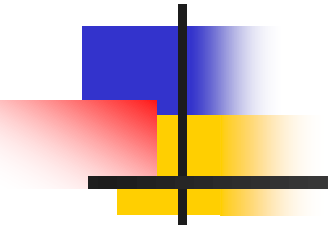


Decision Tree Basics

Dr. Goutam Chakraborty

Model Essentials – Decision Trees

- **Predict new cases.**
- **Select useful inputs.**
- **Optimize complexity.**

Prediction rules

Split search

Pruning

...

Model Essentials – Decision Trees

- **Predict new cases.**
- **Select useful inputs.**
- **Optimize complexity.**

Prediction rules

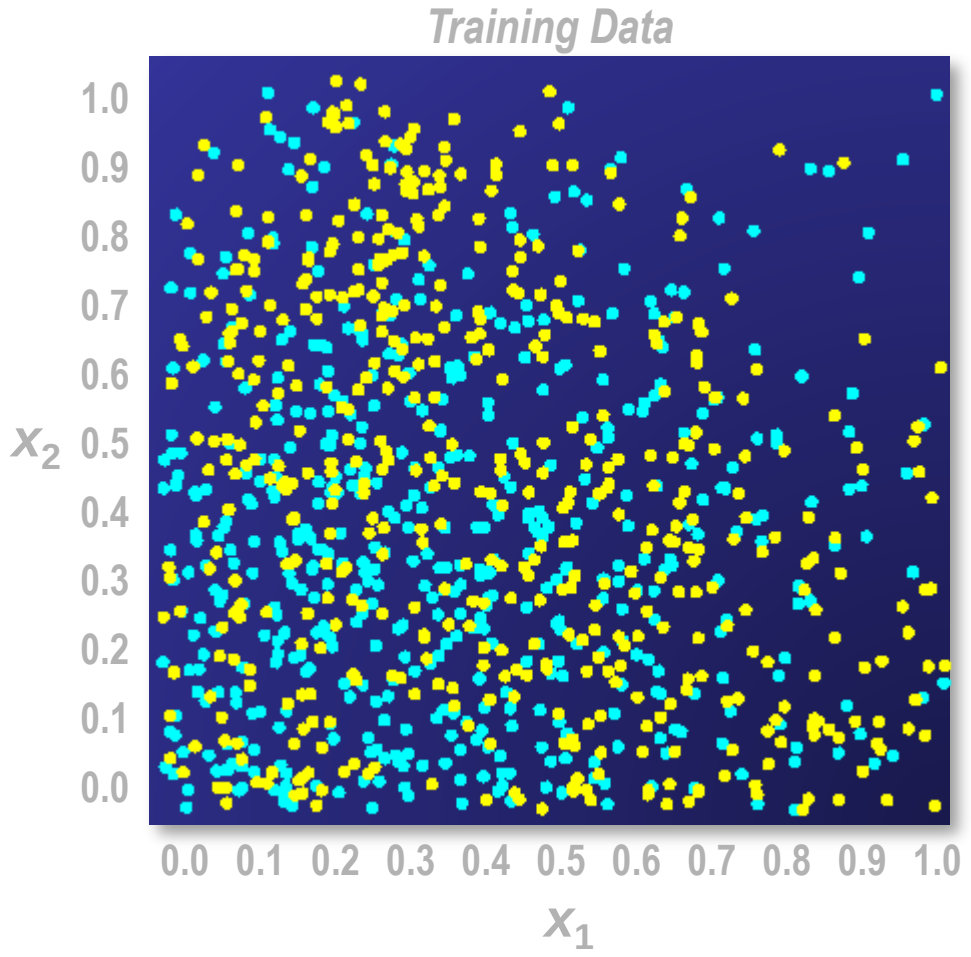
Split search

Pruning

...

Simple Prediction Illustration

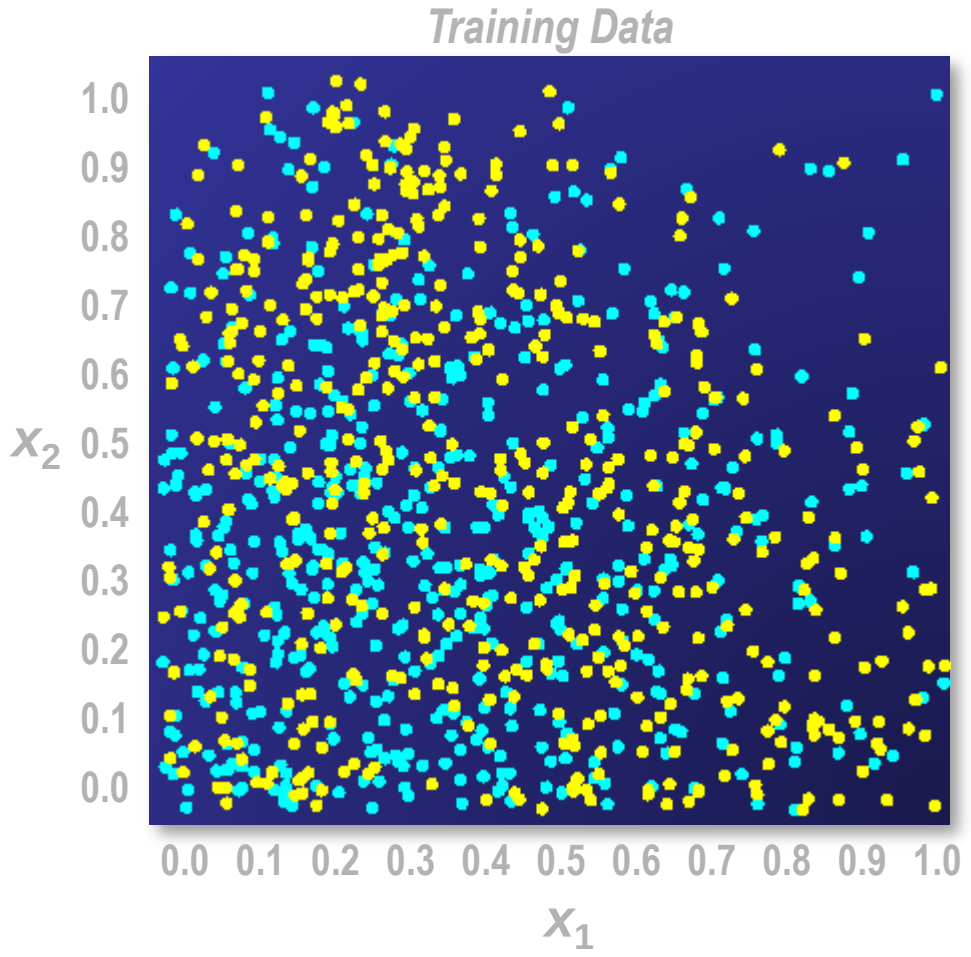
Predict dot color for each x_1 and x_2.



...

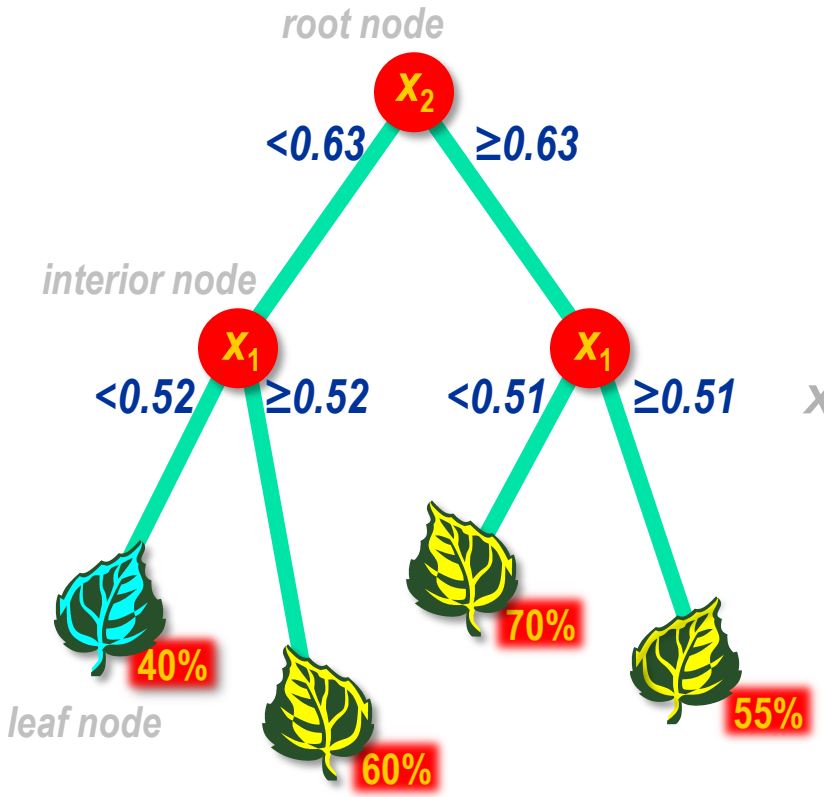
Simple Prediction Illustration

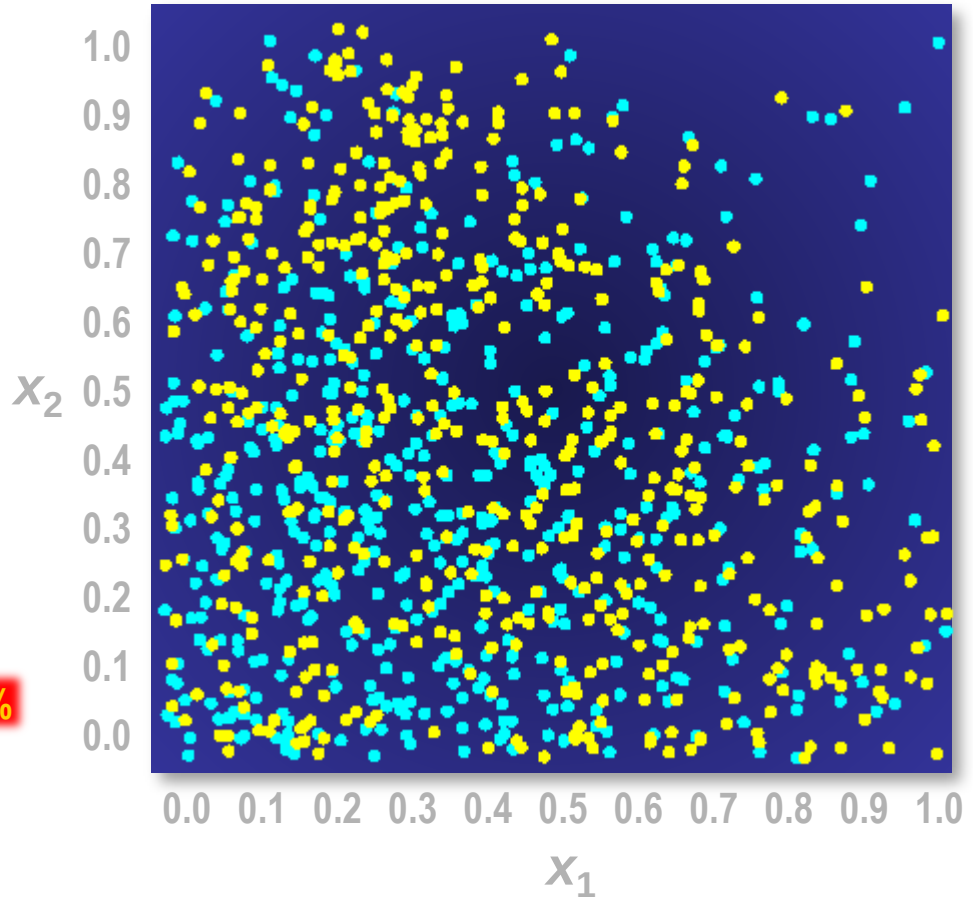
Predict dot color for each x_1 and x_2.



...

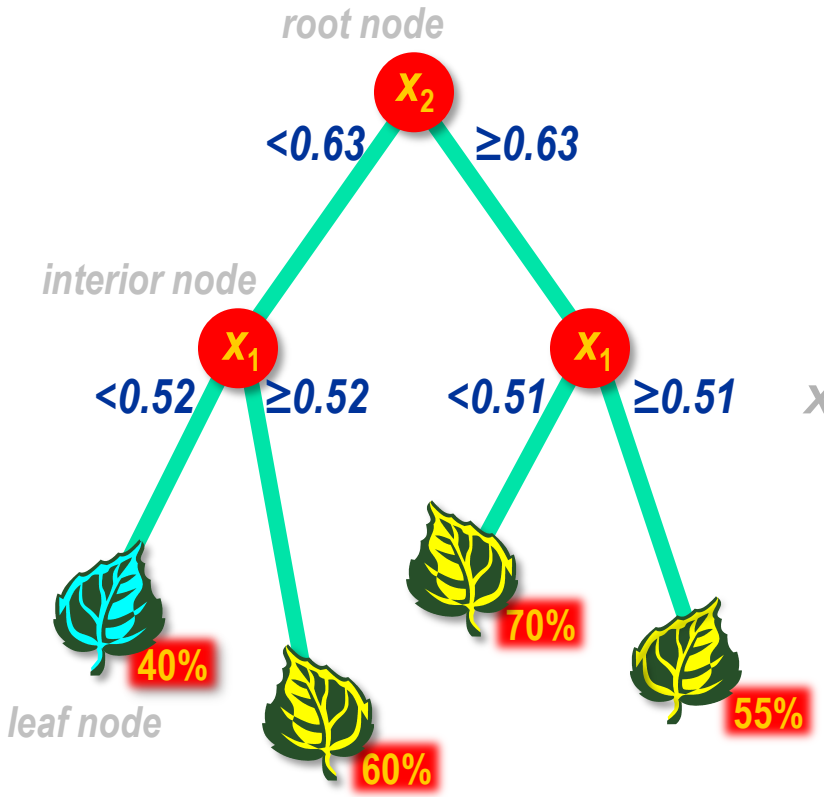
Decision Tree Prediction Rules

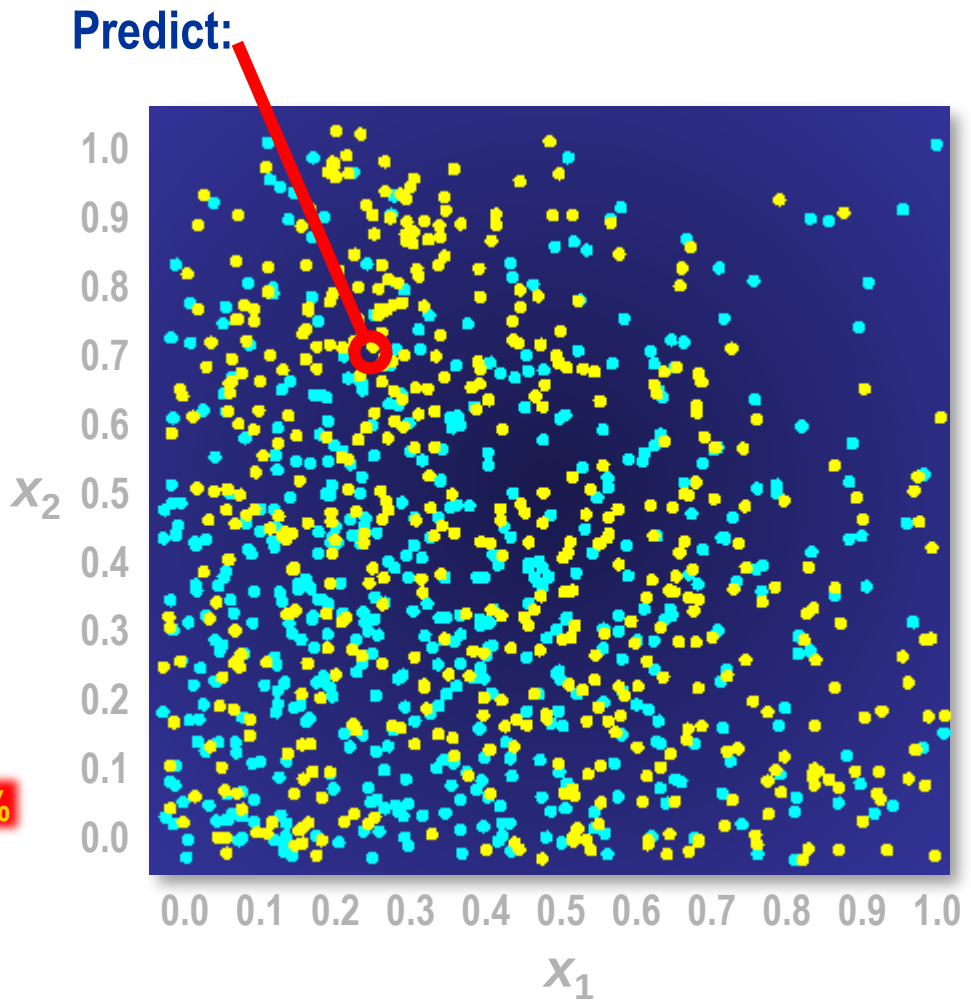




...

