Data Mining: Overfitting

Lecture Notes for Chapter 3

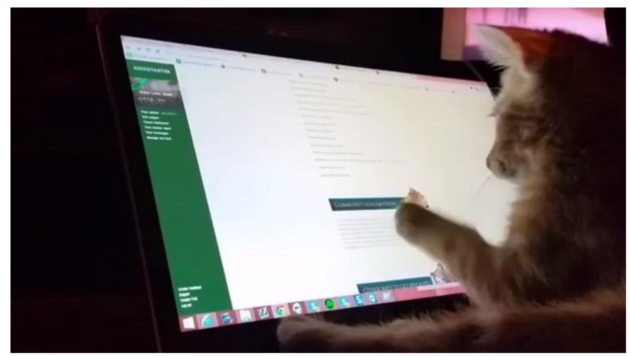
Data Mining

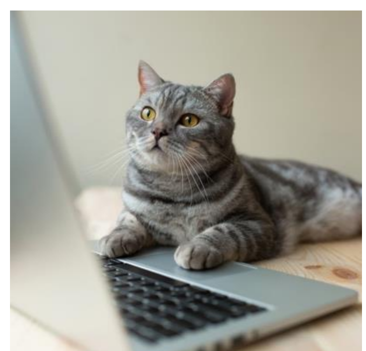
https://ml-graph.github.io/winter-2025/

Yu Wang, Ph.D.
yuwang@uoregon.edu
Assistant Professor
Computer Science
University of Oregon
CS 453/553 – Winter 2025

Course Lecture is very heavily based on
"Introduction to Data Mining"
by Tan, Steinbach, Karpatne, Kumar

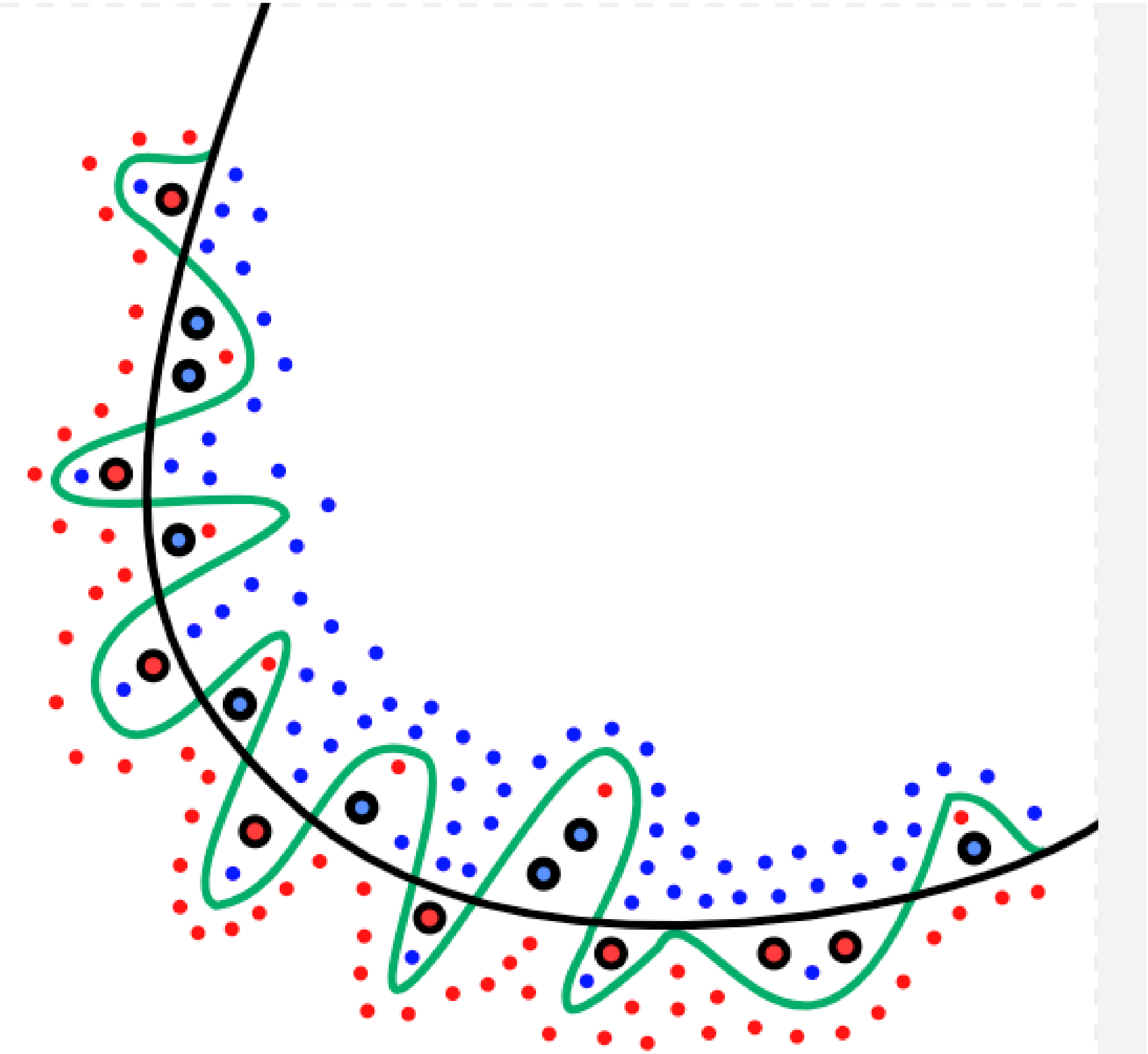
Example





Example

Overfitting

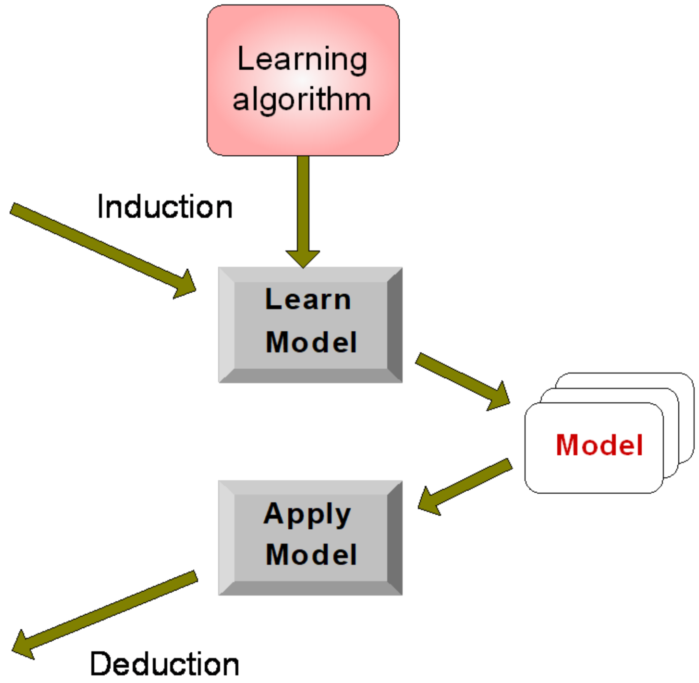


Classification Errors

- **Training errors:** **Errors committed on the training set**
- **Test errors:** **Errors committed on the test set**
- **Generalization errors:** **Expected error of a model over random selection of records from same distribution**

