FIOCRUZ

MINERAÇÃO DE DADOS

PARA NÃO BIOINFORMATAS

AGRUPAMENTO

RICARDO KHOURI

FELIPE TORRES

AGRUPAMENTO

- Agrupamento é a técnica de mineração de dados para descobrir grupos de instâncias relacionadas em conjuntos de dados

- Dividir os dados em grupos que sejam semelhantes ou úteis de alguma forma

- Aprendizado não-supervisionado.

AGRUPAMENTO PARA COMPREENSÃO

- Grupos com instâncias de propriedades compartilhadas.

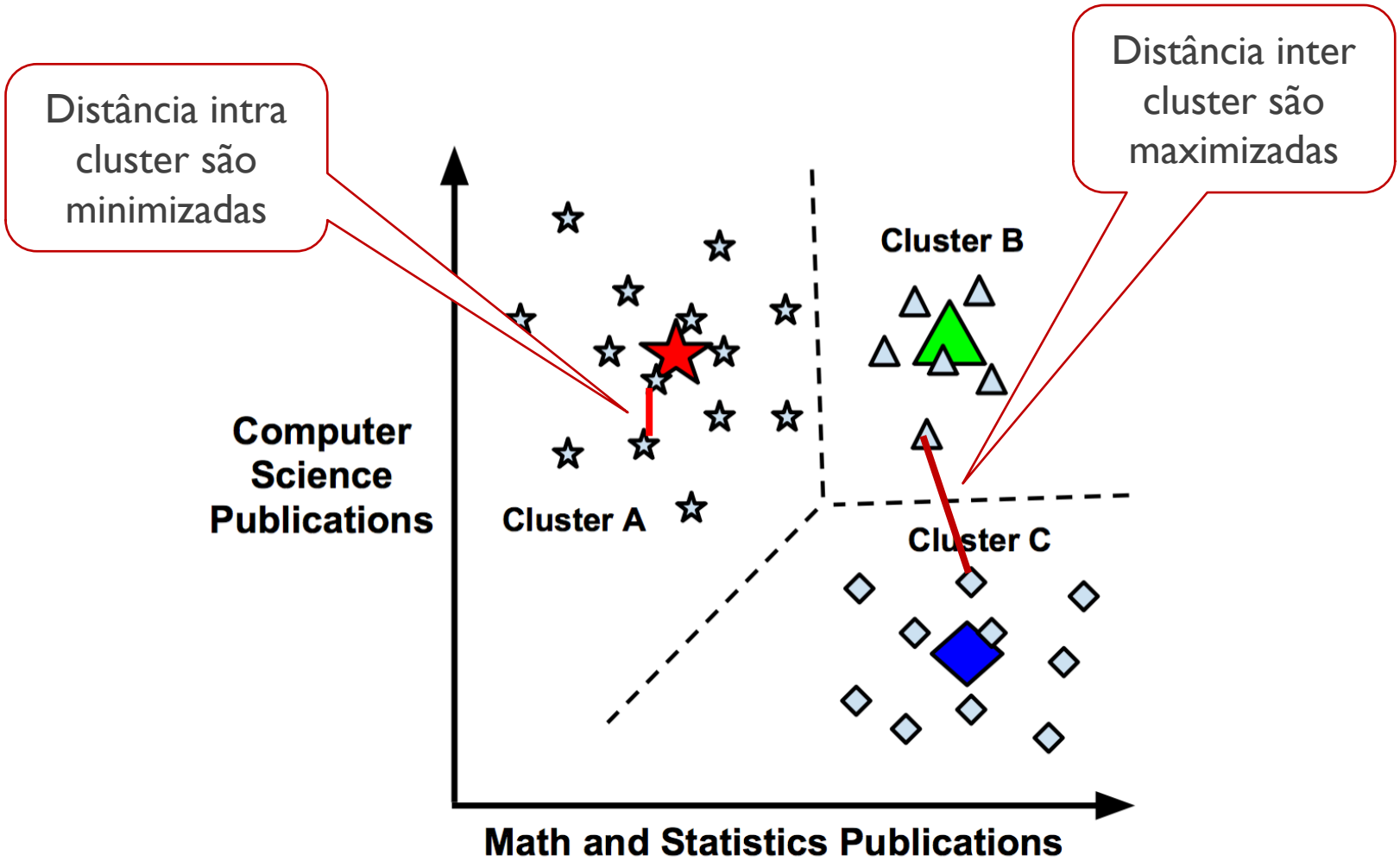
- Este tipo de agrupamento é comum no nosso dia-a-dia...

AGRUPAMENTO PARA UTILIDADE

- Buscar protótipos descritos de grupos

- Sumarização, compressão, vizinhos

AGRUPAMENTO PARA UTILIDADE



APLICAÇÕES DE AGRUPAMENTO

- Encontrar grupos de documentos similares
- Segmentar clientes com hábitos similares
- Descobrir grupos de genes com função similar
- Agrupar pacientes com mesma doença por sintomas
- Outras?

AGRUPAMENTO VS CLASSIFICAÇÃO

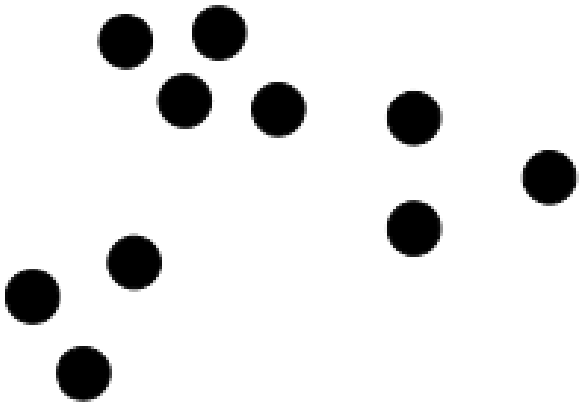
- **Classificação**
  - Aprender um método para predizer as categorias (classes) de padrões não vistos a partir de exemplos pré-rotulados (classificados)

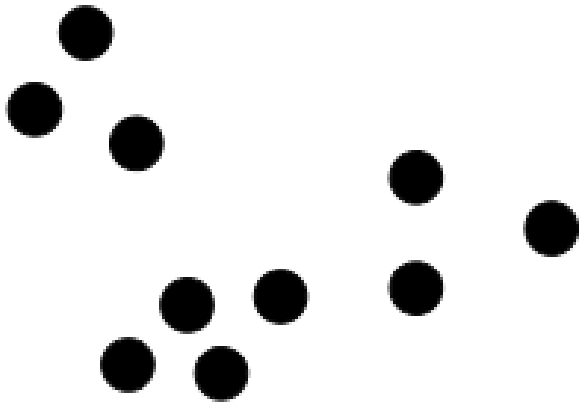
- **Agrupamento**
  - Encontrar os rótulos das categorias (grupos ou **clusters**) e possivelmente o número de categorias diretamente a partir dos dados

O QUE NÃO É AGRUPAMENTO ?

- Classificação supervisionada
  - Possui informação de rótulo de classe
- Segmentação simples
  - Divir alunos em grupos alfabeticamente
- Resultado de uma consulta
  - Agrupamento é resultado de uma especificação externa

QUANTOS GRUPOS EXISTEM ABAIXO ?





QUANTOS GRUPOS EXISTEM ABAIXO ?

Quantos grupos?

Seis grupos

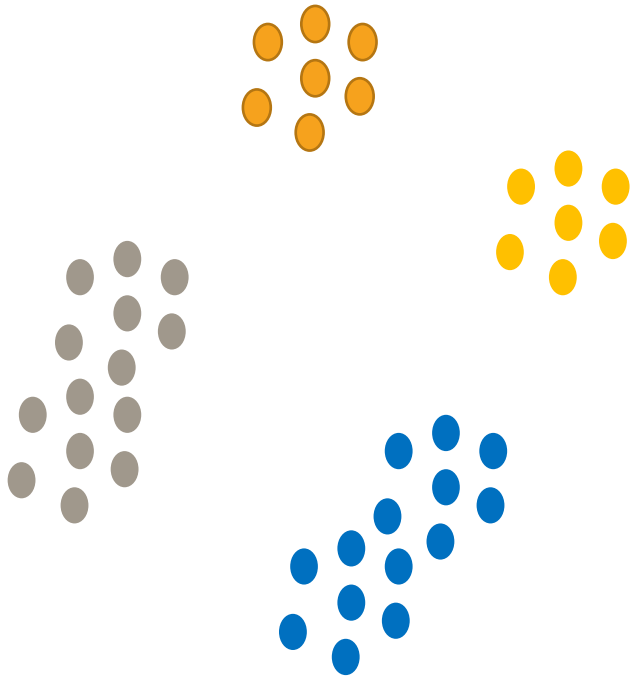
Dois grupos

Quatro grupos

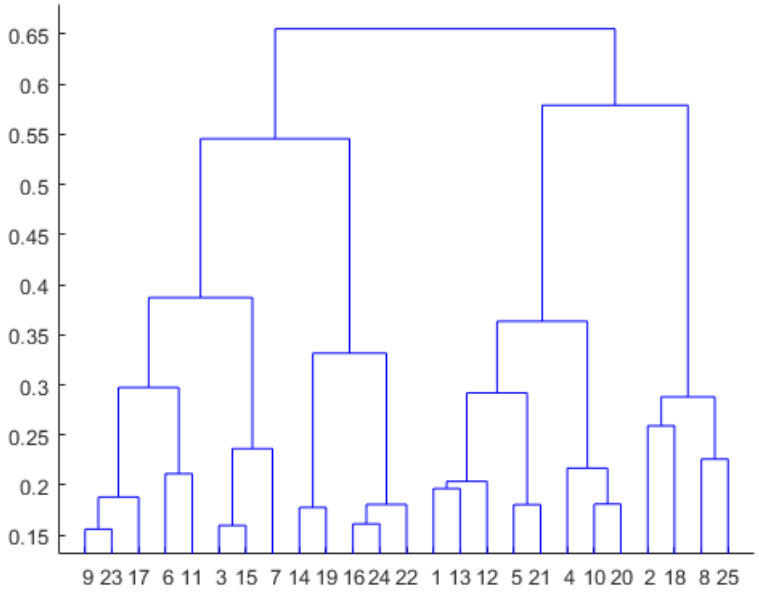
AGRUPAMENTOS

- Um agrupamento (clustering) é um conjunto de grupos (clusters)
- Partitional Clustering
  - Uma divisão dos objetos em subconjuntos (grupos) não-superpostos, cada objeto em exatamente um subconjunto
  - Kmeans, Fuzzy means e QT clustering
- Hierarchical clustering
  - Um conjunto de grupos organizados com uma árvores hierárquica
  - Aglomerador (Top-down) e Divisor (bottom-up)

