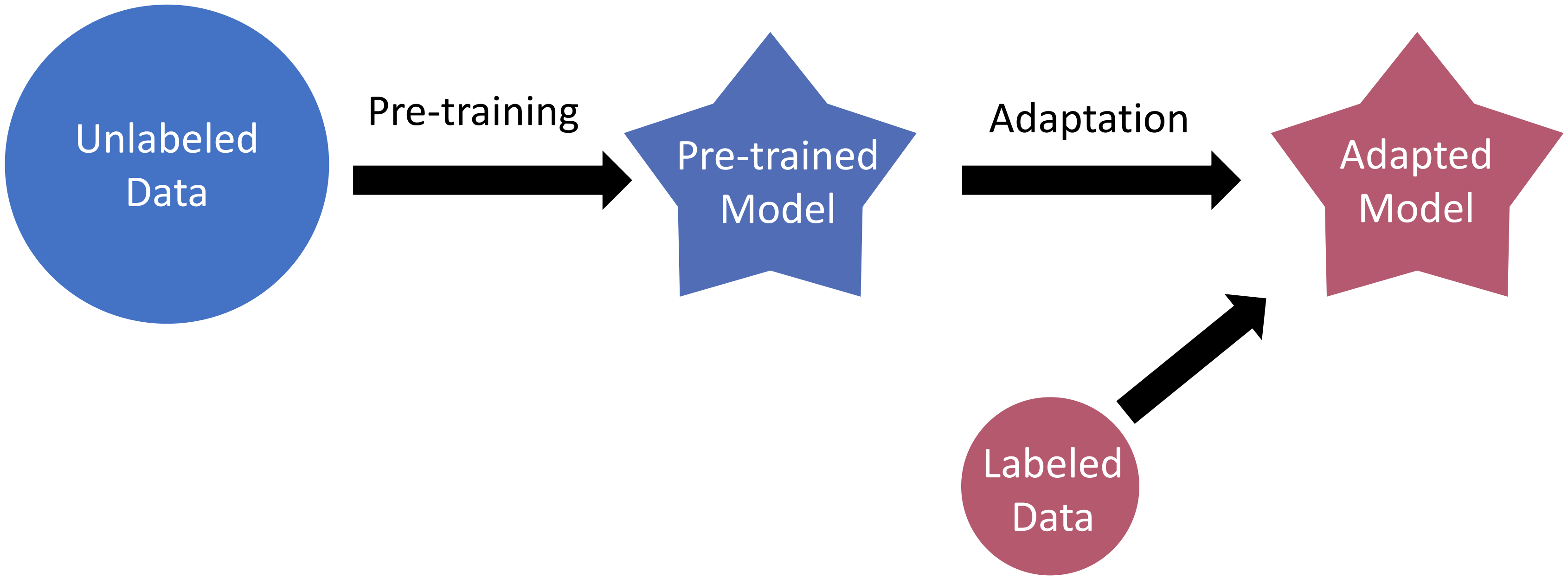
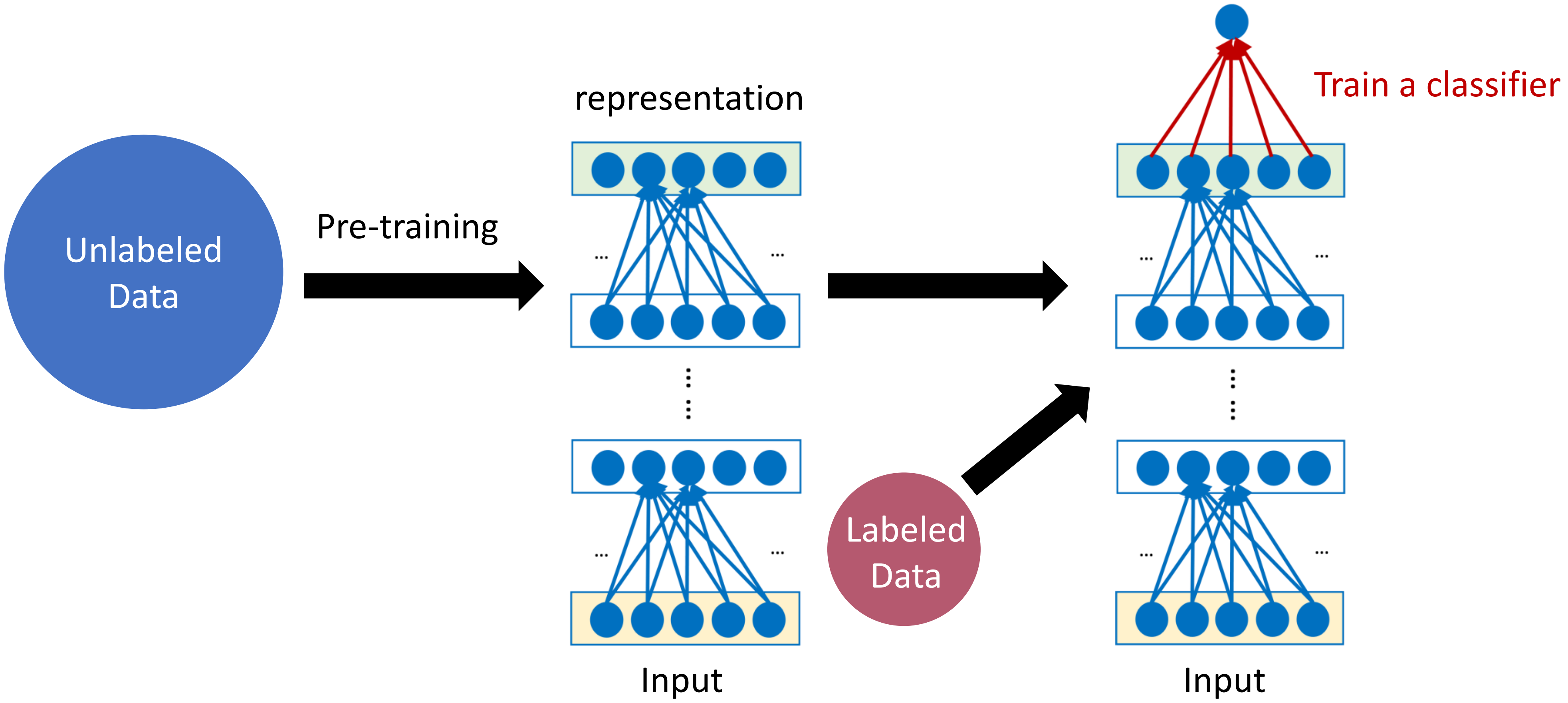
New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning → pre-training + adaptation