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

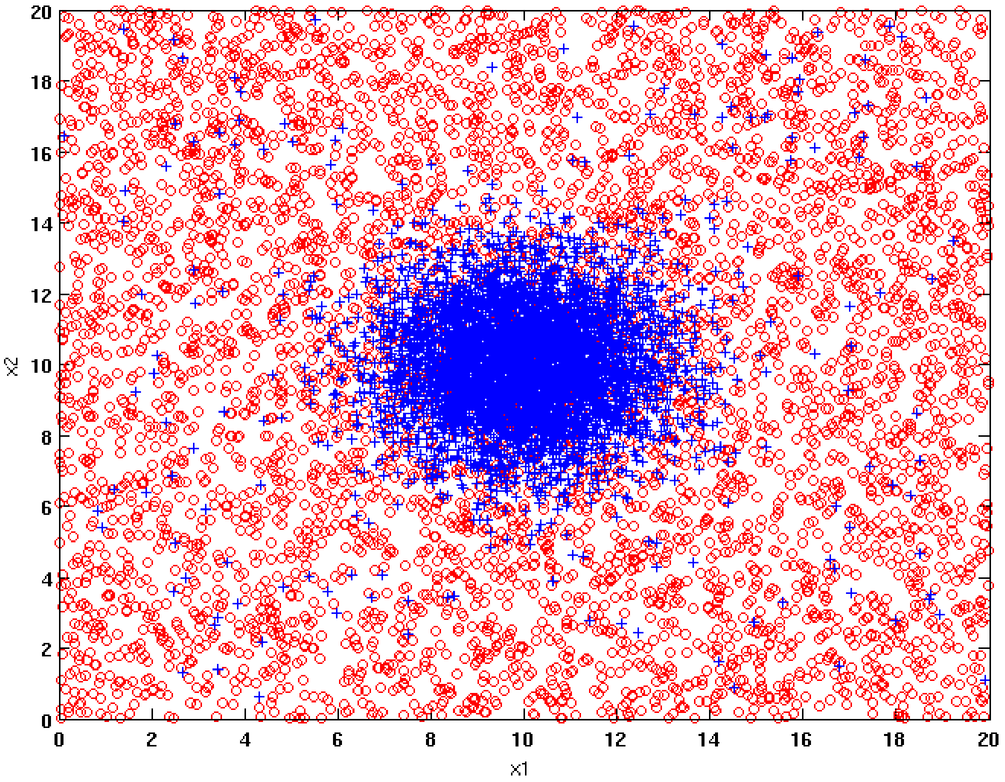
Training Set



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

Example Data Set



Two class problem:

+ : 5400 instances

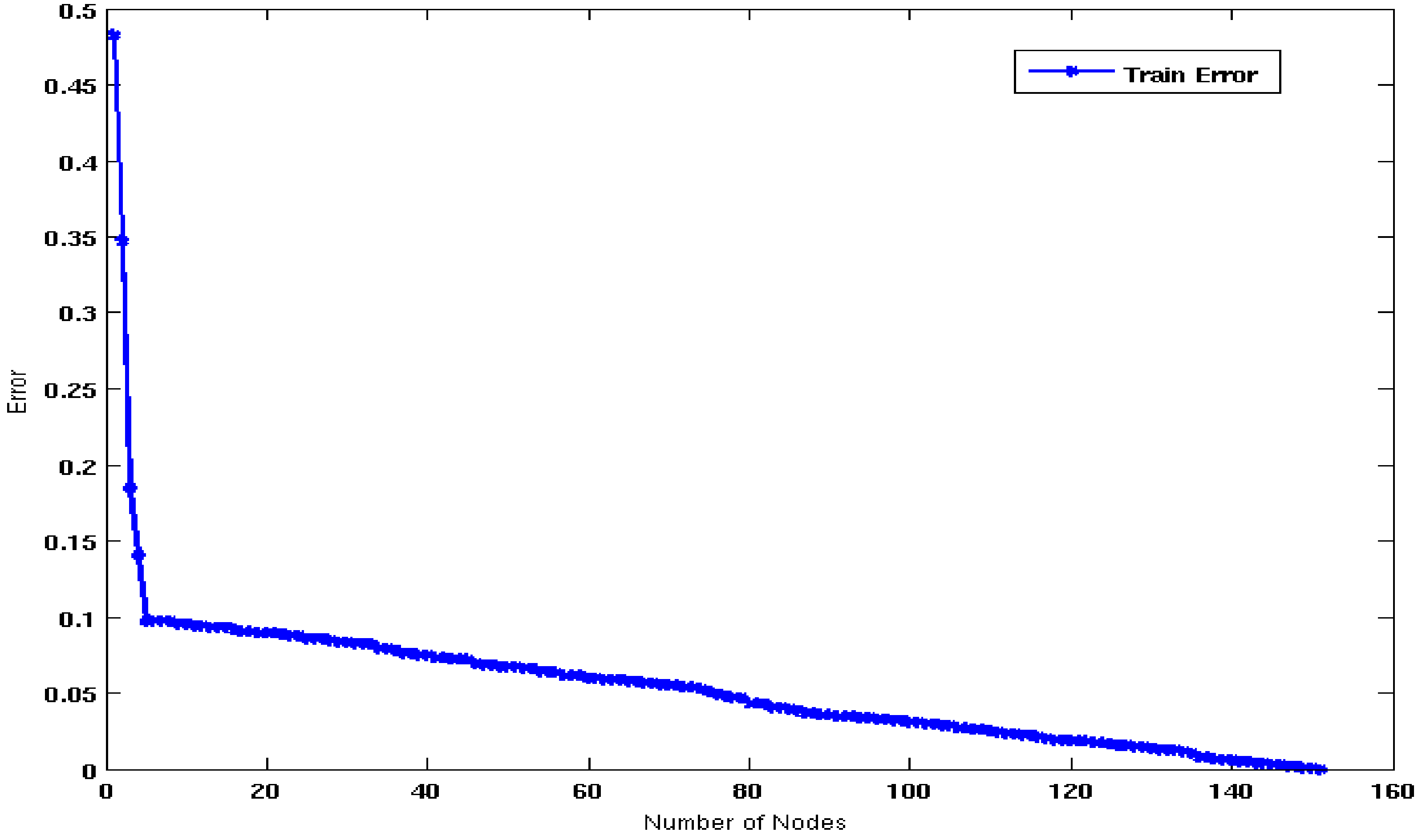
- 5000 instances generated from a Gaussian centered at (10,10)
- 400 noisy instances added

