

Métodos Matriciais e Análise de Clusters

Agrupamentos

Segmentação de Clientes MetLife (seguradora)

Abordagem Tradicional

- Uso de **conhecimentos demográficos**
- **Idade** era usada para segmentar clientes e pacotes de serviços
- Para empresas, **número de funcionários** era o índice de maturidade

Estudo de Segmentação

- Iniciou em 2015 e durou um ano
- Reposicionamento de marca
- Pesquisas feitas abordando **diversos aspectos** (demográficos, “*firmográficos*”, comportamentais e necessidades)
- Mudança de estratégia de segmentação:
 - Basear em características **comportamentais**, em vez da idade



The Segments

The resulting five segments proved attitudinally differentiated and demographically distinct.



	YOUNG ACHIEVERS	CONCERNED MOMS	FINANCIALLY MATURE	HO HUM	SOLO CONTENT
	Young Achievers	Concerned Moms	Financially Mature	Ho Hum	Solo Content
Demographics	Younger Skews male	Young, Middle Age Mostly female	Mature Skews male	Middle Age Mostly female	Mature Male and Female
Attitudes	Early adopters, technical Driven, Risk taker Price sensitive	Use social media, but not otherwise technical Don't know where to begin Price sensitive	Recognize value of insurance Confident about financial matters Least price sensitive	Late adopters Risk averse Not primary decision makers and not thinking about LI	Use social media Mistrustful of financial inst. Least interest in LI

Two segments are primary targets for the Direct Business.

	 YOUNG ACHIEVERS	 CONCERNED MOMS	 FINANCIALLY MATURE	 HO HUM	 SOLO CONTENT
	Young Achievers	Concerned Moms	Financially Mature	Ho Hum	Solo Content
% of US	20%	20%	30%	20%	10%
% of MetLife	50%	30%	10%	5%	5%
Lapse Rate	Low	High	Low	Medium	Medium
Value	High	Medium	Medium	Low	Low

Target

- Large portion of market
- Right for business model

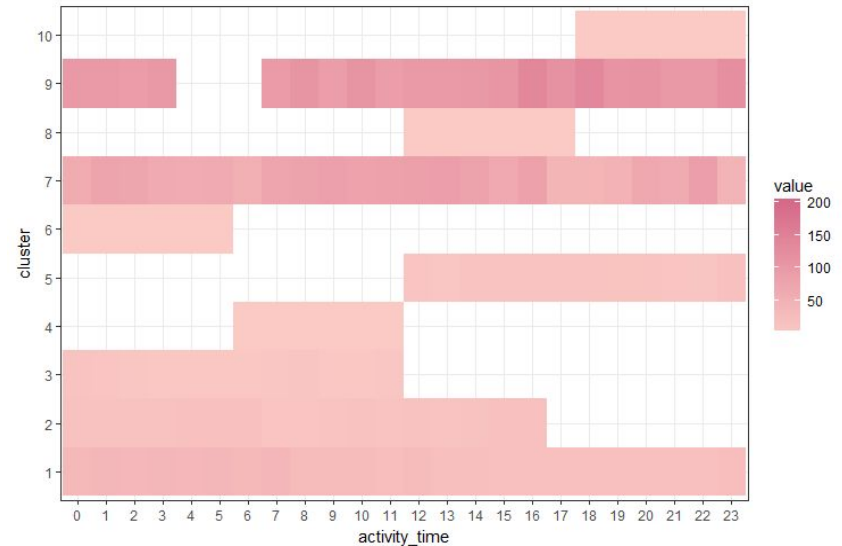
Minimize Cost to Serve

- Prefer face to face
- Low conversion
- Lower value



Segmentação de Clientes Telecom

- Dados da **Telecom Italia em Milão**
 - **SMS**, chamadas e tráfego de dados
- Dados: CDR (Call Detail Records)
 - Horário de início e fim da ligação
 - Terminal de início e fim
 - Região



Identificação de Assuntos em Notícias (Twist)

Caso Rock in Rio 2015



- **Processamento de Linguagem Nat (NLP)**
- Similaridade **semântica** entre textos
- Identificação de **assuntos** sobre um determinado tema



rihanna, cantora, instagram, hotel, riri, fasano, sábado
Last news on: 05/10/2015 19:05:01



brt, terminal, alvorada, ônibus, embarque, consórcio, transporte
Last news on: 30/09/2015 08:53:02



Confira as linhas de ônibus regulares que passam mais próximas à Cidade do Rock

Após longas filas, sistema de ônibus BRT no Rock in Rio deve ter mudanças

Saiba como chegar à Cidade do Rock

Clique no tópico e descubra mais sobre o assunto



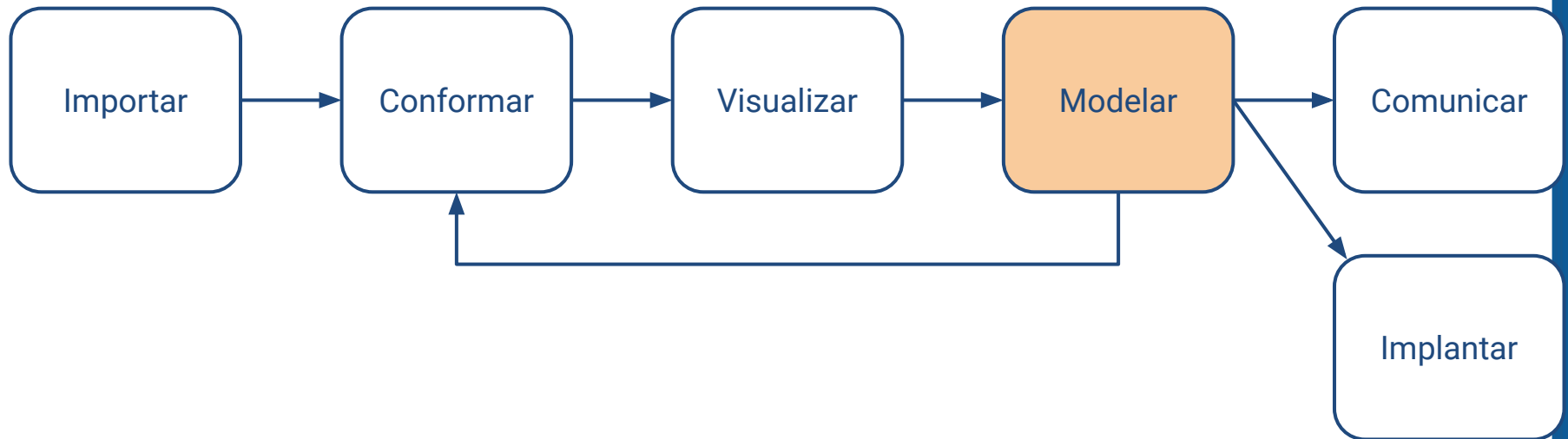
lambert, queen, adam, freddie, brian, mercury, may
Last news on: 29/09/2015 20:32:55



