

Tera

AULA 28

Recommender Systems

Instrutor: [Raphael Ballet](#)

Background:

- Engenheiro de Controle e Automação (IMT)
- Mestre em Sistemas Aeroespaciais e Mecatrônica (ITA)
- Data Scientist - Elo7

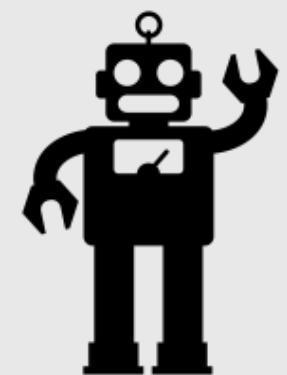
Interesses:



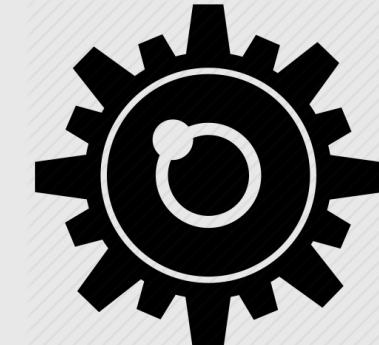
Drones



Aprendizado
de Máquina



Robótica



Visão
Computacional



Processamento de
Linguagem Natural



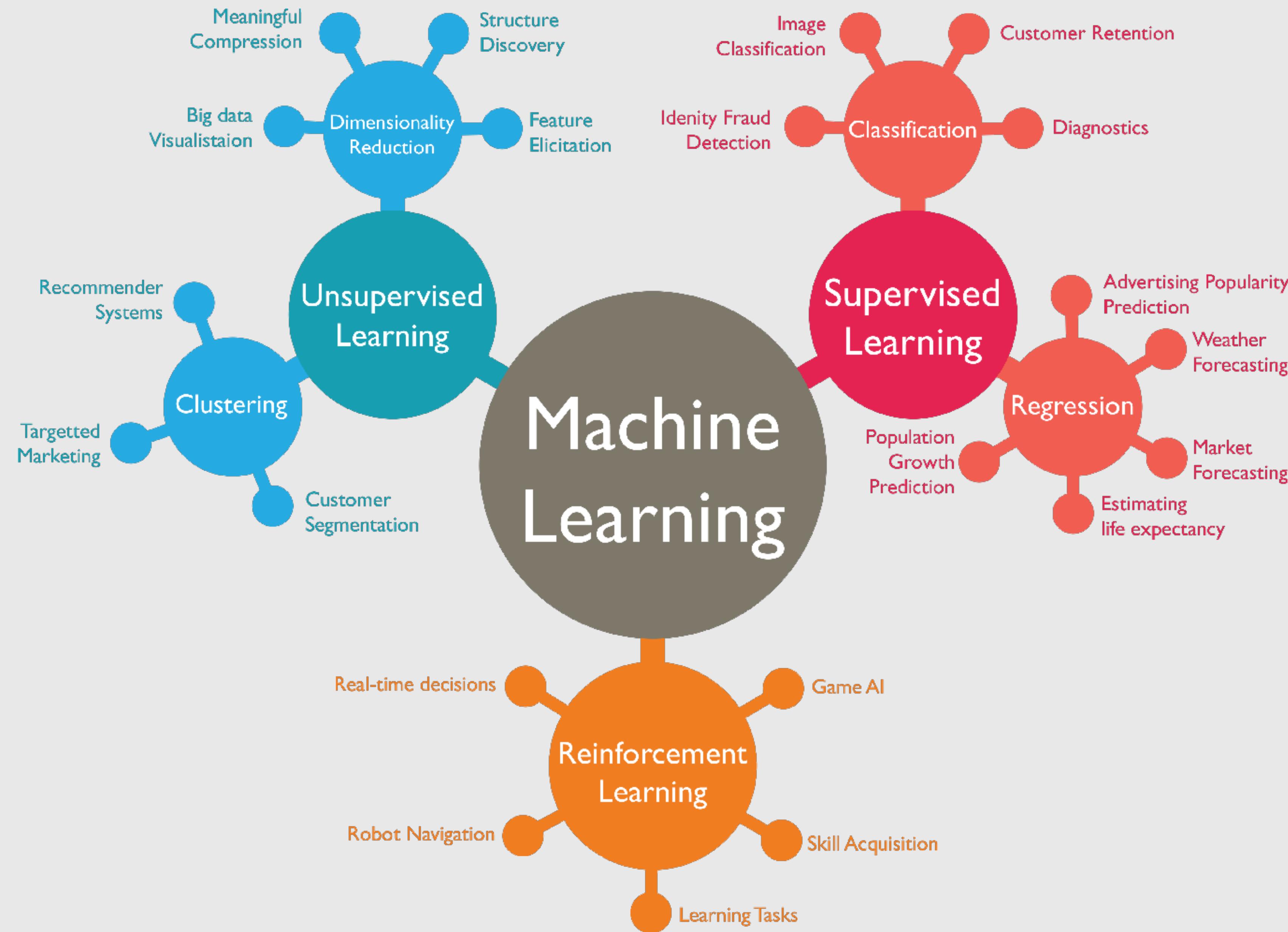
Sistemas de
recomendação

Planejamento:

1. Introdução
2. Sistemas de Recomendação: Conteúdo
3. Redução de Dimensionalidade
4. Visualização Multidimensional: T-SNE
5. Topic Analysis
6. Filtro Colaborativo

T

1. INTRODUÇÃO



I

1. INTRODUÇÃO

- Por que recomendação?

T

1. INTRODUÇÃO

Screenshot of the elo7 website search results for "lembrancinha".

The search bar shows the query "lembrancinha". The results page title is "Lembrancinha" with "1206310 PRODUTOS ENCONTRADOS".

Filter options include:

- Preço: R\$ [] até [] R\$ []
- Cidade: Digite uma cidade []
- Filtrar por: Todos produtos []
- Ordenar por: Relevância []

Product grid:

Image	Name	Description	Price
	LEMBRANCINHA URSINHO ARTICULADO	Atelier Aninha Ioannou	R\$ 15,00
	LEMBRANCINHA CASAMENTO SABONETE	Saboaria Natural Empório do Banho, desde 2000	R\$ 8,46
	Lembrancinha batizado	Amô Atelier	R\$ 6,99
	Tag para lembrancinha	My Party	R\$ 8,00 FRETE GRÁTIS
	Lembrancinha Minions	Vila em Festa	R\$ 3,00
	Bolo de frutas cristalizadas	Patricia Bertola Confeitaria	R\$ 30,00
	Dedoche Turma da Galinha Pintadinha	Encontro dos Pontos	R\$ 38,00
	Vaso com Rosas e Flores - Decoração	Origami Star	R\$ 60,00

I

1. INTRODUÇÃO

- Existe muita demanda e muita oferta, mas como fazer “match”?

I

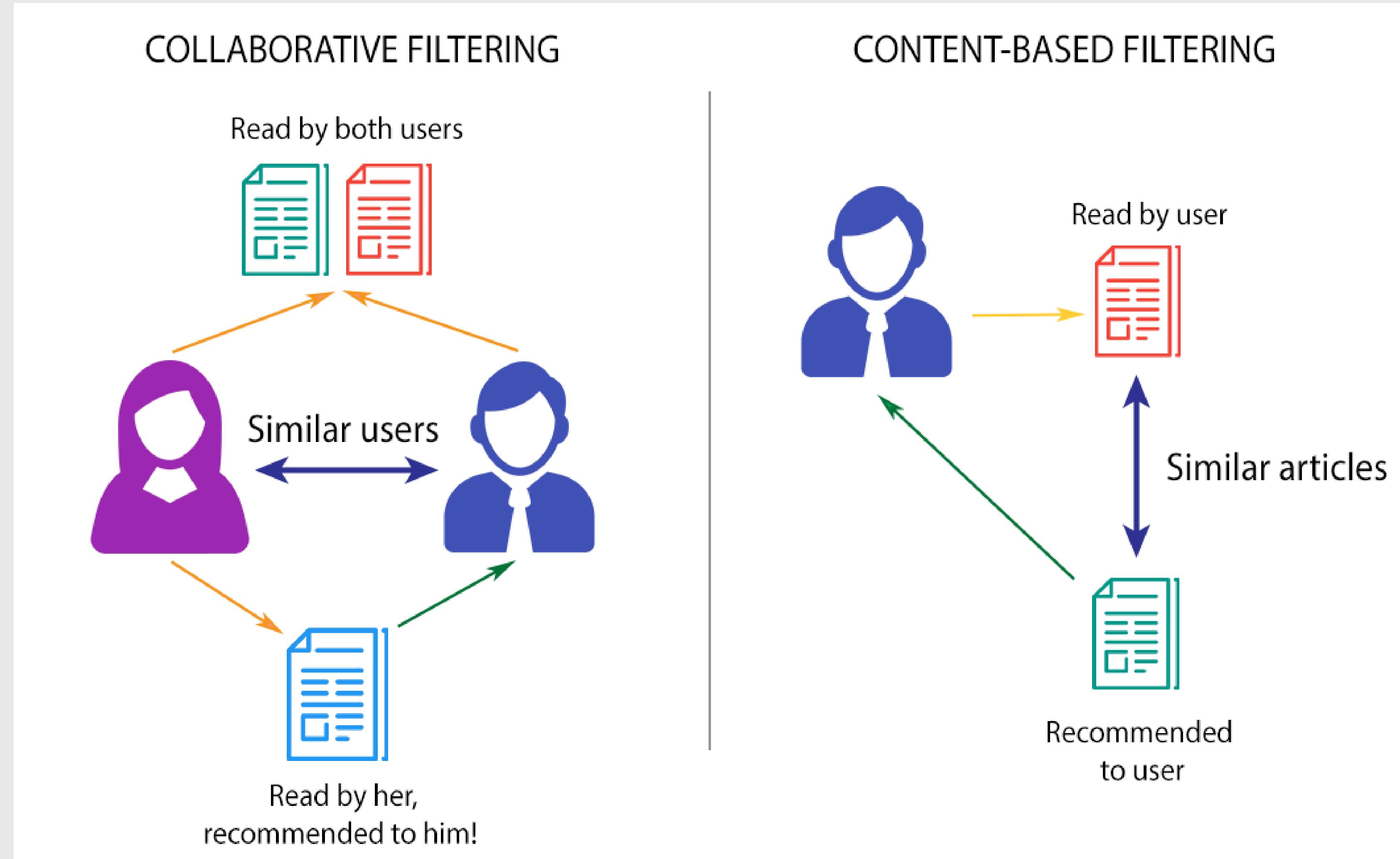
1. INTRODUÇÃO

- Existe muita demanda e muita oferta, mas como fazer “match”?
- Como recomendar produtos que nem mesmo o usuários sabia que queria?

2. Sistemas de Recomendação

- Existem 3 grandes grupos:
 - Proximidade de documentos (produtos, músicas, filmes etc)
 - Proximidade entre usuários (filtro colaborativo)
 - Híbrido → Mistura dos outros 2

2. Sistemas de Recomendação

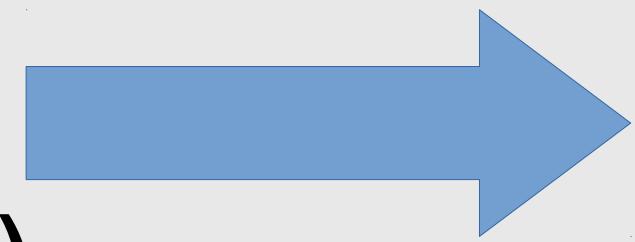


2. Sistemas de Recomendação

- Proximidade de documentos:
 - Distância entre documentos
 - Similaridade de temas (tópicos)

2. Sistemas de Recomendação

- Proximidade de documentos:
 - Distância entre documentos
 - Similaridade de temas (tópicos)



**Clustering / Topic
Analysis**

2. Sistemas de Recomendação

- Exemplo – Recomendação artigos NY Times



leu / gostou



Recomendação



- Esporte
- Baseball
- Campeonato

- Esporte
- Baseball

2. Sistemas de Recomendação

- Vetores muito esparsos:
 - Distâncias semelhantes (vetores/produtos \sim equidistantes)
 - Tempo de cálculo muito alto
 - Excessivo espaço em memória utilizado

2. Sistemas de Recomendação

- Solução:
- **Clustering**
- **Redução de dimensionalidade**
- **Análise de Tópicos**

