



Aquele que enxerga longe!

www.apoemacursos.com

Data Science

Módulo 1

Ed. Enseada Trade Center
Rua Professor Almeida Cousin, 125, Sala 1003, 10º andar
Enseada do Suá, Vitória – ES, CEP: 29050-565

Tópicos

- Motivos para fazer um curso de Data Science
- Ferramentas do Cientista de Dados
- Obtendo ajuda
- Procurando respostas
- Introdução a Programação R

Motivos para fazer um curso de Data Science



Vamos resolver este problema usando Big Data, mas nenhum de nós tem a menor ideia de como fazer isso

Motivos para fazer um curso de Data Science

Análise para Big Data

“O desafio fundamental para as aplicações de Big Data é explorar os grandes volumes de dados e extrair informações úteis ou conhecimento para futuras ações”

Fonte: Rajaraman and Ullman 2012

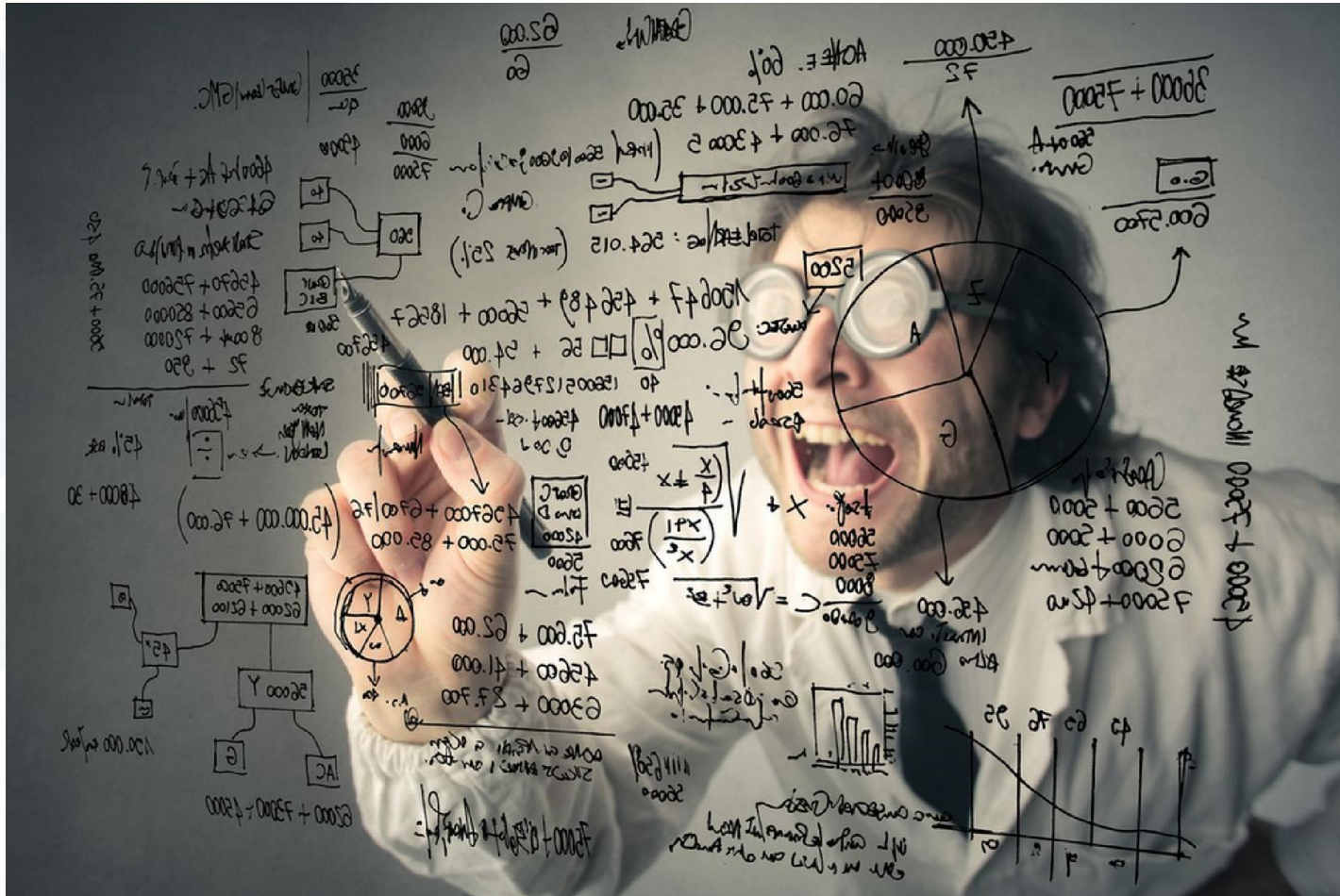
24

[illegible]

