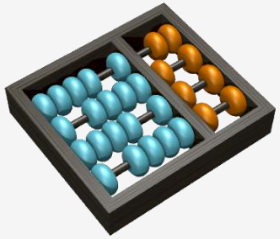


Data Mining

MC536 – Banco de Dados

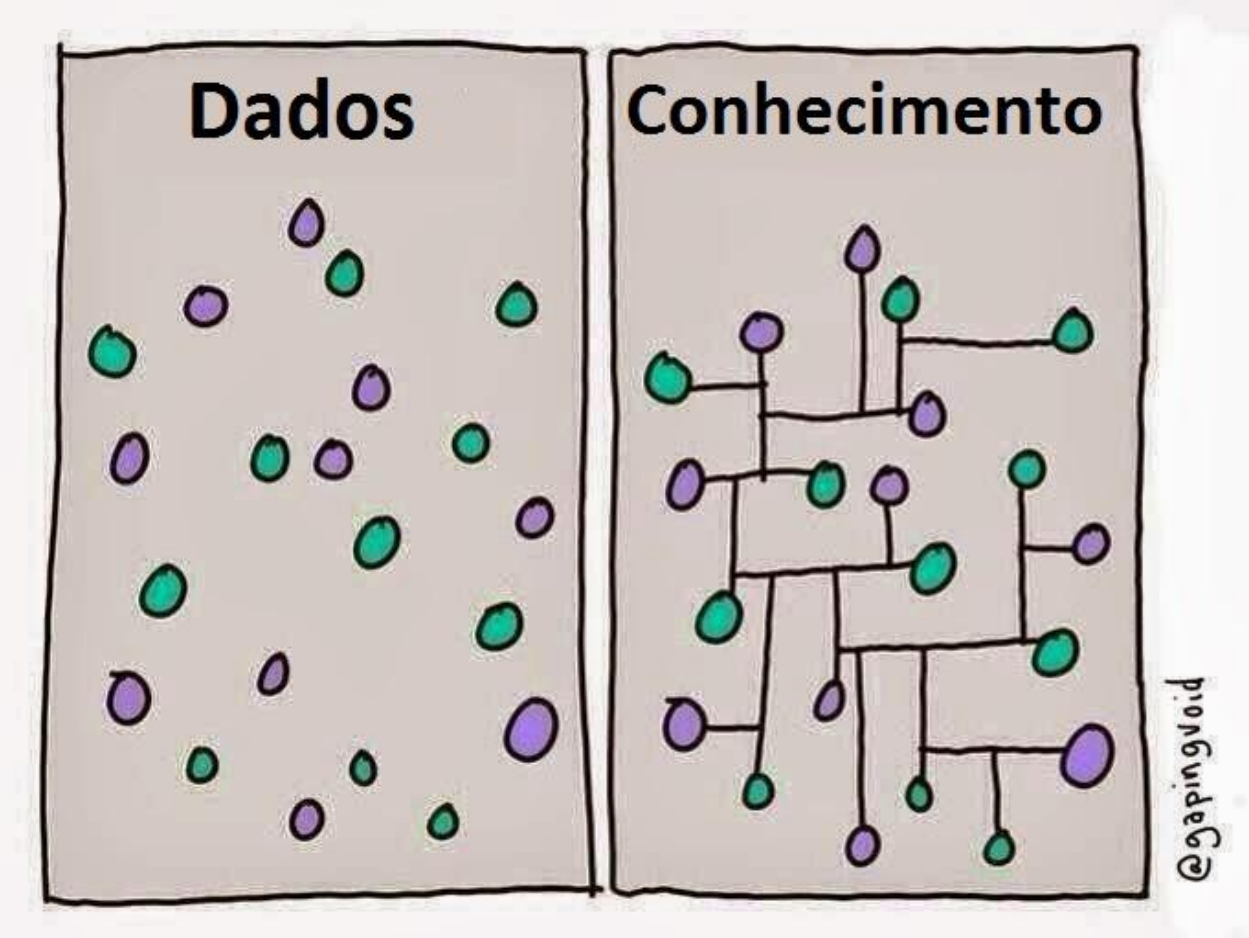
Profº.: André Santanchè

PED: Lucas Oliveira Batista

Introdução

- Dilúvio de dados
- Dados de empresa, sociedade, ciência, engenharia...
- Apenas dados são suficientes?

Introdução



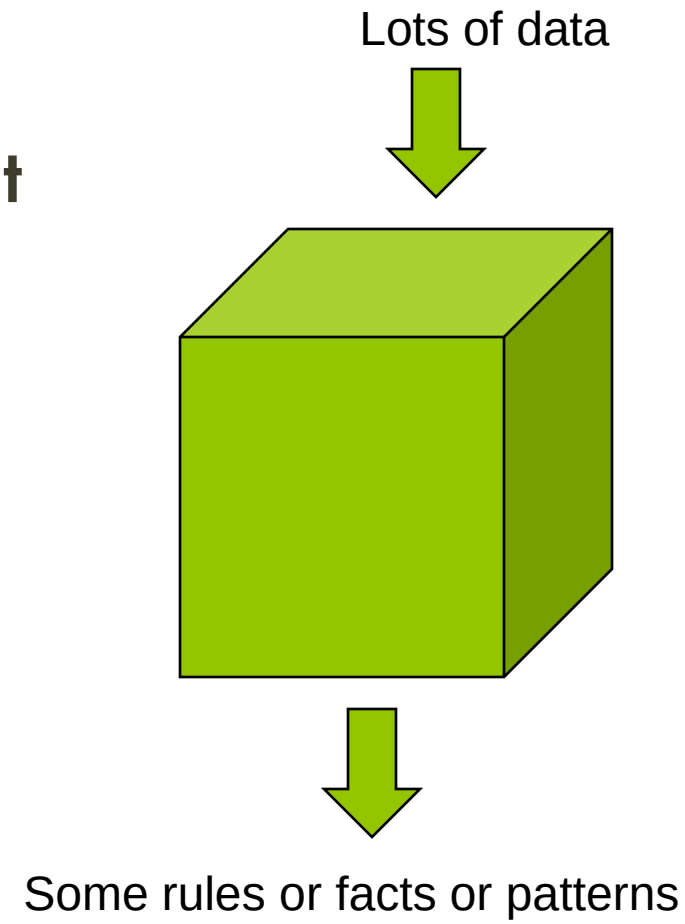
O que é Data Mining?

- "Data mining is the *process* of discovering interesting patterns and knowledge from *large* amounts of data" (Han; Kamber; Pei, 2011)

- Fontes de dados: Banco de Dados relacionais, Banco de Dados em grafos, Data Warehouse, Web…

O que é Data Mining?

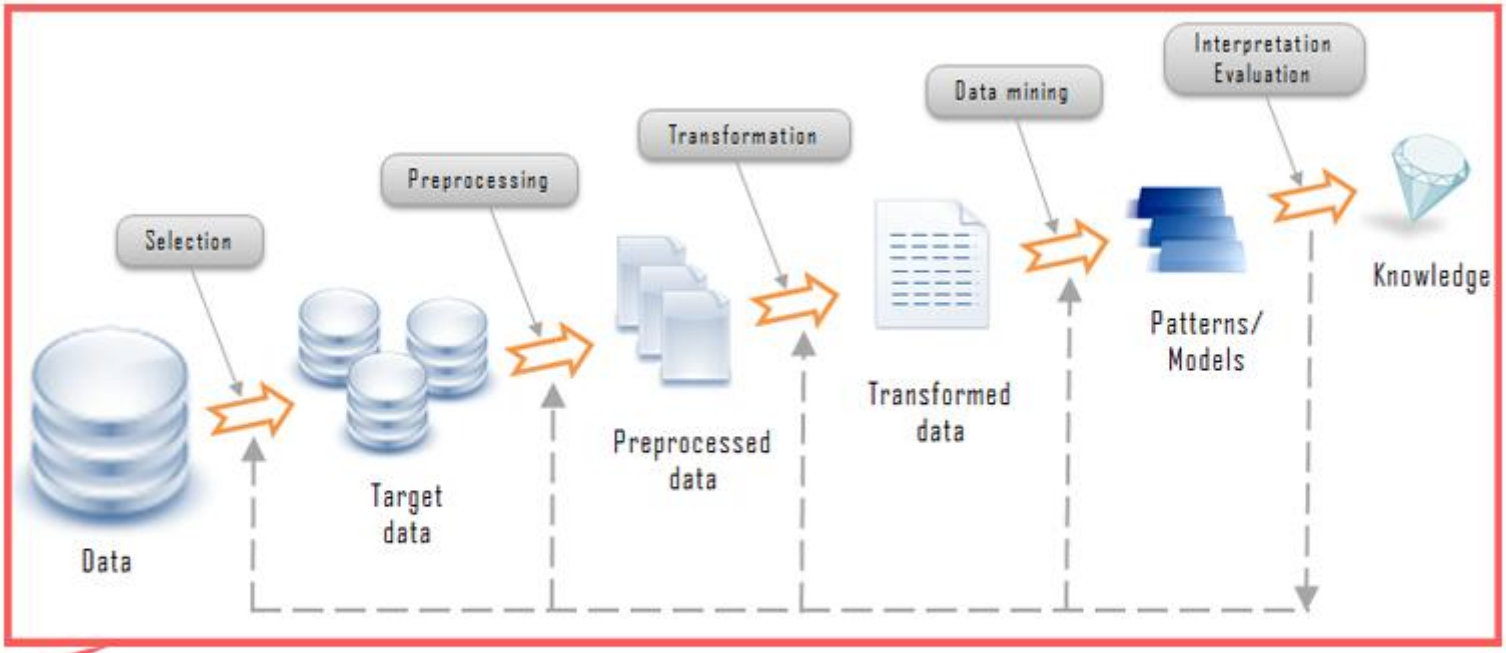
- Lots of raw data **in**
- *Some data mining*
- Facts, rules, patterns **out**



Fonte: Introdução a Data Mining Eamon Keogh

O que é Data Mining?

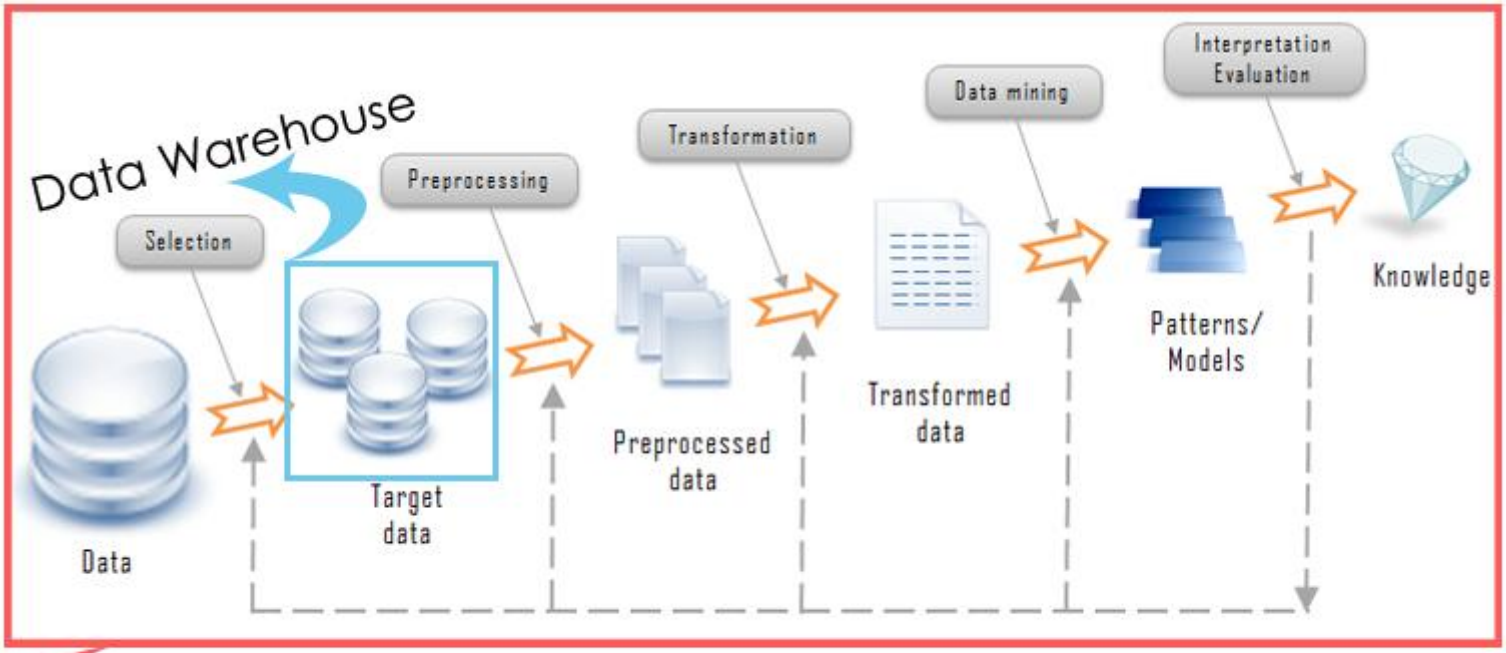
- Data Mining não é Knowledge Discovery from Data



Knowledge Discovery from Data ou KDD

O que é Data Mining?

