1 The Lottery-Ticket Hypothesis (LTH) states that in a dense network, there exists a **winning or lottery subnetwork** which if trained from scratch again, would produce results very close to the dense network

1-a. "Tickets" are these sparse subnetworks within the larger network, defined by a binary mask over the weights. A winning ticket is a subnetwork that, when trained from the same initialization as the original network, can achieve similar performance to the full network.

2 LTH shows that a lot of deep, dense networks are often overparametrized. This challenges the assumption that a bigger network is always better and shows that it is worthwhile to compress your model

3 Early Bird Tickets (EBT) also aims to find emergent subnetworks in a dense network. It builds on the lottery ticket paper by hypothesizing that the emergent subnetworks can be detected earlier during the training process instead of full convergence unlike the lottery ticket hypothesis.

3-a EBT tries to solve the high computational cost of LTH by detecting the subnetworks early. This is done by removing the need for training till full convergence.

4 To detect EBTs we follow the following steps:

(a) Train the network for 10-20% of the way to full convergence

(b) At certain checkpoints create a mask over the weights with the lowest magnitudes to create a sparse subnetwork

(c) Keep track of the masks that we create and observe the change over checkpoints through some similarity measure

(d) If the masks from the past few checkpoints do not change much then we have an EBT

(e) We restart training with the EBT subnetwork

5 5-a The regions in the top left appear darker because initially the masks will change quite rapidly since we might not have a good guess for the EBT

5-b As the training progresses, if an EBT exists, then the mask should stabilize. Hence for later epochs comparison (as you go right and bottom) you see it get yellower

5-c Each entry on the diagonal compares a mask with itself, so the similarity is perfect (distance = 0). Hence, the diagonal is always yellow regardless of training stage.

6 Derivative of a question in Lab 2, we could use **mutual information** of each activation with the labels as a metric of importance. We create a mask on the bottom contributors of mutual information.

Rationale
Neurons that quickly become predictive of labels are likely to form the predictive backbone of the network. Mutual information (MI) directly measures how much a neuron's activation reduces uncertainty about the label, so it is a principled signal of early importance.

6-a We could test this by training a dense network for only a small fraction of the epochs and then recording neuron activations on a held-out probe set. Compute mutual information between each neuron's activations and the labels to rank neurons by importance, prune the least informative ones, and rewind the kept weights to their initialization. Retrain the pruned network fully and compare its accuracy and mask overlap to standard Early-Bird tickets or IMP. If the accuracy is close while requiring less training, the MI method is validated.