CS 4824/ECE 4424: Decision Trees

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

Supervised Function Approximation

- Problem setting
  - Set of possible instances $\boldsymbol{X}$
  - Unknown target function $\boldsymbol{f}: \boldsymbol{X} \rightarrow \boldsymbol{Y}$
  - Set of function hypotheses: $\boldsymbol{H} = \{h \mid h: \boldsymbol{X} \rightarrow \boldsymbol{Y}\}$
- Input
  - Training examples $\{\langle X^{(i)}, Y^{(i)} \rangle\}$ of unknown function $\boldsymbol{f}$
- Output
  - Hypothesis $h \in \boldsymbol{H}$ that best approximates $\boldsymbol{f}$

Supervised Function Approximation: Decision Tree Learning

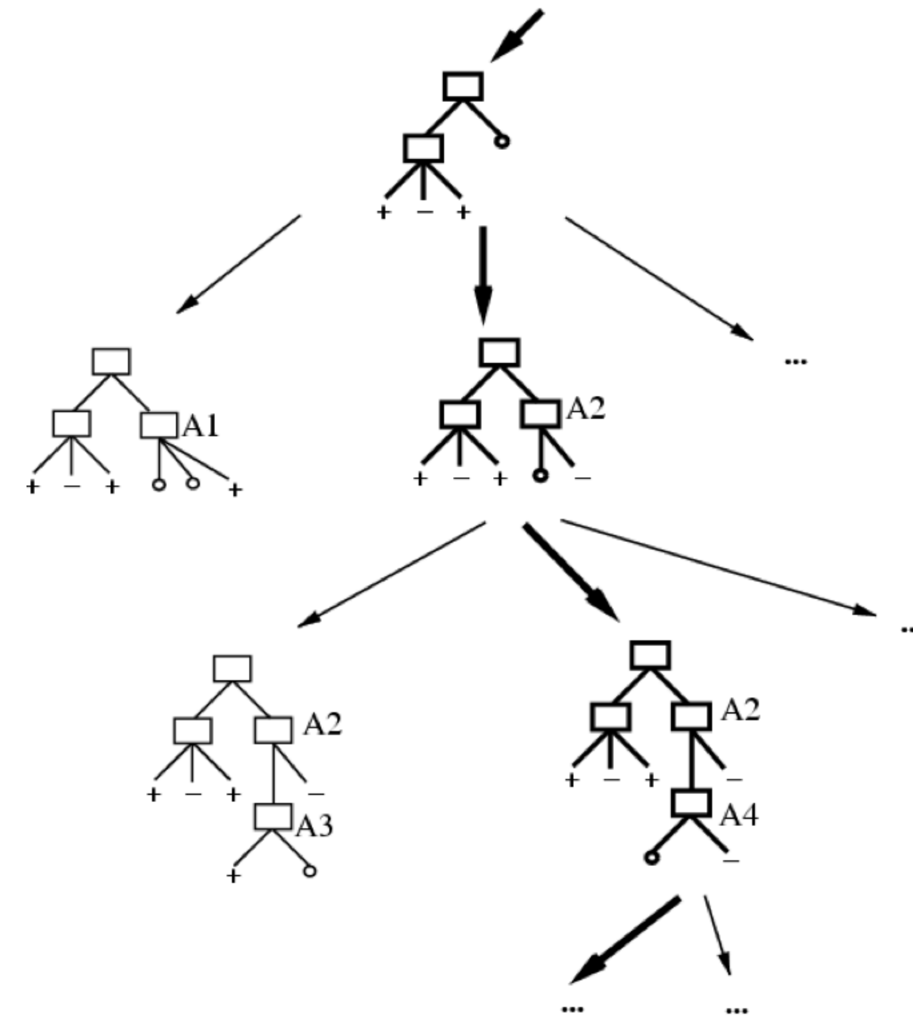
- Problem setting
  - Set of possible instances $\boldsymbol{X}$
    - each instance x in X is a feature vector $x = \langle x_1, x_2, \ldots, x_n \rangle$
  - Unknown target function $\boldsymbol{f}: \boldsymbol{X} \to \boldsymbol{Y}$
    - Y is discrete valued
  - Set of function hypotheses: $\boldsymbol{H} = \{h \mid h: \boldsymbol{X} \to \boldsymbol{Y}\}$
    - each hypothesis h is a decision tree