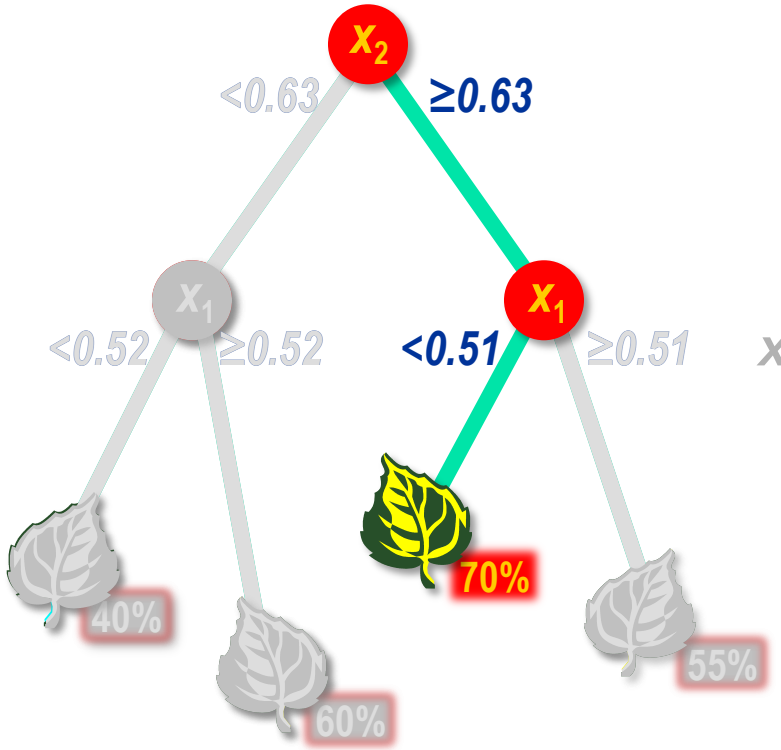
Decision Tree Prediction Rules

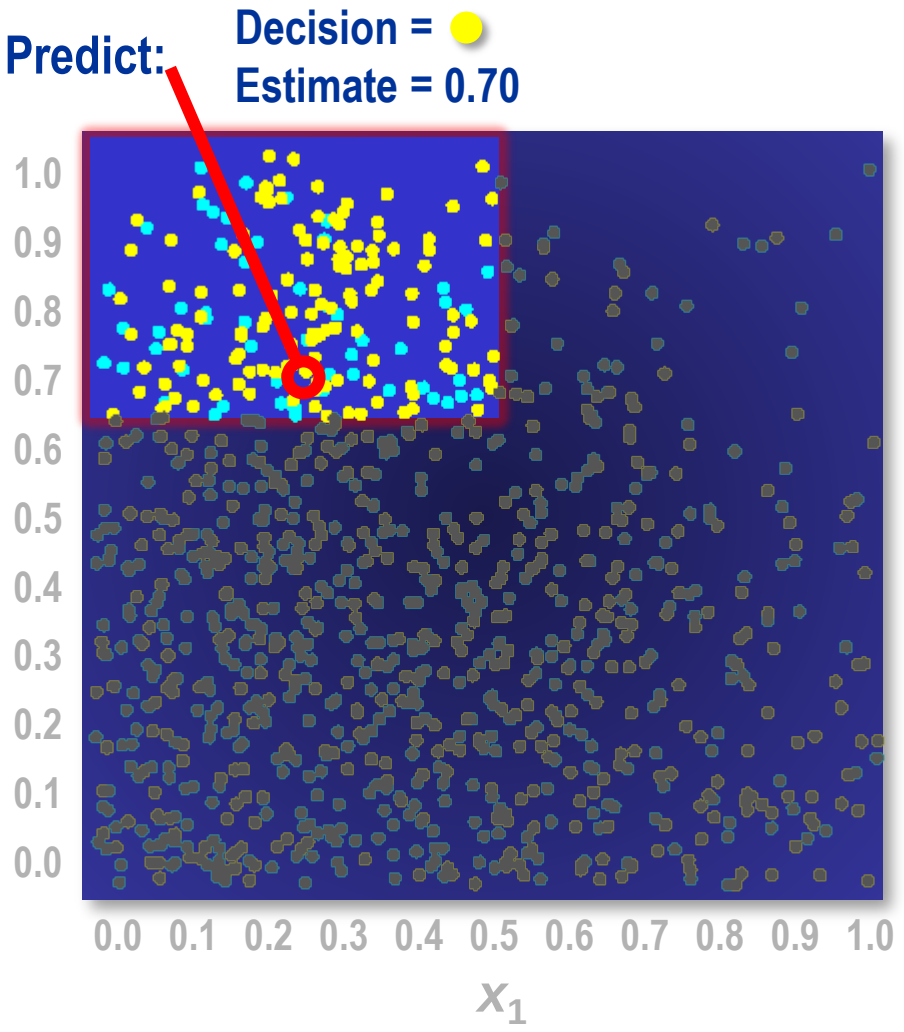




...

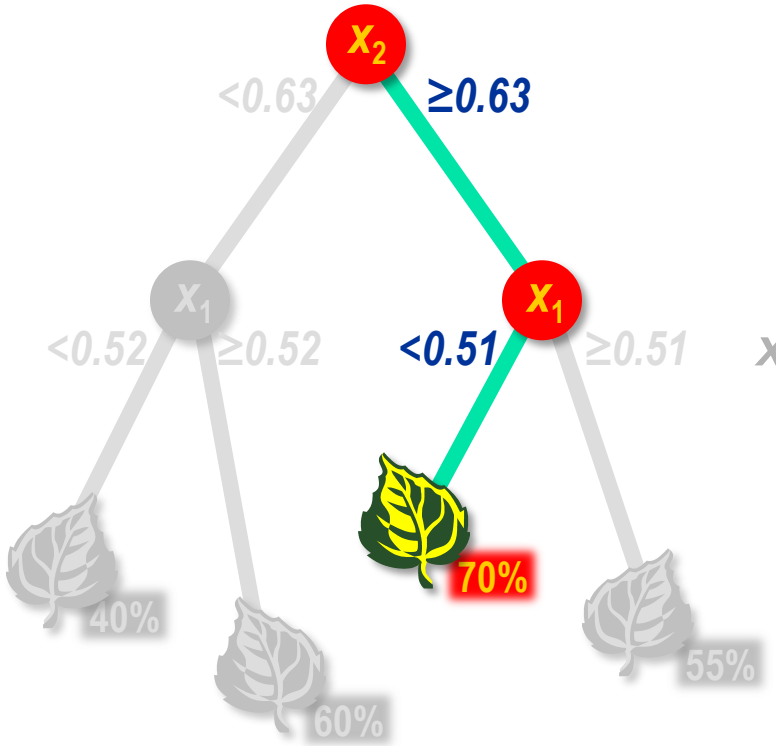
Decision Tree Prediction Rules

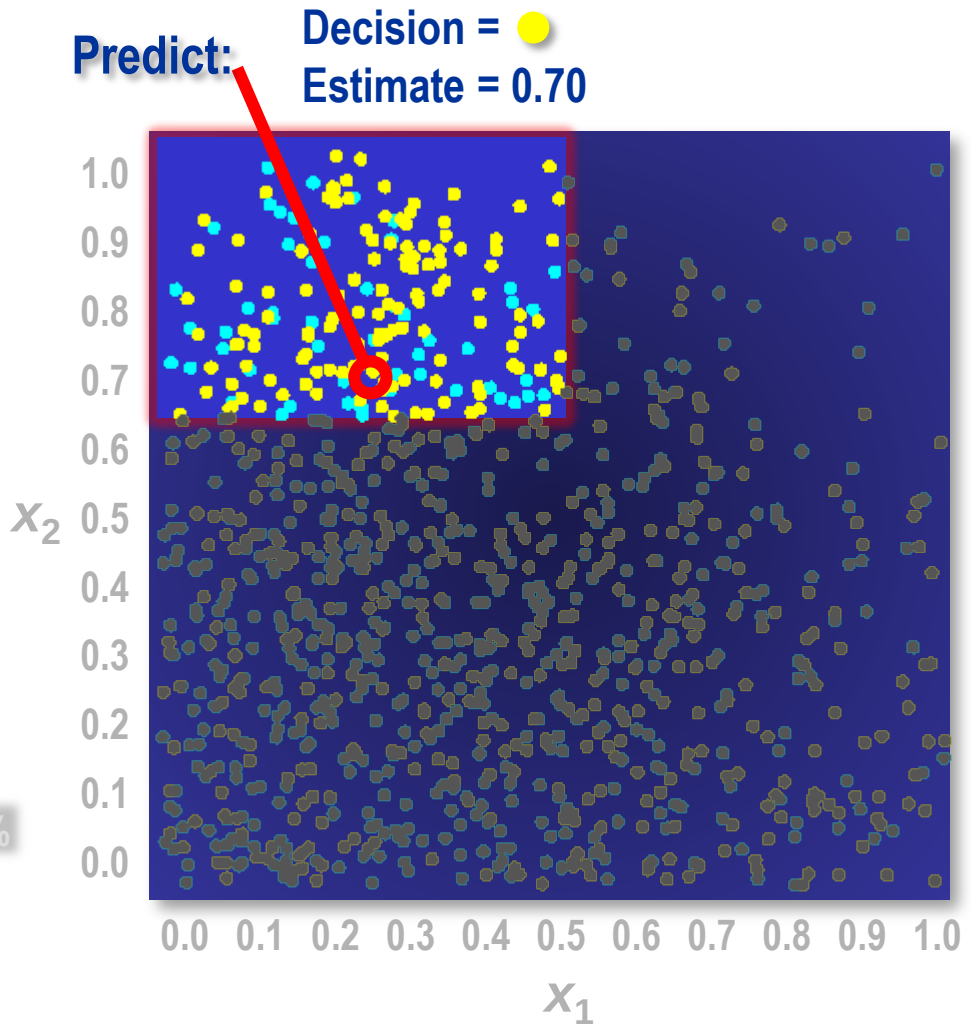




...

Decision Tree Prediction Rules





...

Model Essentials – Decision Trees

- **Predict new cases.**
- Select useful inputs.
- Optimize complexity.

Prediction rules

Split search

Pruning

...

Model Essentials – Decision Trees

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

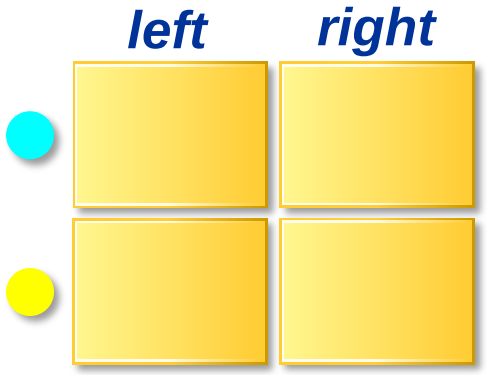
...

Model Essentials – Decision Trees

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

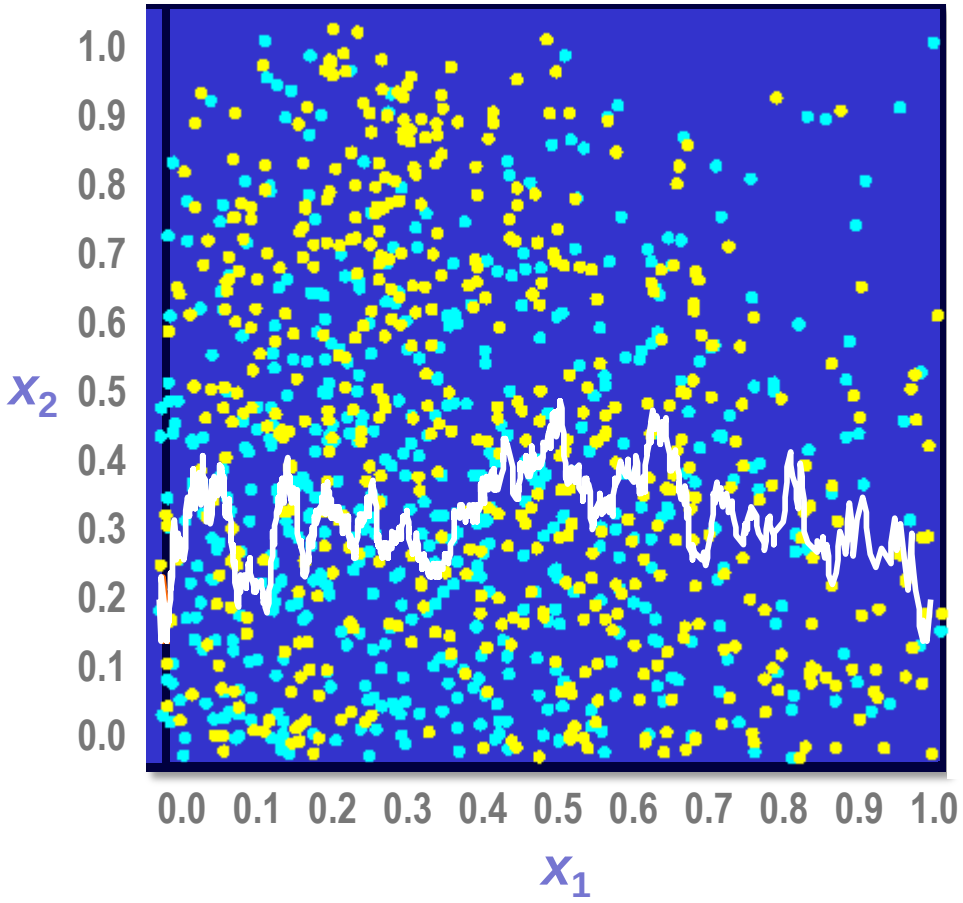
...

Decision Tree Split Search



Classification Matrix

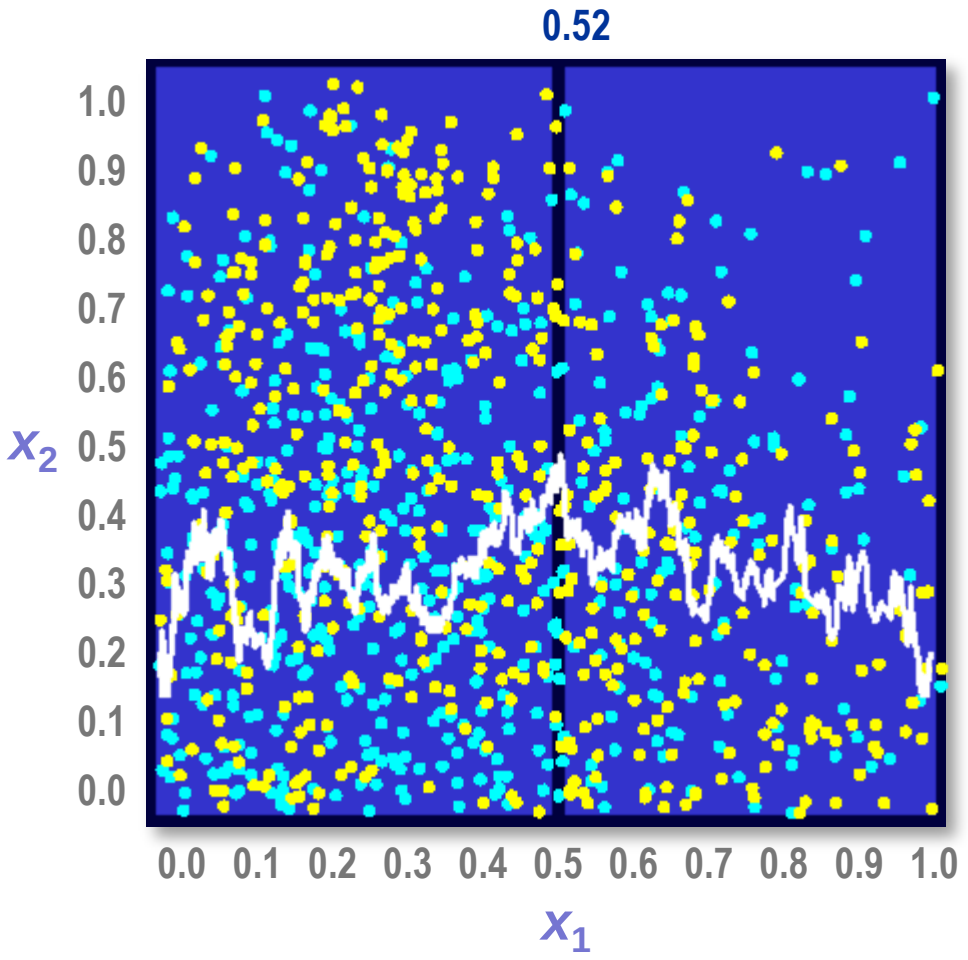
Calculate the *logworth* of every partition on input x_1.



