Efficient Inference

Divyam Madaan

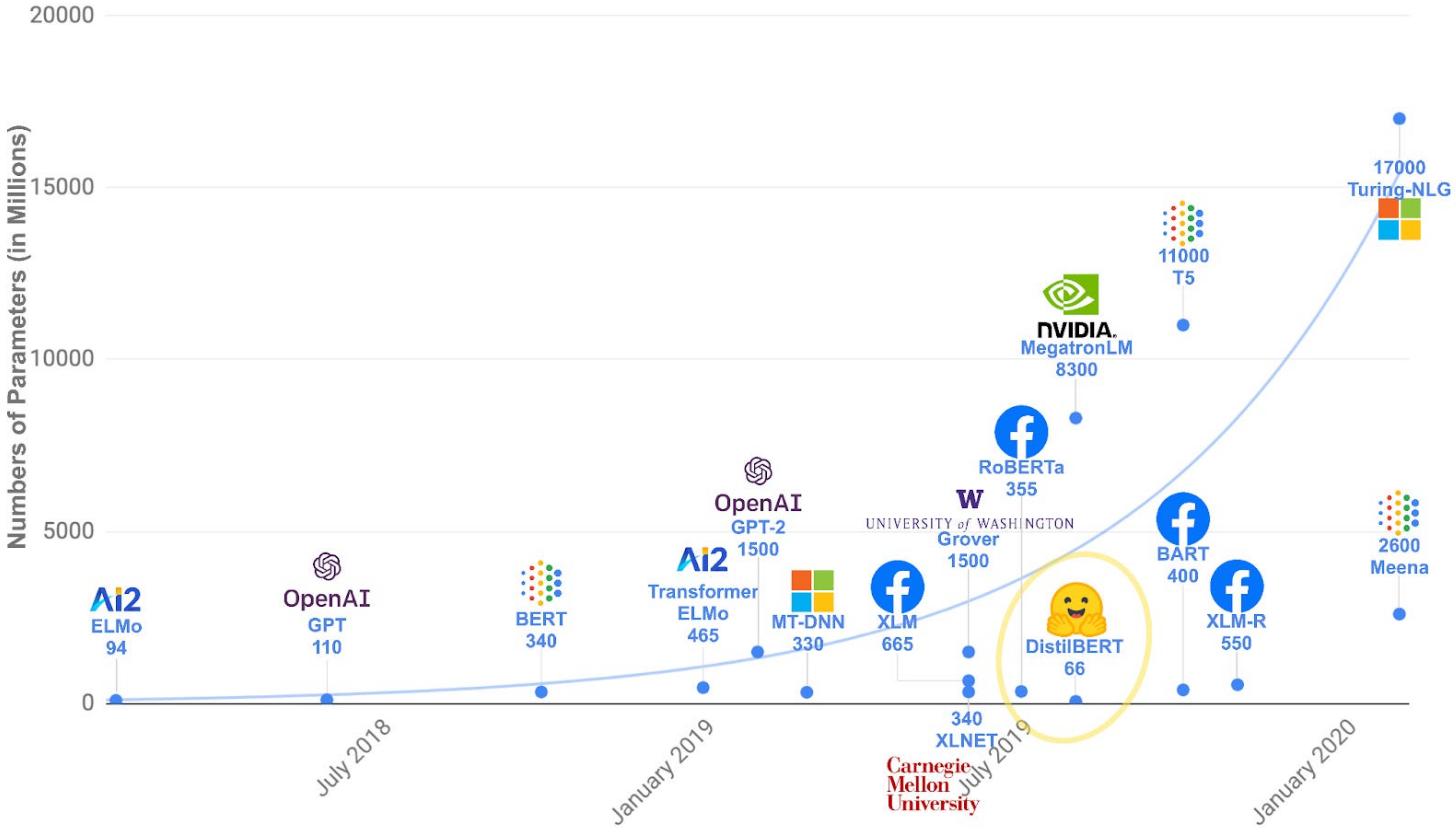
October 17, 2024

Overview of efficient inference

- Efficiency challenge
- Quantization
    - What is quantization and how to quantize?
- Pruning
    - Pruning before, during, after training
- Knowledge distillation
    - Distillation on outputs, weights and features.
- SCP tutorial

Efficiency challenge

Size of models makes ____________________



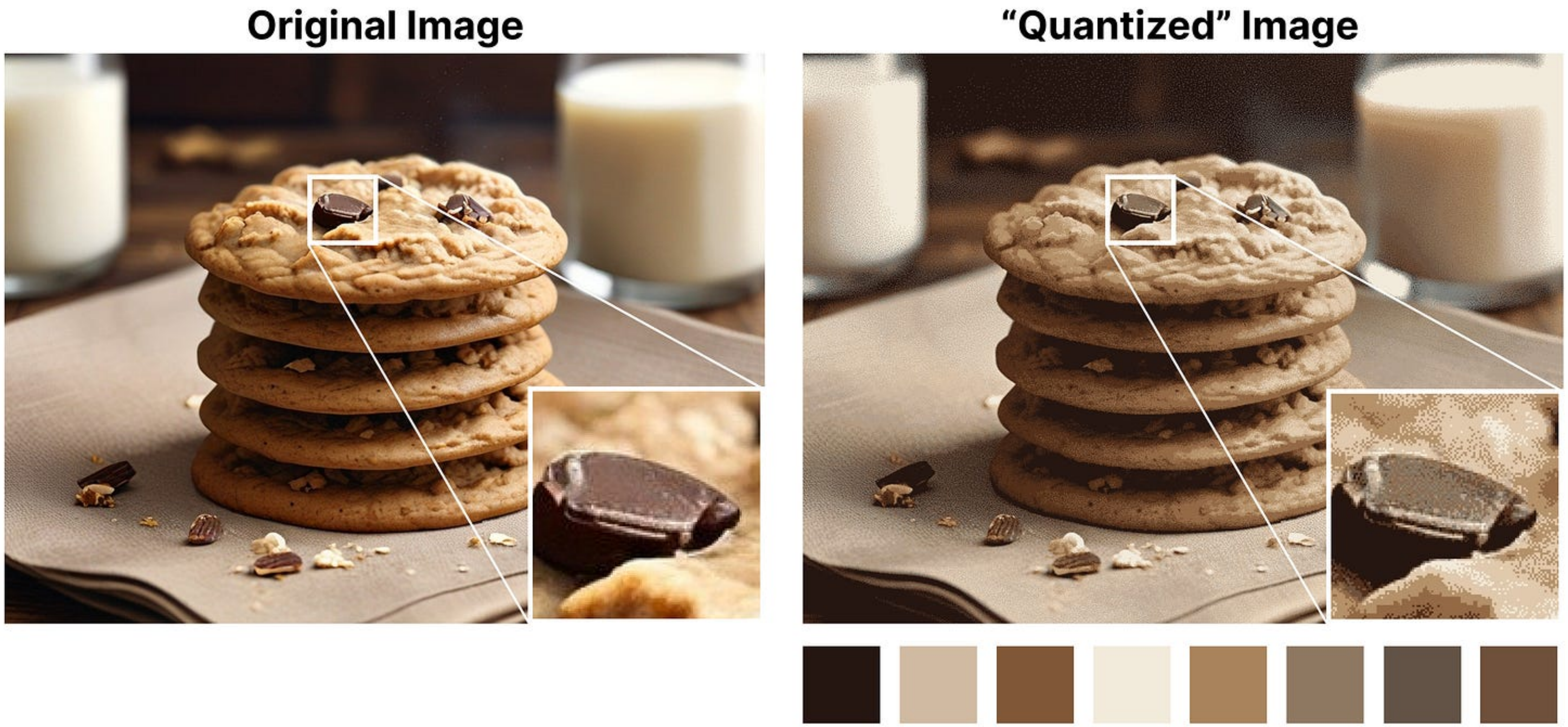
How does quantization look like in real-world?

