

# Decision Trees

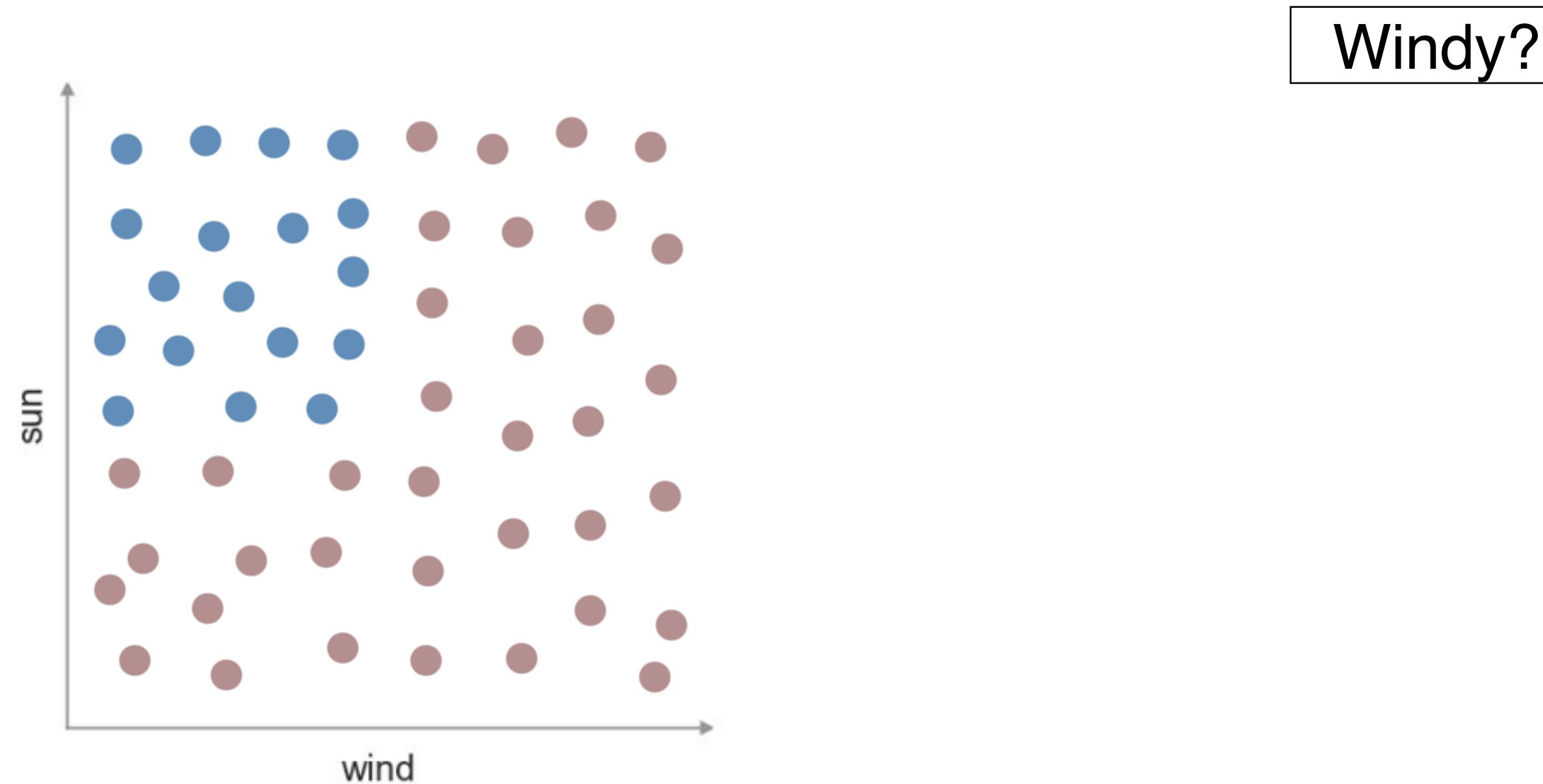
---

- Recall that The Kernel Trick in SVMs allowed us to derive nonlinear decision rules using simple linear techniques
- Decision Trees allow us to come up with nonlinear-ish decision boundaries via a union of simple linear boundaries
- Decision Trees are basis for many popular learning algorithms
- Two of the Top 10 Data Mining Algorithms (as voted by IEEE International Conference on Data Mining in 2006) are Decision Trees

# Simple Example

Decide whether you should play tennis

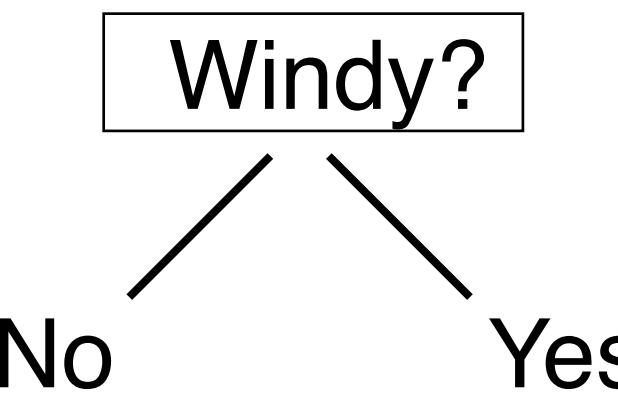
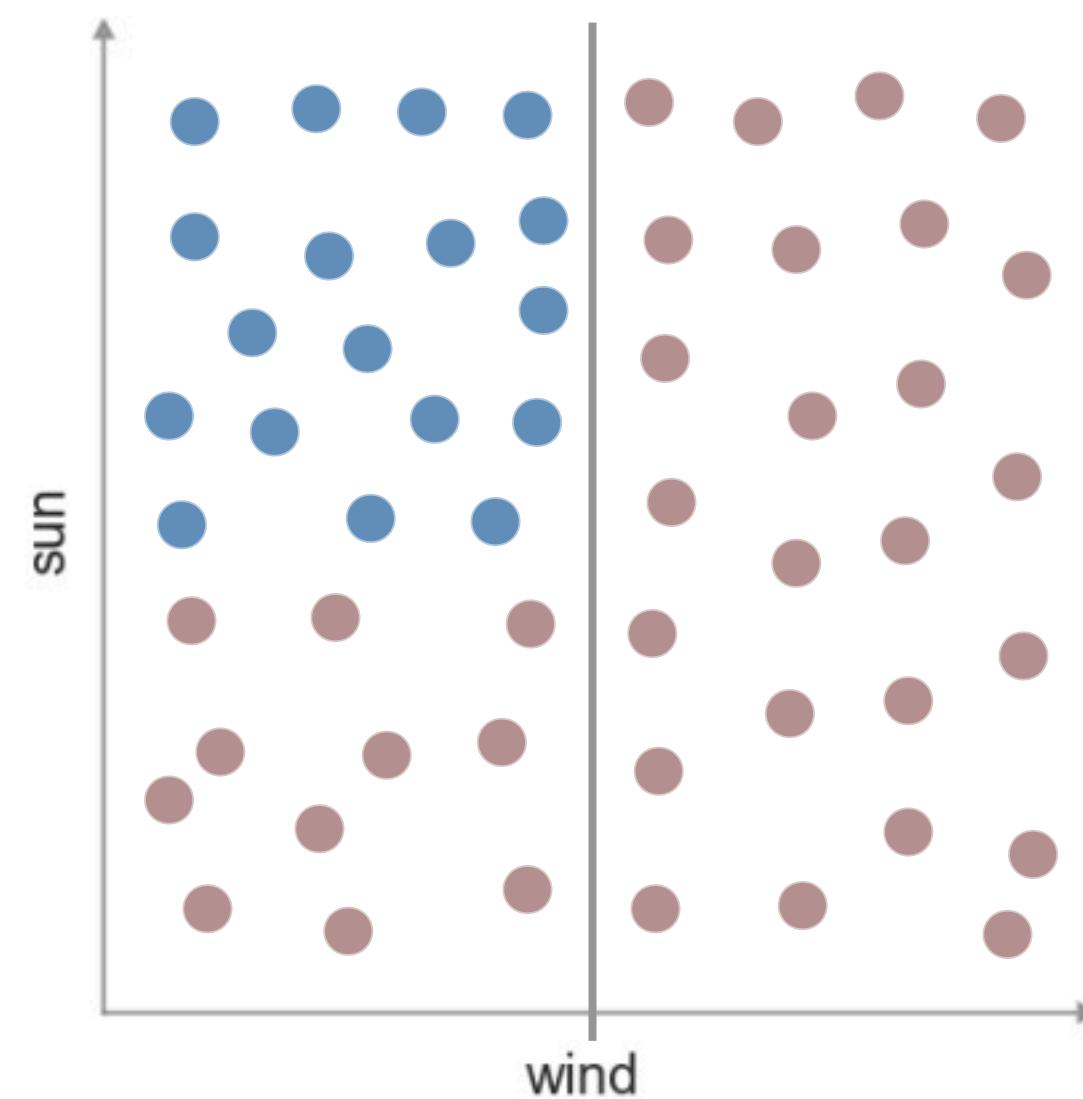
Decision trees allow you to ask multiple linear questions



# Simple Example

Decide whether you should play tennis

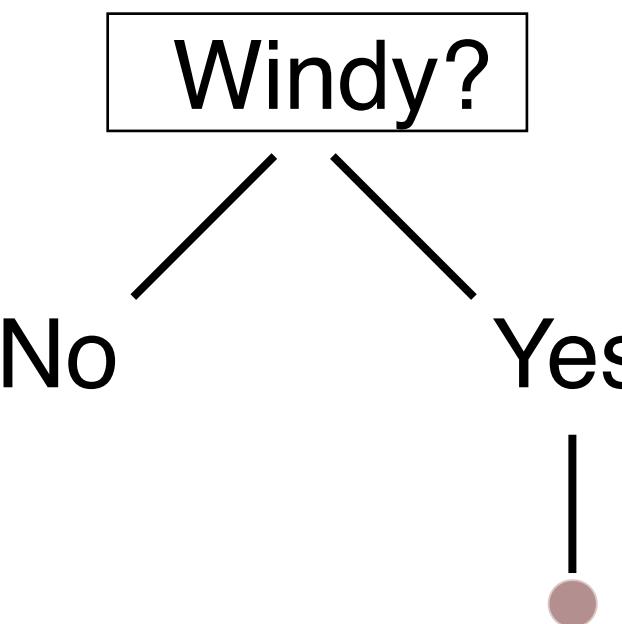
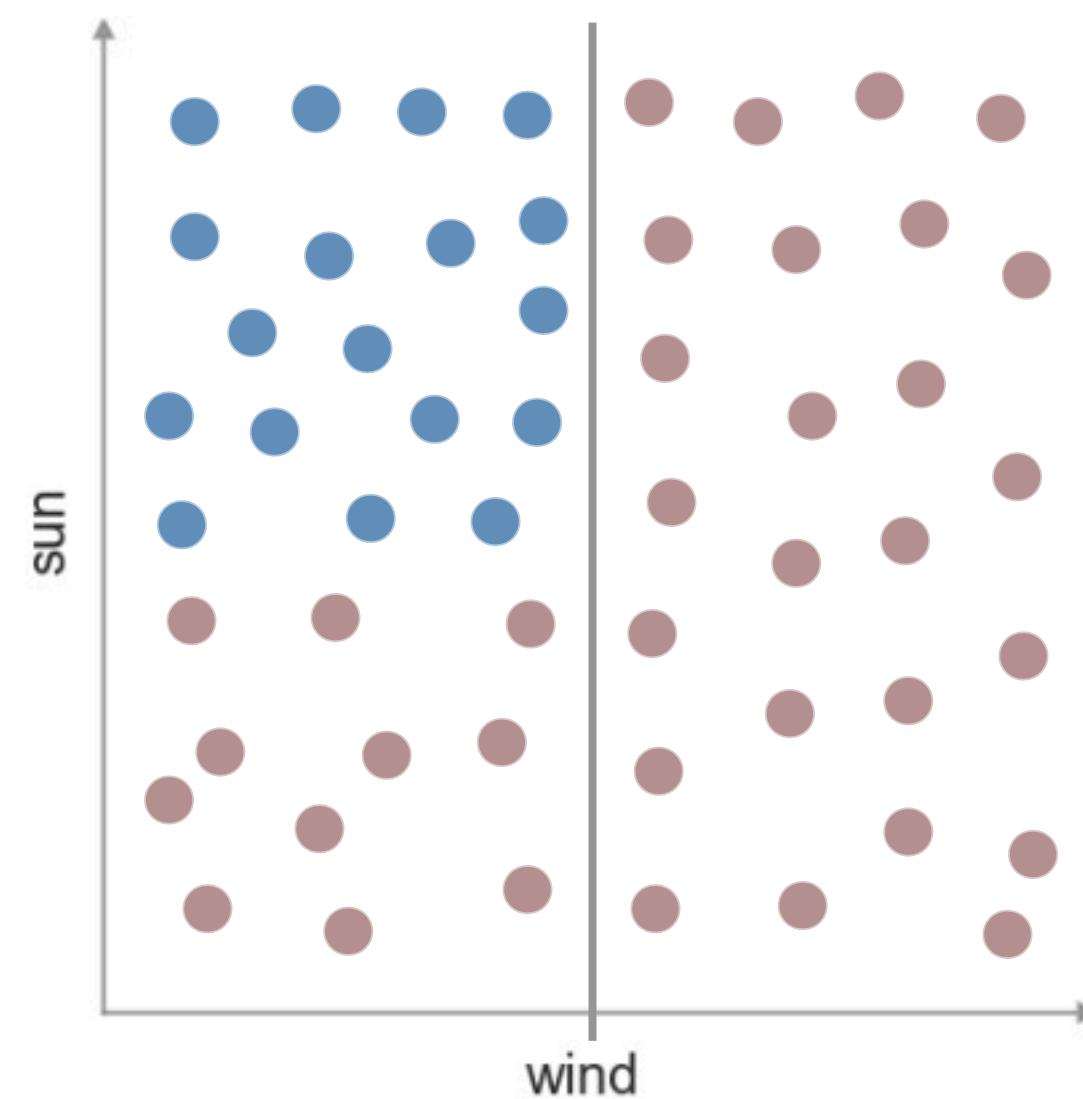
Decision trees allow you to ask multiple linear questions



