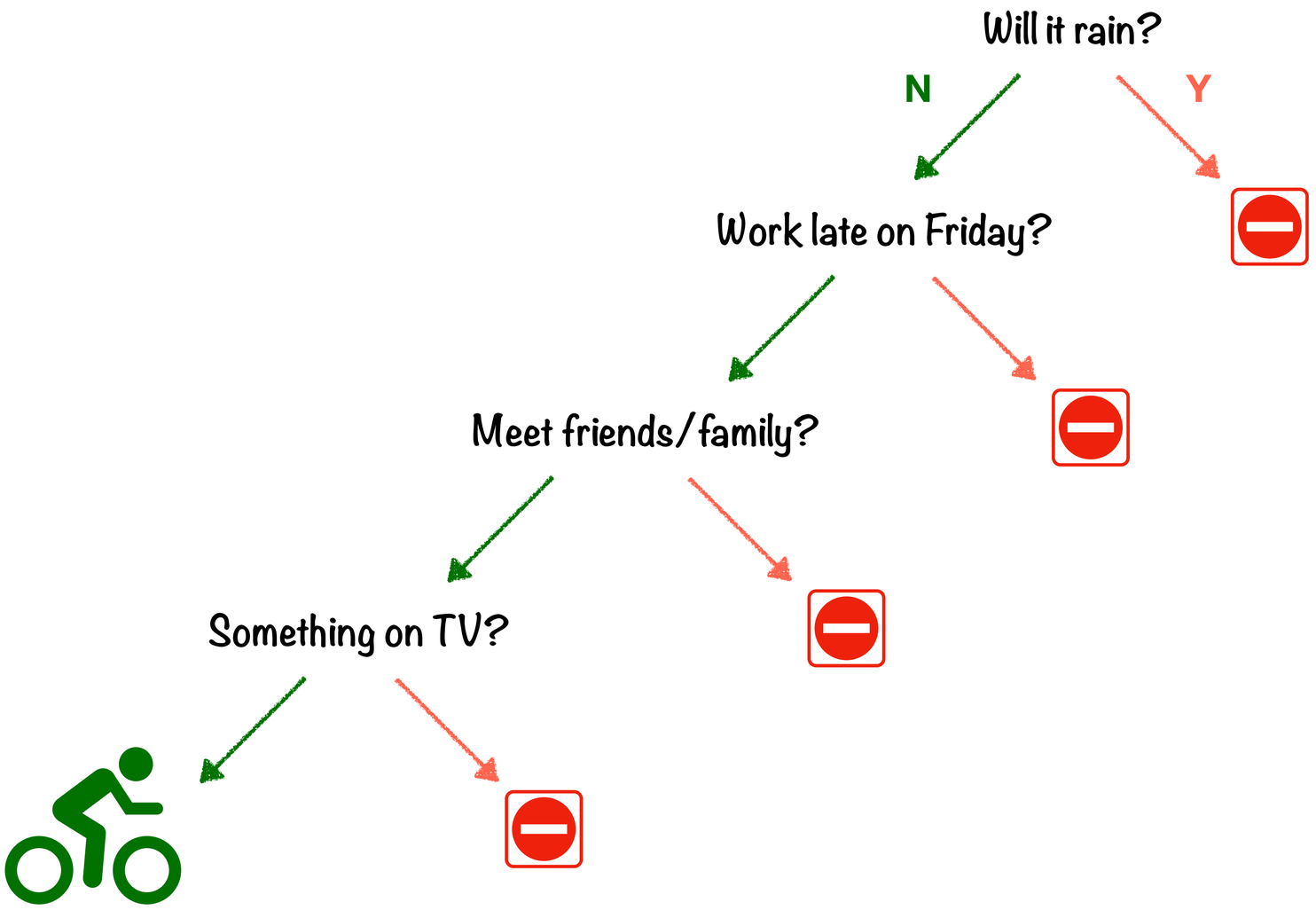
Decision Trees

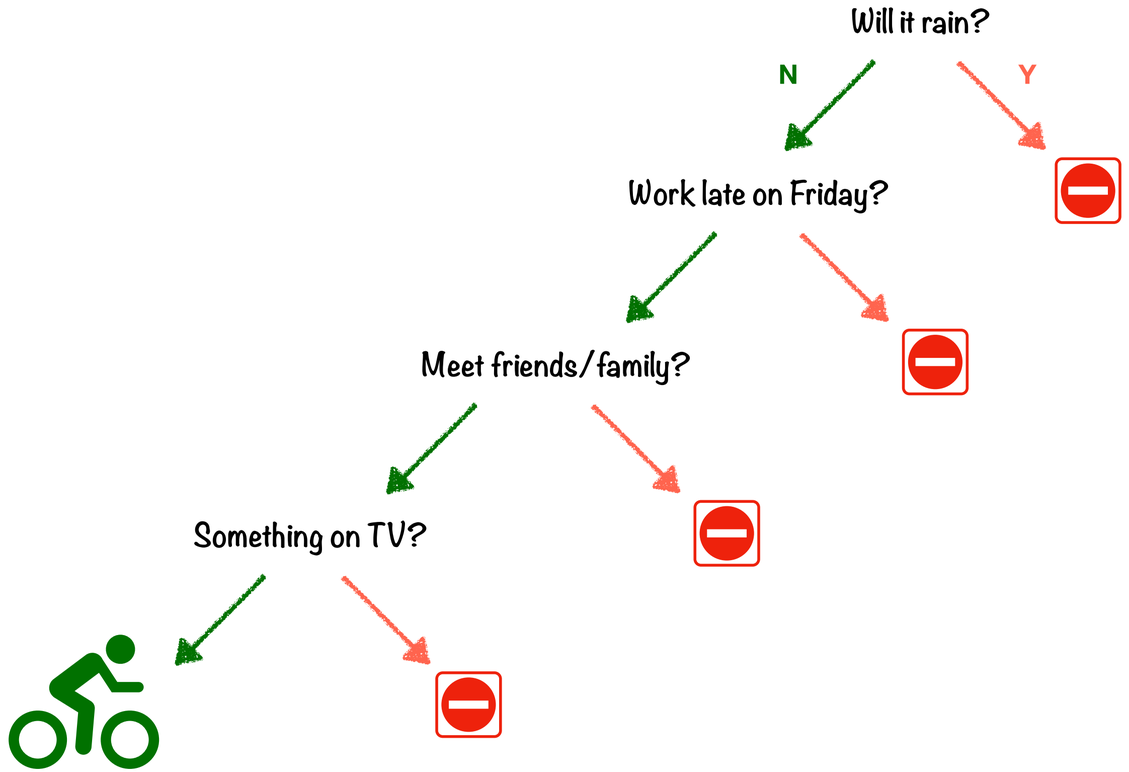
Should I go biking on Saturday morning?

- Will it rain?
- Will I work late on Friday?
- Do I need to meet friends/family?
- Is there something on TV?

Should I go biking on Saturday morning?



