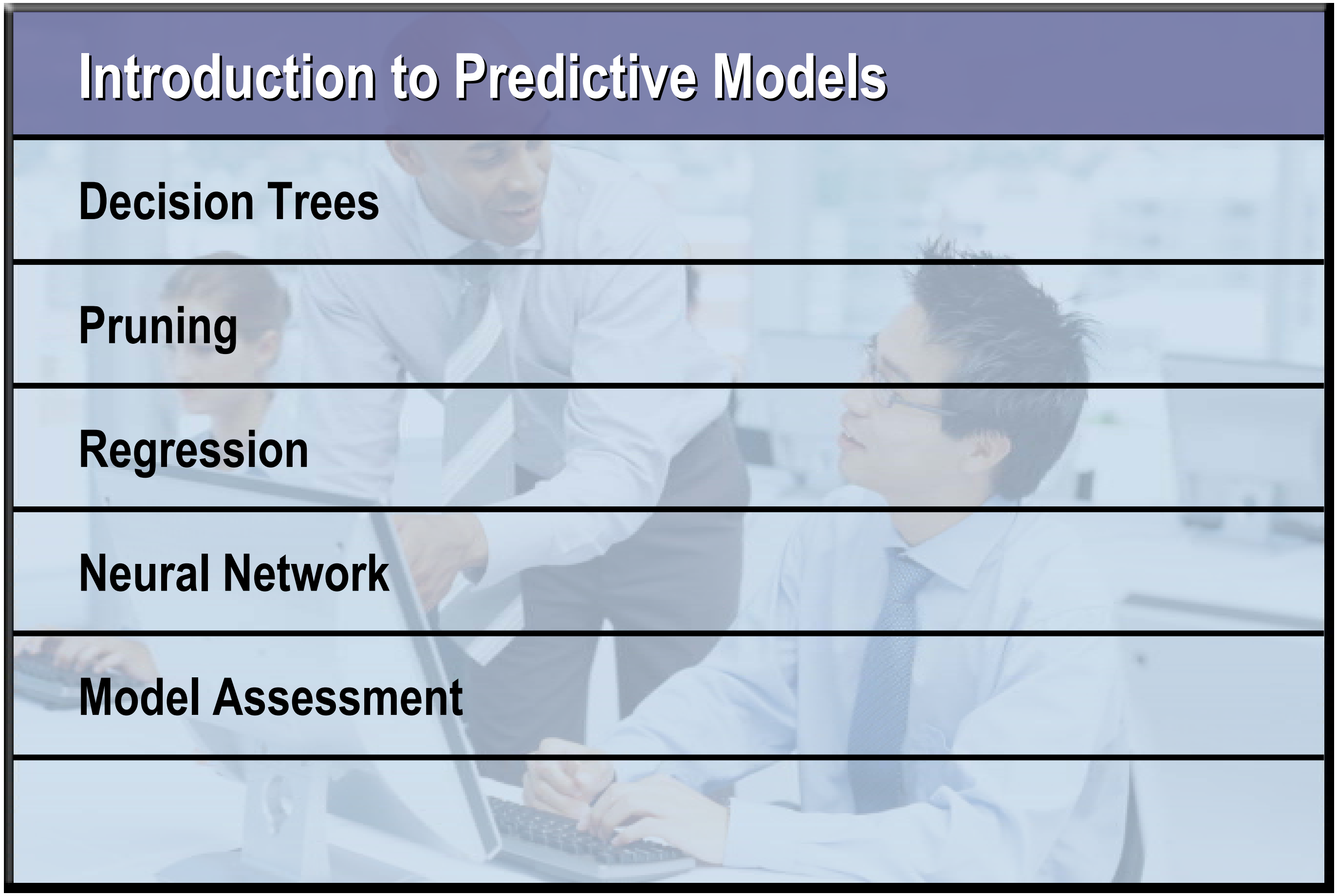
Decision Trees and other predictive models

Mathias Lanner SAS Institute

Agenda



Introduction to Predictive Models

Decision Trees

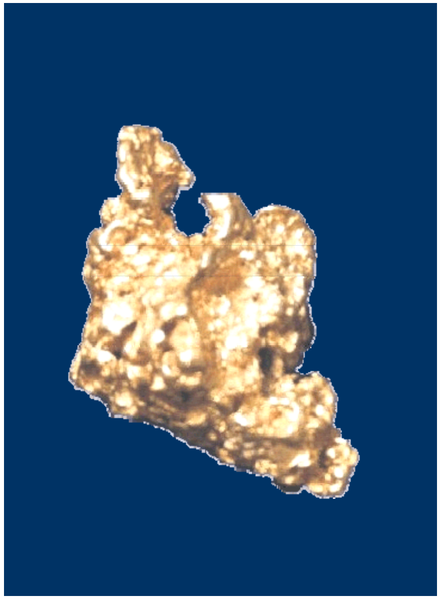
Pruning

Regression

Neural Network

Model Assessment

Predictive Modeling

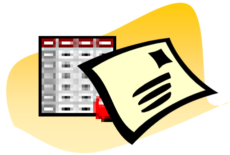


The *Essence* of Data Mining

"Most of the big payoff [in data mining] has been in predictive modeling."

– Herb Edelstein

Predictive Modeling Applications


Database marketing

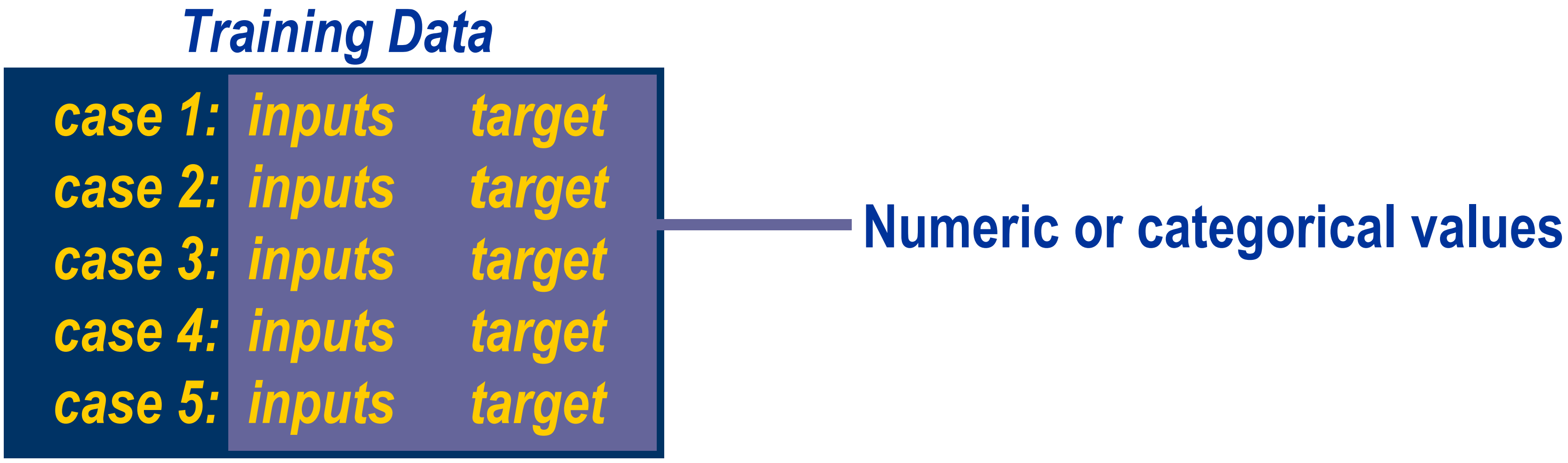
Financial risk management

Fraud detection

Process monitoring

Pattern detection

Predictive Modeling Training Data



Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Numeric or categorical values

Predictive Modeling Score Data

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Score Data

case 1:	*inputs*	*?*
case 2:	*inputs*	*?*
case 3:	*inputs*	*?*
case 4:	*inputs*	*?*
case 5:	*inputs*	*?*

Only input values known

Predictions

Training Data

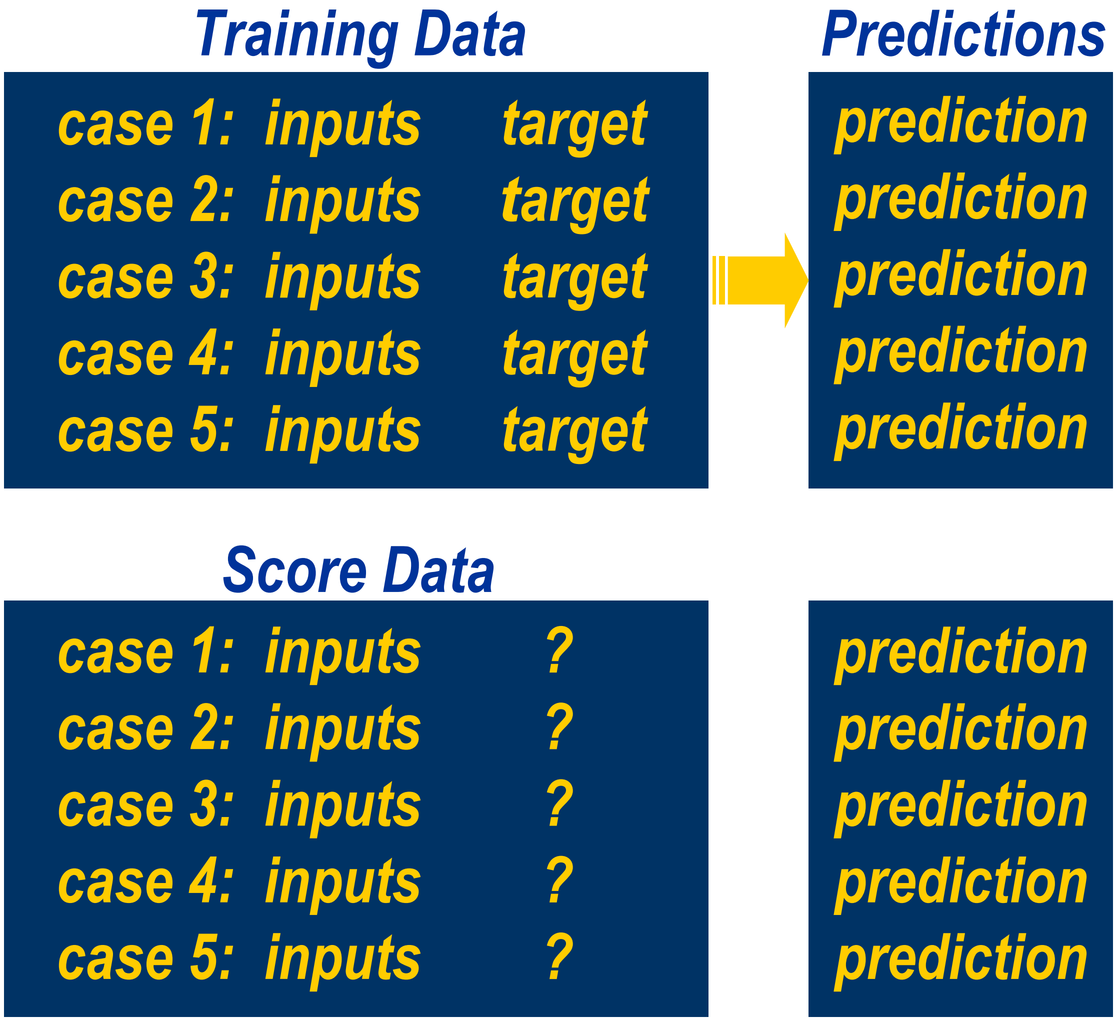
case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Predictions

Score Data

case 1:	*inputs*	*?*
case 2:	*inputs*	*?*
case 3:	*inputs*	*?*
case 4:	*inputs*	*?*
case 5:	*inputs*	*?*

Predictions



Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Predictions

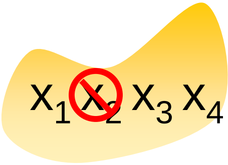
- *prediction*
- *prediction*
- *prediction*
- *prediction*
- *prediction*

Score Data

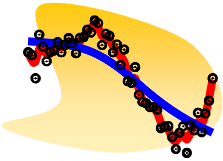
case 1:	*inputs*	*?*	*prediction*
case 2:	*inputs*	*?*	*prediction*
case 3:	*inputs*	*?*	*prediction*
case 4:	*inputs*	*?*	*prediction*
case 5:	*inputs*	*?*	*prediction*

