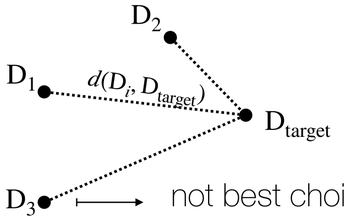
# Geometric Dataset Distances via Optimal Transport

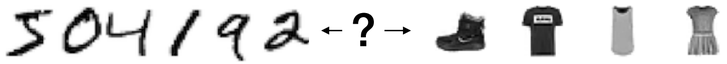
David Alvarez-Melis
Nicolò Fusi

## Distances between Datasets

- Key in various settings: transfer/meta-learning, etc



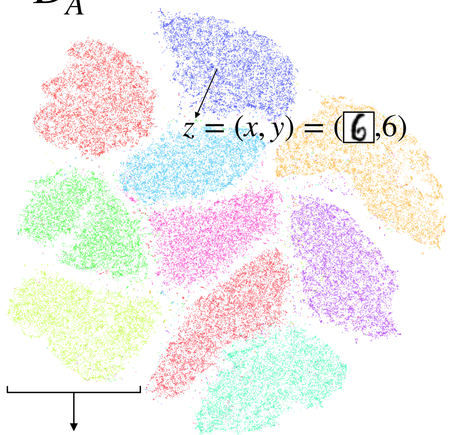

- How to deal with datasets with different label sets?




- Ideally: model agnostic, sound theoretical footing

## The Optimal Transport Dataset Distance

$\mathbf{D}_A$




Labels represented as distribution over features $v_y = \mathcal{N}(\mu_y, \Sigma_y)$

Distance between **feature/label pairs**:

$$d(z, z') = \left( d(x, x')^p + \mathrm{W}_p^p(v_y, v_{y'}) \right)^{1/p}$$

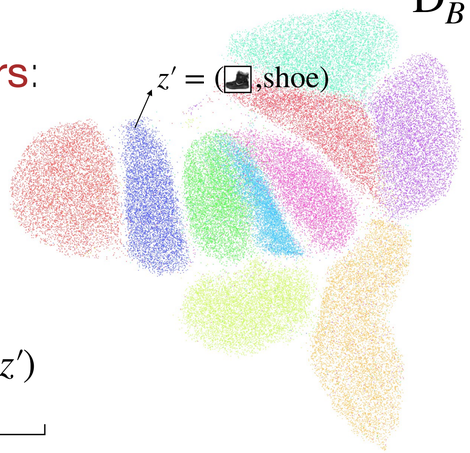
Wasserstein distance

Distance between **datasets**:

$$d_{\mathrm{OT}}(\mathrm{D}_A, \mathrm{D}_B) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{Z}\times\mathcal{Z}} d(z, z')^p \, \mathrm{d}\pi(z, z')$$

Optimal Transport distance: ≈ min-cost matching

$\mathbf{D}_B$




## Properties of OTDD

- A **true metric** in the space of measures over feature-label pairs $\mathscr{P}(\mathscr{X} \times \mathscr{Y})$
- Can be estimated from finite samples
- **Efficient computation** through a Gaussian approximation + online moment estimation

## Application: Predicting Transferability

*NIST Datasets | Text Datasets | MNIST + Augmentations