Decision Tree Split Search

	left	*right*
●	53%	42%
●	47%	58%

max logworth(x_1)
0.95

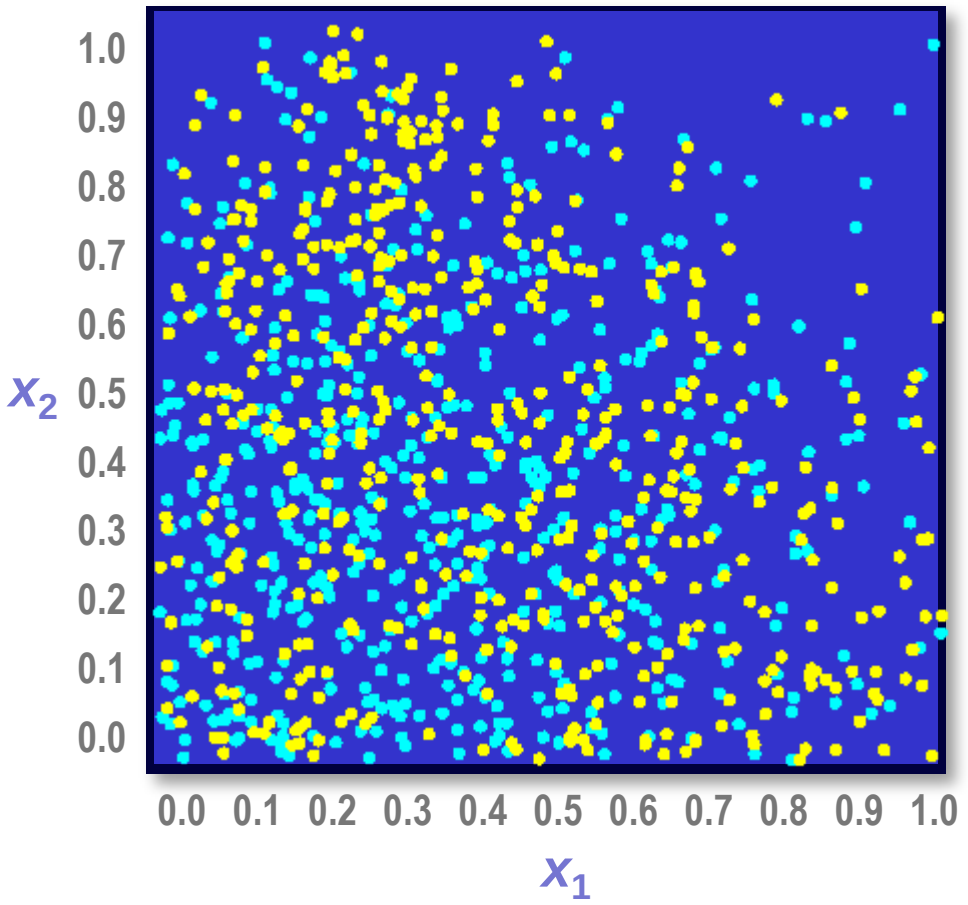


Select the partition with the maximum *logworth*.

Decision Tree Split Search

	left	*right*
○	53%	42%
○	47%	58%

max logworth(x_1) 0.95



Repeat for input x_2.

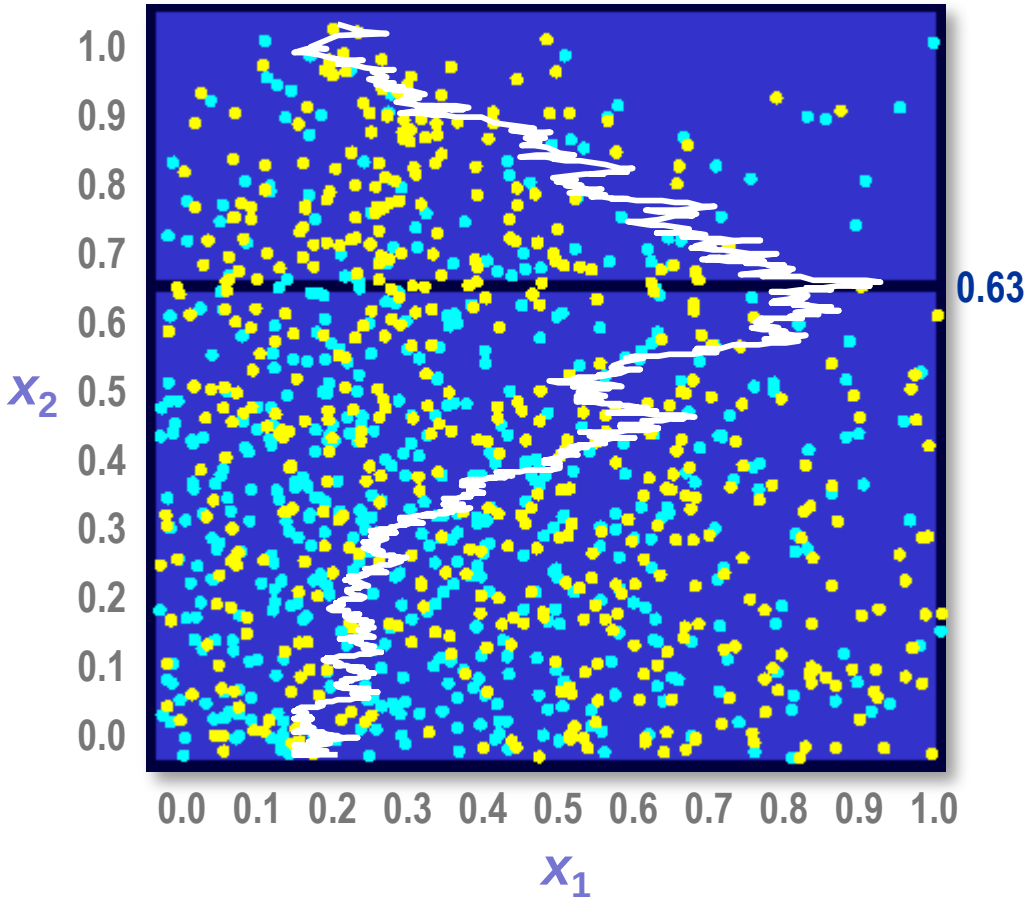
Decision Tree Split Search

	left	*right*
●	53%	42%
●	47%	58%

max logworth(x_1) 0.95

	bottom	*top*
●	54%	35%
●	46%	65%

max logworth(x_2) 4.92



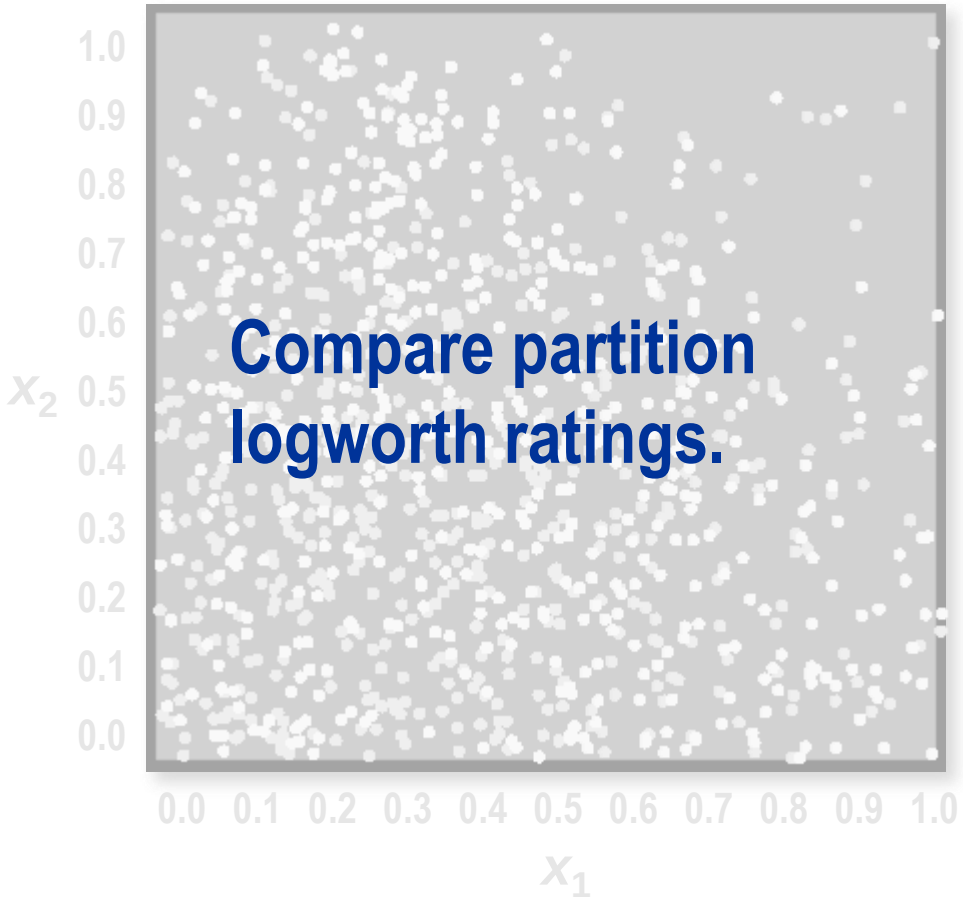
Decision Tree Split Search

	left	*right*
●	53%	42%
●	47%	58%

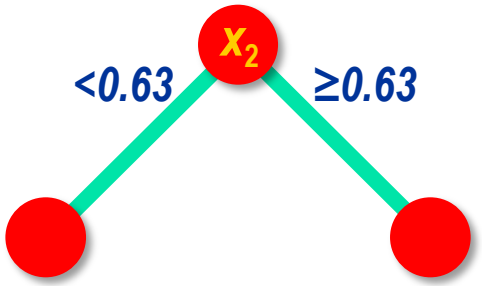
max logworth(x_1)
0.95

	bottom	*top*
●	54%	35%
●	46%	65%

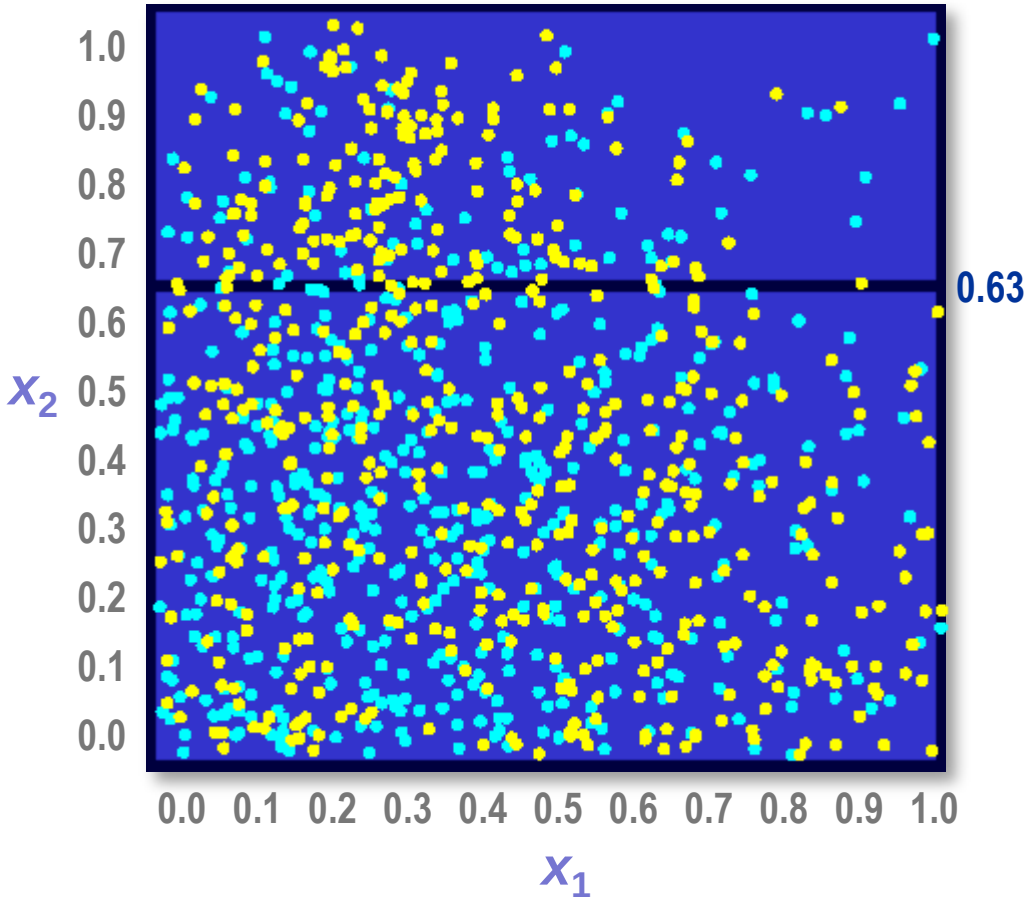
max logworth(x_2)
4.92



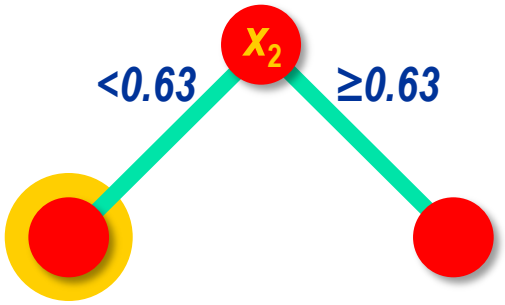
Decision Tree Split Search



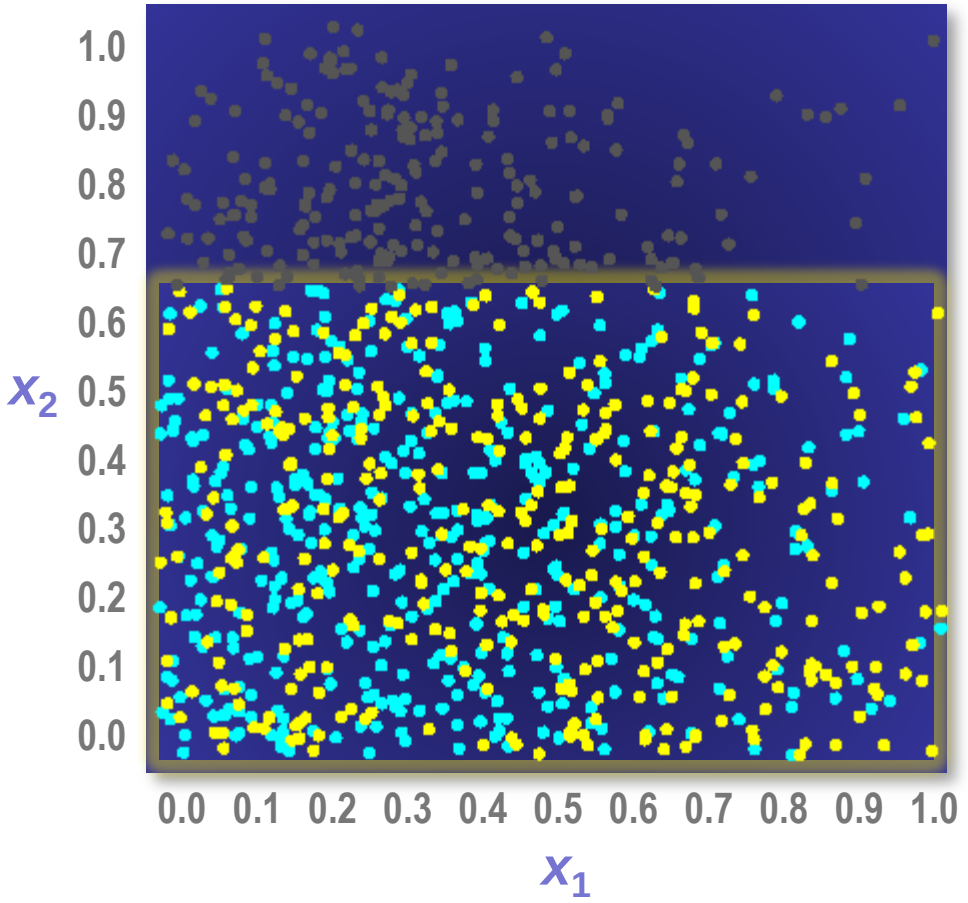
Create a partition rule from the best partition across all inputs.



Decision Tree Split Search



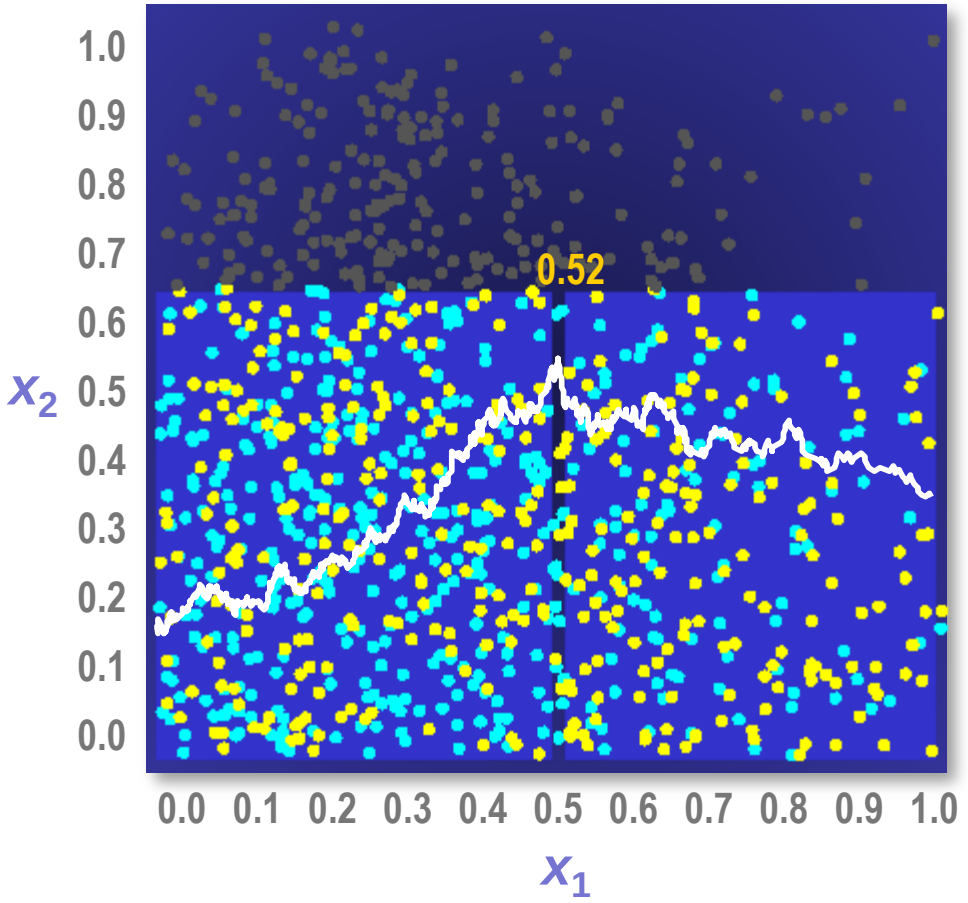
Repeat the process in each subset.



