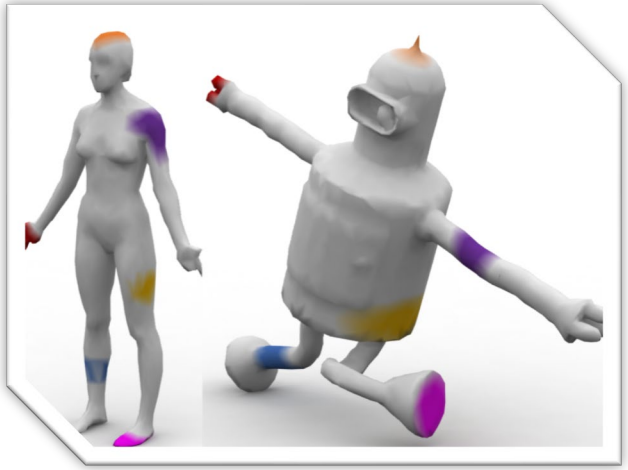
# Gromov-Wasserstein Distances: Entropic Regularization, Duality, and Sample Complexity

Ziv Goldfeld
Cornell University

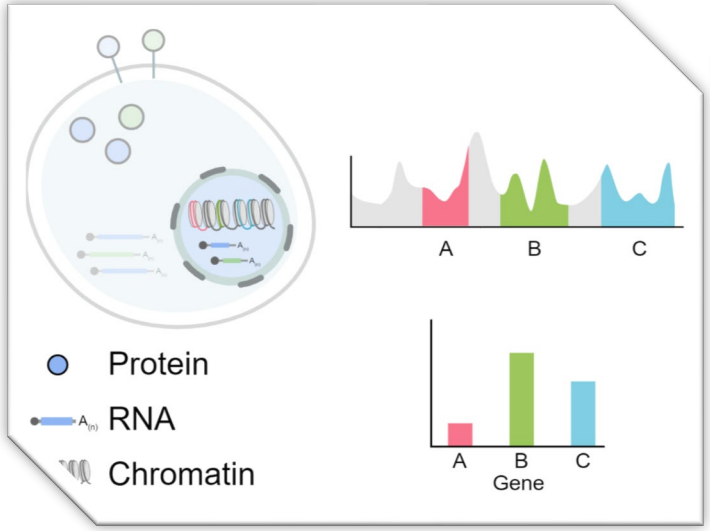
Information Theory and Applications Workshop,
February 13th, 2023

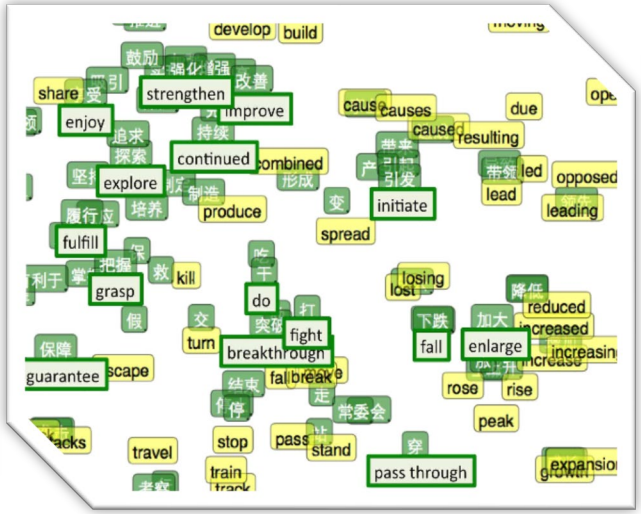
# Heterogeneous & Structured Data

**Dataset Matching:** Various applications require matching heterogeneous & structured datasets



[Solomon-Peyré-Kim-Sra '16]




Garfield as an eldritch horror.

