Artificial Intelligence

National Taipei University

The Theory of Learning and Ensemble Learning

1111AI06
MBA, IM, NTPU (M6132) (Fall 2022)
Wed 2, 3, 4 (9:10-12:00) (B8F40)

https://meet.google.com/miy-fbif-max

aws educate Cloud Ambassador 2020 Cohort

Min-Yuh Day, Ph.D, Associate Professor

Institute of Information Management, National Taipei University

https://web.ntpu.edu.tw/~myday

2022-10-26

Syllabus

Week	Date	Subject/Topics
1	2022/09/14	Introduction to Artificial Intelligence
2	2022/09/21	Artificial Intelligence and Intelligent Agents
3	2022/09/28	Problem Solving
4	2022/10/05	Knowledge, Reasoning and Knowledge Representation; Uncertain Knowledge and Reasoning
5	2022/10/12	Case Study on Artificial Intelligence I
6	2022/10/19	Machine Learning: Supervised and Unsupervised Learning

Syllabus

Week Date Subject/Topics

7 2022/10/26 The Theory of Learning and Ensemble Learning

8 2022/11/02 Midterm Project Report

9 2022/11/09 Deep Learning and Reinforcement Learning

10 2022/11/16 Deep Learning for Natural Language Processing

11 2022/11/23 Invited Talk: AI for Information Retrieval

12 2022/11/30 Case Study on Artificial Intelligence II

Syllabus

Week	Date	Subject/Topics
13	2022/12/07	Computer Vision and Robotics
14	2022/12/14	Philosophy and Ethics of AI and the Future of AI
15	2022/12/21	Final Project Report I
16	2022/12/28	Final Project Report II
17	2023/01/04	Self-learning
18	2023/01/11	Self-learning

The Theory of Learning and Ensemble Learning

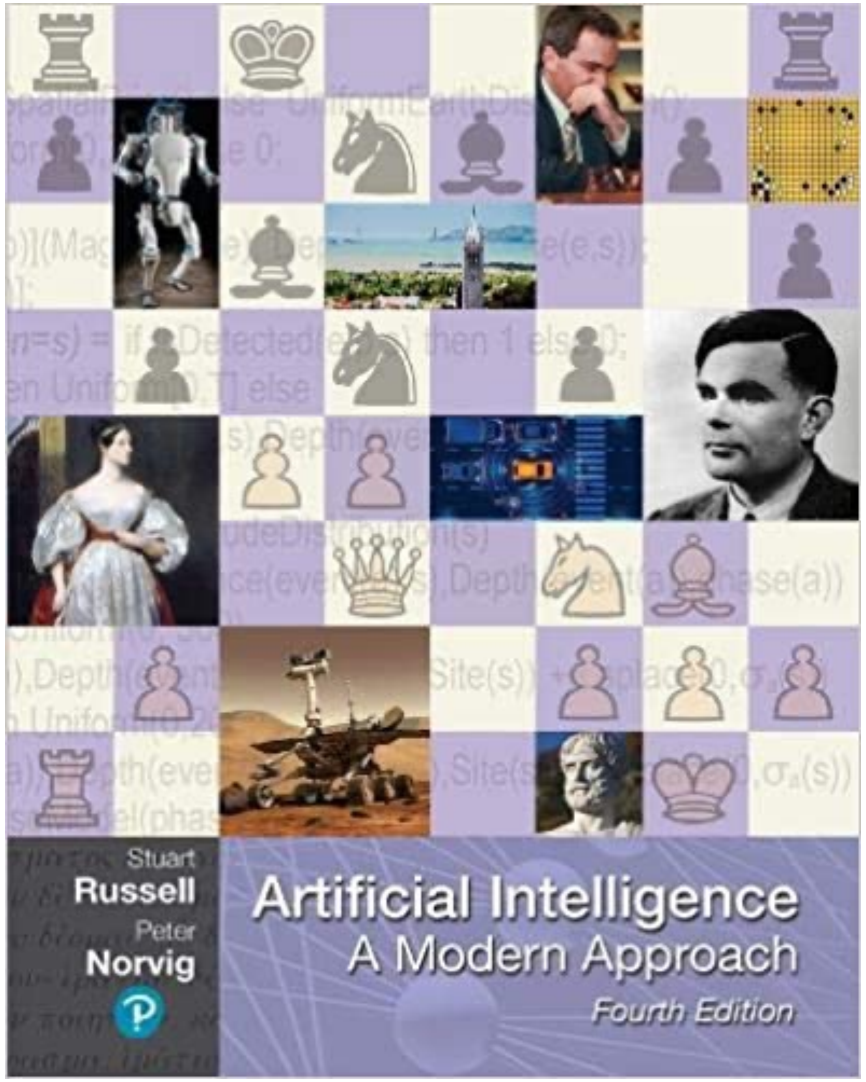
Outline

- The Theory of Learning
  - Computational Learning Theory
  - Probably Approximately Correct (PAC) Learning
- Ensemble Learning
  - Bagging: Random forests
  - Stacking
  - Boosting: Gradient boosting
  - Online learning
- Meta Learning: Learning to Learn

Stuart Russell and Peter Norvig (2020),

Artificial Intelligence: A Modern Approach,

4th Edition, Pearson



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

https://www.amazon.com/Artificial-Intelligence-A-Modern-Approach/dp/0134610997/

Artificial Intelligence: A Modern Approach

1. Artificial Intelligence
2. Problem Solving
3. Knowledge and Reasoning
4. Uncertain Knowledge and Reasoning
5. Machine Learning
6. Communicating, Perceiving, and Acting
7. Philosophy and Ethics of AI

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence:
Machine Learning

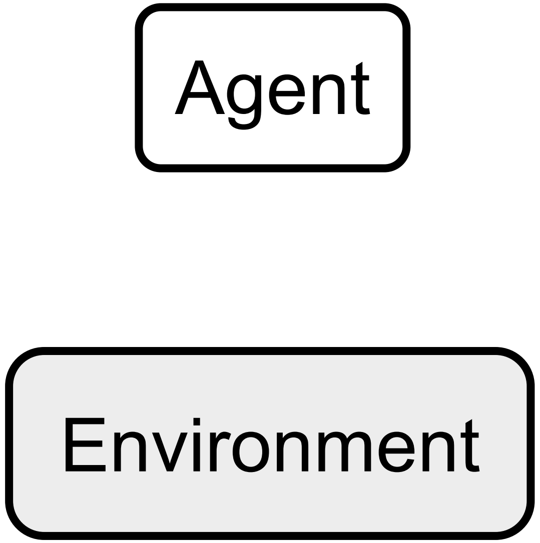
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence: 5. Machine Learning

- Learning from Examples
- Learning Probabilistic Models
- Deep Learning
- Reinforcement Learning

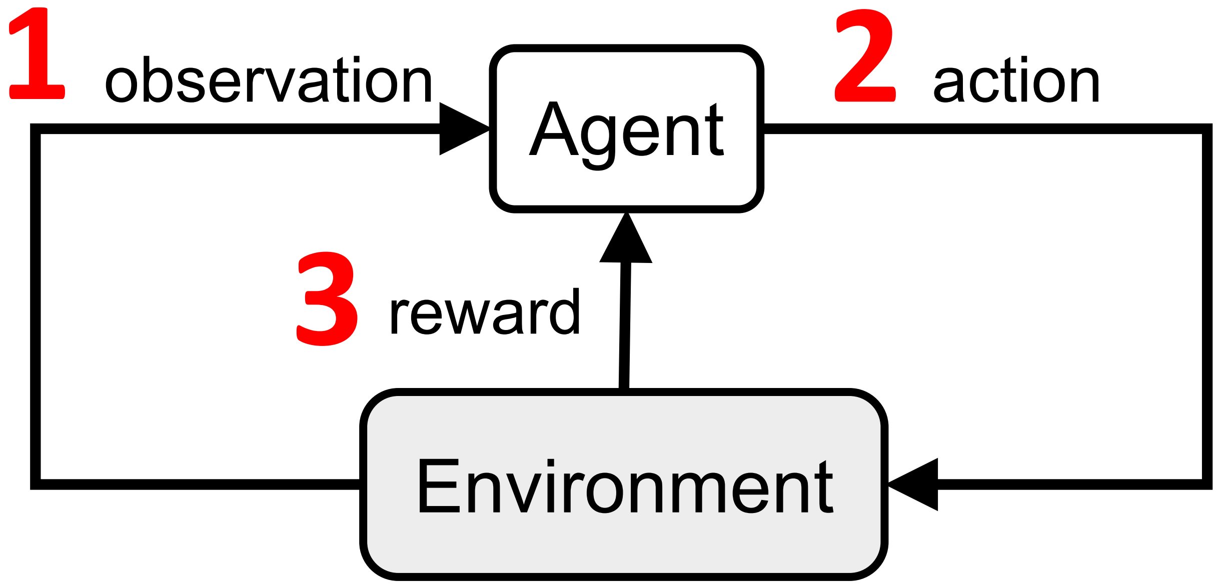
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Reinforcement Learning (DL)



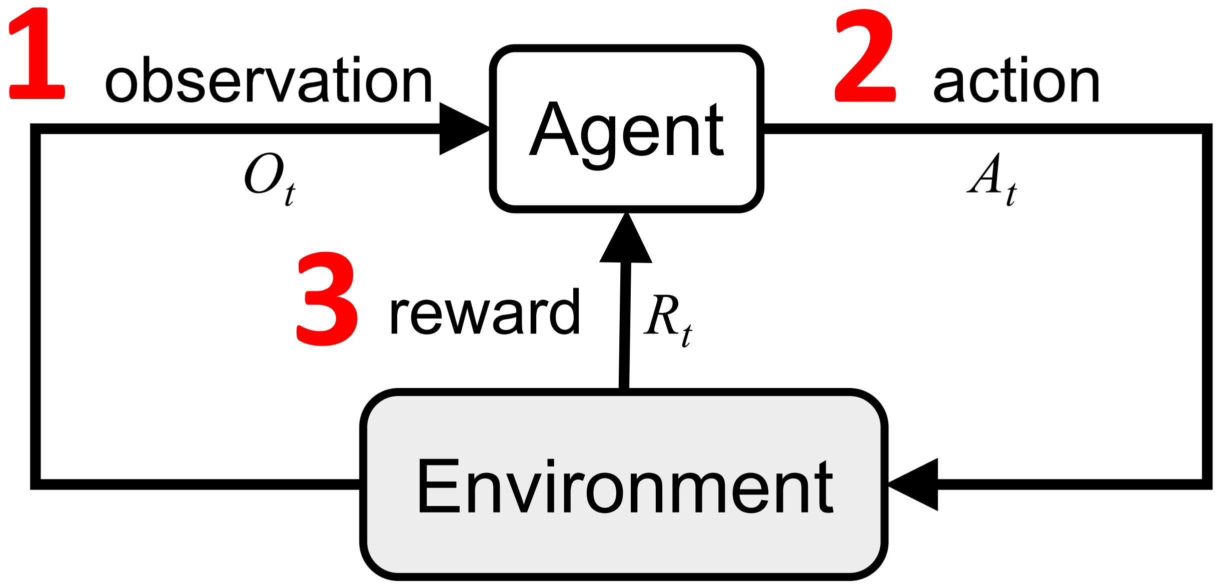
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



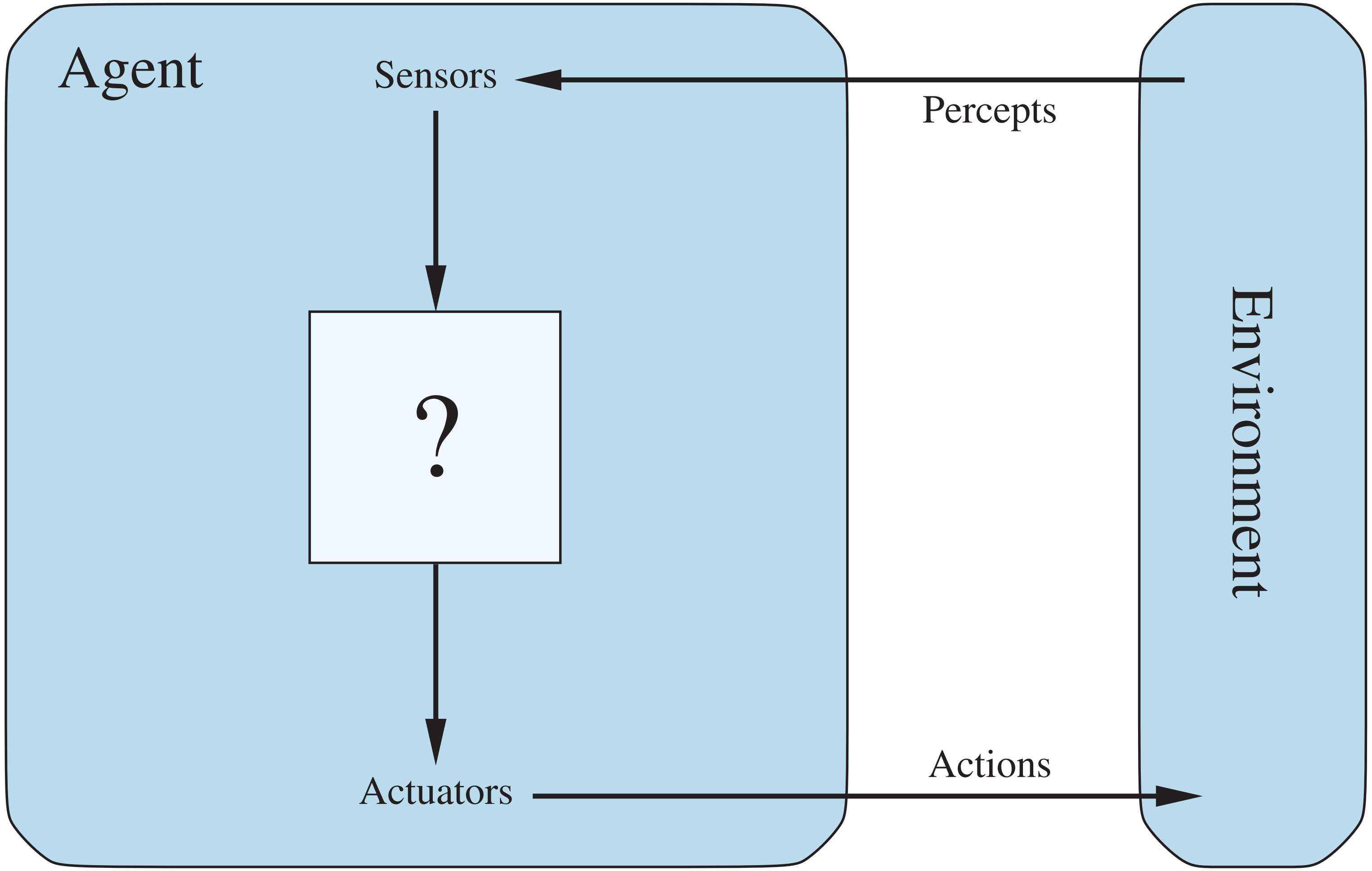
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Agents interact with environments through sensors and actuators



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Machine Learning
Supervised Learning (Classification)
Learning from Examples

$$y = f(x)$$

x	y

Machine Learning
Supervised Learning (Classification)
Learning from Examples

$$y = f(x)$$

x	y
input	*Output*
	label

Machine Learning
Supervised Learning (Classification)
Learning from Examples

$$y = f(x)$$

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
```

Machine Learning
Supervised Learning (Classification)
Learning from Examples

$$y = f(x)$$

Example

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
```

