INTRODUÇÃO AO BIG DATA

Encontro Paraibano de Estatística - EPBEST 2016

Marcus Nunes

19 e 20 de maio de 2016

Universidade Federal do Rio Grande do Norte

QUEM SOU EU?

QUEM SOU EU?

- Sou Marcus Nunes, Ph.D. em Estatística pela Penn State University
- Professor na UFRN
- Meus interesses principais são as aplicações da Estatística em grandes conjuntos de dados, como genética, climatologia e saúde
- marcus.nunes@ccet.ufrn.br
- `http://marcusnunes.me/epbest-2016/`

SOBRE O QUE É ESTE MINICURSO?

SOBRE O QUE É ESTE MINICURSO?

- Uma introdução ao Big Data
- Conceitos e ideias sobre o assunto
- Ninguém vai se tornar um expert em Big Data
- Espero que vocês saiam daqui com mais perguntas do que respostas

SOBRE O QUE É ESTE MINICURSO?

- Definição de Big Data
- O que um Estatístico deveria saber
- Como obter dados para análise
- Visualização de dados
- Agrupamento de Dados
- Aplicação

O QUE É BIG DATA?

O QUE É BIG DATA?

- Não existe consenso a respeito de uma definição sobre o que realmente é big data
- A área ainda é nova; não houve tempo para o conhecimento sedimentar
- Em geral, diz respeito a áreas do conhecimento onde as ferramentas de análise de dados tradicionais não são a melhor escolha possível

O QUE É BIG DATA?

- Big Data são os dados que possuem 3 V:
- Volume
- Velocidade
- Variedade

O QUE É BIG DATA?

- Uma outra definição de Big Data se vale da Estatística para ser formulada
- Podemos considerar um conjunto de dados como Big Data se o tempo que levamos para ajustar um modelo aos dados é maior do que o tempo utilizado para a escolha deste modelo

O QUE É BIG DATA?

- Mike Franklin, da Universidade de Berkeley, diz o seguinte:
- “Big Data é todo conjunto de dados caro para manter e manipular e de onde é difícil extrair informações”
- Esta definição é relativa: para alguns, dados na casa dos terabytes podem ser caros para manter; para outros, dados na casa dos petabytes podem ser baratos para manter

QUEM TRABALHA COM BIG DATA?

QUEM TRABALHA COM BIG DATA?

- Competências de um profissional 100% capacitado para trabalhar com Big Data:
  - Estatística
  - Programação
  - Negócios
  - Conhecer bem a área de atuação (internet, marketing, área financeira etc)

QUEM TRABALHA COM BIG DATA?

- Que tipo de profissionais temos no momento?
  - Bons estatísticos e matemáticos que escrevem códigos sem otimização
  - Bons cientistas da computação que entendem um pouco de estatística e matemática
  - Bons cientistas da computação que entendem um pouco de negócios, depois de muita experiência na área
  - Especialistas em alguma área de atuação
  - Gerentes que sabem fazer estas pessoas trabalharem juntas

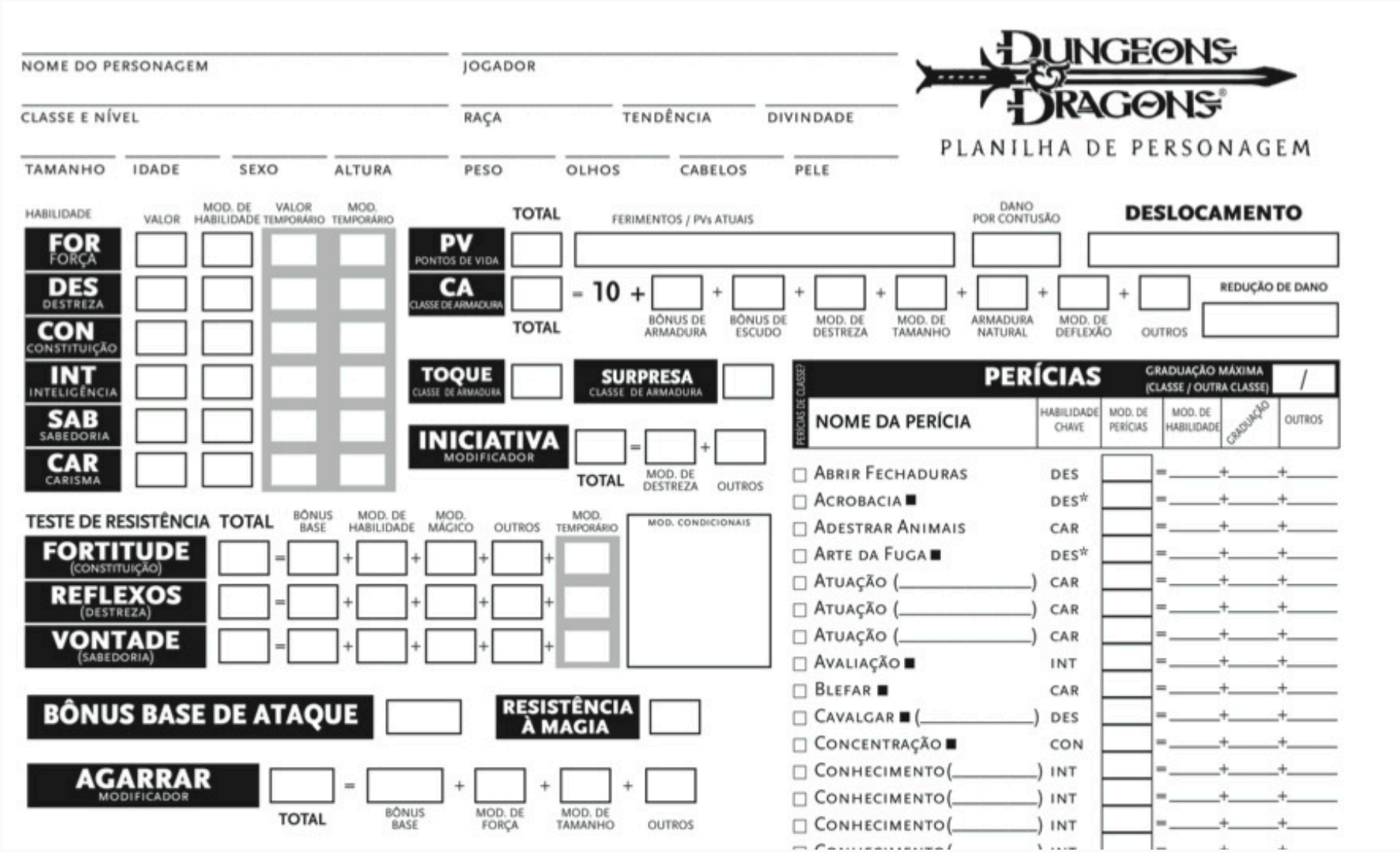
Quem trabalha com Big Data?

- Estatísticos
- Programadores
- Físicos
- Cientistas de Dados

QUEM TRABALHA COM BIG DATA?

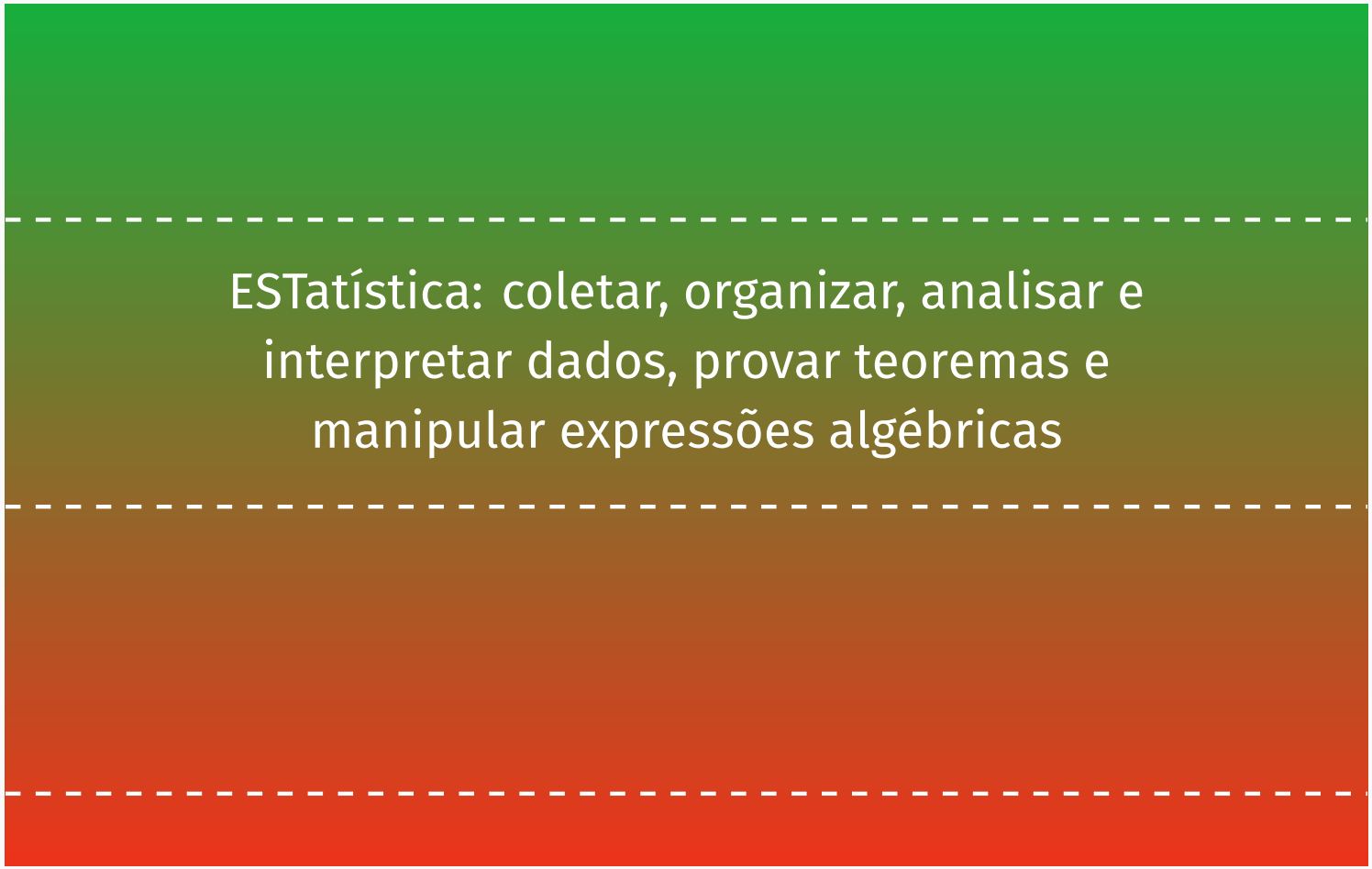
- Cientista de Dados (*Data Scientist*) é um novo nome para Estatístico
- Alguns dizem que o Cientista de Dados é um Estatístico que mora em São Francisco e usa um Mac
- No fundo, ambos são a mesma coisa, embora uma destas profissões trabalhe melhor seu marketing pessoal

QUEM JÁ JOGOU RPG?