**Goals:**

1. Compare how similar/different two datasets are
2. Obtain matching/alignment that respects individual structure

# The Gromov-Wasserstein Distance

**Object matching:** Correspondence between geometric objects

- Represent objects as metric measure spaces

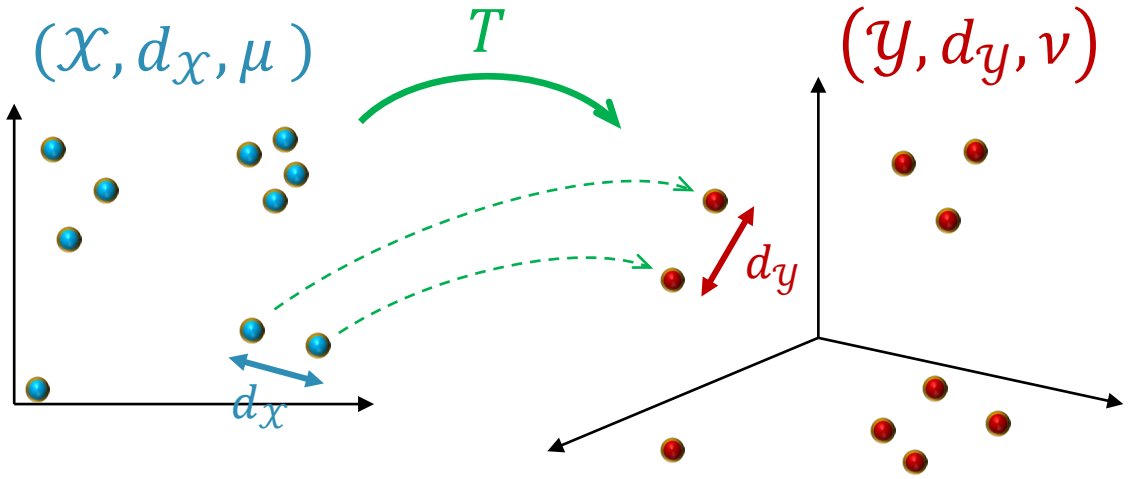
$\Longrightarrow$ $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ & $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$

- Find matching (transport map) $T: \mathcal{X} \to \mathcal{Y}$

$\Longrightarrow$ $\nu = T_{\#}\mu$ (if $X \sim \mu$ then $T(X) \sim T_{\#}\mu$)

- Preserve distances (minimize distortion)

$\Longrightarrow$ $\|x_i - x_j\| \approx \|T(x_i) - T(x_j)\|$




**Gromov-Wasserstein Distance (Memoli '11)**

The $(p, q)$-GW distance btw. mm spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ is

$$\mathrm{D}_{p,q}(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left( \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left| d_{\mathcal{X}}(x, x')^q - d_{\mathcal{Y}}(y, y')^q \right|^p d\pi \otimes \pi(x, y, x', y') \right)^{1/p}$$

# The Gromov-Wasserstein Distance

**Gromov-Wasserstein Distance (Memoli '11)**

$$\mathrm{D}_{p,q}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left( \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left| d_{\mathcal{X}}(x,x')^q - d_{\mathcal{Y}}(y,y')^q \right|^p d\pi \otimes \pi(x,y,x',y') \right)^{1/p}$$

**Comments:** Relaxation of Gromov-Hausdorff distance between measure spaces ($p=\infty, q=1$)

- **Finiteness:** $\mathrm{D}_{p,q}(\mu,\nu) < \infty \; \forall \mu,\nu$ with $\int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,x')^{pq} d\mu \otimes \mu(x,x') < \infty$ & resp. for $\nu$
- **Identification:** $\mathrm{D}_{p,q}(\mu,\nu) = 0 \iff \exists$ isomorphism $T: \mathcal{X} \to \mathcal{Y}$ with $T_{\#}\mu = \nu$
- **Metric:** Metrizes space of isomorphism (equivalence) classes of mm spaces with finite size
- **Computation:** $\mathrm{D}_{p,q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}\right)^p = \frac{1}{n^2} \min_{\sigma\in S_n} \sum_{i,j=1}^n \left| d_{\mathcal{X}}(x_i,x_j)^q - d_{\mathcal{Y}}(y_{\sigma(i)},y_{\sigma(j)})^q \right|^p$

