# Naïve Bayes Classifier

Eamonn Keogh
UCR

This is a high level overview only. For details, see:
Pattern Recognition and Machine Learning, Christopher Bishop, Springer-Verlag, 2006.
Or
Pattern Classification by R. O. Duda, P. E. Hart, D. Stork, Wiley and Sons.

Thomas Bayes
1702 - 1761

We will start off with a visual intuition, before looking at the math…

Grasshoppers

Katydids

Antenna Length

Abdomen Length
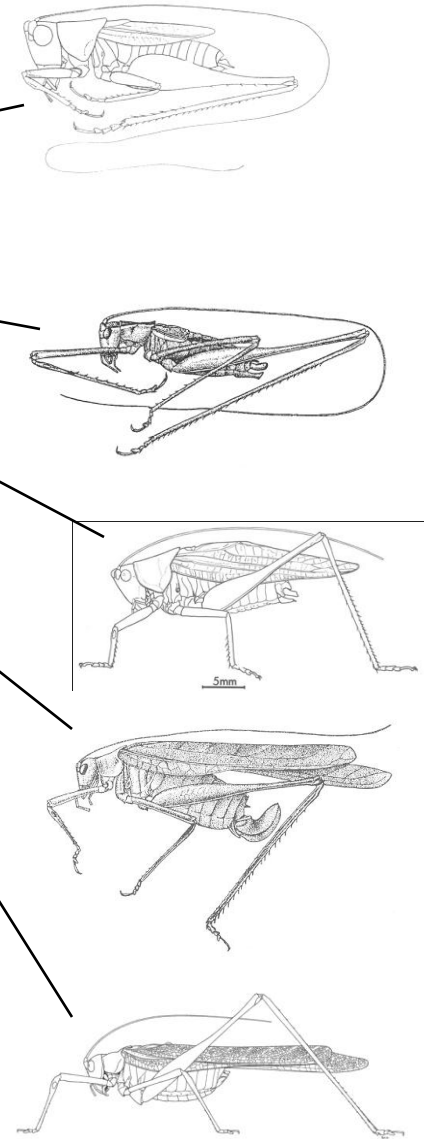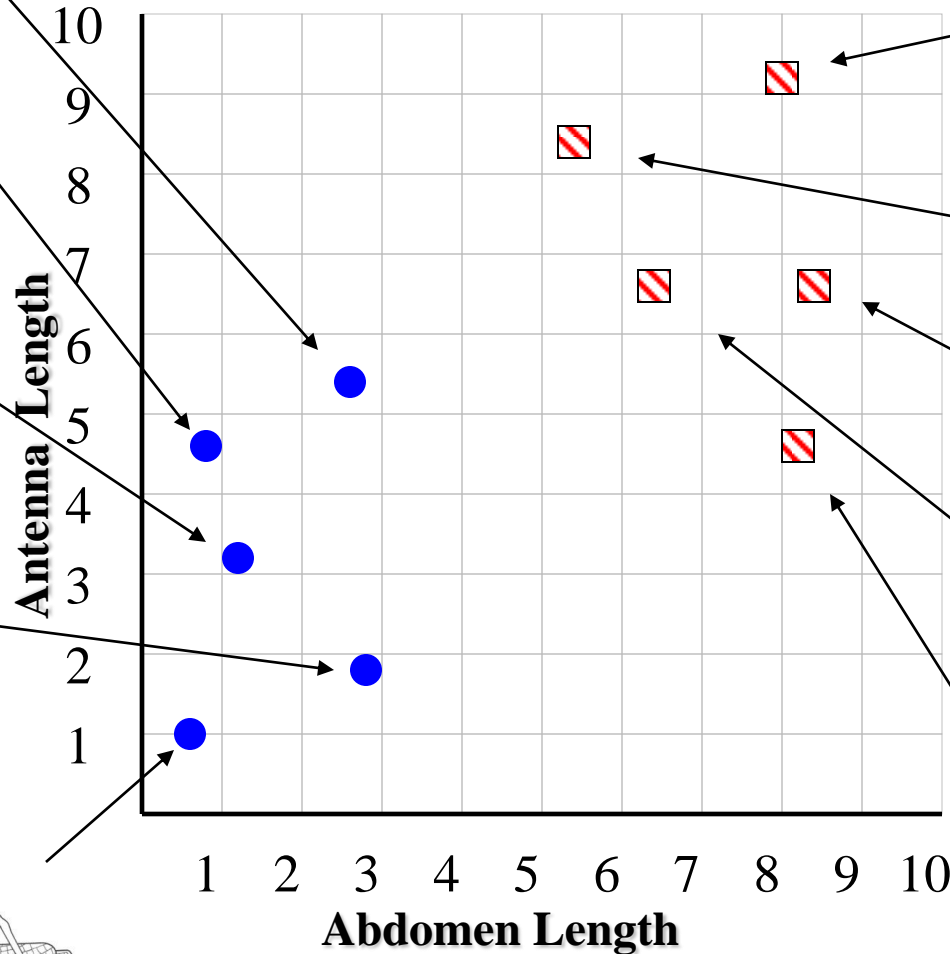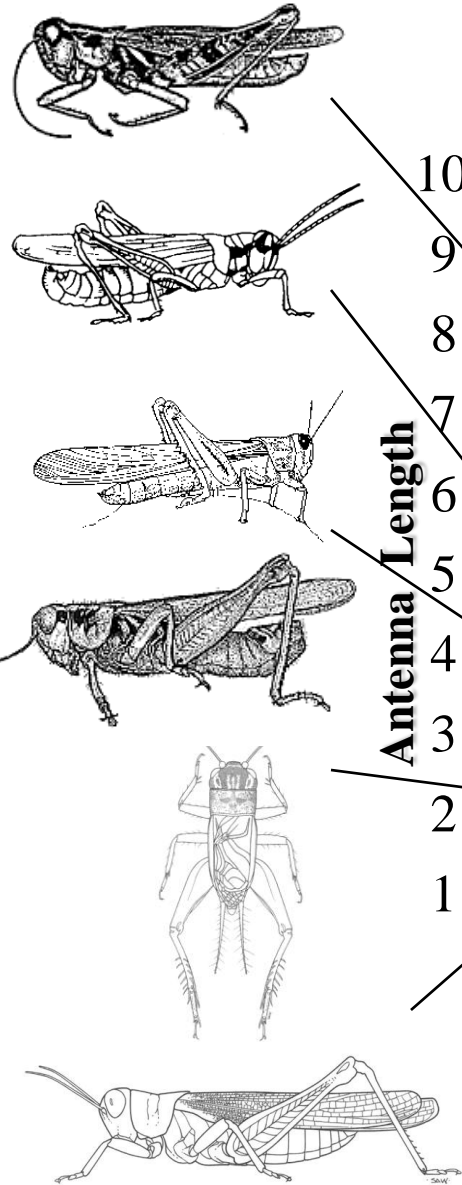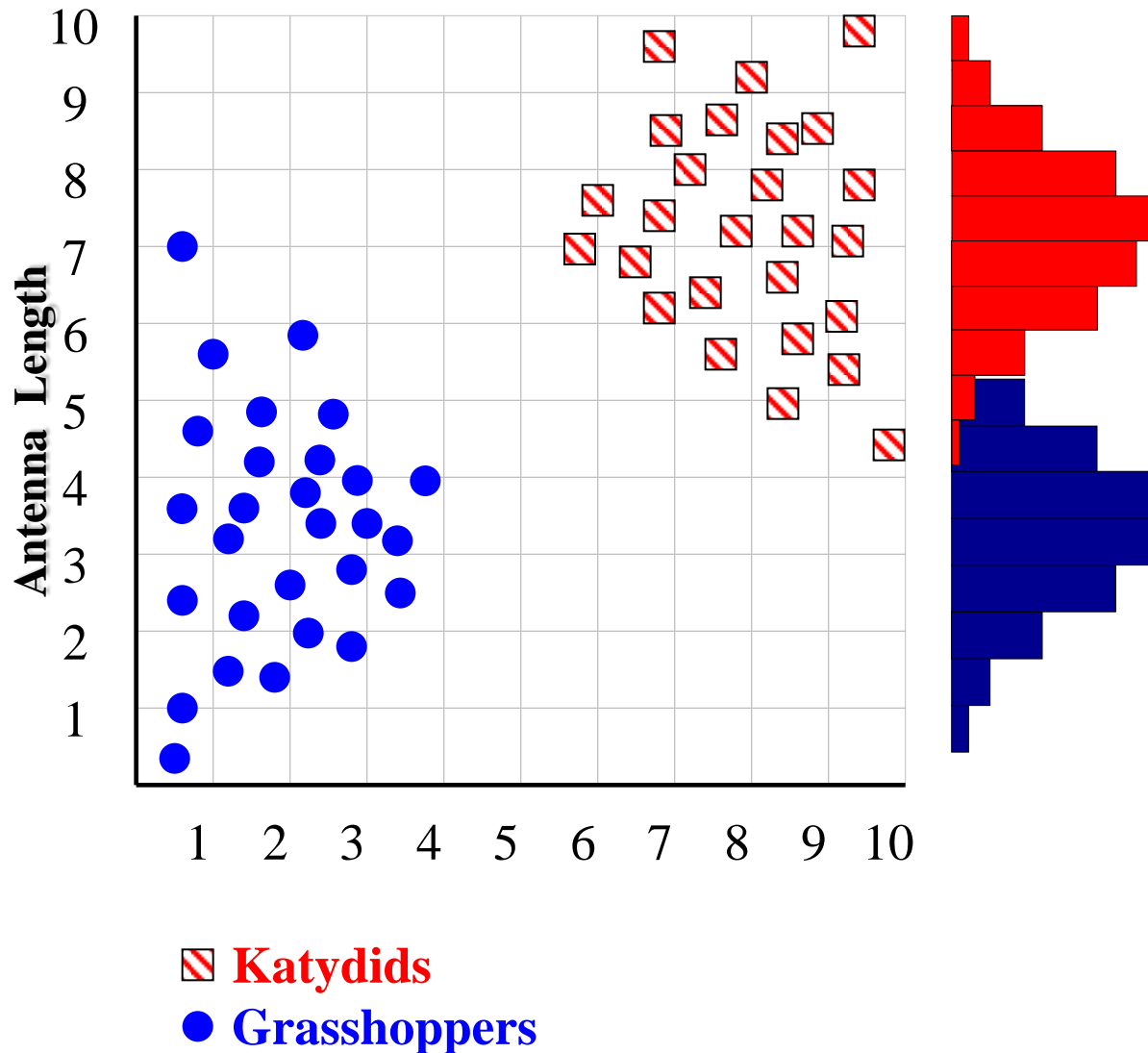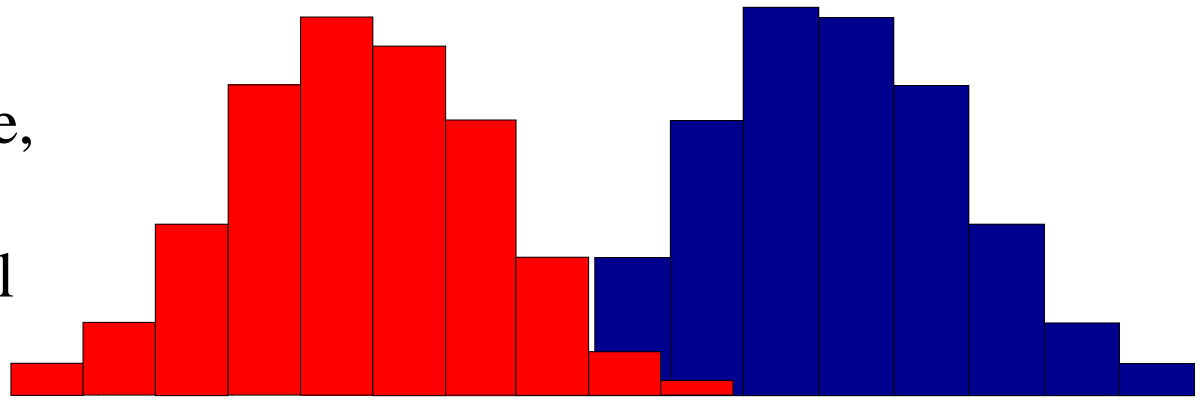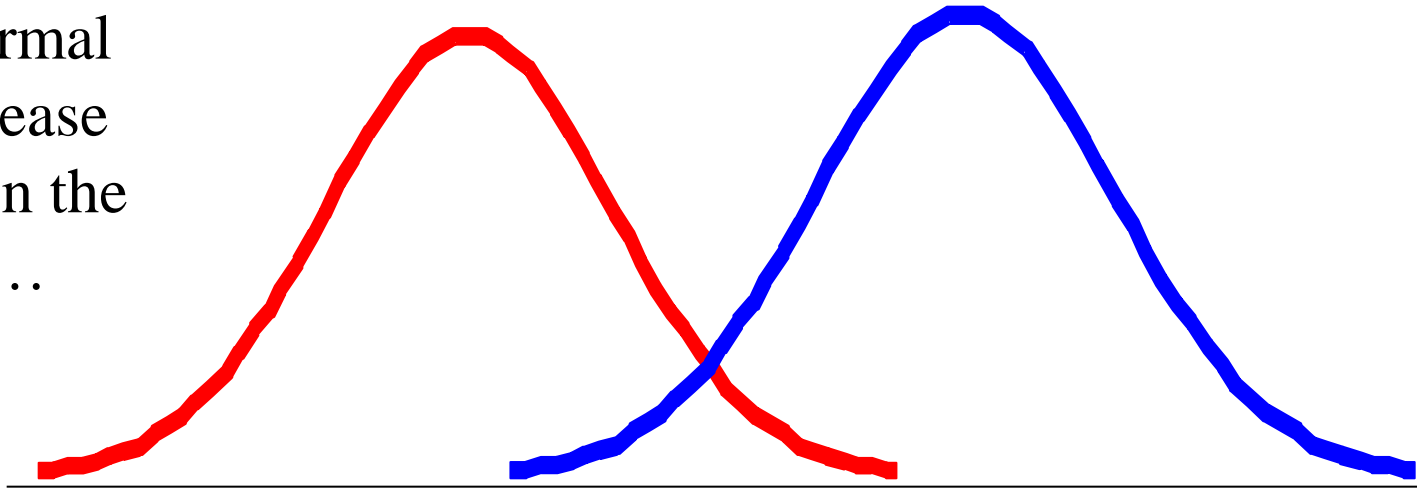
Remember this example?
Let's get lots more data…

