Introduction to Classification, an example of Supervised Machine Learning

Classification: Definition

- Given a collection of examples (*training set*)
  - Each example is represented by a set of *features*, sometimes called *attributes*
  - Each example is to be given a label or class
- Find a *model* for the label as a function of the values of features.
- Goal: previously unseen examples should be assigned a label as accurately as possible.
- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Supervised vs. Unsupervised Learning

- Supervised learning (**classification and other tasks)**
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (includes **clustering)**
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

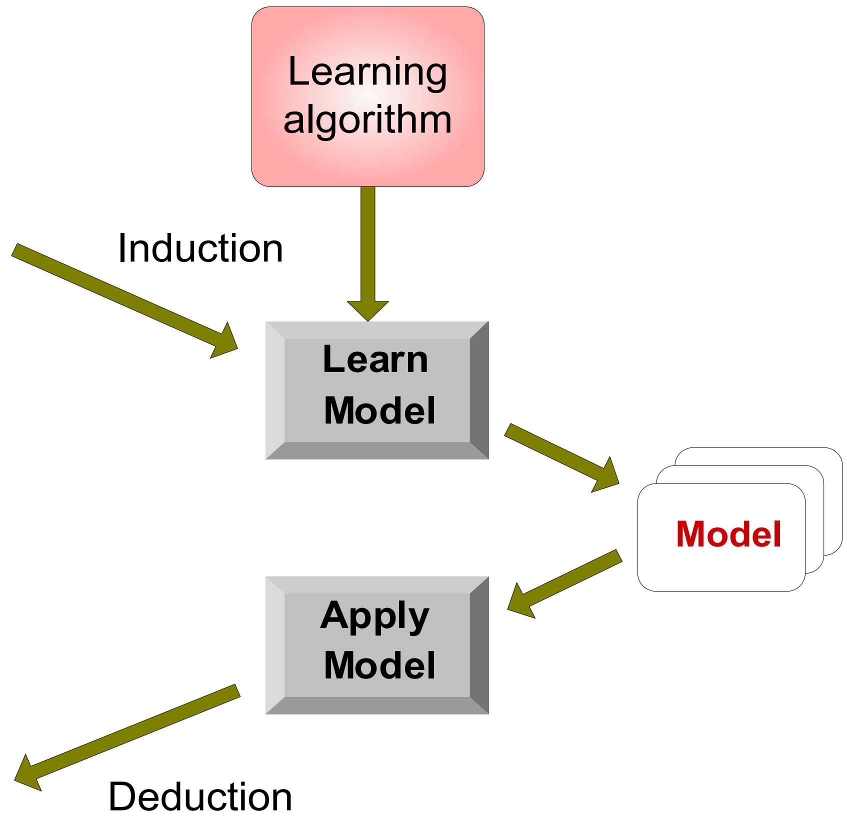
NLP Tasks

- Many NLP tasks can be accomplished either through
  - unsupervised techniques, sometimes also called rule-based or symbolic techniques
  - Supervised techniques, where the task is defined automatically from a training set
- In both cases, the evaluation of the task will most likely use a **training set** to define the technique and a **test set** for evaluation
  - POS tagging uses Hidden Markov Models
  - Parsing uses statistical lexicalized parsers
  - Sentiment analysis uses classification
- The evaluation of these tasks often uses ideas from the evaluation of classificatioin

Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

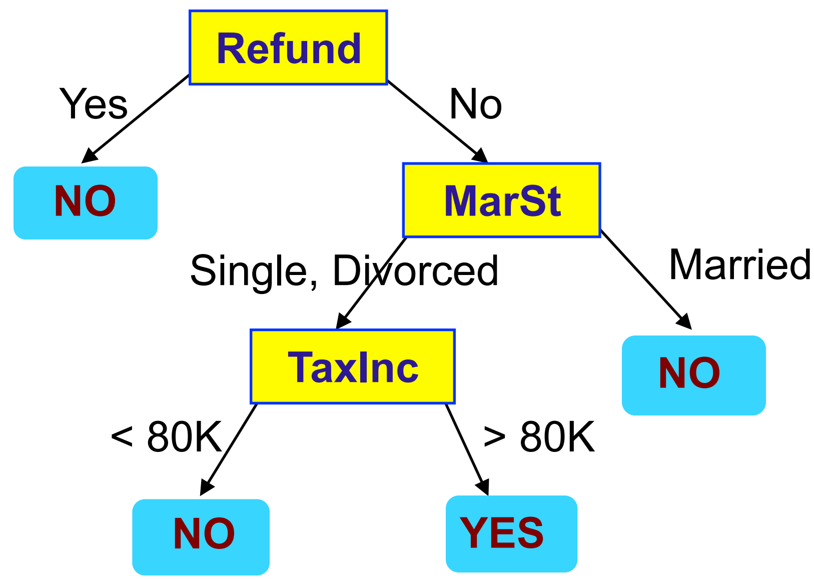
Classification Techniques

- There are a number of different classification algorithms to build a model for classification
  - Decision Tree based Methods
  - Rule-based Methods
  - Memory based reasoning, instance-based learning
  - Neural Networks
  - Genetic Algorithms
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines
- In this introduction, we illustrate classification tasks using Decision Tree methods
- Features can have numeric values (continuous) or a finite set of values (categorical/nominal), including boolean true/false

Example of a Decision Tree

boolean categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



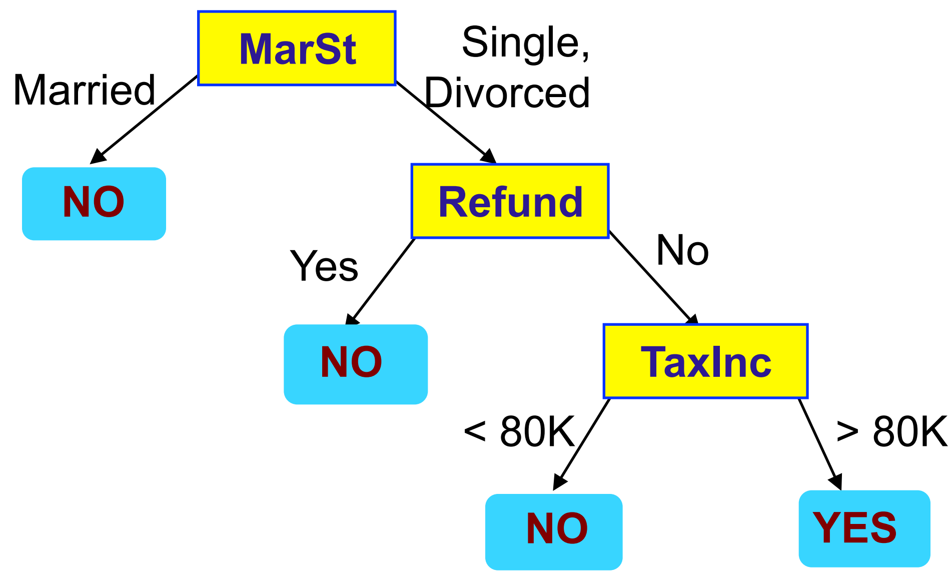
Training Data **Model: Decision Tree**

Example task: Given the marital status, refund status, and taxable income of a person, label them as to whether they will cheat on their income tax.