Goal is to **reduce** the **number of bits (colors)** while **preserving the precision**

Source of images: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization

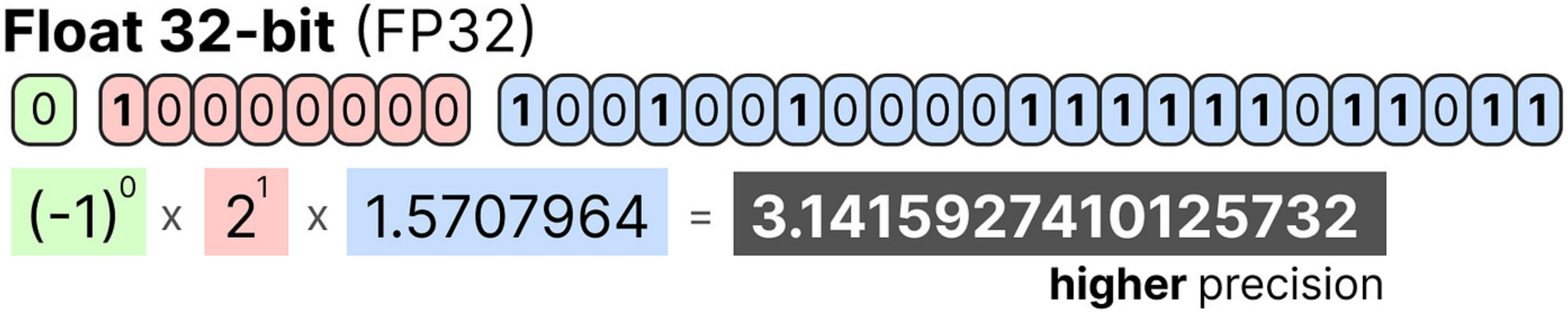
How does quantization look like in real-world?

Reducing the image to use just eight colors leads in a **loss of detail and precision**



Common data types

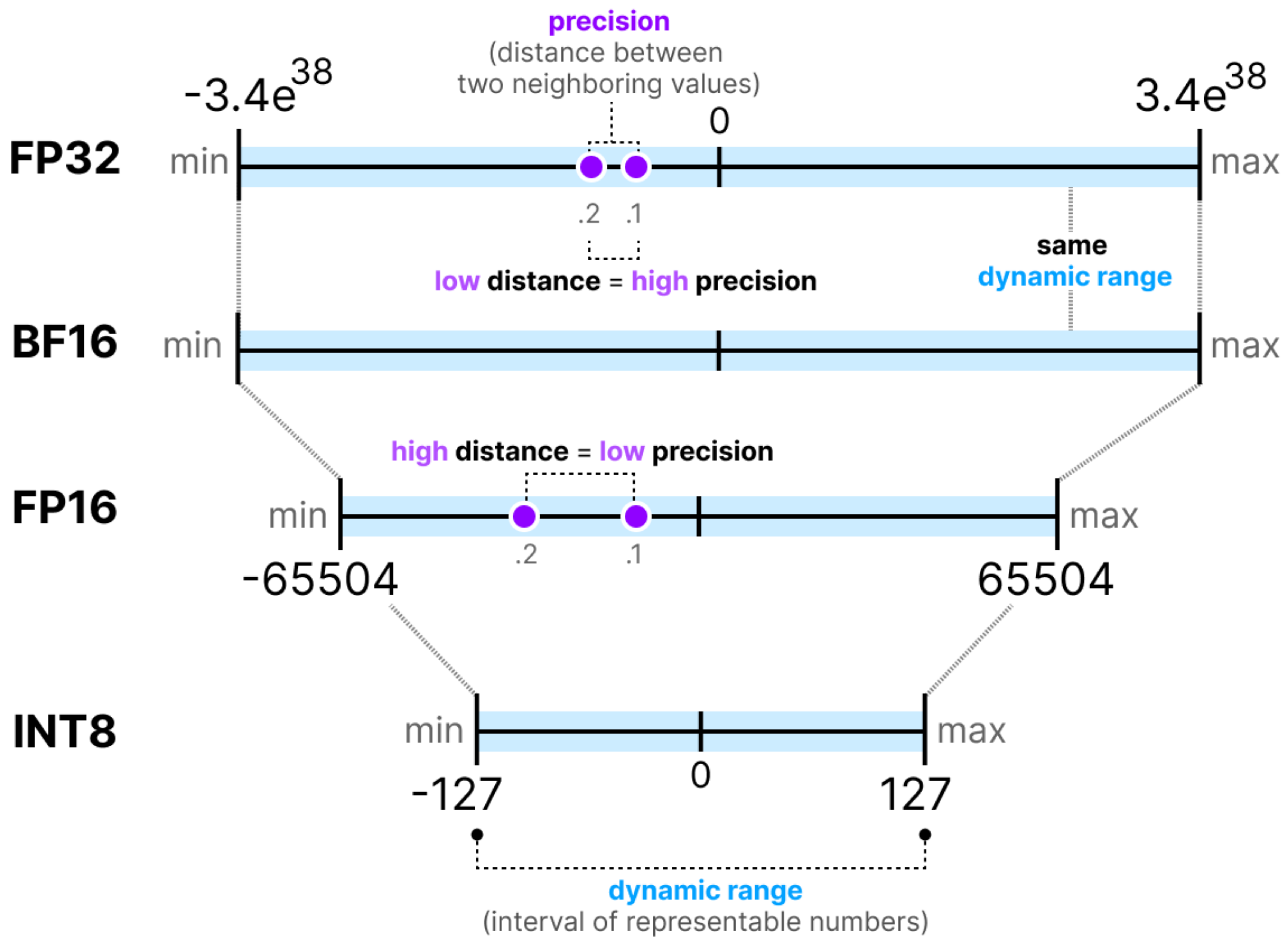
Bits use ________________ to represent a value



$$(-1)^{\text{sign}} \times \text{base}^{\text{exponent}} \times \text{fraction}$$

What is precision?

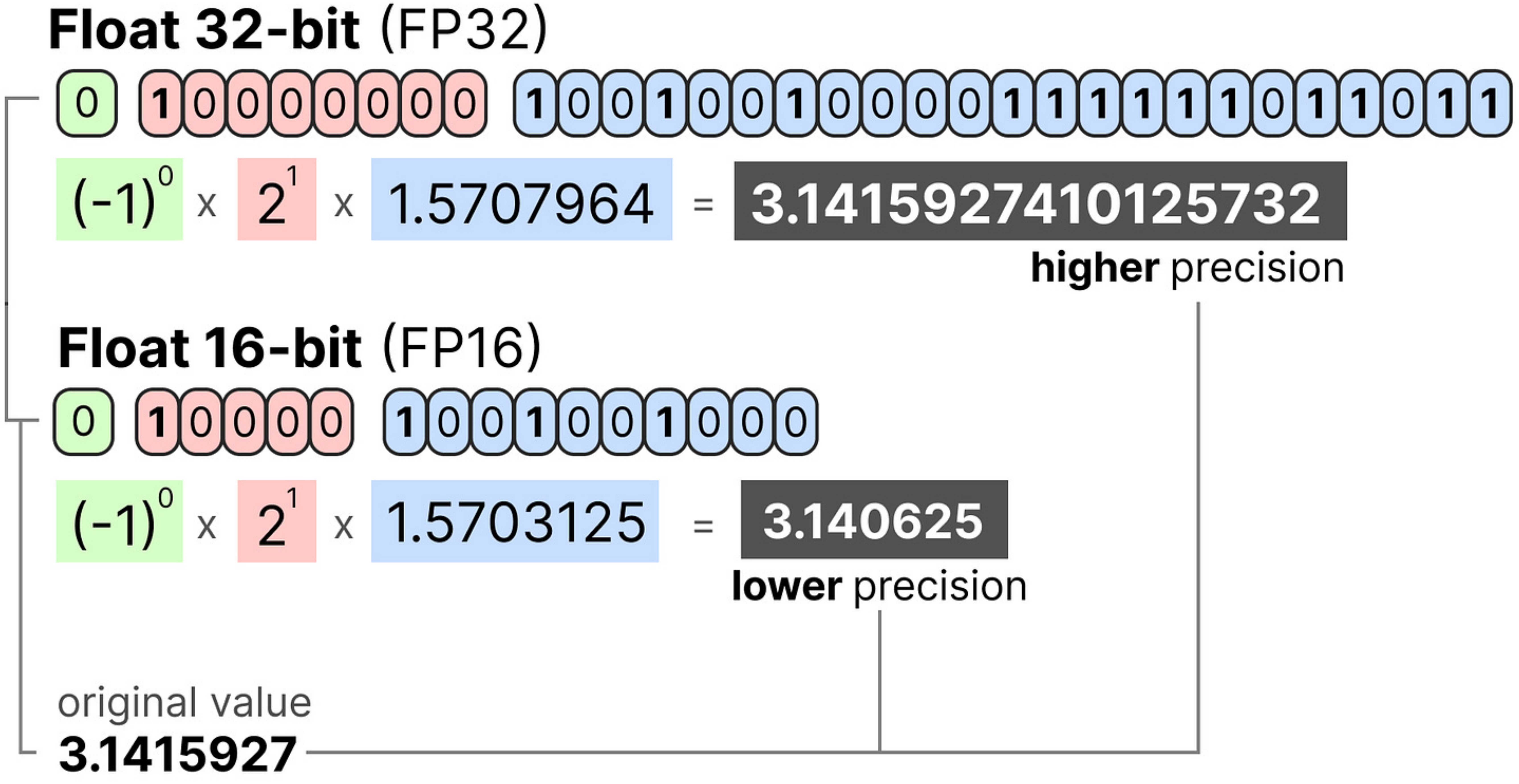
Precision is a measure of how precisely a number can be represented



What is the difference between BF16 and FP16?