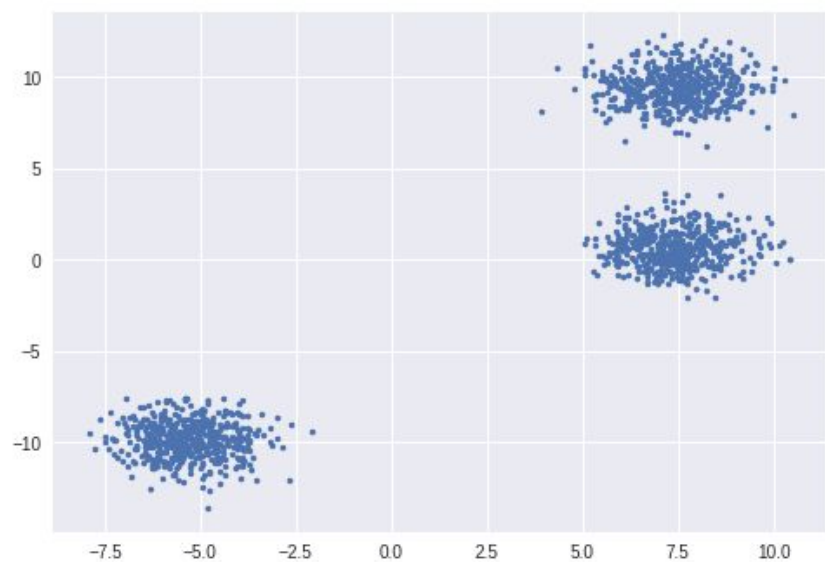
bruna, marquezine, maurício, destri, beijos, paraisópolis, romance
Last news on: 05/10/2015 16:21:15



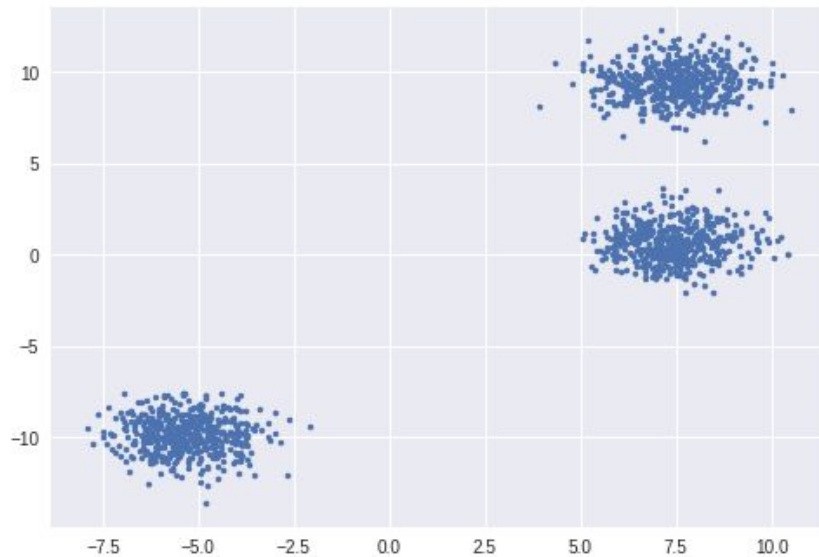
Processo de Ciência de Dados



Clusterização



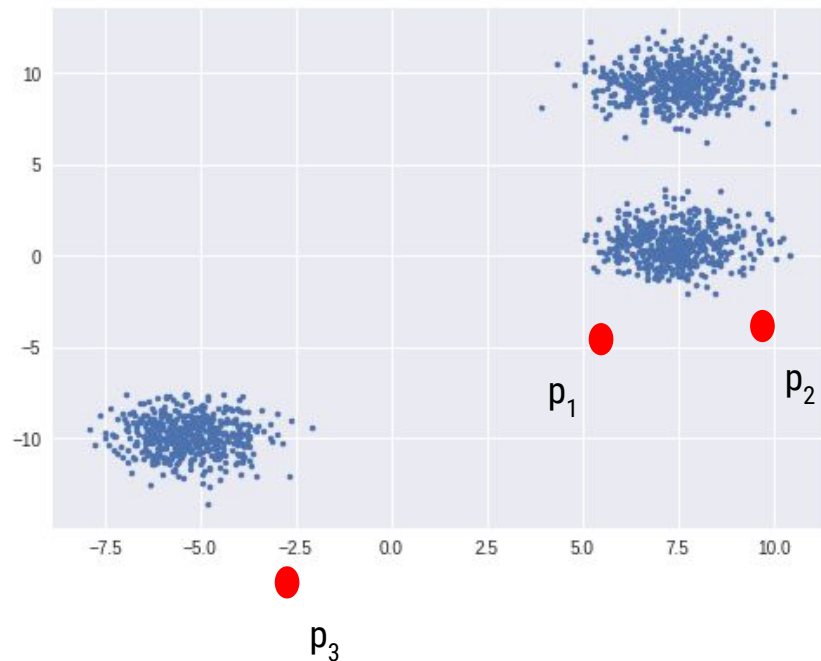
Clusterização



E Clusters em mais que 3 dimensões?

