# Meta-learning: Basics and Recent Advancements

Mertcan Yetkin

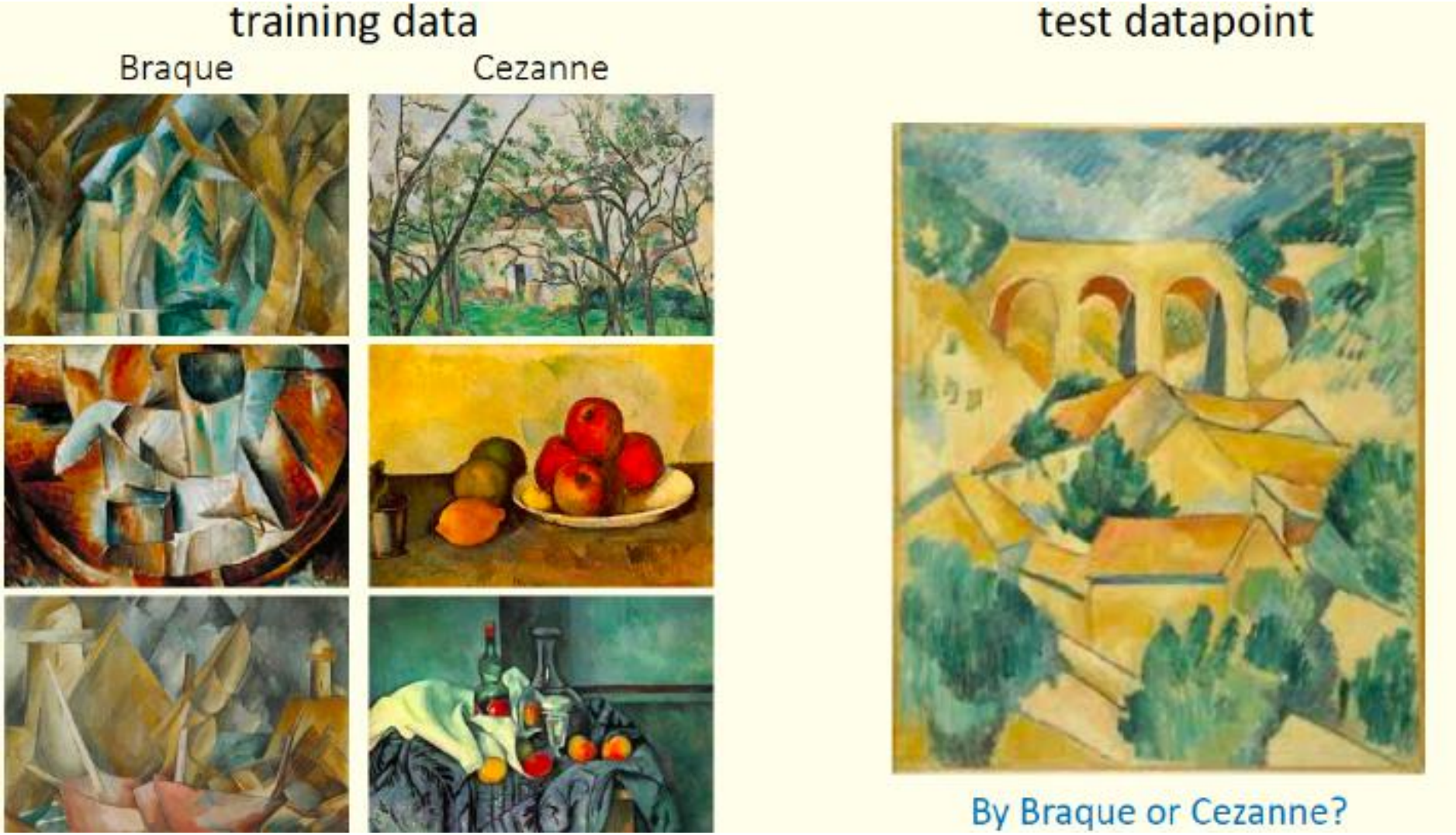
Department of Industrial and Systems Engineering
Lehigh University

OptML meetings
23 September 2020

# Meta-learning

- "Being aware of and taking control of one's own learning" [1]
- Learning to learn
- Understanding how automatic learning can become flexible in solving learning problems
  - Improve the existing learning algorithms
  - Learn the learning algorithm itself

# Motivation


training data
Braque
Cezanne
test datapoint
By Braque or Cezanne?

# Motivation ctd..

- The meta-data is already accumulating

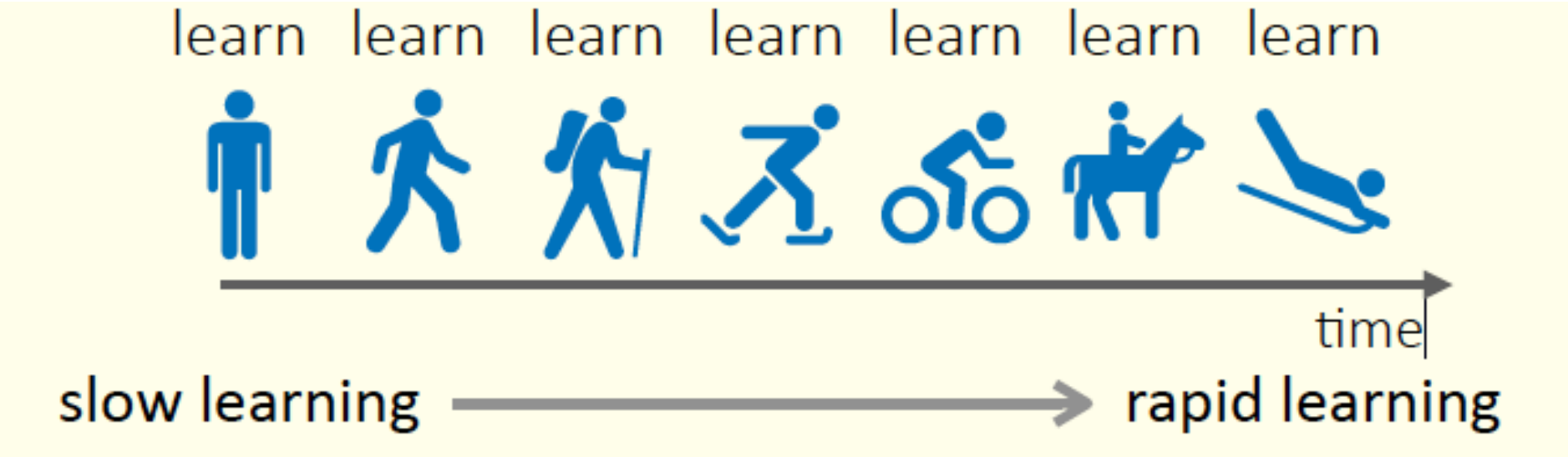



Figure: The ultimate goal

# Formal definition [1]

1. A metalearning system must include a learning subsystem, which adapts with experience.
2. Experience is gained by exploiting metaknowledge extracted
   1. in a previous learning episode on a single dataset, and/or
   2. from different domains or problems.

