Introdução à Mineração de Dados com Aplicações em Ciências Espaciais

Escola de Verão do Laboratório Associado de Computação e Matemática Aplicada

Rafael Santos

Programa

- ***Dia 1:*** Apresentação dos conceitos de mineração de dados, motivação e alguns exemplos.
- ***Dia 2:*** Algoritmos de classificação supervisionada e aplicações.
- ***Dia 3:*** Algoritmos de classificação não-supervisionada e aplicações. Algoritmos de mineração de associações.
- ***Dia 4:*** Visualização e mineração de dados. Outros algoritmos e idéias. Onde aprender mais.

Agrupamento (Clusterização)

Agrupamento (Clusterização)

- Algoritmos para criação de grupos de instâncias
  - Similares entre si,
  - Diferentes de instâncias em outros grupos.
  - Não-supervisionado (?)
- Também conhecidos como algoritmos de aprendizado auto-organizado.
- Diferença entre instâncias e (protótipos de) grupos é dada por um valor: medidas de distância ou similaridade / dissimilaridade.

Agrupamento (Clusterização)

- Duas abordagens gerais:
  - Particionais:
    - Criam grupos de forma iterativa.
    - Reparticiona/reorganiza até atingir um limiar (tempo, erro quadrático, etc).
    - Ao terminar fornece pertinência final de instâncias a grupos.
  - Hierárquicos:
    - *Bottom-up*: cria pequenos grupos juntando as instâncias, repetindo até atingir um critério.
    - *Top-down*: considera todas as instâncias como pertencentes a um grande grupo, subdivide recursivamente este grupo.
  - Podem criar *dendogramas*: agrupamentos hierárquicos com números alternativos de grupos.

K-Médias

- Particional.
- Entrada: instâncias, medida de distância, número de grupos (K).
- Saída: centróides dos grupos, pertinência das instâncias aos grupos, métricas.
- O algoritmo tenta minimizar o erro quadrático calculado entre as instâncias e os centróides dos grupos.

K-Médias: Passos

1. Inicializamos os centróides dos *K* grupos.
2. Marcamos cada instância como pertencente ao grupo (centróide) mais próximo.
3. Recalculamos os centróides dos grupos considerando as pertinências.

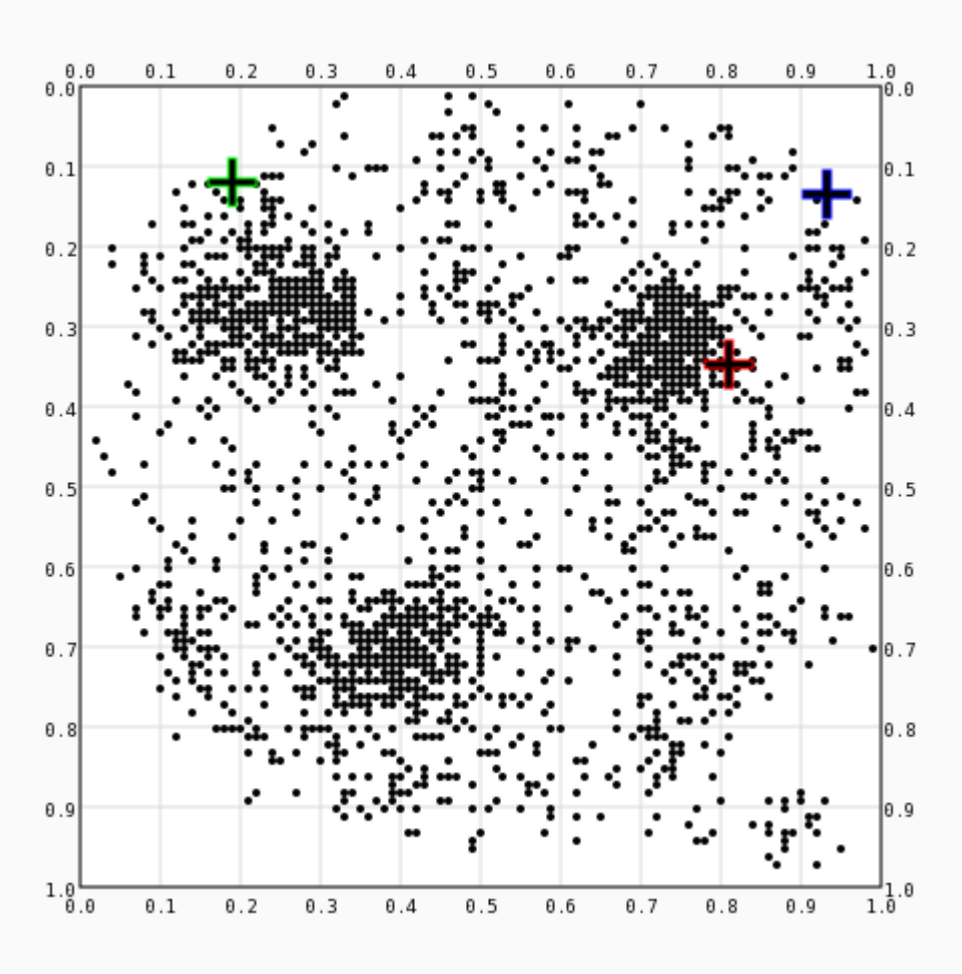
$$v_i = \frac{1}{n_i} \sum_{x_k \in C_i} x_k$$

4. Recalculamos o erro quadrático total.

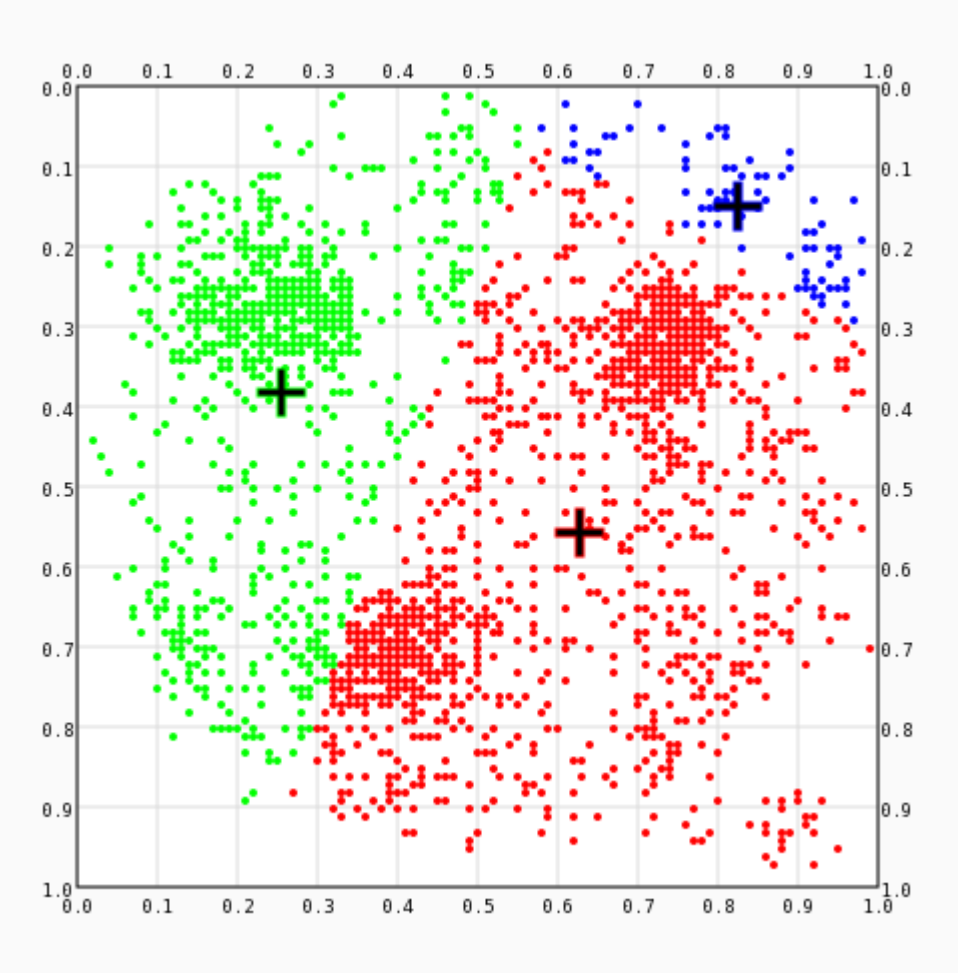
$$J = \sum_{k=1}^{n} \sum_{x_k \in C_i} |x_k - v_i|^2$$

