HUDK 4050:
CORE METHODS
IN EDM

11/13/18 4:08 PM

In the news

Jiaying (Mandy) Li
LenaL
Juye Wang
Sam Thanapornsangsuth

EdSurge

OER is Growing at Religious Colleges, But Raises Unique Challenges

Analytics India MAGAZINE

Data Literacy Is The Biggest Challenge In Analytics Education, Says Qlik's Arun Balasubramanian

HEALTH IT ANALYTICS

Deep Learning for Medical Imaging Fares Poorly on External Data

Longview News-Journal

New NAFSA Data: International Students Contribute $39 Billion to the U.S. Economy

REUTERS

A wagging finger sticking out of your mobile phone is creepy. Why?

Events

Event	Date	URL
People centric approach to optimize Data Science, Commercial impact and Leadership	10:30am November 14	https://events.columbia.edu/cal/event/eventView.do?b=de&calPath=%2Fpublic%2Fcals%2FMainCal&guid=CAL-00bb9e24-655b8449-0165-5e0ea7e9-00001957events@columbia.edu&recurrenceId=
Data Science Behind Uber Eats Dispatch System	6:00pm November 15	https://www.eventbrite.com/e/data-science-behind-uber-eats-dispatch-system-tickets-51734366884
Computing for the Endless Frontier	2:30 November 27	https://events.columbia.edu/cal/event/eventView.do?b=de&calPath=%2Fpublic%2Fcals%2FMainCal&guid=CAL-00bb9e25-66e323b3-0166-e4d9e20e-00002992events@columbia.edu&recurrenceId=
Cross-device User Clustering at Adobe	5:30 November 29	https://events.columbia.edu/cal/event/eventView.do?b=de&calPath=%2Fpublic%2Fcals%2FMainCal&guid=CAL-00bb9e28-655b8cee-0165-5dd5c72b-00001287events@columbia.edu&recurrenceId=
Machine Learning Innovation Summit	December 12-13	https://www.theinnovationenterprise.com/summits/machine-learning-innovation-summit-new-york-2018
Columbia Data Science Day	April 3, 2019	https://www.eventbrite.com/e/data-science-day-columbia-university-2019-tickets-49454358317

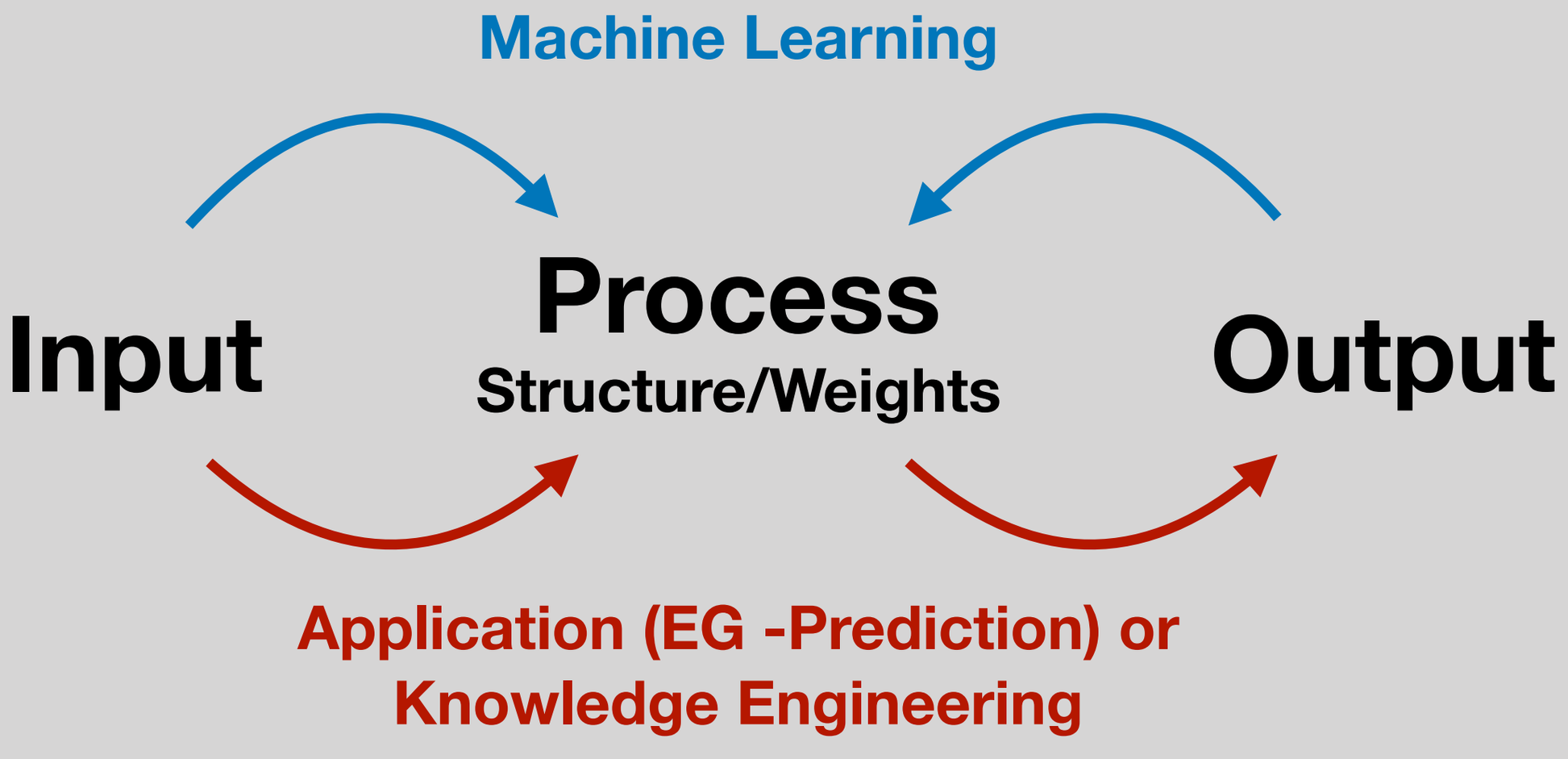
Prediction

Machine Learning

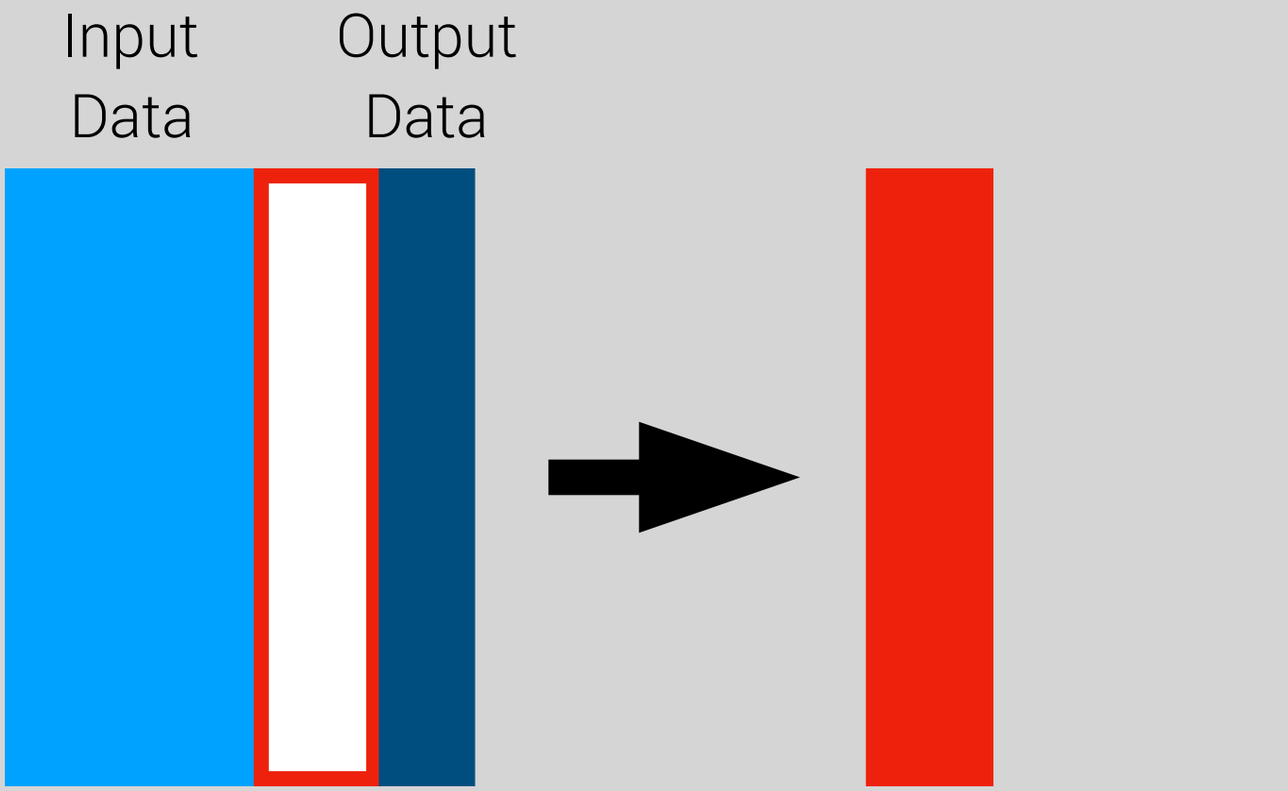
- Samuel built a computer program that could play checkers
- Recognized when it made a mistake and avoided that mistake again ("learning" through prediction)
- Built a tree of all possible moves for a given board
- Maximized a function that described the probability of winning
- Within three weeks it beat Samuel
- In 1956 it beat the world checkers champion

