Curso 2 – CD, AM e DM

Mineração de Dados

Parte 9
Extração de Padrões
Agrupamento baseado em Densidade
Outliers e Anomalias

Prof. Ricardo M. Marcacini
ricardo.marcacini@icmc.usp.br

Agrupamento baseado em Densidade

- Os métodos que estudamos são criticados pelas seguintes limitações:

Agrupamento baseado em Densidade

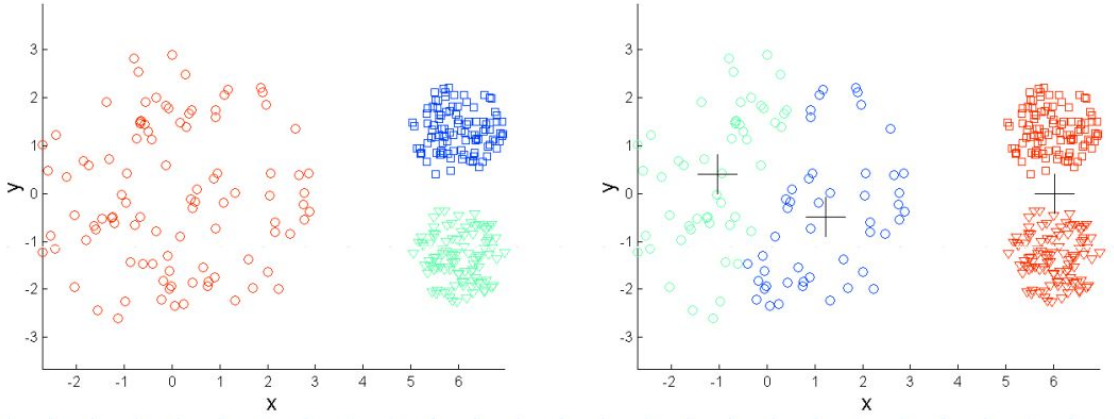
- Os métodos que estudamos são criticados pelas seguintes limitações:
  - **Identificar *clusters* de formatos arbitrários**

Agrupamento baseado em Densidade

- Os métodos que estudamos são criticados pelas seguintes limitações:
  - **Identificar *clusters* de formatos arbitrários**
    - Exemplo: observamos que *k-means tende* a identificar *clusters* de formatos globulares (estudamos na Parte 4)

Agrupamento baseado em Densidade

- Os métodos que estudamos são criticados pelas seguintes limitações:
  - **Identificar *clusters* de formatos arbitrários**
    - Exemplo: observamos que *k-means tende* a identificar *clusters* de formatos globulares (estudamos na Parte 4)



Clusters esperados Clusters obtidos

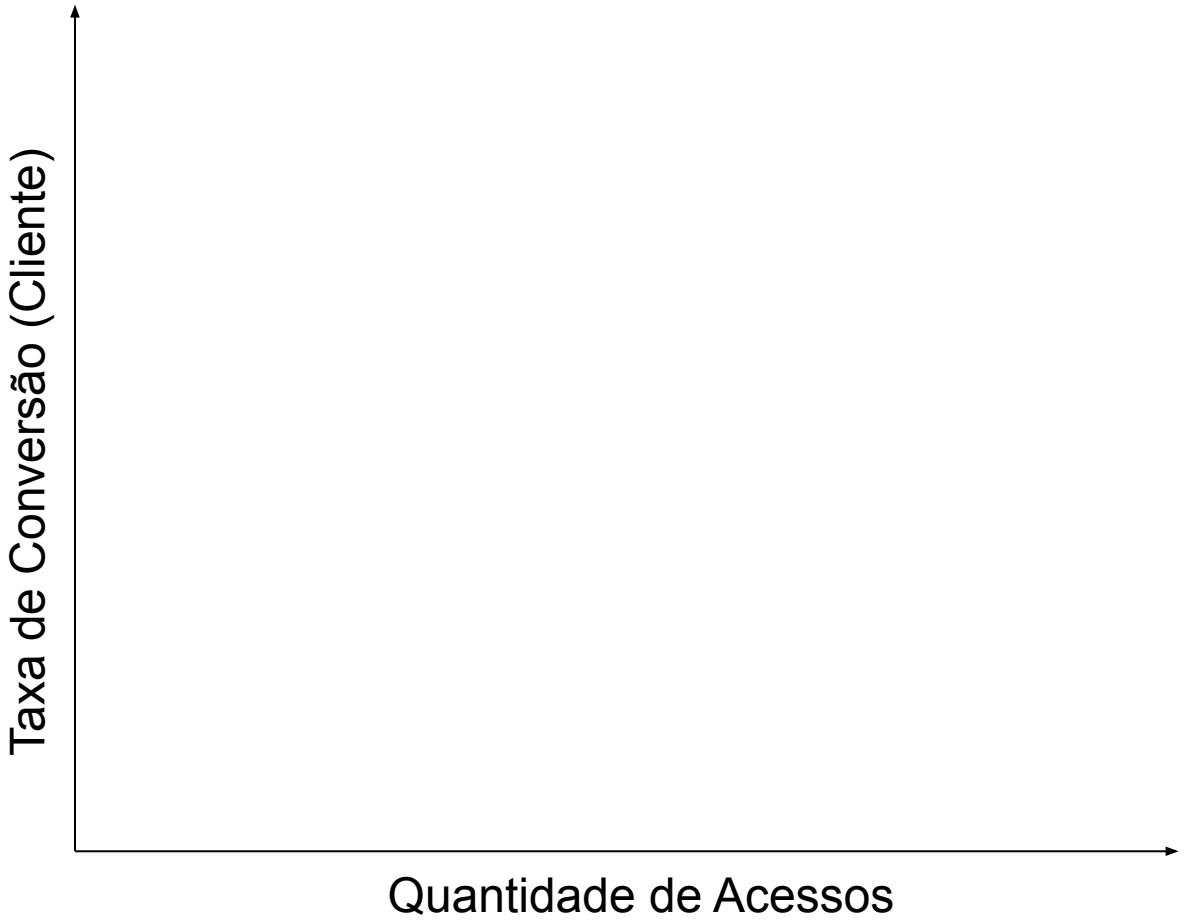
Agrupamento baseado em Densidade

- Os métodos que estudamos são criticados pelas seguintes limitações:
  - **Identificar *clusters* de formatos arbitrários**
    - Exemplo: observamos que *k-means tende* a identificar *clusters* de formatos globulares (estudamos na Parte 4)
  - **Sensíveis à presença de *outliers* e anomalias**

Exemplo:

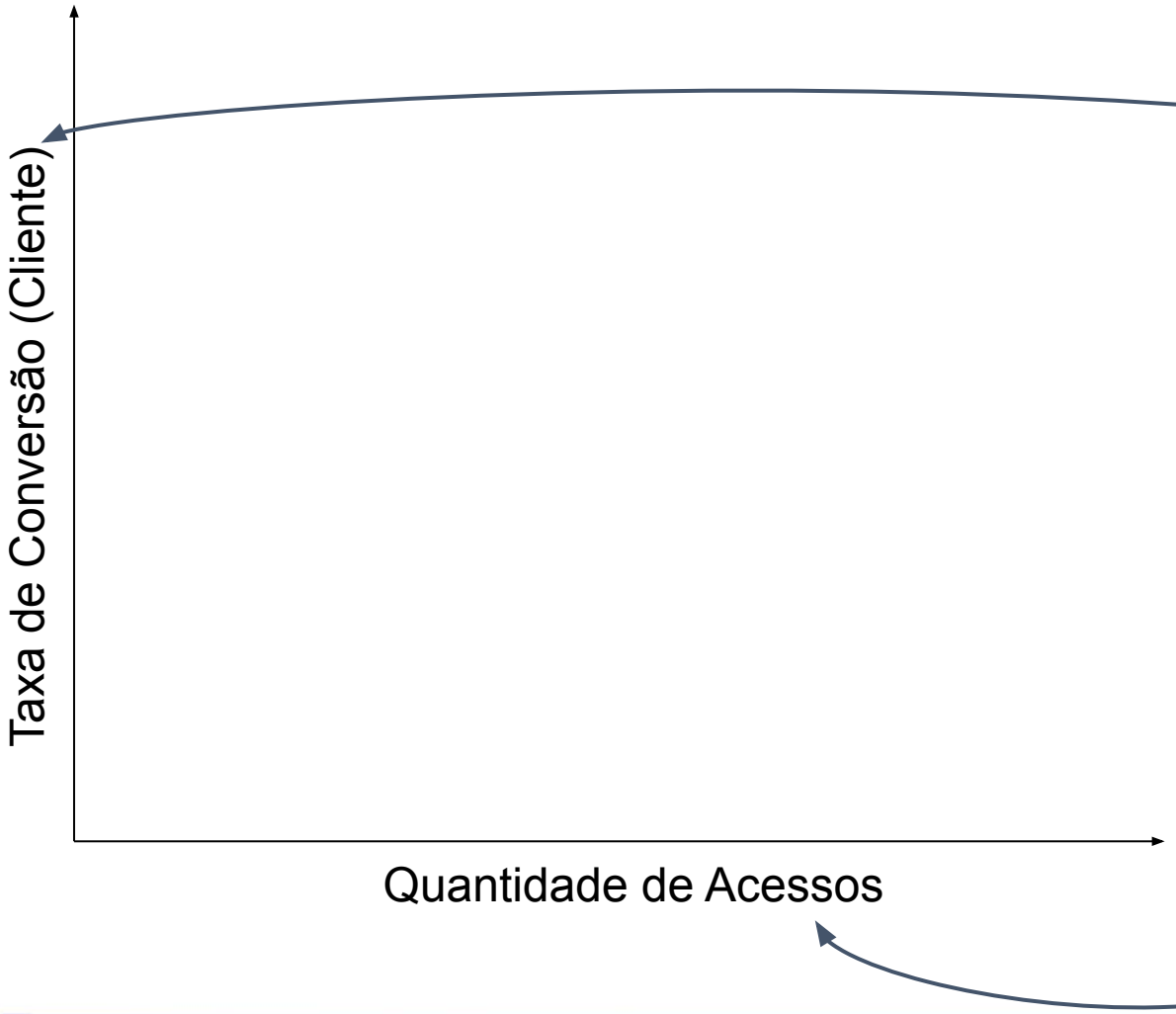
Cliente	Quantidade de Acessos	Taxa de Conversão
1	3	0.2
2	5	0.7
3	11	0.9
…	…	…
n	3	0.4

Exemplo:



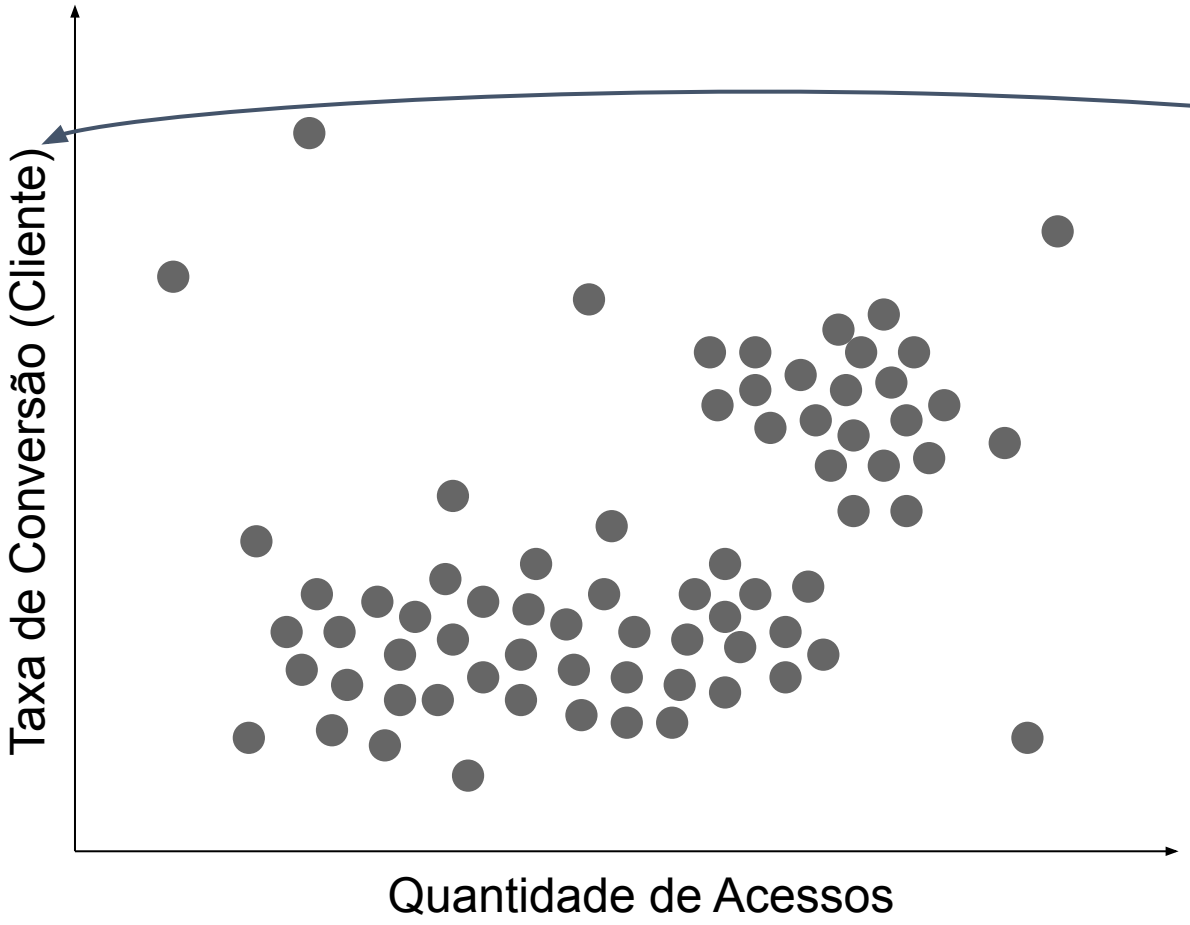
Cliente	Quantidade de Acessos	Taxa de Conversão
1	3	0.2
2	5	0.7
3	11	0.9
…	…	…
n	3	0.4

