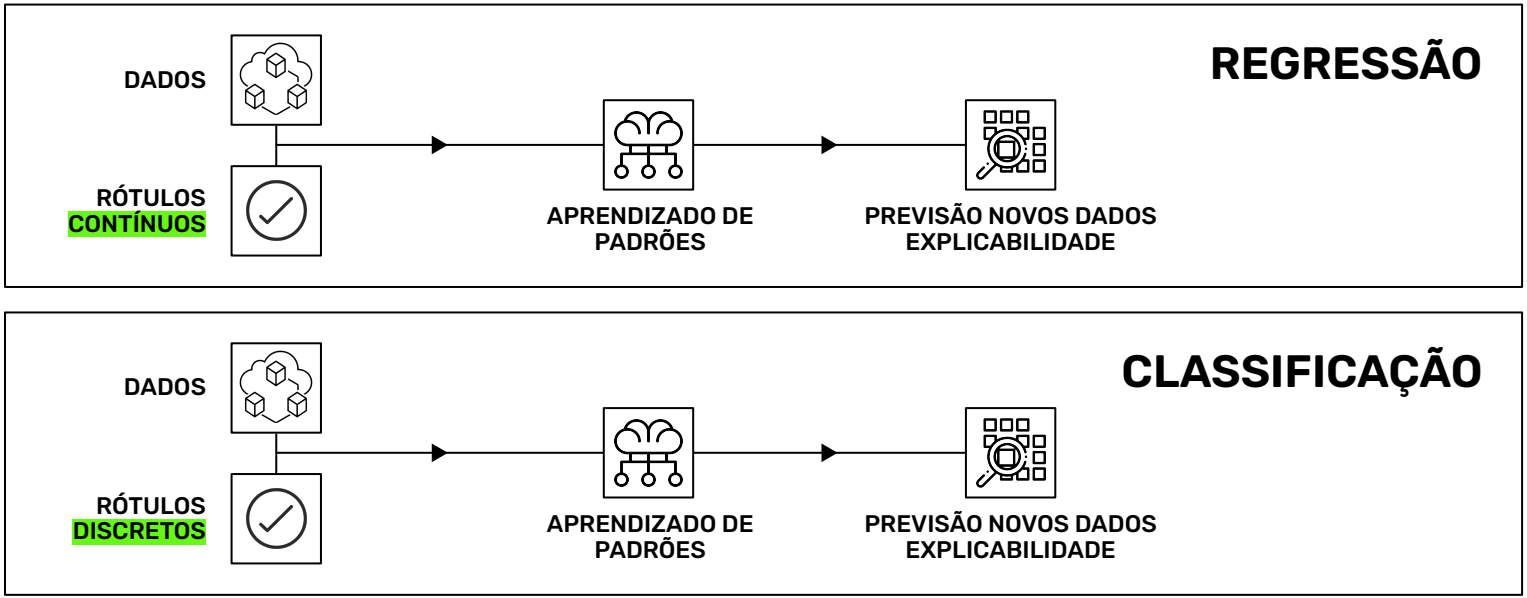
UM POUCO MAIS SOBRE O QUE JÁ SABEMOS



MAS O QUE FAZER QUANDO NÃO SABEMOS O QUE ACERTAR?

PRECISO DE UM EXEMPLO!
PROBLEMA DE RECOMENDAÇÃO DE FILMES

Qual dos filmes indicar para cada usuário?

+ Parasita: drama/ação
+ Coringa: drama/terror
+ Zootopia: fantasia/comédia

NÃO TEMOS MAIS DADOS COM RESPOSTAS DEFINIDAS! O QUE FAZER, ENTÃO?

ID USUÁRIO	NÍVEL DE INTERESSE		
	COMÉDIA	AÇÃO	RECOMENDAÇÃO
1	**1,2**	**3,8**	
2	**1,3**	**3,5**	
3	**4,2**	**1,1**	
4	**3,9**	**1,4**	
5	**1,6**	**3,8**	
6	**1,0**	**3,5**	
7	**1,4**	**3,7**	
8	**4,1**	**1,0**	

PRECISO DE UM EXEMPLO!
PROBLEMA DE RECOMENDAÇÃO DE FILMES

Recomendar cada usuário seria muito trabalhoso!
Como então fazer recomendações massificadas e eficientes?

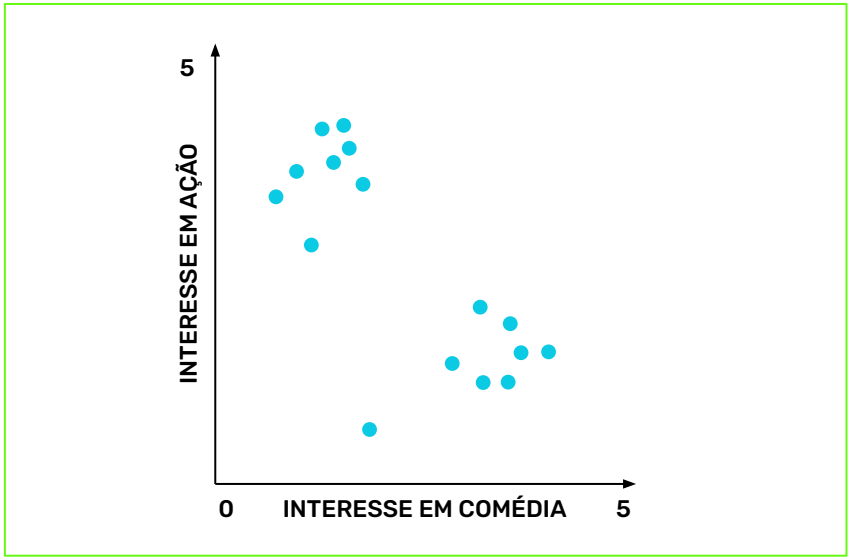
ID USUÁRIO	NÍVEL DE INTERESSE		
	COMÉDIA	AÇÃO	RECOMENDAÇÃO
1	1,2	3,8	Coringa
2	1,3	3,5	
3	4,2	1,1	
4	3,9	1,4	
5	1,6	3,8	
6	1,0	3,5	
7	1,4	3,7	
8	4,1	1,0	

CLUSTERIZAÇÃO (OU AGRUPAMENTO).

