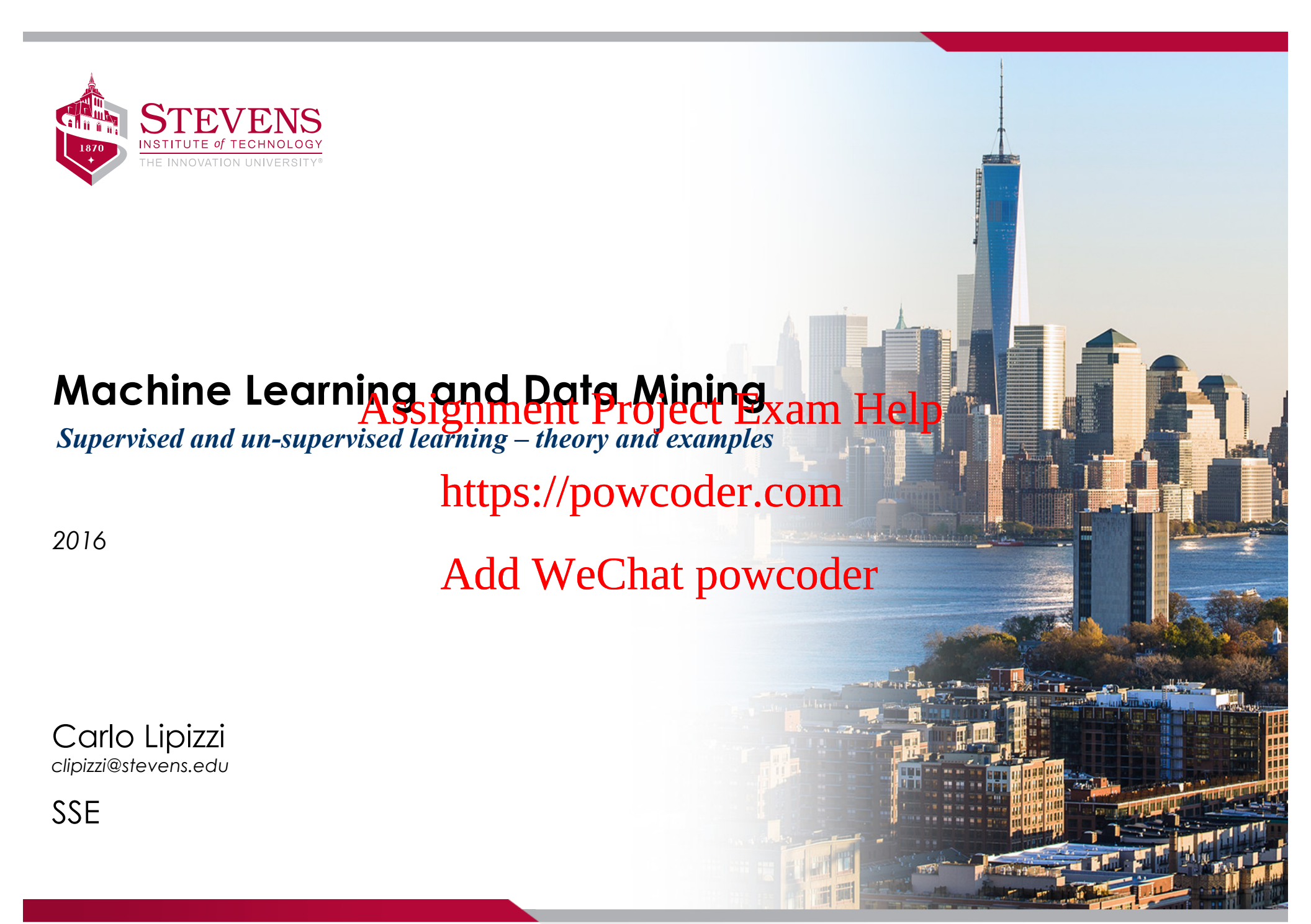


Machine Learning and Data Mining

Supervised and un-supervised learning – theory and examples

2016

Carlo Lipizzi
clipizzi@stevens.edu

SSE

Machine learning and our focus

- Like human learning from past experiences
- A computer does not have "experiences"
- A computer system learns from data, which represent some "past experiences" of an application domain
- Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g.: approve or not-approved, and high-risk or low risk
- The task is commonly called: Supervised learning, classification, or inductive learning

The data and the goal

- Data: A set of data records (also called examples, instances or cases) described by
  - k attributes: A_1, A_2, … A_k.
  - a class: Each example is labelled with a pre-defined class
- Goal: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances

An example: data (loan application)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

An example: the learning task

- Learn a classification model from the data
- Use the model to classify future loan applications into
  - Yes (approved) and
  - No (not approved)
- What is the class for following case/instance?

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

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 data, the task is to establish the existence of classes or clusters in the data

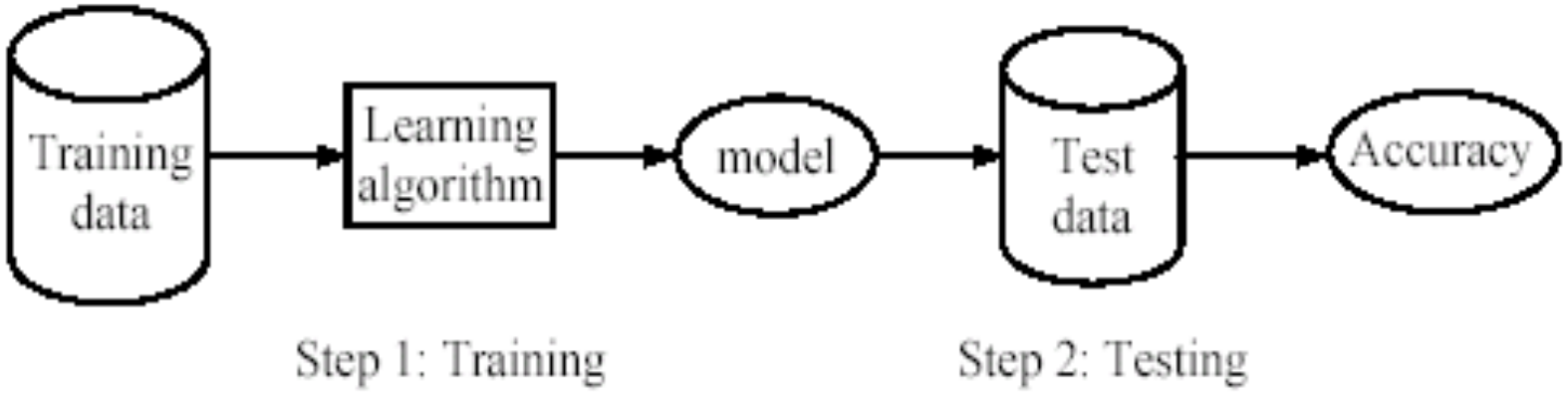
Classification: Definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class
- Find a model for class attribute as a function of the values of other attributes
- Goal: previously unseen records should be assigned a class 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

