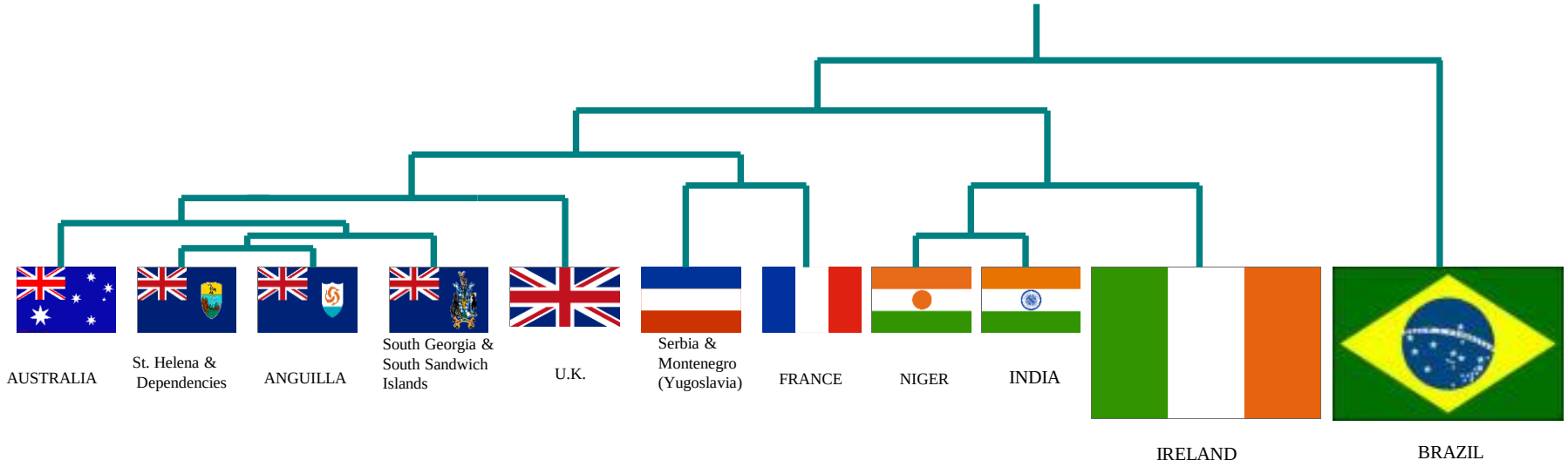
A Gentle Introduction to Machine Learning



Dr. Uzma Jamil

Department of Computer Science

Government College University, Faisalabad.

Data Mining Definition

- Finding hidden information in a database
- Data Mining has been defined as

"*The nontrivial extraction of implicit, previously unknown, and potentially useful information from data**".

- Similar terms
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning
  - Discovery Science
  - Knowledge Discovery

*G. Piatetsky-Shapiro and W. J. Frawley, Knowledge Discovery in Databases, AAAI/MIT Press, 1991.

Database vs. Data Mining

- Query
  - Well defined
  - SQL

■ Output
  - Subset of database

- Field
  - Mature

- Query
  - Poorly defined
  - No precise query language

■ Output
  - Not a subset of database

- Field
  - Still in infancy

Query Examples

Database

- Find all customers that live in Islamabad
- Find all customers that use Mastercard
- Find all customers that missed one payment

Data mining

- Find all customers that are likely to miss one payment (**Classification**)
- Group all customers with simpler buying habits (**Clustering**)
- List all items that are frequently purchased with bicycles (**Association rules**)
- Find any “unusual” customers (**Outlier detection, anomaly discovery**)

The Major Data Mining Tasks

- Classification
- Clustering
- Associations

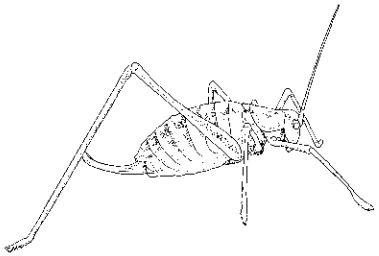
Most of the other tasks (for example, outlier discovery or anomaly detection) make heavy use of one or more of the above.

So in this tutorial we will focus most of our energy on the above, starting with…

The Classification Problem

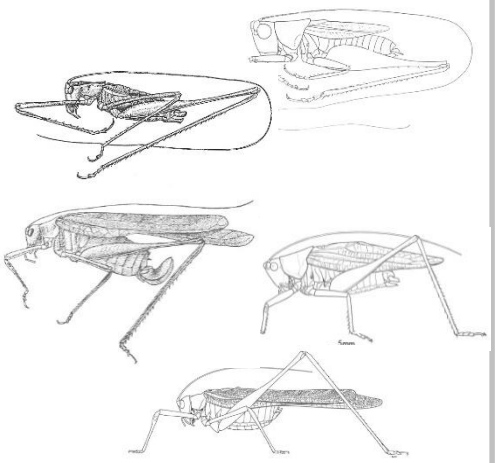
(informal definition)

Given a collection of annotated data. In this case 5 instances of **Katydids** and five of **Grasshoppers**, decide what type of insect the unlabeled example is.

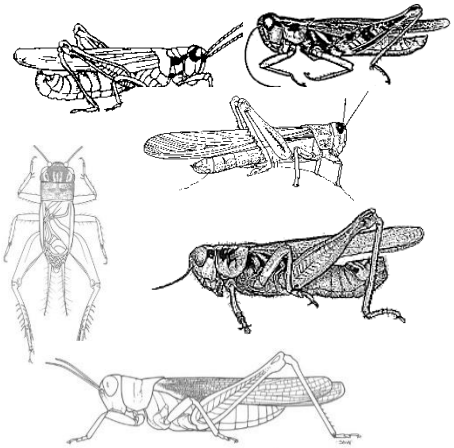


Katydid or **Grasshopper**?

Katydids



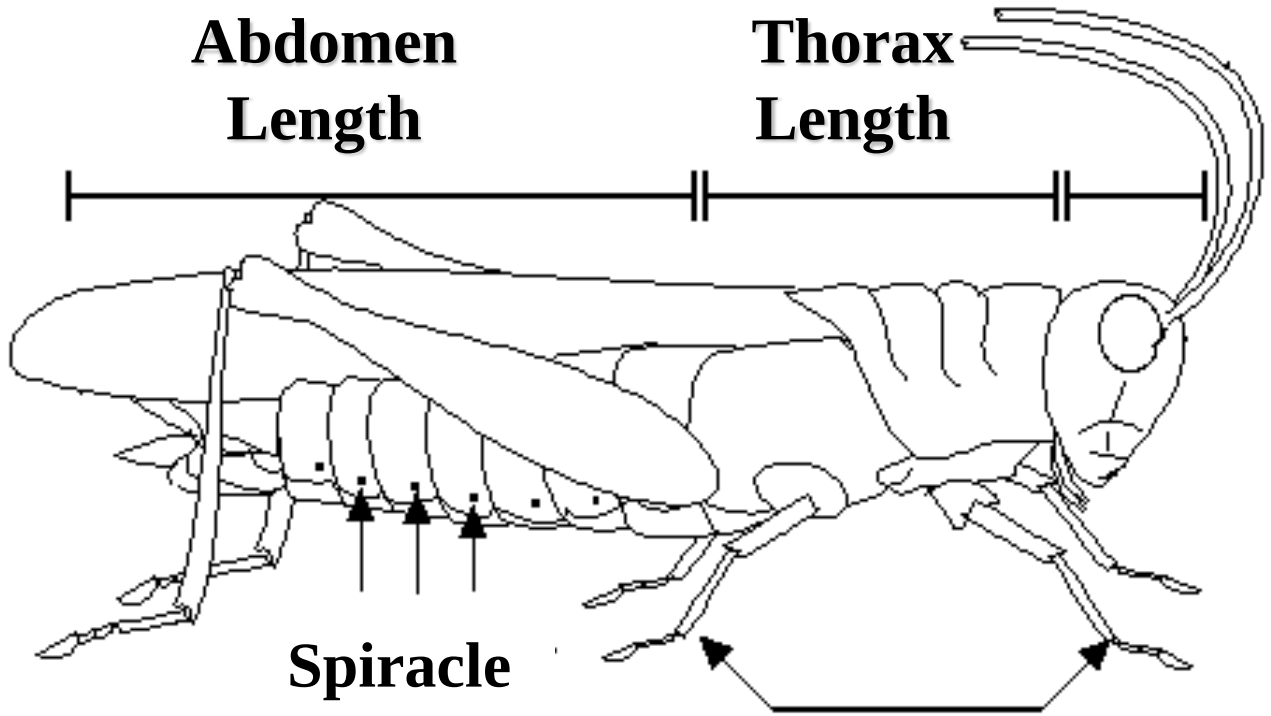
Grasshoppers



For any domain of interest, we can measure *features*

Color {Green, Brown, Gray, Other}

Has Wings?



We can store features in a database.

