



Data Science Academy

# Big Data Analytics com R e Microsoft Azure Machine Learning Módulo 4



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)

De onde importamos os dados para o R?



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)

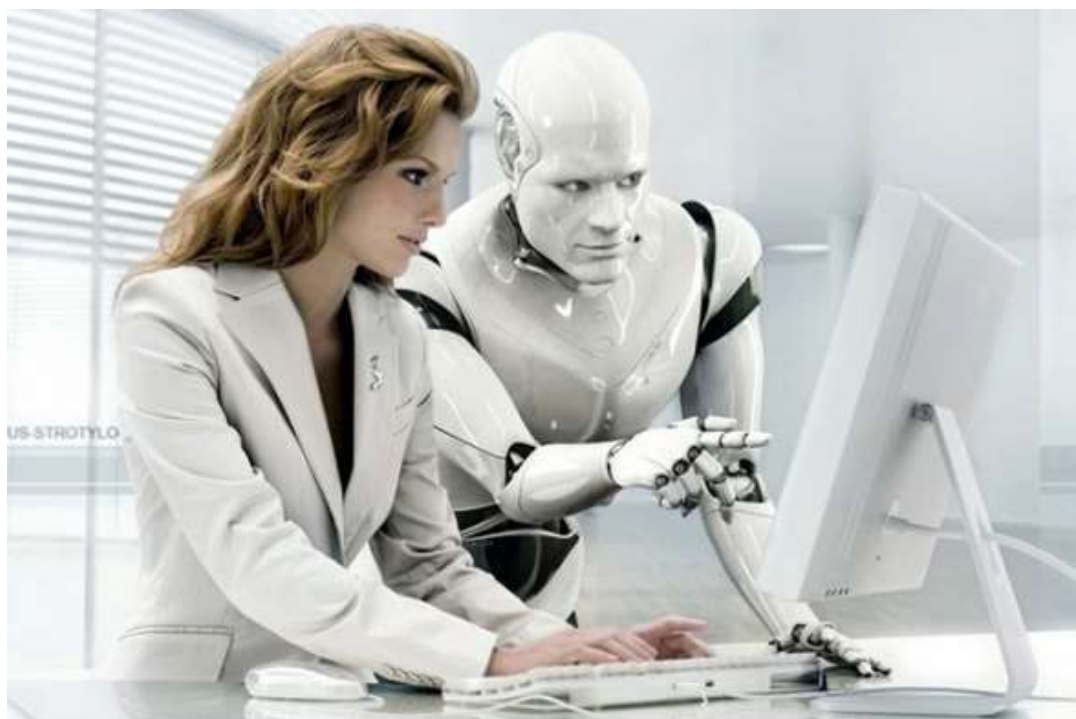


## De onde importamos os dados para o R?

- Arquivos Texto - flat files (txt, csv)
- Arquivos Excel (xls, xlsx)
- Bancos de Dados (Oracle, SQL Server, MySQL, PostgreSQL, SQLite)
- Softwares Estatísticos (SAS, SPSS, Stata)
- Dados da Internet (Web Crawling)



Data Science Academy



A função dos  
algoritmos é  
automatizar os  
processos de análise



Data Science Academy




Aqui estão alguns dos padrões mais comumente observados

- Os cabeçalhos das colunas são valores e não nomes de variáveis
- Diversas variáveis são armazenadas em uma coluna
- As variáveis são armazenados em ambas as linhas e colunas
- Vários tipos de unidade experimental armazenados na mesma tabela
- Um tipo de unidade experimental armazenado em várias tabelas



Data Science Academy



Pacote utils	Pacote readr
<code>read.table()</code>	<code>read_delim()</code>
<code>read.csv()</code>	<code>read_csv()</code>
<code>read.delim()</code>	<code>read_tsv()</code>

Pacote data.table

`fread()`



Data Science Academy



## Pacote utils



O pacote utils, que é automaticamente carregado na sua sessão R, pode importar arquivos simples em diferentes formas, através das funções:



Data Science Academy

# Pacote utils

<code>read.csv</code>	Para valores separados por vírgula e ponto como separador decimal
<code>read.csv2</code>	Para valores separados por ponto e vírgula e vírgula como separador decimal
<code>read.delim</code>	Para valores separados por tab e ponto como separador decimal
<code>read.delim2</code>	Para valores separados por tab e vírgula e vírgula como separador decimal
<code>read.fwf</code>	Para valores com número exato de bytes por coluna

`read.table()`



Data Science Academy





Pacote utils

`read.table()`

Muito útil quando se está fazendo a leitura de arquivos ASCII, que contém dados em formato retangular



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Pacote utils

`read.table()`

```
read.table("arquivo.txt", header = TRUE, sep = ",", stringsAsFactors = FALSE)  
read.table("arquivo.txt", header = TRUE, sep = "\t", stringsAsFactors = FALSE)
```



Data Science Academy



Pacote utils

`read.csv()`

```
read.csv("arquivo.csv", stringsAsFactors = FALSE)
```

```
read.csv2("arquivo.csv", sep = ";", dec = ",", stringsAsFactors = FALSE)
```



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Pacote utils

`read.delim()`

`read.delim("arquivo.txt")`  
`read.delim2("arquivo.txt")`



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)

## Parâmetros

header  
col.names  
na.string  
colClasses  
sep  
stringsAsFactors



Data Science Academy



# Pacote readr

Lançado em Abril/2015 pelos desenvolvedores do RStudio

```
install.packages("readr")
```

```
read_table()
```

```
read_csv ()
```

```
read_delim ()
```



Data Science Academy



Pacote readr

## Pacote readr

```
arq1 <- read_table("bigdatafile.txt", col_names = c("DAY","MONTH","YEAR","TEMP"))
```

```
arq2 <- read.table("bigdatafile.txt", col.names = c("DAY","MONTH","YEAR","TEMP"))
```



Data Science Academy



# Manipulação de Arquivos Excel



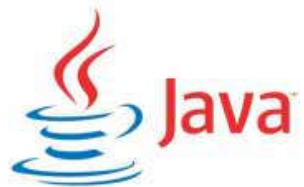
Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# Pacote XLConnect

loadWorkbook()  
getSheets()  
readWorksheet()  
createsheet()  
writeWorksheet()



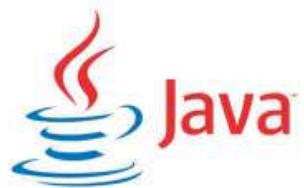
Data Science Academy



## Pacote xlsx

`read.xlsx(file, sheetIndex, header=TRUE, colClasses=NA)`

`read.xlsx2(file, sheetIndex, header=TRUE, colClasses="character")`



Data Science Academy



Pacote readxl

`read_excel()`  
`Excel_sheets()`



Data Science Academy



Pacote gdata

`read.xls()`



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



## Atenção aos Detalhes

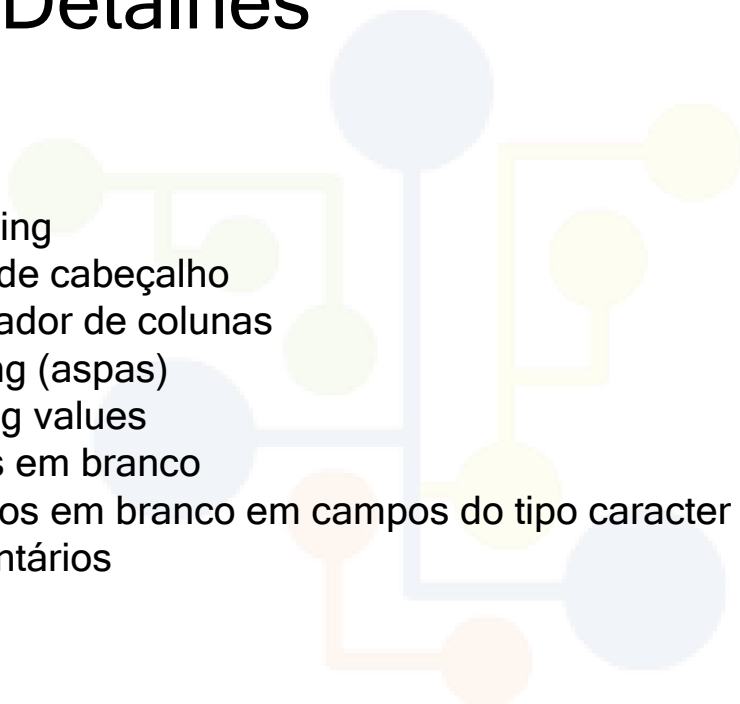
- Em seus arquivos, evite espaços em branco e números como título para as colunas
- Normalmente, a primeira linha de cada arquivo é o cabeçalho, a lista de nomes para cada coluna
- Para a concatenação de palavras, use . ou \_
- Use nomes curtos como título de coluna
- Evite o uso de caracteres especiais
- Dados NA podem existir no seu conjunto de dados e isso será tratado no processo de limpeza



