

## BIG-008 Análise Preditiva na Prática

Profa.: Rosangela de Fátima Pereira Marquesone

Big Data: Inteligência na Gestão dos Dados



#### **Tópicos**

- ☐ Categorias de analytics
- Oportunidades em análise preditiva
- Praticando



#### **Big Data Analytics**

Analytics = aplicação da tecnologia computacional e de estatística
para solucionar problemas de negócios e da indústria

- Tomada de decisões baseadas em dados existentes
  - Não apenas na intuição humana

☐ Oferece meios para obter vantagem competitiva

Permite otimizar novas iniciativas



	Analytics	Big Data Analytics
Custo de armazenamento	Alto	Baixo
Análises	Off line	Tempo real
Uso do Hadoop	Não	Sim
Velocidade do carregamento de dados	Baixa	Alta
Tempo de carregamento dos dados	Longo	50 a 60% mais rápido
Variedade de dados	Estruturados	Não estruturados
Volume de dados	Gigabyte, terabyte	Petabyte, exabyte
Custo com suporte	Alto	Baixo
Tempo de administração	Longo	60% mais rápido

Fonte: Heavy Reading, "Big Data & Advanced Analytics in Telecom" - Huawei



	Ferramentas analíticas legadas	Big Data Analytics	
Custo de armazenamento	Alto	Baixo	
Análises	Off line	Tempo real	
Rig Datadoop	Melhores decisões		
Velocidade de Carregamento de			
Analytics ados Tempo de carregamento dos dados	Decisões en	n tempo real 50 a 60% mais rápido	
Variedade de dados	Estruturados	Não estruturados	
Volume de dados	Gigabyte, terabyte	Petabyte, exabyte	
Custo com suporte	Alto	Baixo	
Tempo de administração	Longo	60% mais rápido	



#### **Quatro abordagens**

Descritiva

Diagnóstica

Preditiva

Prescritiva



#### **Descritiva**

☐ Foco em sumarizar fatos ocorridos

□ Permite compreender como a organização está sendo operada

- ☐ Uso de ferramentas de *Business Intelligence* (BI)
  - Alertas
  - Dashboards
  - Relatórios



O que aconteceu?

#### **Descritiva**

- ☐ Baseado em métricas
  - Visualização de tendências
  - Descoberta de padrões
- □ Análise mais adotada pelas organizações. Mais de 80% da análise de negócios são descritivas

- ☐ Exemplos:
  - Relatório diário/semanal de vendas
  - Lista de cancelamento de assinantes



#### Diagnóstica

- ☐ Foco em determinar o motivo de um evento ter ocorrido
- Demonstra variações de desempenho positivas e negativas
  - Ex.: segmentação de clientes
- Gráfico de controle é uma das técnicas mais utilizadas

- ☐ Uso de outros métodos estatísticos
  - Análise correlacional
  - Análise de variância
  - Testes de hipóteses

Por que aconteceu?



#### **Preditiva**

☐ Foco em **predições de eventos** futuros

□ Extrai padrões encontrados em dados históricos

- ☐ Utilizado para diferentes aplicações
  - Detecção de fraude
  - Gerenciamento de risco
  - Fidelização de clientes



O que acontecerá?



#### **Preditiva**

- ☐ Requer uso de diversos métodos e ferramentas
  - Análises estatísticas
  - Técnicas de simulação
  - Mineração de dados
  - Aprendizado de máquina

Análise descritiva

Relatório climático

Análise preditiva

Previsão do tempo



#### **Prescritiva**

☐ Foco em prever ações futuras e possíveis consequências

 Sugere ações baseadas no conhecimento extraído dos dados

☐ Envolve regras de negócios, conhecimentos matemáticos, mineração de dados



Como fazer acontecer?

#### **Prescritiva**

☐ Segundo a consultora Gartner, somente 3% das empresas utilizam análise prescritiva

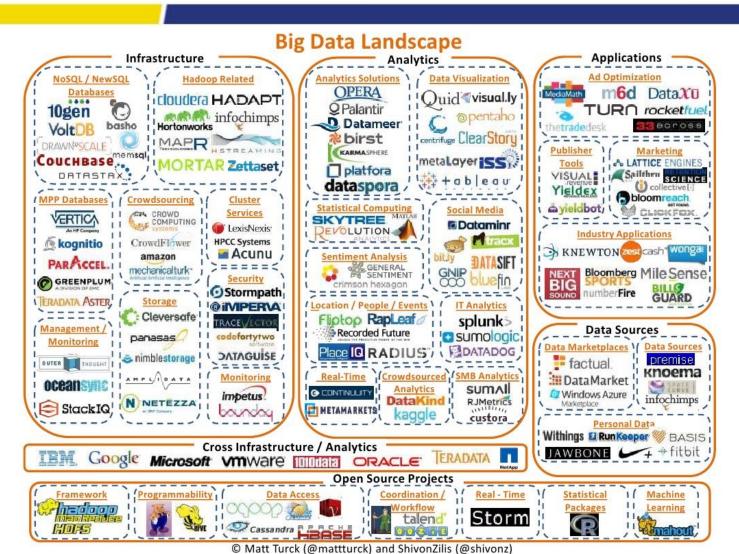
□ Necessário grandes quantidades de dados

- ☐ Exemplos:
  - Ofertas promocionais baseadas na segmentação do cliente
  - Veículos autônomos
  - Manutenção inteligente