I

3. Redução de Dimensionalidade

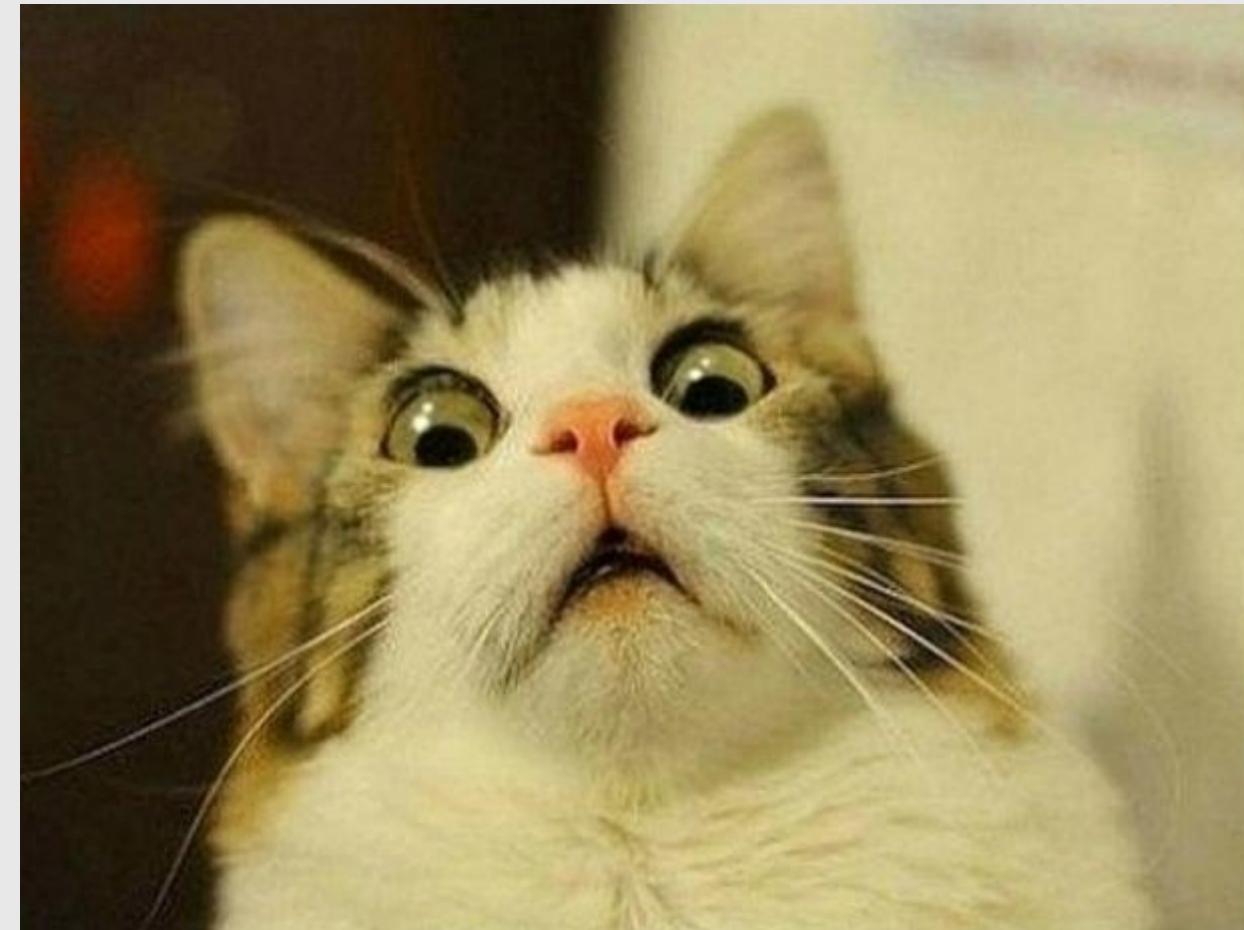
- Dois objetivos principais:
 - Facilitar visualização e intuição
 - Amenizar o problema de similaridade entre observações

I

3. Redução de Dimensionalidade

- Dois objetivos principais:
 - Facilitar visualização e intuição
 - Amenizar o problema de similaridade entre observações

Maldição da dimensionalidade



I

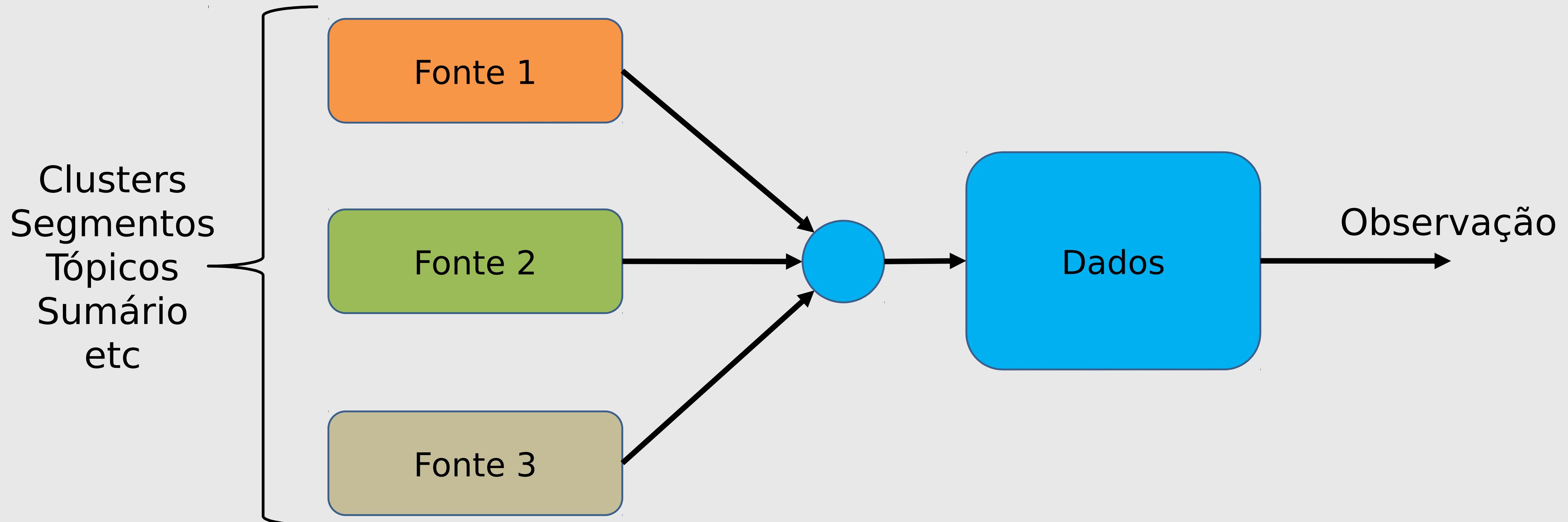
3. Redução de Dimensionalidade

- Técnicas principais:
 - PCA
 - T-SNE (Visualização apenas)
 - Topic Analysis (NMF e LDA)

I

3. Redução de Dimensionalidade

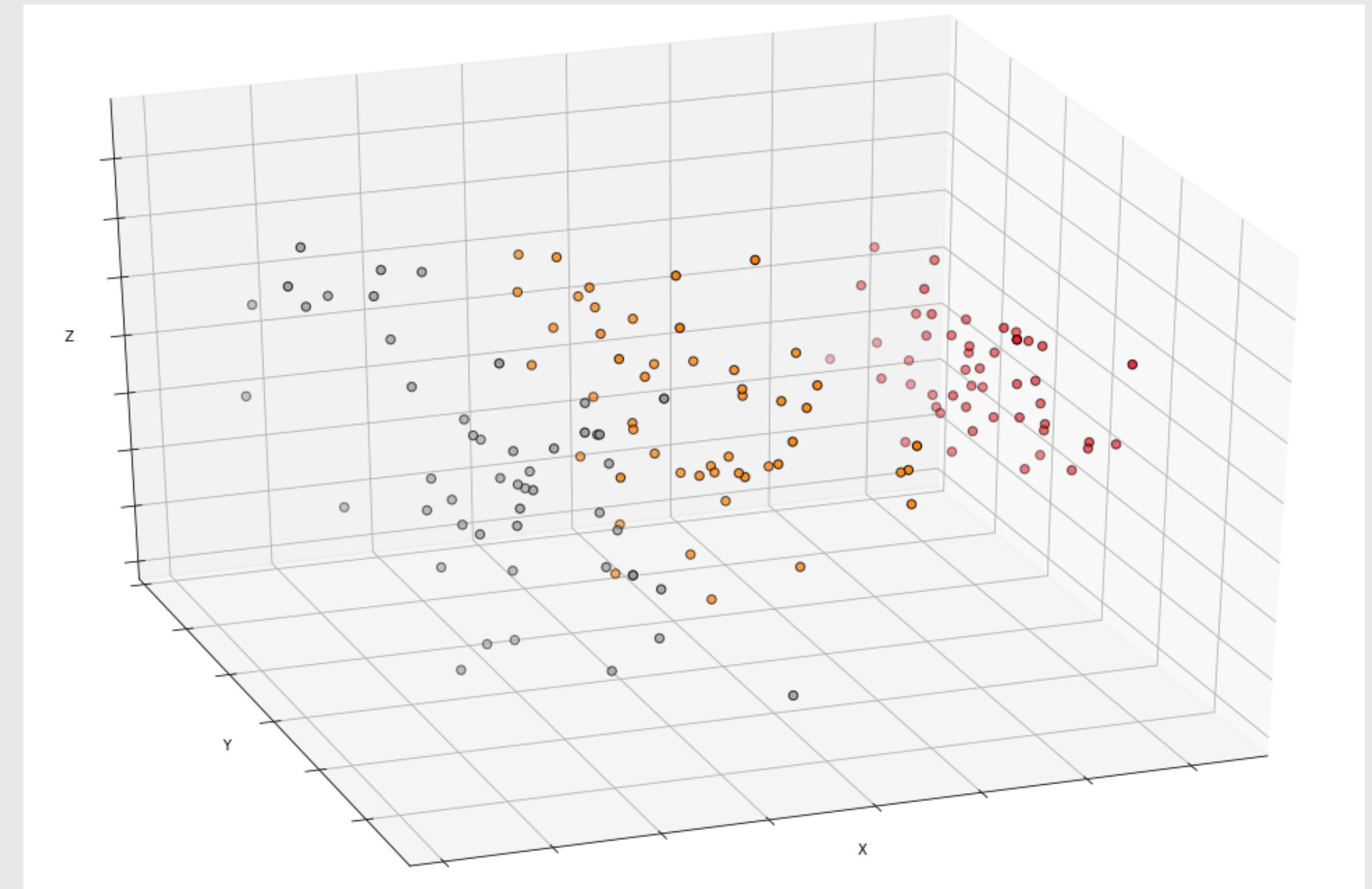
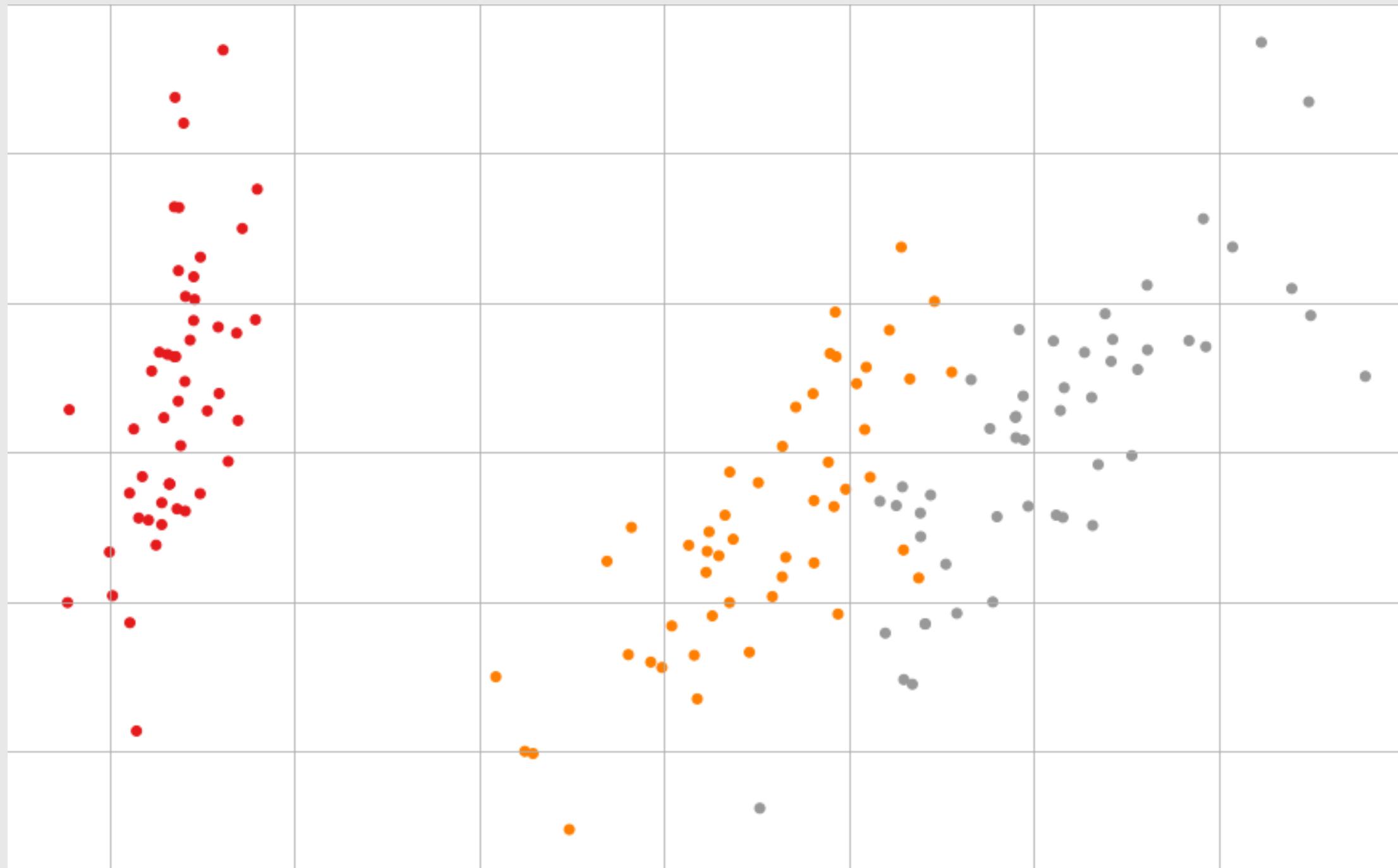
- Documentos



