Slides for Pre-reading

Slides with yellow borders will not be covered in class, but is still testable content - you should review this before class.

Trees & Decisions Trees

Regular Trees

Trees in Computer Science

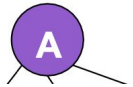
Regular Trees

Trees in Computer Science

- A Decision Tree is a way for a computer to make decisions based on a series of questions.

Trees in Computer Science

- A Decision Tree is a way for a computer to make decisions based on a series of questions.



A **tree** is a **collection of nodes** such that

- One node is the designated ***root***.

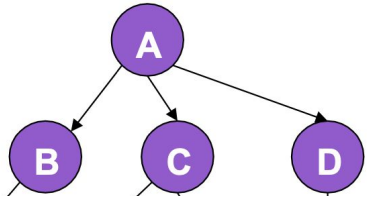
A is the ***root***

Trees in Computer Science

- A Decision Tree is a way for a computer to make decisions based on a series of questions.

A **tree** is a **collection of nodes** such that

- One node is the designated ***root***.
- A node can have zero or more *children*;



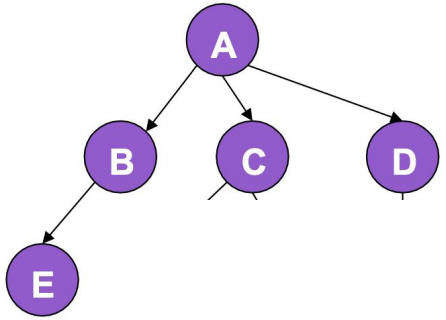
B, C and D are A's children

Trees in Computer Science

- A Decision Tree is a way for a computer to make decisions based on a series of questions.

A **tree** is a **collection of nodes** such that

- One node is the designated ***root***.
- A node can have zero or more *children*;
- a node with zero children is a *leaf*.



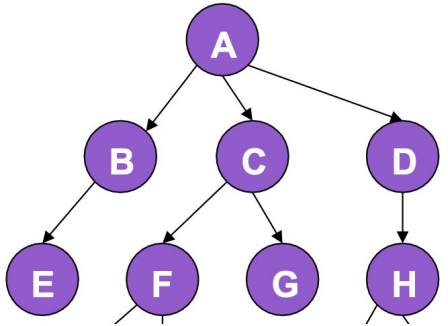
E is a leaf

Trees in Computer Science

- A Decision Tree is a way for a computer to make decisions based on a series of questions.

A **tree** is a **collection of nodes** such that

- One node is the designated ***root***.
- A node can have zero or more *children*;
- a node with zero children is a *leaf*.
- All non-root nodes have a single *parent*.



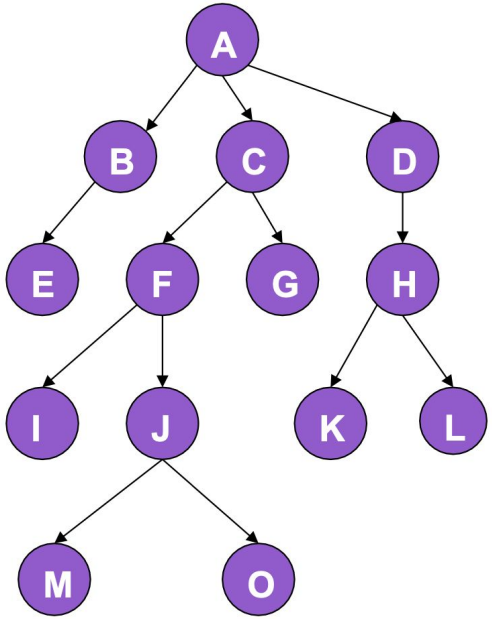
C is a parent to F + G

Trees in Computer Science

- A Decision Tree is a way for a computer to make decisions based on a series of questions.

A **tree** is a **collection of nodes** such that

- One node is the designated ***root***.
- A node can have zero or more *children*;
- a node with zero children is a *leaf*.
- All non-root nodes have a single *parent*.
- *Edges* denote parent-child relationships.
  - Example: The arrows between F → I

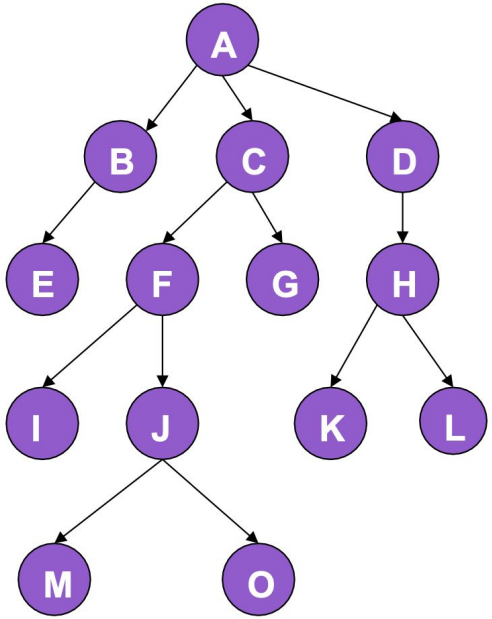


Trees in Computer Science

- A Decision Tree is a way for a computer to make decisions based on a series of questions.

A **tree** is a **collection of nodes** such that

- One node is the designated ***root***.
- A node can have zero or more *children*;
- a node with zero children is a *leaf*.
- All non-root nodes have a single *parent*.
- *Edges* denote parent-child relationships.
- Nodes and/or edges may be labeled by data.
  - Each node on this tree is labeled by a letter



CPSC 100

Computational Thinking

Algorithm, Classifiers and Trees!

