

**Tera**

# AULA 28

Unsupervised Learning:  
Topic Analysis

Instrutor: [Raphael Ballet](#)

## Background:

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

## Interesses:



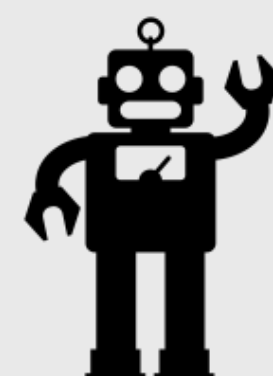
[Drones](#)



[Aprendizado de Máquina](#)



[Processamento de Linguagem Natural](#)



[Robótica](#)



[Visão Computacional](#)

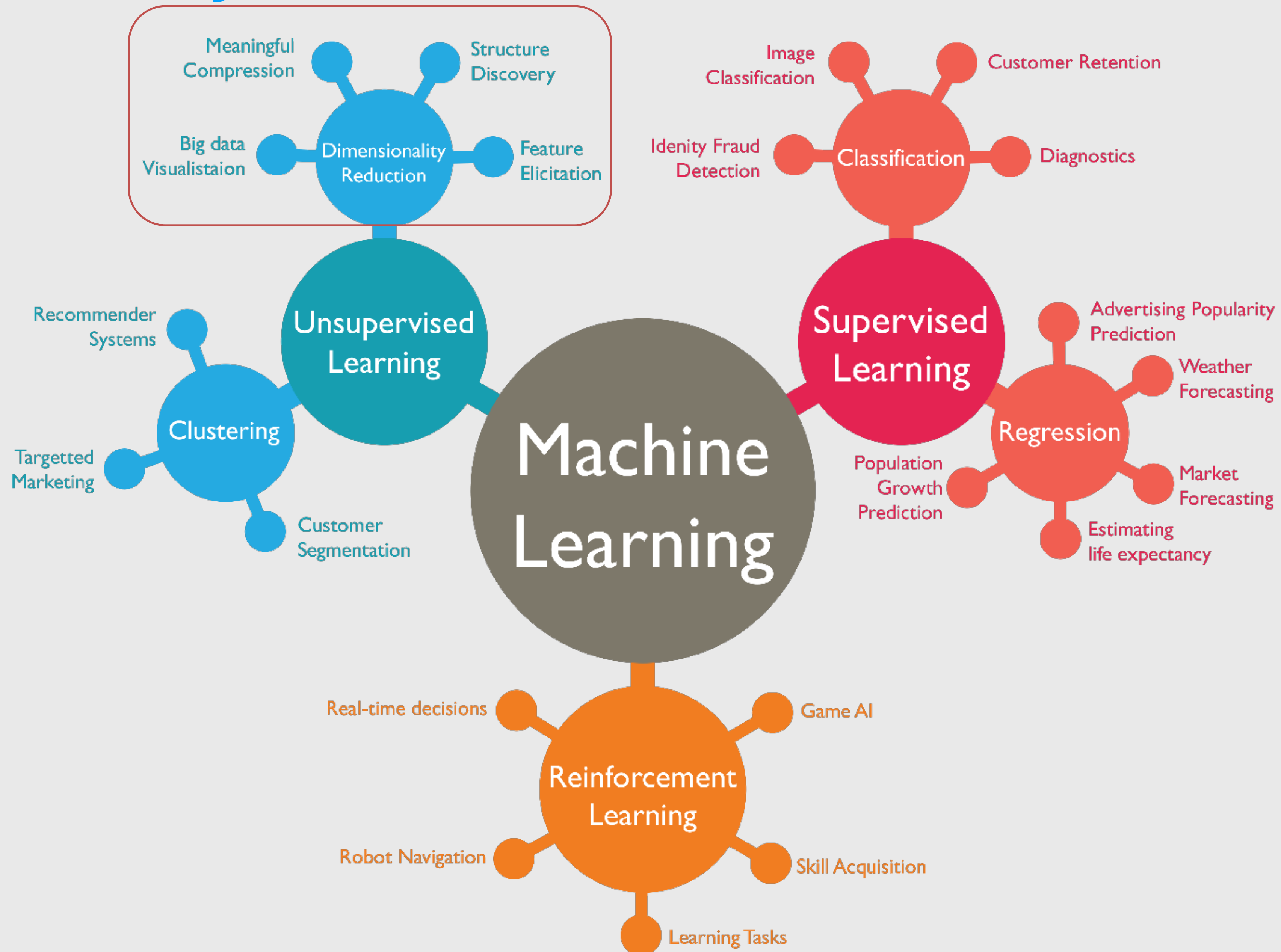


[Sistemas de recomendação](#)

# Planejamento:

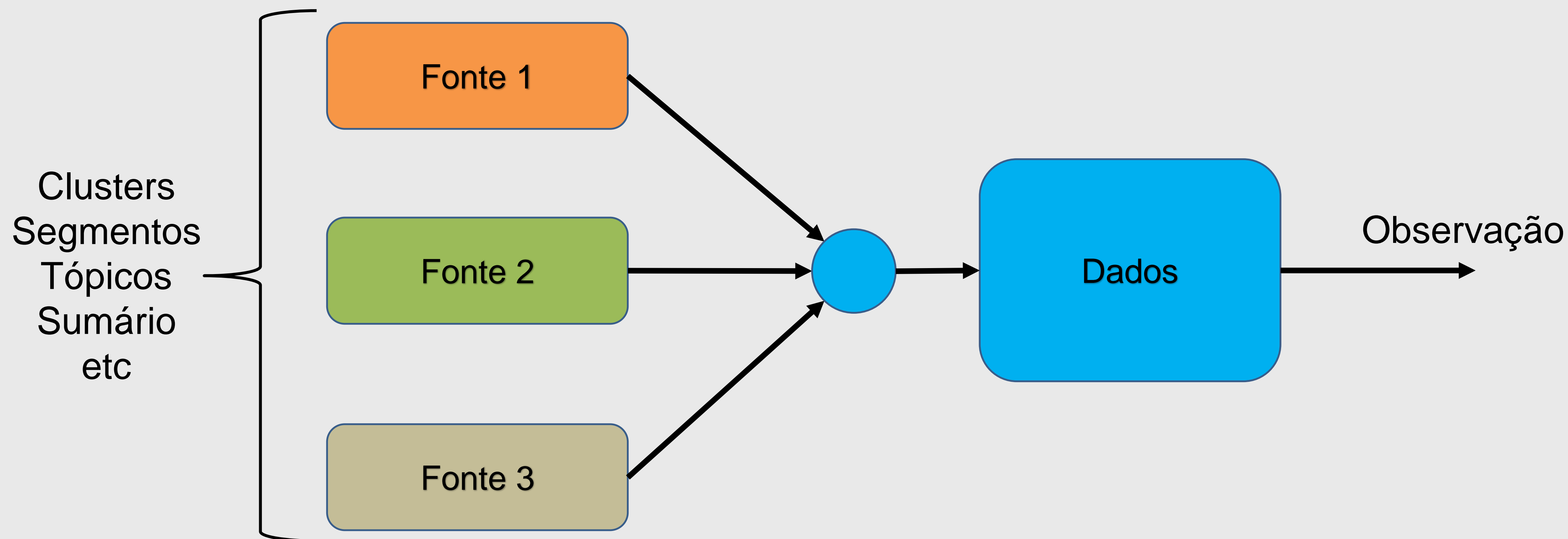
1. Introdução
2. Principal Component Analysis (PCA)
3. T-SNE
4. Topic Analysis:
  - a) Non-Negative Matrix Factorization (NMF)
  - b) Latent Dirichlet Allocation (LDA)
5. Sistemas de recomendação
6. Case

# 1. INTRODUÇÃO



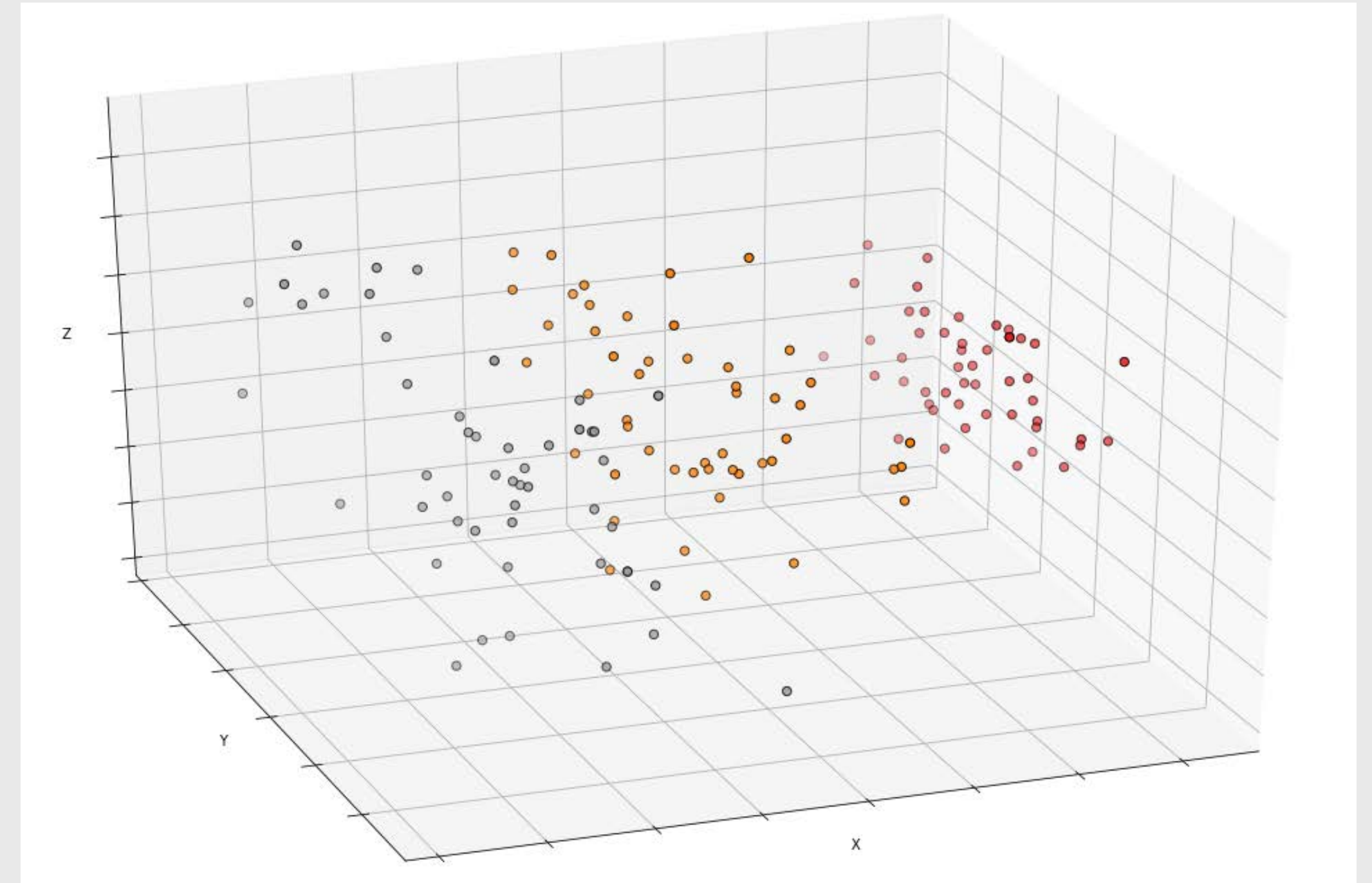
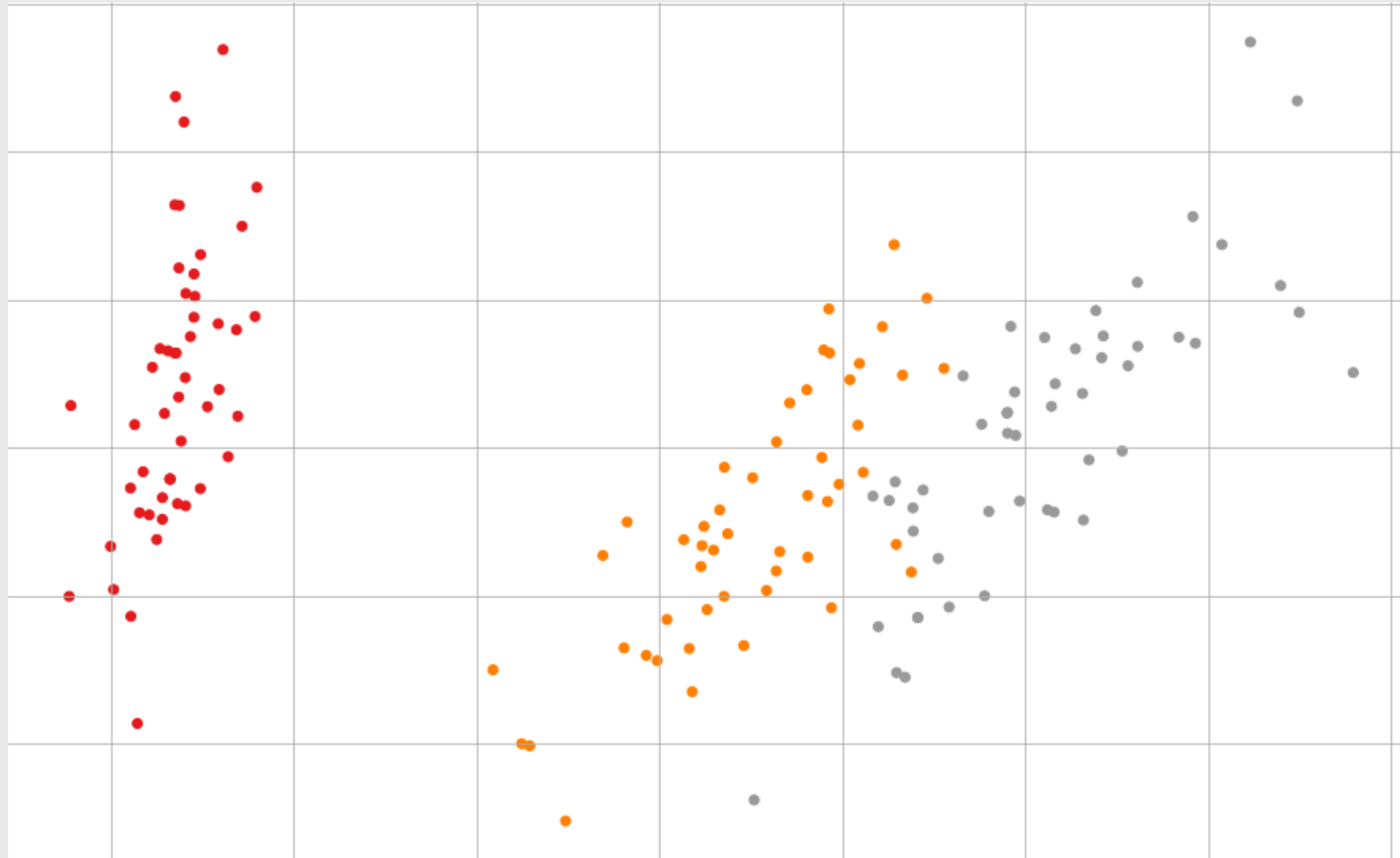
# 1. INTRODUÇÃO

