



Data Science Academy

# R - Fundamentos para Análise de Dados



Data Science Academy

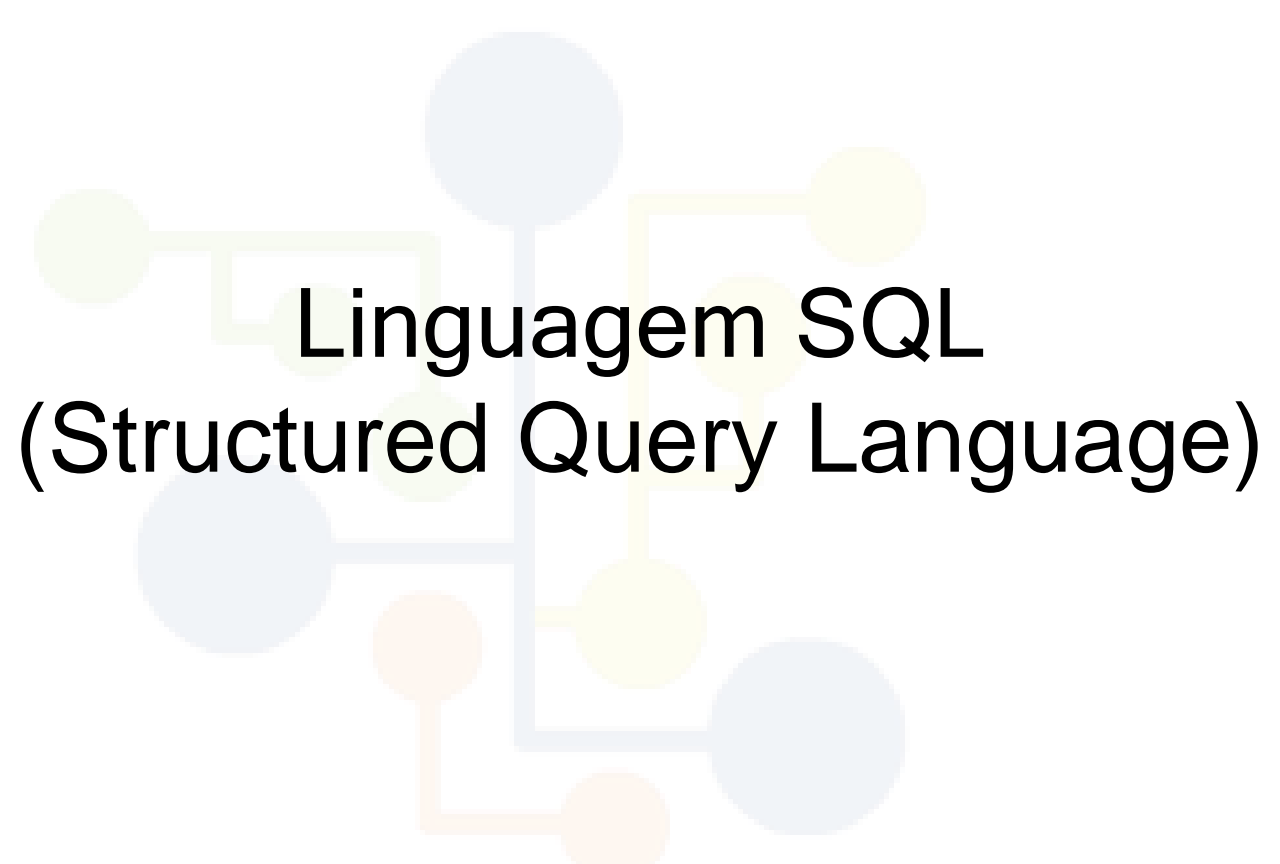

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Seja Bem-Vindo



Data Science Academy

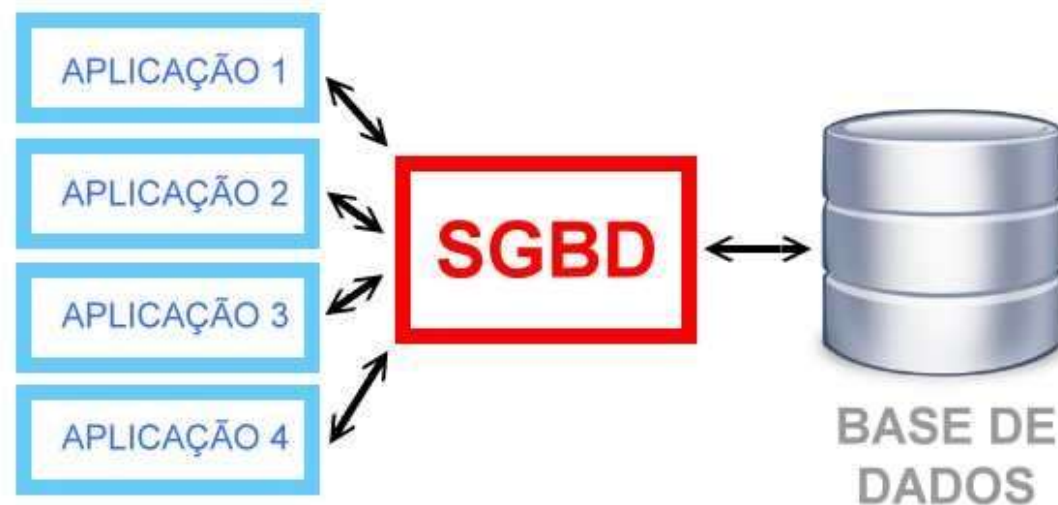


# Linguagem SQL (Structured Query Language)



Data Science Academy

# Sistemas Gerenciadores de Bancos de Dados Relacionais



Data Science Academy

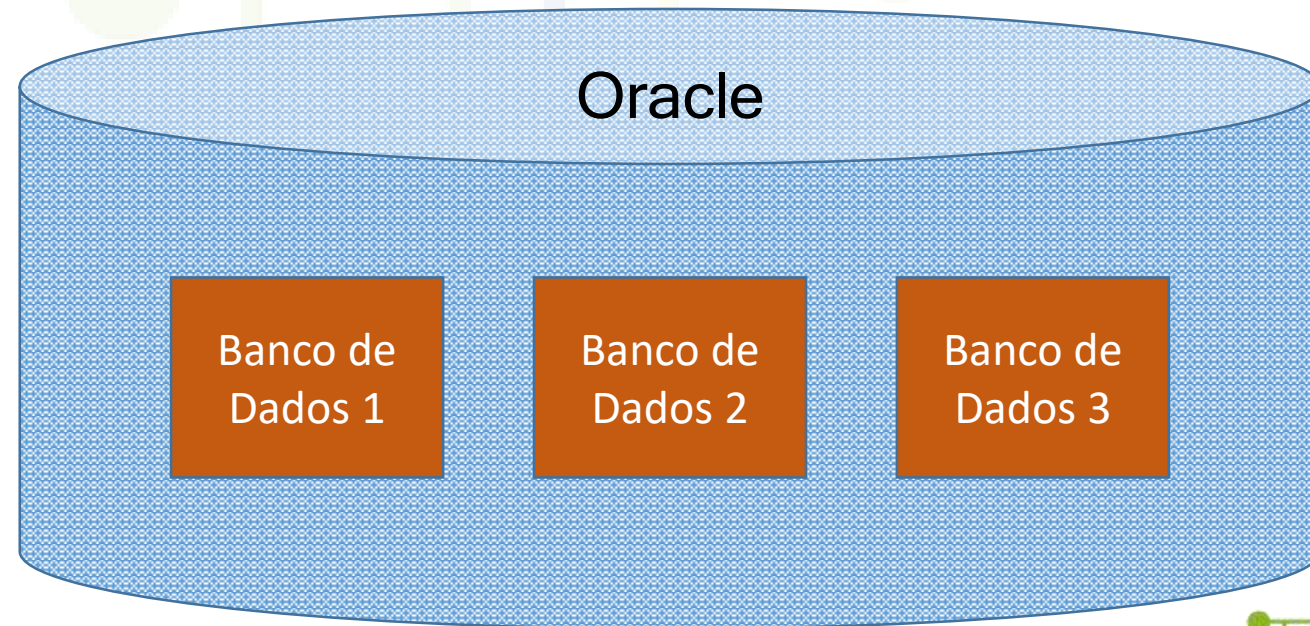
