

The background of the slide is a collage of abstract, technical-looking images. It features a network graph with red lines and green nodes, a grid of small purple arrows, and a scatter plot with orange and blue dots. The text "Overfitting and Tree Pruning" is centered in a large, bold, black font with a white outline. The number "1" is in the bottom right corner.

# Overfitting and Tree Pruning

# Overfitting and Tree Pruning

- ❑ Overfitting: An induced tree may overfit the training data
  - ❑ Too many branches, some may reflect anomalies due to noise or outliers
  - ❑ Poor accuracy for unseen samples
- ❑ Two approaches to avoid overfitting
  - ❑ Prepruning: *Halt tree construction early* - do not split a node if this would result in the goodness measure falling below a threshold
    - ❑ Difficult to choose an appropriate threshold
  - ❑ Postpruning: *Remove branches* from a “fully grown” tree - get a sequence of progressively pruned trees
    - ❑ Use a set of data different from the training data to decide which is the “best pruned tree”