MAS QUE TIPO DE AGRUPAMENTO?

PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

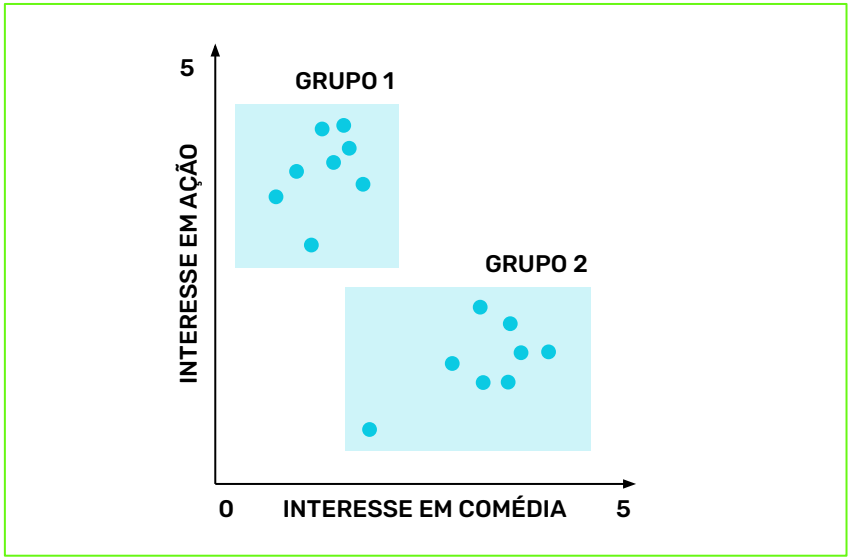
Conseguimos encontrar **grupos** de usuários? Quais?



PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

Conseguimos encontrar **grupos** de usuários? Quais?

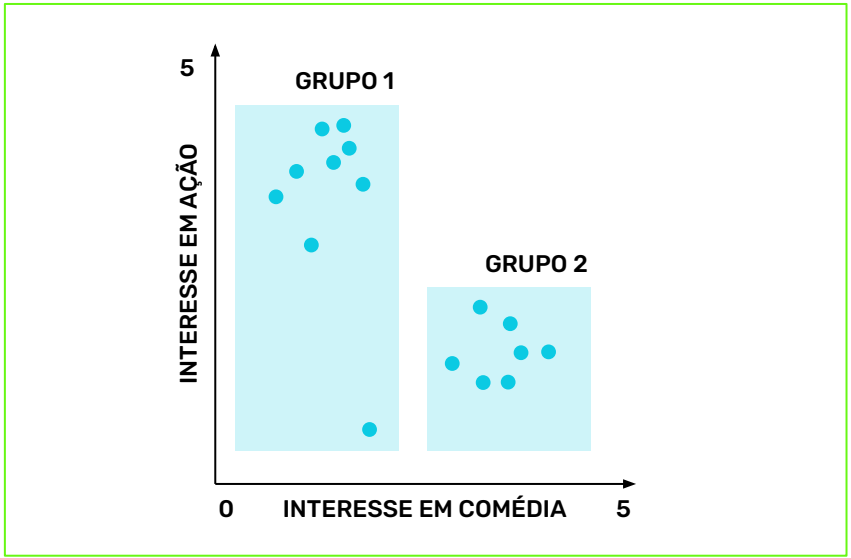
Existe uma **única forma** de agrupamento?



PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

Conseguimos encontrar **grupos** de usuários? Quais?

Existe uma **única forma** de agrupamento?

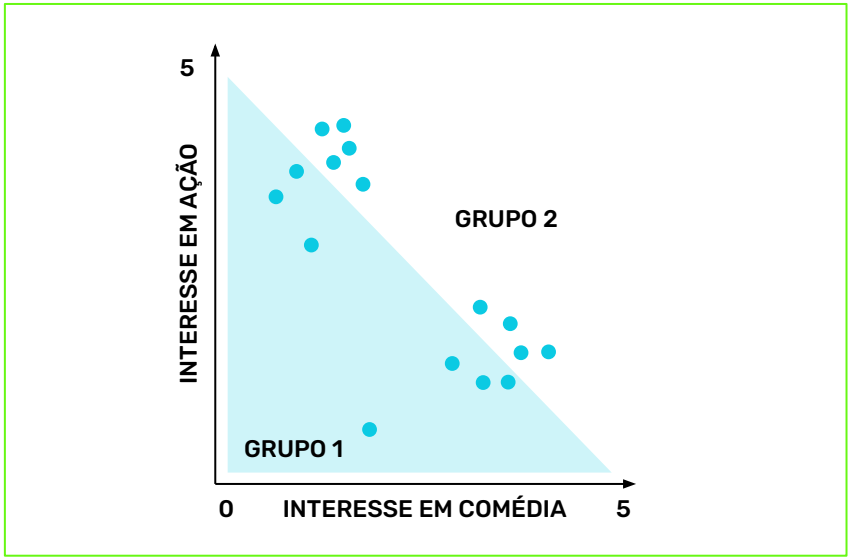


PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

Conseguimos encontrar **grupos** de usuários? Quais?