5. Verificamos condições de parada e repetimos a partir do passo 2.

K-Médias

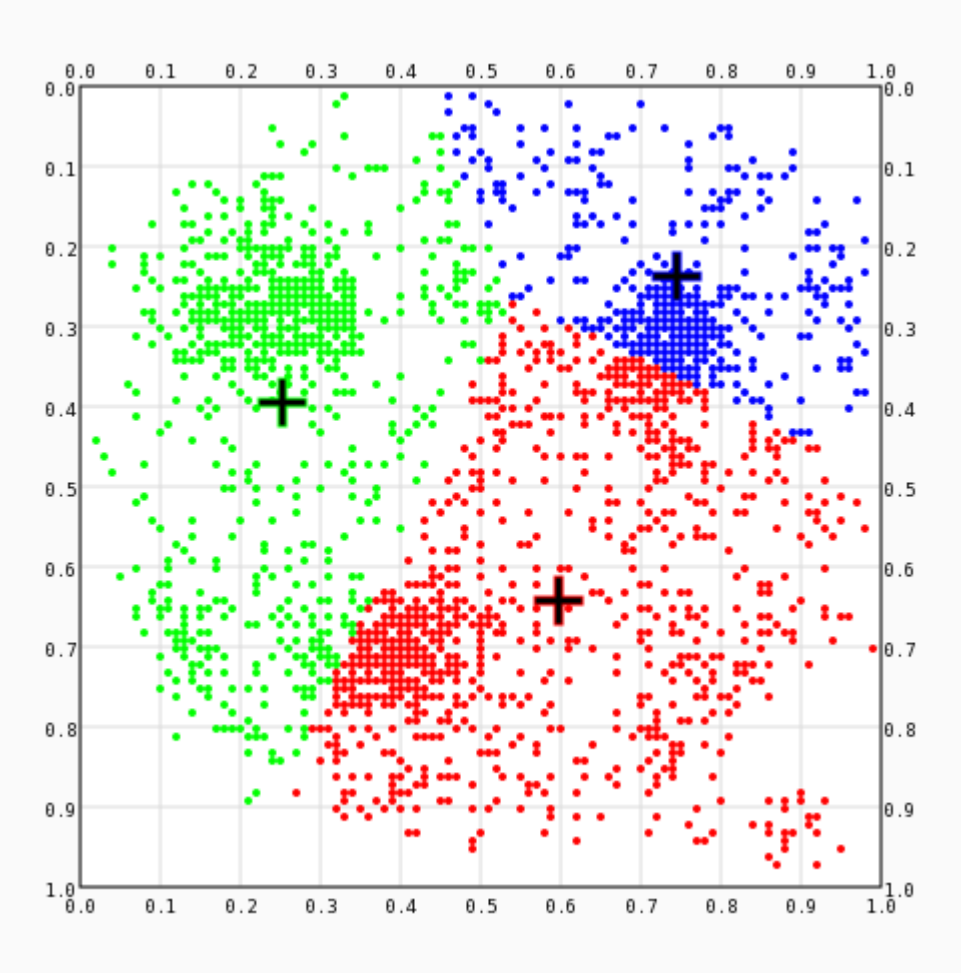


0

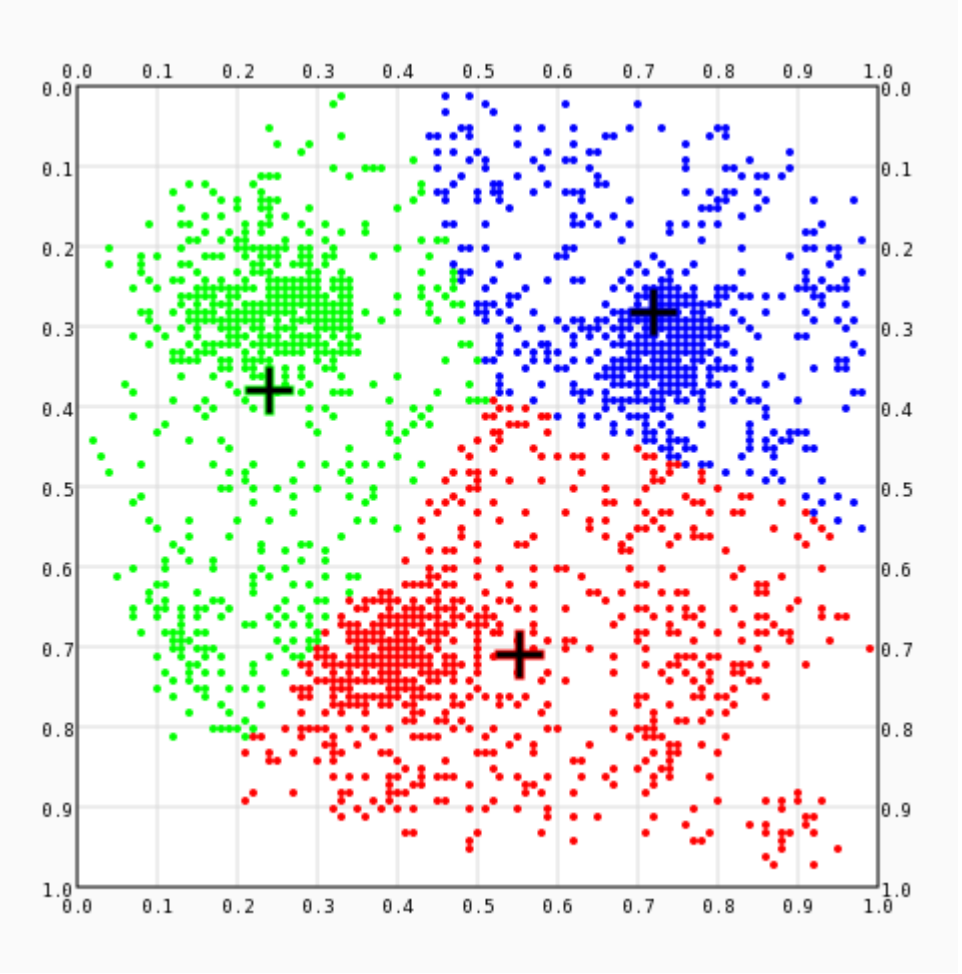


1

K-Médias

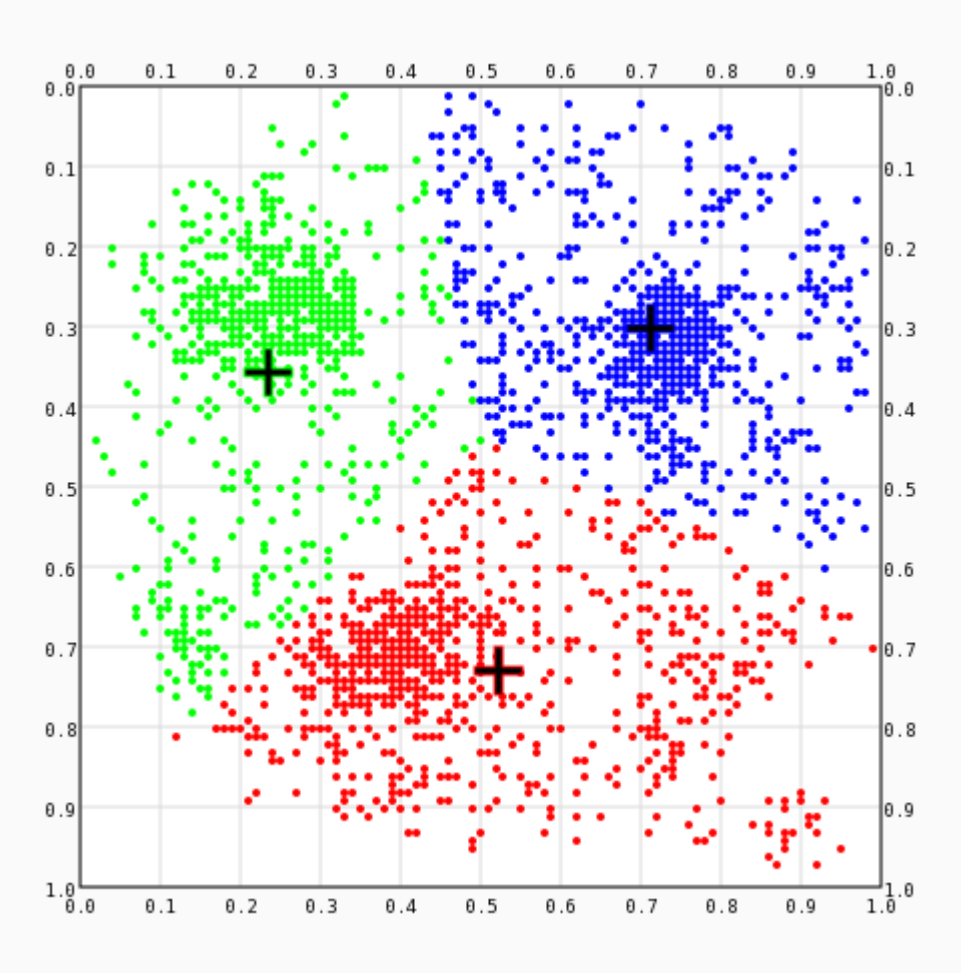


2

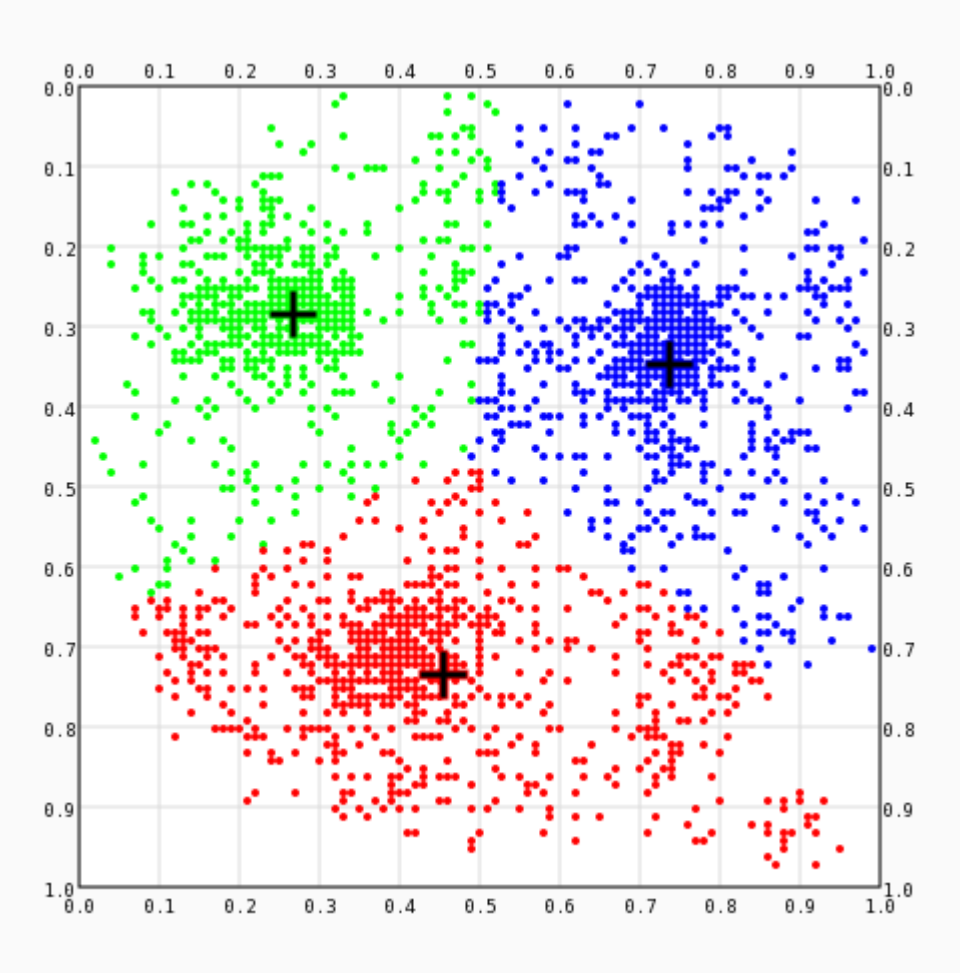


3

K-Médias



4



10

K-Médias

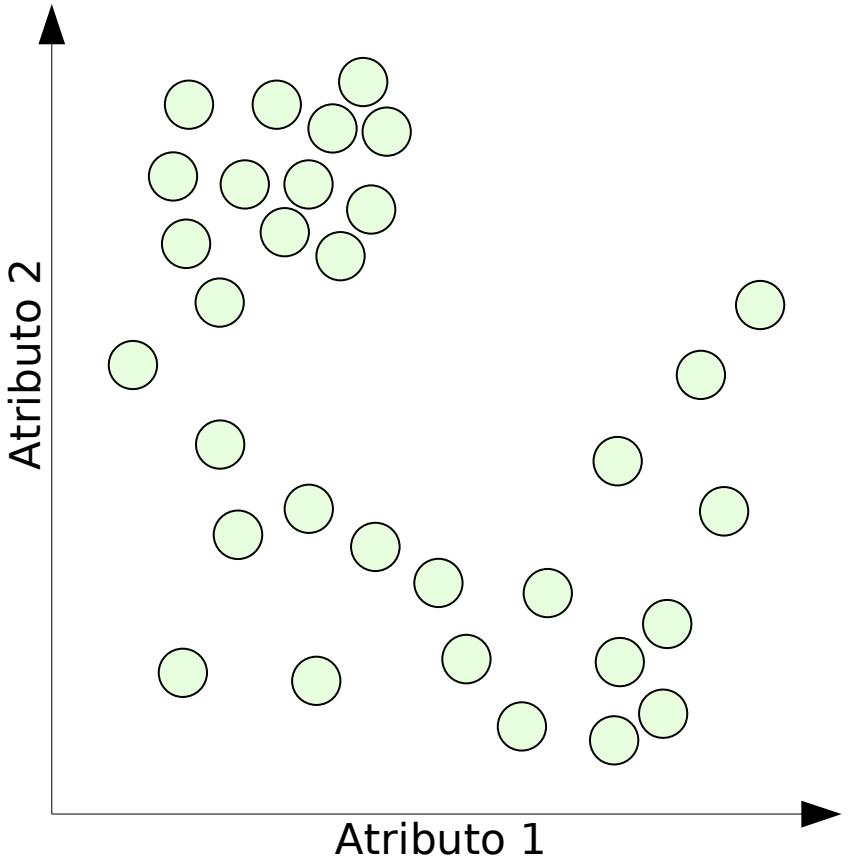
- Problemas:
  - Múltiplas iterações considerando todos os dados: problemas de performance.
  - Inicialização: como escolher centróides iniciais (impacto na convergência).
  - Converge para um mínimo local.
  - Singularidades: grupos sem instâncias relacionadas.
    - Não podemos calcular seus centróides.
  - Escolha de *K*?
    - Existe um *K'* melhor do que o *K*?

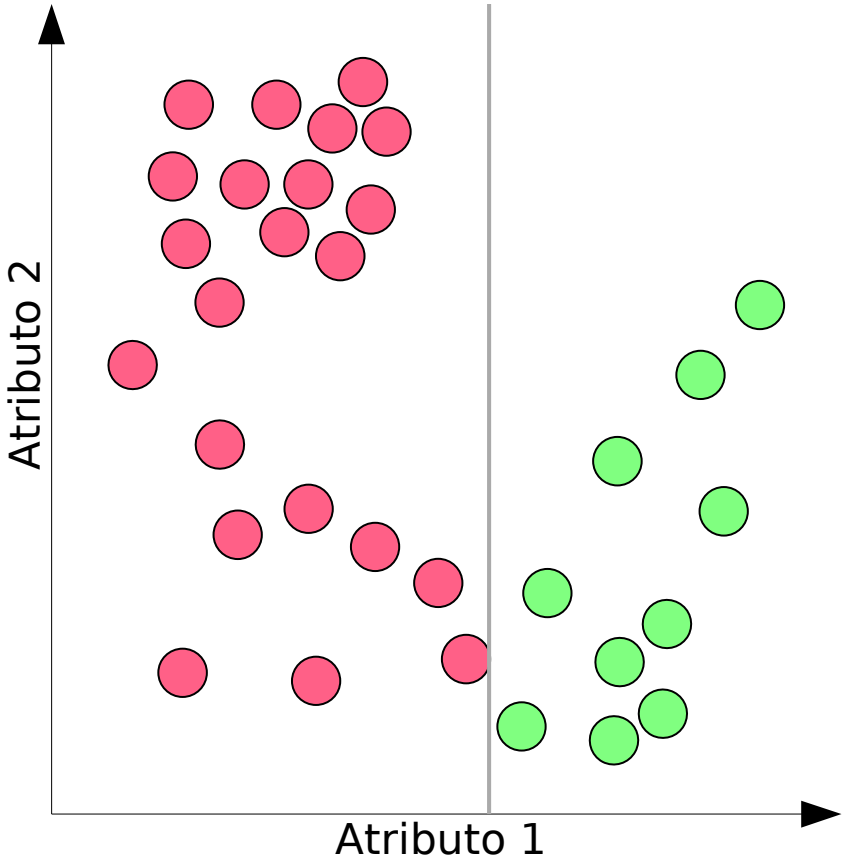
Isodata

- K-Médias mais heurísticas: nada de pequenos grupos, quebraremos grupos com grande variância.
- Mais complexo, demorado do que simples K-Médias.
- Mais parâmetros devem ser especificados, mas por se tratar de uma heurística, estes parâmetros podem ser aproximados.
- Descrição no livro do Carl Looney: 12 passos em 3 páginas.