Another Example of Decision Tree

boolean categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

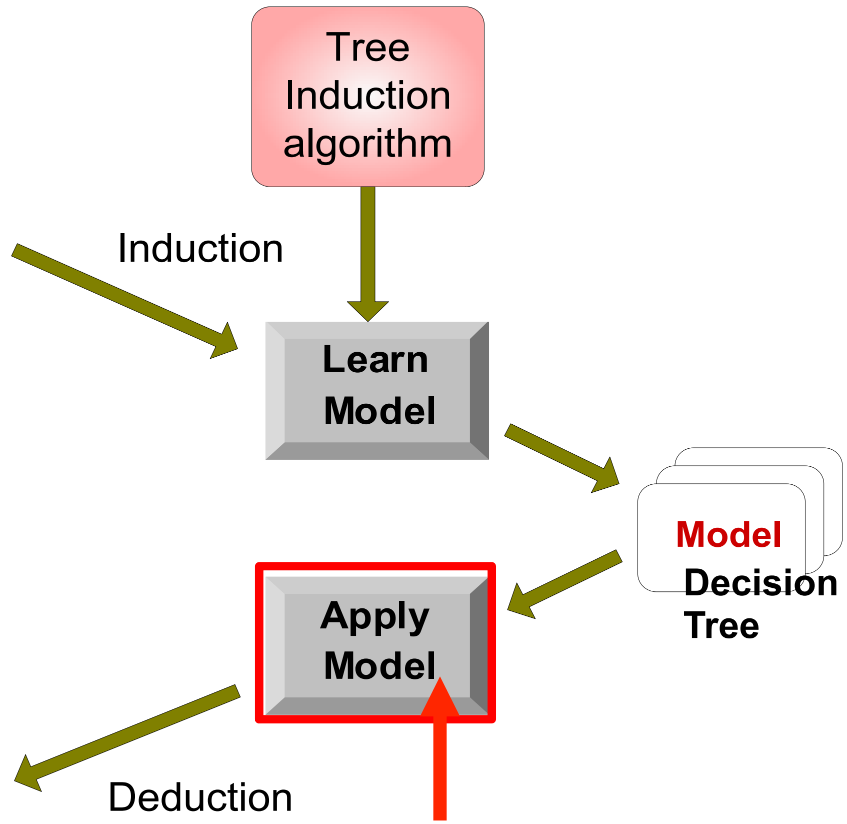


There could be more than one tree that fits the same data!