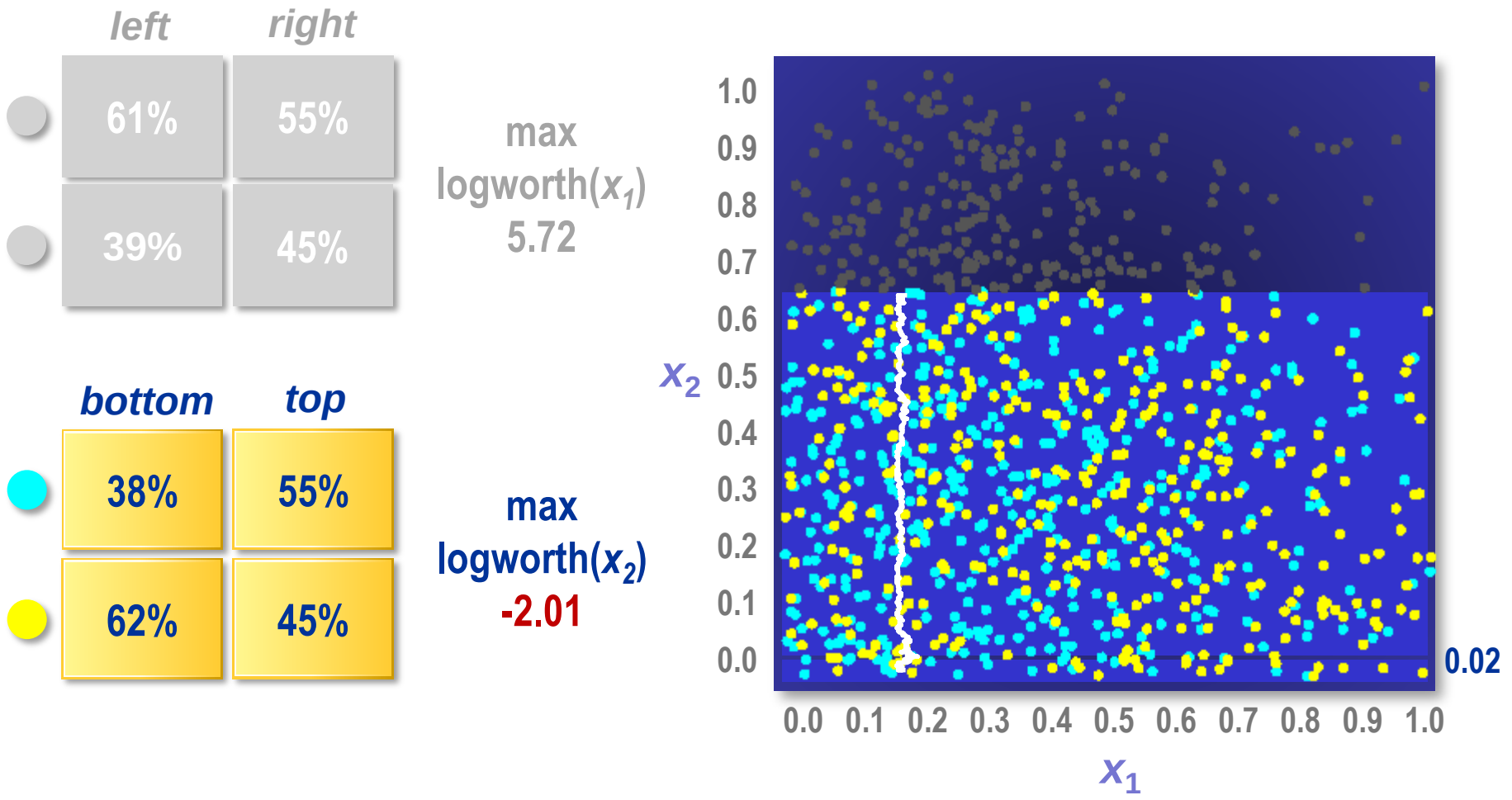
Decision Tree Split Search

	left	*right*
●	61%	55%
●	39%	45%

max logworth(x_1)
5.72



Decision Tree Split Search



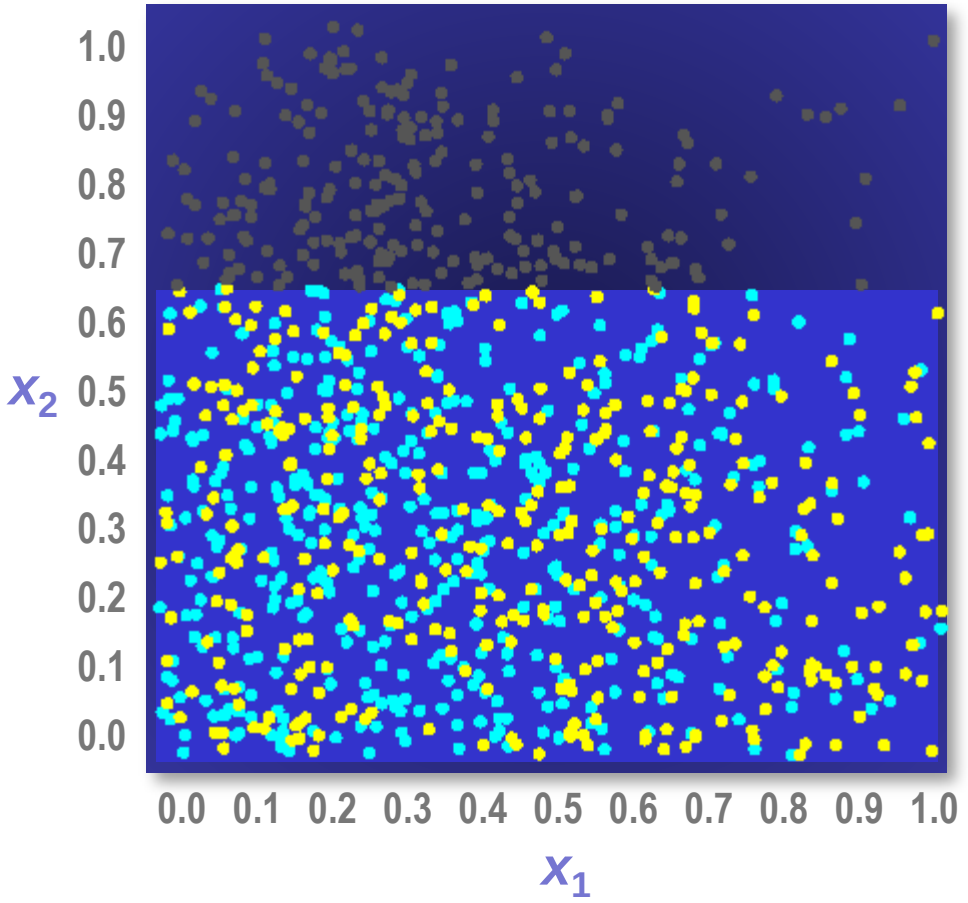
Decision Tree Split Search

	left	*right*
●	61%	55%
●	39%	45%

max logworth(x_1)
5.72

	bottom	*top*
●	38%	55%
●	62%	45%

max logworth(x_2)
-2.01



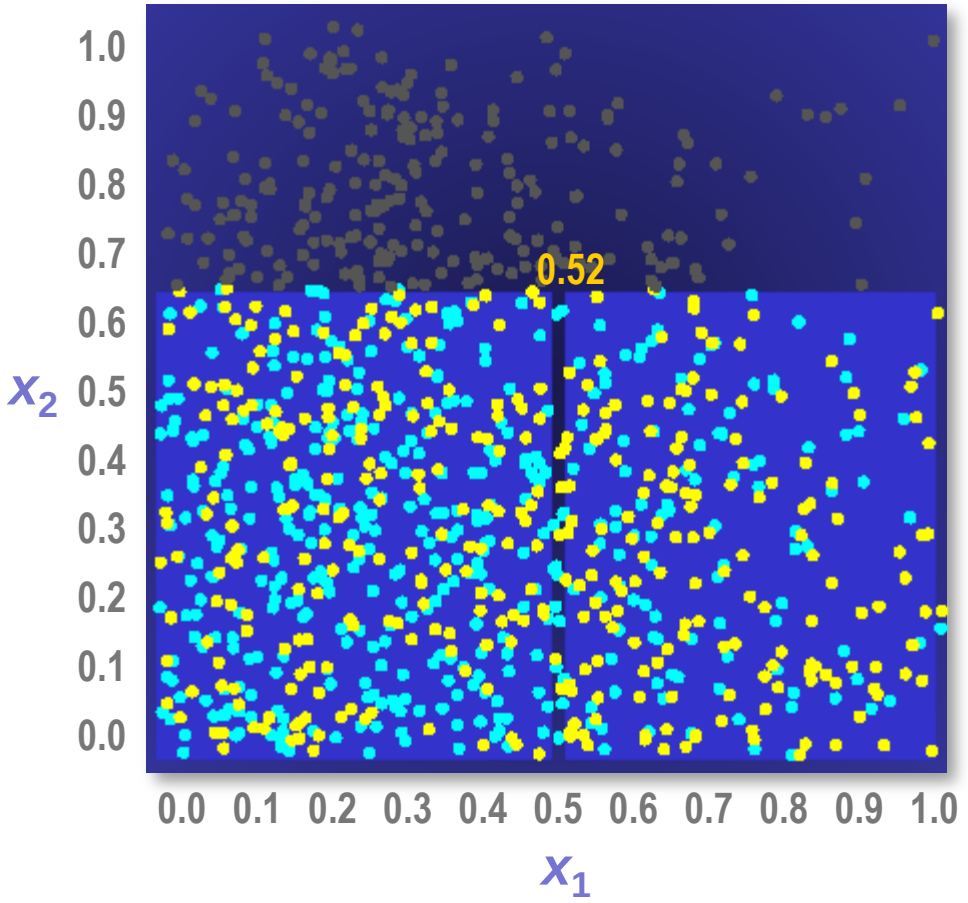
Decision Tree Split Search

	left	*right*
●	61%	55%
●	39%	45%

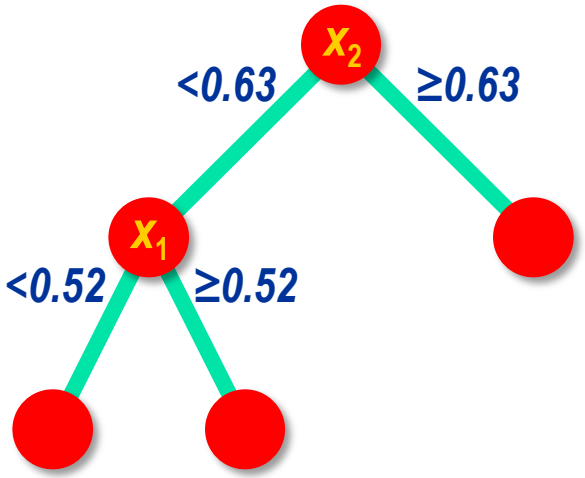
max logworth(x_1)
5.72

	bottom	*top*
●	38%	55%
●	62%	45%

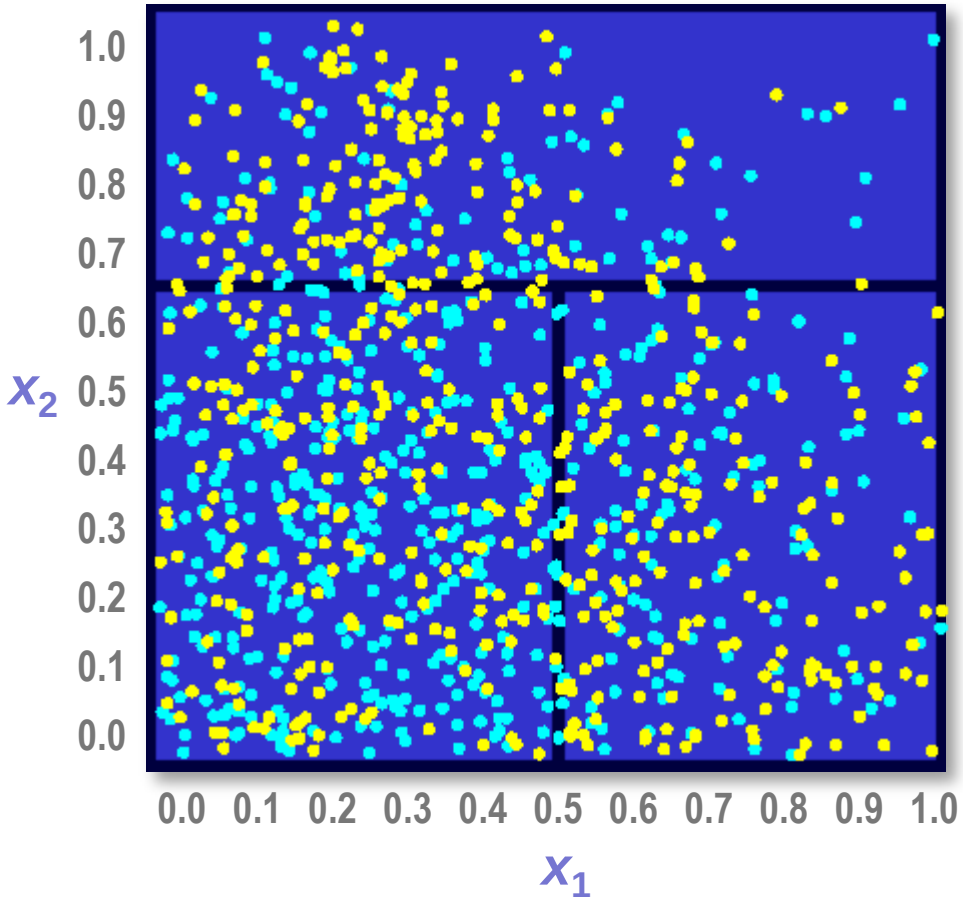
max logworth(x_2)
-2.01



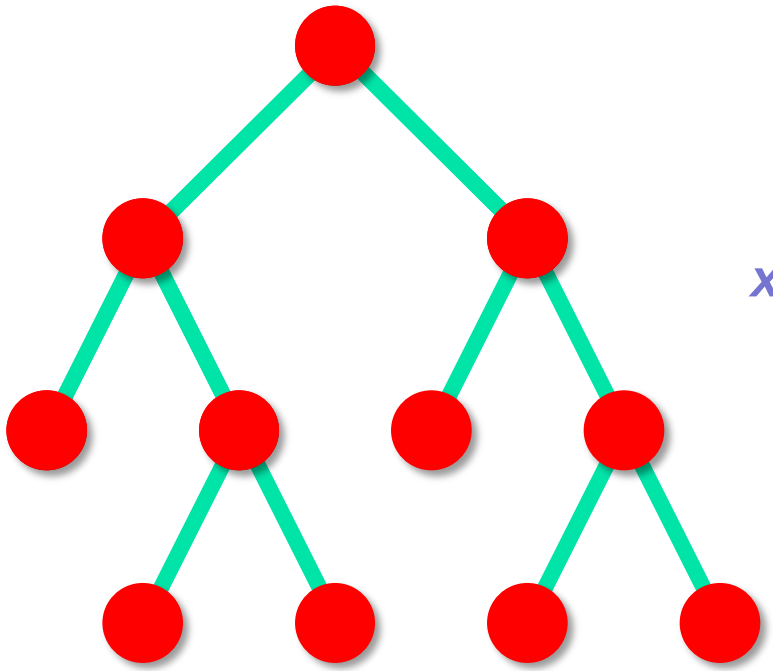
Decision Tree Split Search

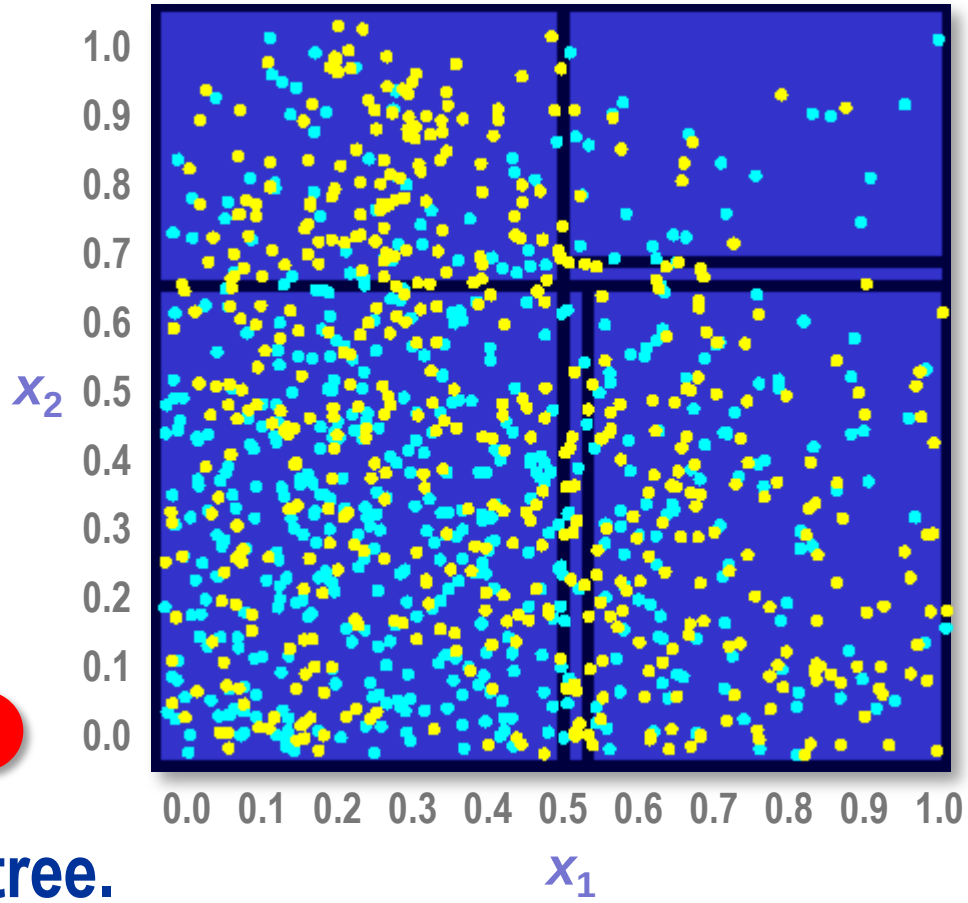


Create a second partition rule.