Lógica Nebulosa (*Fuzzy Logic*)

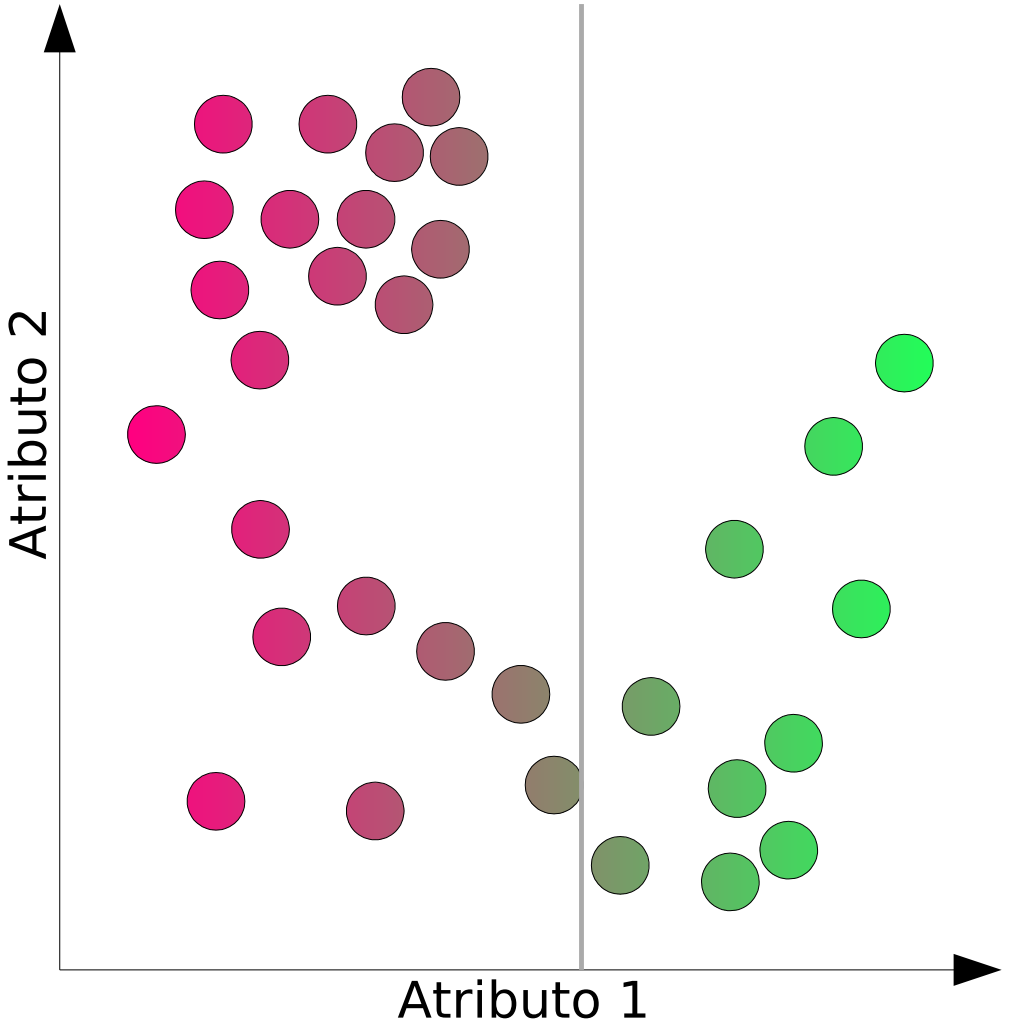
- Consideremos pertinência a classe ou grupo...





Lógica Nebulosa (*Fuzzy Logic*)

- ... não precisa ser estritamente booleana!
  - Cada instância pode pertencer a mais de uma categoria com pertinências *entre* 0 e 1.



Lógica Nebulosa (*Fuzzy Logic*)

- ... não precisa ser estritamente booleana!
  - Cada instância pode pertencer a mais de uma categoria com pertinências *entre* 0 e 1.

- Exemplo:

Instância	Classe A	Classe B	Classe C	Classe D
1	0.31	0.19	0.50	0.00
2	0.08	0.01	0.74	0.17
3	0.25	0.24	0.26	0.25
4	0.99	0.00	0.00	0.01
5	0.50	0.50	0.00	0.00

Fuzzy C-Médias

- Similar ao K-Médias, com mesmas características gerais.
- Cria uma *tabela de pertinência* de cada instância em cada grupo.
  - Tabela provê informações interessantes!

Instância	Classe A	Classe B	Classe C	Classe D
1	0.31	0.19	0.50	0.00
2	0.08	0.01	0.74	0.17
3	0.25	0.24	0.26	0.25
4	0.99	0.00	0.00	0.01
5	0.50	0.50	0.00	0.00

Fuzzy C-Médias: Passos

1. Inicializamos a tabela de pertinência.
2. Calculamos os centróides a partir das pertinências com

$$v_i = \frac{\sum_{k=1}^{n} \mu_{ik}^{m} x_{ik}}{\sum_{k=1}^{n} \mu_{ik}^{m}}$$

3. Calculamos a tabela de pertinências a partir dos centróides valores das instâncias com

$$\mu_{ik} = \frac{\left[\frac{1}{|x_k - v_i|^2}\right]^{1/(m-1)}}{\sum_{j=1}^{c} \left[\frac{1}{|x_k - v_i|^2}\right]^{1/(m-1)}}$$

4. Recalculamos a função objetivo

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^{m} |x_k - v_i|^2$$

5. Verificamos condições de parada e repetimos a partir do passo 2.

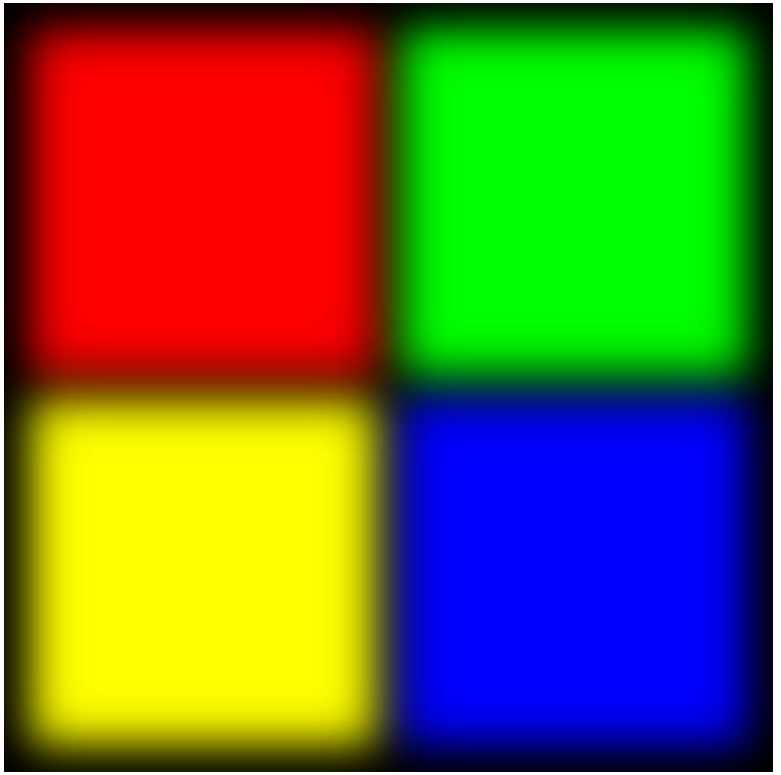
Fuzzy C-Médias

- Exemplo com C=6 e imagem Ikonos.

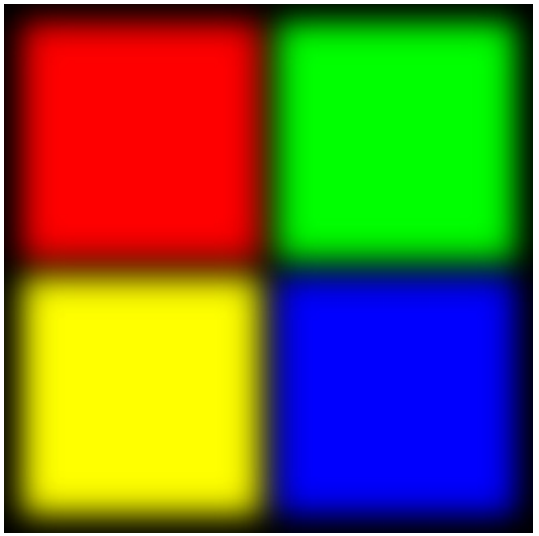
Qual valor de C?

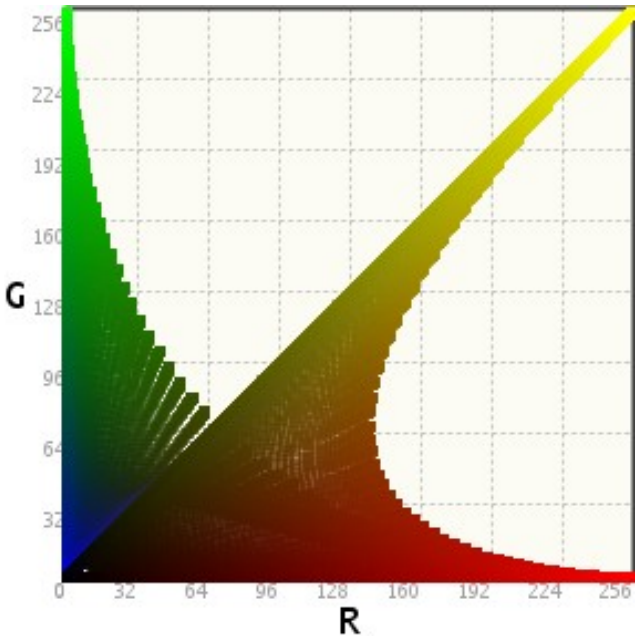
Clusters	Partition Coefficient	Partition Entropy	Compactness and Separation
2	0.677813	0.487545	0.185556
3	0.693175	0.550510	0.082218
4	0.776866	0.456778	0.029484
5	0.814956	0.398648	0.014663
6	0.785108	0.466327	0.190570
7	0.774956	0.502596	0.103595
8	0.780768	0.506613	0.046404
9	0.784015	0.508109	0.032702

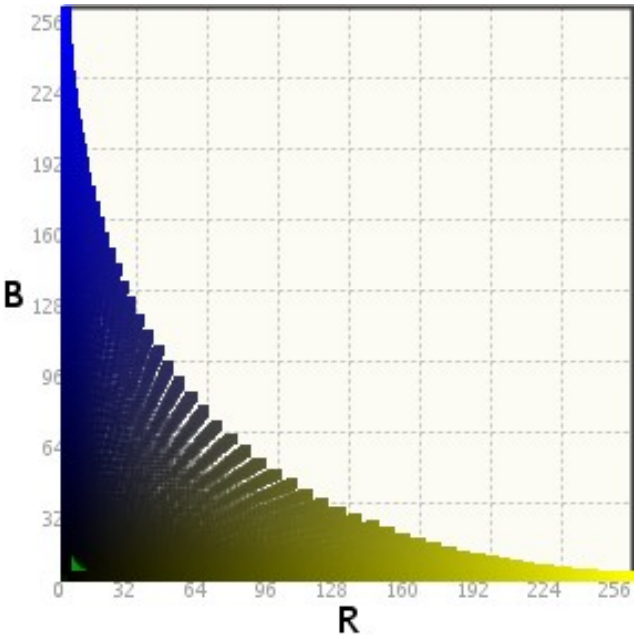
Best number of clusters:
- **according to Partition Coefficient:5**
- **according to Partition Entropy:5**
- **according to Compactness and Separation:5**

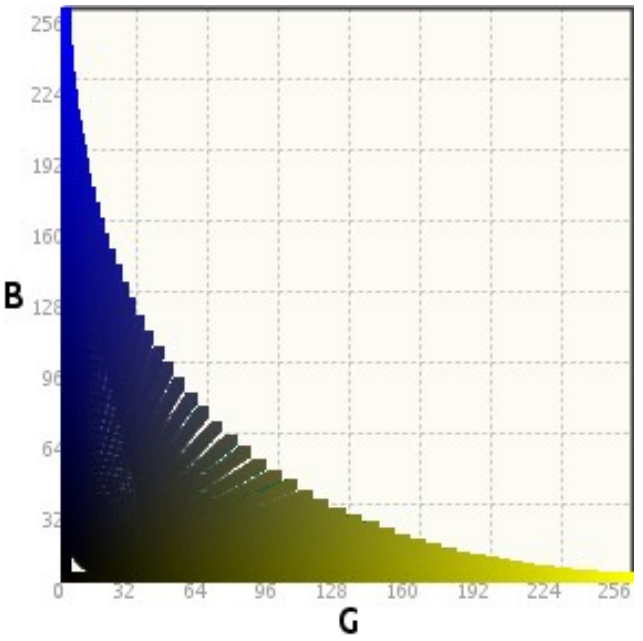


Qual valor de C?







