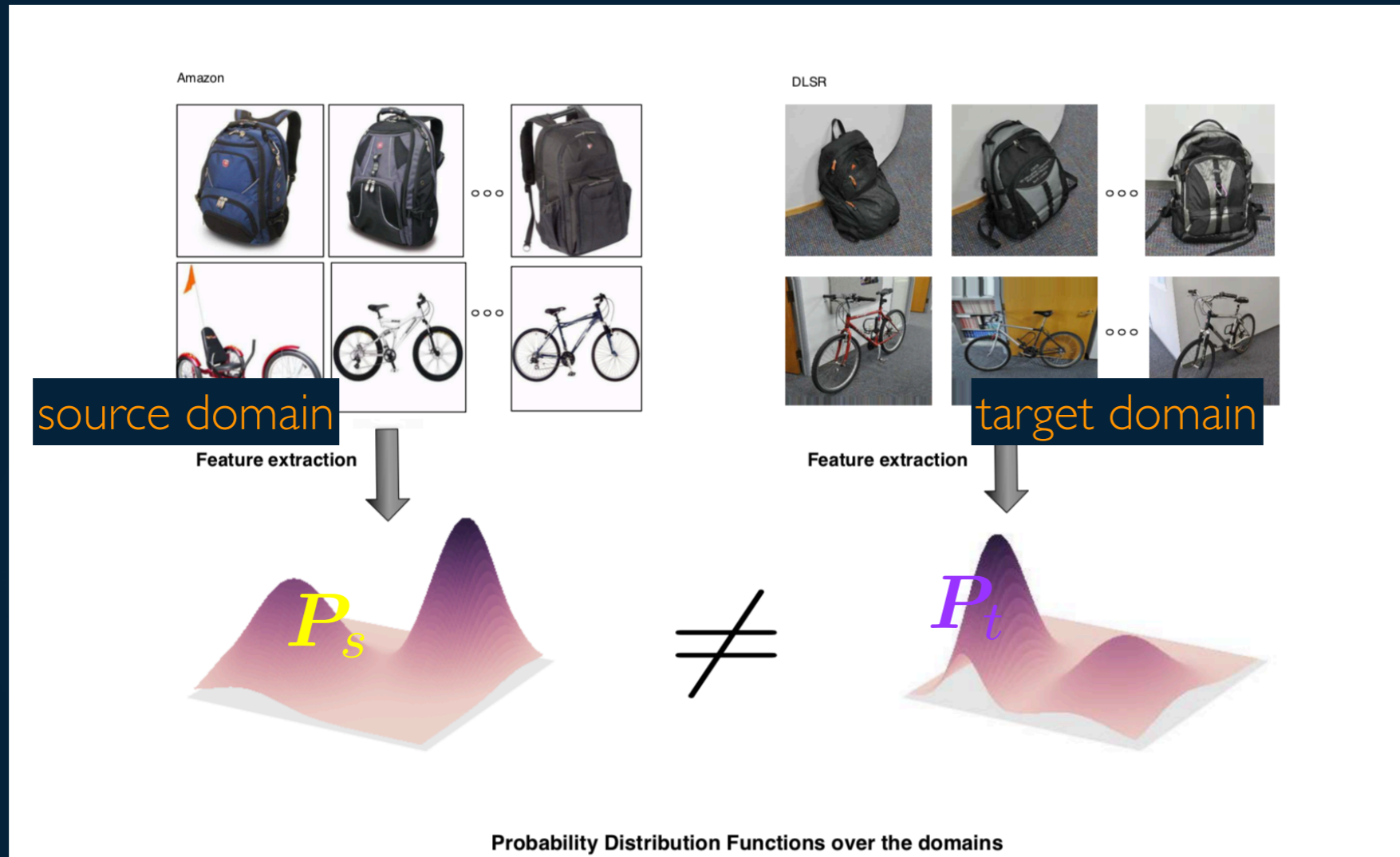
## Generative modelling as a matching distribution problem

- 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].

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

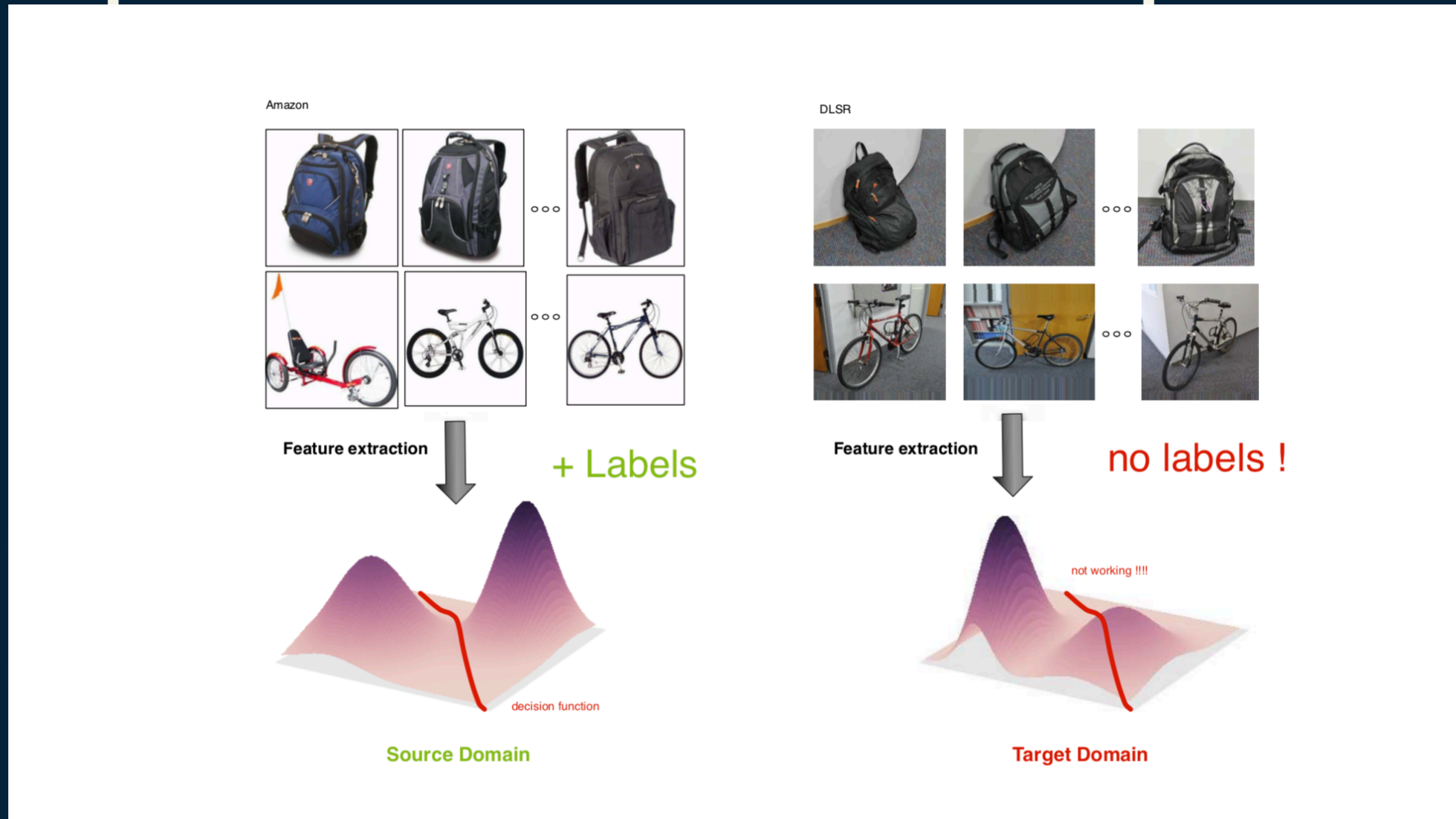
$\{z_i\}$  some random vectors,  $\{x_j\}$  some samples from the target distribution.