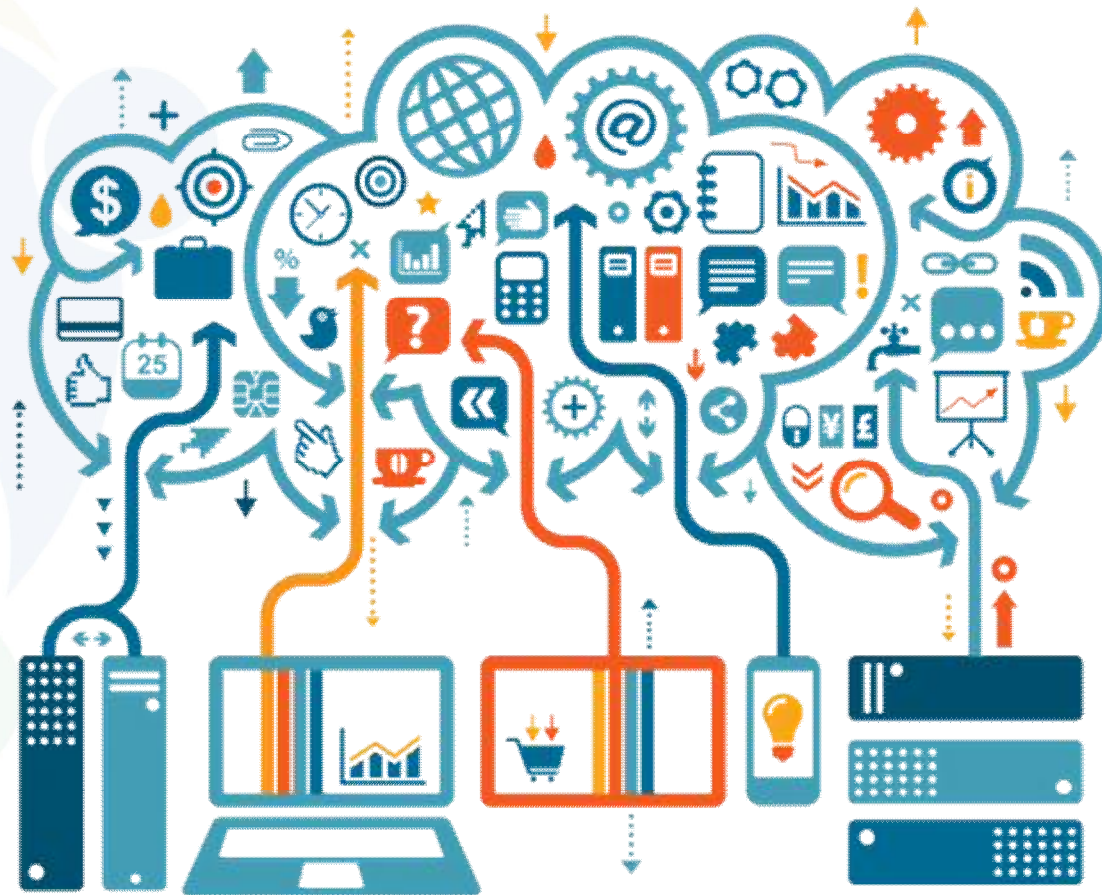
Motiv



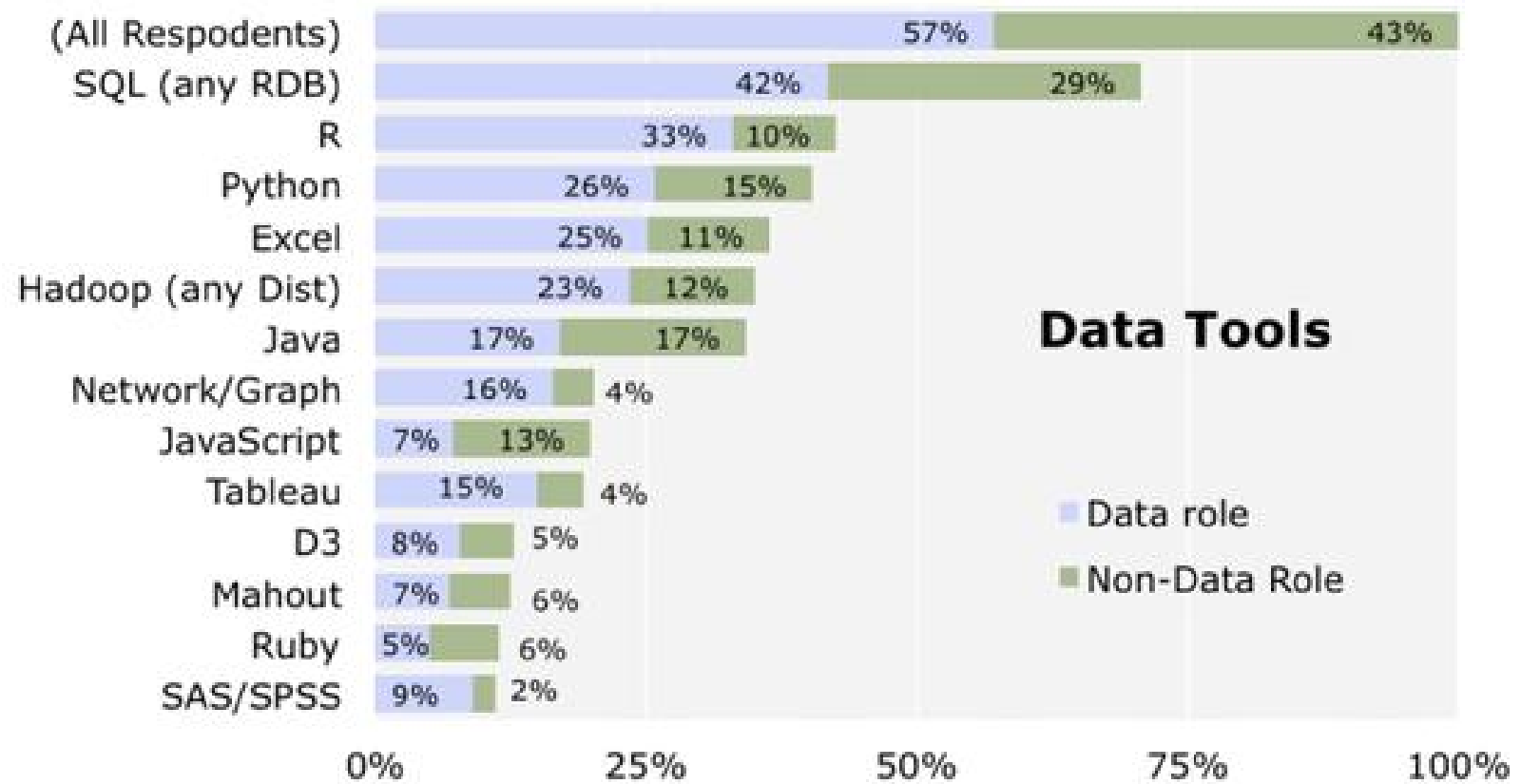
Por que Ciência de Dados?



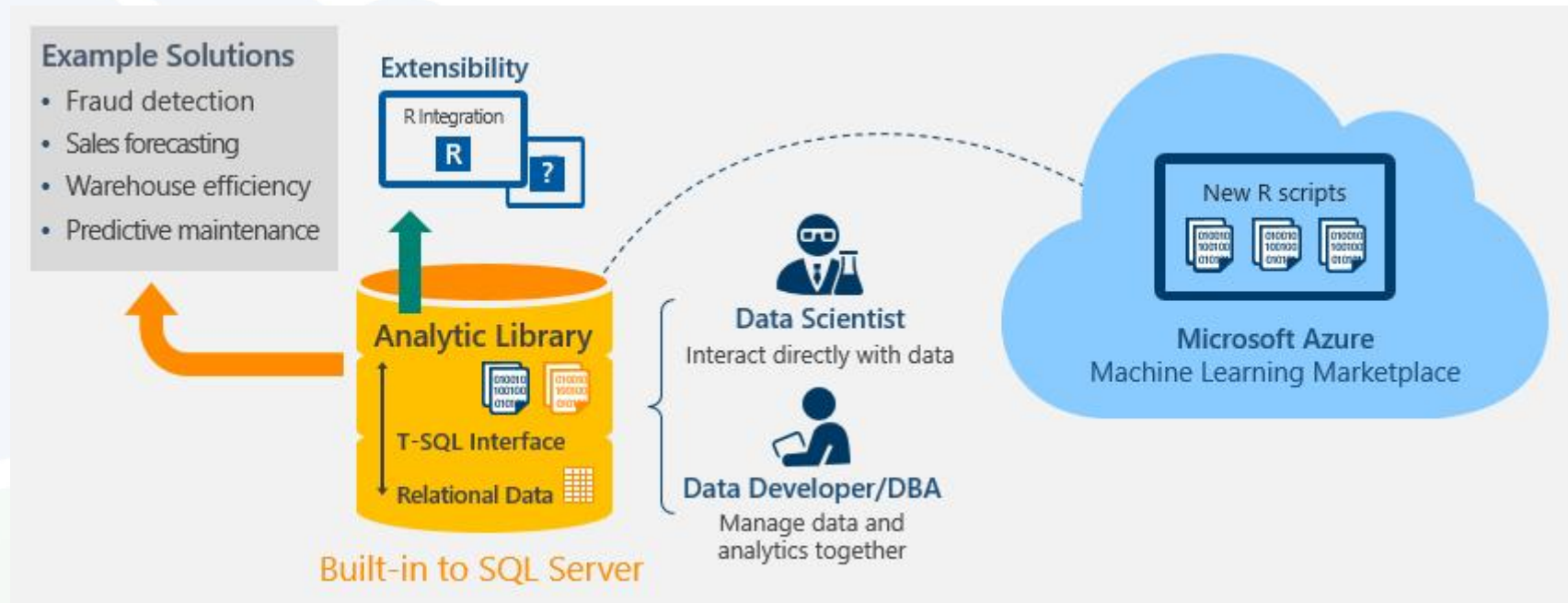
Por que Ciência de Dados?



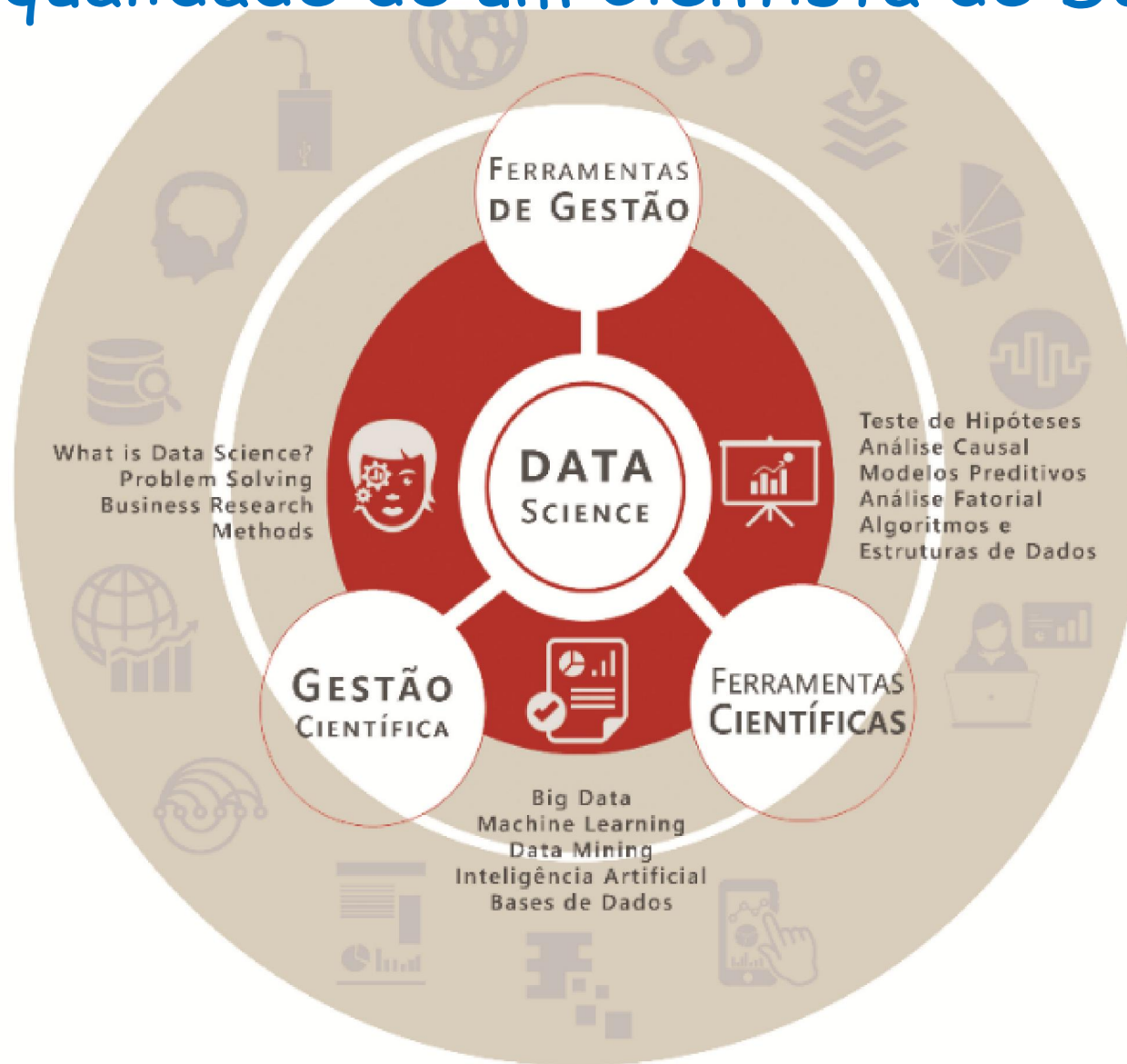
Por que a Linguagem R?