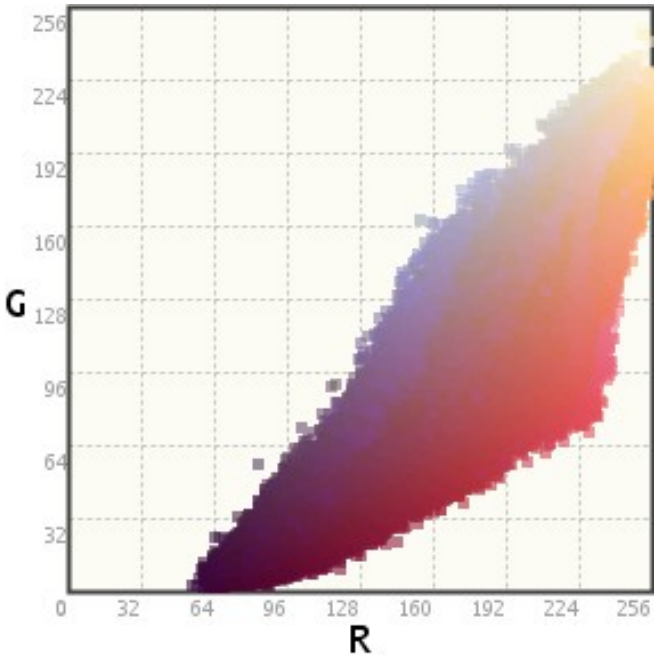


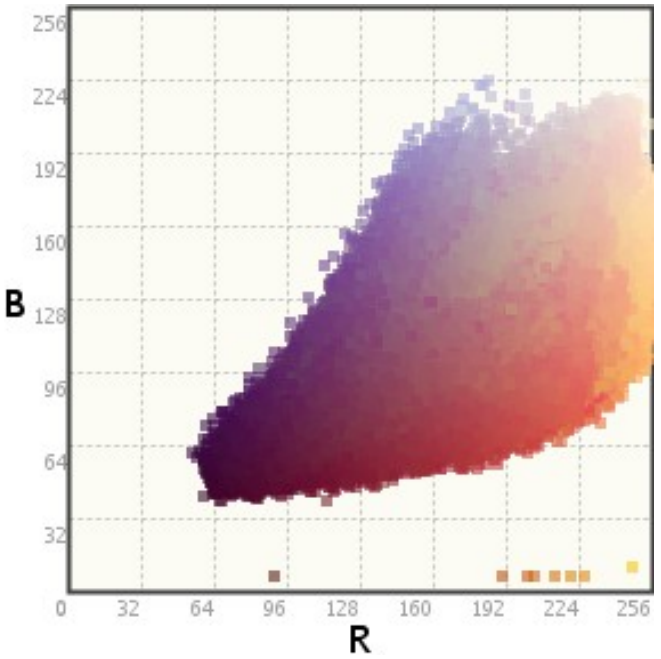
Qual valor de C?

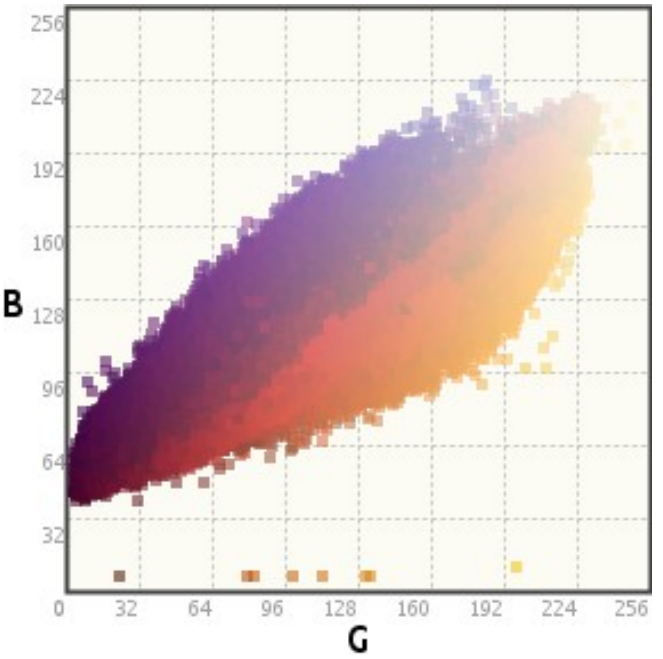
Clusters	Partition Coefficient	Partition Entropy	Compactness and Separation
2	0.809582	0.315675	0.038657
3	0.727024	0.489138	0.055242
4	0.704106	0.570761	0.088028
5	0.659179	0.683212	0.299256
6	0.607616	0.807902	0.365119
7	0.574450	0.900263	1.063374
8	0.550291	0.980936	1.300172
9	0.516148	1.062658	1.442328

Best number of clusters:
- **according to Partition Coefficient:2**
- **according to Partition Entropy:2**
- **according to Compactness and Separation:2**

Qual valor de C?

