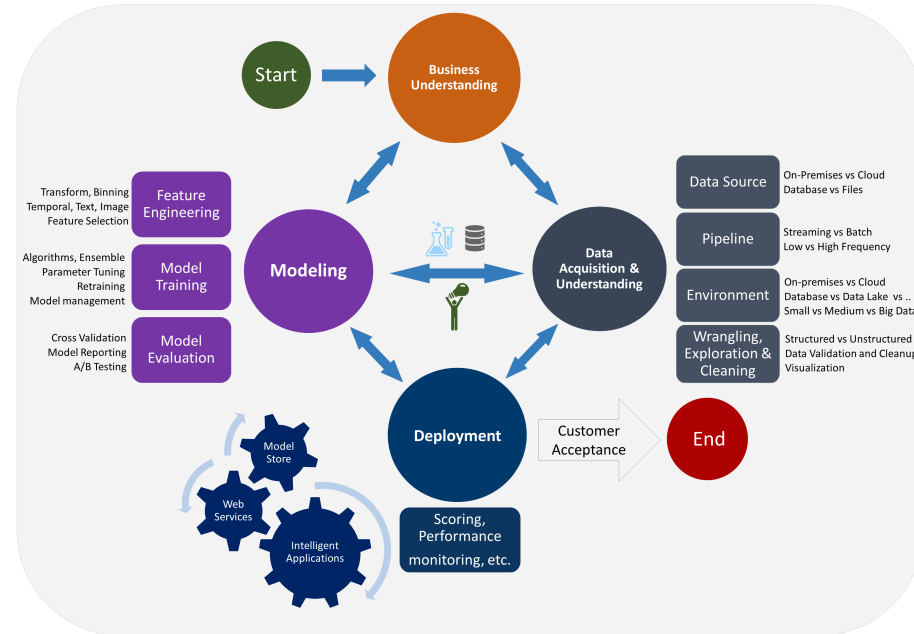


Classificação - Introdução

Mineração de Dados

Ronaldo C. Prati

Data Science Lifecycle



Exemplos de problemas

Perda de clientes: o consumidor irá desistir do serviço?

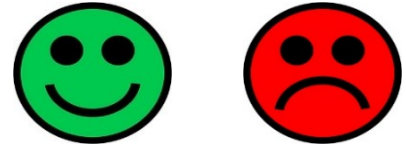


Sistema de relacionamento
Com o cliente (CRM System)
Dados sobre o cliente

- Demograficos
- Comportamento
- Vendas



Modelo



- **Previsão de desistência**
- Probabilidade de Insatisfação
- Propensão de compra/oferta
- Gerenciamento de Campanha
- Segmentação de clientes
- ...

Segmentação de Clientes: qual o grupo cada cliente pertence?



Sistema de relacionamento
Com o cliente (CRM System)
Dados sobre o cliente

- Demograficos
- Comportamento
- Vendas



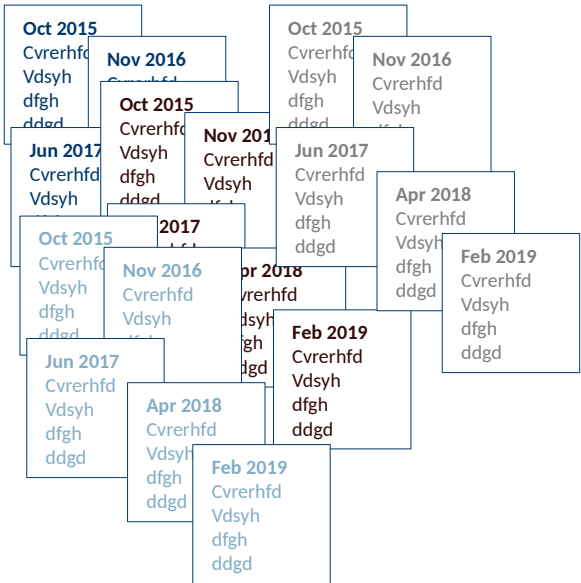
Modelo



- Previsão de desistência
- Probabilidade de Insatisfação
- Propensão de compra/oferta
- Gerenciamento de Campanha
- Segmentação de clientes
- ...

Análise de risco: esse cliente em potencial vai pagar o financiamento?

Histórico do cliente



Modelo



Prognóstico de risco

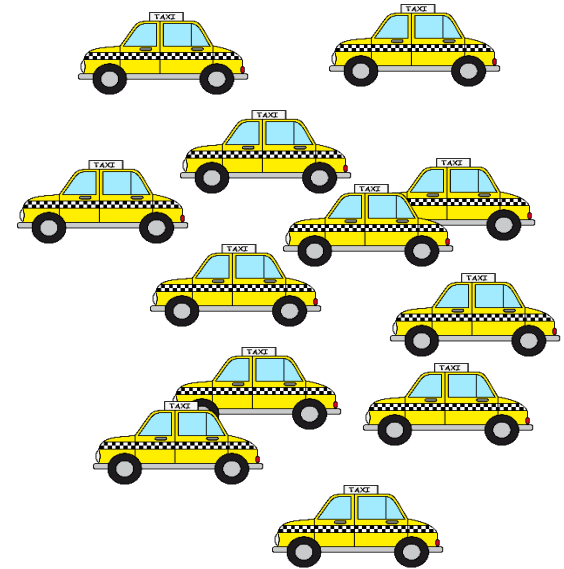
- High Risk
- Low Risk
- High Risk
- Very High Risk
- Very Low Risk
- Medium Risk
- ...

First Name	Last Name
DOB	Gender
dd/MM/yy	
Employment	Credit Risk
	High



Predição de demanda: quantos vai querer o meu serviço/produto?

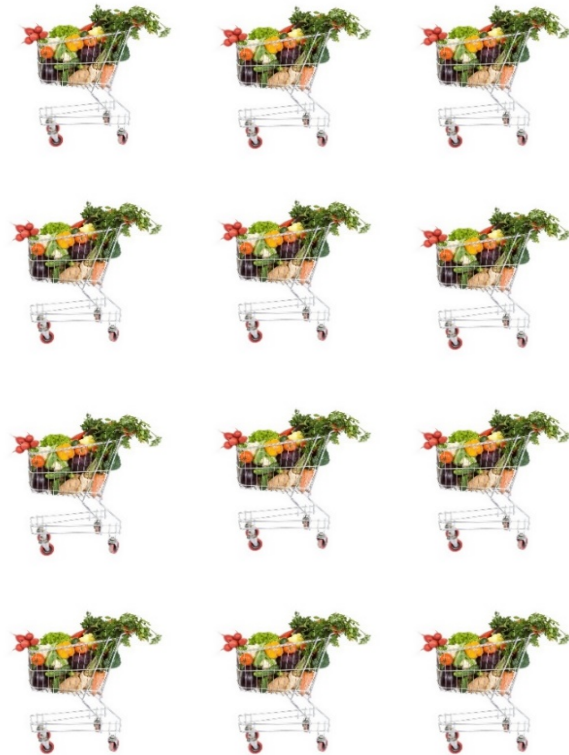
- Quantos carros de aplicativo eu preciso em uma cidade, em uma quarta-feira a noite?
- Quanta energia a cidade irá consumir no próximo mês?
- Quantos cliente irão frequentar meu restaurante hoje?



Modelo

Serviços de recomendação

Sistema de recomendação: pessoas que compraram um item estão interessados em quais outros itens?



Modelo



Recomendação



Deteção de fraude: essa transação é legítima ou fradulenta?



Transações

- Trx 1
- Trx 2
- Trx 3
- Trx 4
- Trx 5
- Trx 6
- ...



Modelo

Transação expúria

[illegible]

Deteção de fraude: essa transação é legítima ou fradulenta?



Transações

- Trx 1
- Trx 2
- Trx 3
- Trx 4
- Trx 5
- Trx 6
- ...



Modelo

Transação expúria

[illegible]

Análise de sentimento: como posso saber a opinião das pessoas/clientes?



Samsung

Samsung Galaxy S7 Edge G935A 32GB Unlocked - Gold Platinum



125 customer reviews | 606 answered questions



Beautiful phone from a wonderful seller!

By

on May 29, 2017

Color: Gold | **Verified Purchase**

This practically new beautiful phone well exceeded my expectations!



One Star

By

on August 3, 2016

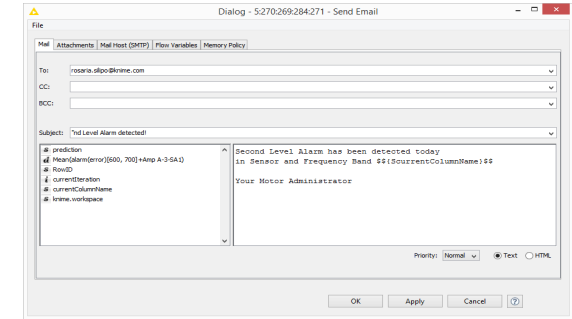
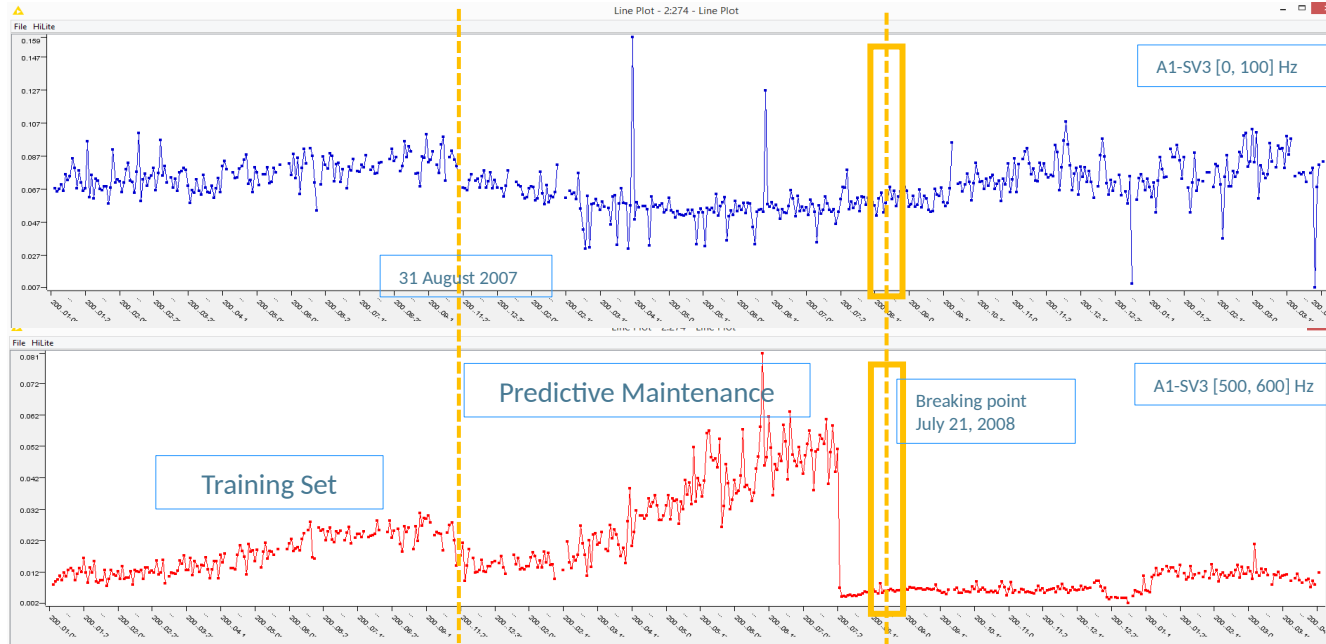
Color: Black Onyx | **Verified Purchase**