AGRUPAMENTO



Partitional Clustering



Hierarchical Clustering

AGRUPAMENTO

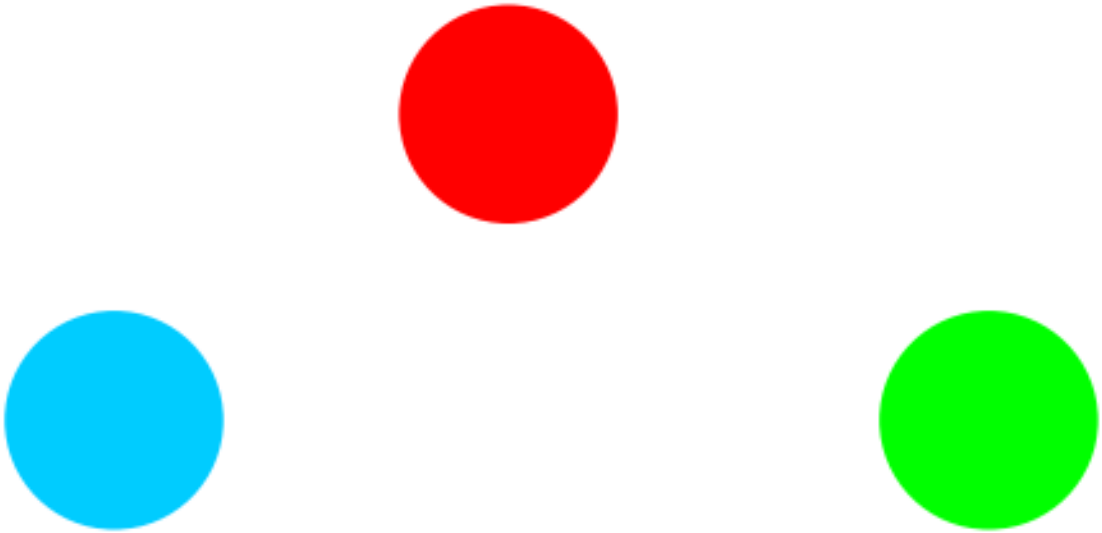
- Exclusivo vs não-exclusivo
  - Cada instância pode pertencer a um único grupo cluster ou a múltiplos grupos
- Fuzzy vs não-fuzzy
  - Em fuzzy clustering, cada instância pertence a cada grupo com pertinência entre 0 e 1
- Parcial vs completo
  - Em alguns casos, desejamos agrupar parte dos dados
- Heterogêneo ou Homogêneo
  - Grupos de tamanhos, formatos ou densidades diferentes ou iguais

TIPOS DE GRUPOS

- Tipos de cluster
  - Clusters bem-separados
  - Clusters baseados em centros
  - Clusters contíguos
  - Clusters baseados em densidade
  - Propriedade ou Conceitual
  - Descrito por uma função objetivo

TIPOS DE GRUPOS

- Clusters bem-separados
  - Qualquer ponto em um grupo é mais próximo a todos pontos do mesmo grupo do que a qualquer ponto fora do grupo.



3 well-separated clusters

TIPOS DE GRUPOS

- Clusters baseados em centros
  - Cada ponto é mais próximo do ‘centro’ do grupo do que do centro de qualquer outro grupo
  - O ‘centro’ do grupo pode ser um centróide, a média de todos pontos do grupo, ou um medóide, o ponto mais ‘representativo’ de um grupo

4 center-based clusters

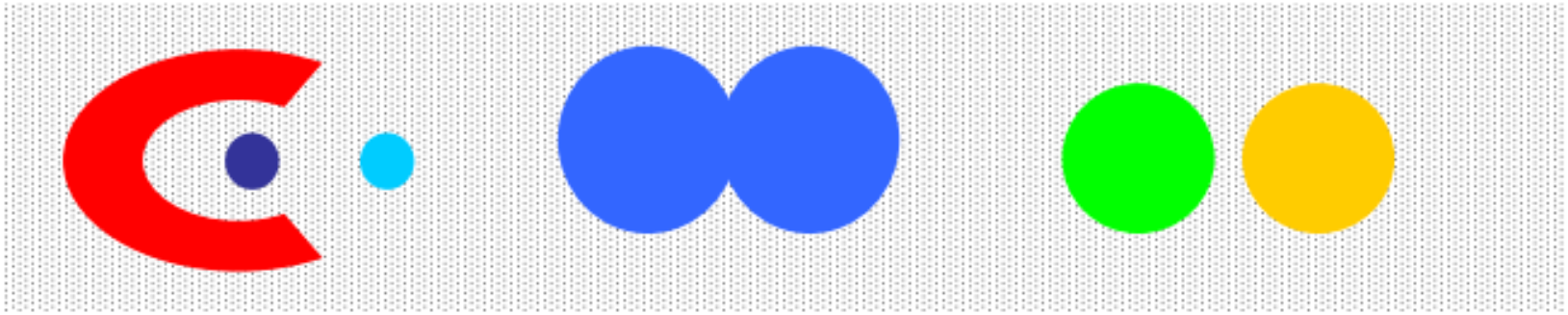
TIPOS DE GRUPOS

- Clusters contíguos (Nearest neighbor ou Transitivo)
  - Um ponto em um grupo é mais próximo de um ou mais pontos no grupo do que qualquer outro ponto for a do grupo.

8 contiguous clusters

TIPOS DE GRUPOS

- Clusters baseados em densidade
  - Um cluster é uma região densa de pontos, separada por regiões de baixa densidade de outras regiões de alta densidade
  - Usado para clusters irregulares ou entrelaçados, e quando há ruído e outliers



6 density-based clusters

TIPOS DE GRUPOS

- Clusters com Propriedade ou Conceito comum
  - Objetos dentro de um clusters compartilham uma propriedade ou representação um conceito particular.
- Clusters baseados em função objetivo
  - Cluster deve maximar/minimizar uma função objetivo
  - Avalie clustering proposto com base nesta função
  - Variação: ajustar os clusters a um modelo pré-definido

CARACTERÍSTICAS IMPORTANTES DOS DADOS

- Tipo de proximidade ou medida de densidade
- Esparcidade
- Cuidados com similaridade
- Tipos de atributos
- Definem tipo de similaridade
- Outras características, ex. Autocorrelação
- Dimensionalidade
- Ruído e Outliers
- Tipo de Distribuição

O BOM AGRUPAMENTO

- O que é um bom agrupamento?
  - alta similaridade intra-cluster: coesão do cluster
  - baixa similaridade inter-cluster: distinção entre cluster
- A qualidade de um agrupamento
  - Normalmente, há uma função de 'qualidade' que mede o quão bom é um grupo
  - É difícil definir 'suficientemente similar' ou 'suficientemente bom'
    - Resposta muito subjetiva

PASSO-A-PASSO BÁSICO PARA O AGRUPAMENTO

- **Seleção de características**
  - É necessário selecionar quais características serão utilizadas para a tarefa de agrupamento.
- **Medida de similaridade**
  - A decisão da medida de similaridade deve ser baseada no tipo de variável do seu dataset.
- **Critério de agrupamento**
  - Qual será o critério utilizado para o agrupamento das instâncias.
- **Algoritmo de agrupamento**
  - Definir qual o algoritmo mais se adequa a sua tarefa.
- **Validação e Interpretação dos resultados**
  - Avaliar os resultados, verificar a presença de problemas no procedimento e corrigi-los quando necessário.

SELEÇÃO DE CARACTERÍSTICAS

- É nessa etapa que a maioria dos erros e viés ou bias acontecem.