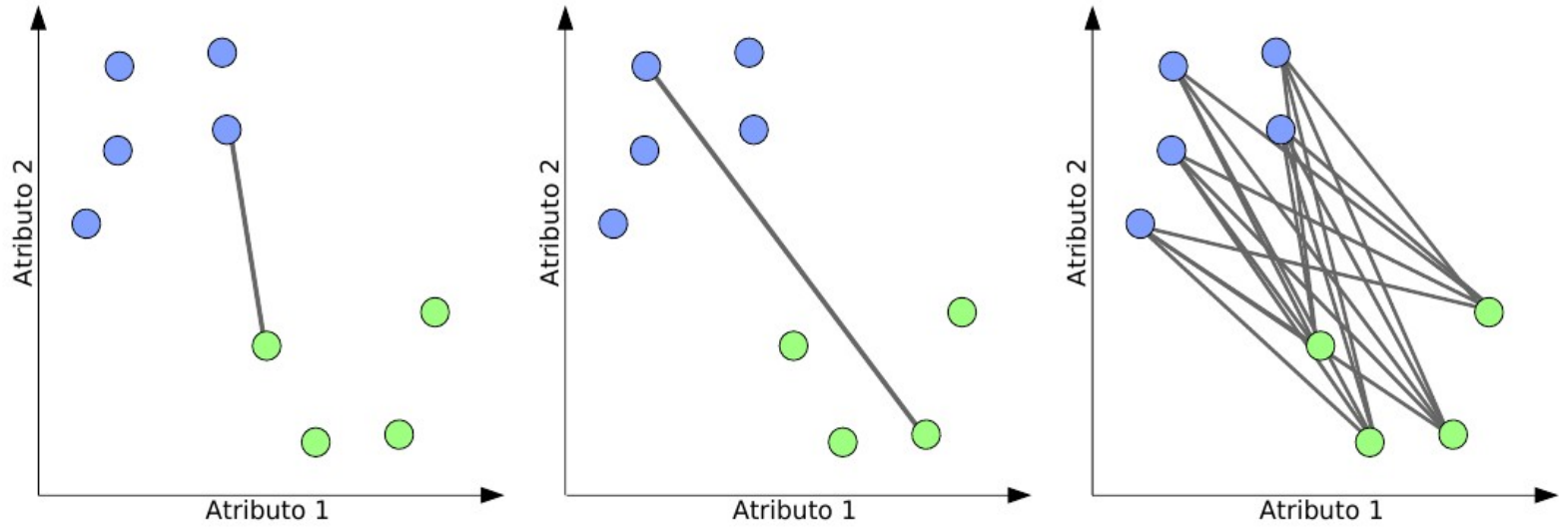




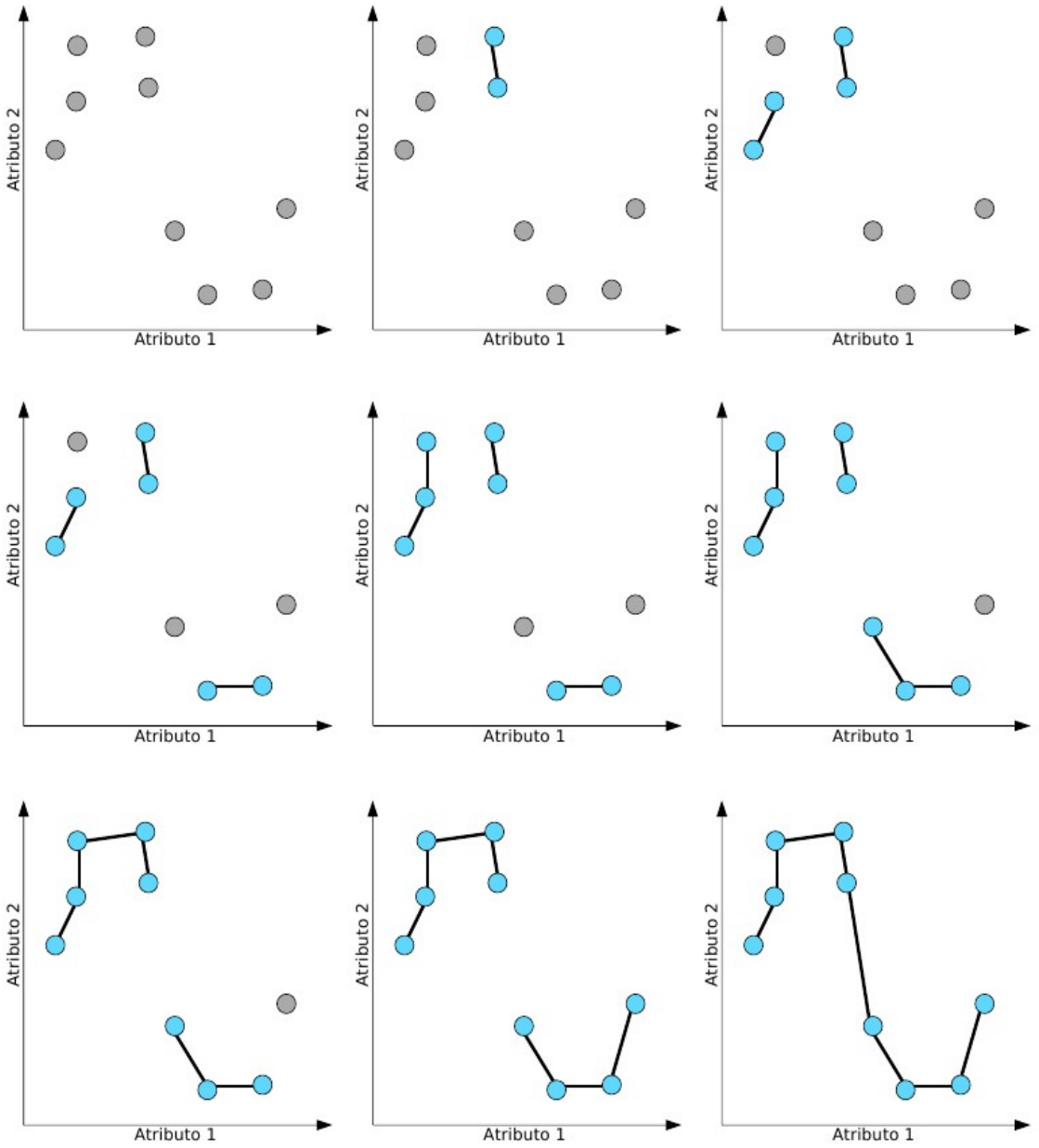


Agrupamento Hierárquico

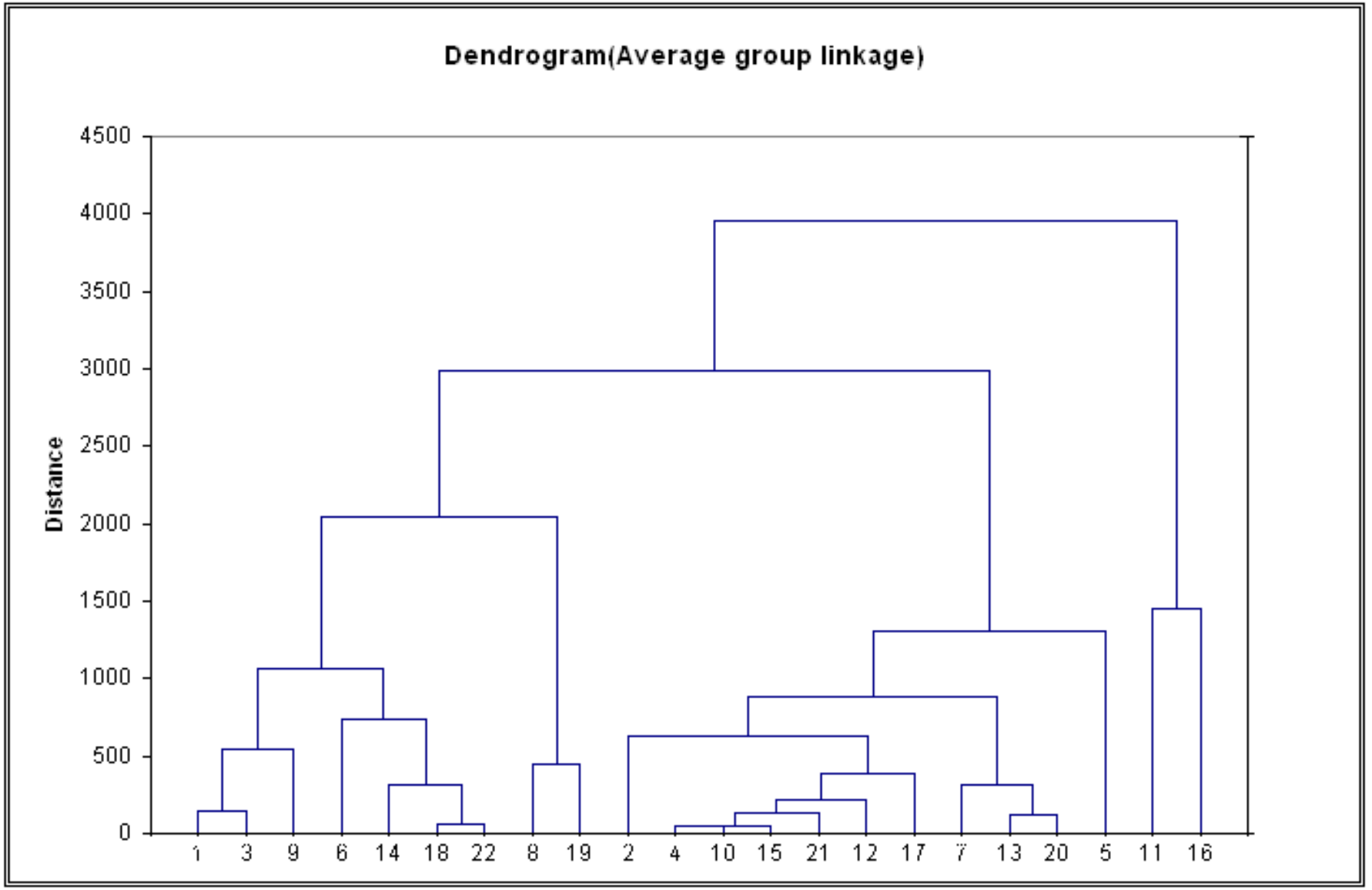
- *Bottom-up*:
  1. Considere todas as instâncias como grupos (centros sãoos valores da própria instância).
  2. Crie uma matriz de distâncias que indique a distância de cada grupo a cada outro grupo.
  3. Localize, nesta matriz, os dois grupos com menor **distância** entre eles, e efetue a união destes grupos.
  4. Se ainda houver dois ou mais grupos, volte ao passo 2.



Agrupamento Hierárquico: Simulação

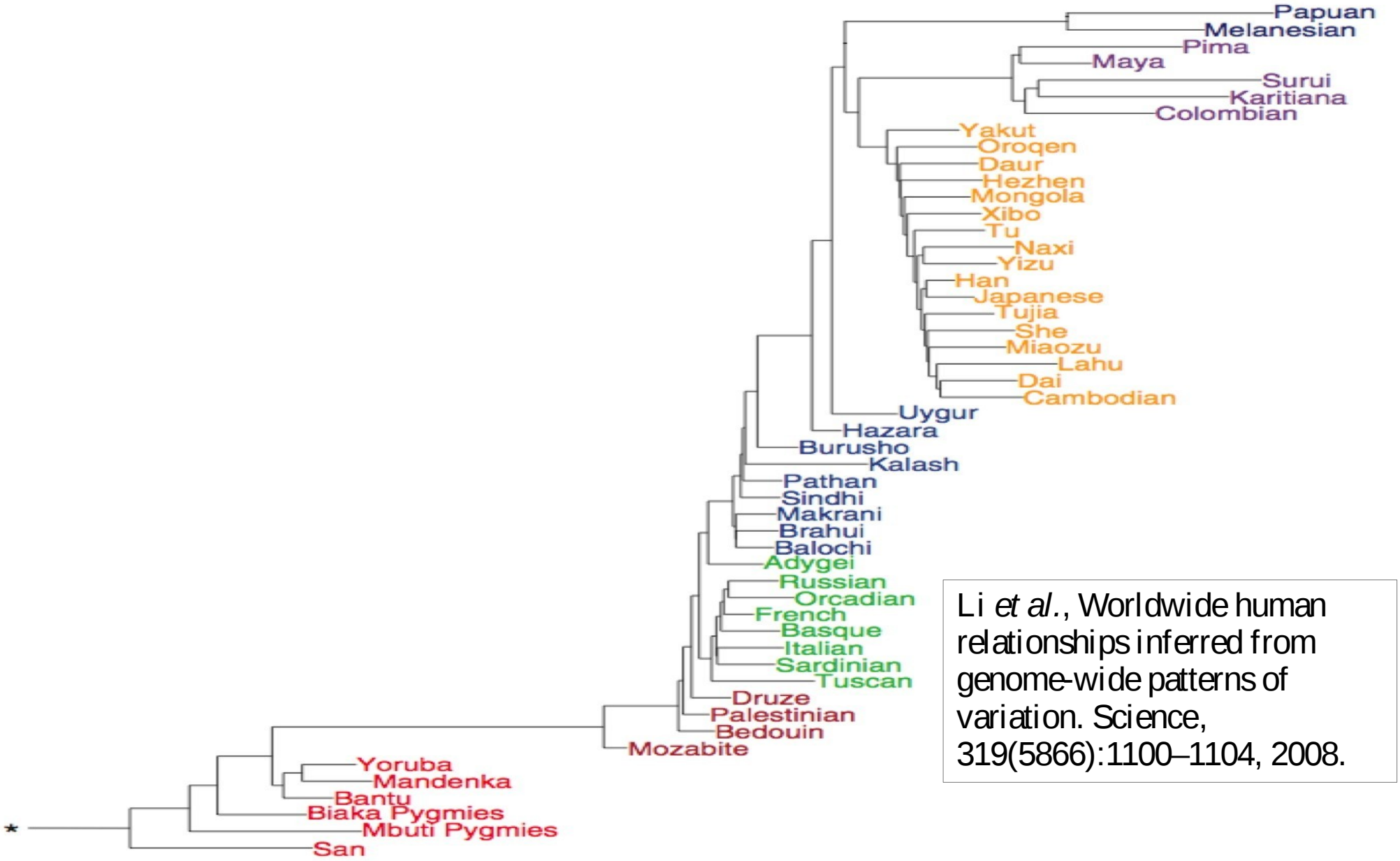


Agrupamento Hierárquico: Dendograma



Fonte: XLMiner http://www.resample.com/xlminer/

Agrupamento Hierárquico: Dendograma



Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. Science, 319(5866):1100–1104, 2008.

Agrupamento Hierárquico

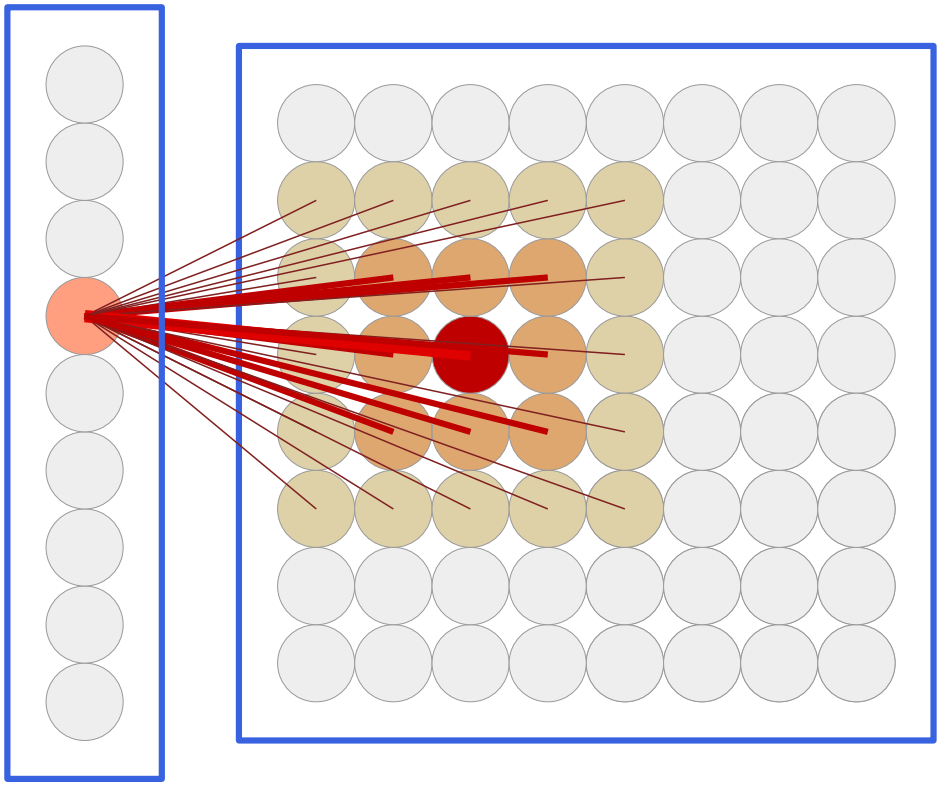
- Vantagens:
  - Número de agrupamentos pode ser determinado experimentalmente ou de forma exploratória.
  - Análise do resultado usando dendograma, que indica a estrutura hierárquica dos agrupamentos.
  - Resultado independe da ordem de apresentação dos dados.
- Problemas:
  - Matriz de distância pode consumir muita memória e seu recálculo é custoso.
  - Nem todos os elementos precisam ser recalculados.
  - Somente diagonal da matriz precisa ser armazenada.
  - O de sempre: como calcular distância não-numérica?

Self-Organizing Maps (SOMs)

- Também conhecidos como redes de Kohonen.
- Mapeiam vetores em N dimensões para 2 ou 3 dimensões, preservando topologia.
- Por extensão, usados para fazer agrupamento e classificação em fase posterior.
- Usados também para redução de dimensionalidade com manutenção de topologia.

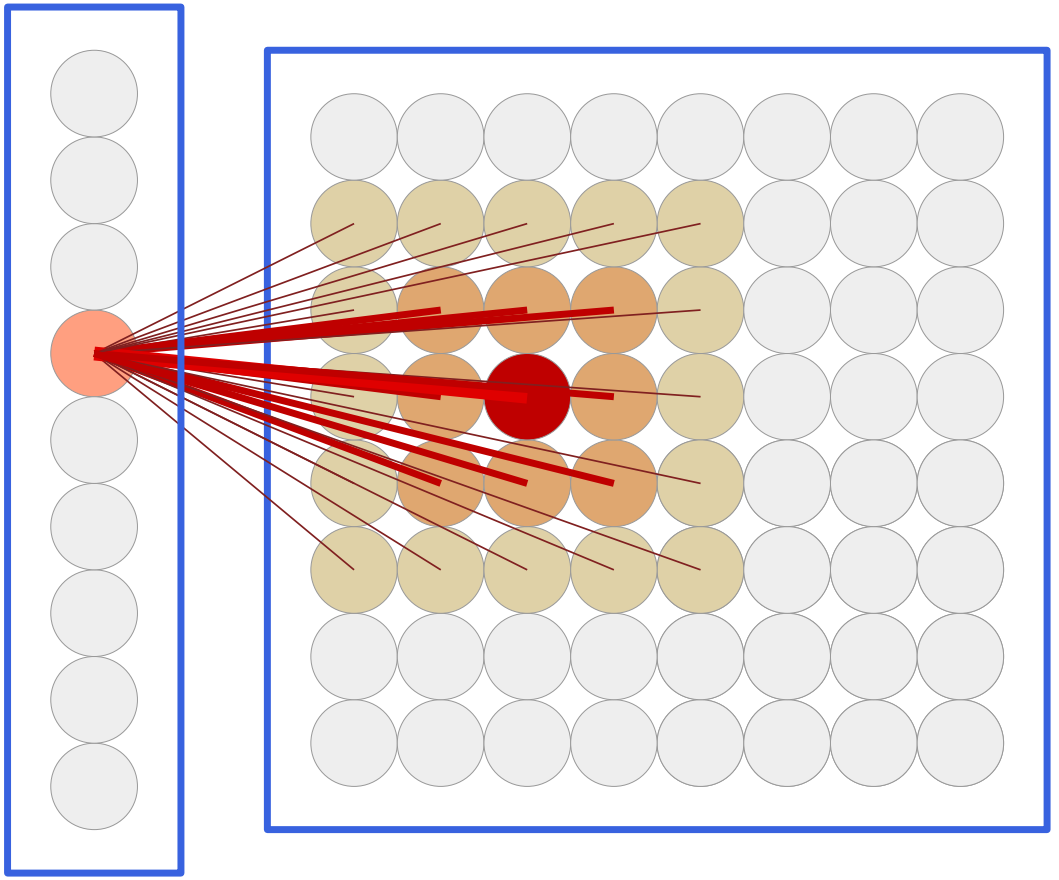
Self-Organizing Maps (SOMs)

- Uma camada de entrada, contendo os dados que serão usados para treinamento.
- Uma camada de neurônios para mapeamento.
- Cada neurônio é um vetor com as mesmas dimensões da entrada.