- Documentos



# 1. INTRODUÇÃO

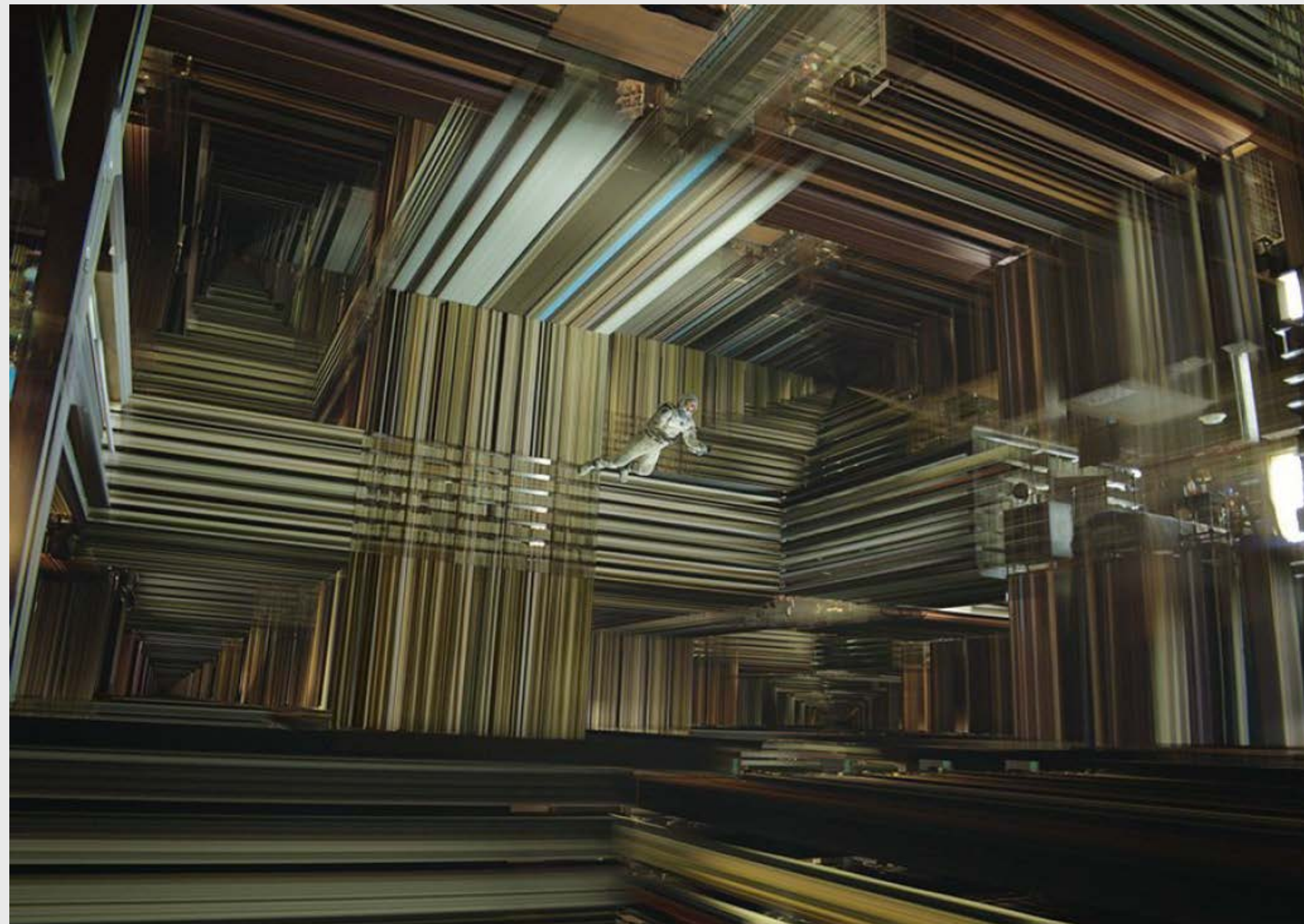
- Visualização dos dados – 2D e 3D





# 1. INTRODUÇÃO

- Visualização mais do que 3D?





