ICC204 - Aprendizagem de Máquina e Mineração de Dados

Redução de Dimesionalidade

Prof. Rafael Giusti
rgiusti@icomp.ufam.edu.br

Maldição da dimensionalidade

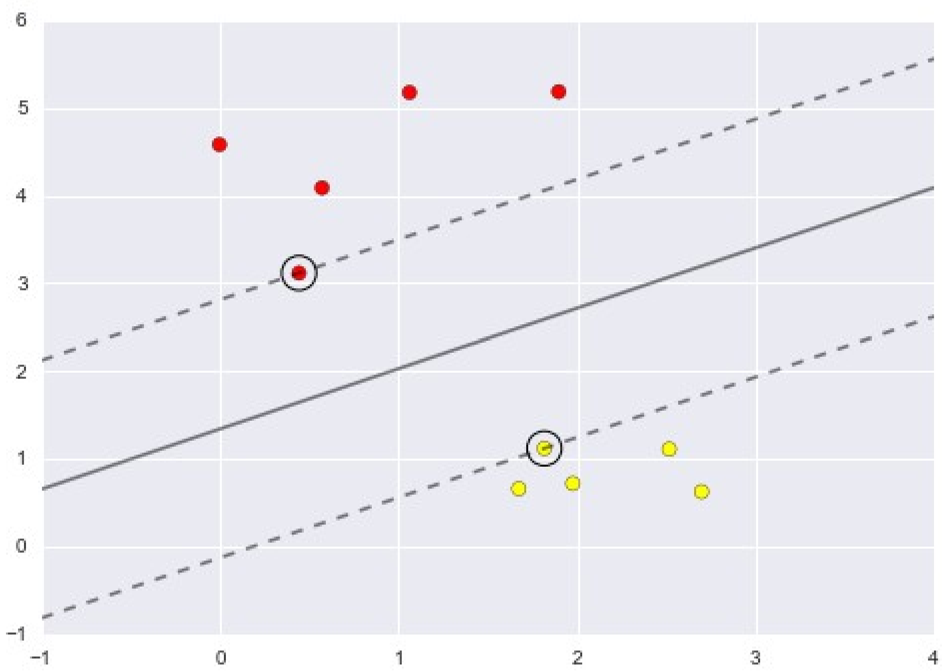
- A **dimensionalidade** de um conjunto de dados é o número de dimensões necessário para representar os exemplos
  - Número de atributos independentes
- A **maldição da dimensionalidade** compreende uma série de dificuldades associadas à alta dimensionalidade dos dados