The classification problem can now be expressed as:

• Given a training database (**My_Collection**), predict the **class** label of a previously unseen instance

My_Collection

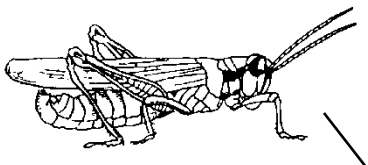
Insect ID	Abdomen Length	Antennae Length	Insect Class
1	2.7	5.5	**Grasshopper**
2	8.0	9.1	**Katydid**
3	0.9	4.7	**Grasshopper**
4	1.1	3.1	**Grasshopper**
5	5.4	8.5	**Katydid**
6	2.9	1.9	**Grasshopper**
7	6.1	6.6	**Katydid**
8	0.5	1.0	**Grasshopper**
9	8.3	6.6	**Katydid**
10	8.1	4.7	**Katydids**

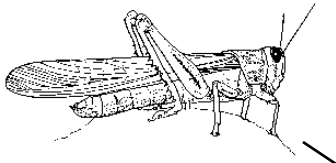
previously unseen instance =

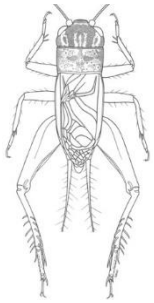
11	5.1	7.0	???????

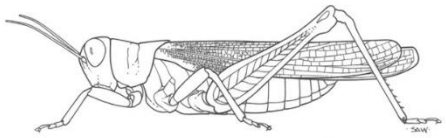
Grasshoppers

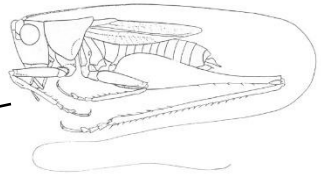
Katydids

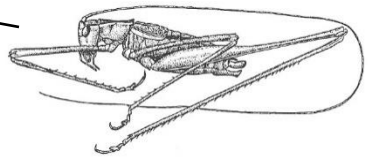


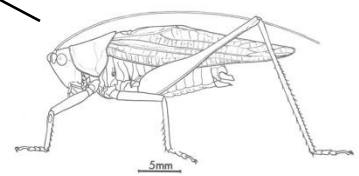


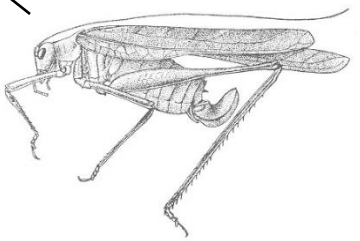


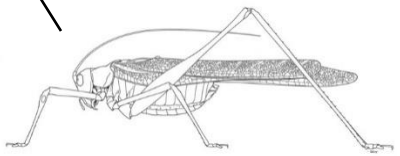


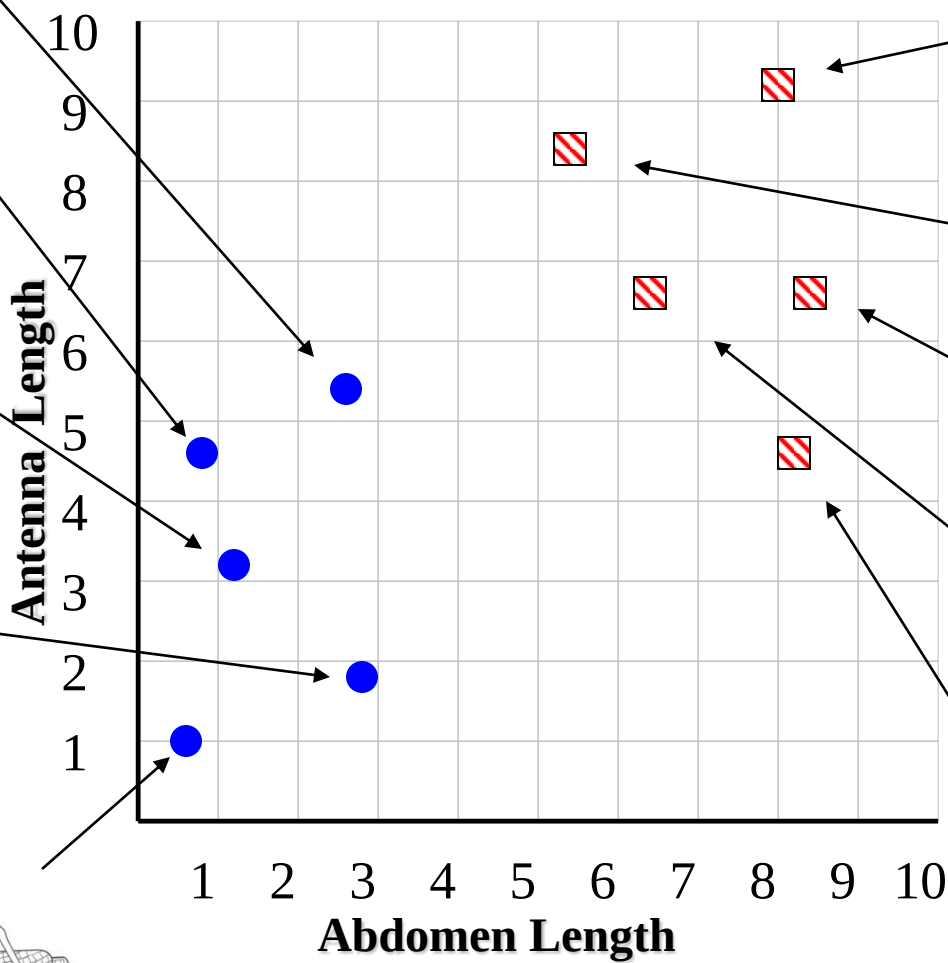




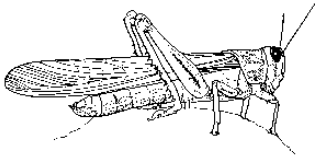




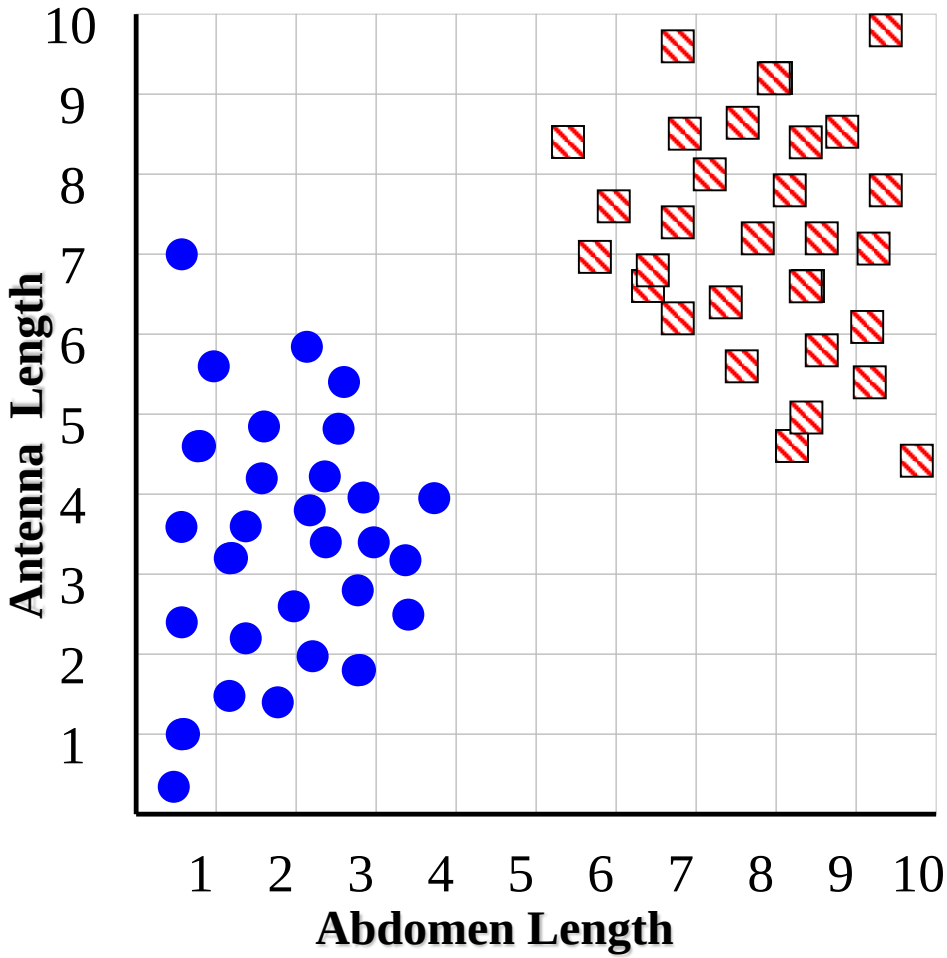




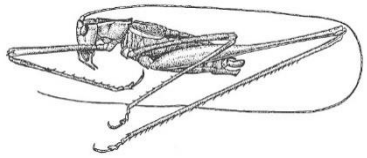
Grasshoppers



We will also use this lager dataset as a motivating example…



Katydids



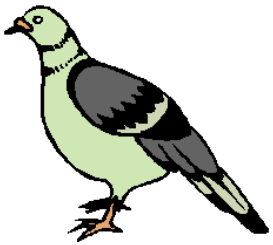
Each of these data objects are called…

- exemplars
- (training) examples
- instances
- tuples

We will return to the previous slide in two minutes. In the meantime, we are going to play a quick game.