I

3. Redução de Dimensionalidade

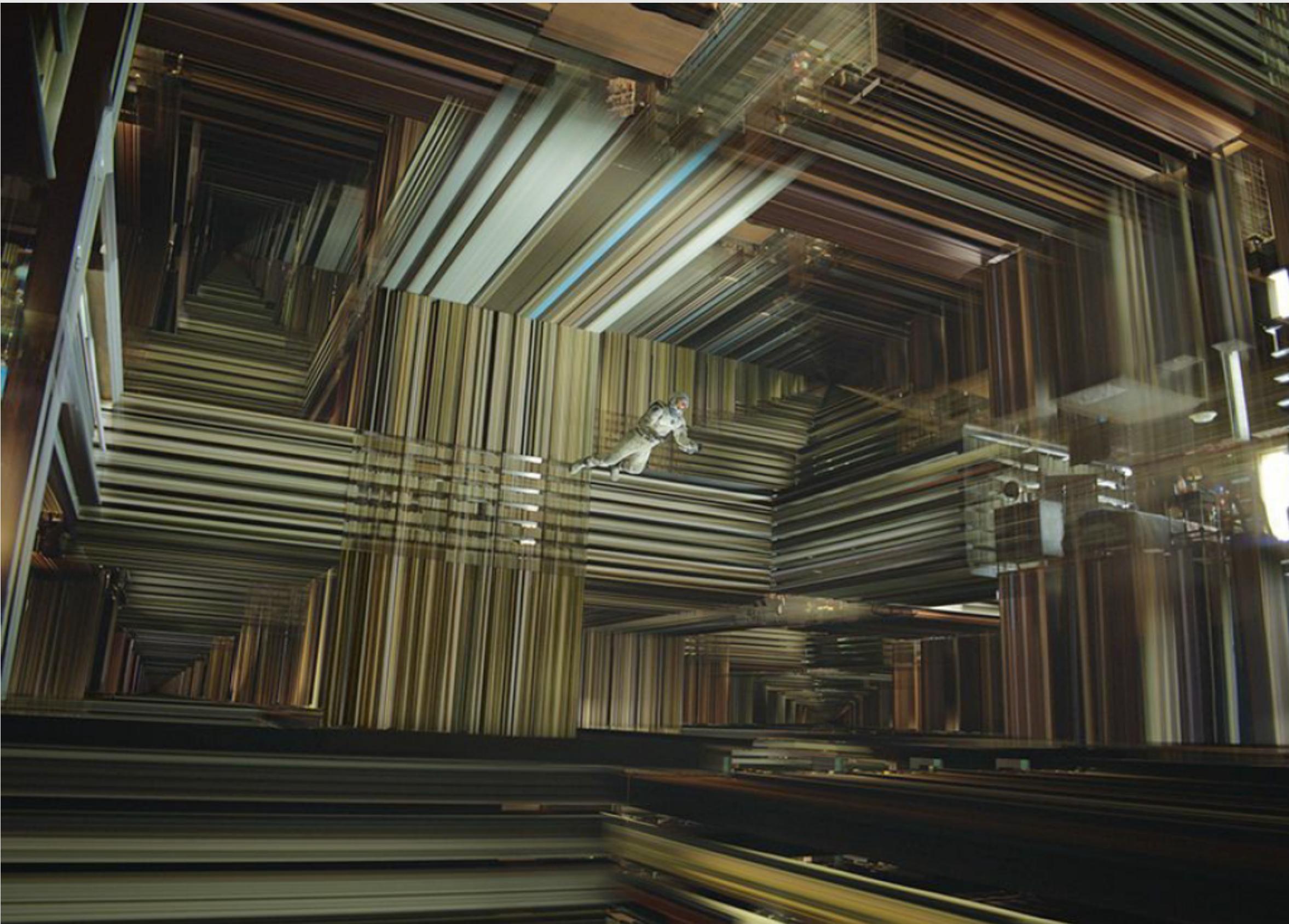
- Visualização dos dados - 2D e 3D



I

3. Redução de Dimensionalidade

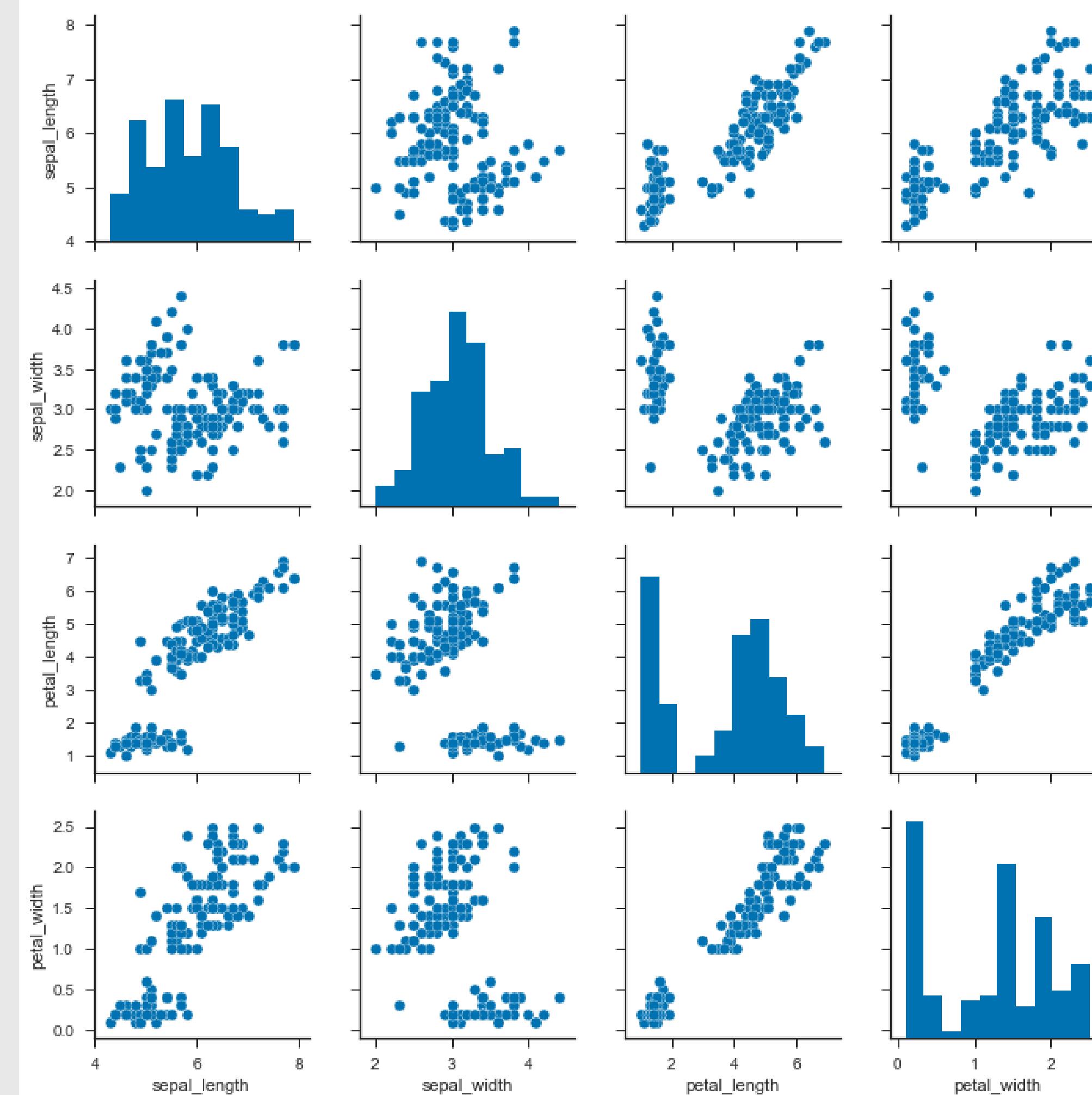
- Visualização mais do que 3D?



I

3. Redução de Dimensionalidade

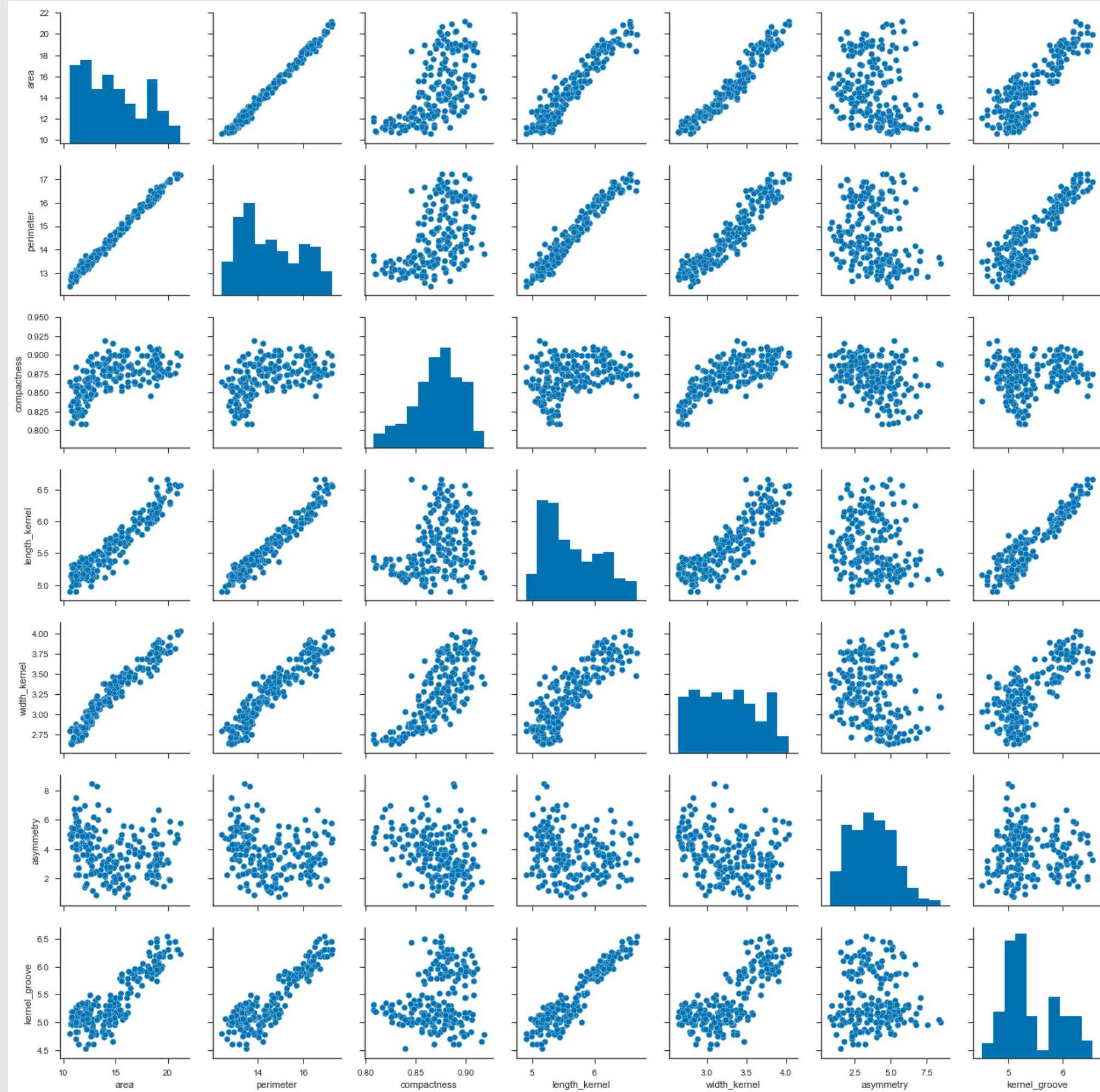
- Visualização 4 dimensões



I

3. Redução de Dimensionalidade

- Visualização 7 dimensões



I

3. Redução de Dimensionalidade

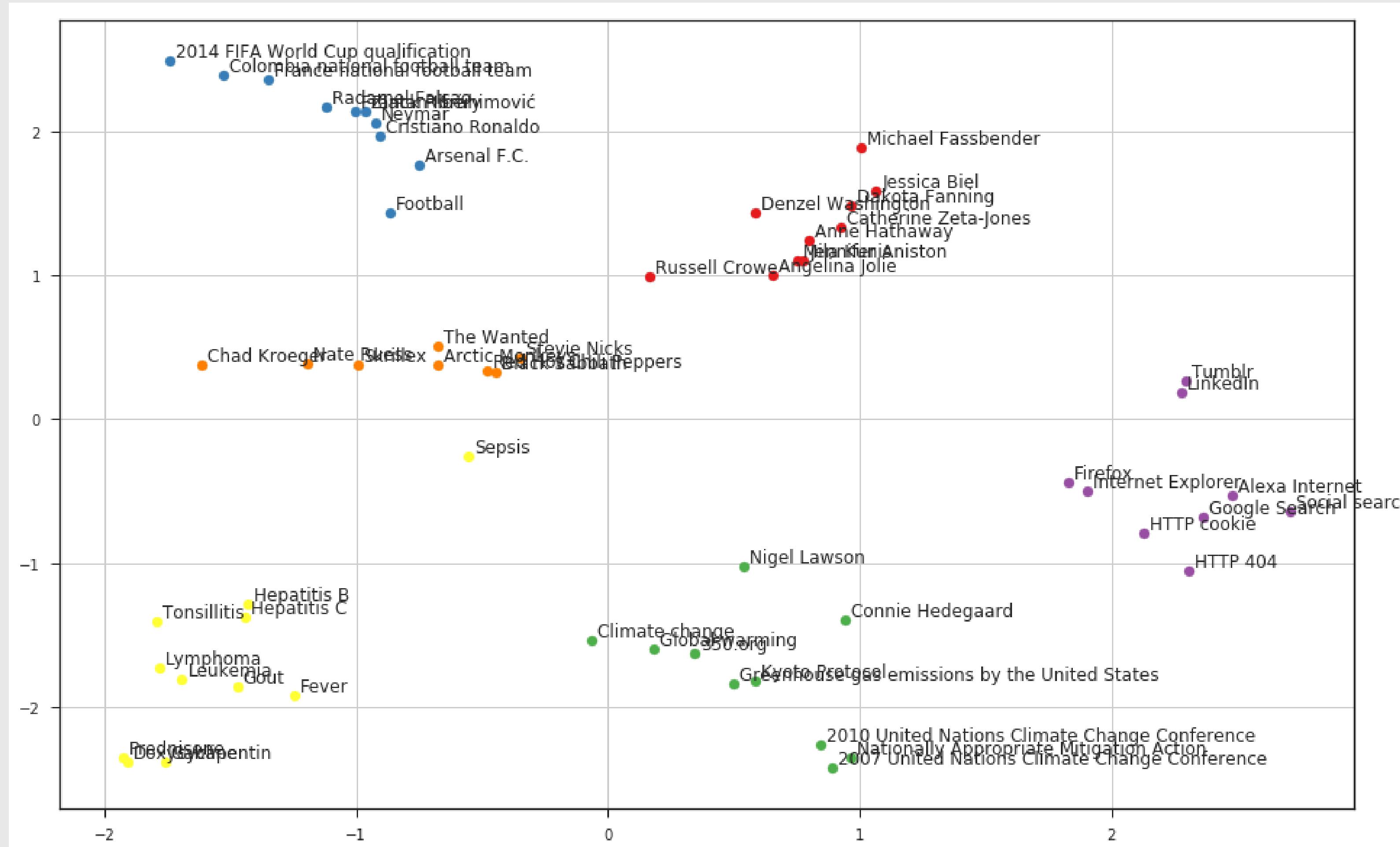
- Visualização NLP (>1k dimensões)

?

