A LINGUAGEM R NA PESQUISA

Filipe J. Zabala

Escola de Ciências
PUCRS

2010-01-08 · 2019-01-10
Porto Alegre · RS · Brasil

PUCRS ESCOLA DE CIÊNCIAS

PUCRS 360°

CIÊNCIAS
EXTENSÃO - CURTA DURAÇÃO

PUCRS 360°

A Linguagem R na Pesquisa

com
Me. Filipe Jaeger Zabala

MARISTA PUCRS DO TAMANHO DO FUTURO

Material

`www.estatisticaclassica.com`

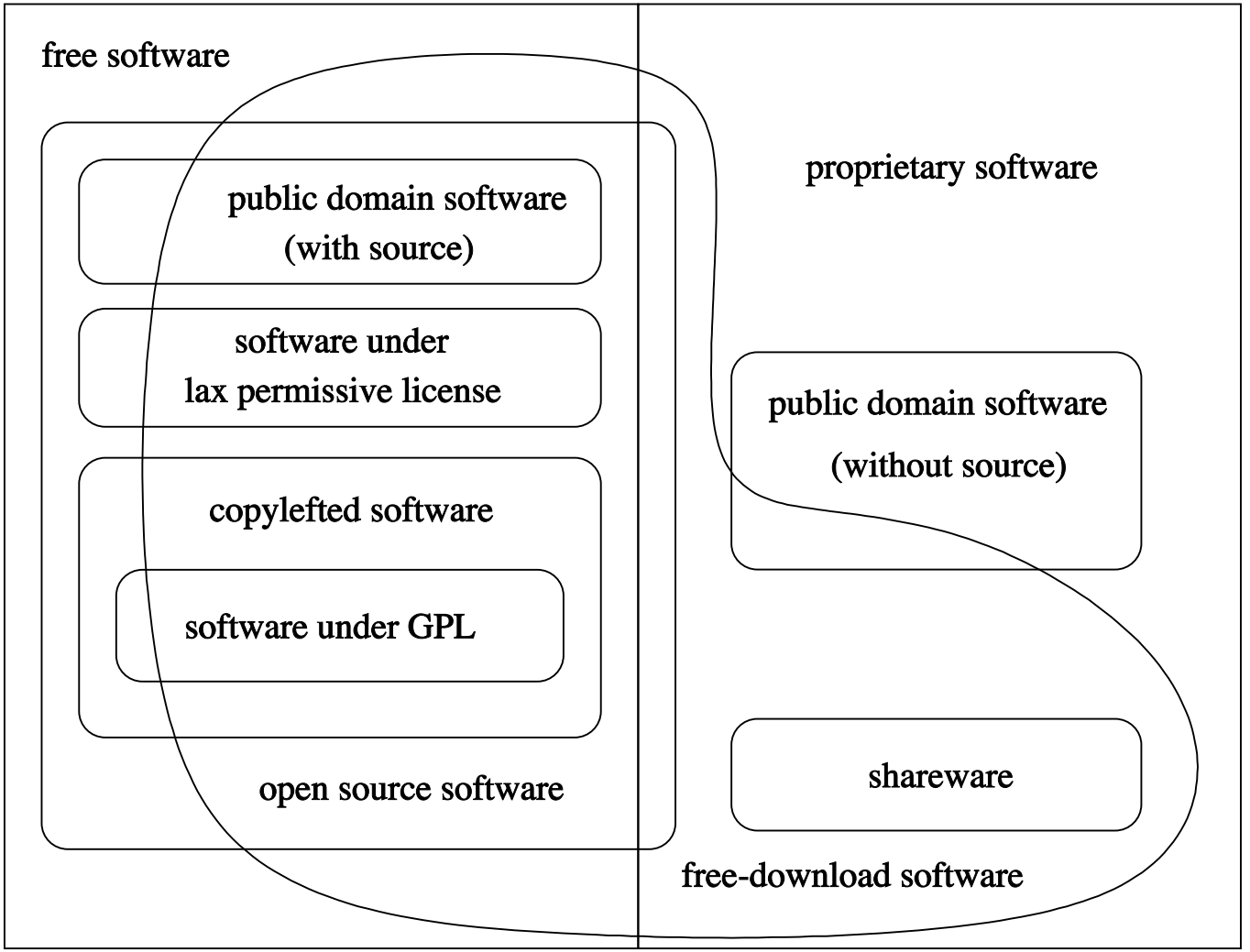
Quem sou

- Filipe Jaeger Zabala · `filipe.zabala@pucrs.br`
  - 2000-2004 Bacharel em Estatística (IM-UFRGS)
  - 2006-2009 Mestre em Ciências (Estatística) (IME-USP)
  - 2009-hoje Sócio da ZN Consultoria Estatística
  - 2010-hoje Professor da Escola de Ciências da PUCRS
  - 2016-hoje Colaborador do LaFa-PUCRS*
  - 2017-hoje Coordenador da Associação Brasileira de Jurimetria no RS
  - 2019-2022 Doutorando do PPG da Psiquiatria da UFRGS

* Laboratório de Áudio e Fonética Acústica

Ferramentas

Categorias de software



gnu.org/philosophy/categories.en.html

As 4 liberdades do software livre

0. A liberdade de *executar* o programa como quiser, para qualquer propósito.

1. A liberdade de *estudar* como o programa funciona, e adaptar às suas necessidades.*

2. A liberdade de *redistribuir* cópias e assim você pode ajudar seu vizinho.

3. A liberdade de *melhorar* o programa, e lançar suas melhorias ao público, de forma a beneficiar toda a comunidade.*

“ *Software livre é uma questão de liberdade, não de preço.*”

~ Richard Stallman

gnu.org

* Acesso ao código fonte é uma precondição para os itens 1 e 3.

