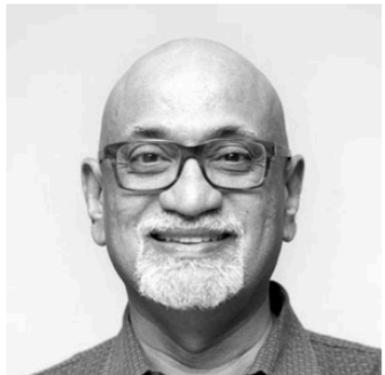


HUDK 4051:  
LEARNING  
ANALYTICS:  
PROCESS &  
THEORY

# Events

<https://events.columbia.edu/cal/event/showEventMore.rdo;jsessionid=FA0759B9138B57B9A20FEF1B2053F1B8>



**February 28 | 12:00PM-1:30PM**

**P. Anandan, CEO, Wadhwanai AI**

*Location: Northwest Corner Building, 14th Floor*

**RSVP**



**March 6 | 12:00PM-1:30PM**

**Zachary Lipton, Assistant Professor, Carnegie Mellon University**

*Location: CEPSR 414*

**RSVP**



**March 13 | 12:00PM-1:30PM**

**Phebe Vayanos, Assistant Professor, Industrial & Systems Engineering and Computer Science, University of Southern California**

*Location: Northwest Corner Building, 14th Floor*

**RSVP**

A blue banner with white text for an IBM info session. It features a city skyline background.

**Marketing Data Science  
at IBM Info Session**

**WHEN: 3:30 PM - 4:30 PM, THURSDAY, FEBRUARY 13TH.**

**WHERE: CLASSROOM 414 CEPSR**

<https://www.eventbrite.com/e/marketing-data-science-at-ibm-info-session-tickets-93576627211>

A blue banner for Data Science Day at Columbia University. It features a crown icon and a network graphic.

**DATA SCIENCE INSTITUTE  
COLUMBIA UNIVERSITY**

**DATA SCIENCE DAY**

**MARCH 31<sup>ST</sup> 2020**

[https://datascience.columbia.edu/?utm\\_source=sendinblue&utm\\_campaign=Events\\_Weekly\\_February\\_11\\_2020&utm\\_medium=email](https://datascience.columbia.edu/?utm_source=sendinblue&utm_campaign=Events_Weekly_February_11_2020&utm_medium=email)

# Recommender Systems\*

# Today

- Prediction background
- The five tribes
- Caret (Weka)
- Activity

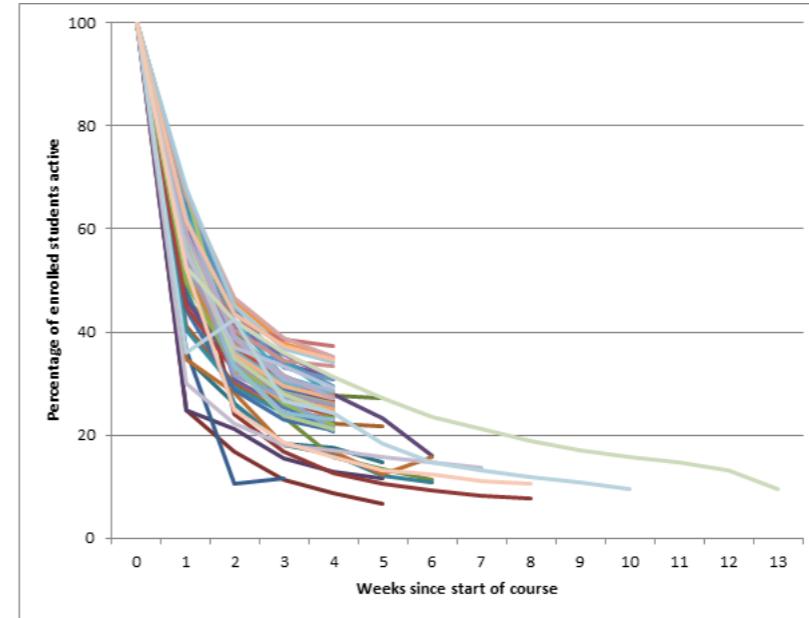
# Prediction

- There are times when we want to automate a process in education
- Action oriented
- Sometimes this is a supervised learning problem
- (Sometimes it is not)



# Prediction

- Common supervised learning problems:
- Drop out, attendance, payment
- Correct/incorrect



