Por que o Cientista de Dados precisa estudar Matemática/Estatística?

Cibele Russo

ICMC USP

22/3/2022

Sobre mim

- Bacharelado em Matemática Aplicada e Computação Científica (ICMC USP, 2005)
- Mestrado em Ciências de Computação e Matemática Computacional (ICMC USP, 2006)
- Doutorado em Estatística (IME USP, 2010)
- Professora de Estatística, Ex-Coordenadora do Bacharelado em Estatística e Ciência de Dados e Professora do MBA em Ciências de Dados (ICMC USP)

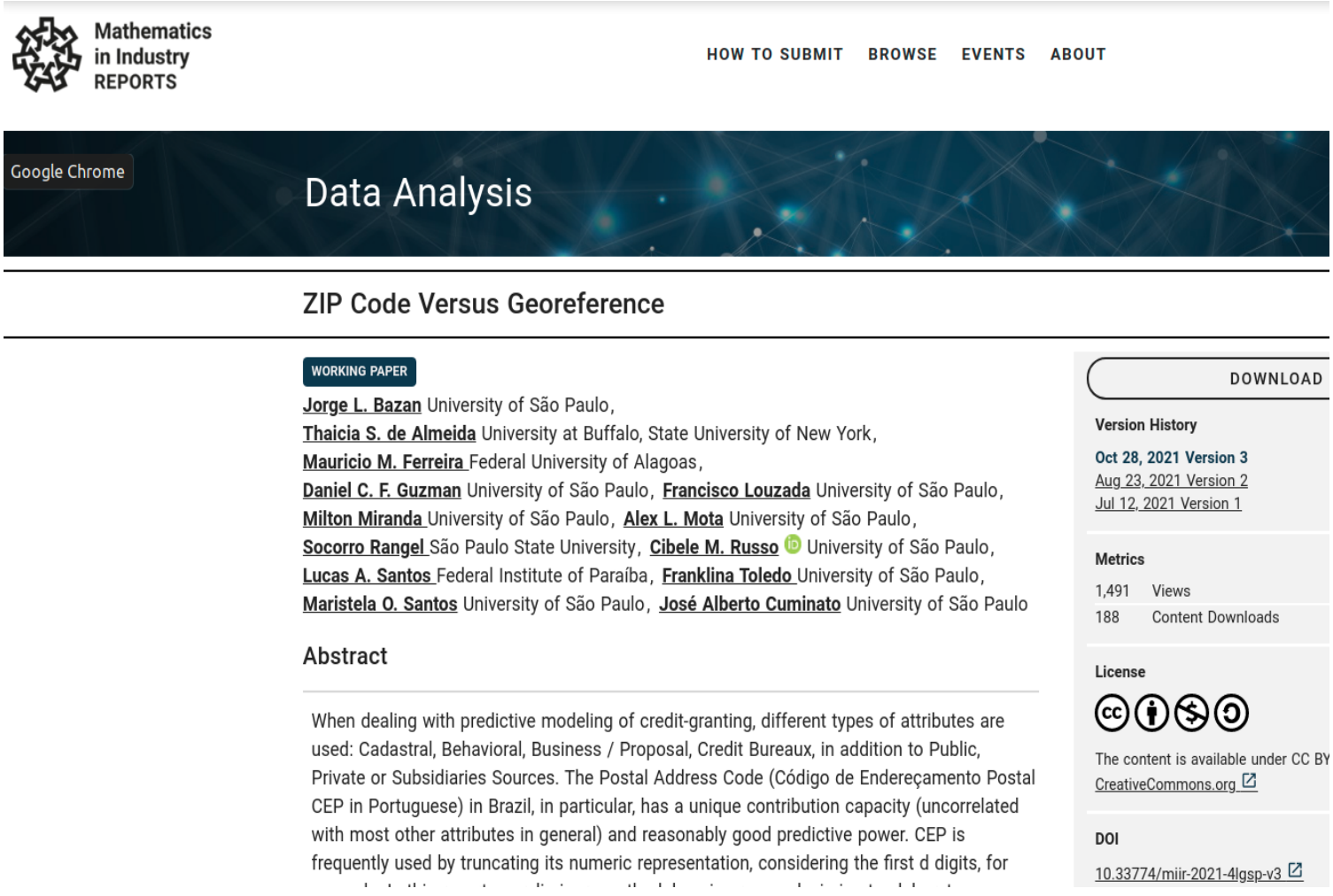
Algumas experiências

- 2004: Estatcamp - Consultoria Estatística em Qualidade, São Carlos SP
- 2006: Itaú - Itaú/Unibanco, São Paulo SP
- 2008: IPq - Hospital das Clínicas, São Paulo SP
- 2013: Biostatistics Dept, Erasmus Medical Center - Rotterdam, Netherlands

Erasmus Medical Center. Fonte: https://www.wildeboer.de/en/references/healthcare/erasmus-medical-center-rotterdam-nl/

Alguns projetos recentes: Estoque Seguro

Alguns projetos recentes: ZIP code versus georeference



Mathematics in Industry REPORTS

HOW TO SUBMIT BROWSE EVENTS ABOUT

Google Chrome

Data Analysis

ZIP Code Versus Georeference

WORKING PAPER

Jorge L. Bazan University of São Paulo,
Thaicia S. de Almeida University at Buffalo, State University of New York,
Mauricio M. Ferreira Federal University of Alagoas,
Daniel C. F. Guzman University of São Paulo, Francisco Louzada University of São Paulo,
Milton Miranda University of São Paulo, Alex L. Mota University of São Paulo,
Socorro Rangel São Paulo State University, Cibele M. Russo University of São Paulo,
Lucas A. Santos Federal Institute of Paraíba, Franklina Toledo University of São Paulo,
Maristela O. Santos University of São Paulo, José Alberto Cuminato University of São Paulo