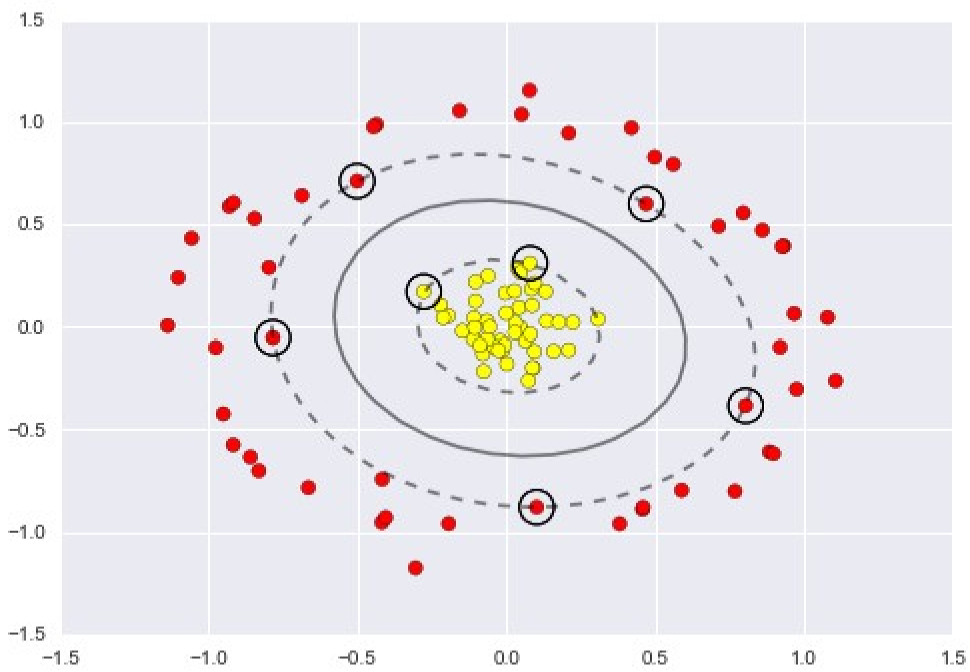
Maldição da dimensionalidade

- **Complexidade temporal**
  - A complexidade temporal dos algoritmos aumenta com o número de atributos e o tamanho dos conjuntos de treino e teste – ambos $\mathcal{O}(n)$
    - k-NN "ingênuo" tem complexidade $\mathcal{O}(k{\cdot}n^2{\cdot}m)$
    - Naive Bayes pode ser treinado em $\mathcal{O}(m{\cdot}n{\cdot}\log n)$
    - SVC é treinado em $\mathcal{O}(n^3{\cdot}m)$

Maldição da dimensionalidade

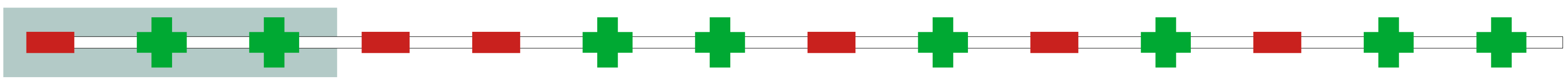
- **Perda da representatividade**
  - O número de exemplos necessário para representar um conceito aumenta exponencialmente com a dimensionalidade





Maldição da dimensionalidade

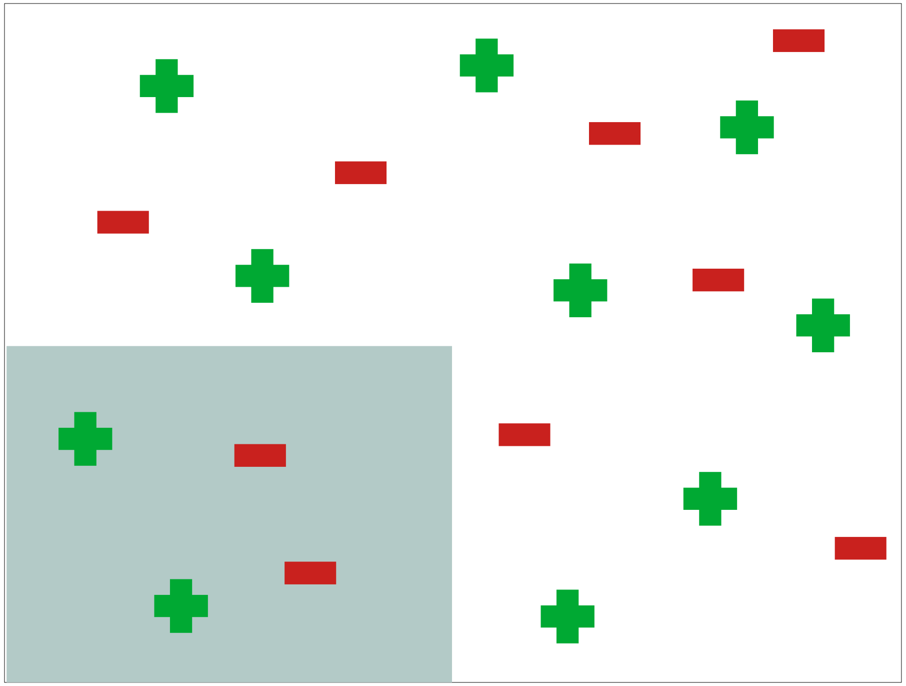
- **Perda da representatividade**
  - O número de exemplos necessário para representar um conceito aumenta exponencialmente com a dimensionalidade