	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
Transação 1	0	1	0	1	1	0	0
Transação 2	1	0	1	1	1	0	0
Transação 3	0	1	0	1	1	0	0
Transação 4	1	1	0	1	1	0	0
Transação 5	0	0	1	0	0	0	0
Transação 6	0	0	0	0	1	0	0
Transação 7	0	0	0	1	0	0	0

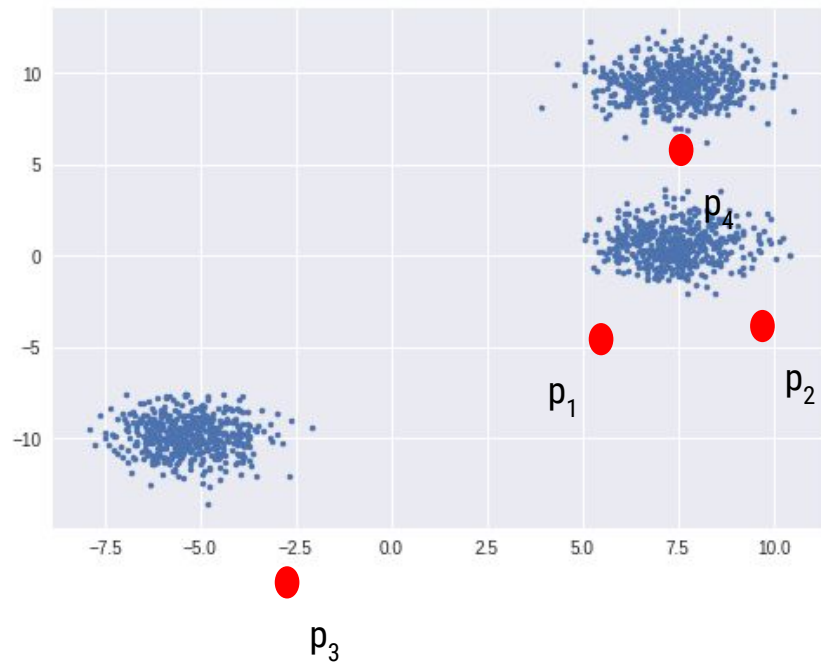
Clusterização



E Clusters em mais que 3 dimensões?

	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
Transação 1	0	1	0	1	1	0	0
Transação 2	1	0	1	1	1	0	0
Transação 3	0	1	0	1	1	0	0
Transação 4	1	1	0	1	1	0	0
Transação 5	0	0	1	0	0	0	0
Transação 6	0	0	0	0	1	0	0
Transação 7	0	0	0	1	0	0	0

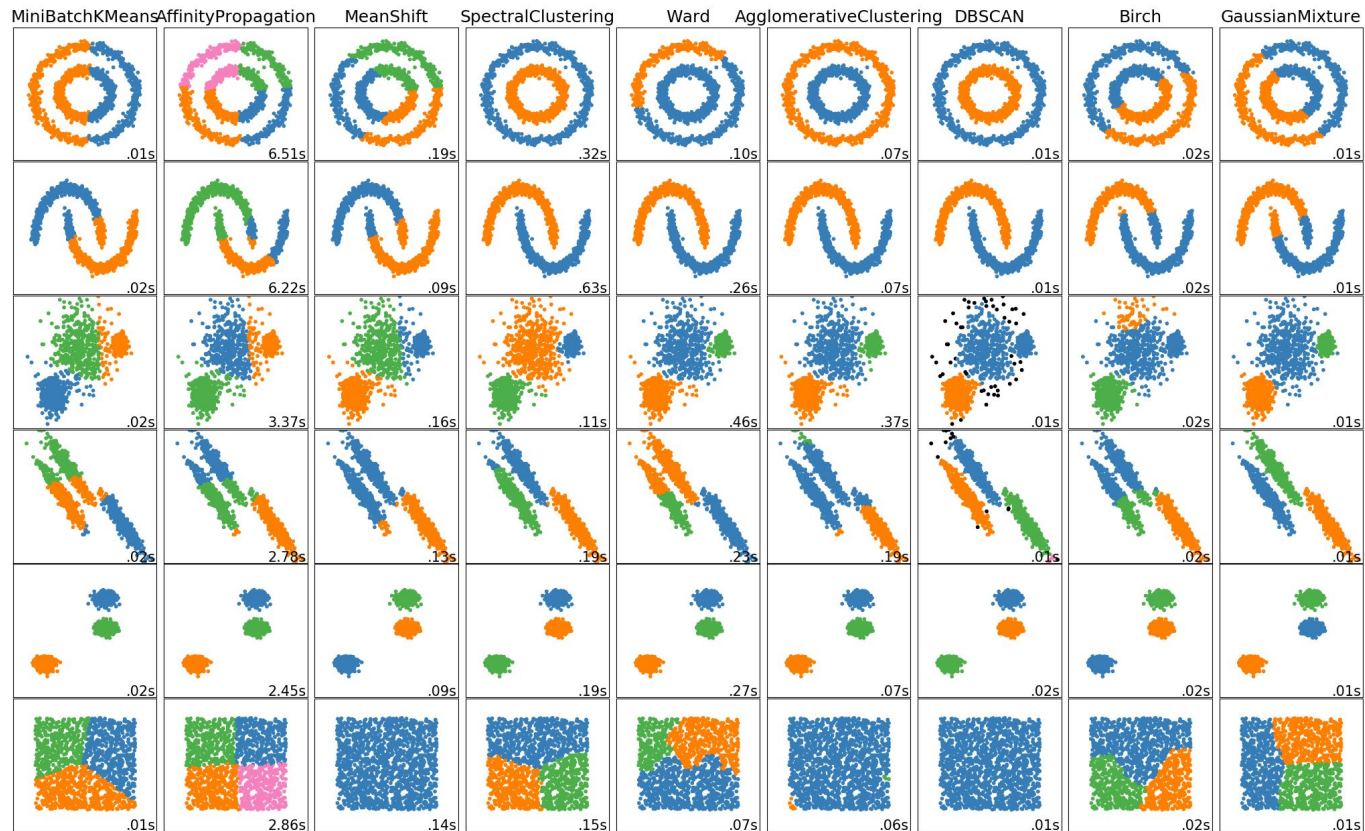
Clusterização



E Clusters em mais que 3 dimensões?

