Introduction to Text Mining

Tom Bohannon

TAIR Conference

February 2013

Objectives

- Define *text mining* and identify text mining applications.
- Survey applications of text mining.
- Use *an example* to illustrate text mining concepts.
- Examine how text mining fits into modern data mining projects.

What Is Text Mining ?

- Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects.
- “SAS defines text mining as the process of investigating a large collection of free-form documents in order to discover and use the knowledge that exists in the collection as a whole.” (*SAS® Text Miner: Distilling Textual Data for Competitive Business Advantage*)

Text Mining – Two General Goals

- Pattern Discovery (Unsupervised Learning)
  - Identify naturally occurring groups (classification*).
  - Derive convenient segments (clustering).
- Prediction (Supervised Learning)
  - Input variables are associated with values of a target variable.
  - Derive a model or set of rules that produces a predicted target value for a given set of inputs.

* Classification with a target variable is prediction.

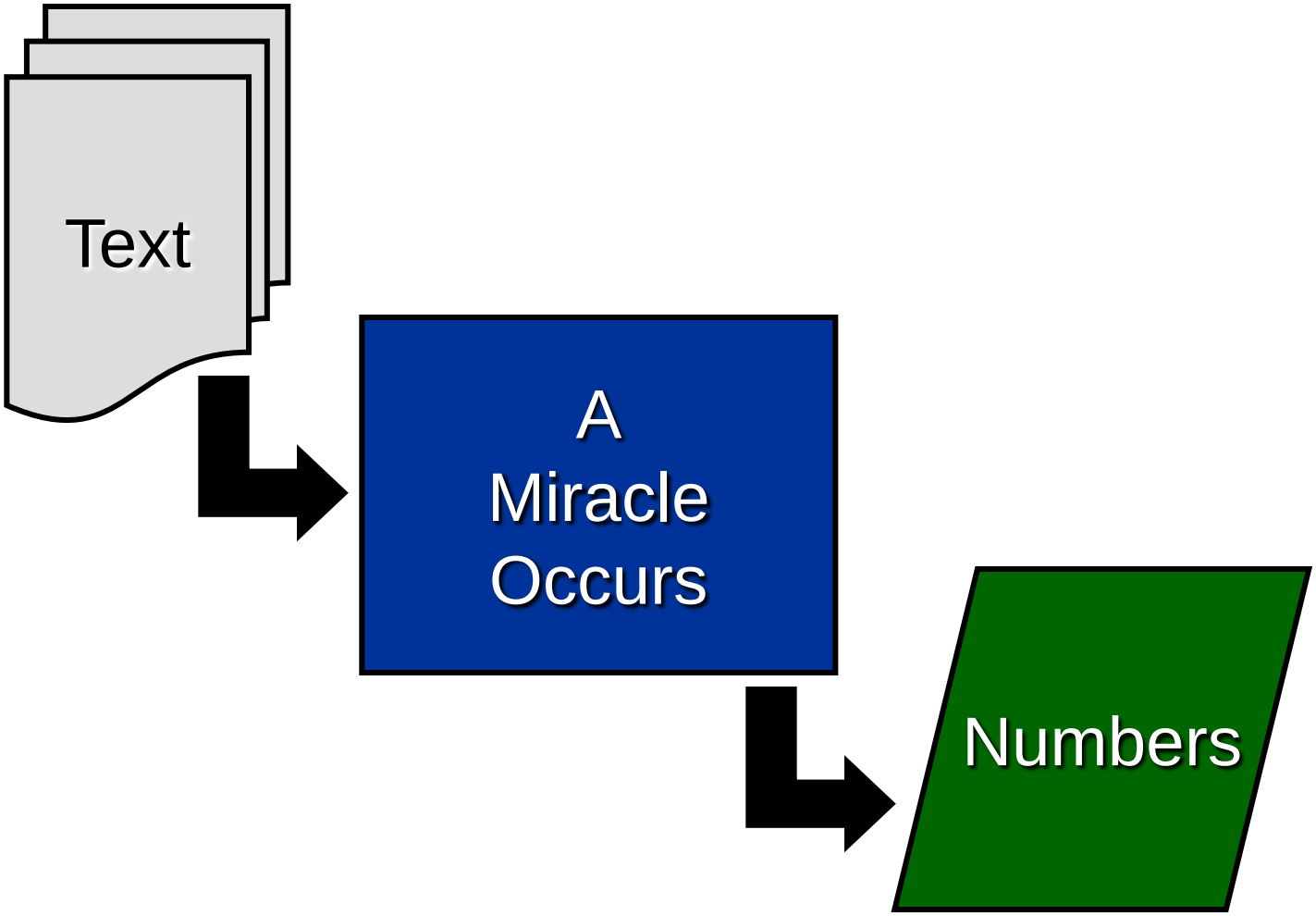
Text Mining

Text mining has the following characteristics:

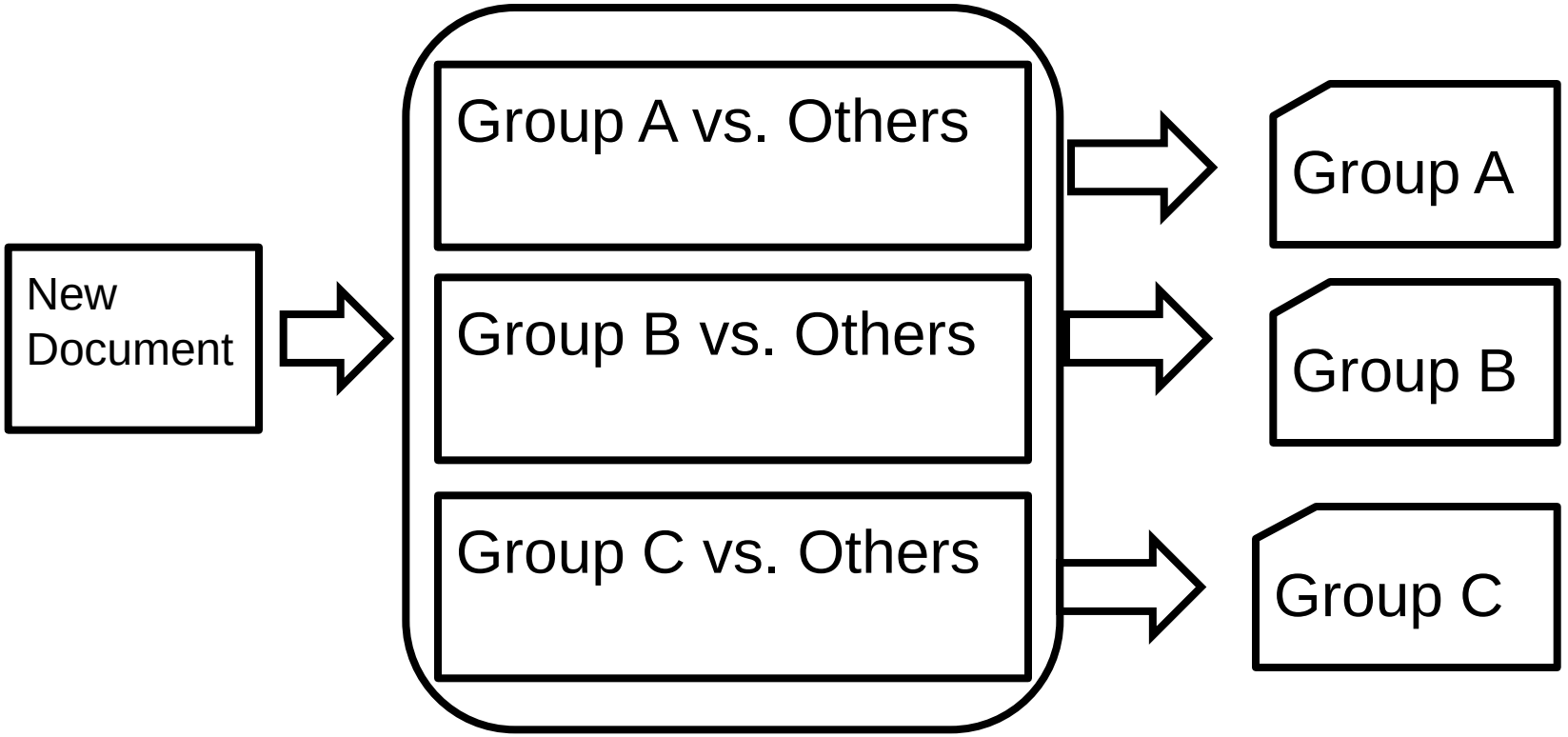
- operates with respect to a corpus of documents
- employs a dictionary to identify relevant terms
- accommodates a variety of metrics to quantify the contents of a document within the corpus
- derives a structured vector* of measurements for each document relative to the corpus
- employs analytical methods applied to the structured vector of measurements based on the goals of the analysis, for example, groups documents into segments

* Some text mining methods use a structured matrix.

Another View of Text Mining



Application: Document Classification



Document Categorization

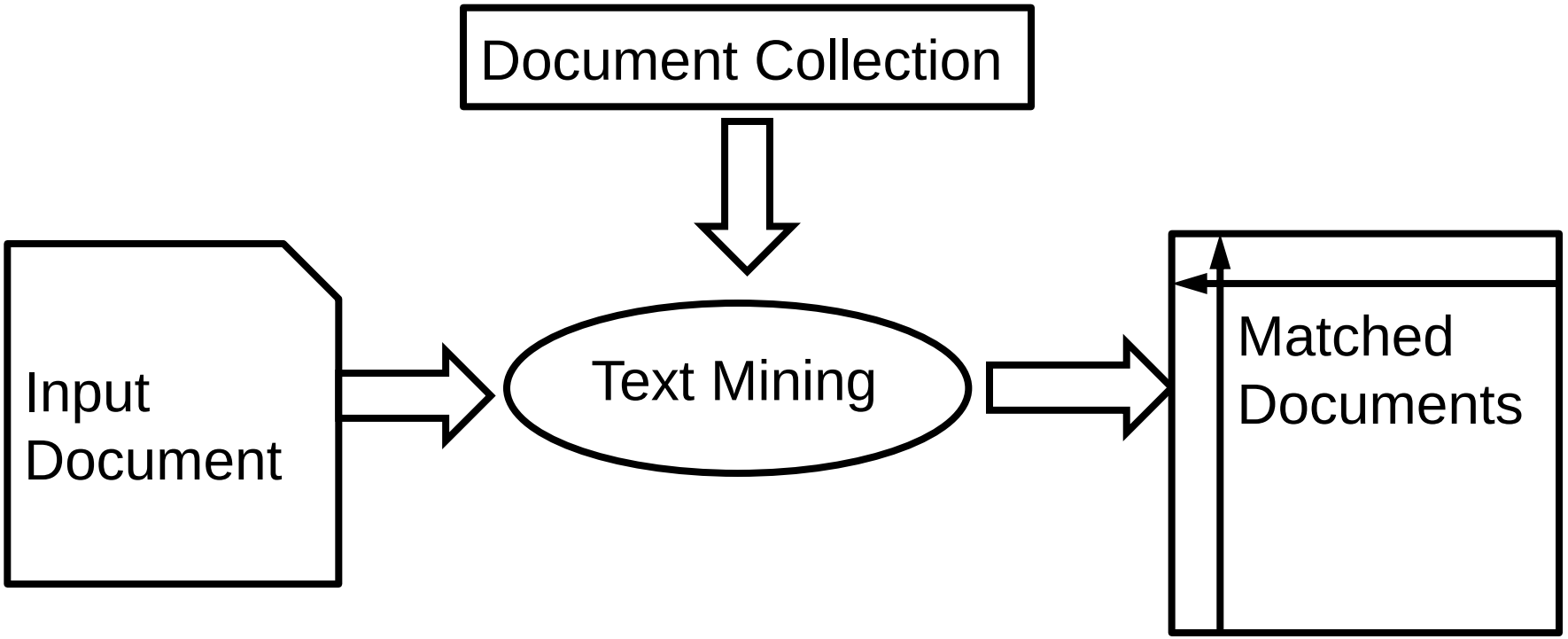
Document categorization

- Assign documents to pre-defined categories

Examples

- Process email into work, personal, junk
- Process documents from a newsgroup into “interesting”, “not interesting”, “spam and flames”
- Process transcripts of bugged phone calls into “relevant” and “irrelevant”

Application: Information Retrieval

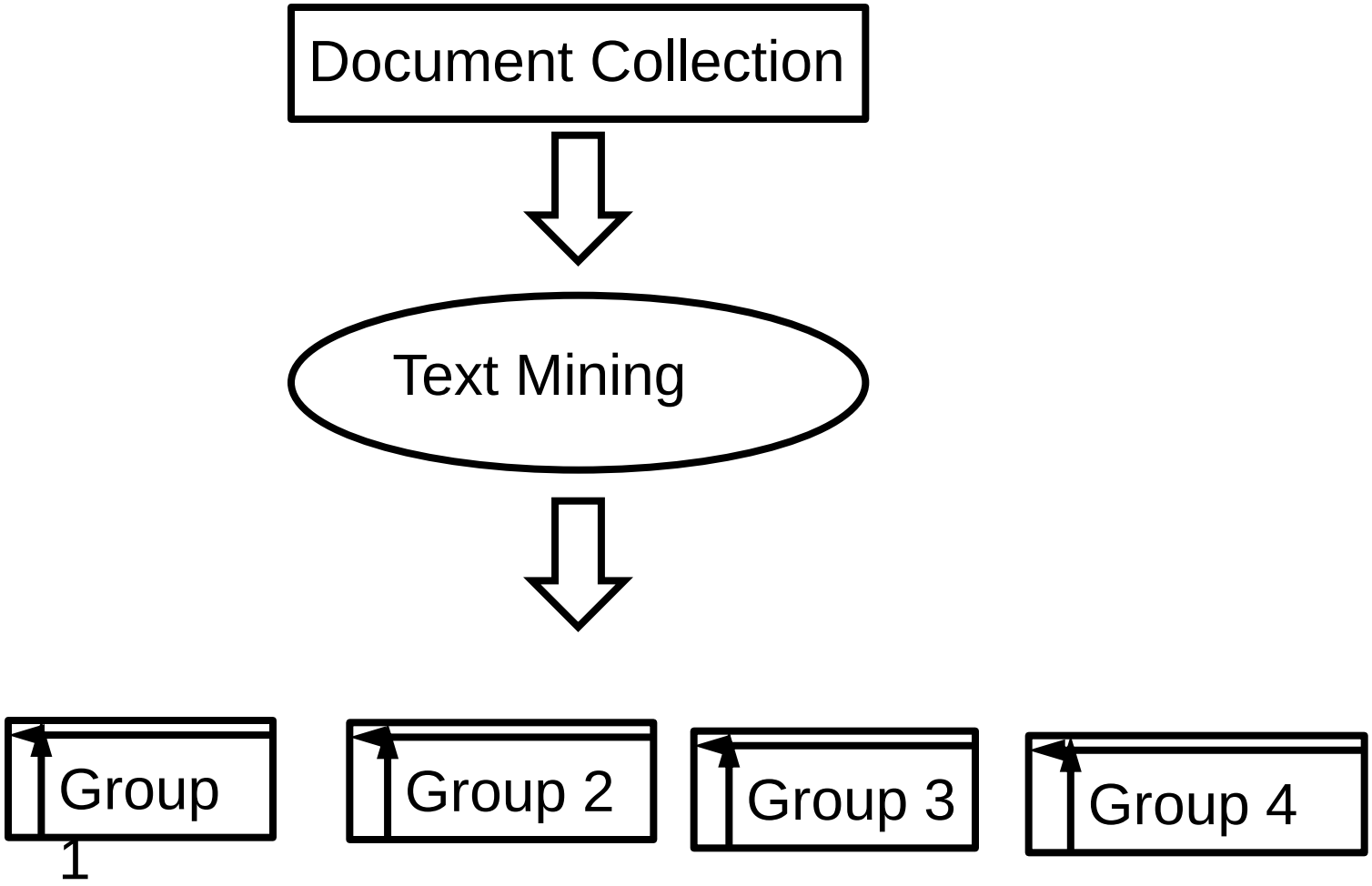


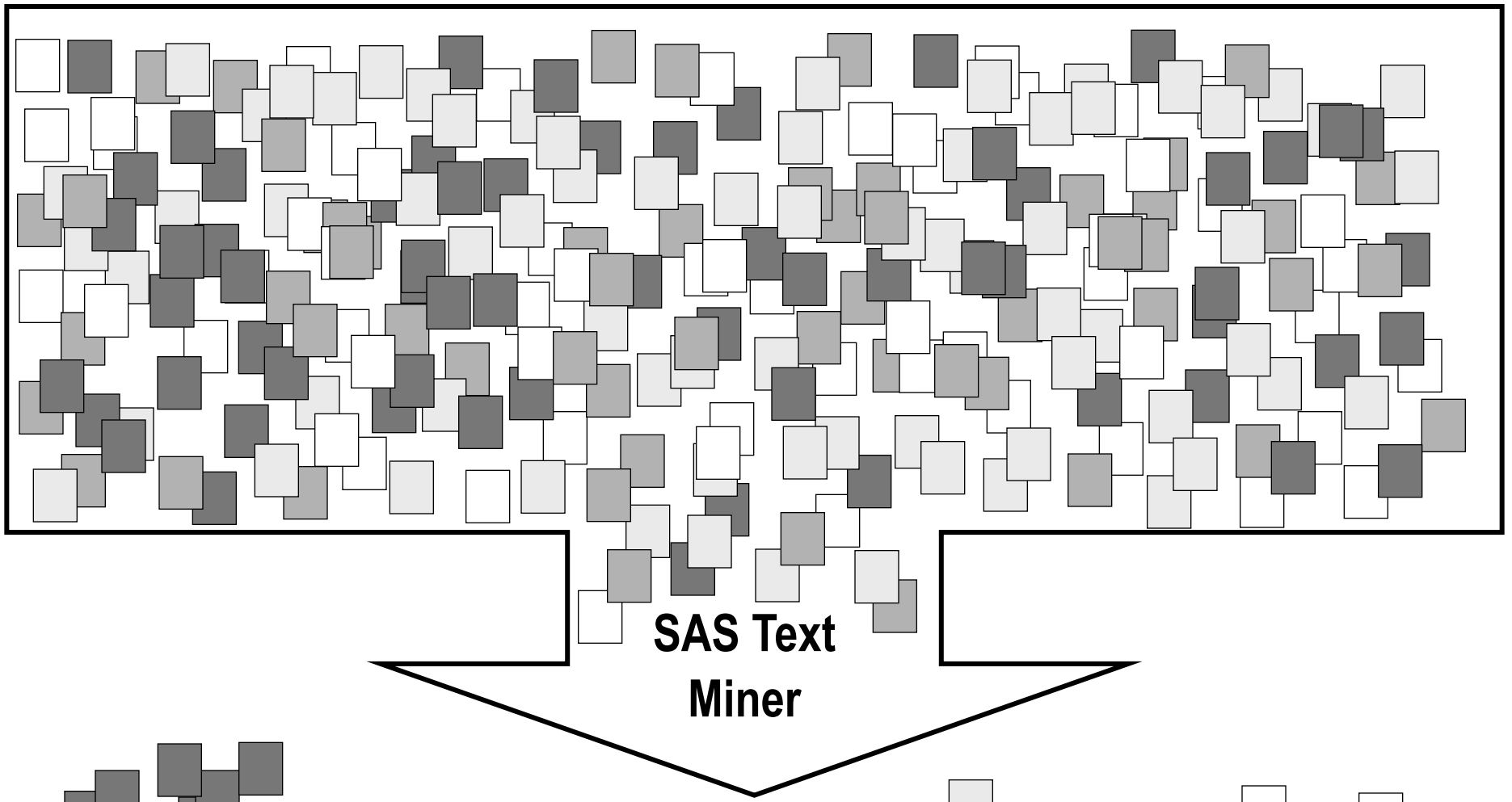
Introduction

How can we retrieve information using a search engine?.