Existe uma **única forma** de agrupamento?

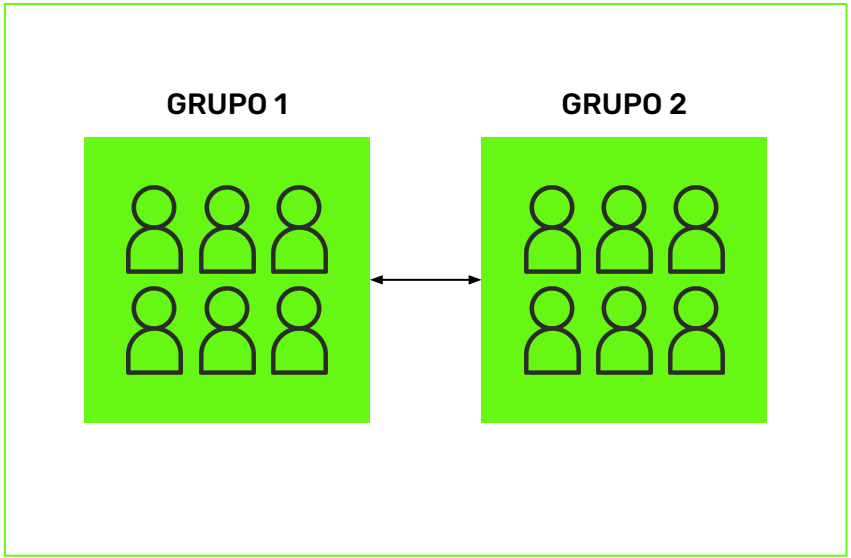
Como avaliar a **qualidade** dos agrupamentos?



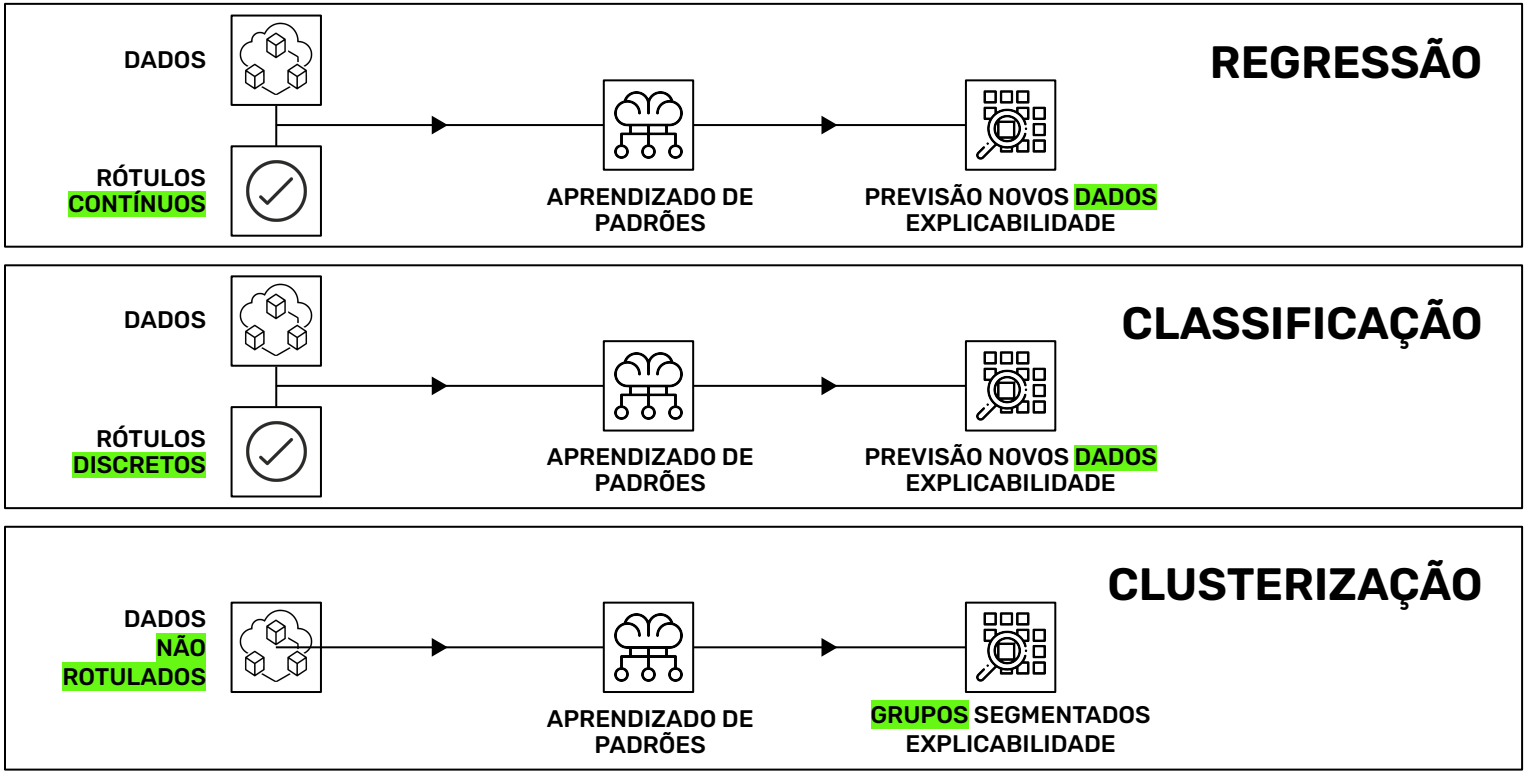
ENTÃO, QUAL É O OBJETIVO DA CLUSTERIZAÇÃO?