With a lot of data, we can build a histogram. Let us just build one for "Antenna Length" for now…

We can leave the histograms as they are, or we can summarize them with two normal distributions.

Let us us two normal distributions for ease of visualization in the following slides…

- We want to classify an insect we have found. Its antennae are 3 units long. How can we classify it?

- We can just ask ourselves, give the distributions of antennae lengths we have seen, is it more *probable* that our insect is a **Grasshopper** or a **Katydid**.
- There is a formal way to discuss the most *probable* classification…

$p(c_j | d)$ = probability of class c_j, *given* that we have observed d

3

Antennae length is **3**
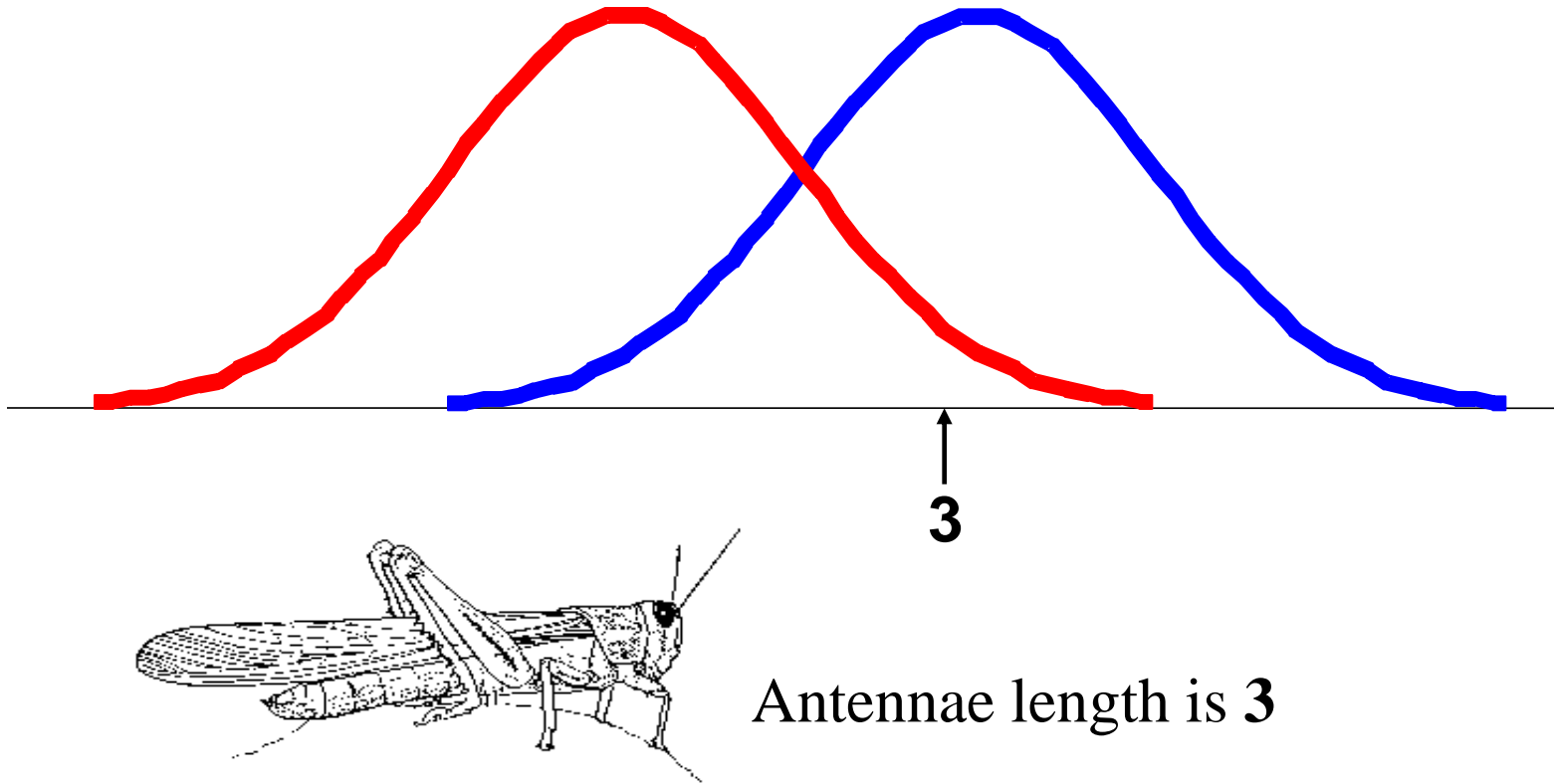
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **3**) = 10 / (10 + 2) = 0.833

P(**Katydid** | **3**) = 2 / (10 + 2) = 0.166
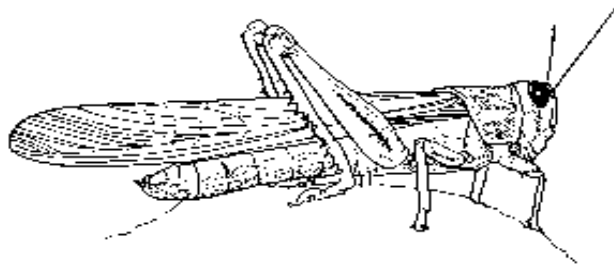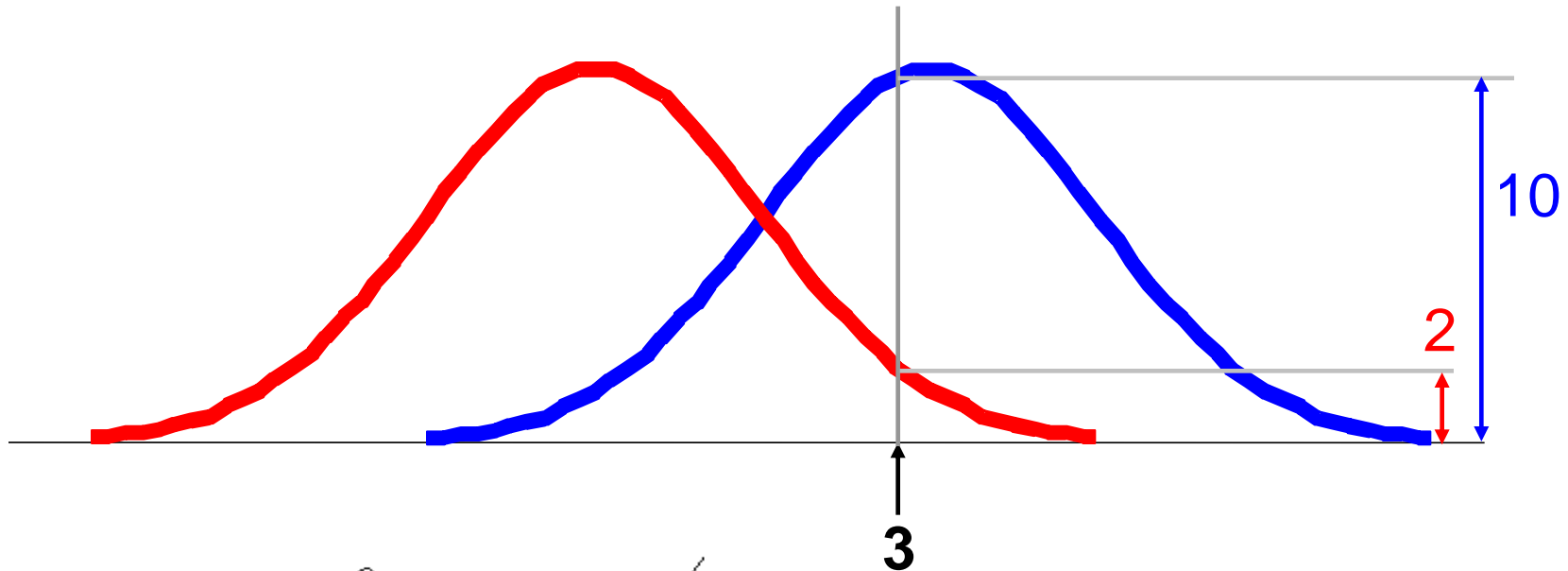
3

Antennae length is **3**

$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **7**) = 3 / (3 + 9)   = 0.250