# Simple Example

Decide whether you should play tennis

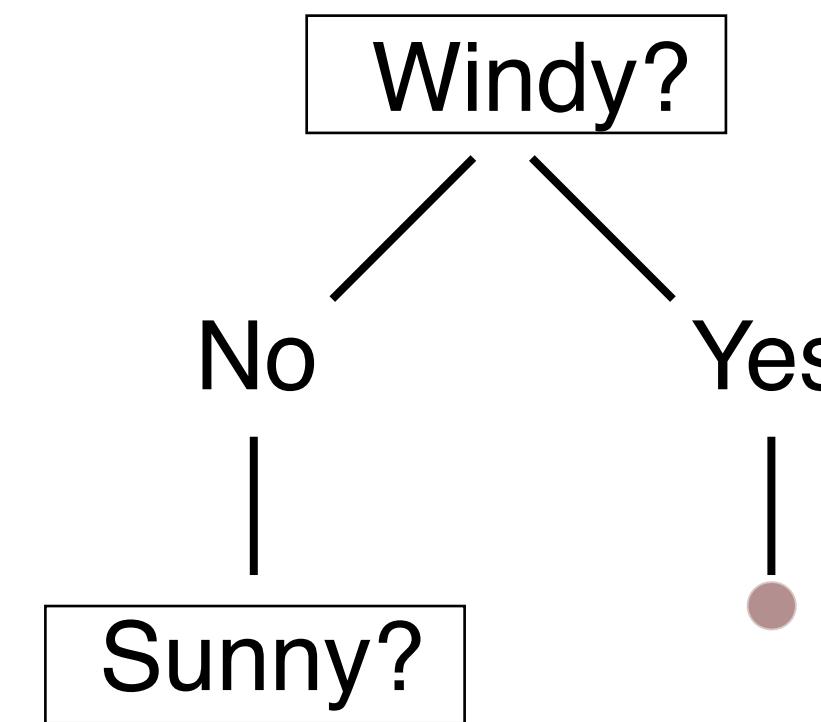
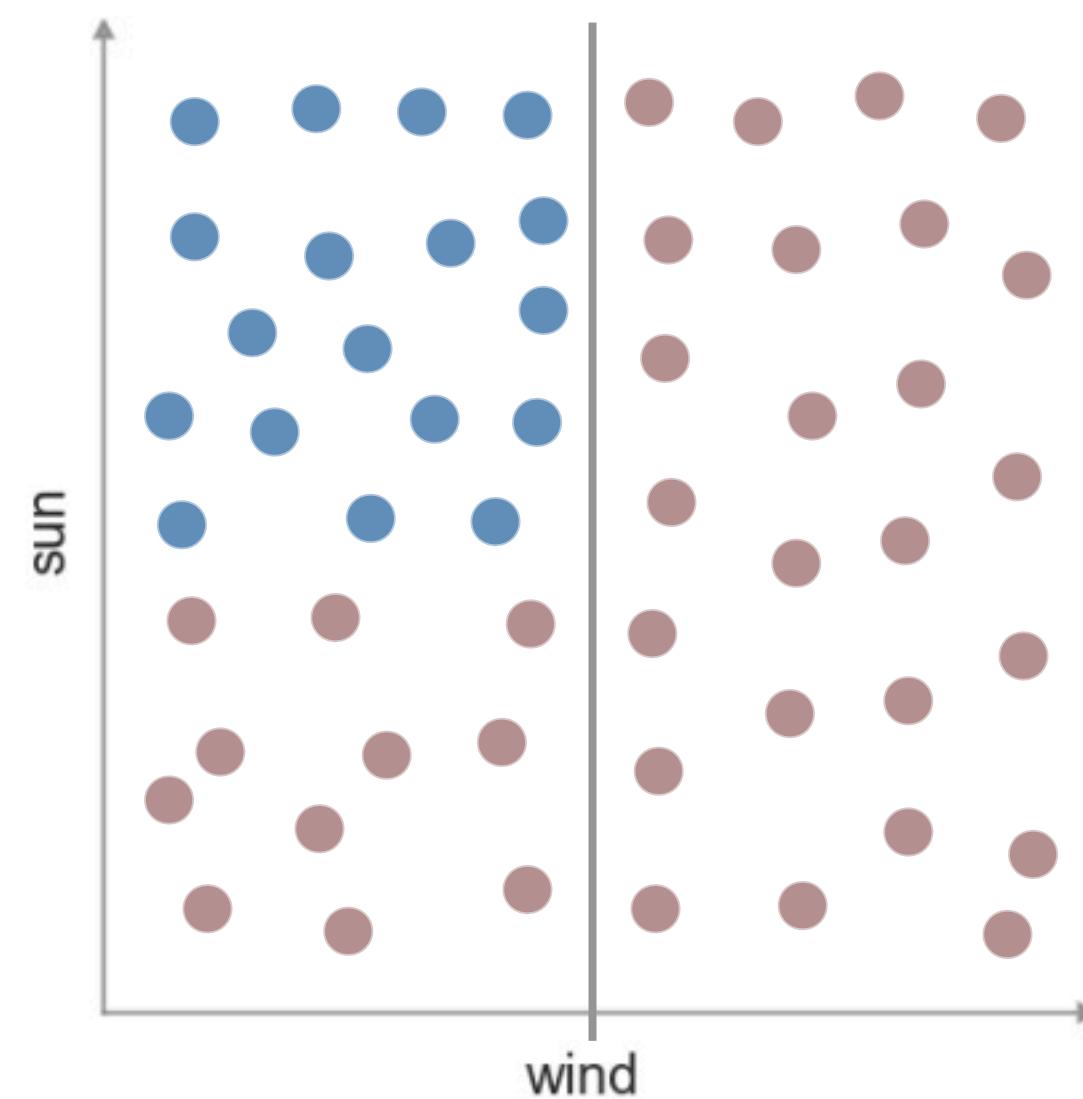
Decision trees allow you to ask multiple linear questions



# Simple Example

Decide whether you should play tennis

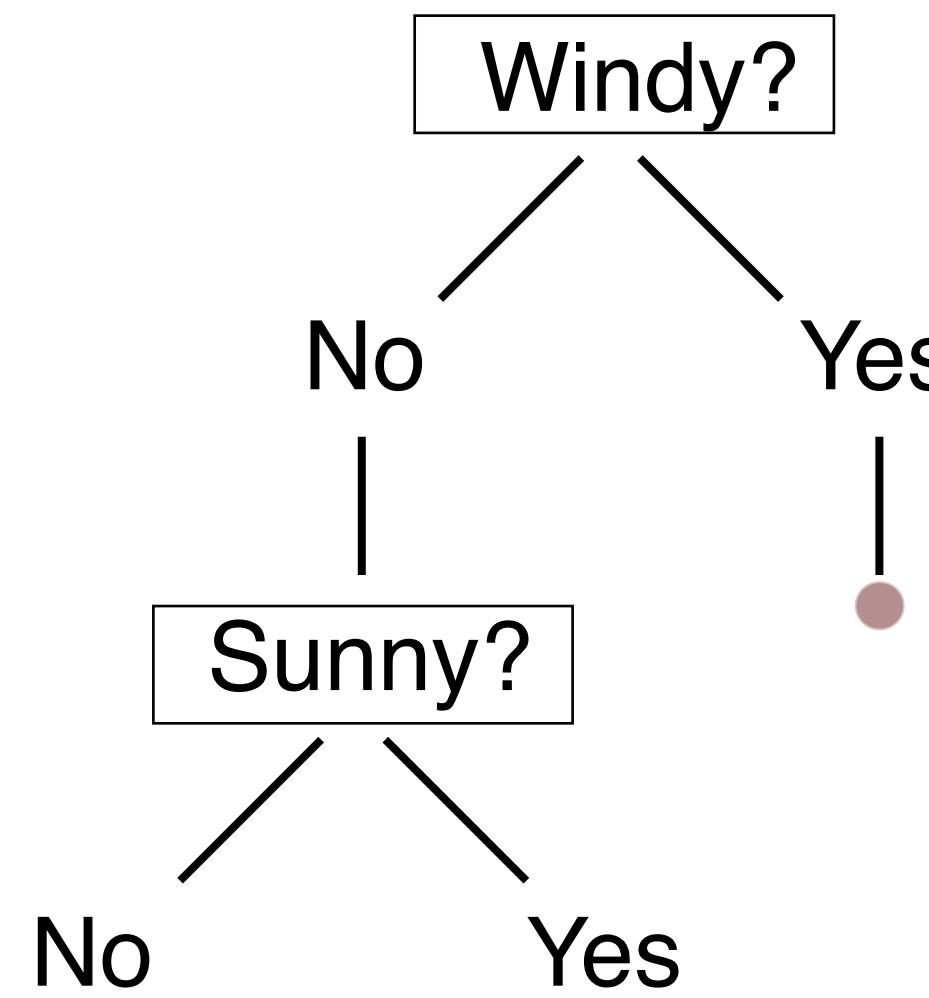
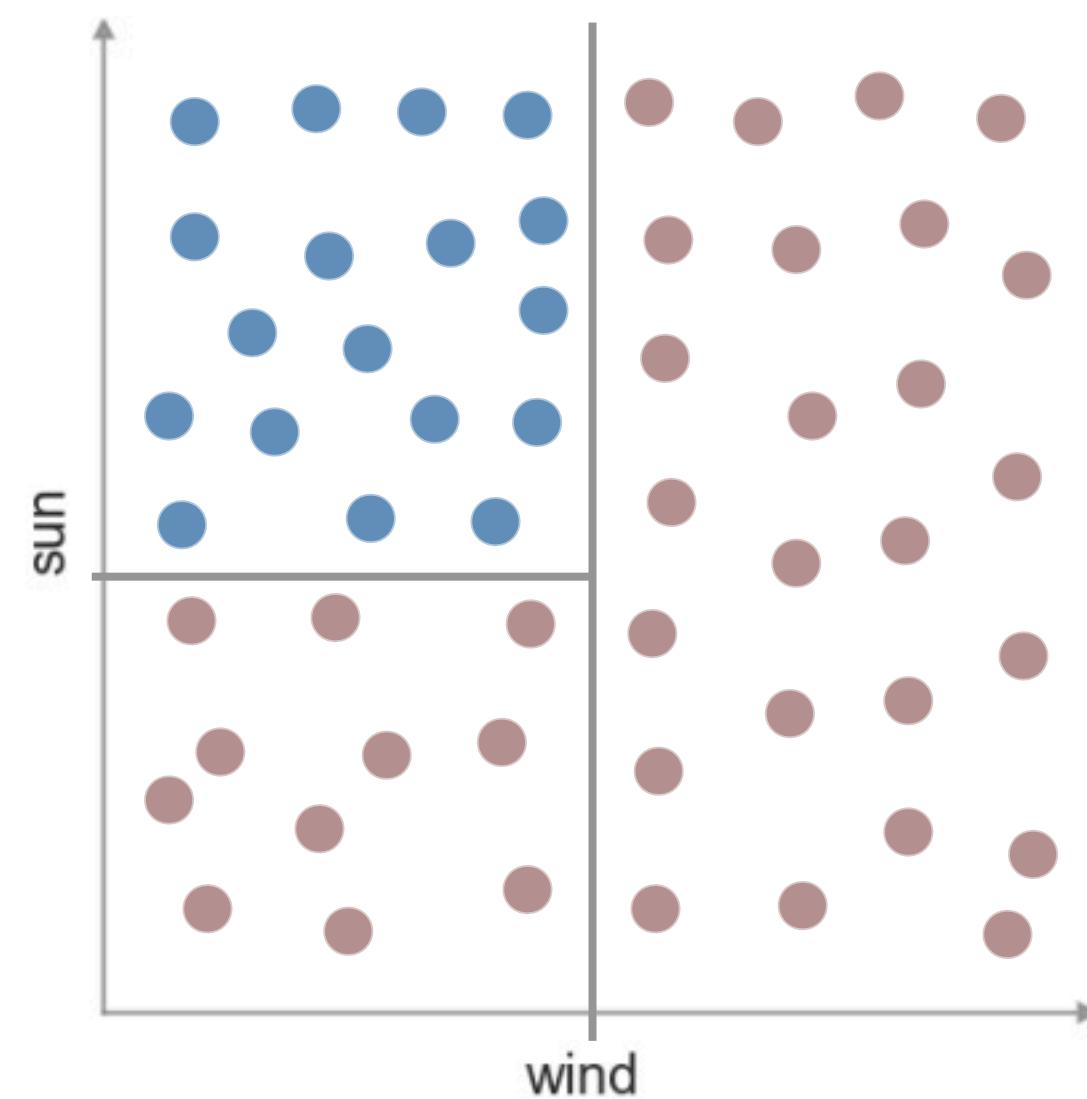
Decision trees allow you to ask multiple linear questions



# Simple Example

Decide whether you should play tennis

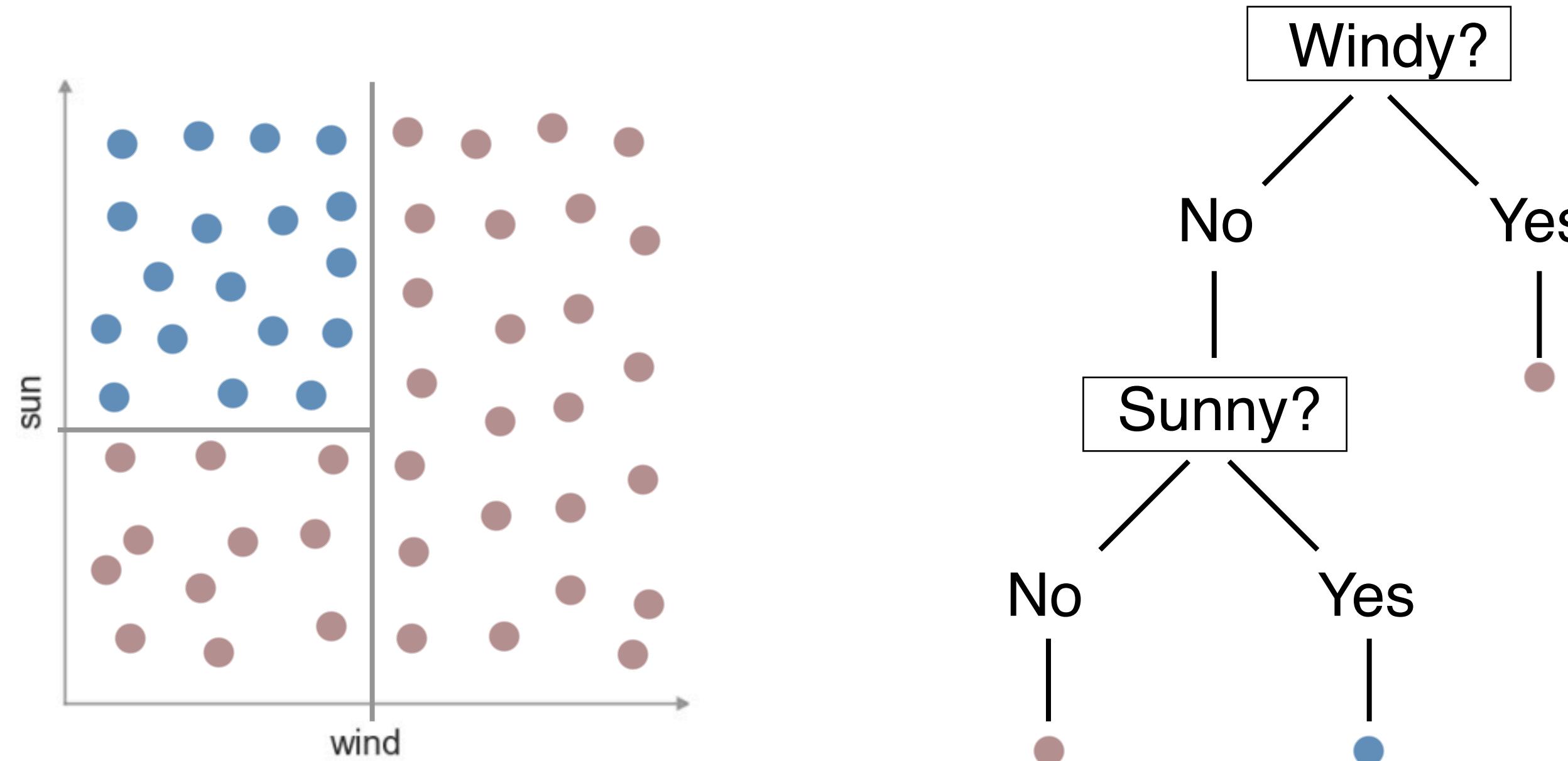
Decision trees allow you to ask multiple linear questions