- We can represent the query and the documents as vectors (*vector space model*)
  - However to construct these vectors we should perform a preliminary document preparation.
- The documents are retrieved by finding the closest distance between the query and the document vector.

Application: Clustering




SAS Text
Miner

Document Classification

Document classification

- Cluster documents based on similarity

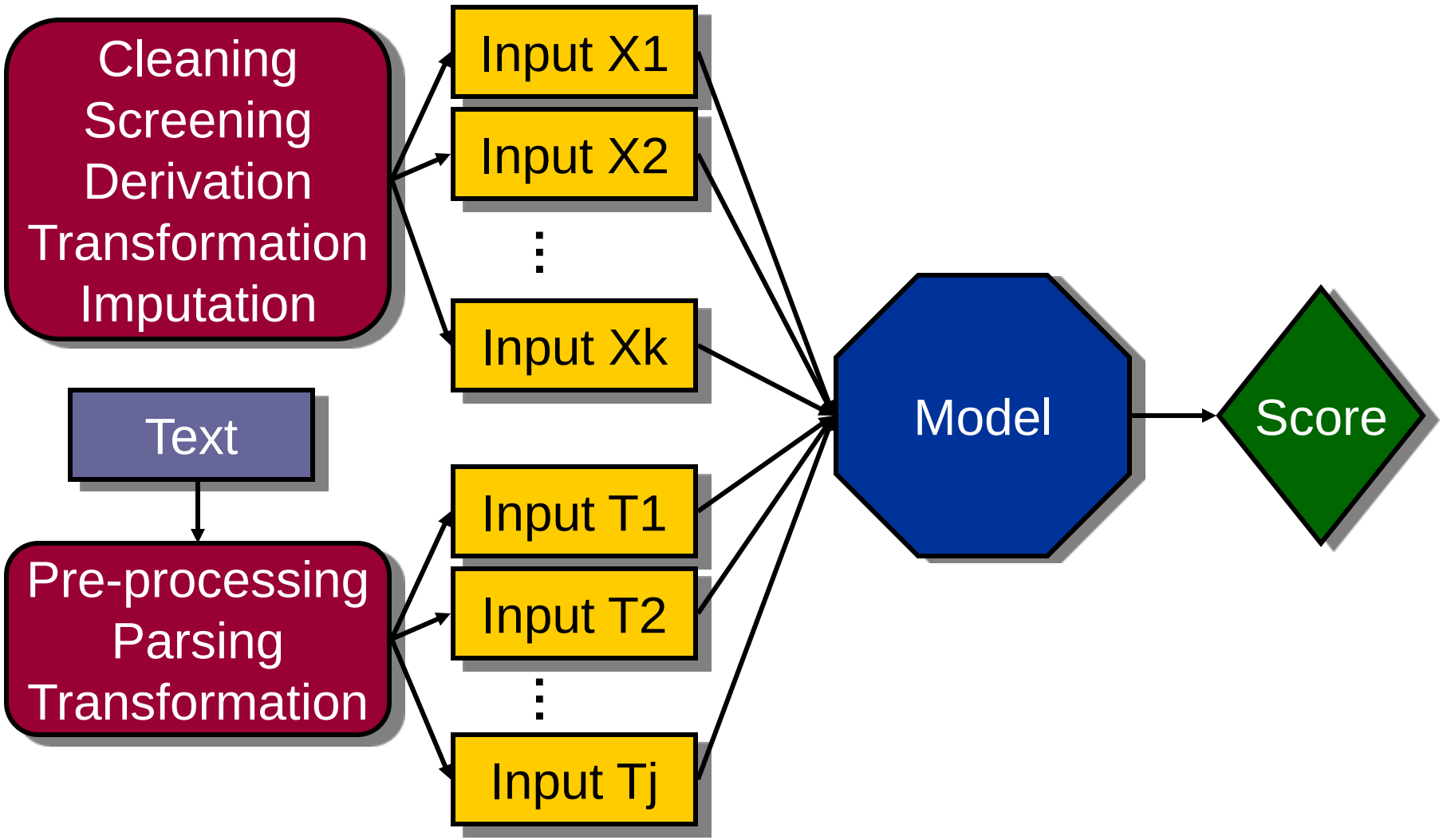
Examples

- Group samples of writing in an attempt to determine author(s)
- Look for “hot spots” in customer feedback
- Find new trends in a document collection (outliers, hard to classify)

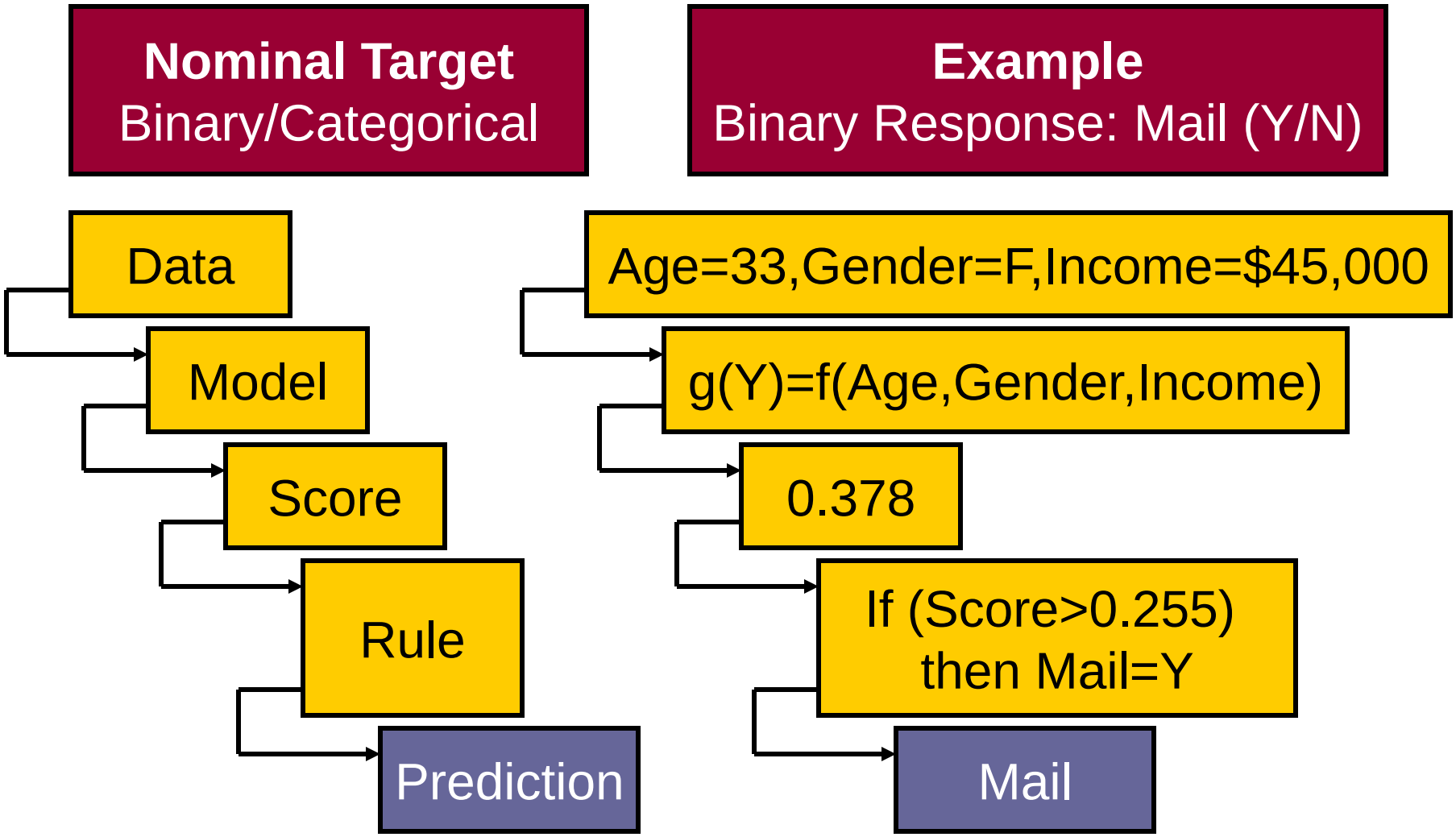
IR Applications Using Text Mining

- Survey Analysis
- Analysis of Student Evaluations of Instructors
- Predictive Modeling
  - Enrollment Models
  - Retention Models

Predictive Modeling



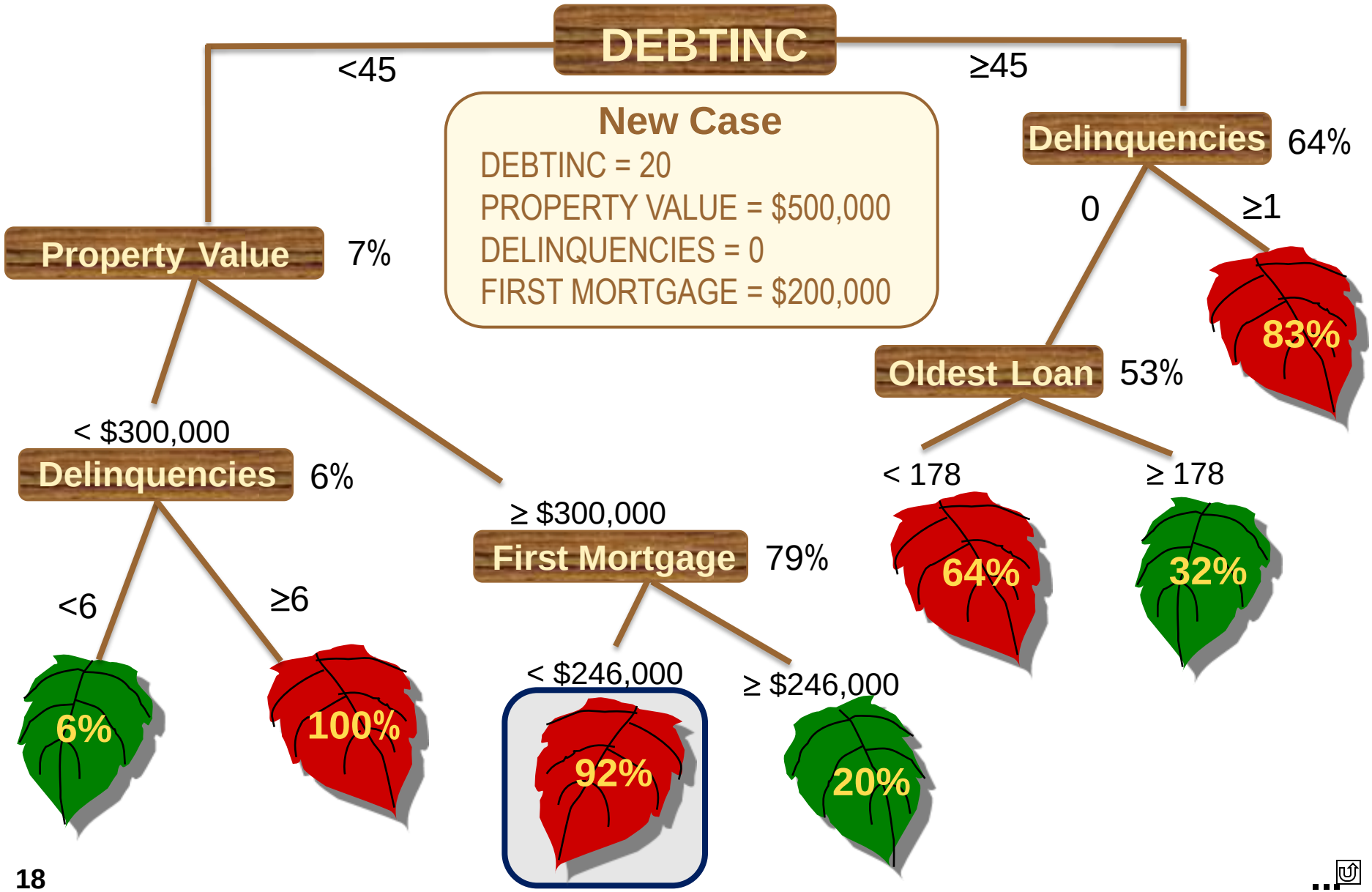
Obtaining the Prediction



Objectives

- Explore the general concept of decision trees.
- Build a decision tree model.
- Examine the model results and interpret these results.

Fitted Decision Tree

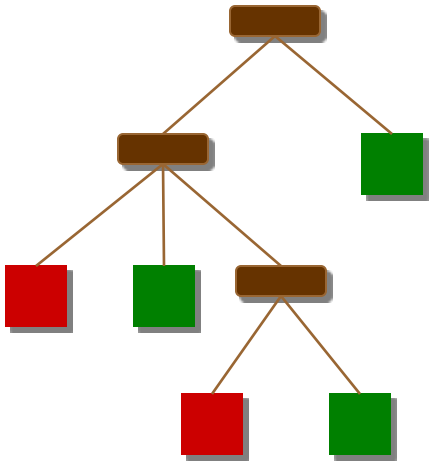


The Cultivation of Trees

- Split Search
  - Which splits are to be considered?
- Splitting Criterion
  - Which split is best?
- Stopping Rule
  - When should the splitting stop?
- Pruning Rule
  - Should some branches be lopped off?

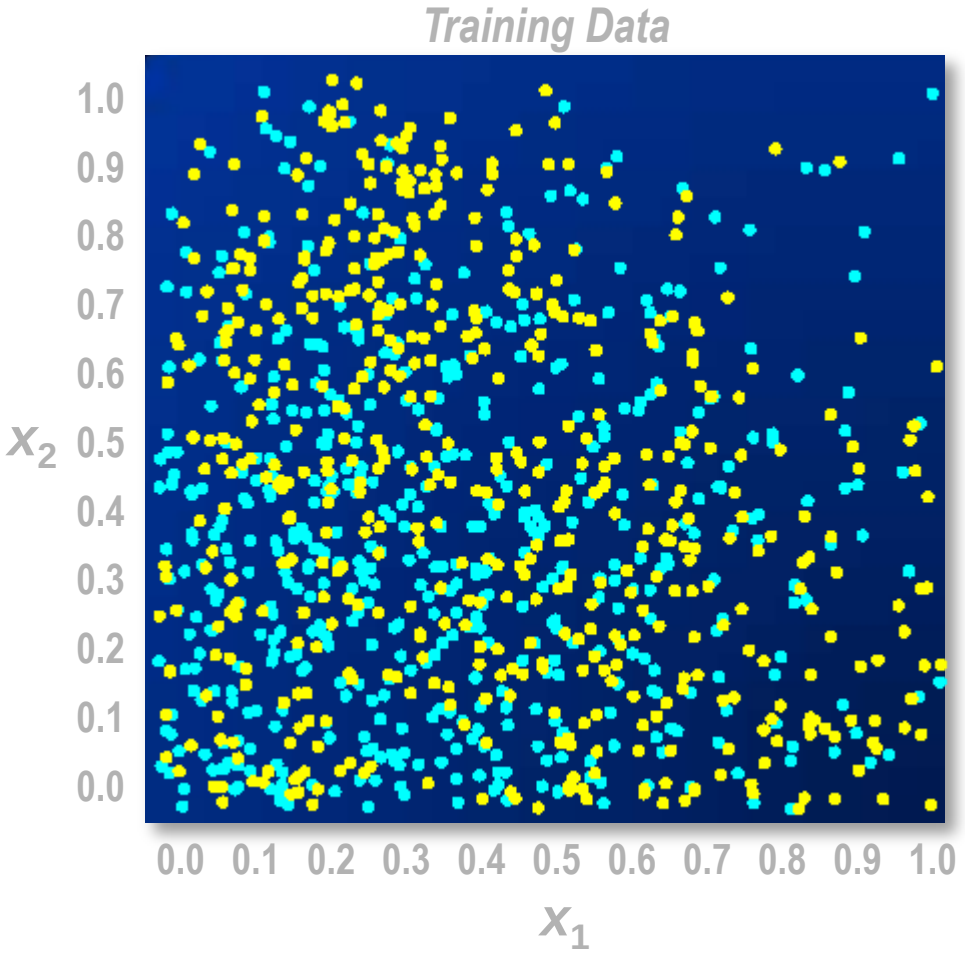
Benefits of Trees

- Interpretability
  - tree-structured presentation
- Mixed Measurement Scales
  - nominal, ordinal, interval
- Regression trees
- Robustness
- Missing Values



Simple Prediction Illustration

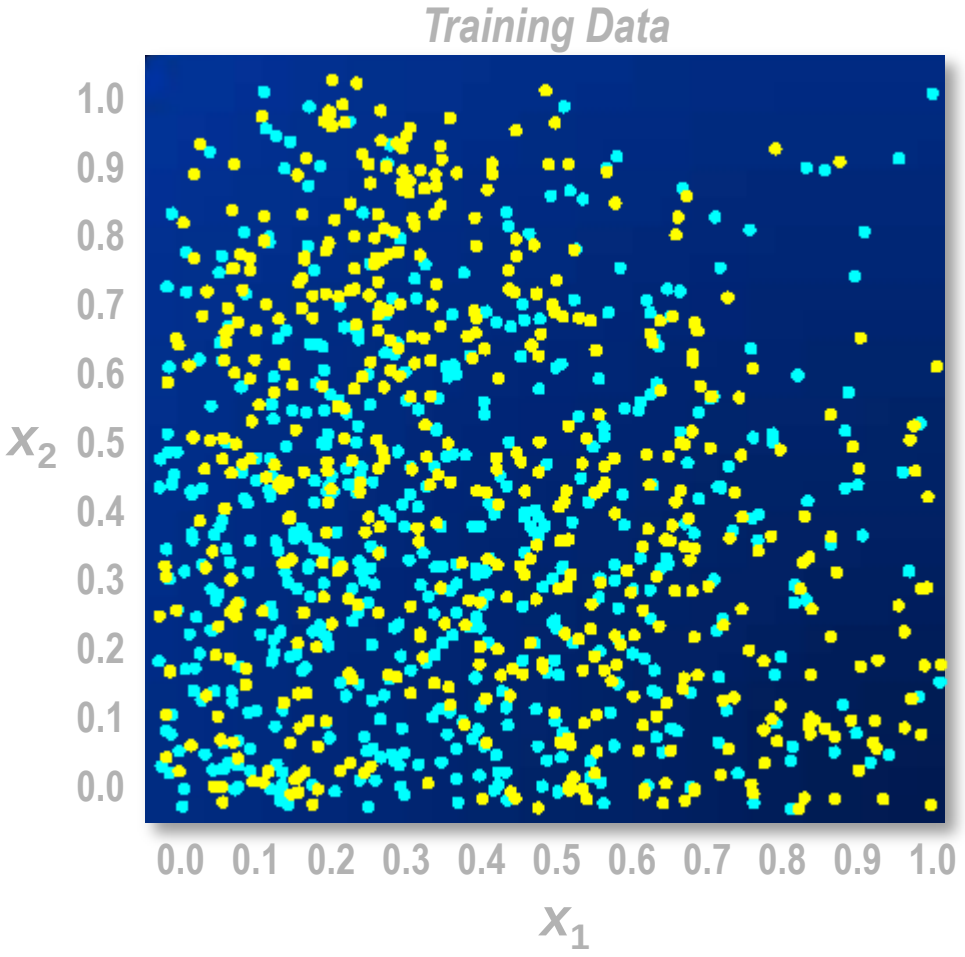
Predict dot color for each x_1 and x_2.



