Introduction to Similarity and Deep Metric Learning

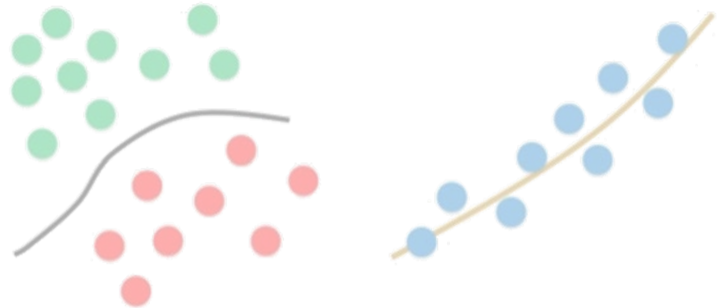
Björn Ommer

Machine Vision & Learning Group

University of Munich

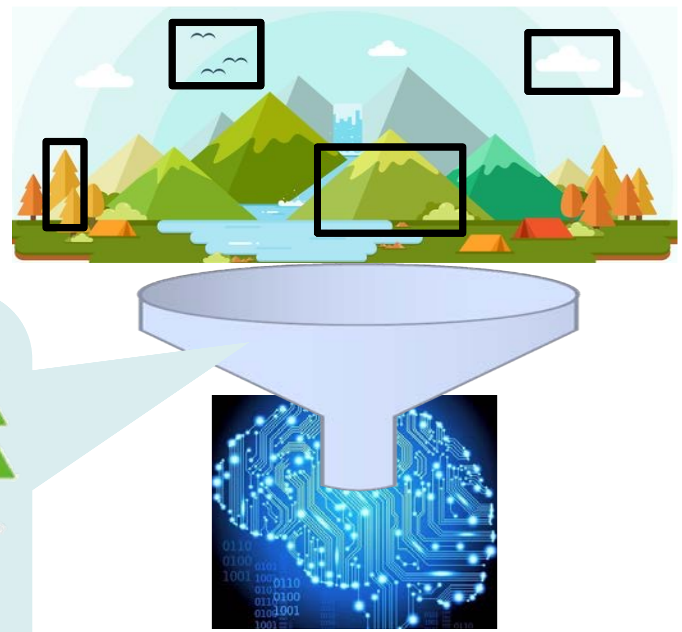
Grand Goal of Machine Learning in CV

~~Learning to solve a task~~

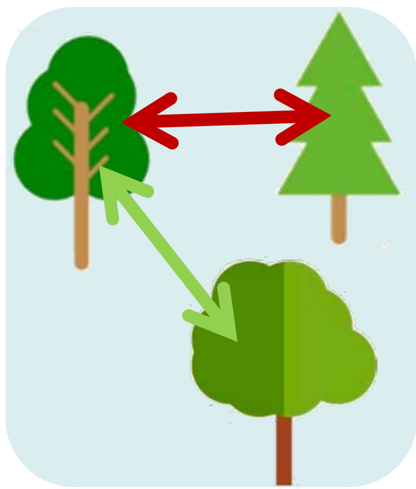


Classification Regression ...