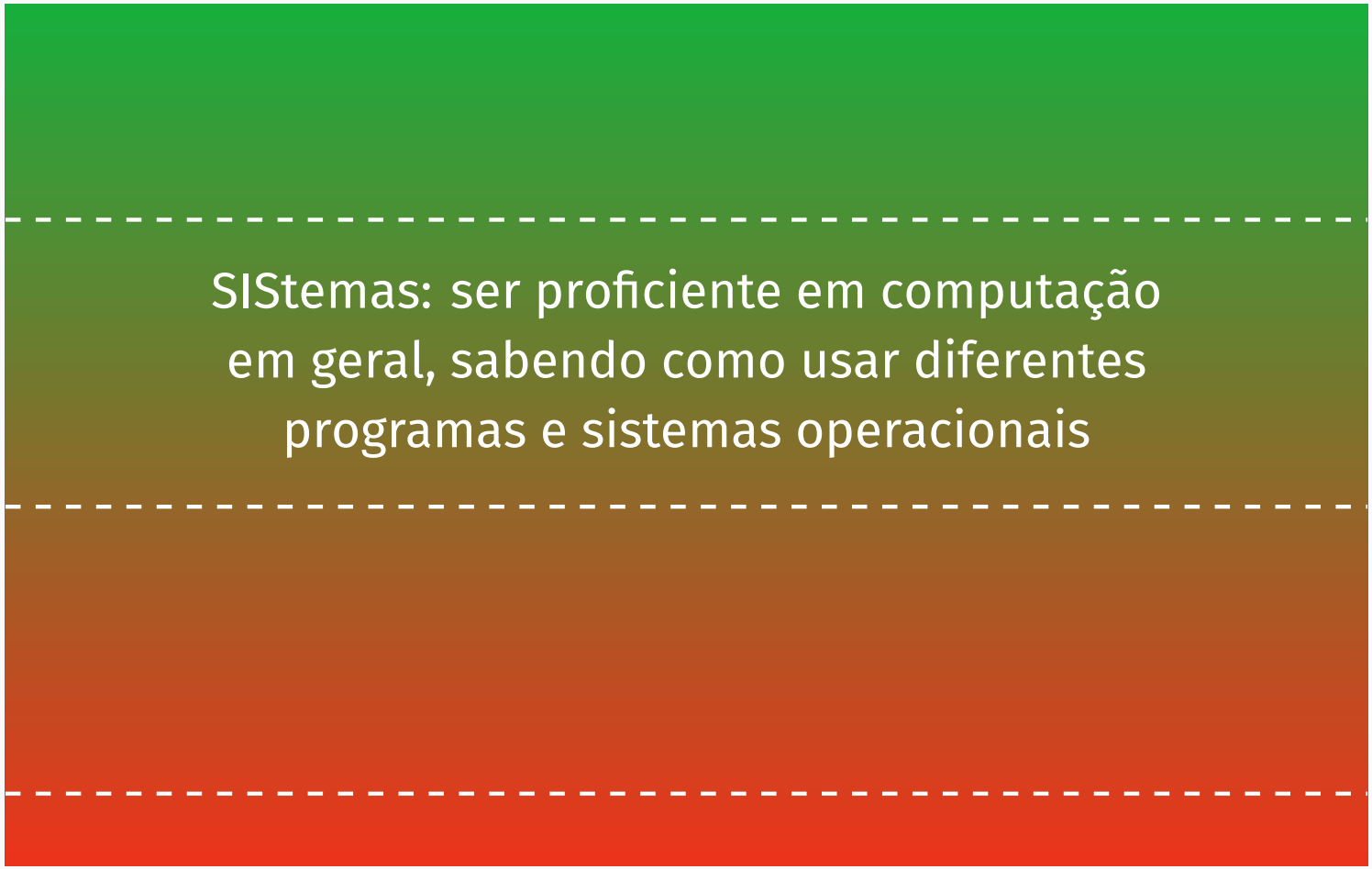
MALCOLM GLADWELL - OUTLIERS

O QUE UM CIENTISTA DE DADOS PRECISA SABER



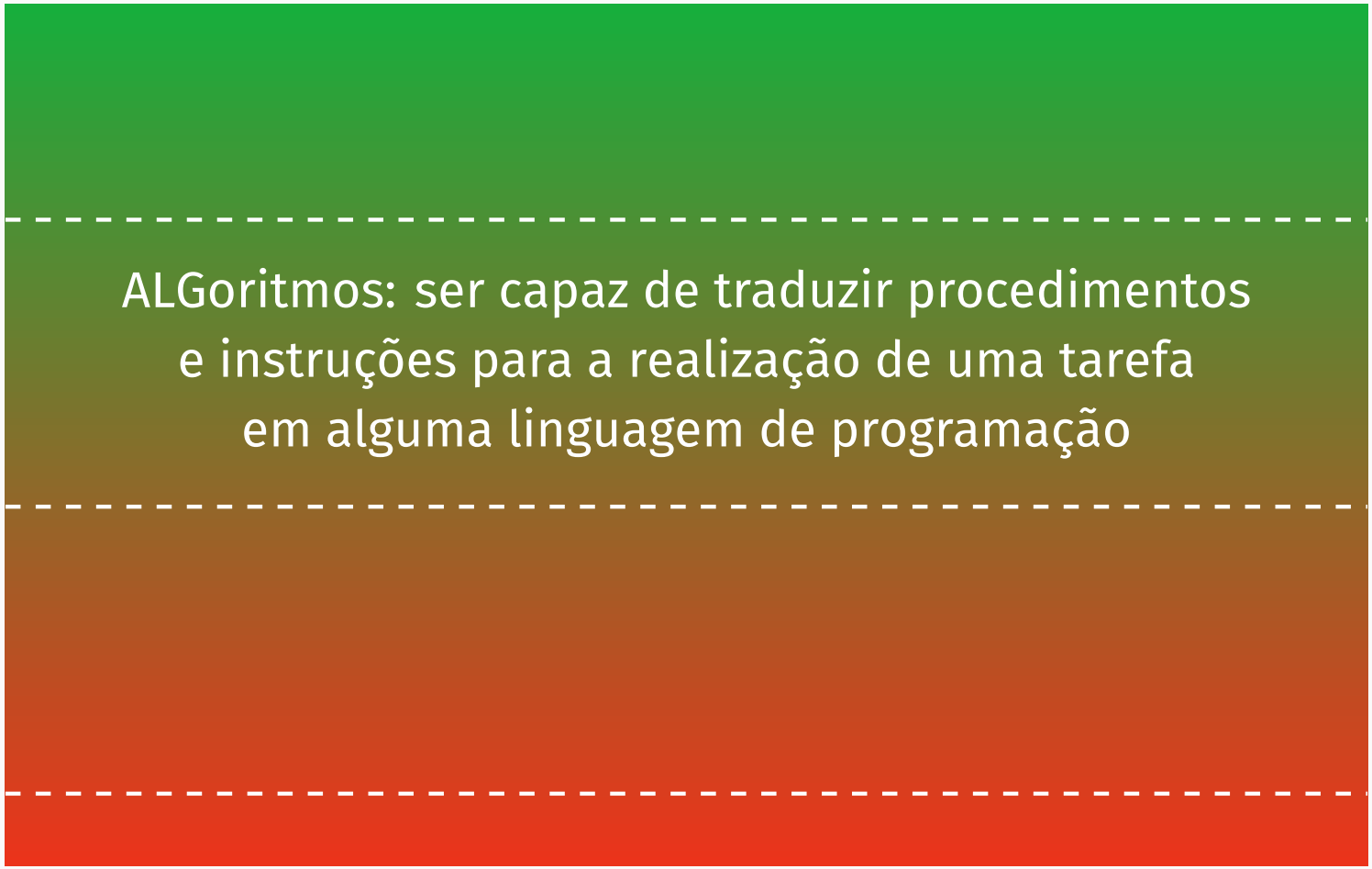
EST

O QUE UM CIENTISTA DE DADOS PRECISA SABER



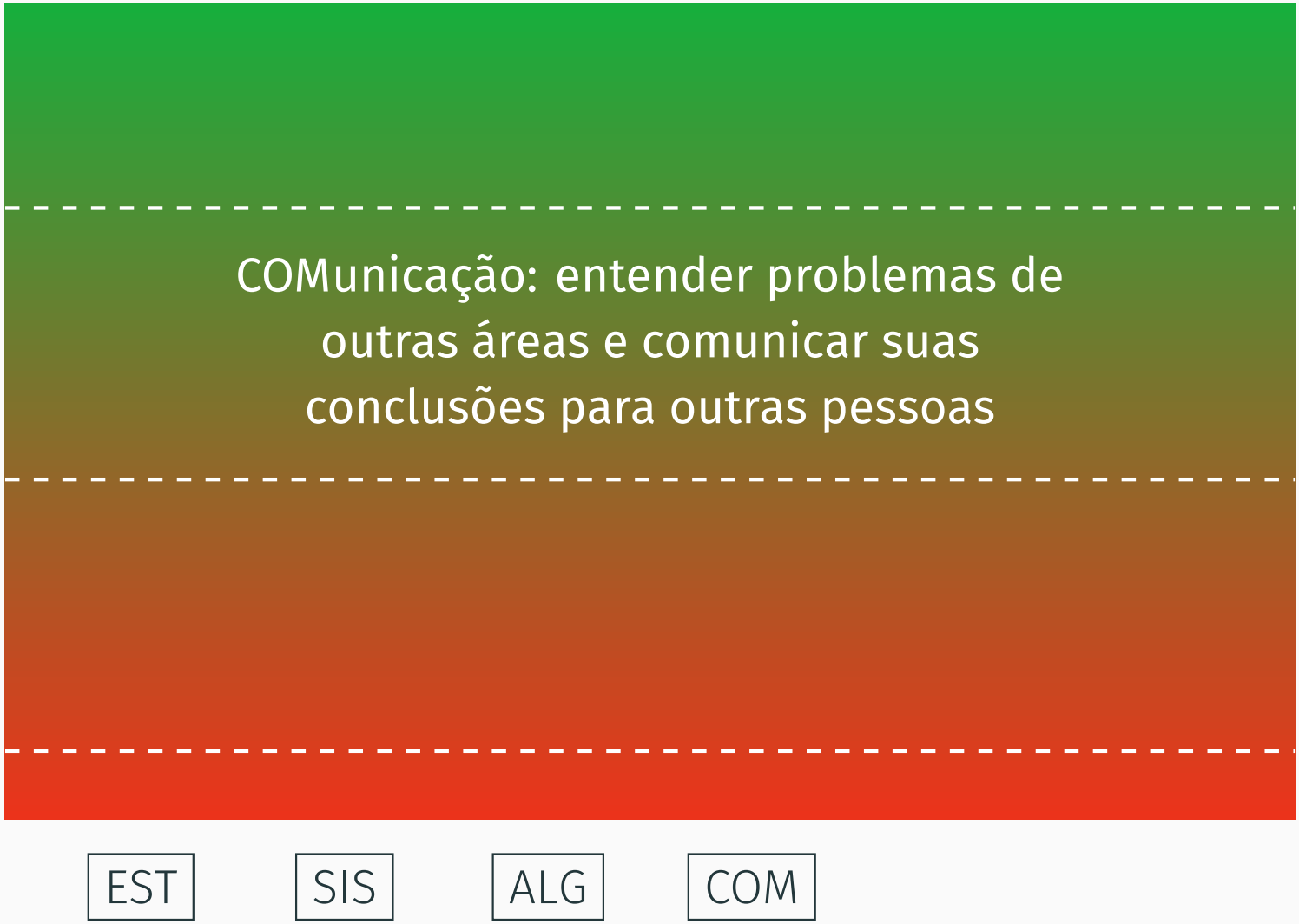
EST SIS

O QUE UM CIENTISTA DE DADOS PRECISA SABER

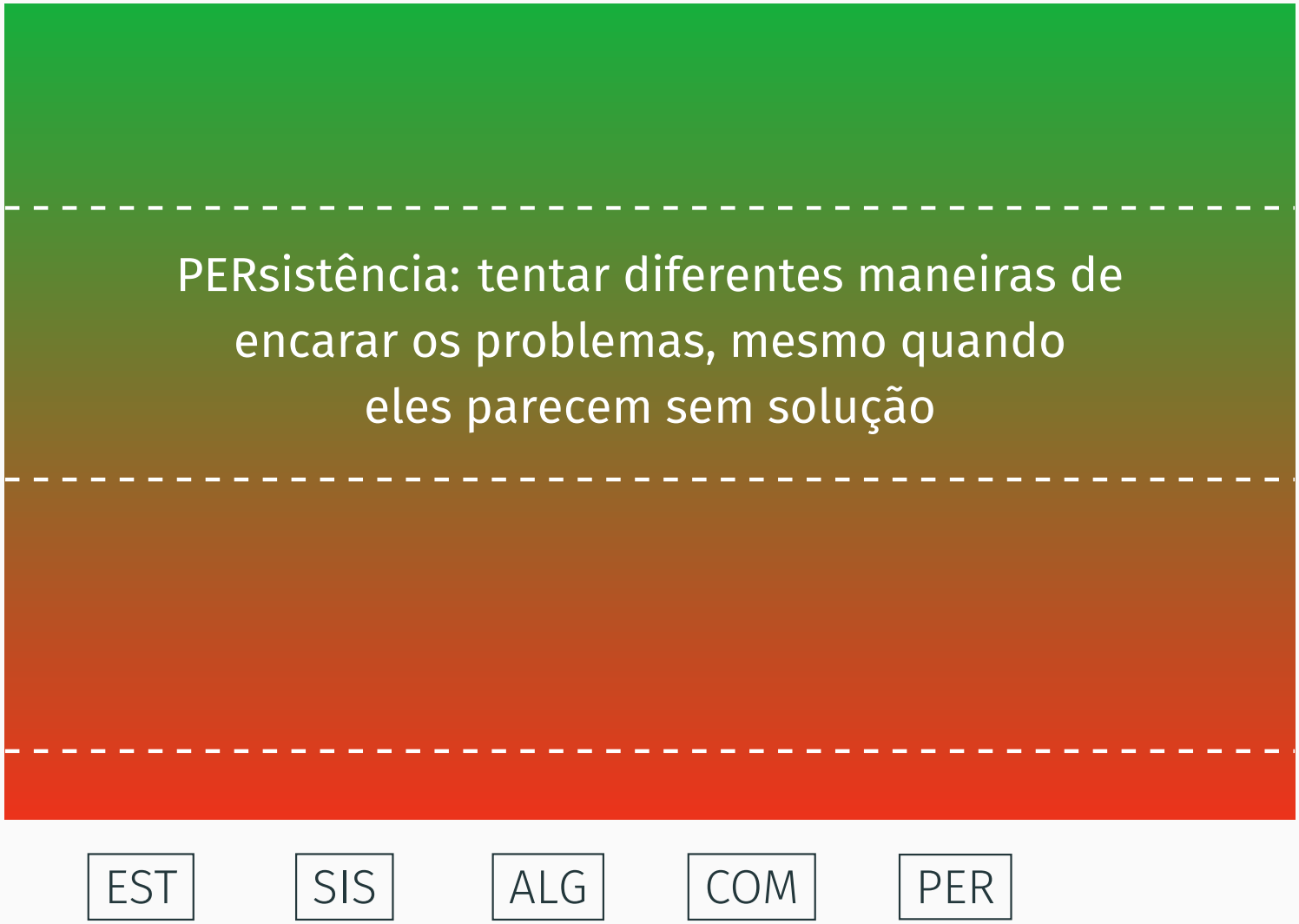


EST SIS ALG

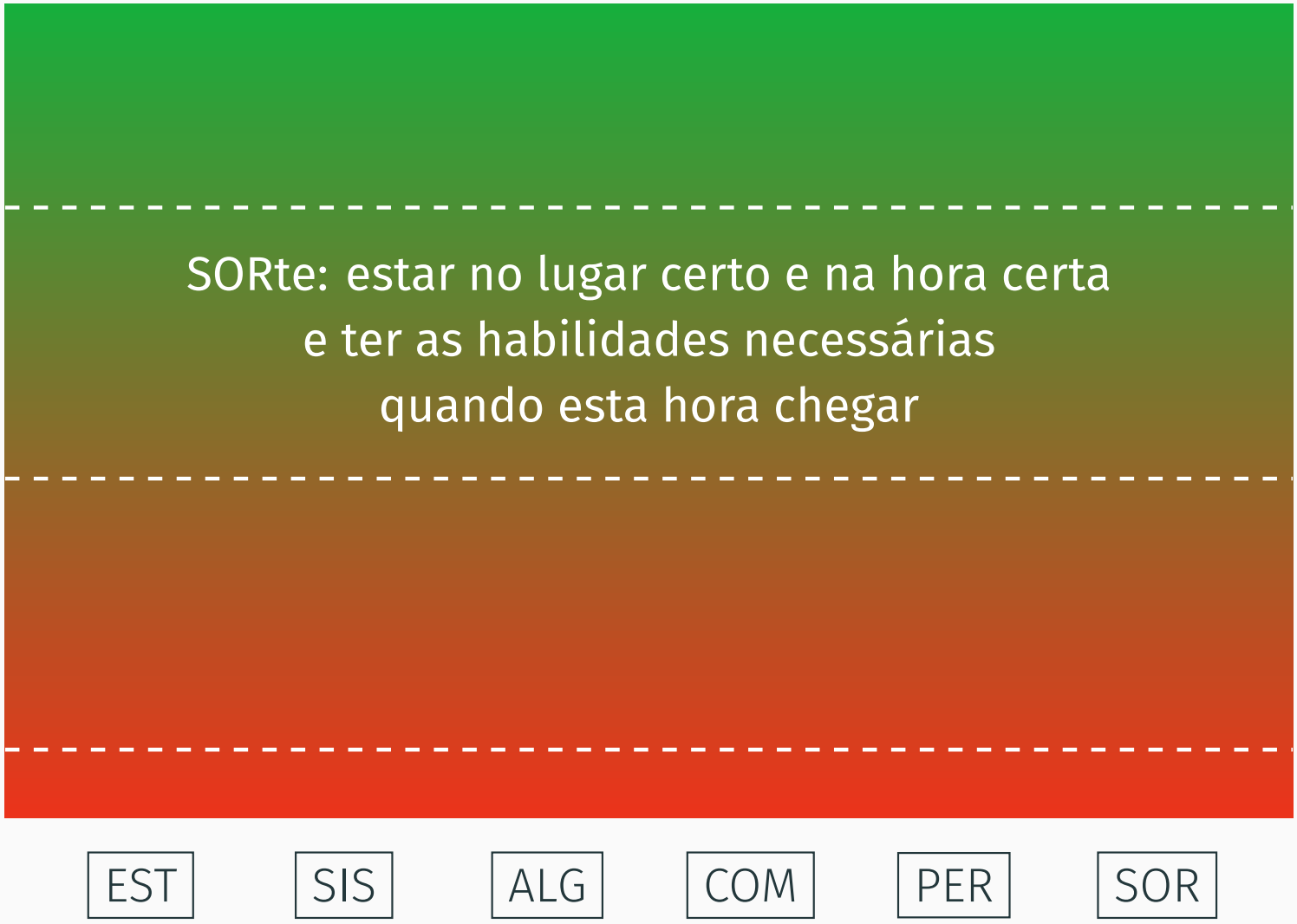
O QUE UM CIENTISTA DE DADOS PRECISA SABER



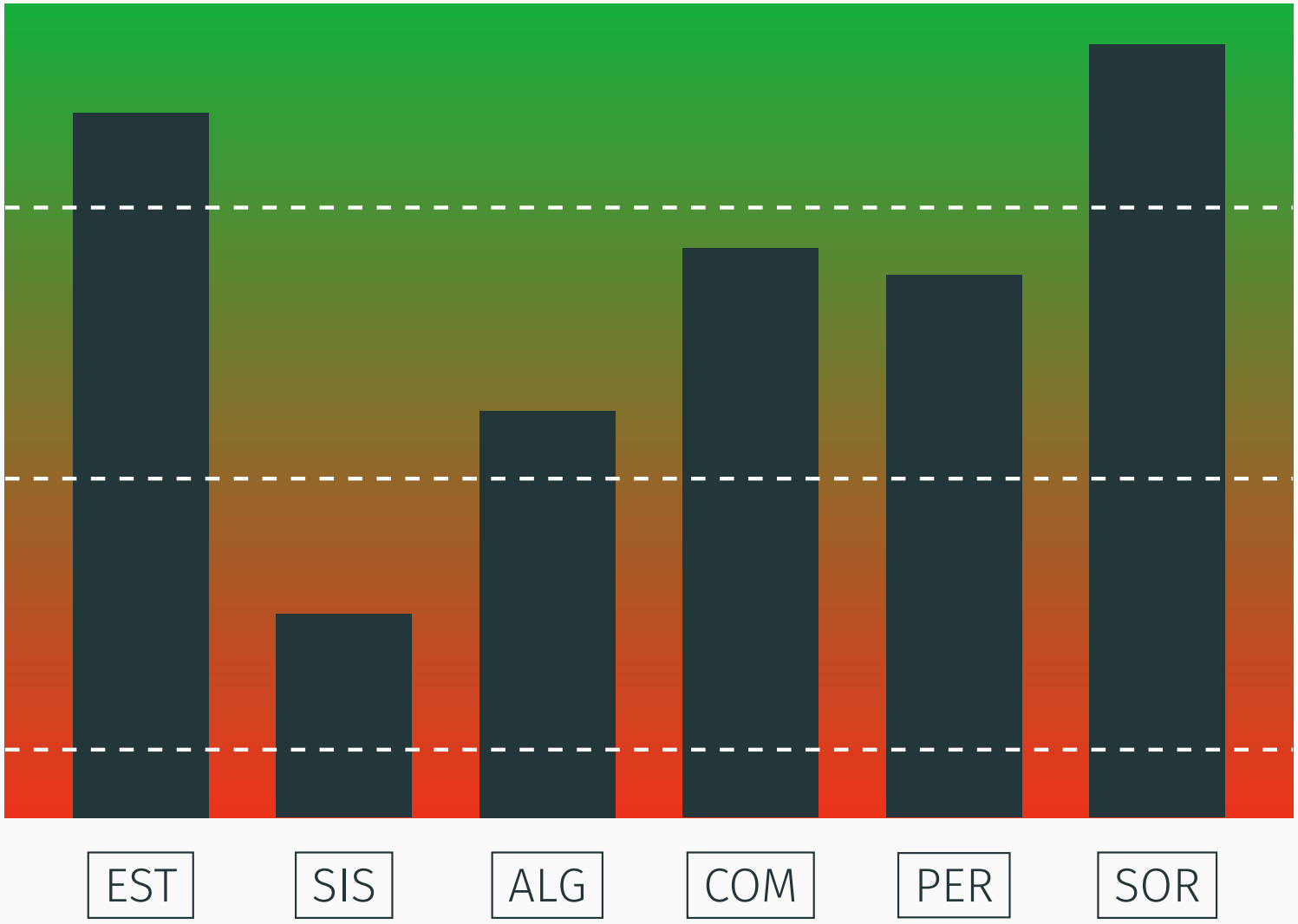
O QUE UM CIENTISTA DE DADOS PRECISA SABER



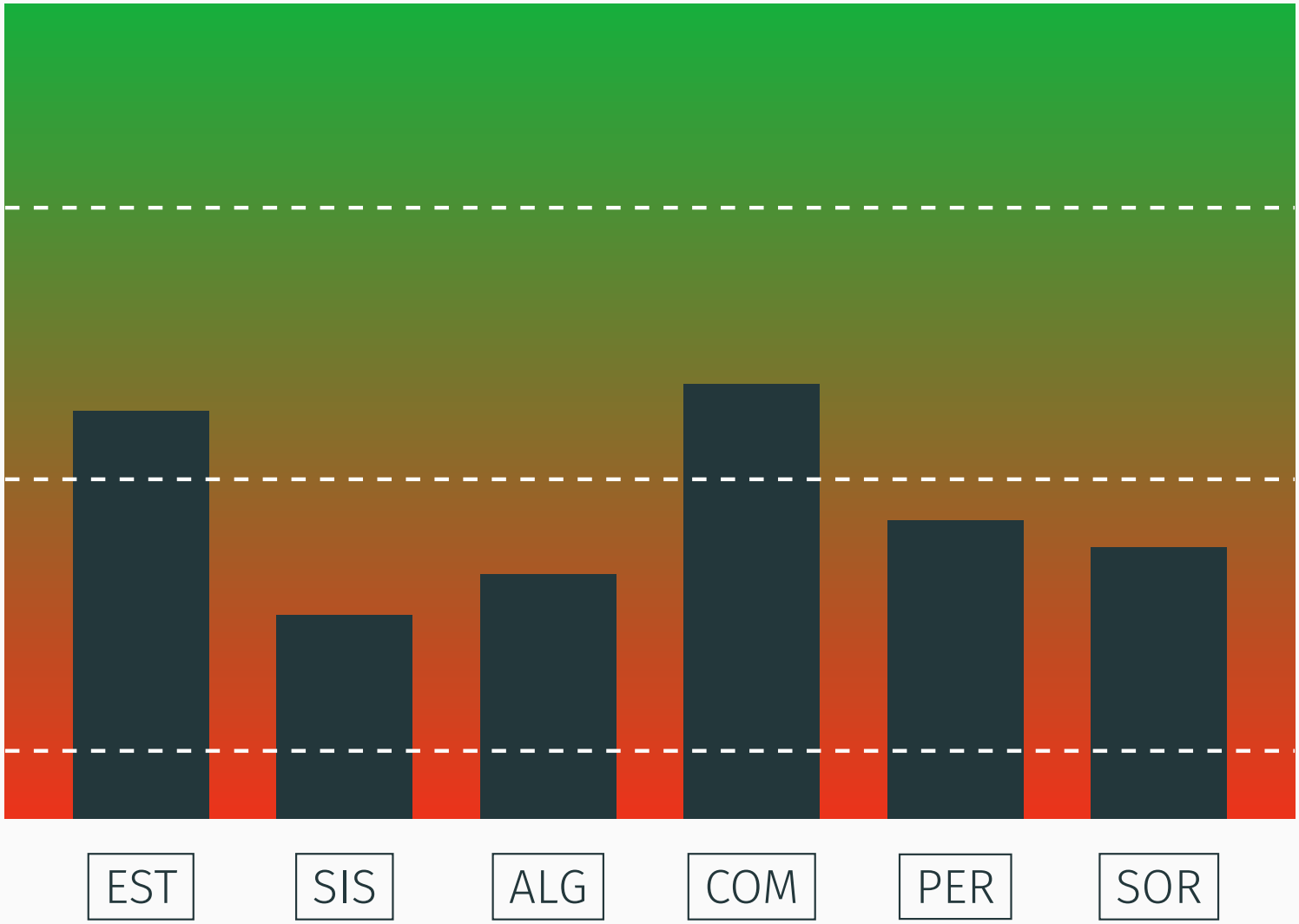
O QUE UM CIENTISTA DE DADOS PRECISA SABER



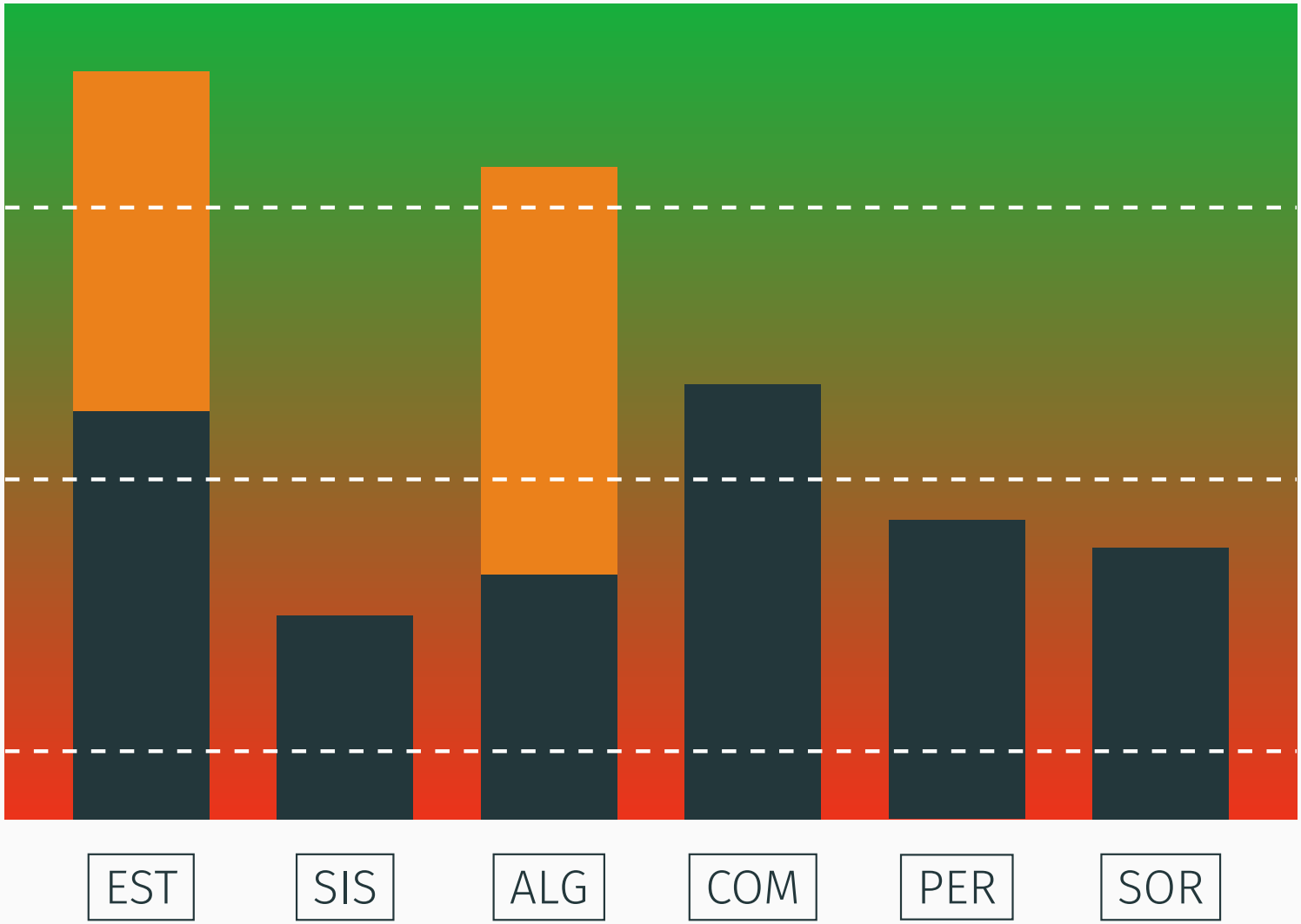
O QUE EU SEI



NÃO É BOM ESTAR NA MÉDIA

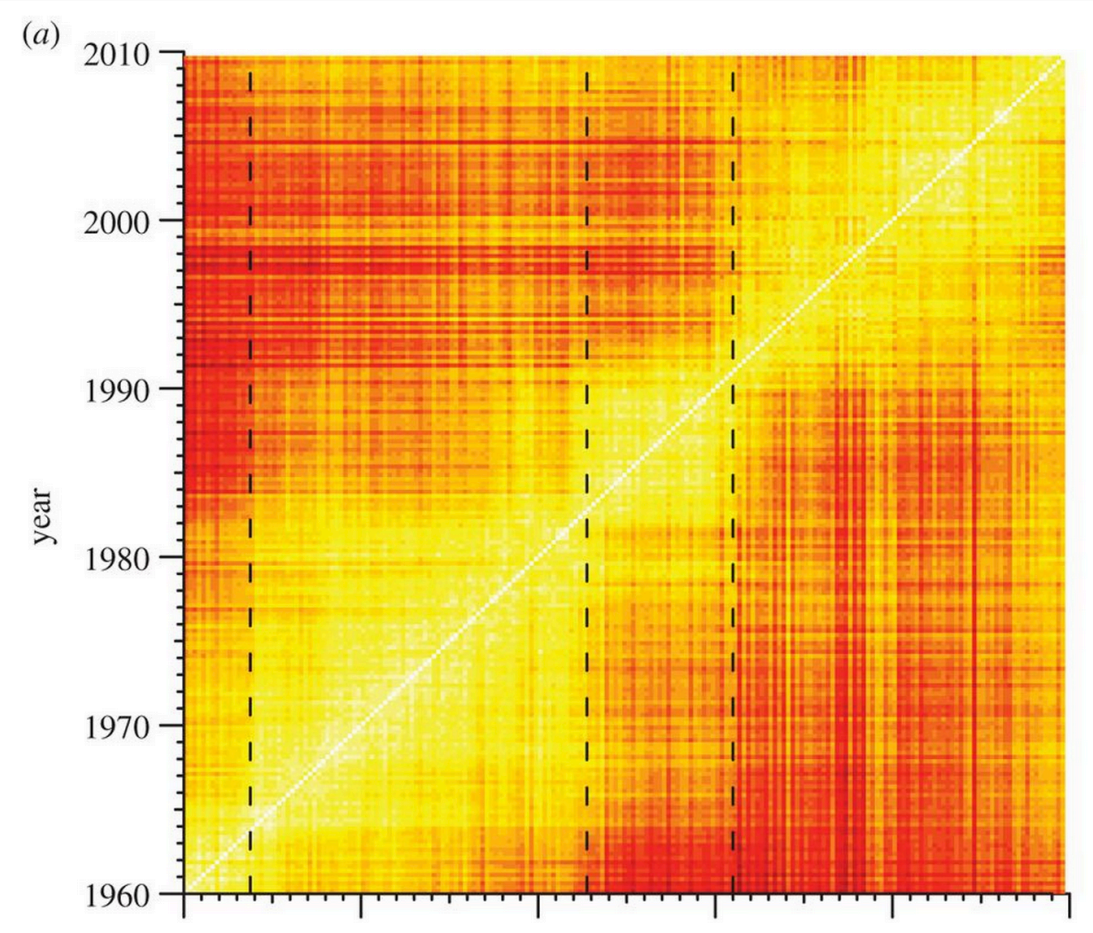


SEJA MUITO BOM EM ALGUMAS ÁREAS



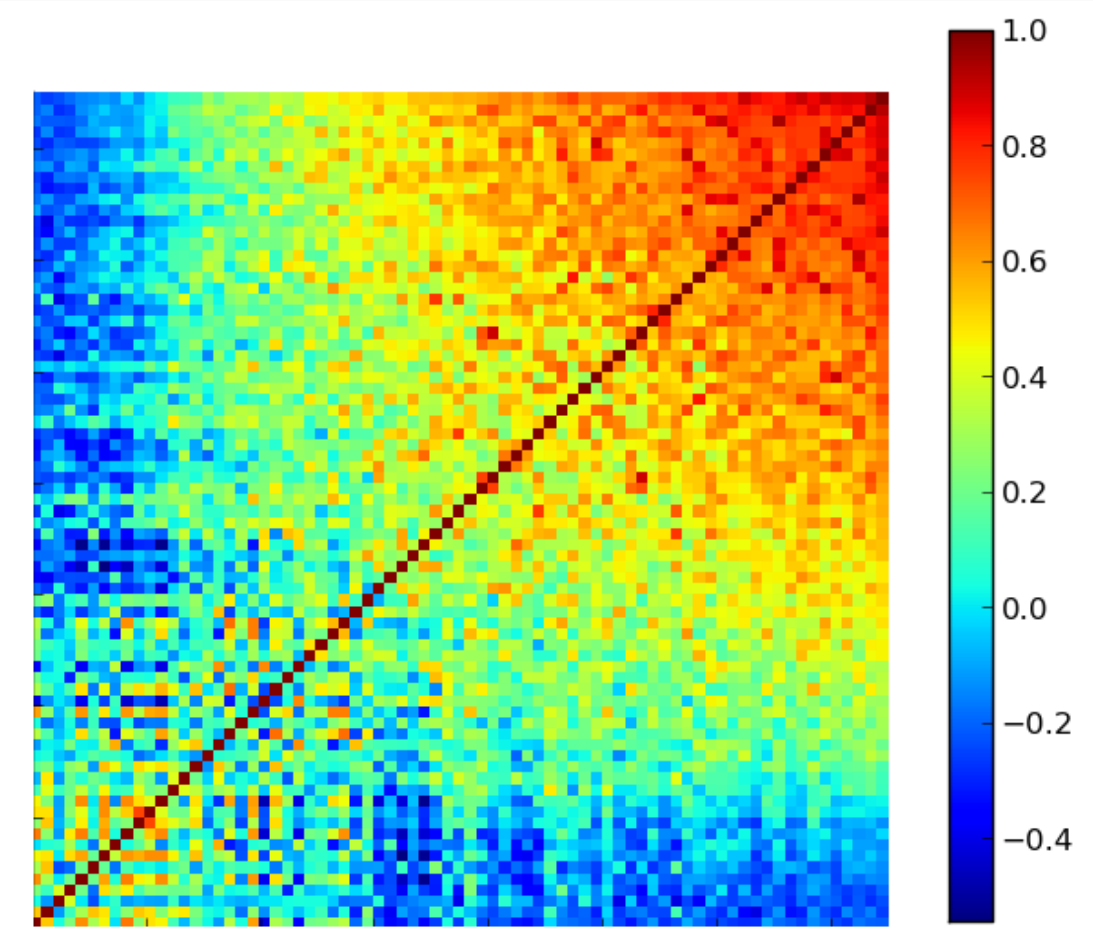
BIG DATA NA PRÁTICA

THE EVOLUTION OF POPULAR MUSIC: USA 1960–2010



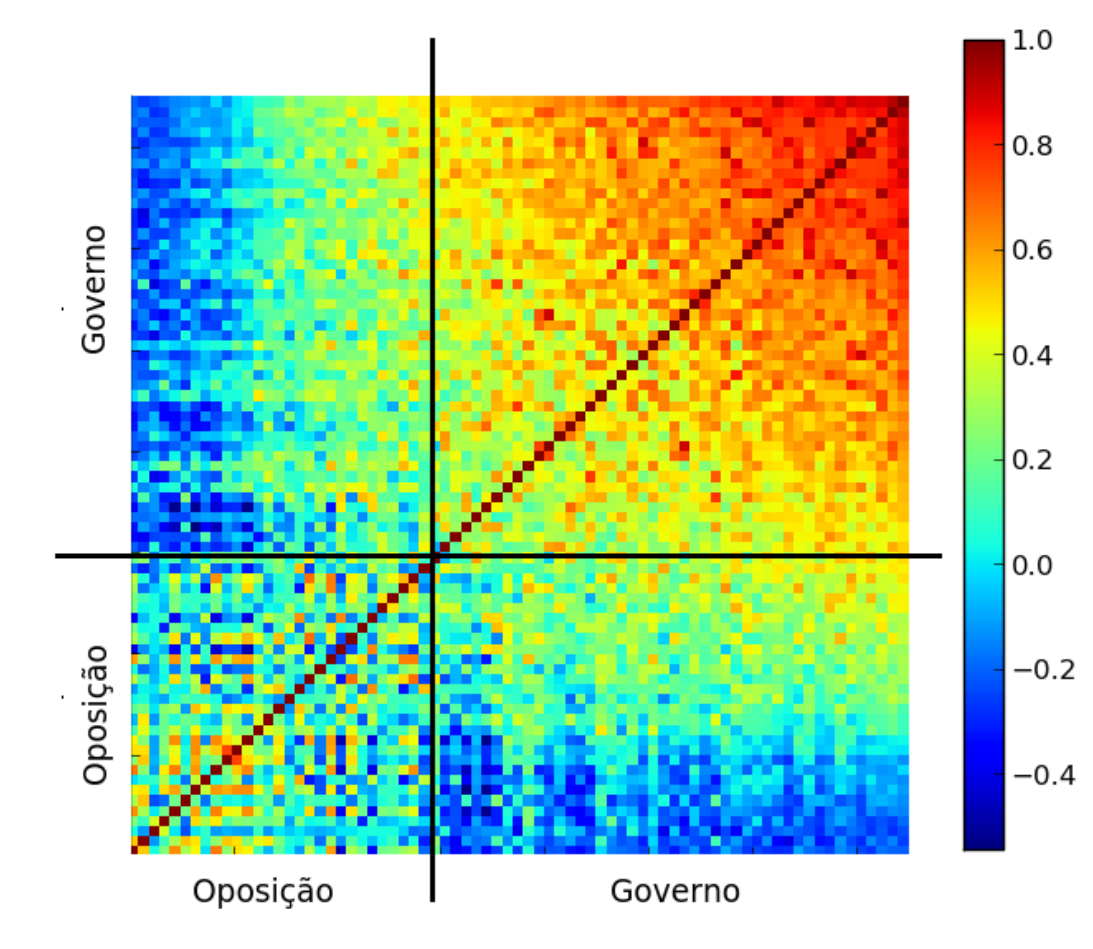
Fonte: http://rsos.royalsocietypublishing.org/content/2/5/150081/