1 dimensão: 20% da população cobre 20% do espaço de atributos

Maldição da dimensionalidade

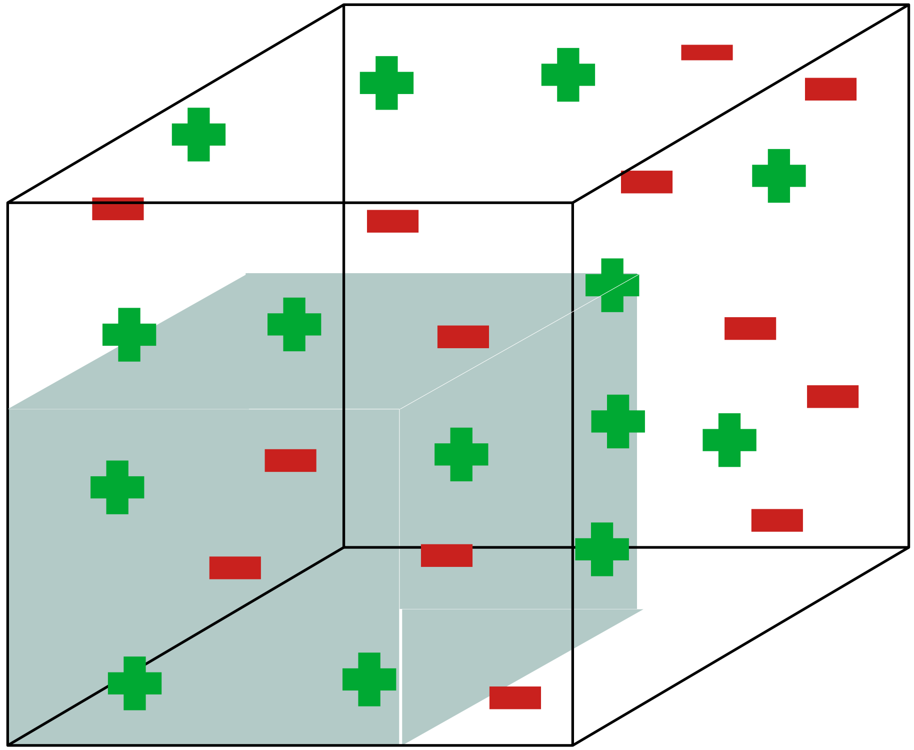
- **Perda da representatividade**
  - O número de exemplos necessário para representar um conceito aumenta exponencialmente com a dimensionalidade



2 dimensões: para cobrir 20% do espaço, precisamos obter 45% da população em cada dimensão

Maldição da dimensionalidade

- **Perda da representatividade**
  - O número de exemplos necessário para representar um conceito aumenta exponencialmente com a dimensionalidade



3 dimensões: para cobrir 20% do espaço, precisamos obter 58% da população em cada dimensão

