THE
DEVELOPER'S
CONFERENCE

Trilha – BIG DATA

Data Analysis com R

Rodrigo Ribeiro Gonçalves

Globalcode

Rodrigo Ribeiro Gonçalves

THE DEVELOPER'S CONFERENCE

- **Atualmente**
  - Administrador de Banco de Dados
  - Professor Universitário

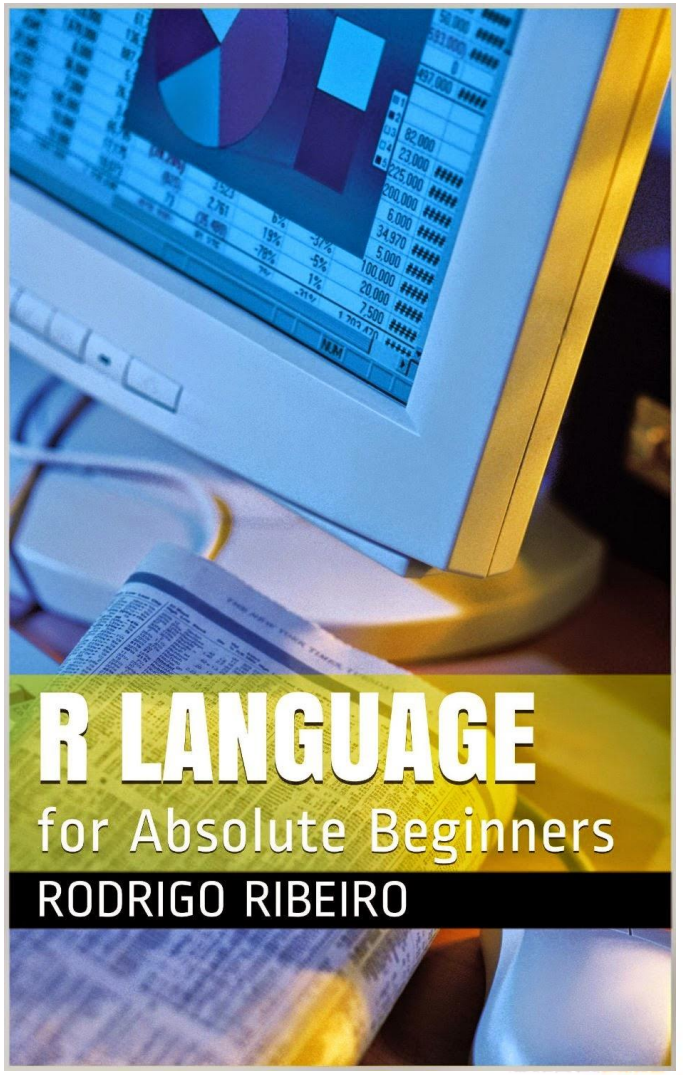
- Email: dbconsultoria@gmail.com

Globalcode

Rodrigo Ribeiro Gonçalves

THE DEVELOPER'S CONFERENCE

- Autor do livro **Linguagem R – Para Iniciantes**
- Autor de outros livros e e-books sobre tecnologia da informação.
- Blog: www.tsqlmaster.blogspot.com



- **Parte Teórica**
  - BIG DATA e Mercado
    - Problema ou solução
  - A Linguagem R
    - O que é e como usar
- **Parte Prática**
  - Usando o R
    - Obtendo dados
    - Analisando Dados
    - Apresentando Resultados
- **Encerramento**
  - Outros recursos
  - Considerações

THE DEVELOPER'S CONFERENCE

Globalcode

THE DEVELOPER'S CONFERENCE

Segundo dados do Gartner Research (2014), mais de 50% dos projetos de Big Data e Analytics falham.

- Estouro de orçamento
- Estouro de prazo
- Má definição de escopo
- Má qualidade da entrega

Fonte: http://www.analytics-magazine.org/july-august-2014/1074-the-data-economy-why-do-so-many-analytics-projects-fail

Globalcode

THE DEVELOPER'S CONFERENCE

Falta de apoio da parte estratégica da empresa.

