

**Marcelo Colunno**  
Cientista de Dados  
Bayer Crop Science

Redução de  
Dimensionalidade

# AGENDA

- **Bloco 1: Introdução**
  - **Abordagens: Data Preparation, Data Visualization, Clustering**
- **Bloco 2: Métodos de Redução de Dimensionalidade**
  - **Lineares (ex. PCA)**
  - **Não lineares (ex. t-SNE, UMAP)**
    - + Intervalo - 10 min
- **Bloco 3: Aplicações Práticas - Jupyter**
- **Dúvidas e reflexões finais**
- **Como foi?**

T



# Introdução

# Definições

- **Dimensionalidade:** número de variáveis de entrada de um modelo, i.e., o número de colunas de um *DataSet* considerado como a quantidade de dimensões ou o número de graus de liberdade
- **Redução de Dimensionalidade:** técnicas que reduzem o número de dimensões de um *DataSet* visando atingir objetivos específicos
- **Maldição da Dimensionalidade:** quanto maior o número de variáveis de entrada, maior é o desafio de modelagem preditiva (*DataSets* muito extensos apresentam problemas, desde dificuldades na captura de padrões como tempo de processamento)
- **Alta Dimensionalidade:** pode ser centenas, milhares ou até milhões de variáveis de entrada



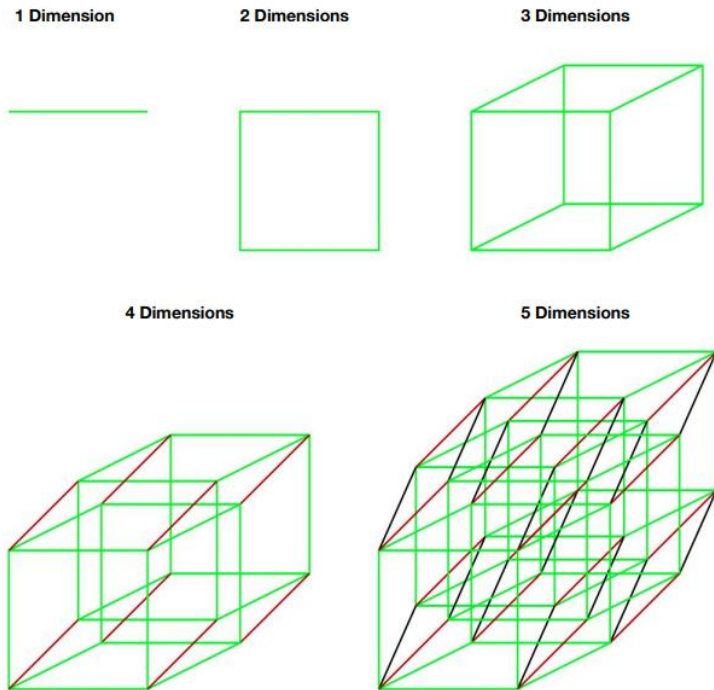
# Benefícios

- **Evitar o problema do *overfitting*:** modelos mais complexos com muitos recursos de entrada tem a tendência de ter problemas com *overfitting* nos dados de treino
- **Remove o ruído dos dados:** eliminar os recursos redundantes melhora a performance do modelo
- **Resolve o problema da multicolinearidade:** combina variáveis independentes altamente correlacionadas entre si em um *DataSet* de variáveis não correlacionadas
- **Muito útil para a visualização de dados:** possibilidade de visualização dos dados em um gráfico 2D ou 3D
- **Pode ser utilizado para compressão de imagens**

**T**

# Benefícios

## Dificuldade de visualizar objetos com mais de 3 dimensões



## One hot encoding: diminuir a complexidade dos dados

[illegible]



# Benefícios

Aumento da **capacidade de processamento** (poder computacional)