VOTOS DOS SENADORES NO BRASIL

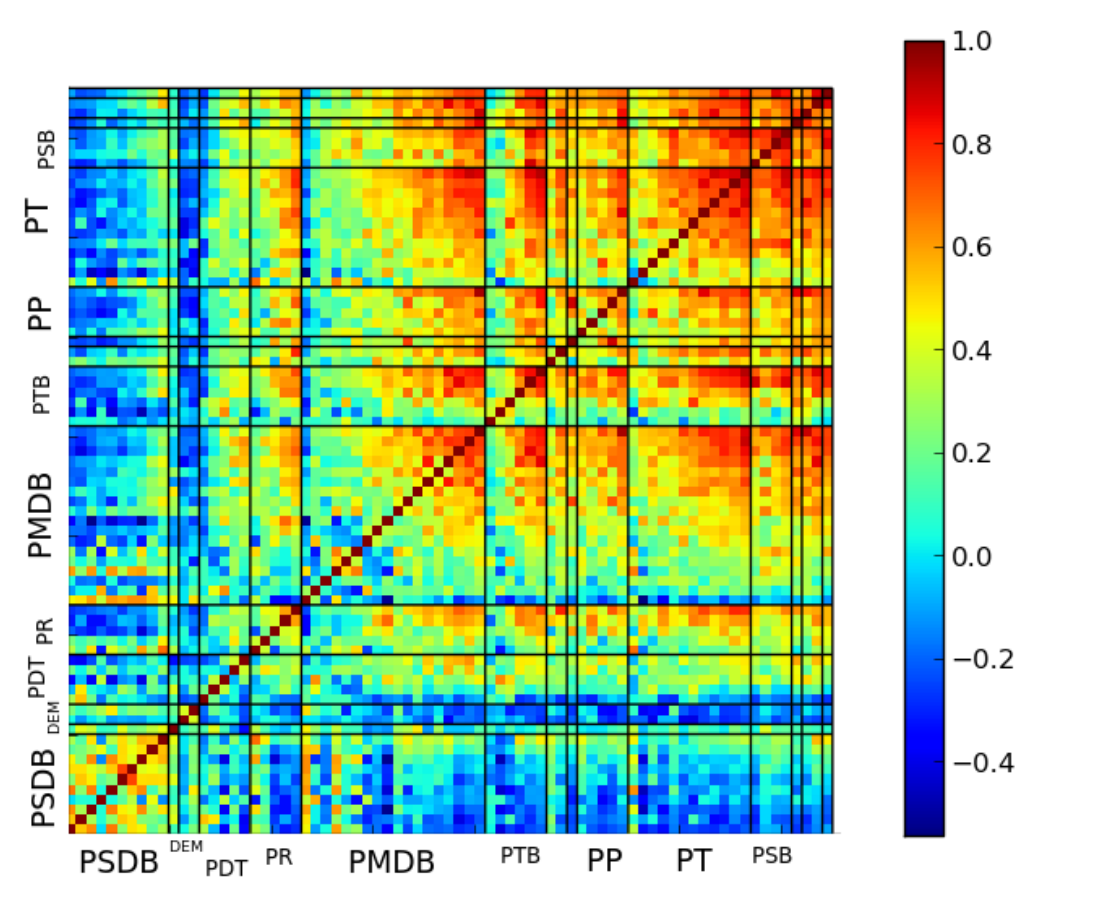


Fonte: http://www.todasasconfiguracoes.com/2013/09/14/ha-partidos-politicos-no-brasil/

VOTOS DOS SENADORES NO BRASIL



VOTOS DOS SENADORES NO BRASIL



TARGET

- How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Fonte:
`http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/`

TARGET

- Compartilhamos dados com empresas sempre que fazemos compras
- As empresas usam estes dados para traçar o perfil dos clientes
- Com isto, podem enviar ofertas direcionadas para cada segmento de público

TARGET

- Certa vez, um pai chegou furioso a uma loja da Target em Minneapolis, exigindo falar com o gerente
- Ele argumentava que a filha recebeu, pelo correio, ofertas de roupas de bebê e berços
- Detalhe: a filha ainda era aluna do Ensino Médio

TARGET

- Na lógica do pai, isto estaria encorajando a filha a engravidar
- O gerente da Target se desculpou e disse que ia averiguar o que tinha acontecido
- Nada de estranho foi descoberto; mesmo assim, o gerente ligou para o pai uns dias depois e se desculpou novamente

TARGET

- Exemplo de perfil traçado pela Target:
- Imagine uma cliente chamada Jenny Ward, 23 anos, moradora de Atlanta
- Se em março ela comprou
    - hidratante de manteiga de cacau
    - uma bolsa grande
    - suplementos de zinco e magnésio
    - um tapete azul claro
- Segundo o modelo da loja, há 87% de chance de ela estar grávida e dar a luz no fim de agosto

TARGET

- No caso específico da Target, o modelo deles é capaz de enviar correspondência para as clientes grávidas em momentos bastante específicos da gestação
- Além disso, serve para indicar como organizar fisicamente as lojas (por exemplo, colocar fraldas ao lado de cerveja)
- É possível descobrir que, quando casam, as pessoas comprar tipos diferentes de cereal

TARGET

- Conclusão: depois da desculpa por telefone do gerente, o pai voltou a ligar para a loja
- Desta vez, ele se desculpou e disse que “haviam eventos em sua casa de que ele não estava completamente a par”
- De fato, a menina estava grávida e deu a luz poucos meses depois do contato da loja

NETFLIX

- How Netflix Reverse Engineered Hollywood
- To understand how people look for movies, the video service created 76,897 micro-genres
- We took the genre descriptions, broke them down to their key words... and built our own new-genre generator
  Fonte: `http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/`

NETFLIX

- Gêneros de filmes em locadoras? Ação, Aventura, Comédia, Drama etc.
- Gêneros de filmes na Netflix? African-American Crime Documentaries, Scary Cult Movies from the 1980s, Feel-good Romantic Spanish-Language TV Shows, Visually Striking Latin American Comedies e mais
- Assim, é mais fácil sugerir filmes que podem agradar o usuário

- Por exemplo, O Tesouro de Sierra Madre e Os Vingadores - A Era de Ultron são ambos filmes de aventura

NETFLIX

- É possível que ambos os filmes estivessem numa mesma prateleira de uma locadora
- Na Netflix, um dos filmes poderia ser classificado como “Aventura Dramática Passada no México na Década de 1920” e o outro como “Aventura com Grupo de Super-Heróis e Grande Orçamento”
- Notem como é possível especificar o gênero do filme tanto quanto quisermos

SPOTIFY

- Serviço de streaming de música
- 20+ milhões de assinantes, 75+ milhões de usuários ativos (Junho de 2015)

SPOTIFY

- Terabytes de dados são produzidos pelos usuários
- É possível relacionar dados dos usuários para sugerir novos artistas e canções
- Críticos não são necessários neste sistema: novas sugestões são criadas a partir de algoritmos

Fonte: `https://labs.spotify.com/tag/big-data/`

Experimento: Séries de TV

SÉRIES DE TV

- Vamos fazer um experimento
- Fãs de Game of Thrones ficarão em um grupo e fãs de The Walking Dead ficarão em outro
- Eu darei minhas opiniões sobre estas séries e vocês julgarão meu conhecimento
- Minhas afirmações serão genéricas a ponto de não estragar a surpresa de quem não assistiu estas séries ainda, mas específicas a ponto de me fazer compreender por quem já as assistiu

SÉRIES DE TV

0 - Nunca assistiu

1 - Assistiu alguns episódios aleatórios

2 - Assistiu alguns episódios conectados

3 - Assistiu muitos episódios

4 - Assistiu a maioria episódios

5 - Assistiu todos os episódios

SÉRIES DE TV - GAME OF THRONES

- De modo geral, a quarta temporada foi a melhor de todas, embora não seja possível dizer que foi muito melhor do que as outras
- As temporadas tendem a começar bem, piorar um pouco e ficarem muitos boas na reta final
- Os melhores episódios ocorreram no final da terceira e da quinta temporadas (todavia, não foram os últimos episódios)
- O pior episódio foi “Unbowed, Unbent, Unbroken”, o sexto episódio da quinta temporada (Sansa Stark casa com Ramsay Bolton)
- A quinta temporada foi a mais irregular

Séries de TV - The Walking Dead

- Não há uma temporada que se destaque das demais
- A terceira temporada começou bem, mas foi piorando com o tempo
- Os melhores episódios na quarta e quinta temporadas
- Entretanto, a quinta temporada começou muito bem, decaiu e se recuperou no final
- O pior episódio foi “Still”, episódio 11 da quarta temporada (Daryl e Beth na floresta)

SÉRIES DE TV

0 - Nunca assistiu

1 - Assistiu alguns episódios aleatórios

2 - Assistiu alguns episódios conectados

3 - Assistiu muitos episódios

4 - Assistiu a maioria episódios

5 - Assistiu todos os episódios

SÉRIES DE TV

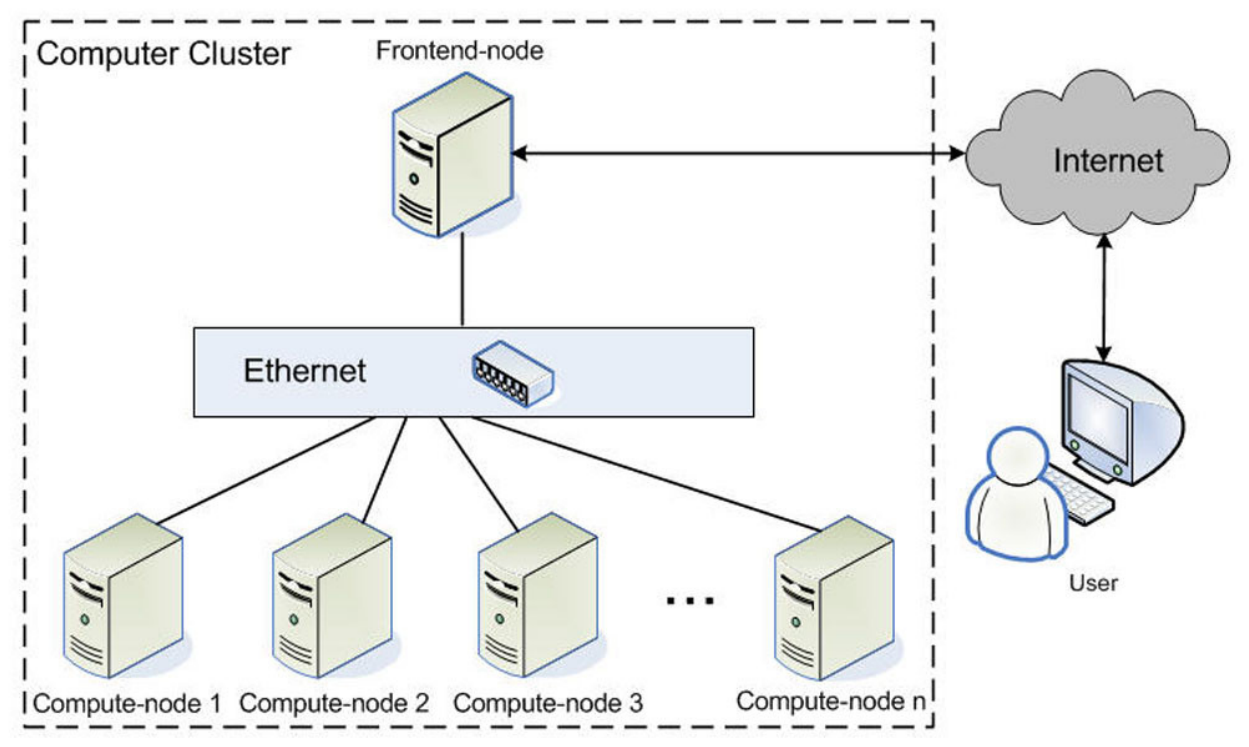
- Game of Thrones: 0 (nunca assisti)
- The Walking Dead: 0 (nunca assisti)
- Tirei minhas afirmações através de uma análise estatística utilizando o site `http://www.imdb.com/`

COMPUTAÇÃO DISTRIBUÍDA

COMPUTAÇÃO DISTRIBUÍDA

- Computação Distribuída ou Computação Paralela
- Otimização do trabalho computacional
- “Dividir para conquistar” - Napoleão Bonaparte

IDEIA BÁSICA



Fonte: BioMed Central

- Rack de computadores

Fonte: `http://racksolutions.com`

- Coleção de racks que formam o National Super Computer Center em Guangzhou, China

Fonte: `http://www.independent.co.uk/`

LISTA DE SUPERCOMPUTADORES

- TOP500 Supercomputer Sites - `http://top500.org`
- Lista atualizada duas vezes por ano, com os maiores supercomputadores do mundo

TOP 500 - NOVEMBRO DE 2015

1. National Super Computer Center in Guangzhou - China
2. DOE/SC/Oak Ridge National Laboratory - Estados Unidos
3. DOE/NNSA/LLNL - Estados Unidos
4. RIKEN Advanced Institute for Computational Science (AICS) - Japão
5. DOE/SC/Argonne National Laboratory - Estados Unidos

TOP 500 - NOVEMBRO DE 2015 - BRASIL

1. Laboratório Nacional de Computação Científica (201)
2. SENAI CIMATEC (242)
3. Laboratório Nacional de Computação Científica (266)
4. Laboratório Nacional de Computação Científica (311)
5. Petróleo Brasileiro S.A. (407)

Mas vocês tem um supercomputador na sua frente!

- (ou quase)
- Todo PC moderno multicore tem a capacidade de realizar trabalhos em paralelo
- Os códigos dos programas devem ser adaptados para isso

COMPUTAÇÃO PARALELA NO PC

COMPUTAÇÃO PARALELA NO PC

- Existem diversas maneiras de paralelizar um código no PC
- Como já dito, um computador com vários núcleos pode paralelizar um programa em si mesmo
- Uma rede de computadores pode paralelizar um programa dentro dela
- Na última década, GPUs tem sido muito utilizadas para computação paralela

COMPUTAÇÃO PARALELA NO PC

- Não é possível executar automaticamente um código tradicional de maneira paralela
- Os códigos devem ser adaptados para isso
- Na maioria das vezes, este processo não é trivial

LEI DE AMDAHL

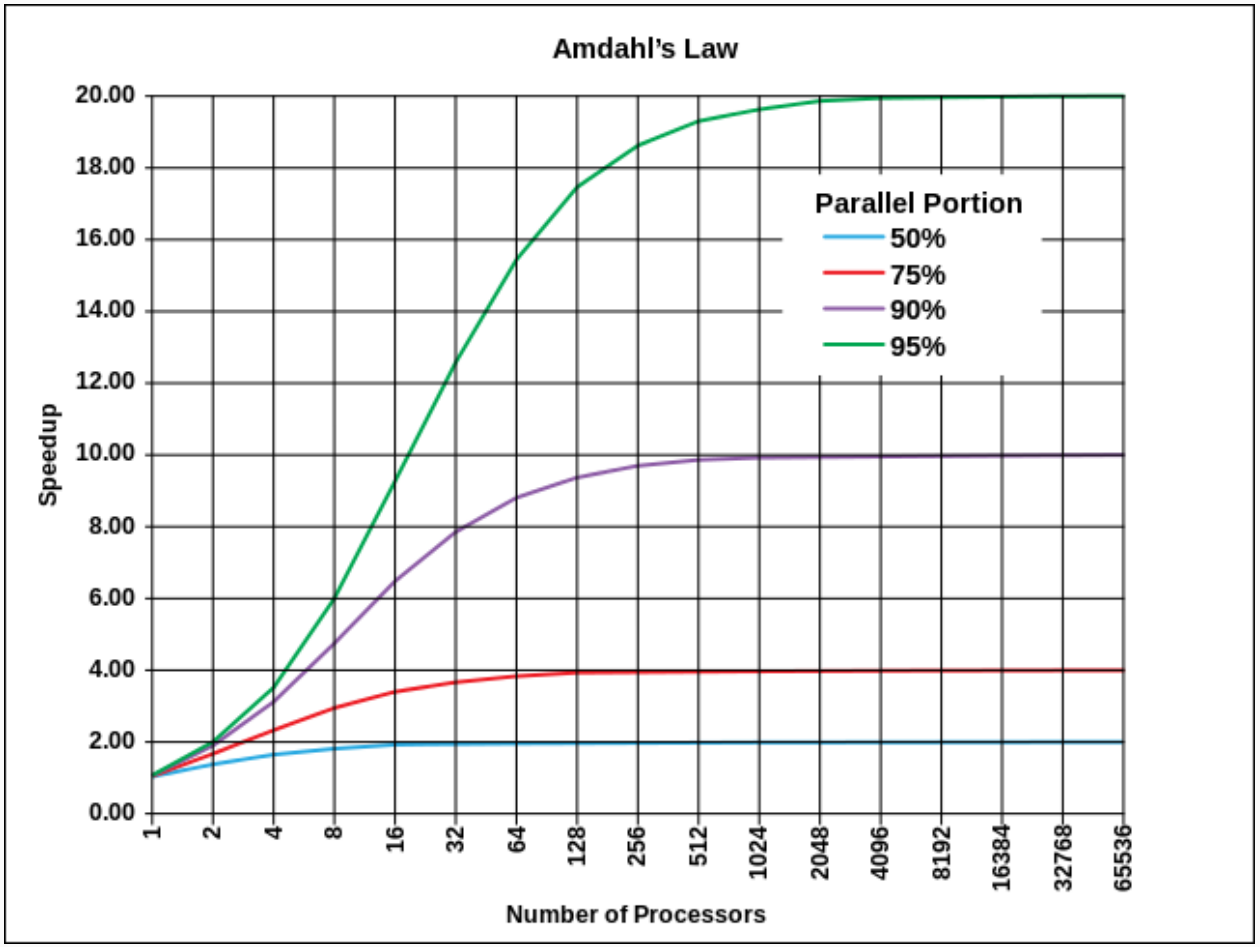
- Não há uma relação totalmente linear entre o número de processadores que um sistema tem e a velocidade de processamento
- É de se esperar que, ao aumentar o número de processadores de um sistema, ocorra o seguinte:

No. de Processadores	Tempo de Execução
1	1000
2	500
4	250
8	125

LEI DE AMDAHL

- Infelizmente, a tabela anterior não é sempre válida
- Em geral, há partes de um programa que não podem ser paralelizadas
- Desta forma, o número de processadores não influencia diretamente no tempo de execução do programa
- Além disso, há um limite teórico máximo para a paralelização de um programa

LEI DE AMDAHL



Fonte: Wikipedia

OPÇÕES DE SOFTWARE

- Não existe uma maneira correta e definitiva de se paralelizar um código
- Há várias alternativas disponíveis no mercado
- Felizmente, a imensa maioria delas é gratuita
- Infelizmente, quase todas são exclusivas para *nix (Linux, Unix, OS X)

OPÇÕES DE SOFTWARE

- MPI (Message Passing Interface)
- Hadoop
- Pacotes no R

COMPUTAÇÃO PARALELA NO R

- Existem diversos pacotes que lidam com computação em paralelo no R
- Por exemplo, `snow`, `multicore`, `Rmpi`, `foreach` e `parallel`
- Veremos algumas aplicações do `parallel` no R

ESTIMAÇÃO DO TEMPO DE PROCESSAMENTO

- É possível medir o tempo de execução de um código no R
- Há duas maneiras de fazermos isto: através dos comando `proc.time` e `system.time`
- Veremos exemplos destes dois métodos a seguir

ESTIMAÇÃO DO TEMPO DE PROCESSAMENTO

```
> n         <- 100
> repl      <- 10
> resultado <- rep(0, repl)
> t.inicial <- proc.time()
> for (j in 1:repl){
+   resultado[j] <- det(matrix(rnorm(n^2), ncol=n))
+ }
> proc.time()-t.inicial

##    user  system elapsed
##   0.018   0.002   0.021
```

ESTIMAÇÃO DO TEMPO DE PROCESSAMENTO

```
> set.seed(1)
> n         <- 100
> repl      <- 10
> resultado <- rep(0, repl)
> system.time(
+   for (j in 1:repl){
+     resultado[j] <- det(matrix(rnorm(n^2), ncol=n))
+   }
+ )

##    user  system elapsed
##   0.013   0.001   0.014
```

PARALELIZAÇÃO UTILIZANDO O PACOTE PARALLEL

- Este pacote é distribuído com o R desde a versão 2.14
- Baseado nos pacotes `multicore` e `snow`
- Particularmente eficiente para rodar códigos do tipo SPMD (single program, multiple data), uma subcategoria do tipo de paralelização MIMD (multiple instruction, multiple data)

- Na técnica SPMD, processadores independentes executam o mesmo código em conjuntos de dados diferentes, sendo estes conjuntos reais ou simulados
- Na técnica MIMD, processadores independentes executam códigos diferentes em conjuntos de dados diferentes, sendo estes conjuntos reais ou simulados
- Por ser menos geral, a técnica SPMD é menos versátil do que a MIMD, mas é mais fácil de compreender e de implementar