Globalcode

Por quê?

THE DEVELOPER'S CONFERENCE

- Analytics envolve olhar para o passado... rever desenvolvimentos já realizados (resistência).
- Para acessar os dados é necessário quebrar/dobrar as barreiras de segurança da informação pré-estabelecidas.
- Interagir com pessoas que estão envolvidas em outros projetos ou que não tem interesse em participar nos projetos de Analytics.

Globalcode

Desafios

THE DEVELOPER'S CONFERENCE

- Descolamento entre expectativa e realidade, devido a pouca cultura de Analytics (Analfabetismo de Dados).

- Um pouco de confusão em relação ao termo BIG DATA e os resultados que podem ser obtidos através de um projeto de BIG DATA e BIG DATA ANALYTICS.

Globalcode

BIG DATA, ANALYTICS e DATA SCIENCE

THE DEVELOPER'S CONFERENCE

- BIG DATA e DATA ANALYTICS **NÃO** é um produto ou um computador
- BIG DATA e DATA ANALYTICS é uma estratégia, visão e arquitetura em constante evolução.
- BIG DATA e DATA ANALYTICS envolve pessoas (Cientistas de dados, Analista de Dados, Analista de Requisitos, DBAs, Infra-estrutura, Investimento em treinamentos, capacitação, software, etc...)

Globalcode

Linguagem R

THE DEVELOPER'S CONFERENCE

Por ser uma tecnologia voltada para estatística, se encaixa bem no nicho de DATA SCIENCE.

- **Data Analysis**
- **Machine Learning**
- **Text Mining**
- **Estatística**
- **Data Visualization**
  - **Gráficos, histogramas... etc**

Globalcode

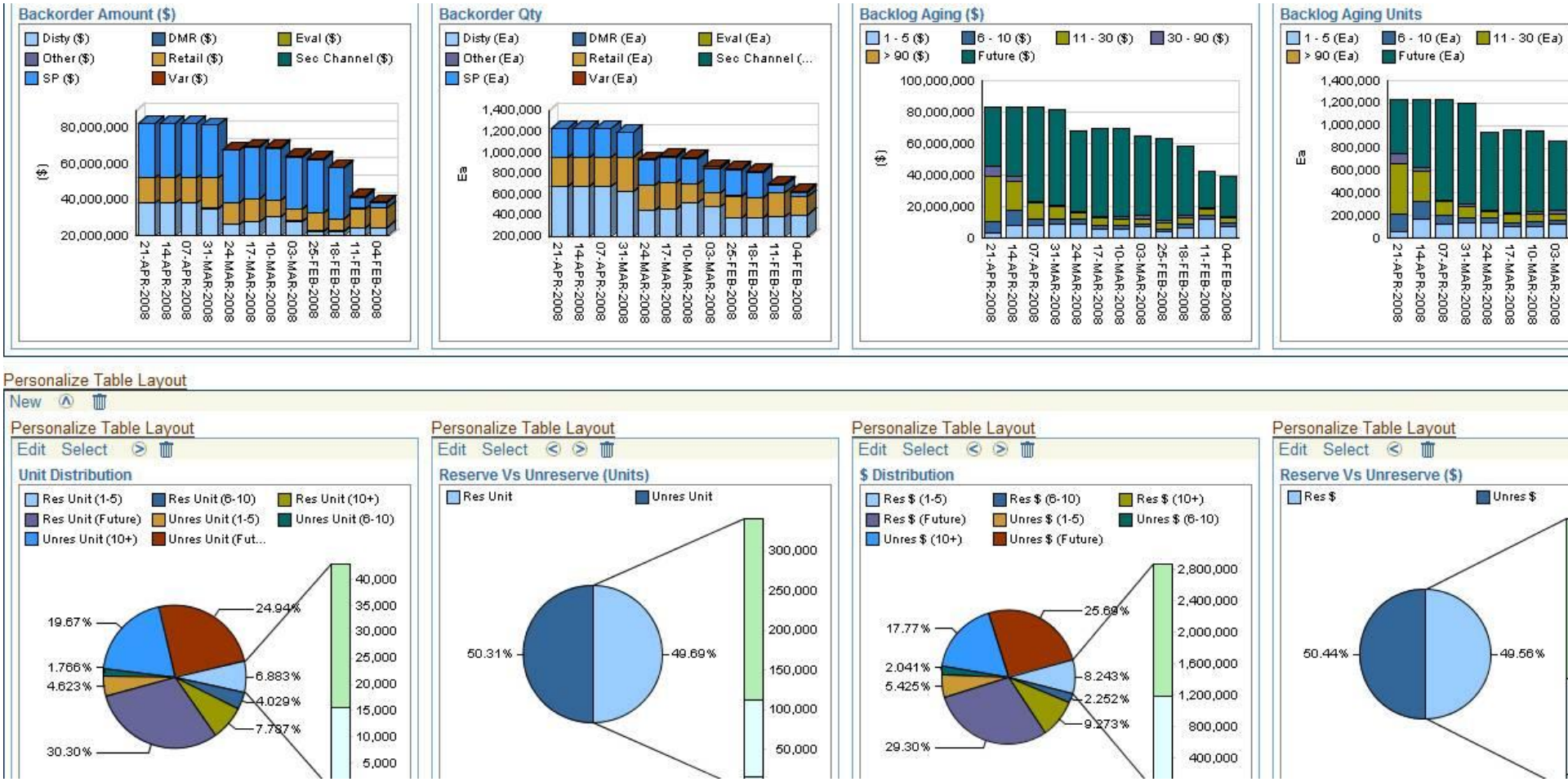
R pode ser usado para BI?