Predictive Modeling Essentials

Predict new cases



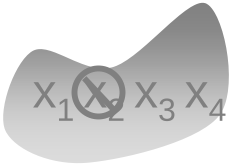
Select useful inputs



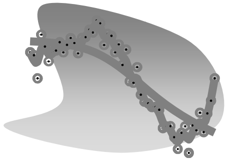
Optimize complexity

Predictive Modeling Essentials

Predict new cases

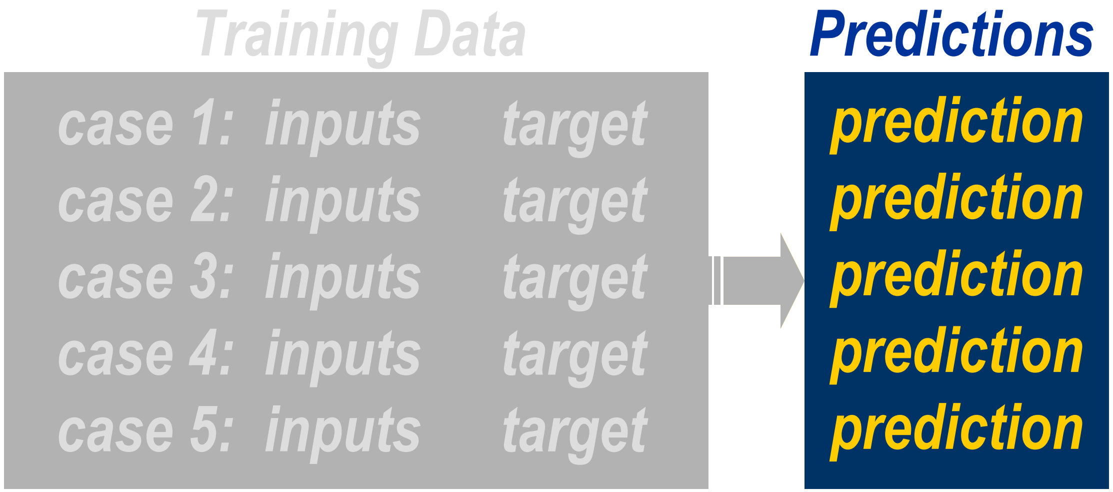


Select useful inputs



Optimize complexity

Three Prediction Types



Training Data

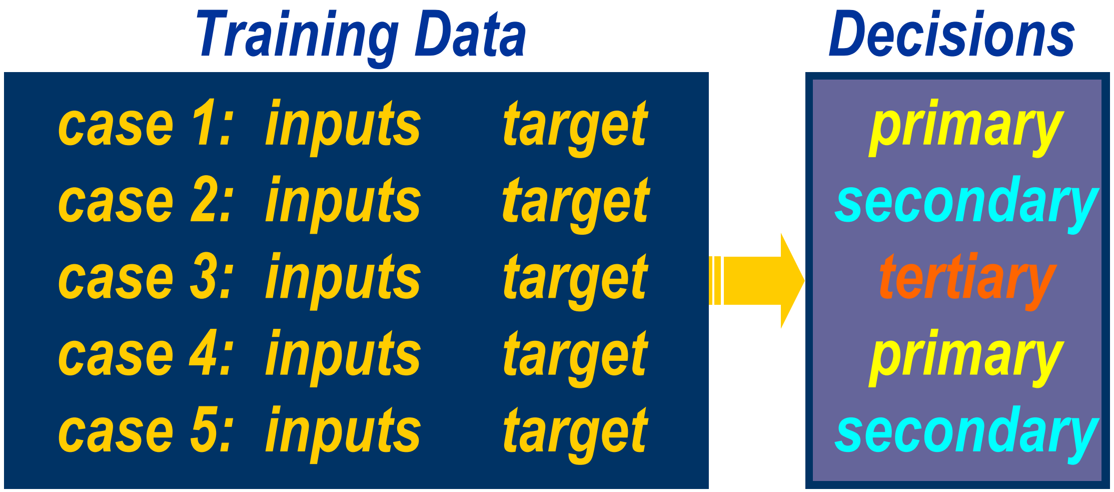
case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Predictions

prediction
prediction
prediction
prediction
prediction

- **Decisions**
- **Rankings**
- **Estimates**

Decision Predictions



Training Data

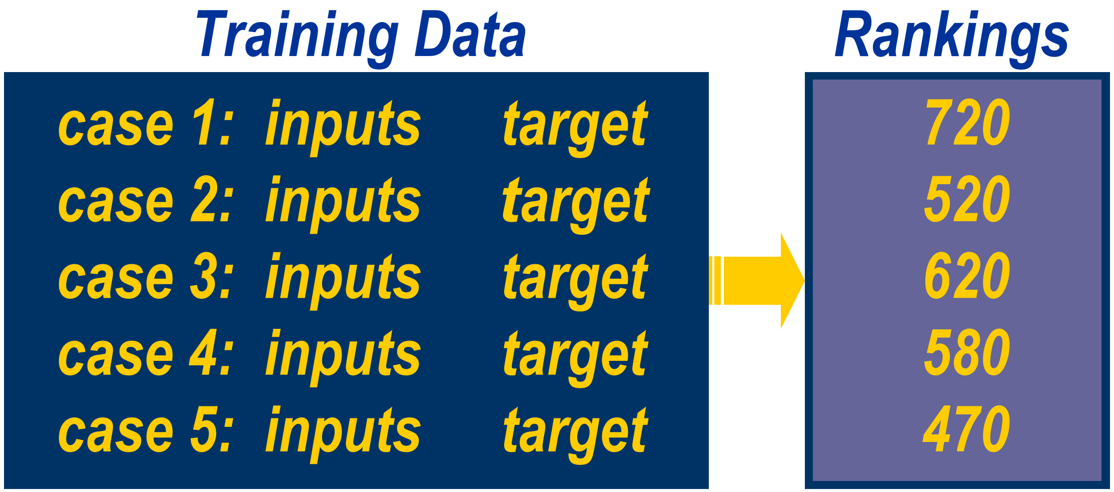
case 1: inputs target
case 2: inputs target
case 3: inputs target
case 4: inputs target
case 5: inputs target

Decisions

primary
secondary
tertiary
primary
secondary

Trained model uses input measurements to make best decision for each case.

Ranking Predictions



Training Data

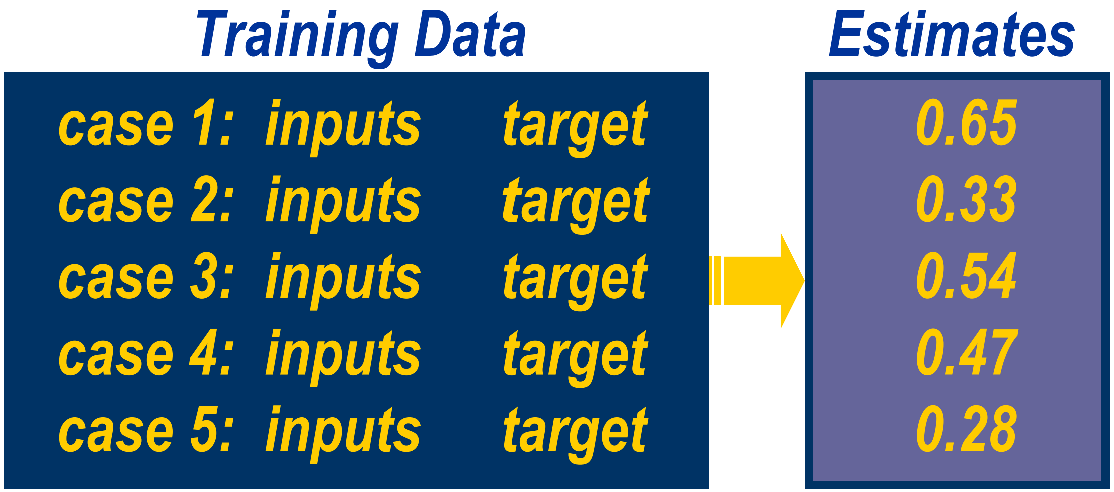
case 1: inputs target
case 2: inputs target
case 3: inputs target
case 4: inputs target
case 5: inputs target

Rankings

720
520
620
580
470

Trained model uses input measurements to optimally rank each case.

Estimate Predictions



Training Data

case 1: inputs target
case 2: inputs target
case 3: inputs target
case 4: inputs target
case 5: inputs target

Estimates

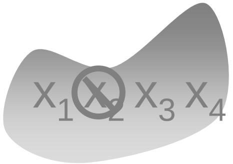
0.65
0.33
0.54
0.47
0.28

Trained model uses input measurements to optimally estimate target value.