...

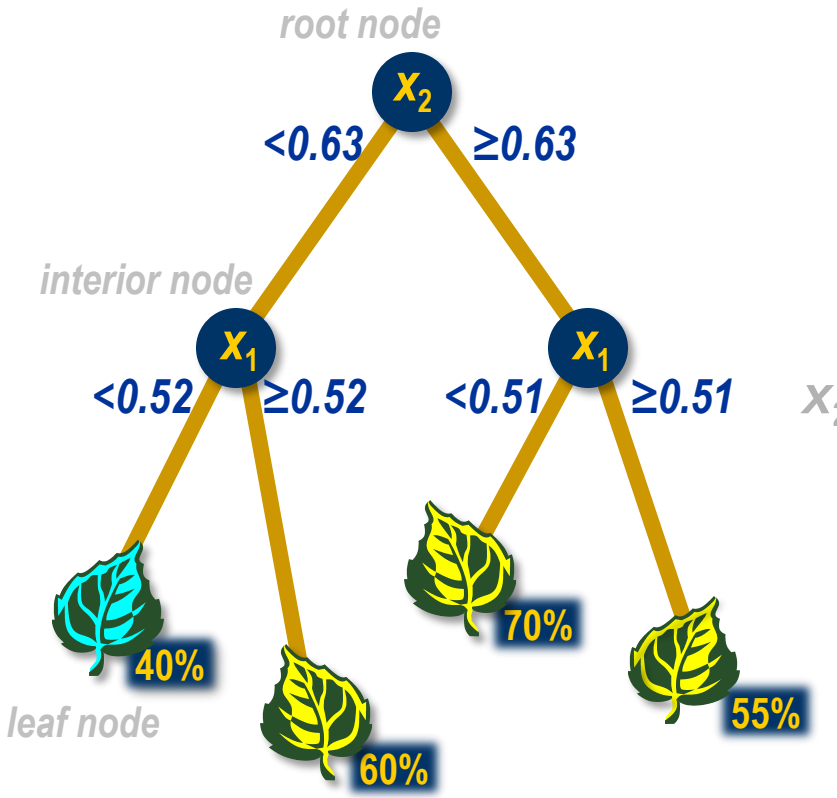
Simple Prediction Illustration

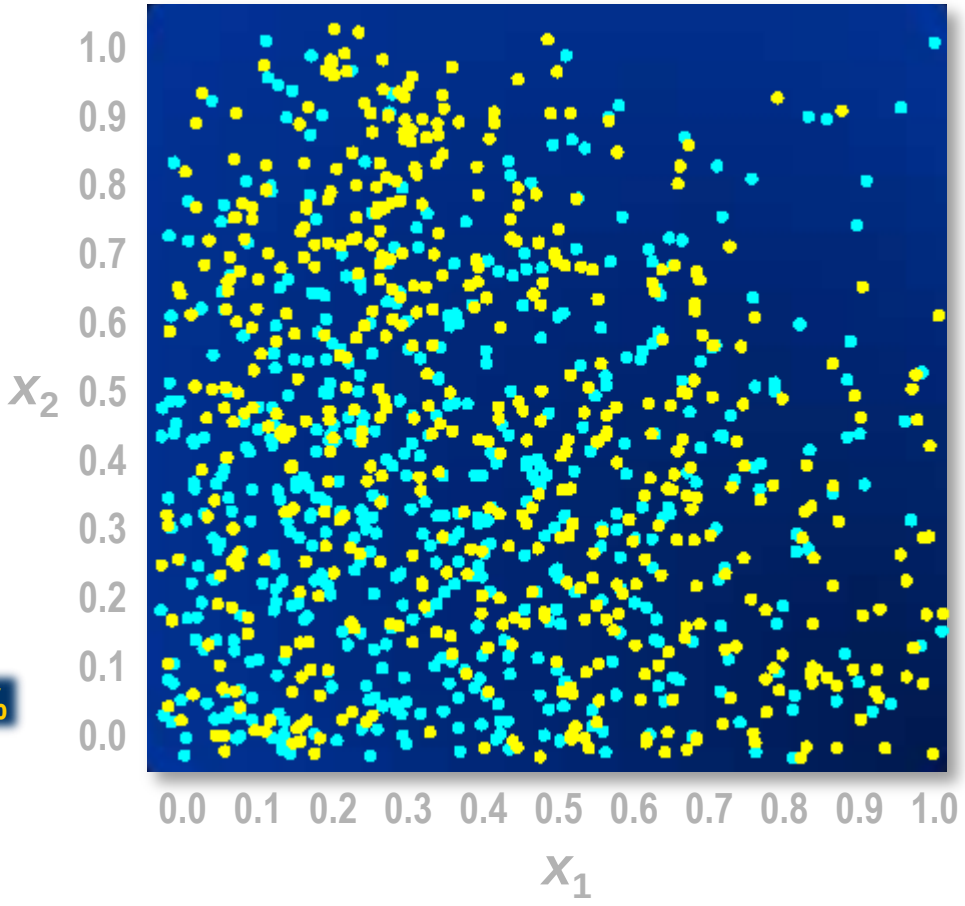
Predict dot color for each x_1 and x_2.



...

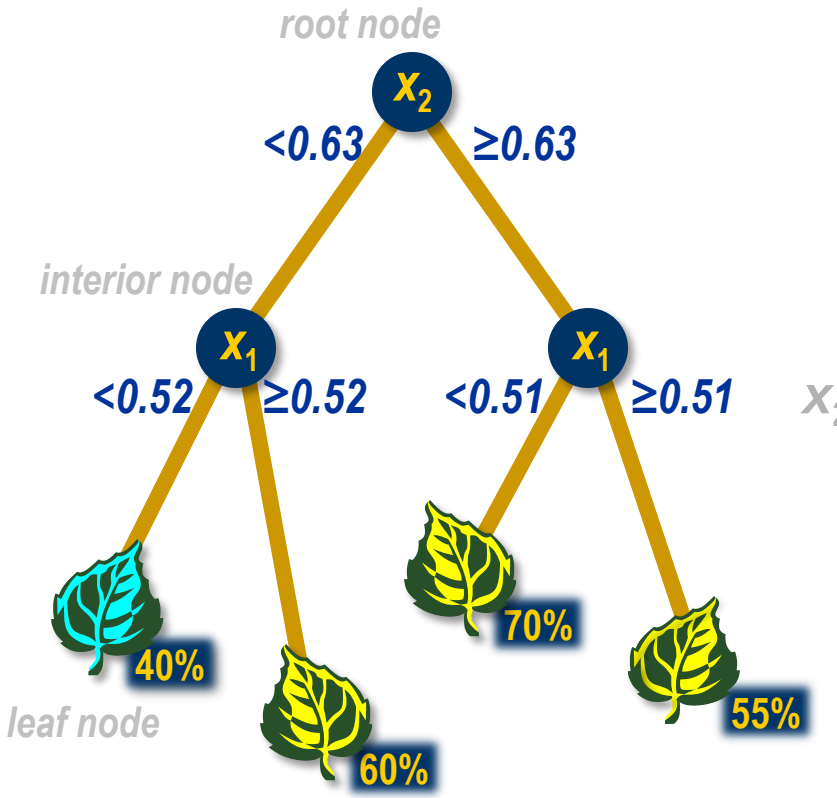
Decision Tree Prediction Rules

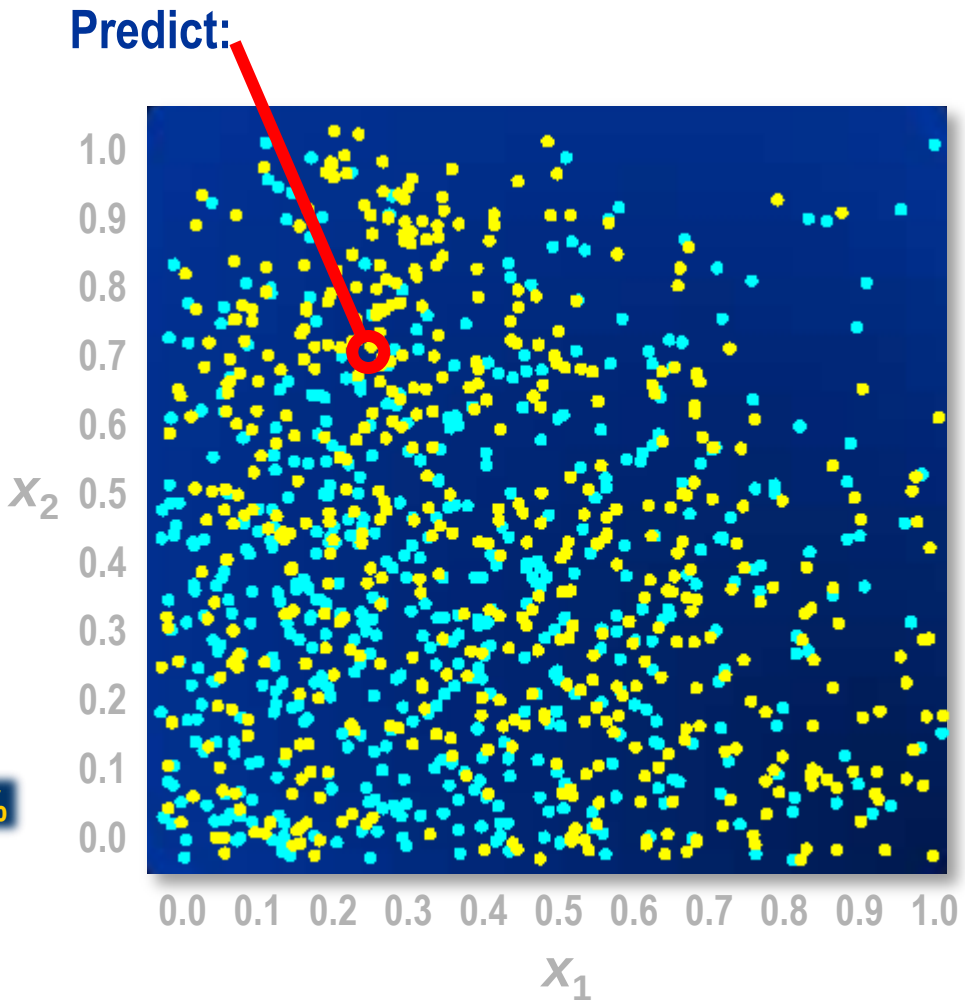




...

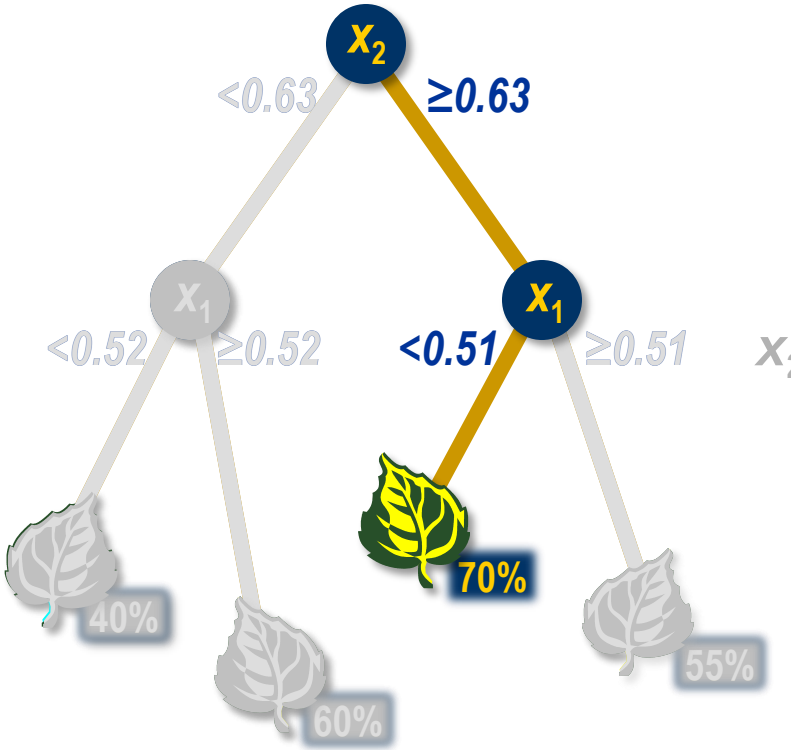
Decision Tree Prediction Rules

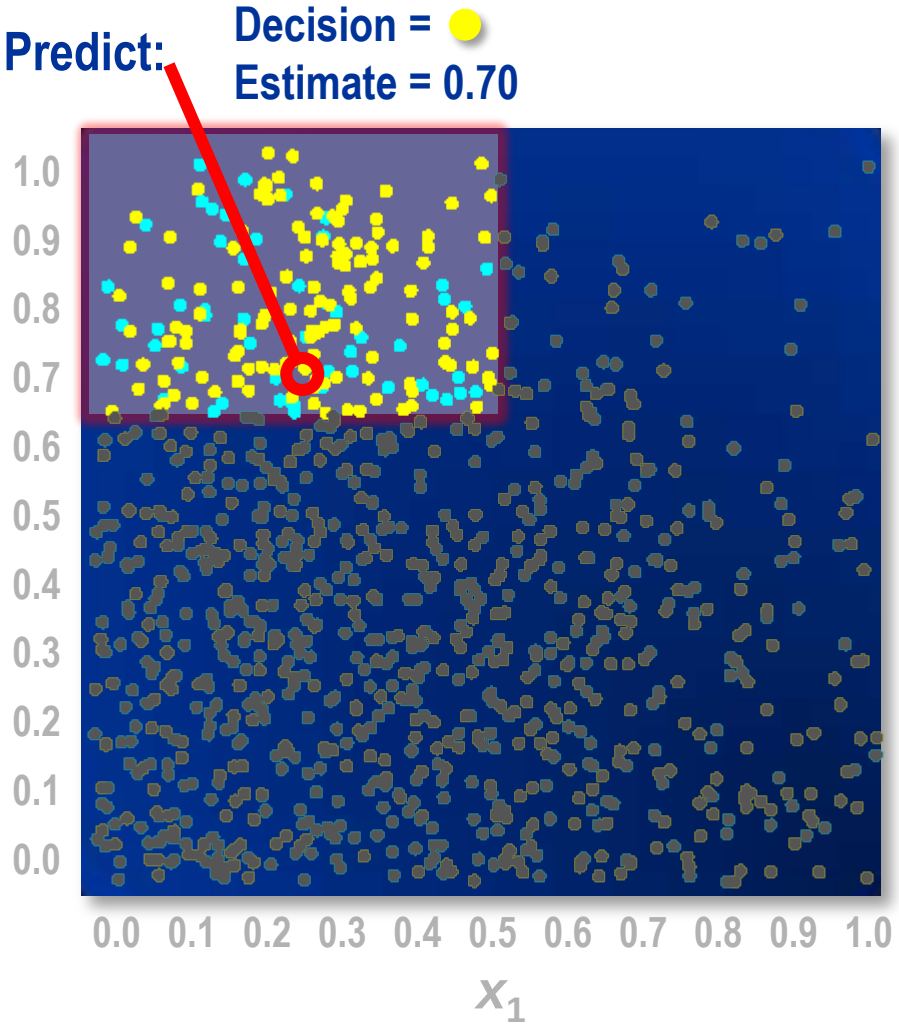




...

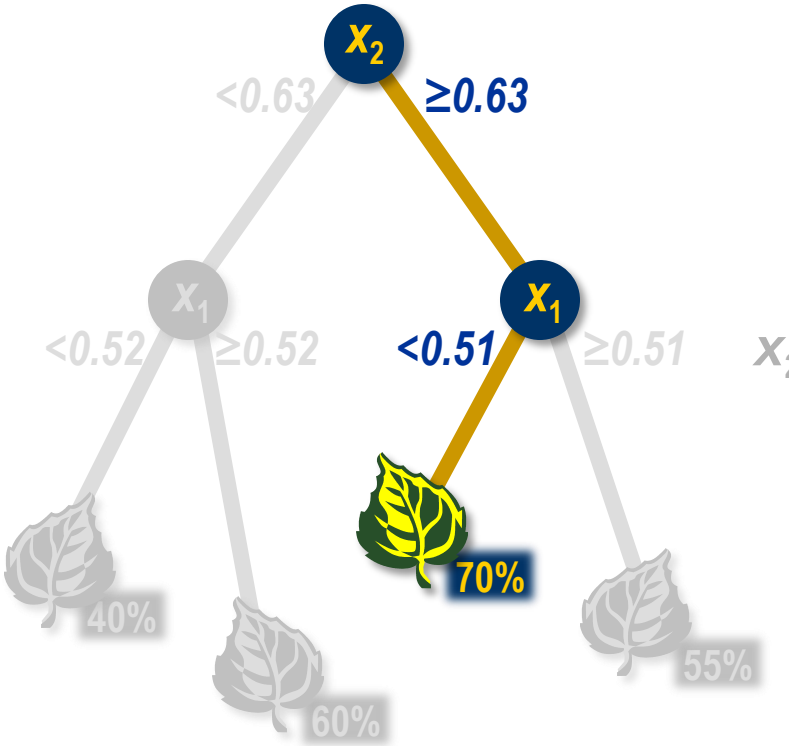
Decision Tree Prediction Rules

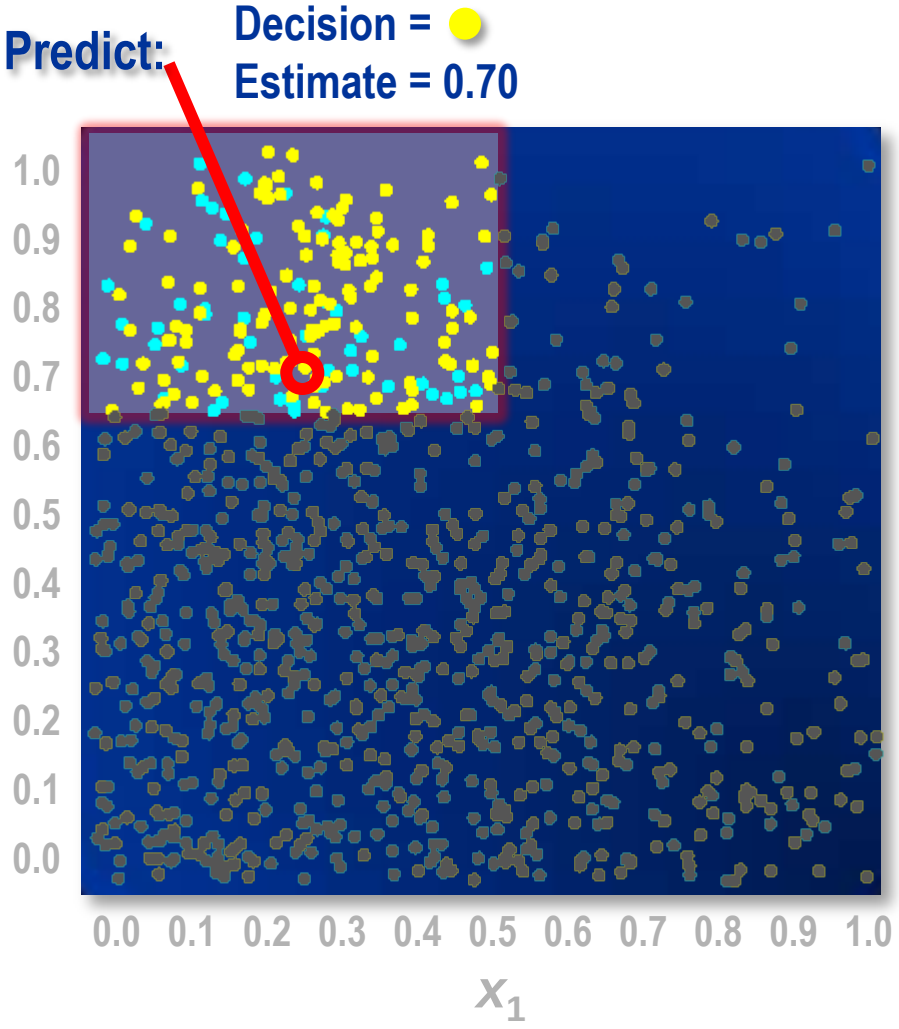




...