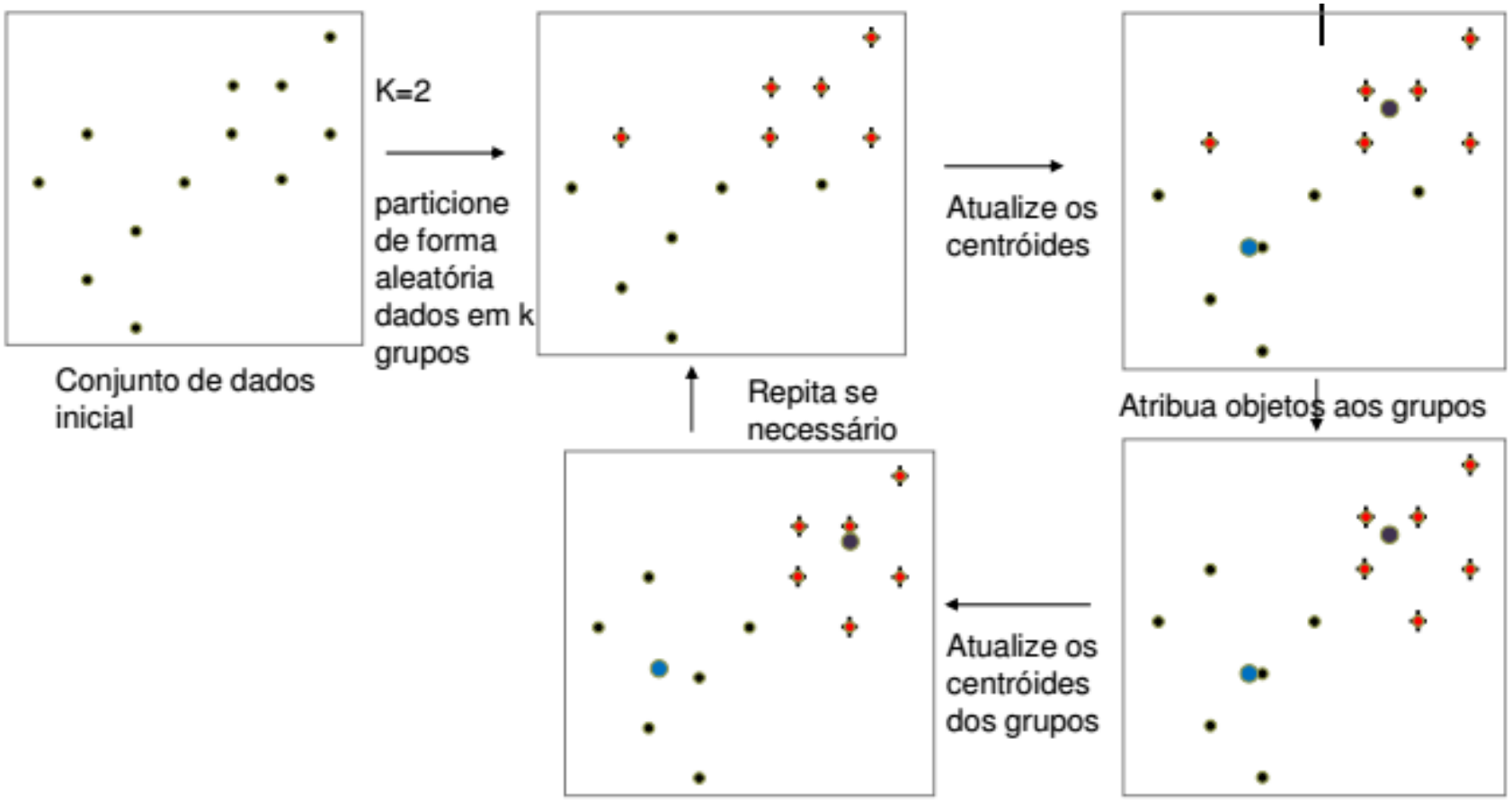
- Se forem selecionados atributos sem critérios e cometimento, o seu resultado será espúrio e não representará a realidade.

- As características selecionadas devem ser observadas e avaliadas utilizando algumas medidas e testes estatísticos.

SELEÇÃO DE CARACTERÍSTICAS – ASPECTOS E AVALIAÇÕES

- A taxa de missing data pode achatar e enviezar os seus resultados.
- É necessário que cada atributo não tenha uma sobreposição semântica entre os demais.
- Deve ser avaliado a correlação dos atributos com a sua hipótese para remover os atributos fracamente correlacionados.
- Em algumas situações o algoritmo exige que você realize algumas transformações nos seus dados.

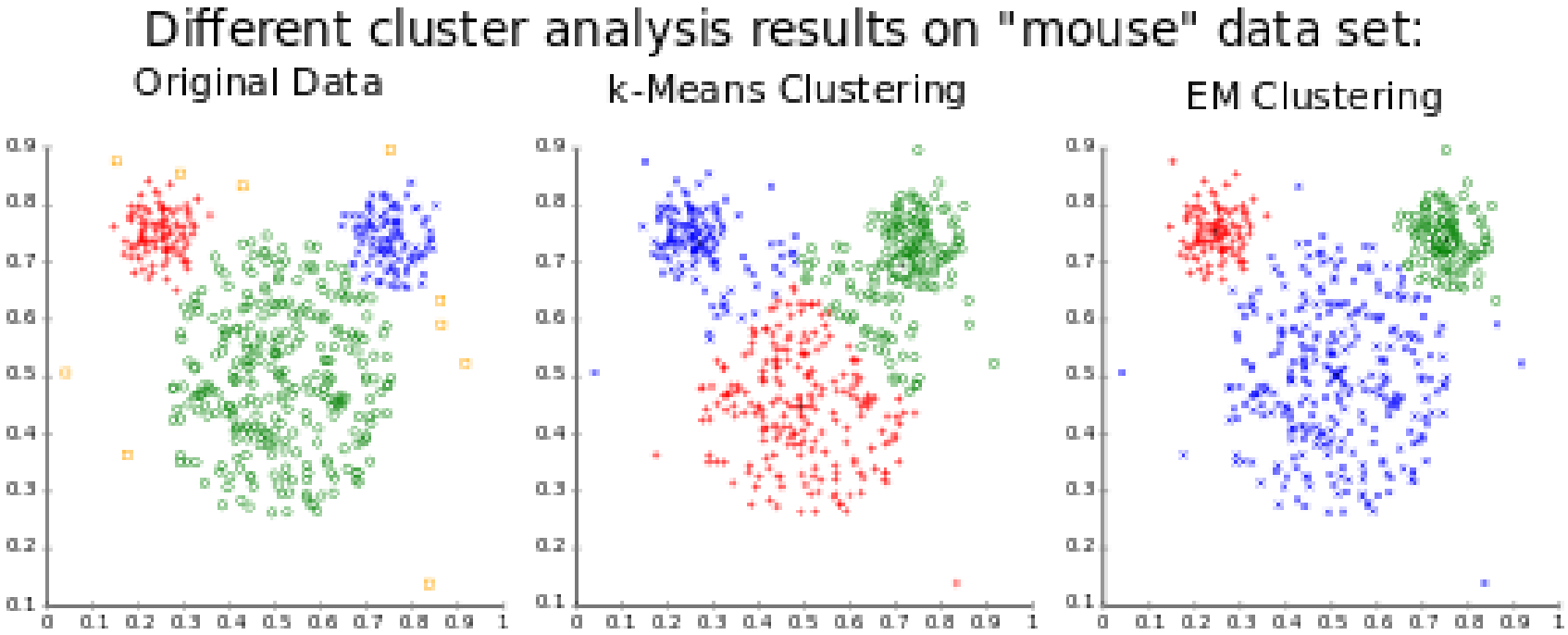
ALGORITMO DE K-MEANS



ALGORITMO DE K-MEANS

- **K-Means**
  - Eficiente: O (tkn), onde n é núm de objetos, k é núm de clusters, e t é núm de iterações. Normalmente, k, t << n.
  - Muitas vezes fica preso em mínimo local
  - Aplicável somente a objetos com atributos contínuos
    - Use k-modes para dados categóricos
  - Precisa especificar k, núm de cluster
  - Sensível a ruído e outliers
  - Não adequado a clusters de formato não convexo

ALGORITMO DE K-MEANS



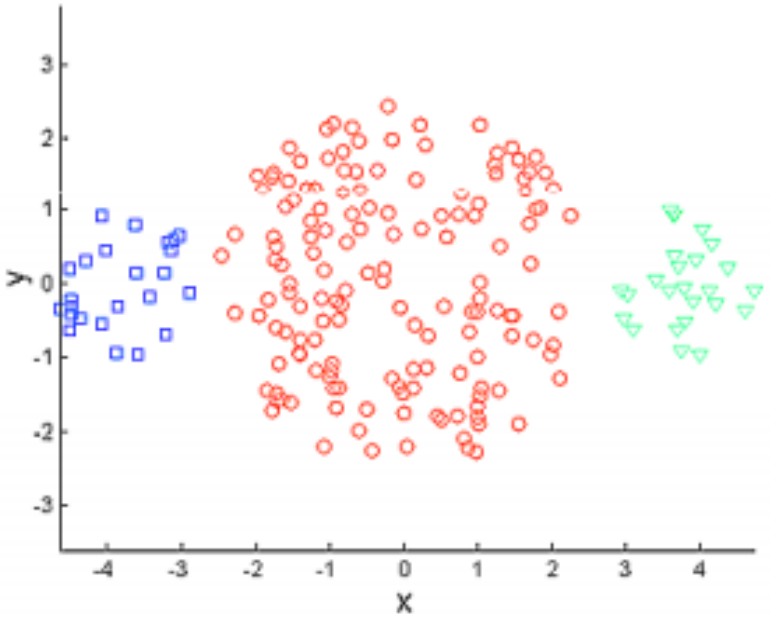
PONTOS QUE MERECEM ATENÇÃO NO USO DO K-MEANS

- K-means tem problema com grupos diferentes em:
  - Tamanho
  - Densidade
  - Formato não-globular

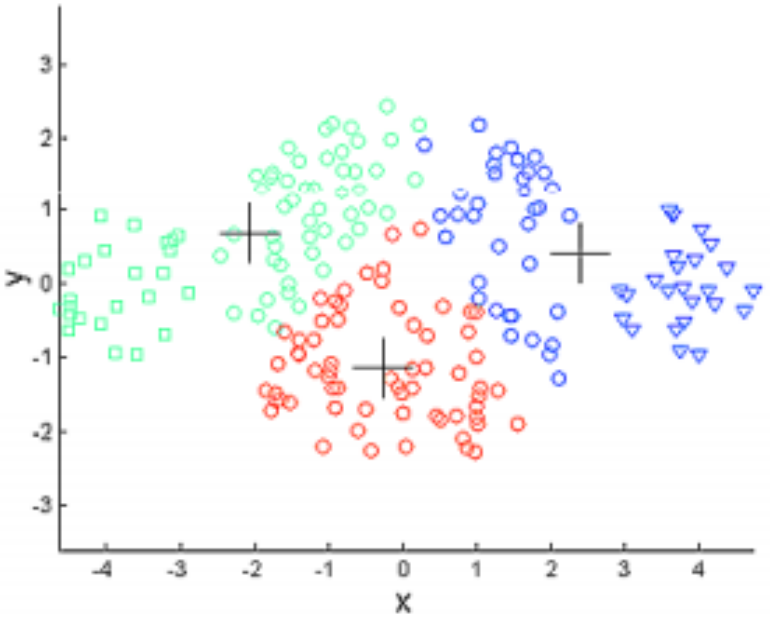
- A forma de contornar esses problemas é aumentando o número de clusters.