Abstract

When dealing with predictive modeling of credit-granting, different types of attributes are used: Cadastral, Behavioral, Business / Proposal, Credit Bureaux, in addition to Public, Private or Subsidiaries Sources. The Postal Address Code (Código de Endereçamento Postal CEP in Portuguese) in Brazil, in particular, has a unique contribution capacity (uncorrelated with most other attributes in general) and reasonably good predictive power. CEP is frequently used by truncating its numeric representation, considering the first d digits, for

DOWNLOAD

Version History

Oct 28, 2021 Version 3
Aug 23, 2021 Version 2
Jul 12, 2021 Version 1

Metrics

1,491 Views
188 Content Downloads

License

DOI

10.33774/miir-2021-4lgsp-v3

Alguns projetos recentes: ZIP code versus georeference

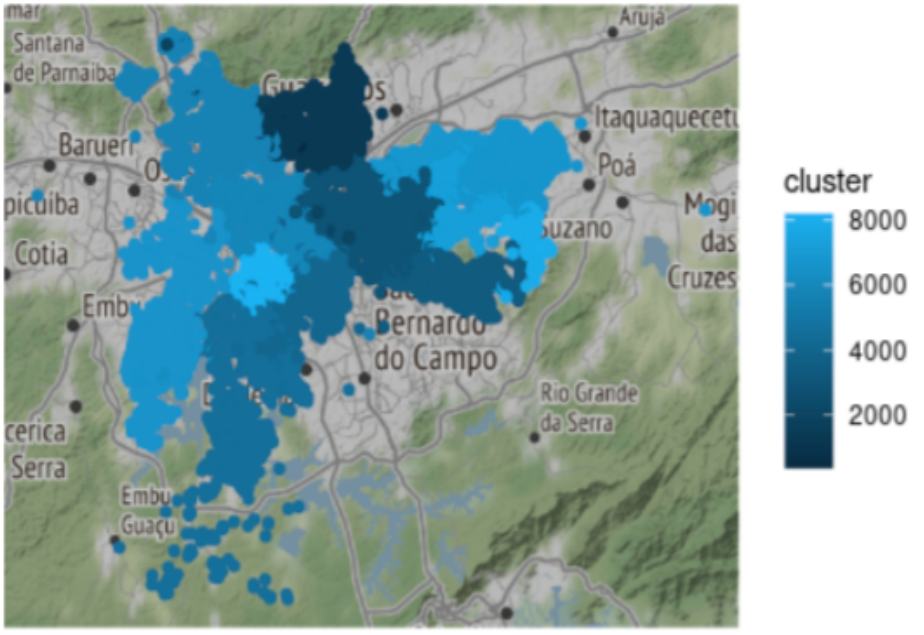


Figure 3. Nearest neighbors by default, with 33 clusters obtained

Alguns projetos recentes: ZIP code versus georeference

E a Matemática? E a Estatística?

- Ferramentas para a **formação sólida** do cientista de dados;
- Ajudam na **curiosidade** pela teoria dos modelos que são populares hoje;
- Motivam o interesse pela **criação de novos modelos**, de melhor performance que os existentes;
- Contribuem para tornar a/o cientista de dados brasileira/o **competitiva/o** onde ela/e quiser estar.

Matemática, Estatística, Ciência de Dados

O que é Ciência de dados?

O que é Ciência? Ciência é Conhecimento. Ciência de dados seria conhecimento profundo sobre dados?

No que a Ciência de Dados se difere da Estatística?

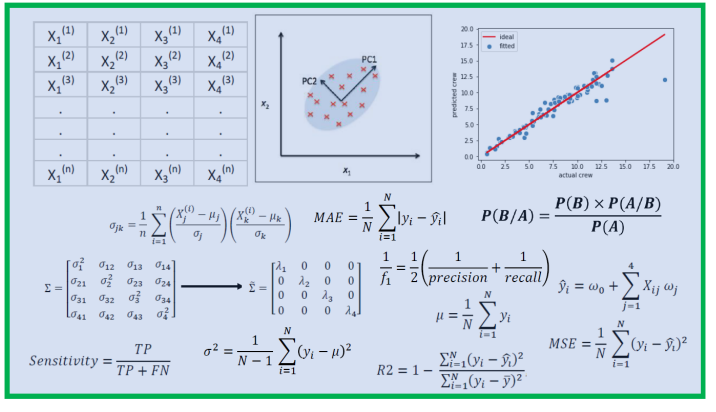
E onde entra a Matemática nisso tudo?