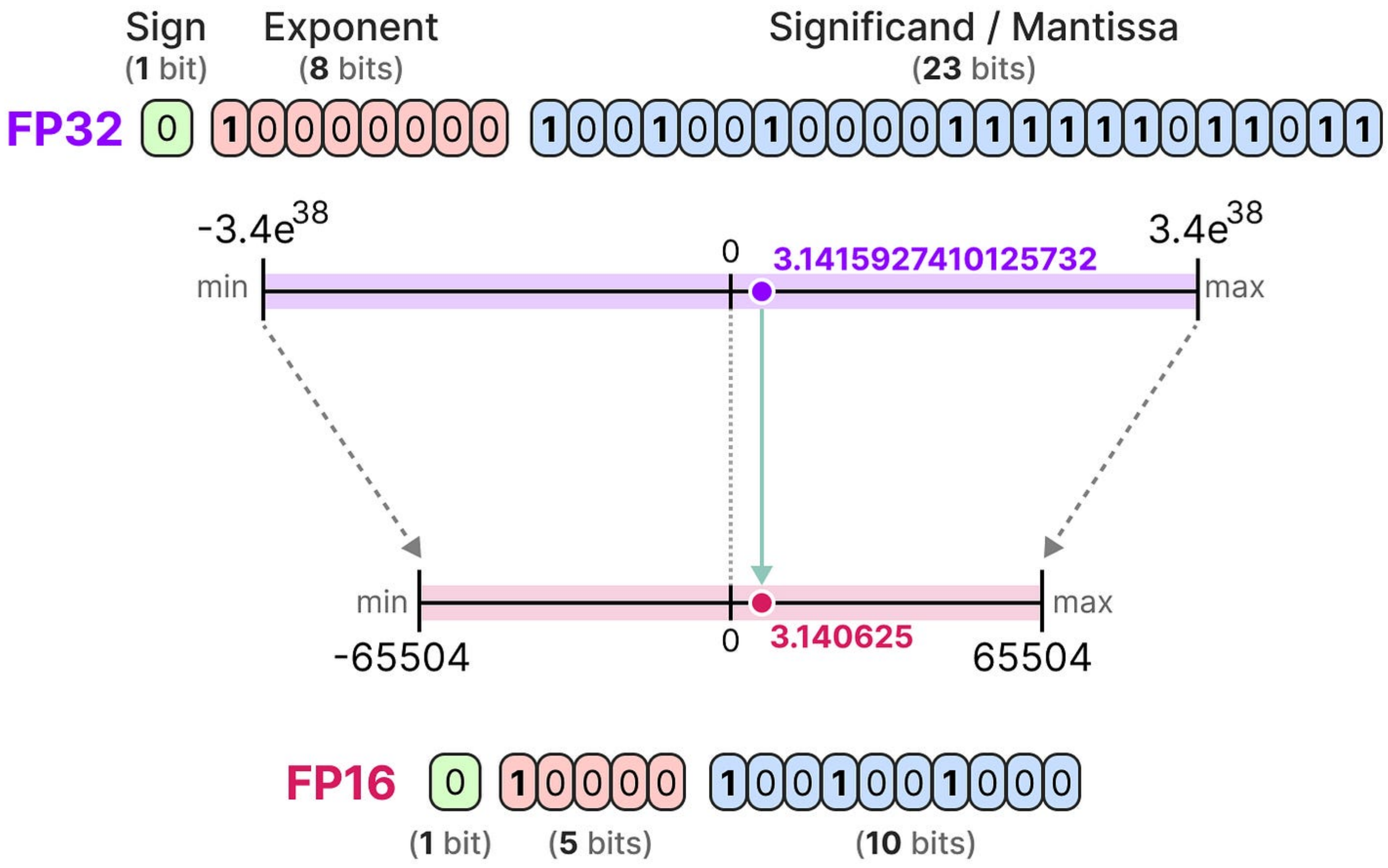
Common data types

Using more bits **increases the precision** of a value



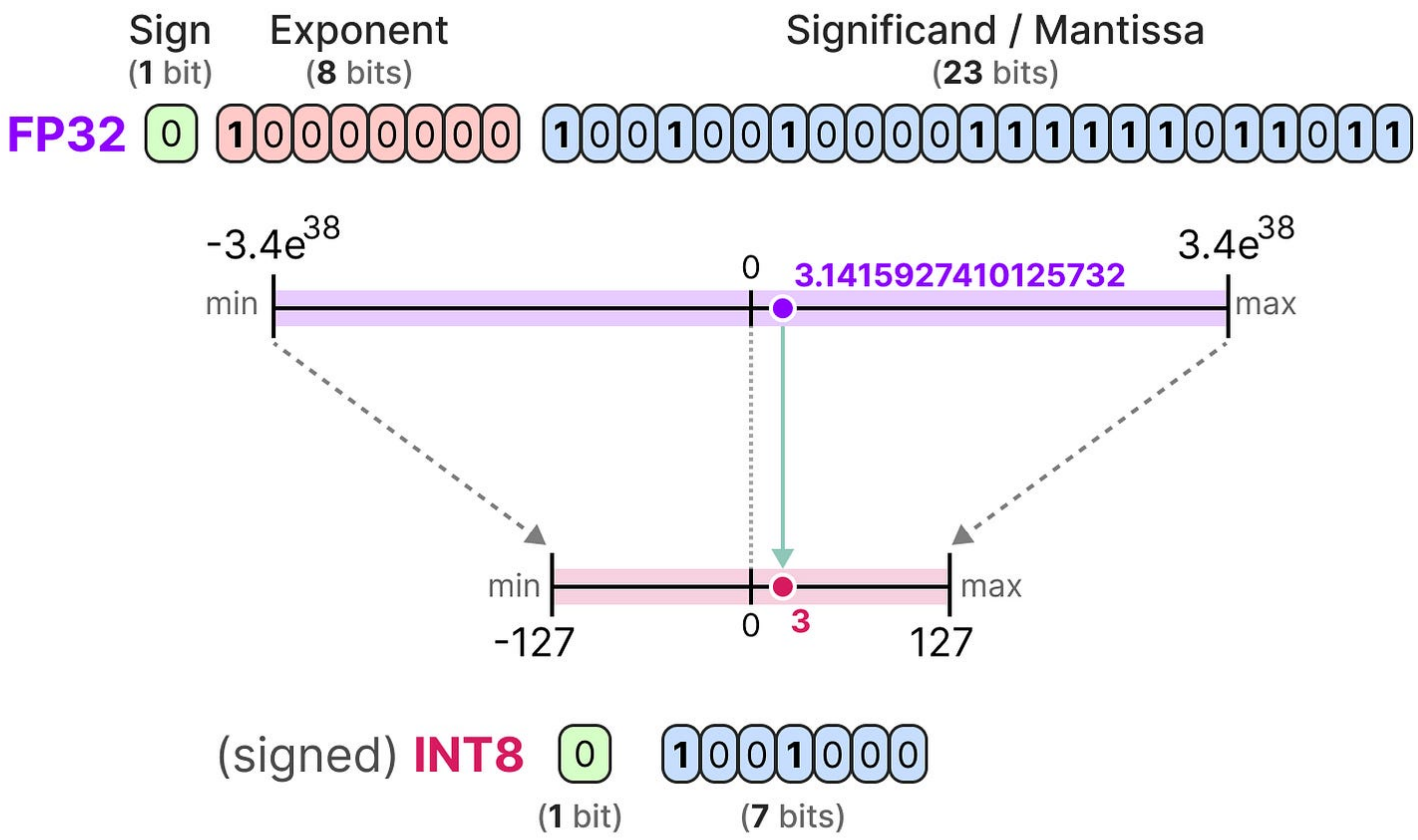
Common data types

The range of values **reduce with quantization**



Common data types

We can convert FP32 to **8-bit integer**-based representations to **save memory**



