CS-C3240 - Machine Learning

Model Regularization

Data Augmentation. Soft Model-Selection. Transfer Learning. Multi-Task Learning. Semi-Supervised Learning.

Alexander Jung

What I want to teach you today:

- basic idea of regularization
- regularization as soft model selection
- basic idea of data augmentation
- equivalence between regularization and data aug.

What is ML ?

informal: learn hypothesis out of a hypothesis space or "model" that incurs minimum loss when predicting labels of datapoints based on their features

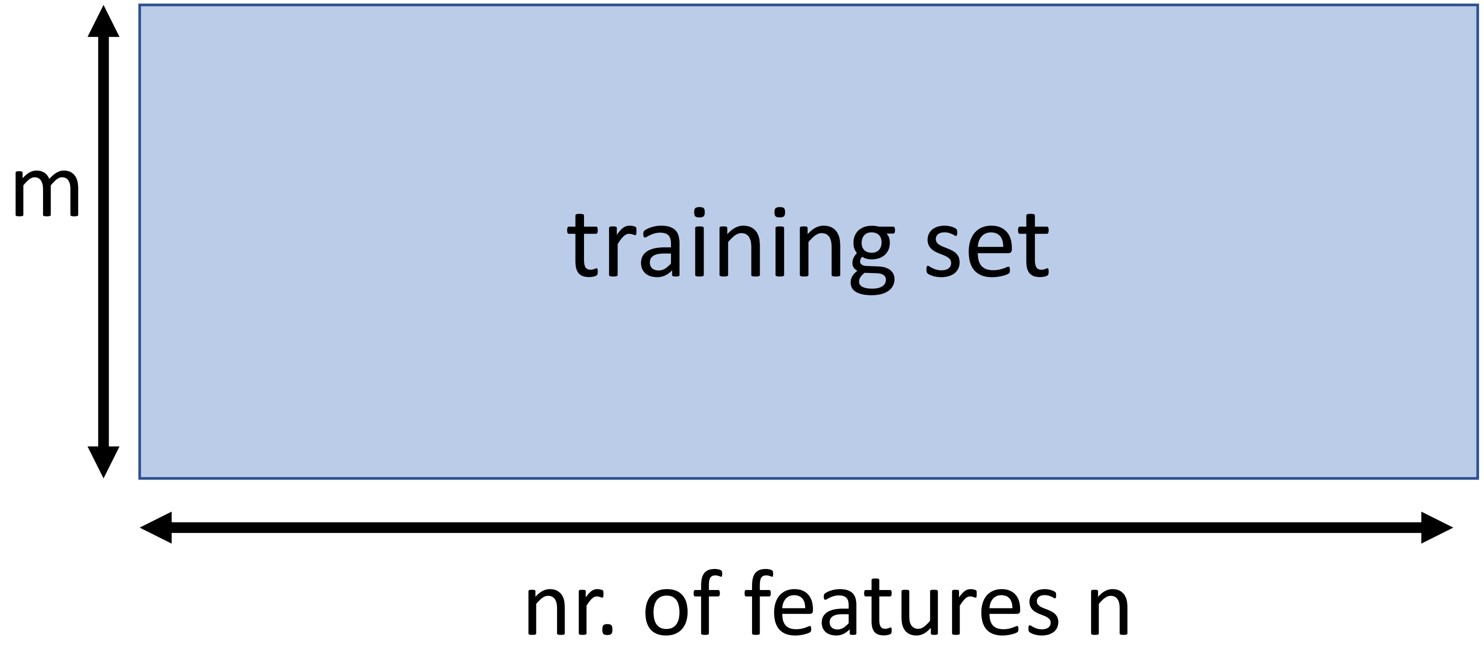
"training error"

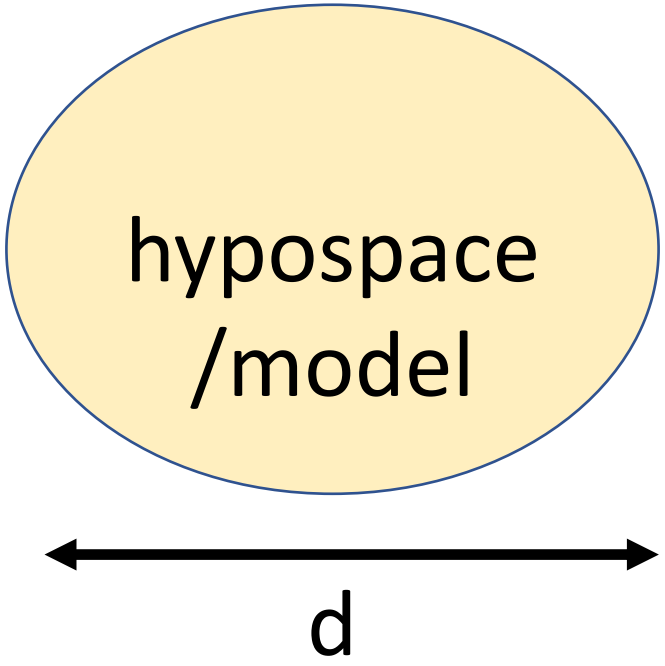
$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \, \mathcal{E}(h|\mathcal{D})$$

$$\stackrel{(\boxed{2.12})}{=} \underset{h \in \mathcal{H}}{\operatorname{argmin}} (1/m) \sum_{i=1}^{m} \mathcal{L}((\mathbf{x}^{(i)}, y^{(i)}), h).$$