# Sistemas Gerenciadores de Bancos de Dados Relacionais

Tudo que fazemos no banco de dados, passa pelo SGBD



Data Science Academy

# Sistemas Gerenciadores de Bancos de Dados Relacionais




Data Science Academy

# Sistemas Gerenciadores de Bancos de Dados Relacionais

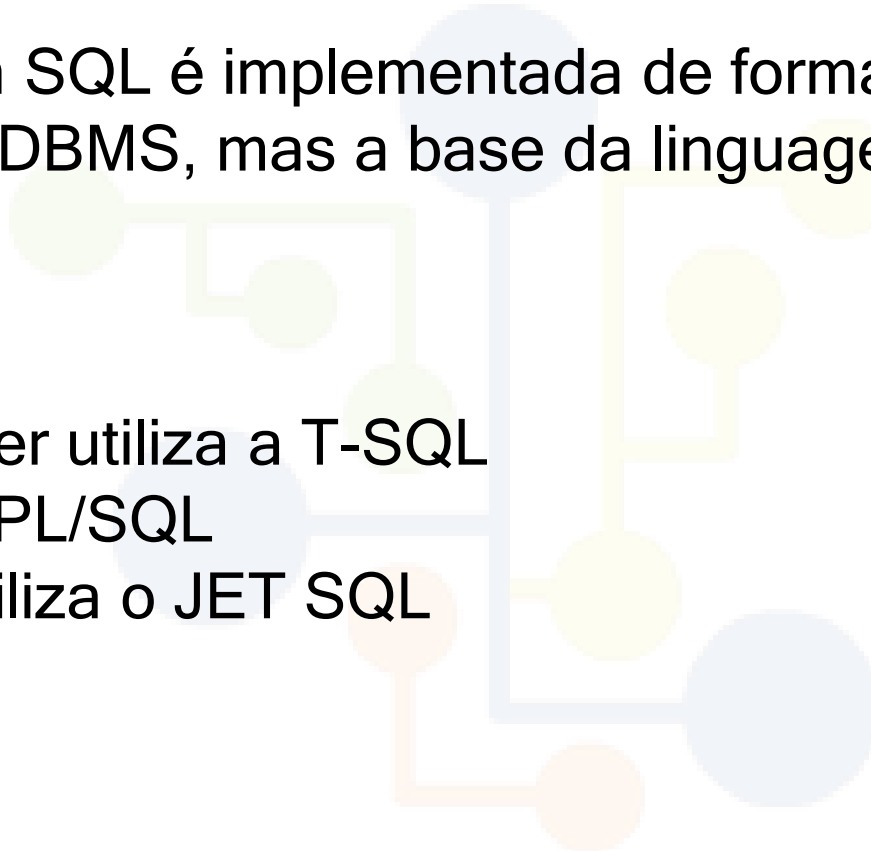
ORACLE®



Data Science Academy



A Linguagem SQL é implementada de forma diferente em diferentes RDBMS, mas a base da linguagem é a mesma

