Few-shot learning

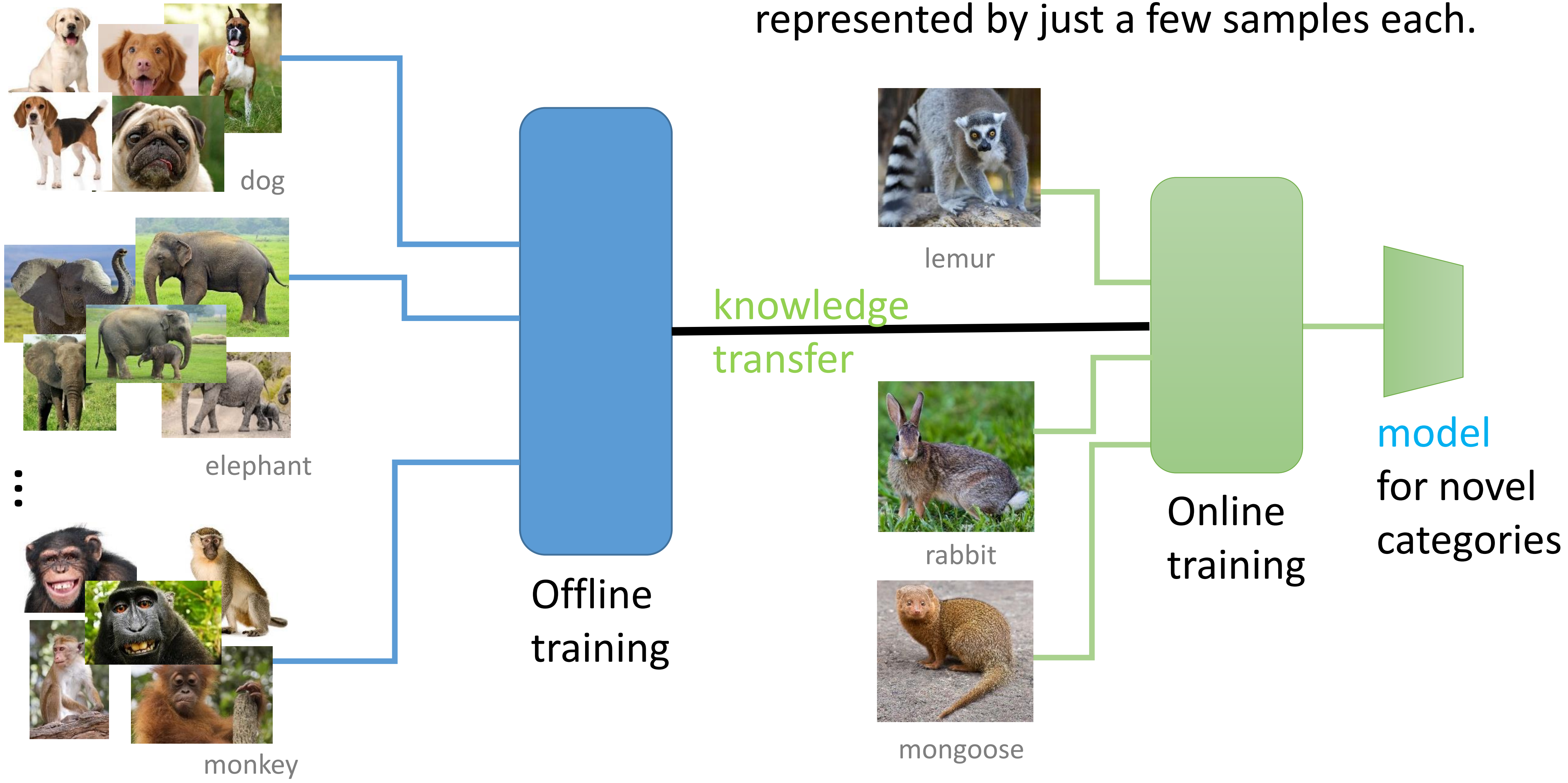
State of the Art

Joseph Shtok
IBM Research AI

The presentation is available at http://www.research.ibm.com/haifa/dept/imt/ist_dm.shtml

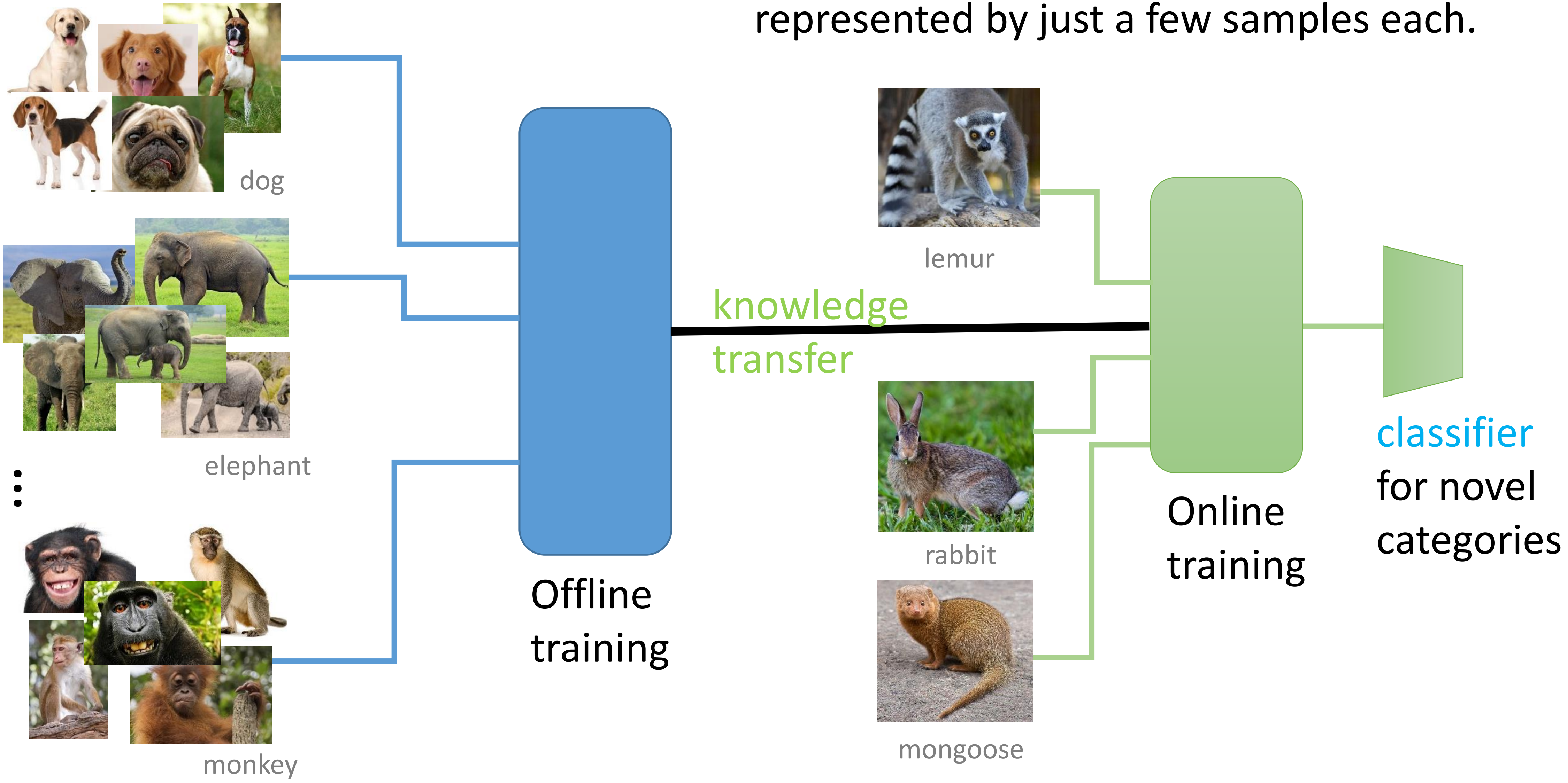
Problem statement

Using a large annotated offline dataset, perform given task for novel categories, represented by just a few samples each.