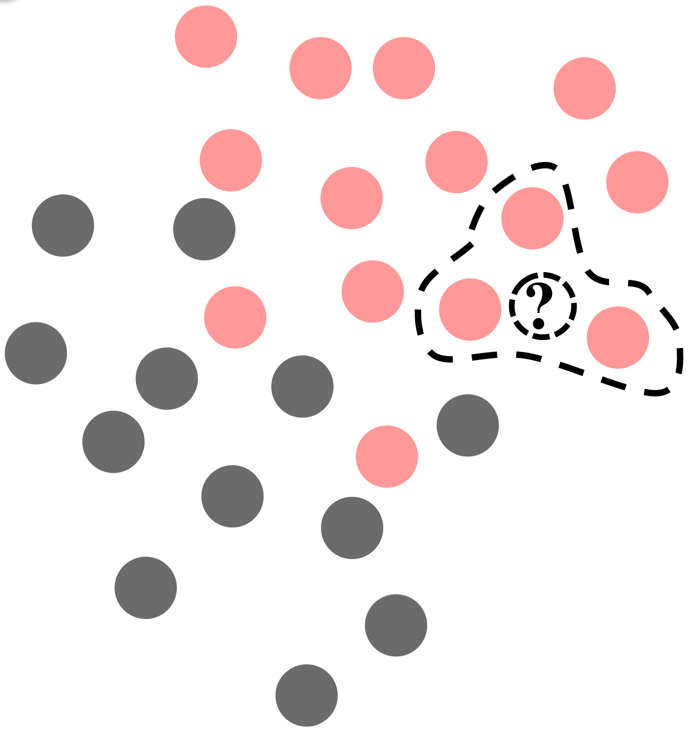
Learning a representation of the world



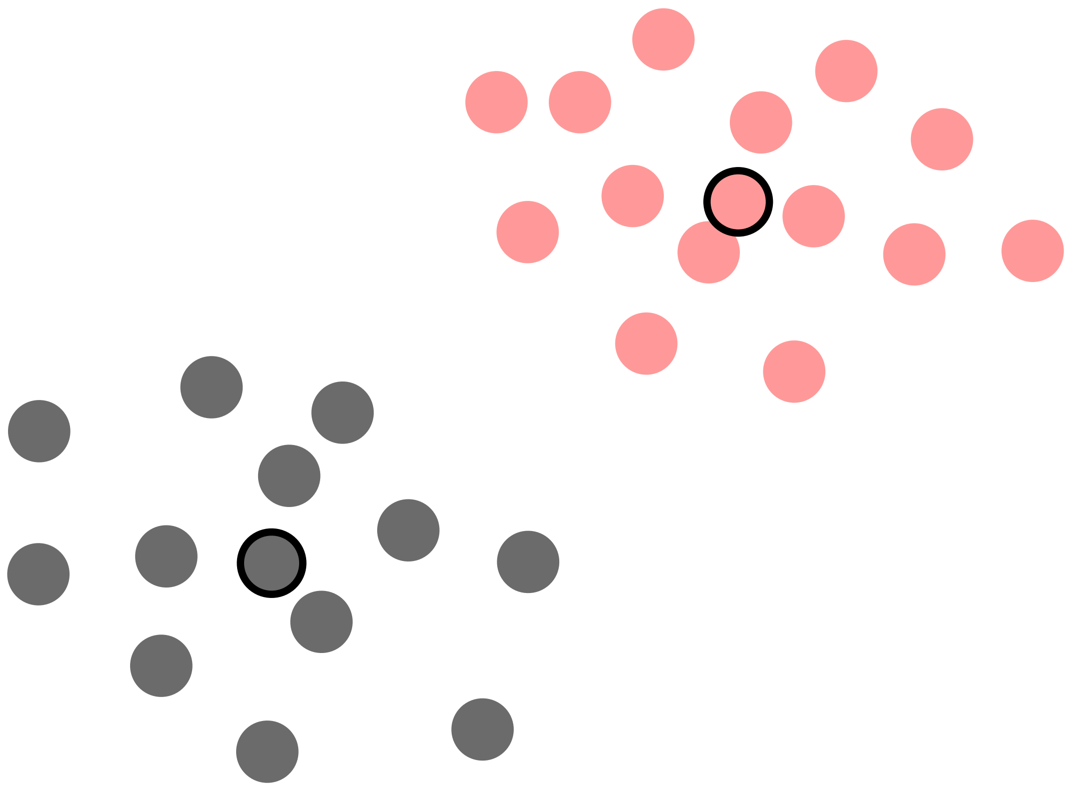
& its semantic interdependencies



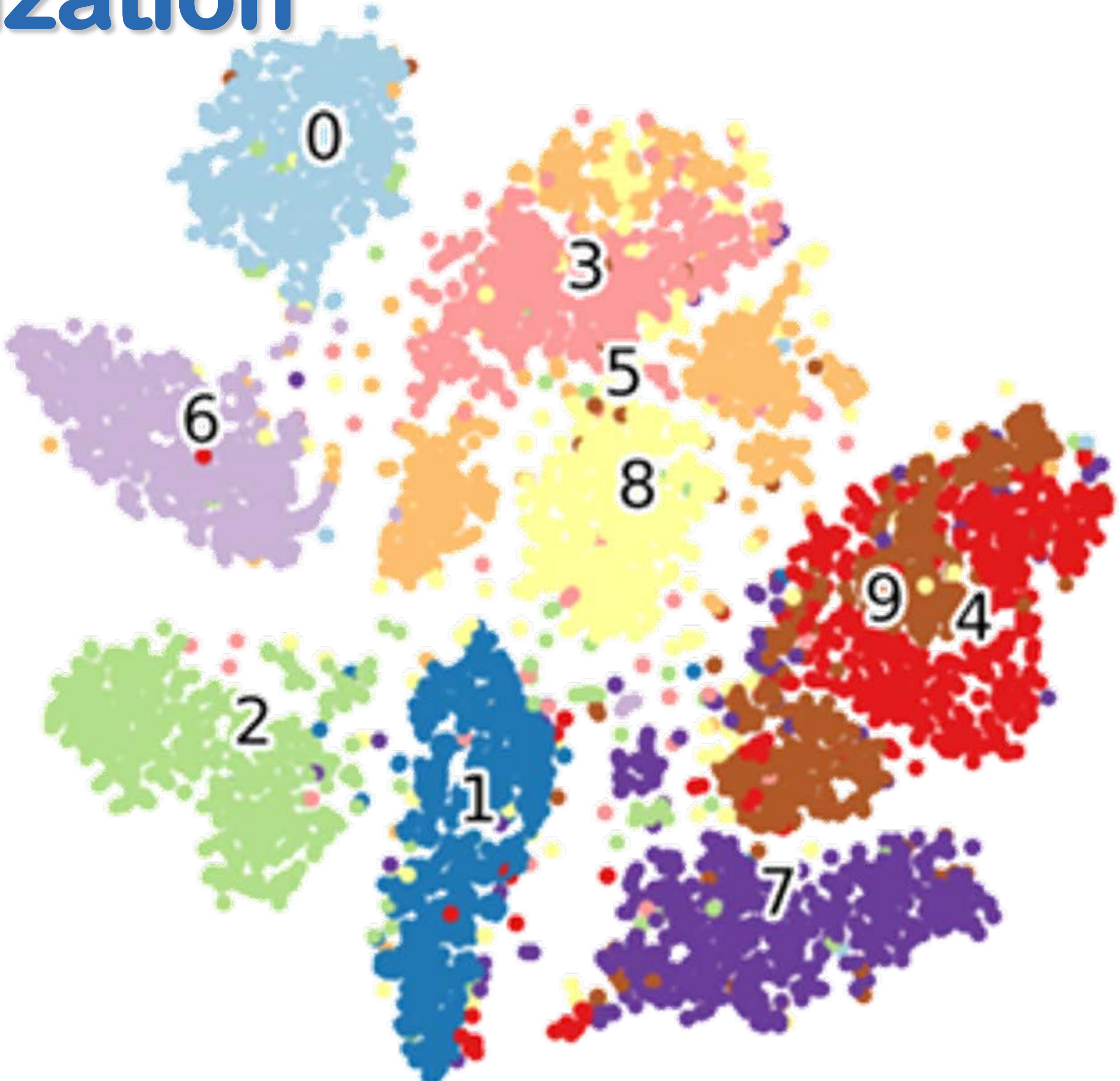
Relations Matter: *Nearest Neighbor* Classification, Density Estimation, Retrieval



Relations Matter: Grouping

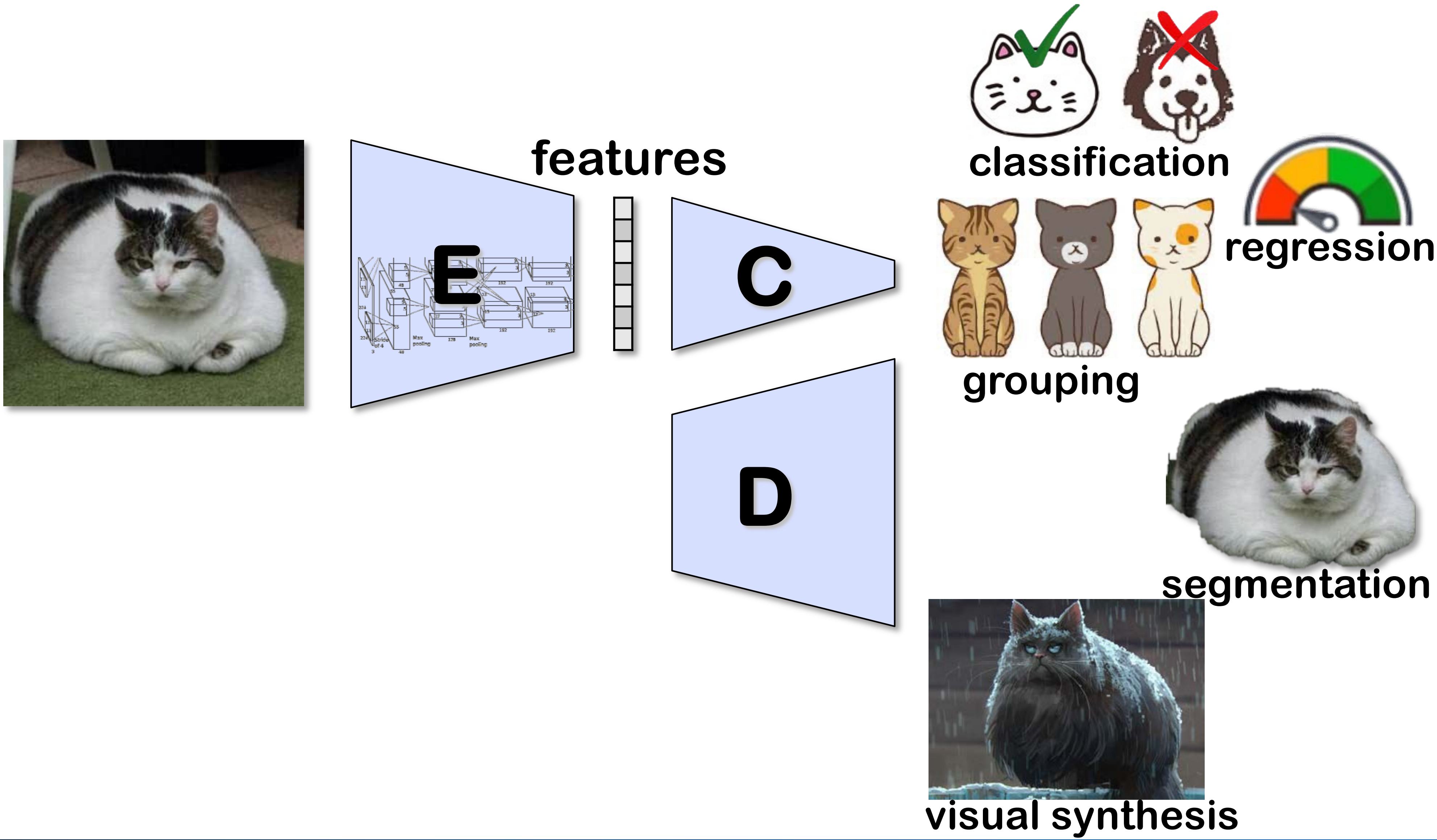


Relations Matter: Data Visualization

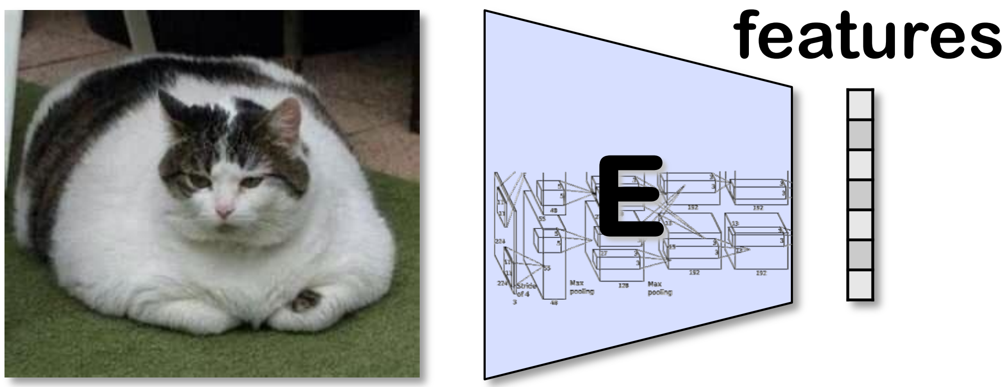


[t-SNE, v. d. Maaten & Hinton, JMLR'08]