I

3. Redução de Dimensionalidade

- Visualização NLP (>1k dimensões): Desejo



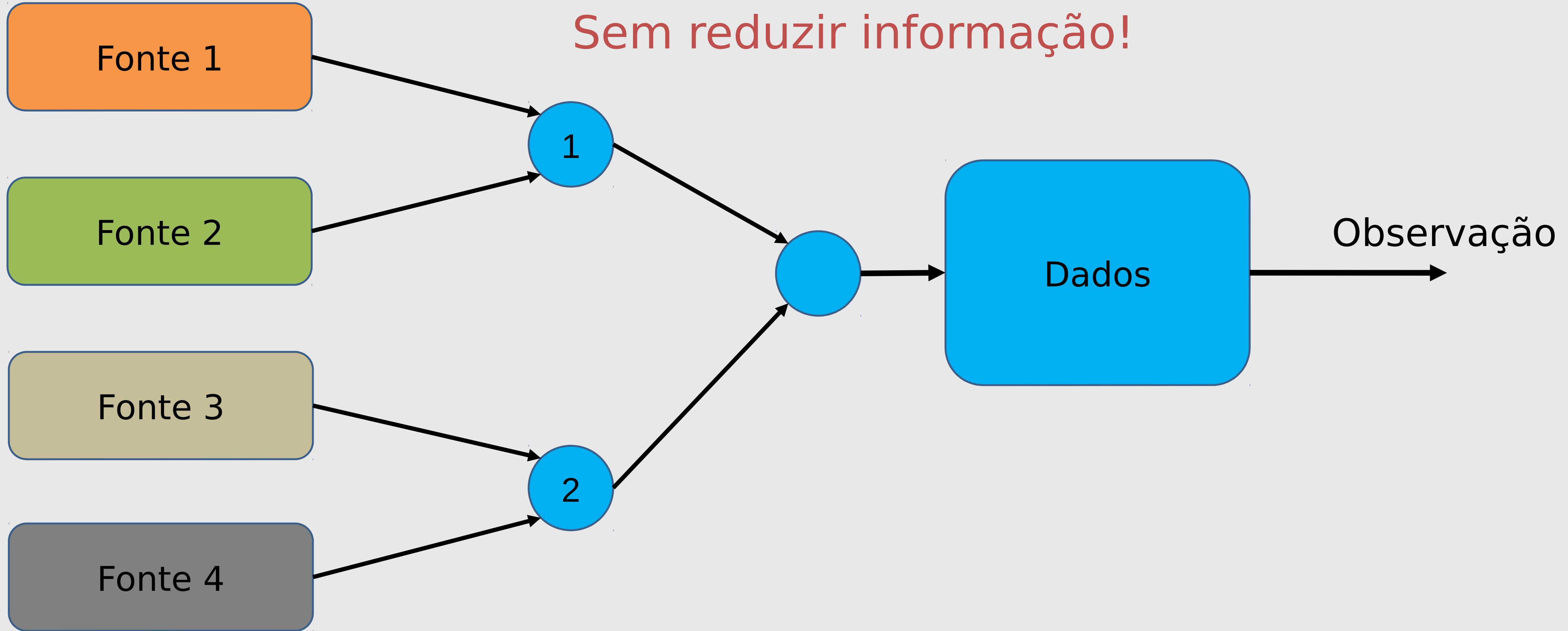
I

3. Redução de Dimensionalidade

- Objetivo

Sumário – Menor dimensão

Sem reduzir informação!



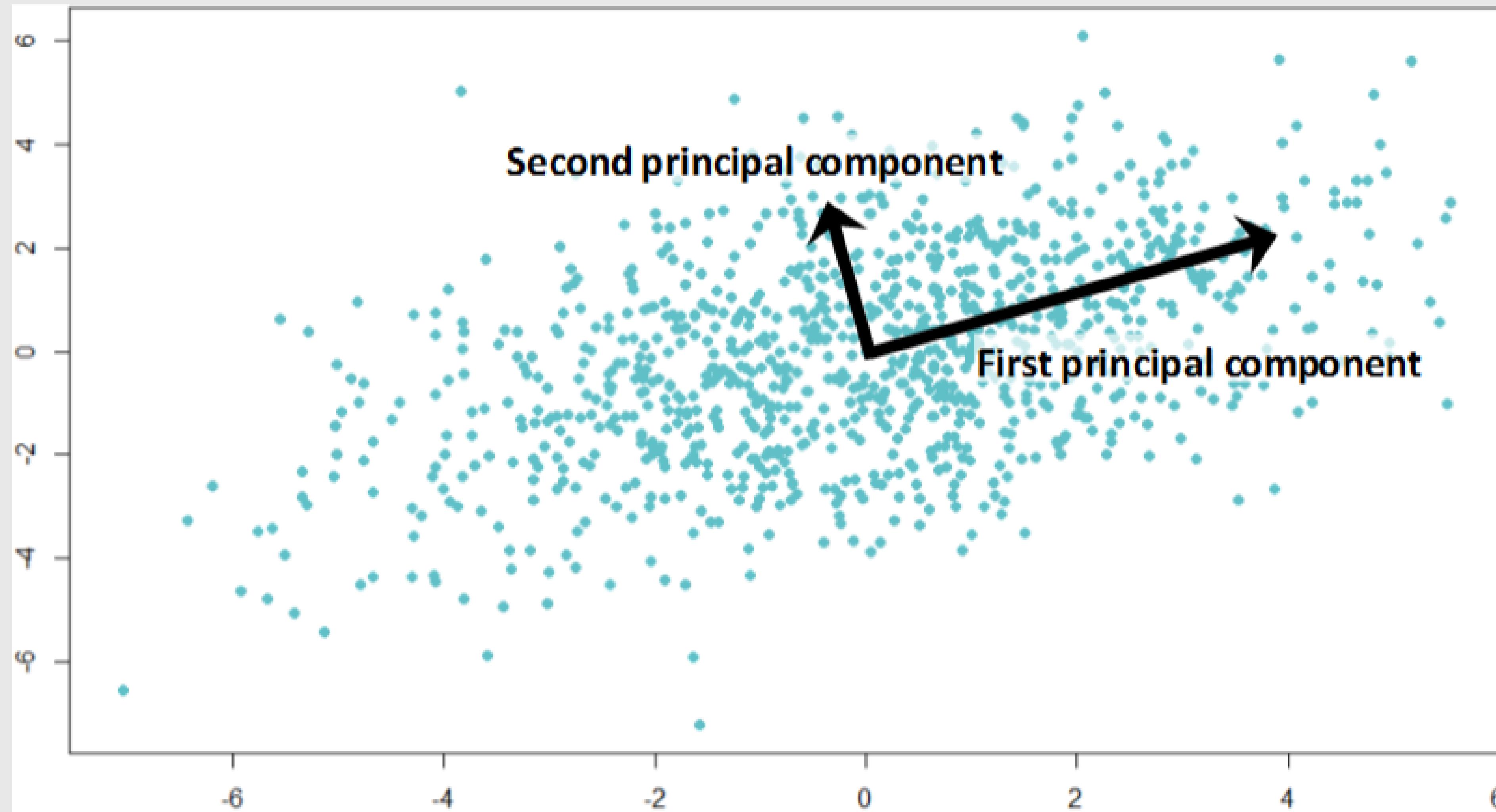
3. PCA

- **Principal Component Analysis:**
 - Encontra atributos de maior variação → “mais importantes”
 - Elimina atributos de menor variação → “menos explicativos”

I

3. PCA

- Componentes principais:



3. PCA

- PCA: Componentes principais:
 - **Primeiro componente:**
 - Direção de maior variação nos dados
 - **Segundo componente:**
 - Direção da segunda maior variação e ortogonal ao primeiro (descorrelacionado do primeiro)

...

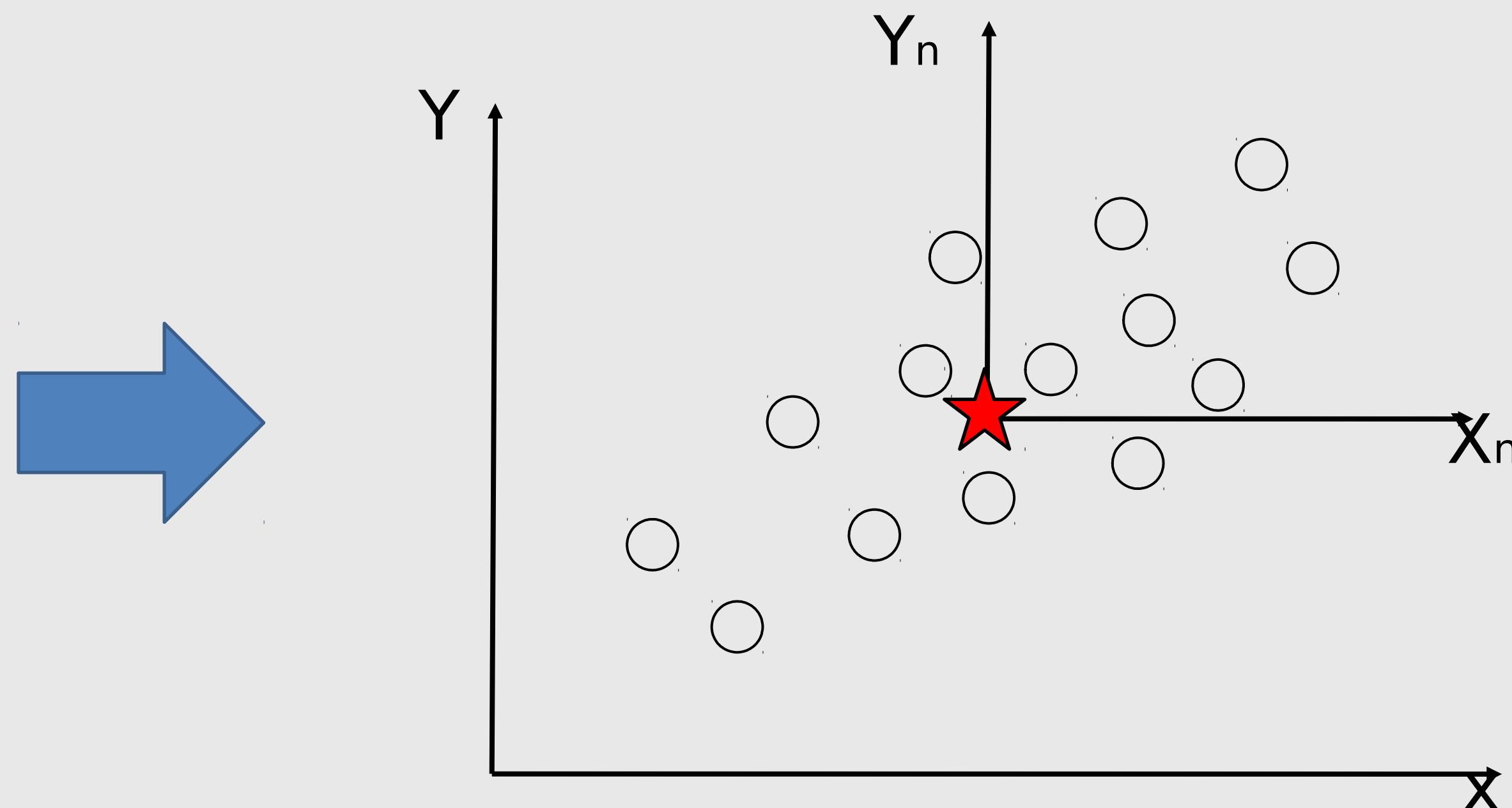
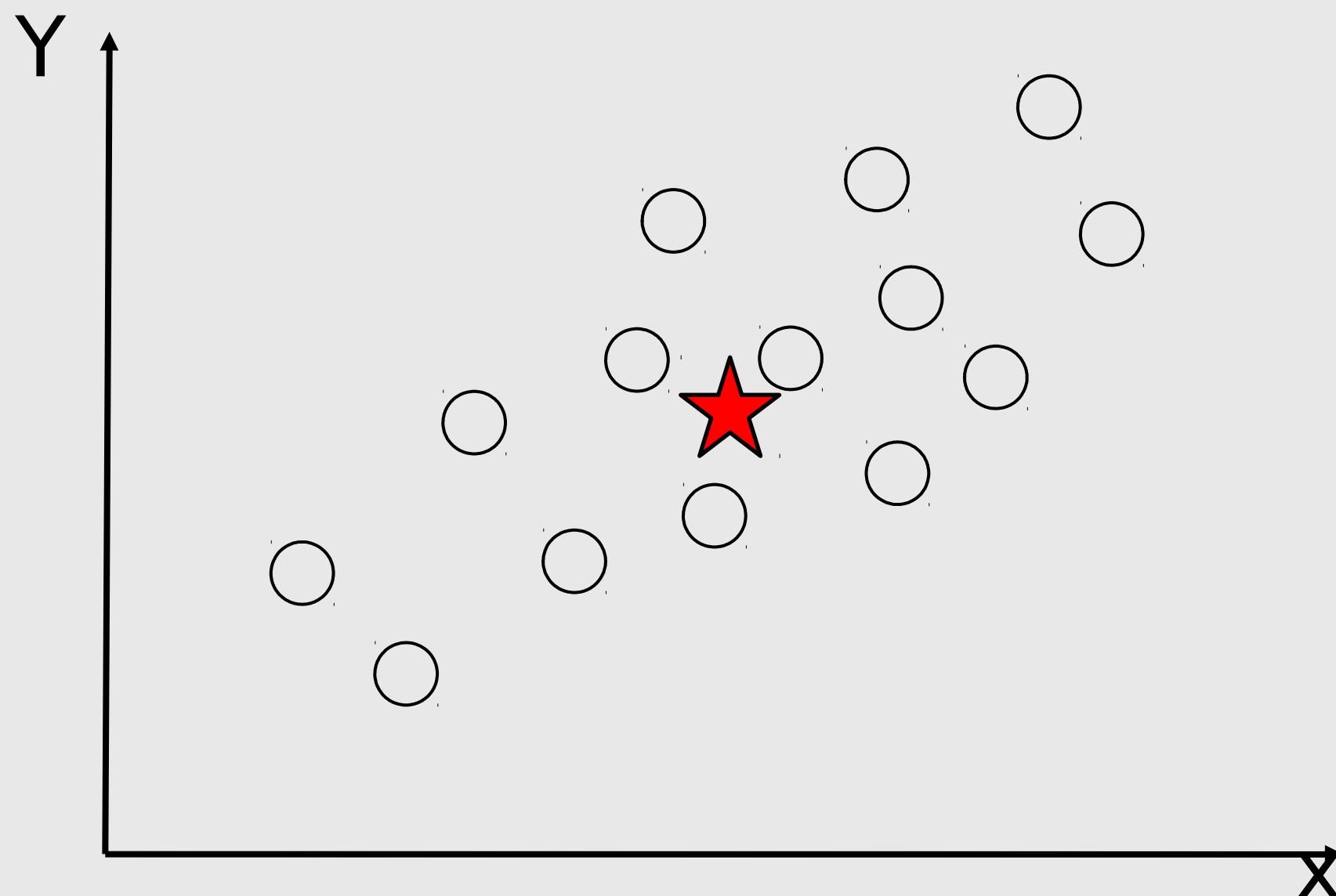
3. PCA