Decision Tree

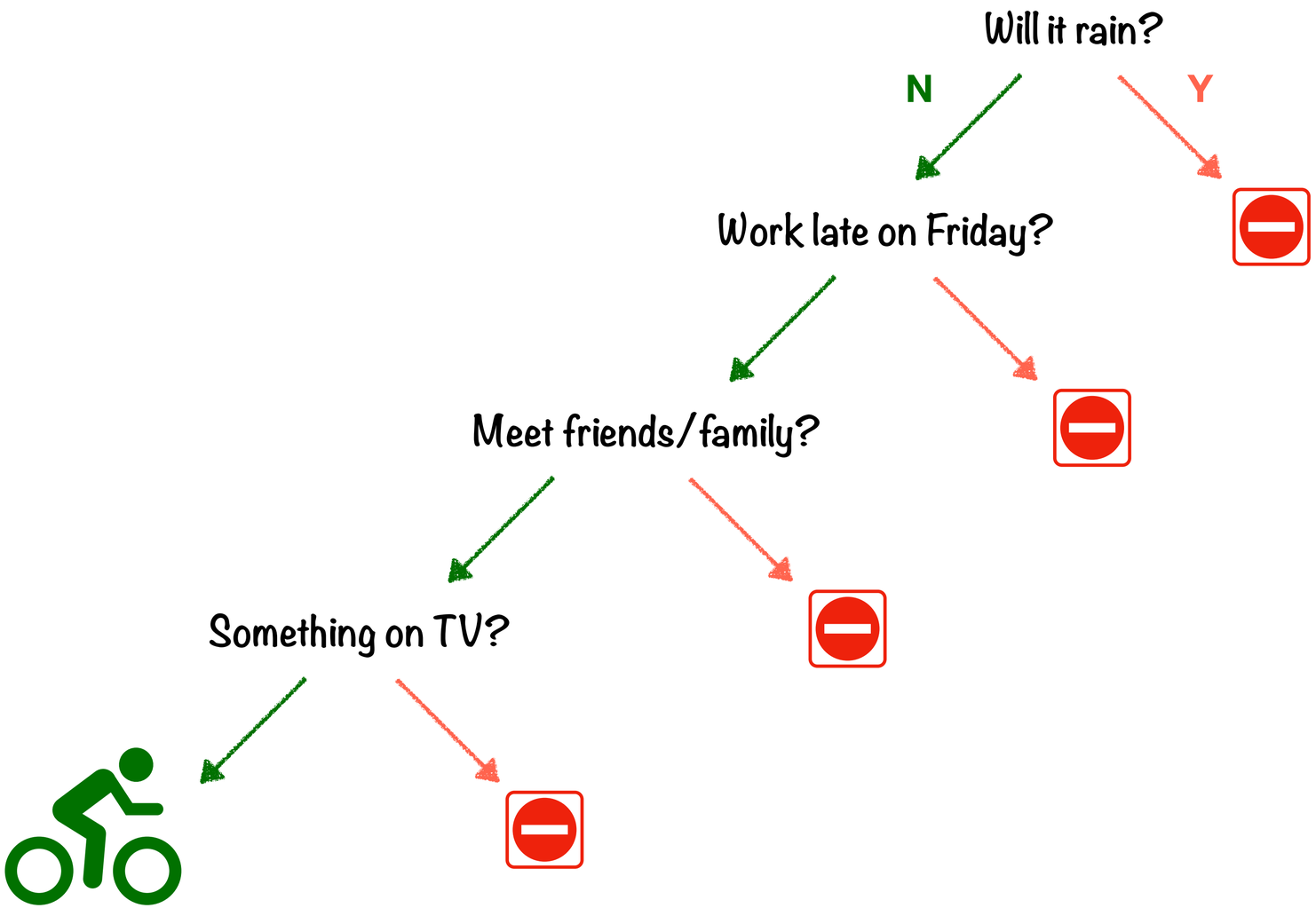


This is exactly what a decision tree looks like

Decision Tree

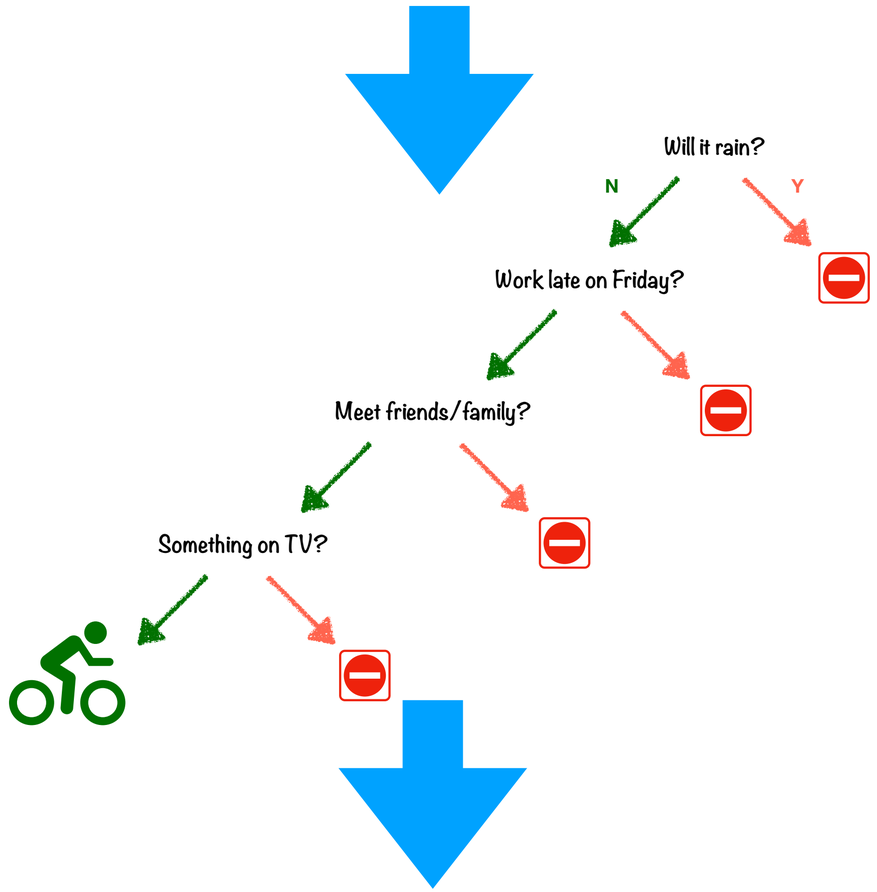
- Helps predict an outcome given a set of inputs
- In business, it represents visually how a decision is taken with inputs and consequences of each decision
- In ML, it predicts the outcome given value of input variables

Decision Tree



Decision Tree

Input Variables/Predictors



Outcome/Output Variables

Decision Tree

- Can be used for classification, just like SVM
- With SVM, can't understand relationship between input variables and outcome
- Decision Trees are not a blackbox

Decision Tree

- Inputs can be categorical or continuous
- Outputs can also be categorical or continuous
  - Called regression tree when outcome is continuous

Will it rain?

N

Y

Work late on Friday?

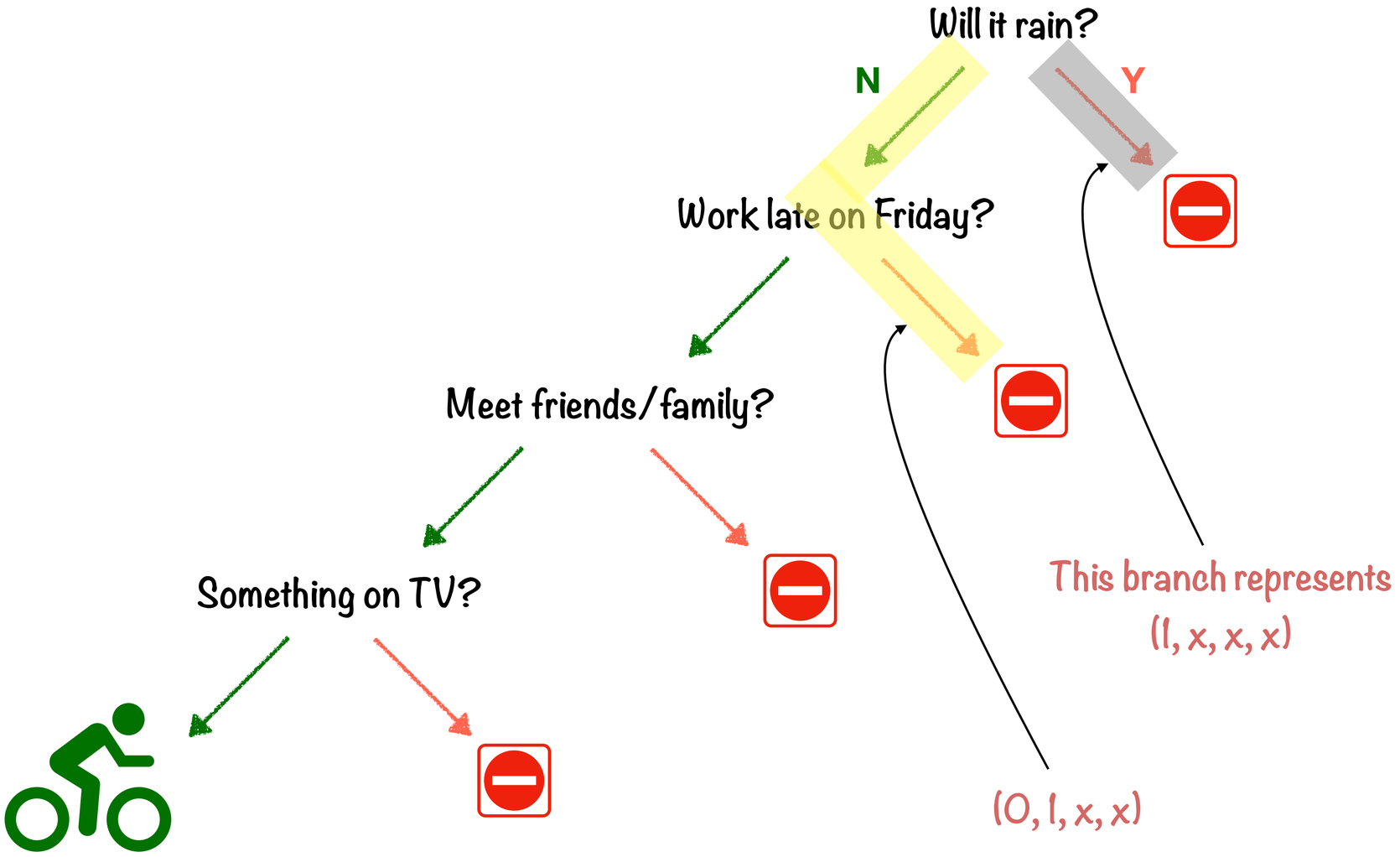
Meet friends/family?

Something on TV?

