10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Decision Trees

Matt Gormley

Lecture 2

January 22, 2018

Reminders

- **Homework 1: Background**
  - **Out: Wed, Jan 17 (today)**
  - **Due: Wed, Jan 24 at 11:59pm**
  - Two parts: written part on Canvas, programming part on Autolab
  - unique policy for this assignment: **unlimited submissions** (i.e. keep submitting until you get 100%)

ML as Function Approximation

Chalkboard

- ML as Function Approximation
  - Problem setting
  - Input space
  - Output space
  - Unknown target function
  - Hypothesis space
  - Training examples

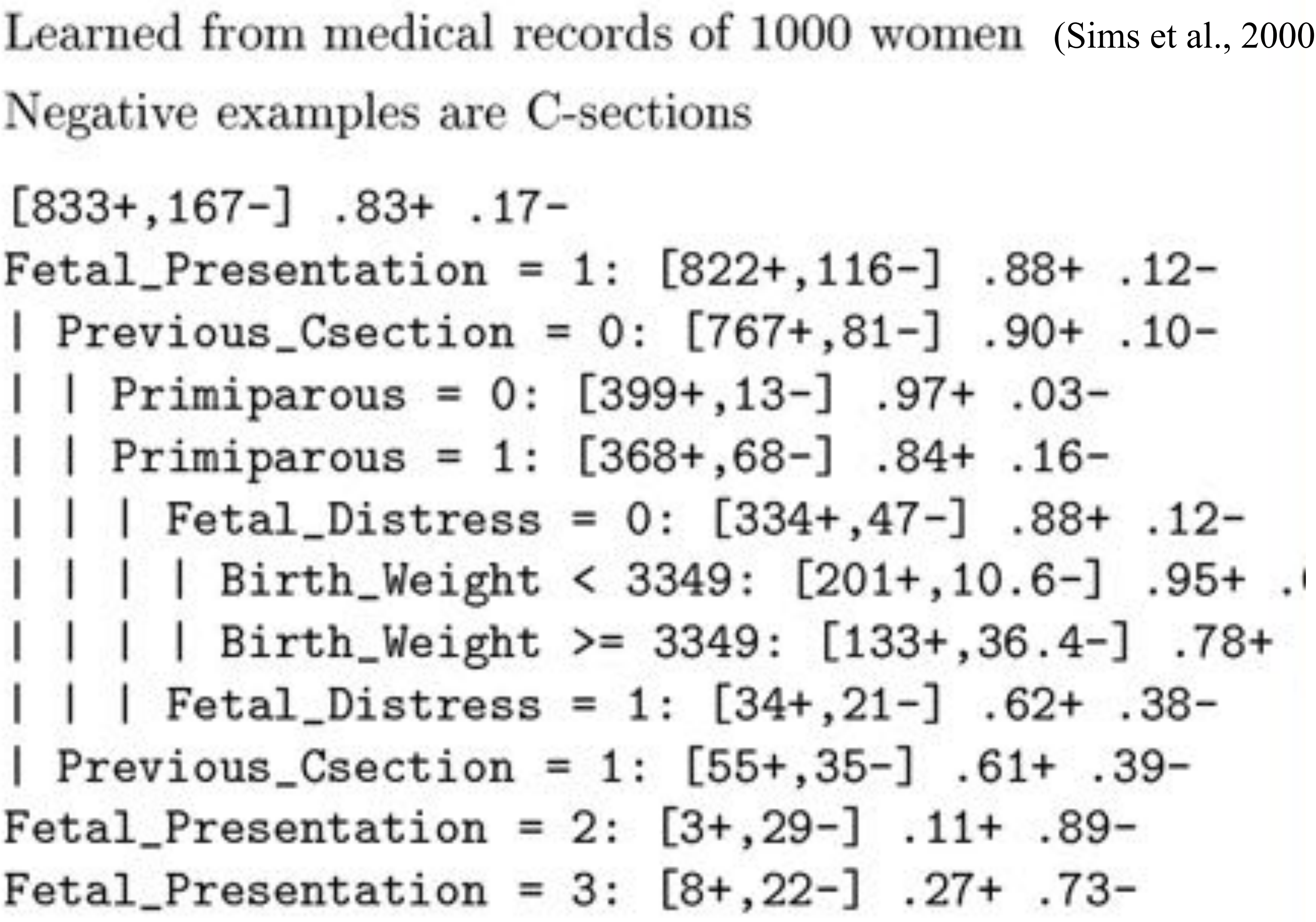
DECISION TREES

Decision Trees

Chalkboard

- – Example: Medical Diagnosis
- – Does memorization = learning?
- – Decision Tree as a hypothesis
- – Function approximation for DTs
- – Decision Tree Learning

Tree to Predict C-Section Risk



Learned from medical records of 1000 women (Sims et al., 2000)

Negative examples are C-sections

```
[833+,167-] .83+ .17-
Fetal_Presentation = 1: [822+,116-] .88+ .12-
| Previous_Csection = 0: [767+,81-] .90+ .10-
| | Primiparous = 0: [399+,13-] .97+ .03-
| | Primiparous = 1: [368+,68-] .84+ .16-
| | | Fetal_Distress = 0: [334+,47-] .88+ .12-
| | | | Birth_Weight < 3349: [201+,10.6-] .95+ .
| | | | Birth_Weight >= 3349: [133+,36.4-] .78+
| | | Fetal_Distress = 1: [34+,21-] .62+ .38-
| Previous_Csection = 1: [55+,35-] .61+ .39-
Fetal_Presentation = 2: [3+,29-] .11+ .89-
Fetal_Presentation = 3: [8+,22-] .27+ .73-
```