Matemática, Estatística, Ciência de Dados

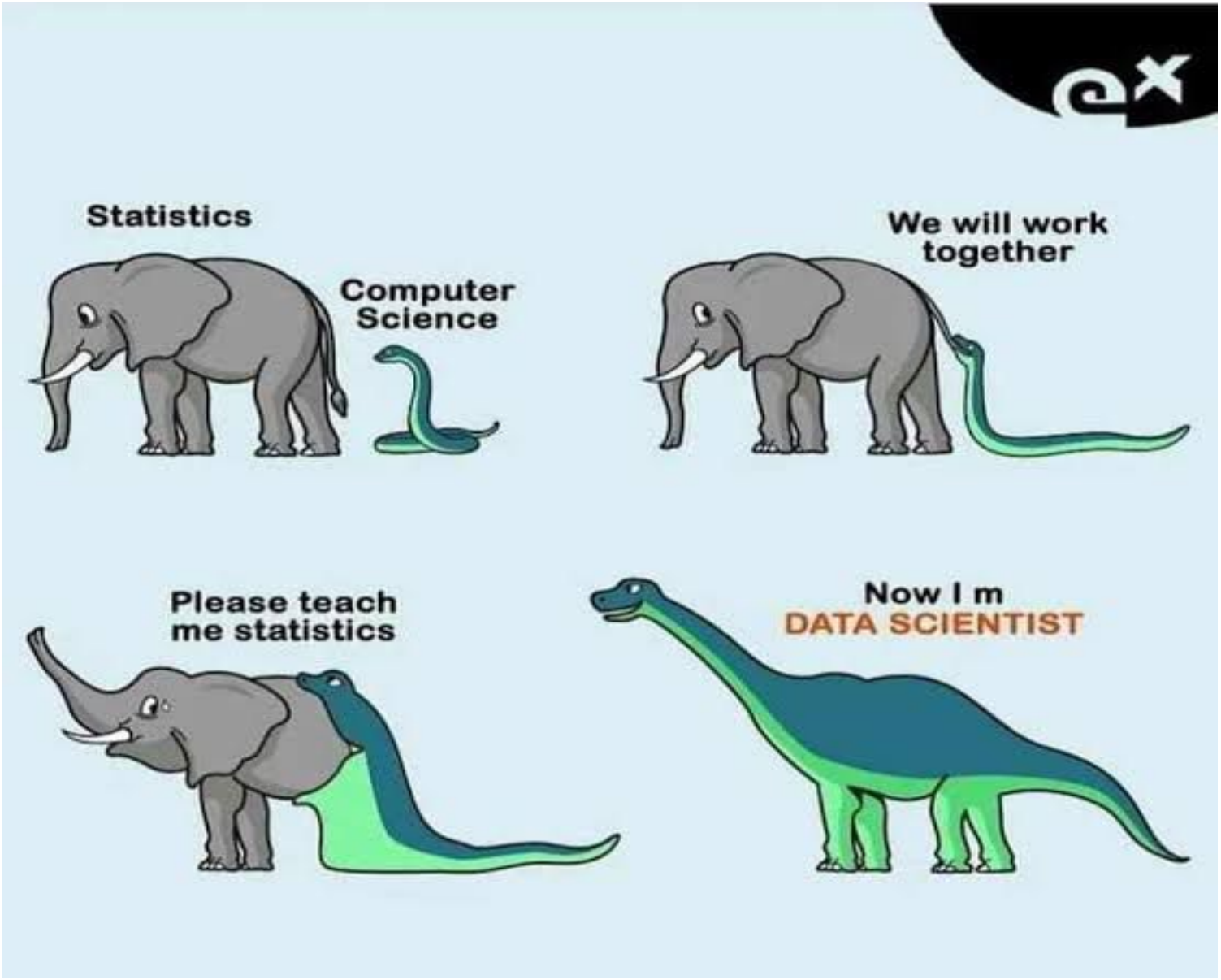
Que Matemática o Cientista de dados precisa aprender?

Principalmente:

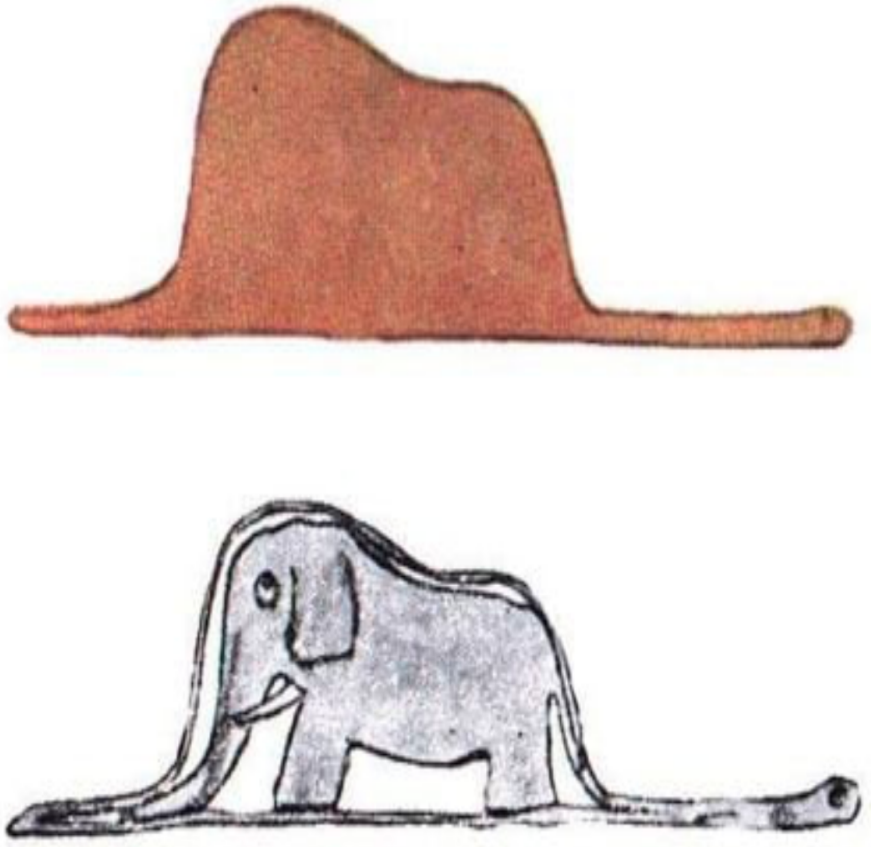
1. Cálculo
2. Álgebra Linear
3. Probabilidade e Estatística
4. Ciência da Computação
5. Otimização
6. O que mais é importante?