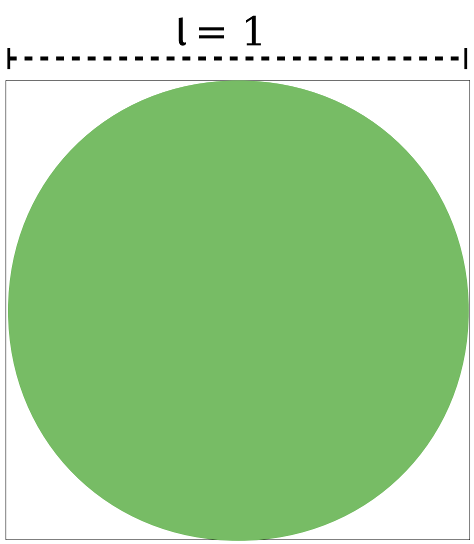
Maldição da dimensionalidade

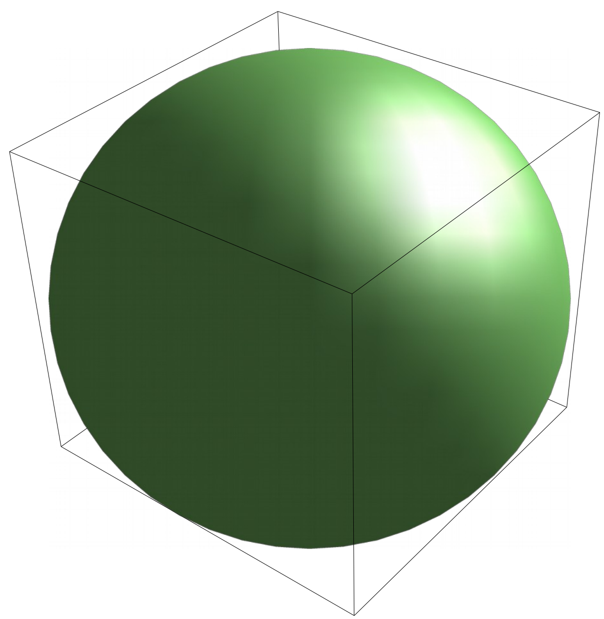
- **Insignificância do conceito de vizinhança**
  - A separação entre o vizinho mais longe e o mais próximo tende a zero

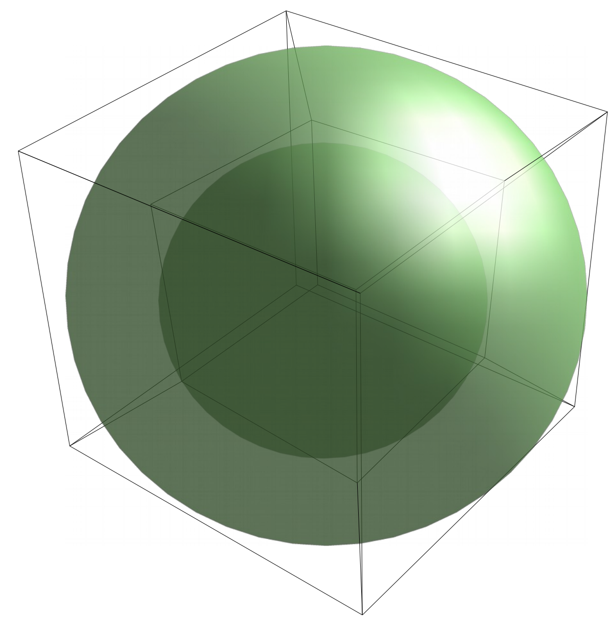
$$\lim_{n \to \infty} \frac{d_{\max} - d_{\min}}{d_{\min}} = 0$$

Maldição da dimensionalidade

- **Insignificância do conceito de vizinhança**
  - A separação entre o vizinho mais longe e o mais próximo tende a zero