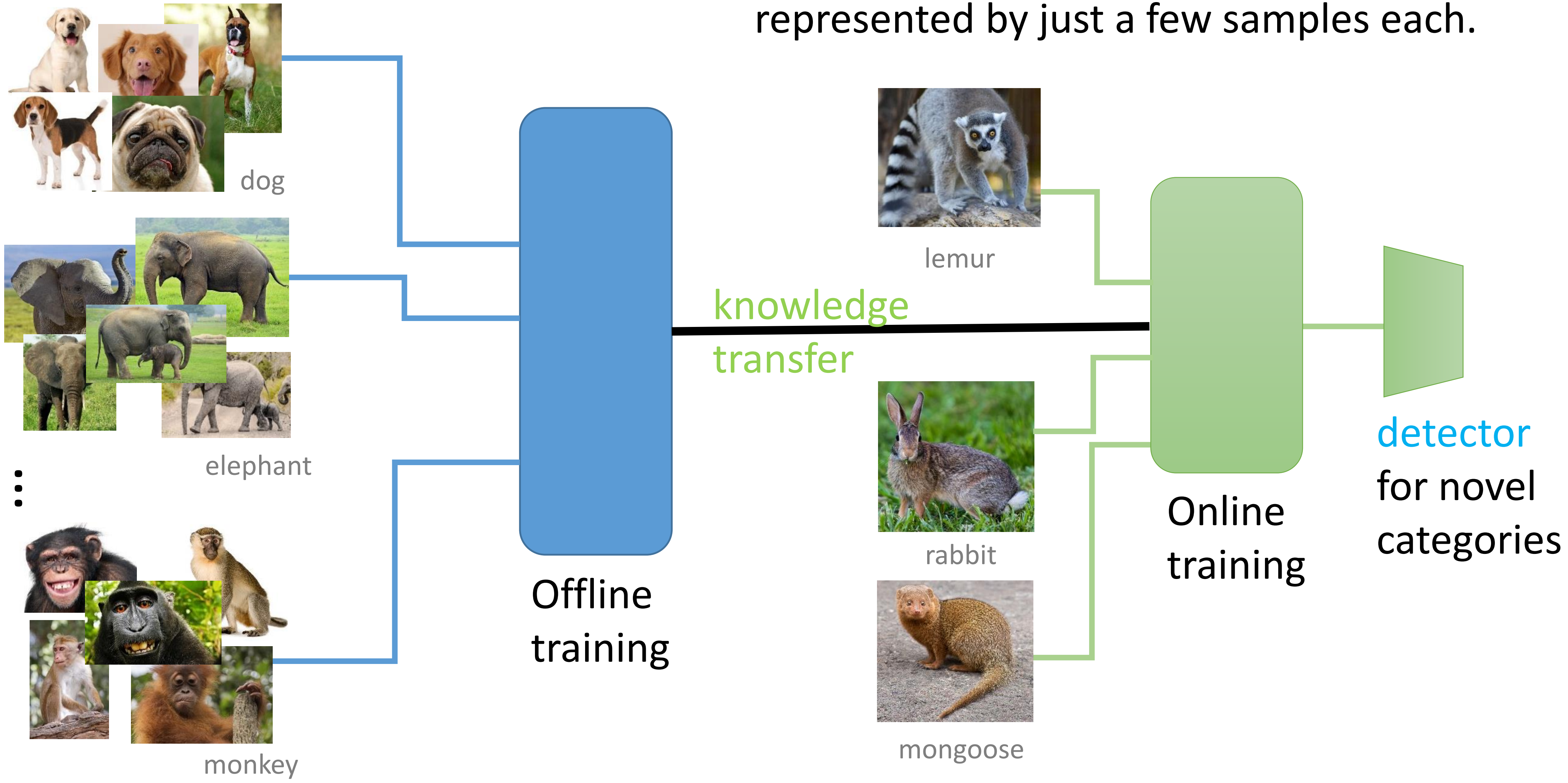
Problem statement

Using a large annotated offline dataset, perform classification for novel categories, represented by just a few samples each.



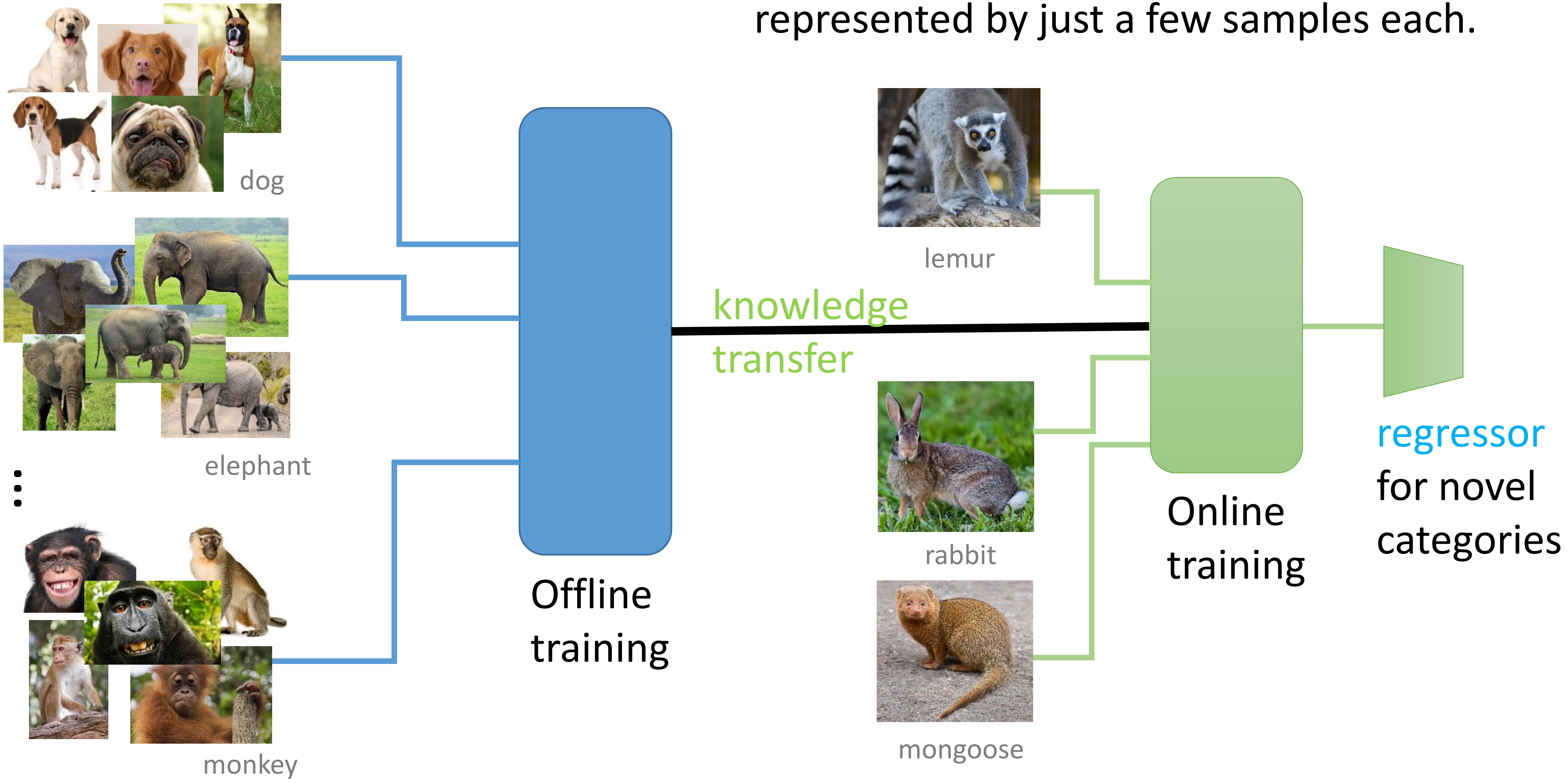
Problem statement

Using a large annotated offline dataset, perform detection for novel categories, represented by just a few samples each.