Machine Learning
Supervised Learning (Classification)
Learning from Examples

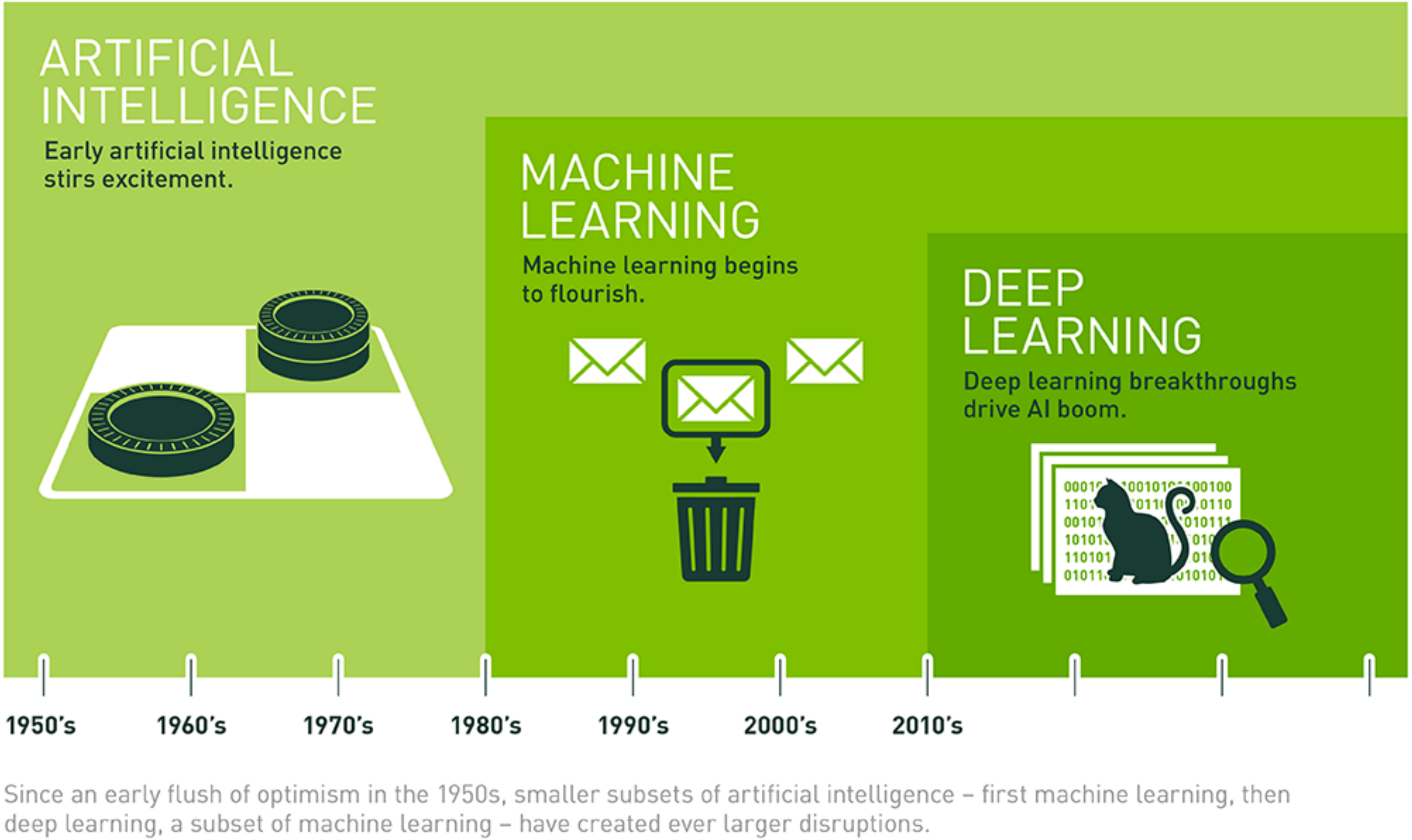
$y = f(x)$

Example

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
```

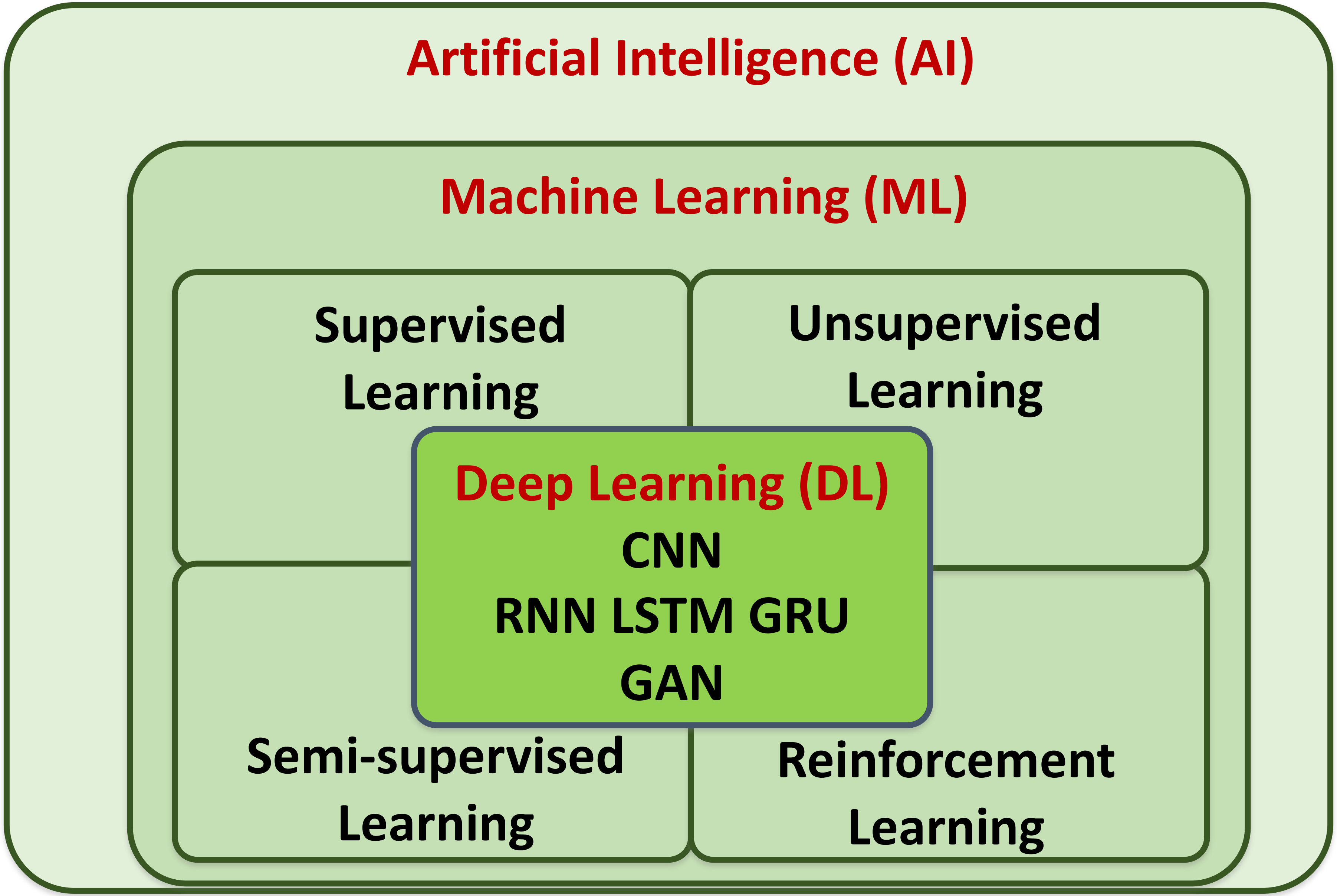
x y

Artificial Intelligence
Machine Learning & Deep Learning



Source: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

AI, ML, DL



Source: https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/deep_learning.html

The Theory of Learning

The Theory of Learning

- **Computational Learning Theory**
- **Probably approximately correct (PAC)**

The Theory of Learning

- How can we be sure that our learned hypothesis will predict well for previously unseen inputs?
  - How do we know that the hypothesis h is close to the target function f if we don't know what is?
- How many examples do we need to get a good h?
- What hypothesis space should we use?
- If the hypothesis space is very complex,
  can we even find the best h or do we have to settle for a local maximum?
- How complex should h be?
- How do we avoid overfitting?

Computational Learning Theory

- Intersection of AI, statistics, and theoretical computer science.
- Any hypothesis that is seriously wrong will almost certainly be "found out" with high probability after a small number of examples.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Probably approximately correct (PAC)

- Any hypothesis that is consistent with a sufficiently large set of training examples is unlikely to be seriously wrong.
- PAC learning algorithm:
  - Any learning algorithm that returns hypotheses that are probably approximately correct.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Linear function

$$y = f(x)$$

$$y = w_1 x + w_0$$

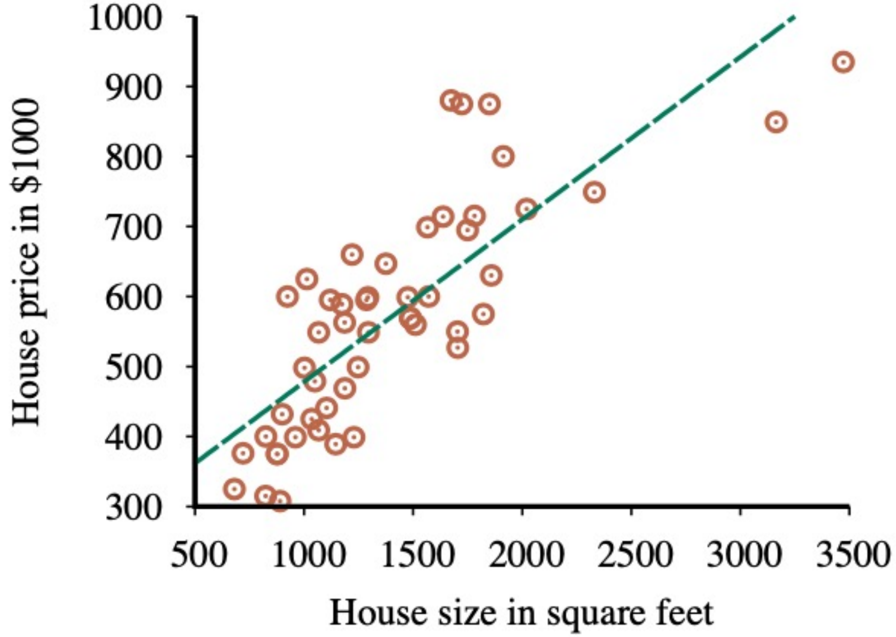
$$h_w(x) = w_1 x + w_0$$

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

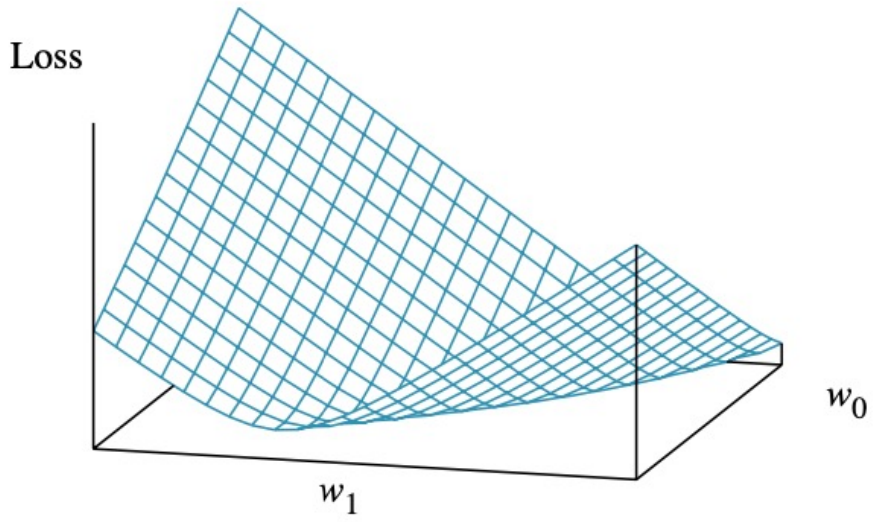
Linear Regression Weight Space

$$h_w(x) = w_1 x + w_0$$

$$w^* = \text{argmin}_w \, Loss(h_w)$$



$y = 0.232\,x + 246$



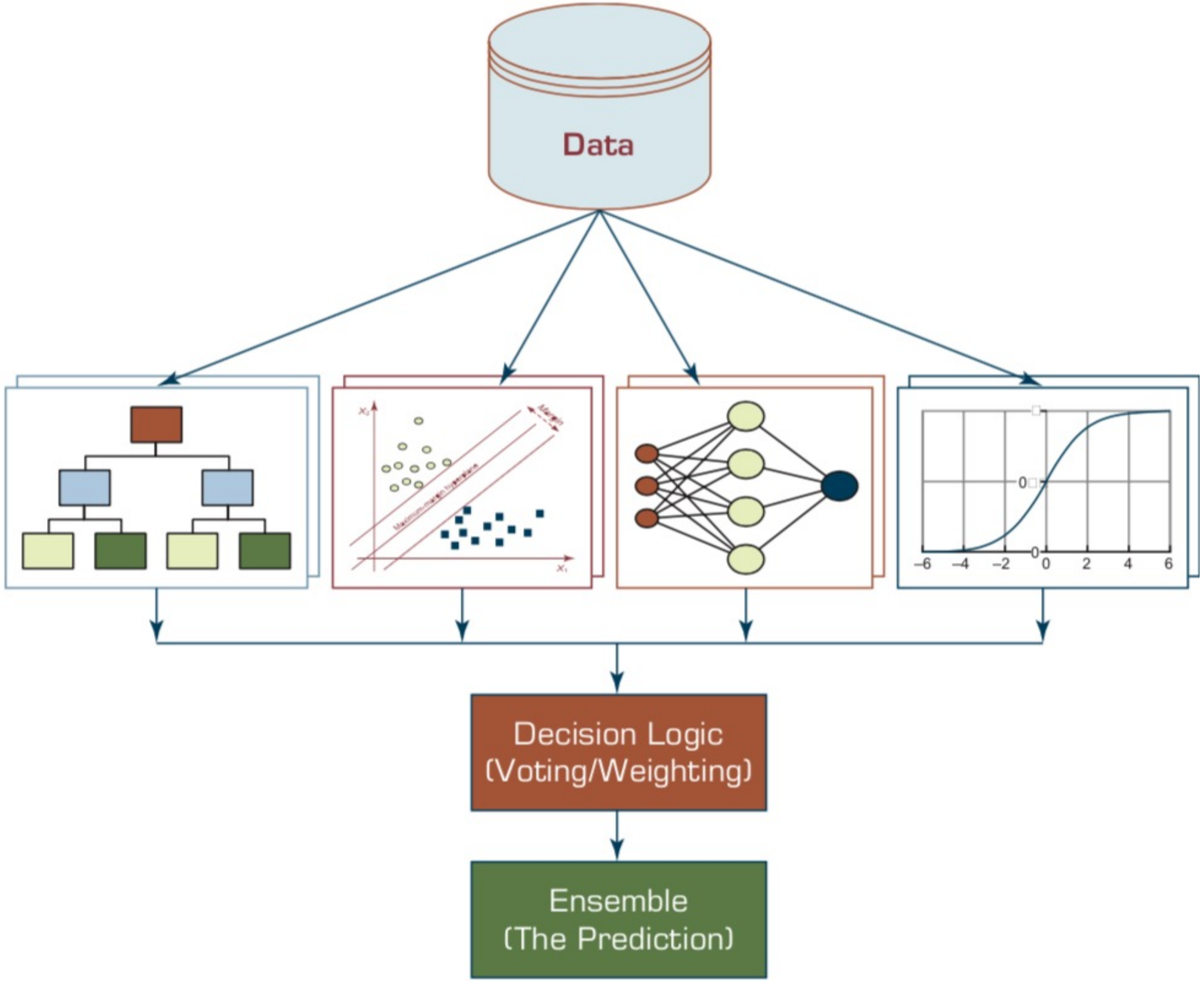
Loss function for Weights (w_1, w_0)

Ensemble Learning

Ensemble Learning

- **Select a collection**, or **ensemble**, of hypotheses, $h_1, h_2, \ldots, h_n$
, and **combine** their predictions
by **averaging, voting**, or
by another level of machine learning.

Ensemble Models
Heterogeneous Ensemble



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Ensemble Learning

- Base model
  - individual hypotheses
  - $h_1, h_2, \ldots, h_n$
- Ensemble model
  - hypotheses combination

Why Ensemble Learning

- **Reduce bias**
- **Reduce variance**

Ensemble Learning

- Bagging
  - Random forests
- Stacking
- Boosting
  - Gradient boosting
- Online learning

Ensemble Learning: Bagging

- **Bagging**
  - Generate distinct training sets by sampling with replacement from the original training set.
- Classification:
  - Plurality Vote (Majority Vote)
- Regression:
  - Average

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Ensemble Learning: Random forests

- **Random forest** model is a form of **decision tree bagging** in which we take extra steps to make the ensemble of trees more diverse, to reduce variance.
- The key idea is to randomly vary the **attribute** choices (rather than the training **examples**)

Ensemble Learning: Random forests

- **Extremely randomized trees (ExtraTrees)**
  - **Use randomness in selecting the split point value**
  - **for each selected attribute, we randomly sample several candidate values from a uniform distribution over the attribute's range**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Ensemble Learning: Stacking

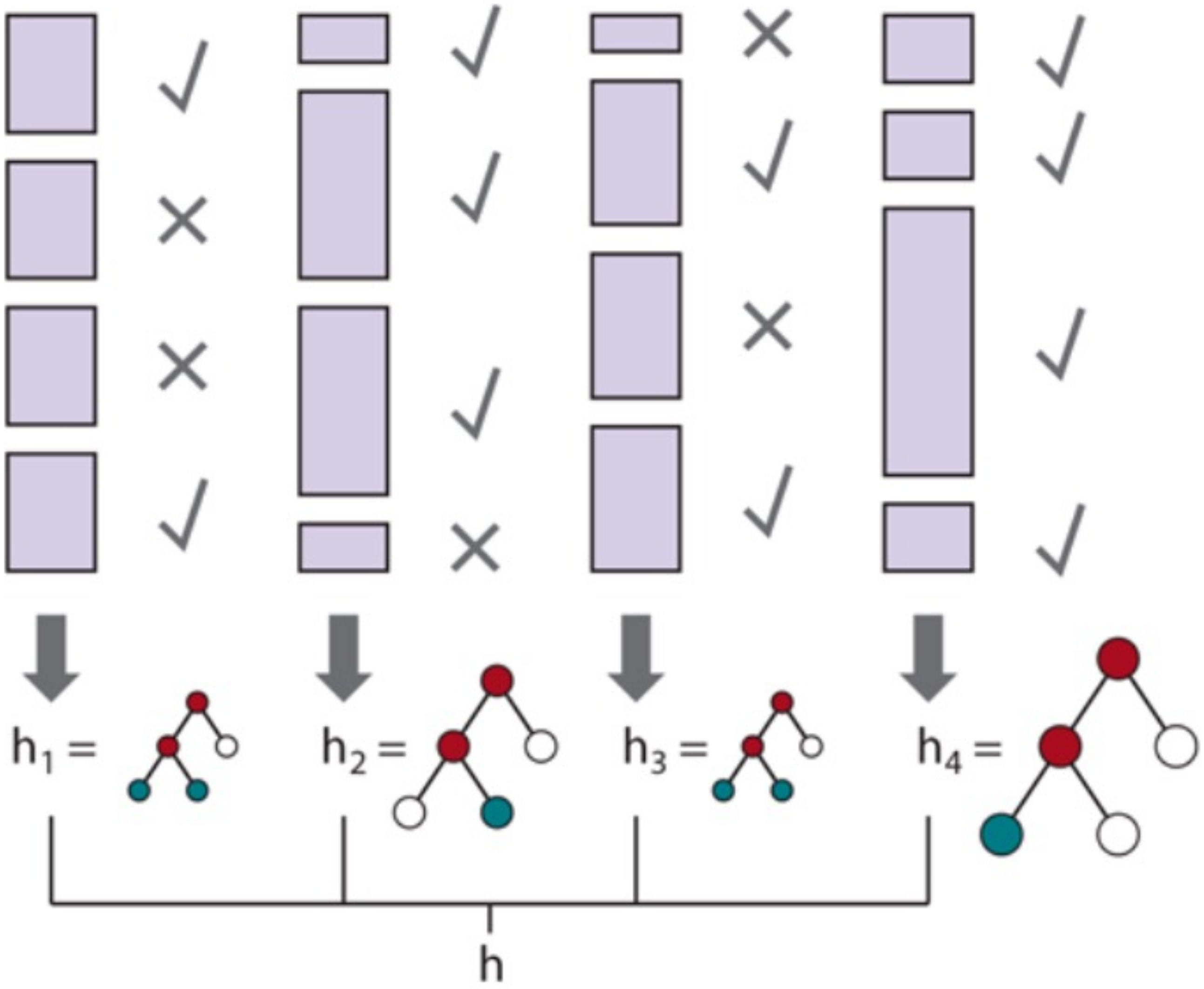
- Staking
  - Stacked generalization combines multiple base models from different model classes trained on the same data.
- Bagging
  - Combines multiple base models of the same model class trained on different data.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Ensemble Learning: Boosting

- **Boosting**
  - **The most popular ensemble method**
- **Weighted training set**

Ensemble Learning: Boosting



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Ensemble Learning: Gradient boosting

- **Gradient boosting**
  - Gradient boosting is a form of boosting using gradient descent
- **Gradient boosting machines (GBM)**
- **Gradient boosted regression trees (GBRT)**
- Popular method for regression and classification of factored tabular data

Ensemble Learning: Online learning

- **Online learning**
  - **Data are not i.i.d. (independent and identically distributed)**
  - **An agent receives an input x_i from nature, predicts the corresponding y_i and then is told the correct answer.**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

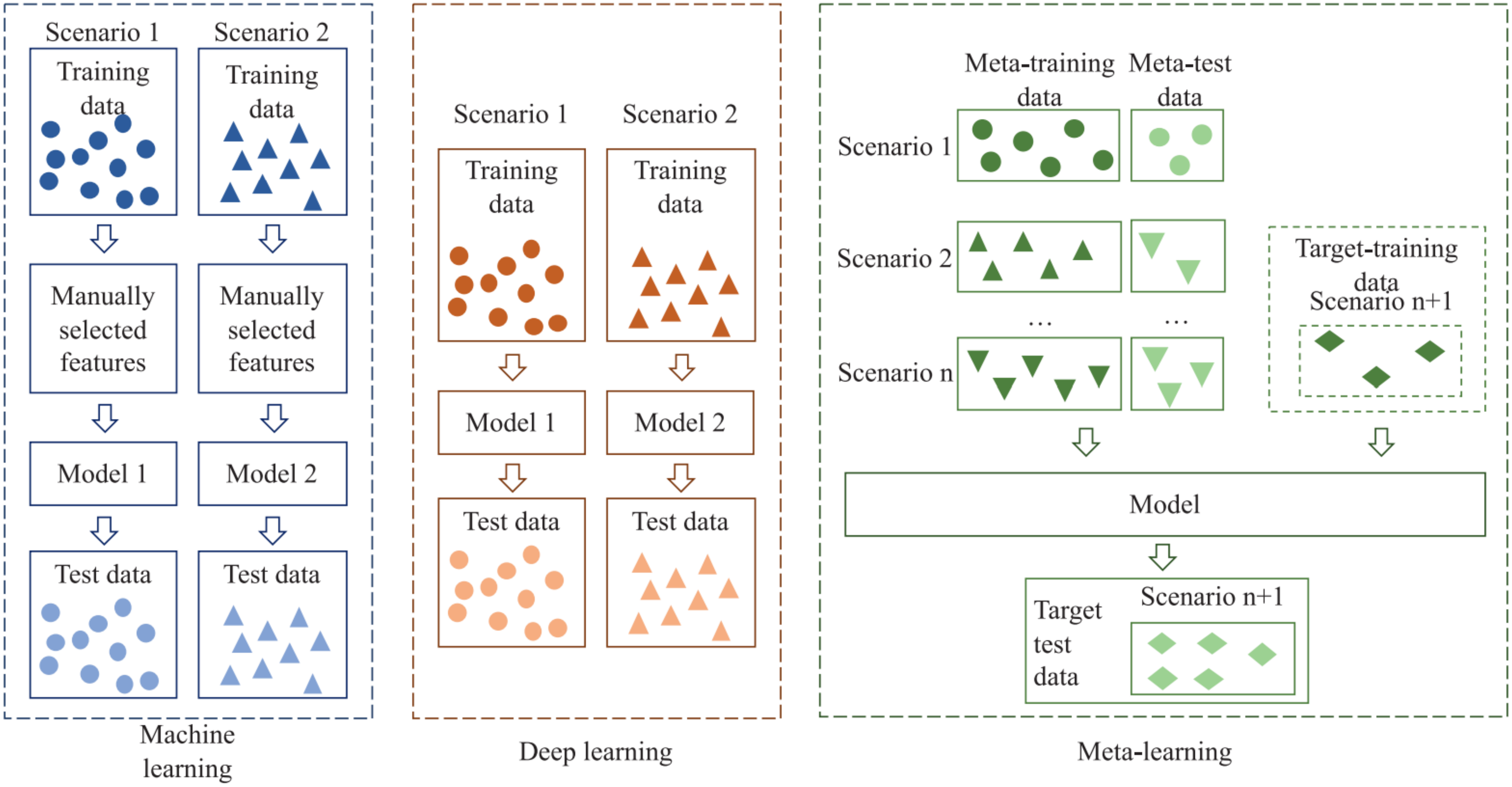
Meta Learning:
Learning to Learn

Deep Learning
Transfer Learning
Few-Shot Learning
Meta Learning

Deep Learning, Transfer Learning, Few-Shot Learning, Meta Learning

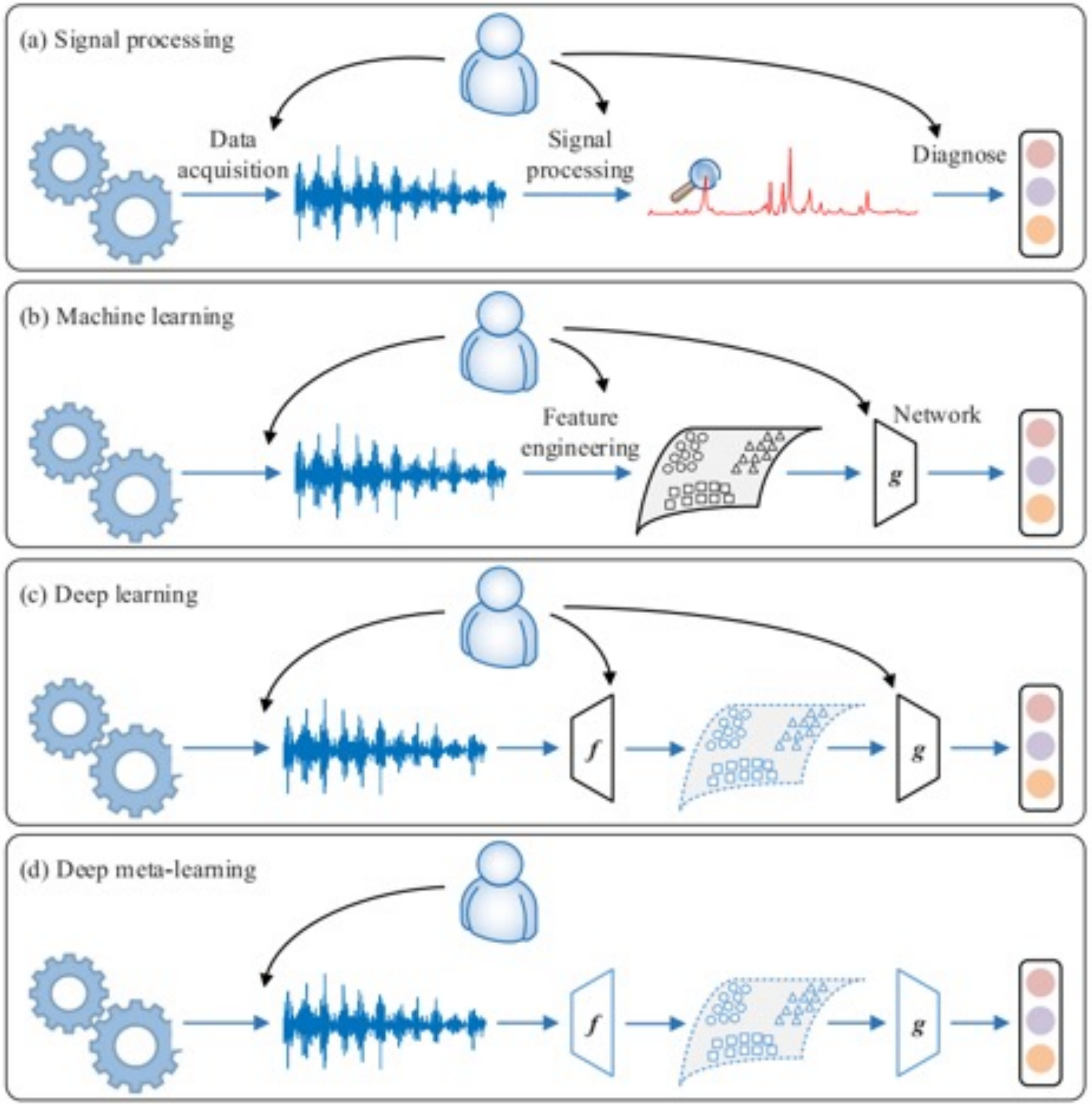
- Deep Learning
  - Transfer Learning
    - Pre-training, Fine-Tuning (FT)
- Meta Learning: Learning to Learn
- Few-Shot Learning (FSL)
- One-Shot Learning (1SL)
- Zero-Shot Learning (0SL)(ZSL)

Machine Learning, Deep Learning, Meta Learning



Source: Luo, Shuai, Yujie Li, Pengxiang Gao, Yichuan Wang, and Seiichi Serikawa. "Meta-seg: A survey of meta-learning for image segmentation." Pattern Recognition (2022): 108586.

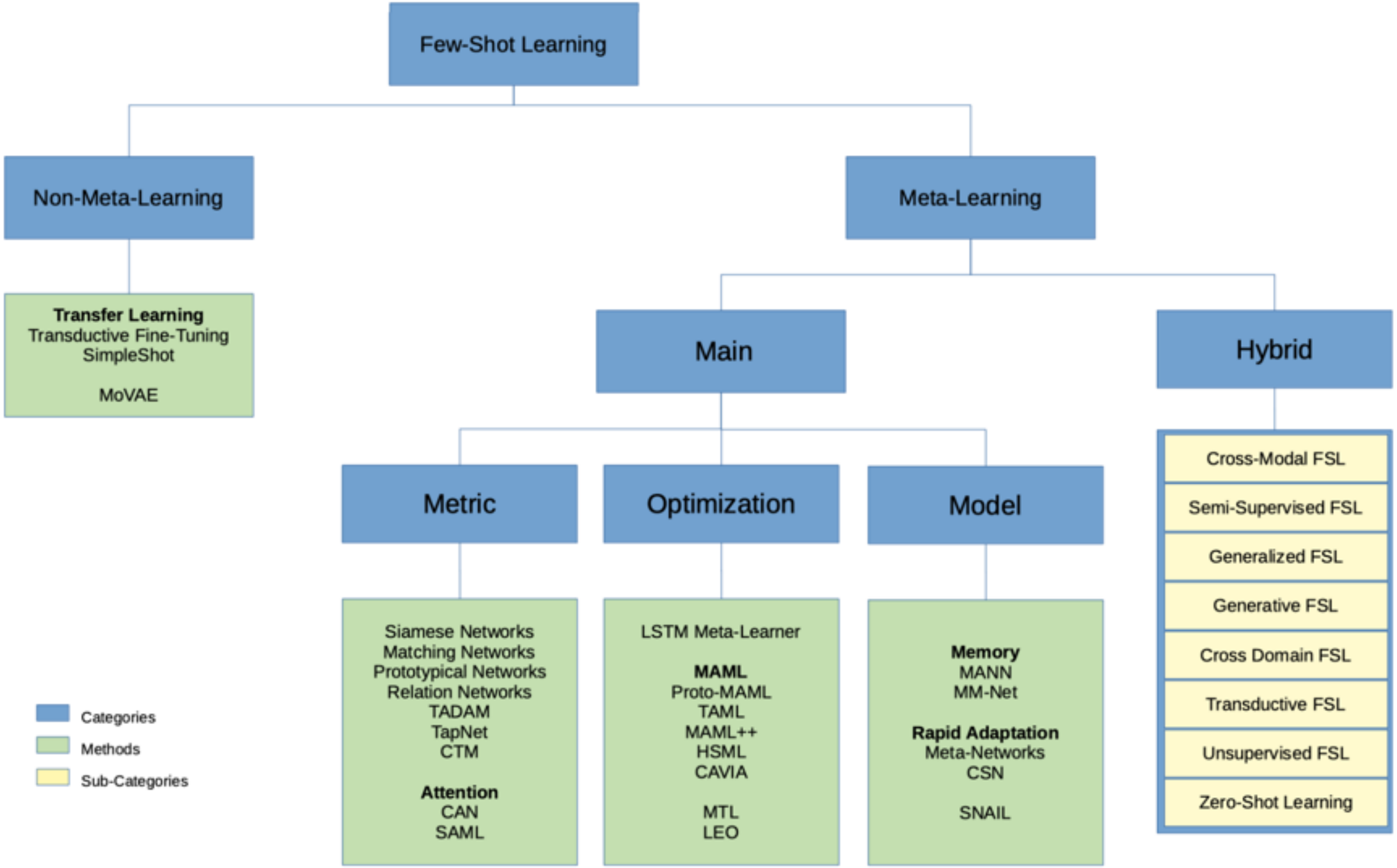
Machine Learning, Deep Learning, Meta Learning



Source: Feng, Yong, Jinglong Chen, Jingsong Xie, Tianci Zhang, Haixin Lv, and Tongyang Pan. "Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects." Knowledge-Based Systems 235 (2022): 107646.

Few-Shot Learning (FSL) and Meta Learning

Machine learning from few training examples



Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Meta Learning, Transfer Learning, Ensemble Learning, Continual Learning, Multi-Task Learning

Features	Method					
	Meta-learning	Transfer learning	Ensemble learning	Continual learning	Multi-task learning	Hierarchical Bayesian models