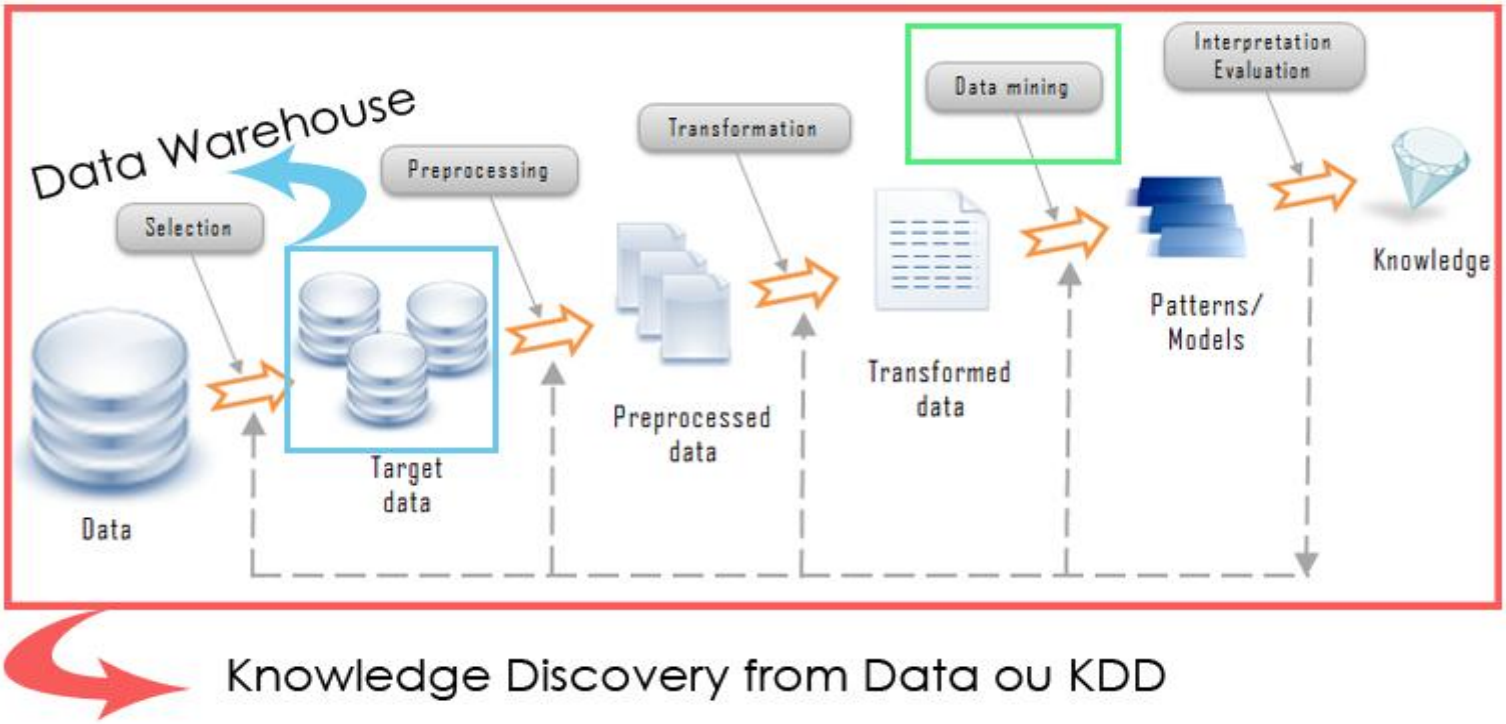
- Data Mining não é Data Warehouse



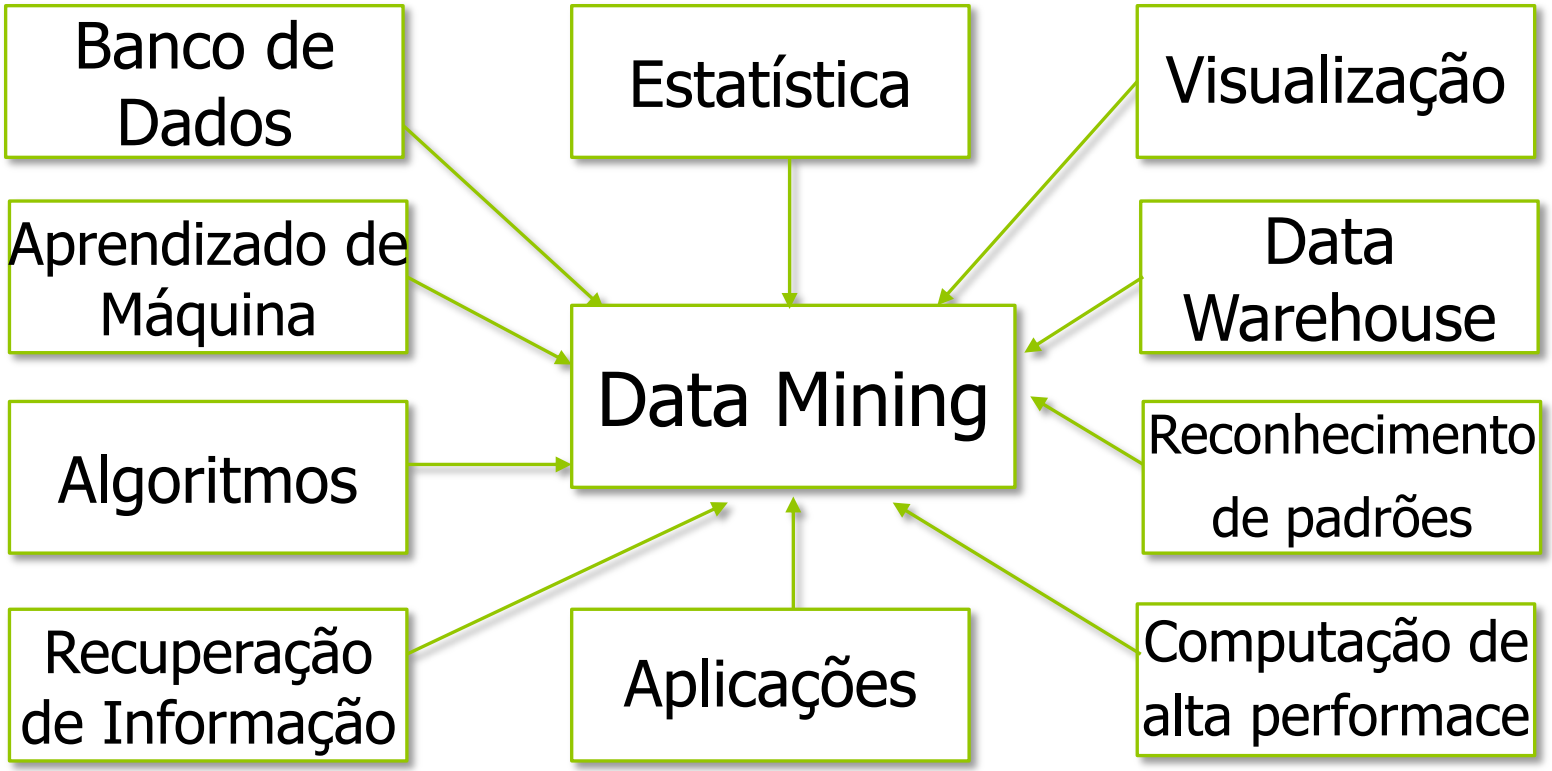
Knowledge Discovery from Data ou KDD

O que é Data Mining?

- Data Mining é um passo da KDD que aplica algoritmos específicos para extrair padrões a partir de dados



Quais técnicas são utilizadas?



Porque utilizar Data Mining?

- Enorme quantidade de dados são coletadas diariamente
- Dificuldade dos humanos em visualizar e entender grande conjunto de dados
- Permite análise de dados automática

Empresa “Um Pouco de Tudo”

- A Um Pouco de Tudo é uma grande empresa de eletrônicos com diversas filiais espalhadas pelo mundo

- Armazena uma grande quantidade de dados sobre filiais, empregados, clientes, produtos, transações de vendas...

Quais métodos são utilizados para gerar padrões?

- Técnicas de Data Mining são divididas em:
  - Descrição de Classes/Conceitos
  - Mineração de padrões frequentes, associações e correlações
  - Classificação e Regressão para análise preditiva
  - Análise de agrupamento
  - Análise de outlier

Quais métodos são utilizados para gerar padrões?

- Técnicas de Data Mining são divididas em:
  - Descrição de Classes/Conceitos
  - Mineração de padrões frequentes, associações e correlações
  - Classificação e Regressão para análise preditiva
  - Análise de agrupamento
  - Análise de outlier

Descrição de Classes/Conceitos

- Associa dados a classes ou conceitos
  - Classes de itens a venda: computadores ou impressoras
  - Conceito de clientes: gastaMuito ou gastaPouco
- Derivados usando caracterização de dados e/ou discriminação de dados

Descrição de Classes/Conceitos: Caracterização

- Características gerais de uma classe
- Um Pouco de Tudo: características de clientes que gastam mais de R$ 5000 por ano

Descrição de Classes/Conceitos: Caracterização

- Características gerais de uma classe

- Um Pouco de Tudo: características de clientes que gastam mais de R$ 5000 por ano

Clientes entre 40 e 50 anos, empregados e com alta taxa de crédito

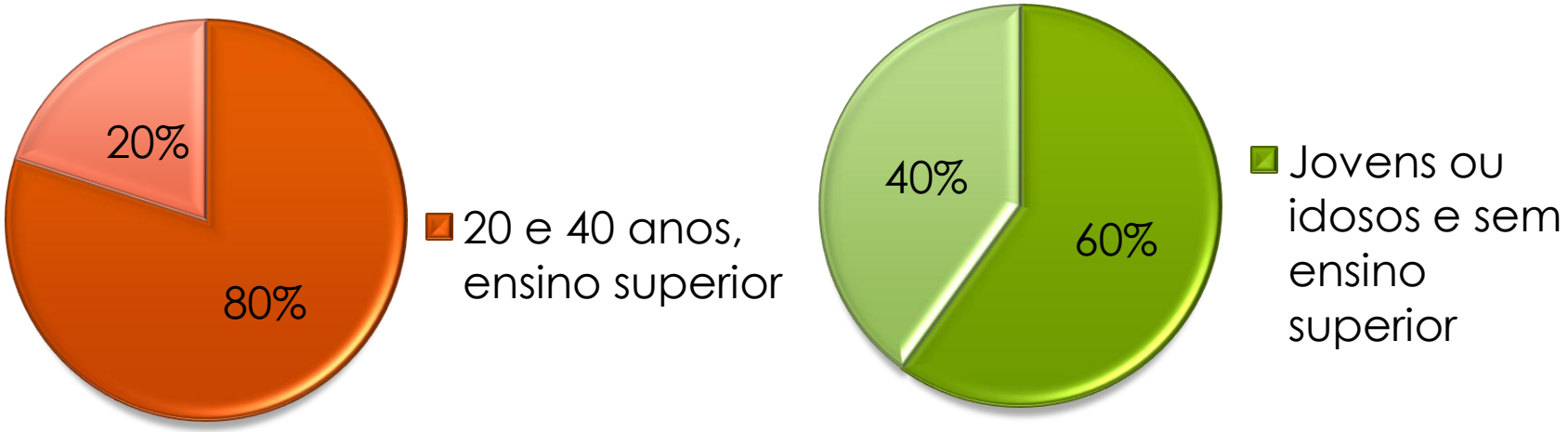
Descrição de Classes/Conceitos: Discriminação

- Comparar características entre classes

- Um Pouco de Tudo: Clientes que compram regularmente X Clientes que raramente compram

Descrição de Classes/Conceitos: Discriminação

- Um Pouco de Tudo: Clientes que compram regularmente X Clientes que raramente compram



Quais métodos são utilizados para gerar padrões?

- Técnicas de Data Mining são divididas em:
  - Descrição de Classes/Conceitos
  - Mineração de padrões frequentes, associações e correlações
  - Classificação e Regressão para análise preditiva
  - Análise de agrupamento
  - Análise de outlier

Padrões frequentes, associações e correlações

- Padrões frequentes gera associações e correlações entre dados
- Um Pouco de Tudo: Quais itens são frequentemente comprados juntos?

Padrões frequentes, associações e correlações

- Padrões frequentes gera associações e correlações entre dados

- Um Pouco de Tudo: Quais itens são frequentemente comprados juntos?

Compra (Cliente, PC) => Compra (Cliente, **Software**) [suporte: **25%** confiança: **50%**]

Padrões frequentes, associações e correlações

- Um Pouco de Tudo: Quais itens são frequentemente comprados juntos?

Compra (Cliente, PC) => Compra (Cliente, **Software**) [suporte: **25%** confiança: **50%**]

Transação 1	PC, DVD, Software
Transação 2	DVD, Cartão Memória
Transação 3	PC, Cartão Memória
Transação 4	Televisão, Som

Padrões frequentes, associações e correlações

- Um Pouco de Tudo: Quais itens são frequentemente comprados juntos?

Compra (Cliente, PC) => Compra (Cliente, **Software**) [**suporte: 25%** confiança: **50%**]

Transação 1	**PC, DVD, Software**
Transação 2	**DVD, Cartão Memória**
Transação 3	**PC, Cartão Memória**
Transação 4	**Televisão, Som**

Padrões frequentes, associações e correlações

- Um Pouco de Tudo: Quais itens são frequentemente comprados juntos?

Compra (Cliente, PC) => Compra (Cliente, **Software**) [suporte: **25%** **confiança: 50%**]

Transação 1	**PC, DVD, Software**
Transação 2	DVD, Cartão Memória
Transação 3	**PC, Cartão Memória**
Transação 4	Televisão, Som

Padrões frequentes, associações e correlações

- Padrões frequentes gera associações e correlações entre dados

- Um Pouco de Tudo: Quais itens são frequentemente comprados juntos?

Compra (Cliente, PC) => Compra (Cliente, **CD**) [suporte: **0.3%** confiança: **5%**]

Padrões frequentes, associações e correlações

- Padrões frequentes gera associações e correlações entre dados

- Um Pouco de Tudo: Quais itens são frequentemente comprados juntos?

  Compra (Cliente, PC) => Compra (Cliente, **CD**) [suporte: **0,2%**, confiança: **5%**]

Exercício 1

- Cite 2 padrões frequentes que poderiam ser minerados considerando o banco de dados abaixo.

Transação 1	Pão, leite, queijo, presunto, desodorante, feijão
Transação 2	Achocolatado, pão, leite
Transação 3	Cebola, laranja, salsa, manga
Transação 4	Carne, presunto, ovos, queijo, pão
Transação 5	Chocolate, pipoca, refrigerante, leite
Transação 6	Caneta, bala, fralda, queijo, leite, pão

Quais métodos são utilizados para gerar padrões?

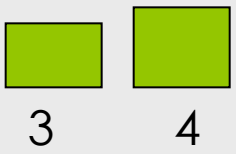
- Técnicas de Data Mining são divididas em:
  - Descrição de Classes/Conceitos
  - Mineração de padrões frequentes, associações e correlações
  - Classificação e Regressão para análise preditiva
  - Análise de agrupamento
  - Análise de outlier

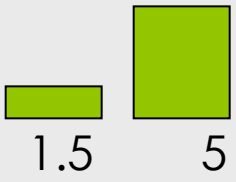
Classificação e Regressão para análise preditiva

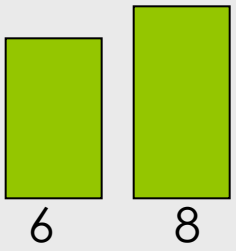
- Classificação: baseada na análise de dados de classes conhecidas

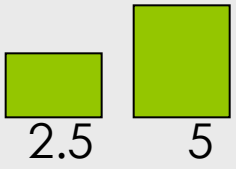
Pigeon Problem 1 (extraído de Eamon Keogh)

Examples of class A

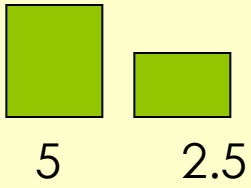


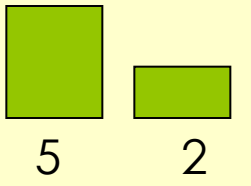


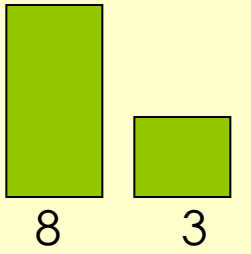


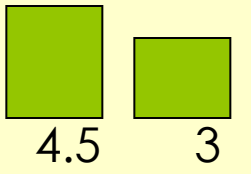


Examples of class B



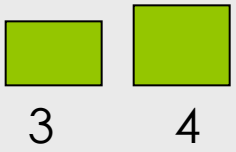


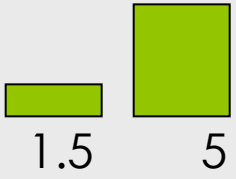


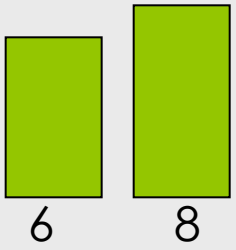


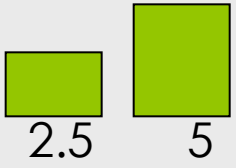
Pigeon Problem 1

Examples of class A

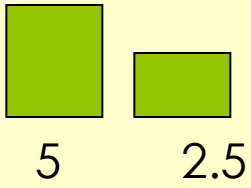


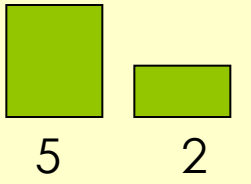


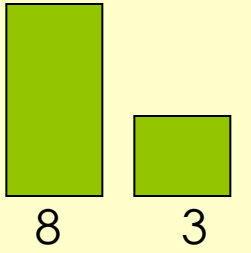


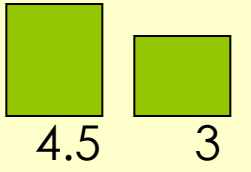


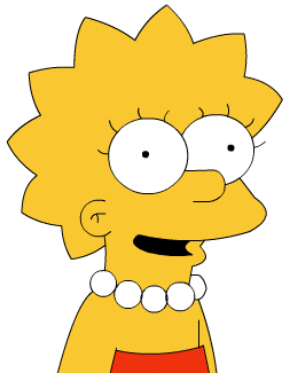
Examples of class B



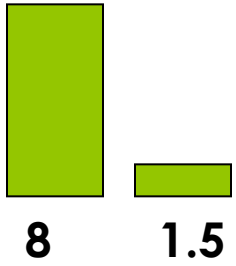




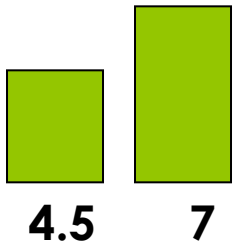




What class is this object?