Arthur Samuel, 1952

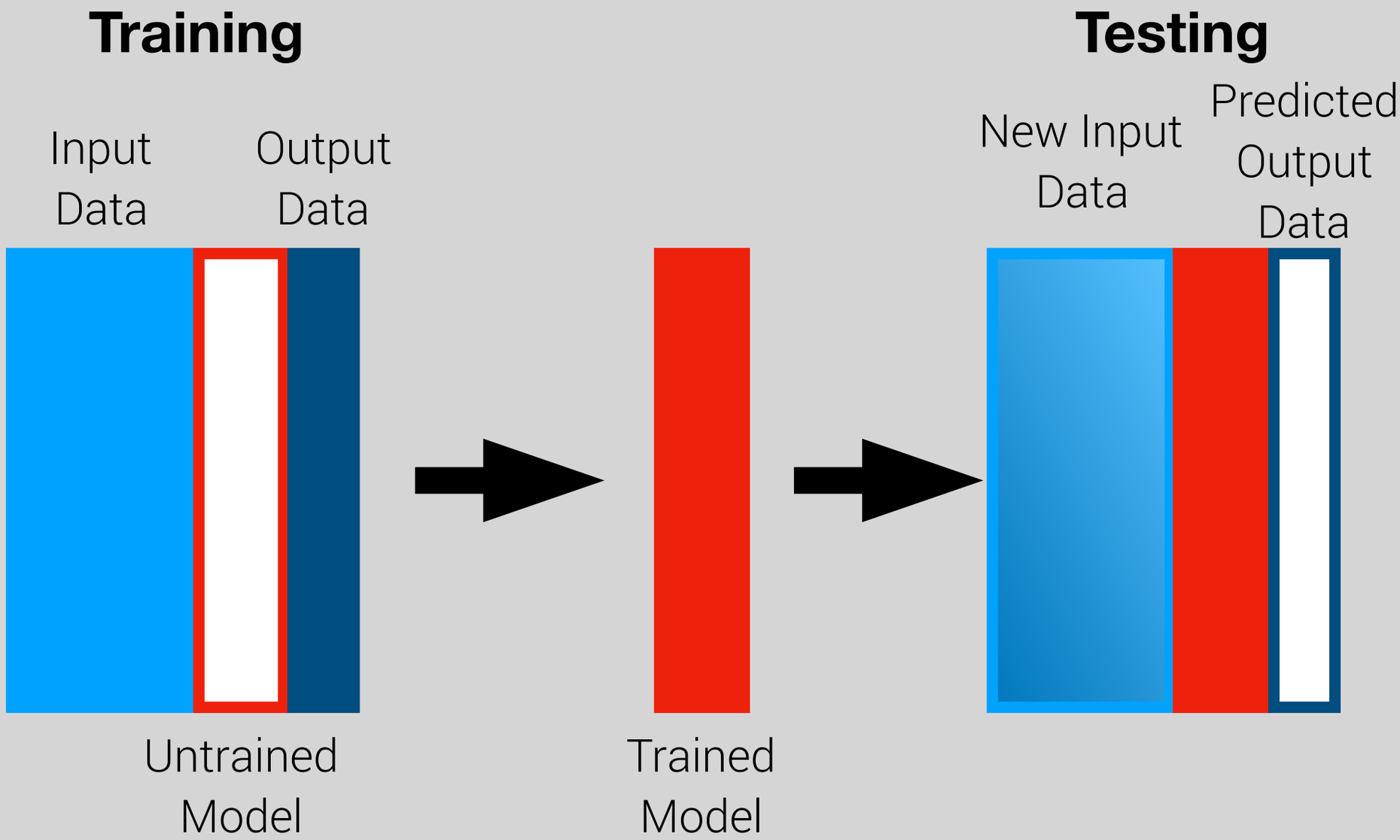
Machine Learning



Educational Statistics



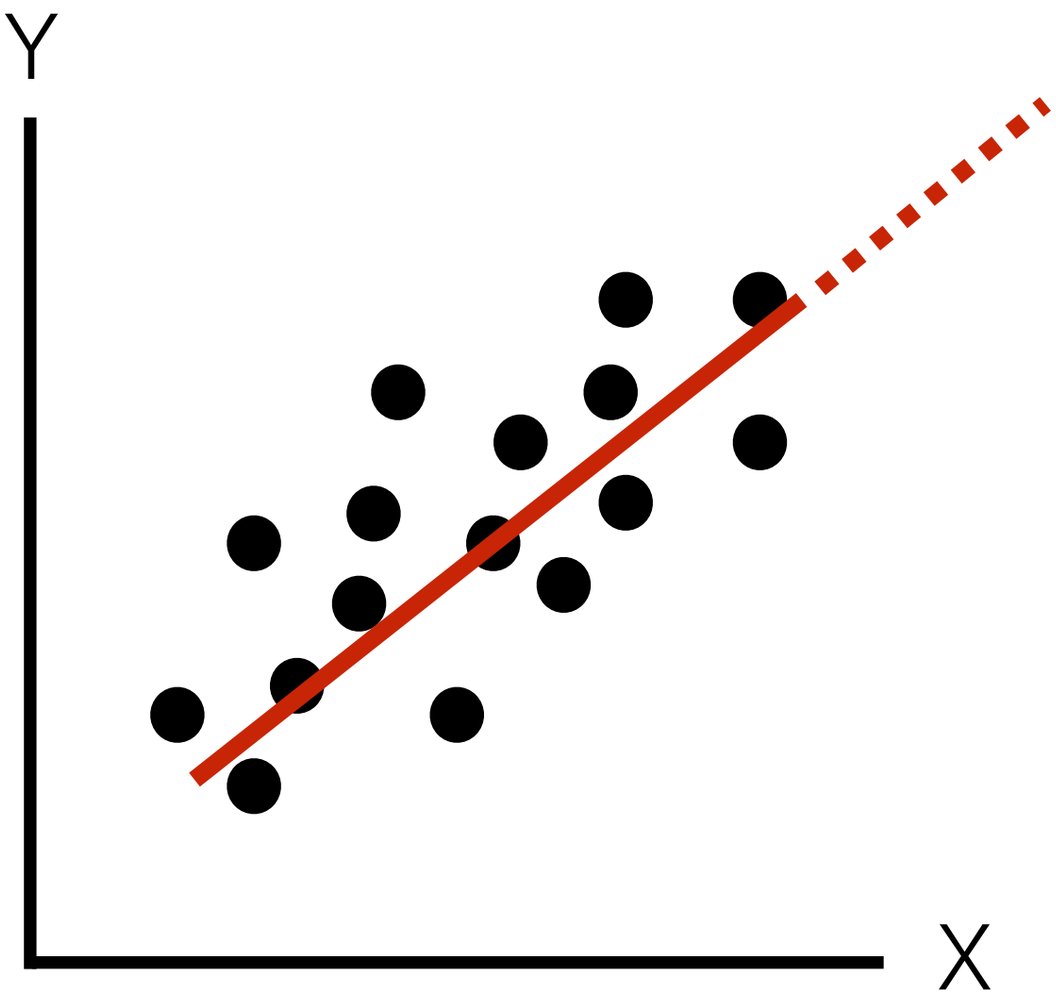
Machine Learning



Prediction

- Cuts to the ❤️ of the difference between machine learning and ed statistics

- Characterize data vs. predict the future



Terminology

Supervised Learning: Techniques used to learn the relationship between independent attributes and a designated dependent attribute (the label). (Have labelled data available that the machine can learn from)

For example: Have images labelled as dog, cat, etc, machine must learn the labels

Unsupervised Learning: Learning techniques that group instances without a pre-specified dependent attribute.

For example: Clustering algorithms

Terminology

Classification: Mapping an unlabeled instance to a discrete class by a classifier

Example: Identify a student as likely to drop out or not based on demographic data

Regression (*as a form of classification*): Mapping from an unlabeled instance to a value within a continuous range

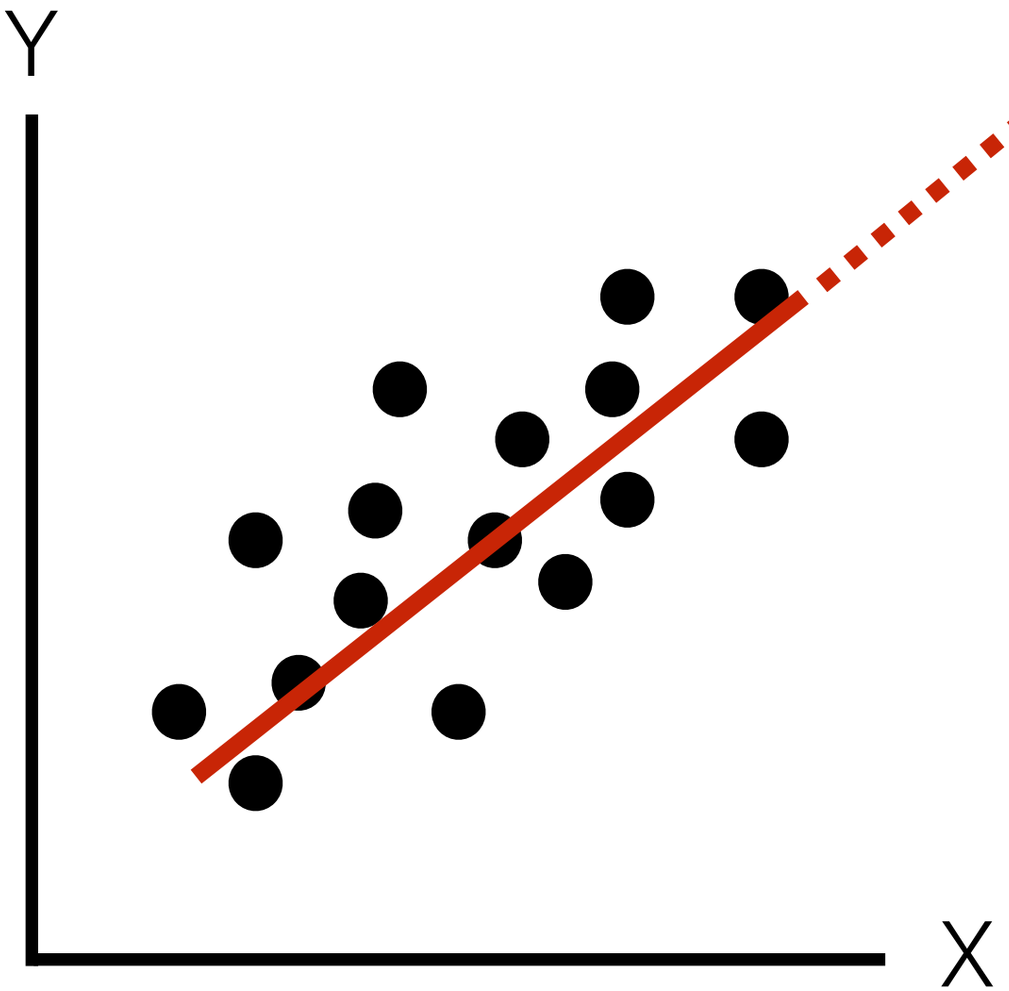
Example: Identify a student as having a math test score of 70 based on online assignment performance