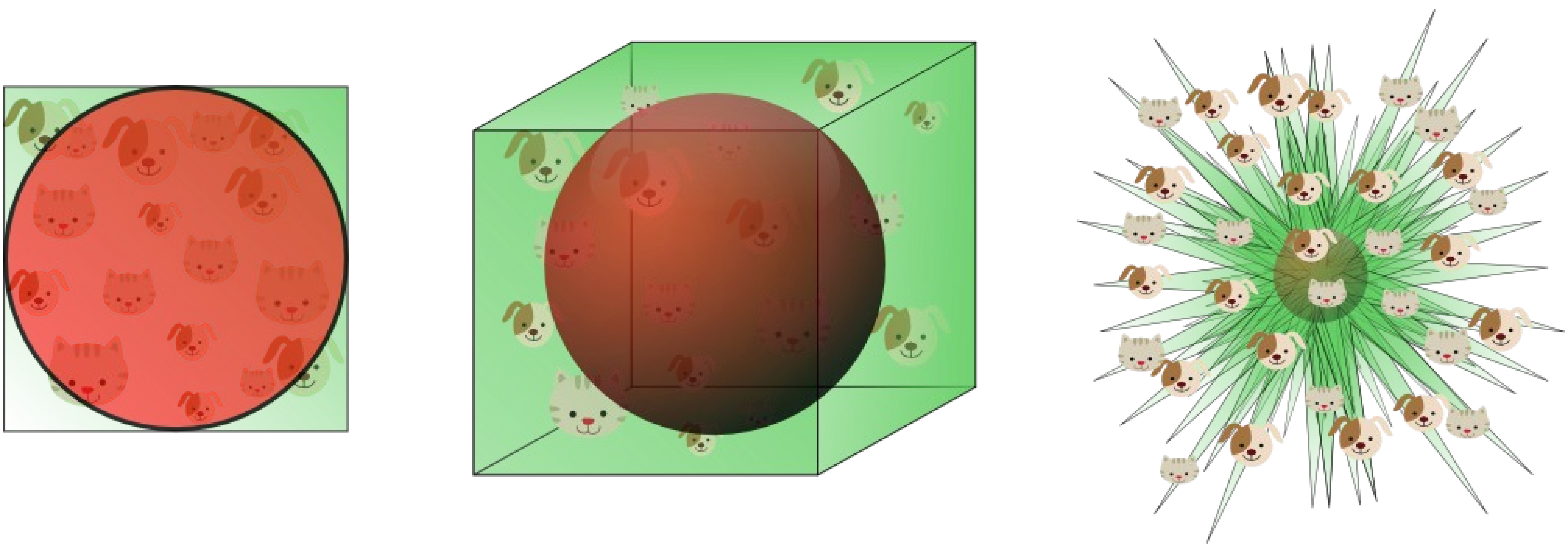
$A_s = \pi r^2 = {}^1/_4\pi$

$A_s = {}^4/_3\pi r^3 = {}^1/_6\pi$

$A_s = {}^1/_2\pi^2 r^4 \approx {}^1/_{10}\pi$

Maldição da dimensionalidade

- **Insignificância do conceito de vizinhança**
  - A separação entre o vizinho mais longe e o mais próximo tende a zero