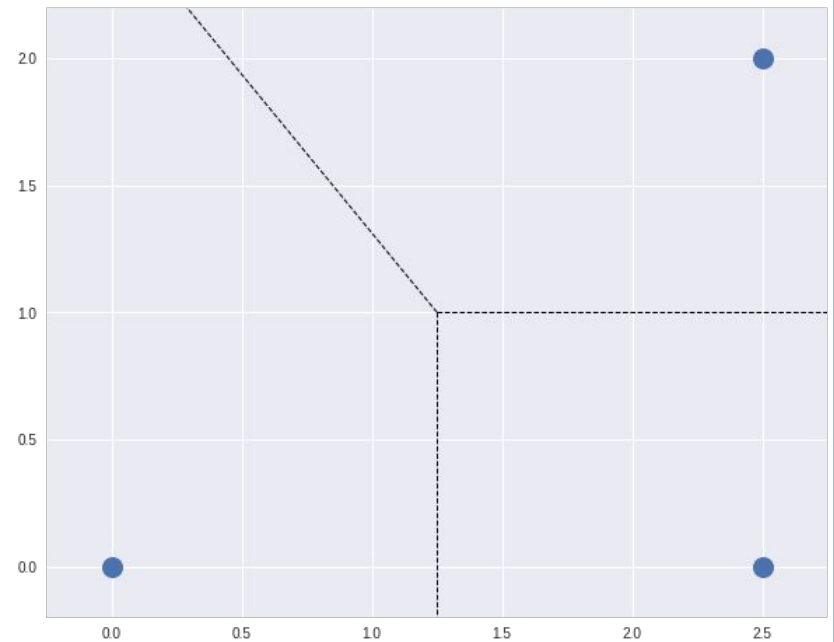
	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
Transação 1	0	1	0	1	1	0	0
Transação 2	1	0	1	1	1	0	0
Transação 3	0	1	0	1	1	0	0
Transação 4	1	1	0	1	1	0	0
Transação 5	0	0	1	0	0	0	0
Transação 6	0	0	0	0	1	0	0
Transação 7	0	0	0	1	0	0	0

Efeito de diferentes algoritmos



K-means

- Divide o espaço em **k partições**
- Clusters representados por **centróides**
- Um ponto pertence ao cluster do **centróide mais próximo**

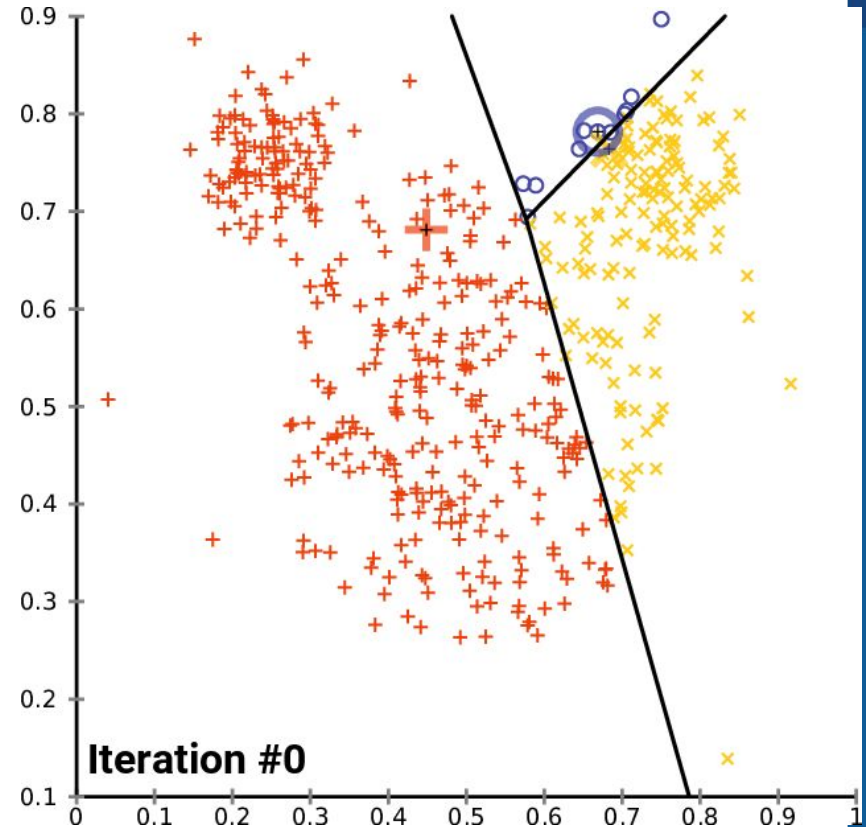


Algoritmo k-means

1. Inicializar k centróides em pontos aleatórios
2. Para cada ponto, encontrar qual o centróide mais próximo
3. Calcular o baricentro dos pontos para cada centroide
4. Mover o centróide na direção do seu baricentro
5. Repetir a partir de 2.