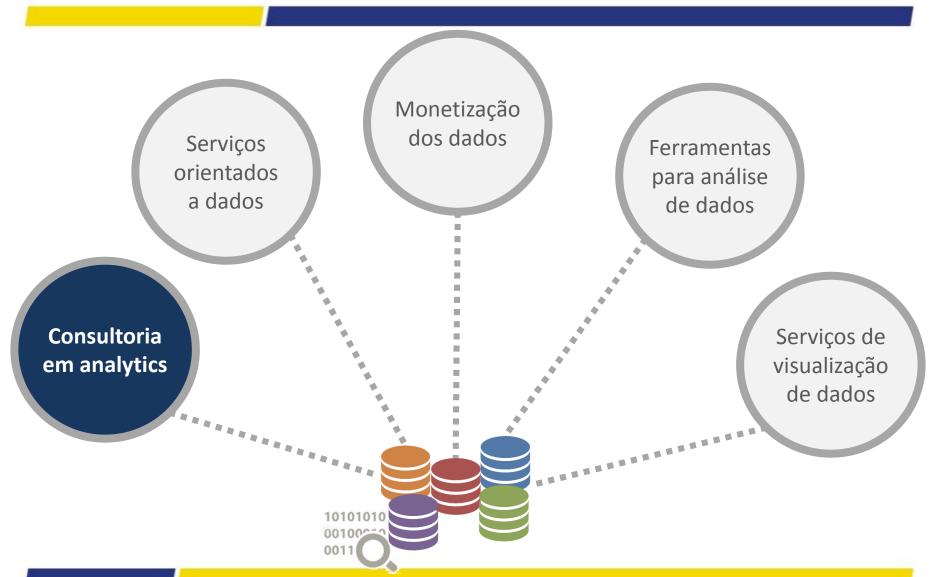




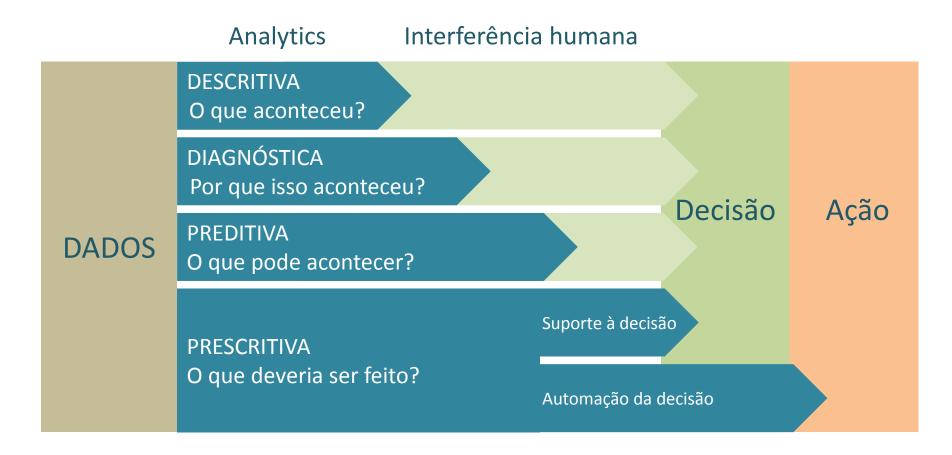


## Quais soluções posso gerar a partir de **Big Data Analytics**?









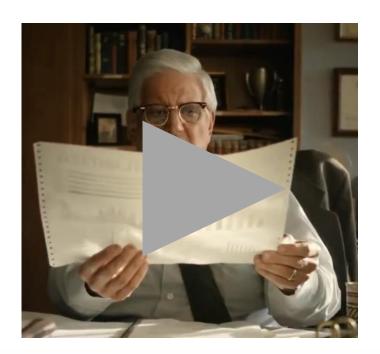


Empresas brasileiras que oferecem consultoria em analytics

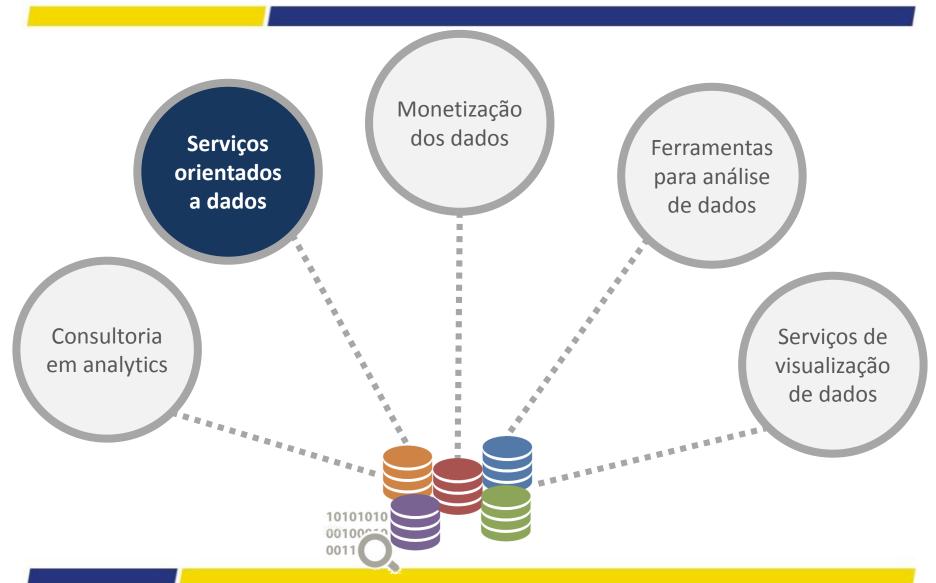




Big Data Analytics pode oferecer informações valiosas, porém, é necessário uma análise criteriosa das informações...