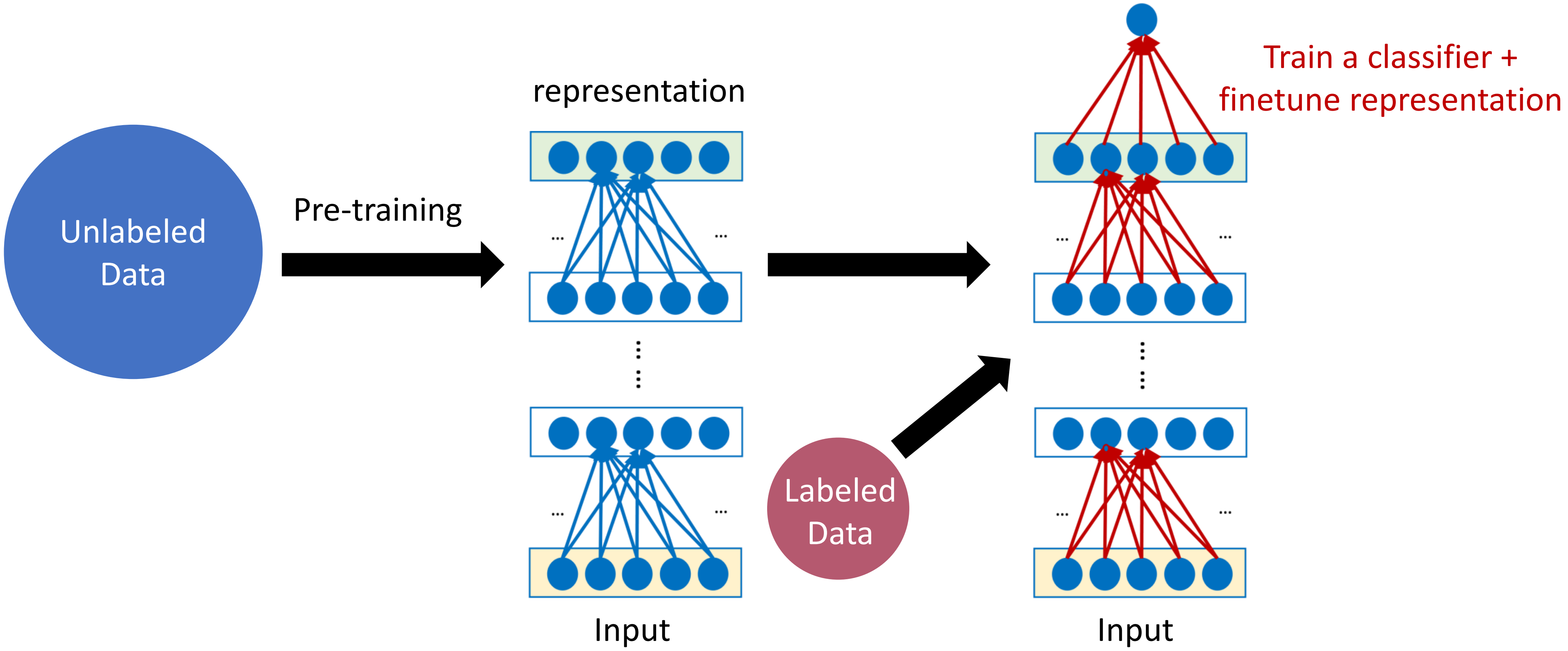
New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning → pre-training + adaptation