🚫 Quadratic assignment problem (non-convex) [Commander '05] $\Longrightarrow$ **NP complete**

# Entropic GW vs. Computational Hardness

**Approach:** Explore variants of the GW problem for computational tractability

- **Sliced GW**: Avg/max of GW btw low-dimensional projections [Vayer-Flamary-Tavenard '20]
- **Unbalanced GW:** Relax marginal constraints via $f$-div. penalty [Séjourné-Vialard-Peyré '23]
- **Entropic GW:** Add entropic penalty to GW cost [Peyré-Cuturi-Solomon '16]

**Entropic Gromov-Wasserstein Distance**

$$\mathrm{S}^{\epsilon}_{p,q}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \iint \left|d_{\mathcal{X}}(x,x')^q - d_{\mathcal{Y}}(y,y')^q\right|^p d\pi \otimes \pi(x,y,x',y') + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$$

⊛ Computed via mirror-descent w/ Sinkhorn iterations [Solomon *et al* '16]

↳ Sinkhorn algorithm time complexity is $\tilde{O}(n^2/\epsilon^2)$ (highly parallelizable) [Lin *et al* '22]

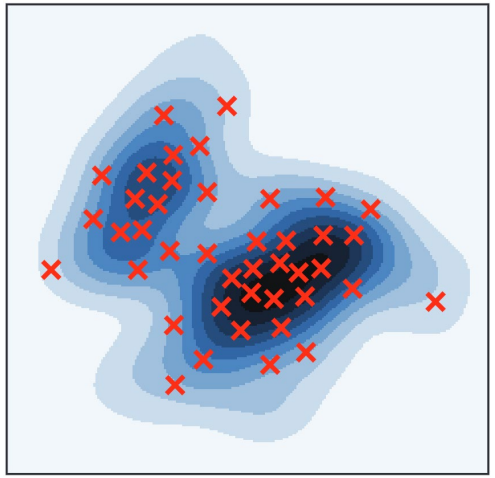
# A Statistical Question

**Question:** $\mu, \nu$ are unknown; we sample $X_1, \dots, X_n \sim \mu$ & $Y_1, \dots, Y_n \sim \nu$

- **Empirical measures:** $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$

$\Longrightarrow$ Can we approximate $\mathrm{D}_{p,q}(\mu, \nu) \approx \mathrm{D}_{p,q}(\hat{\mu}_n, \hat{\nu}_n)$?

... $\mathrm{S}^{\epsilon}_{p,q}(\mu, \nu) \approx \mathrm{S}^{\epsilon}_{p,q}(\hat{\mu}_n, \hat{\nu}_n)$?



**Asymptotic Ans:** **Yes!** For $\mu, \nu$ w/ finite $pq$–size, $\mathrm{D}_{p,q}(\hat{\mu}_n, \hat{\nu}_n) \to \mathrm{D}_{p,q}(\mu, \nu)$ a.s. [Mémoli '11]

**Non-Asymptotic Regime:** What is the **rate** at which $\mathbb{E}\left[\left|\mathrm{D}_{p,q}(\mu, \nu) - \mathrm{D}_{p,q}(\hat{\mu}_n, \hat{\nu}_n)\right|\right]$ decays?

🚫 **Open question:** No available results for either $\mathrm{D}_{p,q}$ or $\mathrm{S}^{\epsilon}_{p,q}$

↳ **Statistical implications:** Principled sample-size selection + further stat. advancements

↳ **Computational implications:** Time complexity depends on sample size

# Duality Theory for (Entropic) GW Distance

**Setting:** $(2,2)$-cost over Euclidean spaces

- **mm-spaces:** $\left(\mathbb{R}^{d_x}, \|\cdot\|, \mu\right)$ and $\left(\mathbb{R}^{d_y}, \|\cdot\|, \nu\right)$ with $M_4(\mu) := \int \|x\|^4 d\mu(x)\,, M_4(\nu) < \infty$
- **Quadratic GW:** $\mathrm{S}_\epsilon(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \iint \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 d\pi \otimes \pi + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$