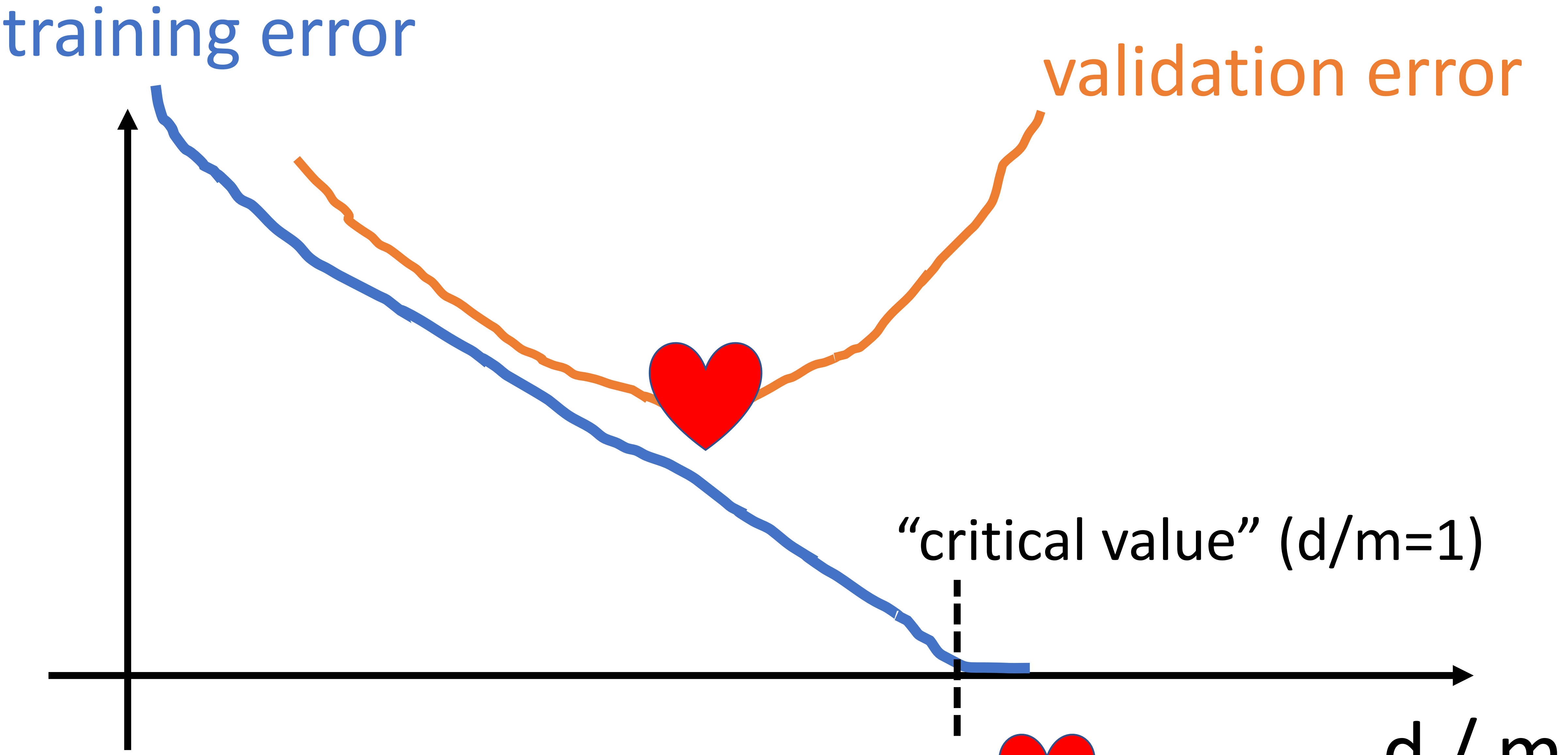
see Ch. 4.1 of mlbook.cs.aalto.fi

Data and Model Size





crucial parameter is the ratio d/m



adjust model and/or data to reach

bring d/m below critical value 1:

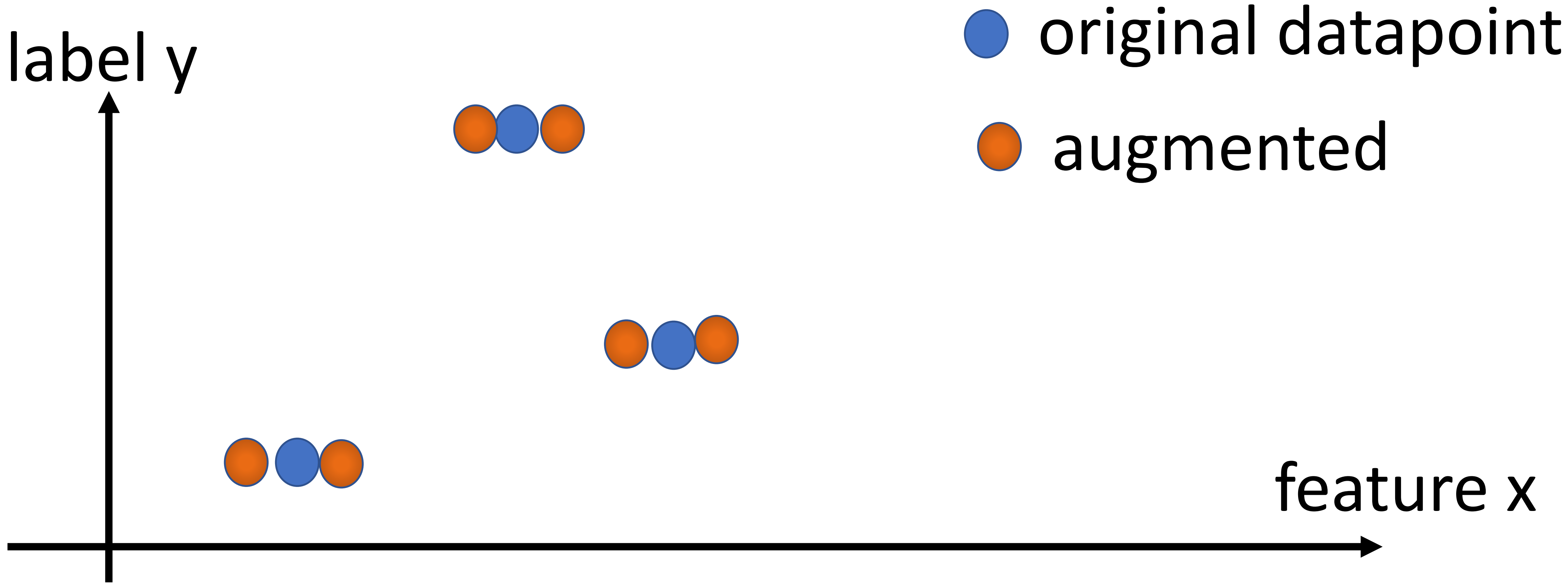
- increase m by using more training data
- decrease d by using smaller hypothesis space

bring d/m below critical value 1:

- increase m by using more training data
- decrease d by using smaller hypothesis space