Exemplo:



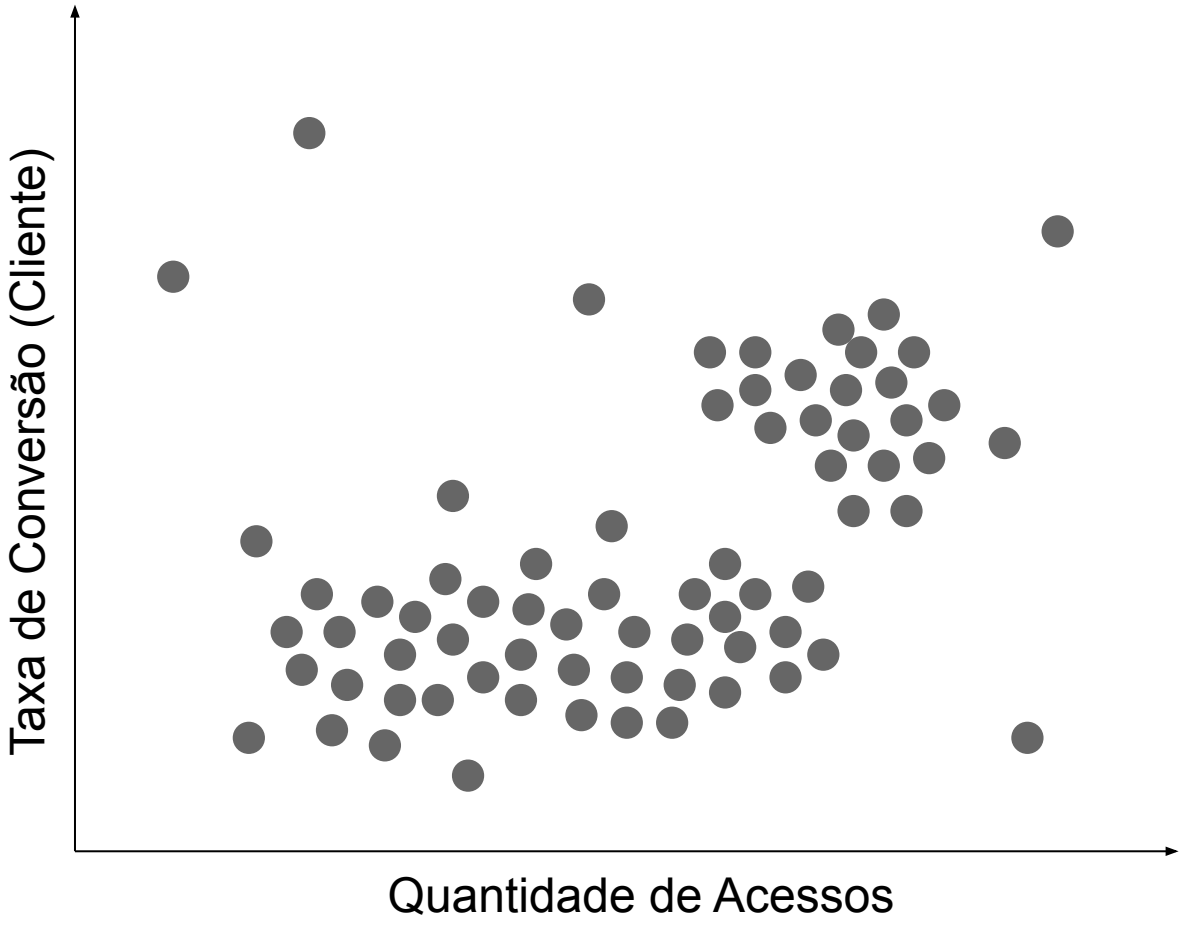
Cliente	Quantidade de Acessos	Taxa de Conversão
1	3	0.2
2	5	0.7
3	11	0.9
…	…	…
n	3	0.4

Exemplo:

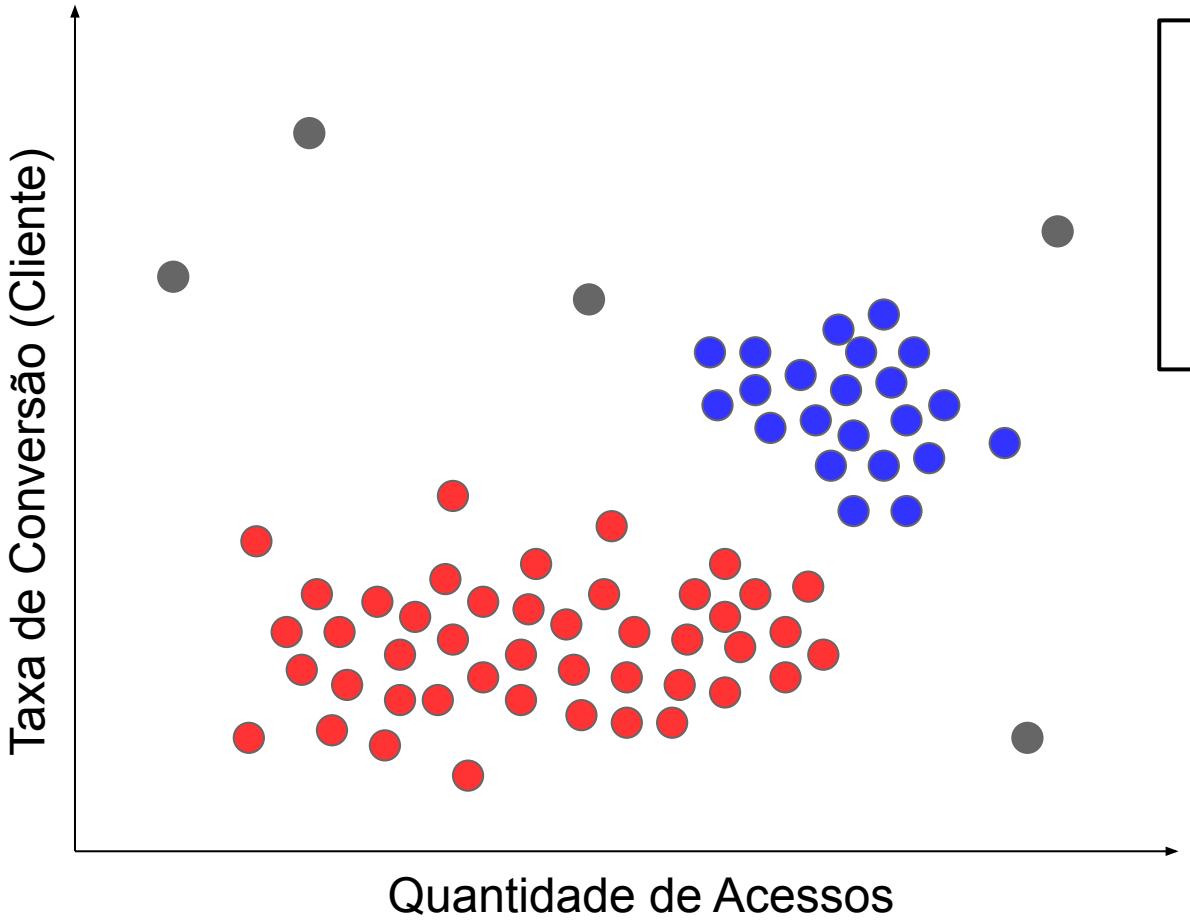


Cliente	Quantidade de Acessos	Taxa de Conversão
1	3	0.2
2	5	0.7
3	11	0.9
…	…	…
n	3	0.4

Exemplo:

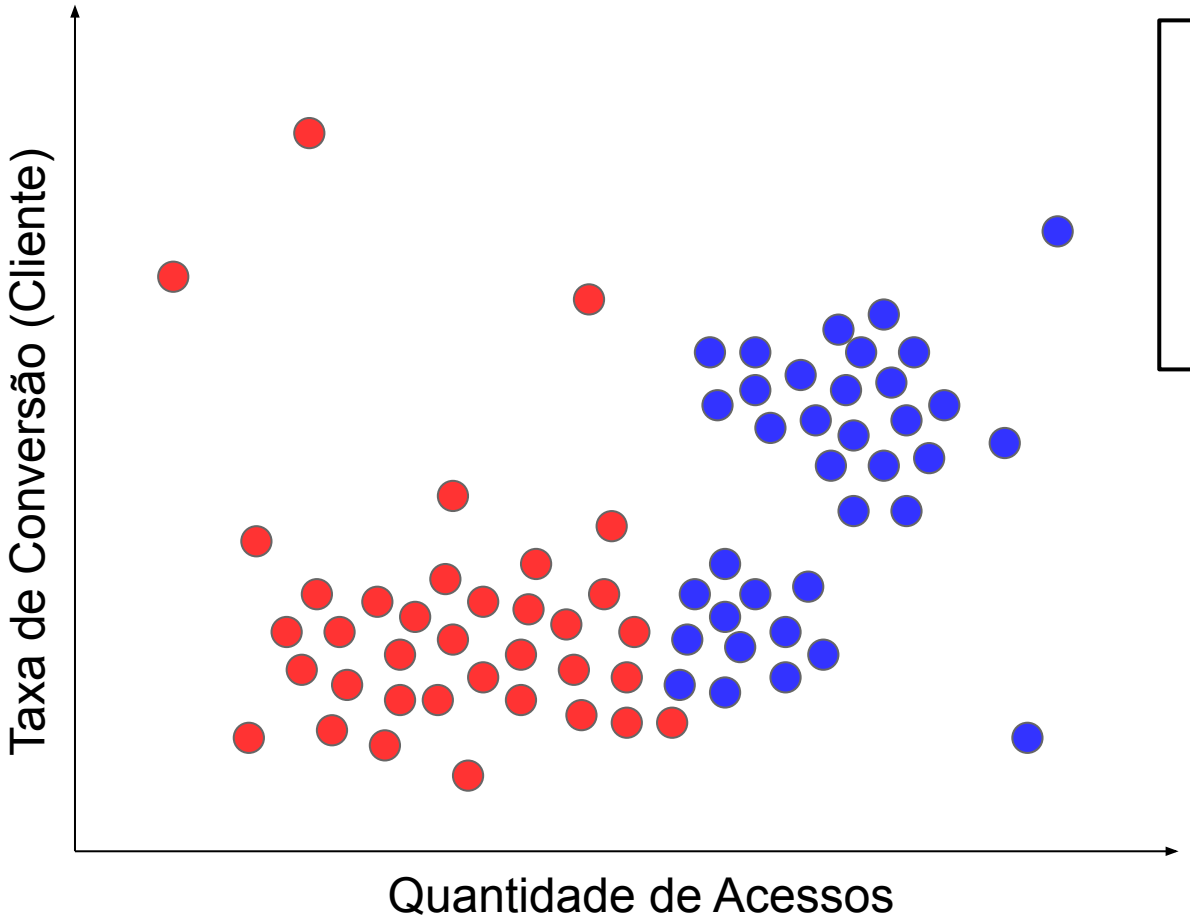


Exemplo:



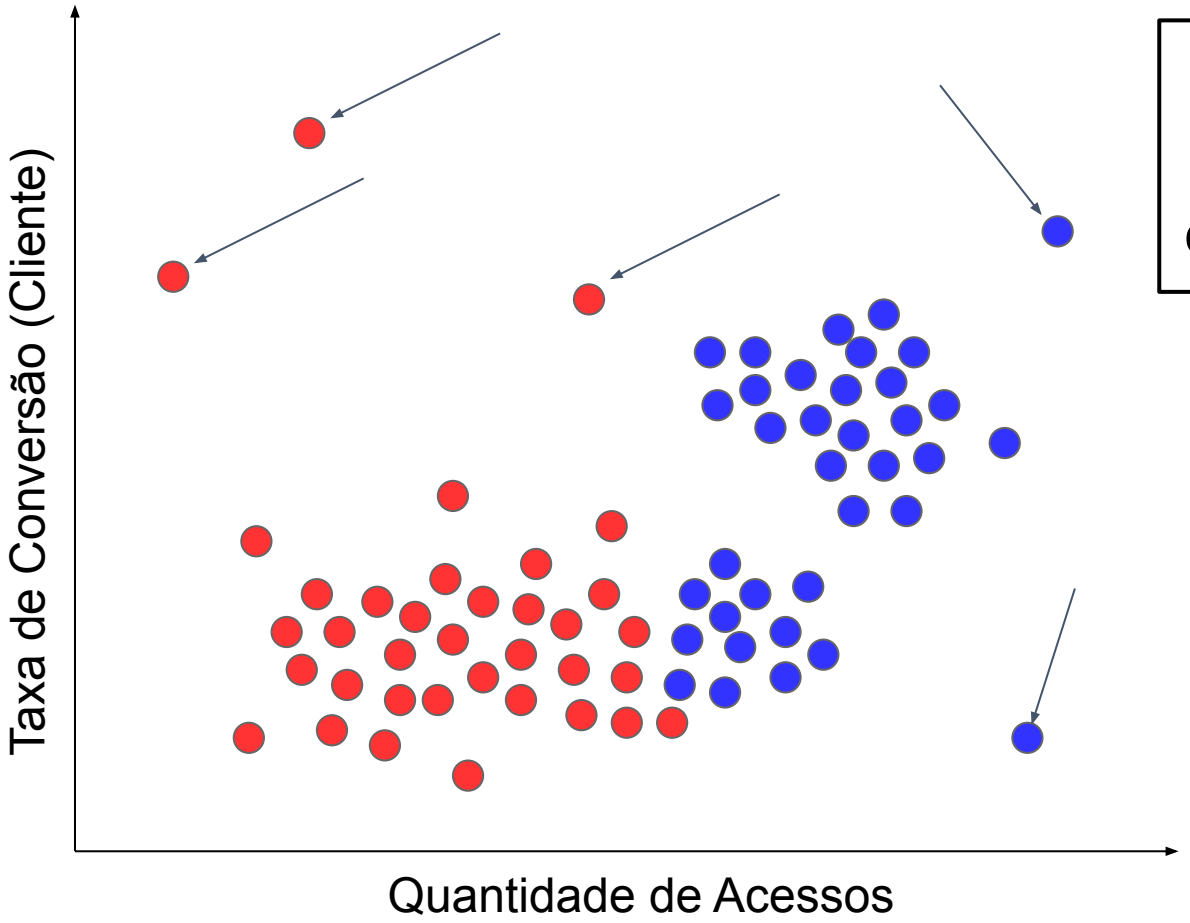
Humanos identificam naturalmente estruturas de grupos conforme regiões de alta densidade