THE DEVELOPER'S CONFERENCE

- **Business Intelligence usa estatística descritiva.**
  - **Foco na visualização e SUMARIZAÇÃO!**

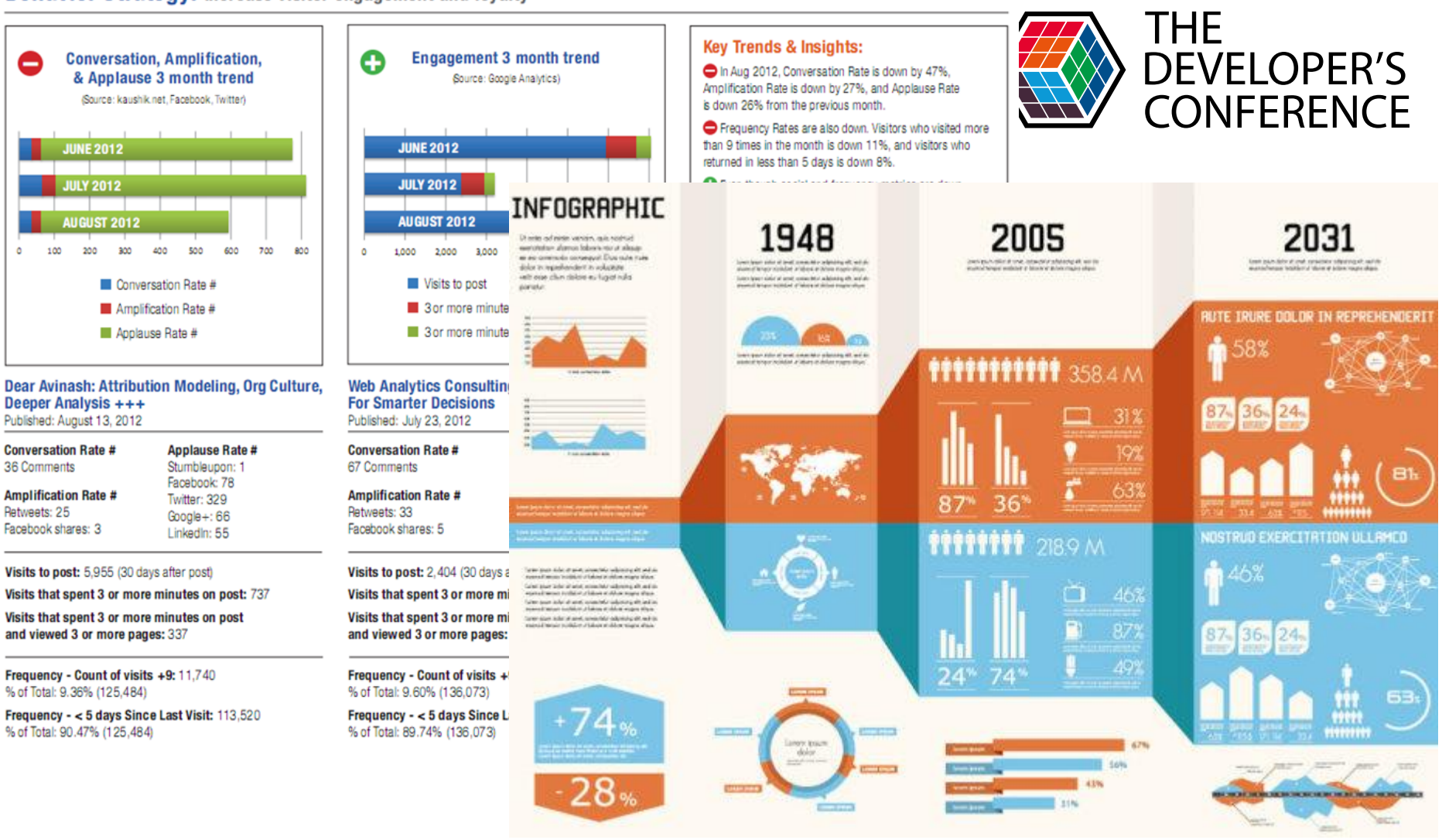
- **BIG DATA e DATA SCIENCE usam estatística dedutiva para analisar e realizar previsões de resultados e de comportamentos**
  - **Foco nos INSIGHTS!**