Encontrar grupos em que:

+ Elementos de grupos diferentes sejam **muito diferentes** entre si
+ Em um mesmo grupo, os elementos **muito parecidos** entre si



NOSSO ARSENAL DE MODELOS, ATUALIZADO!



APLICAÇÕES INTERESSANTES DE CLUSTERING

DIRECIONAMENTO DE AÇÕES E CAMPANHAS DE MARKETING

Ilustração: unblast.com

CONSTRUÇÃO DE PERSONAS DE CONSUMO DE PRODUTOS

Ilustração: unblast.com

SISTEMAS DE RECOMENDAÇÃO

https://towardsdatascience.com/recommender-system-a1e4595fc0f0
https://towardsdatascience.com/recommender-system-a1e4595fc0f0

CONSTRUINDO A PRIMEIRA CLUSTERIZAÇÃO
RECOMENDAÇÃO DE PRODUTOS ONLINE

Qual dos produtos indicar para cada usuário?

+ Geladeira
+ Sofá
+ Máquina de café expresso

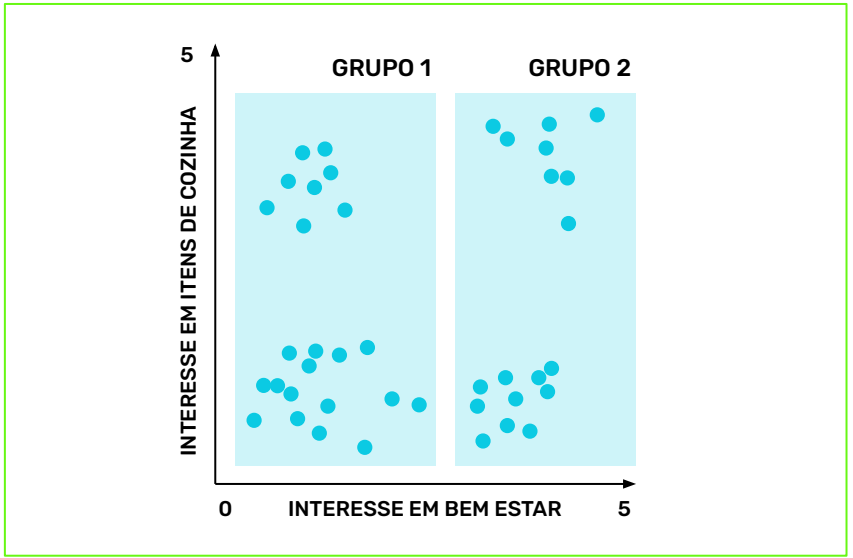
MAIS UMA VEZ NÃO TEMOS RÓTULOS, MAS AGORA SABEMOS COMO ABORDAR O PROBLEMA!

ID USUÁRIO	NÍVEL DE INTERESSE		
	COZINHA	BEM ESTAR	RECOMENDAÇÃO
1	**1,2**	**3,8**	
2	**1,3**	**3,5**	
3	**4,2**	**1,1**	
4	**3,9**	**1,4**	
5	**1,6**	**3,8**	
6	**1,0**	**3,5**	
7	**1,4**	**3,7**	
8	**4,1**	**1,0**	

PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

Conseguimos encontrar **quantos grupos** de usuários?

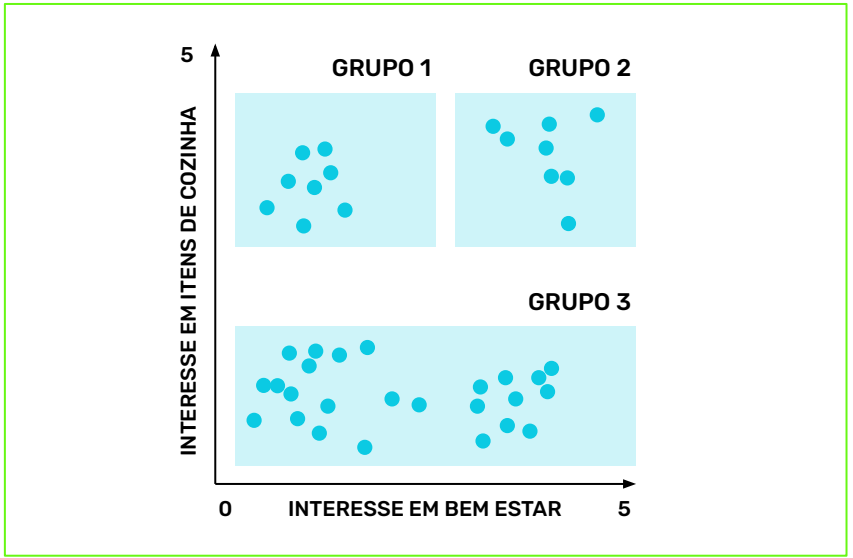
2 grupos?



PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

Conseguimos encontrar
quantos grupos de usuários?