Exemplo:



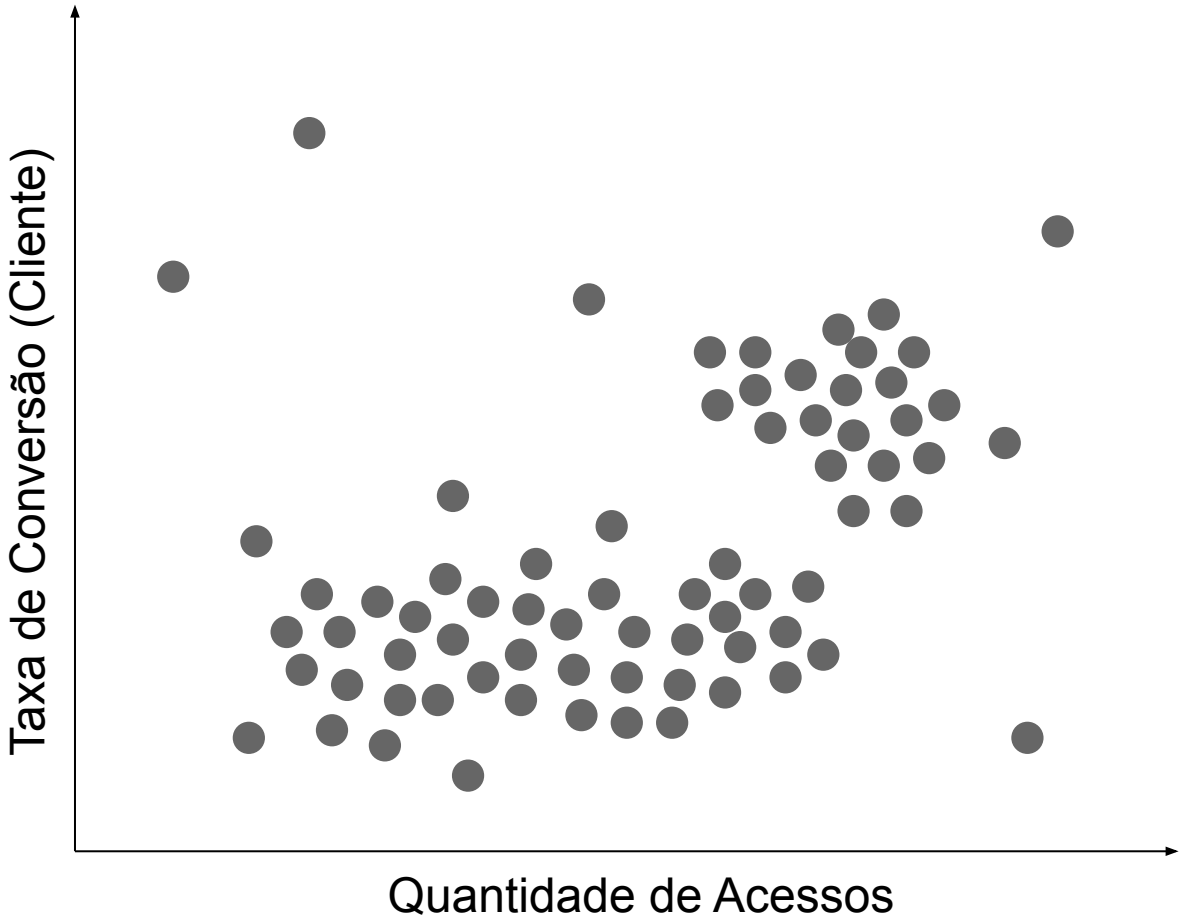
Vários métodos, como o *k-means*, não identificam corretamente os grupos de formatos variados.

Exemplo:



Ainda, *outliers* são considerados e afetam a estrutura de agrupamento

Agrupamento baseado em Densidade



Agrupamento baseado em Densidade

- DBSCAN:
  - Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, 1996 (Vol. 96, No. 34, pp. 226-231).

Agrupamento baseado em Densidade

- DBSCAN:
  - Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, 1996 (Vol. 96, No. 34, pp. 226-231).

2014 SIGKDD TEST OF TIME AWARD

Aug 18 2014 /

2014 SIGKDD Test of Time Award:

The SIGKDD Test of Time award recognizes outstanding papers from past KDD Conferences beyond the last decade that have had an important impact on the data mining research community.

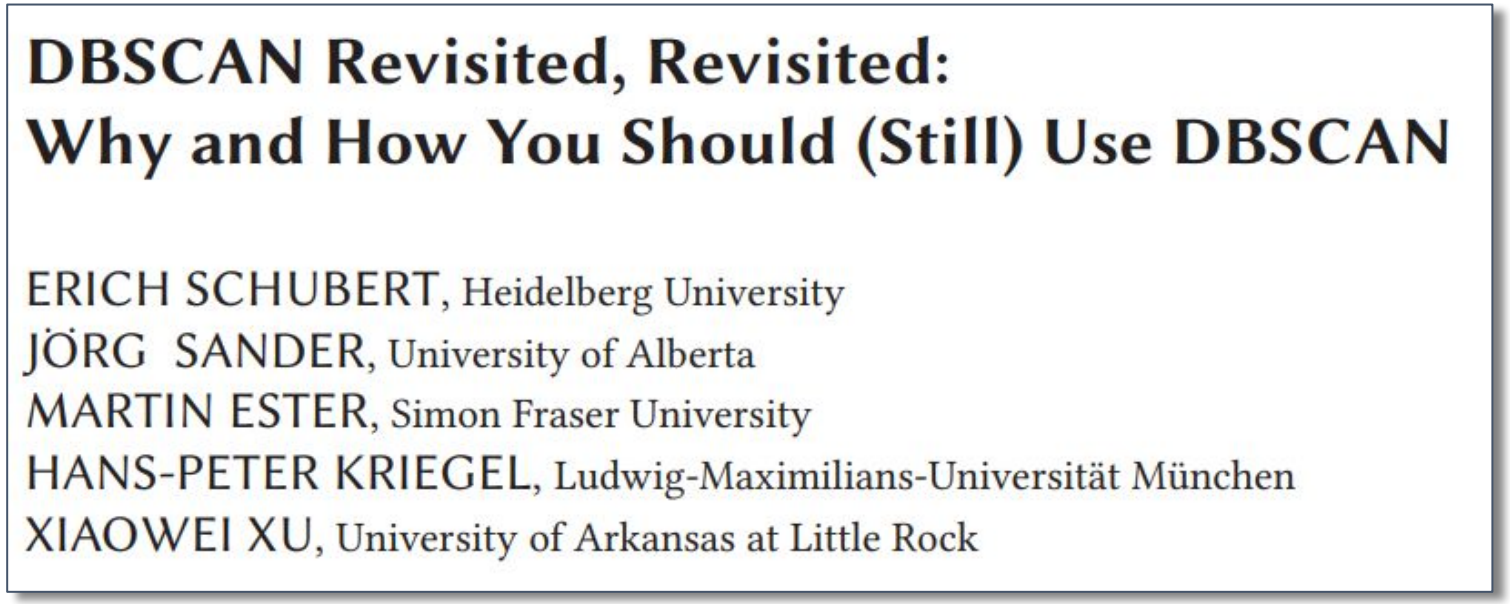
The 2014 Test of Time award recognizes the following influential contributions to SIGKDD that have withstood the test of time:

A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [KDD 1996]

https://www.kdd.org/News/view/2014-sigkdd-test-of-time-award

Agrupamento baseado em Densidade

- DBSCAN:
  - Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, 1996 (Vol. 96, No. 34, pp. 226-231).