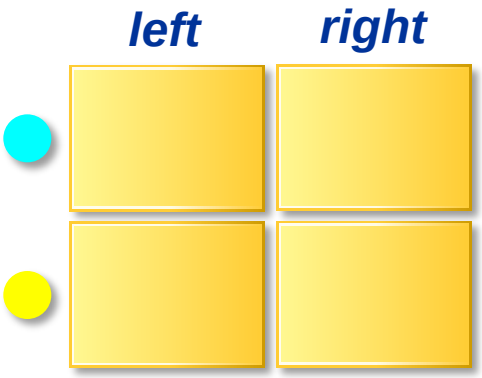
Decision Tree Prediction Rules





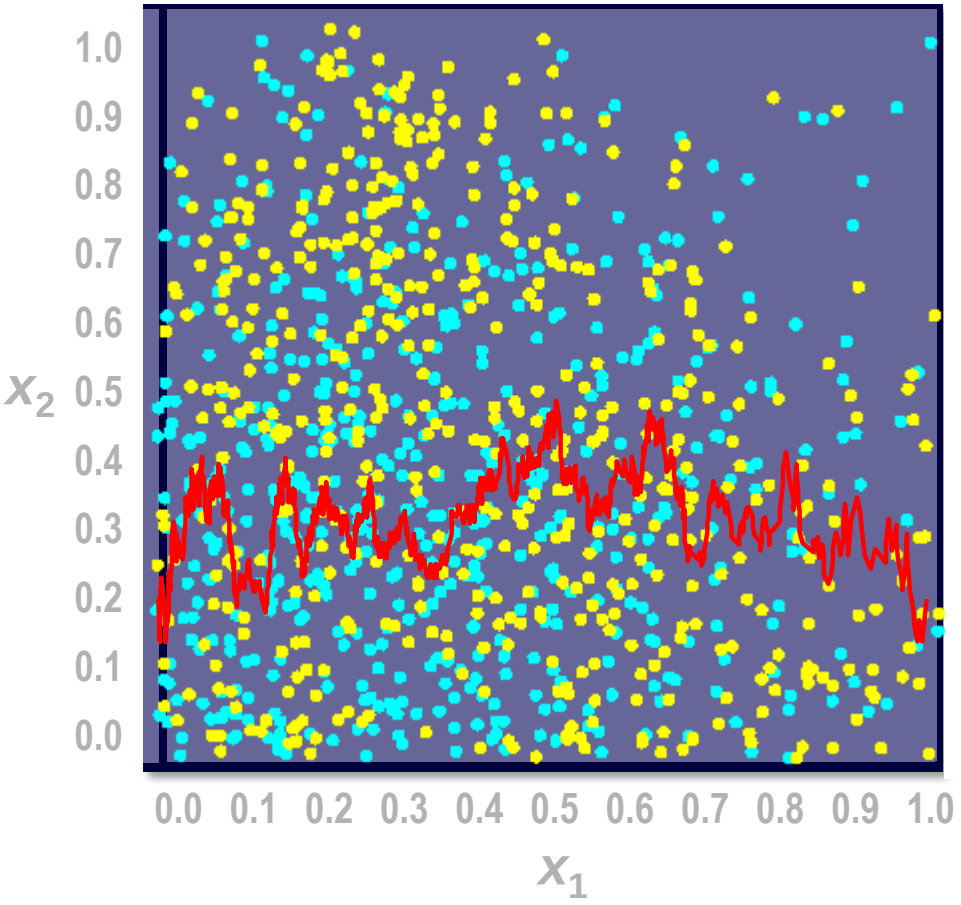
...

Decision Tree Split Search



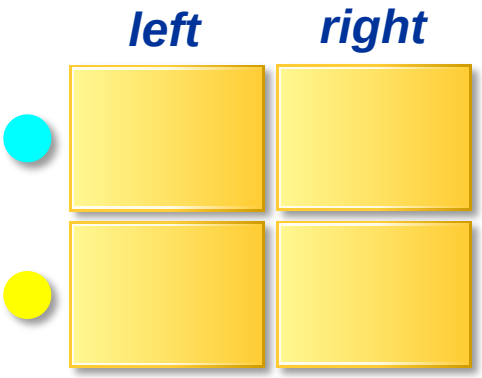
Confusion Matrix

Calculate the *logworth* of every partition on input x_1.

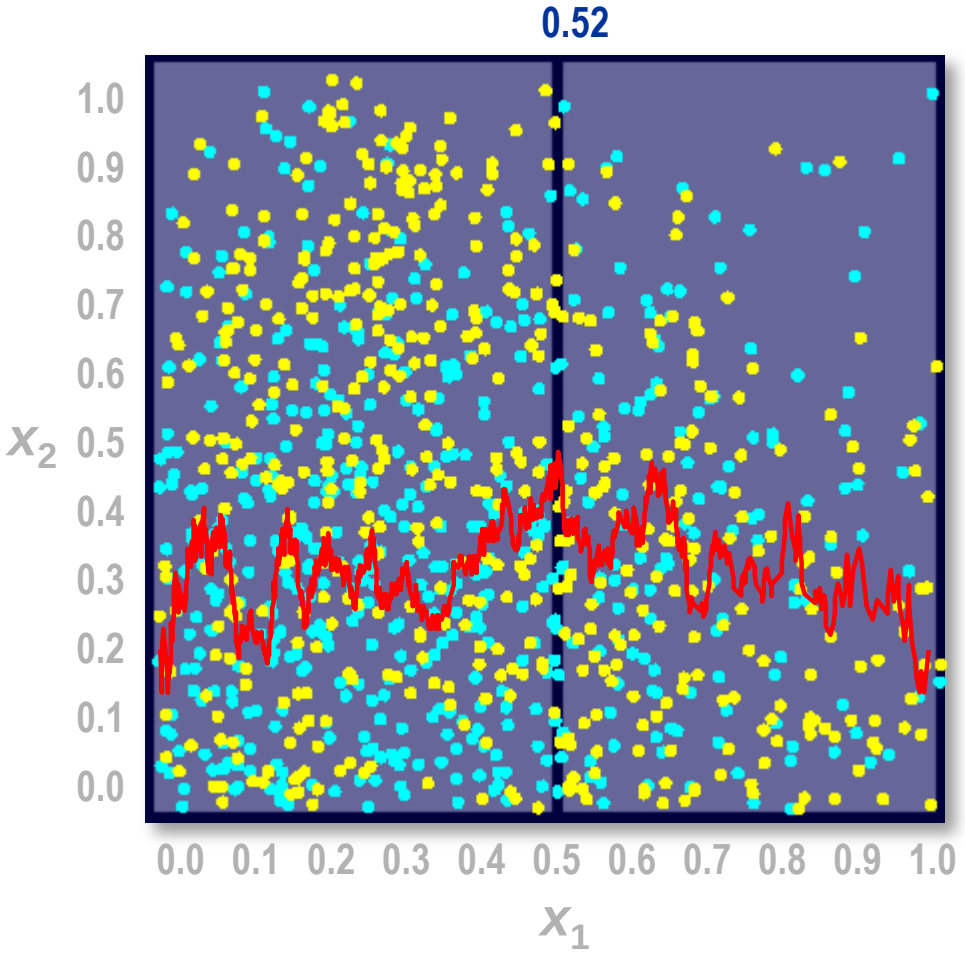


...

Decision Tree Split Search



Calculate the *logworth* of every partition on input x_1.

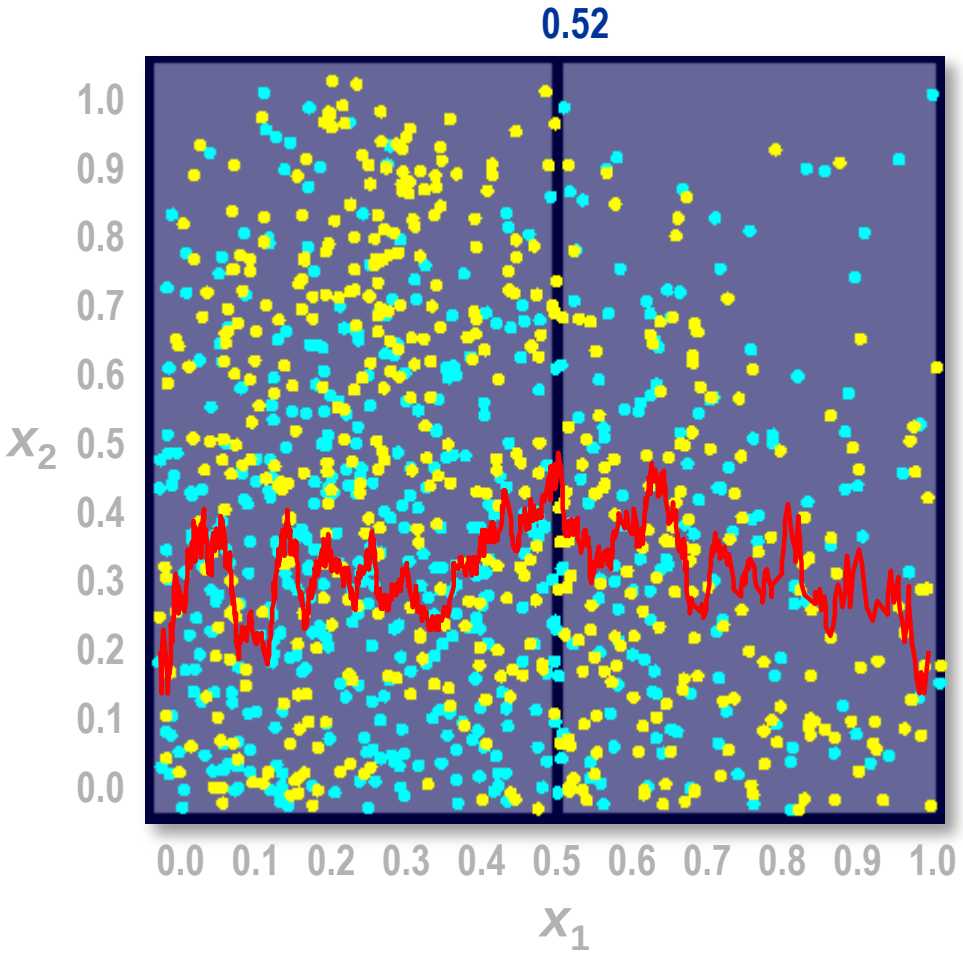


...

Decision Tree Split Search

	left	*right*
●	53%	42%
●	47%	58%

max logworth(x_1)
0.95



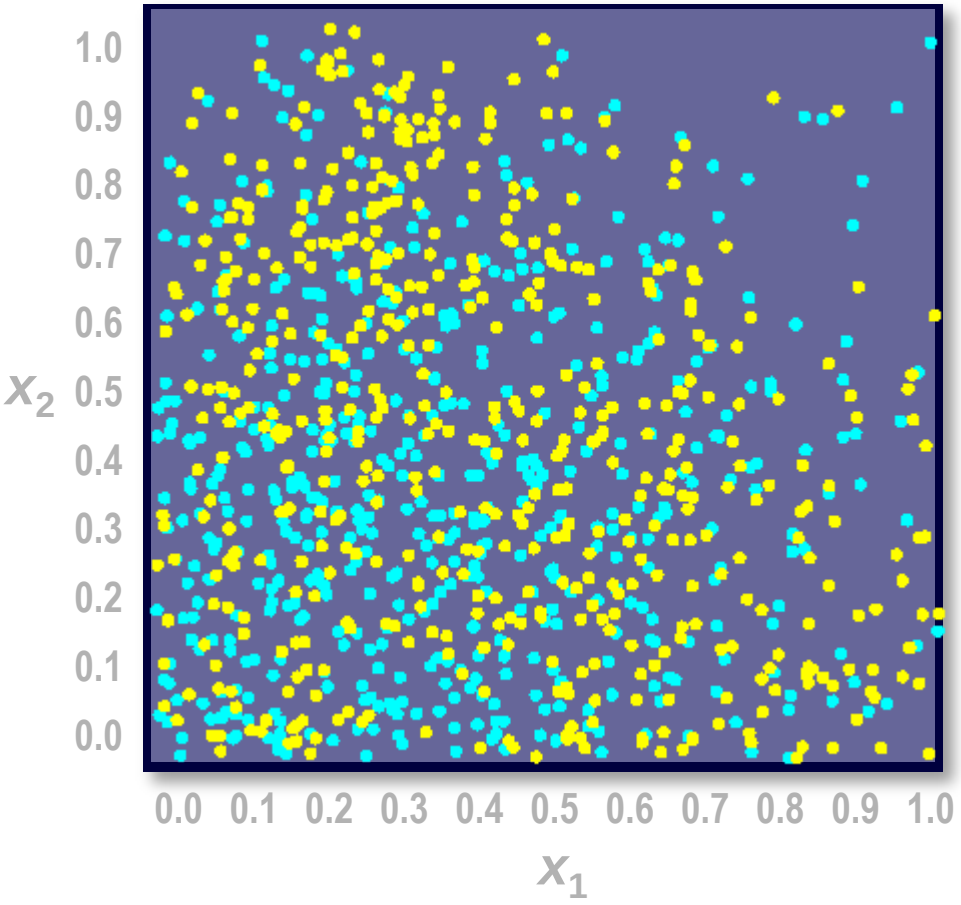
Select the partition with the maximum *logworth*.

...

Decision Tree Split Search

	left	right
○	53%	42%
○	47%	58%

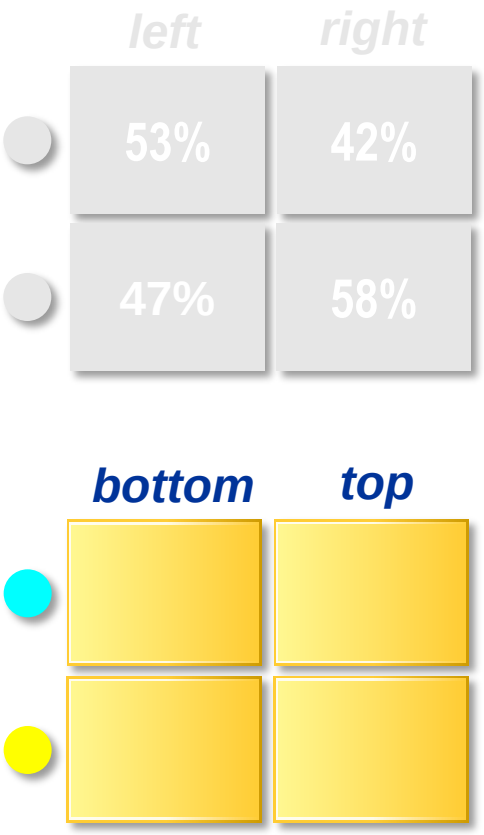
max logworth(x_1) 0.95



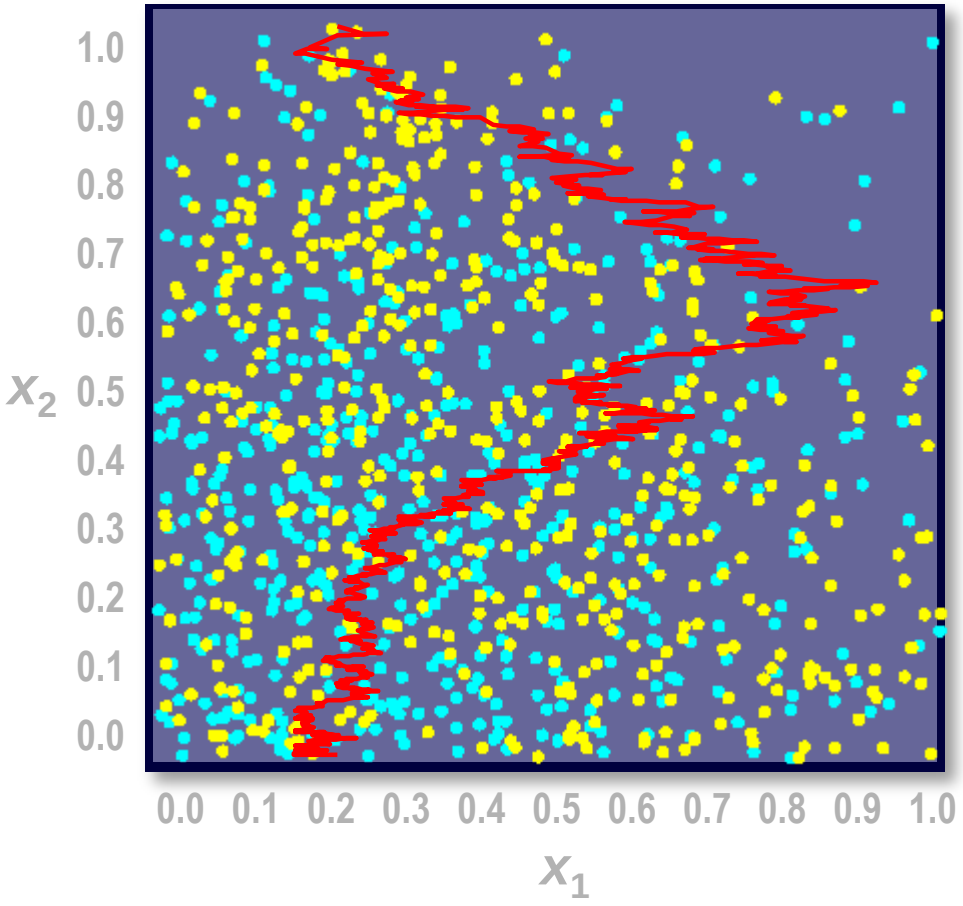
Repeat for input x_2.

...

Decision Tree Split Search



max
logworth(x_1)
0.95



...