Very bad experience

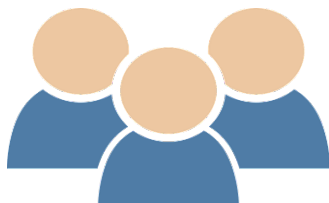


Deteção de anomalia

Prever uma falha mecânica o mais tarde possível, mas antes que ela ocorra



Are there other compounds having this substructure and being a dopaminergic antagonist?

c1ccc2c(c1)nc3ccccc3s2

Enter Search Options

Enter a SMILES to search ChEBI for compounds (substructure based search):

N1C2=CC=CC=C2SC2=CC=CC=C2

Type in a role of your interest and select using the autocomplete function a valid term:

dopamine

dopamine receptor D2 antagonist
EC 1.14.17.1 (dopamine beta-monooxygenase) inhibitor
dopaminergic agent
dopaminergic antagonist
dopamine uptake inhibitor
dopamine agonist

Reset

Apply

Close



Selected Compound

promazine

CN1C=CC2=C(C=C1)N(C)C3=CC=CC=C3C2

chembl: C1709292

charge: 1

molecularweight: 234.135

unprot_weight: 3.1745

definition: A phenothiazine derivative in which the phenothiazine moiety has a 5-(dimethylamino)propyl group at the 10-position.

database_cross_references: CHEBI:142230; DrugBank:000025; KEGG:R000001; LINC01584; LINC01577; PDB:105046; Pubmed:00070566; Reaxys:14601; Refchem:J0450; Wikipedia:Promazine; WDI:Q17775; Pubmed:135954; ChEMBL_C224; CAS:56-40-2; PDB:2BQ205; KEGG:R00040

Showing 1 of 1 entries

Selected compound has following roles assigned:

ChemBL_LABEL	KEGG_LABEL	parent_LABEL	definition, KEGG_LABEL
phenazine	has role	EC 3.4.21.35 (serp) epinephrine inhibitor	Any EC 3.4.21.35 (serp epinephrine) inhibitor that interferes with the action of prolactinase (EC 3.4.21.26).
phenazine	has role	5-HT receptor antagonist	5-HT receptor antagonists are the drugs that selectively bind to but do not activate serotonin (5) receptors, thereby blocking the actions of endogenous histamine.
phenazine	has role	antipsychotic	A drug used to prevent psychosis or, similarly, an antipsychotic may act to a wide range of mechanisms: it might affect the mesolimbic control system (the vomiting centre and the chemoreceptor trigger zone) or affect the peripheral receptors.
phenazine	has role	dopaminergic antagonist	A drug that blocks the role of dopamine dopamine receptors, thereby blocking the actions of dopamine or endogenous agonists.
phenazine	has role	muscarinic antagonist	A drug that binds to but does not activate muscarinic, cholinergic receptors, thereby blocking the actions of endogenous acetylcholine or exogenous agonists.

Network View

Search: 11

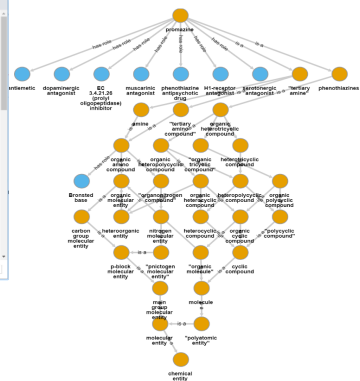
List of compounds matching your query: Select only one!

Select one compound to learn more about from this page (click on the ID)

Show: entries

carphenazine C1709292, "C1709292" (16)	chlorpromazine C1709292, "C1709292" (16)	fluphenazine C1709292, "C1709292" (16)	mesoridazine C1709292, "C1709292" (16)
mesoridazine isoxalide C1709292, "C1709292" (16)	methidimiprazine C1709292, "C1709292" (16)	perazine C1709292, "C1709292" (16)	perphenazine C1709292, "C1709292" (16)

Parent / Apps / Close



Tarefas de Mineração de Dados

Principais tarefas

— Classificação

— Predição de categorias:

- Predizer o diagnóstico de uma doença de acordo com exames (doente/não doente)
- *Qual a nota de crédito do cliente ? Ótima / boa / regular / ruim*
- *Esse cliente vai responder a esse anúncio? Sim / Não*

— Regressão

— Predizer valores numéricos

- *Como o preço da ação vai evoluir?*
- *Qual será o preço desse celular na próxima semana?*

— Agrupamento/Segmentação

- Agrupar registros similares para ter uma visão geral, detectar outliers, ou ter insights sobre a estrutura dos dados
 - *Meus clientes podem ser separados em grupos diferentes?*
 - *Quais genes são expressados em conjunto?*

— **Análise de Associações**

- Encontrar correlações para melhor entender as interdependências entre os atributos
 - Quais produtos um consumidor irá gostar, baseado em seu histórico de compras?
 - *Exista algum padrão de interações entre redes sociais?*

— **Análise de Desvio**

- Conhecendo a tendência dos dados, encontrar subgrupos que se comportam de maneira diferente
 - Quais circunstância o sistema se comporta diferente?
 - Quais as características de consumidores que não são influenciados pela moda tem em comum?

Aprendizado Supervisionado

- Objetivo é prever um atributo alvo, com base em outros
- Premissa: em adição ao conjunto de dados de observações, sabemos o valor do atributo alvo para um conjunto de exemplos

Classificação

- Atributo alvo é **categórico**
 - *Qual categoria o cliente pertence?*

Regressão

- Atributo alvo é **numérico**
 - *Quanto o cliente irá gastar no próximo mês?*

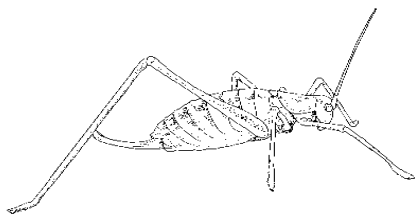
O problema de classificação

O problema de classificação

(definição informal)

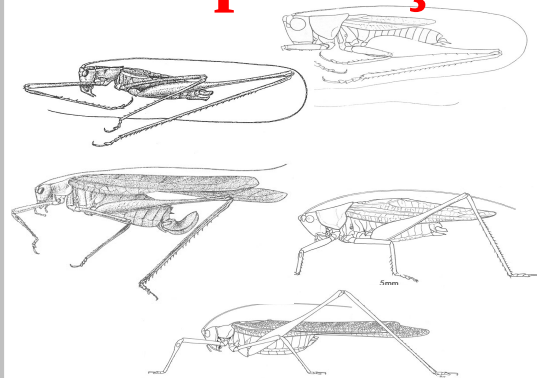
Dada uma coleção de dados detalhados, neste caso 5 exemplos de **Esperança** e 5 do **Gafanhoto**, decida a qual tipo de inseto o exemplo não rotulado pertence.

Obs: **Esperança** : tipo de gafanhoto verde.

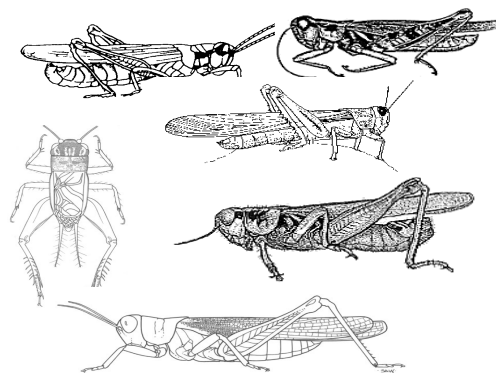


Esperança ou **Gafanhoto**?

Esperança



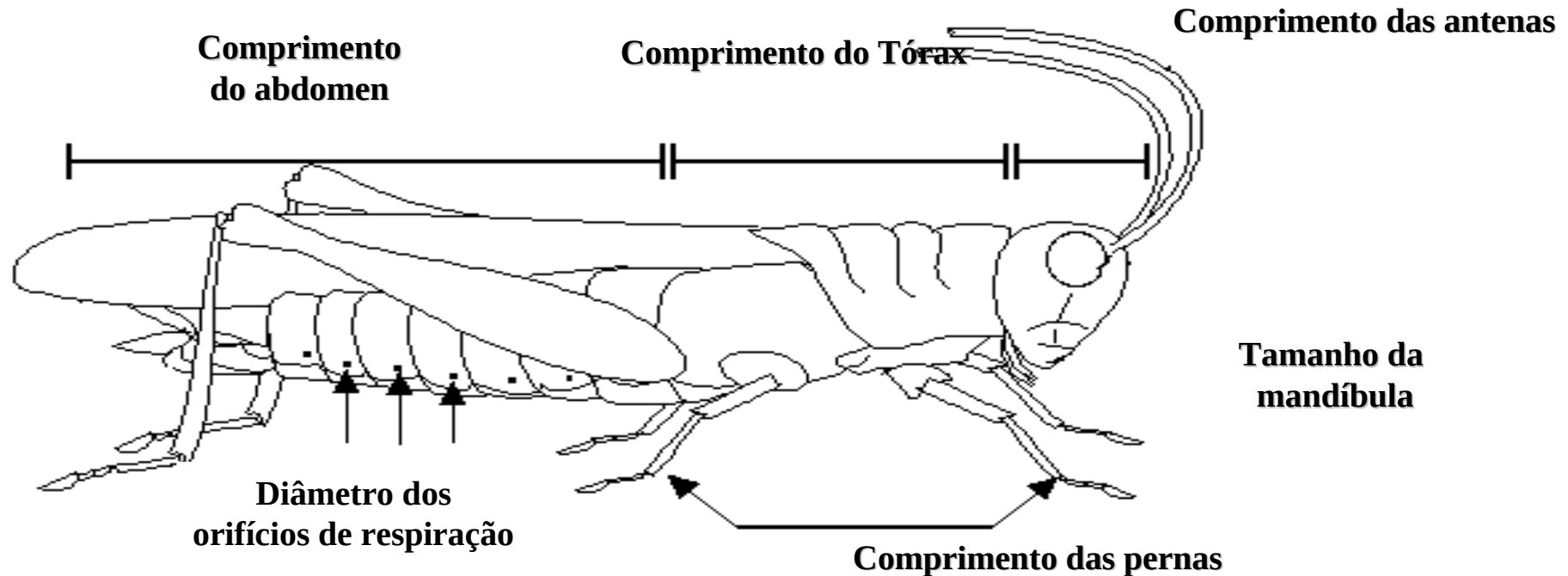
Gafanhoto



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 Base**), prediga o rótulo da **classe dos exemplos ainda não vistos**