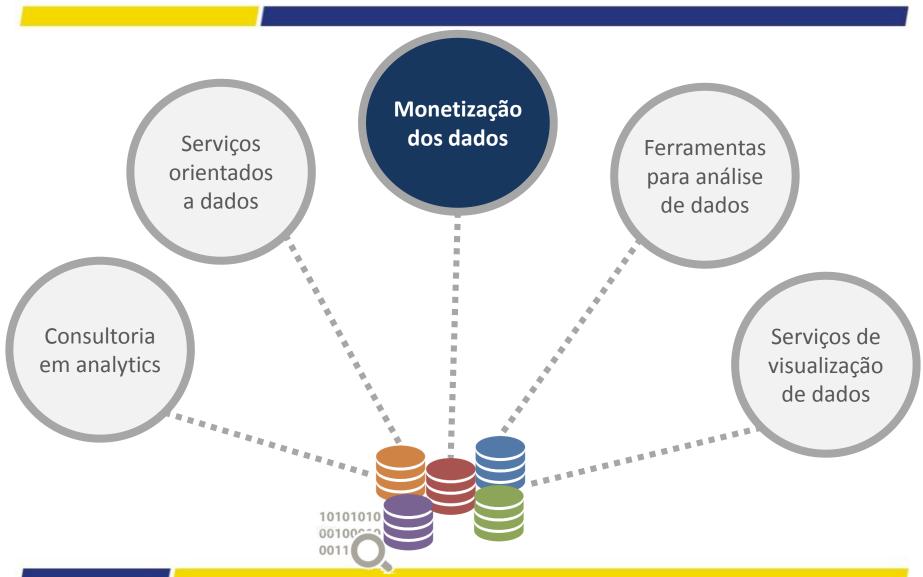














#### Monetização dos dados

Transformar ativos de informação em dinheiro, direta ou indiretamente, por meio de troca, comercialização ou venda direta

#### **Exemplos:**

- Dados de comportamento de usuários para campanhas de marketing
- Métricas de uso de transações de cartão de crédito para comerciantes
- Dados de sensores para seguradoras
- Dados genéticos para pesquisadores



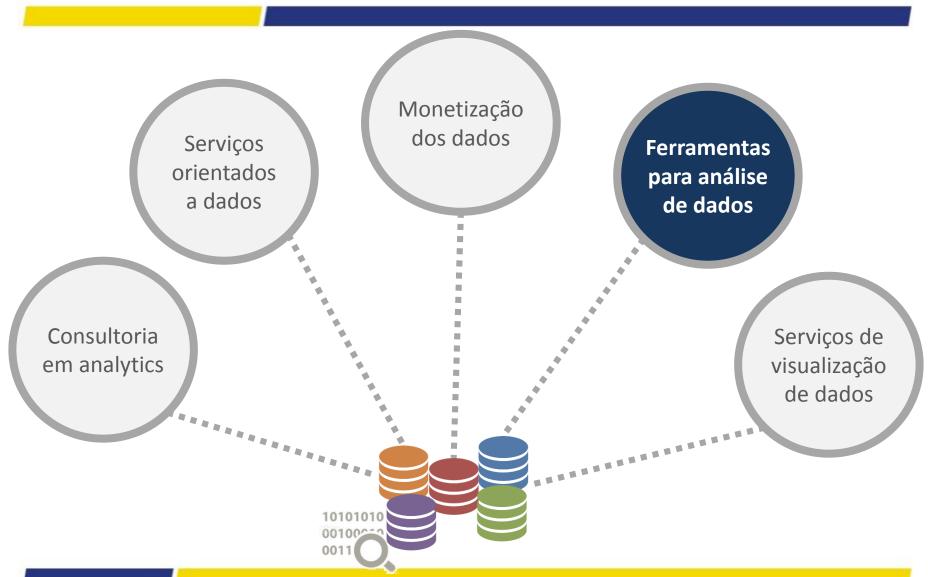


#### Monetização dos dados

Necessidade de avaliar questões relativas à **PRIVACIDADE DOS DADOS** 

- ☐ Quem é o dono dos dados?
- Quais dados podem ser capturados?
- ☐ Dados foram capturados com a anuência do usuário?







Processamento distribuído dos dados

Algoritmos de aprendizado de máquina

Processamento em tempo real

Construção de modelos

Tendência: soluções de Big Data como plataformas e serviços de computação em nuvem (PaaS/SaaS)



#### Ferramentas para análise de dados

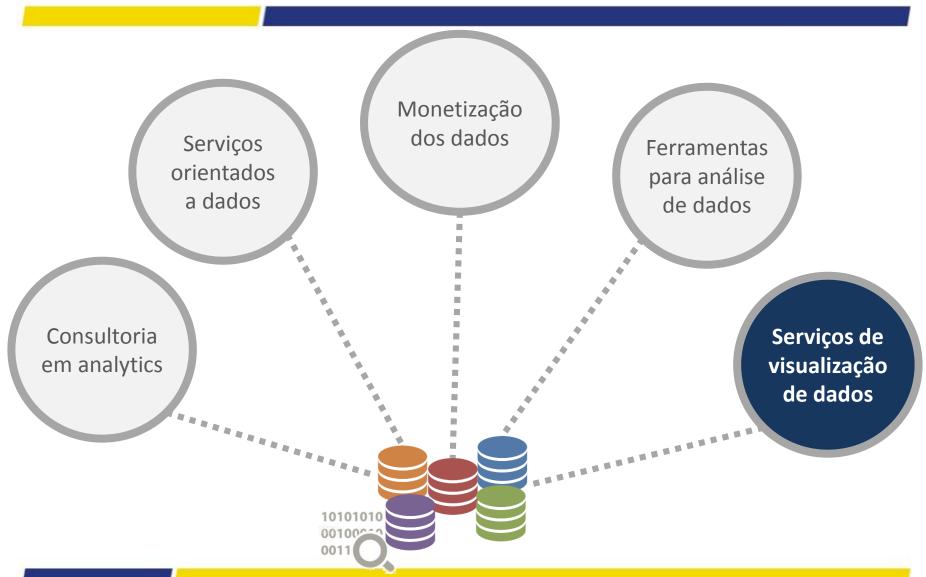
Exemplo: NextBio

Empresa líder na análise e agregação de dados de genomas

Oferece uma plataforma como serviço de nuvem para que pesquisadores acessarem e analisarem mais de 10.000 estudos sobre os genomas em uma base com mais de 100 TB de dados