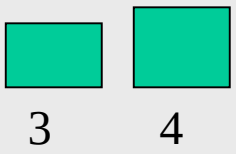
I am going to show you some classification problems which were shown to pigeons!

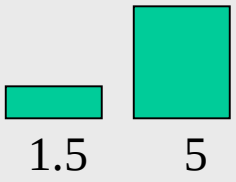
Let us see if you are as smart as a pigeon!

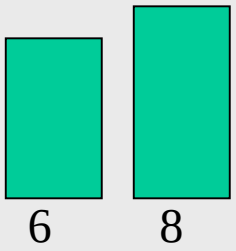


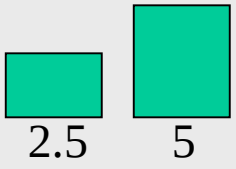
Pigeon Problem 1

Examples of class A

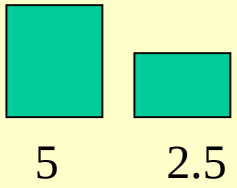


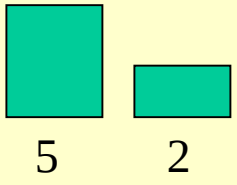


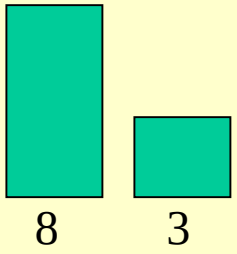


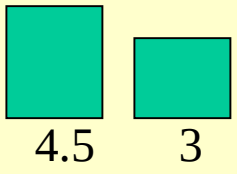


Examples of class B



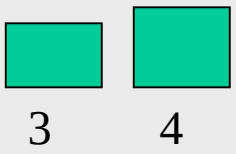


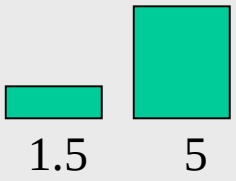


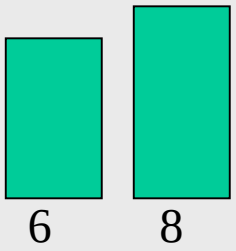


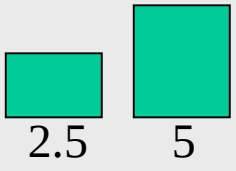
Pigeon Problem 1

Examples of class A

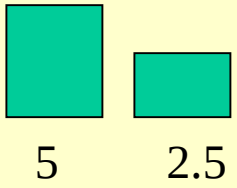


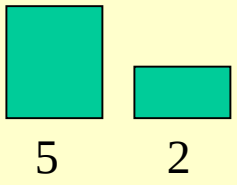


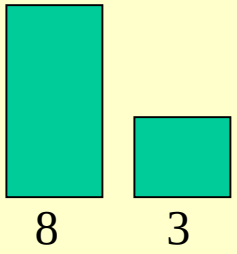


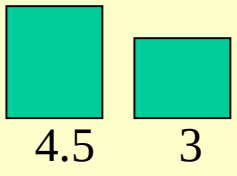


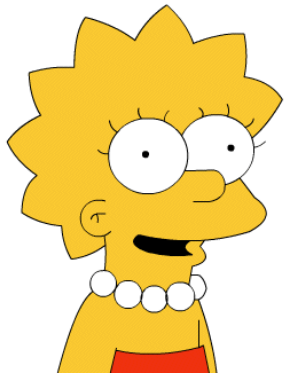
Examples of class B



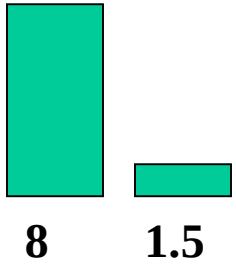




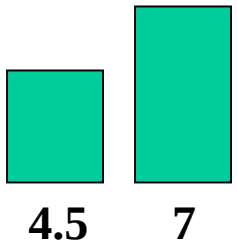




What class is this object?

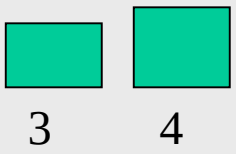


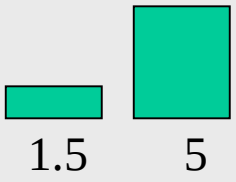
What about this one, A or B?

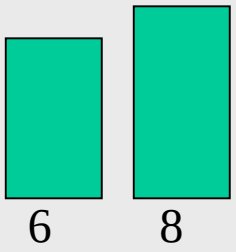


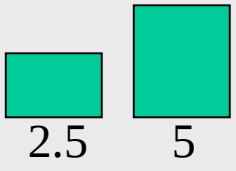
Pigeon Problem 1

Examples of class A

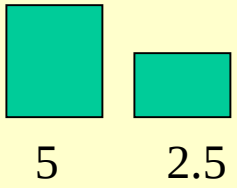


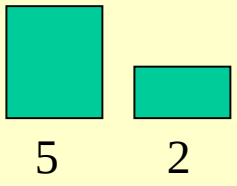


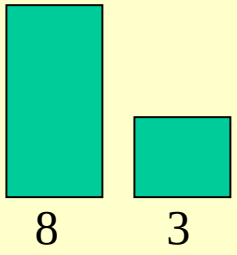


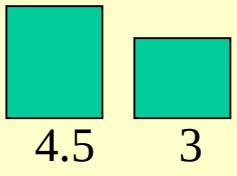


Examples of class B

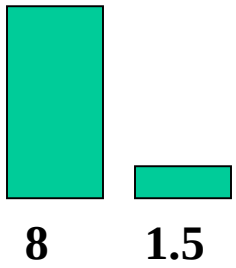








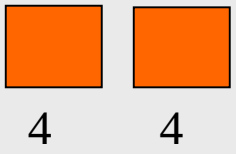
This is a B!

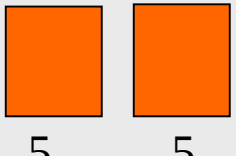


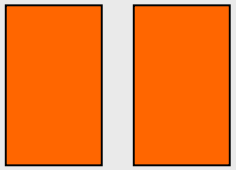
Here is the rule.
If the left bar is smaller than the right bar, it is an A, otherwise it is a B.