Data Science Academy



# Atenção aos Detalhes

- 
- Encoding
  - Linha de cabeçalho
  - Separador de colunas
  - Quoting (aspas)
  - Missing values
  - Linhas em branco
  - Espaços em branco em campos do tipo caracter
  - Comentários



Data Science Academy



## Outros Pacotes para Importação de Arquivos:

- Pacote **rjson** - Leitura de arquivos JSON para o R
- Pacote **XML** - Leitura de arquivos xml
- Pacote **httr** - Leitura de páginas html para o R
- Pacote **Rcurl** - Web Crawling (Capítulo 5)
- Pacote **foreign** - Leitura de arquivos do SPSS, SAS (Capítulo 5)
- Pacote **sas7bdat** - Leitura de arquivos SAS (Capítulo 5)



Data Science Academy



# data.table

Fonece um rápido processo de carga de dados, pois as funções reconhecem automaticamente os parâmetros dos arquivos e decidem a melhor forma de carga

fread()



Data Science Academy



## Resumindo:

### Manipulação de Arquivos txt e csv

Package utils      `read.table()`  
                     `read.csv()`  
                     `read.delim()`

Package readr      `read_table()`  
                     `read_csv()`  
                     `read_delim()`

Package data.table      `fread()`

### Manipulação de Arquivos excel

XLConnect  
xlsx  
readxl  
gdata  
r2excel



Data Science Academy



E como o R se conecta aos SGBD's?



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



## Bancos de Dados e Pacotes R

Banco de Dados	Pacote R
Oracle	ROracle
Microsoft SQL Server	RSQLServer
PostgreSQL	RPostgreSQL
MySQL	RMySQL
SQLite	RSQLite
MongoDB	RMongo
Conexão ODBC	RODBC

Data Science Academy



# Bancos de Dados e Pacotes R



Banco de Dados	Pacote R
Conexão ODBC	RODBC



Data Science Academy



Quais os passos necessários para conectar em um banco de dados usando R:

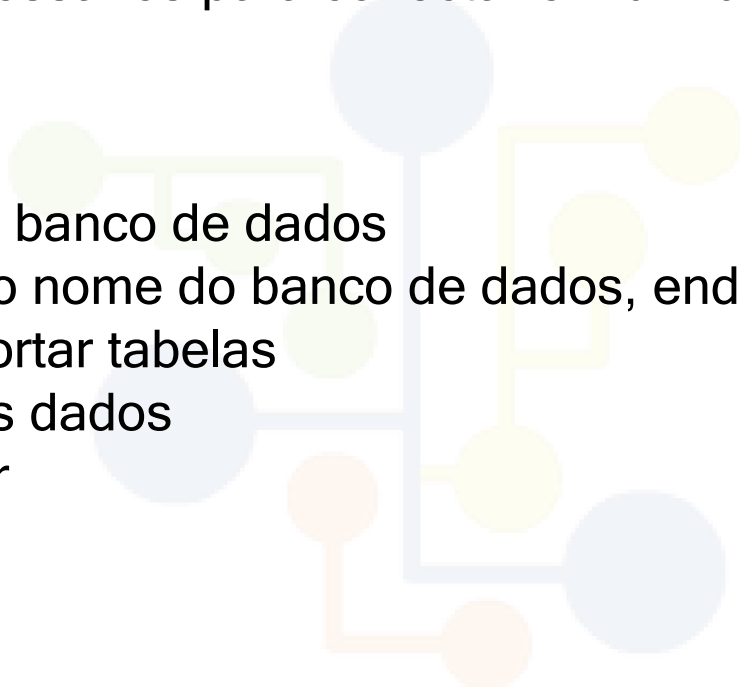
- Conectar ao banco de dados → `DBI.dbConnect ()`
- 