# Examples?

# Neural architecture search

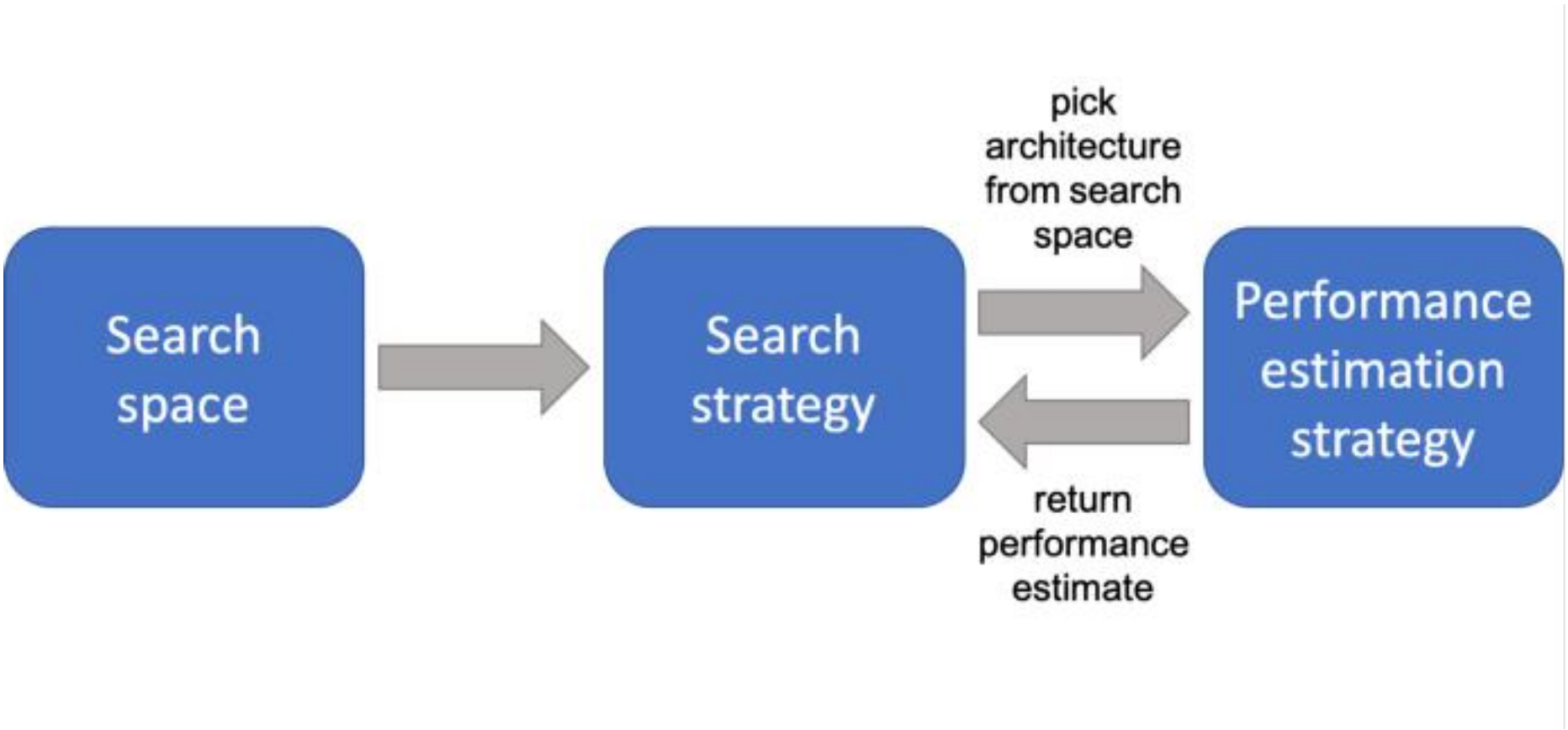
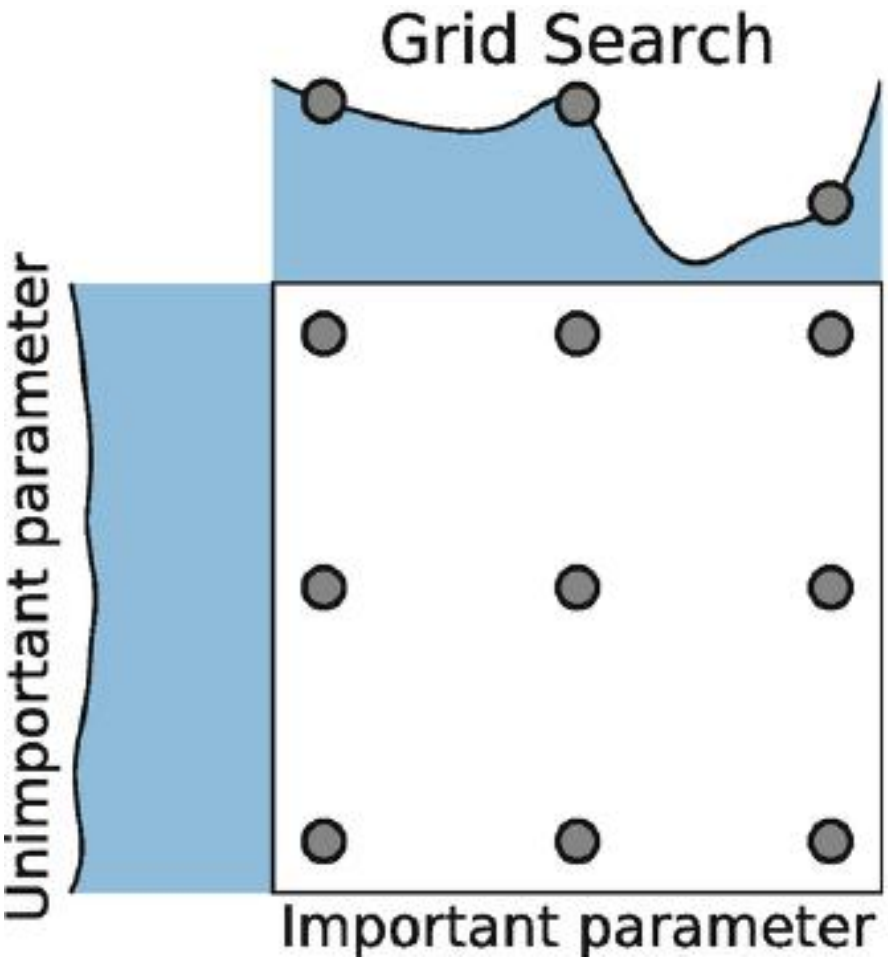