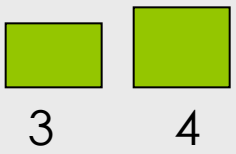


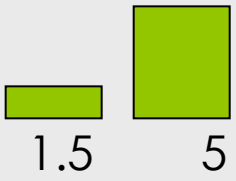
What about this one, A or B?

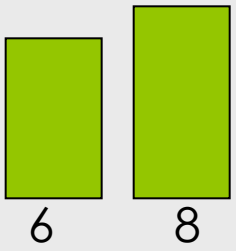


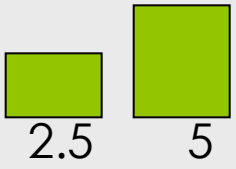
Pigeon Problem 1

Examples of class A

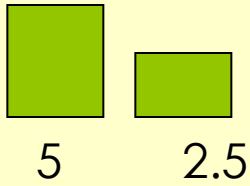


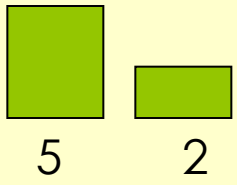


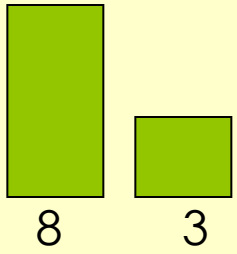


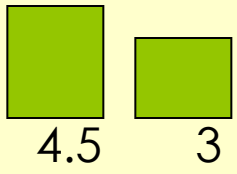


Examples of class B

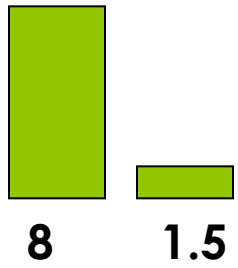








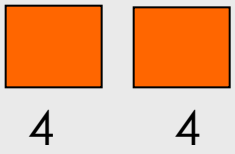
This is a B!

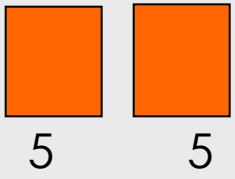


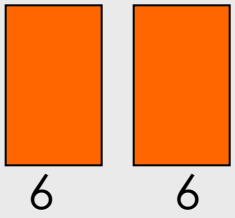
Here is the rule.
If the left bar is smaller than the right bar, it is an A, otherwise it is a B.

Pigeon Problem 2

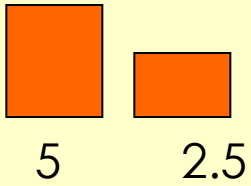
Examples of class A

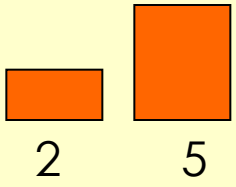


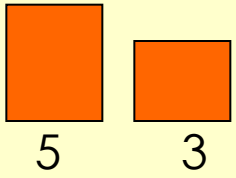


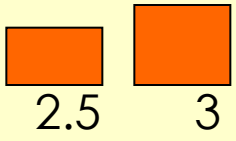


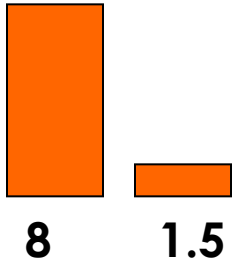
Examples of class B

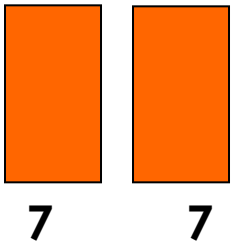






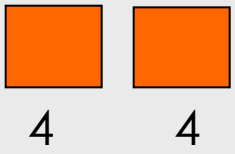


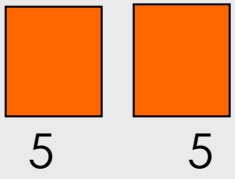


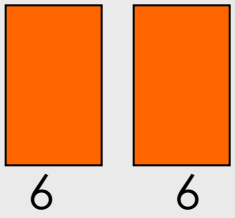


Pigeon Problem 2

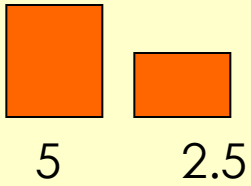
Examples of class A

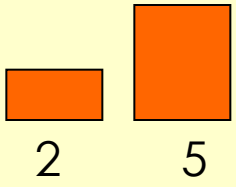


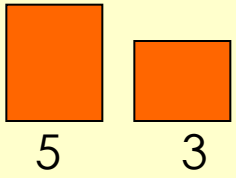


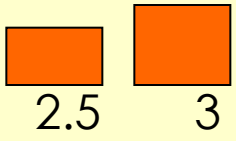


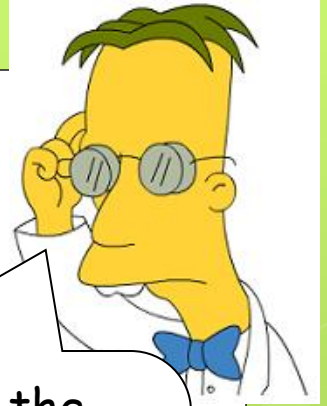
Examples of class B





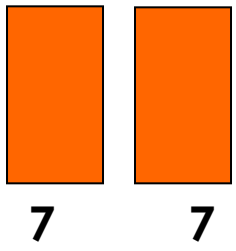






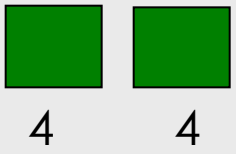
The rule is as follows, if the two bars are equal sizes, it is an A. Otherwise it is a B.

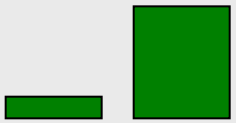
So this one is an A.

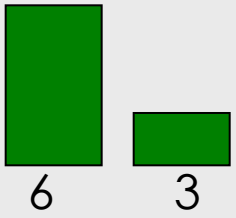


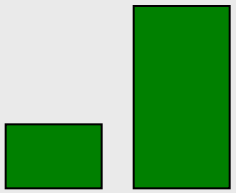
Pigeon Problem 3

Examples of class A

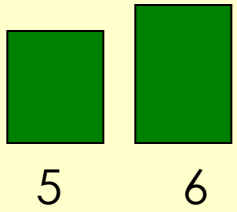


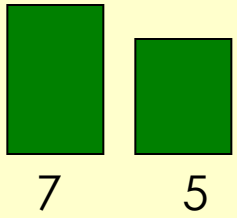


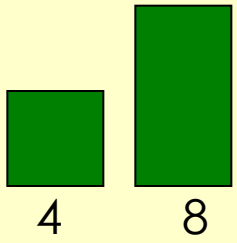


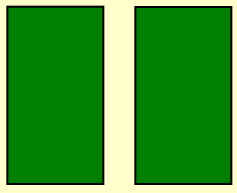


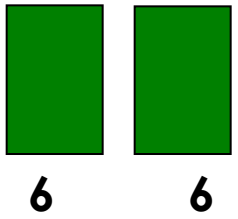
Examples of class B







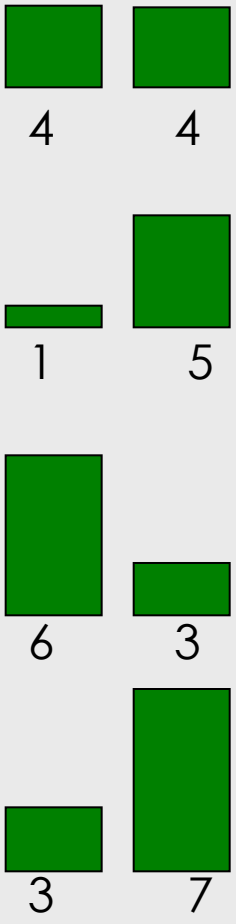




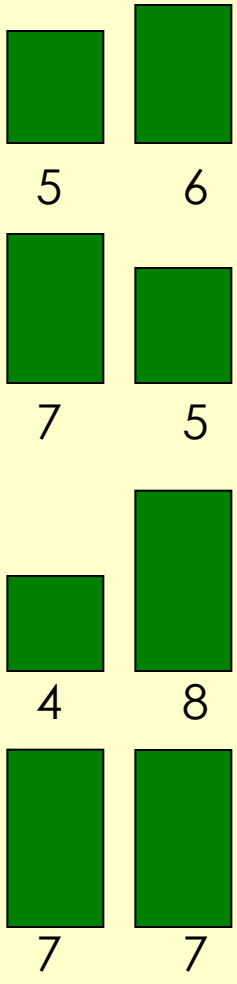
This one is really hard!
What is this, A or B?

Pigeon Problem 3

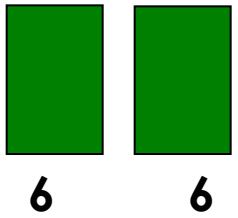
Examples of class A



Examples of class B



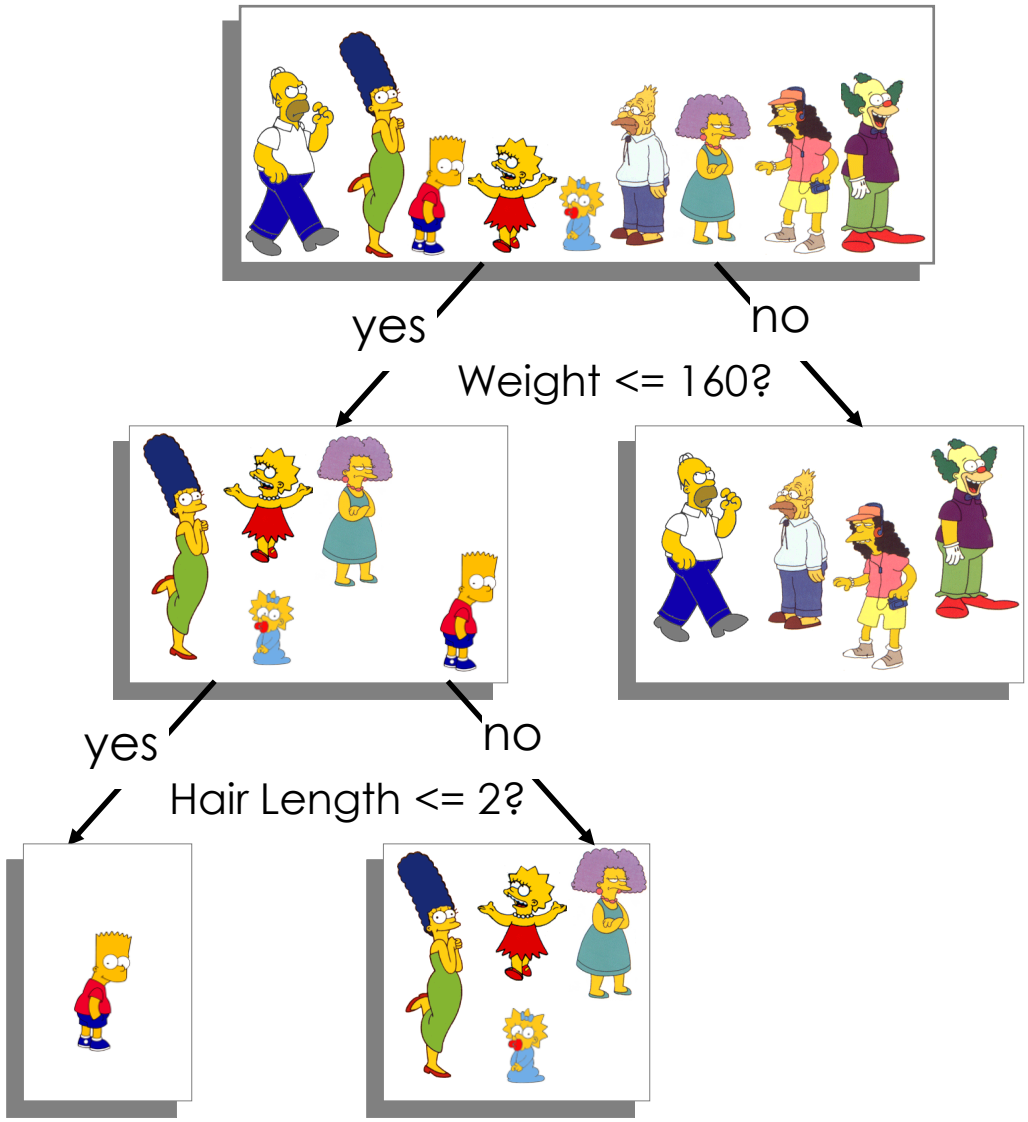
It is a B!