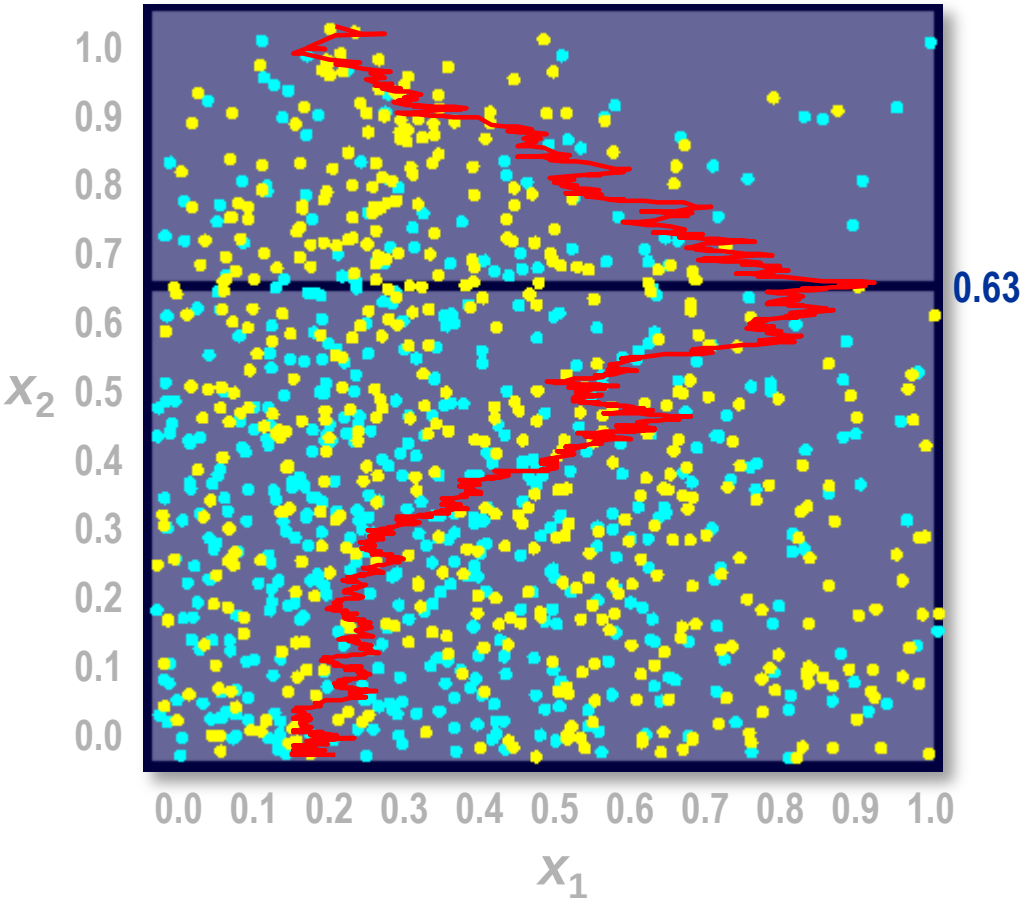
Decision Tree Split Search

	left	right
○	53%	42%
○	47%	58%

max logworth(x_1) 0.95

	bottom	top
●	54%	35%
●	46%	65%

max logworth(x_2) 4.92



...

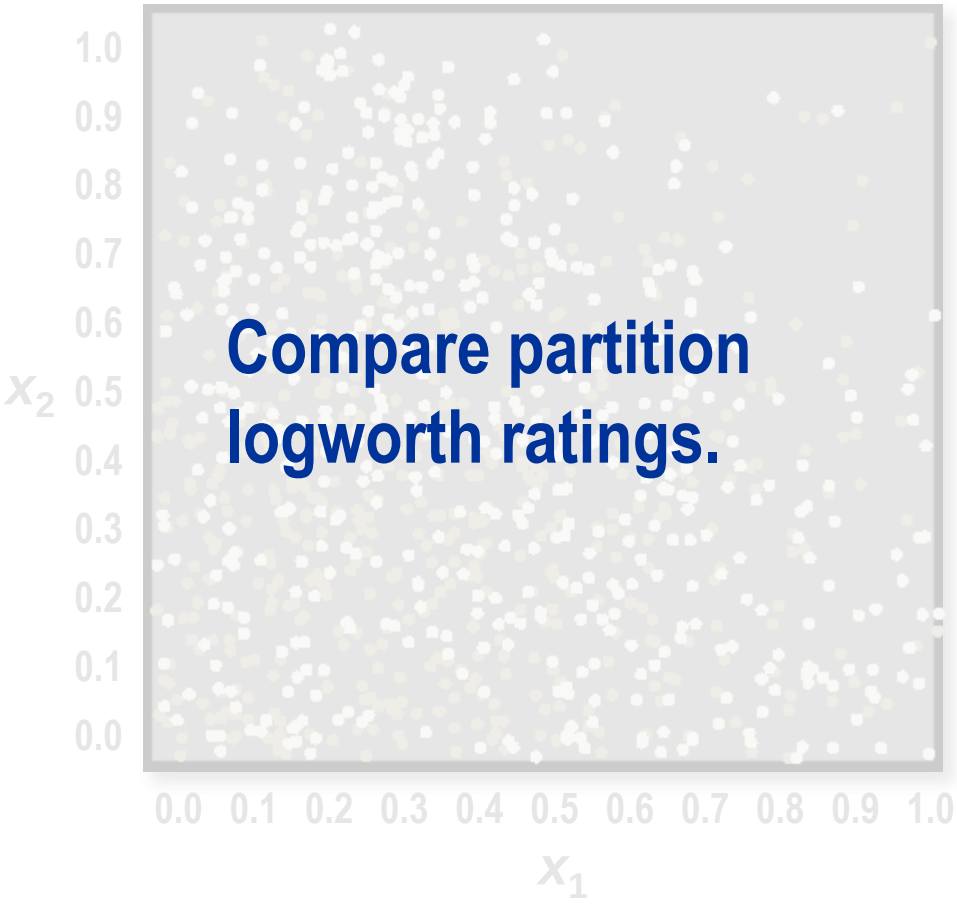
Decision Tree Split Search

	left	*right*
●	53%	42%
●	47%	58%

max logworth(x_1)
0.95

	bottom	*top*
●	54%	35%
●	46%	65%

max logworth(x_2)
4.92



...

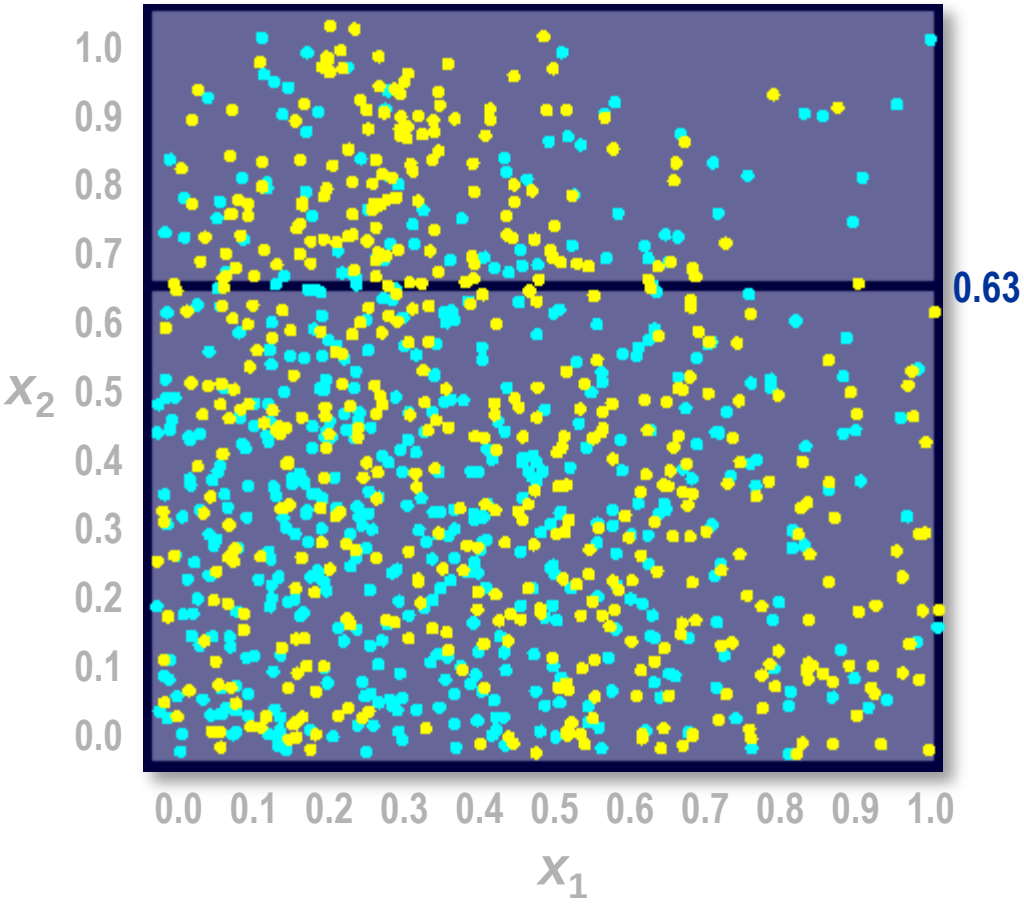
Decision Tree Split Search

	left	right
○	53%	42%
○	47%	58%

max logworth(x_1) 0.95

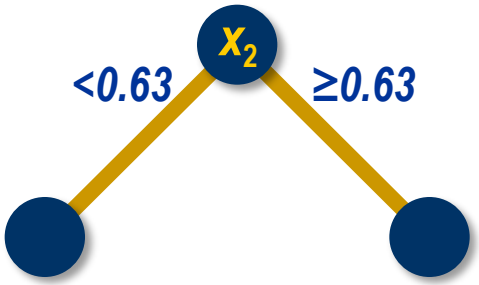
	bottom	top
●	54%	35%
●	46%	65%

max logworth(x_2) 4.92

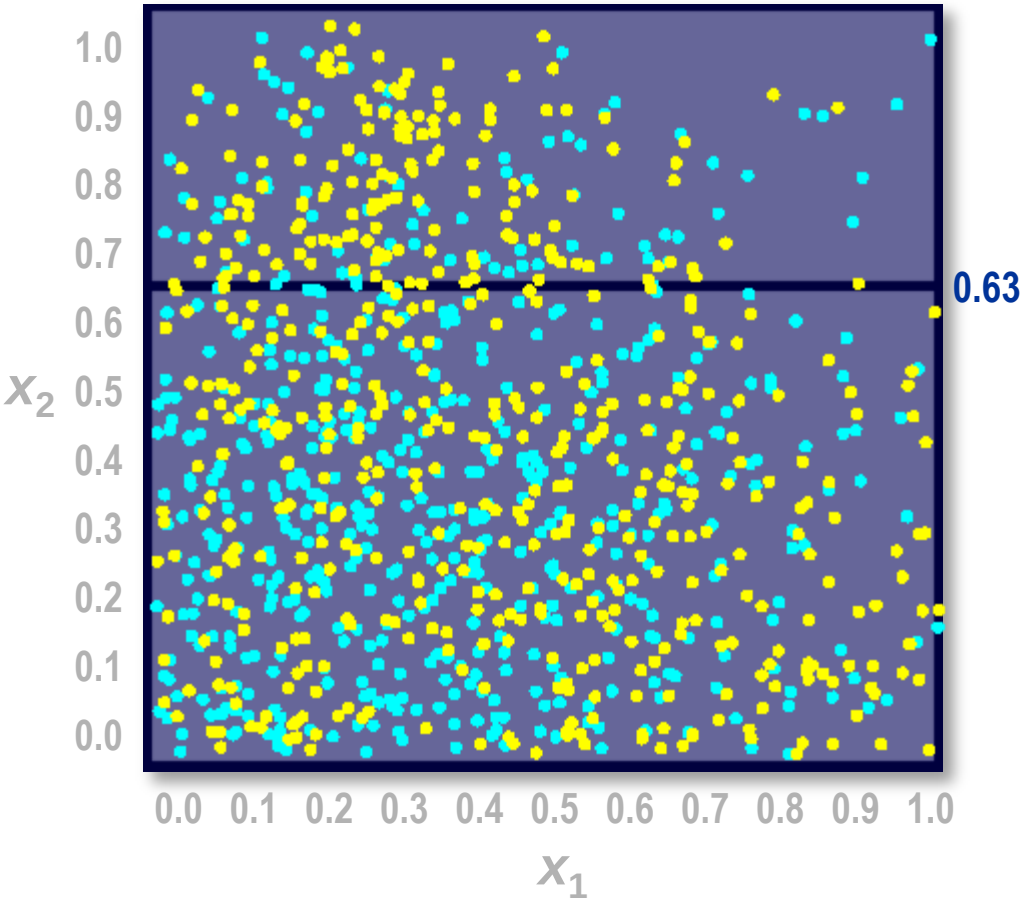


...

Decision Tree Split Search

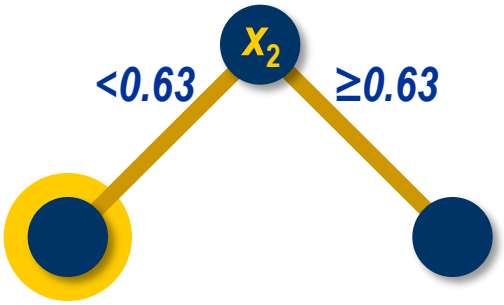


Create a partition rule from the best partition across all inputs.

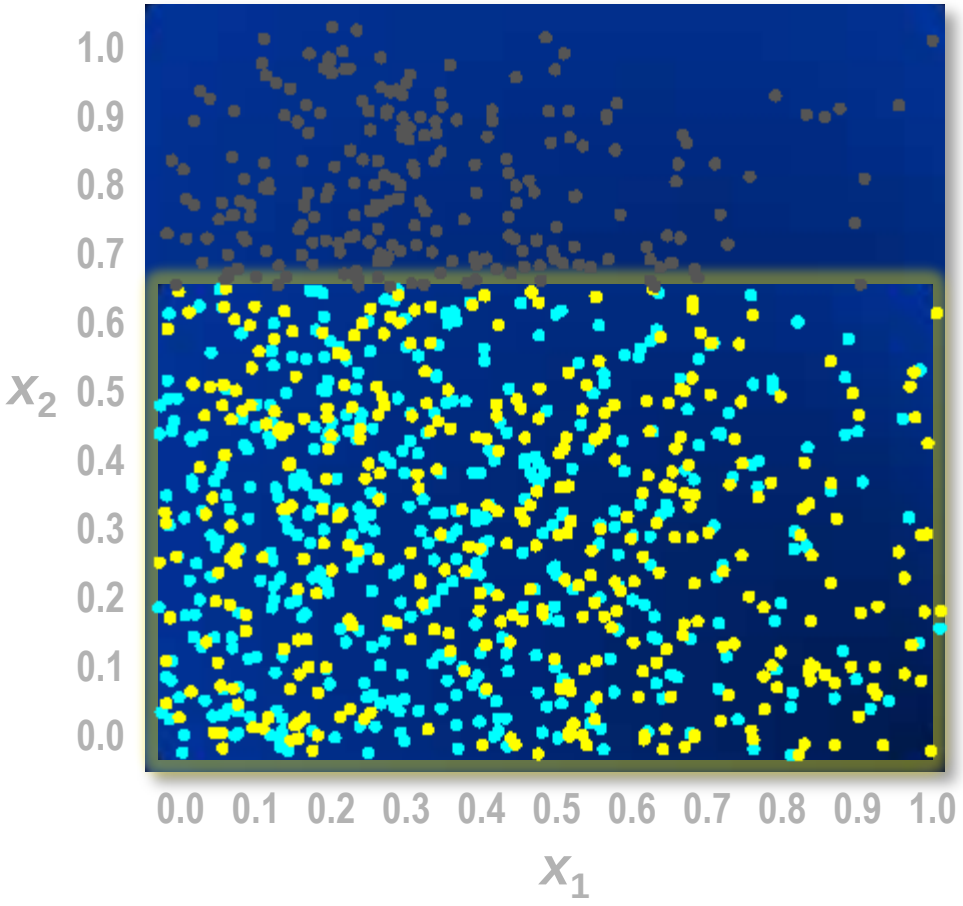


...

Decision Tree Split Search

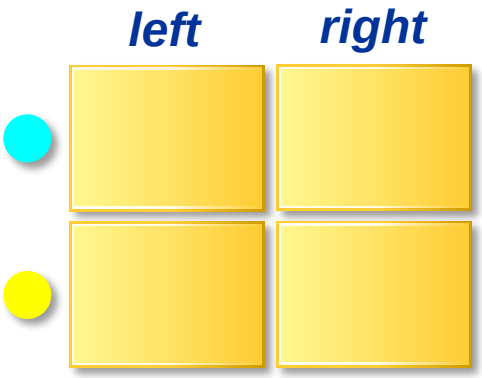


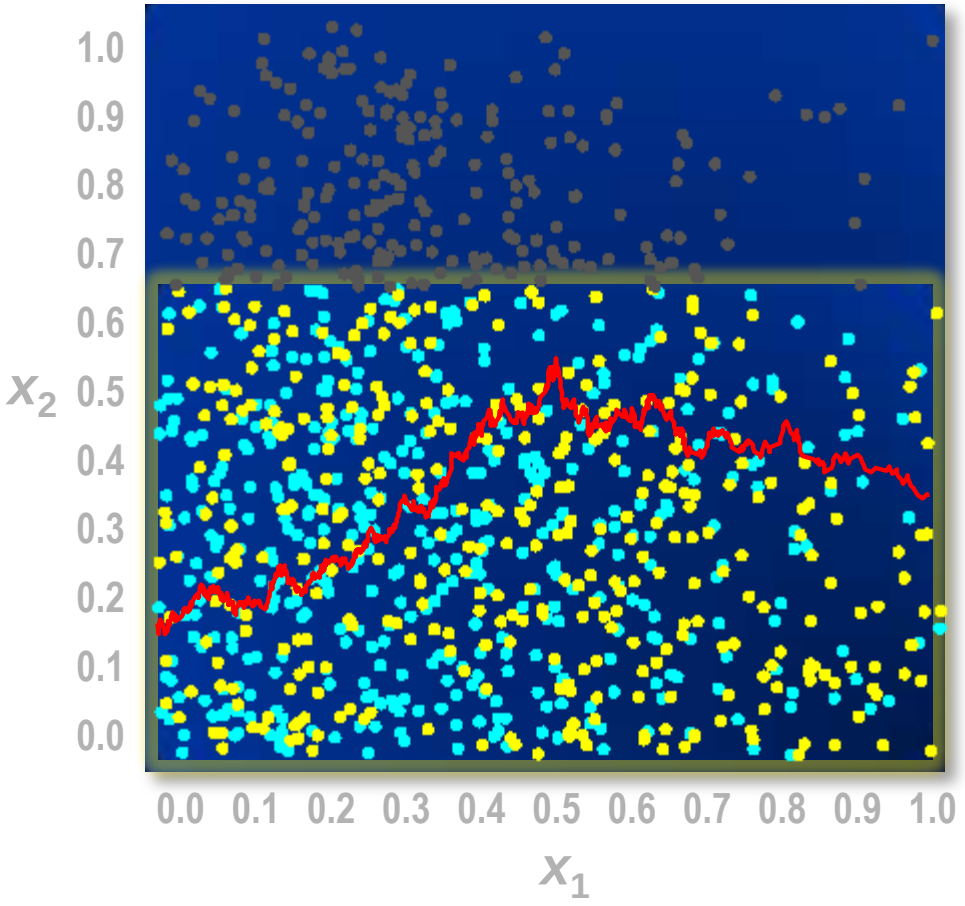
Repeat the process in each subset.



...