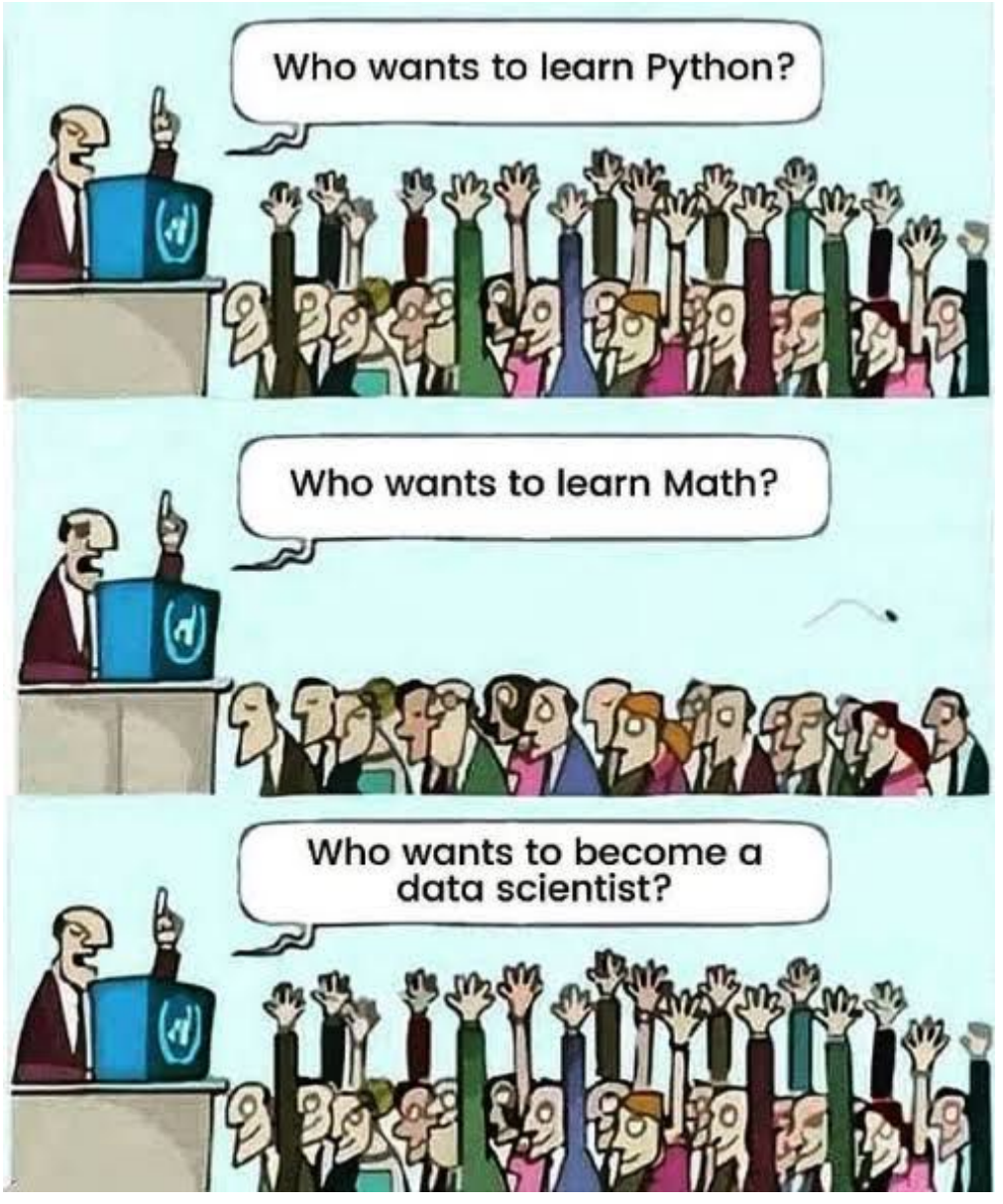
Fonte: https://taylor-mark110.medium.com/mathematics-an-essential-skill-for-aspiring-data-science-professional-c772b484573