Decision Tree Split Search





Repeat to form a maximal tree.

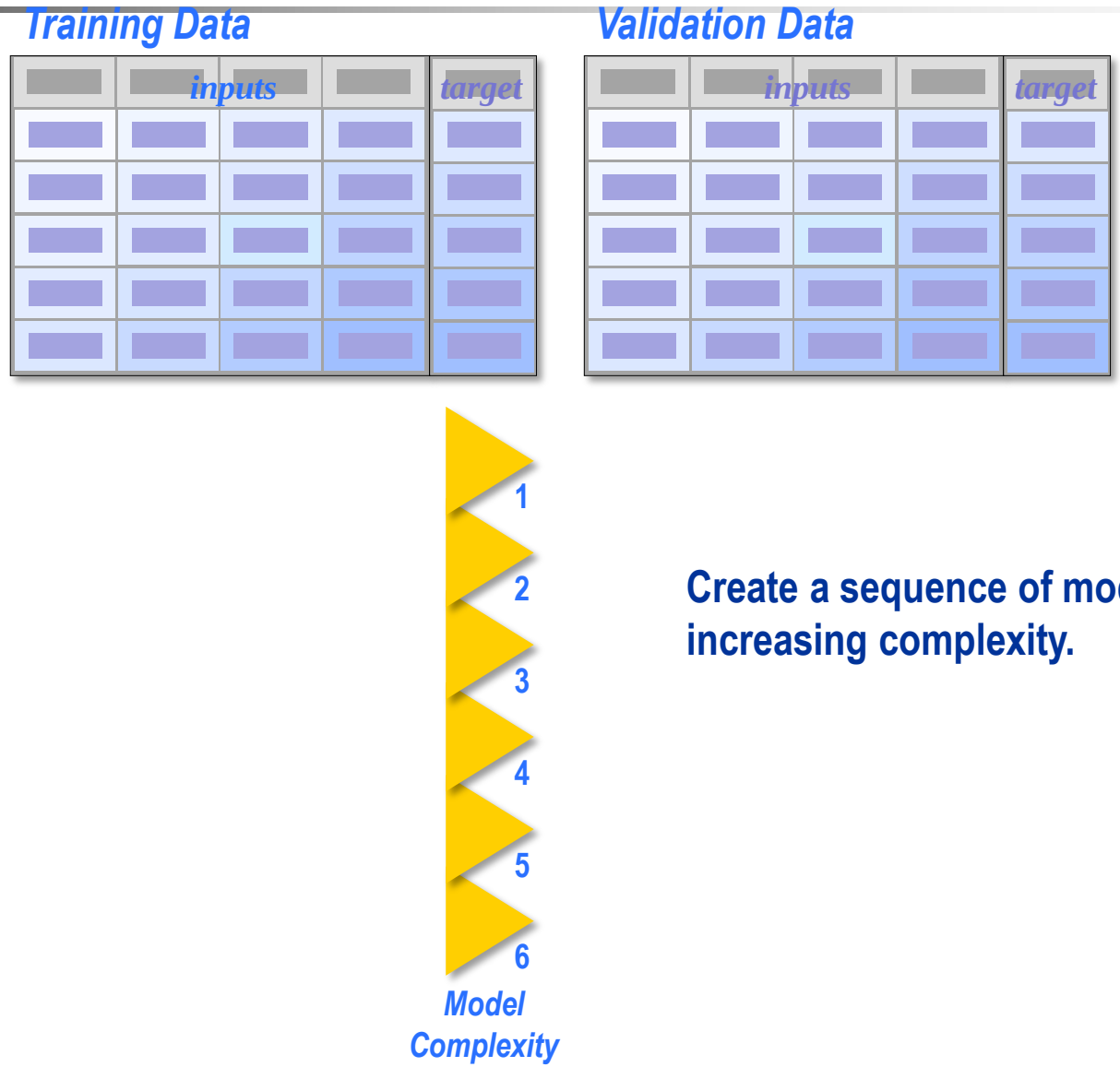
Model Essentials – Decision Trees

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

Model Essentials – Decision Trees

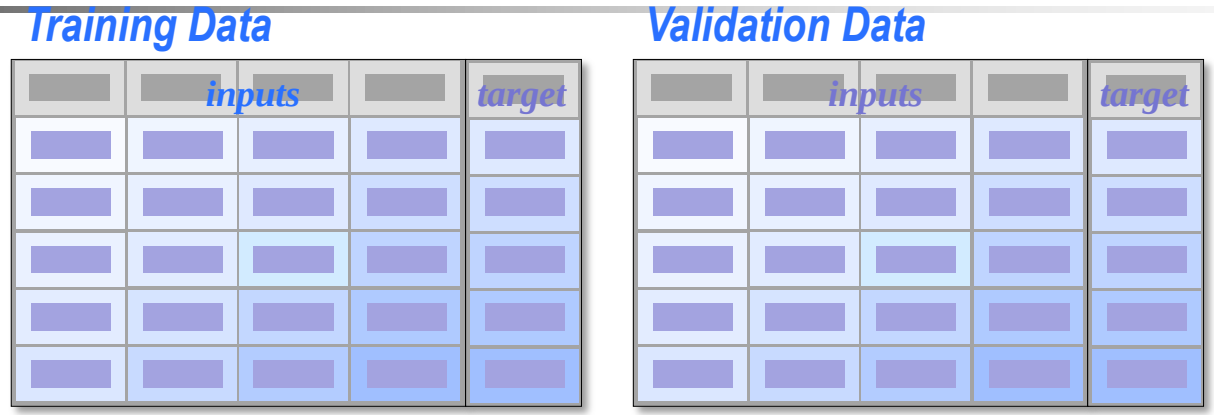
- Predict new cases. — Prediction rules
- Select useful inputs. — Split search
- Optimize complexity. — **Pruning**

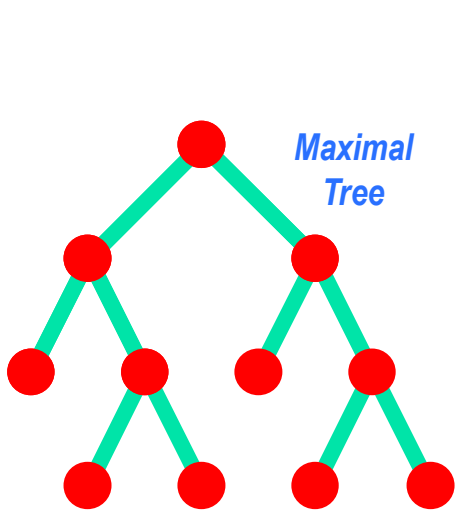
Predictive Model Sequence



Create a sequence of models with increasing complexity.

Maximal Tree

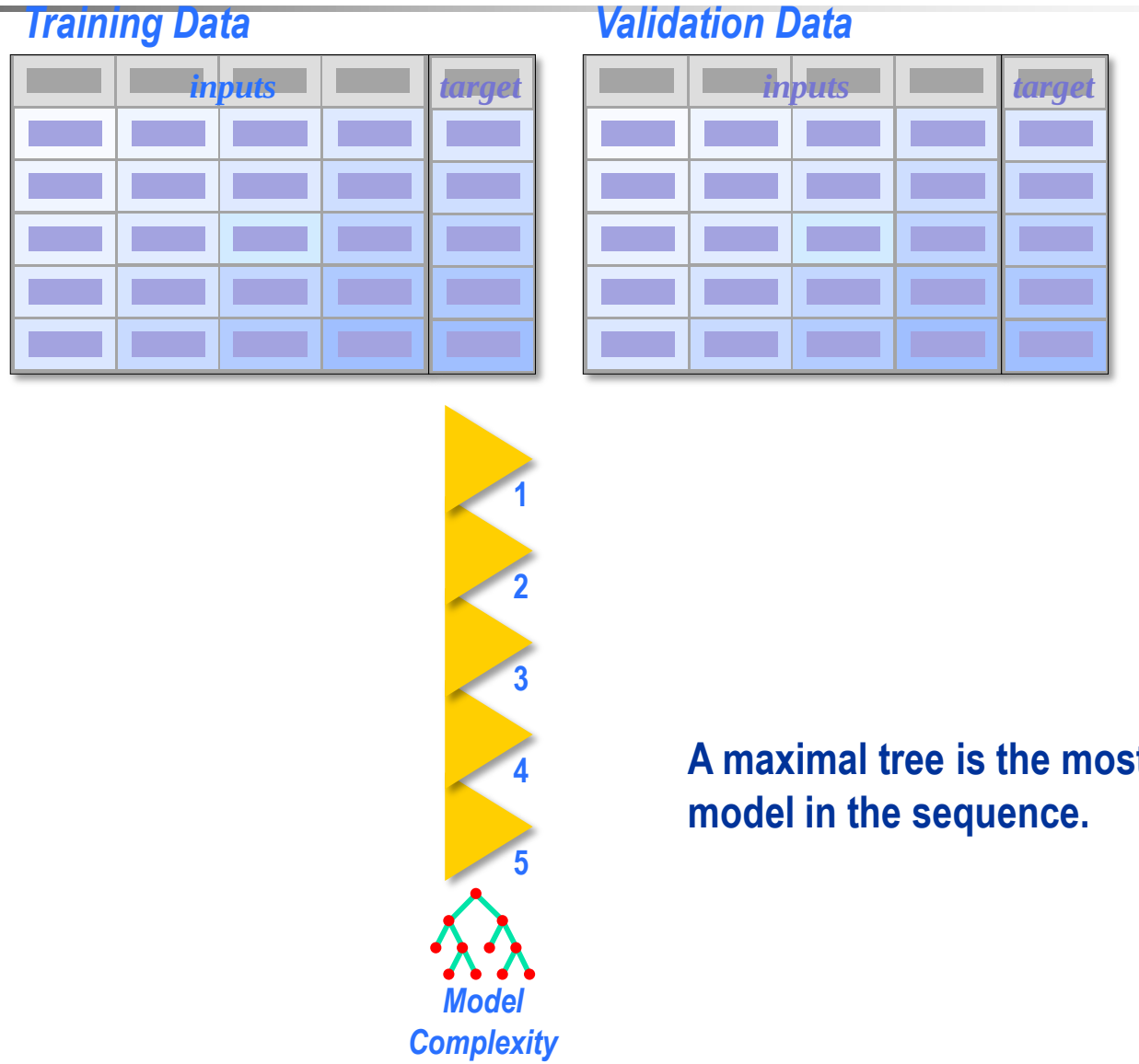




Create a sequence of models with increasing complexity.

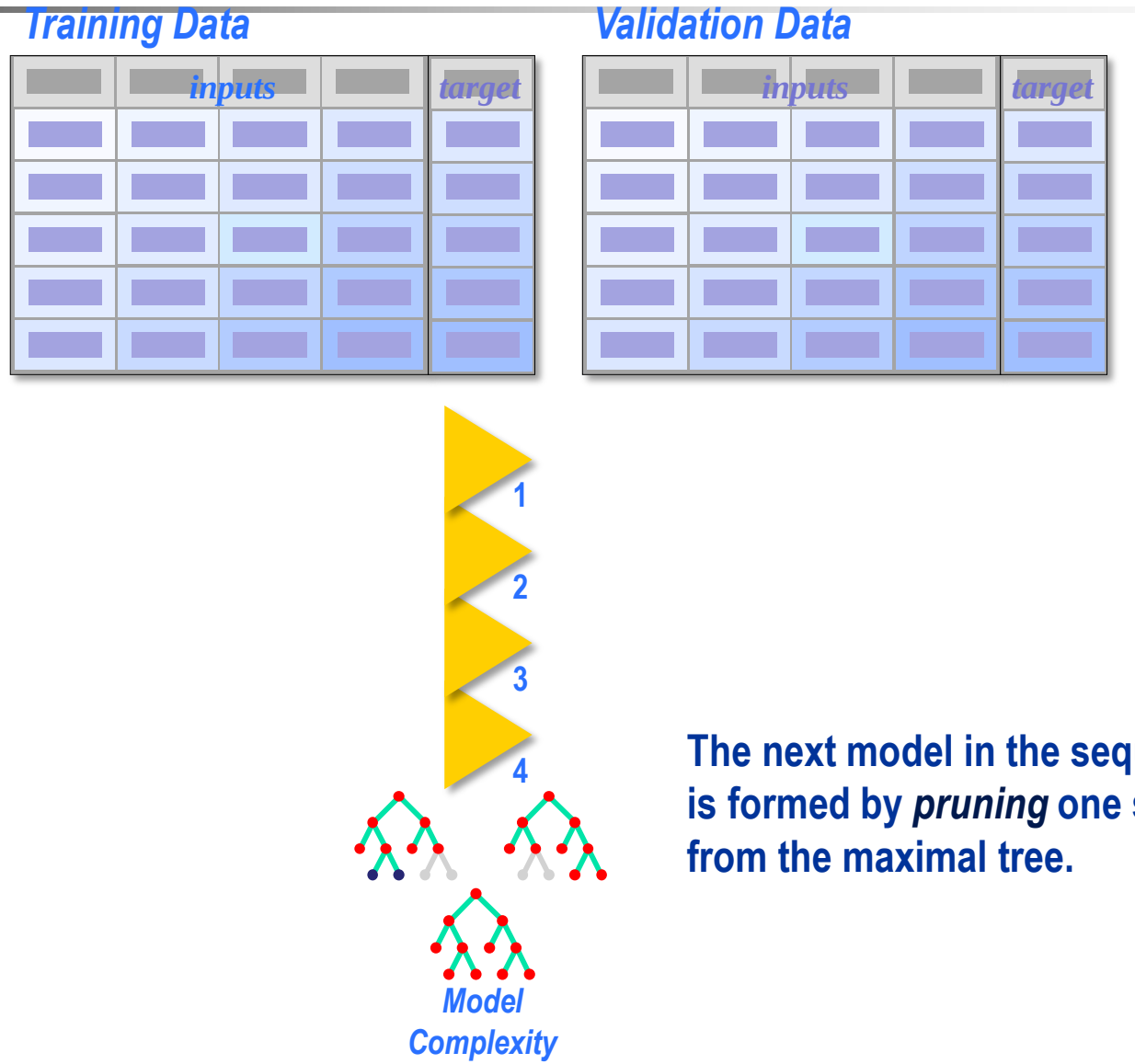
A maximal tree is the most complex model in the sequence.