O algoritmo converge quando o movimento for menor que um valor pré-definido

Visualização Interativa

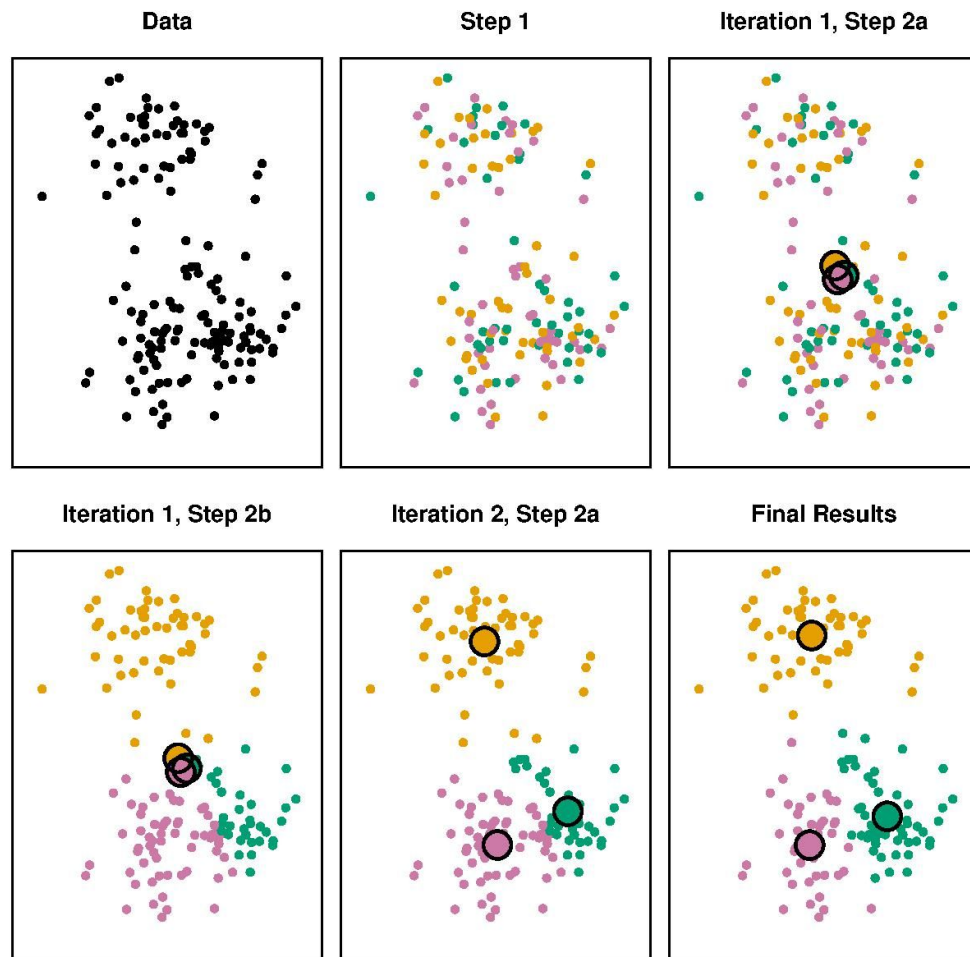


Algoritmo k-means

1. Inicializar k centróides em pontos aleatórios
2. Para cada ponto, encontrar qual o centróide mais próximo
3. Calcular o baricentro dos pontos para cada centroide
4. Mover o centróide na direção do seu baricentro
5. Repetir a partir de 2.

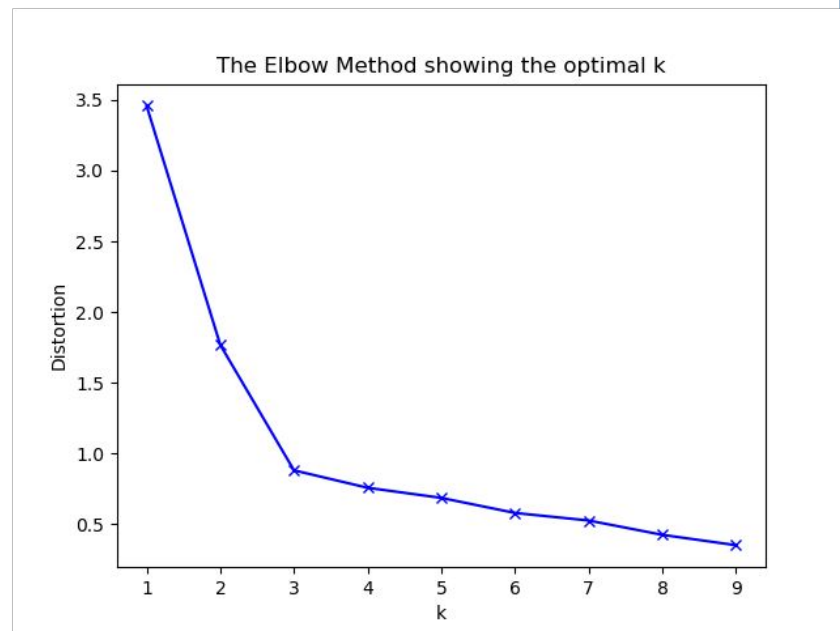
O algoritmo converge quando o movimento for menor que um valor pré-definido

[Visualização Interativa](#)



Algoritmo k-means. Qual k?

- Observar somatório das distâncias de cada ponto ao centróide mais próximo
- Usar k a partir do qual a diferença é pequena.



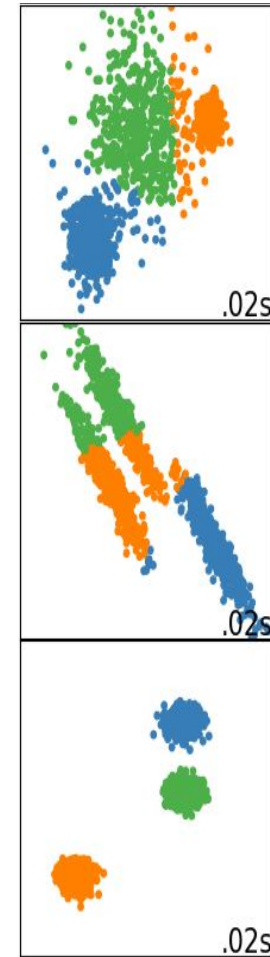
Considerações Práticas para K-means

Vantagens

- Algoritmo **simples** e **rápido**
- Pode ser **paralelizado**
- Amplamente conhecido e utilizado

Pontos de Atenção

- Usa **inicialização aleatória**
- É suscetível a **mínimos locais**
- É dependente da **topologia** dos dados
- É sensível a **outliers**