Data Science Academy



Quais os passos necessários para conectar em um banco de dados usando R:

- Conectar ao banco de dados
  - Determinar o nome do banco de dados, endereço, porta, usuário e senha
  - Listar e importar tabelas
  - Manipular os dados
  - Desconectar
- 



Data Science Academy




# Bancos de Dados NoSQL (Not Only SQL)



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



NoSQL é uma tecnologia de banco de dados projetada para suportar os requisitos de aplicações em nuvem e arquitetado para superar em escala e desempenho as limitações de bancos de dados relacionais (RDBMS)



Data Science Academy



Os principais Bancos de Dados NoSQL são:

Graph	Neo4J
	FlockDB
	GraphDB
	ArangoDB

Key-value	Oracle NoSQL DB
	MemcacheDB
	Redis
	Voldemort


Document	MongoDB
	CouchDB
	RavenDB
	Terrastore

Column	HBase
	Cassandra*
	Hypertable
	Accumulo




Data Science Academy







MongoDB	RDBMS
Database	Database
Collection	Tabela
Document	Linha/Tupla
Field	Coluna
Embedded Documents	Join de Tabelas
Primary Key	Primary Key



Data Science Academy

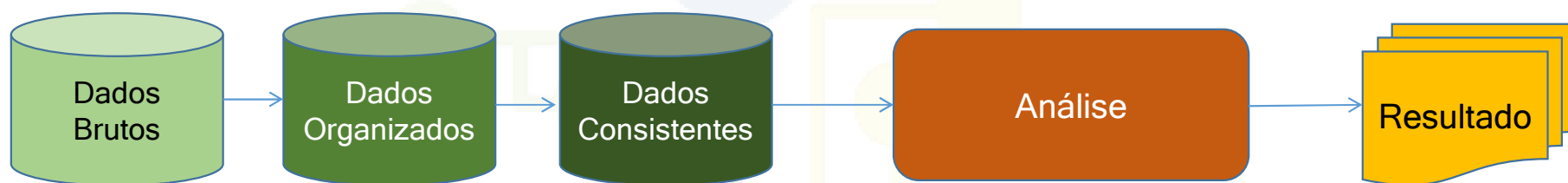


# Data Wrangling (Manipulação de Dados)



Data Science Academy

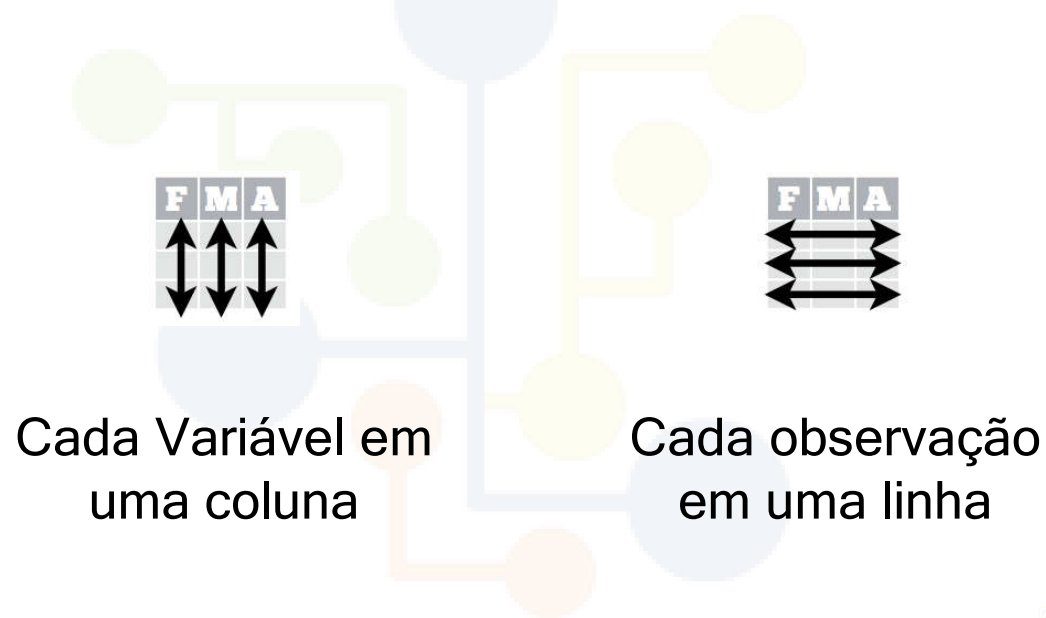
[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