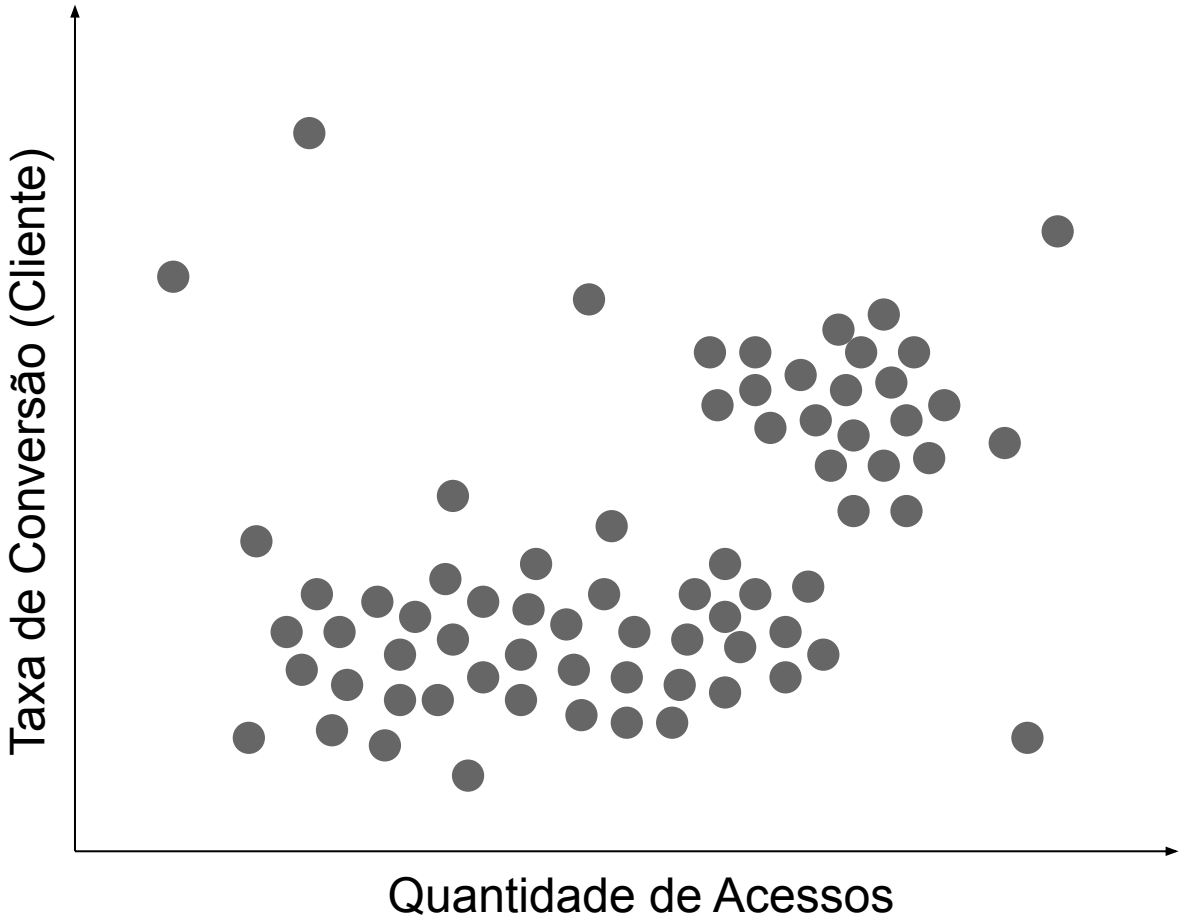


DBSCAN Revisited, Revisited:
Why and How You Should (Still) Use DBSCAN

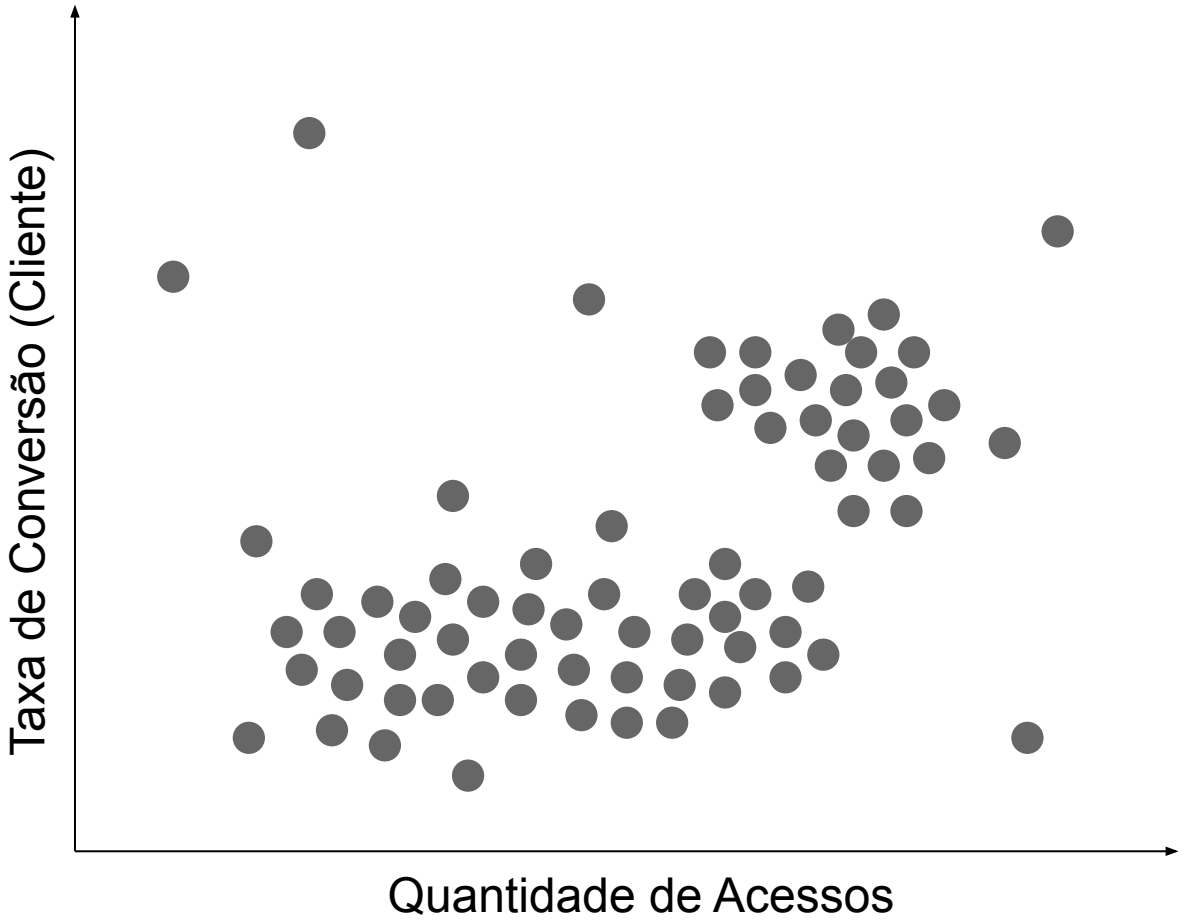
ERICH SCHUBERT, Heidelberg University
JÖRG SANDER, University of Alberta
MARTIN ESTER, Simon Fraser University
HANS-PETER KRIEGEL, Ludwig-Maximilians-Universität München
XIAOWEI XU, University of Arkansas at Little Rock

https://doi.org/10.1145/3068335

Agrupamento baseado em Densidade

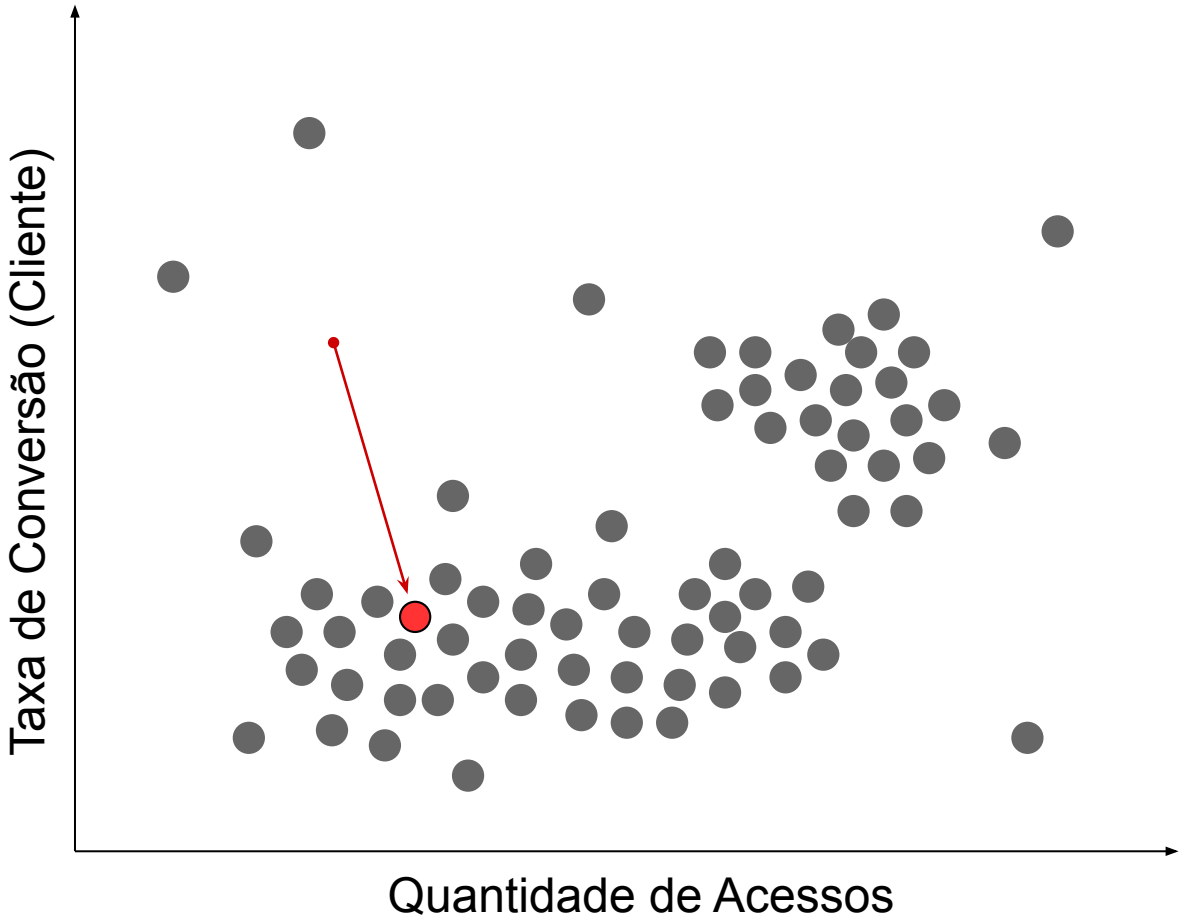


Agrupamento baseado em Densidade



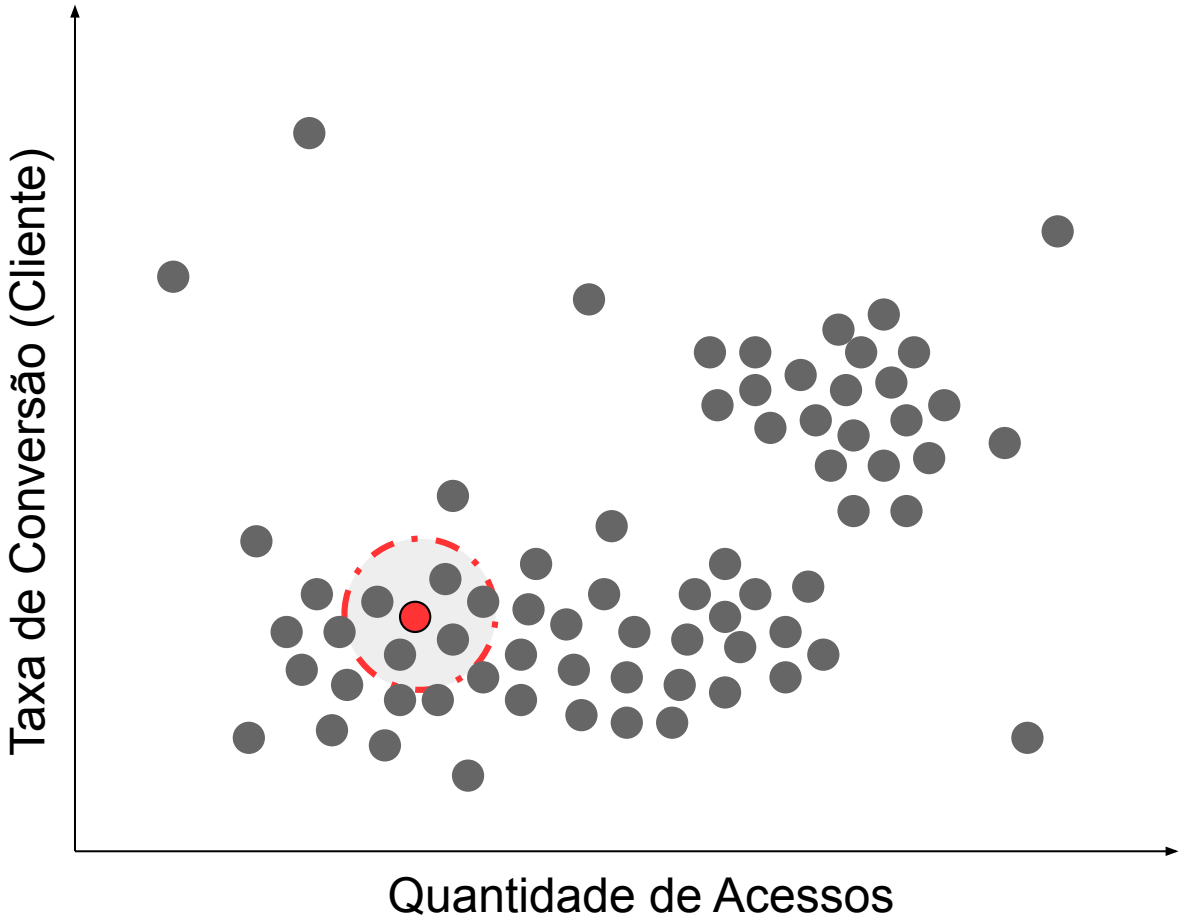
- **Pontos de Núcleo**
- **Pontos de Fronteira**
- **Pontos *Outliers***

Agrupamento baseado em Densidade



Pontos de Núcleo

Agrupamento baseado em Densidade

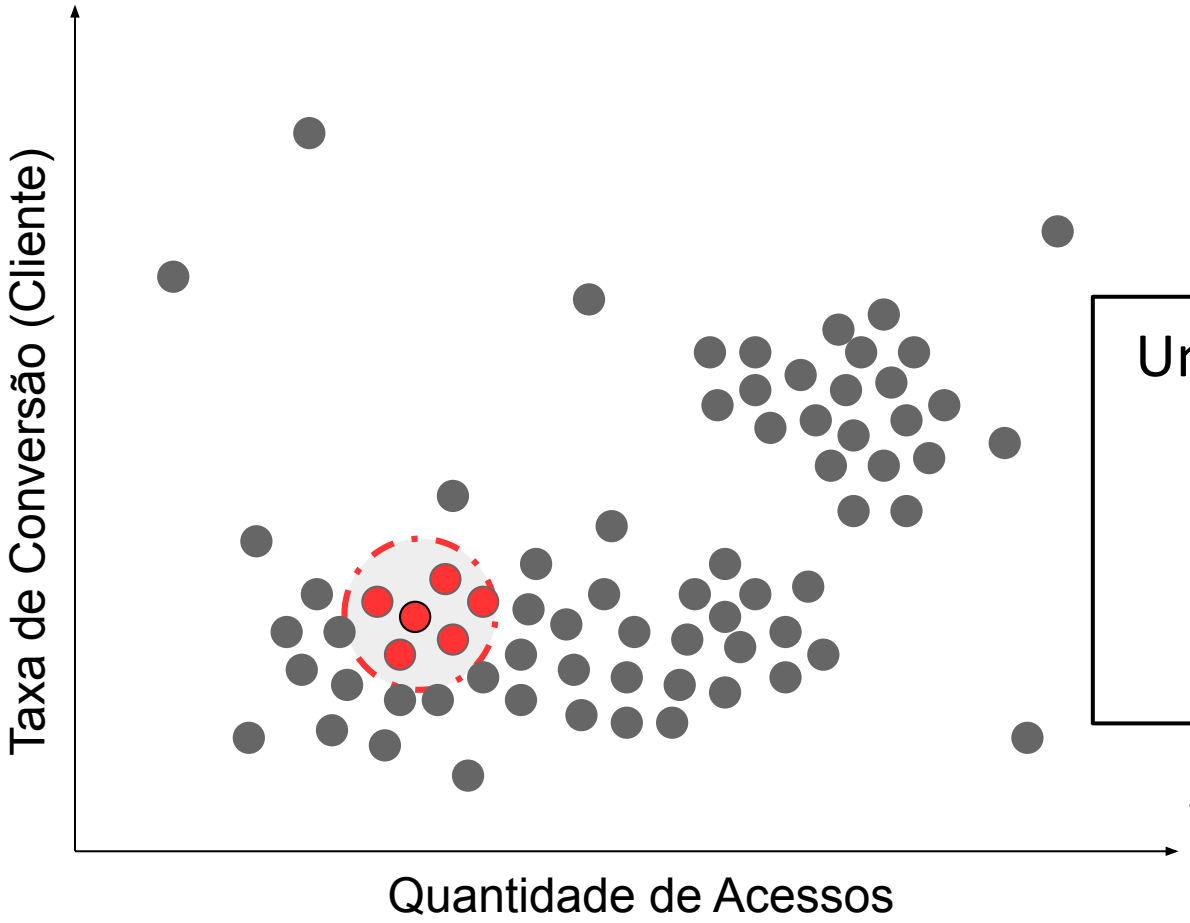


Pontos de Núcleo

Determinar a vizinhança ε de um ponto.

