

HUDK 4050: CORE METHODS IN EDM

In the news

**The One And Only Lesson
To Be Learned From NAEP
Scores**

Forbes



**Pearson Acquires Ed Tech
Company Lumerit**

Deal enhances Pearson's online learning and employability capabilities



The EdTech Genome Project

edScoop



**Can Ed-Tech Democratize Education In Asia
Pacific?**

Events

Title	Date - Time	Location
<u>Data Science Pioneers: Conquering the Next Frontier</u>	11/6 - 6:30pm	Dataiku 920 Broadway
<u>AWS Machine Learning Day</u>	11/6 - 12:00pm	Online
<u>The Color of Surveillance</u>	11/7	Georgetown University
<u>All Tech is Human: NYC</u>	11/9	ThoughtWorks
<u>Xavier Ochoa: Multimodal Analytics</u>	11/12 - 12:00pm	NYU
F1 Student Employment Workshop	11/12 - 2:00pm	HM 138
HUD Happy Hour	11/26 - 5:00pm	E's Bar
<u>Science Communication Workshop</u>	11/20 - 9:30am	Low Library
<u>Citizens and Technology Summit</u>	11/25	Ford Foundation

Prediction

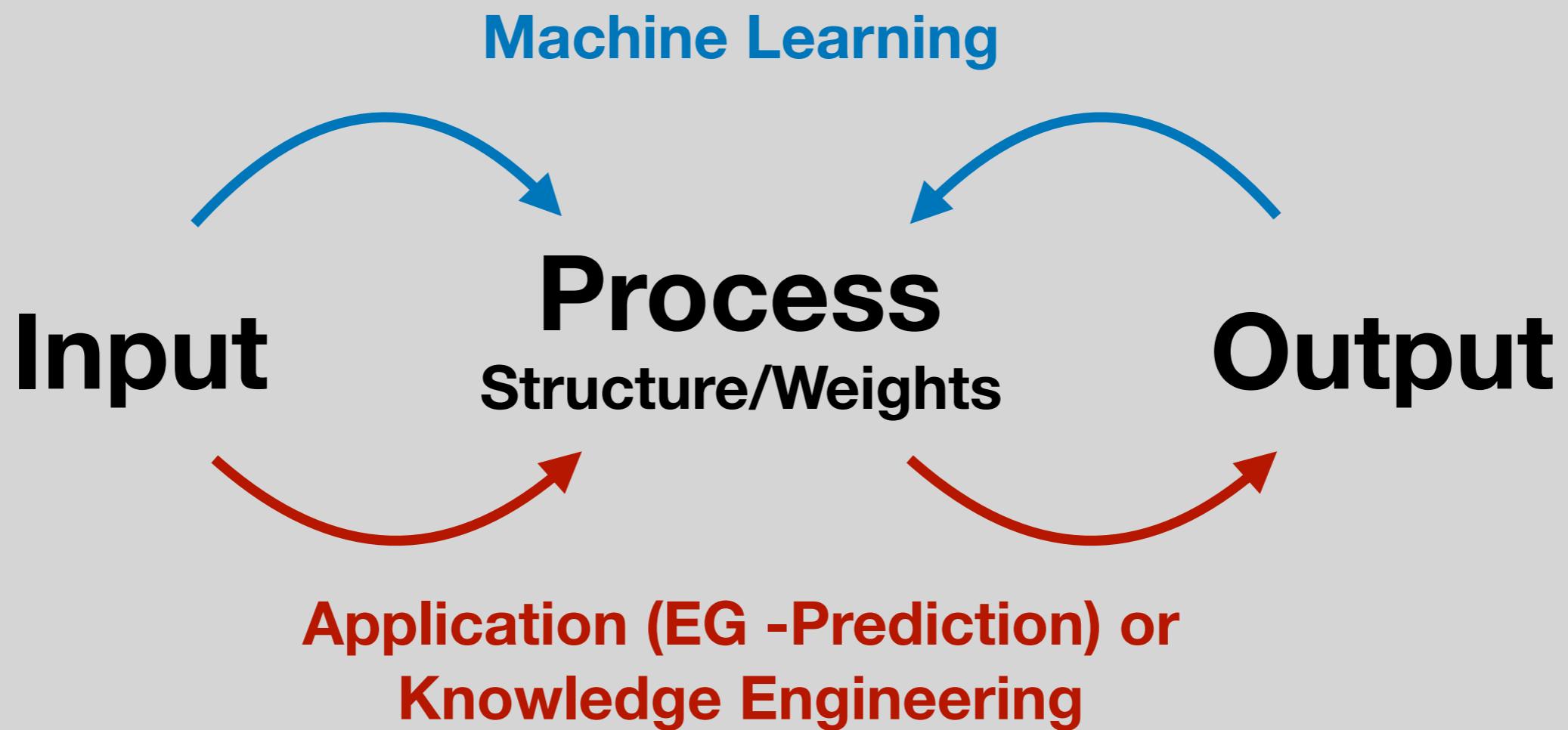
Machine Learning

- 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

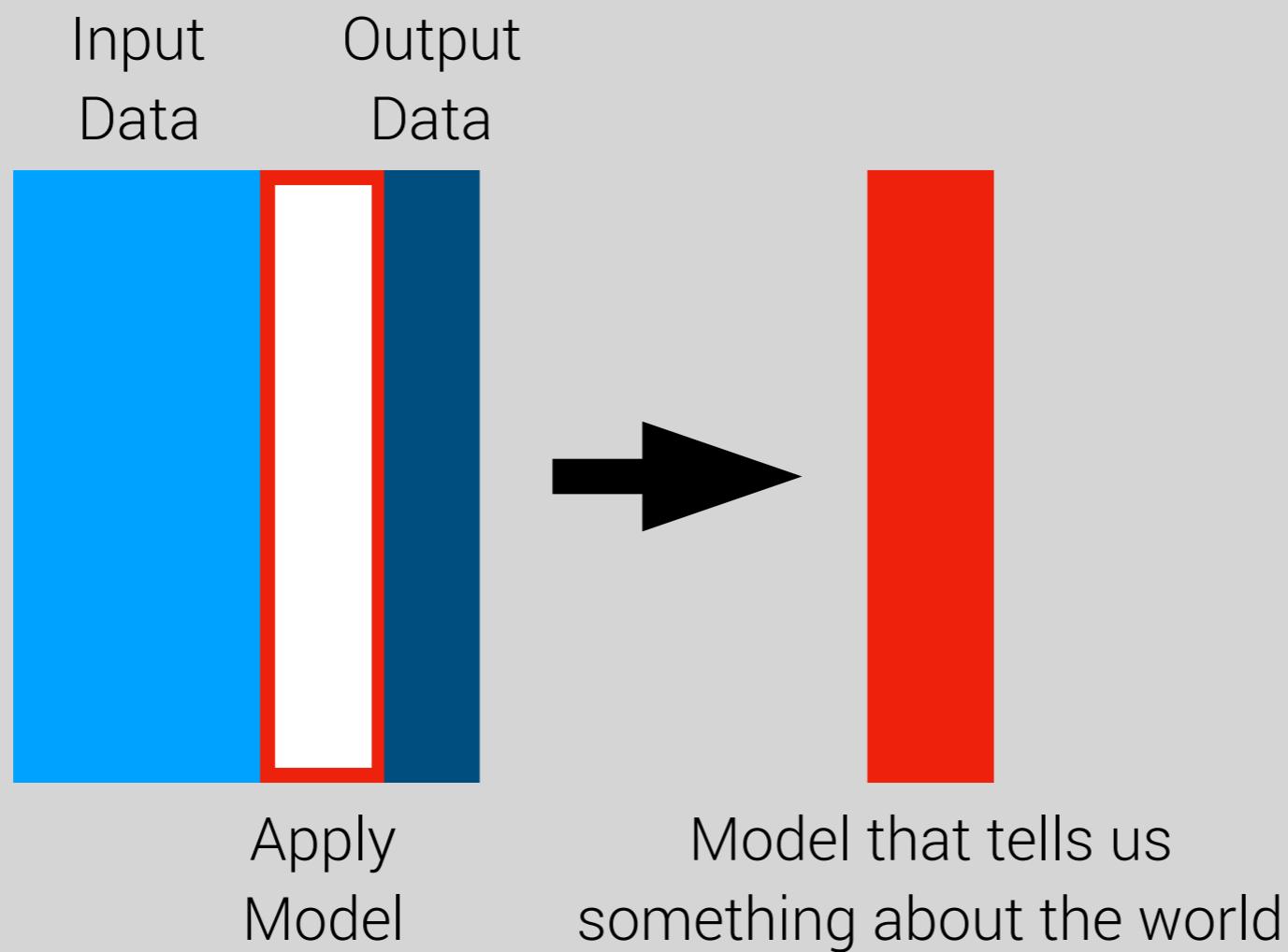


Arthur Samuel, 1952

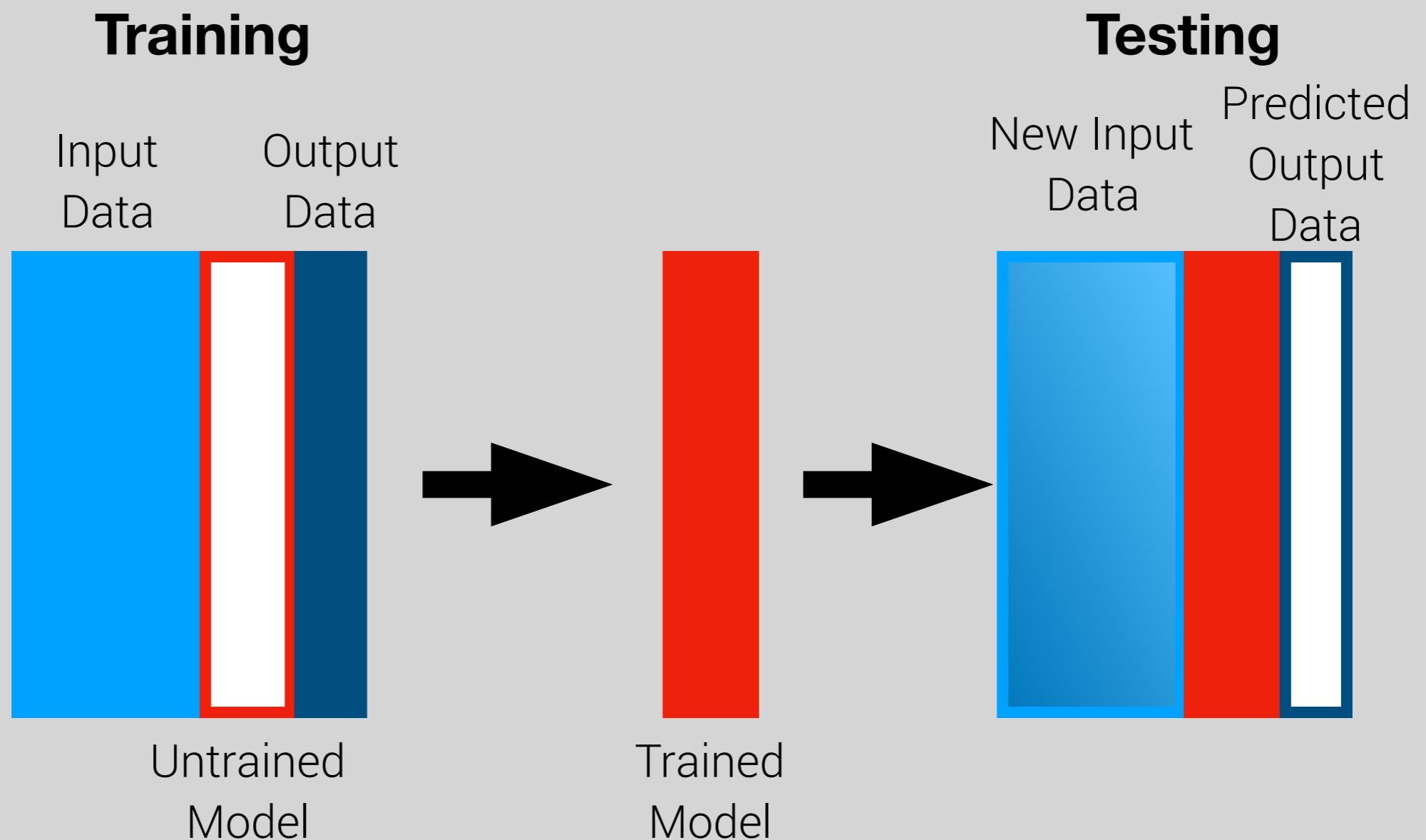
Machine Learning

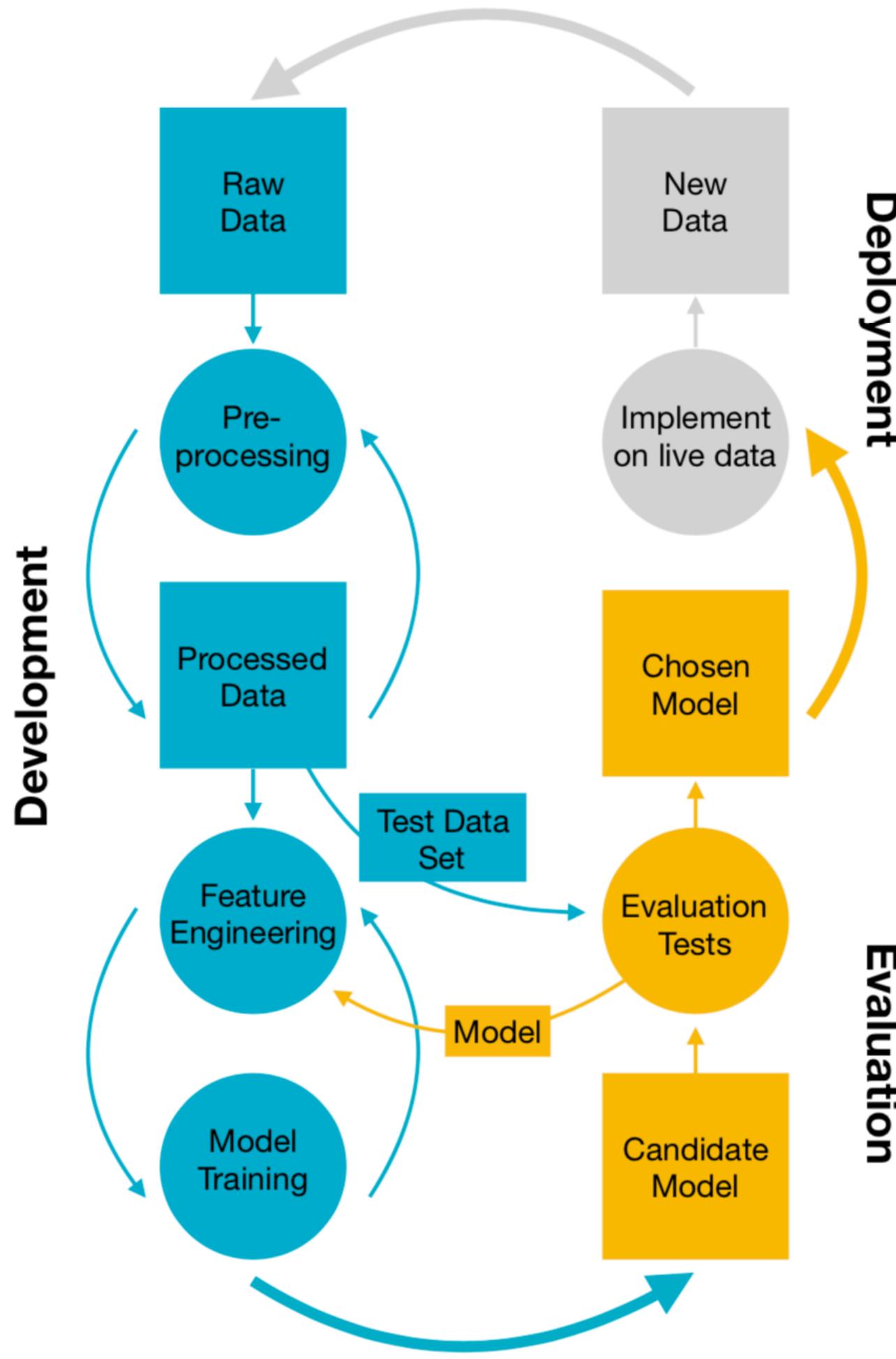


Educational Statistics



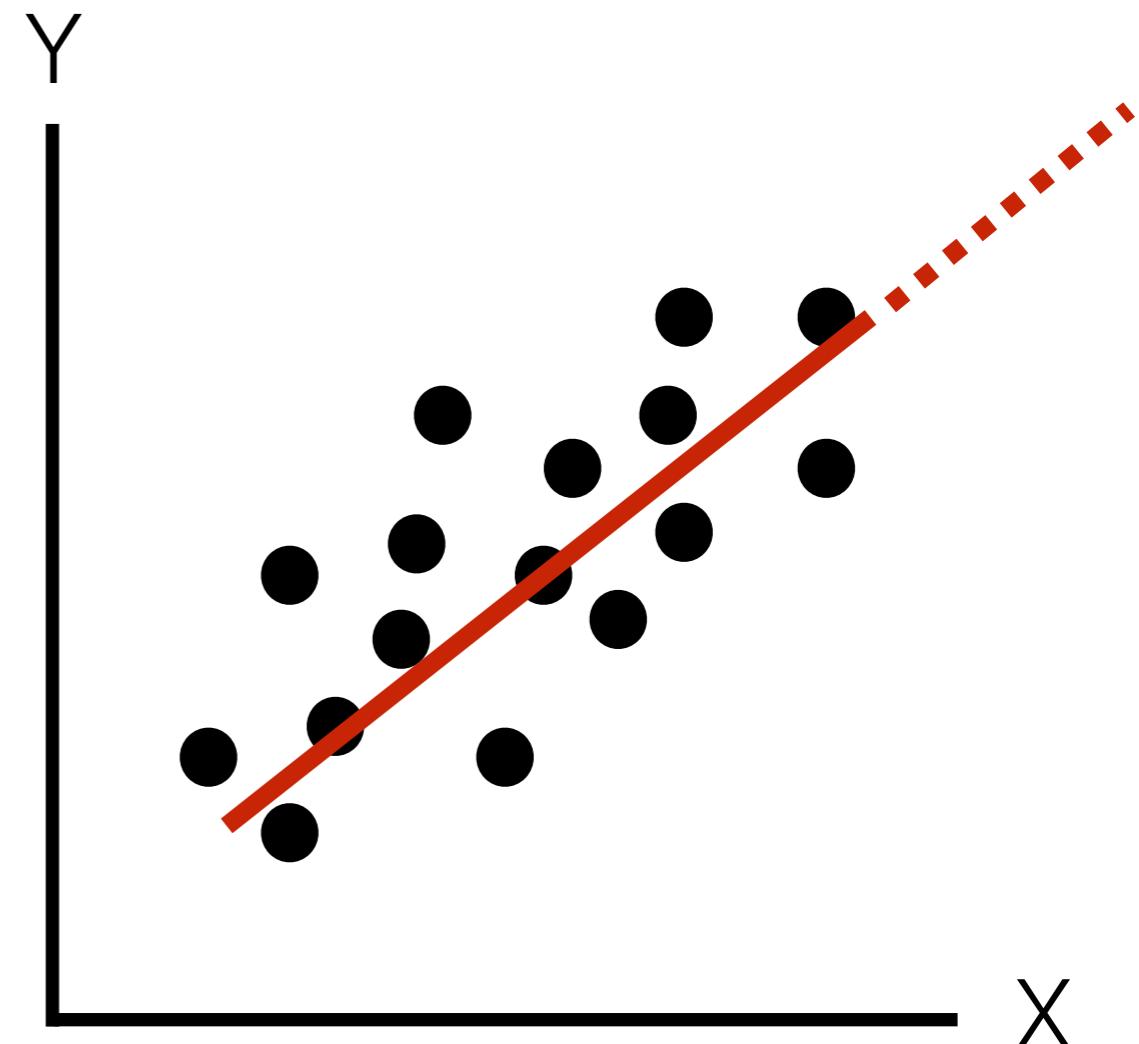
Machine Learning





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