# Simple Example

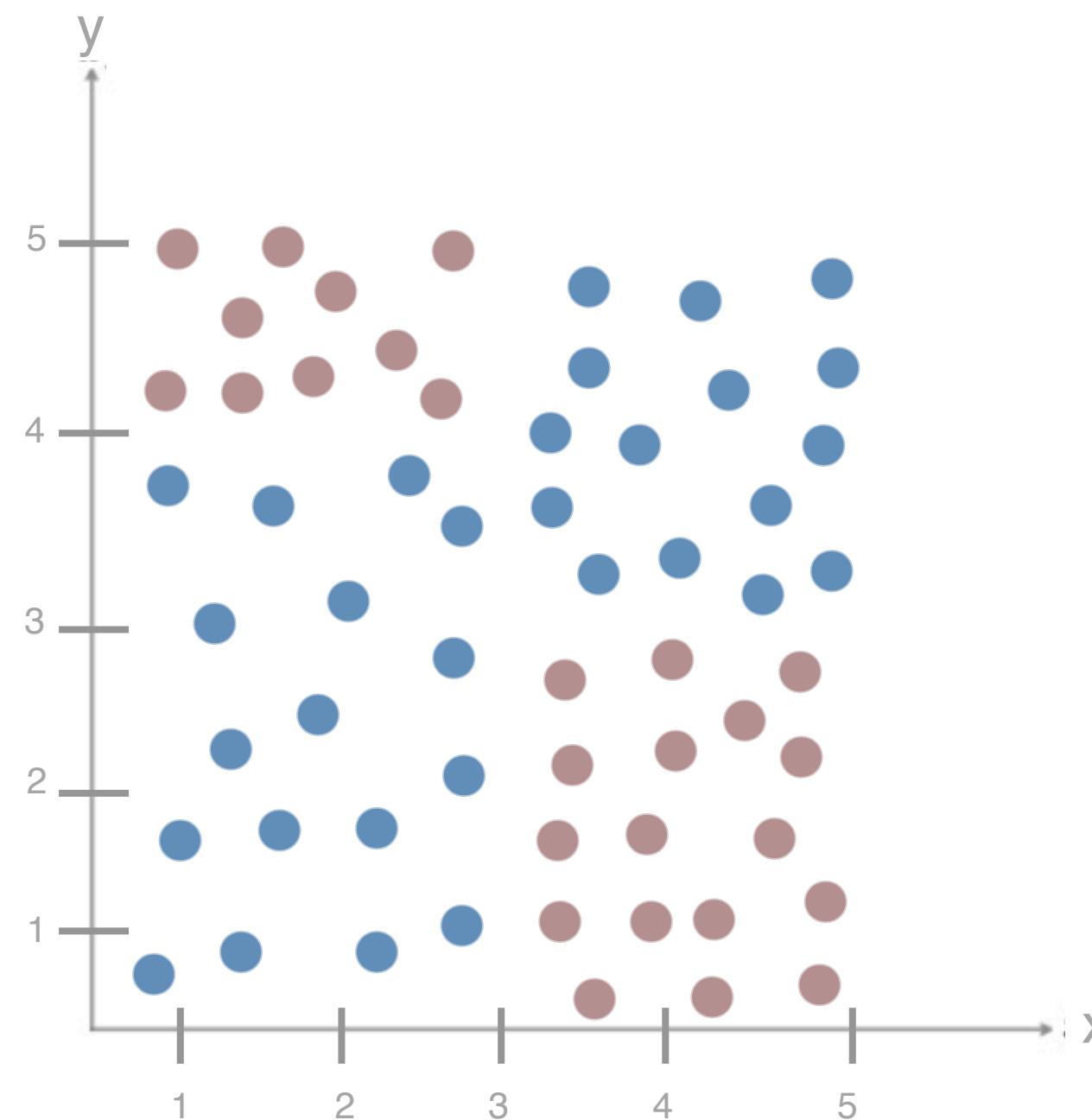
Decide whether you should play tennis

Decision trees allow you to ask multiple linear questions



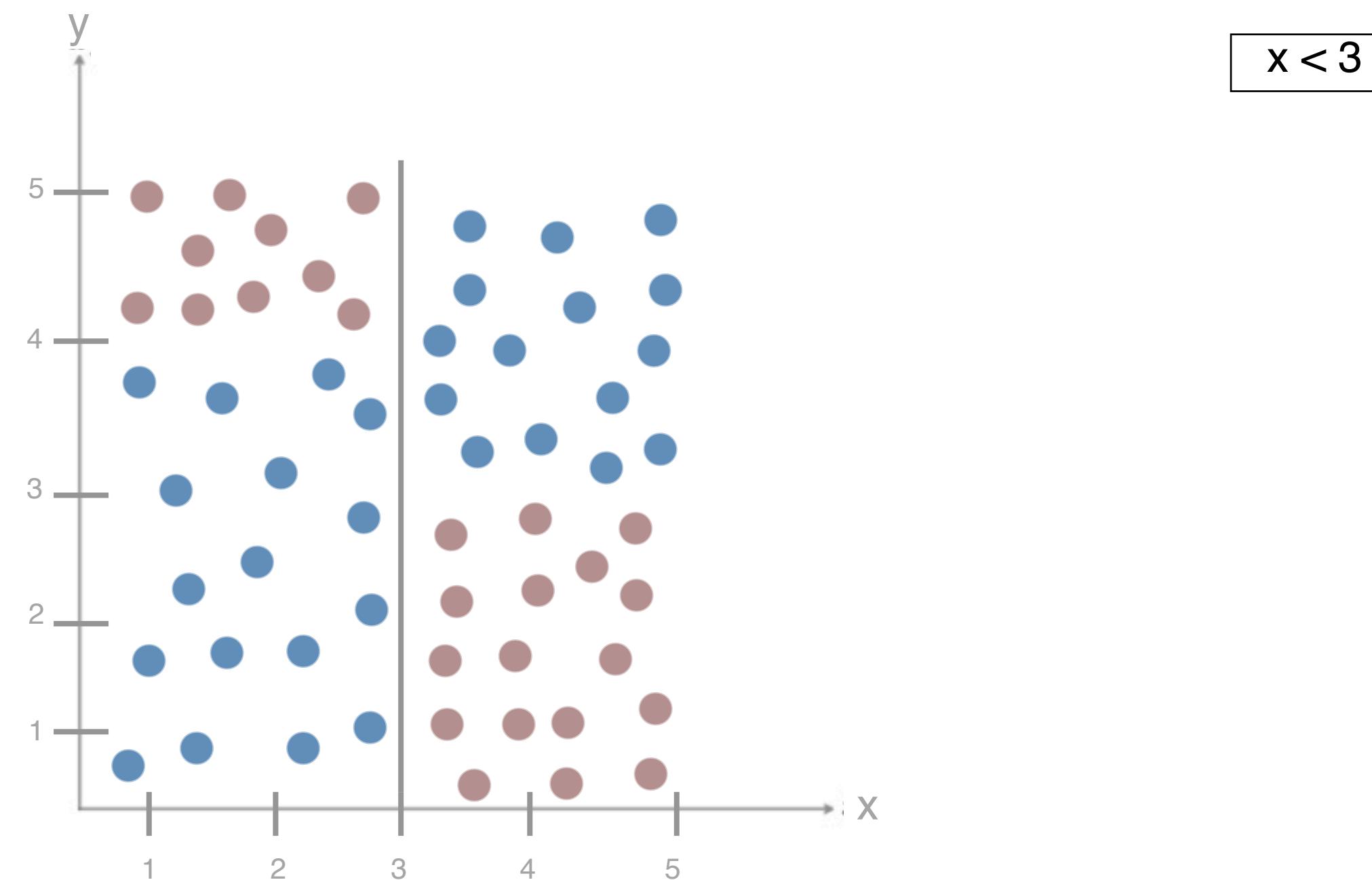
# Slightly Less Simple Example

Decision trees allow you to ask multiple linear questions



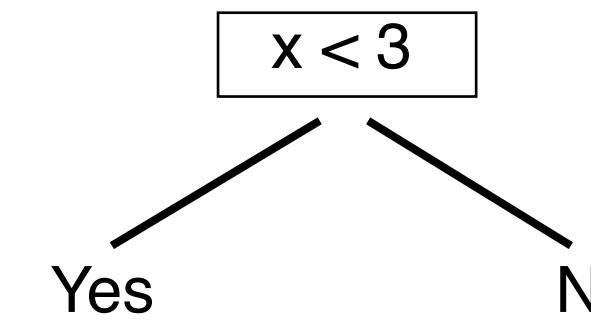
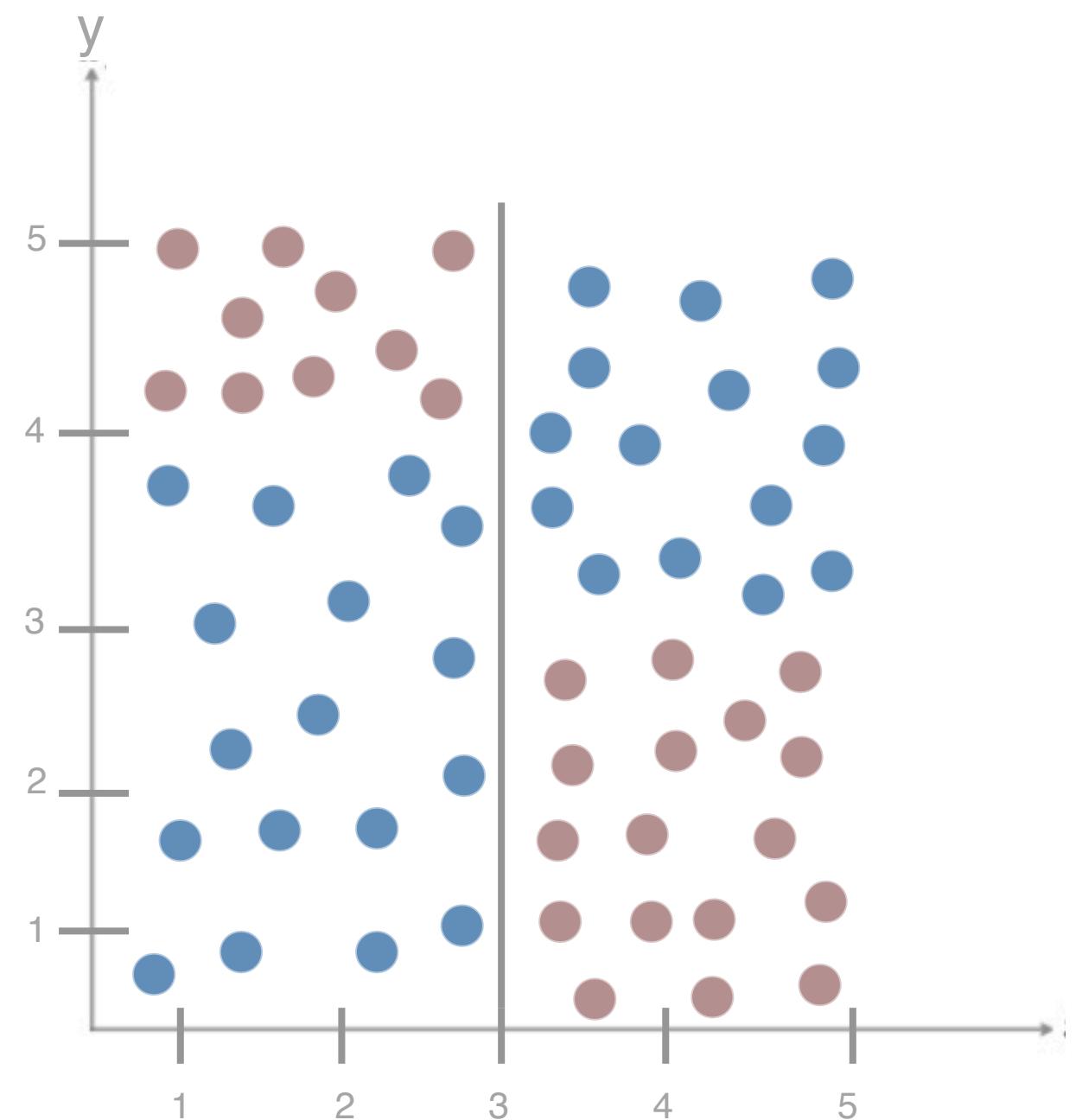
# Slightly Less Simple Example