PARALELIZAÇÃO UTILIZANDO O PACOTE PARALLEL

- A utilização do pacote `parallel` depende da compreensão do que são listas no R
- A principal ideia é paralelizar a criação de listas
- É possível utilizar tanto múltiplos cores da máqina, quanto máquinas diferentes para executar o código

PARALELIZAÇÃO UTILIZANDO O PACOTE PARALLEL

- É possível pedir para o R detectar o número de núcleos presentes na máquina

```
> library(parallel)
> detectCores(logical=FALSE)
```

```
## [1] 2
```

PARALELIZAÇÃO UTILIZANDO O PACOTE PARALLEL

- A seguir, veremos um exemplo de como paralelizar um código em R
- O código a seguir calcula todas as somas dos x primeiros naturais
- Sim, seria possível otimizar o código utilizando funções já implementadas no R, mas não é nosso interesse aqui

```
> f <- function(x){
+   soma <- 0
+   for (j in 1:x){
+       soma <- soma + j
+   }
+   return(soma)
+ }
```

PARALELIZAÇÃO UTILIZANDO O PACOTE PARALLEL

```
> n <- 10000
>
> system.time(
+   resultado.padrao <- lapply(1:n, FUN=f)
+ )

##    user  system elapsed
##  11.251   0.146  11.960

> system.time(
+   resultado.mc01 <- mclapply(X=1:n, FUN=f, mc.cores=1)
+ )

##    user  system elapsed
##  11.095   0.086  11.368
```

PARALELIZAÇÃO UTILIZANDO O PACOTE PARALLEL

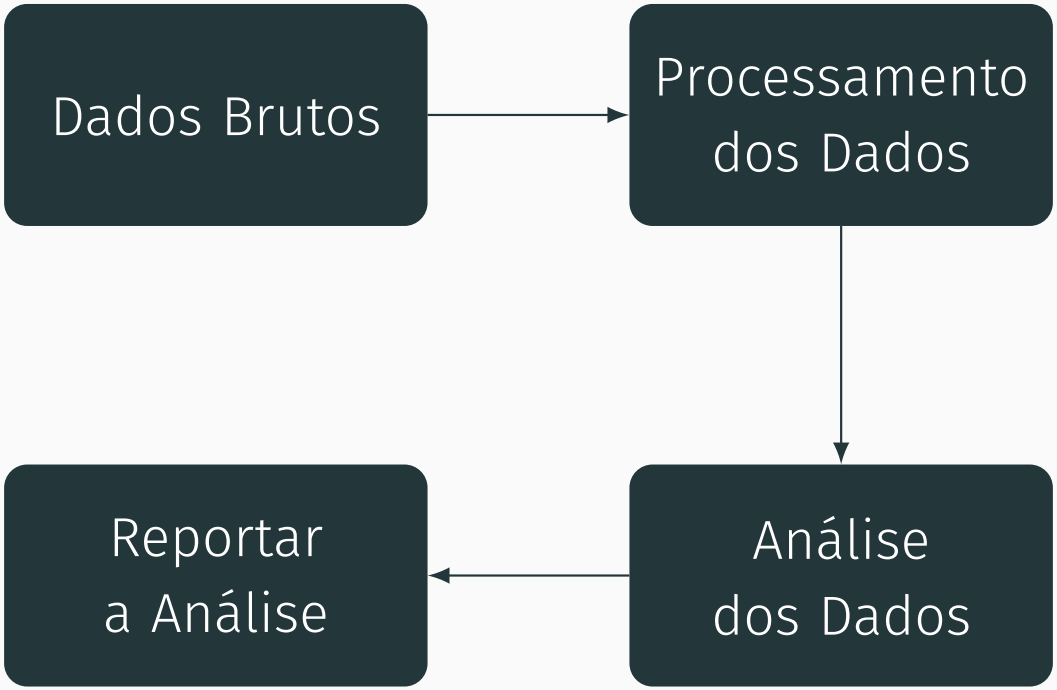
```
> system.time(
+   resultado.mc02 <- mclapply(X=1:n, FUN=f, mc.cores=2)
+ )

##    user  system elapsed
##   6.220   0.104   6.467

> system.time(
+   resultado.mc04 <- mclapply(X=1:n, FUN=f, mc.cores=4)
+ )

##    user  system elapsed
##  23.320   0.343   6.158
```

PARALELIZAÇÃO UTILIZANDO O PACOTE PARALLEL

- Perceba que não há diferença em utilizar os comandos `lapply` (tradicional) ou `mclapply` (`multicore apply` com 1 core)
- A opção com 2 cores é a mais rápida
- Não há ganho de desempenho ao aumentar o número de cores de 2 para 4, pois meu computador só possui 2 cores

ANÁLISE DE DADOS

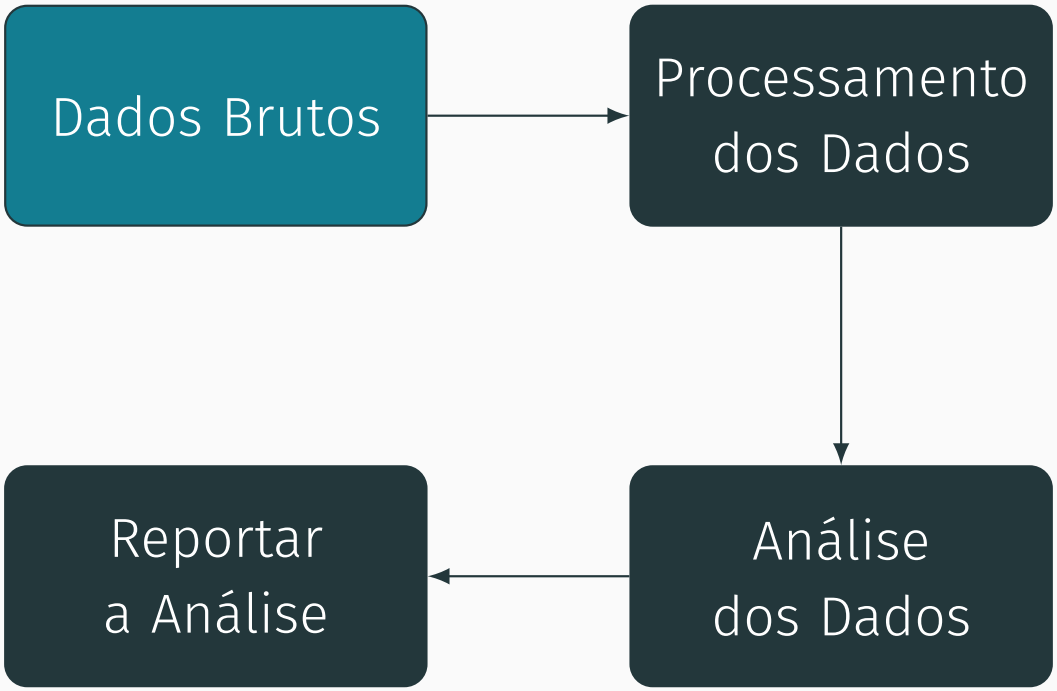
ANÁLISE DE DADOS



Dados Brutos

COMO OBTER DADOS PARA ANÁLISE

- Nem sempre é fácil conseguir dados para análise
- Como é possível extrair diversas informações a partir de análises, muitos vezes o acesso a dados é cobrado
- Veremos aqui como encontrar e baixar dados gratuitos da internet


Dados Brutos
Processamento dos Dados
Análise dos Dados
Reportar a Análise

DADOS BRUTOS

- Dados brutos são dados sem tratamento algum
- São os dados originais, vindos diretamente da fonte
- Através dos dados brutos conseguimos as tabelas que serão utilizadas em nossas análises

DADOS BRUTOS

- Há diversas maneiras de dados brutos serem obtidos a partir da internet
- As maneiras principais são
    1. Download via sites de divulgação
    2. Web scraping
    3. Solicitação direta

DOWNLOAD DE DADOS BRUTOS

- Instituto Brasileiro de Geografia e Estatística (IBGE) - `http://downloads.ibge.gov.br/`
- Banco Central do Brasil - `http://www4.bcb.gov.br/pec/series/port/aviso.asp`
- Portal Brasileiro de Dados Abertos - `http://dados.gov.br`
- Portal da Transparência - `http://www.transparencia.gov.br/`
- Dados Abertos da Câmara dos Deputados - `http://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/dados-abertos-cota-parlamentar`

DOWNLOAD DE DADOS BRUTOS

- Gene Expression Omnibus (GEO) - `http://www.ncbi.nlm.nih.gov/geo/`
- UCI Machine Learning Repository - `http://archive.ics.uci.edu/ml/`
- The home of the U.S. Government's open data - `http://www.data.gov/`
- Analyze Survey Data for Free - `http://www.asdfree.com/`

DOWNLOAD DE DADOS BRUTOS

- IMDb - `http://www.imdb.com/interfaces`
- The Last.fm Dataset | Million Song Dataset - `http://labrosa.ee.columbia.edu/millionsong/lastfm`
- Economic Time Series Page - `http://www.economagic.com/`
- Time Series Data Library - `https://datamarket.com/data/list/?q=provider:tsdl`

DOWNLOAD DE DADOS BRUTOS

- Eu mantenho uma lista de sites favoritos online, onde é possível verificar quais sites de dados já me interessaram em algum momento
- `https://pinboard.in/u:grandeabobora/t:datasets/`
- `https://pinboard.in/u:grandeabobora/t:database/`

WEB SCRAPING

- Termo que significa vasculhar a internet através de dados
- A ideia é coletar e organizar automaticamente os dados que estão espalhados em um ou mais sites
- Logicamente, apenas dados abertos ao público podem ser coletados desta maneira

- O R possui pacotes que realizam este tipo de trabalho
- Existem desde pacotes bastante específicos, como o `twitteR`, até pacotes com usos mais gerais, como o `rvest`
- Infelizmente, o desempenho do R não é o melhor para este tipo de situação

- A opção mais utilizada para web scraping é o `python`
- Há muito tutoriais e material de consulta disponível, embora a maioria seja em inglês
- O endereço para baixar o interpretador da linguagem é `https://www.python.org/`

- Assim como o R possui pacotes, o `python` possui módulos que aumentam as funcionalidades da linguagem
- O módulo de `python` responsável por web scraping é chamado `scrapy`
- Maiores informações podem ser obtidas em `http://scrapy.org/`

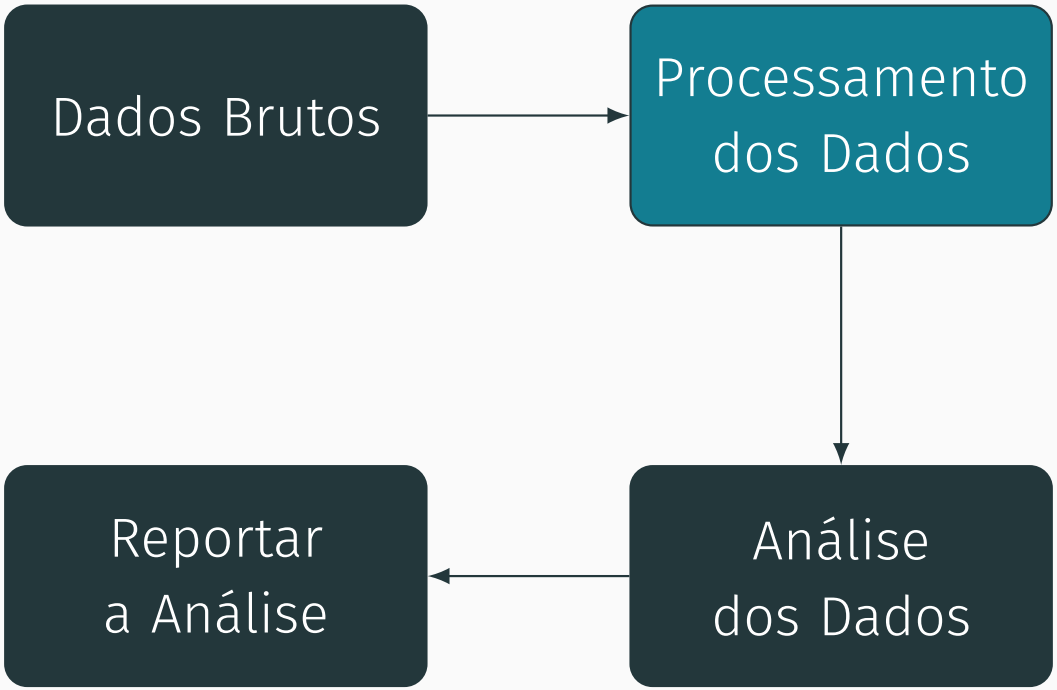
- O link `http://www.python-forum.org/viewtopic.php?f=25&t=1626` possui um tutorial que explica como utilizar o `scrapy`

WEB SCRAPING

- É possível utilizar ferramentas mais amigáveis para extrair dados da internet
- Por exemplo, o site import.io - `http://import.io/`

Processamento dos Dados

PROCESSAMENTO DOS DADOS

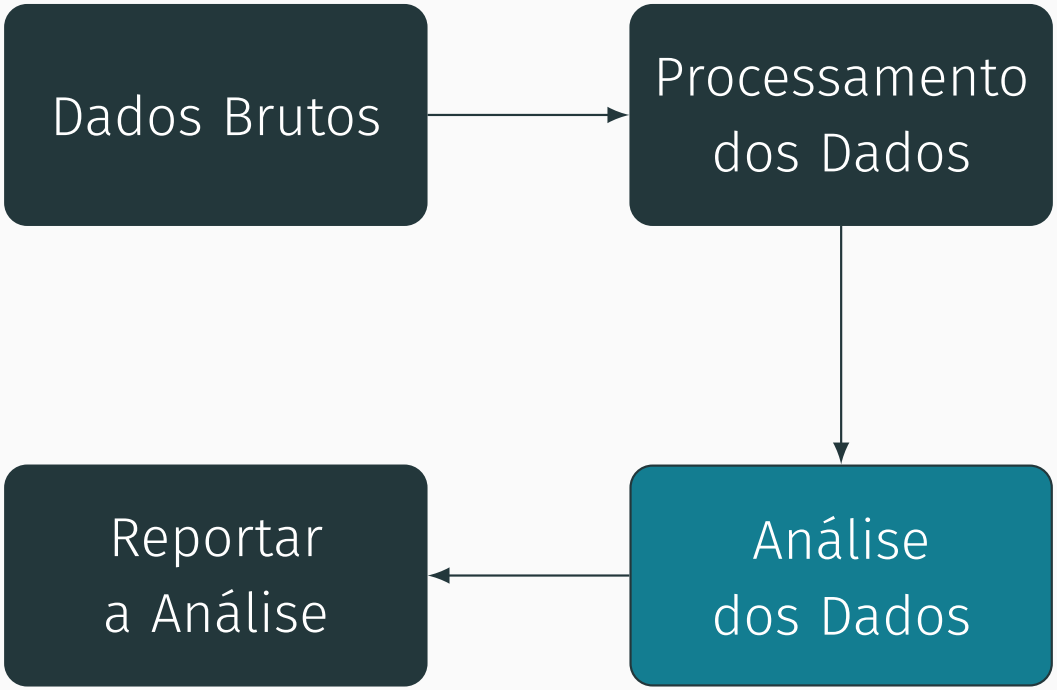


PROCESSAMENTO DOS DADOS

- É a etapa de preparação dos dados brutos para análise
- Existem diversas maneiras de fazer isto, seja utilizando o R ou outras ferramentas, como a linha de comando de sistemas *nix (Linux, Unix, Mac OS)
- Vamos ver como realizar a limpeza de dados em um conjunto fornecido pelo IMDb

Análise dos Dados

ANÁLISE DOS DADOS

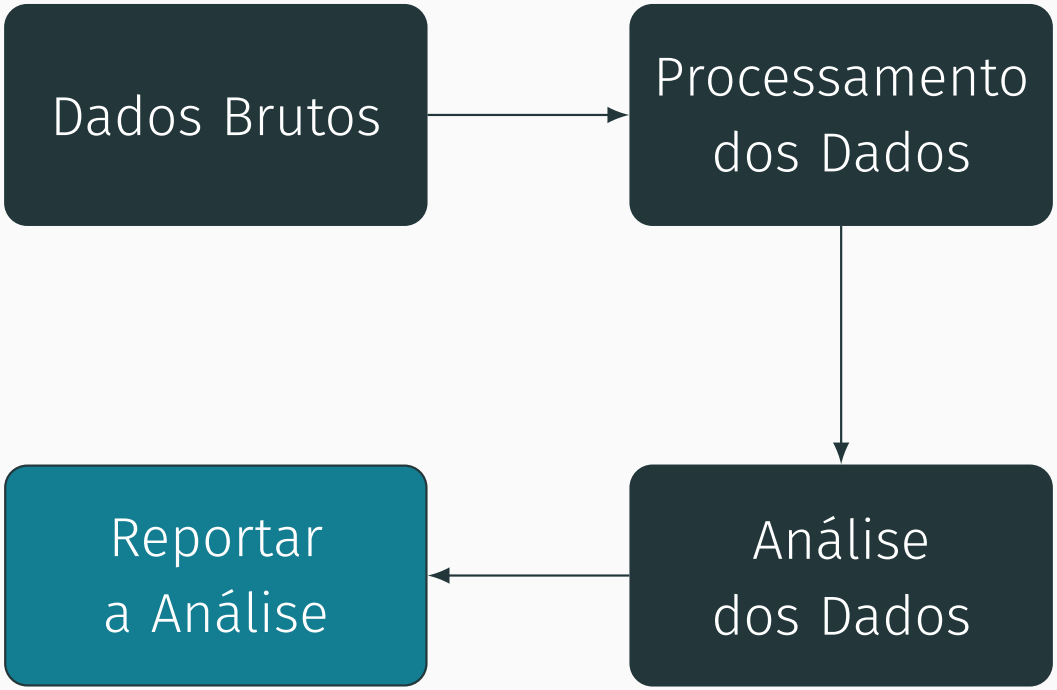


ANÁLISE DOS DADOS

- Fica imensamente mais fácil proceder com a análise dos dados após eles terem sido previamente processados
- Perceba que o banco de dados criado com os dados do IMDb é muito mais fácil de analisar do que o conjunto original
- Ele está pronto para ser trabalhado por um programa como R, `python` ou até mesmo `Excel`

Reportar a Análise

REPORTAR A ANÁLISE



REPORTAR A ANÁLISE

- As conclusões tiradas na análise são reportadas neste passo
- Tente mantê-las interessantes, mas simples; concisas, mas completas; rigorosas, mas leves
- Uma tendência atual é usar dados para contar uma história; tente se valer de uma abordagem assim para manter seu interlocutor interessado

AGRUPAMENTO DE DADOS

K-MEANS

- É um método popular de classificação de dados
- O objetivo é separar N observações em K grupos
- Neste método, cada observação é designada para o grupo com a média mais próxima

- Este método é sensível a outliers
- Os pontos podem se mover de um grupo para outro, mas a resposta final depende da inicialização dos centros
- Se uma observação estiver igualmente perto de dois ou mais centros, então o grupo deve ser decidido aleatoriamente

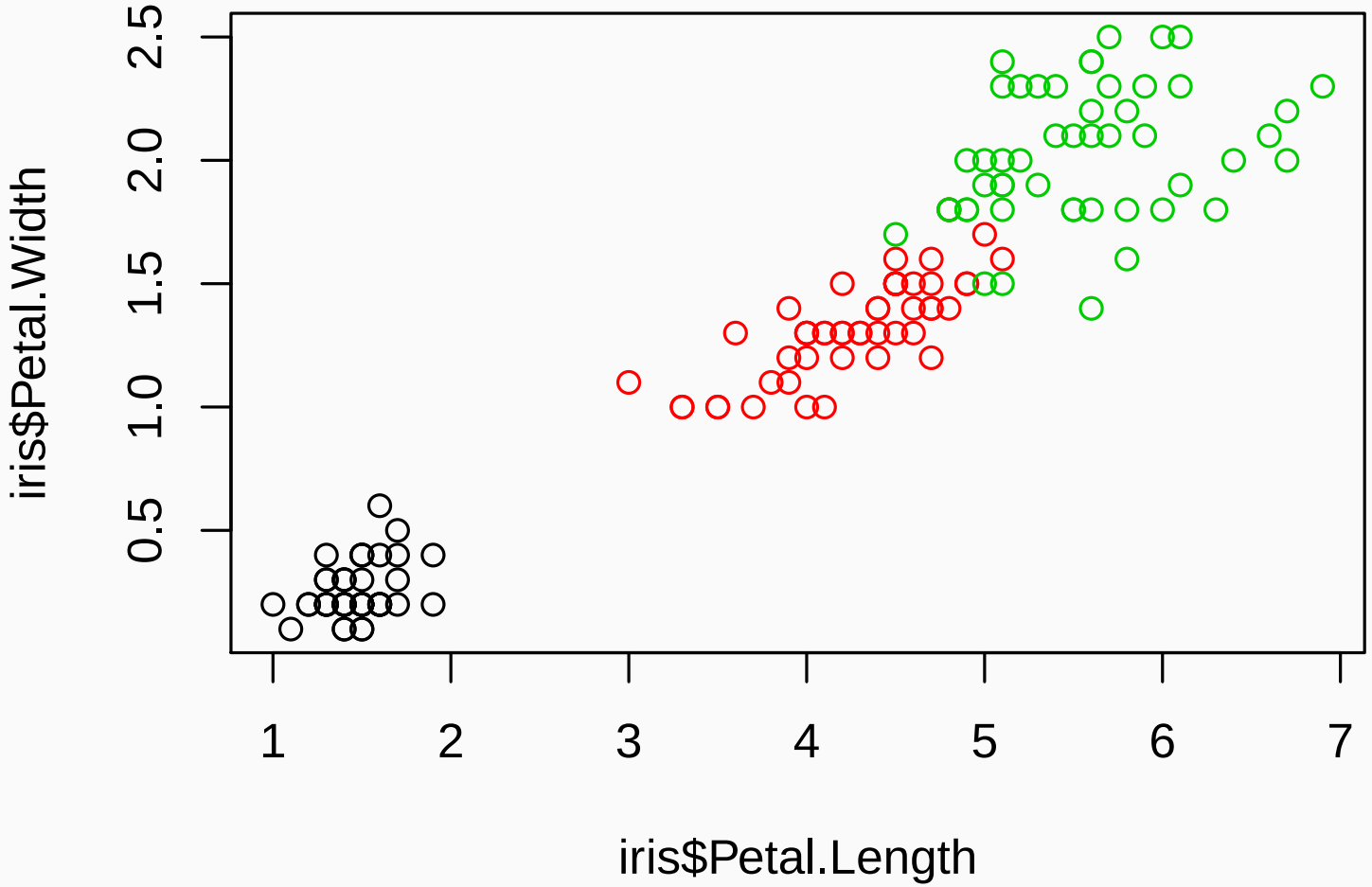
- Pode ser muito efetivo como uma método de previsão *black box*
- Não é útil para entender a natureza da relação entre as características e as classes

K-MEANS

- Vamos utilizar o conjunto de dados `iris`
- Precederemos com a Análise de Componentes Principais e o K-Means