Pigeon Problem 2

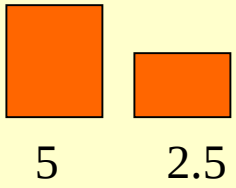
Examples of class A

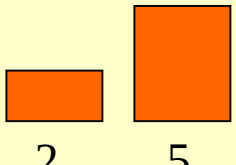


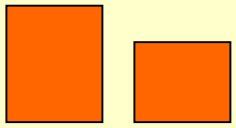


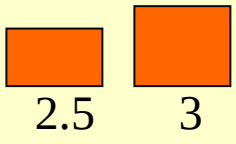


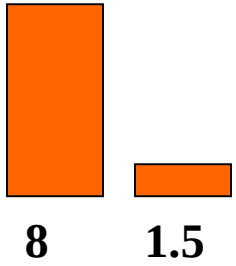
Examples of class B

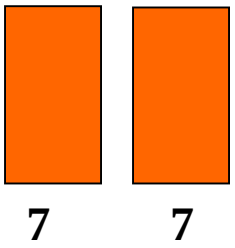






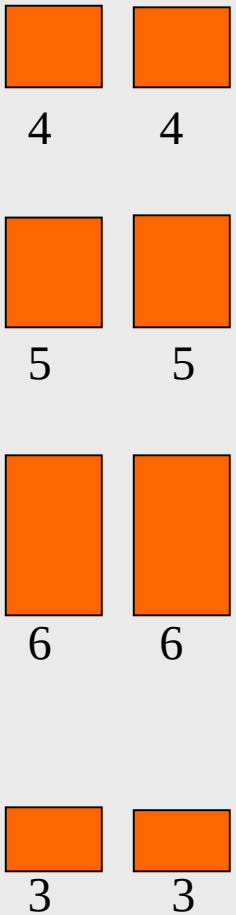




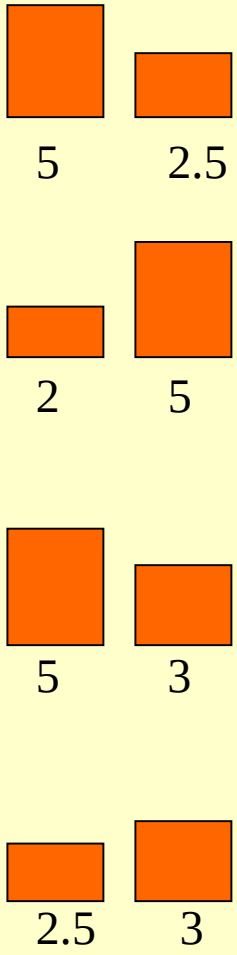


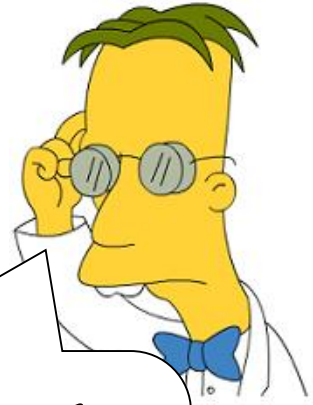
Pigeon Problem 2

Examples of class A



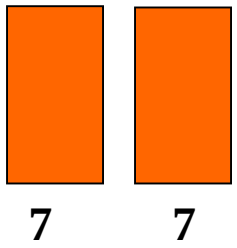
Examples of class B





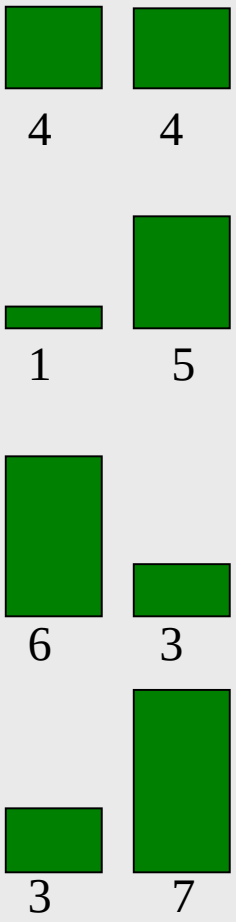
The rule is as follows, if the two bars are equal sizes, it is an A. Otherwise it is a B.

So this one is an A.

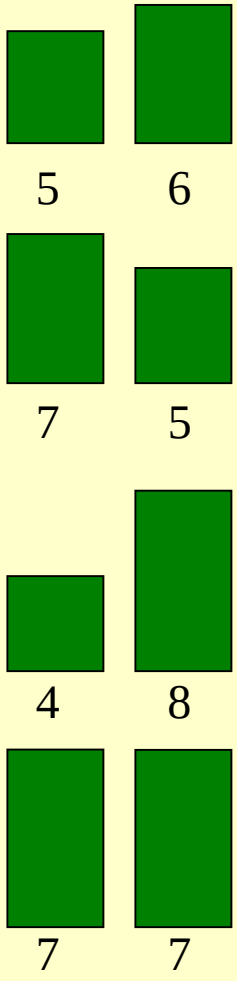


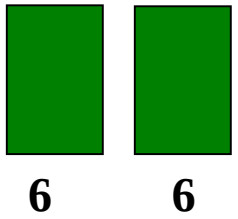
Pigeon Problem 3

Examples of class A



Examples of class B

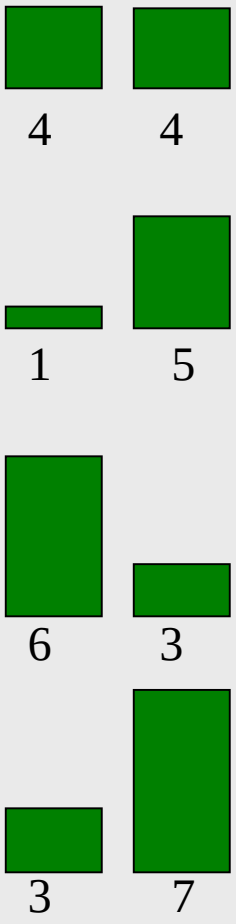




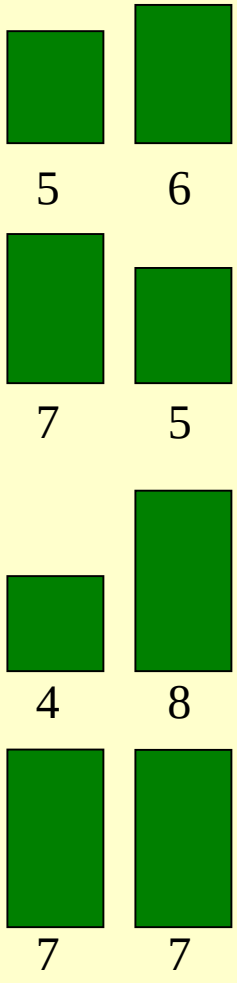
This one is really hard!
What is this, A or B?

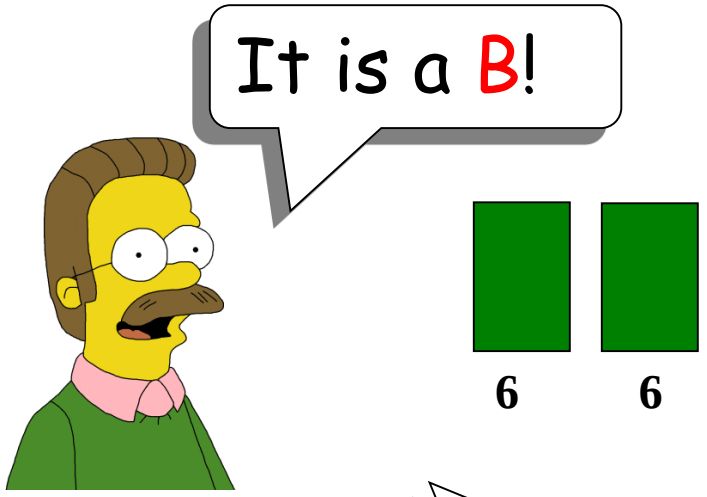
Pigeon Problem 3

Examples of class A



Examples of class B

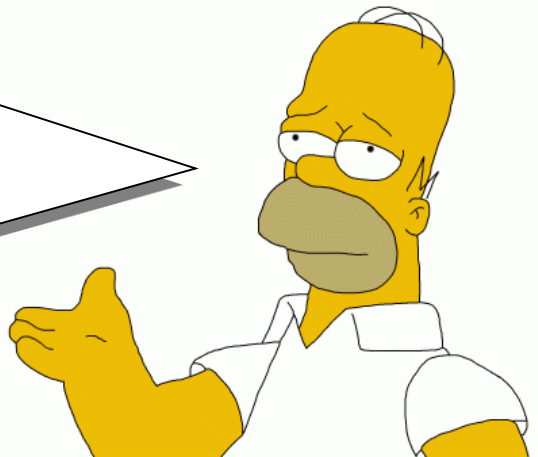




The rule is as follows, if the square of the sum of the two bars is less than or equal to 100, it is an A. Otherwise it is a B.

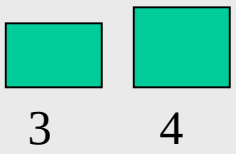
Why did we spend so much time with this game?

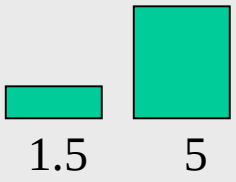
Because we wanted to show that almost all classification problems have a geometric interpretation, check out the next 3 slides…