Learning from prior experience	√	√	✗	√	✗	√
Relationship between source tasks	No limitation	Related	Same	Task streams	Related	Related
Relationship between source tasks and target tasks	No limitation	Related	Same	Related	Related	Related
Considering the requirements of the target task	√	✗	✗	✗	✗	✗

Source: Luo, Shuai, Yujie Li, Pengxiang Gao, Yichuan Wang, and Seiichi Serikawa. "Meta-seg: A survey of meta-learning for image segmentation." Pattern Recognition (2022): 108586.

Meta-Learning and Few-shot Learning
Notations and Terms

Optimization-based Meta-learning

Notation A	Term A
$\mathcal{D}_i^{train}$	Training set for task $\mathcal{T}_i$
$\mathcal{D}_i^{test}$	Test set for task $\mathcal{T}_i$
$\mathcal{D}_{meta-train}$	Meta-training set
$\mathcal{D}_{meta-test}$	Meta-testing set

Metric-based Meta-learning

Notation B	Term B
S_i	Support Set for task $\mathcal{T}_i$
Q_i	Query Set for task $\mathcal{T}_i$
$\mathcal{D}_{train}$	Training Set
$\mathcal{D}_{test}$	Test Set

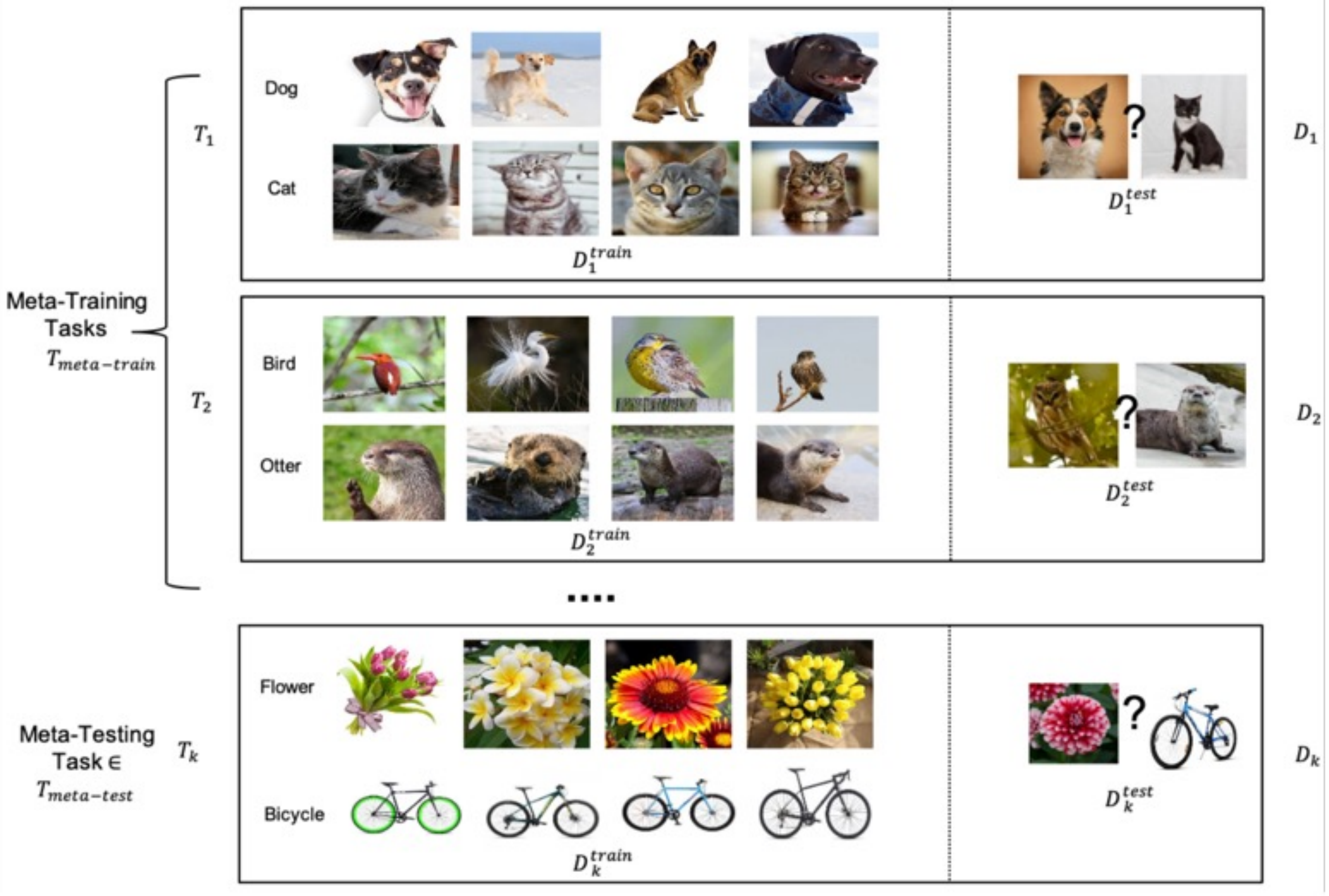
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Meta-Learning Symbols

Symbol	Meaning
$\mathcal{T}_i$	Task i
$\mathcal{L}$	Loss function
(x_k, y_k)	Input-Output pair
f_θ	Model (function) with parameters θ
g_{θ_1}	Embedding function
d_{θ_2} or d	Distance function
g_ϕ	Meta-Learning model with parameters ϕ
$P_\theta(y\|x)$	Output probability of y for input x using model parameters θ
$k_\theta(x_1, x_2)$	Kernel function measuring similarity between two vectors x_1 and x_2
σ	Softmax function
α, β	Learning rates
w	Weights
$\mathbf{v}_c$	Prototype of class c
C	Set of classes present in S
S^c	Subset of S containing all elements (x_k, y_k) such that $y_k = c$
$\bigoplus$	Concatenation operator
B	Number of batches (X_b, Y_b) sampled in inner-loop for a randomly sampled task $\mathcal{T}_i$
I	Number of tasks $\mathcal{T}_i$ sampled in inner-loop
J	Number of outer-loop iterations

Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

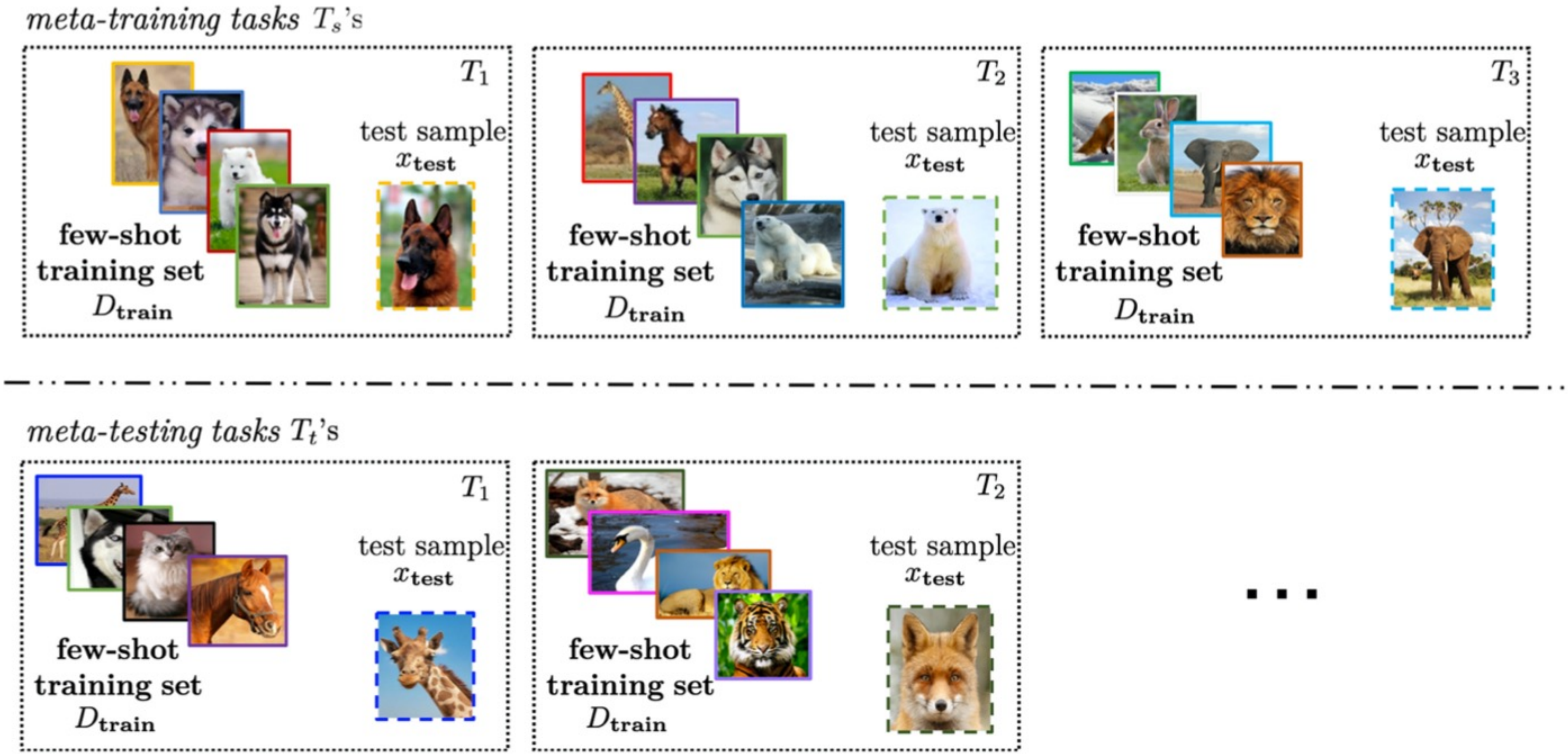
Meta-Learning Example Setup



Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Few-Shot Learning (FSL)

Solving the FSL problem by meta-learning

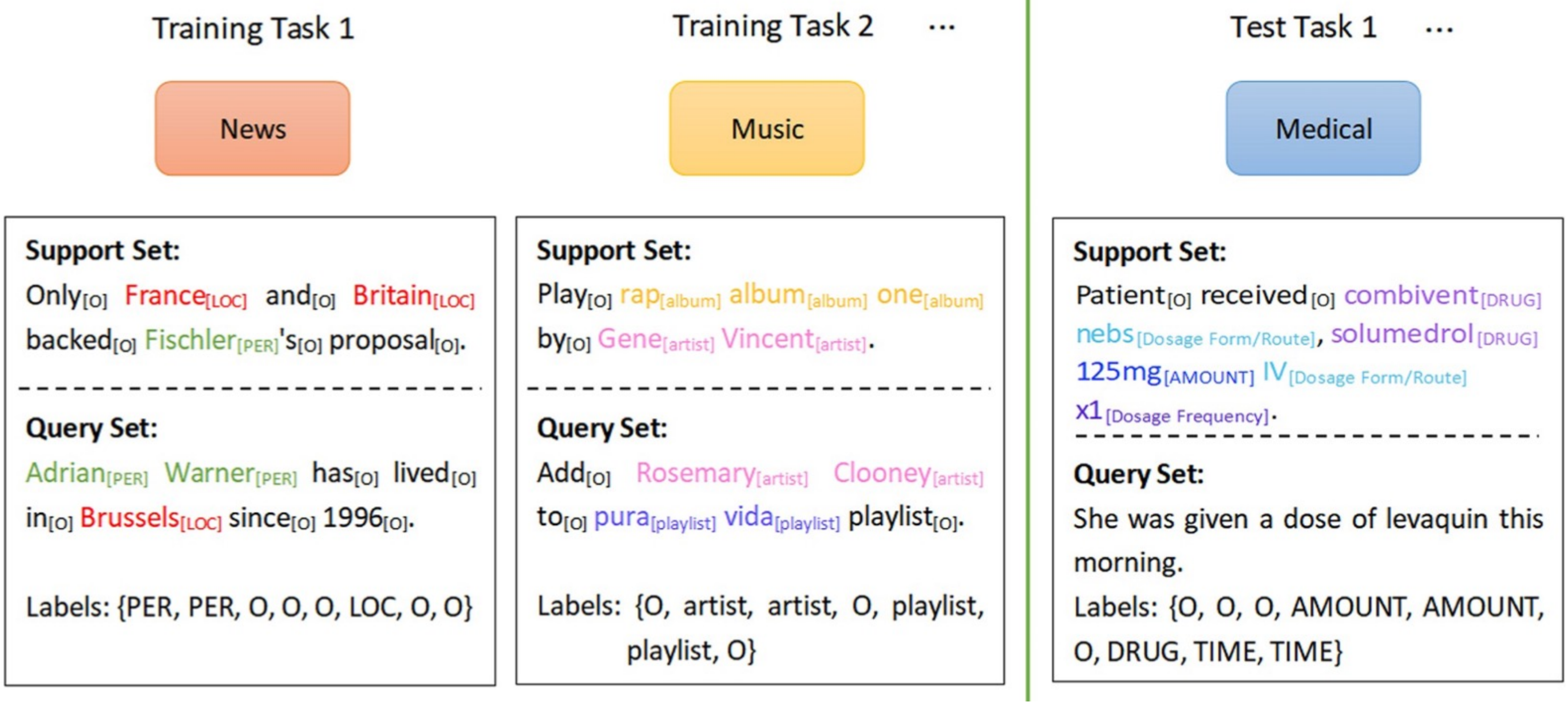


Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

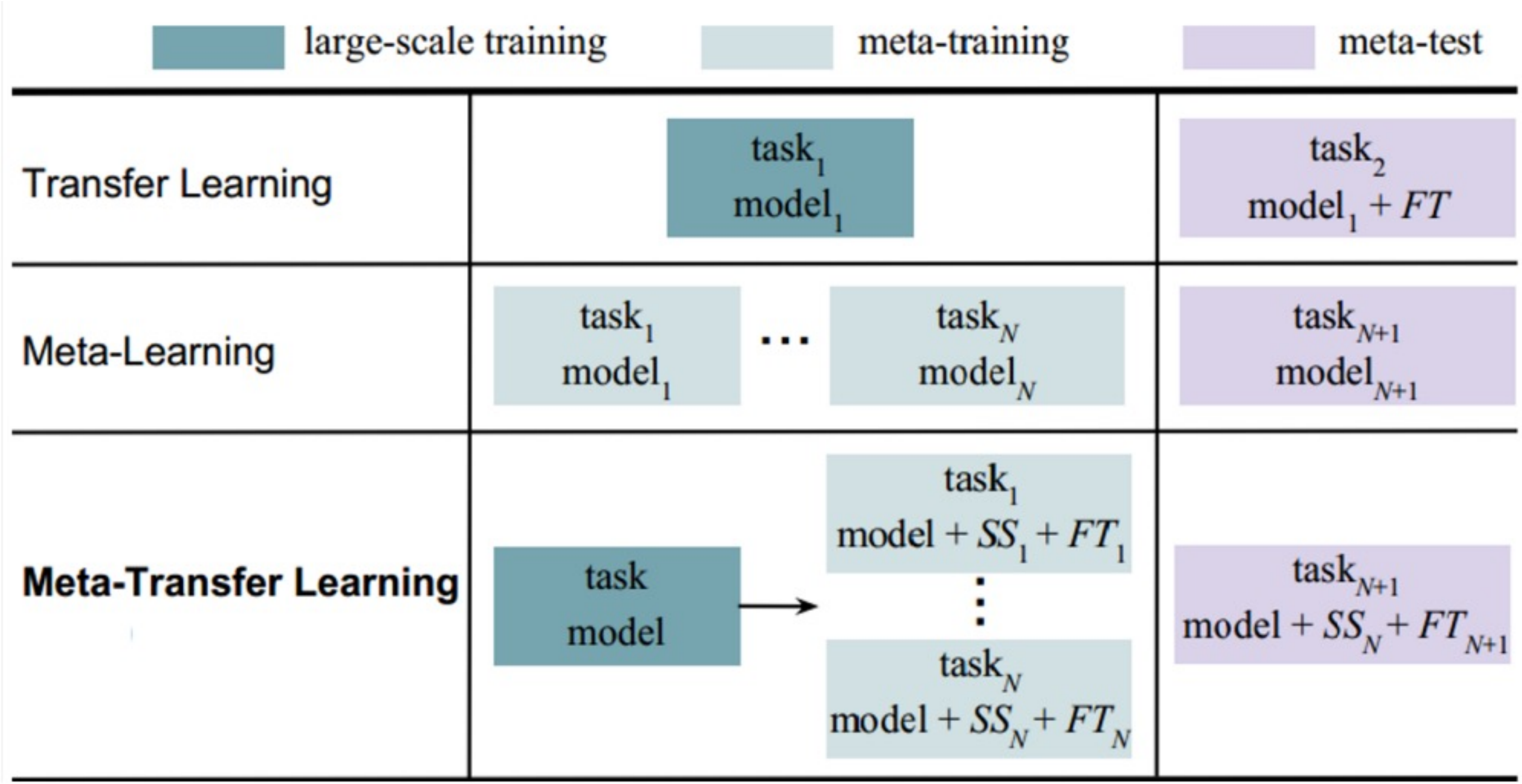
Meta-learning

Each task mimics the few-shot scenario, and can be completely non-overlapping.
Support sets are used to train; query sets are used to evaluate the model



Meta-Task Learning (MTL)

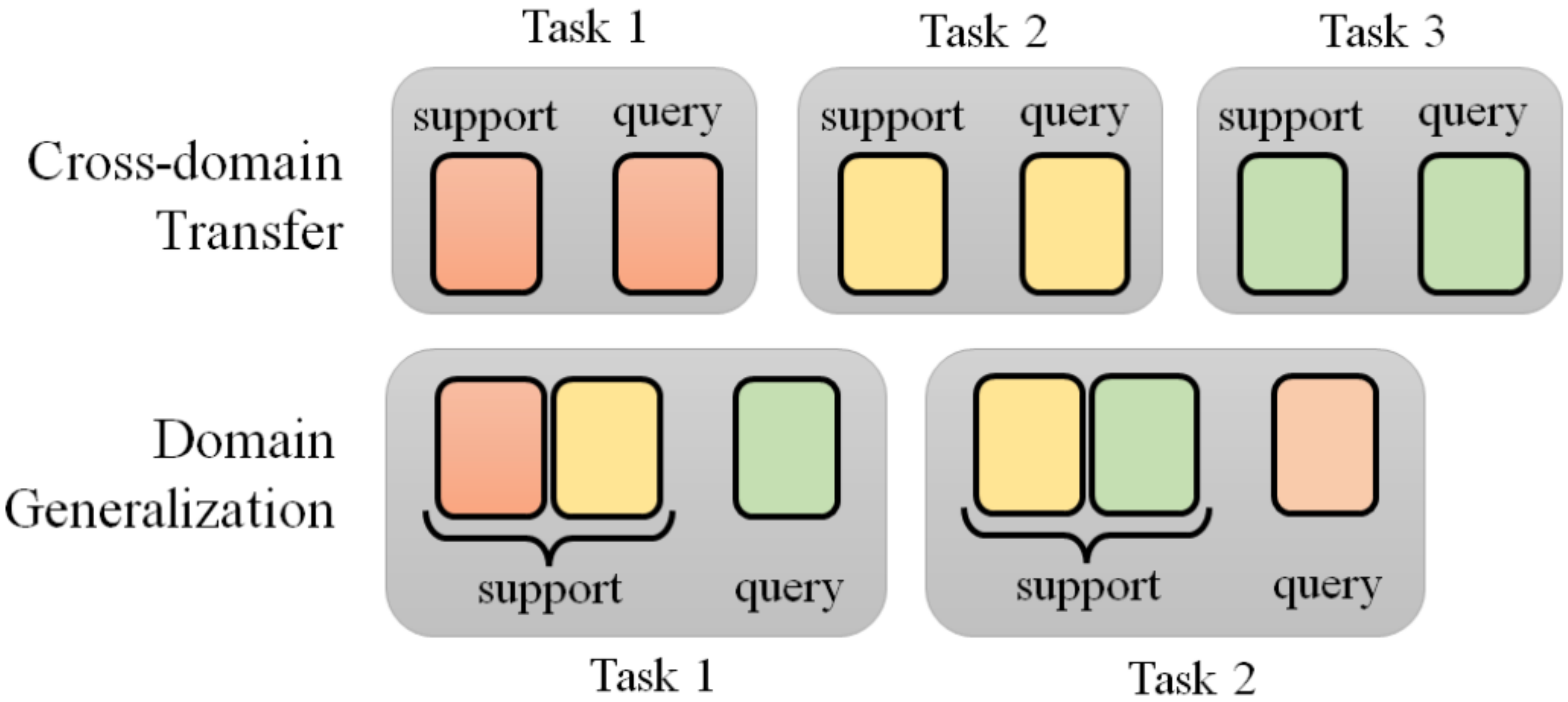
Transfer Learning Strategy for Meta-Learning



Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

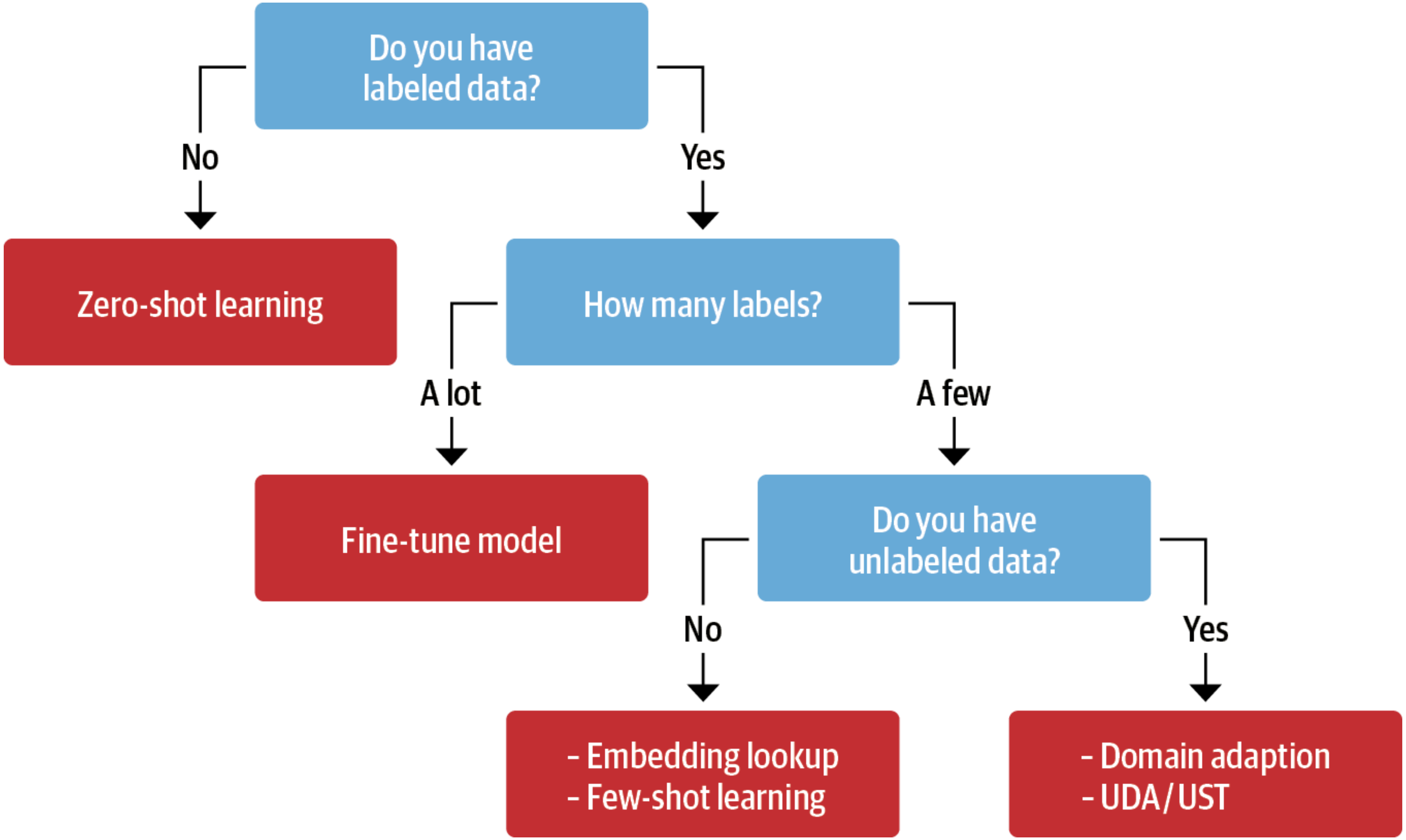
Meta Learning

The task construction of cross-domain transfer and domain generalization



Source: Lee, Hung-yi, Shang-Wen Li, and Ngoc Thang Vu. "Meta Learning for Natural Language Processing: A Survey." arXiv preprint arXiv:2205.01500 (2022).

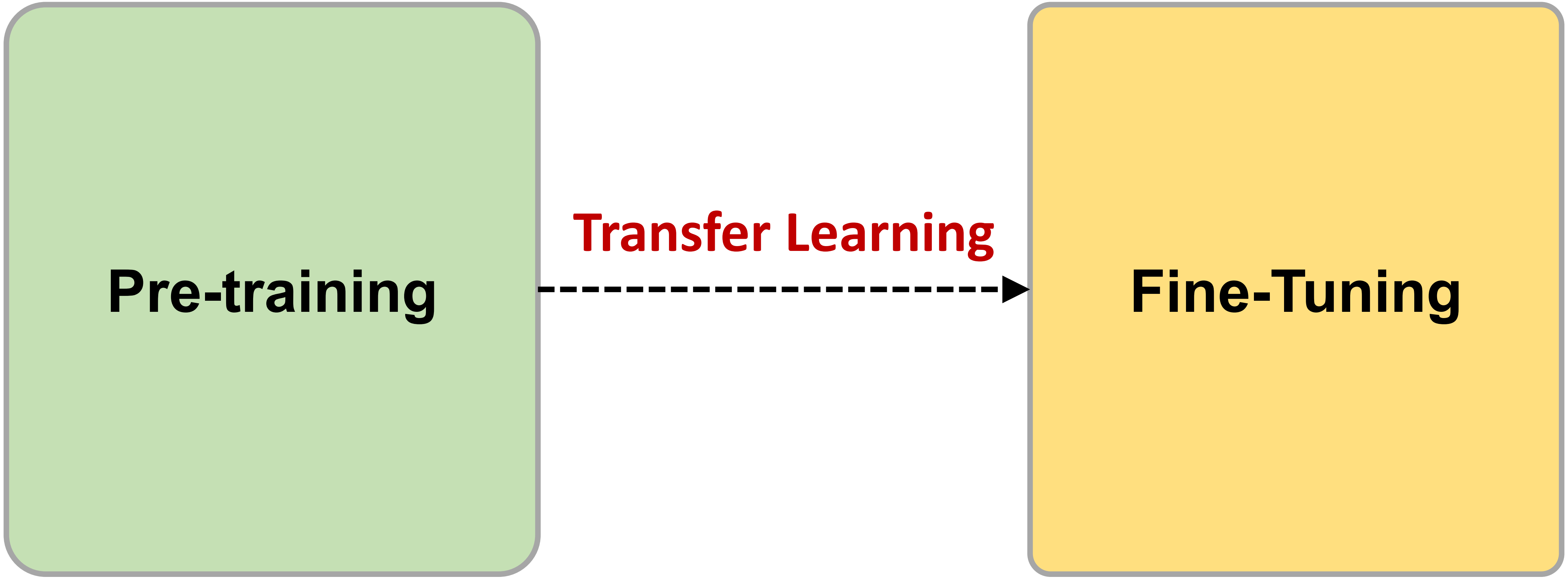
Transfer Learning, Fine-tuning, Few-shot learning



Unsupervised Data Augmentation (UDA)
Uncertainty-aware Self-Training (UST)

Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.

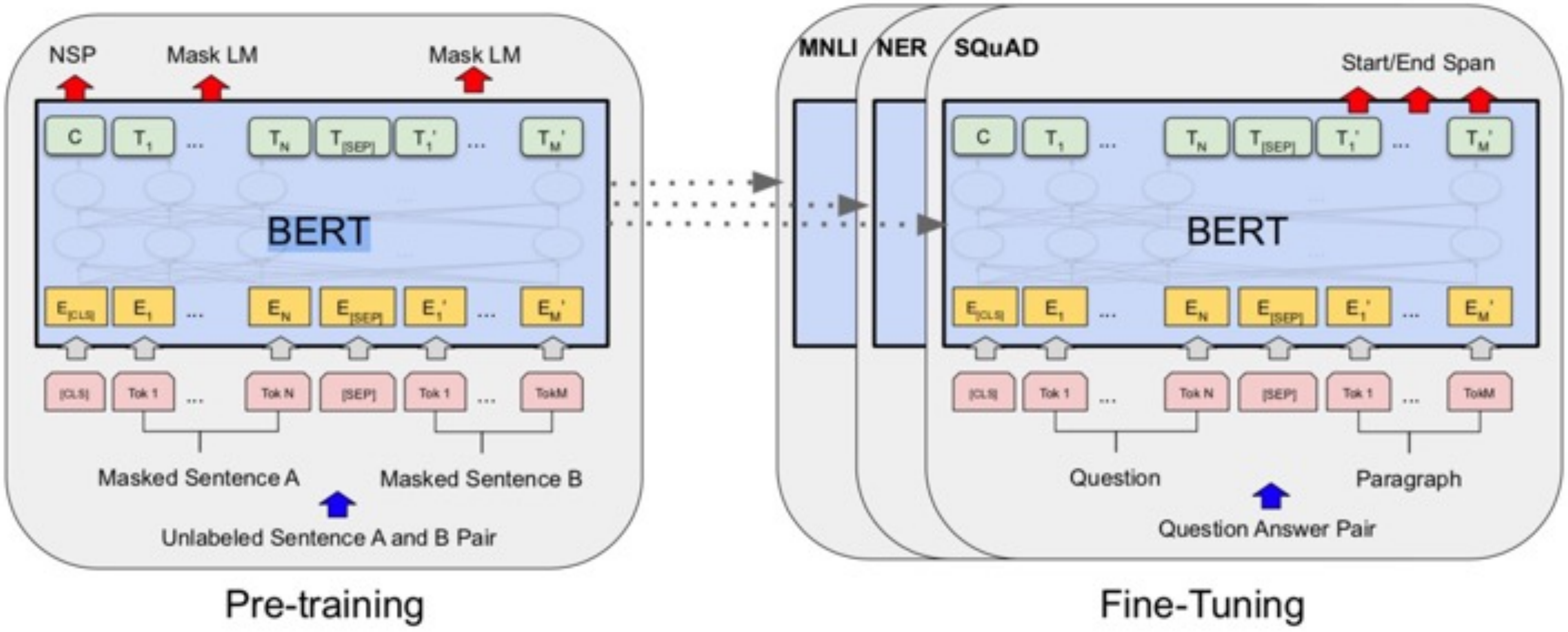
Transfer Learning



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

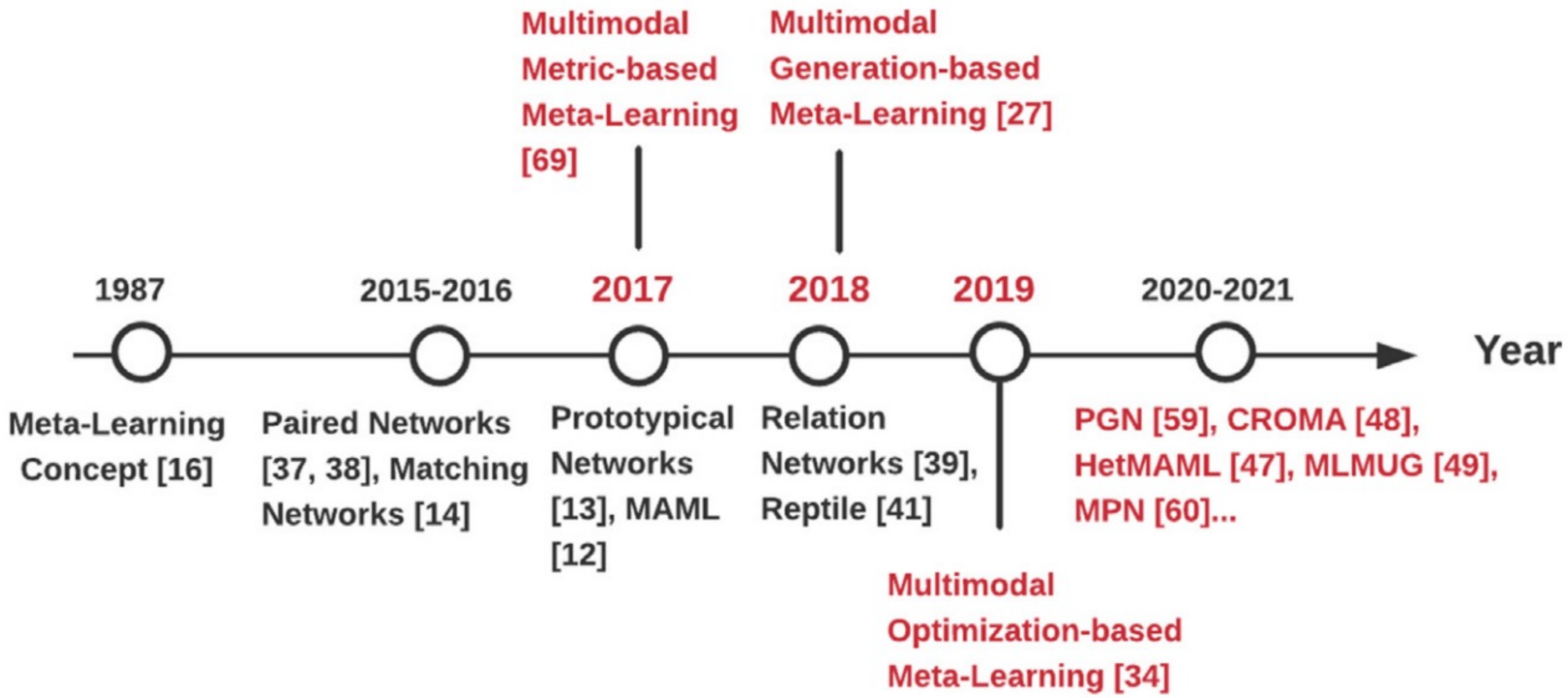
BERT (Bidirectional Encoder Representations from Transformers)

Overall pre-training and fine-tuning procedures for BERT



Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Meta Learning



Source: TMa, Yao, Shilin Zhao, Weixiao Wang, Yaoman Li, and Irwin King. "Multimodality in meta-learning: A comprehensive survey." Knowledge-Based Systems (2022): 108976.

Meta Learning

Year	Achievement	Ref.
1987	(1) A new framework of "learning how to learn" with self-referential learning was proposed. The neural networks in self-referential learning can regard their weights as inputs and update them continuously. (2) Based on the conventional neural network, two types of wights were used to connect the neurons. Each type of weight presents a different learning speed.	[34,35]
1990	A synaptic learning rule, which is biologically plausible, was proposed to automatically study the learning rules.	[36]
1993	A chain of meta-networks was introduced to improve the learning capacity of a recurrent neural network for a dynamic environment.	[37]
1995	A framework was proposed to optimize the learning rule within a parametric learning rule space.	[38]
1996	An improved self-referential model was proposed. Time ratios were used to measure the effects of learning processes on the later learning processes.	[39]
1998	The term "Learning to learn" was proposed to equally represent the concept of meta-learning.	[40]
2001	Gradient descent methods were firstly used in meta-learning instead of evolutionary methods, which were widely used in previous research.	[41,42]
2003	A biologically plausible meta-reinforcement learning algorithm was proposed to tune the parameters of the meta-learning model dynamically and adaptively.	[43]
2004	A new perspective of meta-learning was proposed: exploring the interaction between the learning mechanism and the specific contexts to which the mechanism applies.	[9]
2008	The zero-data learning problem was addressed.	[44]
2010–2012	The breakthrough of deep neural networks marks the beginning of the era of meta-learning.	[45–47]
2013	The relationship between transfer learning and meta-learning was described.	[48]
2016	A meta-learning algorithm named gradient descent by gradient descent was proposed.	[49]
2017	(1) MAML was proposed. (2) A doctoral thesis systematically introduced the concept of meta-learning and corresponding methods.	[50,51]
2018	Reptile, an improved version of MAML, was proposed.	[52]
2019	The Capsule network provides a new method to improve the learning capacity of meta-learning, especially in computer vision.	[53]
2020	Combining auto-encoder and capsule network to focus on the zero-shot learning problem.	[54]

