



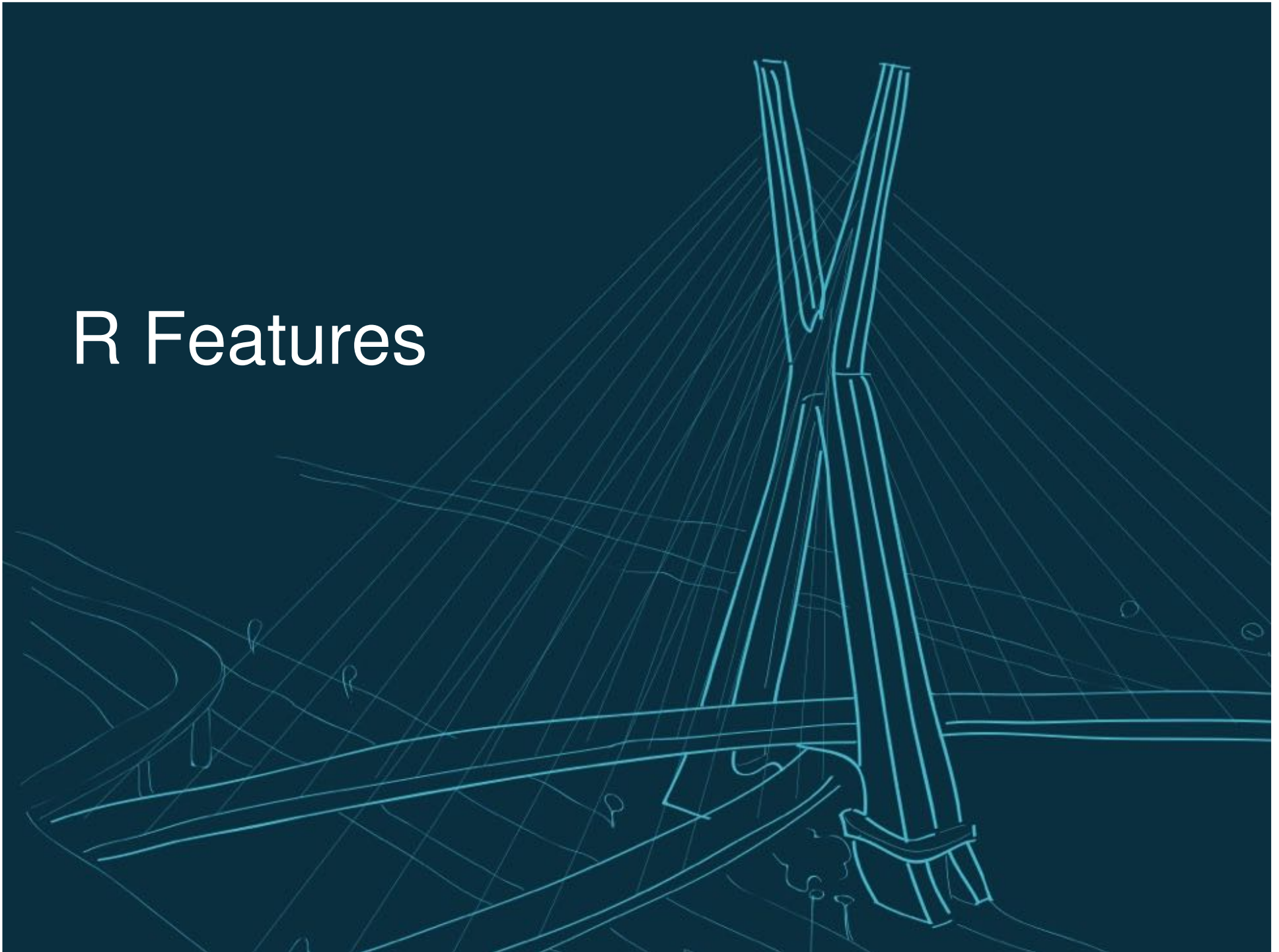
Telefonica

R Tools Evaluation

A review by
Analytics @ Global BI / Local & Regional
Capabilities

Telefónica CCDO
May 2015

R Features



What is ?

