

Data Science

CS301

Decision Trees

Week 11

Fall 2024

Oliver BONHAM-CARTER

Are you here today?!

ATTENDANCE

<https://forms.gle/iaY7zBmxj8KvsDMa8>



Another Topic in Machine Learning

- We have discussed linear and multivariate regression that predict after training with historical data
- Now we will discuss Decision Trees which are used to categorize data based on another type of training.
- Decision trees are a statistical method used to cluster data – to create categories by splitting data.



Kinds of Research?

1. **Predicting customer churn in a subscription-based service:** A decision tree model can help identify key factors influencing whether a customer is likely to cancel their subscription, such as billing issues, service quality problems, or competitor offerings.
2. **Diagnosing diseases based on symptoms:** In medical applications, decision trees are used to help doctors diagnose diseases by identifying patterns in symptoms that lead to specific conditions. For example, a decision tree could help predict whether a patient has pneumonia, based on factors like cough type, chest pain, and fever.
3. **Credit risk assessment for lending institutions:** Decision trees can be employed to analyze an applicant's credit history, income, employment status, and other relevant factors to determine the likelihood of defaulting on a loan. This helps lenders make more informed decisions about extending credit.



Kinds of Research?

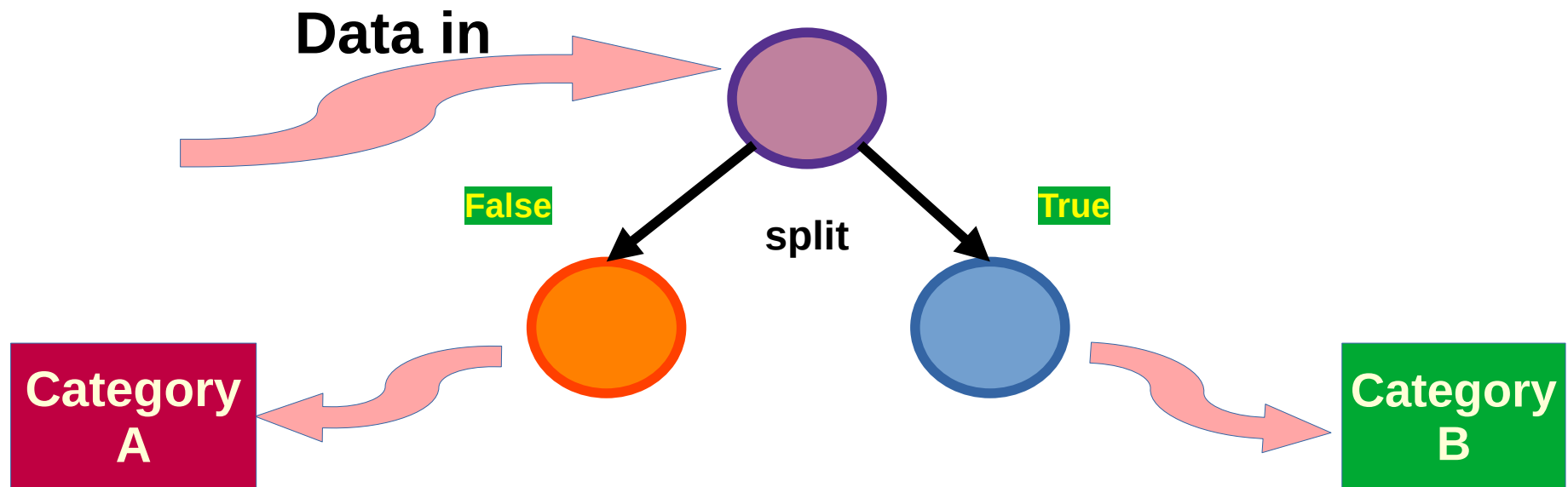
4. **Marketing campaign optimization:** By examining customer data, decision trees can help marketers identify which factors influence whether a customer responds positively to a particular advertising campaign. This information can be used to optimize future campaigns for greater effectiveness.

5. **Stock price prediction and trading strategies:** Decision trees can be utilized to analyze historical market data and make predictions about future stock prices or trends. These models can help traders identify profitable opportunities based on factors like economic indicators, news events, and technical analysis.

6. And many others!

What Are Decision Trees?

- Supervised Machine Learning Models: Using pre-existing labeled data to train and predict outcomes (*categories*) based on input features
- A system to divide data by pushing it through a tree where conditions determine the divisions.
- Decision making based on historical data

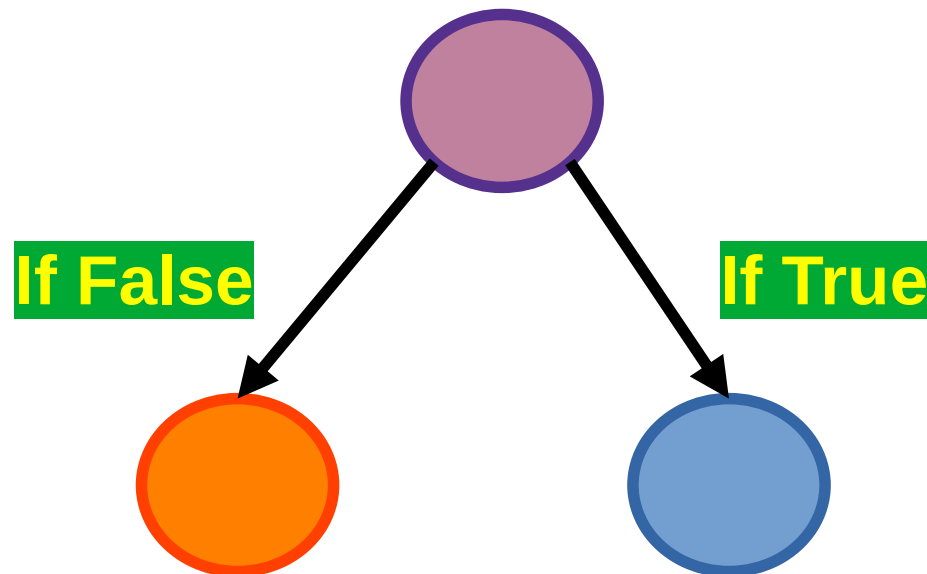




Starting With Decision Trees

Defining the Parts

- What role does the Root play – which leading decision does this node play?
- What roles do the internal nodes (decisions) or leaf nodes (categories) play?
- How to divide tree – how to make categories from data?
- How to measure the accuracy of tree splits?