Source: TMa, Yao, Shilin Zhao, Weixiao Wang, Yaoman Li, and Irwin King. "Multimodality in meta-learning: A comprehensive survey." Knowledge-Based Systems (2022): 108976.

Meta-learning Approaches

	Metric-based	**Optimization-based**	**Model-based**
Key idea	Metric Learning [19]	Gradient Descent	Memory; RNN
How $P_\theta(y\|x)$ is modeled?	$\sum_{(x_k,y_k)\in S} k_\theta(x,x_k)y_k,$	$P_{\theta'}(y\|x),$ where $\theta' = g_\phi(\theta, S)$	$f_\theta(x,S).$
Advantages	Faster Inference. Easy to deploy.	Offers flexibility to optimize in dynamic environments. S can be discarded post-optimization.	Faster inference with memory models. Eliminates the need for defining a metric or optimizing at test.
Disadvantages	Less adaptive to optimization in dynamic environments. Computational complexity grows linearly with size of S at test.	Optimization at inference is undesirable for real-world deployment. Prone to overfitting.	Less efficient to hold data in memory as S grows. Hard to design.

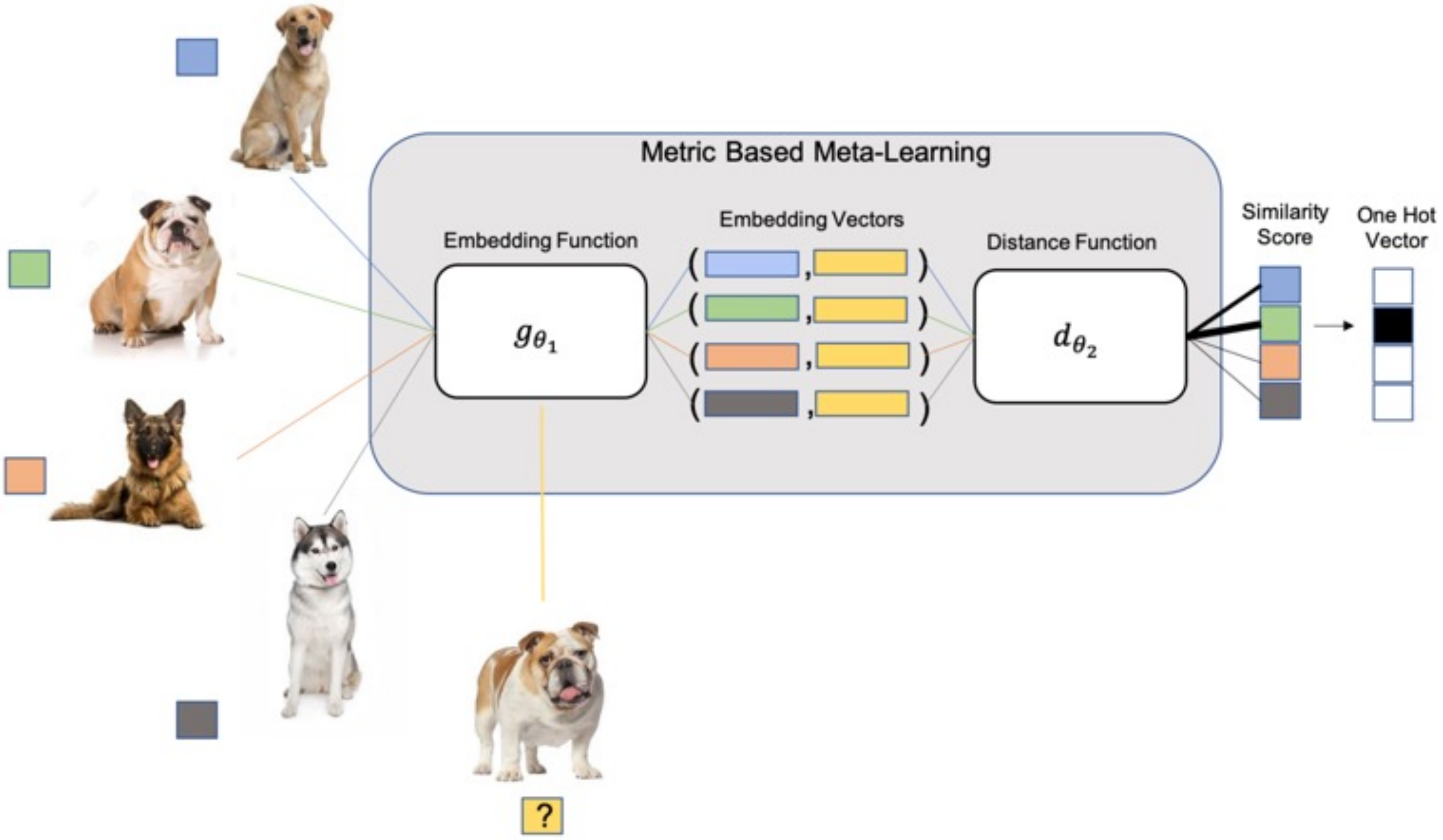
Meta Learning: Learning to Learn

Class	Methods	Reference	Summary
Metric-Based	Siamese Neural Networks Matching Networks Prototype Networks Relation Networks	[32–36] [37–41] [42–46] [47–53]	We show four metric-based meta-learning algorithms, focusing on feature extractors, similarity metrics, and automatic algorithm selection However, the metric-based approaches are sensitive to the dataset and increase the computational expenditure when the number of tasks is large.
Model-Based	Memory-Augmented Neural Networks Meta Networks Simple Neural Attentive Meta-Learner	[54–56] [57,58] [59–65] [66–71]	We display three model-based approaches. MANN combines neural networks with external memory modules, but the model is complex. Meta-Net is computationally intensive and has high memory requirements. SNAIL is relatively simplified, but has to be optimized in terms of automatic parameter tuning and reducing computation.
Optimization-Based	MAML META- LSTM META- SGD	[72–80] [81–86] [87–93]	We present three methods of optimization-based meta-learning. MAML is relatively simple to implement, but the capacity of the model is limited. Meta-LSTM has a large capacity, but a complicated training process. Meta-SGD has improved capacity but still has difficulties in generalization ability.

Source: Tian, Yingjie, Xiaoxi Zhao, and Wei Huang. "Meta-learning approaches for learning-to-learn in deep learning: A survey." Neurocomputing 494 (2022): 203-223.

Metric-based Meta-learning

M-Way K-Shot Task (4-way-1-shot classification task)



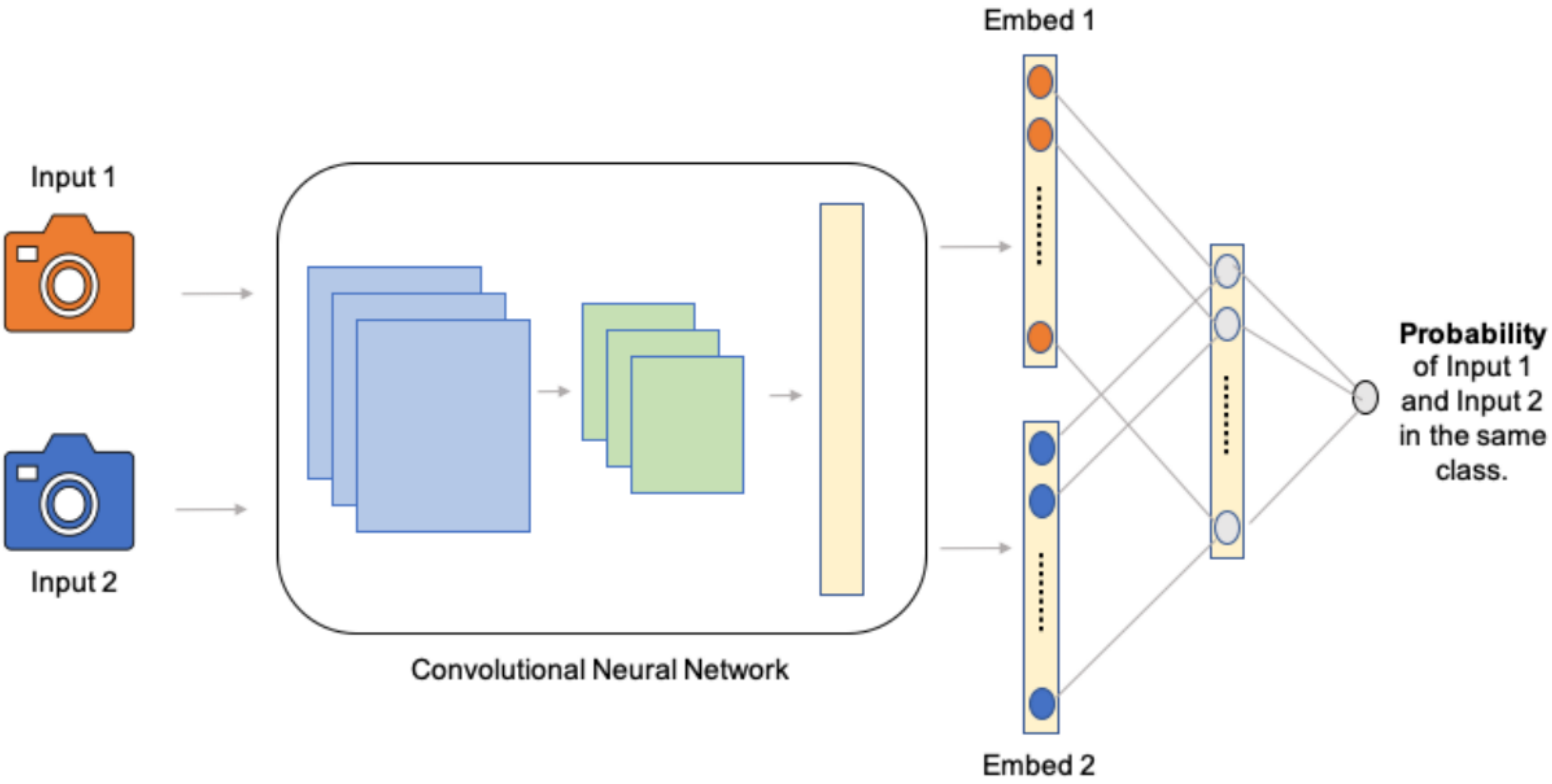
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Metric-based Meta-Learning Methods

Method	T.I	g_{θ_1}	d_{θ_2}	Prediction	Loss
Siamese Networks [20]	Yes	CNN	L1	$v = w \cdot d(g_{\theta_1}(x_1), g_{\theta_2}(x_2))$ $p = \text{sigmoid}(\sum_j v_j)$	$-(y \log(p) + (1-y)(\log(1-p))$
Matching Networks [13]	Yes	CNN + LSTM w/ attention	Cosine Similarity	$\hat{y} = \sum_{k=1}^{t} \sigma(d(f_\theta(\hat{x}), g_{\theta_1}(x_k))y_k$ $P(y = c\|\hat{x}) = \hat{y}_c$	$-\log P$
Prototypical Networks [21]	Yes	CNN	Euclidean	$P(y = c\|x) = \sigma(-d(g_{\theta_1}(x), \mathbf{v}_c))$	$-\log P$
Relation Networks [22]	Yes	CNN	Learned by CNN	$r_c = d_{\theta_2}(g_{\theta_1}(x) \oplus \mathbf{v_c}))$	$\sum_{c \in C} (r_c - \mathbf{1}(y == c))^2$
TADAM [16]	No	ResNet-12	Cosine / Euclidean	$P_\lambda(y = c\|x) = \sigma(-\lambda d(g_{\theta_1}(x, \Gamma), \mathbf{v}_c))$	$-\log P$
TapNet [23]	No	Resnet-12	Euclidean	$P(y = c\|x) = \sigma(-d(\mathbf{M}(g_{\theta_1}(x)), \mathbf{M}(\Phi_c)))$	$-\log P$
CTM [24]	No	Any	Any	-	-

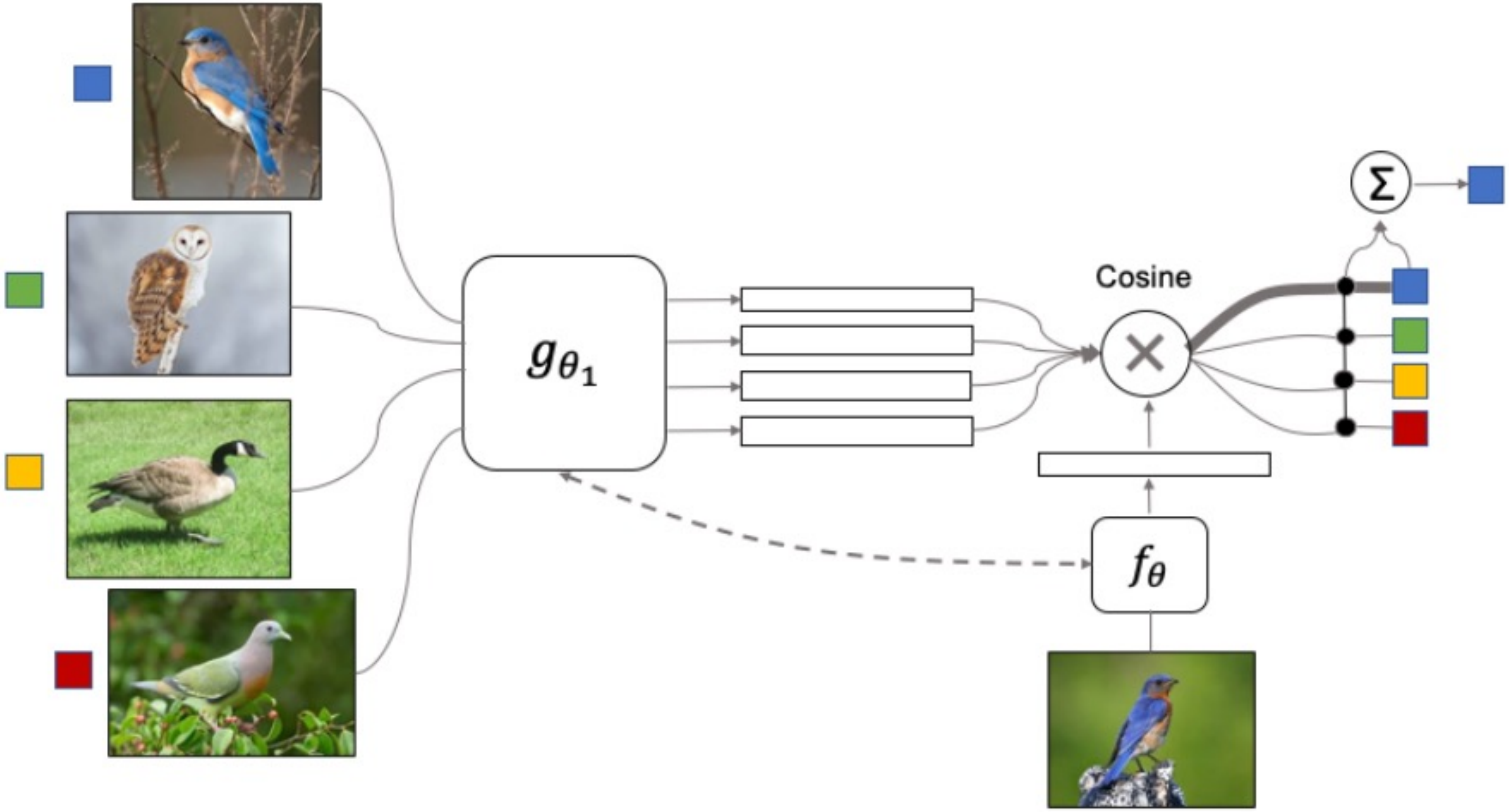
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Convolutional Siamese Network



Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

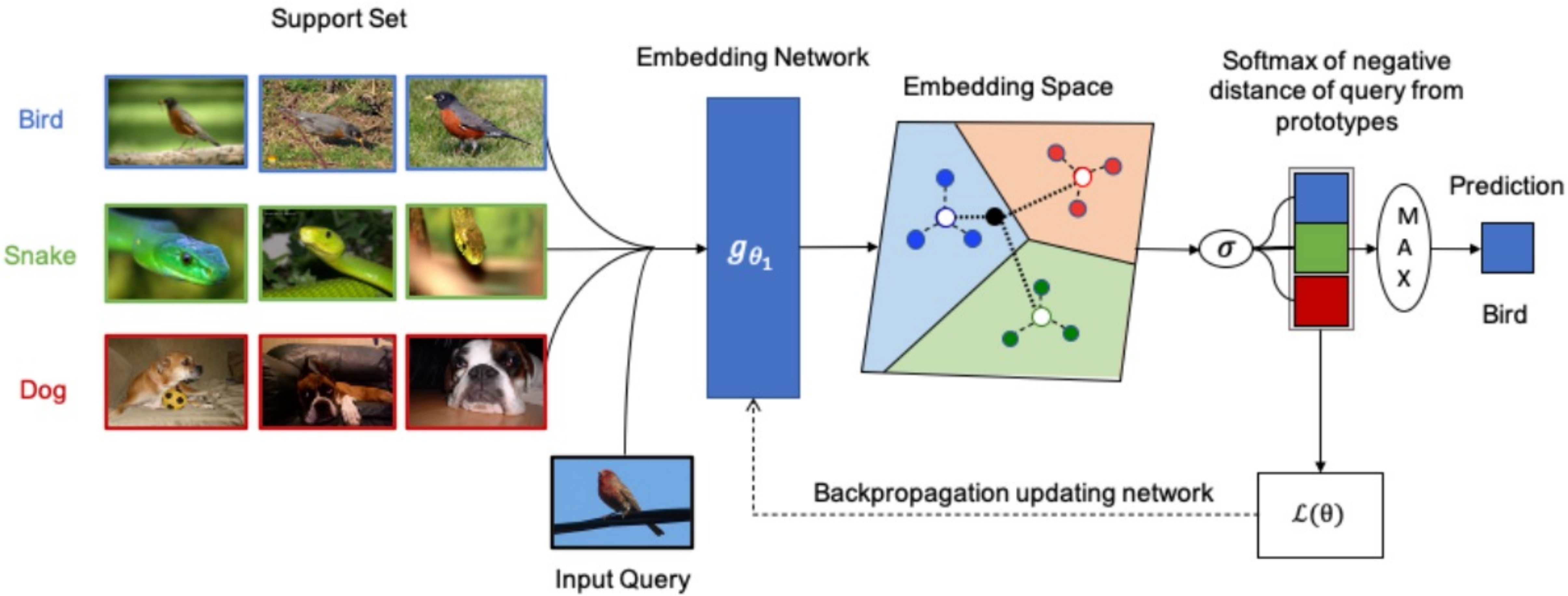
Meta Learning: Matching Networks



Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

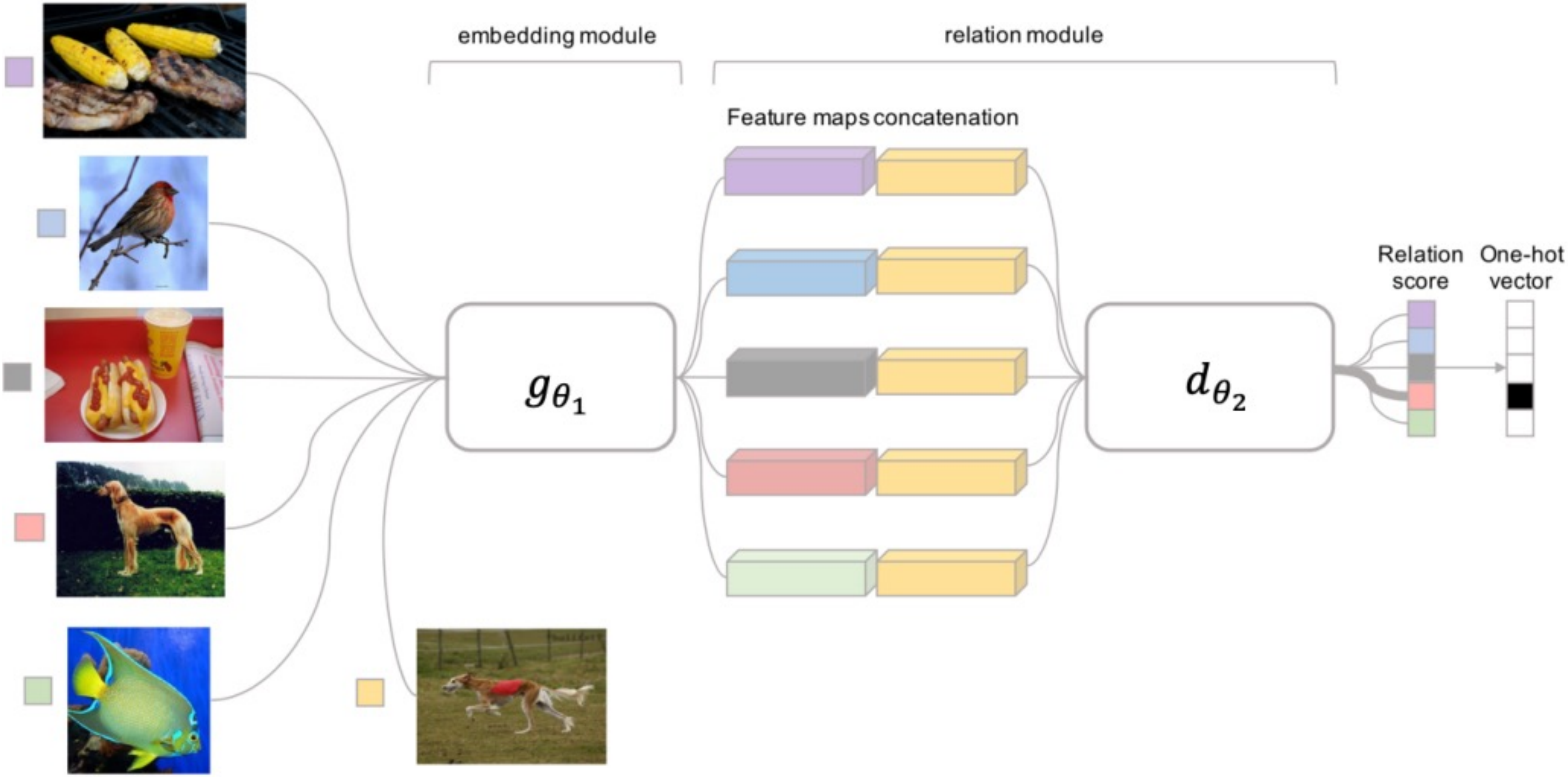
Few-shot Prototypes

v_c are computed as the mean of embedded support examples for each class



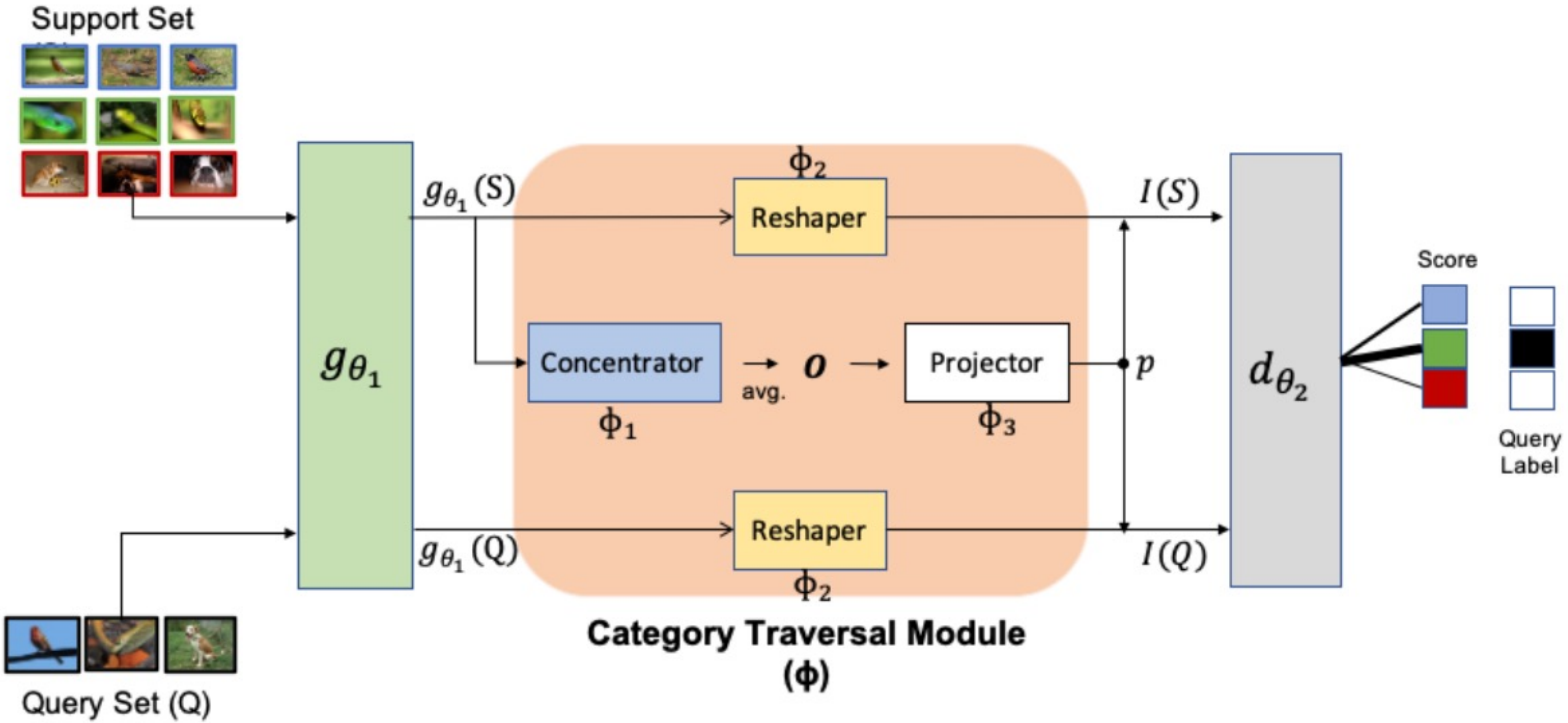
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Meta Learning: Relation Network



Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Meta Learning: Category Traversal Module (CTM)



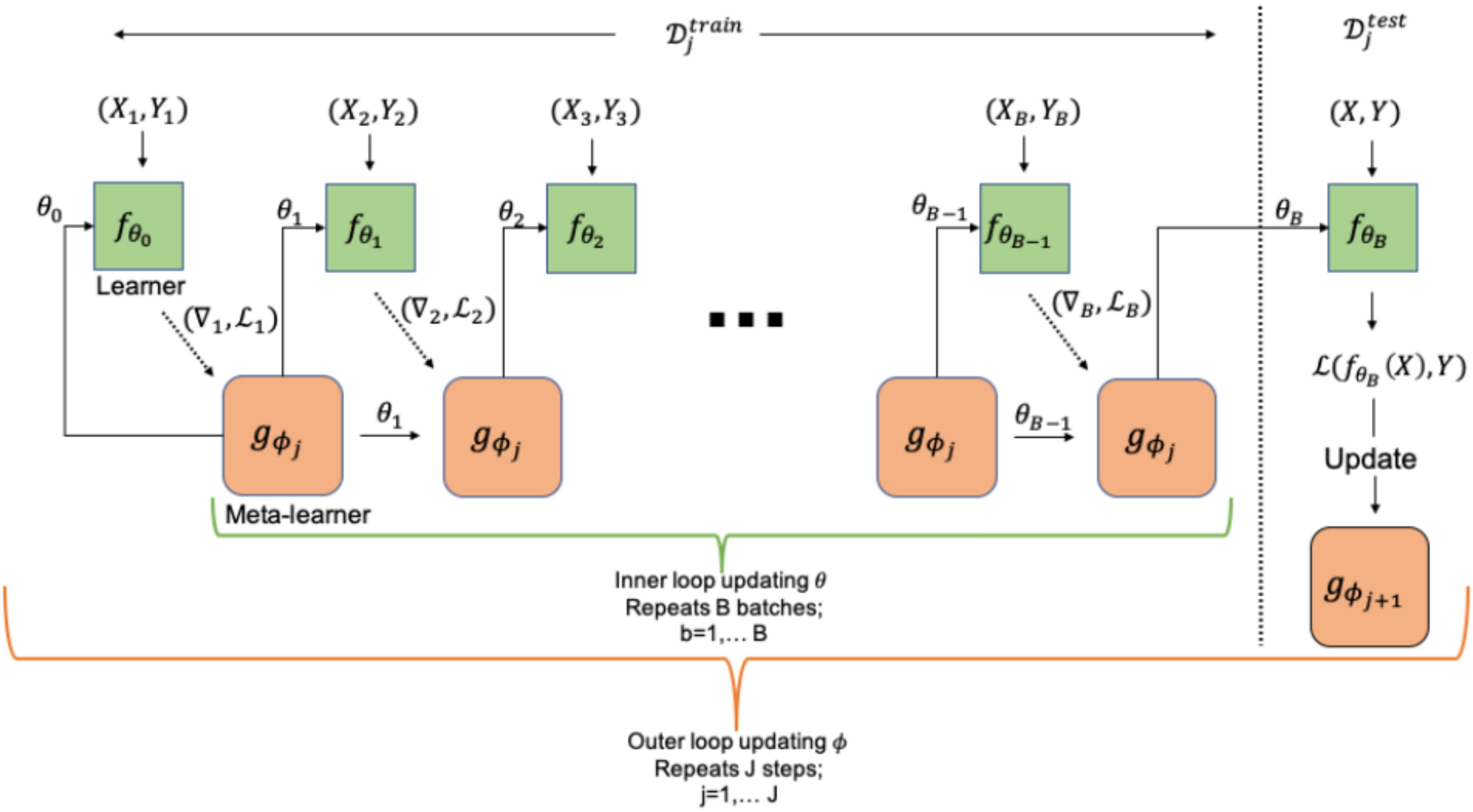
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Optimization-based Meta-Learning Methods

Method	Learner	Meta-Learner
LSTM Meta-Learner [35]	Repeat $\forall b \in [1..B]$ $\mathcal{L}_b \leftarrow \mathcal{L}(f(X_b; \theta_{b-1}), Y_b)$ $\theta_b \leftarrow g((\nabla_{\theta_{b-1}} \mathcal{L}_b, \mathcal{L}_b); \phi_{j-1})$	Repeat $\forall j \in [1..J]$ $\mathcal{L}_j^{test} \leftarrow \mathcal{L}(f(X; \theta_B), Y)$ $\phi_j \leftarrow \phi_{j-1} - \alpha \nabla_{\phi_{j-1}} \mathcal{L}_j^{test}$
MAML [14]	Repeat $\forall i \in [1..I]$ $\mathcal{L}_i^{train} \leftarrow \mathcal{L}(f(\mathcal{D}_i^{train}; \theta_{j-1}))$ $\theta_i^* \leftarrow \theta_{j-1} - \alpha \nabla_{\theta_{j-1}} \mathcal{L}_t^{train}$ $\mathcal{L}_i^{test} \leftarrow \mathcal{L}(f(\mathcal{D}_i^{test}; \theta_t^*))$	Repeat $\forall j \in [1..J]$ $\theta_j \leftarrow \theta_{j-1} - \beta \nabla_{\theta_{j-1}} \sum_{i=1}^{I} \mathcal{L}_i^{test}$
MTL [37]	$\mathcal{L}_i^{train} \leftarrow \mathcal{L}(f(\mathcal{D}_i^{train}; [\theta_{j-1}, \phi_{j-1}, \Theta]))$ $\theta_i^* \leftarrow \theta_{j-1} - \alpha \nabla_{\theta_{j-1}} \mathcal{L}_i^{train}$ $\mathcal{L}_i^{test} \leftarrow \mathcal{L}(f(\mathcal{D}_i^{test}; \theta_i^*))$	$\theta_j \leftarrow \theta_{j-1} - \beta \nabla_{\theta_{j-1}} \sum_{i=1}^{I} \mathcal{L}_i^{test}$ $\phi_j \leftarrow \phi_{j-1} - \beta \nabla_{\phi_{j-1}} \sum_{i=1}^{I} \mathcal{L}_i^{test}$