K Jordan, Open University, 2013



# Prediction

- The aim of prediction is to predict the future and intervene

## **Predictive Modelling in Teaching and Learning**

Christopher Brooks & Craig Thompson

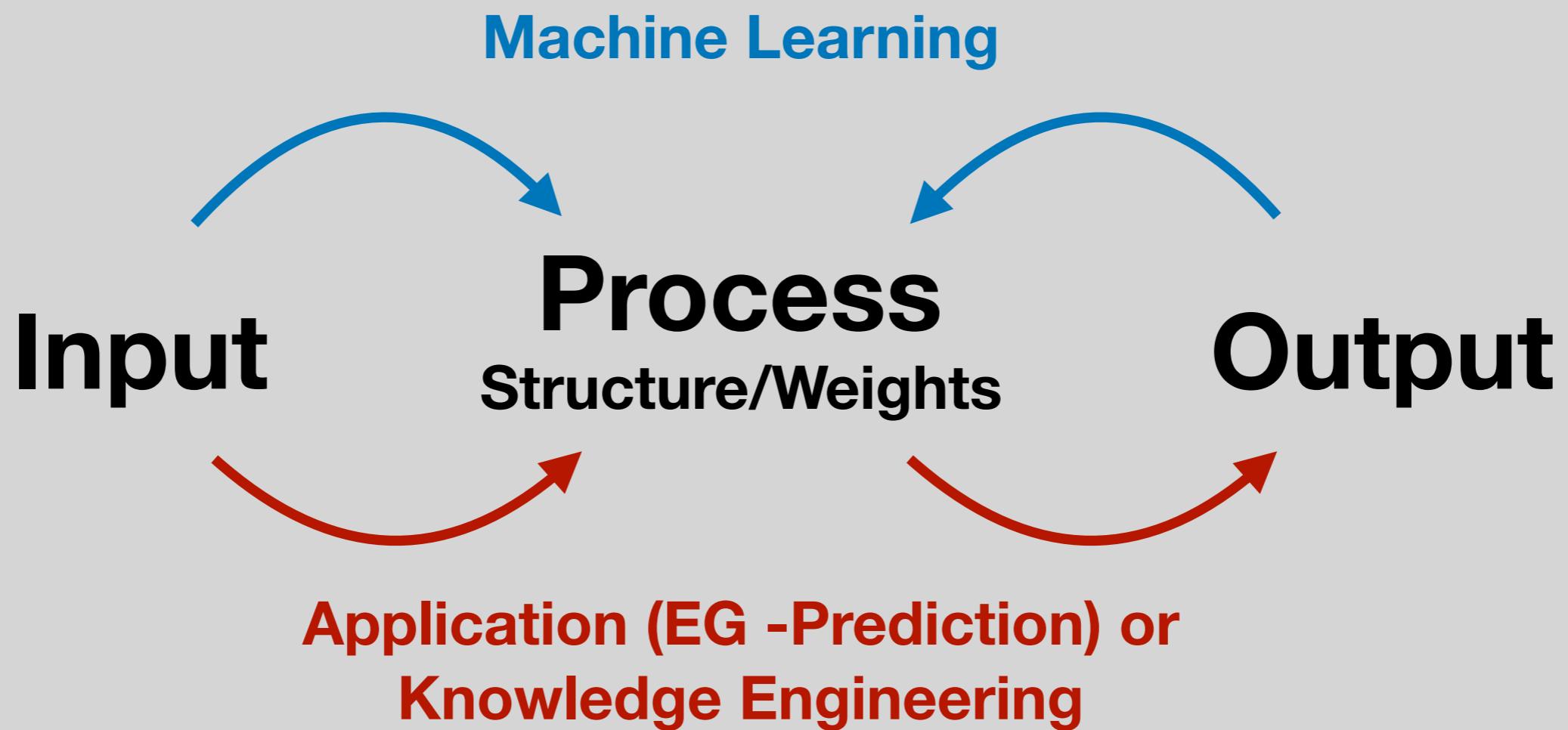
- Key metric is the accuracy of the prediction and how generalizable it is (Purdue)
- JEDM

## **PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS**

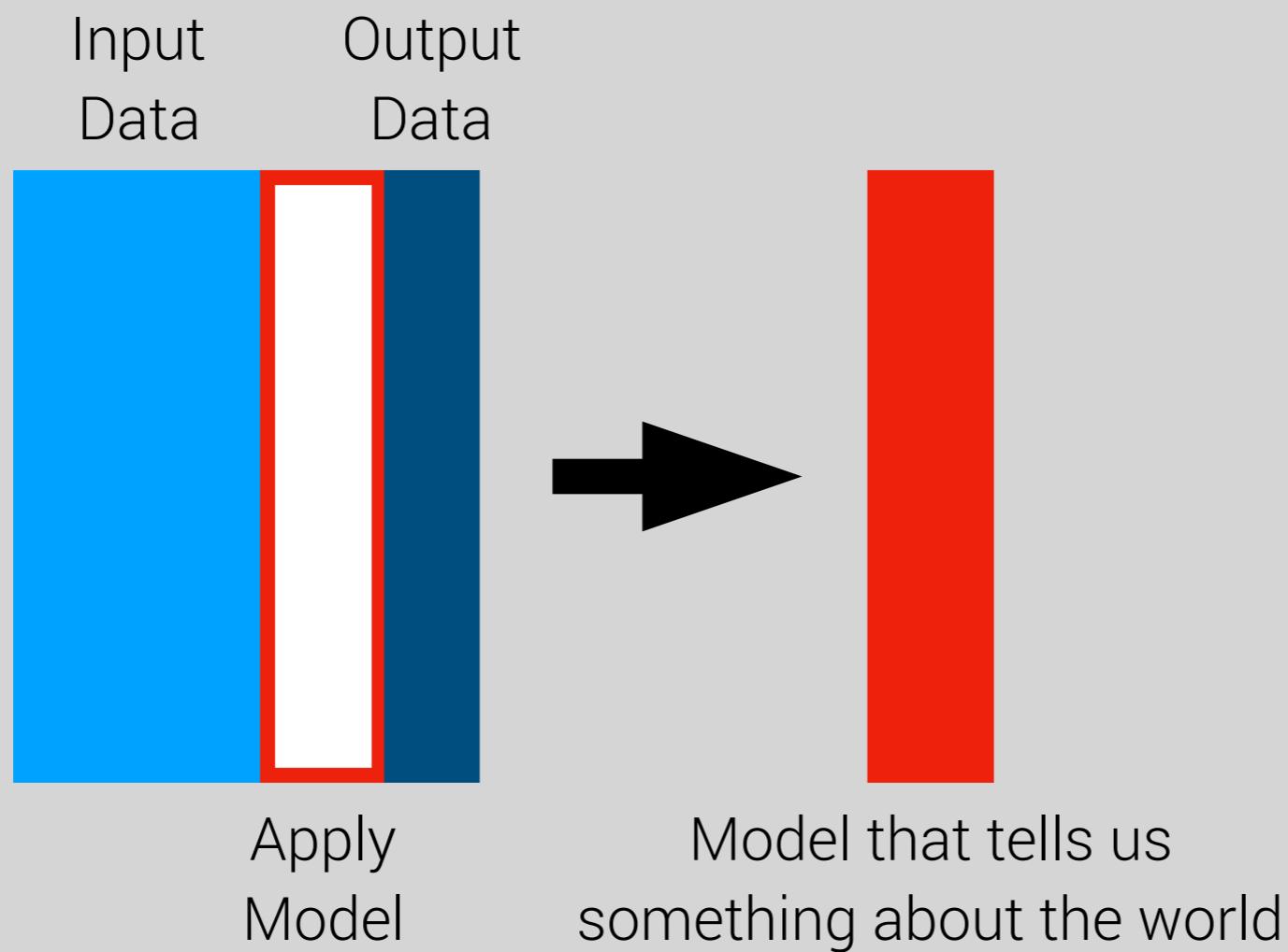
Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao

Department of Computer Engineering,  
Fr. C.R.I.T., Navi Mumbai, Maharashtra, India

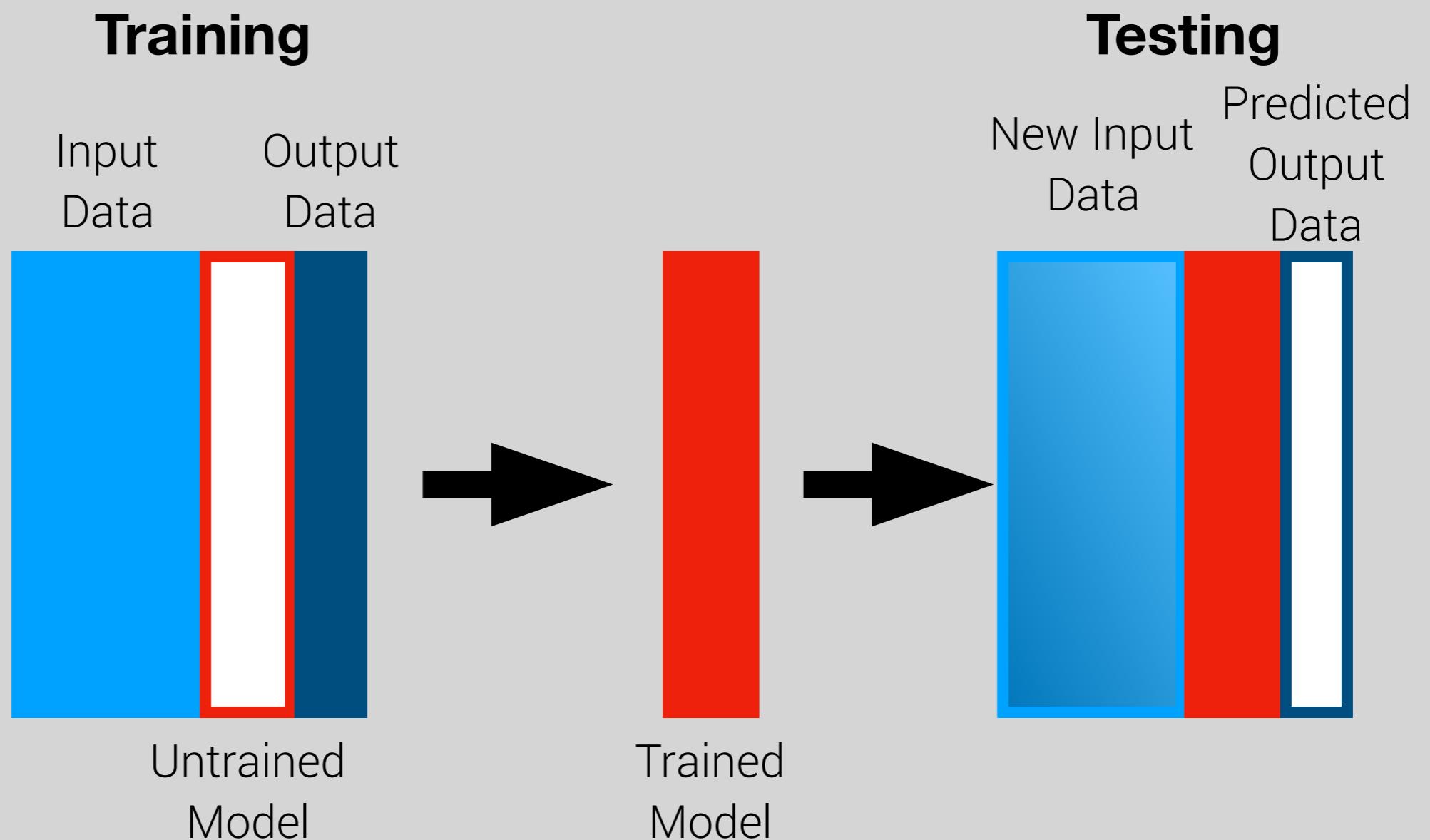
# Machine Learning

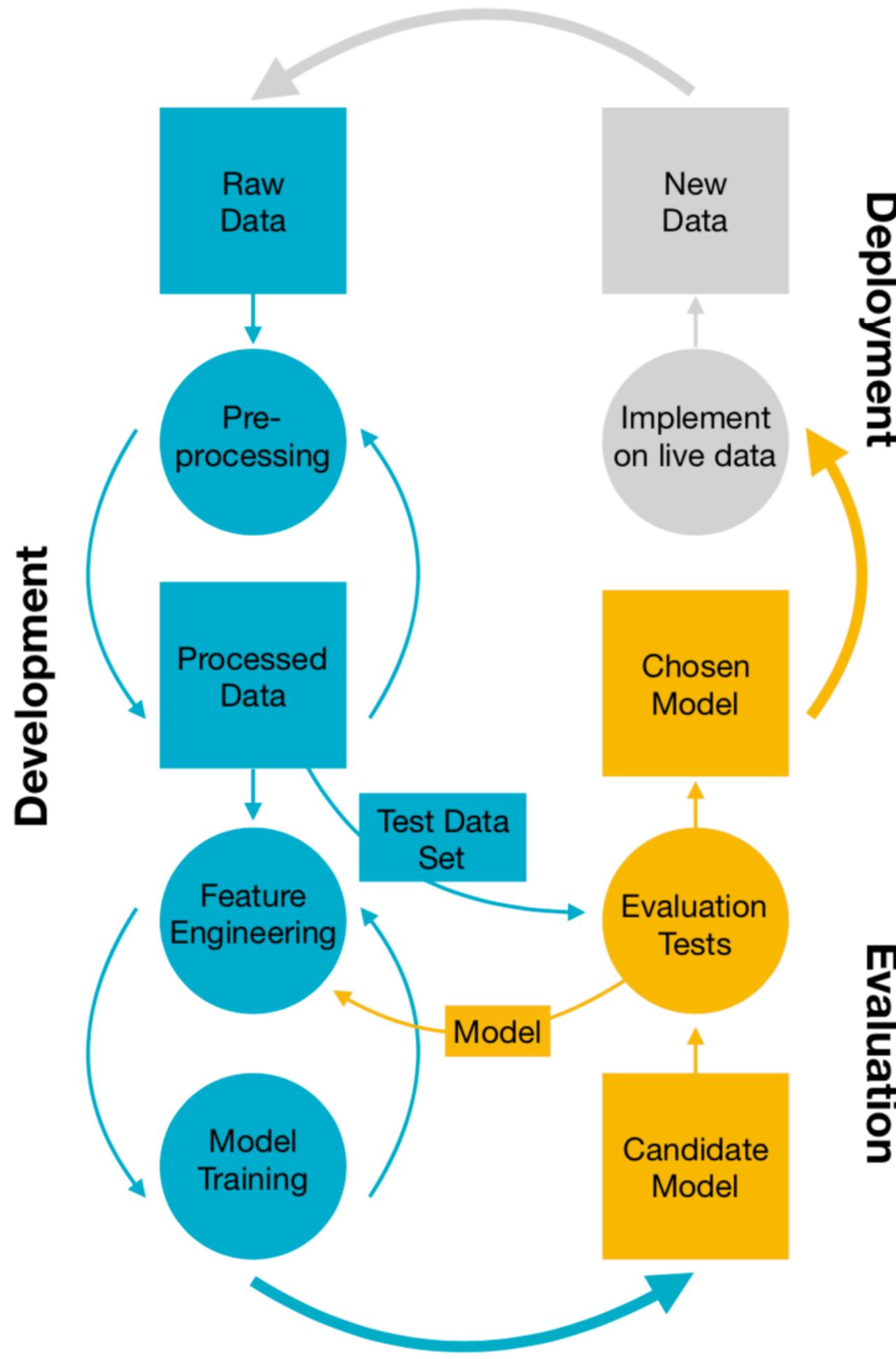


# Educational Statistics



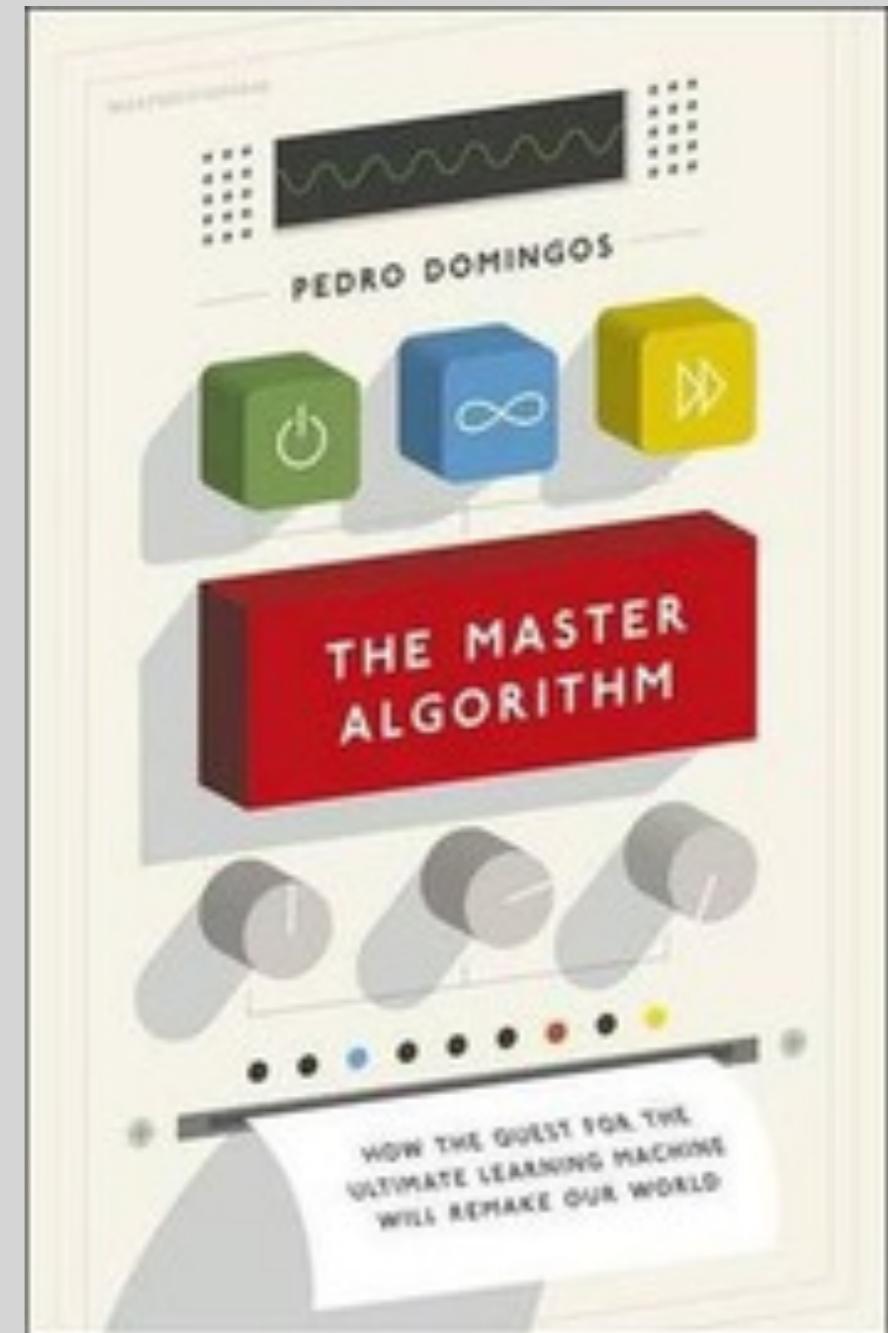