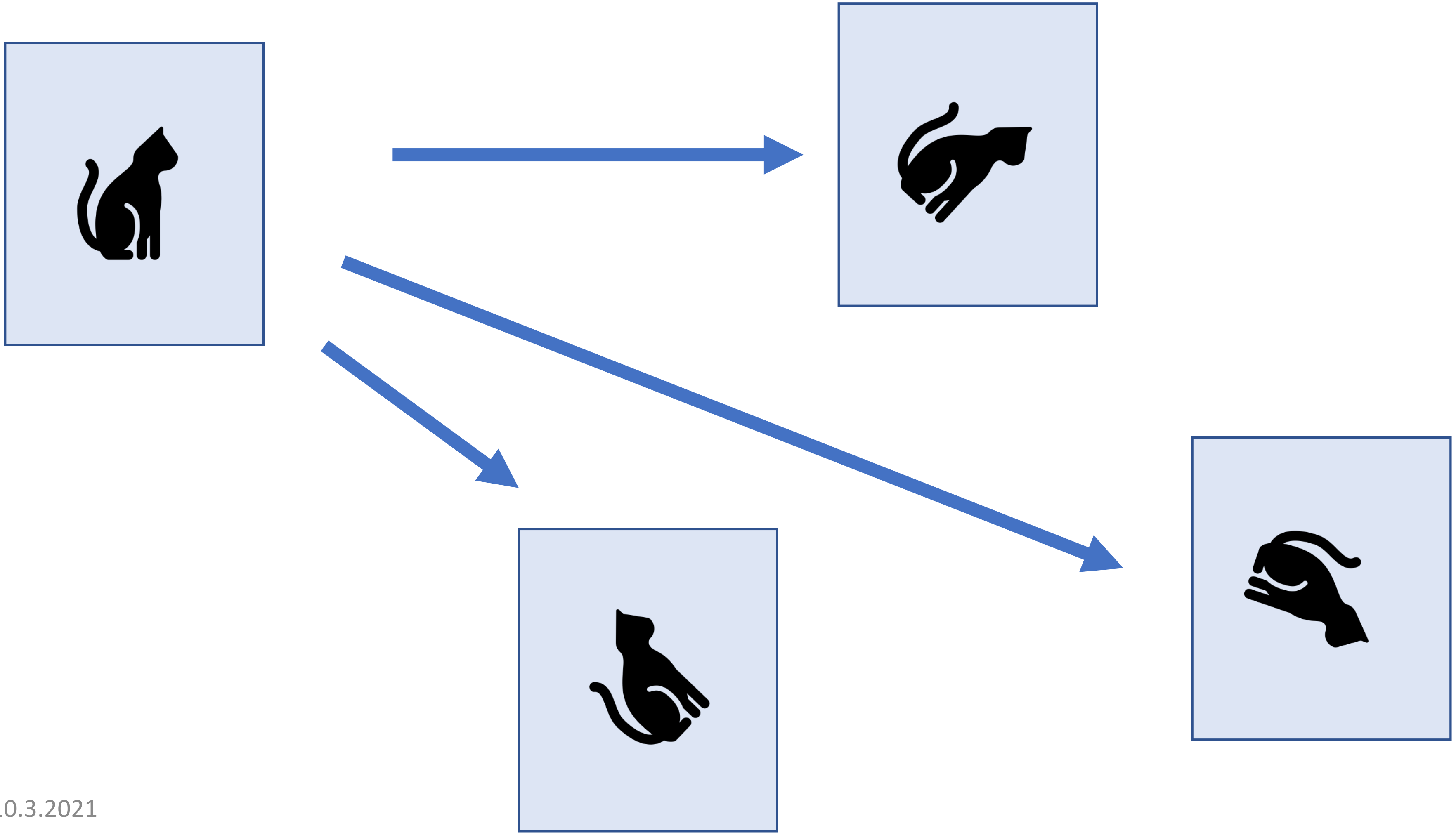
Data Augmentation

add a bit of noise to features

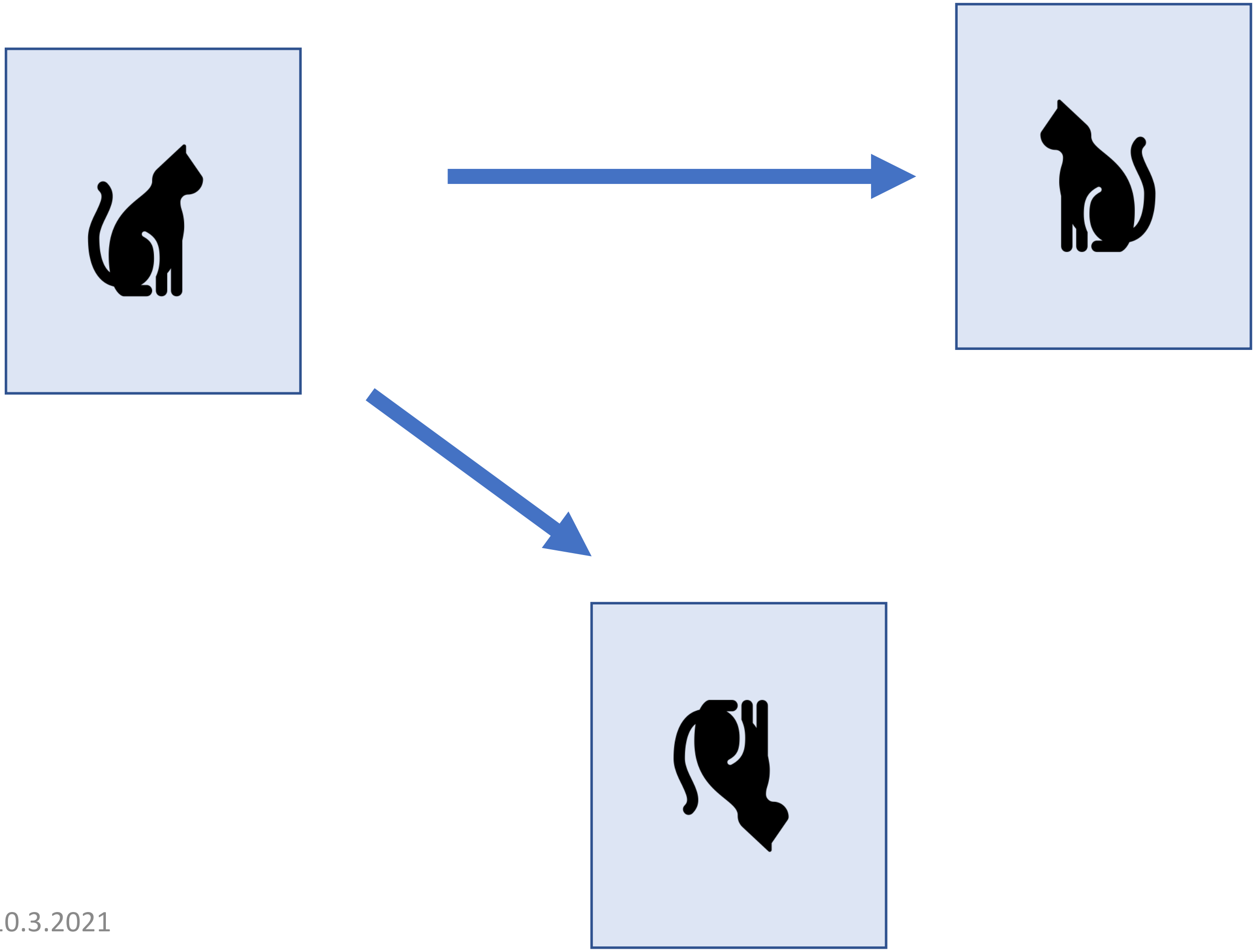


we have increased the dataset by factor 3 !