Learning an Embedding aka Features



Learning an Embedding aka Features



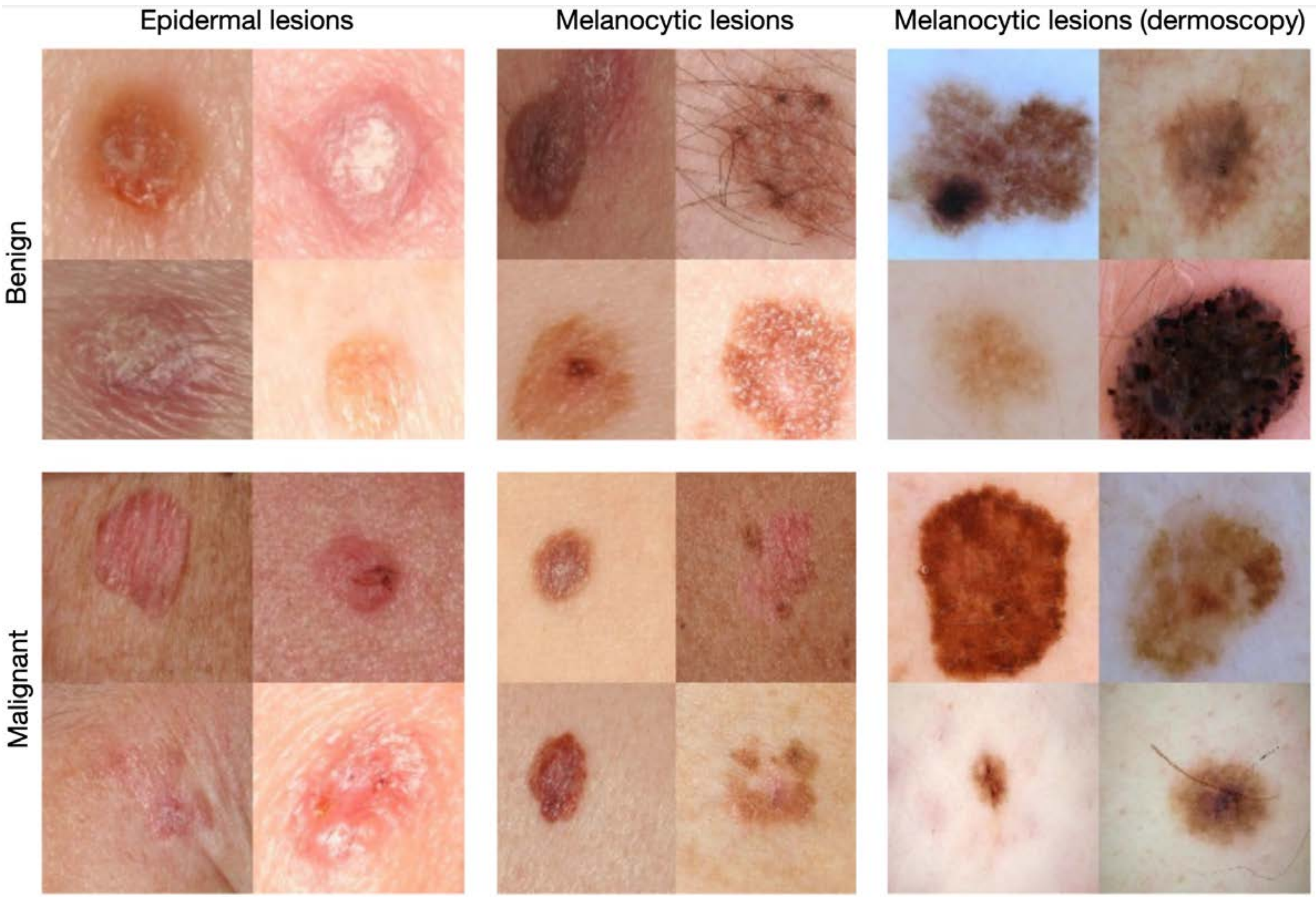
When you have difficulty in classification,
do not look for ever more esoteric mathematical tricks,
instead, **find better features**.

–B.P.K Horn: Robot Vision, 1986

Björn Ommer | b.ommer@lmu.de

Key Challenge of Data Analysis:

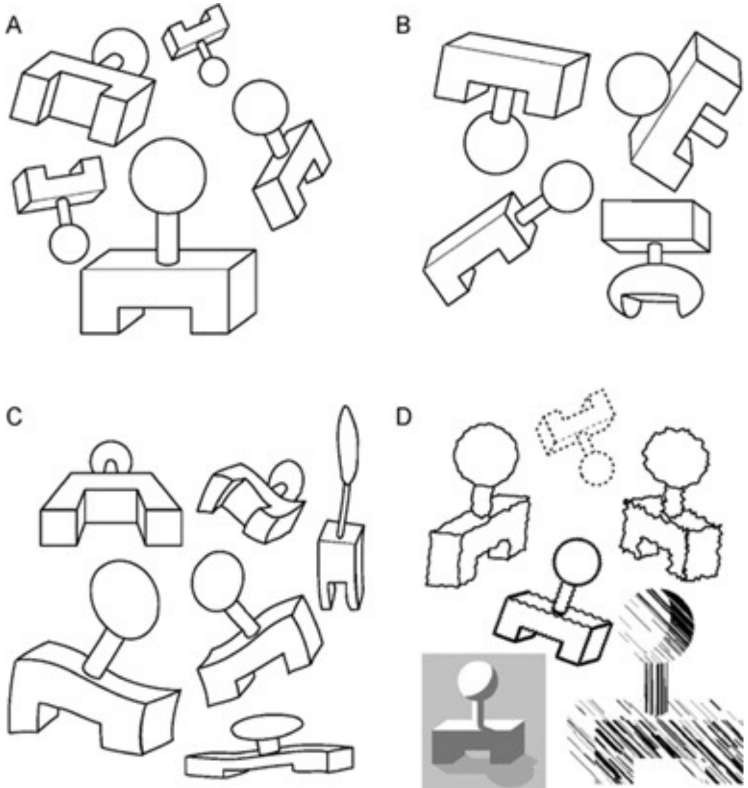
Intra-Class Variability



[Esteva et al., Nature 542, 2017]