# 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

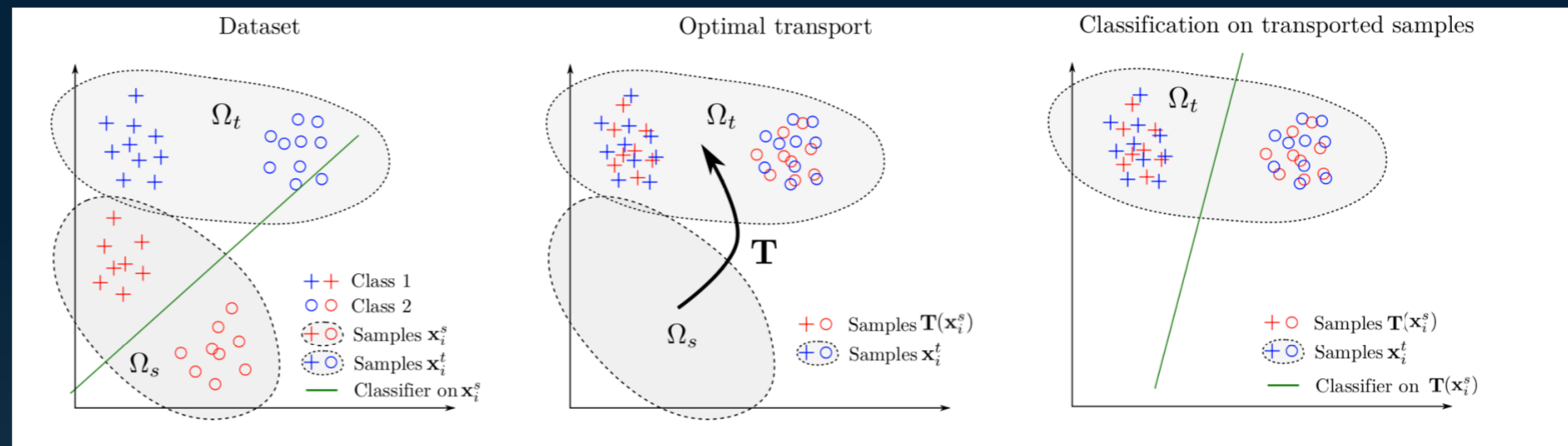
## Assumptions

- There exist a transport  $\mathbf{T}$  in the feature space between the two domains.
- The transport preserves the conditional distributions:

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

## 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.



# 5. (Smoothed, Sliced) OT

# Wasserstein Distance: Curse of Dimensionality

$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left( \inf_{\gamma \in \Pi_{\text{con}}(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^p \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \right)^{1/p}$$

- The Wasserstein distance is often estimated from samples.

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}$$

- The error of these empirical estimates suffers from an exponential dependence on dimension  $d$  that presents an obstacle to sample-efficient bounds.

$$\mathbb{E}[W_p(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu})] \lesssim n^{-1/d}$$

[Altschuler et al., 2017, Weed & Bach, 2019, Lei, 2020]

- Linear programming problem that requires generally  $\mathcal{O}(n^3 \log(n)^2)$  arithmetic operations.

# Wasserstein Distance in One-Dimension

- When  $d = 1$ , the Wasserstein distance can be calculated in closed-form owing to the cumulative distributions of  $\mu$  and  $\nu$ .
- The Wasserstein distance admits the quantile representation