Maximal Tree



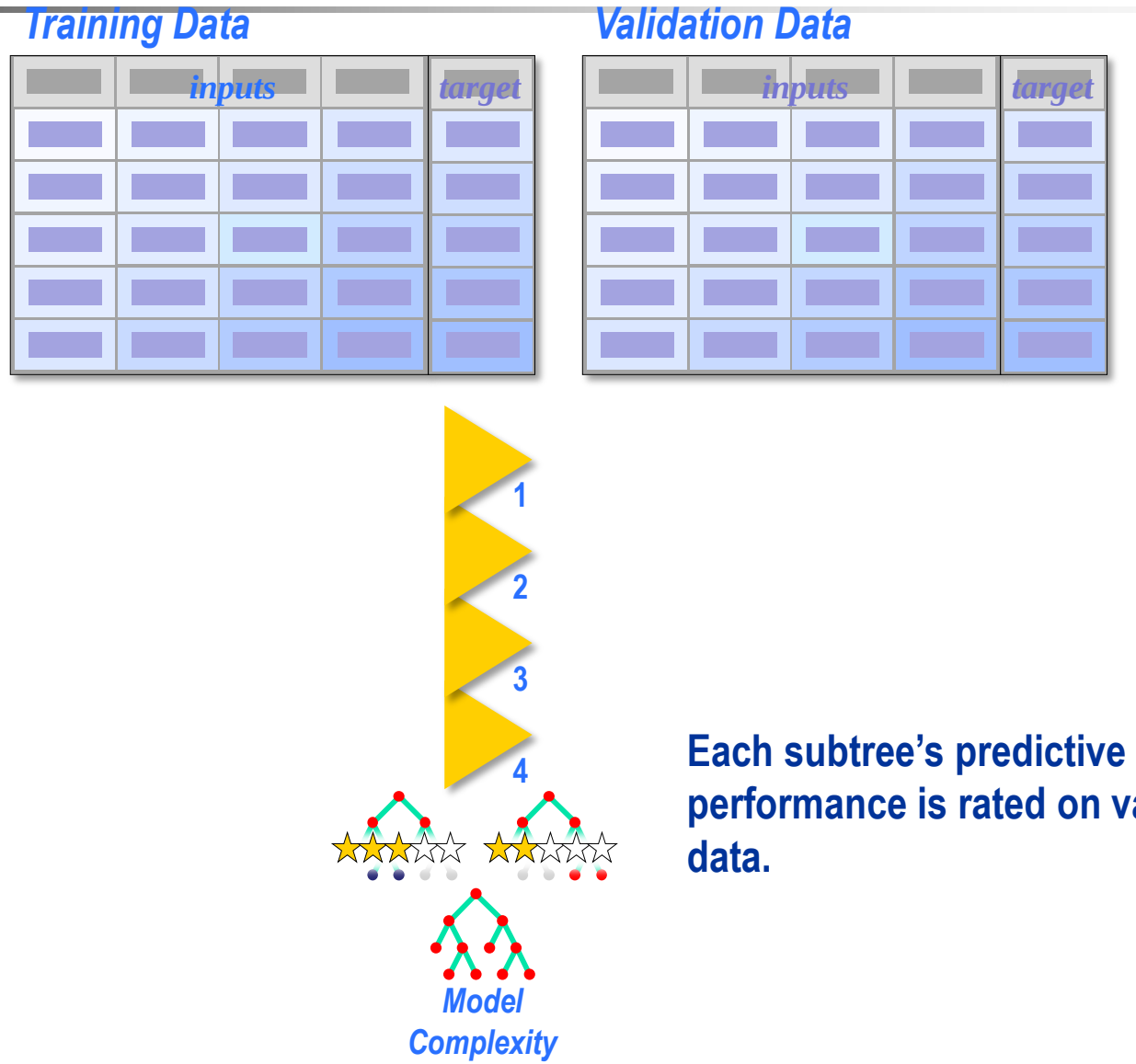
A maximal tree is the most complex model in the sequence.

Pruning One Split



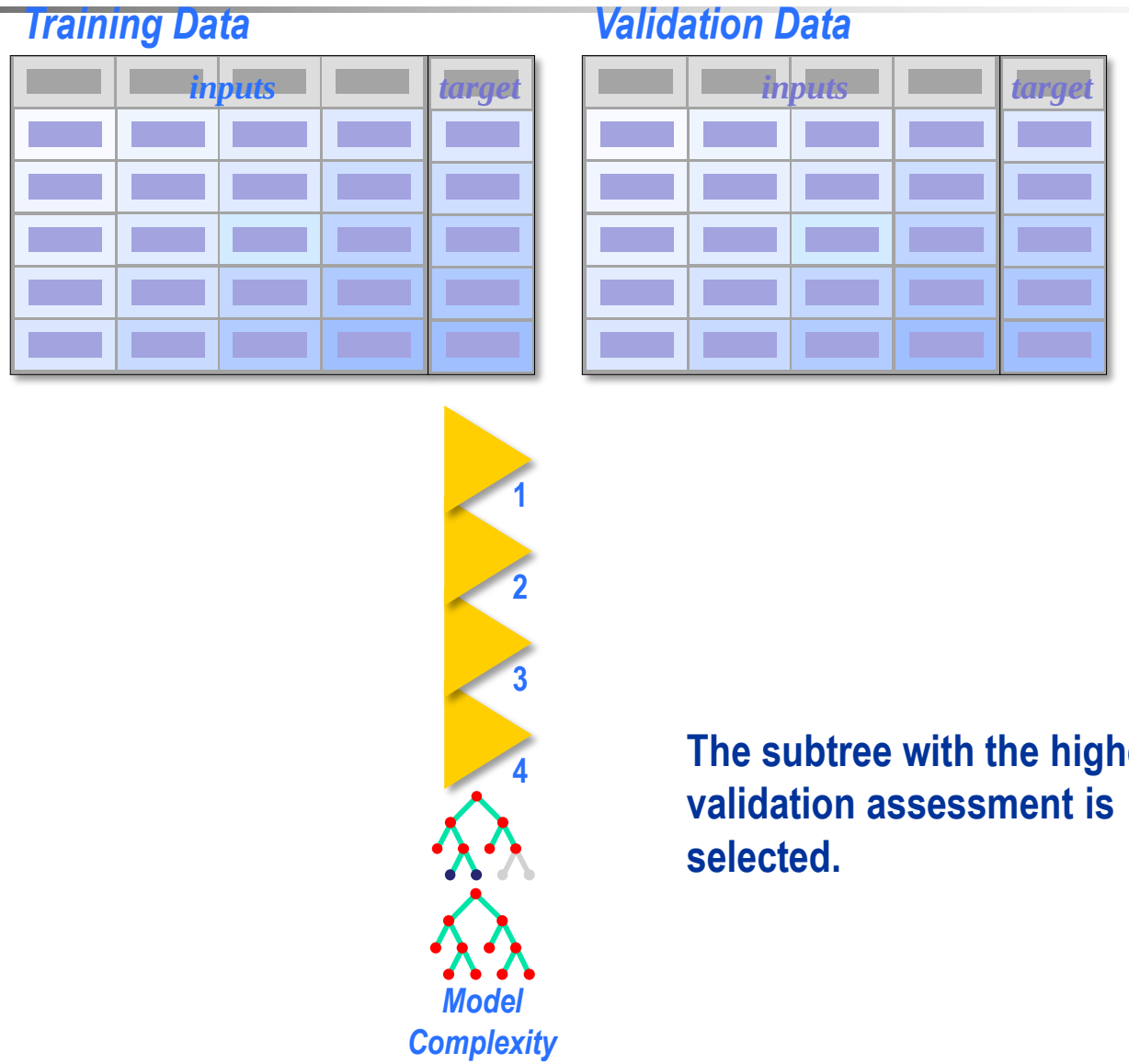
The next model in the sequence is formed by *pruning* one split from the maximal tree.

Pruning One Split



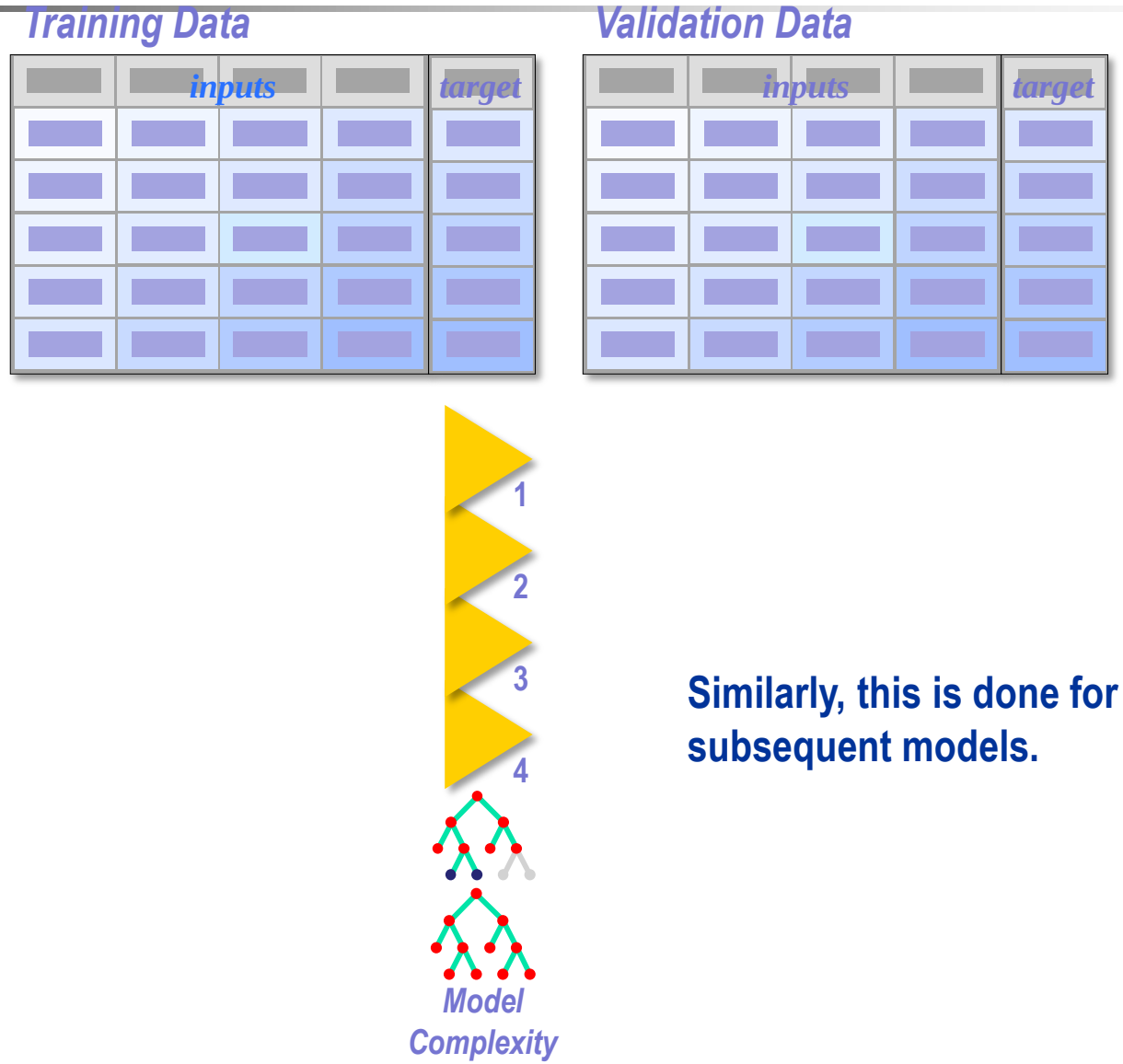
Each subtree's predictive performance is rated on validation data.

Pruning One Split



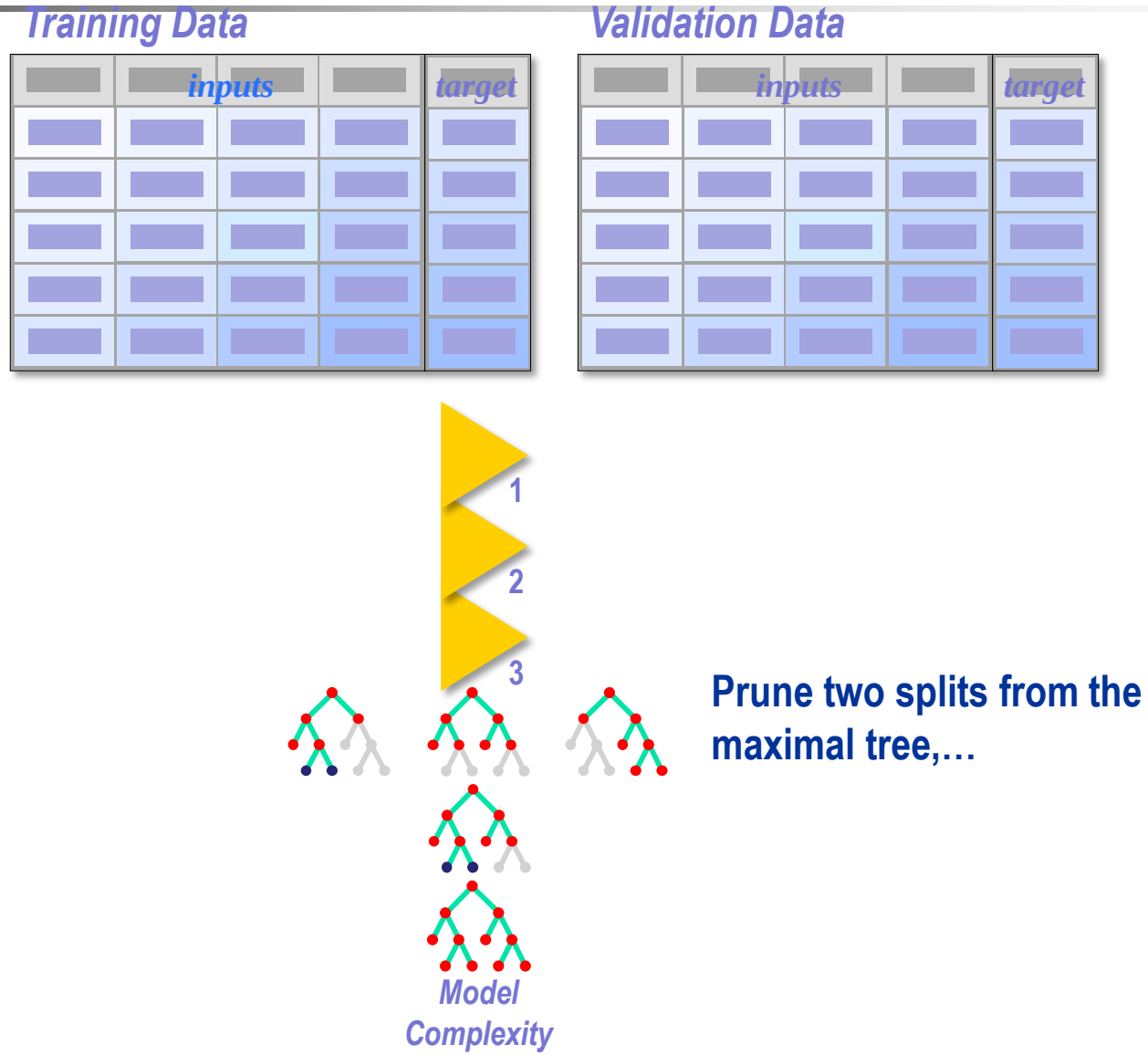
The subtree with the highest validation assessment is selected.

Pruning Two Splits



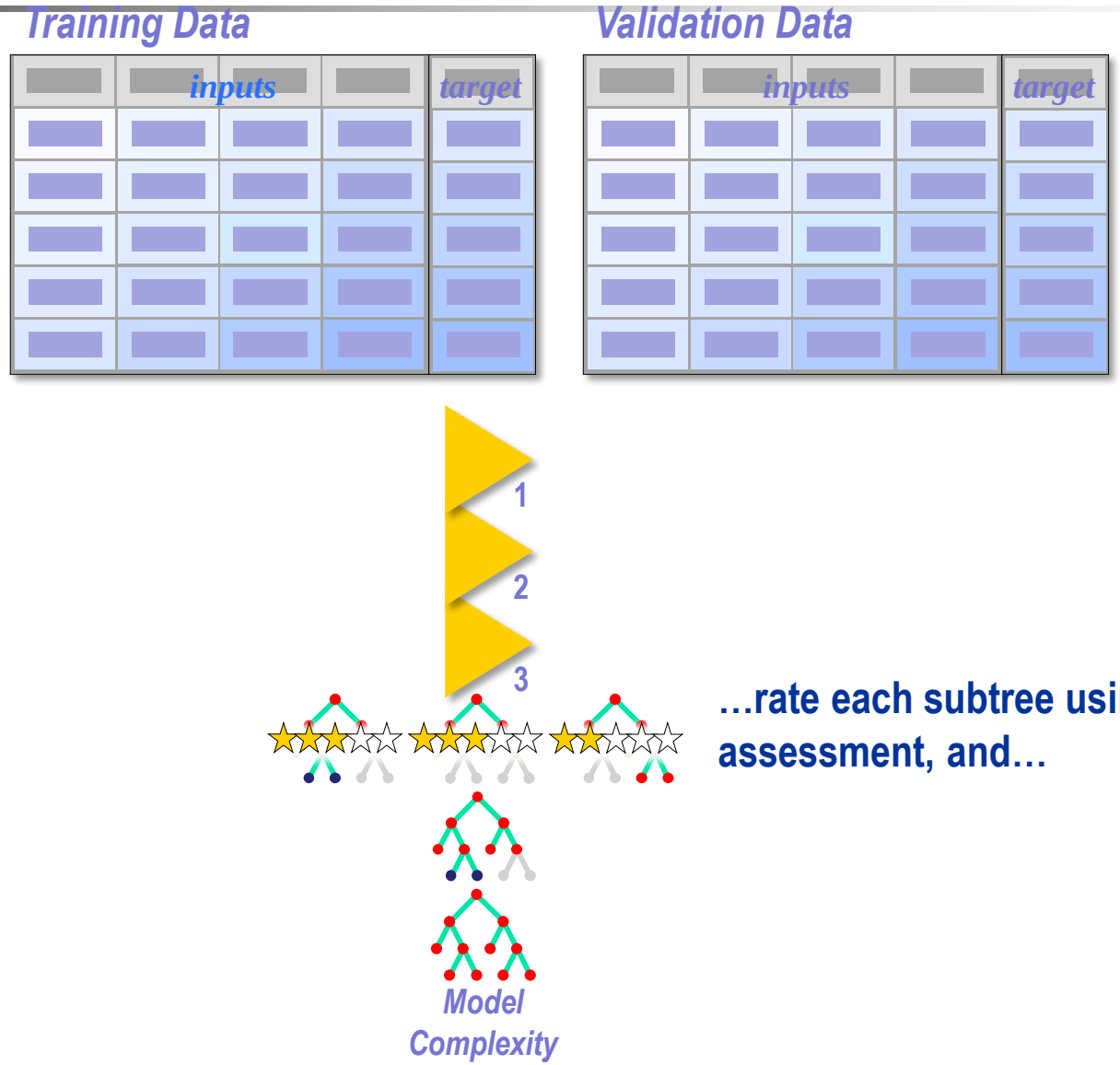
Similarly, this is done for subsequent models.