# Machine Learning





# Five Tribes

- Symbolists
- Connectionists
- Evolutionaries
- Bayesians
- Analogizers



# Symbolists

# Symbolists

- Inverse deduction (deduction)

*Socrates is human*

.....?

*Therefore Socrates is mortal*

- Decision trees

# Symbolists

- Manipulating symbols with an algorithm and choose the ones that work
- It is about learning “rules” that can be applied

# Symbolists

Positives:

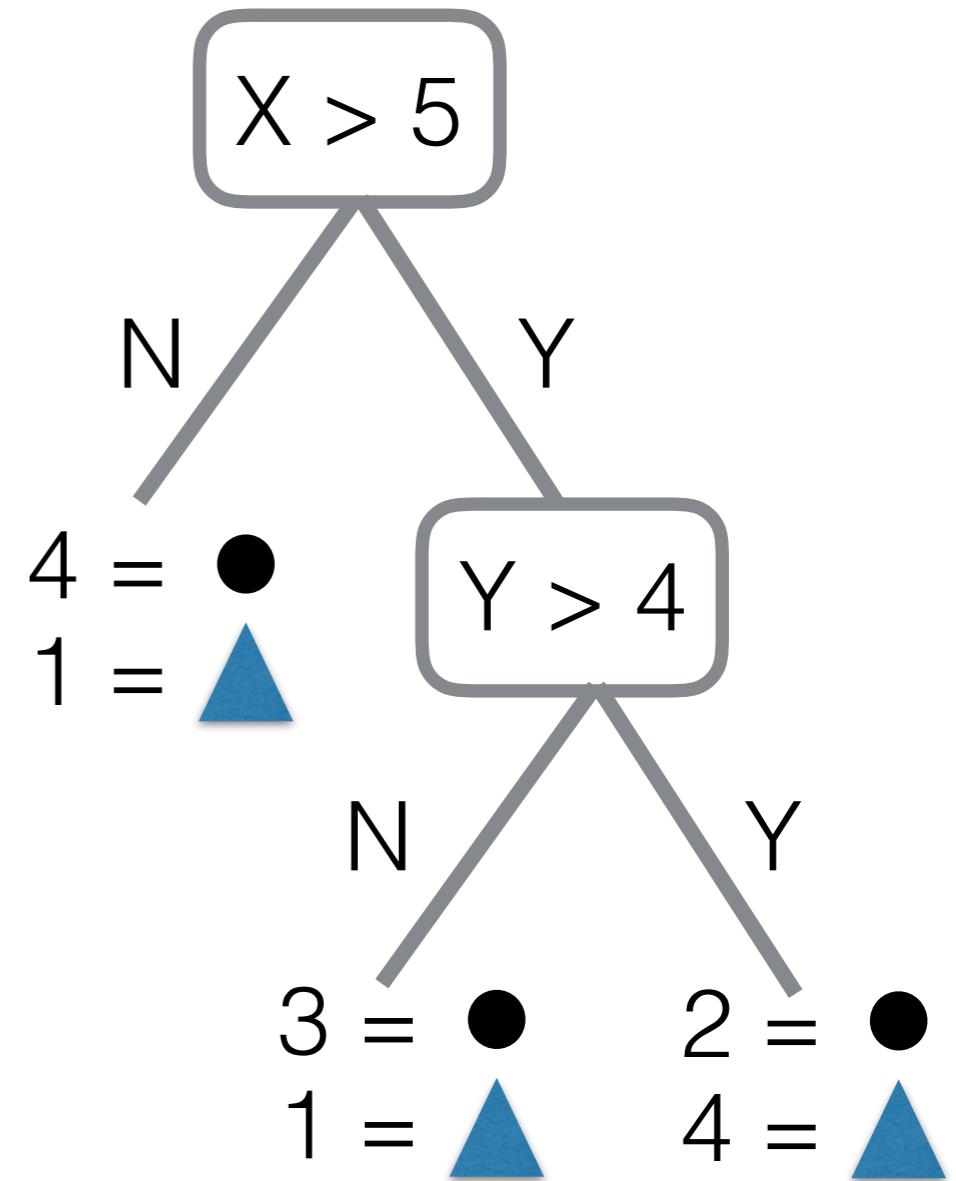
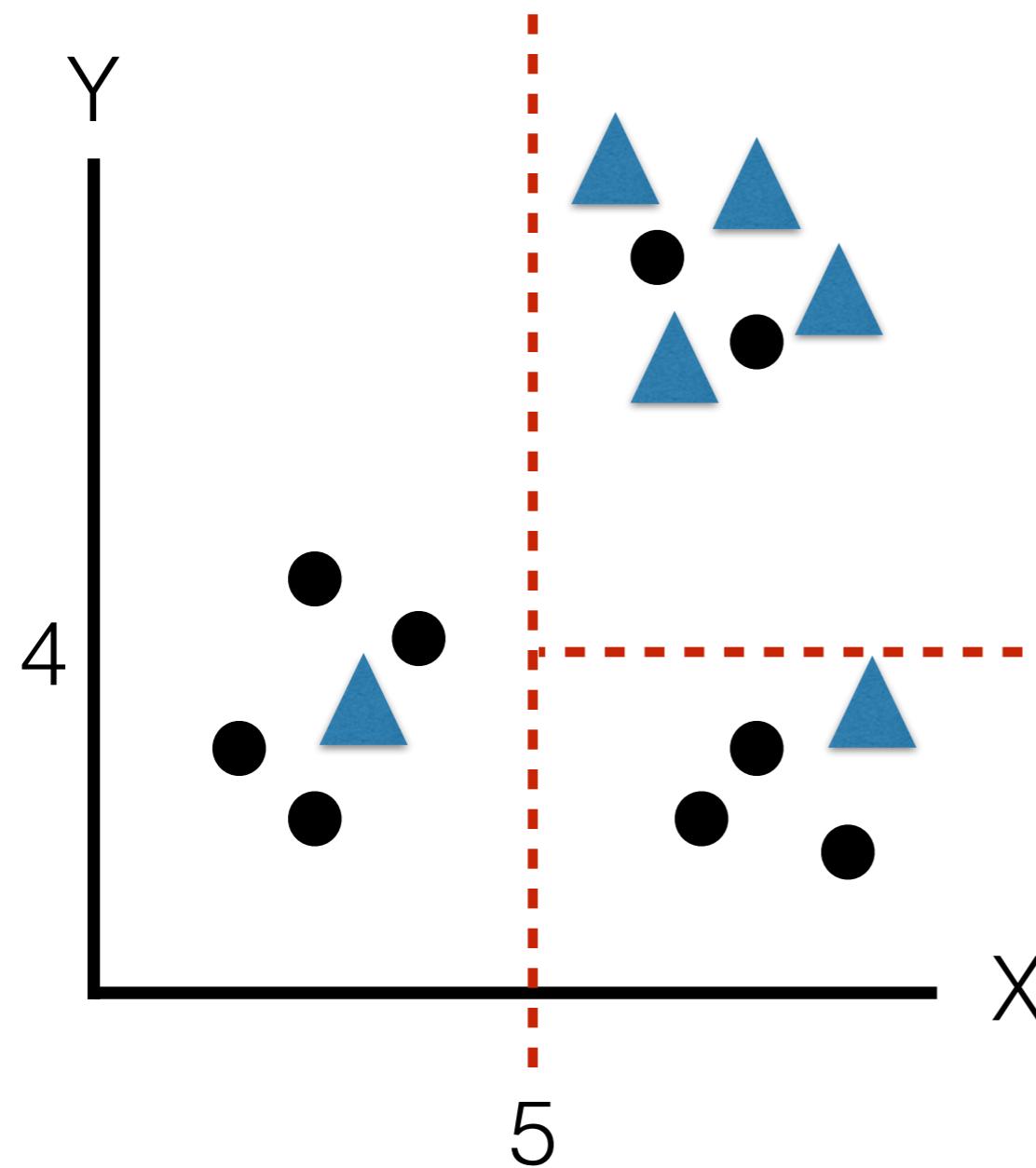
- Easy to interpret
- Fast
- Makes “sense”