Globalcode



- **Típico dashboard de BI!**

Globalcode



- **Relatório de Data Science: Não apenas dados e gráficos, mas também INSIGHTS!**

Linguagem R

THE DEVELOPER'S CONFERENCE

- **É uma linguagem e um ambiente de desenvolvimento.**
- **R é uma evolução da linguagem S, desenvolvida dentro do BELL LABS.**
- **É um software FREE e OPEN SOURCE, distribuído pela GPL Versão 2**
- **Pode ser obtido no repositório do projeto: http://cran.r-project.org (Comprehensive R Archive Network)**
- **Lançada publicamente em 1993**
- **Criado para trabalhar com grandes DATASETS**

Globalcode

REVOLUTION ANALYTICS

THE DEVELOPER'S CONFERENCE

Revolution R OPEN

- **Uma distribuição alternativa do R com *features* adicionais, além de total compatibilidade com o R "original".**

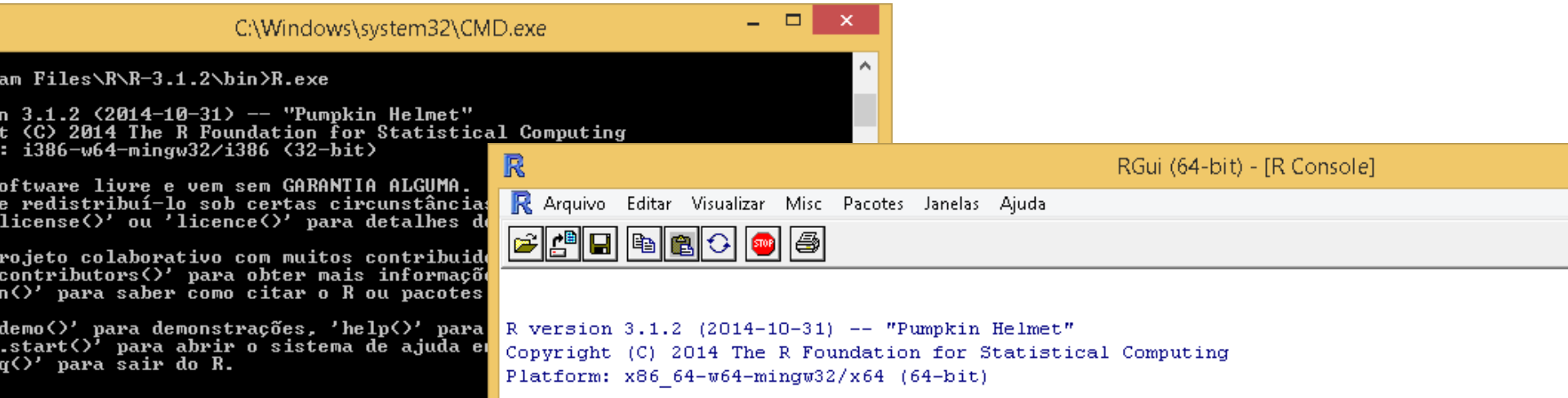
- **A empresa Revolution Analytics acabou de ser comprada pela Microsoft! Fonte**