- Input
  - Training examples $\{\langle X^{(i)}, Y^{(i)} \rangle\}$ of unknown function $\boldsymbol{f}$

- Output
  - Hypothesis $h \in \boldsymbol{H}$ that best approximates $\boldsymbol{f}$

Searching for the best hypothesis

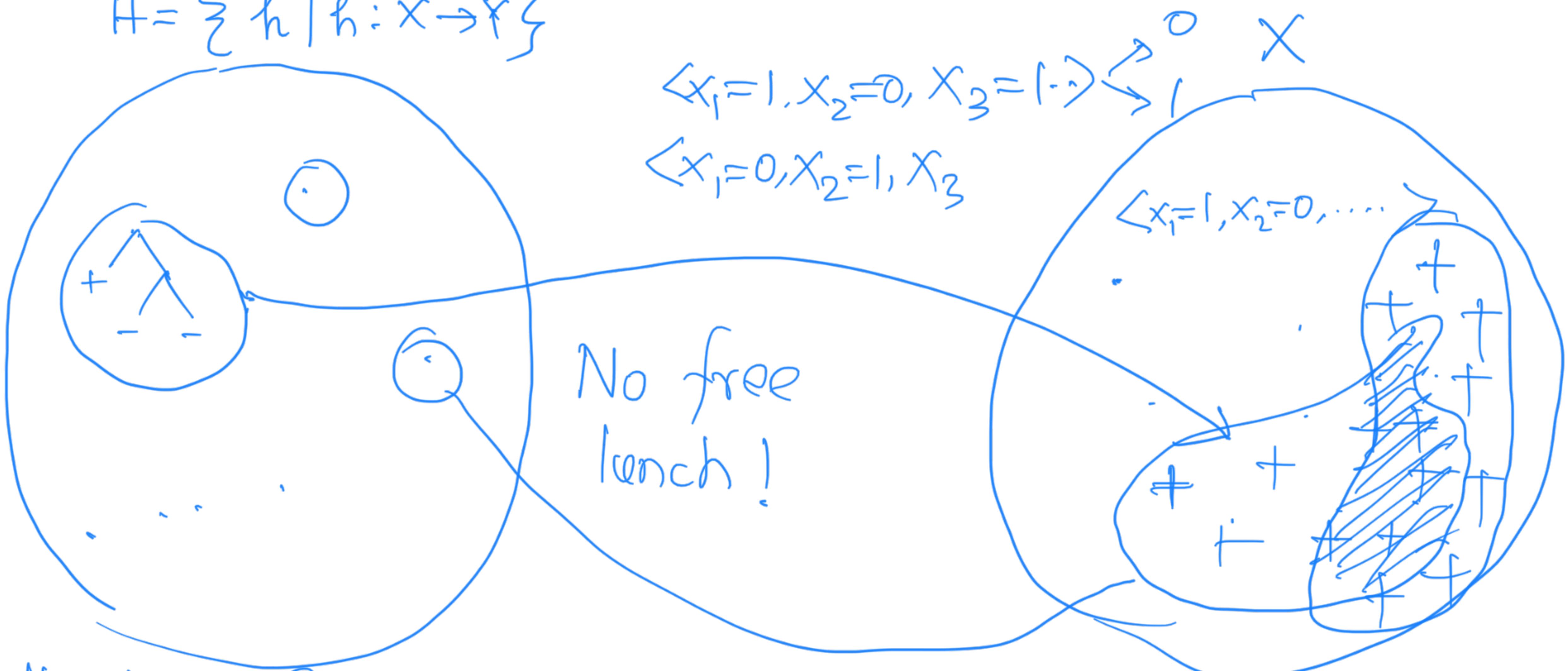
- Decision tree learning
  - performs a heuristic search through space of decision trees



The big picture

$f: X \to Y$ $\quad x = \langle x_1, \dots x_n \rangle \quad x_i \in \{0,1\}, \ Y \in \{0,1\}$

$H = \{ h \mid h: X \to Y \}$



Hypothesis Space $\qquad\qquad$ Instance Space

$|X| = 2^n$

DT's that can represent all possible functions $= 2^{2^n}$

possible teachabe functions $2^{|X|} = 2^{2^n}$

Traing examples we need s.t. there is one DT for $H =$ All of th

No free lunch!