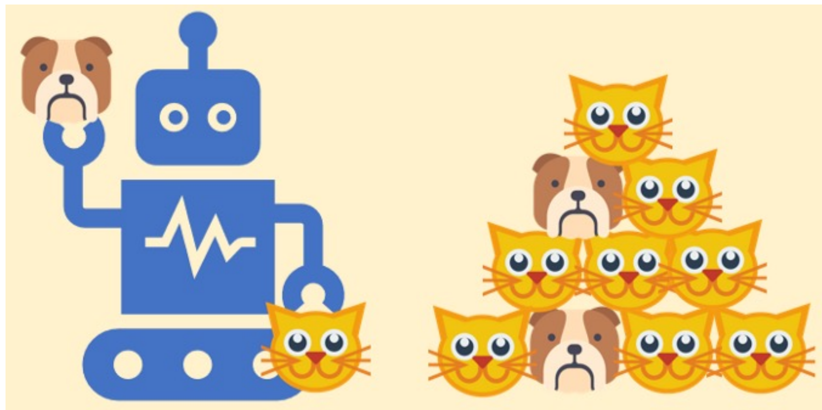
… Find Better Features

- **Invariance to clutter**



Lehar S. (2003): The World In Your Head

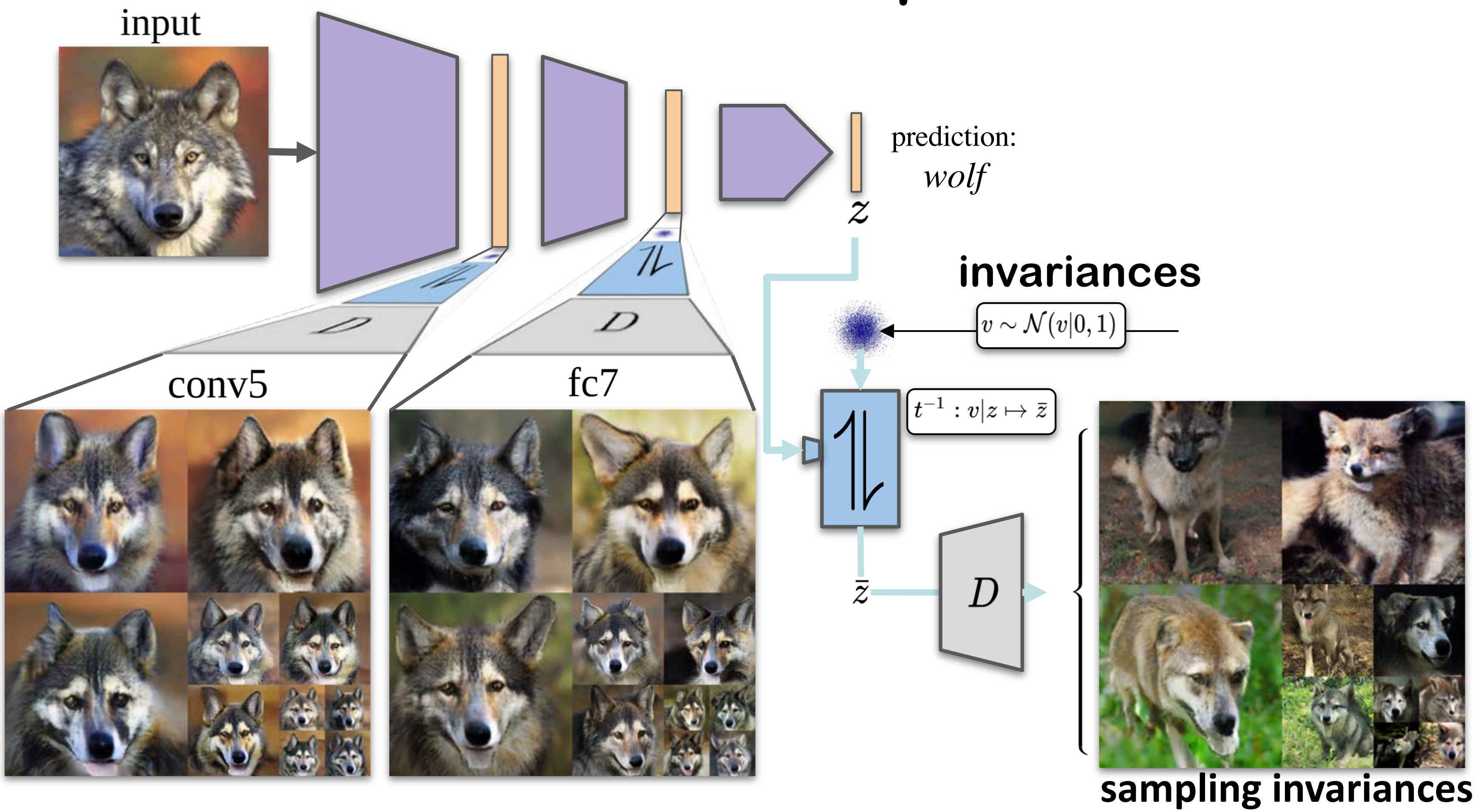
- **BUT:Preserve essential characteristics for task**



- **dim(feature) << dim(input)**

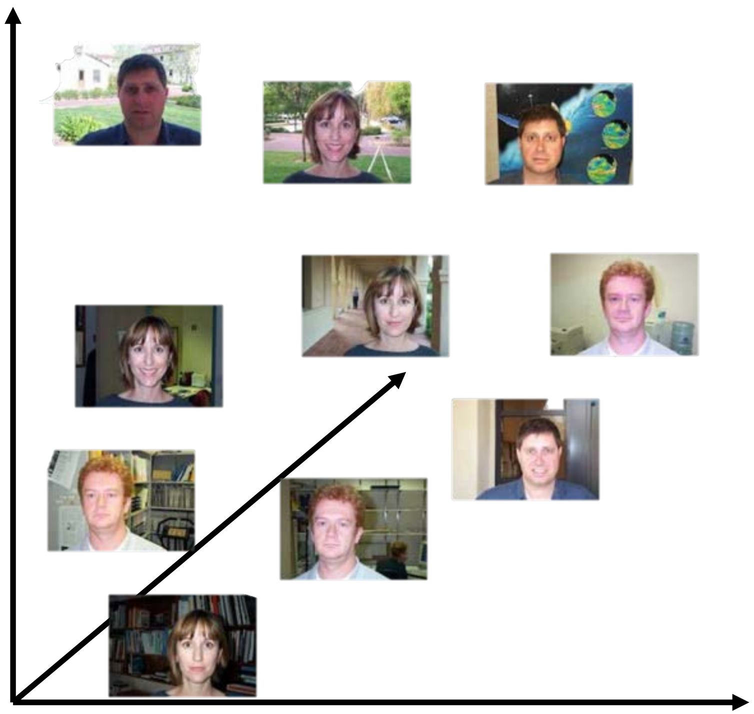
Invariance

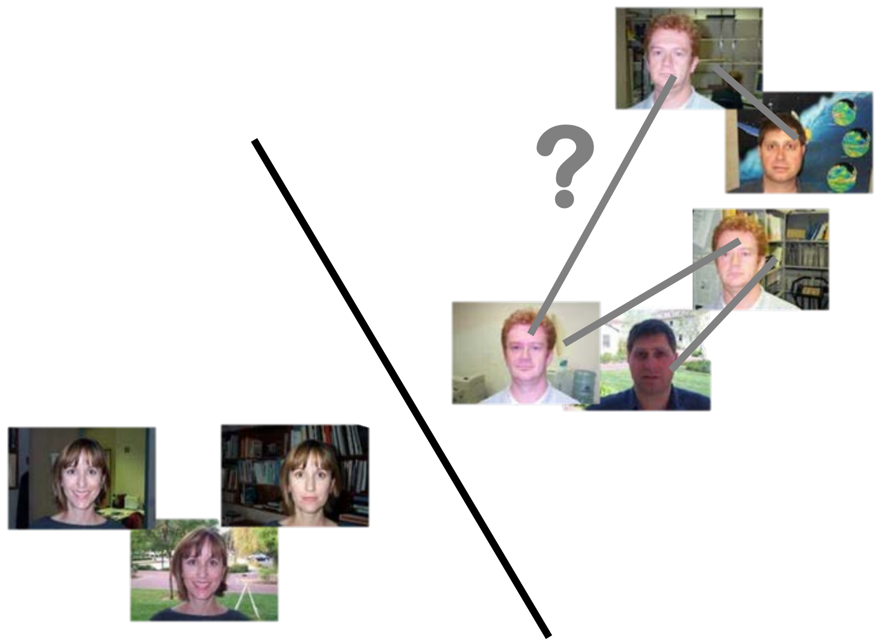
- **Invariance of features ⇨ equivalence classes**



[Rombach, Esser, Ommer, ECCV'20]