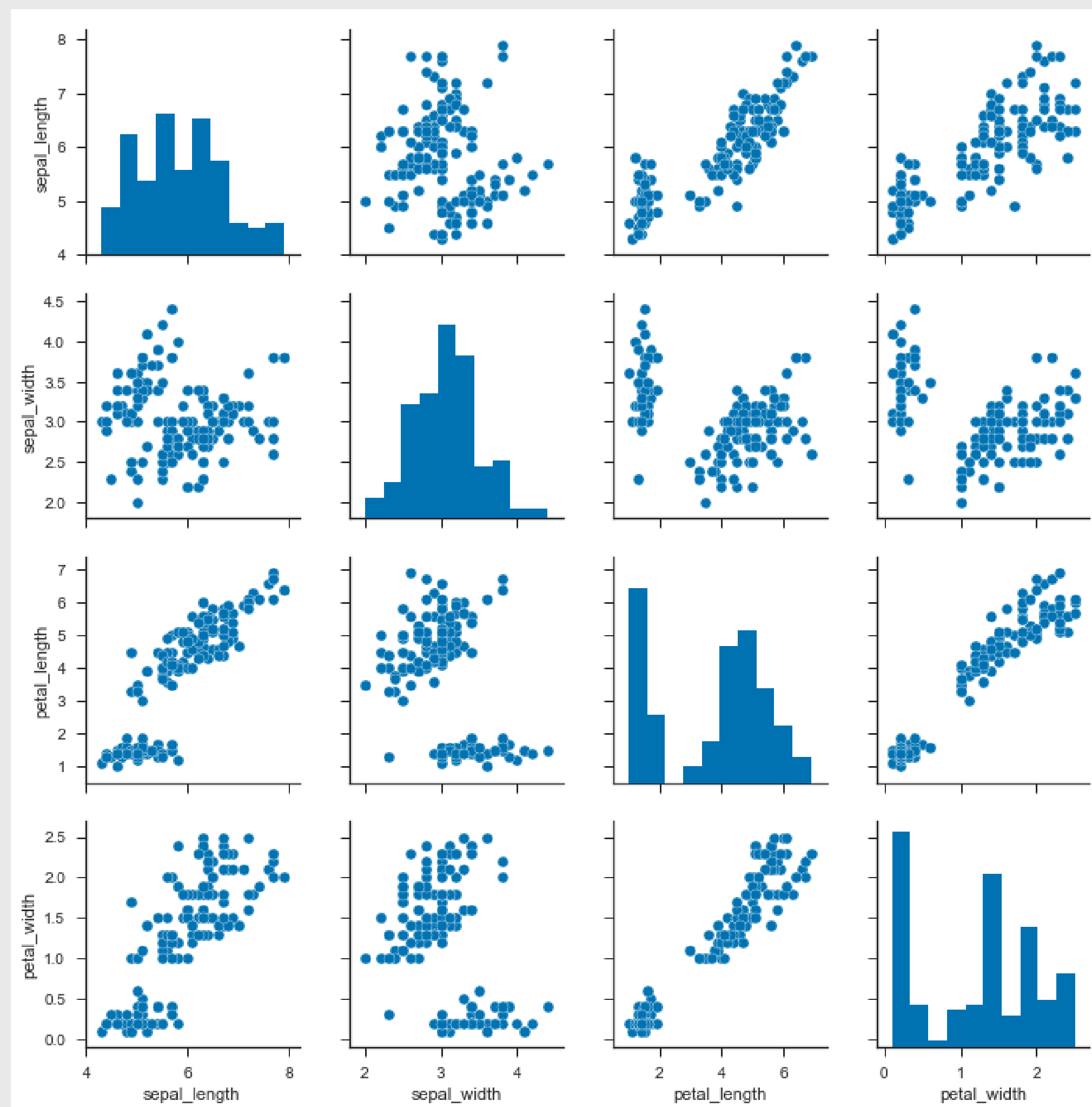
T

# 1. INTRODUÇÃO

- Visualização dos dados: notebook

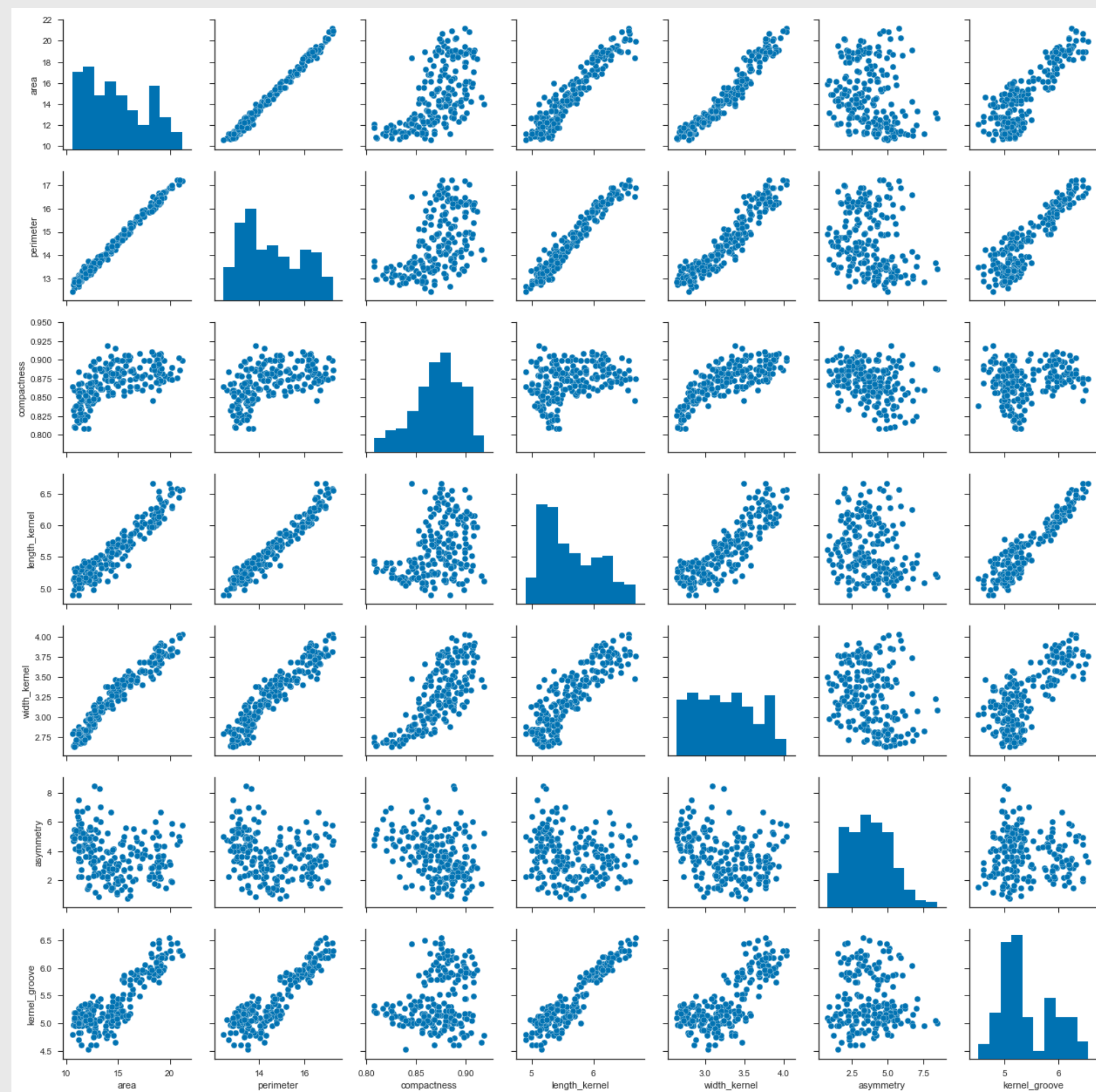
# 1. INTRODUÇÃO

- Visualização 4 dimensões



# 1. INTRODUÇÃO

- Visualização 7 dimensões



# 1. INTRODUÇÃO

- Problemas de NLP
  - Vetor de atributos do texto:

|                                       |   |
|---------------------------------------|---|
|                                       | aardvark<br>apple<br>.<br>.<br>.                  |
| document0<br>document1<br>.<br>.<br>. | 0, 0.1, ... 0.<br><br>word frequencies ("tf-idf") |



T

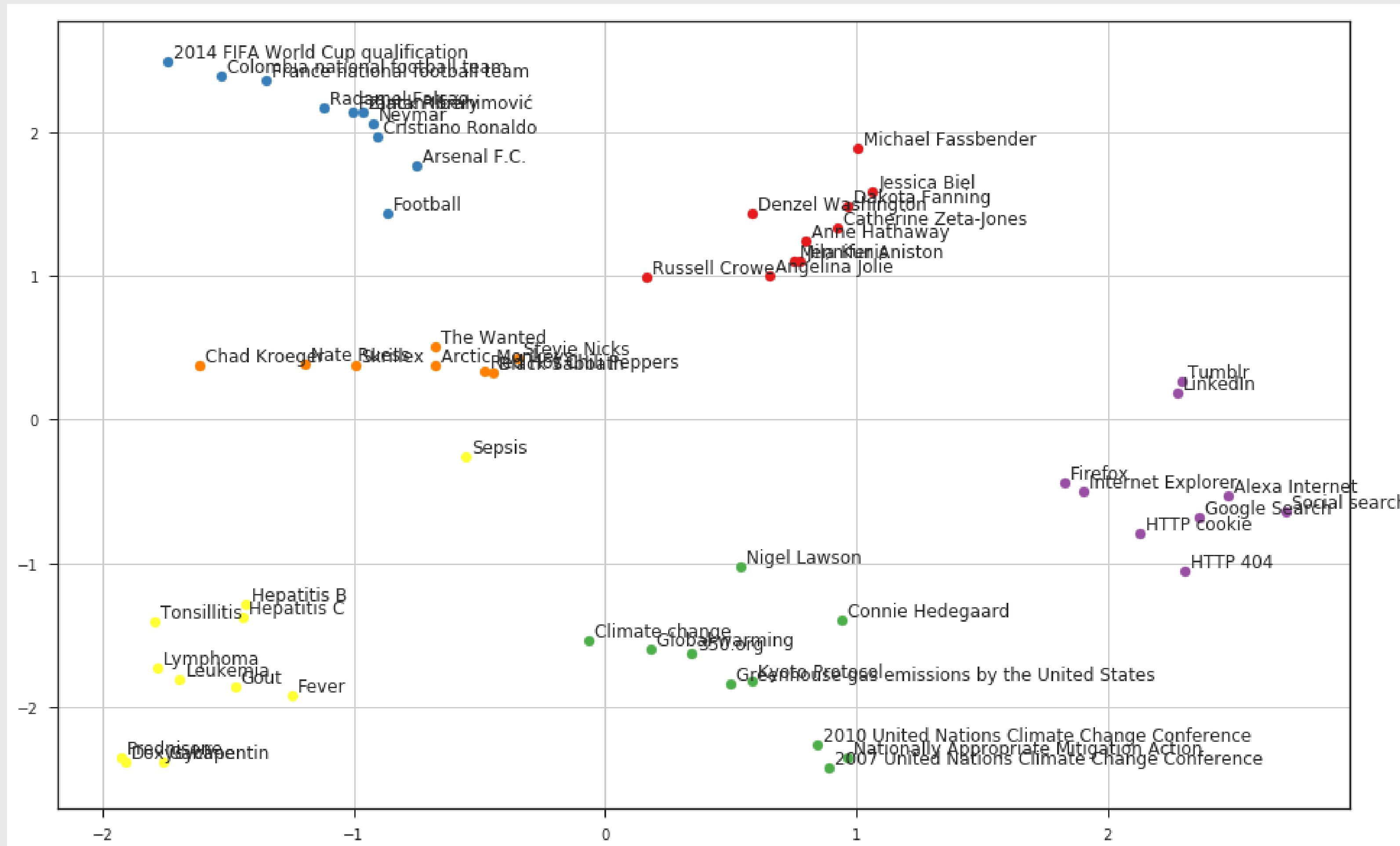
# 1. INTRODUÇÃO

- Visualização NLP (>1k dimensões)



# 1. INTRODUÇÃO

- Visualização NLP (>10k dimensões): **Desejo**

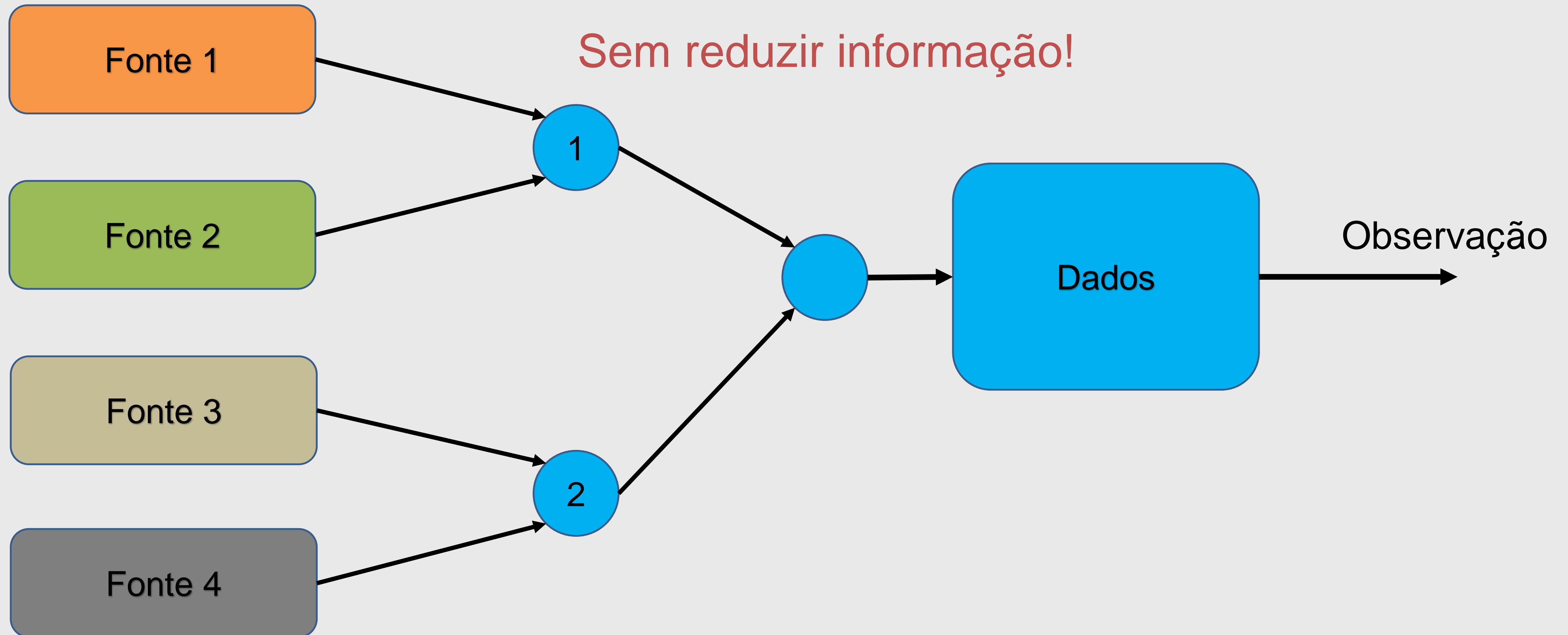


# 1. INTRODUÇÃO

- Objetivo

Sumário – Menor dimensão

Sem reduzir informação!



