Section 1 Model Complexity (10mins + 20mins)

a. Given the default settings in the Tensorflow playground, attendees are encouraged to simplify the following existing network structure (4x2) by reducing the number of hidden layers and number of neurons to achieve a good accuracy (test loss).

Suggested solution: 1 hidden layer with 3 neurons (3x0), while 2x0 is not enough for fitting.

b. Change the dataset from circle data to spiral data which is with a more complicated distribution. Attendees are encouraged to increase the model complexity to fit the data.
Suggested answers: 6x6x4 can be with a semi-good fit, 6x6x6x4 can be good, can be with other nice answers. Another possibility: 6x6x6x8x4x4
Section 2 Activation Function (15 mins)

With the existing spiral data and network setting, attendees are encouraged to try different activation functions, including Tanh (default), ReLU, Sigmoid and Linear and study the differences in classification accuracy and convergence time.
Note: ReLU works better and has a fast convergence speed, where Sigmoid is hard to converge even with higher learning rates. Sigmoid is not for deep neural networks. Tanh is semi-good and Linear is not working since the data is not linear separable (combination of linear is linear). Results may vary due to the randomness in optimization.

Section 3 Learning Rate and Batch Size (15 mins + 10 mins)

a. With the existing spiral data and network setting with ReLU activation function, attendees are encouraged to try different learning rates, including 0.001, 0.01, 0.03, 0.1, 0.3, 1, etc.
Note: A higher learning rate may result in a faster convergence but may also lead to a non-convergence situation.

b. With the existing spiral data and network setting with ReLU activation function, attendees are encouraged to try different batch sizes, including 1, 5, 10, 20, 40, 80, etc.
Note: Stochastic, mini-batch and batch are three types of optimization schemes with respect to different batch sizes. A larger batch size leads to a steeper global descent but may get stuck in a local minimum and has a higher memory usage. A smaller batch size adds more randomness to the optimization and leads to a low convergence speed. The batch size should be decided based on the data size and the hardware constraints.

Section 4 Regularization (10 mins + 10 mins)

a. With the existing spiral data and network setting with ReLU activation function, attendees are encouraged to try L2 based regularization.
Note: L2 helps to generalize the model and make the model less prone to overfitting. The difference between test loss and training loss is much smaller than the ones without using L2 regularization.

b. With the existing spiral data and network setting with ReLU activation function, attendees are encouraged to try L1 based regularization.

Note: L1 also helps to generalize the model and it enables a greater shrinkage to edge weights, where a majority of the edge weights can be with 0 or close to 0 values. The regularization rate
should not be too high. If the rate is high, the model will be oversimplified that could induce oscillations.

**Section 5 Feature Engineering (10 mins + 10 mins)**

a. Simplify the existing model (i.e., delete hidden layers and neurons) while introducing more augmented features ($x_1x_2$, $x_1^2$, etc.) and further simplify the model.

b. Choose circle data and then simplify the model, then delete or involve more existing and augmented features ($x_1x_2$, $x_1^2$, etc.) and simplify the model.
Note: Nicely engineered or augmented features can greatly contribute to the performance. However, in real practice, feature engineering is more like an art than science, since the relationship between the data and potential engineered features can be difficult to pinpoint.