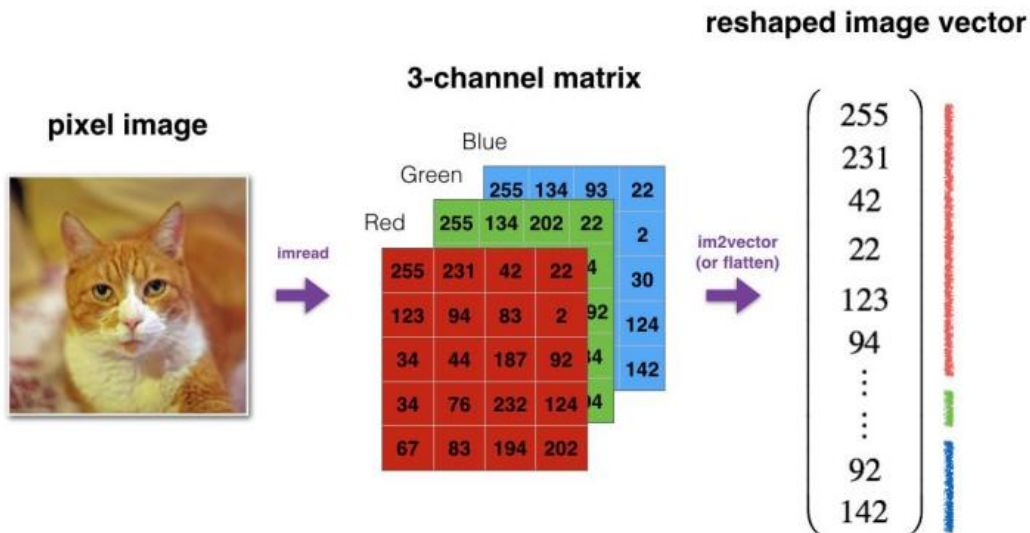


Imagem 100x100 pixels:

Red:  $100 \times 100 = 10.000$  +

Green:  $100 \times 100 = 10.000$  +

Blue:  $100 \times 100 = 10.000$

---

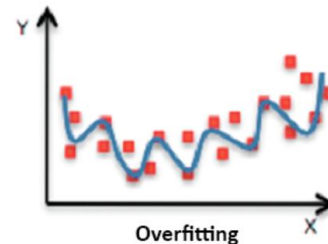
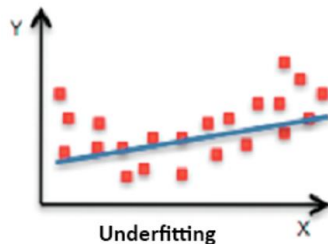
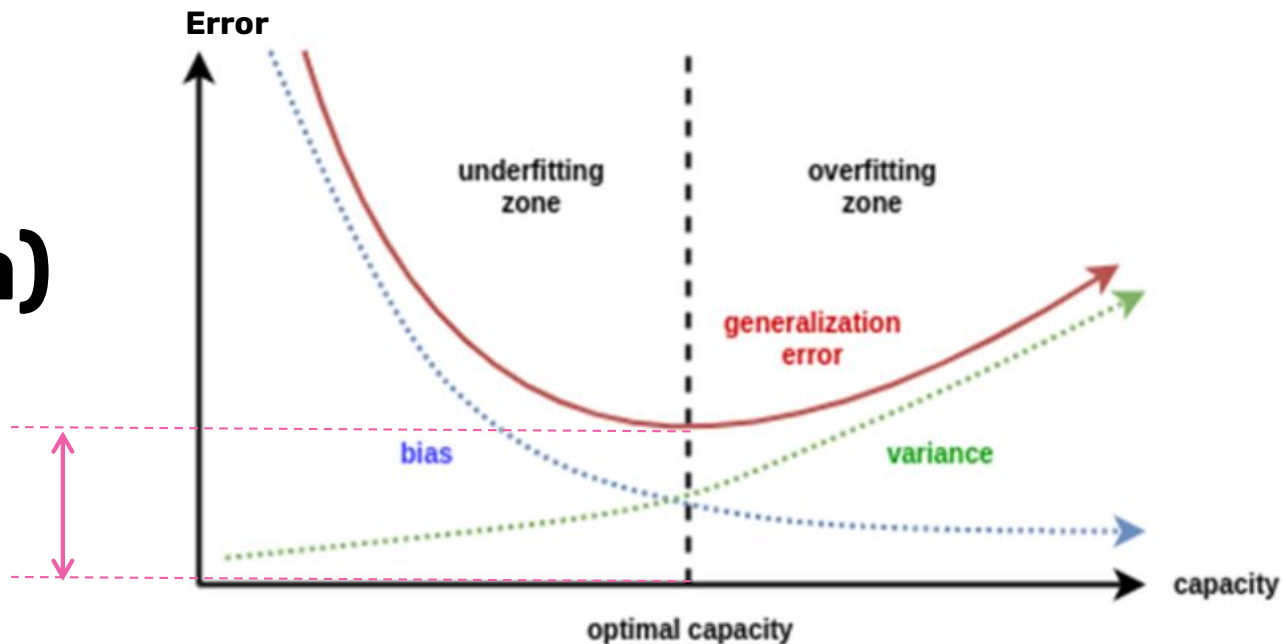
Total: 30.000 features