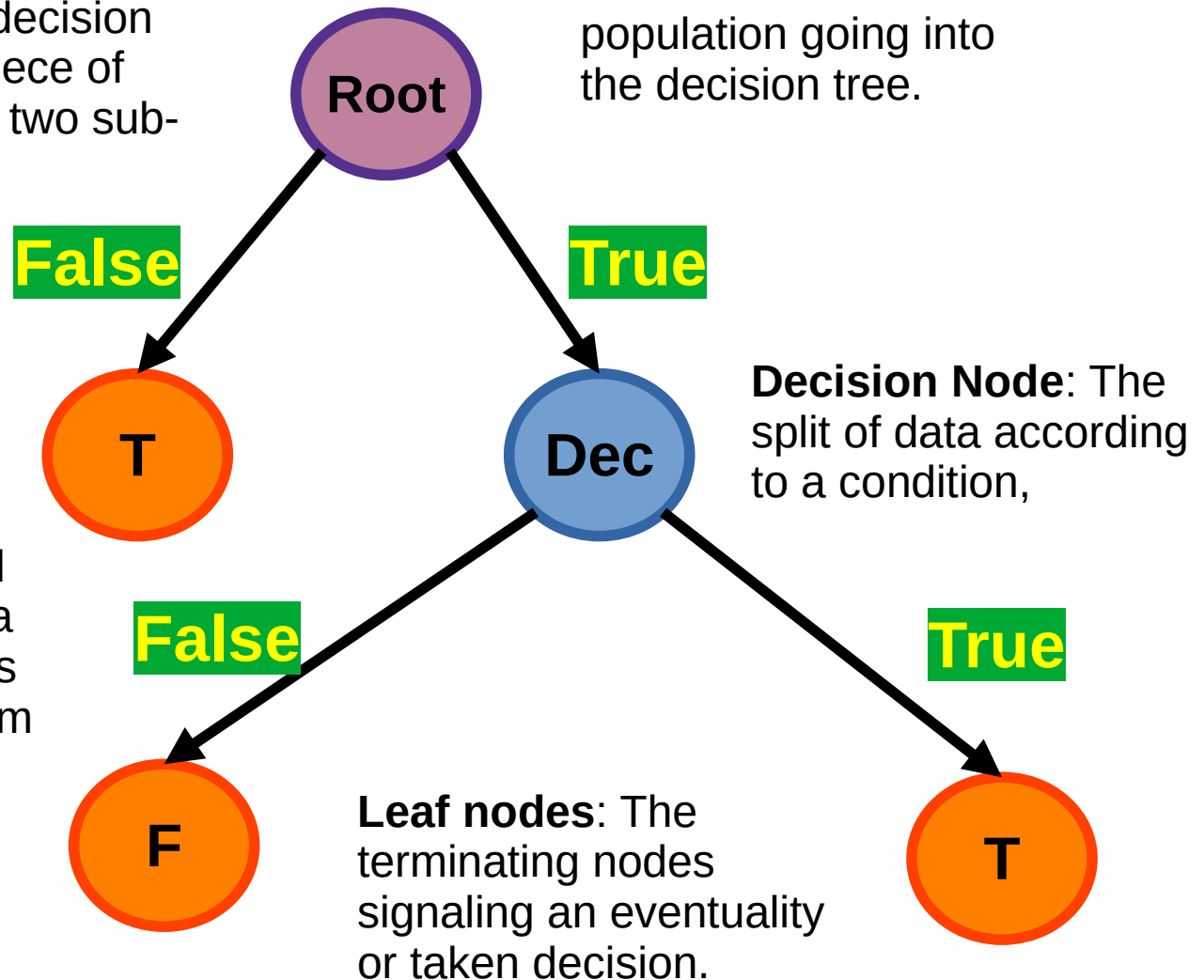




Decision Tree Terms

Node Splitting: After a decision of True or False, the a piece of data must go into one of two sub-nodes,

Root Node: The entire population going into the decision tree.



Example: Pack an Umbrella?

- Want to decide to take an umbrella on a walk during potential rainy day.
- I decide whether to take the umbrella at all.
 - *Too many things in my arms makes carrying an umbrella **cumbersome**. Will it even rain? Do I care if I get wet?*
- ***How many eventualities (categories) are there to create using historical data?***





The Data

N	IsRaining?	WalkingDistance	ThingsCarrying	TakeUmbrella?
1	1	200	4	0
2	1	750	9	1
3	0	500	3	0
4	0	1000	7	0

How many things am
I already carrying?

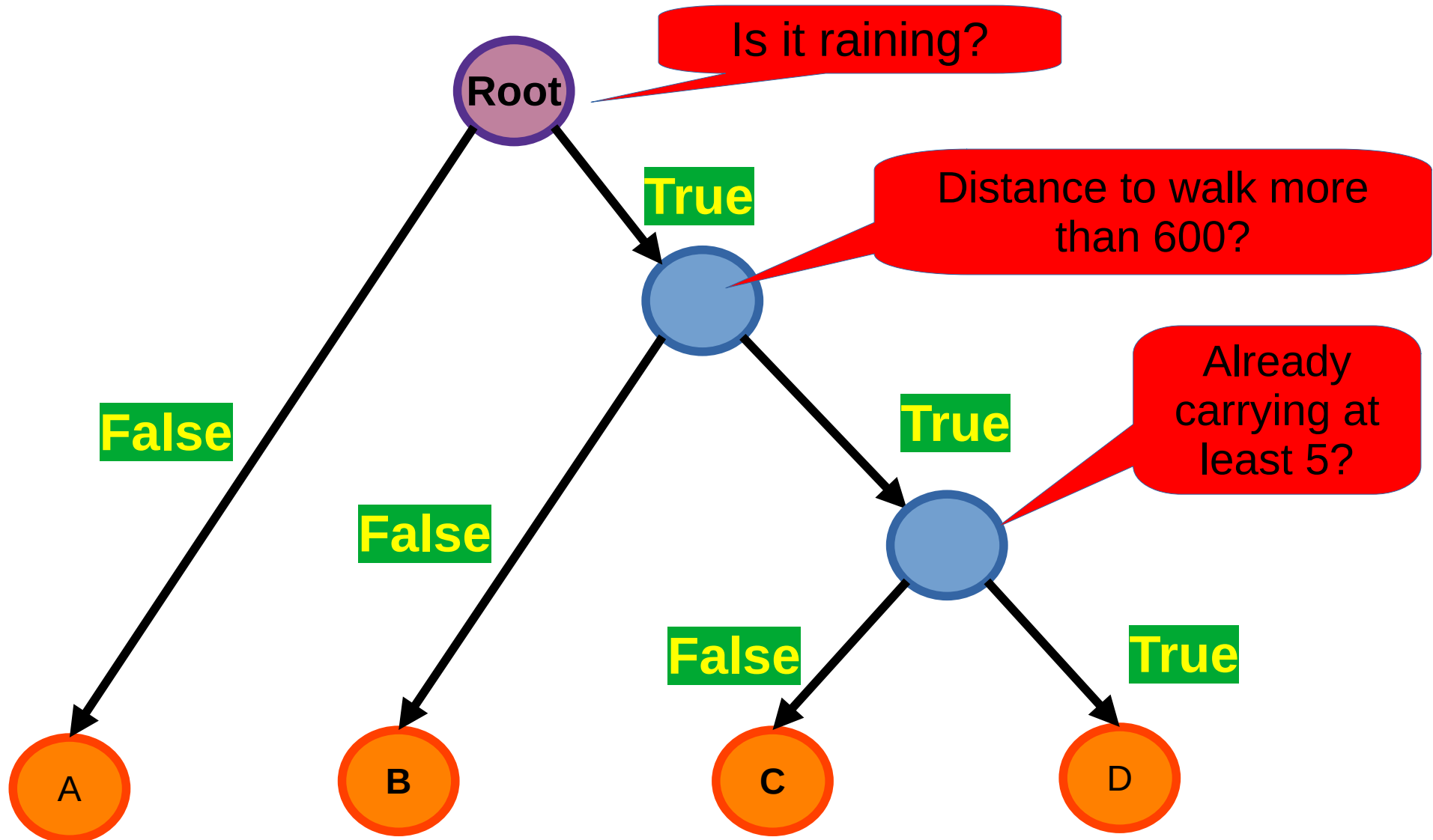
How many things am
I already carrying?