- MS SQL Server utiliza a T-SQL
  - Oracle utiliza PL/SQL
  - MS Access utiliza o JET SQL
- 



Data Science Academy





## Quais o benefícios da Linguagem SQL?

Permite que os usuários acessem dados em sistemas de gerenciamento de bancos de dados relacionais



Data Science Academy



## Quais o benefícios da Linguagem SQL?

Permite a manipulação de dados armazenados em bancos de dados



Data Science Academy



## Quais o benefícios da Linguagem SQL?

Permite a criação e remoção de objetos no banco de dados  
(tabelas, índices, visões, procedimentos armazenados)



Data Science Academy

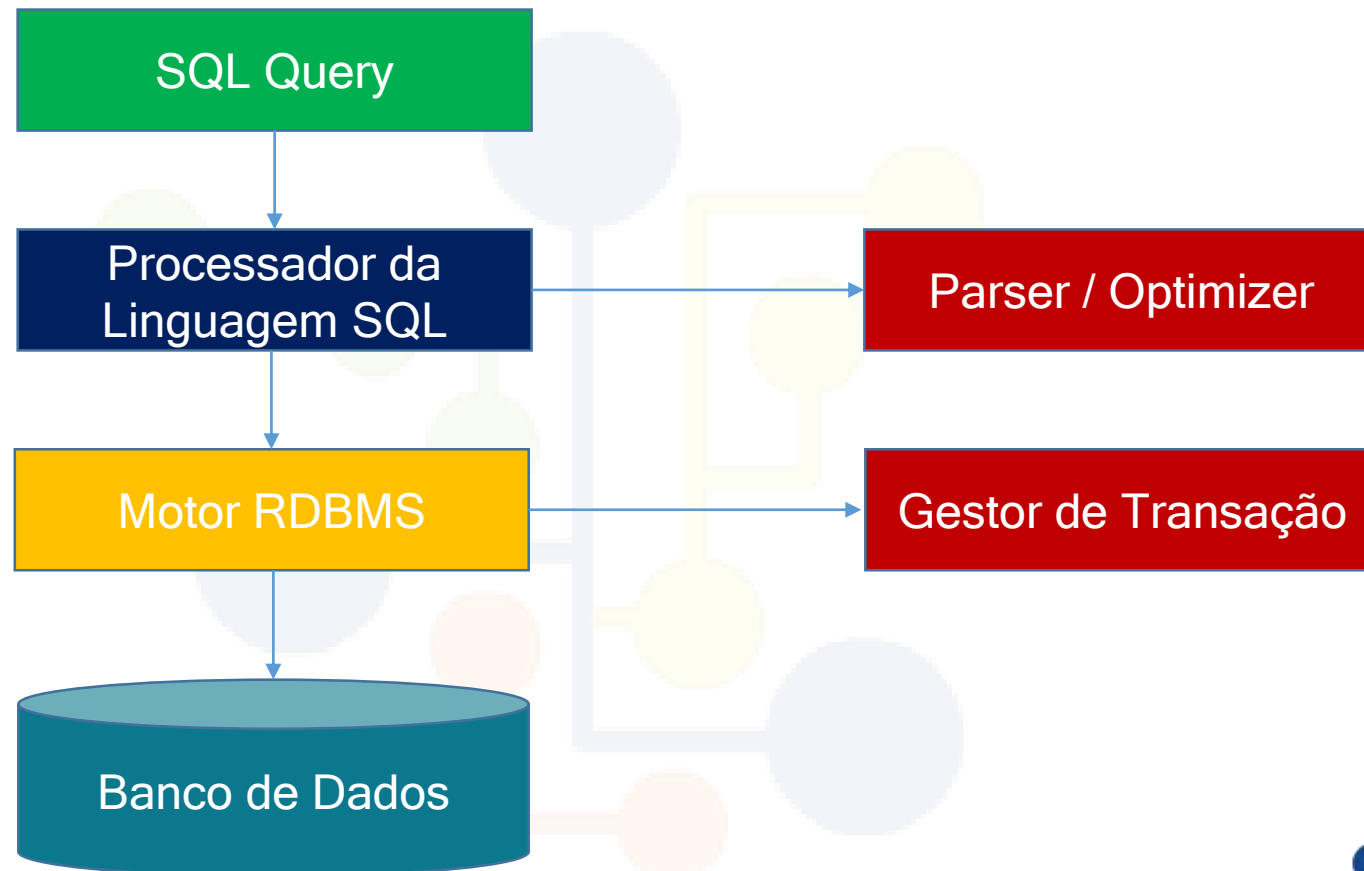