Decision Tree

- The leaves of a tree are the outcomes
- In a classification problem, these are class labels
- The tree gives the most likely outcome given the values of the variables
- The branches are combinations of predictor values leading to an outcome

Decision Tree



Decision Tree Learning

- The process of creating/learning a decision tree from training data
- Start with training data in the form of feature vector, label/outcome
  - (1,0, 1, 1), No biking
  - (0,0,0,0), I bike!
- Obtain a decision tree used to classify/predict a new instance
- A supervised learning approach

Decision Tree Learning

- Recursive partitioning is the most common strategy for decision tree learning
- Decision tree learning algorithms
  - CART
  - ID3
  - C4.5
  - CHAID

Decision Tree Learning

- Algorithm needs to tell us the order in which the predictors are evaluated
- If the predictor is continuous, the tree needs to split the variable into ranges

Greedy Algorithm for Learning a Decision Tree

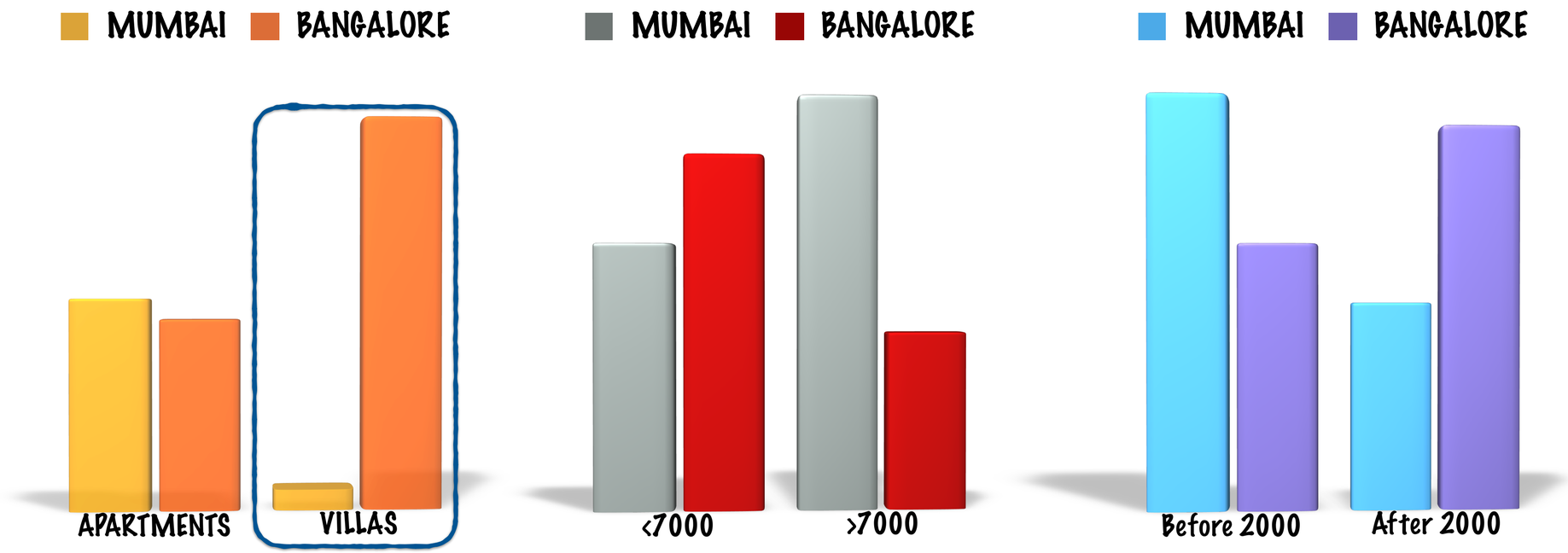
GIVEN THE
TYPE OF HOUSING,
RENT/BEDROOM AND
THE YEAR IT WAS BUILT

PREDICT THE CITY TO
WHICH A RESIDENCE
BELONGS

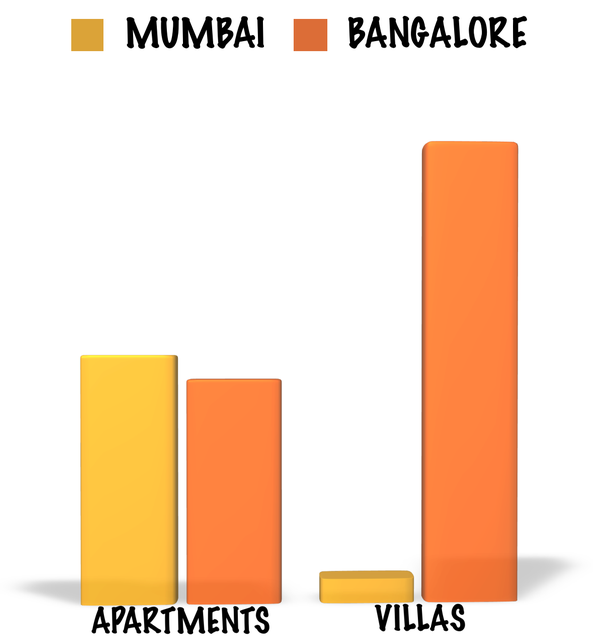
MUMBAI (OR)
BANGALORE

Greedy Algorithm for Learning a Decision Tree

DRAW A HISTOGRAM FOR EACH ATTRIBUTE, FOR RESIDENCES IN EACH CITY



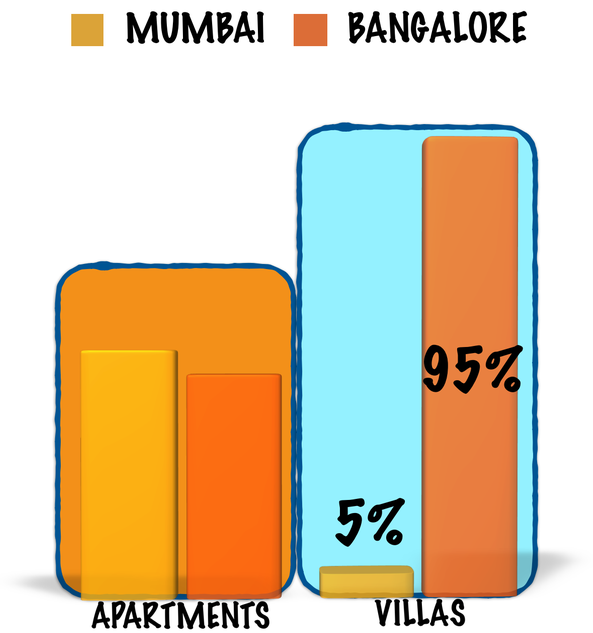
Type **Rent/bedroom** **Year built**



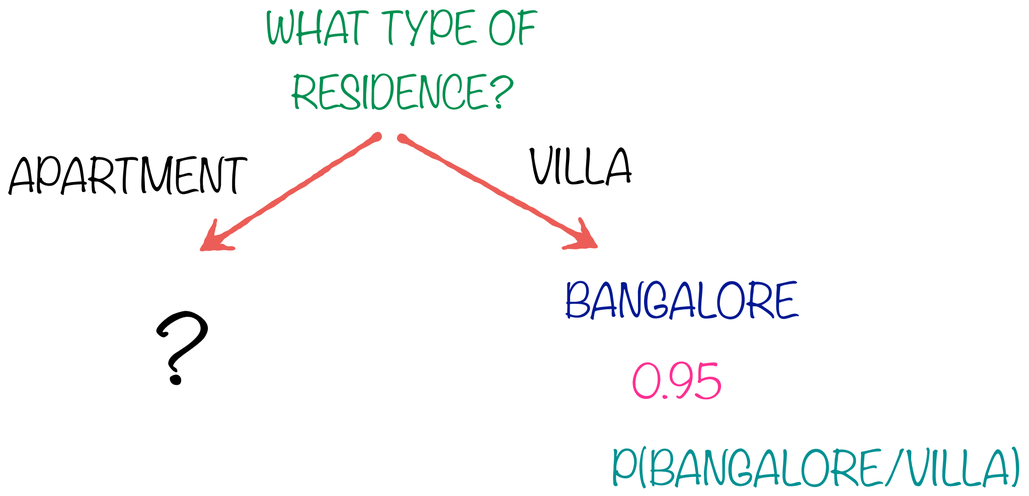
THE TYPE OF HOUSING SEEMS TO BE THE CLEAREST INDICATOR OF WHETHER A RESIDENCE BELONGS TO MUMBAI OR BANGALORE

IF IT IS A VILLA, THEN IT MOST LIKELY BELONGS TO BANGALORE

IF IT IS AN APARTMENT, WE ARE STILL NOT SURE WHICH CITY IT BELONGS TO



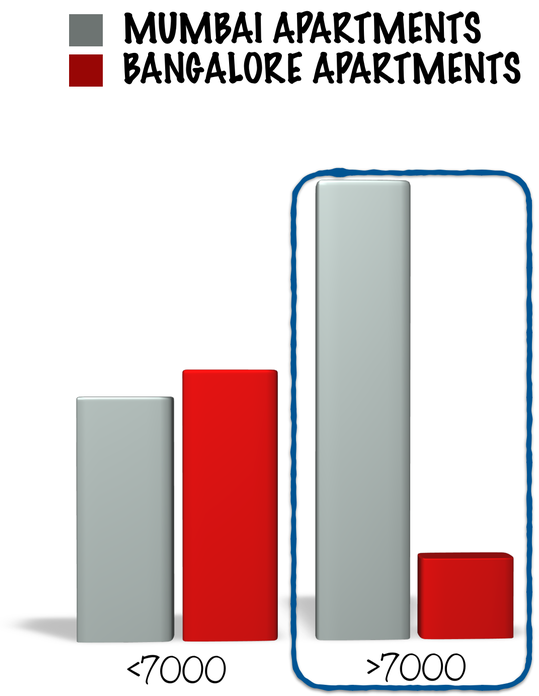
WE NOW HAVE THE FIRST
NODE OF OUR DECISION TREE