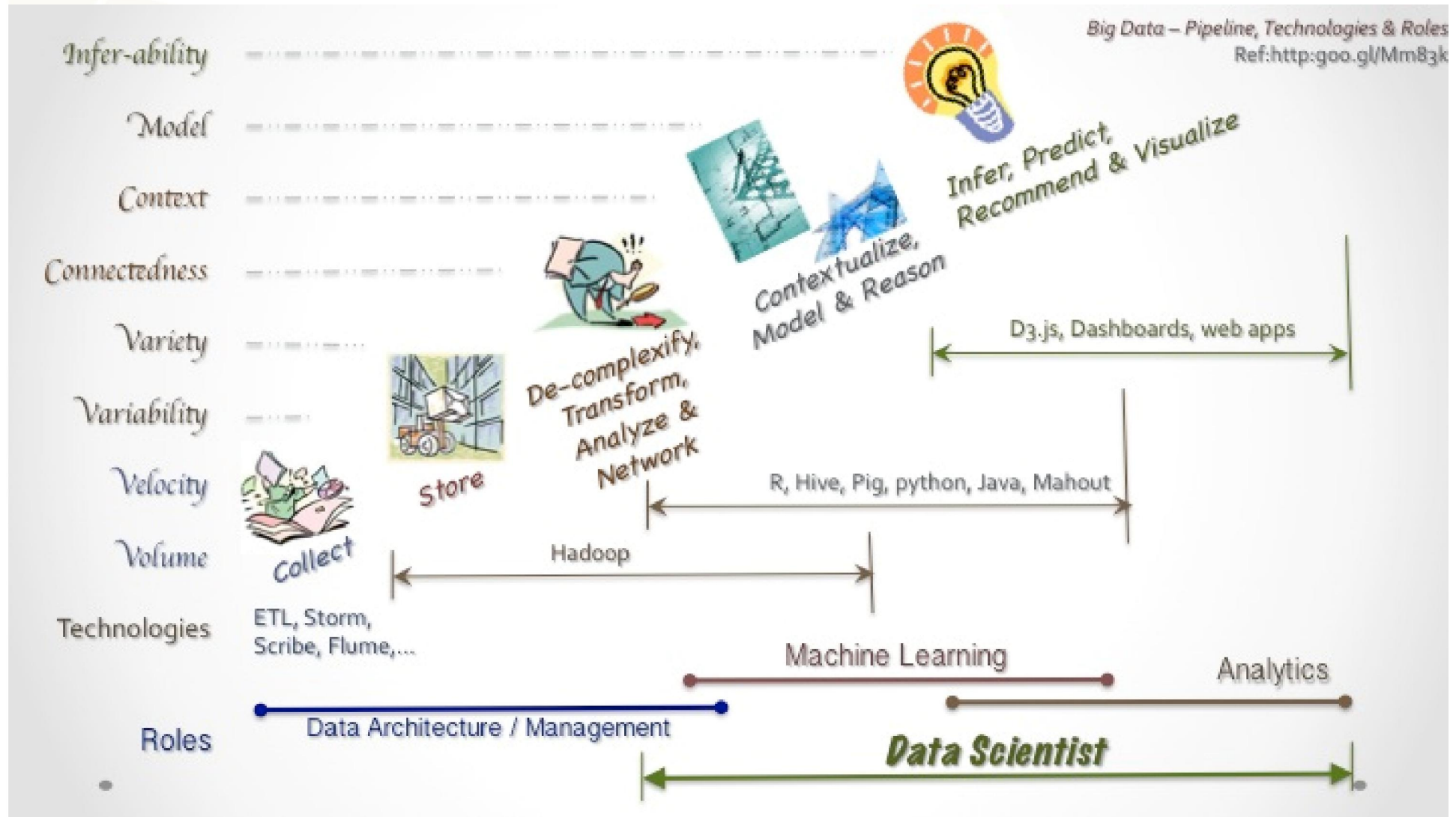
Por que a Linguagem R?



As qualidades de um Cientista de Dados



As qualidade de um Cientista de Dados



Coleta e Limpeza do Dados

- Dados Brutos X Arrumados (já processados)
- Download / Carga
- Lendo dados de várias fontes
 - Excel, XML, JSON, MySQL, HDF5, Web, ...
- Merging Data
- Reshaping Data
- Sumarizando Dados
- Procurando e alterando
- Dados e Recursos

Conectando e Listando base de dados

```
ucscDb <- dbConnect(MySQL(), user = "genome", host = "genome-mysql.cse.ucsc.edu")
result <- dbGetQuery(ucscDb, "show databases;")
dbDisconnect(ucscDb)
result
```

Merging Data

```
mergedData2 <- merge(reviews, solutions, by.x = "solution_id", by.y = "id",  
  all = TRUE)  
head(mergedData2[, 1:6], 3)  
reviews[1, 1:6]
```

Dados brutos x Dados processados

- **Dados não tratados**

- A fonte original dos dados
- Frequentemente difícil de usar para análise de dados
- A análise de dados inclui processamento
- Os dados brutos só precisam ser processados uma vez

- **Dados processados**

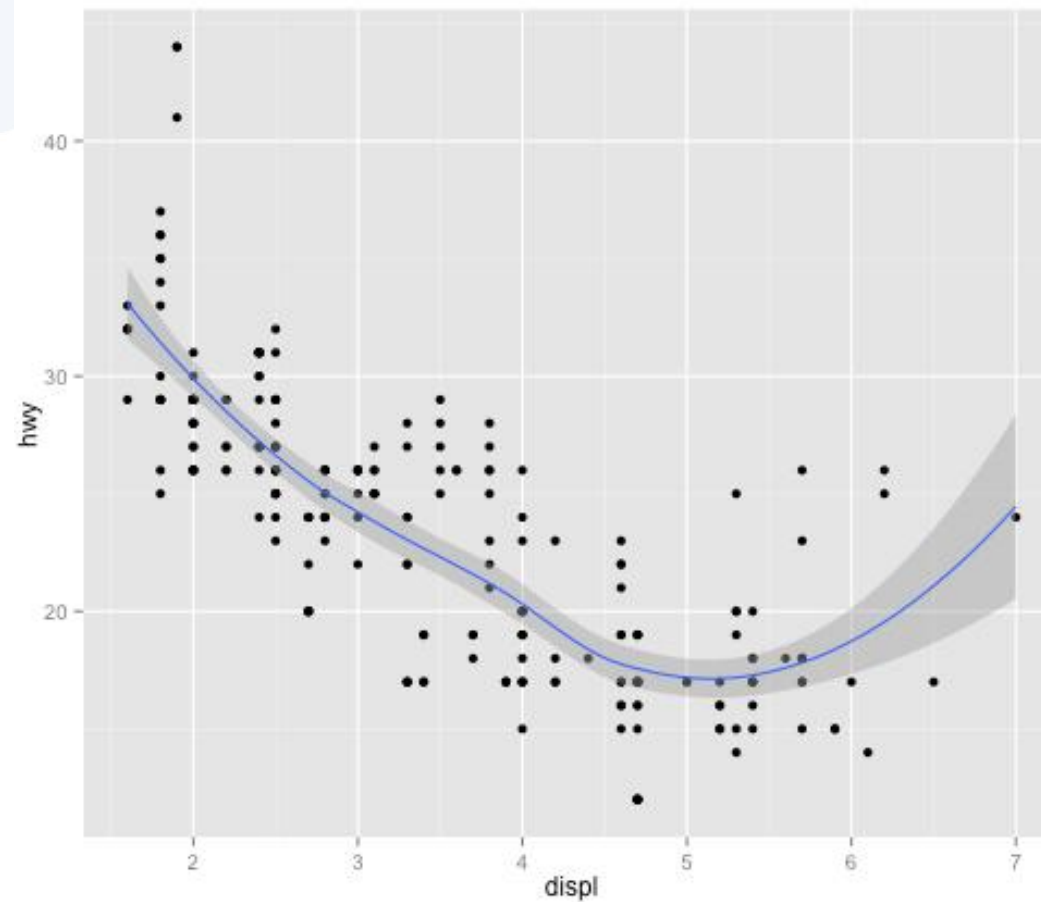
- Dados prontos para análise
- O processamento pode incluir fusão, subconjunto, transformação, etc.
- Podem existir normas
- Todas as etapas devem ser registradas