- Inductive inference
  - Reliable generalization beyond the training data is impossible unless we add more assumptions into the model.

- "Essentially all models are wrong, but some are useful." — George Box

No free lunch!

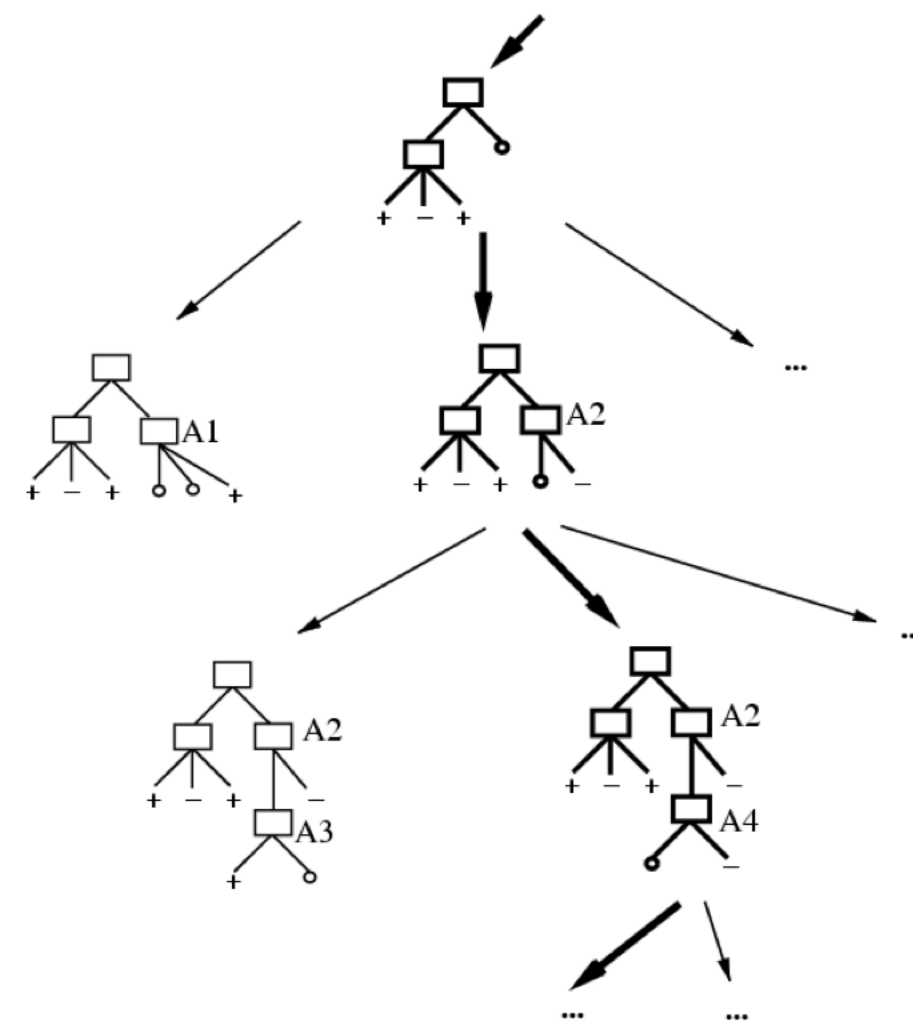
- Inductive inference
  - Reliable generalization beyond the training data is impossible unless we add more assumptions into the model.

- "Essentially all models are wrong, but some are useful."
  — George Box

Q. What was the assumption in decision tree learning?

Which tree should we output?

- Decision tree learning
  - performs a heuristic search through space of decision trees