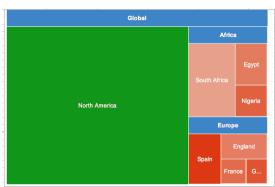


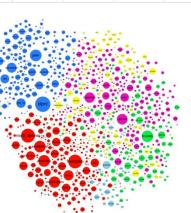


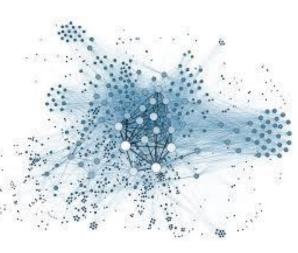


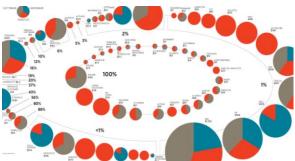


### Visualização de dados = soluções visuais para projetos orientados por dados











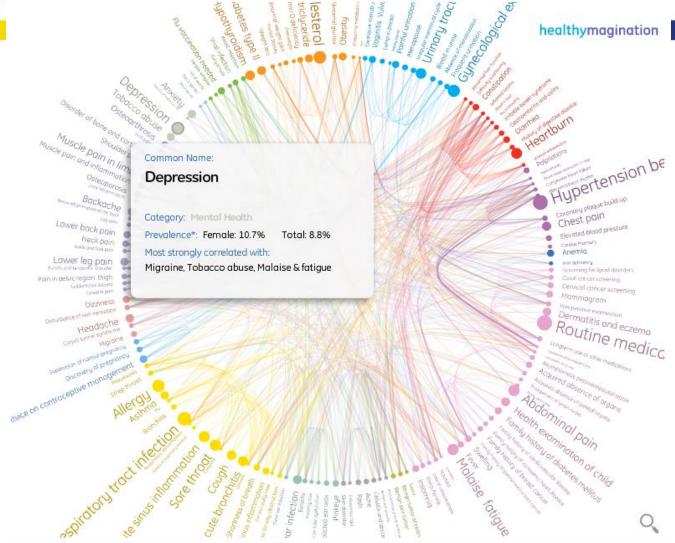


**Exemplo:** GE – visualizações do relacionamento entre sintomas

Base de dados: registros de 7.2 milhões de pacientes capturados nos equipamentos da GE

Apresenta categoria dos sintomas, como sintomas estão relacionados e prevalência entre homens e mulheres







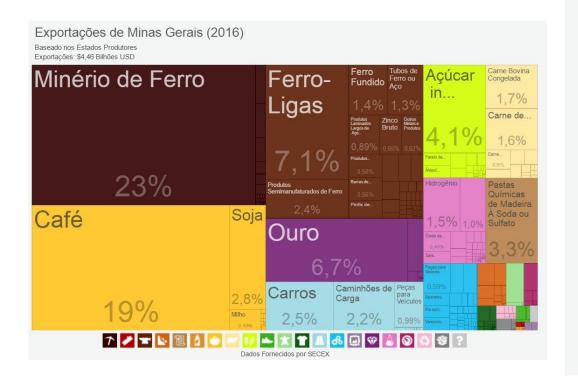
#### **DATAVIVA**

Plataforma aberta de pesquisa, que permite ao usuário acessar mais de 1 bilhão de visualizações com dados socioeconômicos dos mais de 5 mil municípios brasileiros.

Base de dados: Integração e análise de dados disponibilizados pelo Ministério do Trabalho e Previdência Social (MTPS), Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC) e Ministério da Educação (MEC).



#### **DATAVIVA**







# Mas quais ferramentas posso utilizar?























# **Software R**

# Criado por Ross Ihaka e Robert Gentleman nos anos 90

Universidade de Auckland, Nova Zelândia



Fonte: https://www.stat.auckland.ac.nz/~ihaka/downloads/the-r-project.pdf



# Como foi criado o R?

1. Criado inicialmente para testar ideias estatísticas

2. Utilizado posteriormente como ferramenta de ensino de cursos de estatística

 Adotado posteriormente por um grande número de usuários e desenvolvedores



## Características do R

 Divulgado ao público com publicação de um artigo no periódico "The Journal of Computational Statistics and Graphics" em 1996

# R: A Language for Data Analysis and Graphics

Ross IHAKA and Robert GENTLEMAN

In this article we discuss our experience designing and implementing a statistical computing language. In developing this new language, we sought to combine what we felt were useful features from two existing computer languages. We feel that the new language provides advantages in the areas of portability, computational efficiency, memory management, and scoping.

Key Words: Computer language; Statistical computing.



# R oferece suporte à diversas operações de análise de dados

- Classificação
- Agrupamento
- Análise de série temporal
- K-Means
- Modelos lineares e não lineares
- Muito mais...



