Aula 13 – Agrupamento Densidade

1001524 – Aprendizado de Máquina I
2022/1 - Turmas A, B e C
Prof. Dr. Murilo Naldi
PESCD: Bruno Silva Sette

naldi@ufscar.br

Agradecimentos

- Parte do material utilizado nesta aula foi cedido pelos professores Ricardo Campello, Diego Silva e, por esse motivo, o crédito deste material é deles
- Parte do material utilizado nesta aula foi disponibilizado por M. Kumar no endereço: www-users.cs.umn.edu/~kumar/dmbook/index.php
- Agradecimentos a Intel Software e a

  Intel IA Academy pelo material

  disponibilizado e recursos didáticos

Análise de grupos

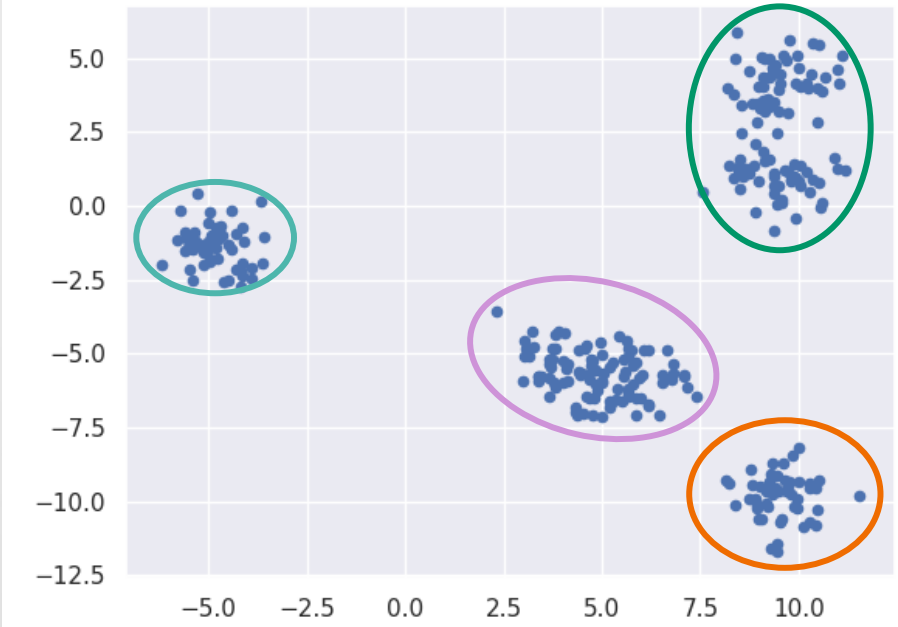
- Grupos por similaridade
  - Objetos do mesmo grupo são os mais similares, enquanto devem ser mais dissimilares de outros grupos
  - É um conceito bastante intuitivo visualmente

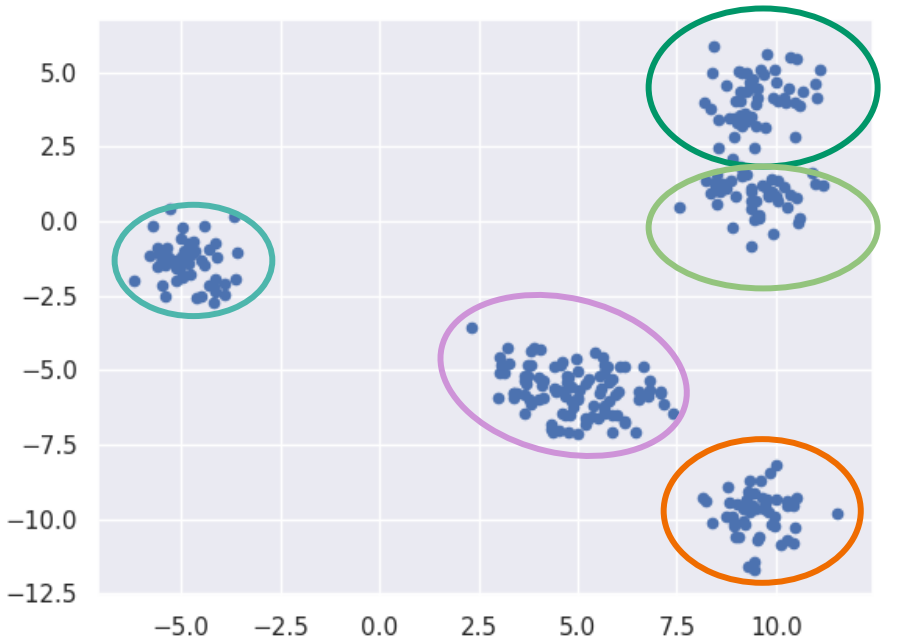
Análise de grupos

- … ou não!

Análise de grupos

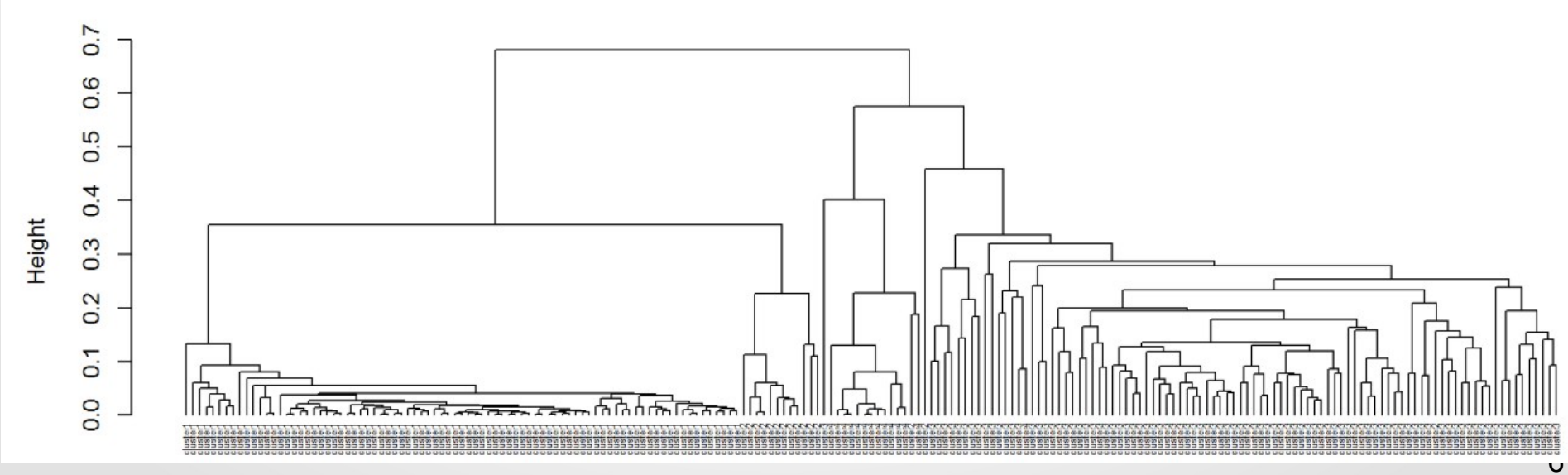
- … ou não!
  - Mas se os dados não estiverem em duas ou três dimensões?





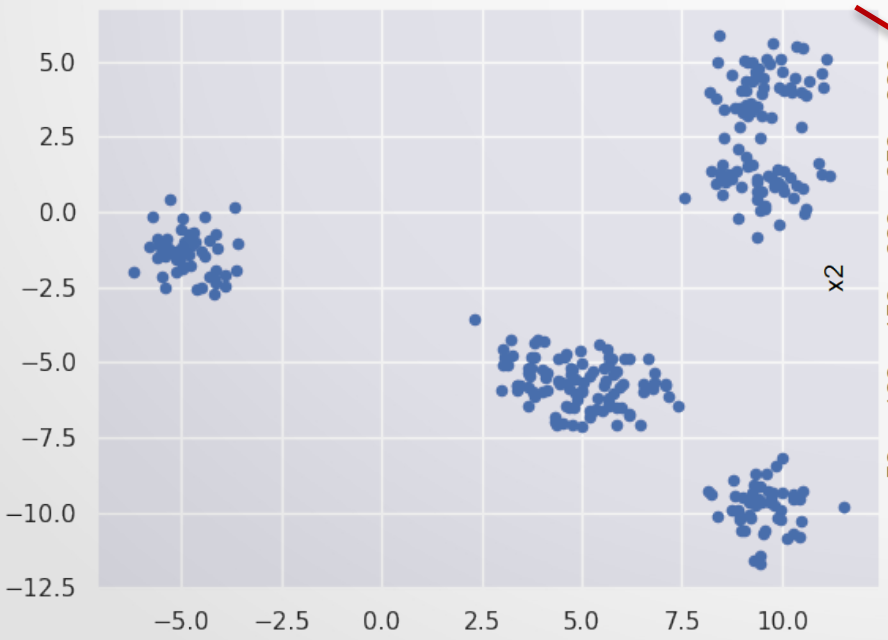
Visualização e análise de grupos

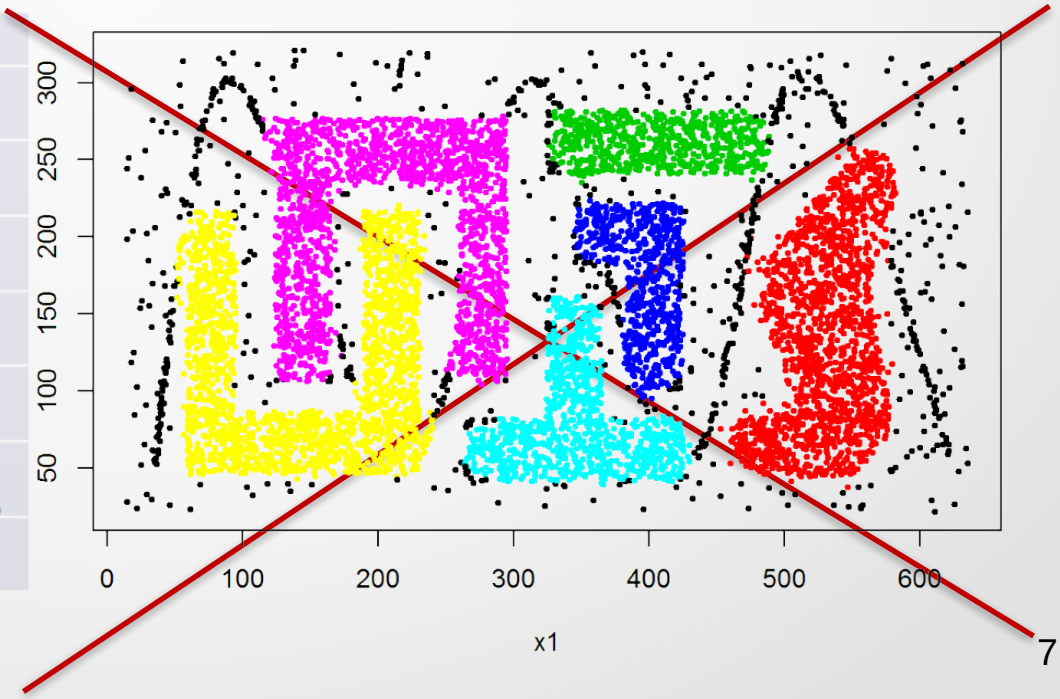
- Ainda assim é possível encontrar rica descrição das relações entre objetos e seus grupos
- “YeastGalactose” Data:
  - Níveis de expressão gênica de um subconjunto de 205 genes selecionados da levedura Saccharomyces cerevisiae de 20 medições diferentes



Formatos globulares

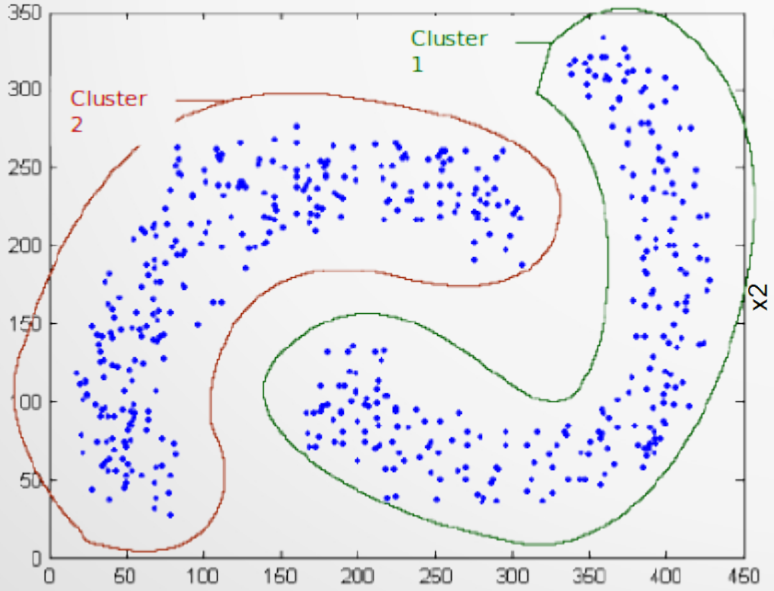
- Métodos focados em agrupamento por similaridade encontram grupos globulares intrinsecamente
  - k-médias, ligação completa, ligação média, Ward's, *bisecting k-means*

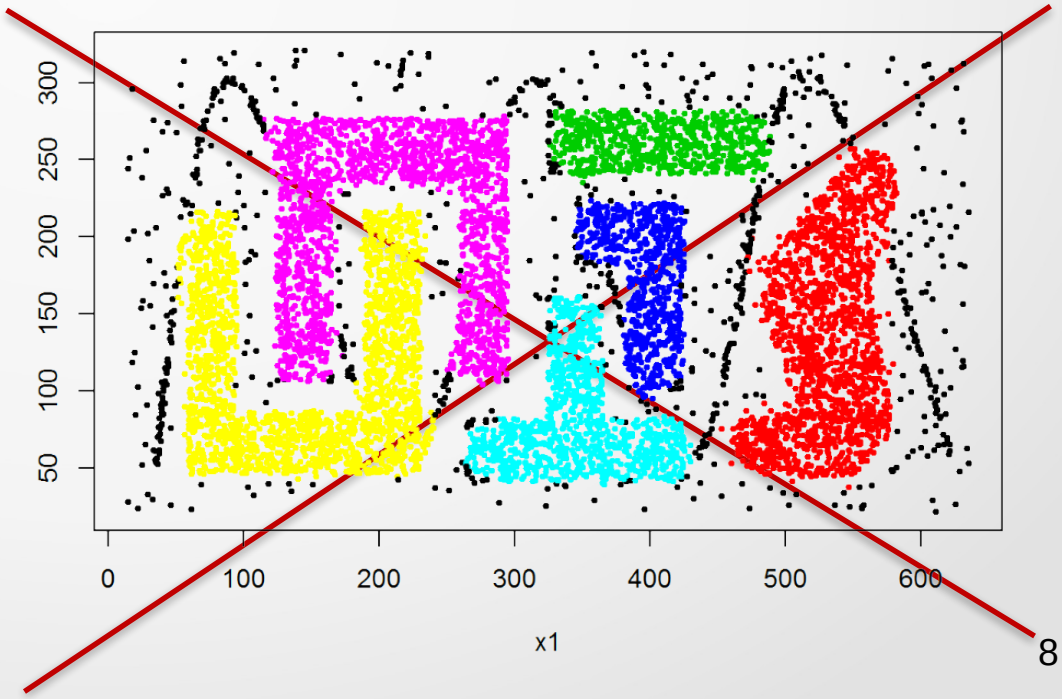




Formatos arbitrários

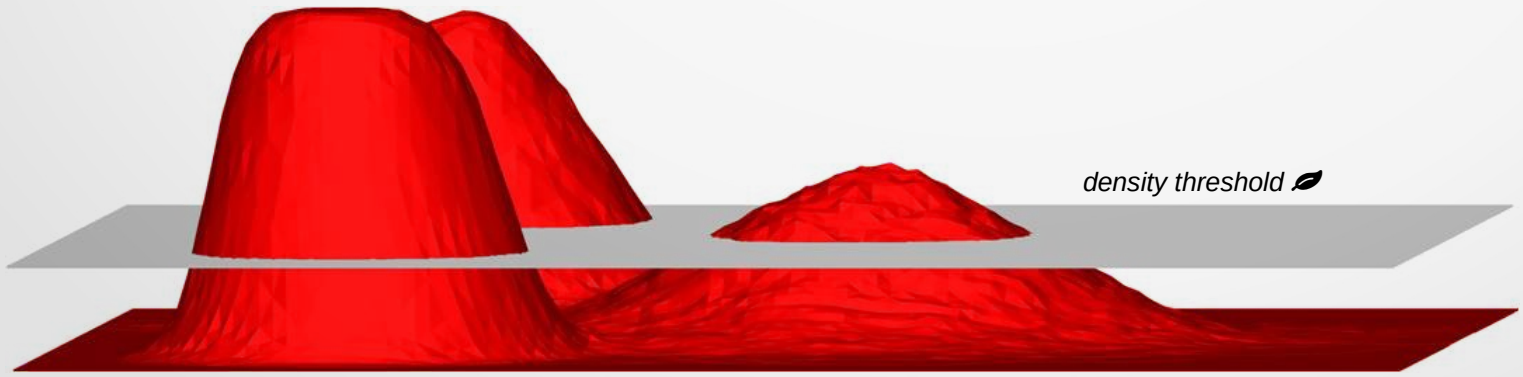
- Algoritmo de ligação simples pode detectar grupos de formatos arbitrários
  - Mas é muito sensível a ruído
    - Problemas como o "*chaining effect*"





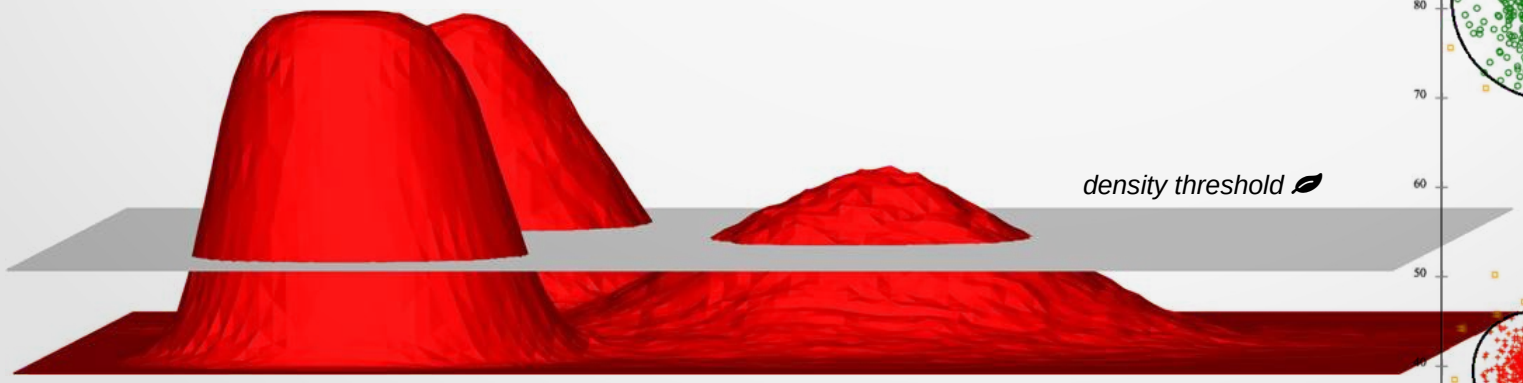
Agrupamento por densidade

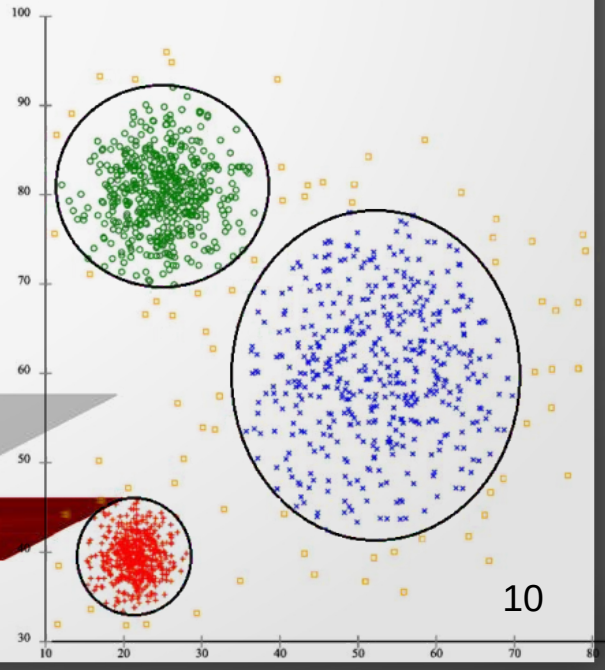
- Conceito estatístico de agrupamento baseado em densidade:
  - Density-Contour Clusters [J.A. Hartigan, “Clustering Algorithms”, J. Wiley & Sons, 1975]
    - Máximo de subconjuntos conectados por densidade em um nível $\{x \mid f(x) \geq \lambda\}$



Agrupamento por densidade

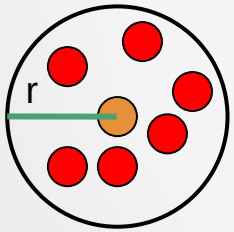
- Conceito estatístico de agrupamento baseado em densidade:
  - Density-Contour Clusters [J.A. Hartigan, “Clustering Algorithms”, J. Wiley & Sons, 1975]
    - Máximo de subconjuntos conectados por densidade em um nível {x | f(x) ≥ λ}
- “Alta densidade separadas por baixa”

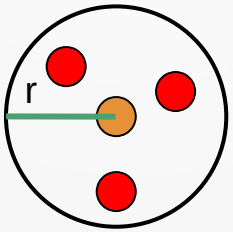


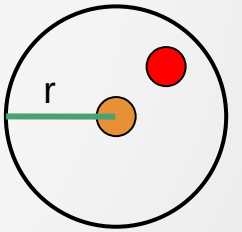


Agrupamento por densidade

- Quais desses pontos (laranjas) estão em regiões densas?
  - Sua resposta foi baseada em quais características?