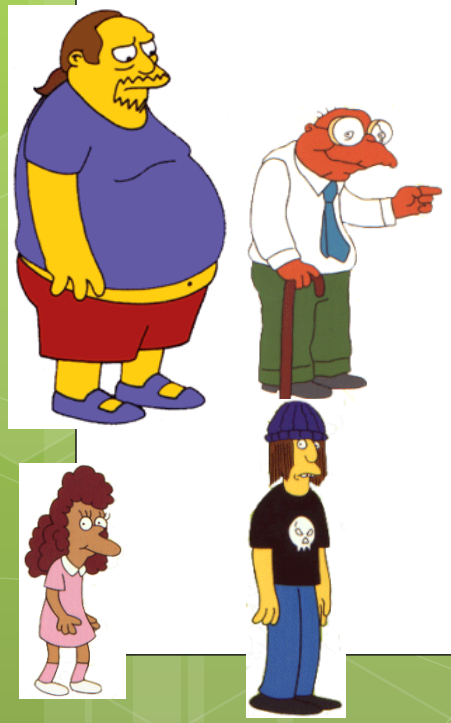
The rule is as follows, if the square of the sum of the two bars is less than or equal to 100, it is an A. Otherwise it is a B.

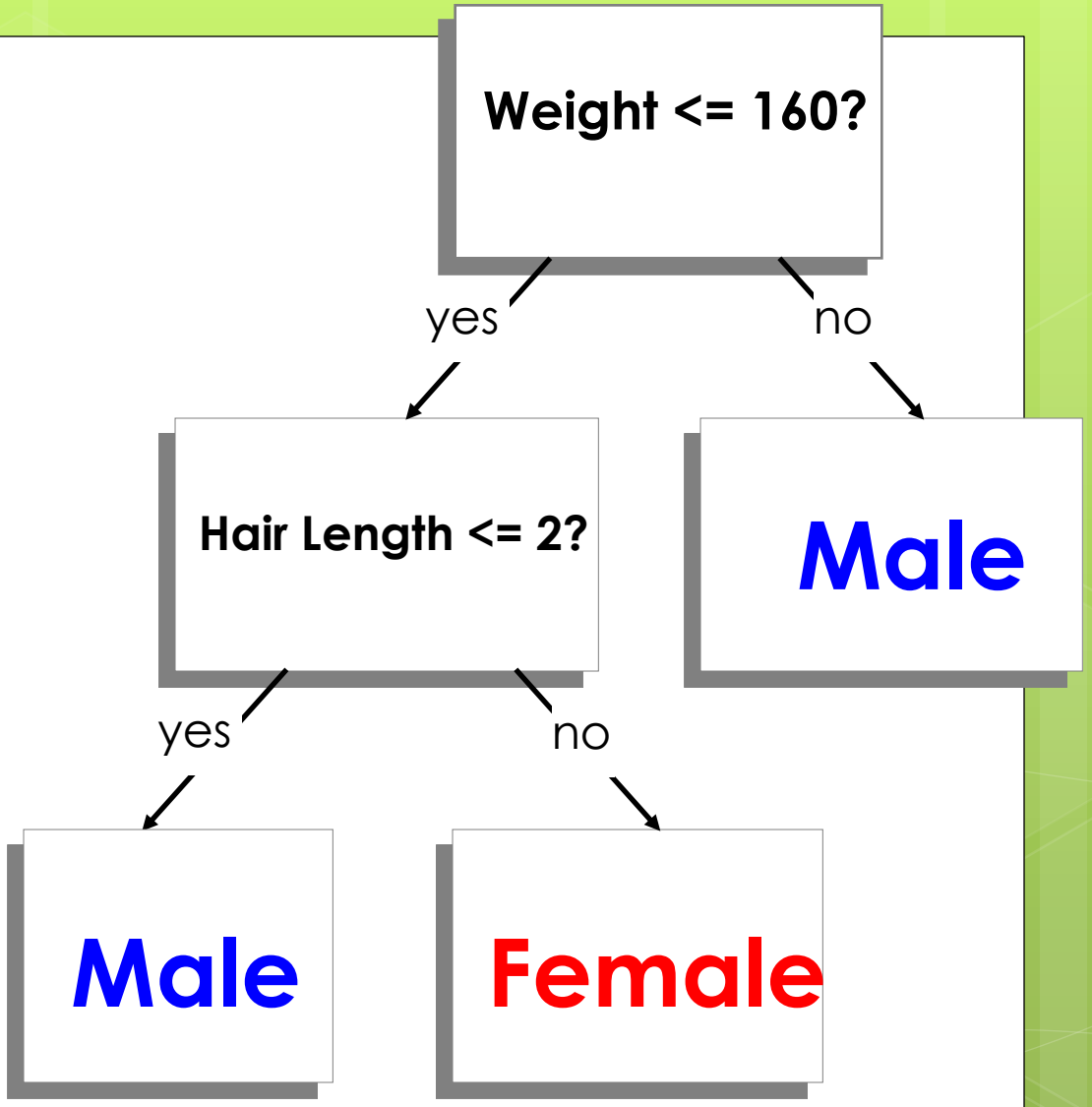
Person	Hair Length	Weight	Age	Class
 Homer	0”	250	36	**M**
 Marge	10”	150	34	**F**
 Bart	2”	90	10	**M**
 Lisa	6”	78	8	**F**
 Maggie	4”	20	1	**F**
 Abe	1”	170	70	**M**
 Selma	8”	160	41	**F**
 Otto	10”	180	38	**M**
 Krusty	6”	200	45	**M**

 Comic	8”	290	38	**?**


yes
no
Weight <= 160?
yes
no
Hair Length <= 2?

How would these people be classified?





Classificação e Regressão para análise preditiva

- Um Pouco de Tudo: Classificar o resultado de um item em promoção em BOM ou RUIM

Classificação e Regressão para análise preditiva

- Um Pouco de Tudo: Classificar o resultado de um item em promoção em BOM ou RUIM

Desconto, marca	19%, ITAUTEC	**Desconto, marca**
20%, DELL 50%, ITAUTEC 30%, DELL 21%, ITAUTEC 10%, DELL 15%, DELL	?	10%, ITAUTEC 15%, ITAUTEC 5%, ITAUTEC 20%, ITAUTEC

Classificação e Regressão para análise preditiva

- Um Pouco de Tudo: Classificar o resultado de um item em promoção em BOM ou RUIM

Desconto, marca

20%, DELL
50%, ITAUTEC
30%, DELL
21%, ITAUTEC
10%, DELL
15%, DELL

19%, ITAUTEC

Padrão:
Se desconto>20%
BOM
SenãoSe marca=DELL
BOM
Senão RUIM

Desconto, marca

10%, ITAUTEC
15%, ITAUTEC
5%, ITAUTEC
20%, ITAUTEC

Classificação e Regressão para análise preditiva

- Um Pouco de Tudo: Classificar o resultado de um item em promoção em BOM ou RUIM

Desconto, marca

20%, DELL
50%, ITAUTEC
30%, DELL
21%, ITAUTEC
10%, DELL
15%, DELL

19%, ITAUTEC

Padrão:
Se desconto>20%
BOM
SenãoSe marca=DELL
BOM
Senão RUIM

Desconto, marca

10%, ITAUTEC
15%, ITAUTEC
5%, ITAUTEC
20%, ITAUTEC

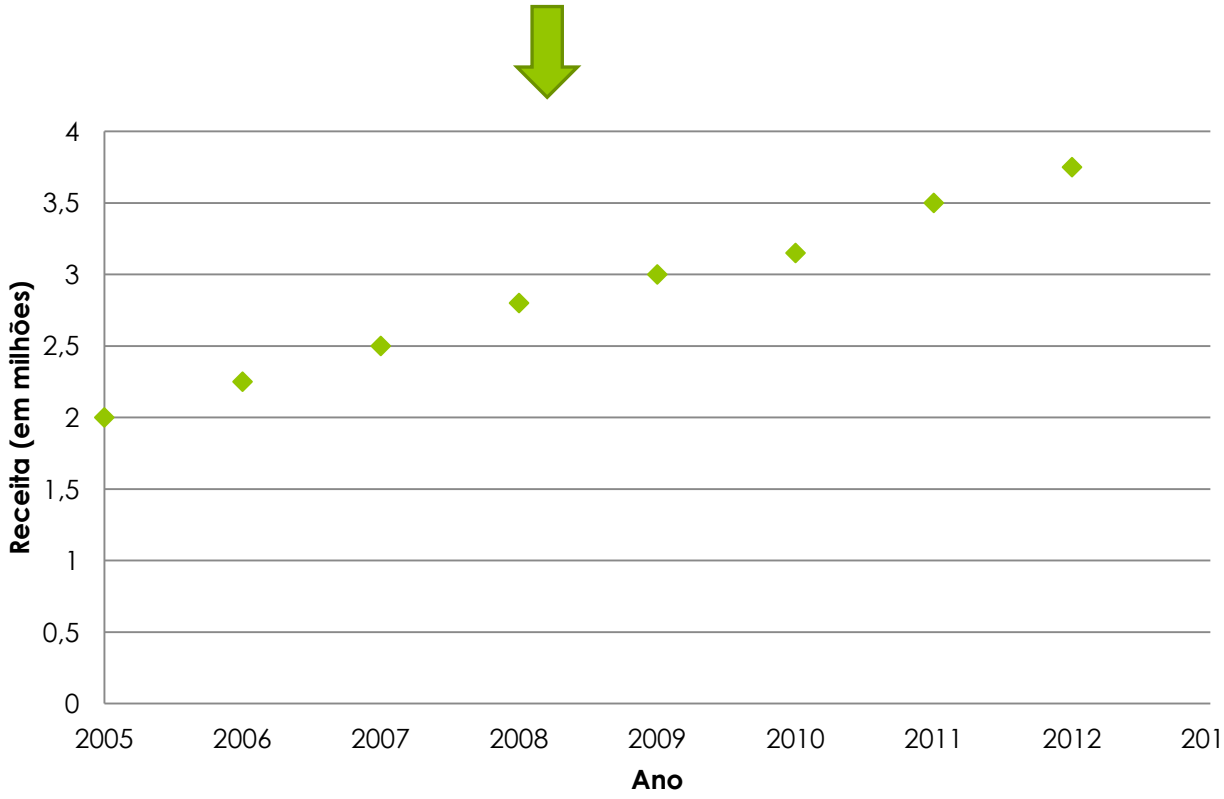
Classificação e Regressão para análise preditiva

- Regressão: prever valores em falta ou não disponíveis

- Um Pouco de Tudo: Prever a receita de um item com base em anos anteriores

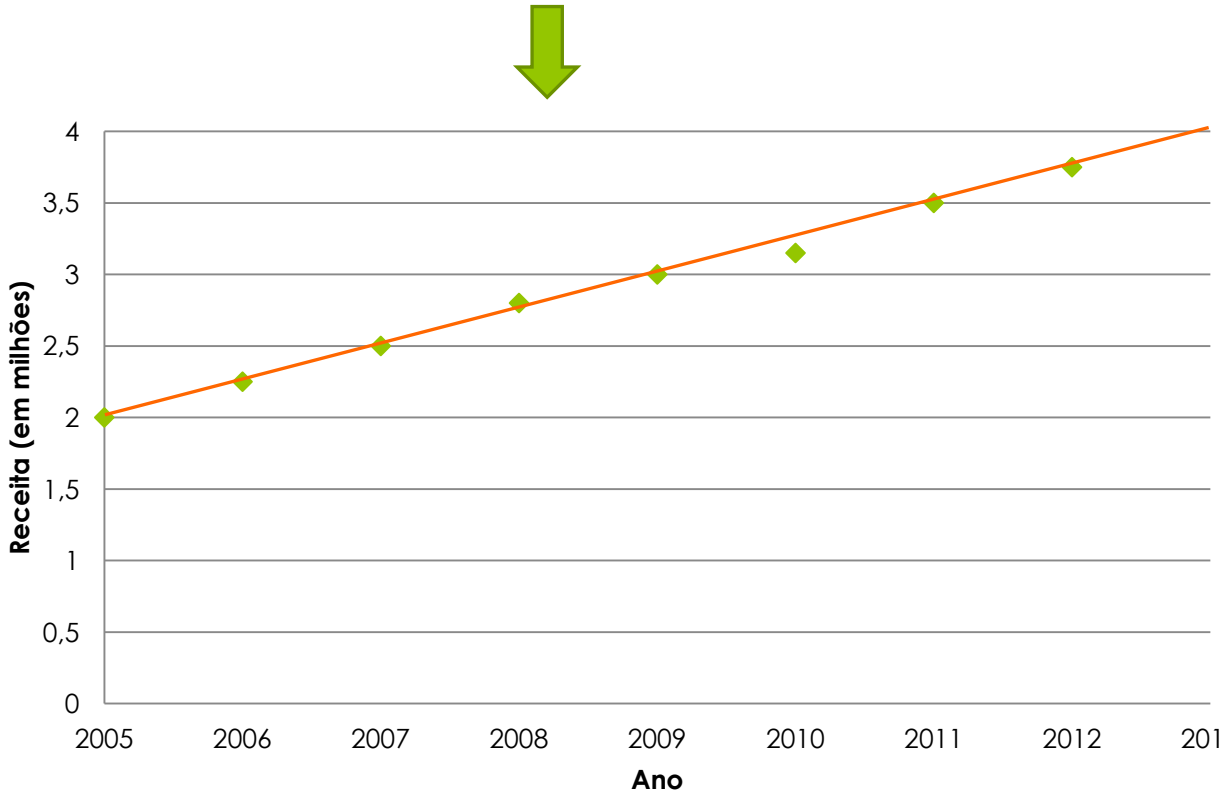
Classificação e Regressão para análise preditiva

- Um Pouco de Tudo: Prever a receita de um item com base em anos anteriores



Classificação e Regressão para análise preditiva

- Um Pouco de Tudo: Prever a receita de um item com base em anos anteriores



Quais métodos são utilizados para gerar padrões?