Atenção: o parâmetro ε é definido pelo usuário

Agrupamento baseado em Densidade

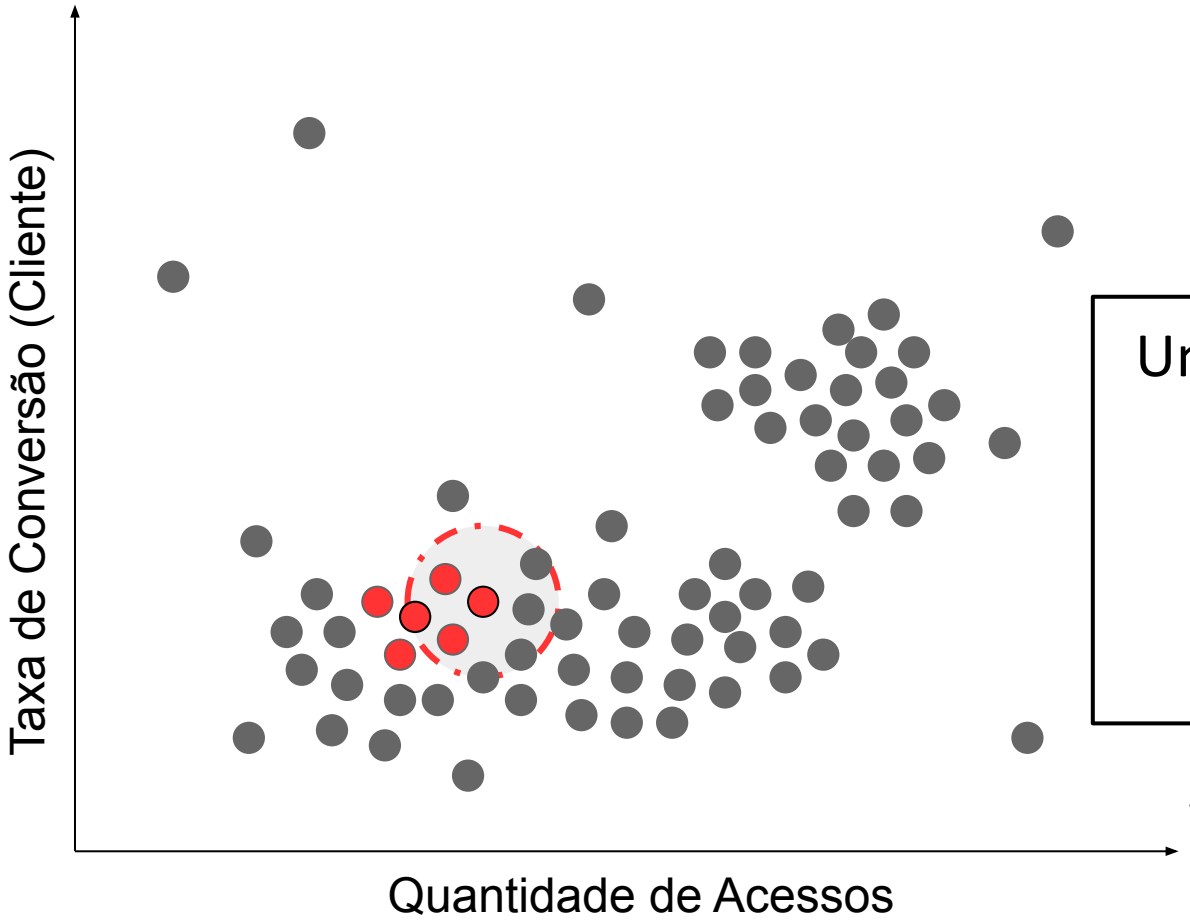


Pontos de Núcleo

Um ponto de núcleo contém um número mínimo de pontos (*minPTS*) em sua vizinhança (incluindo o próprio ponto).

Exemplo: minPTS=4

Agrupamento baseado em Densidade

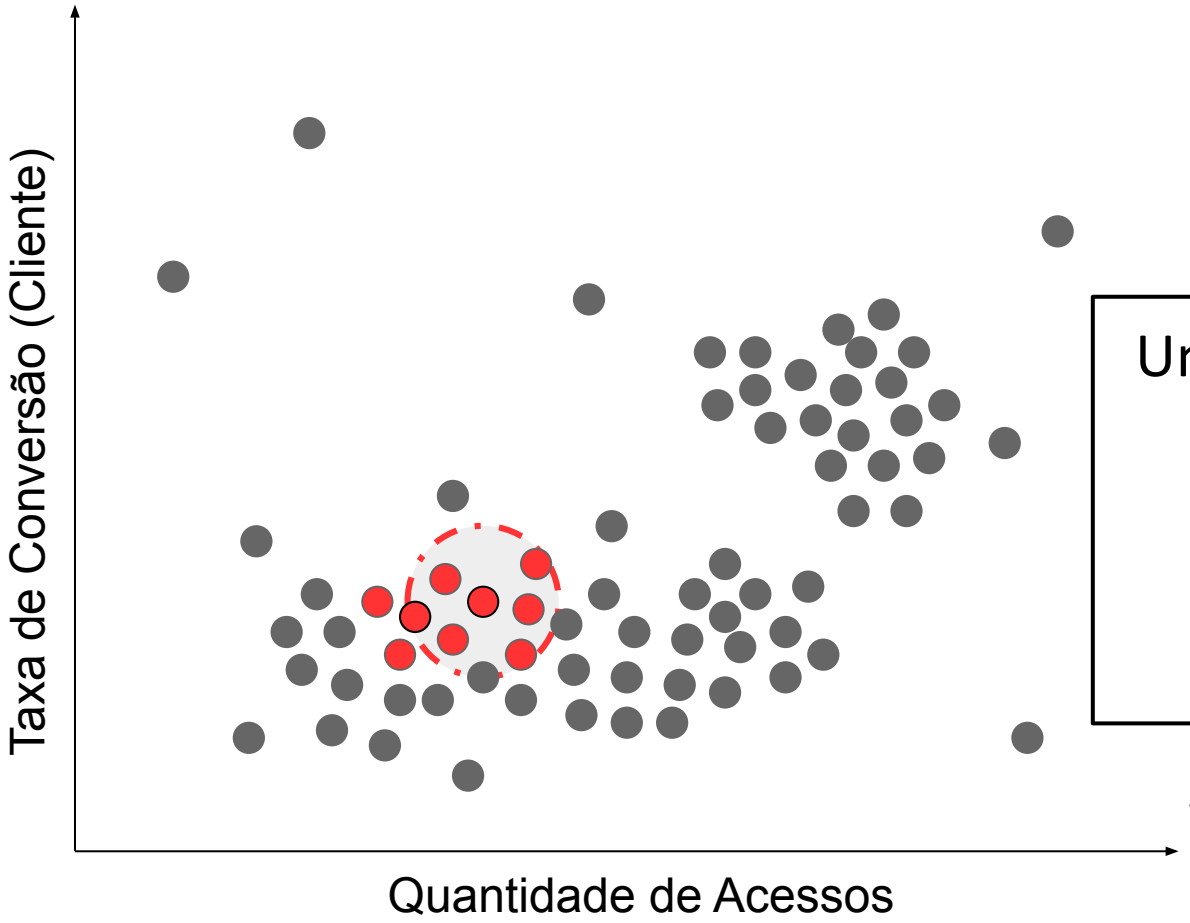


Pontos de Núcleo

Um ponto de núcleo contém um número mínimo de pontos (*minPTS*) em sua vizinhança (incluindo o próprio ponto).

Exemplo: minPTS=4

Agrupamento baseado em Densidade

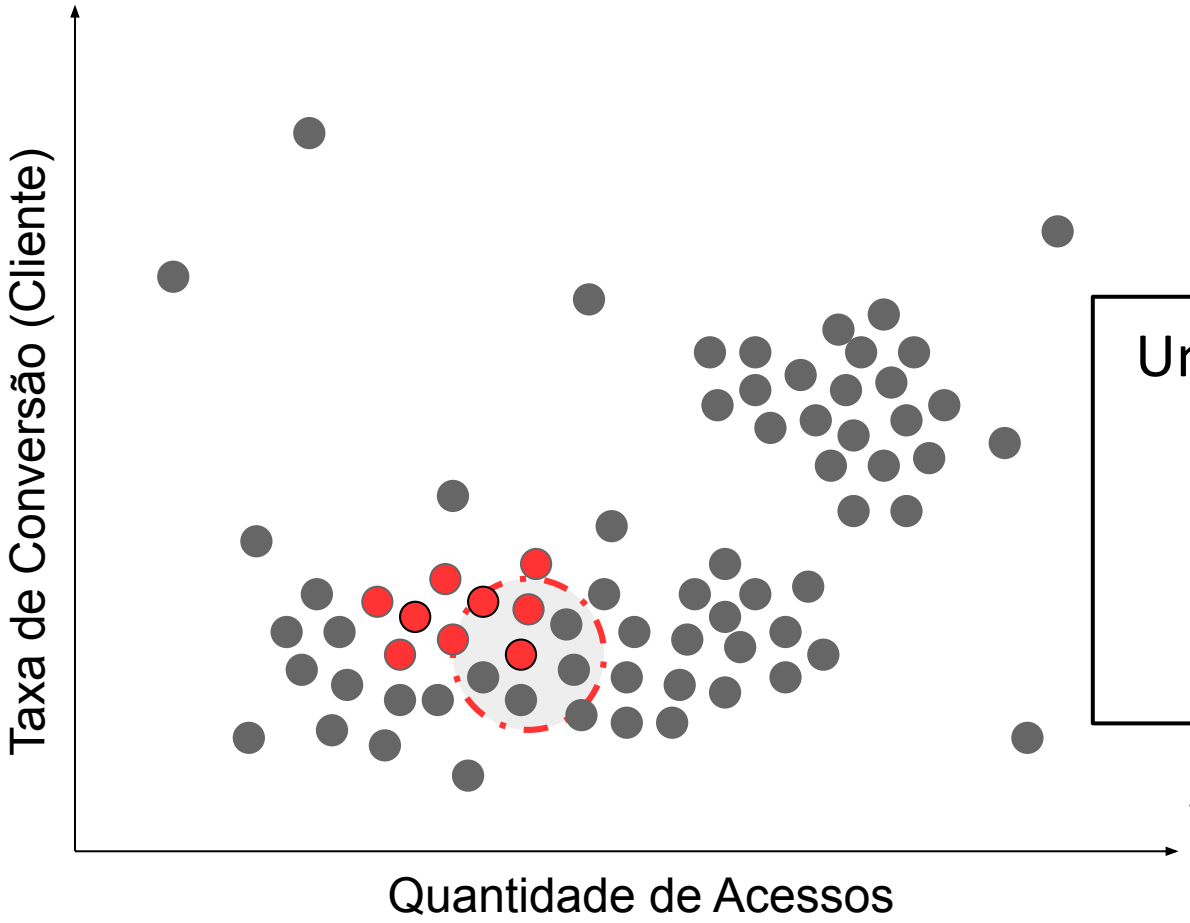


Pontos de Núcleo

Um ponto de núcleo contém um número mínimo de pontos (*minPTS*) em sua vizinhança (incluindo o próprio ponto).

Exemplo: minPTS=4

Agrupamento baseado em Densidade

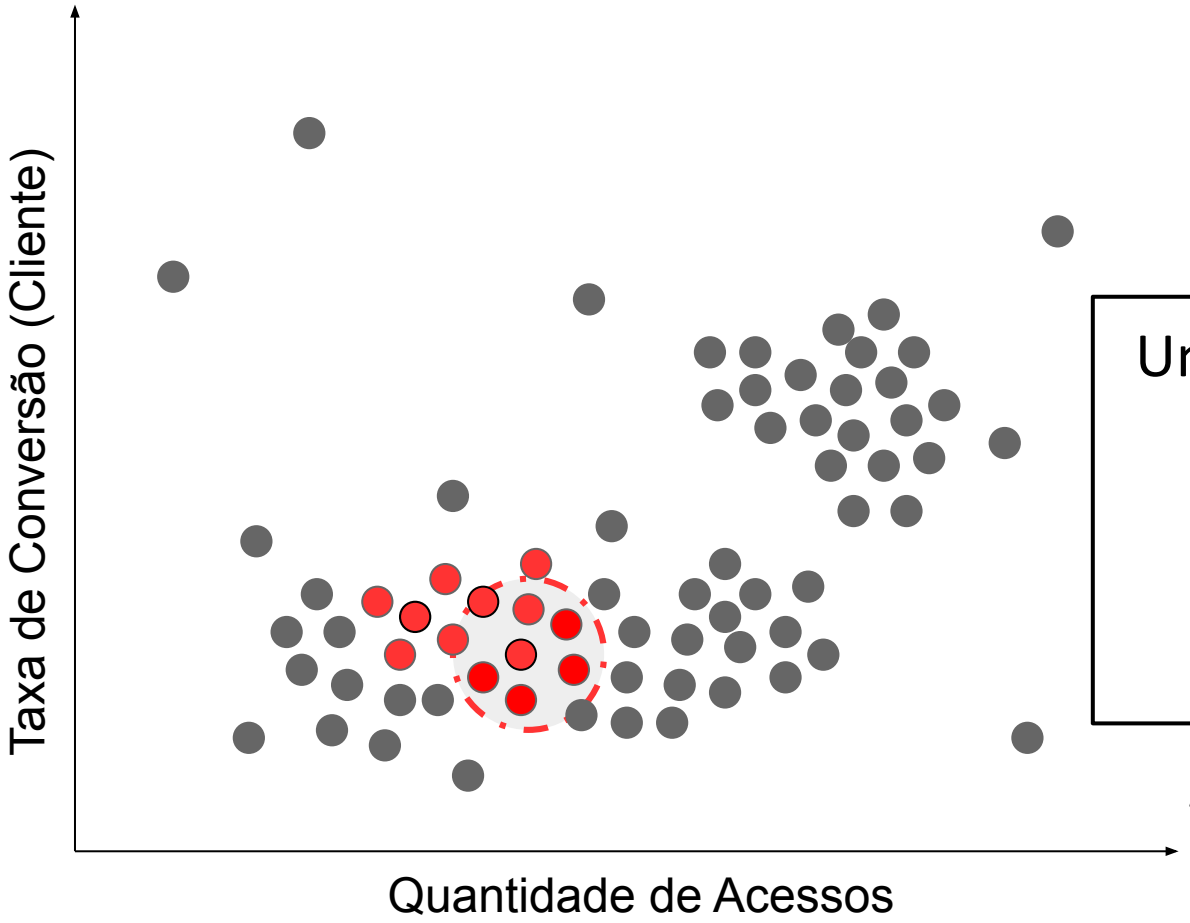


Pontos de Núcleo

Um ponto de núcleo contém um número mínimo de pontos (*minPTS*) em sua vizinhança (incluindo o próprio ponto).