Análise Exploratória de Dados

- Princípios da Analytic Graphics
- Explorando gráficos
- Sistemas de Plotagem em R
 - base
 - lattice
 - ggplot2
- Hierarchical clustering
- K-Means clustering
- Redução de Dimensionalidade

Exemplo de um Gráfico em R

```
qplot(displ, hwy, data = mpg, geom = c("point", "smooth"))
```

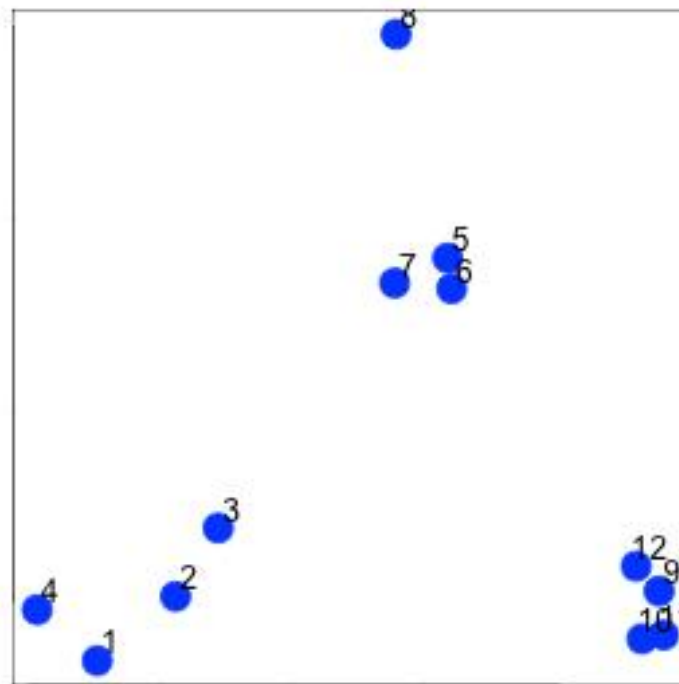


Princípios da Análise Gráfica

- Princípio 1: Mostrar comparações
- Princípio 2: Mostrar causalidade, mecanismo, explicação
- Princípio 3: Mostrar dados multivariados
- Princípio 4: Integrar múltiplos modos de evidência
- Princípio 5: Descrever e documentar as evidências
- Princípio 6: O conteúdo é rei

Exemplo: K-Means Clustering

```
set.seed(1234)
par(mar = c(0, 0, 0, 0))
x <- rnorm(12, mean = rep(1:3, each = 4), sd = 0.2)
y <- rnorm(12, mean = rep(c(1, 2, 1), each = 4), sd = 0.2)
plot(x, y, col = "blue", pch = 19, cex = 2)
text(x + 0.05, y + 0.05, labels = as.character(1:12))
```



Reproduzindo conteúdo de pesquisa

- Estrutura de uma análise de dados
- Organizar uma análise de dados
- LaTeX
- R Markdown
- Análise de dados baseada em evidências
- RPubS

Passos para a Análise de Dados

1. Definir a pergunta
2. Definir o conjunto de dados ideal
3. Determinar quais dados você pode acessar
4. Obter os dados
5. Limpe os dados
6. Análise exploratória de dados
7. Previsão / modelagem estatística
8. Interpretar os resultados
9. Sintetizar / escrever resultados
10. Criar código reproduzível



Arquivos de Análise de Dados

- Data
 - Bruto
 - Processado
- Figuras
 - Exploratorias
 - Finais
- Código R
 - Raw scripts
 - Final scripts
 - R Markdown files
- Texto
 - Readme files
 - Textos de análises

Estatística Inferencial

- Probabilidade básica
- Distribuições comuns
- Assintóticas
- Intervalos de confiança
- Testes de hipóteses
- Poder
- Bootstrapping
- Testes não-paramétricos
- Estatísticas bayesianas básicas

Estatística Inferencial

Apenas um
Exemplo

- Suponha que a proporção de chamadas de ajuda que são tratadas em um dia aleatório por uma linha de ajuda é dada por:

$$f(x) = \begin{cases} 2x & \text{for } 1 > x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Trata-se de uma densidade matematicamente válida?

Distribuição Normal Padrão

- Diz-se que uma variável aleatória segue uma distribuição normal ou gaussiana com média μ e variância σ^2 se a densidade associada for

$$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

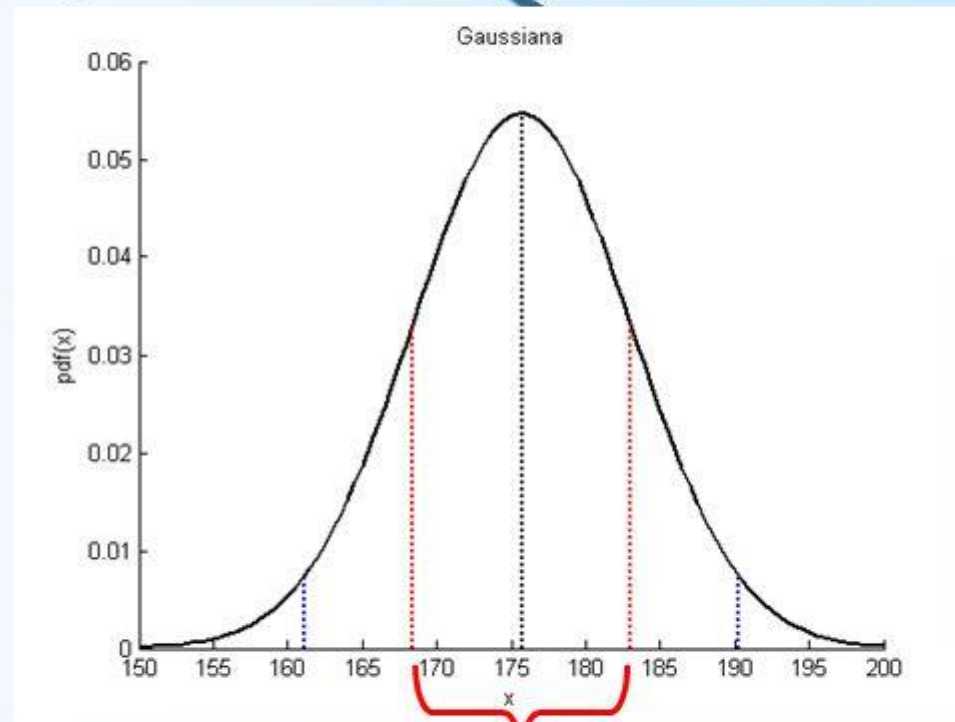
- Se X uma variável aleatória com esta densidade então
 $E[X] = \mu$ e $\text{Var}(X) = \sigma^2$
- Nós escrevemos
 $X \sim N(\mu, \sigma^2)$
- Quando $\mu = 0$ e $\sigma = 1$, a distribuição resultante é chamada **Distribuição Normal Padrão**
- A função de densidade normal padrão é etiquetada como ϕ