Decision Tree Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

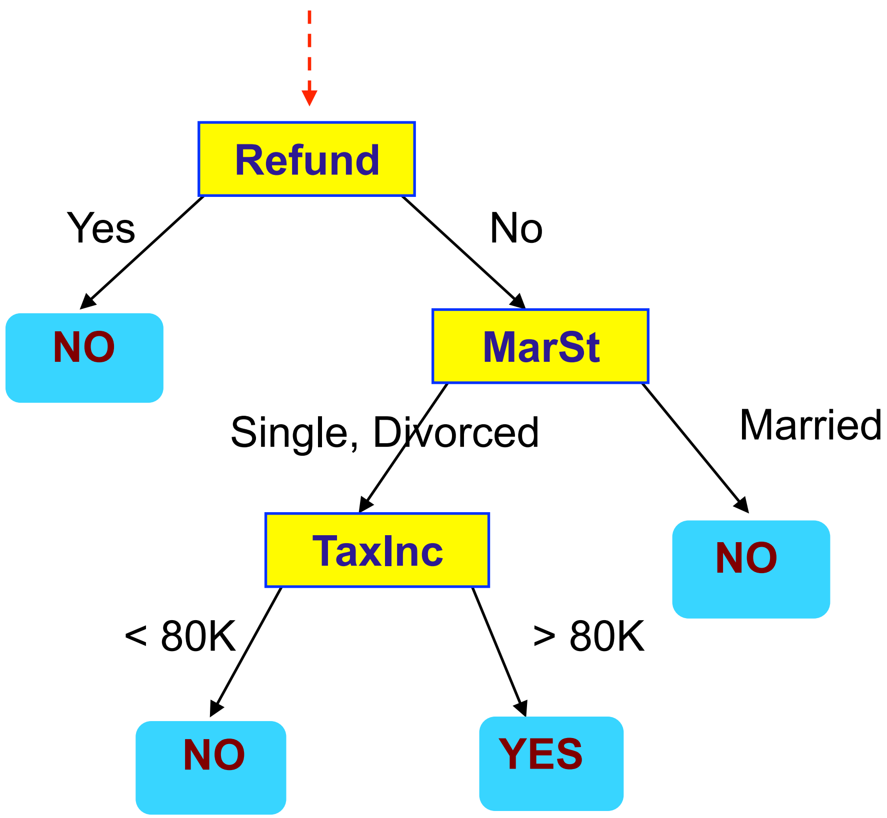


Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Apply Model to Test Data

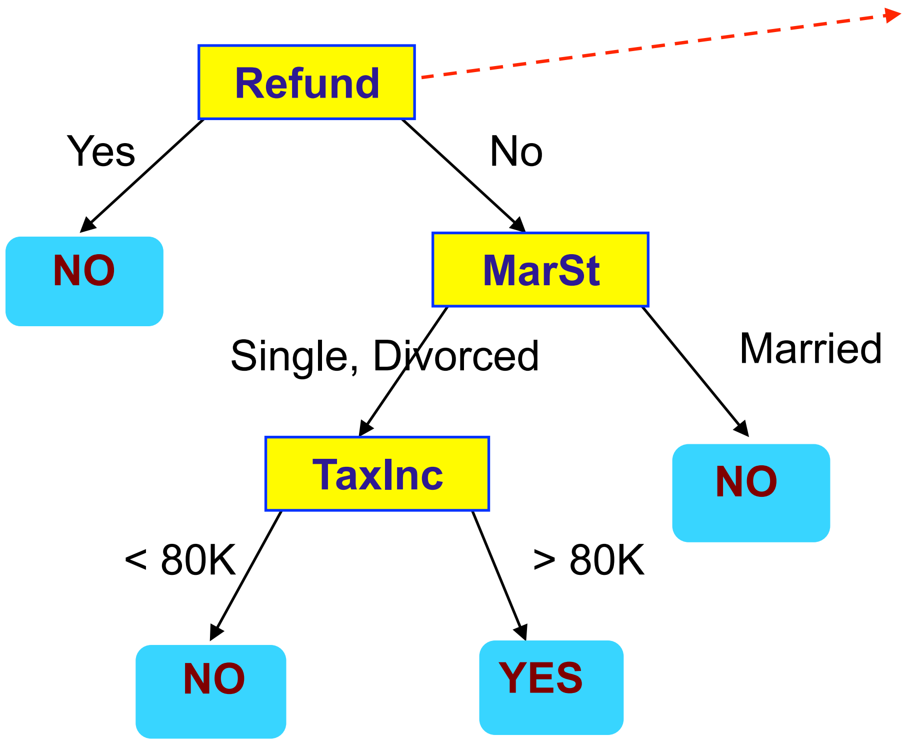
Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

