Mineração de Dados 2018.2

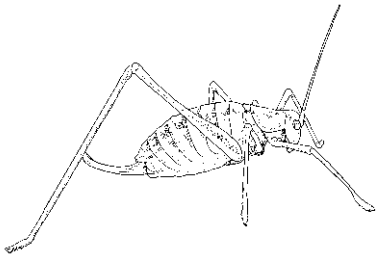
Introdução a classificação de dados

(slides baseados no material do Prof. Carlos Soares e Prof. Eamonn Keogh [eamonn@cs.ucr.edu])

O problema de classificação

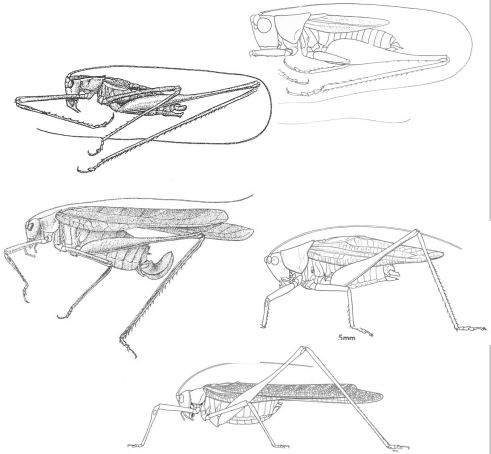
(definição informal)

Dada uma coleção de dados detalhados, neste caso 5 exemplos de **Esperança** e 5 do **Gafanhoto**, decida a qual tipo de inseto o exemplo não rotulado pertence.
Obs: **Esperança** : tipo de gafanhoto verde.

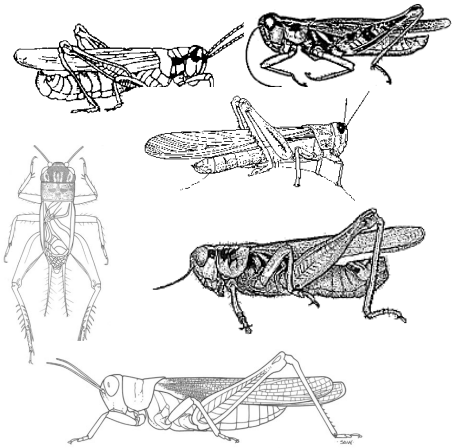


Esperança ou **Gafanhoto?**

Esperança



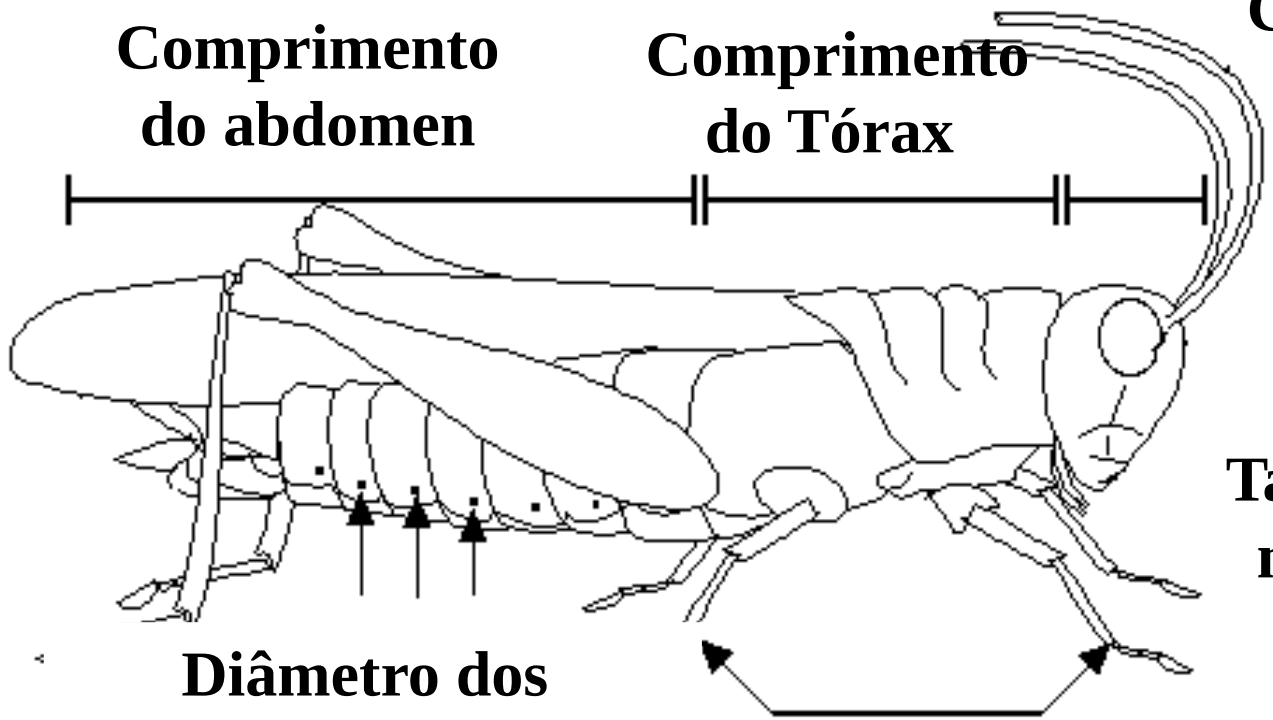
Gafanhoto



Para qualquer domínio de interesse podemos medir *características*

Cor {Verde, Marrom, Cinza, Outra}

Tem asas?



Podemos armazenar as *características* em bases de dados

O problema de classificação agora pode ser expresso da seguinte forma:

- Dada uma base de treinamento(**Minha_Coleção**), prediga o rótulo da **classe dos exemplos ainda não vistos**

Minha_Coleção

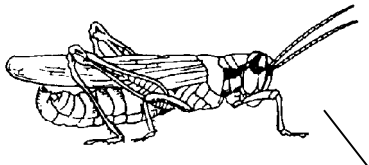
ID do inseto	Comp. do abdômen	Comp. das antenas	Classe do inseto
1	2.7	5.5	**Gafanhoto**
2	8.0	9.1	**Esperança**
3	0.9	4.7	**Gafanhoto**
4	1.1	3.1	**Gafanhoto**
5	5.4	8.5	**Esperança**
6	2.9	1.9	**Gafanhoto**
7	6.1	6.6	**Esperança**
8	0.5	1.0	**Gafanhoto**
9	8.3	6.6	**Esperança**
10	8.1	4.7	**Esperança**

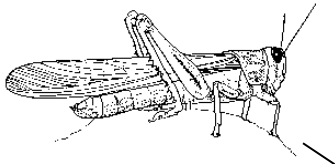
Exemplo não visto =

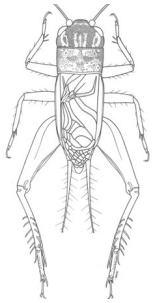
11	5.1	7.0	???????