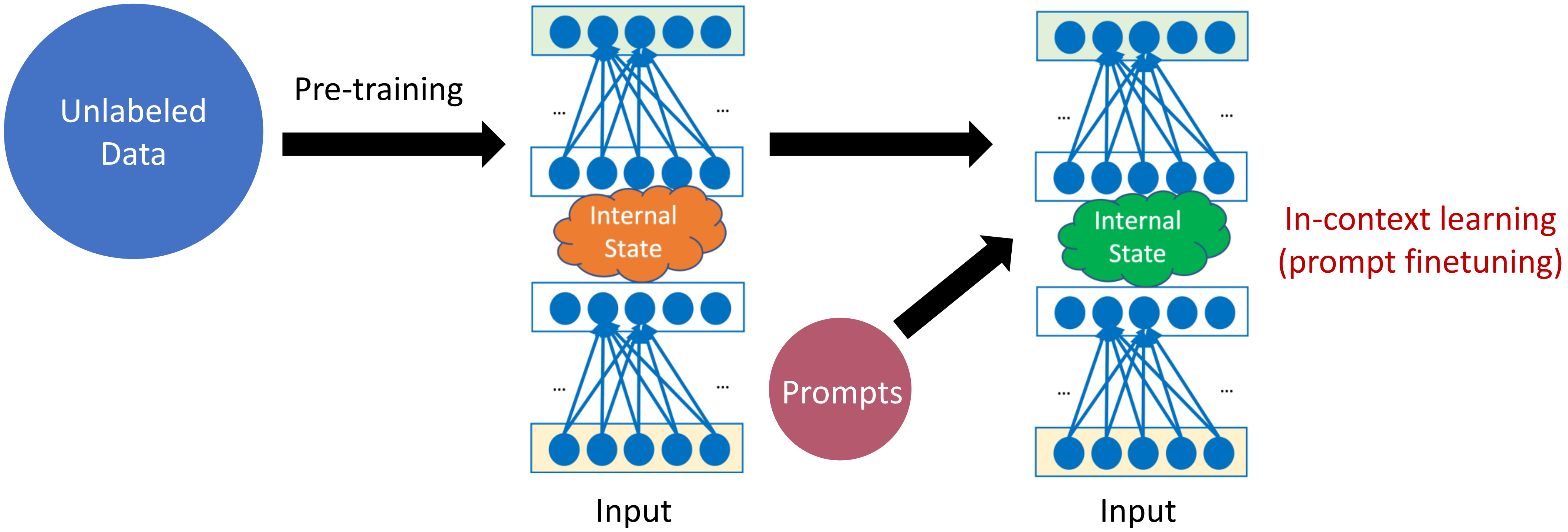
New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning → pre-training + adaptation