Decision trees allow you to ask multiple linear questions



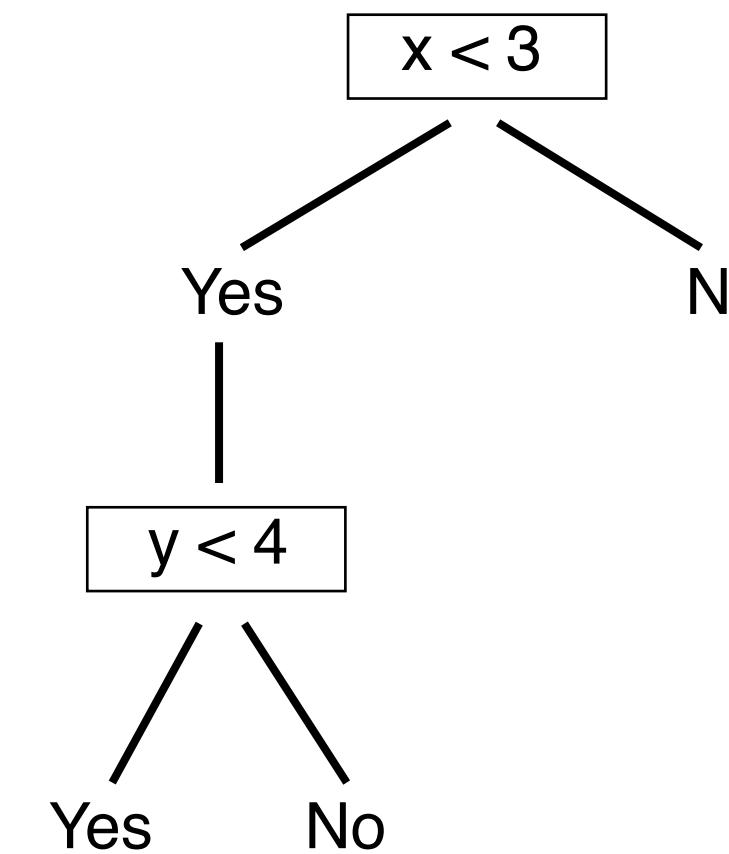
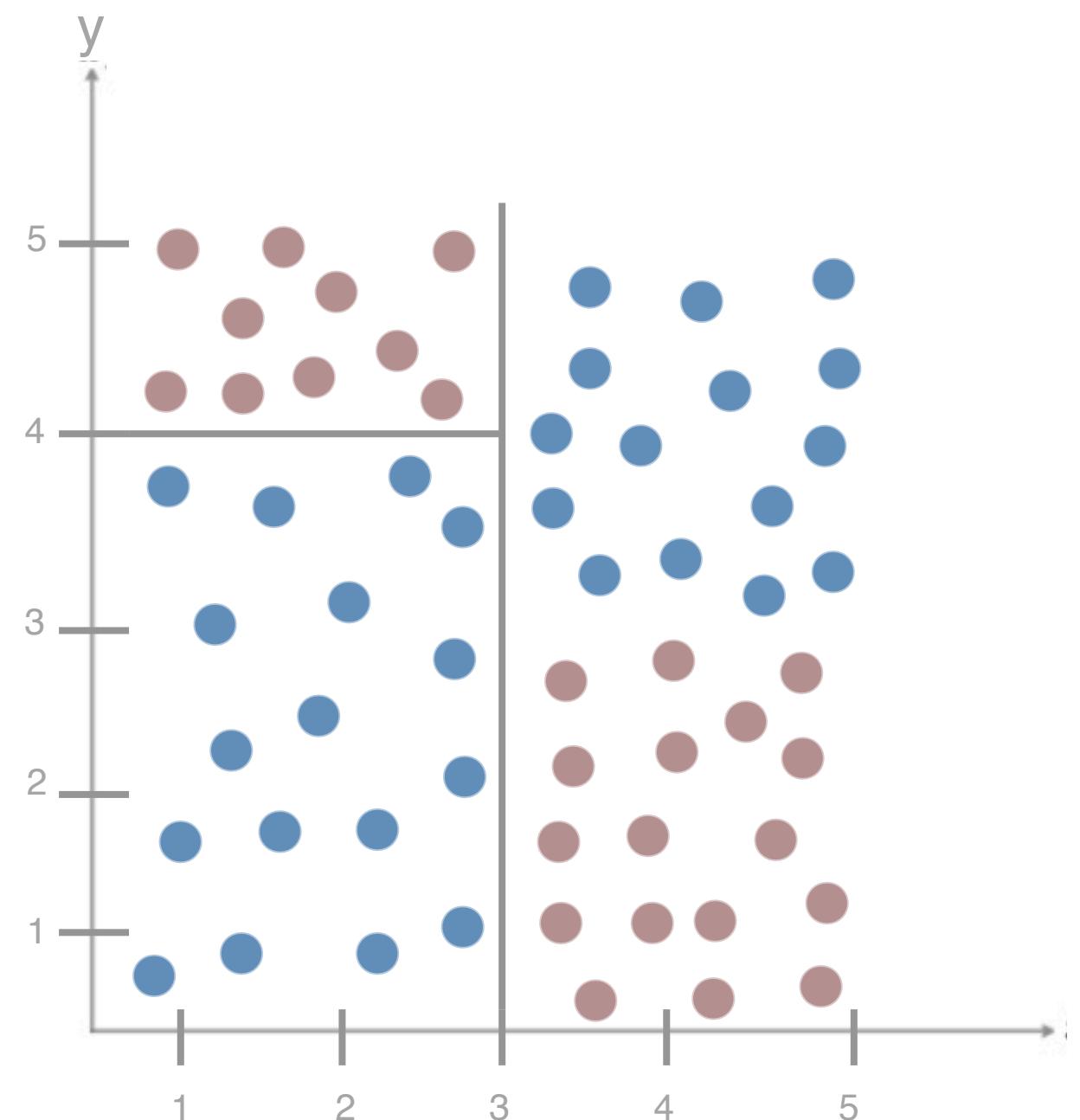
# Slightly Less Simple Example

Decision trees allow you to ask multiple linear questions



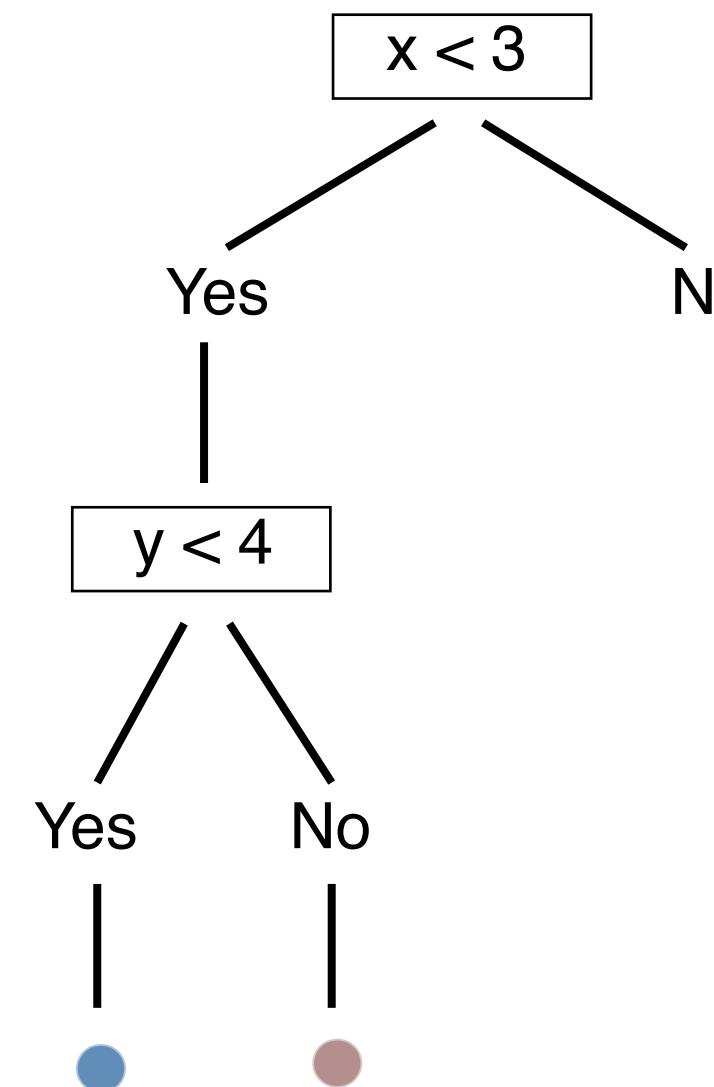
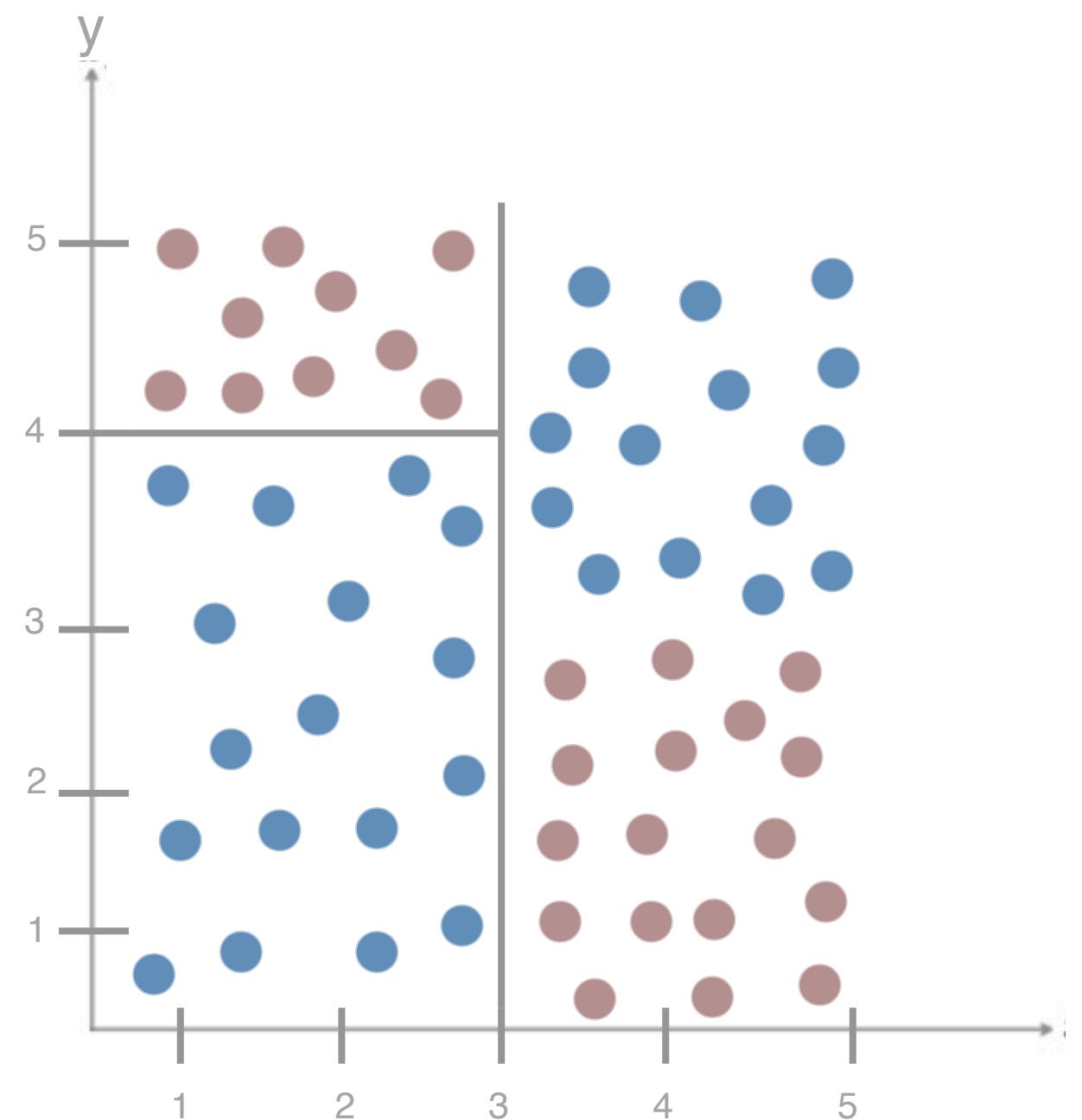
# Slightly Less Simple Example

