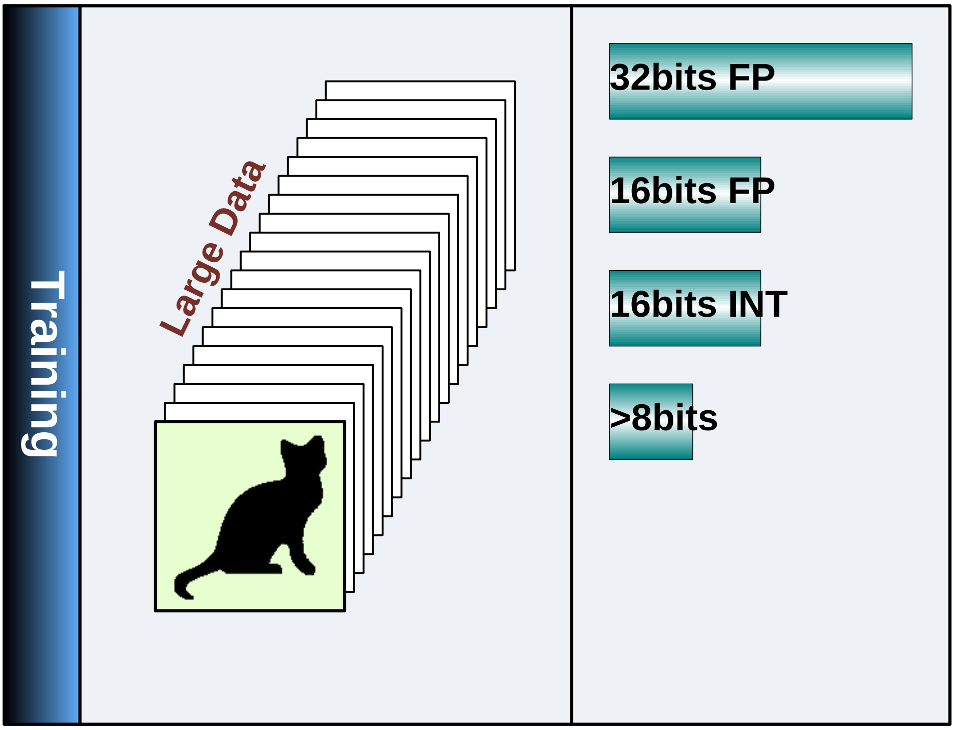
Quantization



Training

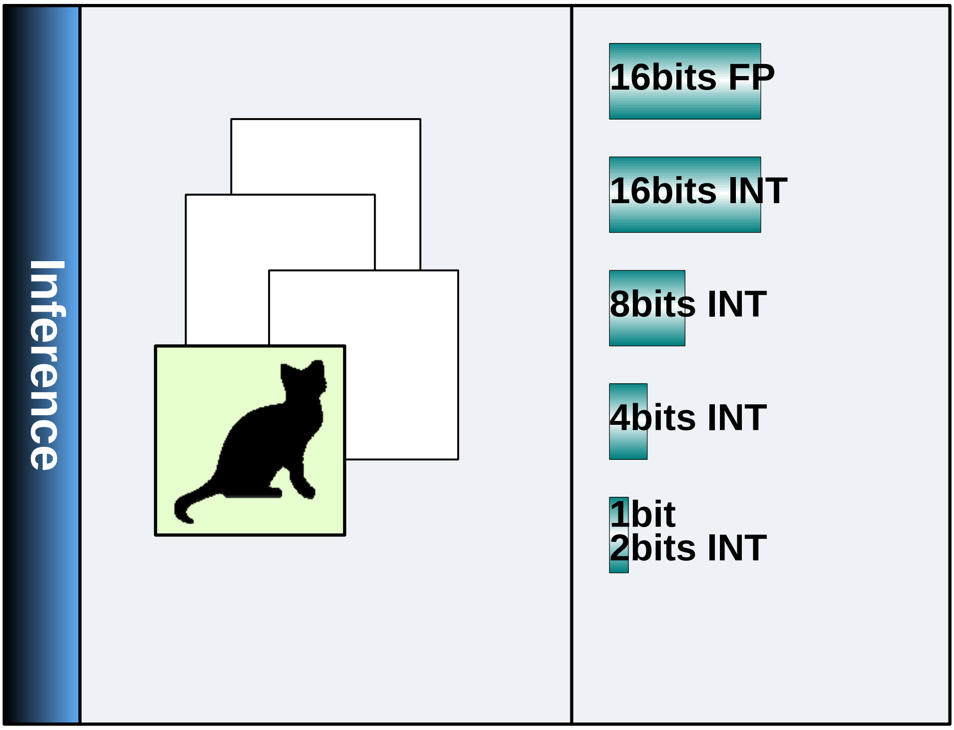
Large Data

32bits FP

16bits FP

16bits INT

>8bits



Inference

16bits FP

16bits INT

8bits INT

4bits INT

1bit
2bits INT