How far will I be
walking (potentially in
the rain)?

Should I bring the
umbrella with me?



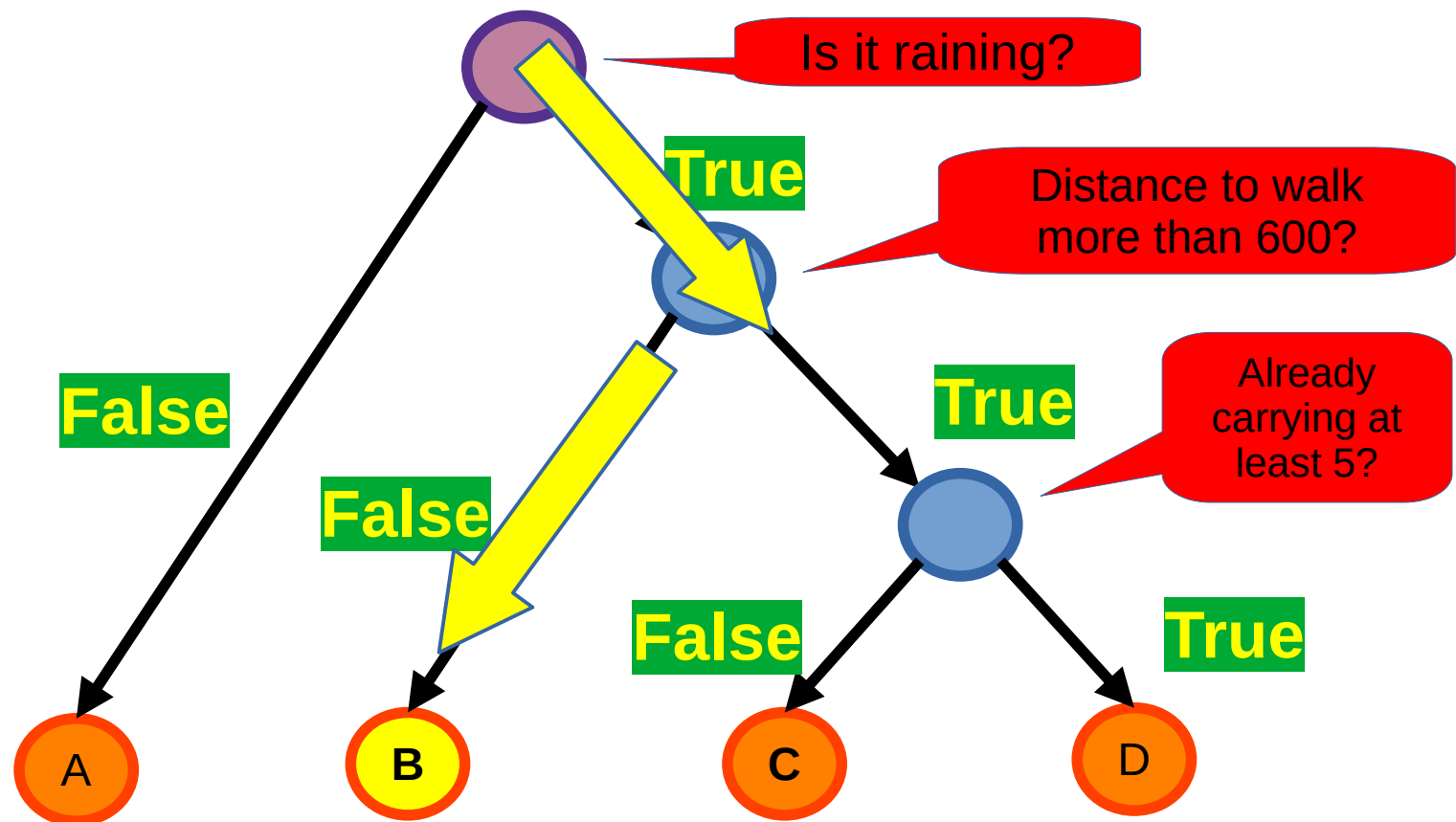
Setting Up a Decision Tree





Running Data (1)