**Decomposition:** Assume wlog that $\mu, \nu$ are centered (invariance to translation); then

$$\mathrm{S}_\epsilon(\mu,\nu) = \mathrm{S}_1(\mu,\nu) + \mathrm{S}_{2,\epsilon}(\mu,\nu)$$

where $\mathrm{S}_1(\mu,\nu) = \int \|x-x'\|^4 d\mu \otimes \mu + \int \|y-y'\|^4 d\nu \otimes \nu - 4\int \|x\|^2\|y\|^2 d\mu \otimes \nu$

$$\mathrm{S}_{2,\epsilon}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi - 8\sum_{\substack{1\le i\le d_x \\ 1\le j\le d_y}} \left(\int x_i y_j d\pi\right)^2 + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$$

$\Longrightarrow$ Derive a dual form for $\mathrm{S}_{2,\epsilon}(\mu,\nu)$ by linearizing quadratic term!

# Duality Theory for (Entropic) GW Distance

**Approach:** Linearize quadratic term using auxiliary variables

$$\mathrm{S}_{2,\epsilon}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi - 8\sum_{\substack{1\le i\le d_x\\ 1\le j\le d_y}} \left(\int x_i y_j d\pi\right)^2 + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi\|\mu\otimes\nu)$$

$$= \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi - 8\sum_{\substack{1\le i\le d_x\\ 1\le j\le d_y}} \inf_{-\frac{M_{\mu,\nu}}{2}\le a_{ij}\le \frac{M_{\mu,\nu}}{2}} \left(a_{ij}^2 - \int a_{ij} x_i y_j d\pi\right) + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi\|\mu\otimes\nu)$$

Optimality at $a_{ij}^*(\pi) = 0.5\int x_i y_j d\pi$ and define $M_{\mu,\nu} = \sqrt{M_2(\mu)M_2(\nu)}$

$$= \inf_{\mathbf{A}\in\mathcal{D}_{M_{\mu,\nu}}} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \underbrace{\inf_{\pi\in\Pi(\mu,\nu)} \int \left(-4\|x\|^2\|y\|^2 - 32 x^{\mathrm{T}}\mathbf{A}y\right) d\pi + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi\|\mu\otimes\nu)}_{=:\ c_{\mathbf{A}}(x,y)} = \mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\mu,\nu)$$

**Theorem (Zhang-G.-Mroueh-Sriperumbudur '23)**

Fix $\epsilon > 0$, $(\mu,\nu) \in \mathcal{P}_4(\mathbb{R}^{d_x}) \times \mathcal{P}_4(\mathbb{R}^{d_y})$, and any $M \ge \sqrt{M_2(\mu)M_2(\nu)}$, we have

$$\mathrm{S}_{2,\epsilon}(\mu,\nu) = \inf_{\mathbf{A}\in\mathcal{D}_M} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \sup_{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)} \int \varphi d\mu + \int \psi d\nu - \epsilon \int e^{\frac{\varphi(x)+\psi(y)-c_{\mathbf{A}}(x,y)}{\epsilon}} d\mu\otimes\nu + \epsilon$$

# Sample Complexity of Entropic GW

**Theorem (Zhang-G.-Mroueh-Sriperumbudur '23)**

Fix $\epsilon > 0$ and let $(\mu, \nu) \in \mathcal{P}(\mathbb{R}^{d_x}) \times \mathcal{P}(\mathbb{R}^{d_y})$ be 4-sub-Weibull with param. $\sigma^2 > 0$. Then

$$\mathbb{E}[|\mathrm{S}_\epsilon(\mu,\nu) - \mathrm{S}_\epsilon(\hat{\mu}_n,\hat{\nu}_n)|] \lesssim_{d_x,d_y} \underbrace{\frac{1+\sigma^4}{\sqrt{n}}}_{\substack{S_1 \text{ rate} \\ + \\ \text{centering bias}}} + \underbrace{\epsilon\left(1+\left(\frac{\sigma}{\sqrt{\epsilon}}\right)^{9\lceil (d_x \vee d_y)/2\rceil + 11}\right)\frac{1}{\sqrt{n}}}_{S_{2,\epsilon} \text{ rate}}$$