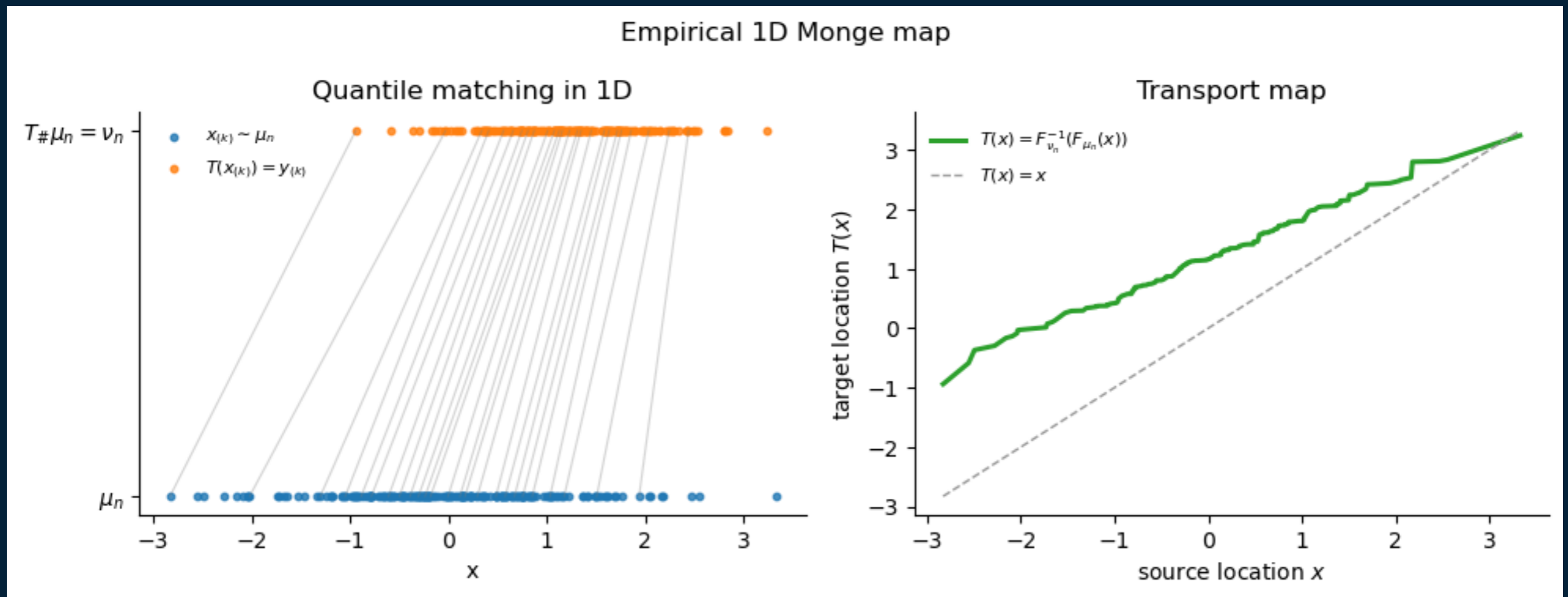
$$W_p(\mu, \nu) = \left( \int_0^1 |F_\mu^{-1}(q) - F_\nu^{-1}(q)|^p dq \right)^{1/p}$$

- For empirical measures with equal weights, this reduces to matching order statistics

$$W_p(\mu_n, \nu_n) = \left( \frac{1}{n} \sum_{k=1}^n |x_{(k)} - y_{(k)}|^p \right)^{1/p}$$

- Hence computing 1D requires only the sorting of the samples, which yields a closed-form formula and an  $\mathcal{O}(n \log(n))$ .

# Wasserstein Distance in One-Dimension



- To derive a metric for high-dimensional distributions based on one-dimensional Wasserstein distance.
- The main idea is to **project high-dimensional probability distributions onto a random one-dimensional space** and then to **compute the Wasserstein distance**.

# Sliced-Wasserstein Distance



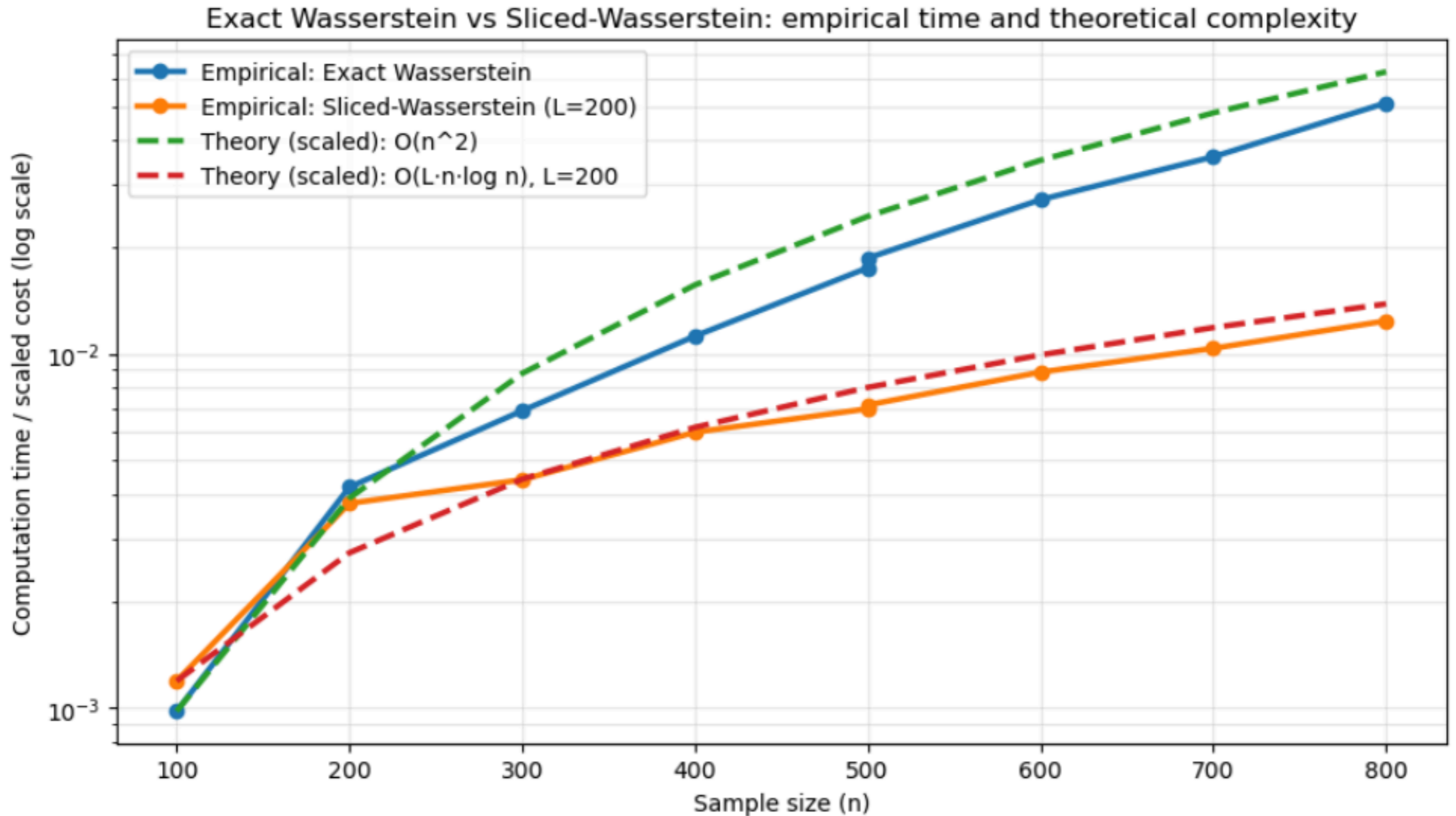
- The sliced Wasserstein distance (SW) reads as

$$SW_p(\mu, \nu) = \left( \int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}_u \mu, \mathcal{R}_u \nu) u_d(\mathbf{u}) d\mathbf{u} \right)^{1/p}.$$

