**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

**Properties of OTDD**

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

**Application: Predicting Transferability**

- *NIST Datasets*
- Text Datasets
- MNIST + Augmentations

**The Optimal Transport Dataset Distance**

$$d(z, z') = \left( d(x, x')^p + W_p^p(d_y, d_{y'}) \right)^{1/p}$$

**Distance between feature/label pairs:**

$$d(z, z') = \left( d(x, x')^p + W_p^p(d_y, d_{y'}) \right)^{1/p}$$

**Distance between datasets:**

$$d_{OT}(D_A, D_B) = \min_{\pi \in \Pi(a, b)} \int_{\mathcal{X} \times \mathcal{Y}} d(z, z') \, d\pi(z, z')$$

Labels represented as distribution over features

$$\nu_y = \mathcal{N}(\mu_y, \Sigma_y)$$

Optimal Transport distance: = min-cost matching

**Geometric Dataset Distances via Optimal Transport**

- David Alvarez-Melis
- Nicolò Fusi

*github.com/microsoft/otdd*

*dmelis.net/projects/otdd*