DBSCAN (1996)

- Density-Based Spatial Clustering of Applications with Noise – DBSCAN
  - Algoritmo de densidade mais popular (> 25.500 citações)
- Baseado na definição anterior de vizinhança
  - Considera pontos em regiões esparsas como ruídos (ou seja, fora de qualquer grupo)
- Capaz de encontrar grupos com formas arbitrárias

Densidade

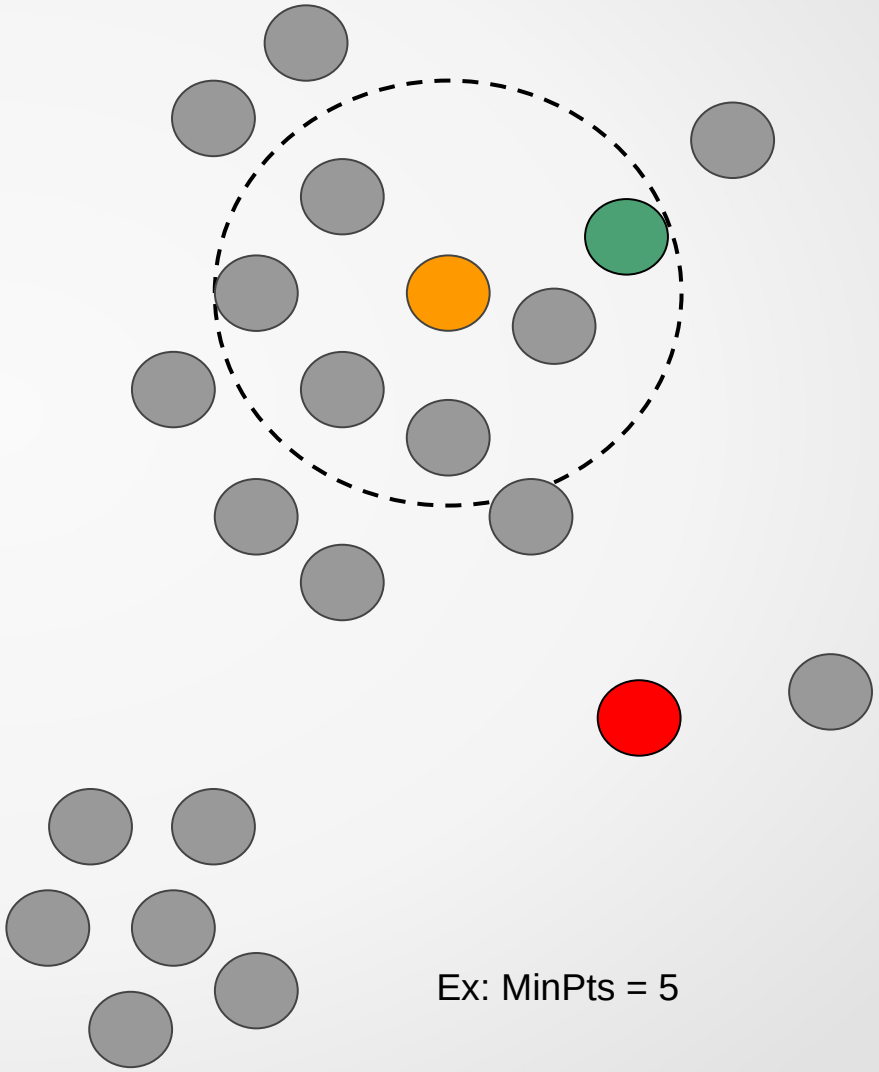
- Dada por dois parâmetros:
  - ε-Vizinhança: conjunto de pontos com distância, no máximo, ε para o ponto de referência p

$$N_\epsilon(p) = \{q | d(p,q) \leq \epsilon\}$$

  - m_{pts}: Número mínimo de pontos na ε-Vizinhança para considerar p como um ponto de região densa

Ponto de núcleo (*core*)

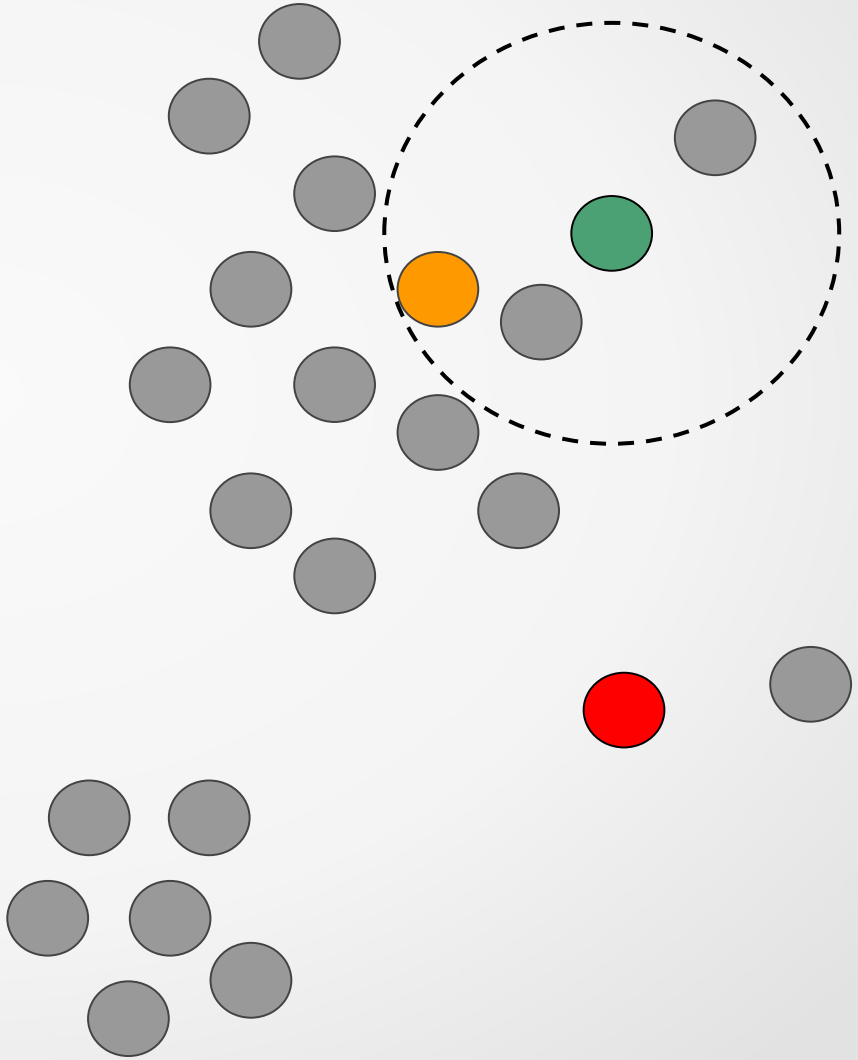
- Ponto em região de alta densidade, ou seja, com m_{pts} ou mais pontos na sua ε-Vizinhança



Ex: MinPts = 5

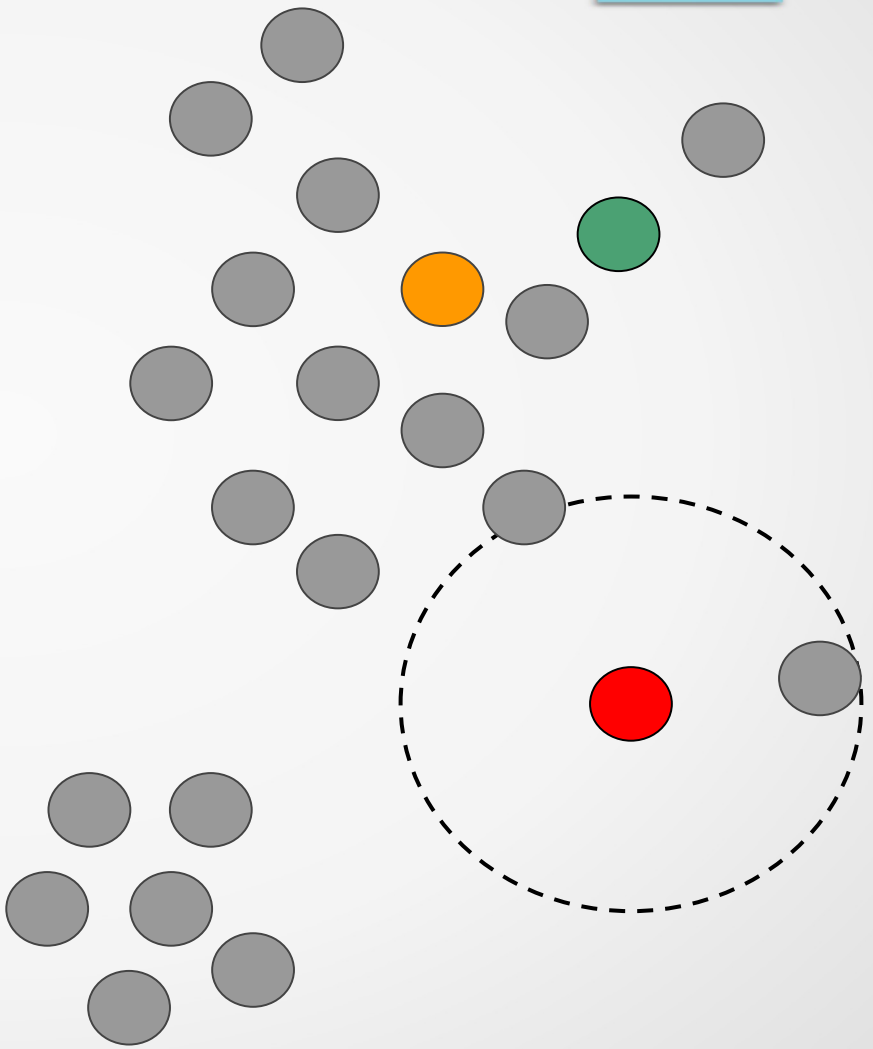
Pontos de borda (*bordeline*)

- Não possui m_{pts} ou mais pontos na sua ε-vizinhança, mas é ε-vizinho de um ponto core



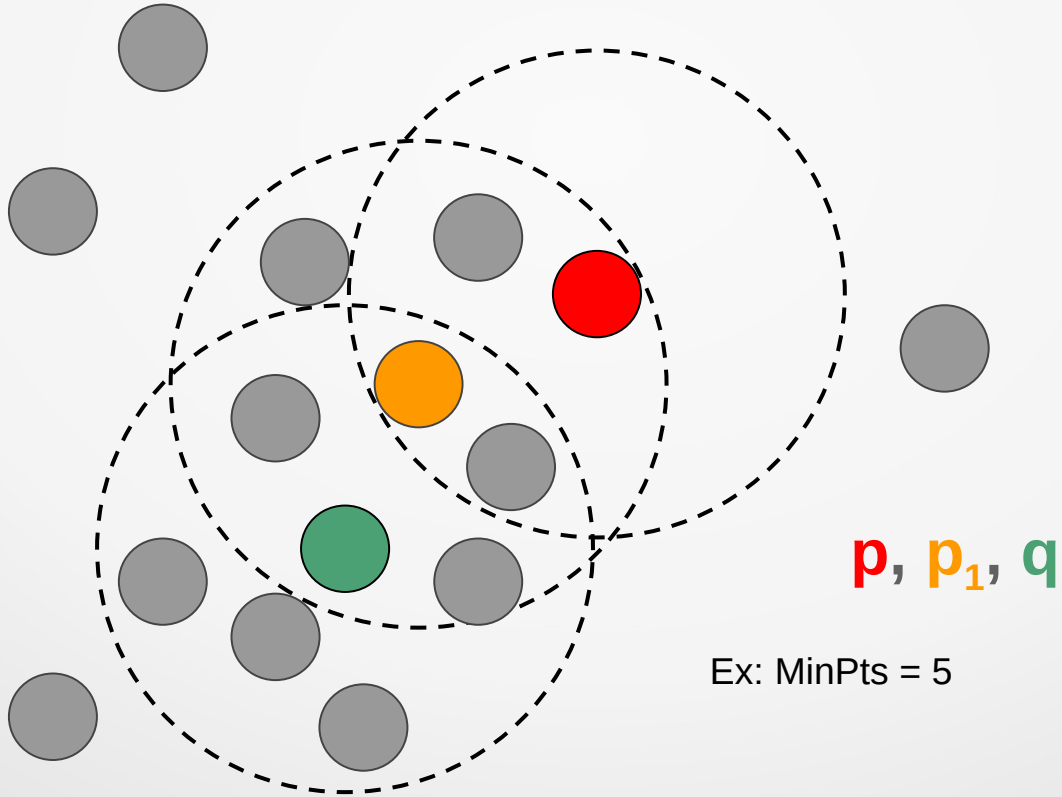
Ruídos e *outliers*

- Os demais pontos



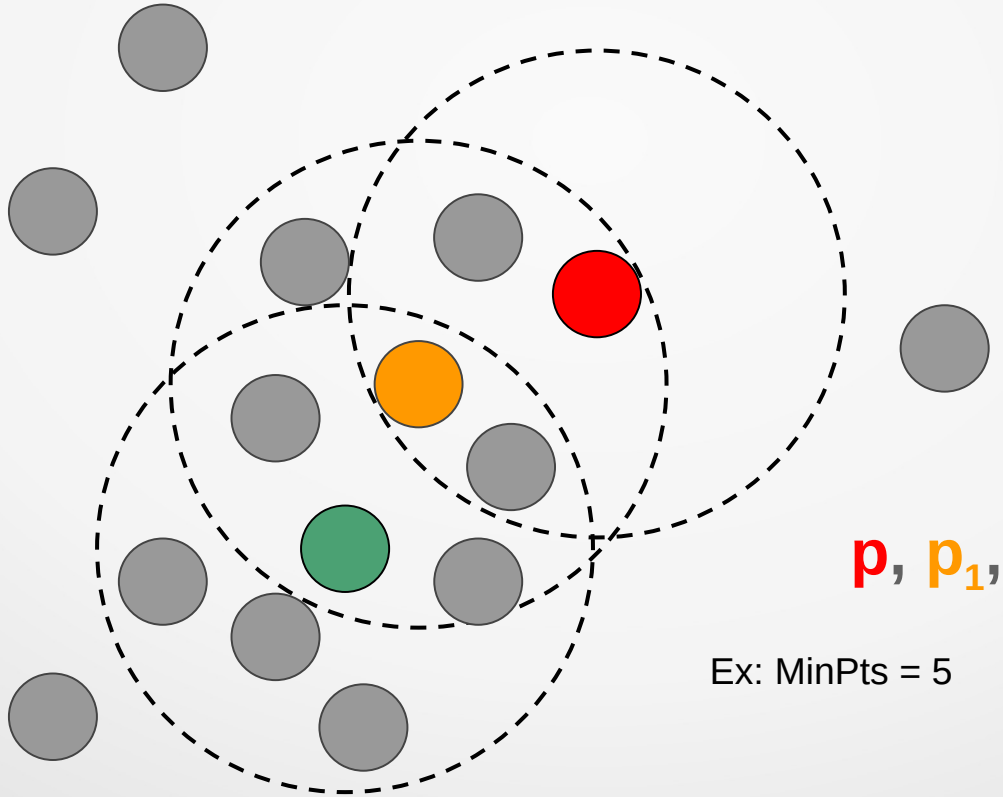
Alcançabilidade (*reachability*)

- Todo ponto na ε-vizinhança de um ponto de núcleo p é dito *diretamente* alcançável (por densidade) por p



Alcançabilidade (*reachability*)

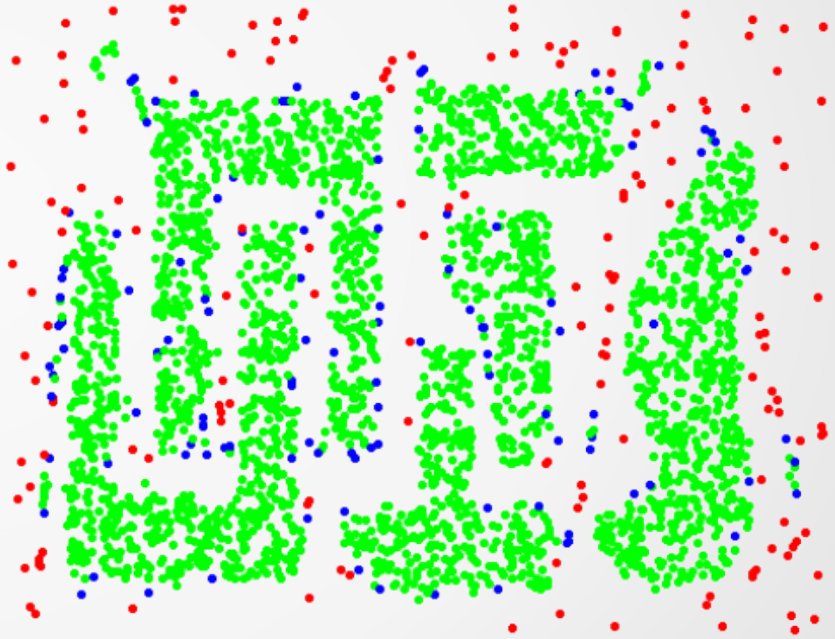
- Caso haja uma corrente de pontos diretamente alcançáveis de p a q, dizemos que q é (*indiretamente*) alcançável por p



DBSCAN - Algoritmo

- Para cada objeto o $\in$ D
  - Se o ainda não foi “classificado”
    - Encontre todos os pontos alcançáveis (por densidade) por o e atribua um novo (e mesmo) grupo a todos esses pontos
  - Senão
    - Atribua o ao conjunto de *outliers*

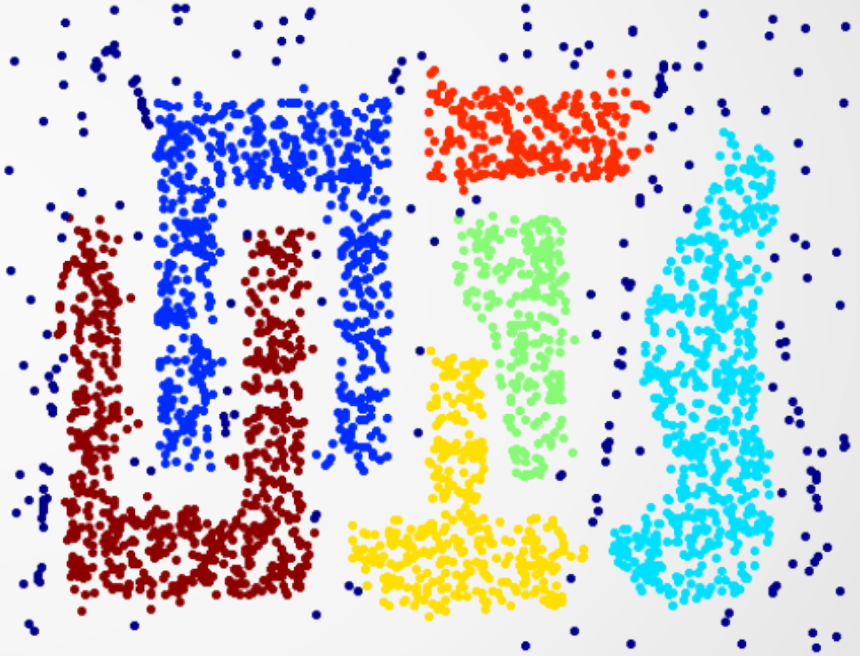
Exemplo



Core, **borderline**, **outlier**

Fonte: https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_density.pdf

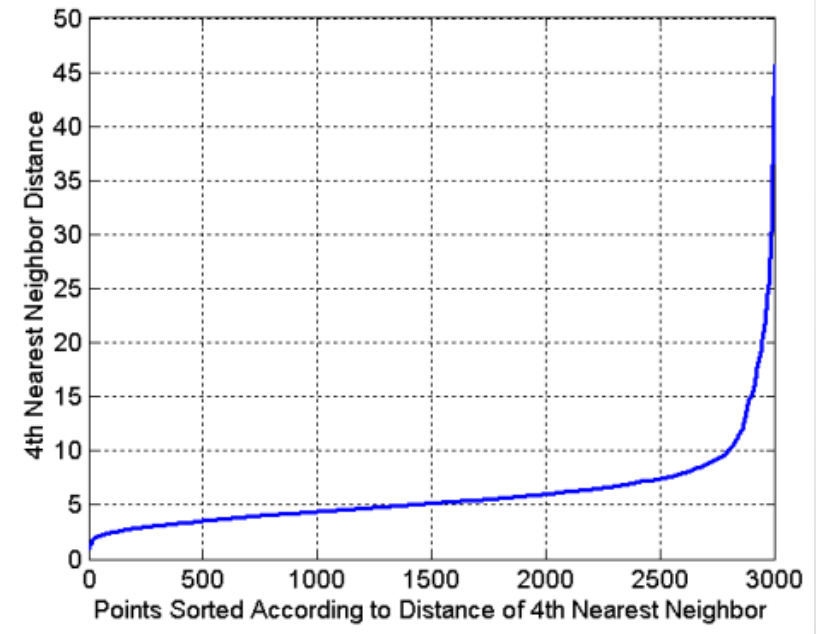
Exemplo – ε ideal



Fonte: https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_density.pdf

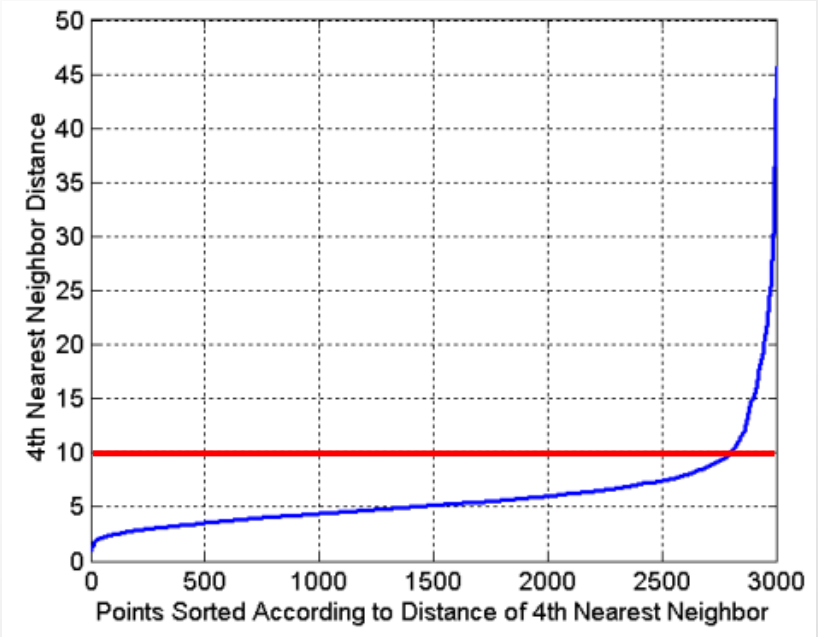
Parâmetros

- A tendência é que os m_{pts} vizinhos mais próximos possuam distâncias parecidas quando estiverem no mesmo grupo
  - *Outliers* possuem a distância para seus vizinhos maior do que pontos *core*
  - Exemplo para $m_{pts} = 4$



Parâmetros

- A tendência é que os m_{pts} vizinhos mais próximos possuam distâncias parecidas quando estiverem no mesmo grupo
  - *Outliers* possuem a distância para seus vizinhos maior do que pontos *core*
  - Exemplo para m_{pts} = 4
    - ε=10 parece ser “ok”



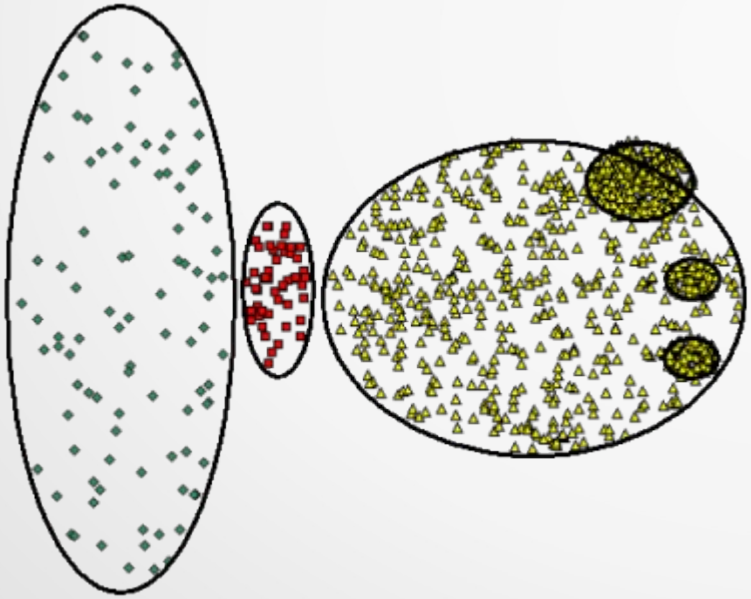
Complexidade

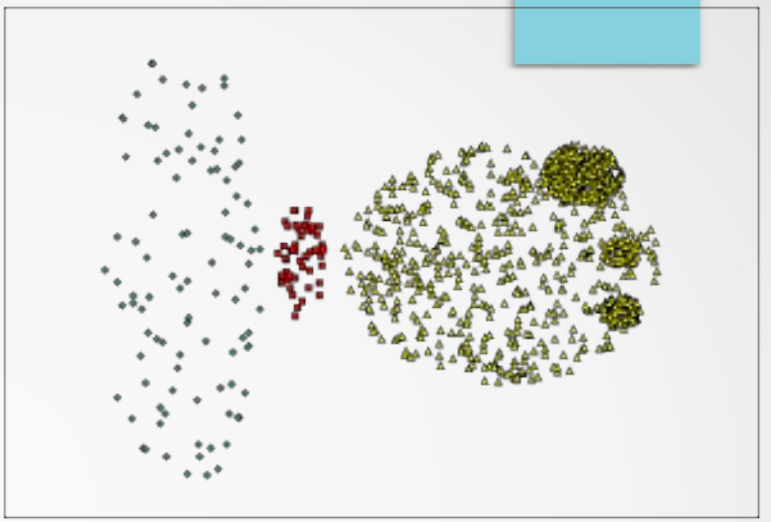
- Espaço: O(n)

- Tempo: O(n^2) pois, para cada ponto, deve-se calcular a distância para os demais
  - Pode-se reduzir a complexidade com aproximações com indexação, space trees, etc : O($nlogn$)

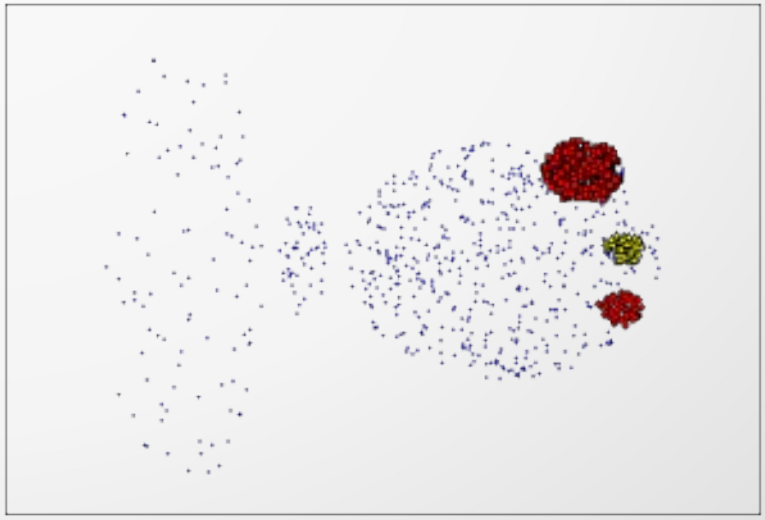
Limitações

- Múltiplas densidades,
  - parametrização,
  - e grandes volumes...