# Exemplos de soluções disponíveis

Nome	Descrição						
arm	Pacote para criação de modelos lineares						
Igraph	Pacote para análise de redes. Utilizado para representar redes sociais						
lubridate	Pacote com diversas soluções para manipulação de datas						
reshape	Pacote para agregação de dados						
tm	Pacote para mineração de texto. Adequado para atuar com dados não estruturados						
XML	Pacote para manipulação de arquivos XML e HTML						



# R é aplicado em diferentes áreas

- Finanças
- Ciências sociais
- Genética
- Medicina
- Redes sociais

**—** ...



# R é utilizado por diversas empresas

















The New York Times



# Benefícios do R

Plataforma única para análise de dados

# Ambiente R

Manipulação de dados

Análise de dados

Visualização de dados



# Benefícios do R

Permite integração com diversos bancos de dados



Arquivos de texto



Banco de dados relacionais



Banco de dados distribuídos

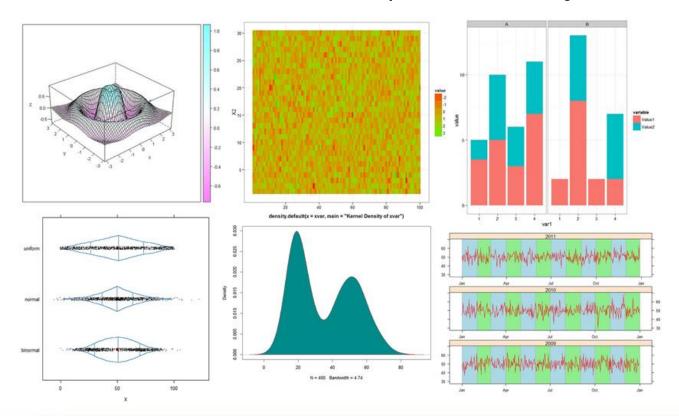


Streaming de dados



# Benefícios do R

Grande variedade de recursos para visualização de dados



# R na prática

# Atividade

O uso de análise preditiva em sistemas de recomendação



A análise preditiva pode ser aplicada de 3 diferentes formas no marketing:

Clustering models (segmentação)

Propensity models (predição)

Collaborative filtering (recomendação)



# Quantas vezes você verificou/comprou itens recomendados por sites de e-commerce?



You viewed



Principles of Data Integration

AnHai Doan, Zachary Ives

★★★☆☆ (2) Hardcover: \$57.83

View or edit your browsing history

Customers who viewed this also viewed



Managing Data in Motion: Data...

> April Reeve

**本本本本本** (2)

Paperback: \$44.28 Kindle Edition: \$26.71



Connecting the Data: Data Integration... Angelo R. Bobak

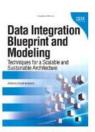
**★★★★** (1) Paperback: \$30.92

Petrolition and Petrolitic Annual Control of Petrolitic Science and Petrolitic Science and

Data Matching: Concepts and...

> Peter Christen

Hardcover: \$48.97 Kindle Edition: \$27.53



Data Integration Blueprint and...

> Anthony Giordano

★★★☆ (7) Hardcover



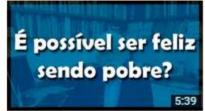
# Quantos vídeos você já assistiu que foram recomendados pelo YouTube?



Como parar de reclamar por Arata Academy 46.826 visualizações • 3 semanas atrás



Steven Tyler - Cryin'
(Acoustic)
por Ismael Quispe
791.667 visualizações • 1 ano atrás



pobre? - Flávio Gikovate por Flávio Gikovate 14.361 visualizações • 1 ano atrás

É possível ser feliz sendo



O Teatro Mágico - Você me bagunça (Legendado) por Monisa Segundo 1.608.098 visualizações • 4 anos atrás Mostrar mais



# Você já encontrou indicações de vagas de emprego que fosse do seu interesse?



Vagas que podem ser de seu interesse









# Por que a recomendação é tão importante atualmente?



30 milhões de músicas 1.23 bi usuários ativos/dia

Facebook

**80 milhões**de fotos
por dia

Instagram

+14 mil filmes

Netflix

4 bilhões Visualizações por dia

Youtube



- ☐ Sistemas de recomendação: área de analytics para gerar recomendações personalizadas a um usuário
- ☐ Filtragem colaborativa: tipo de sistema de recomendação para sugerir itens/produtos/serviços a partir de gostos similares de outros usuários





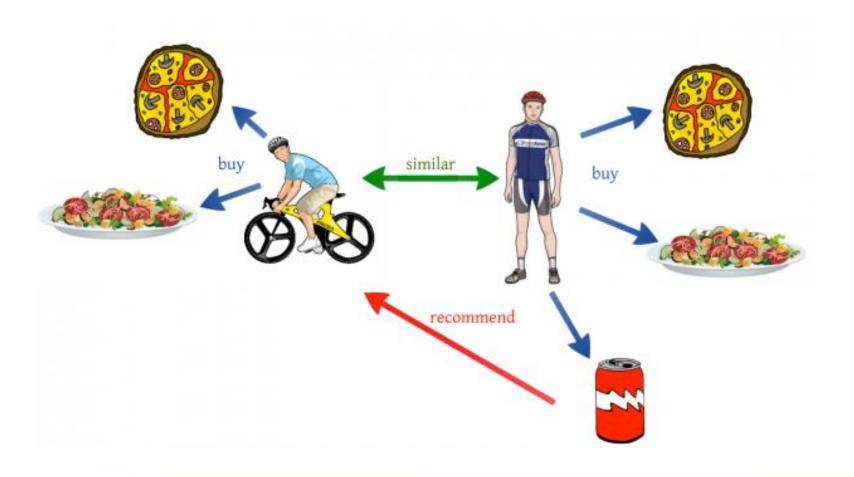
# ☐ Filtragem colaborativa item a item

 Similaridade entre itens i & j é computada isolando os usuários e aplicando uma técnica de cálculo de similaridade