LEO [38]	$\phi_{j-1} = \{\phi_e, \phi_r, \phi_d, \alpha\}$ $\mathbf{z_i} \leftarrow g(\mathcal{D}_i^{train}; [\phi_e, \phi_r, \Theta])$ $\theta_i \leftarrow g(\mathbf{z_i}; \phi_d)$ $\mathcal{L}_i^{train} \leftarrow \mathcal{L}(f(\mathcal{D}_i^{train}; \theta_i))$ $\mathbf{z_i^*} \leftarrow \mathbf{z_i} - \alpha \nabla_{\mathbf{z_i}} \mathcal{L}_i^{train}$ $\theta_i^* \leftarrow g(\mathbf{z_i^*}; \phi_d)$ $\mathcal{L}_i^{test} \leftarrow \mathcal{L}(f(\mathcal{D}_i^{test}; \theta_i^*))$	$\phi_j \leftarrow \phi_{j-1} - \beta \nabla_{\phi_{j-1}} \sum_{i=1}^{I} \mathcal{L}_i^{test}$

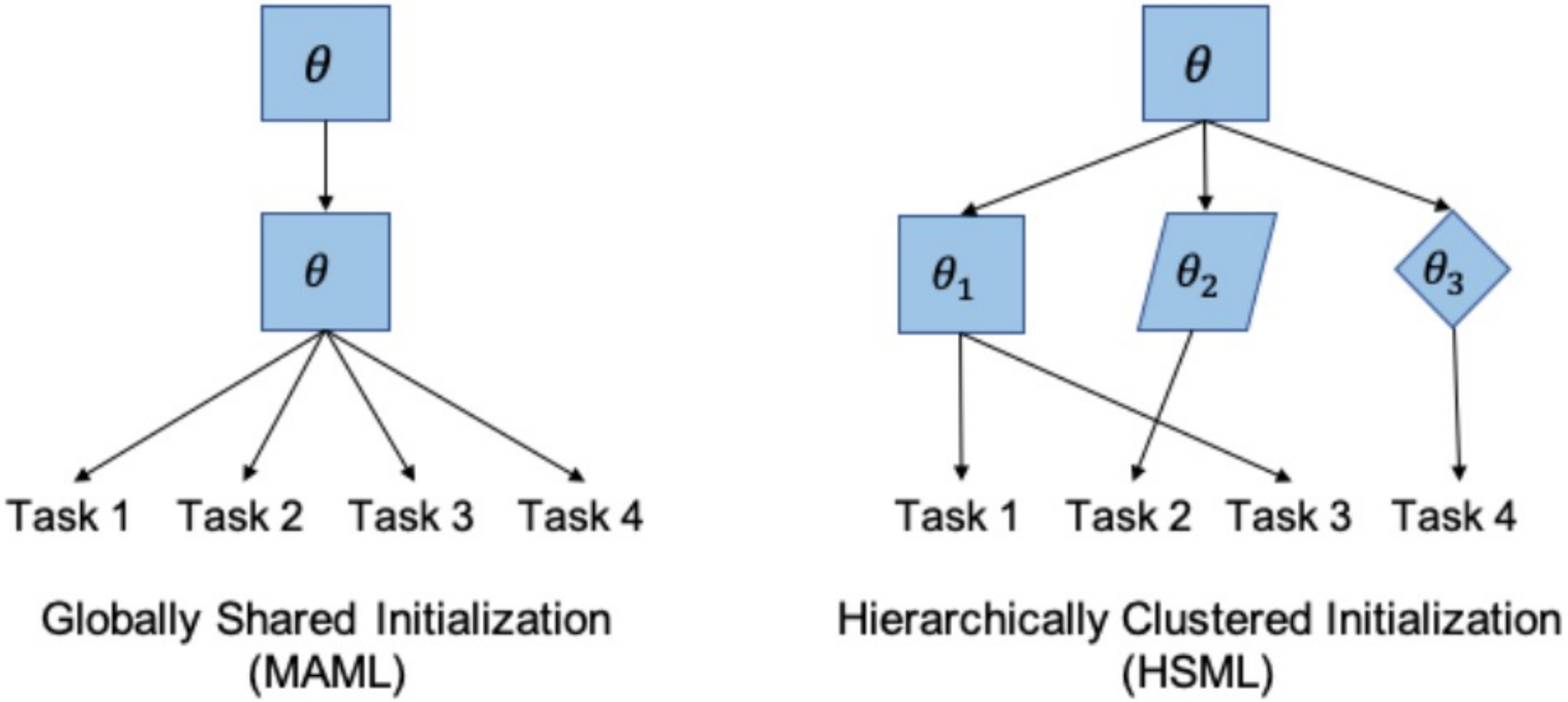
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Computational Graph for the Forward Pass of the Meta-learner



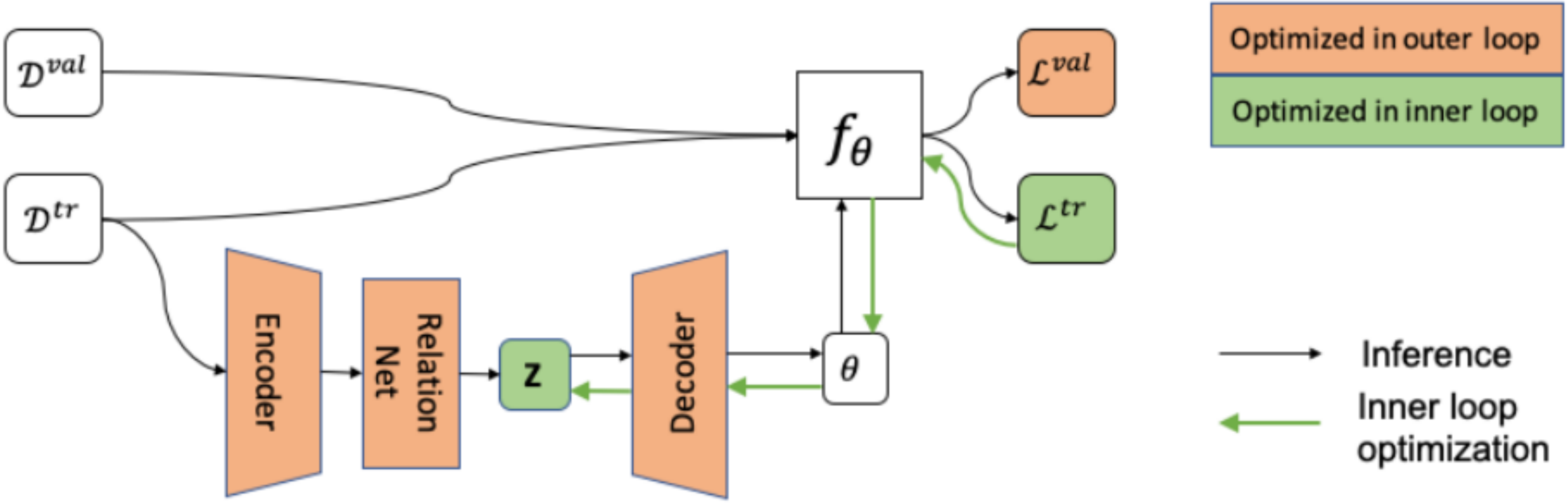
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Model-Agnostic Meta-Learning (MAML)
Hierarchically Structured Meta-Learning (HSML)



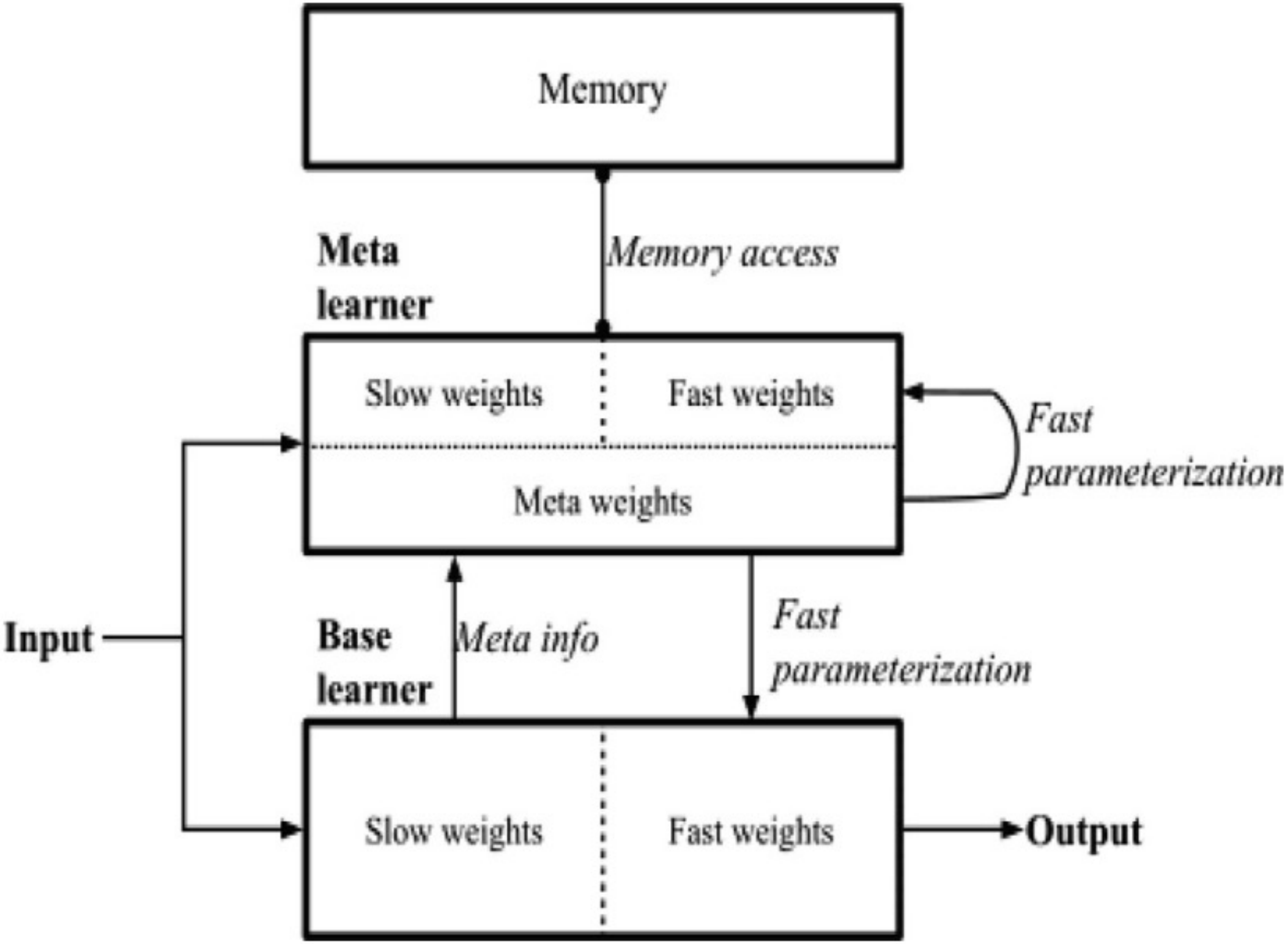
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Latent Embedding Optimization (LEO)



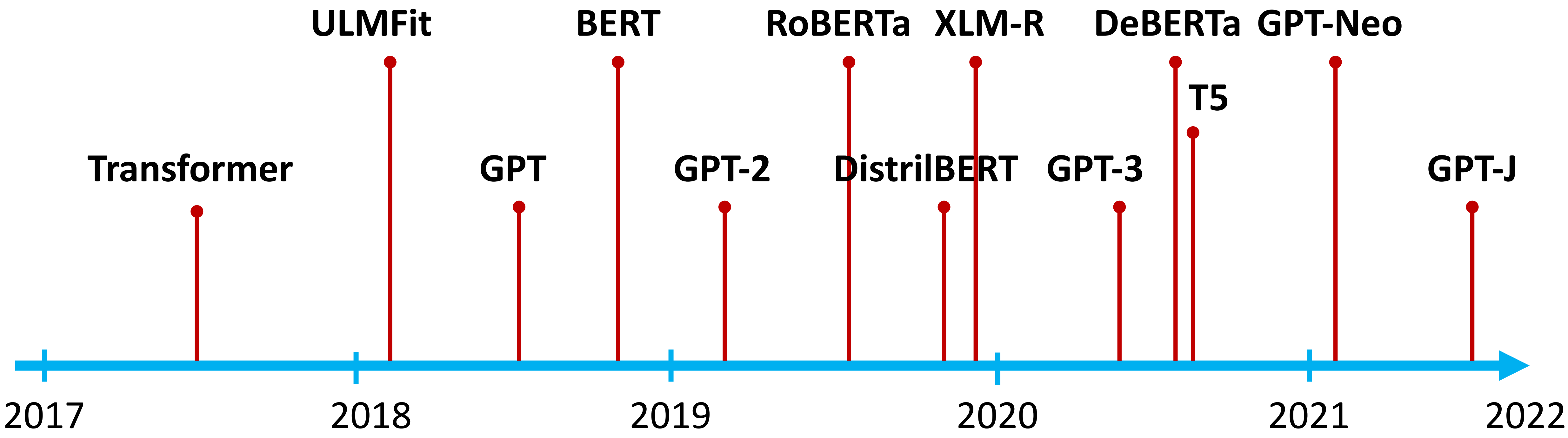
Source: Parnami, Archit, and Minwoo Lee. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." arXiv preprint arXiv:2203.04291 (2022).

Overall Architecture of Meta Networks



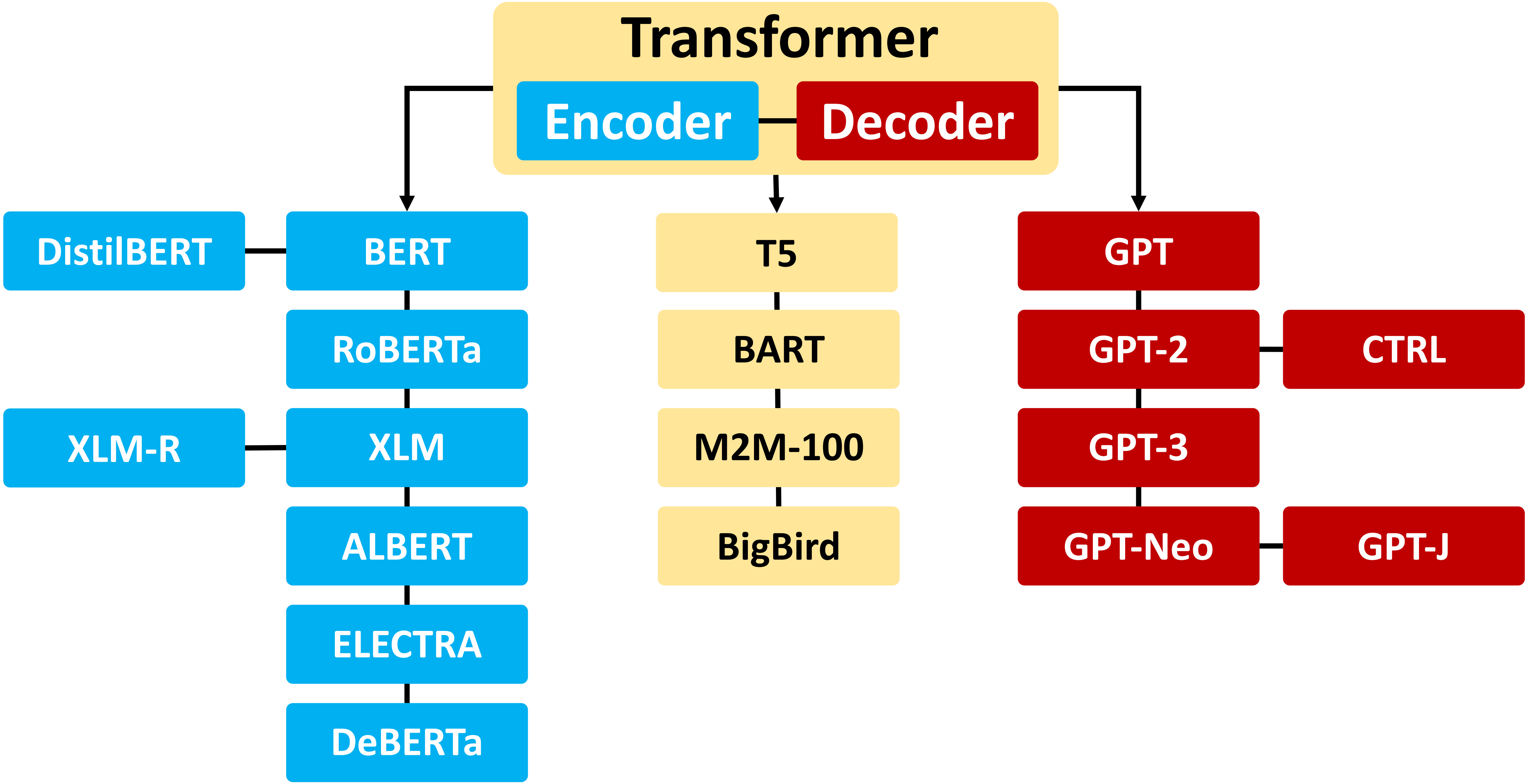
Source: Tian, Yingjie, Xiaoxi Zhao, and Wei Huang. "Meta-learning approaches for learning-to-learn in deep learning: A survey." Neurocomputing 494 (2022): 203-223.

The Transformers Timeline



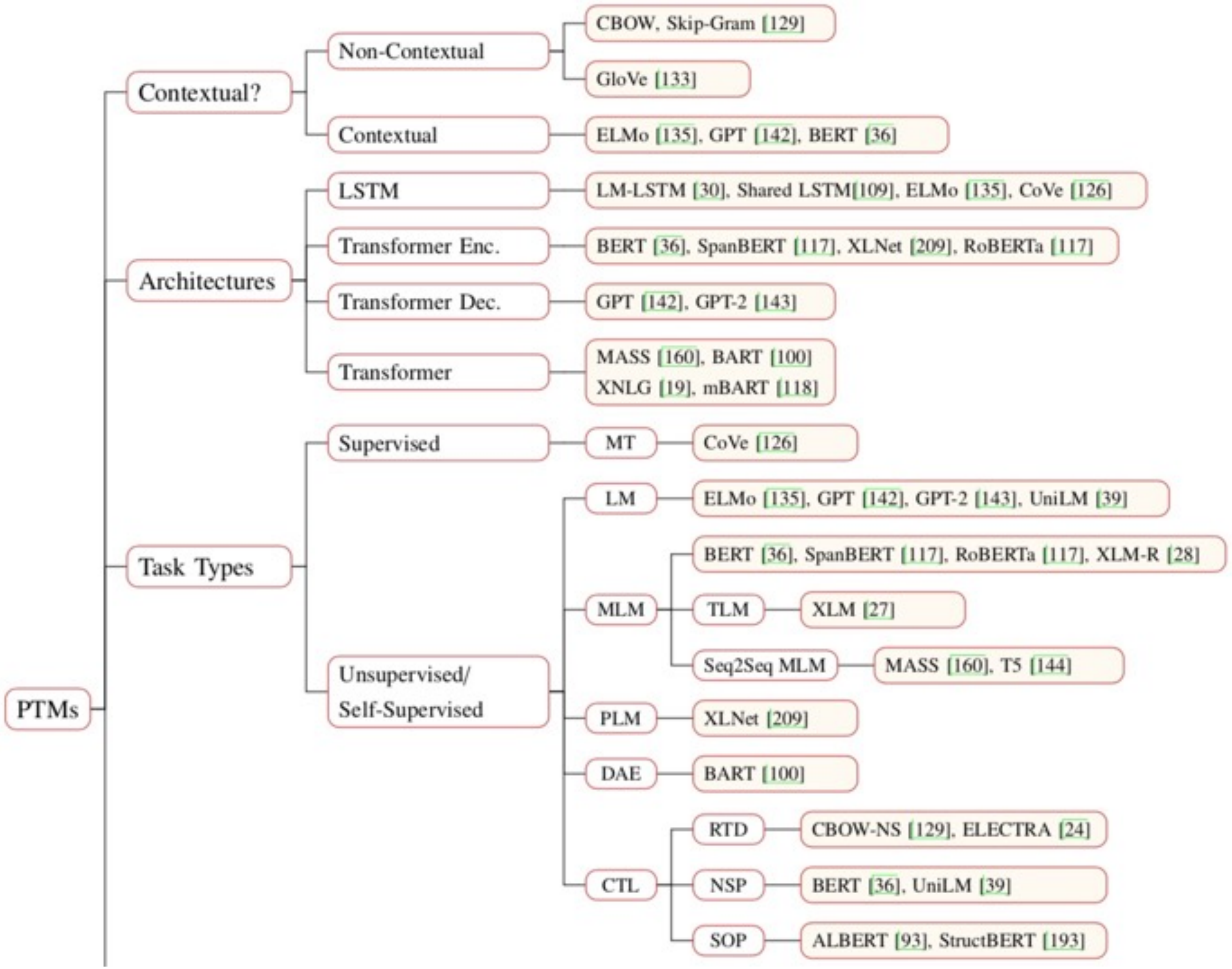
Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.

Transformer Models



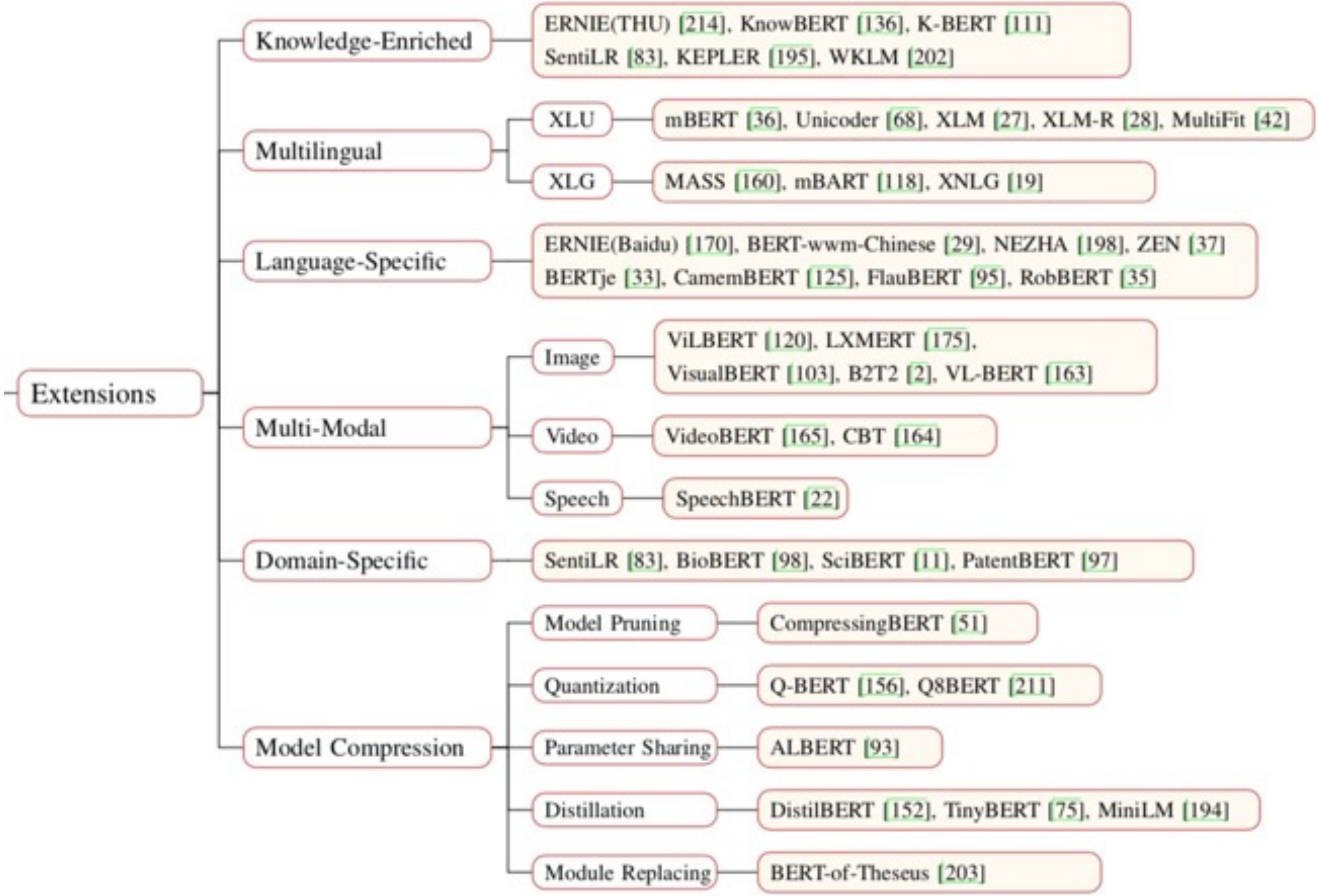
Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.

Pre-trained Models (PTM)



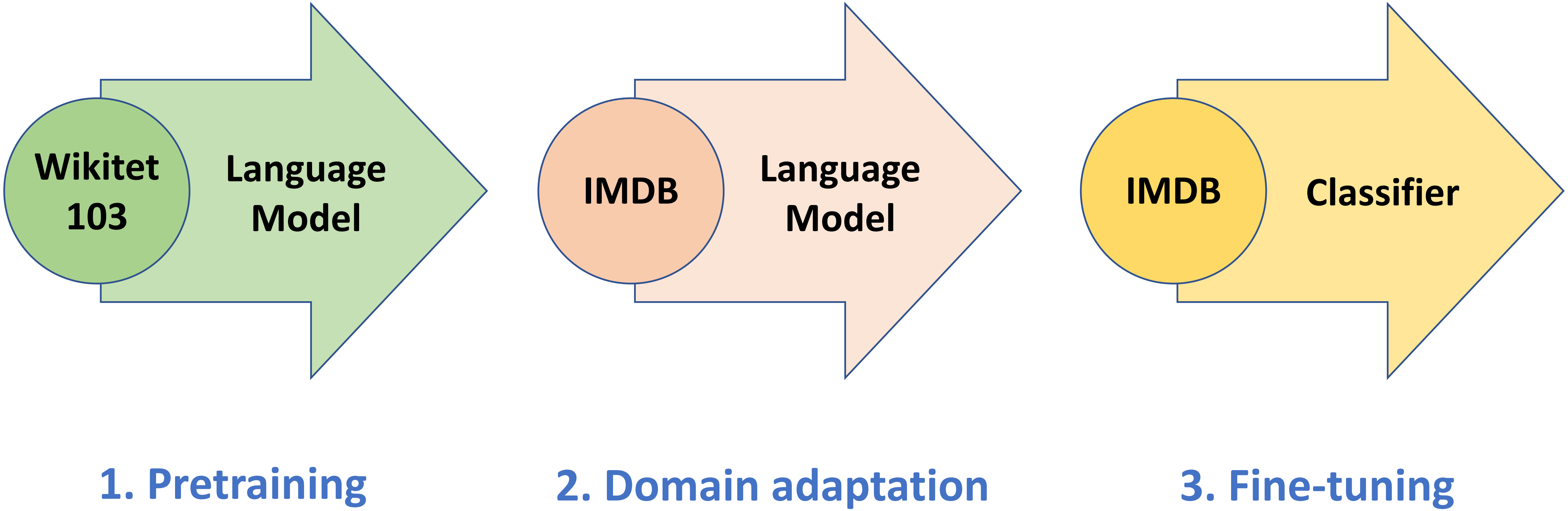
Source: Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained Models for Natural Language Processing: A Survey." arXiv preprint arXiv:2003.08271 (2020).

Pre-trained Models (PTM)



Source: Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained Models for Natural Language Processing: A Survey." arXiv preprint arXiv:2003.08271 (2020).

ULMFiT: 3 Steps
Transfer Learning in NLP



Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.

A typical pipeline for training transformer models

with the Datasets, Tokenizers, and Transformers libraries

Datasets	Tokenizers	Transformers	Datasets
Load and process datasets	Tokenize input texts	Load models, train and infer	Load metrics evaluate models

Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.
https://github.com/nlp-with-transformers/notebooks

Few-Shot Learning (FSL)
Typical Scenarios

- **Acting as a test bed for learning like human**
- **Learning for rare cases**
- **Reducing data gathering effort and computational cost**

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

- **Few-Shot Learning (FSL) is a sub-area in machine learning.**
- **Machine Learning Definition**
  - **A computer program is said to learn from experience E with respect to some classes of task T and performance measure P if its performance can improve with E on T measured by P.**
    - Example: Image classification task (T), a machine learning program can improve its classification accuracy (P) through E obtained by training on a large number of labeled images (e.g., the ImageNet data set).

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Machine Learning

task T	experience E	performance P
image classification [73]	large-scale labeled images for each class	classification accuracy
the ancient game of Go [120]	a database containing around 30 million recorded moves of human experts and self-play records	winning rate

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

- **Few-shot Learning (FSL) is a type of machine learning problems (specified by E, T, and P), where E contains only a limited number of examples with supervised information for the target T.**
  - **Existing FSL problems are mainly supervised learning problems.**
  - **Few-shot classification learns classifiers given only a few labeled examples of each class.**
    - image classification
    - sentiment classification from short text
    - object recognition

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

- **Few-shot classification learns a classifier h, which predicts label y_i for each input x_i.**
- **Usually, one considers the *N-way-K-shot* classification, in which D_{train} contains $I = KN$ examples from N classes each with K examples**

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

- **Few-Shot Learning (FSL)**
  - $K = 10 \sim 100$ **examples**
- **One-Shot Learning (1SL)**
  - $K = 1$ **example**
- **Zero-Shot Learning (0SL)(ZSL)**
  - $K = 0$

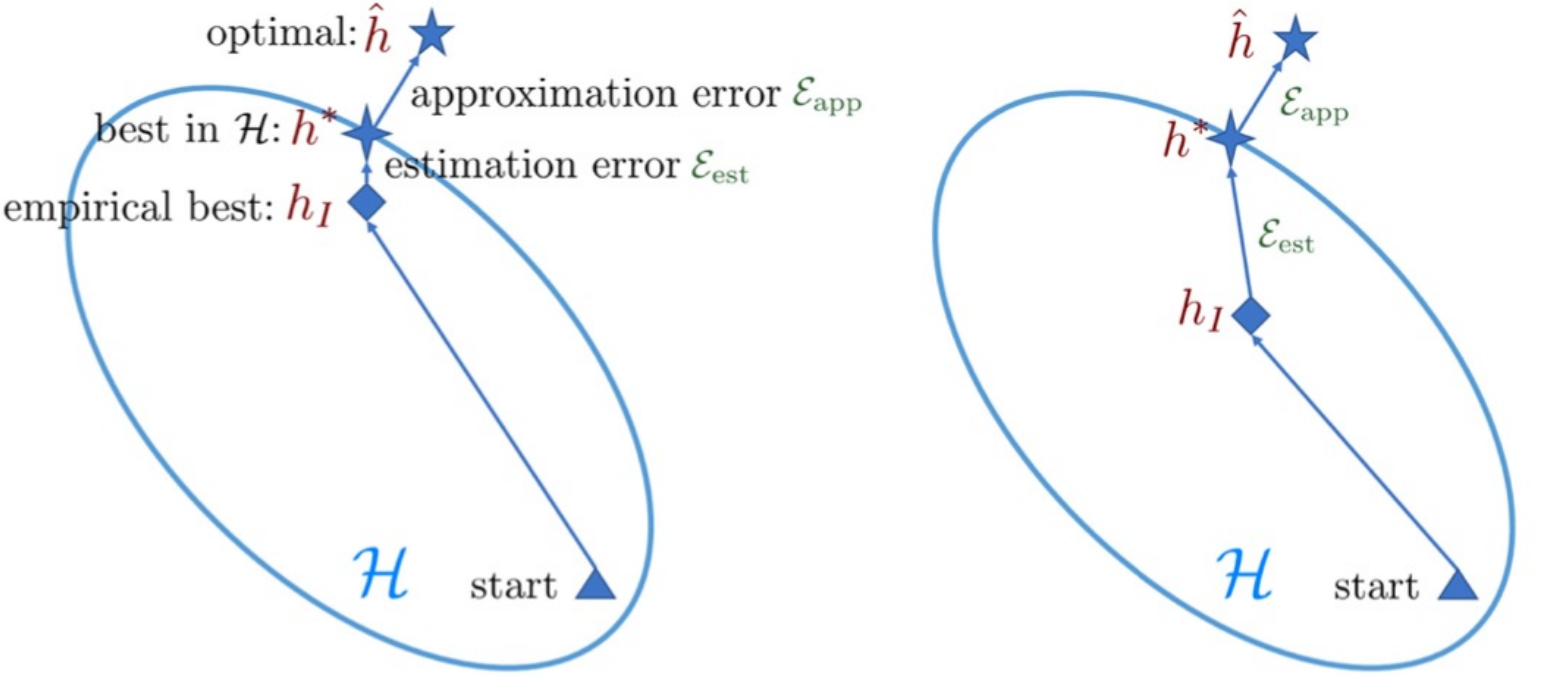
Few-Shot Learning (FSL)

task T	experience E		performance P
	supervised information	prior knowledge	
character generation [76]	a few examples of new character	pre-learned knowledge of parts and relations	pass rate of visual Turing test
drug toxicity discovery [4]	new molecule's limited assay	similar molecules' assays	classification accuracy
image classification [70]	a few labeled images for each class of the target T	raw images of other classes, or pre-trained models	classification accuracy

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Comparison of learning with sufficient and few training samples

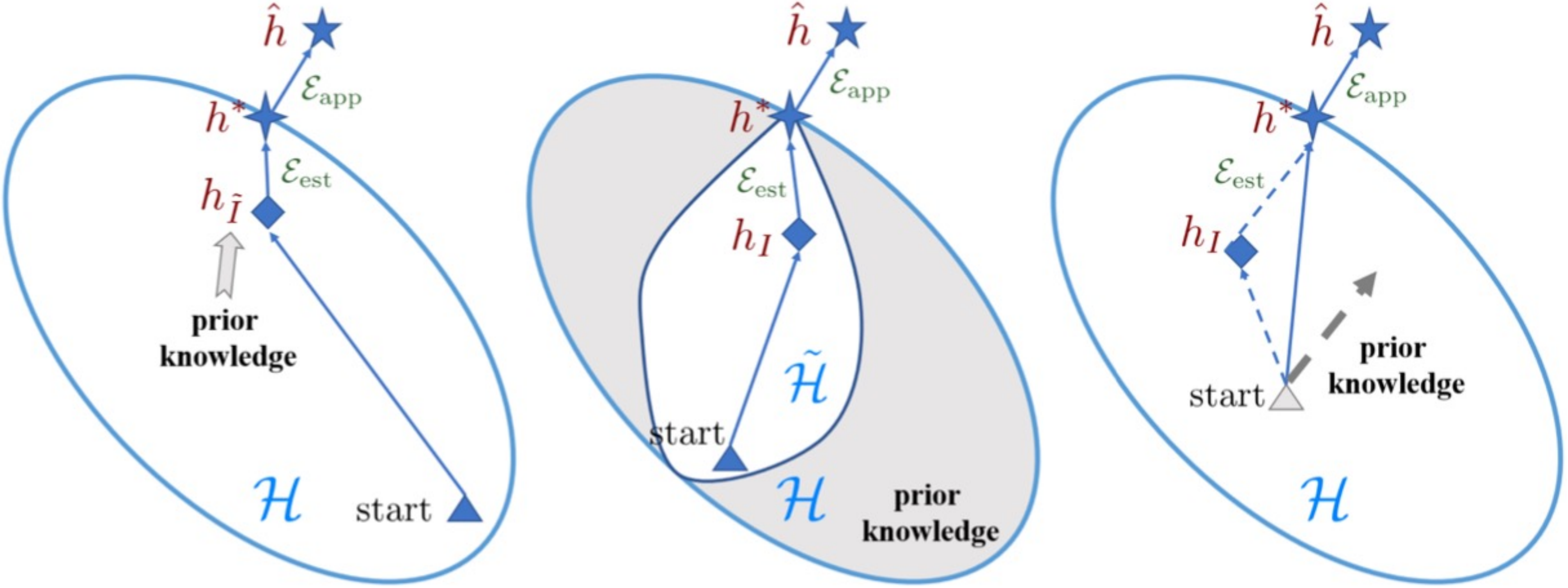


(a) Large I (b) Small I

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Different perspectives on how FSL methods solve the few-shot problem

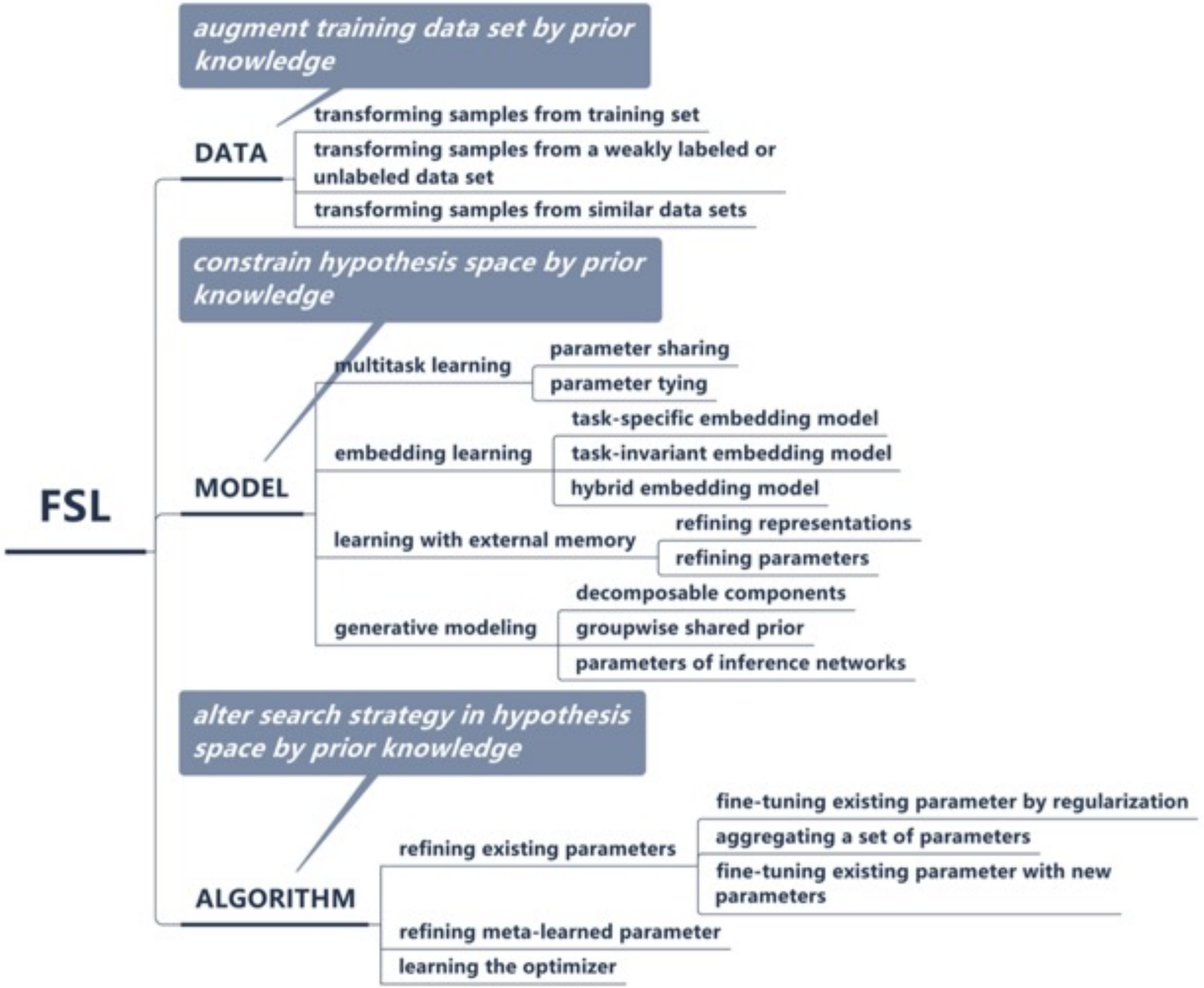


(a) Data. (b) Model. (c) Algorithm.

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

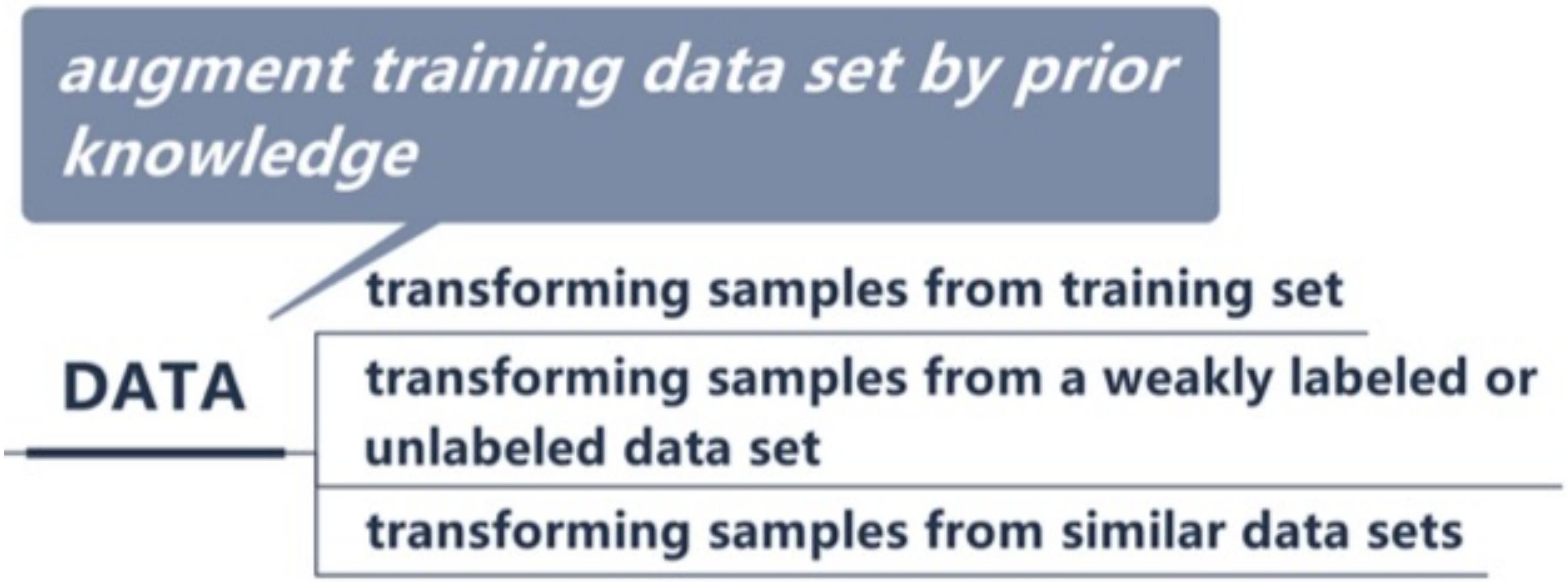
Few-Shot Learning (FSL)

A taxonomy of FSL methods



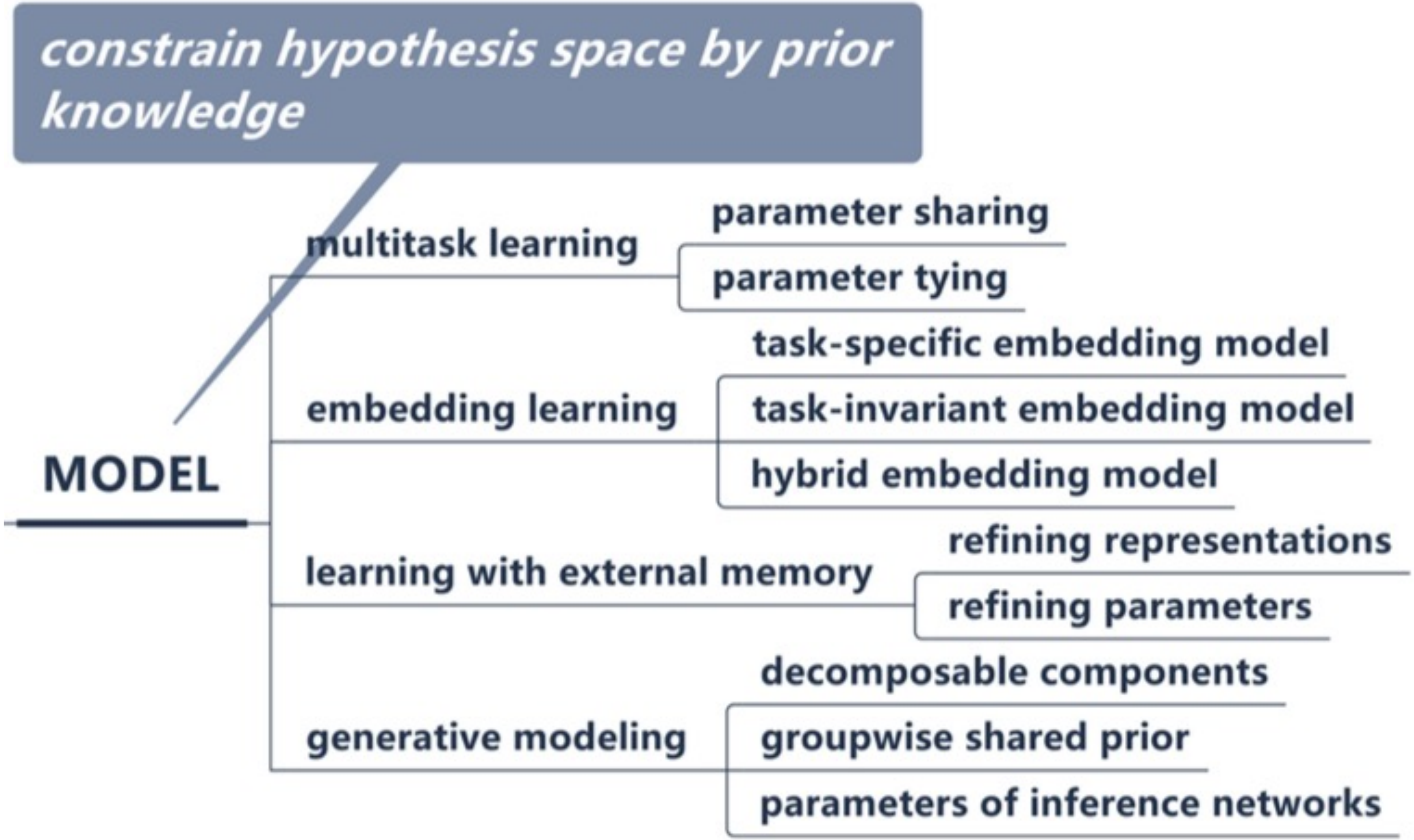
Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

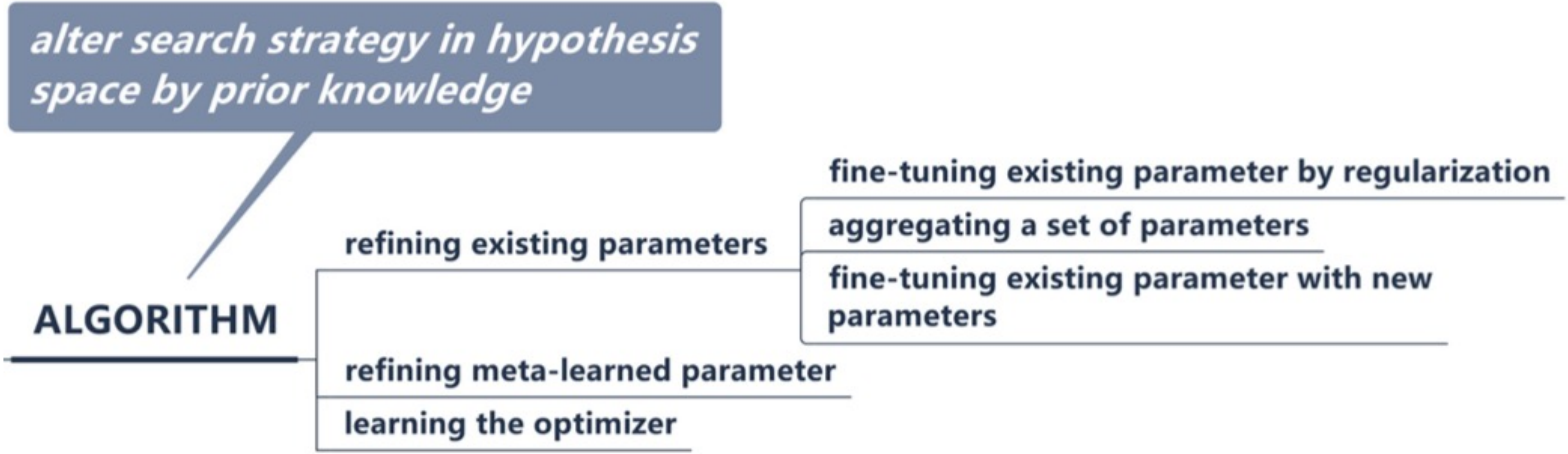


Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

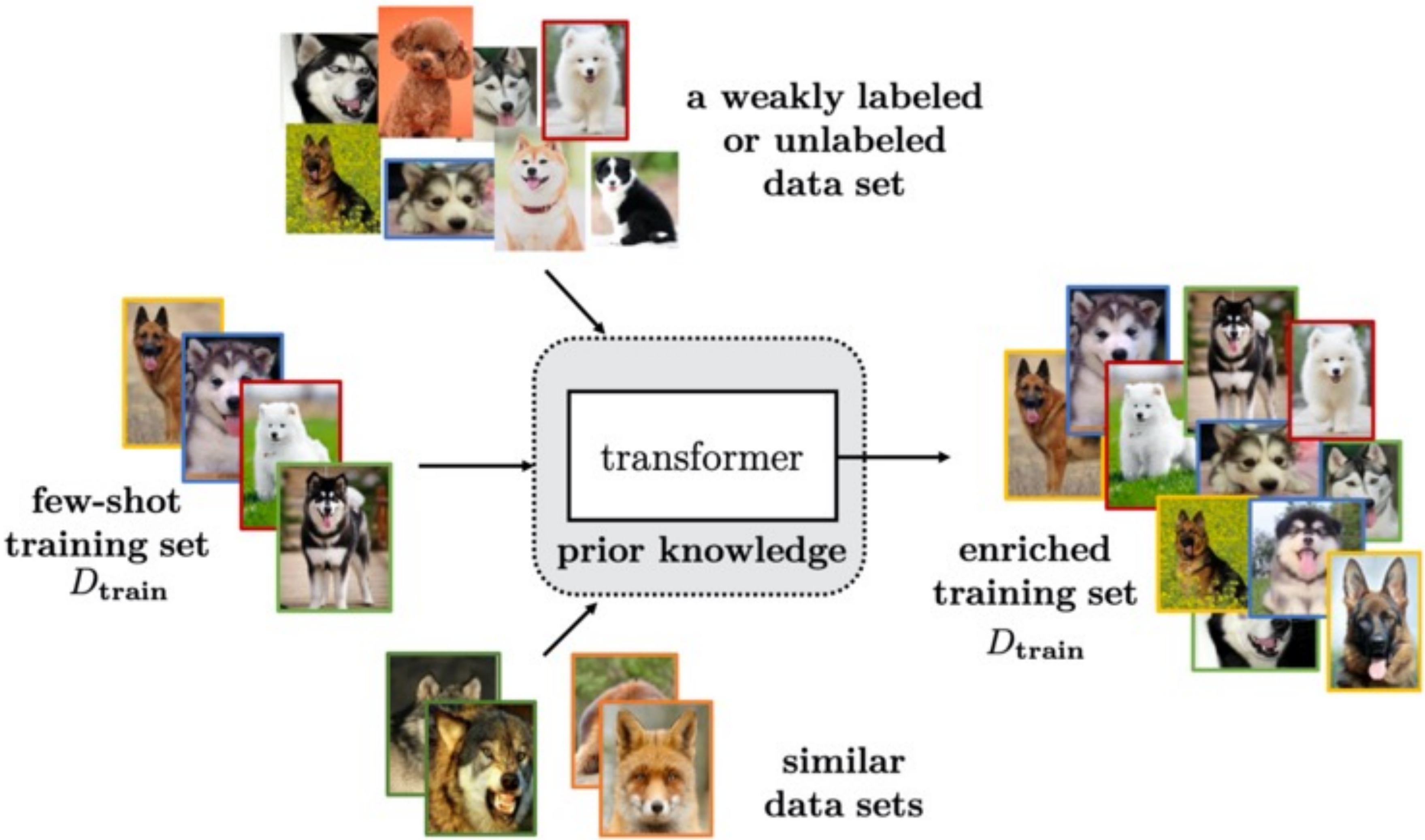


Few-Shot Learning (FSL)



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)
Solving the FSL problem by data augmentation



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Characteristics for FSL Methods Focusing on the Data Perspective

category	input (x, y)	transformer t	output $(\tilde{x}, \tilde{y})$