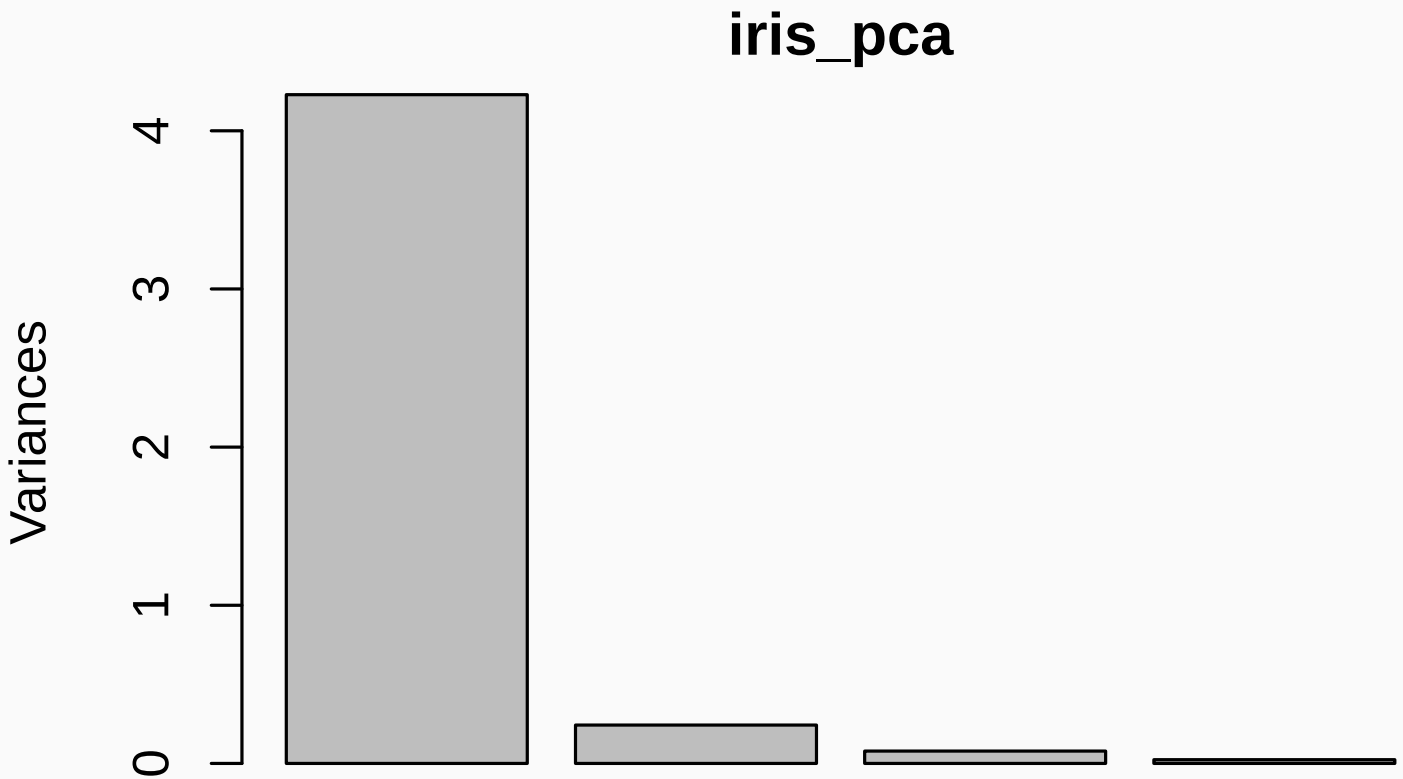
```
> head(iris)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
> plot(iris$Petal.Length, iris$Petal.Width,
+ col=as.numeric(iris$Species))
```



```
> iris2 <- iris[, -5]
> iris_pca <- prcomp(iris2)
> summary(iris_pca)

## Importance of components:
##                           PC1     PC2    PC3     PC4
## Standard deviation     2.0563 0.49262 0.2797 0.15439
## Proportion of Variance 0.9246 0.05307 0.0171 0.00521
## Cumulative Proportion  0.9246 0.97769 0.9948 1.00000
```

```
> plot(iris_pca)
```

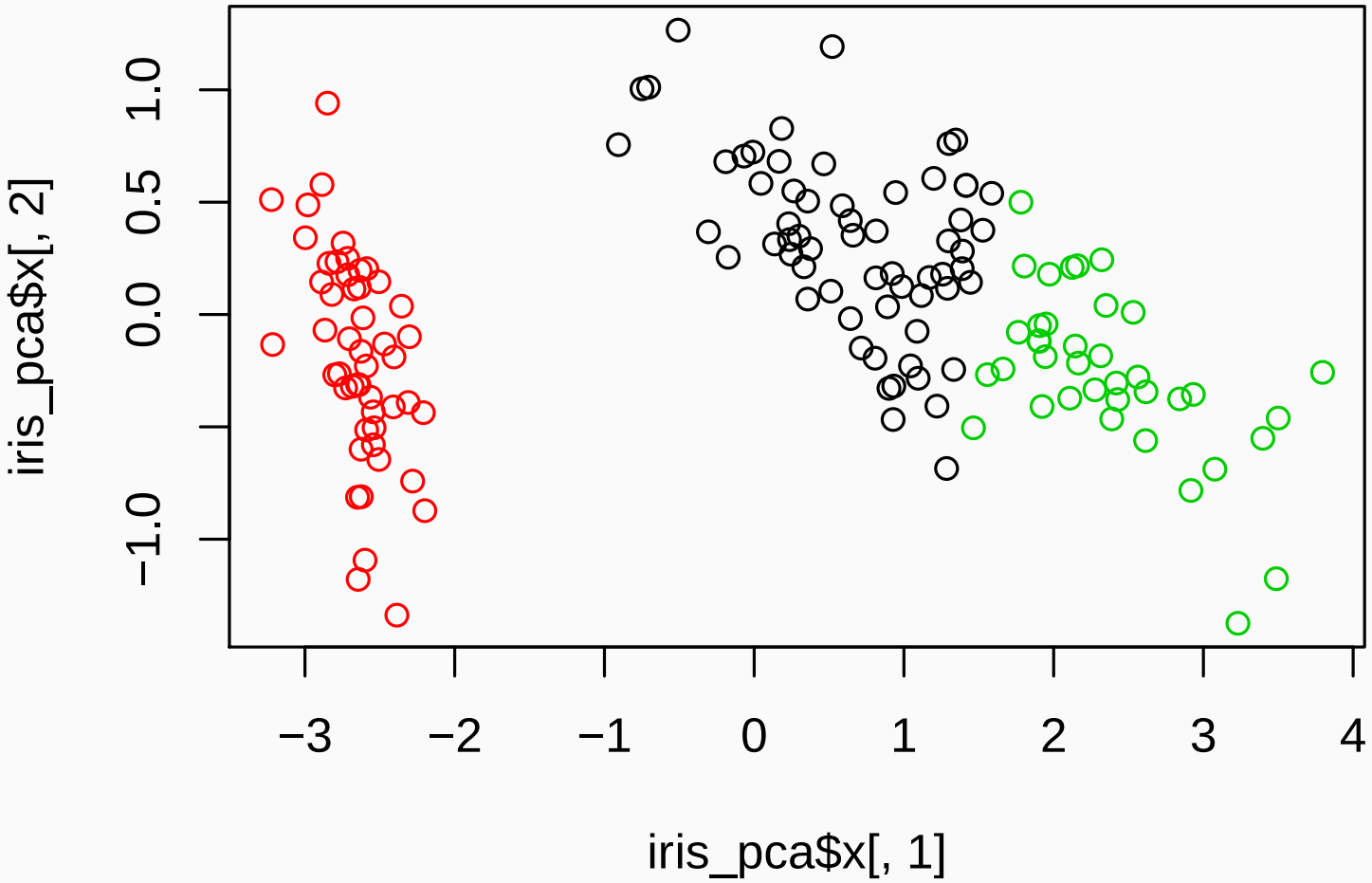


K-MEANS

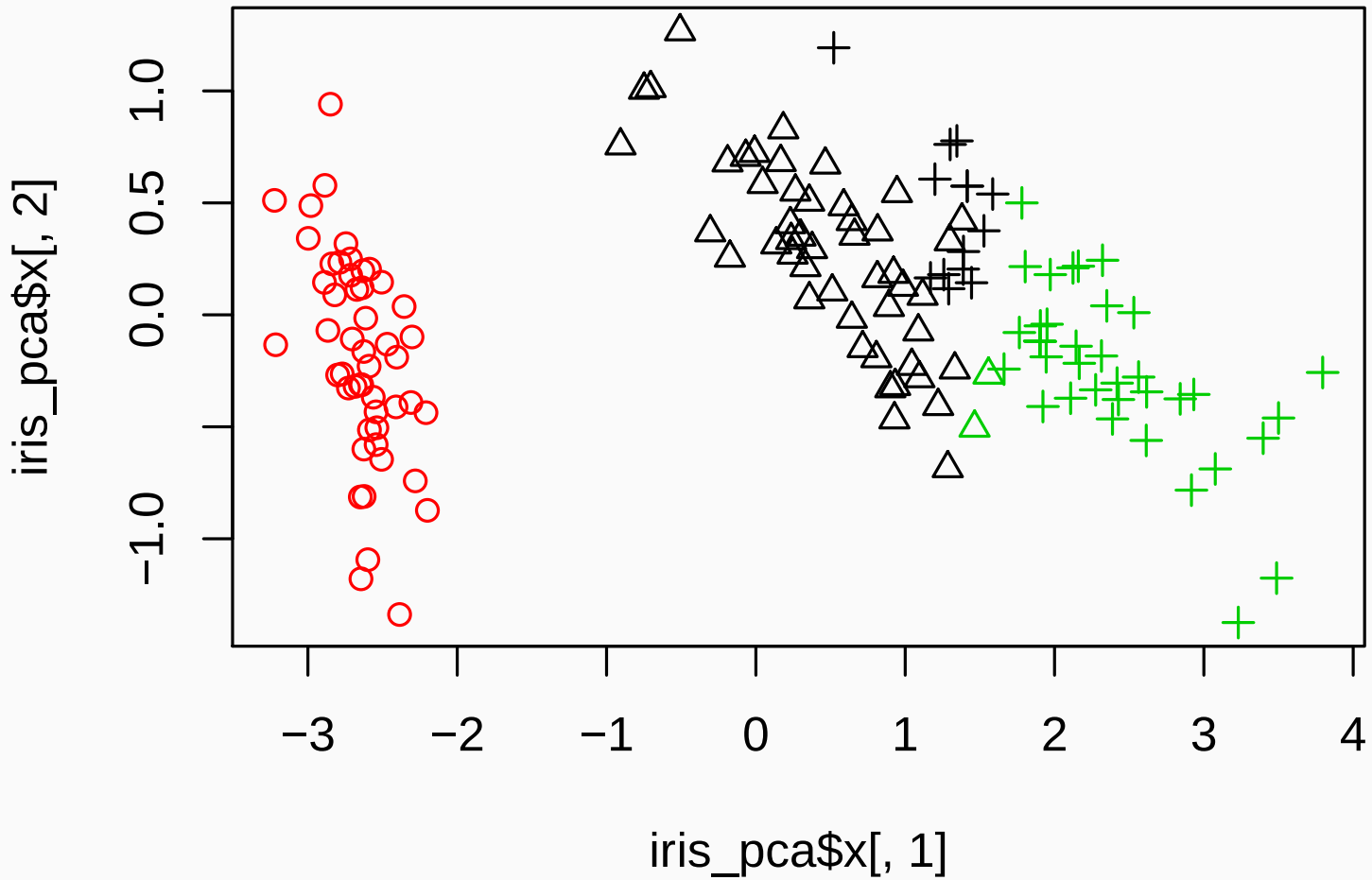
```
> iris_kmeans <- kmeans(iris2, centers=3)
> names(iris_kmeans)

## [1] "cluster"      "centers"      "totss"
## [4] "withinss"     "tot.withinss" "betweenss"
## [7] "size"         "iter"         "ifault"

> iris_kmeans$cluster

##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [28] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 1
##  [55] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1
##  [82] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 3 3 1 3
## [109] 3 3 3 3 3 1 1 3 3 3 3 1 3 1 3 1 3 3 1 1 3 3 3 3 3 1 3
## [136] 3 3 3 1 3 3 3 1 3 3 3 1 3 3 1

> iris_kmeans$size

## [1] 62 50 38
```

```
> resultado <- table(iris$Species, iris_kmeans$cluster)
> resultado

##
##                1  2  3
##   setosa       0 50  0
##   versicolor 48  0  2
##   virginica  14  0 36

> apply(resultado, 2, sum)

##  1  2  3
## 62 50 38
```

```
> plot(iris_pca$x[, 1], iris_pca$x[, 2],
+ col=as.numeric(iris_kmeans$cluster))
```



```
> plot(iris_pca$x[, 1], iris_pca$x[, 2],
+ col=as.numeric(iris_kmeans$cluster), pch=as.numeric(iris$Species))
```



Aplicação

OBJETIVOS

- Trabalho proposto por Marcos Silva, da Embrapa Gado de Leite de Juiz de Fora, MG
- Identificar animais resistentes à infecção por nematóides
- Separar os animais em grupos para posterior acasalamento
- Sequenciar o genoma dos animais resistentes e não-resistentes para identificar os genes mais importantes

POPULAÇÃO

- Dados de contagem de larvas de nematóides em gado F2 Gir x Holandês
- Foram utilizados 4 touros Holandeses e 28 fêmeas Gir para gerar uma população F1
- A partir dela foram selecionados 5 touros e 68 fêmeas para gerar uma população F2

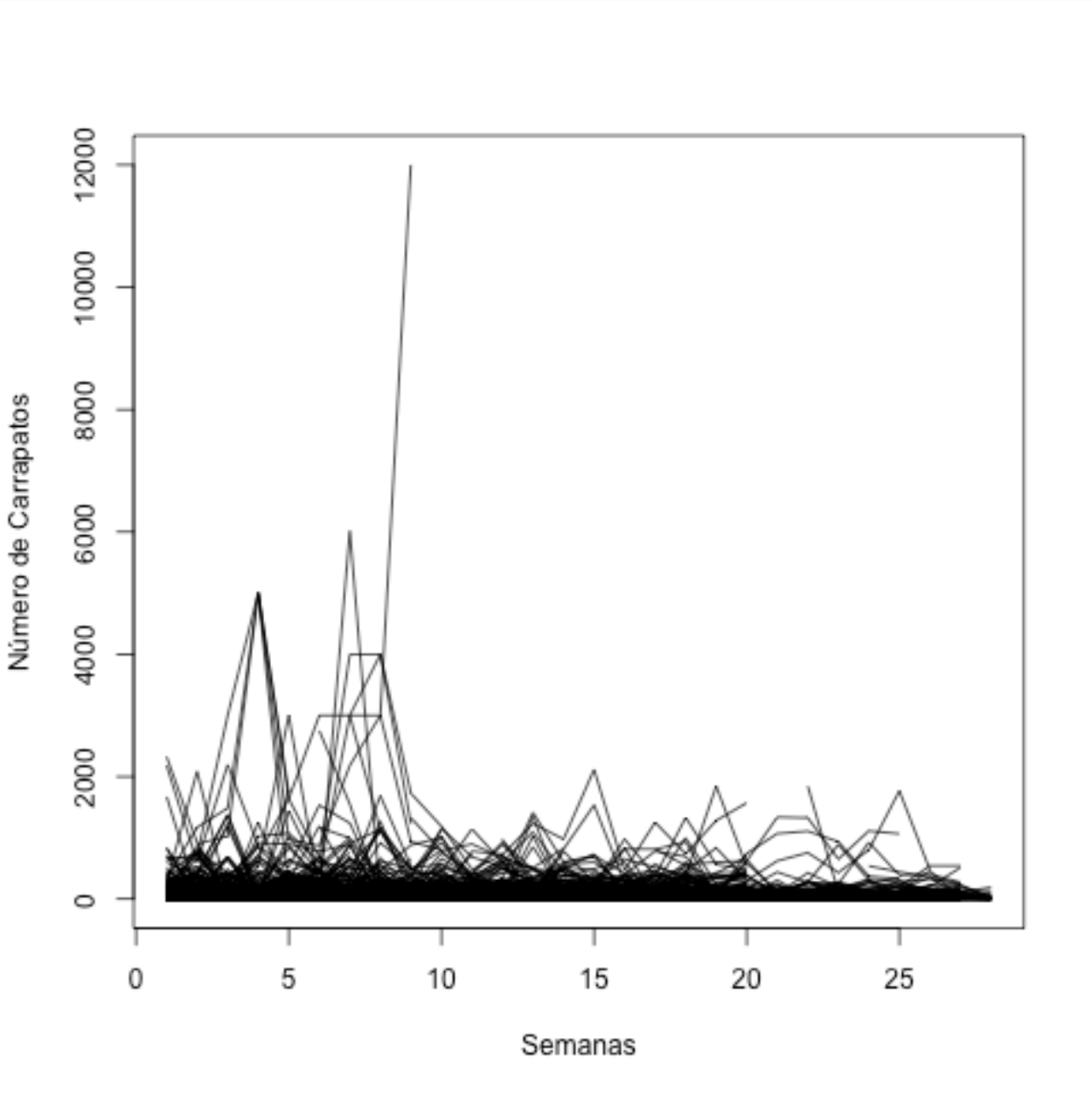
ANÁLISE PRÉVIA

- Contagem de ovos de nematóides em fezes
- Dados discretos
- Coleta durante 28 semanas
- Apresentam autocorrelação
- Dados longitudinais

ESTATÍSTICA DESCRITIVA

- 1552 perfis, totalizando 43.456 observações
- 700 perfis não possuem observação alguma
- 852 perfis para analisar

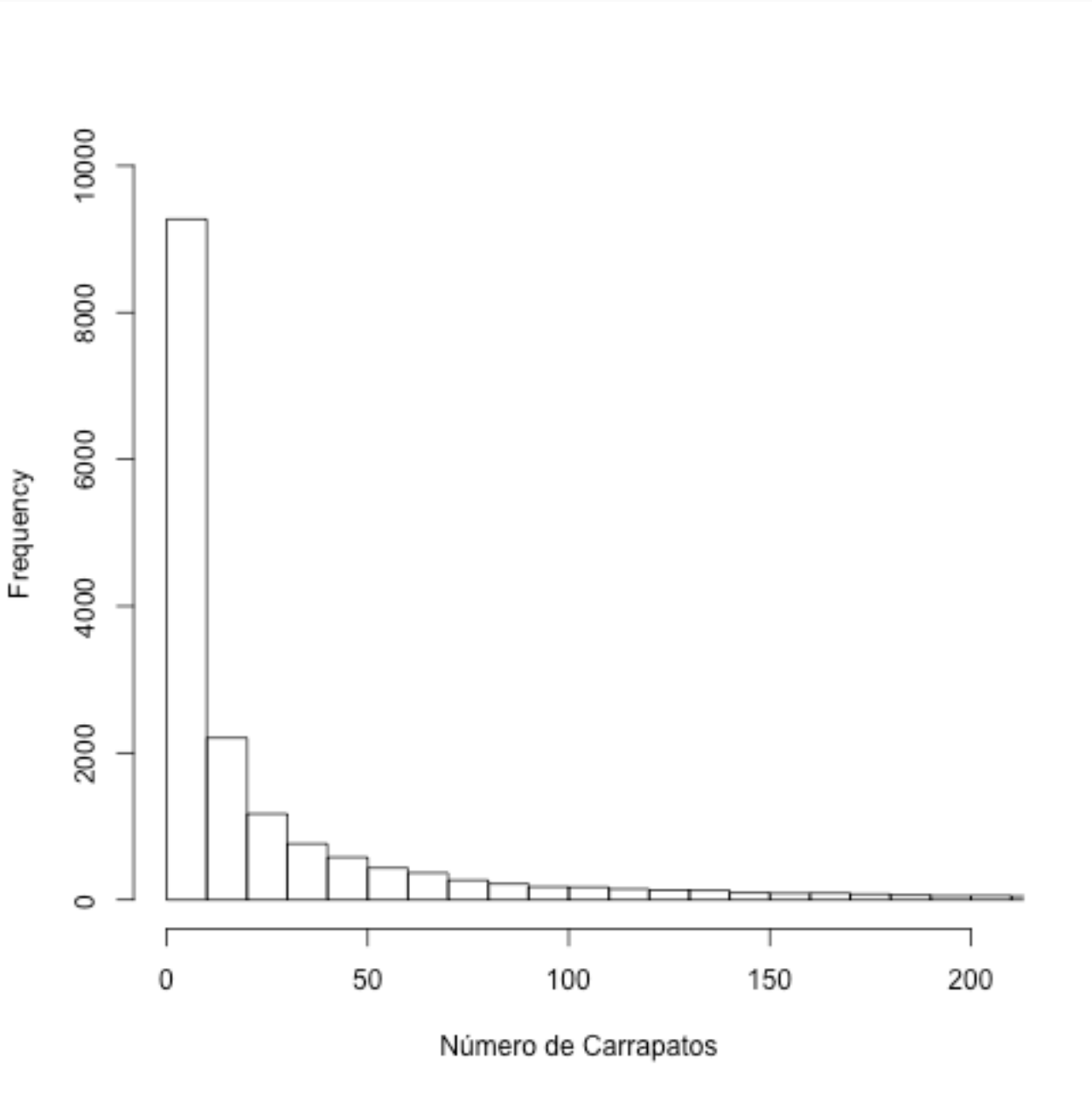
VISUALIZAÇÃO DOS DADOS



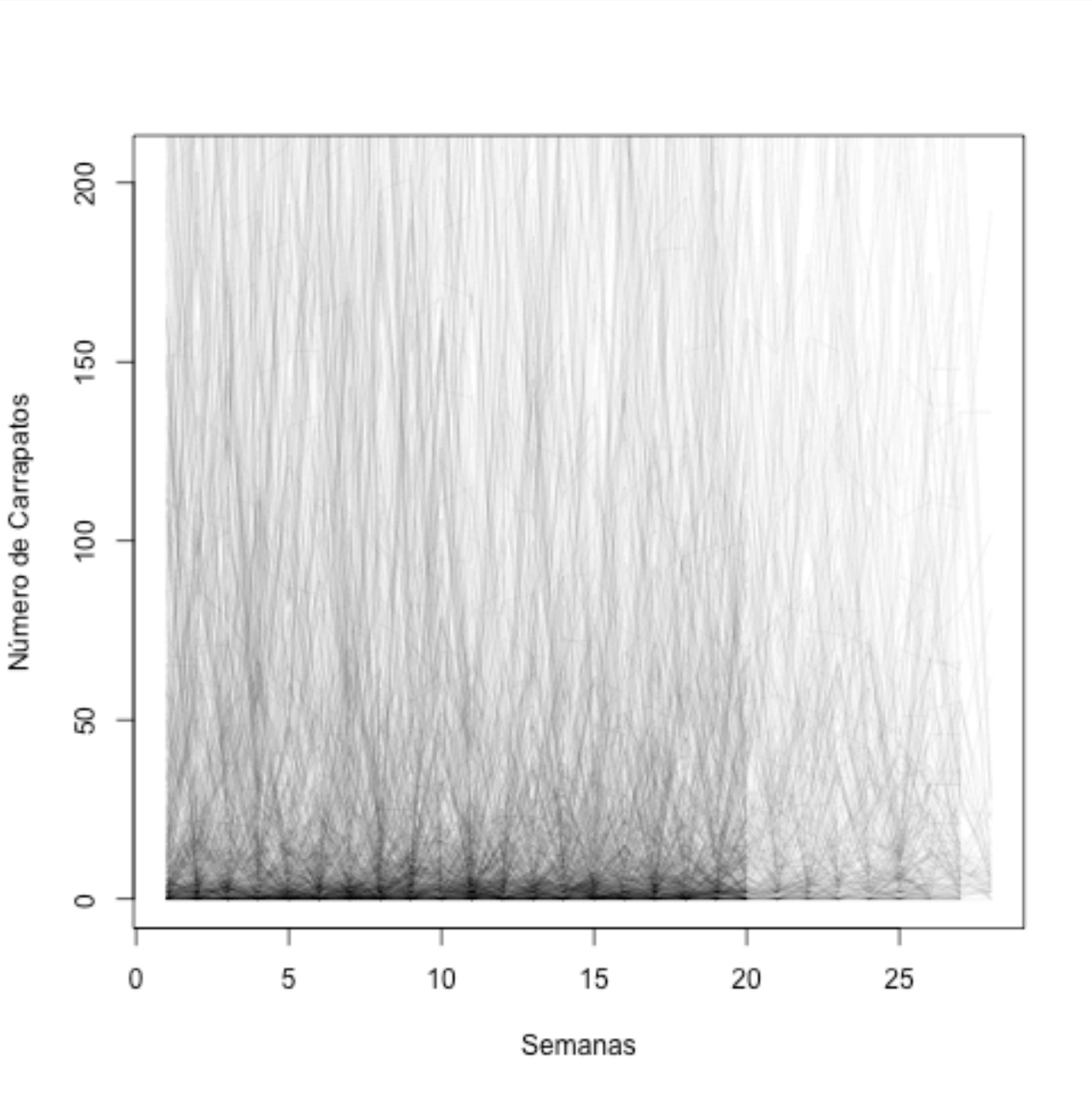
Min	Q_1	Med	$\bar{X}$	Q_3	Max	NA
0	2	9	50,43	35	12.000	6453