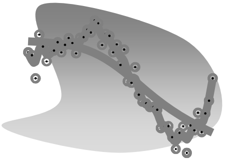
Model Essentials – Predict Review

Predict new cases

Decide, rank, estimate



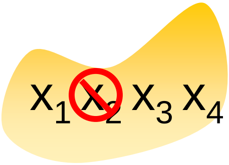
Select useful inputs



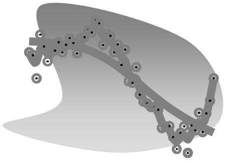
Optimize complexity

Model Essentials – Select Review

Predict new cases

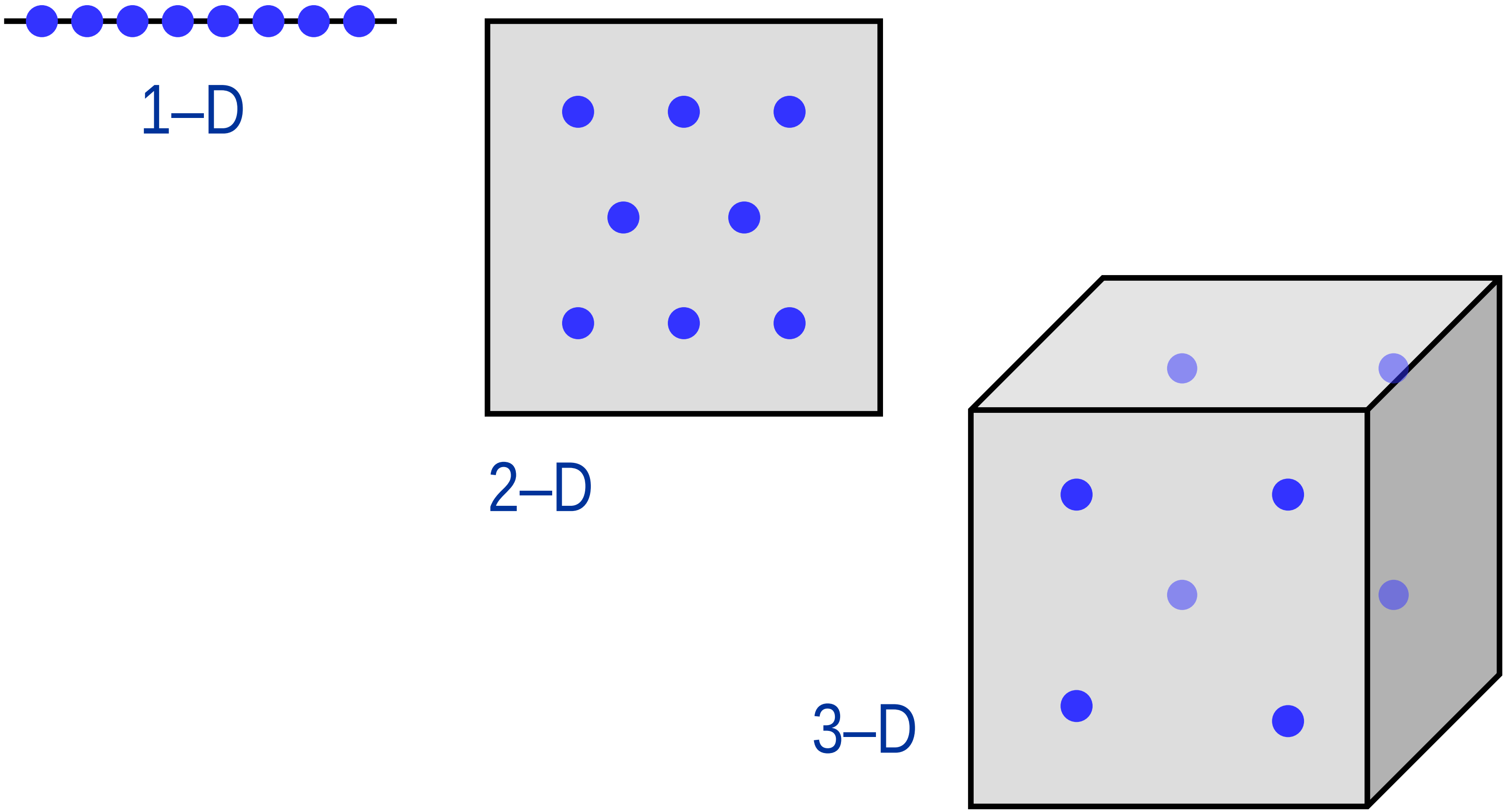


Select useful inputs



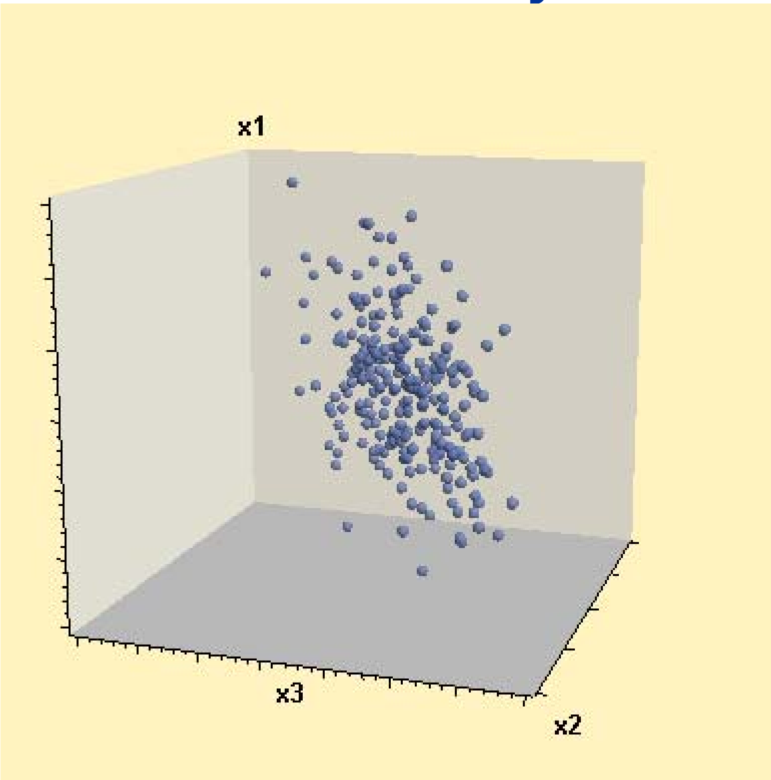
Optimize complexity

Curse of Dimensionality

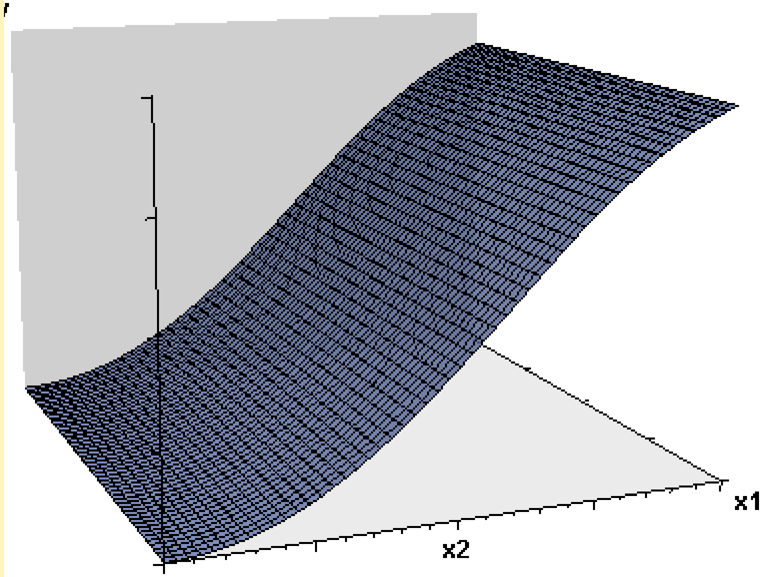


Input Selection

Redundancy



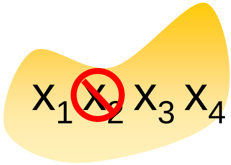
Irrelevancy



Model Essentials – Select Review

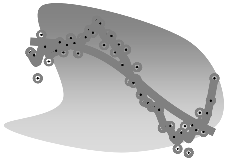
Predict new cases

Decide, rank, estimate



Select useful inputs

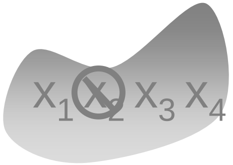
Eradicate redundancies irrelevancies



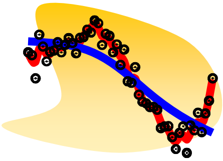
Optimize complexity

Model Essentials – Optimize

Predict new cases



Select useful inputs



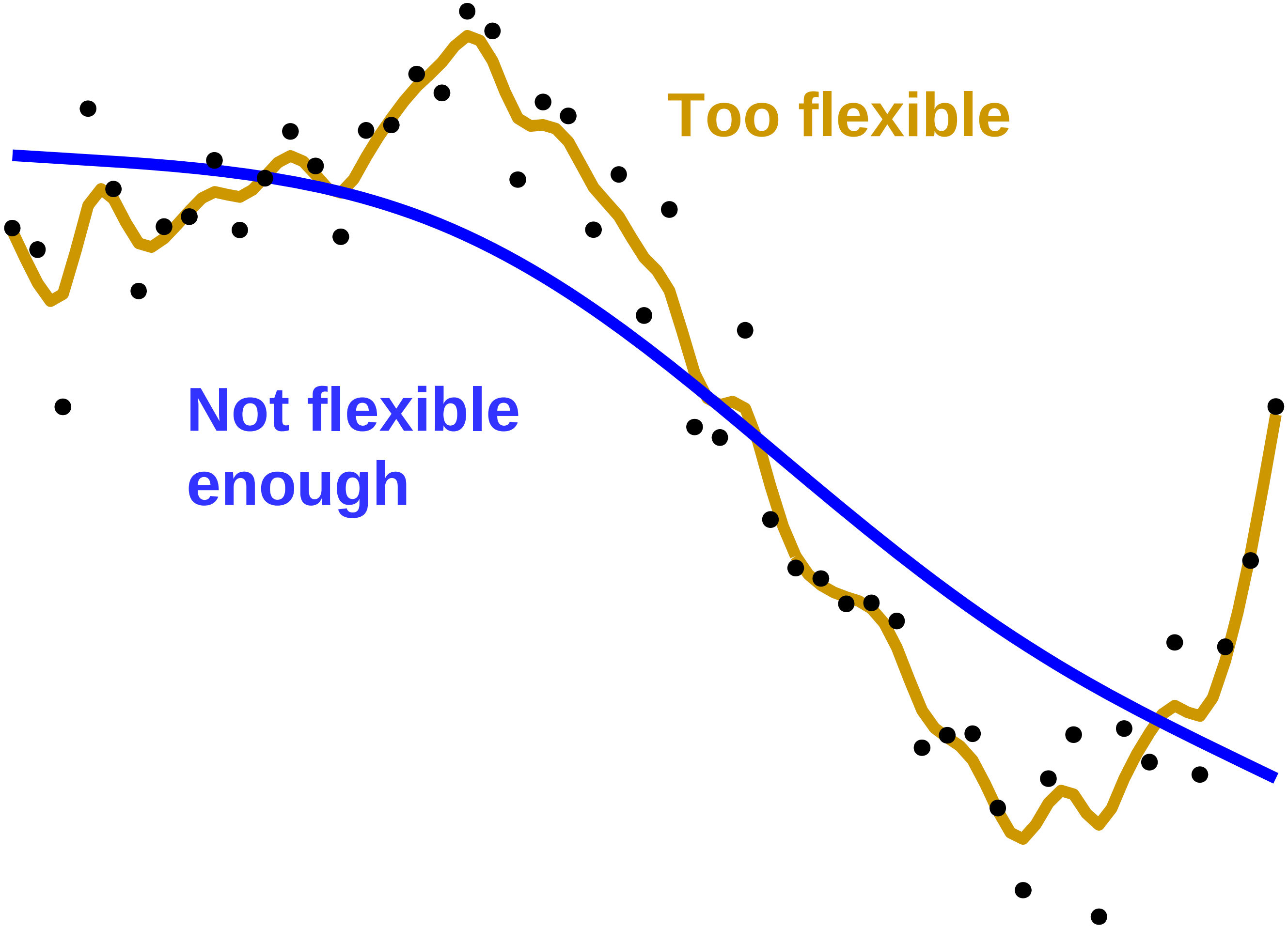
Optimize complexity

Fool's Gold

My model fits the training data perfectly...

I've struck it rich!

Model Complexity



Data Splitting

Training
Validation
Test

Training Data Role

Training Data

case 1: inputs target
case 2: inputs target
case 3: inputs target
case 4: inputs target
case 5: inputs target

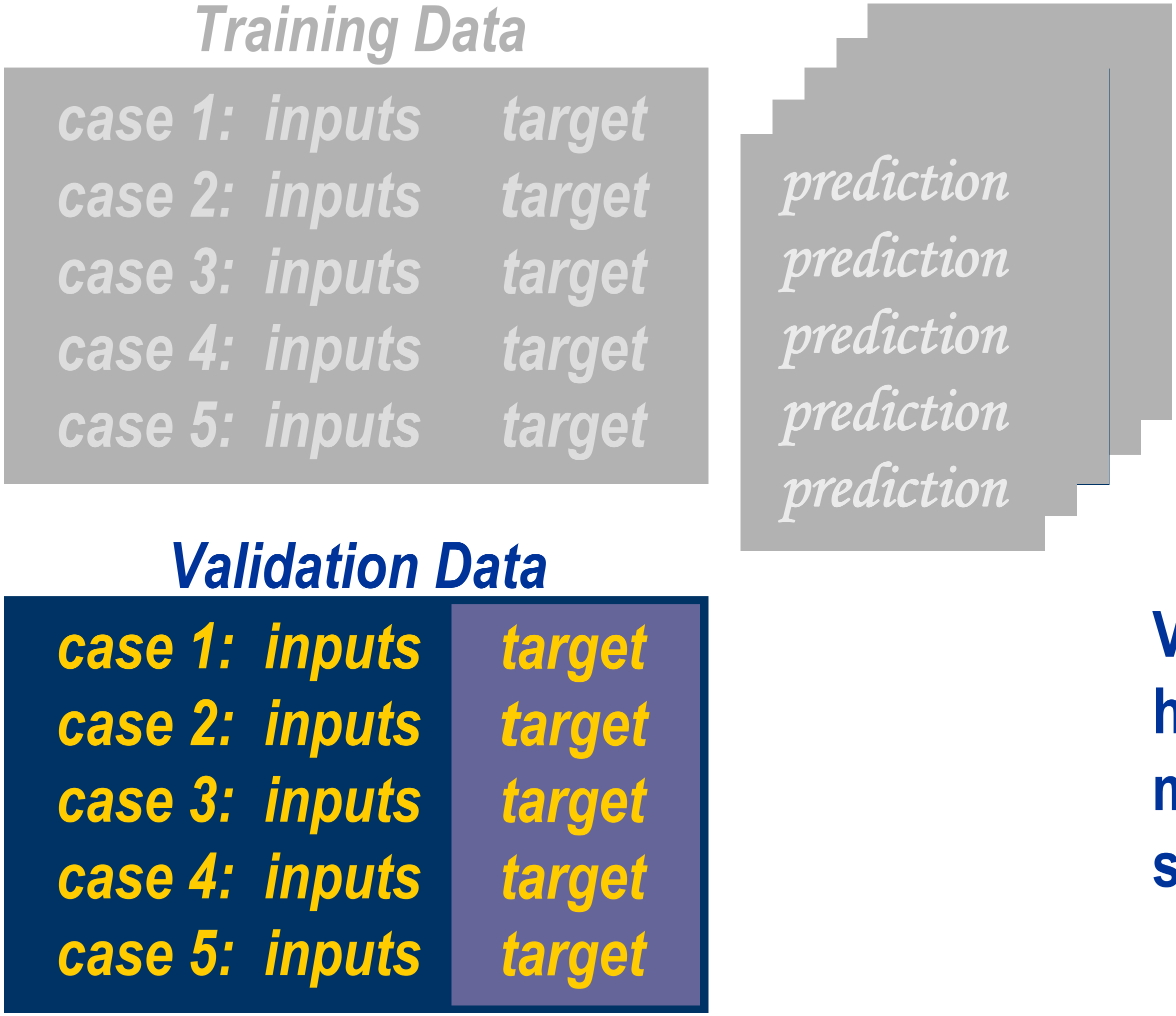
prediction
prediction
prediction
prediction
prediction

Training data gives sequence of predictive models with increasing complexity.