Ferramentas e links

	R	RStudio	Tabula
Caracterização	Linguagem e ambiente	Editor de R (IDE)	Extrator tabelas (PDF)
Versão utilizada	3.5.2	1.2.1206 (*preview*)	1.2.1
Tipo de software	Livre	Código aberto	Livre
Máximo de linhas	Ilimitado*	-	Ilimitado*
Extensão	Pacotes adicionais	Pacotes adicionais	-
Link	`r-project.org`	`rstudio.com`	`tabula.technology`
Suporte	Comunidade**	RStudio†	Comunidade‡

- Bookdown
  `bookdown.org`
- CRAN Task Views
  `cran.r-project.org/web/views`
- Wiki R
  `ufrgs.br/wiki-r`
- Curso R
  `curso-r.com`

* Depende das memórias RAM+SWAP disponíveis.
** r-bloggers.com, stackoverflow.com, stackexchange.com
† support.rstudio.com/hc/en-us
‡ github.com/tabulapdf/tabula/issues

R

- Linguagem e ambiente de programação para cálculos estatísticos e gráficos
- Desenvolvido no departamento de Estatística da Universidade de Auckland
- Código disponível sob a licença GNU* GPL**
- Atualmente a *R Foundation* está sediada na Universidade de Economia e Negócios de Viena, Áustria
- Influenciado por linguagens como *S* e *Scheme*, de conceito minimalista orientado a objeto
- Especifica um pequeno núcleo padrão acompanhado de pacotes para a extensão da linguagem
- Em 2018-01-08 contava com 13644 pacotes†

* **GNU**'s **N**ot **U**nix, gnu.org/gnu/manifesto.html

** A Licença Pública Geral GNU é um tipo de licença utilizada para software livre, que garante aos usuários finais (indivíduos, organizações ou empresas) a liberdade de usar, estudar, compartilhar e modificar o software.

† cran.r-project.org/web/packages

RStudio

- Editor e ambiente de desenvolvimento integrado ao R
- Amplia as funcionalidades básicas da linguagem
- Possibilita a criação de documentos automáticos em html, pdf e docx
- Disponível nas versões *Desktop* e *Server*
- A versão *Preview* possui recursos mais avançados, porém menos estáveis
- Gerencia múltiplas linguagens
  - R
  - LaTeX
  - Markdown
  - Python
  - SQL
  - Stan
  - HTML
  - C++
  - D3

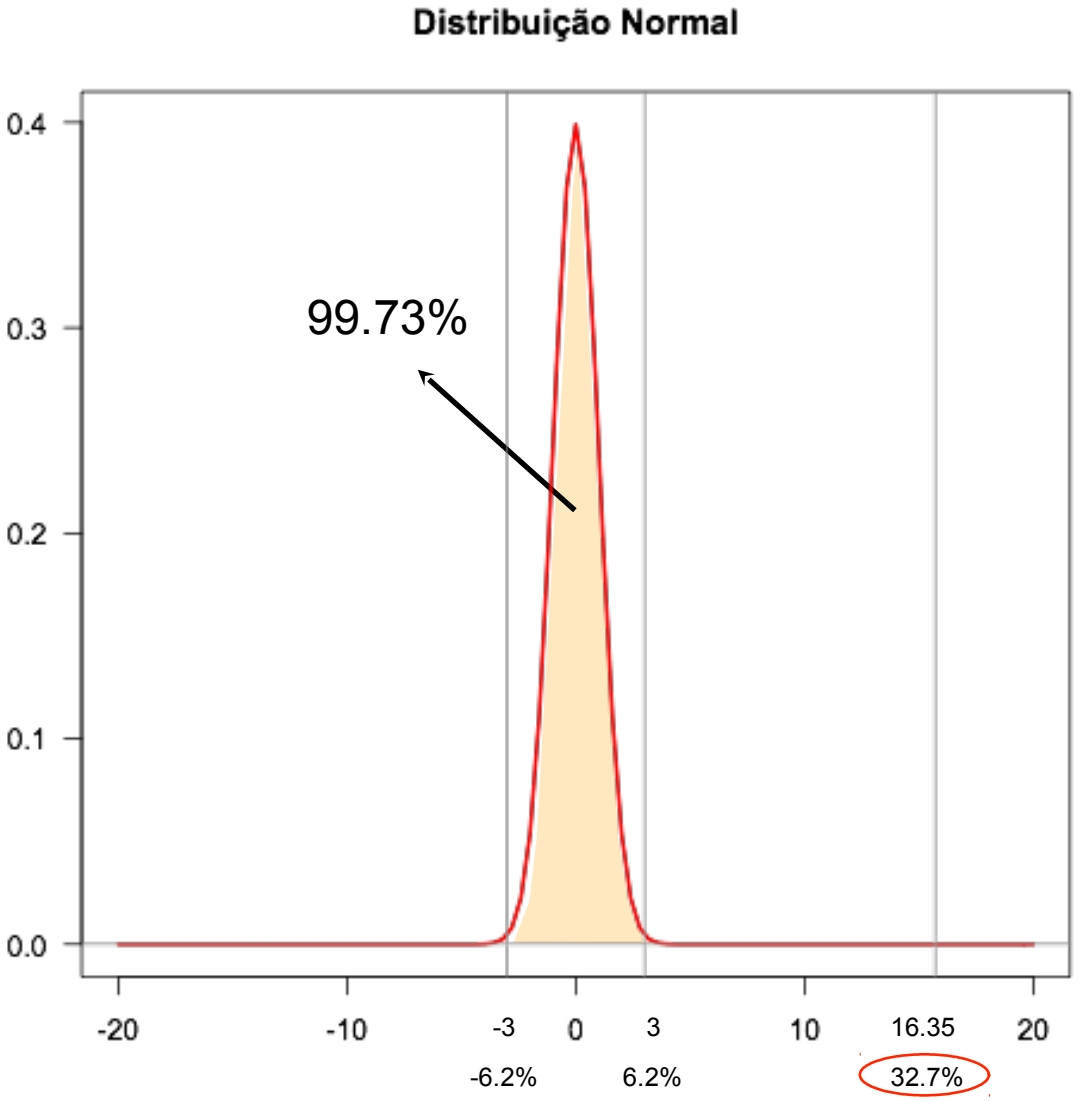
Tabula

- Ferramenta para liberar tabelas de dados trancadas em PDFs
- Compilado para Mac e Windows
- Código-fonte disponível no GitHub
- Usado para gerar reportagens investigativas em organizações de notícias
- Utilizado por organizações de base como a `schoolcuts.org`
- Pesquisadores transformam relatórios PDF em planilhas do Excel, CSVs e arquivos JSON para uso em aplicativos de banco de dados e análise

Exemplos

Jurimetria

Suspeita de racismo pela polícia



46.2% - 13.5% = 32.7%

Observado

População

Diferença

Probabilidade de as paradas supostamente racistas serem devidas ao acaso:

4.349800077e-60

ou

$0.0^{57}4349800077\%$

ou

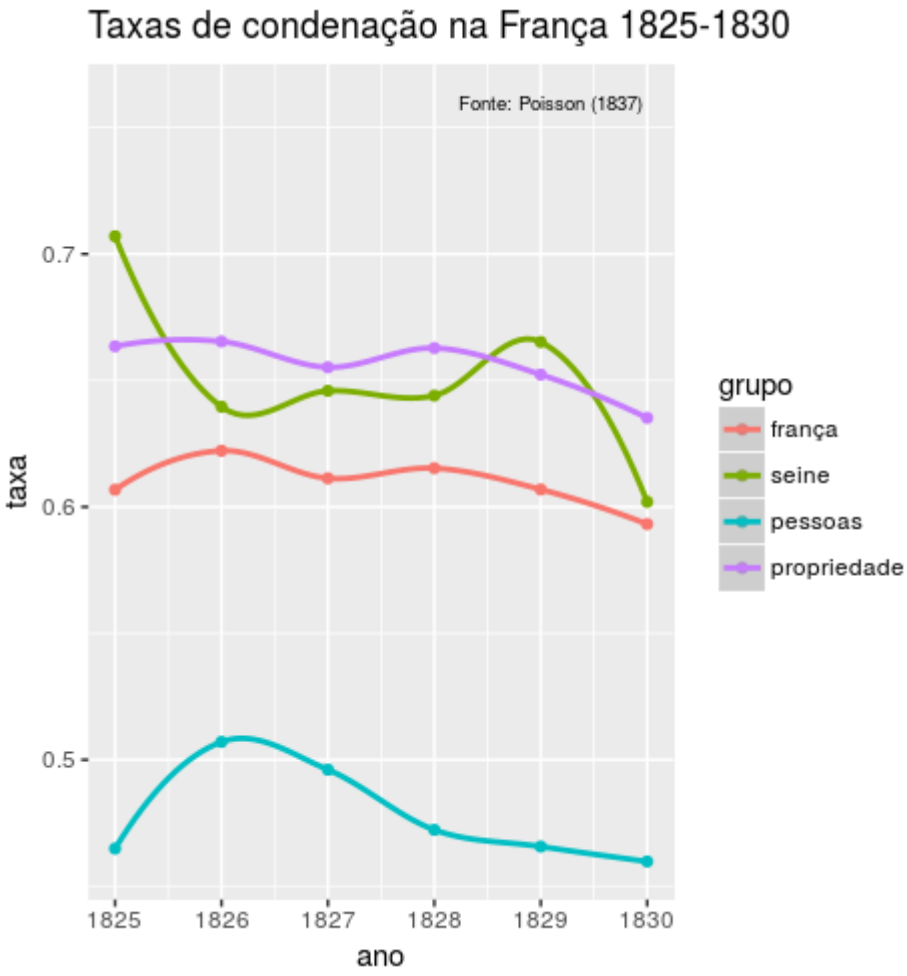
0.00000000000000000000000000000000000000000000000000004349800077%

Fonte: Kadane (2008) - Statistics in the Law.

Taxas de condenação na França

- Quételet 1835
  - Modelos para análise do veredito de júris através das taxas de condenação na França · 1825-1830
  - Objetivo: sociedade
  - Conceito do *Homem Médio (l'Homme Moyen)*

- Poisson 1837
  - Paralelamente a Quételet desenvolve modelos para análise do veredito de júris
  - Objetivo: método
  - Probabilidade de condenação até 1830...
    - na França: 61%
    - em Seine 65%
    - por crime contra a pessoa: 48%
    - por crime contra a propriedade: 66%



Fonte: Poisson (1837) - Recherches sur la probabilité des jugements en matière criminelle et en matière civile.
https://ia800209.us.archive.org/27/items/recherchessurlap00pois/recherchessurlap00pois.pdf