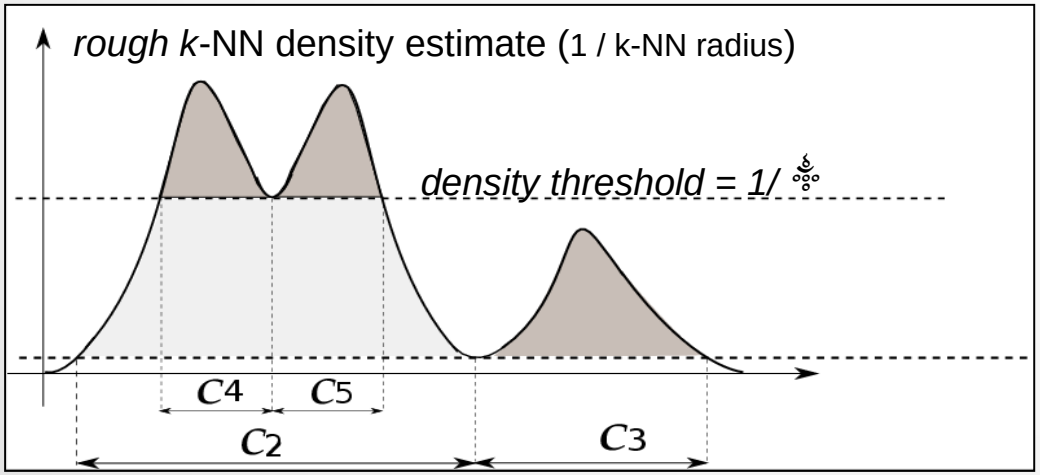
(MinPts=4, Eps=9.92).

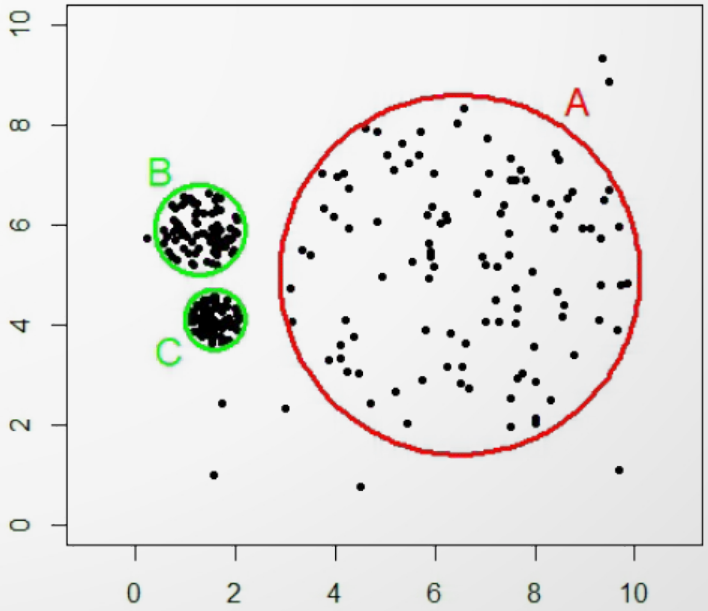


(MinPts=4, Eps=9.75)

Interpretação

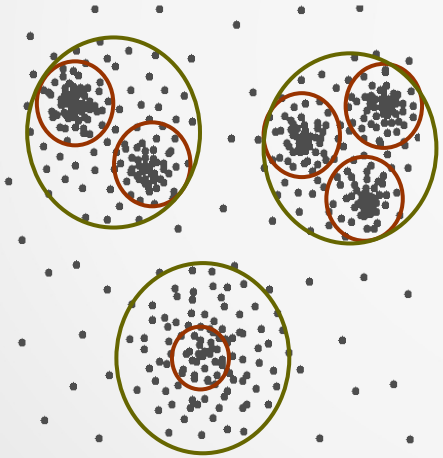
- Estimação de densidade não-paramétrica
  - *k* (m_{pts}) vizinhos é um fator de suavização
    - ε estabelece o limiar de densidade





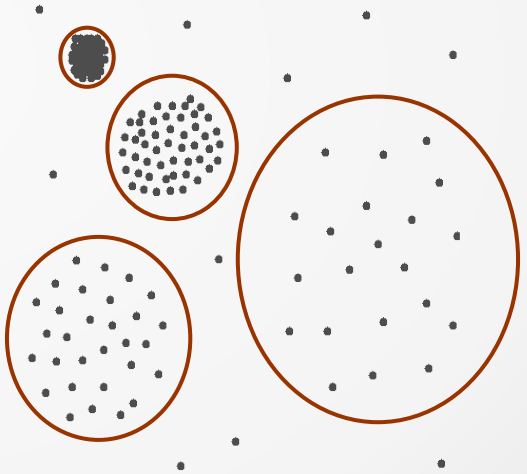
Múltiplas densidades

- Uma solução de única densidade pode não refletir a estrutura dos dados como um todo
  - Particularmente crítico quando temos:



Estrutura naturalmente hierárquica

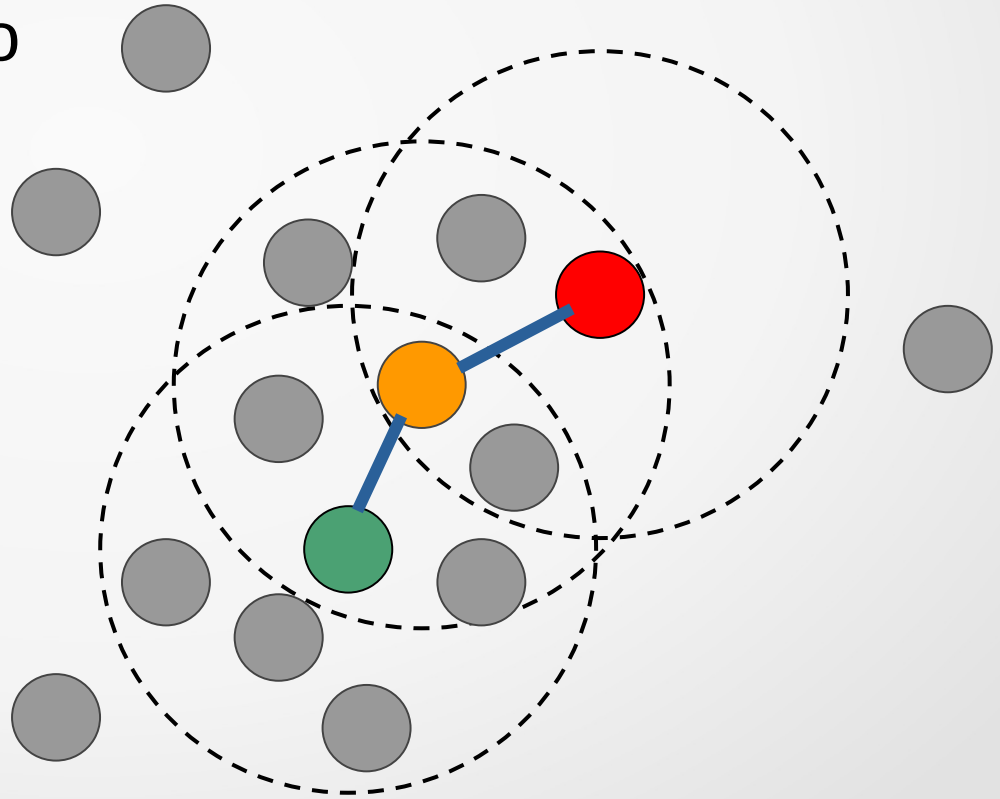
e/ou



Diferentes densidades e tamanhos

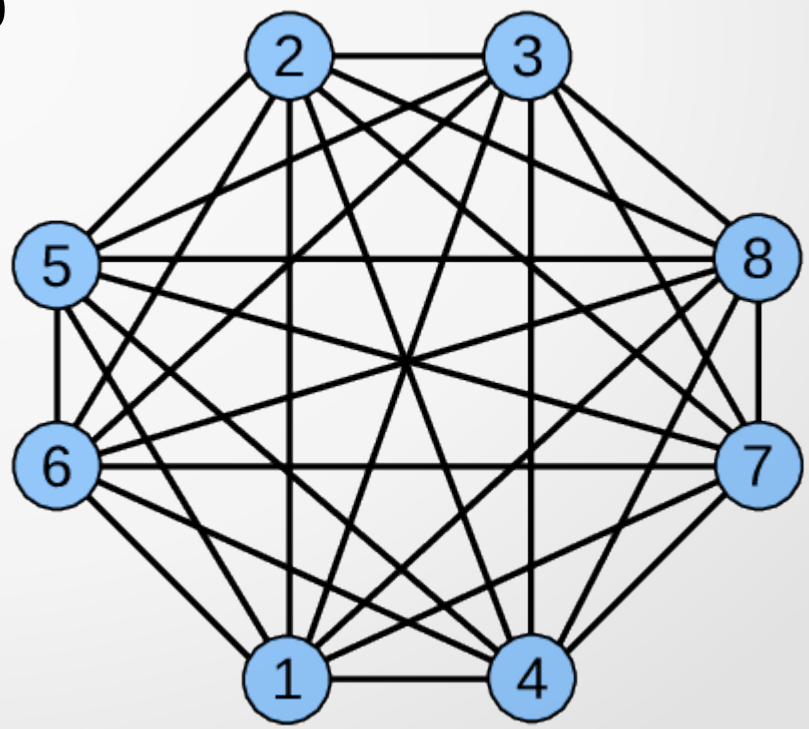
Solução

- Um algoritmo hierárquico baseado em densidade é uma necessidade!
- Sob perspectiva de grafos, o DBSCAN pode ser visto como:
  - Vértices: observações densas
  - Vértices adjacentes: observações diretamente alcançáveis em ε



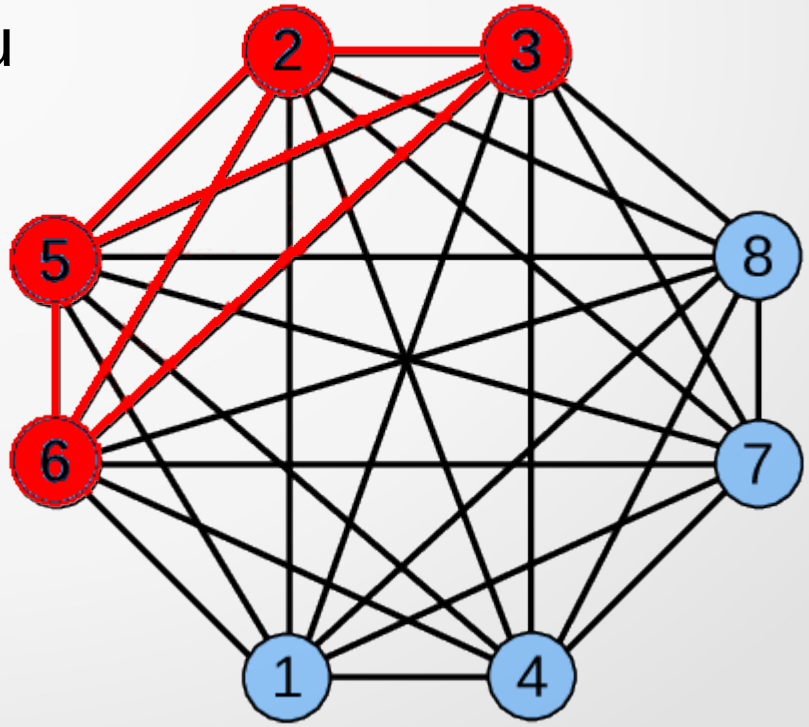
DBSCAN como grafo

- Grafo completo ponderado **G**:
  - vértices: todas as observações
  - pesos das arestas (incluindo arestas próprias):
    - raio mínimo ε_{pq} para que p e q sejam densos e diretamente alcançáveis por ε_{pq}
    - existe um nível de densidade ($\lambda = 1/\varepsilon_{pq}$) abaixo do qual qualquer aresta (p,q) atende a esses critérios



DBSCAN como grafo

- Grupos: são componentes conexos e subgrafos de **G**:
  - arestas: mantêm apenas os vértices menores do que o limiar ε definido pelo usuário
  - vértices: são mantidos apenas o que possuem grau maior que zero (não sejam ruído)



Ideia principal

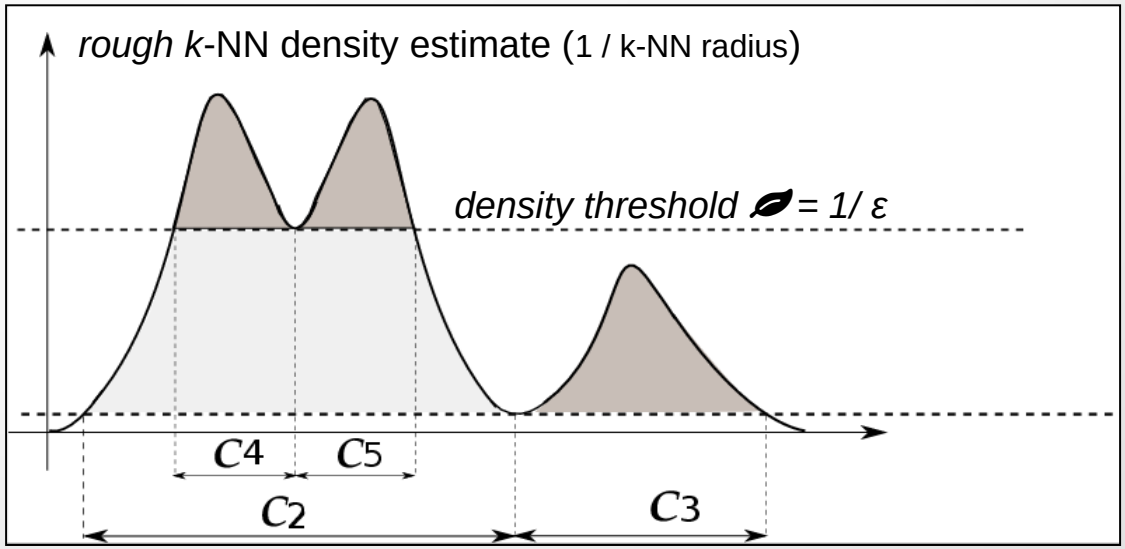
- Dados os conceitos que formulam o DBSCAN como grafo
  - Porque manter o limiar de densidade definido pelo usuário?
    - Se podemos ordenar as arestas de G e removê-las em ordem decrescente de peso (ε)?

Ideia principal

- Dados os conceitos que formulam o DBSCAN como grafo
  - Porque manter o limiar de densidade definido pelo usuário?
    - Se podemos ordenar as arestas de G e removê-las em ordem decrescente de peso (ε)?

Hierarchical DBSCAN (HDBSCAN*)

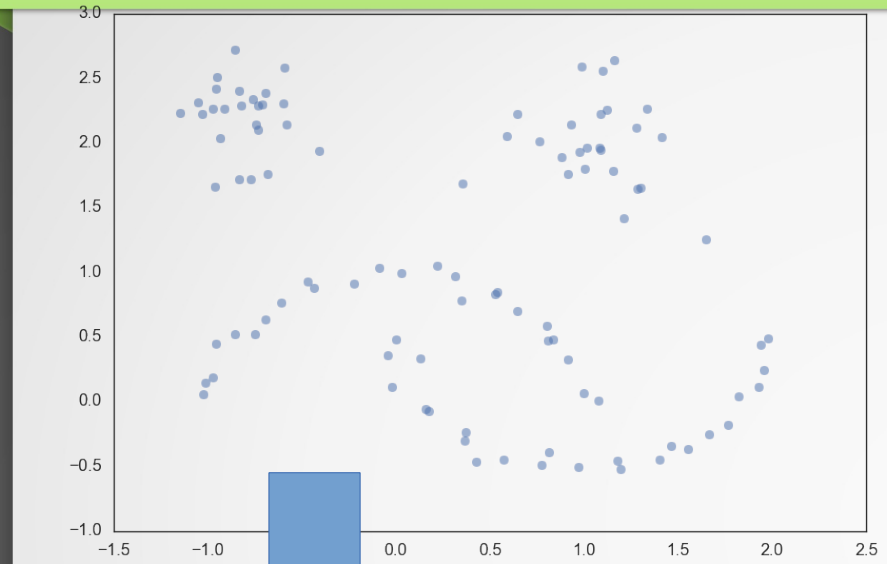
- Ordenar as arestas e removê-las em ordem decrescente de ε
- Componentes conexos aninhados (*clusters*) são criados hierarquicamente ...
  - … para valores crescentes de limiar de densidade ($\lambda = 1/\varepsilon_{pq}$)
- Produz todas as soluções de DBSCAN para $\varepsilon \in [0,\infty)$!

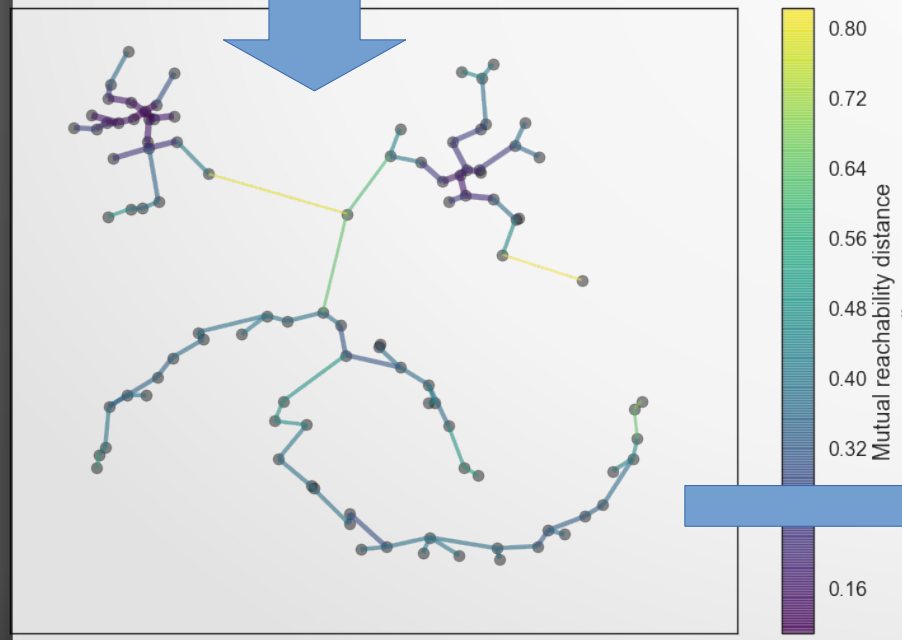


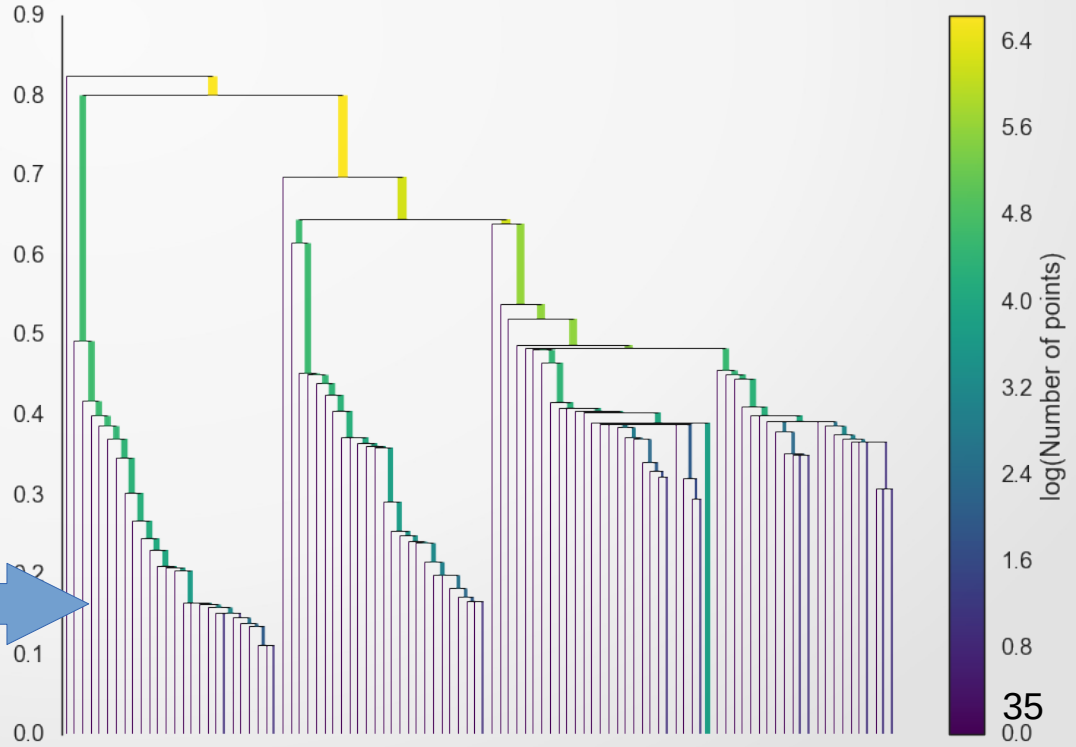
Relação com Ligação Simples

- HDBSCAN* pode ser visto como o algoritmo hierárquico de ligação simples (LS) ...
  - ... sobre o espaço de distâncias de alcance de densidade
- Por isso o HDBSCAN* também é conhecido como LS robusto!
    - E poder ser eficientemente obtido calculando a **MST** *de* **G** *no espaço de distâncias de alcance!*
- Mantêm as mesmas complexidades de espaço e tempo do DBSCAN original

Exemplo

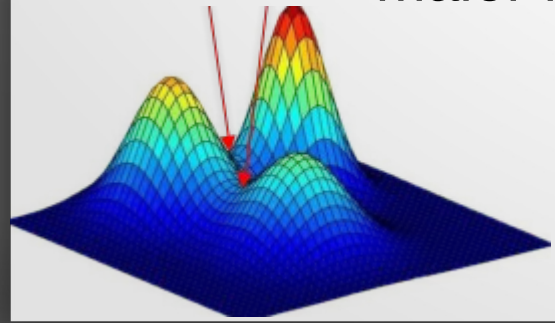

3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
0.0
0.5
1.0
1.5
2.0
2.5


0.80
0.72
0.64
0.56
0.48
0.40
0.32
0.16
Mutual reachability distance


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.1
0.0
distance
6.4
5.6
4.8
4.0
3.2
2.4
1.6
0.8
0.0
log(Number of points)

Resumindo HDBSCAN*

- **Generalização do DBSCAN**: hierárquico, sem limiar de densidade crítico definido pelo usuário
- **Generalização do SL**: capacidade robusta de modelagem de ruído
- **Interpretações estatísticas**:
  - Grupos e árvores com contorno de densidade de Hartigan
  - MST no espaço transformado de distâncias de alcance:
    - Para qualquer p e q: entre todos os caminhos possíveis entre p e q, o caminho de p e q no MST minimiza o valor máximo do peso dentre os caminhos (problema min-max)
      - maior menor nível de densidade conectando p e q



HDBSCAN*

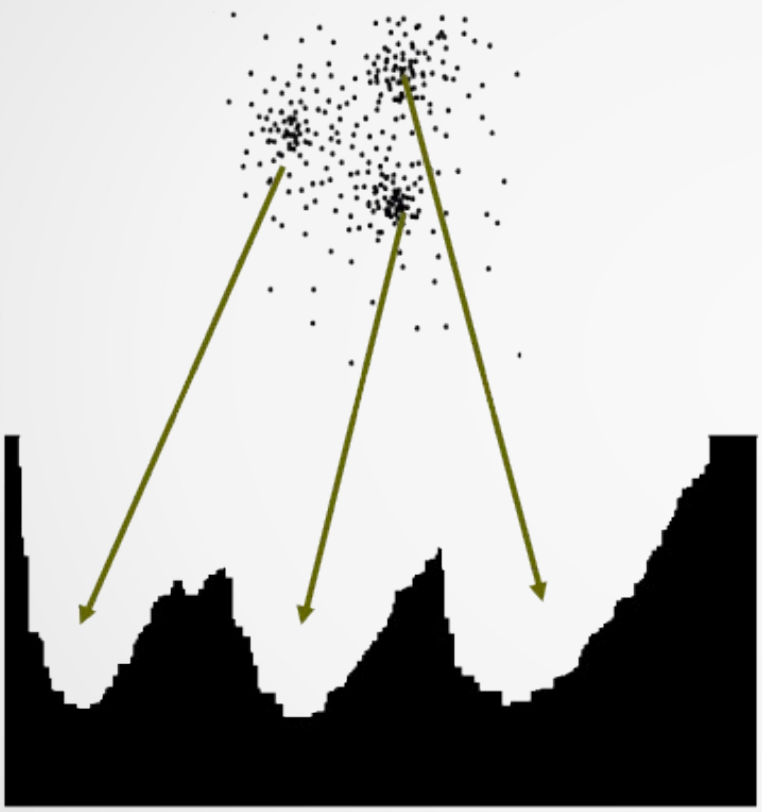
ACM Transactions on Knowledge Discovery from Data, Vol. 10, No. 1, Article 5, Publication date: July 2015.

Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection

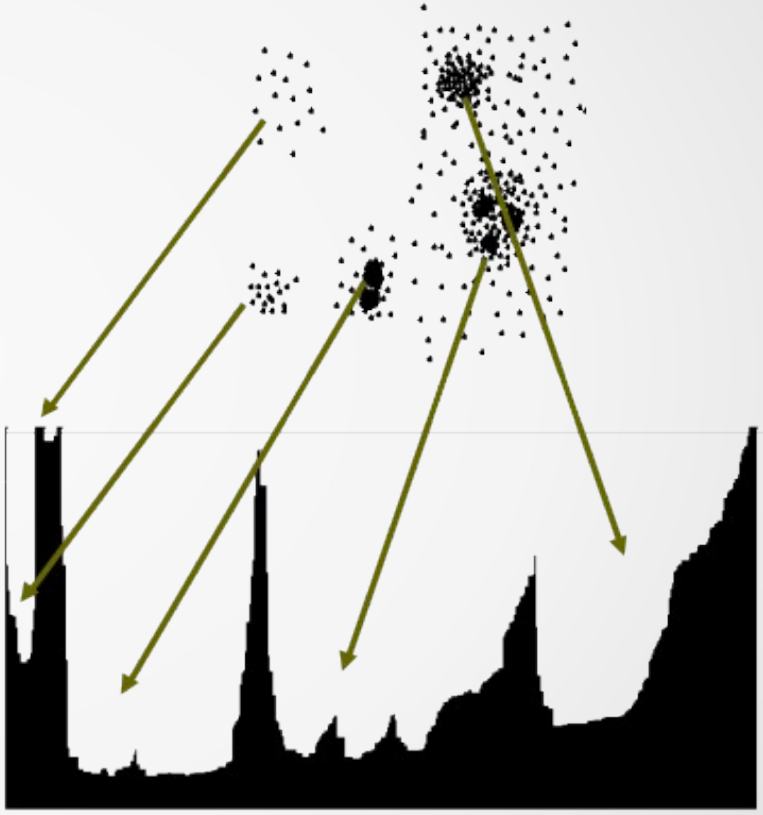
RICARDO J. G. B. CAMPELLO, Department of Computer Sciences, University of São Paulo, Brazil
DAVOUD MOULAVI, Department of Computing Science, University of Alberta, Canada
ARTHUR ZIMEK, Ludwig-Maximilians-Universität München, Germany
JÖRG SANDER, Department of Computing Science, University of Alberta, Canada

An integrated framework for density-based cluster analysis, outlier detection, and data visualization is introduced in this article. The main module consists of an algorithm to compute hierarchical estimates of the level sets of a density, following Hartigan's classic model of density-contour clusters and trees. Such an algorithm generalizes and improves existing density-based clustering techniques with respect to different aspects. It provides as a result a complete clustering hierarchy composed of all possible density-based clusters following the nonparametric model adopted, for an infinite range of density thresholds. The resulting hierarchy can be easily processed so as to provide multiple ways for data visualization and exploration. It can also be further postprocessed so that: (i) a normalized score of "outlierness" can be assigned to each data object, which unifies both the global and local perspectives of outliers into a single definition; and (ii) a "flat" (i.e., nonhierarchical) clustering solution composed of clusters extracted from local cuts through the cluster tree (possibly corresponding to different density thresholds) can be obtained, either in an unsupervised or in a semisupervised way. In the unsupervised scenario, the algorithm corresponding to this postprocessing module provides a global, optimal solution to the formal problem of maximizing the overall stability of the

Visualização: Reachability Plot



cluster ordering

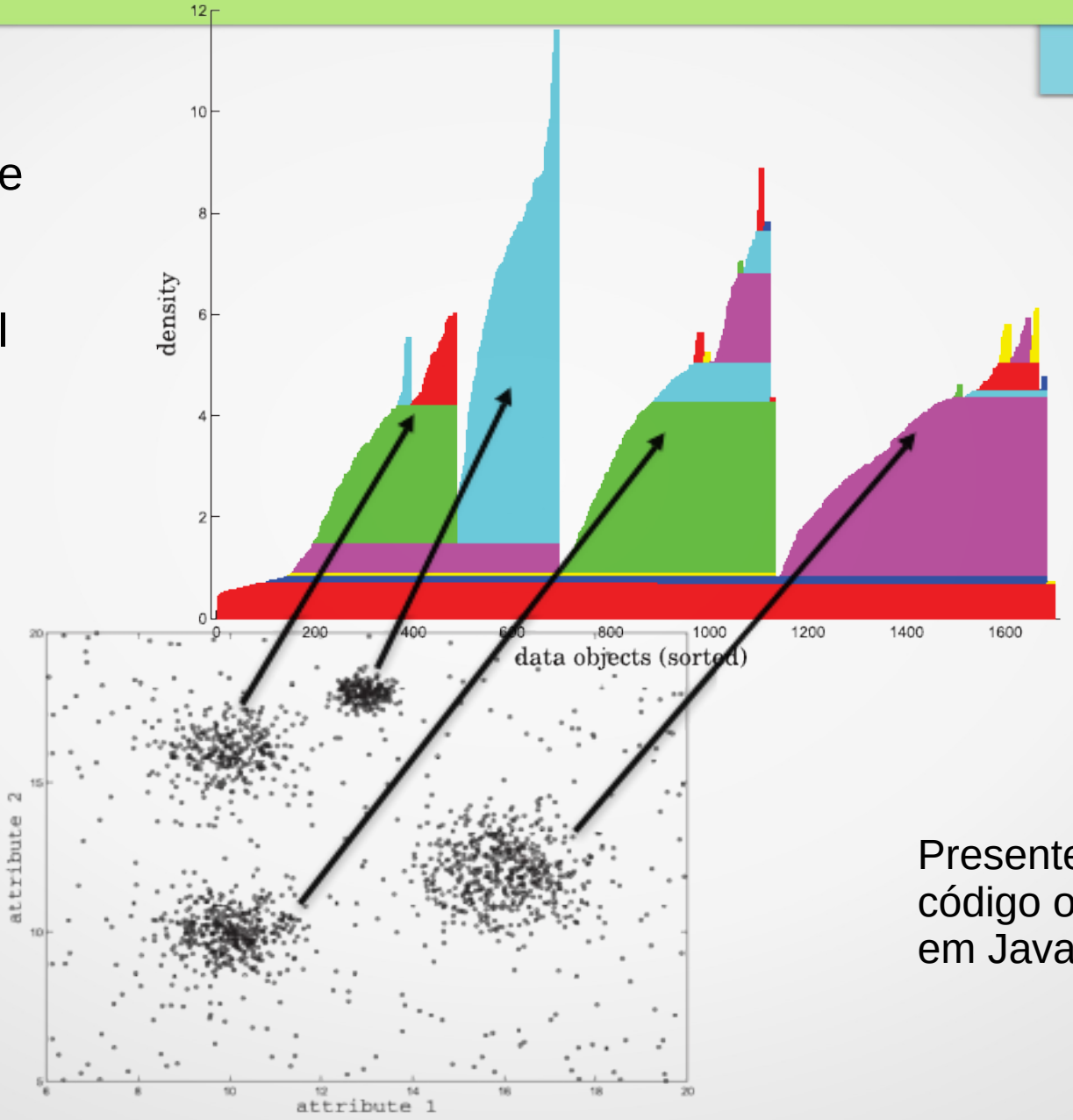


cluster ordering

Presente no código original em Java

Visualização: Silhouette Plots

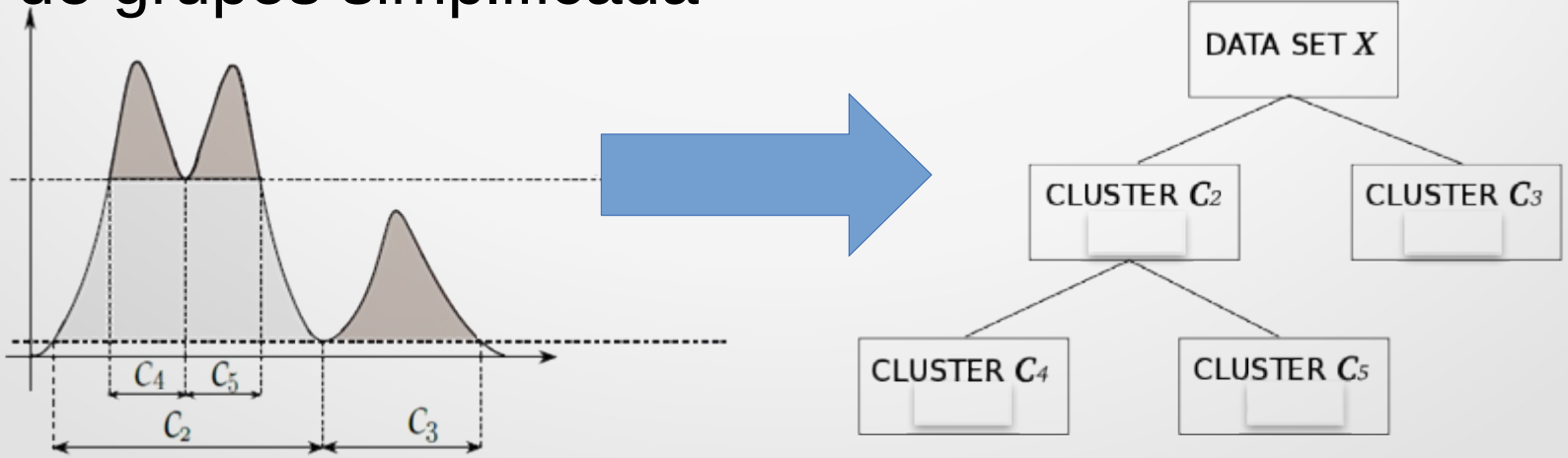
Visualização de estimativa de densidade unidimensional de dados multivariados!



Presente no código original em Java

Simplificação de hierarquia

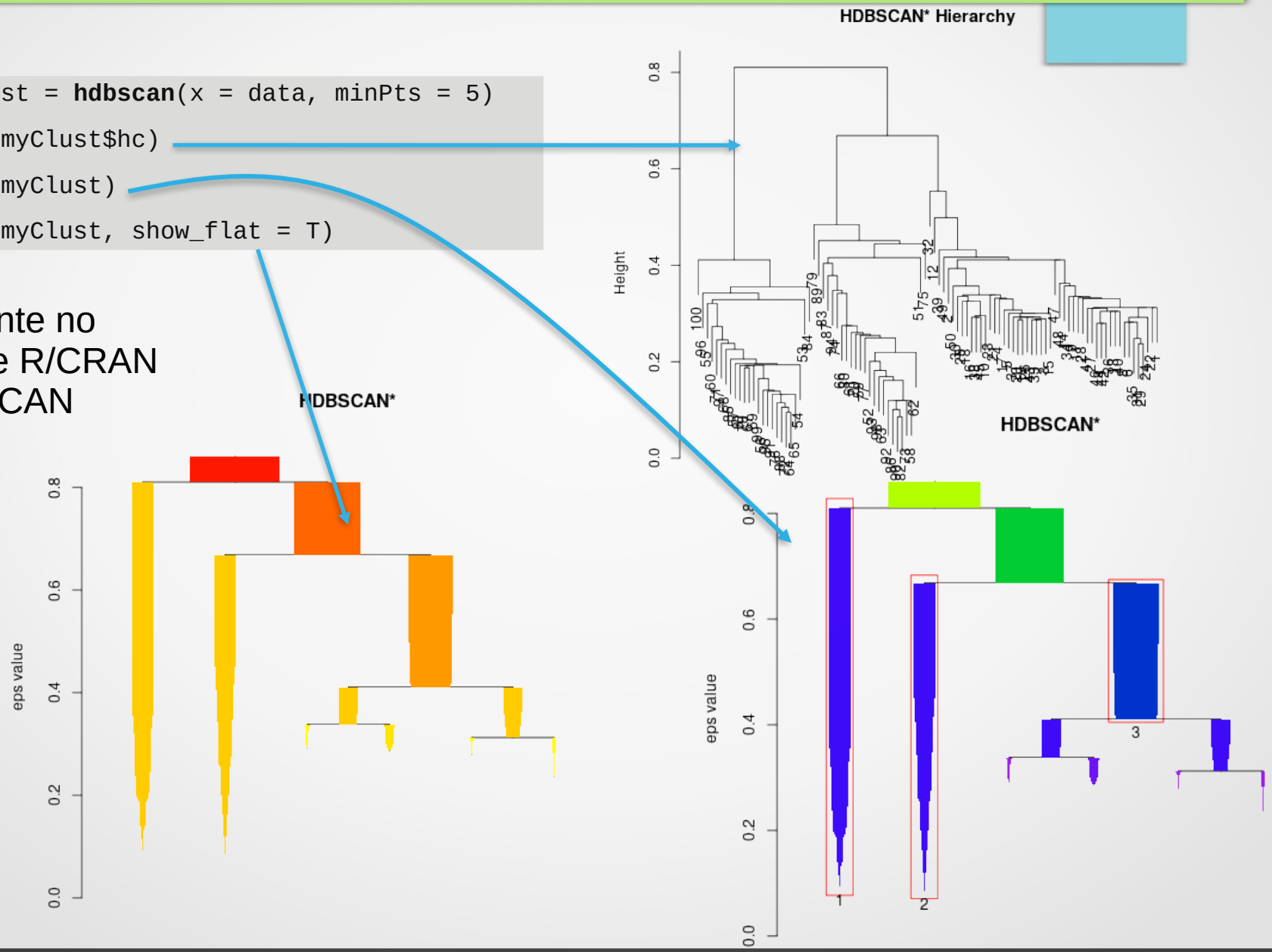
- Percebendo que...
  - a hierarquia do agrupamento é construída aumentando o limiar de densidade ε
- Ideia principal:
  - simplesmente perdendo componentes de ruído/ espúrios, o grupo encolhe (mas não divide)
  - simplifica a hierarquia por ordens de magnitude: árvore de grupos simplificada



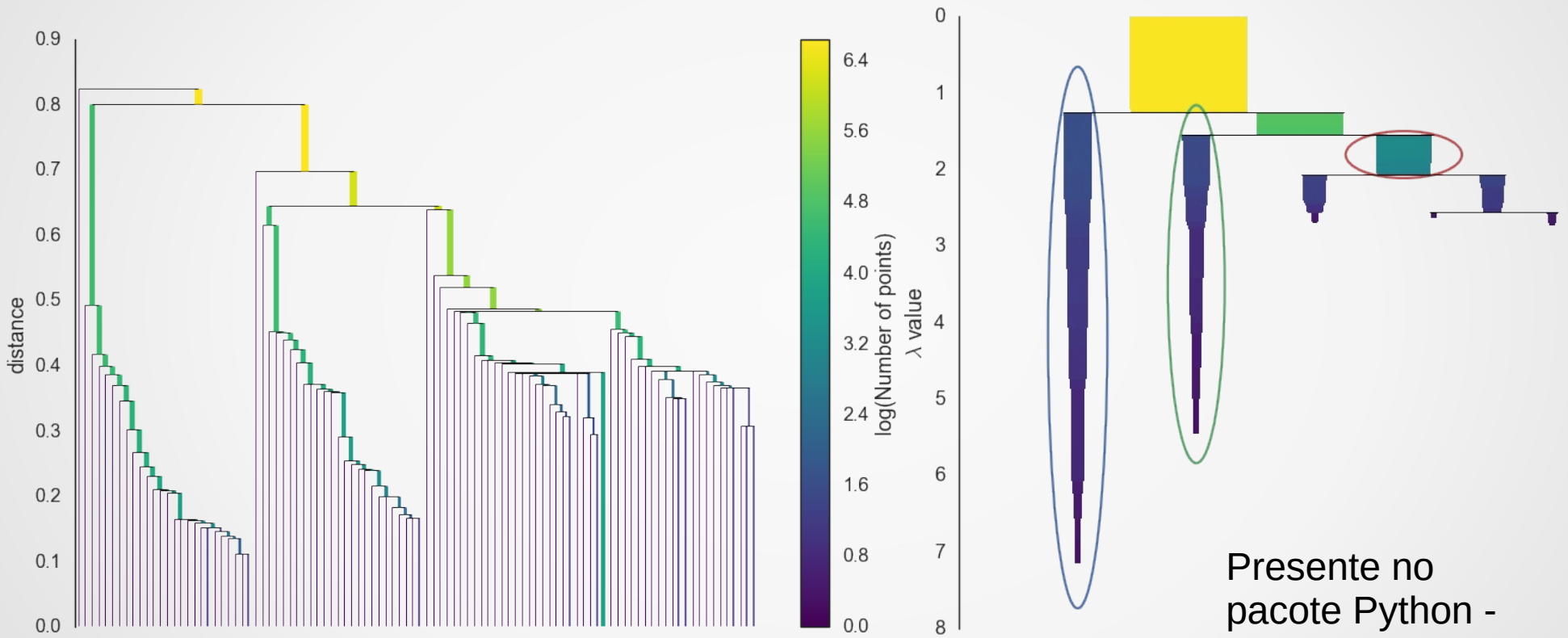
Visualização: Hierarquias

```
myClust = hdbscan(x = data, minPts = 5)
plot(myClust$hc)
plot(myClust)
Plot(myClust, show_flat = T)
```

Presente no pacote R/CRAN - DBSCAN



Visualização: Hierarquias

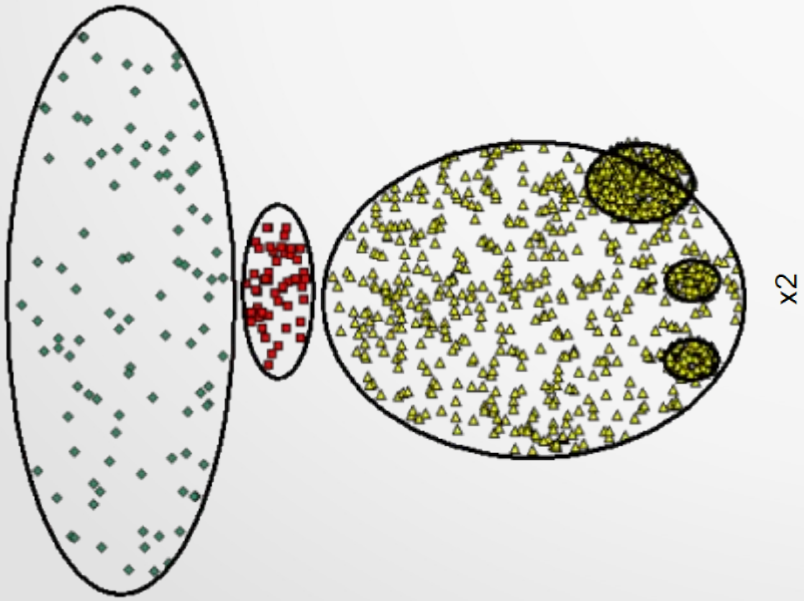


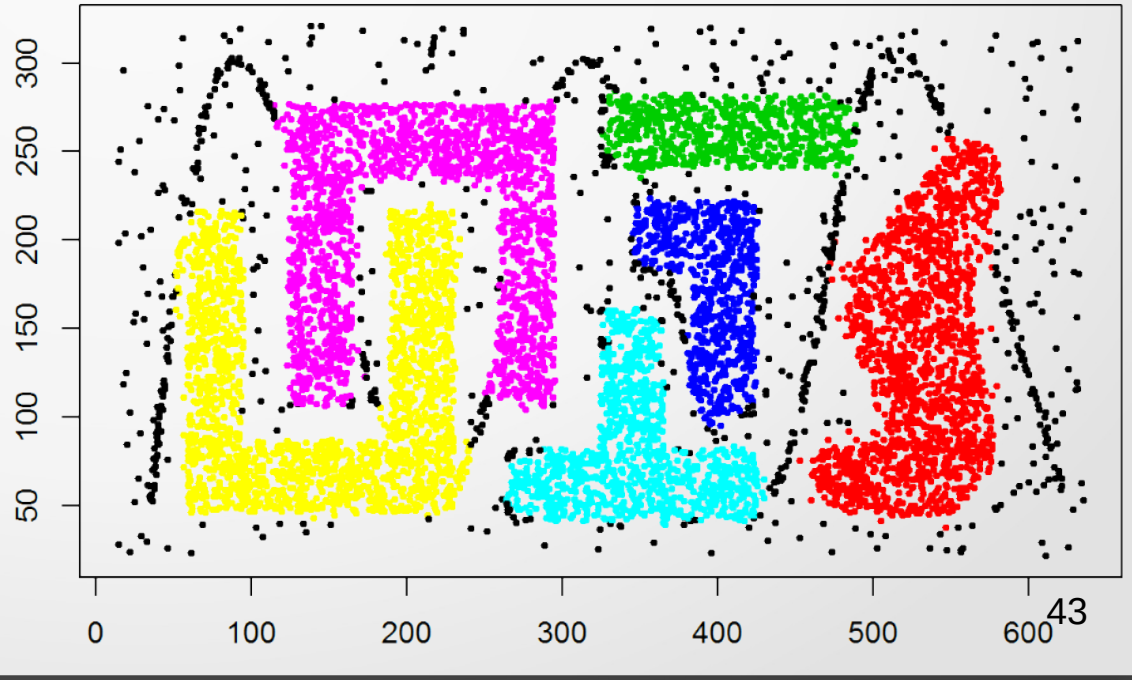
Presente no pacote Python - Scikit-Learn

- SciPy 2016 vídeo no YouTube pelo criador do pacote/notebook Python https://www.youtube.com/watch?v=AgPQ76RIi6A

Mas, e se quiser partição?