Validation Data

case 1: inputs target
case 2: inputs target
case 3: inputs target
case 4: inputs target
case 5: inputs target

Validation Data Role



Training Data

case 1:	inputs	target
case 2:	inputs	target
case 3:	inputs	target
case 4:	inputs	target
case 5:	inputs	target

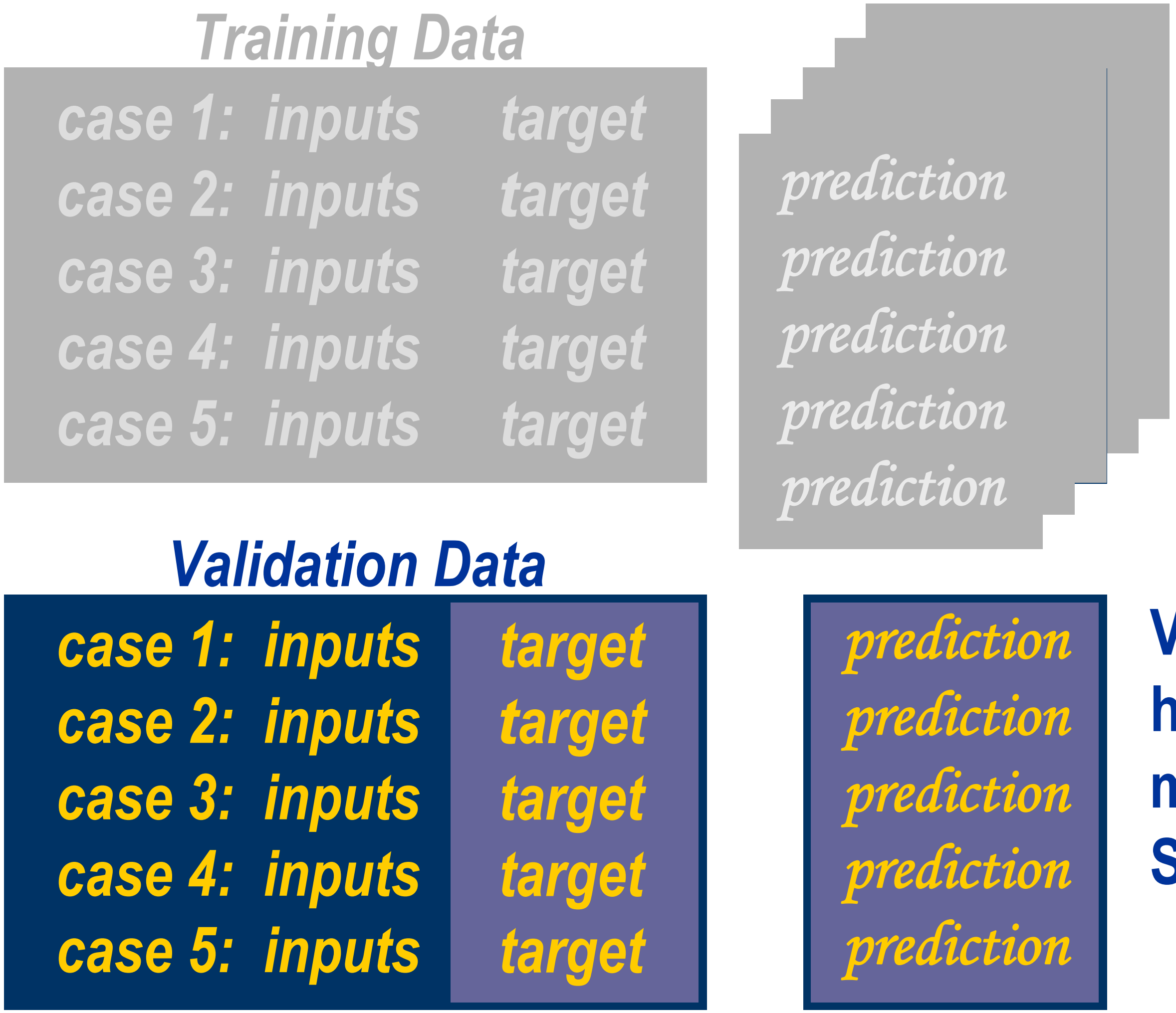
prediction
prediction
prediction
prediction
prediction

Validation Data

case 1:	inputs	target
case 2:	inputs	target
case 3:	inputs	target
case 4:	inputs	target
case 5:	inputs	target

Validation data helps select best model from sequence

Validation Data Role



Training Data

case 1:	inputs	target
case 2:	inputs	target
case 3:	inputs	target
case 4:	inputs	target
case 5:	inputs	target

prediction
prediction
prediction
prediction
prediction

Validation Data

case 1:	inputs	target
case 2:	inputs	target
case 3:	inputs	target
case 4:	inputs	target
case 5:	inputs	target

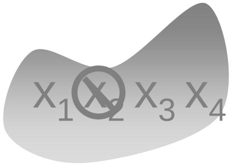
prediction
prediction
prediction
prediction
prediction

Validation data helps select best model from Sequence.

Model Essentials – Optimize

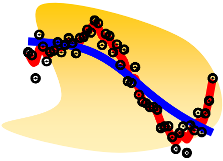
Predict new cases

Decide, rank, estimate



Select useful inputs

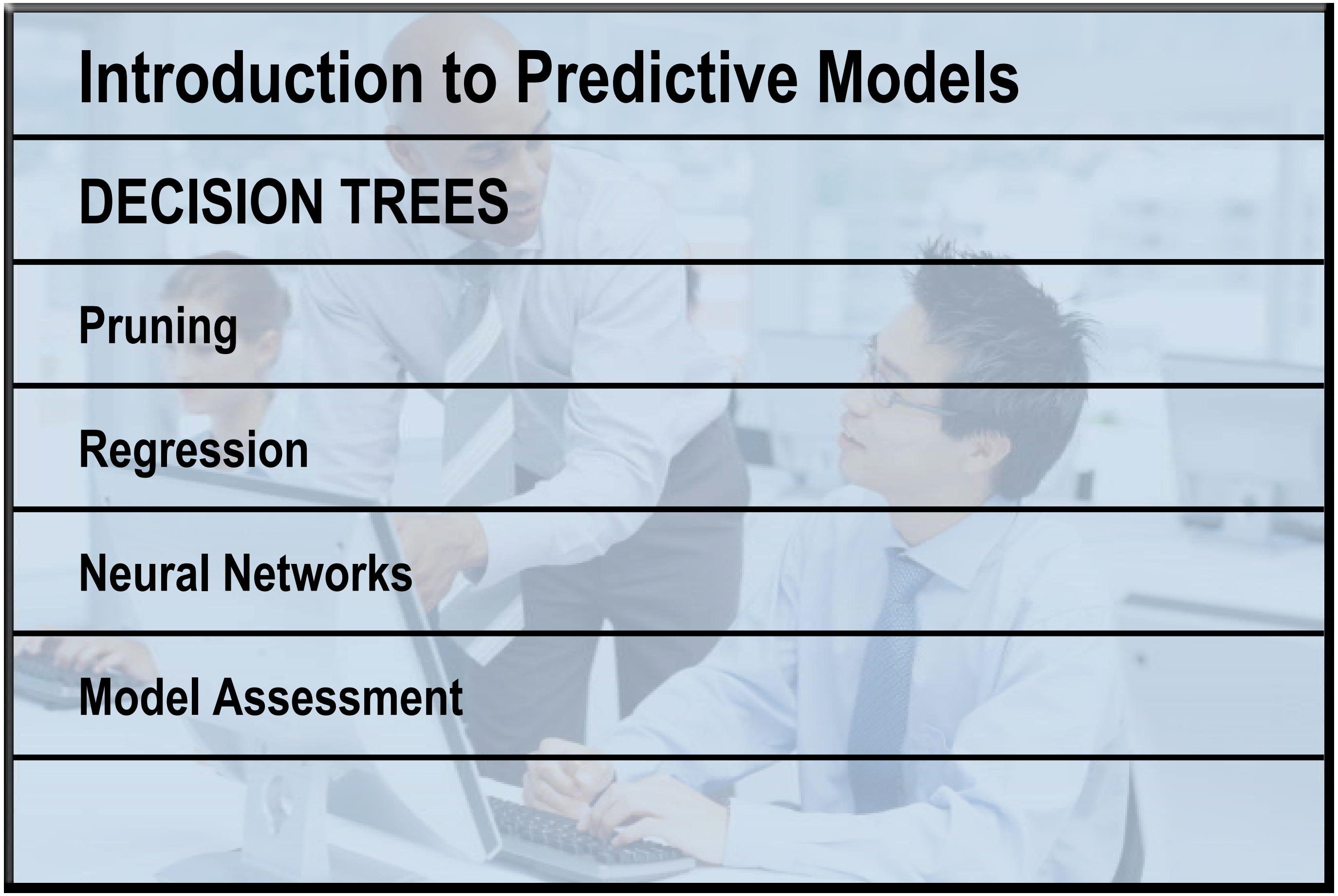
Eradicate redundancies irrelevancies



Optimize complexity

Tune models with validation data

Agenda



Introduction to Predictive Models

DECISION TREES

Pruning

Regression

Neural Networks

Model Assessment

Predictive Modeling Tools

Primary

- Decision Tree
- Regression
- Neural Network

Specialty

- Dmine Regression
- MBR
- AutoNeural
- Rule Induction
- DMNeural

Multiple Model

- Ensemble
- Two Stage

Predictive Modeling Tools

Primary

Decision Tree

Regression

Neural Network

Specialty

Dmine Regression

MBR

AutoNeural

Rule Induction

DMNeural

Multiple Model

Ensemble

Two Stage

Predictive Modeling Tools

Primary

Decision Tree

Regression

Neural Network

Specialty

Dmine Regression

MBR

AutoNeural

Rule Induction

DMNeural

Multiple Model

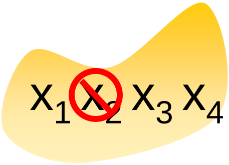
Ensemble

Two Stage

Model Essentials – Decision Trees

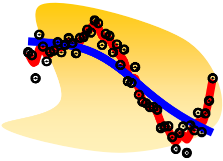
Predict new cases

Prediction rules



Select useful inputs

Split search



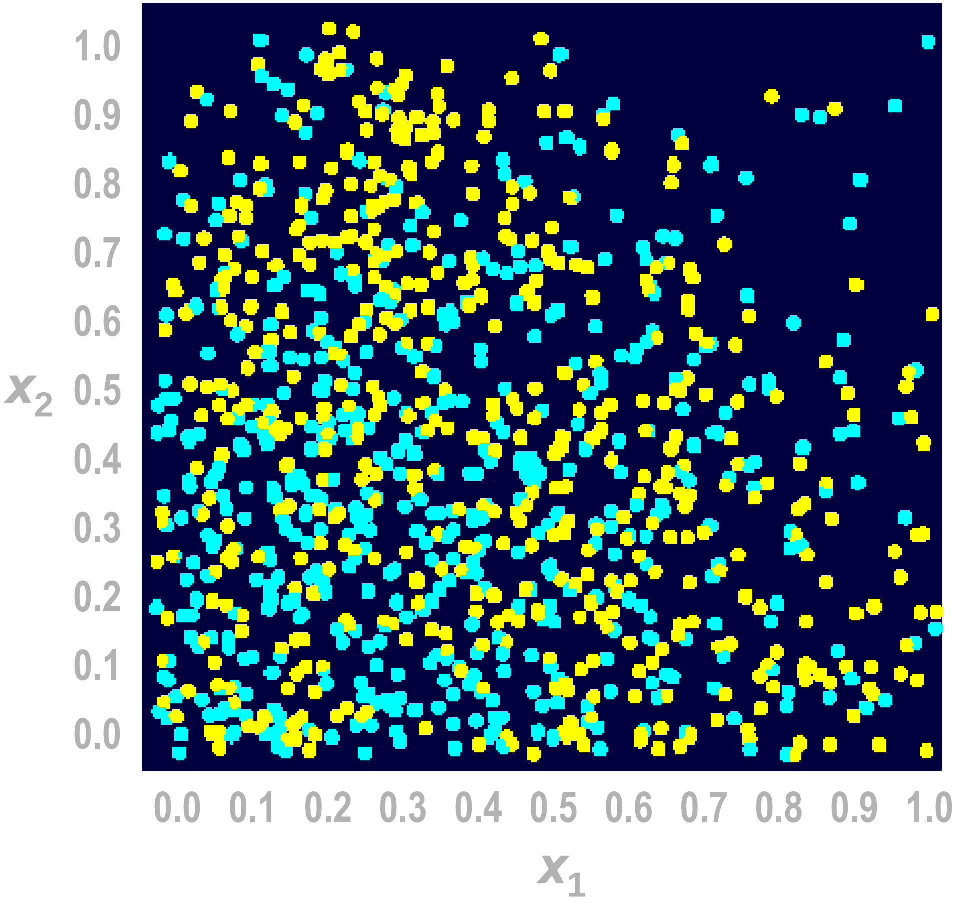
Optimize complexity

Pruning

Simple Prediction Illustration

Analysis goal:

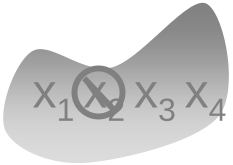
Predict the color of a dot based on its location in a scatter plot.