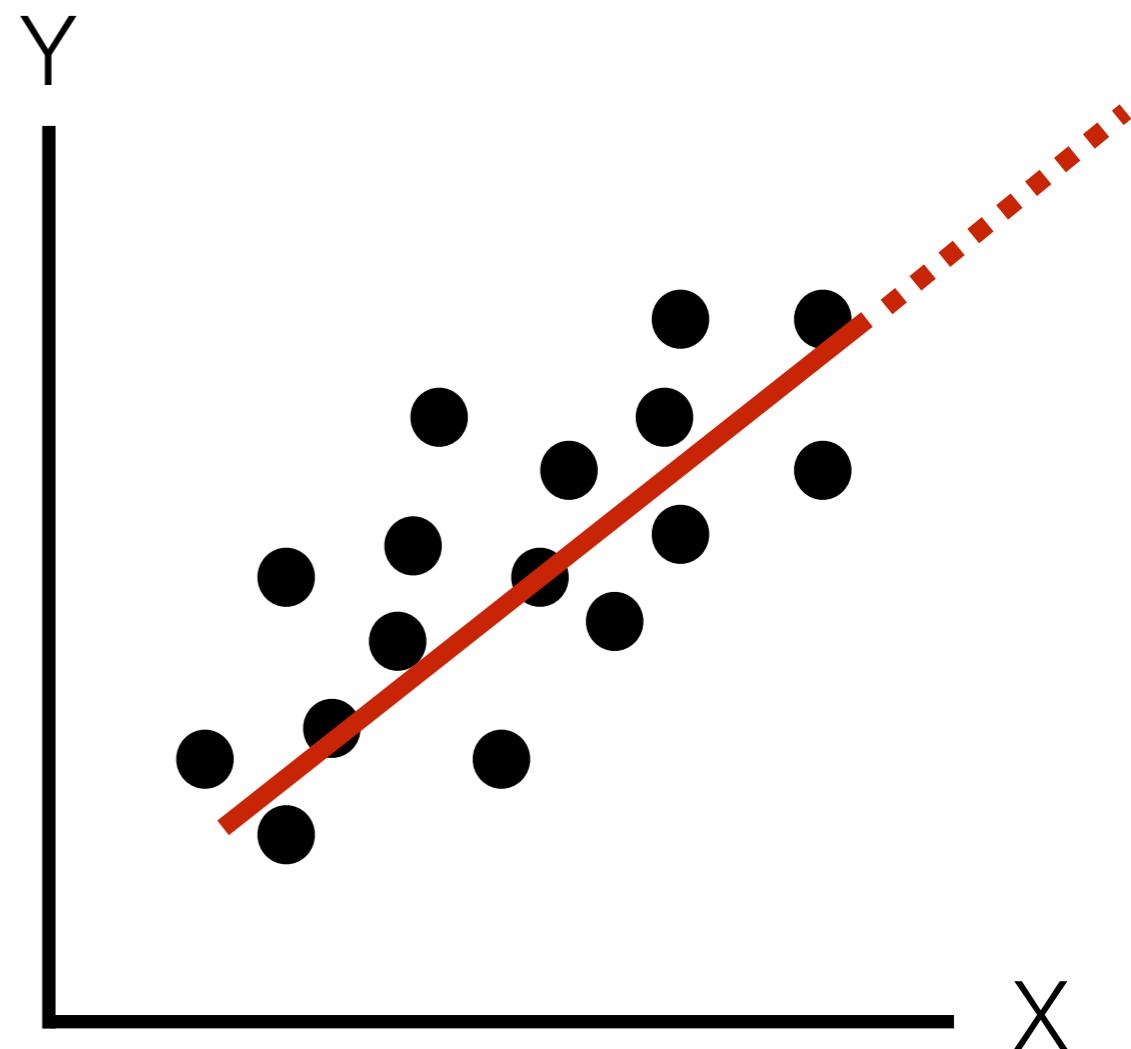
Regression (as a form of classification): Mapping from an unlabeled instance to a value within a continuous range

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

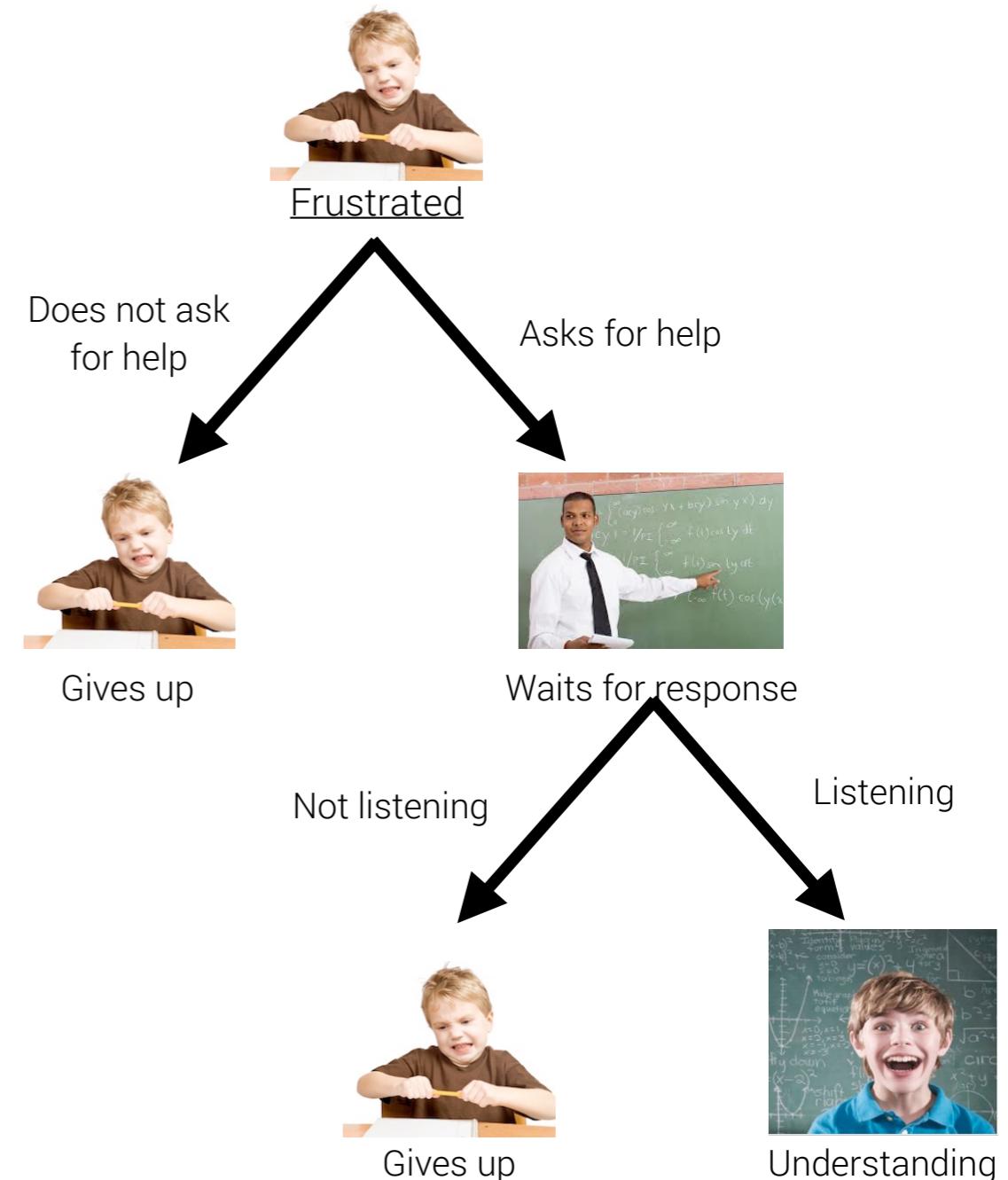
Regression

- 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 as classification
- EG: student states = leaves, student actions = branches



Machine Learning

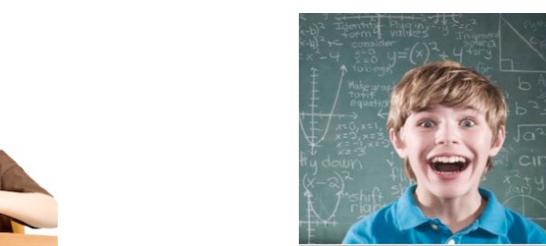
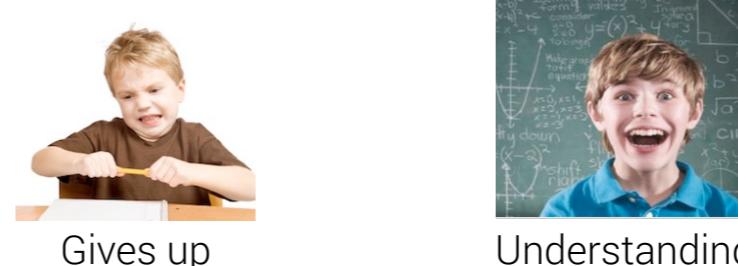
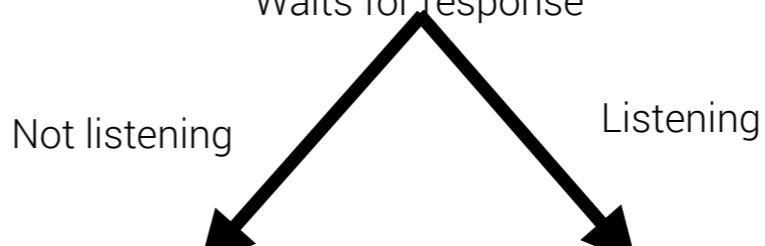
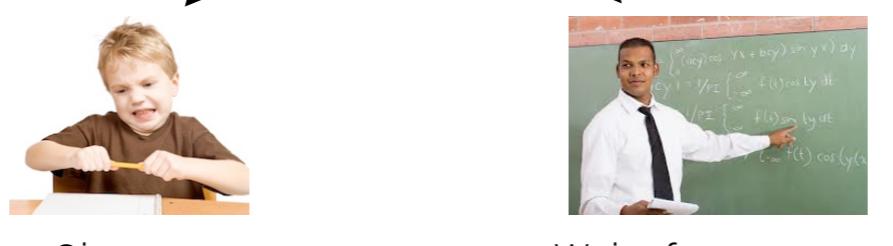
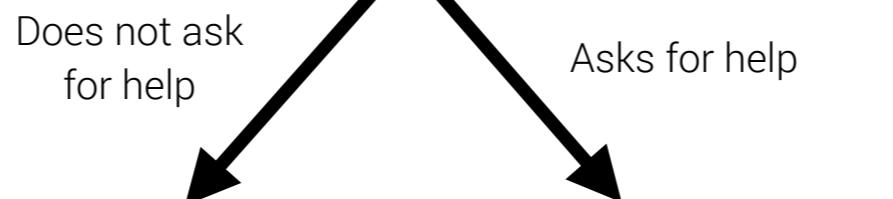
Input