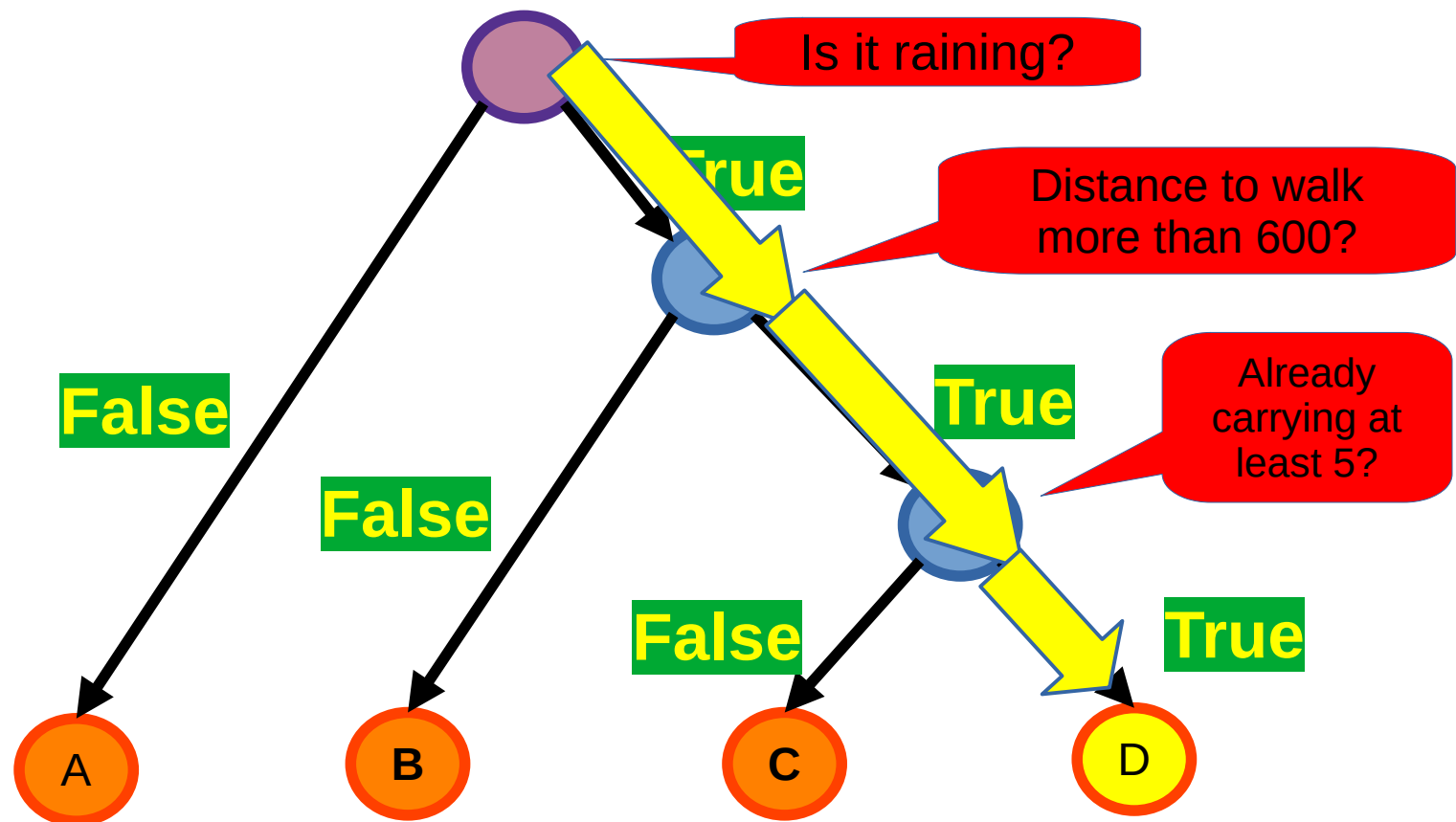
N	IsRaining?	WalkingDistance	ThingsCarrying	TakeUmbrella?
1	1	200	4	0





Running Data (2)

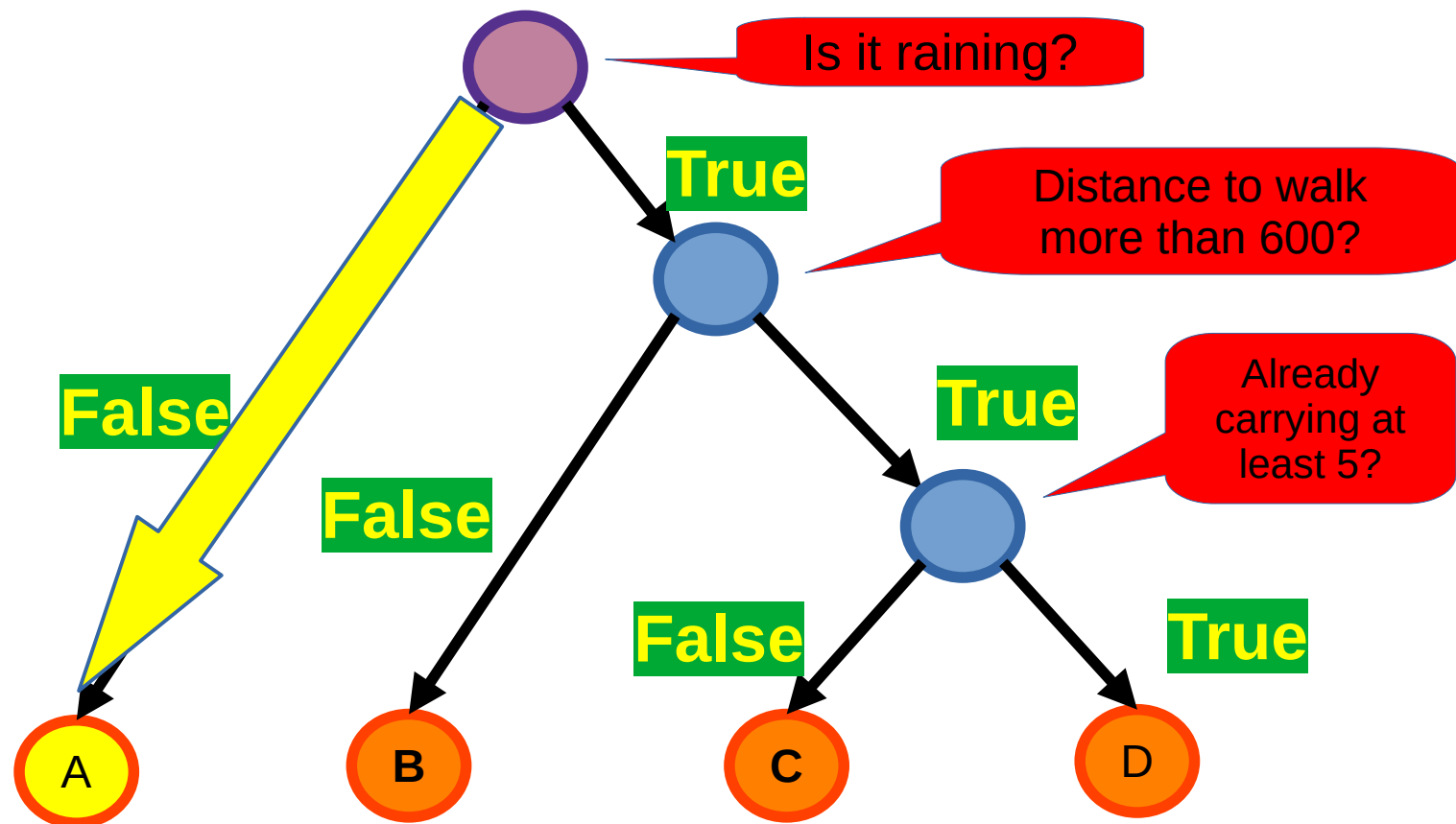
N	IsRaining?	WalkingDistance	ThingsCarrying	TakeUmbrella?
2	1	750	9	1





Running Data (3)