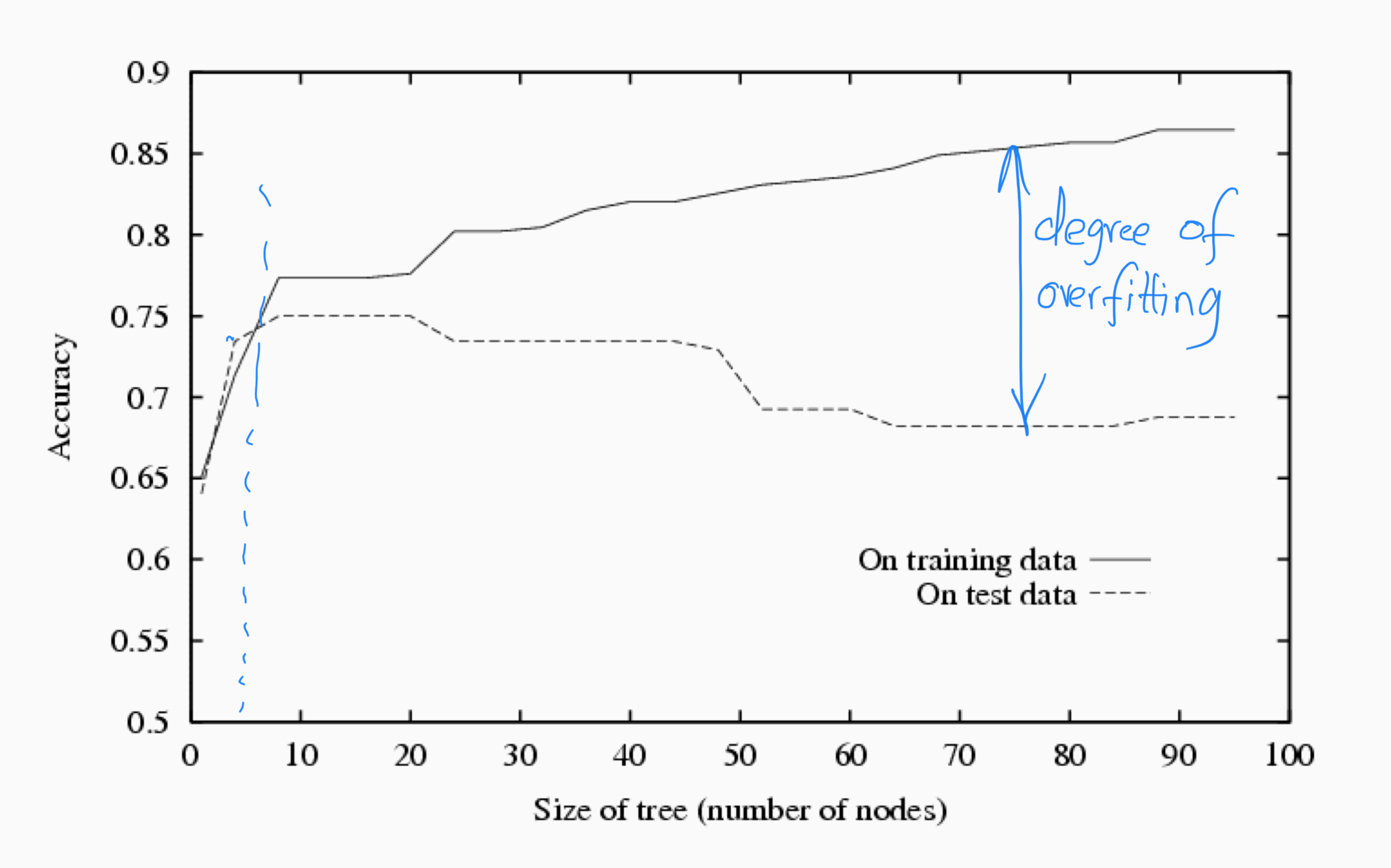
Assumption in decision tree learning

- Stop the top-down greedy growth of decision tree at smallest acceptable tree. Why?
  - Prefer the simplest hypothesis that fits the data (**Occam's Razor**)

Assumptions (or the lack of it) have implications…

- What if we let the decision tree learning algorithm to grow freely at will?
  - This may lead to **overfitting**!

Overfitting in decision tree learning



Avoiding overfitting

- How to avoid overfitting?
  - Stop growing the tree when data split is not significant
  - Grow full tree, then post-prune

- How to select the "best" tree?
  - Measure performance on training dataset
  - Measure perforce on standalone validation dataset

Reduce-error pruning

Split data into *training* and *validation* set

Create tree that classifies *training* set correctly

Do until further pruning is harmful:

1. Evaluate impact on *validation* set of pruning each possible node (plus those below it)
2. Greedily remove the one that most improves *validation* set accuracy

Produces smallest version of most accurate subtree