Model Essentials – Decision Trees

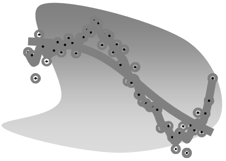
Predict new cases

Prediction rules



Select useful inputs

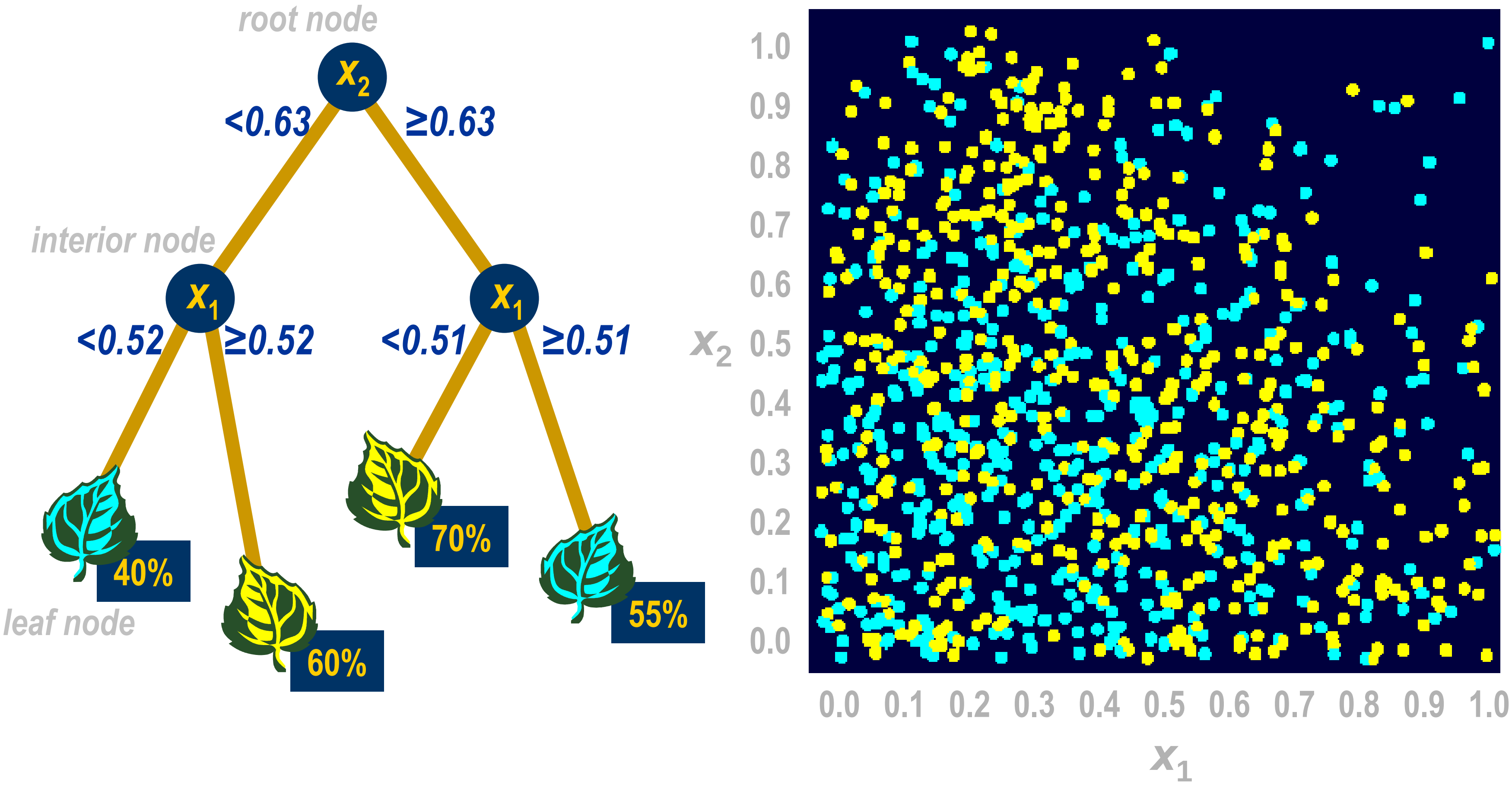
Split search



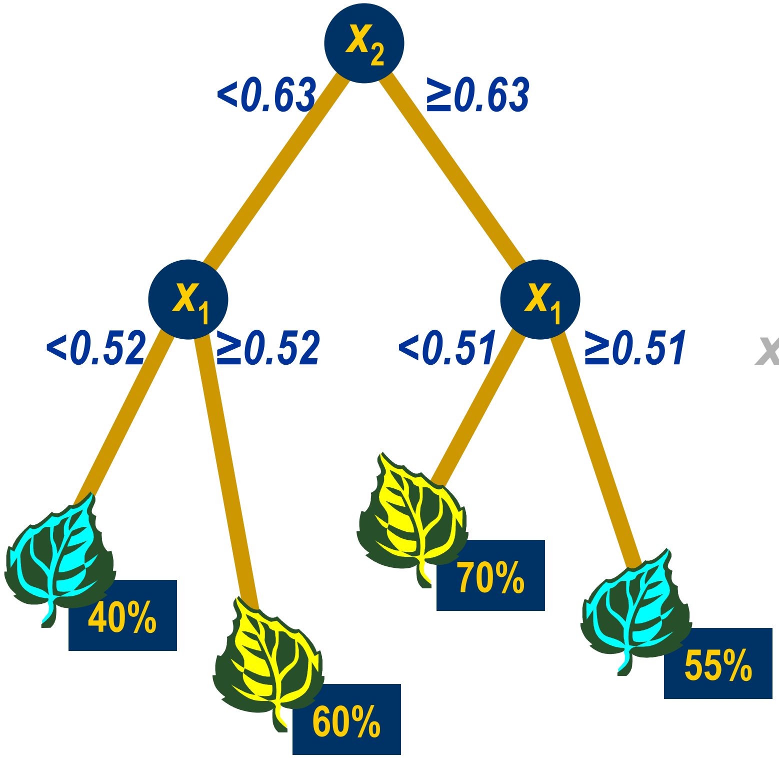
Optimize complexity

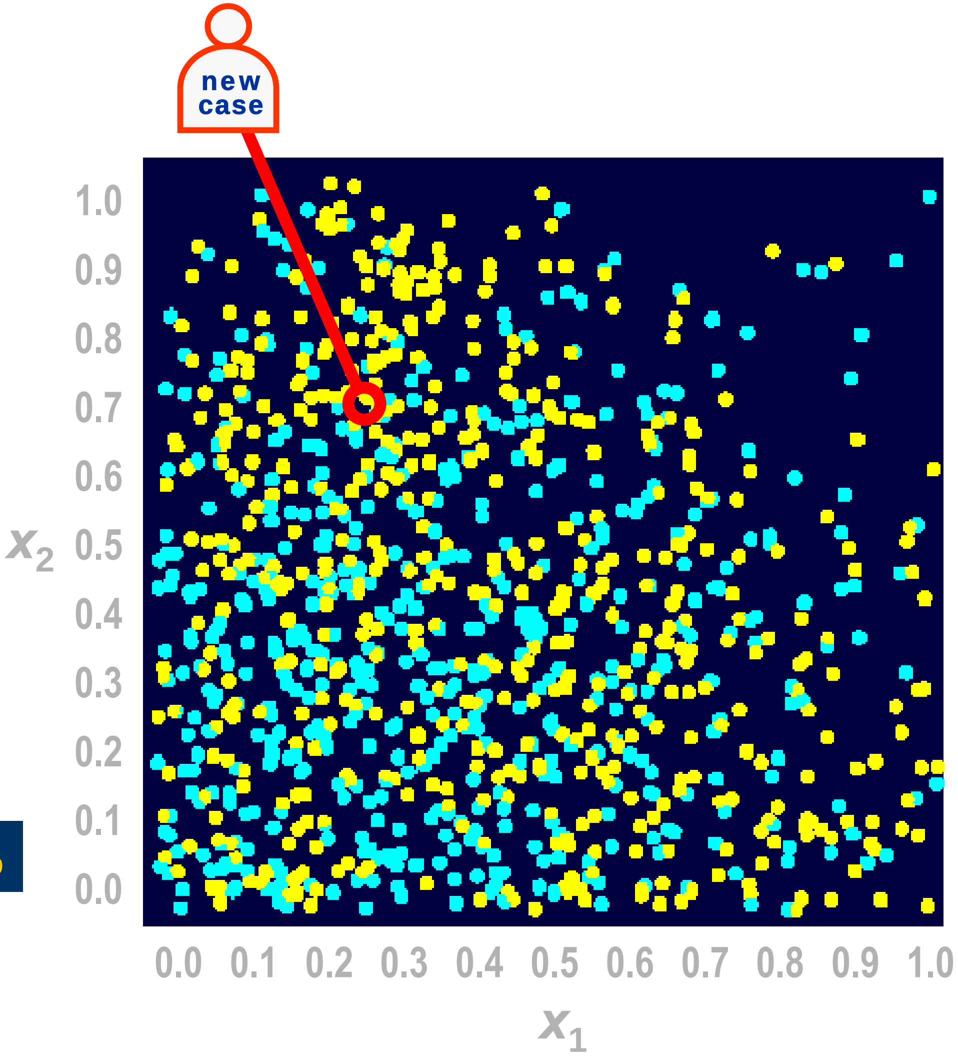
Pruning

Decision Tree Prediction Rules

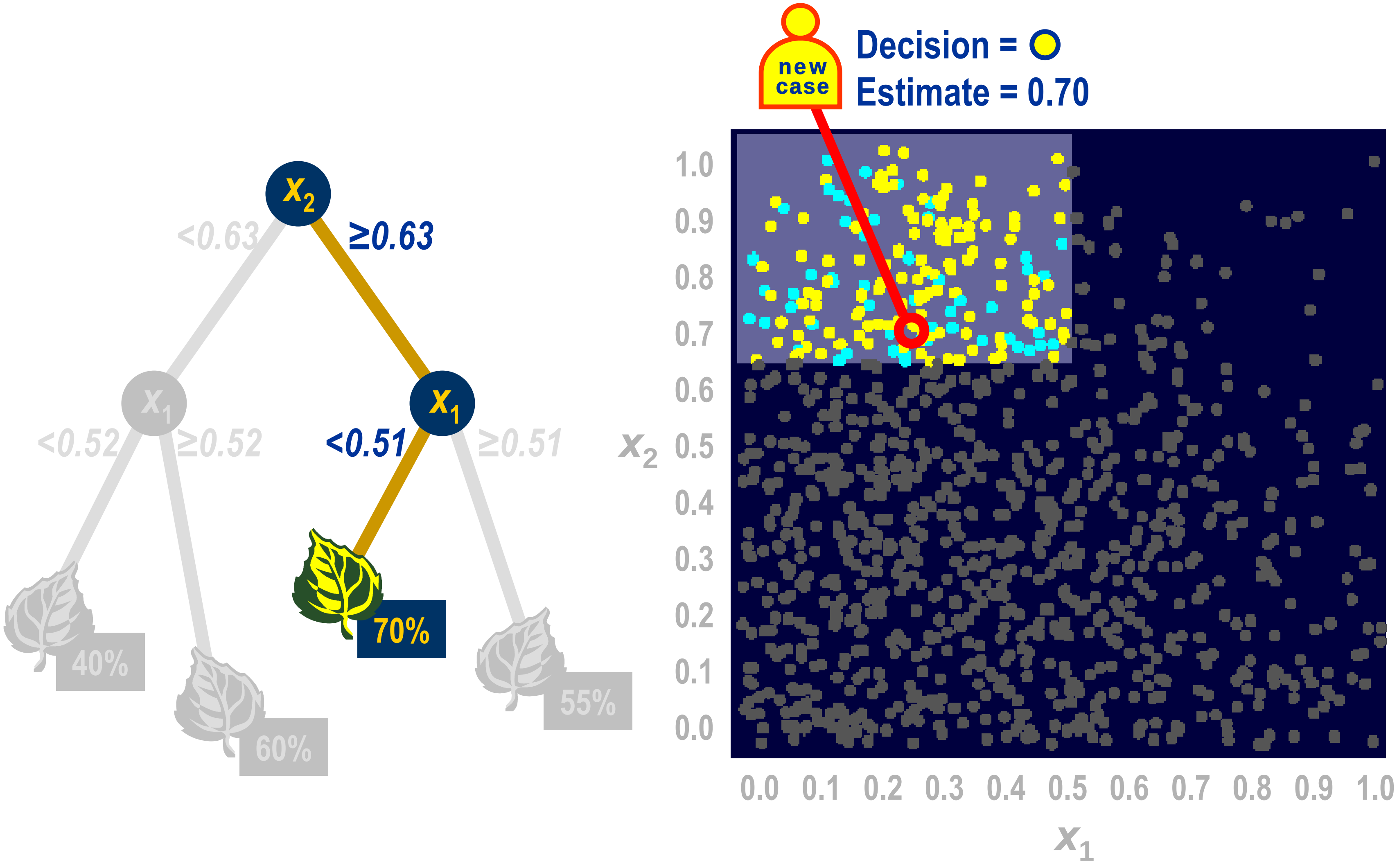


Decision Tree Prediction Rules





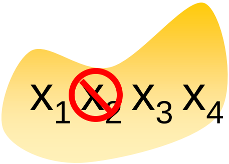
Decision Tree Prediction Rules

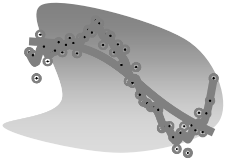


Decision = ○

Estimate = 0.70