Minha Base

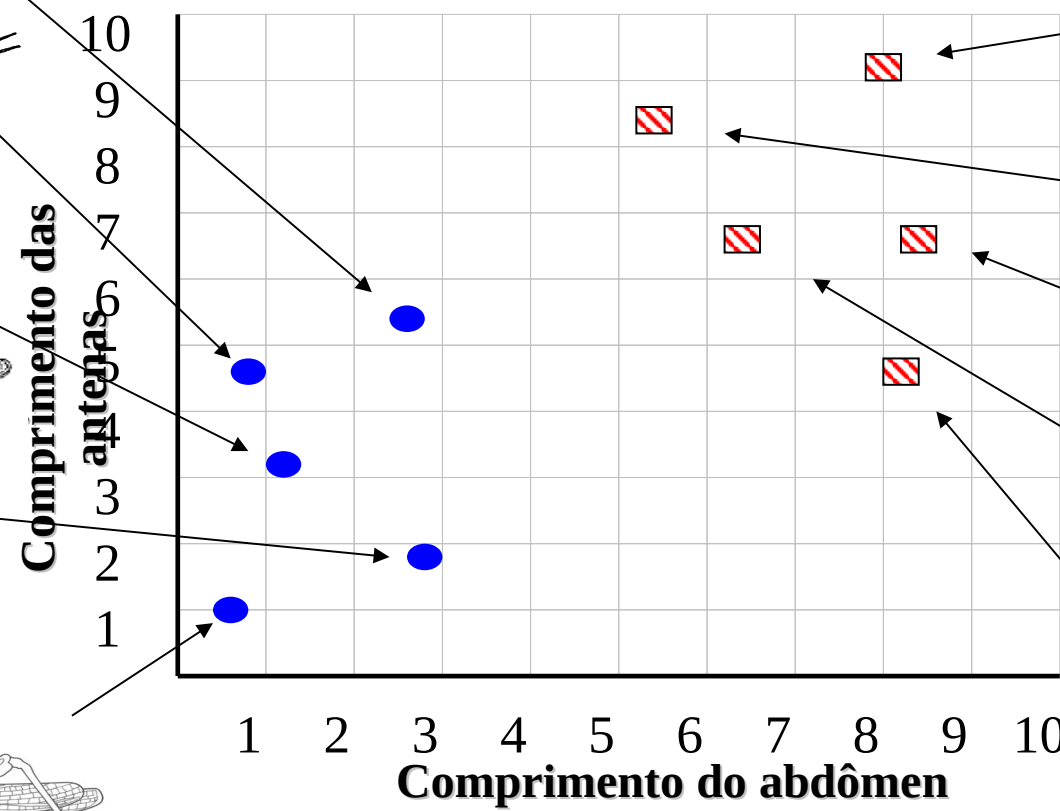
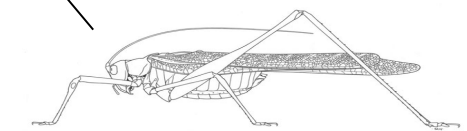
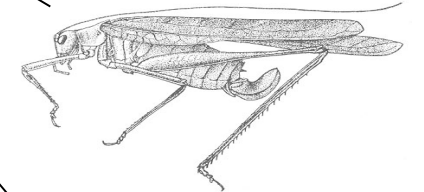
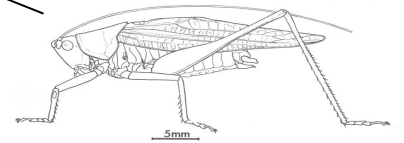
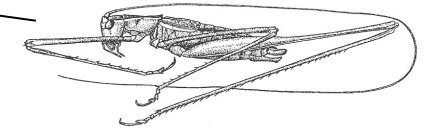
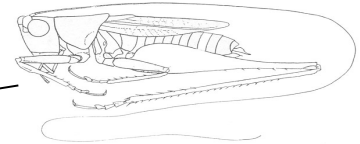
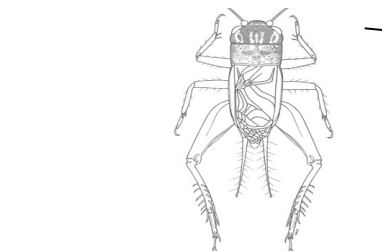
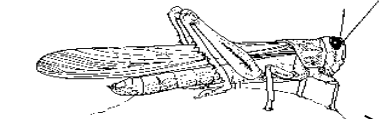
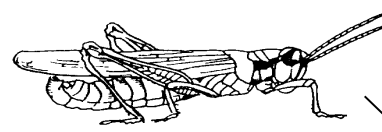
ID do inseto	Comp. do abdômen	Comp. das antenas	Classe do inseto
1	2.7	5.5	Gafanhoto
2	8.0	9.1	Esperança
3	0.9	4.7	Gafanhoto
4	1.1	3.1	Gafanhoto
5	5.4	8.5	Esperança
6	2.9	1.9	Gafanhoto
7	6.1	6.6	Esperança
8	0.5	1.0	Gafanhoto
9	8.3	6.6	Esperança
10	8.1	4.7	Esperança

Exemplo não visto =

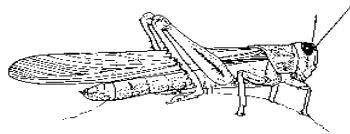
11	5.1	7.0	???????
----	-----	-----	---------

Gafanhoto

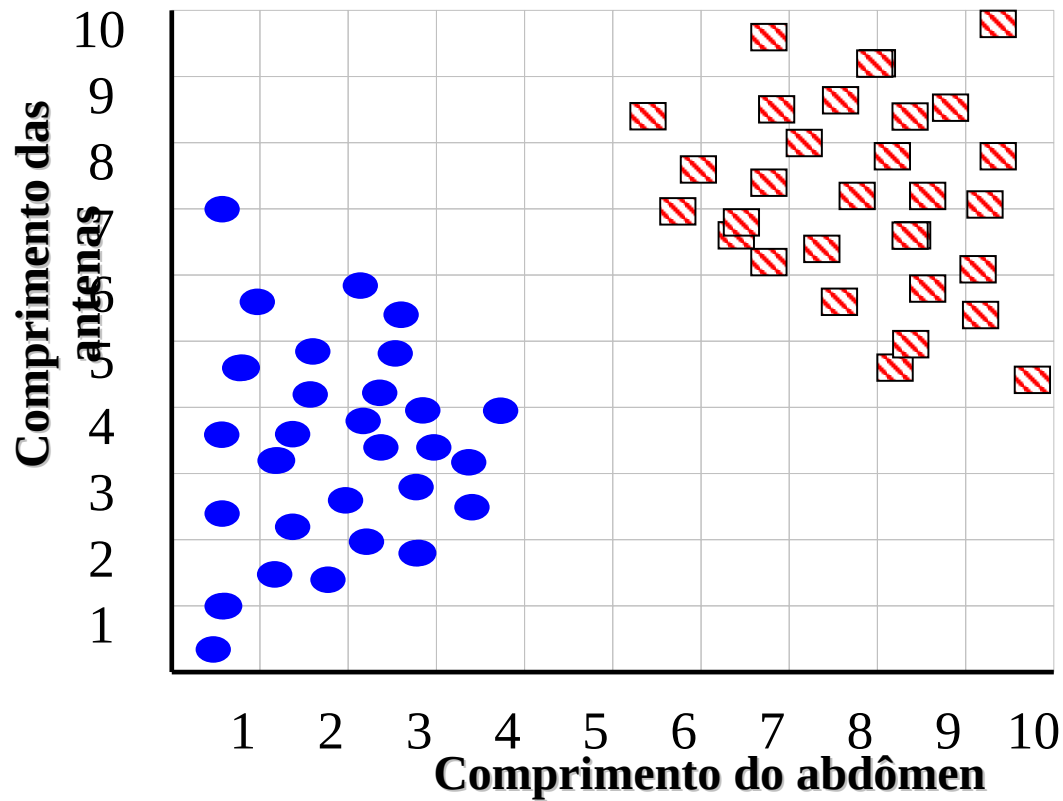
Esperança



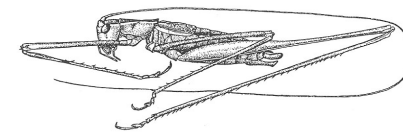
Gafanhoto



Também utilizaremos esta base de dados motivar os conceitos...



Esperança



Cada um destes objetos de dados é chamado de...

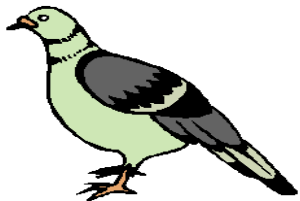
- exemplar
- exemplo (de treinamento)
- instância
- tupla



*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



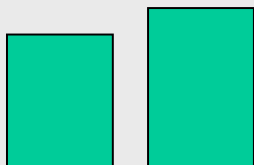
3

4



1.5

5



6

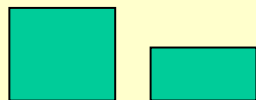
8



2.5

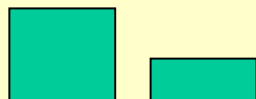
5

Exemplos da classe B



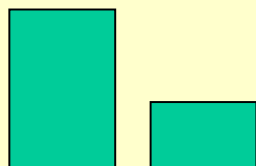
5

2.5



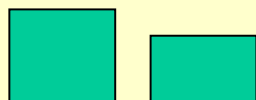
5

2



8

3



4.5

3

Problema do Pombo 1

Exemplos da classe A



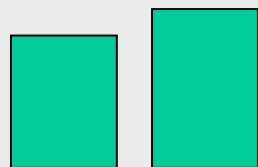
3

4



1.5

5



6

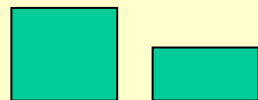
8



2.5

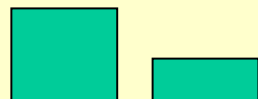
5

Exemplos da classe B



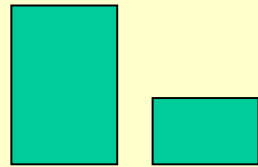
5

2.5



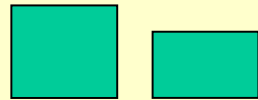
5

2



8

3

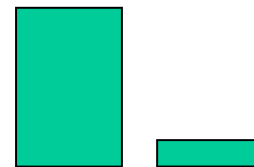


4.5

3



De qual classe é este objeto?