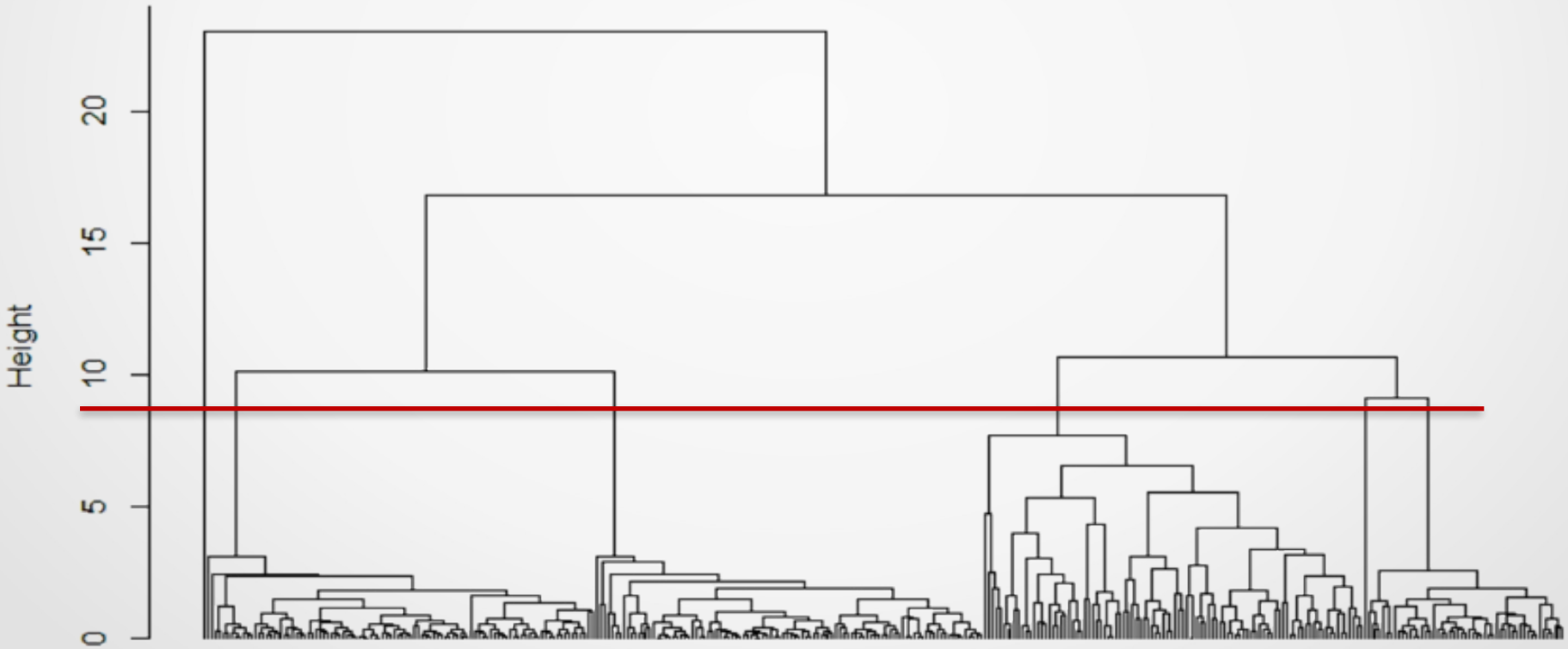
- Apesar da hierarquia conter informação rica sobre a estrutura dos dados, o usuário ainda pode insistir em uma partição dos dados





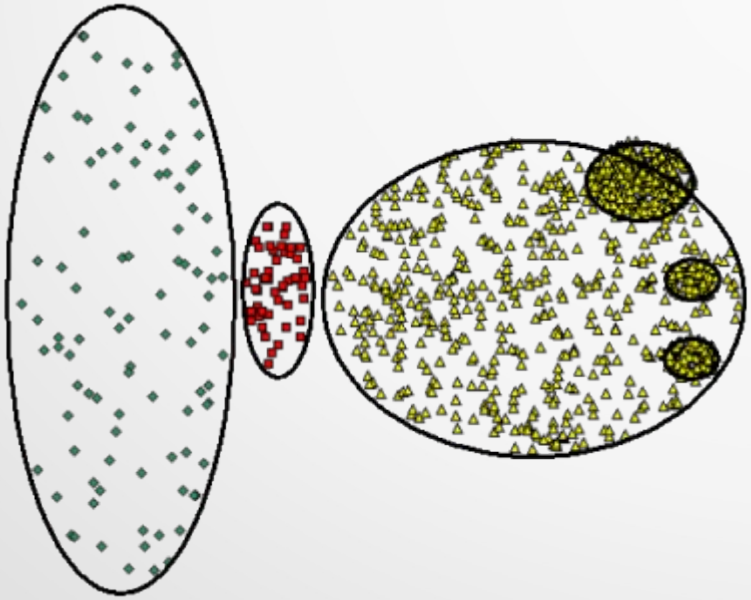
Corte Tradicional

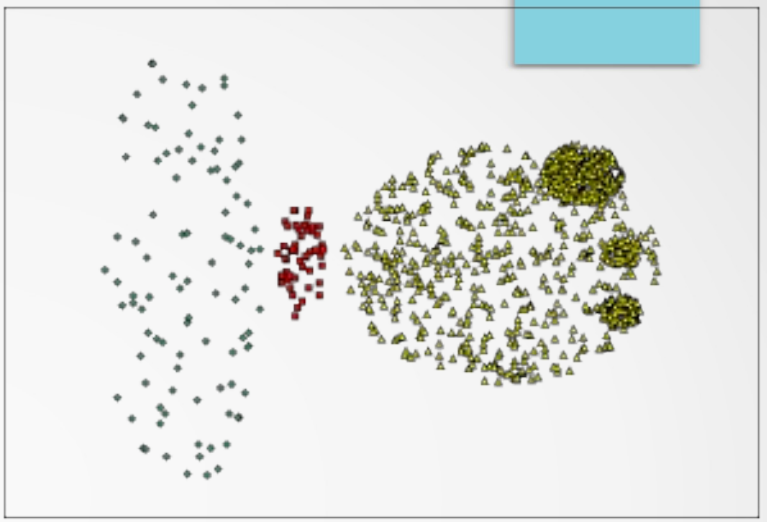
- Horizontal em um nível de densidade ε pré-definido



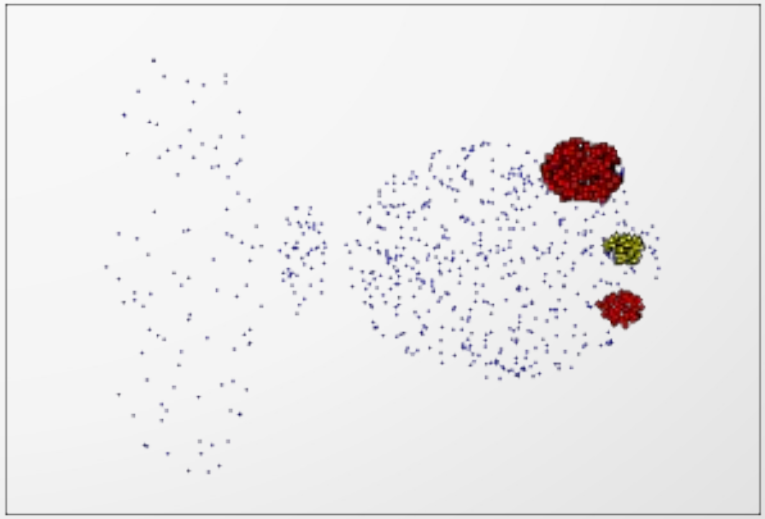
Contudo...

- O resultado é uma partição de DBSCAN
  - Mas já sabemos os problemas dessa abordagem





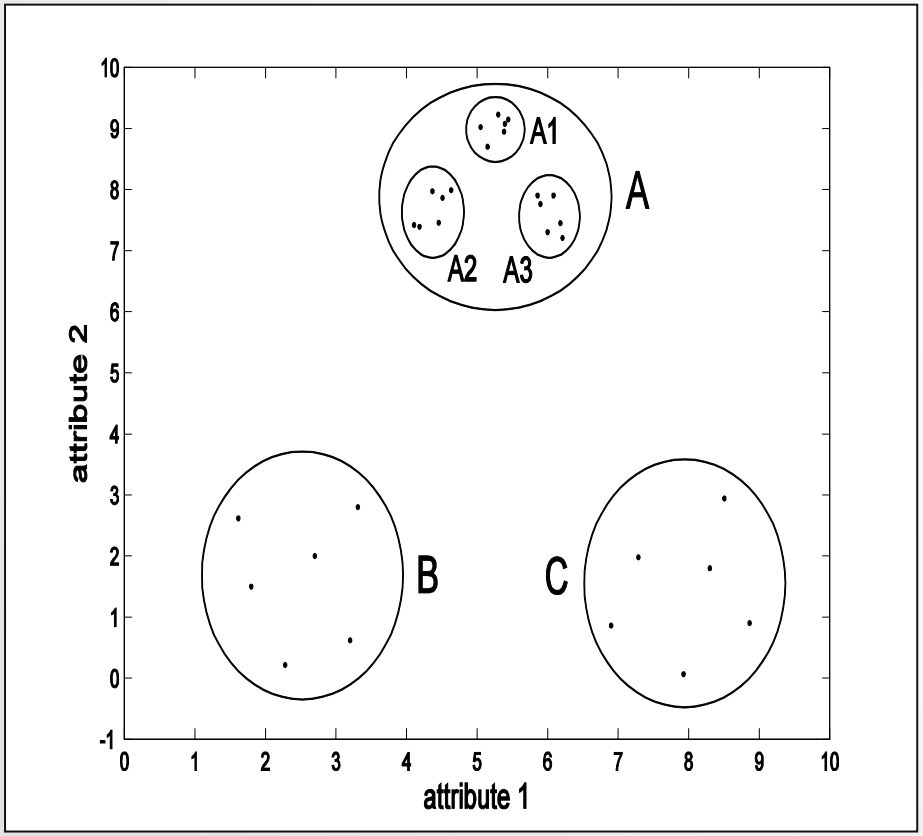
(MinPts=4, Eps=9.92).

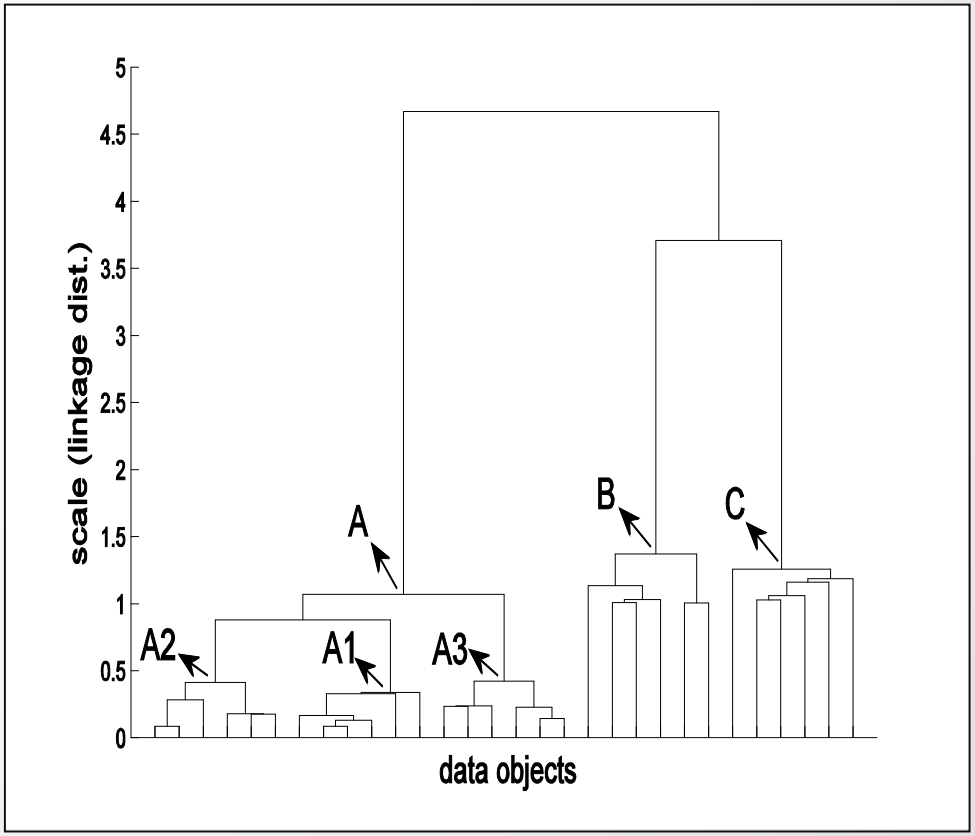


(MinPts=4, Eps=9.75)

Método tradicional é limitado!

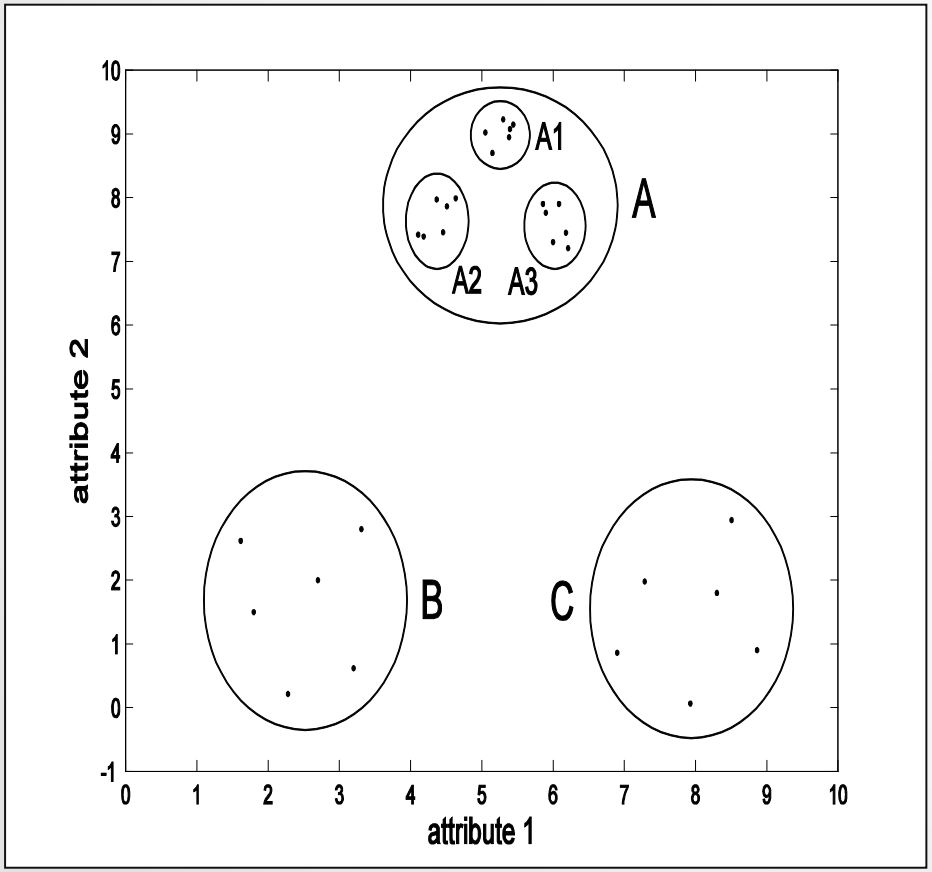
- Queremos que A1, A2, A3, B e C sejam opções viáveis para uma solução!

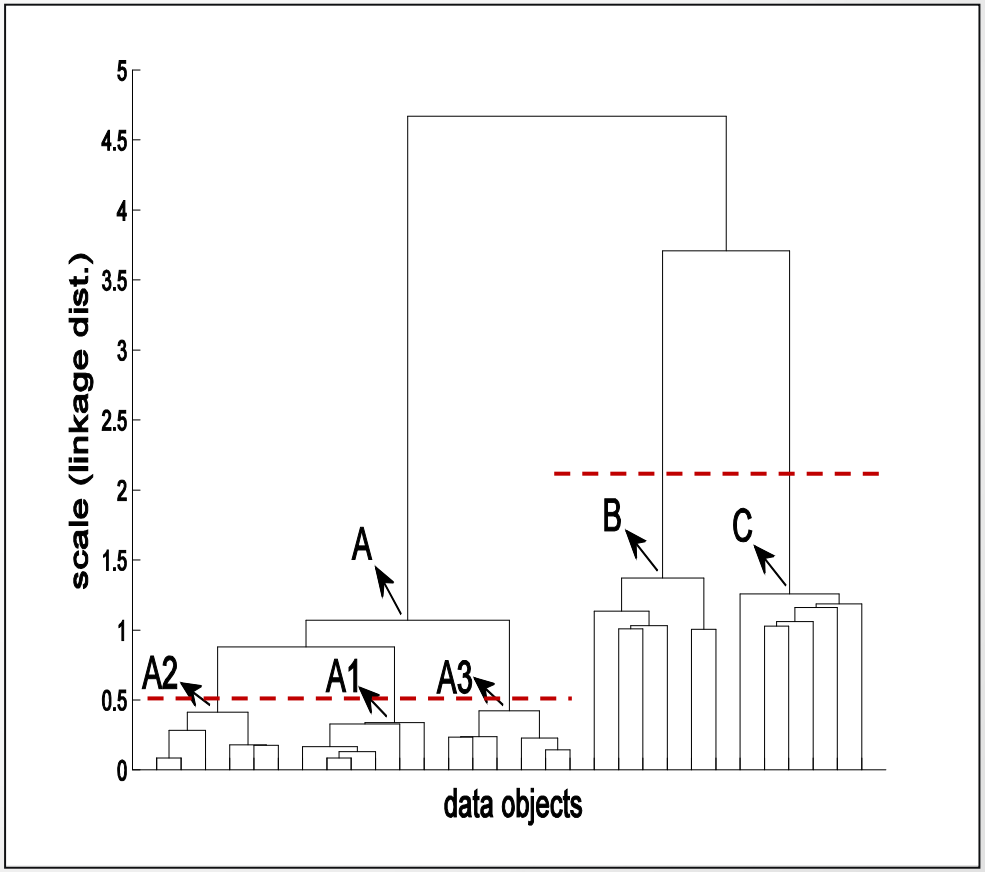




Opção melhor: cortes locais!

- Cortes locais permitem uma solução com múltiplas densidades!





HDBSCAN* utiliza o FOSC

Data Min Knowl Disc (2013) 27:344–371
DOI 10.1007/s10618-013-0311-4

A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies

R. J. G. B. Campello · D. Moulavi · A. Zimek · J. Sander

Received: 3 October 2012 / Accepted: 21 March 2013 / Published online: 4 April 2013

Abstract We introduce a framework for the optimal extraction of flat clusterings from local cuts through cluster hierarchies. The extraction of a flat clustering from a cluster tree is formulated as an optimization problem and a linear complexity algorithm is pre-

Cortes locais como otimização

- Dada uma hierarquia e seja δ_i a escolho do grupo $\mathbf{C}_i$, extrair um conjunto {$\mathbf{C}_i$, $\mathbf{C}_j$, …, $\mathbf{C}_k$} de grupos que maximiza o critério de qualidade:

$$J(\delta_1, \delta_2, \ldots, \delta_n)$$

- Tal que forme uma partição:

$$(\delta_i = 1 \wedge \delta_j = 1) \Rightarrow C_i \cup C_j = \varnothing$$

$$\delta_i = 0 \Rightarrow \exists\, \delta_j = 1 : C_i \subset C_j \vee C_j \subset C_i$$

Supondo (facilitando o problema)

- O critério de otimização pode ser decomposto como uma soma de termos tais que:
  - Cada termo S($\mathbf{C}_i$) representa a medida de qualidade do grupo $\mathbf{C}_i$
  - Cada termo pode ser calculado localmente, isto é:
    - Independentemente de outros grupos
    - E de seleções δ_i ,…, δ_n (ainda a serem feitas)

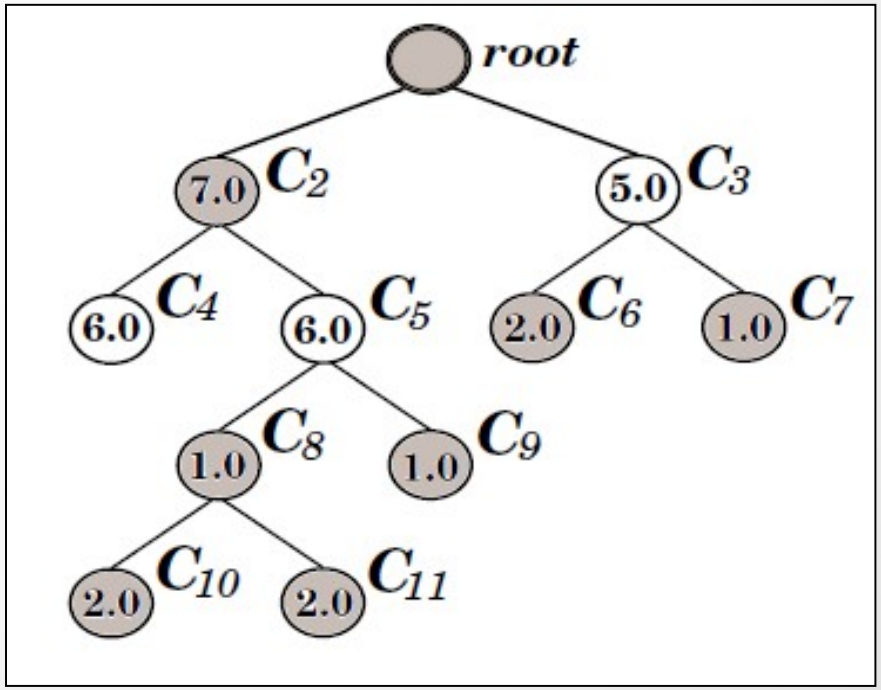
$$J = \sum_{i=1}^{n} \delta_i \cdot S(C_i)$$

Problema básico

- Dados:
  - Uma **hierarquia** de grupos representada por uma árvore
  - Um método para estimar a **qualidade** de cada grupo
  - Um método para **agregar** as qualidades estimadas a serem maximizadas
- Problema:
  - Escolher grupos de formem uma **partição**
  - **Maximizar** a qualidade agregada