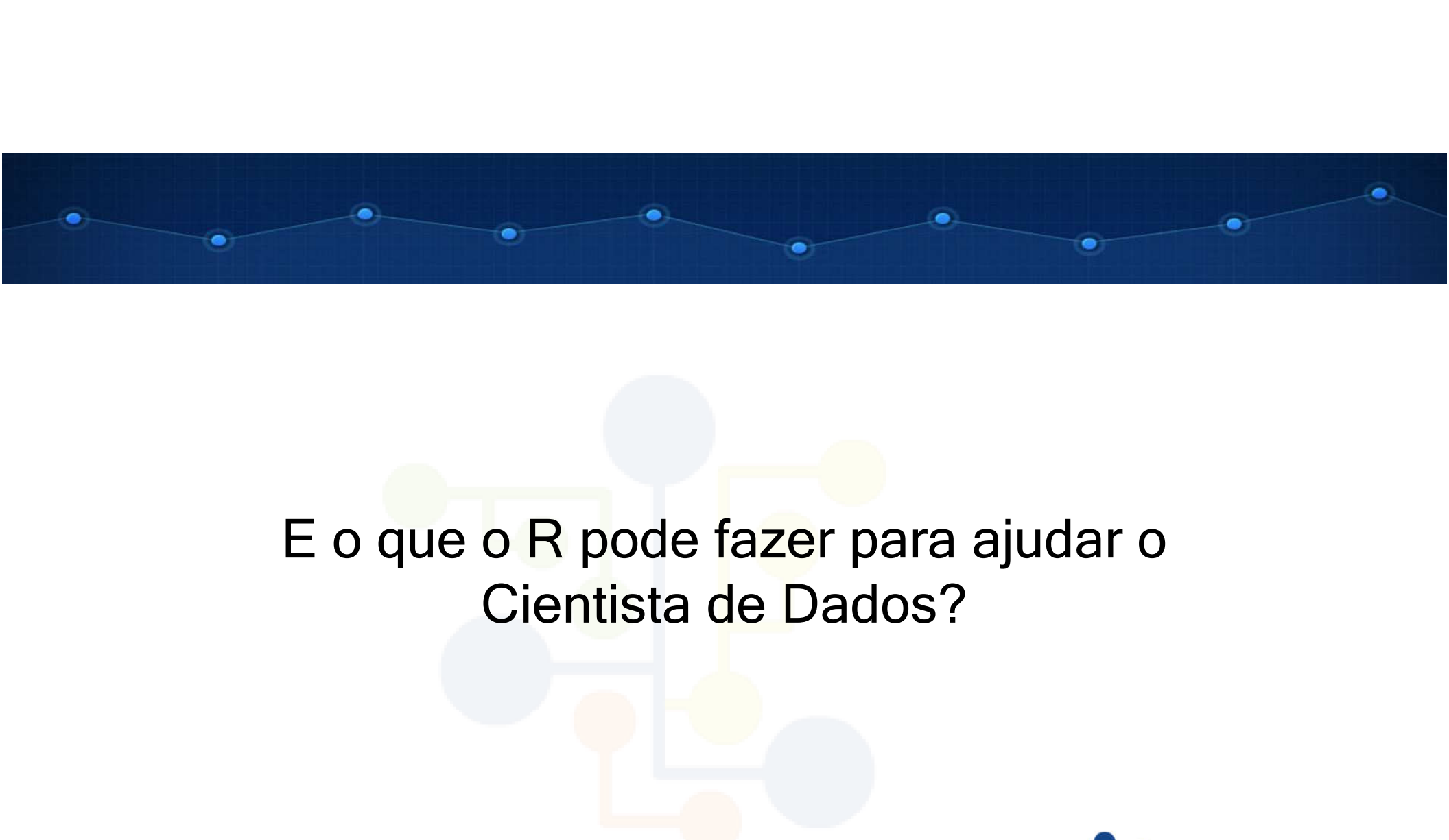
Data Science Academy



## Qual o objetivo do Data Wrangling?



Data Science Academy



E o que o R pode fazer para ajudar o Cientista de Dados?



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



## dplyr

- `select()`
- `filter()`
- `group_by()`
- `summarise()`
- `arrange()`
- `join()`
- `mutate()`

## tidyr

- `gather()`
- `spread()`
- `separate()`
- `unite()`



Data Science Academy





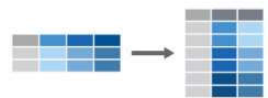
tidyr

# Remodelagem de Dados



Data Science Academy

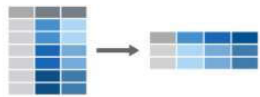
# Funções tidyr



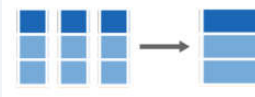
gather()



separate()



spread()

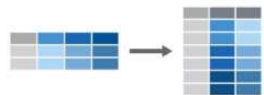


unite()



Data Science Academy

## Funções tidyr



`gather()`

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

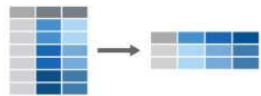