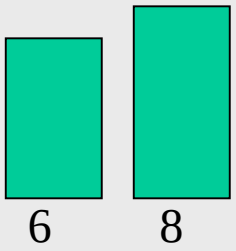


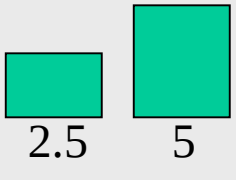
Pigeon Problem 1

Examples of class A

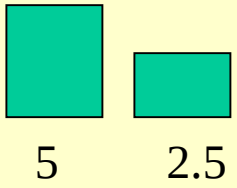


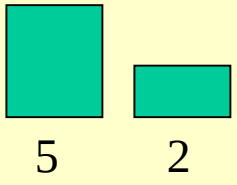


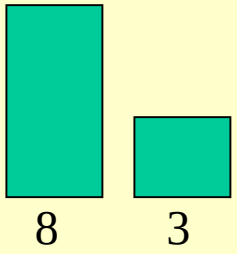


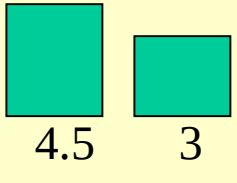


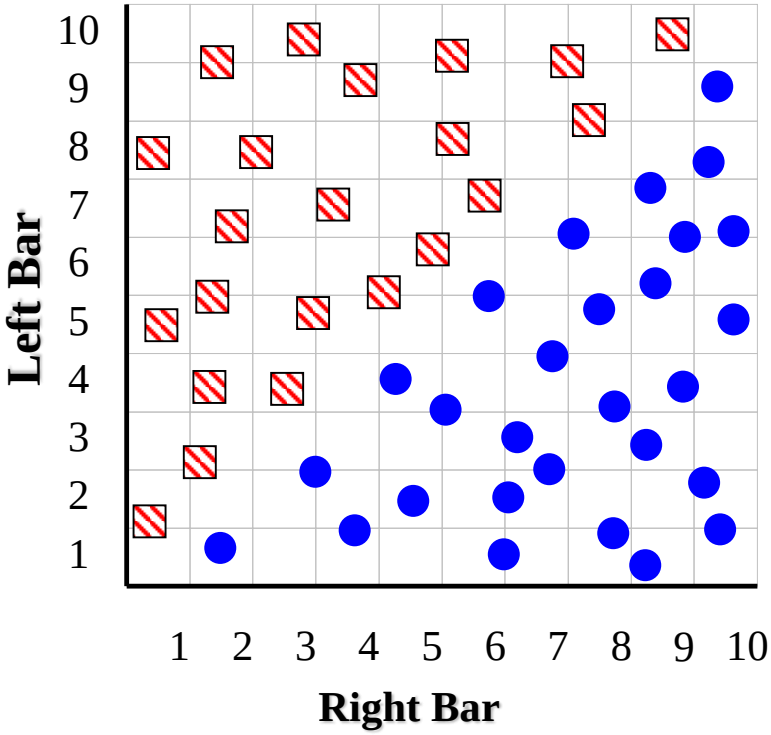
Examples of class B







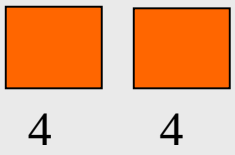


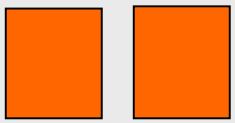


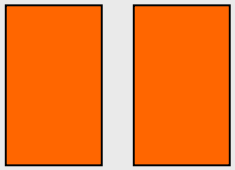
Here is the rule again.
If the left bar is smaller than the right bar, it is an A, otherwise it is a B.

Pigeon Problem 2

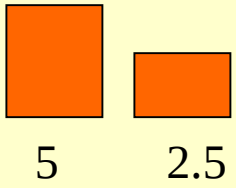
Examples of class A

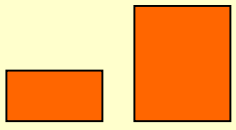


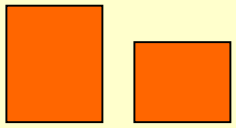


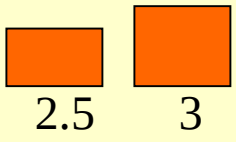


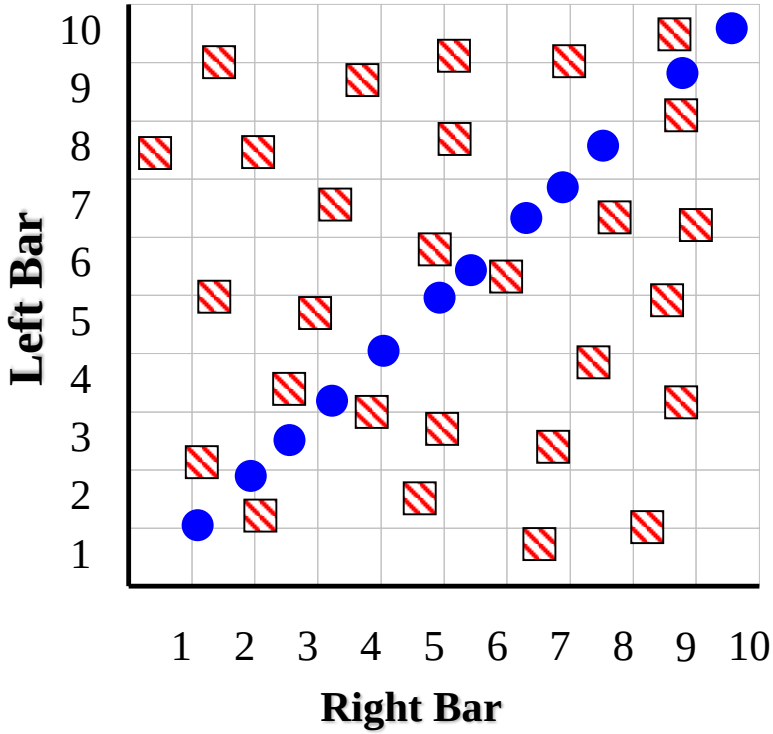
Examples of class B







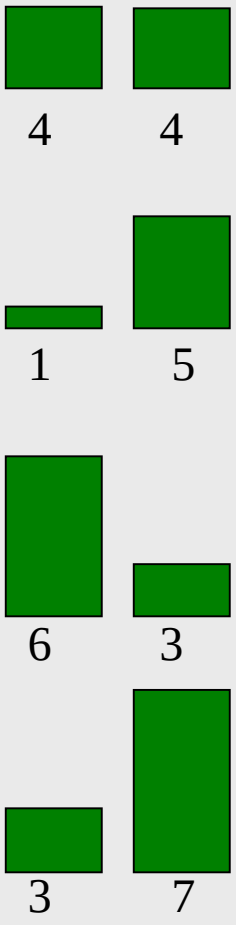




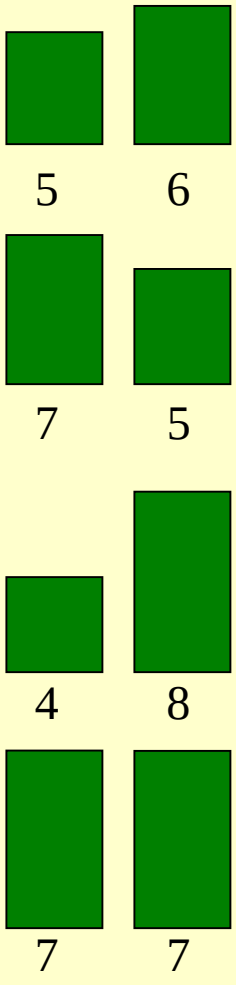
Let me look it up... here it is.. the rule is, if the two bars are equal sizes, it is an A. Otherwise it is a B.

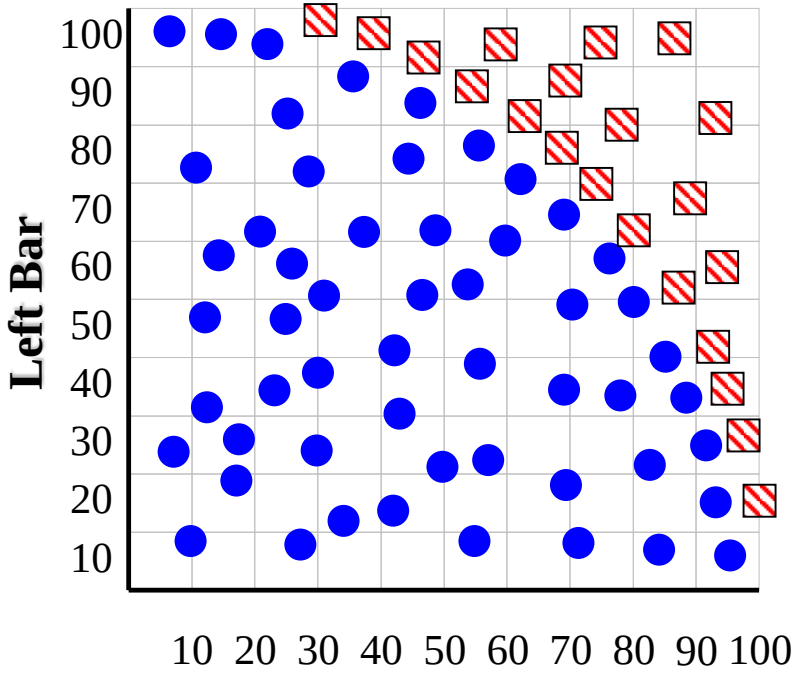
Pigeon Problem 3

Examples of class A



Examples of class B

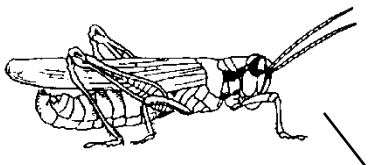


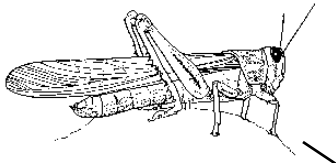


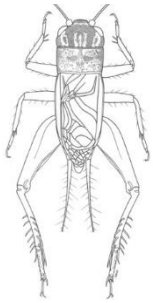
The rule again:
if the square of the sum of the two bars is less than or equal to 100, it is an A. Otherwise it is a B.