PONTOS QUE MERECEM ATENÇÃO NO USO DO K-MEANS

- Diferença de tamanho



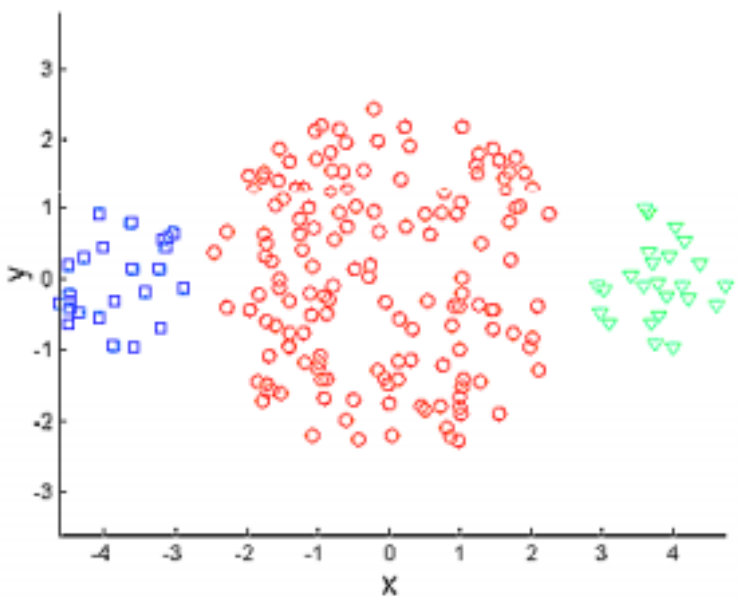
Original Points



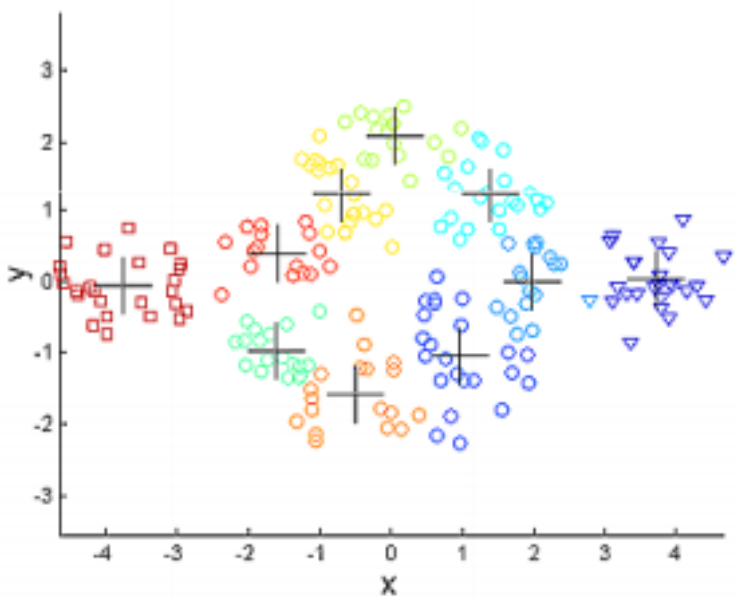
K-means (3 Clusters)

PONTOS QUE MERECEM ATENÇÃO NO USO DO K-MEANS

- Solucionando a diferença de tamanho



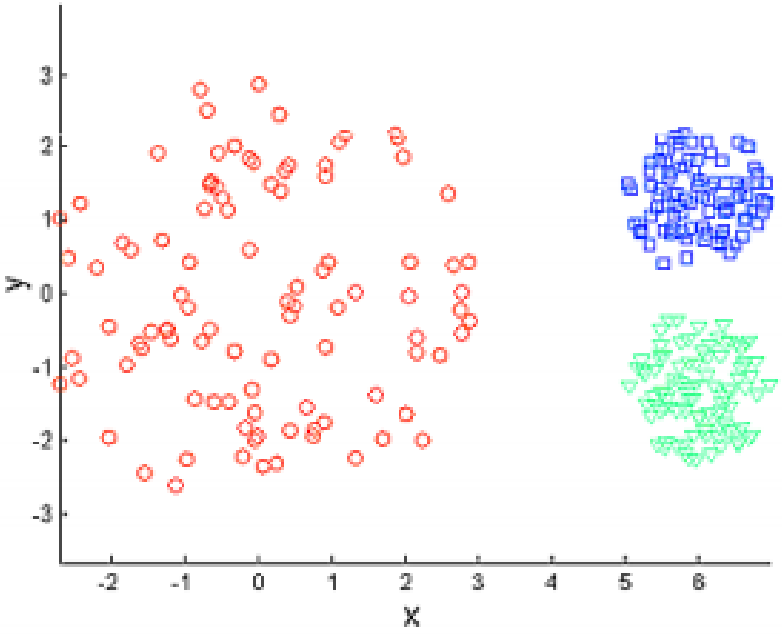
Original Points



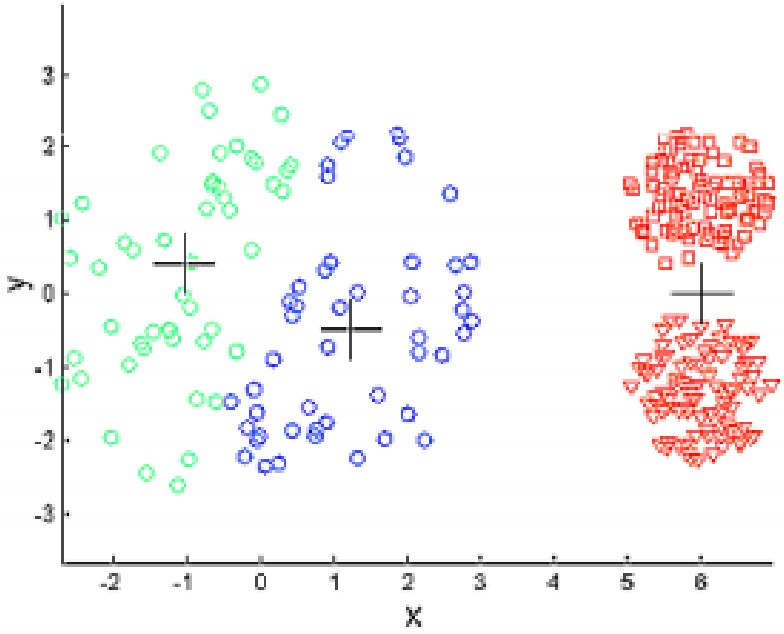
K-means Clusters

PONTOS QUE MERECEM ATENÇÃO NO USO DO K-MEANS

- Diferença de densidade



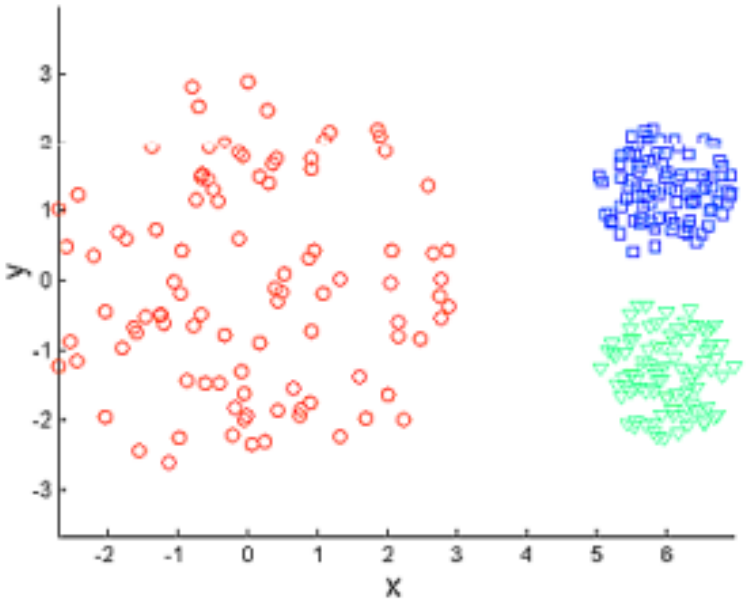
Original Points



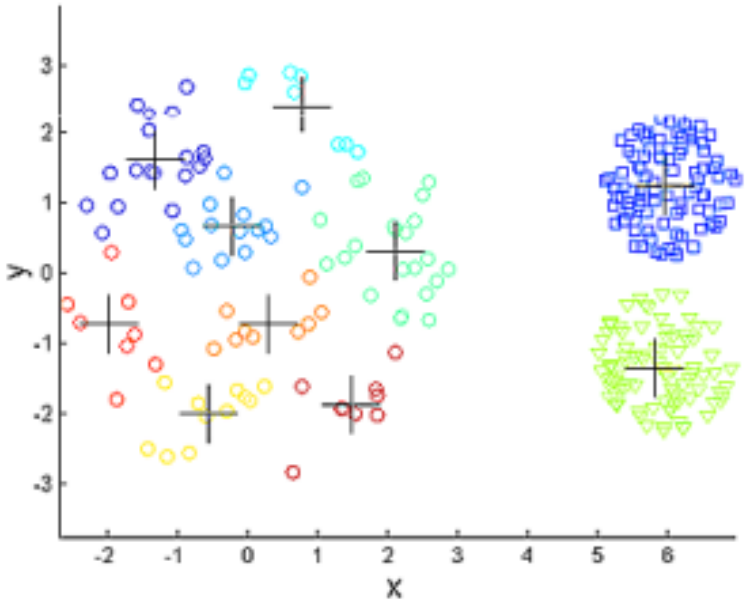
K-means (3 Clusters)