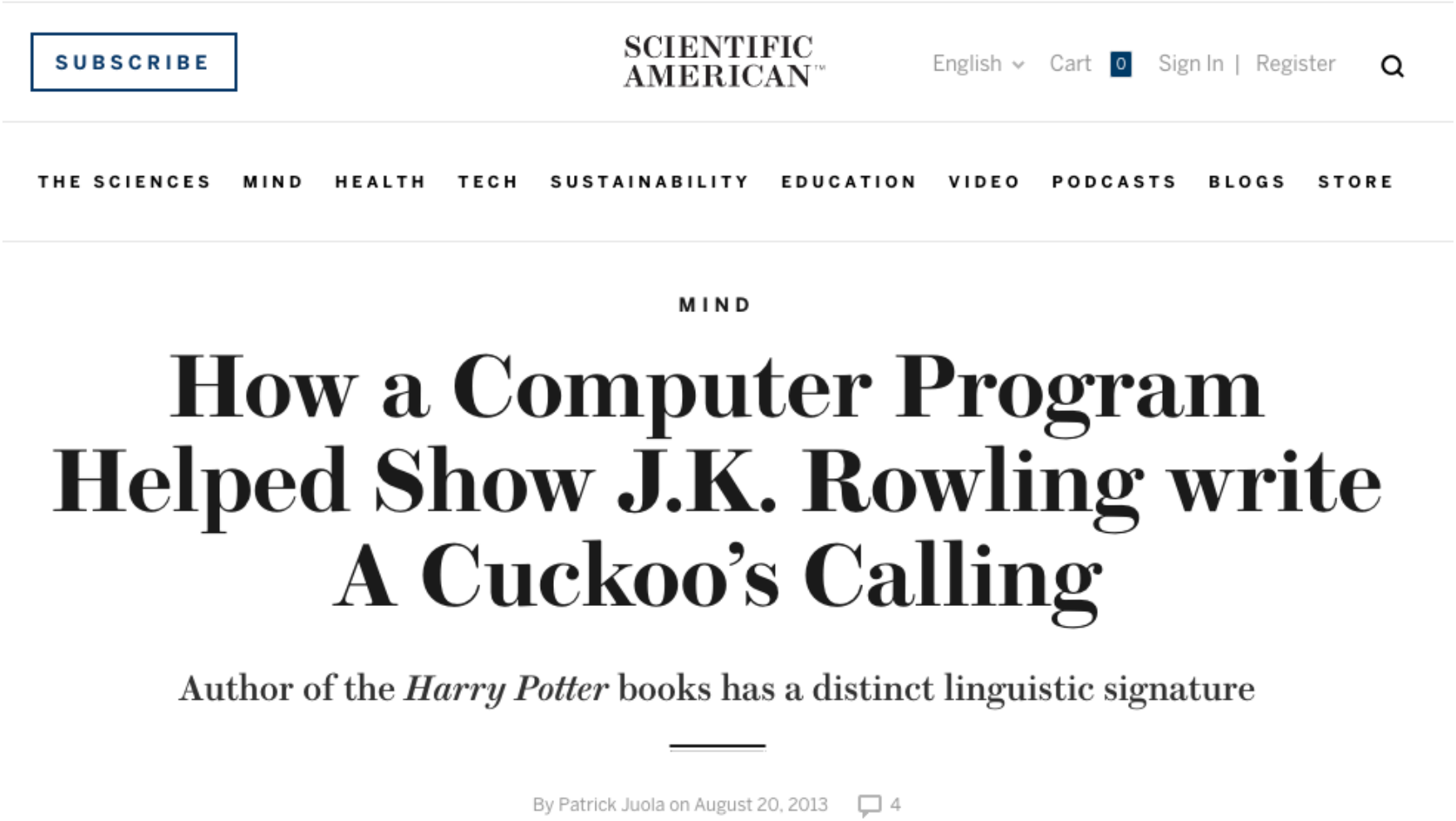
Cautelar de sustação de protesto

- Pedido liminar
  - Empresa X com vendas diárias
  - Maioria dos clientes deixam de comprar se X estiver protestada
- Para cada dia de protesto
  - Calcula-se a chance de prejuízo
  - Estima-se o valor esperado deste prejuízo
- Exemplo
  - Média de 150 consultas/mês
  - Probabilidade (histórica) de efetivação de 35%
  - Ticket médio de R$15680.00/cliente
  - Valor esperado de perda/dia útil de R$ 37418.18