- Passos do algoritmo PCA:
 - 1) Remove média amostral dos dados
 - 2) Rotaciona os eixos para descorrelacionar os atributos
 - 3) Ordena os componentes principais em nível de variância
 - 4) Remove os componentes menos variantes (Opcional)

I

3. PCA

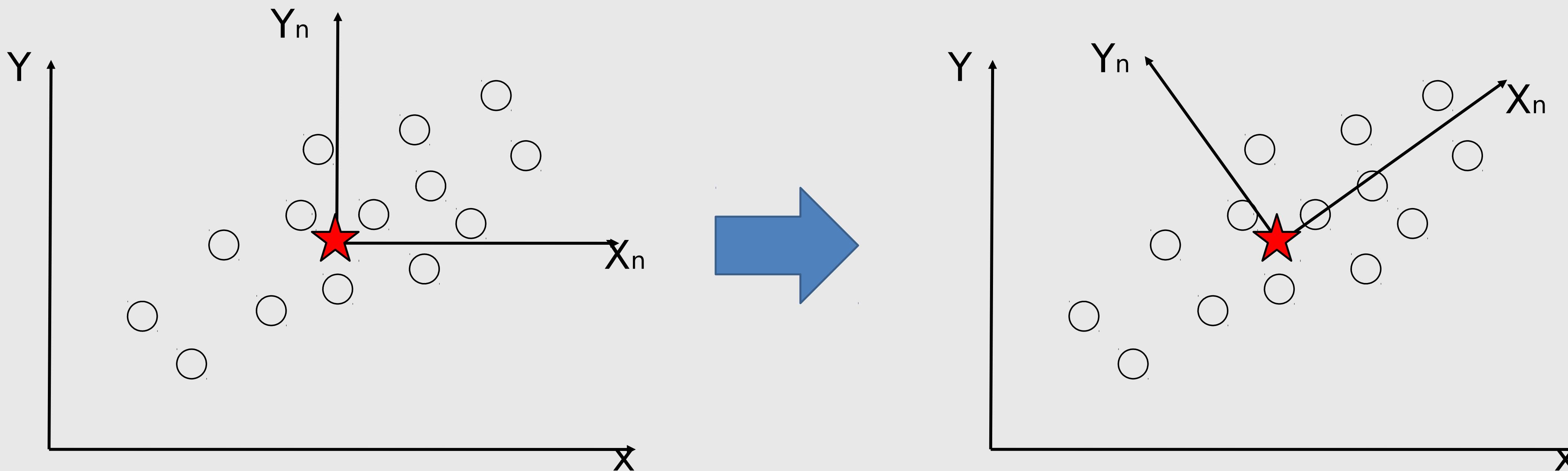
1) Remove média amostral dos dados



T

3. PCA

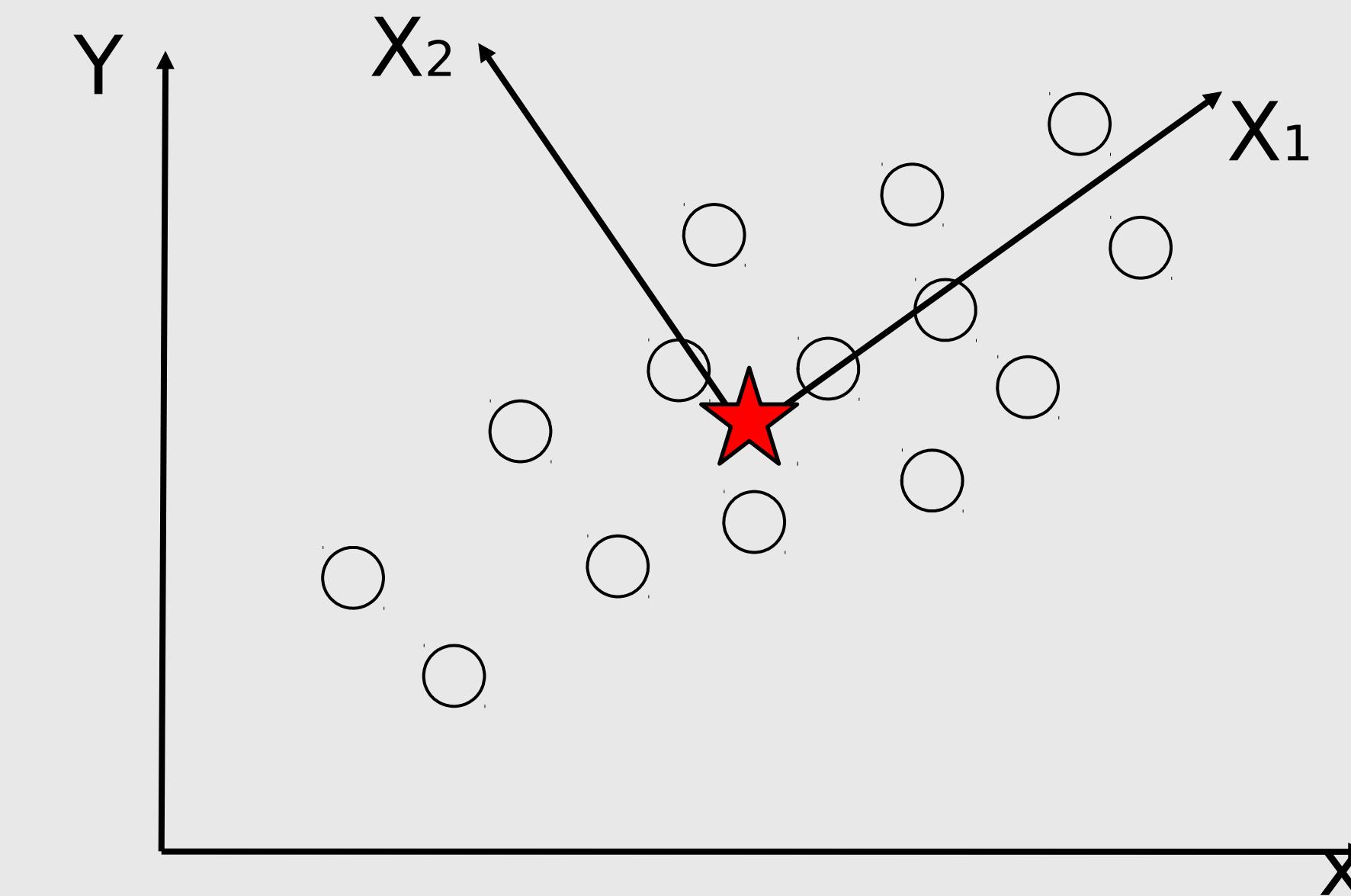
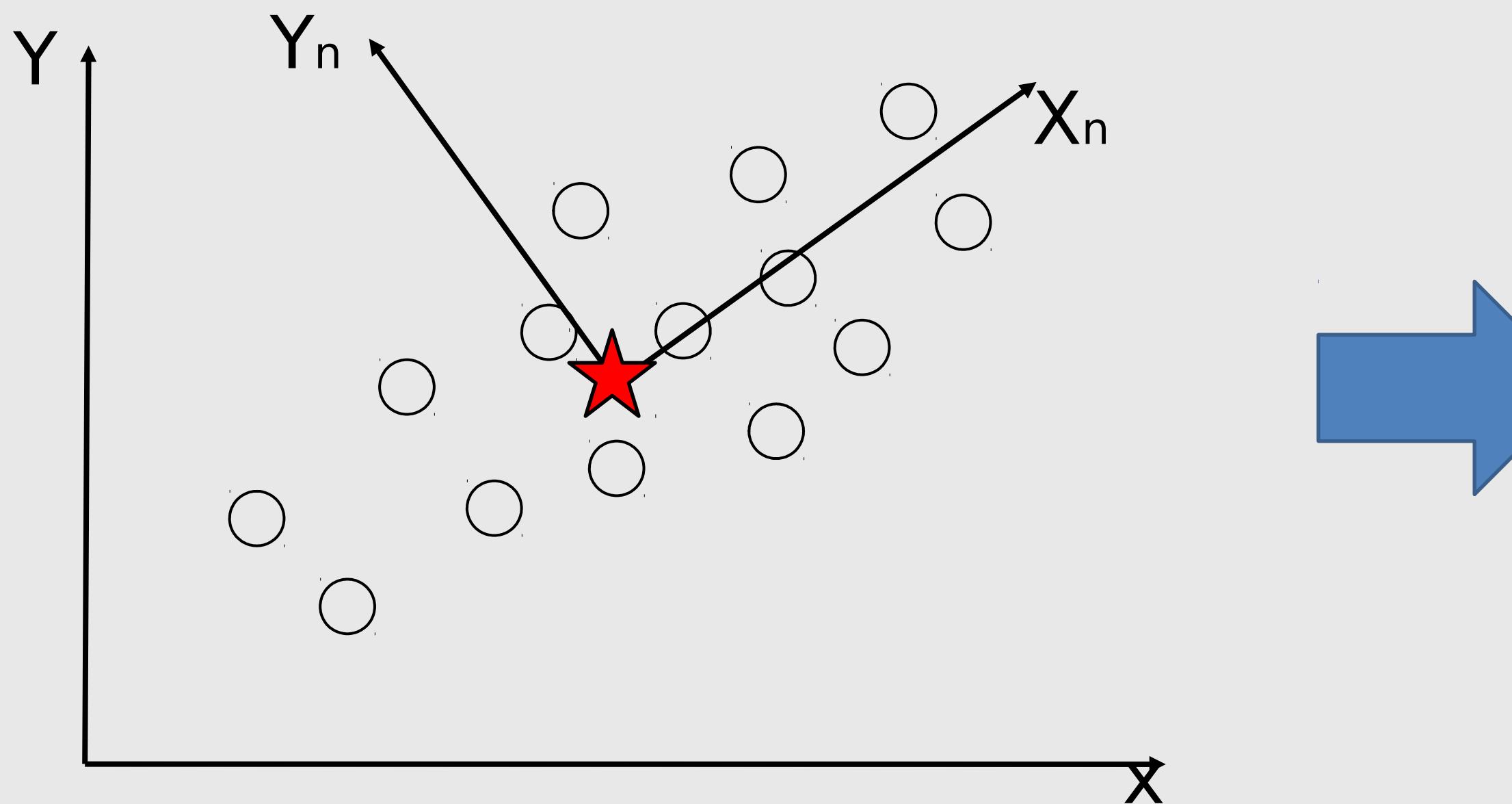
2) Rotaciona os eixos para descorrelacionar os atributos



I

3. PCA

3) Ordena os componentes principais em nível de variância



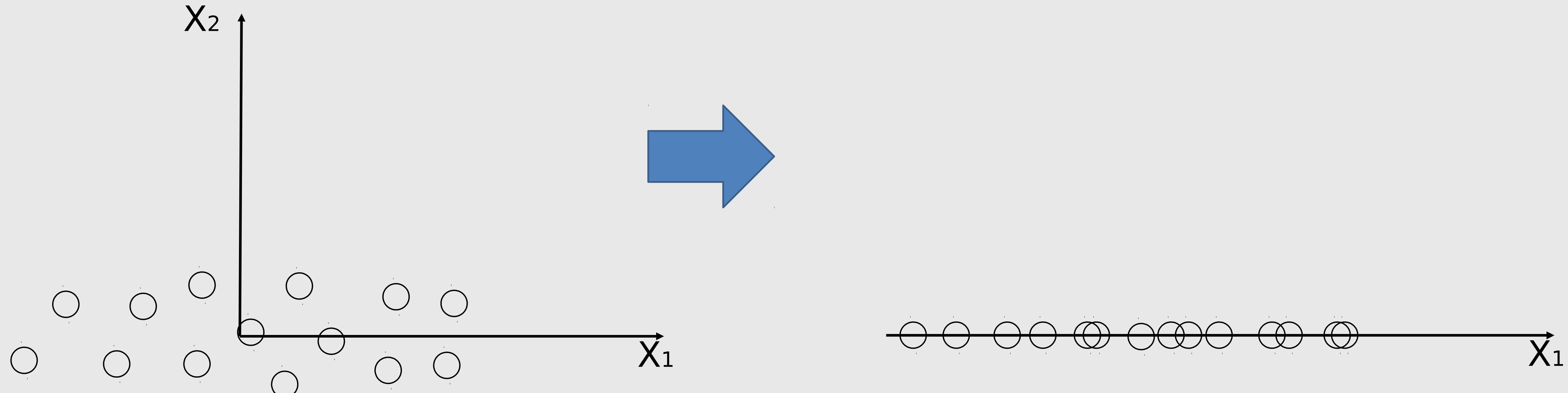
X_1 : Primeiro componente principal

X_2 : Segundo componente principal

I

3. PCA

4) Remove os componentes menos variantes (Opcional)



X_1 : Primeiro componente principal

X_2 : Segundo componente principal

I

3. PCA

- Quando remover componentes principais:
 - Atributos muito correlacionados
 - Componentes secundários pouco variantes
→ baixa **variância explicada**
 - Balanço entre precisão e simplificação
 - Encontrar dimensão intrínseca

T

3. PCA

- Exemplo: notebook

3. PCA

- **Vantagens:**
 - Permite reduzir dimensionalidade do problema sem perder informação
 - Menor dimensionalidade
 - Maior velocidade e menos memória para algoritmos de ML
 - Resultados determinísticos

3. PCA

- **Desvantagens:**
 - Dimensões resultantes (componentes principais) não representam os atributos
 - Perde a “explicabilidade” do algoritmo
 - Má escolha de número de componentes pode prejudicar análise
 - Encontra apenas relações lineares

4. Visualização: T-SNE

- **T-distributed Stochastic Neighbor Embedding**
 - Mapeia N-dimensões em 2 ou 3 dimensões
 - Distância é proporcional a probabilidade de proximidade entre pontos (afinidade)
 - Método iterativo baseado em otimização (gradiente descendente)
 - Procura manter a estrutura dos dados

T

4. T-SNE

- Mas, nós já temos o PCA. Por que usar T-SNE?