Considerações Práticas para K-means



Considerações Práticas para K-means

Vantagens

- Algoritmo **simples** e **rápido**
- Pode ser **paralelizado**
- Amplamente conhecido e utilizado

Pontos de Atenção

- Usa **inicialização aleatória**
- É suscetível a **mínimos locais**
- É dependente da **topologia** dos dados
- É sensível a **outliers**

O que fazer ao aplicar

- Testar **múltiplas inicializações diferentes**. De preferência, na ordem de centenas.
- **Inicializar explicitamente** o gerador de números aleatórios para **reproducibilidade**
- Pode se beneficiar de **normalizações**
 - Variância Unitária
 - PCA
- Cuidado extra com **outliers** durante a limpeza dos dados! Desconfiar de clusters com **poucos** pontos

PRNG: Pseudo-Random Number Generator

- Um PRNG é um algoritmo **determinístico** que gera sequências de números com propriedades de **números aleatórios**
- Inicializado com um **seed**
- Para um mesmo seed, **sempre** gera a **mesma sequência**
- Use o PRNG **ao seu favor**:
 - Inicializar **explicitamente** antes de rodar um algoritmo que use números aleatórios
 - Inicializar com **seeds diferentes** a cada tentativa

Sem fixar o seed

```
> runif(3)
[1] 0.9333853 0.0210864 0.3993138
> runif(3)
[1] 0.83800443 0.47204073 0.03746961
> runif(3)
[1] 0.3844462 0.1631746 0.6489406
```

Fixando o seed

```
> set.seed(12345)
> runif(3)
[1] 0.7209039 0.8757732 0.7609823
> set.seed(12345)
> runif(3)
[1] 0.7209039 0.8757732 0.7609823
> set.seed(12345)
> runif(3)
[1] 0.7209039 0.8757732 0.7609823
```

K-means no R

kmeans(

x,

Dados a serem clusterizados

centers,

Número de centróides (k) ou posições iniciais

iter.max = 10,

Número máximo de iterações

nstart = 1,

Número de inicializações

algorithm =

Variante do algoritmo

"Hartigan-Wong",

trace = FALSE

Salvar métricas para depuração

)

K-means no R

Dados **numéricos contínuos**,
uma linha por observação

kmeans(

x,

Dados a serem clusterizados

centers,

Número de centróides (k) ou
posições iniciais

iter.max = 10,

Número máximo de iterações

nstart = 1,

Número de inicializações

algorithm =

Variante do algoritmo

"Hartigan-Wong",

trace = FALSE

Salvar métricas para depuração

)

Imagem como uma nuvem de pontos

Imagem original (21,003 cores)