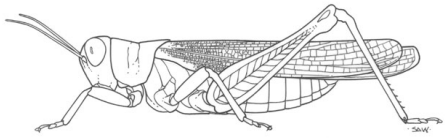
Gafanhoto

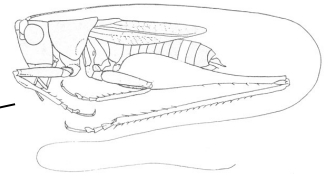
Esperança

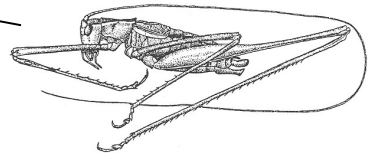


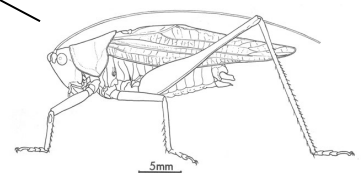


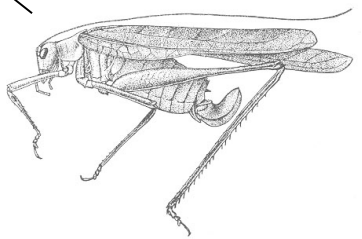


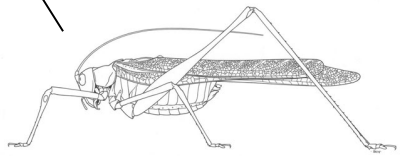


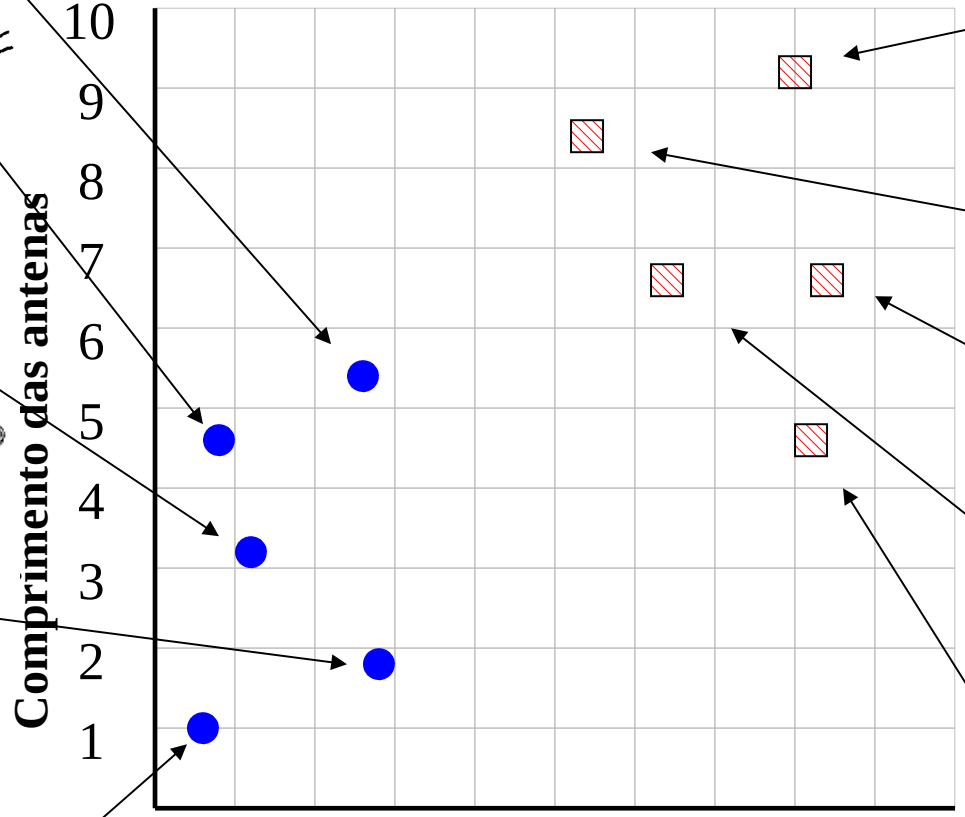




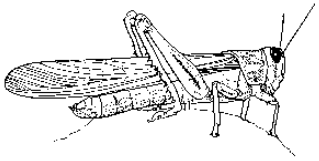




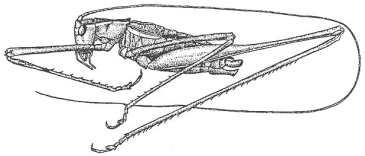




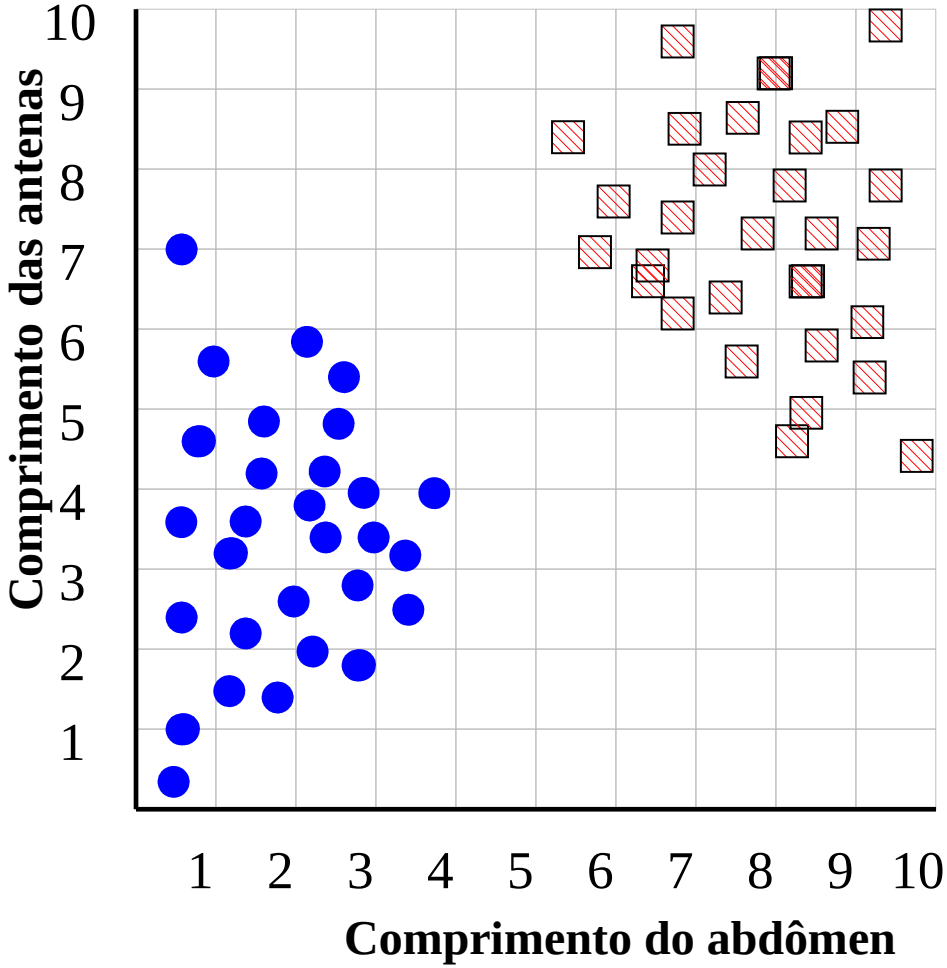
Gafanhoto



Esperança

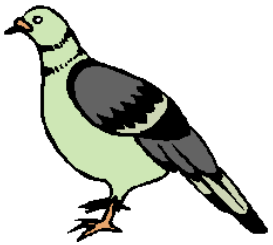


Também utilizaremos esta base de dados maior para motivação …



Cada um destes objetos de dados é chamado de…

- exemplar
- exemplo (de treinamento)
- instância
- tupla



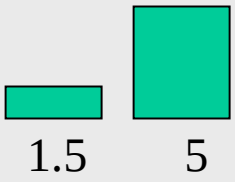
Voltaremos ao slide anterior em dois minutos. Enquanto isso vamos jogar um joguinho rápido.

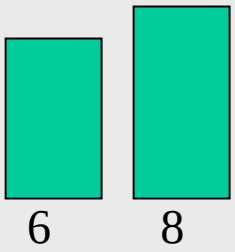
Vou mostrar a vocês alguns problemas de classificação que foram mostrados a pombos!

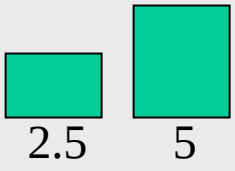
Vamos ver se você é tão esperto quanto um pombo!

Problema do Pombo 1

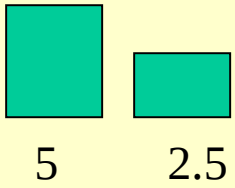
Exemplos da classe A

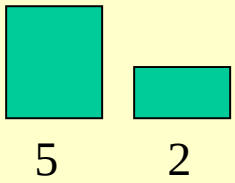


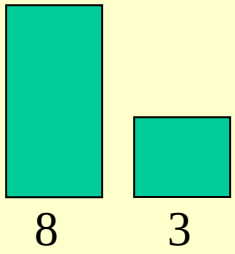


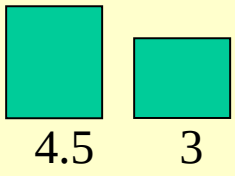


Exemplos da classe B



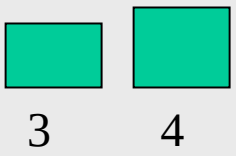


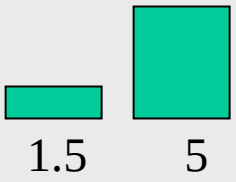


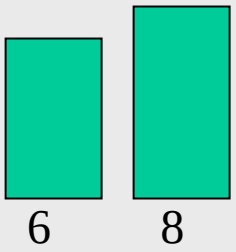


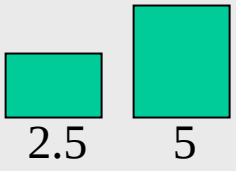
Problema do Pombo 1

Exemplos da classe A

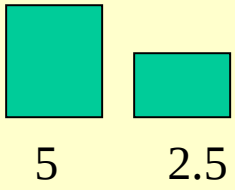


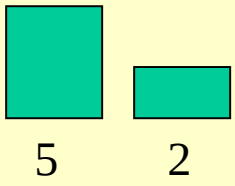


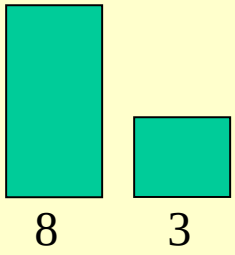


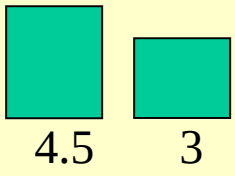


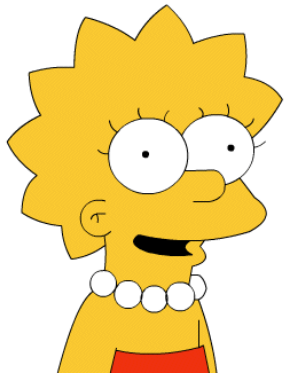
Exemplos da classe B



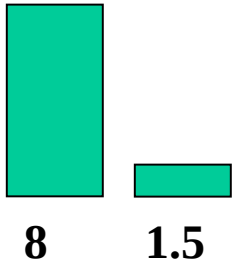




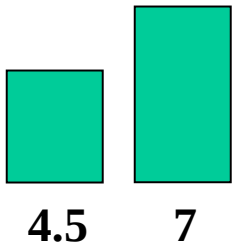




De qual classe é este objeto?

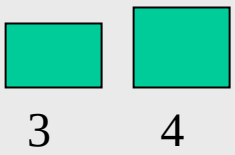


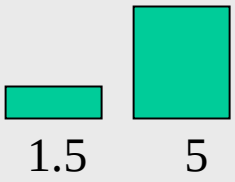
Que tal este, A ou B?