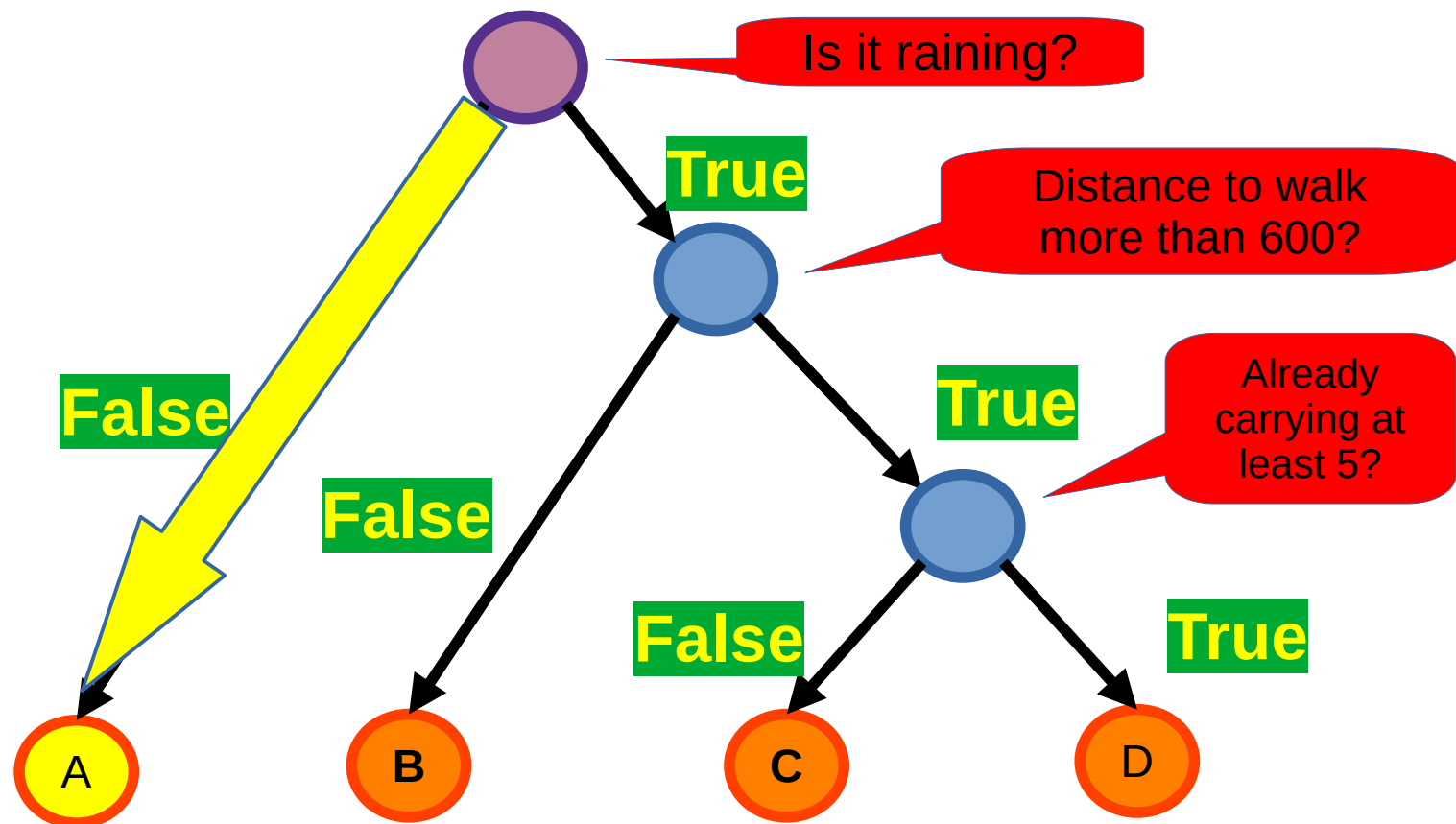
N	IsRaining?	WalkingDistance	ThingsCarrying	TakeUmbrella?
3	0	500	3	0



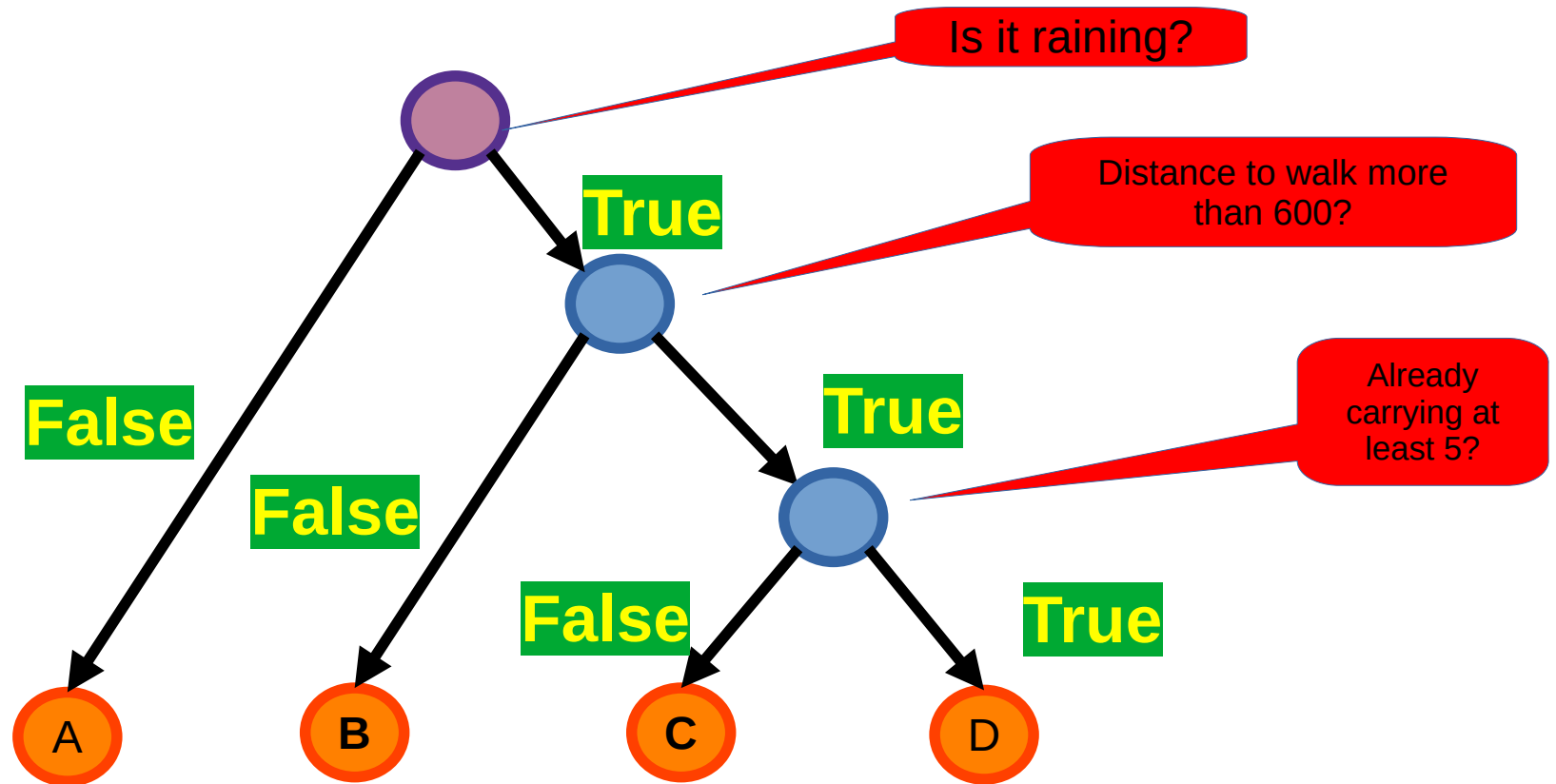


Running Data (4)

N	IsRaining?	WalkingDistance	ThingsCarrying	TakeUmbrella?
4	0	1000	7	0



What Eventualities Exist?