Drawbacks:

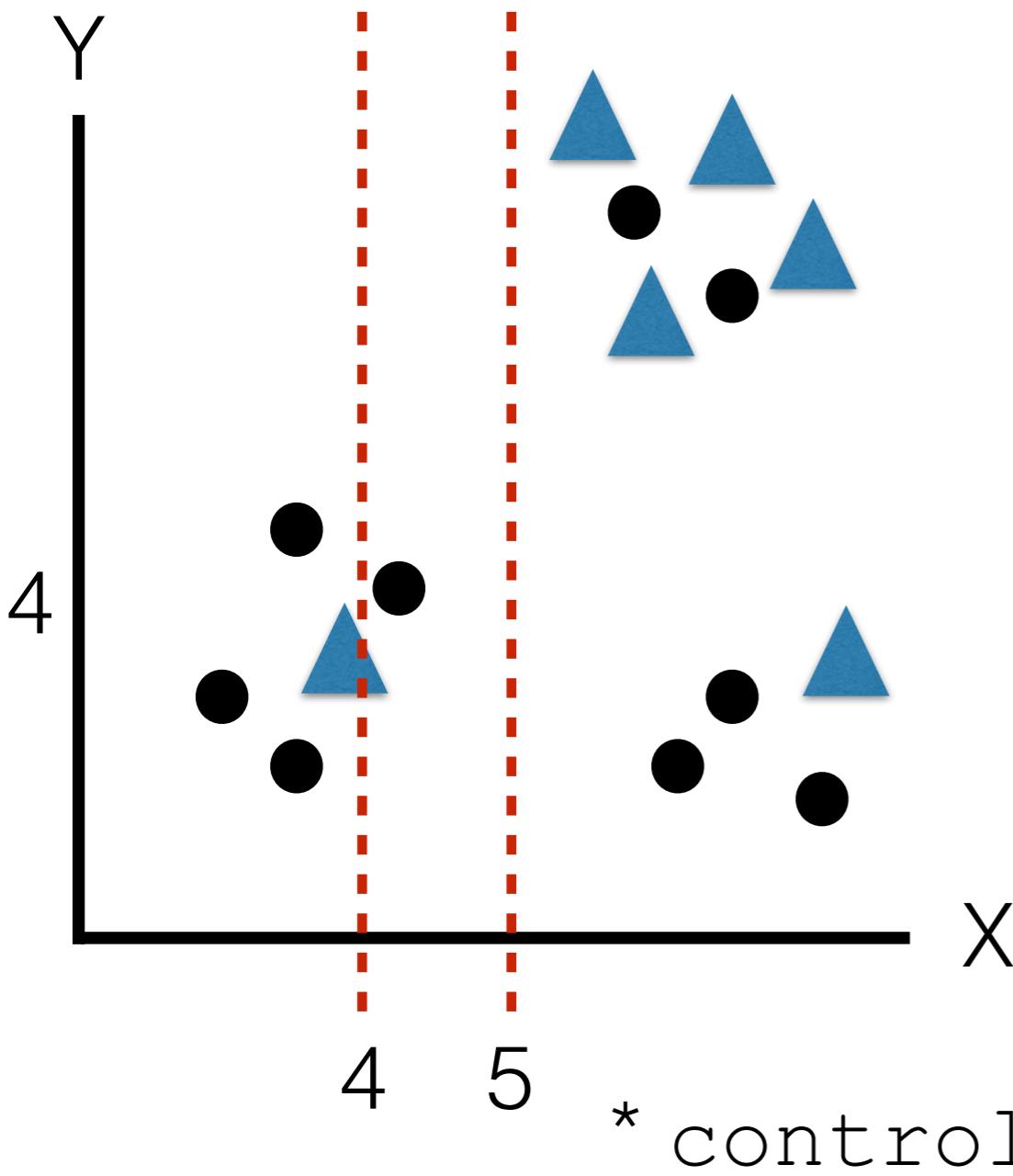
- The number of possible inductions is practically infinite - need to be very specific about the problem space
- If the premise or conclusion are wrong it is over
- Overfitting
- Concepts are rarely cleanly defined: female/male, spam/non-spam - can't incorporate grey areas