Figure: Neural architecture search

# Hyperparameter optimization



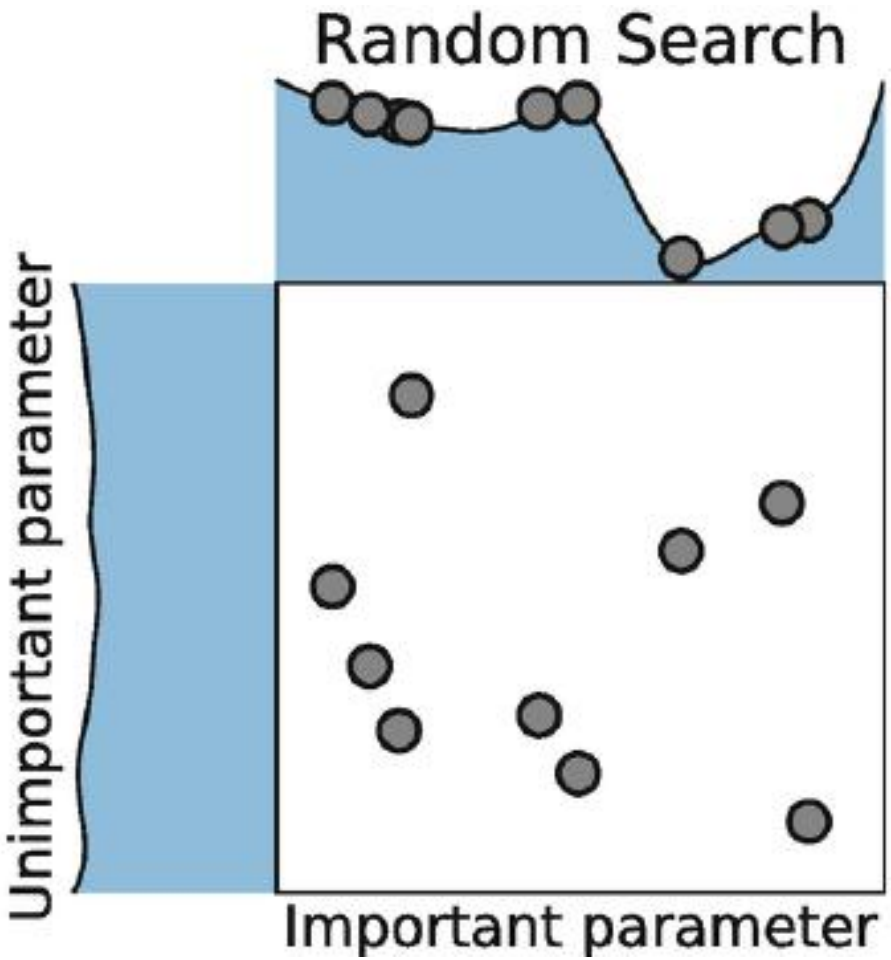
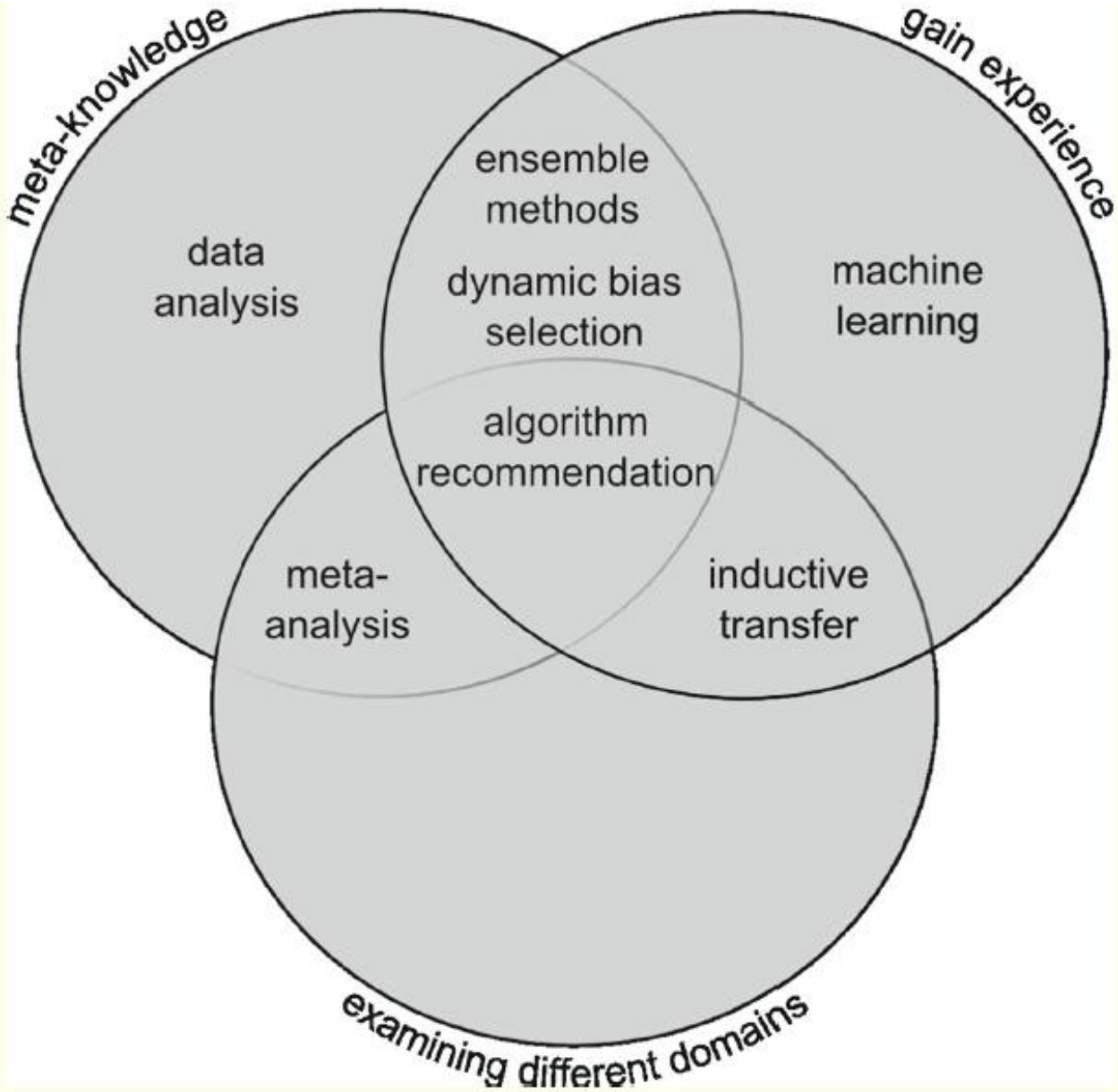




Figure: Hyperparameter optimization

# The big picture



# An example ensemble method

- Bagging:

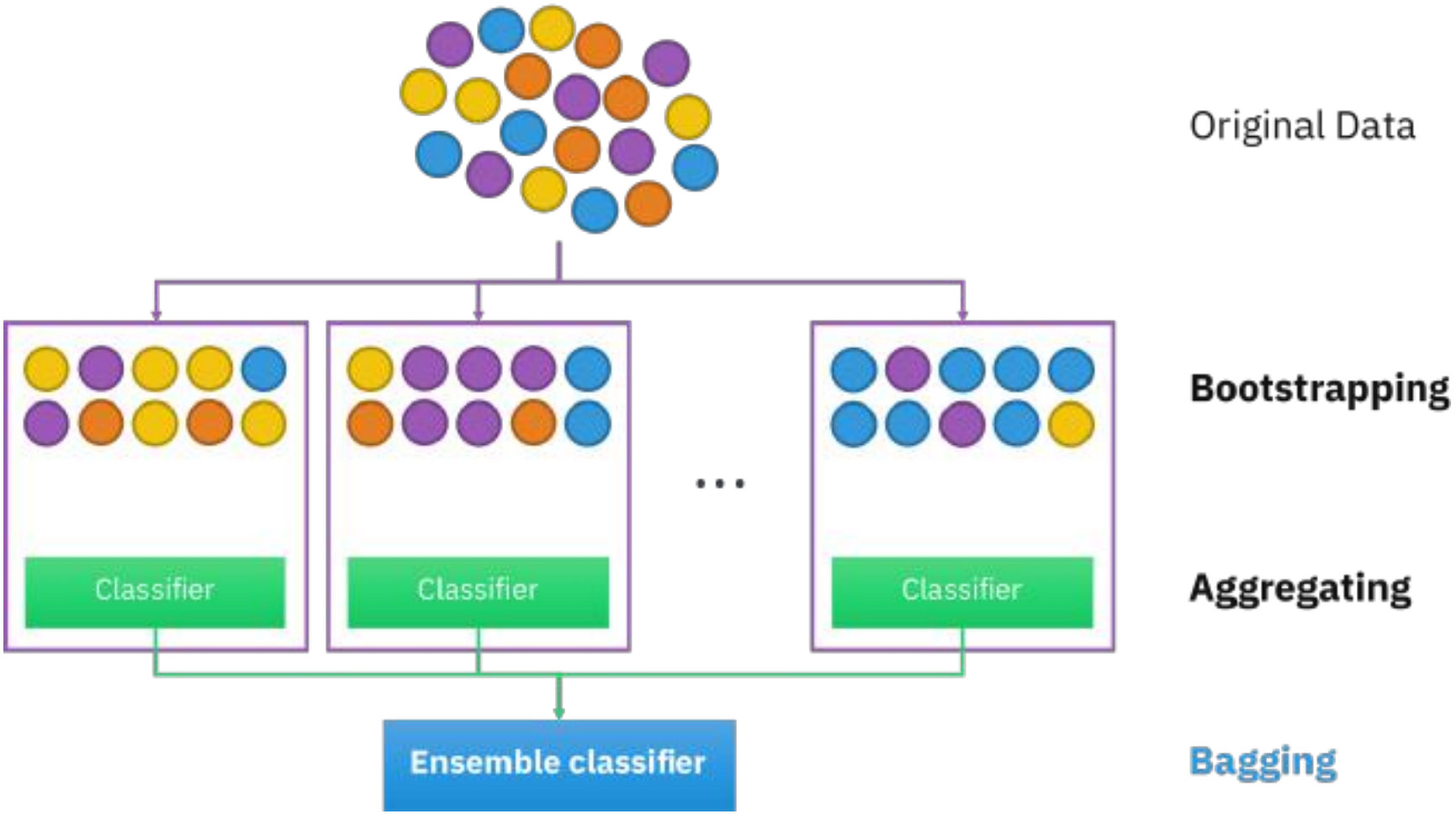



Figure: Bagging

# Another example ensemble method

- Boosting:

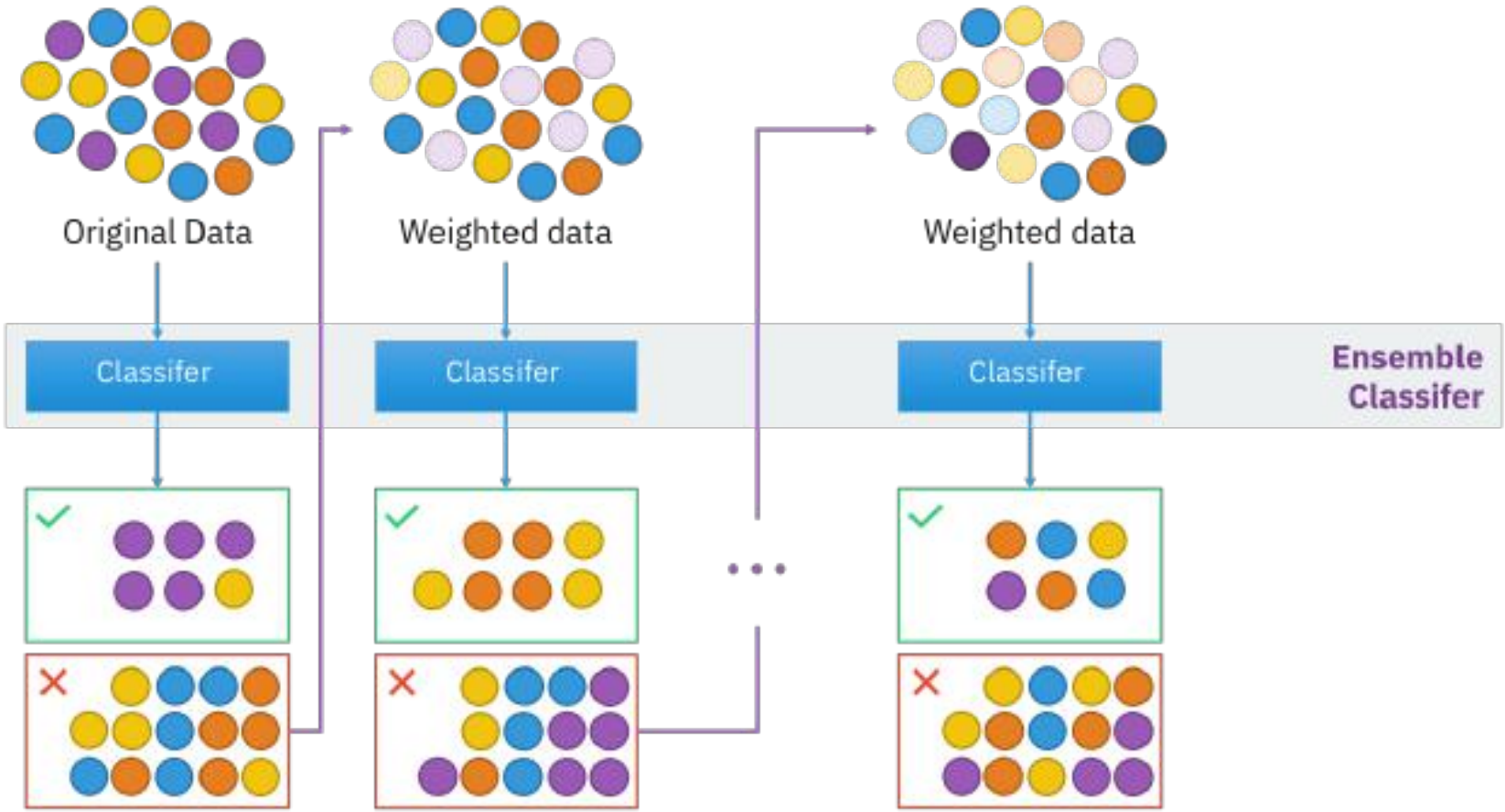



Figure: Boosting

# What is meta-data?

Some examples include:

- Algorithm runtime on a specific task
- Accuracy of a classification task
- Used step sizes in an iterative method
- ... any information obtained from a learning task (from a new experience)

# Bias notion in meta-learning

- set of assumptions influencing the choice of hypotheses for explaining the data
  - declarative bias, i.e. representing hypotheses using neural networks only
  - procedural bias, i.e. preferring hypotheses with smaller runtime
- bias in base-learning is fixed
- meta-learning tries to find the right bias

# Mathematical definition

- conventional machine learning:
  - given training dataset $D = \{x_1, y_1, \ldots, x_N, y_N\}$
  - We can train a predictive model $\hat{y} = f_\theta(x)$, parametrized by $\theta$
  - by solving $\theta^* = \arg\min_\theta L(D; \theta, w)$
- $w$ is included to represent factors such as
  - choice of optimizer for $\theta$
  - function class $f$
  - initial point for $\theta$
  - further settings affecting bias

# Mathematical definition ctd..

- meta-learning:
  - define task $T = \{D, L\}$, given a distribution of tasks $p(T)$
  - learning how to learn becomes $\min_w = \mathbb{E}_{T \sim p(T)} L(D; w)$
  - by solving $\theta^* = \arg\min_\theta L(D; \theta, w)$
  - where $L(D; w)$ measures the performance of a model trained using $w$ on dataset $D$
- in practice, we sample from tasks for meta-training stage $D_{\text{source}} = \{(D^{\text{train}}_{\text{source}}, D^{\text{val}}_{\text{source}})\}$
  - meta-training step: $w^* = \arg\max_w \log p(w|D_{\text{source}})$
  - meta-testing on new task $i$: $\theta^{*^i} = \arg\max_\theta \log p(\theta|w^*, D^{\text{train},i}_{\text{target}})$
  - evaluate the accuracy of meta-learner $\theta^{*^i}$ on $D^{\text{test}}_{\text{target}}$