Instructor: Firas Moosvi
Department of Computer Science
University of British Columbia

Agenda

- Course Admin
- Recap
  - Classifier
  - Decision Trees
- Entropy

Learning Goals

Learning Goals

After this lecture, you should be able to:

- Explain the concept of a **rooted tree** and **decision tree**.
- Describe what the general decisions are in building a decision tree.
  - **Build a decision tree using entropy.**
- Describe **what considerations** are important in building a decision tree.

Course Admin

Clicker Questions (using Agora)

Clicker Question

Q: What is a classifier?

A. *This option is intentionally left blank*

B. A method to predict the future

C. An algorithm that maps input data to a specific category

D. A type of decision tree used for data mining

E. A type of data storage for algorithms

Clicker Question

Q: In classification, how is the accuracy of a classifier evaluated?

iClicker

A. By comparing training data with random data

B. By matching the classifier's results with decisions from test data

C. By ensuring the classifier can handle large datasets

D. By improving the efficiency of the algorithm

Let's build a Decision Tree

Building Decision Trees

- Should you get an ice cream?
- You might start out with the following data

Weather	Wallet	Ice Cream?
Great	Empty	No
Nasty	Empty	No
Great	Full	Yes
Okay	Full	Yes
Nasty	Full	No

Building Decision Trees

- Should you get an ice cream?
- You might start out with the following data

Attributes

Conditions

Weather	Wallet	Ice Cream?
Great	Empty	No
Nasty	Empty	No
Great	Full	Yes
Okay	Full	Yes
Nasty	Full	No

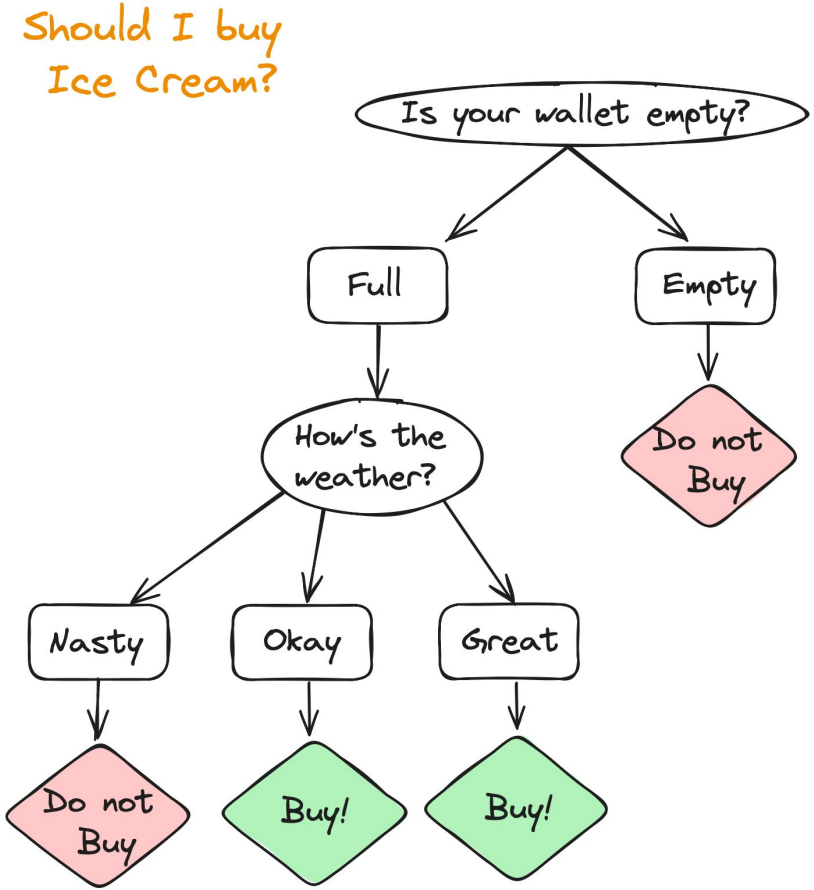
Ice Cream Decision Tree

Should you get an ice cream?

Weather	Wallet	Ice Cream?
Great	Empty	No
Nasty	Empty	No
Great	Full	Yes
Okay	Full	Yes
Nasty	Full	No

Should you get an ice cream?

Weather	Wallet	Ice Cream?
Great	Empty	No
Nasty	Empty	No
Great	Full	Yes
Okay	Full	Yes
Nasty	Full	No



In-class Activity

In-class Activity: Decision Tree

Draw the decision tree with the data on the right for 13 people that are on the job market.

There is data on the salary as well the commute time…

#	Salary ($)	Commute Time	Decision
1	120,000	45 min	Accept
2	105,000	10 min	Accept
3	100,000	25 min	Accept
4	90,000	20 min	Accept
5	75,000	15 min	Accept
6	65,000	60 min	Reject
7	55,000	20 min	Reject
8	50,000	25 min	Reject
9	45,000	70 min	Reject
10	30,000	40 min	Reject
11	95,000	25 min	Accept
12	80,000	10 min	Accept
13	110,000	50 min	Accept

What if it is not so clear ?

Soccer League:

Do we cancel the game?

Soccer League: Cancel Game?