Problem statement

Using a large annotated offline dataset, perform regression for novel categories, represented by just a few samples each.



Why work on few-shot learning?

IBM

1. It brings the DL closer to real-world business use cases.

2. It involves a bunch of exciting cutting-edge technologies.

- Companies hesitate to spend **much** ... **ey** on a ... for a solution that th...
- Relevant objects are cont... d with new ones. DL has to be agile.

Meta-learning methods

Data synthesizers

Neural Turing Machines

GANs

Semantic metric spaces

Graph neural networks

Networks generating networks

Meta-learning

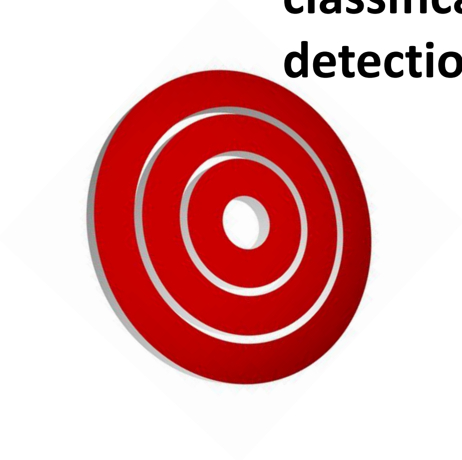
Learn a learning strategy to adjust well to a new few-shot learning task

Metric learning

Learn a \`semantic\` embedding space using a distance loss function

Few-shot learning

Learn to perform classification, detection, regression

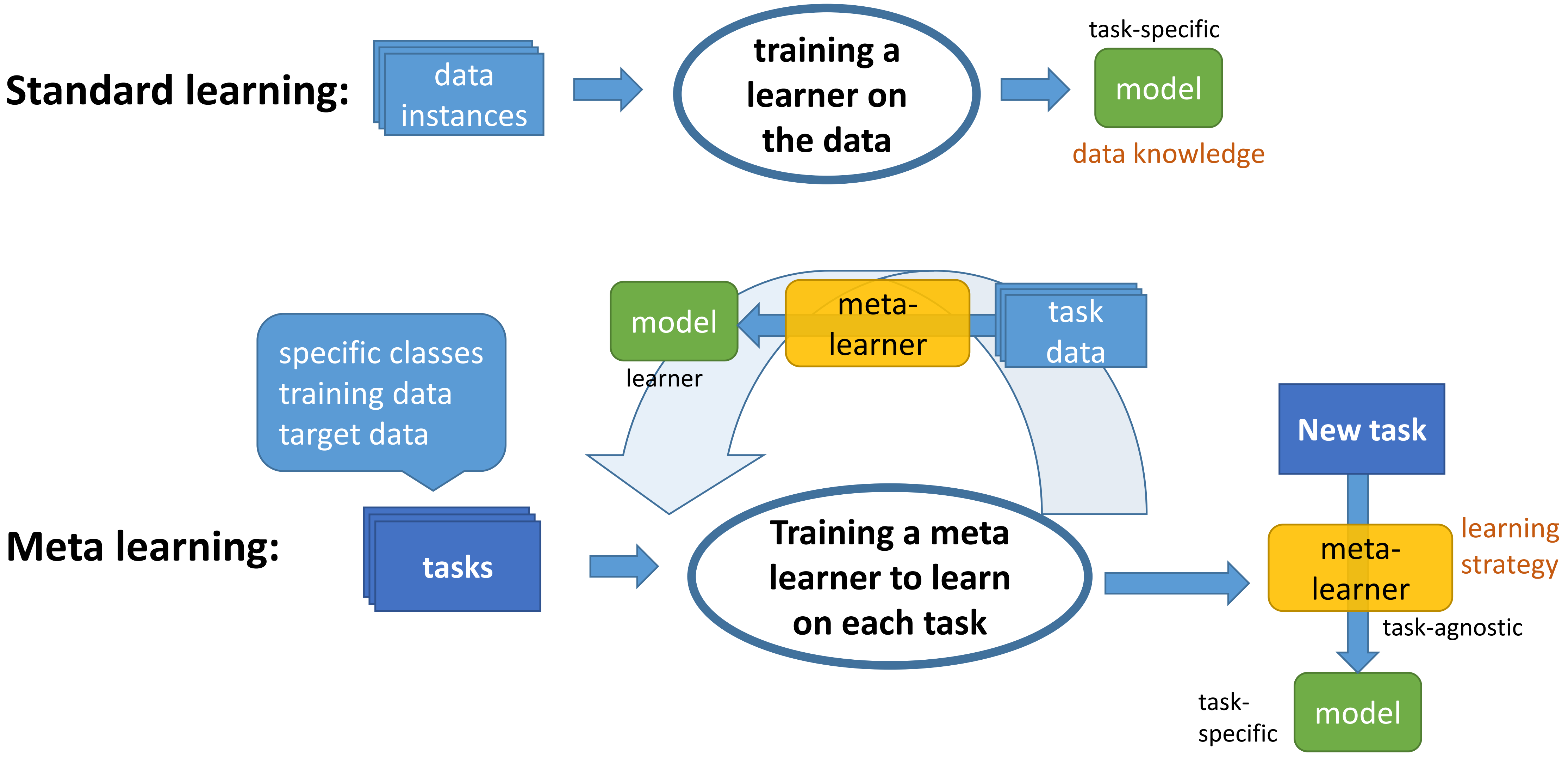


Each category is represented by just a few examples

Data augmentation

Synthesize more data from the novel classes to facilitate the regular learning

Meta-Learning



Recurrent meta-learners

Matching Networks in Vinyals et.al., NIPS 2016

Distance-based classification: based on similarity between the query and support samples in the embedding space *(adaptive metric)*:

$$\hat{y} = \sum_i a(\hat{x}, x_i) y_i \,, \qquad a(\hat{x}, x_i) = similarity\big(f(\hat{x}, S), g(x_i, S)\big)$$

