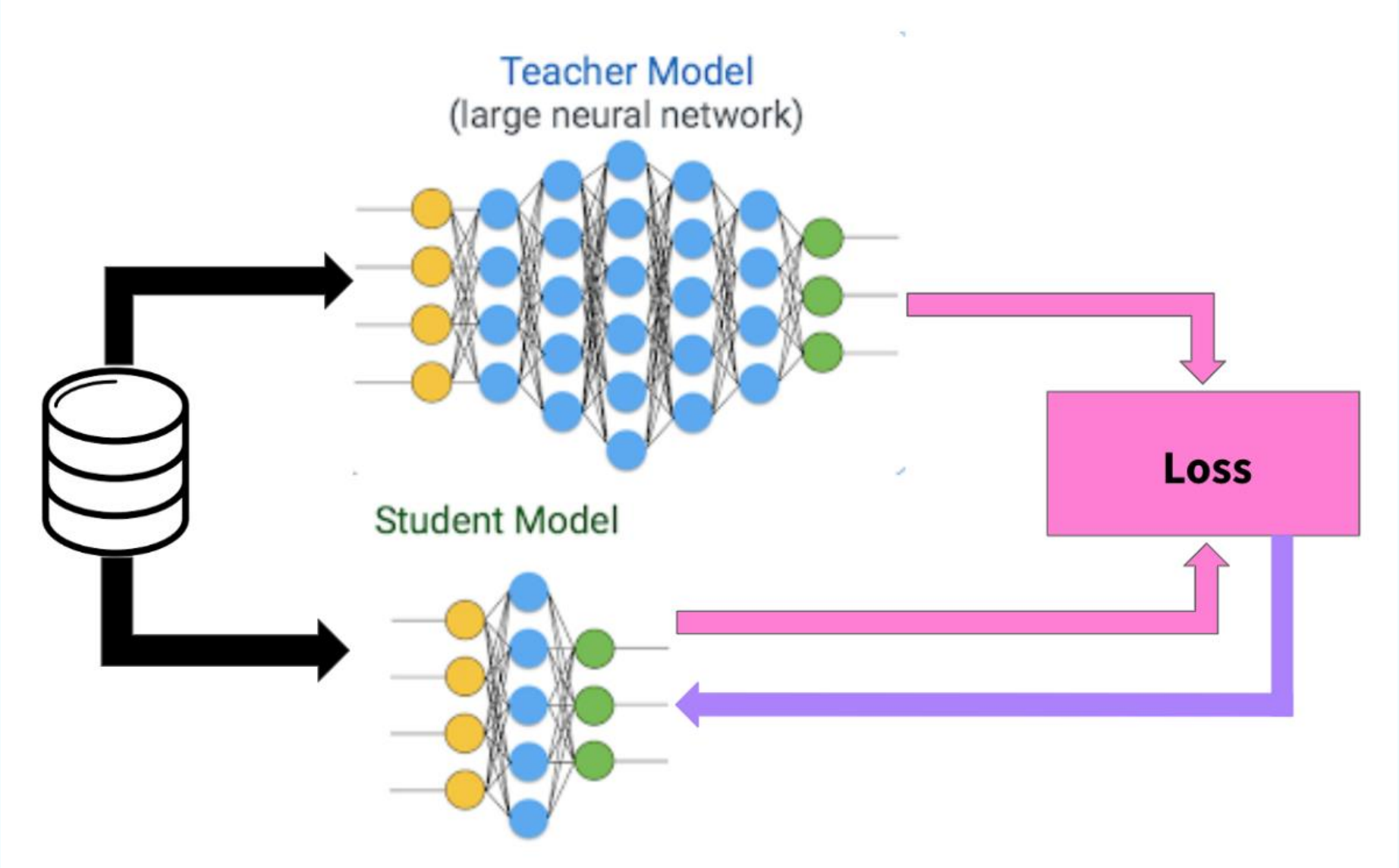
Lecture 6: Model compression, Uncertainty estimation, Active learning