$$E(X) = \frac{150 \times 0.35 \times 15680}{22} = 37418.18$$

Classificador Textual

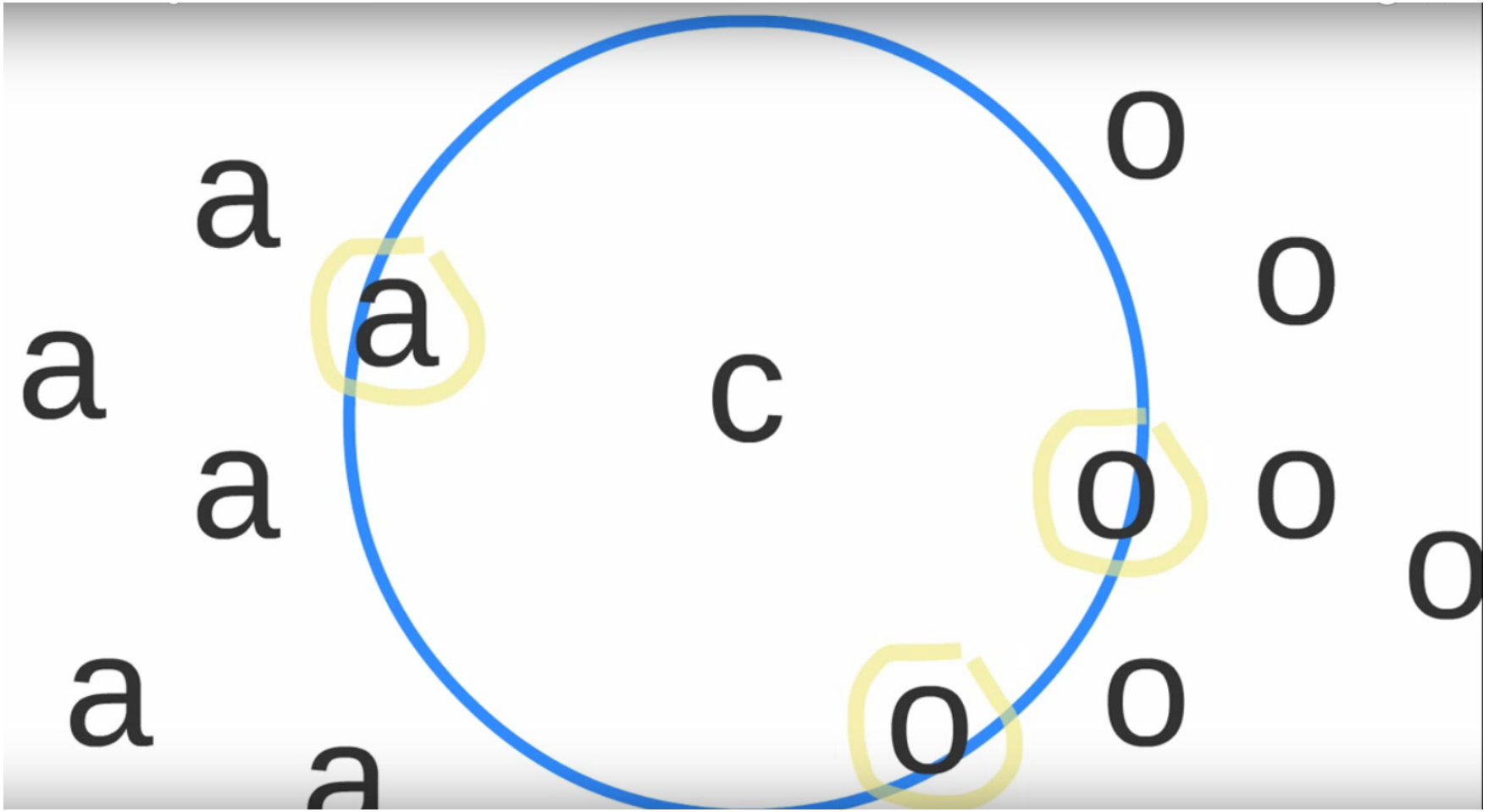
Assinatura Textual



Fonte: http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/

Método kNN

k vizinhos mais próximos
k nearest neighbours

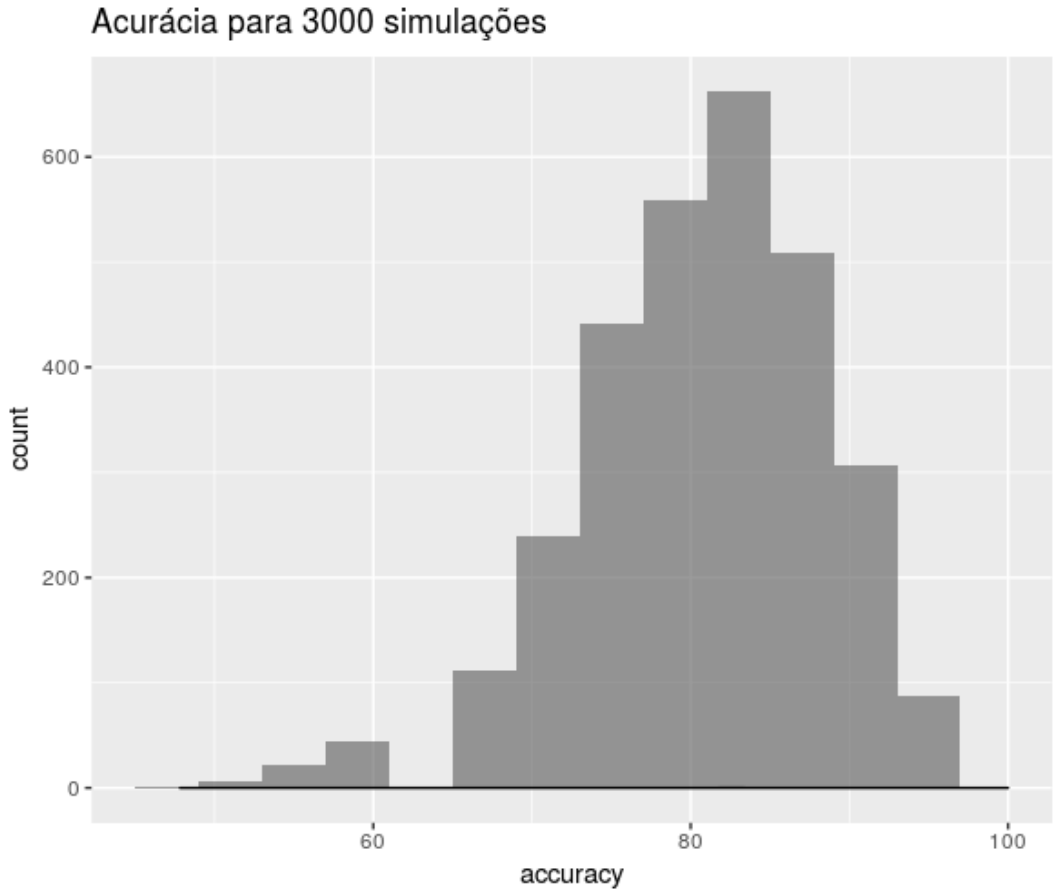


Fonte: Körting (2014) - How kNN algorithm works.

Classificação de discursos

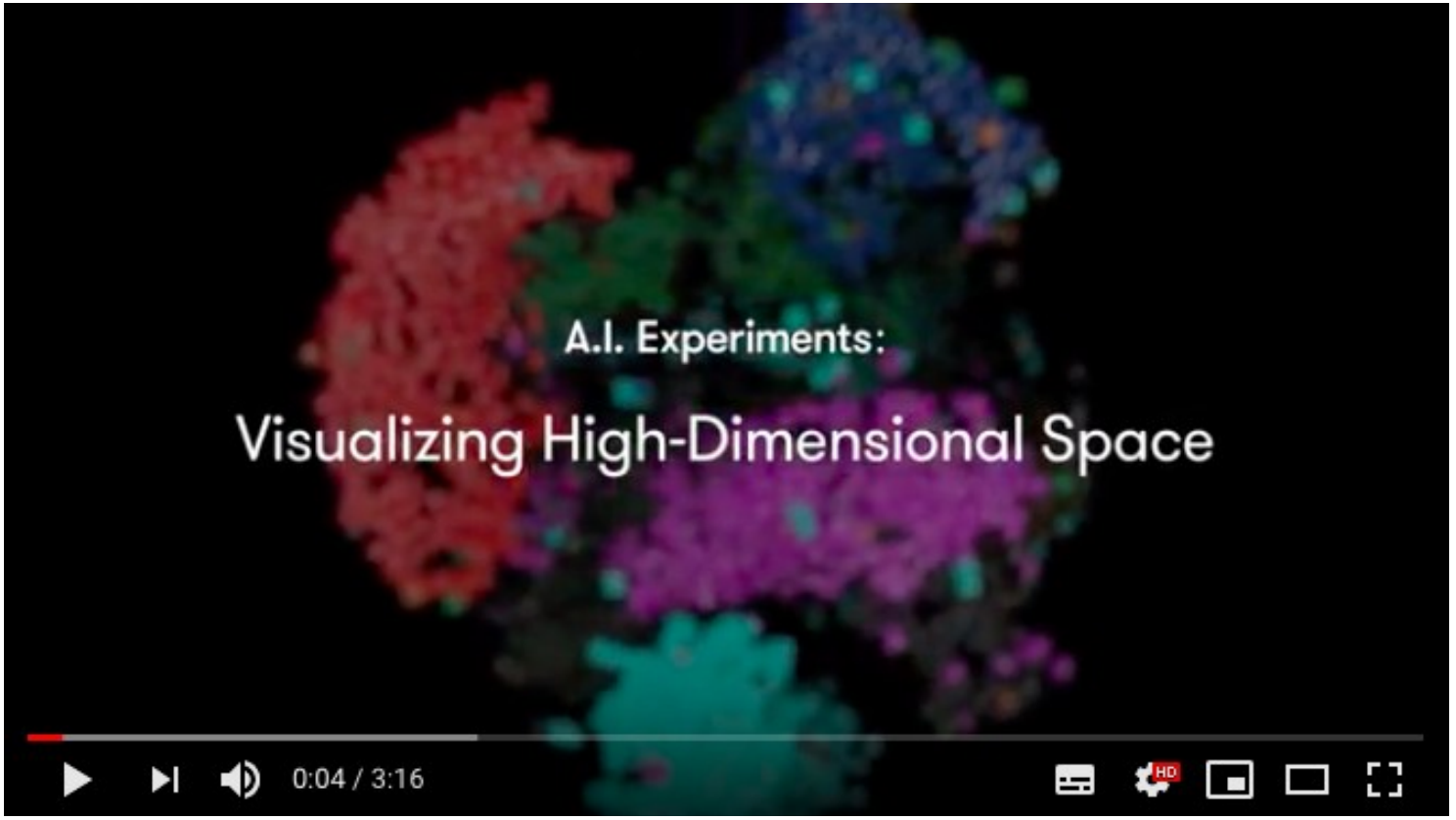
Barack Obama x G.W. Bush

```
             Predictions
Actual        barackobama gwbush
  barackobama          13      1
  gwbush                1      8
> (accuracy <- sum(diag(conf.mat)/length(test.idx) * 100))
[1] 91.30435
```



Fonte: http://obamaspeeches.com
http://www.americanrhetoric.com/speechbanka-f.htm

Classificação textual (e etcetera)