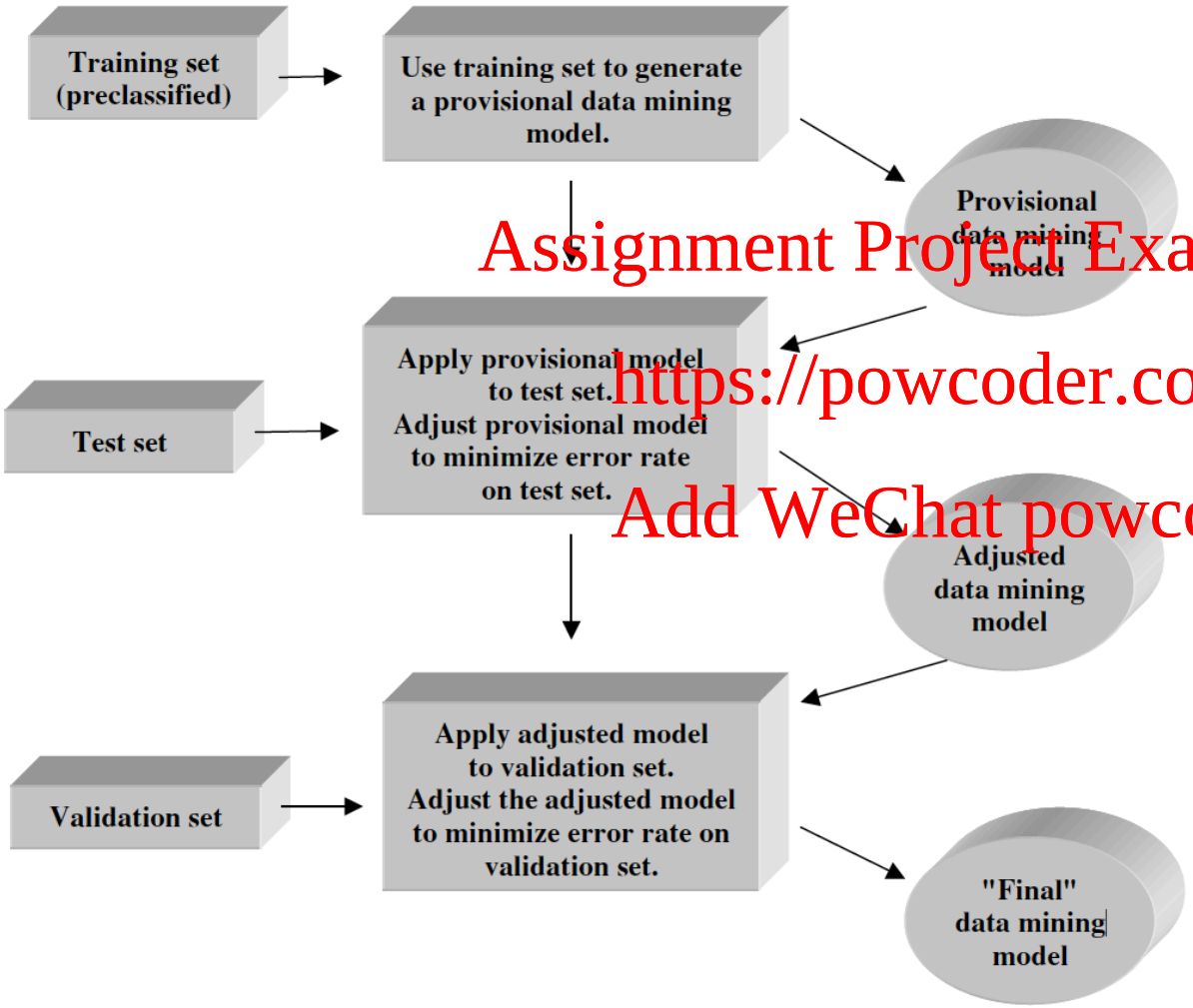
Supervised learning process: two steps

- Learning (training): Learn a model using the training data
- Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$



Supervised learning process: Three steps



- A third step can be added, splitting the original dataset in 3 parts: **Training**, **Testing** and **Validate**
- This is the default for Rattle

Learning in Data Mining

- Given
  - a data set D
  - a task T
  - a performance measure M
  - a computer system is said to learn from D to perform the task T if after learning the system's performance on T improves as measured by M
- In other words, the learned model helps the system to perform T better as compared to no learning

An example

- **Data**: Loan application data
- **Task**: Predict whether a loan should be approved or not
- **Performance measure**: accuracy

No learning: classify all future applications (test data) to the majority class (i.e.: Yes):

Accuracy = 9/15 = 60%

- We can do better than 60% with learning

Exercise 1 "Blind" use of Rattle

- Using the *cars full.txt* dataset
- Question:
  - What are the most influential variables on "time to 60"? Explain

Fundamental assumption of learning

Assumption: The distribution of training examples is identical to the distribution of test examples (including future unseen examples)

- In practice, this assumption is often violated to certain degree
- Strong violations will clearly result in poor classification accuracy
- To achieve good accuracy on the test data, training examples must be sufficiently representative of the test data

Classification: Definition

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

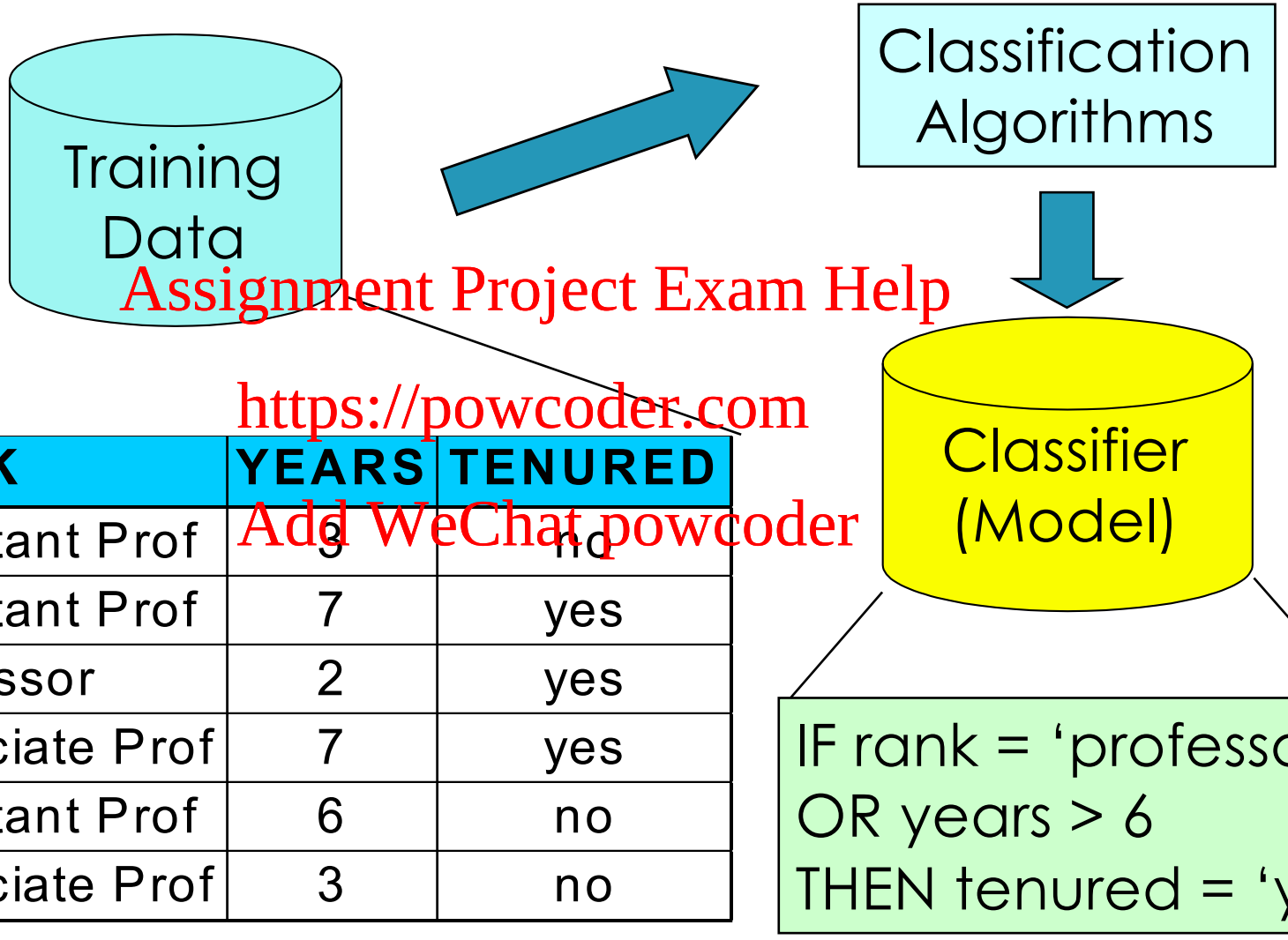
Prediction Problems: Classification vs. Numeric Prediction