Where  $\mathcal{R}_u$  the Radon transform of a probability distribution, i.e.,

$$\mathcal{R}_u \mu(\cdot) = \int_{\mathbb{R}^d} \mu(\mathbf{s}) \delta(\cdot - \mathbf{s}^\top \mathbf{u}) d\mathbf{s}$$

# Sliced-Wasserstein Distance

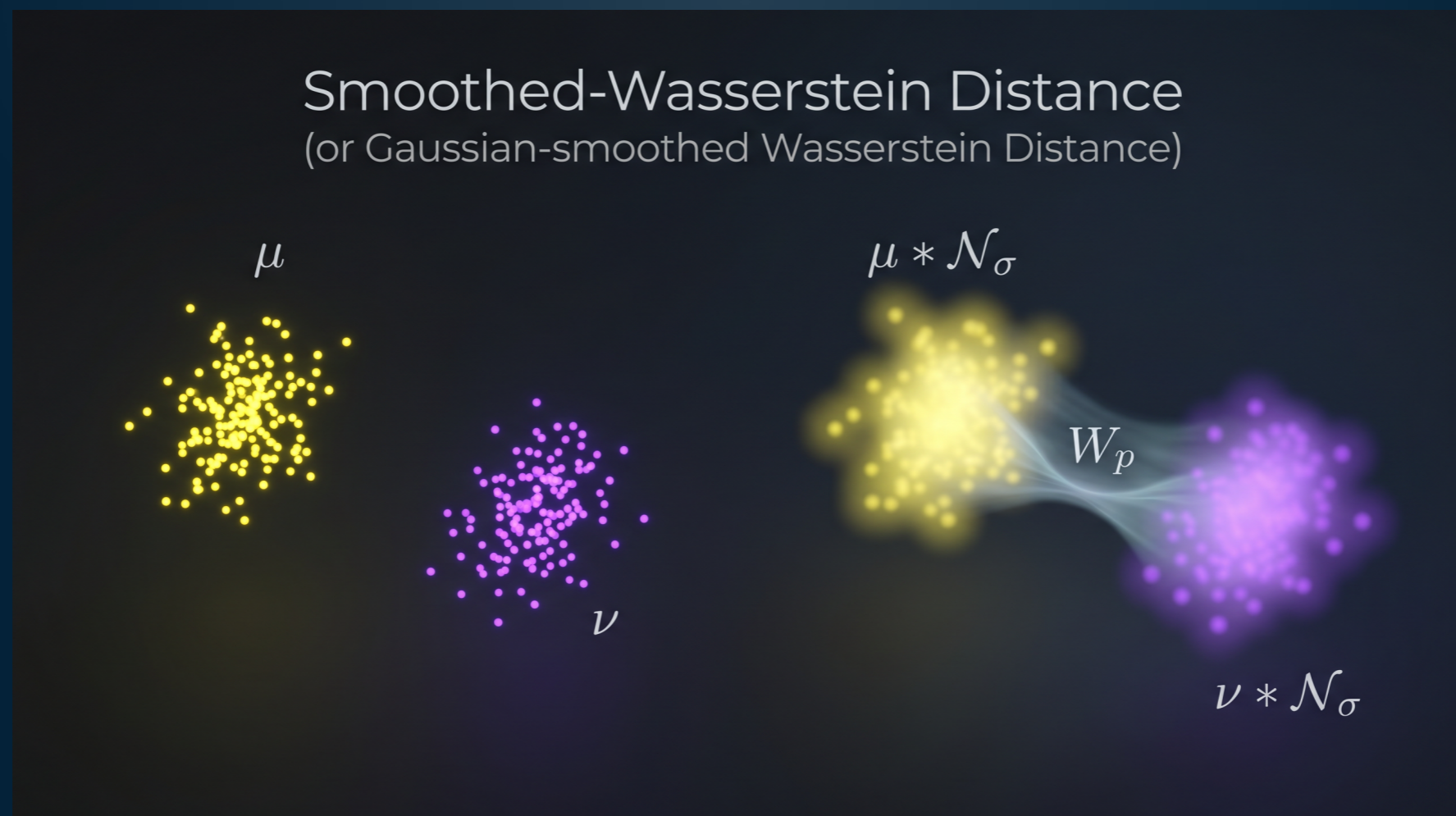


Complexity of SW

$$\mathcal{O}(Ldn + Ln \log n)$$

# Wasserstein for Privacy-Sensitive Data: Smoothed-Wasserstein Distance

- We cannot expose raw samples (medical images, mobility traces, financial records) to define or optimize our objective.
- Instead, we convolve each empirical measure with a **Gaussian**: this hides individual points inside a local cloud of noise.
- We then compare these **blurred** distributions using a smoothed Wasserstein distance [Goldfeld et al., 2020; Goldfeld & Greenewald, 2020].



# Smoothed-Wasserstein Distance

- The  $\sigma$ -smooth  $p$ -Wasserstein distance between probability measures is defined as

$$G_{\sigma} W_p(\mu, \nu) = W_p(\mu * \mathcal{N}_{\sigma}, \nu * \mathcal{N}_{\sigma})$$

- Fast rate of convergence [Nieter et al., 2021]

$$\mathbb{E}[G_{\sigma} W_p(\hat{\mu}_n, \mu)] \lesssim n^{-1/2}$$

- We investigate the theoretical properties of the Gaussian smoothed sliced Wasserstein as well as those of generalized versions denoted as **Gaussian-Smoothed Sliced Divergences**.

$$G_{\sigma} SD_p(\mu, \nu)$$

# Gaussian-Smoothed Sliced Probability Divergences Distance $G_{\sigma}SD_p(\mu, \nu)$

Published in Transactions on Machine Learning Research (11/2024)