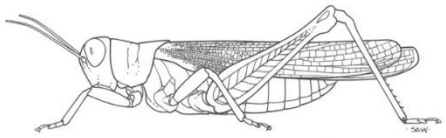
Grasshoppers

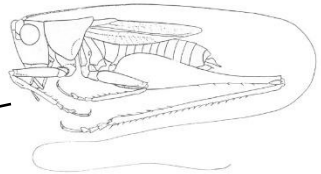
Katydids

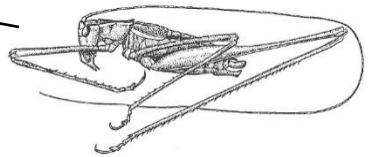


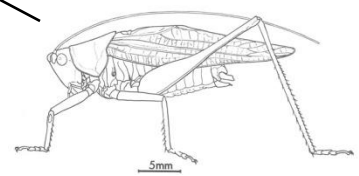


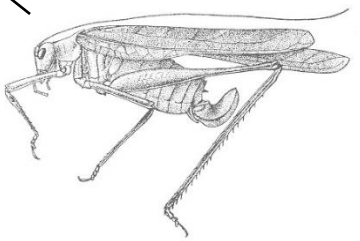


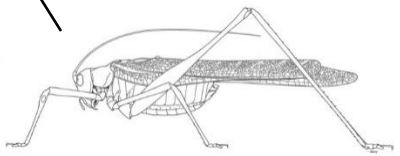


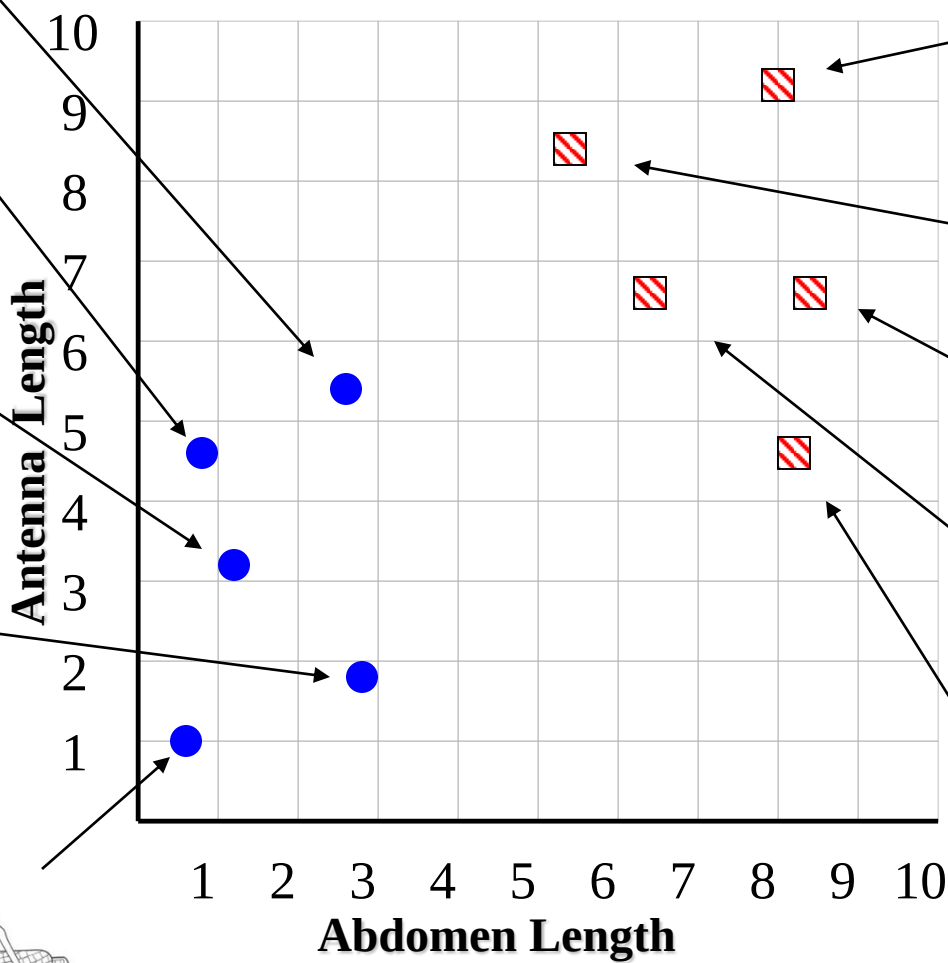




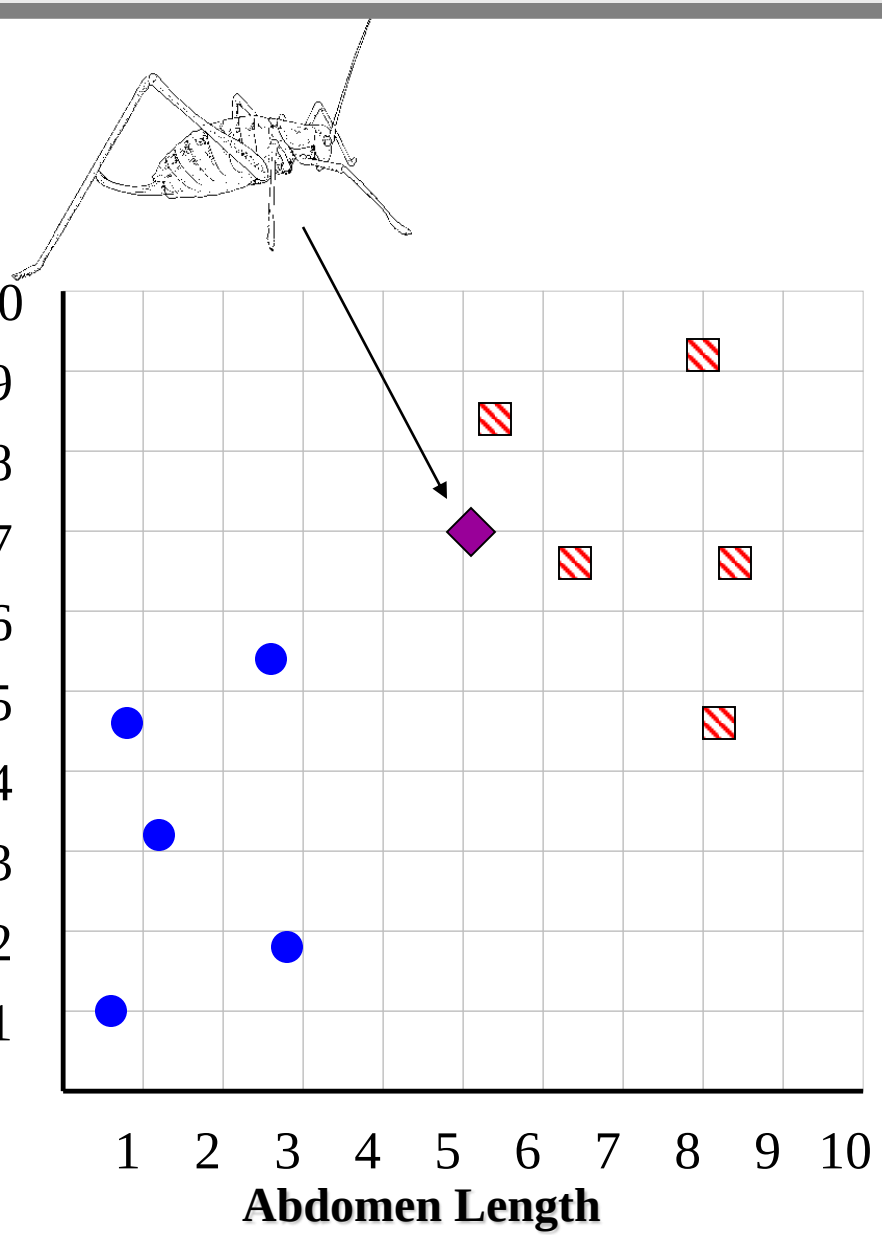








previously unseen instance =	*11*	**5.1**	**7.0**	**???????**

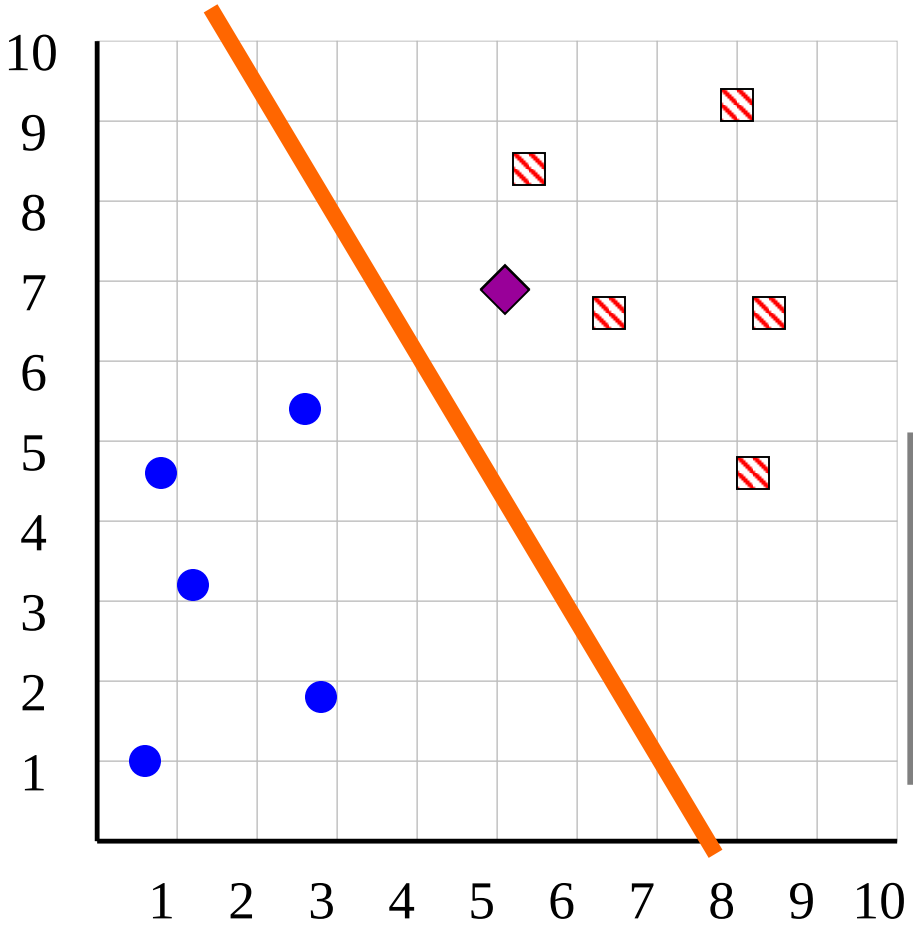


- We can "project" the previously unseen instance into the same space as the database.