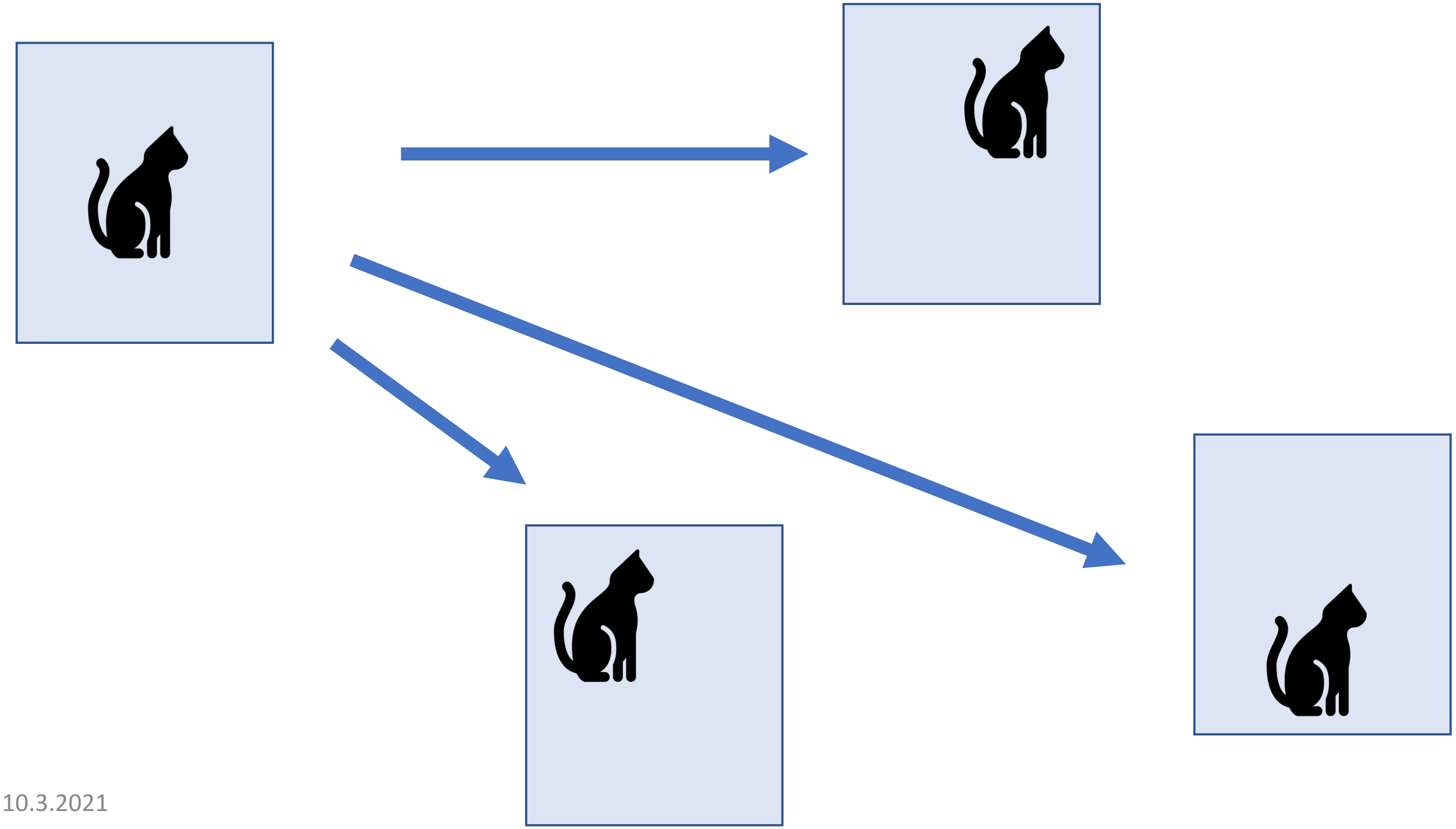
rotated cat image is still cat image



flipped cat image is still cat image



shifted cat image is still cat image



bring d/m below critical value 1:

- increase m by using more training data
- decrease d by using smaller hypothesis space

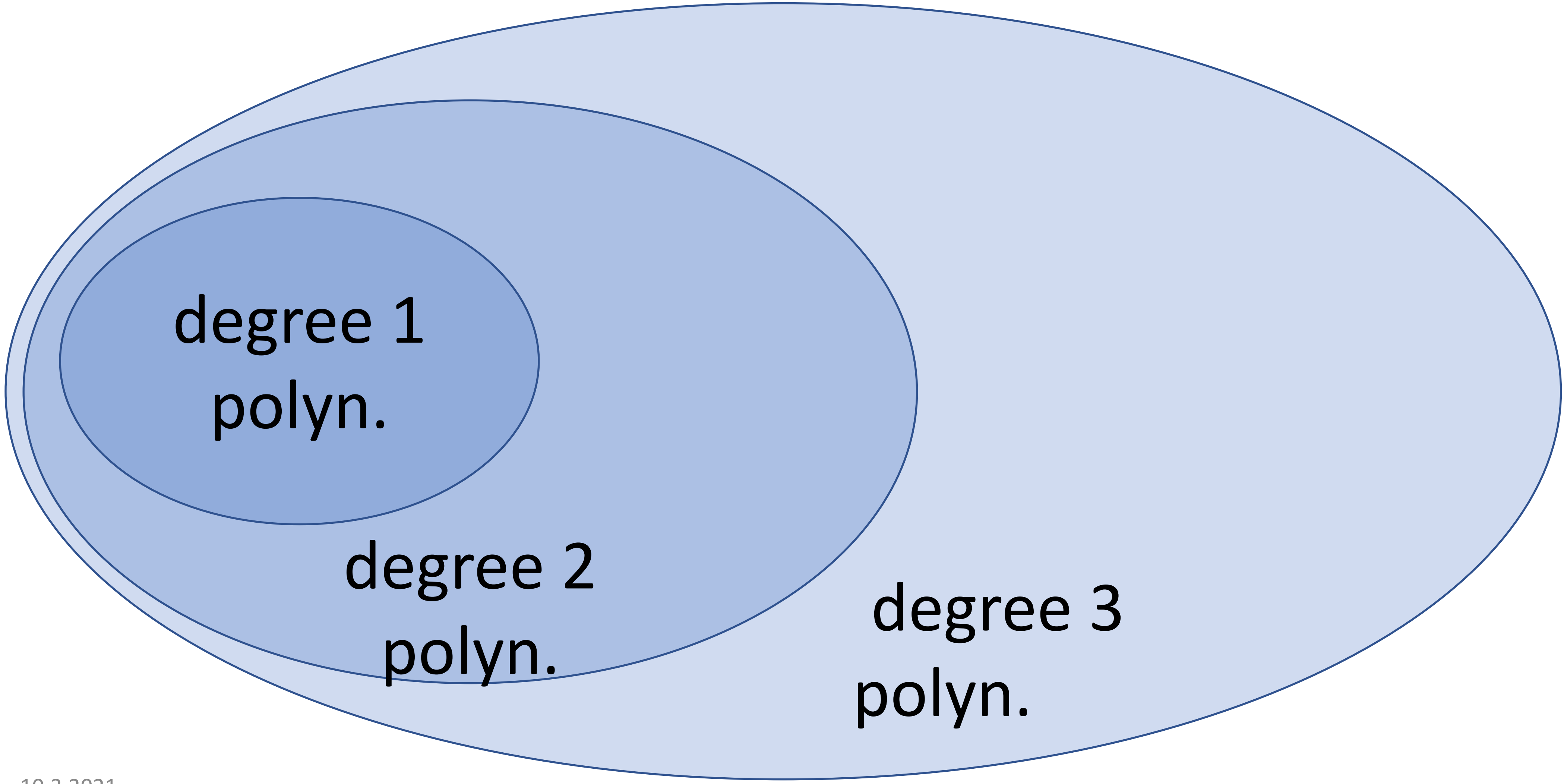
replace original ERM

$$\min_{h \epsilon \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(\left(x^{(i)}, y^{(i)}\right), h\right)$$

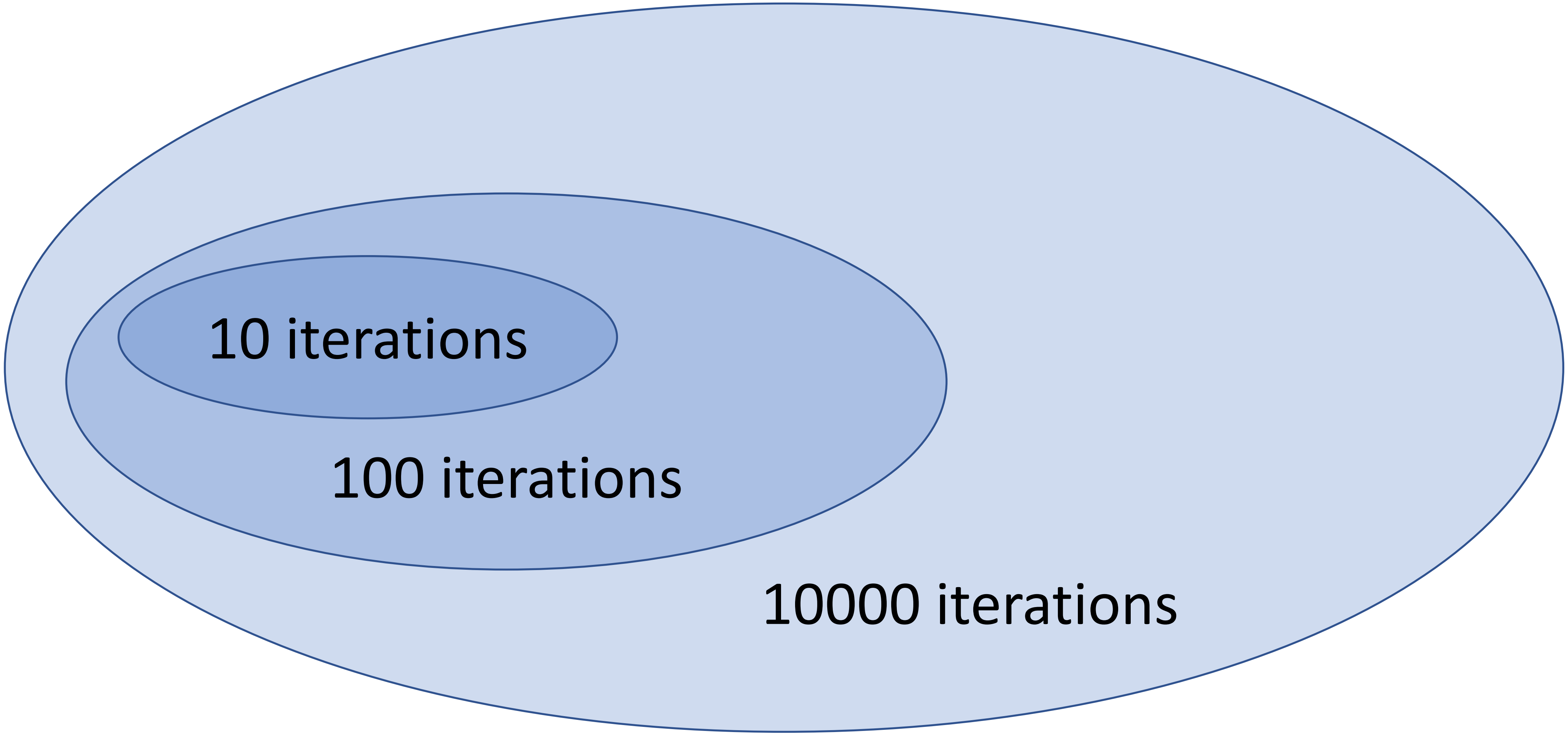
with ERM on smaller $\widehat{\mathcal{H}} \subset \mathcal{H}$