## Gaussian-Smoothed Sliced Probability Divergences

**Mokhtar Z. Alaya**

*Université de Technologie de Compiègne,*

*LMAC (Laboratoire de Mathématiques Appliquées de Compiègne), CS 60 319 - 60 203 Compiègne Cedex*

*alayaelm@utc.fr*

**Alain Rakotomamonjy**

*Criteo AI Lab, Paris, France,*

*a.rakotomamonjy@criteo.com*

**Maxime Berar**

*Univ Rouen Normandie, INSA Rouen Normandie, Université Le Havre Normandie*

*Normandie Univ, LITIS UR4108, Rouen, France*

*maxime.berar@univ-rouen.fr*

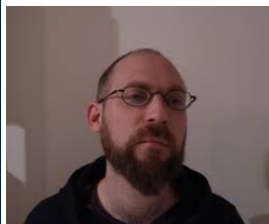
**Gilles Gasso**

*INSA Rouen Normandie, Univ Rouen Normandie, Université Le Havre Normandie,*

*Normandie Univ, LITIS UR4108, Rouen, France*

*gilles.gasso@insa-rouen.fr*

Reviewed on OpenReview: <https://openreview.net/forum?id=weuALLWUV2>



# $G_\sigma \text{SD}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$

## Gaussian-Smoothed Sliced Divergence

The  $\sigma$ -Gaussian-smooth  $p$ -Sliced Divergence between probability measures is defined as

$$G_\sigma \text{SD}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left( \int_{\mathbb{S}^{d-1}} D^p(\mathcal{R}_u \boldsymbol{\mu} * \mathcal{N}_\sigma, \mathcal{R}_u \boldsymbol{\nu} * \mathcal{N}_\sigma) u_d(\mathbf{u}) d\mathbf{u} \right)^{1/p}.$$

- Typical relevant divergences: Sinkhorn divergence or maximum mean discrepancy (MMD)

# $G_\sigma \text{SD}_p(\mu, \nu)$ : Sinkhorn Divergence

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

$$\mathcal{S}_\eta(\mu, \nu) = \min_{P \in \Pi(\mu, \nu)} \{ \langle C, P \rangle - \eta H(P) \} \quad \mathcal{O}(n^2)$$

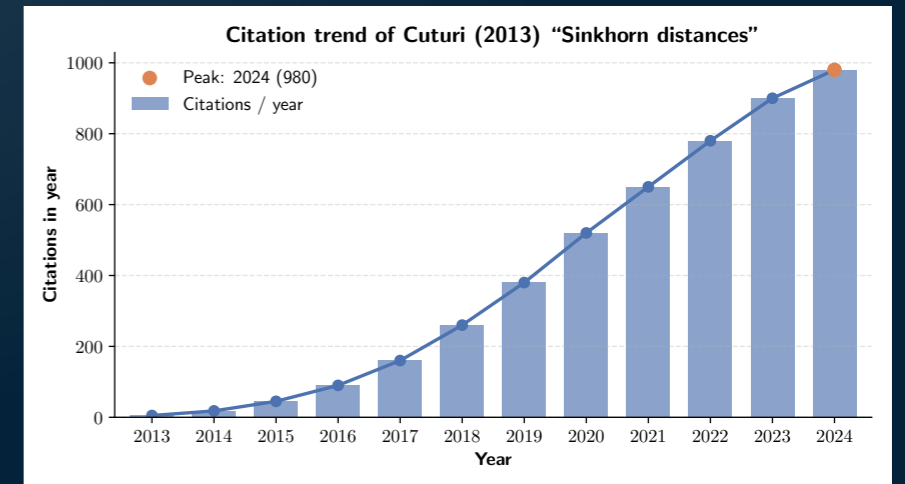
Regularisation parameter

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



Marco Cuturi

**Sinkhorn Distances:  
 Lightspeed Computation of Optimal Transport**  
Conference NeurIPS, 2013  
 Marco Cuturi  
 Graduate School of Informatics, Kyoto University  
 mcuturi@i.kyoto-u.ac.jp



Sinkhorn divergence [Genevay et al., 2018; Peyré & Cuturi, 2019]

$$SKD_\eta(\mu, \nu) = \mathcal{S}_\eta(\mu, \nu) - \frac{1}{2} \mathcal{S}_\eta(\mu, \mu) - \frac{1}{2} \mathcal{S}_\eta(\nu, \nu)$$

# $G_\sigma \text{SD}_p(\mu, \nu)$ : Maximum Mean Discrepancy

- Let  $k : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be a measurable bounded kernel on  $\mathbb{R}$  and consider the reproducing kernel Hilbert space (RKHS)  $\mathcal{H}_k$  associated with  $k$  and equipped with inner product  $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$  and norm  $\| \cdot \|_{\mathcal{H}_k}$ .

- Let  $\mathcal{P}_{\mathcal{H}_k}(\mathbb{R})$  be the set of probability measures  $\eta$  such that  $\forall t, k(t, \cdot) \in \mathcal{H}_k$  and  $\forall f \in \mathcal{H}_k, f(t) = \langle f, k(t, \cdot) \rangle_{\mathcal{H}_k}$   
 $\int_{\mathbb{R}} \sqrt{k(t, t)} d\eta(t) < \infty$ .

- The **kernel mean embedding** is defined as

$$\Phi_k(\eta) = \int_{\mathbb{R}} k(\cdot, t) d\eta(t).$$

- The squared-maximum mean discrepancy (MMD) between  $\mu$  and  $\nu$

$$\text{MMD}^2(\mu, \nu) = \|\Phi_k(\mu) - \Phi_k(\nu)\|_{\mathcal{H}_k}^2$$

$$= \mathbb{E}_{T, T' \sim \mu} [k(T, T')] - 2\mathbb{E}_{T \sim \mu, R \sim \nu} [k(T, R)] + \mathbb{E}_{R, R' \sim \nu} [k(R, R')]$$

[Gretton et al. (2006)]

# $G_\sigma \text{SD}_p(\mu, \nu)$ : Topological Properties

- $G_\sigma \text{SD}_p(\mu, \nu)$  is a proper metric on  $\mathcal{P}_p(\mathbb{R}^d) \times \mathcal{P}_p(\mathbb{R}^d)$ .