Pruning Two Splits



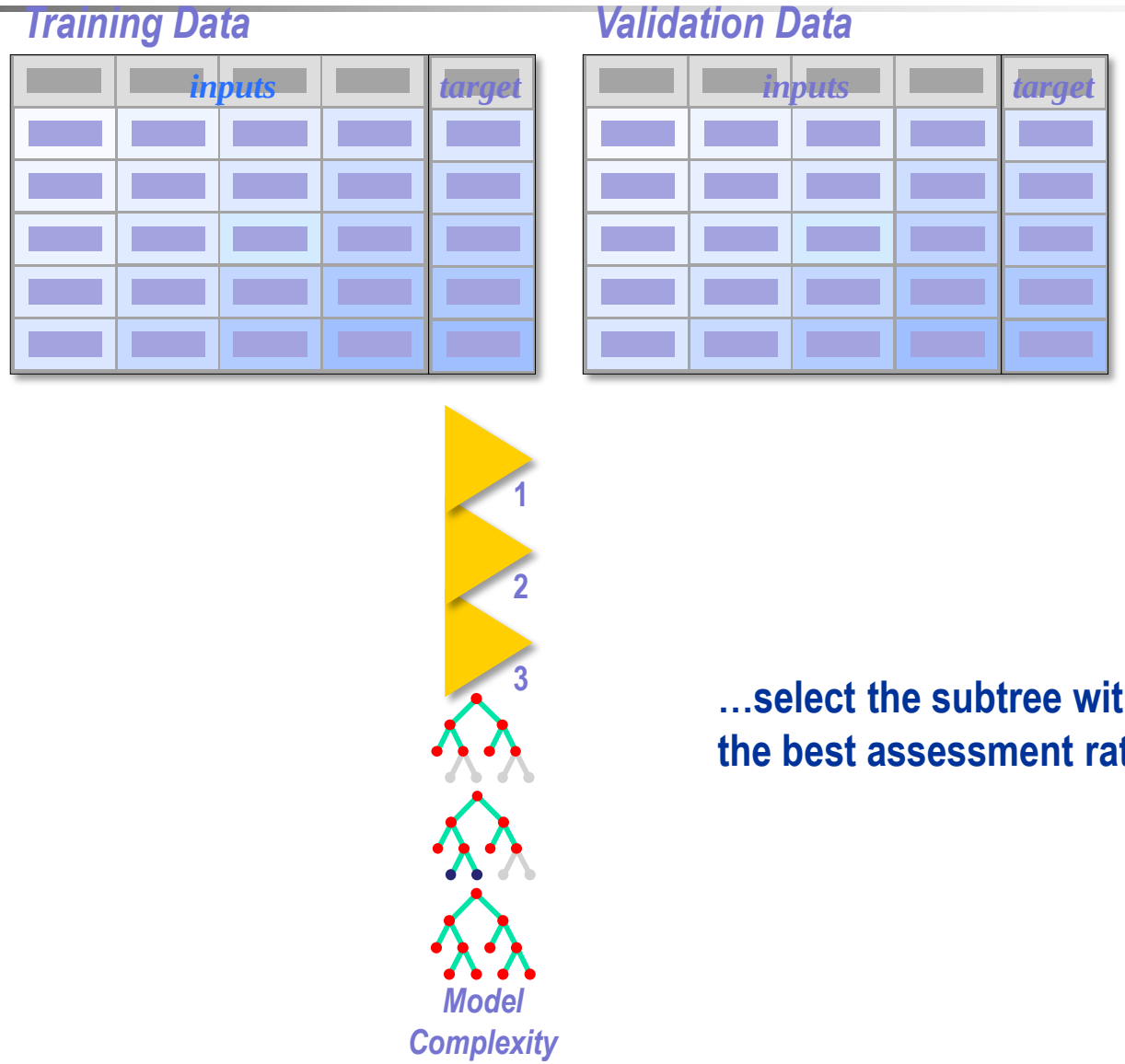
Prune two splits from the maximal tree,…

Pruning Two Splits



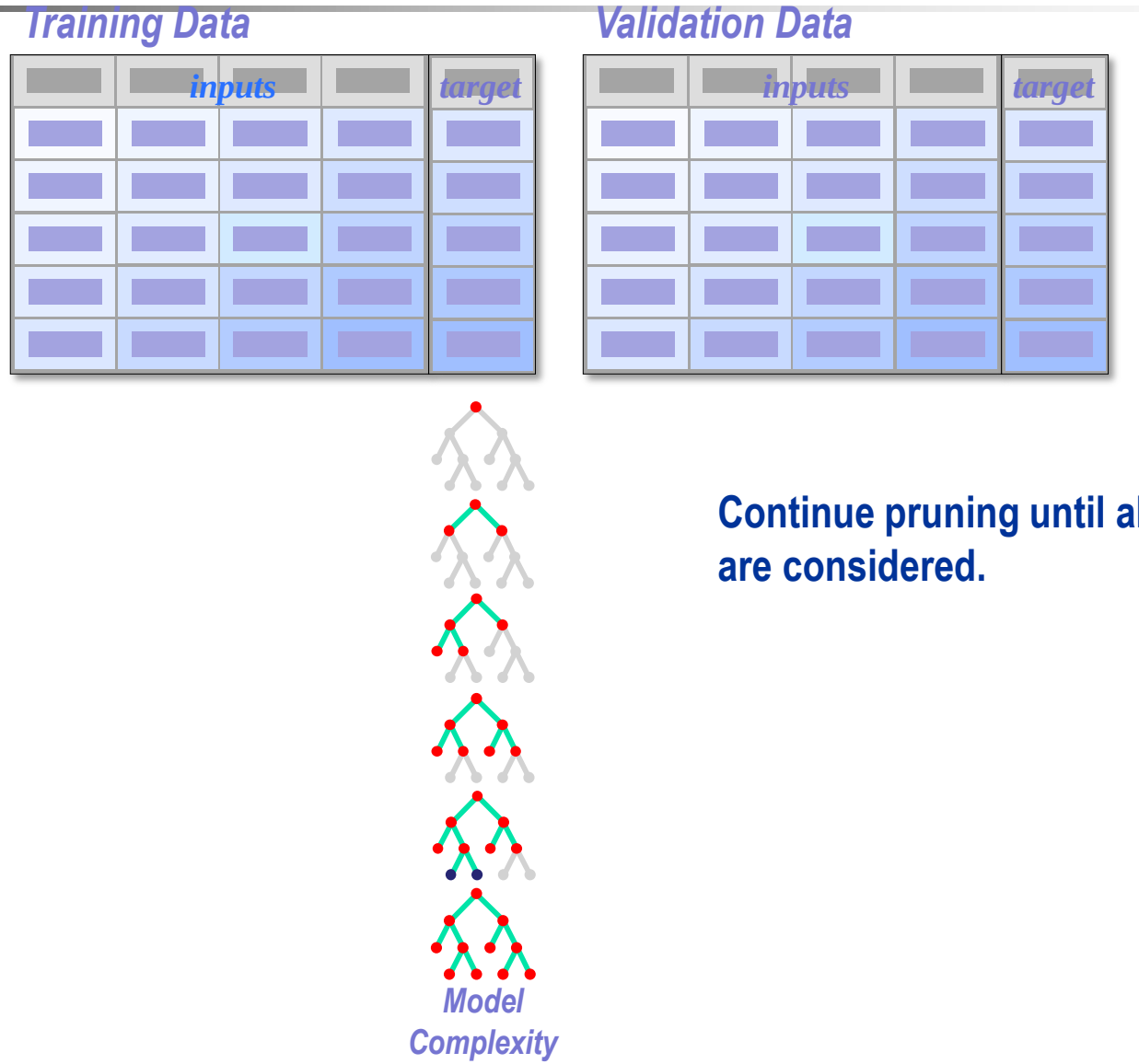
…rate each subtree using validation assessment, and…

Pruning Two Splits



…select the subtree with the best assessment rating.

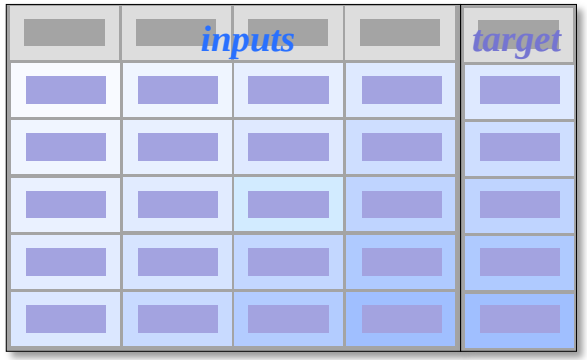
Subsequent Pruning



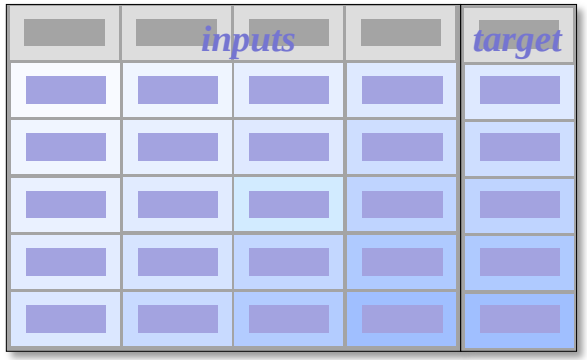
Continue pruning until all subtrees are considered.

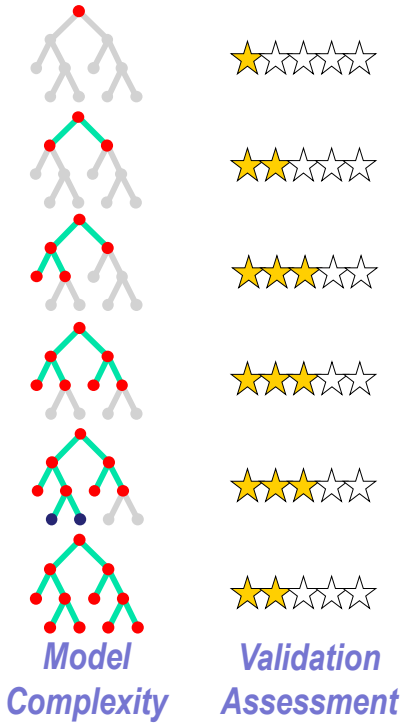
Selecting the Best Tree

- Training Data



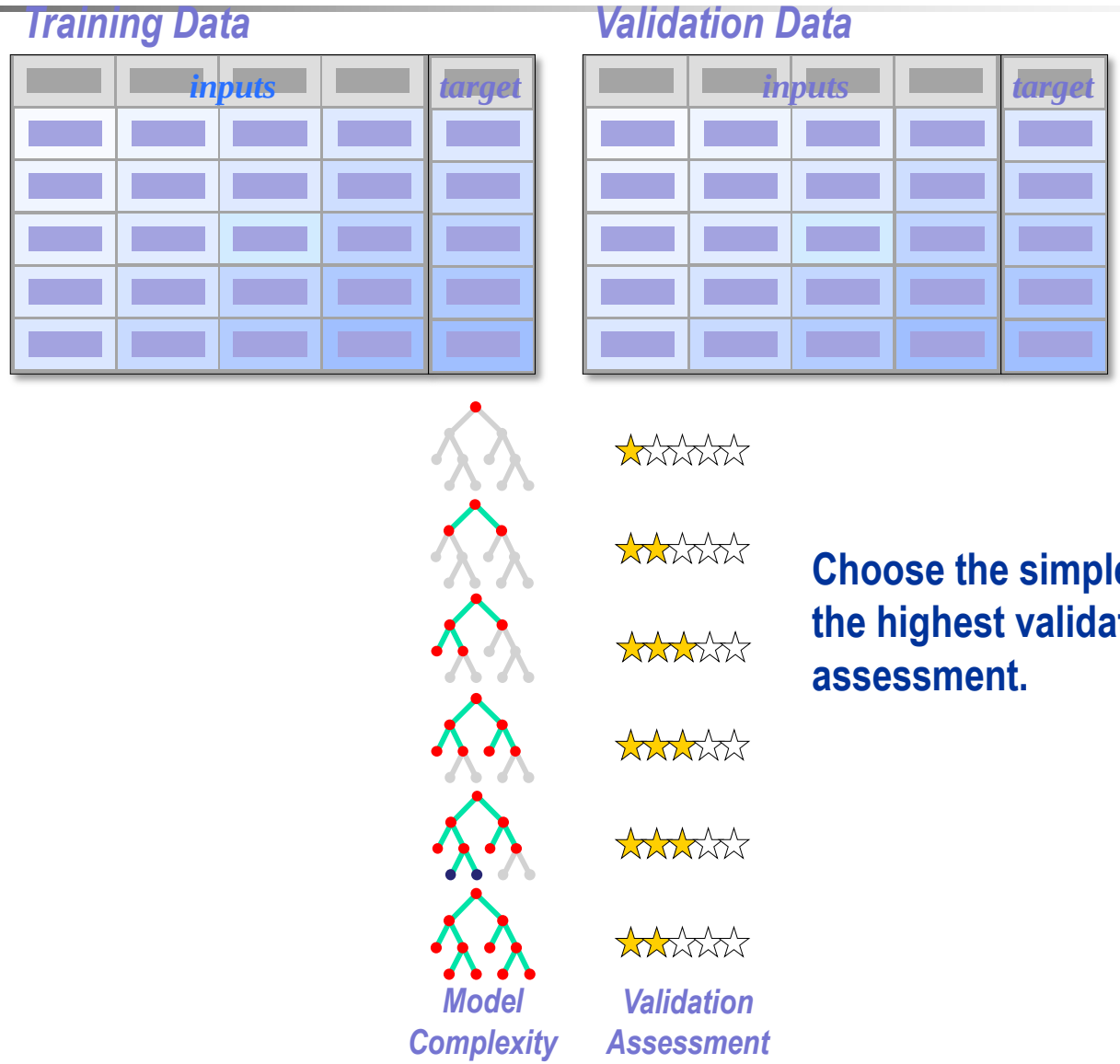
Validation Data





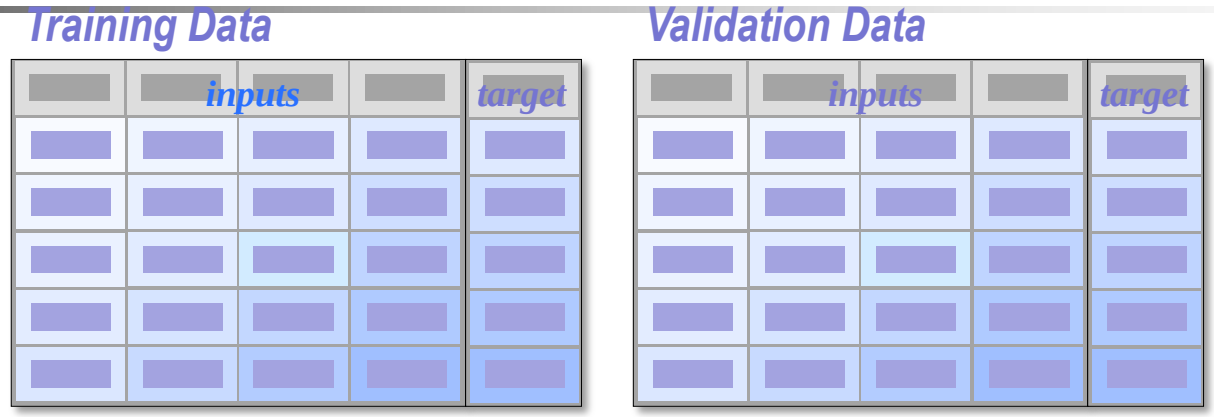
Compare validation assessment between tree complexities.

Validation Assessment



Choose the simplest model with the highest validation assessment.

Validation Assessment



What are appropriate validation assessment ratings?