Figure from Tom Mitchell

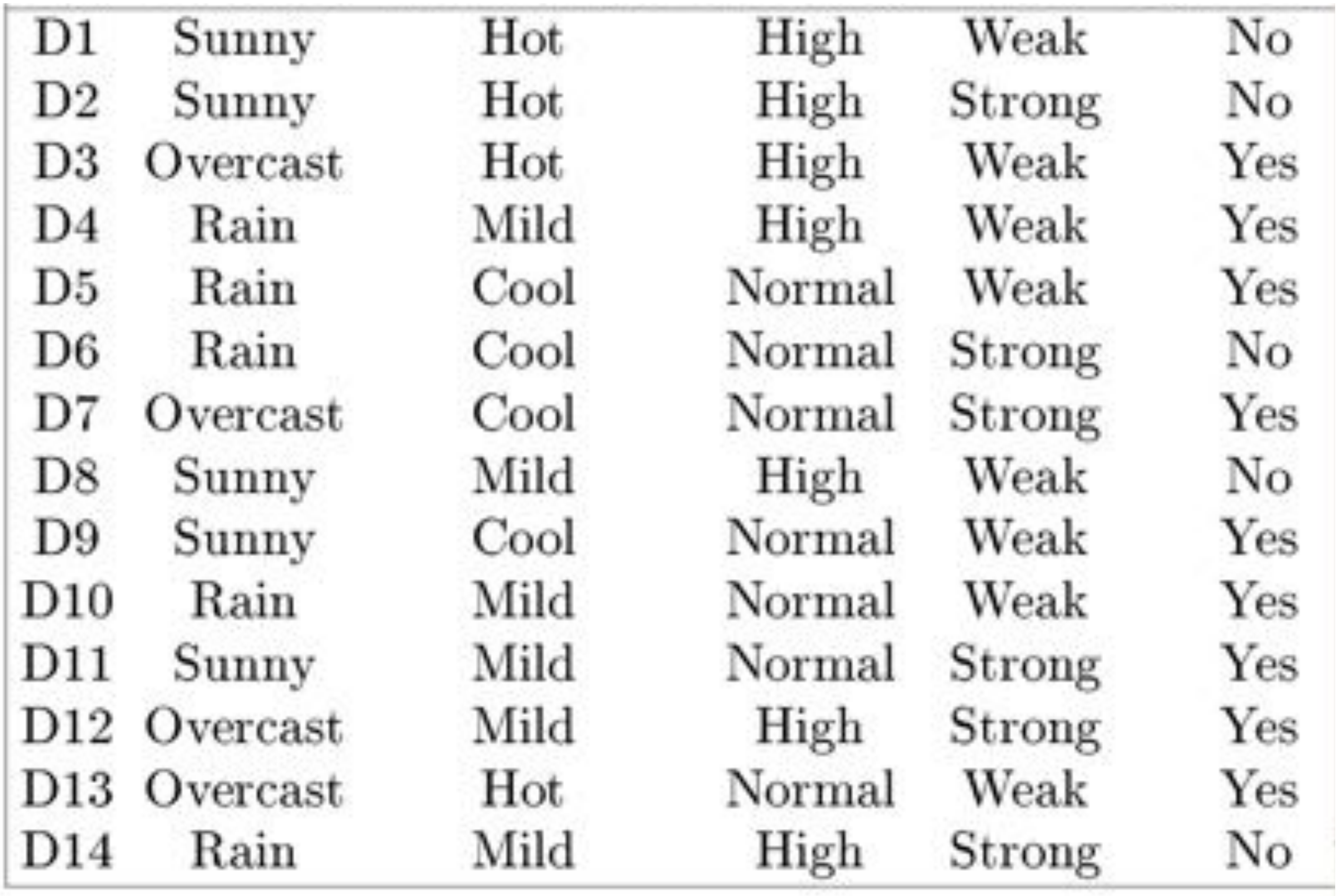
Decision Trees

Chalkboard

- Information Theory primer
  - Entropy
  - (Specific) Conditional Entropy
  - Conditional Entropy
  - Information Gain / Mutual Information
- Information Gain as DT splitting criterion

Tennis Example

Dataset:



Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure from Tom Mitchell

Tennis Example

Which attribute yields the best classifier?