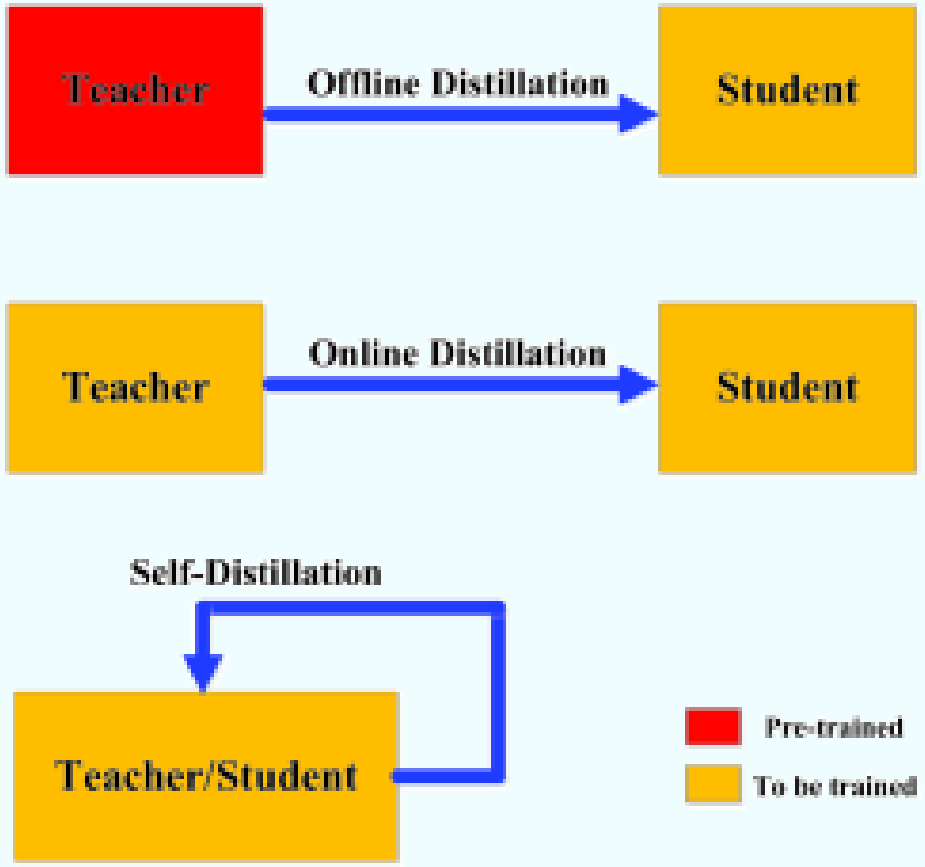
Leonid Sanochkin

Knowledge distillation



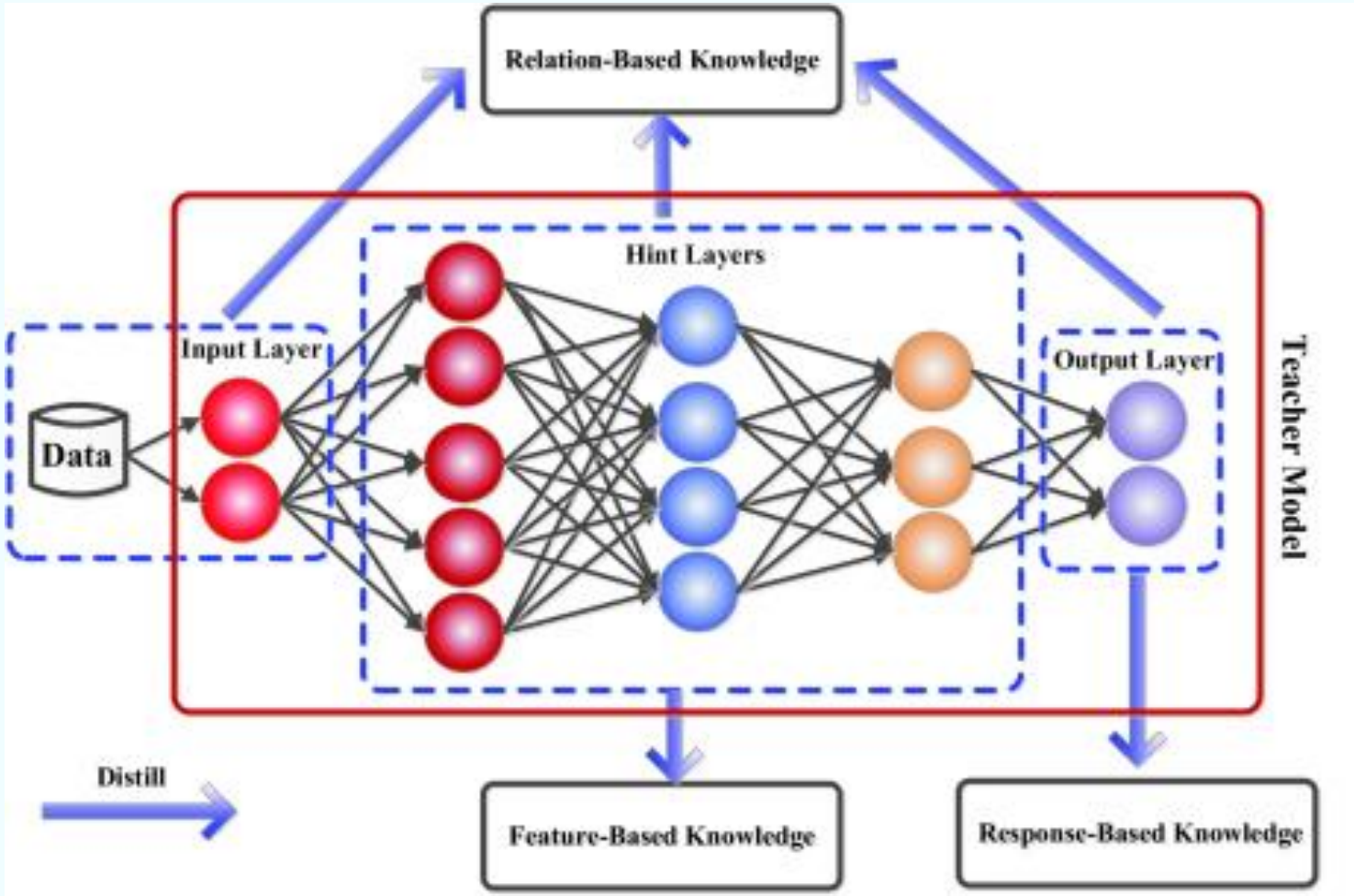
Knowledge distillation