- Which question(s) from the decision tree was/were most relevant?
- What decision rule played the largest role in creating categories (the predicted behaviors)?



What Eventualities Exist?

N	IsRaining?	WalkingDistance	ThingsCarrying	TakeUmbrella?	Eventuality (leaf node)
1	1	200	4	0	B
2	1	750	9	1	D
3	0	500	3	0	A
4	0	1000	7	0	A








- We create a way to summarize the data based a series of steps.
- We note that some of the data may not be as important as others in making decisions ...

Bring

Do not
bring

Example:








Shapes and Colors

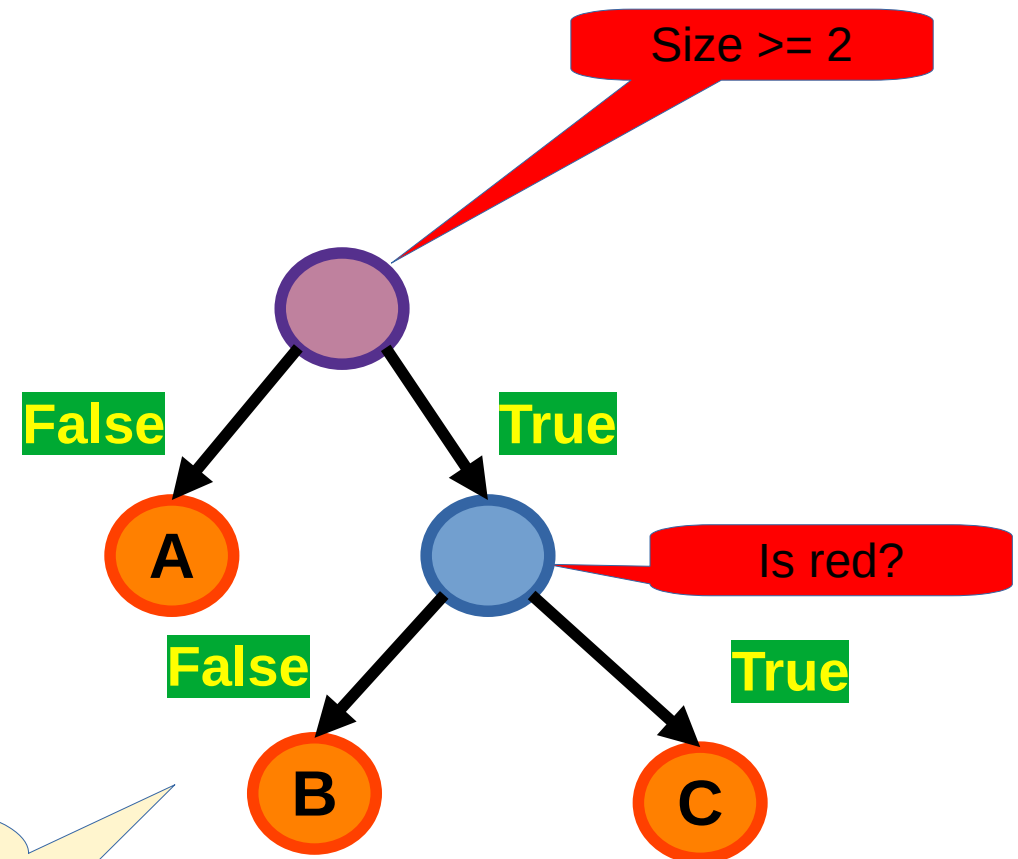
N	Size	Color	ShapeType
1	1	blue	circle 
2	2	red	triangle 
3	1.5	blue	circle 
4	2	red	triangle 
5	2.2	red	triangle 
6	3	red	circle 
7	3.5	blue	circle 

Lets construct some decision rules from a new set of data.
Present features of dataset: Shape, Size and Color`

Testing Two Decision Nodes

Size and Color

N	Size	Color	ShapeType
1	1	blue	circle 
2	2	red	triangle 
3	1.5	blue	circle 
4	2	red	triangle 
5	2.2	red	triangle 
6	3	red	circle 
7	3.5	blue	circle 