- Classification :
  - predicts categorical class labels (discrete or nominal)
  - 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
- Numeric Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values
- Typical applications
  - Credit/loan approval:
  - Medical diagnosis: if a tumor is cancerous or benign
  - Fraud detection: if a transaction is fraudulent
  - Web page categorization: which category it is

Classification—A Two-Step Process

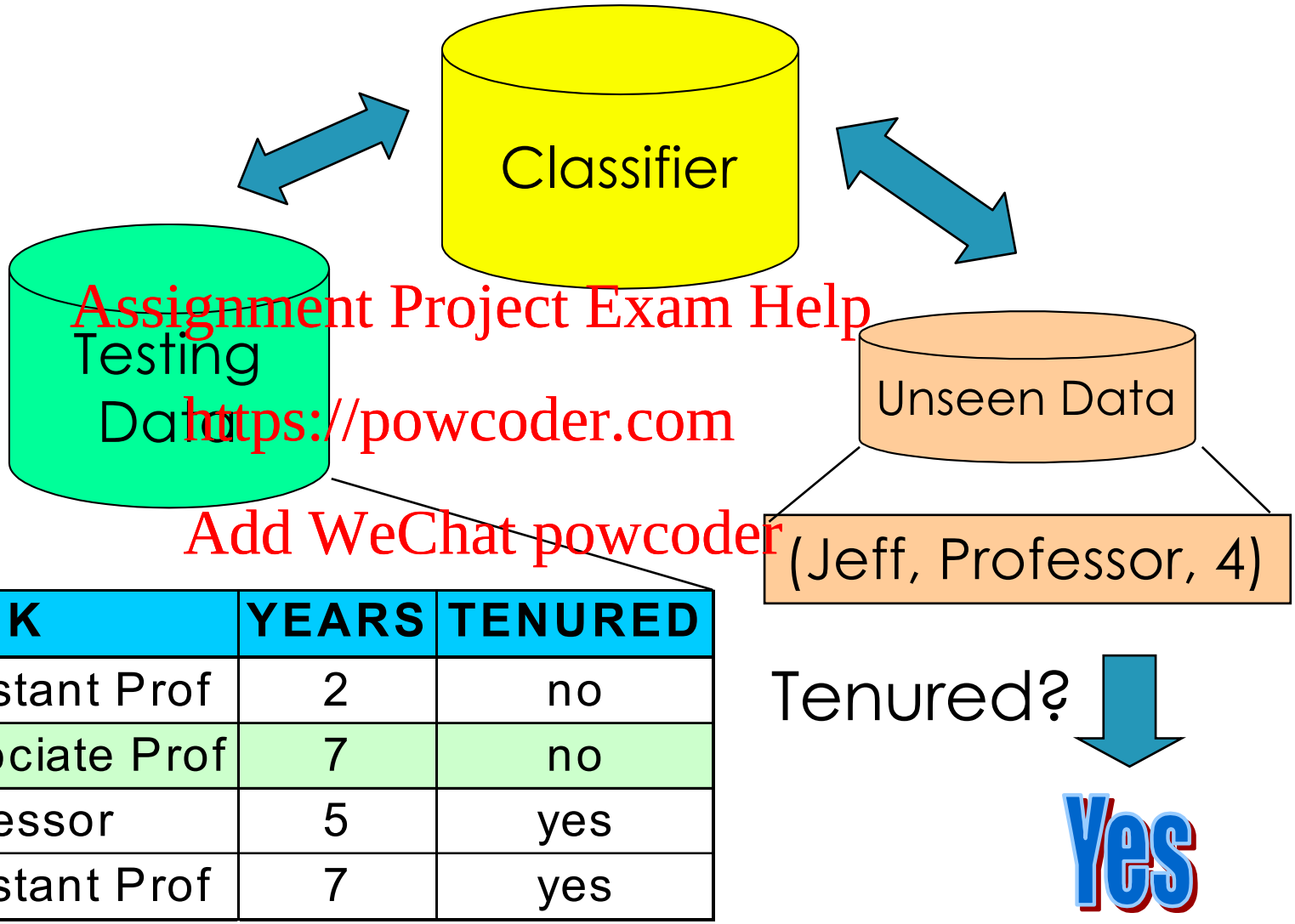
- Model construction: describing a set of predetermined classes
  - Each sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of samples used for model construction is training set
  - The model is represented as classification rules, decision trees, 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
  - If the accuracy is acceptable, use the model to classify new data

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

Using the Model in Prediction

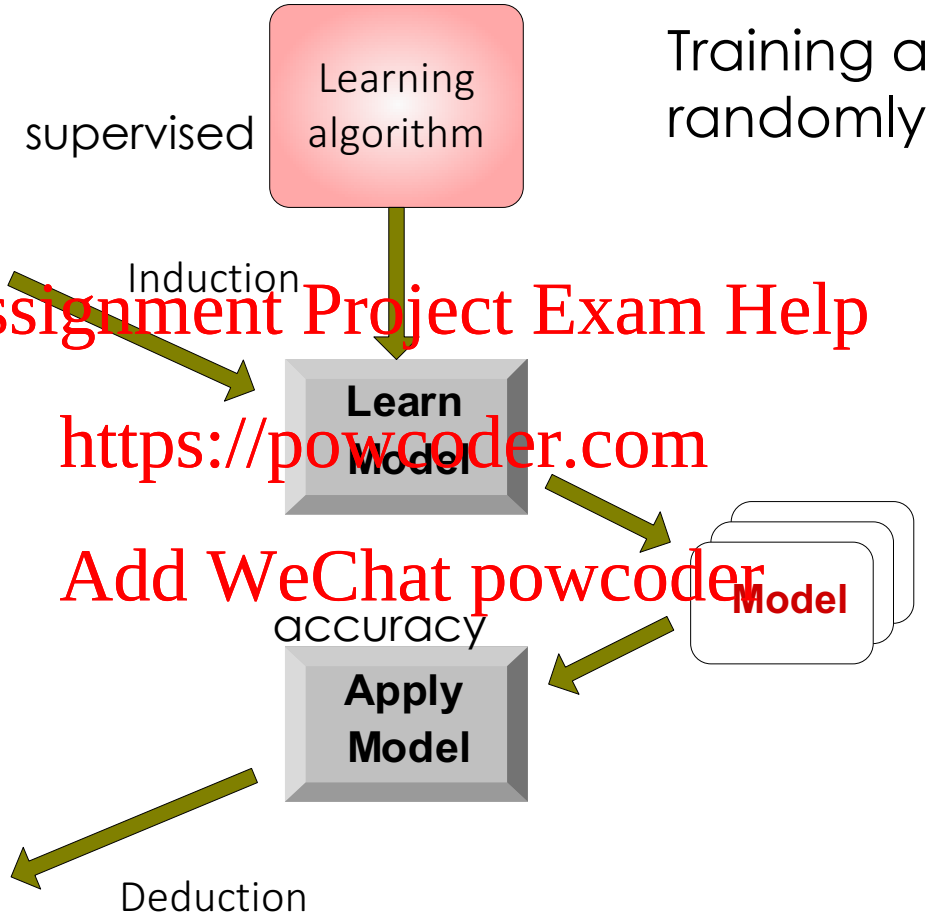


NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

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



Training and Test set are randomly sampled

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

Find a mapping OR function that can predict class label of given sample X

Classification Techniques

- Decision Tree based Methods
- Bayes Classification Methods
- Rule-based Methods
- Nearest-Neighbor Classifier
- Artificial Neural Networks
- Support Vector Machines