to be elaborated later

f, g - LSTM embeddings of x dependent on the support set S

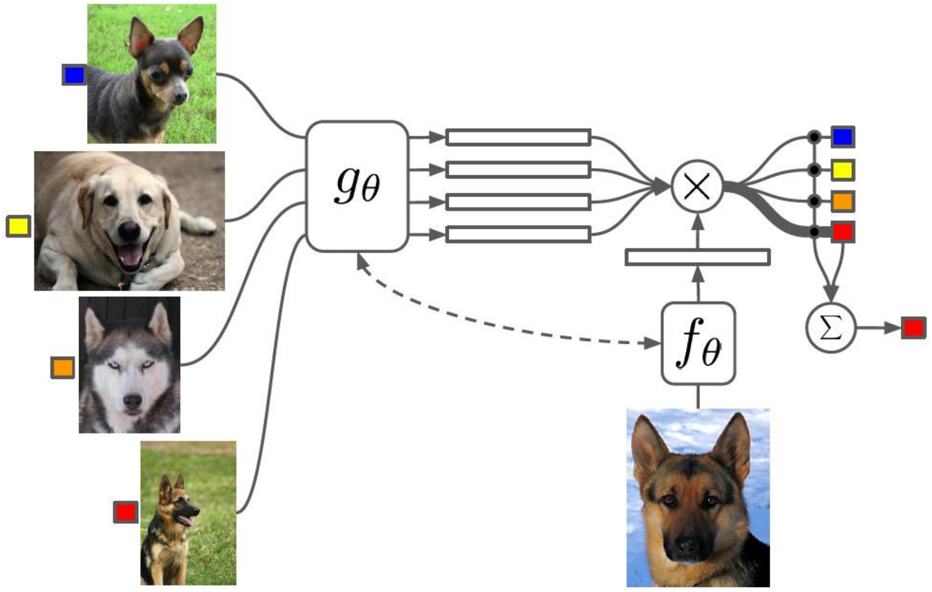


Figure 1: Matching Networks architecture
reprinted from Vinyals et.al., 2016

- Embedding space is class-ag
- LSTM attention mechanism

Method	miniImageNet classification accuracy 1/5 shot
Matching networks	43.56 / 55.31

ncept of episodes: test conditions in the ining.
N new categories
M training examples per category
one query example in {1..N} categories.
Typically, N=5, M=1, 5.

Memory-augmented neural ne
ICML 2016

- Neural Turing Machine = diff
- Learn to predict the distribution $p(y_t | x_t, S_{1:t-1}; \theta)$
- Explicitly store the support samples in the external memory

Optimizers

*Optimize the learner to perform well **after fine-tuning** on the task data done by a single (or few) step(s) of Gradient Descent.*

MAML (Model-Agnostic Meta-Learning) *Finn et.al., ICML 2017*

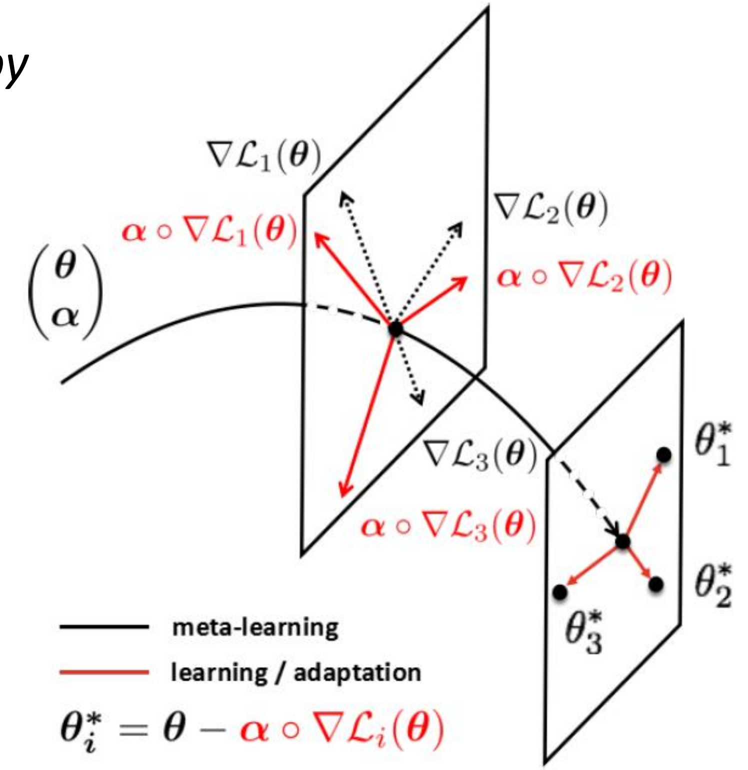
Standard objective (task-specific, for task T):

$$\min_{\theta} \mathcal{L}_T(\theta), \text{ learned via update } \theta' = \theta - \alpha \cdot \nabla_\theta \mathcal{L}_T(\theta)$$

Meta-objective (across tasks):

$$\min_{\theta} \sum_{T\sim p(\mathfrak{I})} \mathcal{L}_T(\theta'), \text{ learn} \ldots \theta'))$$

Method	miniImageNet classification accuracy 1/5 shot
Matching networks	43.56 / 55.31
MAML	48.70 / 63.11
Meta-SGD	54.24 / 70.86



reprinted from Li et.al., 2017

Meta-SGD *Li et.al., 2017*

Remember, α = vector of size θ.

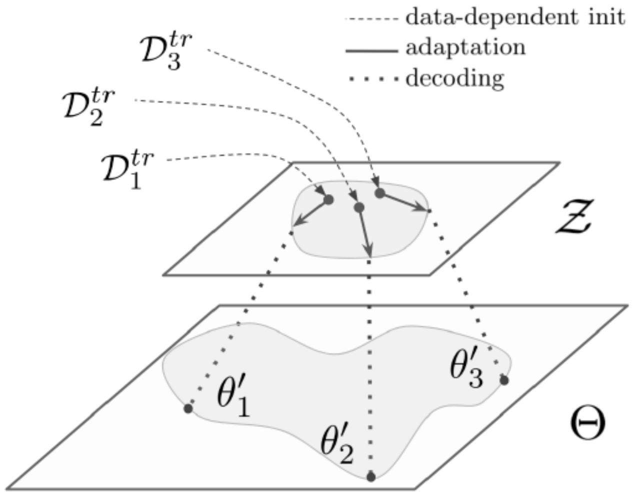
Training objective: minimize the loss $\mathcal{L}_T(\theta')$ with respect to θ = weights initialization, α = update direction and scale, across the tasks.

“Interestingly, the learning process can continue forever, thus enabling life-long learning, and at any moment, the meta-learner can be applied to learn a learner for any new task.”

Optimizers

LEO Rusu et.al., NeurIPS 2018

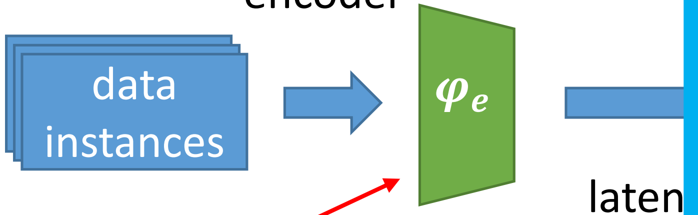
Latent **E**mbedding **O**ptimization: take the optimization problem from high-dim. space of weights $\boldsymbol{\theta}$ to a low-dim. space, for robust Meta-Learning.