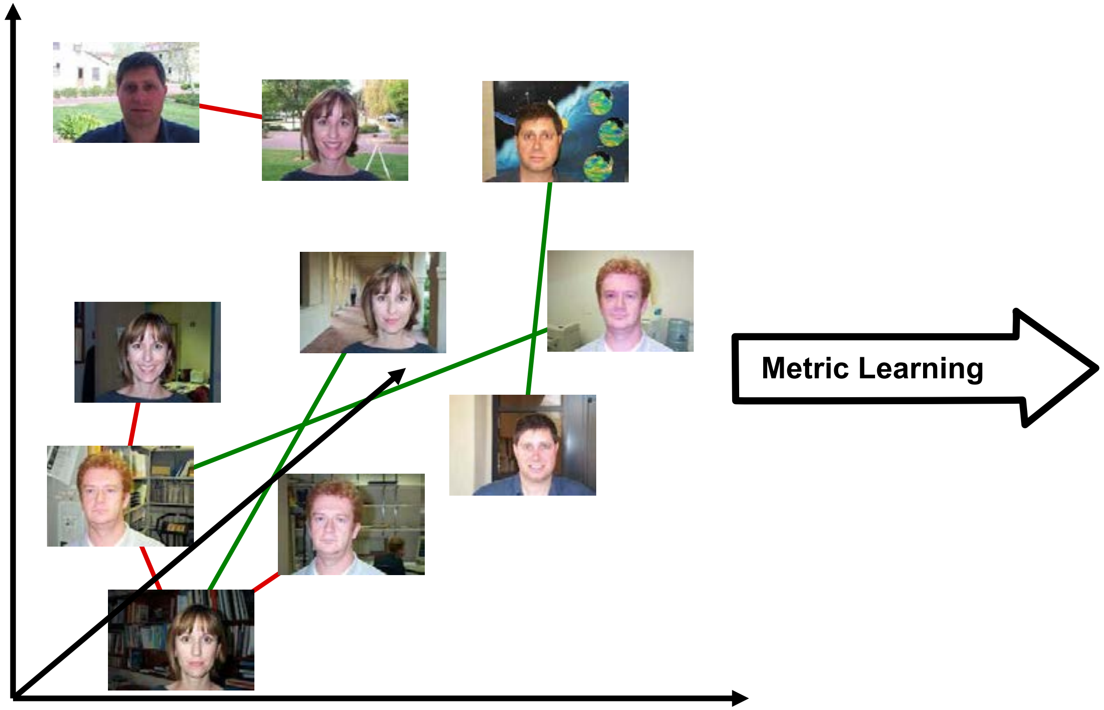
What Characteristics to Retain?

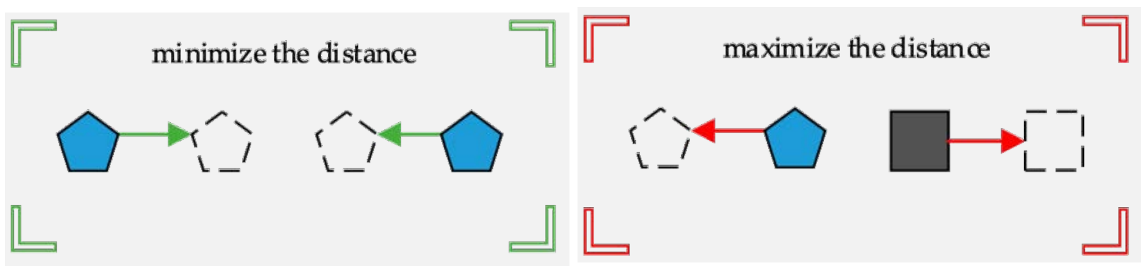




Classification

Want Richer, Fine-grained Structure?

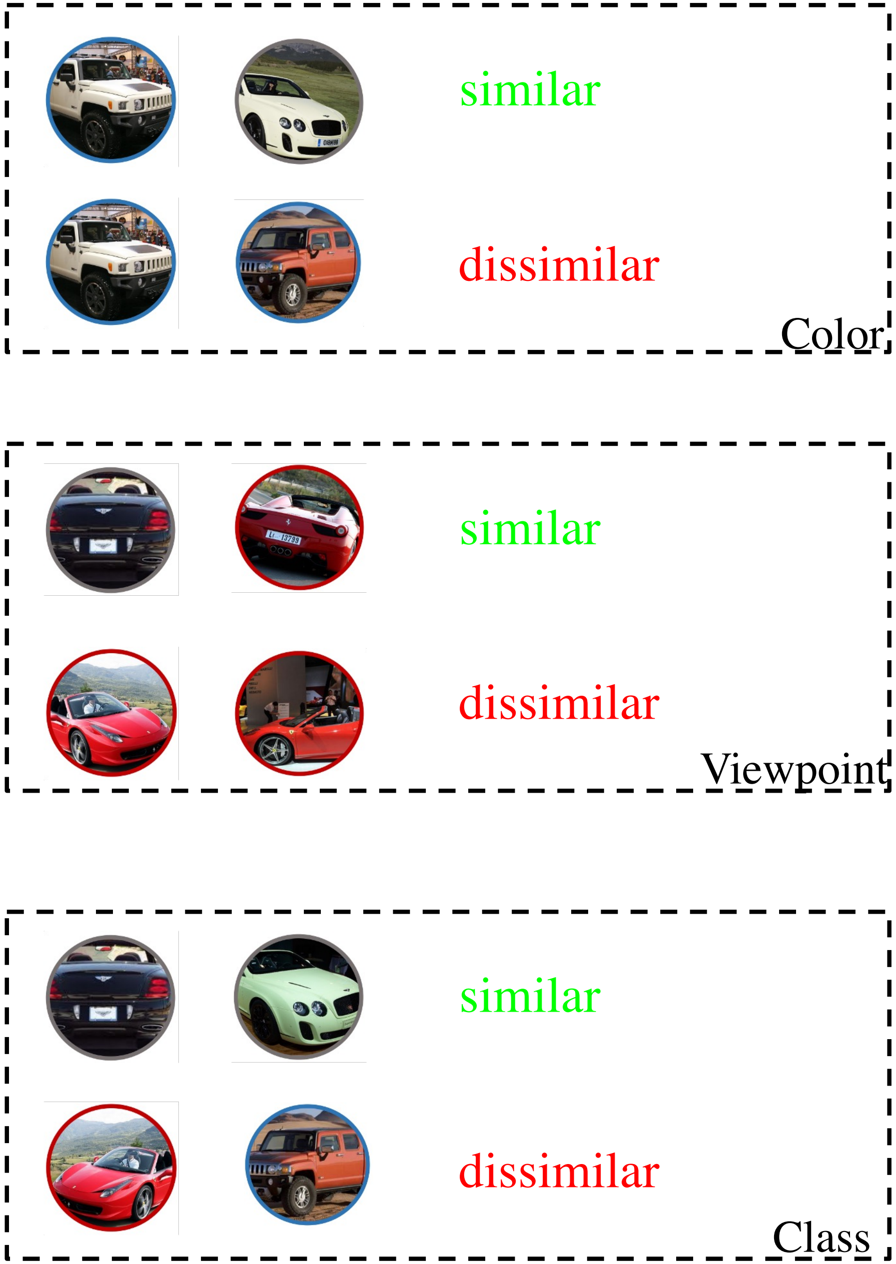




[doi:10.3390/sym11091066]