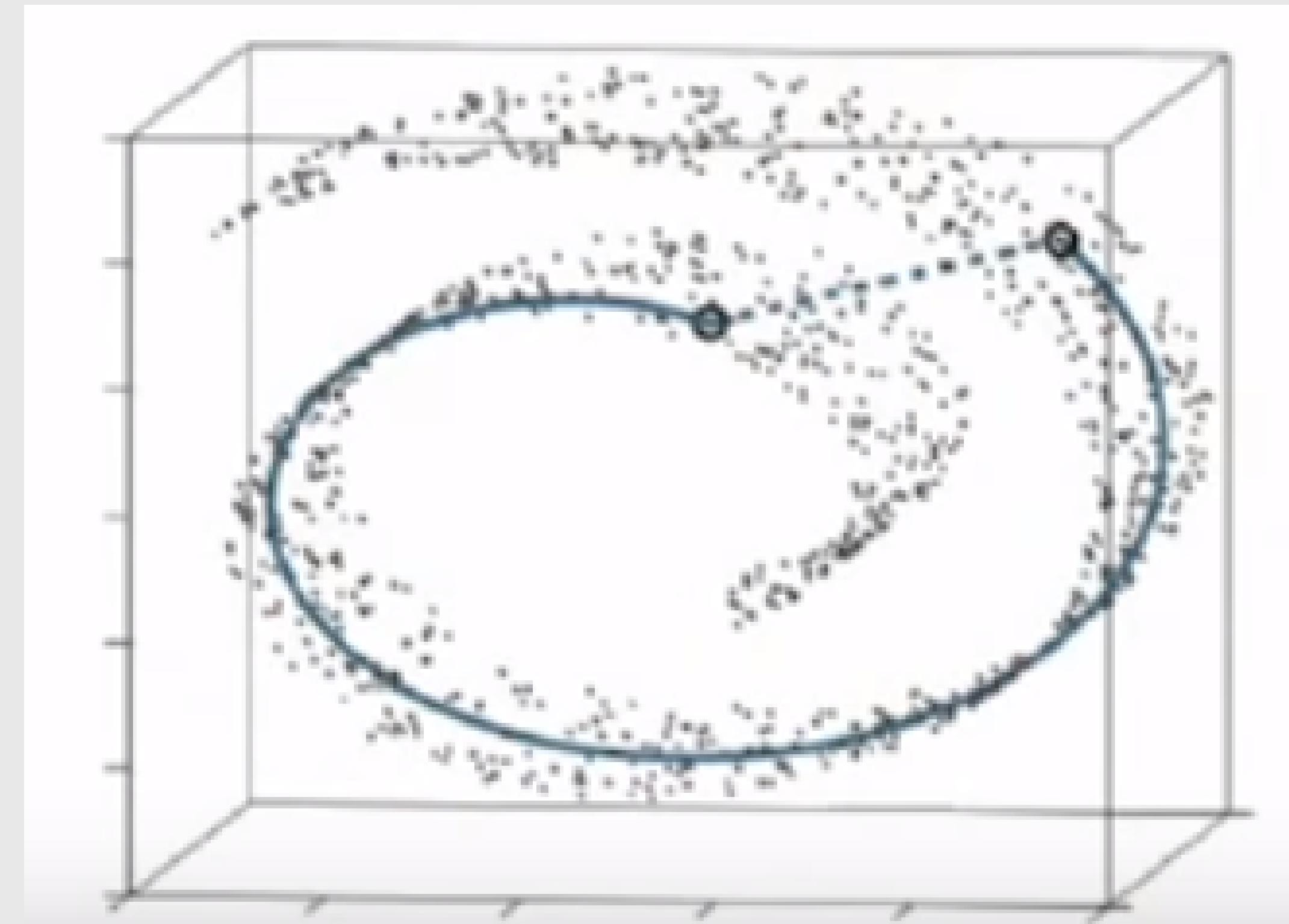
4. T-SNE

- Mas, nós já temos o PCA. Por que usar T-SNE?
 - PCA encontra apenas relações lineares
 - Falha ao encontrar estruturas complexas

T

4. T-SNE

- Mas, nós já temos o PCA. Por que usar T-SNE?
 - PCA encontra apenas relações lineares
 - Falha ao encontrar estruturas complexas



4. T-SNE

- Existem basicamente 2 parâmetros:
 - **Learning rate:** Taxa de aprendizado - gradiente descendente
 - **Perplexity:** Número aproximado de vizinhos de um ponto (observação) - Entre 5 e 50

T

4. T-SNE

- Exemplo: notebook

4. T-SNE

- **Vantagens:**
 - Permite a visualização de relações entre dados multidimensionais
 - Mantém a estrutura dos dados (não-linear)
 - Rápido e eficiente mesmo para grandes dimensões e grande quantidade de observações

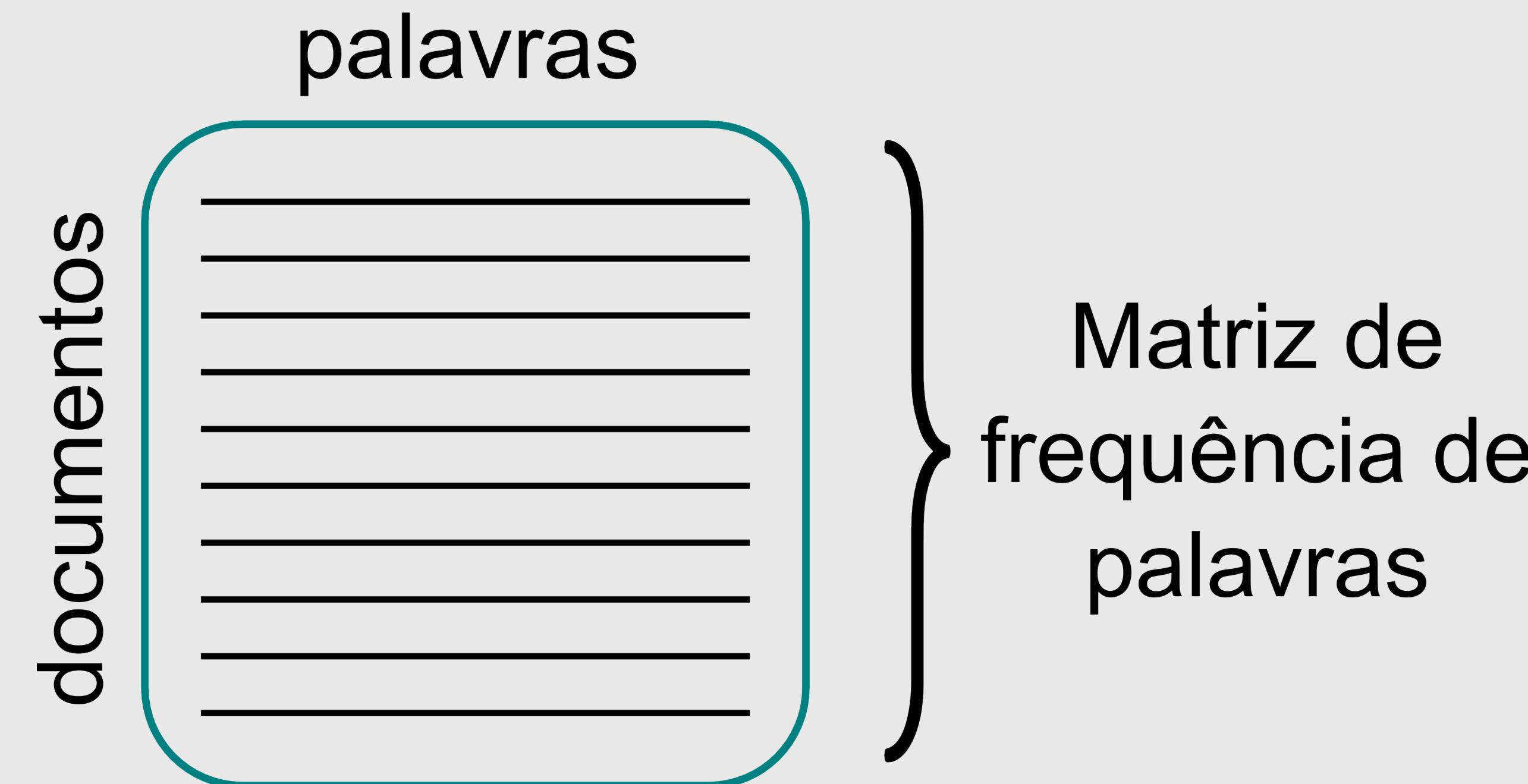
4. T-SNE

- **Desvantagens:**
 - Depende da escolha dos parâmetros (nem sempre fácil)
 - Não possui repitibilidade dos resultados
 - Distâncias entre clusters não significam nada
 - Interpretação dos resultados não trivial

T

5. Topic Analysis

- Problema já conhecido:
 - Documentos + Palavras = Muitas dimensões



T

5. Topic Analysis

- **Solução:**

- Redução de dimensionalidade
- Clustering

...

T

5. Topic Analysis

- **Solução:**
 - Redução de dimensionalidade
 - Clustering
- **Problema:**
 - **Perda de interpretabilidade dos dados**

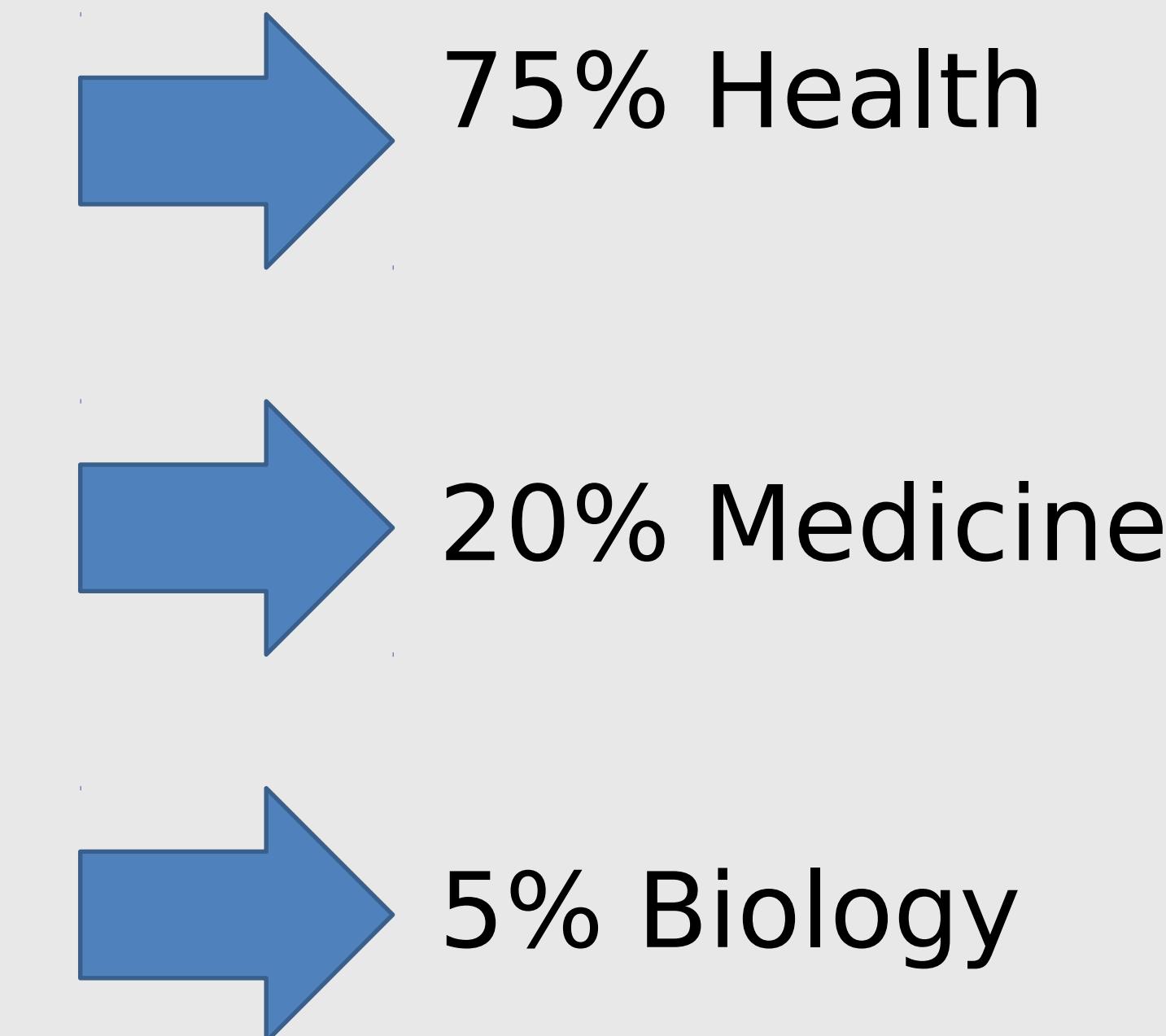
T

5. Topic Analysis

- **Objetivo:**

- Encontrar estrutura implícita nos documentos – Tópicos / Temas

Screenshot of the Wikipedia article on Dengue fever. The page shows the title "Dengue fever", a featured status banner, and a summary of the disease. It includes sections on signs and symptoms (fever, rash, headache, joint pain), clinical course (febrile phase, critical phase), and treatment. A sidebar provides classification and external resources, including ICD-10 codes (A90, 061) and MeSH terms (C02.782.417.214). Below the article is a diagram titled "Symptoms of Dengue fever" showing a human figure with various symptoms labeled.