- Does not ask for help
- Asks for help

- Not listening
- Listening

Process Structure/Weights

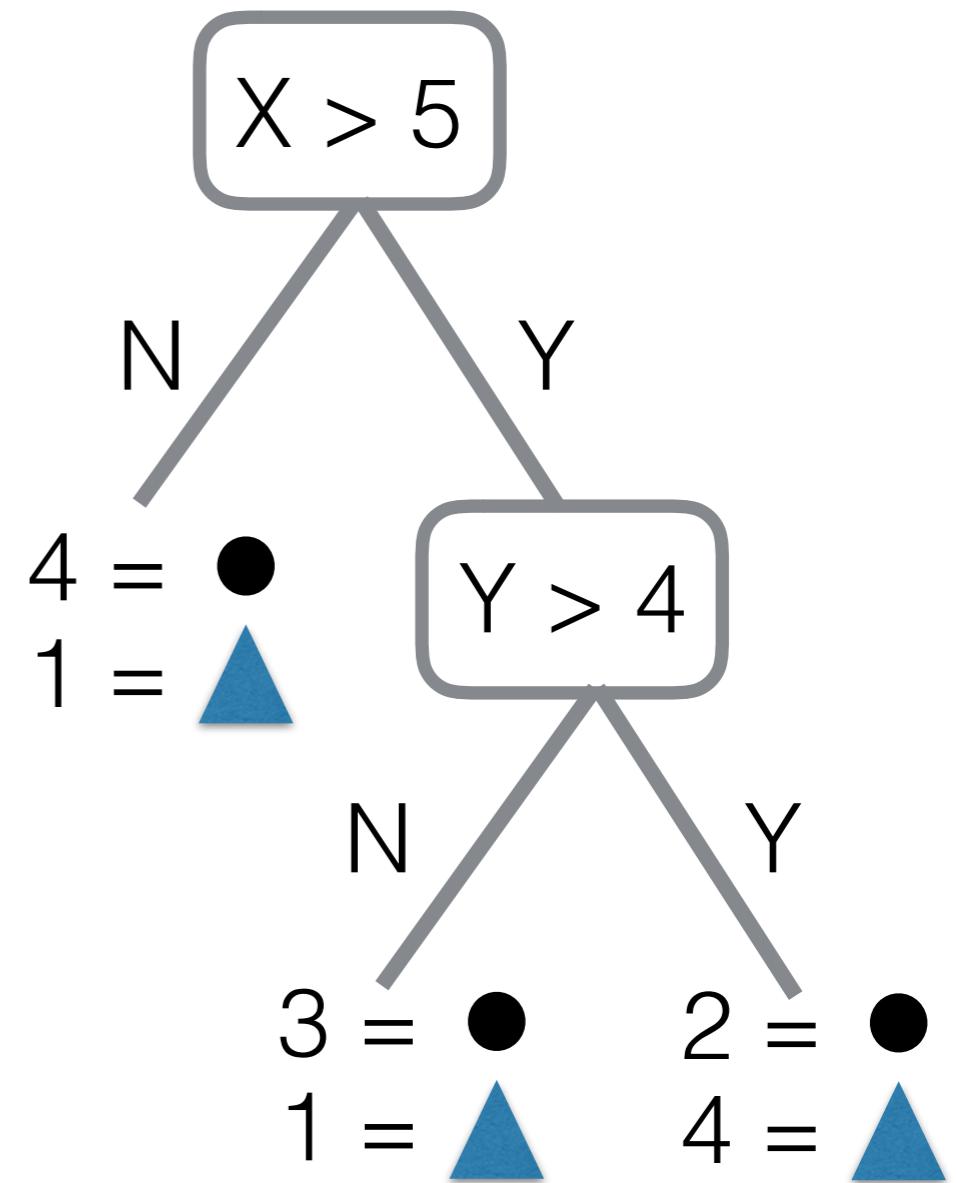
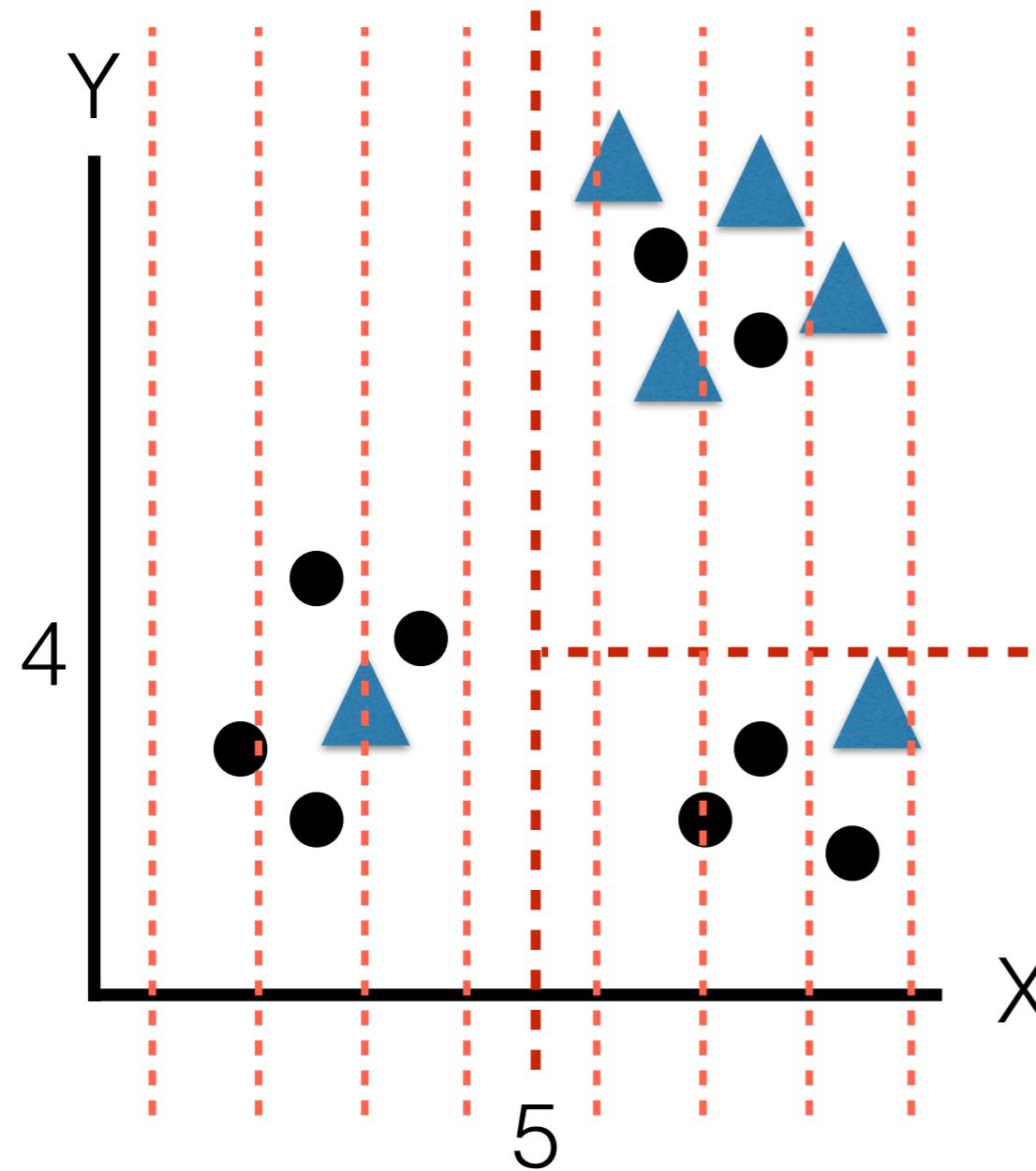
Output



Understanding

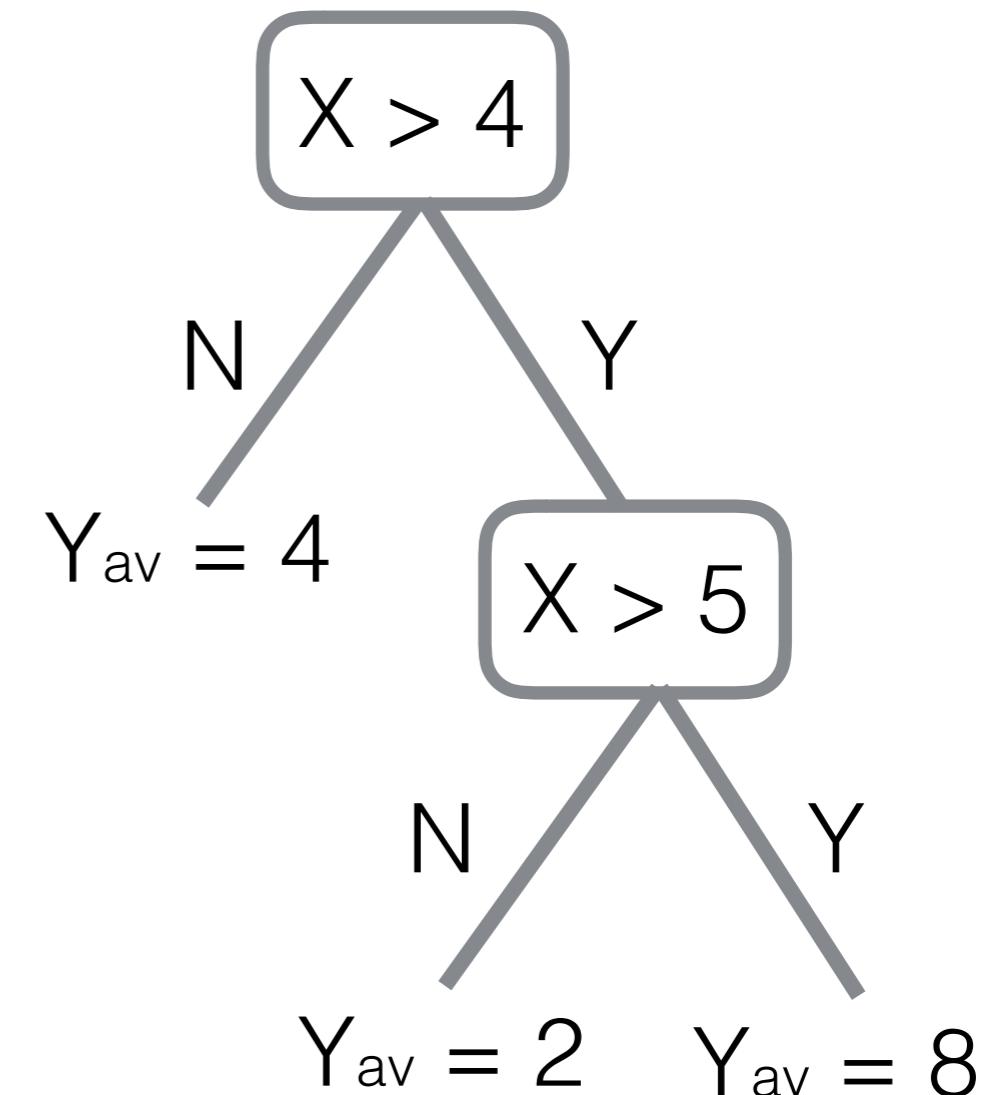
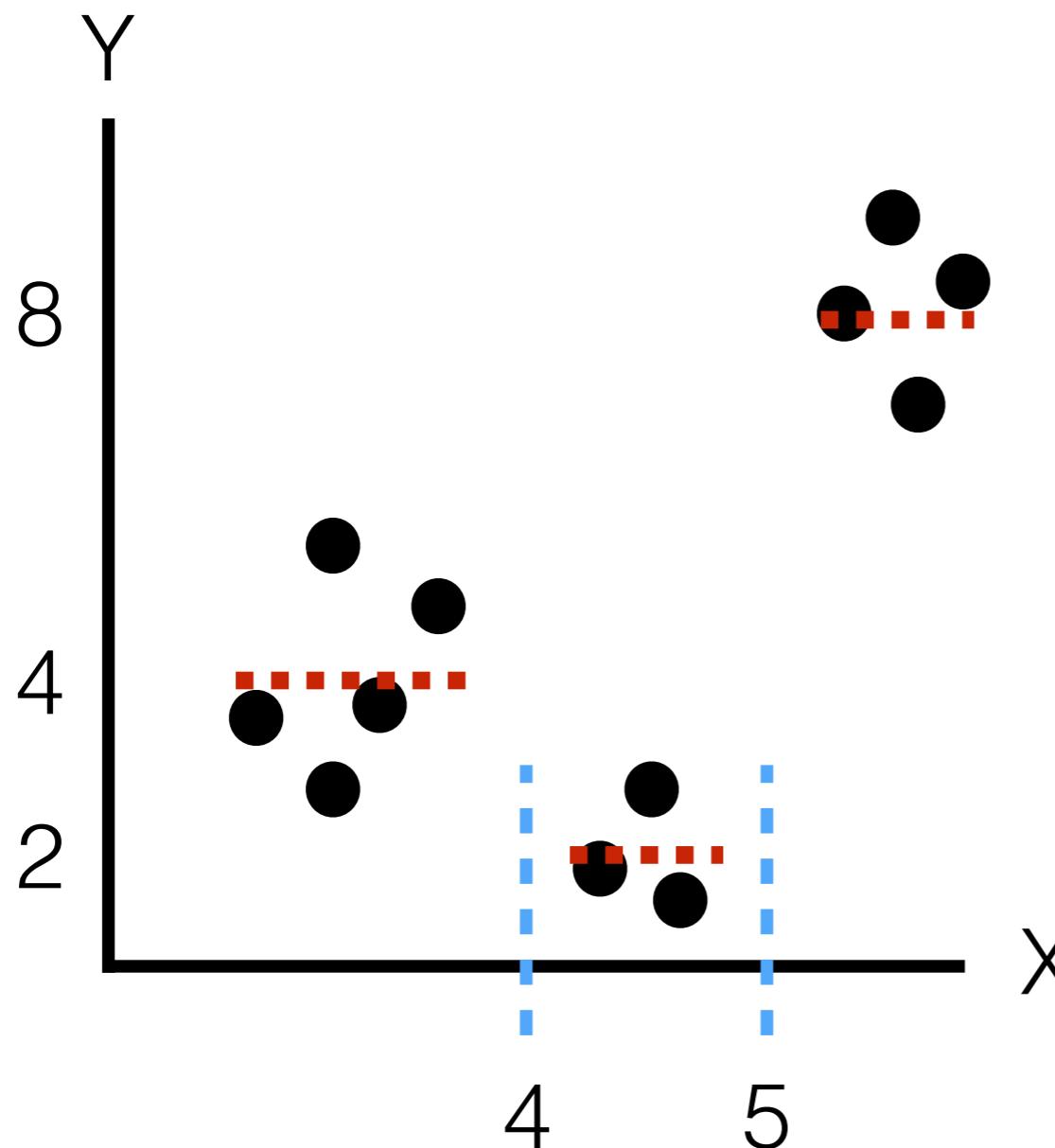
Binary Classification Tree