Distribuições de Probabilidade

2. Distribuição Gaussiana

- Teorema do Limite Central (TLC) [quem é central é o limite, e não o teorema!]
- Ex.: Altura da população masculina adulta ($\mu = 175,7$ cm e $\sigma = 7,3$ cm)

$$pdf(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$\cong 68,27\%$

$\cong 95,45\%$

Modelo de Regressão Linear

- Regressão linear
- Regressão múltipla
- Confounding
- Resíduos e diagnósticos
- Predição usando modelos lineares
- Modelo de especificação incorreta
- Alisamento / splines do Scatterplot
- Aprendizagem de máquina via regressão
- Inferência de remostarem em regressão, bootstrapping, testes de permutação
- Regressão ponderada
- Modelos mistos (interceptações aleatórias)

Uma ideia historicamente famosa, Regressão para a média

- Por que é que as crianças de pais altos tendem a ser altas, mas não tão altas quanto seus pais?
- Por que os filhos de pais baixos tendem a ser baixos, mas não tão baixos quanto seus pais?
- Por que os pais de crianças muito baixos, tendem a ser baixos, mas não tão baixo como seu filho? E o mesmo com os pais de crianças muito altas?
- Por que os melhores atletas este ano tendem a fazer um pouco pior do que o seguinte?

Métodos causais:

Regressão linear em relação aos dados reais

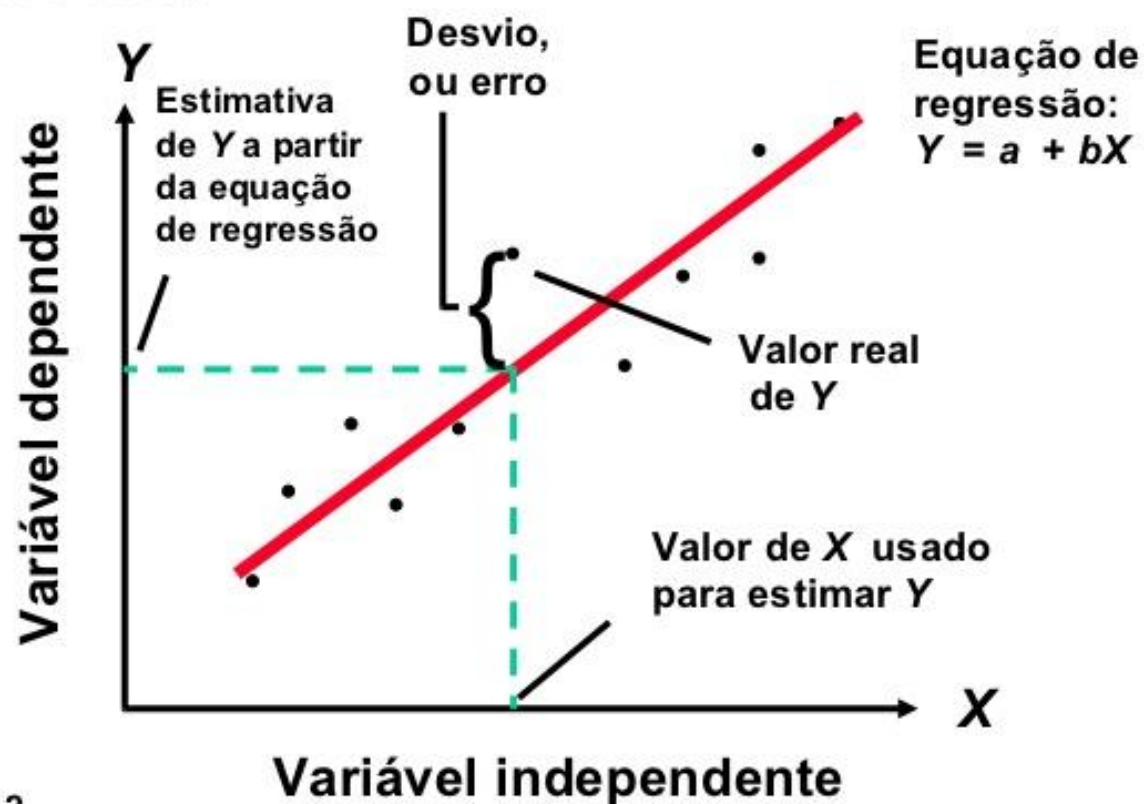
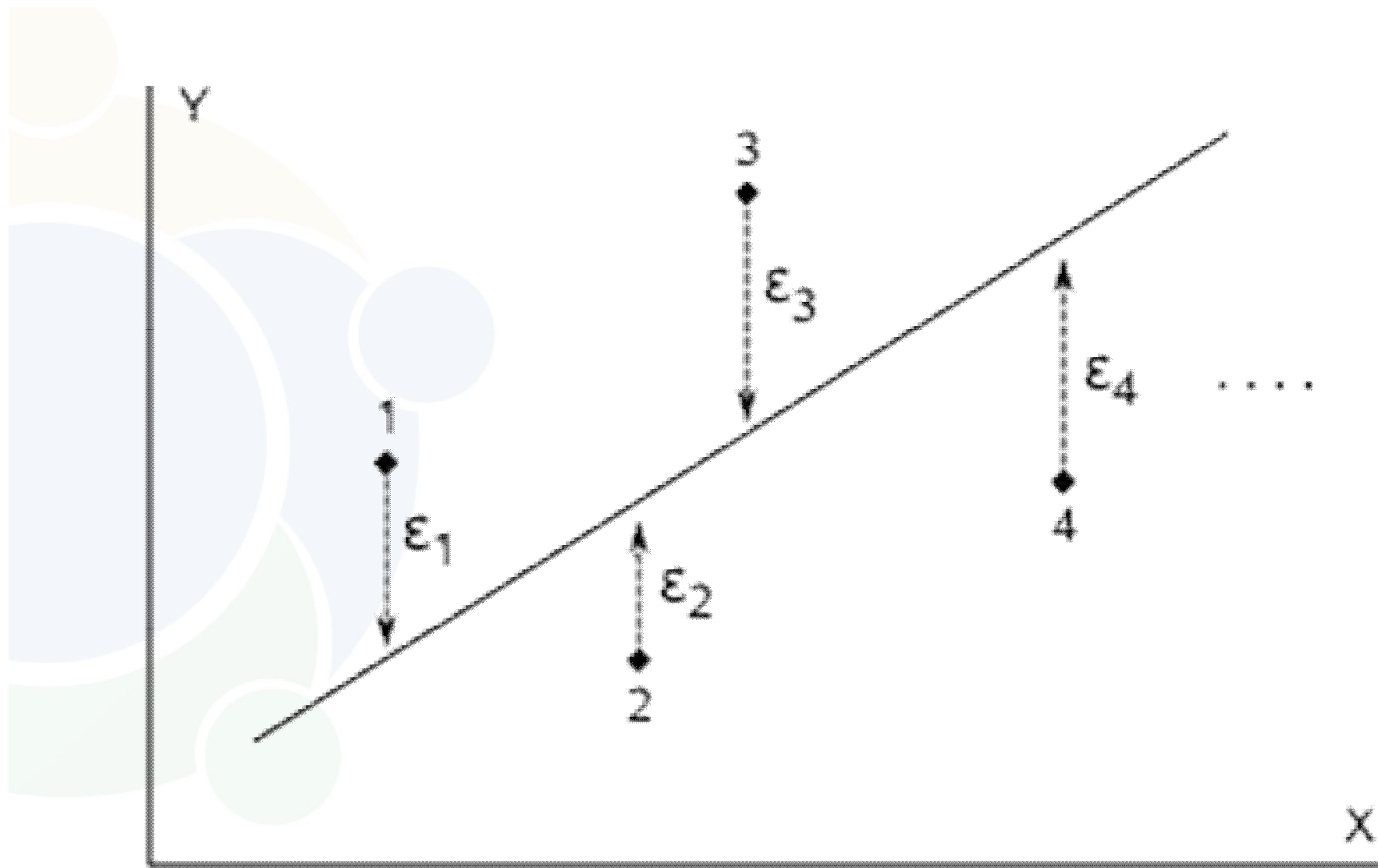
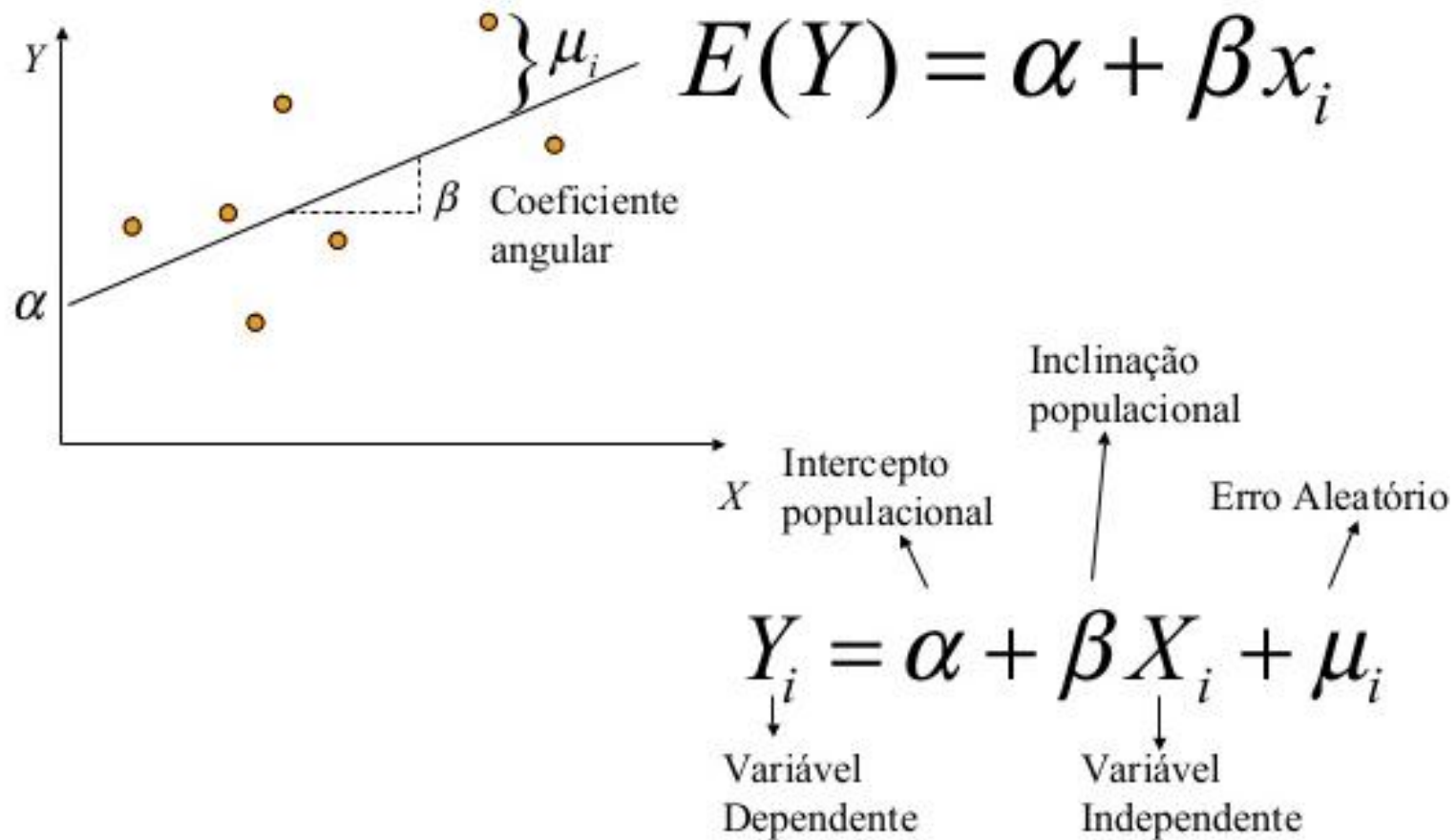