Training Sets: Either supplied by a previous stage of the knowledge discovery process or from some external source

Regression

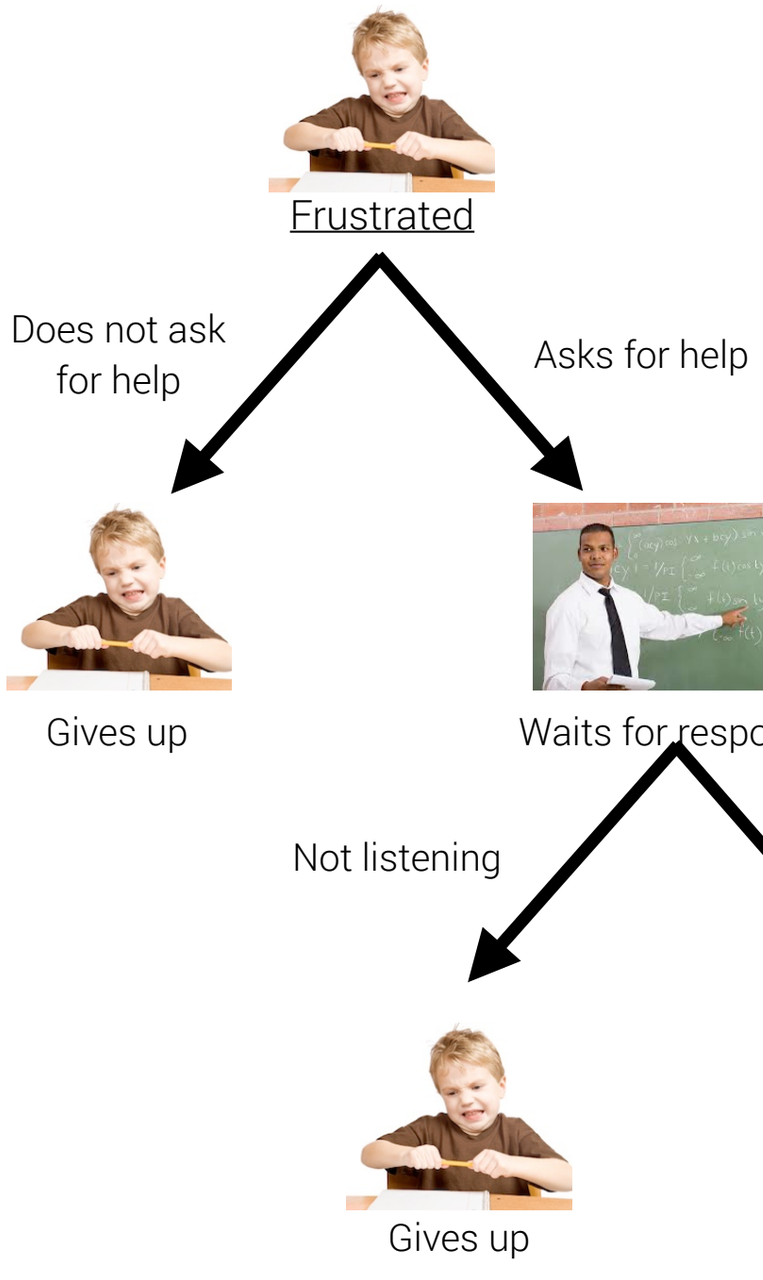
- In Ed Stat = OLS Regression/Logistic Regression (characterize)

- In ML = Mapping from unlabeled instances to a value within a continuous range (future)



Classification Tree

- Decision tree
- Map observations (branches) onto classes (leaves)
- Tree describes the data but can be used classification
- EG: student states = leaves, student actions = branches