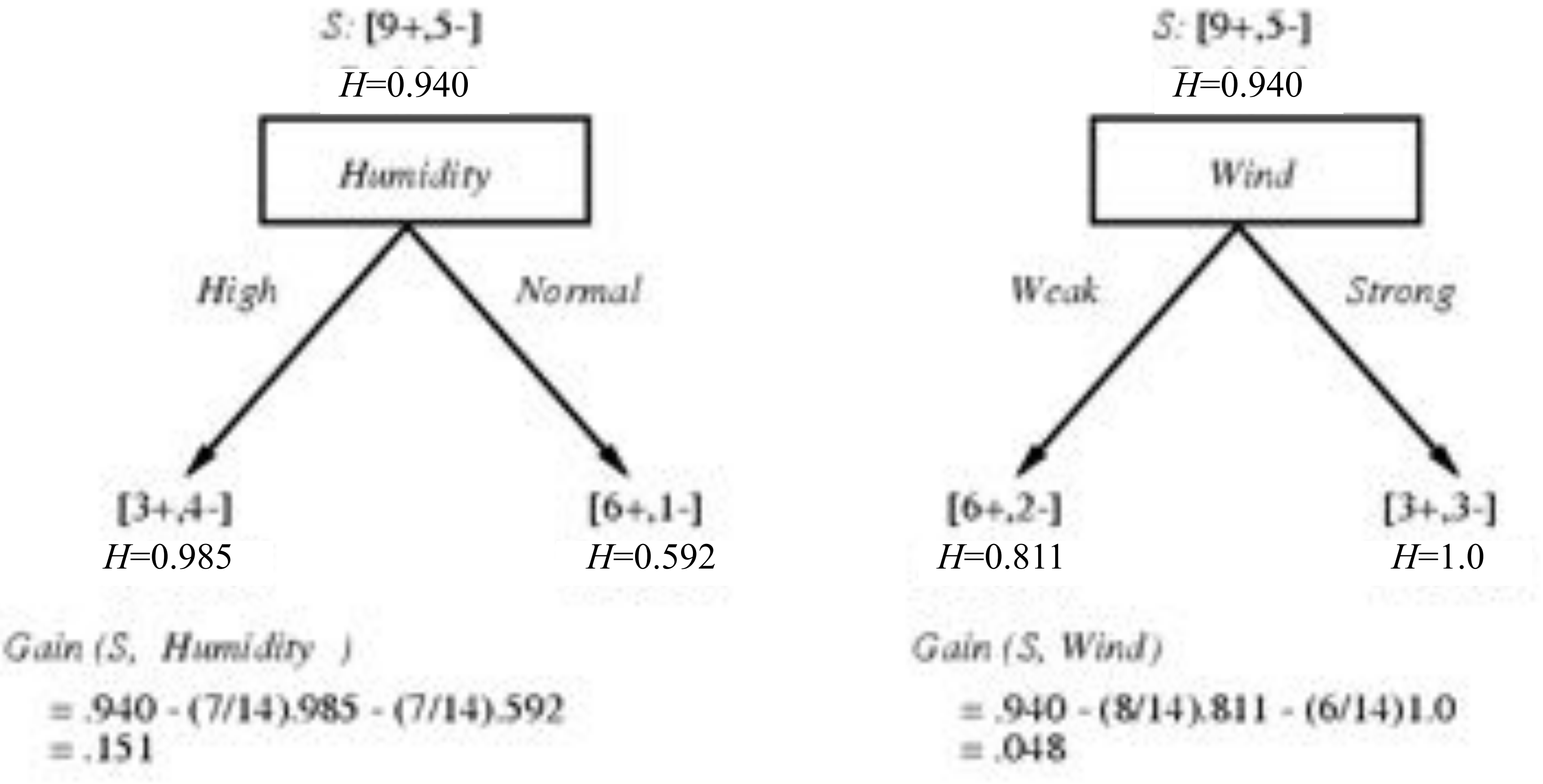