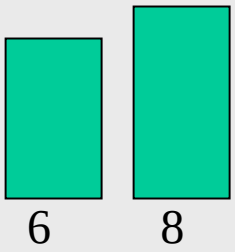


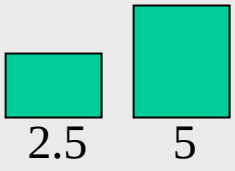
Problema do Pombo 1

Exemplos da classe A

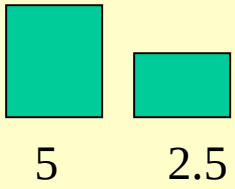


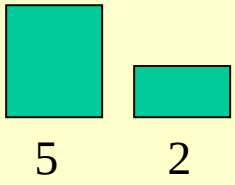


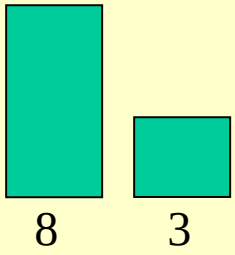


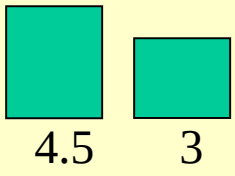


Exemplos da classe B

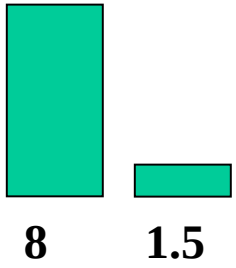








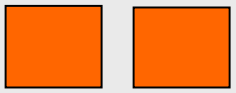
Este é um B!

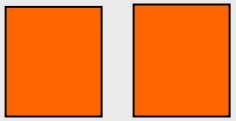


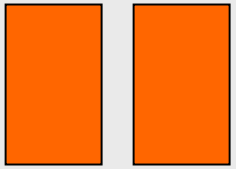
Eis a regra. Se a barra esquerda é menor que a direita, é um A, caso contrário é um B.

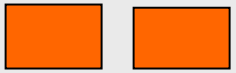
Problema do Pombo 2

Exemplos da classe A

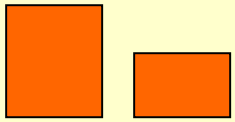

4 4

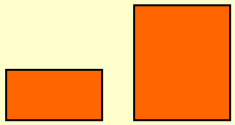

5 5

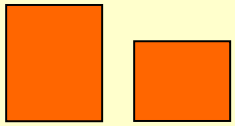

6 6

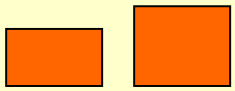

3 3

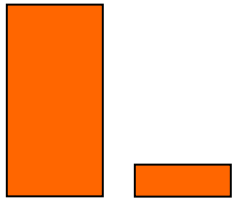
Exemplos da classe B

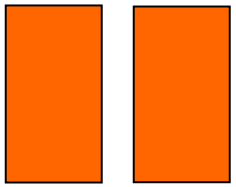

5 2.5


2 5


5 3

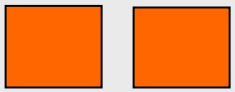

2.5 3

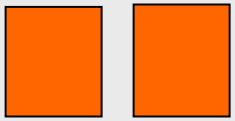

8 1.5

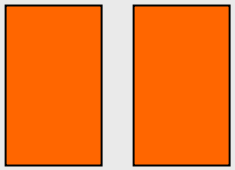

7 7

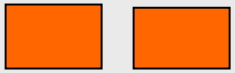
Problema do Pombo 2

Exemplos da classe A

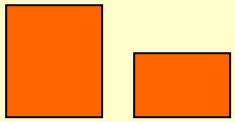

4 4

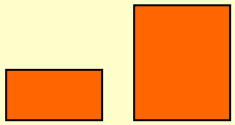

5 5

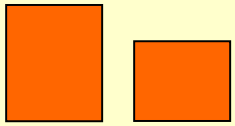

6 6

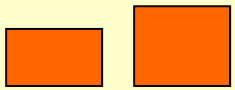

3 3

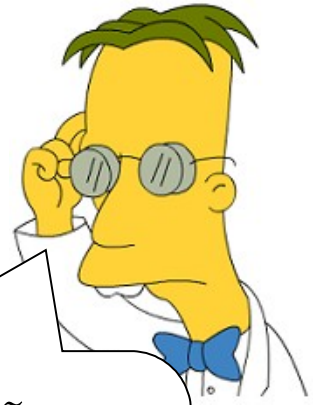
Exemplos da classe B


5 2.5


2 5

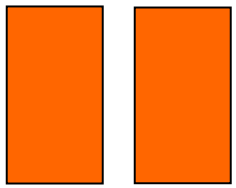

5 3


2.5 3



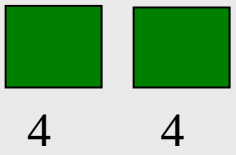
A regra é: se duas barras são iguais em tamanho é um A. Caso contrário é um B.

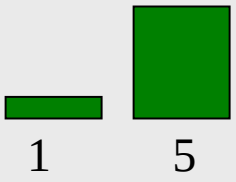
Então este é um A.

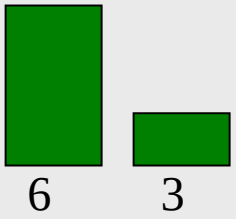

7 7

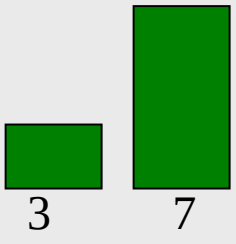
Problema do Pombo 3

Exemplos da classe A

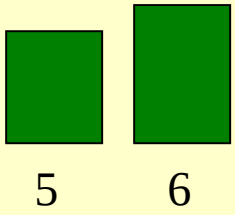


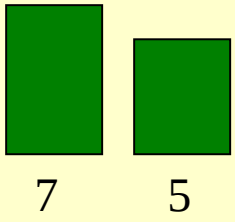


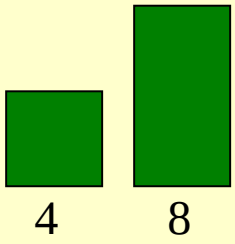


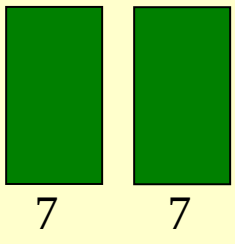


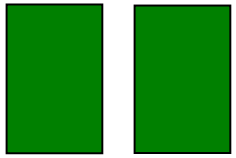
Exemplos da classe B







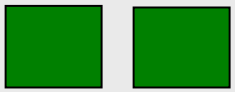




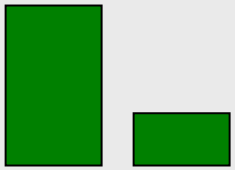
Este é muito difícil!
Qual é este, A ou B?

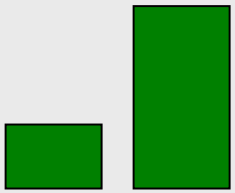
Problema do Pombo 3

Exemplos da classe A

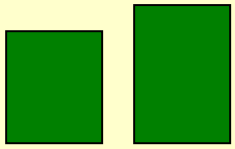

4 4

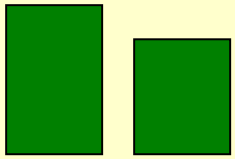

1 5

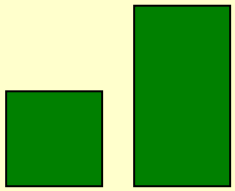

6 3

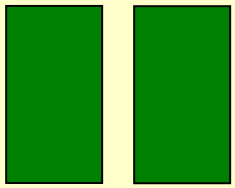

3 7

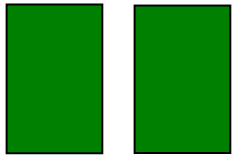
Exemplos da classe B


5 6


7 5


4 8

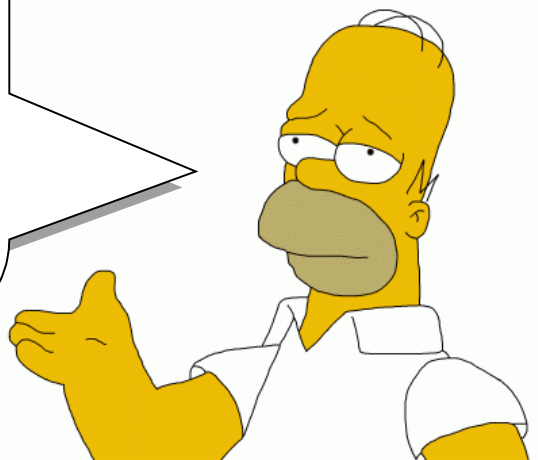

7 7


6 6

A regra é a seguinte, se o quadrado da soma das duas barras é menor ou igual a 100, é um A. Caso contrário é um B.

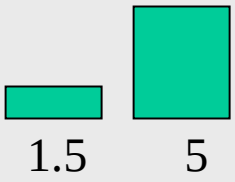
Por que gastamos tanto tempo com este joguinho?