## Quais o benefícios da Linguagem SQL?

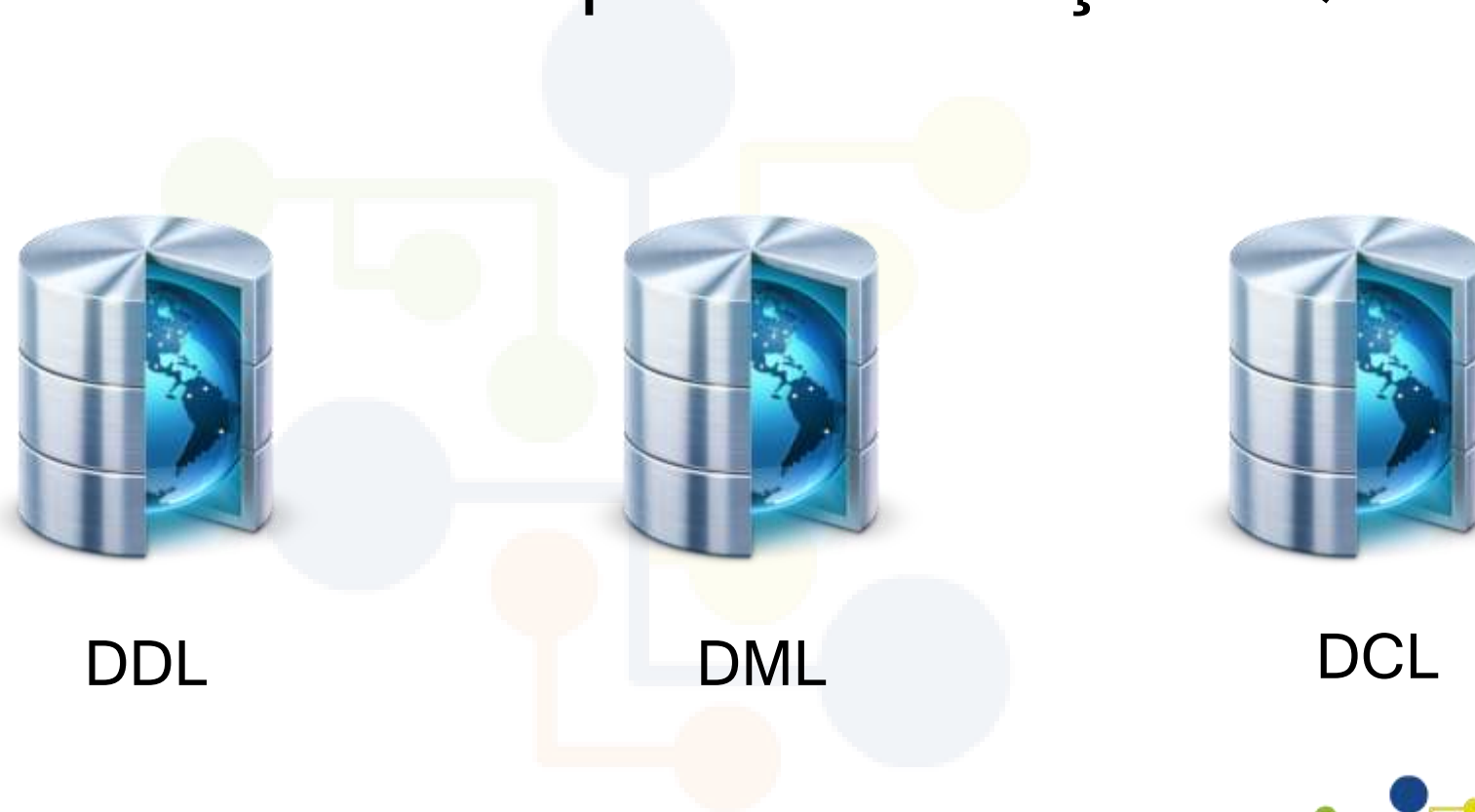
Permite que os usuários possam definir restrições de acesso



Data Science Academy



# Existem 3 Tipos de Instrução SQL



Data Science Academy



DDL

## DDL – Data Definition Language

- Create
- Alter
- Drop



Data Science Academy



DML

## DML – Data Manipulation Language

- Select
- Insert
- Delete
- Update



Data Science Academy





DCL

## DCL – Data Control Language

- Revoke
- Alter



Data Science Academy

# Tabela

Coluna



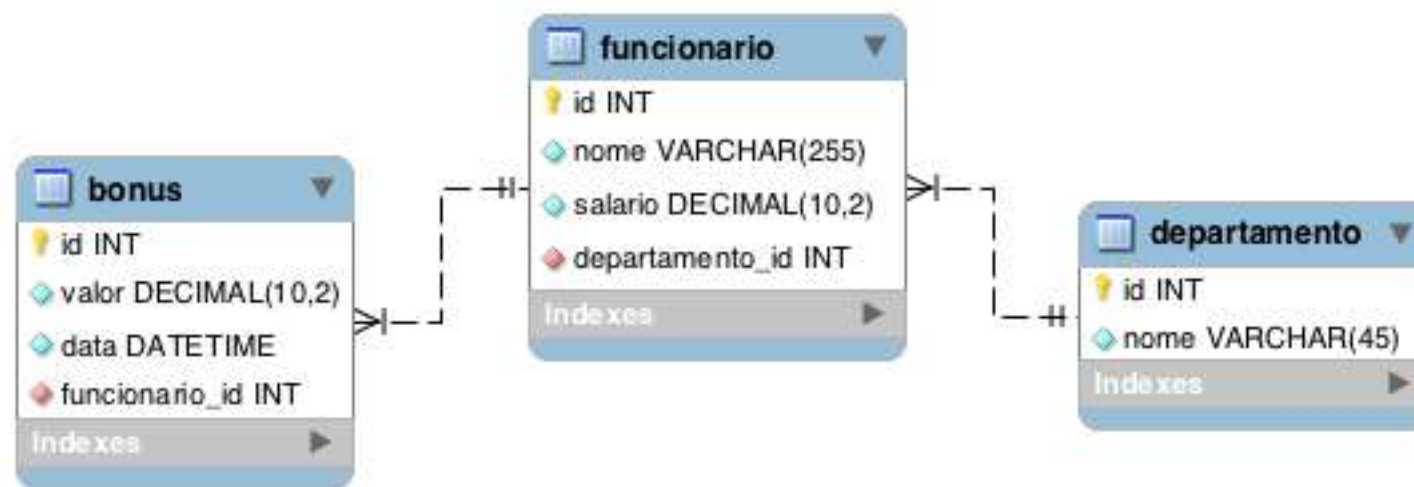
ID	NOME	IDADE	CIDADE
0001	Pele	120	Roma
0002	Zico	110	Paris
0003	Garrincha	105	Vienna