T

5. Topic Analysis

- Principais algoritmos:
 - Non-Negative Matrix Factorization (NMF)
 - Latent Dirichlet Allocation (LDA)

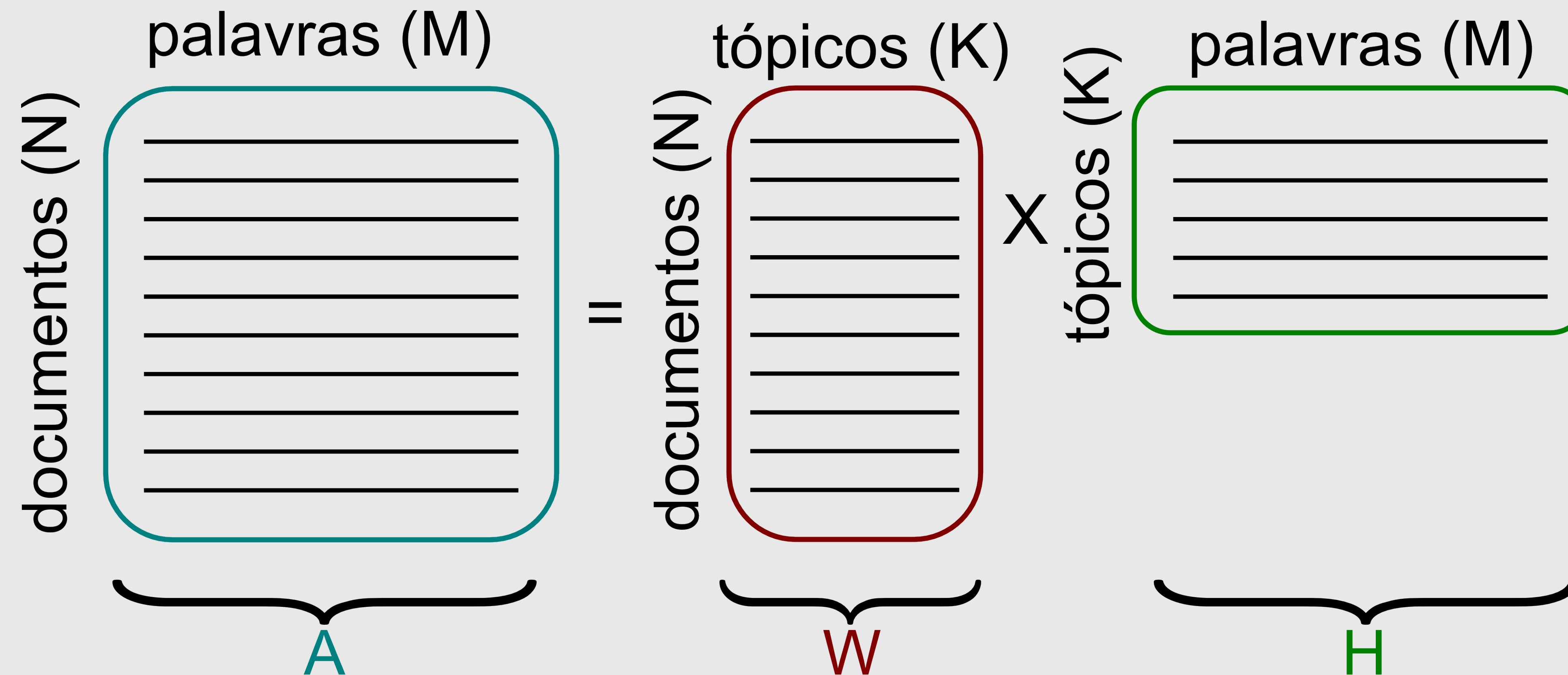
5. Topic Analysis - NMF

- **Non-Negative Matrix Factorization (NMF)**
- Principal objetivo:
 - Decompor a matriz de frequênciа de palavras em representações de tópicos
 - Documentos são compostos de combinações de tópicos
 - Tópicos são compostos de combinações de palavras

T

5. Topic Analysis - NMF

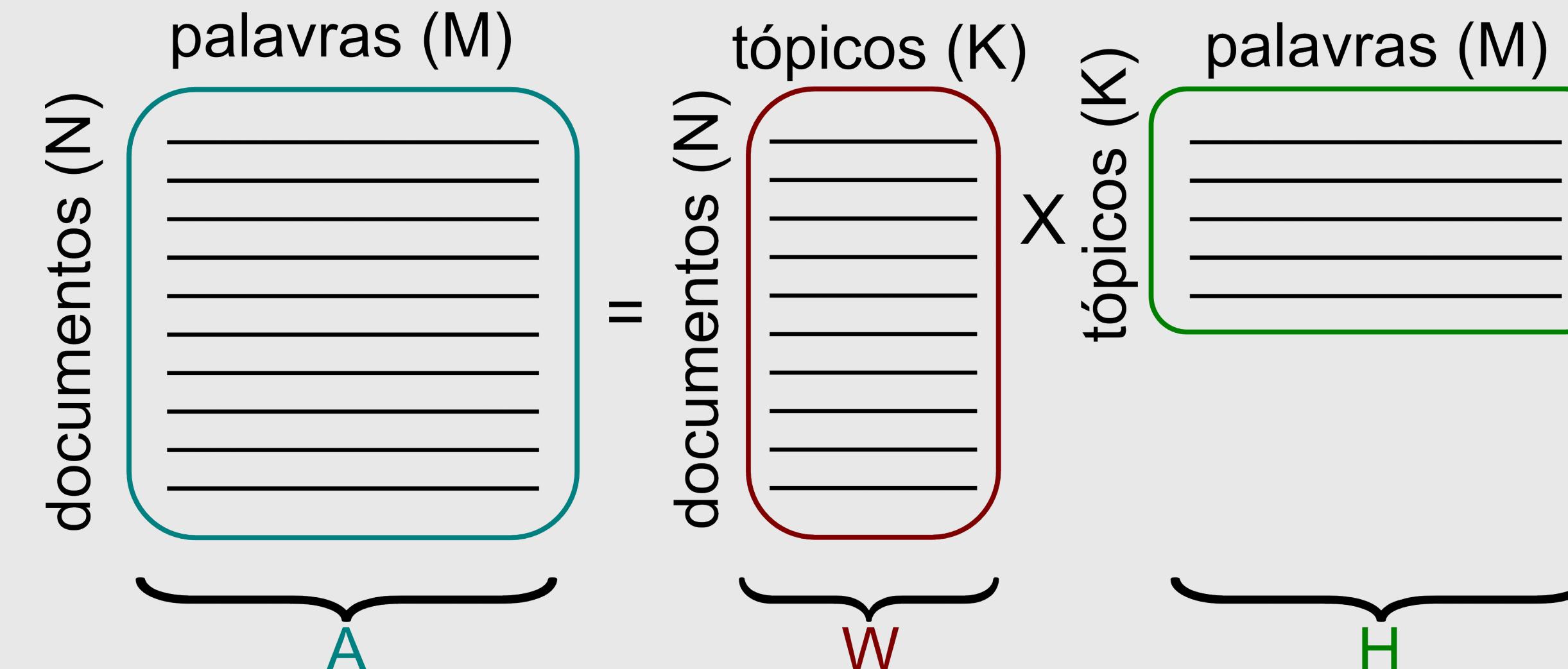
- NMF: Fatoração $\rightarrow A = WH$



T

5. Topic Analysis - NMF

- Matrizes:
 - A: Matriz de frequência de termos (M) em documentos (N)
 - W: Matriz de pesos → distribuição de tópicos (K) nos documentos
 - H: Matriz de atributos → distribuição de palavras nos tópicos



5. Topic Analysis - NMF

- Principais características:
 - Precisa definir o número de tópicos
 - Matrizes A, W e H não podem ter valores negativos
 - Matrizes W e H podem reconstruir matriz A (aprox.)

5. Topic Analysis - NMF

- NMF pode ser utilizado em vários outros cenários:
 - **Segmentação de fontes sonoras do áudio:**
 - Documentos: áudio
 - Features: espectograma do áudio
 - **Segmentação de imagens:**
 - Documentos: imagem
 - Features: pixels

T

5. Topic Analysis - NMF

- Exemplo: notebook

5. Topic Analysis - NMF

- **Vantagens:**
 - Tópicos são interpretáveis
 - Naturalmente agregador (clustering)
 - Pode ser utilizado em outros contextos (ex: imagens, áudio etc)

T

5. Topic Analysis - NMF

- **Desvantagens:**
 - Solução aproximada
 - Pode causar overfitting
 - Limitação de utilizar apenas features positivas

T

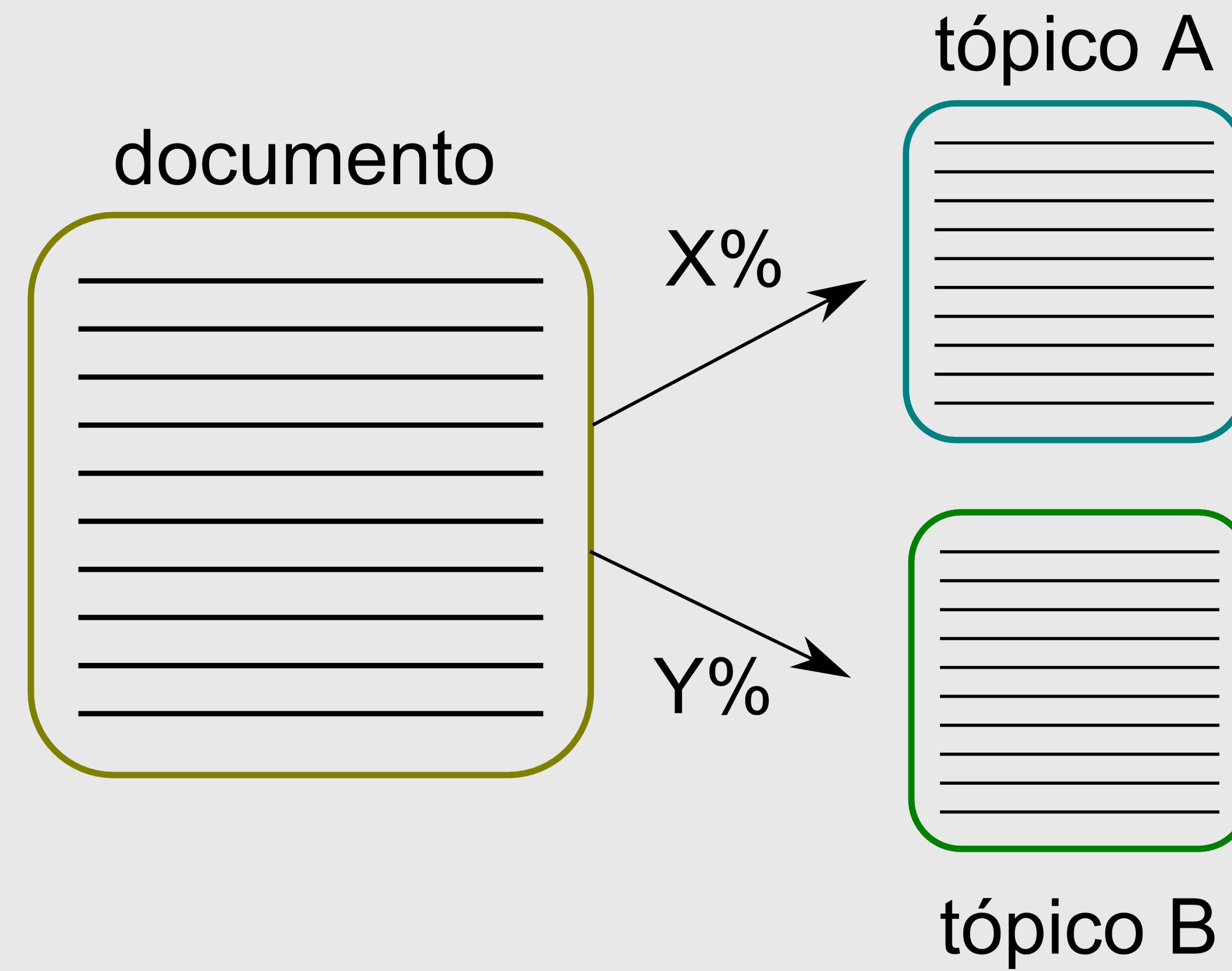
5. Topic Analysis - LDA

- **Latent Dirichlet Allocation (LDA)**
- Método probabilístico
- Representa documentos como uma mistura de tópicos
- Precisa definir o número de tópicos (igual NMF)