Globalcode

Linguagem R

THE DEVELOPER'S CONFERENCE

- **O R pode ser instalado no Windows, Linux e Mac com versões 32 e 64 bits**
- **Depois de instalado, pode ser usado através de uma linha de comando ou através de uma GUI instalada junto com o ambiente de desenvolvimento**

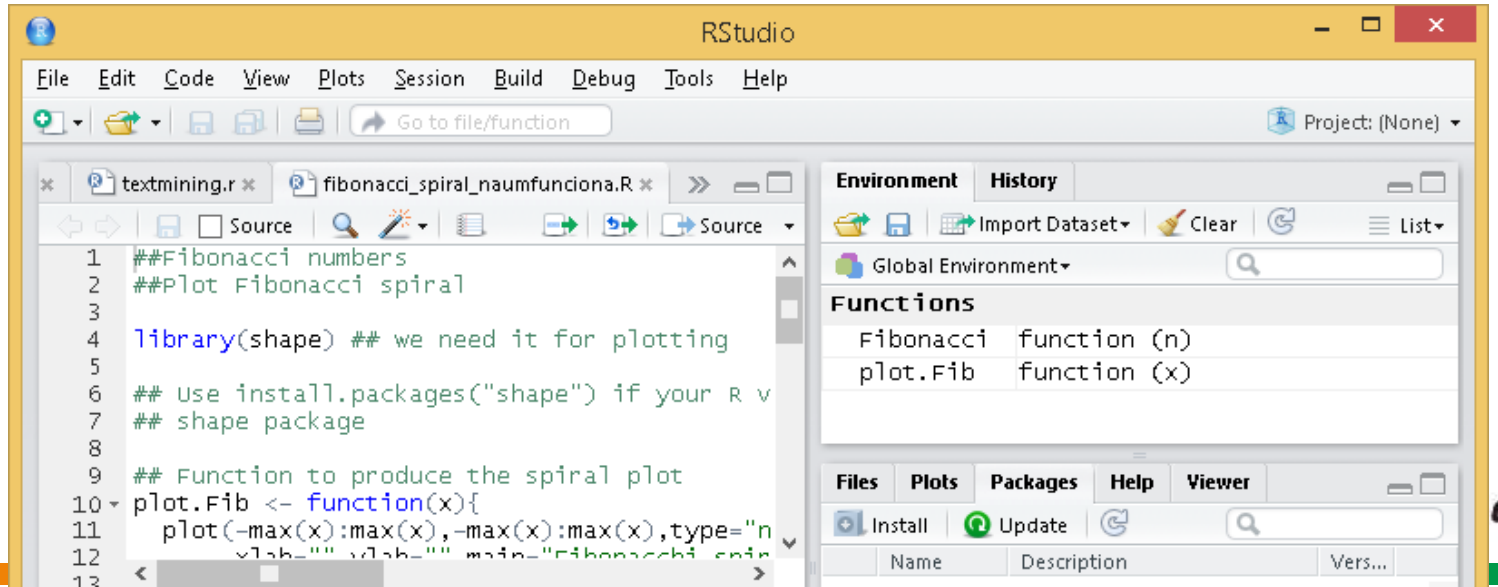


R STUDIO?