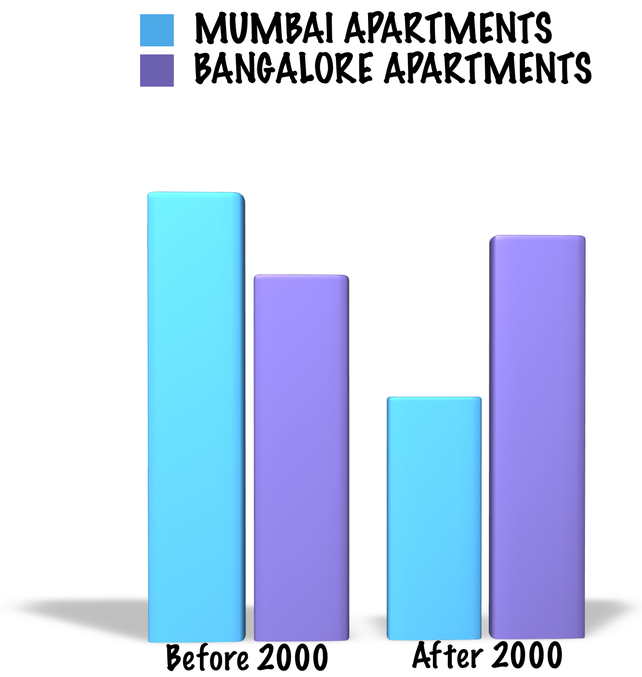
WE'VE DIVIDED OUR DATA INTO
2 SUBSETS - APARTMENTS
AND VILLAS