Machine Learning

Input

Does not ask for help

Asks for help

Not listening

Listening

Process

Structure/Weights

Frustrated

Does not ask for help

Asks for help

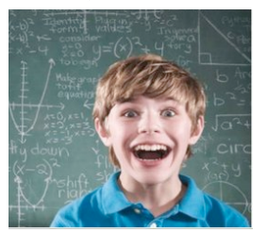
Gives up

Waits for response

Not listening

Listening

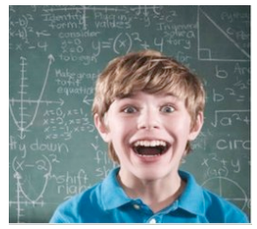
Gives up



Understanding

Output

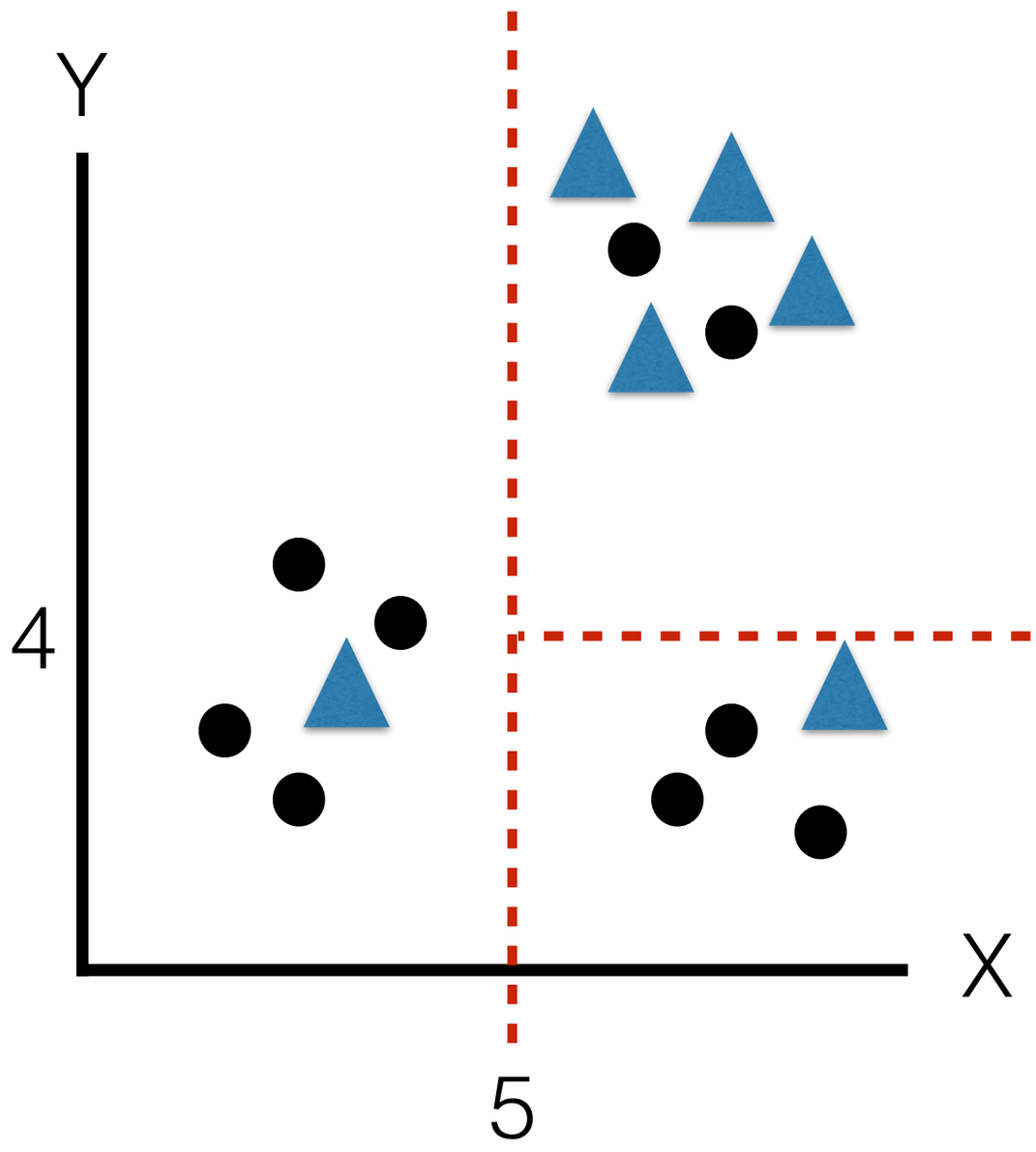
Gives up

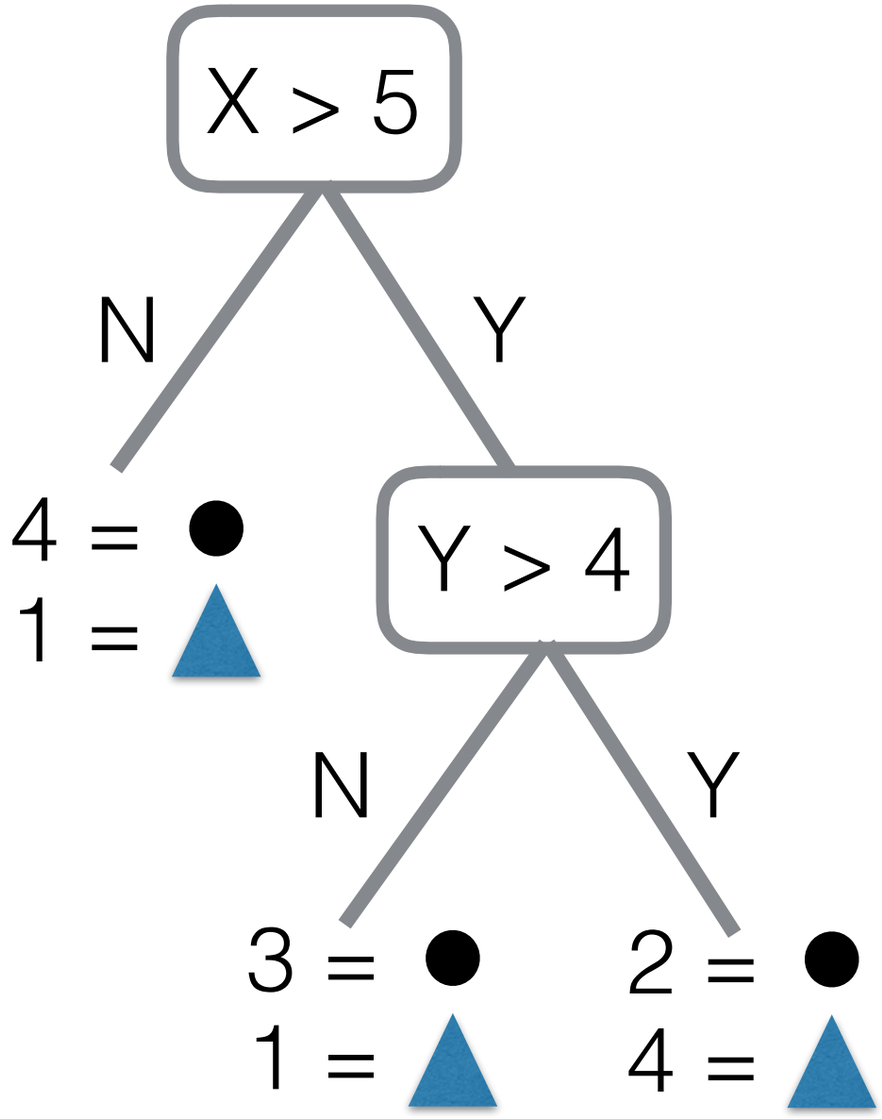


Understanding

Binary Classification Tree

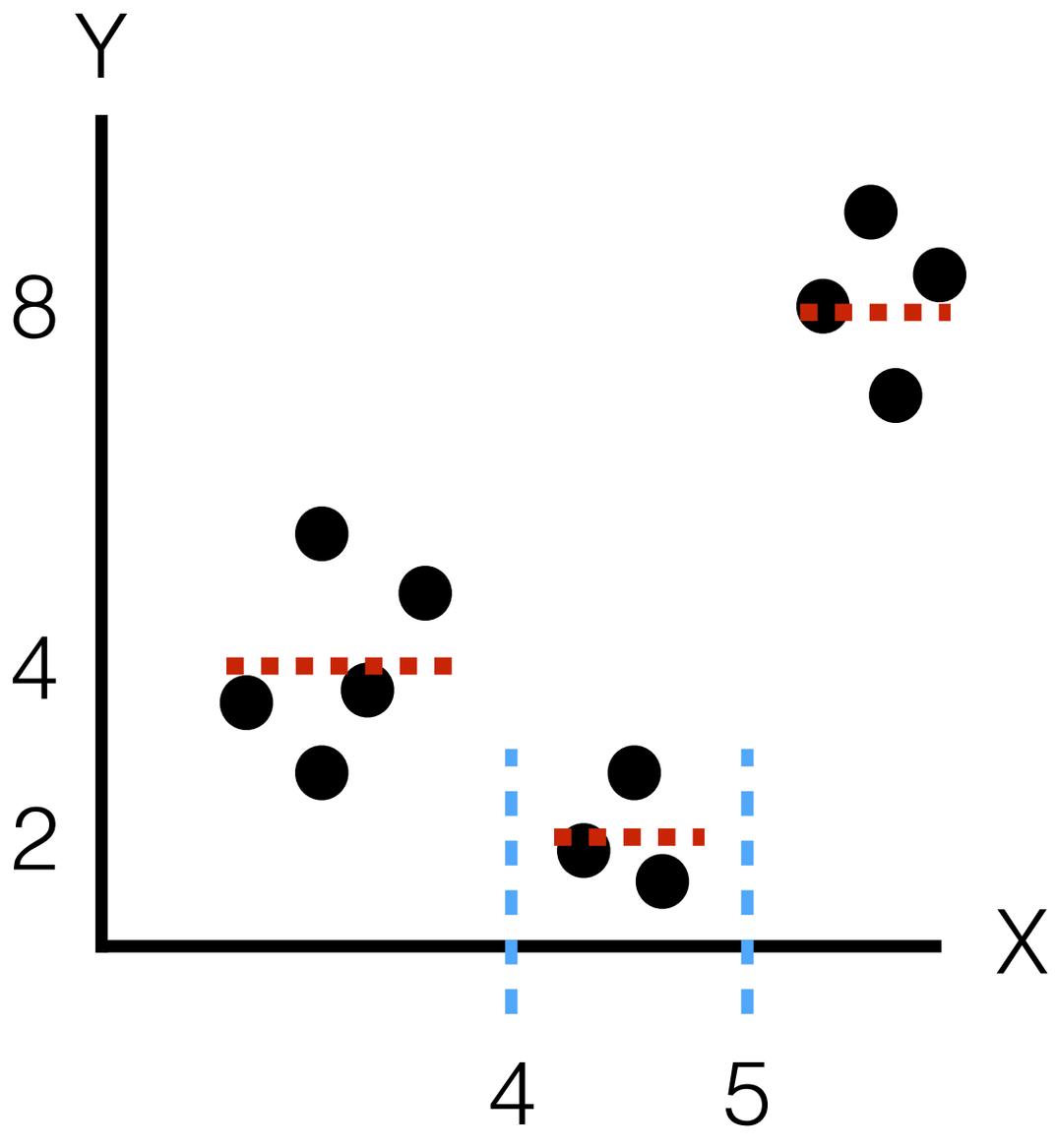
* Minimize the error

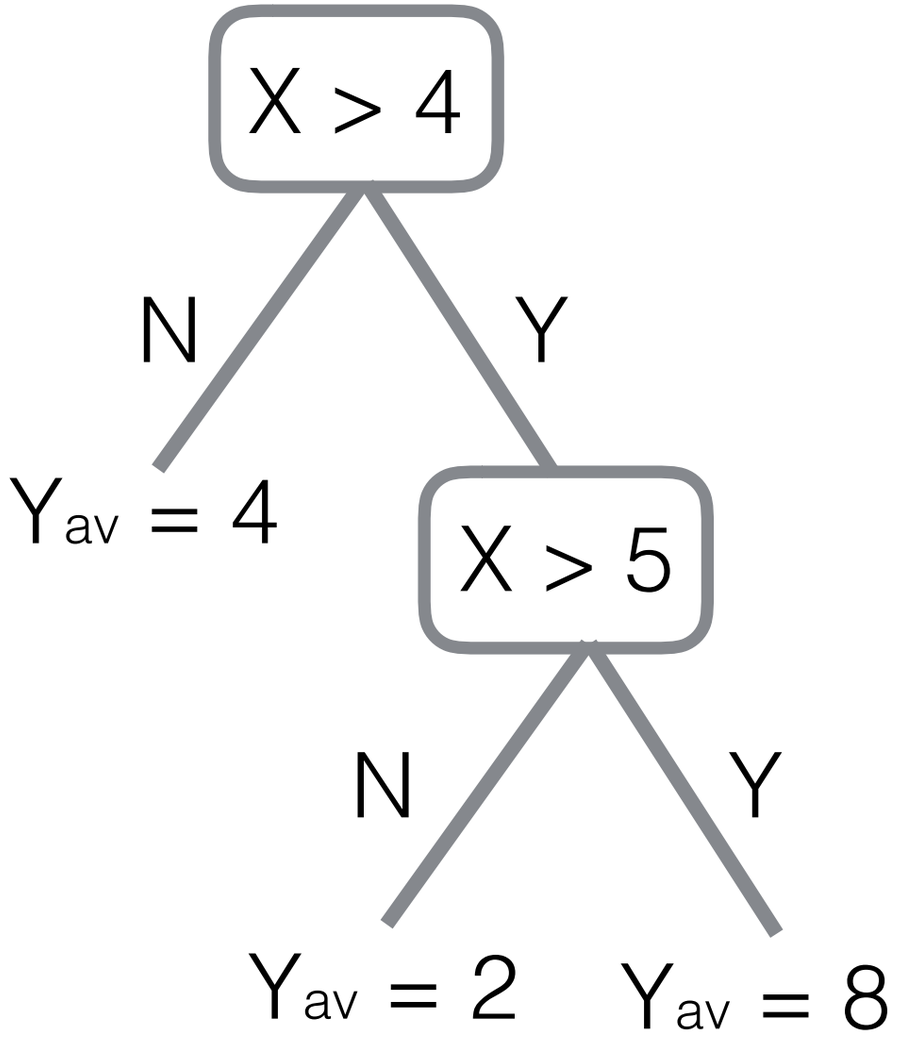




Binary Regression Tree

* Minimize the error