## 2. Redução de Dimensionalidade

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



## 2. Redução de Dimensionalidade

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

Maldição da dimensionalidade



## 2. Redução de Dimensionalidade

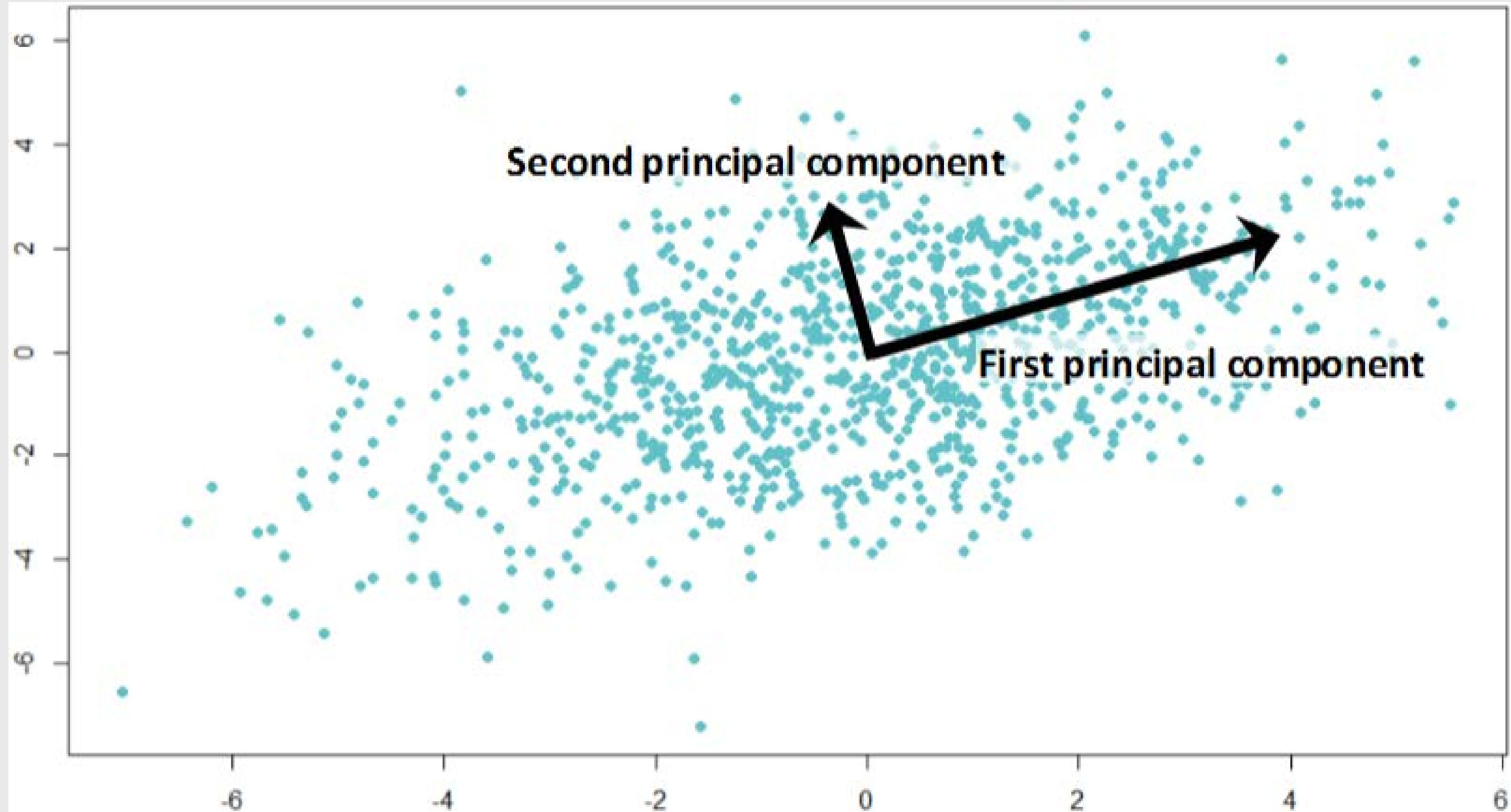
- Técnicas principais:
  - PCA
  - T-SNE (Visualização apenas)
  - Topic Analysis (NMF e LDA)

## 2. PCA

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

## 2. PCA

- Componentes principais:





## 2. PCA

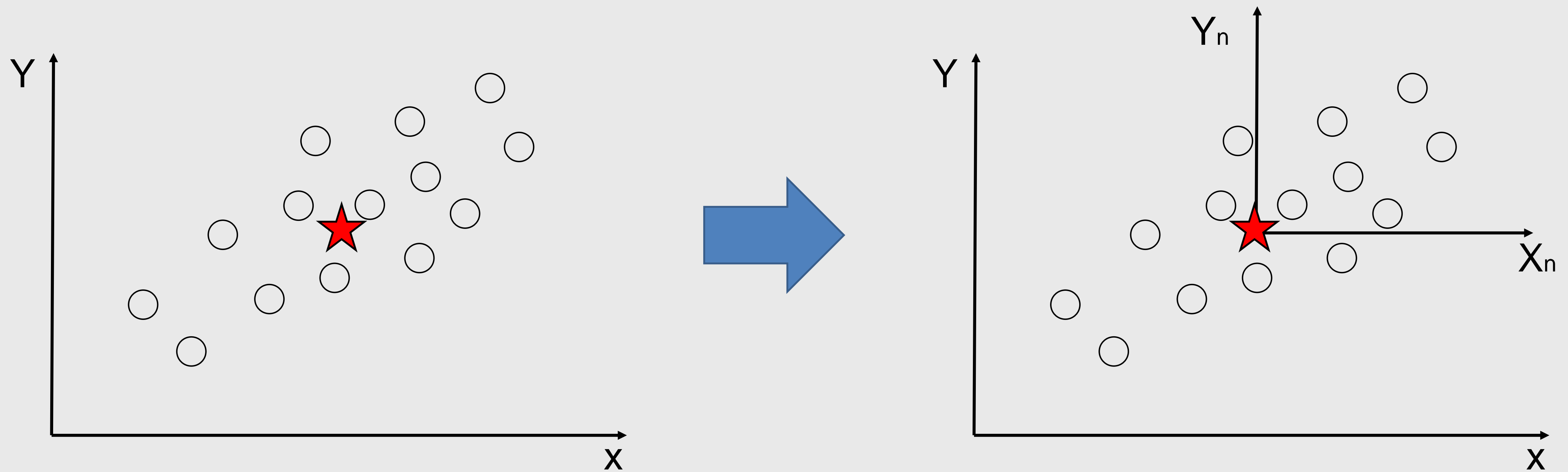
- PCA: Componentes principais:
  - **Primeiro componente:**
    - Direção (combinação linear) de maior variação nos dados
  - **Segundo componente:**
    - Direção da segunda maior variação e ortogonal ao primeiro (descorrelacionado)
  - ...

## 2. 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)

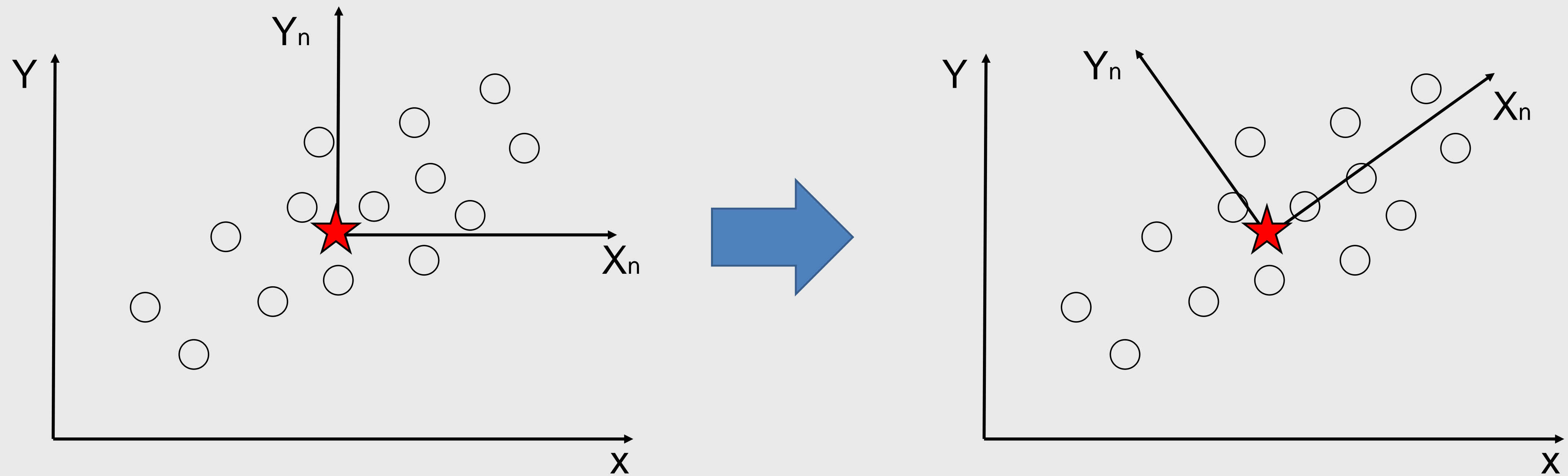
## 2. PCA

1) Remove média amostral dos dados



## 2. PCA

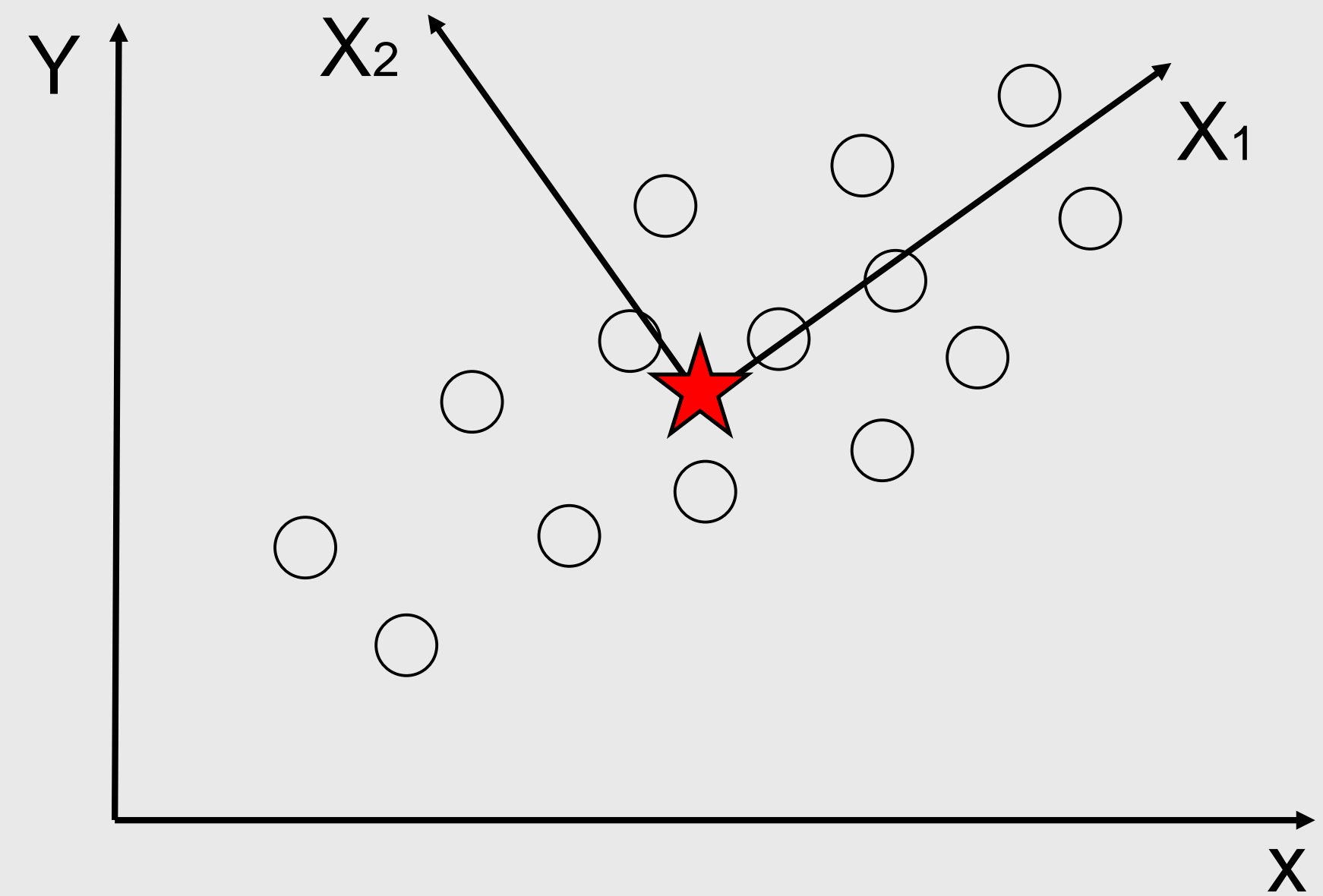
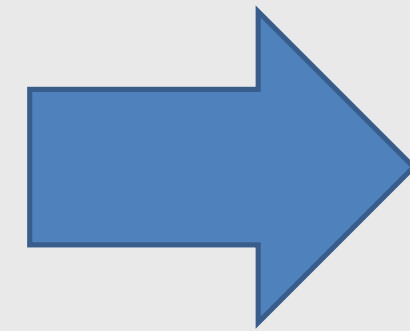
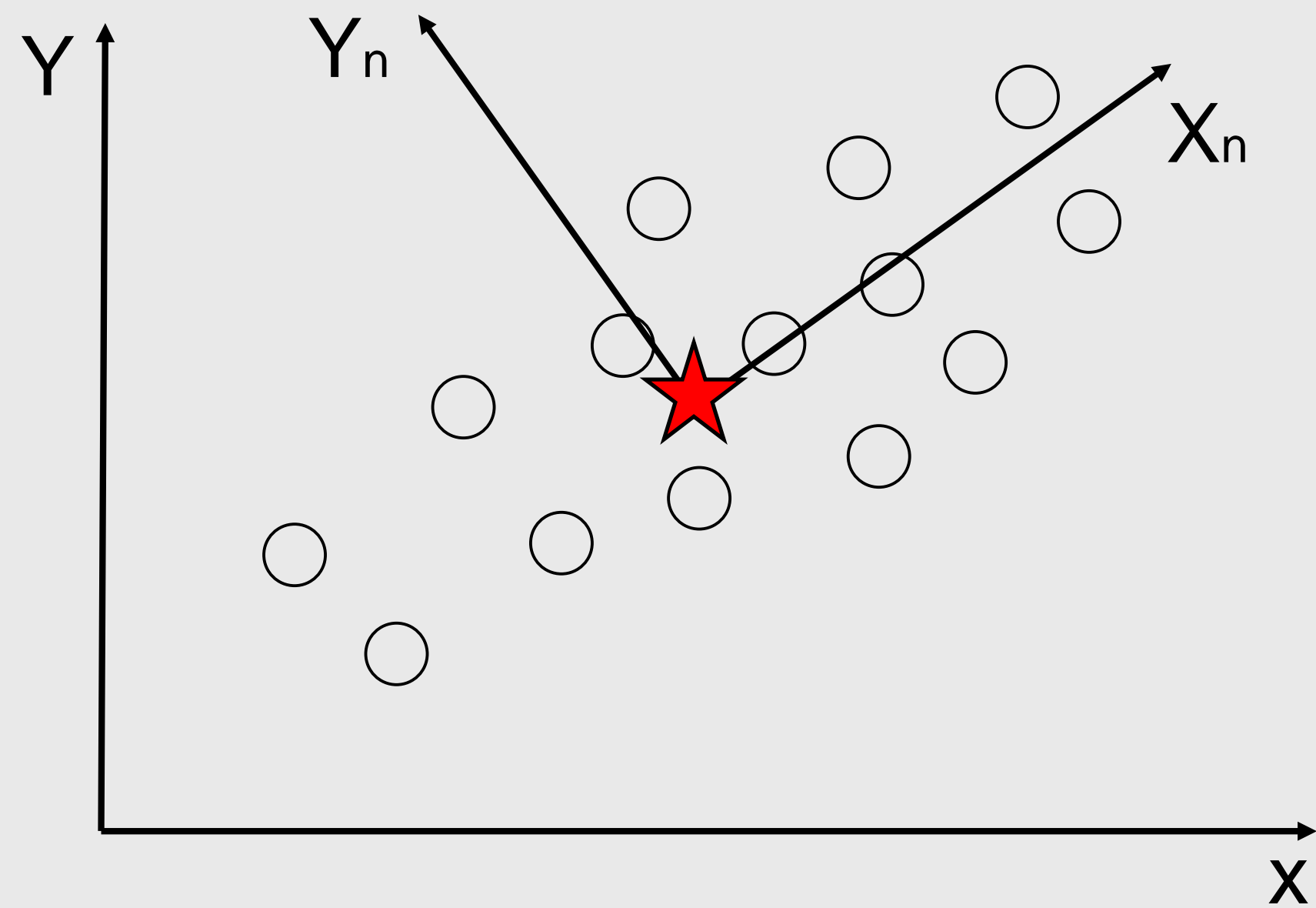
2) Rotaciona os eixos para descorrelacionar os atributos