Learn a **generative distribution** of model parameters $\boldsymbol{\theta}$,
by learning a stochastic **latent space** with an information bottleneck.

$$\varphi^* = arg\min_{\varphi} \sum_{T\sim p(\mathfrak{I})} \mathcal{L}_T\left(\theta' = g_{\varphi_d}(z')\right), \qquad z' = z - \alpha \cdot \nabla_z \mathcal{L}_T(\theta_z),$$

stochastic
encoder

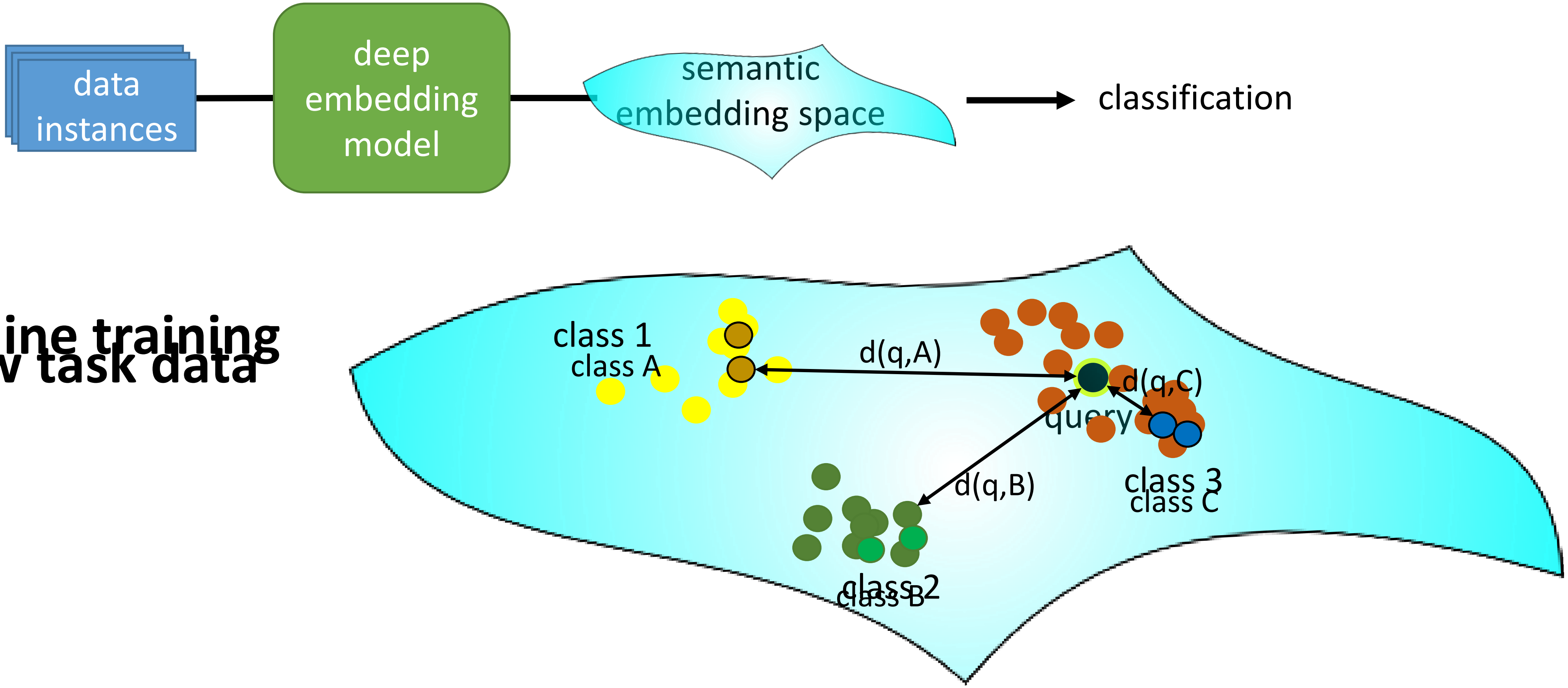


latent
representation

Relation network
encodes distances
between elements of
support set for a new task

Method	miniImageNet classification accuracy 1/5 shot
Matching networks	43.56 / 55.31
MAML	48.70 / 63.11
Meta-SGD	54.24 / 70.86
LEO	**61.76 / 77.59**

Metric Learning



Offline training

New task data

Training: achieve good distributions for offline categories

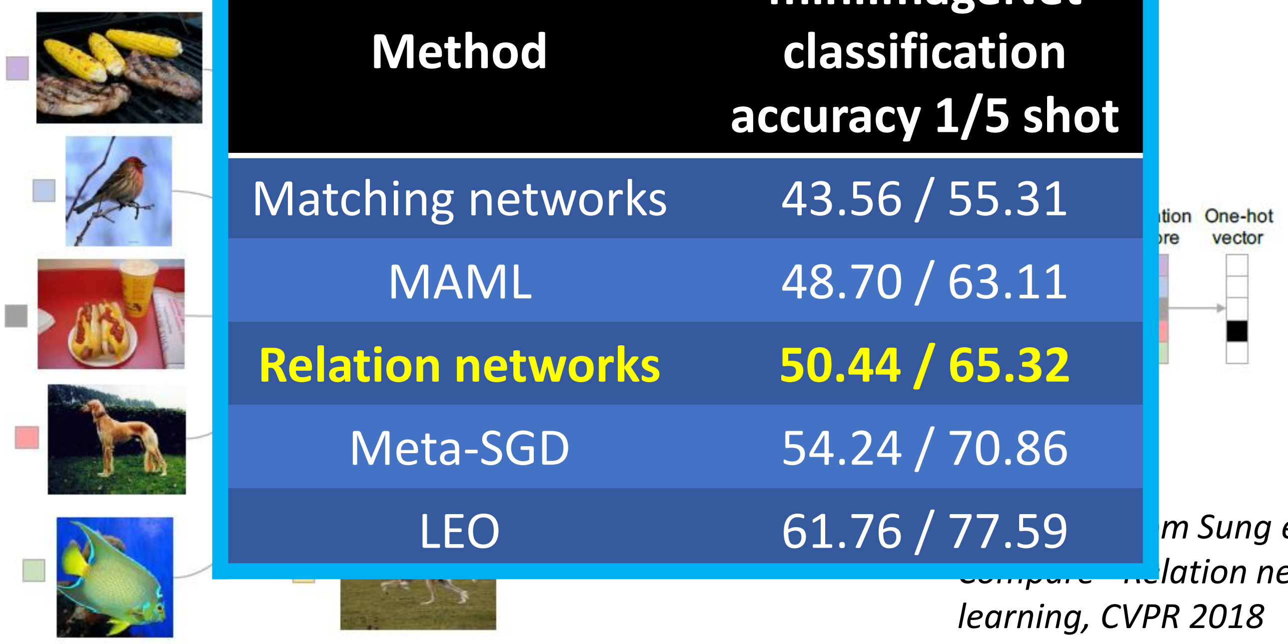
Inference: Nearest Neighbour in the embedding space

Metric Learning

Relation networks, Sung et.al., CVPR 2018

Use the Siamese Networks principle :

- Concatenate embeddings of query and support samples
- Relation module is trained to produces score 1 for correct class and 0 for others
- Extends to zero-shot learning by replacing support embeddings with semantic features.



Method	miniImageNet classification accuracy 1/5 shot
Matching networks	43.56 / 55.31
MAML	48.70 / 63.11
Relation networks	**50.44 / 65.32**
Meta-SGD	54.24 / 70.86
LEO	61.76 / 77.59

...m Sung et.al., Learning to ... elation network for few-shot learning, CVPR 2018

Metric Learning

Matching Networks, Vinyals et.al., NIPS 2016

Objective: maximize the cross-entropy for the non-parametric softmax classifier $\sum_{(x,y)} logP_\theta(y|x,S)$, with

$$P_\theta(y|x,S) = soft$$

Method	miniImageNet classification accuracy 1/5 shot
Matching networks	**43.56 / 55.31**
MAML	48.70 / 63.11
Relation networks	50.44 / 65.32
Prototypical Networks	**49.42 / 68.20**
Meta-SGD	54.24 / 70.86
LEO	61.76 / 77.59

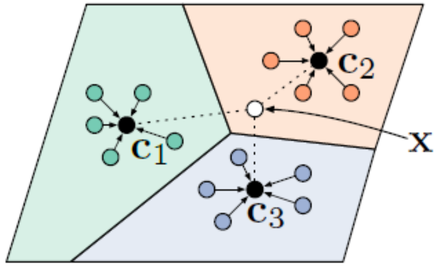
Prototypical Networks, Snell e

Each category is represented b

Objective: maximize the cross-e

probability expression:

$$P_\theta(y|x,C)=s$$



Each category is represented by a single prototype c_i .