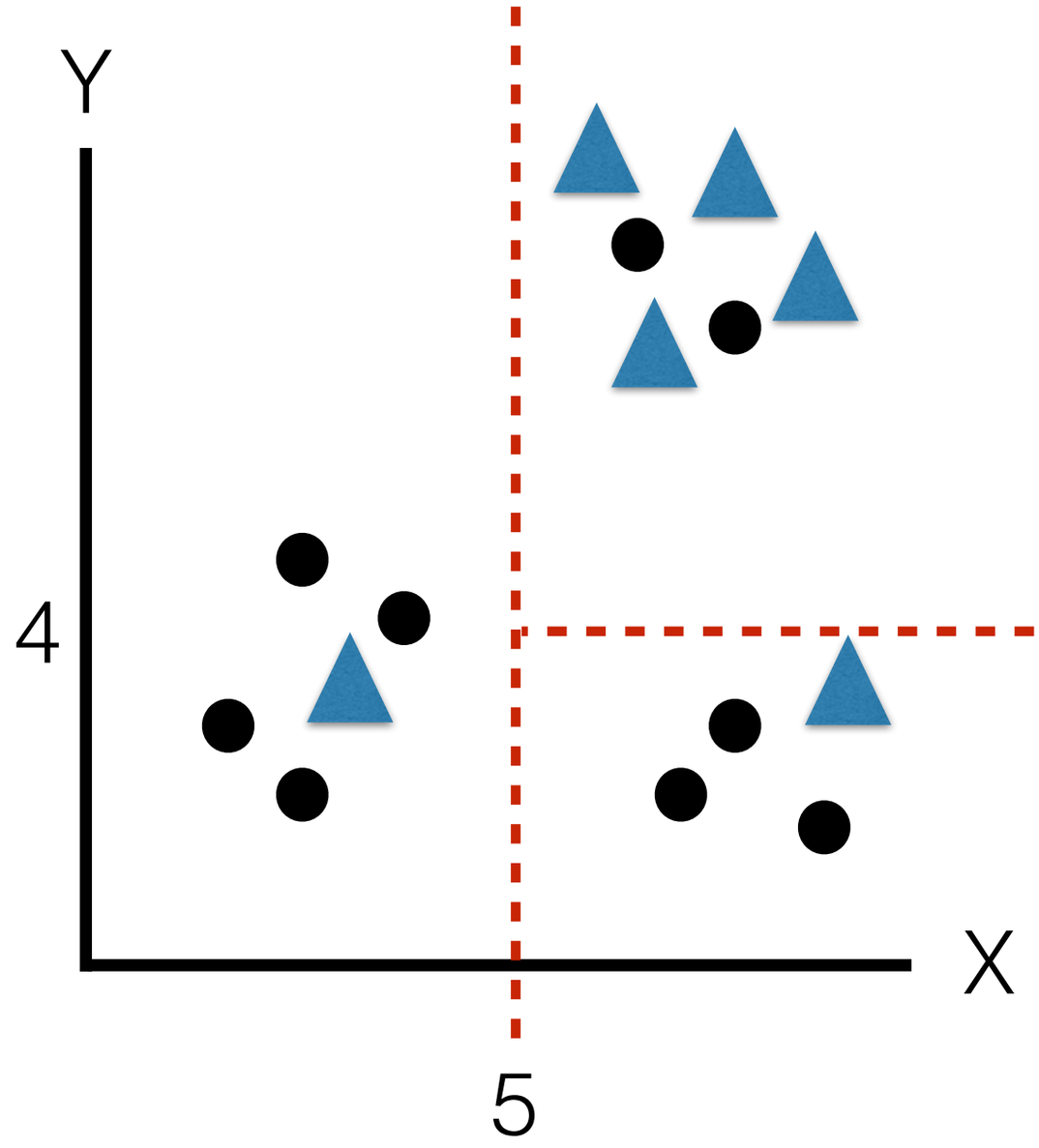


CART Trees

- Leo Breiman invented in the 1970s
- Non-parametric model
- Designed to deal with data that has too many interaction effects
- Trademarked CART (classification & regression trees) so is called rpart (recursive partitioning and regression trees)

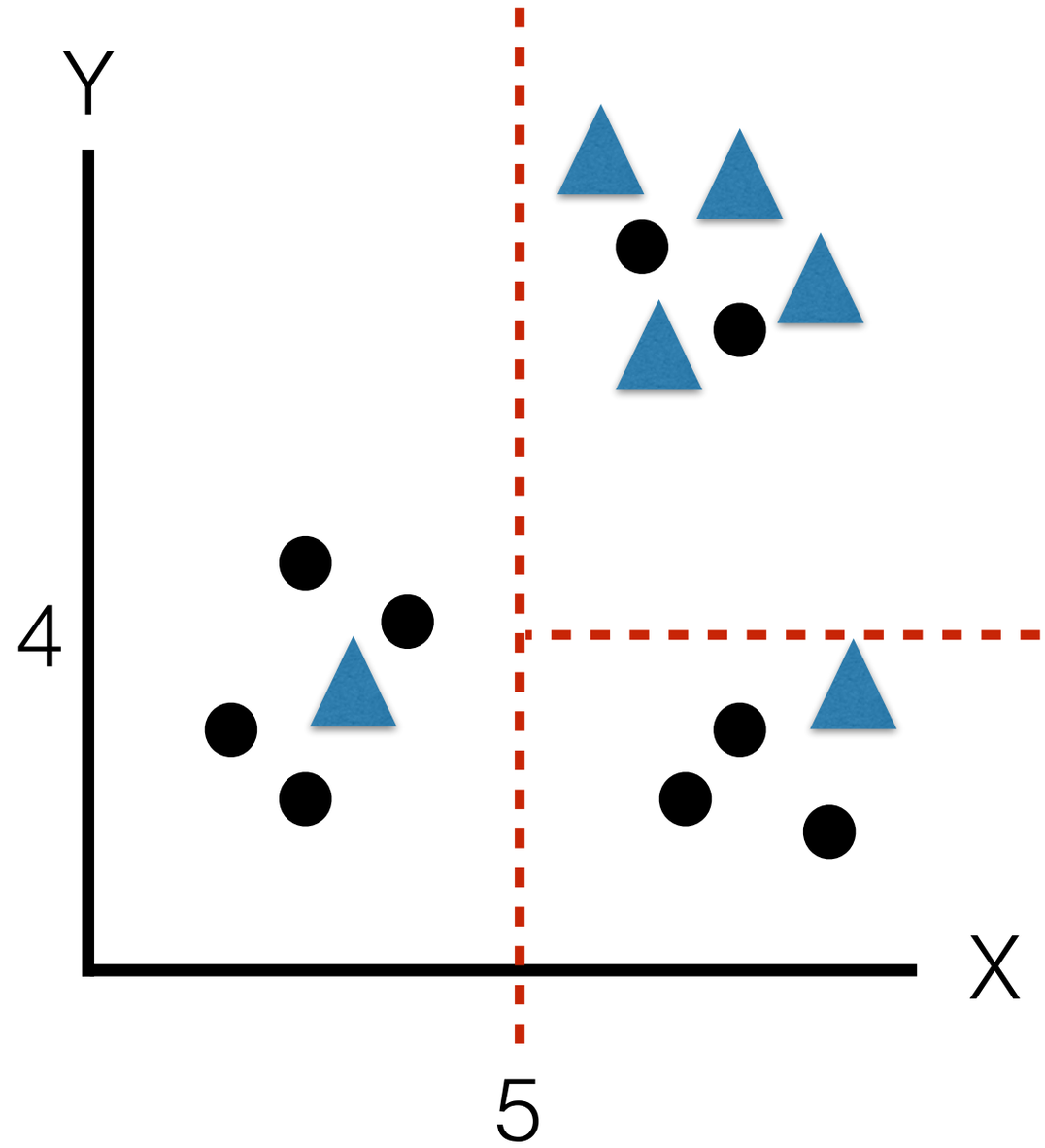
RPART

- Recursive (splits the leaves until you tell it to stop)
- At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the dependent (predicted) variable



Heterogeneity

- Impurity/homogeneity
- leaf has only 1 class, impurity = 0
- Entropy (information gain)
- Gini index
- Classification error