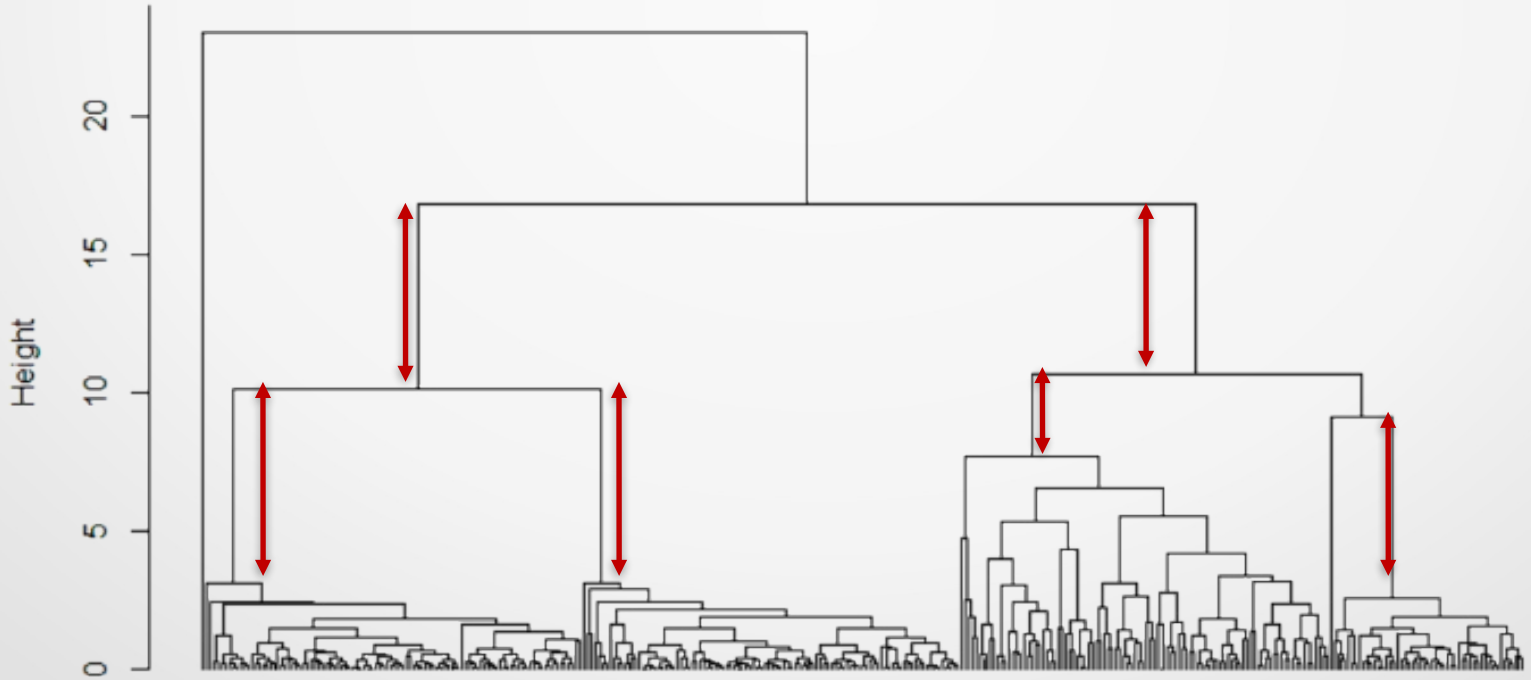
Solução

- Problema de programação dinâmica (de baixo para cima)
  - Achar solução de máximo global
  - Complexidade linear



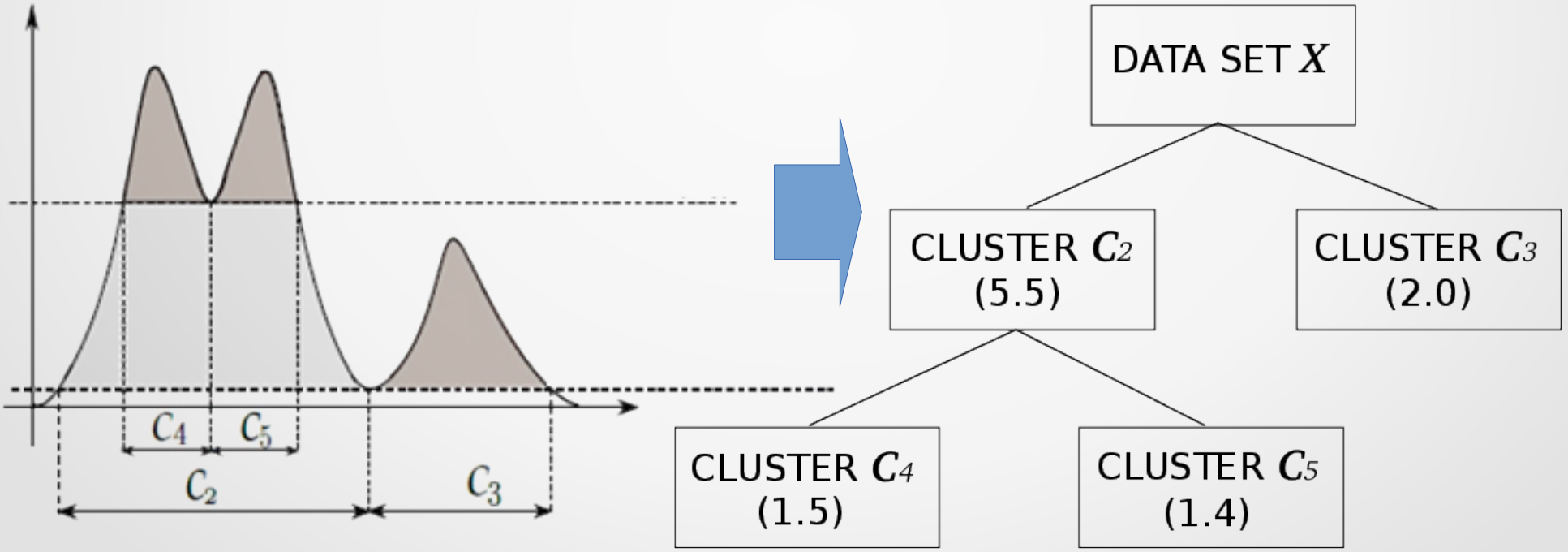
Medida de qualidade aplicável

- **Estabilidade de grupo**:
  - S($\mathbf{C}_i$) = soma dos tempos de vida de todas as observações no grupo $\mathbf{C}_i$



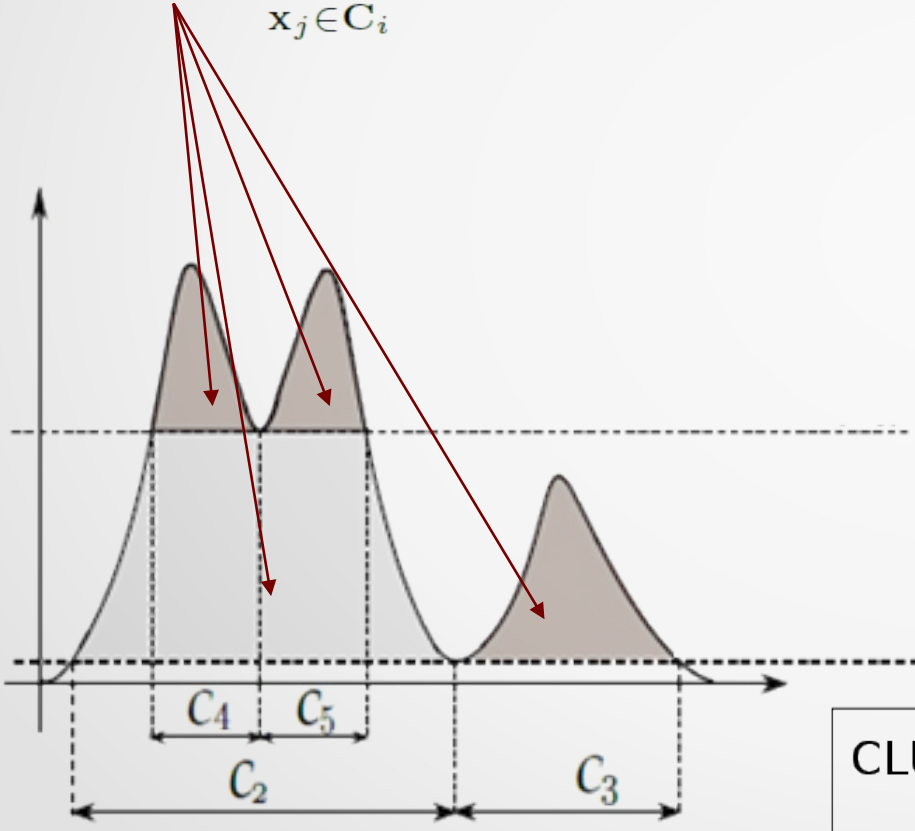
Estabilidade de grupo

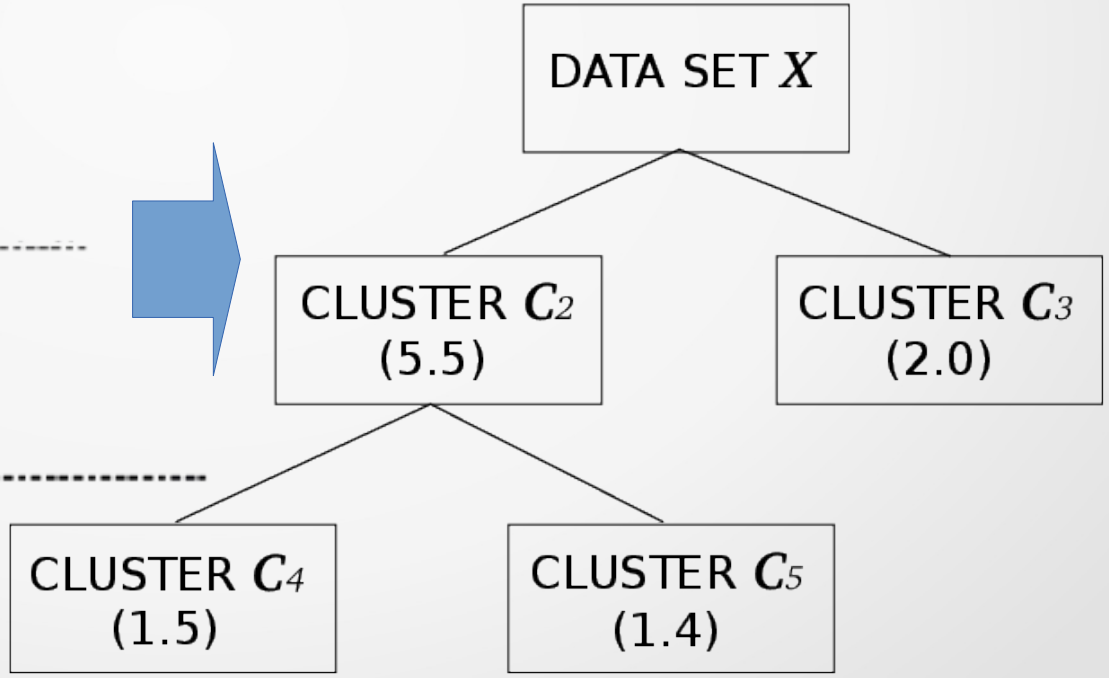
- Estabilidade de grupo para hierarquias baseadas em densidade pode ser definida como:
  - S(**C**$_i$) = excesso de massa relativo de grupo **C**$_i$



Estabilidade de grupo

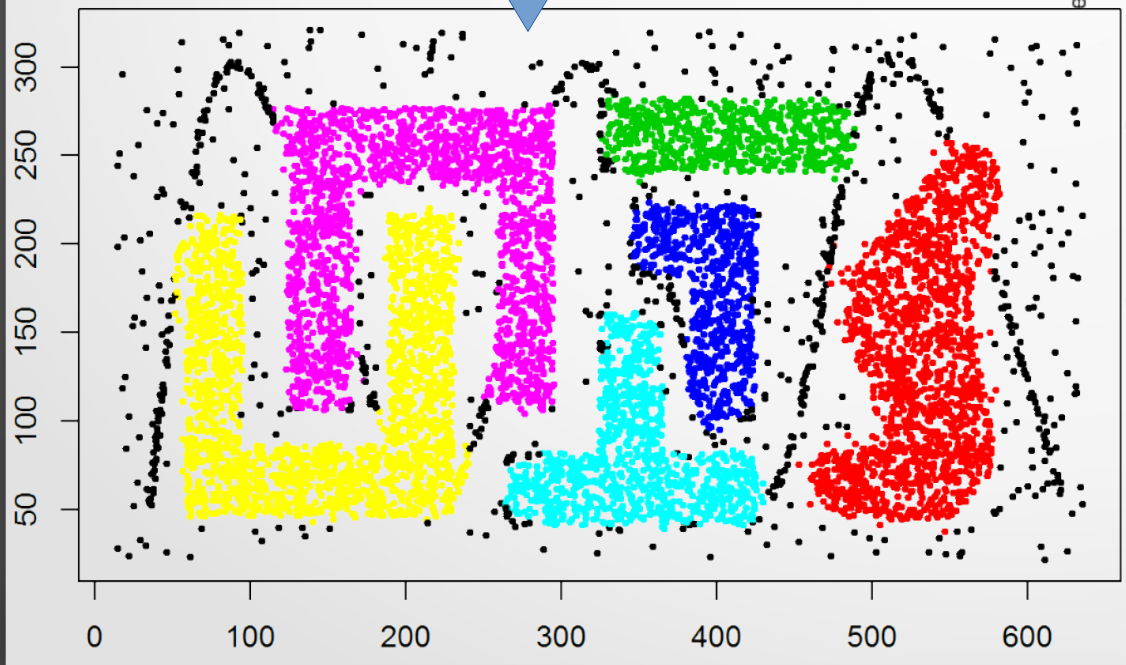
$$S(\mathbf{C}_i) = \sum_{\mathbf{x}_j \in \mathbf{C}_i} \left(\lambda_{max}(\mathbf{x}_j, \mathbf{C}_i) - \lambda_{min}(\mathbf{C}_i) \right) = \sum_{\mathbf{x}_j \in \mathbf{C}_i} \left(\frac{1}{\varepsilon_{min}(\mathbf{x}_j, \mathbf{C}_i)} - \frac{1}{\varepsilon_{max}(\mathbf{C}_i)} \right)$$

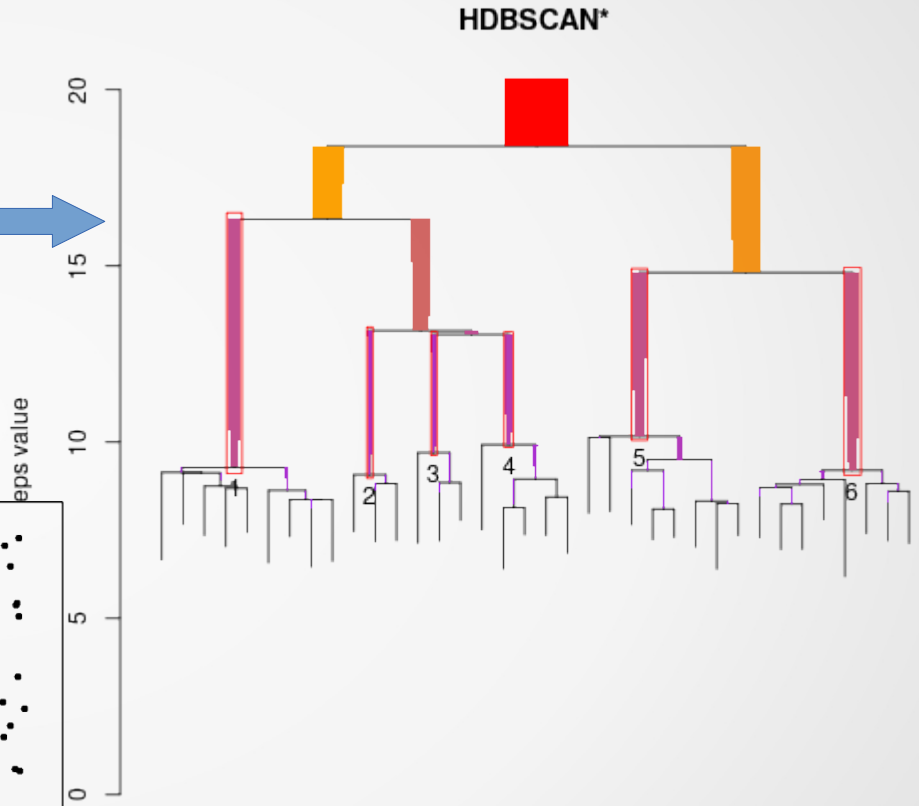




Exemplo de extração

```
myClust = hdbscan(x = DS3, minPts = 25)
plot(DS3, col = color_fun(myClust$cluster))
Plot(myClust, show_flat = T)
```





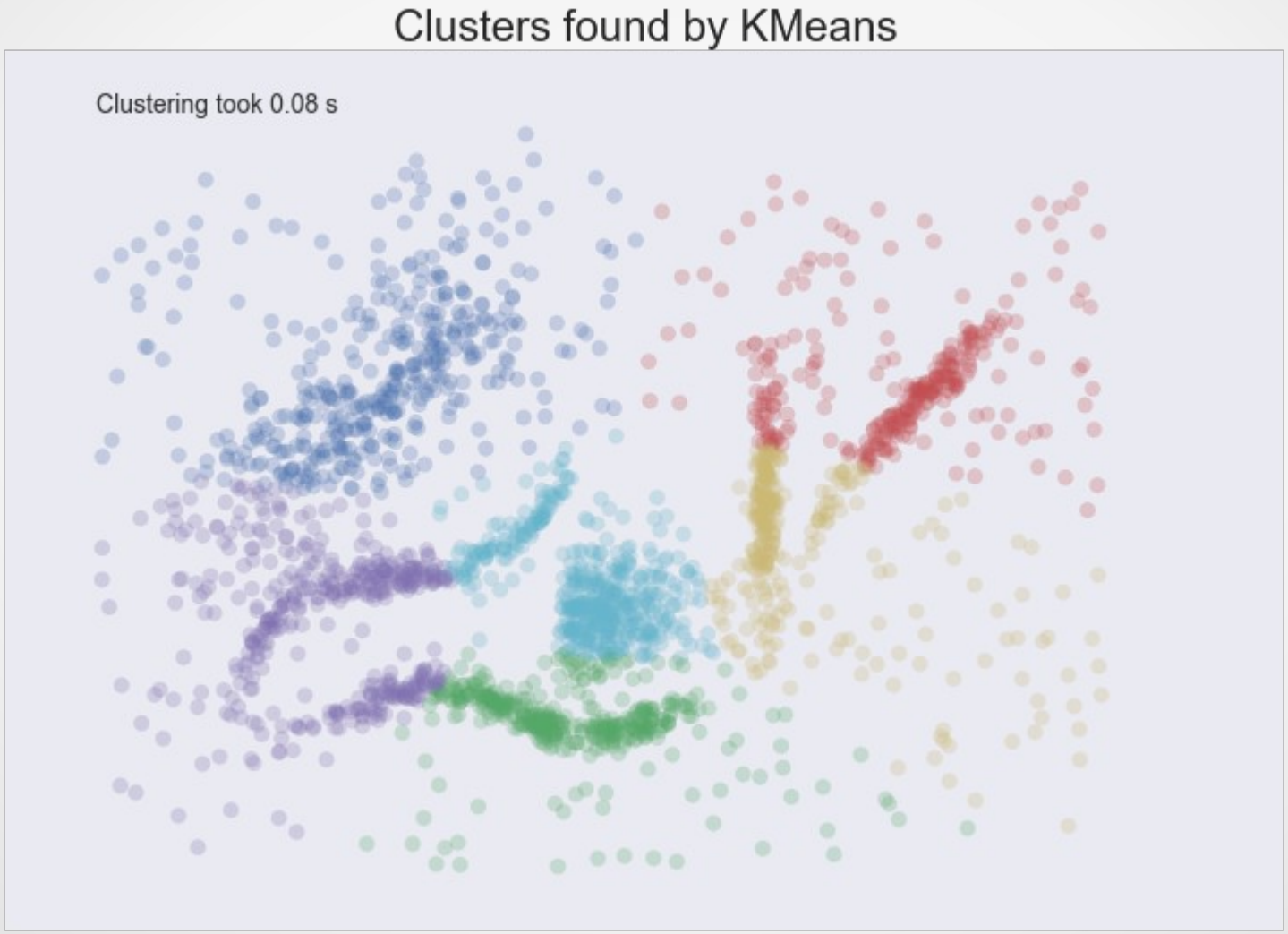
Presente no pacote R/CRAN - DBSCAN

Comparação com outros algoritmos

- Conjunto de dados

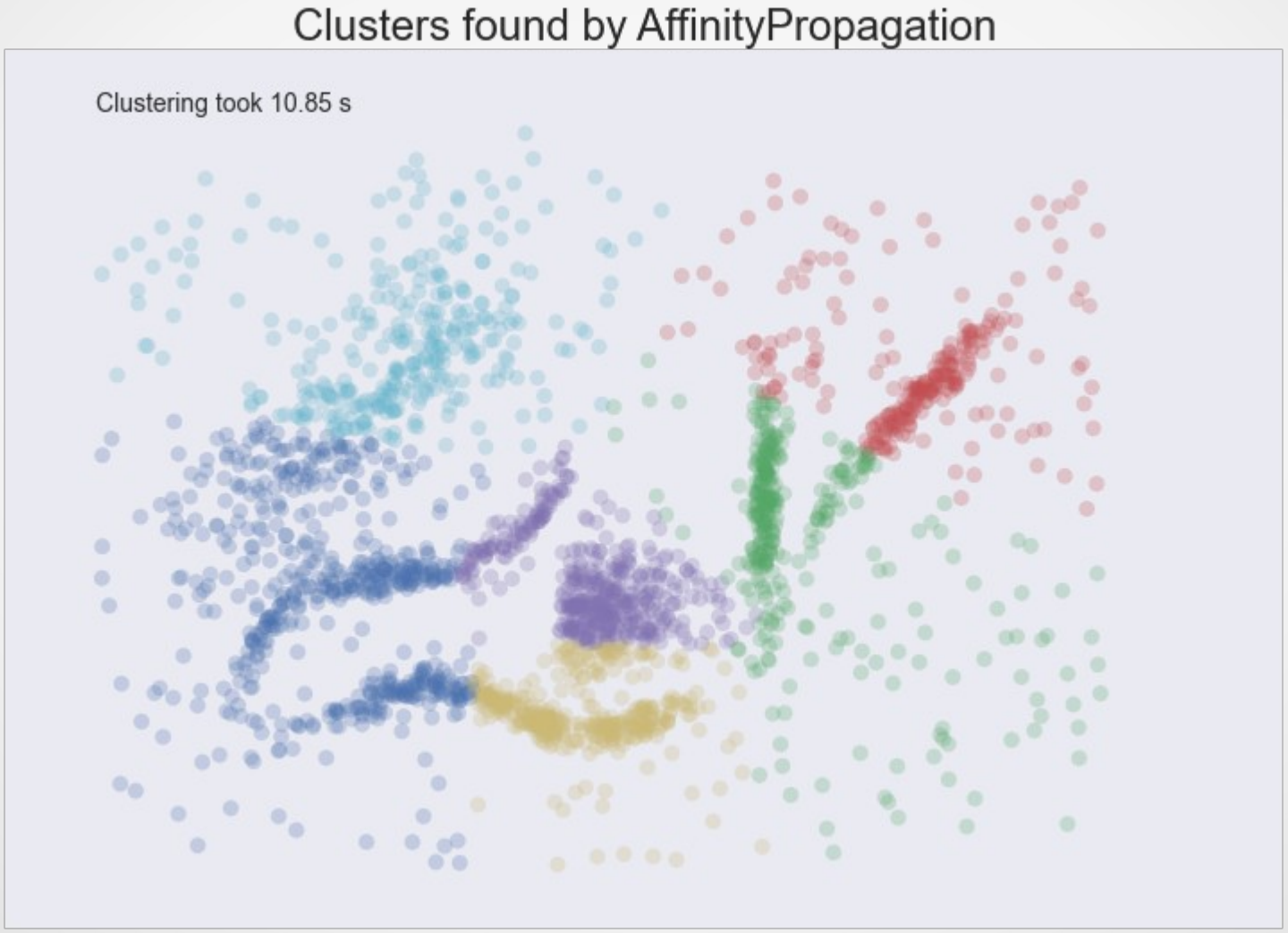
Documentação Pacote Python - hdbscan.readthedocs.io