- Offline distillation
- Online distillation
- Self-distillation

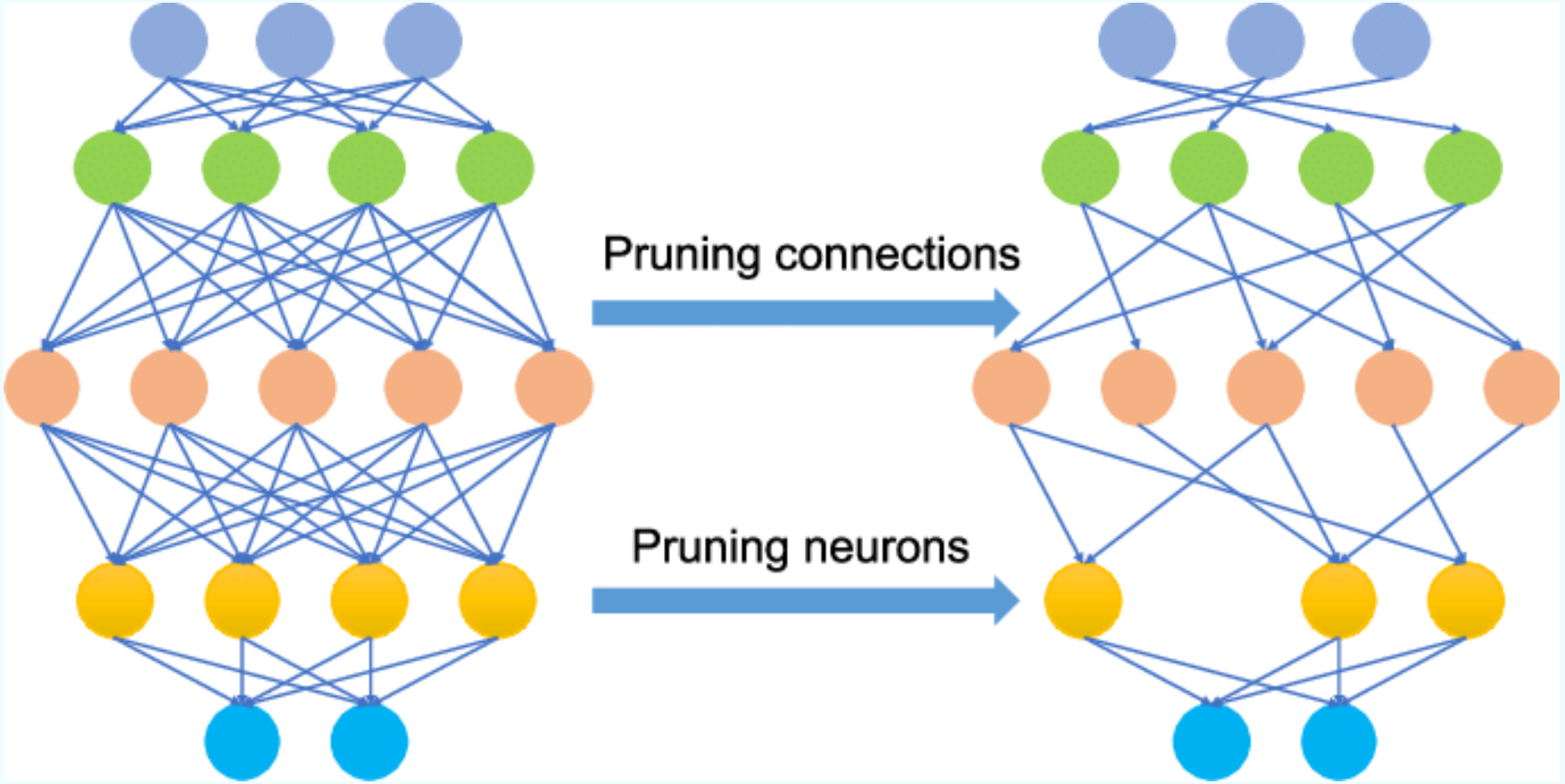


Knowledge distillation

- Response-based
- Feature-based