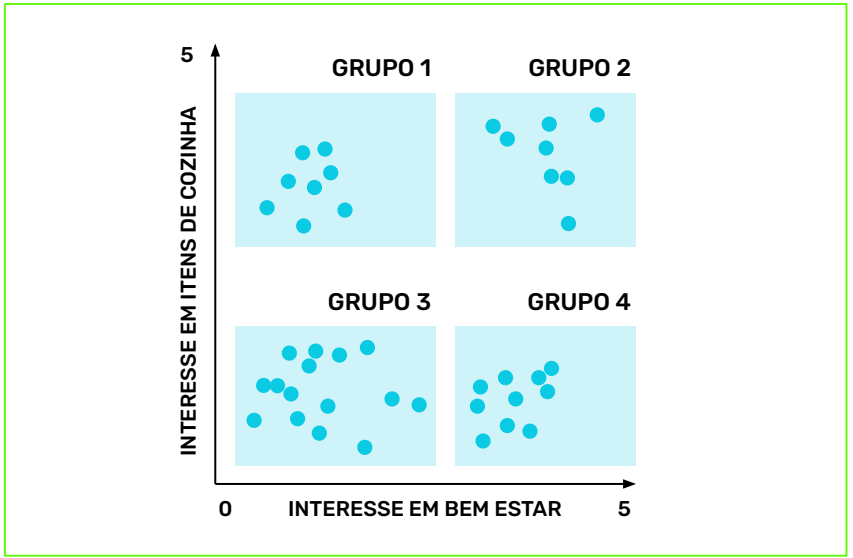
2 grupos?
3 grupos?



PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

Conseguimos encontrar
quantos grupos de usuários?

2 grupos?
3 grupos?
4 grupos?



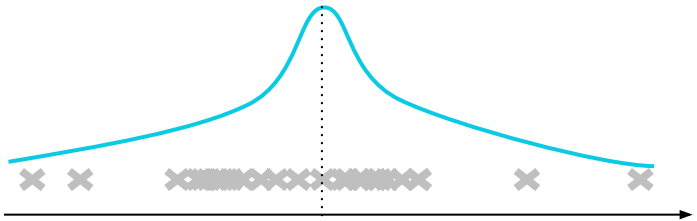
COMO COMPARAR AGRUPAMENTOS? QUAL É O MELHOR, DE FATO?

MEDINDO SEMELHANÇAS E DIFERENÇAS

Qual dos grupos parece ter elementos **mais parecidos,** entre si?

GRUPO 1

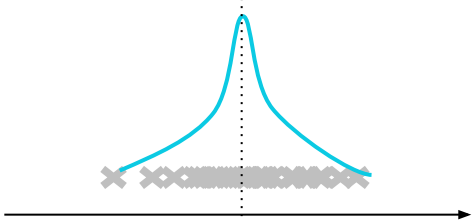
MÉDIA = 2
DESVIO PADRÃO = 1,5



INTERESSE EM ITENS DE BEM ESTAR

GRUPO 2

MÉDIA = 2
DESVIO PADRÃO = 0,5



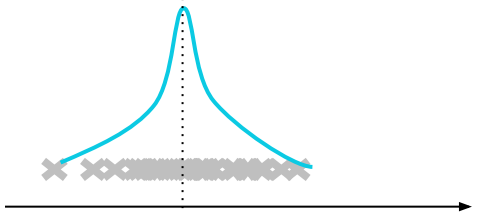
INTERESSE EM ITENS DE BEM ESTAR

MEDINDO SEMELHANÇAS E DIFERENÇAS

Qual dos grupos parece ser mais **diferente** do grupo 2?

GRUPO 2

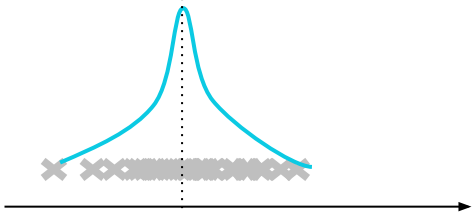
MÉDIA = 2
DESVIO PADRÃO = 0,5



INTERESSE EM ITENS DE BEM ESTAR

GRUPO 3

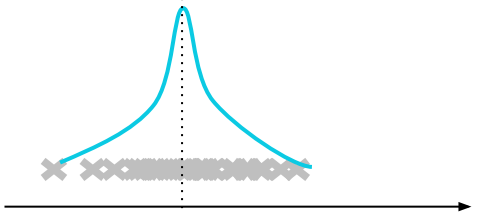
MÉDIA = 3
DESVIO PADRÃO = 0,5



INTERESSE EM ITENS DE BEM ESTAR

GRUPO 4

MÉDIA = 4
DESVIO PADRÃO = 0,5

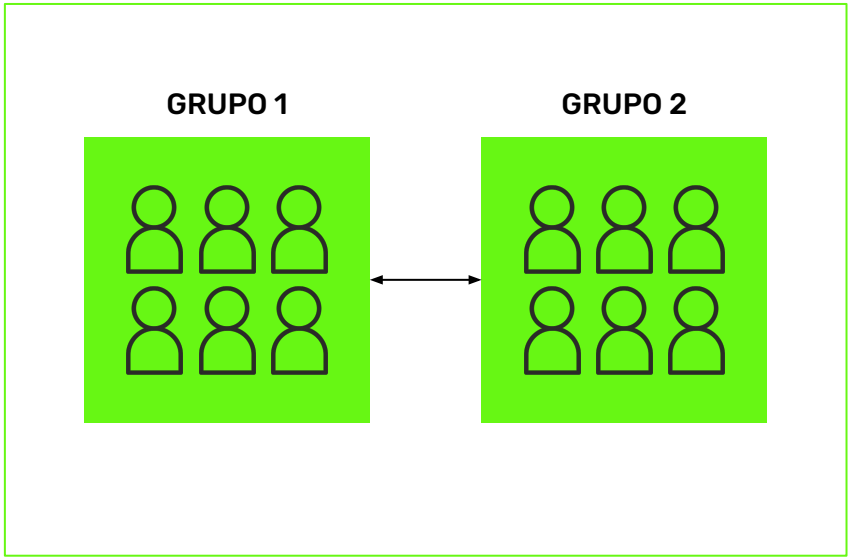


INTERESSE EM ITENS DE BEM ESTAR

ENTÃO, QUAL É O OBJETIVO DA CLUSTERIZAÇÃO?