Exemplo: minPTS=4

Agrupamento baseado em Densidade

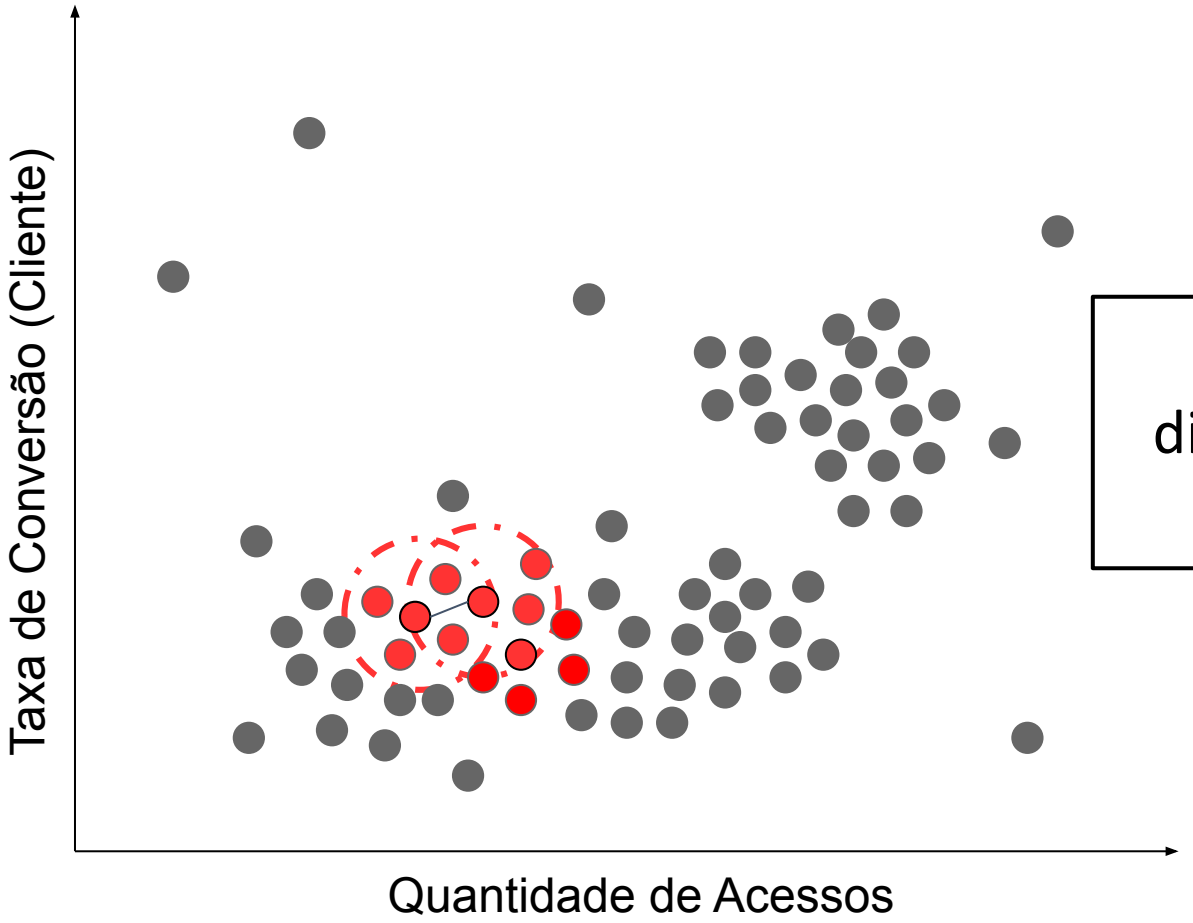


Pontos de Núcleo

Um ponto de núcleo contém um número mínimo de pontos (*minPTS*) em sua vizinhança (incluindo o próprio ponto).

Exemplo: minPTS=4

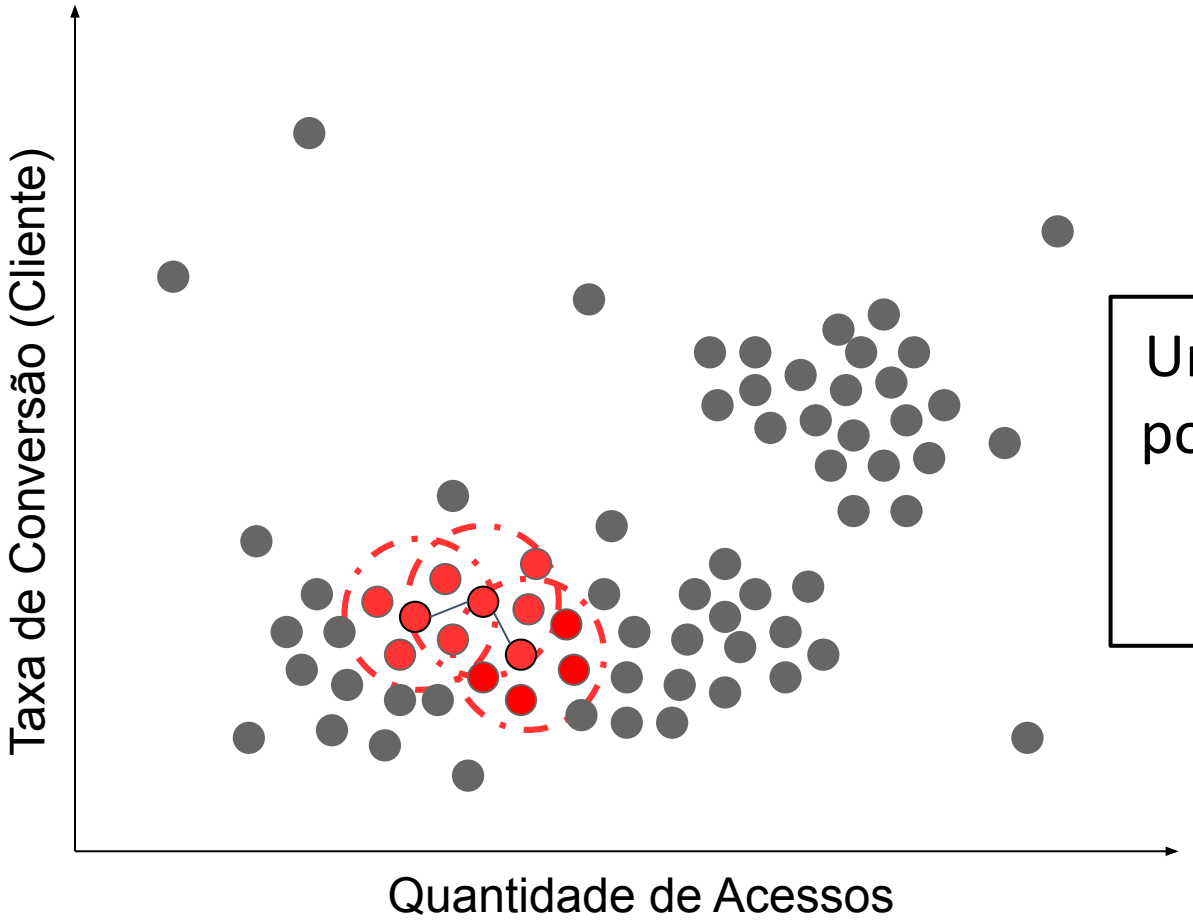
Agrupamento baseado em Densidade



Pontos de Núcleo

Um ponto pode ser diretamente alcançável por um ponto de núcleo.

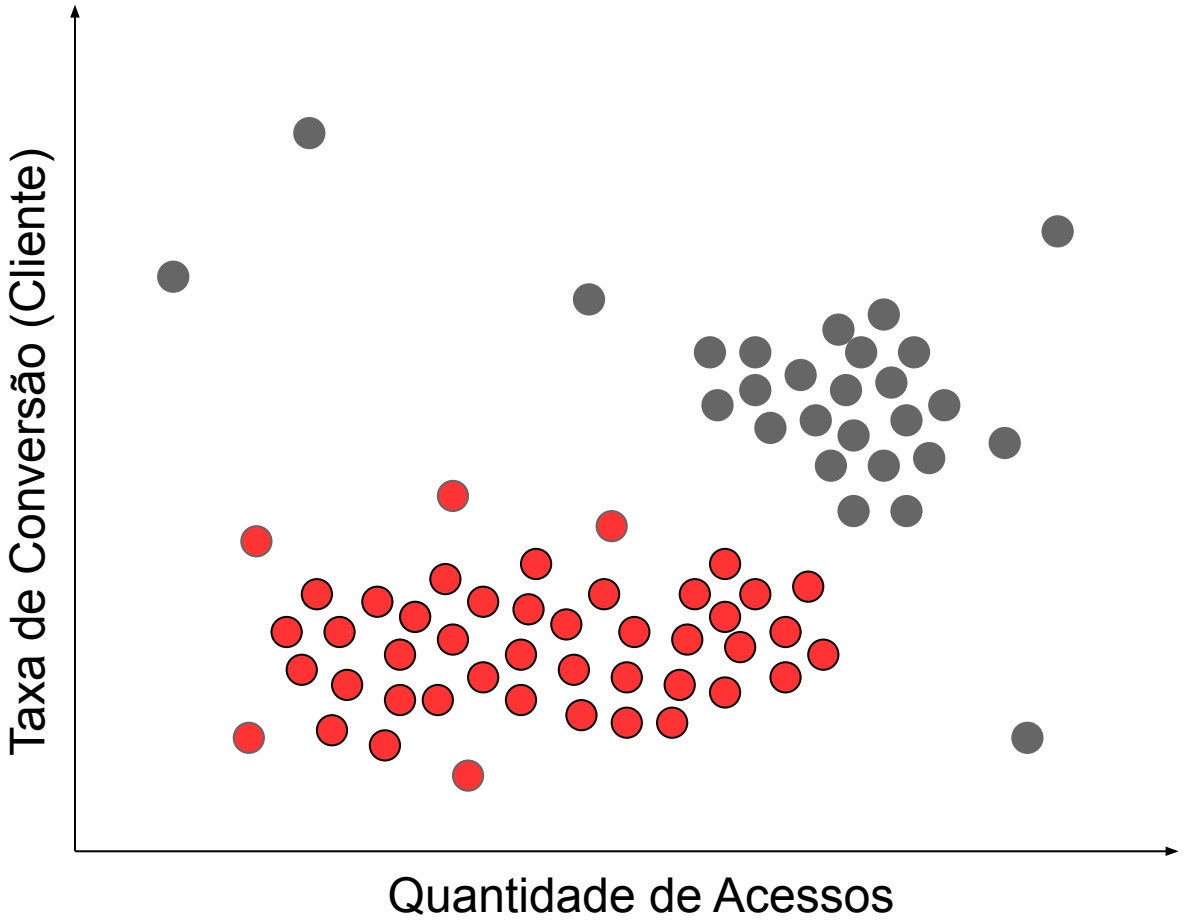
Agrupamento baseado em Densidade



Pontos de Núcleo

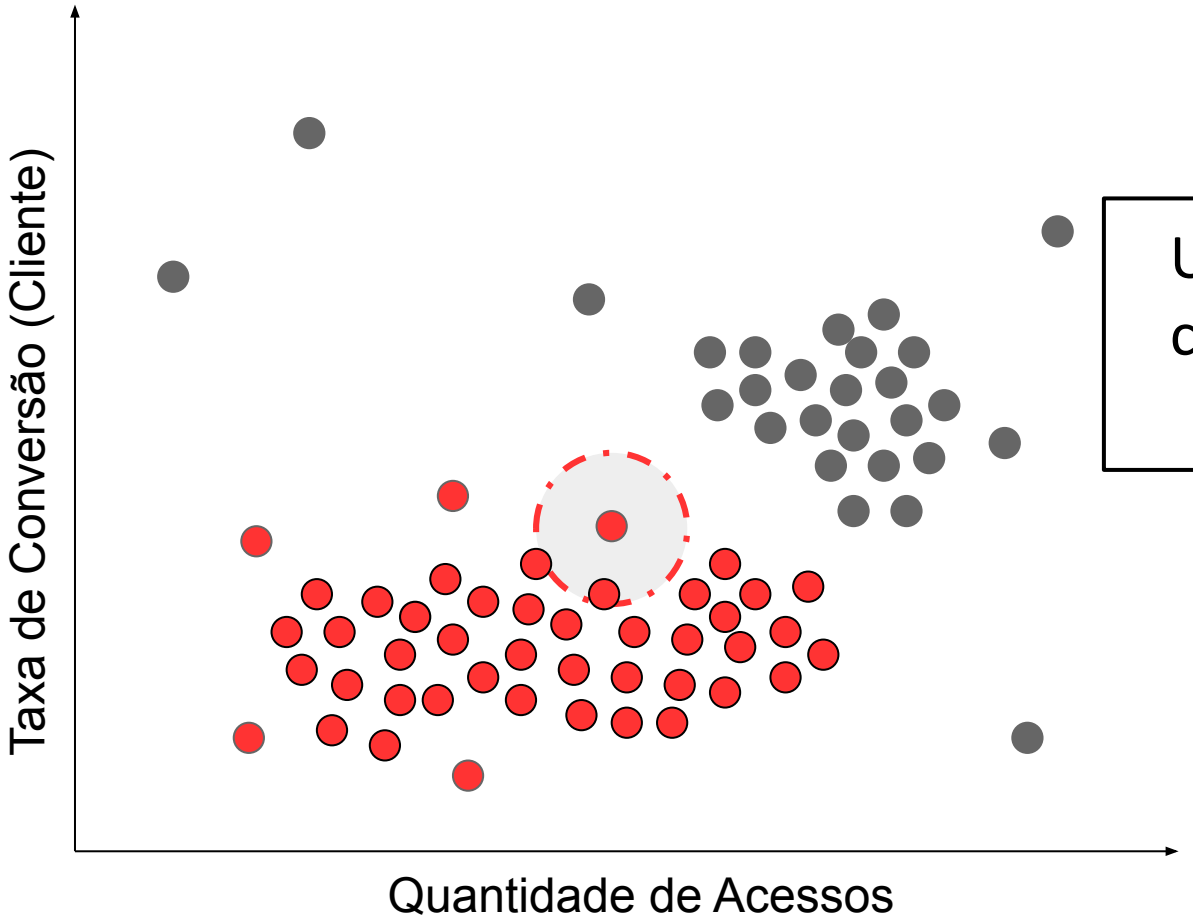
Um ponto pode ser alcançável por um ponto de núcleo, se há um caminho do ponto de núcleo até ele.