transforming samples from D_{train}	original (x_i, y_i)	learned transformation function on x_i	$(t(x_i), y_i)$
transforming samples from a weakly labeled or unlabeled data set	weakly labeled or unlabeled $(\bar{x}, -)$	a predictor trained from D_{train}	$(\bar{x}, t(\bar{x}))$
transforming samples from similar data sets	samples $\{(\hat{x}_j, \hat{y}_j)\}$ from similar data sets	an aggregator to combine $\{(\hat{x}_j, \hat{y}_j)\}$	$(t(\{\hat{x}_j\}), t(\{\hat{y}_j\}))$

The transformer $t(\cdot)$ takes input (x, y) and returns synthesized sample $(\tilde{x}, \tilde{y})$ to augment the few-shot D_{train}.

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

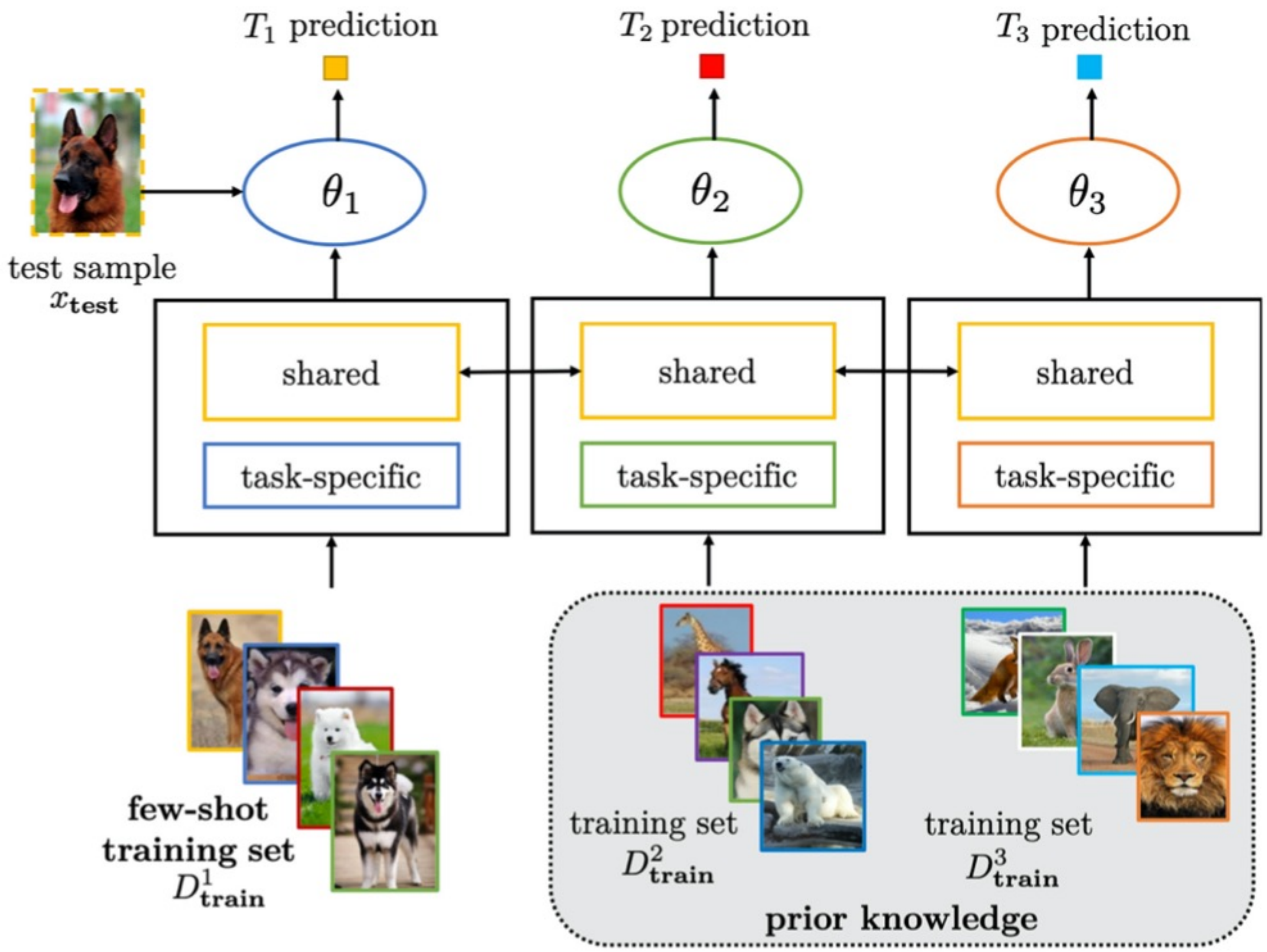
Characteristics for FSL Methods Focusing on the Model Perspective

strategy	prior knowledge	how to constrain $\mathcal{H}$
multitask learning	other T's with their data sets D's	share/tie parameter
embedding learning	embedding learned from/together with other T's	project samples to a smaller embedding space in which similar and dissimilar samples can be easily discriminated
learning with external memory	embedding learned from other T's to interact with memory	refine samples using key-value pairs stored in memory
generative modeling	prior model learned from other T's	restrict the form of distribution

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

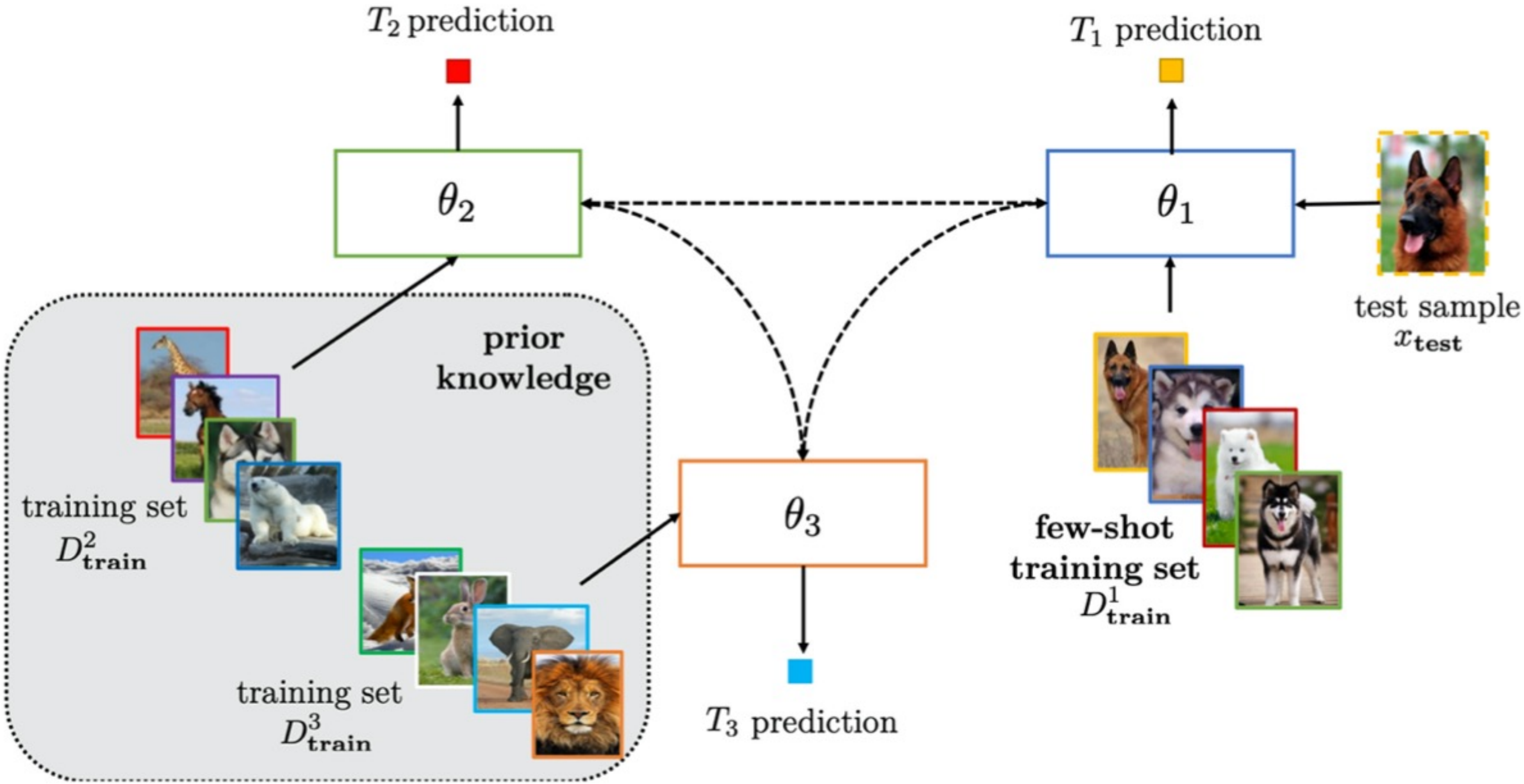
Solving the FSL problem by multitask learning with parameter sharing



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Solving the FSL problem by multitask learning with parameter tying



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

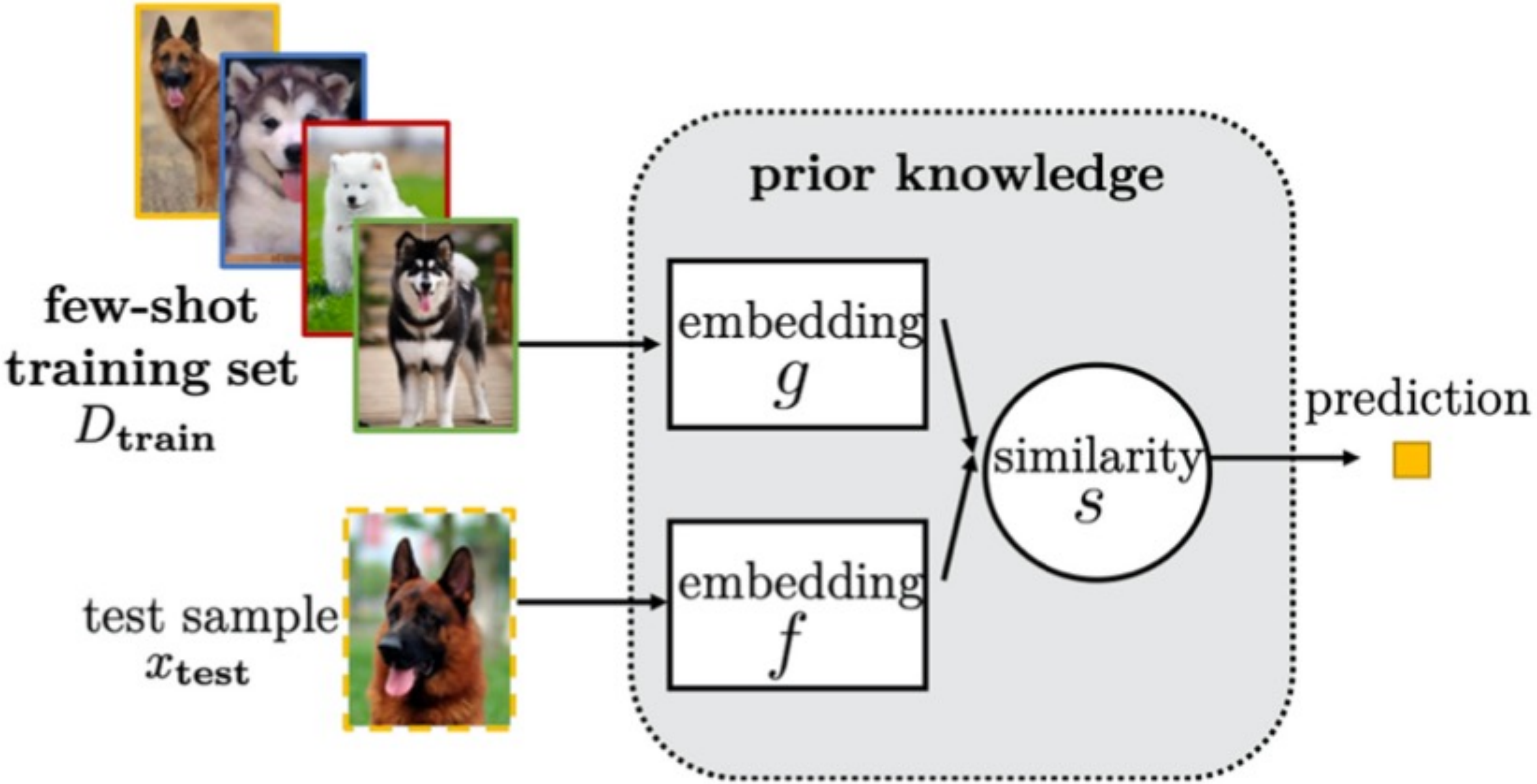
Characteristics of Embedding Learning Methods

category	method	embedding function f for x_{test}	embedding function g for D_{train}	similarity measure s
task-specific	mAP-DLM/SSVM[130]	CNN	the same as f	cosine similarity
task-invariant	class relevance pseudo-metric [36]	kernel	the same as f	squared ℓ_2 distance
	convolutional siamese net [70]	CNN	the same as f	weighted ℓ_1 distance
	Micro-Set[127]	logistic projection	the same as f	ℓ_2 distance
	Matching Nets [138]	CNN, LSTM	CNN, biLSTM	cosine similarity
	resLSTM [4]	GNN, LSTM	GNN, LSTM	cosine similarity
	Active MN [8]	CNN	biLSTM	cosine similarity
	SSMN [24]	CNN	another CNN	learned distance
	ProtoNet [121]	CNN	the same as f	squared ℓ_2 distance
	semi-supervised ProtoNet[108]	CNN	the same as f	squared ℓ_2 distance
	PMN [141]	CNN, LSTM	CNN, biLSTM	cosine similarity
	ARC [119]	LSTM, biLSTM	the same as f	-
	Relation Net [126]	CNN	the same as f	-
	GNN [115]	CNN, GNN	the same as f	learned distance
	TPN [84]	CNN	the same as f	Gaussian similarity
	SNAIL [91]	CNN	the same as f	-
hybrid	Learnet [14]	adaptive CNN	CNN	weighted ℓ_1 distance
	DCCN [162]	adaptive CNN	CNN	-
	R2-D2 [13]	adaptive CNN	CNN	-
	TADAM [100]	adaptive CNN	the same as f	squared ℓ_2 distance

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

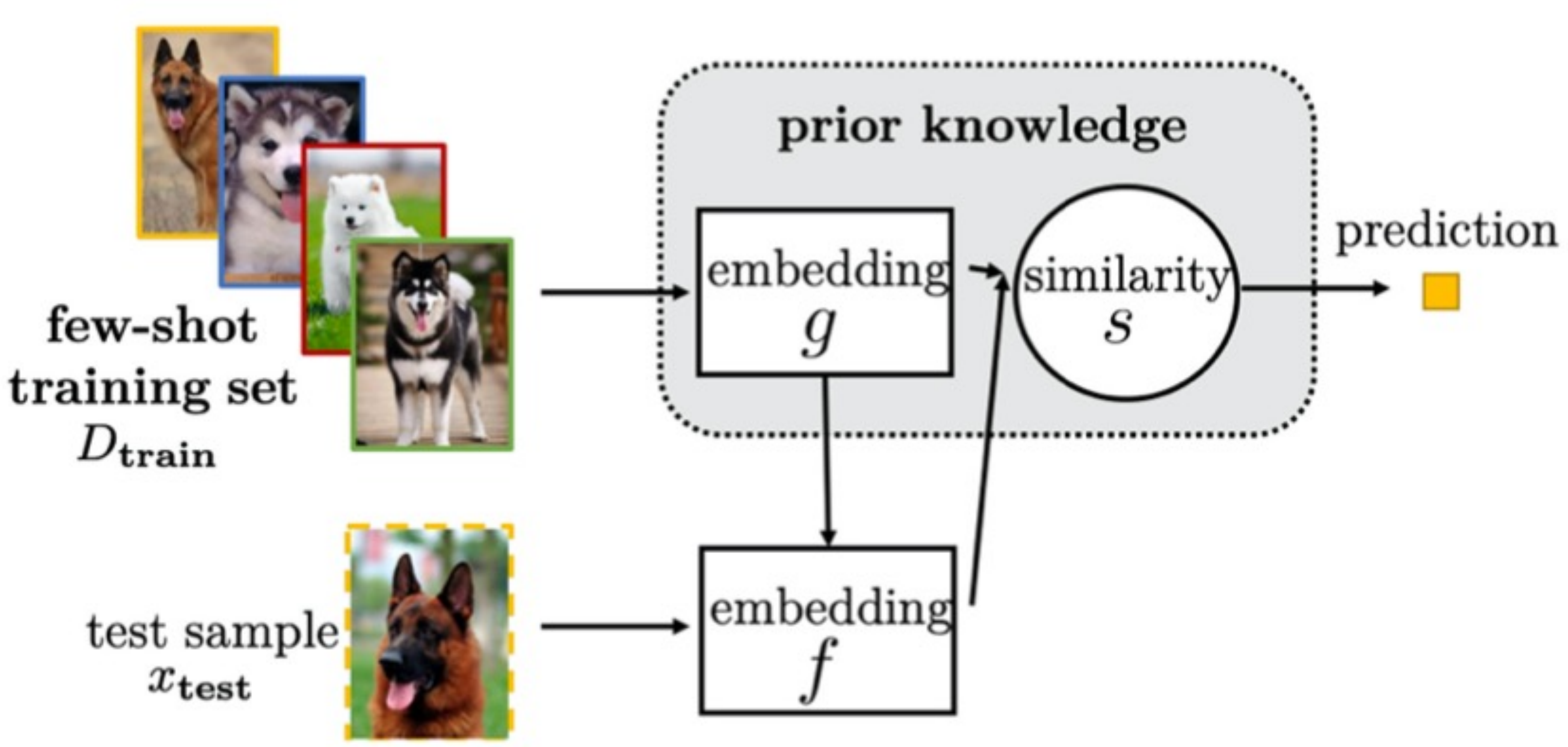
Solving the FSL problem by task-invariant embedding model



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

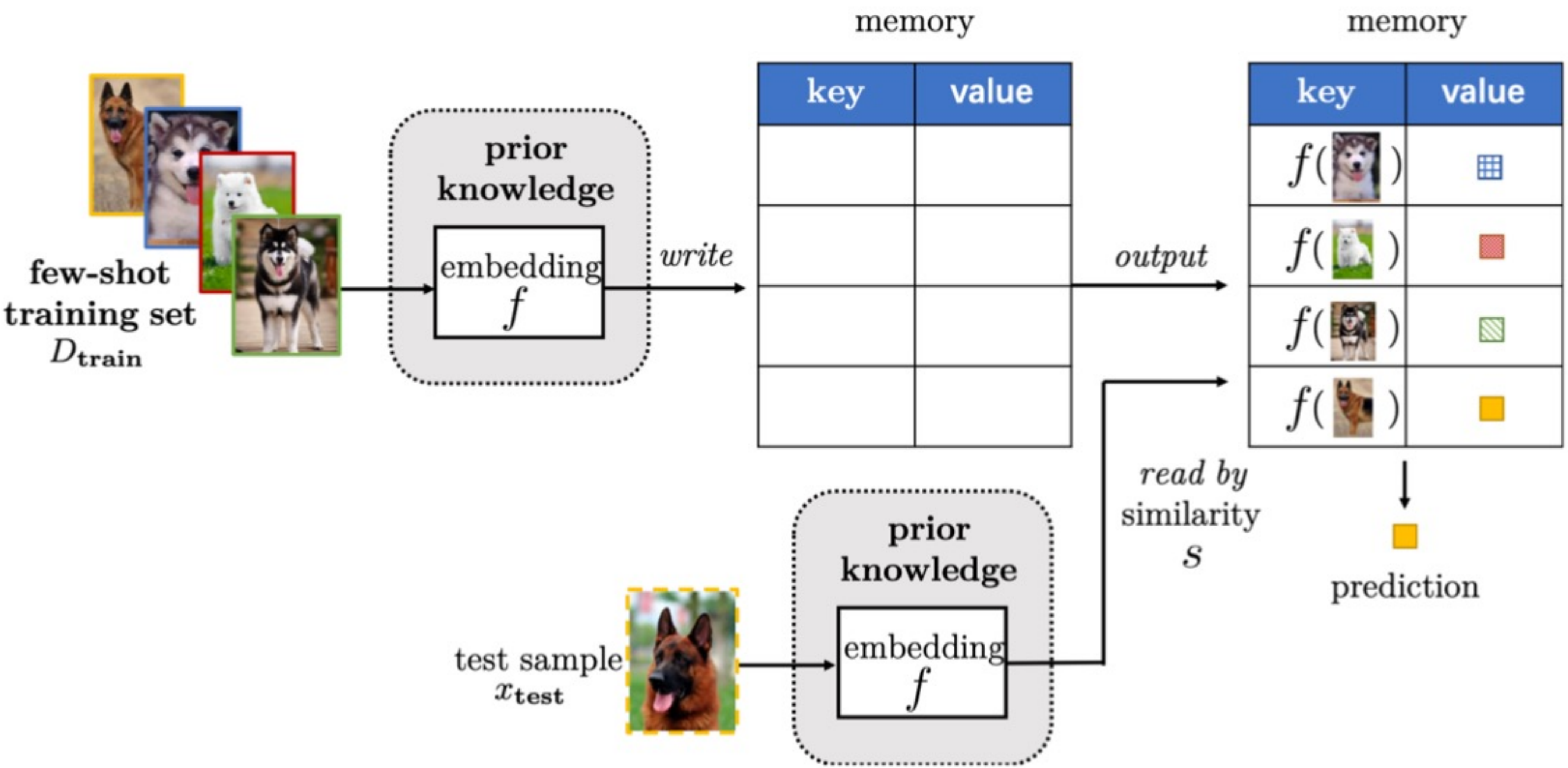
Solving the FSL problem by hybrid embedding model



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Solving the FSL problem by learning with external memory



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Characteristics of FSL Methods
Based on Learning with External Memory

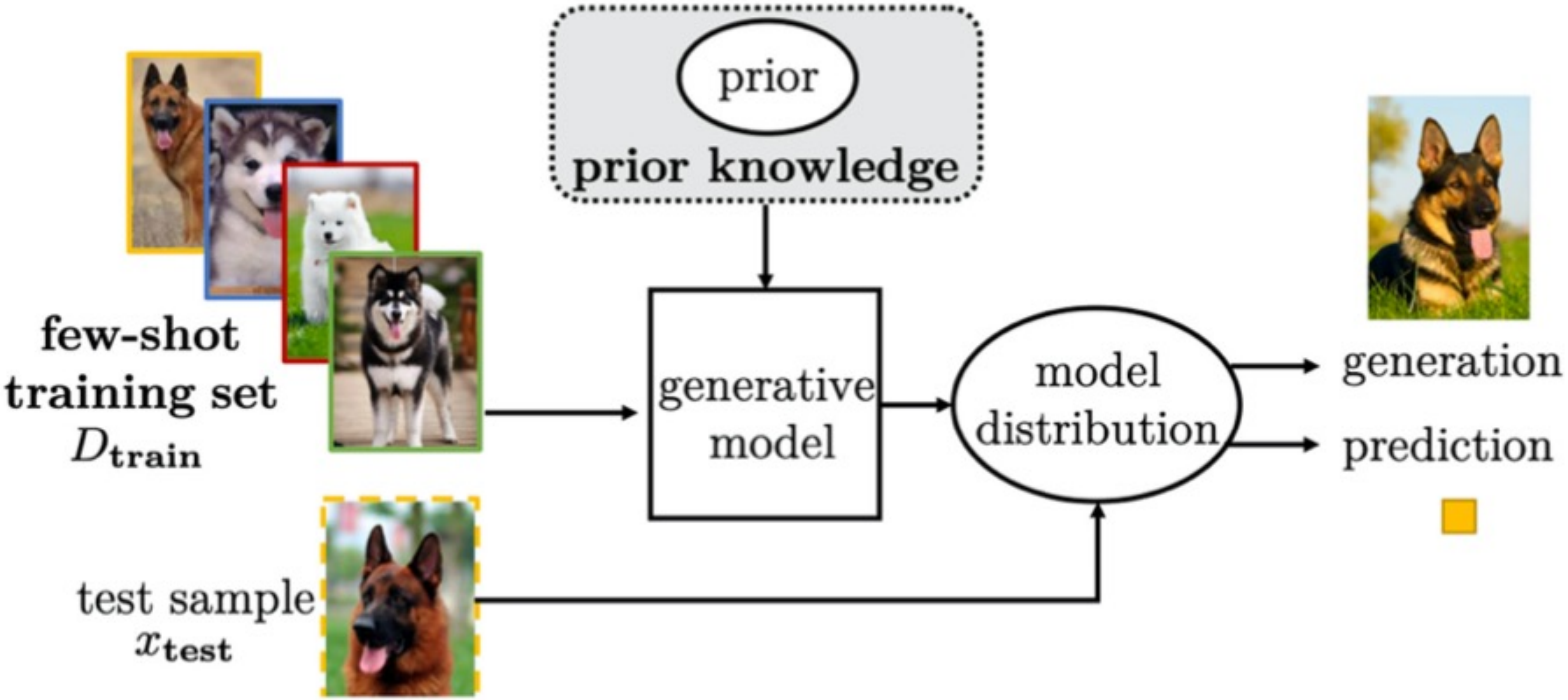
category	method	memory M		similarity s
		key M_{key}	value M_{value}	
refining representations	MANN [114]	$f(x_i, y_{i-1})$	$f(x_i, y_{i-1})$	cosine similarity
	APL [104]	$f(x_i)$	y_i	squared ℓ_2 distance
	abstraction memory [149]	$f(x_i)$	word embedding of y_i	dot product
	CMN [164]	$f(x_i)$	y_i, age	dot product
	life-long memory [65]	$f(x_i)$	y_i, age	cosine similarity
	Mem2Vec [125]	$f(x_i)$	word embedding of y_i, age	dot product
refining parameters	MetaNet [96]	$f(x_i)$	fast weight	cosine similarity
	CSNs [97]	$f(x_i)$	fast weight	cosine similarity
	MN-Net [22]	$f(x_i)$	y_i	dot product

Here, f is an embedding function usually pre-trained by CNN or LSTM.

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

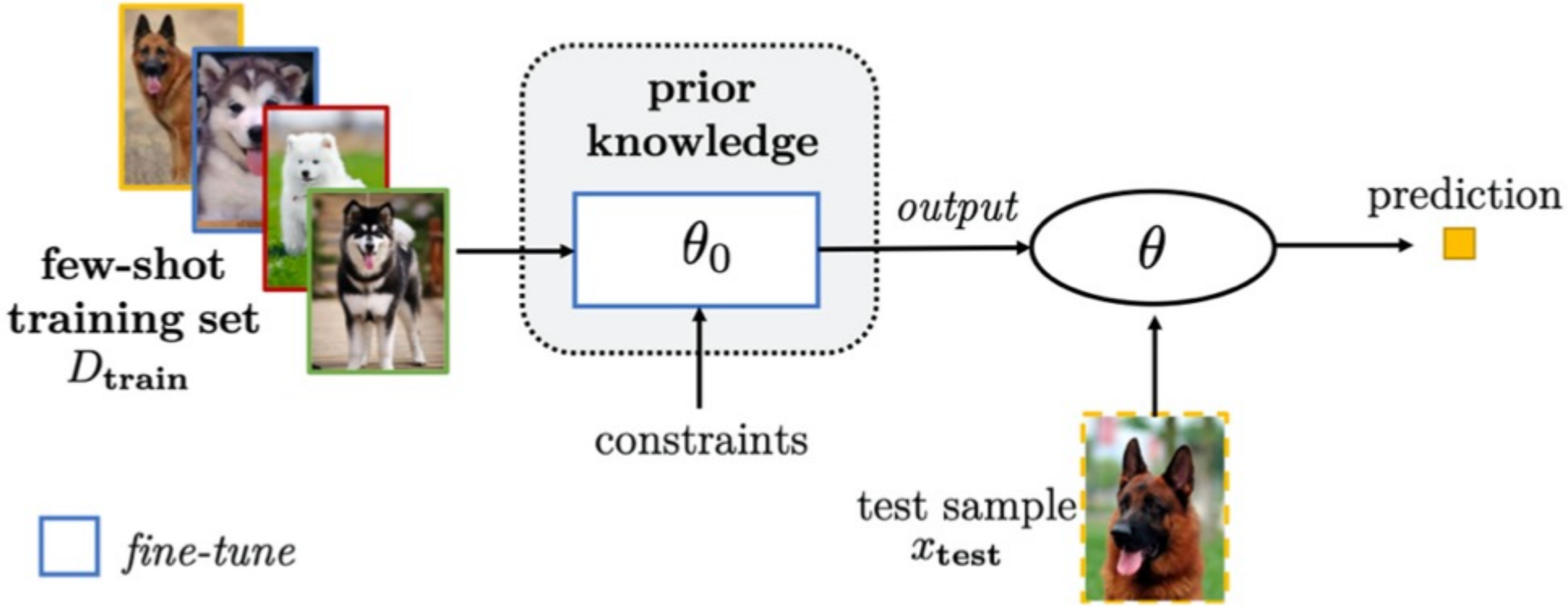
Solving the FSL problem by generative modeling



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Solving the FSL problem by fine-tuning existing parameter θ_0 by regularization



Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Characteristics for FSL Methods

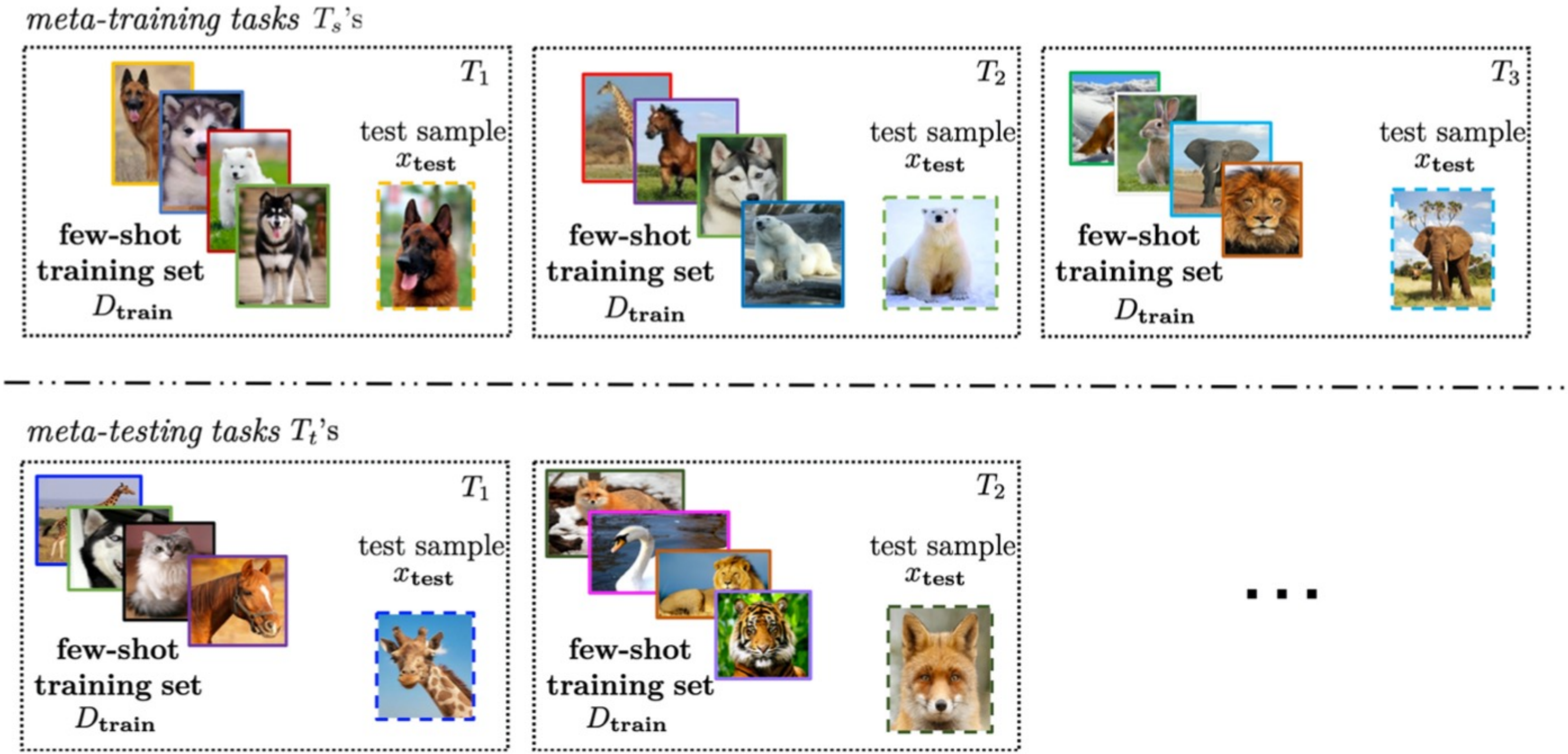
Focusing on the Algorithm Perspective

strategy	prior knowledge	how to search θ of the h^* in $\mathcal{H}$
refining existing parameters	learned θ_0	refine θ_0 by D_{train}
refining meta-learned parameters	meta-learner	refine θ_0 by D_{train}
learning the optimizer	meta-learner	use search steps provided by the meta-learner

Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

Solving the FSL problem by meta-learning

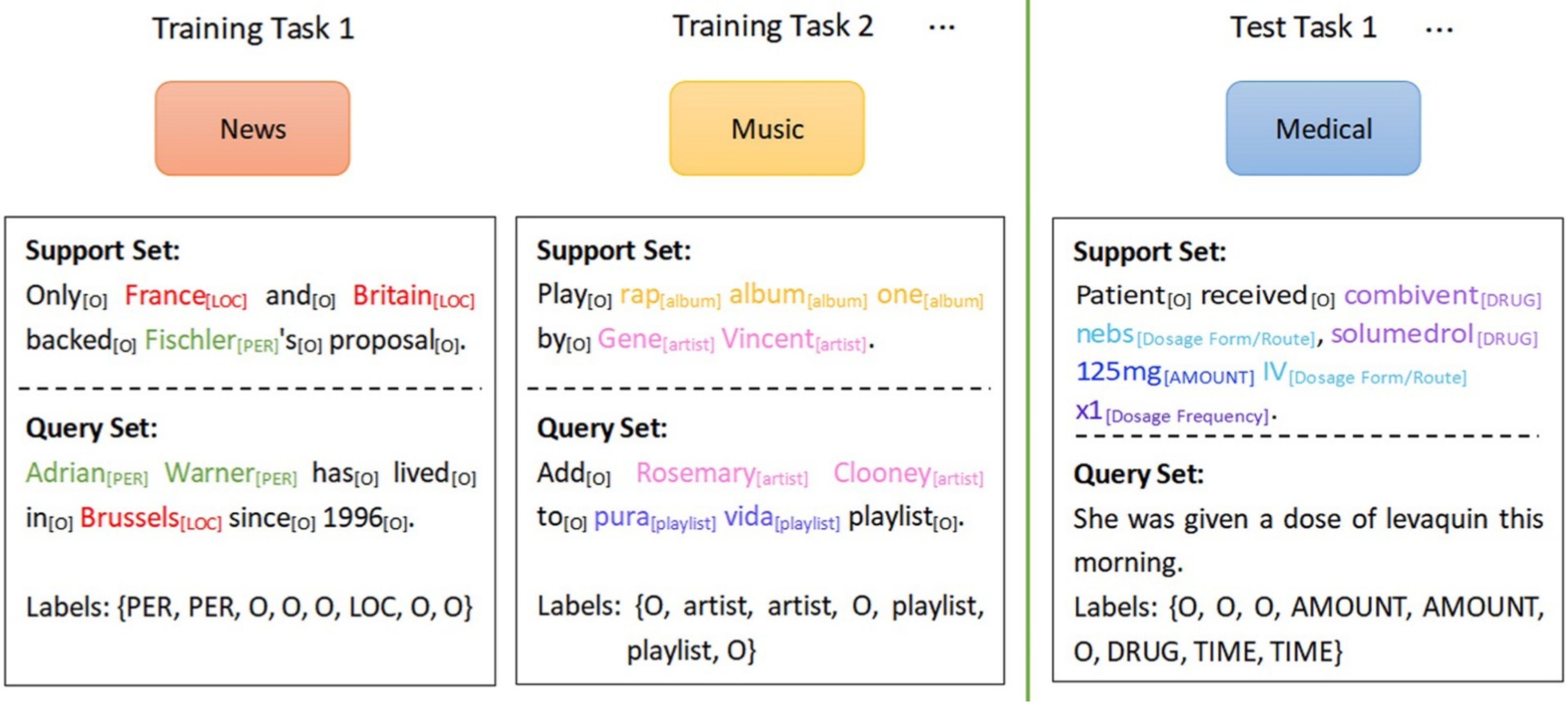


Source: Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a few examples: A survey on few-shot learning." ACM computing surveys (csur) 53, no. 3 (2020): 1-34.

Few-Shot Learning (FSL)

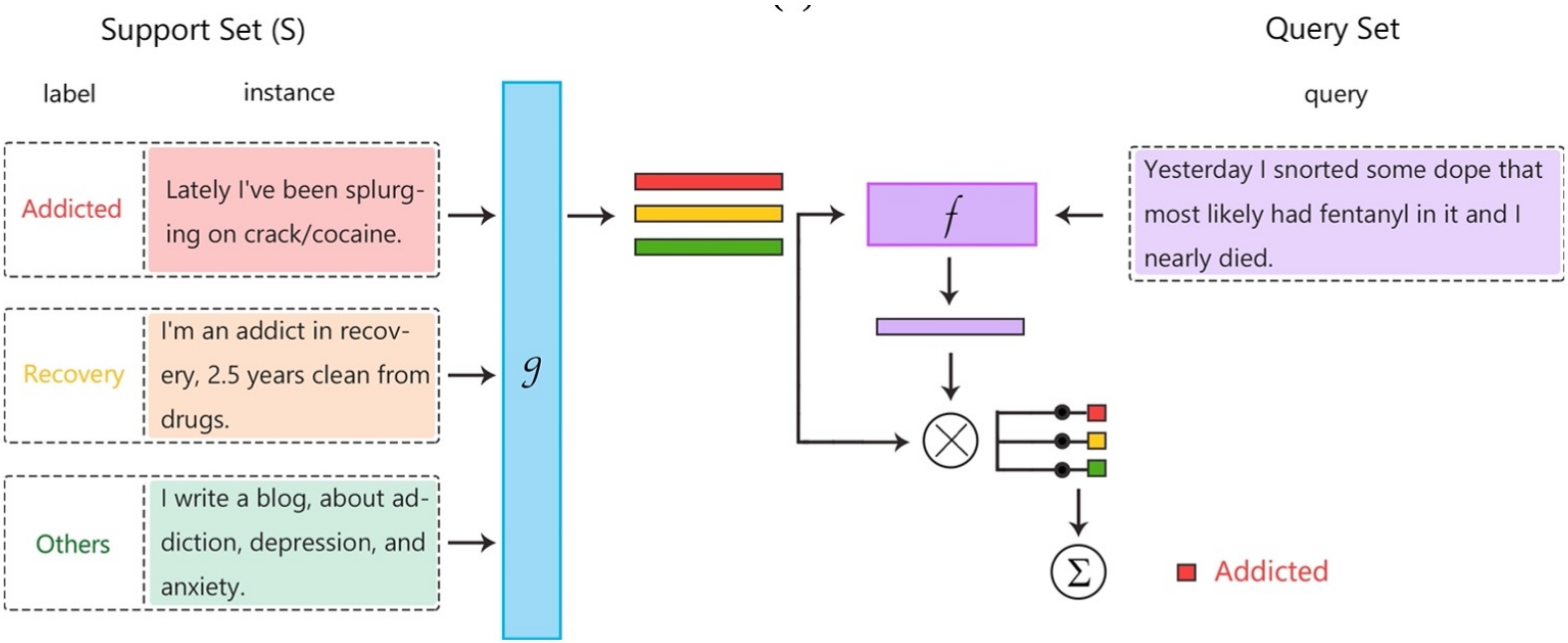
Meta-learning

Each task mimics the few-shot scenario, and can be completely non-overlapping.
Support sets are used to train; query sets are used to evaluate the model



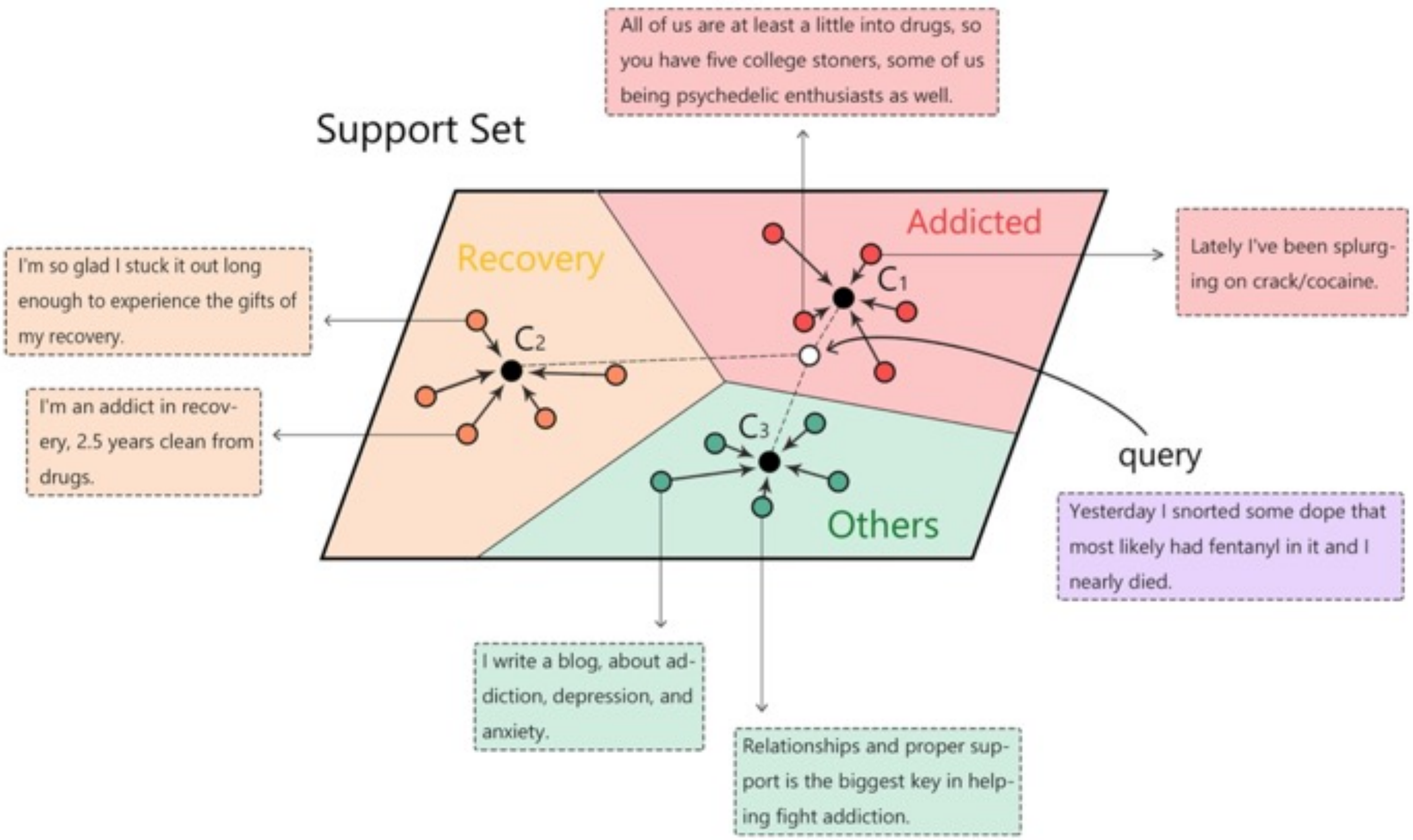
Few-Shot Learning (FSL)

Matching networks



Few-Shot Learning (FSL)

Prototypical network



Source: Yao Ge, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Abeed Sarker (2022). "Few-shot learning for medical text: A systematic review." arXiv preprint arXiv:2204.14081 (2022).

Few-Shot Learning (FSL) for medical text

Study	Year	Data source	Research aim	Size of training set	Number of entities / classes	Entity type of training domain	Entity type of test domain