PONTOS QUE MERECEM ATENÇÃO NO USO DO K-MEANS

- Solucionando a diferença de densidade



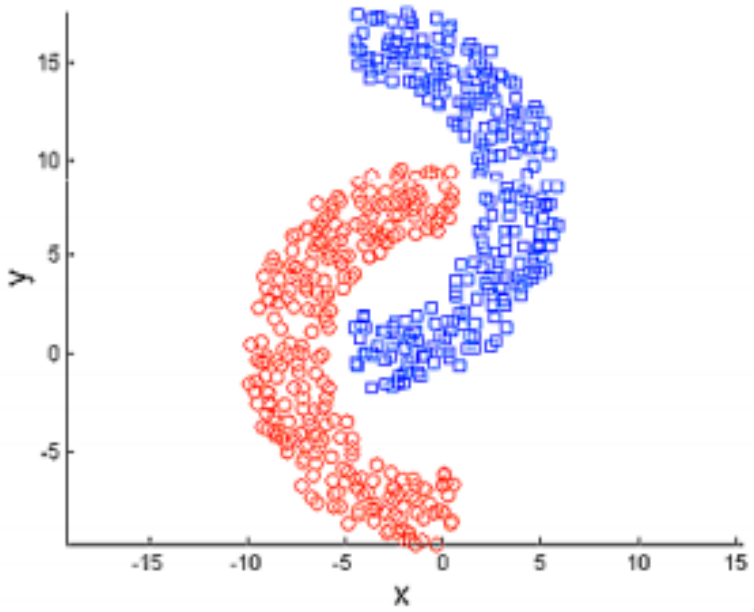
Original Points



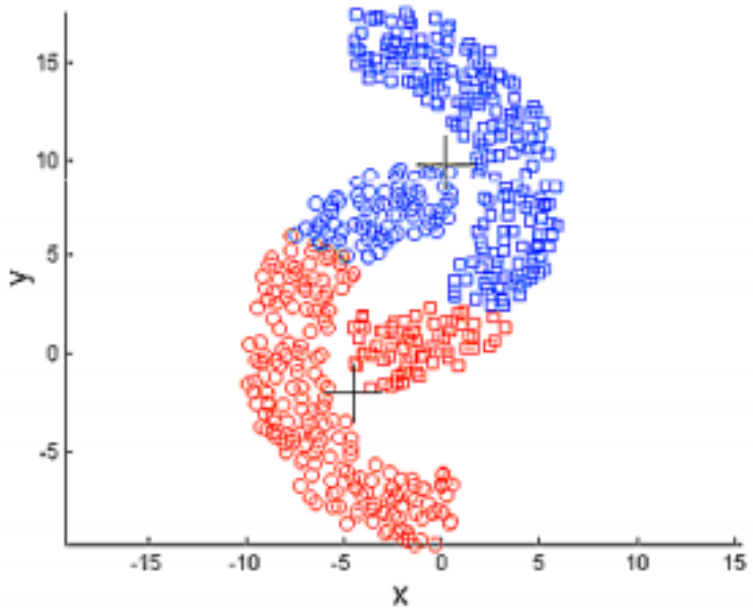
K-means Clusters

PONTOS QUE MERECEM ATENÇÃO NO USO DO K-MEANS

- Fórmula não-globular



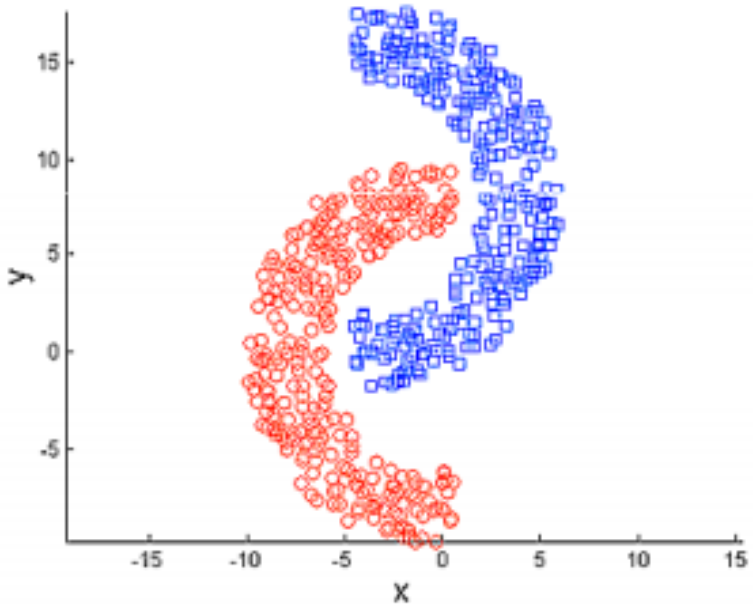
Original Points



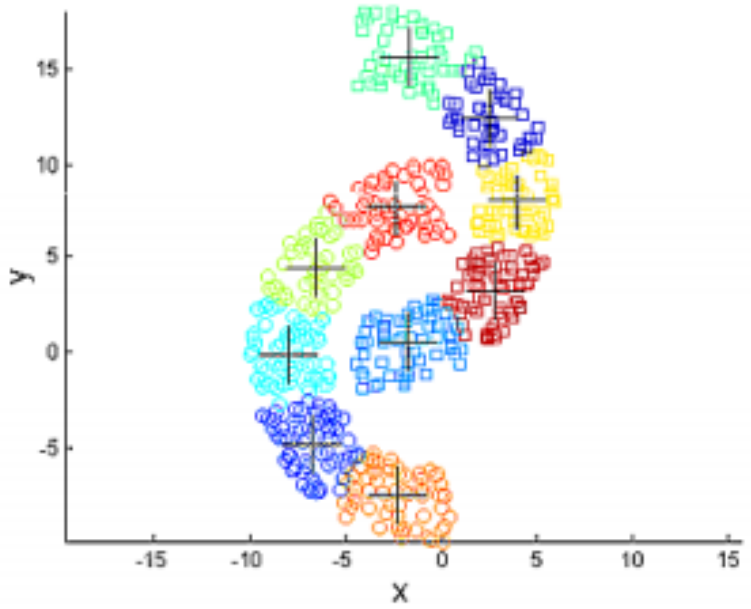
K-means (2 Clusters)

PONTOS QUE MERECEM ATENÇÃO NO USO DO K-MEANS

- Solucionando a fórmula não-globular



Original Points



K-means Clusters

DICAS DE AGRUPAMENTO COM K-MEANS

- A escolha dos centros iniciais é crítico
- Múltiplas execuções ajuda, mas a chance é pequena
- Amostre e use hierarchical clustering para determinar centróides iniciais
- Selecione mais que k centróides iniciais e selecione entre estes centróides iniciais (os mais distantes entre si)
- Pós-processamento
- Bisecting K-means

DICAS DE AGRUPAMENTO COM K-MEANS

- Pré-processamento
  - Normalize os dados
  - Elimine outliers
- Pós-processamento
  - Elimine clusters pequenos (possível outlier)
  - Divida clusters com alto SSE relativo
  - Junte clusters próximos e com baixo SSE relativo
  - Use estes passos durante processo de clustering