T

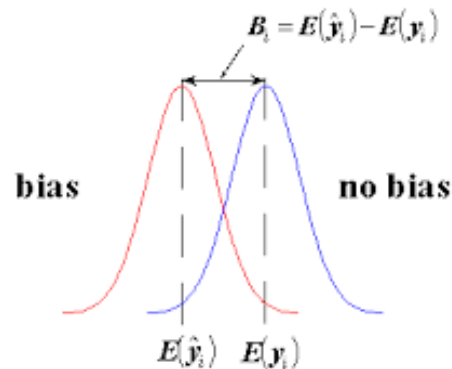
# Redução do Ruído (Viés Vs. Variância)

**Ruído = Erro Irreduzível**



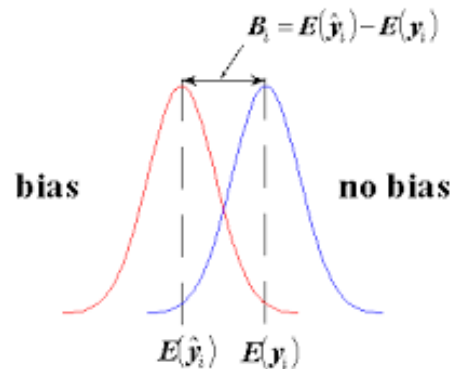
# Ruído (Viés Vs. Variância)

- **Viés** é a diferença entre a previsão média e o valor correto. Também é conhecido como erro de viés ou erro devido ao viés.
  - ❑ **Modelos de Low-Bias:** k-vizinhos mais próximos ( $k=1$ ), árvores de decisão e SVM
  - ❑ **Modelos de High-Bias:** Regressão Linear e Regressão Logística.

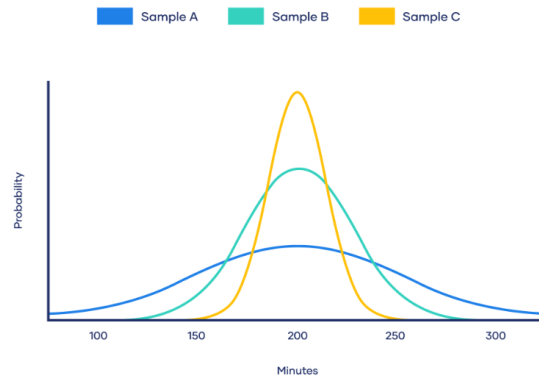


# Ruído (Viés Vs. Variância)

- **Viés** é a diferença entre a previsão média e o valor correto. Também é conhecido como erro de viés ou erro devido ao viés.
  - ❑ **Modelos de Low-Bias:** k-vizinhos mais próximos ( $k=1$ ), árvores de decisão e SVM
  - ❑ **Modelos de High-Bias:** Regressão Linear e Regressão Logística.
- **Variância** é a quantidade que a previsão mudará se diferentes conjuntos de dados de treinamento forem usados (dados dispersos ou inconsistentes). Também é conhecido como Erro de Variância ou Erro devido à Variância.
  - ❑ **Modelos de Baixa Variância:** Regressão Linear e Regressão Logística.
  - ❑ **Modelos de alta variância:** k-vizinhos mais próximos ( $k=1$ ), árvores de decisão e SVM