- We have now abstracted away the details of our particular problem. It will be much easier to talk about points in space.

▨ **Katydids**

● **Grasshoppers**

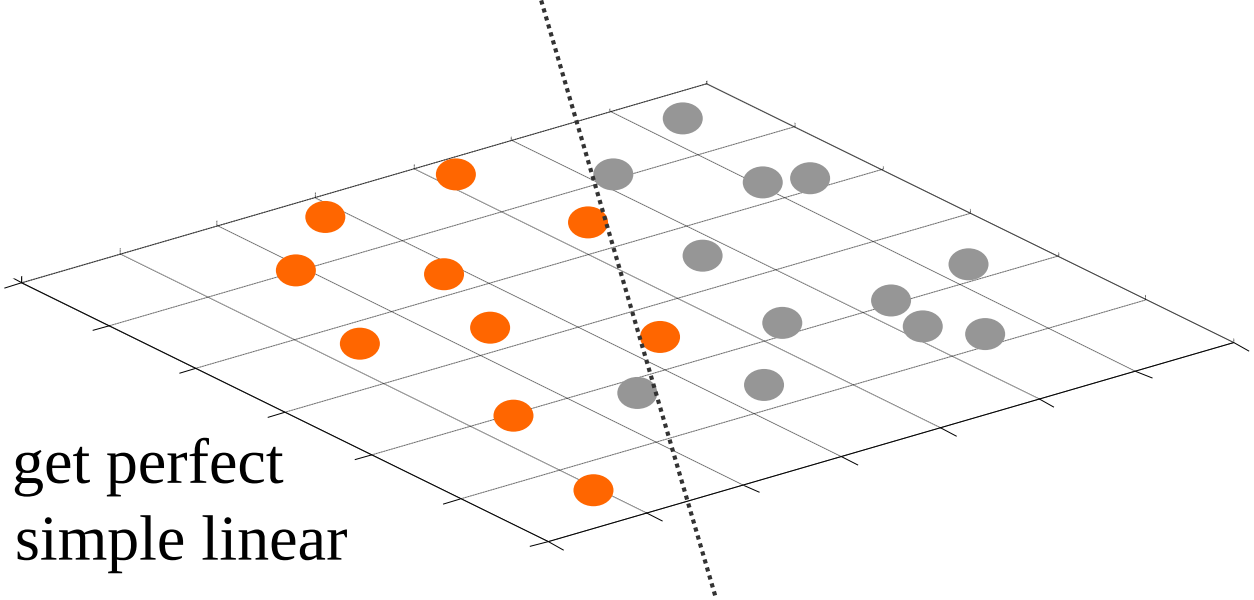
Simple Linear Classifier



R.A. Fisher
1890-1962

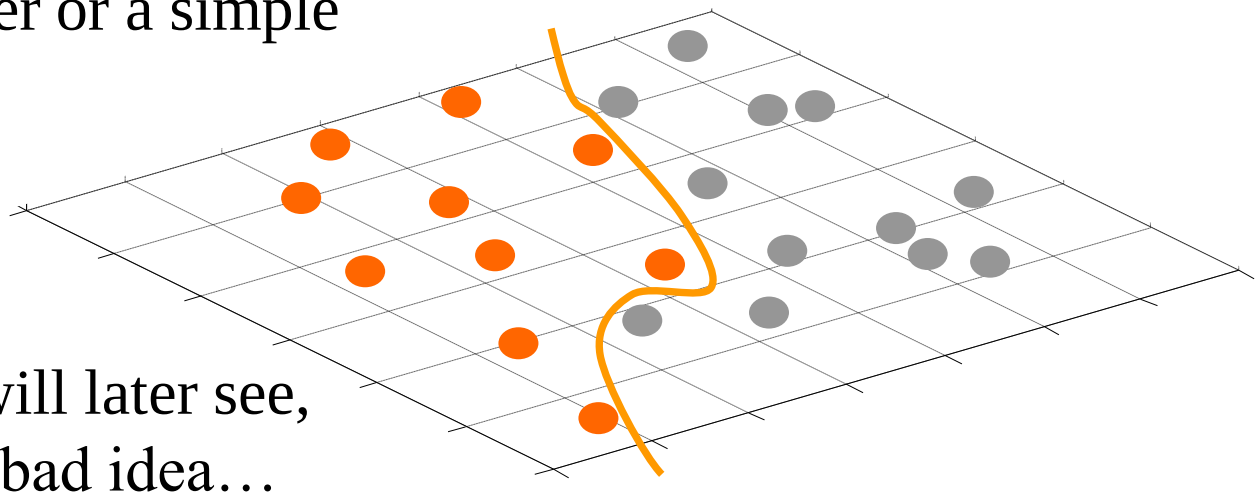
If previously unseen instance above the line
then
class is **Katydid**
else
class is **Grasshopper**

Katydids
Grasshoppers



We can no longer get perfect accuracy with the simple linear classifier…

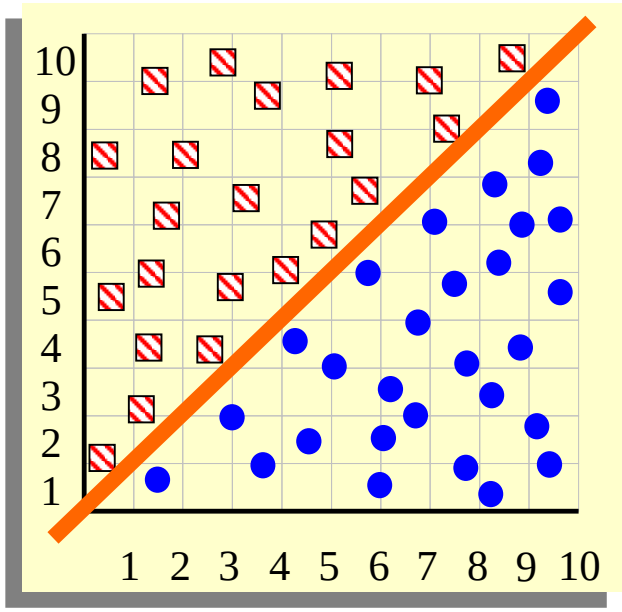
We could try to solve this problem by user a simple *quadratic* classifier or a simple *cubic* classifier..