P(**Katydid** | **7**)   = 9 / (3 + 9)   = 0.750

9

3

7
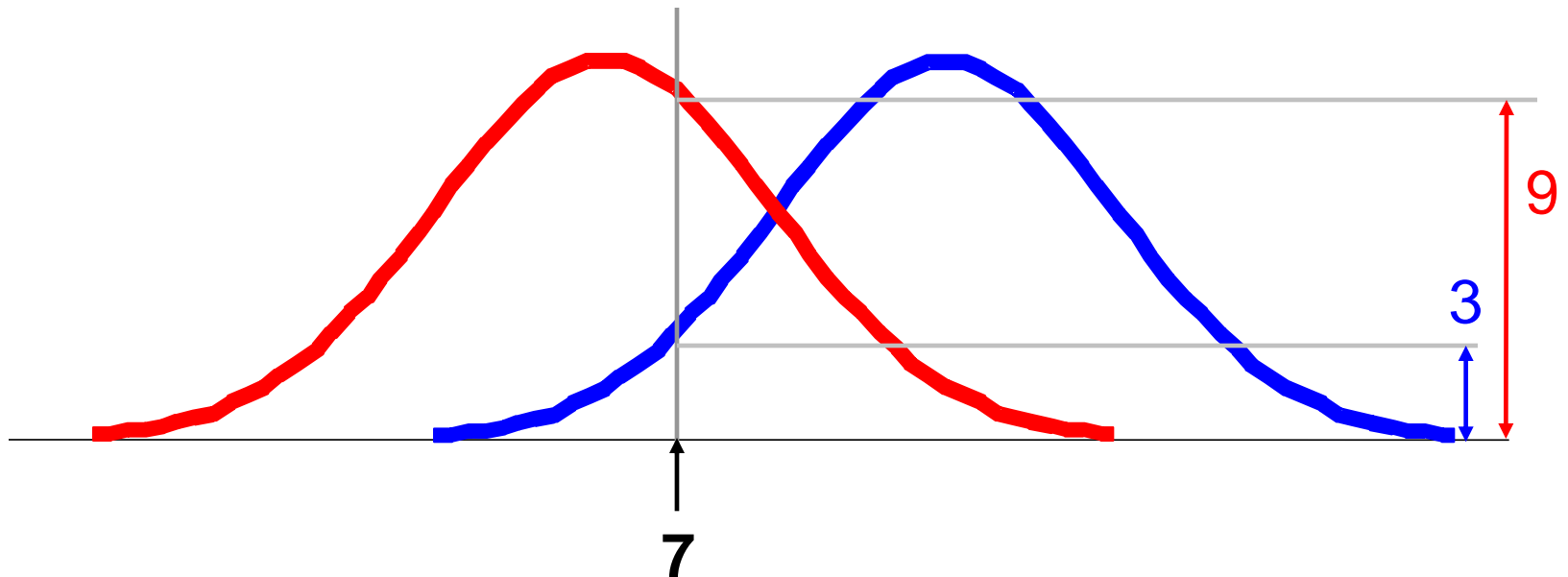
Antennae length is **7**
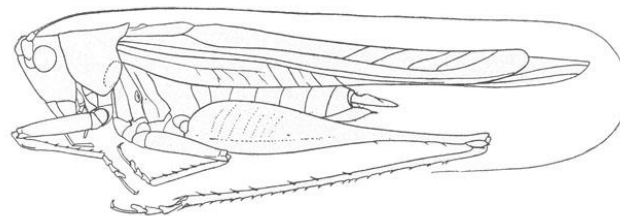
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **5**) = 6 / (6 + 6)   = 0.500

P(**Katydid** | **5**)   = 6 / (6 + 6)   = 0.500

6 6

5

Antennae length is **5**
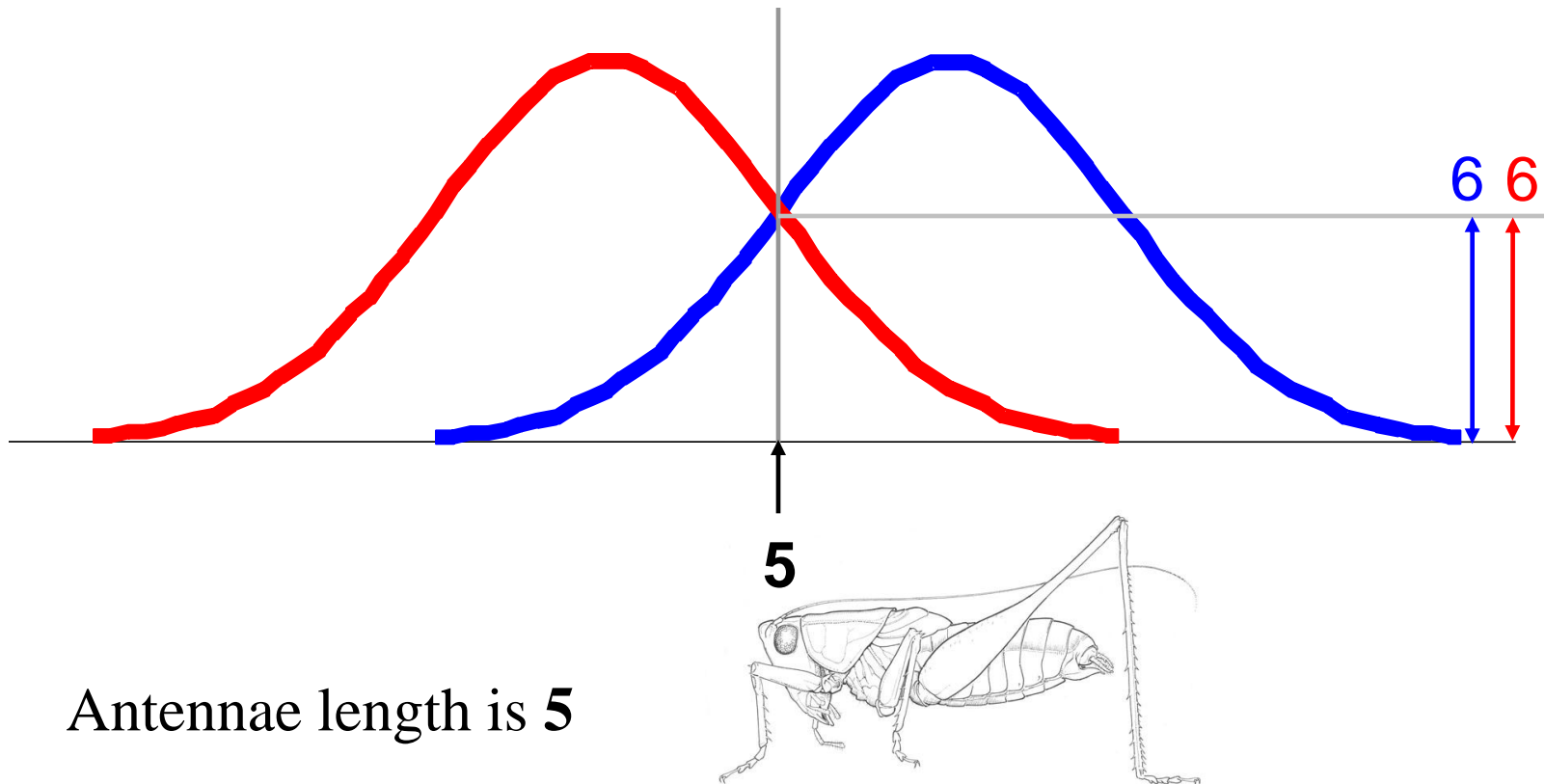
Bayes Classifiers

That was a visual intuition for a simple case of the Bayes classifier, also called:

- Idiot Bayes
- Naïve Bayes
- Simple Bayes

We are about to see some of the mathematical formalisms, and more examples, but keep in mind the basic idea.

Find out the probability of the previously unseen instance belonging to each class, then simply pick the most probable class.

Bayes Classifiers

- Bayesian classifiers use **Bayes theorem**, which says

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

- $p(c_j \mid d)$ = probability of instance d being in class c_j,
  This is what we are trying to compute

- $p(d \mid c_j)$ = probability of generating instance d given class c_j,
  We can imagine that being in class c_j, causes you to have feature d with some probability

- $p(c_j)$ = probability of occurrence of class c_j,
  This is just how frequent the class c_j, is in our database

- $p(d)$ = probability of instance d occurring

  This can actually be ignored, since it is the same for all classes

Assume that we have two classes

c_1 = male, and c_2 = female.

We have a person whose sex we do not know, say "*drew*" or d.

Classifying *drew* as male or female is equivalent to asking is it more probable that *drew* is male or female, I.e which is greater $p(\text{male} \mid drew)$ or $p(\text{female} \mid drew)$

(Note: "Drew can be a male or female name")

Drew Barrymore

Drew Carey

What is the probability of being called "*drew*" given that you are a male?

What is the probability of being a male?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male}) \, p(\text{male})}{p(drew)}$$

What is the probability of being named "*drew*"?

(actually irrelevant, since it is that same for all classes)

This is Officer Drew (who arrested me in 1997). Is Officer Drew a Male or Female?

Luckily, we have a small database with names and sex.

We can use it to apply Bayes rule…

Officer Drew

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Officer Drew

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

$$p(\text{male} \mid drew) = \frac{1/3 \ast 3/8}{3/8} \qquad = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 \ast 5/8}{3/8} \qquad = \frac{0.250}{3/8}$$

Officer Drew is more likely to be a Female.

Officer Drew IS a female!

Officer Drew

$p(\text{male} \mid drew) = \dfrac{1/3 \, * \, 3/8}{3/8} \qquad = \dfrac{0.125}{3/8}$

$p(\text{female} \mid drew) = \dfrac{2/5 \, * \, 5/8}{3/8} \qquad = \dfrac{0.250}{3/8}$

So far we have only considered Bayes Classification when we have one attribute (the "*antennae length*", or the "*name*"). But we may have many features.

How do we use all the features?

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Over 170CM	Eye	Hair length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

- To simplify the task, **naïve Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * ....* p(d_n|c_j)$$

The probability of class c_j generating instance d, equals….

The probability of class c_j generating the observed value for feature 1, multiplied by..

The probability of class c_j generating the observed value for feature 2, multiplied by..

- To simplify the task, **naïve Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * ….* p(d_n|c_j)$$

$$p(\text{officer drew}|c_j) = p(\text{over\_170}_{cm} = \text{yes}|c_j) * p(\text{eye} = blue|c_j) * ….$$
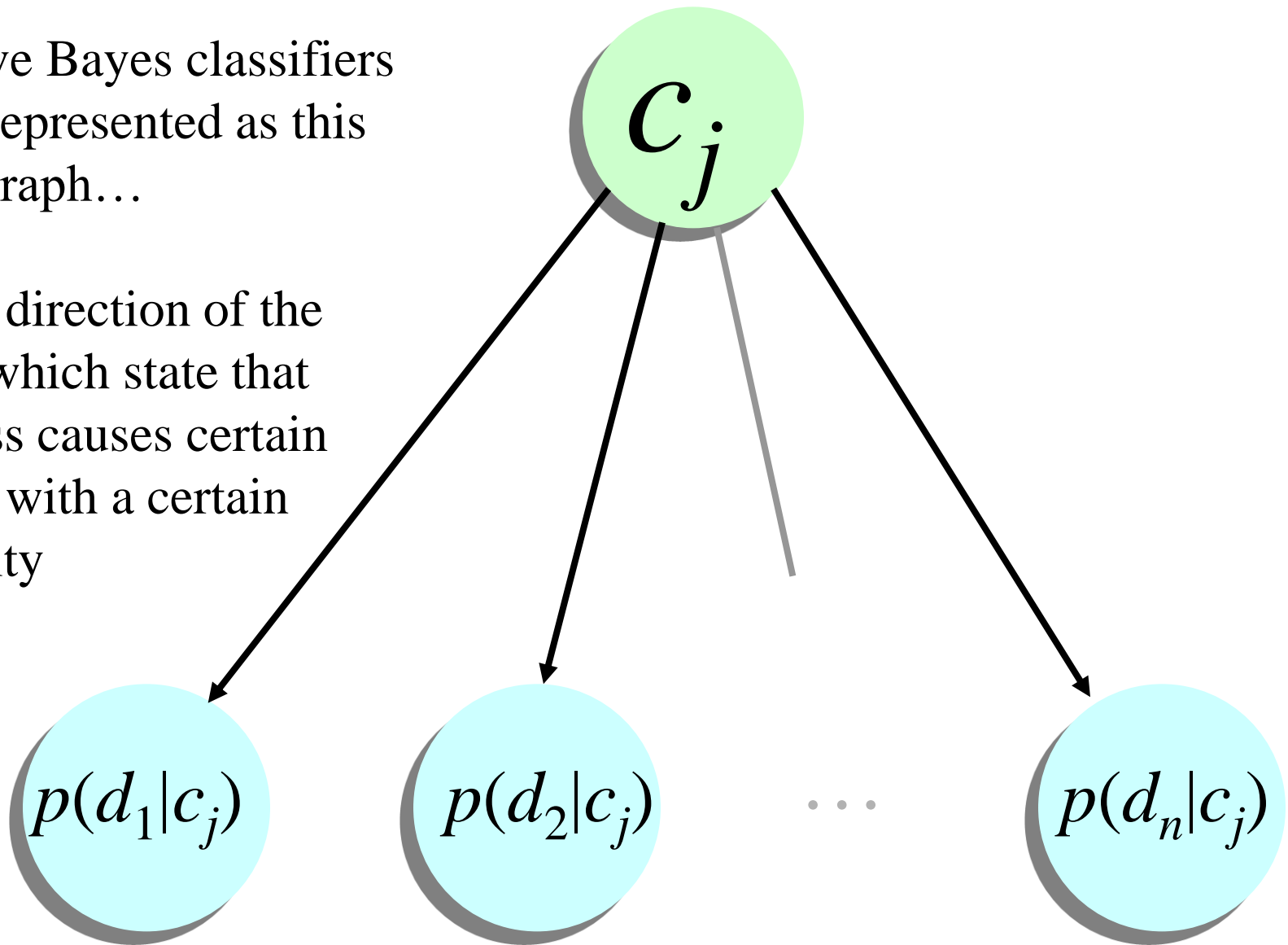