Comparação com outros algoritmos



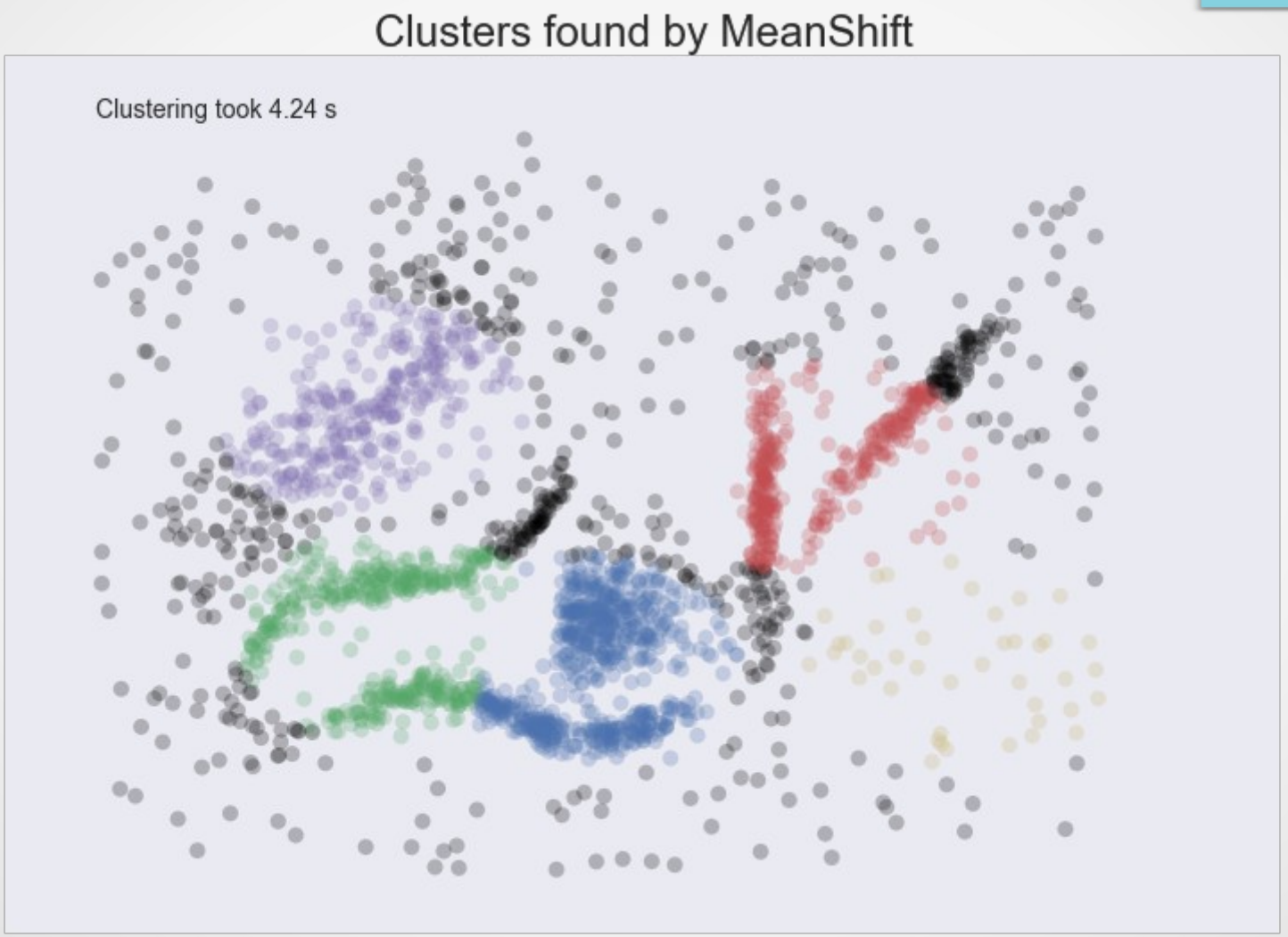
Documentação Pacote Python - hdbscan.readthedocs.io

Comparação com outros algoritmos



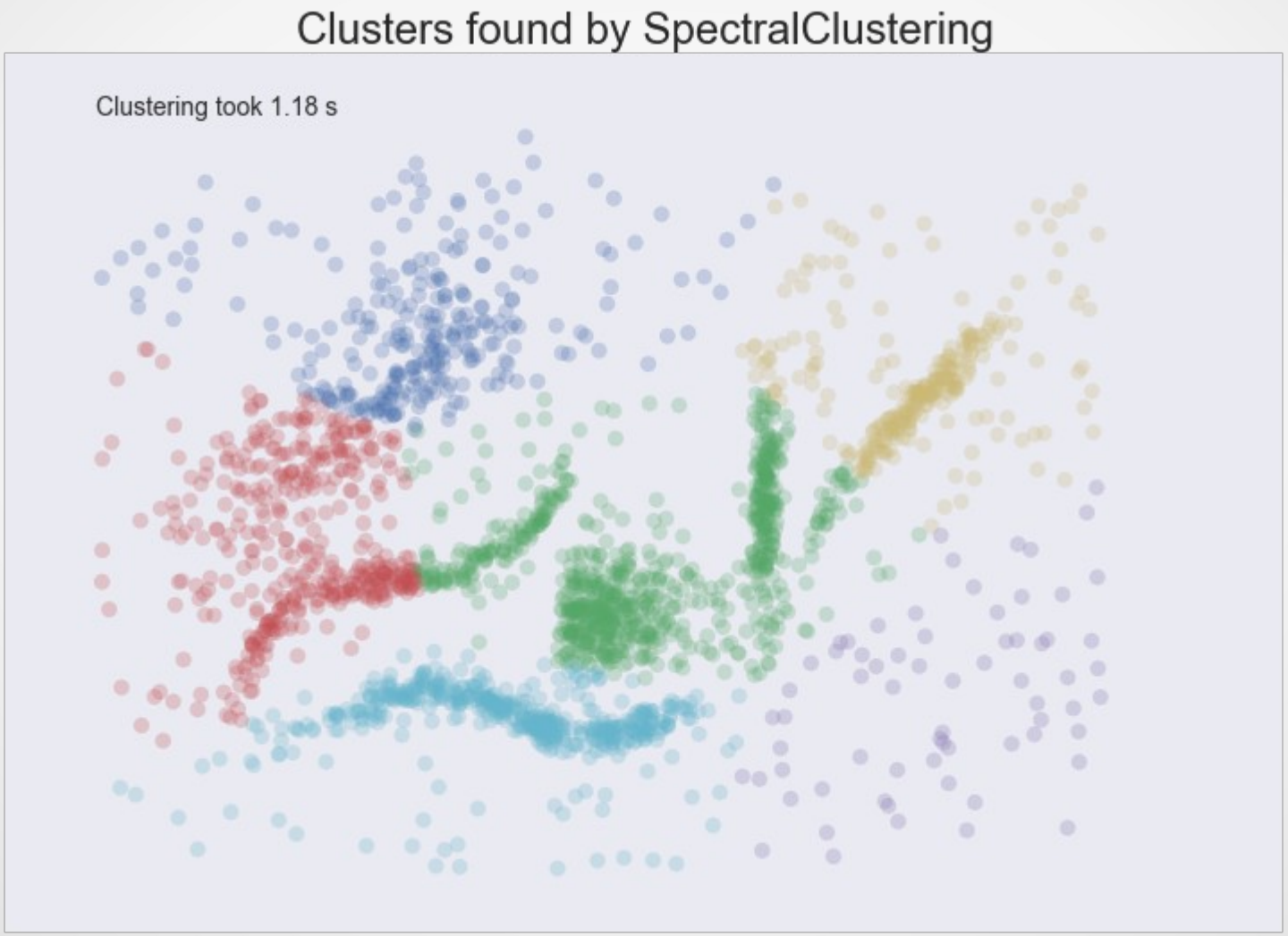
Documentação Pacote Python - hdbscan.readthedocs.io

Comparação com outros algoritmos



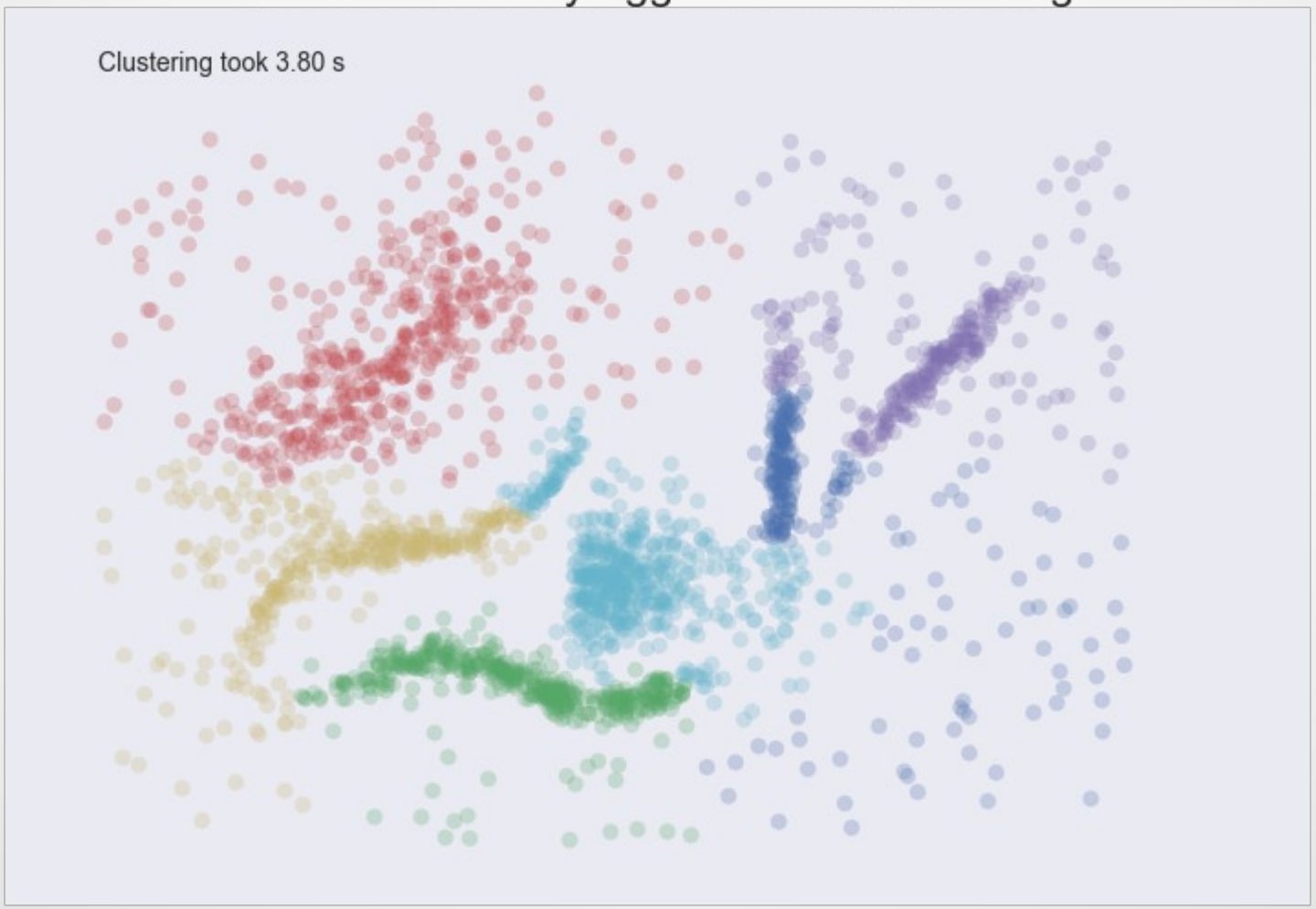
Documentação Pacote Python - hdbscan.readthedocs.io

Comparação com outros algoritmos



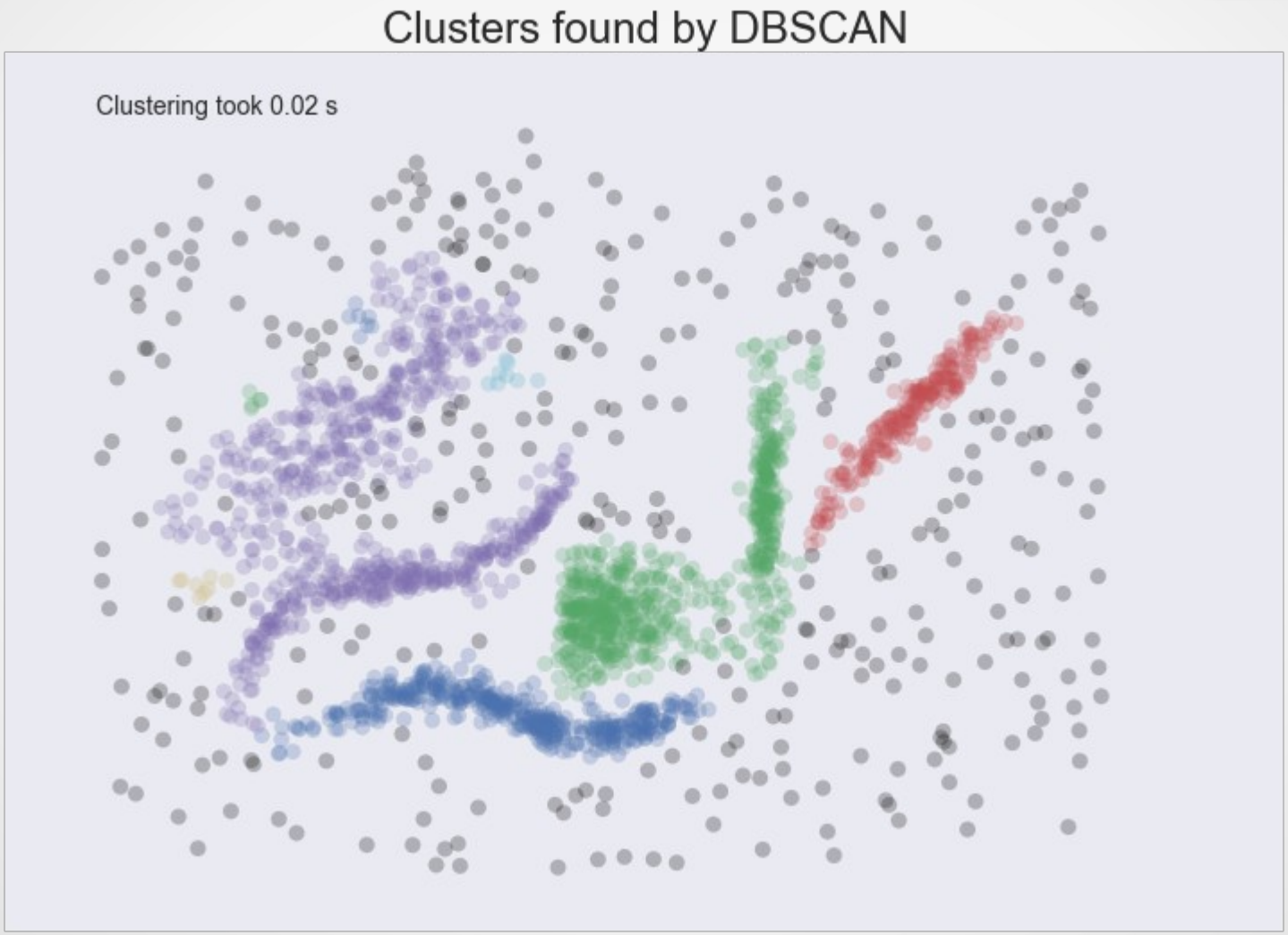
Documentação Pacote Python - hdbscan.readthedocs.io

Comparação com outros algoritmos



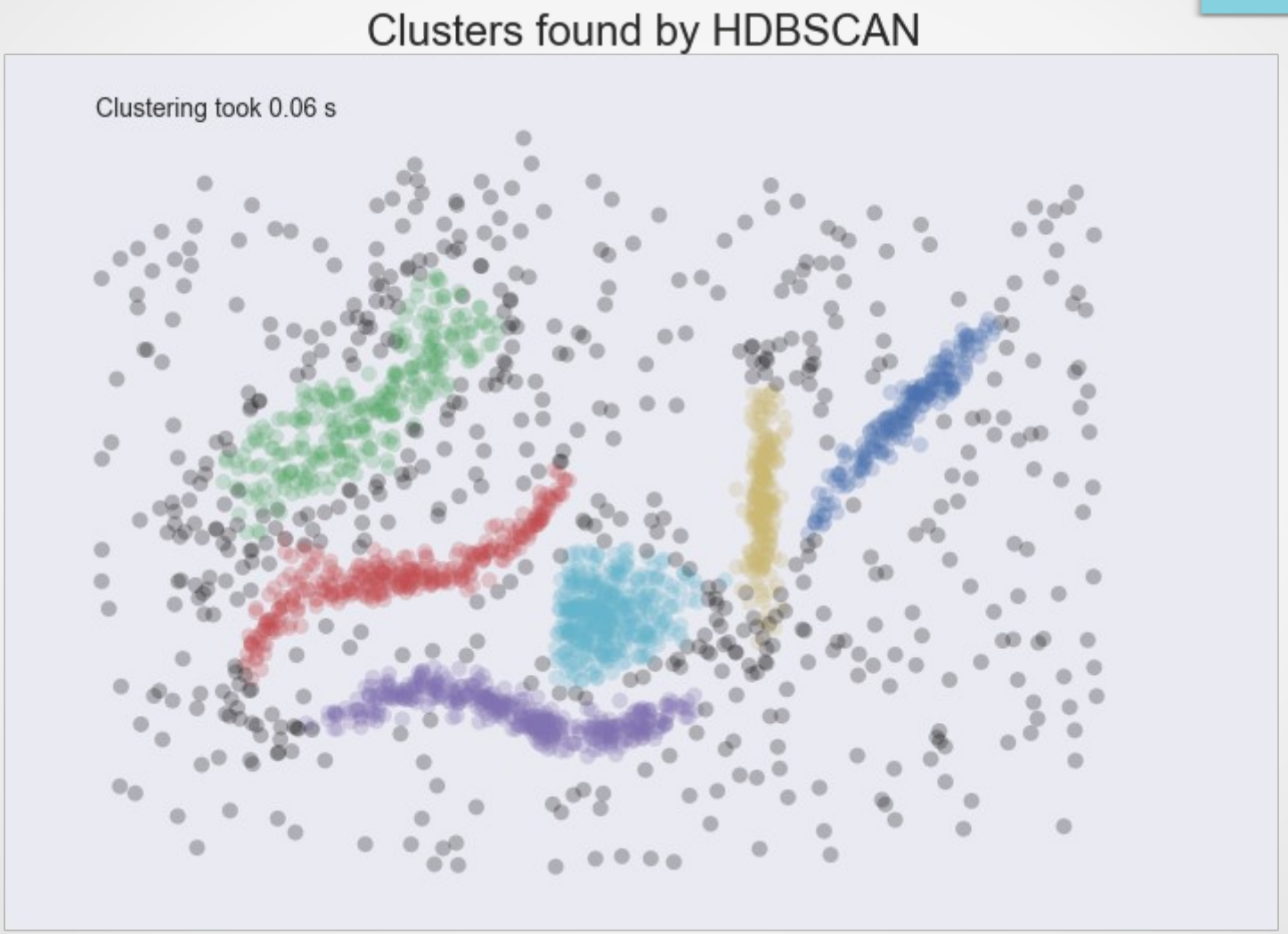
Documentação Pacote Python - hdbscan.readthedocs.io

Comparação com outros algoritmos



Documentação Pacote Python - hdbscan.readthedocs.io

Comparação com outros algoritmos



Documentação Pacote Python - hdbscan.readthedocs.io

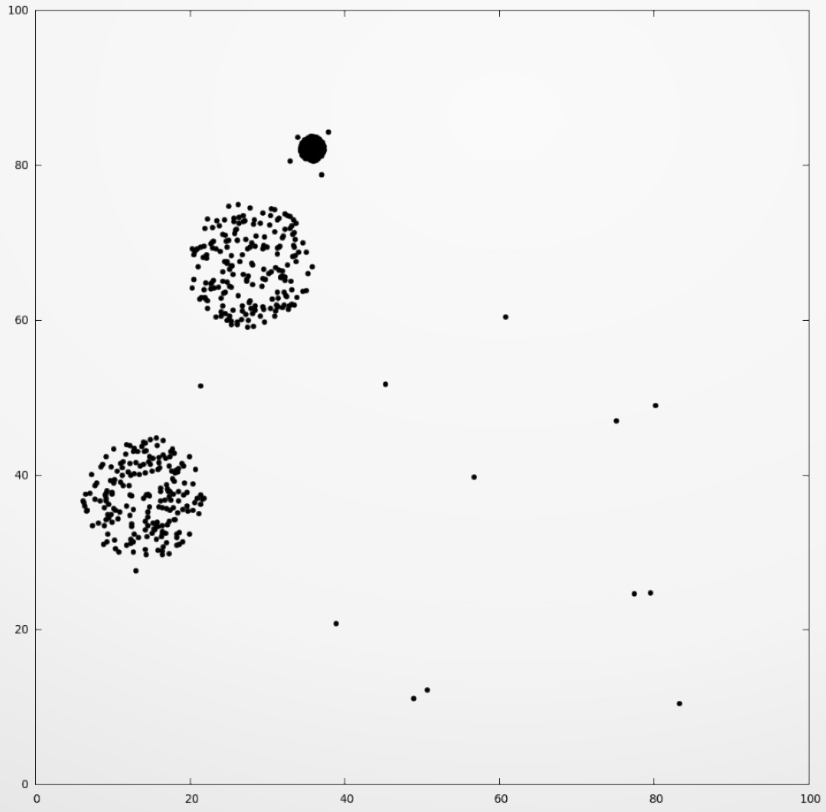
Outras medidas FOSC

- Adaptado para restrições
  - Aprendizado semi-supervisionado
- Utilização Modularidade Q
  - Detecção de comunidade ou agrupamento de grafos

- F. A. Anjos, J. Sander, R. J. G. B. Campello, “A Modularity-Based Measure for Cluster Selection from Clustering Hierarchies”, AusDM 2018.