o : 5400 instances

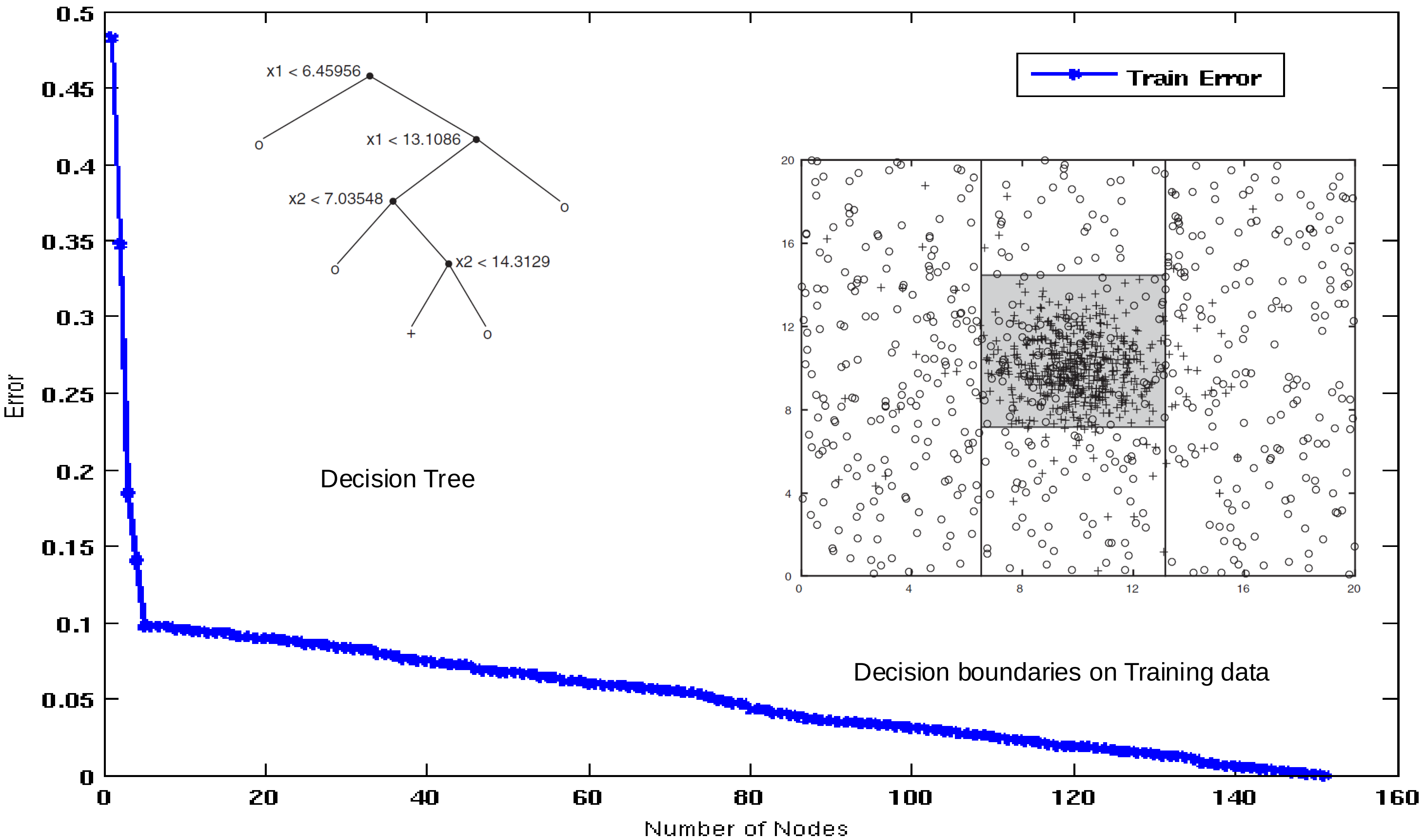
- Generated from a uniform distribution

10 % of the data used for training and 90% of the data used for testing

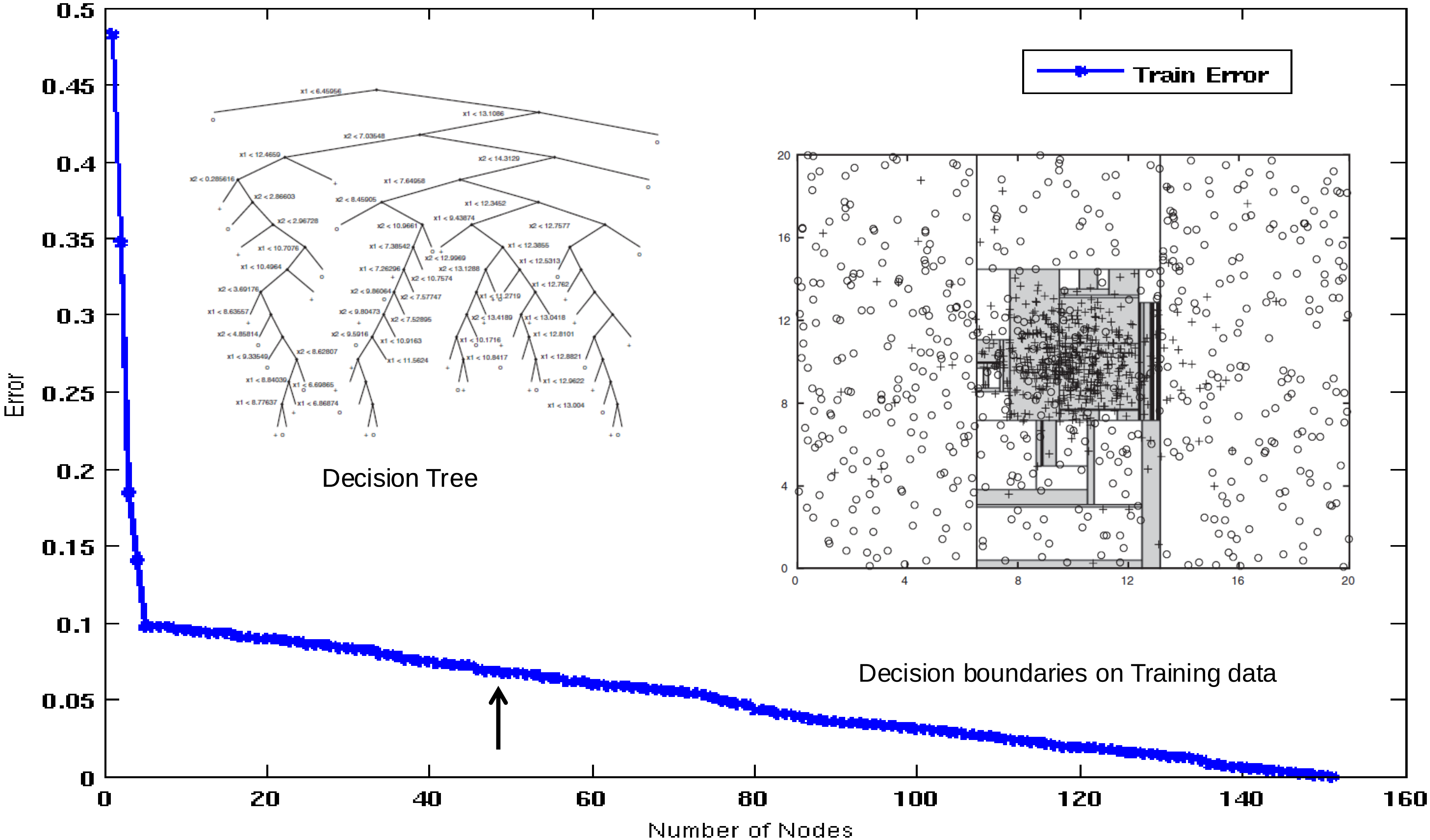
Increasing number of nodes in Decision Trees