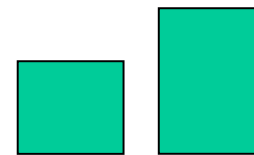


8

1.5



Que tal este, *A* ou *B*?

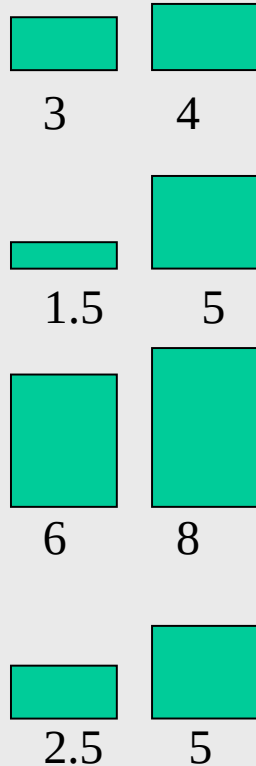


4.5

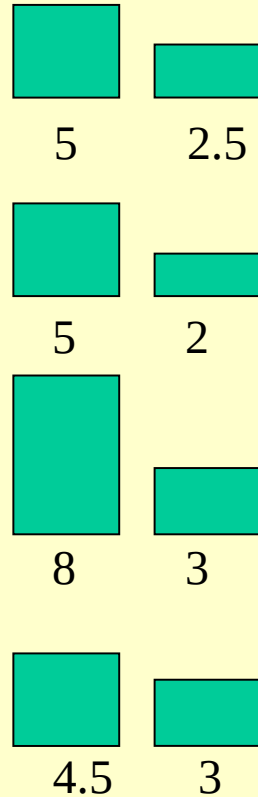
7

Problema do Pombo 1

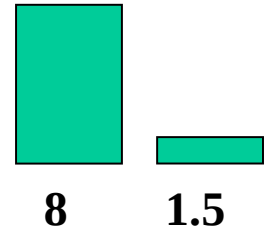
Exemplos da classe A



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



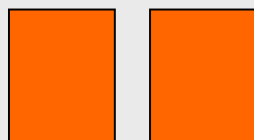
4

4



5

5



6

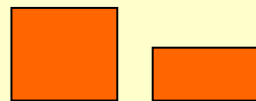
6



3

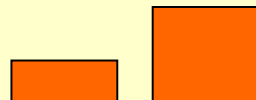
3

Exemplos da classe B



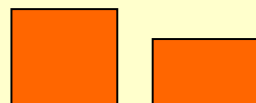
5

2.5



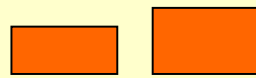
2

5



5

3

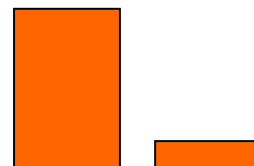


2.5

3



Oh! Este aqui é difícil!

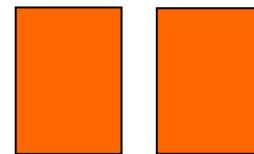


8

1.5



Até eu sei este!



7

7

Problema do Pombo 2

Exemplos da classe A



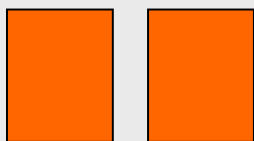
4

4



5

5



6

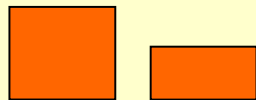
6



3

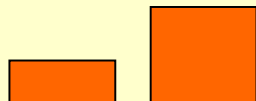
3

Exemplos da classe B



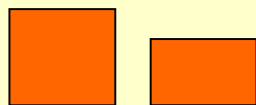
5

2.5



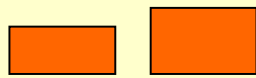
2

5



5

3

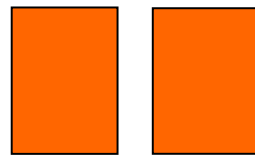


2.5

3

A regra é: se duas barras são iguais em tamanho é um $\mathcal{A}$. Caso contrário é um $\mathcal{B}$.

Então este é um $\mathcal{A}$.



7

7

Problema do Pombo 3

Exemplos da classe A



4

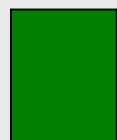
4



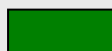
1



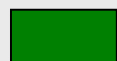
5



6



3

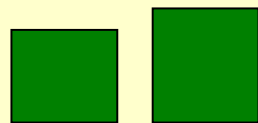


3



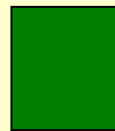
7

Exemplos da classe B



5

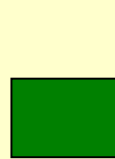
6



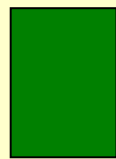
7



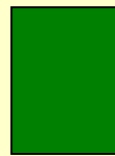
5



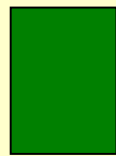
4



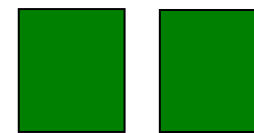
8



7



7



6

6

Este é muito difícil!

*Qual é este, **A** ou **B**?*

Problema do Pombo 3

Exemplos da classe A



4

4



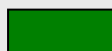
1



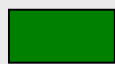
5



6



3

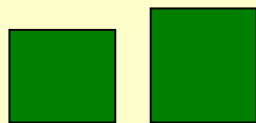


3



7

Exemplos da classe B



5



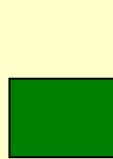
6



7



5



4



8

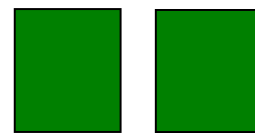


7



7

É um **B**!



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**.



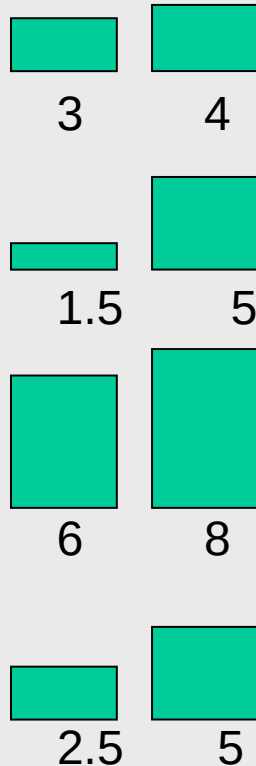
Por que gastamos tanto tempo com este joguinho?

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

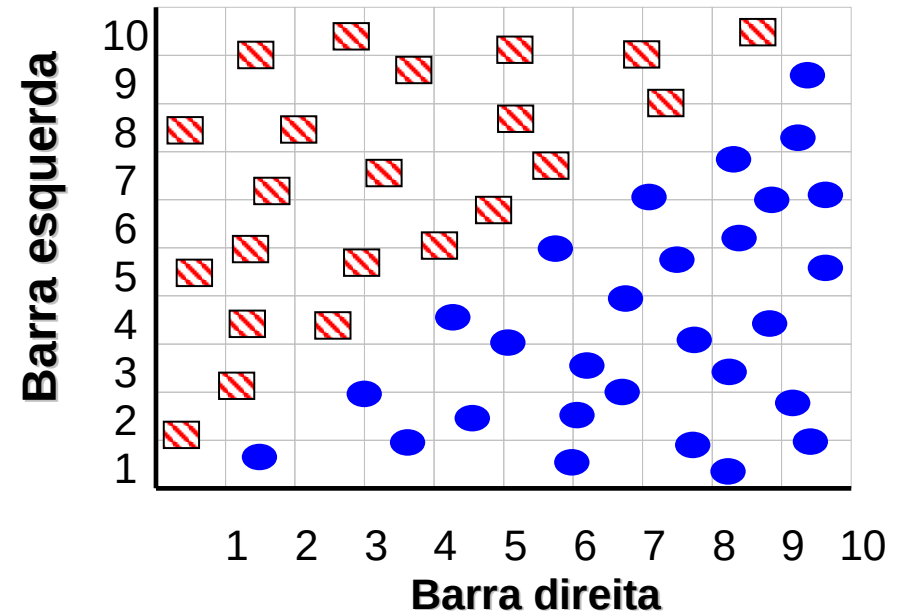
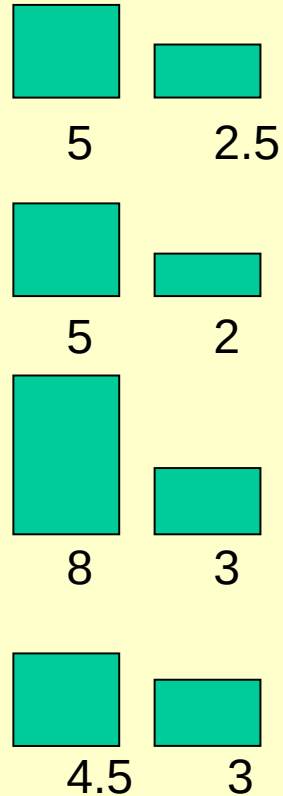