75%	90%	95%	99%
35	112	205	662

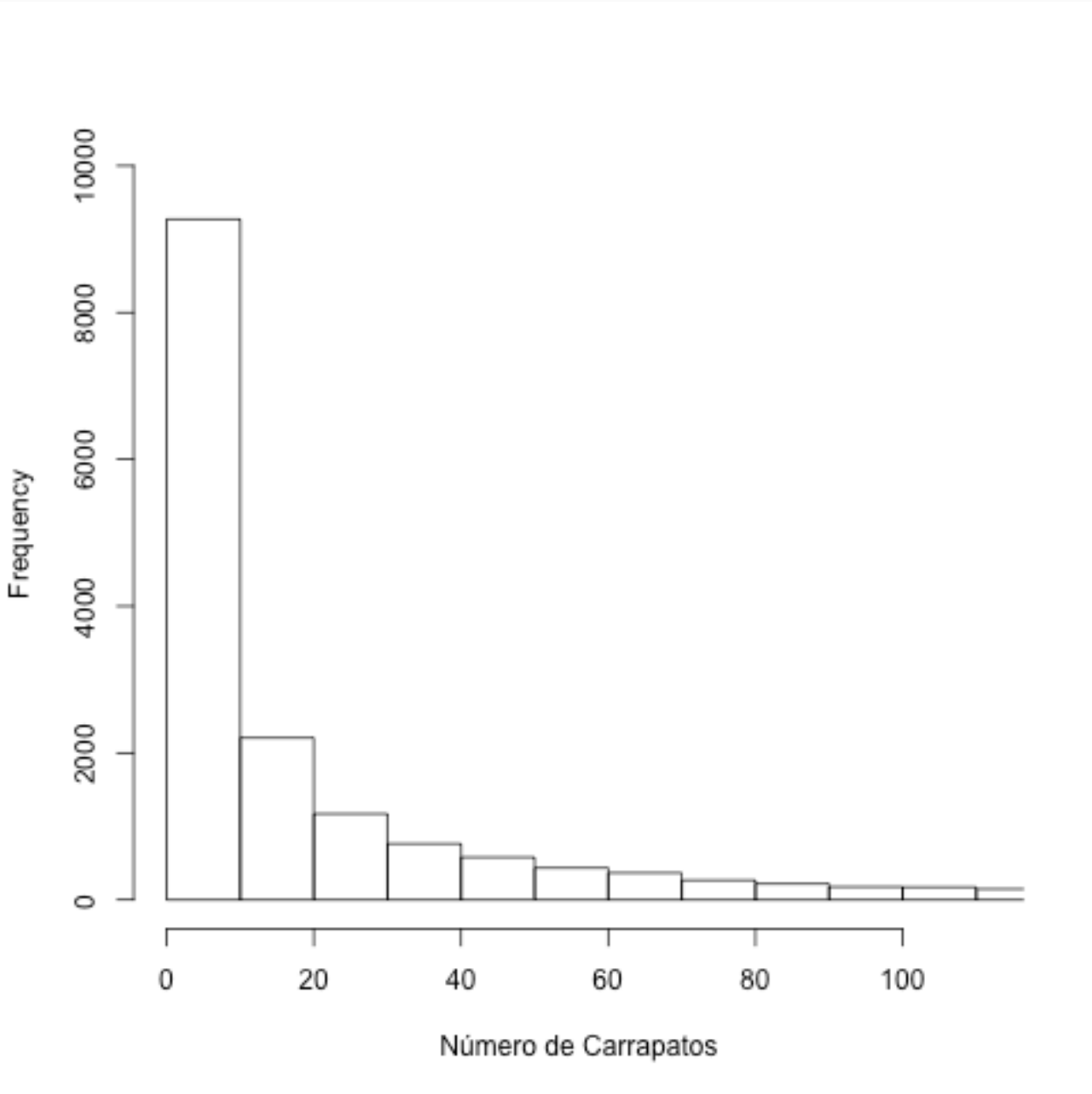
HISTOGRAMA DE 95% DOS DADOS



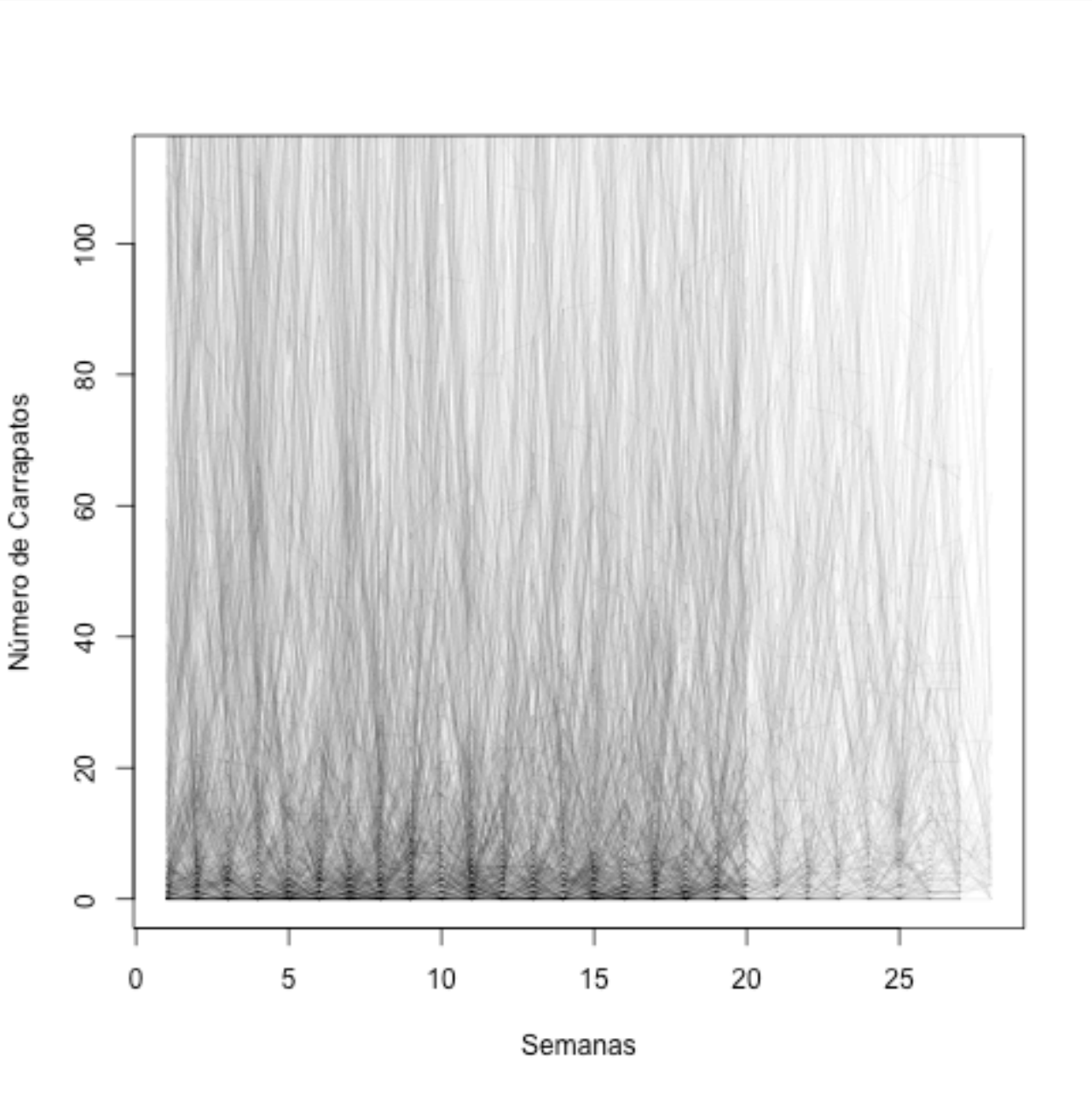
VISUALIZAÇÃO DOS PERFIS DE 95% DADOS



HISTOGRAMA DE 90% DOS DADOS



VISUALIZAÇÃO DOS PERFIS DE 90% DADOS



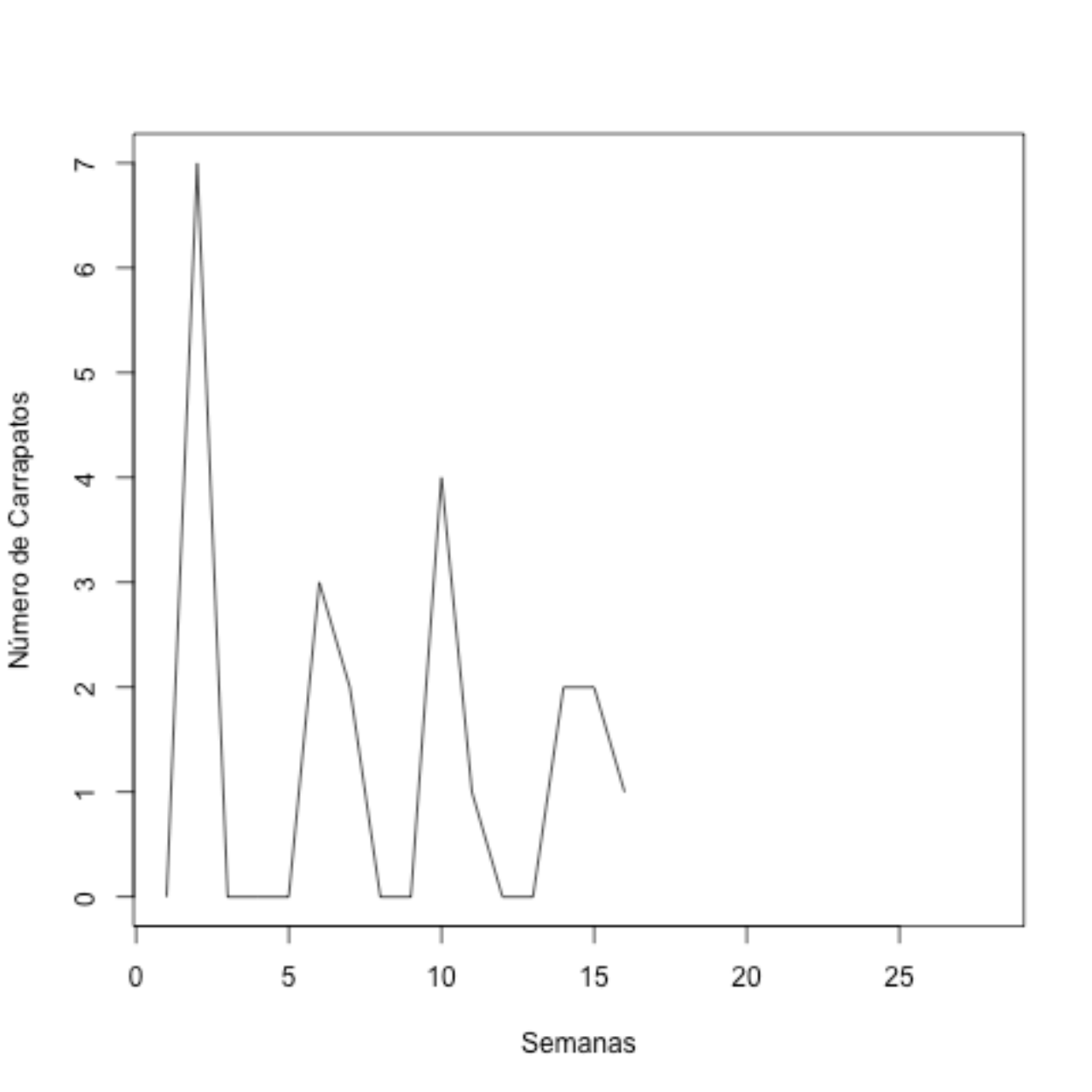
DADOS REAIS COMPLETOS

- 852 perfis
- Alguns animais foram medidos mais de uma vez, com resultados diferentes
- 480 animais diferentes
- Até 28 observações em cada perfil
- Dados faltantes

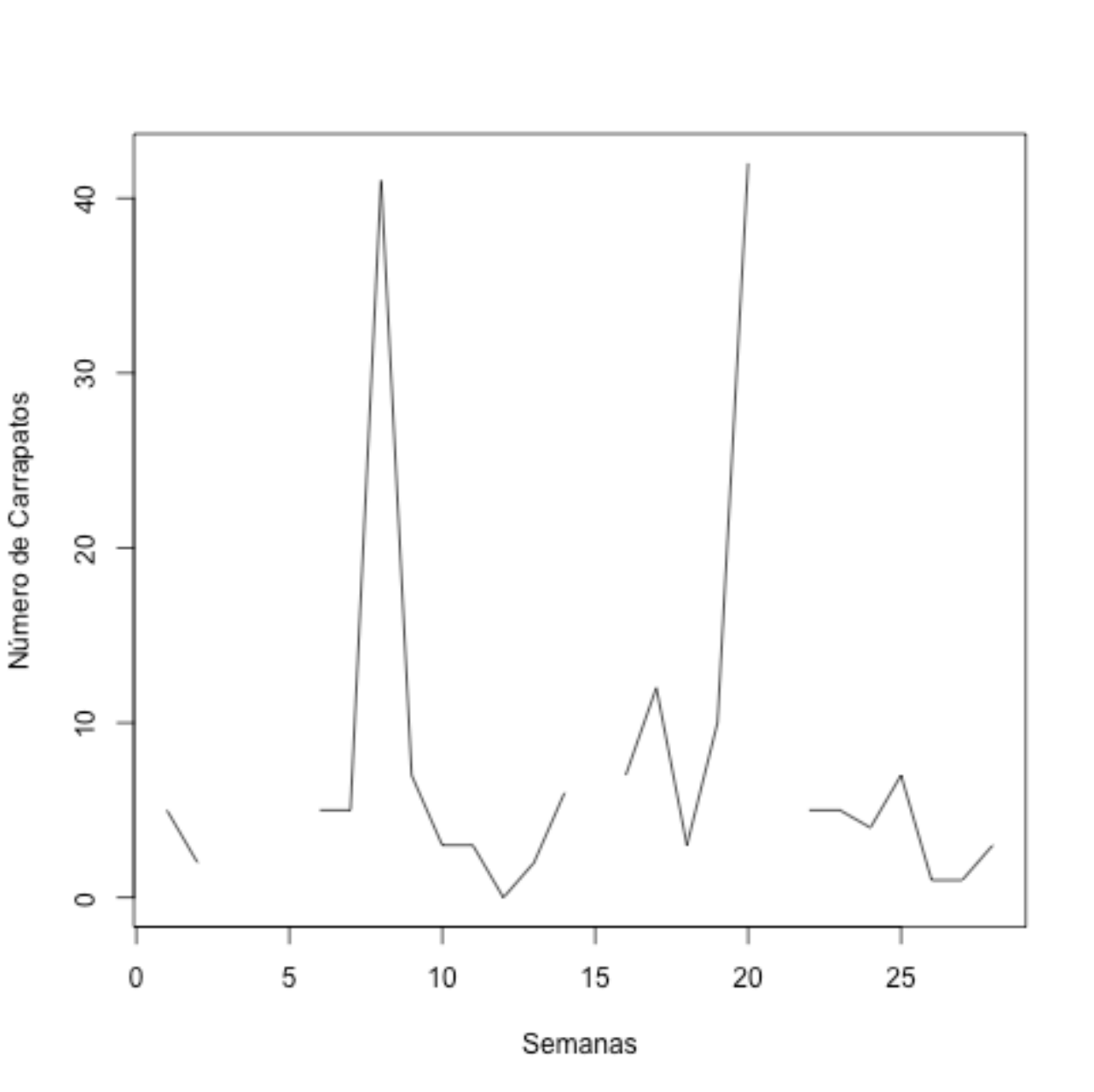
KML: K-MEANS PARA DADOS LONGITUDINAIS

- Criado especialmente para lidar com dados longitudinais
- Método não-supervisionado
- Bons resultados inclusive para curvas não-polinomiais

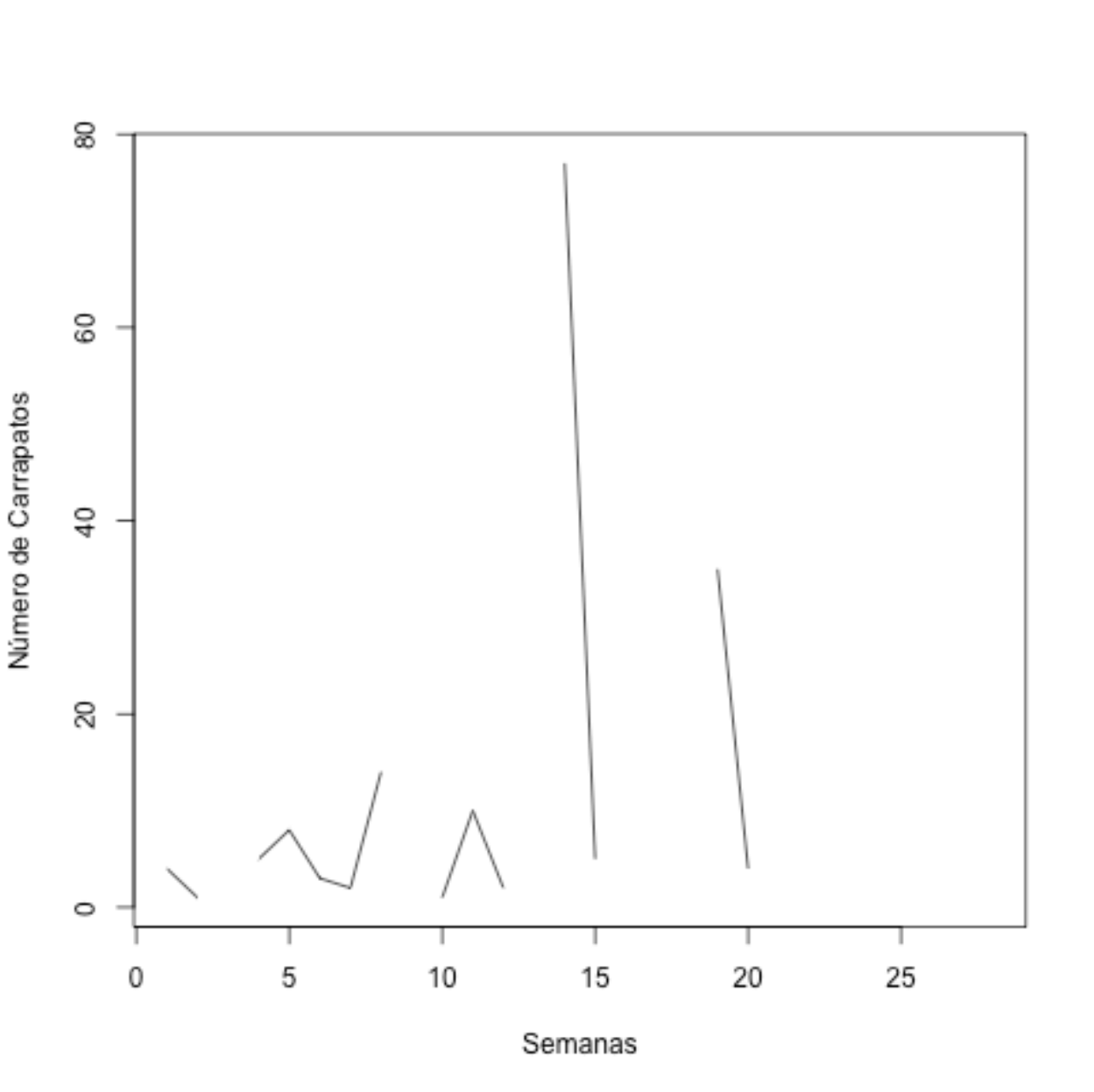
PERFIL DO ANIMAL 99948



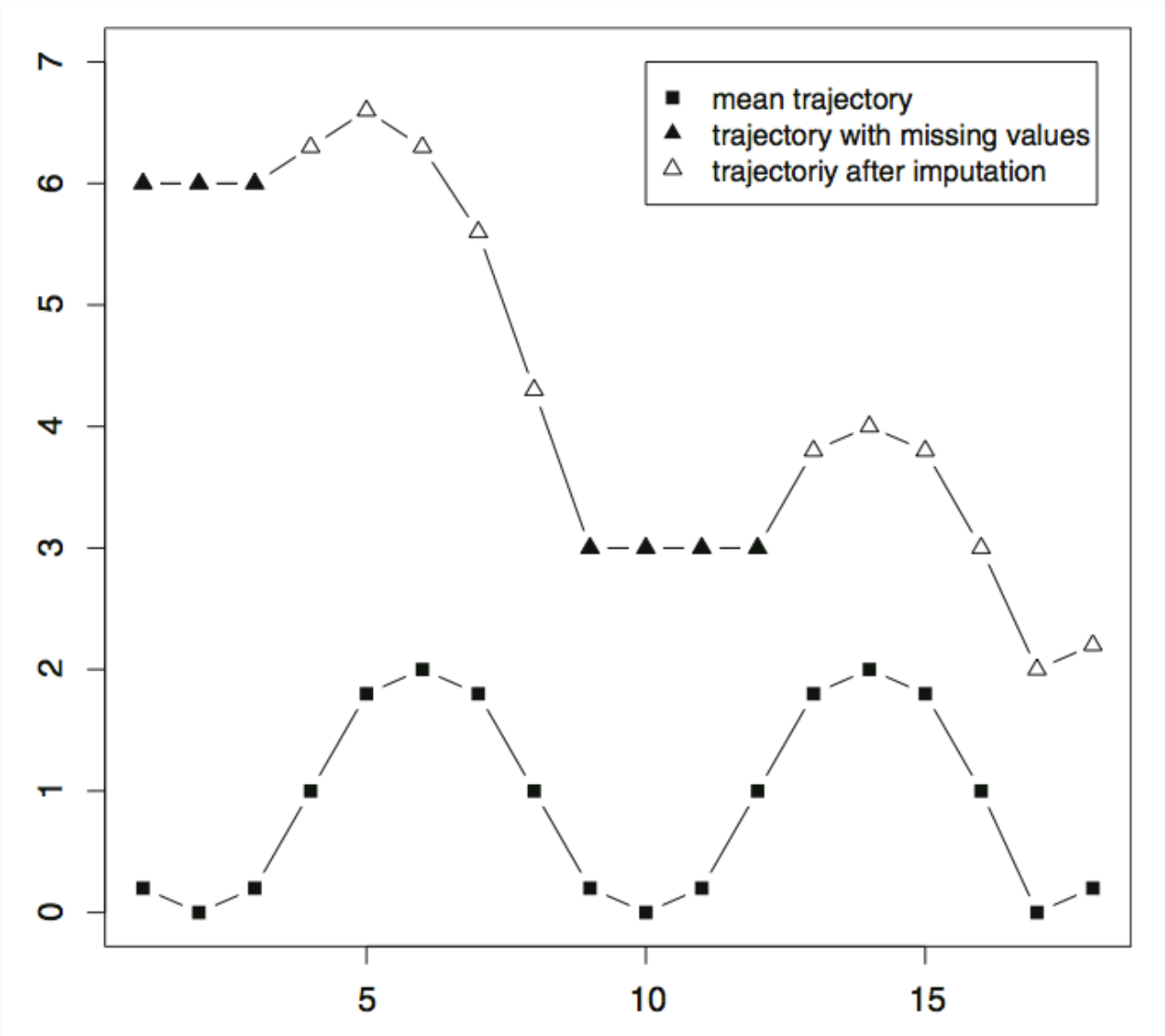
PERFIL DO ANIMAL 3662



PERFIL DO ANIMAL 2756



DADOS FALTANTES

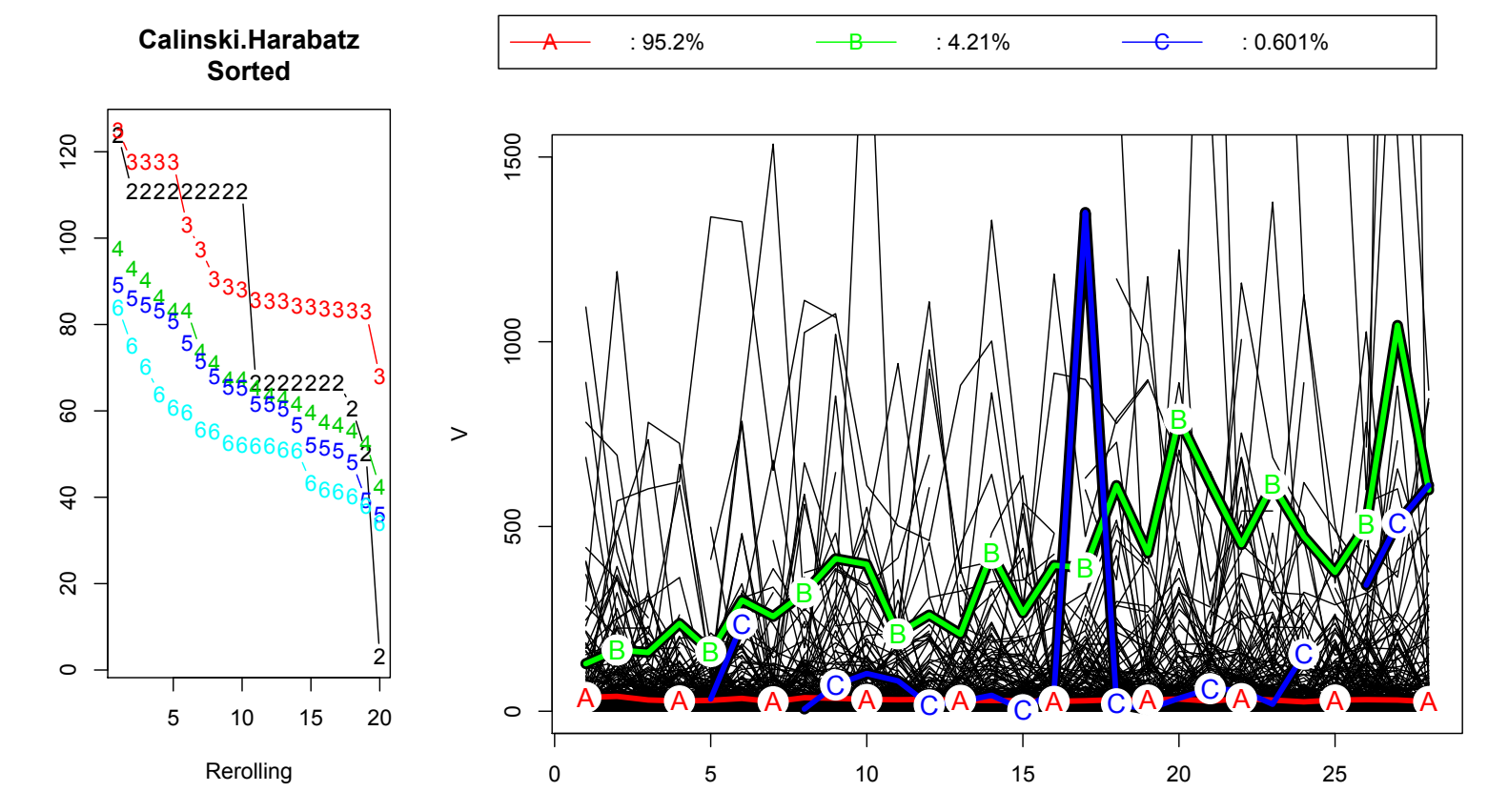


Fonte: Genolini e Falissard (2009)