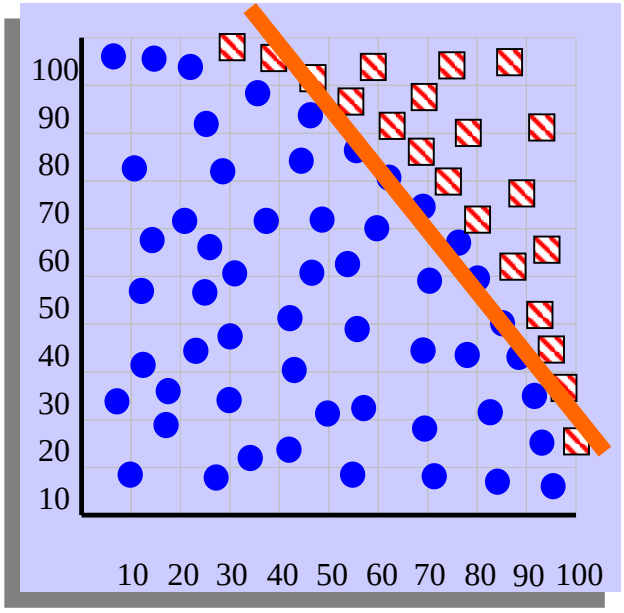


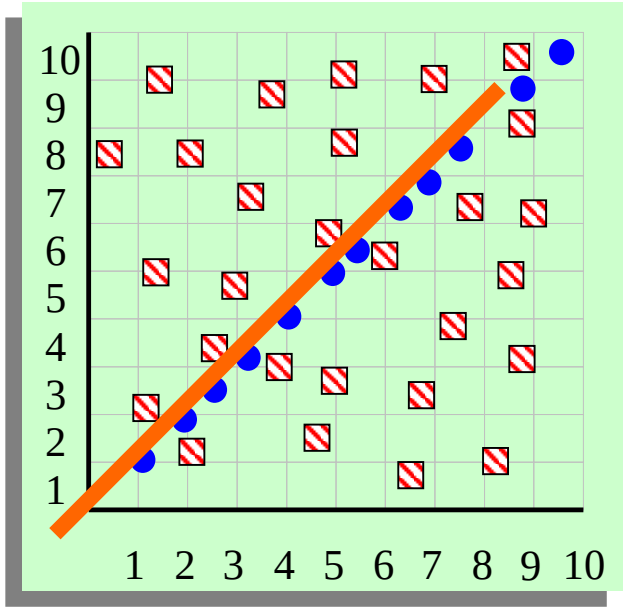
However, as we will later see, this is probably a bad idea…

Which of the “Pigeon Problems” can be solved by the Simple Linear Classifier?

1) Perfect
2) Useless
3) Pretty Good







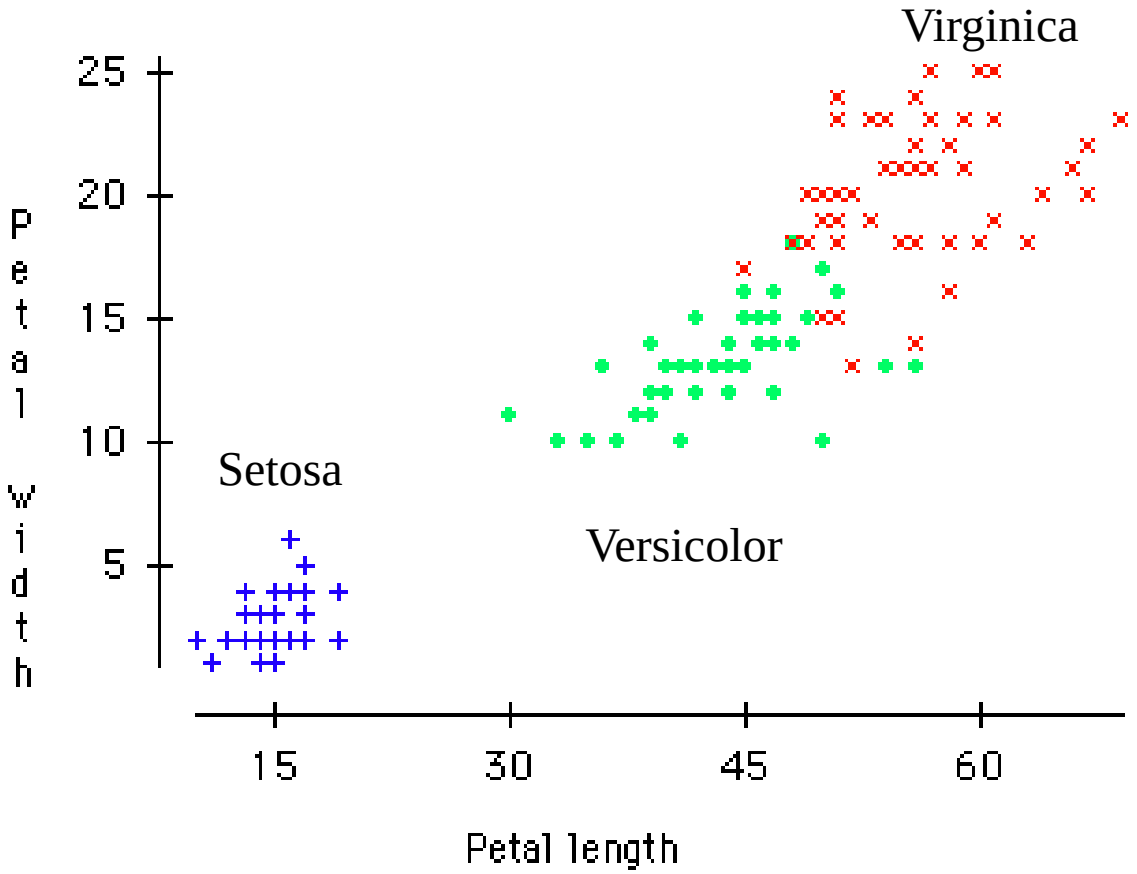
Problems that can be solved by a linear classifier are call **linearly separable**.

A Famous Problem

R. A. Fisher's Iris Dataset.

- 3 classes
- 50 of each class

The task is to classify Iris plants into one of 3 varieties using the Petal Length and Petal Width.

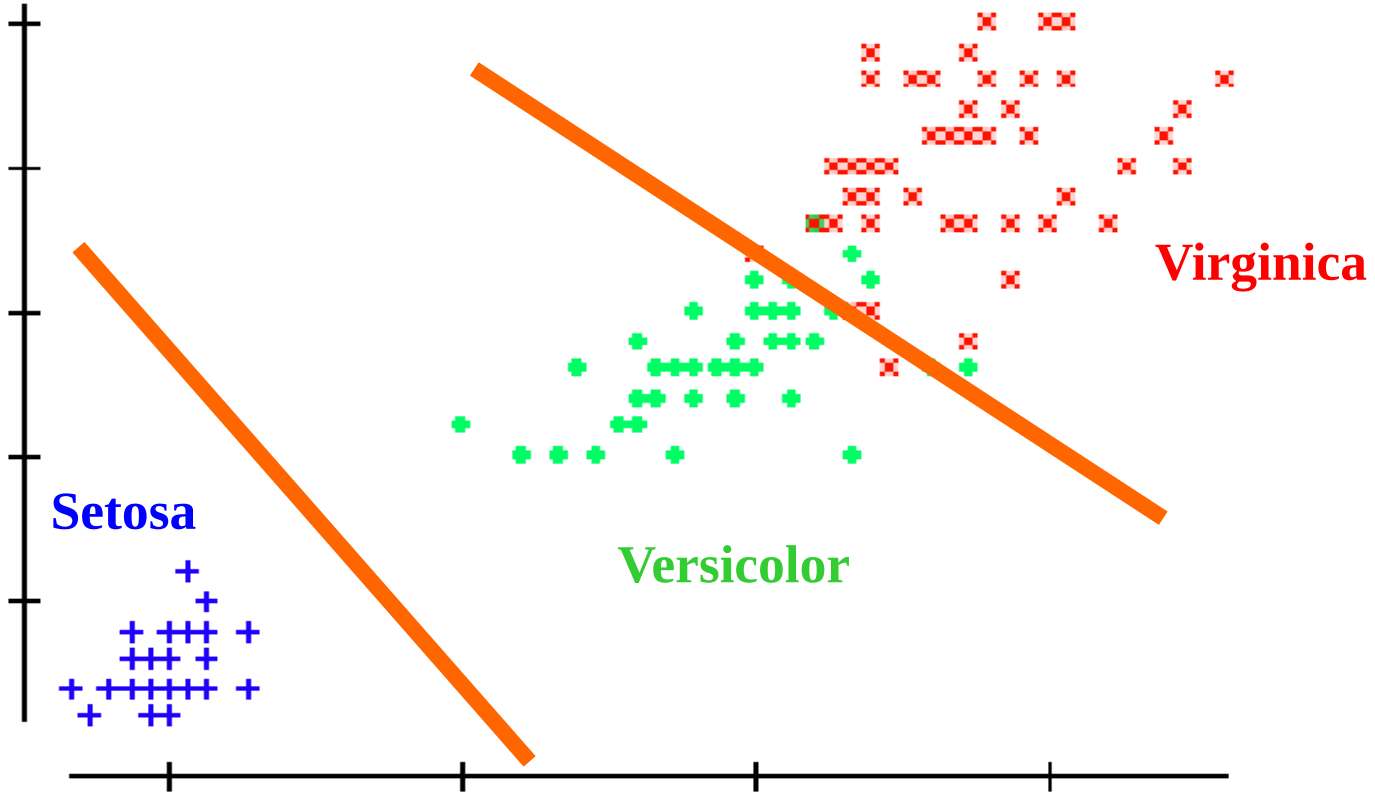


Iris Setosa

Iris Versicolor

Iris Virginica

We can generalize the piecewise linear classifier to N classes, by fitting N-1 lines. In this case we first learned the line to (perfectly) discriminate between **Setosa** and **Virginica**/**Versicolor**, then we learned to approximately discriminate between **Virginica** and **Versicolor**.



If petal width > 3.272 – (0.325 * petal length) **then** class = **Virginica**
Elseif petal width…