Problema do Pombo 1

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



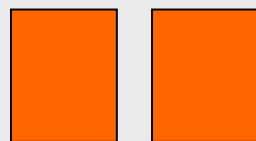
4

4



5

5



6

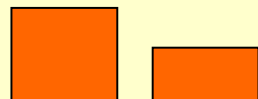
6



3

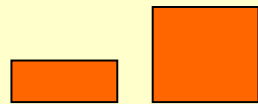
3

Exemplos da classe B



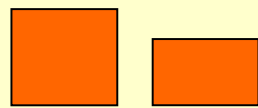
5

2.5



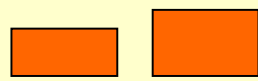
2

5



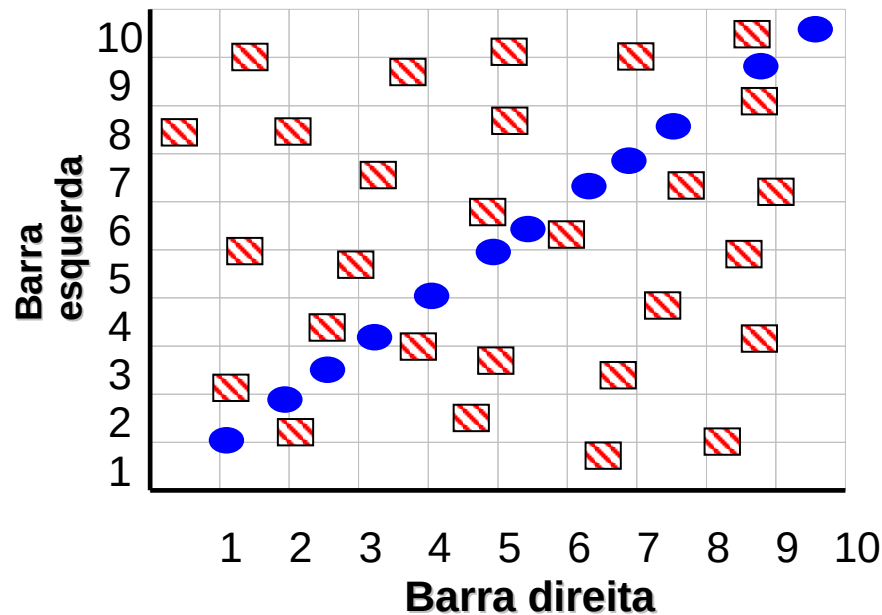
5

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

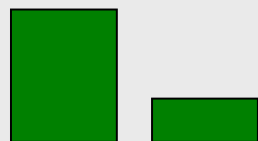
Exemplos da classe A



4 4



1 5



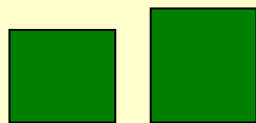
6

3

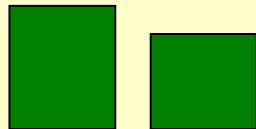


3 7

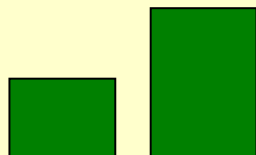
Exemplos da classe B



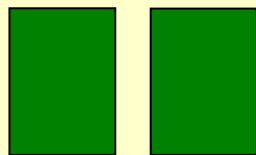
5 6



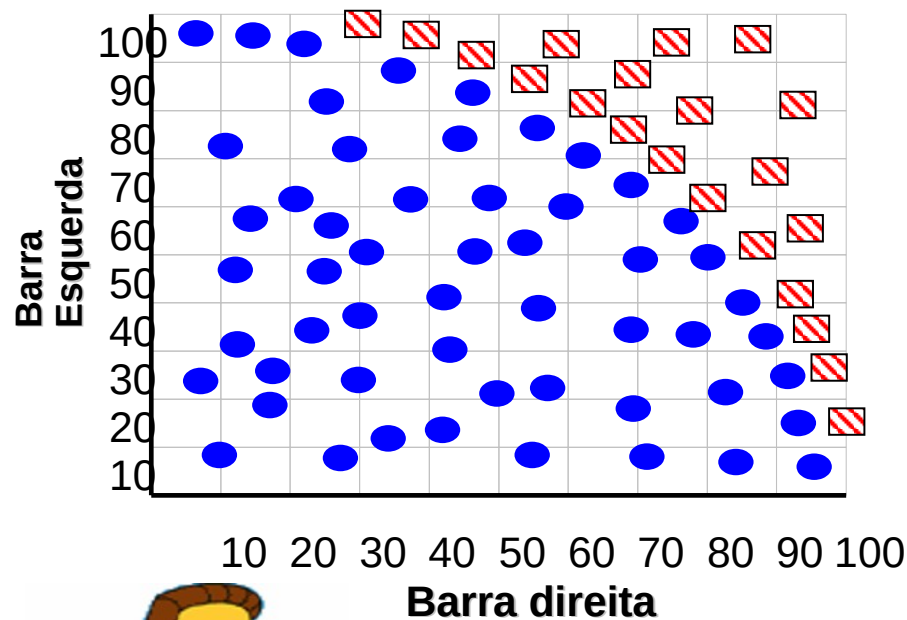
7 5



4 8



7 7

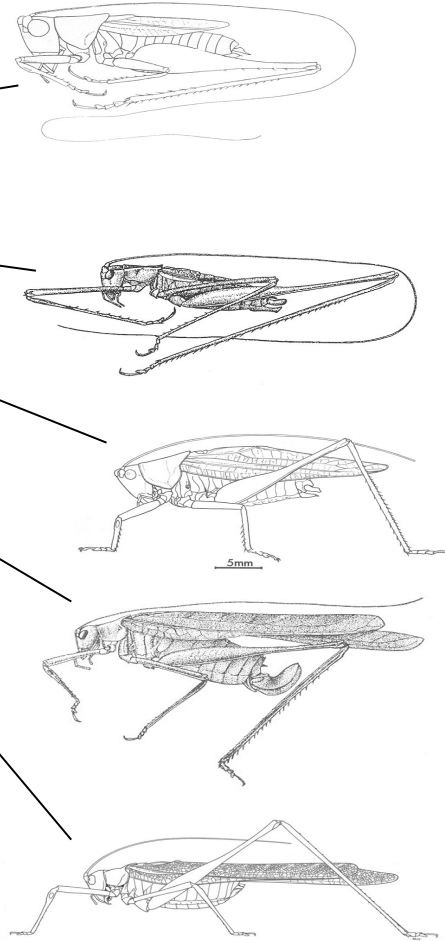
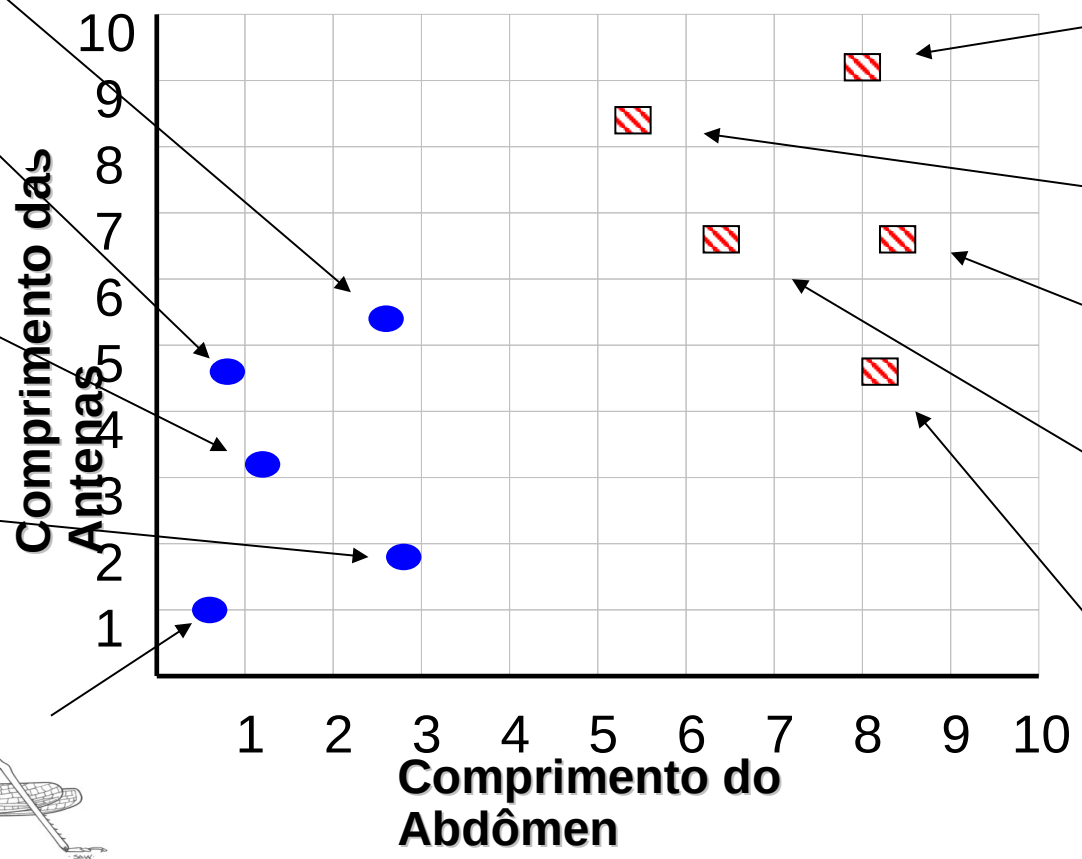
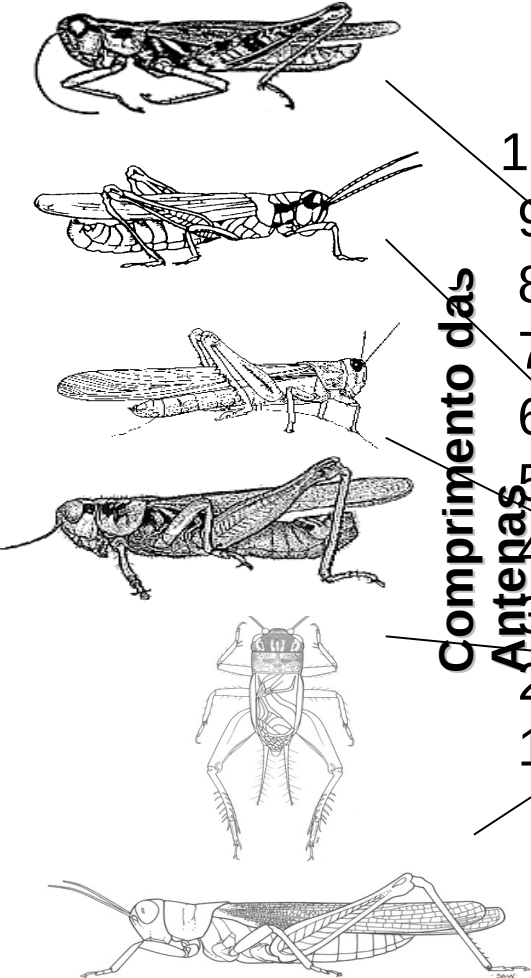


A regra novamente:

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

Gafanhoto

Esperança



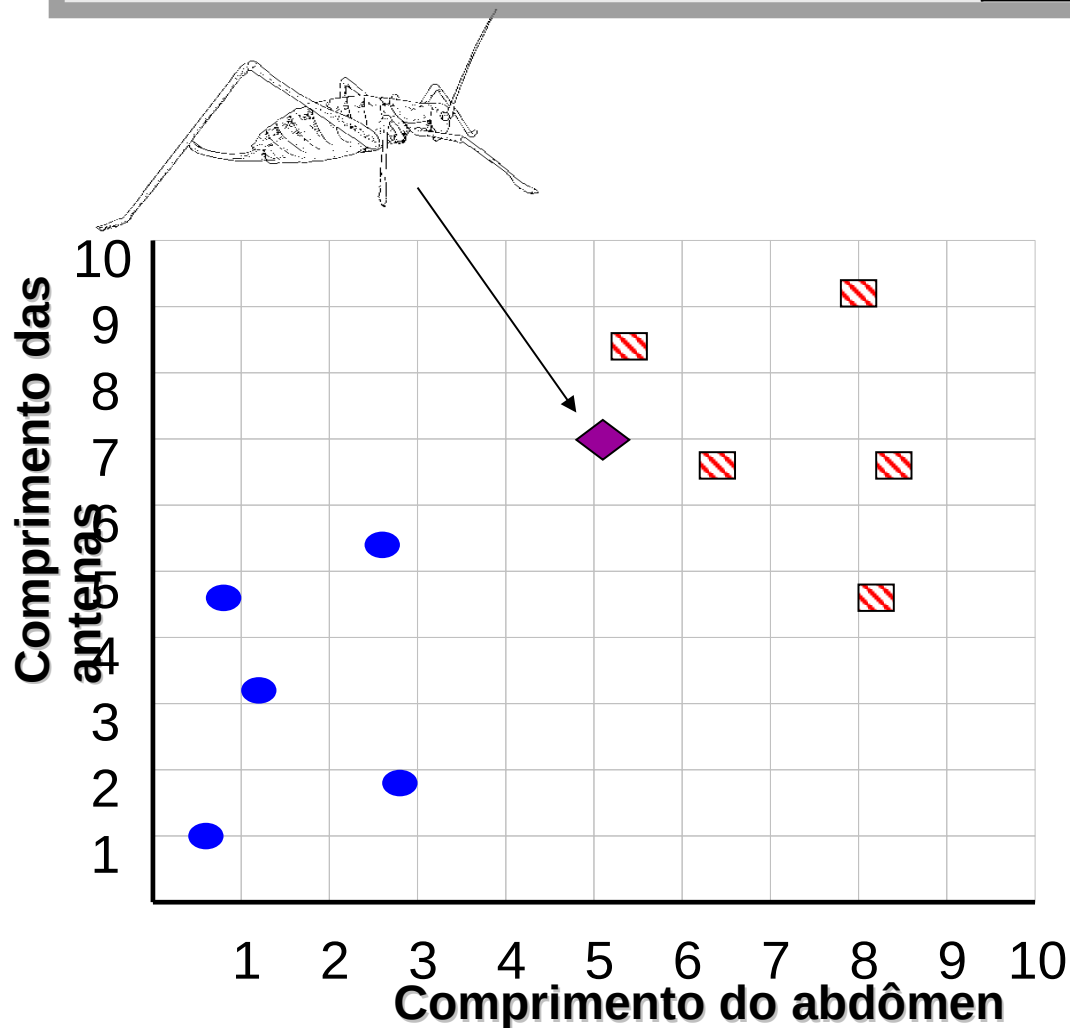
Exemplo não visto antes =

11

5.1

7.0

???????



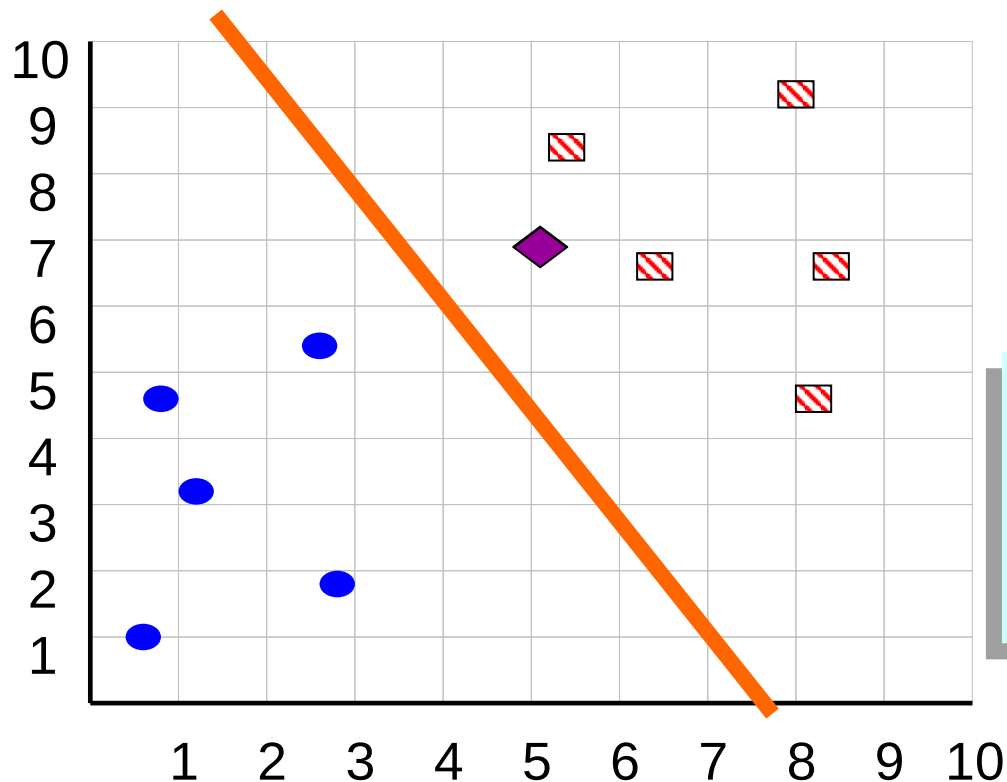
- Podemos “projetar” o **exemplo não visto antes** dentro do mesmo espaço que a base de dados.
- Acabamos de abstrair os detalhes do nosso problema particular. Será muito mais fácil conversar sobre pontos no espaço.