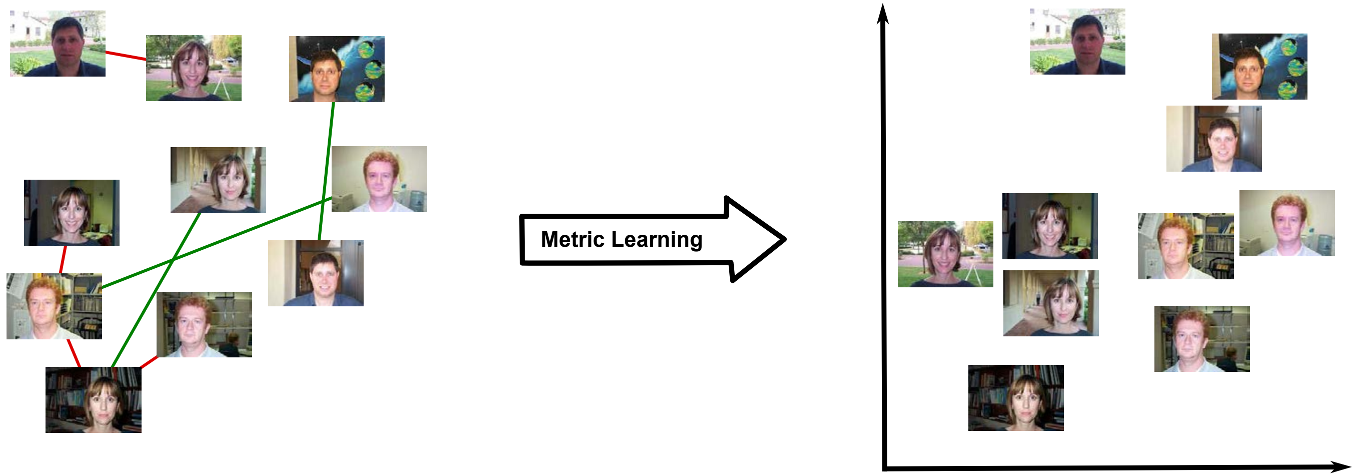
Different Notions of Similarity?

Similarity?



Metric / Similarity / Representation Learning

Goal: Learn sematic relations btw datapoints



Idea: Learn a mapping $\psi(x_i)$ s.t.:

semantic relations btw. $x_i, x_i \in \mathcal{X}$ $\xrightarrow{\psi(\cdot)}$ metric distances

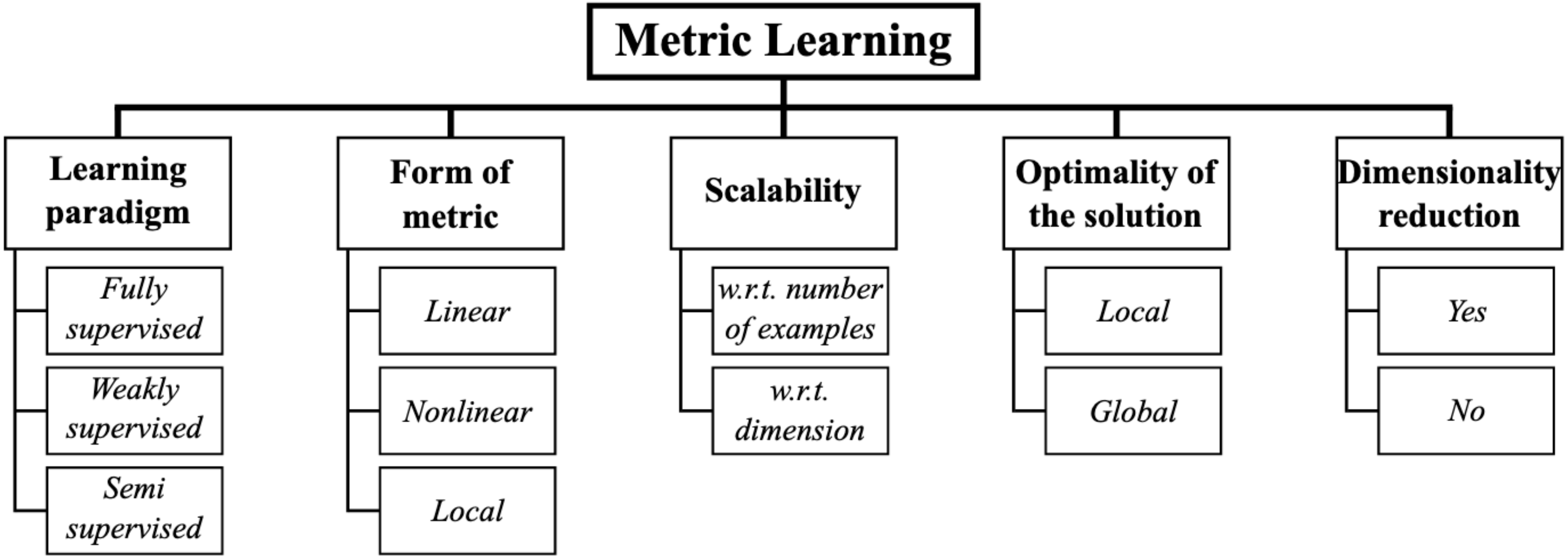
$$D_\psi = \left\|\psi(x_i) - \psi(x_j)\right\|_2^2$$

Classical approaches to learning (lin.) embedding: PCA, LDA, Conv Opt. …

(Pseudo) Metric $\mathrm{d}(\cdot,\cdot) \coloneqq \Delta(\psi_\theta(\cdot), \psi_\theta(\cdot))$

- $d\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}_{\geq 0}$
- **Pseudo:** $d(x,x) = 0$ **| metric:** $d(x,y) = 0 \Leftrightarrow x = y$
- **Symmetry:** $d(x,y) = d(y,x)$
- **Subadditivity:** $d(x,z) \leq d(x,y) + d(y,z)$

Metric Learning: Research Directions



[Bellet et al.: arXiv:1306.6709]

Linear Metric Learning: Mahalanobis Dist

$$d_M(x,y) = \sqrt{(x-y)^{\mathrm{T}} M (x-y)}$$
$$= \sqrt{(x-y)^{\mathrm{T}} L^{\mathrm{T}} L (x-y)}$$
$$= \sqrt{(Lx-Ly)^{\mathrm{T}} (Lx-Ly)}$$

$M = \Sigma^{-1}$: **Mahalanobis**
$M = \mathbb{I}$: **Euclidean**

Typically: $\mathrm{rank}(M) < \dim(\mathcal{X})$
⇨ **Low dim. embedding**

- **Challenges** [Xing et al., NIPS'02]
  - **Assuring** M **is PSD** ⇨ $\mathcal{O}(\dim(\mathcal{X})^3)$
  - **Rank constraint or regularization on** M ⇨ **NP-hard**
- **Alternative: no PSD (violate axioms) ⇨ bilinear form:** $d_M(x,y) = x^T M y$

[Xing et al. Distance Metric Learning with Application to Clustering with Side-Information. NIPS'02]