- * Minimize the error



Binary Regression Tree

- * Minimize the error



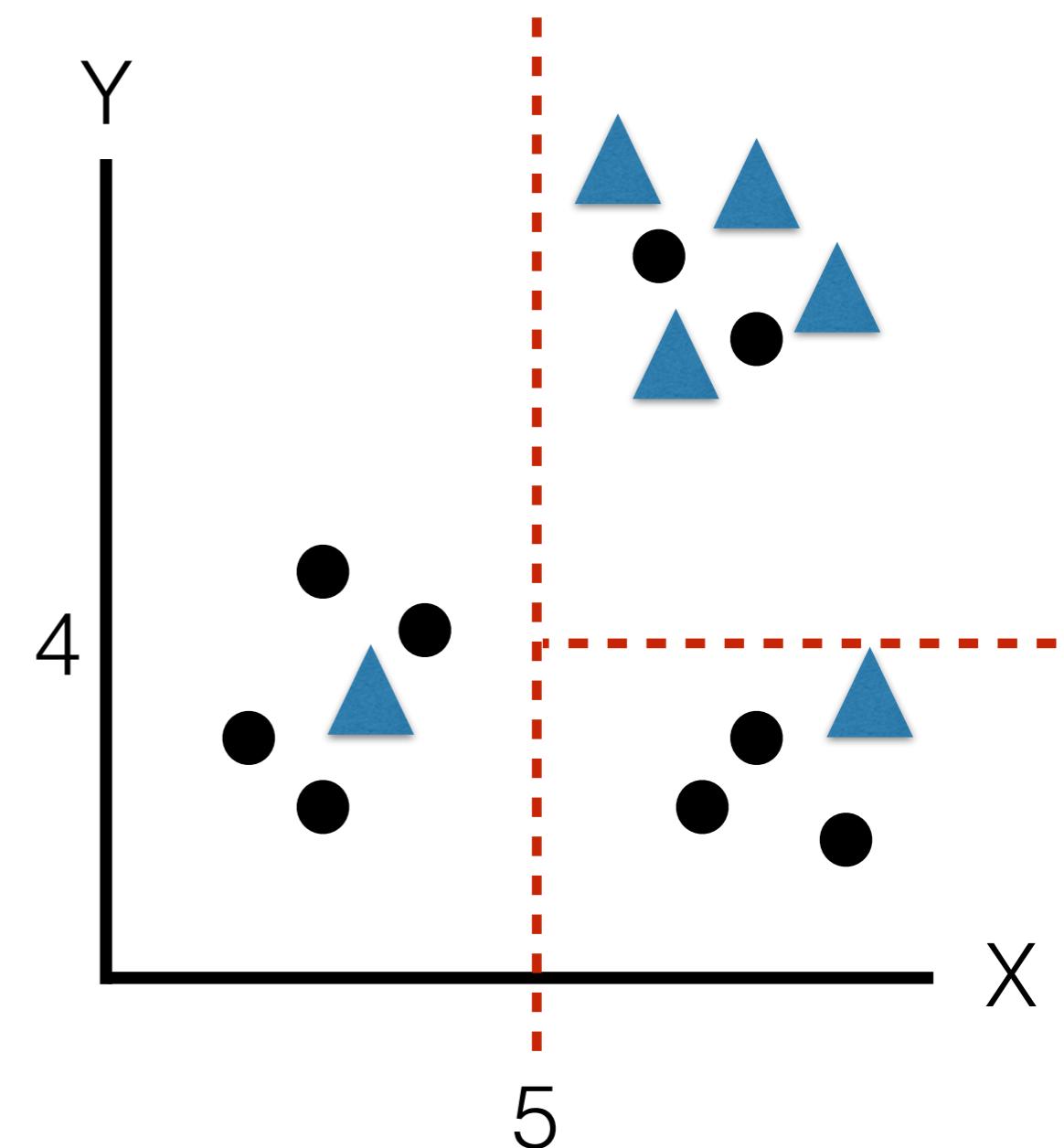
CART Trees

- 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)