■ **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**

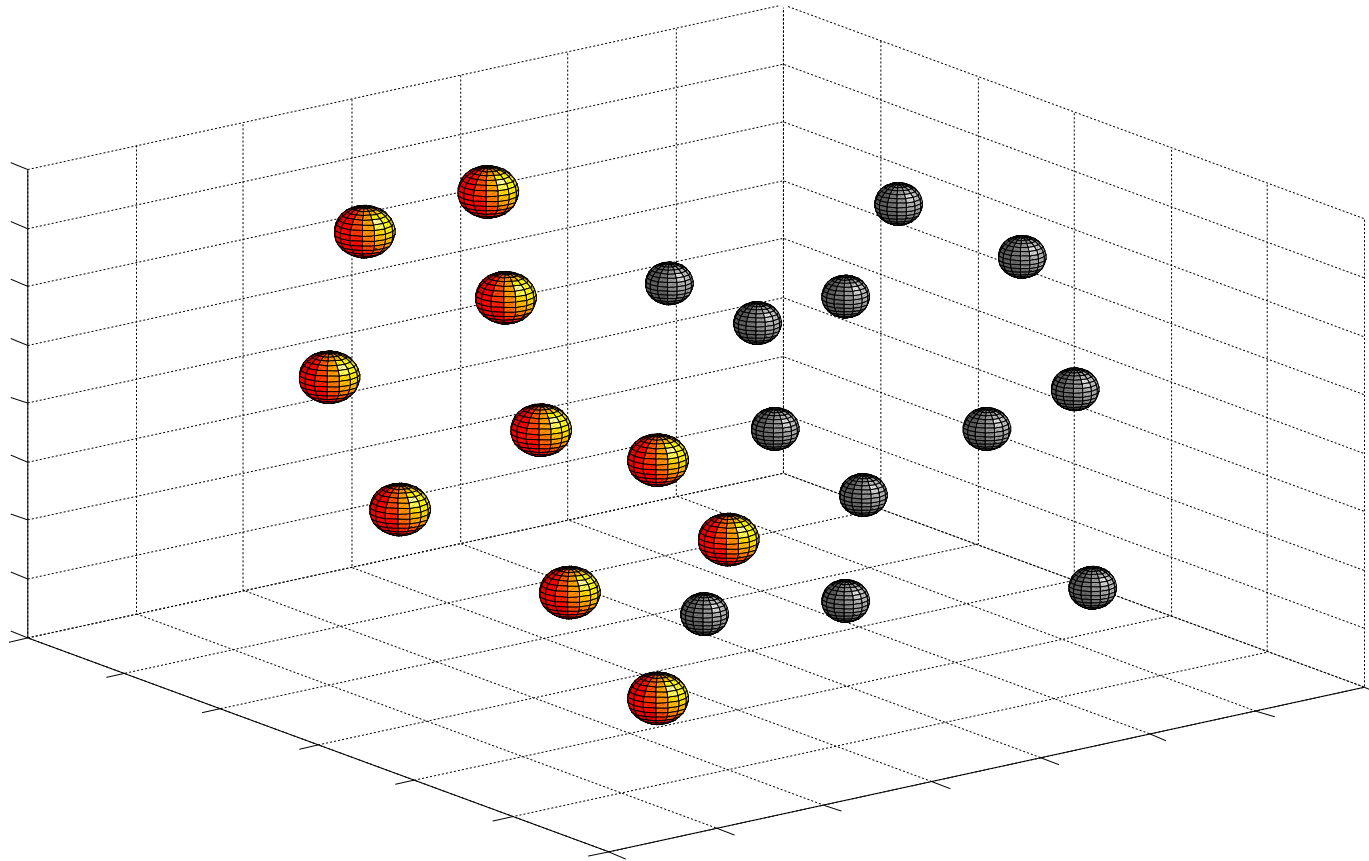


Esperança

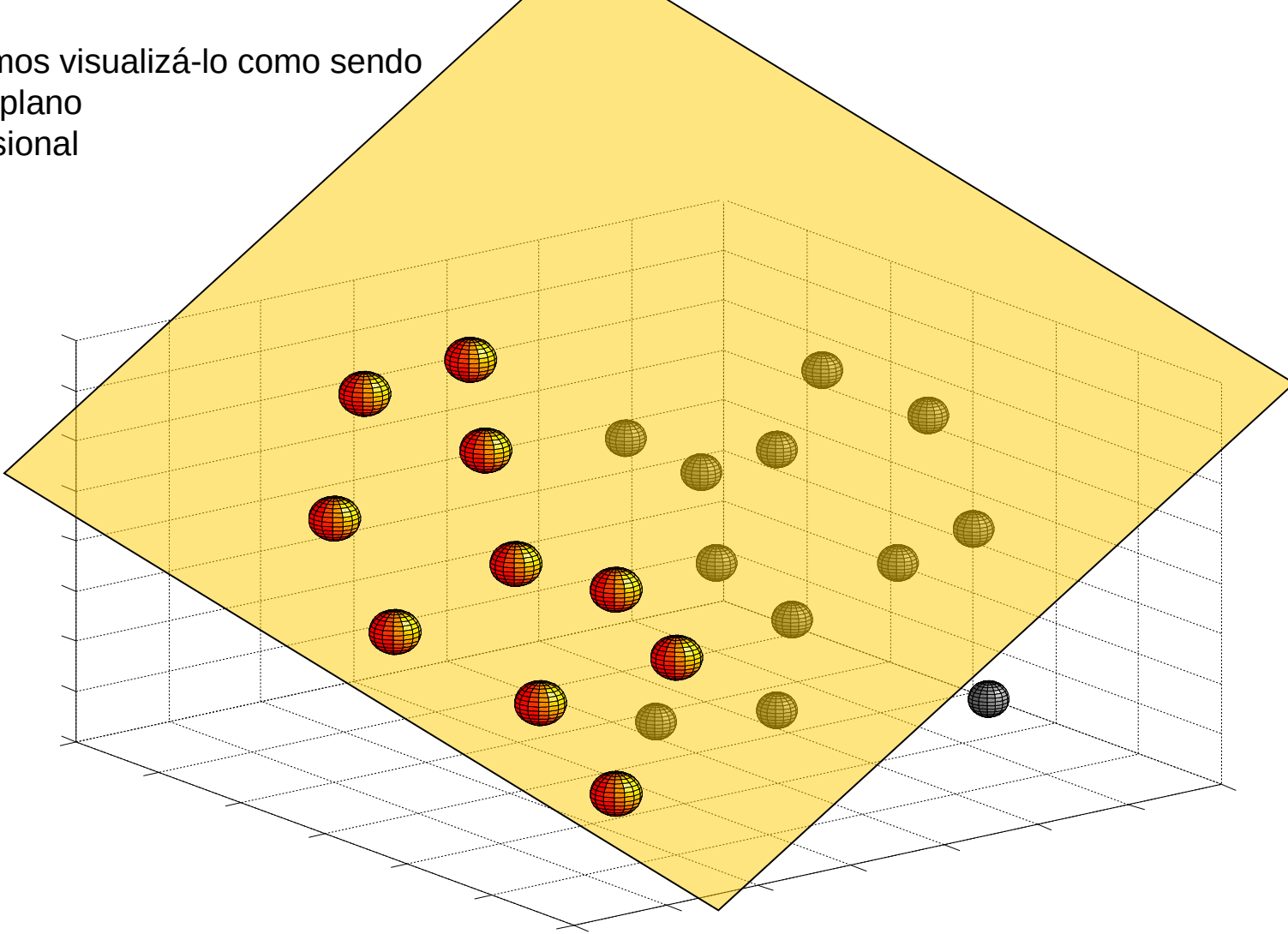


Gafanhoto

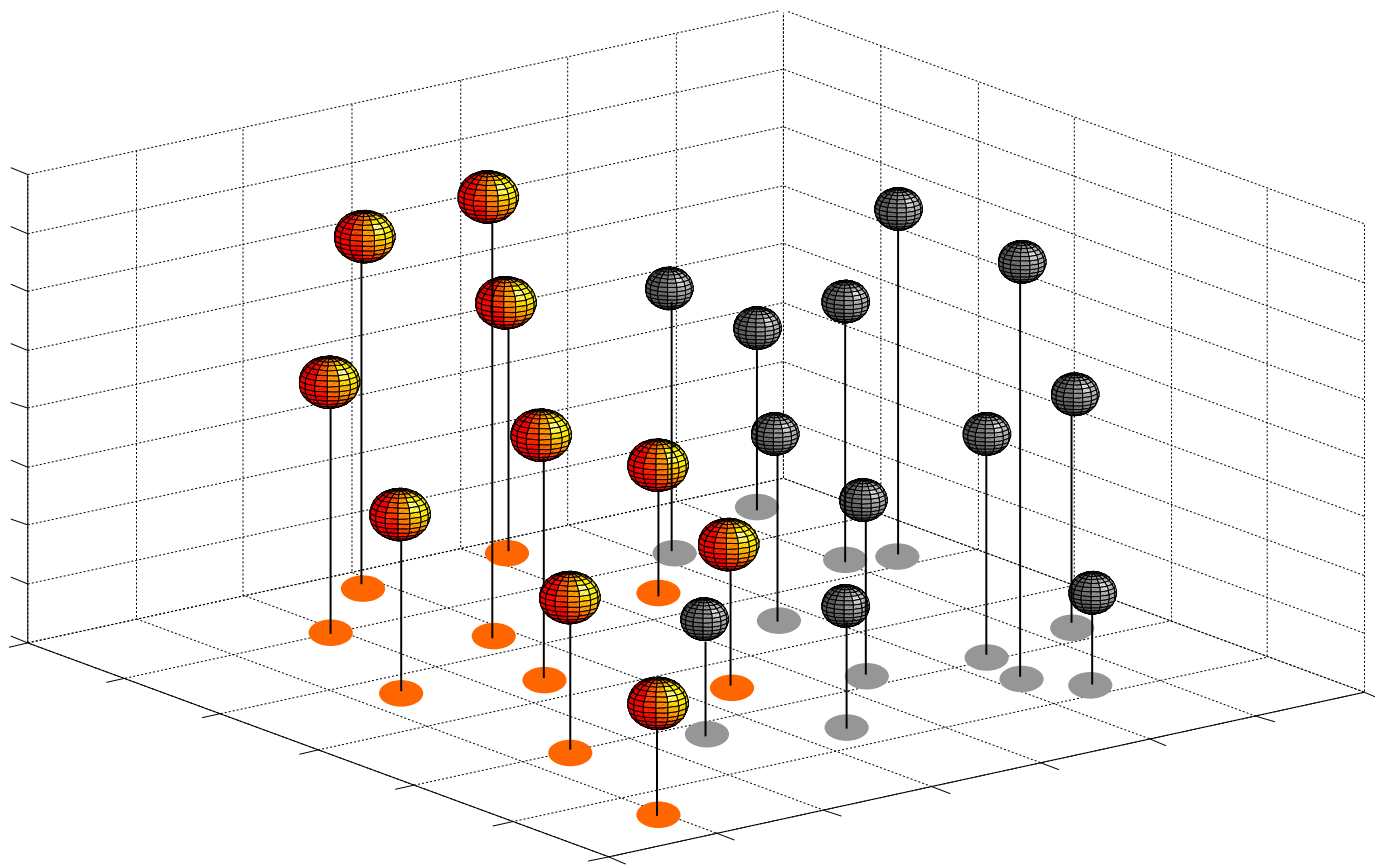
O classificador linear simples
é definido para espaços dimensionais maiores...

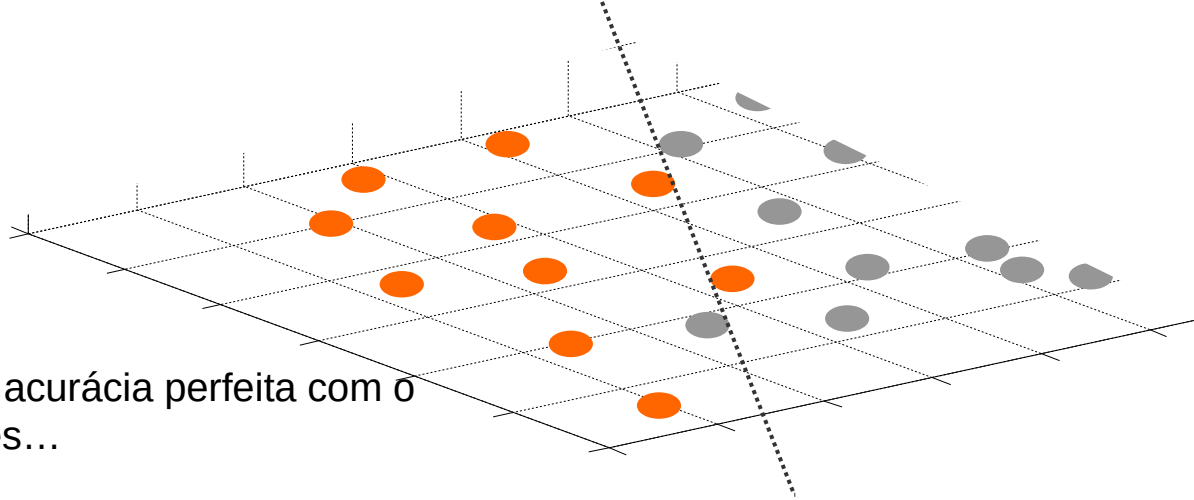


... podemos visualizá-lo como sendo
um hiperplano
n-dimensional



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

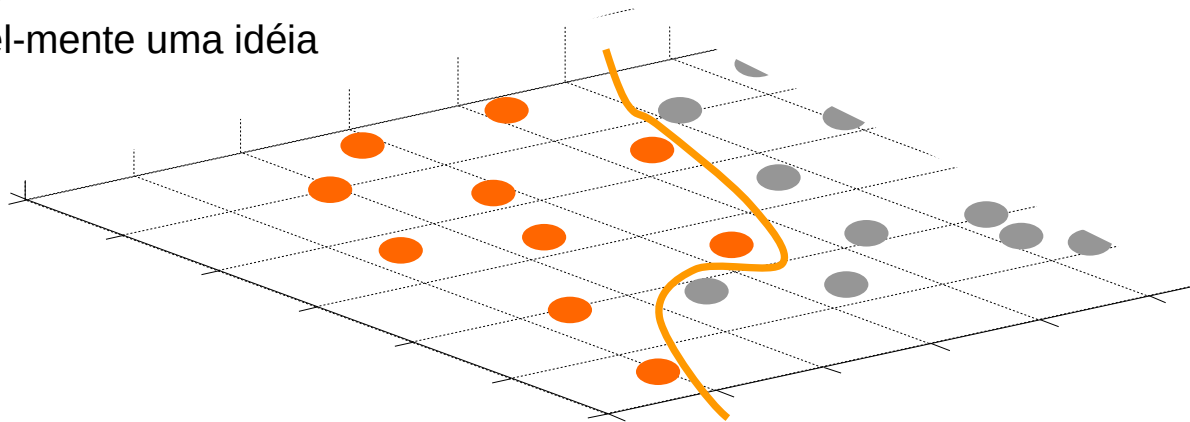




Não podemos mais obter acurácia perfeita com o classificador linear simples...