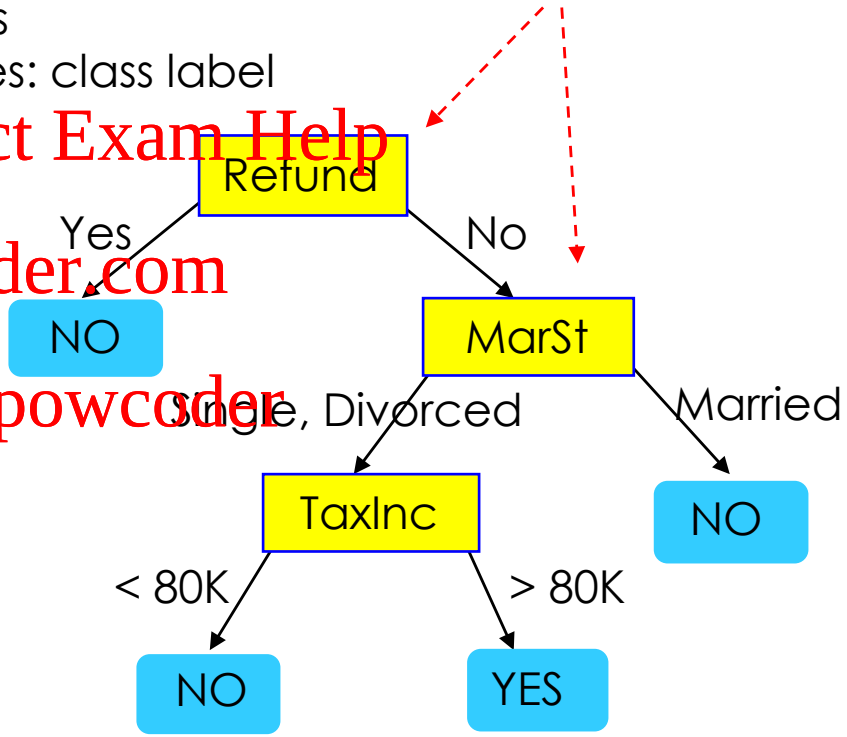
Example of a Decision Tree

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

categorical, categorical, continuous, class

Training Data

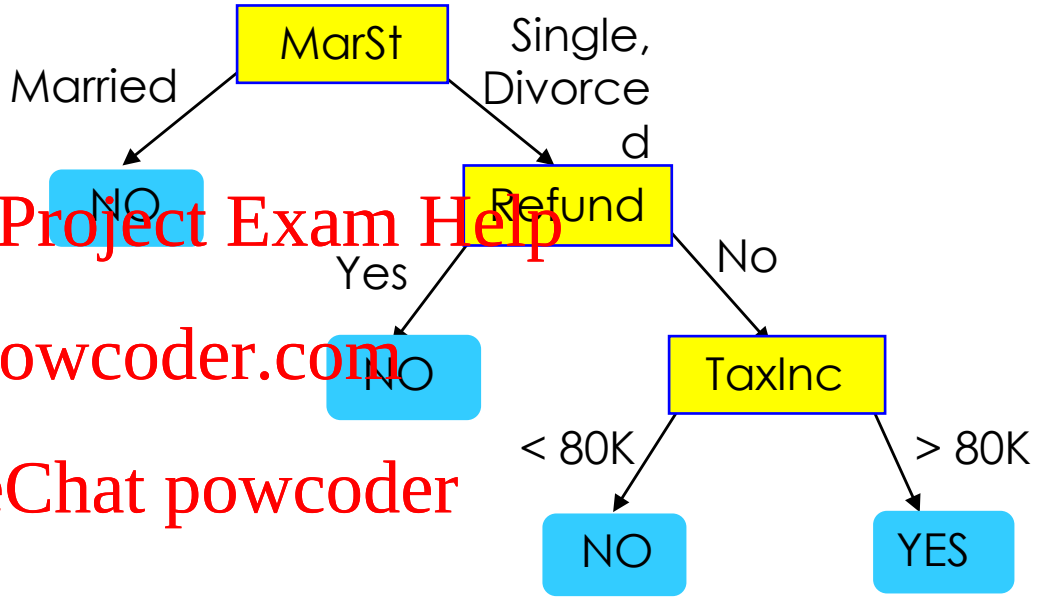
Root node:
Internal nodes: attribute tests (*Splitting Attributes*)
conditions
Leaf nodes: class label



Model: Decision Tree

Another Example of Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
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