Absolute maximum (absmax) quantization

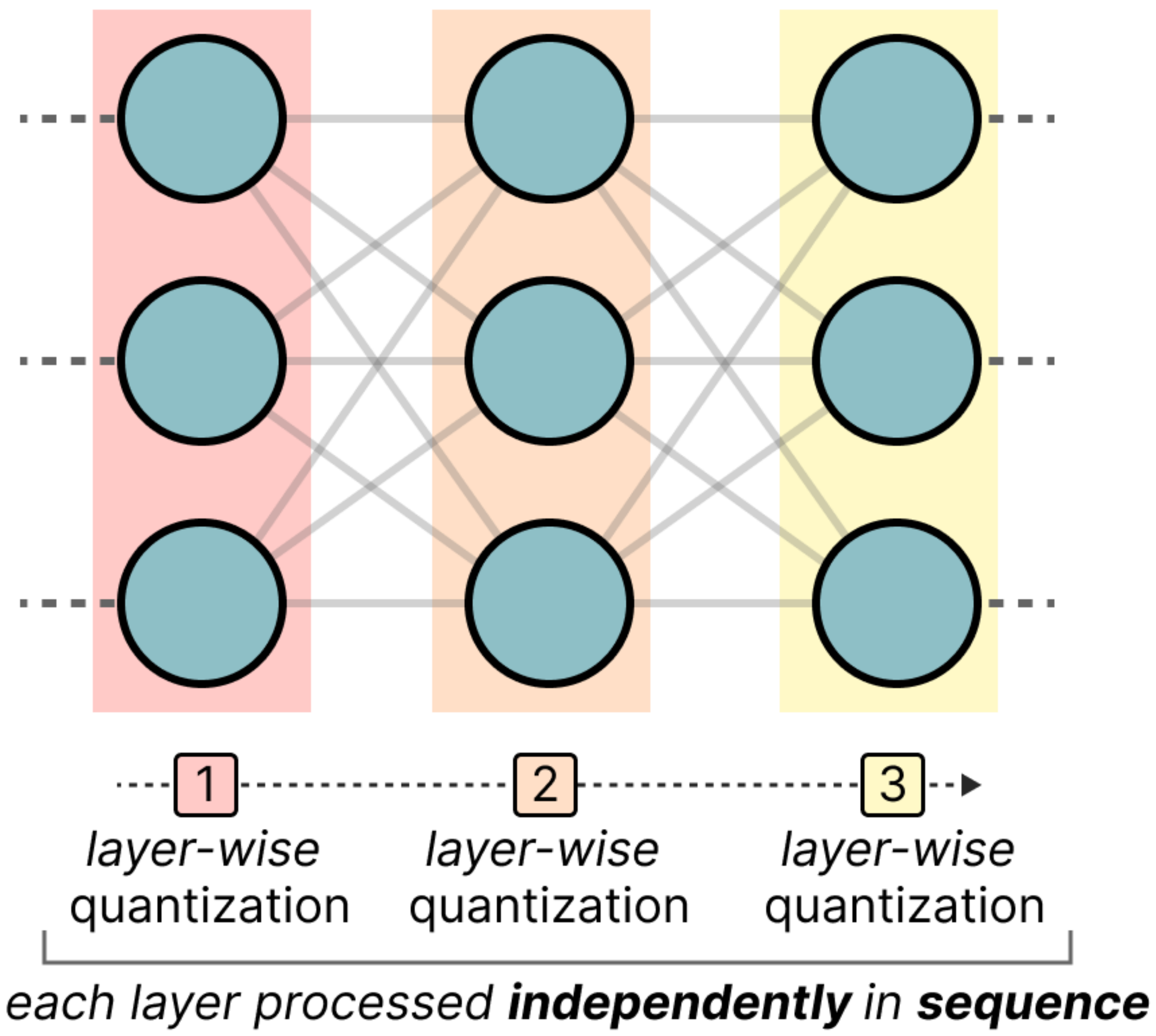
Original number is ________________ to scale it into the range [-127, 127].

$$\mathbf{X}_{\text{quant}} = \text{round}\left(\qquad\qquad \right)$$

$$\mathbf{X}_{\text{dequant}} = \frac{\max|\mathbf{X}|}{127} \cdot \mathbf{X}_{\text{quant}}$$

GPTQ

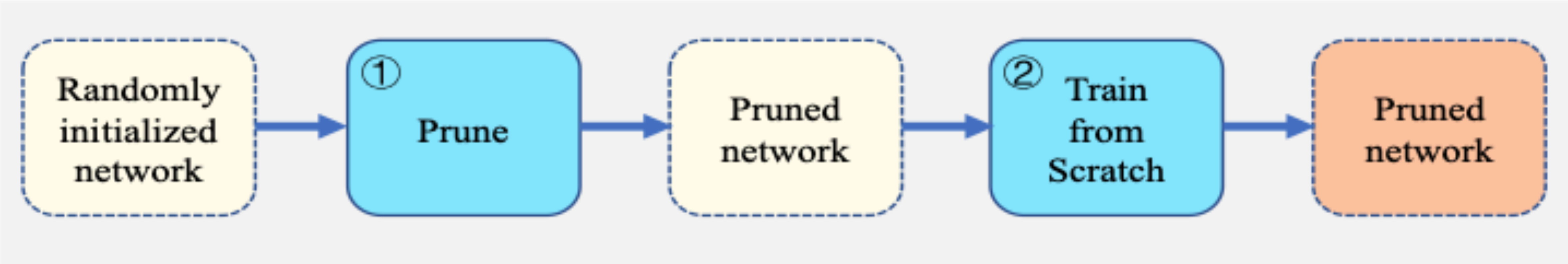
Each layer is **quantized independently**. Given a layer l with a weight matrix W_l, find quantized weights $\widehat{W_l}$