REPEAT THIS PROCESS RECURSIVELY,
FOR EACH SUBSET

DRAW A HISTOGRAM FOR EACH OF THE REMAINING ATTRIBUTES FOR ONLY THE APARTMENTS IN EACH CITY



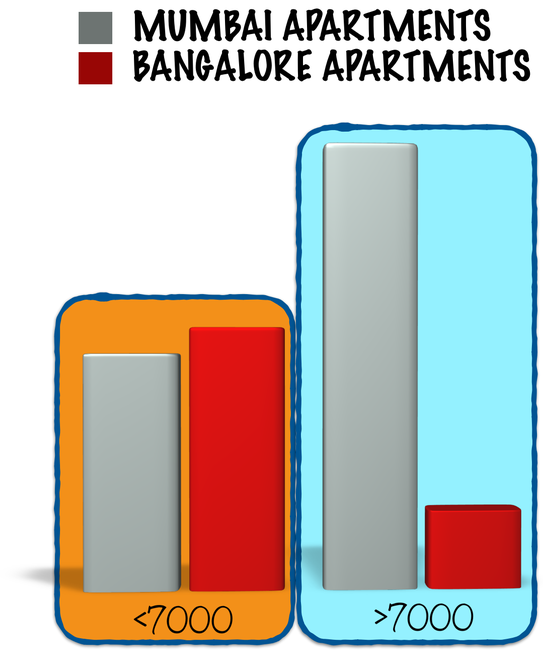
Rent/bedroom



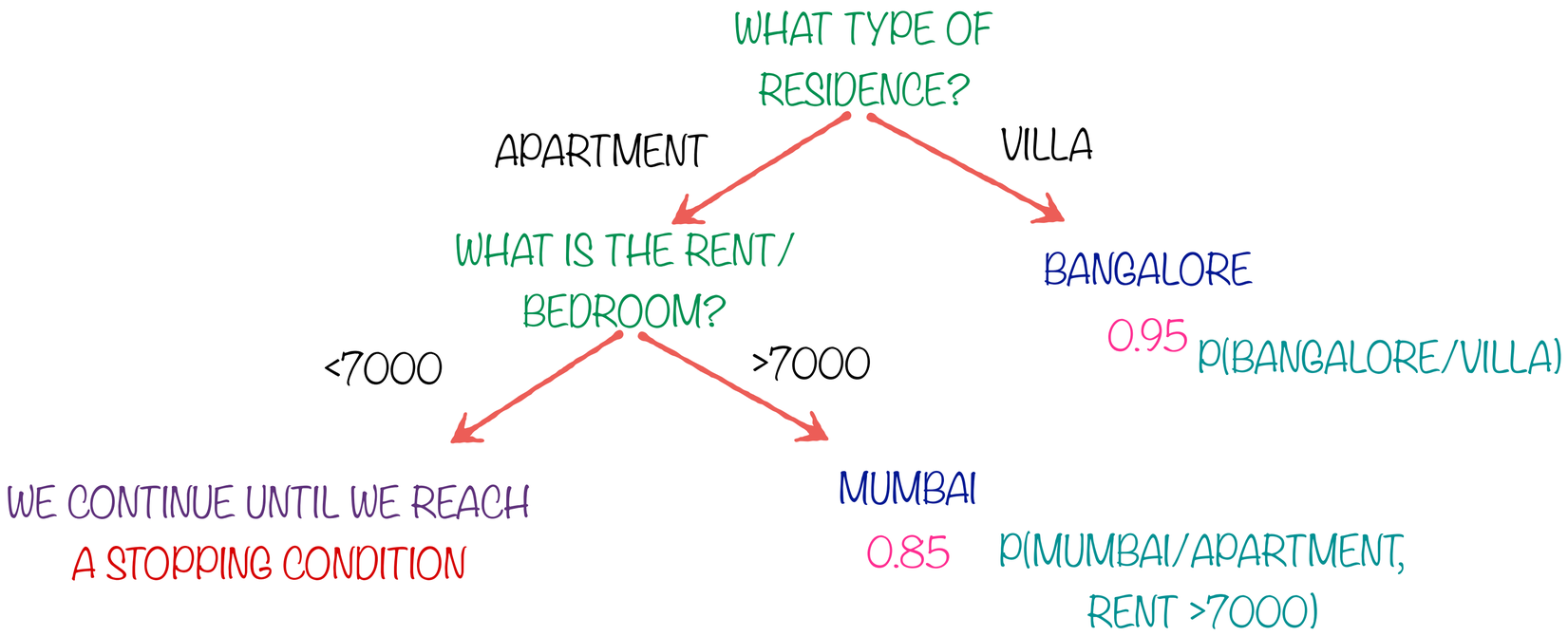
Year built

WHEN YOU LOOK ONLY AT APARTMENTS, THE RENT/ BEDROOM SEEMS TO BE A GOOD PREDICTOR OF THE CITY

NOW, WE'LL DIVIDE THE APARTMENTS INTO TWO SUBSETS, RENT >7000, RENT <7000



Our Decision Tree so far

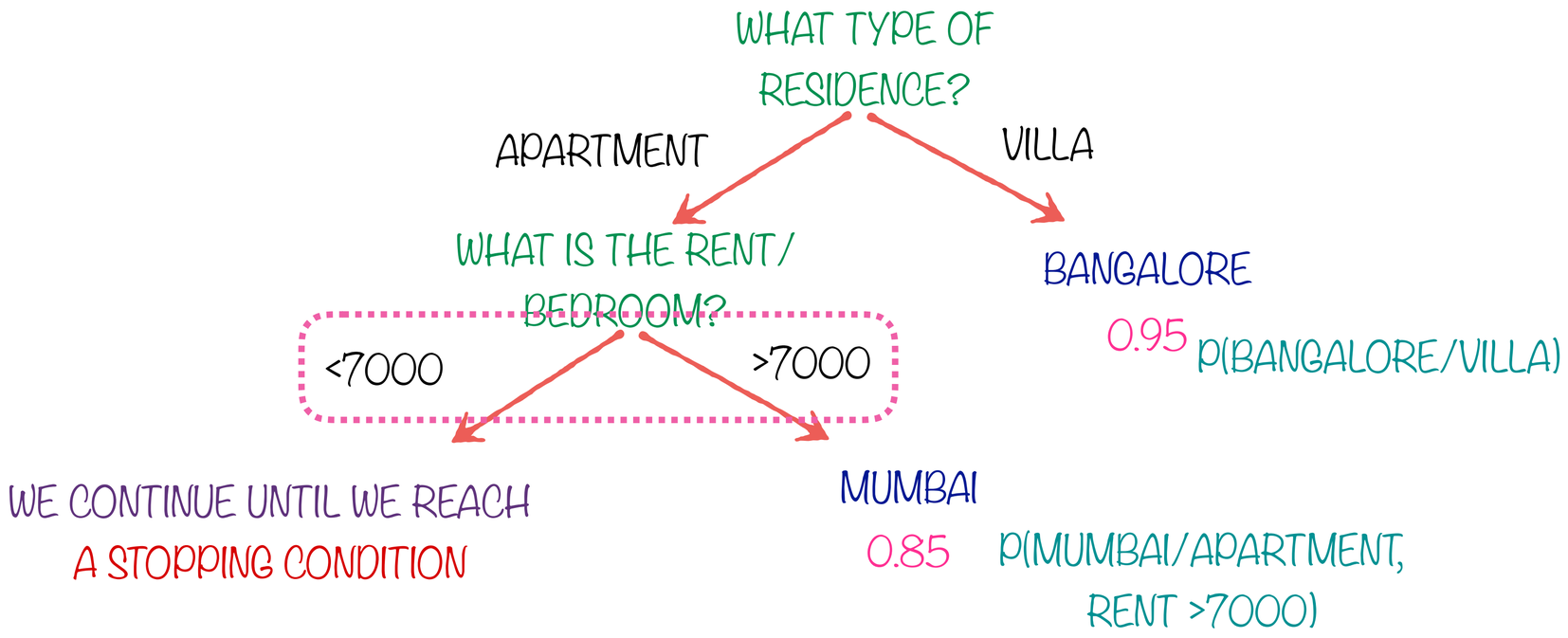


This is Recursive Partitioning

The Stopping Condition could be

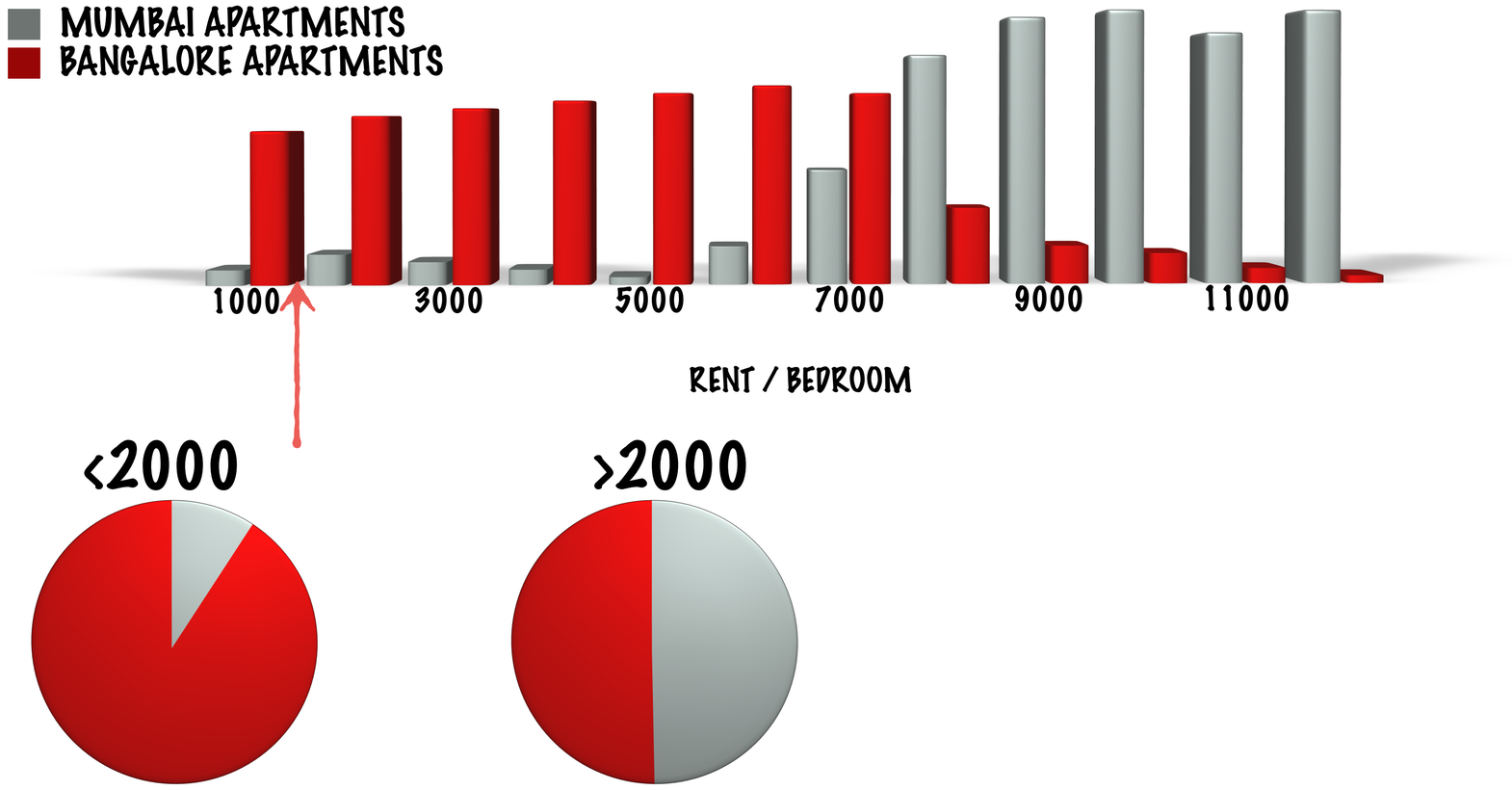
- All our subsets are mostly homogeneous
- We have run out of attributes
- The tree is too large

The Best Split for a Continuous Input Variable

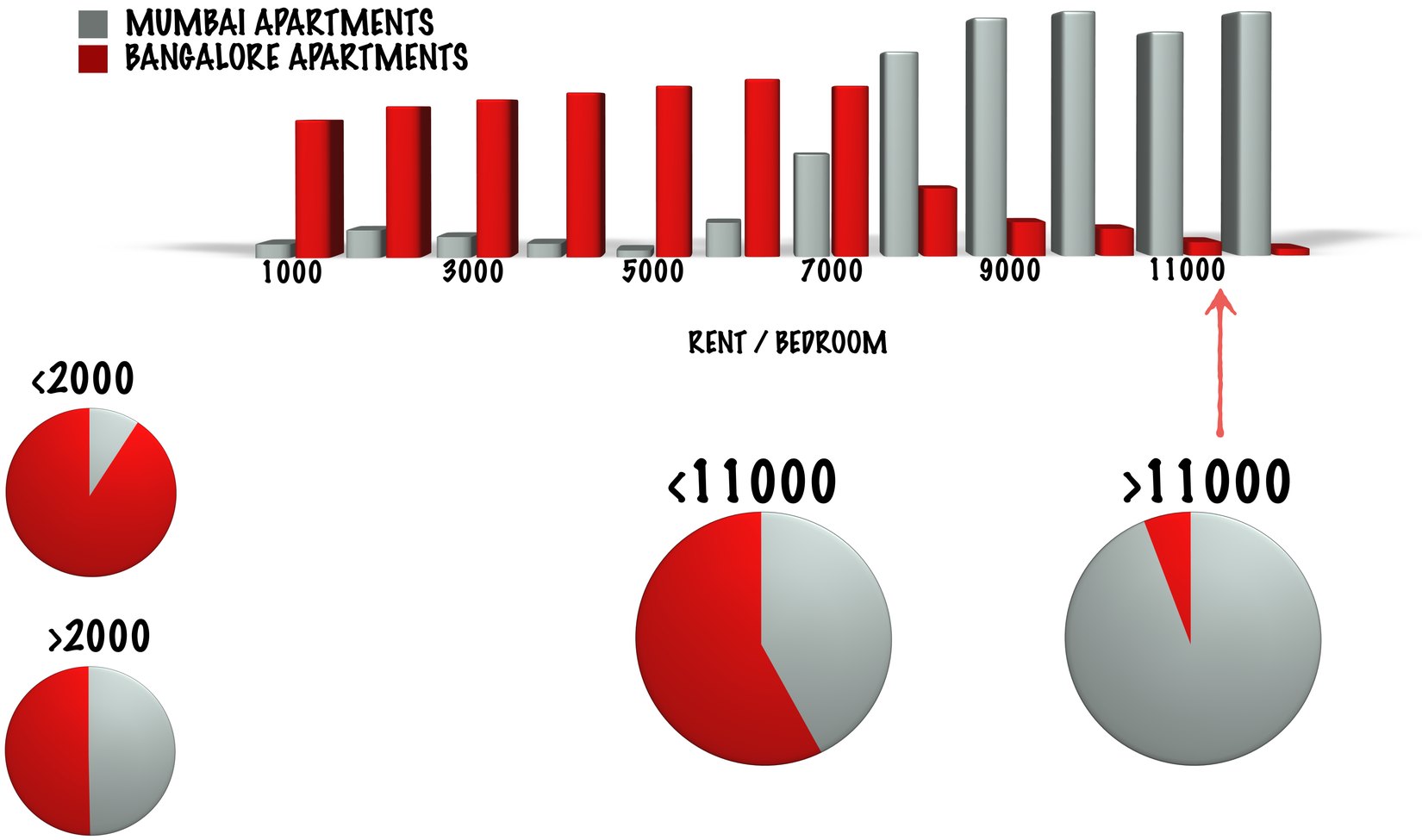


The Best Split for a Continuous Input Variable

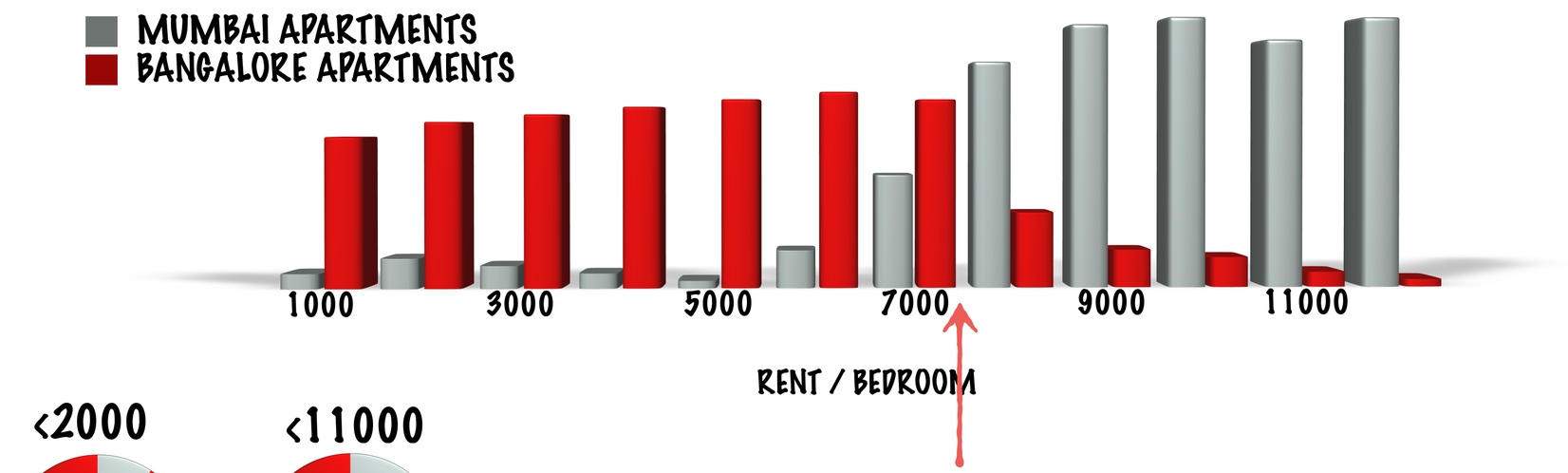
FIRST WE DRAW A HISTOGRAM FOR RENT FOR APARTMENTS IN EACH CITY