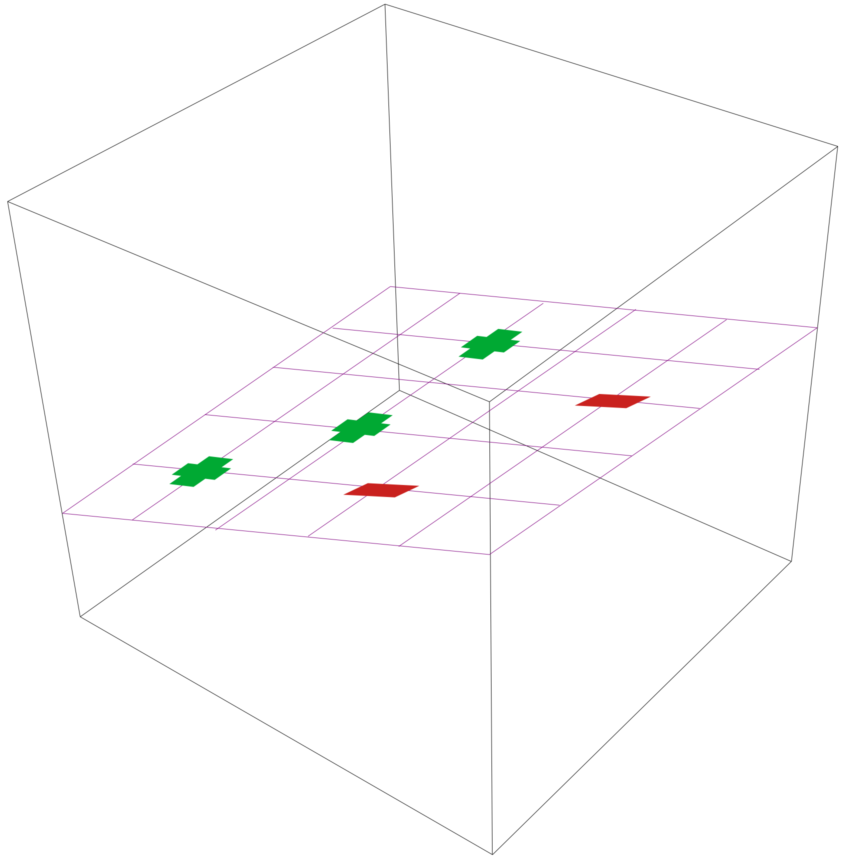


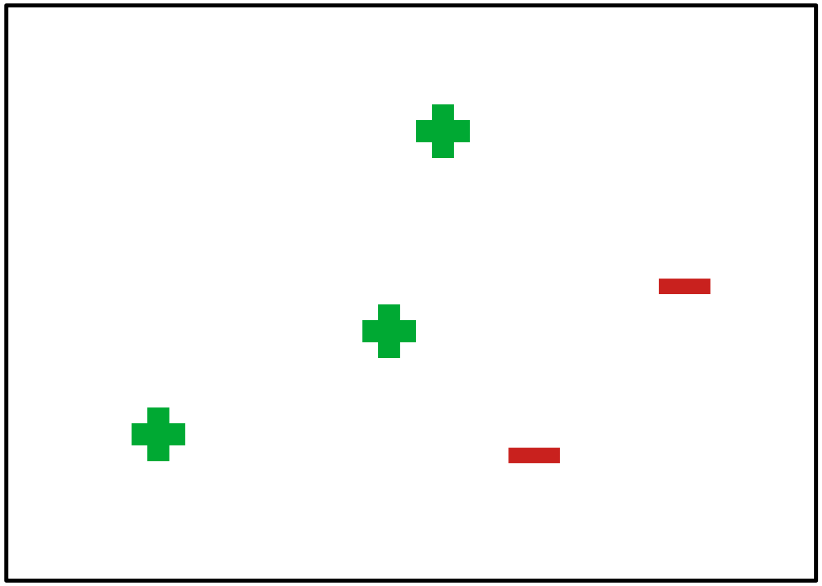
https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Redução de dimensionalidade

- Transportar exemplos de um espaço $\mathcal{R}^N$ para um espaço de menor dimensionalidade $\mathcal{R}^M$, tal que $M \ll N$
- De modo geral, a redução de dimensionalidade incorre um **erro de reconstrução**
  - Mas o erro pode ser zero se a dimesionalidade intrínseca aos dados for M
  - Os exemplos estavam "encapsulados" (*embedded*) em um espaço de maior dimensionalidade