RPART

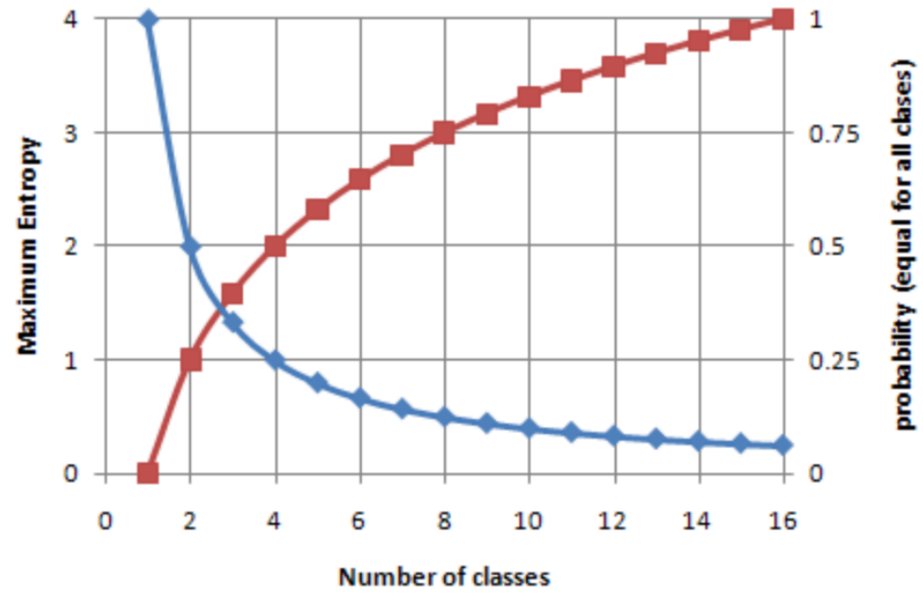
- `parms`

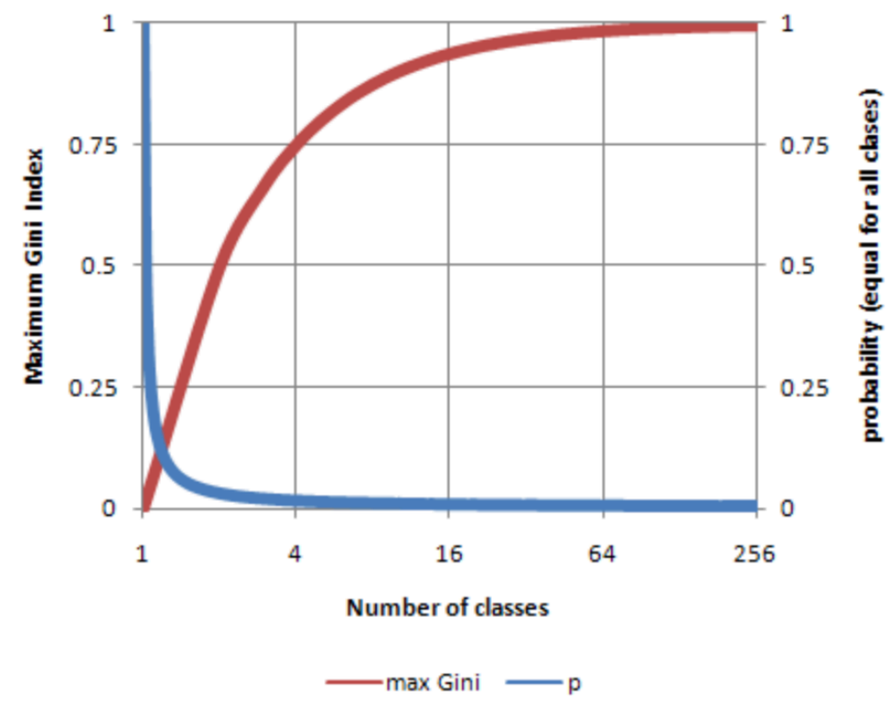
- Entropy (information gain)