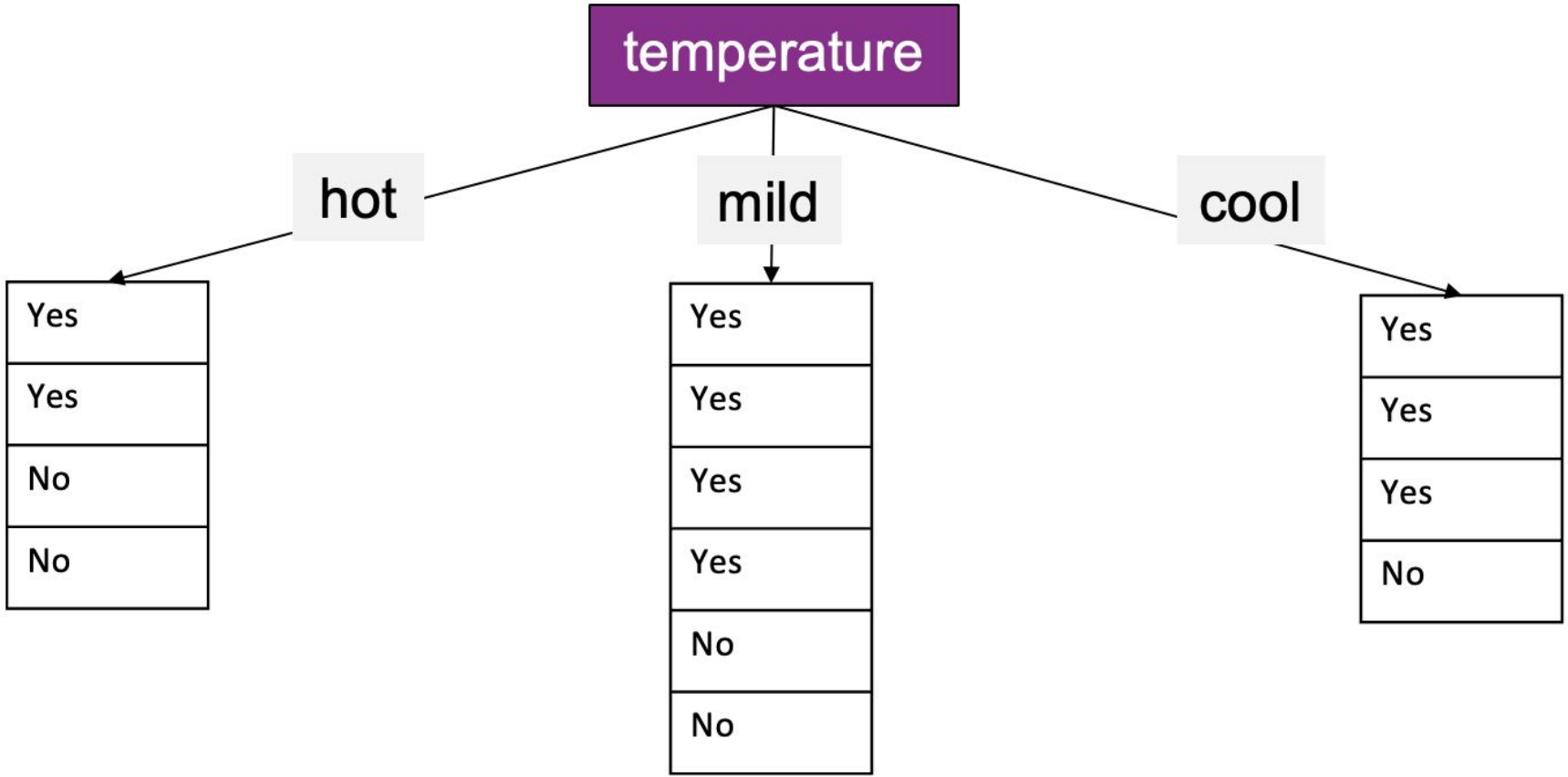
- Build a **decision tree** to help officials decide
- Assume that decisions are the same given the same information
- The *leaf nodes* should be whether or not to play
- The *non-leaf* nodes should be **attributes** (e.g., Outlook, Windy)
- The edges should be **conditions** (e.g., sunny, hot, normal)

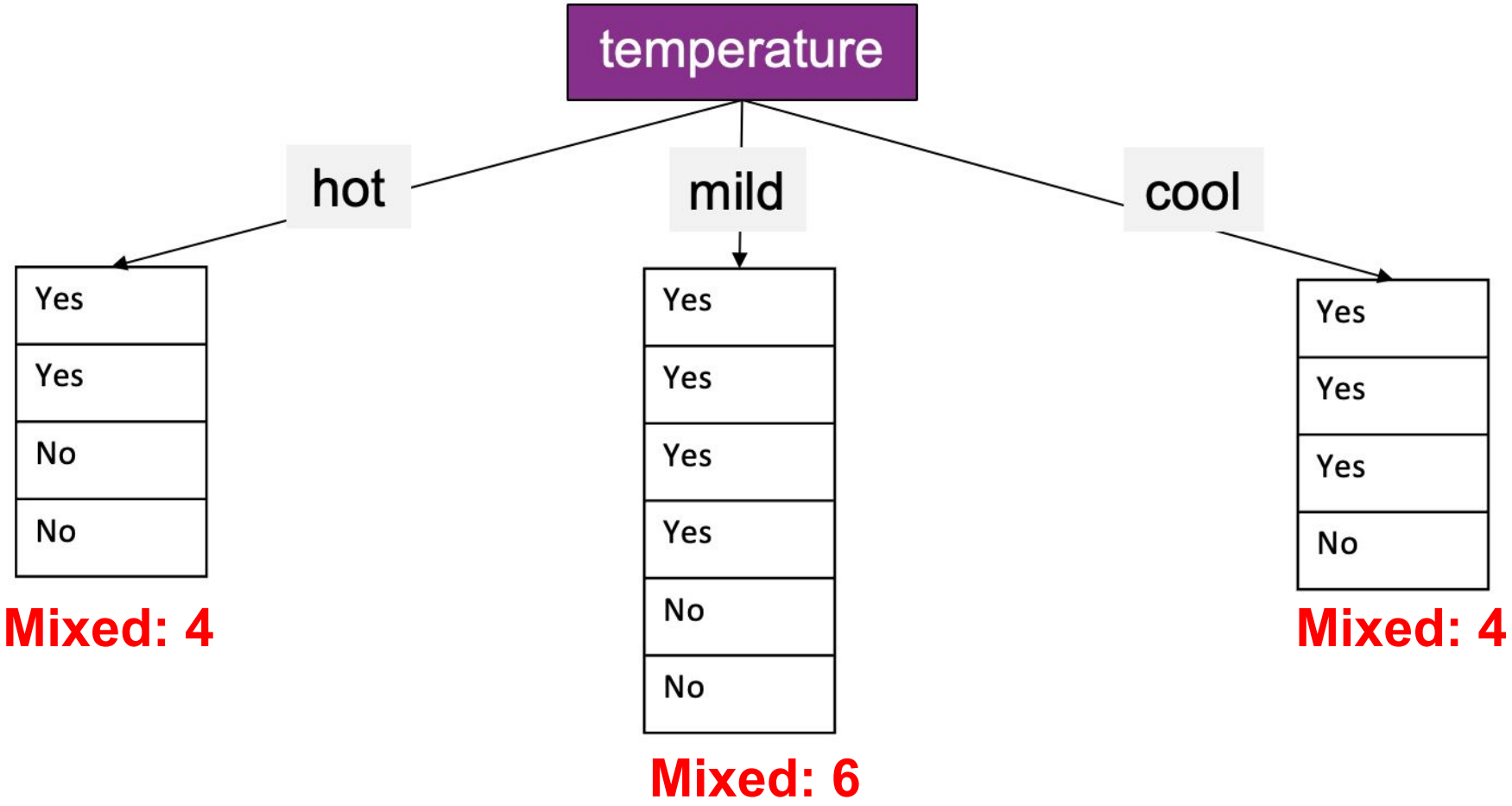
Want to have as few mixed “Yes” and “No” answers together in groups as possible.