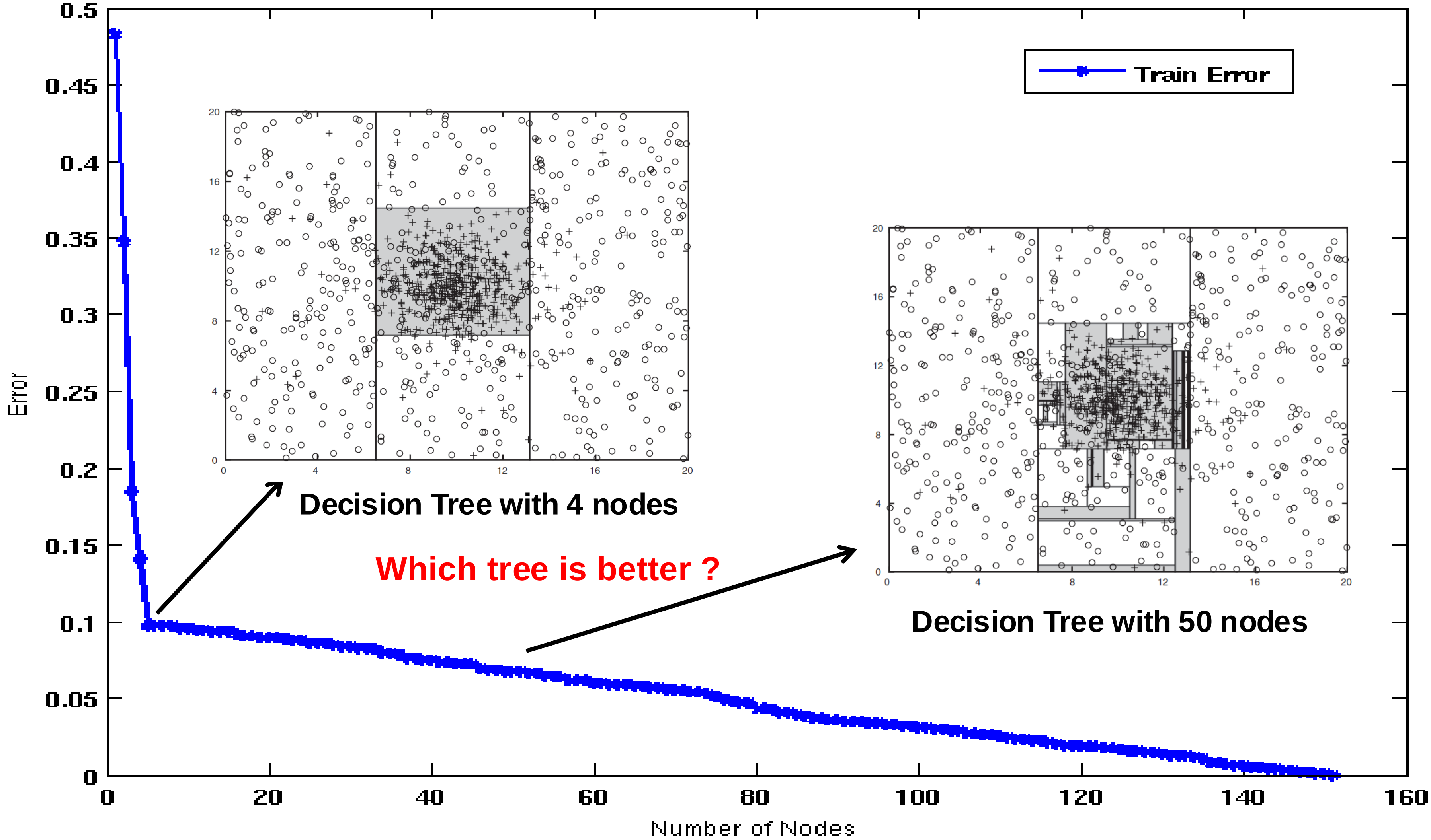
Decision Tree with 4 nodes



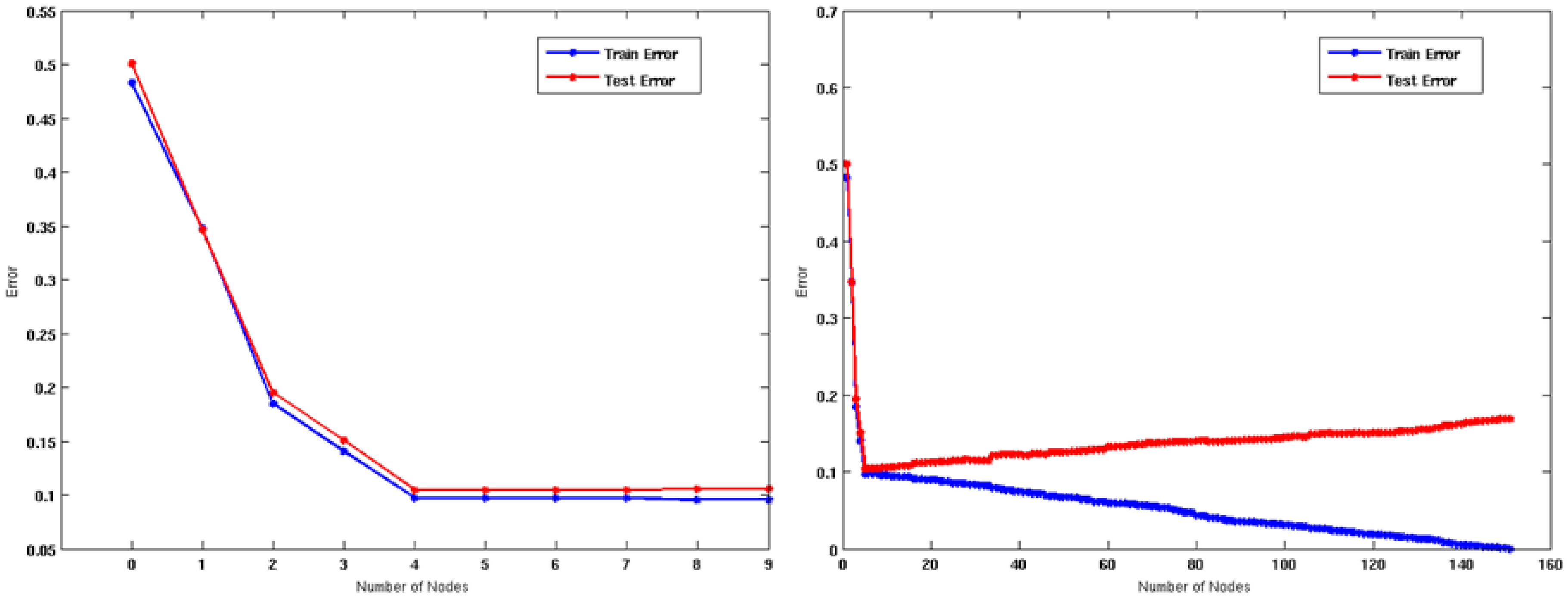
Decision Tree with 50 nodes



Which tree is better?