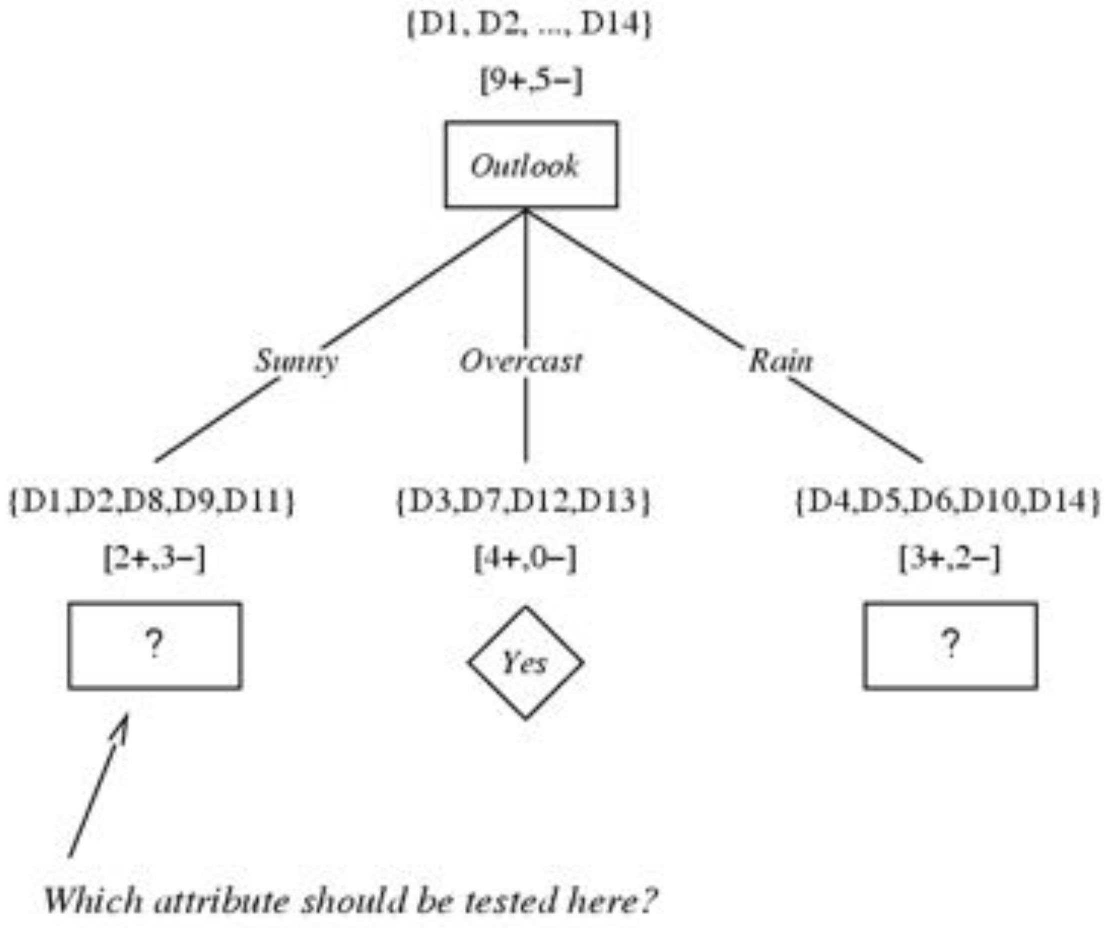