Officer Drew is blue-eyed, over 170$_{cm}$ tall, and has long hair

$$p(\text{officer drew}| \text{Female}) = 2/5 * 3/5 * ….$$
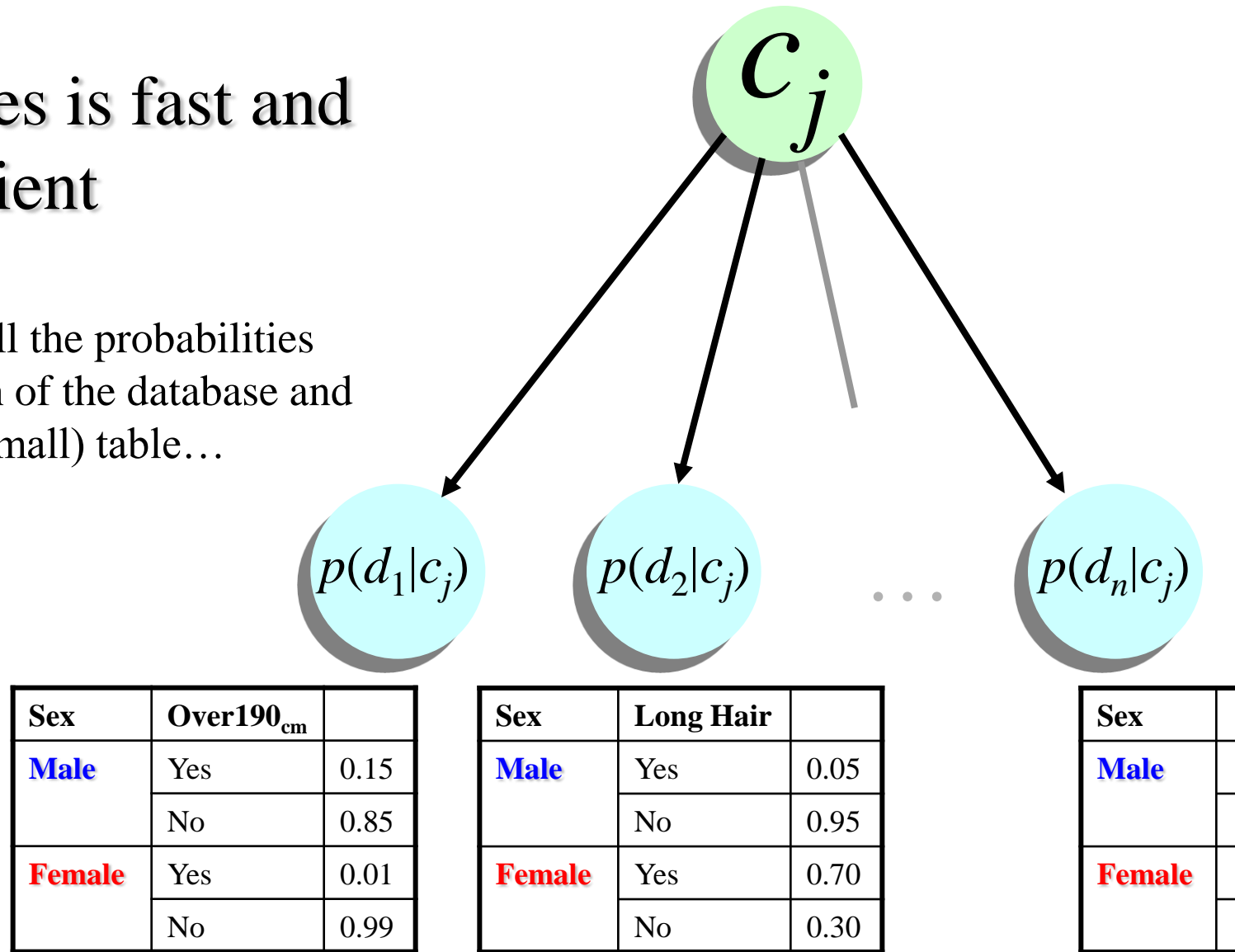$$p(\text{officer drew}| \text{Male}) = 2/3 * 2/3 * ….$$

The Naive Bayes classifiers is often represented as this type of graph…

Note the direction of the arrows, which state that each class causes certain features, with a certain probability

$$c_j$$

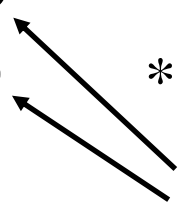$$p(d_1|c_j) \qquad p(d_2|c_j) \qquad \cdots \qquad p(d_n|c_j)$$

Naïve Bayes is fast and space efficient

We can look up all the probabilities with a single scan of the database and store them in a (small) table…

c_j

$p(d_1|c_j)$ $p(d_2|c_j)$ . . . $p(d_n|c_j)$

Sex	Over190$_{cm}$	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

Sex	Long Hair	
Male	Yes	0.05
	No	0.95
Female	Yes	0.70
	No	0.30

Sex	
Male	
Female	

Naïve Bayes is NOT sensitive to irrelevant features...

Suppose we are trying to classify a persons sex based on several features, including eye color. (Of course, eye color is completely irrelevant to a persons gender)

$$p(\text{Jessica} \mid c_j) = p(\text{eye} = \text{brown} \mid c_j) * p(\text{wears\_dress} = \text{yes} \mid c_j) * \ldots.$$

$p(\text{Jessica} \mid \text{Female}) = \;\; 9{,}000/10{,}000 \qquad * \;\; 9{,}975/10{,}000 \;\; * \ldots.$

$p(\text{Jessica} \mid \text{Male}) \;\; = \;\; 9{,}001/10{,}000 \qquad * \;\; 2/10{,}000 \qquad * \ldots.$

Almost the same!

However, this assumes that we have good enough estimates of the probabilities, so the more data the better.

An obvious point. I have used a simple two class problem, and two possible values for each example, for my previous examples. However we can have an arbitrary number of classes, or feature values

c_j

$p(d_1|c_j)$ $p(d_2|c_j)$ . . . $p(d_n|c_j)$
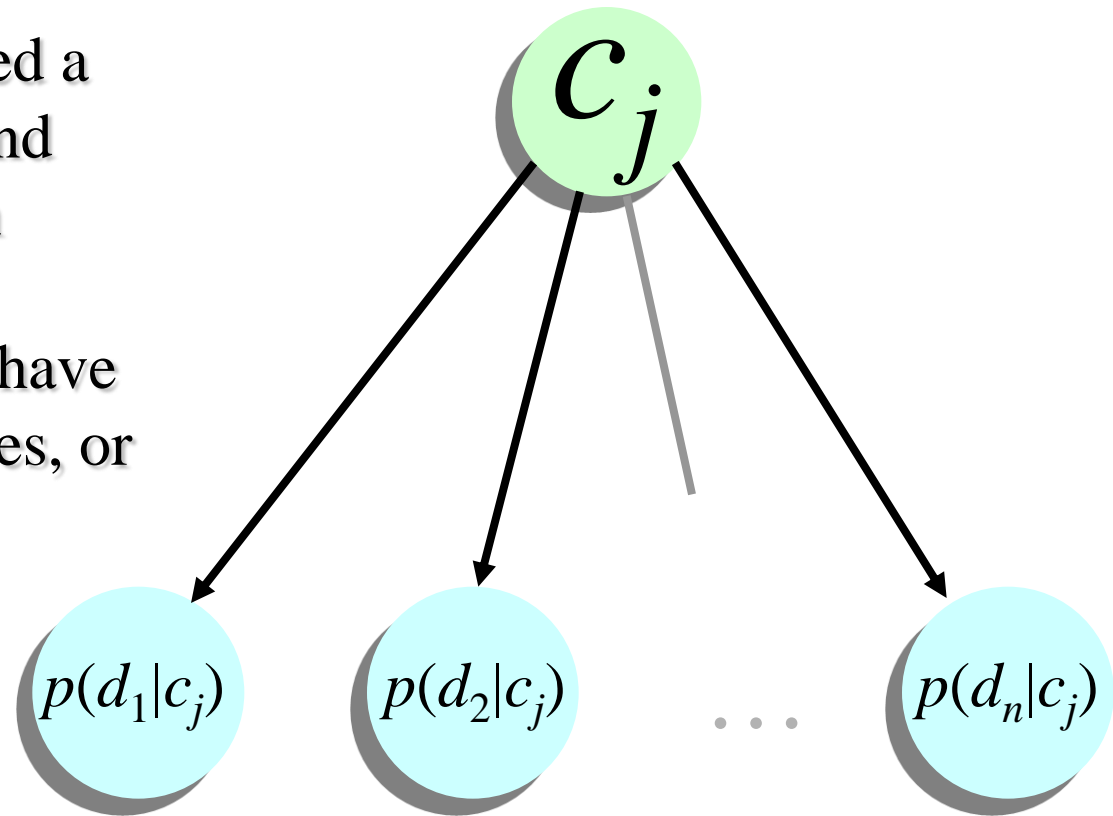
Animal	Mass >10$_{kg}$	
Cat	Yes	0.15
	No	0.85
Dog	Yes	0.91
	No	0.09
Pig	Yes	0.99
	No	0.01

Animal	Color	
Cat	Black	0.33
	White	0.23
	Brown	0.44
Dog	Black	0.97
	White	0.03
	Brown	0.90
Pig	Black	0.04
	White	0.01

Animal
Cat
Dog
Pig

Problem!