Agrupamento baseado em Densidade



Pontos de Fronteira

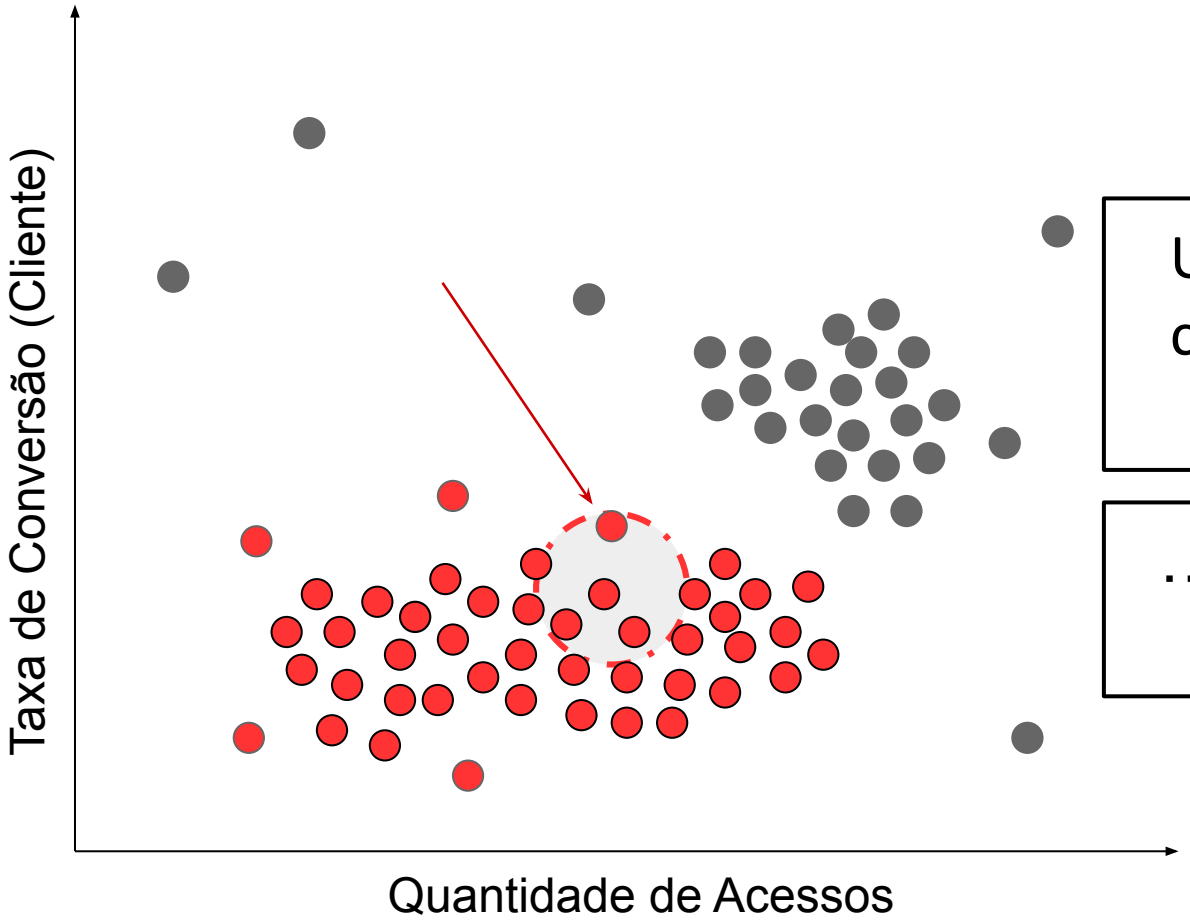
Agrupamento baseado em Densidade



Pontos de Fronteira

Um ponto de fronteira não contém *minPTS* pontos em sua vizinhança…

Agrupamento baseado em Densidade

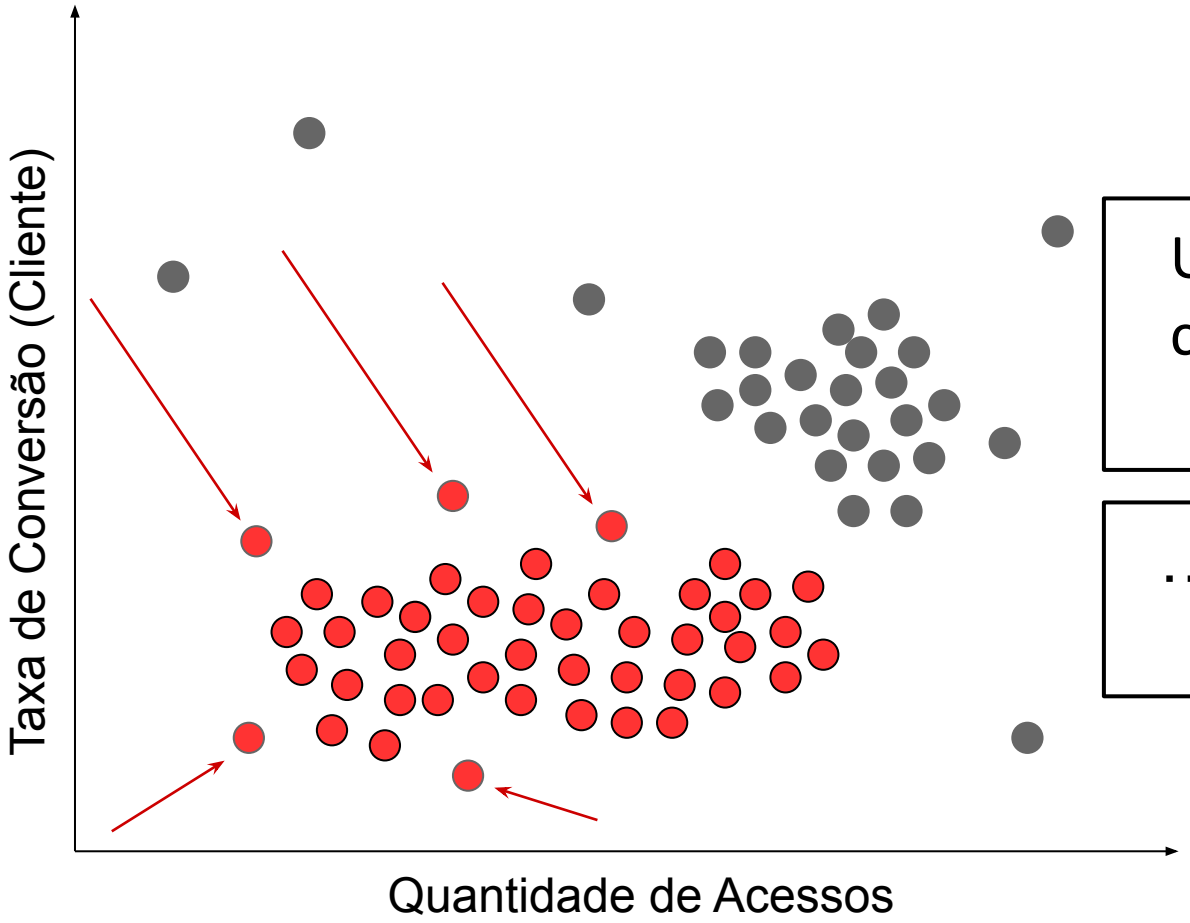


Pontos de Fronteira

Um ponto de fronteira não contém *minPTS* pontos em sua vizinhança…

… mas é alcançável por um ponto de núcleo.

Agrupamento baseado em Densidade

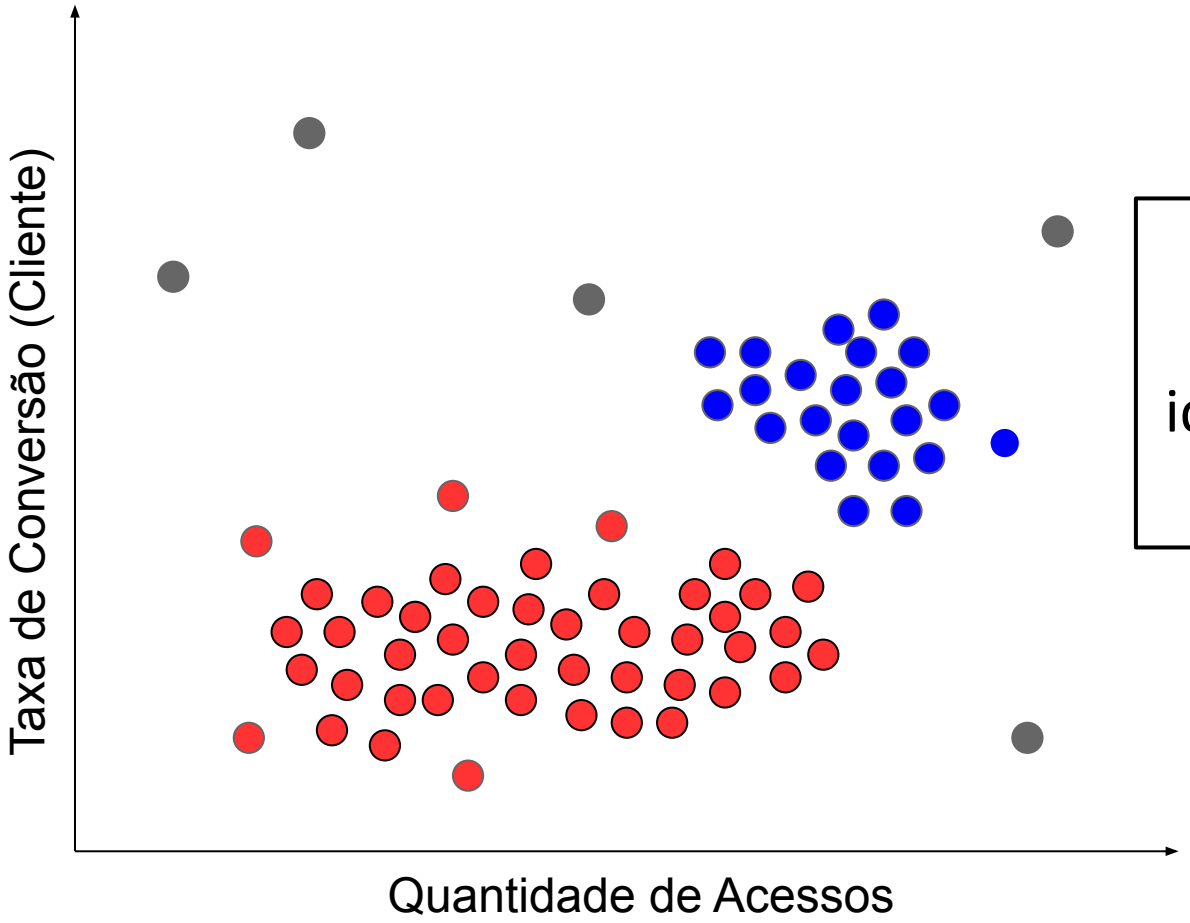


Pontos de Fronteira

Um ponto de fronteira não contém *minPTS* pontos em sua vizinhança…

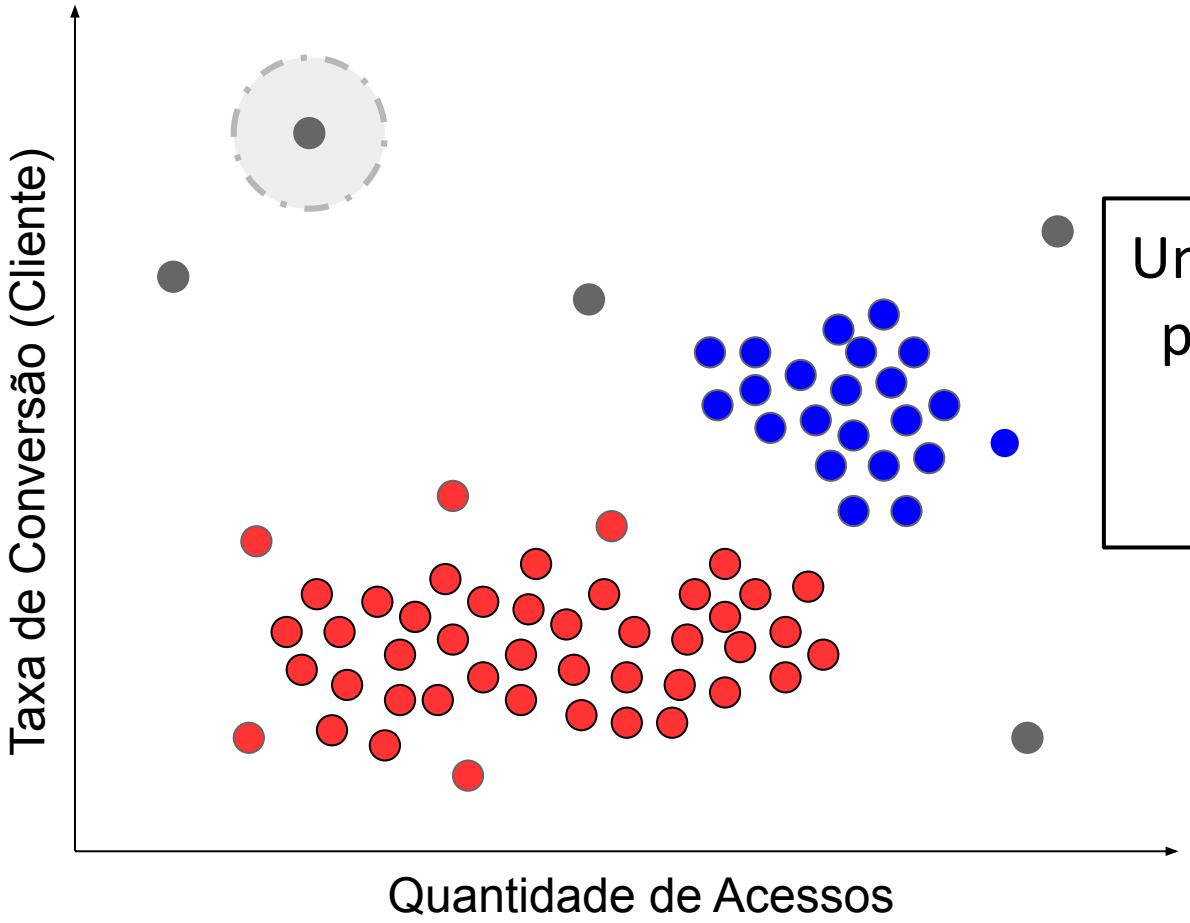
… mas é alcançável por um ponto de núcleo.