# Example in computer vision

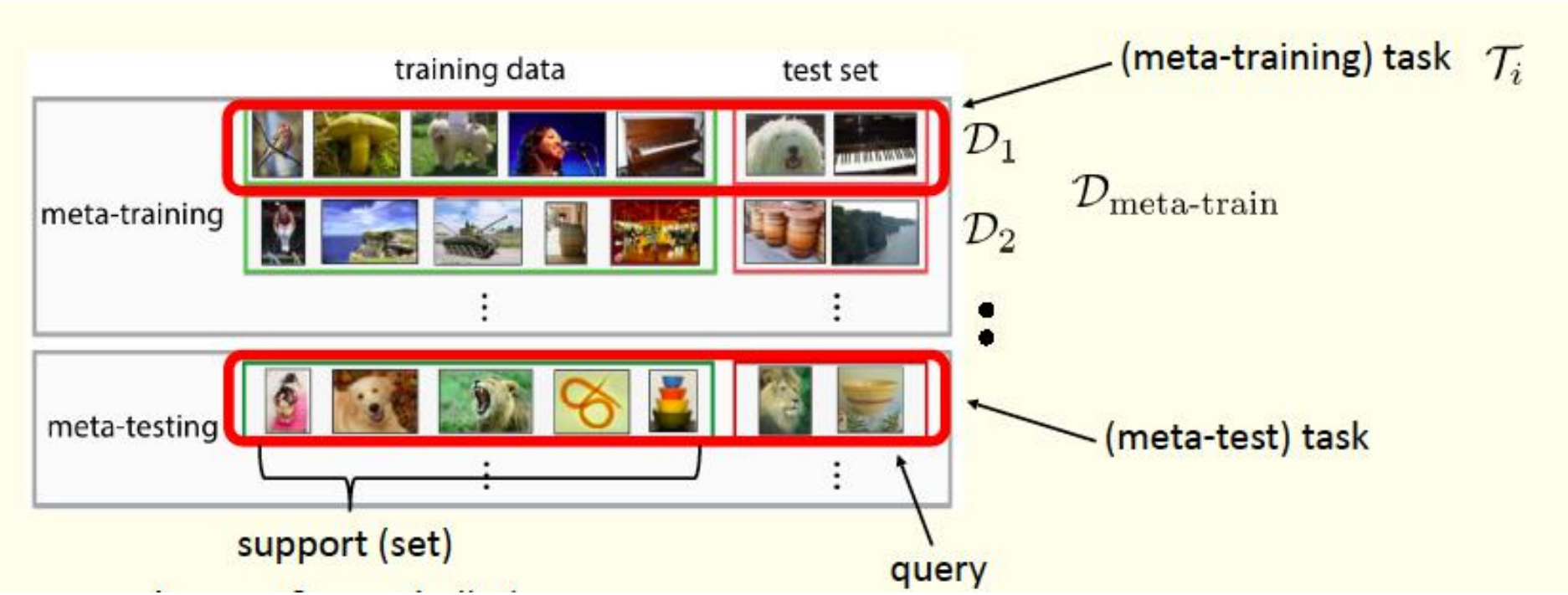



Figure: Data structure in few-shot classification

# An example benchmarking dataset

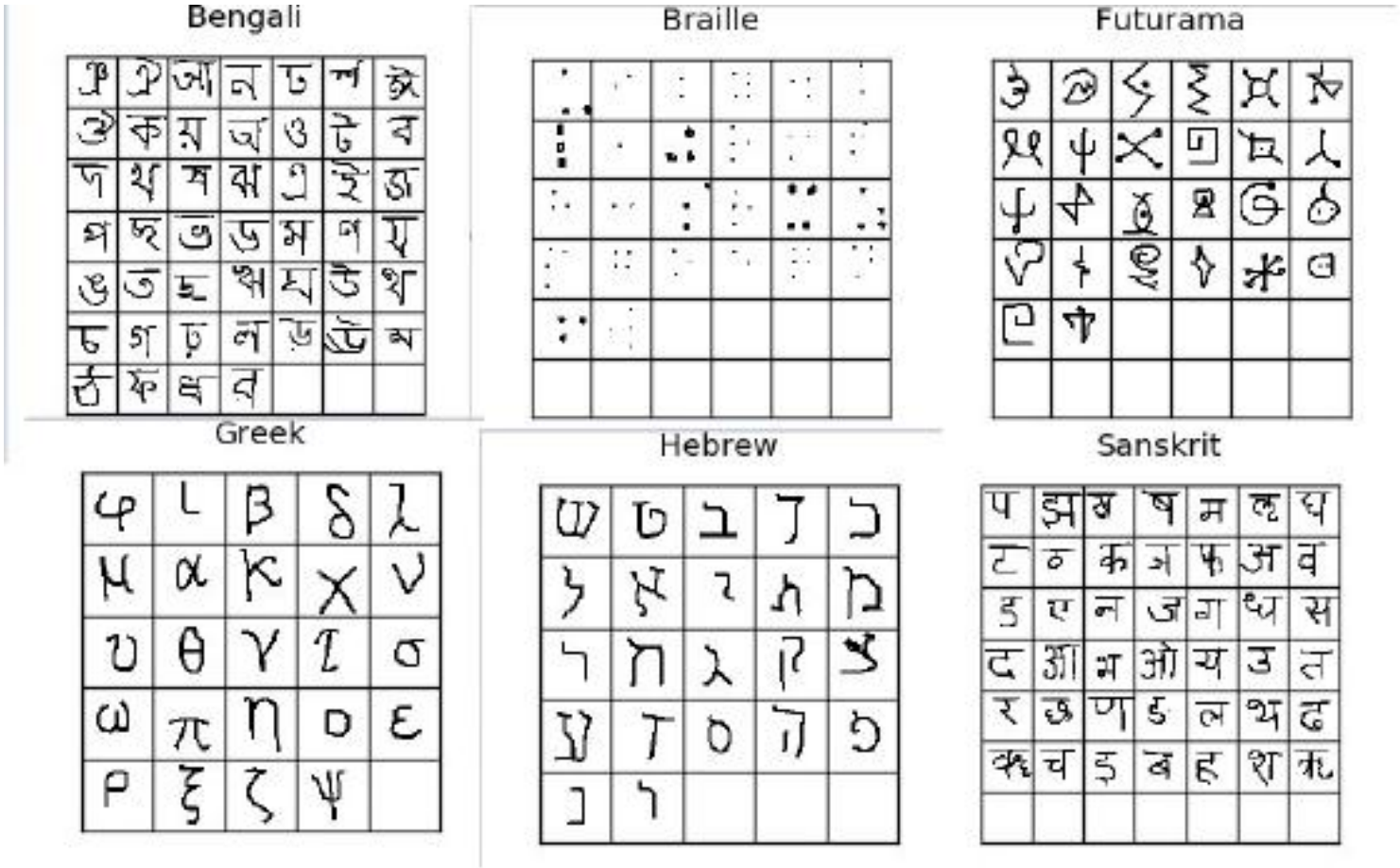



Figure: A glance at the Omniglot dataset

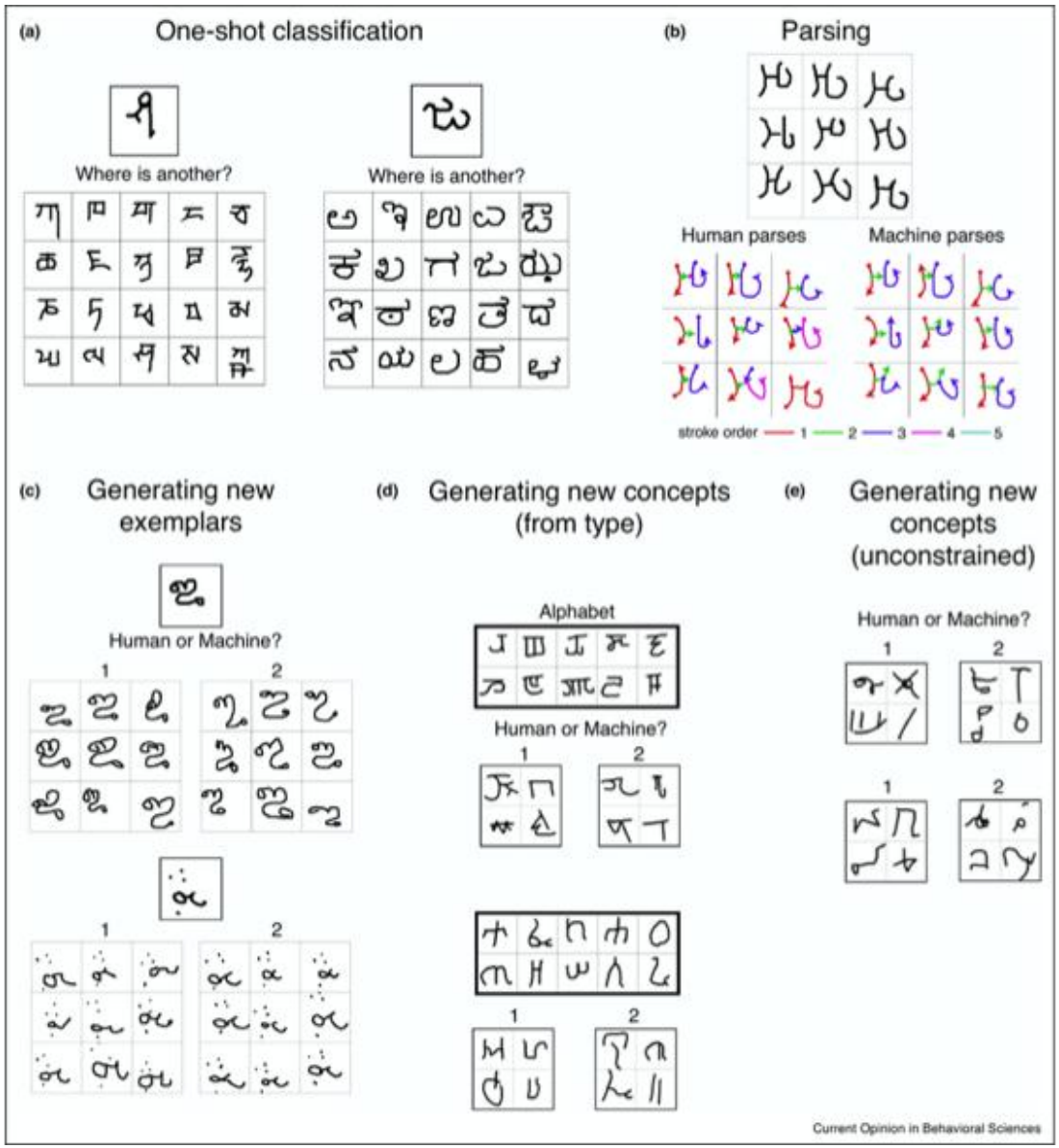



Figure: Some example tasks within Omniglot dataset

# Meta-overfitting

- Corresponds to learning a bias $w$ that constrains the hypothesis space of $\theta$ too tightly around solutions to the source tasks.