Metric Learning

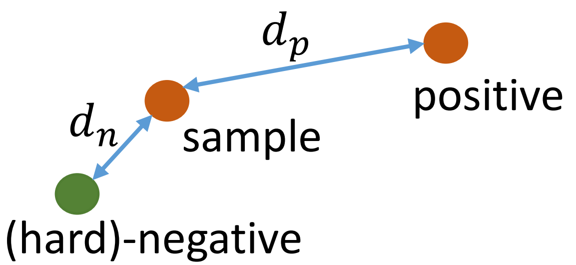
Large Margin Meta-Lea

Regularize the cross-e in
objective:

$$\mathcal{L} =$$

Method	miniImageNet classification accuracy 1/5 shot
Matching networks	43.56 / 55.31
MAML	48.70 / 63.11
Relation networks	50.44 / 65.32
Prototypical Networks	49.42 / 68.20
Large-margin	**51.08 / 67.57**
Meta-SGD	54.24 / 70.86
LEO	61.76 / 77.59

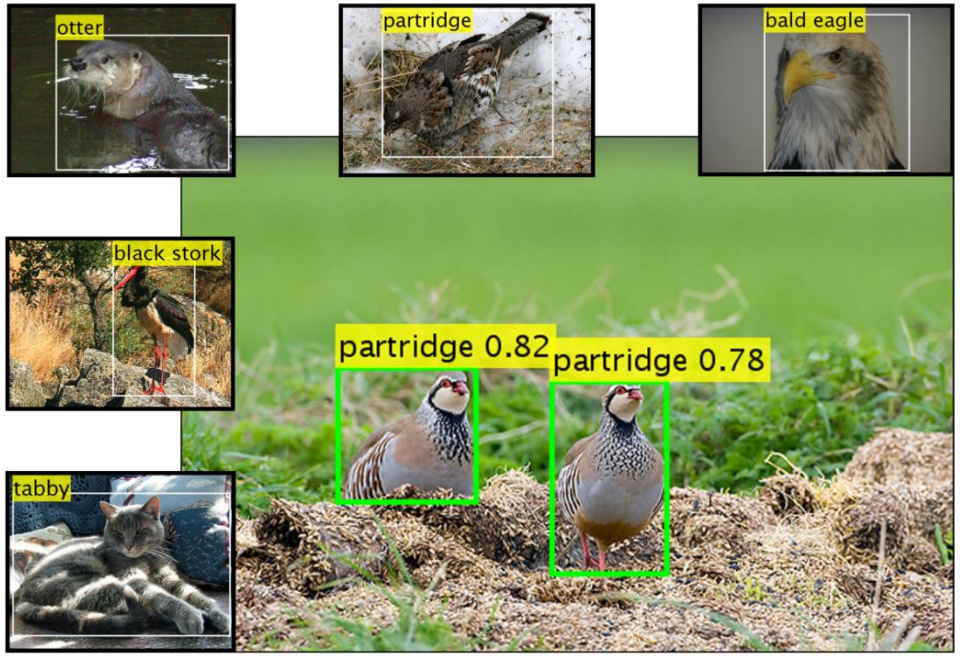
Distances for triplet loss



RepMet: Few-shot det

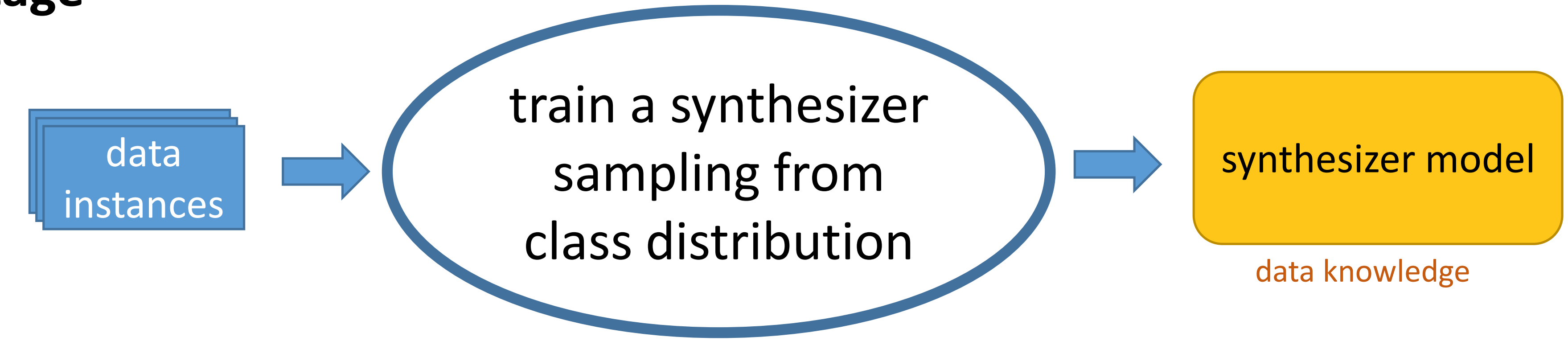
- Equip a standard
- Introduce class re
- Learn an embedding space using the objective

$$\mathcal{L} = \mathcal{L}_{CE} + \left|\min_j d(E, R_{ij}) - \min_{j,k\neq i} d(E, R_{kj}) + \alpha\right|_+$$