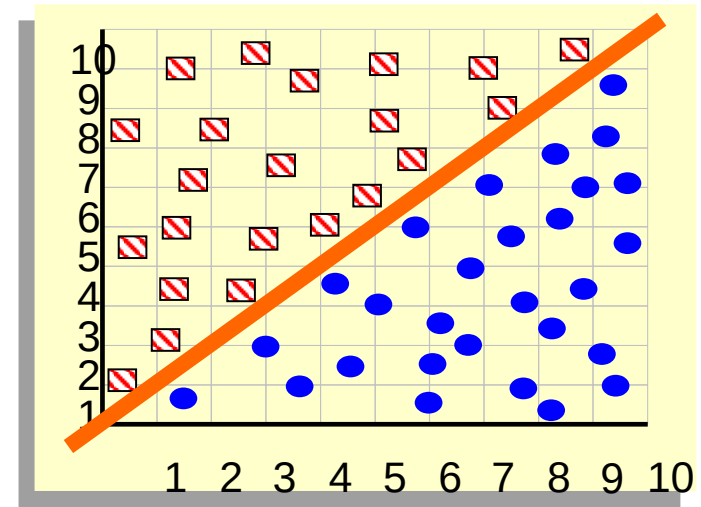
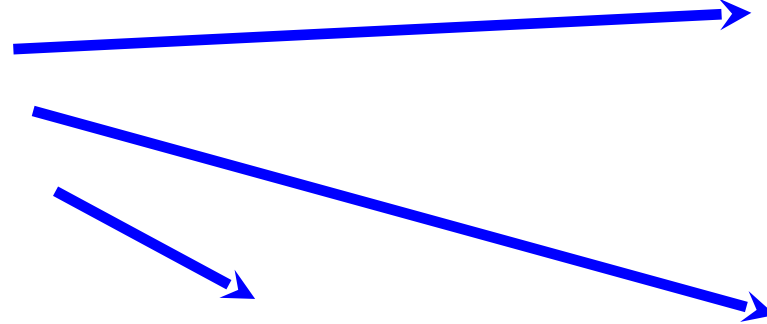
Podemos tentar resolver este problema usando um classificador *quadrático* simples ou um classificador *cúbico* simples...

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

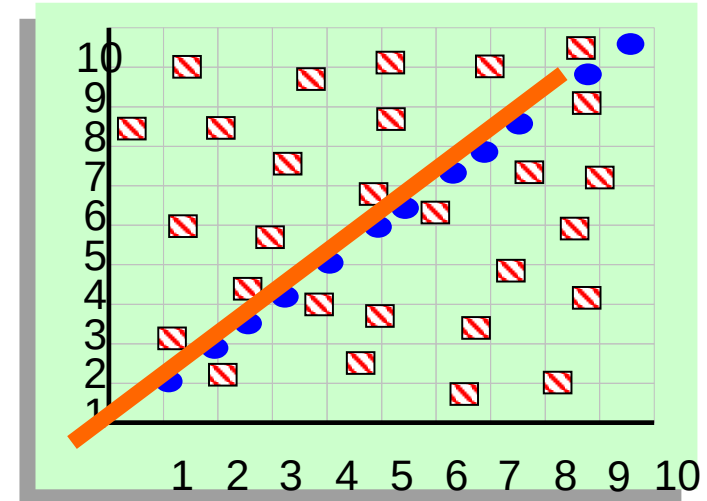
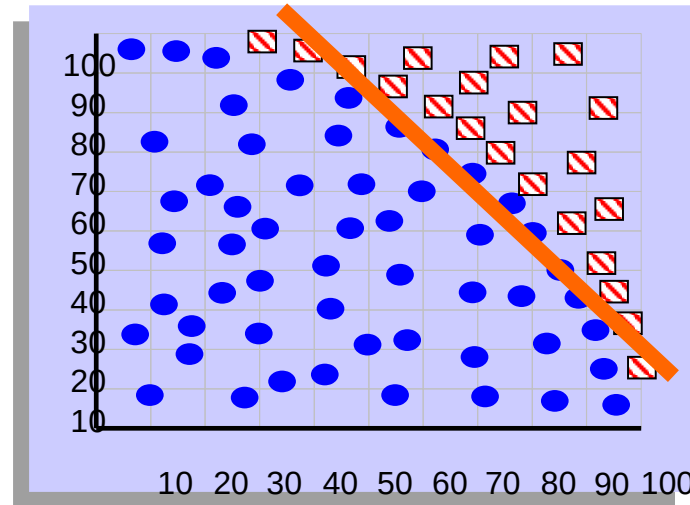


Quais dos “Problemas do Pombo” podem ser resolvidos pelo Classificador Linear Simples?

- 1) Perfeito
- 2) Inútil
- 3) Muito bom



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

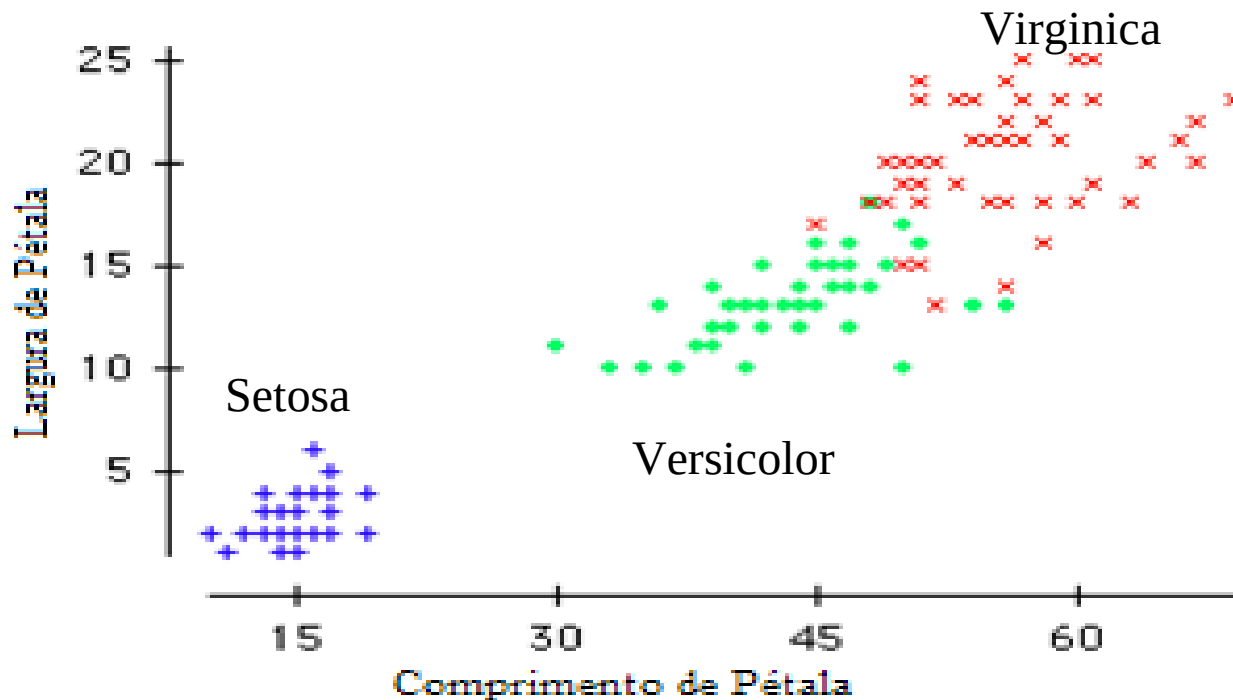


Um problema famoso

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

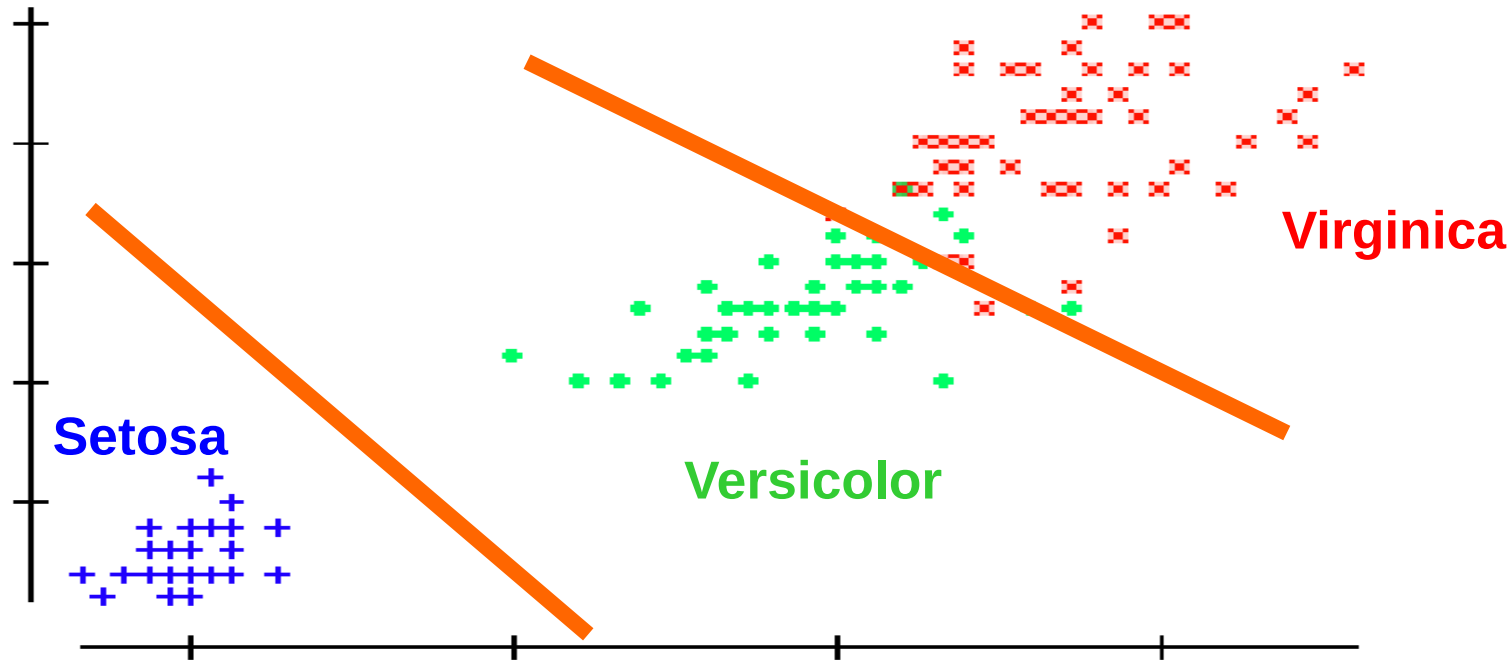


Iris Versicolor



Iris Virginica

Podemos generalizar o classificador linear relativo a variáveis a N classes, combinando N-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 $> 3.272 - (0.325 * \text{comp. de pétala})$
Então classe = **Virginica** Senão Se largura de pétala...

Referências

- Slides baseados parcialmente no tutorial do prof. Eamon Keogh (<http://www.cs.ucr.edu/~Eeamonn/tutorials.html>) e no livro Guide to Intelligent Data Science (<https://www.datascienceguide.org/teaching-material.html>)