**Comments:**

- **Optimality:** Rate is parametric and hence minimax optimal
- **Entropic OT:** Rate is matches that for EOT (assuming compact support or sub-Gaussianity)
- **Constants:** May not be optimal but matches best known dependence on $\epsilon, \sigma$ for EOT
- **One-sample:** When only $\mu$ is estimated, rate is similar but with $d_x$ instead of $d_x \vee d_y$

# Sample Complexity of Entropic GW: Proof Outline

**Decomposition:** Split $S_\epsilon$ into $S_1 + S_{2,\epsilon}$ and center empirical measures

$$\mathbb{E}[|S_\epsilon(\mu,\nu) - S_\epsilon(\hat{\mu}_n,\hat{\nu}_n)|] \le \mathbb{E}[|S_1(\mu,\nu) - S_1(\hat{\mu}_n,\hat{\nu}_n)|] + \mathbb{E}\left[\left|S_{2,\epsilon}(\mu,\nu) - S_{2,\epsilon}(\hat{\mu}_n,\hat{\nu}_n)\right|\right] + \frac{\sigma^2}{\sqrt{n}}$$

**$S_1$ Analysis:** Involves only estimation of moments

**$S_{2,\epsilon}$ Analysis:** Hinges on dual form + regularity analysis of optimal potentials

1. **EOT reduction:** $\left|S_{2,\epsilon}(\mu,\nu) - S_{2,\epsilon}(\hat{\mu}_n,\hat{\nu}_n)\right| \le \sup_{\mathbf{A}\in\mathcal{D}_M} \left|\mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\mu,\nu) - \mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\hat{\mu}_n,\hat{\nu}_n)\right|$ ⊛
2. **Potentials:** For each $\mathbf{A} \in \mathcal{D}_M$: $|\varphi(x)| \le C_{d_x,d_y}(1+\tilde{\sigma}^5)(1+\|x\|^4)$

   $|D^\alpha\varphi(x)| \le C_{\alpha,d_x,d_y}\left(1+\tilde{\sigma}^{4.5|\alpha|}\right)\left(1+\|x\|^{3|\alpha|}\right), \forall \alpha \in \mathbb{N}_0^{d_x}$

$\Longrightarrow$ $\forall \mathbf{A} \in \mathcal{D}_M \quad (\varphi^*,\psi^*) \in \mathcal{F}_s \times \mathcal{G}_s$ for Hölder classes of arbitrary smoothness

3. **Reduction to emp. Process:** $\mathbb{E}[⊛] \le \mathbb{E}\left[\sup_{\varphi\in\mathcal{F}_s} |(\mu-\hat{\mu}_n)\varphi|\right] + \mathbb{E}\left[\sup_{\psi\in\mathcal{G}_s} |(\mu-\hat{\mu}_n)\psi|\right]$

$\Longrightarrow$ Partition $\mathbb{R}^{d_x}$ into compact shells **&** bound ent. integral of $\mathcal{F}_s$ (Hölder) with $s = \left\lceil \frac{d_x}{2} \right\rceil + 1$

# Standard GW: Duality & Sample Complexity

**Theorem (Zhang-G.-Mroueh-Sriperumbudur '23)**

For $(\mu,\nu) \in \mathcal{P}_4(\mathbb{R}^{d_x}) \times \mathcal{P}_4(\mathbb{R}^{d_y})$ and any $M \geq \sqrt{M_2(\mu)M_2(\nu)}$, we have

$$\mathrm{S}_{2,\epsilon}(\mu,\nu) = \inf_{\mathbf{A}\in\mathcal{D}_M} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \sup_{\substack{(\varphi,\psi)\in C_b(\mathbb{R}^{d_x})\times C_b(\mathbb{R}^{d_x}) \\ \varphi(x)+\psi(y)\leq c_{\mathbf{A}}(x,y)}} \int \varphi d\mu + \int \psi d\nu$$

where $\mathrm{c}_{\mathbf{A}}(x,y) \coloneqq -4\|x\|^2\|y\|^2 - 32x^{\mathrm{T}}\mathbf{A}y$.