Alicia Lara-Clares and Ana Garcia-Serrano[44]	2019	MEDDOCAN shared task dataset[45]	NER	500 clinical cases, with no reconstruction	29	Clinical	Clinical
Ferré et al.[46]	2019	BB-norm dataset from the Bacteria Biotope 2019 Task[47]	Entity Normalization	Original dataset with no reconstruction and zero-shot	Not mentioned *	Biological	Biological
Hou et al.[48]	2020	Snips dataset[49]	Slot Tagging (NER)	1-shot and 5-shot	7	Six of Weather, Music, PlayList, Book (including biomedical), Search Screen (including biomedical), Restaurant and Creative Work.	The remaining one
Sharaf et al.[50]	2020	ten different datasets collected from the Open Parallel Corpus (OPUS)[51]	Neural Machine Translation (NMT)	Sizes ranging from 4k to 64k training words (200 to 3200 sentences), but reconstructed	N/A †	Bible, European Central Bank, KDE, Quran, WMT news test sets, Books, European Medicines Agency (EMEA), Global Voices, Medical (ufal-Med), TED talks	Bible, European Central Bank, KDE, Quran, WMT news test sets, Books, European Medicines Agency (EMEA), Global Voices, Medical (ufal-Med), TED talks
Lu et al.[52]	2020	MIMIC II[22] and MIMIC III[23], and EU legislation dataset[53]	Multi-label Text Classification	5-shot for MIMIC II and III, 50-shot for EU legislation	MIMIC II: 9 MIMIC III: 15 EU legis-lation: 5	Medical	Medical

Source: Yao Ge, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Abeed Sarker (2022). "Few-shot learning for medical text: A systematic review." arXiv preprint arXiv:2204.14081 (2022).

Few-Shot Learning (FSL) for medical text

Study	Year	Data source	Research aim	Size of training set	Number of entities / classes	Entity type of training domain	Entity type of test domain
Lu et al.[80]	2021	Constructed and shared a novel dataset[††] based on Weibo for the research of few-shot rumor detection, and use PHEME dataset[81]	Rumor Detection (NER)	For the Weibo dataset: 2-way 3-event 5-shot 9-query; for PHEME dataset: 2-way 2-event 5-shot 9-query	Weibo: 14 PHEME: 5	Source posts and comments from Sina Weibo related to COVID-19	Source posts and comments from Sina Weibo related to COVID-19
Ma et al.[82]	2021	CCLE, CERES-correctedCRISPR gene disruption scores, GDSC1000 dataset, PDTC dataset and PDX dataset[‡‡]	Drug-response Predictions	1-shot, 2-shot, 5-shot and 10-shot	N/A[†]	Biomedical	Biomedical
Kormilitzin et al.[83]	2021	MIMIC-III[23] and UK-CRIS datasets[30, 31]	NER	25%, 50%, 75% and 100% of the training set, with no reconstruction	7	Electronic health record	Electronic health record
Guo et al.[84]	2021	Abstracts of biomedicalliteratures (from relation extraction task of BioNLP Shared Task 2011 and 2019[47]) and structured biological datasets	NER	100%, 75%, 50%, 25%, 0% of training set, with no reconstruction	Not mentioned [*]	Biomedical entities	Biomedical entities
Lee et al.[85]	2021	COVID19-Scientific[86], COVID19-Social[87] (fact-checked by journalists from a website called Politi-fact.com), FEVER[88] (Fact Extraction and Verification, generated by altering sentences extracted from Wiki-pedia to promote research onfact-checking systems)	Fact-Checking (close to Text Classification)	2-shot, 10-shot and 50-shot	Not mentioned [*]	Facts about COVID-19	Facts about COVID-19

Source: Yao Ge, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Abeed Sarker (2022). "Few-shot learning for medical text: A systematic review." arXiv preprint arXiv:2204.14081 (2022).

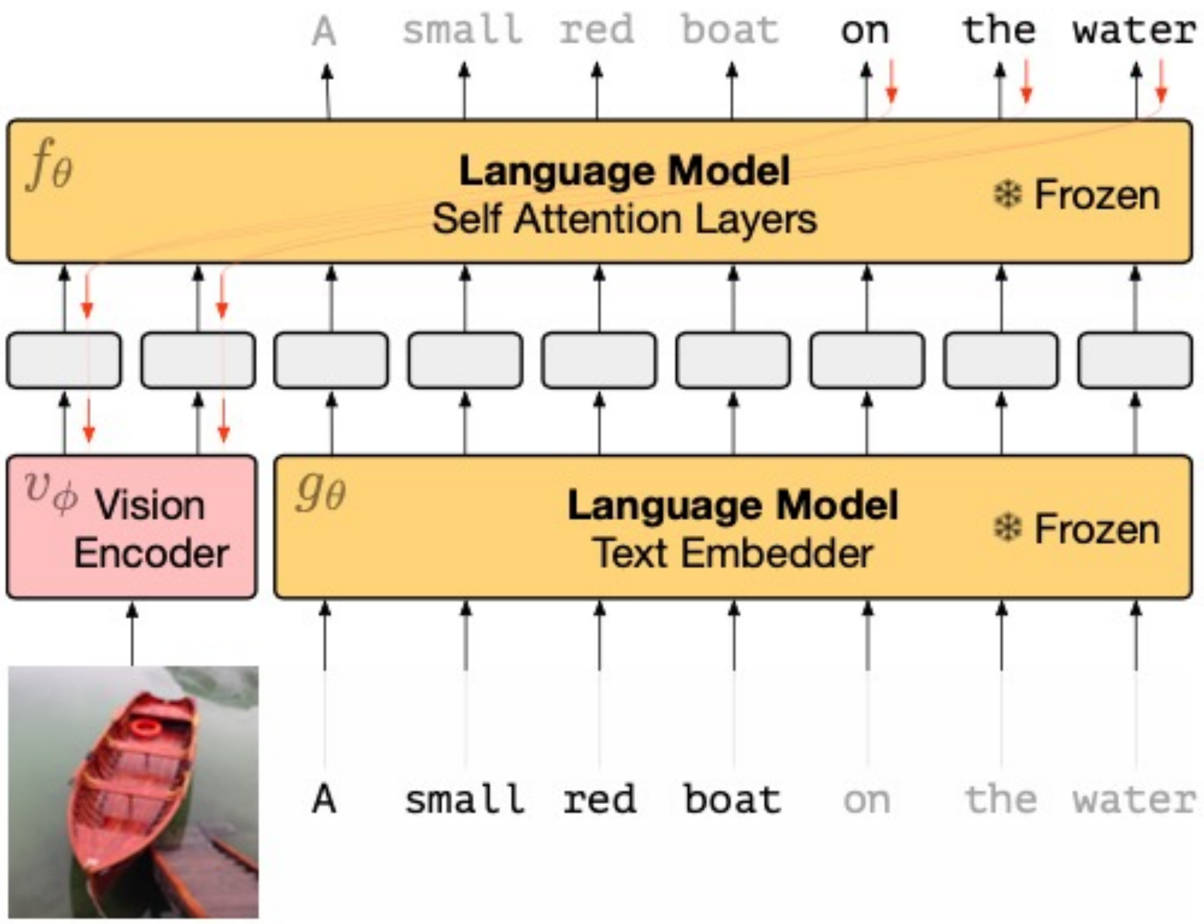
Multimodal Few-Shot Learning with Frozen Language Models

Curated samples with about five seeds required to get past well-known language model failure modes of either repeating text for the prompt or emitting text that does not pertain to the image.

These samples demonstrate the ability to generate open-ended outputs that adapt to both images and text, and to make use of facts that it has learned during language-only pre-training.

Source: Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill (2021). "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.

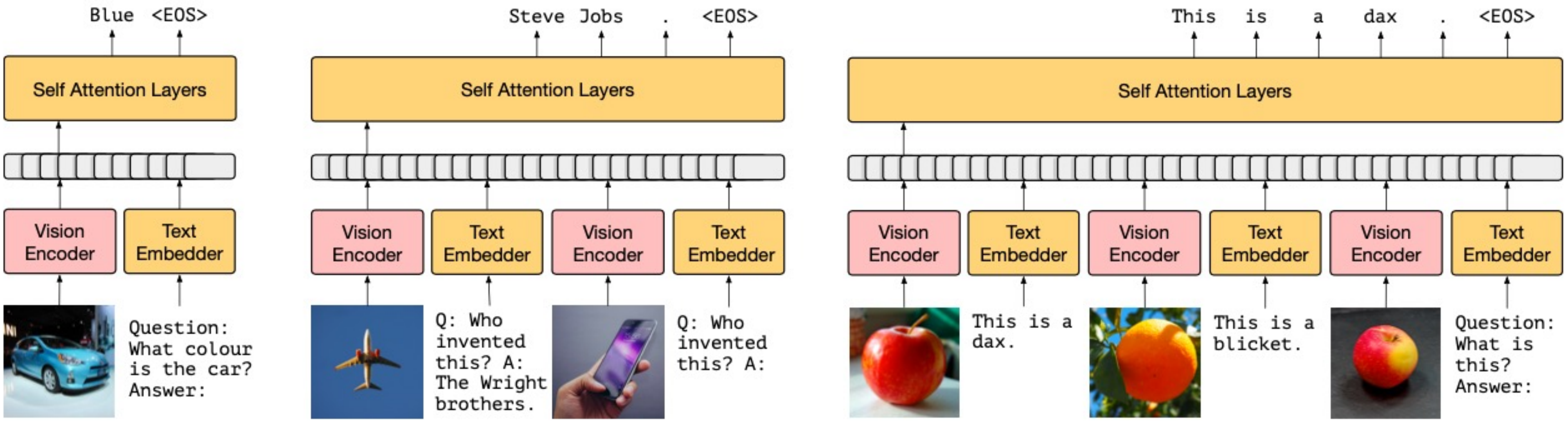
Multimodal Few-Shot Learning with Frozen Language Models



Gradients through a frozen language model's self attention layers are used to train the vision encoder.

Source: Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill (2021). "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.

Multimodal Few-Shot Learning with Frozen Language Models



(a) **0-shot VQA** (b) **1-shot outside-knowledge VQA** (c) **Few-shot image classification**

Inference-Time interface for *Frozen*. The figure demonstrates how we can support (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via in-context learning.

Source: Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill (2021). "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.

Multimodal Few-Shot Learning with Frozen Language Models

Examples of (a) the Open-Ended miniImageNet evaluation (b) the Fast VQA evaluation.

Source: Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill (2021). "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.

GPT-3: Language Models are Few-Shot Learners

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam

Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss

Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh

Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

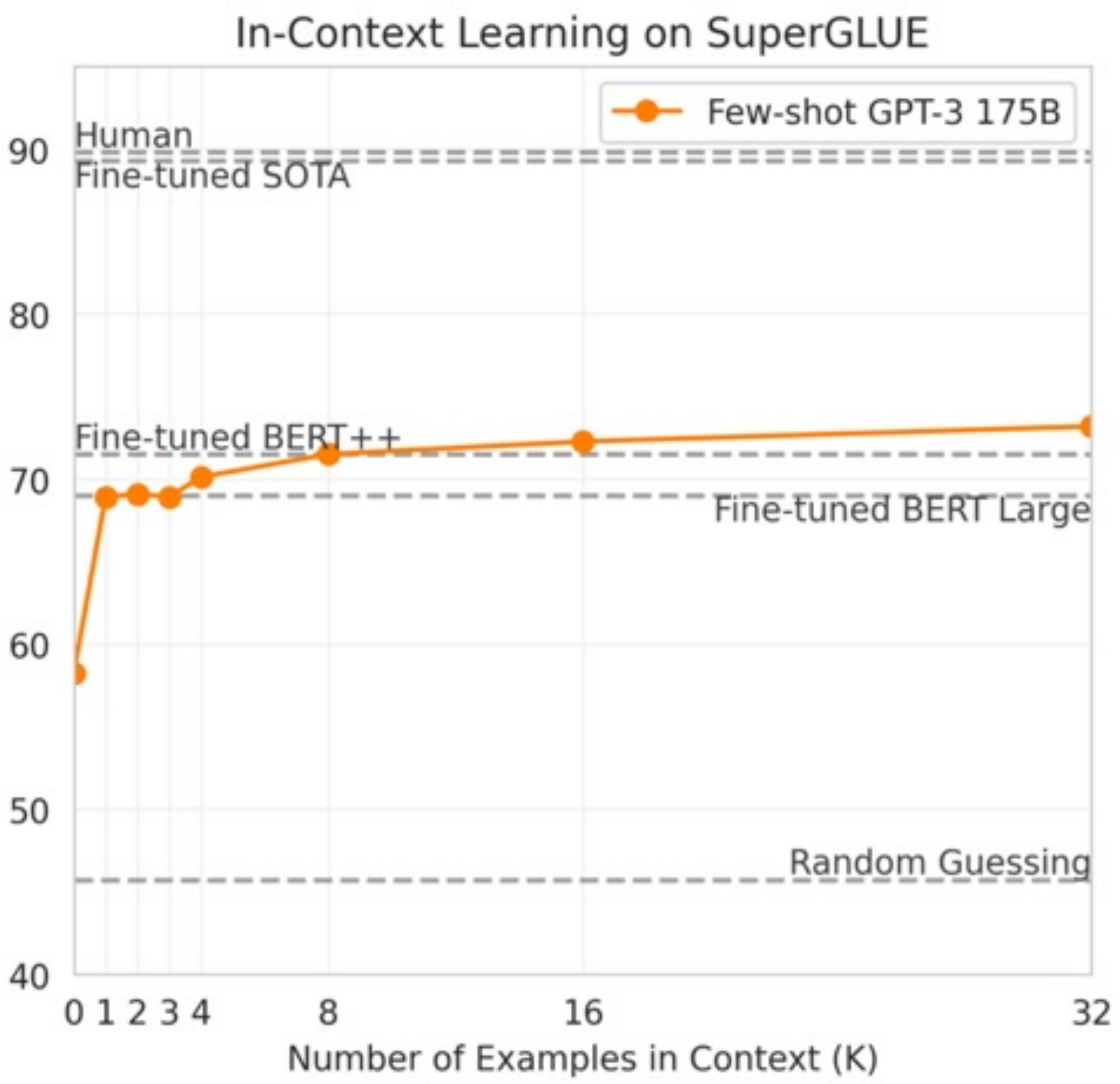
This work was funded by **OpenAI**.
All models were trained on **V100** GPU's on part of a high-bandwidth cluster provided by Microsoft.

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

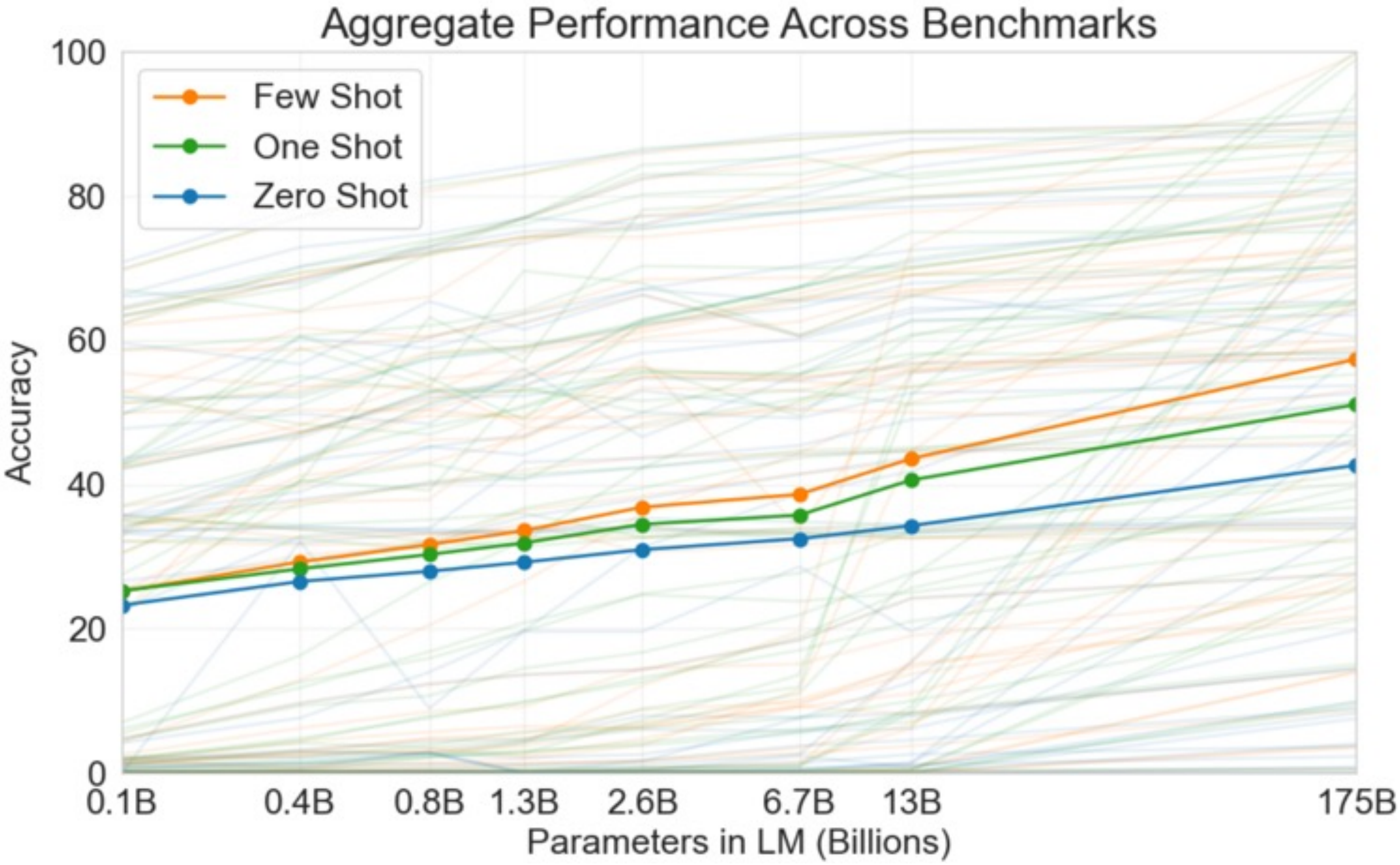
GPT-3: Language Models are Few-Shot Learners



Performance on SuperGLUE increases with model size.
A value of K = 32 means that our model was shown 32 examples per task, for 256 examples total divided across the 8 tasks in SuperGLUE.

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners



Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

Performance on cloze and completion tasks.

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets.

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