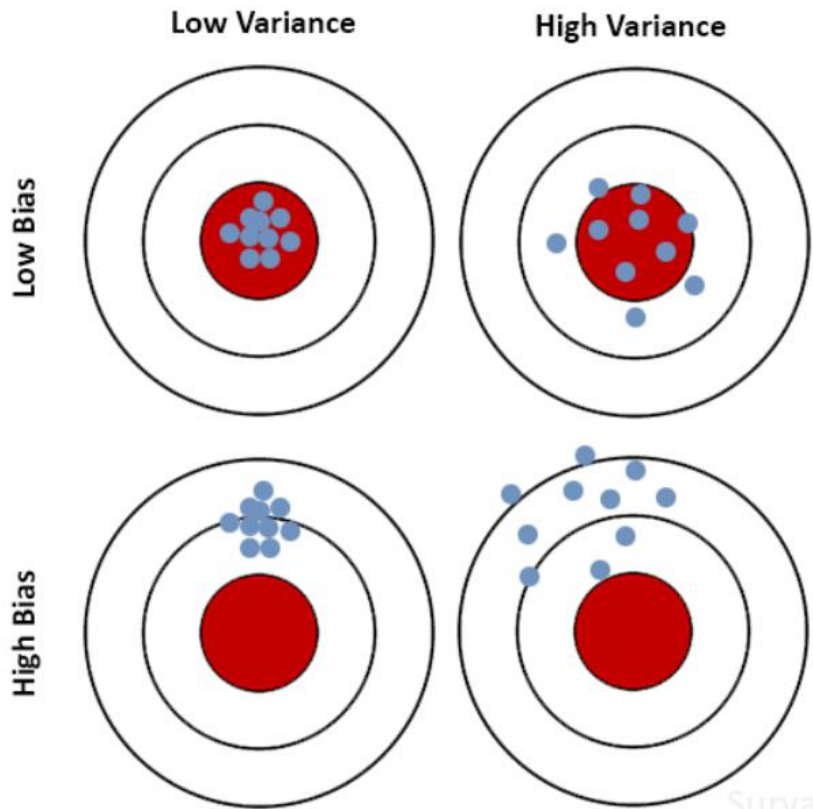


Average phone use per day in minutes



# Ruído (Viés Vs. Variância)

- **Low Bias - Low Variance:** É um modelo ideal. Mas, não podemos conseguir isso.
- **Low Bias - High Variance (Overfitting):** As previsões são inconsistentes e precisas em média. Isso pode acontecer quando o modelo usa um grande número de parâmetros.
- **High Bias - Low Variance (Underfitting):** As previsões são consistentes, mas imprecisas em média. Isso pode acontecer quando o modelo usa poucos parâmetros.
- **High Bias - High Variance:** As previsões são inconsistentes e imprecisas em média.



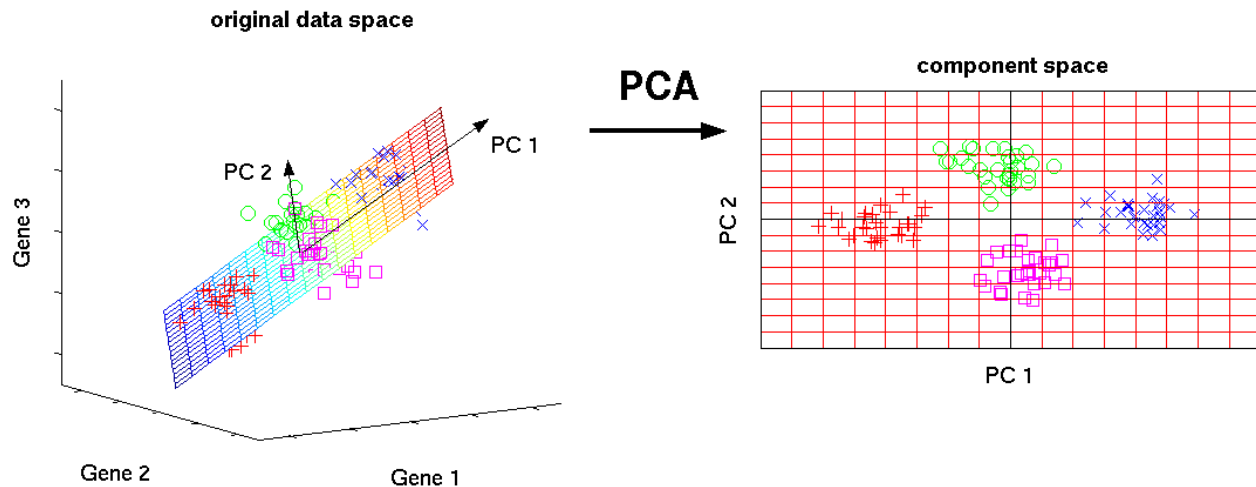
# Ruído (Viés Vs. Variância)

- **Erro irreduzível:** não pode ser reduzido independentemente dos modelos. É uma medida da quantidade de ruído em nossos dados devido a variáveis desconhecidas.
- **Como se relacionam viés, variância e erros irreduzíveis?**
- **Erro = erro redutível + erro irreduzível**
- **Erro redutível = viés<sup>2</sup> + variância**
- **Erro = Viés<sup>2</sup> + Variância + Erro Irreduzível**

$$\underbrace{(\hat{f}(x) - f(x))^2}_{\text{Erro}^2} = \underbrace{(E[\hat{f}(x)] - f(x))^2}_{\text{Viés}^2} + \underbrace{E[(\hat{f}(x) - E[\hat{f}(x)])^2]}_{\text{Variância}} + \underbrace{\sigma_e^2}_{\text{Erro Irreduzível}}$$