Dimesionalidade intrínseca





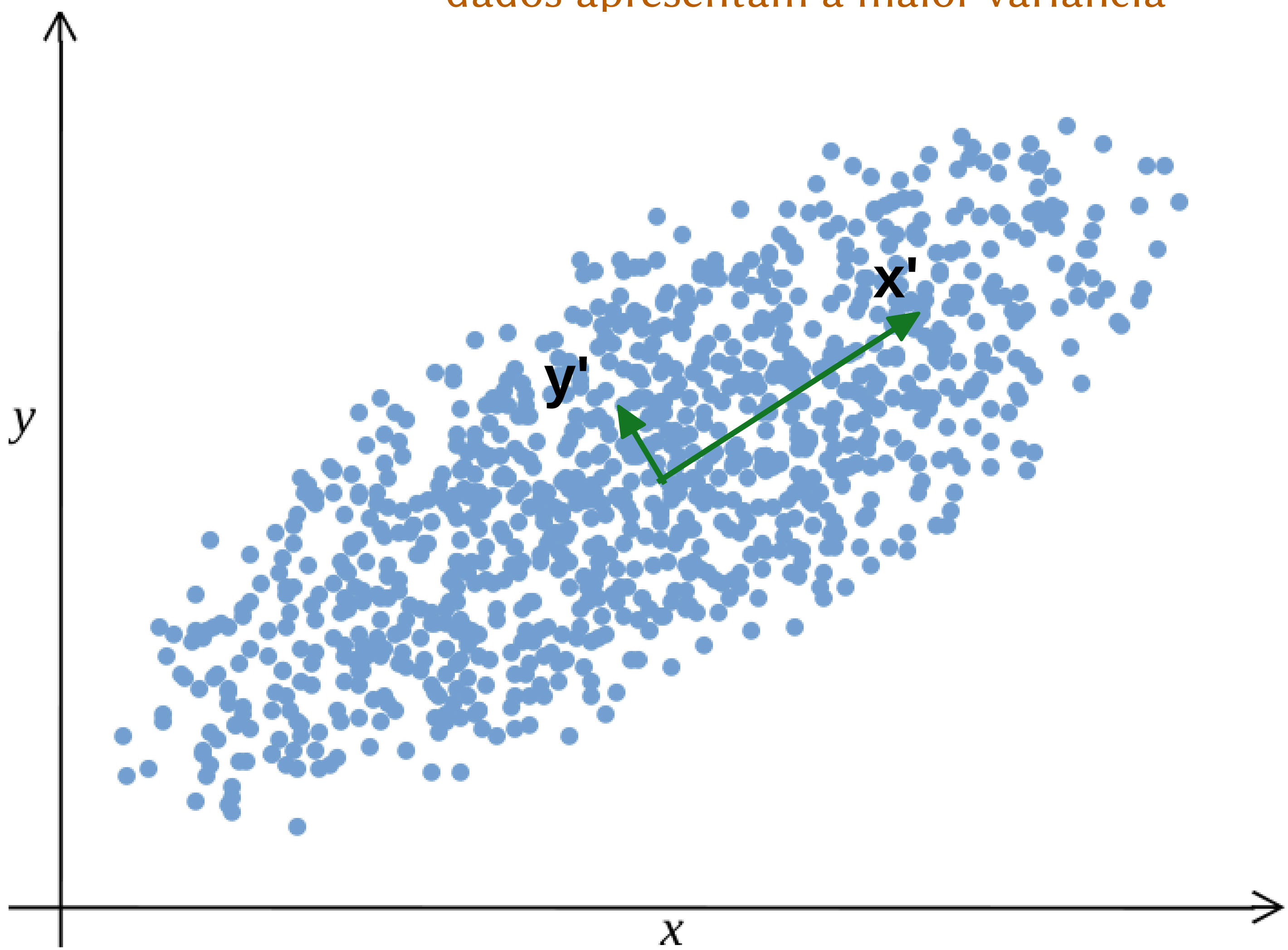
Os exemplos à esquerda estão contidos em um espaço de três dimensões, mas na verdade ocupam apenas a região de um plano (dimensão inerente d=2)

PCA

- O **PCA** é um método de análise de dados não supervisionado que pode ser utilizado para redução de dimensionalidade
  - *Principal Component Analysis*
  - Análise de Componentes Principais
- O PCA procura as **componentes principais** dos dados, que são os eixos de **maior variância** e faz uma **rotação** para gerar um novo espaço

PCA

Os vetores x' e y' indicam as direções na qual os dados apresentam a maior variância



PCA

Os dados são rotacionados de modo que os vetores x' e y' formam a nova base do espaço no qual os exemplos são representados