Results on three Open-Domain QA tasks

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

GPT-3 is shown in the few-, one-, and zero-shot settings, as compared to prior SOTA results for closed book and open domain settings.
TriviaQA few-shot result is evaluated on the wiki split test server.

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

GPT-3 results on a selection of QA / RC tasks.

Setting	ARC (Easy)	ARC (Challenge)	CoQA	DROP
Fine-tuned SOTA	**92.0**[a]	**78.5**[b]	**90.7**[c]	**89.1**[d]
GPT-3 Zero-Shot	68.8	51.4	81.5	23.6
GPT-3 One-Shot	71.2	53.2	84.0	34.3
GPT-3 Few-Shot	70.1	51.5	85.0	36.5

CoQA and DROP are F1 while ARC reports accuracy.
See the appendix for additional experiments. a[KKS+20] b[KKS+20] c[JZC+19] d [JN20]

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

Few-shot GPT-3 outperforms previous unsupervised NMT work by 5 BLEU when translating into English reflecting its strength as an English LM.

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA

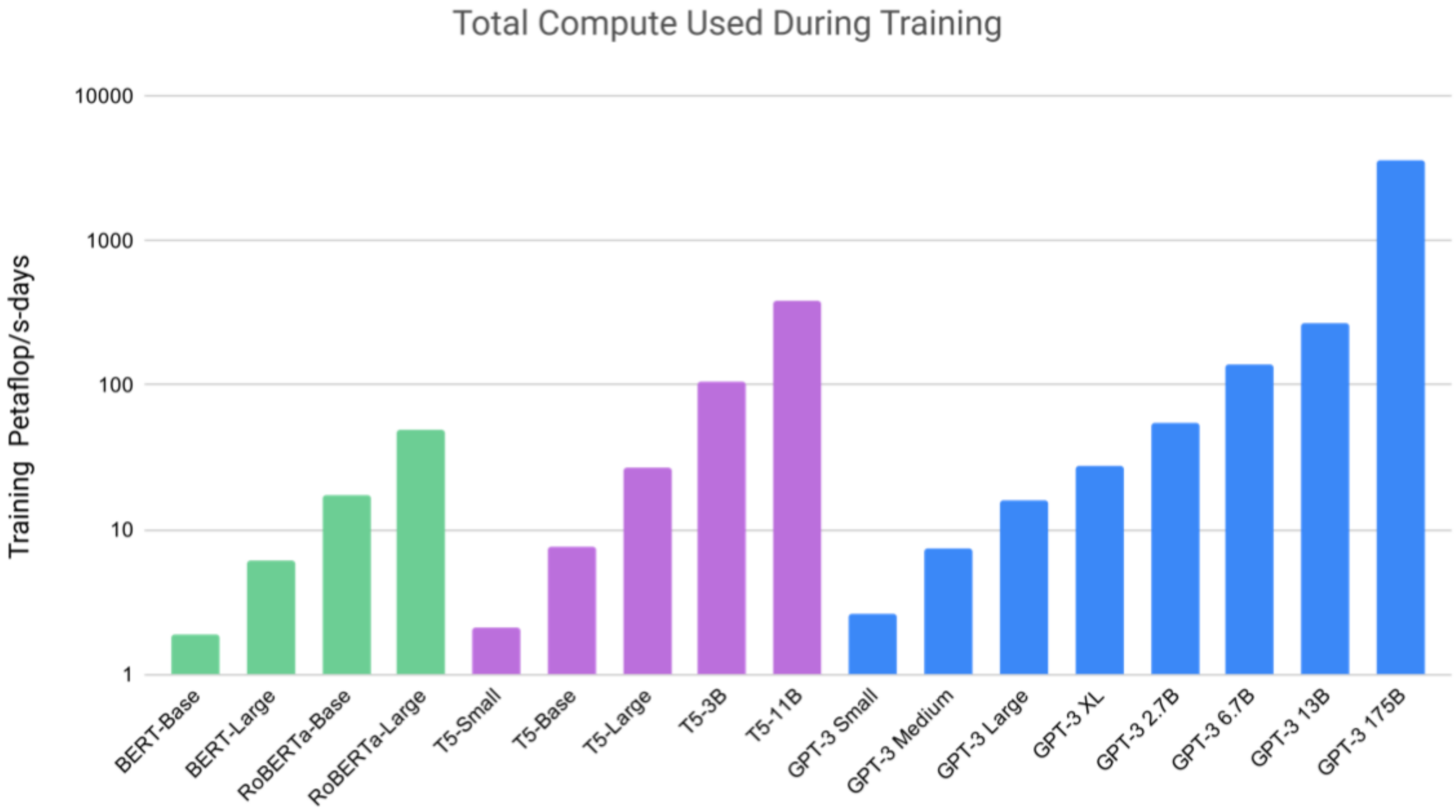
	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners



GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params),
both models took roughly 50 petaflop/s-days of compute during pre-training

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

Human accuracy in identifying whether short (~200 word) news articles are model generated

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

This table compares mean accuracy between five different models, and shows the results of a two-sample T-Test for the difference in mean accuracy between each model and the control model (an unconditional GPT-3 Small model with increased output randomness).

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

GPT-3: Language Models are Few-Shot Learners

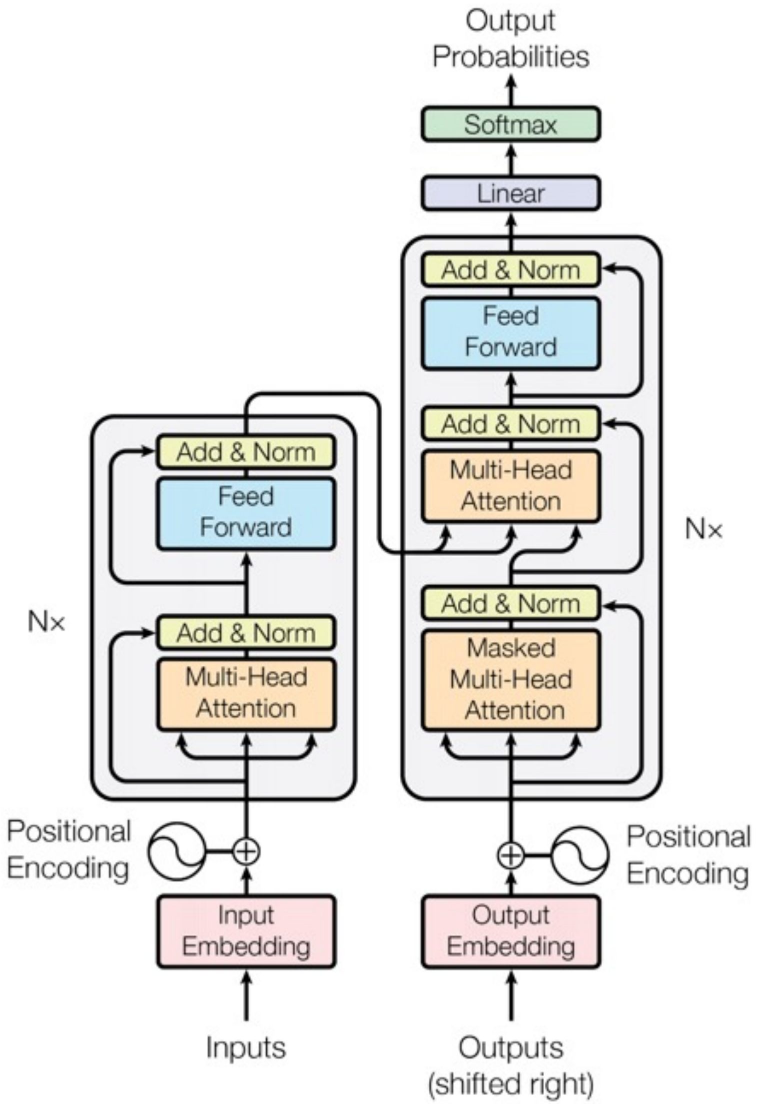
The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%)

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Source: Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. (2020) "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901

Transformer (Attention is All You Need)

(Vaswani et al., 2017)

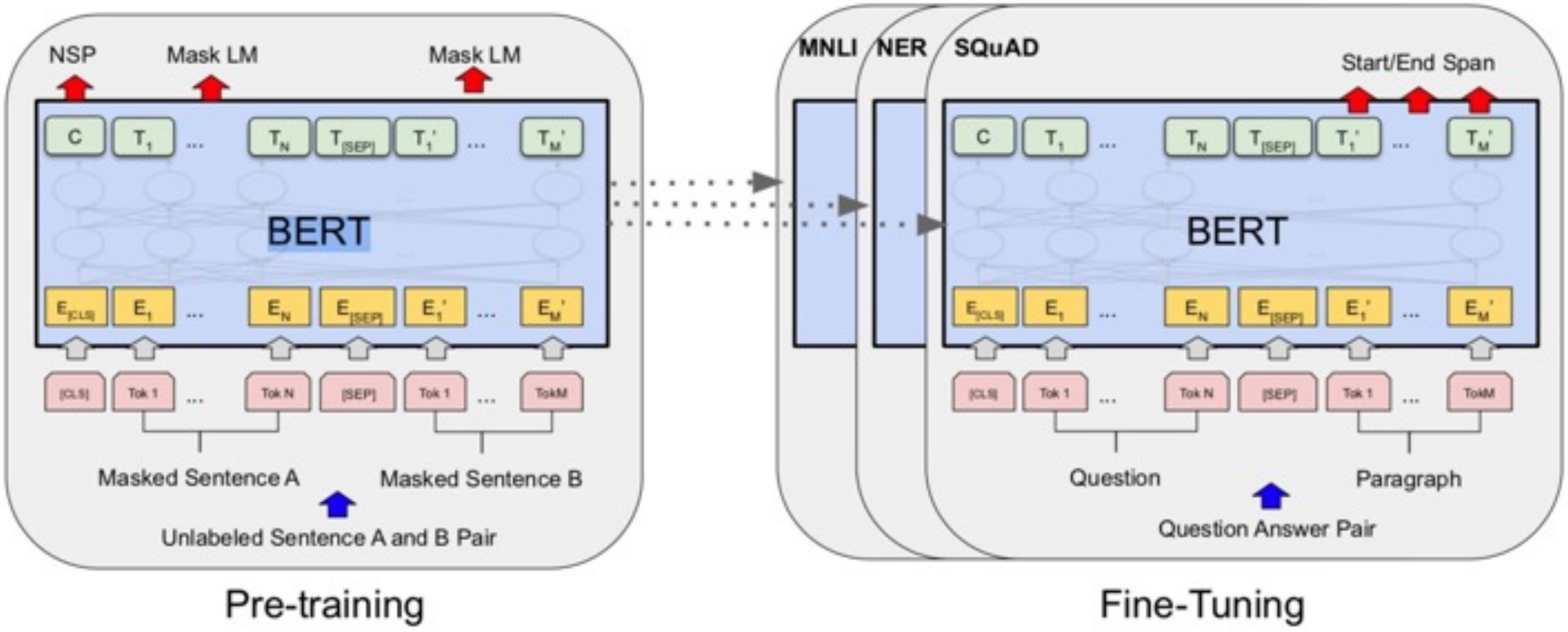


Source: Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

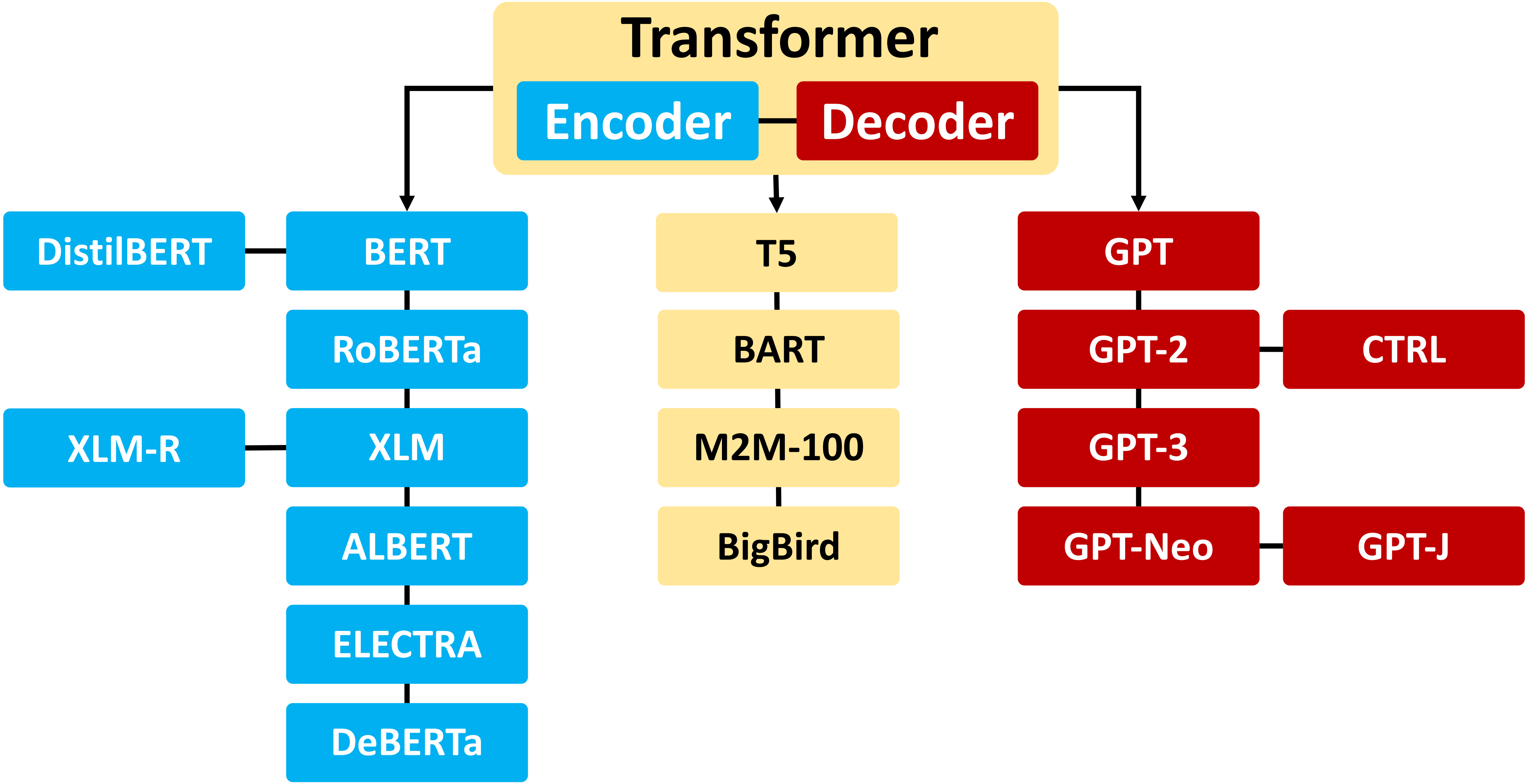
BERT (Bidirectional Encoder Representations from Transformers)

Overall pre-training and fine-tuning procedures for BERT



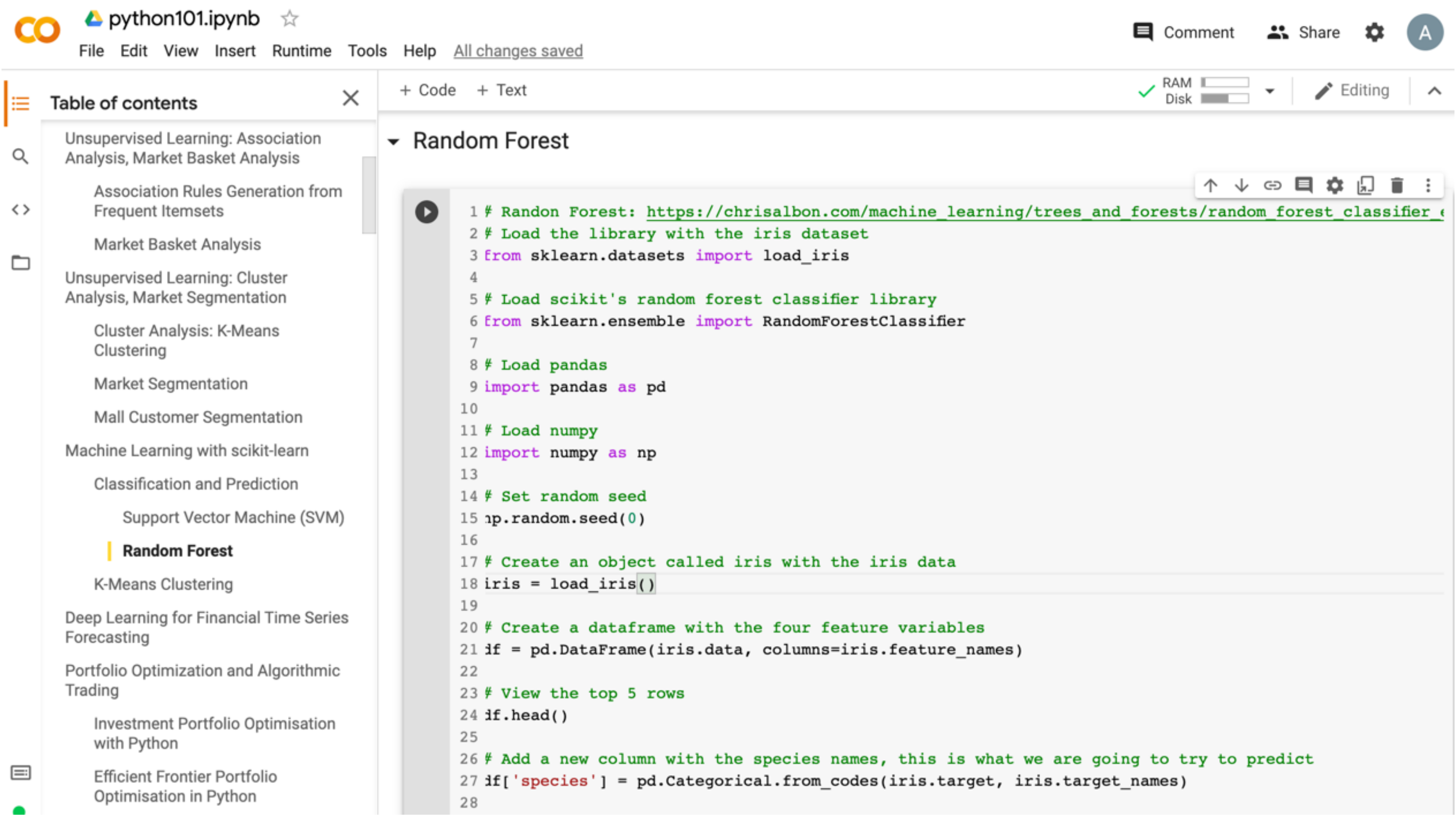
Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Transformer Models



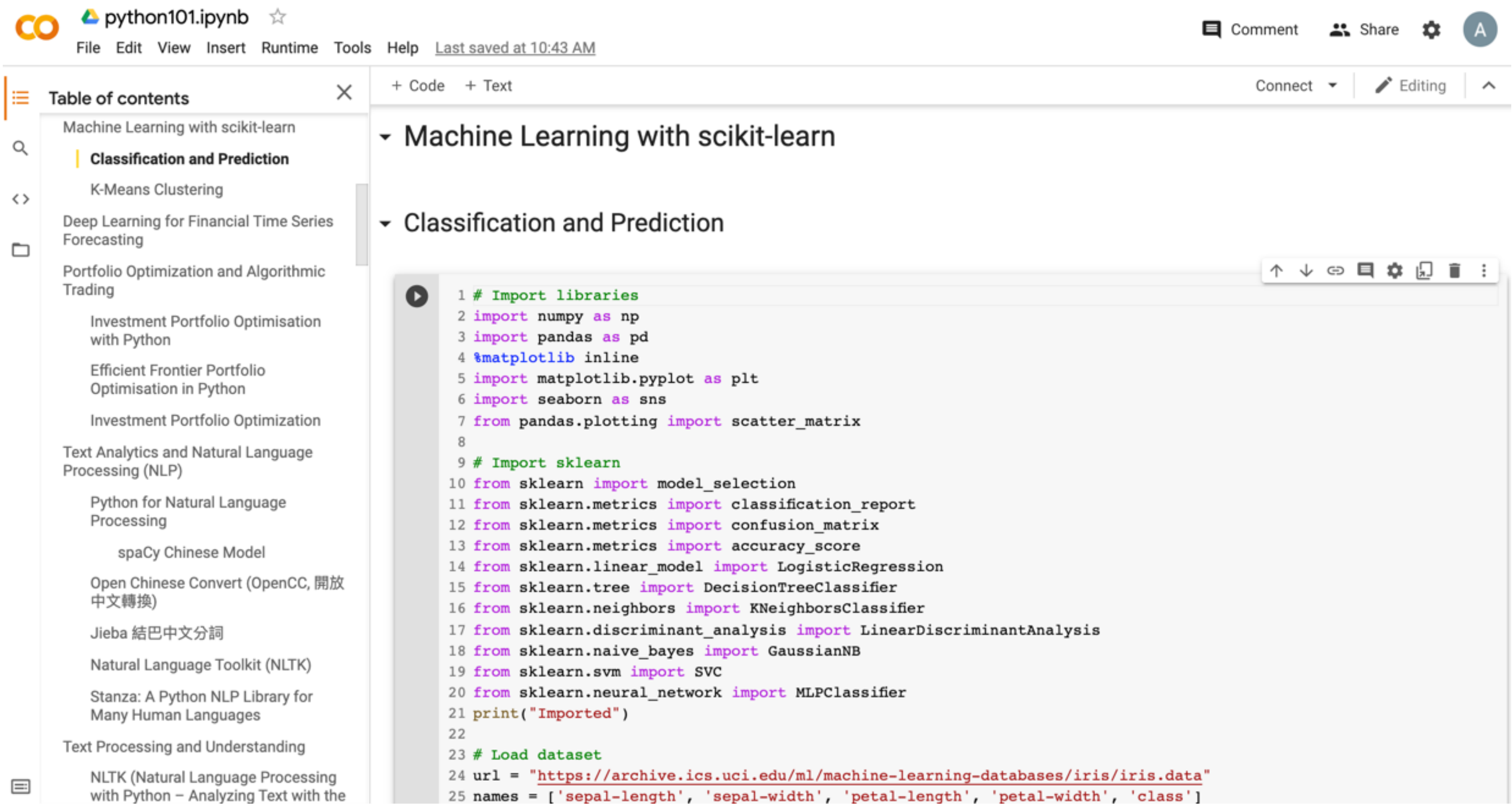
Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.

Machine Learning: Ensemble Learning Random Forest



https://tinyurl.com/aintpupython101

Machine Learning: Supervised Learning Classification and Prediction

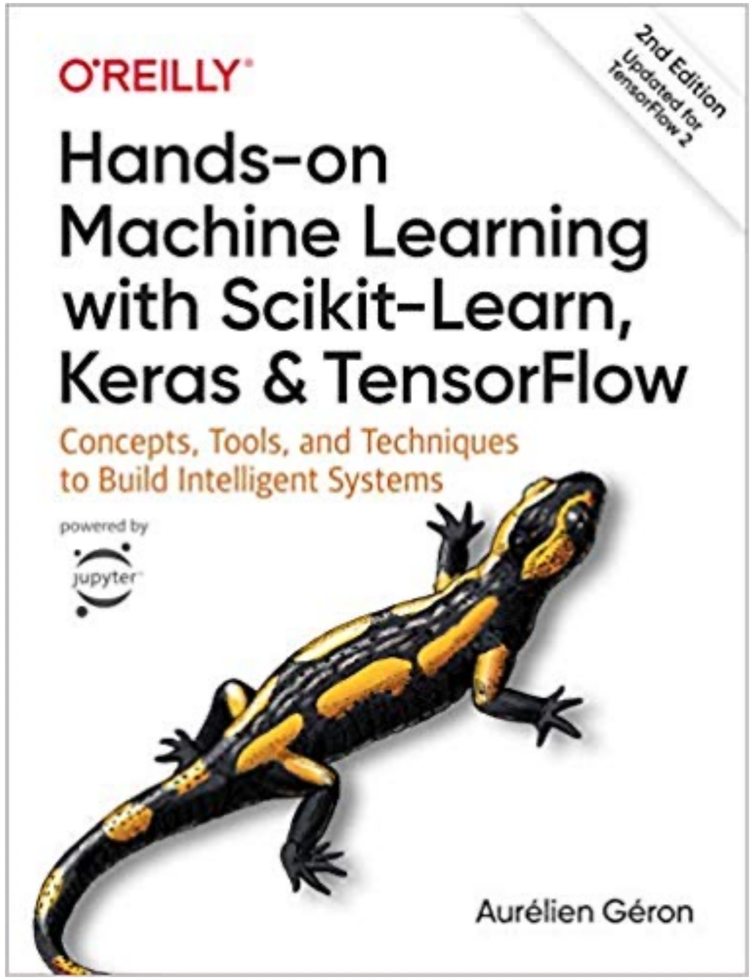


https://tinyurl.com/aintpupython101

Aurélien Géron (2019),

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition

O'Reilly Media, 2019



https://github.com/ageron/handson-ml2

Source: https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1492032646/

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

Notebooks

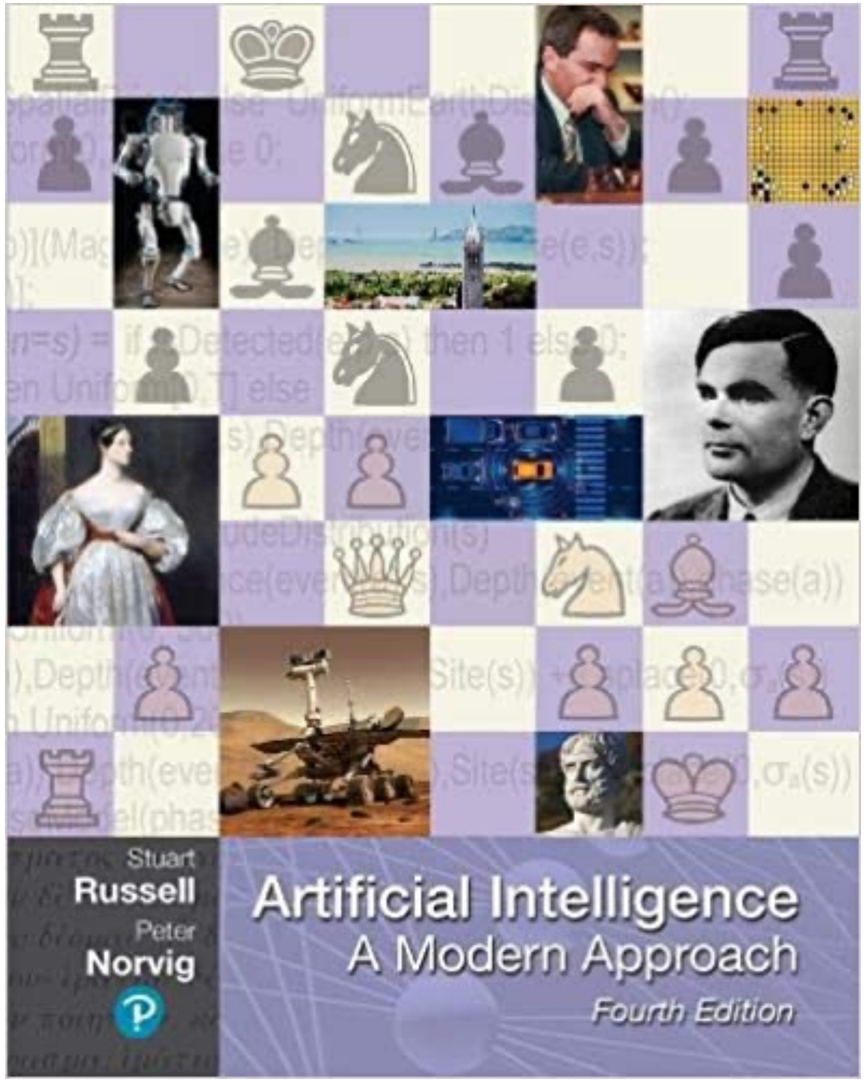
1. The Machine Learning landscape
2. End-to-end Machine Learning project
3. Classification
4. Training Models
5. Support Vector Machines
6. Decision Trees
7. Ensemble Learning and Random Forests
8. Dimensionality Reduction
9. Unsupervised Learning Techniques
10. Artificial Neural Nets with Keras
11. Training Deep Neural Networks
12. Custom Models and Training with TensorFlow
13. Loading and Preprocessing Data
14. Deep Computer Vision Using Convolutional Neural Networks
15. Processing Sequences Using RNNs and CNNs
16. Natural Language Processing with RNNs and Attention
17. Representation Learning Using Autoencoders
18. Reinforcement Learning
19. Training and Deploying TensorFlow Models at Scale

https://github.com/ageron/handson-ml2

Stuart Russell and Peter Norvig (2020),

Artificial Intelligence: A Modern Approach,

4th Edition, Pearson



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

https://www.amazon.com/Artificial-Intelligence-A-Modern-Approach/dp/0134610997/

Artificial Intelligence: A Modern Approach (AIMA)

- **Artificial Intelligence: A Modern Approach (AIMA)**
  - http://aima.cs.berkeley.edu/
- **AIMA Python**
  - http://aima.cs.berkeley.edu/python/readme.html
  - https://github.com/aimacode/aima-python
- **Learning**
  - http://aima.cs.berkeley.edu/python/learning.html

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence: A Modern Approach (AIMA)

- ⌂ US Edition
- ⌂ Global Edition
- Acknowledgements
- Code
- Courses
- Editions
- Errata
- Exercises
- Figures
- Instructors Page
- Pseudocode
- Reviews

Artificial Intelligence: A Modern Approach, 4th US ed.

by Stuart Russell and Peter Norvig

The authoritative, most-used AI textbook, adopted by over **1500** schools.

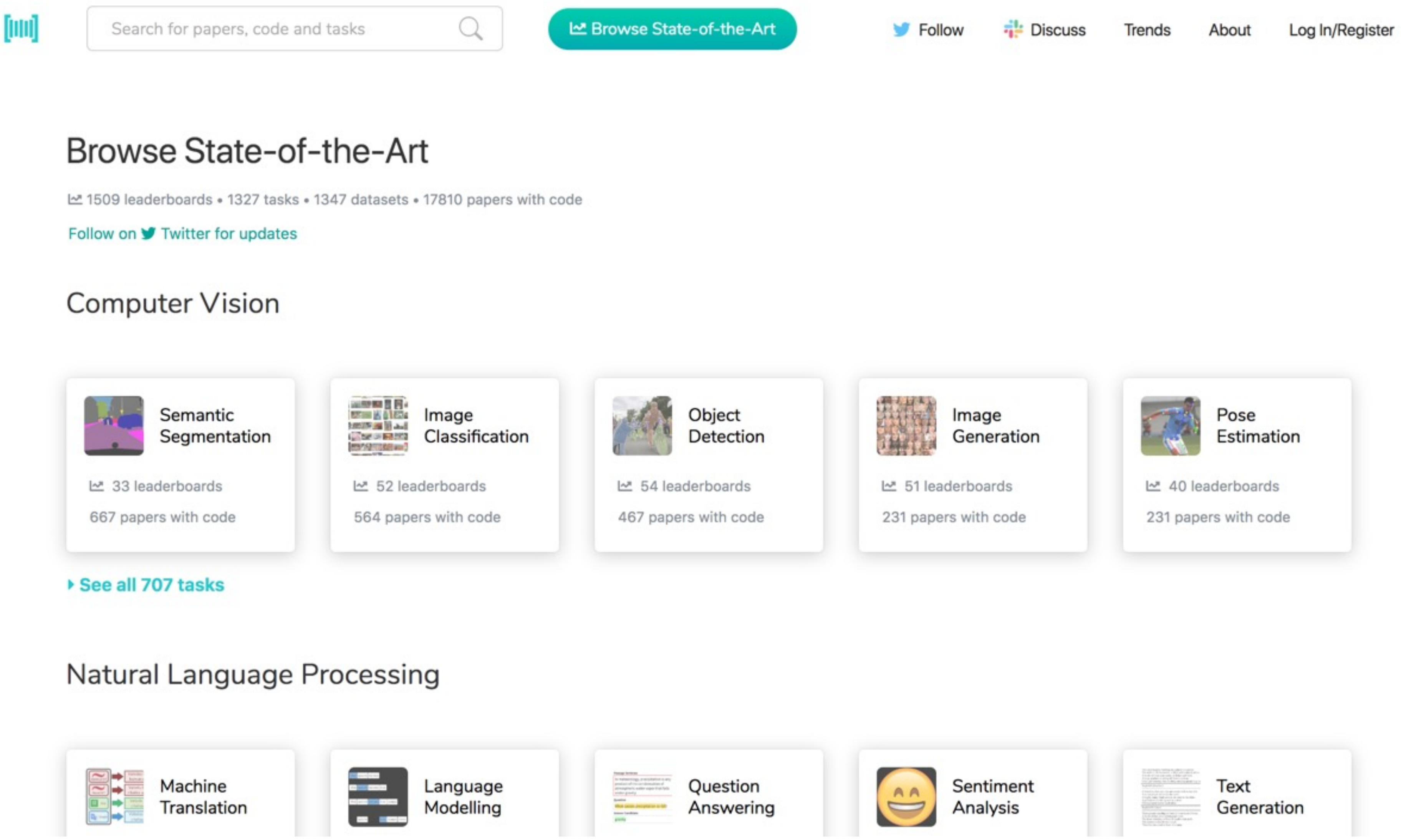
Table of Contents for the US Edition (or see the Global Edition)

Exercises (website)
Figures (pdf)
Code (website); Pseudocode (pdf)
Covers: US, Global

http://aima.cs.berkeley.edu/

Papers with Code

State-of-the-Art (SOTA)



https://paperswithcode.com/sota

Summary

- The Theory of Learning
  - Computational Learning Theory
  - Probably Approximately Correct (PAC) Learning
- Ensemble Learning
  - Bagging: Random forests
  - Stacking
  - Boosting: Gradient boosting
  - Online learning
- Meta Learning: Learning to Learn

References

- Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson.
- Aurélien Géron (2019), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition, O’Reilly Media.
- Steven D'Ascoli (2022), Artificial Intelligence and Deep Learning with Python: Every Line of Code Explained For Readers New to AI and New to Python, Independently published.
- Nithin Buduma, Nikhil Buduma, Joe Papa (2022), Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms, 2nd Edition, O'Reilly Media.
- Min-Yuh Day (2022), Python 101, https://tinyurl.com/aintpupython101