Linha  
ou  
Registro



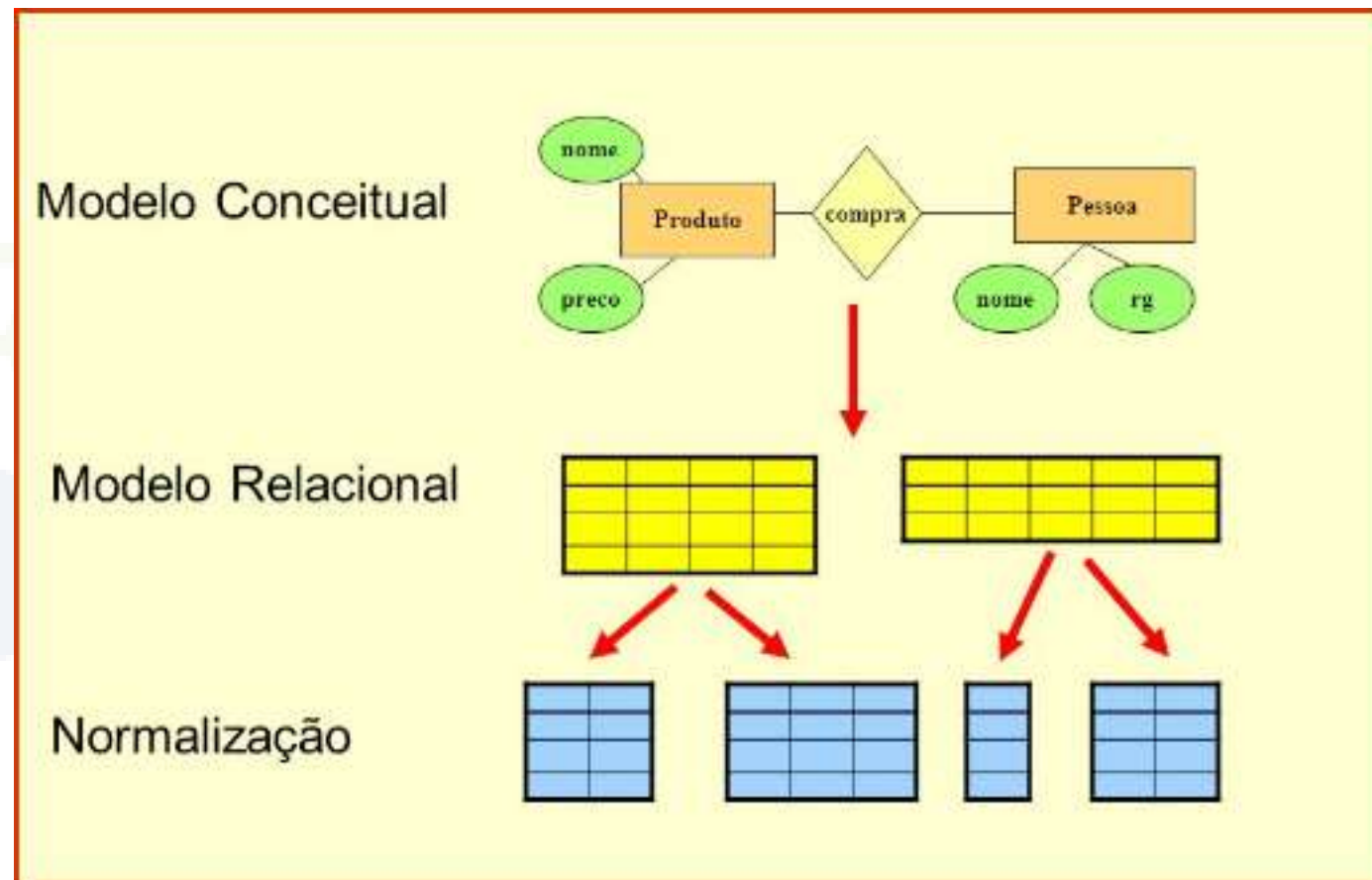
Data Science Academy

# Constraints - Integridade Referencial



Data Science Academy

# Normalização



Data Science Academy



Data Science Academy



# Importação e Manipulação de Dados de Bancos de Dados Relacionais



Data Science Academy



Bancos de Dados são Coleções de Tabelas



Data Science Academy



DataFrames em R são estruturas  
semelhantes a Tabelas

Observações => Linhas  
Variáveis => Colunas



Data Science Academy





Como acessamos dados em tabelas?

Linguagem SQL



Data Science Academy



# Sistemas Gerenciadores de Bancos de Dados



Data Science Academy



E como o R se conecta aos SGBD's?