$$\widehat{W_l^*} = \arg\min_{W_l} \| W_l X - \widehat{W_l} X \|^2$$

Pruning before training

Procedure: __________

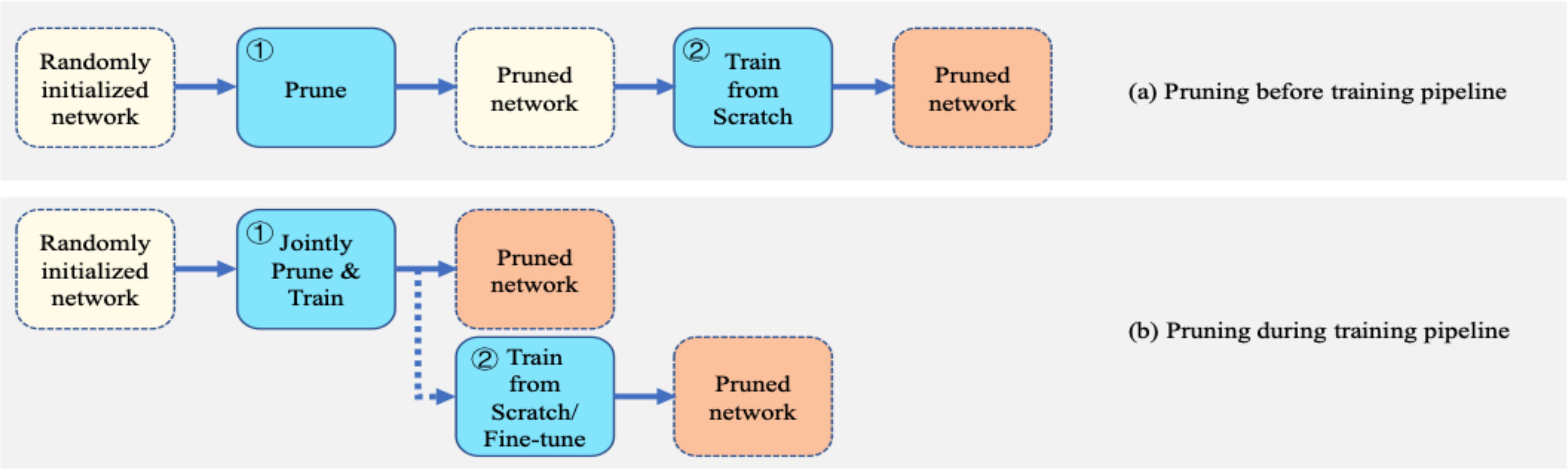


(a) Pruning before training pipeline

$$f(x_t, W_0 \odot M') \rightarrow f(x_t, W_t \odot M')$$

Pruning during training

Procedure: __________

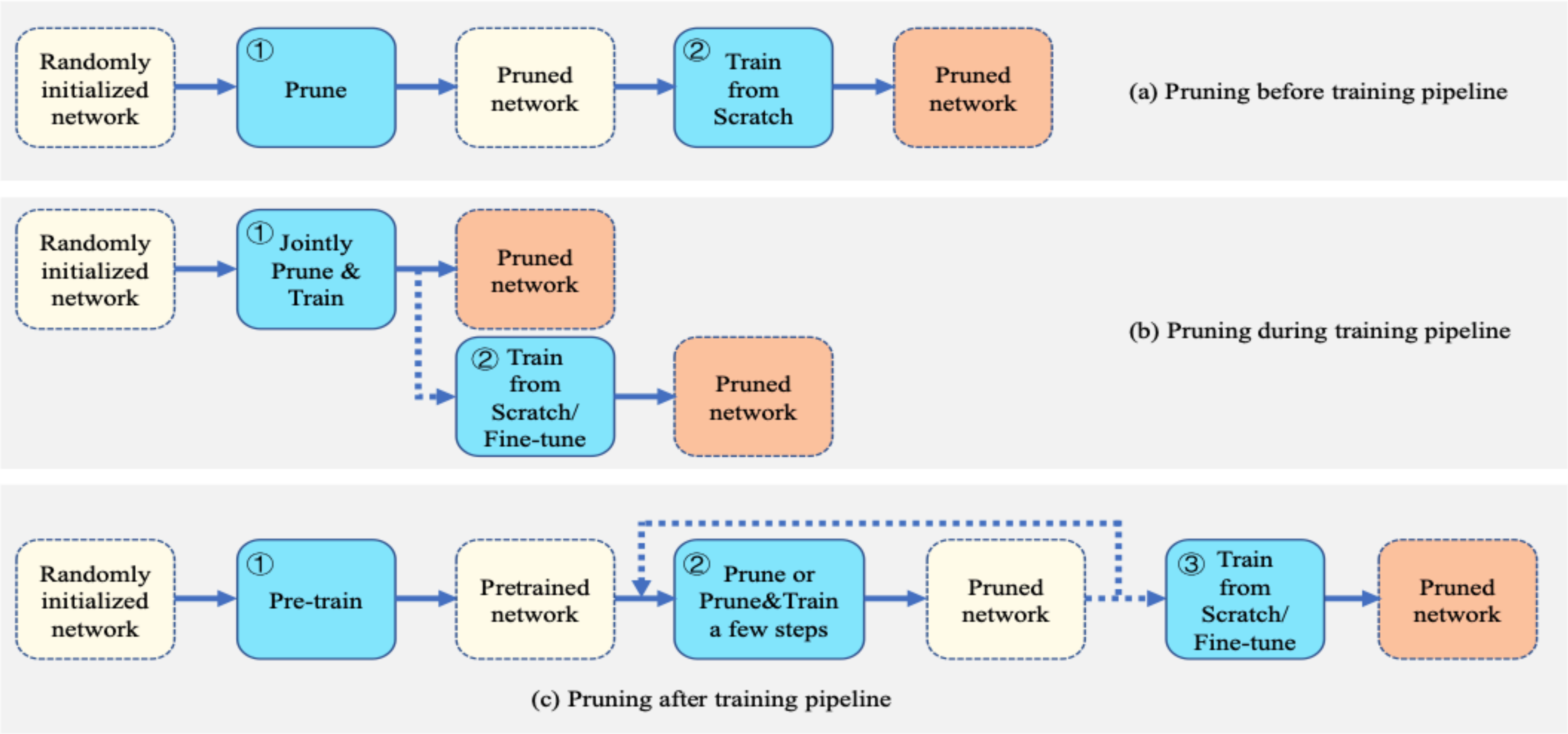


$$f(x_t, W_0 \odot M') \rightarrow f(x_t, W_t \odot M')$$

$$f(x_t, W_0) \rightarrow f(x_t, W_t \odot M_t)$$

Pruning after training

Procedure: __________



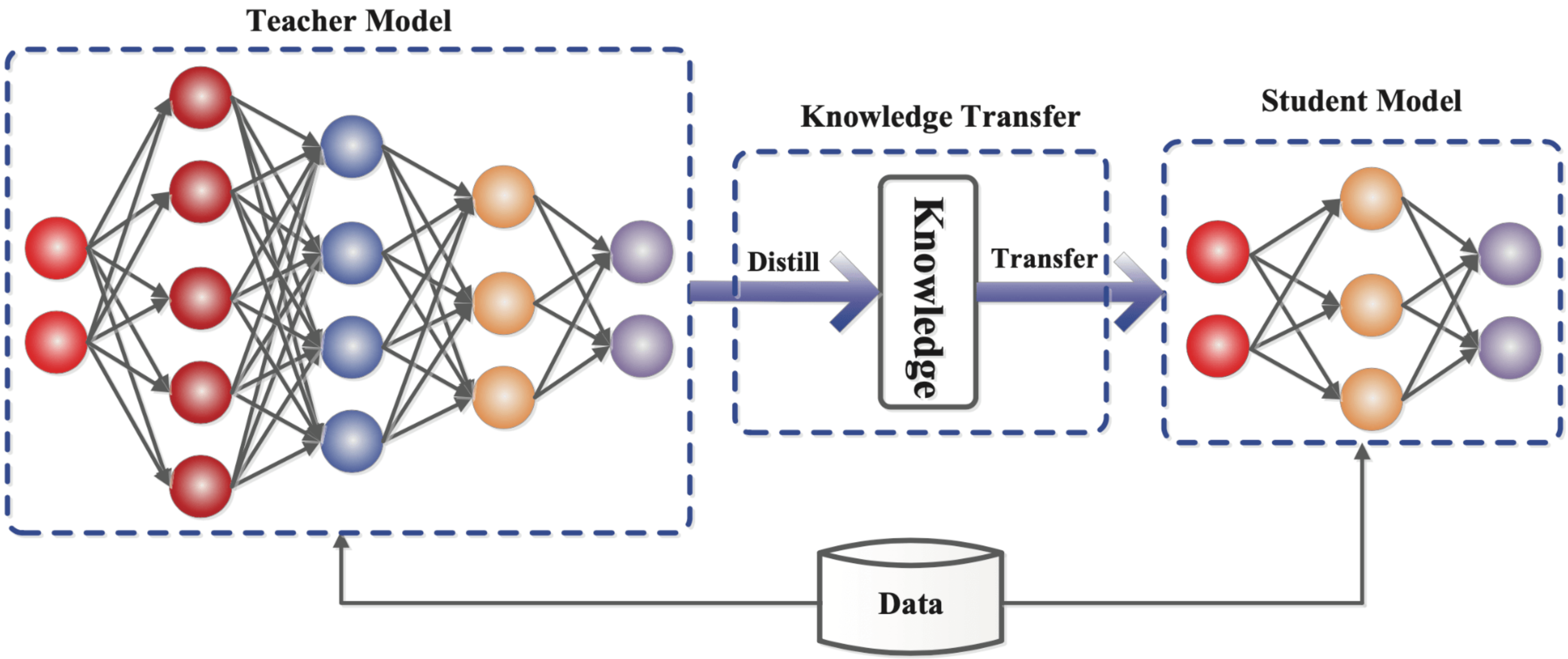
$$f(x_t, W_0 \odot M') \to f(x_t, W_t \odot M')$$