Agrupamento baseado em Densidade



Os *clusters* são obtidos aplicando tais critérios para identificar pontos conectados por densidade

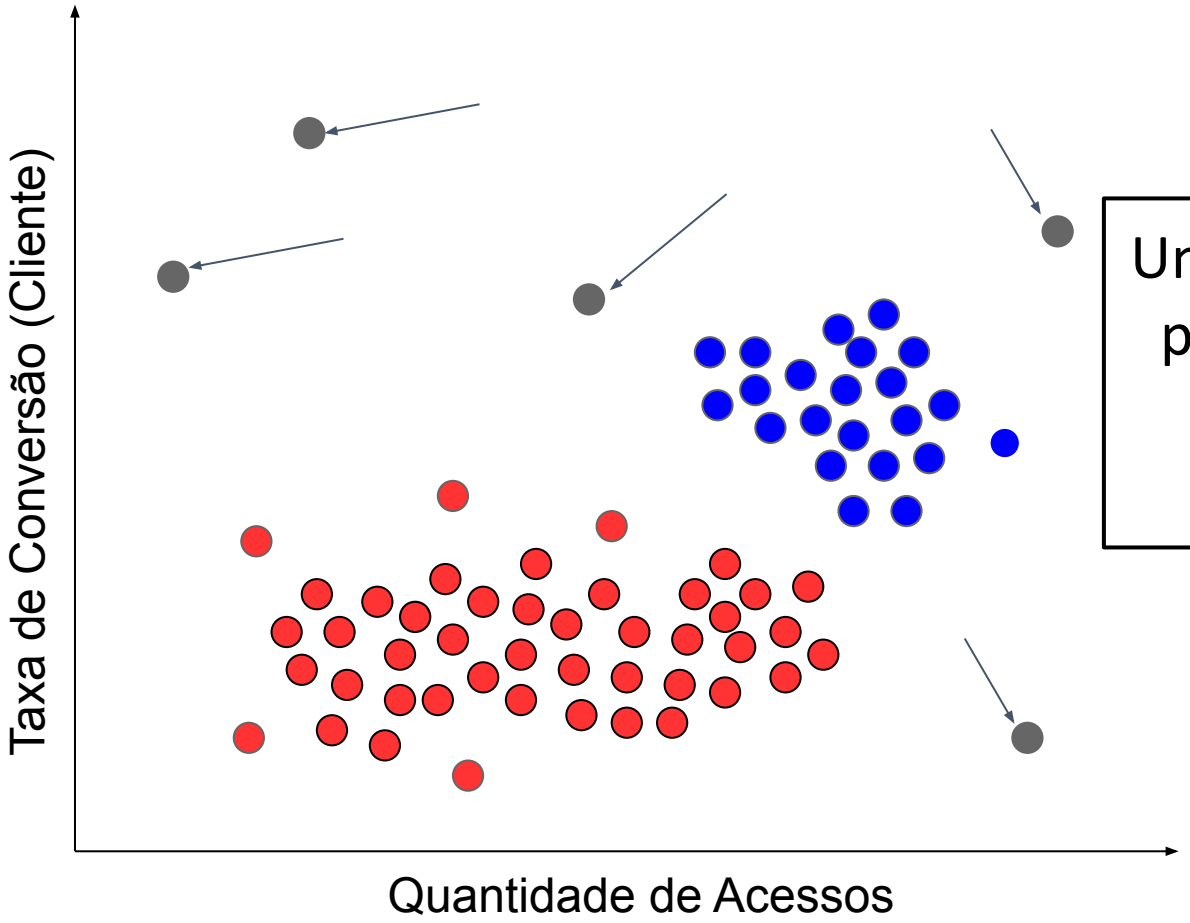
Agrupamento baseado em Densidade



Pontos *Outliers*

Um *outlier* não possui *minPTS* pontos em sua vizinhança e não é alcançável por um ponto de núcleo.

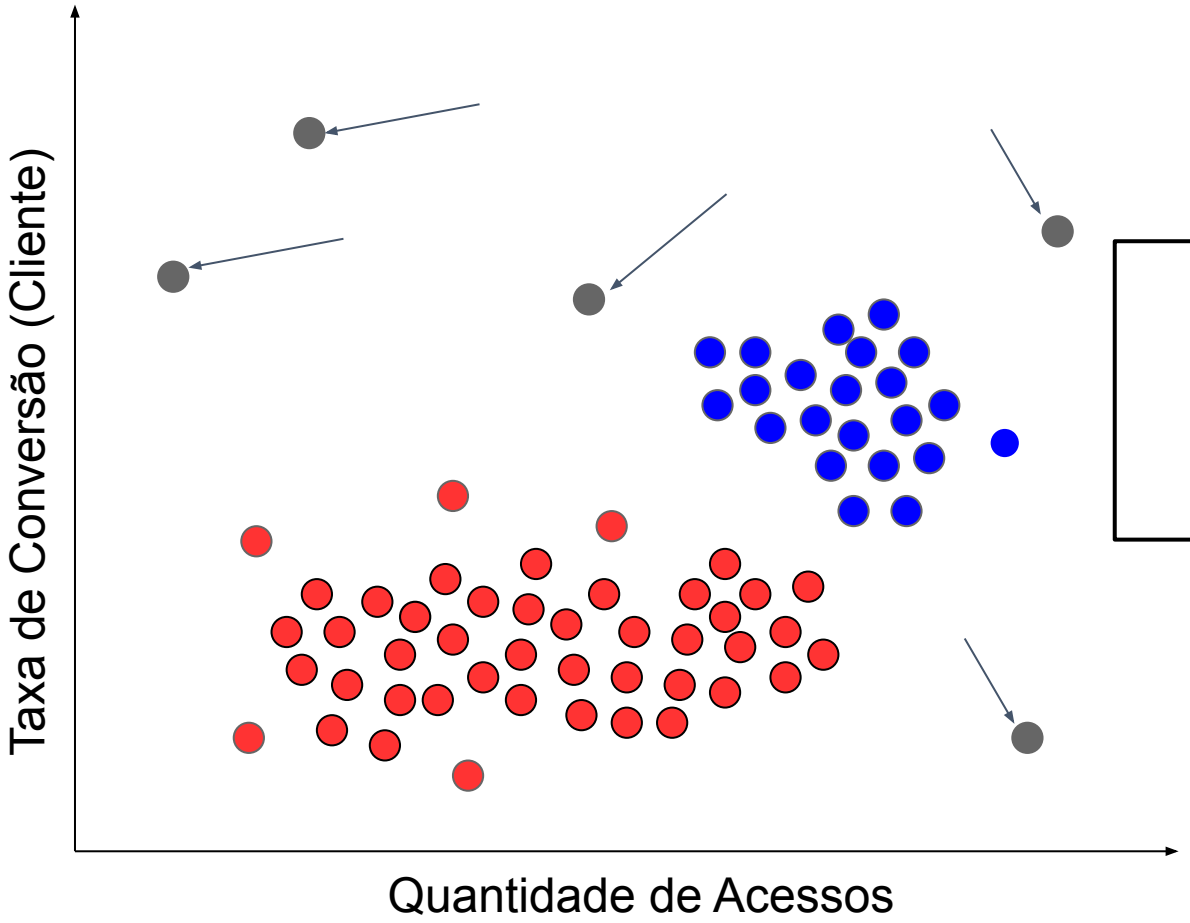
Agrupamento baseado em Densidade



Pontos *Outliers*

Um *outlier* não possui *minPTS* pontos em sua vizinhança e não é alcançável por um ponto de núcleo.

Agrupamento baseado em Densidade



Pontos *Outliers*

Como interpretar?

Ruído?
Anomalias?

Agrupamento baseado em Densidade

- Ferramenta *web* para demonstrar o DBSCAN
  - https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Agrupamento baseado em Densidade

- Vantagens

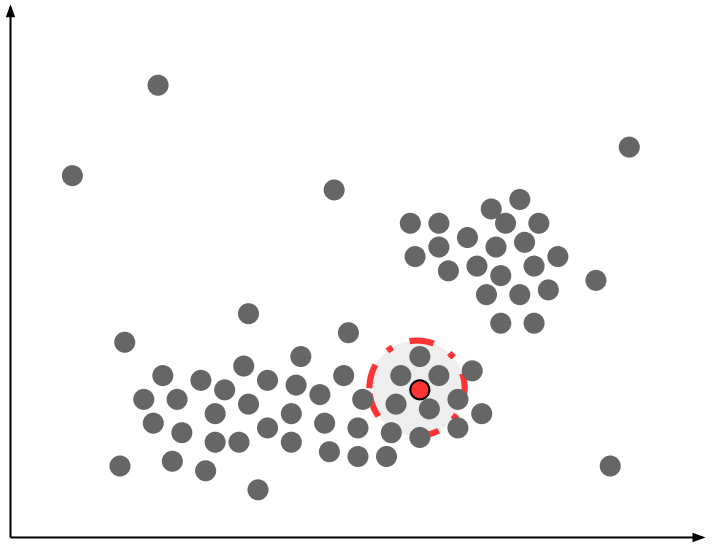
- Desvantagens

Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
- Desvantagens

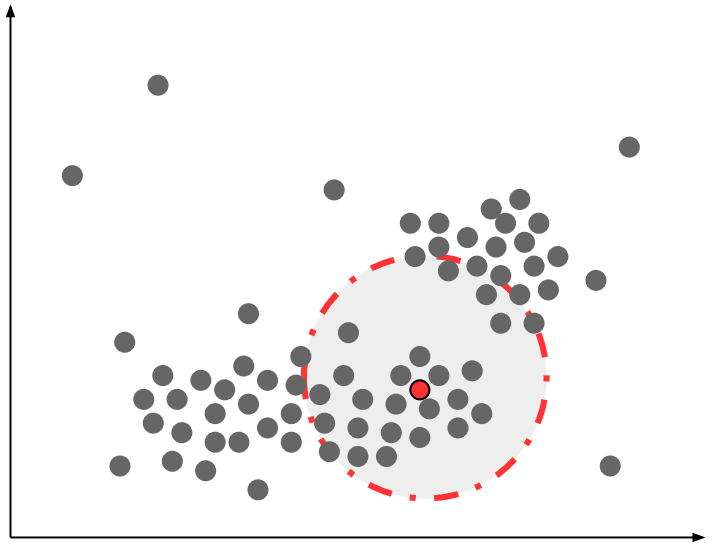
Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
- Desvantagens
  - **Precisamos definir os parâmetros ε e *minPTS***



Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
- Desvantagens
  - **Precisamos definir os parâmetros ε e *minPTS***



Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
  - **Robusto a *outliers***
- Desvantagens
  - **Precisamos definir os parâmetros ε e *minPTS***

Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
  - **Robusto a *outliers***
  - **Encontra *clusters* de diferentes formatos**
- Desvantagens
  - **Precisamos definir os parâmetros ε e *minPTS***

Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
  - **Robusto a *outliers***
  - **Encontra *clusters* de diferentes formatos**
- Desvantagens
  - **Precisamos definir os parâmetros ε e *minPTS***
  - **Custo computacional**

Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
  - **Robusto a *outliers***
  - **Encontra *clusters* de diferentes formatos**

- Desvantagens
  - **Precisamos definir os parâmetros ε e *minPTS***
  - **Custo computacional**

Podemos mitigar essa limitação ao usar estratégias para acelerar a busca dos vizinhos mais próximos!

Agrupamento baseado em Densidade

- Vantagens
  - **Não precisamos definir o número de *clusters***
  - **Robusto a *outliers***
  - **Encontra *clusters* de diferentes formatos**

- Desvantagens
  - **Precisamos definir os parâmetros ε e *minPTS***
  - **Custo computacional**

Podemos mitigar essa limitação ao usar estratégias para acelerar a busca dos vizinhos mais próximos!

Já estudamos esse assunto 😎

Bibliografia

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD (Vol. 96, No. 34, pp. 226-231). https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 1-21. https://doi.org/10.1145/3068335

Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. (2016). *Introduction to Data Mining (2nd Edition)*. Pearson.