Data Science Academy

# 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

# 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:



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 → `dbListTables()`
- 

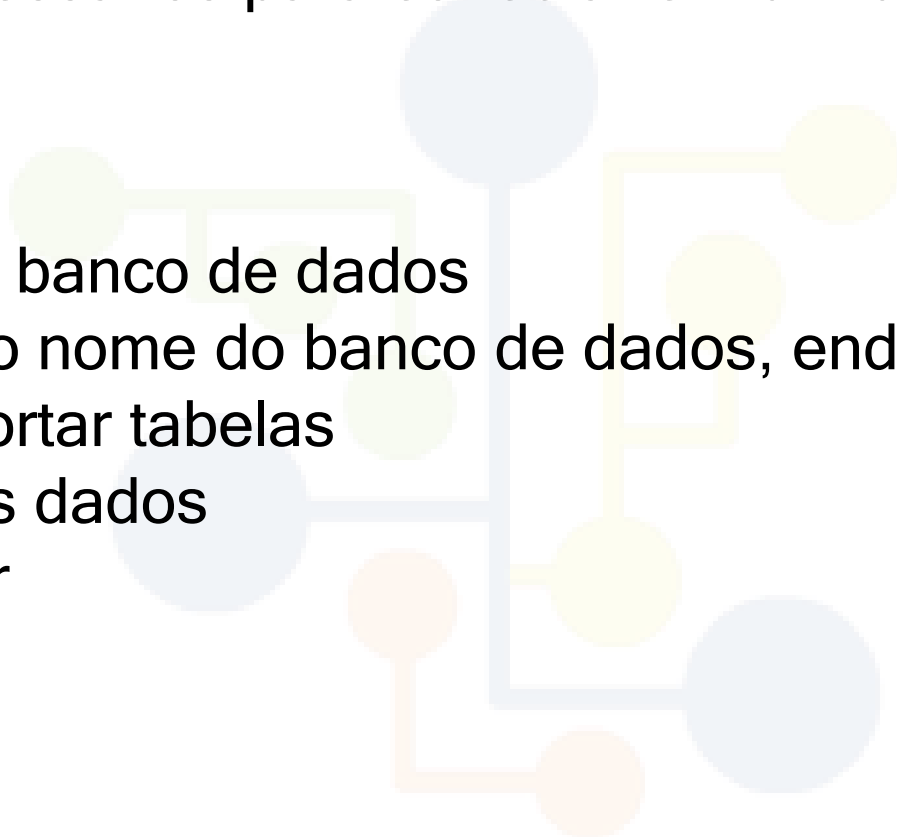


Viu como vetores  
são importantes?





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



Data Science Academy



# Importação e Manipulação de Dados de Bancos de Dados NoSQL



Data Science Academy



# Bancos de Dados NoSQL (Not Only SQL)



Data Science Academy

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







E por que devo aprender a usar um  
banco de dados NoSQL?