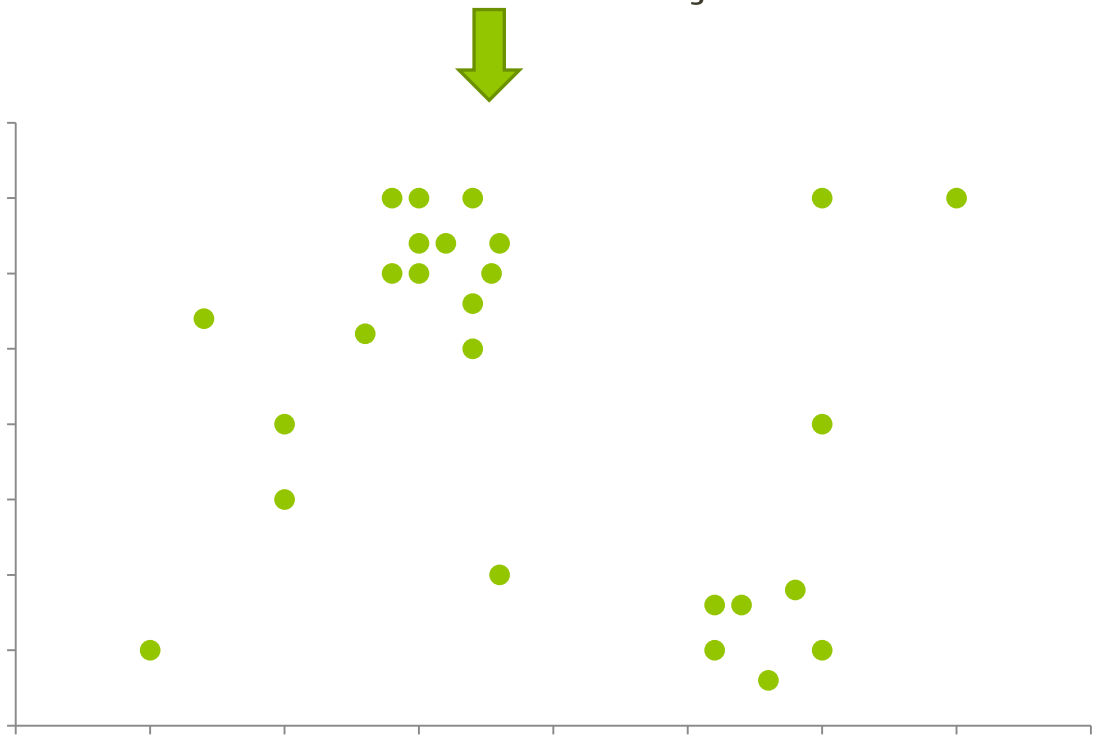
- Técnicas de Data Mining são divididas em:
  - Descrição de Classes/Conceitos
  - Mineração de padrões frequentes, associações e correlações
  - Classificação e Regressão para análise preditiva
  - Análise de agrupamento
  - Análise de outlier

Análise de Agrupamento

- Analisa o conjunto de dados sem conhecer as classes que pertencem
- Um Pouco de Tudo: Agrupar os clientes de acordo com seu endereço

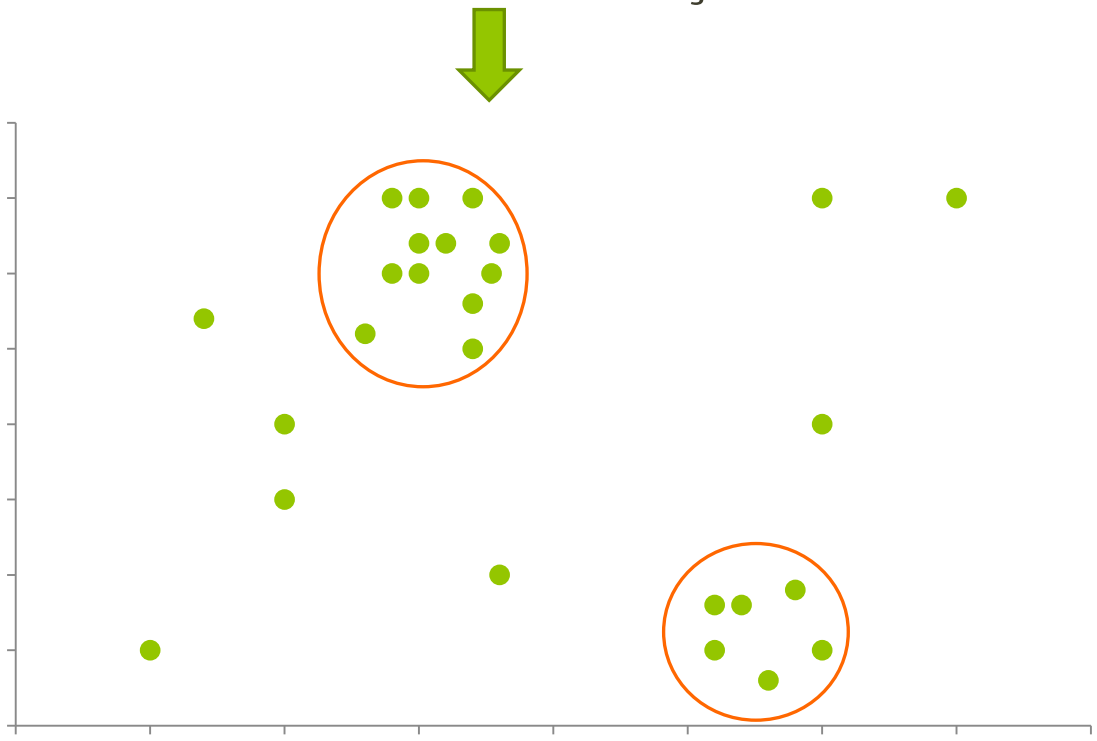
Análise de Agrupamento

- Um Pouco de Tudo: Agrupar os clientes de acordo com seu endereço



Análise de Agrupamento

- Um Pouco de Tudo: Agrupar os clientes de acordo com seu endereço



Quais métodos são utilizados para gerar padrões?

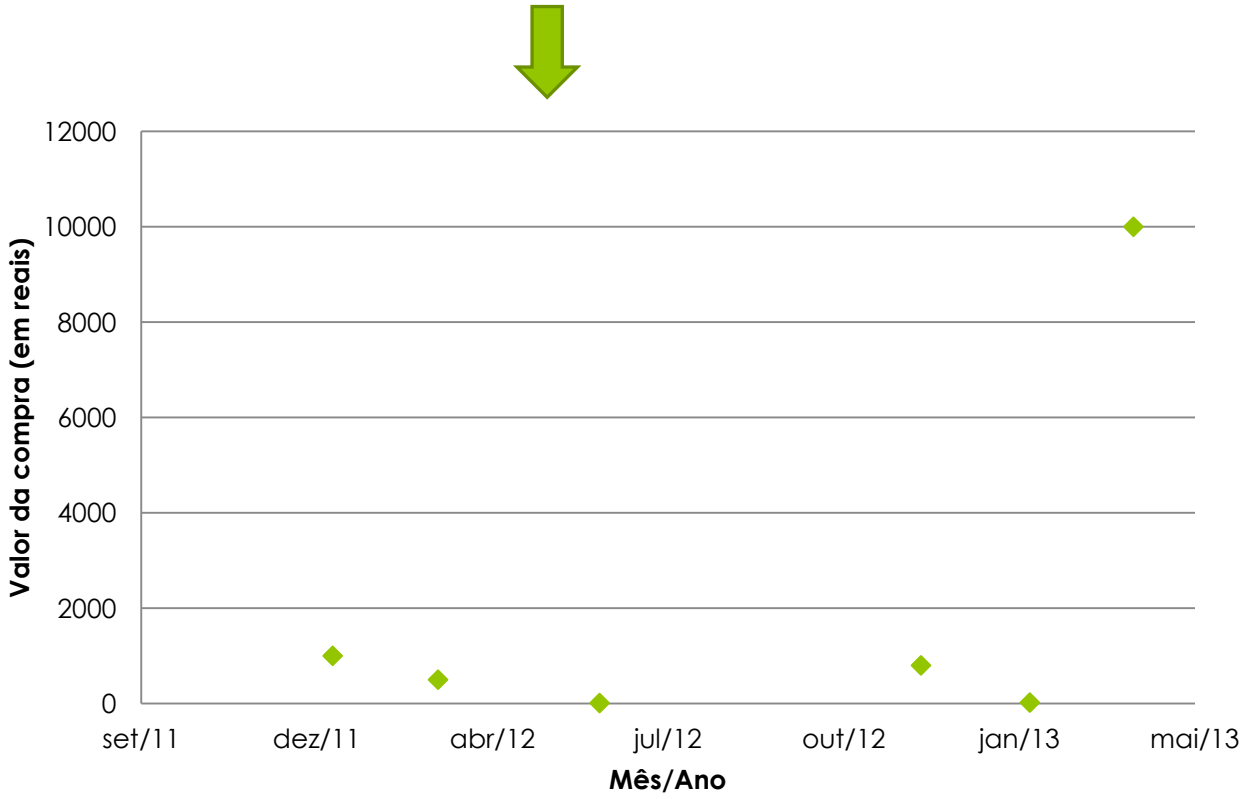
- Técnicas de Data Mining são divididas em:
  - Descrição de Classes/Conceitos
  - Mineração de padrões frequentes, associações e correlações
  - Classificação e Regressão para análise preditiva
  - Análise de agrupamento
  - Análise de outlier

Análise de Outlier

- Analisa dados com comportamento muito diferente dos demais dados
- Um Pouco de Tudo: Detectar fraudes no uso do cartão de crédito

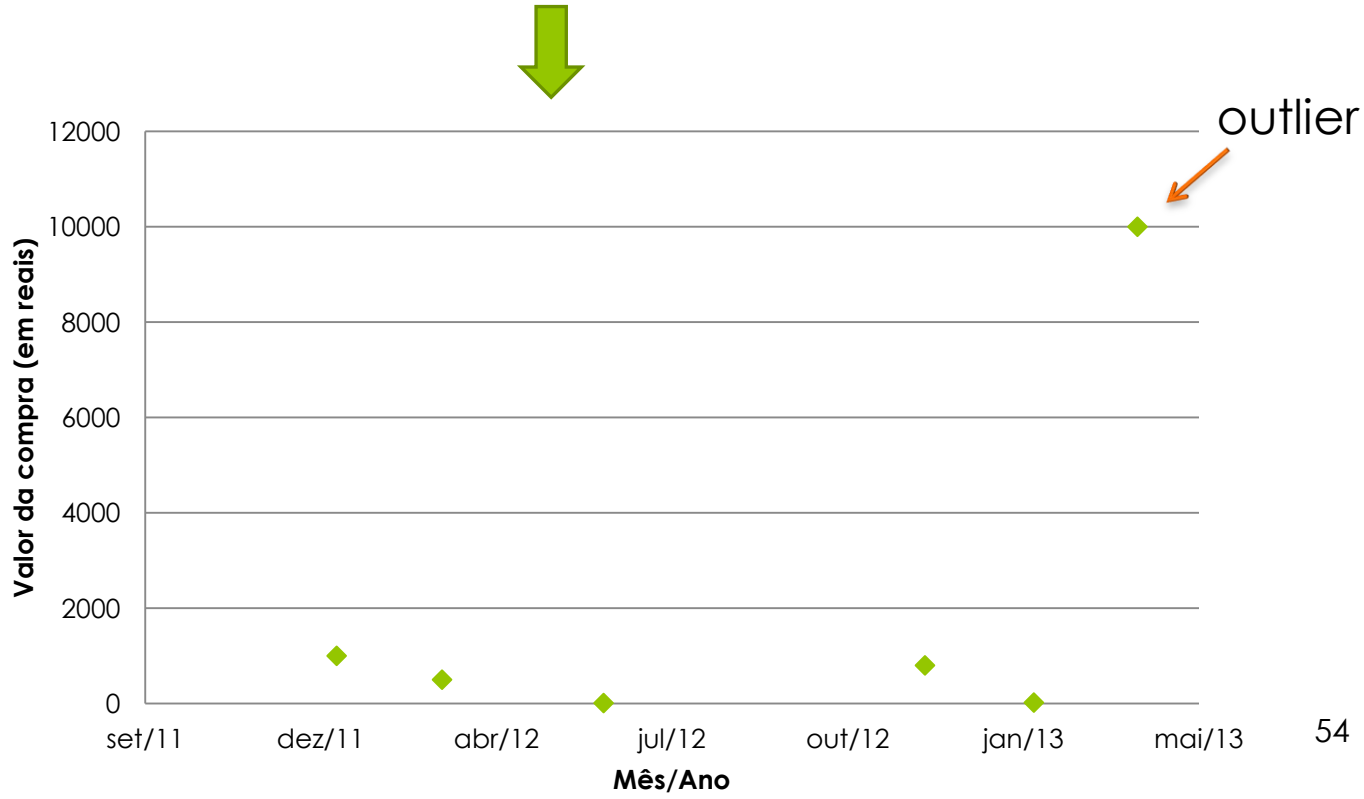
Análise de Outlier

- Um Pouco de Tudo: Detectar fraudes no uso do cartão de crédito



Análise de Outlier

- Um Pouco de Tudo: Detectar fraudes no uso do cartão de crédito



Todos os padrões são interessantes?

- Não!
- Milhões de padrões podem ser gerados e pequena fração de padrões interessam ao usuário
- Padrão interessante:
  - Facilmente compreendido por humanos
  - Válido com um determinado grau de certeza
  - Potencialmente útil
  - Novo

  ou
  - Valida a hipótese do usuário

Exercício 2

- Dentre os conceitos de Data Mining apresentados, quais conceitos você utilizaria no banco de dados da sua rede social? Justifique.

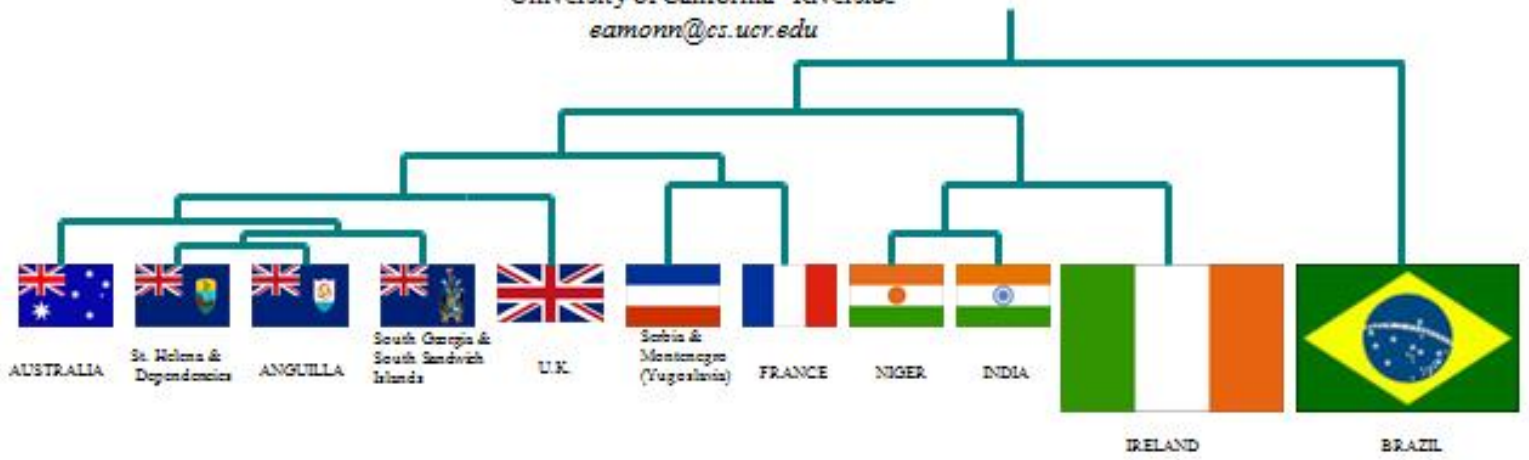
UCR UNIVERSITY OF CALIFORNIA, RIVERSIDE

A Gentle Introduction to Machine Learning and Data Mining for the Database Community

Dr Eamonn Keogh

University of California - Riverside

eamonn@cs.ucr.edu



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
  - Maturing

Query Examples

Database

- Find all customers that live in Boa Vista
- 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**)

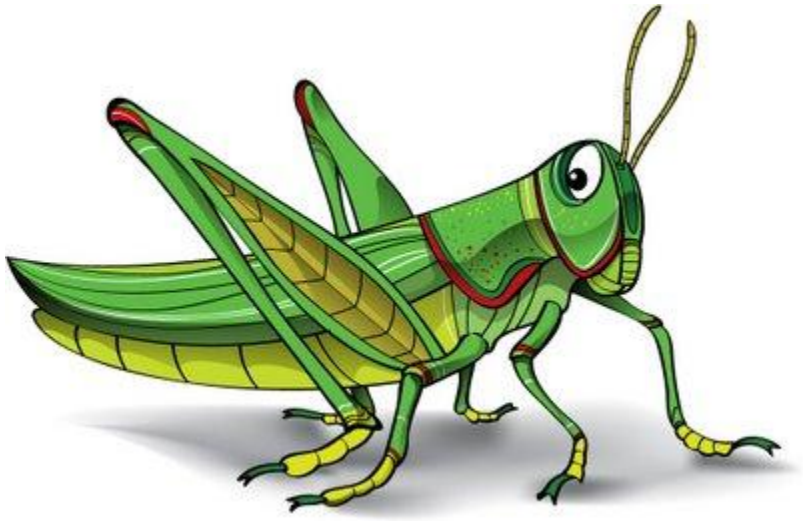
Why is Data Mining Hard?