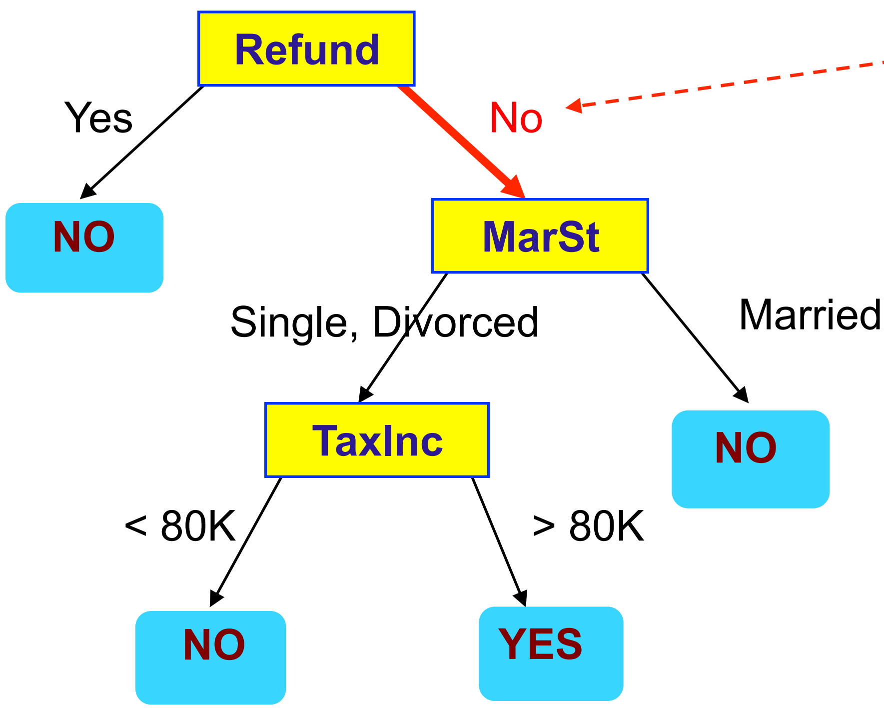
Apply Model to Test Data



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data



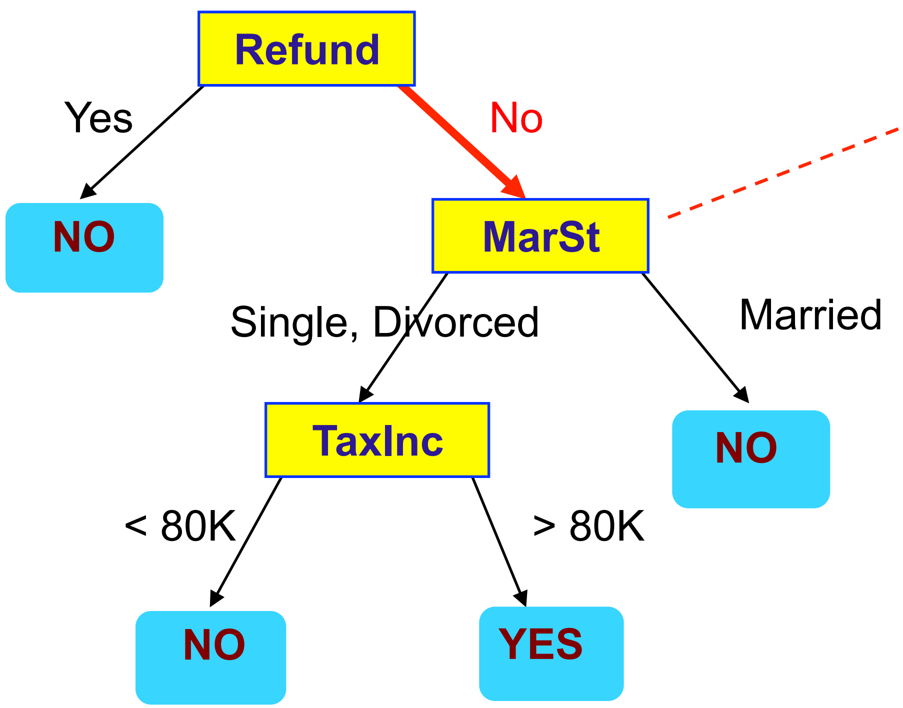
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

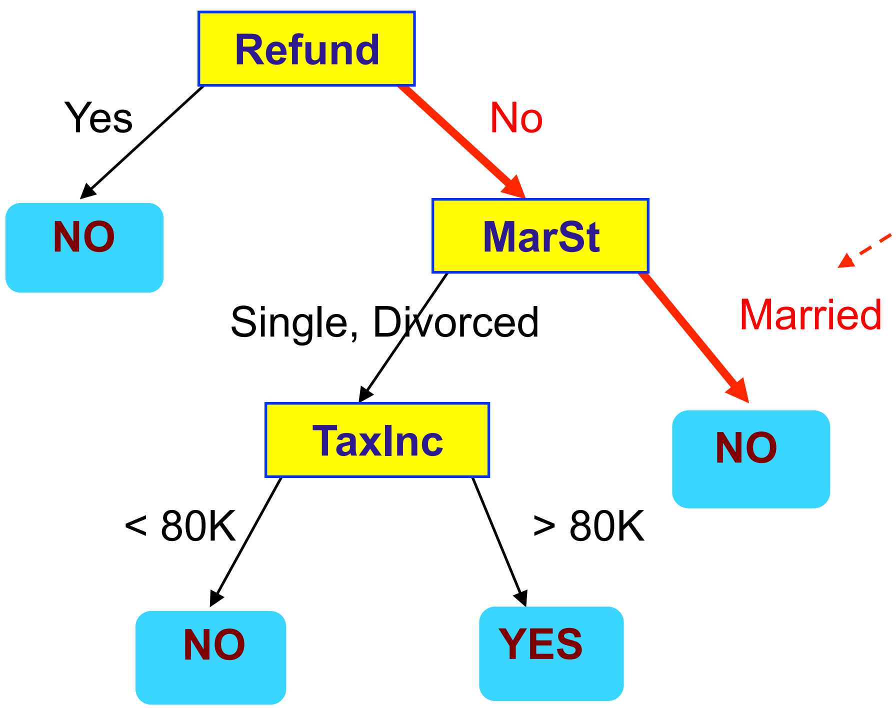
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