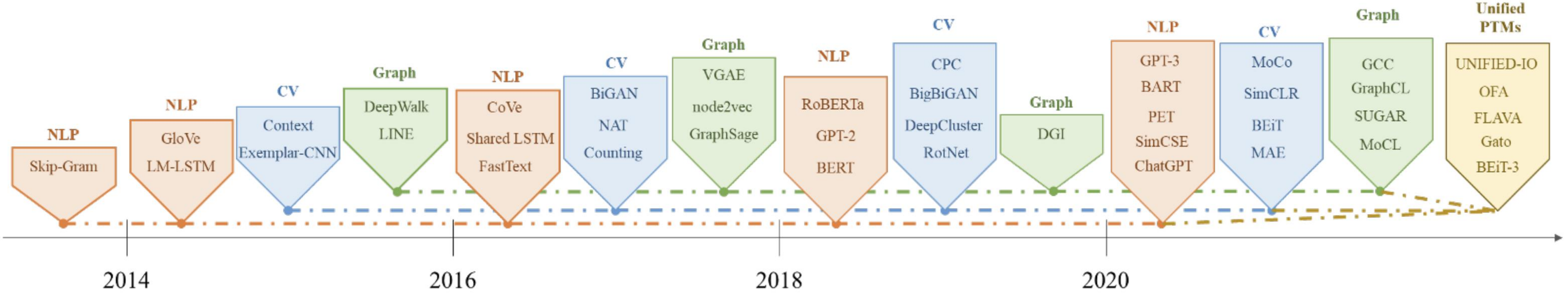
New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning → pre-training + adaptation



New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning → pre-training + adaptation

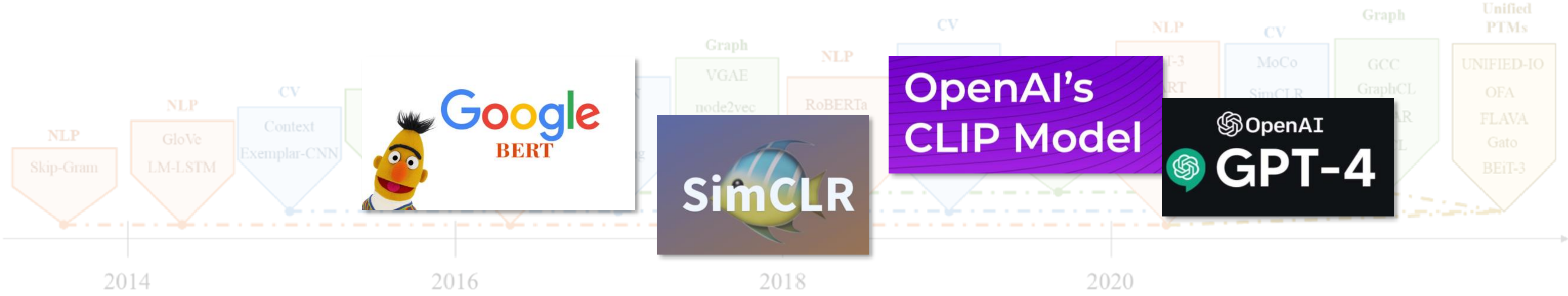


The history and evolution of pre-trained models

[ZLL+23] Zhou, Li, Li, et al. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. 2023.