- Scalability
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis
- Privacy issues

Fonte: Introdução a Data Mining Eamon Keogh

Data Mining x Sociedade

- Qual o impacto do data mining na sociedade?
  - Violação de privacidade, direitos autorais
- Data mining ajuda em pesquisas científicas, gerenciamento empresarial
  - Divulgação imprópria de dados, violação de privacidade
- Data mining invisível



Aplicações

Em empresas

- Diversas empresas utilizam data mining para marketing, investimento, detecção de fraude...
  - Google, Facebook, Walmart, Visa, Mastercard...

Caso Target

- Segunda maior rede varejista dos Estados Unidos
- "Aumentou alguns bilhões de dólares no seu faturamento anual, apenas criando estratégias de venda com base nas informações extraídas da mineração de dados."
- A Target sabia que uma adolescente estava grávida antes mesmo dos pais dela

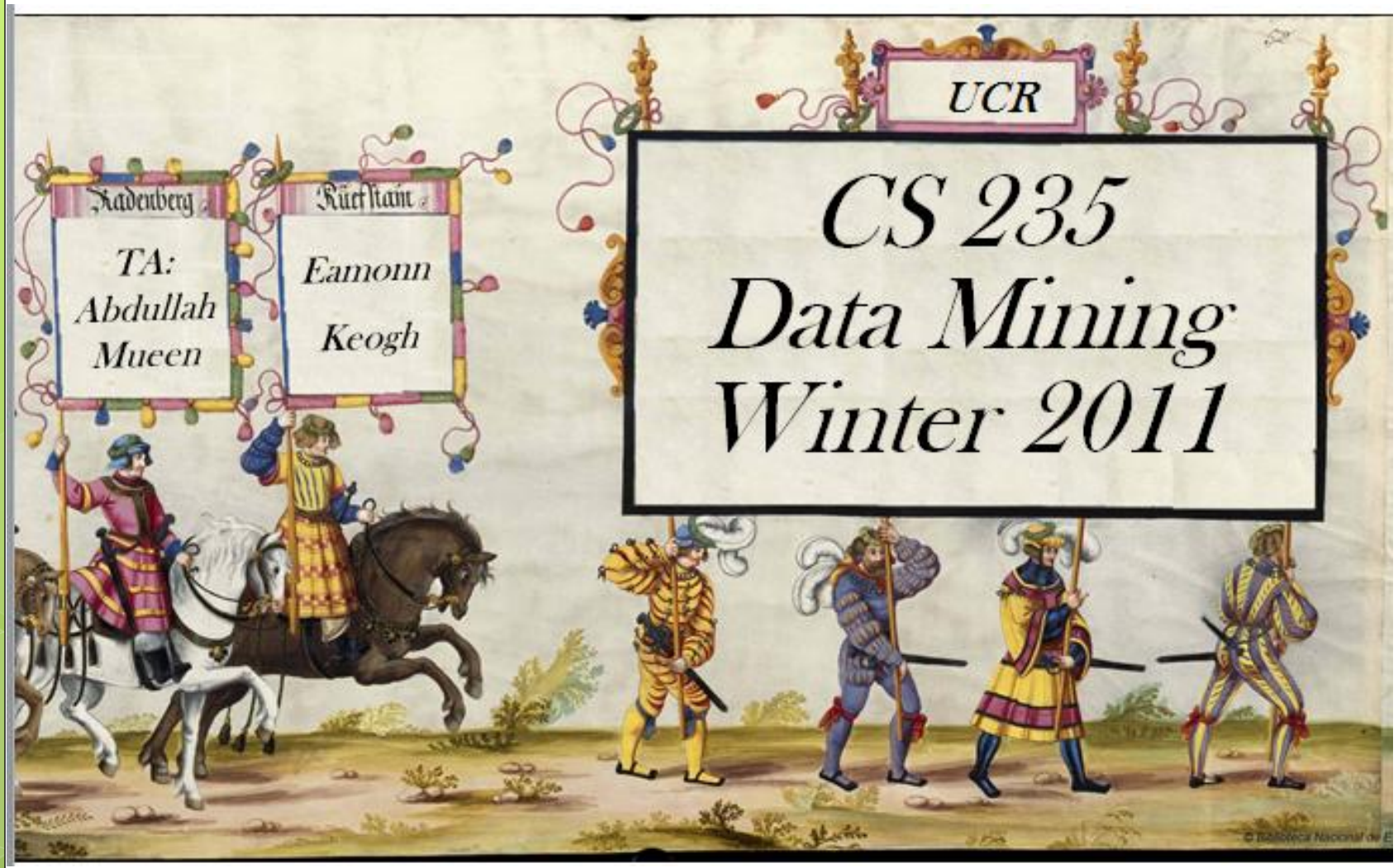
Caso Target

- "Conforme o computador analisava os dados, ele foi capaz de identificar cerca de 25 produtos que, quando analisados em conjunto, lhe permitiram atribuir a cada cliente uma pontuação de "previsão de gravidez". Mais importante, ele também poderia estimar a data do parto para dentro de um pequeno intervalo de tempo, assim a Target poderia enviar cupons programados para estágios muito específicos de sua gravidez."

Fonte: http://tecnoblog.net/151635/potencial-whatsapp-mineracao-de-dados/

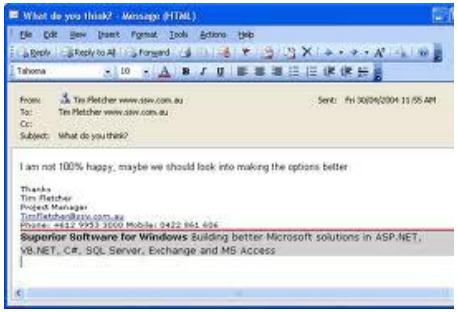
Em pesquisas

- Eamon Keogh
  - Mineração de séries temporais
  - Classificação de insetos, folhas....
  - http://www.cs.ucr.edu/~eamonn/


UCR
CS 235
Data Mining
Winter 2011
TA:
Abdullah
Mueen
Eamonn
Keogh

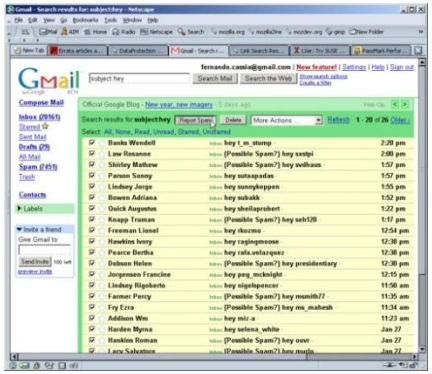
The Classification Problem

Given a collection of annotated data…

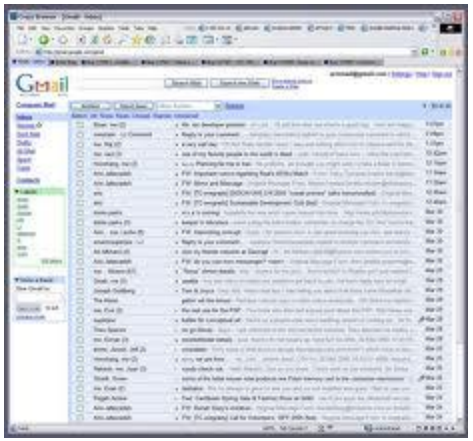


Spam or **email**?

spam



email



The Classification Problem

Given a collection of annotated data…

Spanish or **Polish**?

Spanish



Polish

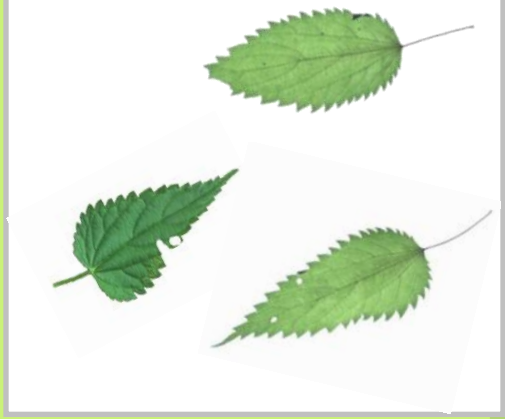



The Classification Problem

Given a collection of annotated data…

Stinging Nettle or **False Nettle**?

Stinging Nettle



False Nettle

The Classification Problem

Given a collection of annotated data…

Greek

Gunopulos
Papadopoulos
Kollios
Dardanos

Irish

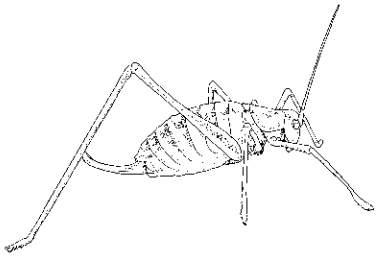
Keogh
Gough
Greenhaugh
Hadleigh

Tsotras

Greek or Irish?

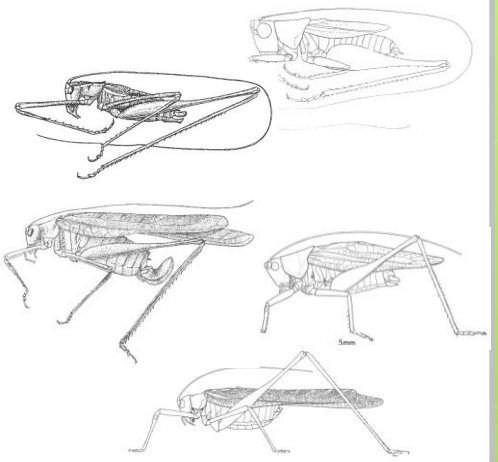
The Classification Problem
(informal definition)

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

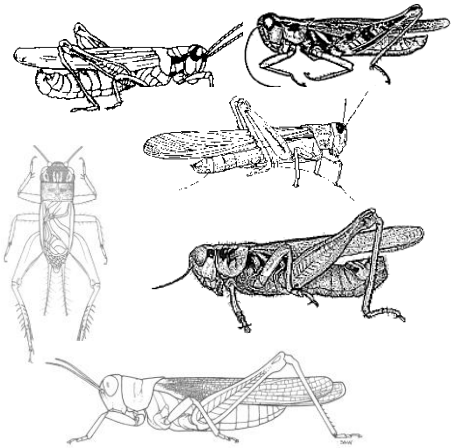


Katydid or **Grasshopper**?

Katydids



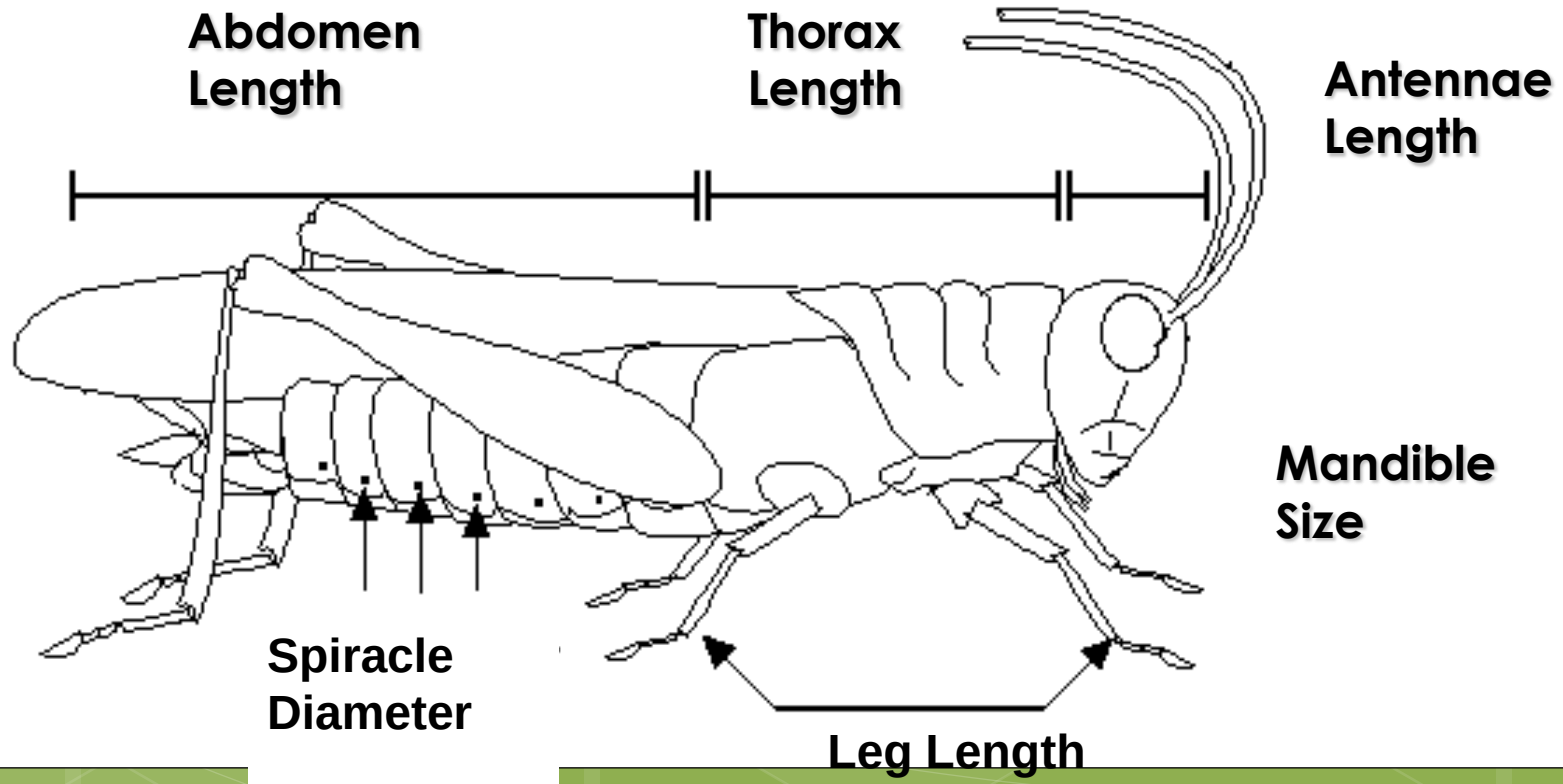
Grasshoppers



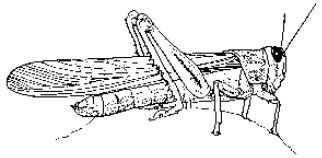
For any domain of interest, we can measure *features*

Color {Green, Brown, Gray, Other}

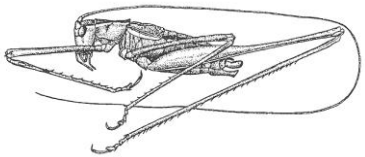
Has Wings?