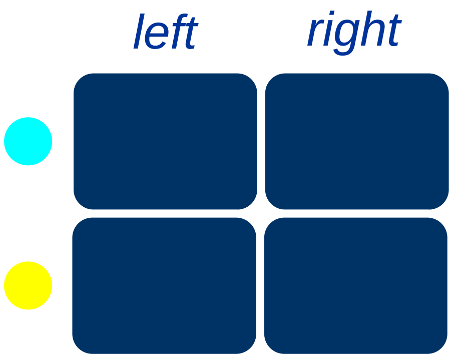
Model Essentials – Decision Trees

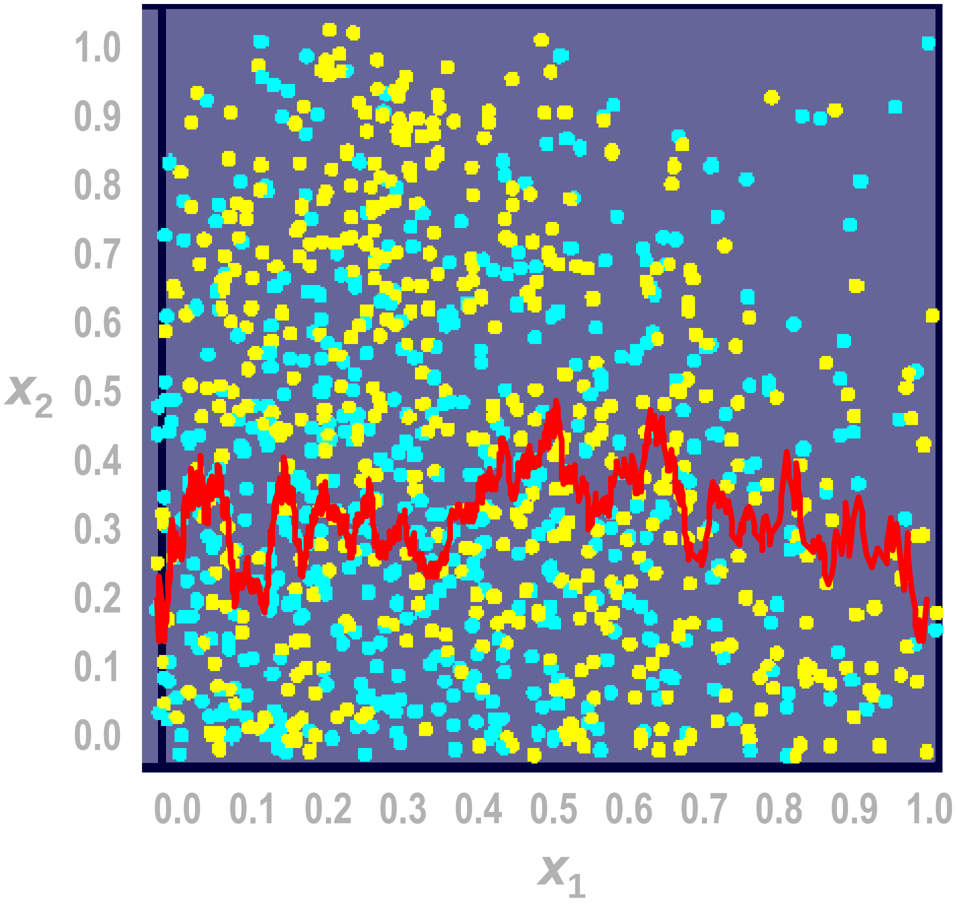




Predict new cases	Prediction rules
Select useful inputs	**Split search**
Optimize complexity	Pruning

Decision Tree Split Search



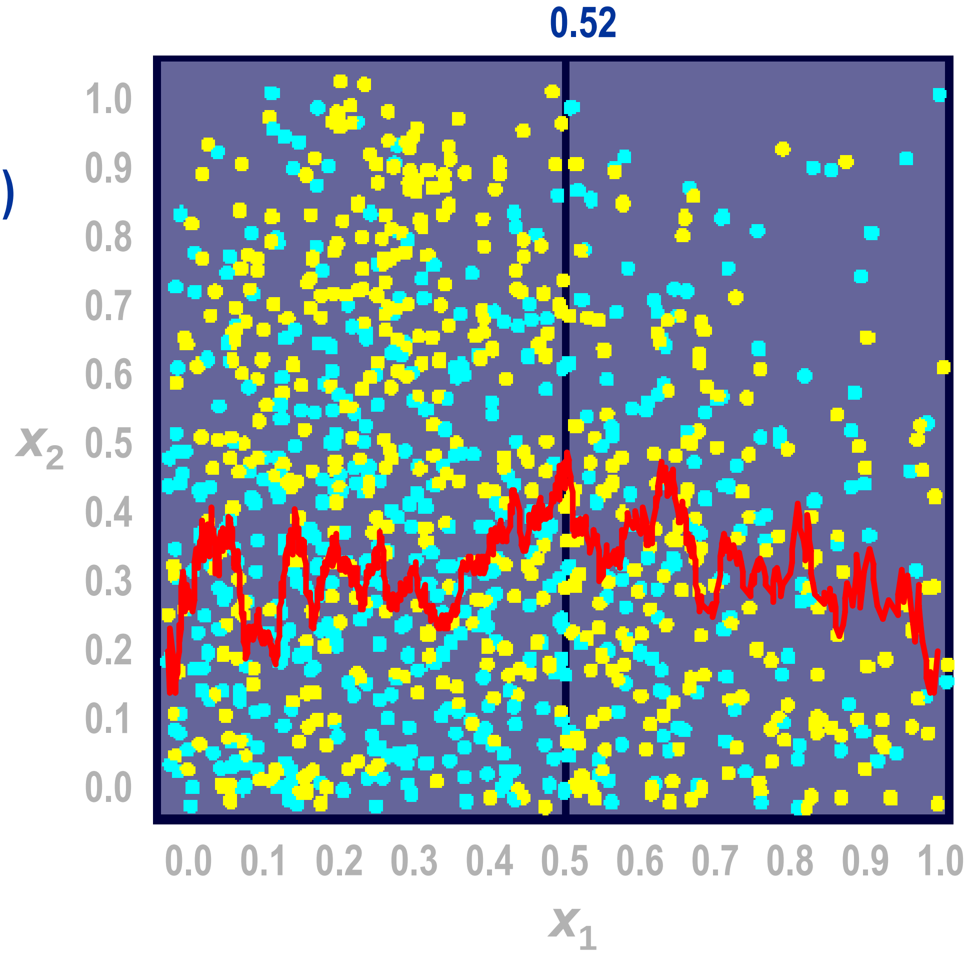


Calculate the *logworth* of every partition on input x_1.

Decision Tree Split Search

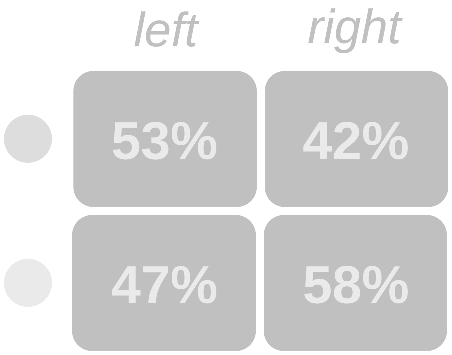
	left	*right*
🔵	53%	42%
🟡	47%	58%

max
logworth(x_1)
0.95



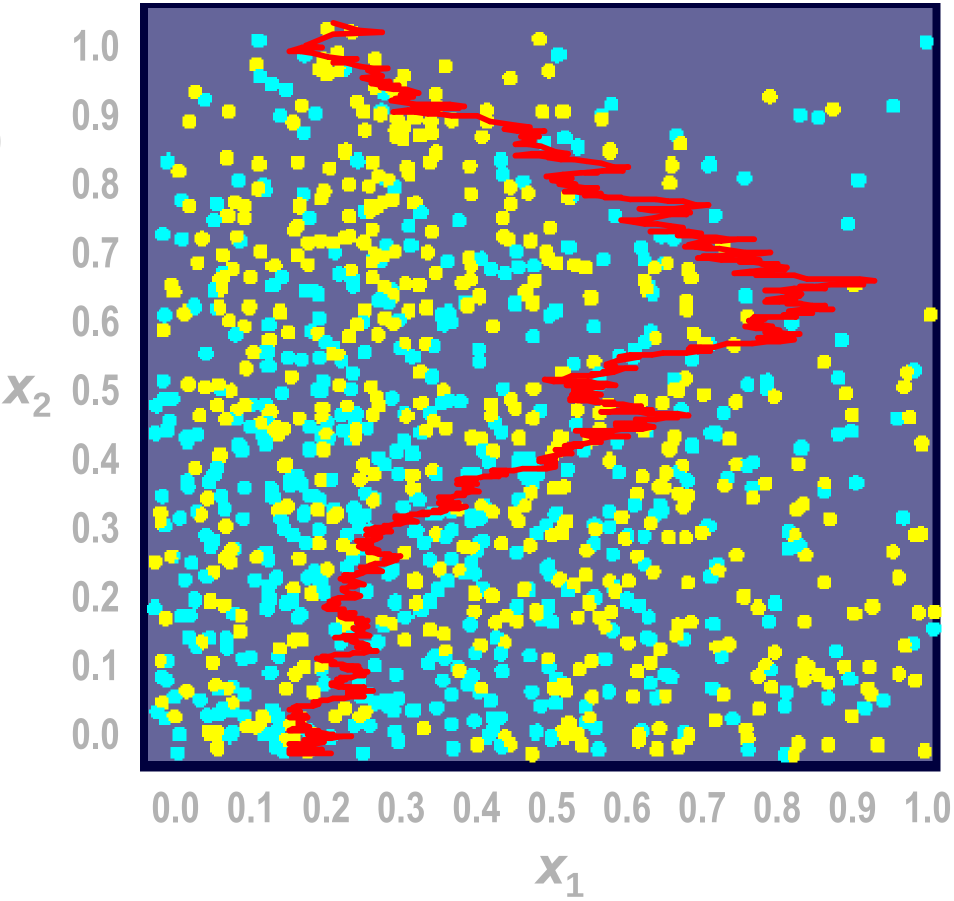
Select the partition with maximum *logworth*.