Decision trees allow you to ask multiple linear questions



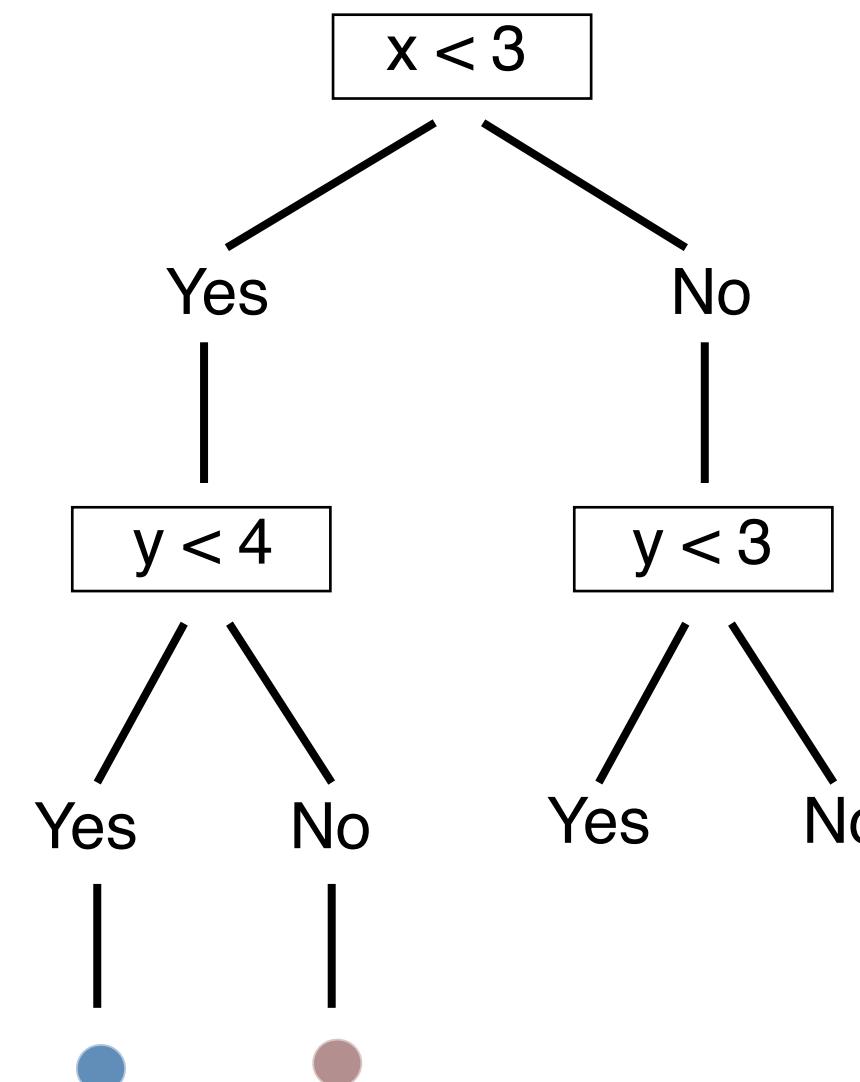
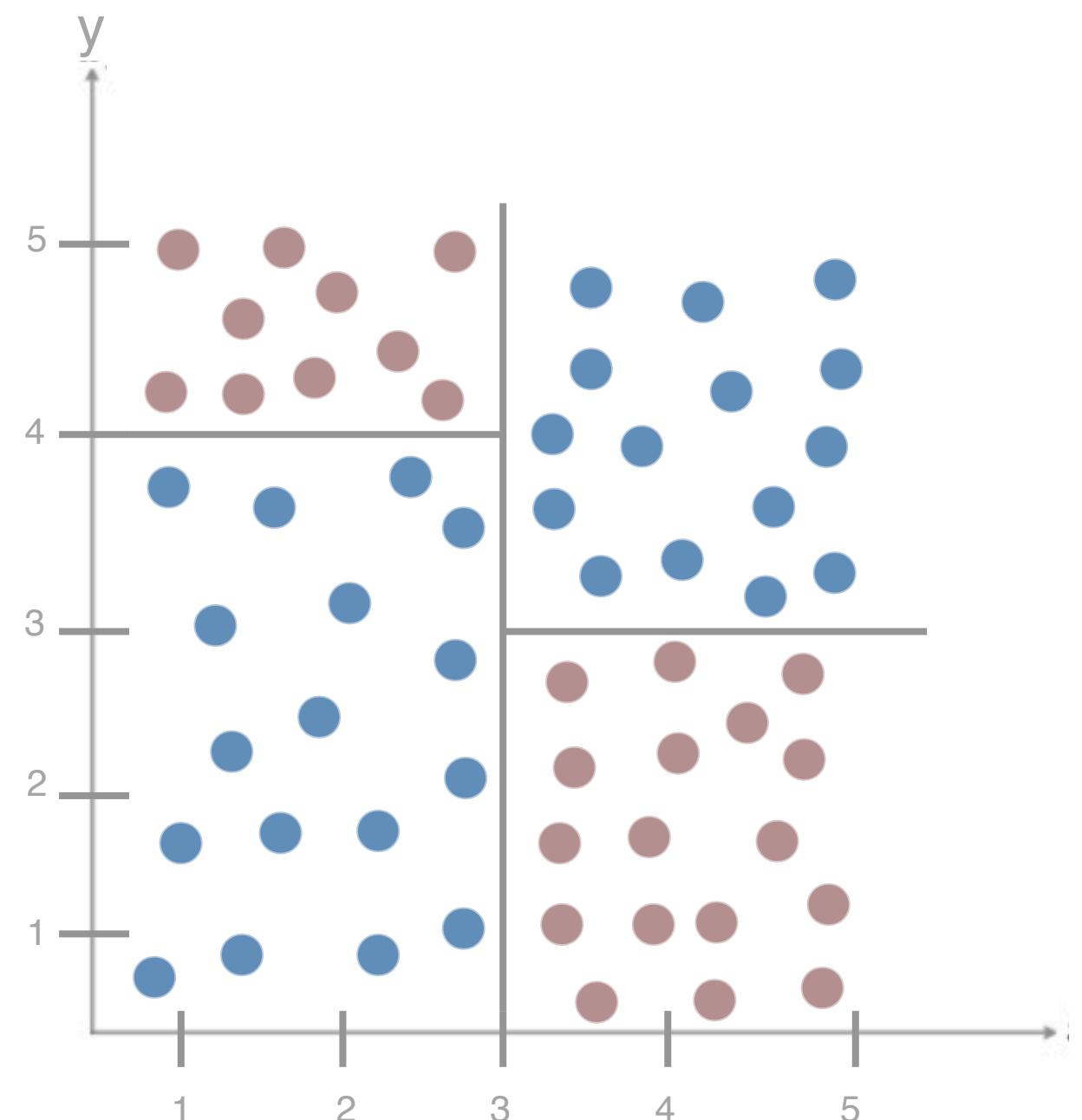
# Slightly Less Simple Example

Decision trees allow you to ask multiple linear questions



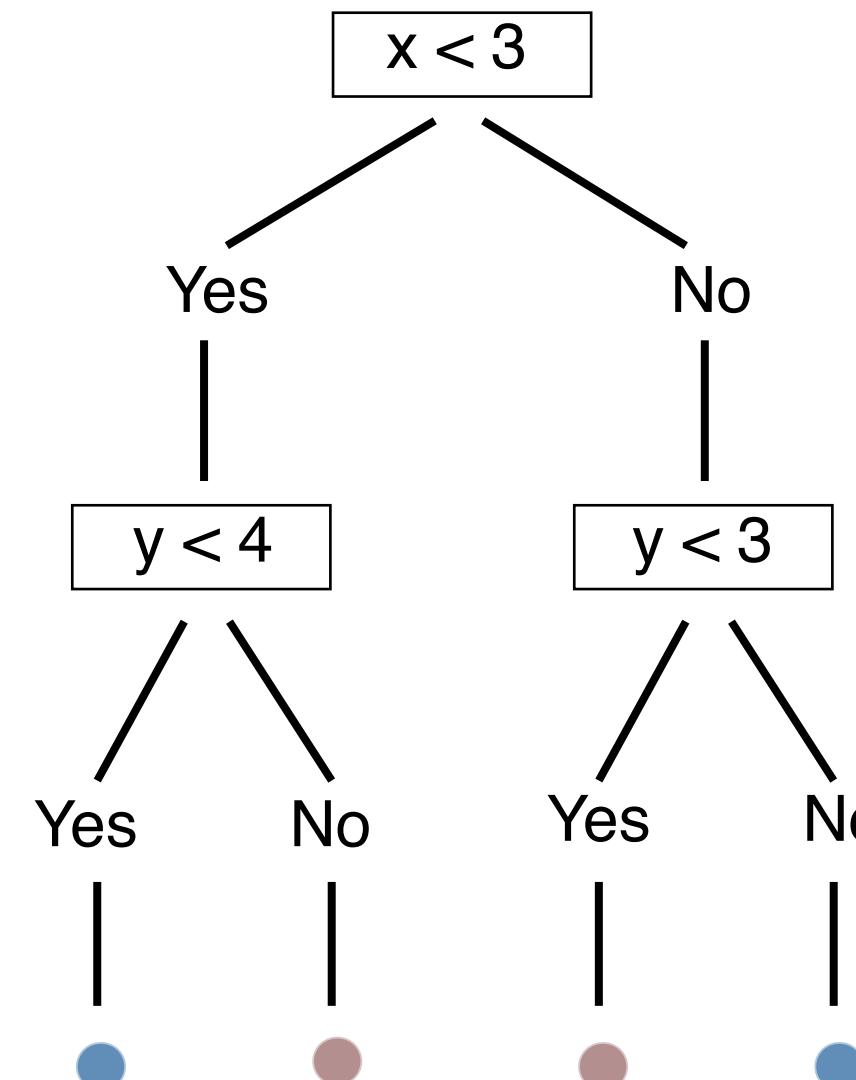
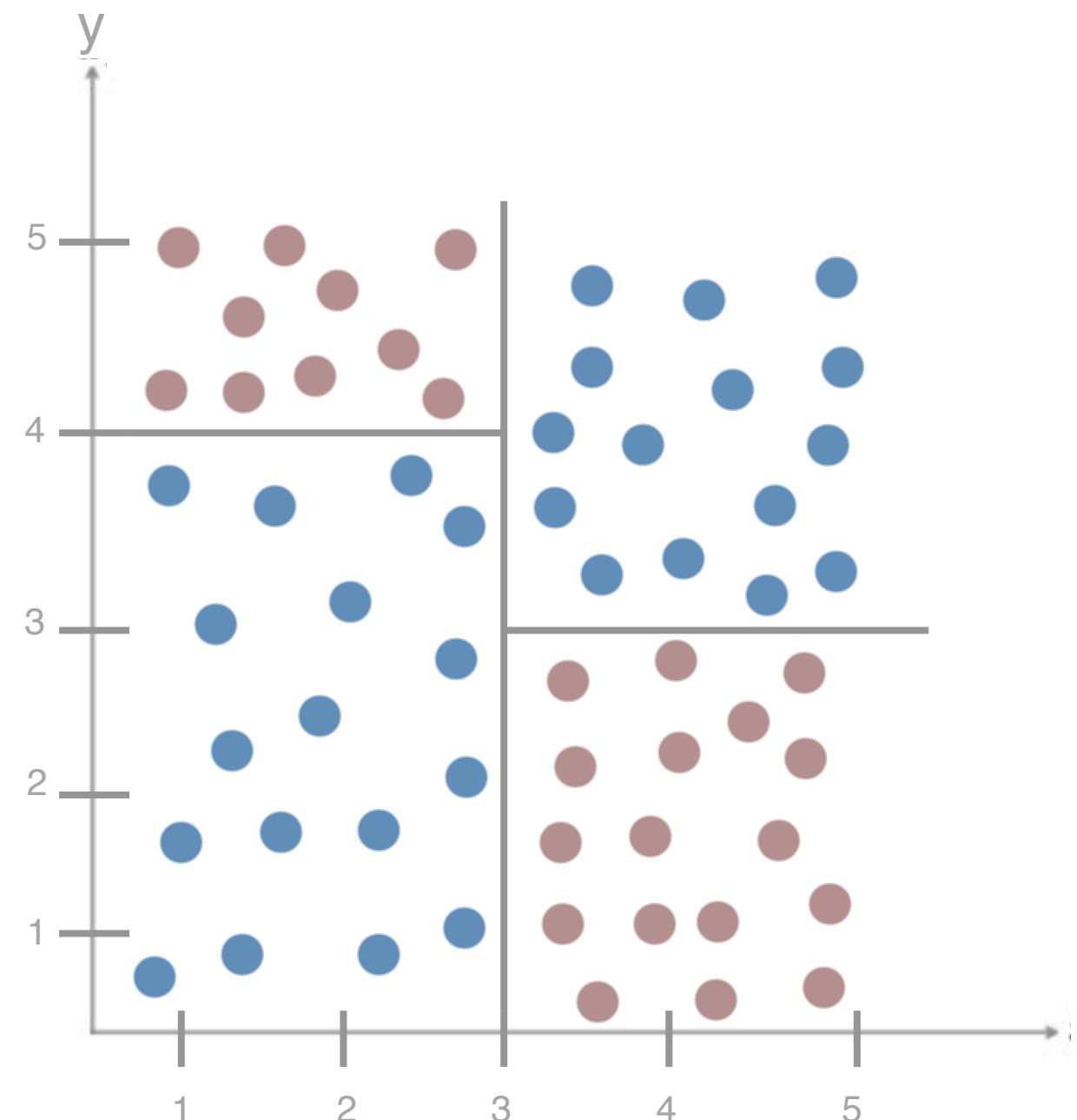
# Slightly Less Simple Example

Decision trees allow you to ask multiple linear questions



# Slightly Less Simple Example

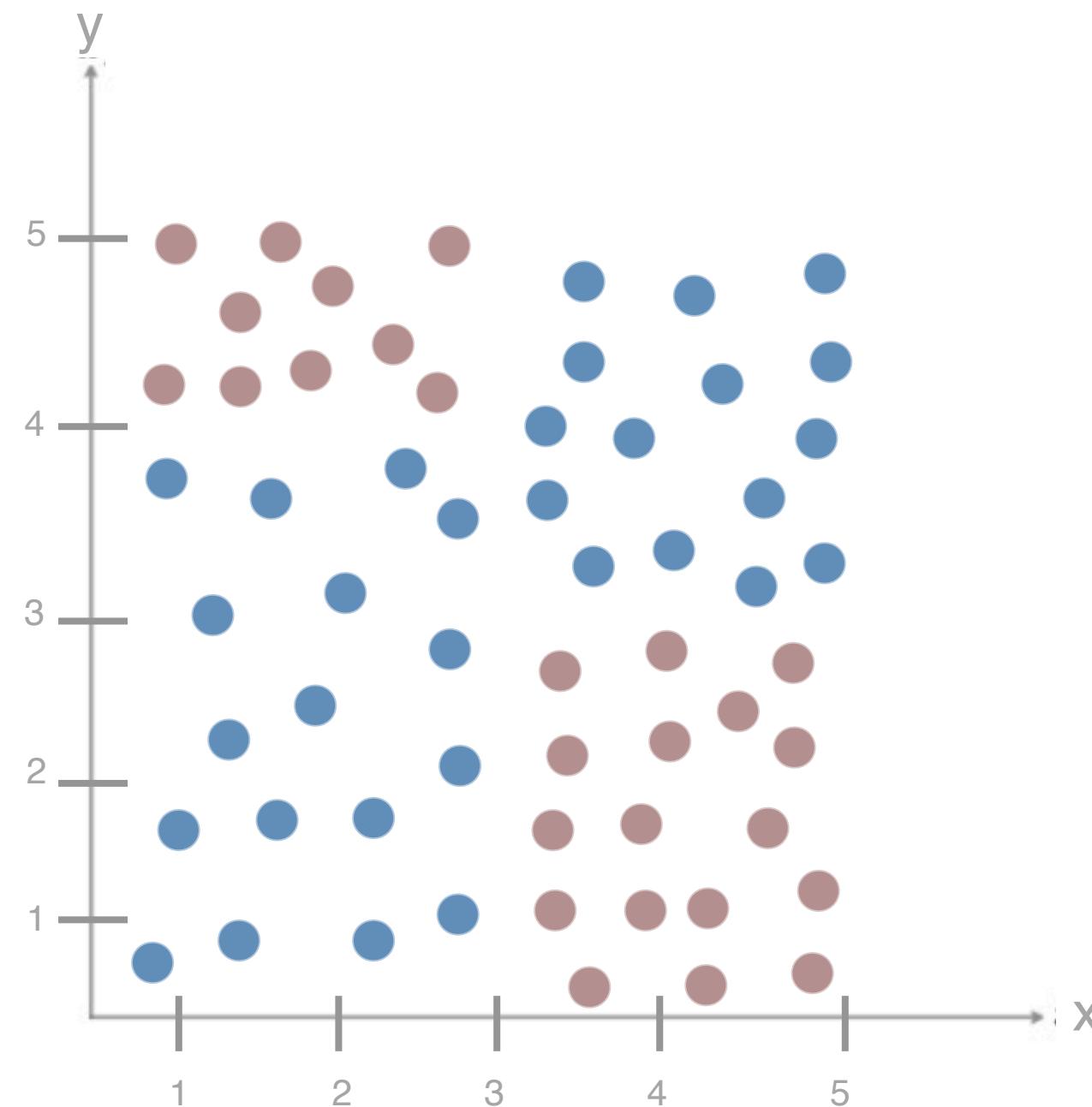
Decision trees allow you to ask multiple linear questions



