I would say, Deep learning models are notable for requiring enormous amounts of training data to reach state-of-the-art performance. For example, the ImageNet Large Scale Visual Recognition Challenge on which teams challenge their image recognition models, contains 1.2 million training images hand-labeled with 1000 object categories. Without large scale training data, deep learning models will not converge on their optimal settings and will not perform well on complex tasks such as speech recognition or machine translation. This data requirement only grows when a single neural network is used to solve a problem end-to-end; that is, taking raw audio recordings of speech as the input and outputting text transcriptions of the speech. This is in contrast to using multiple networks each providing intermediate representations (e.g. raw speech audio input → phonemes → words → text transcript output; or raw pixels from a camera mapped directly to steering commands). If we want AI systems to solve tasks where training data is particularly challenging, costly, sensitive, or time-consuming to procure, it’s important to develop models that can learn optimal solutions from less examples (i.e. one or zero-shot learning). When training on small data sets, challenges include overfitting, difficulties in handling outliers, differences in the data distribution between training and test. An alternative approach is to improve learning of a new task by transferring knowledge a machine learning model acquired from a previous task using processes collectively referred to as transfer learning.

A related problem is building smaller deep learning architectures with state-of-the-art performance using a similar number or significantly less parameters. Advantages would include more efficient distributed training because data needs to be communicated between servers, less bandwidth to export a new model from the cloud to an edge device, and improved feasibility in deploying to hardware with limited memory.

**Applications:** Training shallow networks by learning to mimic the performance of deep networks originally trained on large, labeled training data; architectures with fewer parameters but equivalent performance to deep models (e.g. Squeeze Net); machine translation.

**Companies:** Geometric Intelligence/Uber, DeepScale.ai, Microsoft Research, Curious AI Company, Google, Bloomsbury AI.