- Most widely used data analysis software
 - Used by 2M+ data scientists, statisticians and analysts
- Most powerful statistical programming language
 - Flexible, extensible and comprehensive for productivity
- Create beautiful and unique data visualizations
 - As seen in New York Times, Twitter and Flowing Data
- Thriving open-source community
 - Leading edge of analytics research
- Fills the talent gap
 - New graduates prefer R

Text from



Importance of

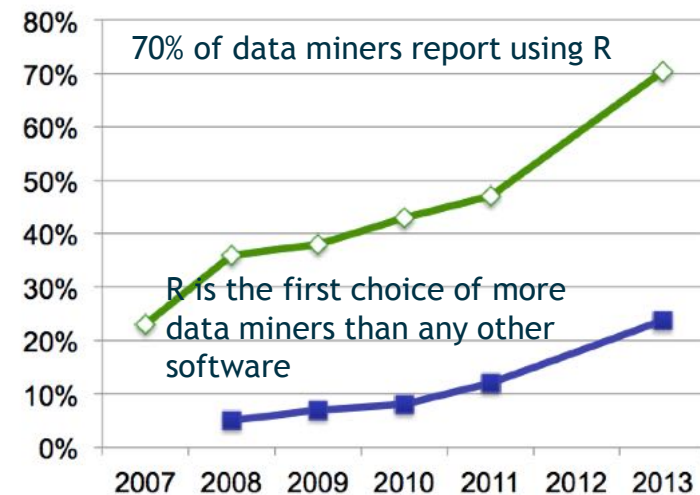
- R is the **highest paid** IT skill
- R **most-used** data science language after SQL
- R is used by **70%** of data miners
- R is **#15** of all programming languages
- R **growing faster** than any other data science language
- R is the **#1 Google Search** for Advanced Analytics software
- R has more than **2 million users** worldwide

Text from



R Usage Growth

Rexer Data Miner Survey, 2007-2013



Source: www.rexeranalytics.com

Data import with

- Data collection (multiple connectors)

- CSV ₍₁₎ Text files delimited or fixed, xml, ₍₂₎ json ... ₍₃₎

- Other analytics formats files (Excel, SPSS, SAS, Stata, Systat ...)

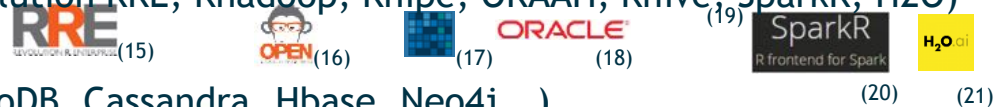


- ODBC/JDBC connectors ₍₉₎ ₍₁₀₎

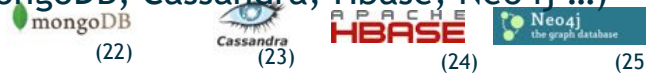
- Native relational database connectors (Oracle, Teradata, SQL Server, MySQL ...)



- Hadoop connectors (Revolution RRE, Rhadoop, Rhipe, ORAAH, Rhive, SparkR, H2O)

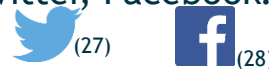


- No SQL connectors (MongoDB, Cassandra, Hbase, Neo4j ...)



- Http (SOA, WS, REST) and ftp connectors ₍₂₆₎

- Social networks connectors (Twitter, Facebook...)