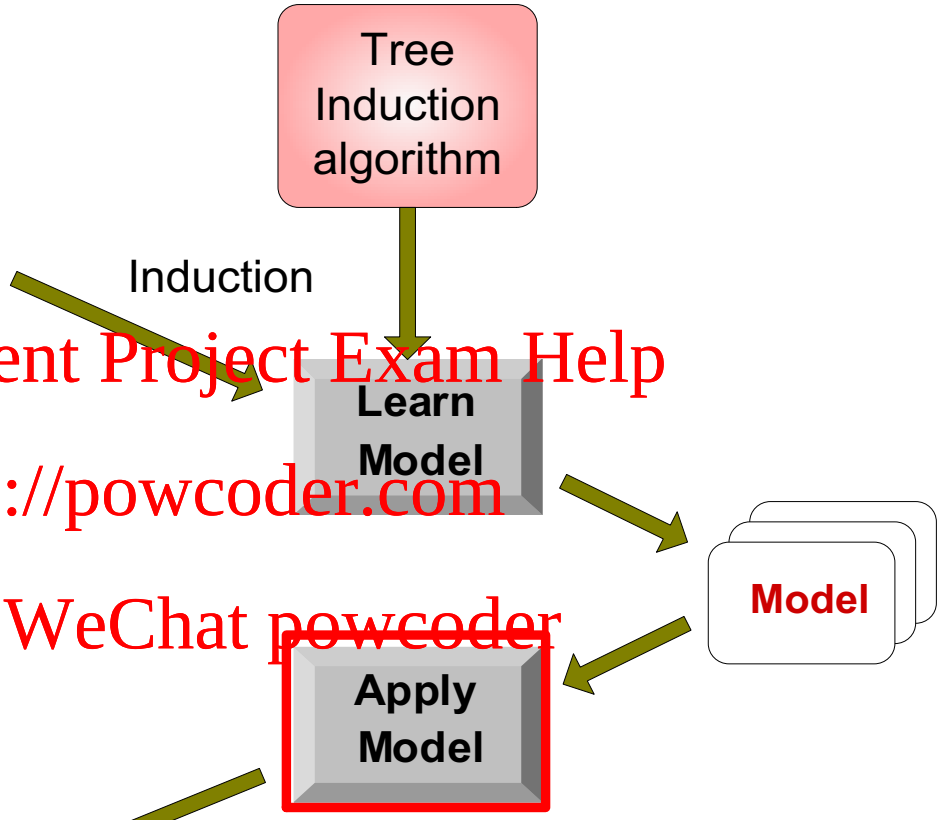


There could be more than one tree that fits the same data

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

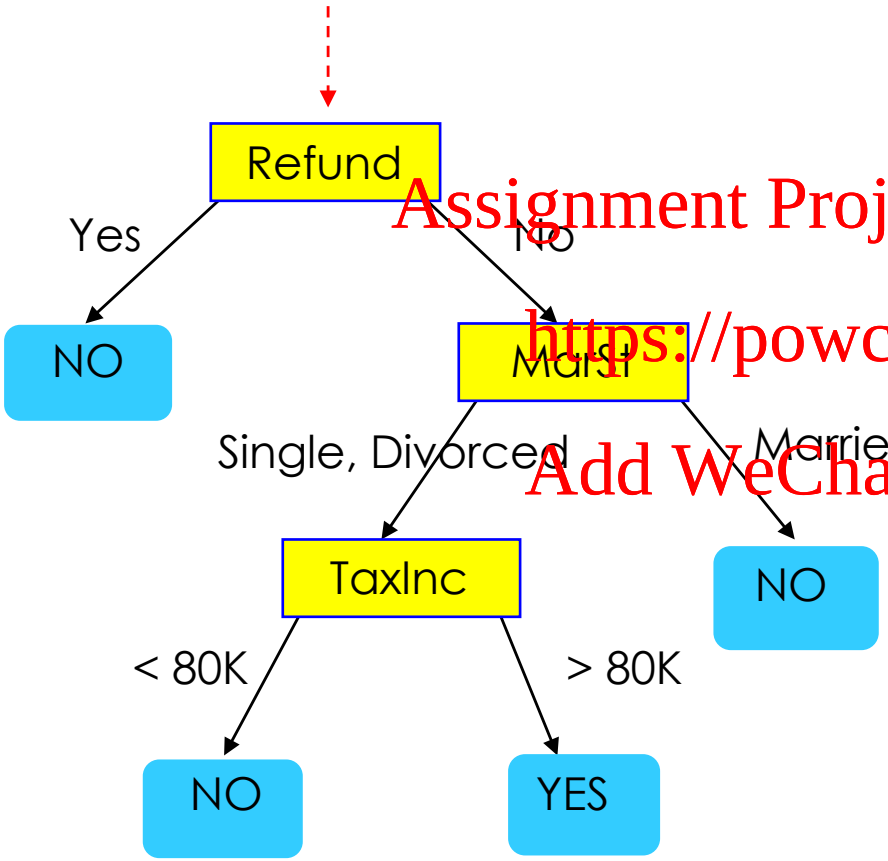


Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Apply Model to Test Data

Start from the root of tree.



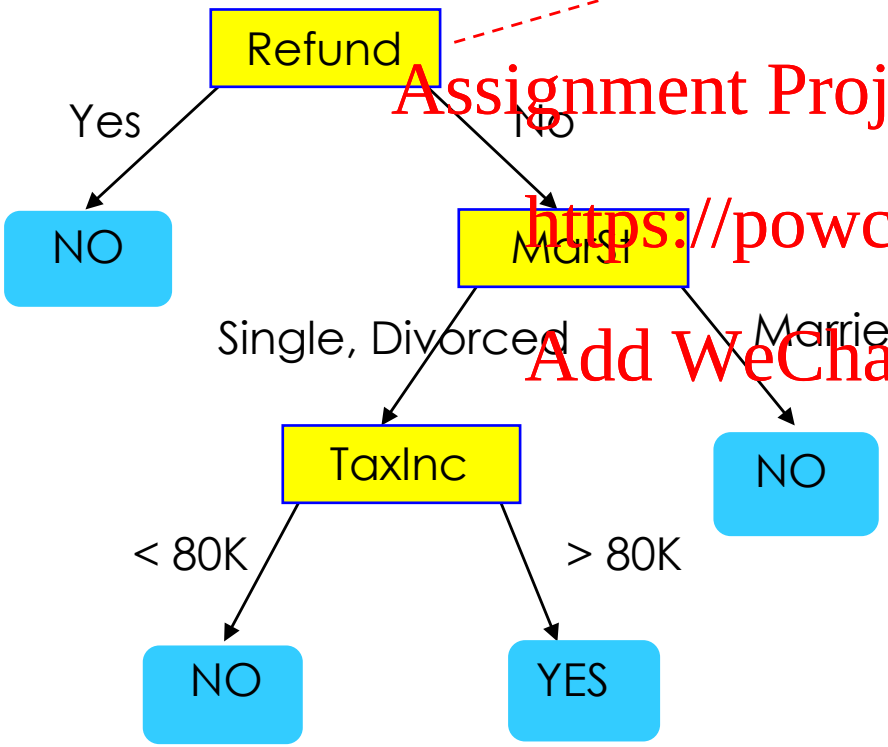
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

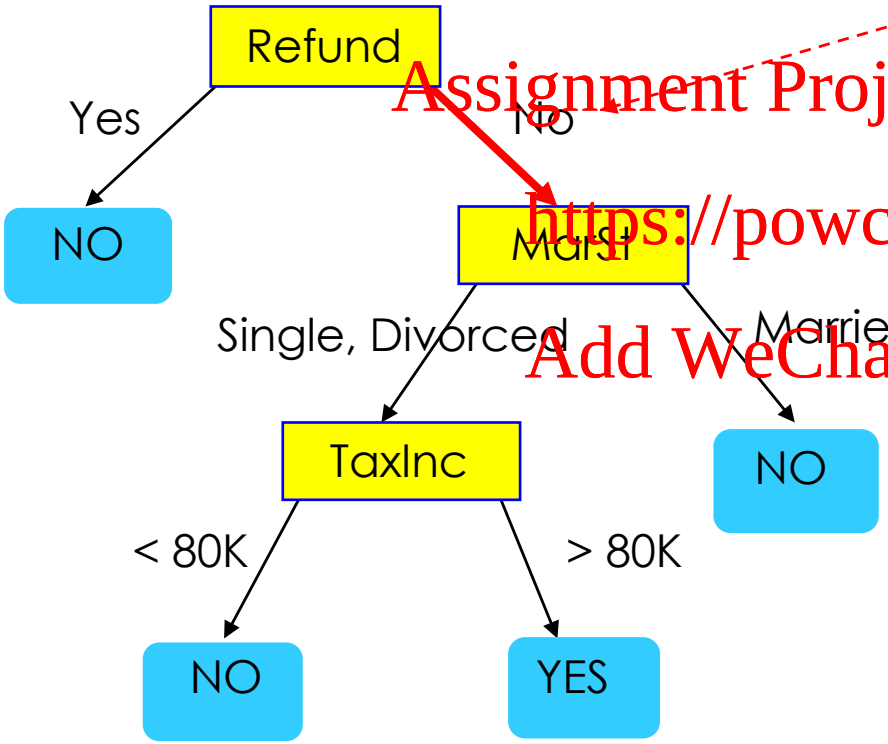
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