**Theorem** For any  $\sigma > 0, p \geq 1$ , the following properties hold:

1. if  $D(\cdot, \cdot)$  is non-negative (or symmetric), then  $G_\sigma \text{SD}_p(\cdot, \cdot)$  is non-negative (or symmetric);
2. if  $D(\cdot, \cdot)$  satisfies the identity of indiscernibles, i.e. for  $\mu', \nu' \in \mathcal{P}(\mathbb{R})$ ,  $D(\mu', \nu') = 0$  if and only if  $\mu' = \nu'$ , then this identity also holds for  $G_\sigma \text{SD}_p(\cdot, \cdot)$  for any  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ ;
3. if  $D(\cdot, \cdot)$  satisfies the triangle inequality then  $G_\sigma \text{SD}_p(\cdot, \cdot)$  satisfies the triangle inequality.

- $G_\sigma \text{SD}_p(\mu, \nu)$  metrizes the weak topology.

**Theorem** Let  $\sigma > 0, p \geq 1$ ,  $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ , and  $\{\mu_k \in \mathcal{P}_p(\mathbb{R}^d)\}_{k \in \mathbb{N}}$  a sequence of distributions. Assume that the divergence  $D$  is bounded and metrizes the weak topology on  $\mathcal{P}(\mathbb{R})$ . Then,  $\lim_{k \rightarrow \infty} G_\sigma \text{SD}_p(\mu_k, \mu) = 0$  if and only if  $\mu_k \Rightarrow \mu$ .

- $G_\sigma \text{SD}_p(\mu, \nu)$  is lower semi-continuous.

**Proposition** Let  $\sigma > 0, p \geq 1$  and assume that the base divergence  $D$  is lower semi-continuous w.r.t. the weak topology in  $\mathcal{P}(\mathbb{R})$ . Then,  $G_\sigma \text{SD}_p$  is lower semi-continuous with respect to the weak topology in  $\mathcal{P}_p(\mathbb{R}^d)$ .

# $G_\sigma \text{SD}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$ : Statistical Properties (I)

- The smoothed Gaussian sliced divergence between the empirical probability measures  $\hat{\boldsymbol{\mu}}_n$  and  $\hat{\boldsymbol{\nu}}_n$

$$G_\sigma \text{SD}_p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\nu}}_n) = \left( \int_{\mathbb{S}^{d-1}} D^p(\mathcal{R}_{\mathbf{u}} \hat{\boldsymbol{\mu}}_n * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}} \hat{\boldsymbol{\nu}}_n * \mathcal{N}_\sigma) u_d(\mathbf{u}) d\mathbf{u} \right)^{1/p}.$$

- Remark that for fixed direction  $\mathbf{u} \in \mathbb{S}^{d-1}$ , the distributions  $\mathcal{R}_{\mathbf{u}} \hat{\boldsymbol{\mu}}_n * \mathcal{N}_\sigma$  and  $\mathcal{R}_{\mathbf{u}} \hat{\boldsymbol{\nu}}_n * \mathcal{N}_\sigma$  are **continuous**, in particular they are a mixture of Gaussian distributions centered on the projected samples with variance  $\sigma^2$ .

## Lemma

Conditionally on the samples  $\{X_i\}_{i=1,\dots,n}$  and  $\{Y_i\}_{i=1,\dots,n}$  one has:

$$\mathcal{R}_{\mathbf{u}} \hat{\boldsymbol{\mu}}_n * \mathcal{N}_\sigma = \frac{1}{n} \sum_{i=1}^n \mathcal{N}(\mathbf{u}^\top X_i, \sigma^2) \quad \mathcal{R}_{\mathbf{u}} \hat{\boldsymbol{\nu}}_n * \mathcal{N}_\sigma = \frac{1}{n} \sum_{i=1}^n \mathcal{N}(\mathbf{u}^\top Y_i, \sigma^2)$$

# $G_\sigma \text{SD}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$ : Statistical Properties (II)

- We further need to sample with respect to the continuous mixture Gaussian measures in order to get a **fully** empirical measure version of  $G_\sigma \text{SD}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$ .
- Let  $T_1^x, \dots, T_n^x$  and  $T_1^y, \dots, T_n^y$  be i.i.d. observations of  $\mathcal{R}_u \hat{\boldsymbol{\mu}}_n * \mathcal{N}_\sigma$  and  $\mathcal{R}_u \hat{\boldsymbol{\nu}}_n * \mathcal{N}_\sigma$  respectively.
- Sampling i.i.d.  $\{T_i^x\}_{i=1, \dots, n}$  is given by the following scheme: we first choose the component  $\mathcal{N}(\mathbf{u}^\top X_i, \sigma^2)$  then we generate  $T_i^x = \mathbf{u}^\top X_i + Z_i^x$ , where  $Z_i^x \sim \mathcal{N}_\sigma$ .

# $G_\sigma \text{SD}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$ : Statistical Properties (II)

- Hence, we set, for a given direction  $\mathbf{u}$

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^n \delta_{T_i^x} = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{u}^\top X_i + Z_i^x}$$

- The measure  $\hat{\boldsymbol{\mu}}_n \in \mathcal{P}(\mathbb{R})$  defines an empirical version of the continuous  $\mathcal{R}_{\mathbf{u}} \hat{\boldsymbol{\mu}}_n * \mathcal{N}_\sigma$ .
- We define the **double empirical smoothed Gaussian sliced divergence** as:

## Double Empirical Divergence

The double empirical smoothed Gaussian sliced divergence reads as

$$\hat{G}_\sigma \text{SD}_p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\nu}}_n) = \left( \int_{\mathbb{S}^{d-1}} D^p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\nu}}_n) u_d(\mathbf{u}) d\mathbf{u} \right)^{1/p}$$

# $G_\sigma \text{SD}_p(\mu, \nu)$ : Statistical Properties (II)

- The empirical distributions are derived from a **double sampling process**, which leads to consider a double expectations, wrt the origin distribution  $\mathbb{E}_{\mu^{\otimes n}}$  and wrt the sampling from the Gaussian smoothing  $\mathbb{E}_{\mathcal{N}_\sigma^{\otimes n}}$ .
- We first consider the conditional expectation given the samples  $\mathbb{E}_{\mathcal{N}_\sigma^{\otimes n}}[\cdot | X_1, \dots, X_n]$  then apply  $\mathbb{E}_{\mu^{\otimes n}}$ . We denote by

$$\mathbb{E}_{\mu^{\otimes n} | \mathcal{N}_\sigma^{\otimes n}}[\cdot] = \mathbb{E}_{\mu^{\otimes n}}[\mathbb{E}_{\mathcal{N}_\sigma^{\otimes n}}[\cdot | X_1, \dots, X_n]].$$