Decision Tree Split Search



max logworth(x_1) 0.95

Repeat for input x_2.



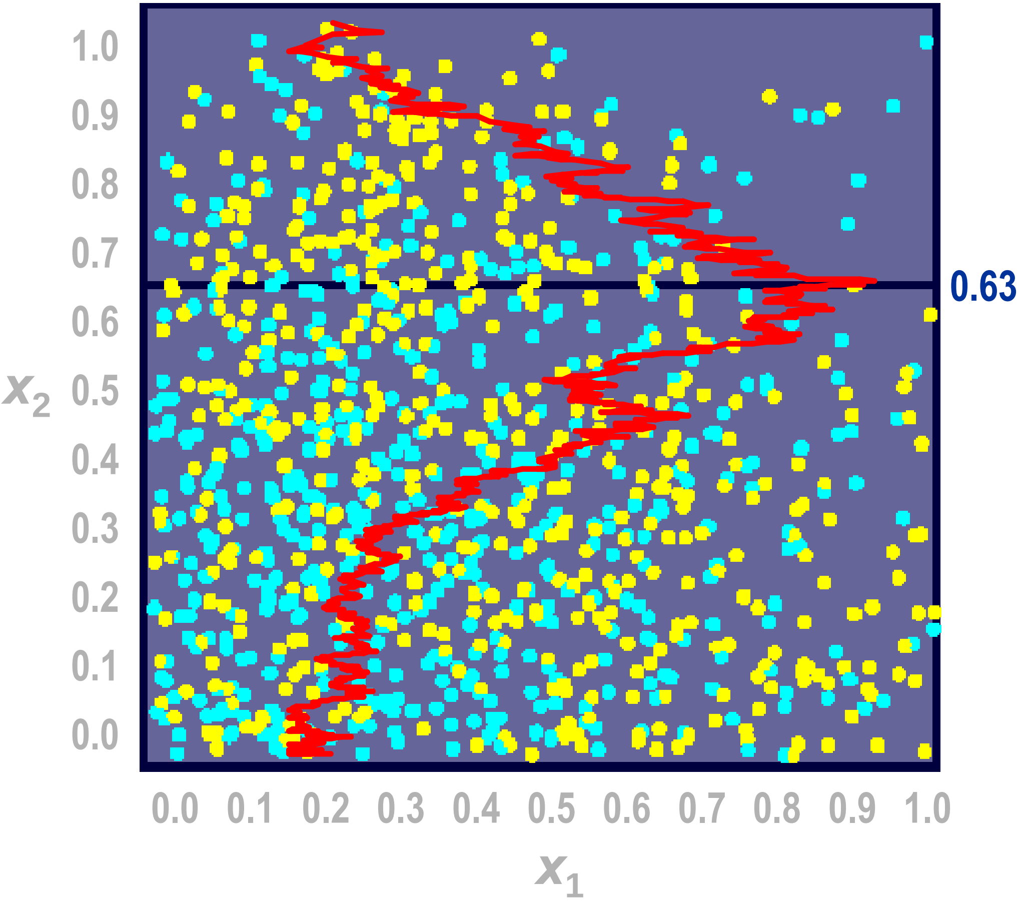
Decision Tree Split Search

	left	*right*
●	53%	42%
●	47%	58%

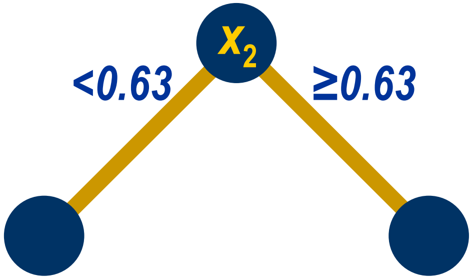
max
logworth(x_1)
0.95

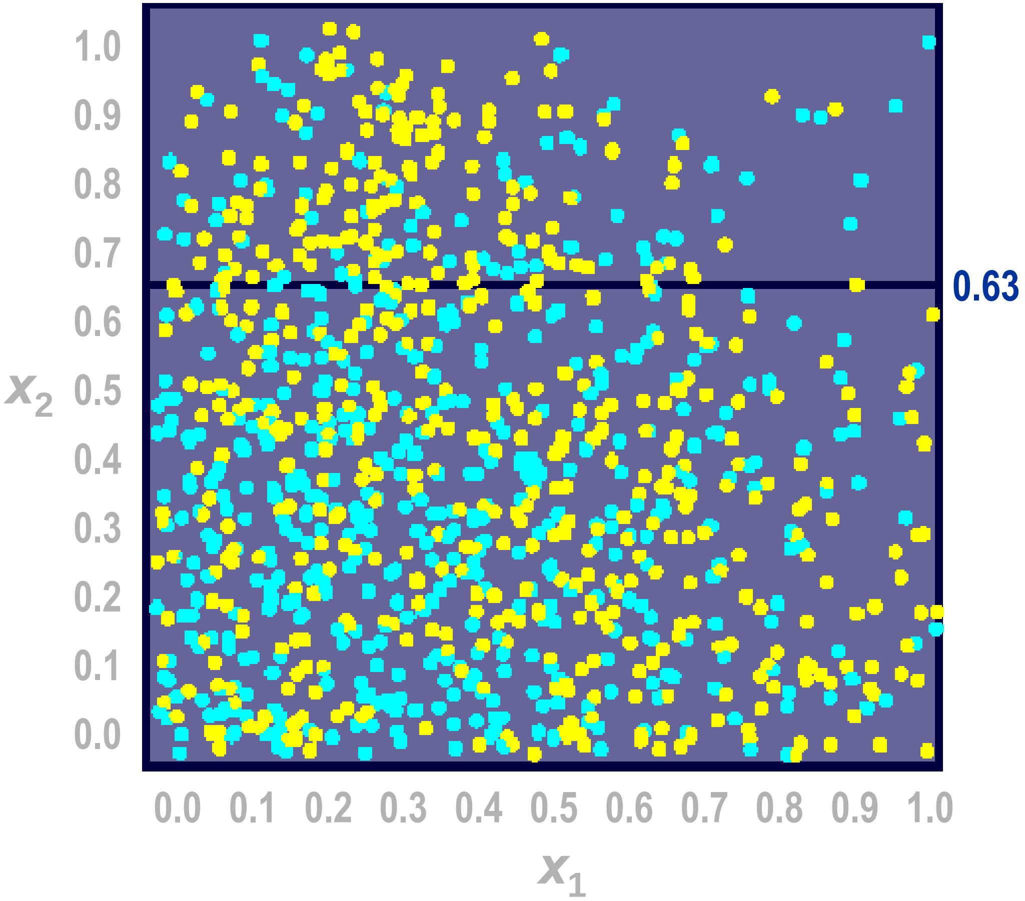
	bottom	*top*
●	54%	35%
●	46%	65%

max
logworth(x_2)
4.92



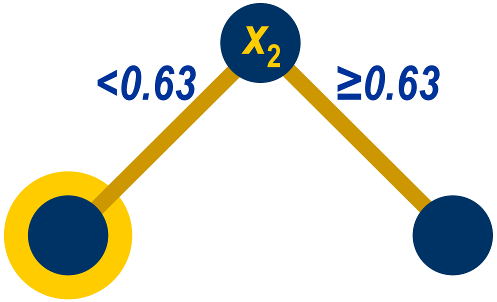
Decision Tree Split Search

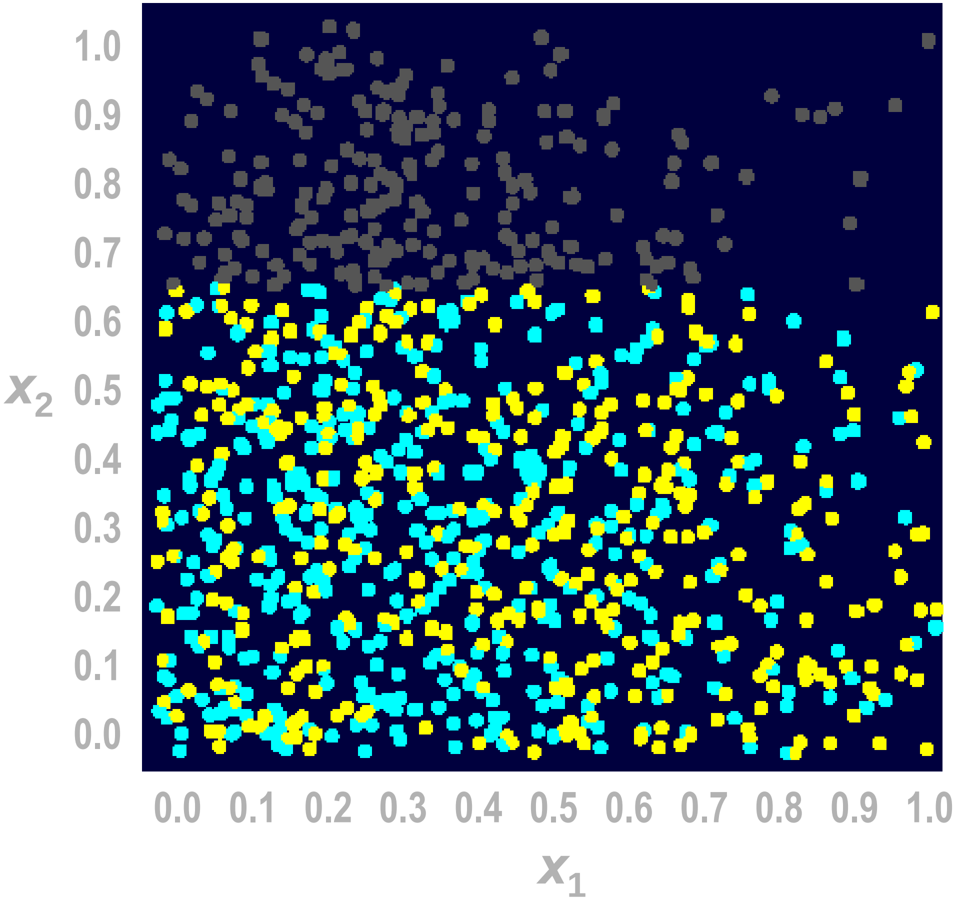




Create partition rule from best partition across all inputs.

Decision Tree Split Search



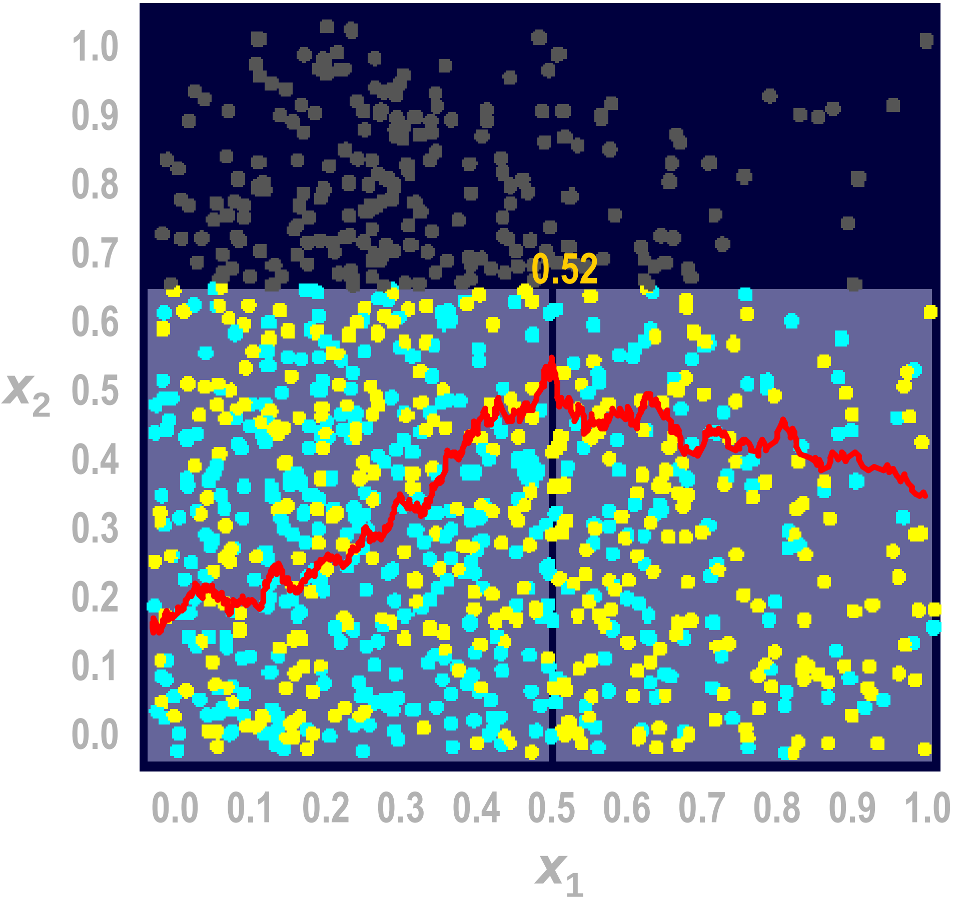


Repeat process in each subset.