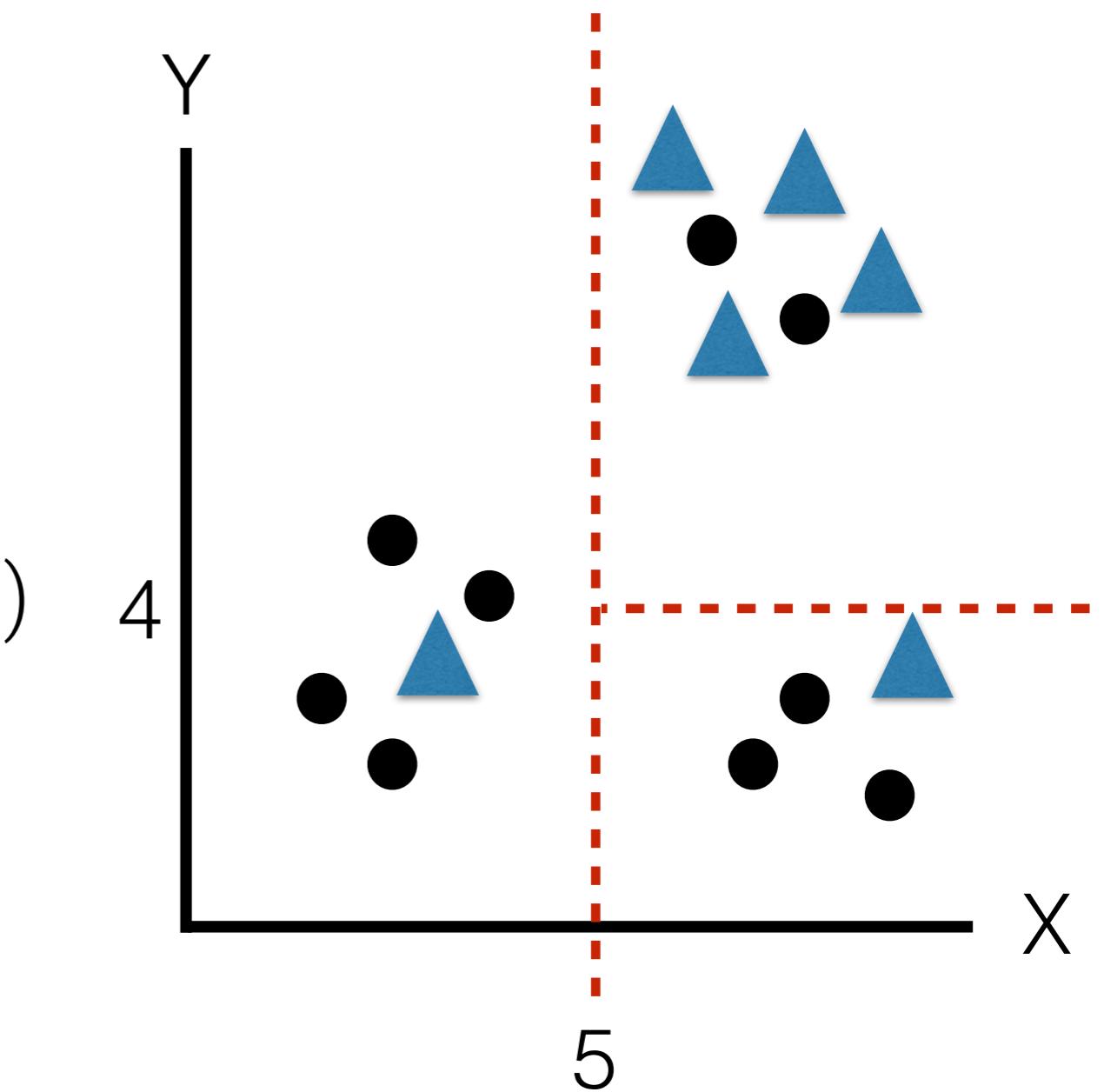
RPART

- 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



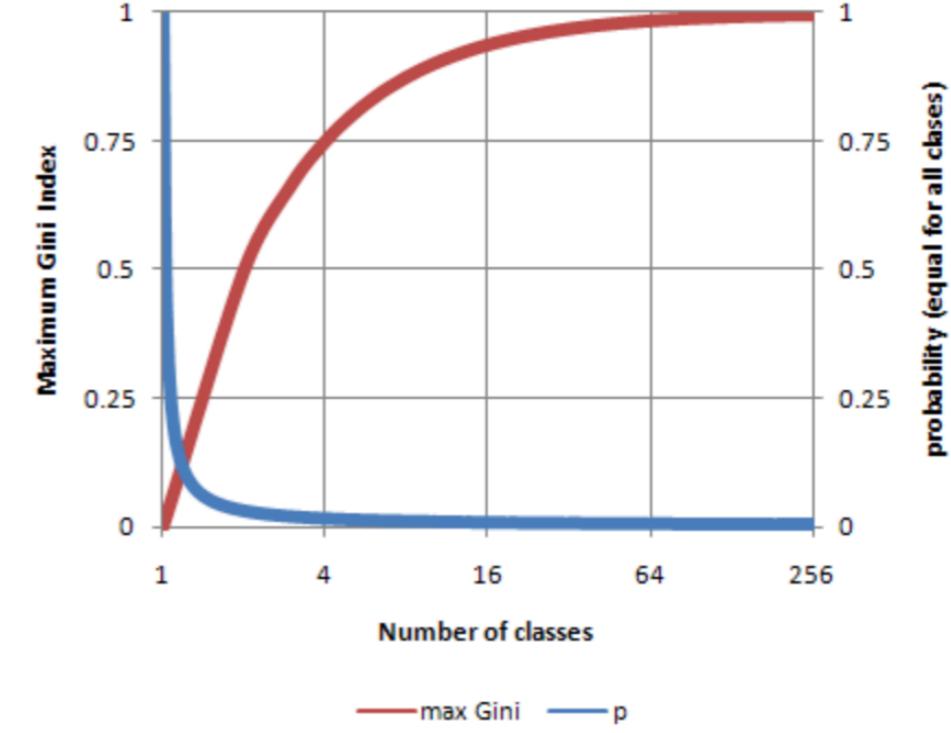
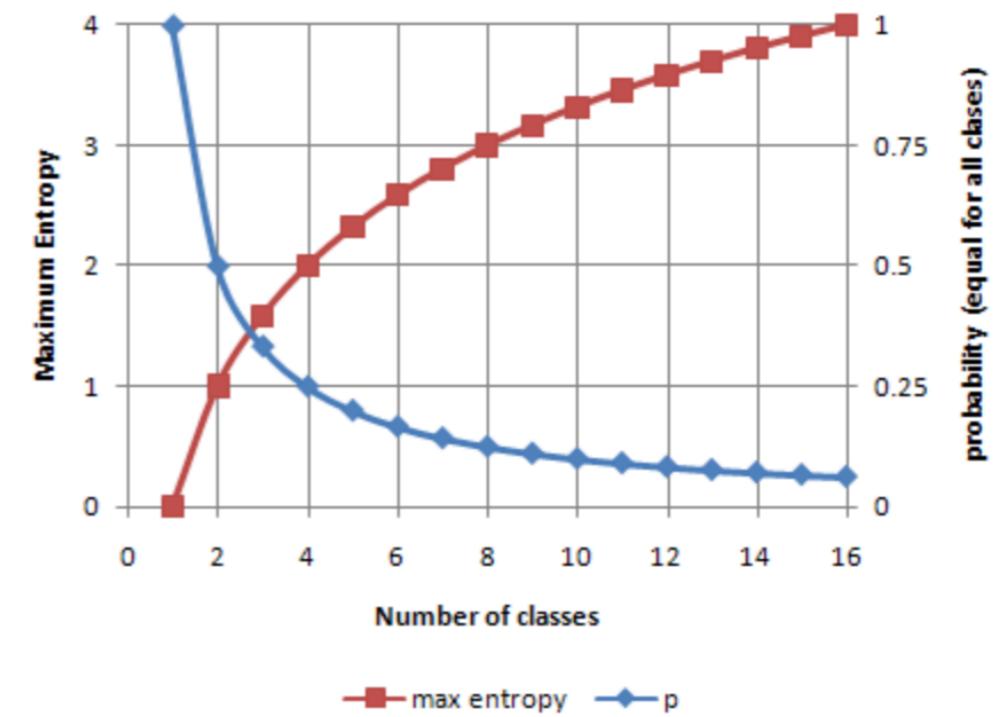
RPART

- parms
- Entropy (information gain)

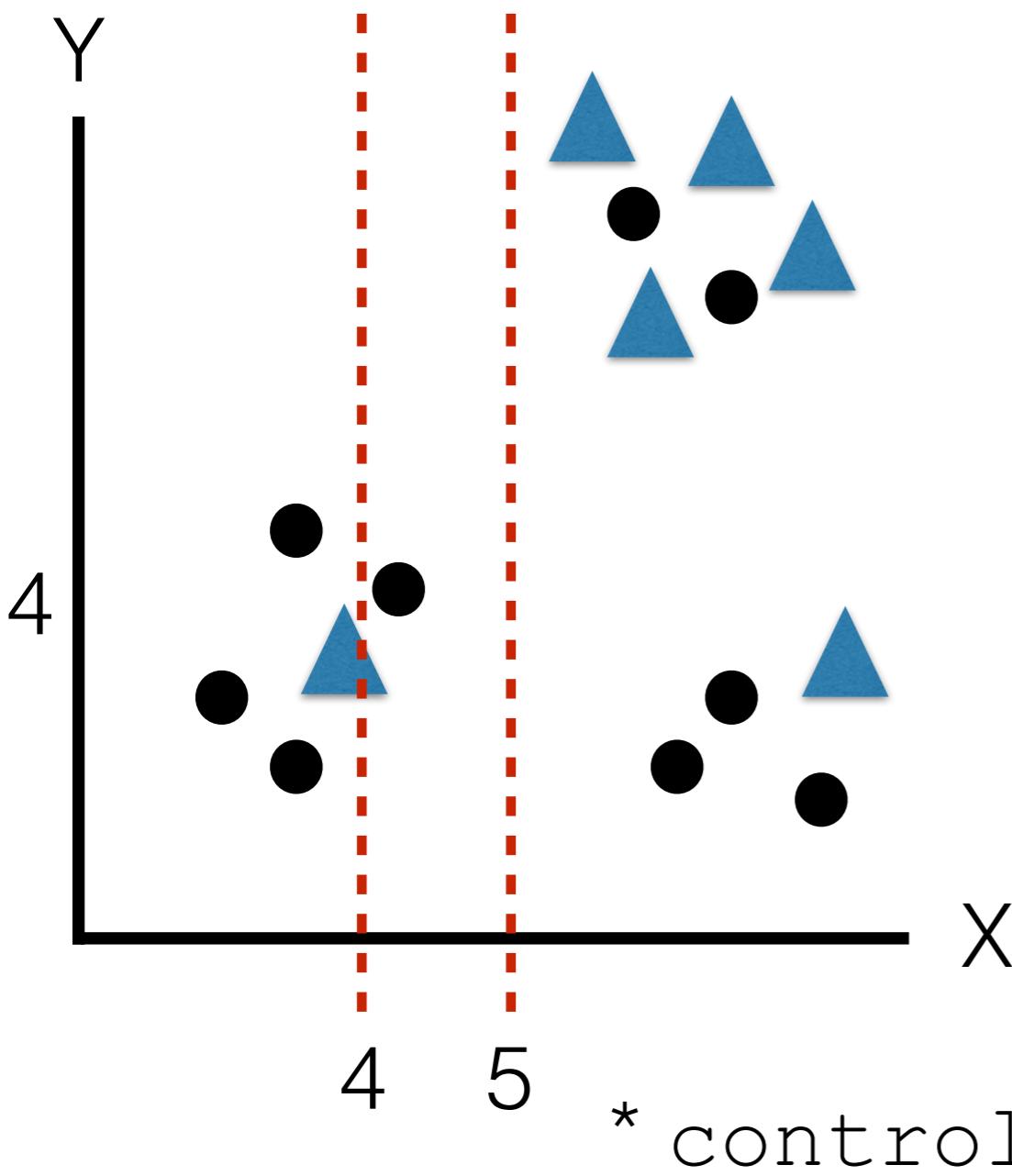
$$\text{Entropy} = \sum_j -p_j \log_2 p_j$$