Beyond Linearity

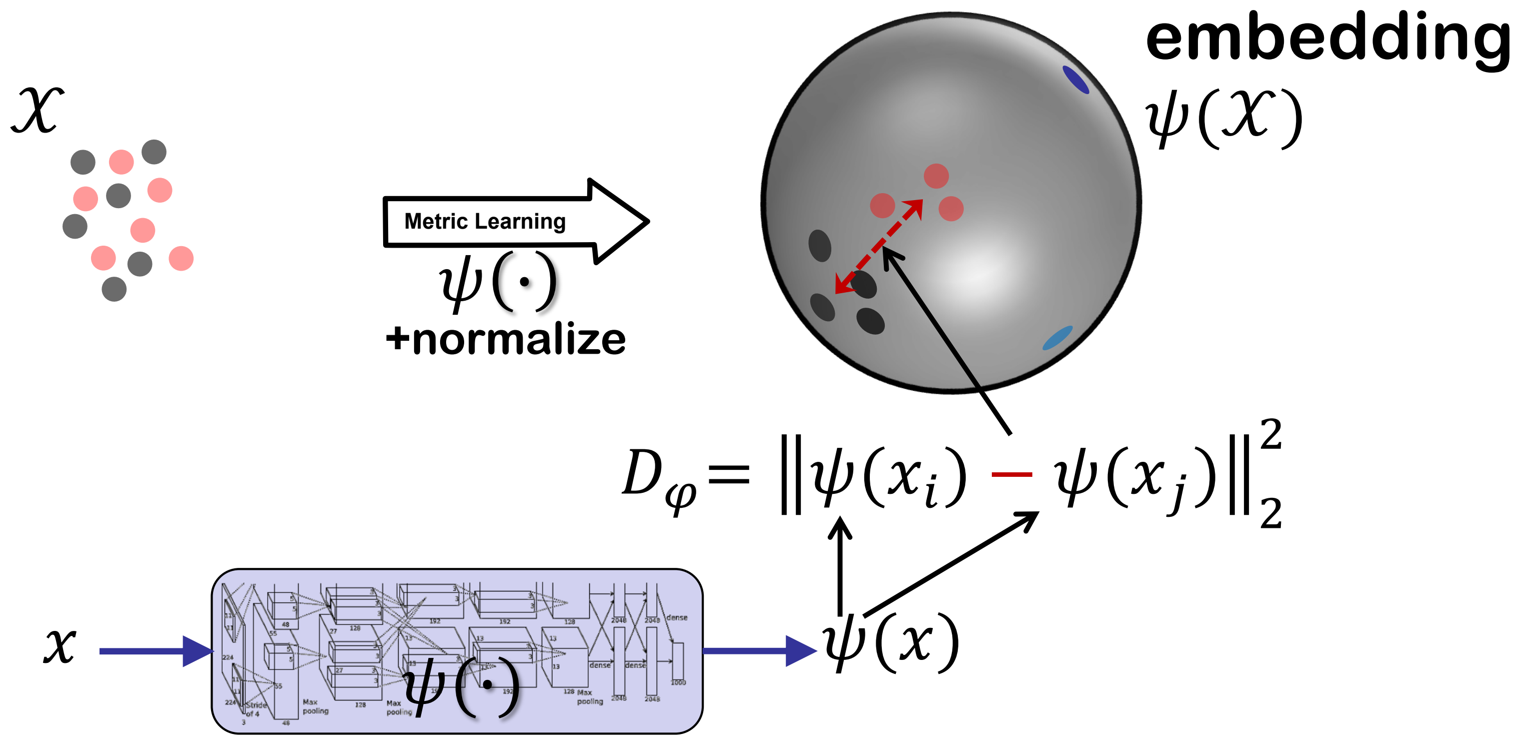
- **Linearity: Convexity & robustness to overfitting**
- **Representing non-linear structure**
  - **Kernel trick: linear metric learning after non-lin embedding into kernel space**
    - **Kernel** $\mathrm{k}(x, x') =< \phi(x), \phi(x') >$
    - $\boldsymbol{\Phi} = [\phi(x_1), \dots, \phi(x_n)]$, **let** $L^T = \Phi U^T \Rightarrow M = U^T U$

    $\Rightarrow \mathrm{d}_M^2\big(\phi(x), \phi(x')\big) = (K - K')^T M (K - K')$

    $K = \boldsymbol{\Phi}^T \phi(x) = [k(x_1, x), \dots, k(x_n, x)]^T$
  - **BUT:** $\mathcal{O}(\mathrm{n}^2)$ **params & only inner products**

[Chatpatanasiri et al. A new kernelization framework for Mahalanobis distance learning algorithms , Neurocomputing, 2010]

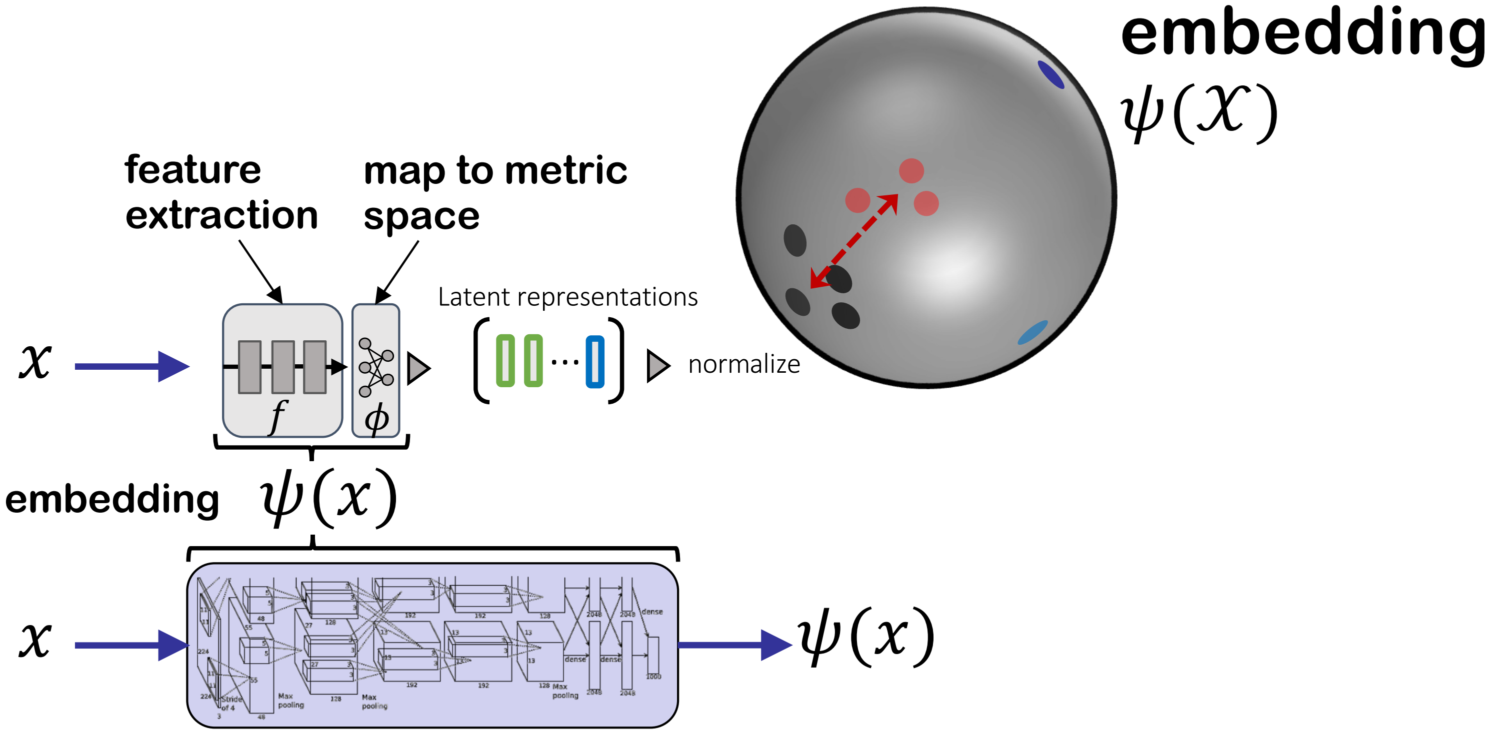
Deep Metric & Representation Learning



DML: find representation of semantic relations

[Bautista et al NIPS‘16,
Sanakoyeu et al CVPR‘19,
Milbich et al. PAMI’20, Pattern Recogn’20,
Roth et al. ICCV‘19]

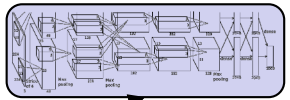
Deep Metric & Representation Learning



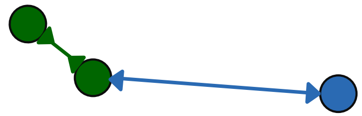
DML: find representation of semantic relations

[Bautista et al NIPS'16,
Sanakoyeu et al CVPR'19,
Milbich et al. PAMI'20, Pattern Recogn'20,
Roth et al. ICCV'19]

DML in a Nutshell



- **Choose a parametrized embedding fct** ψ_θ
- **Pick a distance measure** Δ **for the embedding space, e.g.** $\Delta(\psi_\theta(x_i), \psi_\theta(x_j)) = \left\|\psi_\theta(x_i), \psi_\theta(x_j))\right\|_2^2$

- **Gather data** $\mathcal{X} = \{x_i\}$ **& similarity judgements**
  - $S = \{(x_i, x_j) | x_i, x_j \text{ are similar}\}$ 
  - $D = \{(x_i, x_j) | x_i, x_j \text{ are dissimilar}\}$ 
  - $T = \{(x_i, x_j, x_k) | x_i \text{ is more similar to } x_j \text{ than to } x_k\}$ 