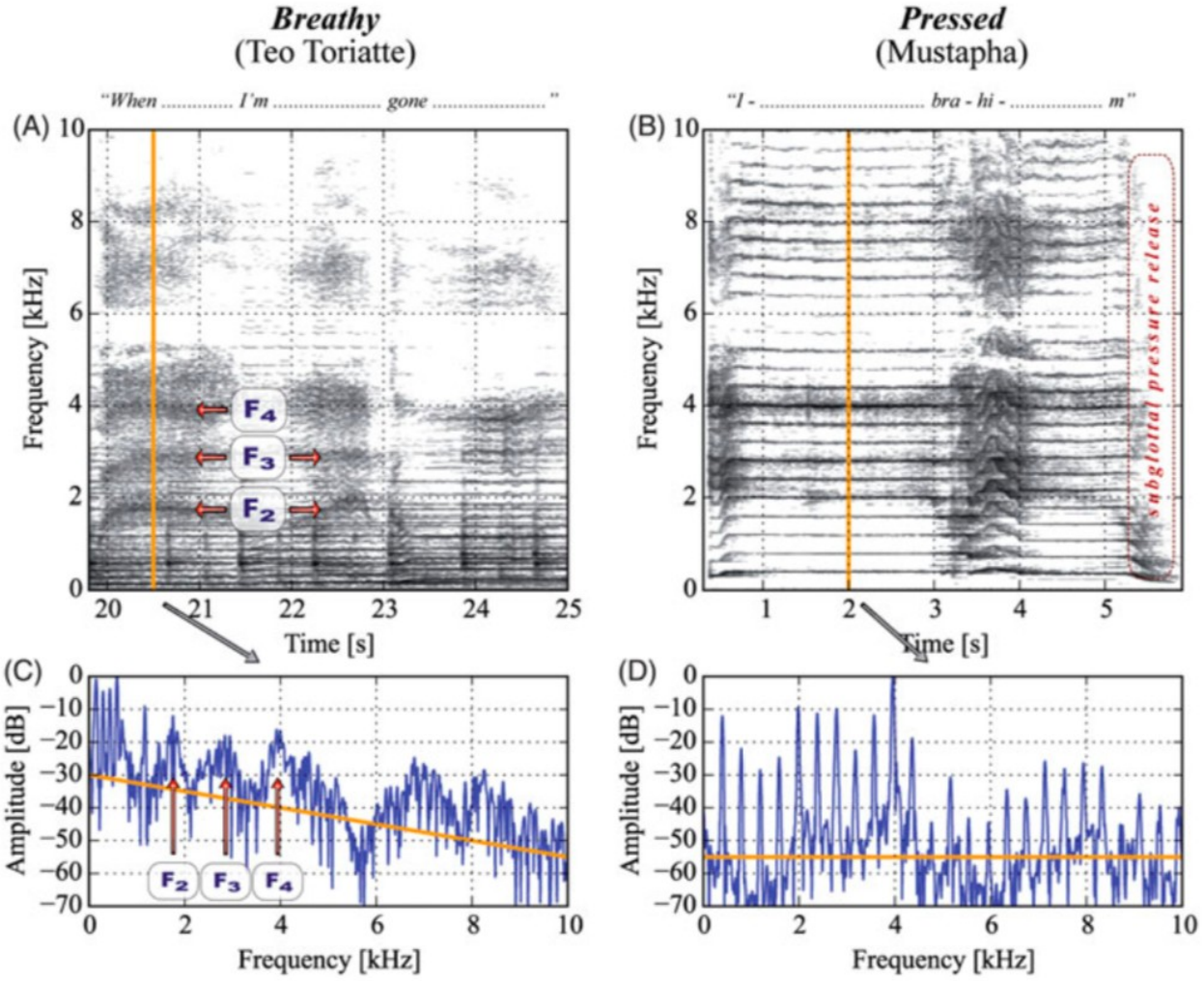


Fonte: https://www.youtube.com/watch?v=wvsE8jm1GzE

Reconhecimento de voz

Reconhecimento de voz

- Espectrograma da voz de Freddie Mercury



Fonte: Herbst et al (2016) - Freddie Mercury – Acoustic analysis of speaking fundamental frequency vibrato and subharmonics.

Reconhecimento de voz

- Algoritmo desenvolvido no LaFa-PUCRS
  - 2 homens, 1 mulher, 5 repetições de texto guiado
  - Classificador SVM (*Support Vector Machine*)
  - Relação treino-teste 50-50, cost = 8, gamma = 0.5
  - Função `svm` da biblioteca `e1071` do R

Fonte: Os autores.

Reconhecimento de voz

- Algoritmo desenvolvido no LaFa-PUCRS
  - 2 homens, 1 mulher, 5 repetições de texto guiado
  - Classificador SVM (*Support Vector Machine*)
  - Relação treino-teste 50-50, cost = 8, gamma = 0.5
  - Função `svm` da biblioteca `e1071` do R


```
fit50 <- classSVM(dfPorNome,0.5)

$tab.total
      true
pred   ana1 ana2 ana3 ana4 ana5 den1 den2 den3 den4 den5 fil1 fil2 fil3 fil4 fil5
  ana1  174   43   27   27   20    0    0    0    1    0    1    1    0    3    2
  ana2   27  177   32   27   15    0    0    0    1    1    1    1    1    1    1
  ana3   24   16  161   31   26    2    1    0    0    0    0    1    0    0    1
  ana4   34   22   27  139   21    0    0    0    0    0    0    1    2    2    1
  ana5   20   30   28   31  169    0    0    0    0    0    0    0    0    1    0
  den1    0    1    0    0    1  193   27   16   23   26    8   12   11    7   12
  den2    0    0    2    1    0   14  193   14   16    9    3    5    3    1    5
  den3    2    0    0    0    0   20   32  171   21   35    8    2    3    1    7
  den4    0    1    1    1    0   13   15   27  177   16    7   10    6    2    5
  den5    1    1    0    2    0   15   24   24   24  177    6   11    7    5    8
  fil1    1    2    0    8    1    9    5    1    9    6  195   32   23   15   29
  fil2    0    0    0    1    0    6    3    4    1    6   19  149   37   18   16
  fil3    2    0    0    2    0    7    3    3    2    1   19   16  148   20   17
  fil4    0    1    1    3    1    2    5    6    5    5   17   23   26  169   22
  fil5    1    2    0    1    2    4    2    6    6    6   13   19   26   18  168
```


Fonte: Os autores.