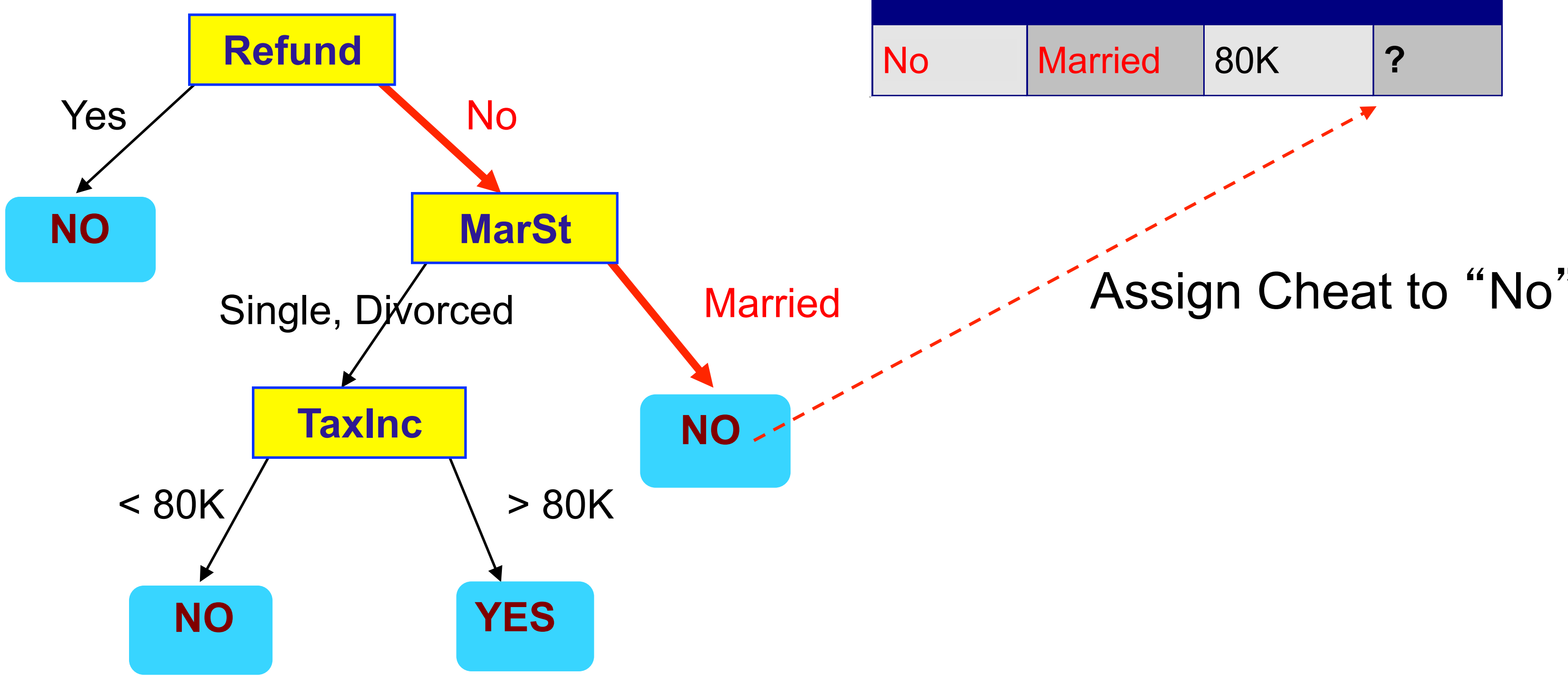
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