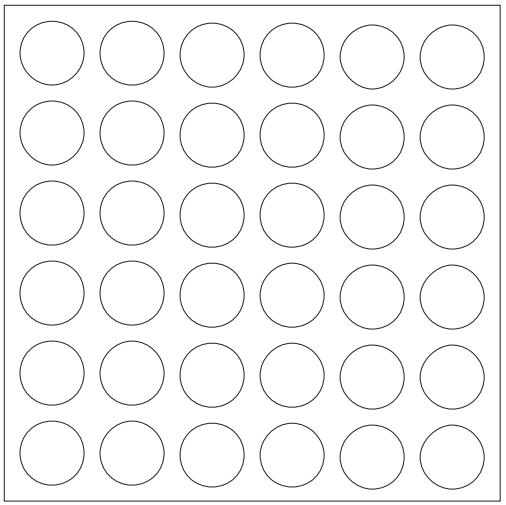
Self-Organizing Maps (SOMs)

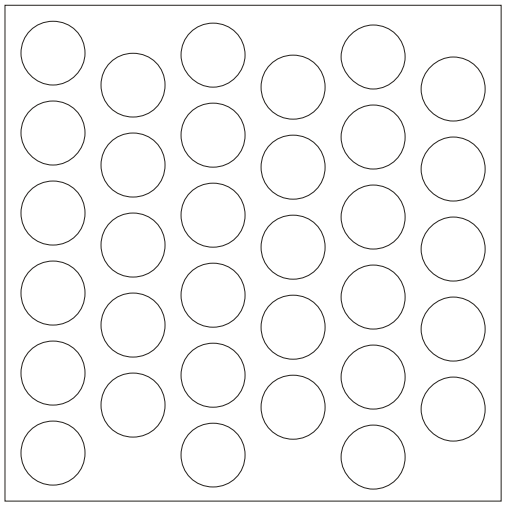
- Entrada: Vetores de dados, rede (considerar arquitetura), parâmetros de treinamento.
- Saída: rede treinada, neurônios se assemelham a vetores apresentados.

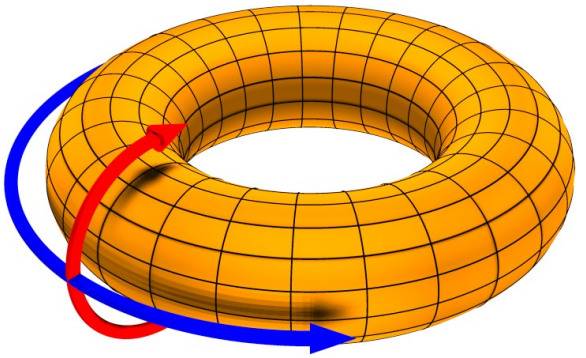


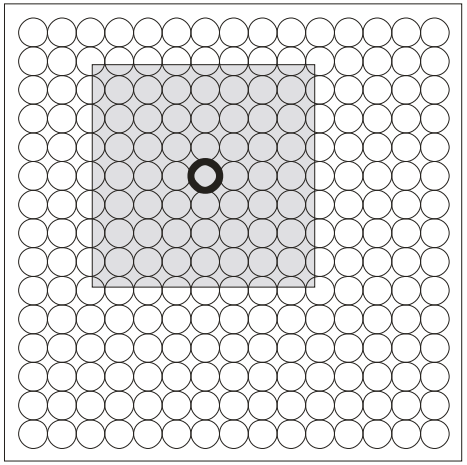
Self-Organizing Maps (SOMs)

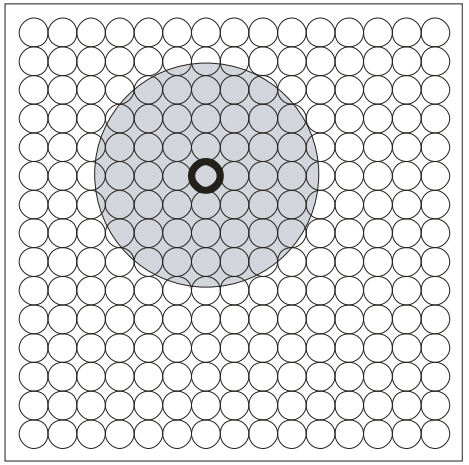
- Topologia e vizinhança

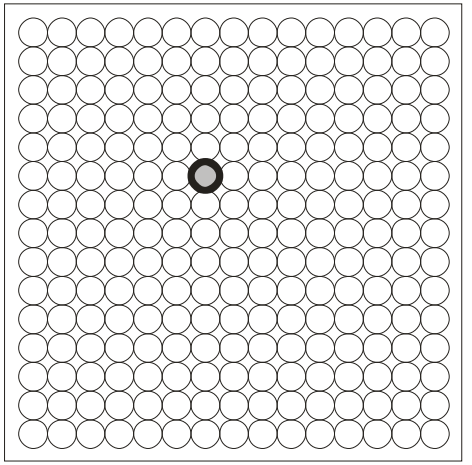


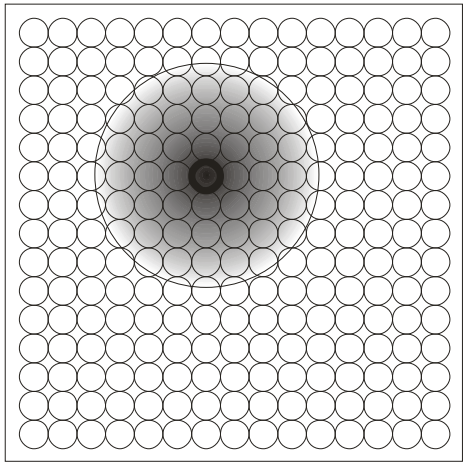












Self-Organizing Maps (SOMs)

1. Inicializar vetores da rede (neurônios) com valores aleatórios.
2. Escolher uma amostra (vetor) de dados.
3. Encontrar o neurônio mais semelhante:
   Aquele cuja distância no espaço de atributos seja a menor para o vetor de dados = o "mais parecido" ou vencedor (*Best Matching Unit*).
4. Atualizar os valores do neurônio vencedor e de seus vizinhos para que fiquem mais similares aos do vetor de entrada.

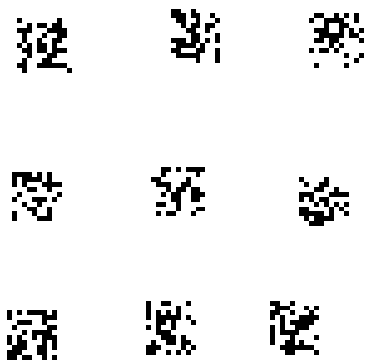
$$W_{t\ 1} \quad W_t \quad L_t R_t \left| W_t \quad D \right|$$

5. Verificar critérios de parada, retornar ao passo 2 se for o caso, atualizar valores para treinamento.

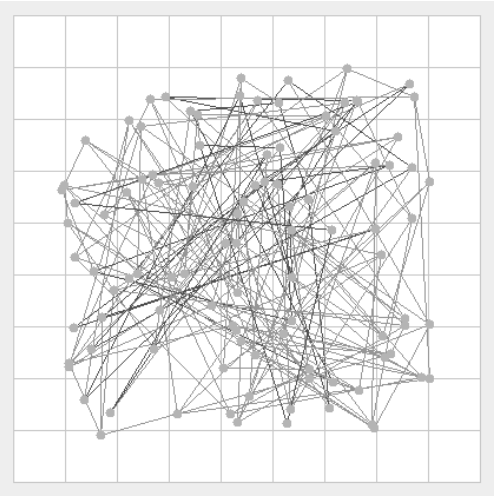
Self-Organizing Maps (SOMs)

- Taxa de aprendizado (*learning rate* L):
  - Valor multiplicador que indica o quanto os valores de um neurônio serão aproximados do dado de entrada.
  - Deve decrescer à medida em que a rede é treinada até um valor mínimo.
- Raio da vizinhança (R).
  - Limiar/valor que indica se um neurônio próximo ao vencedor será considerado vizinho do mesmo.
  - Deve decrescer à medida em que a rede é treinada até um valor mínimo.
  - Aplicável somente à algumas vizinhanças.