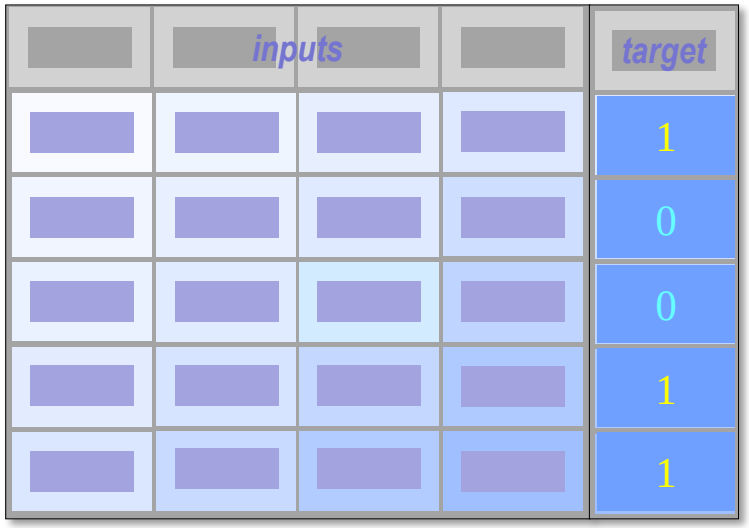
Assessment Statistics

Validation Data

Ratings depend on…

- target measurement (binary, continuous, and so on)
- prediction type (decisions, rankings, estimates)

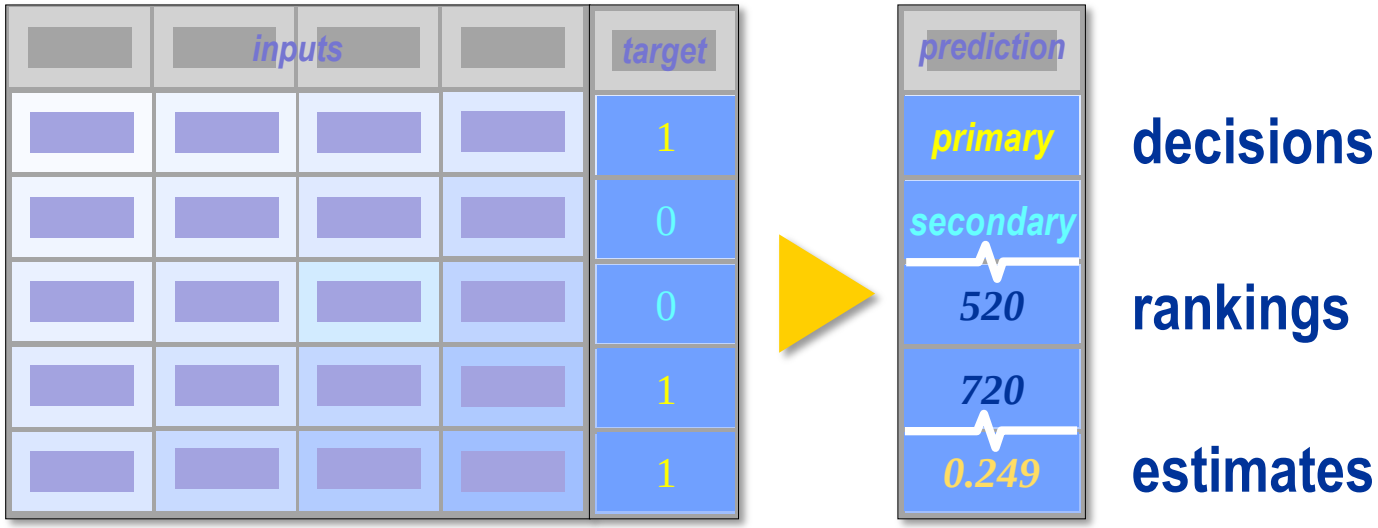
Binary Targets



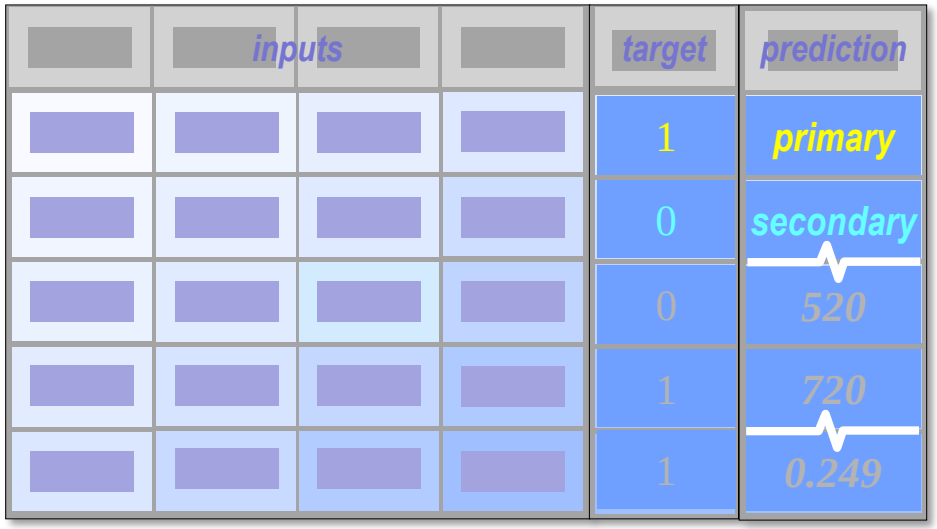
primary outcome

secondary outcome

Binary Target Predictions

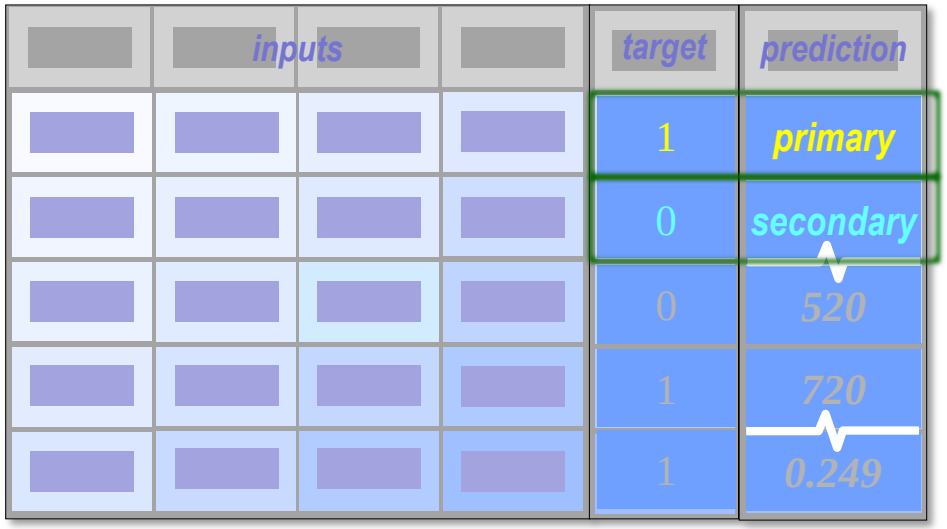


Decision Optimization



decisions

Decision Optimization: Accuracy

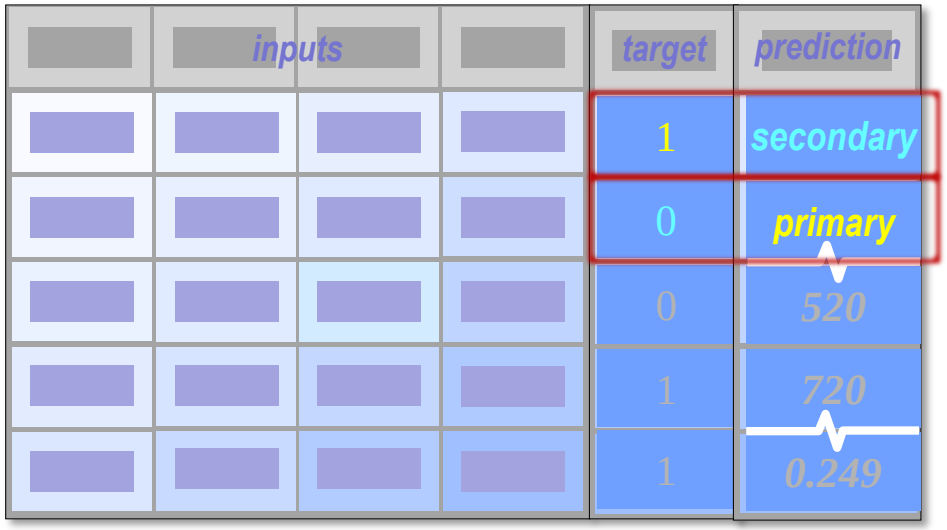


true positive

true negative

Maximize *accuracy*: agreement between outcome and prediction

Decision Optimization: Misclassification

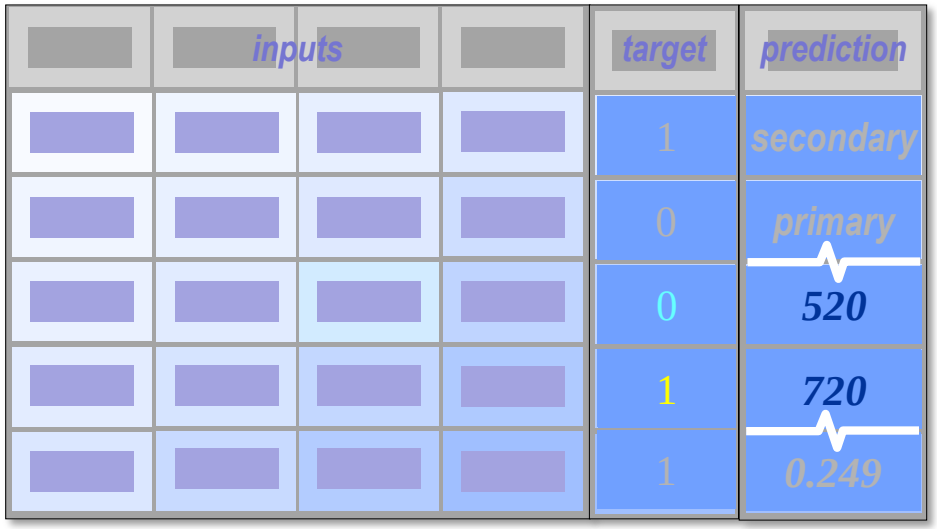


false negative

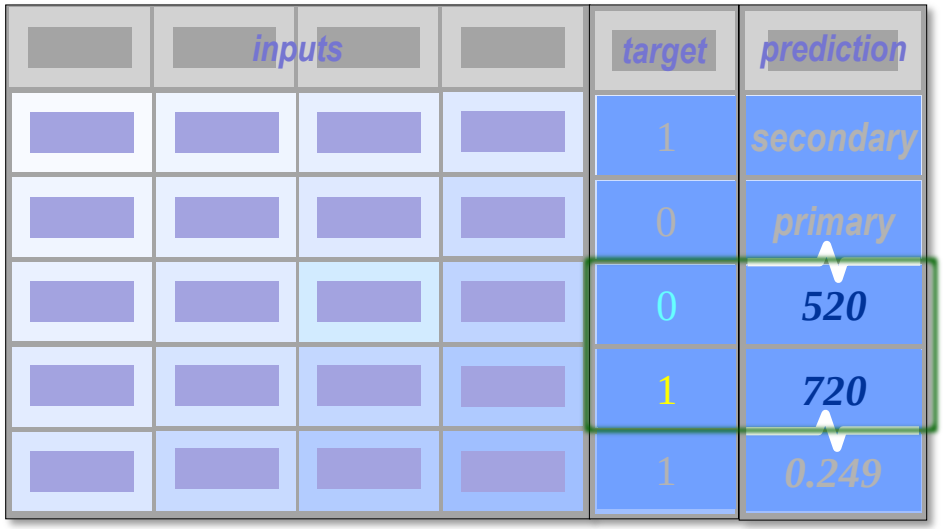
false positive

Minimize *misclassification*: disagreement between outcome and prediction

Ranking Optimization



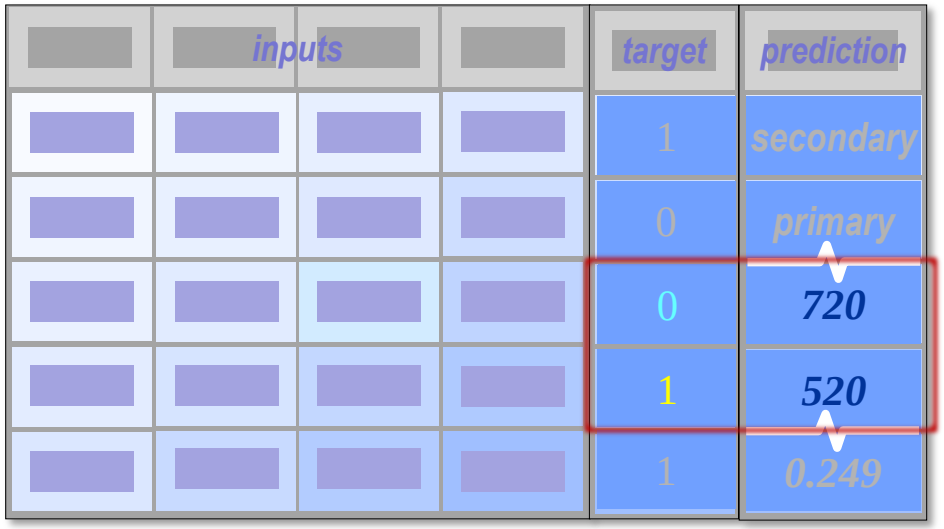
Ranking Optimization: Concordance



target=0→low score
target=1→high score

Maximize *concordance*: proper ordering of primary and secondary outcomes

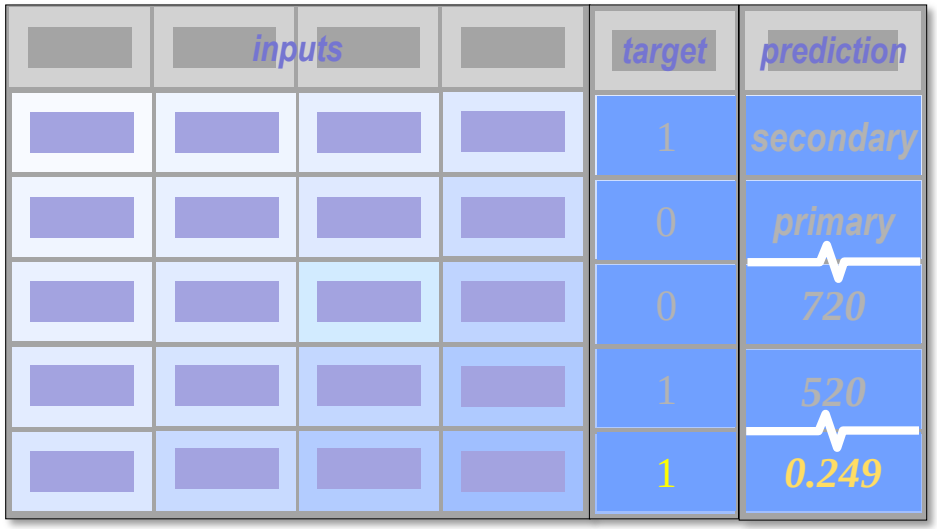
Ranking Optimization: Discordance



target=0→high score
target=1→low score

Minimize *discordance*: improper ordering of primary and secondary outcomes

Estimate Optimization



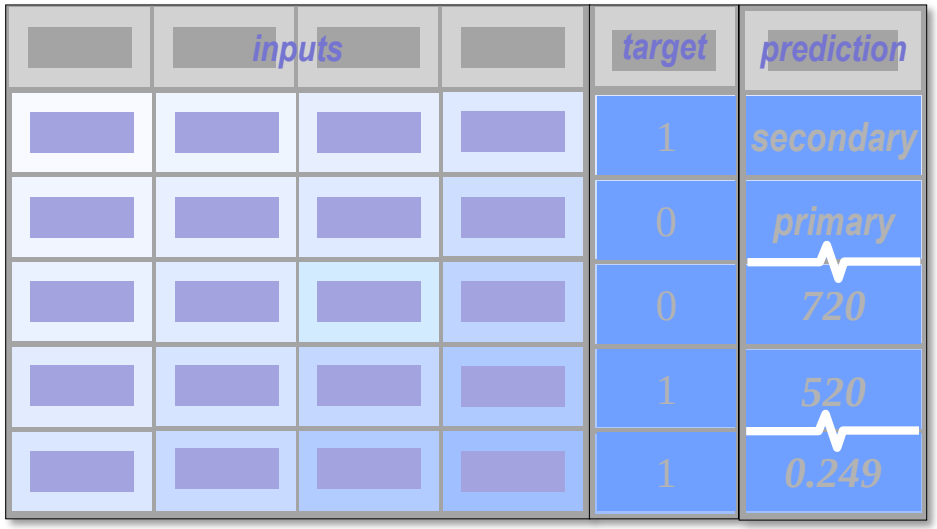
Estimate Optimization: Squared Error

	inputs			target	prediction
				1	secondary
				0	primary
				0	720
				1	520
				1	0.249

(target – estimate)^2

Minimize *squared error*:
squared difference between target and prediction

Complexity Optimization: Summary



decisions

accuracy / misclassification

rankings

concordance / discordance

estimates

squared error