# Slightly Less Simple Example

Notice that in this example we first split on x

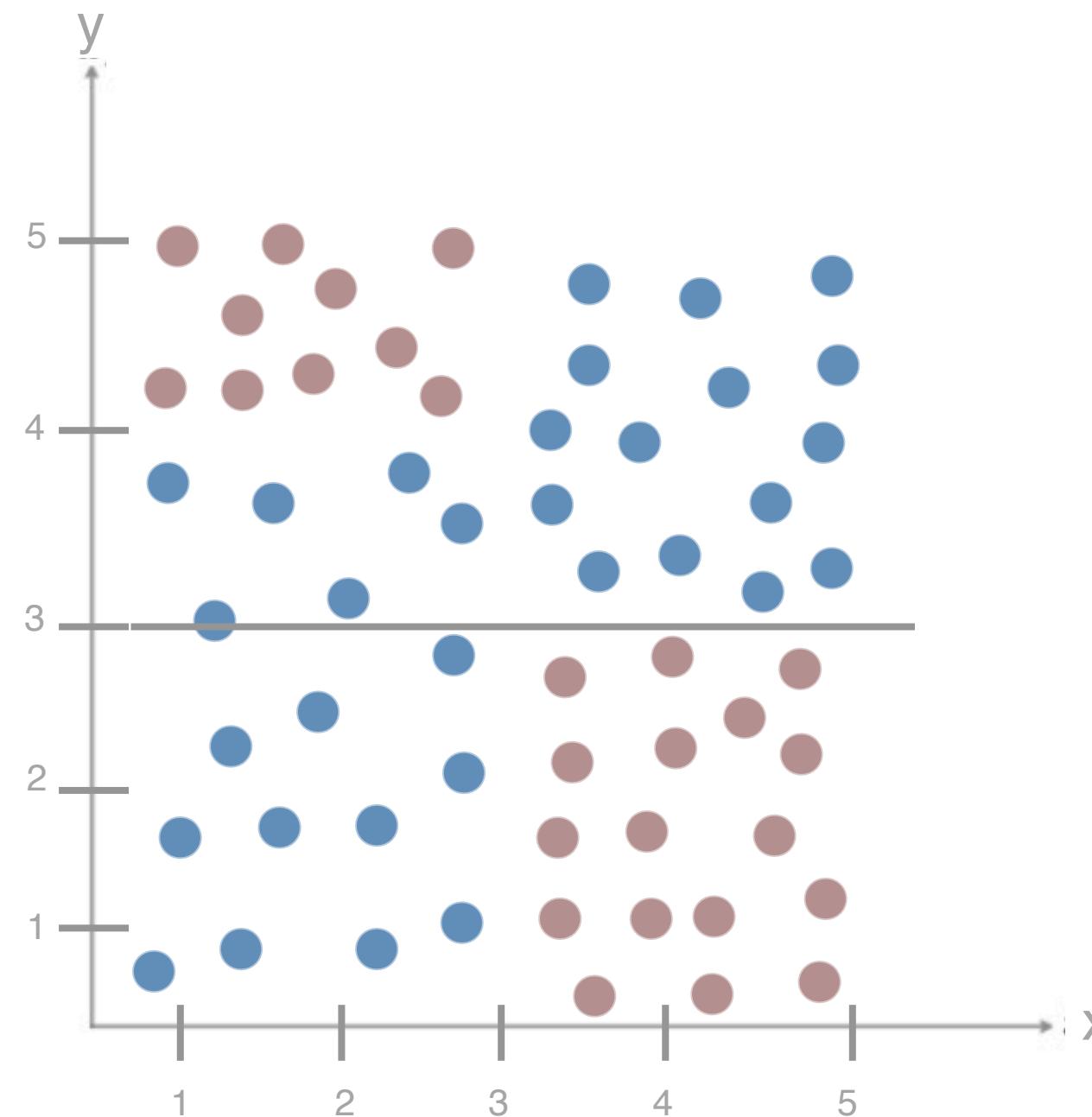
What if we split on y first?



# Slightly Less Simple Example

Notice that in this example we first split on x

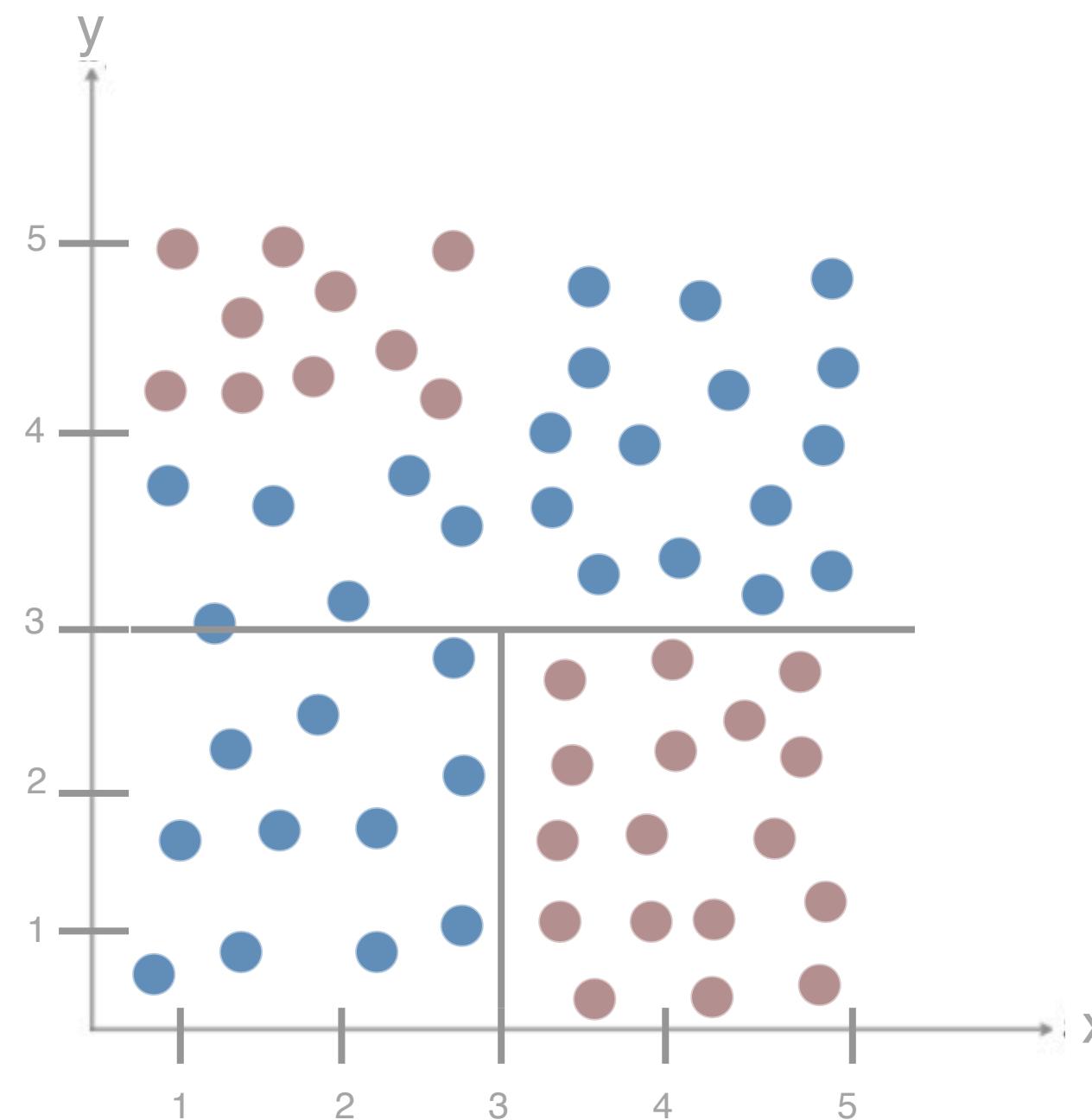
What if we split on y first?



# Slightly Less Simple Example

Notice that in this example we first split on x

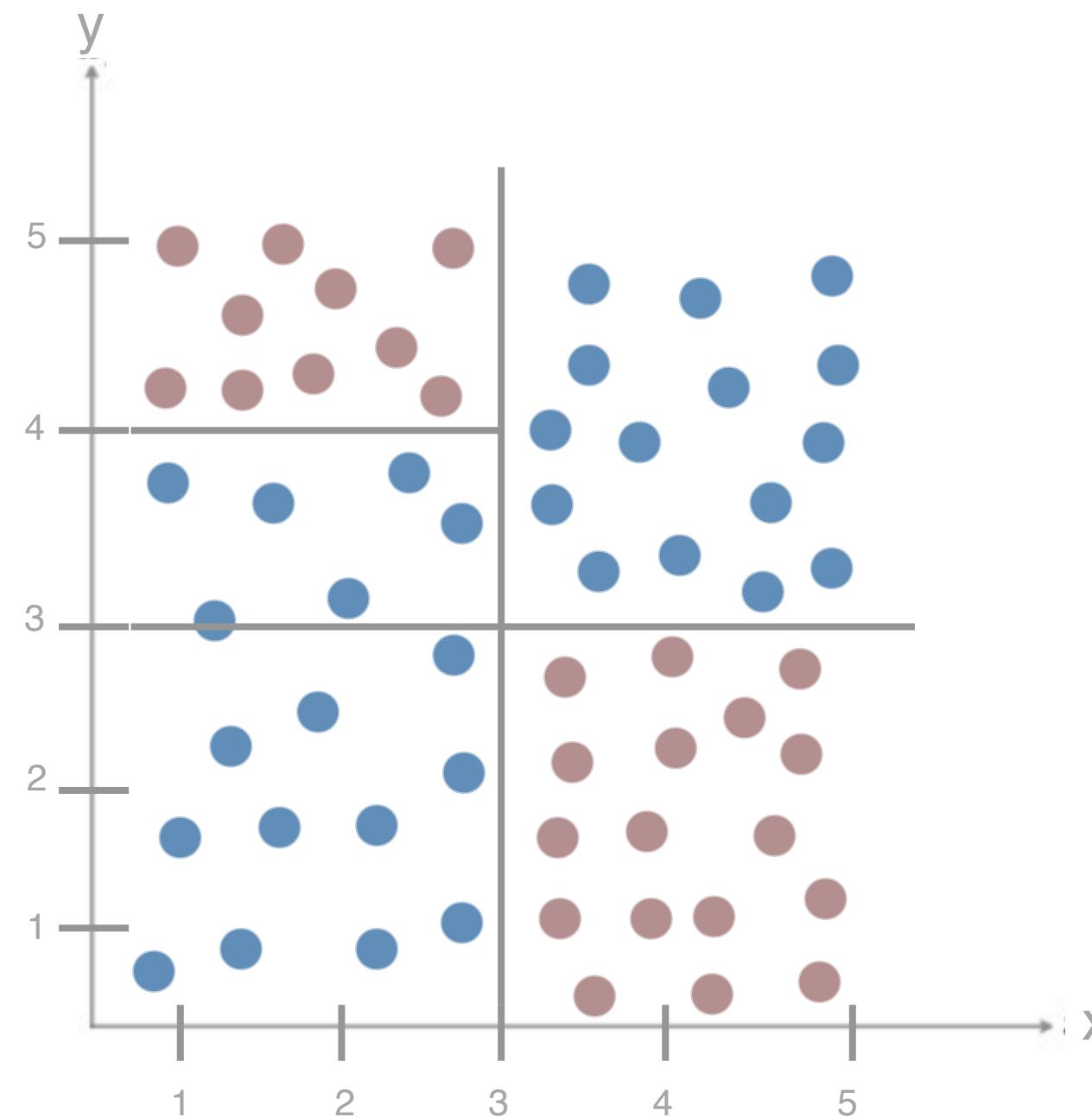
What if we split on y first?