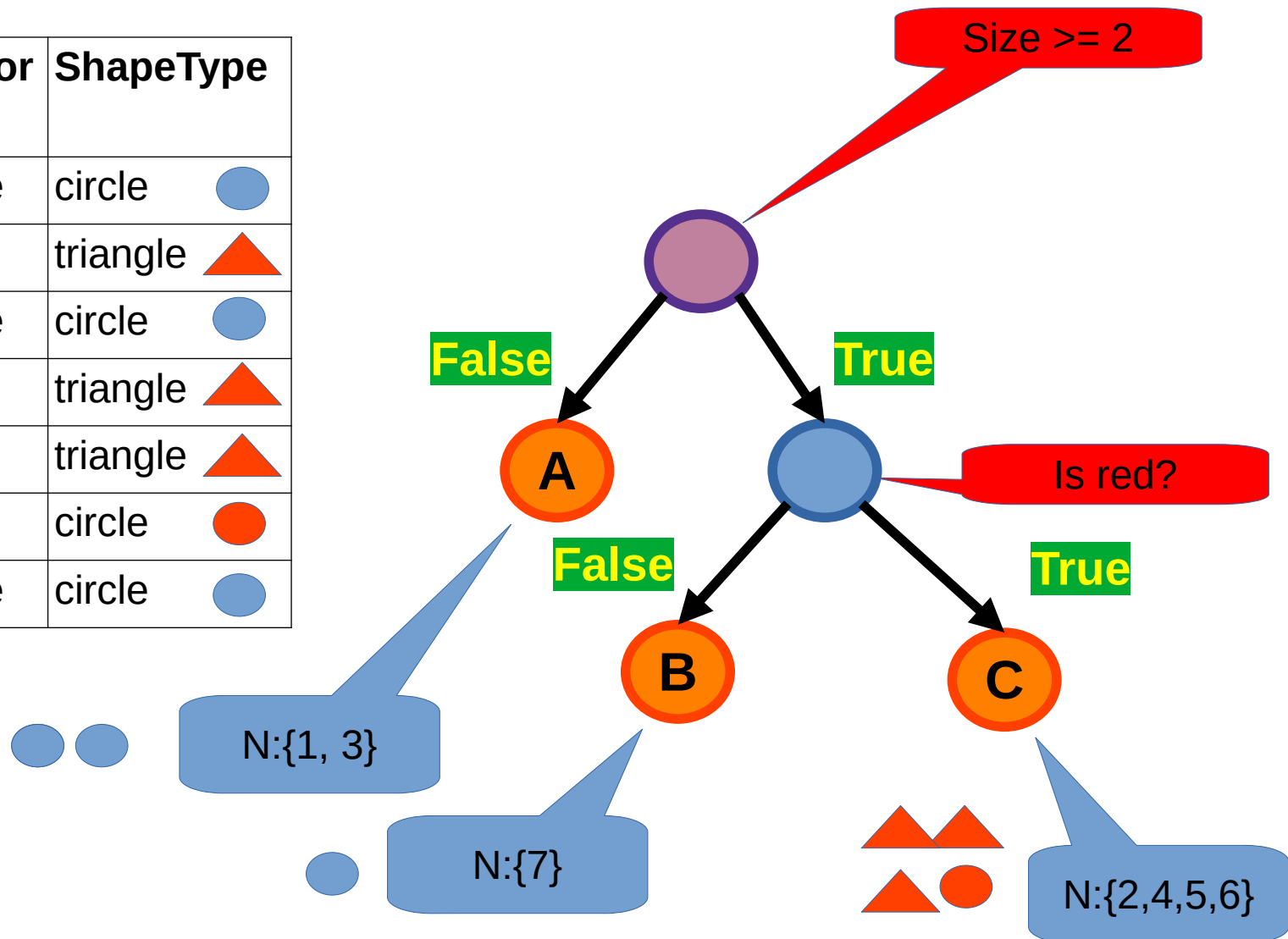
Which decision question works best to create the **purest** categories of data (What are these pure types?)

Size followed by **Color**,
Color followed by **Size**?

Two Main Decision Nodes

Size Then Color

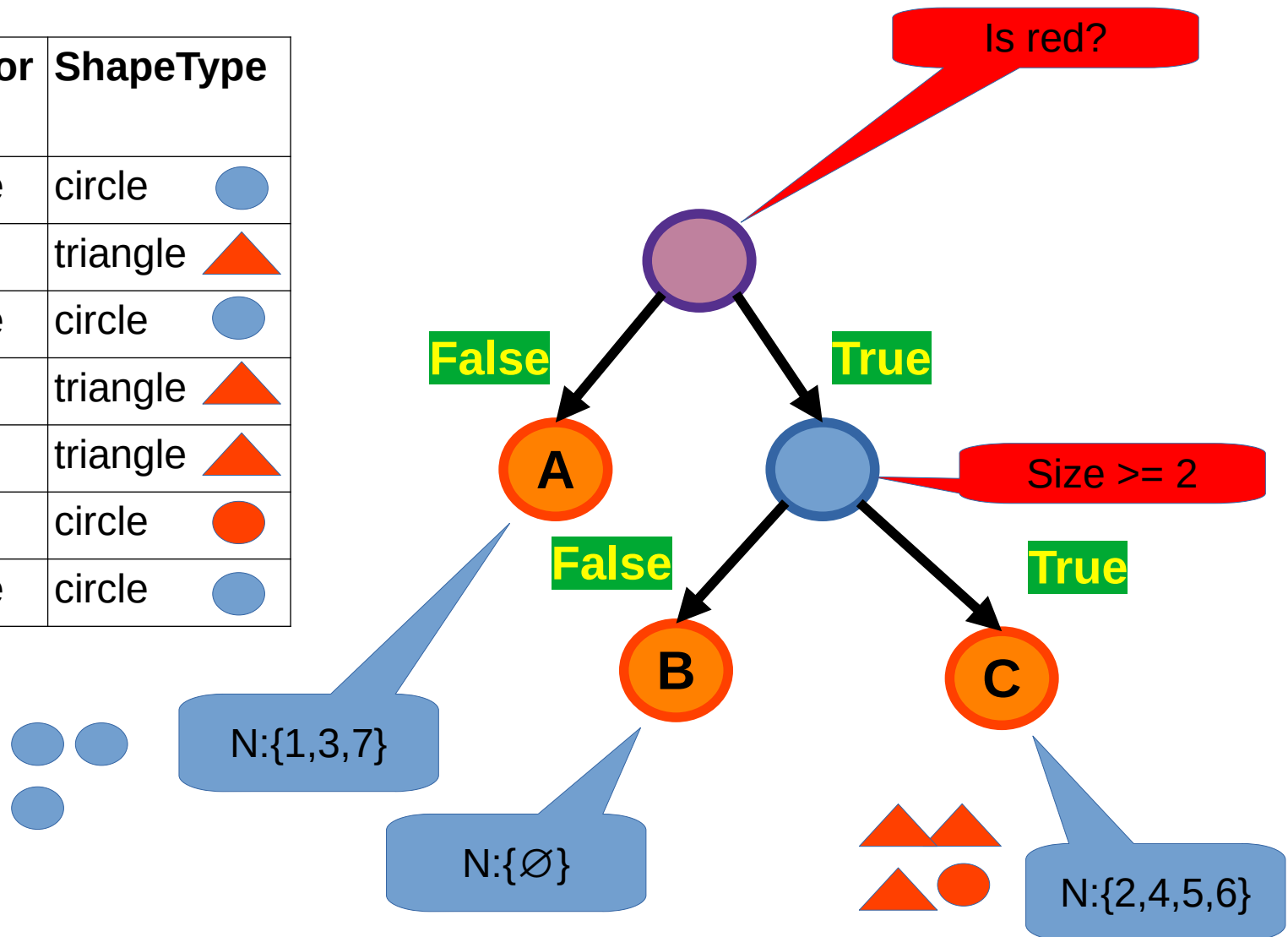
N	Size	Color	ShapeType
1	1	blue	circle
2	2	red	triangle
3	1.5	blue	circle
4	2	red	triangle
5	2.2	red	triangle
6	3	red	circle
7	3.5	blue	circle



Two Main Decision Nodes

Color Then Size

N	Size	Color	ShapeType
1	1	blue	circle
2	2	red	triangle
3	1.5	blue	circle
4	2	red	triangle
5	2.2	red	triangle
6	3	red	circle
7	3.5	blue	circle