Decision Tree Split Search



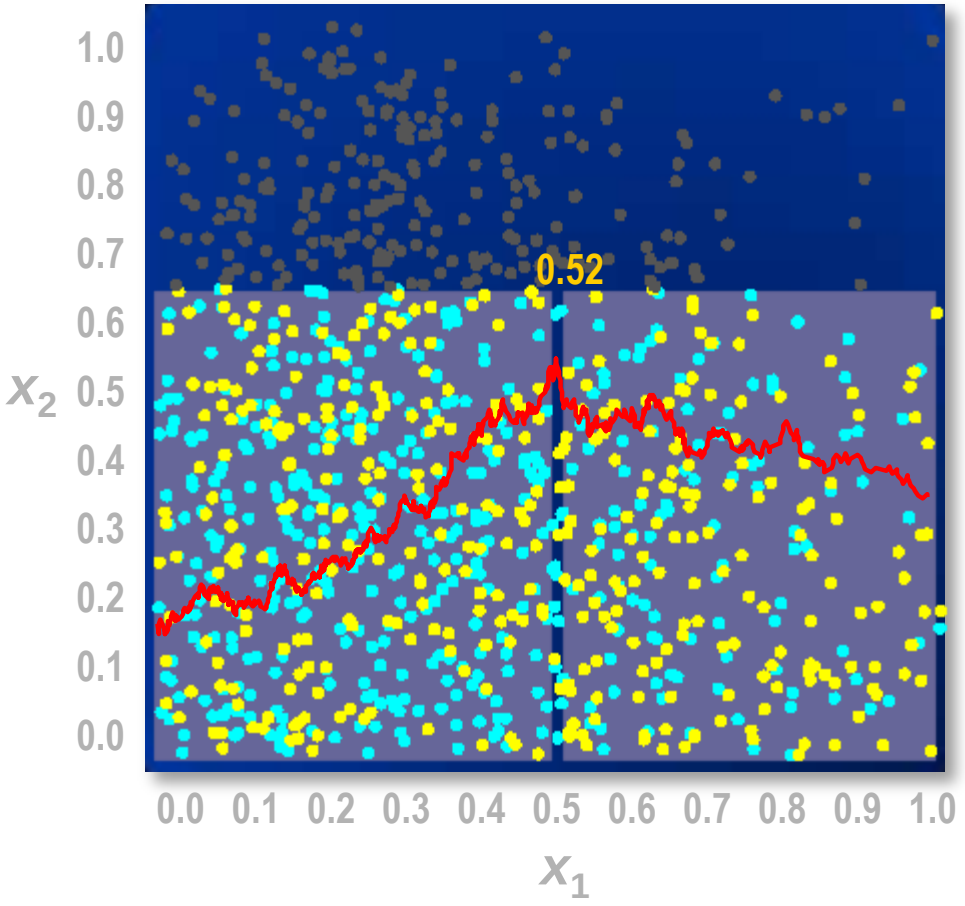


...

Decision Tree Split Search

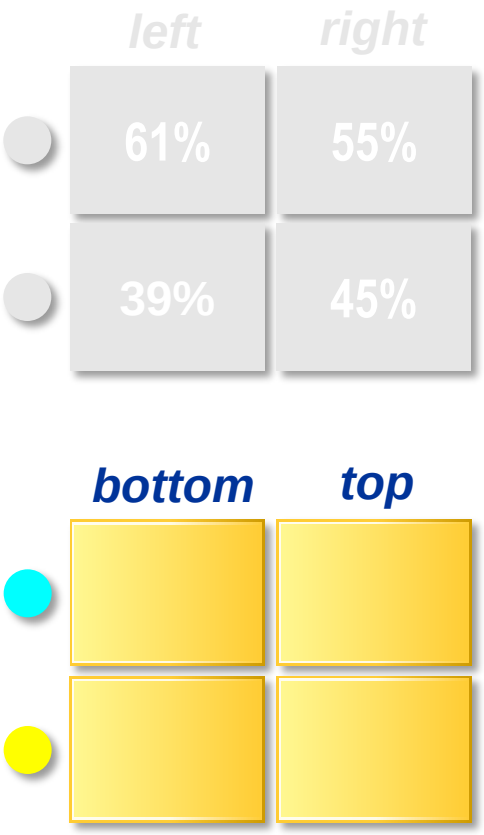
	left	*right*
●	61%	55%
●	39%	45%

max
logworth(x_1)
5.72

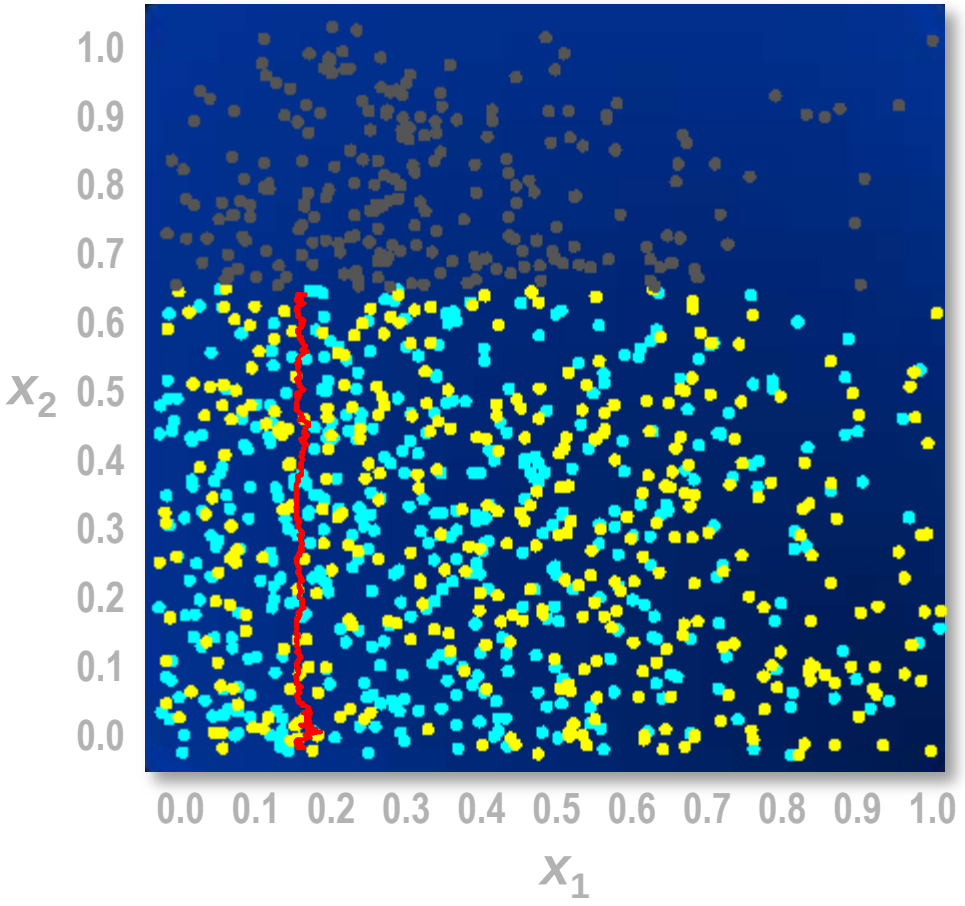


...

Decision Tree Split Search



max
logworth(x_1)
5.72



...

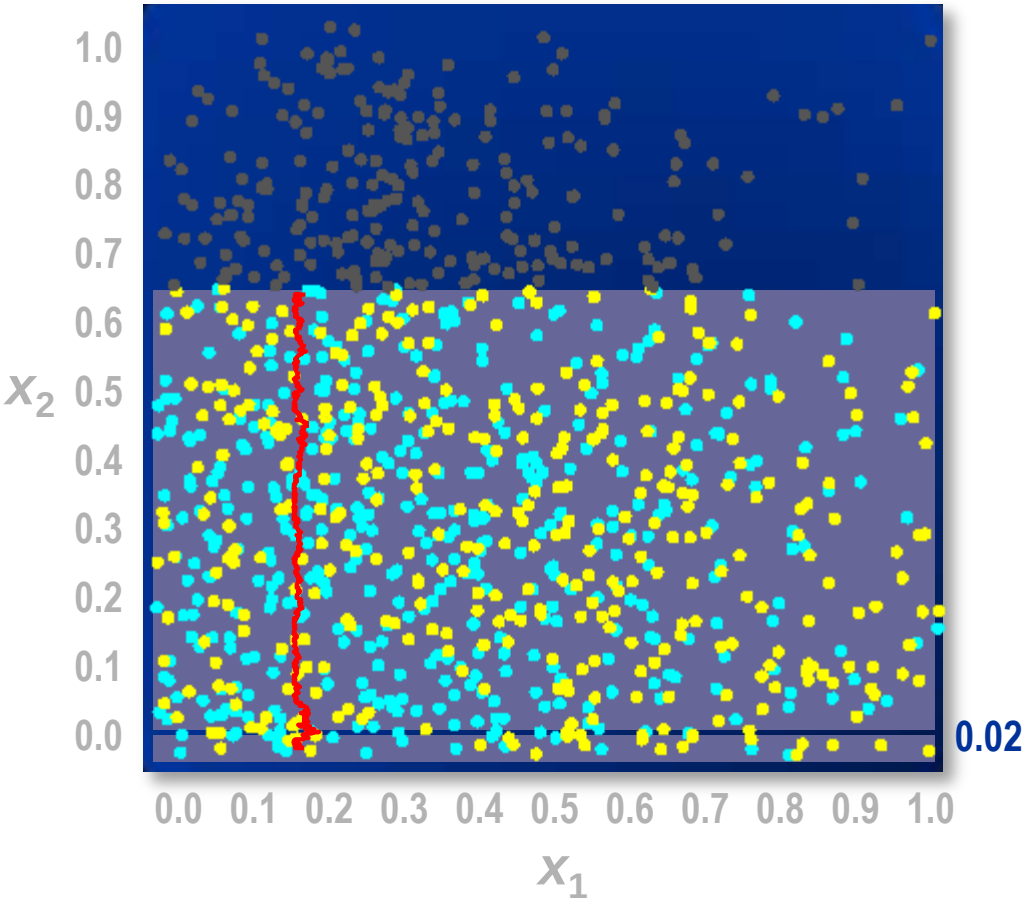
Decision Tree Split Search

	left	right
○	61%	55%
○	39%	45%

max logworth(x_1) 5.72

	bottom	top
●	38%	55%
●	62%	45%

max logworth(x_2) -2.01



...

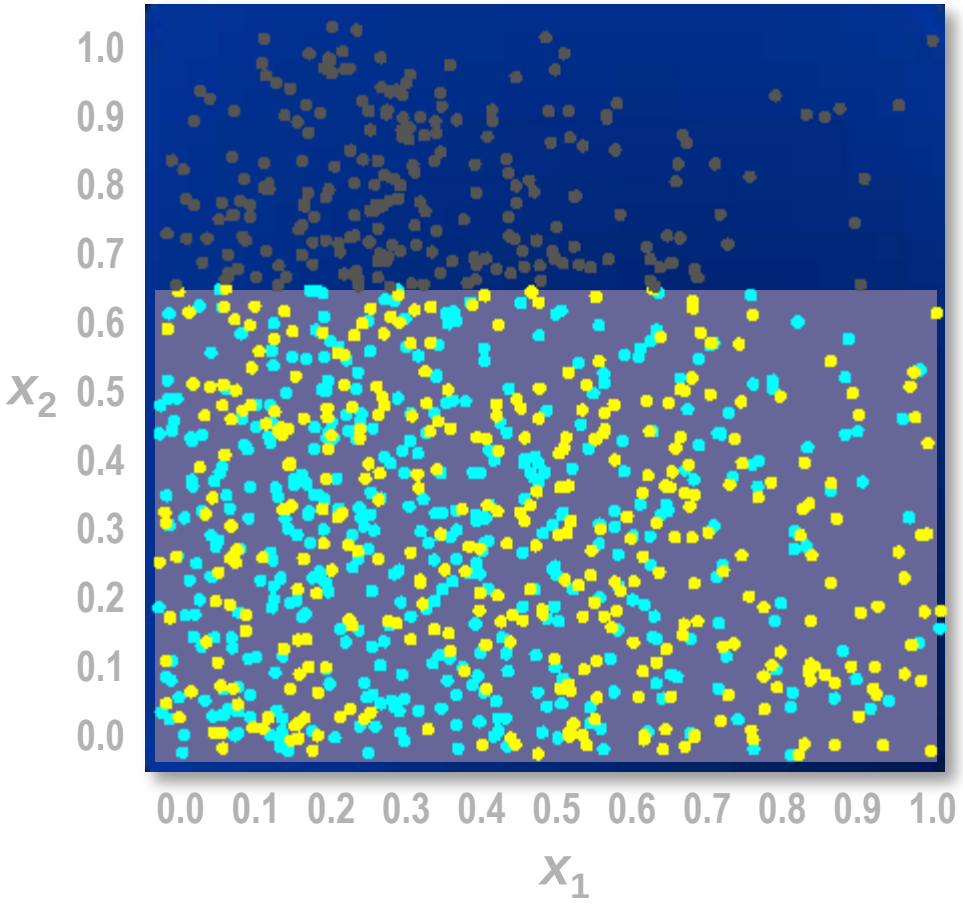
Decision Tree Split Search

	left	*right*
●	61%	55%
●	39%	45%

max logworth(x_1)
5.72

	bottom	*top*
●	38%	55%
●	62%	45%

max logworth(x_2)
-2.01



...

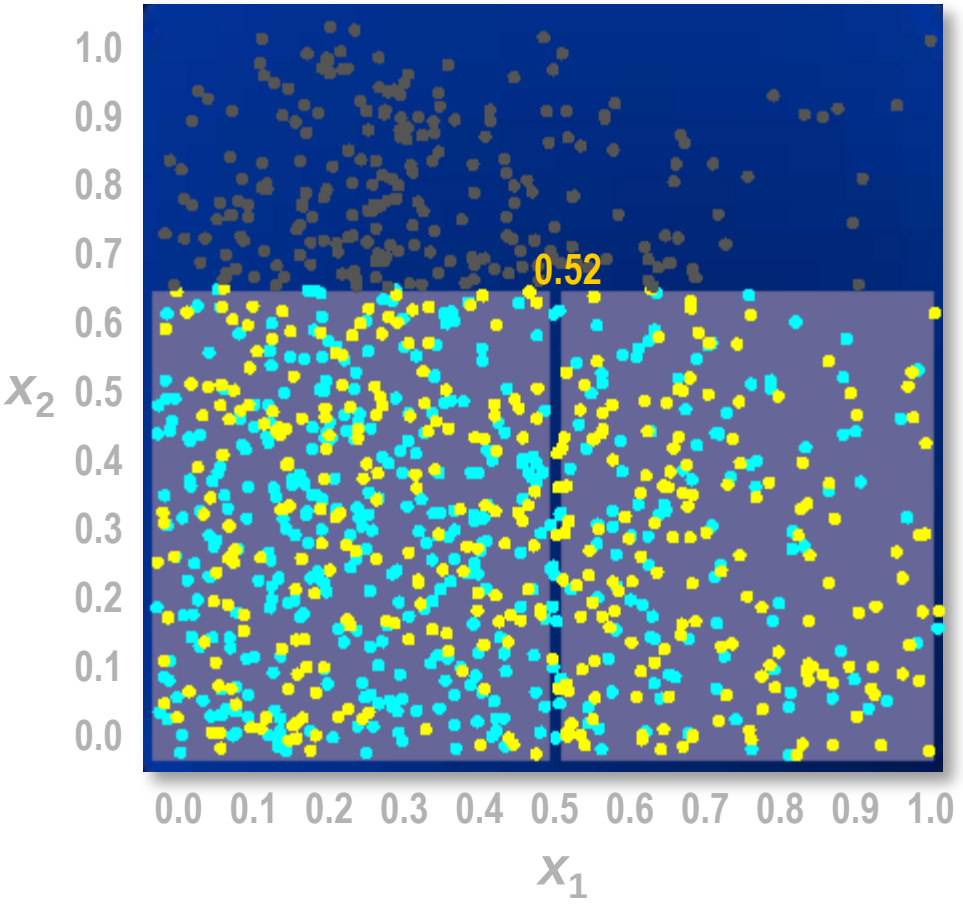
Decision Tree Split Search

	left	*right*
●	61%	55%
●	39%	45%

max logworth(x_1)
5.72

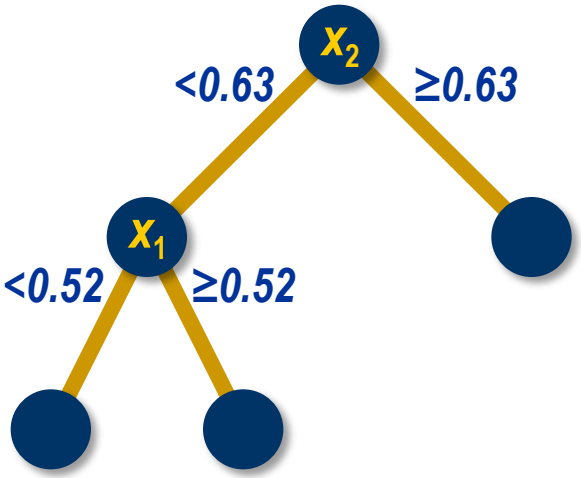
	bottom	*top*
●	38%	55%
●	62%	45%

max logworth(x_2)
-2.01

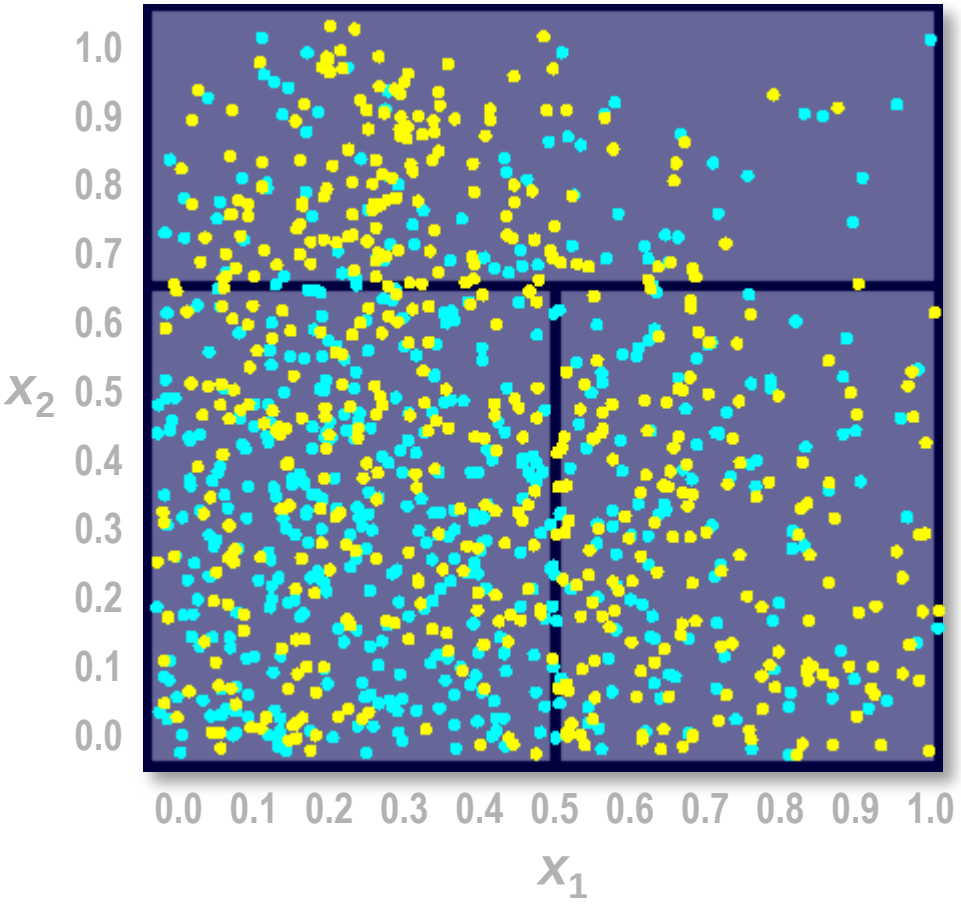


...