Figura 9.2



Conceitos fundamentais

Regressão Linear Simples



Aprendizagem Prática em Machine Learning

- Projeto de estudo de previsão
- Tipos de erros
- Validação cruzada
- O pacote de intercalação
- Traçar para previsão
- Pré-processando
- Previsão com regressão
- Previsão com árvores
- Boosting
- Bagging
- Model blending
- Previsão



Termos básicos

Em geral, Positivo = identificado e negativo = rejeitado. Assim sendo:

- Verdadeiro positivo = corretamente identificado
- Falso positivo = incorretamente identificado
- Negativo verdadeiro = rejeitado corretamente
- Falso negativo = rejeitado incorretamente

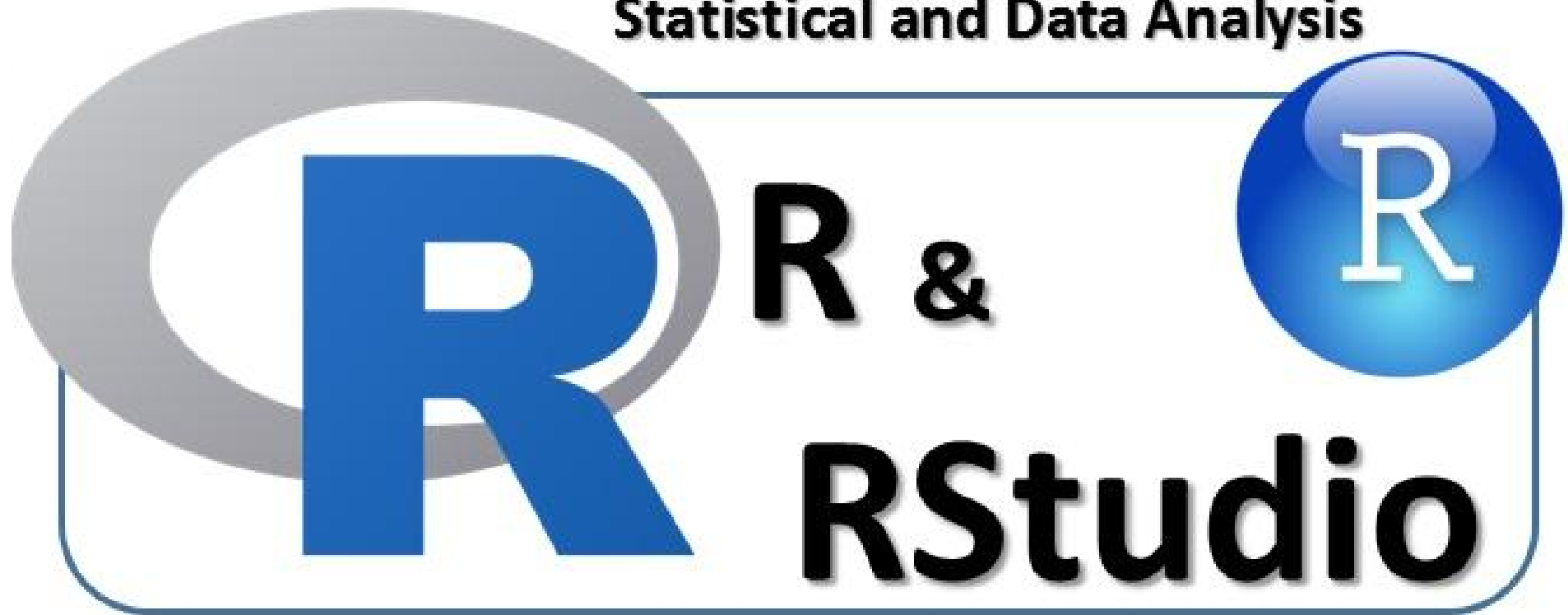
Exemplo em um teste médico:

- Verdadeiro positivo = Pessoas doentes diagnosticadas corretamente como doentes
- Falso positivo = Pessoas saudáveis identificadas incorretamente como doentes
- Verdadeiro negativo = Pessoas saudáveis corretamente identificadas como saudáveis
- Falso negativo = Pessoas doentes identificadas incorretamente como saudáveis.

