Some Example Questions

Precision:

Will very-low precision training (1-4 bit weights, 1-4 bit activations) work in general across all problems we care about?

Sparsity and embeddings: How should we handle:

Dynamic routing like the sparsely-gated Mixture of Experts work (ICLR'17)

Very large embeddings for some problems (e.g. 1B items x 1000D)

Batch size:

Should we build machines for very large batch sizes? Or batch size 1?

Training algorithms:

Will SGD-like algorithms remain the dominant training paradigm?

Or will large-batch second-order methods like K-FAC be better?