Recomenda os top-k-vizinhos mais próximos

Recomendação é composta por itens que os usuários gostaram







□ Recomendação baseada no histórico de consumo de músicas dos usuários





# ■ Exemplo



Legião Urbana

16.447.186 execuções (235.570 ouvintes)

Parecido com: Renato Russo, Cazuza, Capital Inicial, Os Paralamas Do Sucesso, Titãs

- rock



#### Pitty

8.809.898 execuções (251.256 ouvintes)

Parecido com: Agridoce, Marjorie Estiano, Megh Stock, NX Zero, Luxúria

• rock



#### Caetano Veloso

12.782.155 execuções (440.153 ouvintes)

Parecido com: Gal Costa, Gilberto Gil, Maria Bethânia, Chico Buarque, Tom Zé

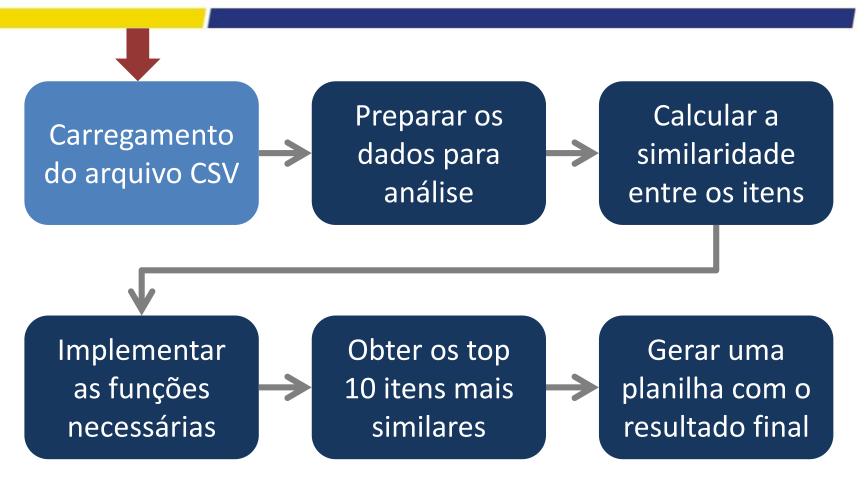
• mpb



- ☐ Base de dados: tabela usuário/artista com as preferências dos usuários da Last.fm
  - Cada linha representa um usuário (Total: 1257)
  - Cada coluna representa um artista (Total: 285)
  - Conteúdo da matriz indica as preferências de cada usuário
     User, abba, ac.dc, adam green, aerosmith, afi, air

1,	0,	0,	0,	0,	0, 0
33,	0,	0,	1,	0,	0, 0
42,	0,	0,	0,	0,	0, 0
51,	0,	0,	0,	0,	0, 0
62,	0,	0,	0,	0,	0, 0
<b>75.</b>	0.	0.	0.	0,	0, 0







☐ Configurando o diretório de trabalho

## Console

> setwd("C:/Users/<nome\_usuario>/Documents/Rnapratica")



☐ Verificar o diretório de trabalho atual

## Console

> getwd()

[1] "C:/Users/Big Data/Documents/Rnapratica"