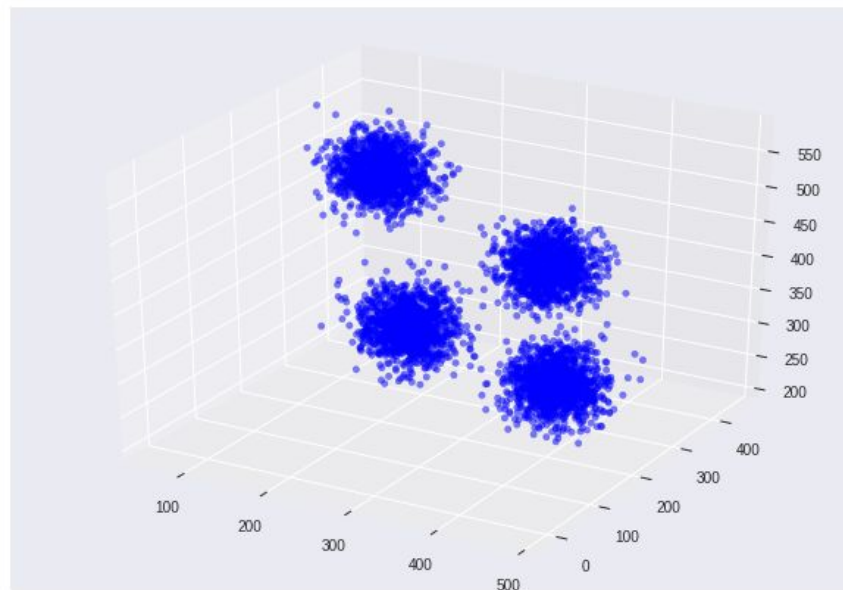
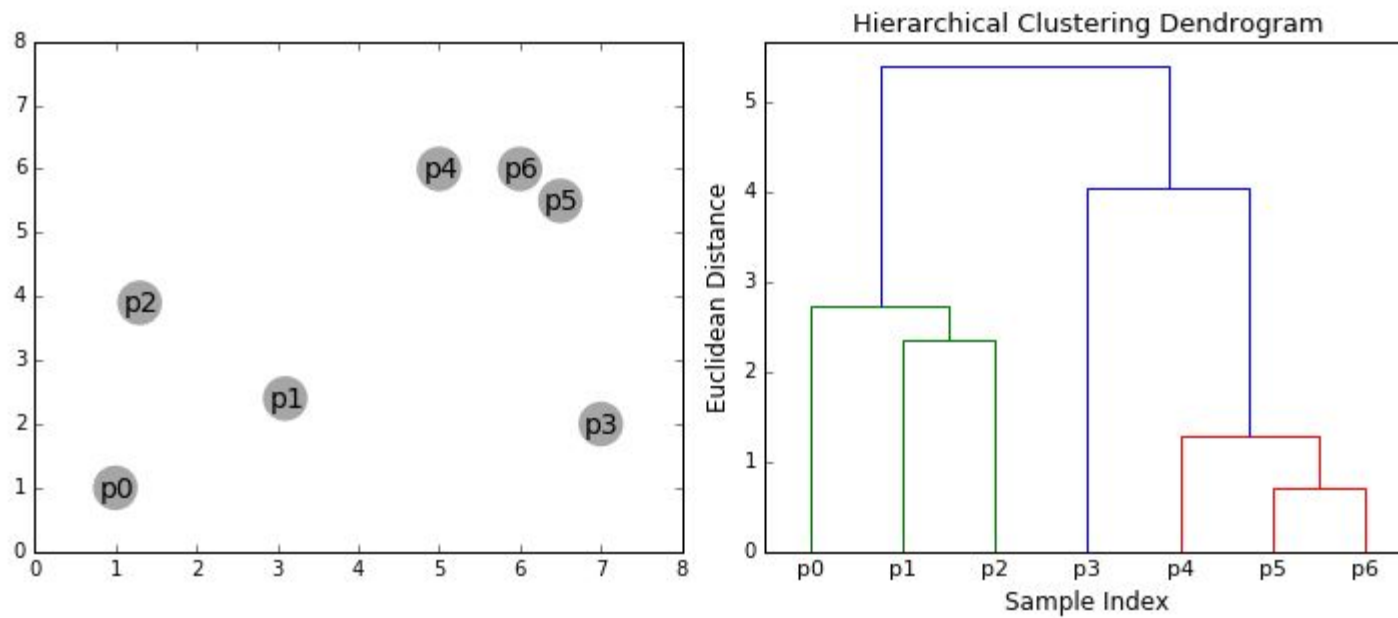
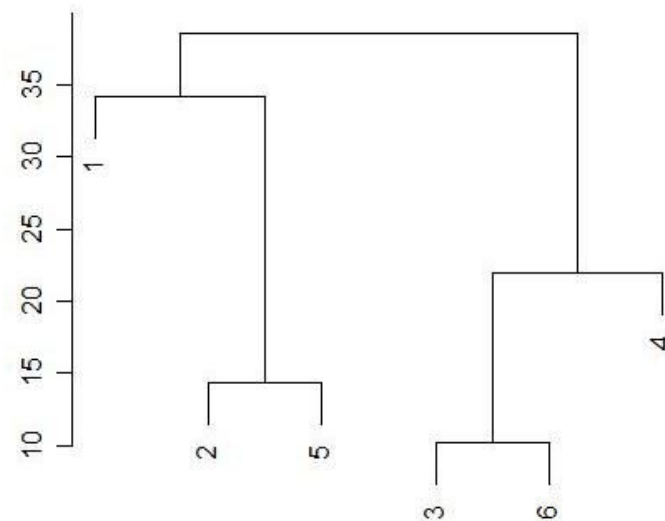
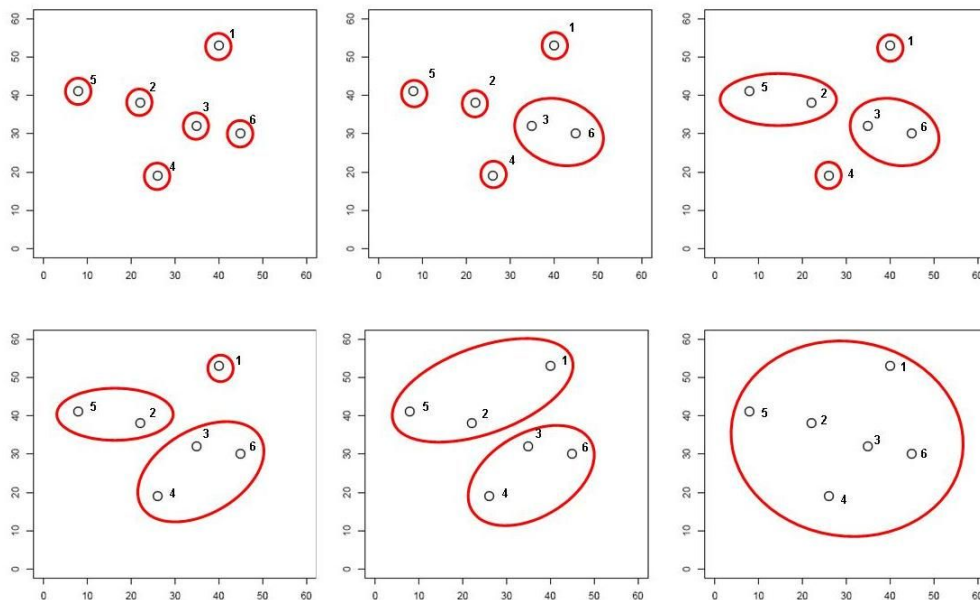


Imagem alterada (4 cores, K-Means)