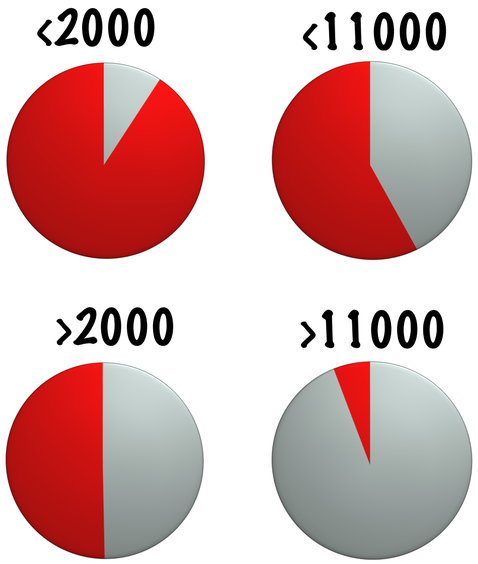


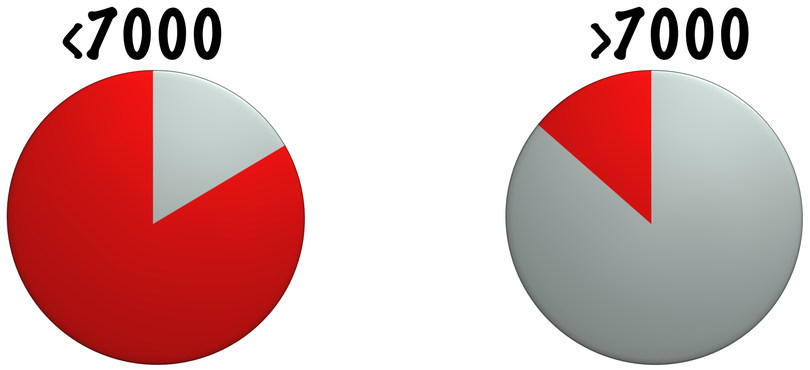
The Best Split for a Continuous Input Variable



The Best Split for a Continuous Input Variable

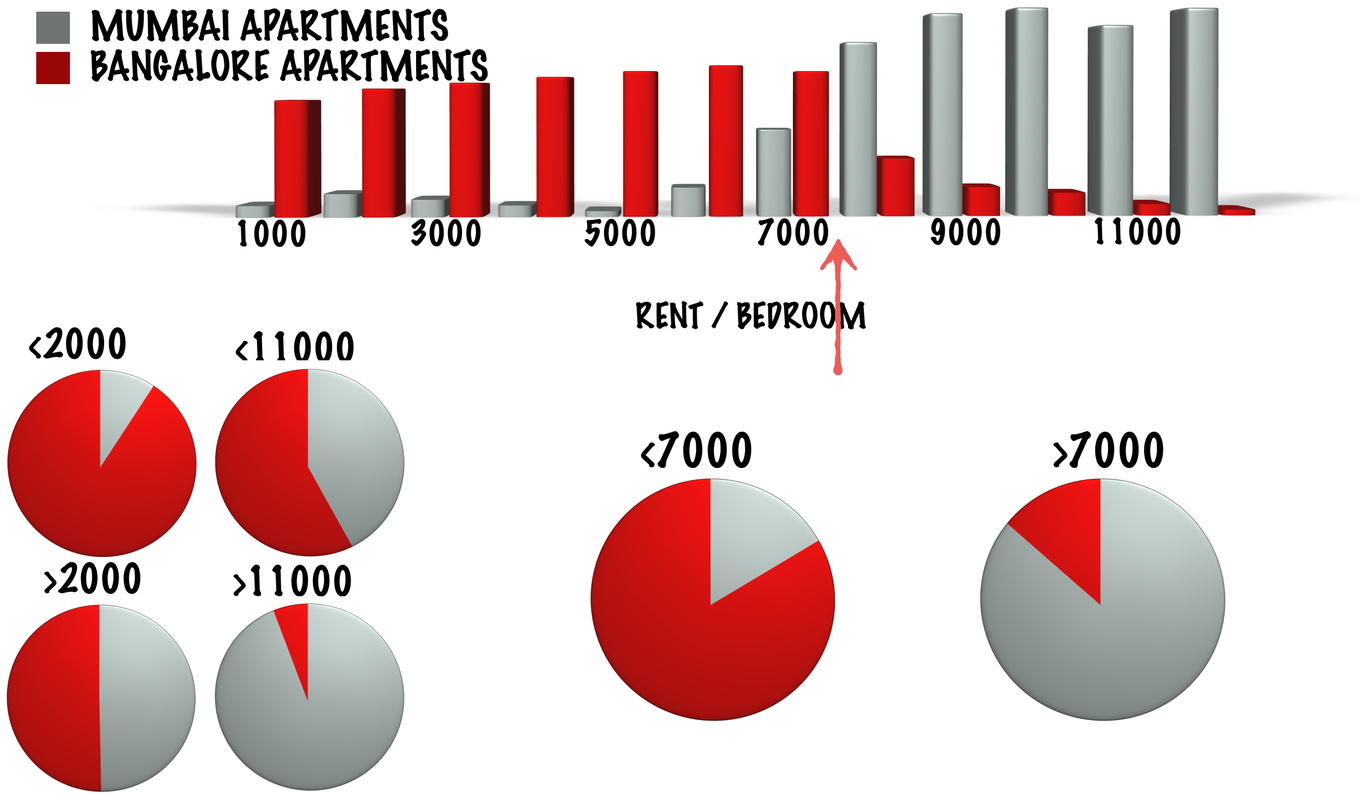






The Best Split for a Continuous Input Variable

The point where the subsets we get are mostly homogeneous



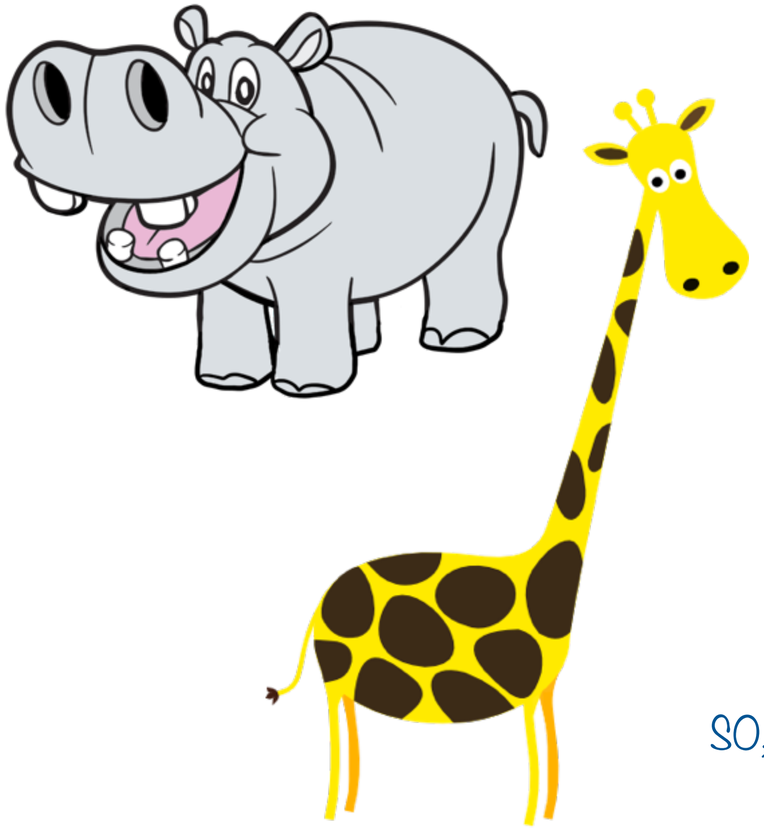
Information Gain

ANY STATEMENT , NEWS OR MESSAGE
CONTAINS INFORMATION

SOME HAVE MORE INFORMATION AND
SOME LESS

THE IDEA OF INFORMATION GAIN IS TO REDUCE
ENTROPY AND MAXIMIZE INFORMATION

LET'S SAY YOU HAVE TO CLASSIFY AN ANIMAL AS A GIRAFFE OR A HIPPO



IF YOU WERE TOLD, THIS ANIMAL HAS 4 LEGS

THIS IS BASICALLY USELESS! BOTH GIRAFFES AND HIPPOS HAVE 4 LEGS, SO THIS STATEMENT GIVES US NO INFORMATION

BUT IF YOU WERE TOLD, THIS ANIMAL IS 10 FEET TALL

THIS IS USEFUL INFORMATION!