**Proposition** Fix  $\sigma > 0, p \geq 1$  and  $\vartheta > \sqrt{2}$ . For  $X \sim \mu$ , assume that  $\int_0^\infty e^{\frac{2\xi^2}{\sigma^2\vartheta^2}} \mathbf{P}[\|X\| > \xi] d\xi < \infty$ . Then,

Convergence with a rate  $\mathcal{O}(n^{-1/2p})$

$$\mathbb{E}_{\mu^{\otimes n} | \mathcal{N}_\sigma^{\otimes n}}[\hat{G}_\sigma \text{SW}_p(\hat{\mu}_n, \mu)] \leq \Xi_{p,\sigma,\vartheta} \frac{1}{n^{1/2p}} + \Upsilon_{p,\sigma,\mu} \frac{(\log n)^{1/p}}{n^{1/p}},$$

where  $\Xi_{p,\sigma,\vartheta} = \frac{2^{\frac{5}{2}-\frac{5}{4p}}}{\pi^{1/2p}} \sigma^{1-\frac{1}{4p}} \vartheta^{1+\frac{1}{p}} (\Gamma(p + \frac{1}{2}) (\sqrt{\frac{4\pi\sigma^2\vartheta^2}{\vartheta^2-2}} + 4 \int_0^\infty e^{\frac{2\xi^2}{\sigma^2\vartheta^2}} \mathbf{P}[\|X\| > \xi] d\xi))^{1/2p}$  and  $\Upsilon_{p,\sigma,\mu} = \frac{2^{2-\frac{1}{2p}} C_p}{\pi^{1/2p}} \sigma^2 (\Gamma(p + \frac{1}{2}) \sum_{k=0}^\infty \frac{(-p)_k}{(\frac{1}{2})_k} \frac{(-1)^k}{(2\sigma^2)^k k!} M_{2k}(\mu))^{1/p}$  with  $C_p$  is a positive constant depending only on  $p$ .

# Domain adaptation with $G_\sigma$ SW

$$\min_{g,h} \{ \mathcal{L}_c(h(g(\mathbf{X}_s)), \mathbf{y}_s) + \mathcal{D}(g(\mathbf{X}_s), g(\mathbf{X}_t)) \}$$

- $\mathcal{L}_c$  can be the **cross-entropy loss or a quadratic loss** and  $\mathcal{D}$  a **divergence between empirical distributions** (Gaussian-smoothed sliced divergence). We solve this problem through stochastic gradient descent.
- When performing such model adaptation, a **privacy/utility trade-off** that has to be handled. In practice, one would prefer the most private model while not hurting its performance. Hence, one would seek the largest noise level  $\sigma > 0$  to use while preserving accuracy on target domain.
- We have considered two datasets: a handwritten digit recognition (USPS/MNIST) and Office 31 datasets.

# Domain adaptation with $G_{\sigma}SW$ : MNIST(source) to USPS (target)



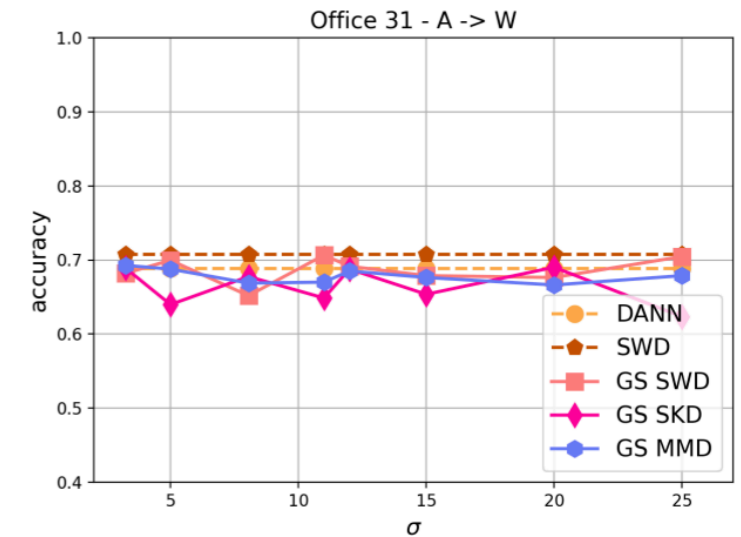
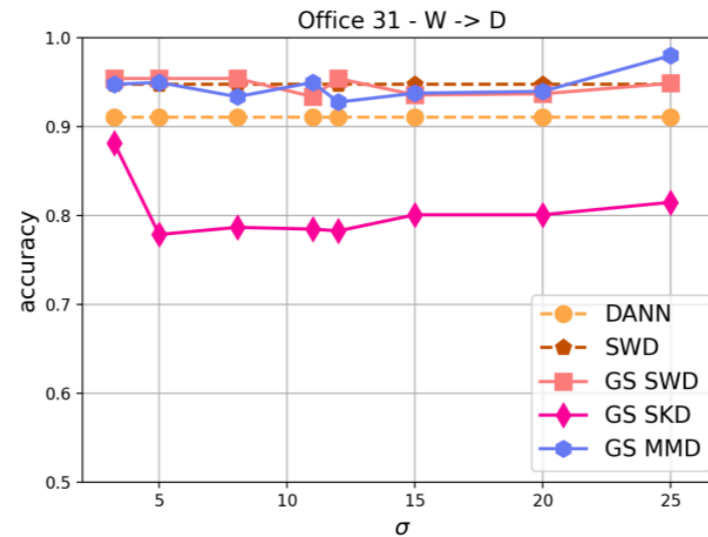
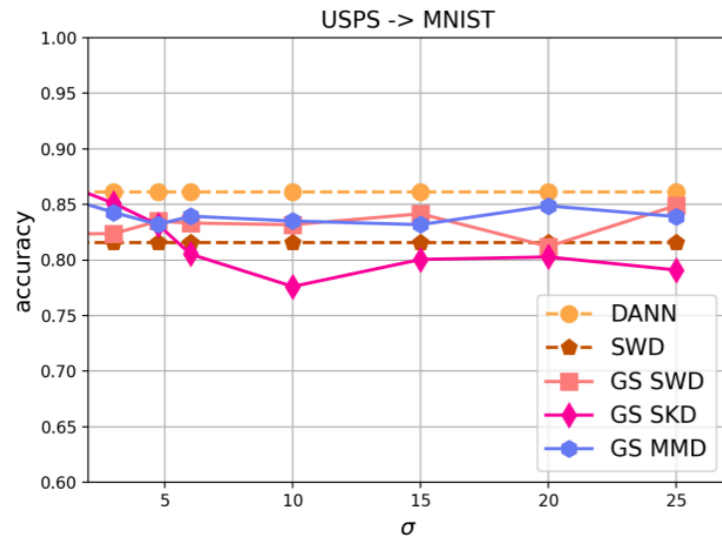
MNIST