Data Science Academy





Data Science Academy

# Preparação



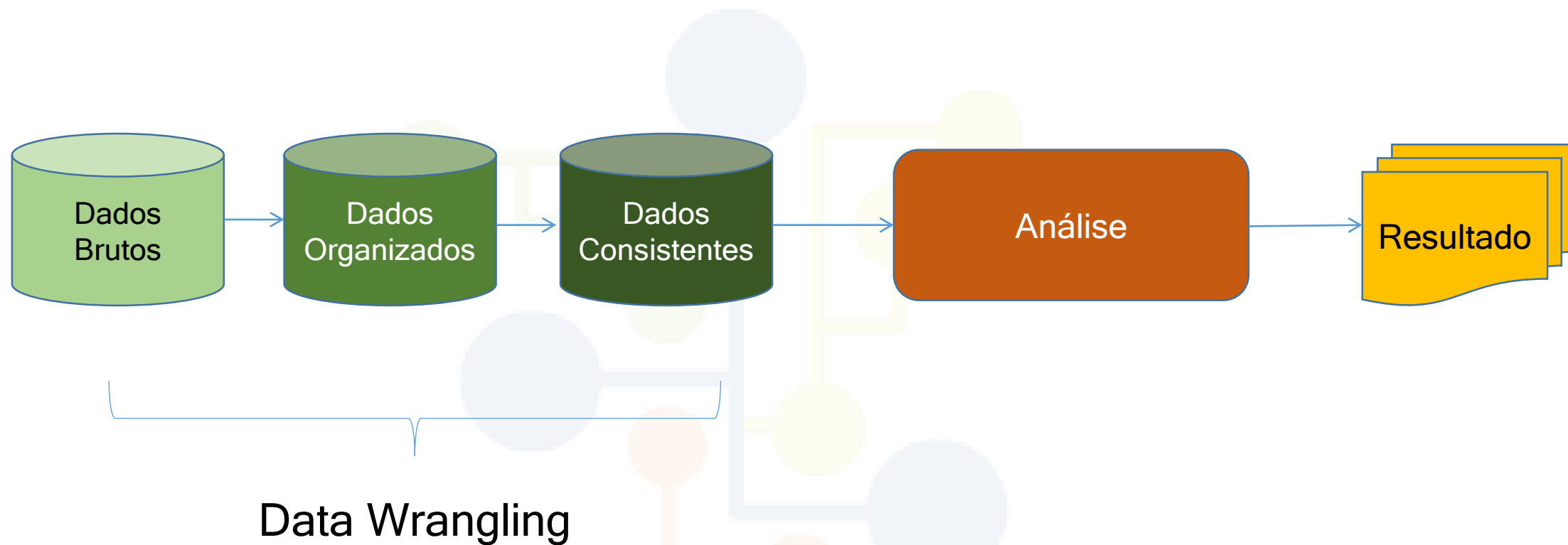
Data Science Academy



# Data Wrangling (Manipulação de Dados)



Data Science Academy







Como o cliente explicou o que queria



Como o gerente do projeto entendeu



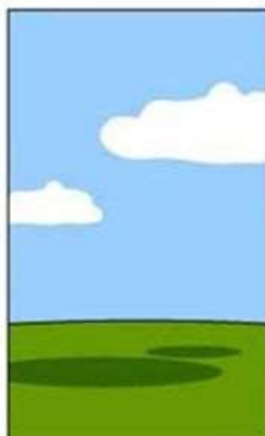
Como foi idealizado



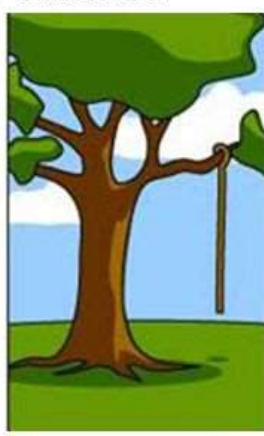
Como foi planejado



Como o gerente o explicou ao cliente



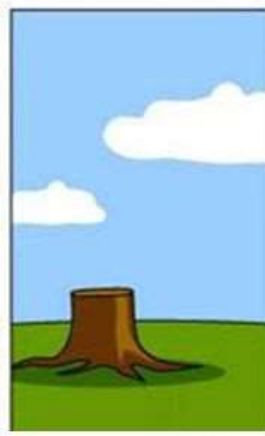
Como o projeto foi documentado



Como o projeto foi entregue



Como o cliente foi cobrado



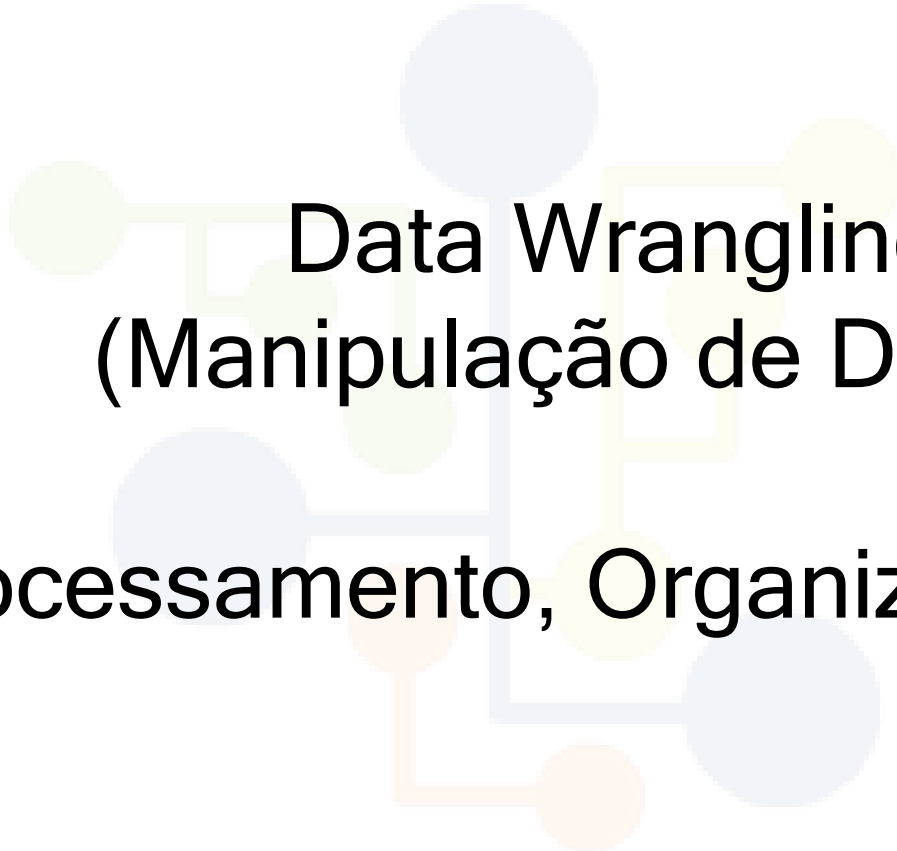

Como o projeto foi apoiado



O que o cliente realmente precisava



Data Science Academy



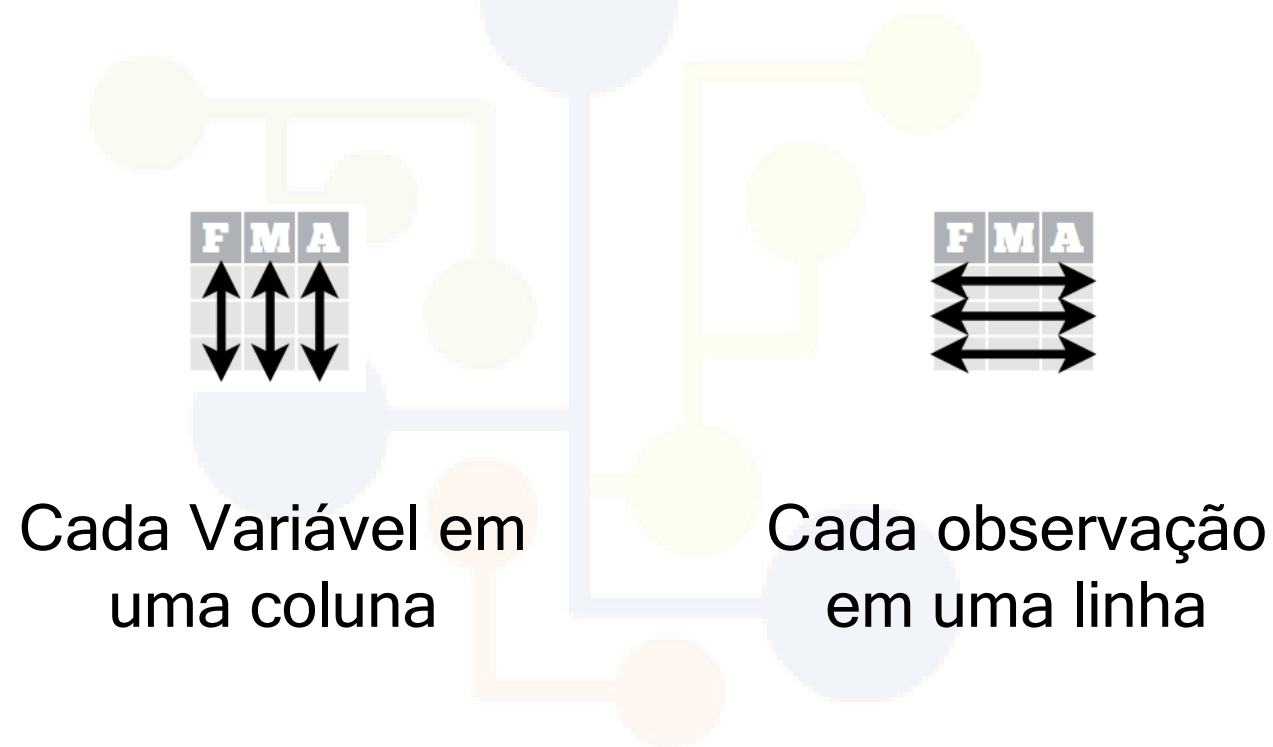
# Data Wrangling (Manipulação de Dados)