`gather()`

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000



Data Science Academy

## Funções tidyr

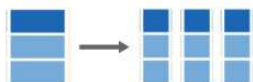


`spread()`

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

## Funções tidyr



`separate()`

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

`separate()`

storm	wind	pressure	year	month	day
Alberto	110	1007	2000	08	12
Alex	45	1009	1998	07	30
Allison	65	1005	1995	06	04
Ana	40	1013	1997	07	1
Arlene	50	1010	1999	06	13
Arthur	45	1010	1996	06	21



Data Science Academy

## Funções tidyr



`unite()`

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

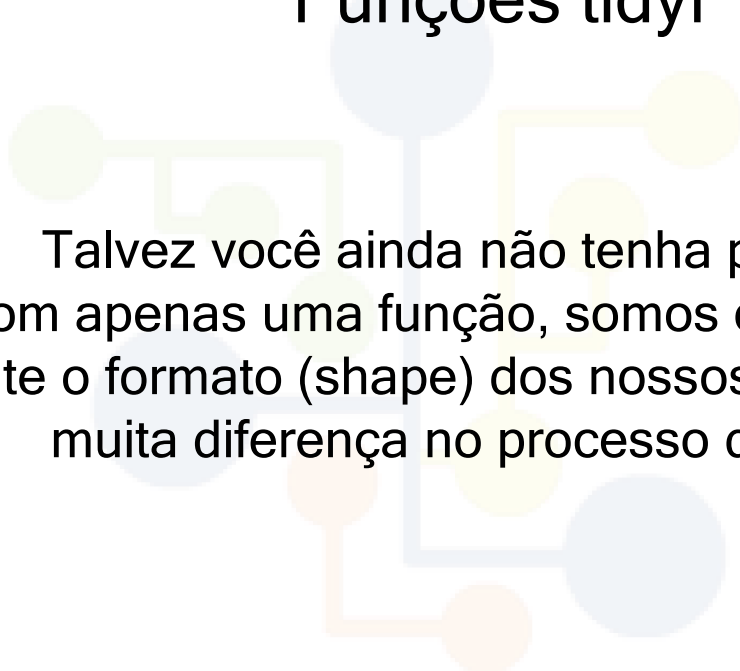
storm	wind	pressure	year	month	day
Alberto	110	1007	2000	08	12
Alex	45	1009	1998	07	30
Allison	65	1005	1995	06	04
Ana	40	1013	1997	07	1
Arlene	50	1010	1999	06	13
Arthur	45	1010	1996	06	21