Figure from Tom Mitchell

Tennis Example



S_{sunny} = {D1,D2,D8,D9,D11}

$Gain\ (S_{sunny}\ , Humidity) = .970 - (3/5)\,0.0 - (2/5)\,0.0 = .970$

$Gain\ (S_{sunny}\, , Temperature) = .970 - (2/5)\,0.0 - (2/5)\,1.0 - (1/5)\,0.0 = .570$

$Gain\ (S_{sunny}, Wind) = .970 - (2/5)\,1.0 - (3/5)\,.918 = .019$

Figure from Tom Mitchell

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
0	1	0
0	1	0
1	1	0
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1

In-Class Exercise

1. Which attribute would **misclassification rate** select for the next split?
2. Which attribute would **information gain** select for the next split?
3. *Justify your answers.*

Decision Trees

Chalkboard

- ID3 as Search
- Inductive Bias of Decision Trees
- Occam’s Razor

Overfitting

Consider a hypothesis h and its

- Error rate over training data: $error_{train}(h)$
- True error rate over all data: $error_{true}(h)$

We say h overfits the training data if

$$error_{true}(h) > error_{train}(h)$$

Amount of overfitting =

$$error_{true}(h) - error_{train}(h)$$

Slide from Tom Mitchell

Overfitting in Decision Tree Learning

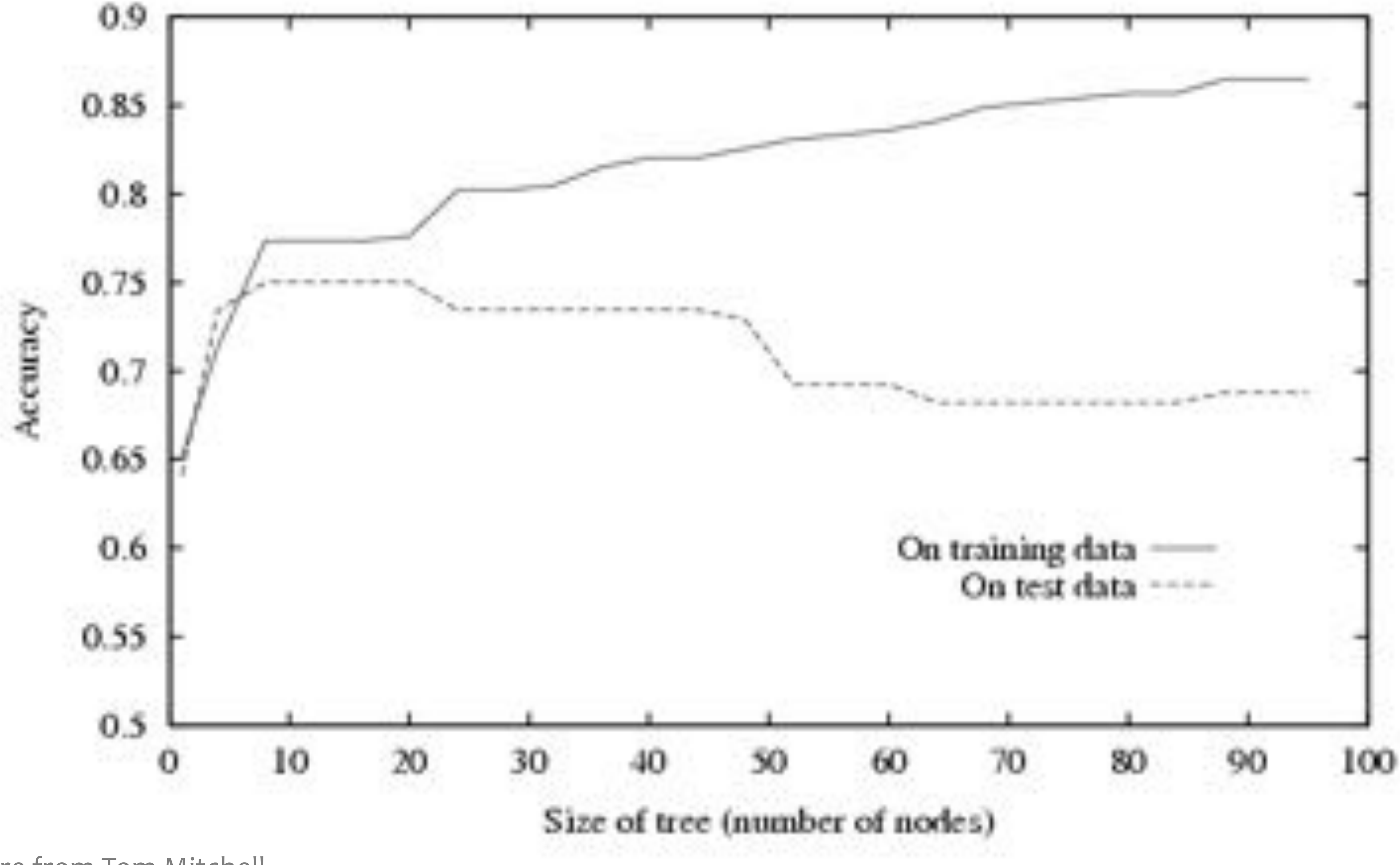
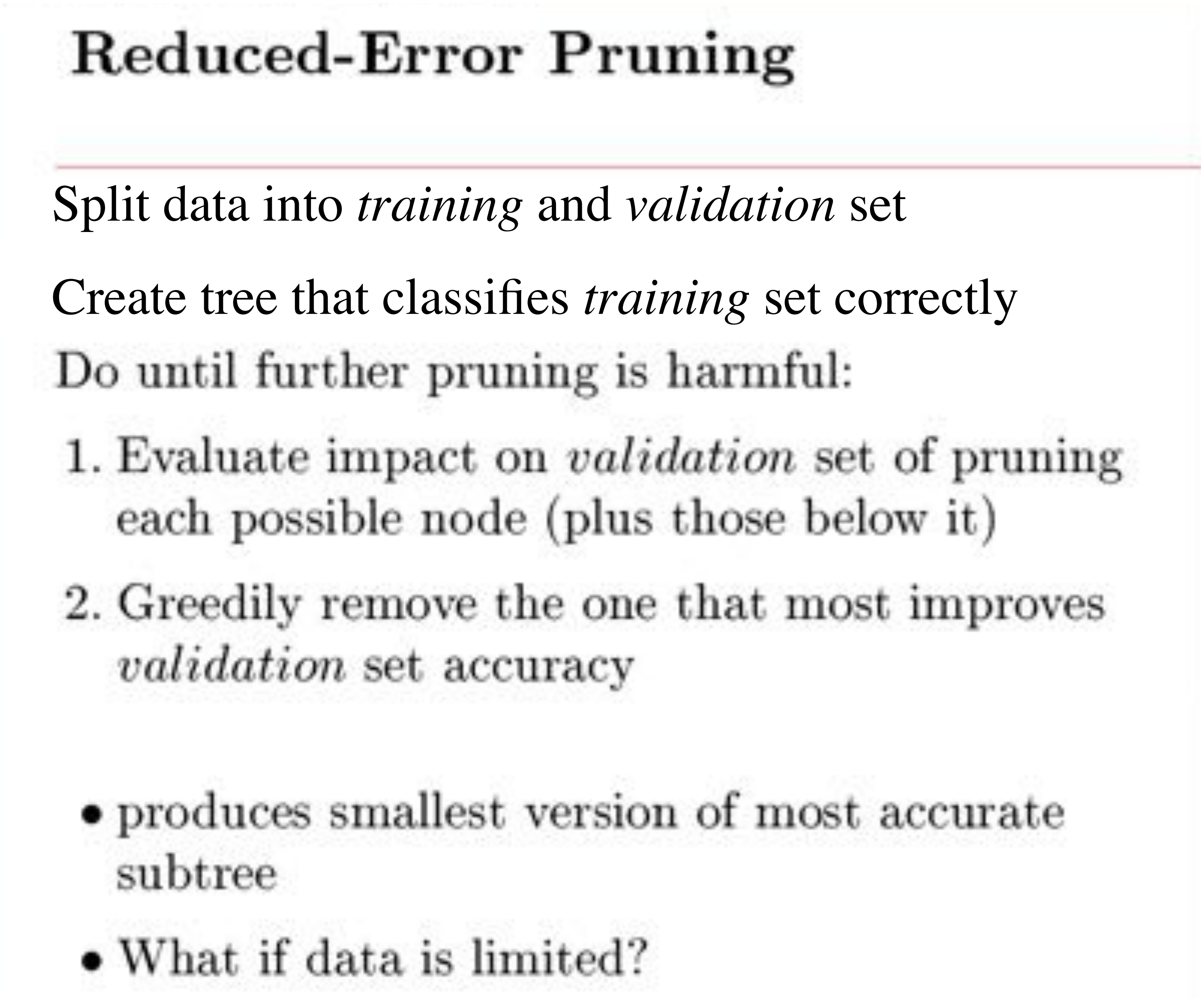