KML: K-MEANS PARA DADOS LONGITUDINAIS

- $B = \sum_{m=1}^{g} n_m (\overline{Y}_m - \overline{Y})(\overline{Y}_m - \overline{Y})'$
- $W = \sum_{m=1}^{g} \sum_{k=1}^{n_m} (Y_{mk} - \overline{Y})(Y_{mk} - \overline{Y})'$
- n_m é o número de trajetórias no cluster m
- $\overline{Y}_m$ é a trajetória média do cluster m
- $\overline{Y}$ é a trajetória média do conjunto inteiro de dados
- Maximizar $C(g) = \dfrac{\mathrm{tr}(B)}{\mathrm{tr}(W)} \dfrac{n_m - g}{g - 1}$

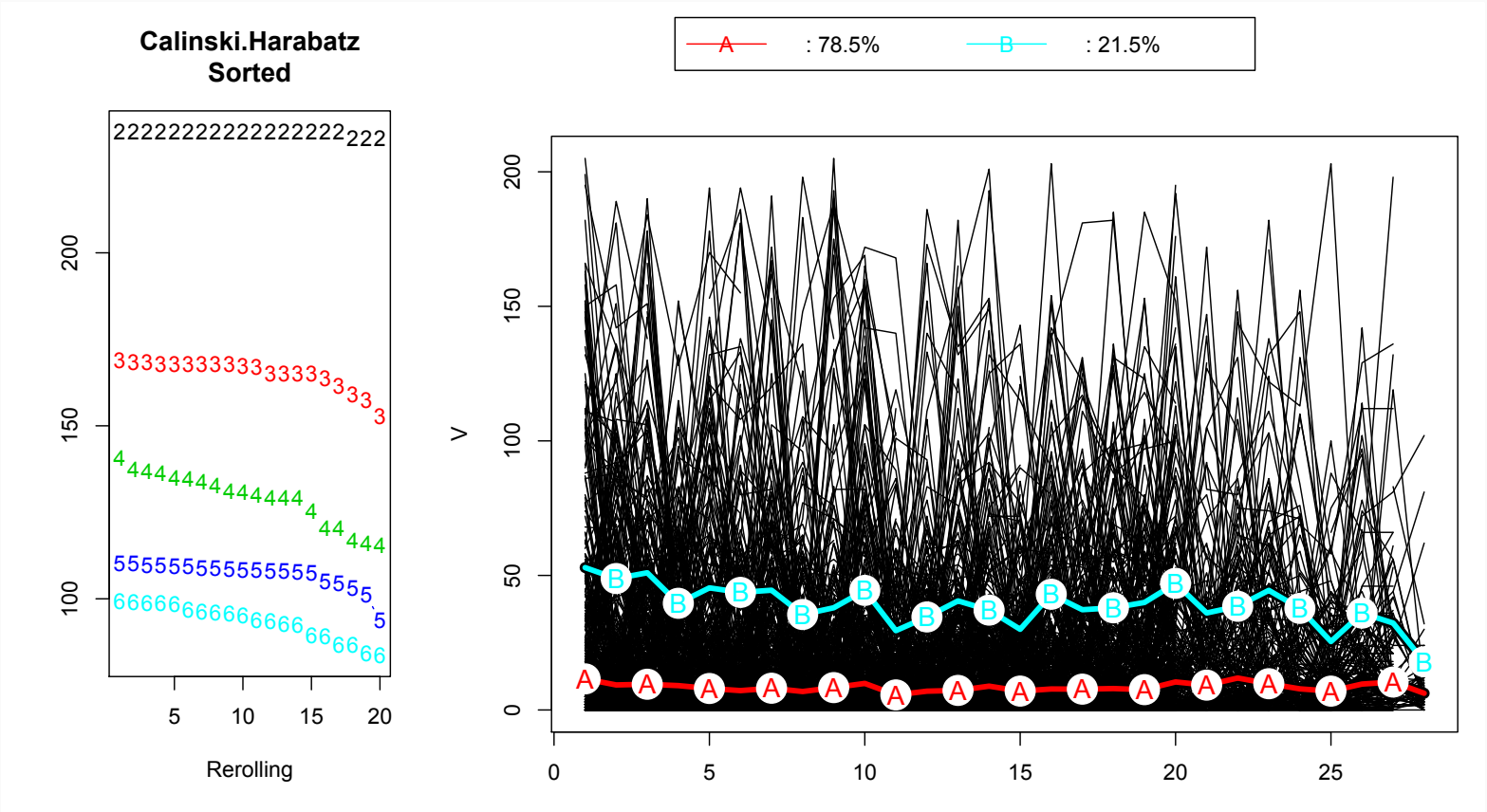
DADOS REAIS COMPLETOS



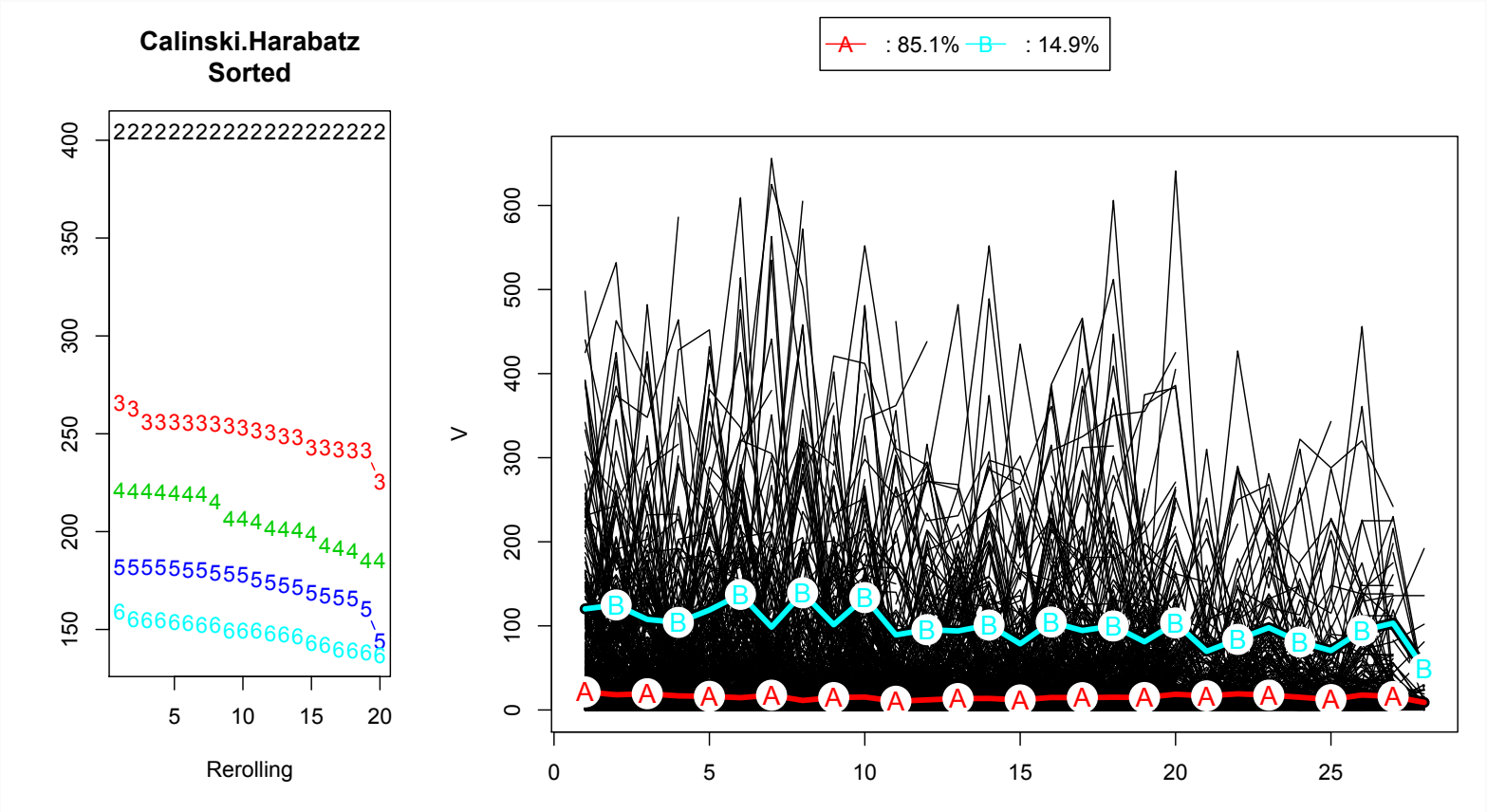
DADOS REAIS

- Processamento prévio
- Manter apenas os animais com todas as observações inferiores ao 95º ou 99º percentil ou manter somente as observações inferiores a 1000

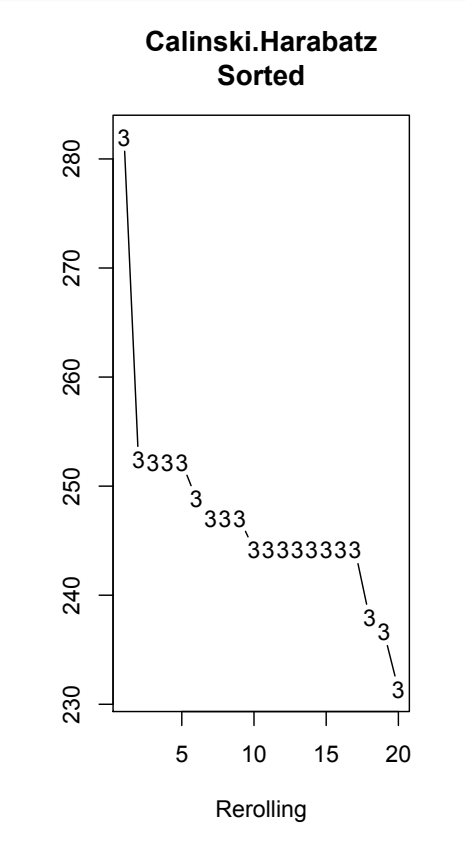
DADOS REAIS INFERIORES AO 95º PERCENTIL

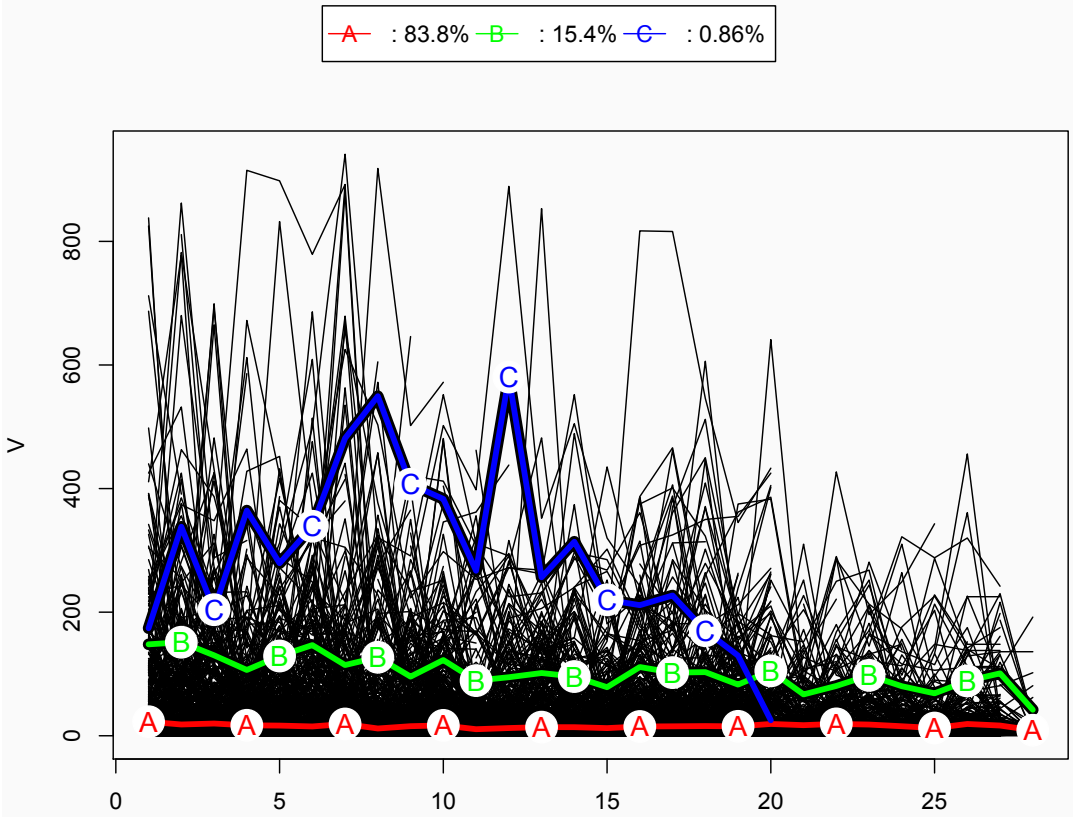


DADOS REAIS INFERIORES AO 99º PERCENTIL



DADOS REAIS INFERIORES A 1000

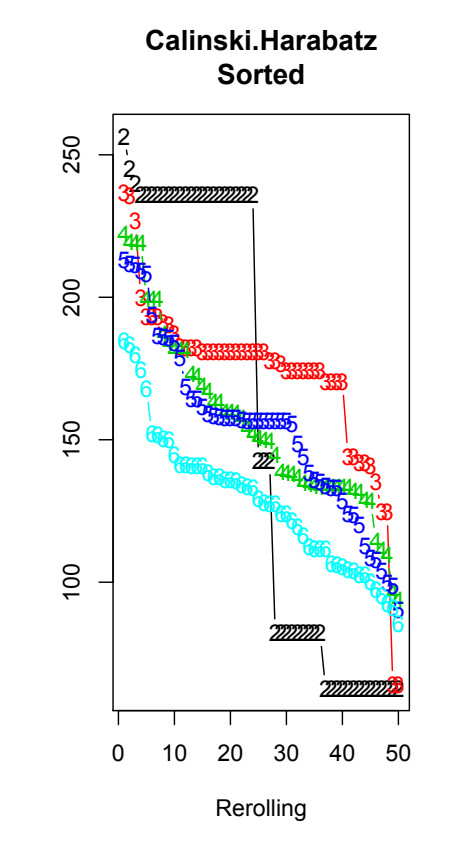


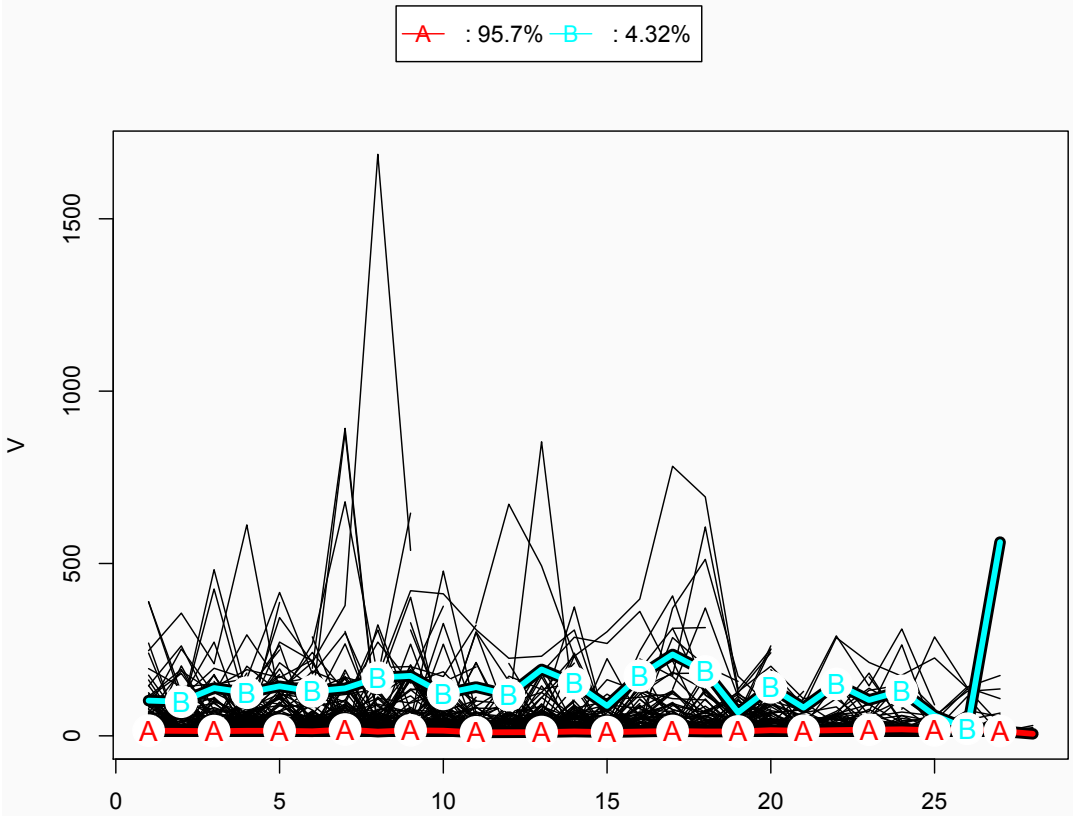


Métodos de coleta dos dados

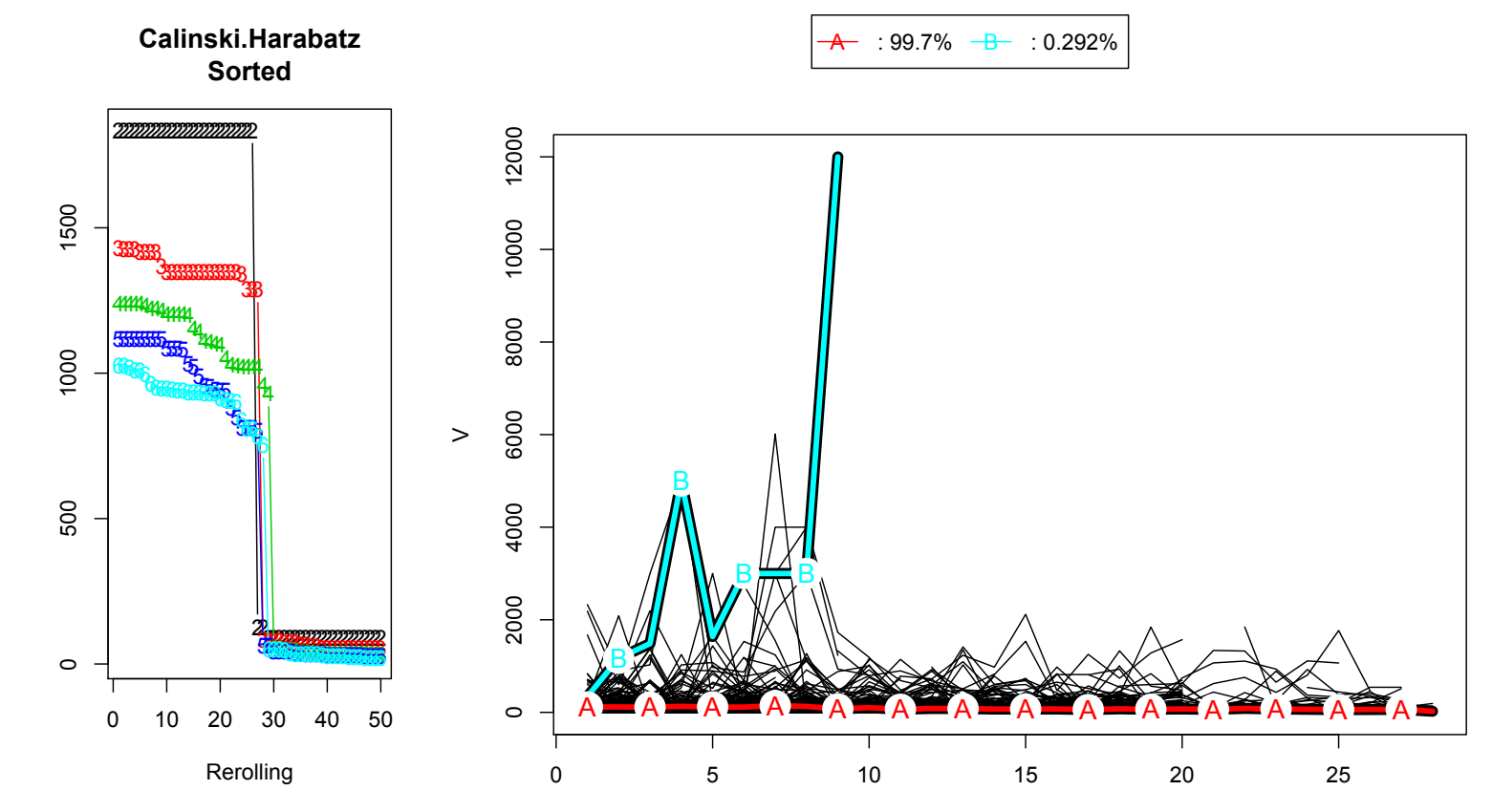
- Bastão (menos preciso)
- Lamínula (mais preciso)