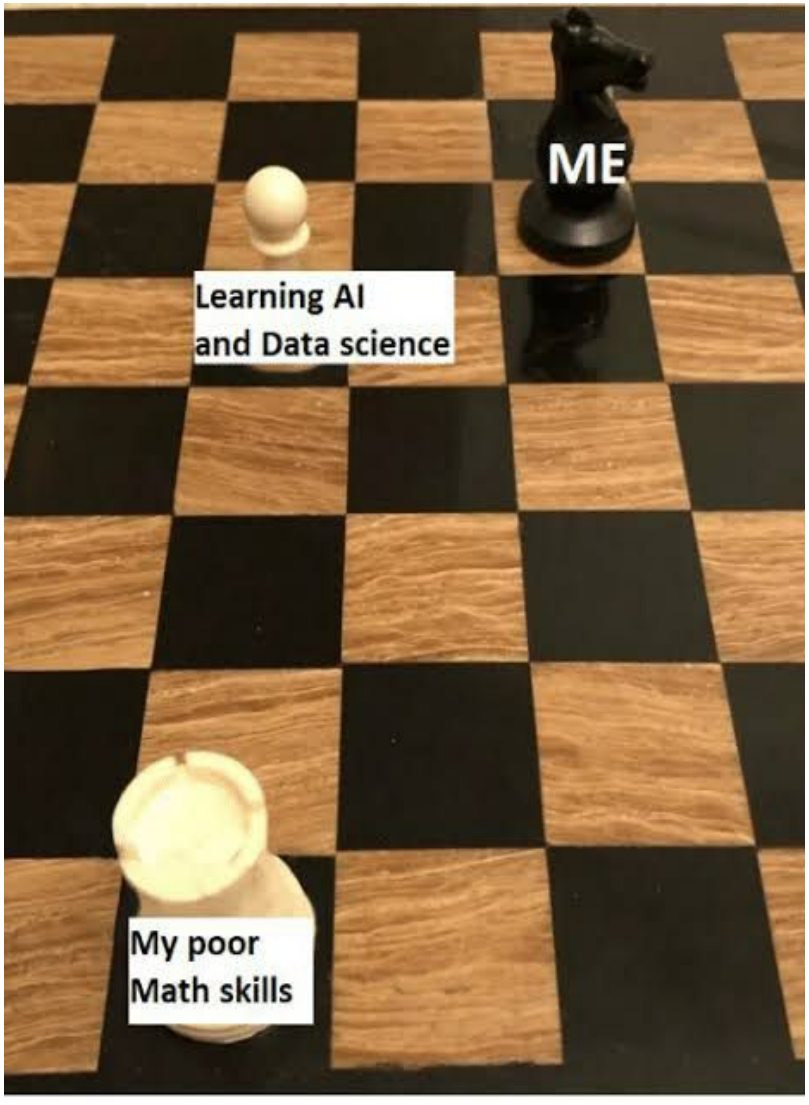
Fonte: incerta



Fonte: DE SAINT-EXUPÉRY, Antoine. **Le petit prince: avec des aquarelles de l'auteur**. Ernst Klett Sprachen GmbH, 2015.



Fonte: incerta



Fonte: incerta

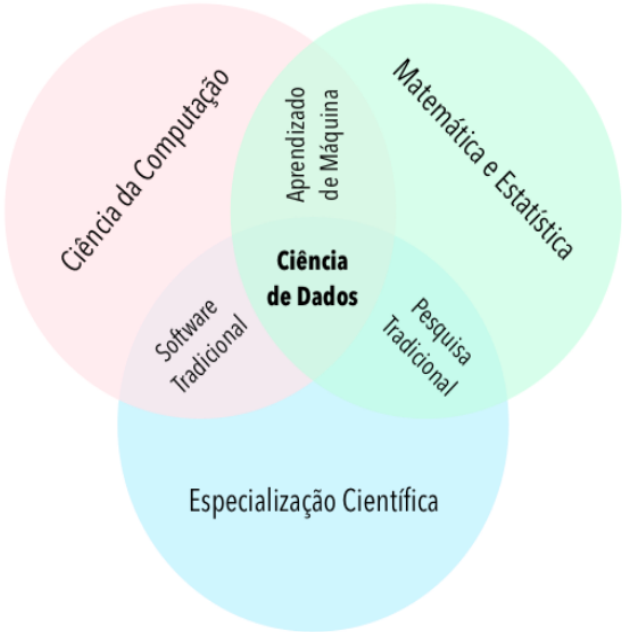


Diagrama da Ciência de Dados (Data Science Venn Diagram por Drew Conway)

Uma área completamente nova que combina:

1. **Ciência da Computação:** armazenar, obter e tratar dados
2. **Matemática e estatística:** filtrar e minerar
3. **Design Gráfico:** visualizar e refinar
4. **Especialização Científica:** perguntar

Fonte: https://alfredbaudisch.medium.com/o-que-%C3%A9-ci%C3%AAncia-de-dados-data-science-7af5bdac101a

Data Science

"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data."
(https://en.wikipedia.org/wiki/Data_science)

Fonte: http://getc.com.tn/formation/big-data-data-science/

Data Science

(https://en.wikipedia.org/wiki/Data_science)

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyze actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

Data Scientist

(https://en.wikipedia.org/wiki/Data_science)

A data scientist is someone who creates programming code, and combines it with statistical knowledge to create insights from data.

Machine learning

1. Aprendizado supervisionado (Modelos de regressão)

    - regressão: target ou valor alvo (variável resposta);
    - classificação: modelos em que a variável resposta é categórica;

2. Aprendizado não-supervisionado (Modelos de análise multivariada)

    - clustering (análise de agrupamentos);
    - análise de componentes principais;
    - análise de correspondência

3. Aprendizado dinâmico (Modelos para séries de tempo)

4. Aprendizado profundo (pode ser visto com interpretação probabilística)

... Outros

Modelo de regressão linear

Objetivos

Predizer Y a partir do conhecimento de variáveis em $X = x$.

Em notação matricial, um modelo linear geral é dado por

$$Y = X\beta + \epsilon,$$

em que

- Y é a **variável resposta** (vetor de variáveis aleatórias observáveis),
- X contém **variáveis preditoras** (matriz conhecida, ou seja, não-aleatória),
- β é um **vetor de parâmetros de interesse**, que queremos estimar,
- ϵ é o **erro aleatório** (vetor de erros aleatórios não observáveis).

Modelo de regressão linear

$$Y = X\beta + \epsilon,$$

em que

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

ou seja,

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}.$$

No modelo

$$Y = X\beta + \epsilon.$$

com as suposições

- $E(\epsilon) = 0$,
- $Var(\epsilon) = \sigma^2 I$,
- $(\epsilon \sim N(0, \sigma^2 I))$