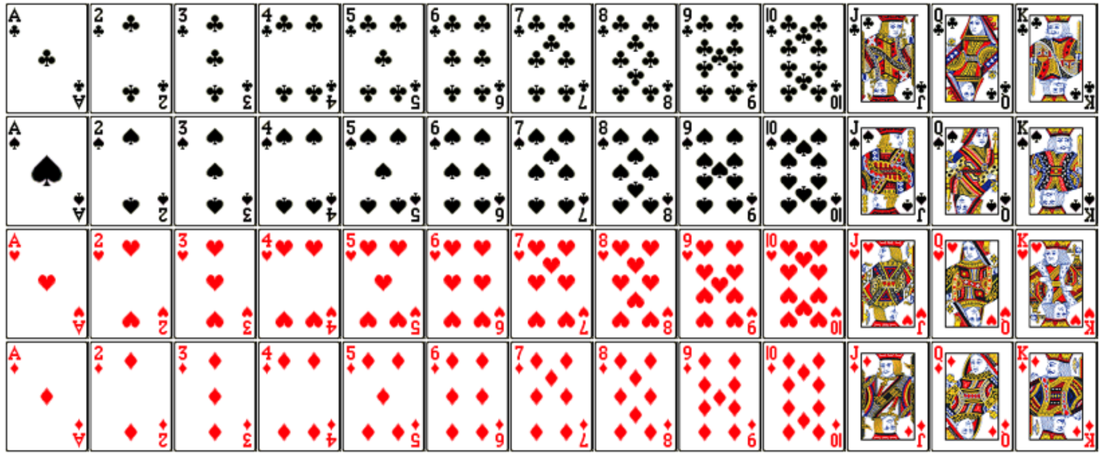
IT TELLS YOU THAT THE ANIMAL IS VERY LIKELY A GIRAFFE

SO, CLEARLY - THE VALUES OF SOME ATTRIBUTES GIVE US MORE INFORMATION THAN OTHERS

AND THERE IS A MATHEMATICAL WAY TO MEASURE THIS INFORMATION

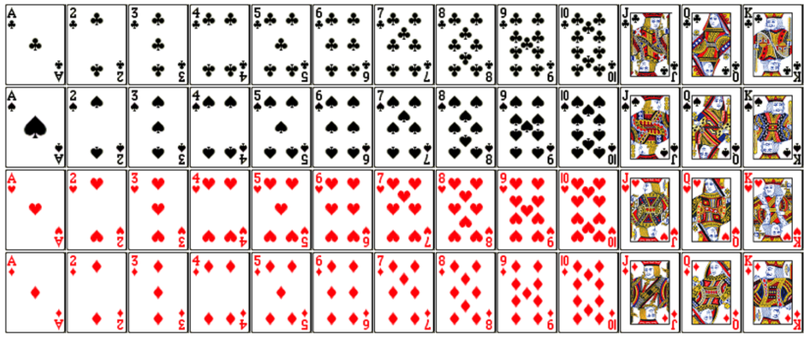
GUESS THE CARD YOUR OPPONENT HOLDS

YOU ARE ALLOWED TO ASK THEM YES/NO QUESTIONS



INITIALLY, THERE ARE 52 POSSIBLE OUTCOMES IN ALL

INITIALLY, THERE ARE 52 POSSIBLE OUTCOMES IN ALL

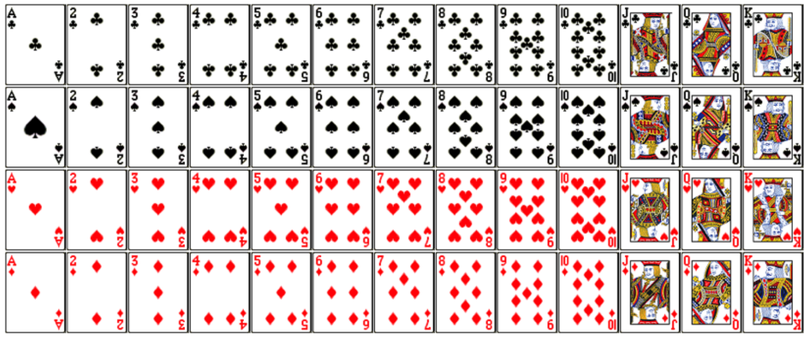


YES

YOU ASK

IS THE CARD AN ACE?

INITIALLY, THERE ARE 52 POSSIBLE OUTCOMES IN ALL



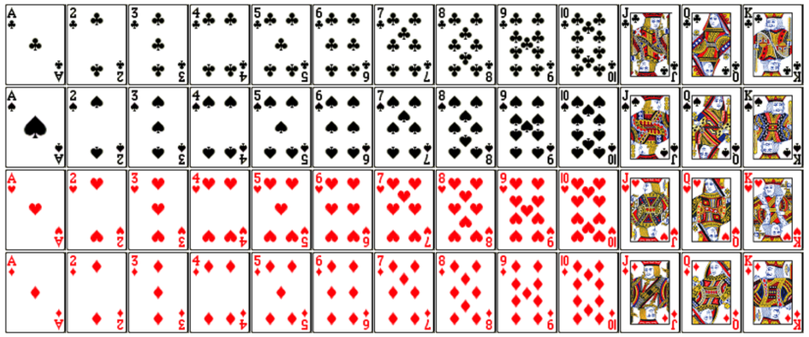
YES

NO

YOU ASK

IS THE CARD AN ACE?

INITIALLY, THERE ARE 52 POSSIBLE OUTCOMES IN ALL



YOU ASK

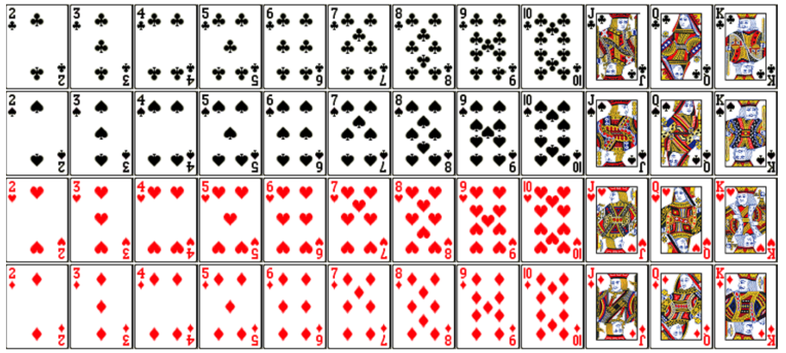
IS THE CARD AN ACE?

YES

NO

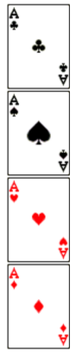
LEFT WITH 4 POSSIBLE OUTCOMES

LEFT WITH 48 POSSIBLE OUTCOMES

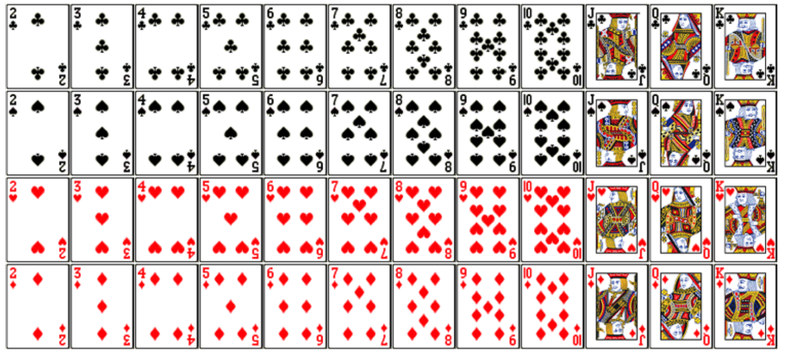


THE ANSWER "YES" GIVES US MORE INFORMATION THAN THE ANSWER "NO"

YES



NO



P(YES)=4/52

P(NO)=48/52

IS THE CARD AN ACE?

THE ANSWER "YES" GIVES US MORE INFORMATION THAN THE ANSWER "NO"

THE ANSWER "YES" HAS A LOWER PROBABILITY

IF X IS A RANDOM VARIABLE THAT REPRESENTS THE ANSWER TO OUR QUESTION

THE LOWER THE PROBABILITY OF THE ANSWER, THE MORE INFORMATION YOU GET

INFORMATION CONTENT OF (X=YES) = -LOG(P(X=YES))

INFORMATION CONTENT OF (X=NO) = -LOG(P(X=NO))

INFORMATION CONTENT OF (X=x) = -LOG(P(X=x))

Information Gain

ANY STATEMENT , NEWS OR MESSAGE
CONTAINS INFORMATION

SOME HAVE MORE INFORMATION AND
SOME LESS

THE IDEA OF INFORMATION GAIN IS TO REDUCE
ENTROPY AND MAXIMIZE INFORMATION

Information Gain

ANY STATEMENT , NEWS OR MESSAGE CONTAINS INFORMATION

SOME HAVE MORE INFORMATION AND SOME LESS

THE IDEA OF INFORMATION GAIN IS TO REDUCE ENTROPY
AND MAXIMIZE INFORMATION

IF YOU WERE TO GUESS, WHAT NEWS YOU'LL HEAR, BEFORE IT HAPPENS

SOME GUESSES ARE EASY

WHAT HOUR WILL THE SUN RISE?

SOME GUESSES ARE HARD

WHAT WILL BE THE RESULT OF A COIN TOSS?