$$\min_{h \epsilon \widehat{\mathcal{H}}} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(\left(x^{(i)}, y^{(i)}\right), h\right)$$

Nested Models



Prune Hypospace by Early Stopping



Soft Model Pruning via Regularization

Regularized ERM

learn hypothesis h out of
model (hypospace) $\mathcal{H}$ by minimizing

$$\underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(\left(x^{(i)},y^{(i)}\right),h\right)}_{\text{average loss on training set (empirical risk of h)}}+\underbrace{\lambda\mathcal{R}(h)}_{\text{loss increase for datapoints outside training set}}$$

average loss on training set
(empirical risk of h)

loss increase for datapoints
outside training set

Regularized Linear Regression

- squared error loss
- linear hypothesis map $h(x) = w^T x = w_1 x_1 + \cdots + w_n x_n$

$$\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - w^T x^{(i)}\right)^2 + \lambda\mathcal{R}(w)$$

- ridge regression uses $\mathcal{R}(w) = \|w\|_2^2 = w_1^2 + \cdots + w_n^2$
- Lasso uses $\mathcal{R}(w) = \|w\|_1 = |w_1| + \cdots + |w_n|$

Regularization = Implicit Pruning!

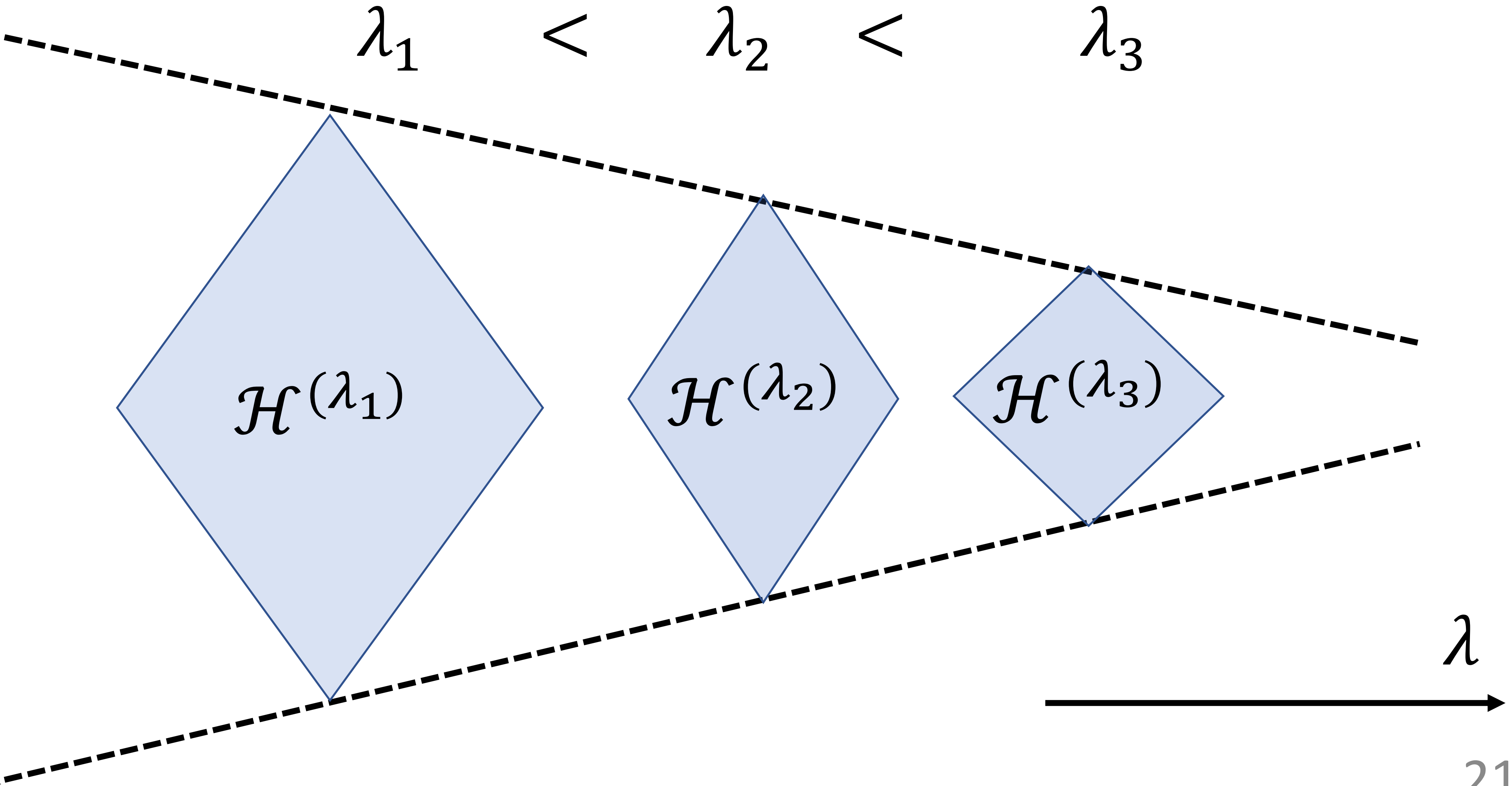
$$\min_{h\epsilon\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(\left(x^{(i)}, y^{(i)}\right), h\right) + \lambda\mathcal{R}(h)$$

equivalent to

$$\min_{h\epsilon\mathcal{H}^{(\lambda)}} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(\left(x^{(i)}, y^{(i)}\right), h\right)$$