Decision Tree Split Search

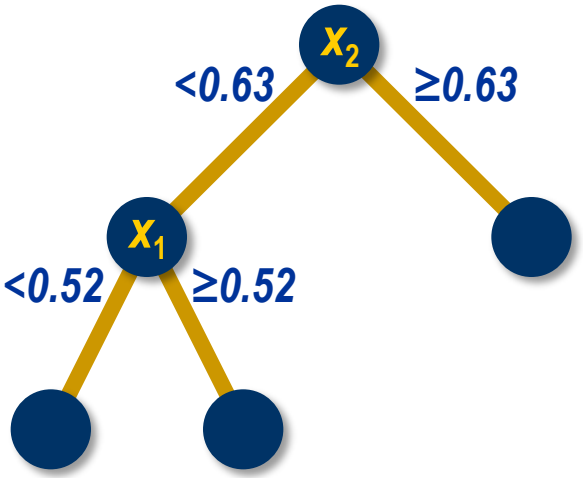


Create a second partition rule.

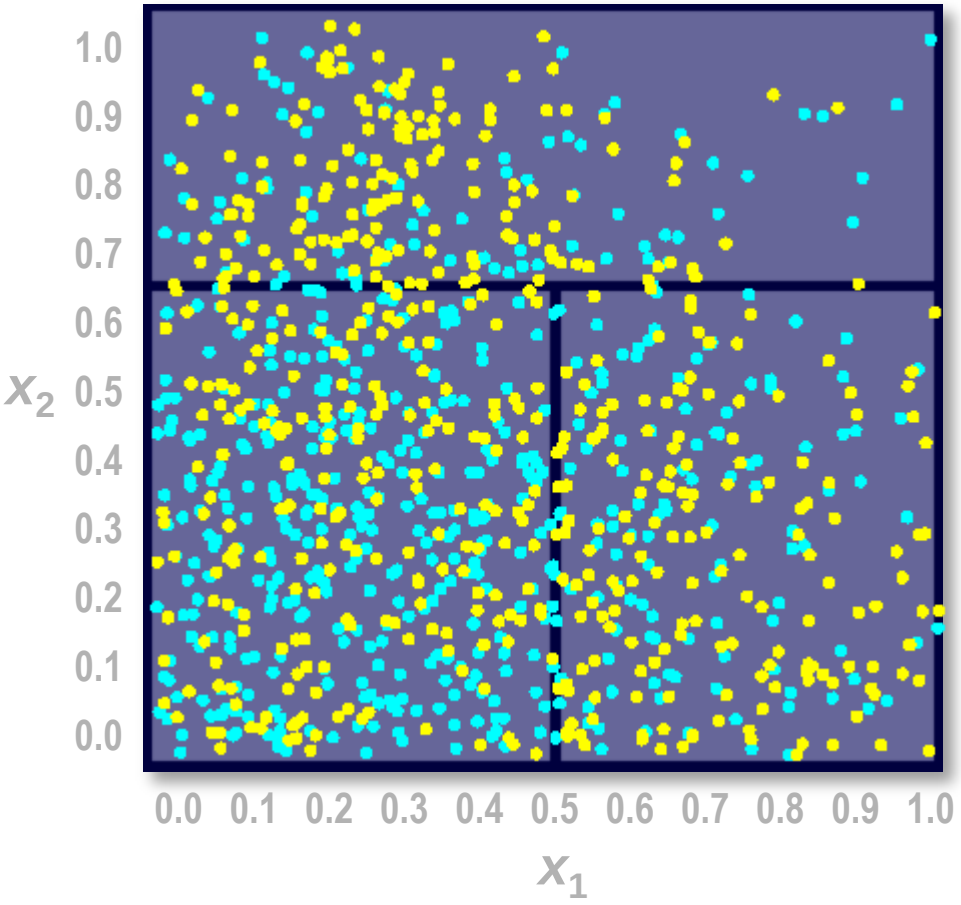


...

Decision Tree Split Search

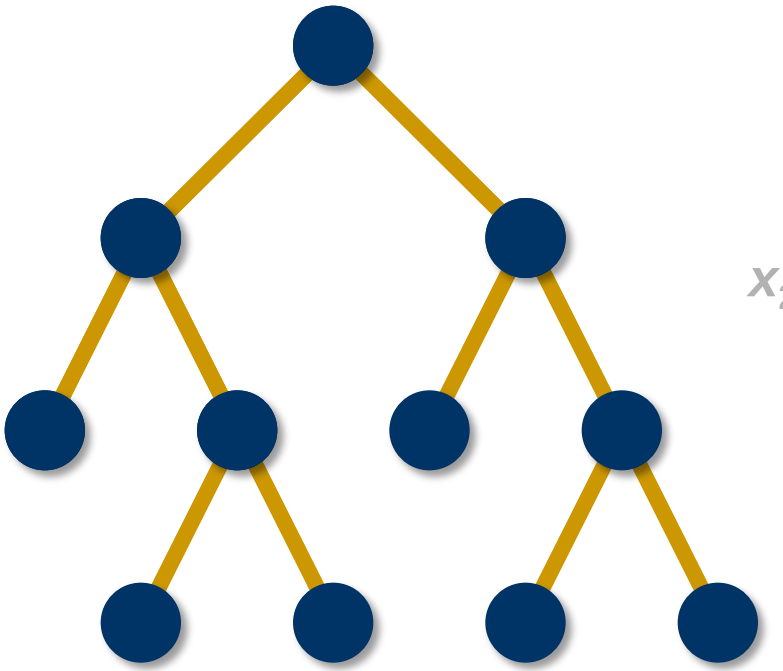


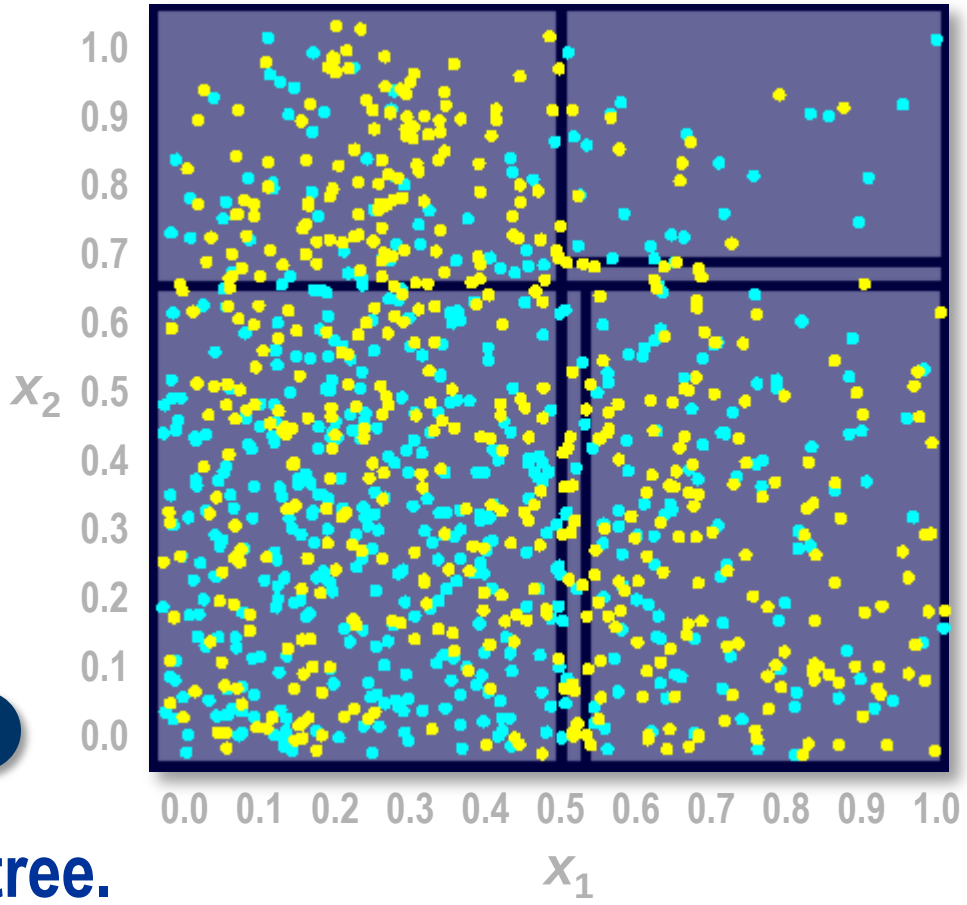
Create a second partition rule.



...

Decision Tree Split Search





Repeat to form a maximal tree.

...

Example

- Two Year School on Texas & Mexico Border
- Strong in Mathematics and Sciences
- Weak in the Arts
- Half of the students are from newly emigrated families

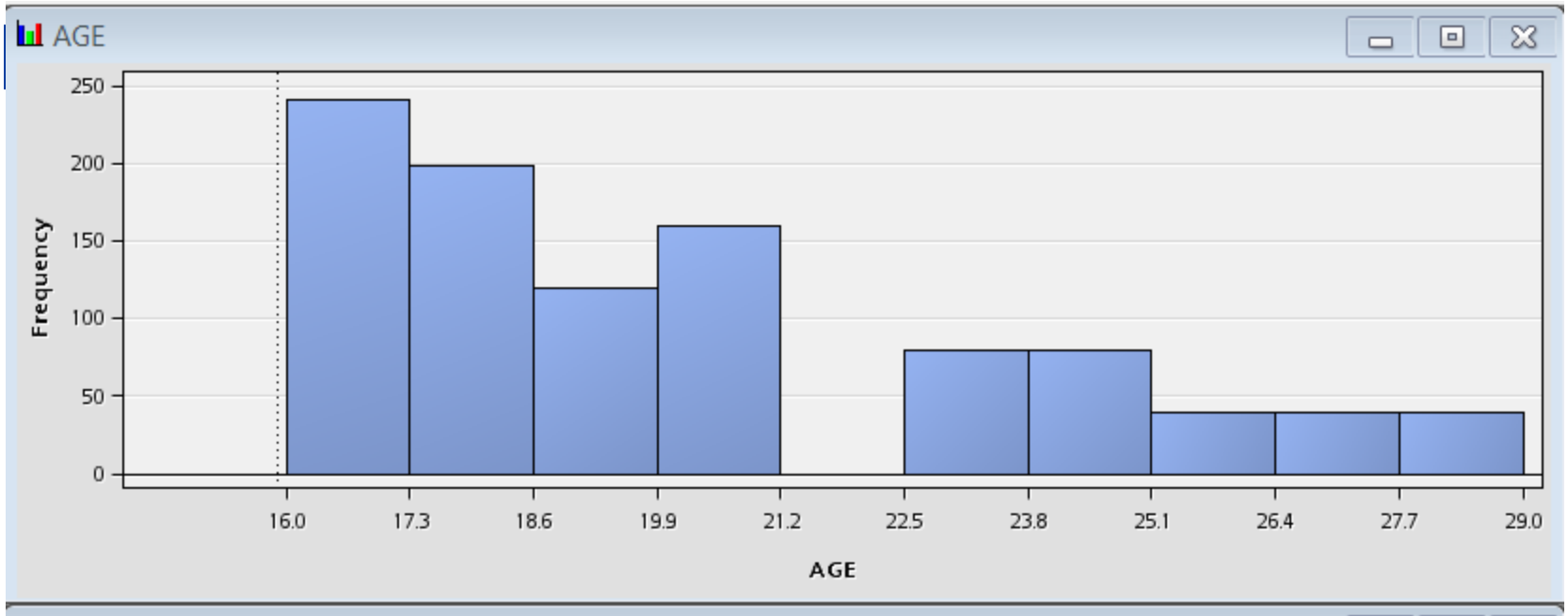
Objective

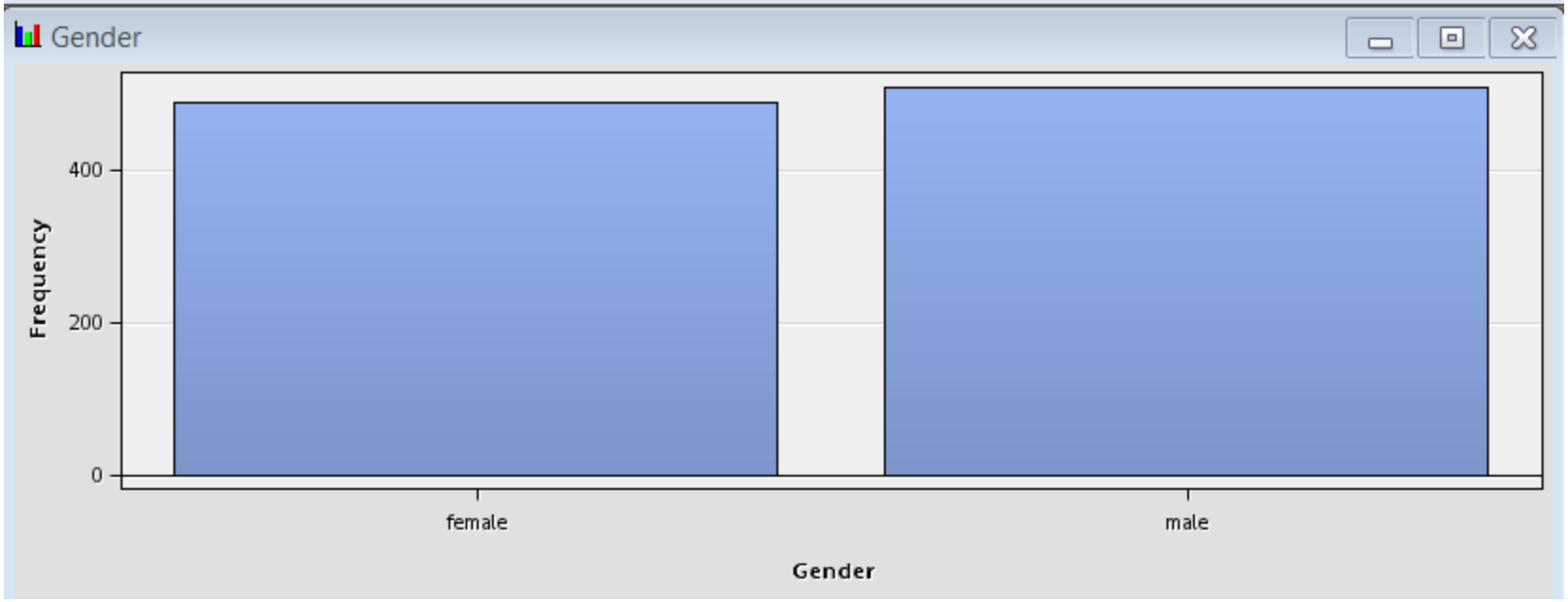
- Improve Graduation Rate
- Identify Students Most Likely Not to Graduate
- Collect Data and Build a Predictive Model
- Determine What Intervention is Approximate

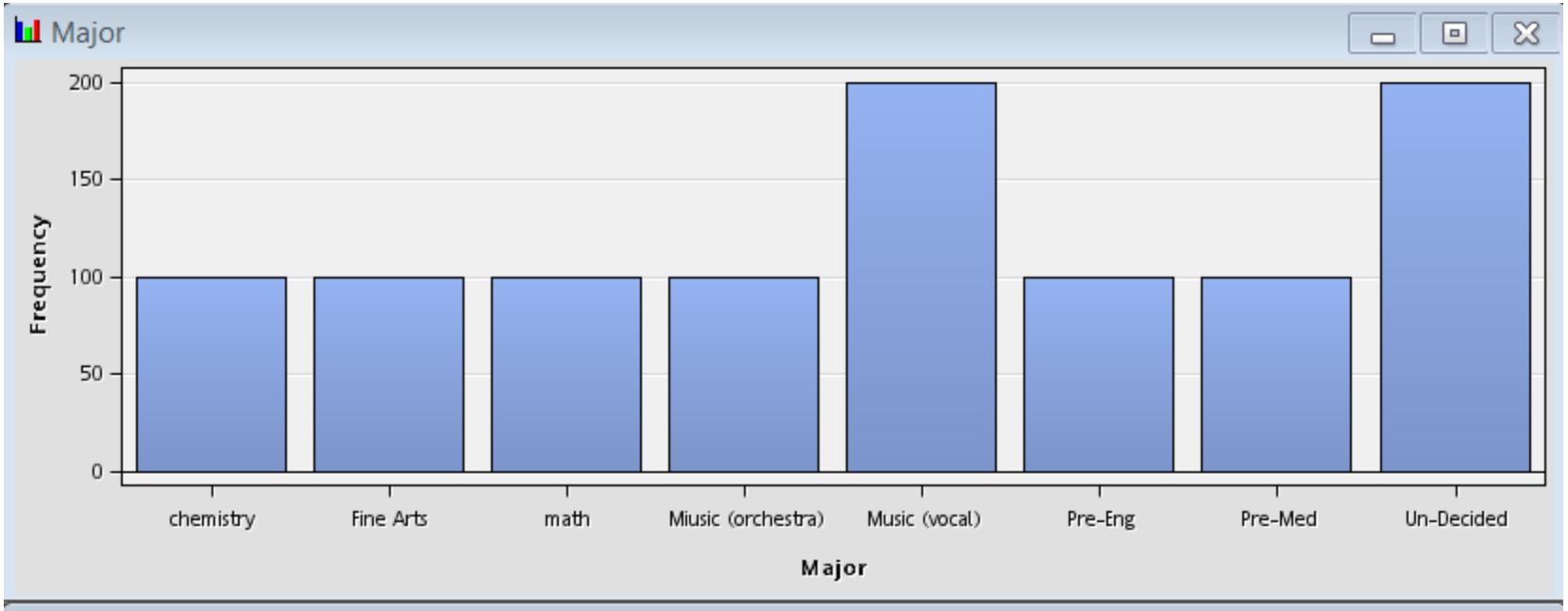
Hypothetical Data

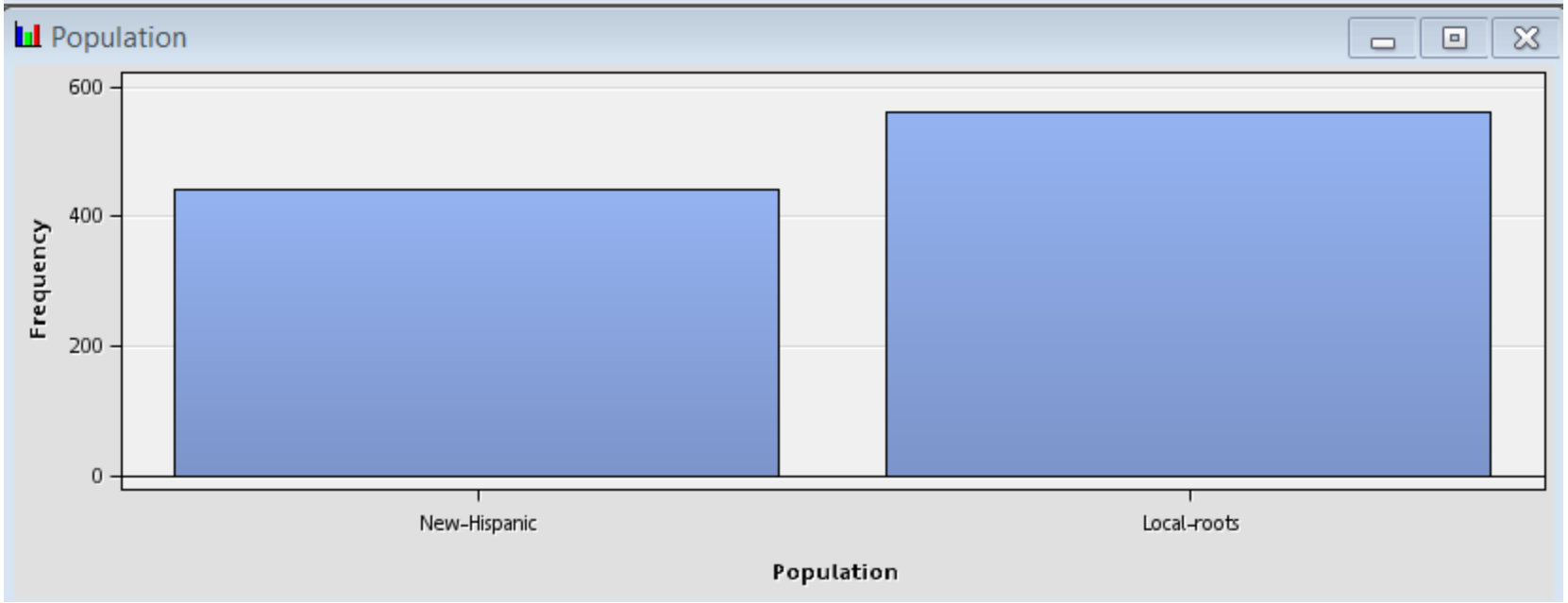
- Sample of 1000 Students Entering Fall 2010
- Determine Which Students Had Left by Fall 2013
- Data Fields