- Relation-based



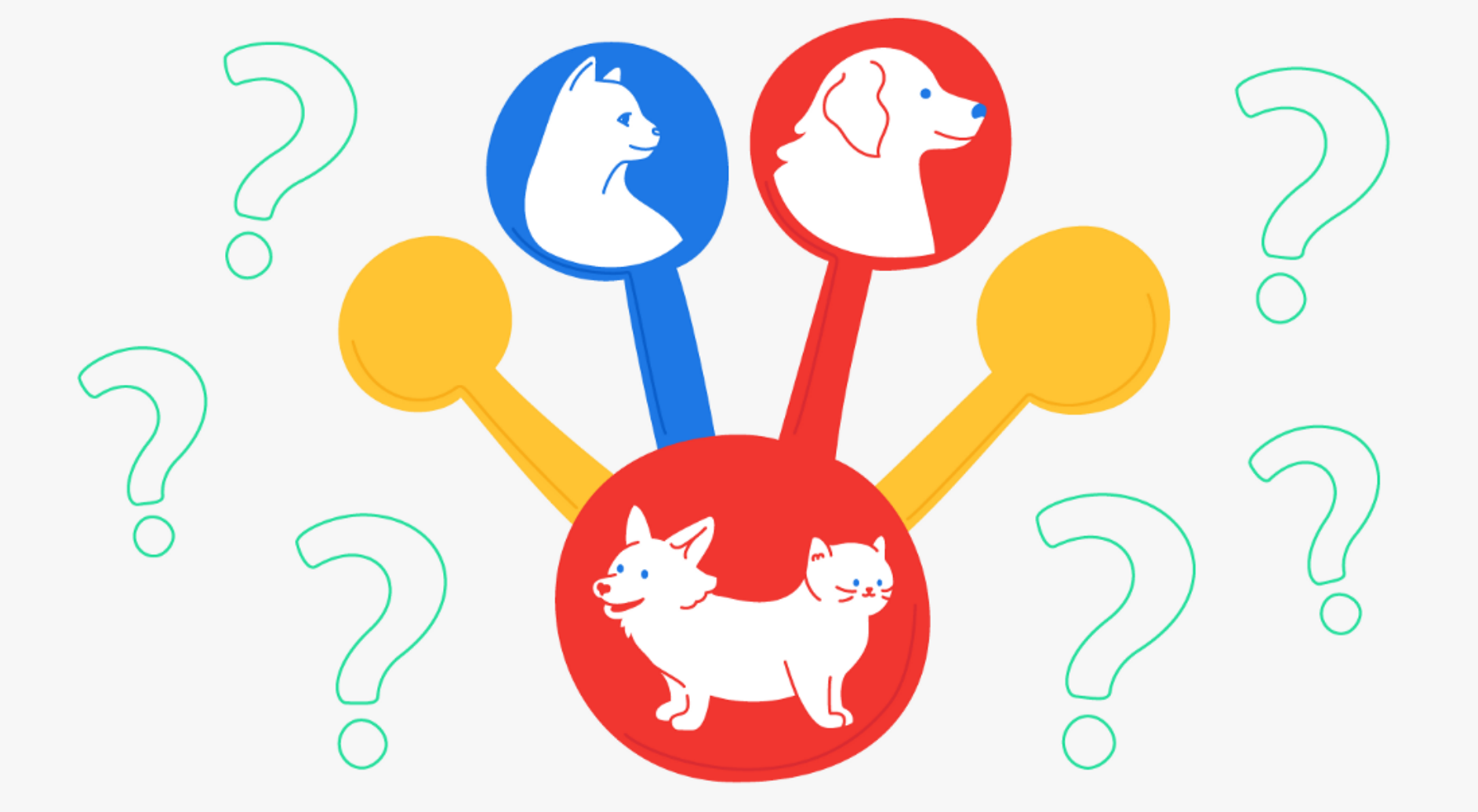
Pruning



Quantization

Quantization Modes	Data Requirements	Inference Latency	Inference Accuracy Loss
Dynamic Quantization	No Data	Usually Faster	Smallest
Static Quantization	Unlabeled Data	Fastest	Smaller
Quantization Aware Training	Labeled Data	Fastest	Smallest