T

5. Topic Analysis - LDA

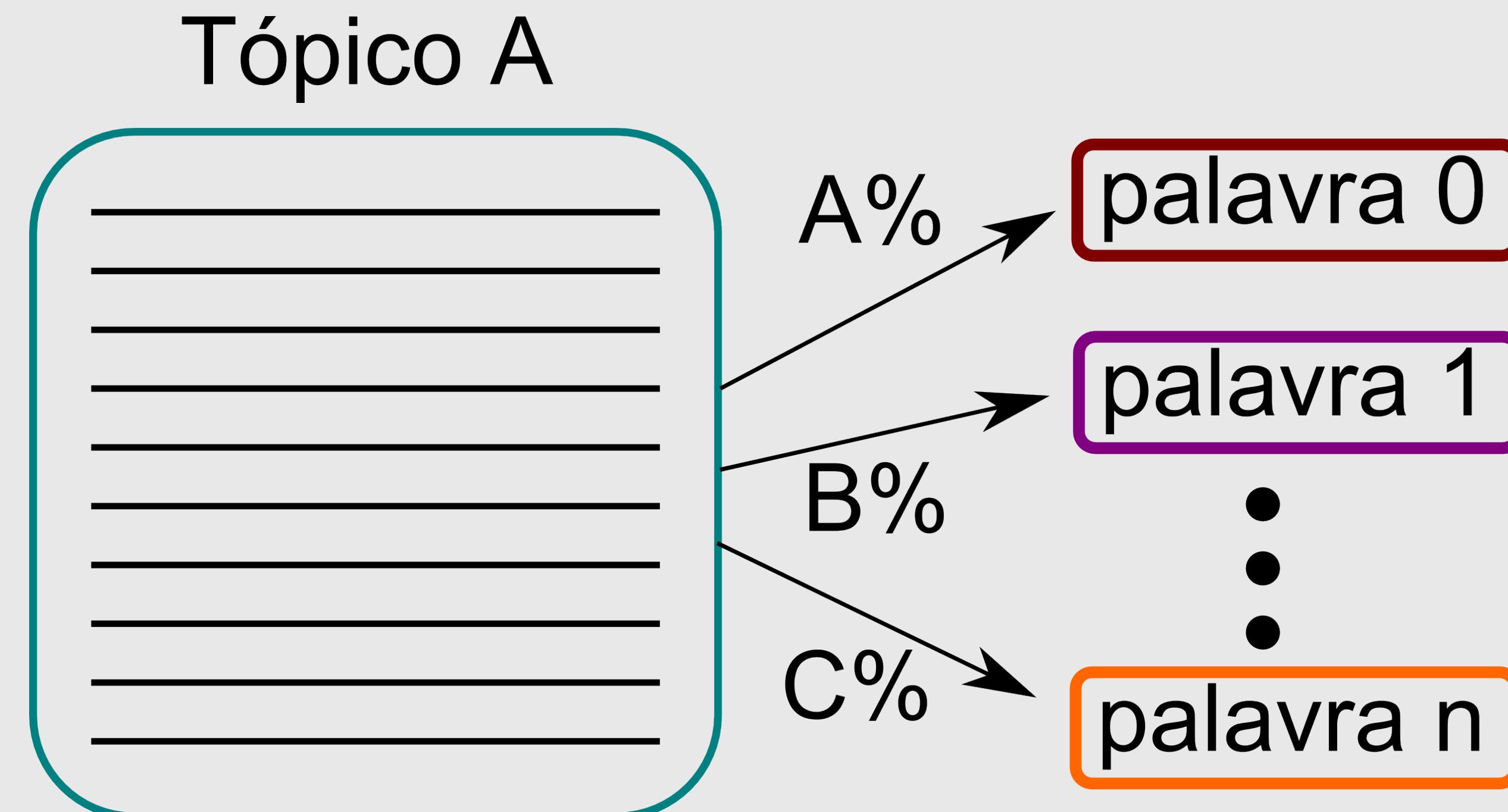
- Documento → Místura de tópicos



I

5. Topic Analysis - LDA

- Tópicos → Mistura de palavras



I

5. Topic Analysis - LDA

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

I

5. Topic Analysis - LDA

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Tópico A: Comida
- Tópico B: Animais

5. Topic Analysis - LDA

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Documento 1: Apenas tópico A
- Documento 2: Apenas tópico B
- Documento 3: Mistura dos tópicos A e B

I

5. Topic Analysis - LDA

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Qual o tópico associado a palavra “Fish” no documento 3?
 - $P('Fish' | \text{tópico A}) = 0.75 (3 - A, 1 - B)$
 - $P('Fish' | \text{tópico B}) = 0.25$

I

5. Topic Analysis - LDA

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Qual a probabilidade de cada tópico no documento 3?

➤ $P(\text{tópico A} \mid \text{Documento 3}) = P(\text{tópico B} \mid \text{Documento 3}) = 0.5$

I

5. Topic Analysis - LDA

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Portanto, podemos concluir que “Fish” está contido no tópico A.

T

5. Topic Analysis - LDA

- O método é repetido para todas as palavras múltiplas vezes
- O algoritmo para quando não houver mais variação (convergência)

T

5. Topic Analysis - LDA

- Exemplo: notebook

I

5. Topic Analysis - LDA

- **Vantagens:**

- Tópicos são interpretáveis
- Permite variação de tópicos e palavras (distribuição)
- Permite gerar documentos novos

T

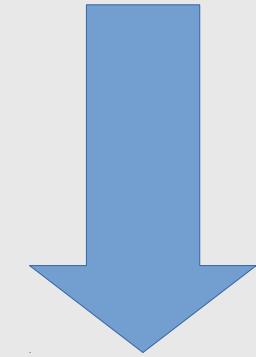
5. Topic Analysis - LDA

- **Desvantagens:**
 - Mesmas desvantagens do NMF

T

6. Filtro Colaborativo

- Proximidade entre usuários (filtro colaborativo):
 - Usuários semelhantes consomem documentos semelhantes



**Clustering
Topic Analysis**

6. Filtro Colaborativo

- Proximidade entre usuários (filtro colaborativo):
 - **Documentos:**
 - Histórico de consumo do usuário (compra, avaliação etc)
 - **Atributos / Features:**
 - Lista de itens de consumo (produtos, livros, filmes etc)

T

6. Filtro Colaborativo

- Exemplo: Recomendação de filmes

[Close](#)

Other Movies You Might Enjoy

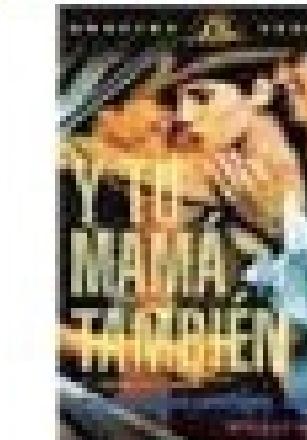
[Amelie](#)



[Add](#)

[Not Interested](#)

[Y Tu Mama Tambien](#)



[Add](#)

[Not Interested](#)

[Guys and Balls](#)



[Add](#)

[Not Interested](#)

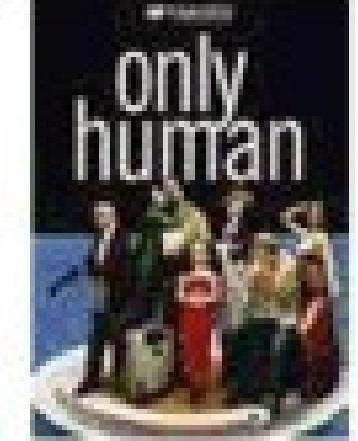
[Mostly Martha](#)



[Add](#)

[Not Interested](#)

[Only Human](#)



[Add](#)

[Not Interested](#)

[Russian Dolls](#)



[Add](#)

[Not Interested](#)

Eiken has been added to your Queue at position 2.

This movie is available now.

[Move To Top Of My Queue](#)

[Continue Browsing](#) [Visit your Queue](#)

[Close](#)

T

6. Filtro Colaborativo

- Exemplo: Recomendação de filmes



T

6. Filtro Colaborativo

- Exemplo: Recomendação de filmes

Recomendação!



6. Filtro Colaborativo

- **Recomendação:**

Encontrar items / usuários semelhantes

→ Menor distância entre vetores

- **Problema:**

Vetores muito esparsos: muitas dimensões sem valores

I

6. Filtro Colaborativo

- Exemplo: Recomendação produtos Elo7

	Histórico de compras				
	prod0	prod1	prod2	prod3	...
user0	0	1	0	0	...
user1	0	0	1	0	...
user2	1	0	0	0	...

→ ~8 milhões
de produtos

Milhões de compradores diferentes

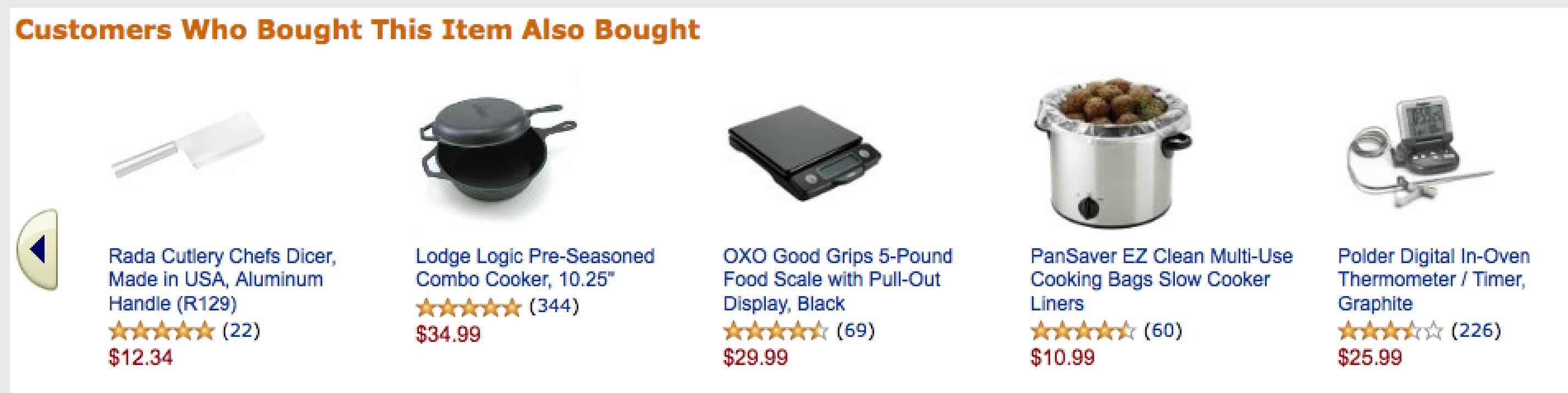
T

6. Filtro Colaborativo

- Diferentes métodos:

- **Memória:**

- a) Item-Item: “Quem comprou isso também comprou ...”



- b) User-Item: “Usuários semelhantes a você compraram ...”

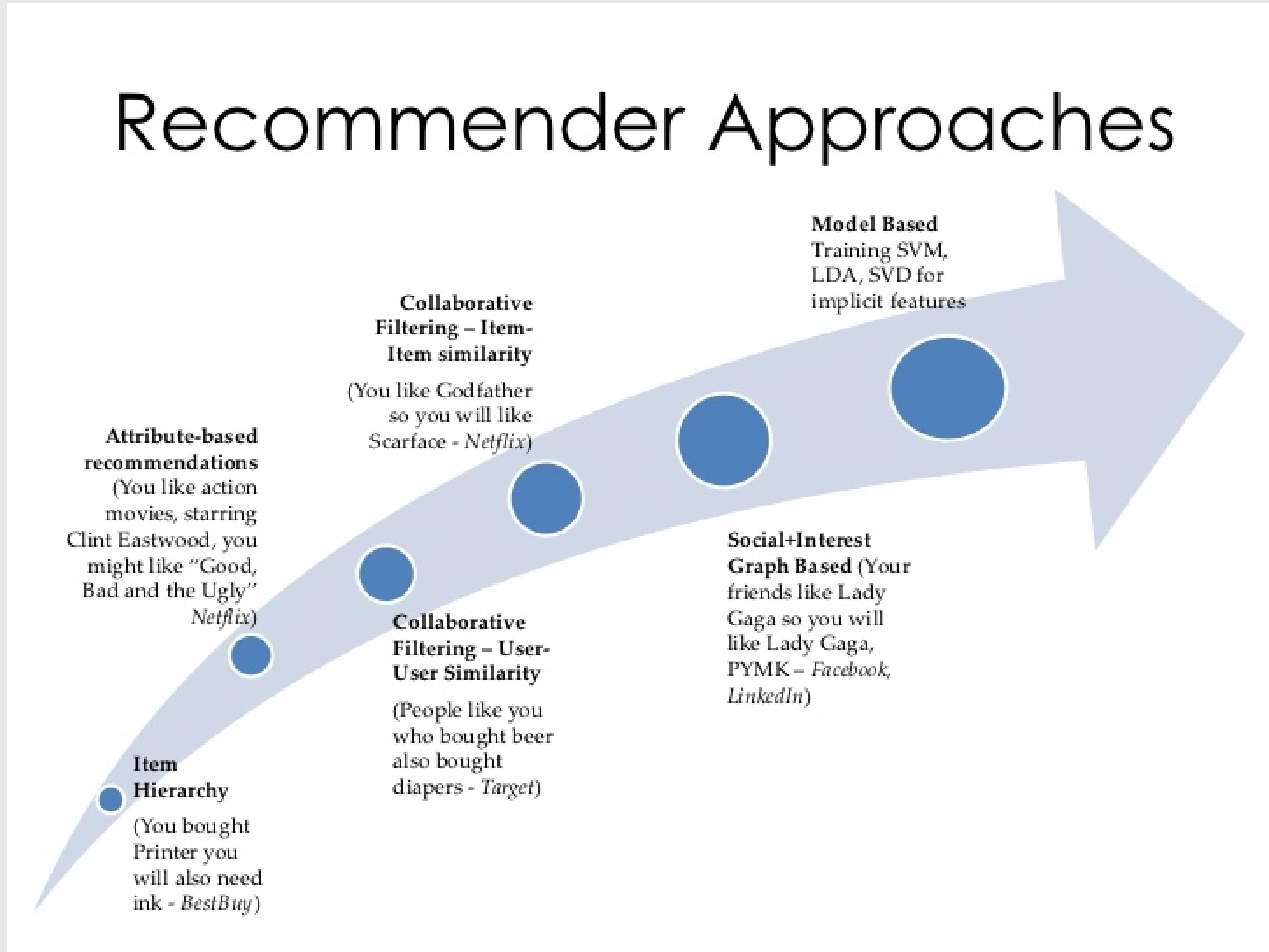
T

6. Filtro Colaborativo

- Diferentes métodos:
 - **Modelo:**
 - a) Fatorização de Matrizes: SVD, PCA, NMF etc

I

6. Filtro Colaborativo



T

6. Filtro Colaborativo

- Case MovieLens:

T

OBRIGADO!