THE DEVELOPER'S CONFERENCE

- **O R também pode ser usado através de uma interface também FREE, chamada de RSTUDIO distribuído pela AGPL V3**

- **WWW.RSTUDIO.COM**



Globalcode

Extensões

THE DEVELOPER'S CONFERENCE

- **É uma linguagem bastante extensível através de pacotes desenvolvidos pela comunidade, vantagem que outras tecnologias não tem.**

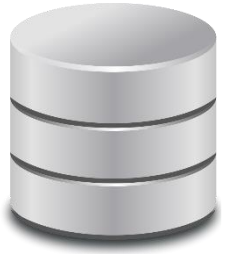
- **http://cran.r-project.org**

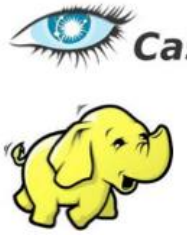
Globalcode

Linguagem R

THE DEVELOPER'S CONFERENCE

- **Linguagem de SCRIPT interpretada**
- **Linguagem Orientada a Objetos**
- **Pode incorporar com relativa facilidade, dados de diferentes origens, como arquivo txt, excel, web, banco de dados relacionais, NoSQL, etc...**





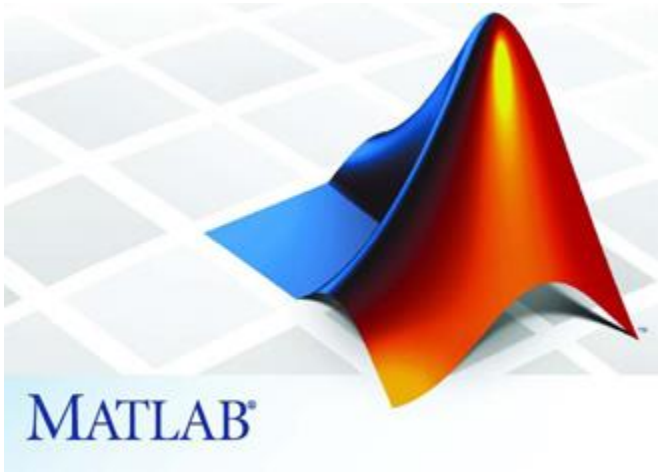
Cassandra

mongoDB

CouchDB
relax

Globalcode

Concorrentes



- **Matlab: U$ 2.650,00**
- **STATA: U$ 1.195,00**

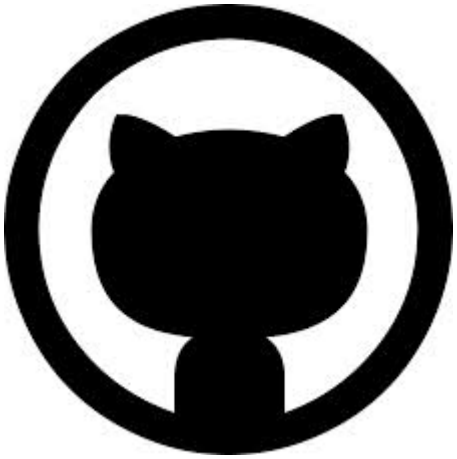
- **Não apenas o licenciamento é diferente...**