## 2. PCA

3) Ordena os componentes principais em nível de variância



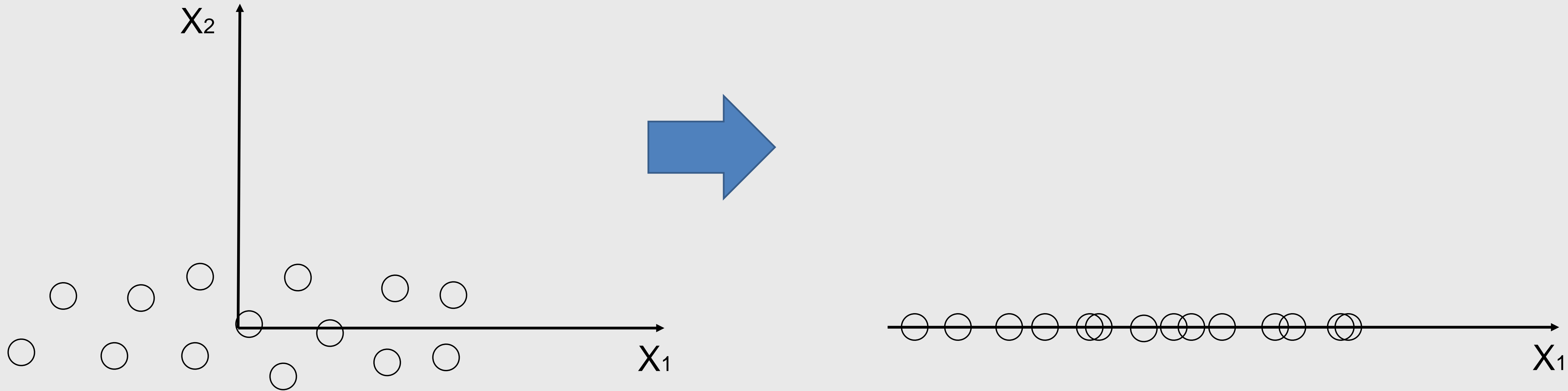
$X_1$ : Primeiro componente principal

$X_2$ : Segundo componente principal

T

## 2. PCA

4) Remove os componentes menos variantes (Opcional)



$X_1$ : Primeiro componente principal

$X_2$ : Segundo componente principal

## 2. PCA

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



## 2. PCA

- Exemplo: notebook

## 2. PCA

- **Vantagens:**

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

## 2. 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



## 3. Visualização: T-SNE

- **T**-distributed **S**tochastic **N**eighbor **E**mbedding
  - 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

## 3. T-SNE

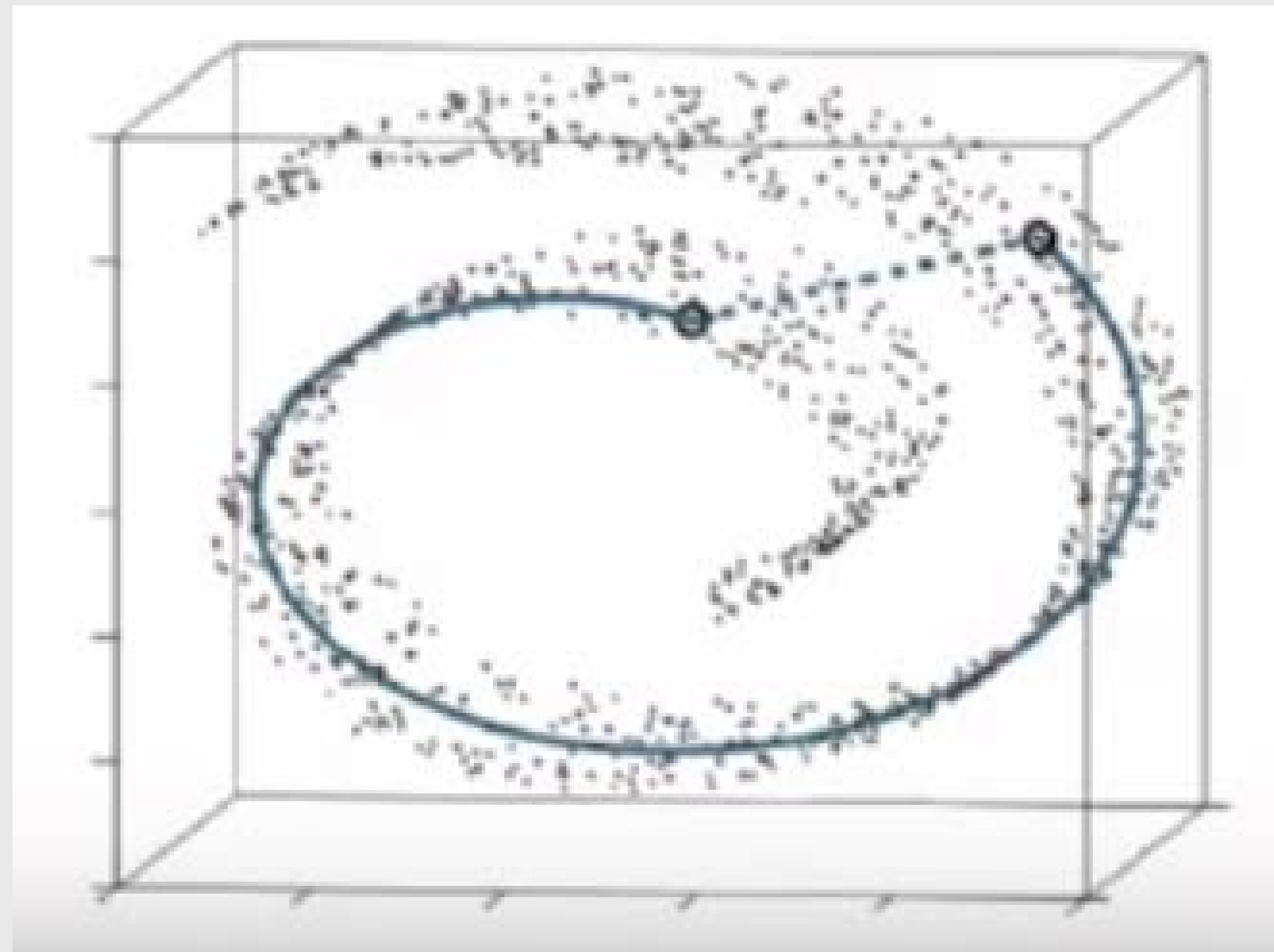
- Mas, nós já temos o PCA. Por que usar T-SNE?

## 3. T-SNE

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

## 3. T-SNE

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



## 3. 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

## 3. T-SNE

- Exemplo: notebook
- [Projeto T-SNE - Tensorflow](#)



## 3. 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

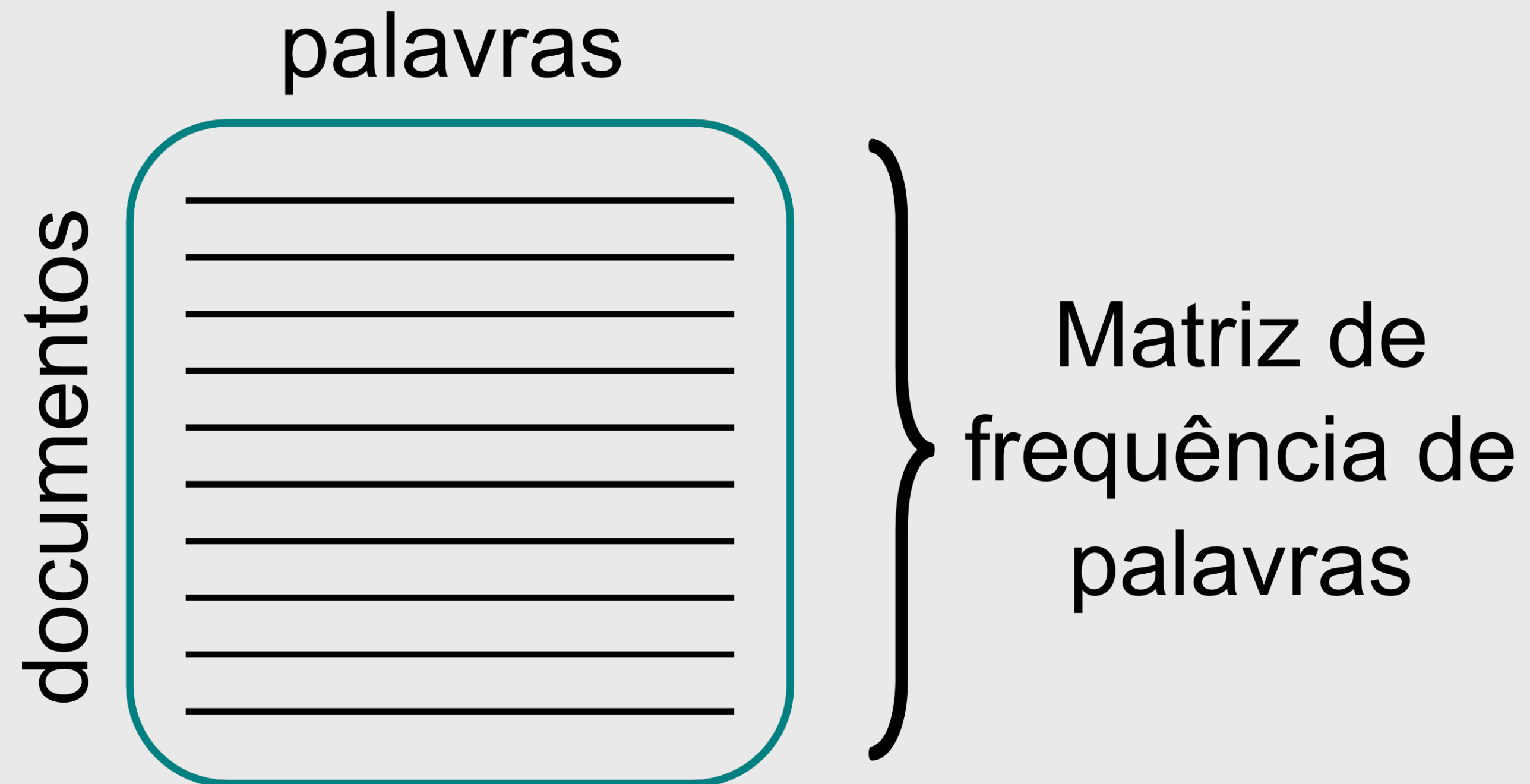
## 3. T-SNE

- **Desvantagens:**

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

## 4. Topic Analysis

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



# 4. Topic Analysis

- **Solução:**
  - Redução de dimensionalidade
  - Clustering
  - ...