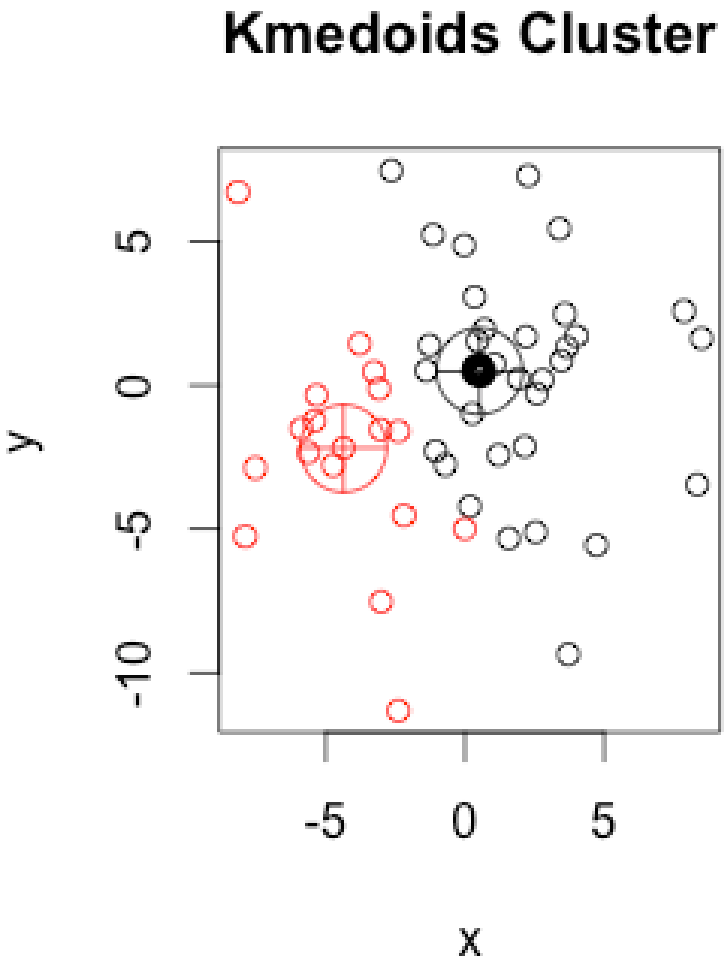
VARIAÇÕES DO ALGORITMO DE K-MEANS

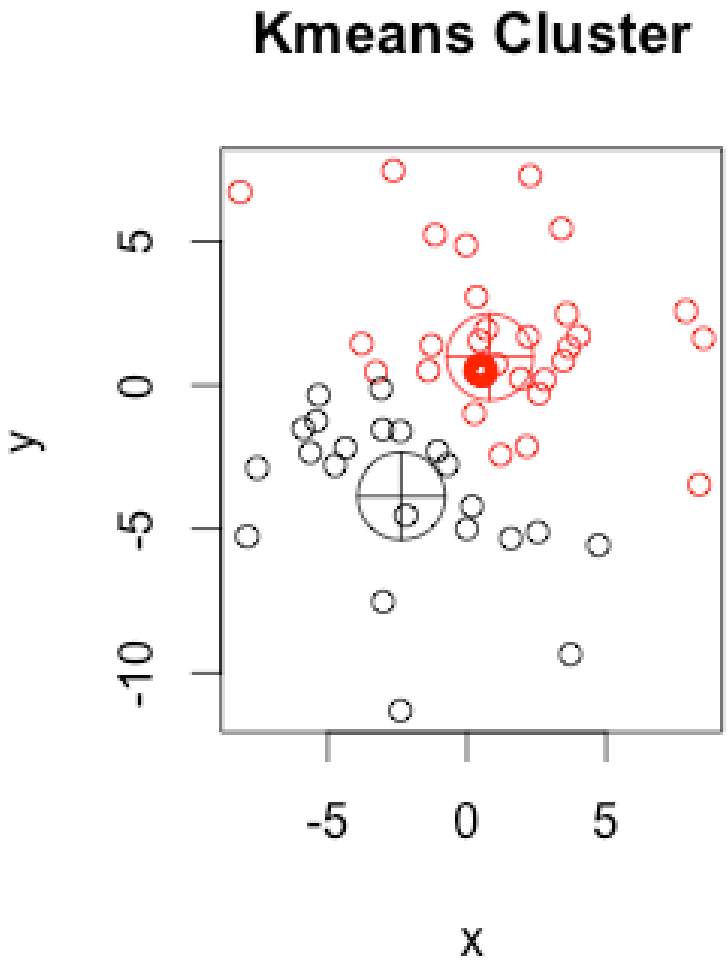
- Seleção inicial dos centroides:
  - Escolha aleatória de objetos da base, sorteio valores aleatórios, etc
  - Cálculo de dissimilaridade
  - Estratégias de cálculo do centroide

ALGORITMO DE K-MEDOIDS

- **K-Medoids**
  - Menos sensível a ruído e outliers
  - Ao invés da média, use o objeto mais centralmente posicionado
  - PAM (Partitioning Around Medoids, Kaufmann & Rousseeuw 1987)
    - Começa com medóides iniciais e iterativamente troca um dos medóides com por um não-medóide se melhorar a distância total do agrupamento resultante
    - somente bases de dados pequenas, alta complexidade computacional
  - Melhoras de eficiência: CLARA (amostras), CLARANS (comparação por amostragem)

ALGORITMO DE K-MEDOIDS

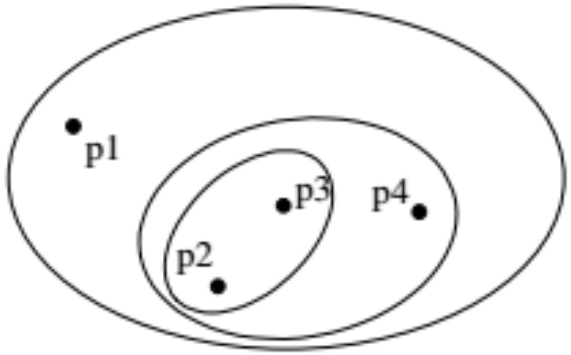




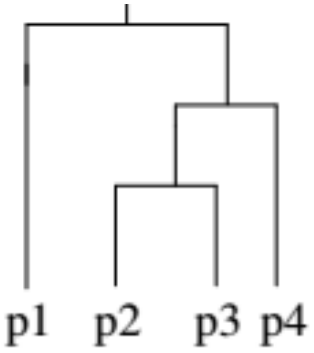
AGRUPAMENTO HIERÁRQUICO

- Um agrupamento hierárquico e representado por uma árvore.
- Os nós folhas são os objetos.
- Cada nó intermediário representa o agrupamento que contém todos os objetos de seus descendentes.

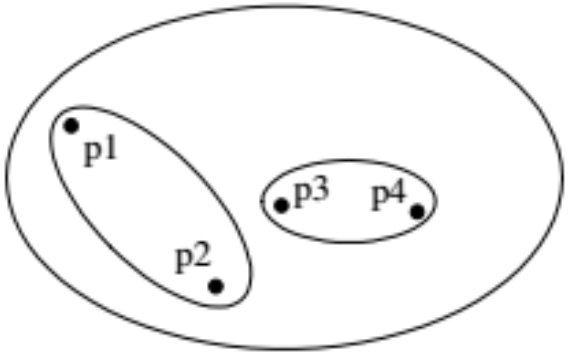
AGRUPAMENTO HIERÁRQUICO



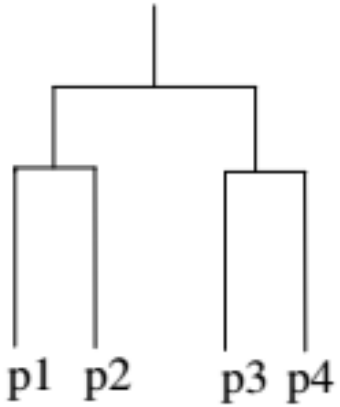
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering

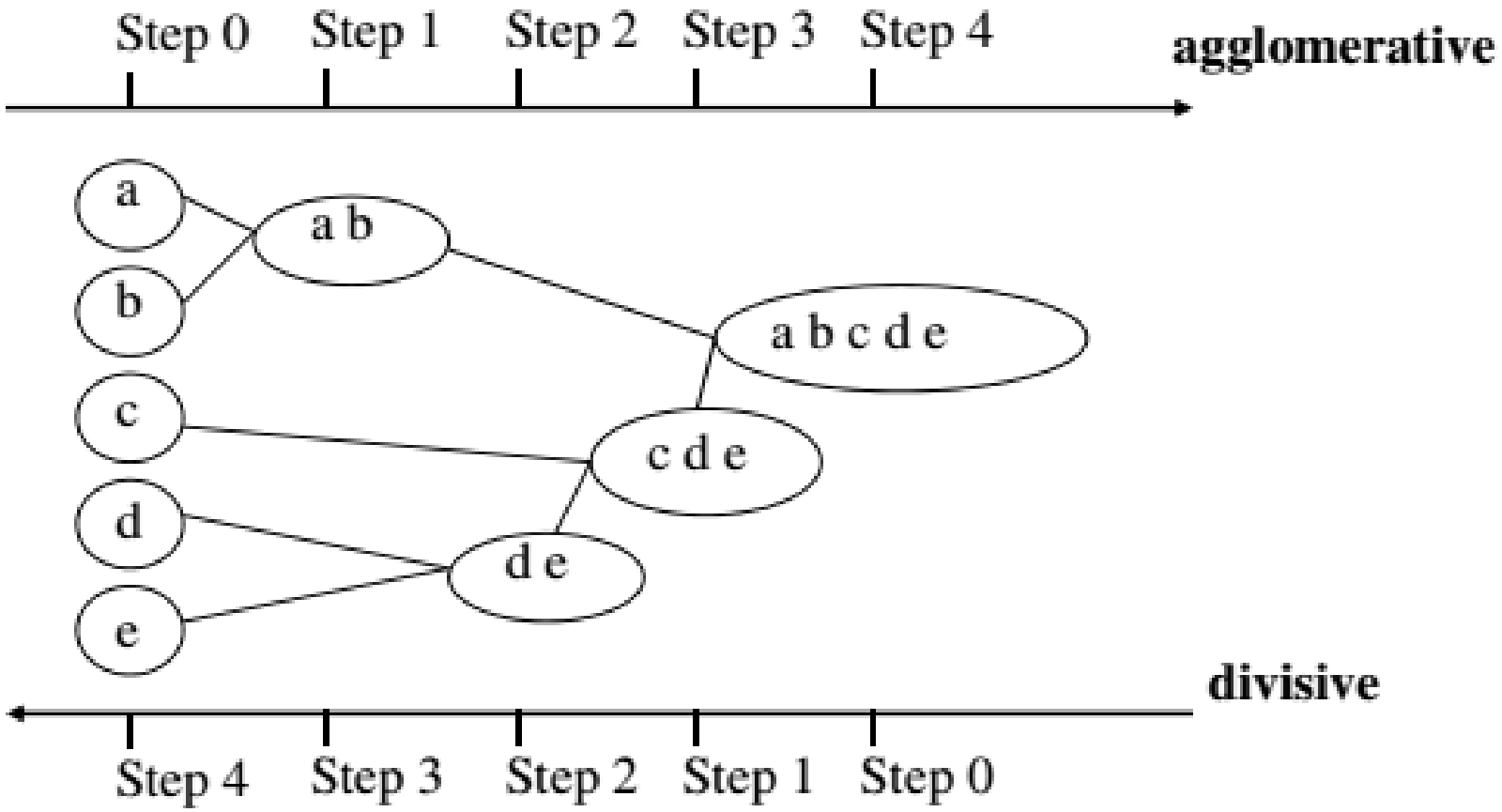


Non-traditional Dendrogram

AGRUPAMENTO HIERÁRQUICO

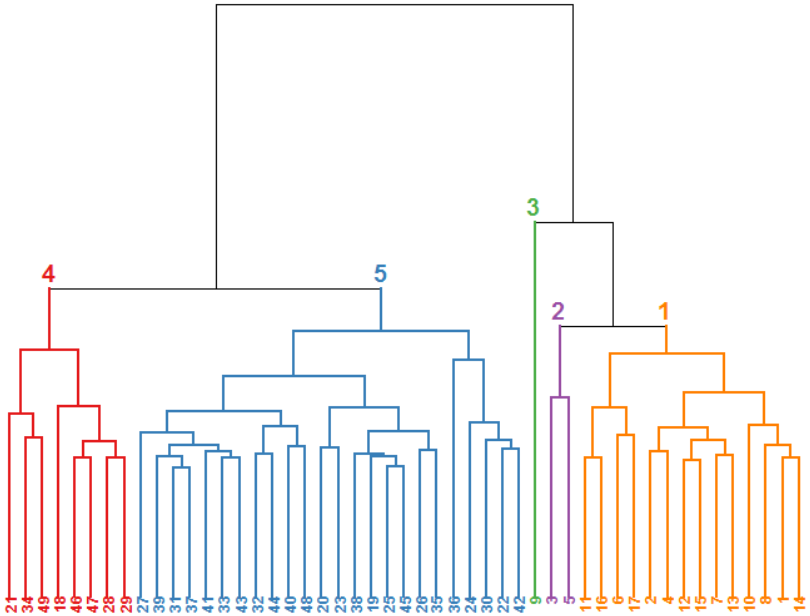
- Hierarchical Clustering
  - Não necessita definir um número de cluster previamente: corte no dendrograma
  - Pode corresponder a uma taxonomia significante
- Duas abordagens principais
  - Aglomerativa: pontos individuais formam clusters, clusters são combinados
  - Divisiva: começa com um cluster com todos, dividi-se cluster em outros clusters