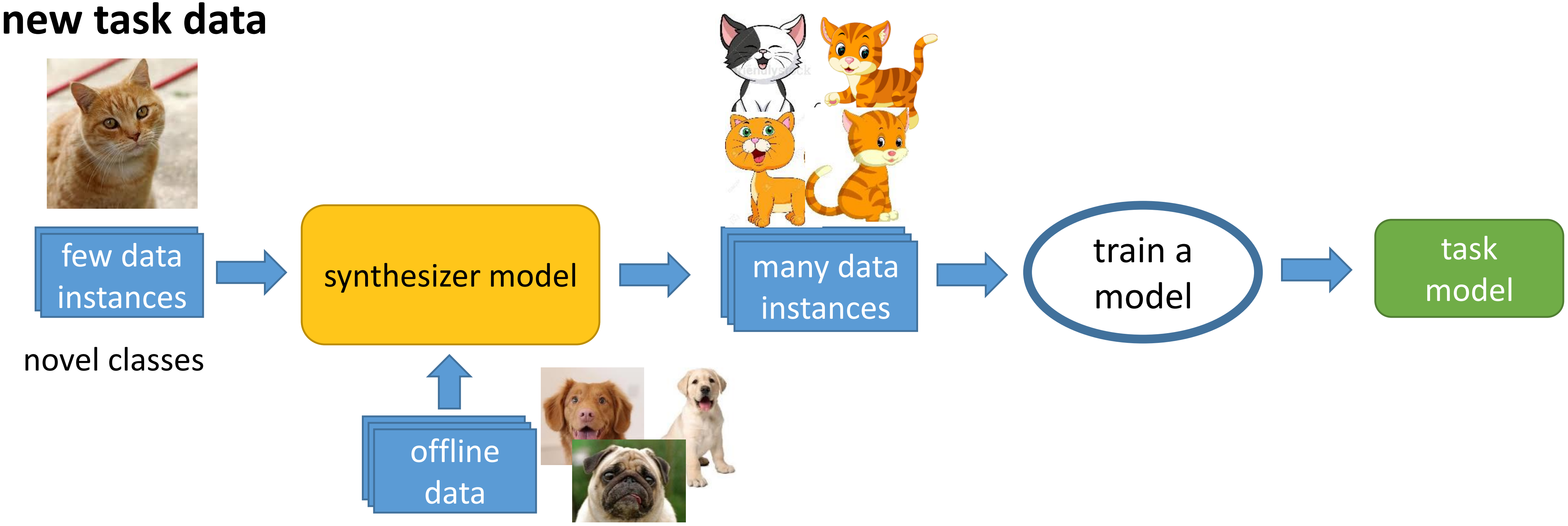


Sample synthesis

Offline stage



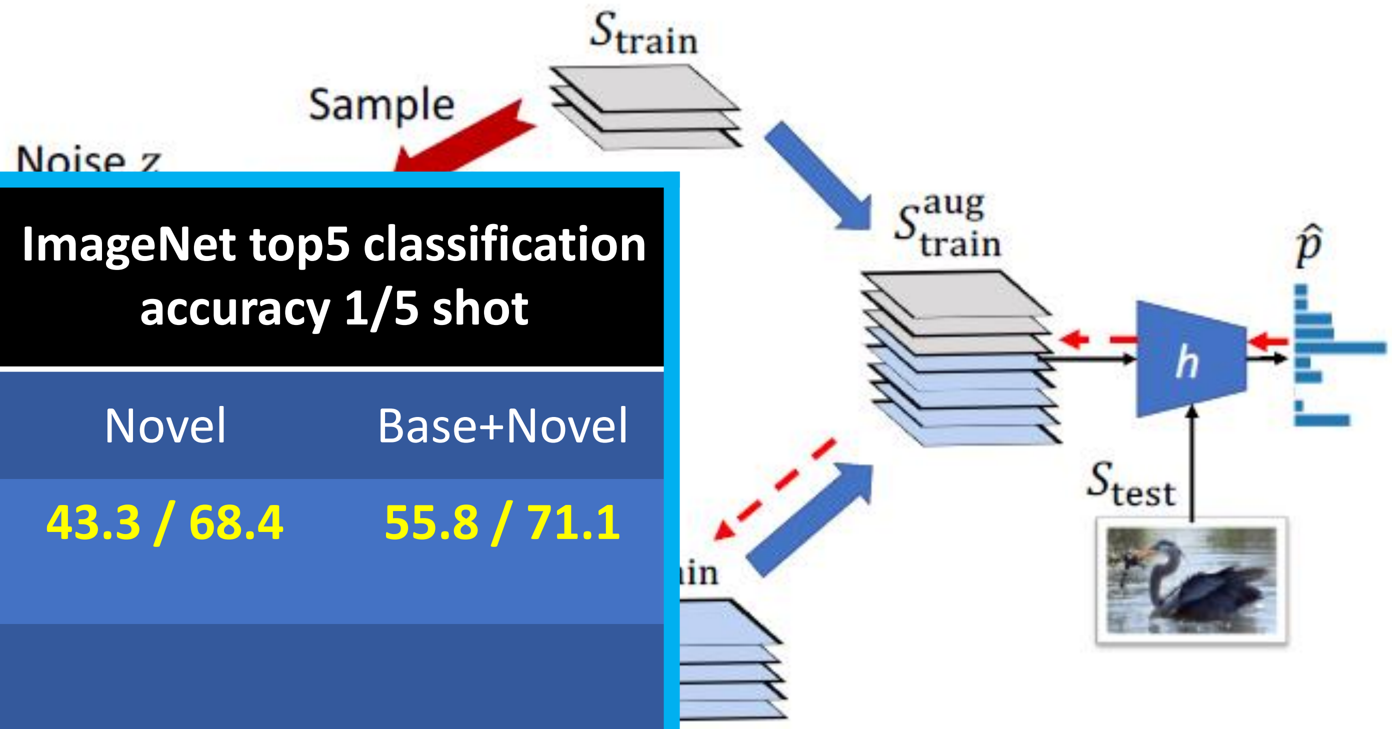
On new task data



Synthesizer optimized for classification

Low-Shot Learning from Imaginary Data
Wang et.al., 2018

- The synthesizer is a part of classifier pipeline, trained end-to end
- The classifier h is differentiable w.r.t. the training data
- In each episode, the backp gradient updates the synth



Method	ImageNet top5 classification accuracy 1/5 shot	
	Novel	Base+Novel
Wang et.al., 2018	43.3 / 68.4	55.8 / 71.1

reprinted from Wang et.al., 2018

More augmentation approaches

Δ-encoder Schwartz et.al., Ne

- Use a variant of autoenco
between two class sample

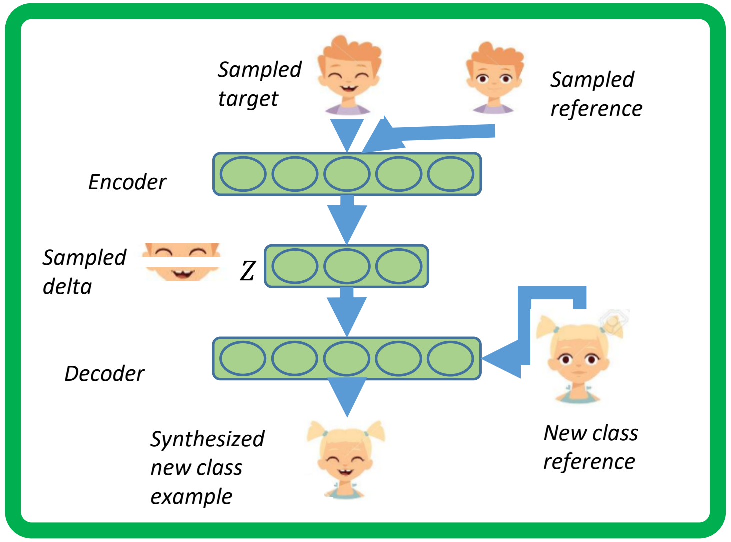
- Transfer class distributions

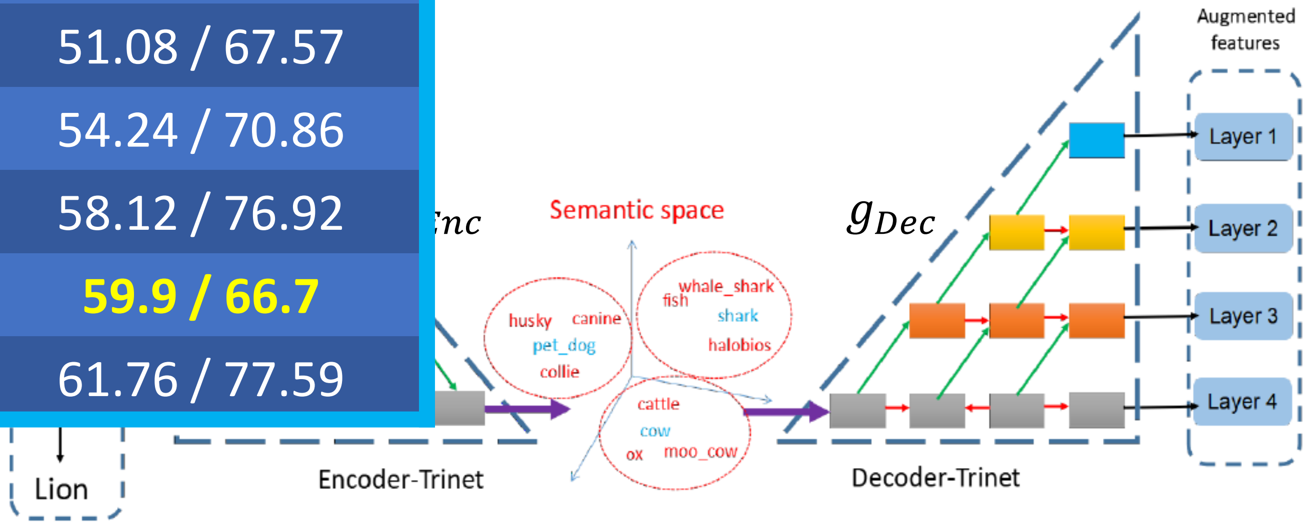
Semantic Feature Augmentat
Learning, Chen et.al., 2018

- Synthesize samples by add
autoencoder's bottleneck

- Make it into a semantic sp
embeddings or visual attri
objective's fidelity term.