## 4. Topic Analysis

- **Solução:**

- Redução de dimensionalidade
- Clustering

- **Problema:**

- **Perda de interpretabilidade dos dados**

# 4. Topic Analysis

## • Objetivo:

- Encontrar estrutura implícita nos documentos – Tópicos / Temas

The screenshot shows the Wikipedia article for "Dengue fever". The article is a featured article. The main text describes the disease, its symptoms, and its transmission. A table of classification and external resources is provided, including ICD-10, ICD-9, DiseasesDB, MedlinePlus, eMedicine, and MeSH. A diagram of symptoms is also shown, illustrating the febrile phase (sudden-onset fever, headache, mouth and nose bleeding) and the critical phase (hypotension, pleural effusion, ascites, gastrointestinal bleeding).

**Dengue fever**  
A featured article from Wikipedia, the free encyclopedia

*For other uses, see [Dengue fever \(disambiguation\)](#).*

**Dengue fever** (UK /dɛnɡɛt/ or US /dɛnɡiː/), also known as **breakbone fever**, is an infectious [tropical disease](#) caused by the [dengue virus](#). Symptoms include [fever](#), [headache](#), [muscle](#) and [joint pains](#), and a characteristic [skin rash](#) that is [similar to measles](#). In a small proportion of cases the disease develops into the life-threatening **dengue hemorrhagic fever**, resulting in [bleeding](#), [low levels of blood platelets](#) and blood plasma leakage, or into **dengue shock syndrome**, where [dangerously low blood pressure](#) occurs.

Dengue is transmitted by several species of [mosquito](#) within the [genus](#) *Aedes*, principally *A. aegypti*. The virus has four different types; infection with one type usually gives lifelong [immunity](#) to that type, but only short-term immunity to the others. Subsequent infection with a different type increases the risk of severe complications. As there is no [vaccine](#), prevention is sought by reducing the habitat and the number of mosquitoes and limiting exposure to bites.

Treatment of acute dengue is supportive, using either oral or intravenous [rehydration](#) for mild or moderate disease, and [intravenous fluids](#) and [blood transfusion](#) for more severe cases. The [incidence](#) of dengue fever has increased dramatically since the 1960s, with around 50–100 million people infected yearly. Early descriptions of the condition date from 1779, and its viral cause and the transmission were elucidated in the early 20th century. Dengue has become a global problem since the [Second World War](#) and is [endemic](#) in more than 110 countries. Apart from eliminating the mosquitoes, work is ongoing on a vaccine, as well as medication targeted directly at the virus.

**Contents** [\[show\]](#)

### Signs and symptoms [\[edit\]](#)

Typically, people infected with dengue virus are [asymptomatic](#) (80%) or only have mild symptoms such as an uncomplicated fever.<sup>[1][2][3]</sup> Others have more severe illness (5%), and in a small proportion it is life-threatening.<sup>[1][3]</sup> The [incubation period](#) (time between exposure and onset of symptoms) ranges from 3–14 days, but most often it is 4–7 days.<sup>[4]</sup> Therefore, travelers returning from endemic areas are unlikely to have dengue if fever or other symptoms start more than 14 days after arriving home.<sup>[5]</sup> Children often experience symptoms similar to those of the [common cold](#) and [gastroenteritis](#) (vomiting and diarrhea),<sup>[6]</sup> and generally have less severe symptoms than adults,<sup>[7]</sup> but are more susceptible to the severe complications.<sup>[5]</sup>

### Clinical course [\[edit\]](#)

The characteristic symptoms of dengue are sudden-onset fever, headache (typically located behind the eyes), muscle and joint pains, and a rash. The alternative name for dengue, "break-bone fever", comes from the associated muscle and joint pains.<sup>[1][8]</sup> The course of infection is divided into three phases: febrile, critical, and recovery.<sup>[9]</sup>

The febrile phase involves high fever, often over 40 °C (104 °F), and is associated with generalized pain and a headache;<sup>[1][10]</sup>

| Dengue fever   |                                   |
|--|-----------------------------------|
| Classification and external resources                          |                                   |
| <span></span> <div>The typical rash seen in dengue fever</div> |                                   |
| <b>ICD-10</b>  | A90 <a href="#">↗</a>             |
| <b>ICD-9</b>   | 061 <a href="#">↗</a>             |
| <b>DiseasesDB</b>  | 3564 <a href="#">↗</a>            |
| <b>MedlinePlus</b>   | 001374 <a href="#">↗</a>          |
| <b>eMedicine</b>   | med/528 <a href="#">↗</a>         |
| <b>MeSH</b>  | C02.782.417.214 <a href="#">↗</a> |

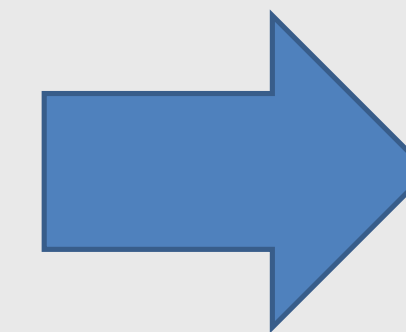
Symptoms of **Dengue fever**

Febrile phase

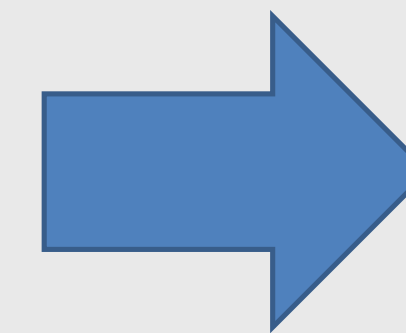
- sudden-onset fever
- headache
- mouth and nose bleeding

Critical phase

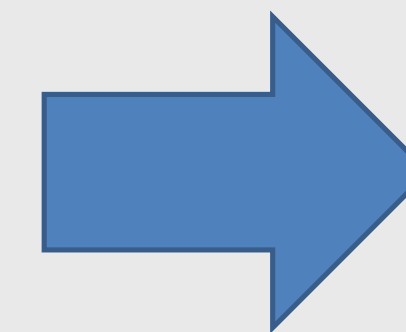
- hypotension
- pleural effusion
- ascites
- gastrointestinal bleeding



75% Health



20% Medicine



5% Biology



## 4. Topic Analysis

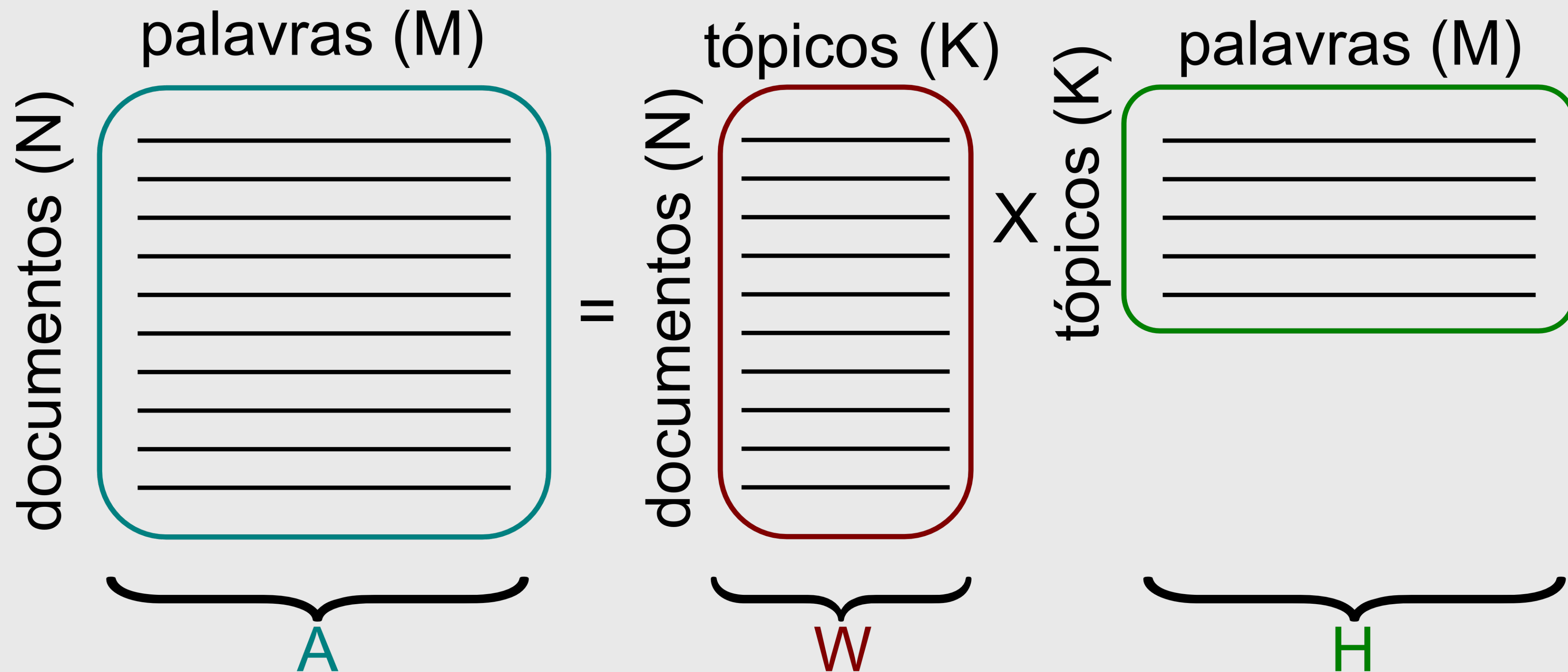
- Principais algoritmos:
  - Non-Negative Matrix Factorization (NMF)
  - Latent Dirichlet Allocation (LDA)