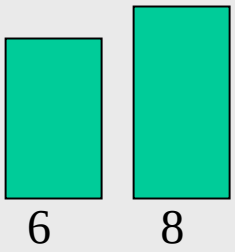
Porque queríamos mostrar que quase todos os problemas de classificação tem uma interpretação geométrica. Confira os próximos 3 slides…

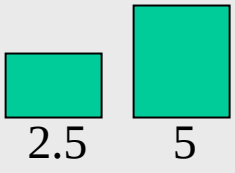


Problema do Pombo 1

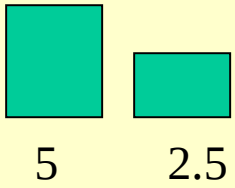
Exemplos da classe A

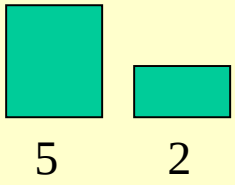


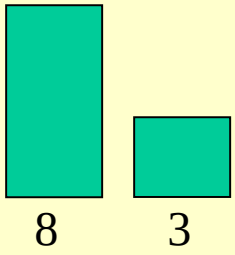


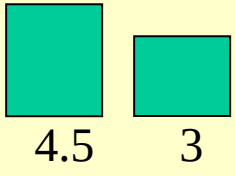


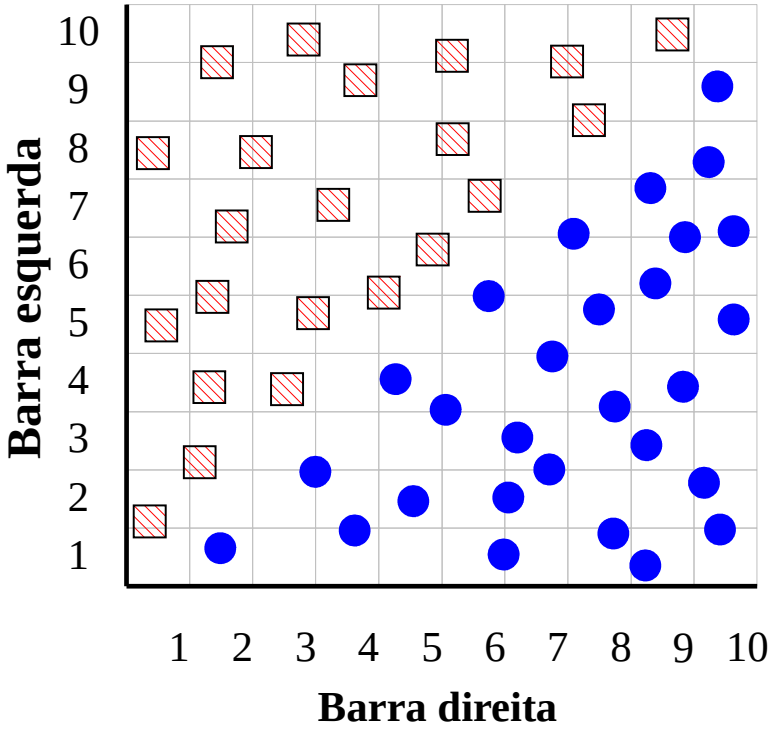
Exemplos da classe B







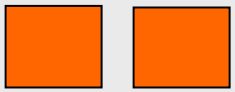


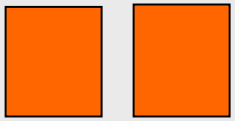


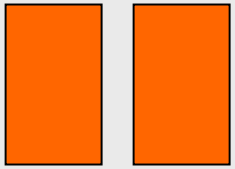
Eis a regra novamente. Se a barra esquerda é menor que a direita, é um A, caso contrário é um B.

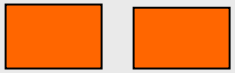
Problema do Pombo 2

Exemplos da classe A

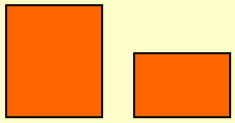

4 4

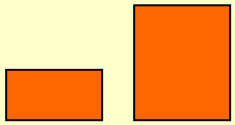

5 5

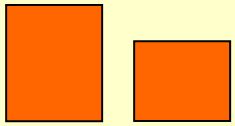

6 6

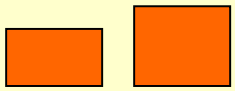

3 3

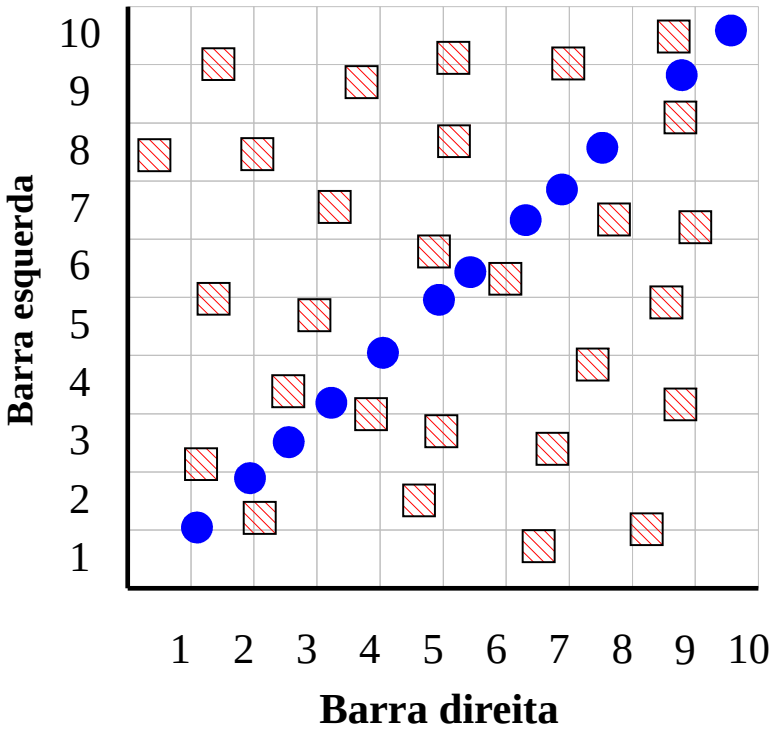
Exemplos da classe B


5 2.5


2 5


5 3

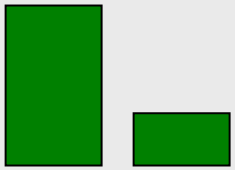

2.5 3

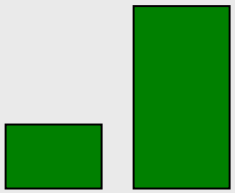


Deixe-me procurar… aqui está… a regra é, se as duas barras têm tamanhos iguais, é um A. Senão é um B.

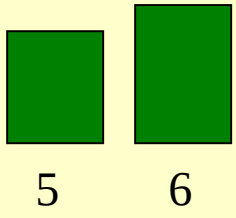
Problema do Pombo 3

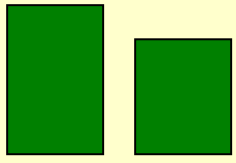
Exemplos da classe A

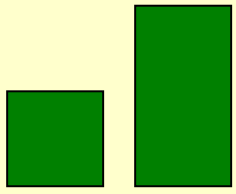


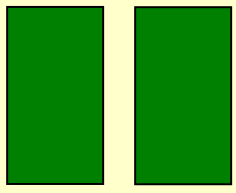


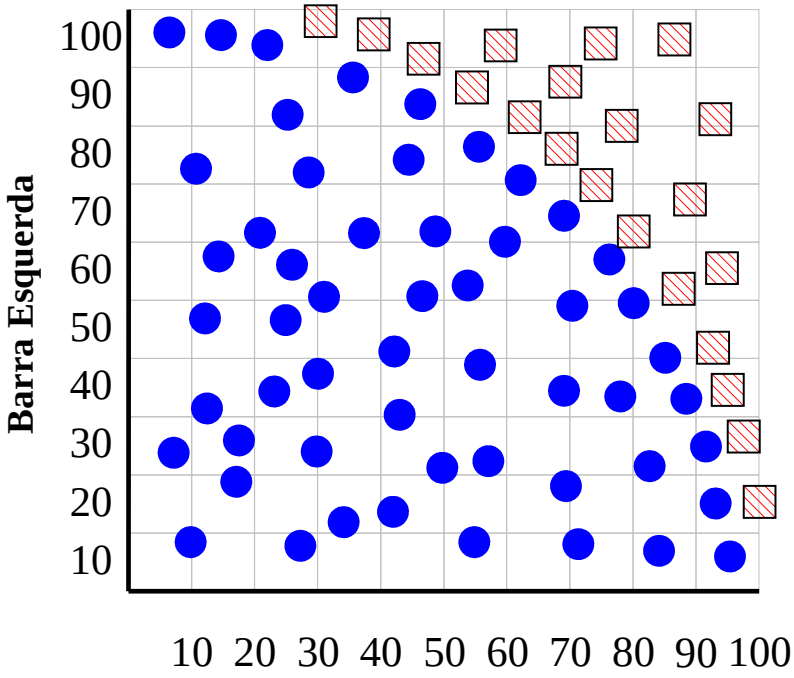
Exemplos da classe B







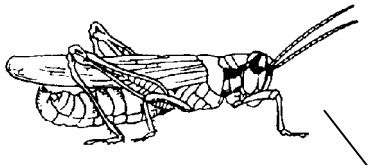


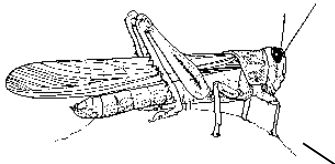


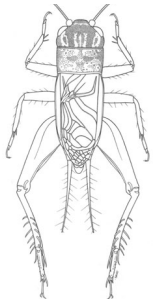
A regra novamente:
Se o quadrado da soma das duas barras é menor ou igual a 100, é um A. Senão é um B.