- Gini index

$$\text{Gini Index} = 1 - \sum_j p_j^2$$



RPART



$X > 5$

N Y

4 = ●
1 = ▲

5 = ●
5 = ▲

$$\text{ENT}_5 = -0.8 \cdot \log_2(0.8) + -0.5 \cdot \log_2(0.5) = 0.75$$

$X > 4$

N Y

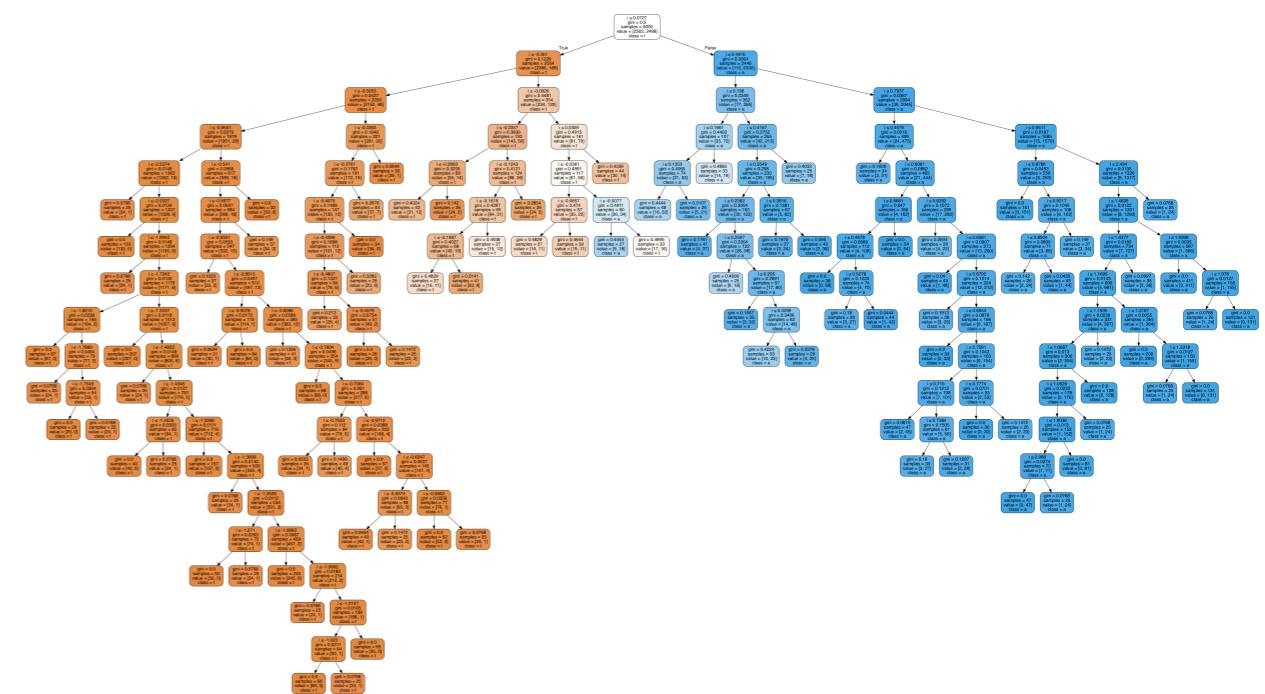
3 = ●
1 = ▲

6 = ●
5 = ▲

$$\text{ENT}_4 = -0.75 \cdot \log_2(0.75) + -0.55 \cdot \log_2(0.55) = 0.76$$

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



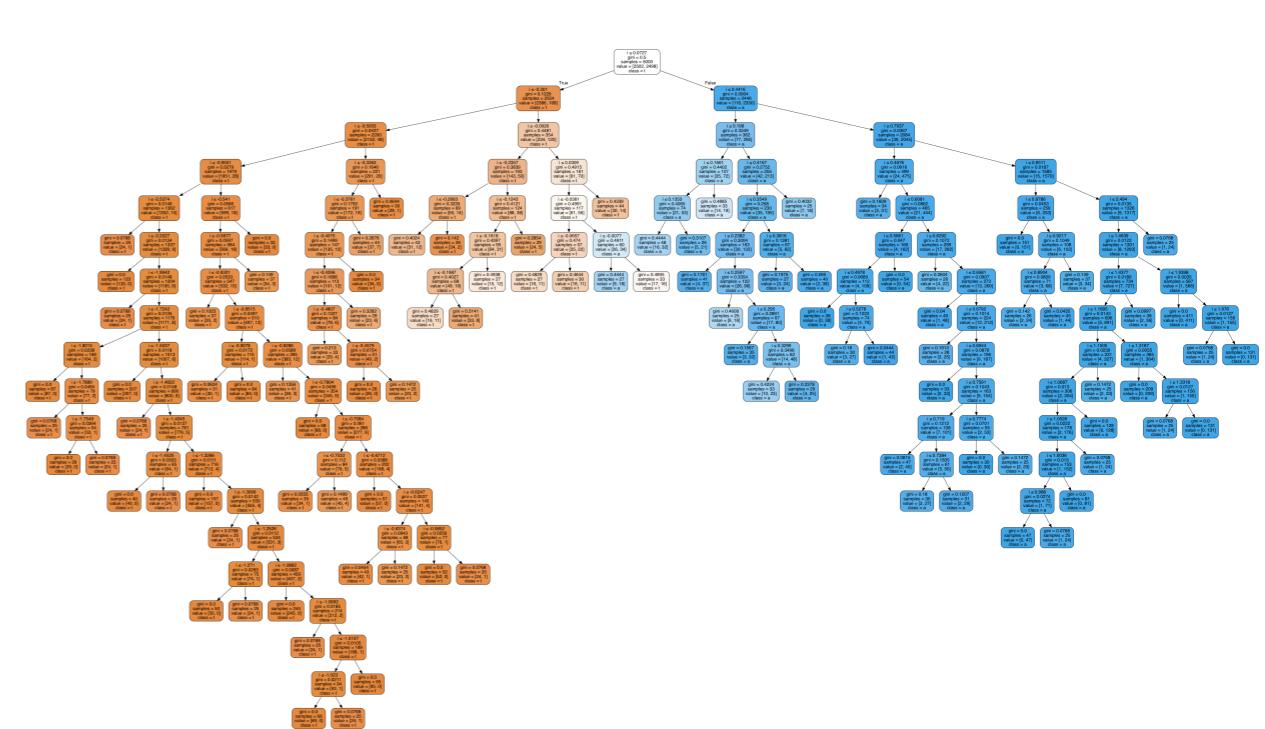
RPART

- RPART prunes
- Uses a cost function:

Number of leaves

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

**Misclassified
instances**



RPART

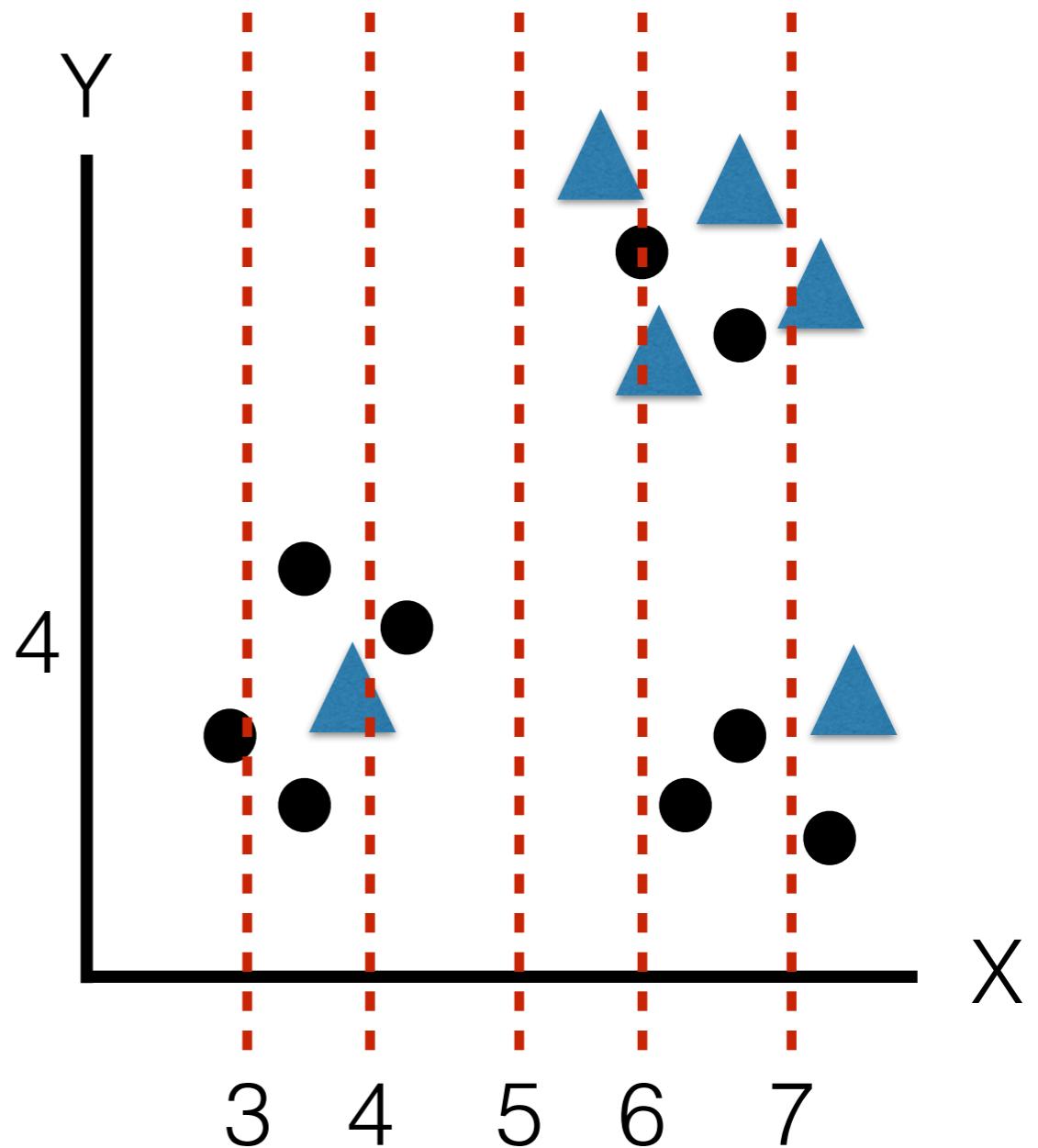
Gotchas

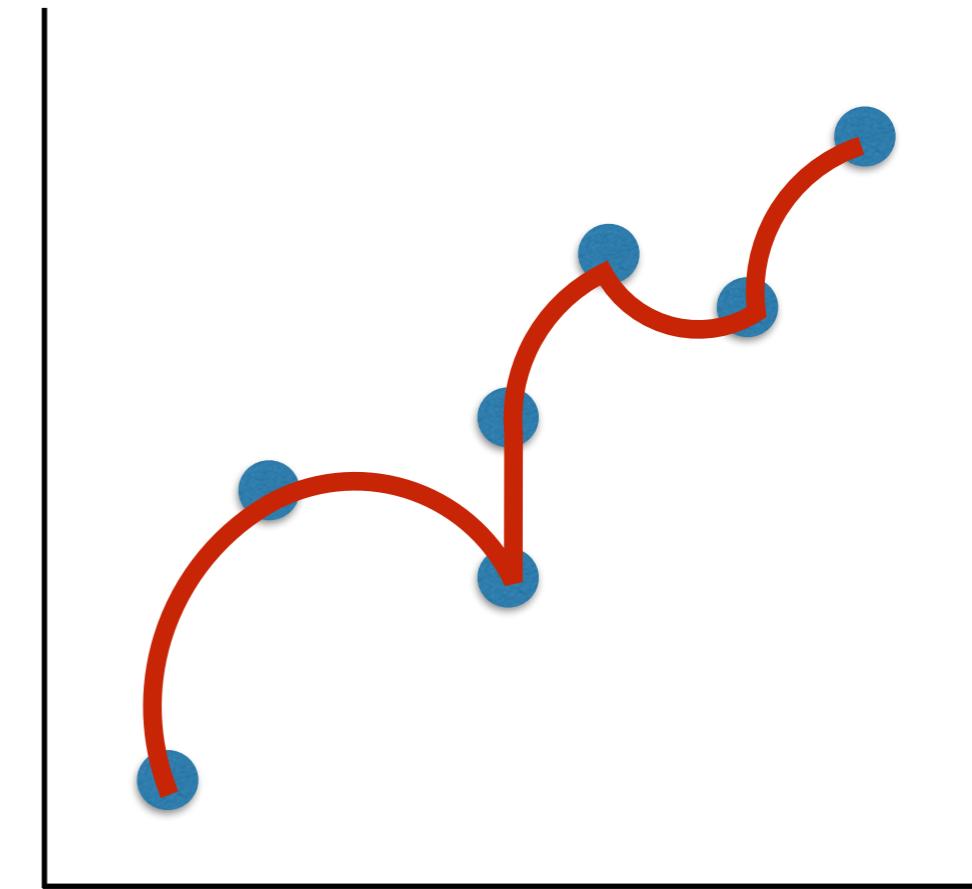
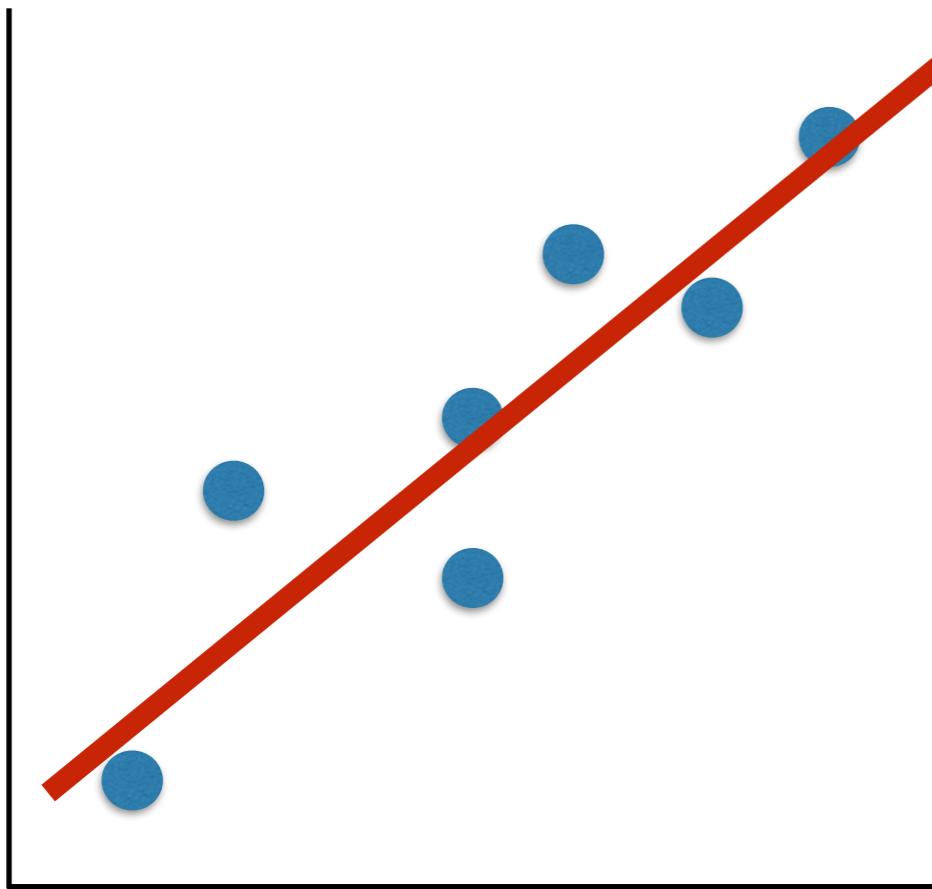
- Overfitting
- Local overfitting
- Sensitive to test data
- Selection bias toward covariates with many possible splits



