COURSE NAME:
DATA WAREHOUSING & DATA MINING

LECTURE 16
TOPICS TO BE COVERED:

- What is classification?
- What is prediction?
- Issues regarding classification and prediction

CLASSIFICATION & PREDICTION

- **Databases are rich with hidden information that can be used for intelligent decision making.**
- Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.

CLASSIFICATION VS. PREDICTION

- Classification:
  - predicts categorical class labels
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction:
  - models continuous-valued functions, i.e., predicts unknown or missing values

APPLICATIONS

- Typical Applications
  - Credit Approval
  - Target Marketing
  - Medical Diagnosis
  - Treatment Effectiveness Analysis
  - Performance Prediction

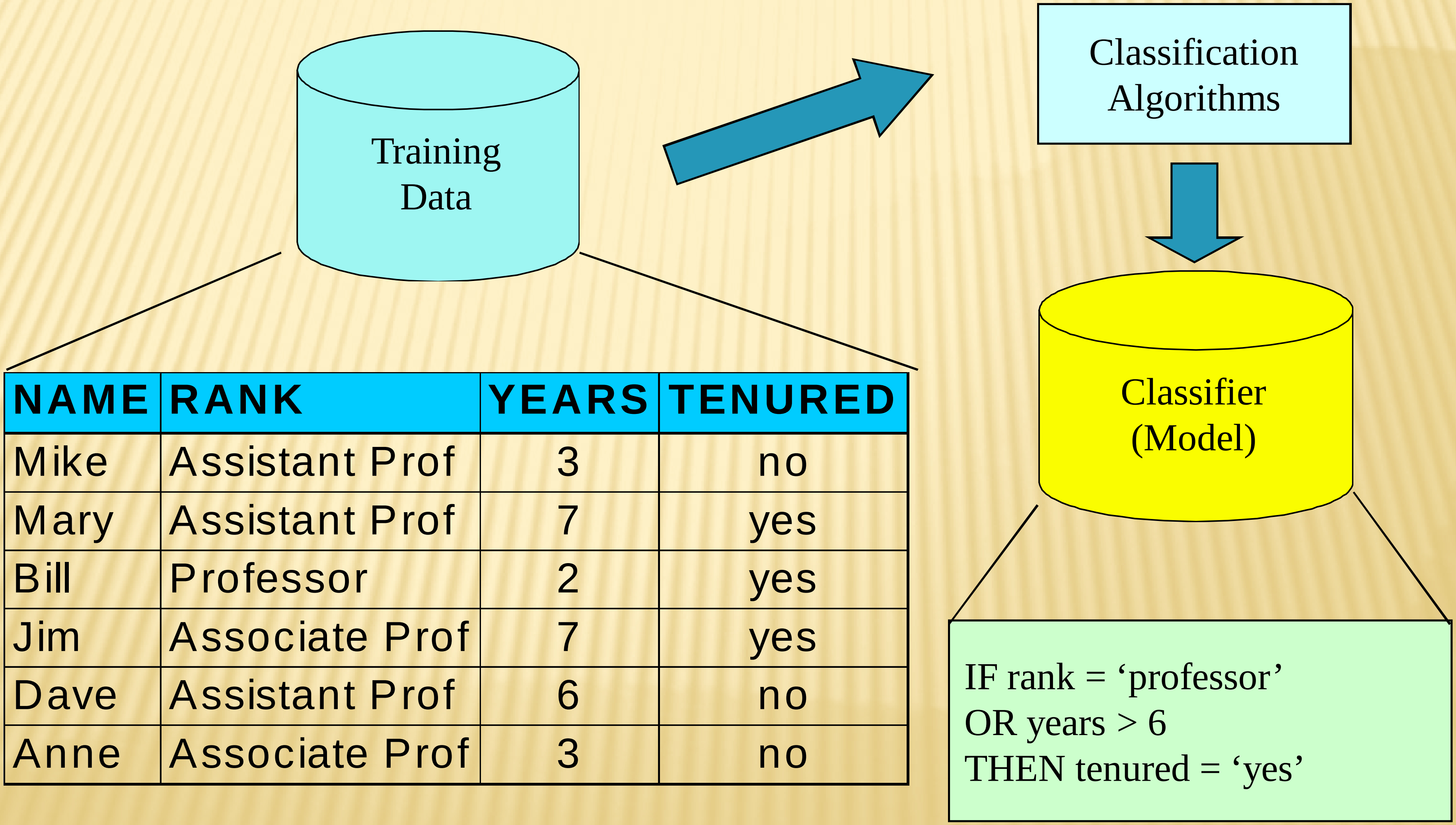
EXAMPLE

- **A bank loans officer needs analysis of her data in order to learn which loan applicants are “safe”andwhichare “risky” for the bank**
- A marketing manager at *AllElectronicsneeds data* analysis to help guess whether a customer with a given profile will buy a new computer.