Model Underfitting and Overfitting

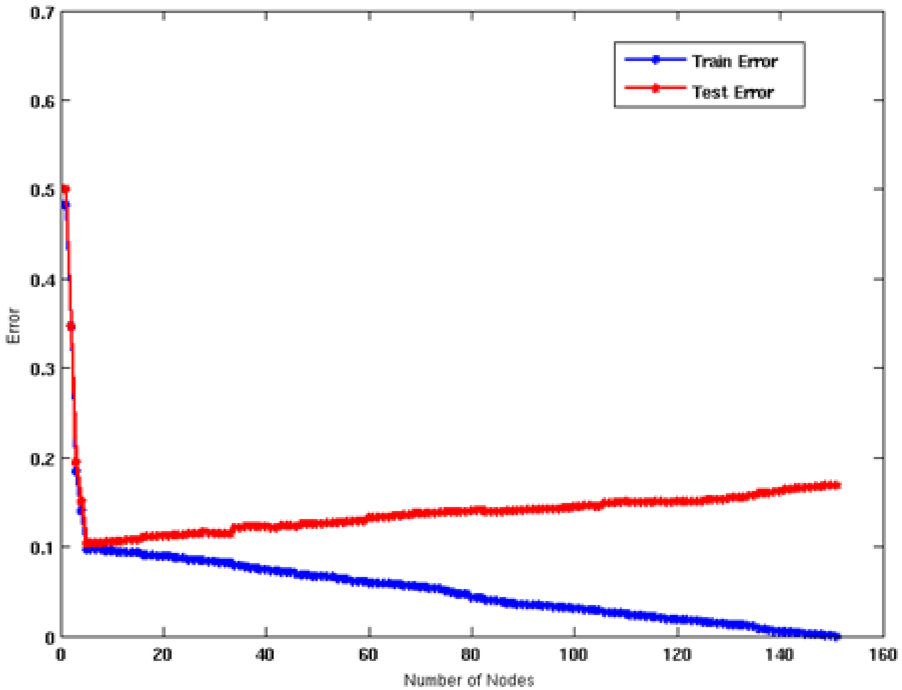


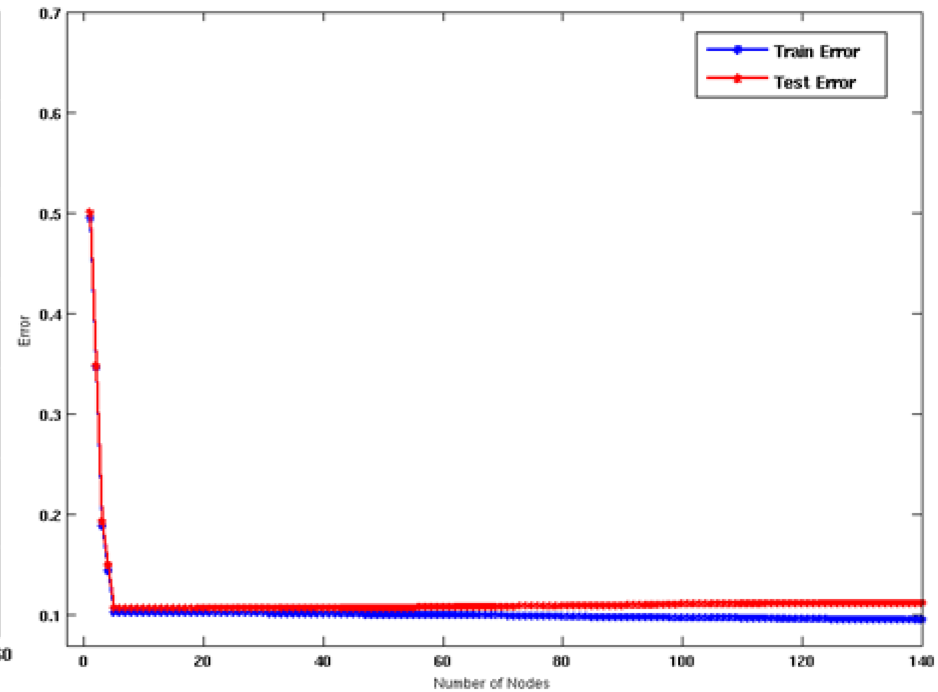
•As the model becomes more and more complex, test errors can start increasing even though training error may be decreasing

Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large

Model Overfitting – Impact of Training Data Size

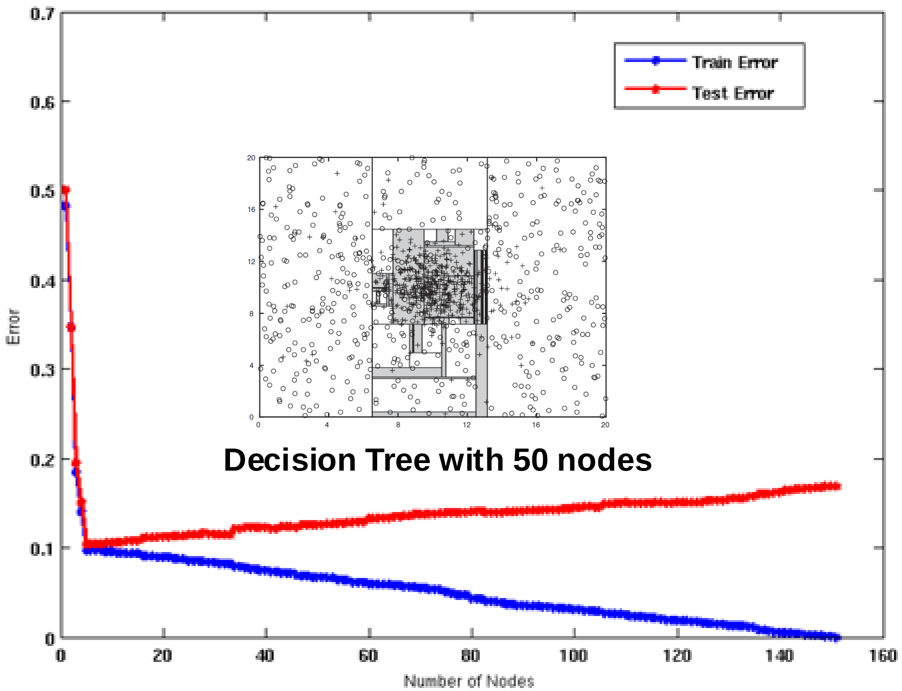


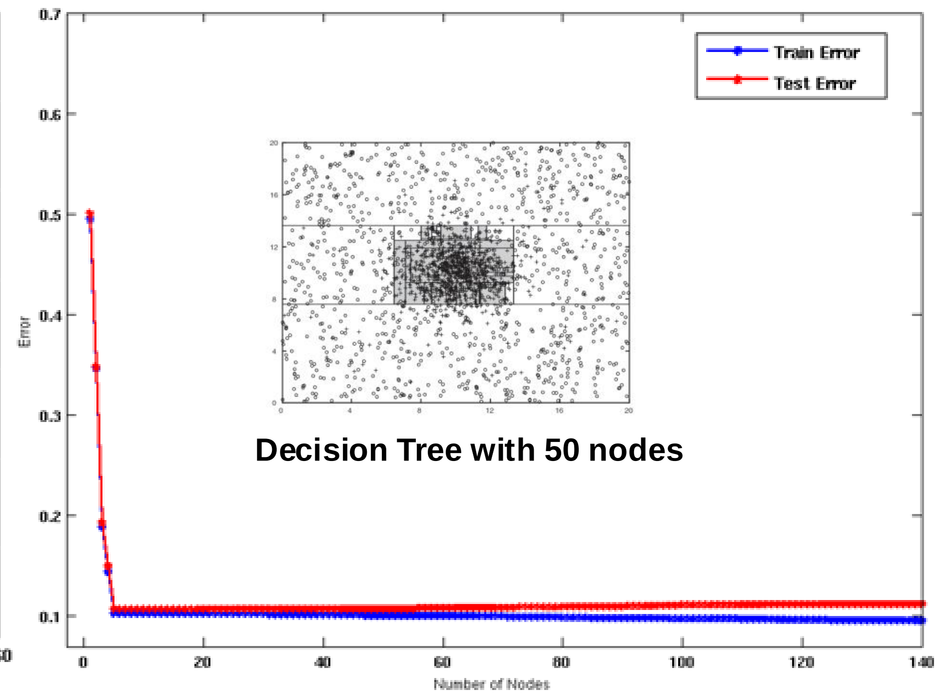


Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model

Model Overfitting – Impact of Training Data Size





Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model

Reasons for Model Overfitting

- Not enough training data

- High model complexity
  - Multiple Comparison Procedure

Reasons for Model Overfitting

Leaderboard for ogbn-arxiv

The classification accuracy on the test and validation sets. The higher, the better.

Package: >=1.1.1

Rank	Method	Ext. data	Test Accuracy	Validation Accuracy	Contact	References	#Params	Hardware	Date
1	SimTeG+TAPE+RevGAT	Yes	0.7803 ± 0.0007	0.7846 ± 0.0004	Keyu Duan	Paper, Code	1,386,219,488	4 * A100-XMS4 (40GB GPU)	Aug 7, 2023
2	TAPE+RevGAT	Yes	0.7750 ± 0.0012	0.7785 ± 0.0016	Xiaoxin He (NUS)	Paper, Code	280,283,296	4 NVIDIA RTX A5000 24GB GPUs	May 31, 2023
3	SimTeG+TAPE+GraphSAGE	Yes	0.7748 ± 0.0011	0.7789 ± 0.0008	Keyu Duan	Paper, Code	1,381,593,403	4 * A100-XMS4 (40GB GPU)	Aug 7, 2023
4	LD+REVGAT	Yes	0.7726 ± 0.0017	0.7762 ± 0.0008	Zhihao Shi (MIRA Lab, USTC & CityBrain Lab, Alibaba Cloud)	Paper, Code	140,438,868	GeForce RTX 3090 (24GB GPU)	Sep 27, 2023
5	GraDBERT & RevGAT+KD	Yes	0.7721 ± 0.0031	0.7757 ± 0.0009	Costas Mavromatis (UMN & AWS)	Paper, Code	1,304,912	GeForce RTX 3090 (24GB GPU)	Apr 20, 2023
6	GLEM+RevGAT	Yes	0.7694 ± 0.0025	0.7746 ± 0.0018	Jianan Zhao (Mila & MSRA Team)	Paper, Code	140,469,624	Tesla V100 (32GB)	Oct 27, 2022
7	GIANT-XRT+AGDN+BoT+self-KD	Yes	0.7637 ± 0.0011	0.7719 ± 0.0008	Chuxiong Sun	Paper, Code	1,309,760	Tesla V100 (16GB GPU)	Sep 2, 2022
8	GIANT-XRT+RevGAT+KD+DCN	Yes	0.7636 ± 0.0013	0.7699 ± 0.0002	Xiaojun Guo(xjguo)	Paper, Code	1,304,912	GeForce GTX 1080 Ti(12GB GPU)	Apr 24, 2023
9	GIANT-XRT+R-RevGAT+KD	Yes	0.7635 ± 0.0006	0.7692 ± 0.0010	LeeXue (HIT Team)	Paper, Code	1,500,712	TITAN RTX (24GB GPU)	Sep 30, 2022
10	GIANT-XRT+DRGAT+KD	Yes	0.7633 ± 0.0008	0.7725 ± 0.0006	anonymous_zhang(anonymous)	Paper, Code	2,685,527	Tesla P100-PCIE-16GB	Jan 14, 2022
11	GIANT-XRT+AGDN+BoT	Yes	0.7618 ± 0.0016	0.7724 ± 0.0006	Chuxiong Sun	Paper, Code	1,309,760	Tesla V100 (16GB GPU)	Sep 2, 2022
12	GIANT-XRT+RevGAT+KD (use raw text)	Yes	0.7615 ± 0.0010	0.7716 ± 0.0009	Eli Chien (UIUC)	Paper, Code	1,304,912	Tesla T4 (16GB GPU)	Nov 8, 2021
13	GIANT-XRT+DRGAT	No	0.7611 ± 0.0009	0.7716 ± 0.0008	anonymous_zhang(anonymous)	Paper, Code	2,685,527	Tesla P100-PCIE-16GB	Jan 17, 2022
14	GIANT-XRT+RevGAT (use raw text)	Yes	0.7590 ± 0.0019	0.7701 ± 0.0009	Eli Chien (UIUC)	Paper, Code	1,304,912	Tesla T4 (16GB GPU)	Nov 8, 2021
15	LGGNN+LabelReuse+C&S	No	0.7570 ± 0.0018	0.7687 ± 0.0005	Shichao Ma(Topo@OppoResearch)	Paper, Code	1,161,640	Tesla V100 (32GB)	Nov 3, 2022
15	GIANT-XRT+LGGNN+LabelReuse+C&S	Yes	0.7570 ± 0.0018	0.7687 ± 0.0005	Shichao Ma(Topo@OppoResearch)	Paper, Code	1,161,640	Tesla V100 (32GB)	Nov 3, 2022