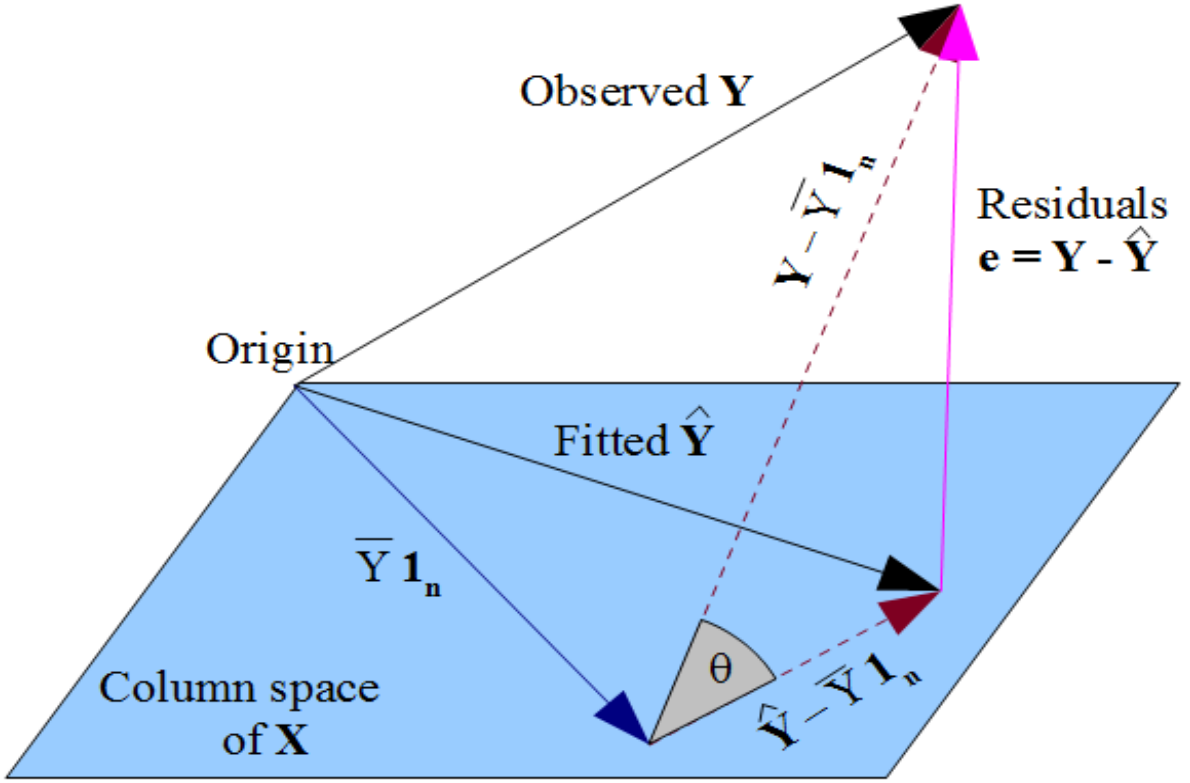
o estimador de mínimos quadrados (máxima verossimilhança) é dado por

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y.$$

O valor ajustado é

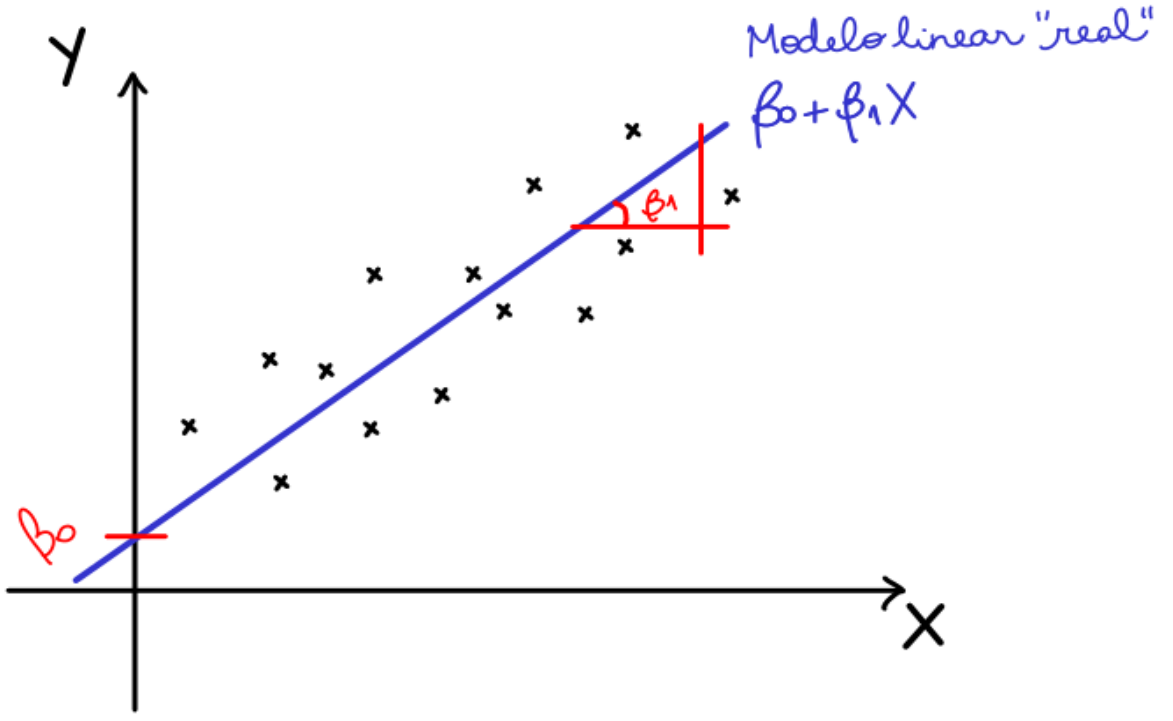
$$\widehat{Y} = X\widehat{\beta} = X(X^\top X)^{-1} X^\top Y = HY$$

Mas o que significa isso?