USPS

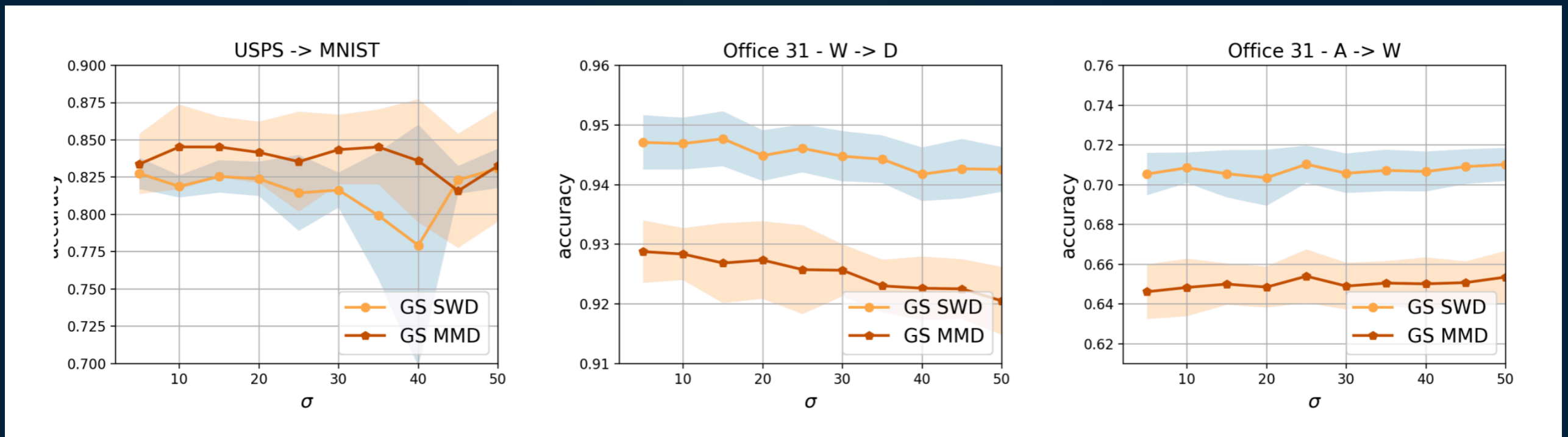
# Domain adaptation with $G_\sigma$ SW

$$\min_{g,h} \{ \mathcal{L}_c(h(g(\mathbf{X}_s)), \mathbf{y}_s) + \mathcal{D}(g(\mathbf{X}_s), g(\mathbf{X}_t)) \}$$



Domain adaptation performances using different divergences on distributions with respect to the Gaussian smoothing. (Left) USPS to MNIST. (Middle) Office-31 Webcam to DSLR. (Right) Office-31 Amazon to Webcam.

# Domain adaptation with $G_\sigma$ SW



Domain adaptation performances using different divergences on distributions with respect to the Gaussian smoothing using one-epoch-fine-tuned models. (Left) USPS to MNIST. (Middle) Office-31 Webcam to DSLR. (Right) Office-31 Amazon to Webcam.

# 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 !
- We provide properties of Gaussian-smoothed sliced divergences for comparing distributions.
- An important direction for future research is considering non Gaussian smoothing distribution enjoying this property.

# 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, *JMLR 2021*.

## POT: Python Optimal Transport

### Contents

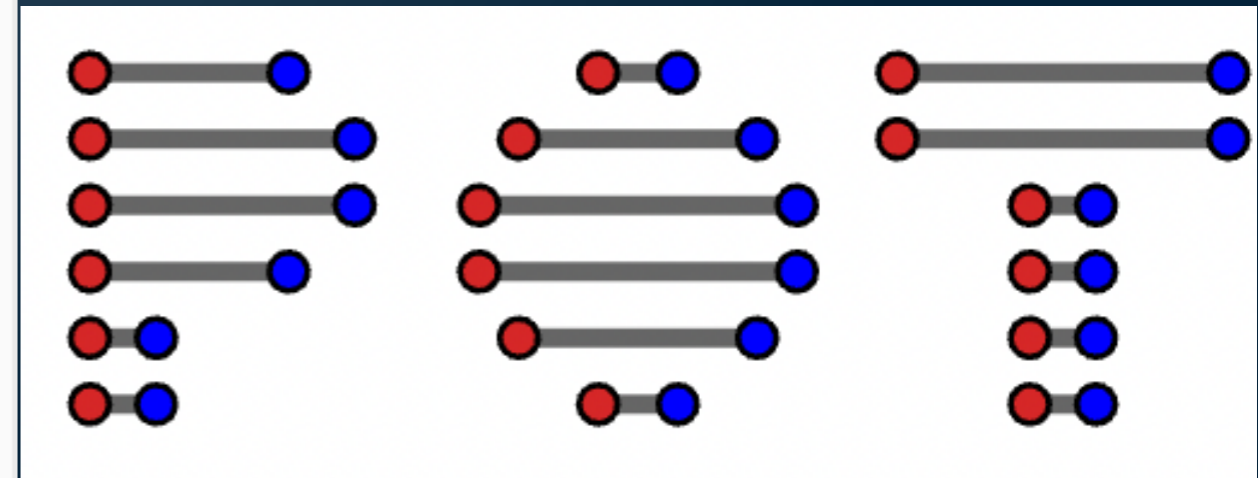
- [POT: Python Optimal Transport](#)
- [Quick start guide](#)
- [API and modules](#)
- [Examples gallery](#)
- [Releases](#)

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>



Thank You for your attention!