Naïve Bayes assumes independence of features…

$p(d|c_j)$

Naïve Bayesian Classifier

$p(d_1|c_j)$

$p(d_2|c_j)$

$p(d_n|c_j)$
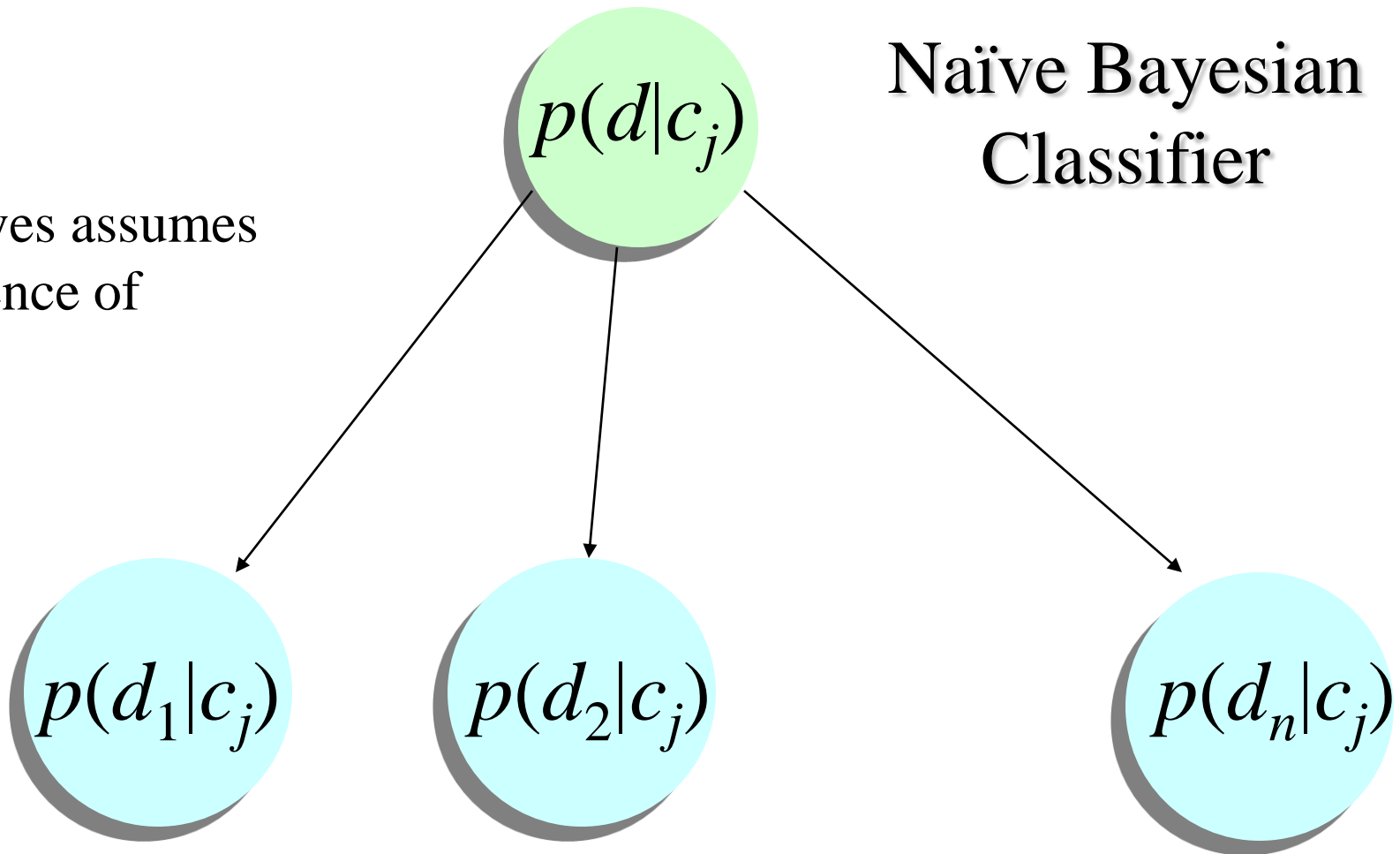
Sex	Over 6 foot	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

Sex	Over 200 pounds	
Male	Yes	0.11
	No	0.80
Female	Yes	0.05
	No	0.95

Solution

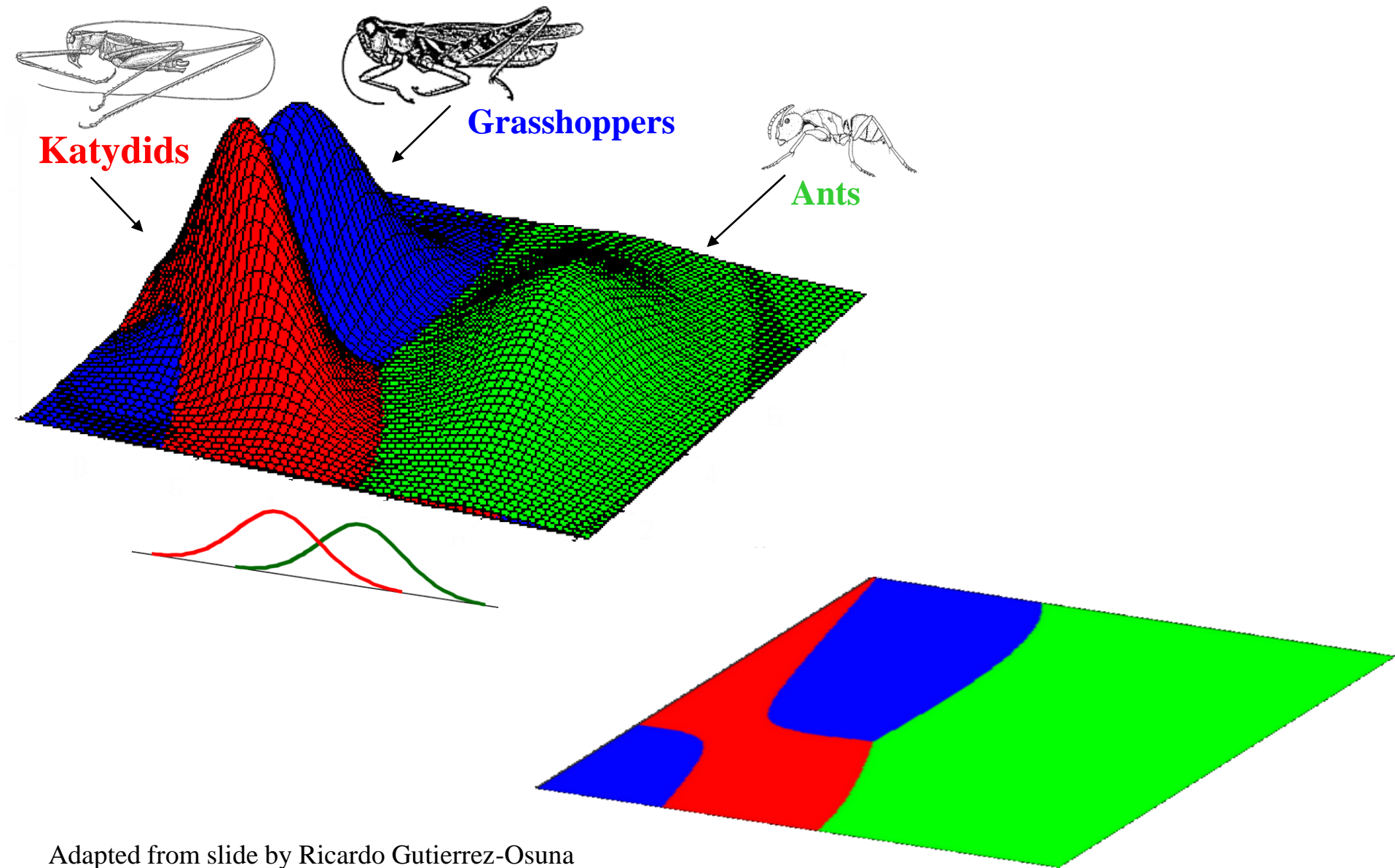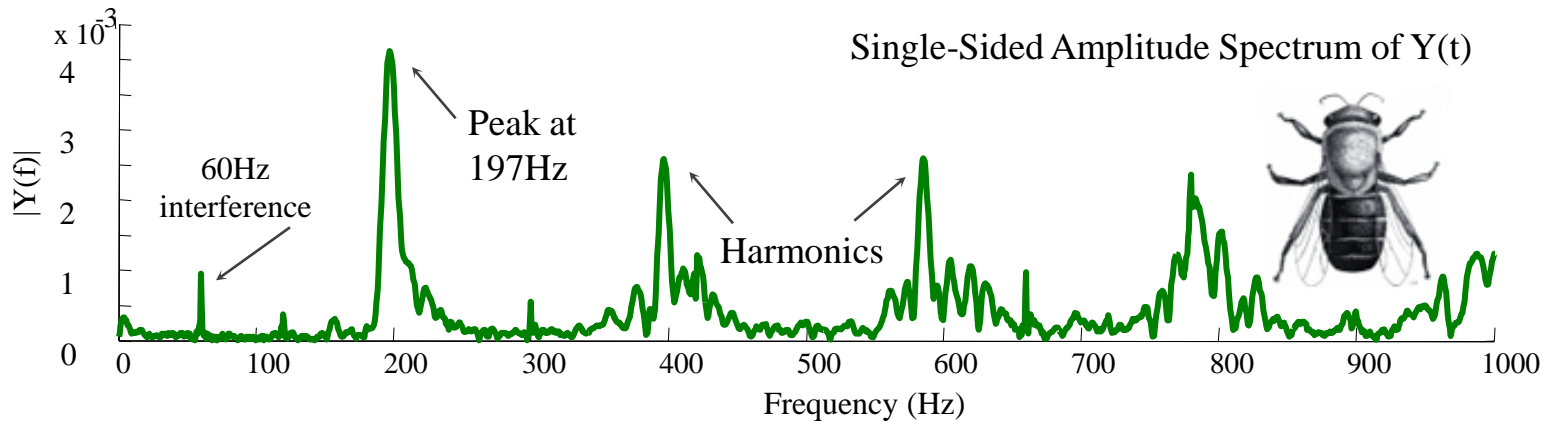Consider the relationships between attributes…

Naïve Bayesian Classifier

$$p(d|c_j)$$

$$p(d_1|c_j) \qquad p(d_2|c_j) \qquad p(d_n|c_j)$$
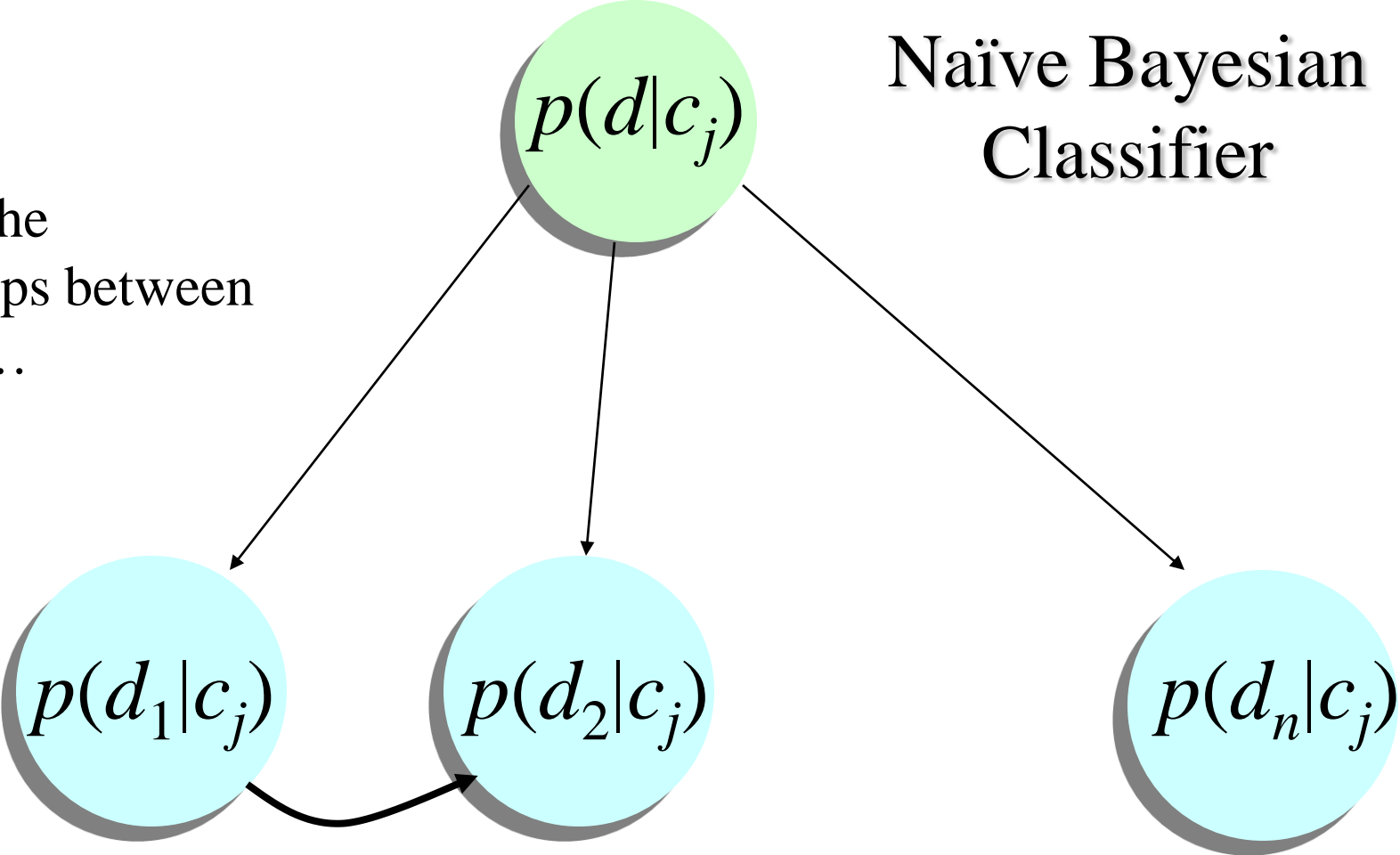
Sex	Over 6 foot	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