$$f(x_t, W_0) \to f(x_t, W_t \odot M_t)$$

$$f(x_t, W_0) \to f(x_t, W_t) \to f(x_t, W_t' \odot M')$$

Knowledge Distillation

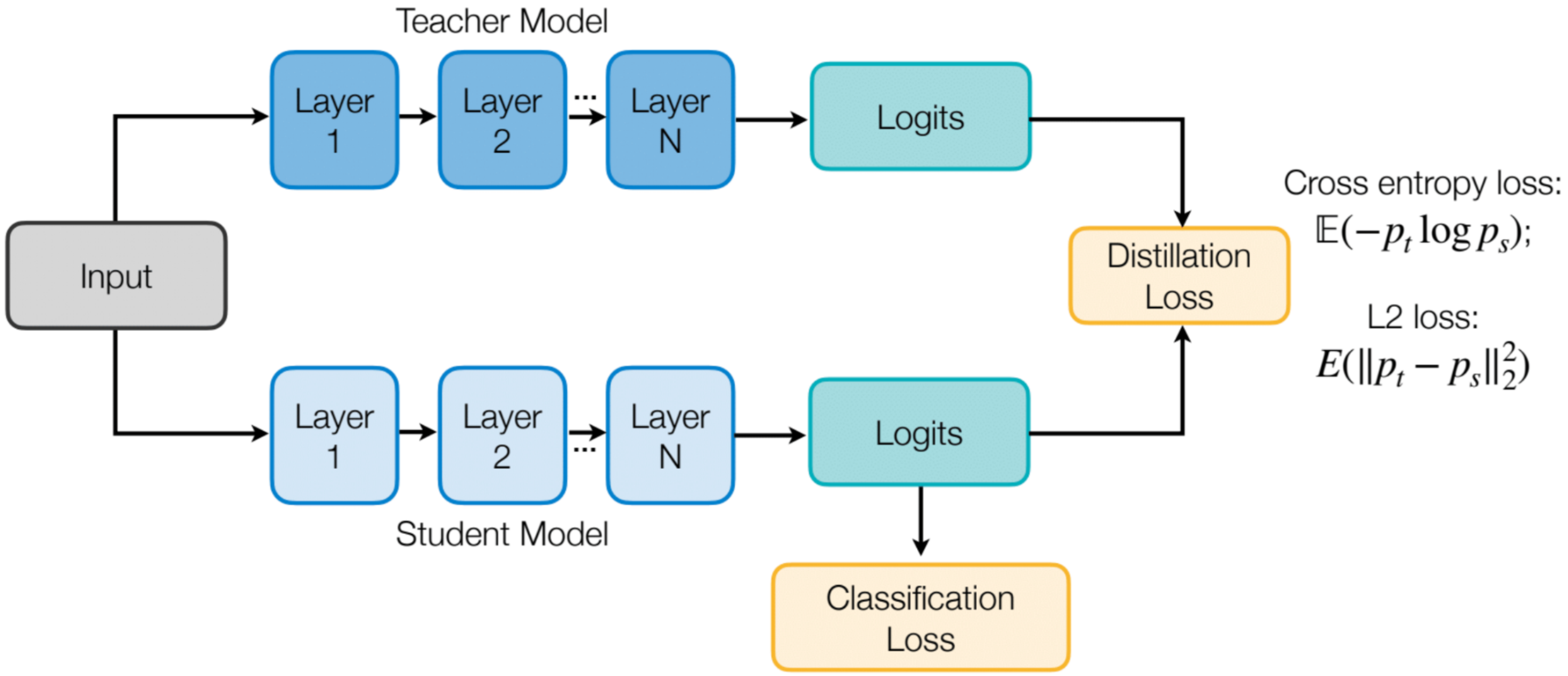
Goal is to **transfer knowledge** from a **larger model** to a **student model**



Why not train the **student model from scratch**?