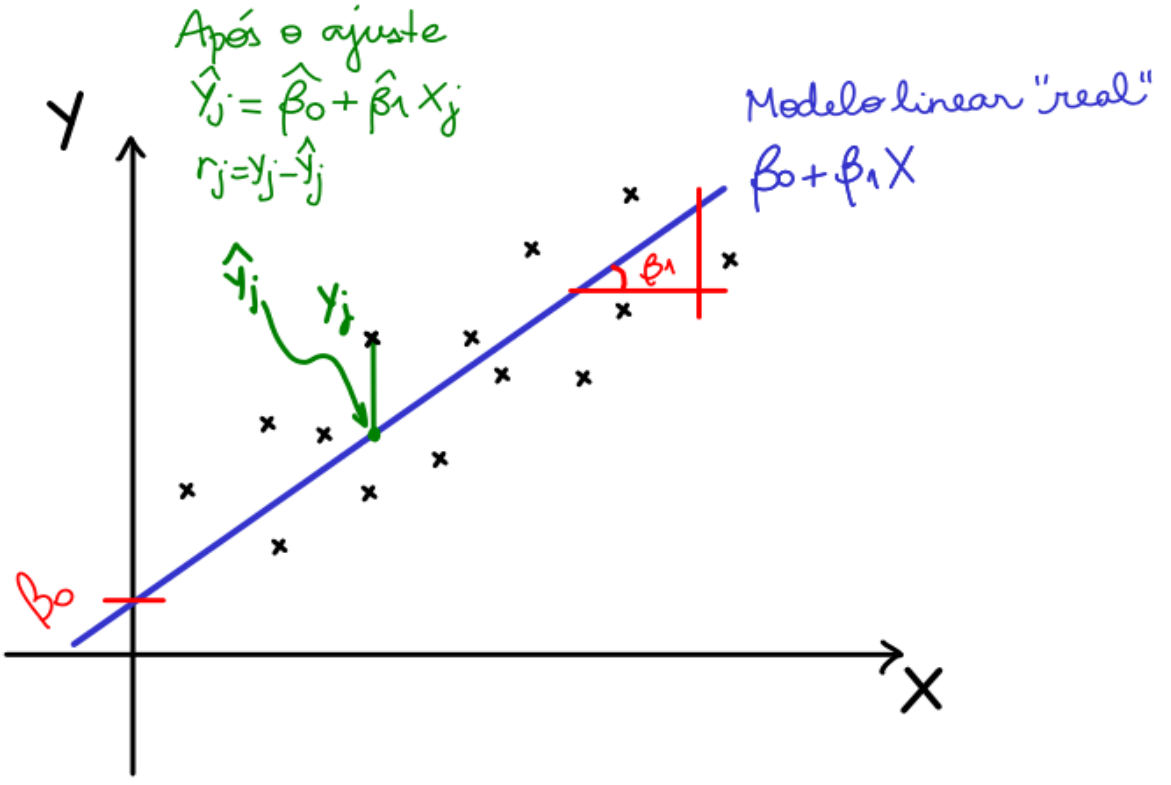


Fonte: https://stats.stackexchange.com/questions/123651/geometric-interpretation-of-multiple-correlation-coefficient-r-and-coefficient

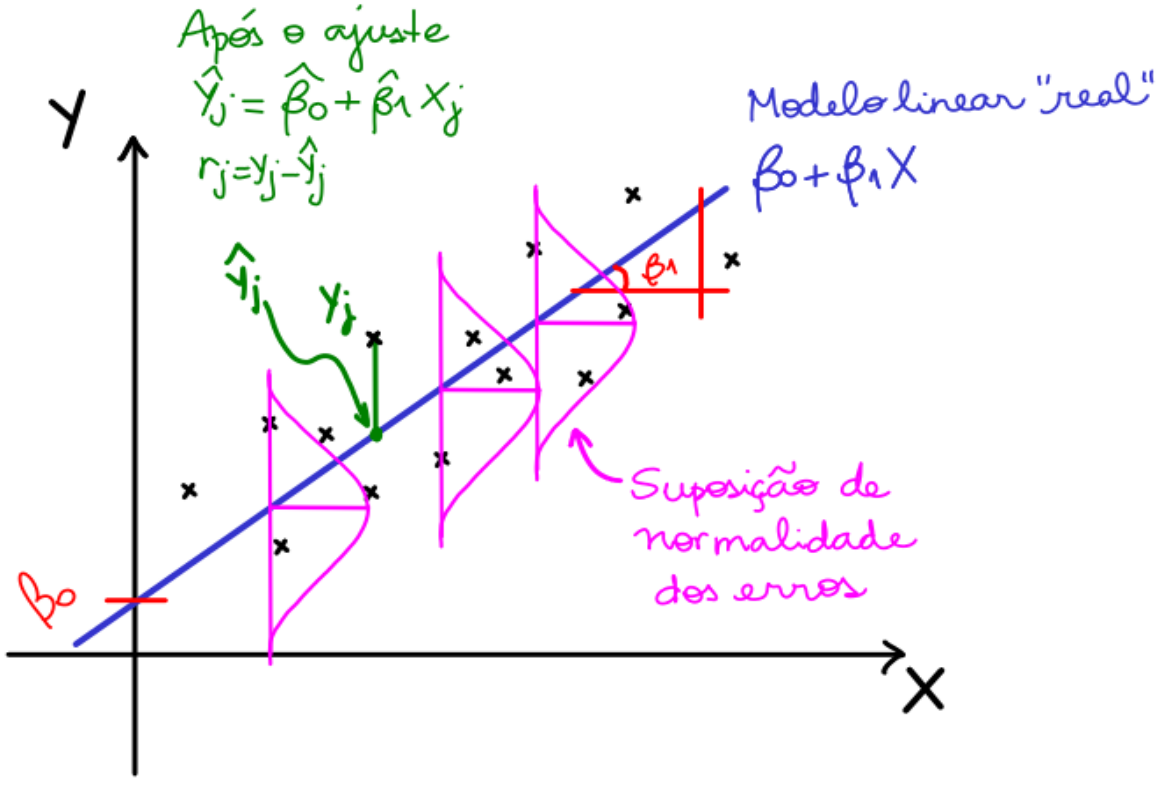
Vamos pensar no modelo de regressão linear simples.



Vamos pensar no modelo de regressão linear simples.