Componentes principais

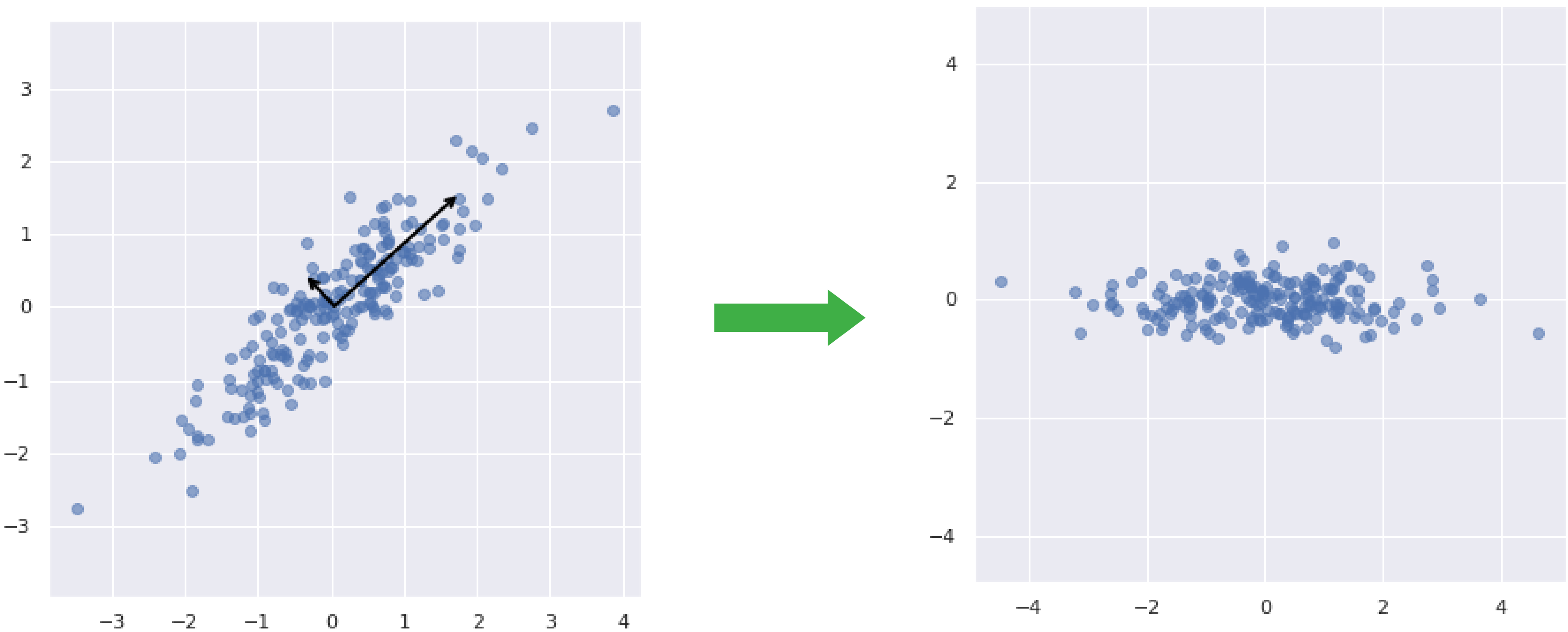
- Os vetores obtidos pelo método PCA são denominados **componentes principais** (ou autovetores)
- Eles são dados em ordem de significância
  - A primeira componente principal corresponde ao eixo de maior variância dos dados
  - Cada componente subsequente corresponde a eixos variância menor que os anteriores
- A redução de dimensionalidade pode ser obtida descartando as componentes de baixa variância

PCA passo a passo

- Se os dados são os vetores coluna
  - $X = (X_1, X_2, X_3)^T$
- Então primeiro centraliza-se os dados
  - $X_{zero} = X - (\overline{X}_1, \overline{X}_2, \overline{X}_3)^T$
- Em seguida obtém-se a matriz de auto-covariância
  - $X_{cov} = Cov(X_{zero}, X_{zero})$
- As componentes principais serão os autovetores de X_{cov}
  - $(\lambda, U) = \mathit{Eigendecomposition}(X_{cov})$

PCA passo a passo

- Para obter a representação dos dados no espaço das componentes principais, basta empregar a matriz U como uma matriz de mudança de base
  - $X_{rot} = X \cdot U$



Redução de dimensionalidade

- Para reduzir a dimensionalidade de um conjunto com dimensão N para dimensão $p < N$
- Obtenha os *loadings* e as componentes principais
  - $(\lambda, U) = \mathit{Eigendecomposition}(X_{cov})$
- Ordene as colunas de U por *loadings* e selecione as p colunas de maior variância
- Faça a transformação com U_p
  - $X_{proj} = X \cdot U_p$

Redução de dimensionalidade

- Aproximação do dígito zero e componentes principais do conjunto DIGITS