## 4. Topic Analysis – NMF

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

## 4. Topic Analysis – NMF

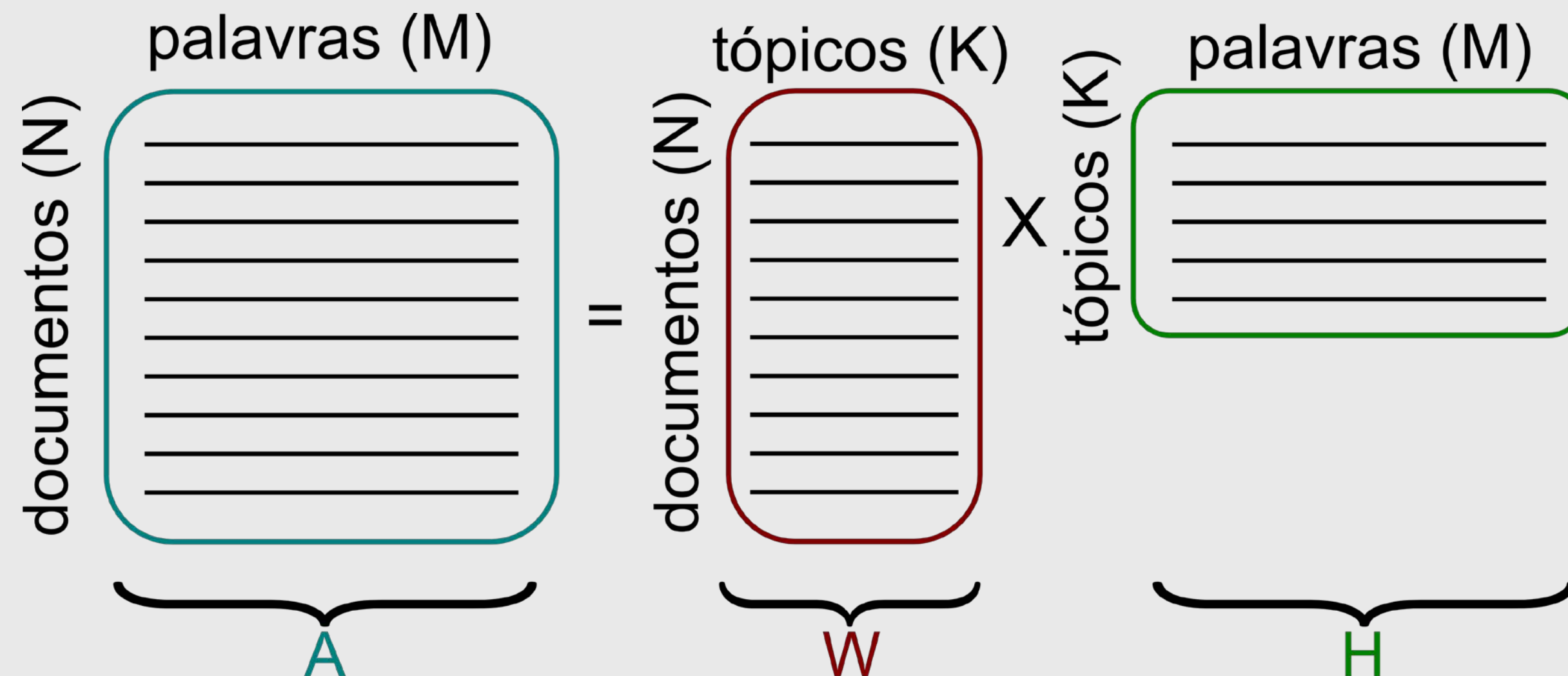
- NMF: Fatoração  $\rightarrow A = WH$



## 4. 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



## 4. 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.)

## 4. 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

## 4. Topic Analysis – NMF

- Exemplo: notebook



## 4. Topic Analysis – NMF

- **Vantagens:**

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

## 4. Topic Analysis – NMF

- **Desvantagens:**

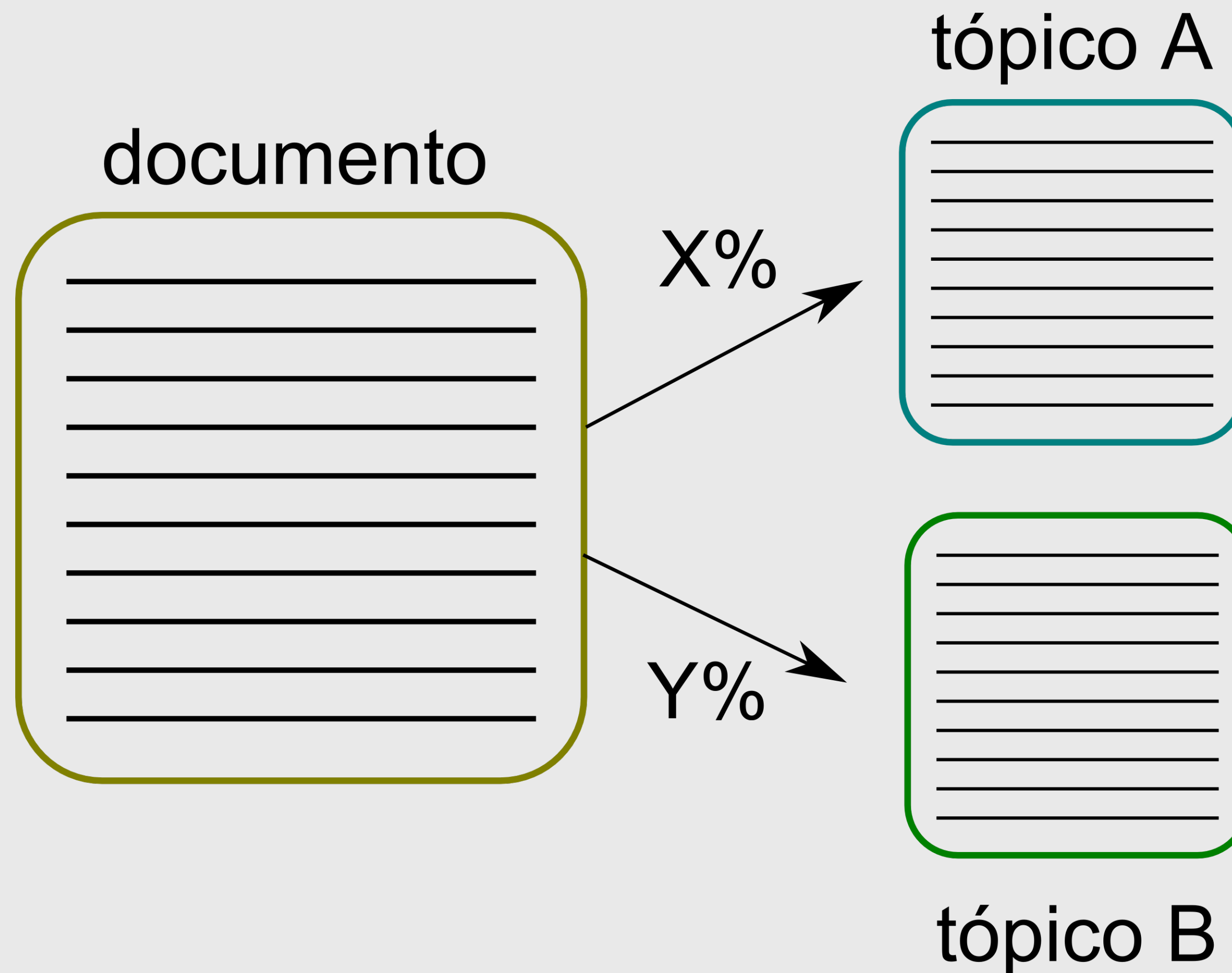
- Solução aproximada
- Pode causar overfitting
- Limitação de utilizar apenas features positivas

## 4. Topic Analysis – LDA

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

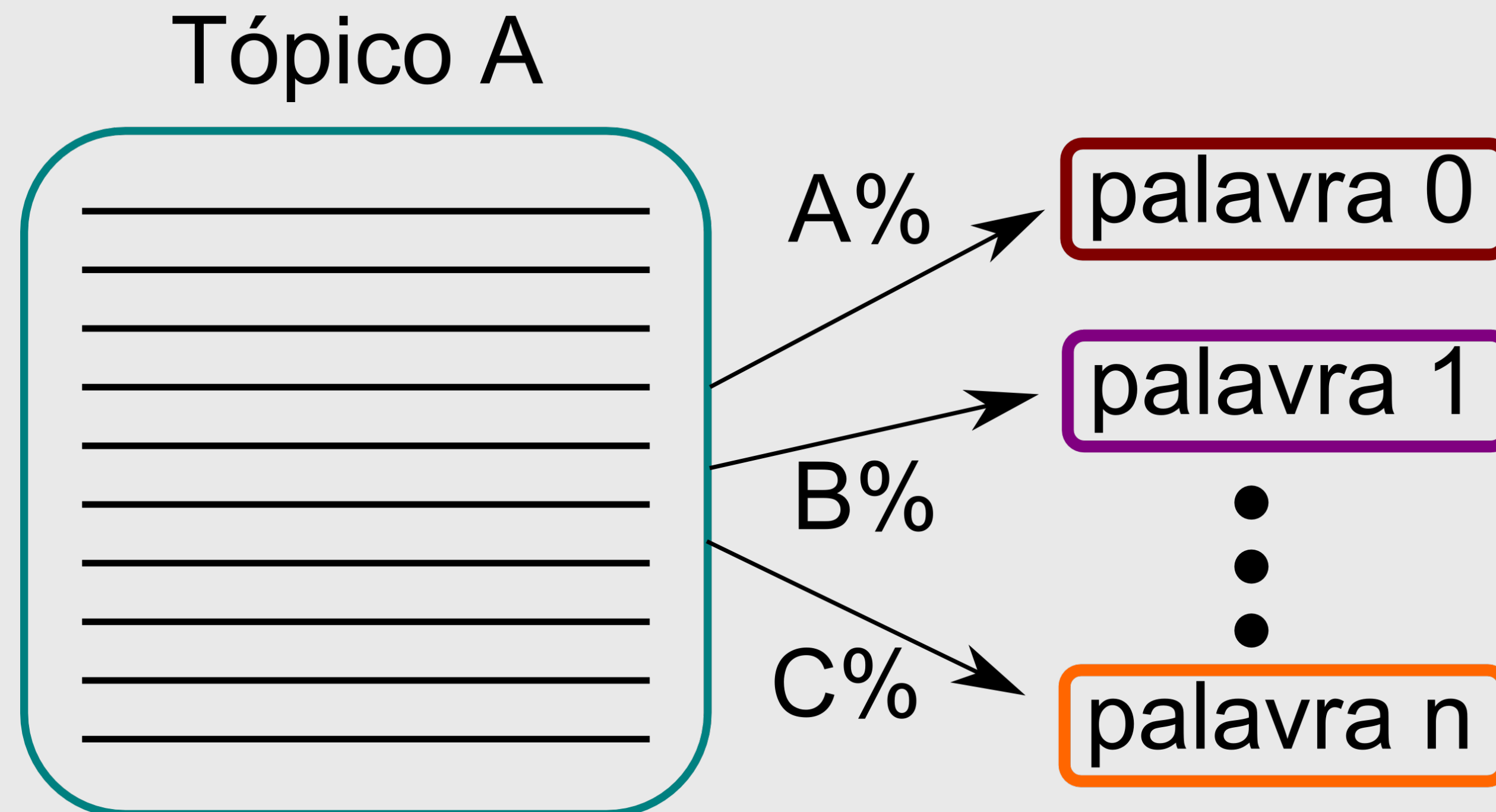
## 4. Topic Analysis – LDA

- Documento → Mistura de tópicos



## 4. Topic Analysis – LDA

- Tópicos → Mistura de palavras



## 4. 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 |

## 4. 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

## 4. 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



## 4. 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?
  - $P(\text{'Fish'} \mid \text{tópico A}) = 0.75$  ( $3 - A, 1 - B$ )
  - $P(\text{'Fish'} \mid \text{tópico B}) = 0.25$

## 4. 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?
  - $P(\text{tópico A} \mid \text{Documento 3}) = P(\text{tópico B} \mid \text{Documento 3}) = 0.5$

## 4. 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.

## 4. Topic Analysis – LDA

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

## 4. Topic Analysis – LDA

- Método gerador:
  - Supõe que os documentos são gerados por um modelo probabilístico
  - Objetivo é aproximar esse modelo

## 4. Topic Analysis – LDA

- Método gerador: Vantagens
  - Podemos amostrar a partir do modelo encontrado
  - Em outras palavras, podemos gerar novos documentos “artificiais”