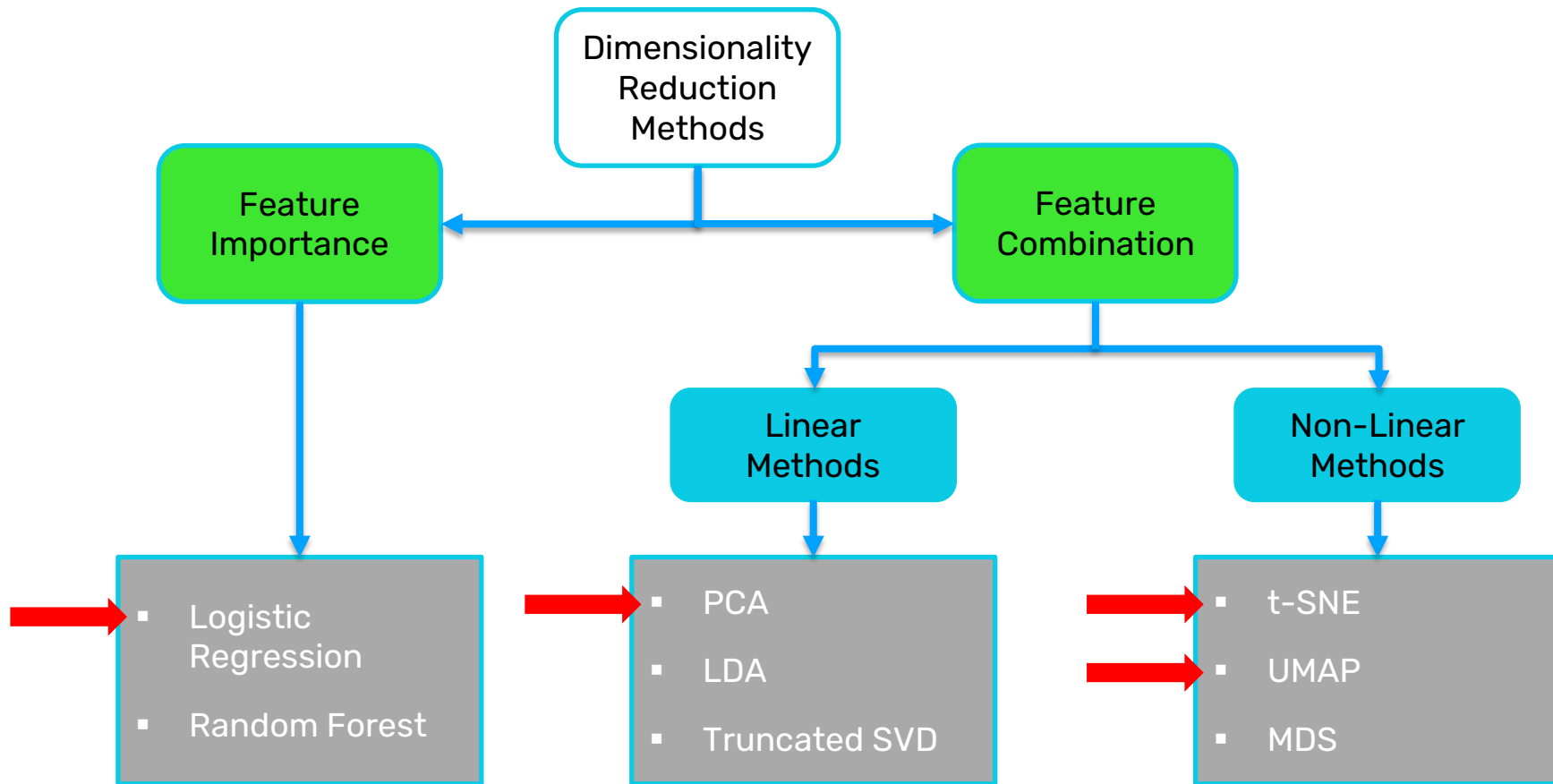
# Abordagens

- **Data Preparation:** Redução dos Dados de Entrada/redução de ruído
- **Data Visualization:** 2D e 3D
- **Cluster Analysis:** agrupamento



A vertical bar with a gradient from green at the top to blue at the bottom.