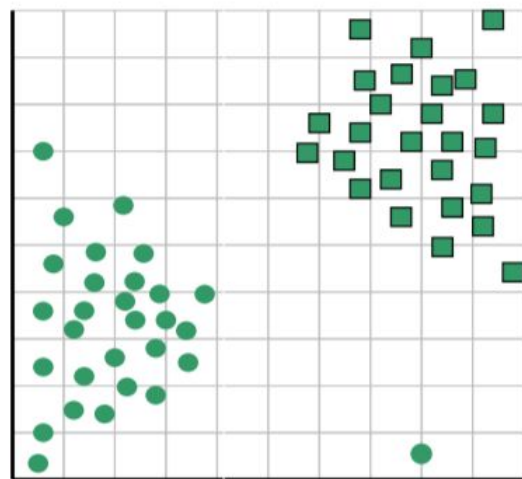




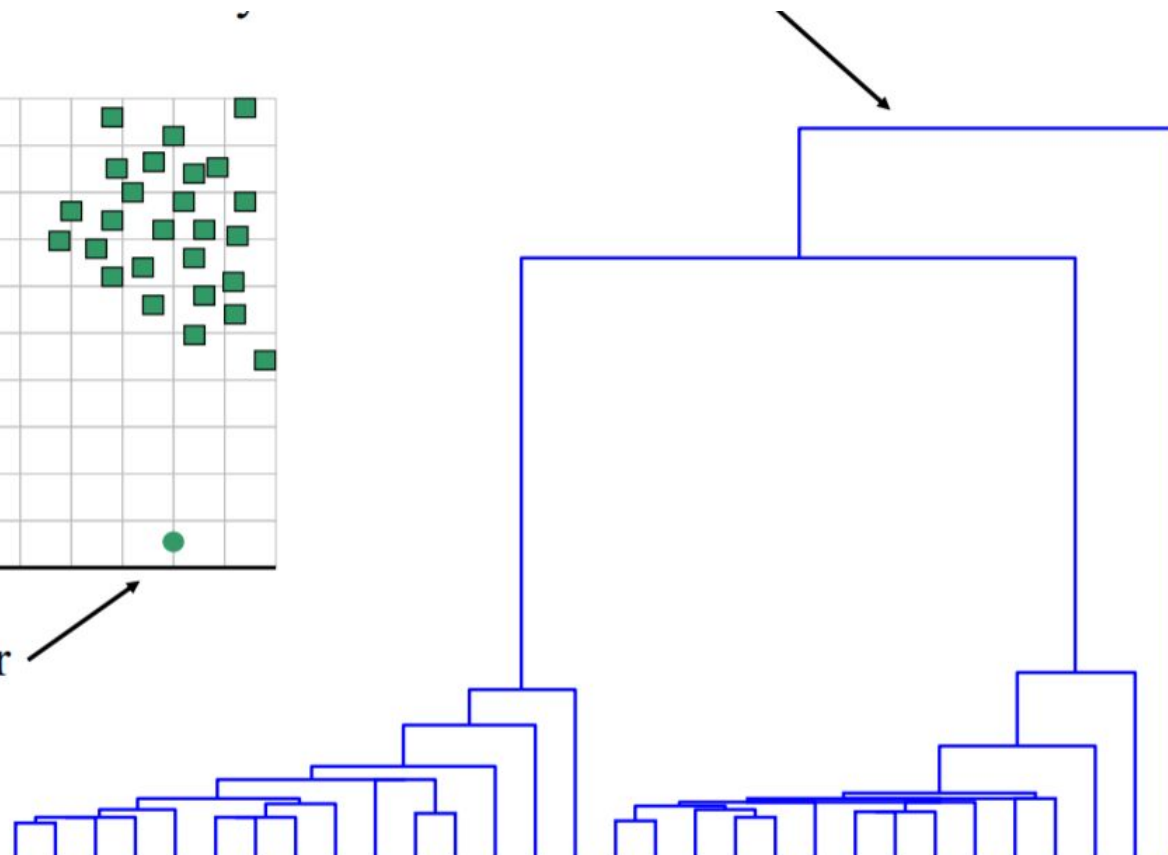
Conectividade



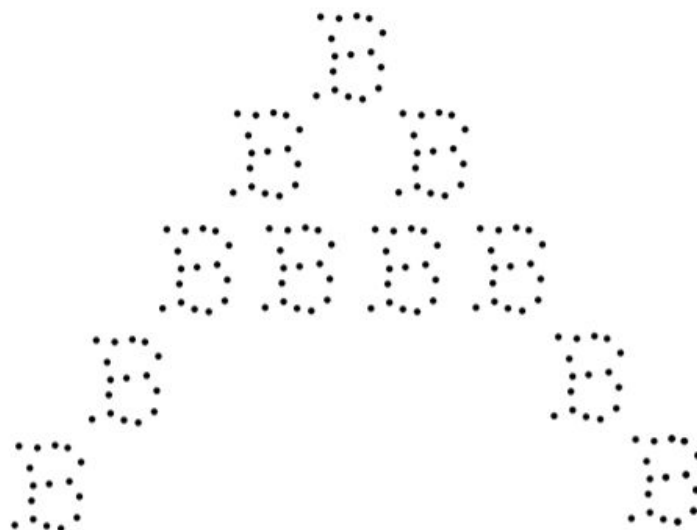
Outlier



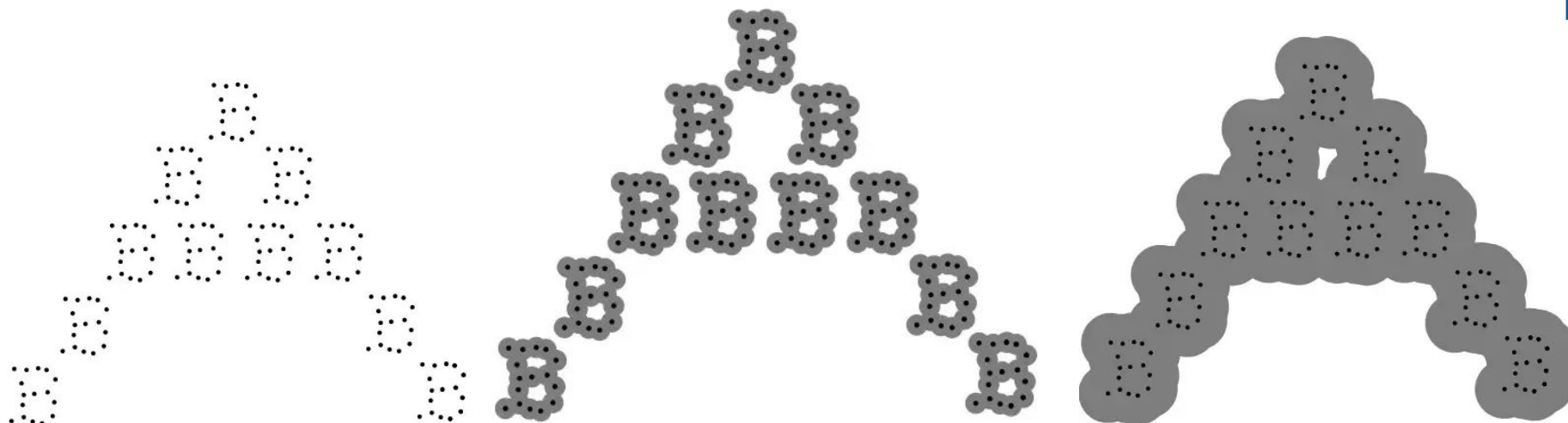
Outlier



O que você vê?



50 pontos, 11 letras, ou 1 letra?



Exemplo original