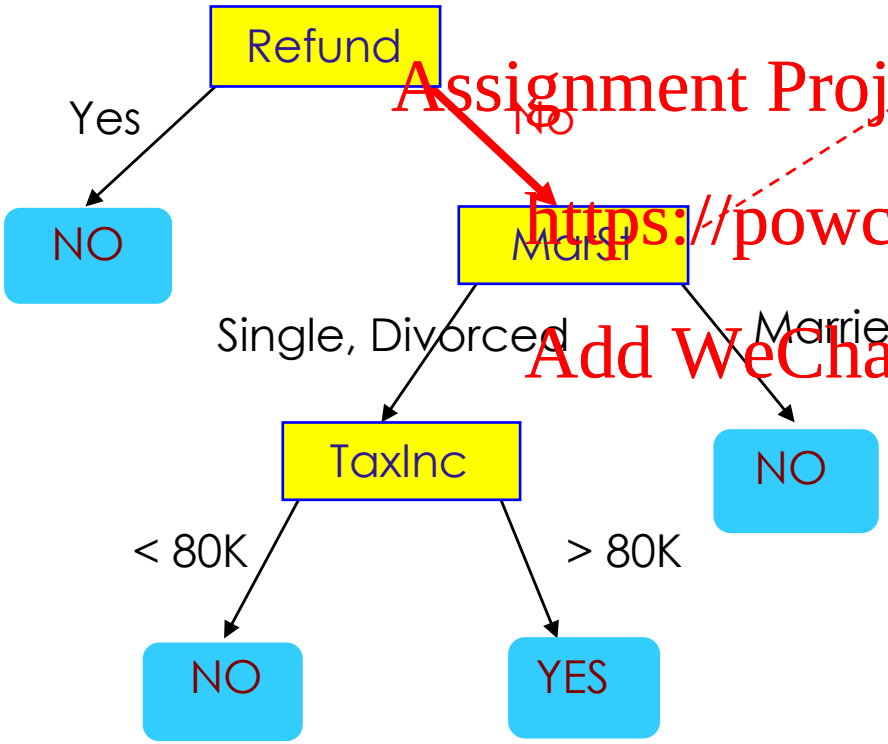
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

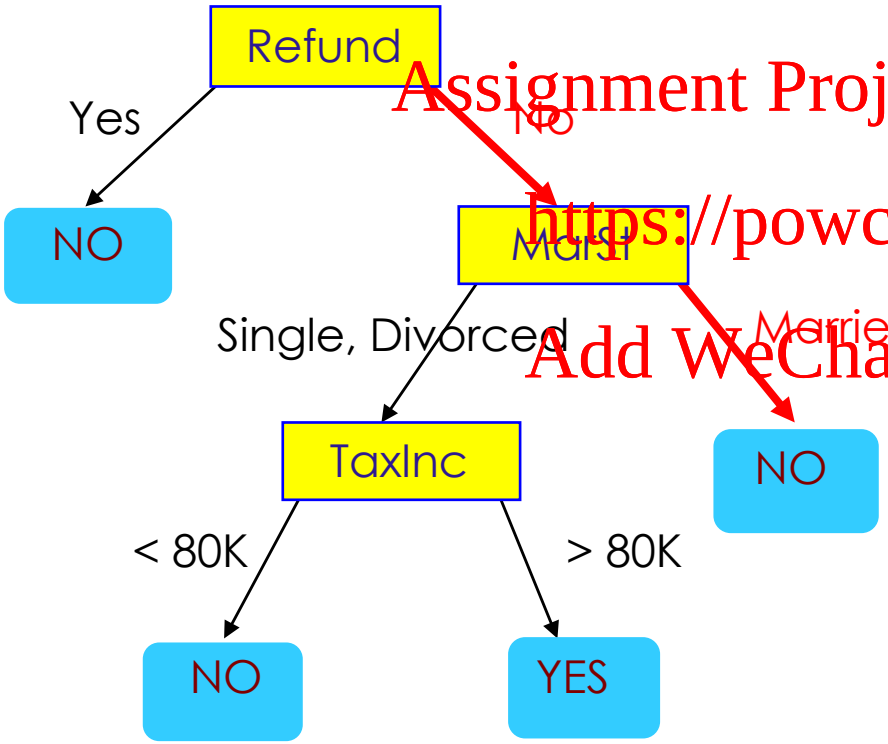
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

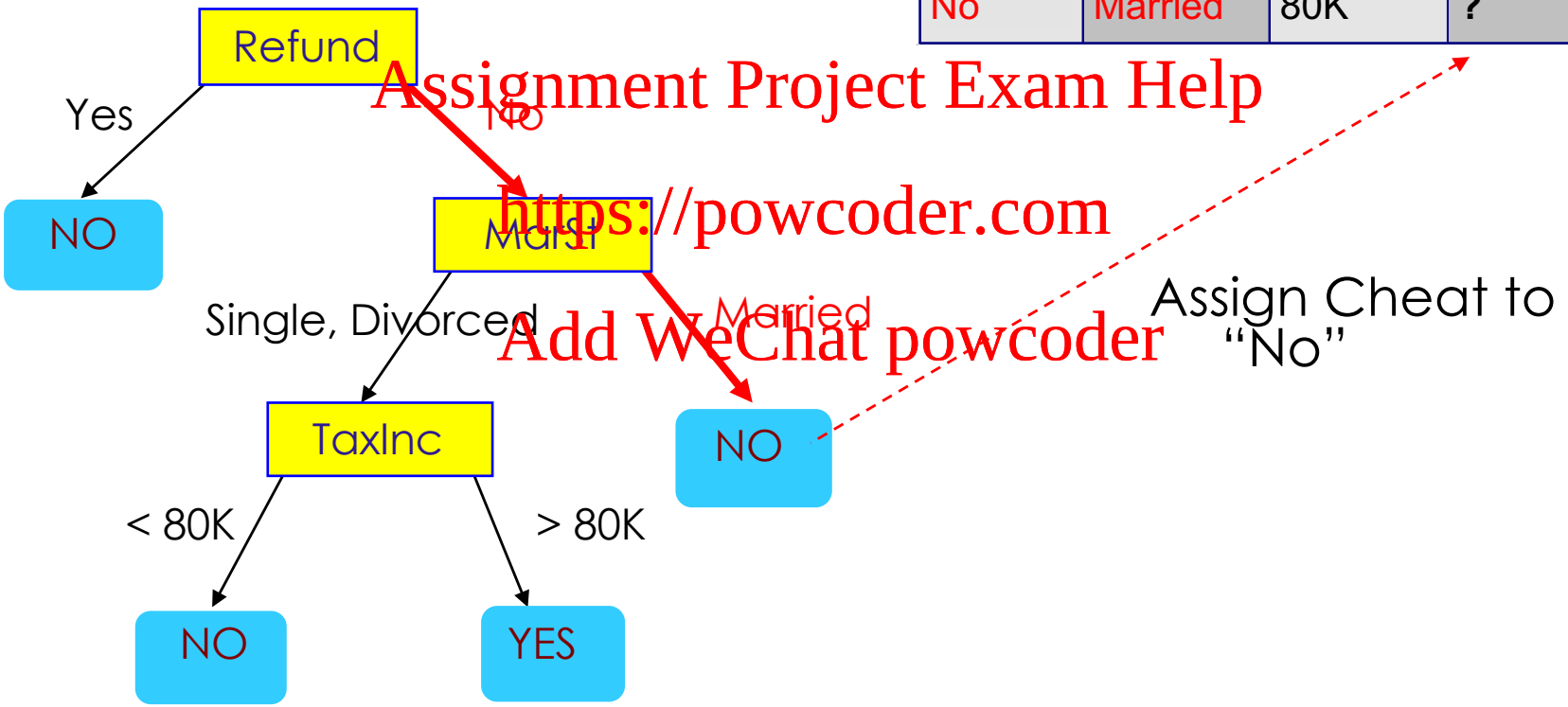
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