Encontrar grupos em que:

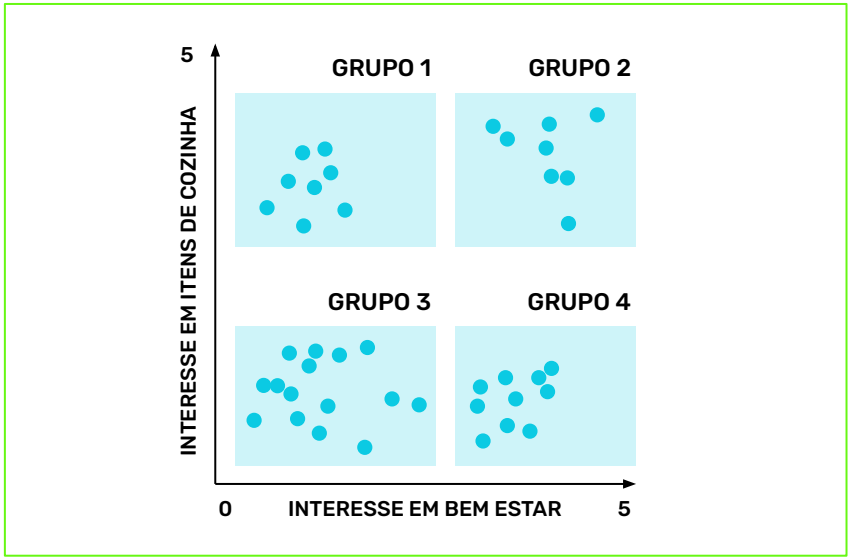
+ Elementos de grupos diferentes sejam **muito diferentes** entre si = grandes diferenças de **médias** dos grupos diferentes
+ Em um mesmo grupo, os elementos **muito parecidos** entre si = **desvio padrão** de elementos do grupo com valores pequenos



PARA AGRUPAR OS DADOS, PRECISAMOS PRIMEIRO VER OS DADOS!

Conseguimos encontrar
quantos grupos de usuários?

2 grupos?
3 grupos?
4 grupos?

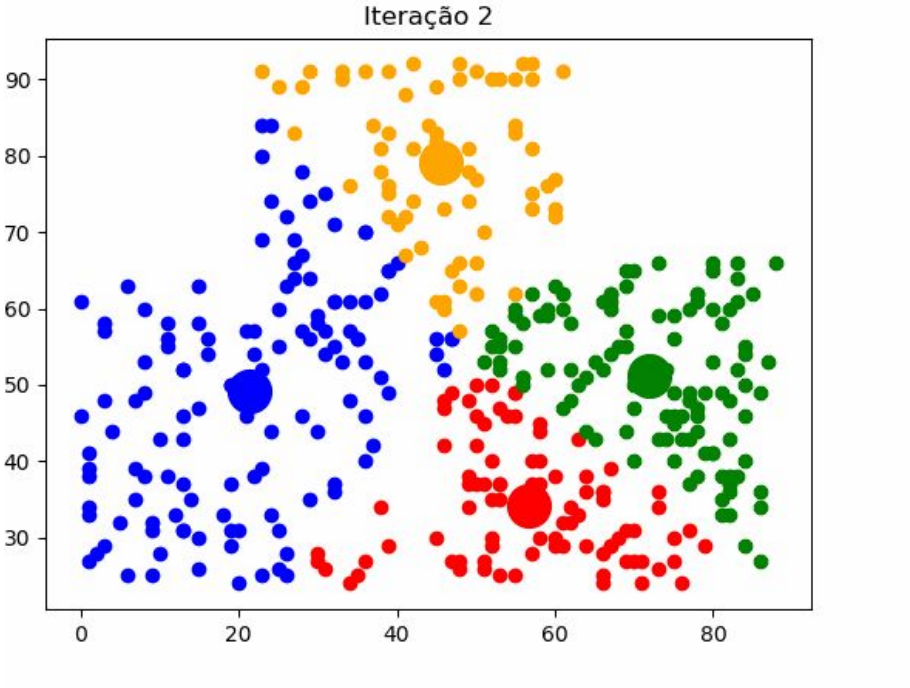


CLUSTERIZAÇÃO É MUITO ÚTIL, MAS TRABALHOSO!

PODEMOS USAR APRENDIZADO AUTOMÁTICO?

SIM! VAMOS ENTENDER A INTUIÇÃO DO K-MEANS

Fonte: https://bioinfo.com.br/inteligencia-artificial-aplicada-a-bioinformatica/



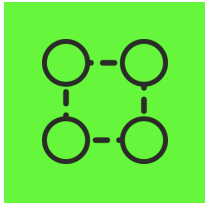
MAS É TUDO AUTOMÁTICO? NÃO!