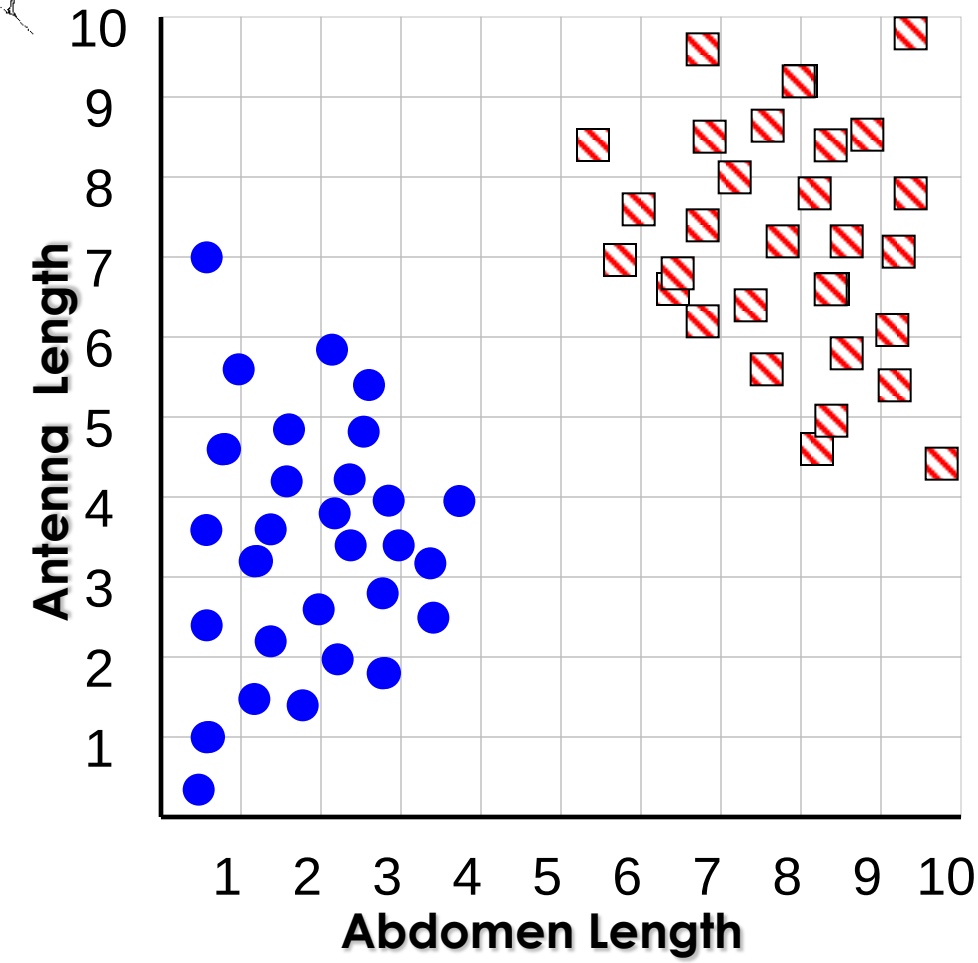


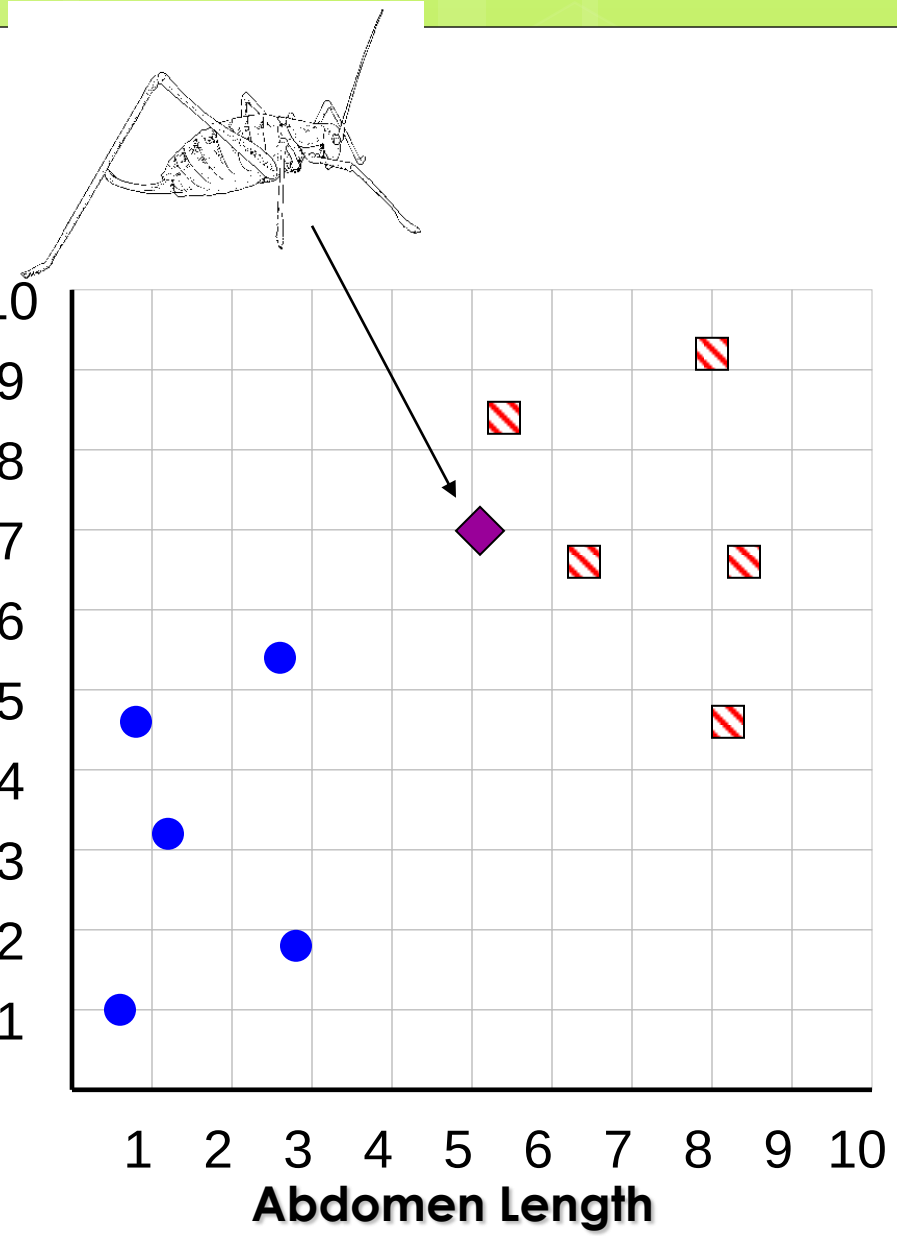
Grasshoppers



Katydids



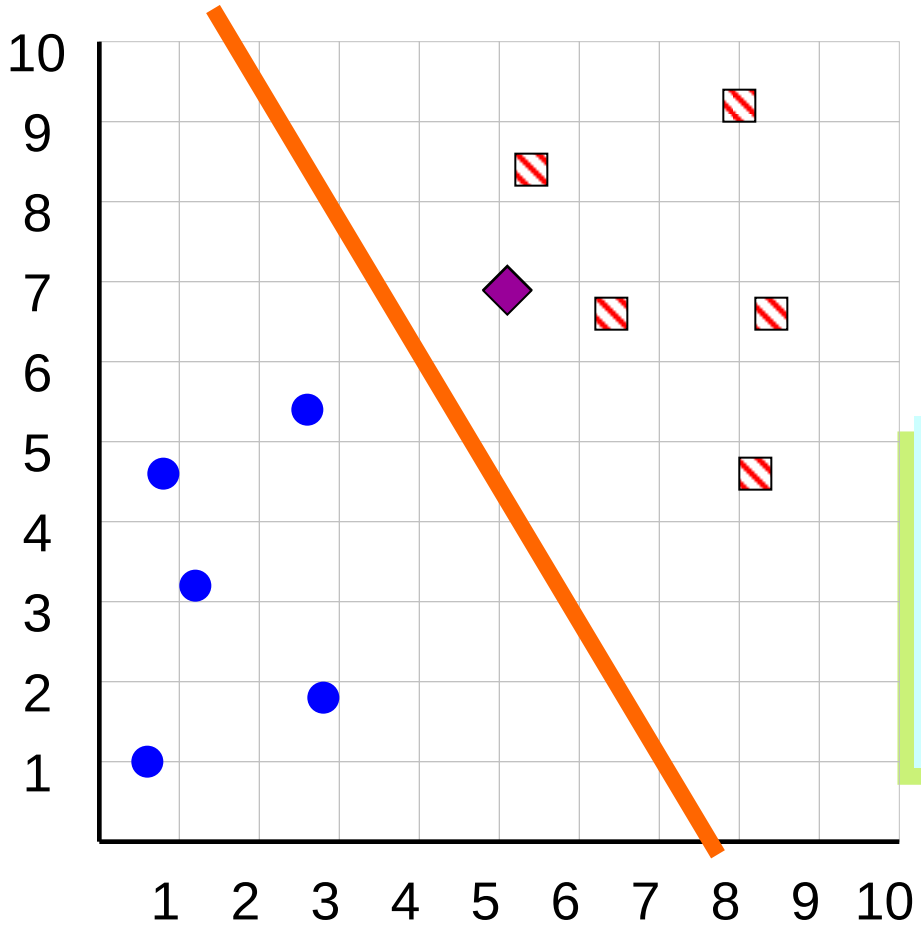




Katydids

Grasshoppers

Simple Linear Classifier

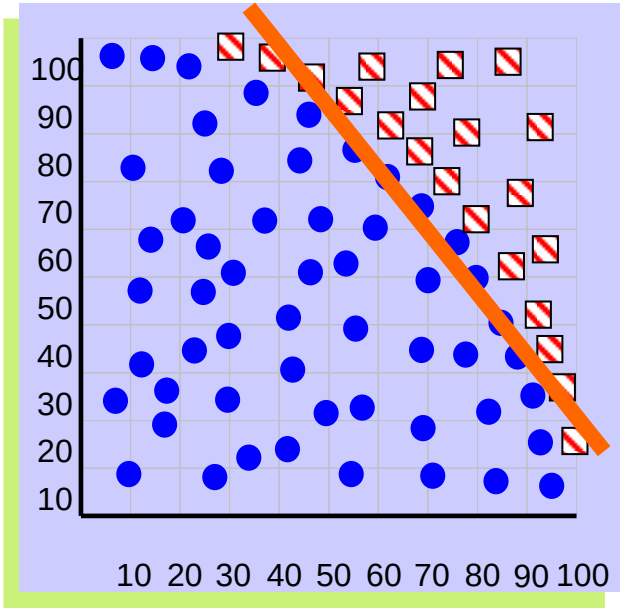


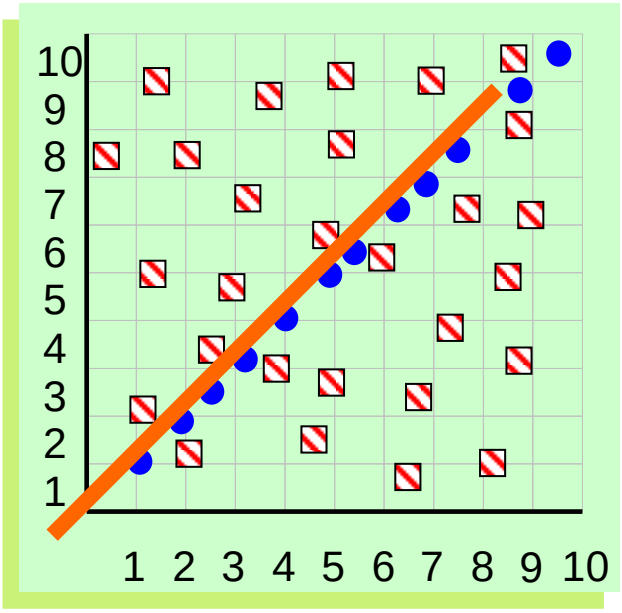
R.A. Fisher
1890-1962

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

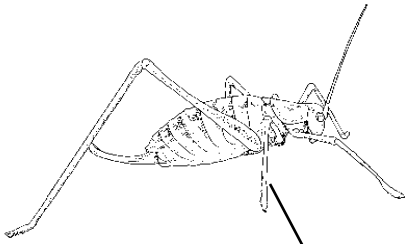
Katydids
Grasshoppers

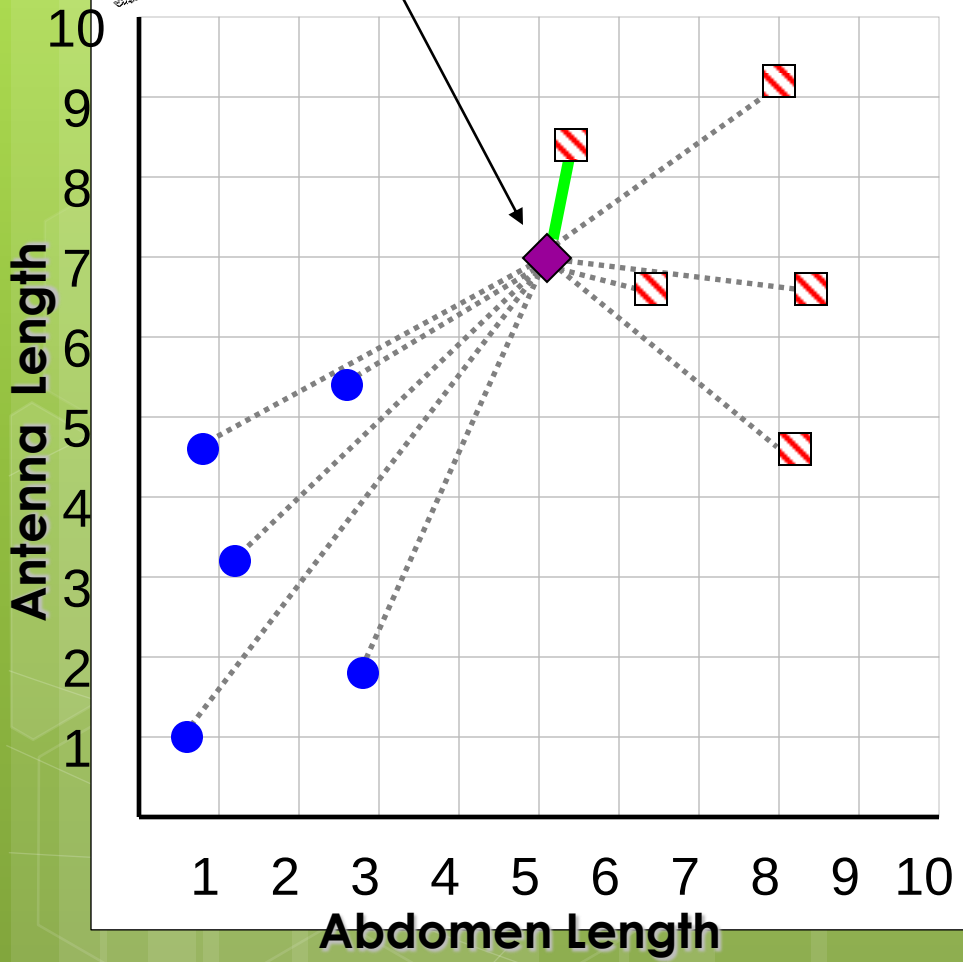
Problemas


100
90
80
70
60
50
40
30
20
10
10 20 30 40 50 60 70 80 90 100

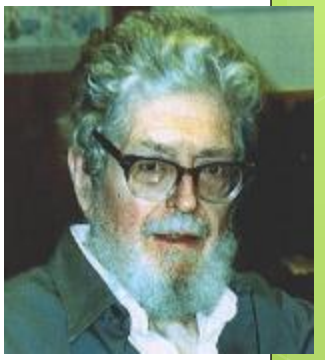

10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10

Nearest Neighbor Classifier





Evelyn Fix
1904-1965



Joe Hodges
1922-2000

If the **nearest** instance to the previously unseen instance **is a** **Katydid**