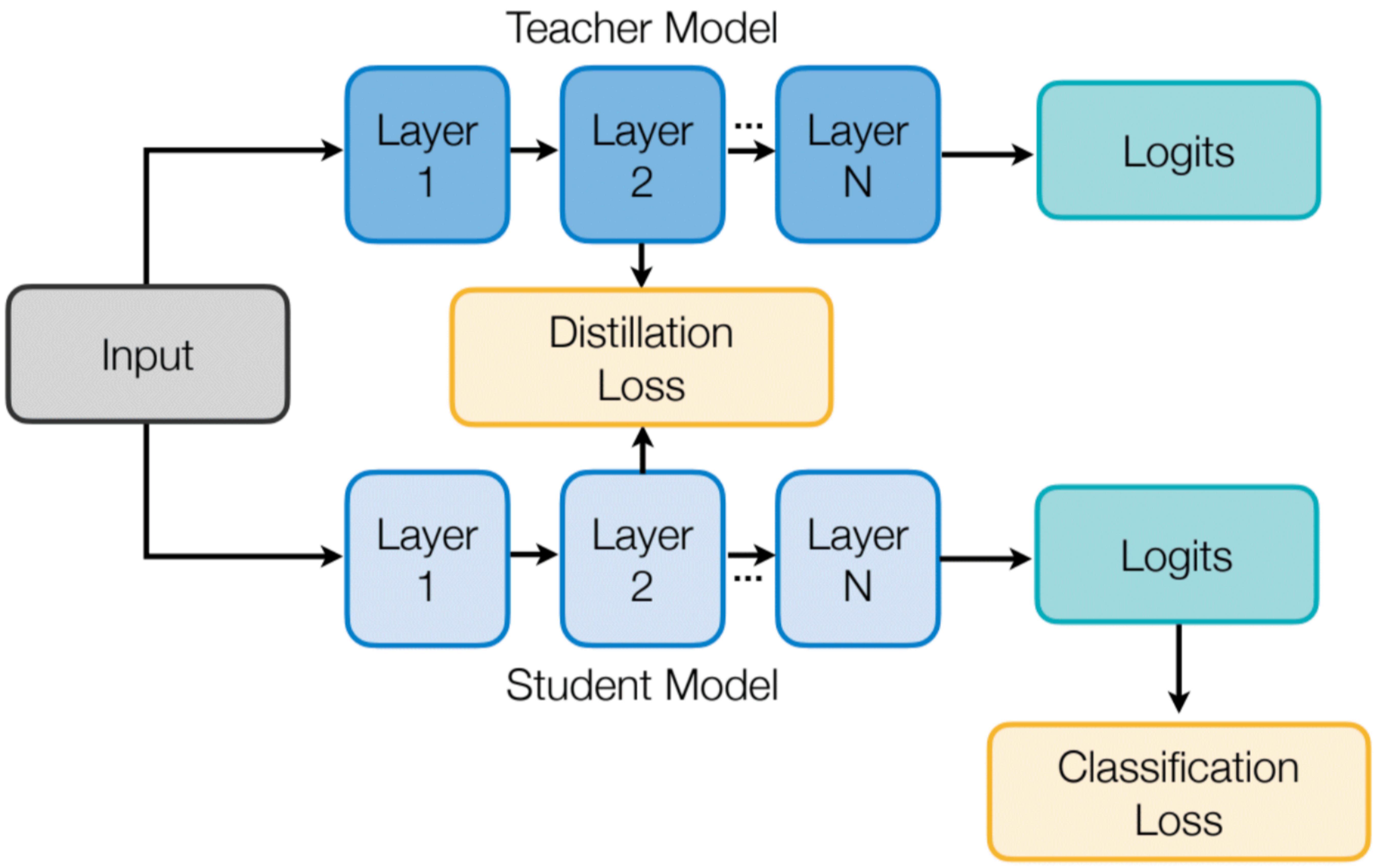
Knowledge Distillation

Simplest way is to **match the outputs** using a distance metric



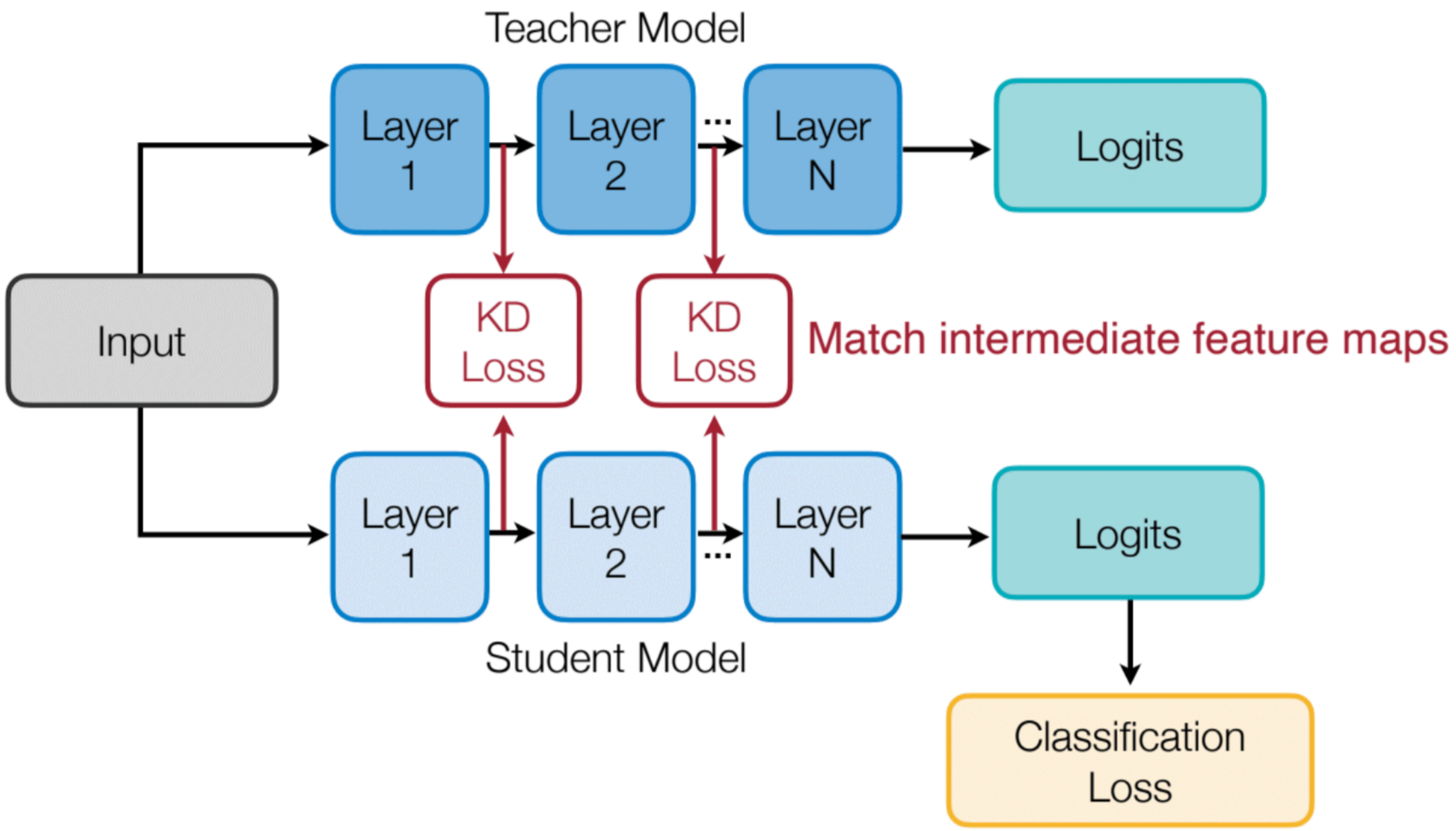
Knowledge Distillation

Match the **weights in intermediate layers** using a distance metric



Knowledge Distillation

Match the **intermediate feature maps** using a distance metric

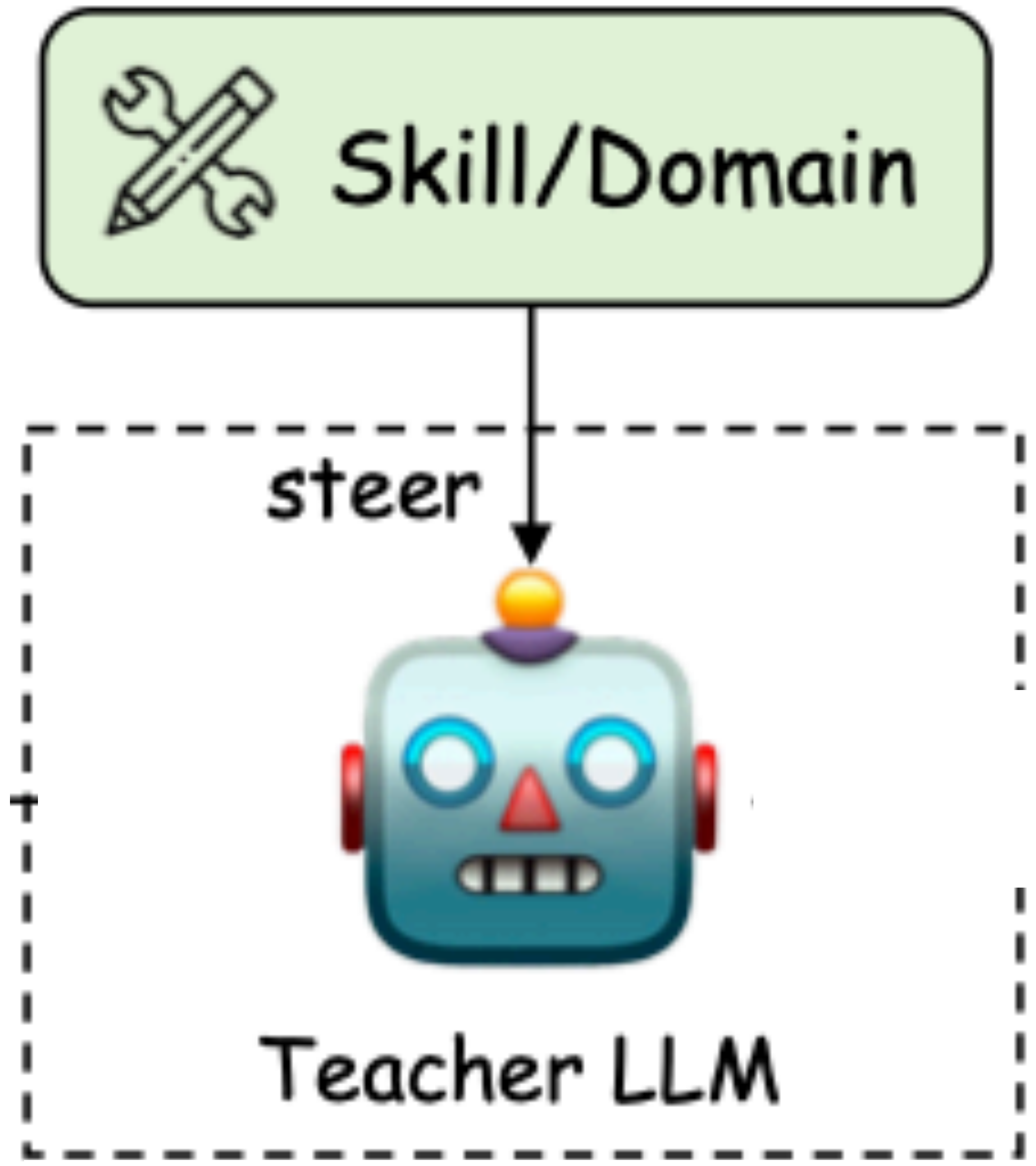


Knowledge Distillation in LLM Era

Due to **inaccessible parameters**, we want to **transfer knowledge** from LLMs