- **Optimize** θ **s.t.** $\mathrm{d}(\cdot,\cdot) := \Delta(\psi_\theta(\cdot), \psi_\theta(\cdot))$ **best agrees with judgements** $\mathrm{argmin}_\theta L(\psi_\theta, \Delta, S, D, T) + \lambda\mathcal{R}(\psi_\theta)$

loss regularization

[Bellet et al.: arXiv:1306.6709]

Main Topics in DML

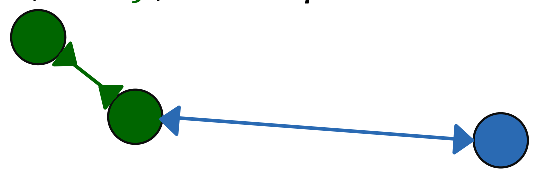
- **Objective function L_φ**
  - **Ranking-based**
    - **Contrastive w/ margin**
    - **Multi-similarity loss**
    - **…**

$$\varphi \leftarrow \underset{\varphi}{\operatorname{argmin}} L_\varphi$$

$$(x_i, x_j, x_k)$$

Anchor *Positive* *Negative*

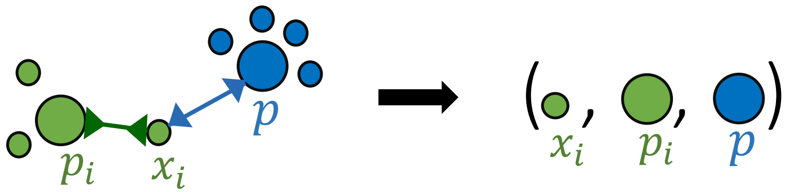
$$L_\varphi = \left[D_\varphi\left(x_i, x_j\right) - D_\varphi(x_i, x_k) + \gamma\right]_+$$



[Wu et al., ICCV’17],
[Wang et al., CVPR’19]

Main Topics in DML

- **Objective function**
  - Ranking-based
  - **Proxy-based**
    - **ProxyNCA**

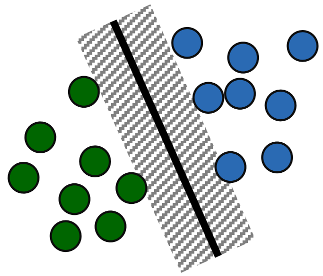


$$L = \log \frac{\exp -D_{\varphi}(x_i, p_i)}{\sum_{p \in P \setminus \{p_i\}} \exp -D_{\varphi}(x_i, p)}$$

[Movshovitz-Attias et al., ICCV'17],
[Goldberger et al., NIPS'04],
[Kim et al. CVPR'20],
[Qian et al., ICCV'19]

Main Topics in DML

- **Objective function**
  - Ranking-based
  - Proxy-based
  - **Classification-based**



$$L = \sum_{j \neq i} \max[0, D_{\varphi}(x_i, x_j) + \gamma]$$

[Deng et al., et al., CVPR'19],
[Liu et al. CVPR'17],
[Liu et al. ICML'16],
[Wang et al. CVPR'18]

Main Topics in DML

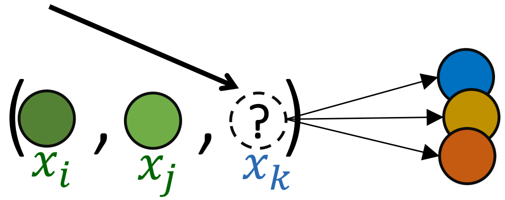
- Objective function
- **Sampling matters**
  - Local (mini-batch) vs. global mining
  - (Semi-)Hard-negatives
  - Hardness-aware
  - Easy positives
  - Adversarial negative synth.

  ...

Cannot train on all $\mathcal{O}(N^3)$ triplets

⇨ **Define sampling distrib.**

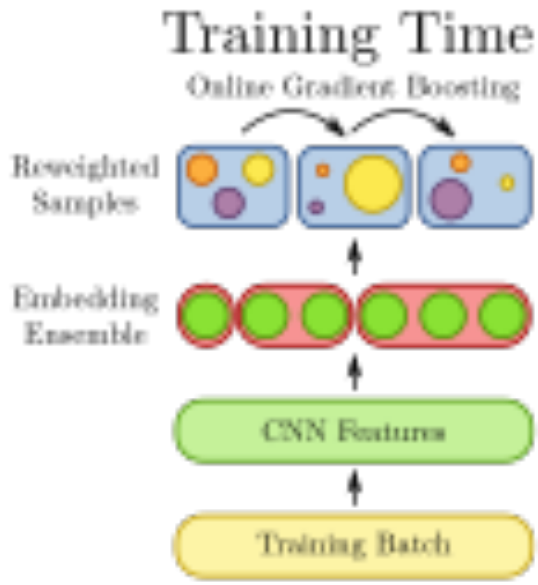
$$p\left(x_k \middle| x_i, x_j, y_i = y_j \neq y_k\right)$$



[Wu et al., ICCV'17],
[Huang et al. ECCV'18],
[Harwood et al., ECCV'17],
[Iscen etal., CVPR'18]

Main Topics in DML

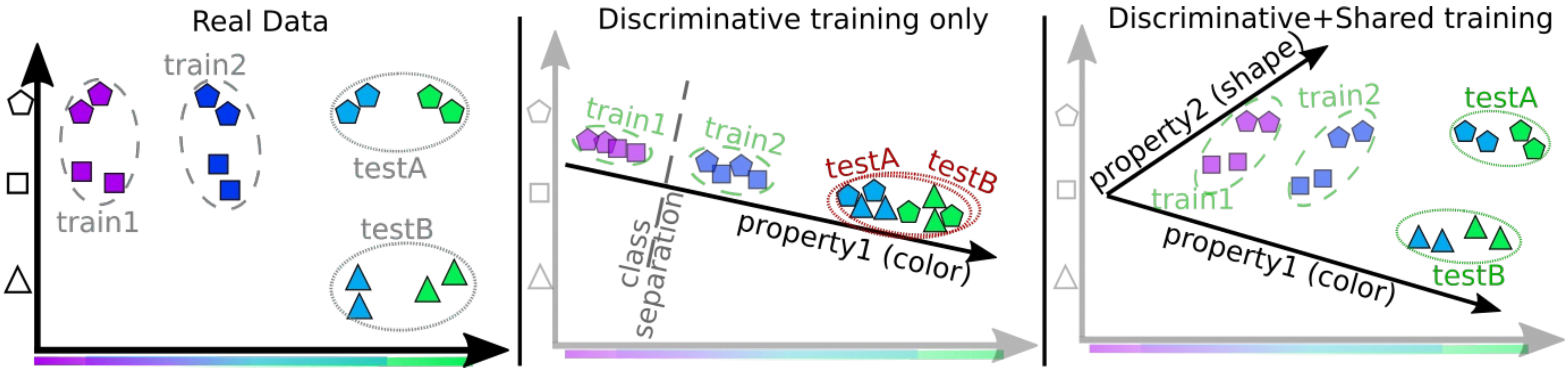
- Objective function
- Sampling matters
- **Ensemble methods**
  - Combining multiple (local) embeddings



[Freund, Schapire, JCSS'97],
[Guo, Gould, arXiv:1506.07224],
[Opitz et al., ICCV'17],
[Yuan et al., ICCV'17],
[Sanakoyeu et al. PAMI.2021.3113270]

Main Topics in DML

- Objective function
- Sampling matters
- Ensemble methods
- **Generalization**



[Sharing Matters for Generalization in Deep Metric Learning, PAMI 2020], [Characterizing generalization under out-of-distribution shifts in deep metric learning, NeurIPS'21]

Metric Learning: Summary

- Similarity measures basis for numerous CV&ML tasks
- Learning richly structured, low-dim embeddings: fine-grained relationships
- Metric learning:
  - Linear, kernelized, non-linear with neural network
- DML main direction:
  - Objective function
  - Sampling strategies
  - Ensemble methods
  - Generalization
- Capturing semantic similarity: holy grail of CV&ML