IF X IS A RANDOM VARIABLE THAT REPRESENTS THE ANSWER TO OUR QUESTION

INFORMATION CONTENT OF (X=x) = -LOG(P(X=x))

AVERAGE VALUE OF THE INFORMATION CONTENT (ALSO CALLED THE EXPECTED VALUE)=

$$\Sigma P(X=x)\,(-\log(P(X=x)))$$

ENTROPY H(X)

ENTROPY IS THE AMOUNT OF UNCERTAINTY/UNPREDICTABILITY THERE IS IN THE ANSWER

ENTROPY INCREASES WITH

1) NUMBER OF POSSIBLE ANSWERS
2) THE EVENNESS OF THE PROBABILITY DISTRIBUTION

P(YES) = 0 => THERE IS NO UNCERTAINTY => ENTROPY =0

YES AND NO HAVE EQUAL PROBABILITY => VERY HIGH ENTROPY

INFORMATION CONTENT OF (X=x) = -LOG(P(X=x))

ENTROPY H(X) = AVERAGE VALUE OF THE INFORMATION CONTENT (ALSO CALLED THE EXPECTED VALUE)= $\sum P(X=x)\,(-\log(P(X=x))$

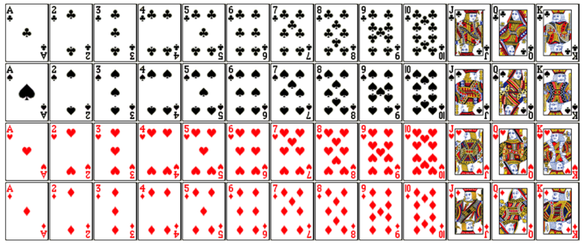
THE GAME IS TO ANSWER THE QUESTION

WHICH CARD DO YOU HOLD?

BEFORE WE HAVE ASKED ANY YES/NO QUESTIONS, THE UNCERTAINTY (ENTROPY) IN OUR GUESS IS VERY HIGH

INITIALLY, THERE ARE 52 POSSIBLE OUTCOMES IN ALL

EACH HAS SAME PROBABILITY =1/52



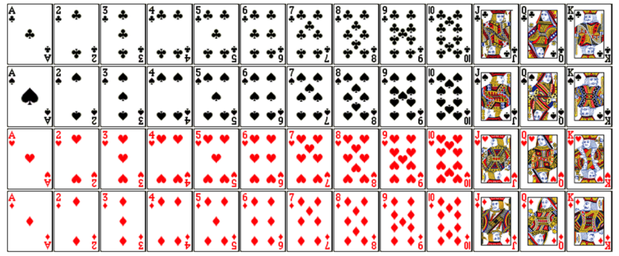
ENTROPY=H(X)=

$\sum(1/52)(-\log(1/52)=$

$\log(52)$

THERE ARE 52 POSSIBLE OUTCOMES IN ALL EACH HAS PROBABILITY =1/52

BEFORE WE HAVE ASKED ANY YES/NO QUESTIONS, THE UNCERTAINTY (ENTROPY) IN OUR GUESS IS VERY HIGH



ENTROPY=H(X)= LOG(52)

ONCE YOU ASK

IS THE CARD AN ACE?

YES

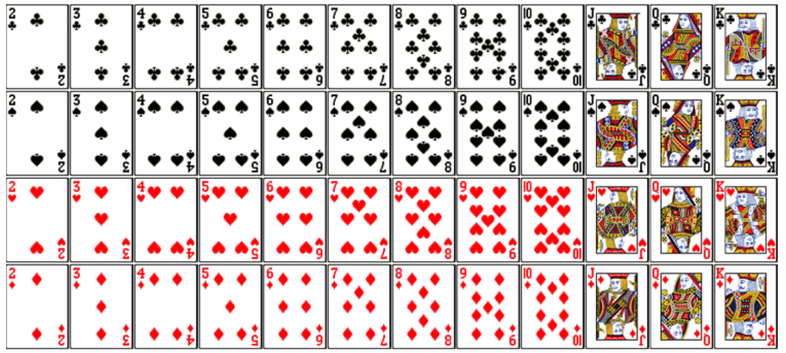
NO

ENTROPY = H(X/QI=YES)= LOG(4)

P(YES)=4/52

ENTROPY = H(X/QI=NO)= LOG(48)

P(NO)=48/52



WITHIN EACH GROUP, THE ENTROPY HAS DECREASED

THE MORE HOMOGENOUS EACH GROUP IS, THE LOWER THE ENTROPY

BEFORE WE HAVE ASKED ANY YES/NO QUESTIONS, THE UNCERTAINTY (ENTROPY) IN OUR GUESS IS VERY HIGH

ENTROPY = H(X) = LOG(52)

ONCE YOU ASK IS THE CARD AN ACE?

ENTROPY = H(X/Q1=YES)= LOG(4)

P(YES)=4/52

ENTROPY = H(X/Q1=NO)= LOG(48)

P(NO)=48/52

ENTROPY AFTER Q1 = H(X/Q1) =

P(YES)*H(X/Q1=YES)+P(NO)*H(X/Q1=NO) =

4/52* LOG(4) + 48/52 * LOG(48)

BEFORE WE HAVE ASKED ANY YES/NO QUESTIONS, THE UNCERTAINTY (ENTROPY) IN OUR GUESS IS VERY HIGH

ENTROPY = H(X)

ONCE YOU ASK IS THE CARD AN ACE?

ENTROPY AFTER Q1 = H(X/Q1)

INFORMATION GAIN = REDUCTION IN ENTROPY OVERALL = H(X) - H(X/Q1)

AS YOU SAW, WHENEVER YOU ASK A QUESTION, SUBSETS ARE FORMED

WHEN EACH OF THOSE SUBSETS ARE HOMOGENOUS, THE INFORMATION GAIN IS MAXIMUM

ENTROPY = H(X)

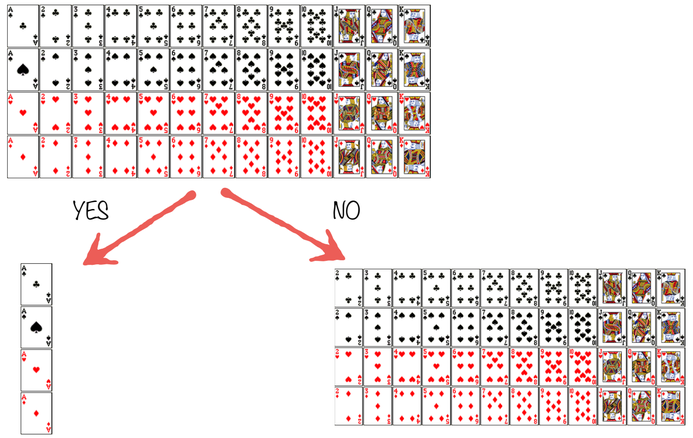
ENTROPY AFTER Q1 = H(X/Q1)

INFORMATION GAIN = REDUCTION IN ENTROPY OVERALL = H(X) - H(X/Q1)

AS YOU SAW, WHENEVER YOU ASK A QUESTION, SUBSETS ARE FORMED

WHEN EACH OF THOSE SUBSETS ARE HOMOGENOUS, THE INFORMATION GAIN IS MAXIMUM

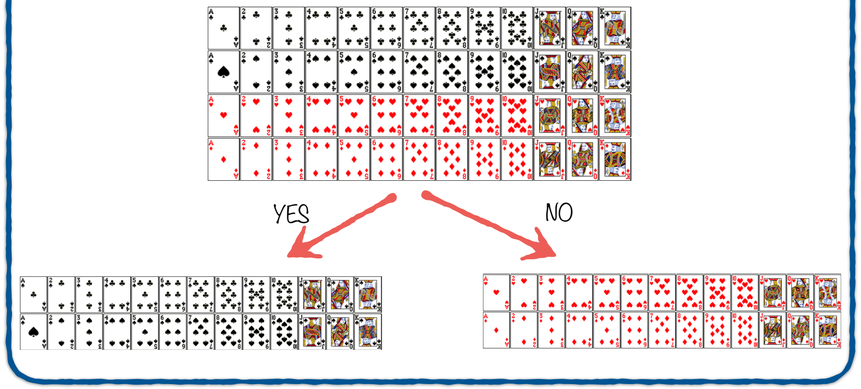
IS IT AN ACE?



H(X/Q1) = 4/52* LOG(4) + 48/52 * LOG(48)

IG= H(X)-H(X/Q1) = 0.12

IS IT A BLACK?



H(X/Q1) = LOG(26)

IG= H(X)-H(X/Q1) = LOG(52)-LOG(26) =0.30

Decision Tree Learning

- Recursive partitioning is the most common strategy for decision tree learning
- Decision tree learning algorithms
  - CART
  - ID3
  - C4.5
  - CHAID

EACH HAS A SLIGHTLY DIFFERENT WAY OF ARRIVING AT THE BEST ATTRIBUTE (OR) MEASURING THE HOMOGENEITY OF A SUBSET

Decision Tree Learning

- Recursive partitioning is the most common strategy for decision tree learning
- Decision tree learning algorithms
  - CART
  - ID3
  - C4.5
  - CHAID

INFORMATION GAIN

EACH HAS A SLIGHTLY DIFFERENT WAY OF ARRIVING AT THE BEST ATTRIBUTE (OR) MEASURING THE HOMOGENEITY OF A SUBSET

Decision Tree Learning

- Recursive partitioning is the most common strategy for decision tree learning
- Decision tree learning algorithms
  - CART
  - ID3
  - C4.5
  - CHAID

GINI IMPURITY

CART IS ANOTHER DECISION TREE LEARNING METHOD

(CLASSIFICATION AND REGRESSION TREES)

IT USES A DIFFERENT WAY TO CHOOSE AN ATTRIBUTE

MINIMIZING GINI IMPURITY

THE IDEA BEHIND, GINI IMPURITY IS SIMPLE

CHOOSE THE ATTRIBUTE SUCH THAT - IF YOU STOP THE DECISION TREE WITH THAT ATTRIBUTE AND GO NO FURTHER

THE PROBABILITY OF A FALSE LABEL IS MINIMIZED