New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning → pre-training + adaptation



The history and evolution of pre-trained models

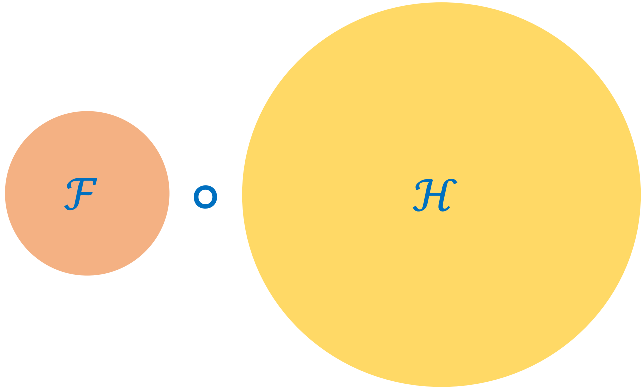
[ZLL+23] Zhou, Li, Li, et al. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. 2023.

Label Efficiency

- Great performance with limited labeled data in downstream tasks

How to quantify the benefit of knowledge transfer?

- Pre-train $h \in \mathcal{H}$, then learn a classifier $f \in \mathcal{F}$ to get final model $f \circ h$
- Pre-train minimizes an unsupervised loss to $\leq \epsilon_{pre}$
- Without pre-train: $\mathcal{F} \circ \mathcal{H}$



[GL20] Garg and Liang. Functional Regularization for Representation Learning: A Unified Theoretical Perspective. NeurIPS'2020.

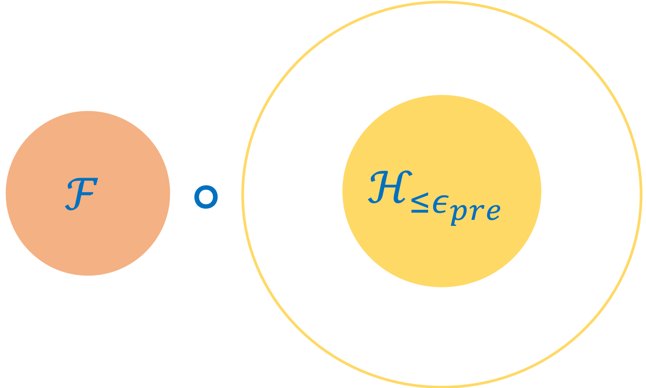
Label Efficiency

- Great performance with limited labeled data in downstream tasks

How to quantify the benefit of knowledge transfer?

- Pre-train $h \in \mathcal{H}$, then learn a classifier $f \in \mathcal{F}$ to get final model $f \circ h$
- Pre-train minimizes an unsupervised loss to $\leq \epsilon_{pre}$

- Without pre-train: $\mathcal{F} \circ \mathcal{H}$
- With pre-train: $\mathcal{F} \circ \mathcal{H}_{\leq \epsilon_{pre}}$



[GL20] Garg and Liang. Functional Regularization for Representation Learning: A Unified Theoretical Perspective. NeurIPS'2020.