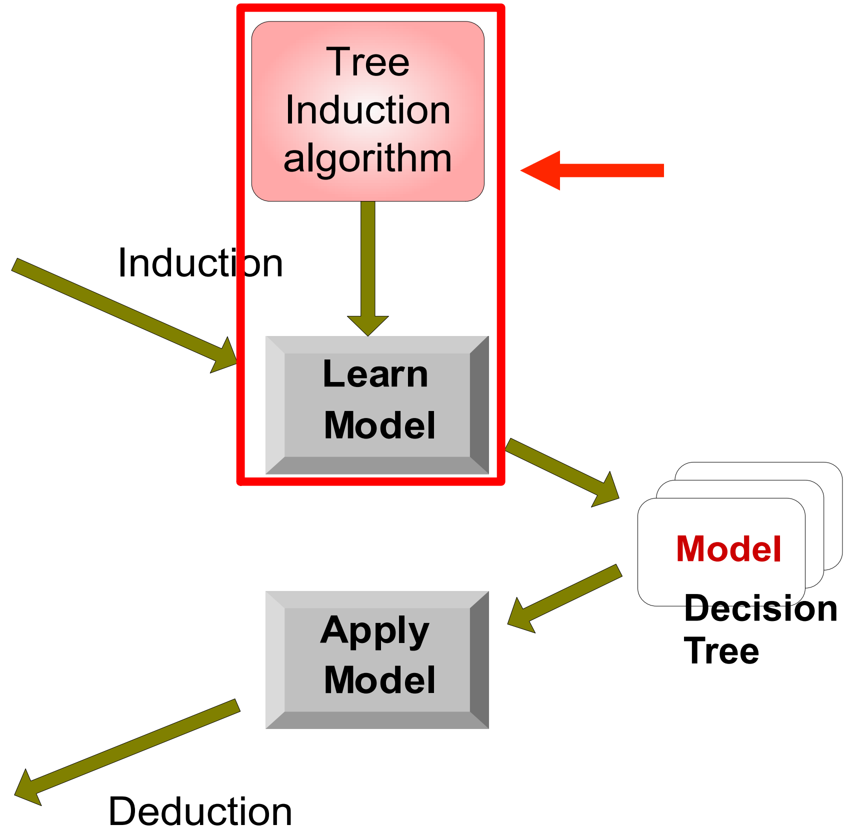
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Decision Tree Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix for a binary classifier (two labels) on test set:

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Classifier Accuracy Measures

- Another widely-used metric: Accuracy of a classifier M is the percentage of test set that are correctly classified by the model M

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

	Yes - C_1	No - C_2
Yes - C_1	a: True positive	b: False negative
No - C_2	c: False positive	d: True negative

classes	buy_computer = yes	buy_computer = no	total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
total	7366	2634	10000

Other Classifier Measures

- Alternative accuracy measures (e.g., for cancer diagnosis or information retrieval)

sensitivity = t-pos/pos /* true positive recognition rate */

specificity = t-neg/neg /* true negative recognition rate */

precision = t-pos/(t-pos + f-pos)

recall = t-pos/(t-pos + f-neg)

accuracy = sensitivity * pos/(pos + neg) + specificity * neg/(pos + neg)

Multi-Class Classification

- Most classification algorithms solve binary classification tasks, while many tasks are naturally multi-class, i.e. there are more than 2 labels
- **Multi-Class problems are solved by training a number of binary classifiers and combining them to get a multi-class result**
- Confusion matrix is extended to the multi-class case
- Accuracy definition is naturally extended to the multi-class case
- Precision and recall are defined for the binary classifiers trained for each label