Reconhecimento de voz

- Algoritmo desenvolvido no LaFa-PUCRS
  - 2 homens, 1 mulher, 5 repetições de texto guiado
  - Classificador SVM (*Support Vector Machine*)
  - Relação treino-teste 50-50, cost = 8, gamma = 0.5
  - Função `svm` da biblioteca `e1071` do R


```
fit50 <- classSVM(dfPorNome,0.5)
```

$tab.total

pred	true ana1	ana2	ana3	ana4	ana5	den1	den2	den3	den4	den5	fil1	fil2	fil3	fil4	fil5
ana1	174	43	27	27	20	0	0	0	1	0	1	1	0	3	2
ana2	27	177	32	27	15	0	0	0	1	1	1	1	1	1	1
ana3	24	16	161	31	26	2	1	0	0	0	0	1	0	0	1
ana4	34	22	27	139	21	0	0	0	0	0	0	1	2	2	1
ana5	20	30	28	31	169	0	0	0	0	0	0	0	0	1	0
den1	0	1	0	0	1	193	27	16	23	26	8	12	11	7	12
den2	0	0	2	1	0	14	193	14	16	9	3	5	3	1	5
den3	2	0	0	0	0	20	32	171	21	35	8	2	3	1	7
den4	0	1	1	1	0	13	15	27	177	16	7	10	6	2	5
den5	1	1	0	2	0	15	24	24	24	177	6	11	7	5	8
fil1	1	2	0	8	1	9	5	1	9	6	195	32	23	15	29
fil2	0	0	0	1	0	6	3	4	1	6	19	149	37	18	16
fil3	2	0	0	2	0	7	3	3	2	1	19	16	148	20	17
fil4	0	1	1	3	1	2	5	6	5	5	17	23	26	169	22
fil5	1	2	0	1	2	4	2	6	6	6	13	19	26	18	168

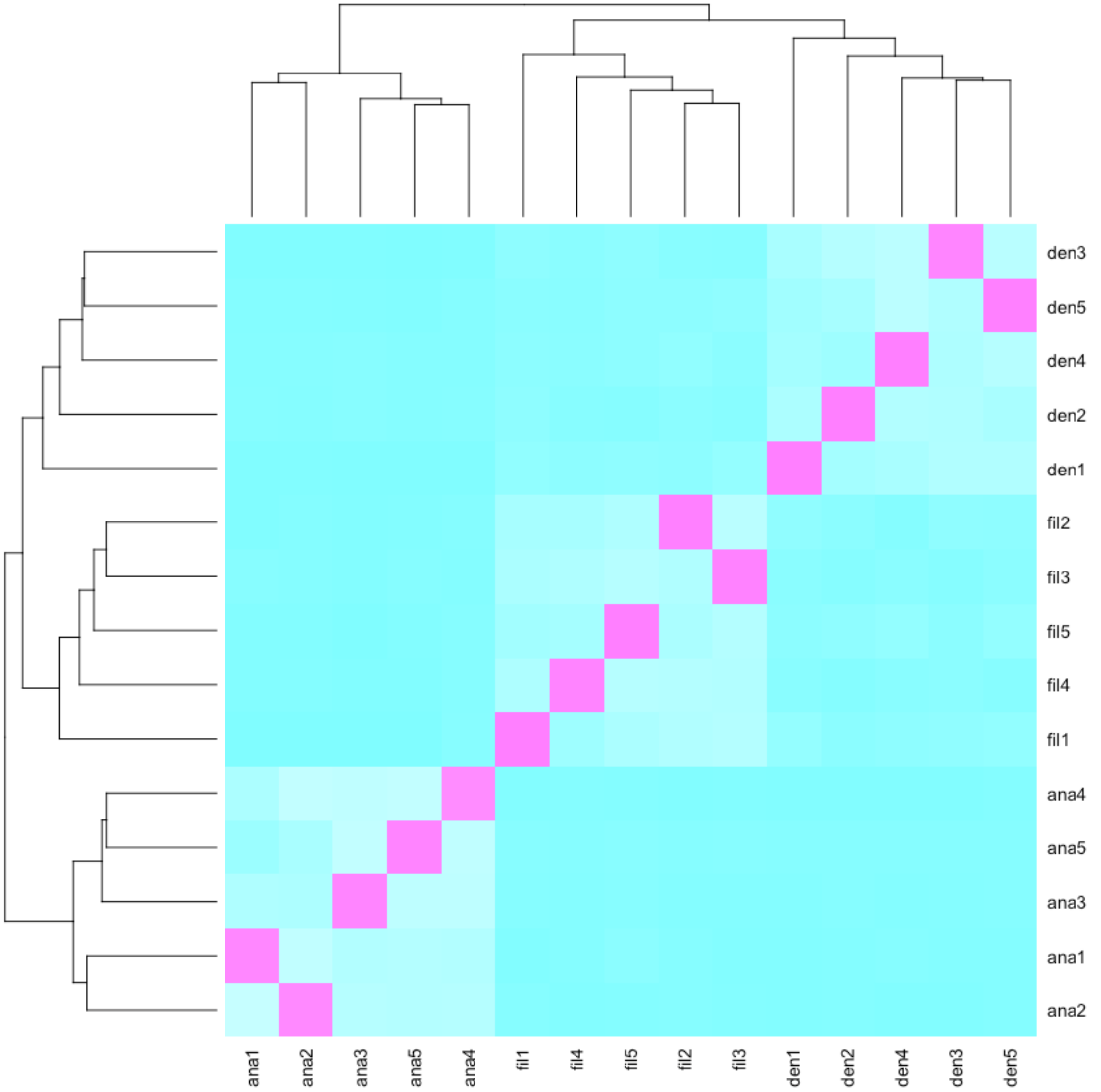
$tab.simbolico

pred	true a1	a2	a3	a4	a5	d1	d2	d3	d4	d5	f1	f2	f3	f4	f5
ana1	$	$	$	$	$	.	.	.	.	.	.	.	.	-	-
ana2	$	$	$	$	+	.	.	.	.	.	.	.	.	.	.
ana3	$	+	$	$	$	-	.	.	.	.	.	.	.	.	.
ana4	$	$	$	$	$	.	.	.	.	.	.	.	-	-	.
ana5	$	$	$	$	$	.	.	.	.	.	.	.	.	.	.
den1	.	.	.	.	.	$	$	+	$	$	+	+	+	+	+
den2	.	.	-	.	.	+	$	+	+	+	-	+	-	.	+
den3	-	.	.	.	.	$	$	$	$	$	+	-	-	.	+
den4	.	.	.	.	.	+	+	$	$	+	+	+	+	-	+
den5	.	.	.	-	.	+	$	$	$	$	+	+	+	+	+
fil1	.	-	.	+	.	+	+	.	+	+	$	$	$	+	$
fil2	.	.	.	.	.	+	-	+	.	+	$	$	$	+	+
fil3	-	.	.	-	.	+	-	-	-	.	$	+	$	$	+
fil4	.	.	.	-	.	-	+	+	+	+	+	$	$	$	$
fil5	.	-	.	.	-	+	-	+	+	+	+	$	$	+	$

Fonte: Os autores.