# Slightly Less Simple Example

Notice that in this example we first split on x

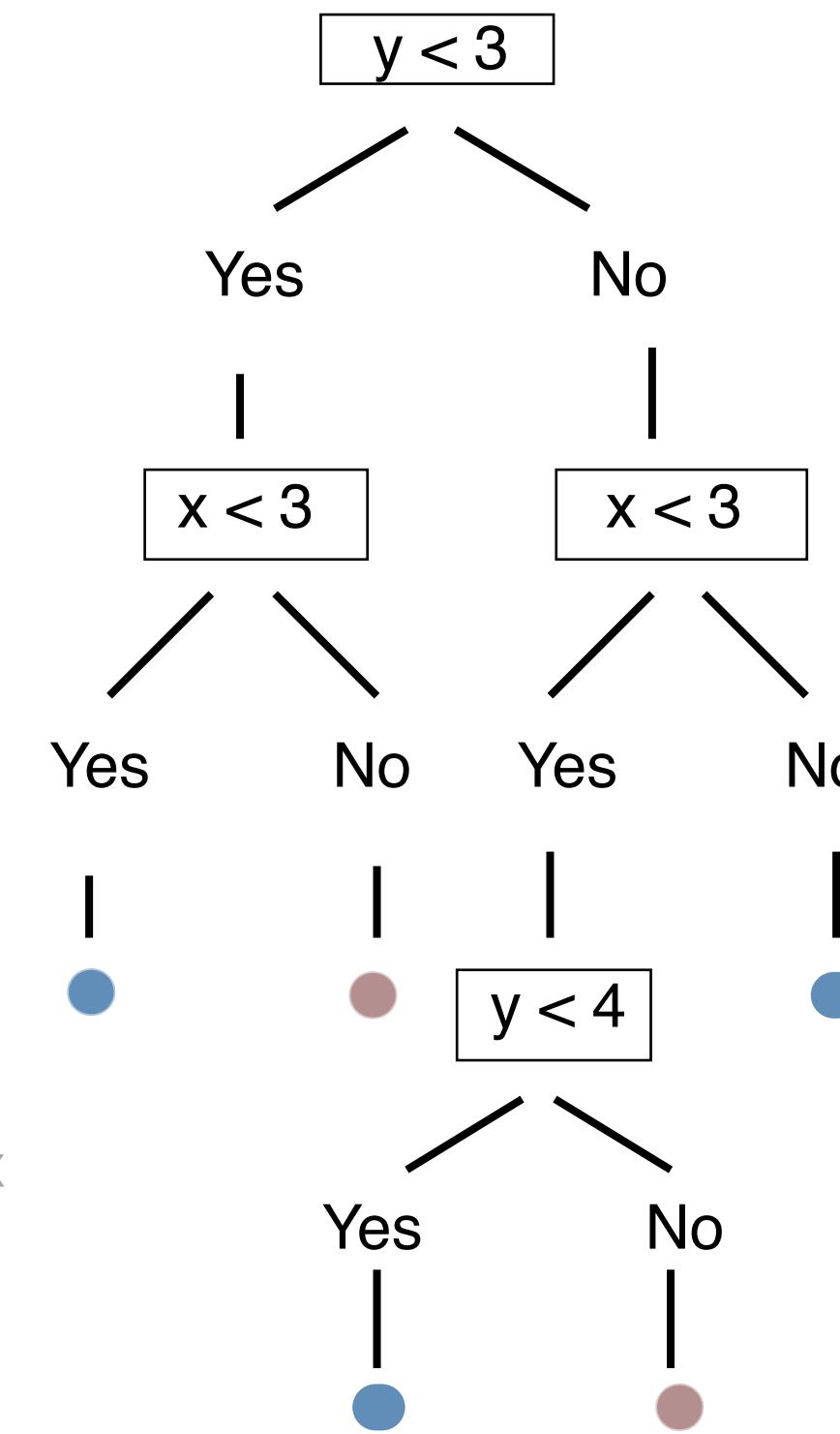
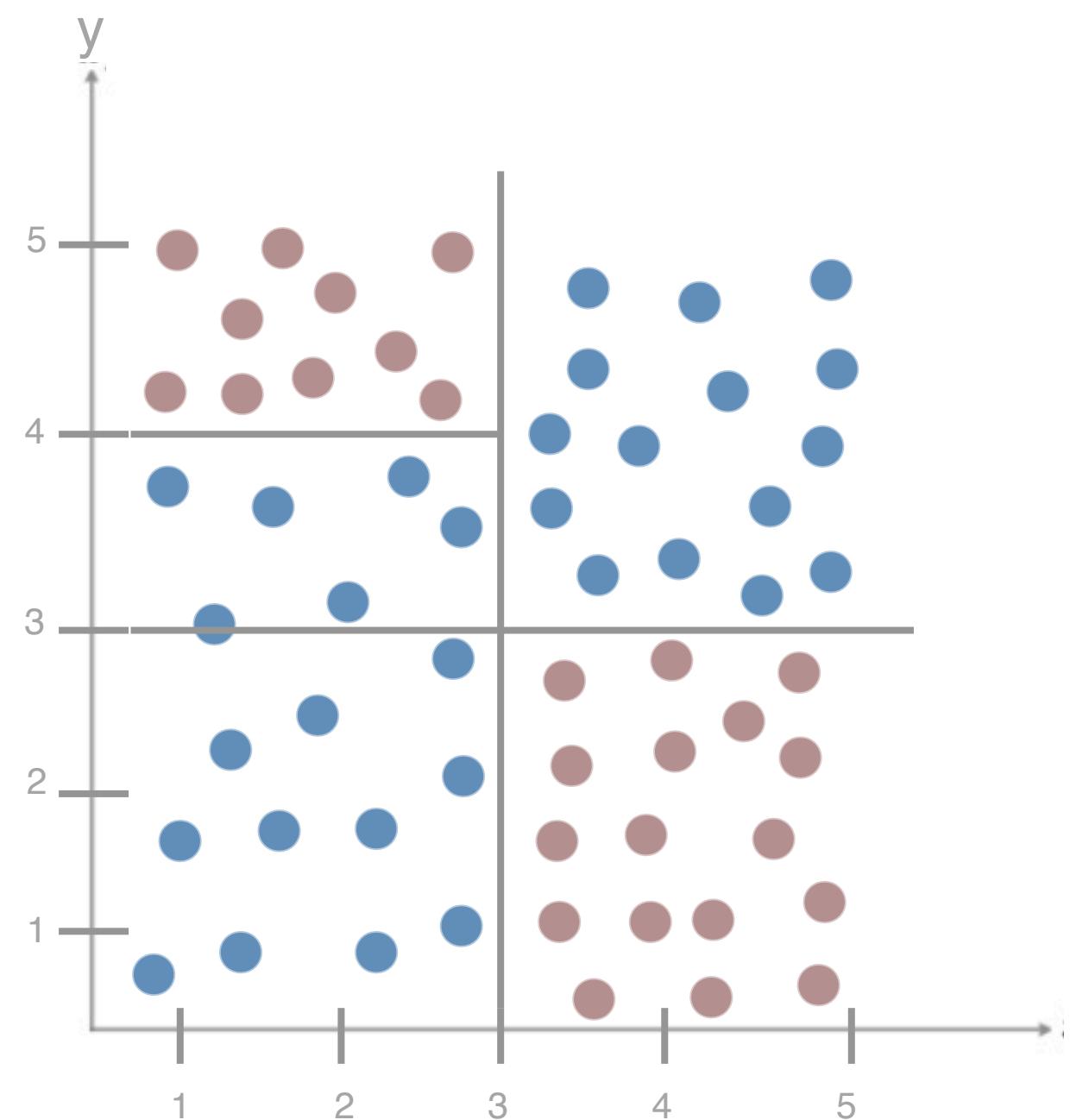
What if we split on y first?



# Slightly Less Simple Example

Notice that in this example we first split on x

What if we split on y first?



# Splitting Strategy

---

Can't check all possible partitionings

Instead we'll take a greedy approach

Decide which feature to best split on

Perform split

Repeat on each child node

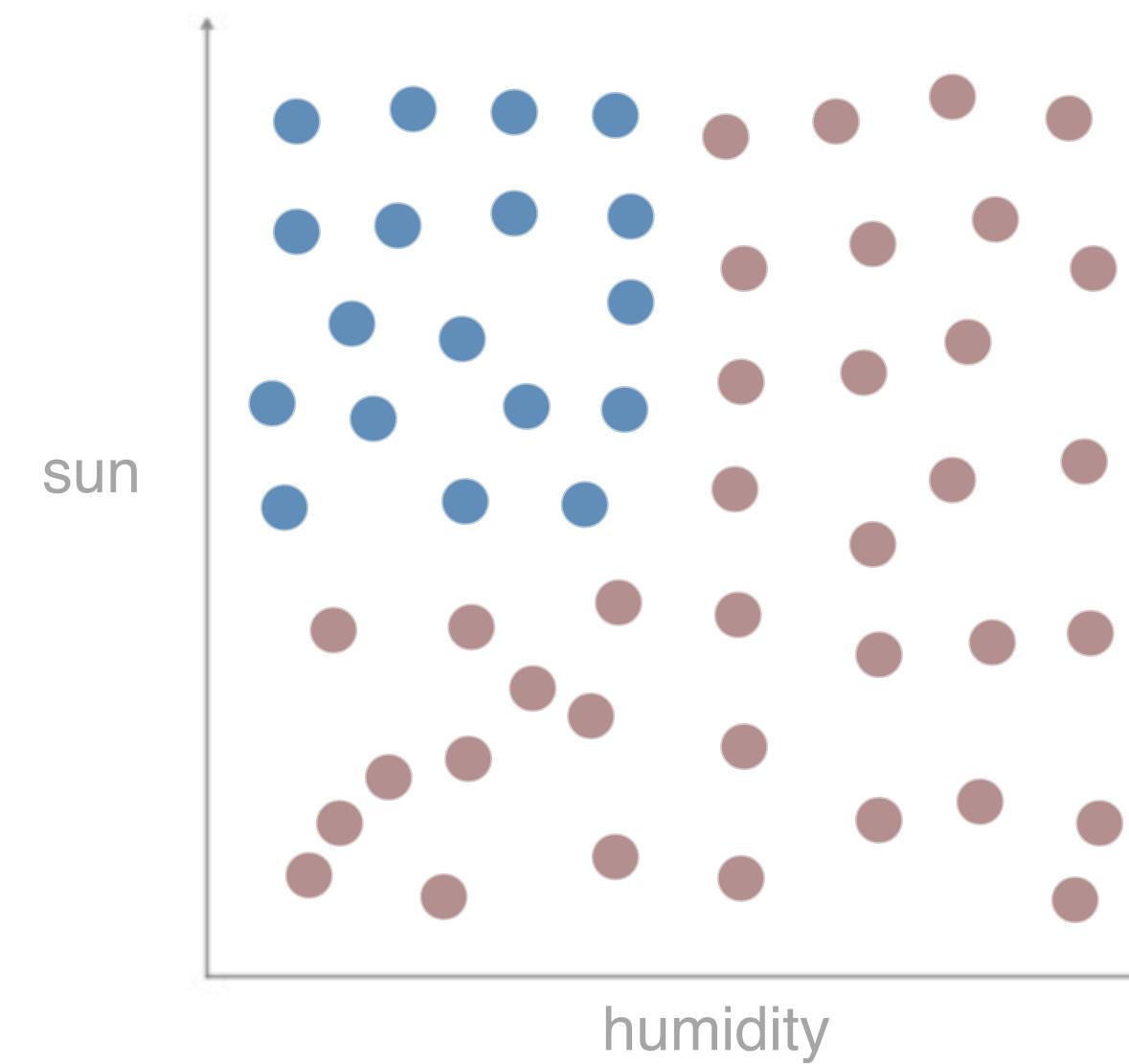
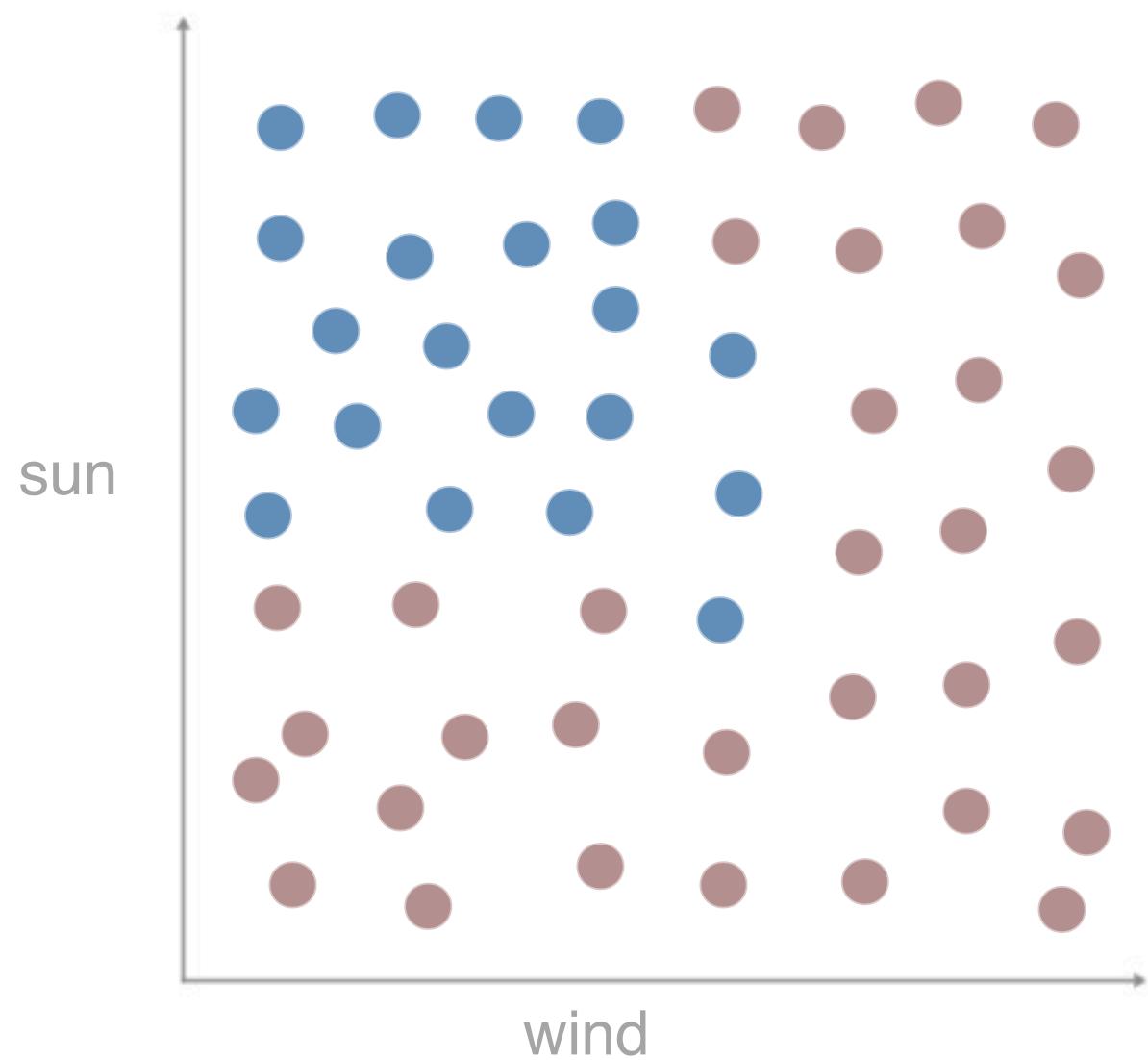
Terminate each branch when all training examples are same class

Splitting based on minimizing **Entropy**

**Entropy:** Measure of impurity of set of examples

# Splitting Strategy

**Entropy:** Measure of impurity of set of examples



# Entropy and Best Splitting

---

**Entropy:** Measure of impurity of set of examples (are there others??)

**Mathematical Representation:**

$$\text{entropy} = \sum_c -p_c \log_2(p_c)$$

where  $p_c$  is the fraction of examples in class c.