Mesmo em aprendizado automático, precisamos definir algumas características **antes do aprendizado.** São os chamados **hiperparâmetros.**

PRINCIPAIS HIPERPARÂMETROS DO K-MEANS

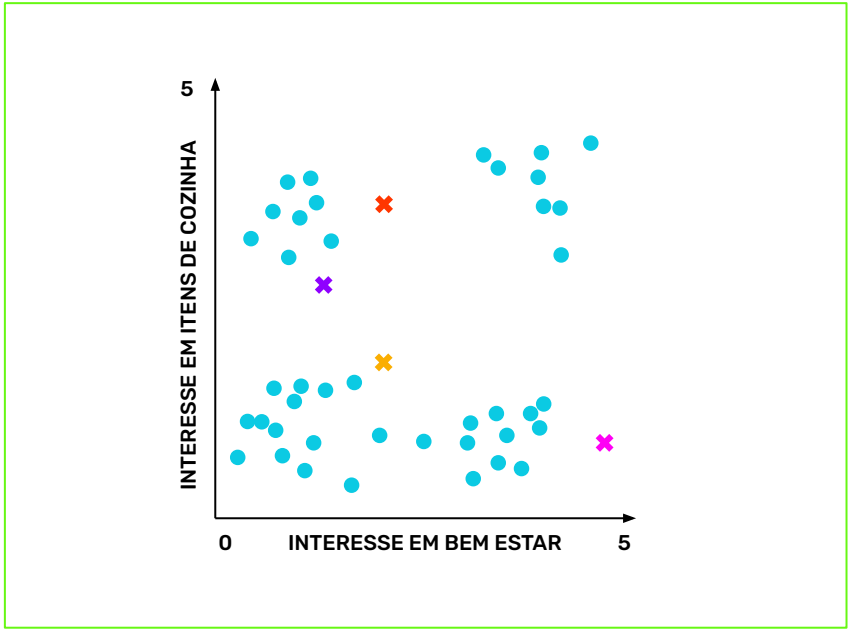
VARIÁVEIS CONSIDERADAS (FEATURES)

NÚMERO DE CLUSTERS (GRUPOS)



OK, MAS COMO FUNCIONA, ENTÃO?

+ Em um gráfico dos dados, inicializamos aleatoriamente 4 pontos, chamados **Centróides**

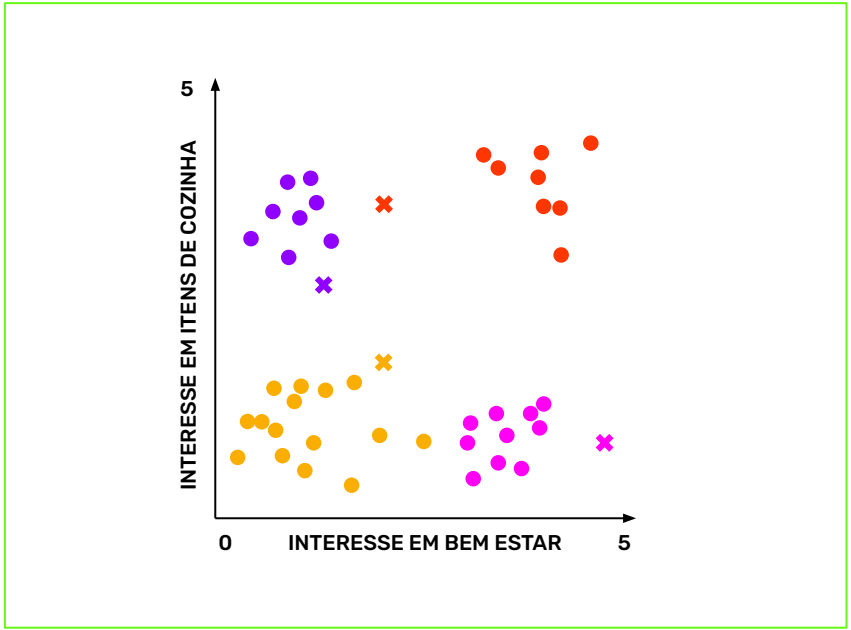


INTERESSE EM ITENS DE COZINHA E DE BEM ESTAR

4 CLUSTERS

OK, MAS COMO FUNCIONA, ENTÃO?

+ Em um gráfico dos dados, inicializamos aleatoriamente 4 pontos, chamados **Centróides**
+ Atribuímos cada dado ao centróide **mais próximo**

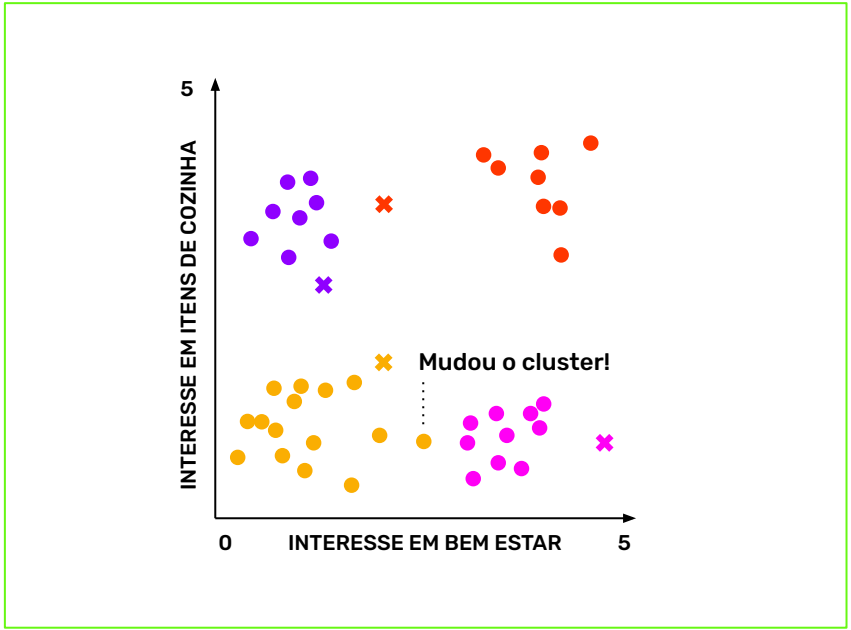


INTERESSE EM ITENS DE COZINHA E DE BEM ESTAR

4 CLUSTERS

OK, MAS COMO FUNCIONA, ENTÃO?