T

## 4. Topic Analysis – LDA

- Exemplo: notebook

## 4. Topic Analysis – LDA

- **Vantagens:**

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



## 4. Topic Analysis – LDA

- **Desvantagens:**

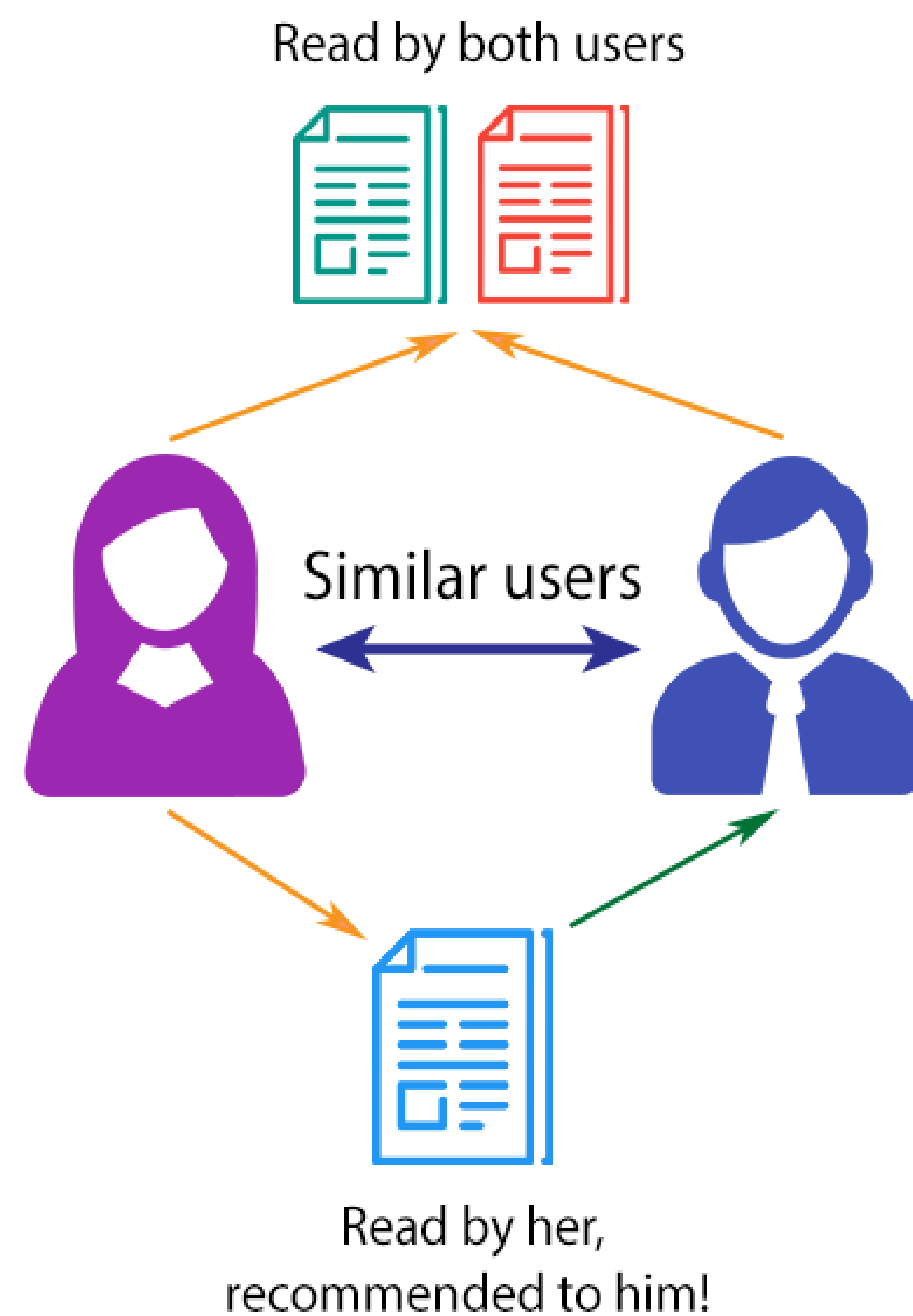
- Mesmas desvantagens do NMF

## 5. Sistemas de Recomendação

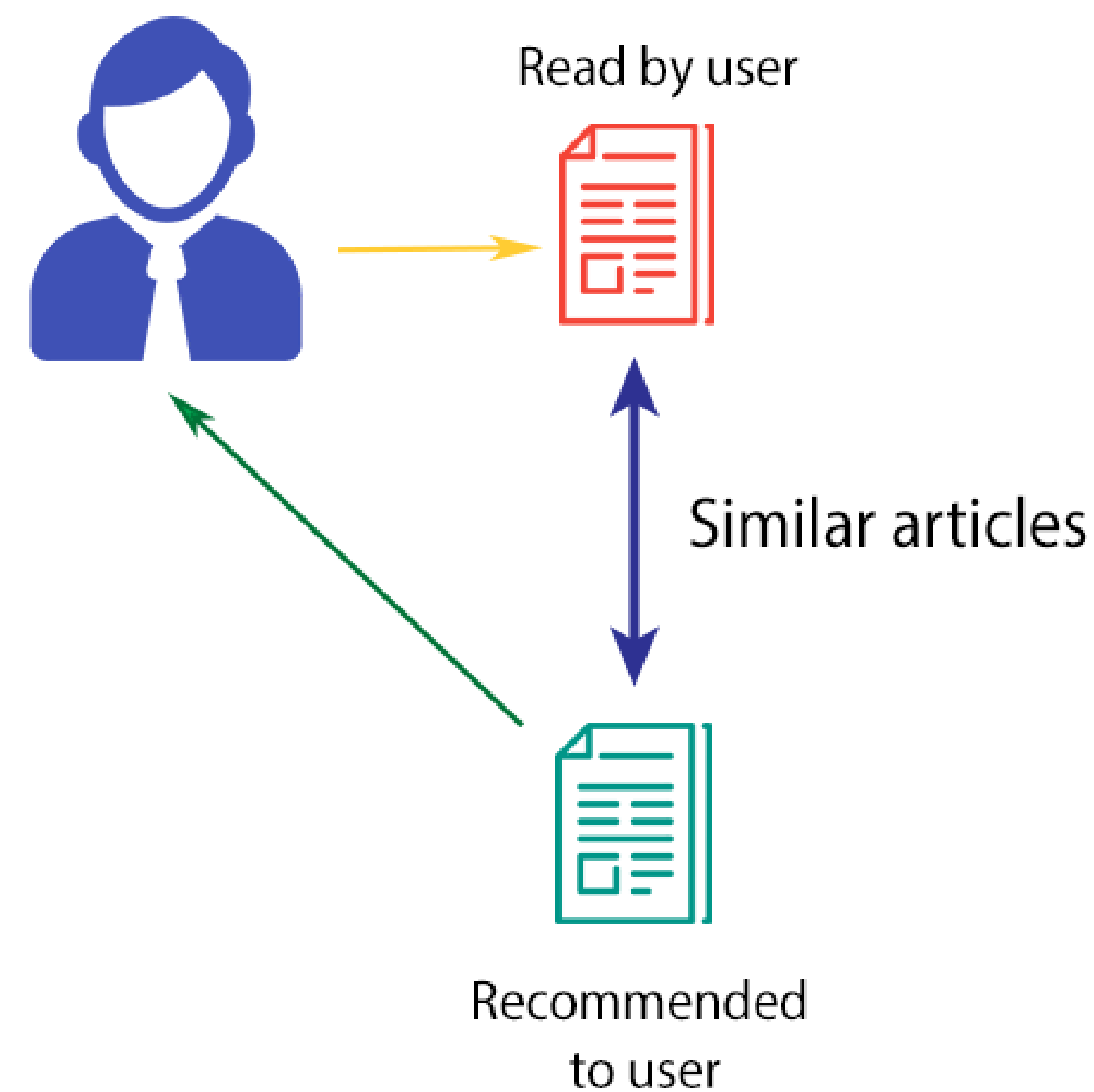
- Existem 2 grandes grupos:
  - Proximidade de documentos (produtos, músicas, filmes etc)
  - Proximidade entre usuários (filtro colaborativo)

## 5. Sistemas de Recomendação

### COLLABORATIVE FILTERING



### CONTENT-BASED FILTERING

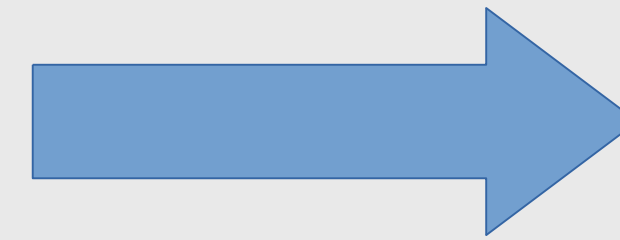


## 5. Sistemas de Recomendação

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

## 5. Sistemas de Recomendação

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



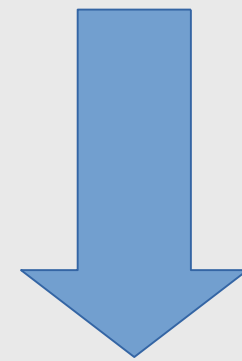
**Clustering / Topic Analysis**

## 5. Sistemas de Recomendação

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

## 5. Sistemas de Recomendação

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



**Clustering Topic  
Analysis**

## 5. Sistemas de Recomendação

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




## 5. Sistemas de Recomendação

- Exemplo: Recomendação de filmes

Close

### Other Movies You Might Enjoy

[Amelie](#)

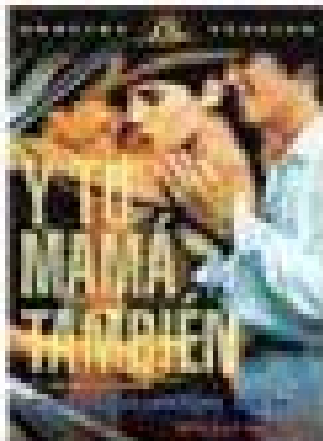


**Add**

★★★★☆

Not Interested

[Y Tu Mama Tambien](#)




**Add**

★★★★☆

Not Interested

[Guys and Balls](#)

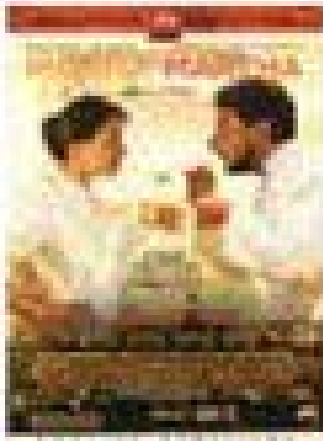


**Add**

★★★★☆

Not Interested

[Mostly Martha](#)

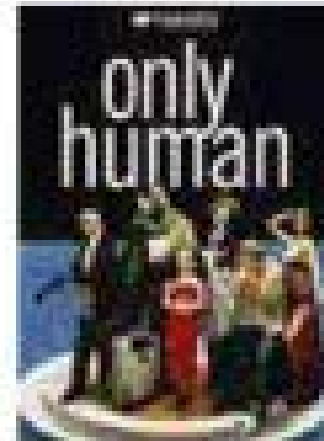


**Add**

★★★★☆

Not Interested

[Only Human](#)

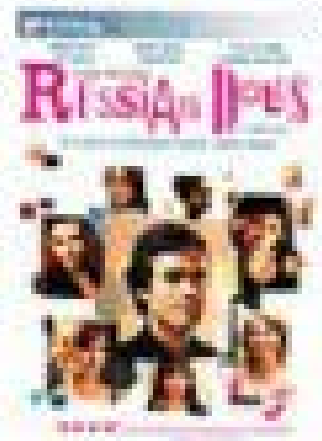


**Add**

★★★★☆

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

## 5. Sistemas de Recomendação

- Exemplo: Recomendação de filmes

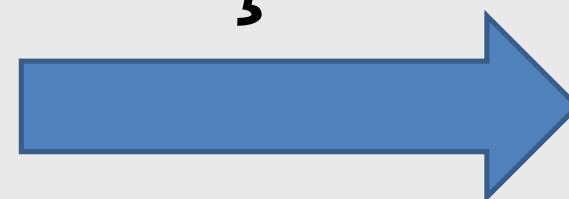


|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 2 |   |   | 4 | 5 |   |
| 5 |   | 4 |   |   | 1 |
|   |   | 5 |   | 2 |   |
|   | 1 |   | 5 |   | 4 |
|   |   | 4 |   |   | 2 |
| 4 | 5 |   | 1 |   |   |

## 5. Sistemas de Recomendação

- Exemplo: Recomendação de filmes

Recomendação!



|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
|    | 2 |   |   | 4 | 5 |   |
|  | 5 |   | 4 |   |   | 1 |
|  |   |   | 5 |   | 2 |   |
|  |   | 1 |   | 5 |   | 4 |
|  |   |   | 4 |   |   | 2 |
|  | 4 | 5 |   | 1 |   |   |

T

OBRIGADO!