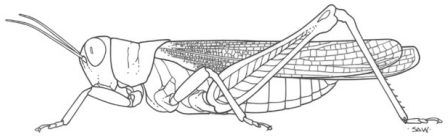
Gafanhoto

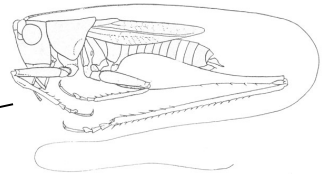
Esperança

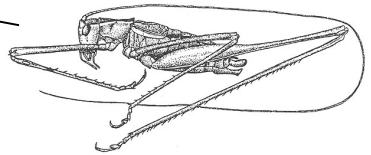


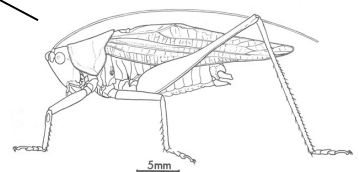


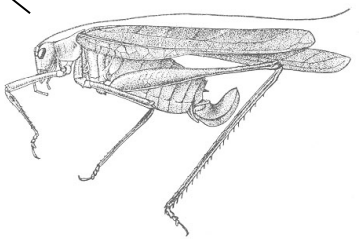


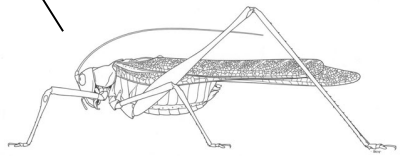


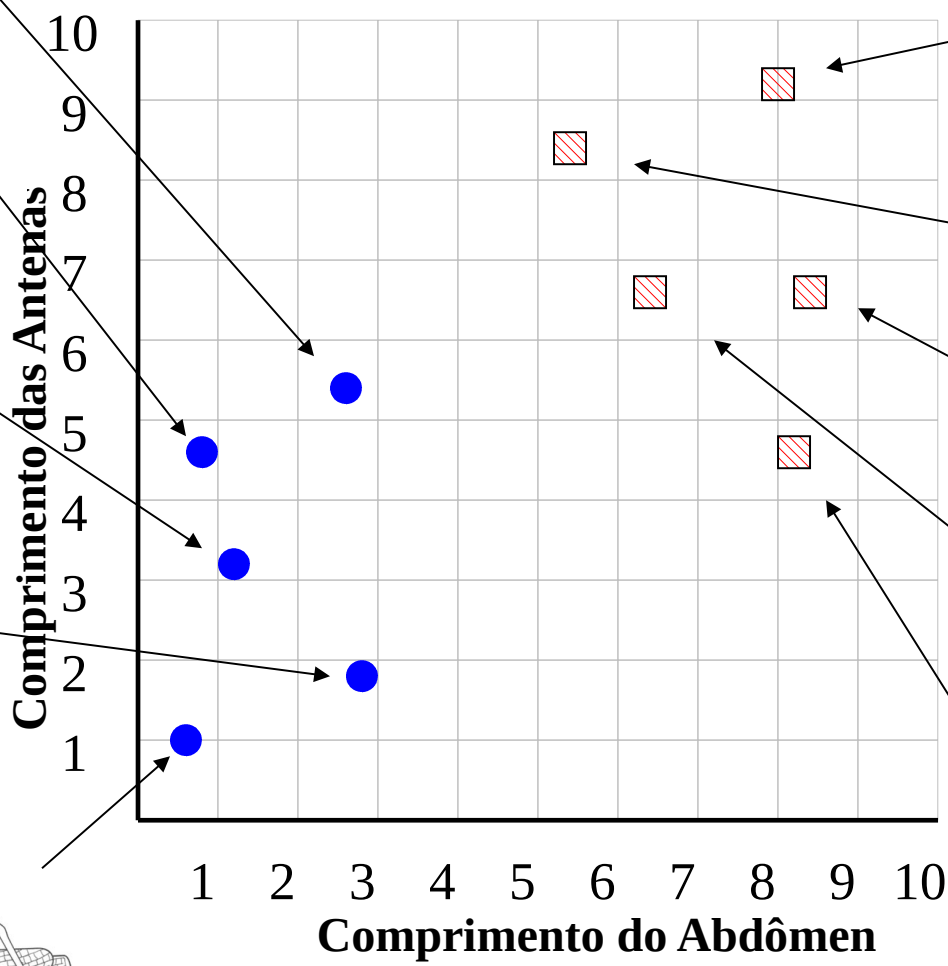






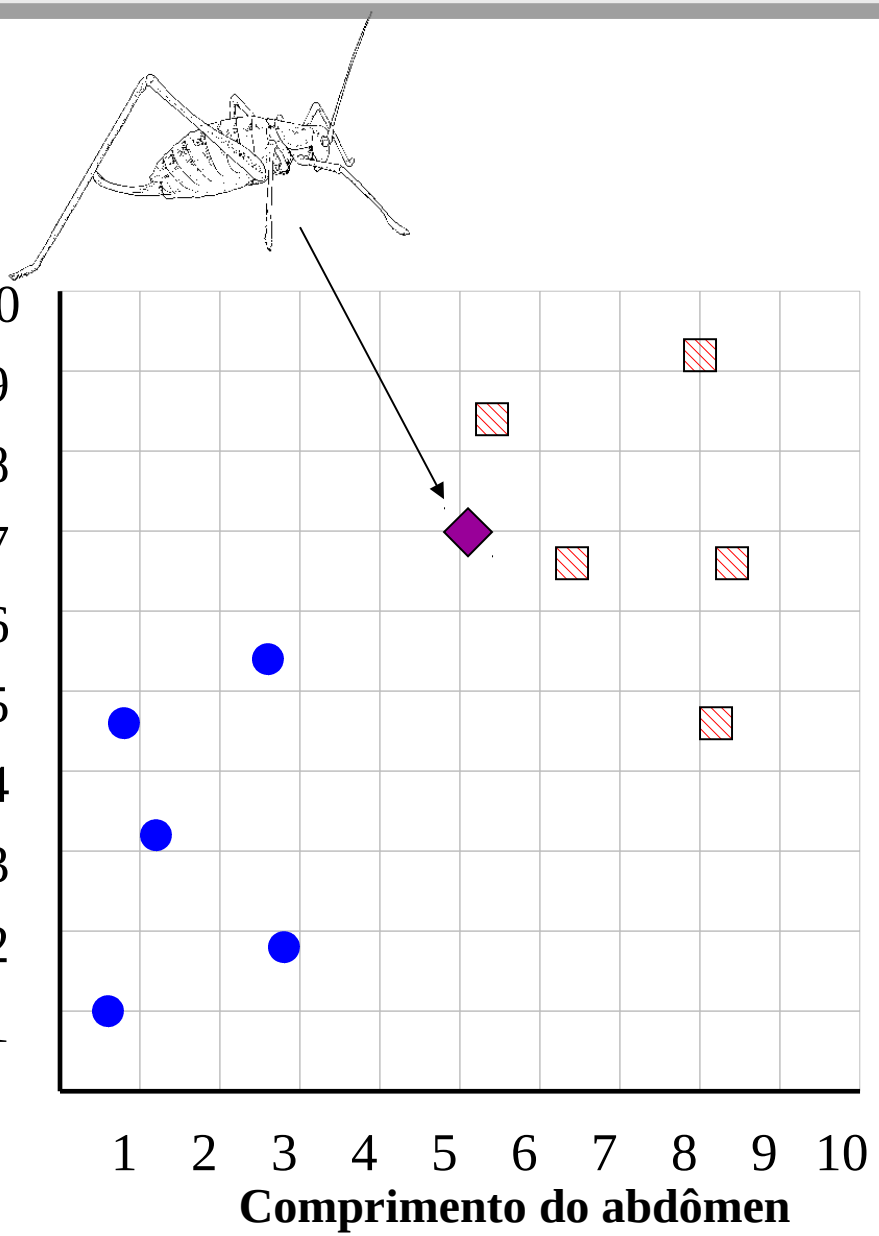






Exemplo não visto antes =

11	**5.1**	**7.0**	**??????**

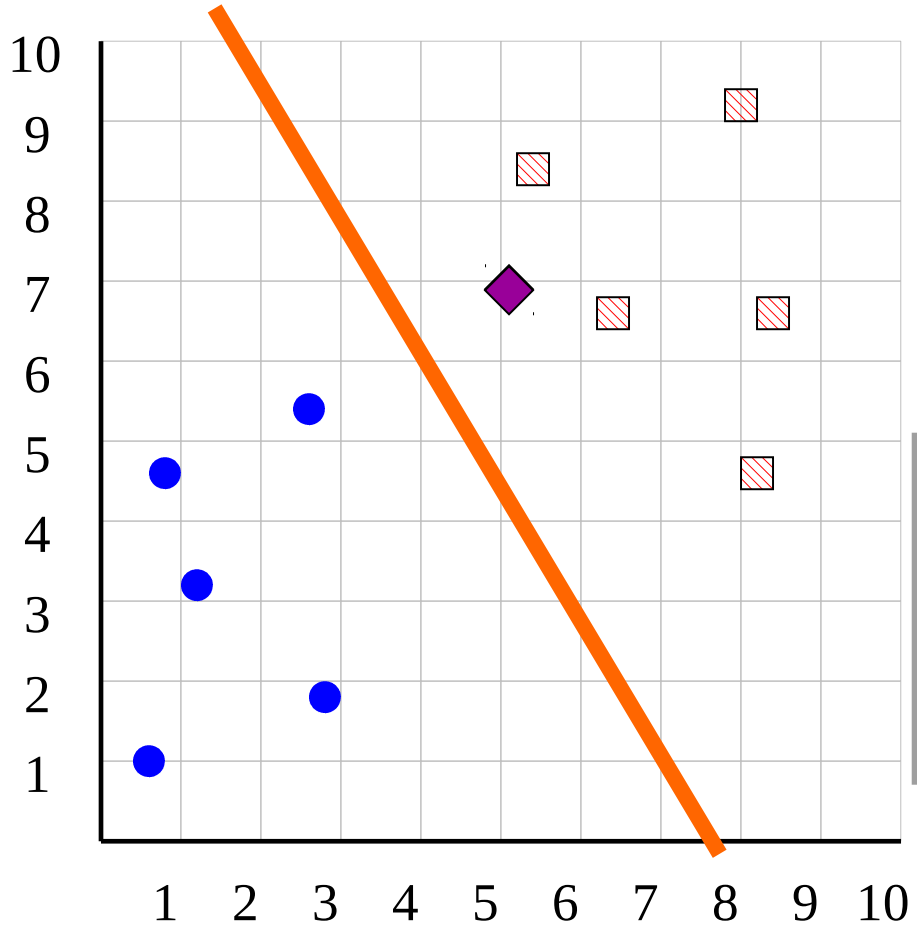


- Podemos “projetar” o exemplo não visto antes dentro do mesmo espaço que a base de dados.
- Acabamos de abstrair os detalhes do nosso problema particular. Será muito mais fácil conversar sobre pontos no espaço.