+ Em um gráfico dos dados, inicializamos aleatoriamente 4 pontos, chamados **Centróides**
+ Atribuímos cada dado ao centróide **mais próximo**
+ **Atualizamos** a posição de cada centróide para a **média (means)** dos pontos atribuídos
+ **Repetimos** os passos anteriores até que os centróides atinjam um ponto de equilíbrio



INTERESSE EM ITENS DE COZINHA E DE BEM ESTAR

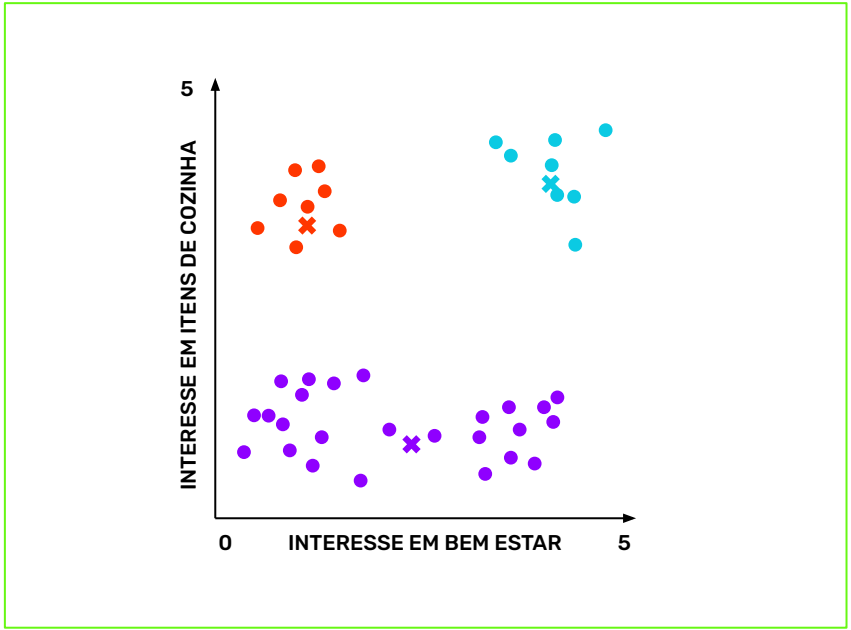
4 CLUSTERS

COMO QUALQUER ALGORITMO, O KMEANS EXIGE ALGUNS CUIDADOS!

giphy.com/gifs

CUIDADOS COM O K-MEANS

+ Não há um **número ótimo** de clusters

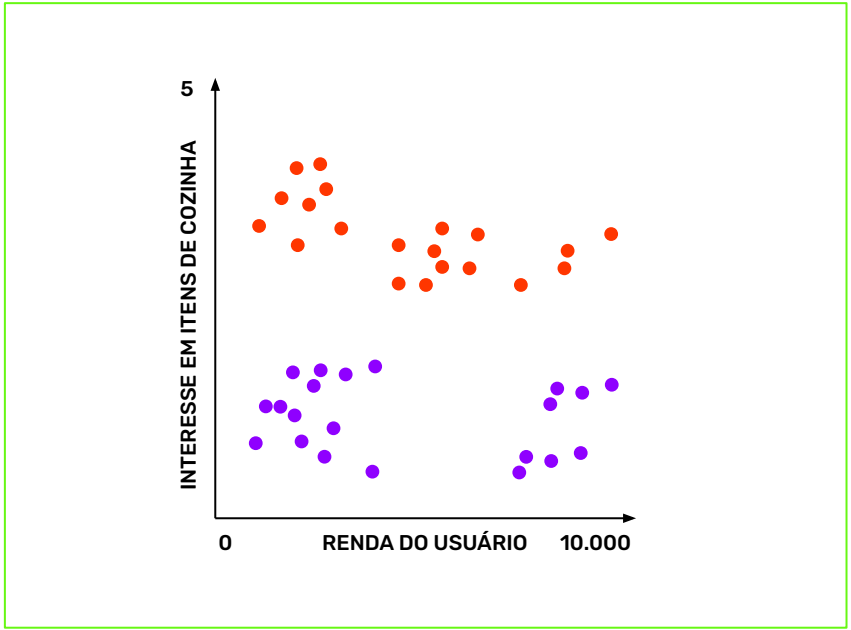


INTERESSE EM ITENS DE COZINHA E DE BEM ESTAR

4 CLUSTERS

CUIDADOS COM O K-MEANS

+ Não há um **número ótimo** de clusters
+ Cuidado com as **ordens de grandeza** das variáveis!



INTERESSE EM ITENS DE COZINHA E DE BEM ESTAR

4 CLUSTERS

E É POSSÍVEL AVALIAR A QUALIDADE DOS CLUSTERS?

tenor.com

INÉRCIA
Distância média entre os pontos de um cluster e seu centróide (quanto menor, melhor)

SILHUETA
Comparação de similaridades dos pontos com seus clusters e com outros clusters (quanto maior, melhor)