☐ Carregando o arquivo do tipo CSV no R

## Console

> base <- read.csv(file="lastfm.csv")</pre>



## ☐ Visualizando os dados

Obtendo os resultados das primeiras 7 linhas e das colunas 1, 3, 4, 5,
6, 7 e 8



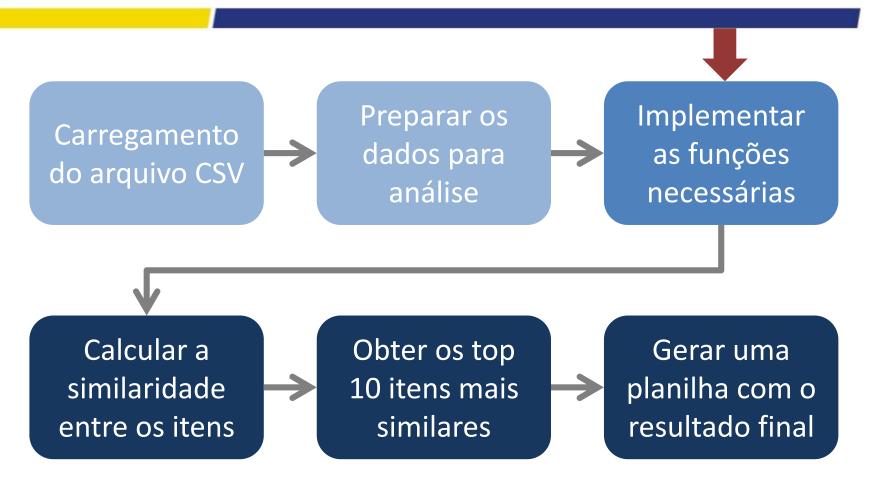


☐ Retirando a coluna de usuários

## Console

> base.msc <- (base[,!(names(base) %in% c("user"))])</pre>



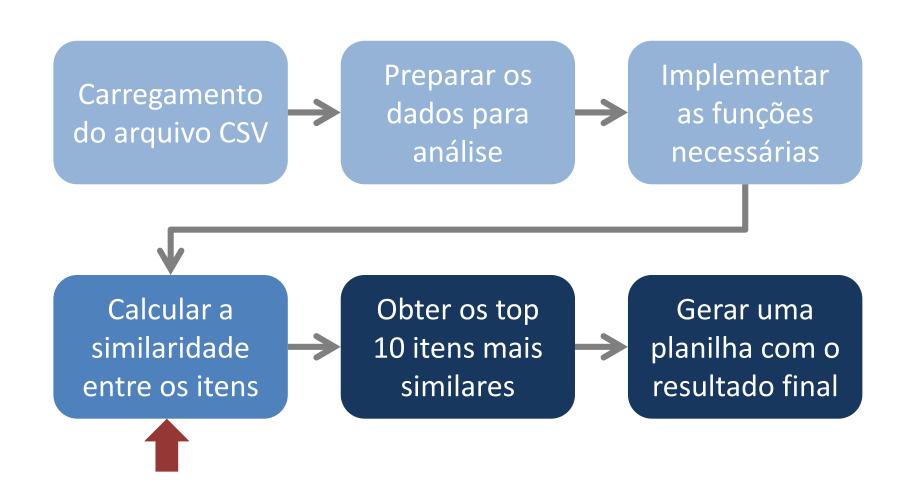


- ☐ Função para cálculo do cosseno entre dois vetores
  - Oferece uma medida de quão similares são dois itens

#### Console

```
> getCosseno<- function(x,y)
{;
  cosseno <- sum(x*y) / (sqrt(sum(x*x)) * sqrt(sum(y*y)));
  return(cosseno);
};</pre>
```







☐ Criar uma matriz para armazenar as medidas de similaridade entre os itens

#### Console

> base.msc.cosseno <- matrix(NA, nrow=ncol(base.msc),ncol=ncol(base.msc),dimnames=list(colnames(base.msc),colnames(base.msc)))



☐ Criar uma matriz para armazenar as medidas de similaridade entre os itens

			_	aerosmith 🗦	afi 🗦	air 🖁	alanis.morissette 🗦	alexisonfire 🗦		cia
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	,
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
NA	NA	NA	NA	NA	NA	NA	NA	NA		
	NA N	NA         NA           NA         NA	NA         NA         NA           NA         NA         NA	NA         NA         NA           NA         NA         NA	NA         NA         NA         NA         NA           NA         NA         NA         NA         NA	NA         NA<	NA         NA<	NA         NA<	NA         NA<	NA         NA<

howing 1 to 14 of 285 entries



☐ Calcular o cosseno de similaridade entre todas as colunas em uma matriz

```
Console
```

```
> for(i in 1:ncol(base.msc)) {;
    for(j in 1:ncol(base.msc)) {;
       base.msc.cosseno[i,j] <-
    getCosseno(as.matrix(base.msc[i]),as.matrix(base.msc[j]))
    };
};</pre>
```



☐ Calcular o cosseno de similaridade entre todas as colunas em uma matriz

	a.perfect.circle ‡	abba <sup>‡</sup>	ac.dc ‡	adam.green 🖣	aerosmith †	afi 🗦	air ‡	alanis.morissette $^{\circ}$	a
a.perfect.circle	1.00000000	0.00000000	0.01791723	0.05155393	0.06277648	0.00000000	0.05175492	0.06071767	^
abba	0.00000000	1.00000000	0.05227877	0.02507061	0.06105625	0.00000000	0.01677890	0.02952693	
ac.dc	0.01791723	0.05227877	1.00000000	0.11315371	0.17715300	0.06789420	0.07572991	0.03807625	
adam.green	0.05155393	0.02507061	0.11315371	1.00000000	0.05663655	0.00000000	0.09338592	0.00000000	
aerosmith	0.06277648	0.06105625	0.17715300	0.05663655	1.00000000	0.00000000	0.11371471	0.10005560	
afi	0.00000000	0.00000000	0.06789420	0.00000000	0.00000000	1.00000000	0.00000000	0.05751973	
air	0.05175492	0.01677890	0.07572991	0.09338592	0.11371471	0.00000000	1.00000000	0.09165445	
alanis.morissette	0.06071767	0.02952693	0.03807625	0.00000000	0.10005560	0.05751973	0.09165445	1.00000000	
alexisonfire	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.08362420	0.03553345	0.00000000	
alicia.keys	0.00000000	0.00000000	0.08833316	0.02541643	0.06189845	0.05337605	0.05103104	0.17960530	
all.that.remains	0.13012001	0.00000000	0.02039967	0.00000000	0.00000000	0.03081668	0.00000000	0.00000000	
amon.amarth	0.04293388	0.00000000	0.10769589	0.01936734	0.07074999	0.02033630	0.01296190	0.02280990	
amv.macdonald	0 00000000	0 11742785	0.01514283	0 00000000	0 00000000	0.06862635	0.01458030	0 10263160	`



☐ Converter a matriz de similaridade em um data frame

#### Console

> base.msc.cosseno <- as.data.frame(base.msc.cosseno)