Method	miniImageNet classification accuracy 1/5 shot
Matching networks	43.56 / 55.31
MAML	48.70 / 63.11
Relation networks	50.44 / 65.32
Prototypical Networks	49.42 / 68.20
Large-margin	51.08 / 67.57
Meta-SGD	54.24 / 70.86
Semantic Feat. Aug.	58.12 / 76.92
Δ-encoder	**59.9 / 66.7**
LEO	61.76 / 77.59





Augmentation with GANs

IBM

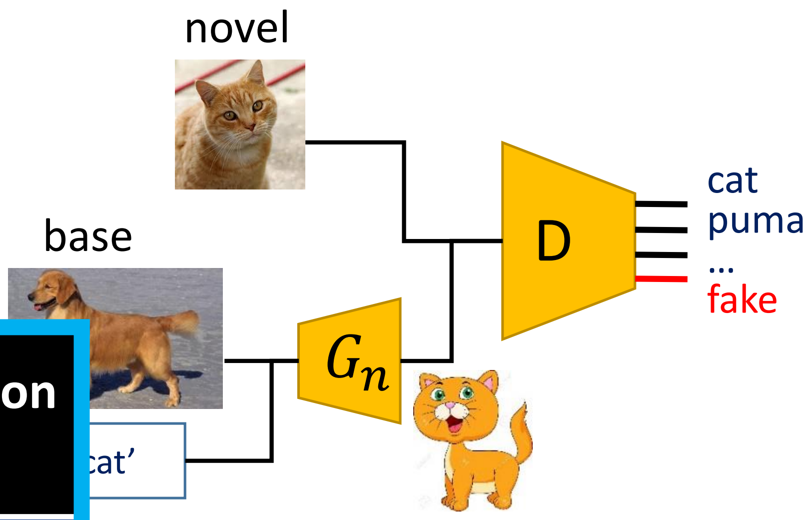
Covariance-Preserving Adversarial Augmentation Networks

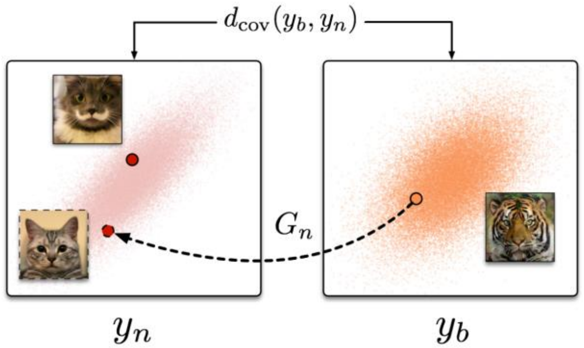
Gao et.al., NeurIPS 2018

- Act in in feature space. Generate samples for novel categories from offline categories, selected by proximity of samples.
- Discriminate by classica
- Cycle-consistency const
- Preserve the category d
  Objective: penalize the difference between the two covariance matrices via *Ky Fan* m-*norm*, *i.e.*, the sum of singular values of m-truncated SVD:

$$d_{\text{cov}}(y_b, y_n) = \left\| [\Sigma_{\mathbf{x}}(\mathbb{P}_{y_b}) - \Sigma_G(\mathbb{P}_{y_n})]_m \right\|_*$$

Method	ImageNet top5 classification accuracy 1/5 shot	
	Novel	Base+Novel
Wang et.al., 2018	43.3 / 68.4	55.8 / 71.1
Gao et.al., 2018	**48.4 / 70.2**	**58.5 / 73.5**



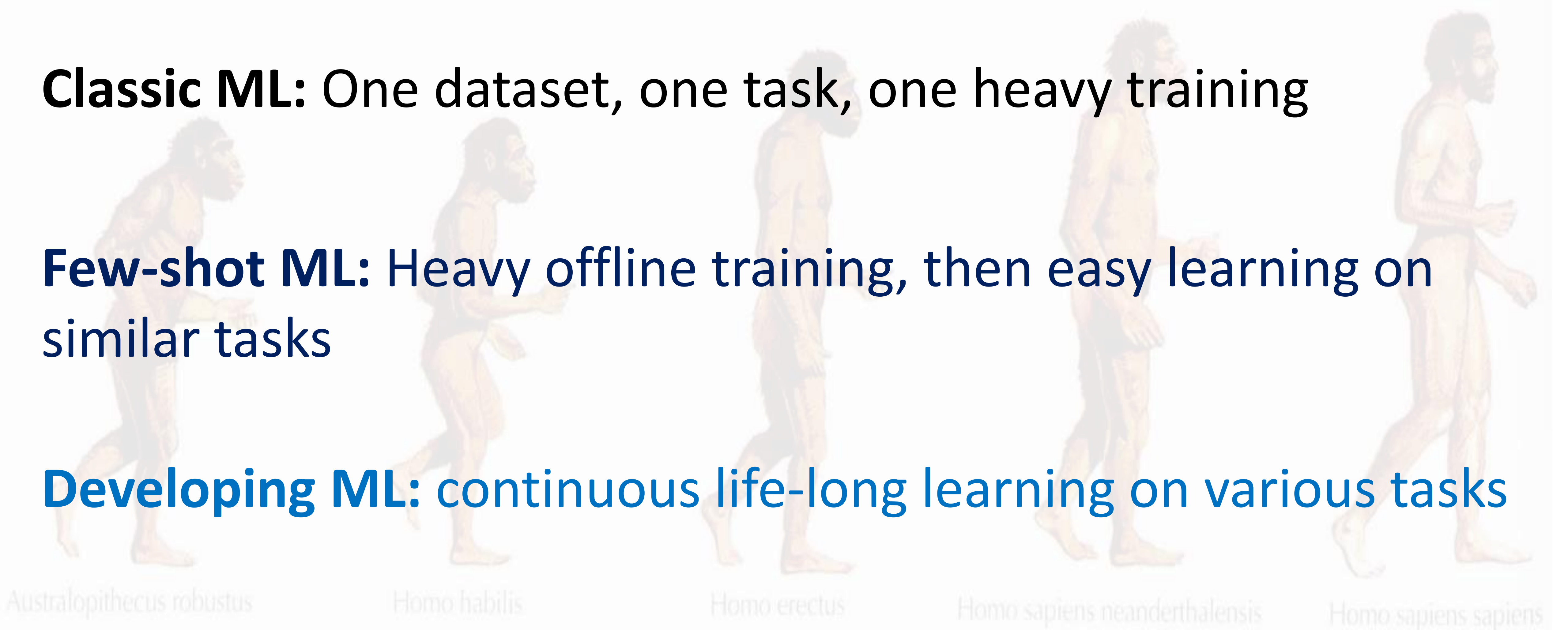


My personal view on the evolution of Machine Learning

Classic ML: One dataset, one task, one heavy training

Few-shot ML: Heavy offline training, then easy learning on similar tasks

Developing ML: continuous life-long learning on various tasks



THANK YOU

The presentation is available at http://www.research.ibm.com/haifa/dept/imt/ist_dm.shtml