Note that for binary classification, let  $p$  be fraction in pos class, then

$$\text{entropy} = -p \log_2(p) - (1 - p) \log_2(1 - p)$$

**Question:** When is entropy the largest / smallest it can be?

# Entropy and Best Splitting

---

$$\text{entropy} = -p \log_2(p) - (1 - p) \log_2(1 - p)$$

**Question:** When is entropy the largest / smallest it can be?

**Answer:**

- When all examples are same class, entropy = 0
- When samples equally balanced, entropy = 1

# Entropy and Best Splitting

Example, consider the tennis problem now with binary features

<b>sun</b>	<b>wind</b>	<b>humidity</b>	<b>tennis</b>
sunny	windy	not humid	tennis
sunny	not windy	not humid	tennis
not sunny	not windy	humid	no tennis
sunny	windy	humid	no tennis

# Entropy and Best Splitting

Example, consider the tennis problem now with binary features

sun	wind	humidity	tennis
$S$	$W$	$\neg H$	$T$
$S$	$\neg W$	$\neg H$	$T$
$\neg S$	$\neg W$	$H$	$\neg T$
$S$	$W$	$H$	$\neg T$

**Question:** What is the entropy of the root node?

**Easy:** The root node is balanced in  $T$  and  $\neg T$  so the entropy is 1

# Entropy and Best Splitting

Example, consider the tennis problem now with binary features

<b>sun</b>	<b>wind</b>	<b>humidity</b>	<b>tennis</b>
$S$	$W$	$\neg H$	$T$
$S$	$\neg W$	$\neg H$	$T$
$\neg S$	$\neg W$	$H$	$\neg T$
$S$	$W$	$H$	$\neg T$

Let's check the math anyway. Denoting the fraction of  $T$  in the training set by  $p$ , we have  $p = \frac{1}{2}$

$$\text{entropy} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

# Entropy and Best Splitting

Example, consider the tennis problem now with binary features

<b>sun</b>	<b>wind</b>	<b>humidity</b>	<b>tennis</b>
$S$	$W$	$\neg H$	$T$
$S$	$\neg W$	$\neg H$	$T$
$\neg S$	$\neg W$	$H$	$\neg T$
$S$	$W$	$H$	$\neg T$

Let's check the math anyway. Denoting the fraction of  $T$  in the training set by  $p$ , we have  $p = \frac{1}{2}$

$$\text{entropy} = -\frac{1}{2}\log_2 2^{-1} - \frac{1}{2}\log_2 2^{-1}$$

# Entropy and Best Splitting

Example, consider the tennis problem now with binary features

<b>sun</b>	<b>wind</b>	<b>humidity</b>	<b>tennis</b>
$S$	$W$	$\neg H$	$T$
$S$	$\neg W$	$\neg H$	$T$
$\neg S$	$\neg W$	$H$	$\neg T$
$S$	$W$	$H$	$\neg T$

Let's check the math anyway. Denoting the fraction of  $T$  in the training set by  $p$ , we have  $p = \frac{1}{2}$

$$\text{entropy} = +\frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = 1$$

# Information Gain and Best Splitting

---

**Goal:** Pick a feature and split that decreases impurity the most

**Information Gain:** Consider a parent node and it's children after a split on feature  $\mathbf{x}_i$ . Define the following

- $D_{par}$ : training subset of the parent node
- $D_{left}$ : training subset of the left child node
- $D_{right}$ : training subset of the right child node
- $I$ : An impurity function (for now, entropy)

**Idea:** Information gain is the difference between impurity at the parent and (weighted average) of impurity at the children

# Information Gain and Best Splitting

**Goal:** Pick a feature and split that decreases impurity the most

**Information Gain:** Consider a parent node and it's children after a split on feature  $\mathbf{x}_i$ . Define the following

- $D_{par}$ : training subset of the parent node
- $D_{left}$ : training subset of the left child node
- $D_{right}$ : training subset of the right child node
- $I$ : An impurity function (for now, entropy)

$$IG(D_{par}, \mathbf{x}_i) = I(D_{par}) - \frac{|D_{left}|}{|D_{par}|} I(D_{left}) - \frac{|D_{right}|}{|D_{par}|} I(D_{right})$$

# Information Gain and Best Splitting

---

$$IG(D_{par}, \mathbf{x}_i) = I(D_{par}) - \frac{|D_{left}|}{|D_{par}|} I(D_{left}) - \frac{|D_{right}|}{|D_{par}|} I(D_{right})$$

For each feature  $\mathbf{x}_i$  compute it's information gain

Split on feature with largest information gain

# Information Gain and Best Splitting

Determine the best first splitting for the Tennis data

<b>sun</b>	<b>wind</b>	<b>humidity</b>	<b>tennis</b>
$S$	$W$	$\neg H$	$T$
$S$	$\neg W$	$\neg H$	$T$
$\neg S$	$\neg W$	$H$	$\neg T$
$S$	$W$	$H$	$\neg T$

We need to compute  $IG(D_{par}, \mathbf{x}_i)$  for each feature, where  $D_{par}$  is the full training set

# Information Gain and Best Splitting

---

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

**Q:** What is  $I(D_{par})$ ?

# Information Gain and Best Splitting

---

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

**Q:** What is  $I(D_{par})$ ?

**A:** Already computed this,  $I(D_{par}) = 1$

# Information Gain and Best Splitting

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

$$I(D_{par}) = 1,$$

**Q:** What are the weights in the average:  $\frac{|D_{left}|}{|D_{par}|}$  and  $\frac{|D_{right}|}{|D_{par}|}$  ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{3}{4}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{4}$$

**Q:** What is the entropy  $I(D_{left})$ ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{3}{4}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{4}$$

$$p = \frac{2}{3} \Rightarrow I(D_{left}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.918$$

# Information Gain and Best Splitting

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

$$I(D_{par}) = 1, \frac{|D_{left}|}{|D_{par}|} = \frac{3}{4}, \frac{|D_{right}|}{|D_{par}|} = \frac{1}{4}, I(D_{left}) = 0.918$$

**Q:** What is  $I(D_{right})$ ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{3}{4}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{4}, \quad I(D_{left}) = 0.918$$

$$p = 0 \Rightarrow I(D_{right}) = 0$$

# Information Gain and Best Splitting

$IG(D_{par}, \text{sun})$ :

- $D_{par} = \{(S, T), (S, T), (\neg S, \neg T), (S, \neg T)\}$
- $D_{left} = \{(S, T), (S, T), (S, \neg T)\}$
- $D_{right} = \{(\neg S, \neg T)\}$

**Q:** What is  $IG(D_{par}, \text{sun})$ ?

$$\begin{aligned} IG(D_{par}, \text{sun}) &= I(D_{par}) - \frac{|D_{left}|}{|D_{par}|} I(D_{left}) - \frac{|D_{right}|}{|D_{par}|} I(D_{right}) \\ &= 1 - \frac{3}{4} \cdot 0.918 - \frac{1}{4} \cdot 0 \\ &= 0.3112 \end{aligned}$$

# Information Gain and Best Splitting

Determine the best first splitting for the Tennis data

<b>sun</b>	<b>wind</b>	<b>humidity</b>	<b>tennis</b>
$S$	$W$	$\neg H$	$T$
$S$	$\neg W$	$\neg H$	$T$
$\neg S$	$\neg W$	$H$	$\neg T$
$S$	$W$	$H$	$\neg T$

We need to compute  $IG(D_{par}, \mathbf{x}_i)$  for each feature, where  $D_{par}$  is the full training set

# Information Gain and Best Splitting

---

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

**Q:** What is  $I(D_{par})$ ?

# Information Gain and Best Splitting

---

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

**Q:** What is  $I(D_{par})$ ?

**A:** Already computed this,  $I(D_{par}) = 1$

# Information Gain and Best Splitting

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

$$I(D_{par}) = 1,$$

**Q:** What are the weights in the average:  $\frac{|D_{left}|}{|D_{par}|}$  and  $\frac{|D_{right}|}{|D_{par}|}$  ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}$$

**Q:** What is the entropy  $I(D_{left})$ ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}$$

$$p = \frac{1}{2} \Rightarrow I(D_{left}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

# Information Gain and Best Splitting

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}, \quad I(D_{left}) = 1$$

**Q:** What is  $I(D_{right})$ ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}, \quad I(D_{left}) = 1$$

$$p = \frac{1}{2} \Rightarrow I(D_{right}) = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2} = 1$$

# Information Gain and Best Splitting

$IG(D_{par}, \text{wind})$ :

- $D_{par} = \{(W, T), (\neg W, T), (\neg W, \neg T), (W, \neg T)\}$
- $D_{left} = \{(W, T), (W, \neg T)\}$
- $D_{right} = \{(\neg W, \neg T), (\neg W, T)\}$

**Q:** What is  $IG(D_{par}, \text{wind})$ ?

$$\begin{aligned} IG(D_{par}, \text{wind}) &= I(D_{par}) - \frac{|D_{left}|}{|D_{par}|} I(D_{left}) - \frac{|D_{right}|}{|D_{par}|} I(D_{right}) \\ &= 1 - \frac{1}{2} \cdot 1 - \frac{1}{2} \cdot 1 \\ &= 0 \end{aligned}$$

# Information Gain and Best Splitting

Determine the best first splitting for the Tennis data

<b>sun</b>	<b>wind</b>	<b>humidity</b>	<b>tennis</b>
$S$	$W$	$\neg H$	$T$
$S$	$\neg W$	$\neg H$	$T$
$\neg S$	$\neg W$	$H$	$\neg T$
$S$	$W$	$H$	$\neg T$

We need to compute  $IG(D_{par}, \mathbf{x}_i)$  for each feature, where  $D_{par}$  is the full training set

# Information Gain and Best Splitting

---

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{right} = \{(\neg H, T), (\neg H, T)\}$

**Q:** What is  $I(D_{par})$ ?

# Information Gain and Best Splitting

---

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{par} = \{(\neg H, T), (\neg H, T)\}$

**Q:** What is  $I(D_{par})$ ?

**A:** Already computed this,  $I(D_{par}) = 1$

# Information Gain and Best Splitting

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{right} = \{(\neg H, T), (\neg H, T)\}$

$$I(D_{par}) = 1,$$

**Q:** What are the weights in the average:  $\frac{|D_{left}|}{|D_{par}|}$  and  $\frac{|D_{right}|}{|D_{par}|}$  ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{right} = \{(\neg H, T), (\neg H, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}$$

**Q:** What is the entropy  $I(D_{left})$ ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{right} = \{(\neg H, T), (\neg H, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}$$

$$p = 1 \Rightarrow I(D_{left}) = 1 \log_2 1 - 0 \log_2 0 = 0$$

# Information Gain and Best Splitting

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{right} = \{(\neg H, T), (\neg H, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}, \quad I(D_{left}) = 1$$

**Q:** What is  $I(D_{right})$ ?

# Information Gain and Best Splitting

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{right} = \{(\neg H, T), (\neg H, T)\}$

$$I(D_{par}) = 1, \quad \frac{|D_{left}|}{|D_{par}|} = \frac{1}{2}, \quad \frac{|D_{right}|}{|D_{par}|} = \frac{1}{2}, \quad I(D_{left}) = 1$$

$$p = 0 \Rightarrow I(D_{right}) = 0 \log_2 0 - 1 \log_2 1 = 0$$

# Information Gain and Best Splitting

$IG(D_{par}, \text{humid})$ :

- $D_{par} = \{(\neg H, T), (\neg H, T), (H, \neg T), (H, \neg T)\}$
- $D_{left} = \{(H, \neg T), (H, \neg T)\}$
- $D_{right} = \{(\neg H, T), (\neg H, T)\}$

**Q:** What is  $IG(D_{par}, \text{humid})$ ?

$$\begin{aligned} IG(D_{par}, \text{humid}) &= I(D_{par}) - \frac{|D_{left}|}{|D_{par}|} I(D_{left}) - \frac{|D_{right}|}{|D_{par}|} I(D_{right}) \\ &= 1 - \frac{1}{2} \cdot 0 - \frac{1}{2} \cdot 0 \\ &= 1 \end{aligned}$$

# Information Gain and Best Splitting

---

So we have ...

$$IG(D_{par}, \text{sun}) = 0.3112$$

$$IG(D_{par}, \text{wind}) = 0$$

$$IG(D_{par}, \text{humid}) = 1$$

**Q:** Which feature should we split on?

# Information Gain and Best Splitting

---

So we have ...

$$IG(D_{par}, \text{sun}) = 0.3112$$

$$IG(D_{par}, \text{wind}) = 0$$

$$IG(D_{par}, \text{humid}) = 1$$

**Q:** Which feature should we split on?

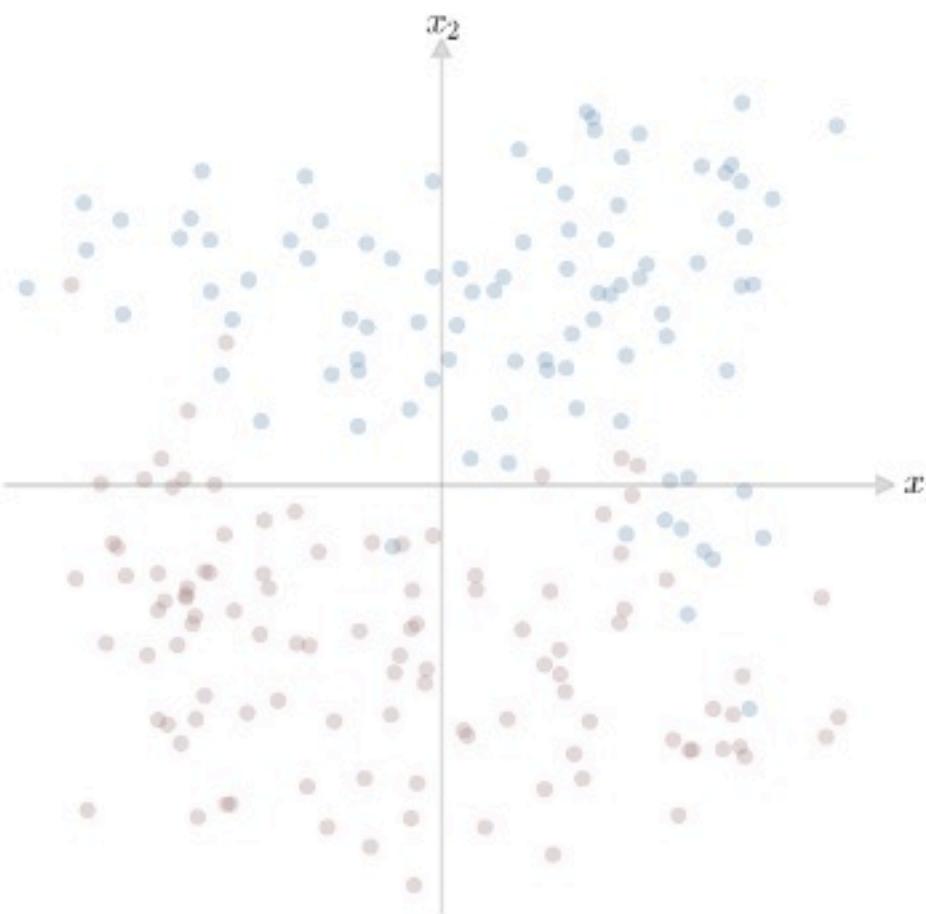
**A:** We split on **humid** because it gives the largest information gain!

# So Many Questions ...

---

- How well do these things work anyway?

In general, Decision Trees are very high variance methods and prone to overfitting



# So Many Questions ...

---

- How well do these things work anyway?

In general, Decision Trees are very high variance methods and prone to overfitting. How do we combat this?