# Binary Classification Tree

- \* Minimize the error



# 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

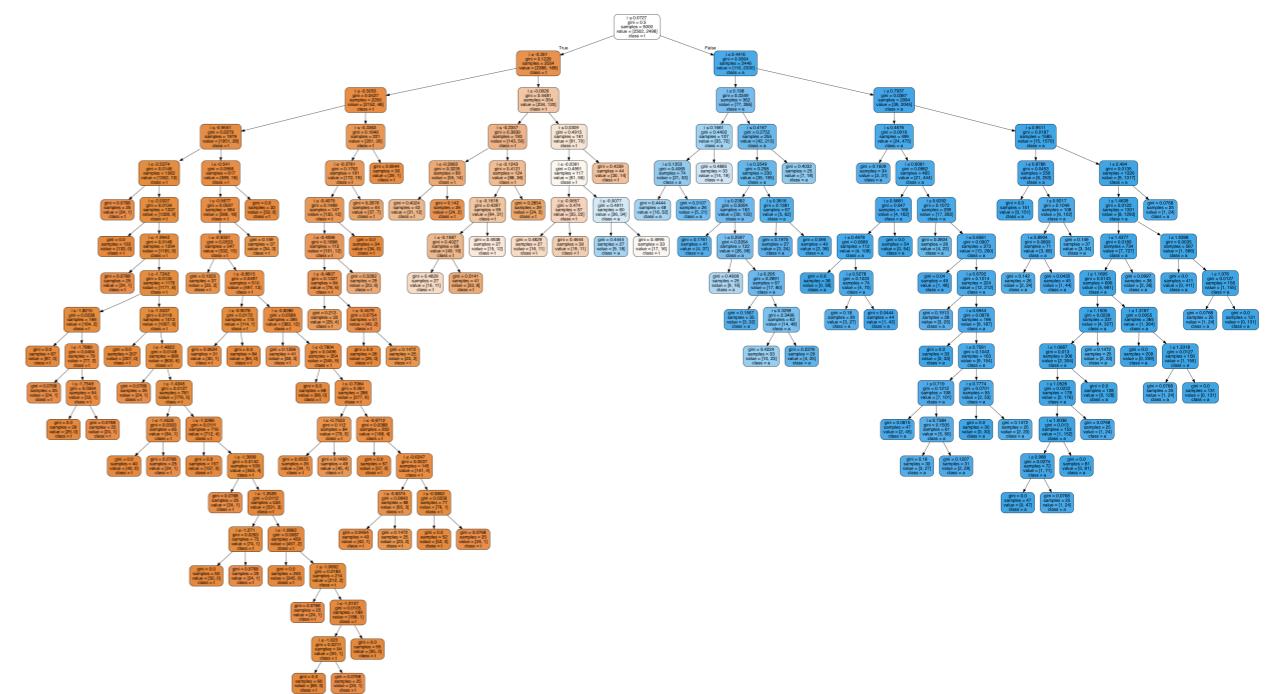
3 = ●  
1 = ▲

6 = ●  
5 = ▲

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

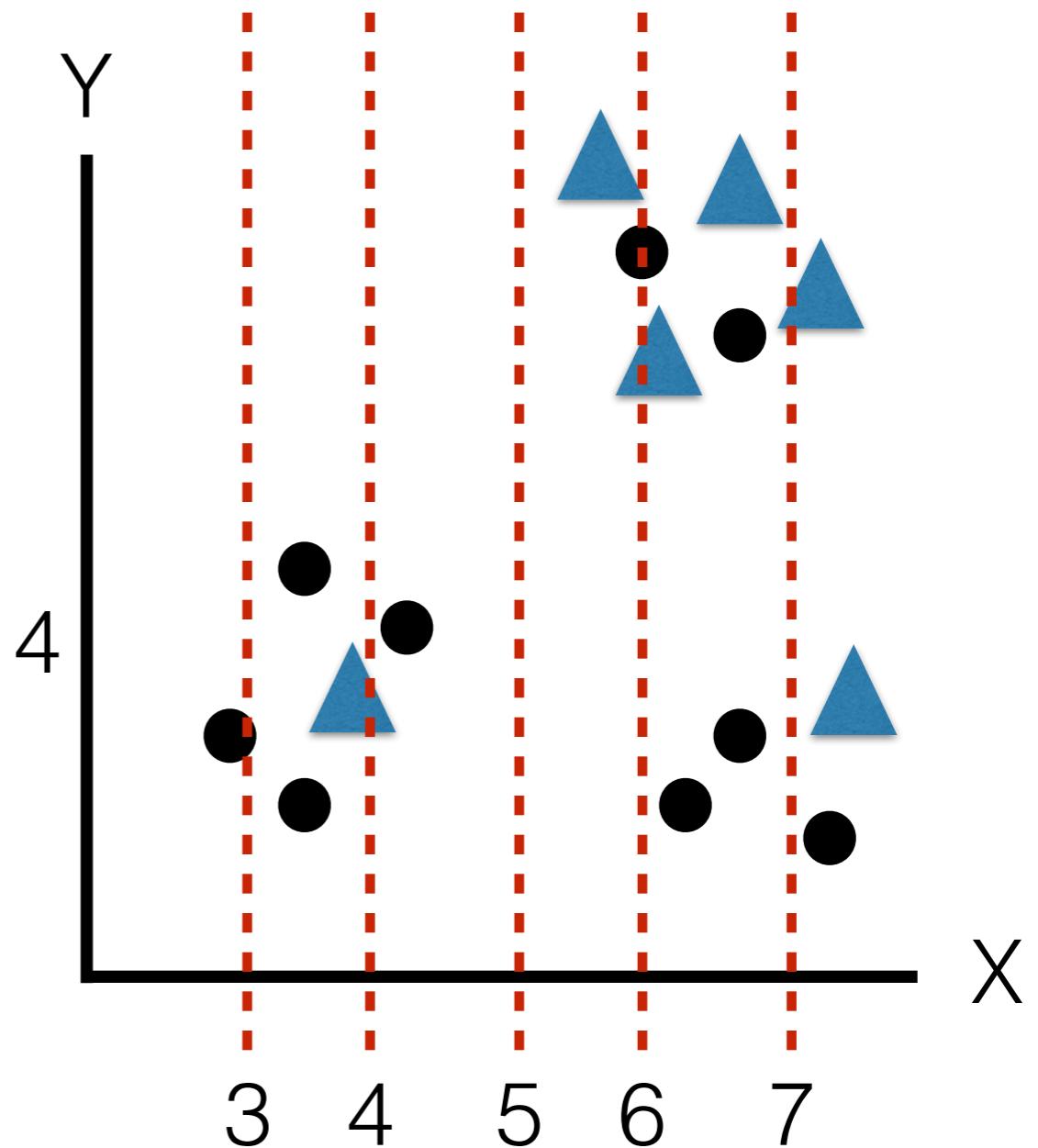
# RPART

- 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



# 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



# Bayesians

# Bayesians

$$P(\theta | \mathbf{D}) = P(\theta) \frac{P(\mathbf{D} | \theta)}{P(\mathbf{D})}$$

- Probabilistic (just grey areas)
- Conditional probabilities shrink the problem space
- Often we know the probabilities of the effects given causes, what we want is the probabilities of the causes given the effects (EG - medical diagnoses)
- Conditional independence assumption

# Bayesians

## Positives

- Computationally simple
- Empirically accurate
- Can handle ambiguity

## Negatives

- Conditional independence assumption
- Susceptible to exponential blowup/Bayesian networks become intractable as variables ↑
- There is no true hypothesis = have to calculate everything
- Can't generate new hypotheses on the fly

# Analogizers

# Analogizers

- Representation = your data
- Find the thing closest to the thing you are looking for: nearest neighbor
- EG - John Snow Cholera Map (1854)
- Collaborative Filters
- k-nearest neighbors, Support Vector Machines, stepwise regression

# Analogizers

## Positives

- Fast and at one time accurate as Neural Nets for complex feature sets OTB
- Can do transfer learning
- High dimensional space works well

## Negatives

- Can't handle class overlap well
- Run time is dependent on data size
- Probabilities are generated by cross validation

# Connectionists

# Connectionists

- Hebb's Rule: neuron's that fire together, wire together
- One concept = many neurons
- Sigmoid curve
- Backpropagation

# Connectionists

## Positives

- Can learn very complex data sets