Figure from Tom Mitchell

How to Avoid Overfitting?

For Decision Trees…

1. Do not grow tree beyond some **maximum depth**
2. Do not split if splitting criterion (e.g. Info. Gain) is **below some threshold**
3. Stop growing when the split is **not statistically significant**
4. Grow the entire tree, then prune



Reduced-Error Pruning

Split data into *training* and *validation* set

Create tree that classifies *training* set correctly

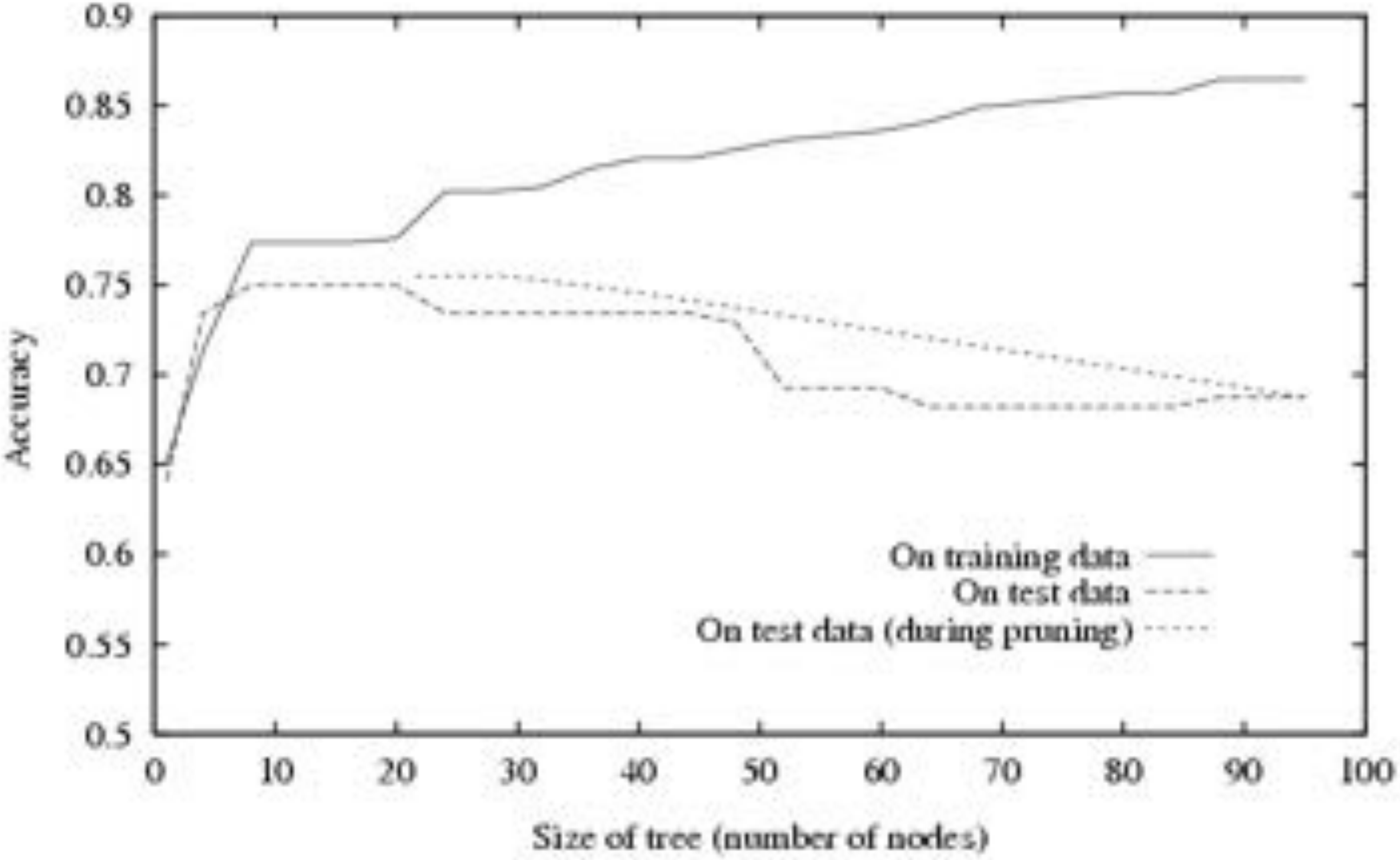
Do until further pruning is harmful:

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
- What if data is limited?

Slide from Tom Mitchell

Effect of Reduced-Error Pruning



Slide from Tom Mitchell

Questions

- Will ID3 always include all the attributes in the tree?
- What if some attributes are real-valued? Can learning still be done efficiently?
- What if some attributes are missing?

Learning Objectives

You should be able to…

1. Implement Decision Tree training and prediction
2. Use effective splitting criteria for Decision Trees and be able to define entropy, conditional entropy, and mutual information / information gain
3. Explain the difference between memorization and generalization [CIML]
4. Describe the inductive bias of a decision tree
5. Formalize a learning problem by identifying the input space, output space, hypothesis space, and target function
6. Explain the difference between true error and training error
7. Judge whether a decision tree is "underfitting" or "overfitting"
8. Implement a pruning or early stopping method to combat overfitting in Decision Tree learning