class is **Katydid**

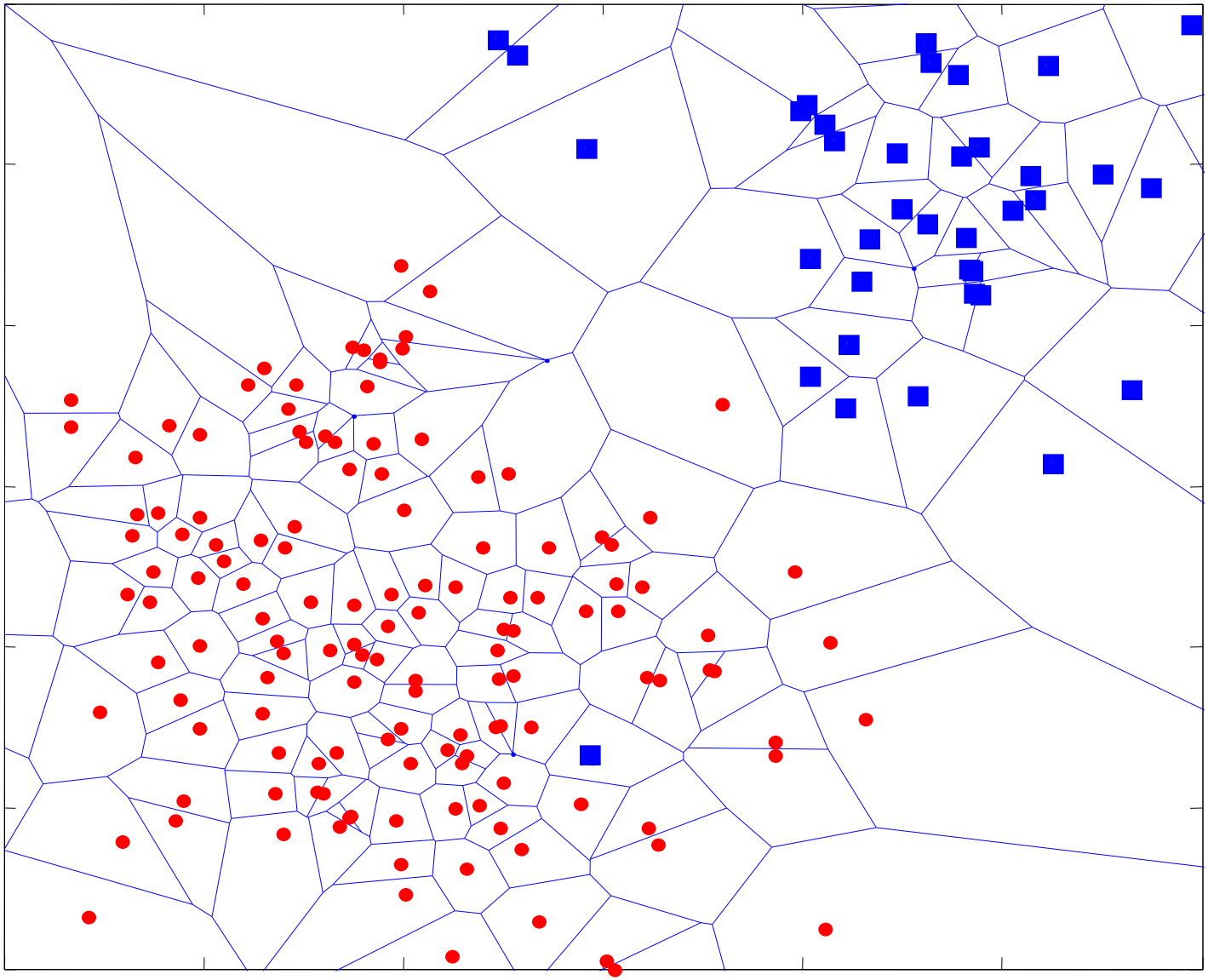
else

class is **Grasshopper**

Katydids

Grasshoppers

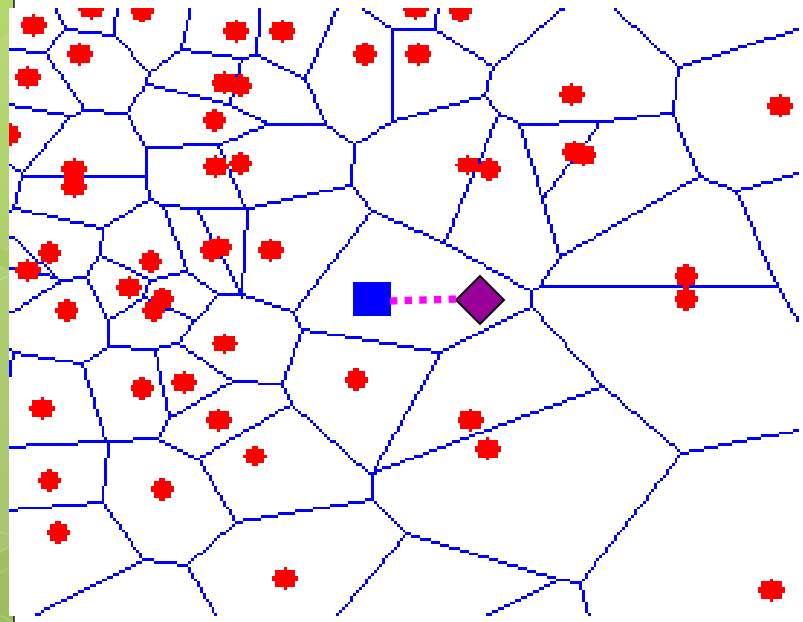
The nearest neighbor algorithm is sensitive to outliers…



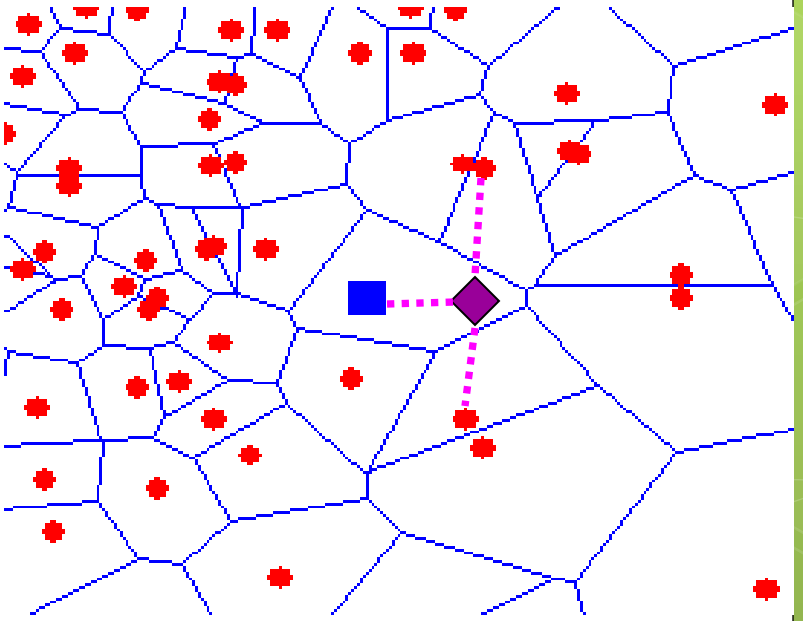
The solution is to…

We can generalize the nearest neighbor algorithm to the K- nearest neighbor (KNN) algorithm.

We measure the distance to the nearest K instances, and let them vote. K is typically chosen to be an odd number.



K = 1



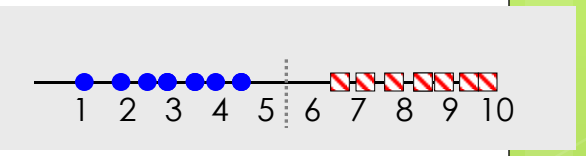
K = 3

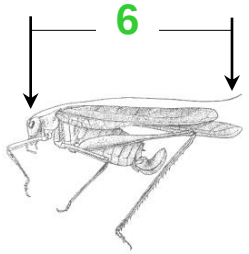
The nearest neighbor algorithm is sensitive to irrelevant features…

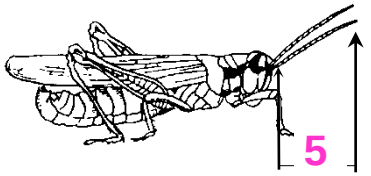
Suppose the following is true, if an insects antenna is longer than 5.5 it is a **Katydid**, otherwise it is a **Grasshopper**.

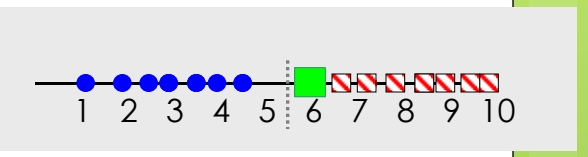
Using just the antenna length we get perfect classification!

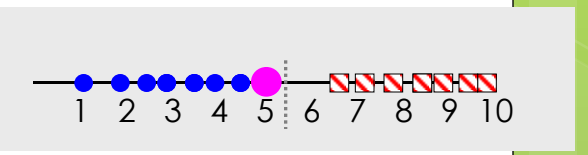
Training data

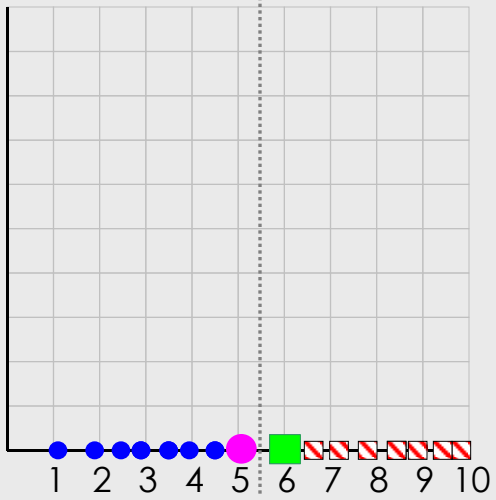


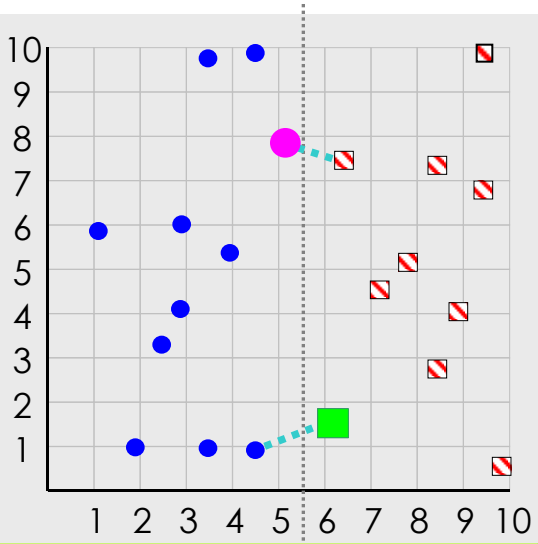












Suppose however, we add in an **irrelevant** feature, for example the insects mass.

Using both the antenna length and the insects mass with the 1-NN algorithm we get the wrong classification!

Algumas ferramentas

"Weka is a collection of machine learning algorithms for data mining tasks"
http://www.cs.waikato.ac.nz/ml/weka/

"The Apache Mahout™ project's goal is to build a scalable machine learning library. […]
Currently Mahout supports mainly three use cases: […]Recommendation, Classification and Clustering"
https://mahout.apache.org/

Data Mining software to bussiness
http://www.pentaho.com/

Referências

- Fayyad, Ussama; Piatetsky-Shapiro, Gregory; SMYTH, Padhraic. **From Data Mining to knowledge Discovery in Databases.** AI Magazine, vol 17, nº3. AAAI, 1996.
- Fayyad, Ussama. **Data Mining and Knowledge Discovery in Databases:** Implications for Scientific Databases. SSDM, 1997.
- HAO, Yuan; CAMPANA, Bilson; KEOGH, Eamonn. **Monitoring and Mining Insect Sounds in Visual Space.** SDM 2012.
- HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** Concepts and Techniques. 3ª ed. Elsevier, 2011.
- KEOGH, Eamonn. **Introduction to Data Mining.** Apresentação. Data Mining Winter 2011.
- KEOGH, Eamonn. **A Gentle Introduction to Machine Learning and Data Mining for the Database Community**

Obrigado!

Exercício 1- Resposta

- Cite 2 padrões frequentes considerando o banco de dados abaixo.

Transação 1	Pão, leite, queijo, presunto, desodorante, feijão
Transação 2	Achocolatado, pão, leite
Transação 3	Cebola, laranja, salsa, manga
Transação 4	Carne, presunto, ovos, queijo, pão
Transação 5	Chocolate, pipoca, refrigerante, leite
Transação 6	Caneta, bala, fralda, queijo, leite, pão

Exercício 1- Resposta

Compra (Cliente, Pão) => Compra (Cliente, **Leite**) [suporte: **50%** confiança: **75%**]

Compra (Cliente, Queijo) => Compra (Cliente, **Presunto**) [suporte: **33%** confiança: **66%**]

Exercício 2 - Respota

- Possíveis respostas:
  - Agrupar usuários de acordo com o gênero musical escolhido. Sugestões de bandas similares podem ser feitas para o grupo.
  - Classificar as bandas de acordo com a popularidade.