Limpeza, Processamento, Organização e Manipulação



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



## dplyr

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

## tidyr

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





tidyr

# Remodelagem de Dados



Data Science Academy

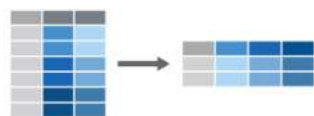
# Funções tidyr



gather()



separate()



spread()

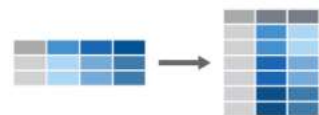


unite()



Data Science Academy

# Funções tidyr



`gather()`

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

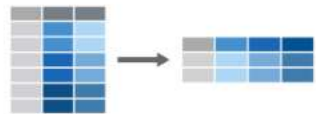
`gather()`

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



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

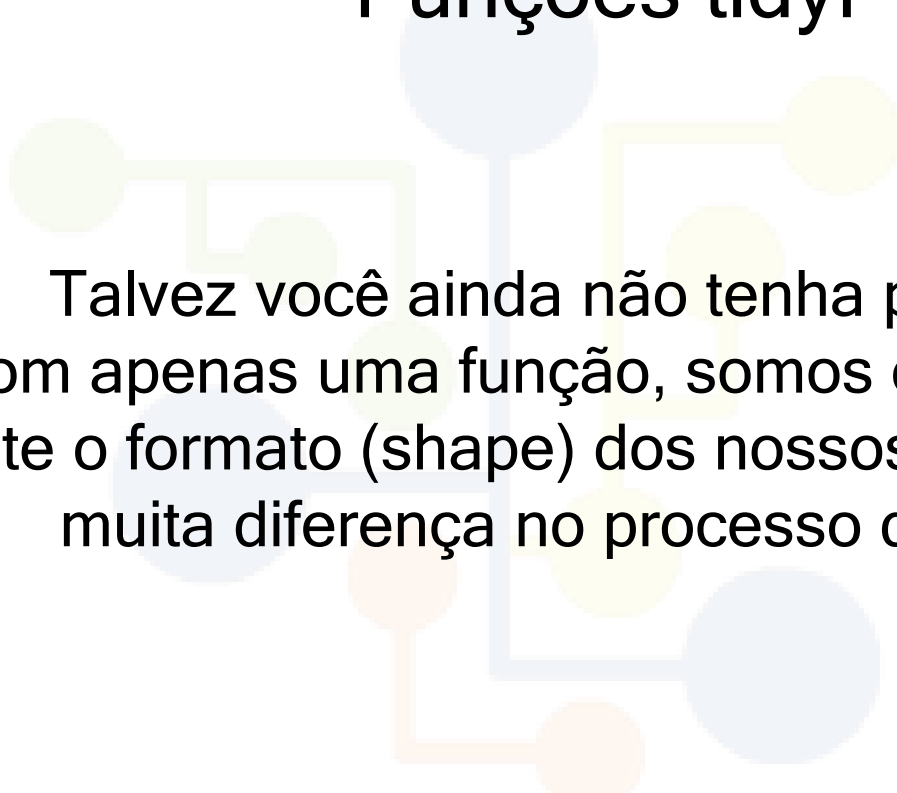


Data Science Academy





## 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



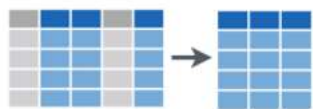
diplyr

# 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

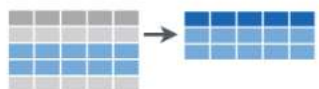


wind	pressure	date
110	1007	2000-08-12
45	1009	1998-07-30
65	1005	1995-06-04
40	1013	1997-07-01
50	1010	1999-06-13
45	1010	1996-06-21



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



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

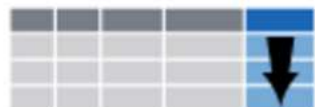


Species	Mean	SD	n
setosa	5.006	0.352	50
versicolor	5.936	0.516	50
virginica	6.588	0.636	50



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



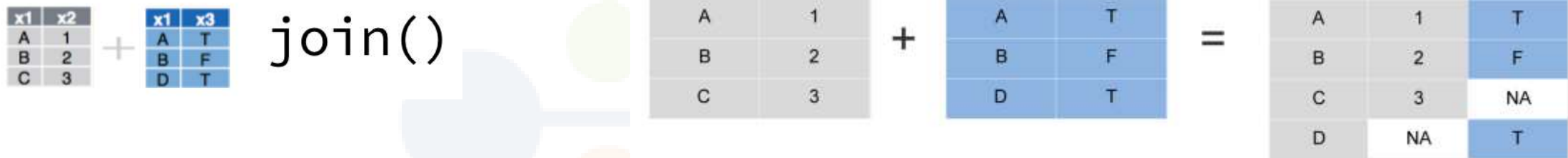
Sepal Area
17.85
14.70
15.04
14.26
18.00
21.06



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



Data Science Academy



Data Science Academy



Operador %>%

`filter(data, variable == numeric_value)`

ou

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



Data Science Academy



Data Science Academy



Obrigado!



Data Science Academy