- A medical researcher wants to analyze breast cancer data in order to predict which one of three specific treatments a patient should receive.
- In each of these examples, the data analysis task is classification, where a model or classifier is constructed to predict *categorical labels, such as “safe” or “risky” for the loan application data; “yes” or “no” for the marketing* data; or “treatment A,” “treatment B,” or “treatment C” for the medical data. These categories can be represented by discrete values, where the ordering among values has no meaning.
- For example, the values 1, 2, and 3 may be used to represent treatments A, B, and C, where there is no ordering implied among this group of treatment regimes.

CLASSIFICATION—A TWO-STEP PROCESS

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction: training set
  - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur

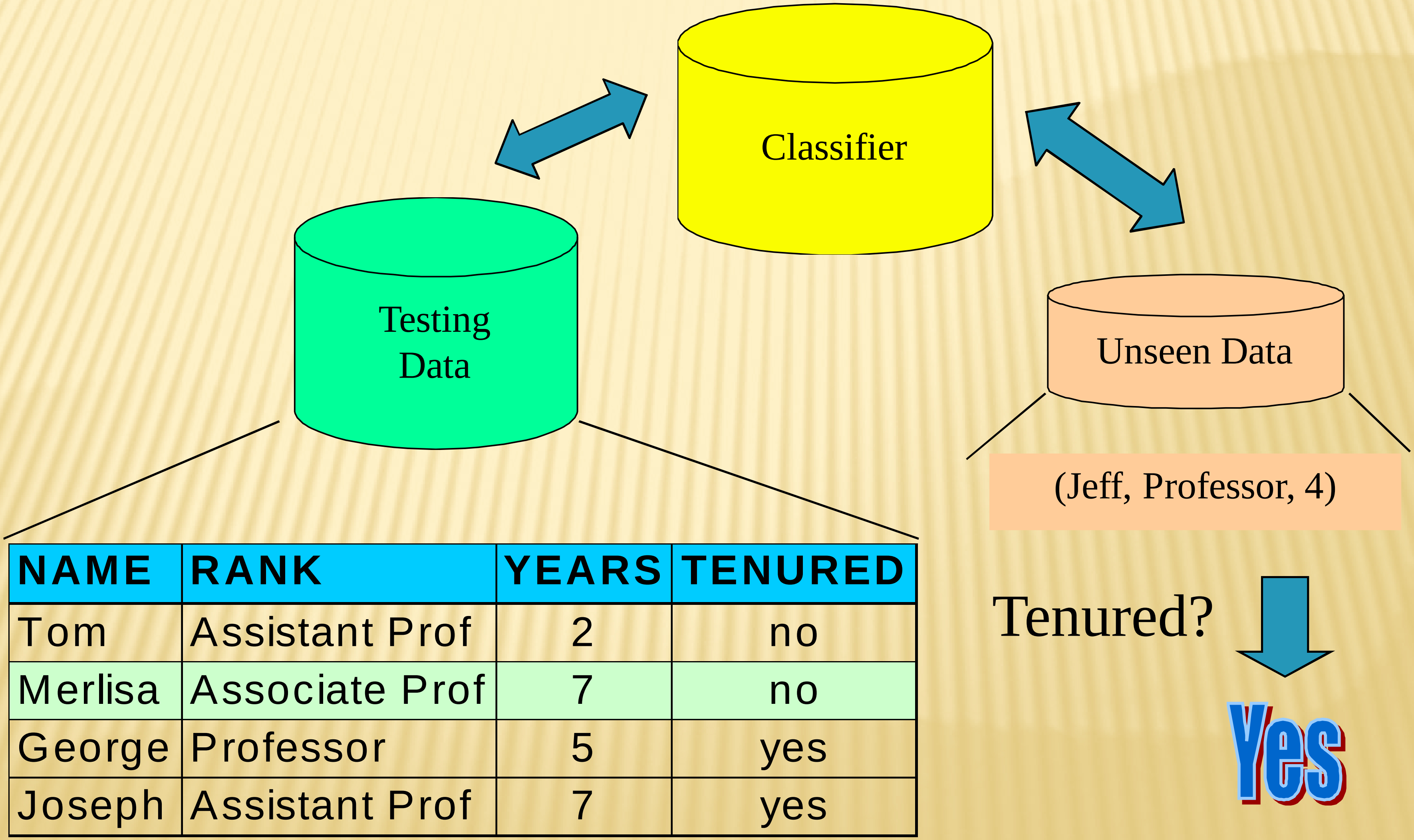
CLASSIFICATION PROCESS (1): MODEL CONSTRUCTION