Uncertainty estimation



Uncertainty estimation

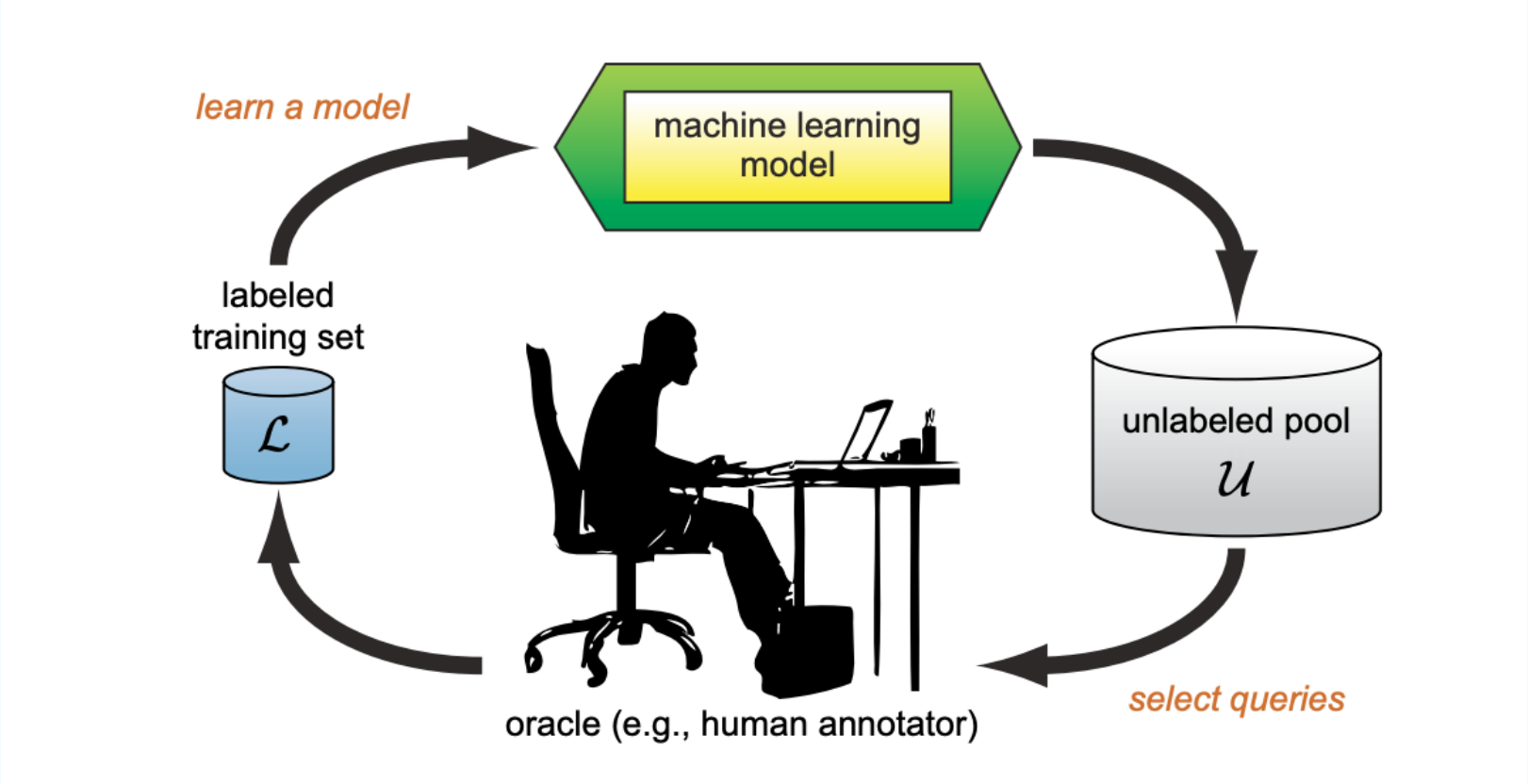
Aleatoric uncertainty is introduced by noise in the data (e.g. sensor data, noise in the measurement process) and it can be input-dependent or input-independent. It is generally considered as irreducible since there is missing information about the ground truth.

Epistemic uncertainty refers to the uncertainty within the model parameters and therefore we do not know whether the model can best explain the data. This type of uncertainty is theoretically reducible given more data