	MLP	GCN	Graph-MLP	ES-GNN	GraphSAGE	LINKX	ES-MLP (ours)
Cora	$76.95_{1.00}$	$\mathbf{88.46_{0.83}}$	$86.64_{1.14}$	$87.30_{0.43}$	$88.26_{0.50}$	$83.15_{0.59}$	$88.15_{1.85}$
CiteSeer	$72.10_{1.12}$	$77.41_{0.95}$	$\mathbf{77.79_{0.10}}$	$74.27_{1.50}$	$76.54_{0.73}$	$73.23_{0.85}$	$75.67_{0.92}$
PubMed	$87.49_{0.90}$	$\mathbf{89.63_{0.79}}$	$87.06_{2.41}$	$88.81_{0.49}$	$89.60_{0.41}$	$87.47_{0.29}$	$87.56_{1.23}$
Actor	$35.81_{0.62}$	$29.24_{0.47}$	$36.03_{0.98}$	$38.91_{0.45}$	$32.24_{0.76}$	$33.92_{1.11}$	$\mathbf{39.73_{0.37}}$
Roman	$60.50_{0.88}$	$41.40_{1.58}$	$64.94_{0.25}$	$60.41_{1.90}$	$62.47_{1.90}$	$65.40_{0.37}$	$\mathbf{65.44_{0.92}}$
Amazon	$44.05_{0.54}$	$46.27_{0.67}$	$37.07_{0.80}$	$46.53_{0.34}$	$44.83_{1.16}$	$39.25_{0.51}$	$\mathbf{47.85_{1.23}}$
Minesweeper	$50.54_{0.49}$	$71.44_{0.74}$	$50.99_{0.35}$	$68.23_{1.10}$	$\mathbf{88.90_{2.37}}$	$51.61_{1.4}$	$50.87_{2.03}$

Effect of Multiple Comparison Procedure

- Consider the task of predicting whether stock market will rise/fall in the next 1 trading days

- Random guessing:

  $P(correct) = 0.5$

Day 1	Up
Day 2	Down
Day 3	Down
Day 4	Up
Day 5	Down
Day 6	Down
Day 7	Up
Day 8	Up
Day 9	Up
Day 10	Down

Effect of Multiple Comparison Procedure

- Approach:
  - Get 50 analysts
  - Each analyst makes 1 random guesses
  - Choose the analyst that makes the most number of correct predictions

- Probability that at least one analyst makes at correct predictions

Effect of Multiple Comparison Procedure

- Many algorithms employ the following greedy strategy:
  - Initial model: M
  - Alternative model: M' = M $\cup$ γ, where γ is a component to be added to the model (e.g., a test condition of a decision tree)
  - Keep M' if improvement, Δ(M,M') > α

- Often times, γ is chosen from a set of alternative components, $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$

- If many alternatives are available, one may inadvertently add irrelevant components to the model, resulting in model overfitting

Model Selection

- Performed during model building
- Purpose is to ensure that model is not overly complex (to avoid overfitting)
- Need to estimate generalization error
  - Using Validation Set
  - Incorporating Model Complexity

Validation Set

- Divide <u>training</u> data into two parts:
  - Training set:
    - use for model building
  - Validation set:
    - use for estimating generalization error
    - Note: validation set is not the same as test set

- Drawback:
  - Less data available for training

Model Selection:
Incorporating Model Complexity

- Rationale: Occam's Razor
  - Given two models of similar generalization errors, one should prefer the simpler model over the more complex model
  - A complex model has a greater chance of being fitted accidentally
  - Therefore, one should include model complexity when evaluating a model

Estimating the Complexity of Decision Trees

- **Pessimistic Error Estimate of decision tree T with k leaf nodes:**

  – **err(T): error rate on all training records**

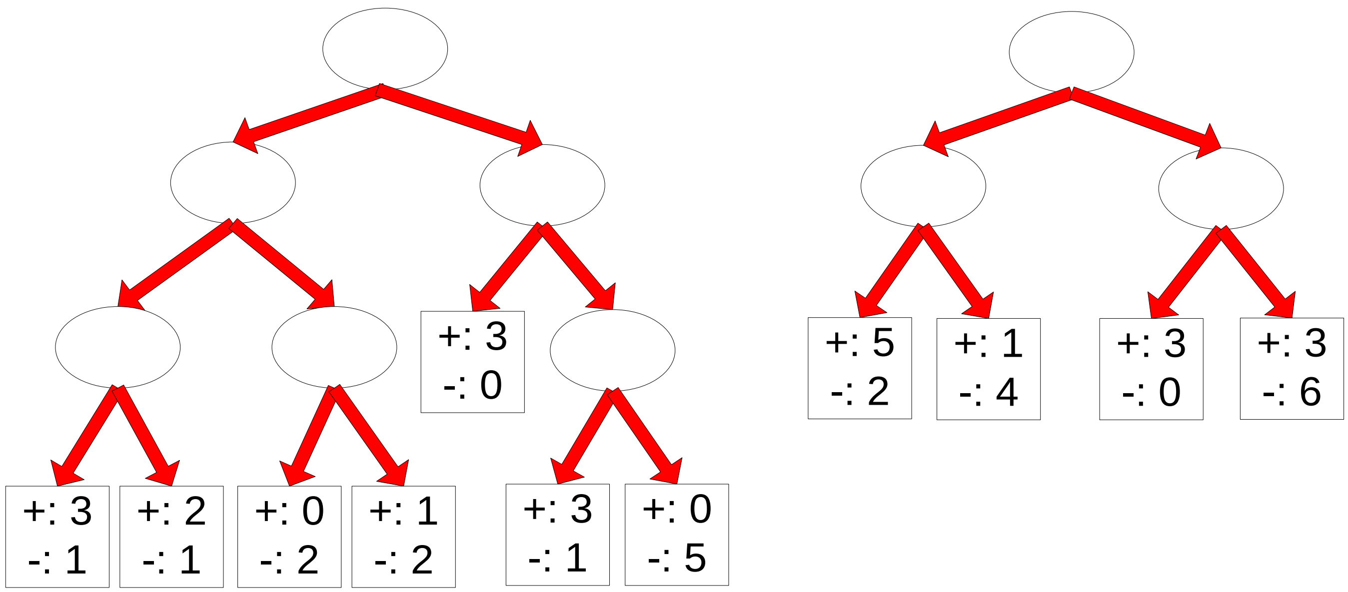
  – **Ω: trade-off hyper-parameter (similar to)**

    - **Relative cost of adding a leaf node**

  – **k: number of leaf nodes**

  – **N_{train}: total number of training records**

Estimating the Complexity of Decision Trees



Decision Tree, T_L

Decision Tree, T_R

$e(T_L) = 4/24$

$e(T_R) = 6/24$

$\Omega = 1$

$e_{gen}(T_L) = 4/24 + 1*7/24 = 11/24 = 0.458$

$e_{gen}(T_R) = 6/24 + 1*4/24 = 10/24 = 0.417$

Model Selection for Decision Trees

- **Pre-Pruning (Early Stopping Rule)**
  - Stop the algorithm before it becomes a fully-grown tree
  - Typical stopping conditions for a node:
    - Stop if all instances belong to the same class
    - Stop if all the attribute values are the same
  - More restrictive conditions:
    - Stop if number of instances is less than some user-specified threshold
    - Stop if class distribution of instances are independent of the available features
    - Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain).
    - Stop if estimated generalization error falls below certain threshold

Model Selection for Decision Trees

- Post-pruning
  - Grow decision tree to its entirety
  - Subtree replacement
    - Trim the nodes of the decision tree in a bottom-up fashion
    - If generalization error improves after trimming, replace sub-tree by a leaf node
    - Class label of leaf node is determined from majority class of instances in the sub-tree