Reconhecimento de voz

```
heatmap(fit50$tab.total, col = cm.colors(256))
```



Fonte: Os autores.

Alteração de voz

- VoCo Project (‘Photoshop for voice’)

Simulação de Monte Carlo e a Agulha de Buffon

Simulação de Monte Carlo

- Método para estimar quantidades a partir de simulações
- Cassinos de Monte Carlo (Mônaco)
- Estimação de π com

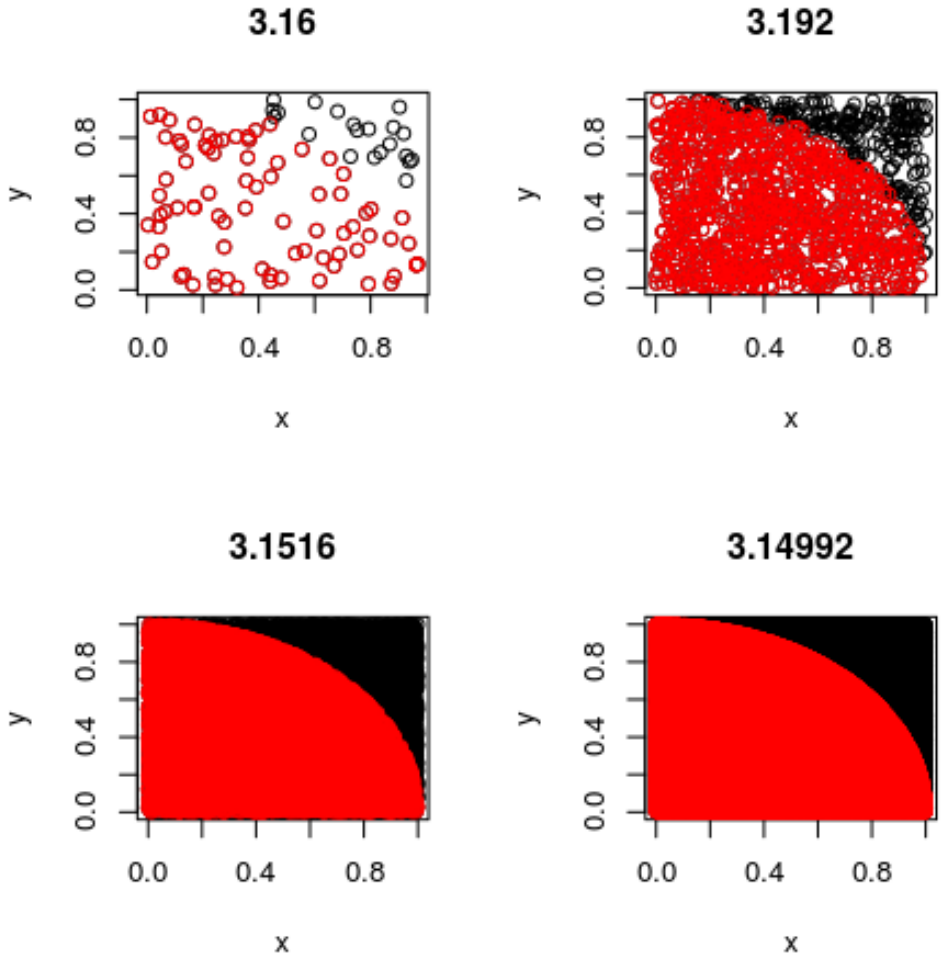
  n = 100

  n = 1 000

  n = 10 000

  n = 1 000 000

$$\frac{\#\text{ dardos no quadrante}}{\#\text{ dardos no quadrado}} = \frac{\frac{1}{4}\pi r^2}{r^2} = \frac{\pi}{4} \quad (1)$$



Fonte: Sullivan (2012) - Introduction to Data Mining for the Life Sciences

Agulha de Buffon

- A probabilidade de uma agulha ($l < t$) cruzar a linha

$$P = \int_{\theta=0}^{\frac{\pi}{2}} \int_{x=0}^{(l/2)\sin\theta} \frac{4}{t\pi}\, dx\, d\theta = \frac{2l}{t\pi}$$

- n agulhas, h cruzadas de linha

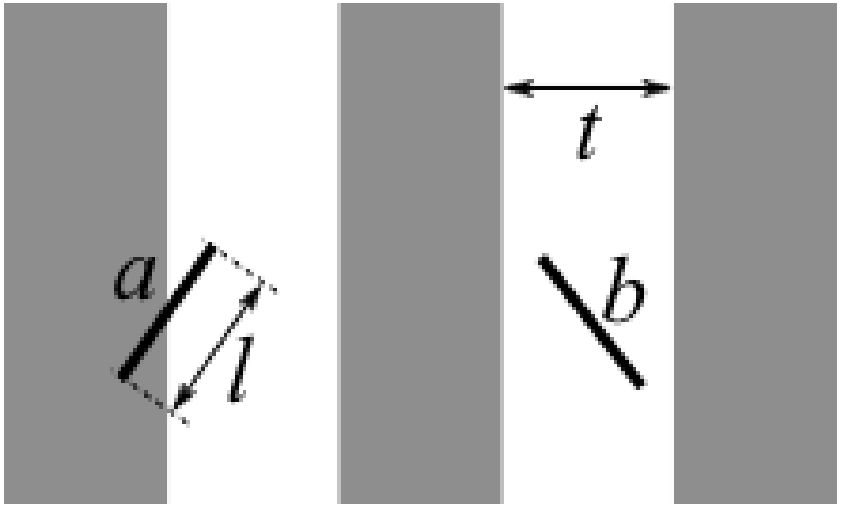
$$P \approx \frac{h}{n}$$

- Aproximação de π

$$\pi \approx \frac{2l \cdot n}{th}$$

- Se $l = t$

$$\pi \approx \frac{2n}{h}$$



Fonte: https://en.wikipedia.org/wiki/Buffon%27s_needle


Tweet
Tom Pair
@TomPair2
UBS ran 10,000 simulations and forecasted Germany to win the World Cup. Goldman Sachs ran 1,000,000 simulations & predicted Brazil as Champions. Neither team made it to the WC Semi-Finals. . . Both banks use same tech simulations for market predictions. Just saying!
Traduzido do inglês
10/07/2018 08:39
Tweete sua resposta

A LINGUAGEM R NA PESQUISA

Filipe J. Zabala

Escola de Ciências
PUCRS

2010-01-08 · 2019-01-10
Porto Alegre · RS · Brasil

PUCRS ESCOLA DE CIÊNCIAS

PUCRS 360°