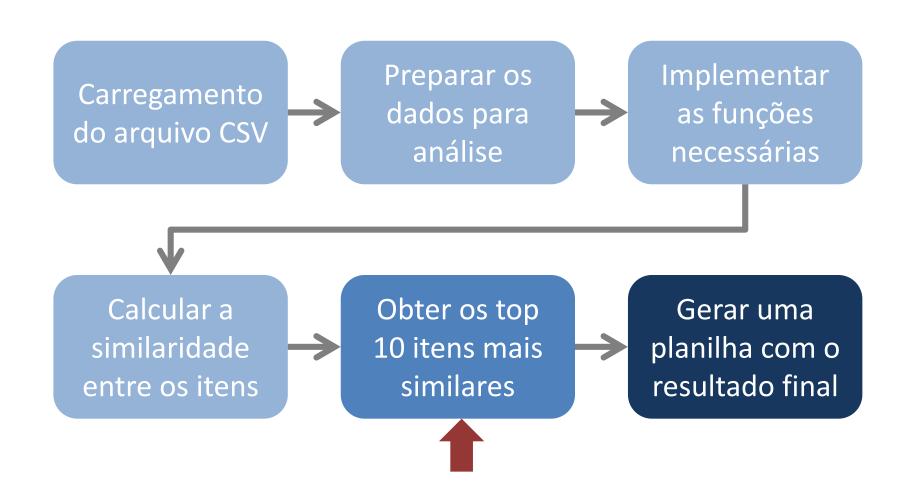
☐ Visualizar o conteúdo do data frame

#### Console

## > head(base.msc.cosseno[,c(1,3:5)])

	a.perfect.circle	ac.dc	adam.green	aerosmith
a.perfect.circle	1.0000000	0.01791723	0.05155393	0.06277648
abba	0.00000000	0.05227877	0.02507061	0.06105625
ac.dc	0.01791723	1.00000000	0.11315371	0.17715300
adam.green	0.05155393	0.11315371	1.00000000	0.05663655
aerosmith	0.06277648	0.17715300	0.05663655	1.00000000
afi	0.00000000	0.06789420	0.00000000	0.00000000







☐ Criar uma nova matriz para armazenar os top-10 vizinhos mais próximos de cada item

#### Console

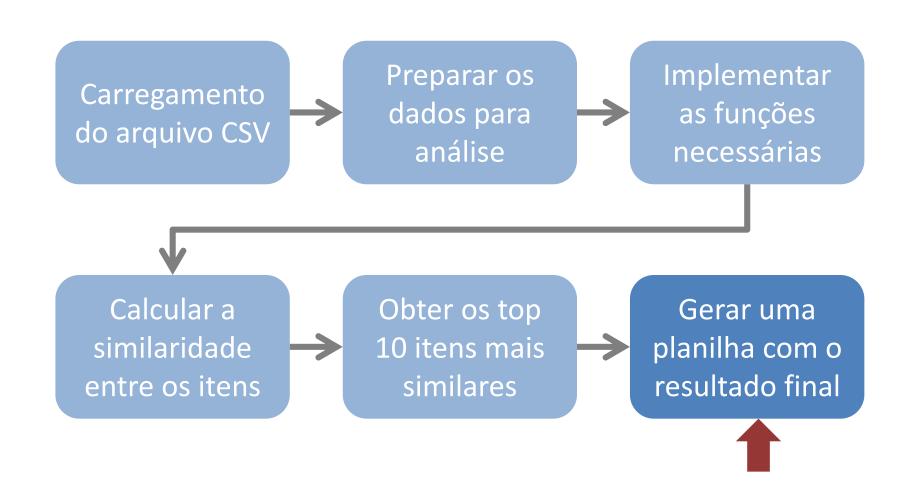
> base.vizinhos <- matrix(NA, nrow=ncol(base.msc.cosseno),ncol=11,dimnames=list(colname s(base.msc.cosseno)))

Ordenar a matriz em ordem decrescente de similaridade

#### Console

```
> for(i in 1:ncol(base.msc))
{;
  base.vizinhos[i,] <-
(t(head(n=11,rownames(base.msc.cosseno[order(base.msc.cosseno[,i],decreasing=TRUE),][i]))));
};</pre>
```





☐ Gravar o resultado em um arquivo csv

#### Console

> write.csv(file="lastfm-top10.csv",x=base.vizinhos[,1])



☐ Abrir o arquivo e visualizar o resultado

a.perfect.circle "a.perfect.circle" "abba" abba "ac.dc" ac.dc adam.green "adam.green" "aerosmith" aerosmith

"tool" "madonna" "red.hot.chilli.peppers" "metallica" "the.libertines" "u2"

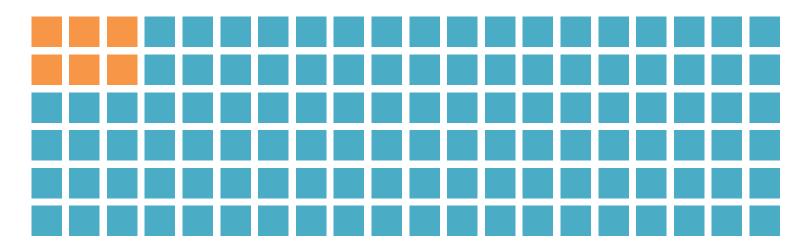
"dredg" "robbie.williams" "the.strokes" "led.zeppelin"



## CONSIDERAÇÕES FINAIS



Big Data está apenas em seu início



A IDC estima que somente 0.5% dos dados globais são analisados



## Ainda há muita oportunidade de inovação!

EDUCAÇÃO MARKETING

**ASTRONOMIA** 

**ESPORTES** 

FINANÇAS VAREJO

**ENERGIA** 

RH

**SEGUROS** 

SETOR PÚBLICO

**BIOLOGIA** 

SAÚDE

**TURISMO** 

**AGRICULTURA** 

**TRÂNSITO** 

MEIO AMBIENTE

**AUTOMOTIVO** 

**ALIMENTAÇÃO** 

ENTRETENIMENTO

**MEDICINA** 

LOGÍSTICA

**MANUFATURA** 

CONSTRUÇÃO CIVIL

**TELECOMUNICAÇÃO** 

**TECNOLOGIA** 



# Um BIG Obrigada!

### Meus dados ©

Email: <a href="mailto:romarquesone@gmail.com">romarquesone@gmail.com</a>

LinkedIn: <a href="https://br.linkedin.com/in/rosangelafpm">https://br.linkedin.com/in/rosangelafpm</a>

Twitter: hadoop\_girl

Lattes: <a href="http://lattes.cnpq.br/1095618344499722">http://lattes.cnpq.br/1095618344499722</a>

Github: rosangelapereira