Sex	Over 200 pounds	
Male	Yes and **Over 6 foot**	0.11
	No and **Over 6 foot**	0.59
	Yes and NOT **Over 6 foot**	0.05
	No and NOT **Over 6 foot**	0.35
Female	Yes and **Over 6 foot**	0.01

Solution

Consider the relationships between attributes…

$p(d|c_j)$

Naïve Bayesian Classifier

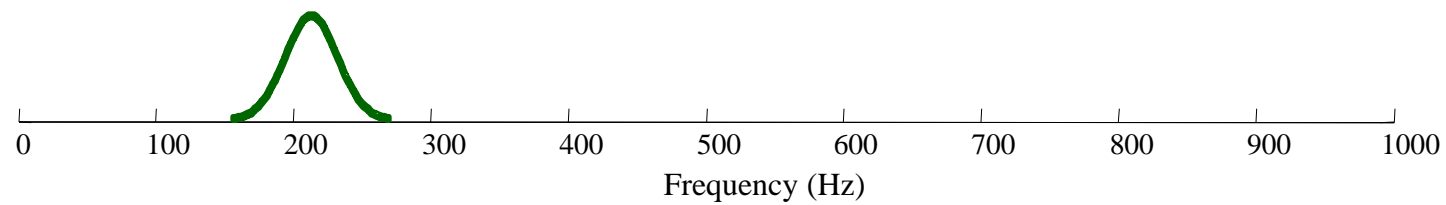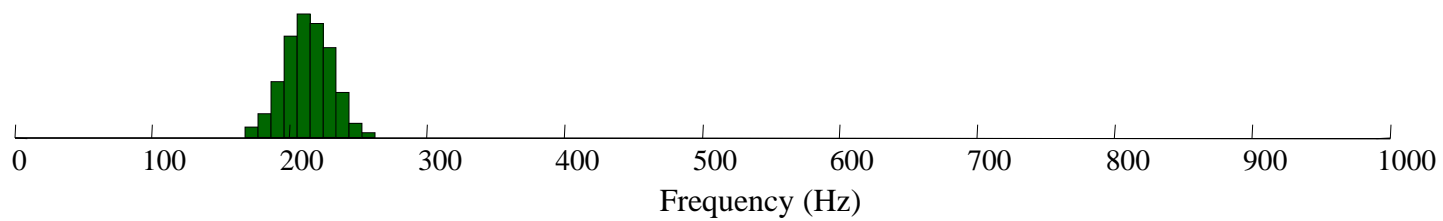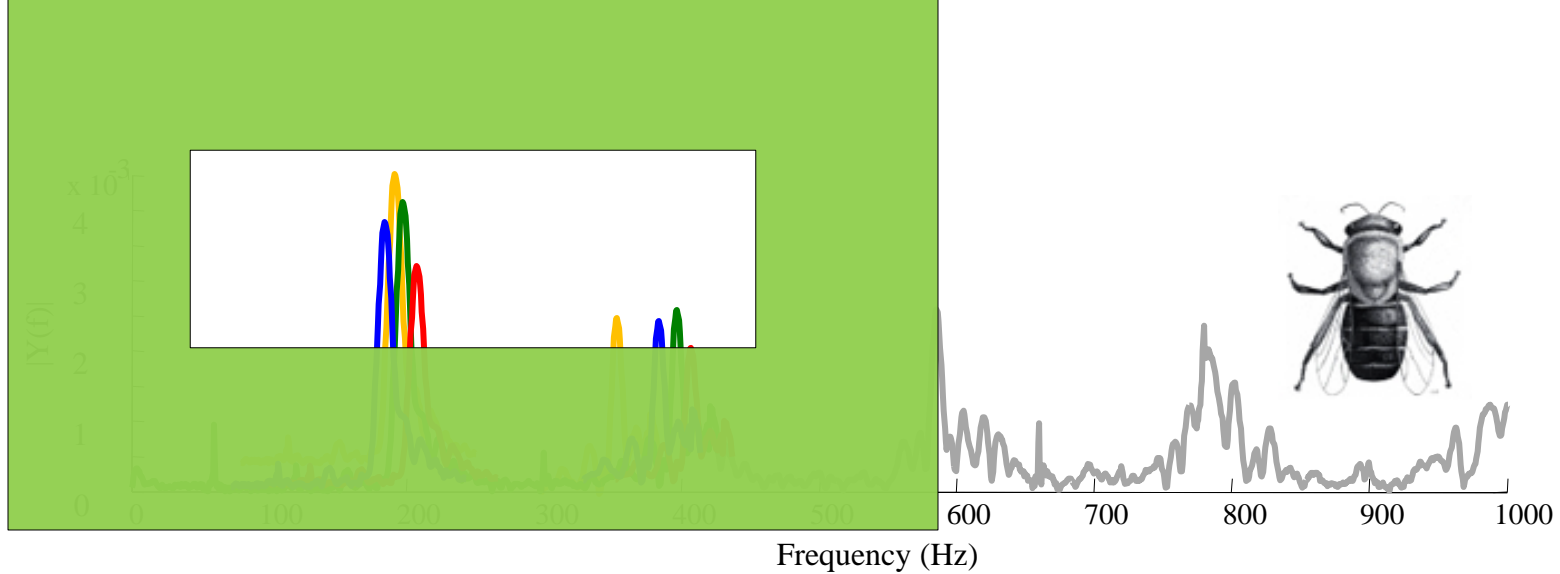$p(d_1|c_j)$ $p(d_2|c_j)$ $p(d_n|c_j)$
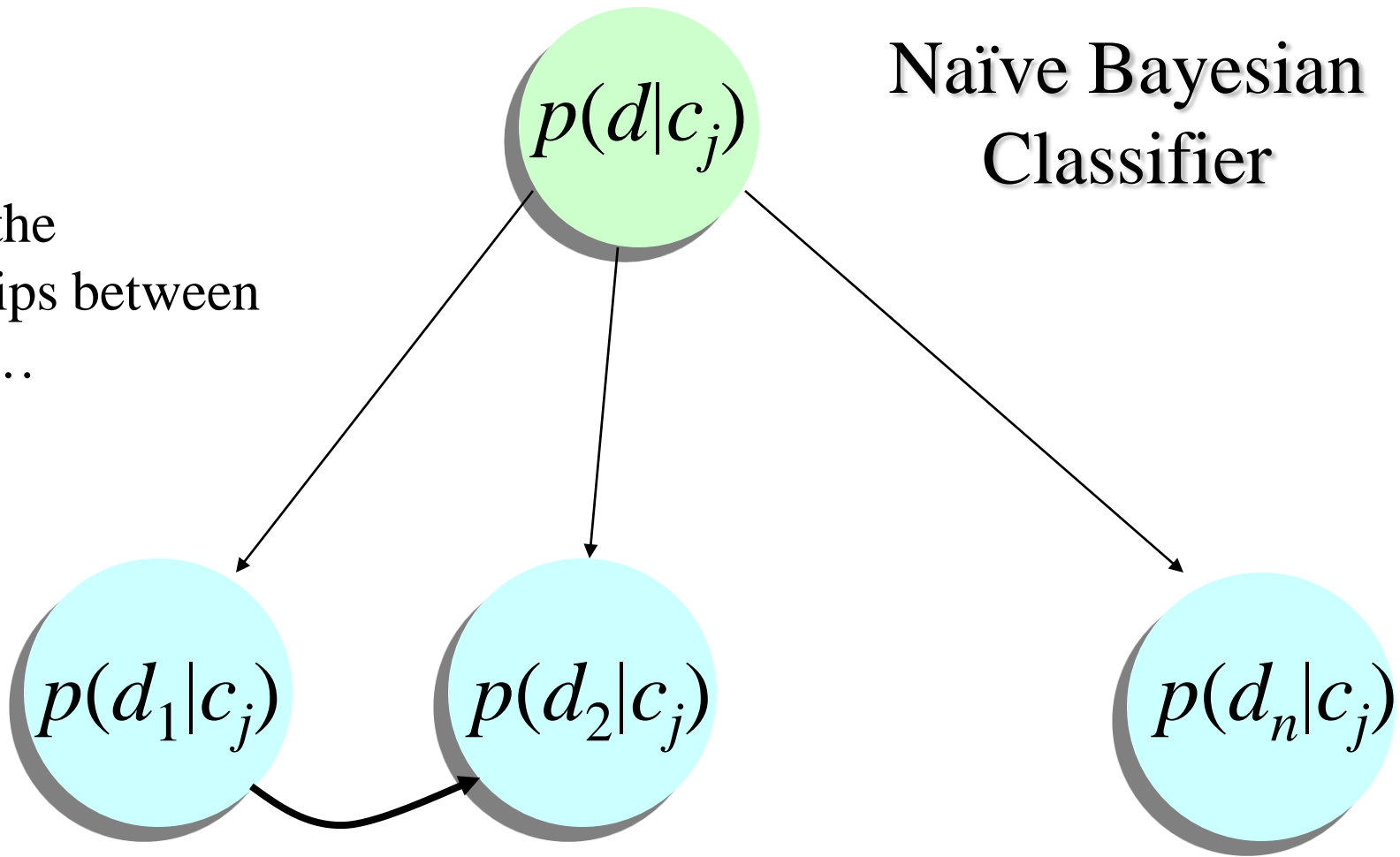
But how do we find the set of connecting arcs??