AGRUPAMENTO HIERÁRQUICO



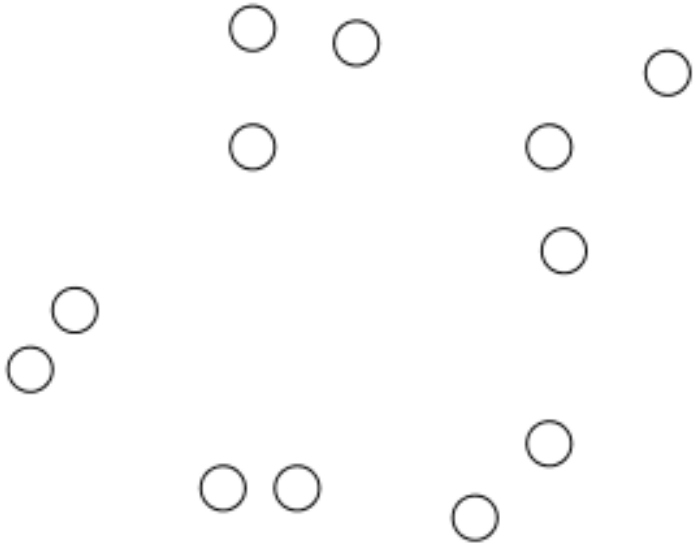
DENDROGRAMA

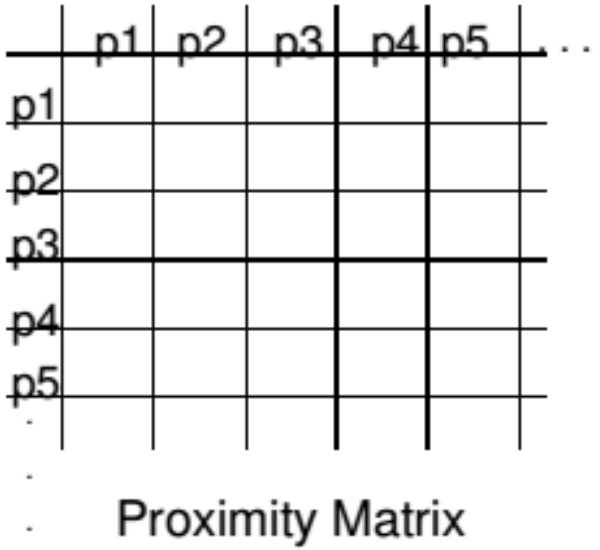
- Dendrograma: decomposição de instâncias em vários níveis de partições aninhadas (árvore de clusters)
- Um agrupamento (clustering) destas instâncias é obtido cortando o dendrograma no nível desejado, então cada componentes conectado é um cluster



ALGORITMO AGLOMERATIVO

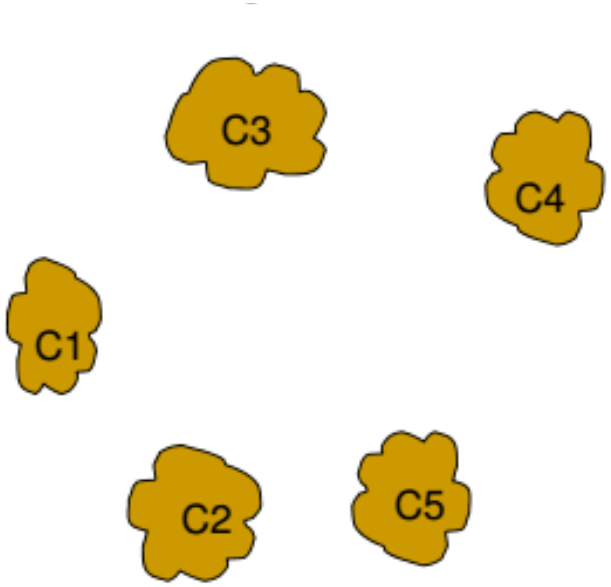
- Inicie com clusters de pontos individuais e matriz de proximidade.





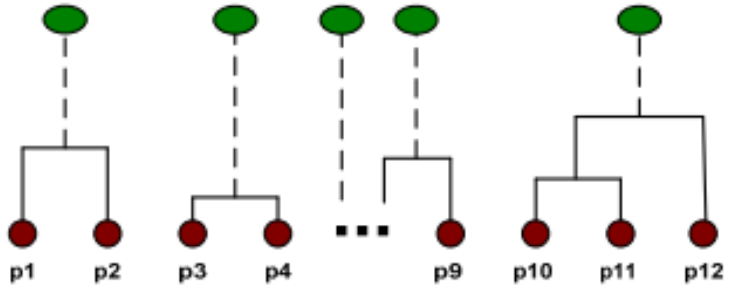
ALGORITMO AGLOMERATIVO

- Será calculada a distância entre cada elemento e os mais próximos formaram clusters.



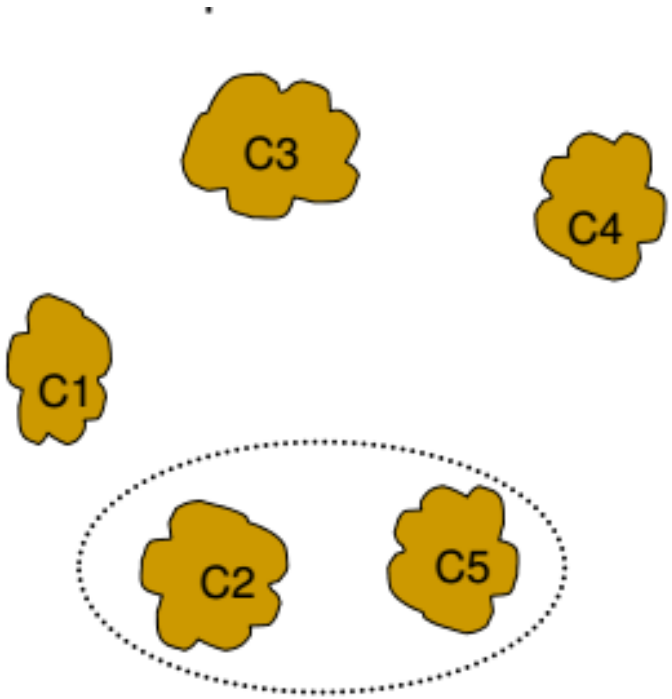
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

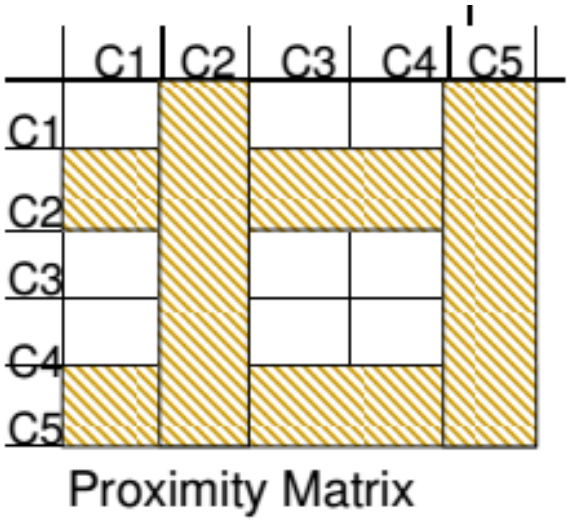
Proximity Matrix

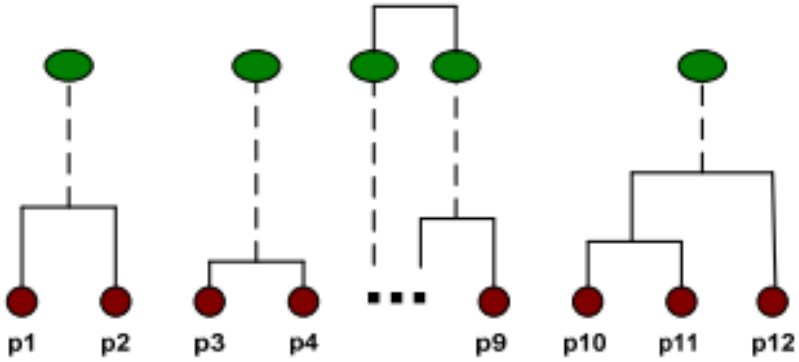


ALGORITMO AGLOMERATIVO

- Depois dessa etapa o algoritmo avalia os clusters mais próximos e realiza a união de clusters.