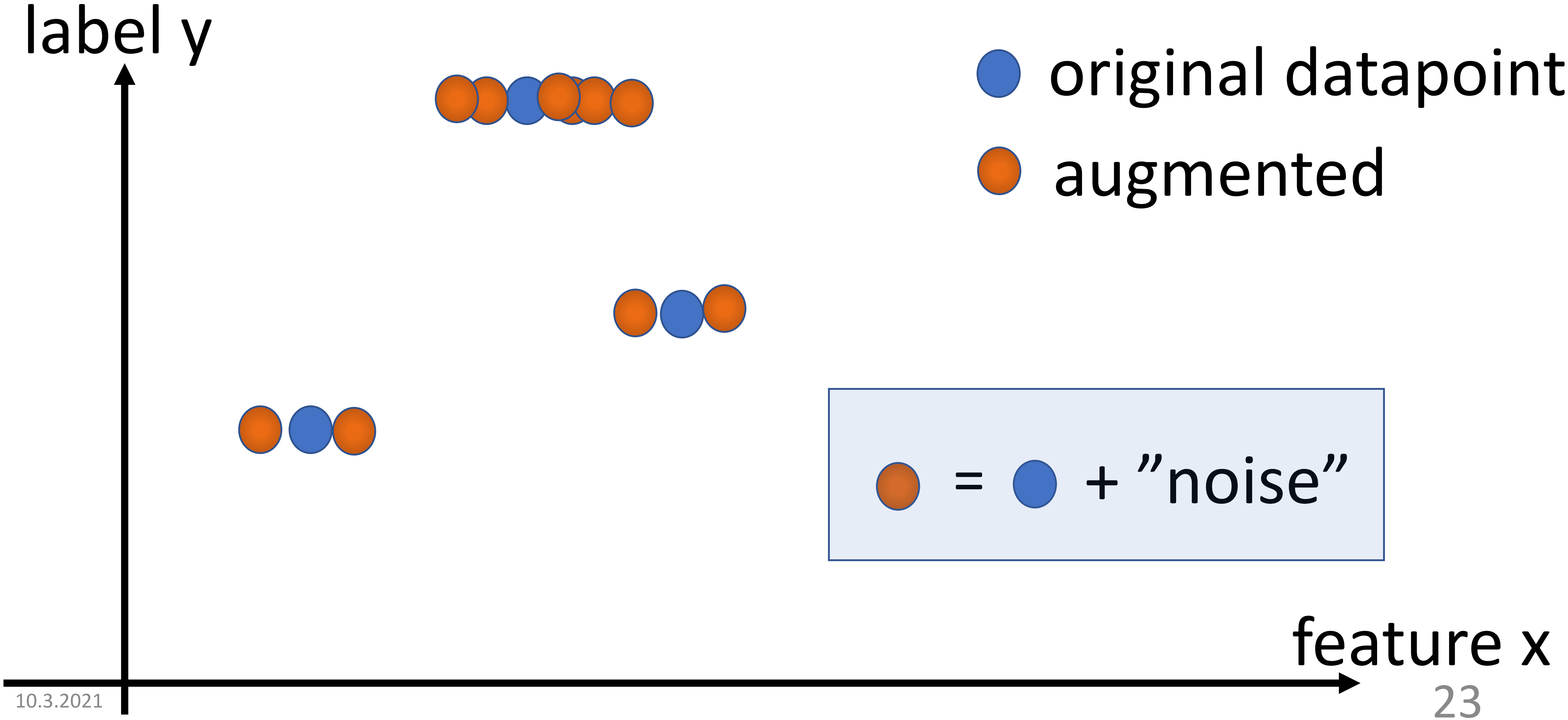
with pruned model $\mathcal{H}^{(\lambda)} \subset \mathcal{H}$

Regularization = “Soft” Model Selection

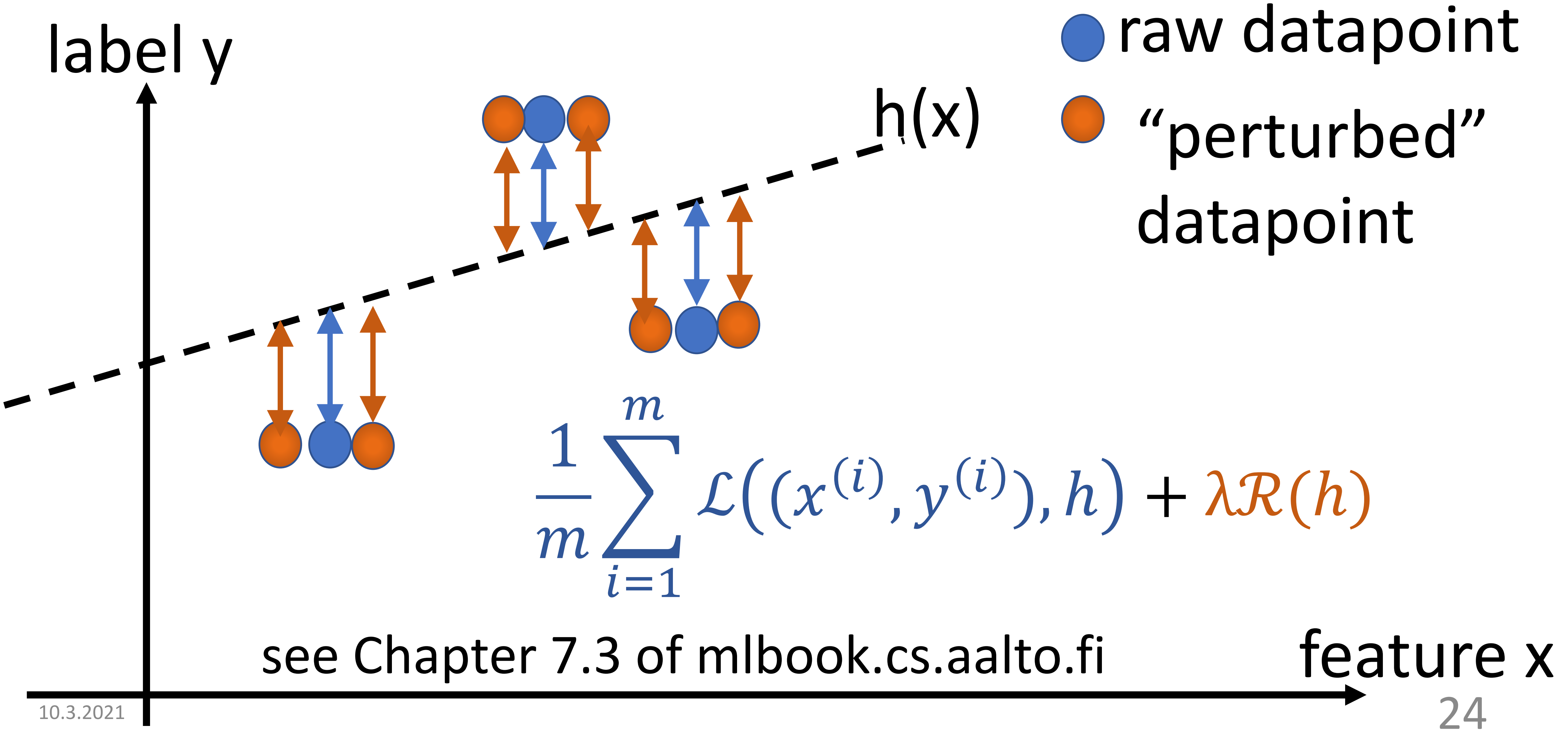


Regularization does implicit Data Augmentation

augment with (infinitely many) realizations of RV!



Regularization =Implicit Data Aug.