## Negatives

- Hyperspace is ~infinite, you will likely find a local minima
- Weights are not interpretable
- Can't do adaptive reasoning (rule chaining)

# Evolutionaries

# Evolutionaries

- John Holland (first PhD in CS)
- Objective, program, fitness function, sex
- Selective breeding + immortality
- EG - Spam filter that looks at every word in an email
- Mostly work at the sub-routine level

# Evolutionaries

## Positives

- Combines neural nets with rule based system
- Maybe it can create any kind of machine?

## Negatives

- No empirical reason to have the sex step
  - And maybe a reason not to (mixability)
- Is it the evolutionary nature or just brute force that leads to success?
- Needs a lot of computing power

# caret

- Standard syntax for comparing many models
- Generate training and testing data sets
- Run several model types
- Run resampling algorithms and alter parameters to generate the best model
- Compare using the same diagnostic metrics
- <https://topepo.github.io/caret/>

# caret

## Generate Training/Test Data Sets

```
trainData <- createDataPartition(  
  y = data$thing, ## the outcome data are needed  
  p = .75, ## The percentage of data in the  
  training set  
  list = FALSE)
```

#Generates a list of index numbers for the sample

```
training <- DATA[ trainData, ]  
testing <-DATA[-trainData, ]
```

# caret

## K-Fold Cross Validation

```
ctrl <- trainControl(method = "cv", repeats = 3)
```

# caret

## Train Model

```
fit1 <- train(  
  thing ~ .,  
  data = training,  
  method = "model", ## Center and scale the  
  predictors for the training set and all future  
  samples.  
  preProc = c("center", "scale")  
  trControl = ctrl #add cross validation specs  
  metric = "ROC"  
)
```

# caret

## Test Model

```
pred1 <- predict(fit1, newdata = testing)
```

```
confusionMatrix(data = pred1, DATA$thing)
```

la-process-and-theory/prediction-activity