Ideia básica por trás do boosting

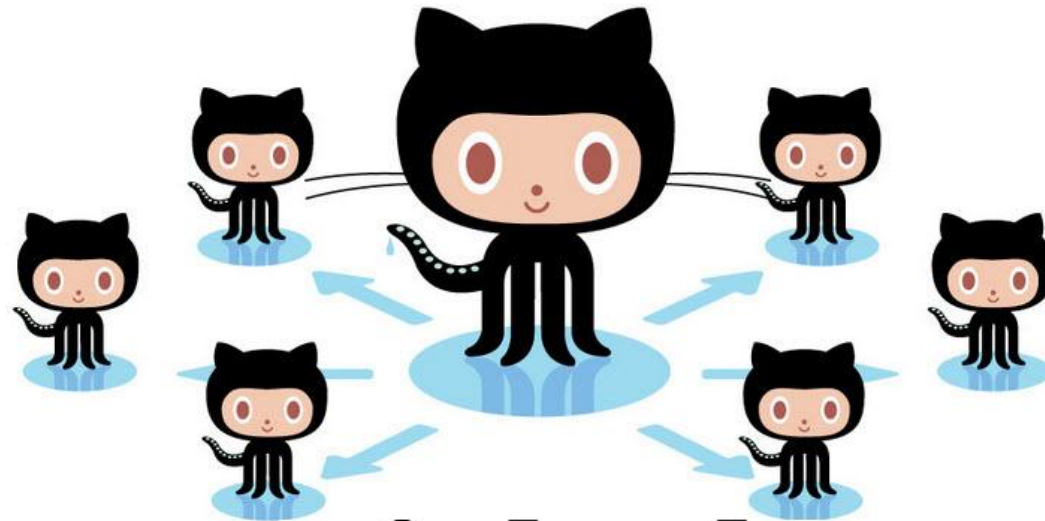
- O objetivo é minimizar o erro (no conjunto de treinamento)
- Iterativo
- Calcula pesos com base em erros

Statistical and Data Analysis



**A GUIDE TO INSTALLING R
FOR WINDOWS, LINUX AND MAC OS X**






github
SOCIAL CODING

Seguro | <https://git-scm.com/download/linux>

Home do Aluno | Learning English - 6 | Google Tradutor | Jogos de computador | Google Maps | IT eBooks - Free Dow | e-Database: Cardinali

Git is a member project of Software Freedom Conservancy, which handles legal and financial needs for the project. Conservancy is currently raising funds to continue its mission. Consider [becoming a supporter!](#)



git --everything-is-local

Search entire site...

About
Documentation
Blog
Downloads
 GUI Clients
 Logos
Community

Download for Linux and Unix

It is easiest to install Git on Linux using the preferred package manager of your Linux distribution. If you prefer to build from source, you can find the tarballs [on kernel.org](https://kernel.org).

Debian/Ubuntu

```
$ apt-get install git
```

Fedora

```
$ yum install git (up to Fedora 21)  
$ dnf install git (Fedora 22 and later)
```

Git is a member project of Software Freedom Conservancy, which handles legal and financial needs for the project. Conservancy is currently raising funds to continue its mission. Consider [becoming a supporter](#)!



Search entire site...

About

Documentation

Blog

Downloads

GUI Clients

Logos

Community

The entire [Pro Git book](#) written by Scott Chacon and Ben Straub is available to [read online for free](#). Dead tree versions are available on [Amazon.com](#).

GUI Clients

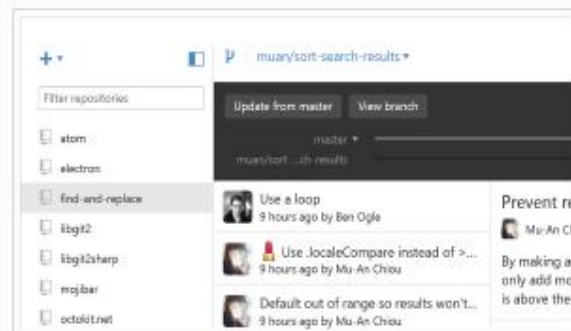
Git comes with built-in GUI tools for committing ([git-gui](#)) and browsing ([gitk](#)), but there are several third-party tools for users looking for platform-specific experience.

All

Windows

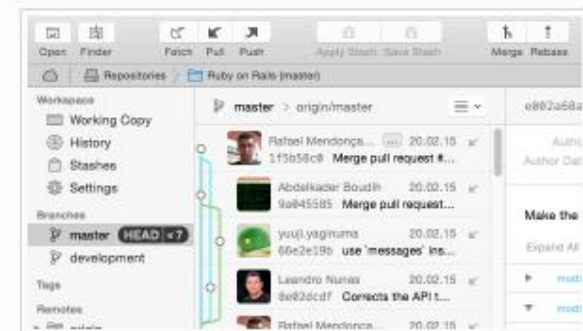
Mac

Linux



GitHub Desktop

Platforms: Windows, Mac
Price: Free



Tower

Platforms: Windows, Mac
Price: \$79/user (Free 30 day trial)

TheScientistBr / DataScienceTraining

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

No description, website, or topics provided.

Edit

New Add topics

1 commit

1 branch

0 releases

1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download



dbranquinho Initial commit

Latest commit 4a5e1aa just now

.gitignore

Initial commit

just now

LICENSE

Initial commit

just now

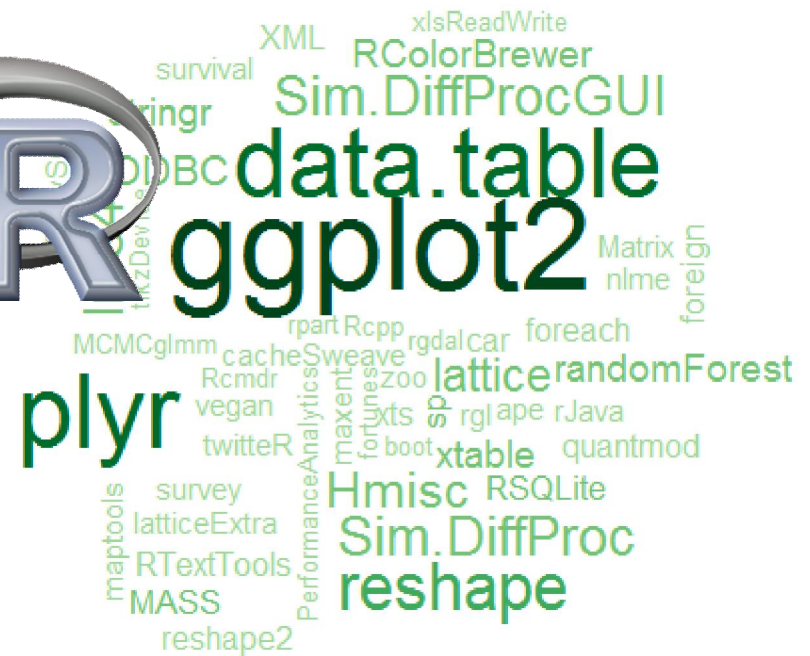
README.md

Initial commit

just now

README.md

DataScienceTraining

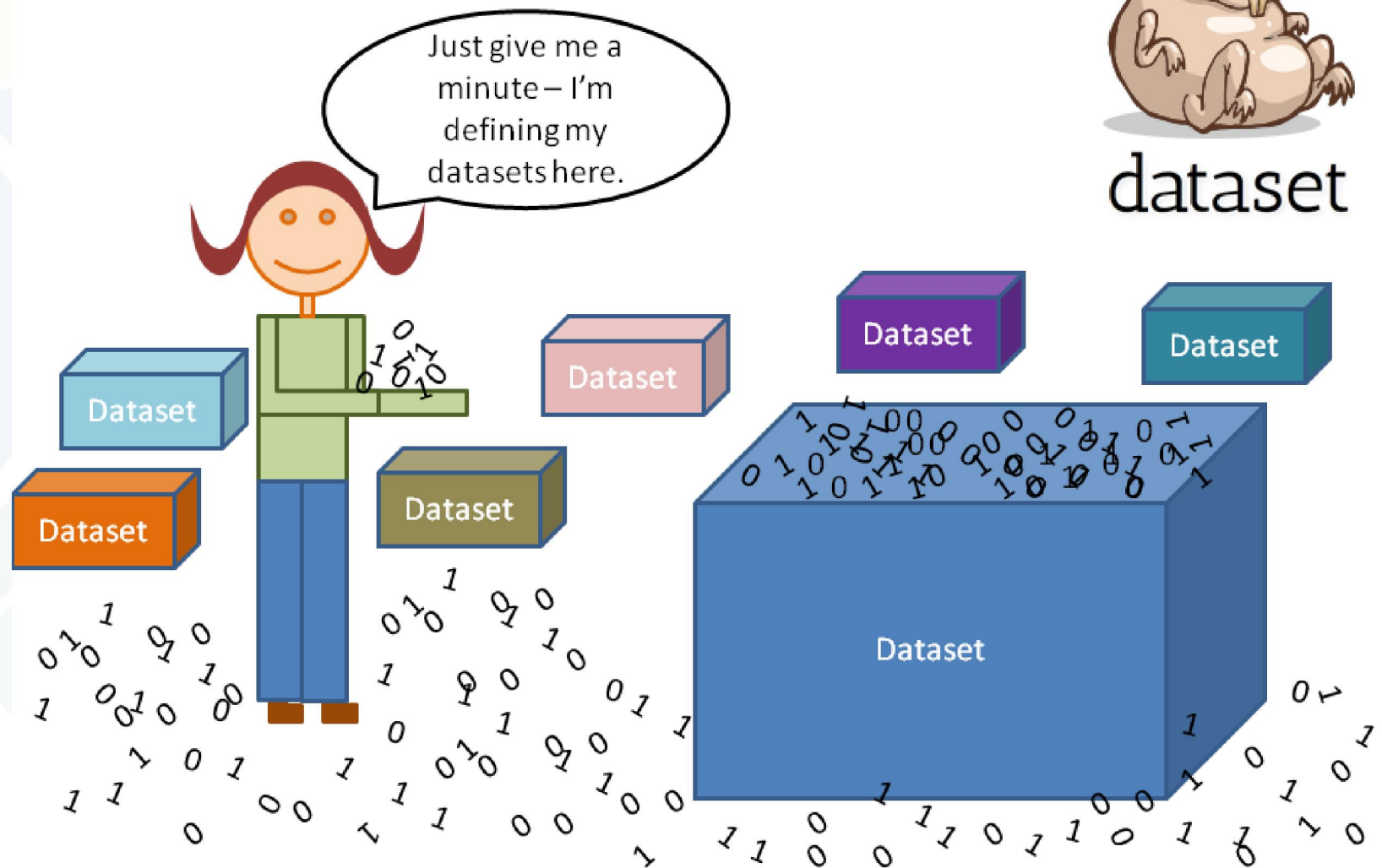


```
a <- available.packages()
head(rownames(a), 3)    ## Show the names of the first few packages
```