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

CLASSIFICATION PROCESS (2): USE THE MODEL IN PREDICTION



Classifier

Testing Data

Unseen Data

(Jeff, Professor, 4)

Tenured?

Yes

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

SUPERVISED VS. UNSUPERVISED LEARNING

- Supervised learning (classification)
  - 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 (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

WHAT IS PREDICTION?

- Prediction is similar to classification
  - First, construct a model
  - Second, use model to predict unknown value
    - Major method for prediction is regression
      - Linear and multiple regression
      - Non-linear regression
- Prediction is different from classification
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions

PREDICTIVE MODELING IN DATABASES

- Predictive modeling: Predict data values or construct generalized linear models based on the database data.
- One can only predict value ranges or category distributions
- Method outline:
  - Minimal generalization
  - Attribute relevance analysis
  - Generalized linear model construction
  - Prediction
- Determine the major factors which influence the prediction
  - Data relevance analysis: uncertainty measurement, entropy analysis, expert judgement, etc.
- Multi-level prediction: drill-down and roll-up analysis

REGRESS ANALYSIS AND LOG-LINEAR MODELS IN PREDICTION

- Linear regression: $Y = \alpha + \beta X$
  - Two parameters , α and β specify the line and are to be estimated by using the data at hand.
  - using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$.
  - Many nonlinear functions can be transformed into the above.

ISSUES REGARDING CLASSIFICATION AND PREDICTION

ISSUES (1): DATA PREPARATION

- Data cleaning
  - Preprocess data in order to reduce noise and handle missing values
- Relevance analysis
  - Remove the irrelevant or redundant attributes.
  - Correlation analysis can be used to identify whether any two given attributes are statistically related.
  - Attribute subset selection can be used in these cases to find a reduced set of attributes such that resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.

ISSUES (1): DATA PREPARATION

  - Hence, relevance analysis, in the form of correlation analysis and attribute subset selection, can be used to detect attributes that do not contribute to the classification or prediction task
- Data transformation
  - Generalize and/or normalize data
  - Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as -1:0 to 1:0, or 0:0 to 1:0. In methods that use distance measurements.
  - The data can also be transformed by *generalizing it to higher-level concepts.* This is particularly useful for continuous valued attributes. For example, numeric values for the attribute *income can be generalized* to discrete ranges, such as *low, medium, and high*.

COMPARING CLASSIFICATION AND PREDICTION METHODS

Classification & Prediction methods can be compared and Evaluated according following criteria:

- Predictive accuracy
  - The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data (i.e., tuples without class label information).
  - the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.
- Speed and scalability
  - time to construct the model
  - time to use the model

COMPARING CLASSIFICATION AND PREDICTION METHODS

- Robustness
  - handling noise and missing values
- Scalability
  - efficiency in disk-resident databases
- Interpretability:
  - understanding and insight provded by the model
- Goodness of rules
  - decision tree size
  - compactness of classification rules