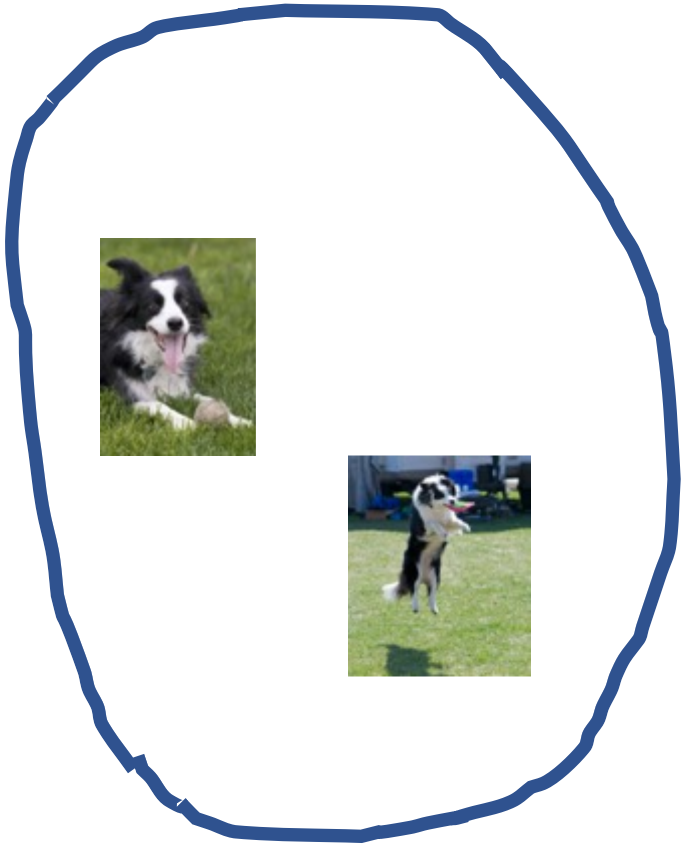


$$\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(\left(x^{(i)},y^{(i)}\right),h\right)+\lambda\mathcal{R}(h)$$

see Chapter 7.3 of mlbook.cs.aalto.fi

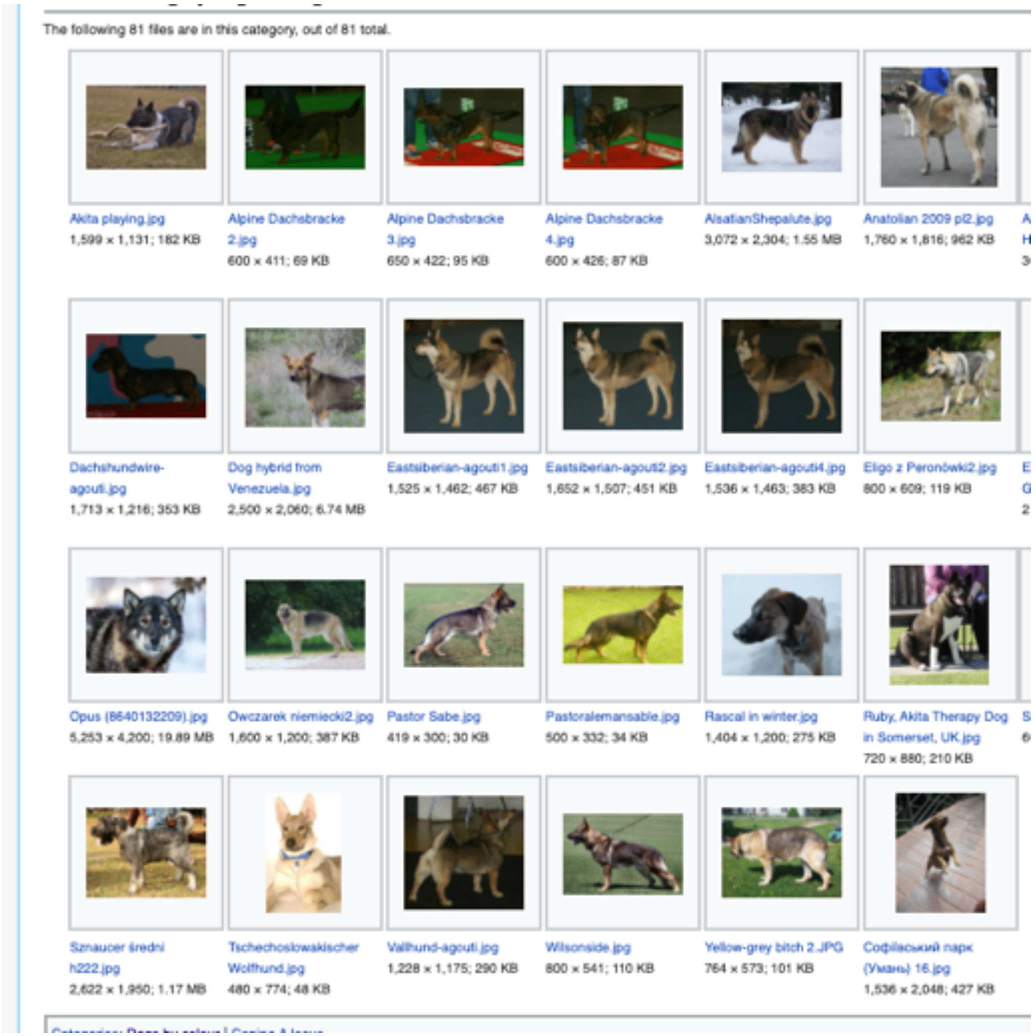
Transfer Learning via Regularization

- Problem I: classify image as “shows border collie” vs. “not”
- Problem II: classify image as “shows a dog” vs. “not”
- ML Problem I is our main interest
- only little training data $\mathcal{D}^{(1)}$ for Problem I
- much more labeled data $\mathcal{D}^{(2)}$ for Problem II
- pre-train a hypothesis on $\mathcal{D}^{(2)}$, fine-tune on $\mathcal{D}^{(1)}$



$\mathcal{D}^{(1)}$

learn h by fine-tuning $\hat{h}$



$\mathcal{D}^{(2)}$

pre-train hypothesis $\hat{h}$

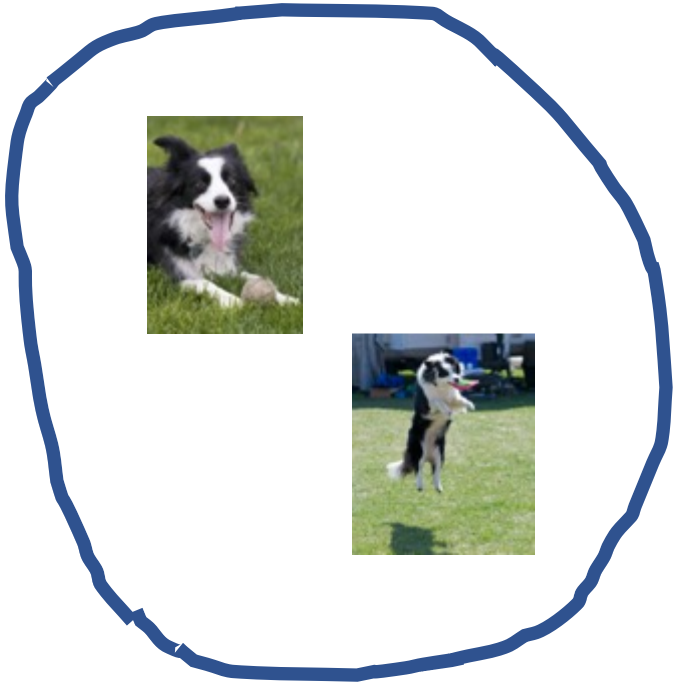
$$\underbrace{\min_{h \epsilon \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(\left(x^{(i)}, y^{(i)}\right), h\right)}_{\text{fine tuning on } \mathcal{D}^{(1)}} + \underbrace{\lambda d(h, \hat{h})}_{\text{distance to hypothesis } \hat{h} \text{ which is pre-trained on } \mathcal{D}^{(2)}}$$