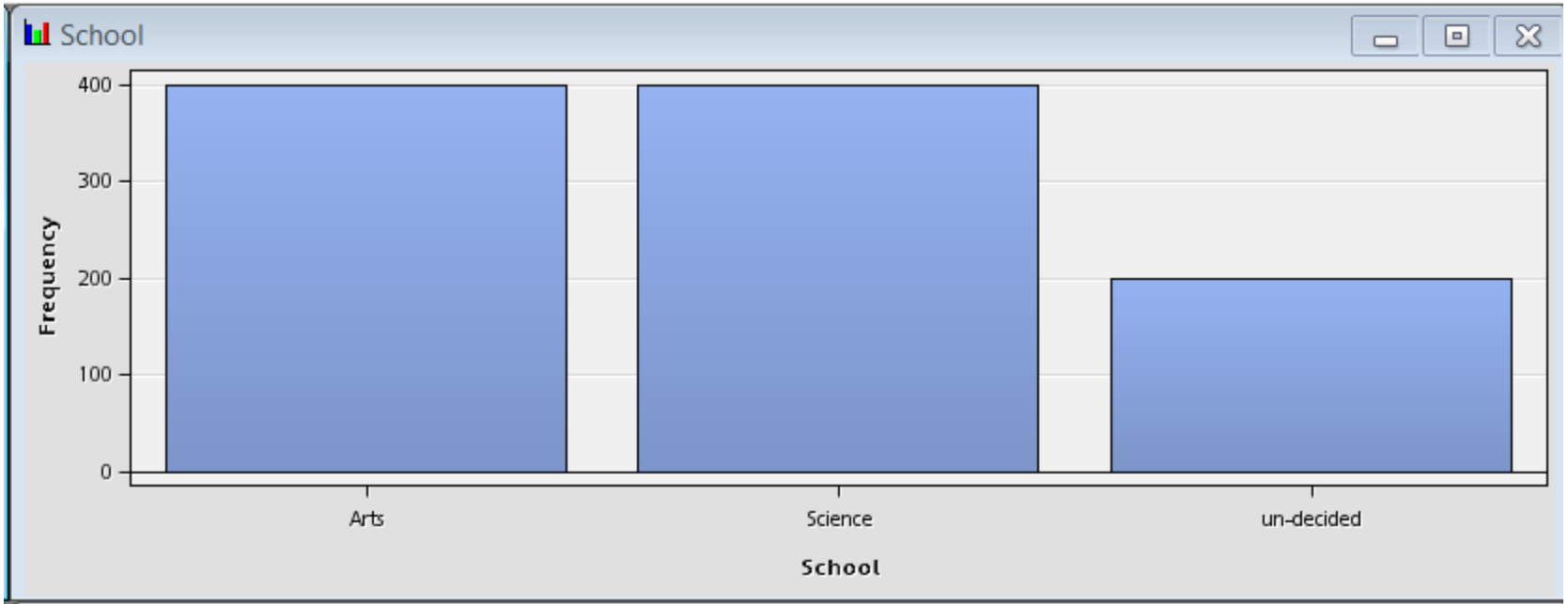
1. Student ID
2. Age
3. Gender
4. Major
5. Population
6. School
7. Enrollment Statement
8. Target

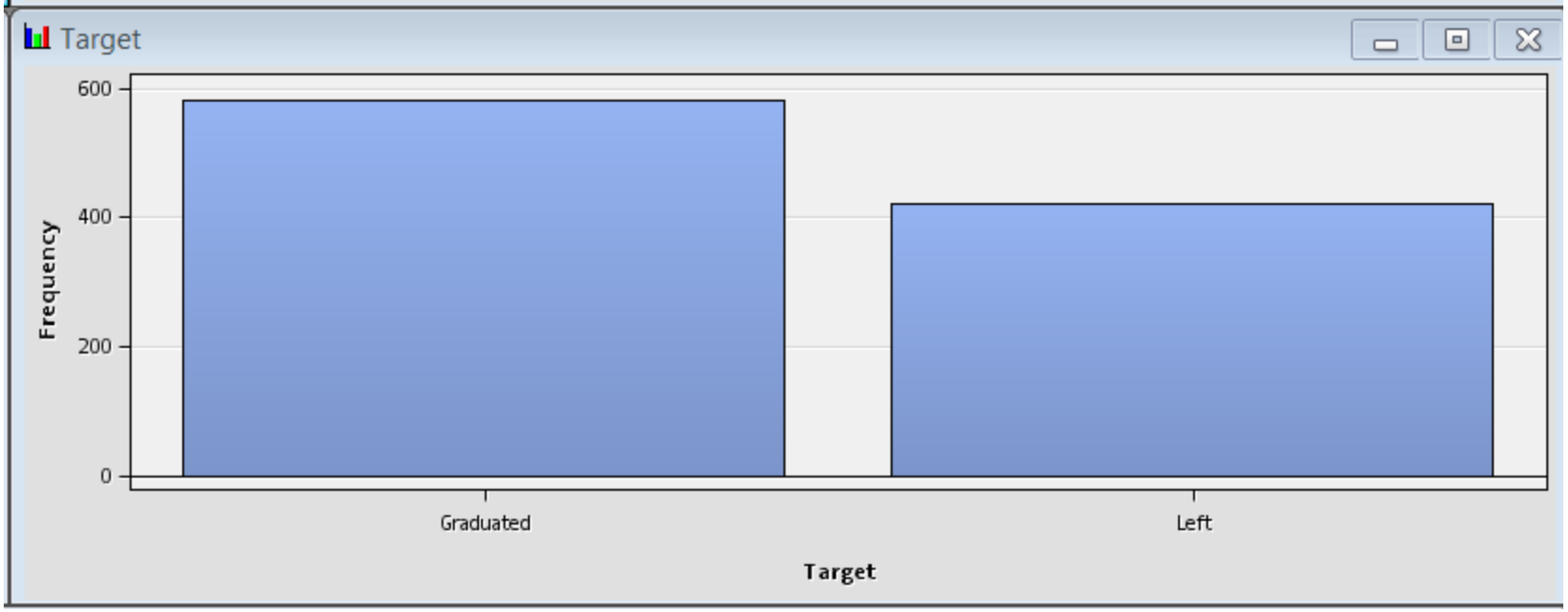

AGE
250
200
150
100
50
0
Frequency
16.0
17.3
18.6
19.9
21.2
22.5
23.8
25.1
26.4
27.7
29.0
AGE


Gender
400
200
0
Frequency
female
male
Gender

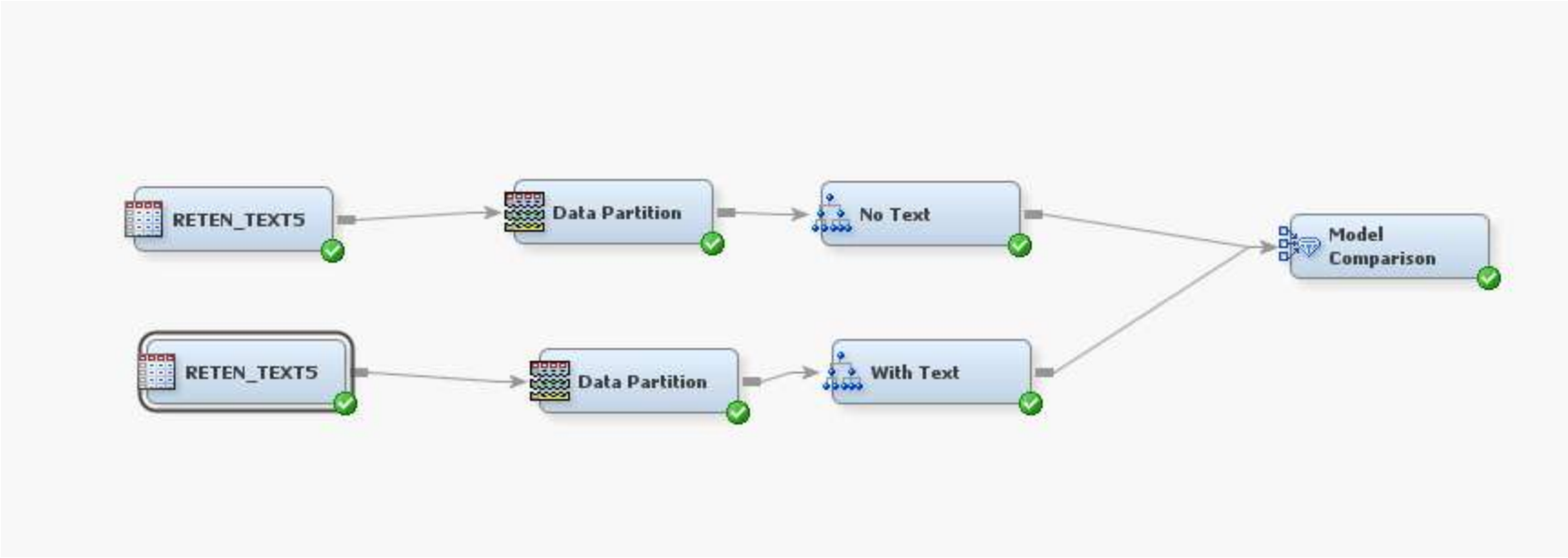

Major
200
150
100
50
0
Frequency
chemistry
Fine Arts
math
Miusic (orchestra)
Music (vocal)
Pre-Eng
Pre-Med
Un-Decided
Major


Population
600
400
200
0
Frequency
New-Hispanic
Local-roots
Population

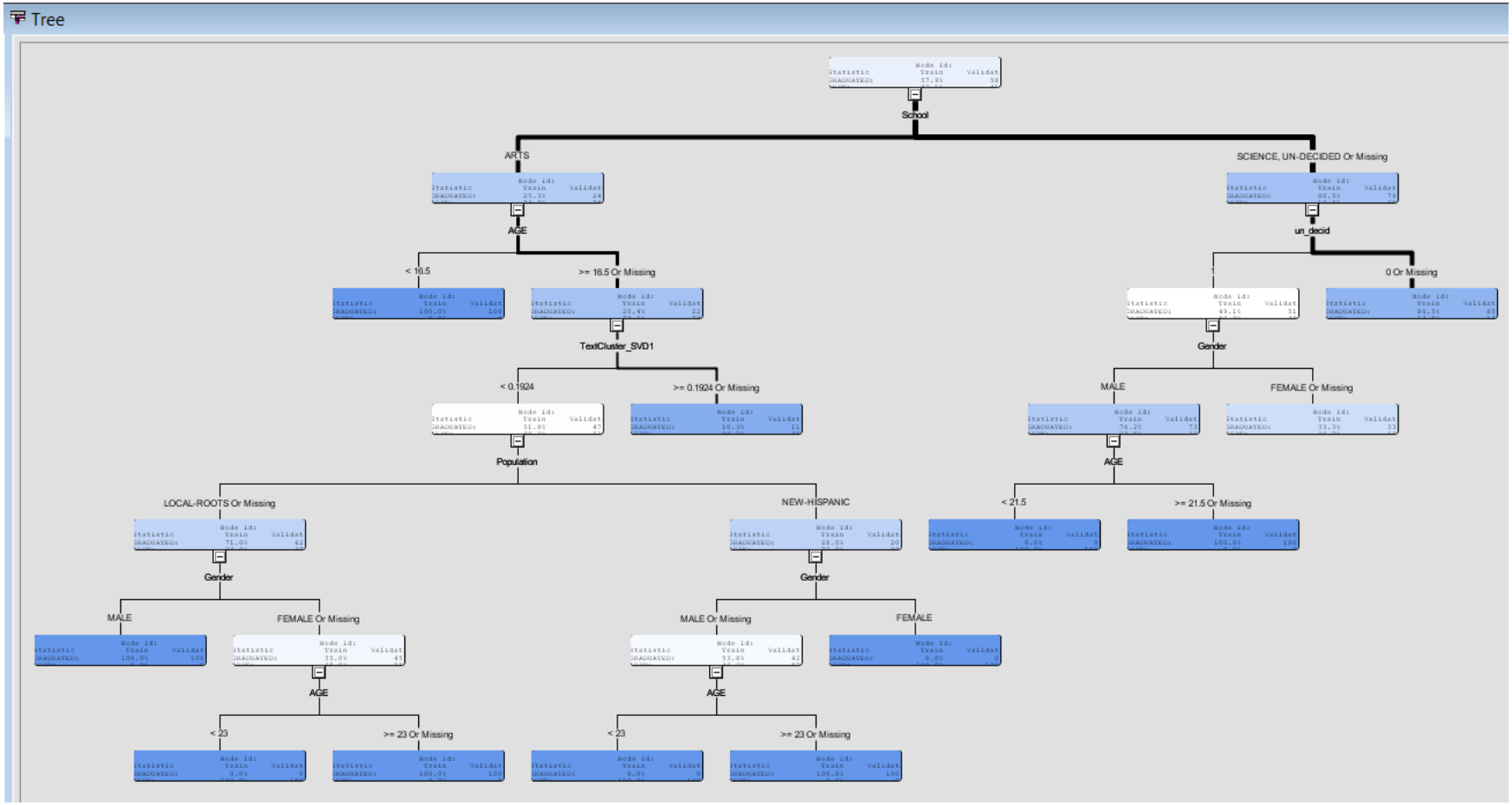

School
400
300
200
100
0
Frequency
Arts
Science
un-decided
School


Target
600
400
200
0
Frequency
Graduated
Left
Target

Process Flow



Decision Tree Model



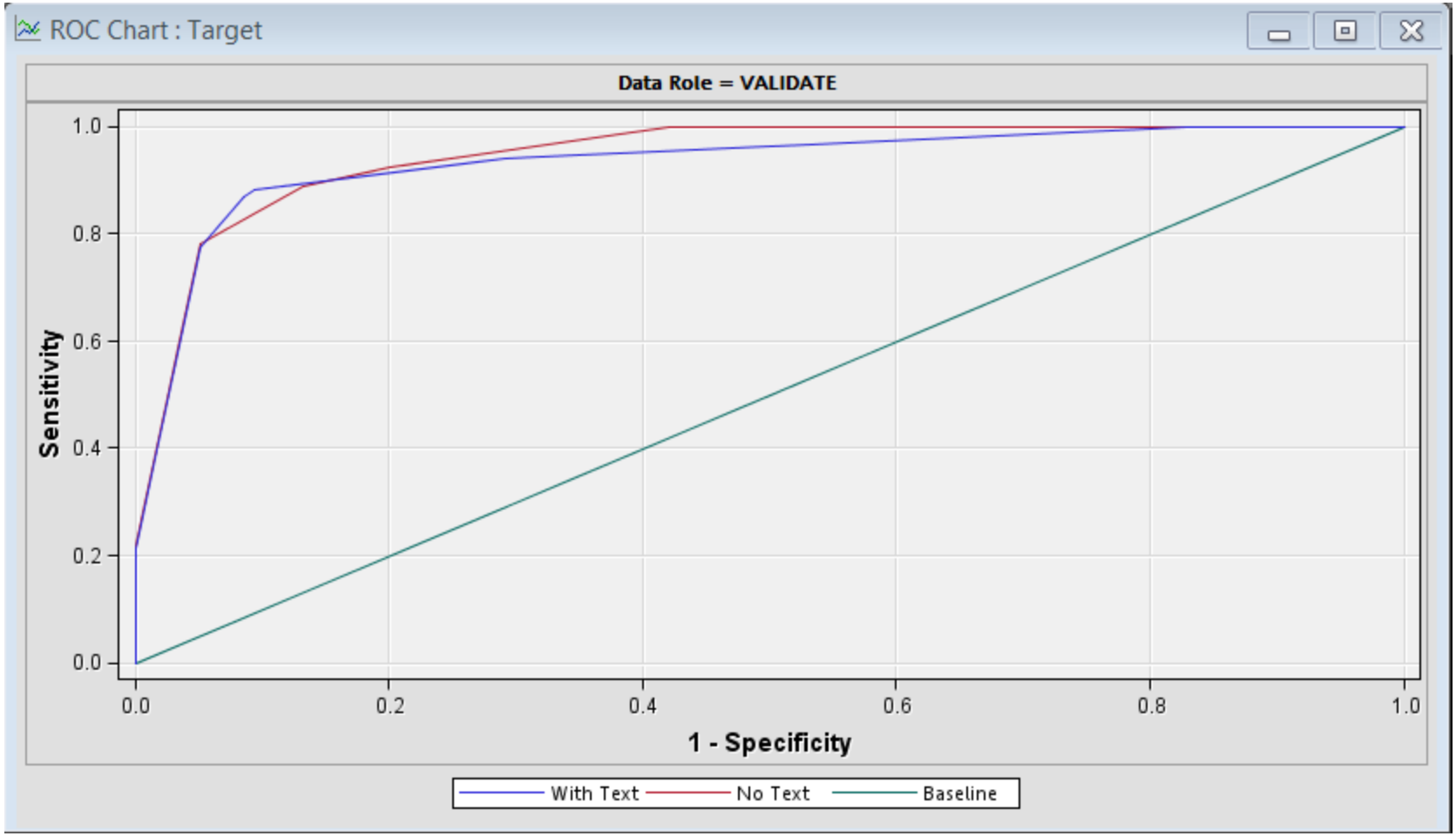
Fit Statistics

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	With Text	0.10474	0.077413	0.09683	0.086645
	Tree	No Text	0.12469	0.069562	0.10017	0.089264

ROC Curve


ROC Chart : Target
Data Role = VALIDATE
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
With Text
No Text
Baseline

Next Steps

- Score Students Entering in Fall 2013 With Model
- Distribute Scoring Information to Approximate People
- Evaluate Model After Two Years