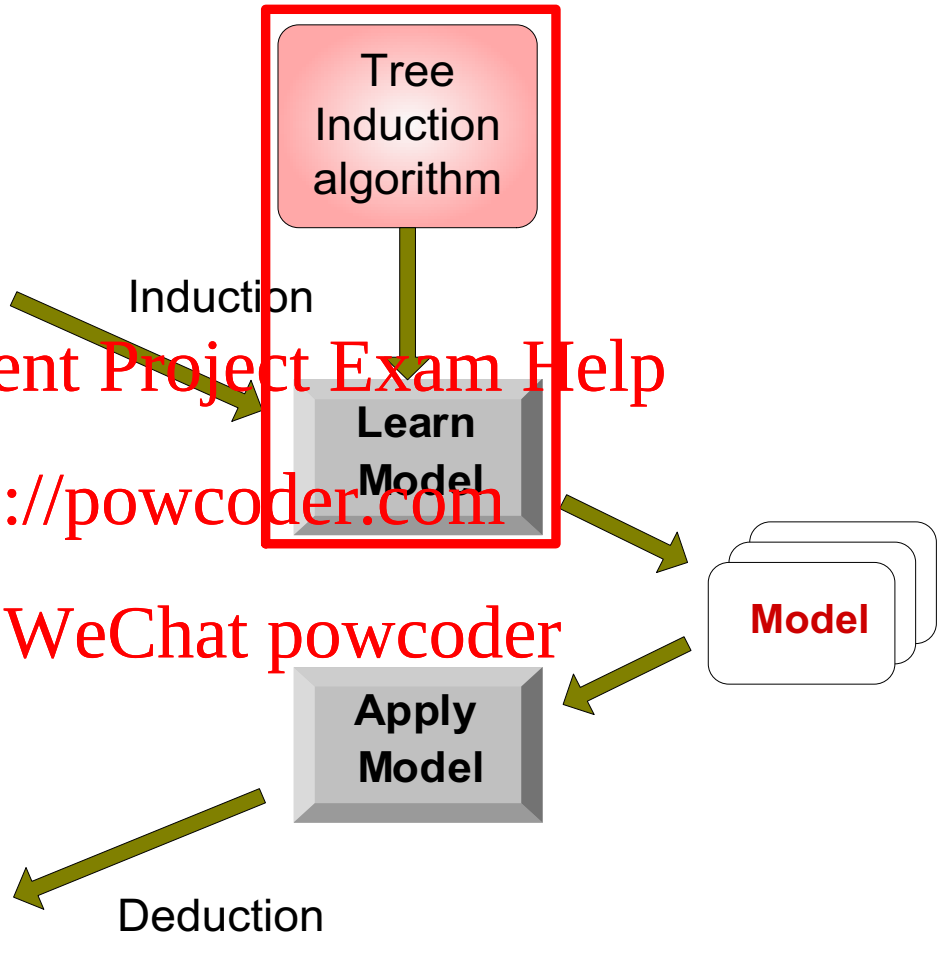
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



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

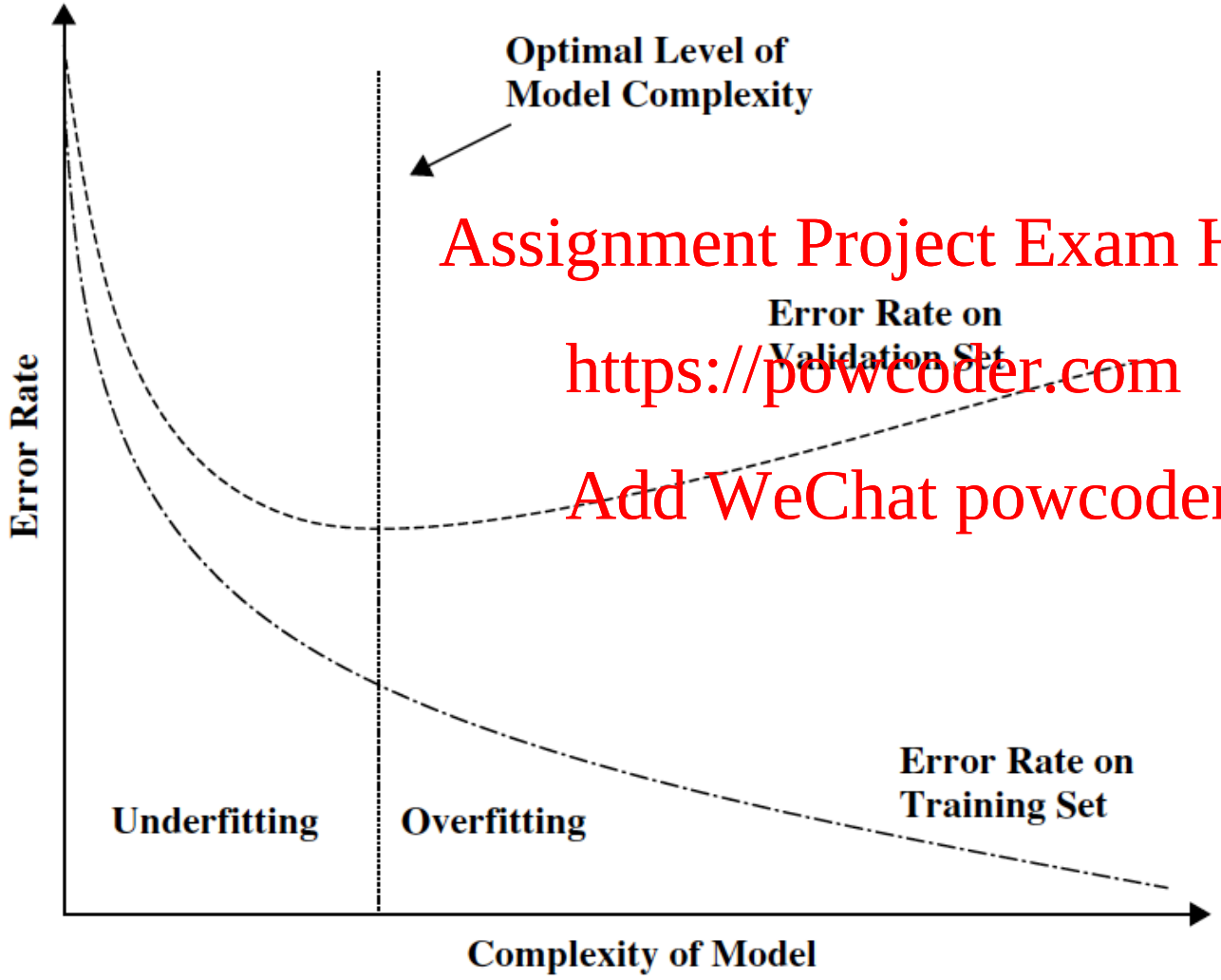
Decision Tree Induction

- Many Algorithms:
  - Hunt's Algorithm
  - ID3, C4.5
  - CART
  - SLIQ,SPRINT

Exercise 2 - “Blind” use of Rattle

- Using the *audit.csv* dataset
- Questions:
  - Examine the data:
    - Are there missing values or outliers?
    - Can you tell anything about the quality of data?
  - What are the most influential variables on “Target Adjusted”? Explain