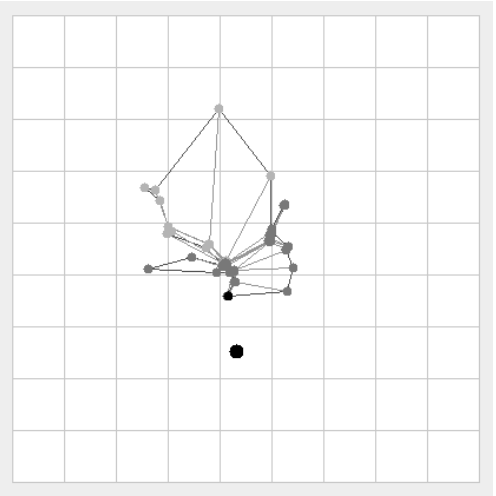
Self-Organizing Maps (SOMs): Exemplos



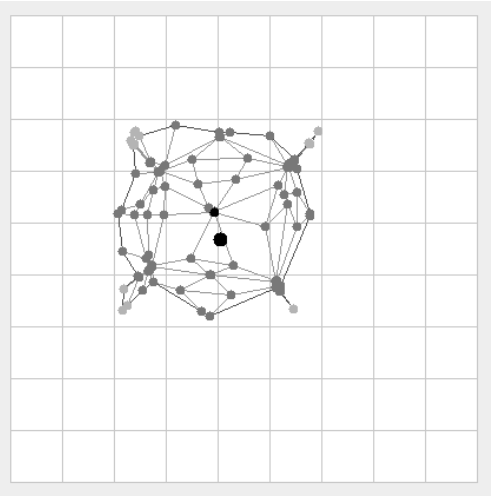
Dados Originais



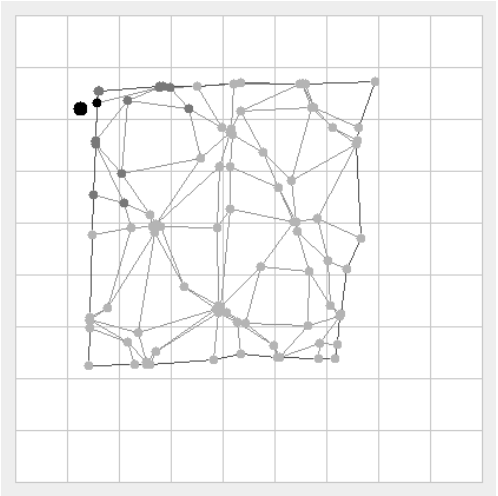
0 iterações



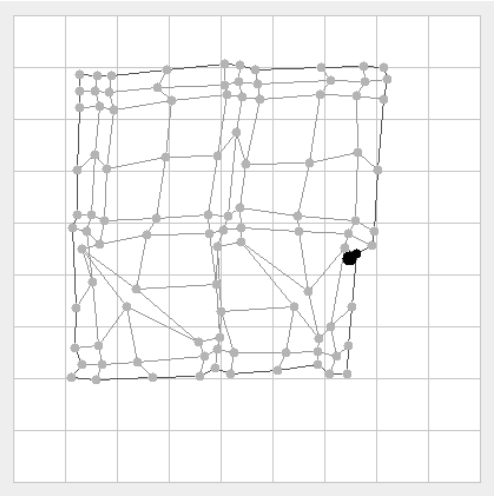
50000 iterações



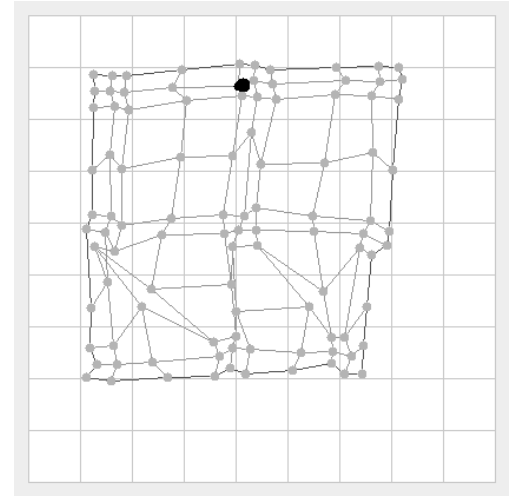
100000 iterações



200000 iterações



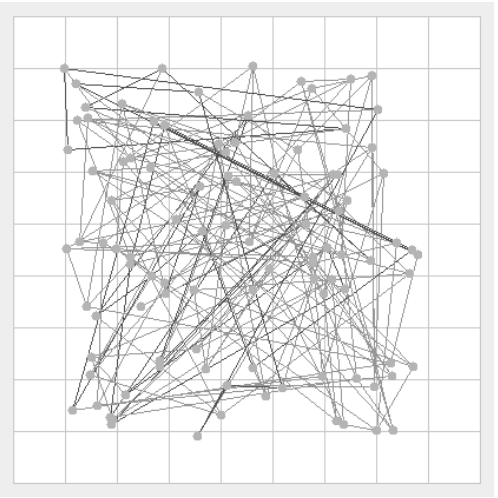
300000 iterações



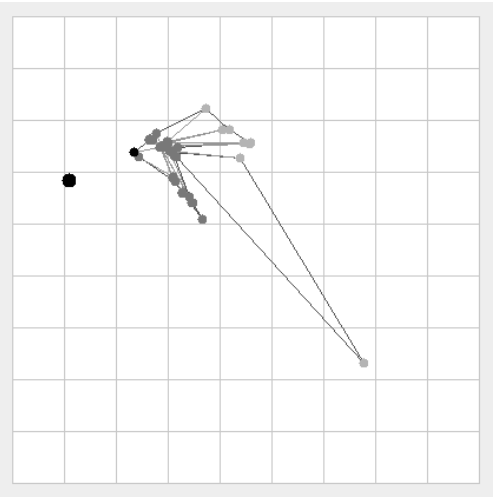
400000 iterações

Self-Organizing Maps (SOMs): Exemplos

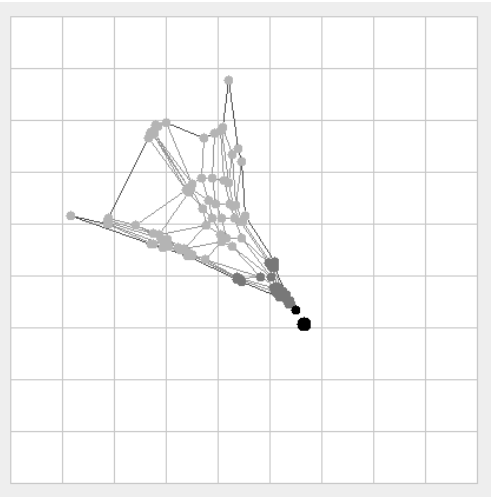
Dados Originais



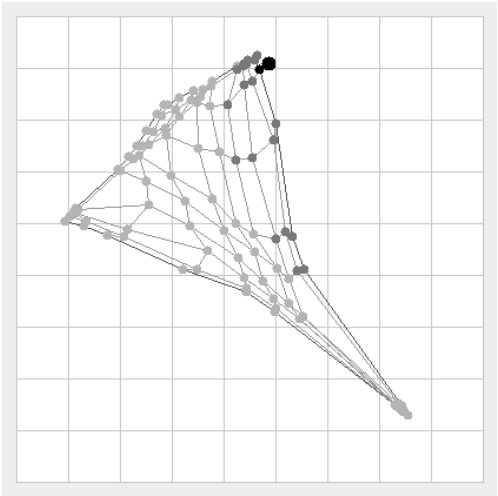
0 iterações



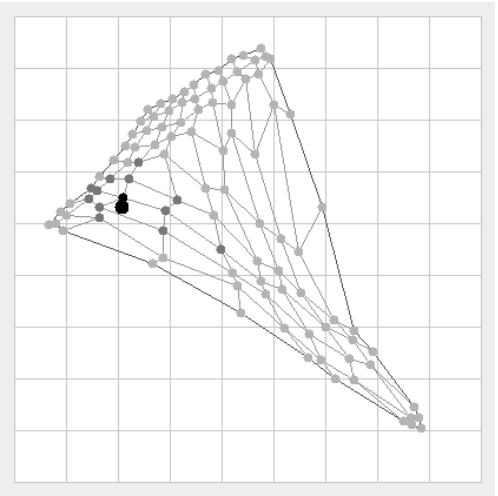
50000 iterações



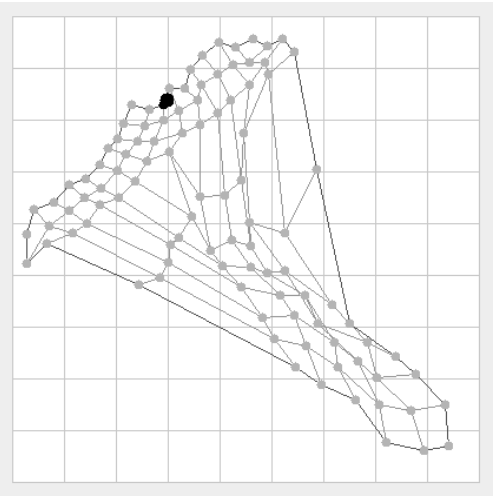
100000 iterações



150000 iterações

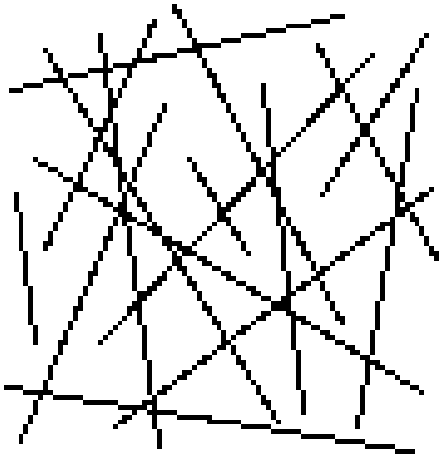


200000 iterações

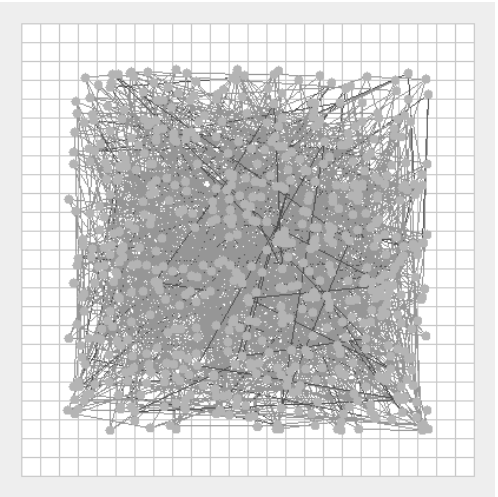


300000 iterações

Self-Organizing Maps (SOMs): Exemplos

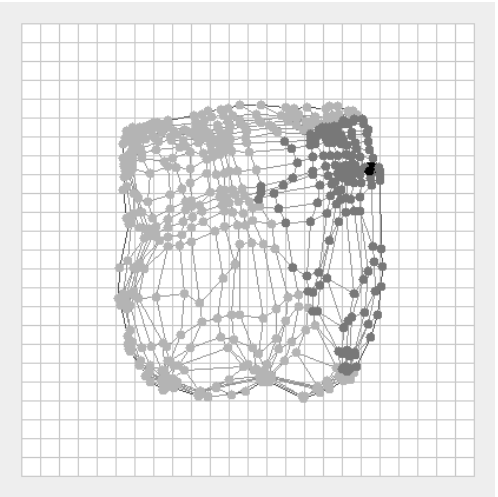


Dados Originais

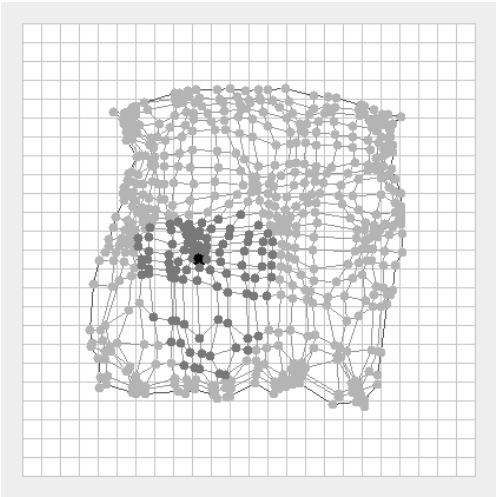


0 iterações

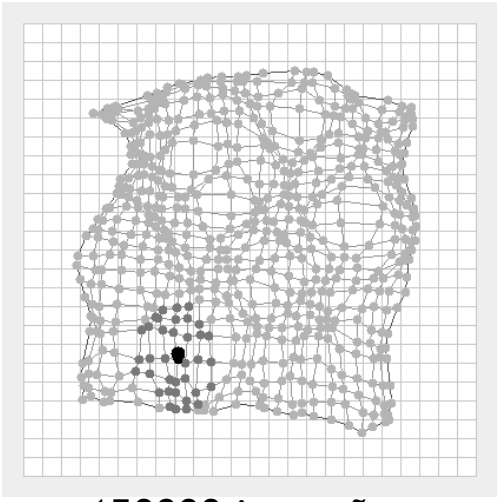
(25x25)



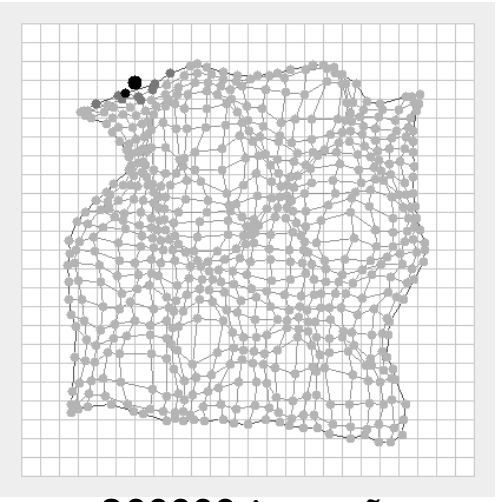
50000 iterações



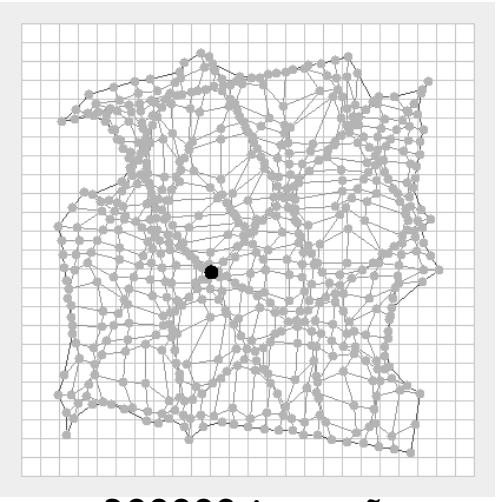
100000 iterações



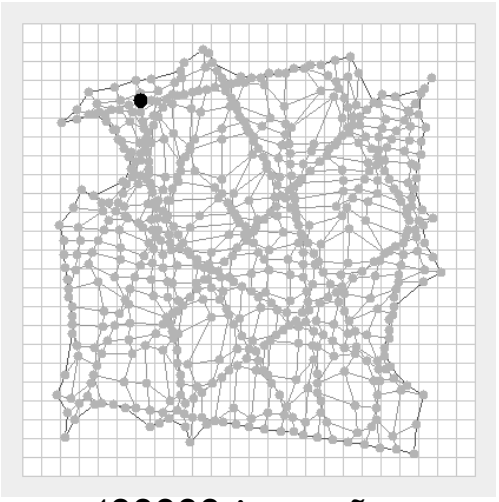
150000 iterações



200000 iterações



300000 iterações



400000 iterações

Regras de Associação

Regras de Associação

- Regras sobre relações e co-ocorrências em bases de dados:
- Se ***X*** ocorre na base de dados, então ***Y*** também ocorre (com alguma relação a ***X***).
- Co-ocorrência: se ***X***, ***Y*** e ***Z*** ocorrem na base de dados então ***A*** também ocorre (com alguma relação a ***X***, ***Y*** e ***Z***).
  - ***X***, ***Y*** e ***Z*** são os antecedentes da associação; ***A*** é o conseqüente.
  - Ocorrências consideradas em escopo limitado: não queremos dizer que se ***X*** ocorre em qualquer "local" da base de dados, ***Y*** também ocorrerá em qualquer "local".
- Muito usado para verificar associações em tabelas de transações ("carrinhos de compra")

Regras de Associação

- Exemplo simples:

Transação	Itens
1	leite, ovos, café, açúcar, fraldas, manteiga
2	leite, café, farinha
3	leite, ovos, açúcar
4	café, açúcar
5	fraldas
6	manteiga, ovos, leite
7	café, açúcar, leite, ovos
8	farinha, manteiga, ovos
9	manteiga, ovos, leite, café, açúcar
10	fraldas, café, cerveja

Regras de Associação

- Conclusões simples sobre a base de dados da tabela:
  - Quem compra leite quase sempre compra ovos.
    - Como definir “quase sempre”? Quantas vezes isso ocorre na base de dados?
  - Quem compra ovos e açúcar sempre compra leite.
    - Mas quantas compras contém ovos e açúcar? O que causa a compra de leite?
  - Quem compra cerveja sempre compra fraldas.
    - Quantas vezes isso ocorre na base de dados?
      Isso é relevante?

Transação	Itens
1	leite, ovos, café, açúcar, fraldas, manteiga
2	leite, café, farinha
3	leite, ovos, açúcar
4	café, açúcar
5	fraldas
6	manteiga, ovos, leite
7	café, açúcar, leite, ovos
8	farinha, manteiga, ovos
9	manteiga, ovos, leite, café, açúcar
10	fraldas, café, cerveja

Regras de Associação

- Muitos que compram café também compram açúcar.
- Ninguém compra só leite.
  - Muitas outras associações negativas existem: quem compra fraldas não compra farinha, quem compra farinha não compra cerveja.
- Quais associações negativas são significativas?