$$Entropy = \sum_j -p_j log_2 p_j$$

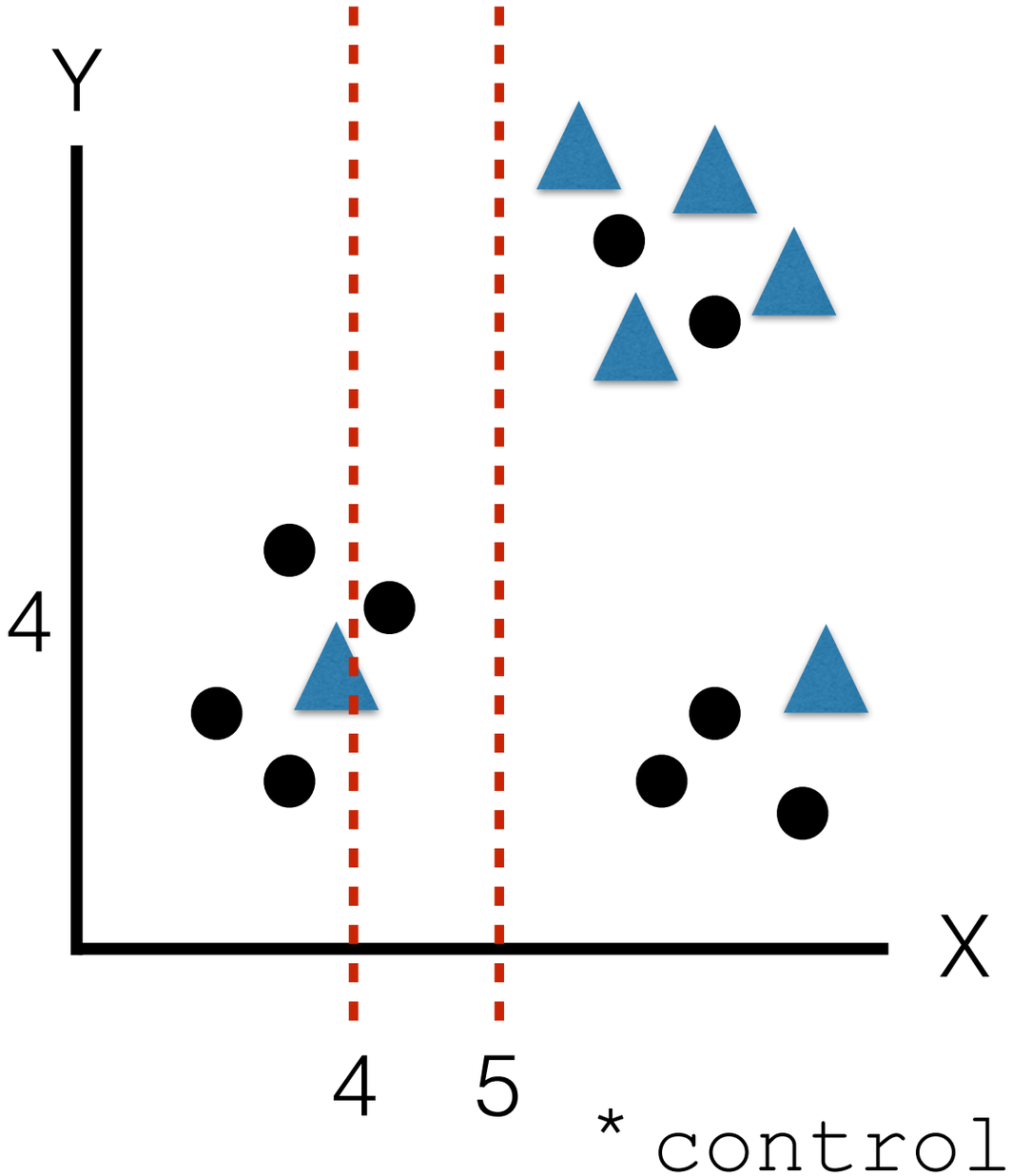
- Gini index

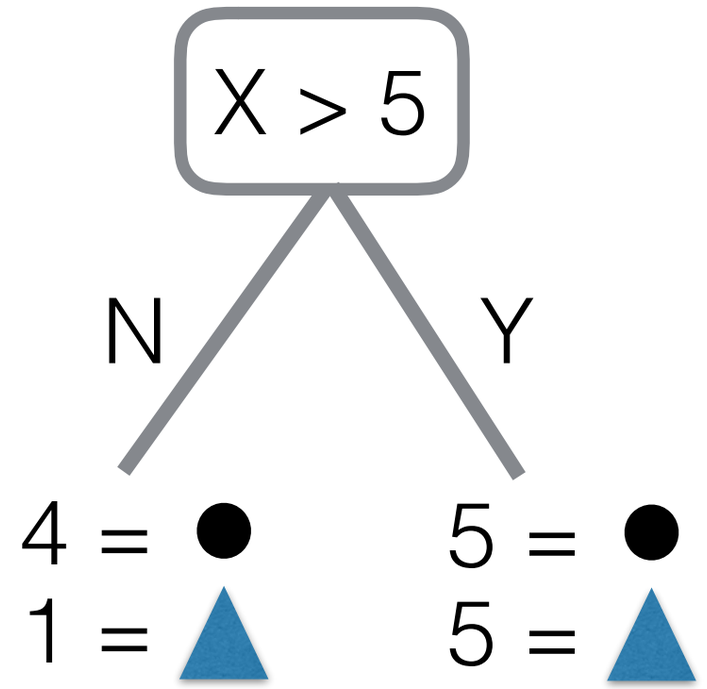
$$Gini\ Index = 1 - \sum_j p_j^2$$



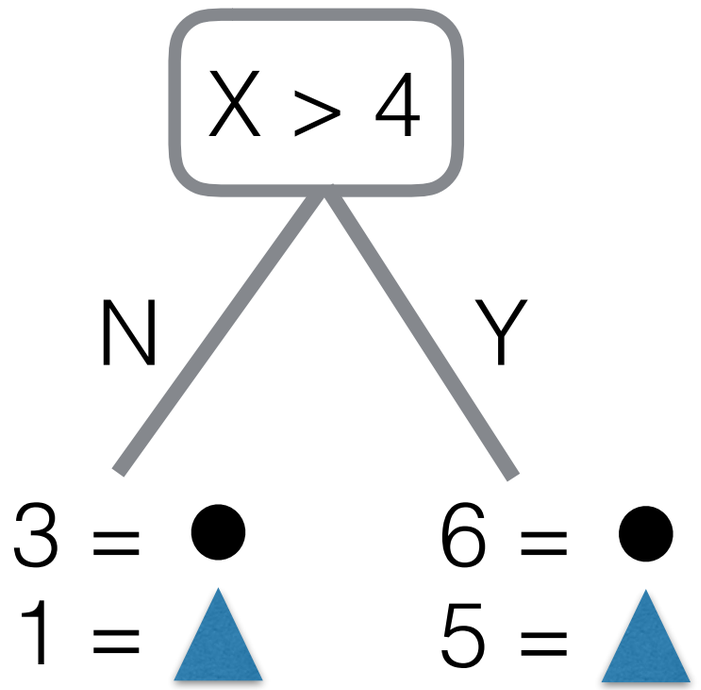


RPART





$ENT_5 = -0.8*\log_2(0.8) + -0.5*\log_2(0.5) = 0.75$



$ENT_4 = -0.75*\log_2(0.75) + -0.55*\log_2(0.55) = 0.76$

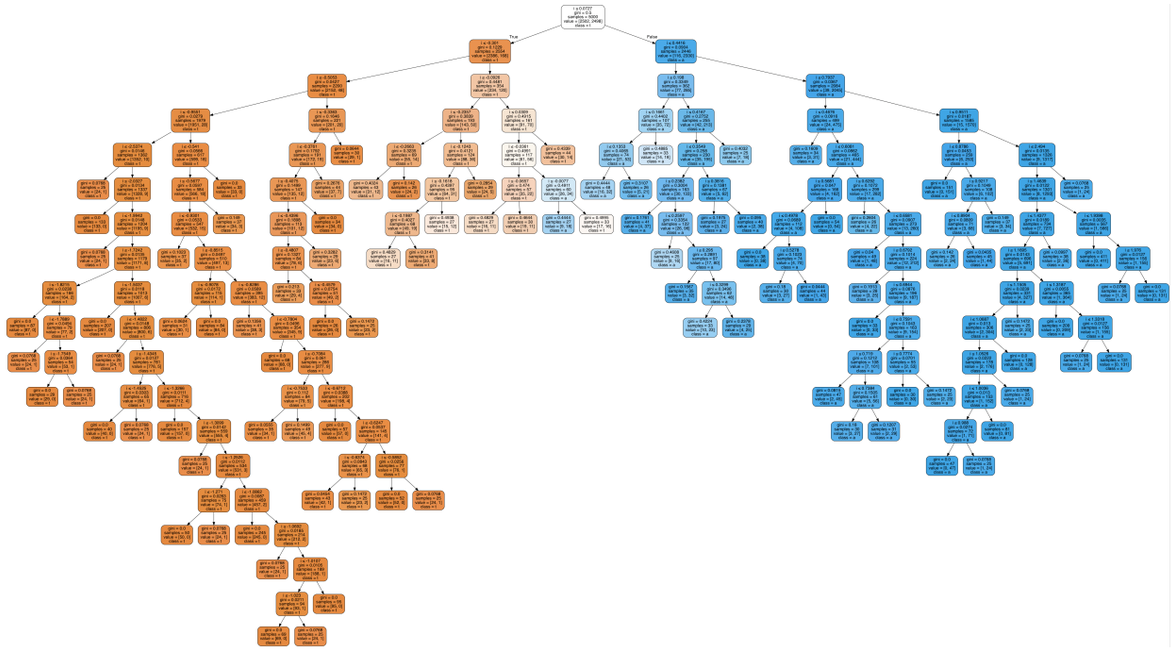
RPART

- Tree chooses the optimal fit at each leaf - NOT the overall best fit for the data

- Therefore, there is a danger of overfitting the tree

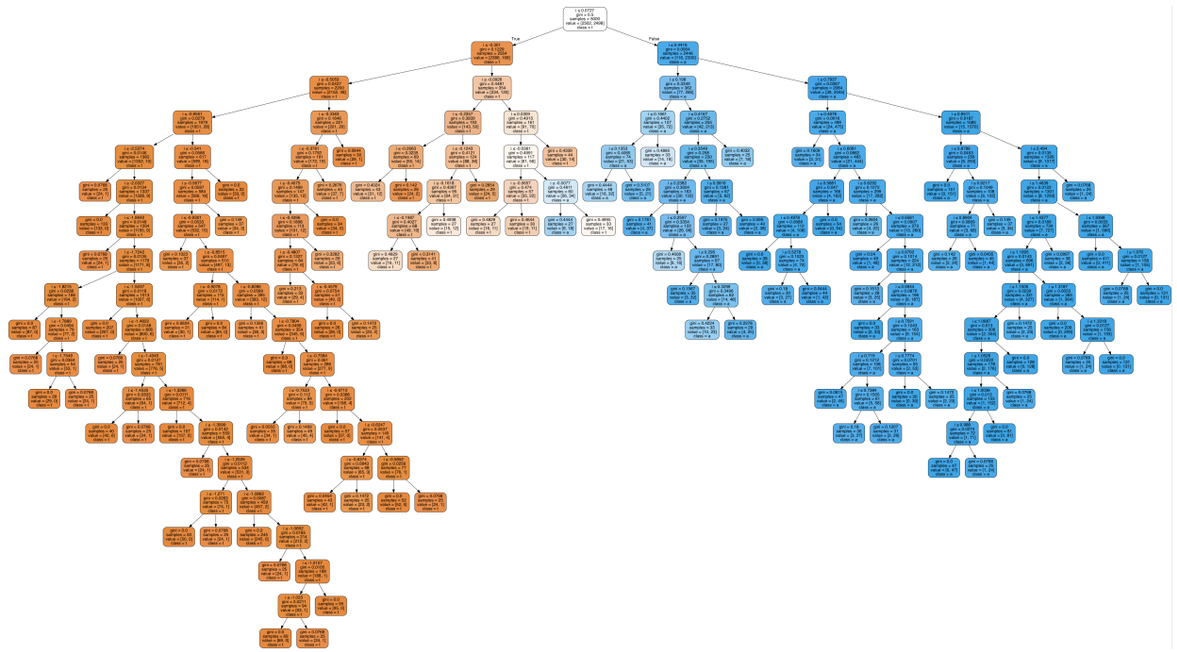
- Tree is too specific to training data to be able to predict new data

- Therefore: stop the tree at a certain number of nodes OR prune



RPART

- RPART prunes



- Uses a cost function:

Number of leaves

$$C_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

Misclassified instances

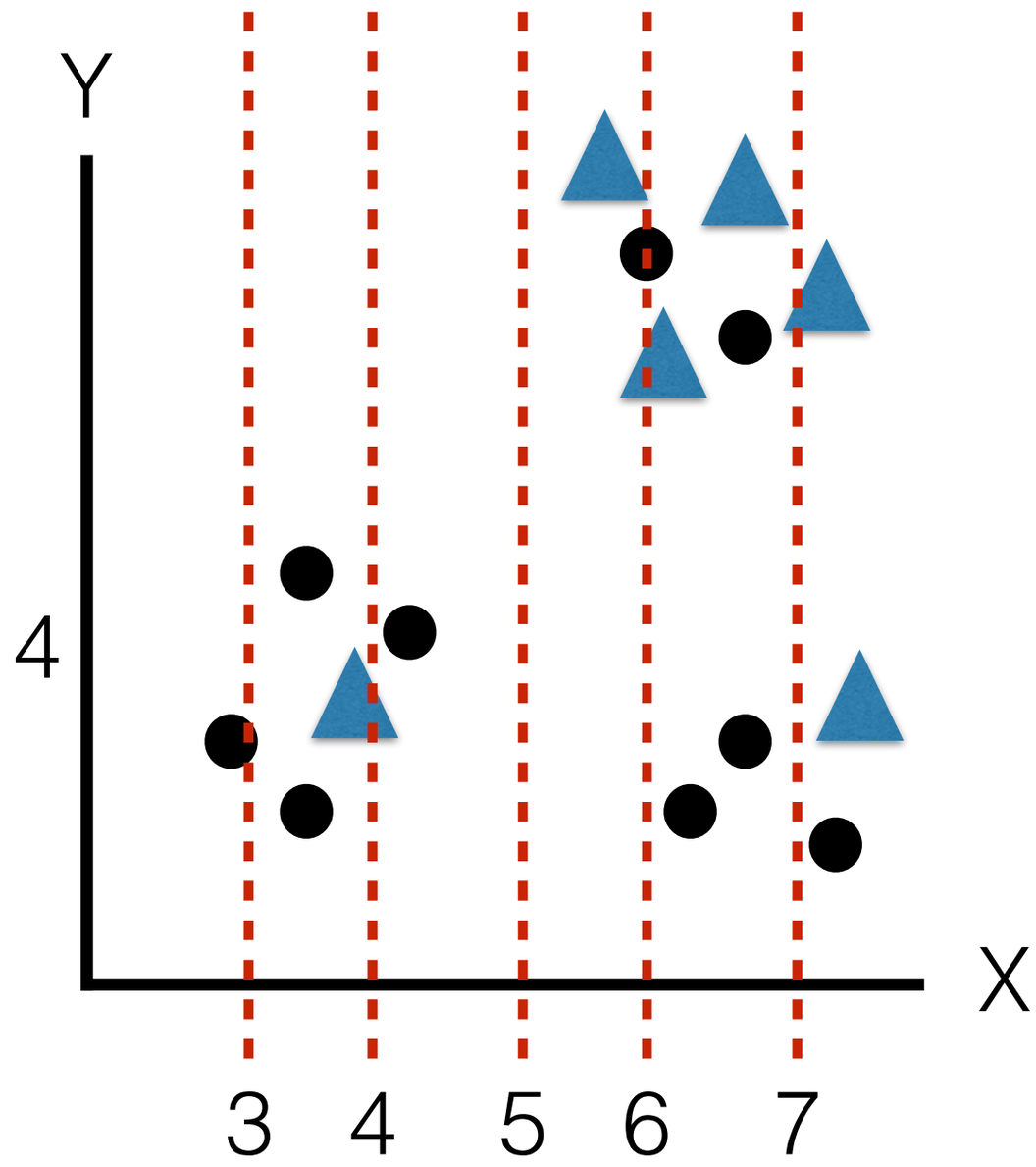
RPART

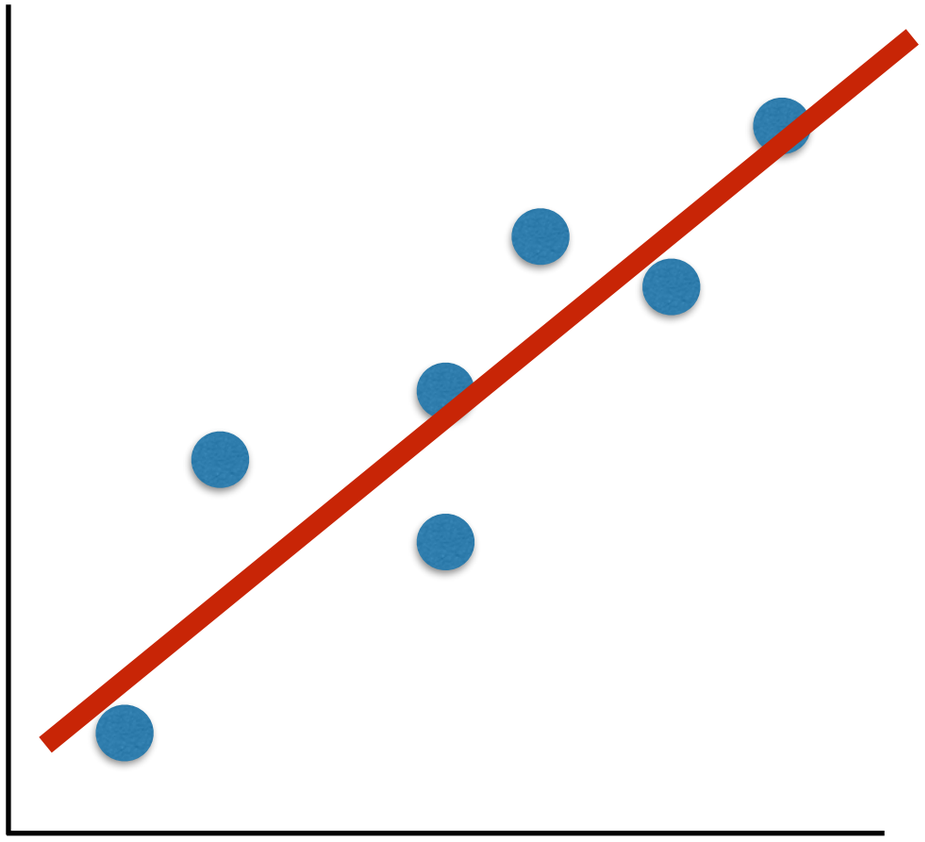
<u>Gotchas</u>

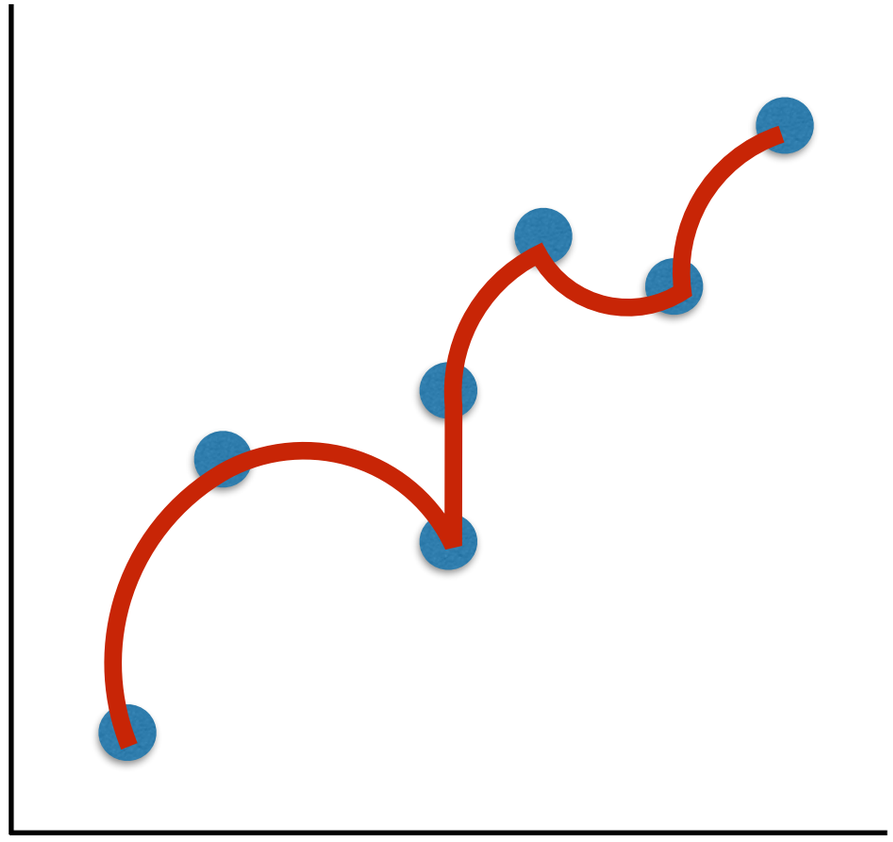
- Overfitting
- Local overfitting
- Sensitive to test data
- Selection bias toward covariates with many possible splits

PARTY

- “part(y)itioning” 😎
- Conditional Inference Tree
- Look at correlation between X and shape and Y and shape
- Statistically test H_0: there is no relationship
- Choose the variable with the highest correlation
- Split on that variable
- Stop when H_0 cannot be rejected







Which is more “accurate”?

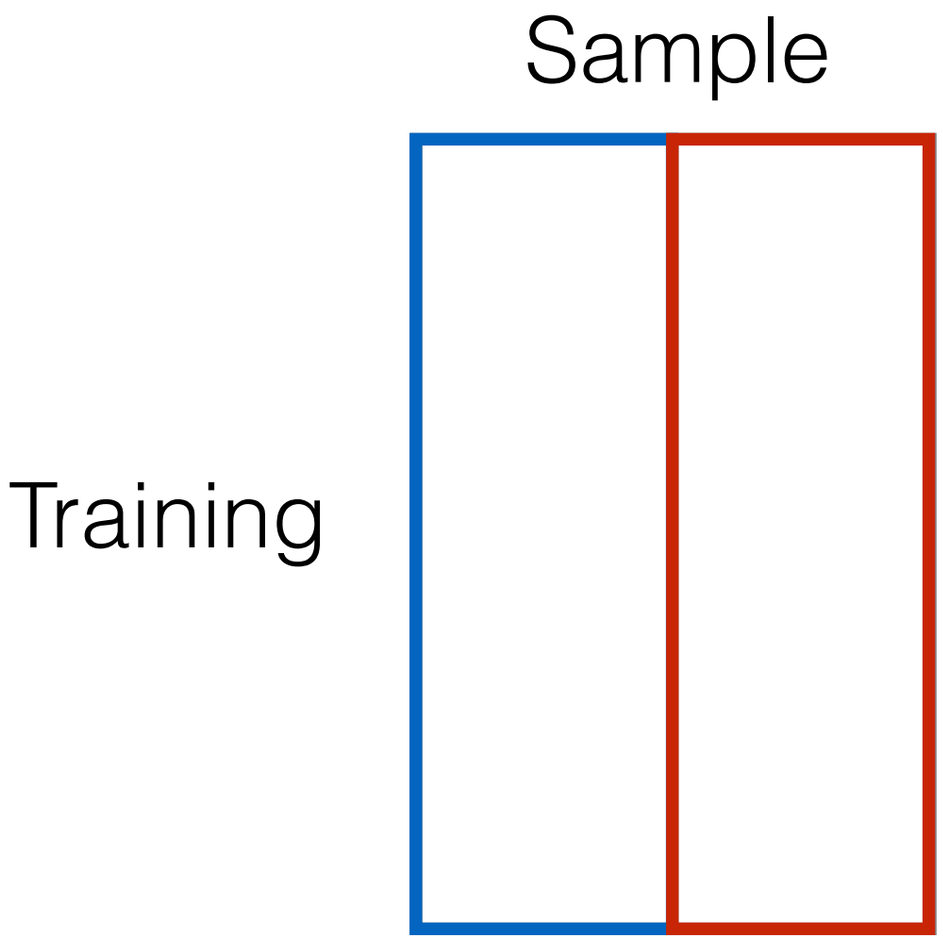
Which is more “useful”?

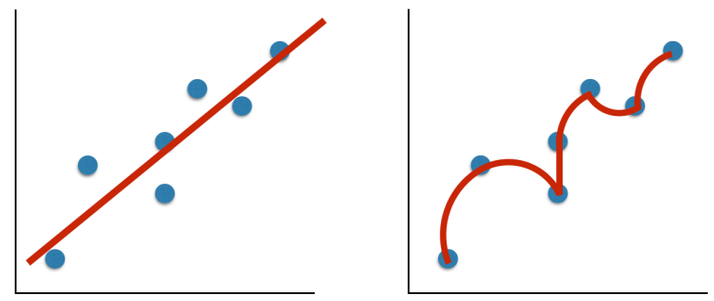
How can we tell?

Cross Validation

- Estimate how accurately a predictive model will perform in practice
- Give an insight on how the model will generalize to an independent dataset

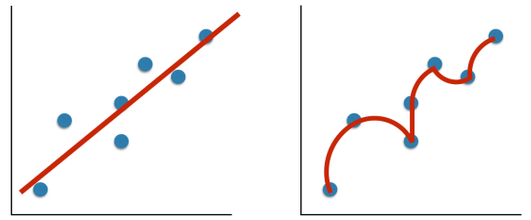
Hold-out Validation



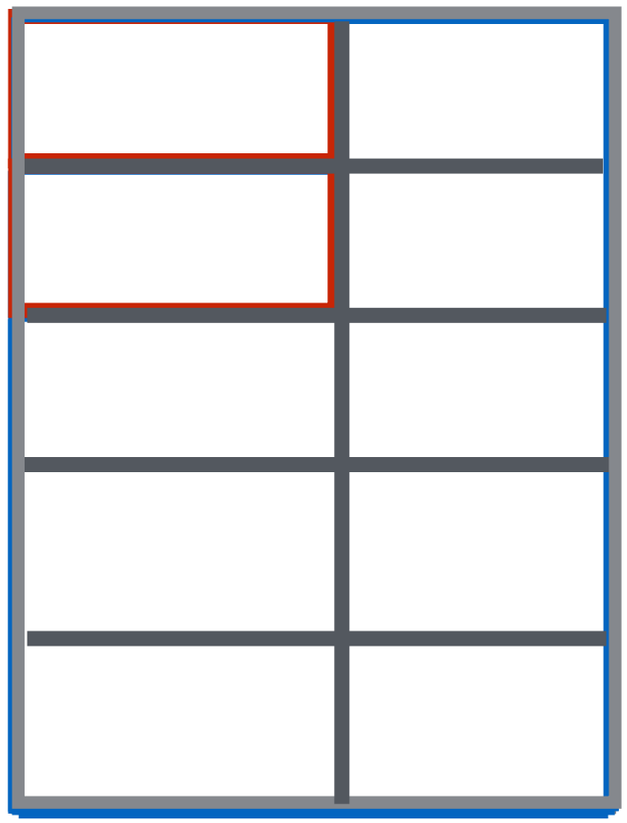


Problem: very dependent on which data are in each group

K-Fold Cross Validation



Sample



Test 1	Training 1	5	2
Test 2	Training 2	4	2
Test 3	Training 3	3	1
Test 4	Training 4	5	4
Test 5	Training 5	4	2
		4.2	2.2

Calculate how accurate we are in each “fold” and average the answer