fine tuning on $\mathcal{D}^{(1)}$

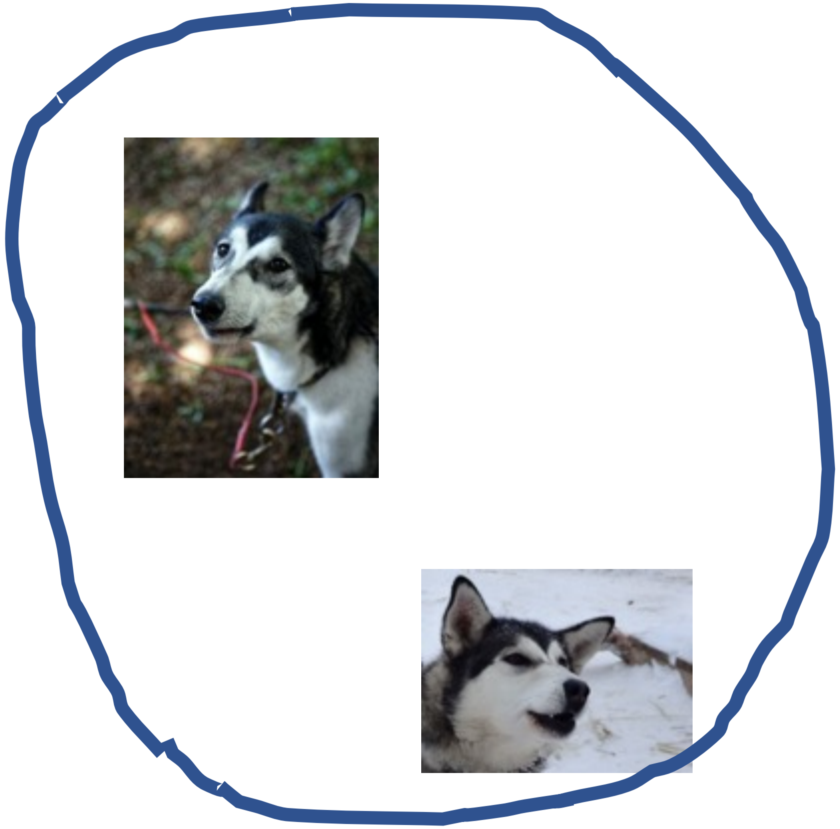
distance to hypothesis $\hat{h}$ which is pre-trained on $\mathcal{D}^{(2)}$

Multi-Task Learning via Regularization

- Problem I: classify image as “shows border colly” vs. “not”
- Problem II: classify image as “shows husky” vs. “not”
- training data $\mathcal{D}^{(1)}$ for Problem I and $\mathcal{D}^{(2)}$ for Problem II
- jointly learn hypothesis $h^{(1)}$ on $\mathcal{D}^{(1)}$ and $h^{(2)}$ on $\mathcal{D}^{(2)}$
- require $h^{(1)}$ to be “similar” to $h^{(2)}$



$\mathcal{D}^{(1)}$



$\mathcal{D}^{(2)}$

jointly learn similar
$h^{(1)}$ and $h^{(2)}$ for each dataset

training error of $h^{(1)}$

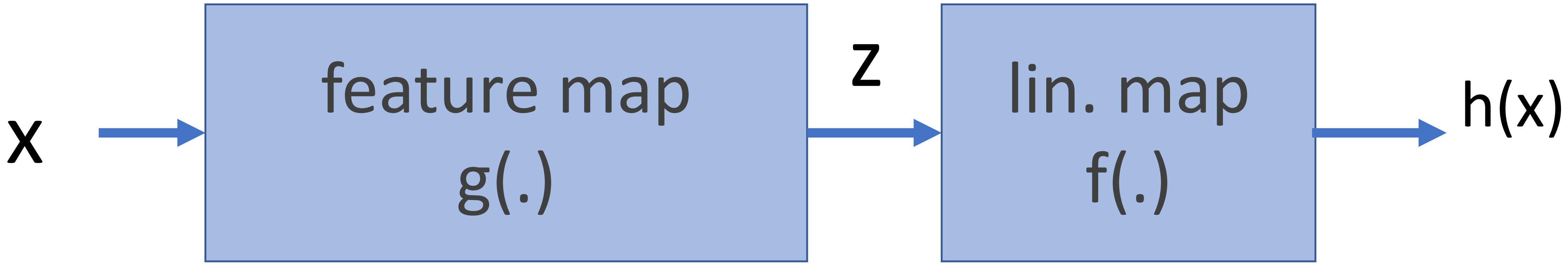
training error of $h^{(2)}$

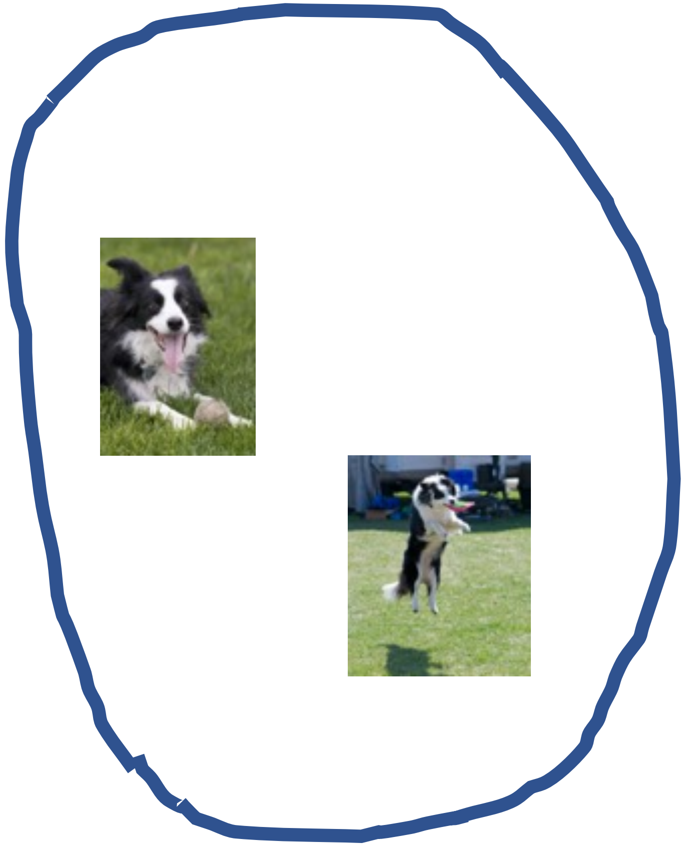
$$\min_{h^{(1)},h^{(2)}} \mathcal{E}\left(h^{(1)} \middle| \mathcal{D}^{(1)}\right) + \mathcal{E}\left(h^{(2)} \middle| \mathcal{D}^{(2)}\right) + \lambda d\left(h^{(1)}, h^{(2)}\right)$$

“distance” between $h^{(1)}$ and $h^{(2)}$

Semi-Supervised Learning via Regularization

- classify image as “shows border colly” vs. “not”
- small labeled dataset $\mathcal{D}^{(1)}$
- massive image database $\mathcal{D}^{(2)}$ with unlabeled images
- train hypothesis h(.) on $\mathcal{D}^{(1)}$ with following structure:





$\mathcal{D}^{(1)}$

learn linear classifier f(.)

$\mathcal{D}^{(2)}$

learn feature map g(.)

$$\underbrace{\min_{h \epsilon \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(\left(x^{(i)}, y^{(i)}\right), h\right)}_{\text{use training error to fine tune f(.)}} + \lambda \, \underbrace{\mathcal{E}\left(g \middle| \mathcal{D}^{(2)}\right)}_{\text{learn feature map g(.) using large unlabeled database } \mathcal{D}^{(2)}}$$

use training error
to fine tune f(.)

learn feature map g(.)
using large unlabeled
database $\mathcal{D}^{(2)}$

To sum up,

- regularization is a soft model pruning
- regularization does implicit data augmentation
- special cases of regularization
  - transfer learning
  - multi-task learning
  - semi-supervised learning

Questions ?