Transação	Itens
1	leite, ovos, café, açúcar, fraldas, manteiga
2	leite, café, farinha
3	leite, ovos, açúcar
4	café, açúcar
5	fraldas
6	manteiga, ovos, leite
7	café, açúcar, leite, ovos
8	farinha, manteiga, ovos
9	manteiga, ovos, leite, café, açúcar
10	fraldas, café, cerveja

Regras de Associação

- Métricas:
- Significância em uma associação: ela pode existir mas ser muito rara em uma base de dados (ex. cerveja → fraldas).
  - **Suporte X → Y**: número de casos que contém ***X*** e ***Y*** dividido pelo número total de registros.
- Confiança em uma associação: o antecedente pode ocorrer várias vezes na base de dados mas nem sempre com o mesmo conseqüente associado.
  - **Confiança X → Y**: número de registros que contém ***X*** e ***Y*** dividido pelo número de registros que contém ***X***.

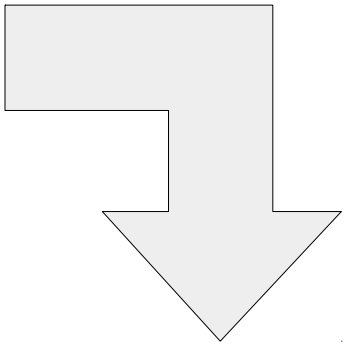
Regras de Associação

- Algoritmo Apriori:
  1. Entrada: coleção de dados associados, suporte mínimo, confiança mínima.
  2. Considerar *K* = 1 para criação de *K-itemsets*
  3. Analisar os dados associados e criar uma tabela de *K-itemsets* com suporte acima do suporte mínimo.
  4. Criar com os *itemsets* filtrados um conjunto de candidatos a (*K* + *1*) *itemsets*.
  5. Usar propriedades do Apriori para eliminar *itemsets* infreqüentes.
  6. Repetir desde o passo 3 até que o conjunto gerado seja vazio.
  7. Listar regras de associação (com permutações) e aplicar limite de confiança.

Regras de Associação

- Simulação do Apriori com suporte mínimo 25% e confiança 75%:

Transação	Itens
1	leite, ovos, café, açúcar, fraldas, manteiga
2	leite, café, farinha
3	leite, ovos, açúcar
4	café, açúcar
5	fraldas
6	manteiga, ovos, leite
7	café, açúcar, leite, ovos
8	farinha, manteiga, ovos
9	manteiga, ovos, leite, café, açúcar
10	fraldas, café, cerveja

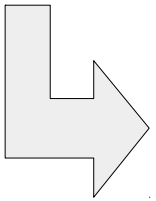


Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1

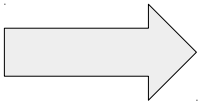
Regras de Associação

- Simulação do Apriori com suporte mínimo 25% e confiança 75%:

Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1



1-*itemsets*	Suporte
leite	60%
ovos	60%
café	60%
açúcar	50%
fraldas	30%
manteiga	40%
farinha	20%
cerveja	10%



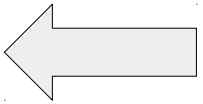
2-*itemsets*	Suporte
[leite,ovos]	50%
[leite,café]	40%
[leite,açúcar]	40%
[leite,fraldas]	10%
[leite,manteiga]	30%
[ovos,café]	30%
[ovos,açúcar]	40%
[ovos,fraldas]	10%
[ovos,manteiga]	40%
[café,açúcar]	40%
[café,fraldas]	20%
[café,manteiga]	20%
[açúcar,fraldas]	10%
[açúcar,manteiga]	20%
[fraldas,manteiga]	10%

Regras de Associação

- Simulação do Apriori com suporte mínimo 25% e confiança 75%:

Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1

3-*itemsets*	Suporte
[leite,ovos,café]	30%
[leite,ovos,açúcar]	40%
[leite,ovos,manteiga]	30%
[leite,café,açúcar]	30%
[leite,café,manteiga]	20%
[leite,açúcar,manteiga]	20%
[ovos,café,açúcar]	30%
[ovos,café,manteiga]	20%
[ovos,açúcar,manteiga]	20%
[café,açúcar,manteiga]	20%



2-*itemsets*	Suporte
[leite,ovos]	50%
[leite,café]	40%
[leite,açúcar]	40%
[leite,fraldas]	10%
[leite,manteiga]	30%
[ovos,café]	30%
[ovos,açúcar]	40%
[ovos,fraldas]	10%
[ovos,manteiga]	40%
[café,açúcar]	40%
[café,fraldas]	20%
[café,manteiga]	20%
[açúcar,fraldas]	10%
[açúcar,manteiga]	20%
[fraldas,manteiga]	10%

Regras de Associação

- Simulação do Apriori com suporte mínimo 25% e confiança 75%:

Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1

3-*itemsets*	Suporte
[leite,ovos,café]	30%
[leite,ovos,açúcar]	40%
[leite,ovos,manteiga]	30%
[leite,café,açúcar]	30%
[leite,café,manteiga]	20%
[leite,açúcar,manteiga]	20%
[ovos,café,açúcar]	30%
[ovos,café,manteiga]	20%
[ovos,açúcar,manteiga]	20%
[café,açúcar,manteiga]	20%

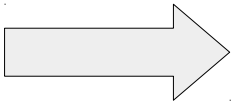
4-*itemsets*	Suporte
[leite,ovos,café,açúcar]	30%
[leite,ovos,café,manteiga]	20%
[leite,ovos,açúcar,manteiga]	20%
[leite,café,açúcar,manteiga]	20%
[ovos,café,açúcar,manteiga]	20%

Regras de Associação

- Simulação do Apriori com suporte mínimo 25% e confiança 75%:

Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1

2-*itemsets*	Suporte
[leite,ovos]	50%
[leite,café]	40%
[leite,açúcar]	40%
[leite,fraldas]	10%
[leite,manteiga]	30%
[ovos,café]	30%
[ovos,açúcar]	40%
[ovos,fraldas]	10%
[ovos,manteiga]	40%
[café,açúcar]	40%
[café,fraldas]	20%
[café,manteiga]	20%
[açúcar,fraldas]	10%
[açúcar,manteiga]	20%
[fraldas,manteiga]	10%

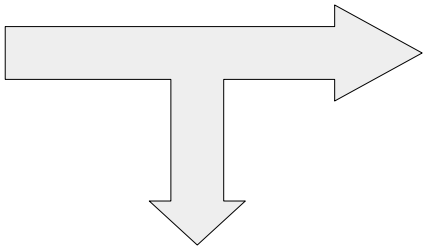


Regra	Suporte	Confiança
[ovos → leite]	50%	83%
[leite → ovos]	50%	83%
[café → leite]	40%	66%
[leite → café]	40%	66%
[açúcar → leite]	40%	80%
[leite → açúcar]	40%	66%
[manteiga → leite]	30%	75%
[leite → manteiga]	30%	50%
[café → ovos]	30%	50%
[ovos → café]	30%	50%
[açúcar → ovos]	40%	80%
[ovos → açúcar]	40%	66%
[manteiga → ovos]	40%	100%
[ovos → manteiga]	40%	66%
[açúcar → café]	40%	80%
[café → açúcar]	40%	66%

Regras de Associação

- Simulação do Apriori com suporte mínimo 25% e confiança 75%:

3-*itemsets*	Suporte
[leite,ovos,café]	30%
[leite,ovos,açúcar]	40%
[leite,ovos,manteiga]	30%
[leite,café,açúcar]	30%
[leite,café,manteiga]	20%
[leite,açúcar,manteiga]	20%
[ovos,café,açúcar]	30%
[ovos,café,manteiga]	20%
[ovos,açúcar,manteiga]	20%
[café,açúcar,manteiga]	20%



Regra	Suporte	Confiança
[café, ovos → leite]	30%	100%
[ovos, café → leite]	30%	100%
[ovos, leite → café]	30%	60%
[café, leite → ovos]	30%	75%
[leite, café → ovos]	30%	75%
[leite, ovos → café]	30%	60%
[açúcar, ovos → leite]	40%	100%
[ovos, açúcar → leite]	40%	100%
[ovos, leite → açúcar]	40%	80%
[açúcar, leite → ovos]	40%	100%
[leite, açúcar → ovos]	40%	100%
[leite, ovos → açúcar]	40%	80%

Regra	Suporte	Confiança
[manteiga, ovos → leite]	30%	75%
[ovos, manteiga → leite]	30%	75%
[ovos, leite → manteiga]	30%	60%
[manteiga, leite → ovos]	30%	100%
[leite, manteiga → ovos]	30%	100%
[leite, ovos → manteiga]	30%	60%
[açúcar, café → leite]	30%	75%
[café, açúcar → leite]	30%	75%
[café, leite → açúcar]	30%	75%
[açúcar, leite → café]	30%	75%
[leite, açúcar → café]	30%	75%
[leite, café → açúcar]	30%	75%
[açúcar, café → ovos]	30%	75%
[café, açúcar → ovos]	30%	75%
[café, ovos → açúcar]	30%	100%
[açúcar, ovos → café]	30%	75%
[ovos, açúcar → café]	30%	75%
[ovos, café → açúcar]	30%	100%

Regras de Associação

- Simulação do Apriori com suporte mínimo 25% e confiança 75%:

Transação	leite	ovos	café	açúcar	fraldas	manteiga	farinha	cerveja
1	1	1	1	1	1	1	0	0
2	1	0	1	0	0	0	1	0
3	1	1	0	1	0	0	0	0
4	0	0	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0
8	0	1	0	0	0	1	1	0
9	1	1	1	1	0	1	0	0
10	0	0	1	0	1	0	0	1

4-*itemsets*	Suporte
[leite,ovos,café,açúcar]	30%
[leite,ovos,café,manteiga]	20%
[leite,ovos,açúcar,manteiga]	20%
[leite,café,açúcar,manteiga]	20%
[ovos,café,açúcar,manteiga]	20%

Regra	Suporte	Confiança
[açúcar, café, ovos → leite]	30%	100%
[café, açúcar, ovos → leite]	30%	100%
[café, ovos, açúcar → leite]	30%	100%
[café, ovos, leite → açúcar]	30%	100%
[açúcar, ovos, café → leite]	30%	100%
[ovos, açúcar, café → leite]	30%	100%
[ovos, café, açúcar → leite]	30%	100%
[ovos, café, leite → açúcar]	30%	100%
[açúcar, ovos, leite → café]	30%	75%
[ovos, açúcar, leite → café]	30%	75%
[ovos, leite, açúcar → café]	30%	75%
[ovos, leite, café → açúcar]	30%	100%
[açúcar, café, leite → ovos]	30%	100%
[café, açúcar, leite → ovos]	30%	100%
[café, leite, açúcar → ovos]	30%	100%
[café, leite, ovos → açúcar]	30%	100%
[açúcar, leite, café → ovos]	30%	100%
[leite, açúcar, café → ovos]	30%	100%
[leite, café, açúcar → ovos]	30%	100%
[leite, café, ovos → açúcar]	30%	100%
[açúcar, leite, ovos → café]	30%	75%
[leite, açúcar, ovos → café]	30%	75%
[leite, ovos, açúcar → café]	30%	75%
[leite, ovos, café → açúcar]	30%	100%

Não vimos casos de conseqüentes múltiplos
(ex. [ovos, leite → café, açúcar] tem 60% de confiança).

Não calculamos associações negativas
(ex.[açúcar → não cerveja], com suporte 50% e confiança 100%).

Regras de Associação: Considerações

- Muitos problemas podem ser representados em matrizes binárias (ou variantes): enorme aplicabilidade.
- Associações negativas podem ser tão importantes quanto positivas.
- **Cuidado**! Na vida real as combinações e permutações podem ser muitas, e as regras quase redundantes!
  - Muitas regras geradas: **mineração de regras**.

Data Mining: Outras Técnicas

- Muitas outras técnicas podem ser usadas:
- Pesquisa Operacional, Inteligência Artificial e outras.
- Outros modelos de redes neurais, *Rough Sets*, *Support Vector Machines*, etc.
- Técnicas de algoritmos genéticos, *Particle Swarm Optimization*, etc.
- Técnicas baseadas em sistemas imunes artificiais, biologia/vida artificial, etc.

Próximos passos...

- ***Dia 1:*** Apresentação dos conceitos de mineração de dados, motivação e alguns exemplos.
- ***Dia 2:*** Algoritmos de classificação supervisionada e aplicações.
- ***Dia 3:*** Algoritmos de classificação não-supervisionada e aplicações. Algoritmos de mineração de associações.
- ***Dia 4:*** Visualização e mineração de dados. Outros algoritmos e idéias. Onde aprender mais.

Mais informações em...

- http://www.lac.inpe.br/~rafael.santos
  - http://www.lac.inpe.br/~rafael.santos/dmapresentacoes.jsp
  - http://www.lac.inpe.br/~rafael.santos/cap359-2010.jsp
- http://www.lac.inpe.br/ELAC/index.jsp