At the start we have 14 mixed Yes’s/No’s

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

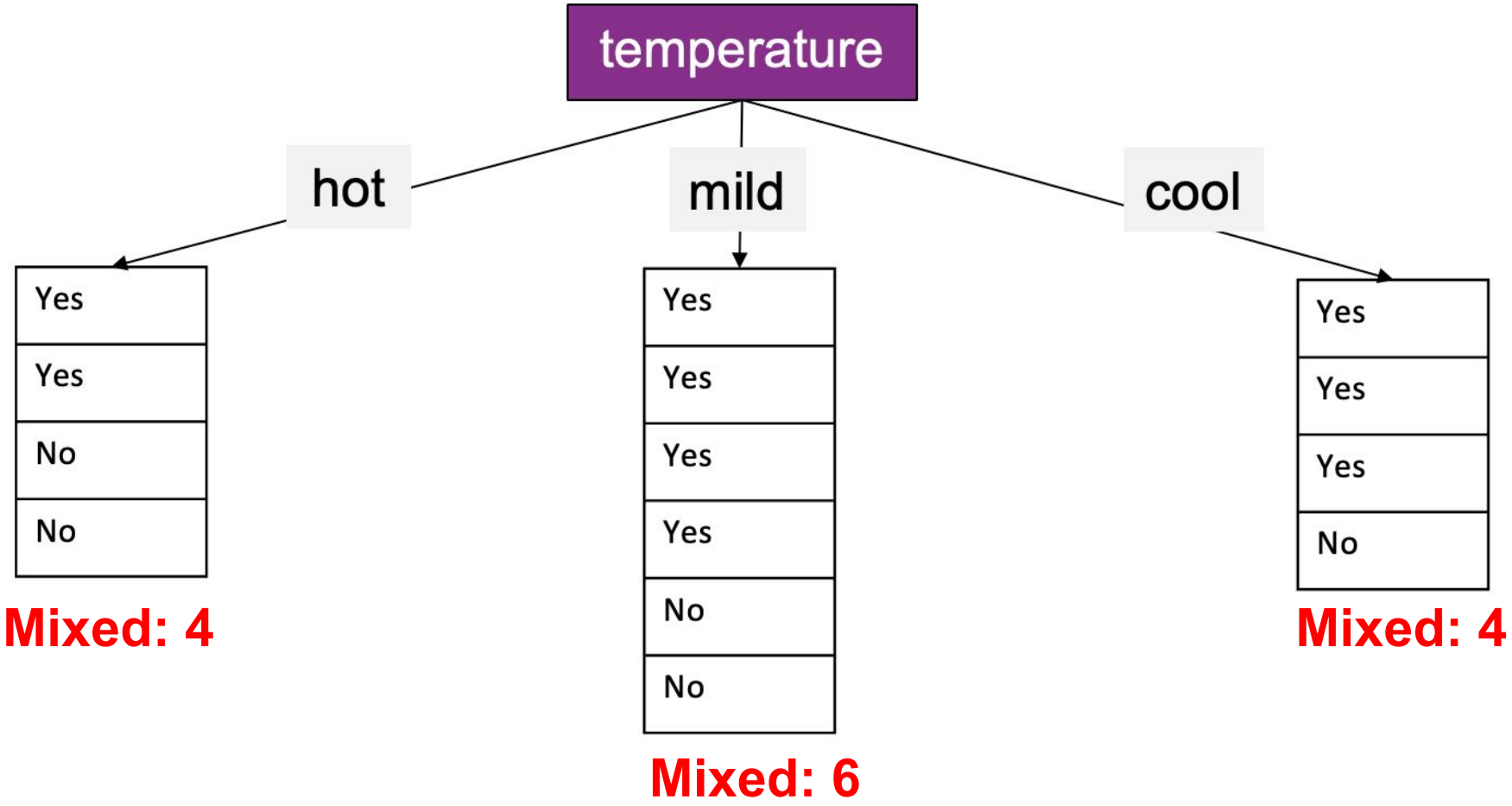
What happens if we split data on Temperature?