The Naïve Bayesian Classifier has a piecewise quadratic decision boundary

Katydids

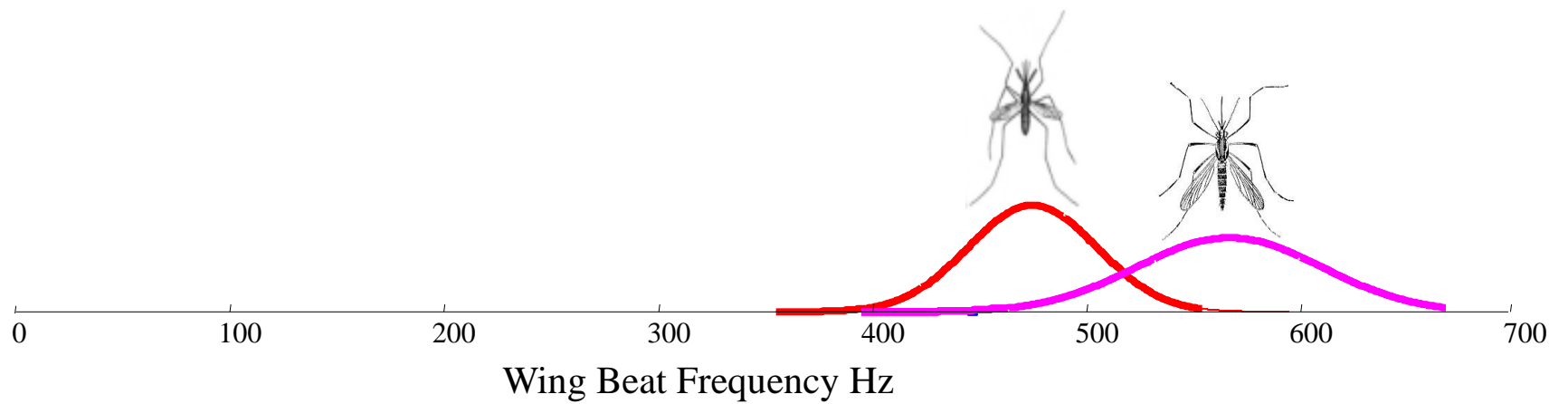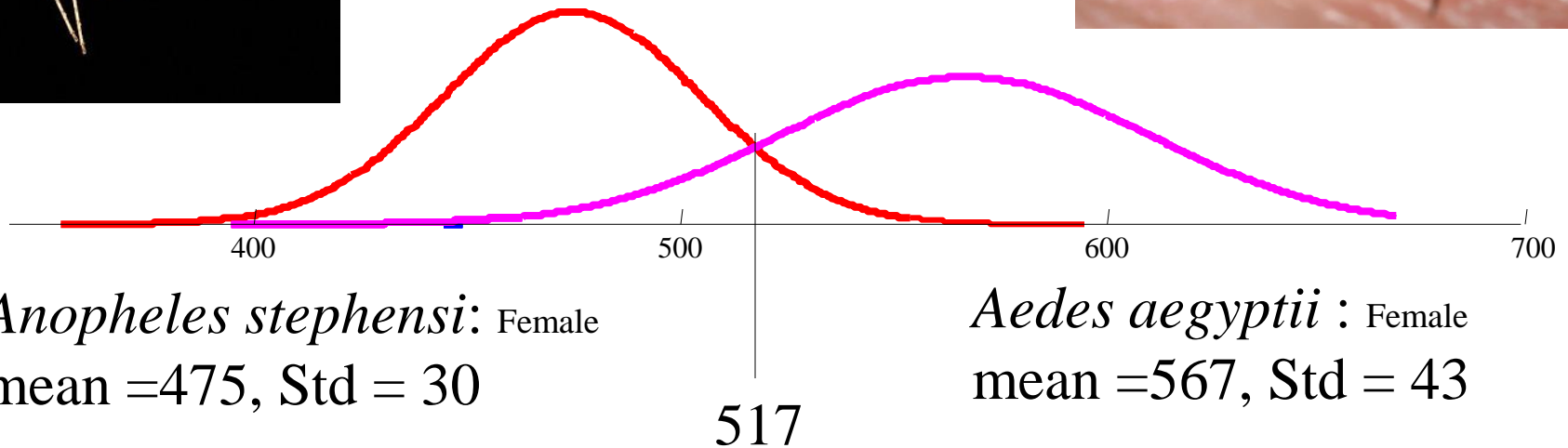**Grasshoppers**

Ants

One second of audio from the laser sensor. Only
Bombus impatiens (Common Eastern Bumble
Bee) is in the insectary.

0.2

0.1

0

-0.1

-0.2

x 10⁴

0   0.5   1   1.5   2   2.5   3   3.5   4   4.5

Background noise

Bee begins to cross laser

Single-Sided Amplitude Spectrum of Y(t)

x 10³

4

3

2

1

0

|Y(f)|

60Hz
interference

Peak at
197Hz

Harmonics

0   100   200   300   400   500   600   700   800   900   1000

Frequency (Hz)

$|Y(f)|$

$\times 10^3$

4

3

2

1

0

0   100   200   300   400   500   600   700   800   900   1000

Frequency (Hz)

0   100   200   300   400   500   600   700   800   900   1000

Frequency (Hz)

0   100   200   300   400   500   600   700   800   900   1000

Frequency (Hz)

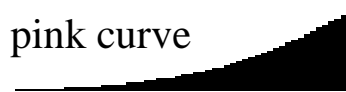$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Wing Beat Frequency Hz

Wing Beat Frequency Hz

Anopheles stephensi: Female
mean =475, Std = 30

Aedes aegyptii : Female
mean =567, Std = 43

517

If I see an insect with a wingbeat frequency of 500, what is it?

$$P(Anopheles|wingbeat = 500) = \frac{1}{\sqrt{2\pi}\ 30}e^{-\frac{(500-475)^2}{2\times 30^2}}$$

517
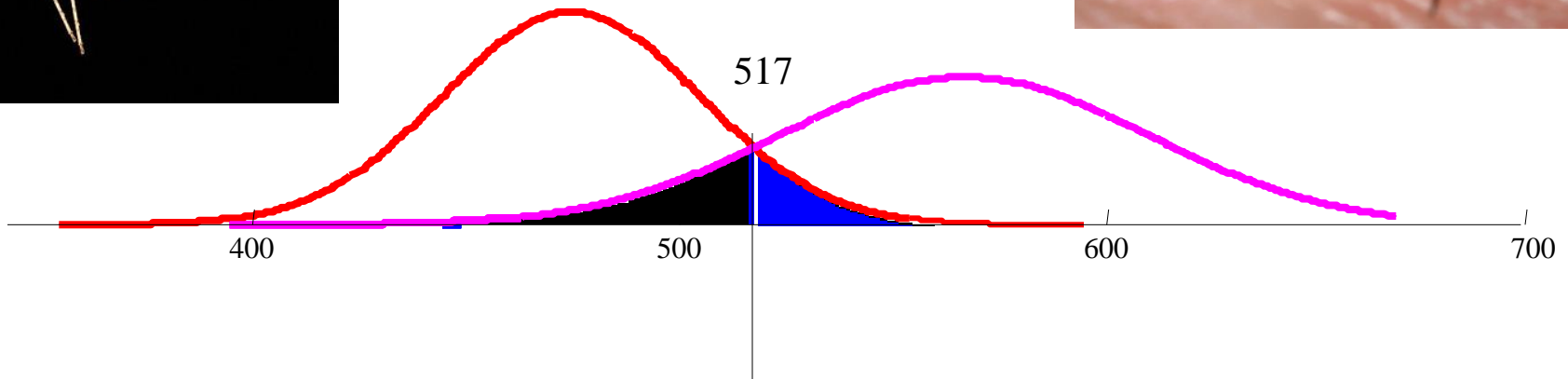
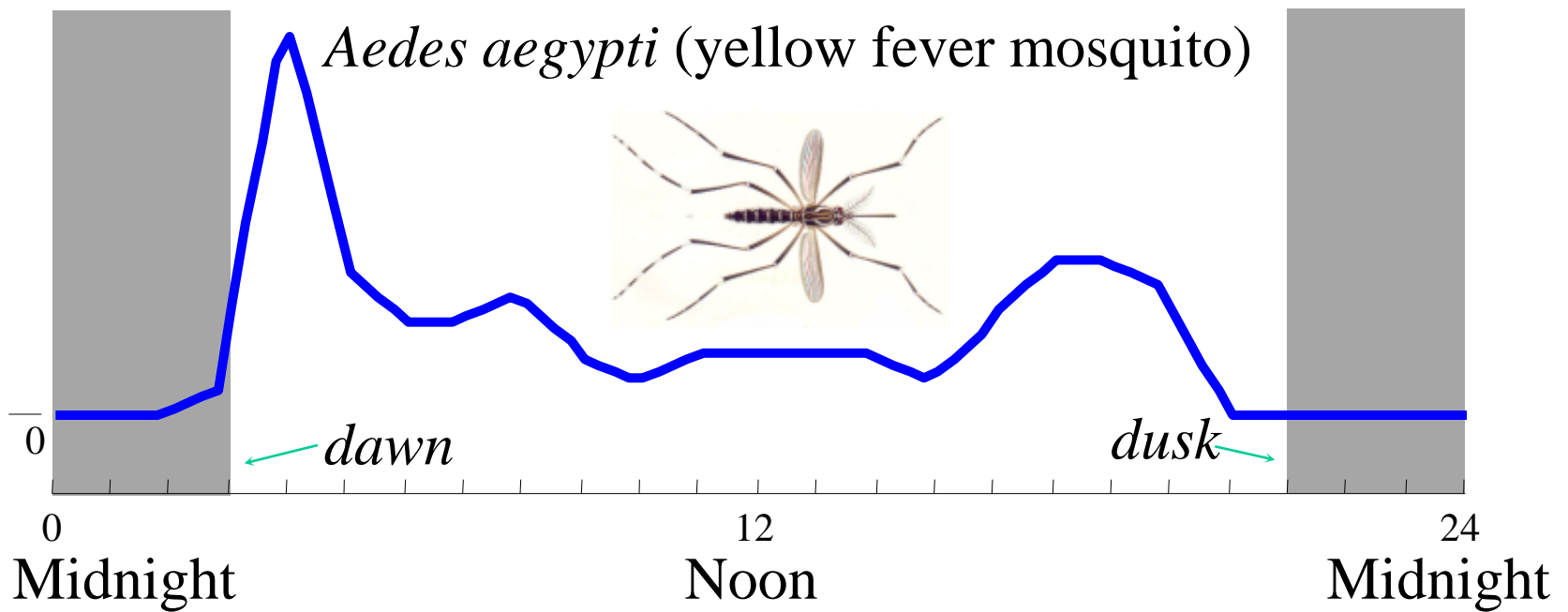What is the error rate?