Universality

- Generally applicable to different tasks

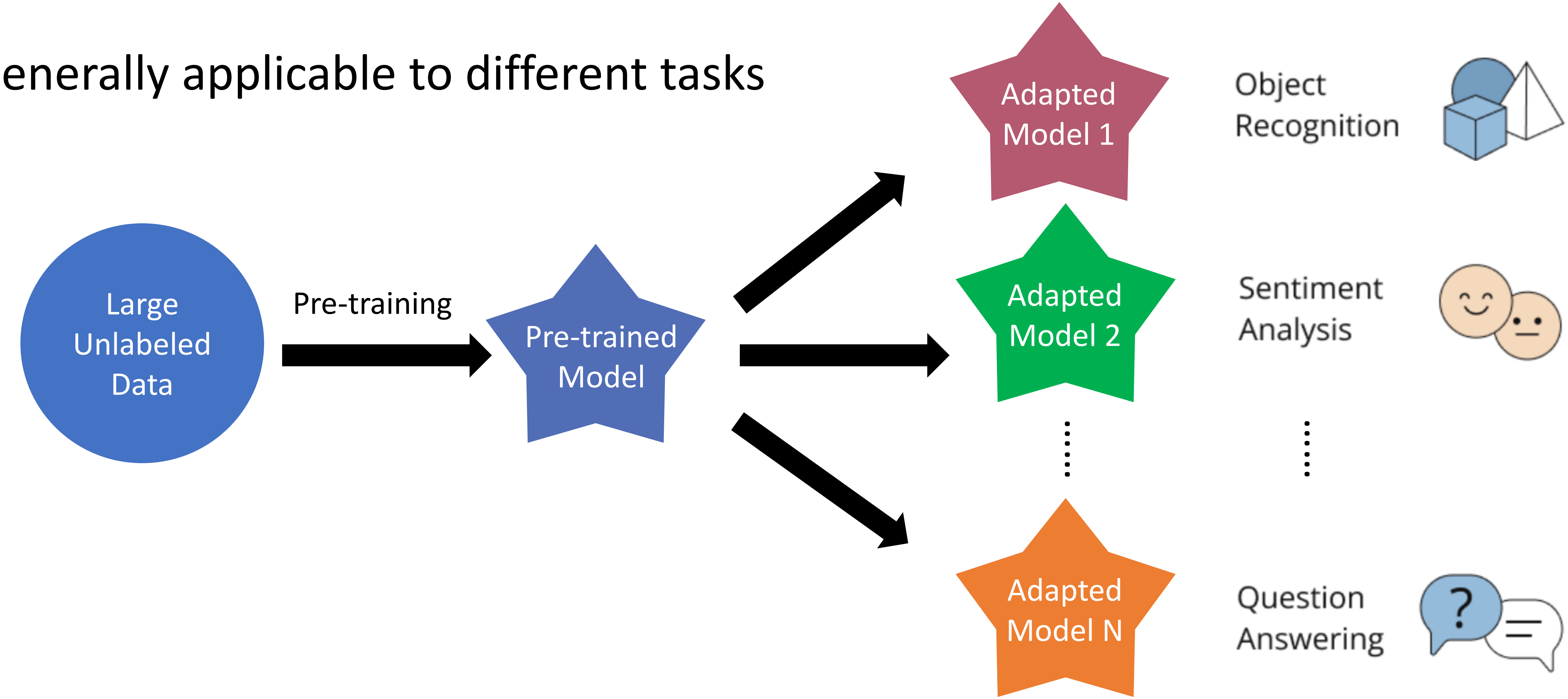
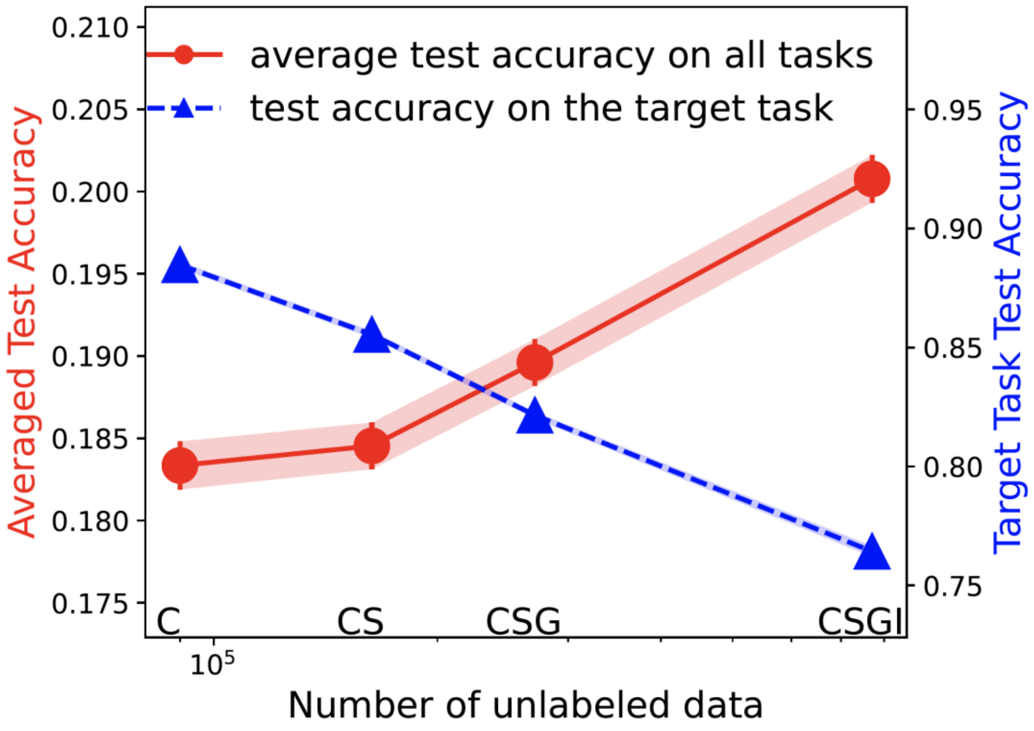