temperature
hot
mild
cool
Yes
Yes
No
No
Yes
Yes
Yes
Yes
No
No
Yes
Yes
Yes
No

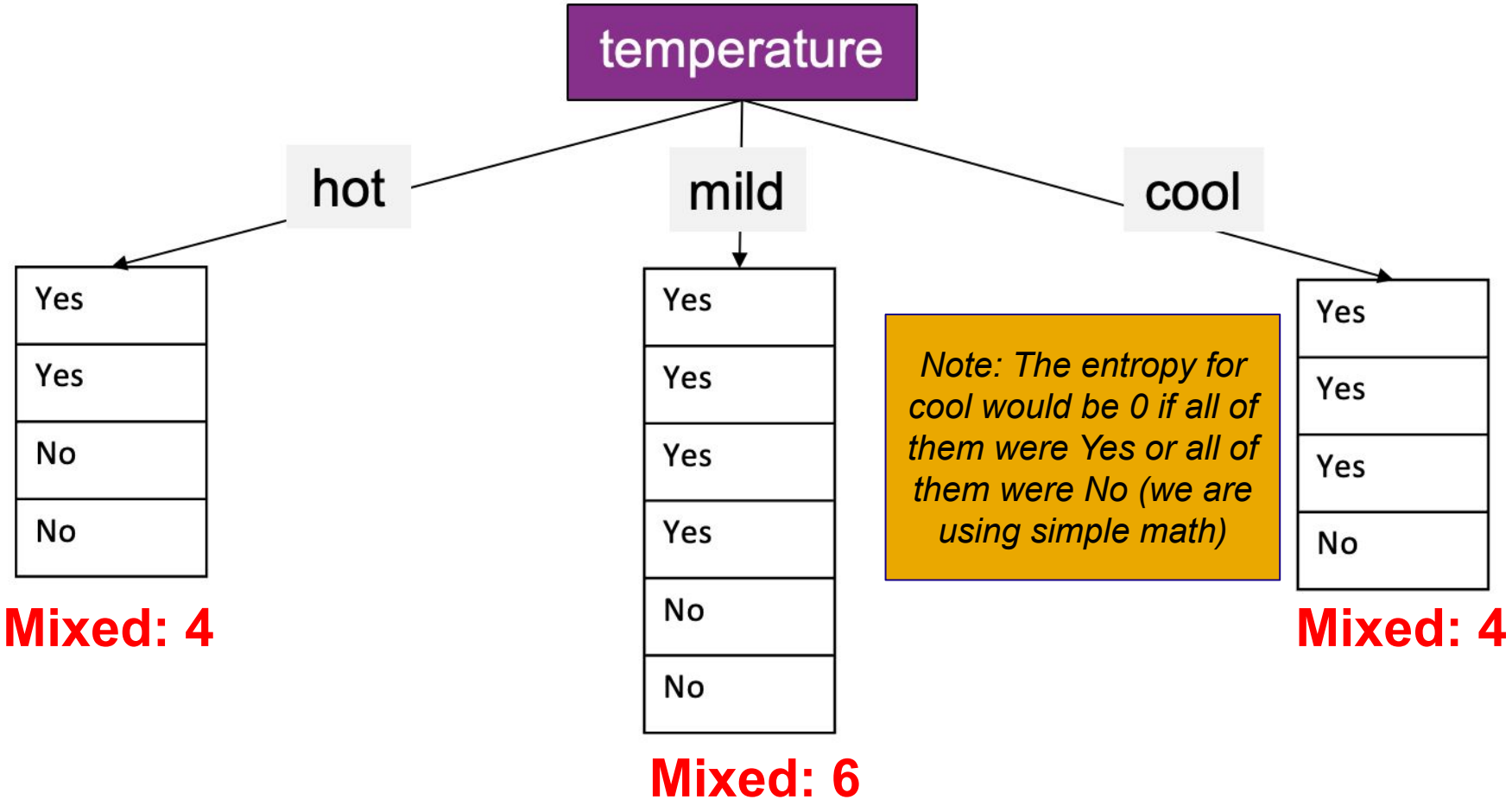

temperature
hot
mild
cool
Yes
Yes
No
No
Yes
Yes
Yes
Yes
No
No
Yes
Yes
Yes
No
Mixed: 4
Mixed: 6
Mixed: 4

What is the uncertainty (entropy) in our data?

Overall entropy = 4 + 4 + 6 = 14



Overall entropy = 4 + 4 + 6 = 14



In-class Activity

[Groups of 3-4]

What's the entropy if you split on the Outlook?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Clicker Question

Q: What's the entropy if you split on Outlook?