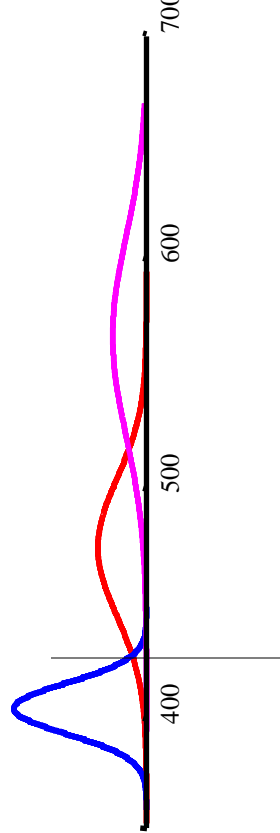12.2% of the area under the pink curve

8.02% of the area under the red curve
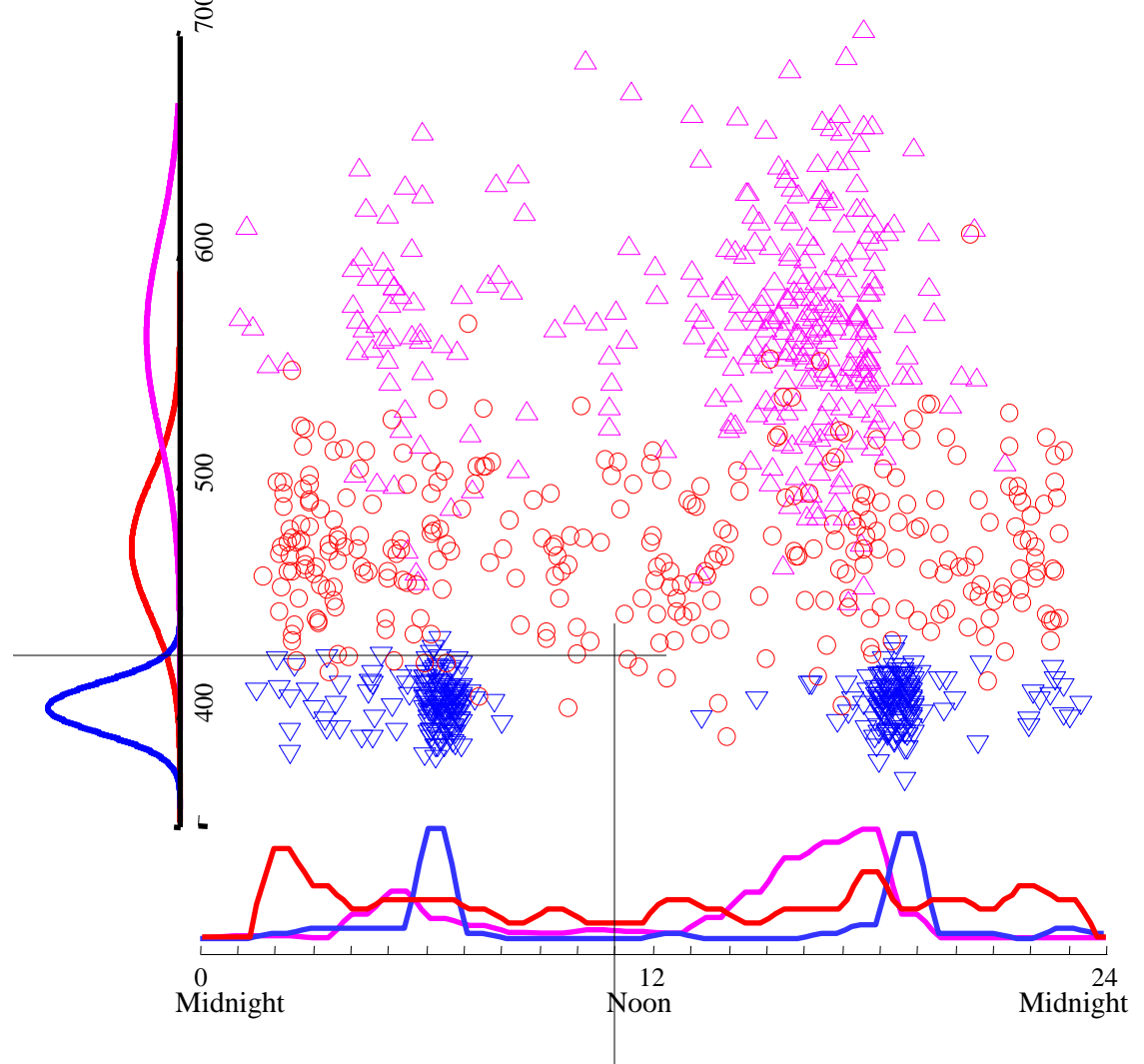
Can we get more features?

Circadian Features

Aedes aegypti (yellow fever mosquito)
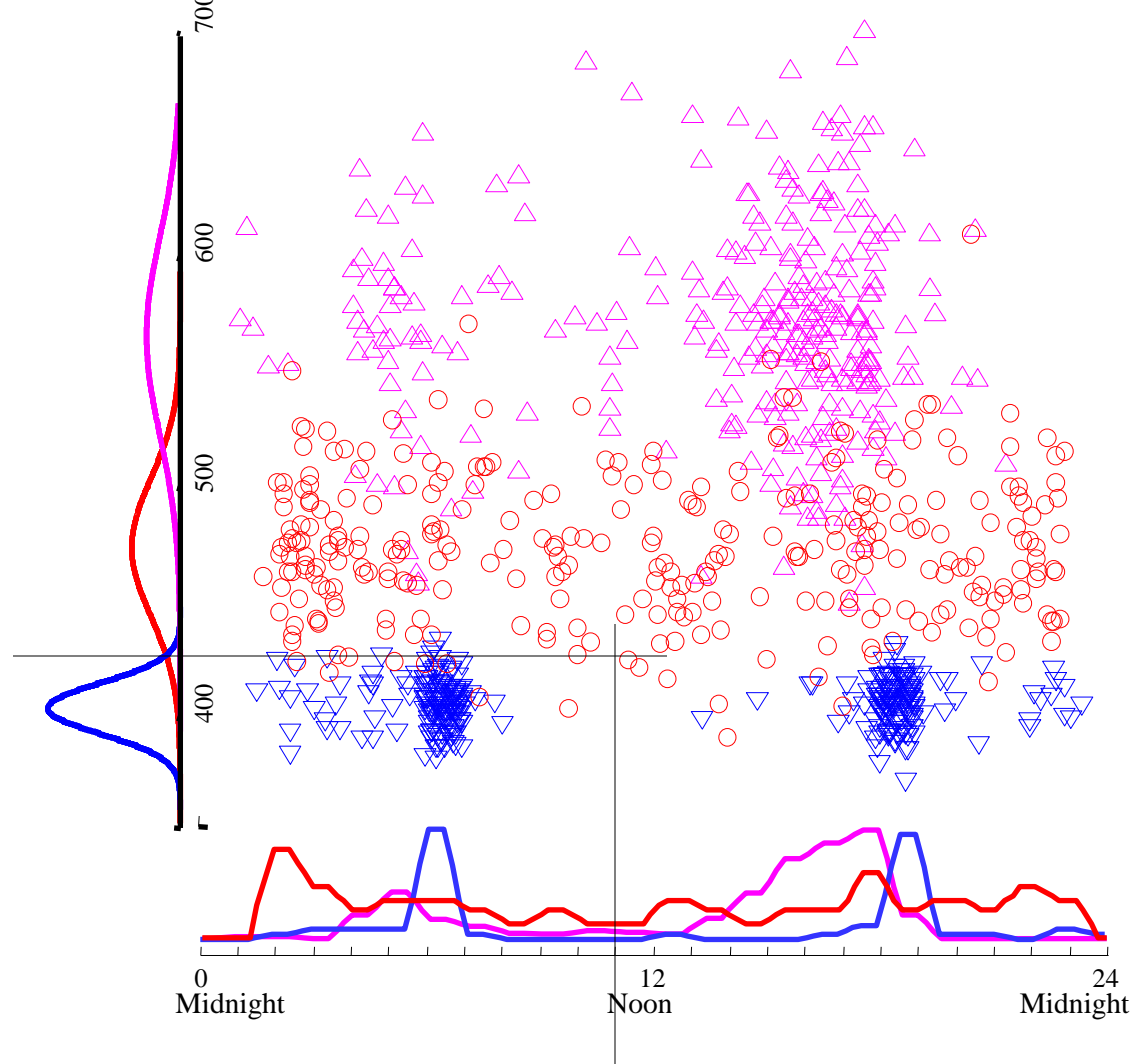
0

dawn

dusk

0
Midnight

12
Noon

24
Midnight

Suppose I observe an
insect with a wingbeat
frequency of 420Hz

What is it?

Suppose I observe an insect with a wingbeat frequency of 420Hz **at 11:00am**

What is it?

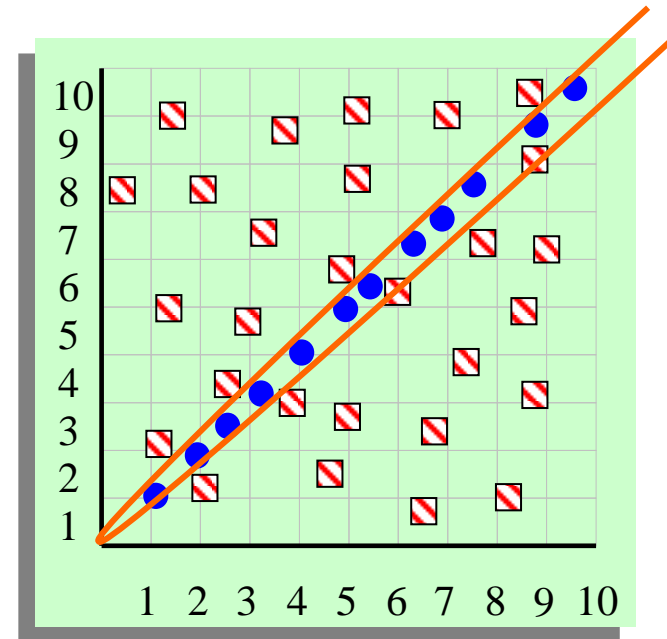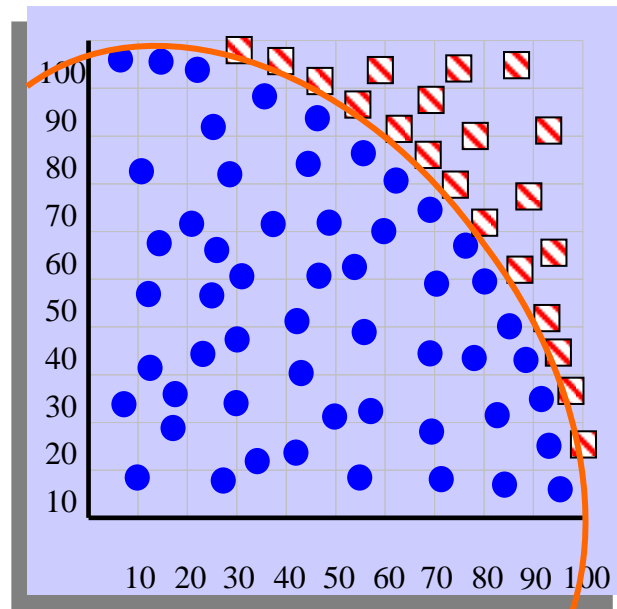Suppose I observe an insect with a wingbeat frequency of 420 at 11:00am
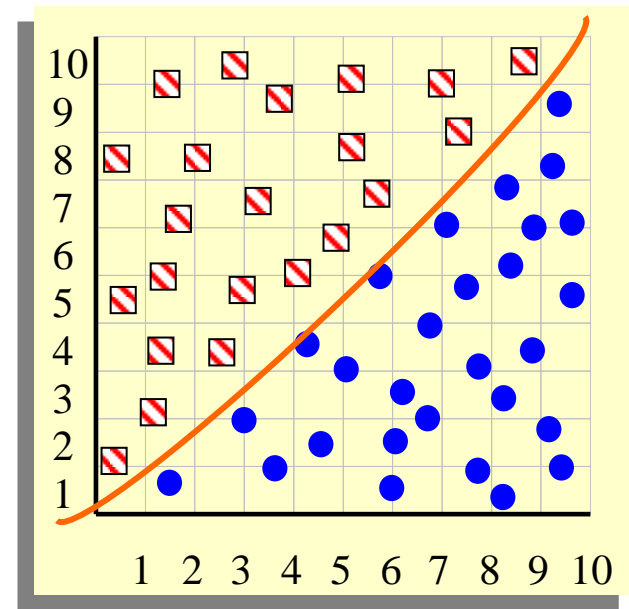
What is it?

(Culex | [420Hz,11:00am]) = (6/ (6 + 6 + 0)) * (2/ (2 + 4 + 3)) = 0.111

(Anopheles | [420Hz,11:00am]) = (6/ (6 + 6 + 0)) * (4/ (2 + 4 + 3)) = 0.222

(Aedes | [420Hz,11:00am]) = (0/ (6 + 6 + 0)) * (3/ (2 + 4 + 3)) = 0.000

Which of the "Pigeon Problems" can be solved by a decision tree?

Advantages/Disadvantages of Naïve Bayes

- Advantages:
  - Fast to train (single scan). Fast to classify
  - Not sensitive to irrelevant features
  - Handles real and discrete data
  - Handles streaming data well
- Disadvantages:
  - Assumes independence of features