**Proof:** Same argument as before but apply **standard OT duality** in the last step

$$\mathrm{S}_{2,\epsilon}(\mu,\nu) = \cdots = \inf_{\mathbf{A}\in\mathcal{D}_{M_{\mu,\nu}}} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \underbrace{\boxed{\inf_{\pi\in\Pi(\mu,\nu)} \int \left(-4\|x\|^2\|y\|^2 - 32x^{\mathrm{T}}\mathbf{A}y\right)d\pi}}_{= \mathrm{OT}_{c_{\mathbf{A}}}(\mu,\nu)}$$

# Standard GW: Duality & Sample Complexity

**Theorem (Zhang-G.-Mroueh-Sriperumbudur '23)**

Let $(\mu,\nu) \in \mathcal{P}(\mathbb{R}^{d_x}) \times \mathcal{P}(\mathbb{R}^{d_y})$ have compact support with diameter bounded by $R > 0$. Then

$$\mathbb{E}[|\mathrm{D}(\mu,\nu)^2 - \mathrm{D}(\hat{\mu}_n,\hat{\nu}_n)^2|] \lesssim_{d_x,d_y,R} \underbrace{\frac{R^4}{\sqrt{n}}}_{S_1 \text{ rate + centering bias}} + \underbrace{(1+R^4)n^{-\frac{2}{d_x \vee d_y \vee 4}}(\log n)^{\mathbb{1}\{d_x \vee d_y = 4\}}}_{S_{2,0} \text{ rate}}$$

**Proof:** Similar argument using Lipschitness **&** concavity of optimal potentials (via cost concavity)

- **Low dimension:** Potential class is Donsker for $d_x \vee d_y \leq 3$ [Hundrieser *et al* '22]

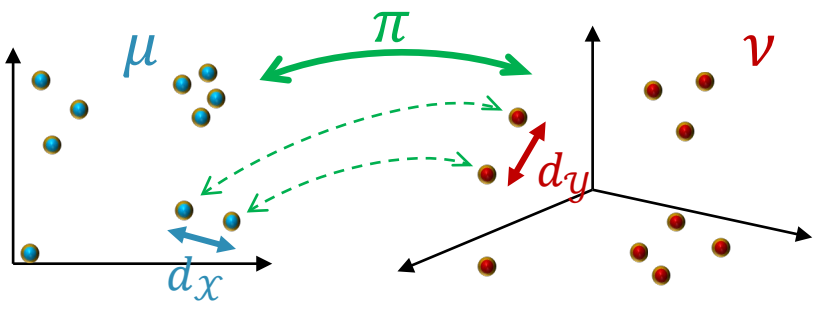
**Comments:**

- **OT:** Rate is matches that for empirical OT with compact support [Manole-Niles Weed '22]
- **Unbdd. support:** [Manole-Niles Weed '22] have argument for OT under strong assumptions
- **Non-squared GW:** If $\mathrm{D}(\mu,\nu) > 0$ then the same rates hold for empirical $\mathrm{D}$ itself
- **One-sample:** When only $\mu$ is estimated, rate is similar but with $d_x$ instead of $d_x \vee d_y$

# Summary

**Gromov-Wasserstein Distance:** Quantifies discrapency between mm spaces

- Applications in ML and beyond for heterogeneous data
- Foundational statistical **&** computational questions open




**Contributions:** Duality and first steps towards statistical theory

- Dual form using auxiliary matrix-valued variable
- First sample complexity results for GW and EGW (quadratic cost over Euclidean spaces)
- Additional results: stability or GW cost and coupling in reg. parameter

**Directions:** New optimization algorithms, limit distribution theory, GW gradient flow, etc.

[*] Zhang, Goldfeld, Mroueh, Sriperumbudur, "Gromov-Wasserstein distances: entropic regularization, duality, and sample complexity", ArXiv: 2212.12848

**Thank you!**