Break out session

How can ML help databases

1) Auto-admin: select indexes and other.
2) Query optimization: e.g., Jenny's talk
3) Cardinality estimation: can you train the model offline, how to deal with continuous learning because data changes, how do we debug on the customer side?
4) Debuggability is hard and goes beyond databases.
5) What happens to query consistency: Models can cause inconsistency in query execution times. Models can make query execution time runtimes even less consistent.
6) Maybe models should not be super black boxes.
7) Maybe instead of ML alone, then ML that produces heuristics that we can then use consistently and predictably
8) Related to Andy Pavlo's Peloton's work.
9) Today, there are too many knobs. Can the ML make all those knobs go away?
10) Do we need new DBAs that know how to work with ML or maybe get rid of DBAs all together
11) ML for cardinality, optimization, tuning, indexing
12) ML also for logging (when to log and what to log?), data replication, and data garbage collection, and garbage collection of indexes and materialized views. In the cloud world, use ML for capacity planning and to scale up the resources as demand increases.
13) Reinforcement learning for federated query optimization
14) ML includes: knowledge basis, expert systems, etc. Not only deep learning.
15) Orchestrating federated queries could also use ML. Any situation where something is a black box is a good context to apply ML

Can we help ML with AQP

1. ML people always assume that data is in a nice table. But data is in many tables. So it is very slow to manage the data when doing ML.
2. What can we express in SQL? Can we express more ML in SQL? Recursive queries for example are very slow. The full support isn't there a lot of the time. How can we transform ML algorithms into something that will run efficiently?
3. Often data scientists download a sample of the data onto their laptop because it is easier to do.
4. DS want to use Python. The more seamless the Python API, the better. But then go back and forth. Can you optimize entire workbooks? Need to push Python computation into the database as a user-defined computation.
5. Because viz are slow, can you start fuzzy and refine it as more data arrives from the database. Maybe approximate answer is OK? Especially if the visualization is interactive.

Data viz is a prime candidate for approximate query processing because there is only so much that one can display on the screen. This is like online aggregation.

How can DB help ML

1. ML produces a model. A model is essentially a generative model of data. A deep network that recognizes images learns about cats, dogs, elephants, etc. But ML people use these models in very limited ways: Use model to classify image or use model to generate an image. But what if we think of a model as a database that we can query? List all animals that you know about ordered by height. Basically summarize data with model and query that model. If there is a querying capability, we can also use it to explain what the model is doing.
2. There is a lot of work out there on visualizing models. “Explanations of ML”
3. What is a model? Is it the code? Is it the data?