Figure credit: James John Williams, digt.com

Trade-off of Label Efficiency and Universality

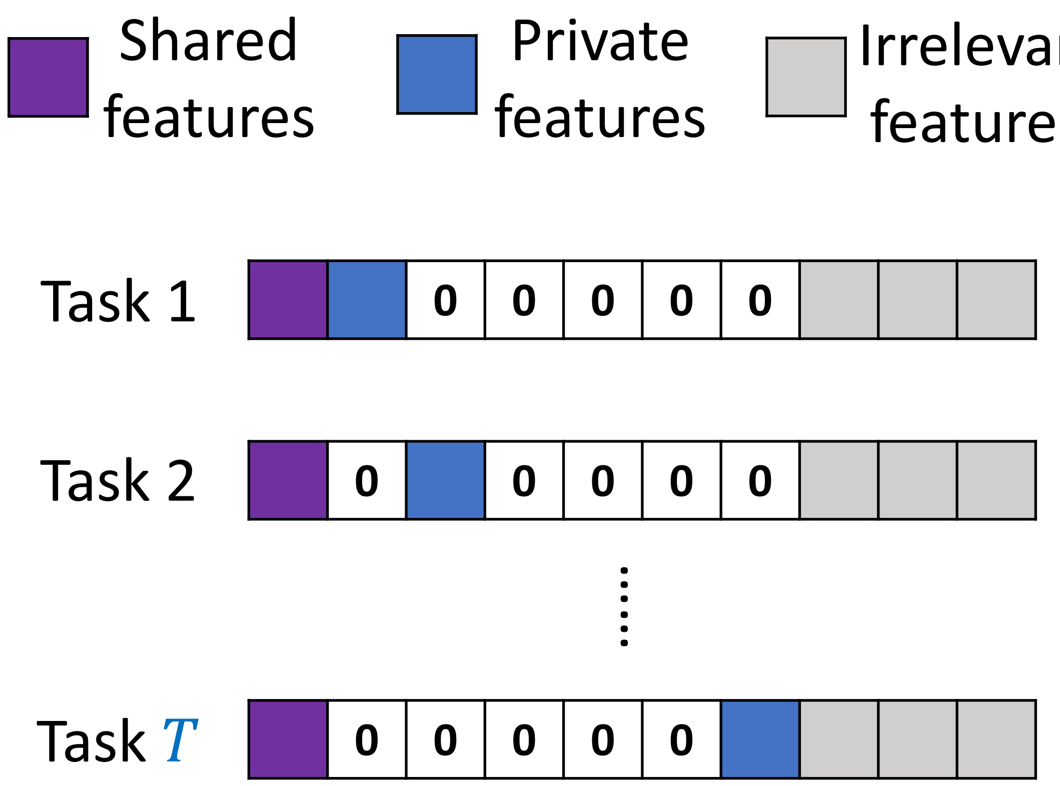
Contrastive learning ResNet18 backbone via MoCo, then classify on CIFAR10.

From left to right, incrementally add to pre-training: CINIC-10 (C), SVHN (S), GTSRB (G), and ImageNet32 (I)



[SCL+23] Shi, Chen, Li, Raghuram, Wu, Liang, Jha. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning. ICLR'2023.

Trade-off Comes from Feature Weighting



- Input: linearly generated from features
- Label: linear on shared/private features

- Pre-trained on Task 1:
  - Recover features for Task 1 but not for others
  - Good prediction on Task 1 but not on others
- Pre-trained on mixture of all tasks:
  - Recover all shared/private features
  - Up-weights the shared features by $O(\sqrt{T})$
  - $O(\sqrt{T})$ worse on Task 1 but better on average

[SCL+23] Shi, Chen, Li, Raghuram, Wu, Liang, Jha. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning. ICLR'2023.