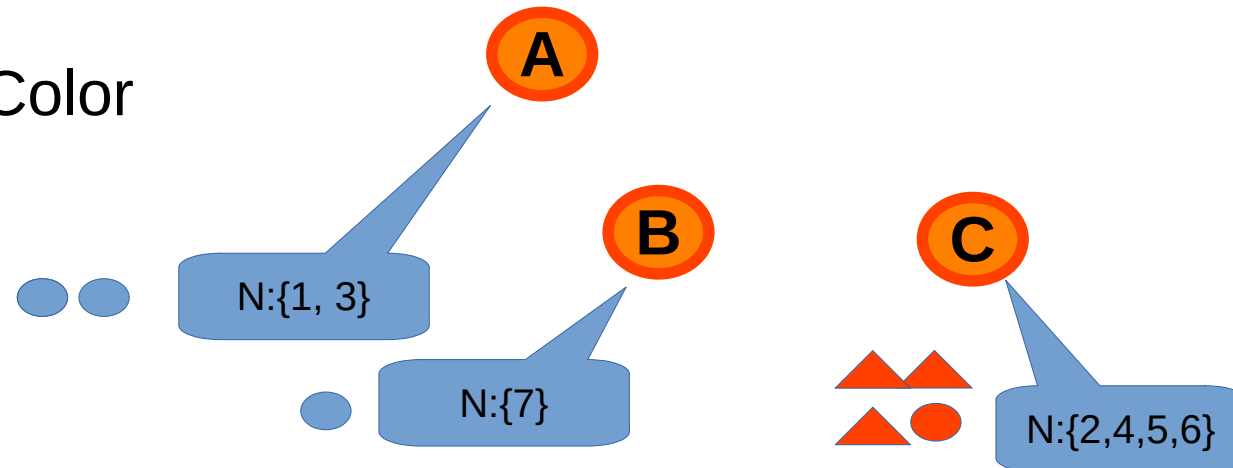




Conclusions

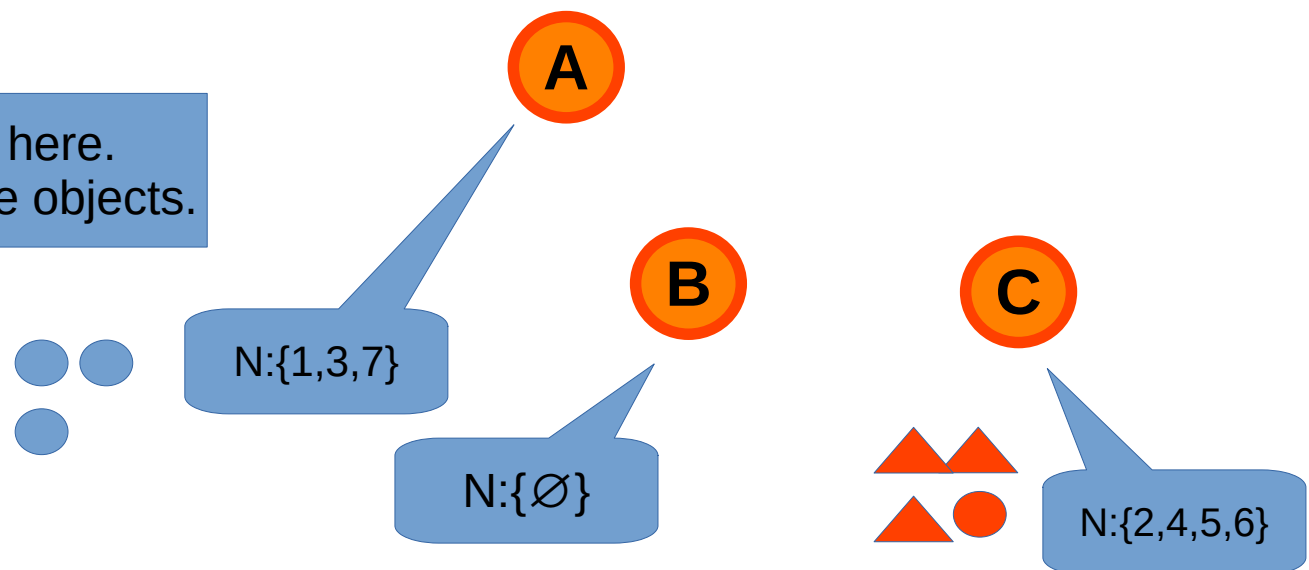
The Order of Decisions Matters

Size then Color



Color then Size

There is are two groups here.
One group contains all same objects.

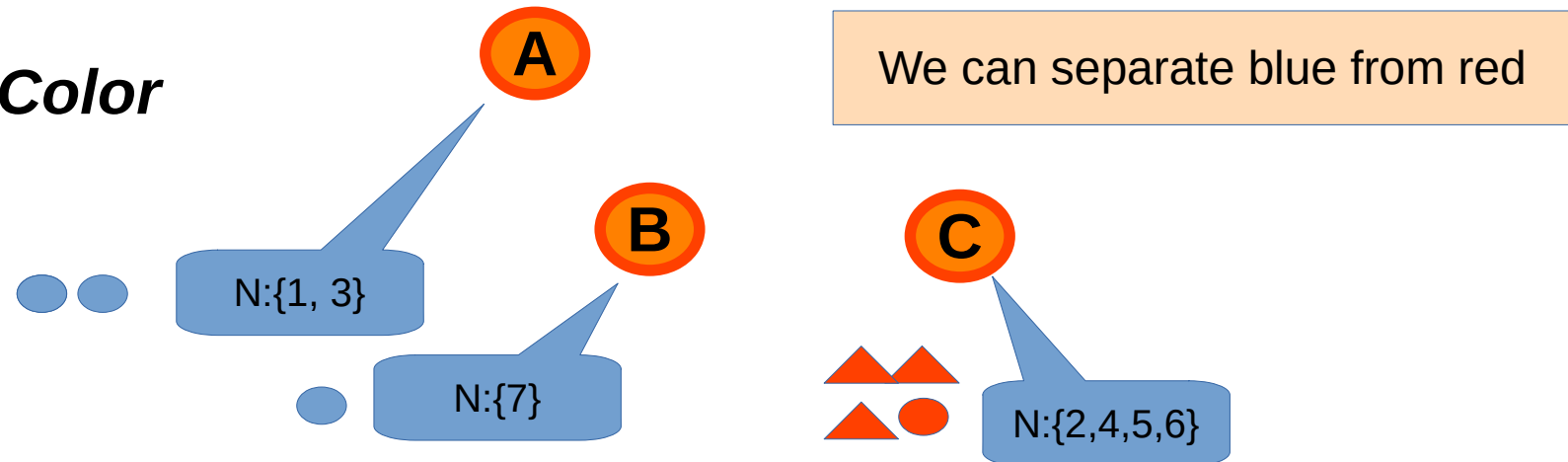




Conclusions

Types of Purity

Size then Color



Color then Size