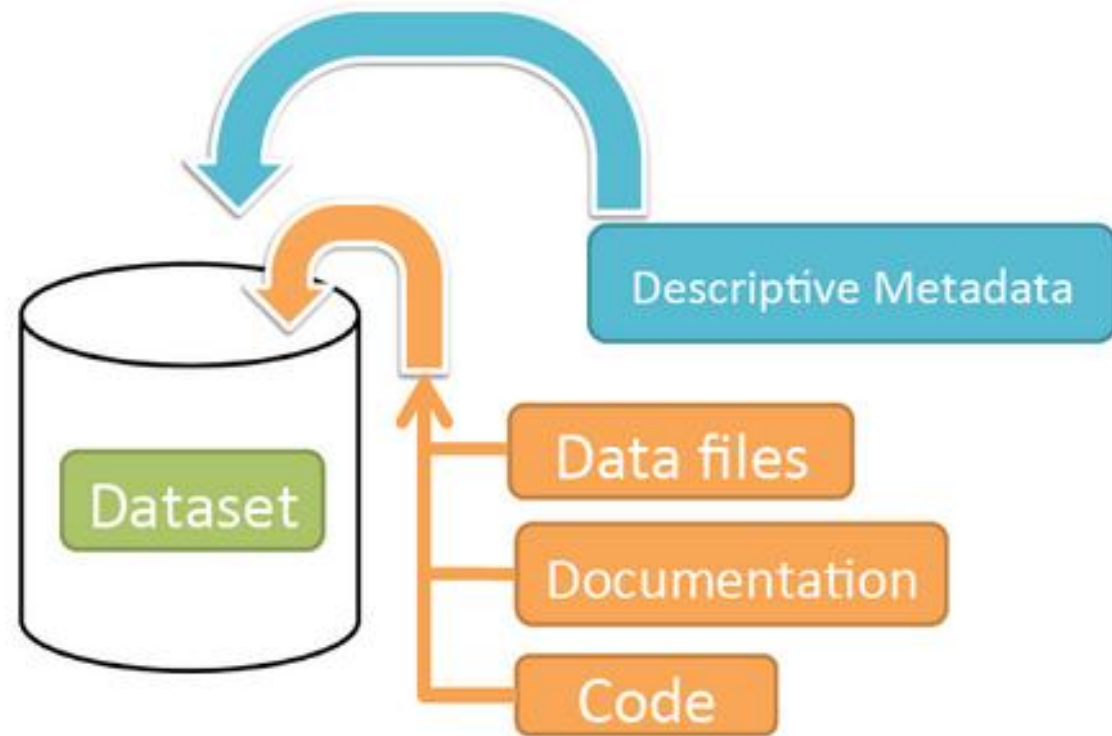
```
## [1] "A3" "abc" "abcdeFBA"
```



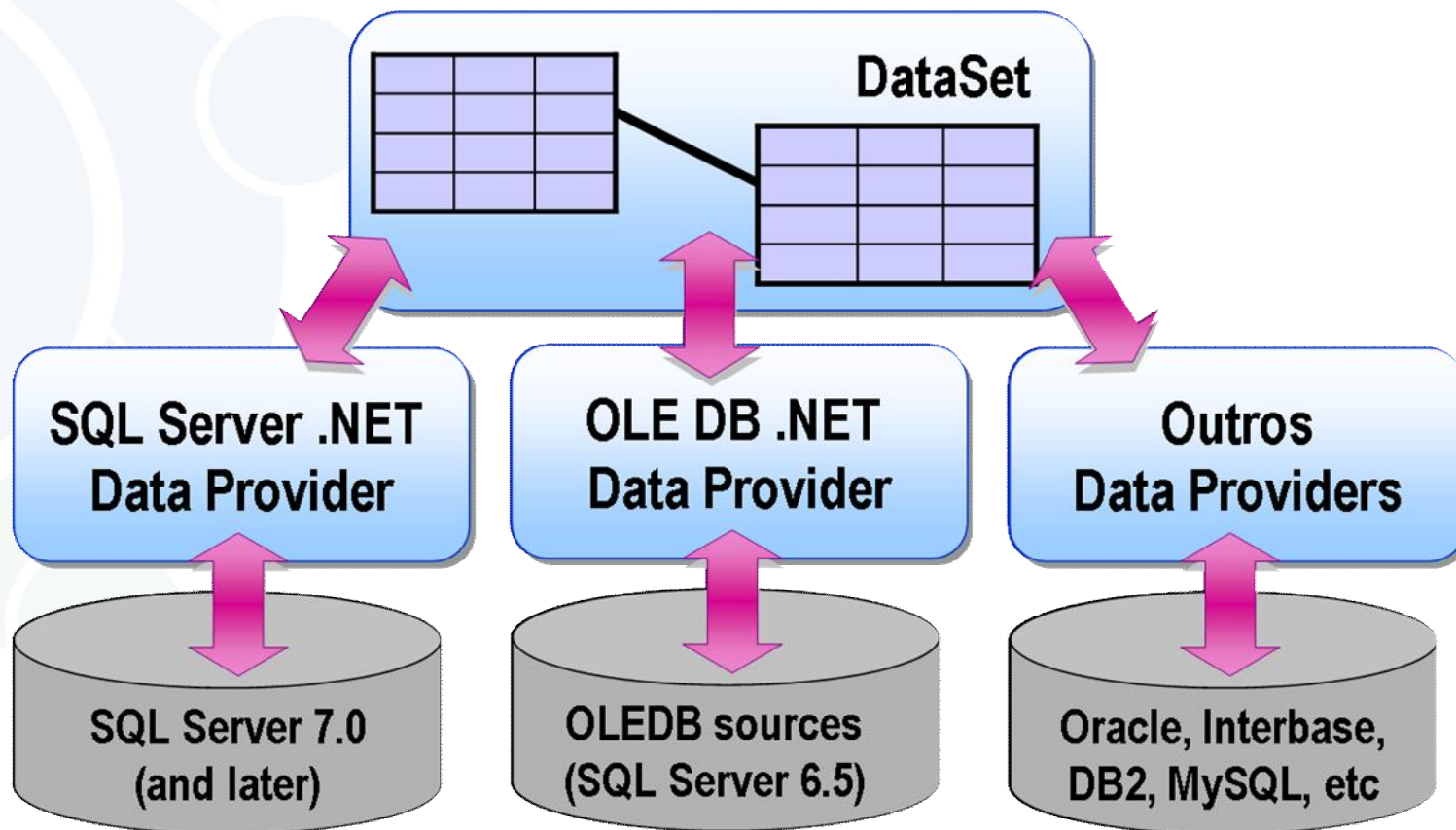

dataset



Schematic Diagram of a **Dataset**



Container for your data, documentation, and code.



Dataset definition