▨ **Esperança**

• **Gafanhoto**

Classificador Linear Simples

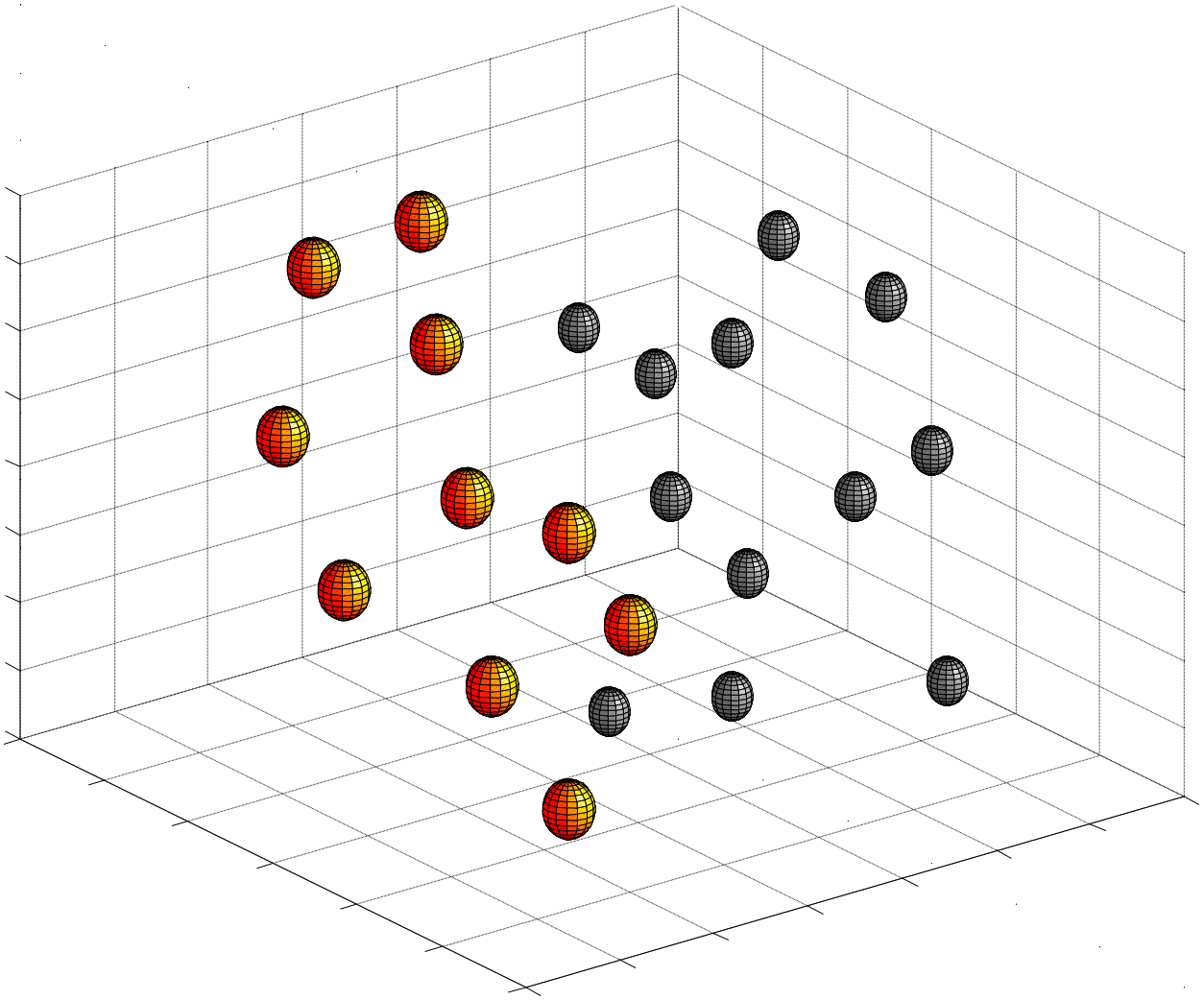


R.A. Fisher
1890-1962

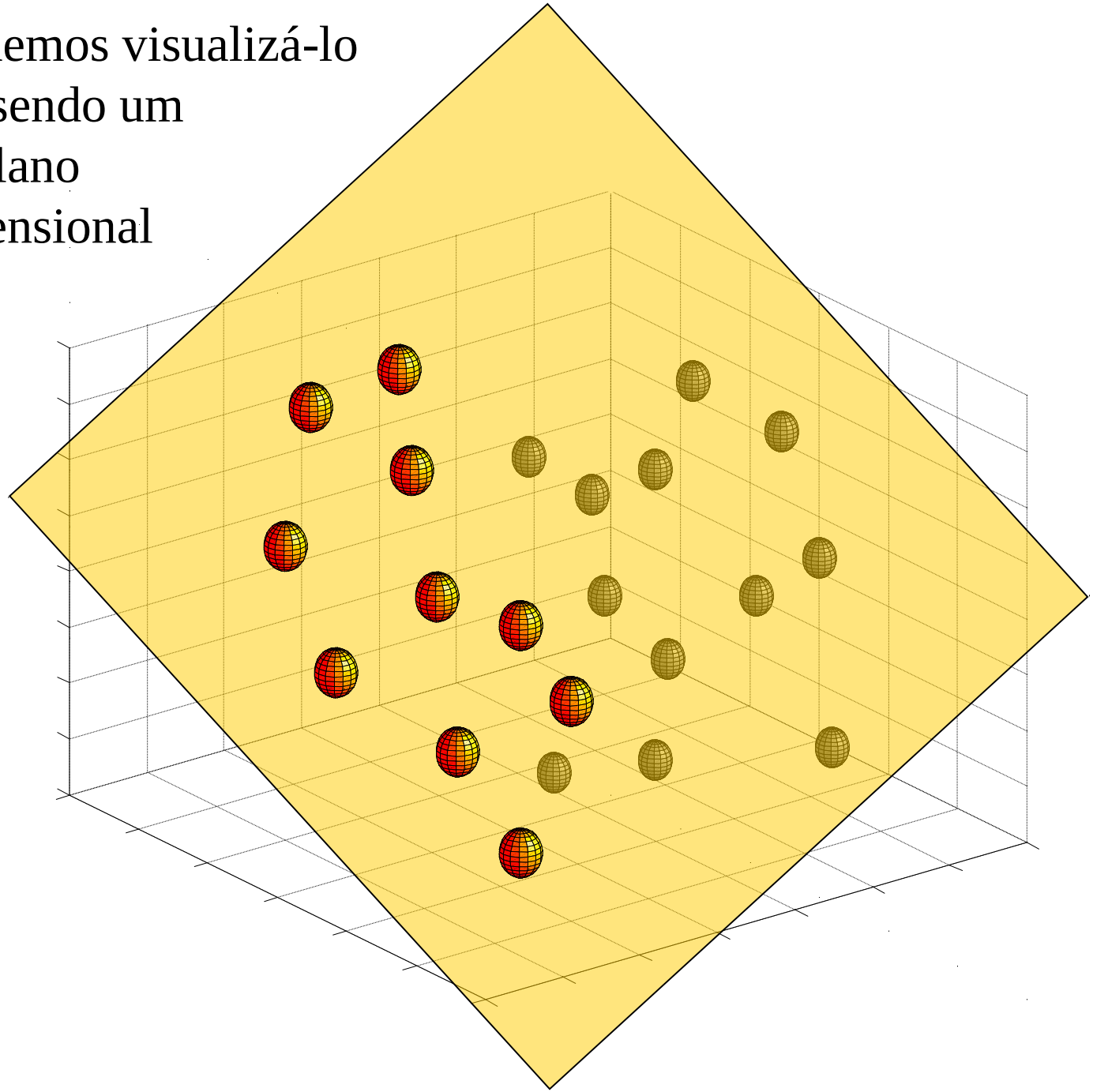
Se exemplo não visto antes está acima da linha
Então
classe é **Esperança**
senão
classe é **Gafanhoto**