Linguagem R

- **Limitação com volumes de dados (até 1 milhão de linhas)**

THE
DEVELOPER'S
CONFERENCE

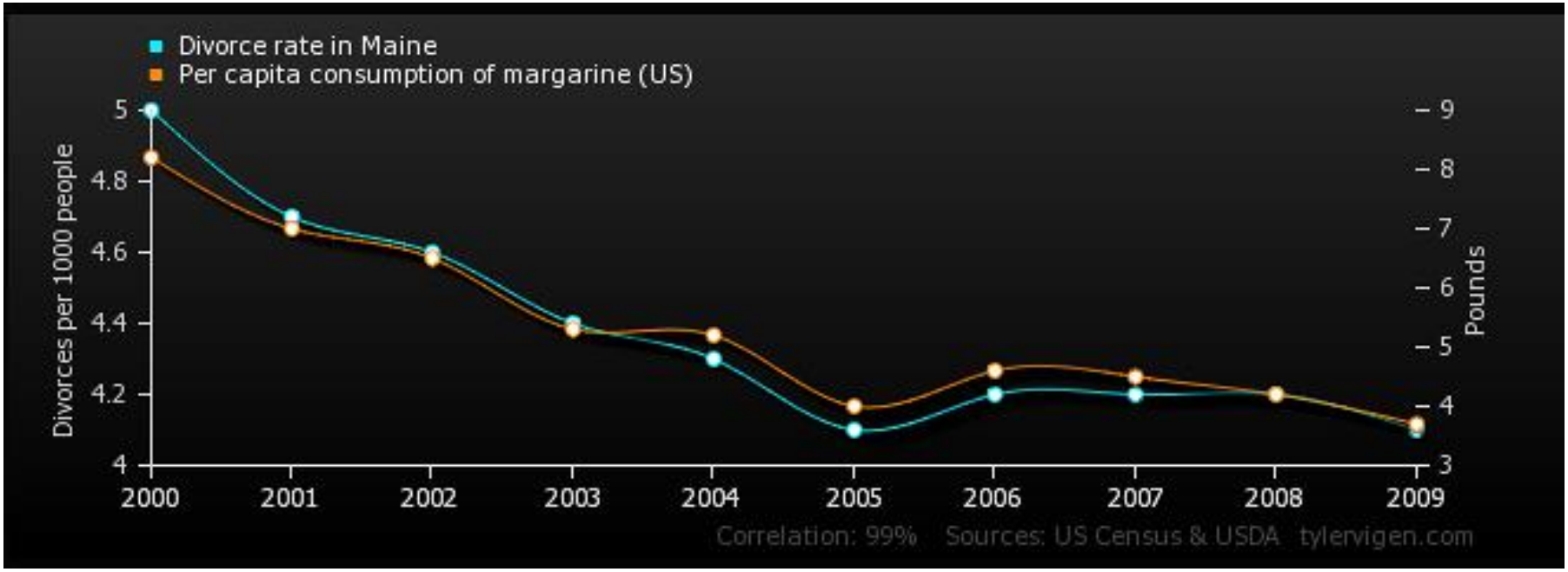
https://github.com/dbconsultoria/tdc2015



Globalcode

Linguagem R

- [“Correlation dosen't imply causation”]



Exemplo 1

THE DEVELOPER'S CONFERENCE

Atribuições e operações básicas.

Globalcode

Exemplo 2

THE DEVELOPER'S CONFERENCE

Vetores e funções úteis.

Globalcode

Exemplo 3

THE DEVELOPER'S CONFERENCE

Data frames.

Globalcode

Exemplo 4

THE DEVELOPER'S CONFERENCE

Importando arquivos CSV.

Globalcode

Exemplo 5

THE DEVELOPER'S CONFERENCE

Importando arquivos CSV da internet.

Globalcode

Exemplo 6

THE DEVELOPER'S CONFERENCE

Instalando pacotes (RMySQL)e conectando com o MySQL.

Globalcode

Exemplo 7

THE DEVELOPER'S CONFERENCE

Gráficos básicos usando o plot.

Globalcode

Exemplo 8

THE DEVELOPER'S CONFERENCE

Gráficos de linha com o Dataset Oranges.

Globalcode

Exemplo 9

THE DEVELOPER'S CONFERENCE

Histogramas e curva normal com o Dataset MTCARS.

Globalcode

Exemplo 10

THE DEVELOPER'S CONFERENCE

Dotplots com o Dataset MTCARS.

Globalcode

Exemplo 11

THE DEVELOPER'S CONFERENCE

PieChart com dados da UCLA.

Globalcode

Outros Recursos

- **R-Bloggers:** www.r-bloggers.com
- **StackOverflow:** http://stackoverflow.com/questions/tagged/r
- **Listas:** r-br@listas.c3sl.ufpr.br

THE
DEVELOPER'S
CONFERENCE

Globalcode

Perguntas e Dúvidas

THE DEVELOPER'S CONFERENCE

Globalcode