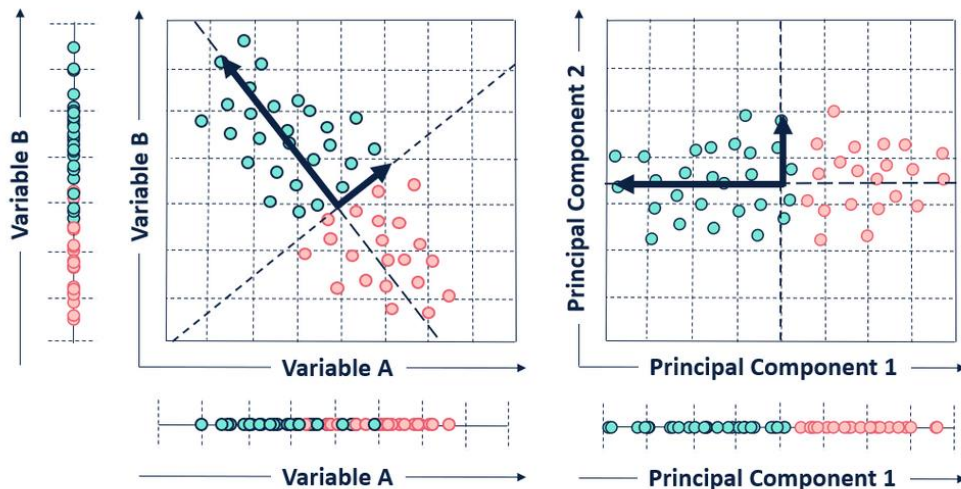
# Métodos de Redução de Dimensionalidade





# T Métodos Lineares: PCA

**Principal Component Analysis (PCA):** projeção linear de um conjunto de variáveis correlacionadas em um número menor de variáveis não correlacionadas chamadas componentes principais, mantendo o máximo possível da variância no conjunto de dados original

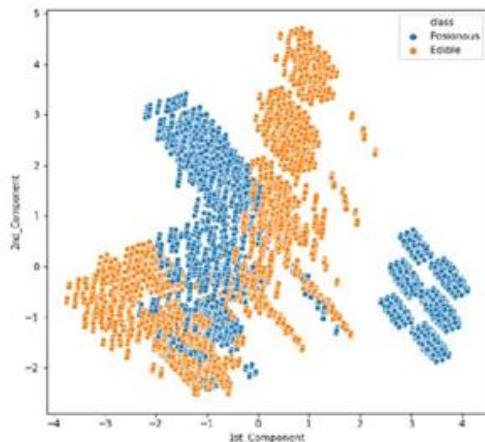


<https://youtu.be/FgakZw6K1QQ>

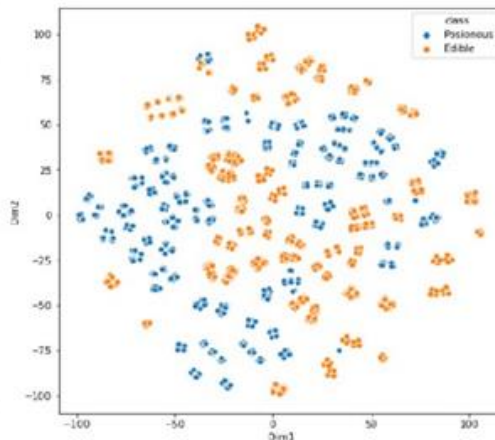
# T Métodos Não Lineares: t-SNE

**t-SNE:** *t-distribution Stochastic Neighborhood Embedding*

- a principal desvantagem do PCA é que ele não mantém as estruturas locais do conjunto de dados