Vamos pensar no modelo de regressão linear simples.



E Aprendizado não-supervisionado?

Objetivo principal: Extrair informação útil de dados que não foram rotulados previamente.

Em outras palavras: Analisar dados sem uma variável específica que os categorize.

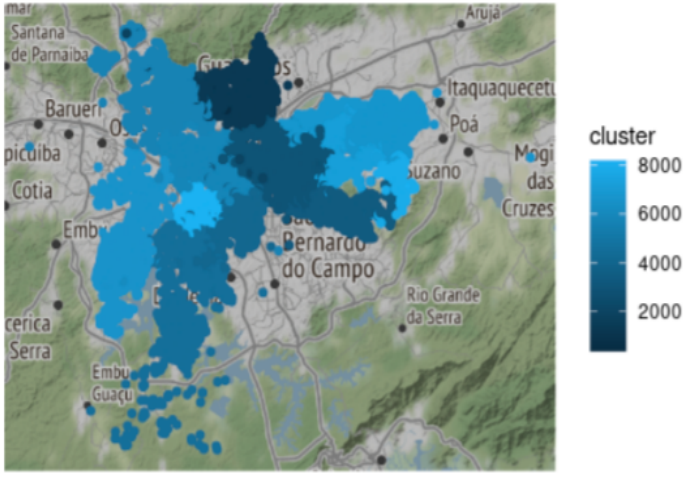


Figure 3. Nearest neighbors by default, with 33 clusters obtained

Exemplo agrupamentos. Fonte: https://www.cambridge.org/engage/miir/article-details/614501d66fc3a8bd14a2b6b2

Análise de Componentes Principais

A Análise de componentes principais (ACP ou PCA, de principal component analysis) é uma técnica que transforma linearmente um conjunto de p variáveis correlacionadas em um conjunto de k variáveis não correlacionadas (com $k < p$), que explicam uma parcela substancial das informações do conjunto original.

Objetivos principais

- Reduzir a dimensionalidade dos dados.
- Obter combinações interpretáveis das variáveis originais.
- Descrever e compreender a estrutura de correlação das variáveis originais.

Contexto

Seja X um vetor aleatório de dimensão $p \times 1$ com vetor de médias (populacionais) $\mu_{p\times 1}$ e matriz de variâncias e covariâncias (populacionais) de $\Sigma_{p\times p}$.

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $Cov(X_i, X_j), i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em reduzir a dimensionalidade do problema, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que **poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais**. Isso pode significar a **redução de custos como tempo computacional e espaço para armazenamento de dados**.

Análise de componentes principais

Seja $X \sim (\mu, \Sigma)$. Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os autovalores de Σ, com autovetores correspondentes $e_1, \ldots, e_p$, tais que

- $e_i^\top e_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
- $e_i^\top e_i = 1$, para $i = 1, \ldots, p$,
- $\Sigma e_i = \lambda_i e_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (e_1, \ldots, e_p)$.

Então $Y_{p \times 1} = O^\top X$ é o vetor de componentes principais de Σ.

Análise de componentes principais

Interpretação geométrica (p=2)

Análise de componentes principais

Interpretação geométrica (p=2)

Análise de componentes principais

Propriedades:

- A j-ésima componente principal de Σ é dada por

$$Y_j = {e_j}^\top X.$$

- $E(Y_j) = {e_j}^\top \mu$.
- $Var(Y_j) = {e_j}^\top \Sigma e_j = \lambda_j$.
- $Cov(Y_i, Y_j) = Cor(Y_i, Y_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
- A proporção da variância total de X que é explicada pela j-ésima componente principal é

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}.$$

Estimação das componentes principais

Como em geral a matriz Σ é desconhecida, utiliza-se a matriz S, de variâncias e covariâncias amostrais, para estimar as componentes principais.

Considere $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ os autovalores de S, com autovetores correspondentes padronizados $\hat{e}_1, \ldots, \hat{e}_p$.

A j-ésima componente principal amostral é dada por

$$\hat{Y}_j = \hat{e}_j^\top X.$$

Exemplo Tempo computacional

Produto de Kronecker

Como a matemática pode melhorar o desempenho dos algoritmos?

Sejam $A_{m\times n} = (a_{ij})$ e $B_{p\times q} = (b_{ij})$.

O **produto de Kronecker** entre A e B é definido por

$$(A \otimes B)_{mp\times nq} = (a_{ij}B).$$

```
Sigma = matrix(c(3,1,2,1,2,1,2,1,4),byrow=T,nrow=3)
A = kronecker(diag(2000),Sigma)
start_time <- Sys.time()
B = solve(A)
end_time <- Sys.time()
end_time - start_time
```

```
# Time difference of 35.79782 secs
```

```
start_time <- Sys.time()
C = solve(kronecker(diag(2000),Sigma))
end_time <- Sys.time()
end_time - start_time
```

```
# Time difference of 37.59477 secs
```

```
start_time <- Sys.time()
C = kronecker(solve(diag(2000)), solve(Sigma))
end_time <- Sys.time()
end_time - start_time
```

```
# Time difference of 1.839286 secs
```

Os caminhos para a/o Cientista de Dados

- O que estudar?
- Como começar?
- Disposição para aprender?
- Como manter a curva de aprendizado com derivada positiva?
- Quais os benefícios?

Para onde seguir depois?

- Especialização? MBA?
- Mestrado?
- Outra graduação? Por que não?

Obrigada!

- **https://icmc.usp.br/pessoas/cibele**.
- **https://www.youtube.com/c/CibeleRussoUSP**.
- **https://github.com/cibelerusso**.
- **https://www.linkedin.com/in/cibelerusso/**.
- **https://www.researchgate.net/profile/Cibele-Russo**.