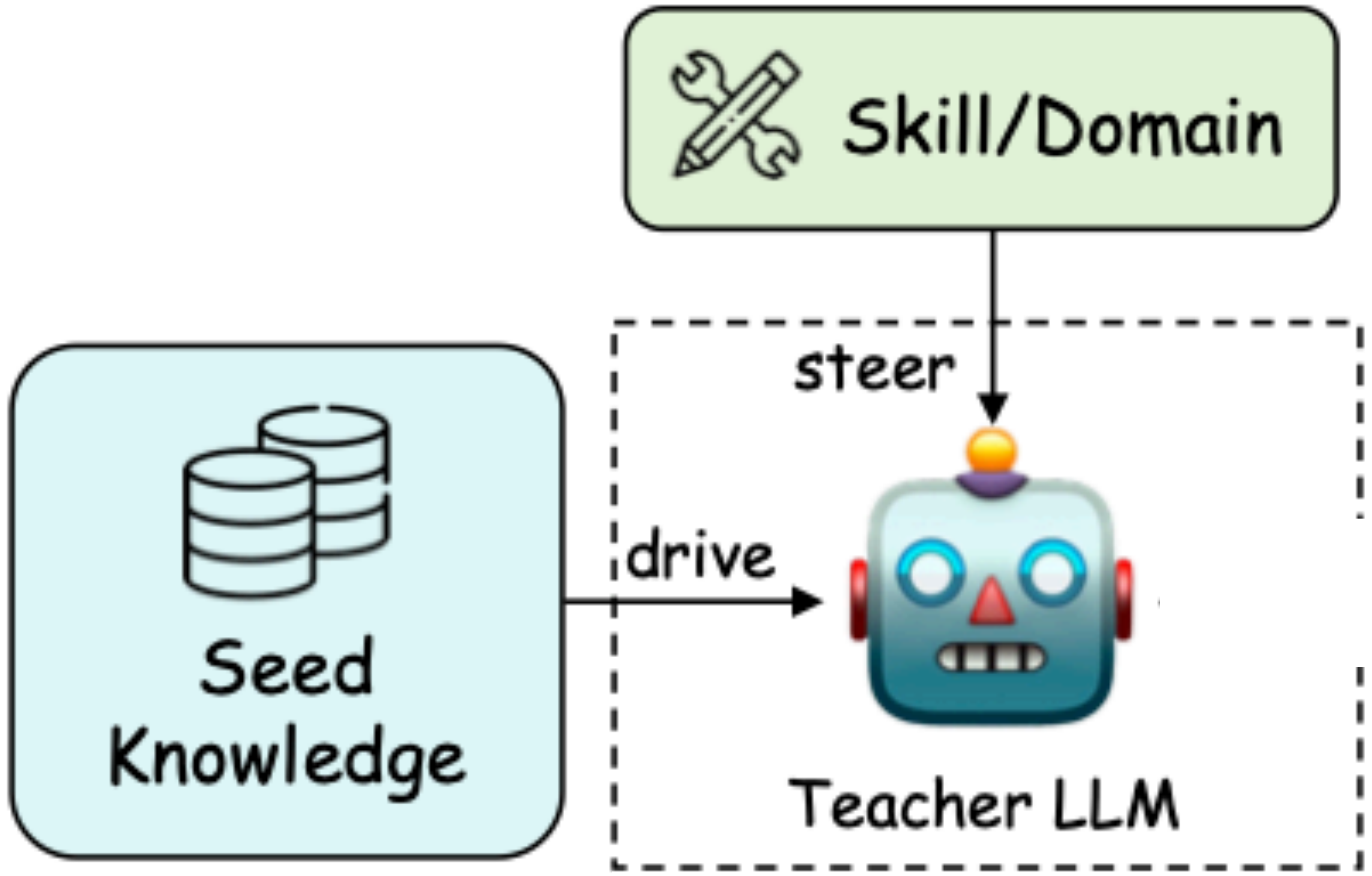
Knowledge Distillation in LLM Era

Steer the LLM for a target skill or domain



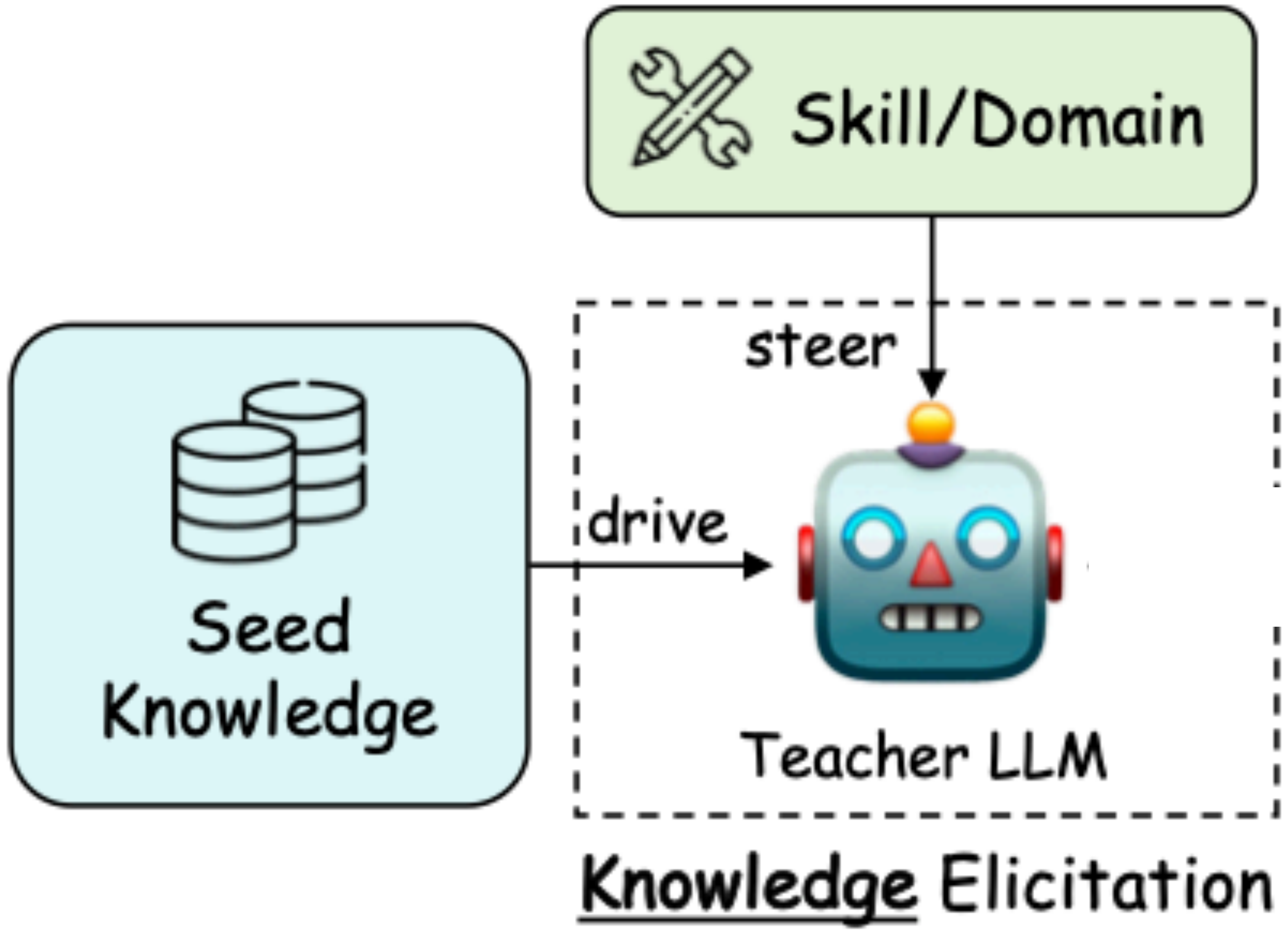
Knowledge Distillation in LLM Era

Feed the LLM with a small data as **seed knowledge**



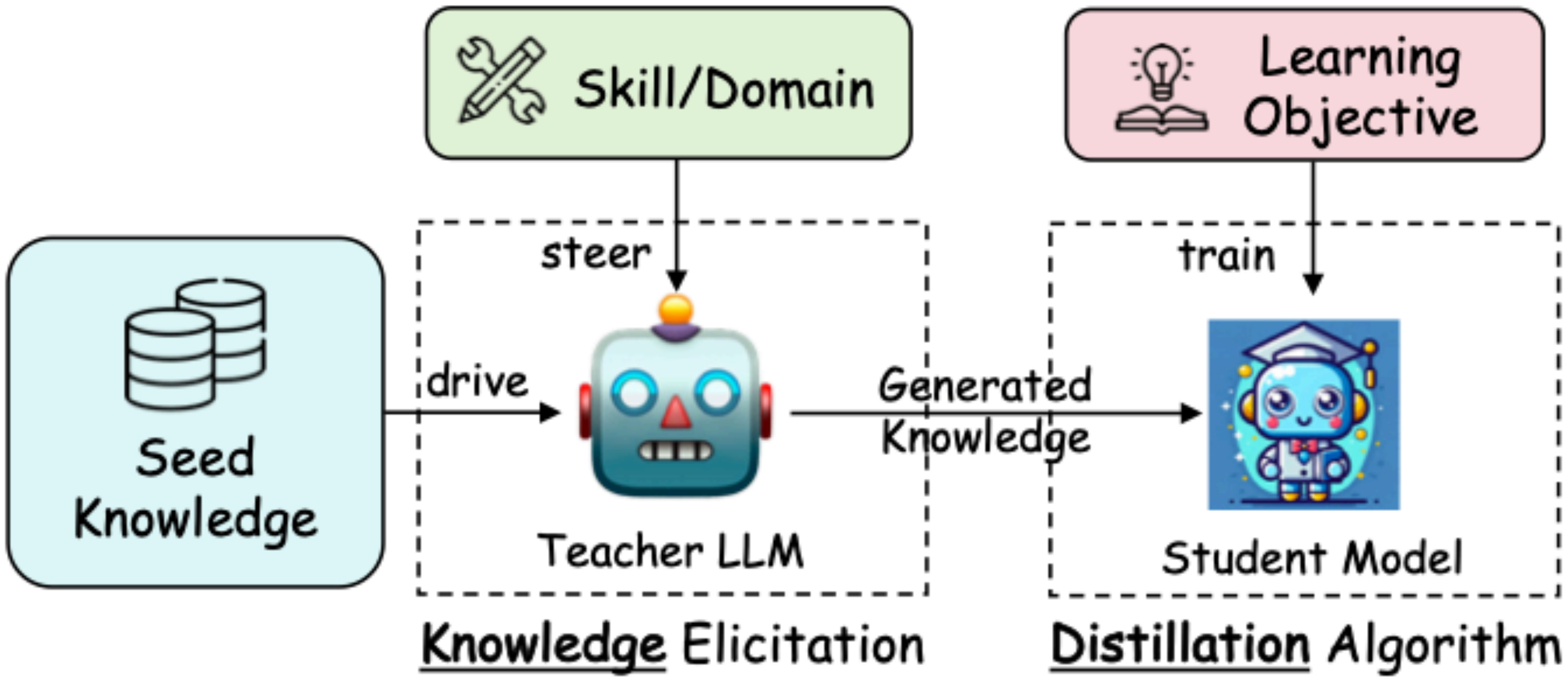
Knowledge Distillation in LLM Era

Feed the LLM with a small data as **seed knowledge**



Knowledge Distillation in LLM Era

Generate knowledge from the teacher LLM and **emulate teacher skills**



Takeaways