Decision Tree Split Search

	left	*right*
●	61%	40%
●	39%	60%

logworth(x_1)
5.72



Select the partition with maximum *logworth* on input x_1.

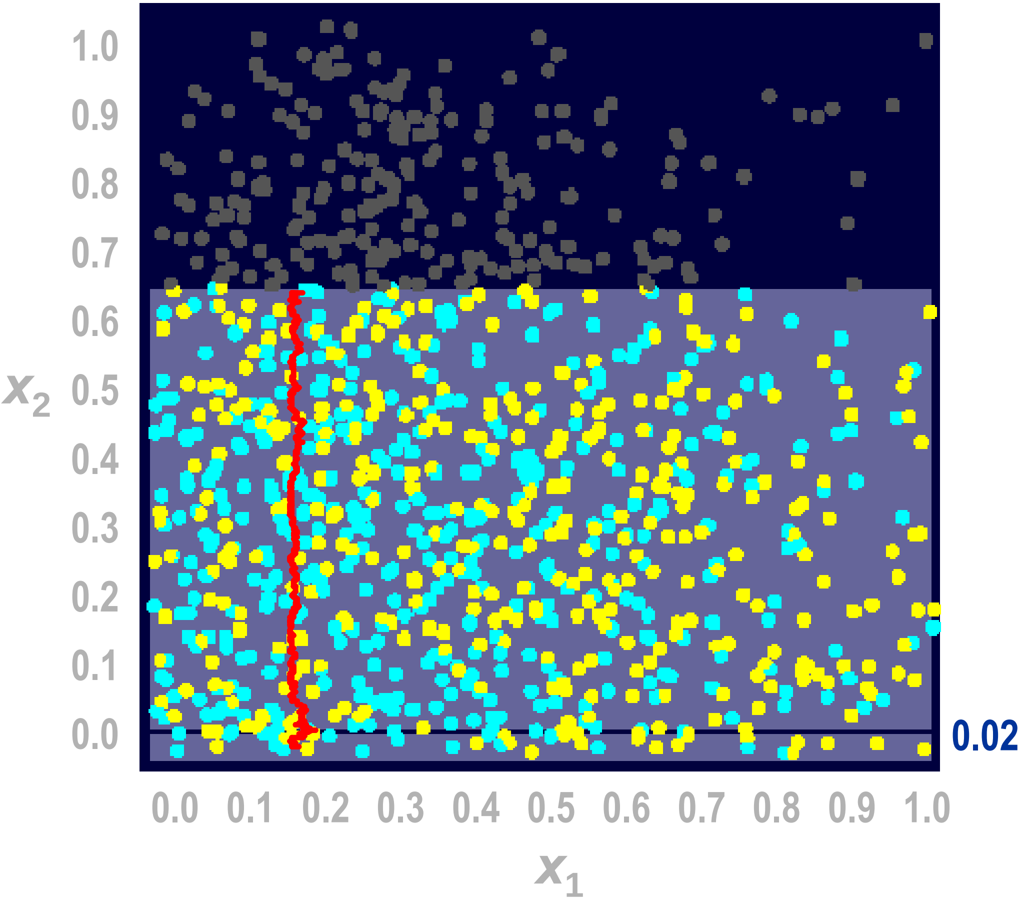
Decision Tree Split Search

	left	*right*
●	61%	40%
●	39%	60%

logworth(x_1)
5.72

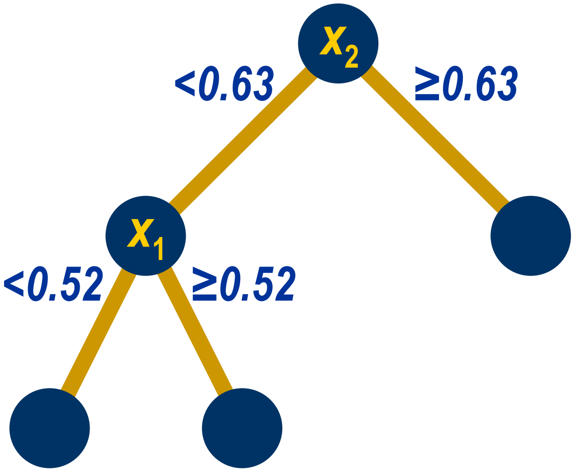
	bottom	*top*
●	38%	55%
●	62%	45%

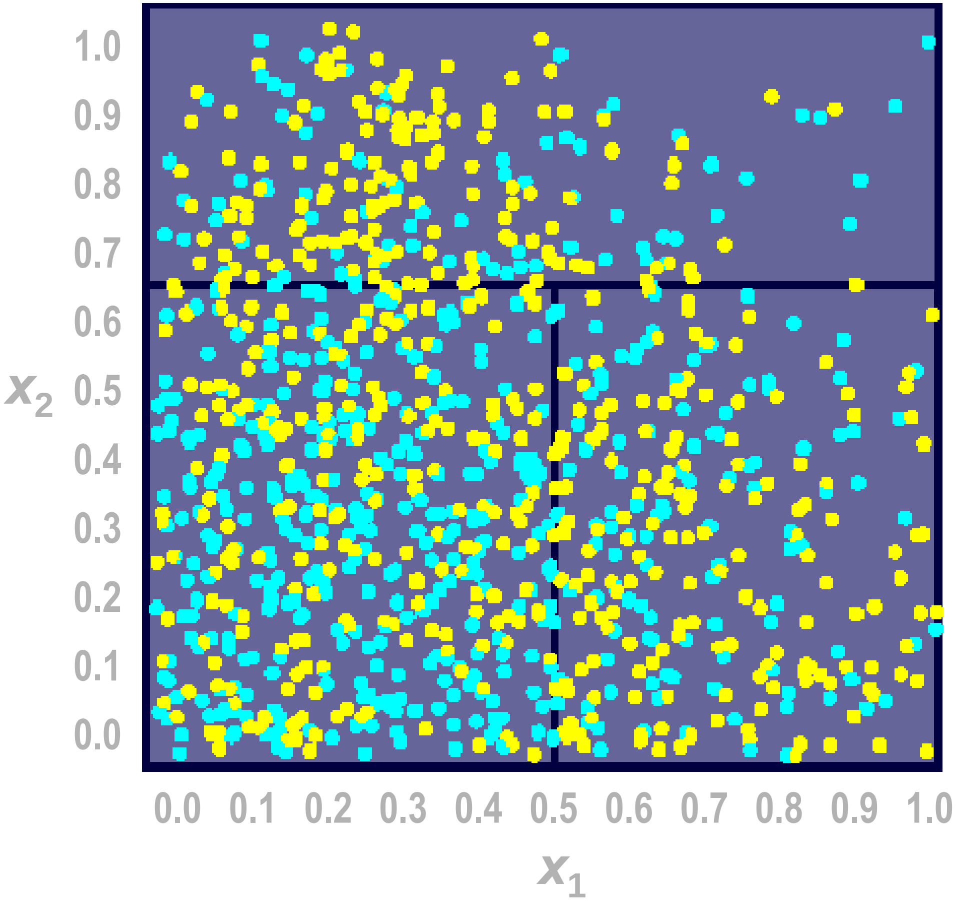
logworth(x_2)
-2.01



Repeat for input x_2.

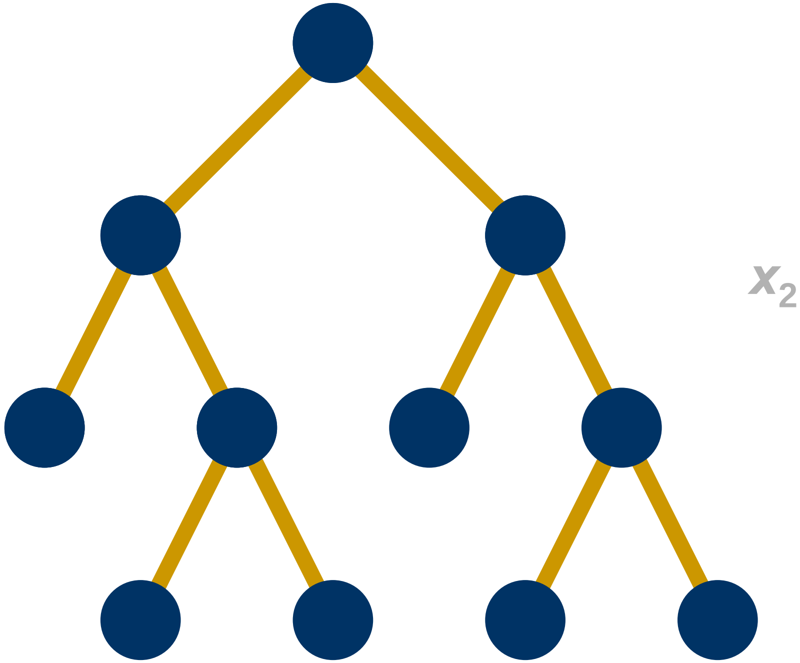
Decision Tree Split Search

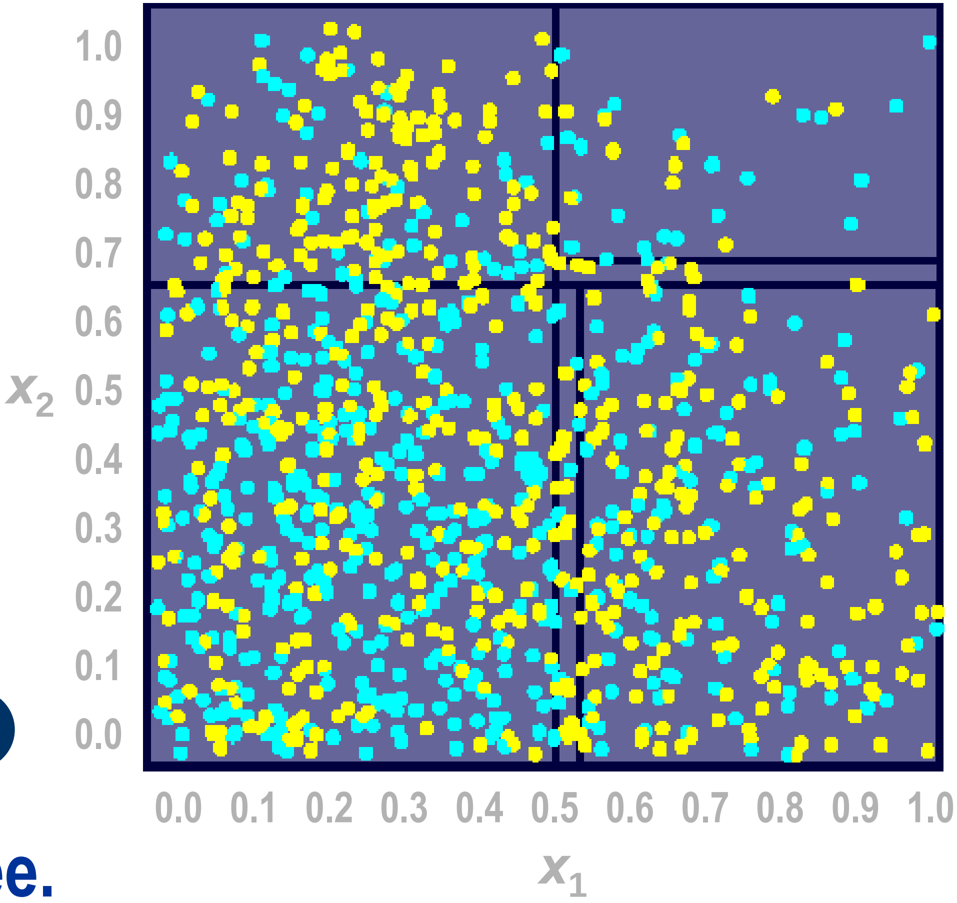




Create second partition rule.

Decision Tree Split Search





Repeat to form maximal tree.

Demo Decision Tree