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”?

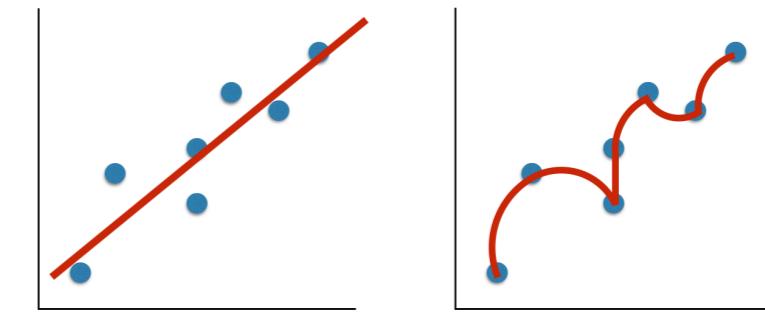
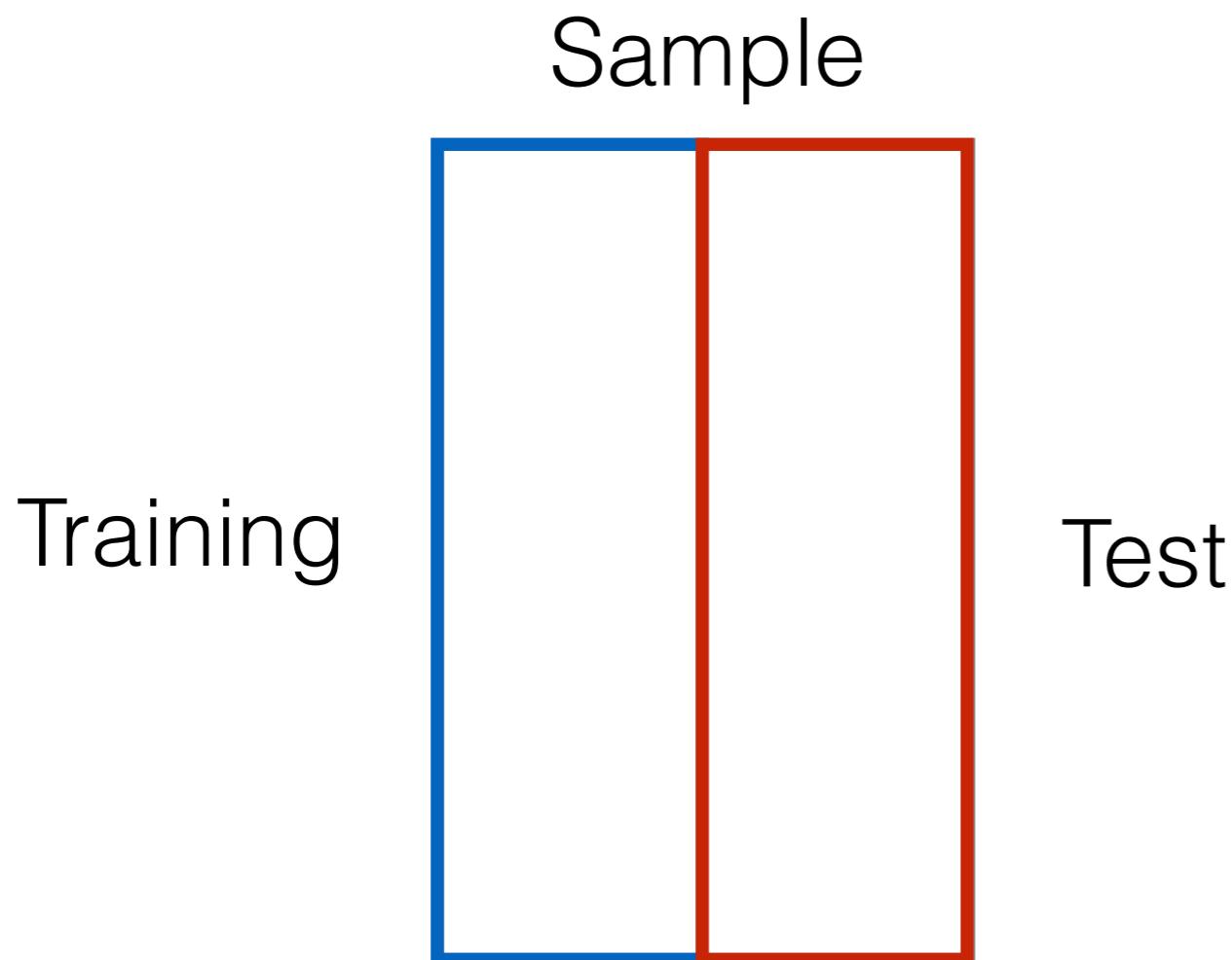
Which is more “useful”?

How can we tell?

Cross Validation

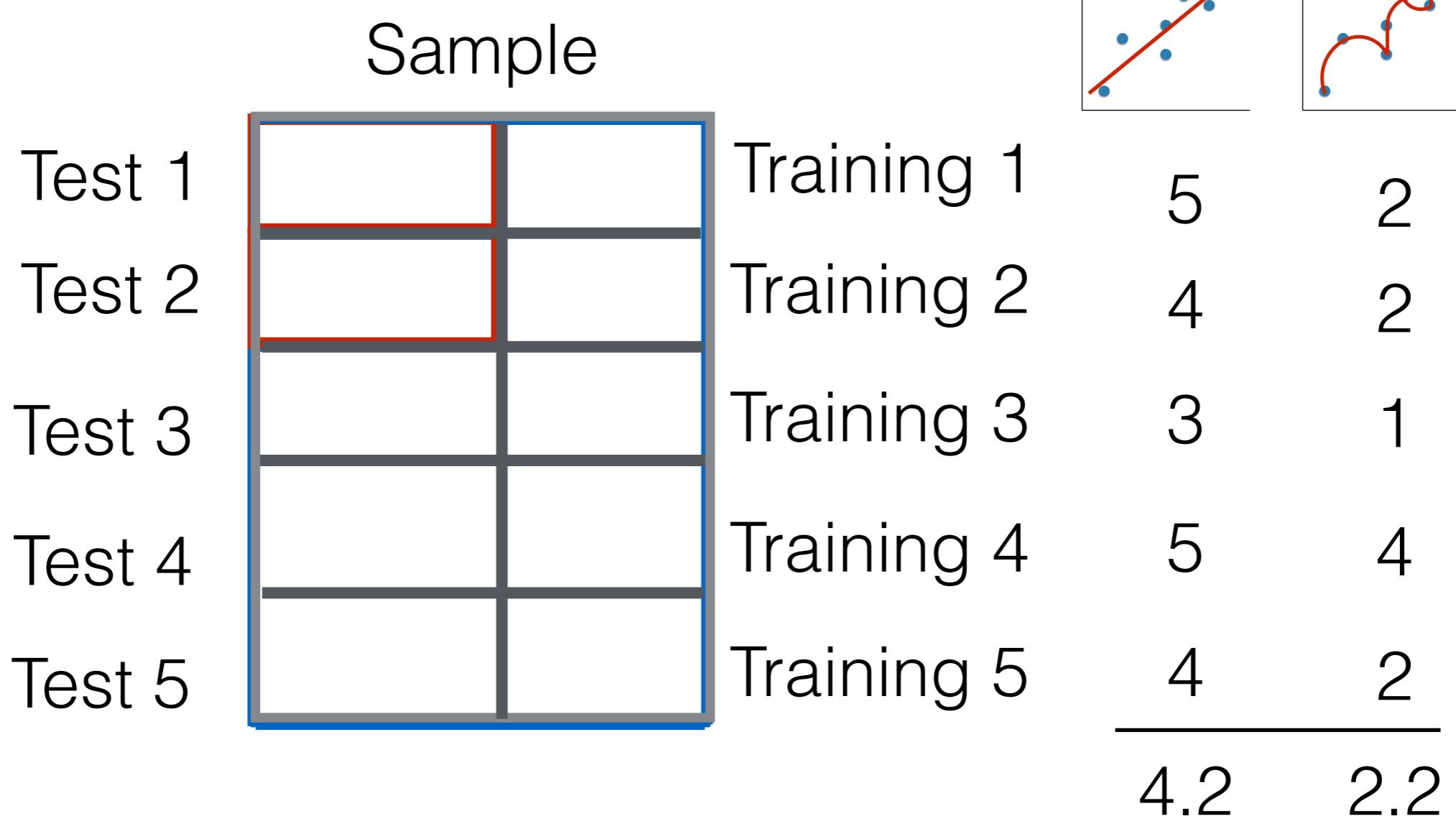
- Estimate how accurately a predictive model will perform in practice
- Give an insight on how the model will generalize to an independent dataset

Hold-out Validation



Problem: very dependent on which data are in each group

K-Fold Cross Validation



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