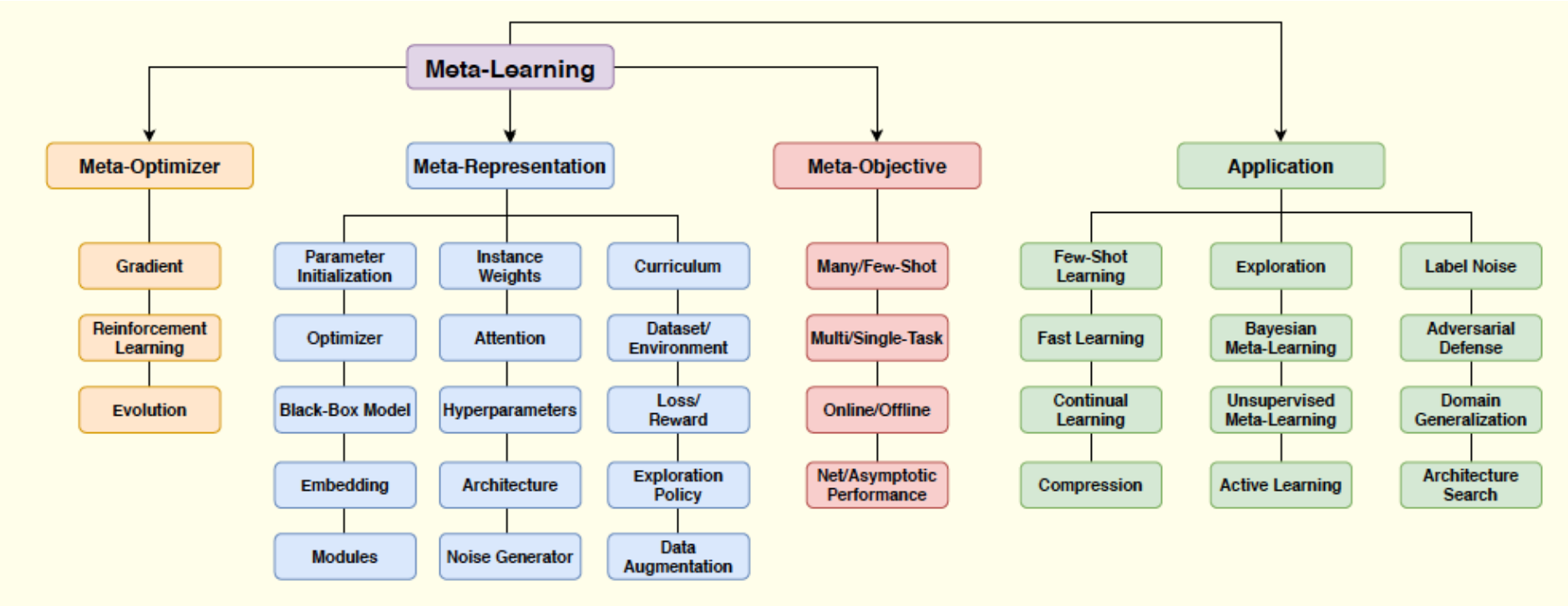
# Bilevel optimization view

- To solve the meta-training step

$$w^* = \arg\min_{w} \sum_{i=1}^{M} L^{\text{meta}}(\theta^{*^i}(w), w, D^{\text{val},i}_{\text{source}})$$

$$\text{s.t.} \theta^{*^i}(w) = \arg\min_{\theta} L^{\text{task}}(\theta, w, D^{\text{train,i}}_{\text{source}})$$

# Meta-learning landscape [2]



# Main categories of meta-learning [3]

- learning from model evaluations
  - task-independent recommendations (ranking of configurations)
  - configuration space design (i.e. hyperparameters)
  - configuration transfer for a new task (i.e. surrogate models)

# Main categories ctd..

- learning from task properties
  - meta-features
  - warm-starting optimization from similar tasks
  - to tune or not to tune?

| Name | Formula | Rationale |
|---|---|---|
| Nr instances | $n$ | Speed, Scalability (Michie et al., 1994) |
| Nr features | $p$ | Curse of dimensionality (Michie et al., 1994) |
| Nr classes | $c$ | Complexity, imbalance (Michie et al., 1994) |
| Nr missing values | $m$ | Imputation effects (Kalousis, 2002) |
| Nr outliers | $o$ | Data noisiness (Rousseeuw and Hubert, 2011) |
| Skewness | $\frac{E(X-\mu_X)^3}{\sigma_X^3}$ | Feature normality (Michie et al., 1994) |
| Kurtosis | $\frac{E(X-\mu_X)^4}{\sigma_X^4}$ | Feature normality (Michie et al., 1994) |
| Correlation | $\rho_{X_1X_2}$ | Feature interdependence (Michie et al., 1994) |
| Covariance | $cov_{X_1X_2}$ | Feature interdependence (Michie et al., 1994) |
| Concentration | $\tau_{X_1X_2}$ | Feature interdependence (Kalousis and Hilario, 2001) |
| Sparsity | sparsity(X) | Degree of discreteness (Salama et al., 2013) |
| Gravity | gravity(X) | Inter-class dispersion (Ali and Smith-Miles, 2006a) |
| ANOVA p-value | $p_{val_{\mathbf{X}_1X_2}}$ | Feature redundancy (Kalousis, 2002) |
| Coeff. of variation | $\frac{\sigma_Y}{\mu_Y}$ | Variation in target (Soares et al., 2004) |
| PCA $\rho_{\lambda_1}$ | $\sqrt{\frac{\lambda_1}{1+\lambda_1}}$ | Variance in first PC (Michie et al., 1994) |
| PCA skewness | | Skewness of first PC (Feurer et al., 2014) |
| PCA 95% | $\frac{dim_{95\%var}}{p}$ | Intrinsic dimensionality (Bardenet et al., 2013) |
| Class probability | $P(\mathbf{C})$ | Class distribution (Michie et al., 1994) |
| Class entropy | $H(\mathbf{C})$ | Class imbalance (Michie et al., 1994) |
| Norm. entropy | $\frac{H(\mathbf{X})}{log_2 n}$ | Feature informativeness (Castiello et al., 2005) |
| Mutual inform. | $MI(\mathbf{C},\mathbf{X})$ | Feature importance (Michie et al., 1994) |
| Uncertainty coeff. | $\frac{MI(\mathbf{C},\mathbf{X})}{H(\mathbf{C})}$ | Feature importance (Agresti, 2002) |
| Equiv. nr. feats | $\frac{H(C)}{MI(C,X)}$ | Intrinsic dimensionality (Michie et al., 1994) |
| Noise-signal ratio | $\frac{H(X)-MI(C,X)}{MI(C,X)}$ | Noisiness of data (Michie et al., 1994) |

# Main categories ctd..

- learning from prior models
  - transfer learning
  - few-shot learning (model-agnostic meta learning)
  - other unsupervised learning

# Black-box adaptation

- Key idea: Train a neural network to represent $p(\theta^i | D_{\text{source}}^{\text{train},i}, w)$
- Use deterministic $\theta^i = f_w(D_{\text{source}}^{\text{train},i})$
- Some common forms for $f_w$
  - LSTM
  - Neural turing machine (NTM)
  - Self-attention
  - 1D convolutions

# Optimization-based approach

- Consider the problem of initializing good weights $\theta$ for different tasks
- meta-learning: $\min_w \sum_{i=1}^{M} L^{\text{meta}}(w - \alpha \nabla_w L^{\text{task}}(w, D_{\text{source}}^{\text{train,i}}), D_{\text{source}}^{\text{val,i}}))$
- meta-testing: $\theta \leftarrow w - \alpha \nabla_w L^{\text{task}}(w, D_{\text{target}}^{\text{train,i}})$