Example of Post-Pruning

Class = Yes	20
Class = No	10
Error = 10/30	

Training Error (Before splitting) = 10/30

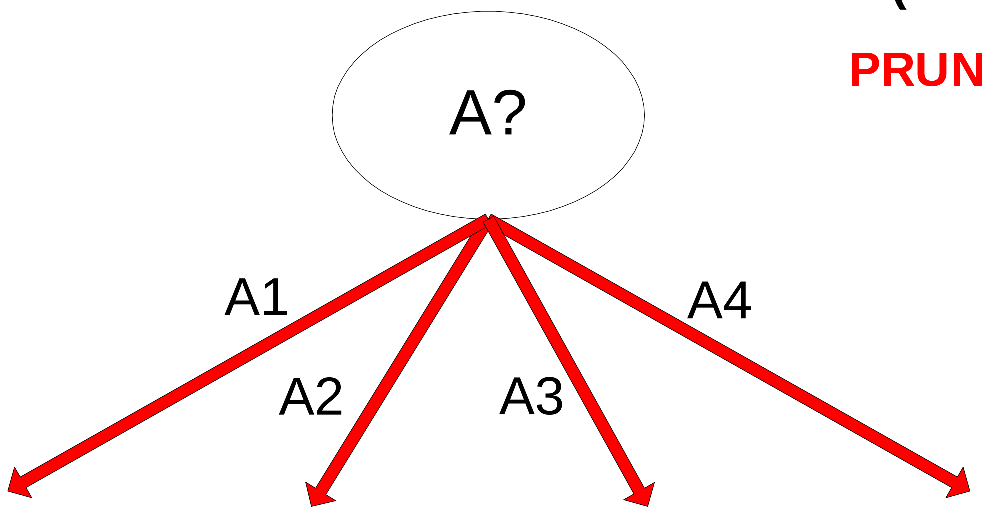
Pessimistic error = (10 + 1)/30 = 11/30

Training Error (After splitting) = 9/30

Pessimistic error (After splitting)

$$= (9 + 4 \times 1)/30 = 13/30$$

PRUNE!



Class = Yes	8
Class = No	4

Class = Yes	3
Class = No	4

Class = Yes	4
Class = No	1

Class = Yes	5
Class = No	1

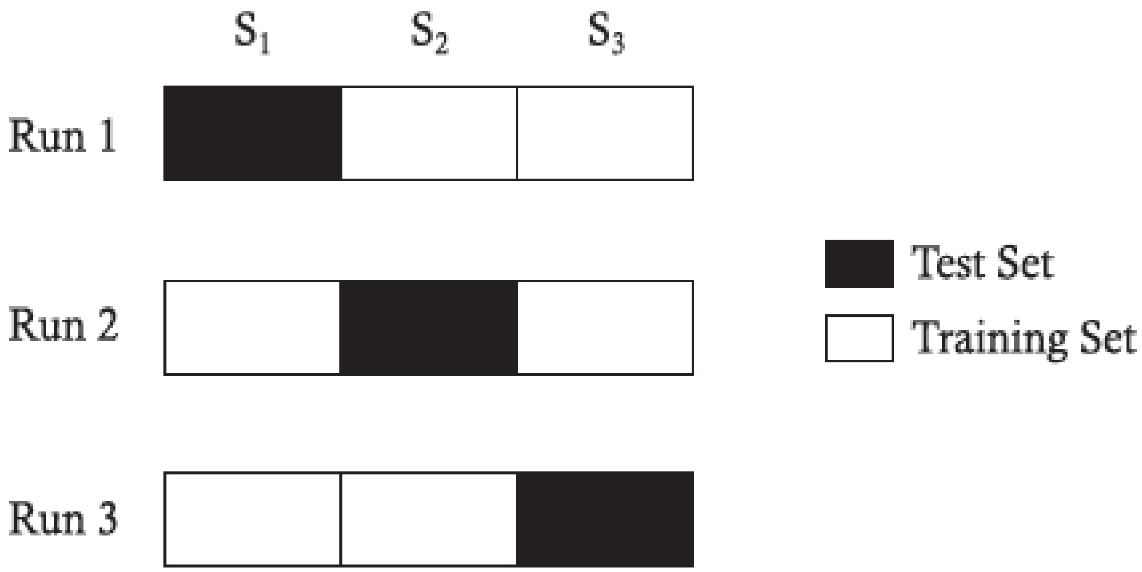
Model Evaluation

- **Purpose:**
  - **To estimate performance of classifier on previously unseen data (test set)**

- **Holdout**
  - **Reserve k% for training and (100-k)% for testing**
  - **Random subsampling: repeated holdout**
- **Cross validation**
  - **Partition data into k disjoint subsets**
  - **k-fold: train on k-1 partitions, test on the remaining one**
  - **Leave-one-out: k=n**

Model Evaluation

- 3-fold cross-validation



Model Evaluation

- **Repeated cross-validation**
  - **Perform cross-validation a number of times**
  - **Gives an estimate of the variance of the generalization error**
- **Stratified cross-validation**
  - **Guarantee the same percentage of class labels in training and test**
  - **Important when classes are imbalanced and the sample is small**
- **Use nested cross-validation approach for model selection and evaluation**

Question?

1. "Judge a man by his questions rather than by his answers."
   – Voltaire
2. "If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions."
   – Albert Einstein
3. "The art and science of asking questions is the source of all knowledge."
   – Thomas Berger
4. "Asking the right questions takes as much skill as giving the right answers."
   – Robert Half
5. "The wise man doesn't give the right answers, he poses the right questions."
   – Claude Lévi-Strauss
6. "Great questions make great companies."
   – Peter Drucker

General thing regarding to quiz

- **1 hrs**

- **Multi-choice QA and writing QA**
  - Example: Among the following attribute, which one is the nominal?
  - Write down four different types of attributes and their key differences.

- I will not tell you about this knowledge, what kind of questions I will give you. All I want to check is whether you **understand** the concept rather than **memorize** it

- Do not worry about the final grade, I am very friendly but I am not a person who is easy to give a student A+ (a tiny portion of the questions would be very hard but very very few)

Review

- **Basics:**
  - **Definition of Data Mining**
  - **What are some exemplary tasks in data mining?**

- **Data:**
  - **Basic Components of Data and their definitions and examples**
  - **What are some basic types of attributes and their properties?**
  - **What are some characteristics of data?**
  - **Curse of Dimensionality**
  - **What types will you have for your data and how will you model each type in the computer?**
  - **What data quality issue will you encounter and how will you typically solve this issue.**

Review

- **Distance and Similarity measure:**
  - Different types of similarity and distance measure
  - Common properties of distance and similarity
  - Properties of distance and similarity
    - Invariant to scaling?
    - Invariant to translation?

- **Application:** Given a specific problem setting, can you choose the right way to model the data and use the right similarity/distance measure to quantify?

Review

- **Basics of classification**
  - Input/output of classification
  - Workflow
  - What is underfitting and what is overfitting?

- **Decision tree**
  - Understanding Training/testing process of Decision-tree
  - When should we stop tree expansion?
  - How should we split the tree?
  - What are some general ways to compute node impurity and how to compute and how do you understand the metric?
  - Smart way to select the best way to split when the attribute is continuous
  - Advantages and disadvantages of decision-tree based classification
  - Tree expansion and tree pruning in decision tree and