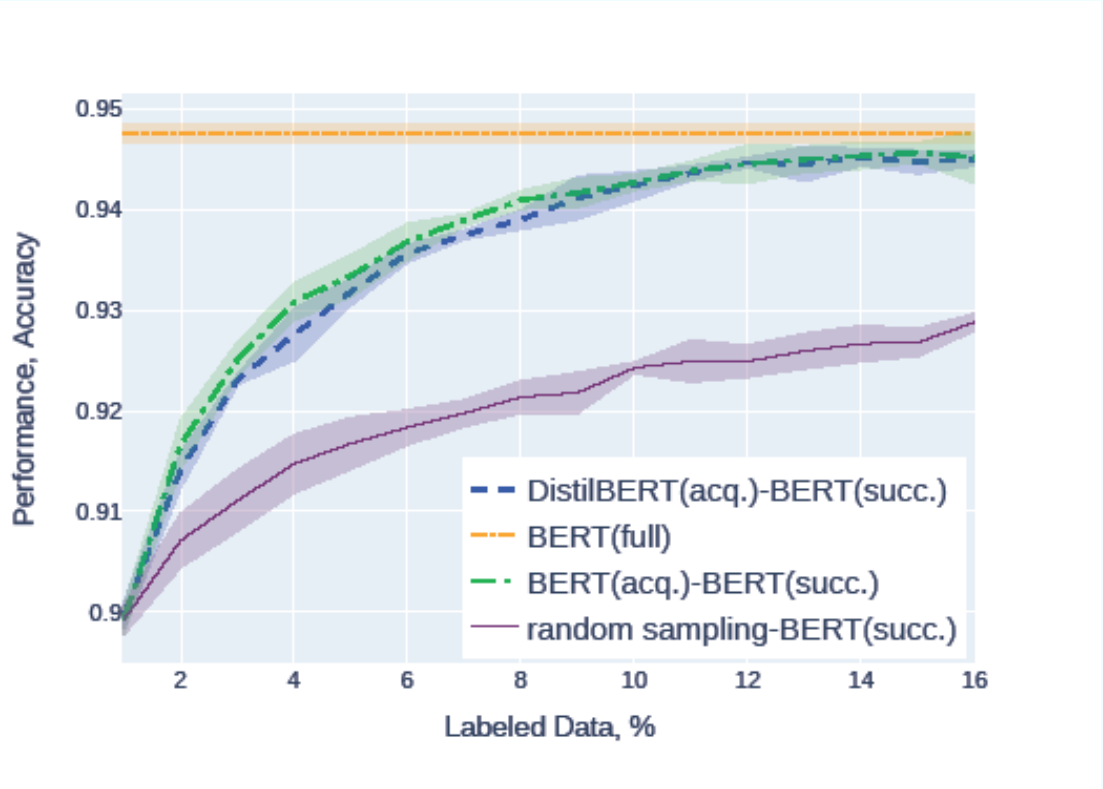
Uncertainty estimation

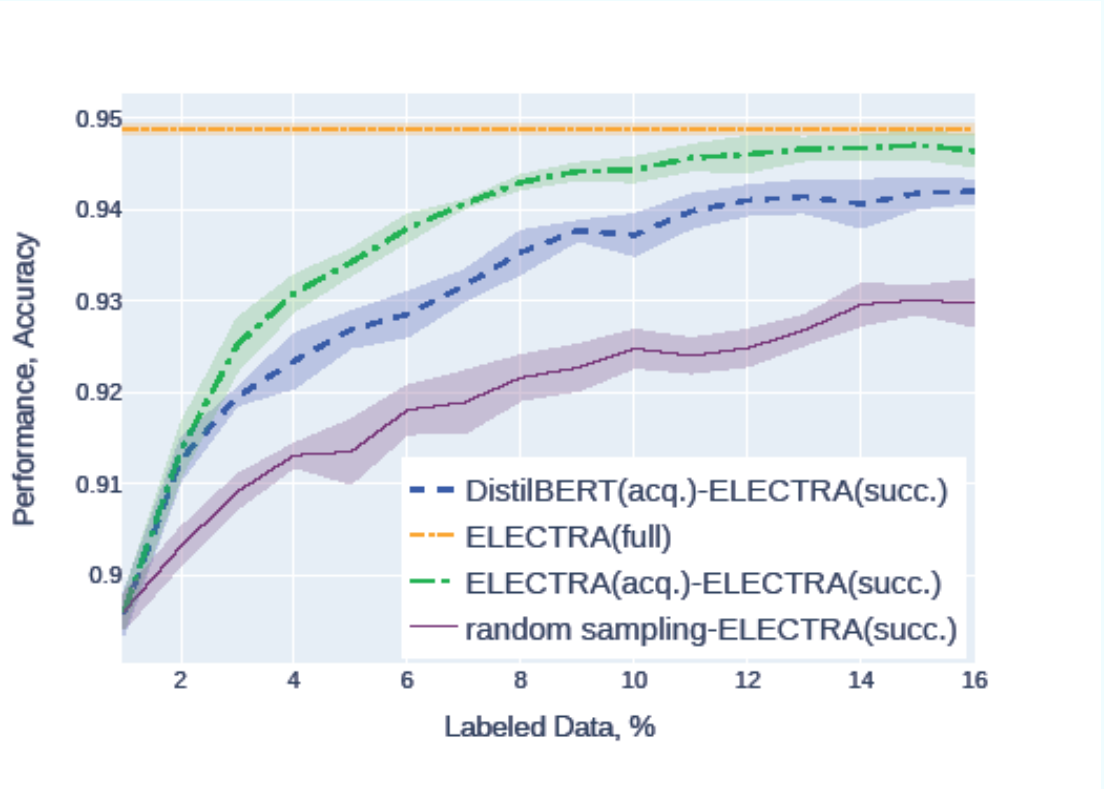
- Least confidence
- Monte-Carlo Dropout
- Deep Ensemble

Active learning



Active learning





Active learning

- Uncertainty
- Diversity
- Hybrid

Summary

- Neural networks can be overparametrized and could be compressed to smaller models without performance decrease
- Active learning can help to reduce cost of data labeling

airi.net