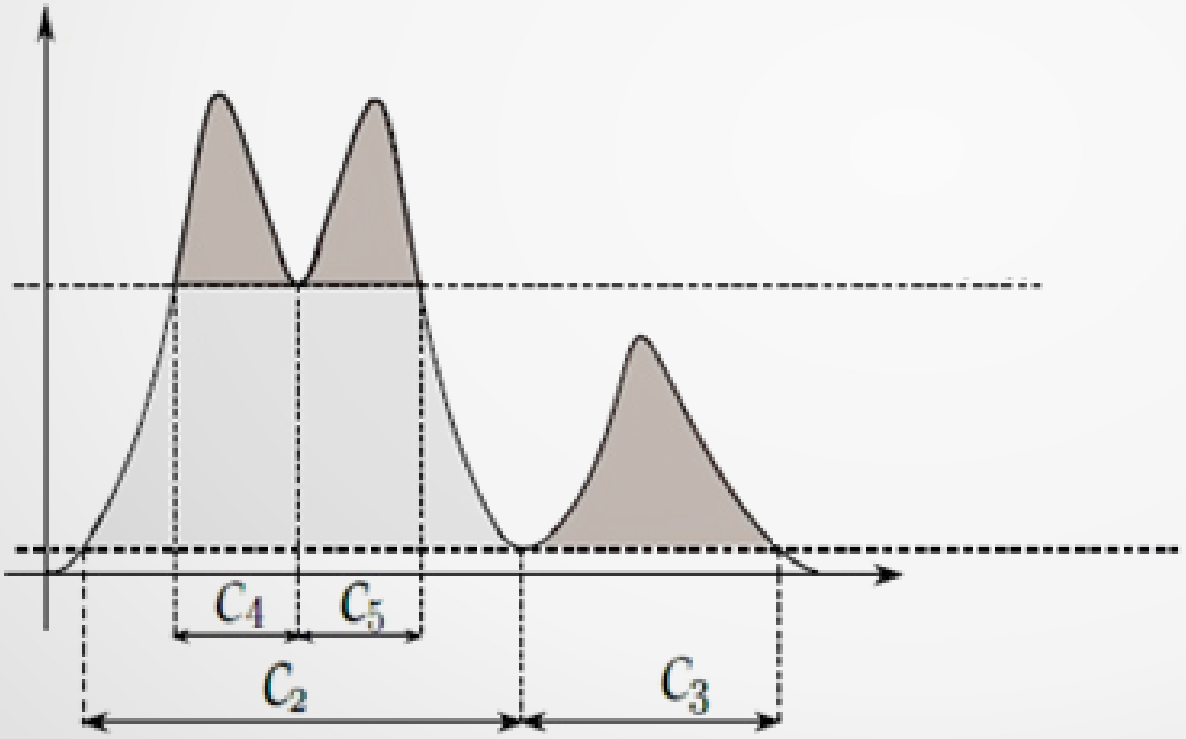
Detecção de *Outliers*

- “Um *outlier* é uma observação que se desvia tanto das outras observações que levanta suspeitas de que foi gerada por um mecanismo diferente.” [Hawkins 1980]



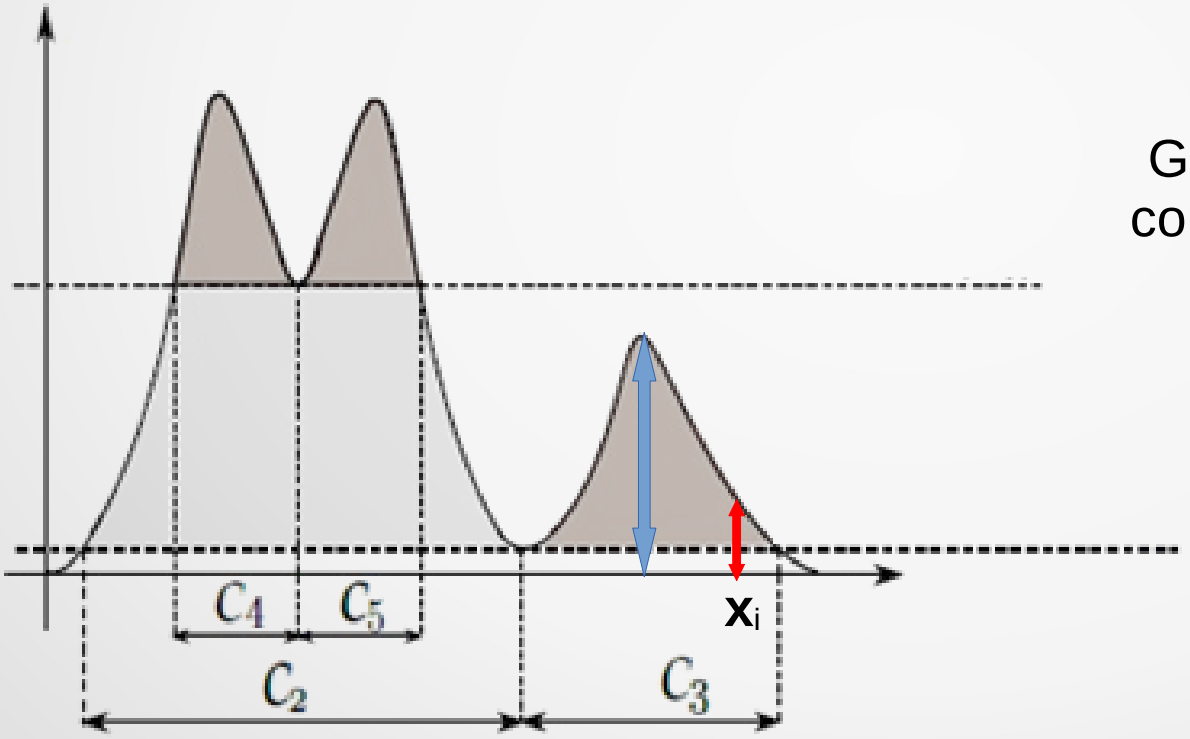
GLOSH

- Global & Local Outlier Scores



GLOSH

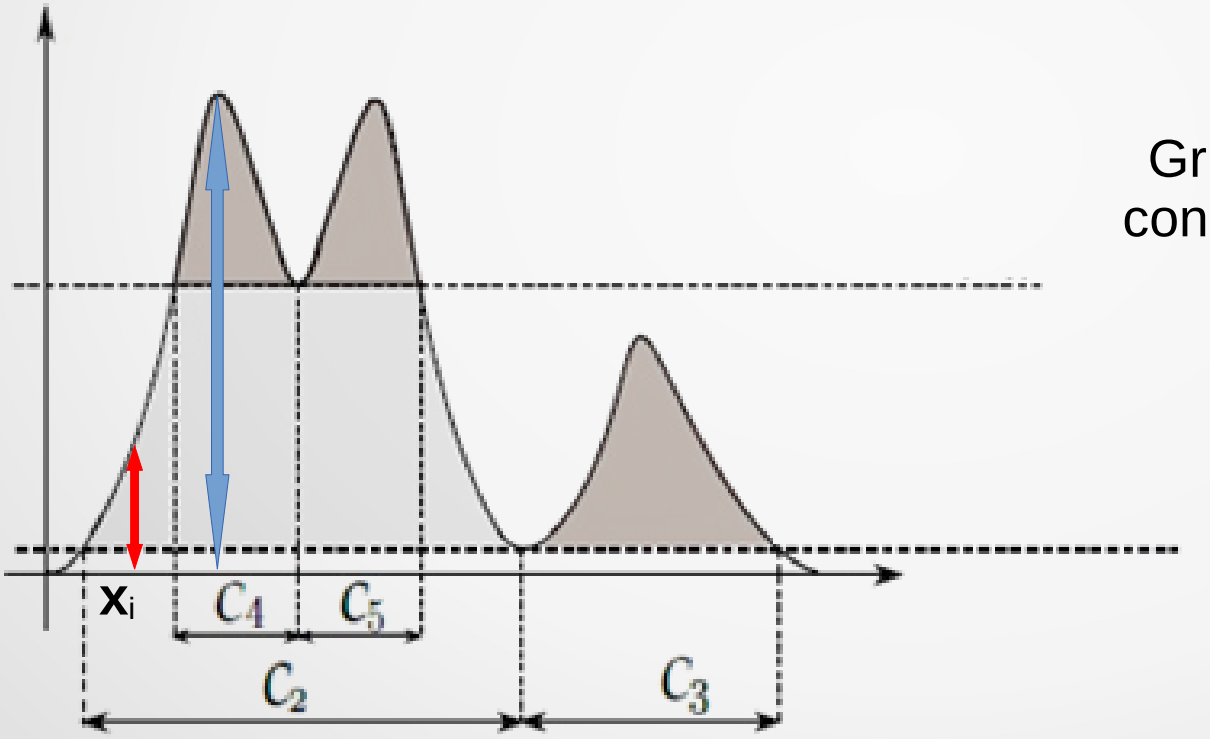
- Global & Local Outlier Scores



Grupo $\mathbf{C}_3$ ao qual $\mathbf{x}_i$ está conectado por densidade

GLOSH

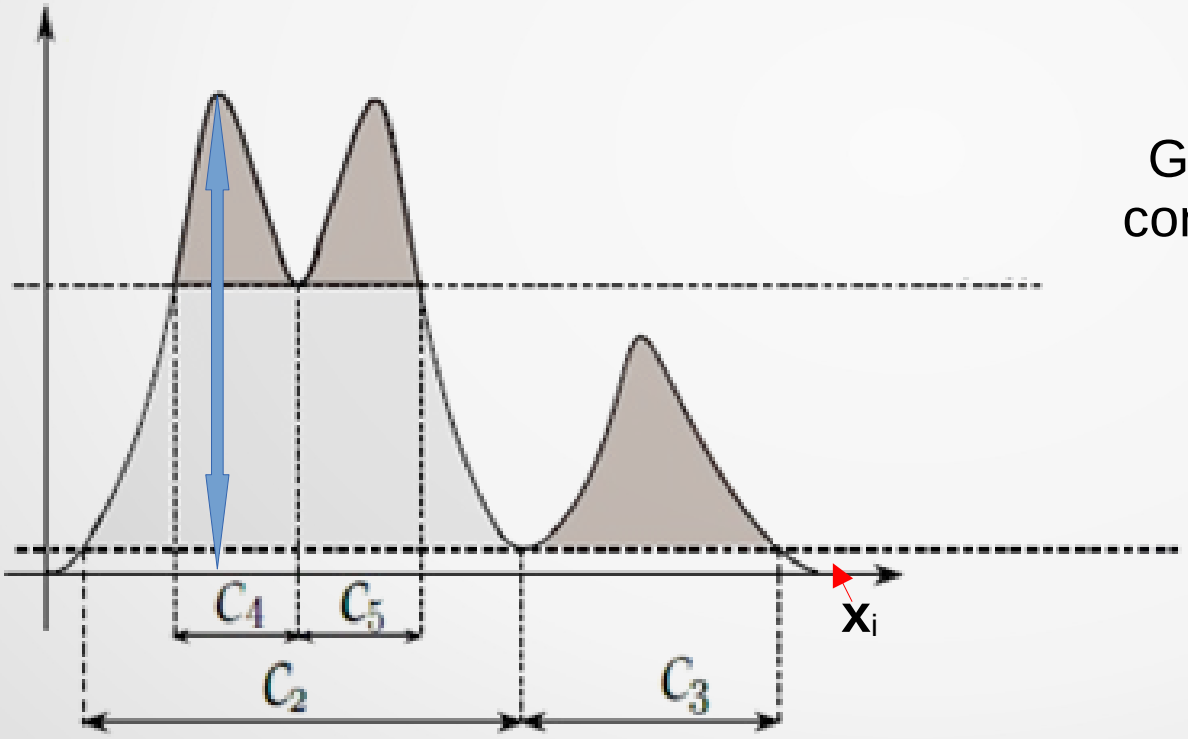
- Global & Local Outlier Scores



Grupo $\mathbf{C}_2$ ao qual $\mathbf{x}_i$ está conectado por densidade

GLOSH

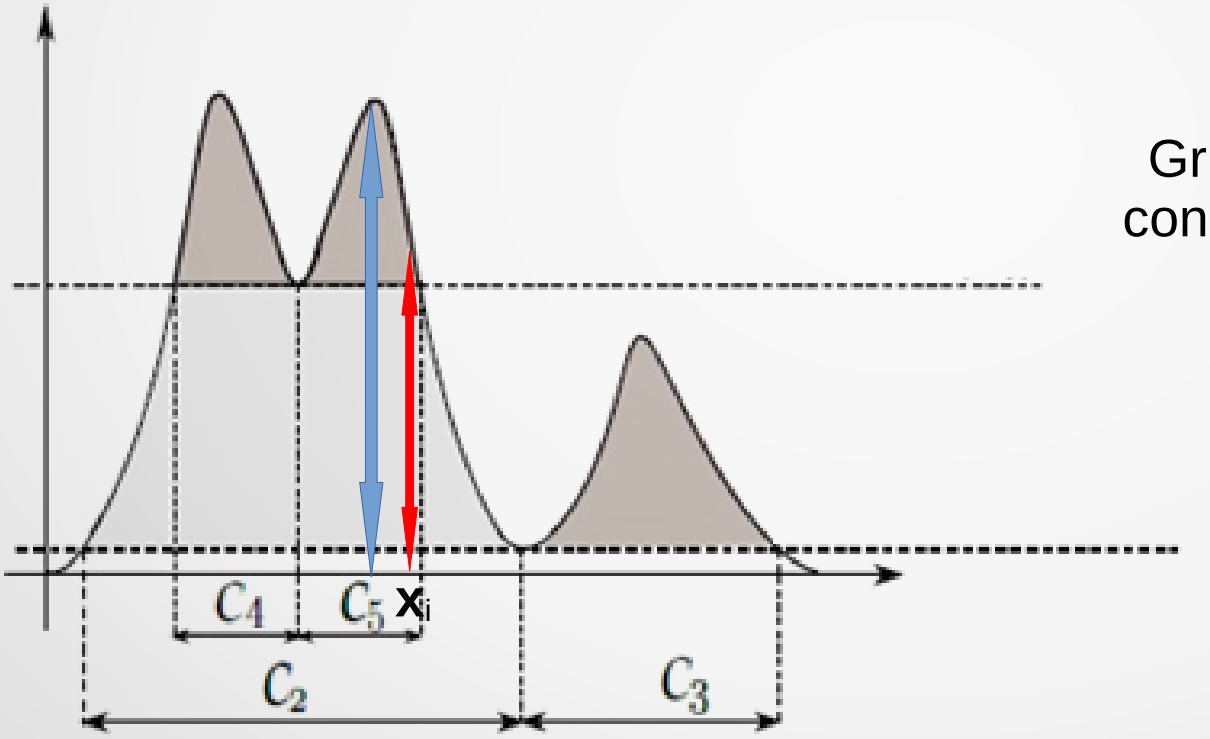
- Global & Local Outlier Scores



Grupo $\mathbf{C}_1$ ao qual $\mathbf{x}_i$ está conectado por densidade

GLOSH

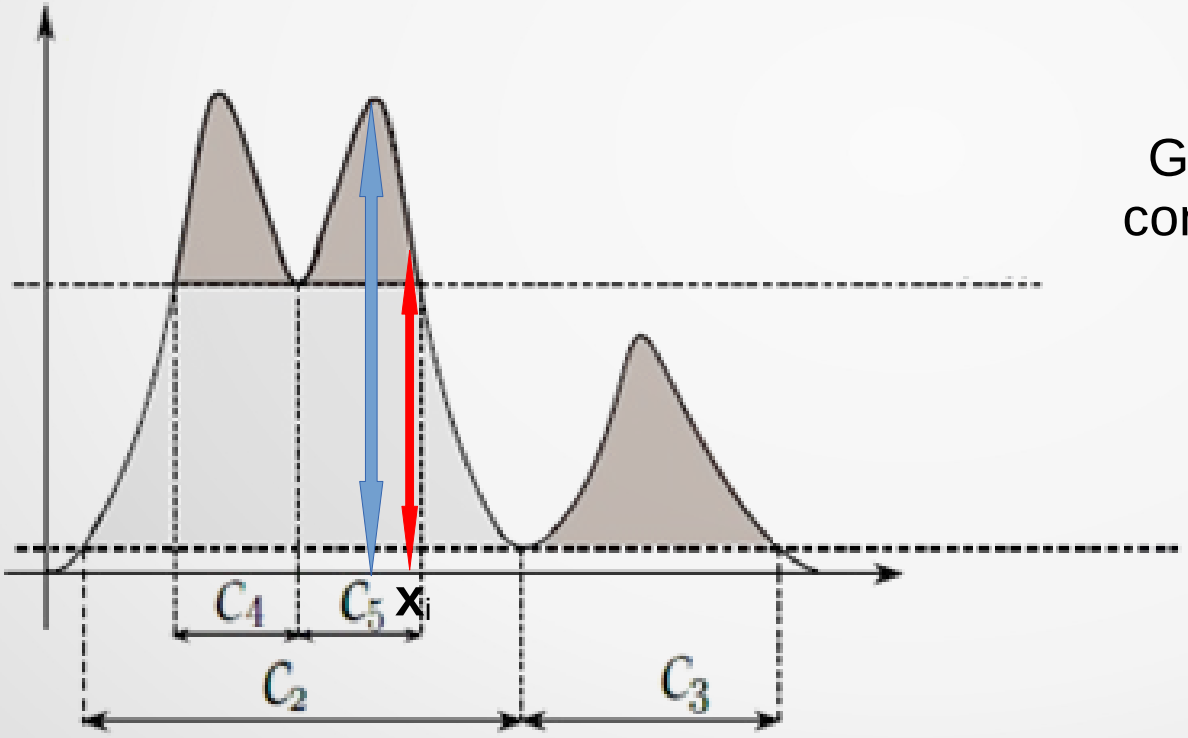
- Global & Local Outlier Scores



Grupo $\mathbf{C_5}$ ao qual $\mathbf{x_i}$ está conectado por densidade

GLOSH

$$\mathrm{GLOSH}(\mathbf{x}_i) = \frac{\lambda_{\max}(\mathbf{x}_i) - \lambda(\mathbf{x}_i)}{\lambda_{\max}(\mathbf{x}_i)} = 1 - \frac{\varepsilon_{\max}(\mathbf{x}_i)}{\varepsilon(\mathbf{x}_i)}$$

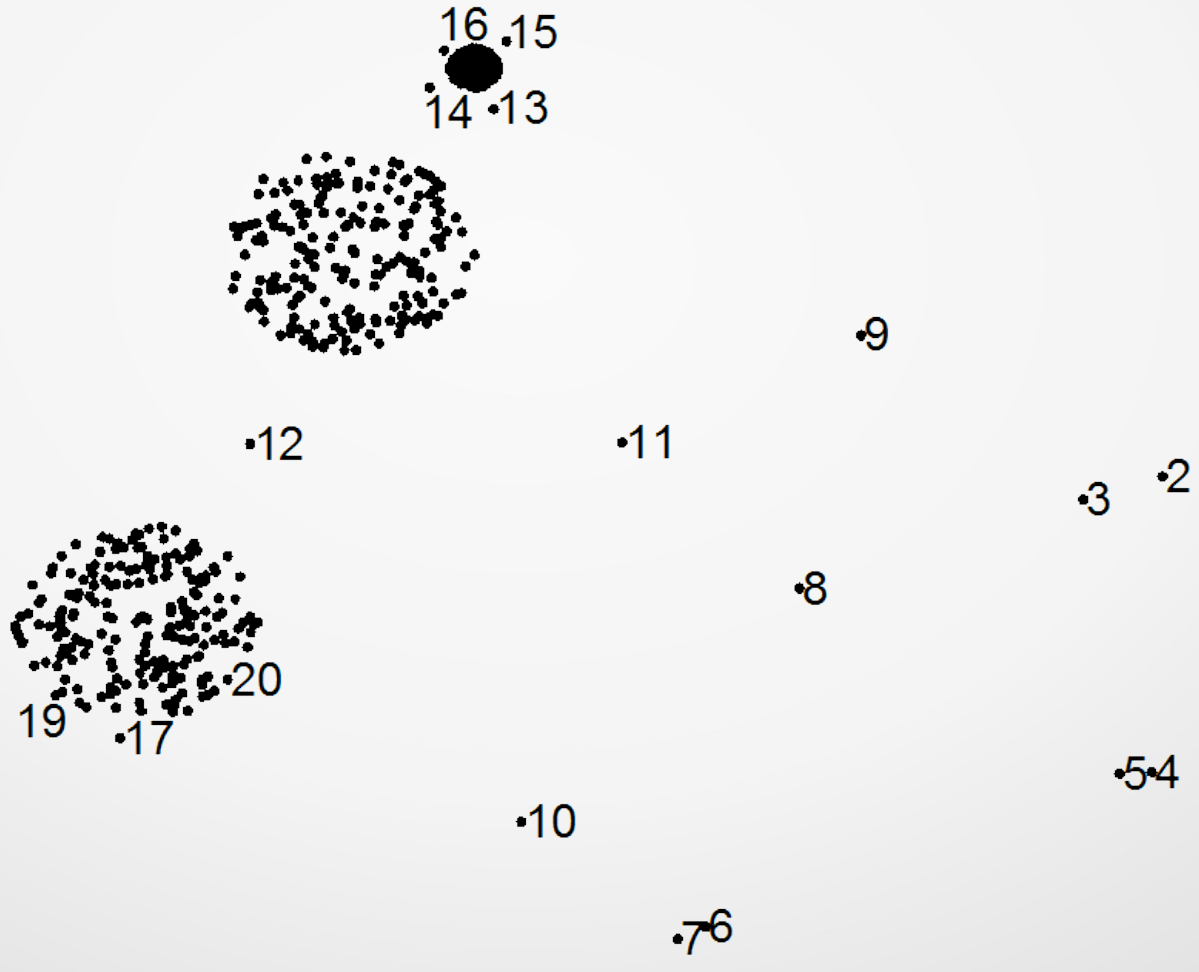


Grupo $\mathbf{C}_5$ ao qual $\mathbf{x}_i$ está conectado por densidade

Detecção de *outliers* com HDBSCAN*

- Exemplo de ranqueamento usando GLOSH

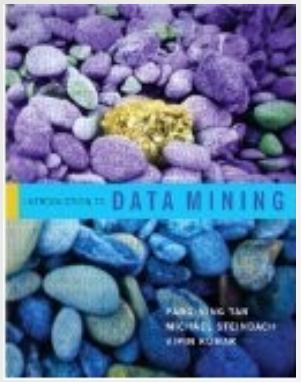
Rank	Score
1	0.996
2	0.995
3	0.994
4	0.994
5	0.994
6	0.993
7	0.993
8	0.992
9	0.991
10	0.989
11	0.985
12	0.974
13	0.891
14	0.877
15	0.854
16	0.803
17	0.680
18	0.573
19	0.568
20	0.562
...	...
615	0.000



Extensões (selecionadas)

- Visualizações – MustaCHE
  - *A. C. A. Neto, J. Sander, R. J. G. B. Campello and M. A. Nascimento, "Efficient Computation and Visualization of Multiple Density-Based Clustering Hierarchies," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3075-3089, 1 Aug. 2021*
- Classificação e agrupamento semi-supervisionada
  - *Castro Gertrudes, J., Zimek, A., Sander, J. et al. A unified view of density-based methods for semi-supervised clustering and classification. Data Min Knowl Disc 33, 1894–1952 (2019).*
- Map-Reduce BigData Version of HDBSCAN*
  - *J. A. d. Santos, T. I. Syed, M. C. Naldi, R. J. G. B. Campello and J. Sander, "Hierarchical Density-Based Clustering Using MapReduce," in IEEE Transactions on Big Data, vol. 7, no. 1, pp. 102-114, 1 March 2021*
- Múltiplas soluções com diferentes m_{pts} – CORE-SG
  - *A. C. A. Neto, M. C. Naldi, R. J. G. B. Campello and J. Sander, "CORE-SG: Efficient Computation of Multiple MSTs for Density-Based Methods," 2022 IEEE 38th International Conference on Data Engineering (ICDE), 2022, pp. 951-964*

Bibliografia

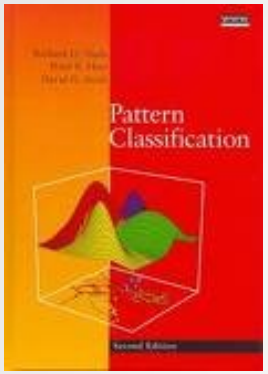


V. TAN, STEINBACH, M., KUMAR, P. Introdução ao Data Mining (Mineração de Dados). Edição 1. Ciência Moderna 2009. ISBN 9788573937619.

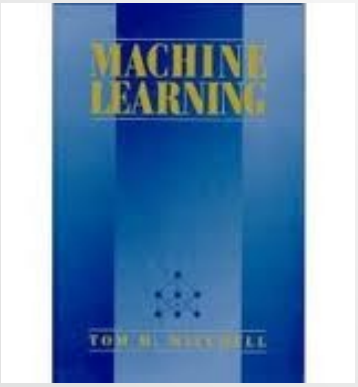
Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Katti Faceli, Ana Carolina Lorena, João Gama, André C. P. L. F. de Carvalho. Grupo Gen 2011

Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993

Referencias



Duda, R.O., Hart, P. E. and Stork, D. G. Pattern Classification (2nd Edition). Wiley-Interscience



MITCHELL, T. Machine Learning, McGraw Hill, 1997.

Referências

- Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988
- Kaufman, L., Rousseeuw, P. J., Finding Groups in Data – An Introduction to Cluster Analysis, Wiley, 2005.
- Tan, P.-N., Steinbach, M., and Kumar, V., Introduction to Data Mining, Addison-Wesley, 2006
- Wu, X. and Kumar, V., The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC, 2009
- D. Steinley, K-Means Clustering: A Half-Century Synthesis, British J. of Mathematical and Stat. Psychology, V. 59, 2006