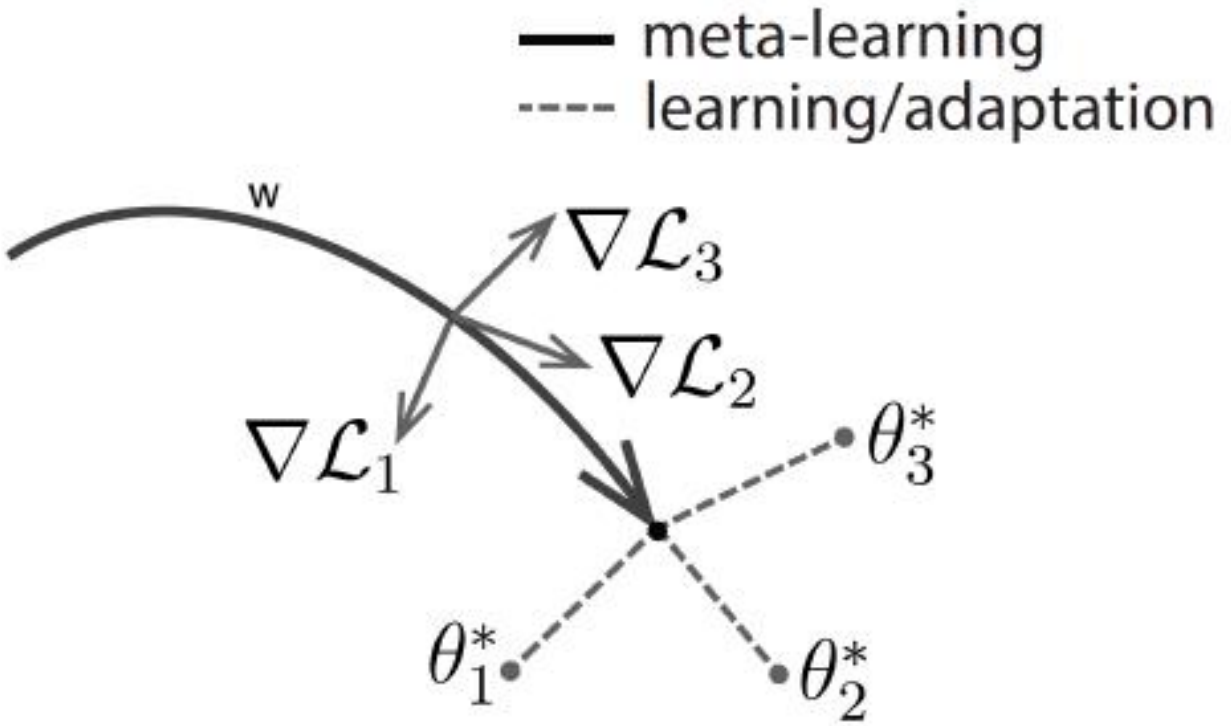
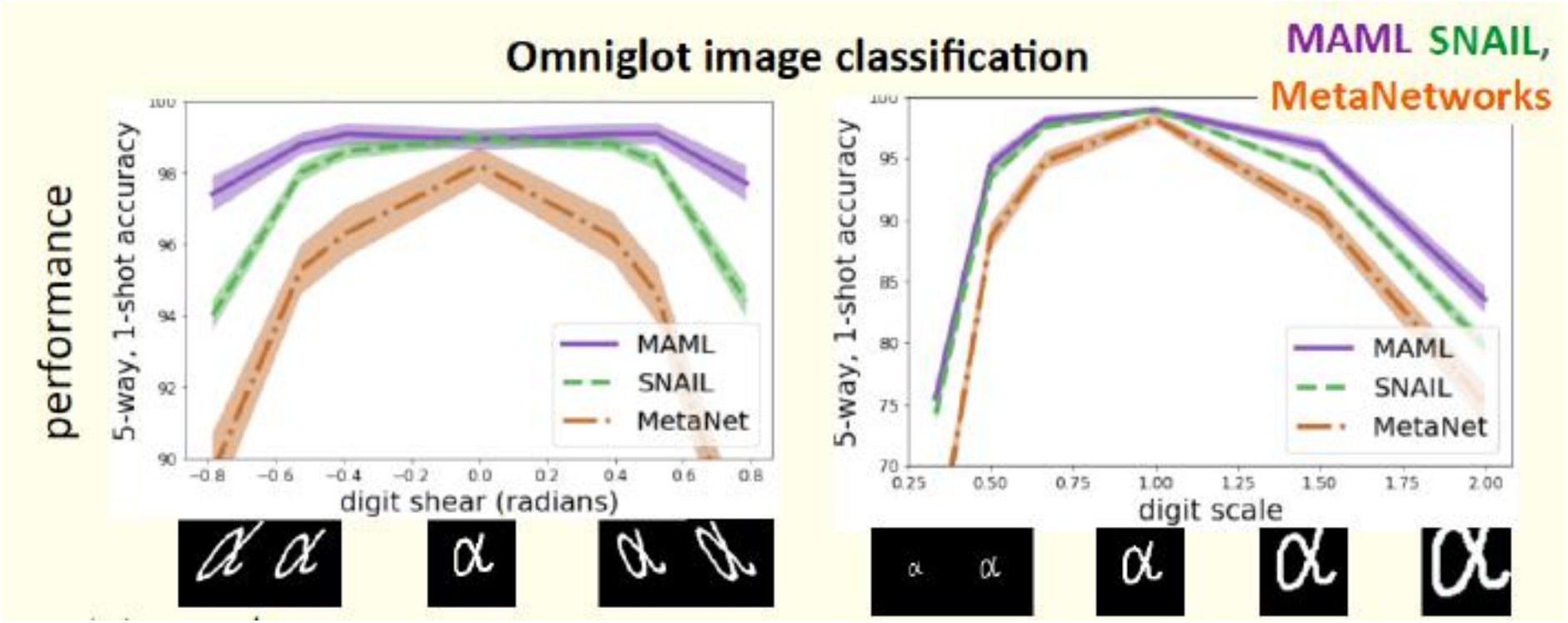



Figure: MAML idea

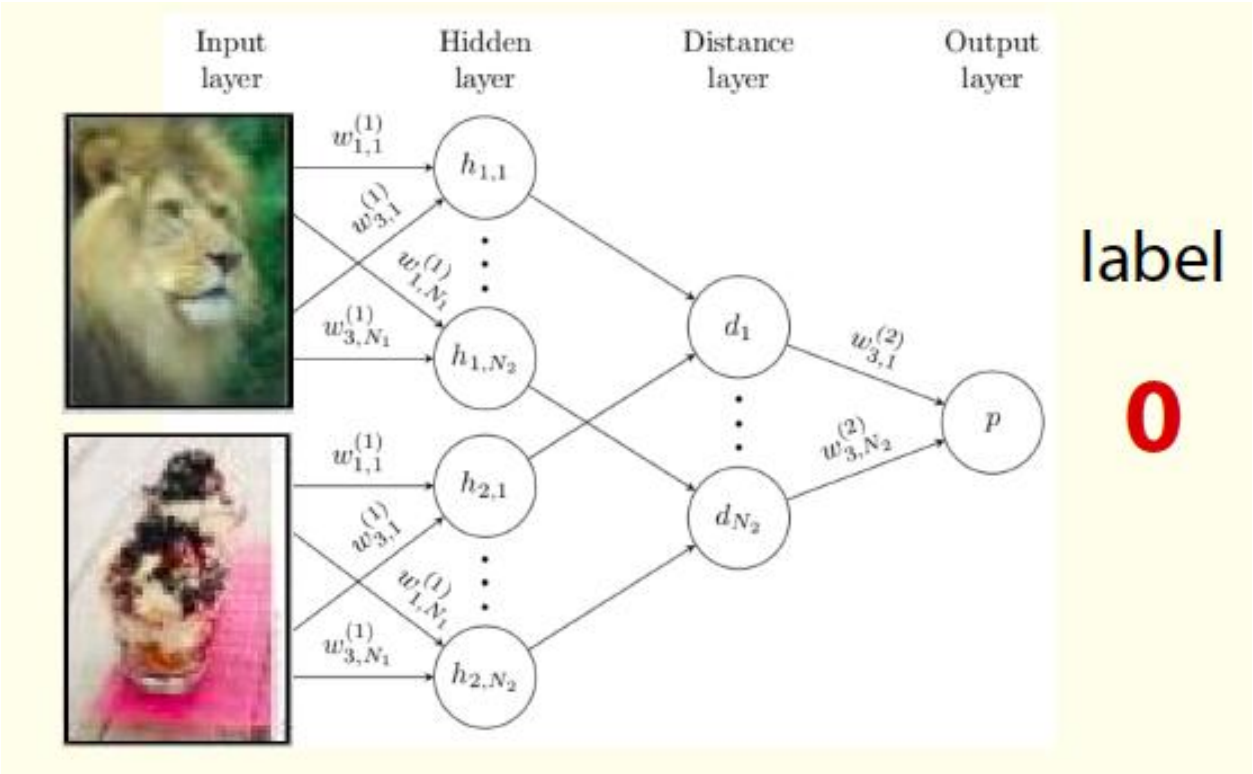
# Optimization vs black-box approaches

Generalization of learning procedures to extrapolated tasks:



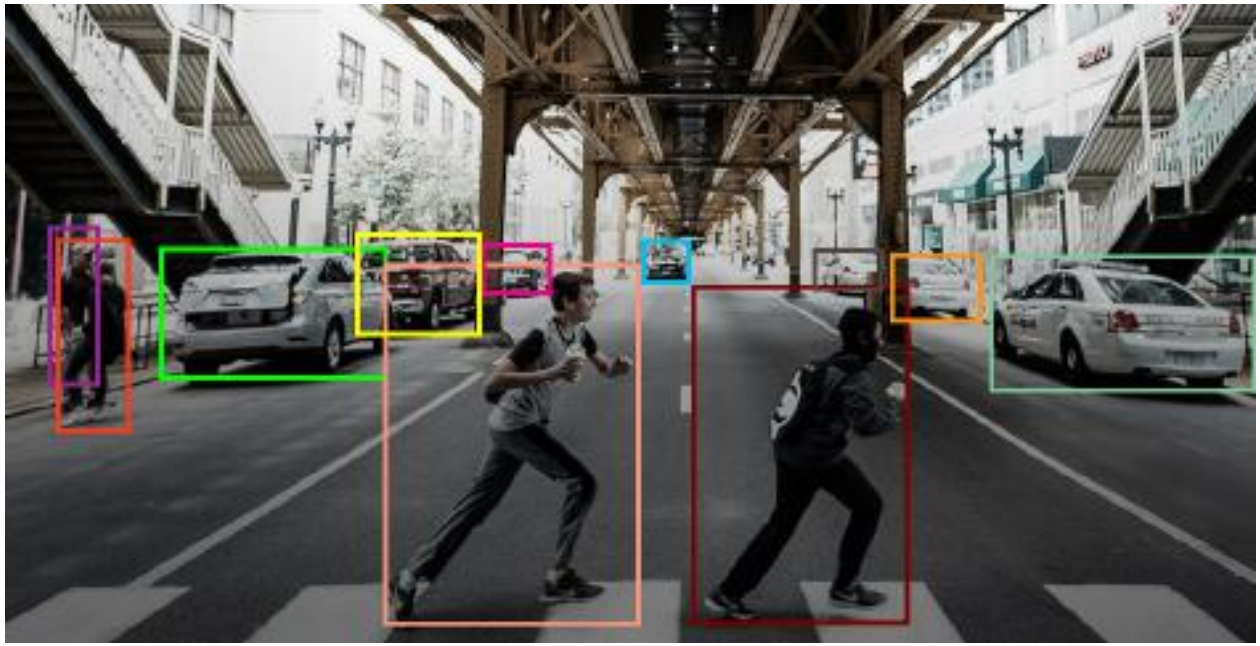
# Other non-parametric approaches

- Obtain task similarity
  - Siamese networks to predict 2 images belong to same class
  - K-nearest neighbours

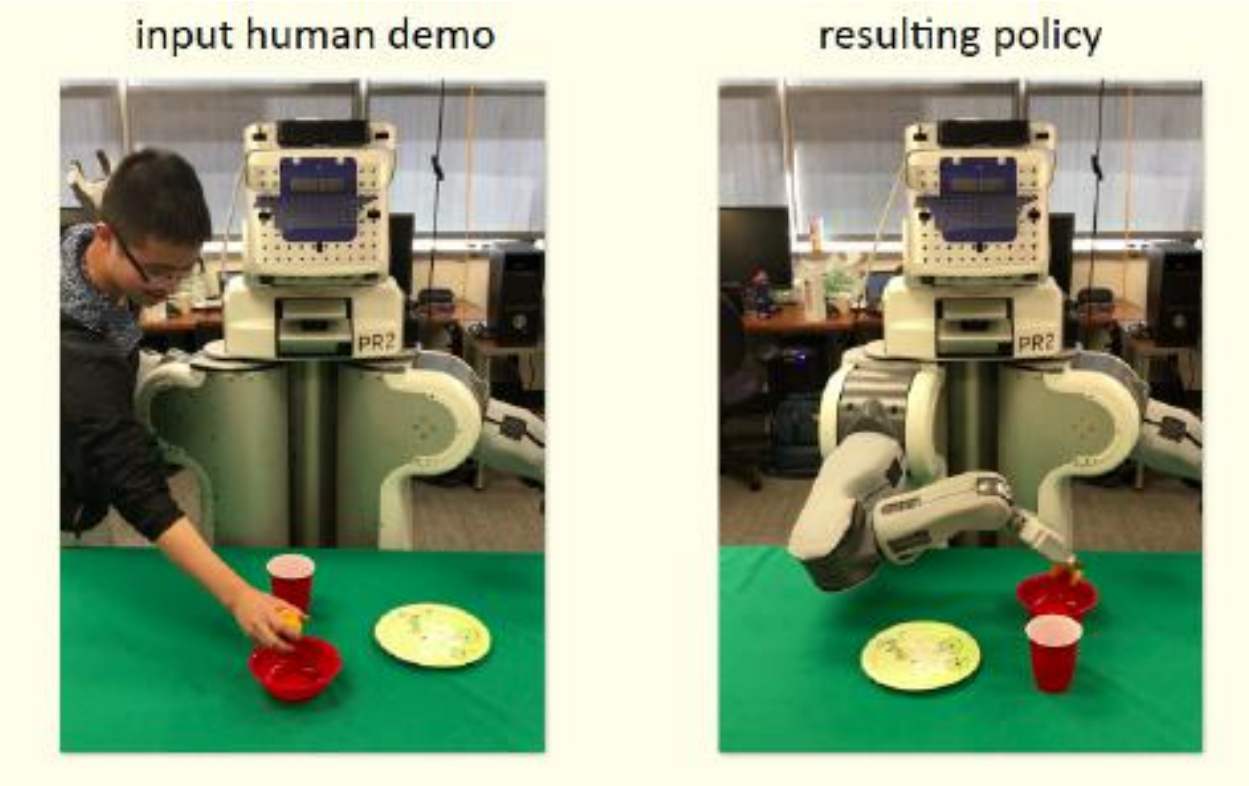


# Application areas

- computer vision: few-shot learning methods
  1. classification
  2. object detection
  3. object segmentation
  4. density estimation



- meta reinforcement learning & robotics
  1. exploration
  2. optimization
  3. knowledge-transfer? (opening jar vs opening door)



- environment learning (simulator)
- continual learning (learning new things, without forgetting, DNN)
- language and speech
  - language modelling (filling in missing words)
  - speech recognition (different accents)
- systems
  - learning with label noise
  - adversarial attacks and defenses

# Challenges

- meta-generalization (new tasks, conflicting gradients)
- one solution for all (distribution over tasks)
- computation cost (many inner loops) $\longrightarrow$ few-shot learning
- cross-modal transfer (visual imitation learning)

# Optimization related readings

- Meta-learning with differentiable convex optimization [4]
- Bilevel programming for hyperparameter optimization and meta-learning [5]
- Model-agnostic meta-learning (MAML) [6]
- Probabilistic model-agnostic meta-learning [7]
- Convergence theory for gradient-based MAML algorithms [8]

# References I



Christiane Lemke, Marcin Budka, and Bogdan Gabrys.
Metalearning: a survey of trends and technologies.
*Artificial intelligence review*, 44(1):117–130, 2015.


Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey.
Meta-learning in neural networks: A survey.
*arXiv preprint arXiv:2004.05439*, 2020.


Joaquin Vanschoren.
Meta-learning: A survey.
*arXiv preprint arXiv:1810.03548*, 2018.

# References II



Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto.
Meta-learning with differentiable convex optimization.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.


Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimilano Pontil.
Bilevel programming for hyperparameter optimization and meta-learning.
*arXiv preprint arXiv:1806.04910*, 2018.


Chelsea Finn, Pieter Abbeel, and Sergey Levine.
Model-agnostic meta-learning for fast adaptation of deep networks.
*arXiv preprint arXiv:1703.03400*, 2017.

# References III



Chelsea Finn, Kelvin Xu, and Sergey Levine.
Probabilistic model-agnostic meta-learning.
In *Advances in Neural Information Processing Systems*, pages 9516–9527, 2018.


Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar.
On the convergence theory of gradient-based model-agnostic meta-learning algorithms.
In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092, 2020.