Esperança
Gafanhoto

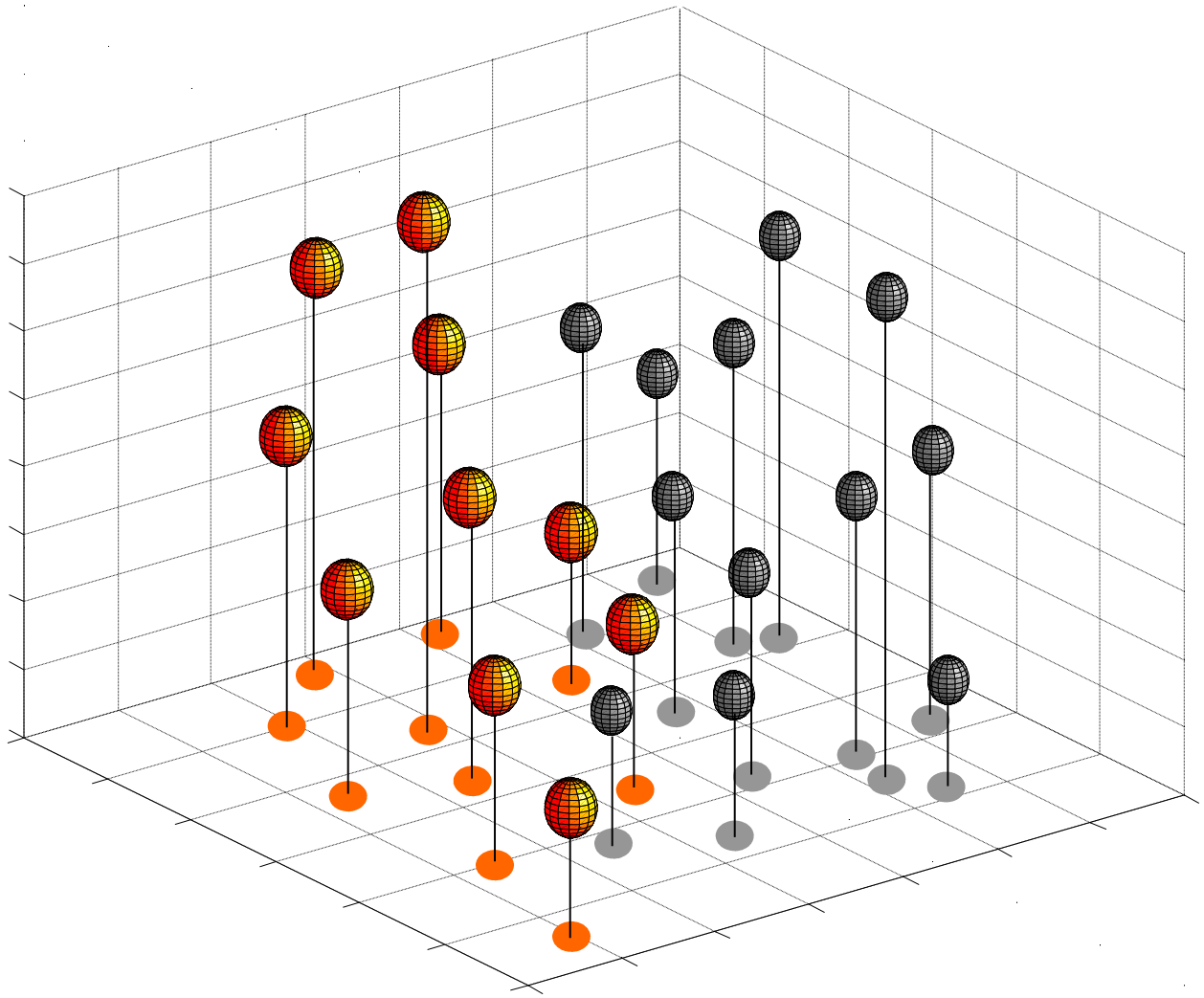
O classificador linear simples é definido para espaços dimensionais maiores…

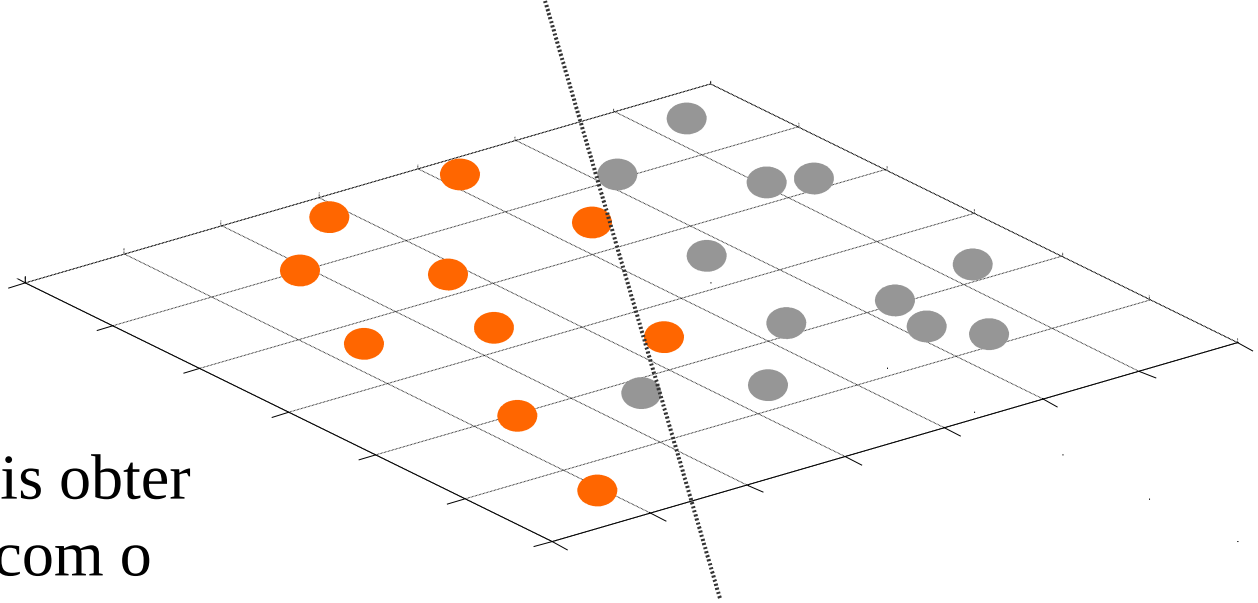


… podemos visualizá-lo como sendo um hiperplano n-dimensional



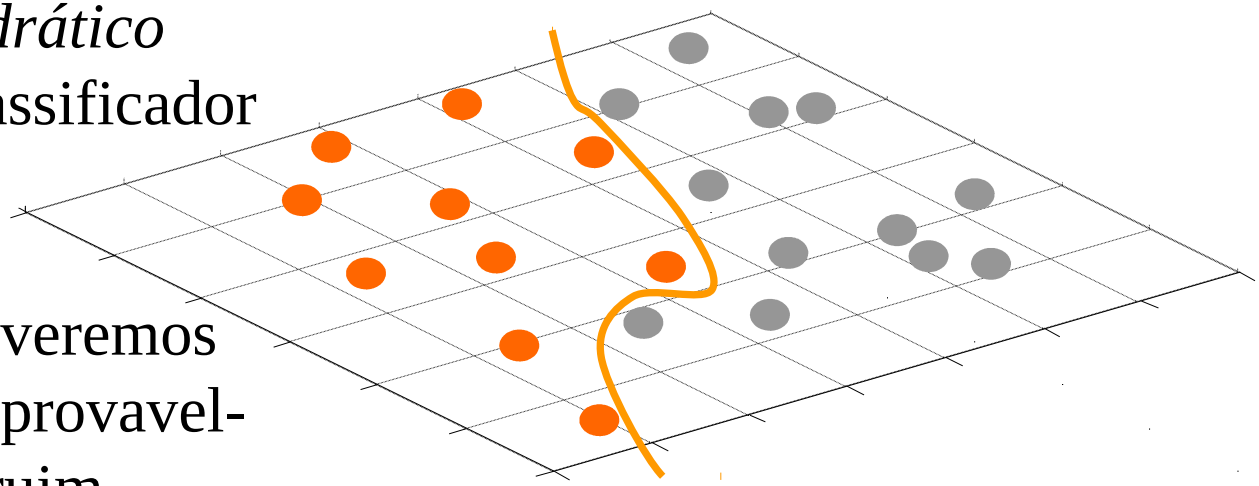
É interessante pensar no que aconteceria neste exemplo se não tivéssemos a terceira dimensão…





Não podemos mais obter acurácia perfeita com o classificador linear simples…

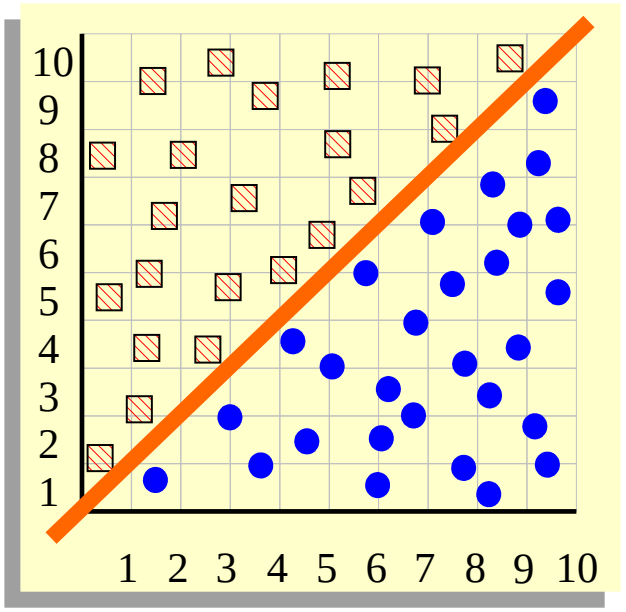
Podemos tentar resolver este problema usando um classificador *quadrático* simples ou um classificador *cúbico* simples…