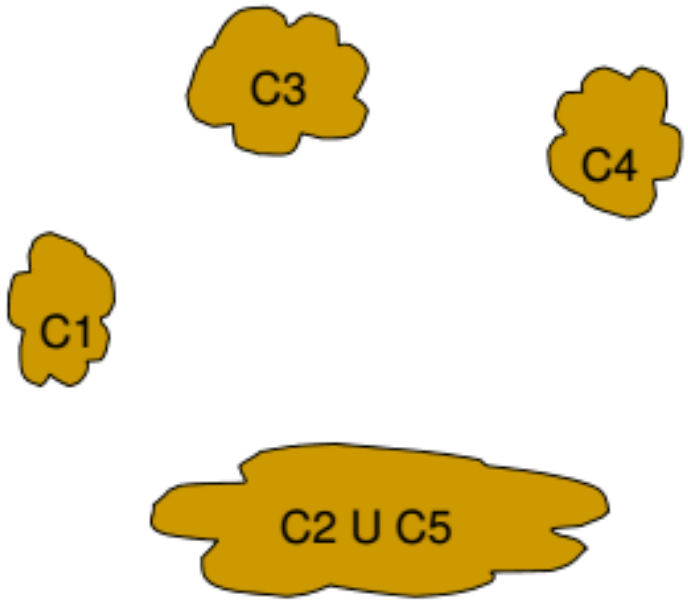






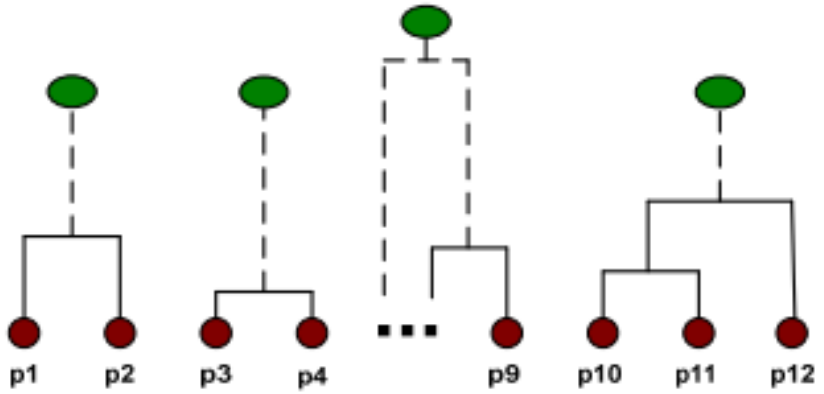
ALGORITMO AGLOMERATIVO

- Depois dessa etapa o algoritmo avalia os clusters mais próximos e realiza a união de clusters.



	C1	C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix

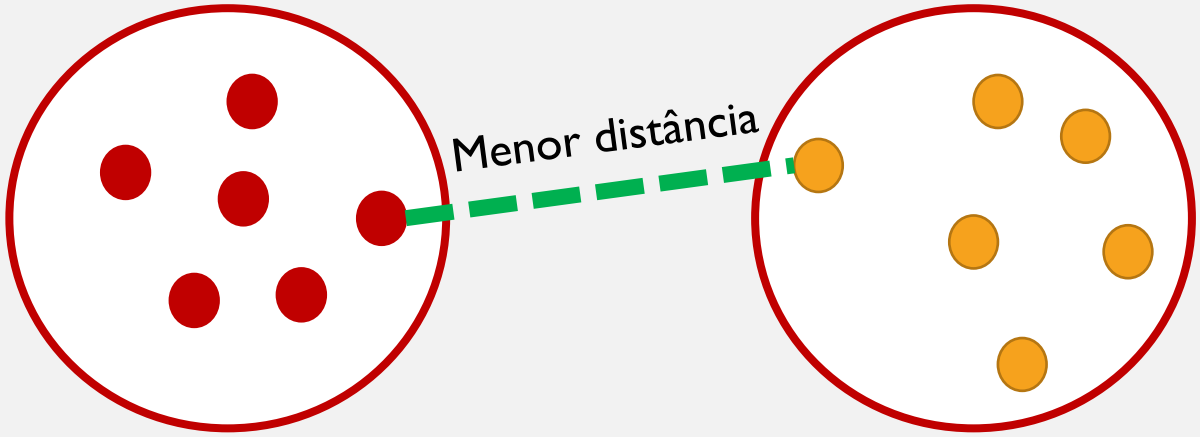


COMO DEFINIR A DISTÂNCIA ENTRE CLUSTERS ?

- MIN
- MAX
- Média dos grupos
- Distância entre centróides

COMO DEFINIR A DISTÂNCIA ENTRE CLUSTERS ?

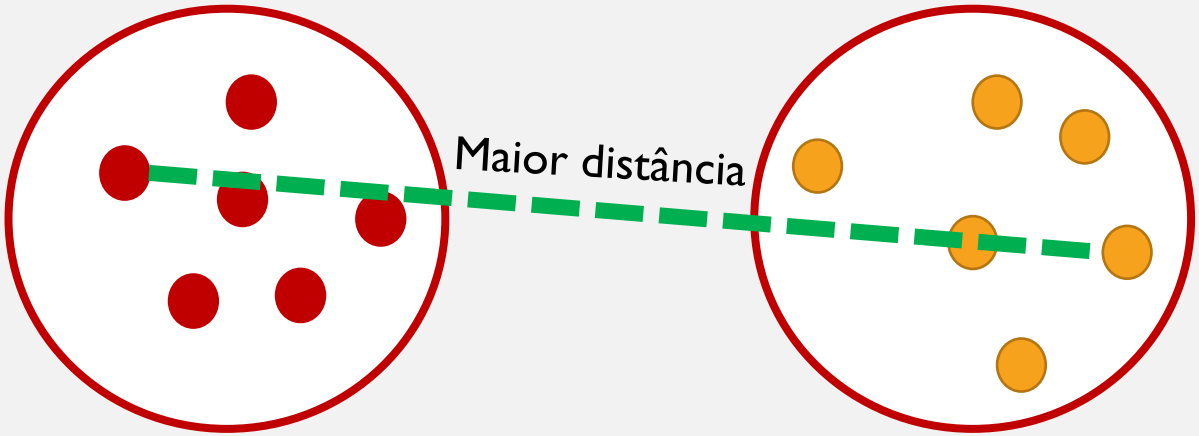
- MIN



- Pode lidar com formatos não elípticos.
- **É sensível a ruídos e outliers.**

COMO DEFINIR A DISTÂNCIA ENTRE CLUSTERS ?

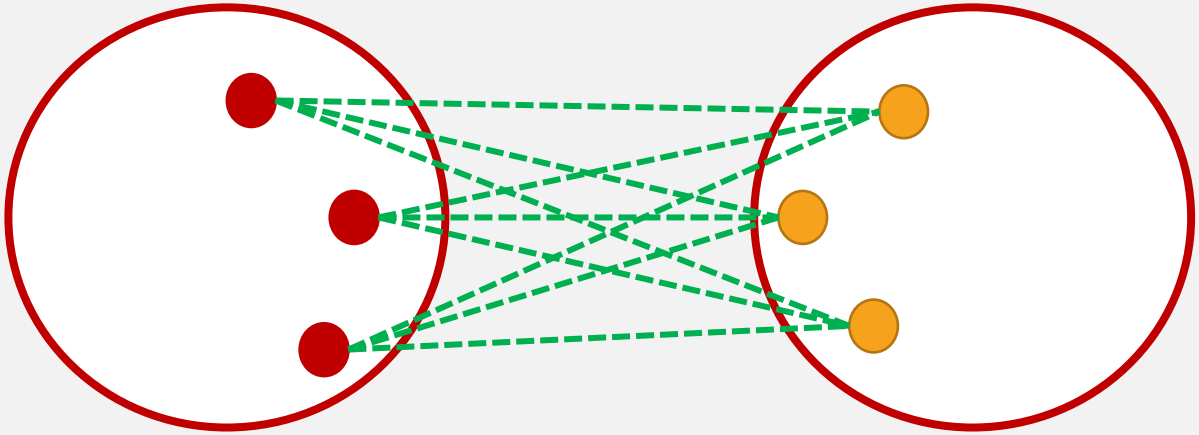
- MAX



- Menos suceptível a ruído e outliers.
- **Tende a quebrar clusters grandes.**
- **Viés para clusters globulares.**

COMO DEFINIR A DISTÂNCIA ENTRE CLUSTERS ?

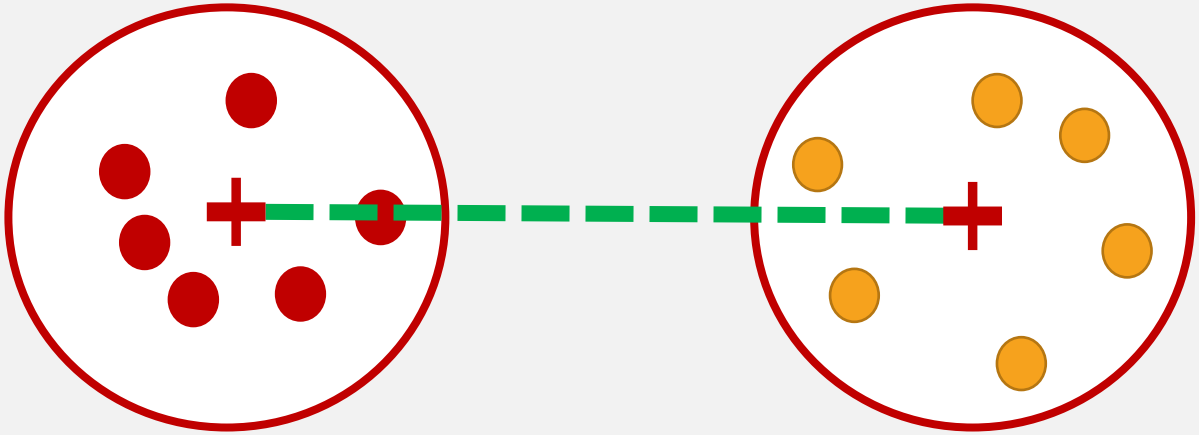
- Média dos grupos



- Menos suceptível a ruído e outliers.
- **Viés para clusters globulares.**

COMO DEFINIR A DISTÂNCIA ENTRE CLUSTERS ?

- Distância entre centróides



LIMITAÇÕES DOS AGRUPAMENTOS HIERÁRQUICOS

- Uma vez tomada a decisão de juntar dois clusters, não pode ser desfeita
- Nenhuma função objetivo objetivo é diretamente diretamente minimizada
- Pode ter sensibilidade a ruído/outliers
- Pode ter dificuldade com clusters de tamanhos diferentes e convexos
- Quebra clusters grandes

FIOCRUZ

MINERAÇÃO DE DADOS

PARA NÃO BIOINFORMATAS

AGRUPAMENTO

RICARDO KHOURI

FELIPE TORRES