Data Science Academy



## Funções tidyr



Talvez você ainda não tenha percebido.  
Mas com apenas uma função, somos capazes de mudar  
completamente o formato (shape) dos nossos dados e isso pode fazer  
muita diferença no processo de análise



Data Science Academy



dplyr

# Transformação de Dados



Data Science Academy



## Funções dplyr



`select()`

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

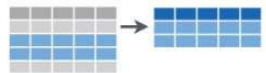


wind	pressure	date
110	1007	2000-08-12
45	1009	1998-07-30
65	1005	1995-06-04
40	1013	1997-07-01
50	1010	1999-06-13
45	1010	1996-06-21



Data Science Academy

## Funções dplyr



`filter()`

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21



storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Allison	65	1005	1995-06-04



Data Science Academy

## Funções dplyr



`group_by()`

country	year	sex	cases
Afghanistan	1999	female	1
Afghanistan	1999	male	1
Afghanistan	2000	female	1
Afghanistan	2000	male	1
Brazil	1999	female	2
Brazil	1999	male	2
Brazil	2000	female	2
Brazil	2000	male	2
China	1999	female	3
China	1999	male	3
China	2000	female	3
China	2000	male	3

country	year	sex	cases
Afghanistan	1999	female	1
Afghanistan	1999	male	1
Afghanistan	2000	female	1
Afghanistan	2000	male	1
Brazil	1999	female	2
Brazil	1999	male	2
Brazil	2000	female	2
Brazil	2000	male	2
China	1999	female	3
China	1999	male	3
China	2000	female	3
China	2000	male	3



Data Science Academy

## Funções dplyr

summary  
function

summarise()

head(iris)

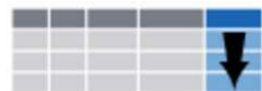
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Species	Mean	SD	n
setosa	5.006	0.352	50
versicolor	5.936	0.516	50
virginica	6.588	0.636	50



Data Science Academy

## Funções dplyr



`arrange()`

`head(iris)`

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa



Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
7.9	3.8	6.4	2.0	virginica
7.7	3.8	6.7	2.2	virginica
7.7	2.6	6.9	2.3	virginica
7.7	2.8	6.7	2.0	virginica
7.7	3.0	6.1	2.3	virginica
7.6	3.0	6.6	2.1	virginica



Data Science Academy

# Funções dplyr



`mutate()`

`head(iris)`

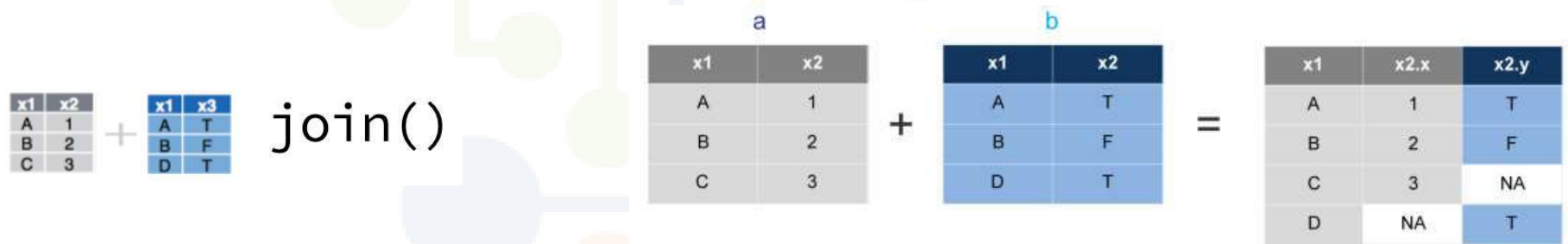
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Sepal Area
17.85
14.70
15.04
14.26
18.00
21.06



Data Science Academy

## Funções dplyr



Data Science Academy



## Funções dplyr

Existem outras funções e variações destas funções

O pacote dplyr permite que se realize operações complexas com dataframes e matrizes, utilizando apenas uma instrução



Data Science Academy





Operador %>%

`filter(data, variable == numeric_value)`

ou

`data %>% filter(variable == numeric_value)`



Data Science Academy