PCA



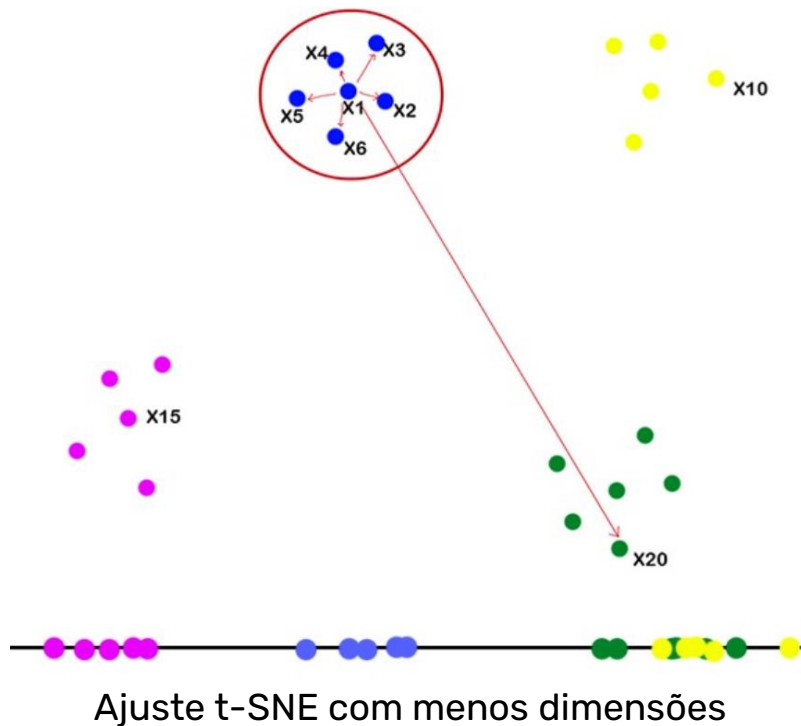
t-SNE



# Métodos Não Lineares: t-SNE

- O t-SNE cria uma distribuição de probabilidade escolhendo um ponto de dados aleatório e calculando a **distância euclidiana** com outros pontos de dados ( $|x_i - x_j|$ ). Os pontos de dados **próximos** do ponto de dados selecionado terão **mais valor de similaridade** e os pontos de dados que estão **longe** do ponto de dados selecionado **terão menos valor de similaridade**.
- Em seguida, **converte a distância de similaridade calculada em probabilidade** conjunta de acordo com a distribuição Normal.

Representação visual de um dataset multidimensional

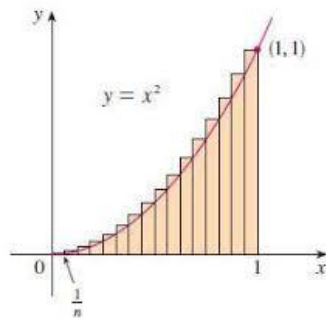




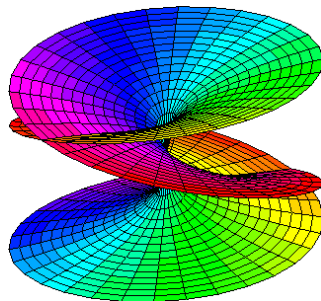
# Métodos Não Lineares: UMAP

**UMAP:** *Uniform Manifold Approximation and Projection*

- A UMAP é construída a partir de um arcabouço teórico baseado na **geometria de Riemann** (obtenção de grandezas por integração diferencial) e na topologia algébrica.
- O resultado é um **algoritmo escalável prático** que se aplica a dados do mundo real.
- O algoritmo UMAP é competitivo com o t-SNE para qualidade de visualização e, sem dúvida, **preserva mais da estrutura global com desempenho de tempo de execução superior**.
- Além disso, UMAP **não possui restrições computacionais** na dimensão de incorporação, tornando-o viável como uma técnica de redução de dimensão de propósito geral para aprendizado de máquina.