A. 0

B. 5

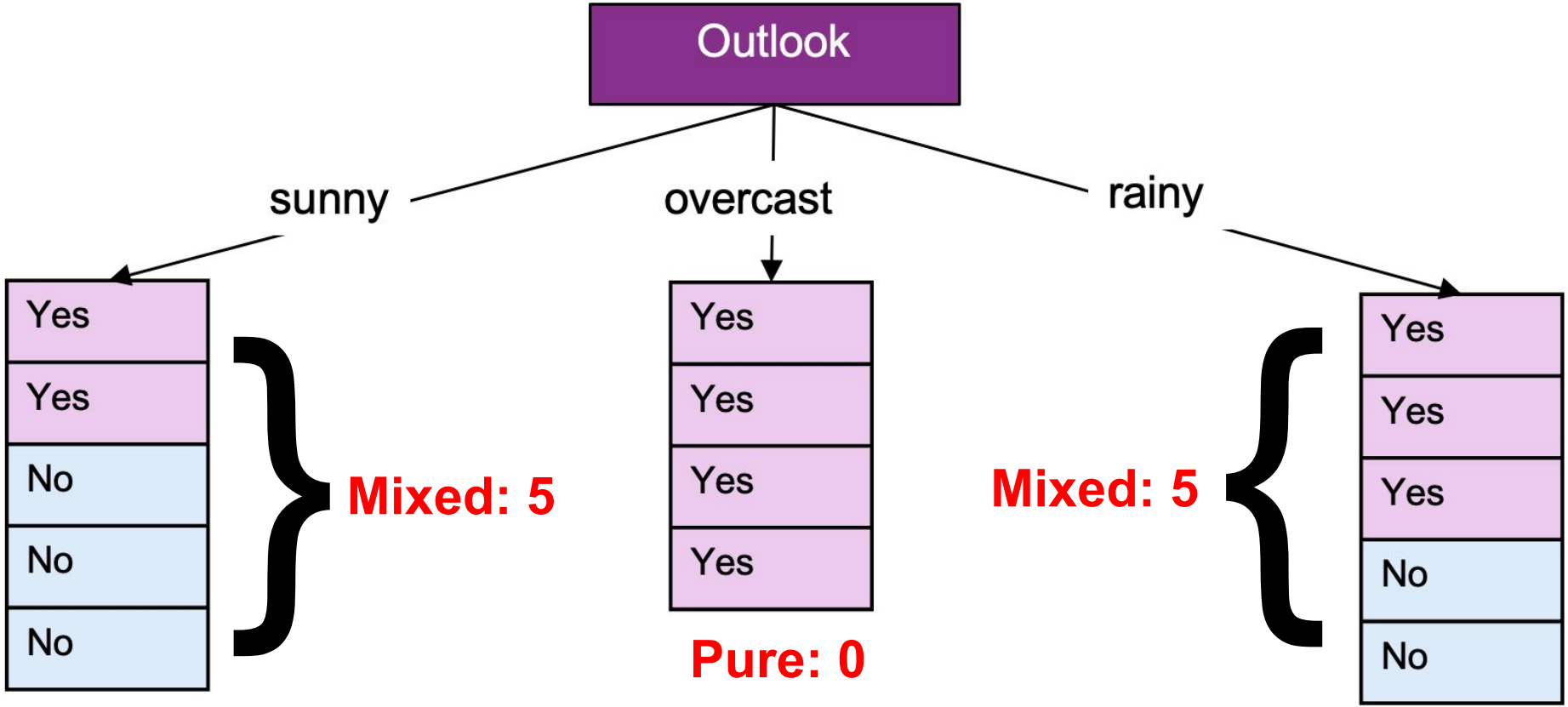
C. 10

D. 14

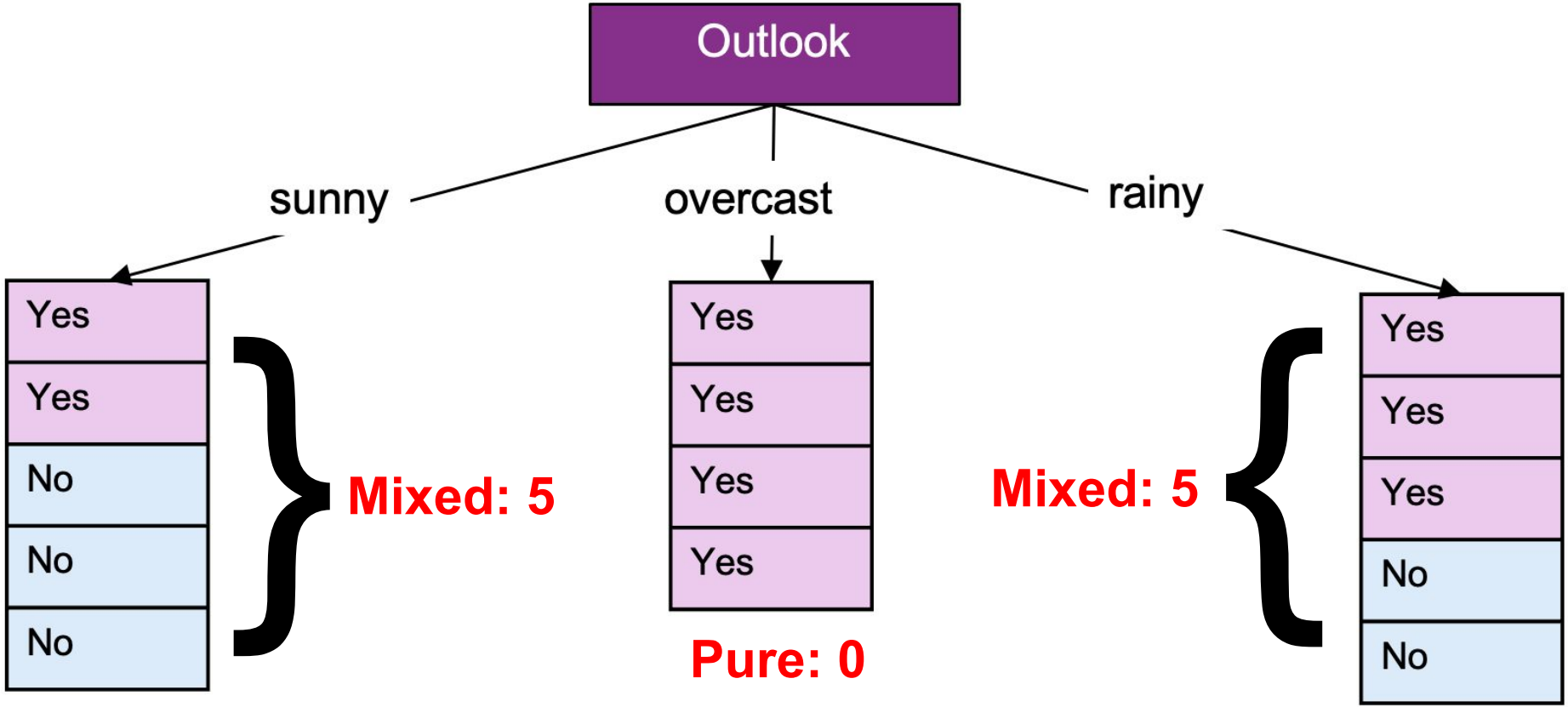
E. None of the above

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Q: What's the entropy if you split on Outlook?