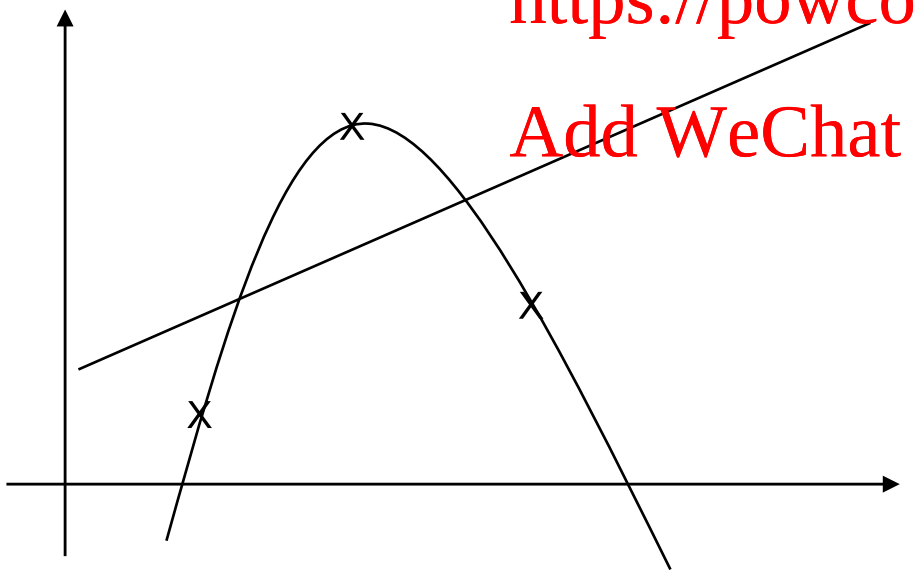
Optimizing the model



- The optimal level of model complexity is at the minimum error rate on the validation set

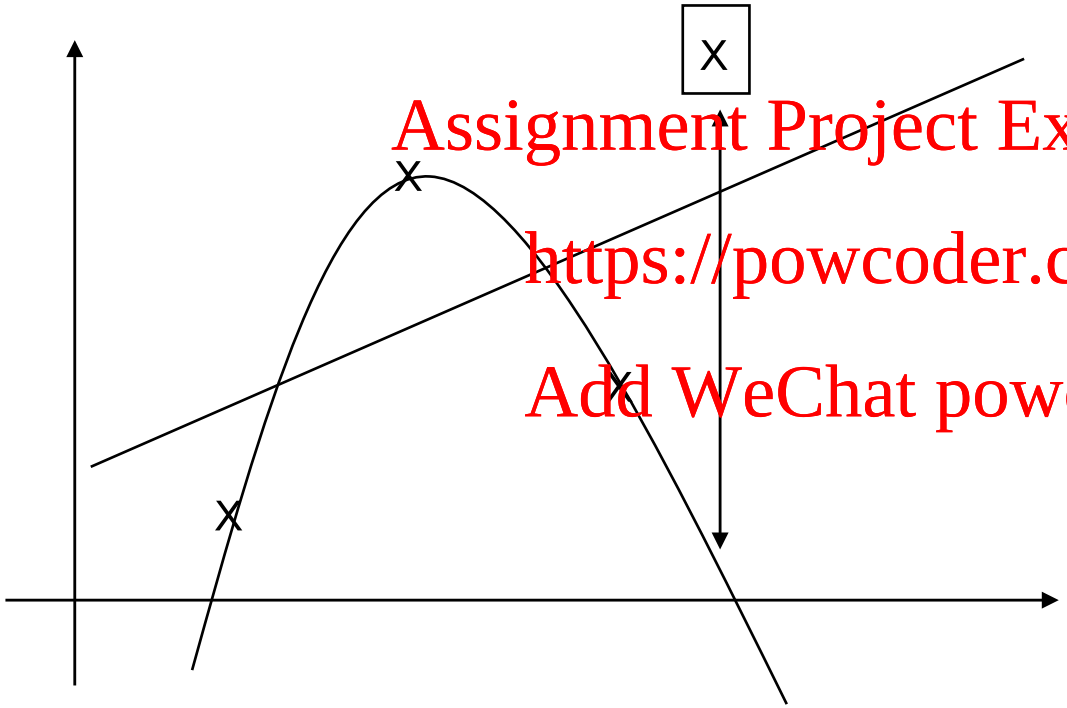
Overfitting: Definition

- The algorithm finds model that fits the training data and performs well on the trained data, but performs poorly on real world new data it has not seen before: the algorithm may pick up details in the data that are characteristics of the training sample, but not the actual problem being modeled



- Assume that the 3 x's represent the noisy data from a linear model, represented by the straight line in the figure. If we fit the 3 points to a quadratic model, we could get a perfect fit, represent by the concave curve

Overfitting: Why is a problem



- The overfitted model has very poor predictive power
- For example, the new point, represented by the "x" in the box, is faraway from the prediction by the overfitted quadratic model. In contrast, the difference between the linear model and the new points is much smaller

Overfitting: Causes and Remediation

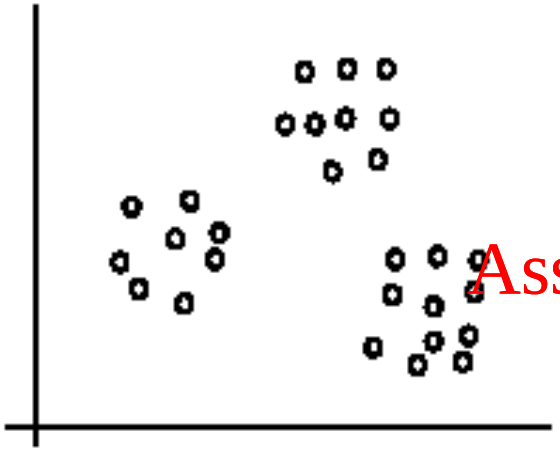
- Causes:
  - Noise in the system -> Greater variability in data
  - Complex model -> many parameters -> higher degree of freedom -> greater variability

- Remediation:
  - For any algorithm:
    - Read your data and your model, with a subject matter expert view
    - Split the training data in several parts, using each part but one as training
  - For classification trees pre or post "pruning" methods: reduce the number of splits forcing the split threshold or consider a limited number of splits after the model has been created

Clustering

- Clustering is a technique for finding similarity groups in data, called clusters
  - it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters
- Clustering is an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given
- Due to historical reasons, clustering is often considered synonymous with unsupervised learning

Examples



- The data set on the left has three natural groups of data points, i.e., 3 natural clusters
- **Marketing**: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records
- **Biology**: classification of plants and animals given their features

- **Insurance**: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds
- **City-planning**: identifying groups of houses according to their house type, value and geographical location
- **Earthquake studies**: clustering observed earthquake epicenters to identify dangerous zones
- **WWW**: document classification; clustering weblog data to discover groups of similar access patterns

Additional Examples

- Groups people of similar sizes together to make "small", "medium" and "large" T-Shirts
  - Tailor-made for each person: too expensive
  - One-size-fits-all: does not fit all
- Given a collection of text documents, we want to organize them according to their content similarities
  - To produce a topic hierarchy
- Clustering is one of the most utilized data mining techniques
  - It has a long history, and used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.
  - In recent years, due to the rapid increase of online documents, text clustering becomes important

Aspects of clustering

- A clustering algorithm
  - Partitional clustering
  - Hierarchical clustering
  - …
- A distance (similarity, or dissimilarity) function
- Clustering quality
  - Inter-clusters distance ⇒ maximized
  - Intra-clusters distance ⇒ minimized
- The quality of a clustering result depends on the algorithm, the distance function, and the application

Exercise 3 - "Blind" use of Rattle

- Using the *cereals.txt* dataset
- Questions:
  - Import the dataset into Rattle (possible intermediate step)
  - Examine the data
  - Create clusters and read the results