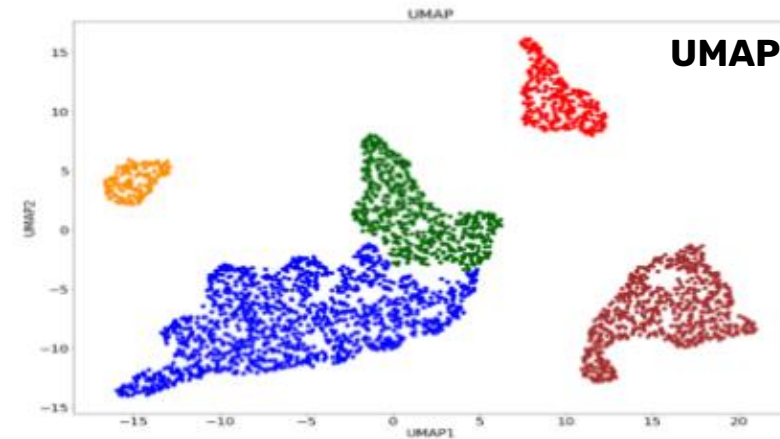
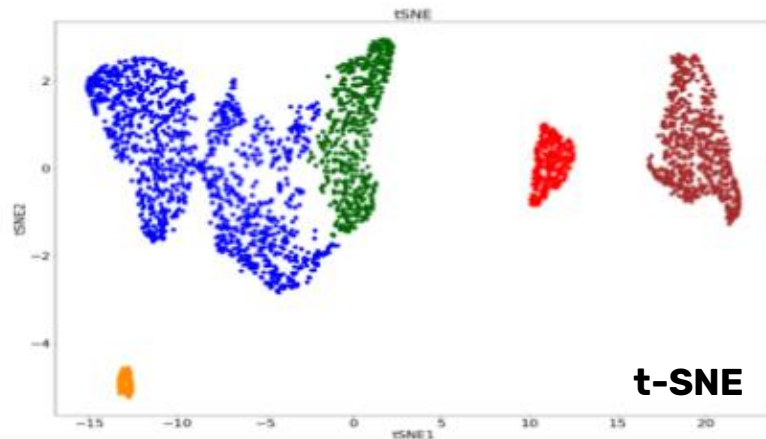
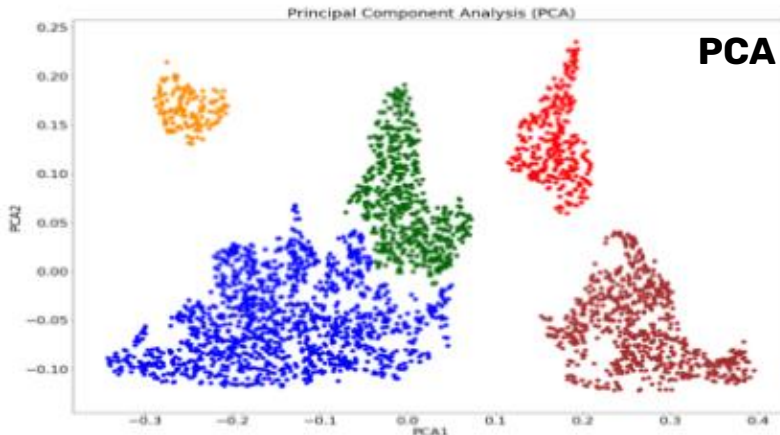
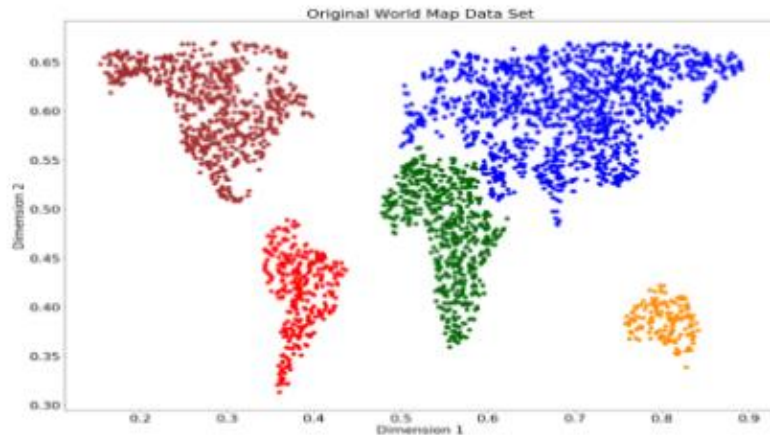
Somas de Riemann e Integração



Superfícies de Riemann

**T**

# Comparação entre Métodos



# t-SNE Vs. UMAP

## t-SNE

- sempre começa com uma **inicialização randômica**, o que causa resultados diferentes para um mesmo DataSet
- **move cada ponto individualmente** a cada interação de maneira reduzida

## UMAP

- inicializa com Spectral Embedding, que **garante sempre a mesma inicialização** (índices de similaridade relativizando a posição dos pontos)
- **move um ou um subconjunto de pontos por vez**, que permite uma melhor escala para DataSets maiores

T

A vertical bar with a gradient from light green at the top to light blue at the bottom.

**ALGUMA DÚVIDA ATÉ  
AQUI?**



# Aplicações Práticas: Jupyter



# INTERVALO 10 MIN



## **APROVEITE PARA:**


- Fazer anotações do que viu até agora (aprendizados, insights, dúvidas)
- Levantar-se, esticar os braços e as pernas, relaxar por mais tempo
- Comer algo para voltar com energia renovada
- Ir ao toalete

T



# DÚVIDAS FINAIS

T

A vertical bar with a gradient from green at the top to blue at the bottom.

# COMO FOI?