DADOS REAIS - BASTÃO

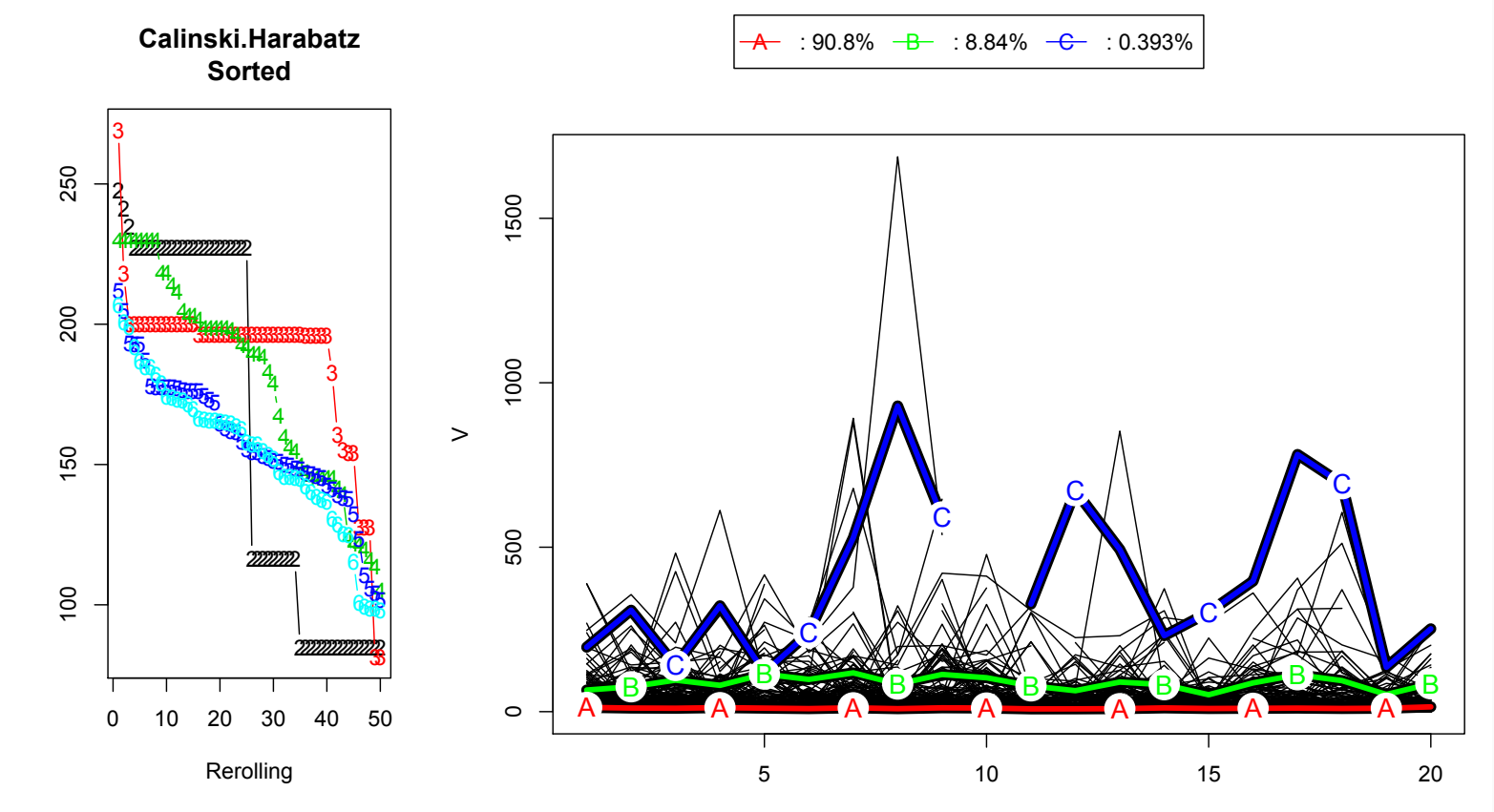




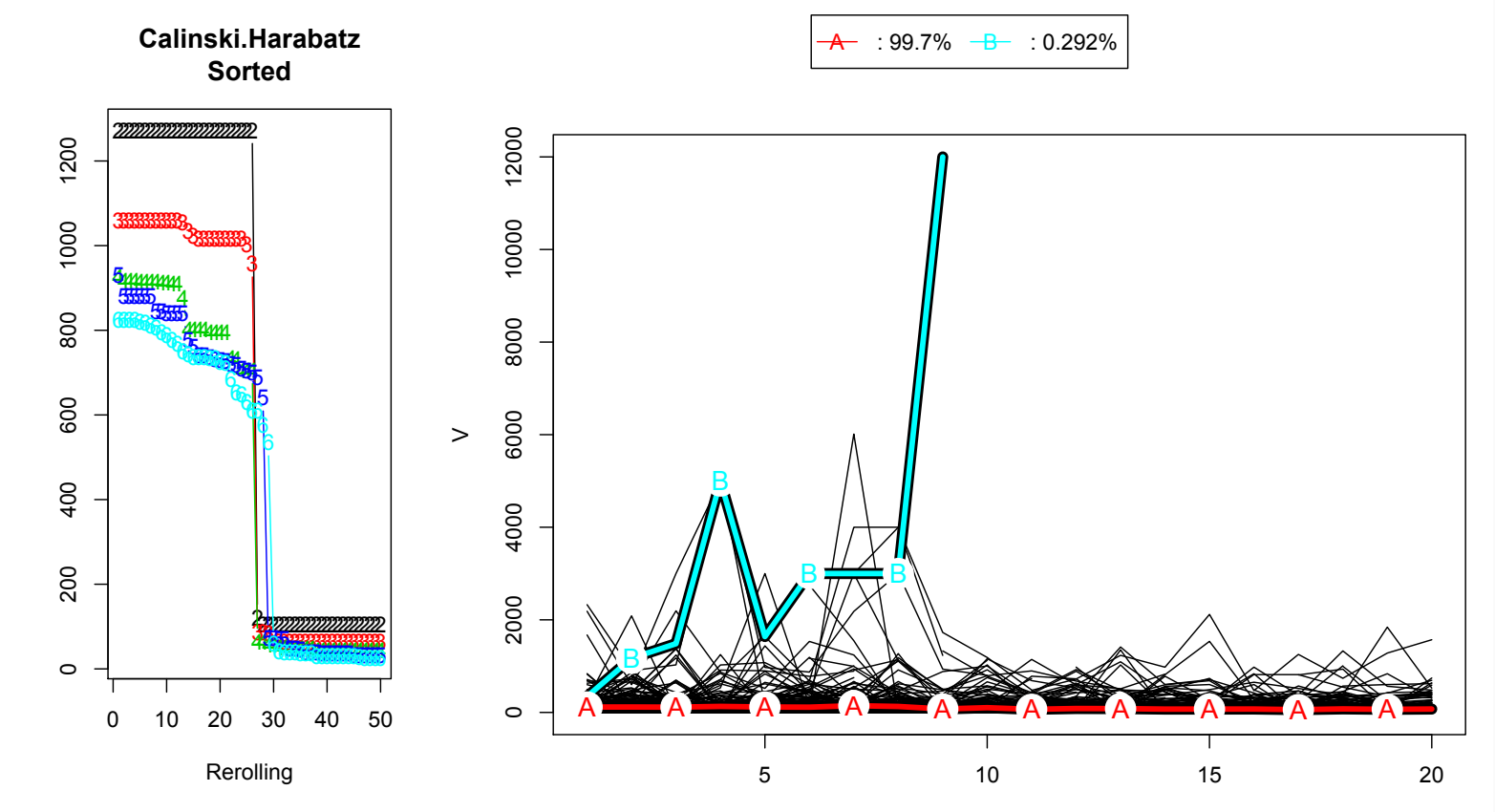
DADOS REAIS - LAMÍNULA



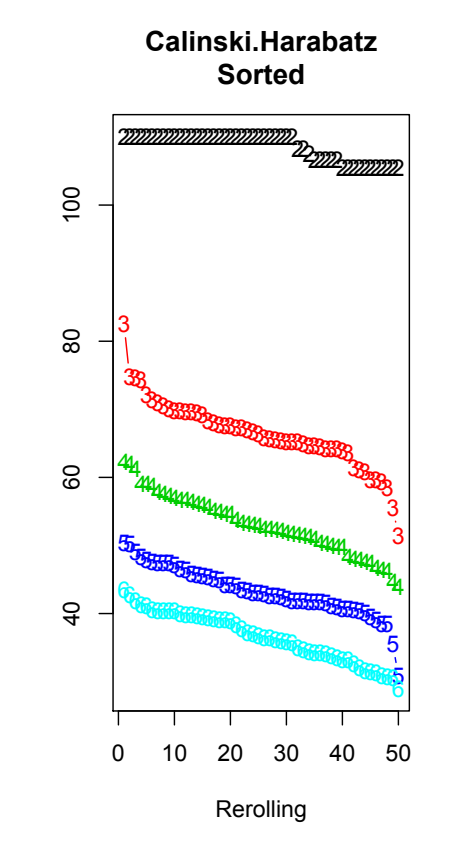
DADOS REAIS - BASTÃO - 20 SEMANAS

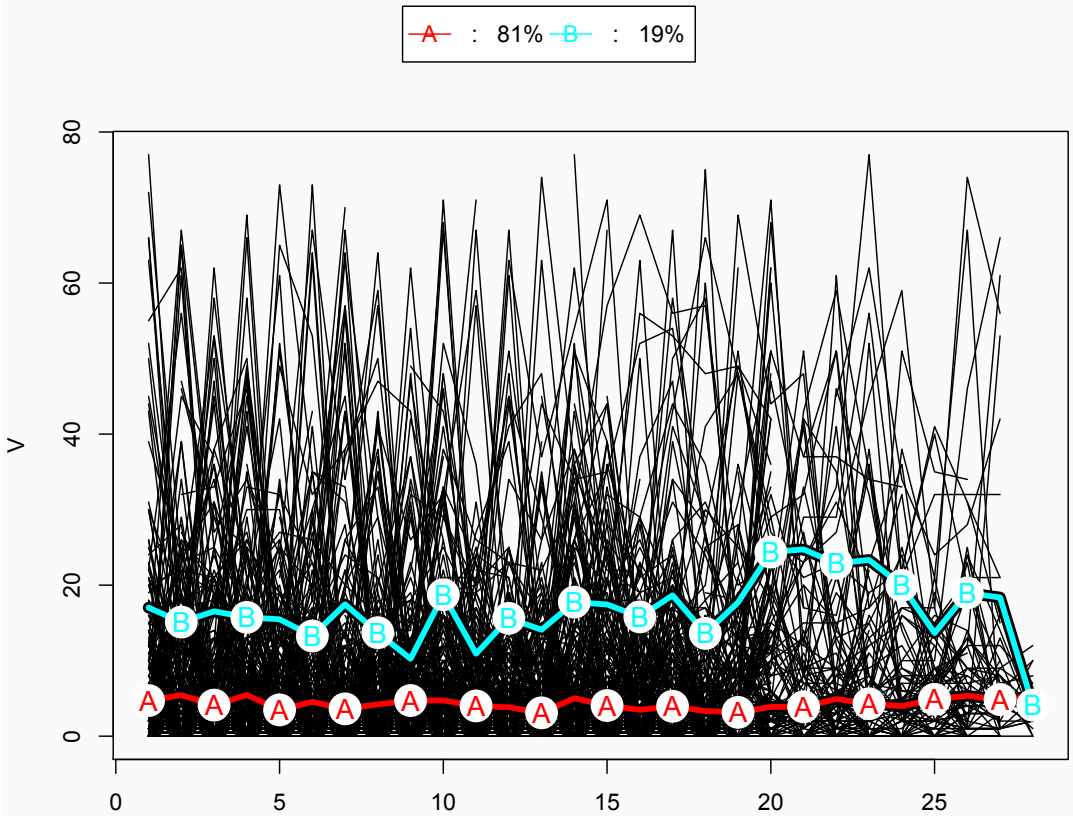


DADOS REAIS - LAMÍNULA - 20 SEMANAS

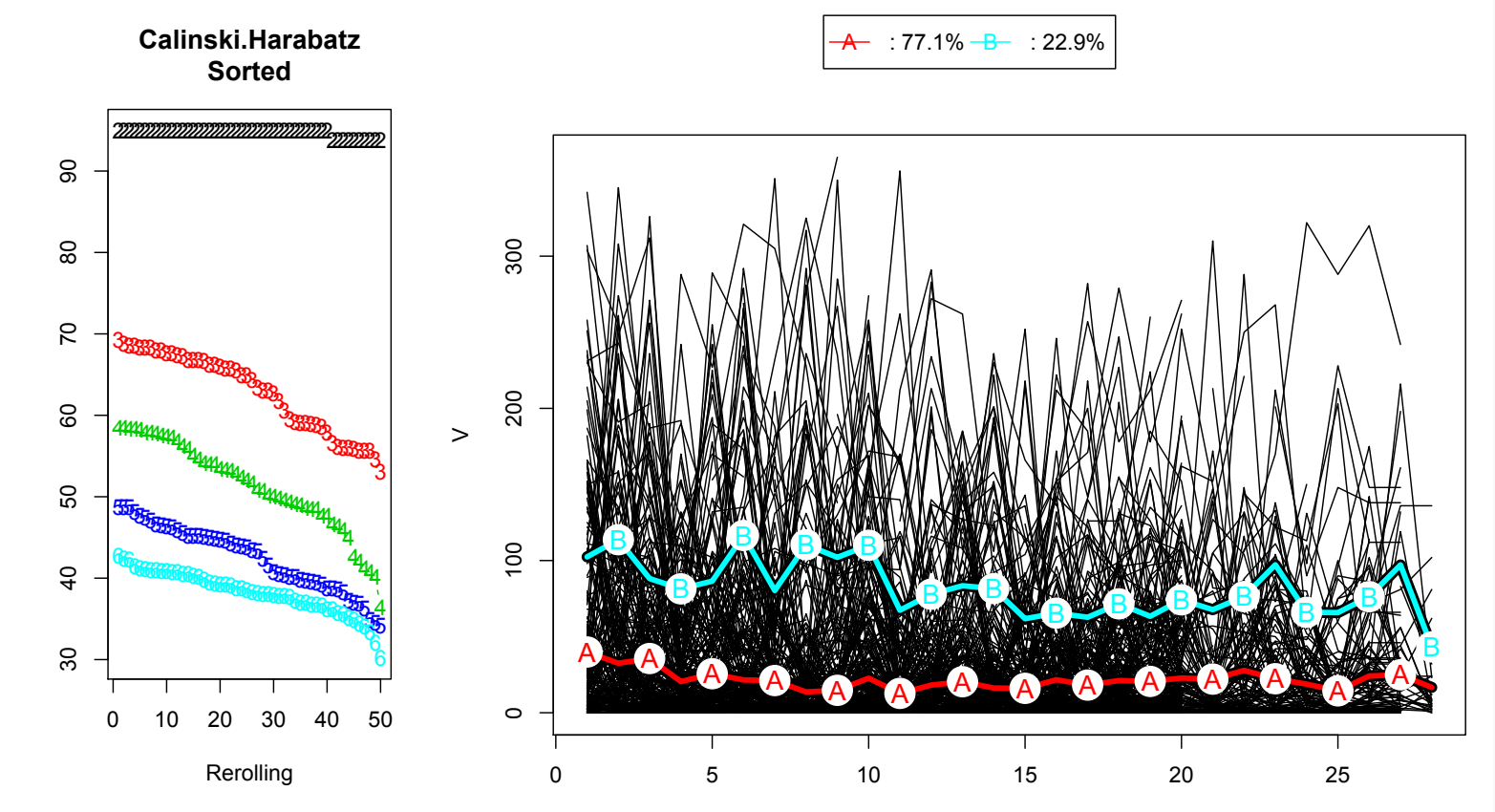


DADOS REAIS - BASTÃO - 95º PERCENTIL

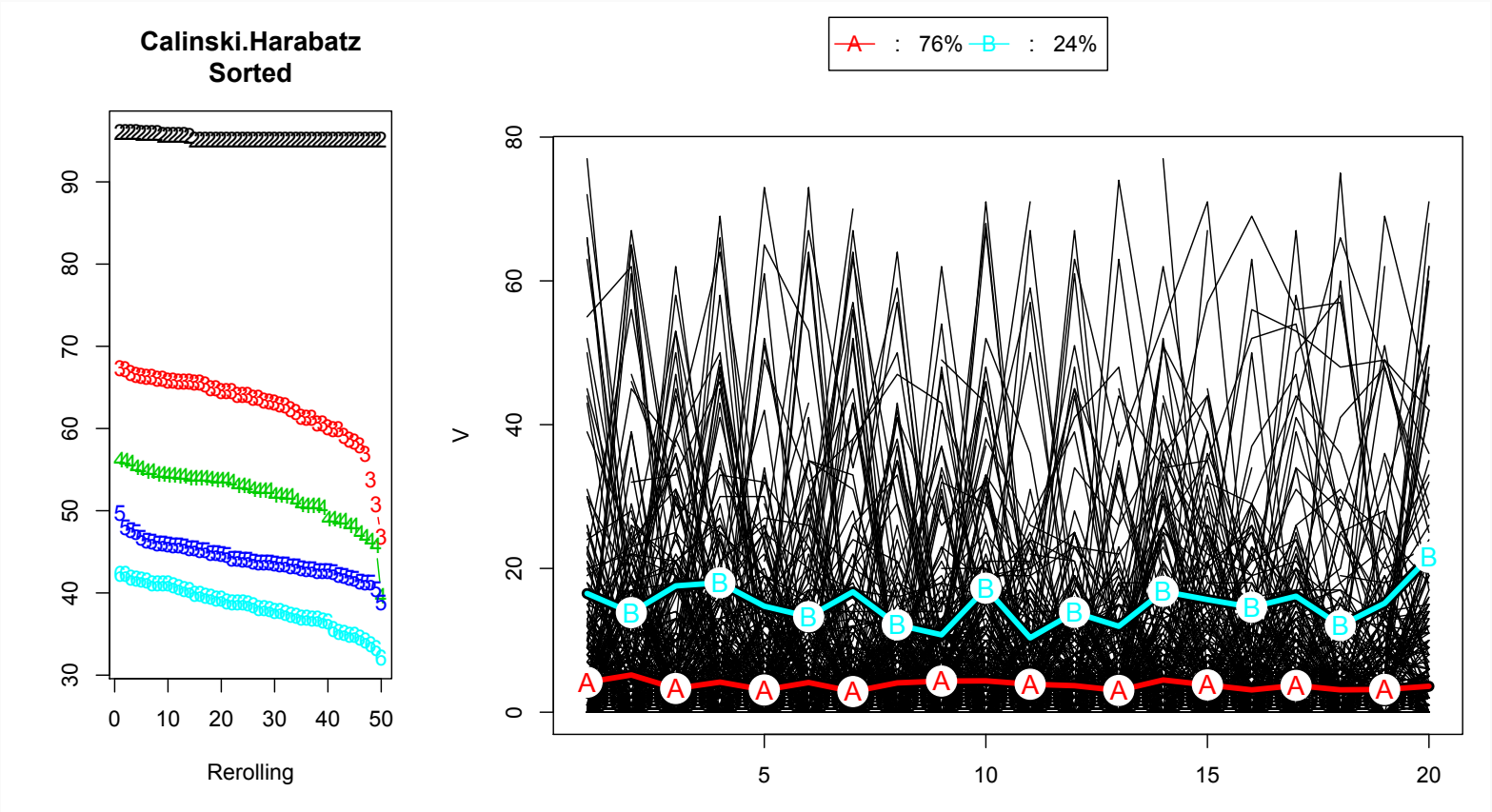




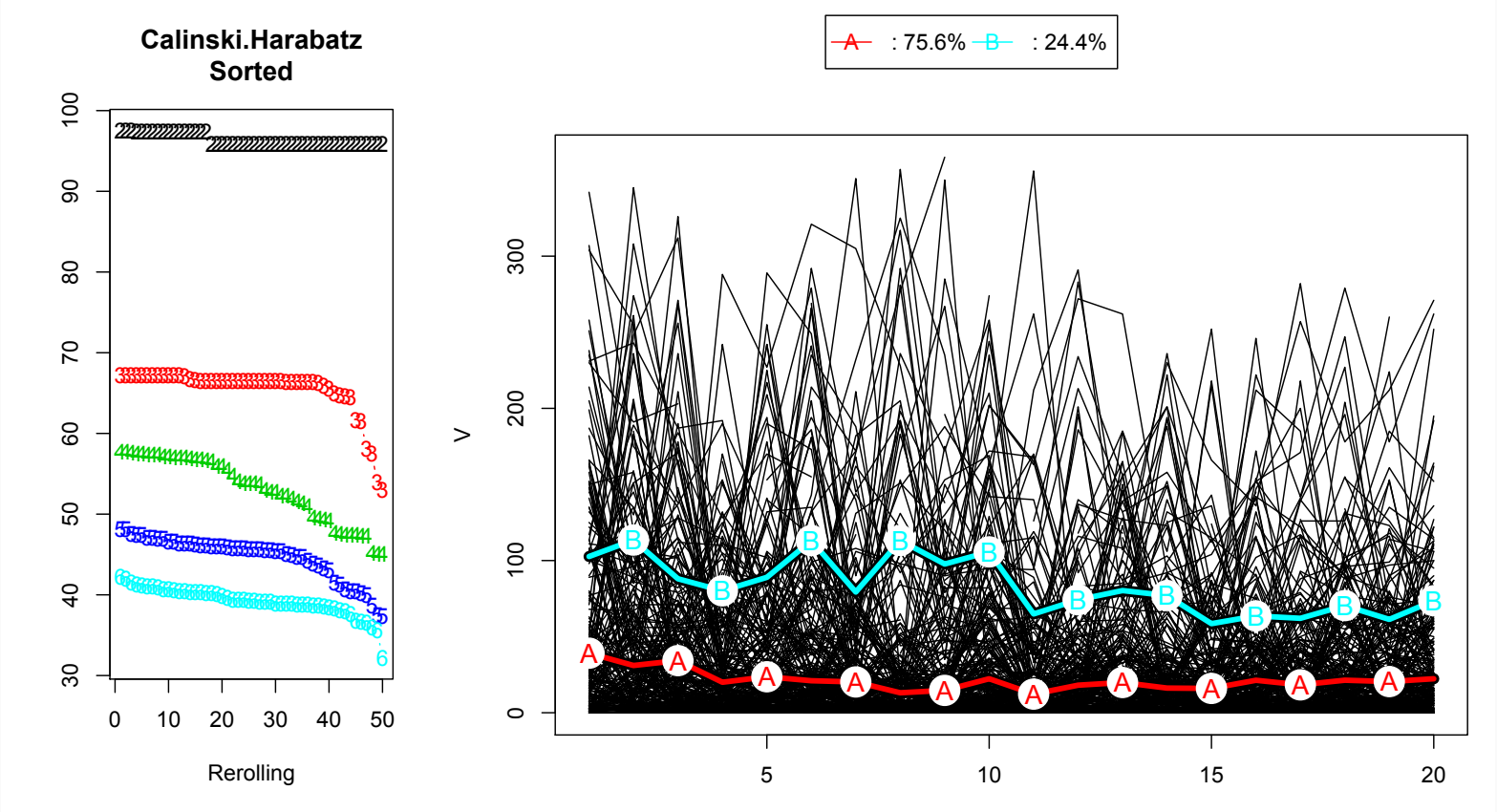
DADOS REAIS - LAMÍNULA - 95º PERCENTIL



DADOS REAIS - BASTÃO - 20 SEMANAS E 95º PERCENTIL



DADOS REAIS - LAMÍNULA - 20 SEMANAS E 95º PERCENTIL



PRÓXIMOS PASSOS

- Fazer a validação cruzada desta análise de conglomerados
- Realizar um GWAS para comparar o método proposto com o método naïve
- Ajustar um modelo linear frequentista com resposta discreta e medições repetidas
- Ajustar um modelo linear bayesiano com resposta discreta e medições repetidas
- Delinear um experimento para coleta de material genético

CONCLUSÃO

CONCLUSÃO

- Big Data são 3 V: Volume, Velocidade e Variedade
- Há muita coisa ainda para ser analisada e descoberta
- Existe uma infinidade de dados prontos para serem analisados e mais dados ainda sendo produzidos cada vez mais
- Espero que vocês tenham gostado do que viram nas últimas 4 horas
- Venham para a UFRN!

INTRODUÇÃO AO BIG DATA

Encontro Paraibano de Estatística - EPBEST 2016

Marcus Nunes

19 e 20 de maio de 2016

Universidade Federal do Rio Grande do Norte