Overall entropy = 5 + 0 + 5 = 10

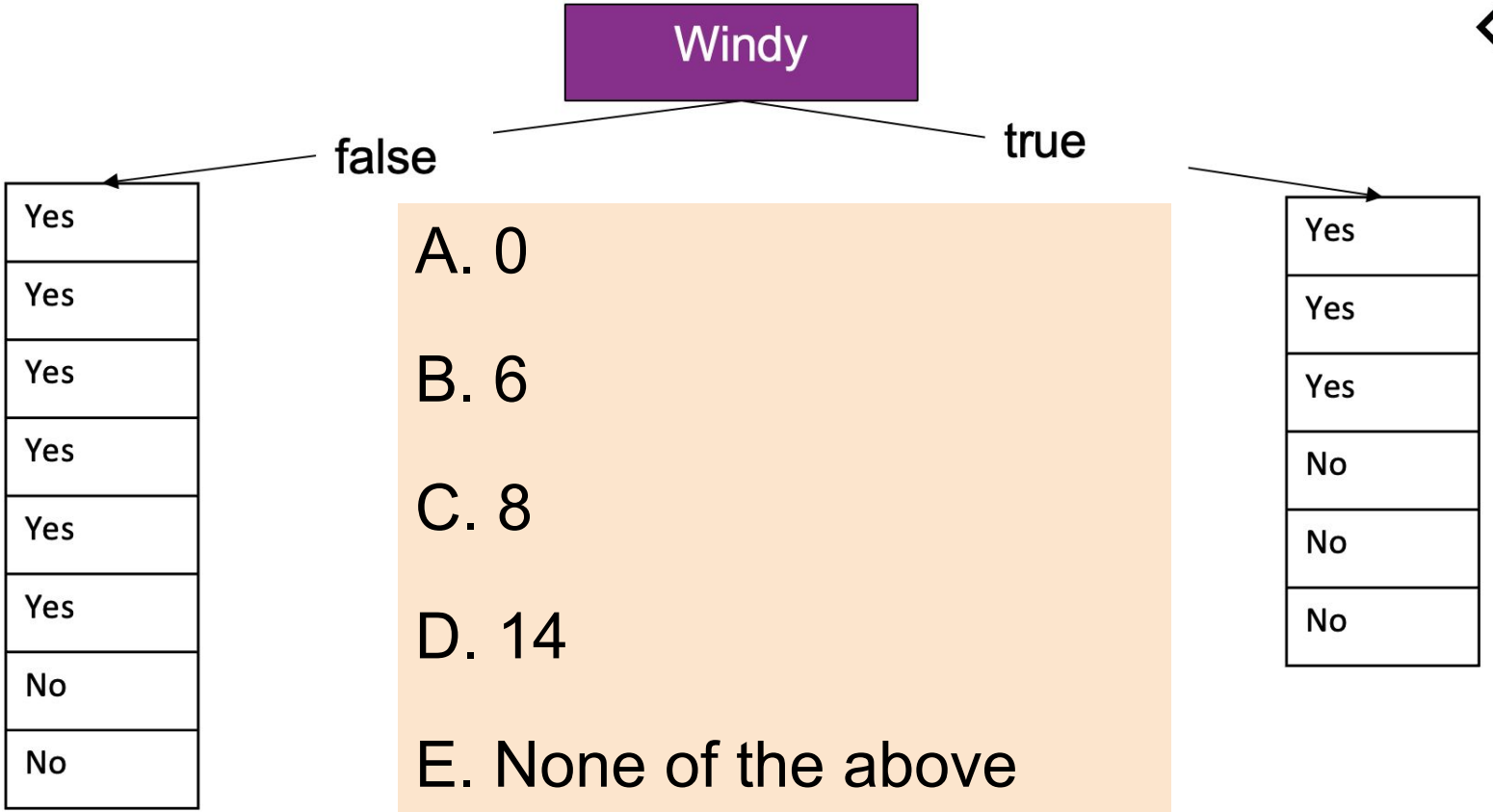


What's the entropy if you split on Windy?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Clicker Question

Q: What's the entropy if you split on Windy?