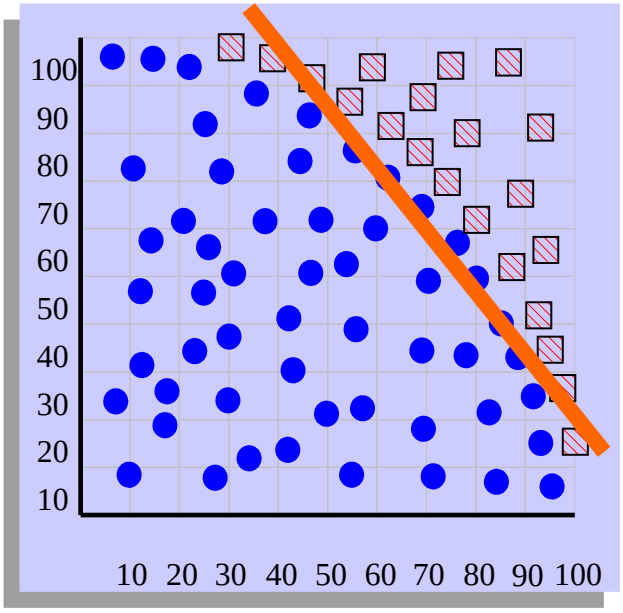


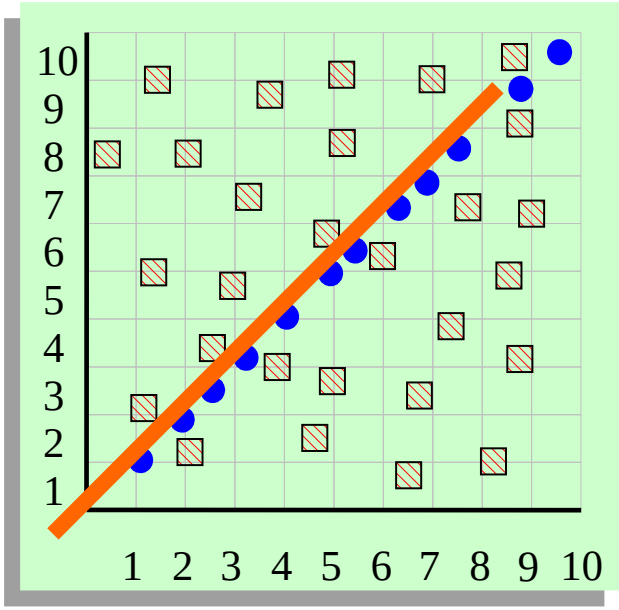
Entretanto, como veremos mais tarde, esta é provavel-mente uma idéia ruim…

Quais dos “Problemas do Pombo” podem ser resolvidos pelo Classificador Linear Simples?

1) Perfeito
2) Inútil
3) Muito bom







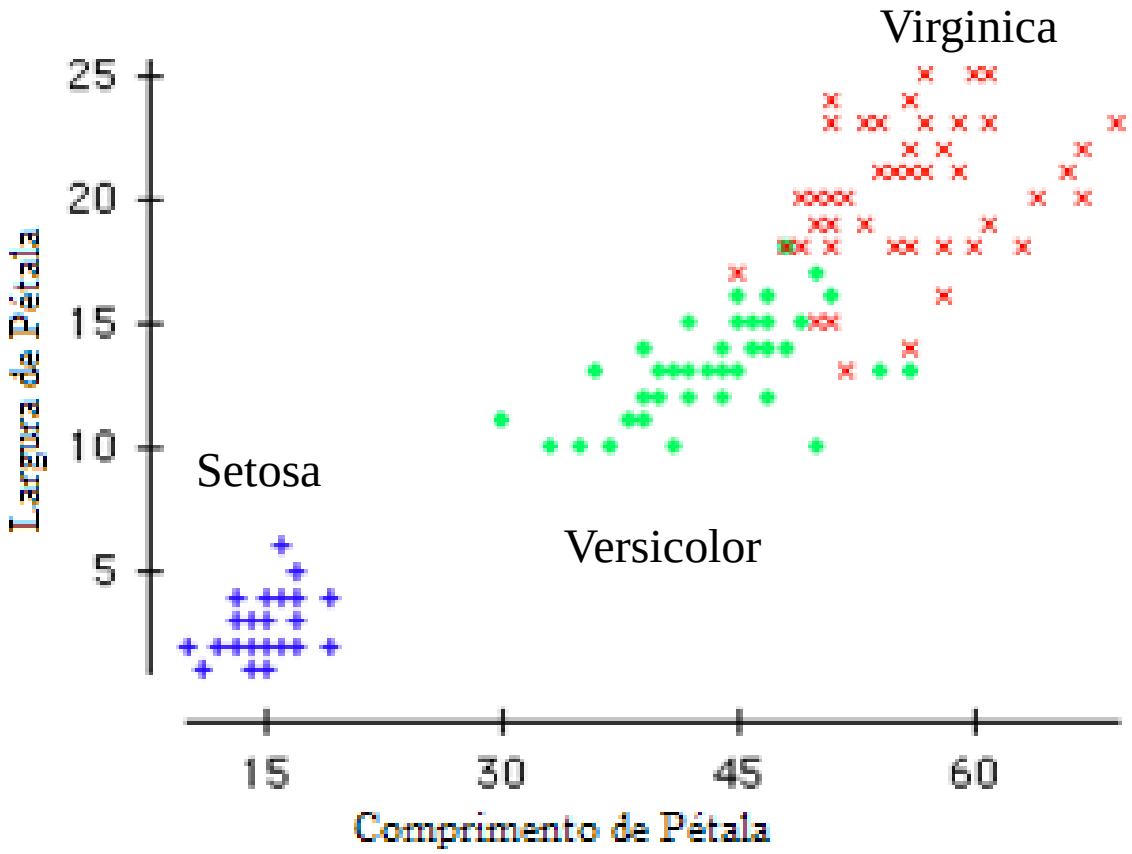
Problemas que podem ser resolvidos por um classificador linear são chamados de **linearmente separáveis**.

Um problema famoso

R. A. Fisher’s Iris Dataset.

- 3 classes
- 50 exemplos de cada classe

A tarefa é classificar as plantas em uma das 3 variedades usando comprimento de pétala e largura de pétala.

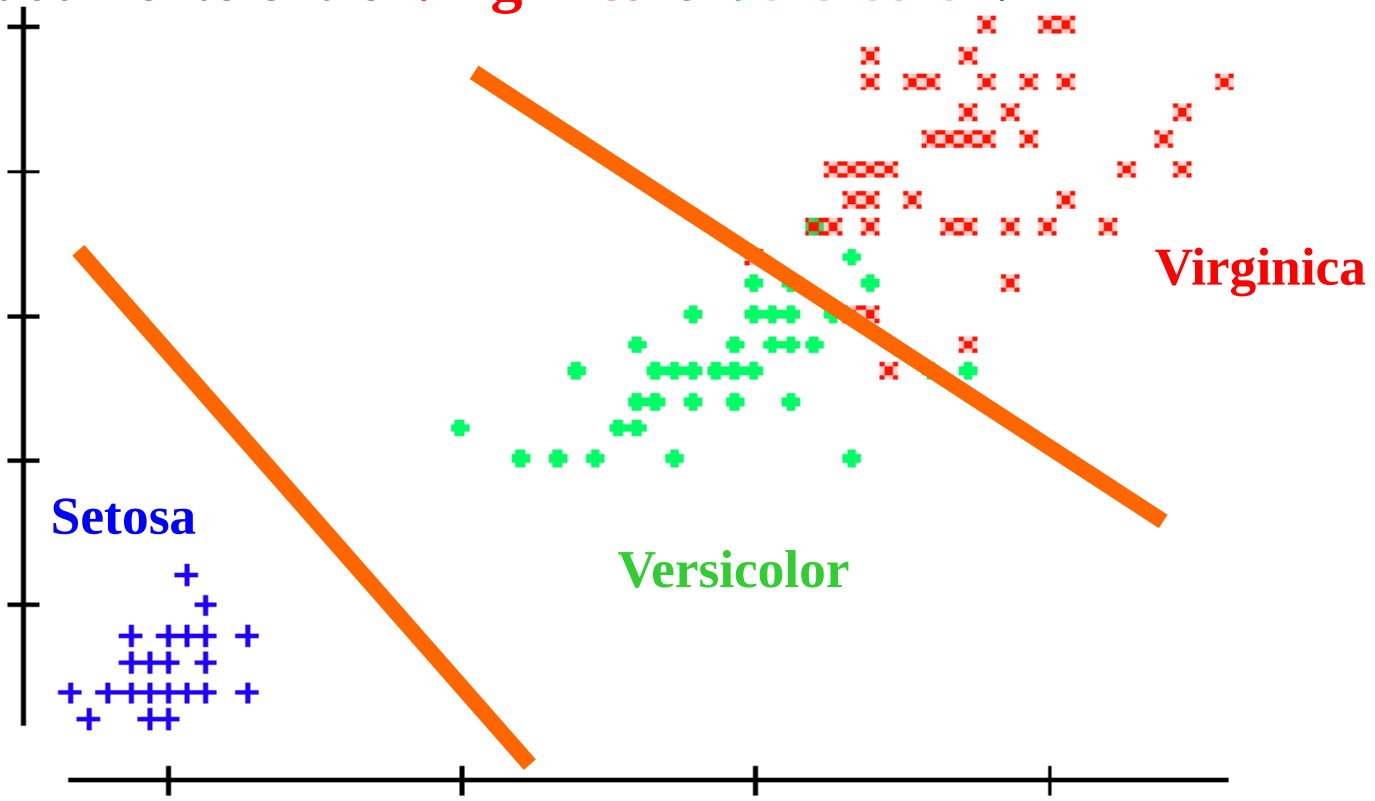


Iris Setosa

Iris Versicolor

Iris Virginica

Podemos generalizar o classificador linear relativo a variáveis a *C* classes, combinando *C*-1 linhas. Neste caso primeiramente aprendemos a linha para (perfeitamente) discriminar entre **Setosa** e **Virginica**/**Versicolor**, então aprendemos a discriminar aproximadamente entre **Virginica** **e** **Versicolor**.



Se comp. de pétala > 3.272 – (0.325 * comp. de pétala)
Então classe = **Virginica** **Senão Se** largura de pétala…