A. 0

B. 6

C. 8

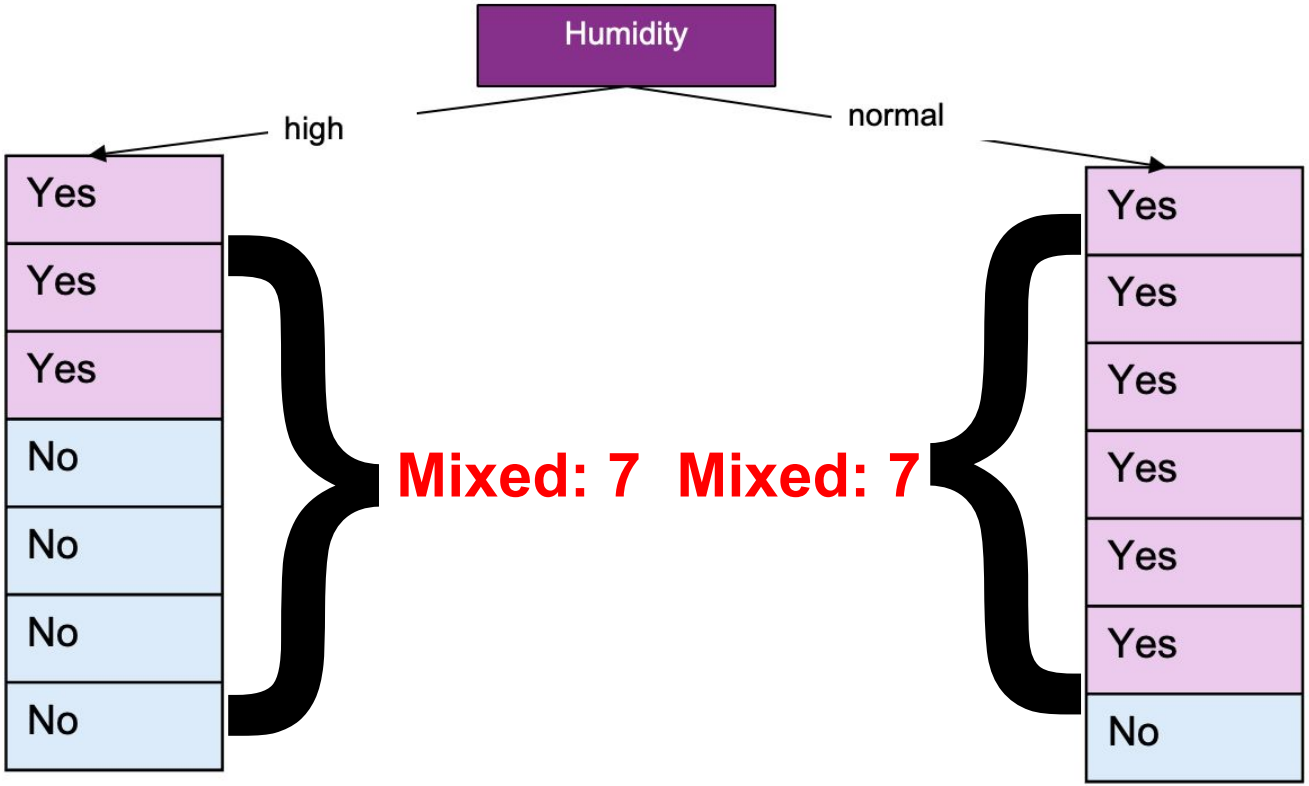
D. 14

E. None of the above

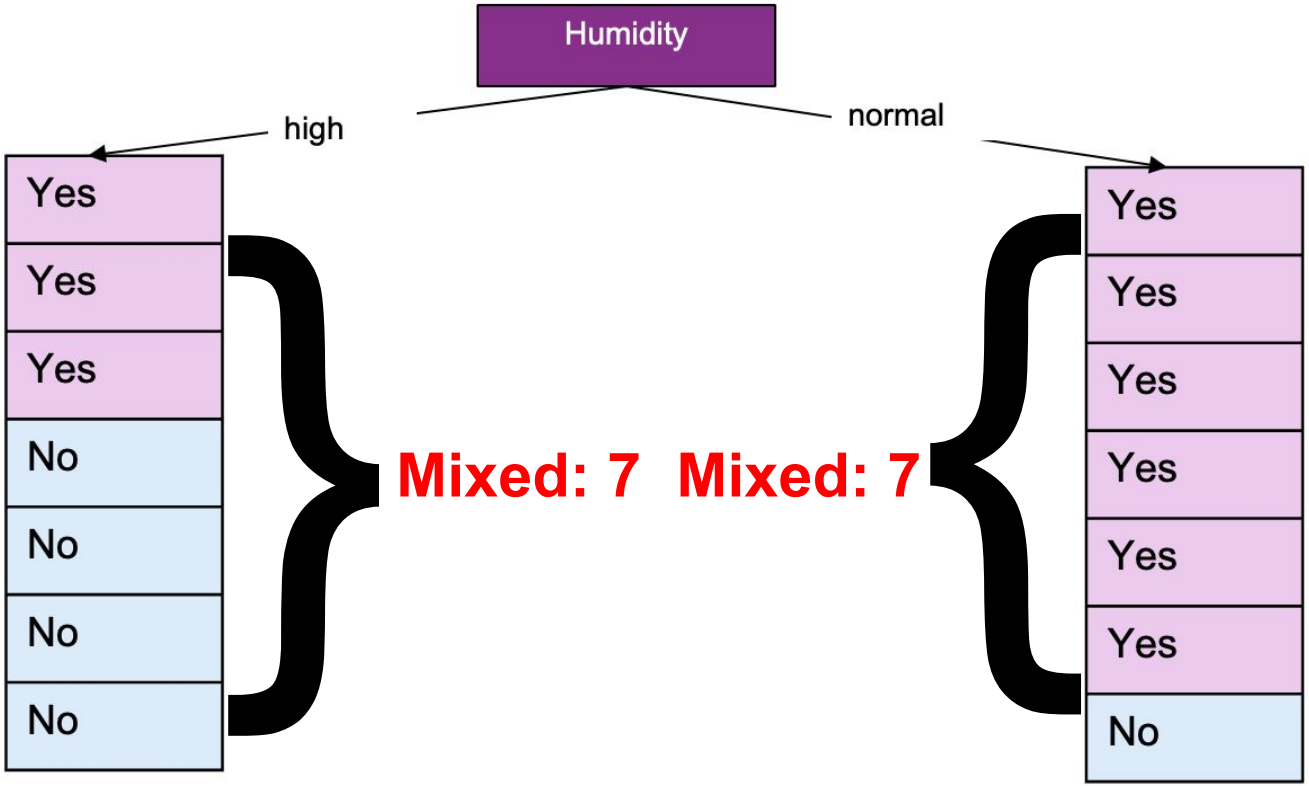
What's the entropy if you split on Humidity?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Q: What's the entropy if you split on Humidity?



Overall entropy = 7 +7 = 14



Debrief

What is the best attribute to split on?

- Entropy if we split on Temperature = 14
- Entropy if we split on Outlook = 10
- Entropy if we split on Windy = 14
- Entropy if we split on Humidity = 14

Why?

What is the best attribute to split on?

- Entropy if we split on Temperature = 14
- **Entropy if we split on Outlook = 10**
- Entropy if we split on Windy = 14
- Entropy if we split on Humidity = 14

Why? It does the best job of **minimizing** entropy

Wrap up