- Other enterprise tools connectors (SAP/R3, Salesforce, Splunk)



() Packages reference, see last slide

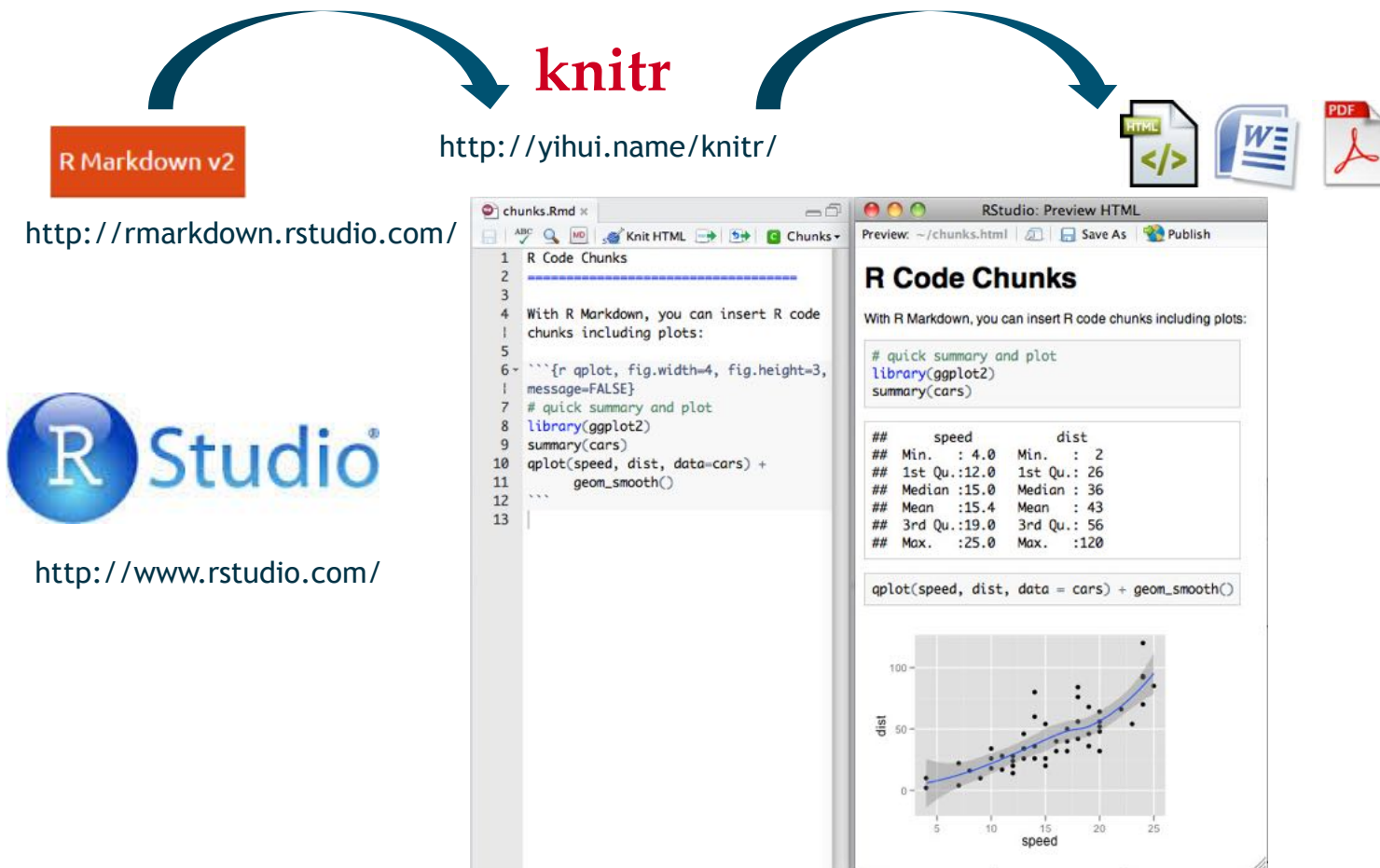
Data preparation with



- Variable creation and transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort
- Merge & Join
- Split
- Aggregate (means, sums)
- Reshape
- ...

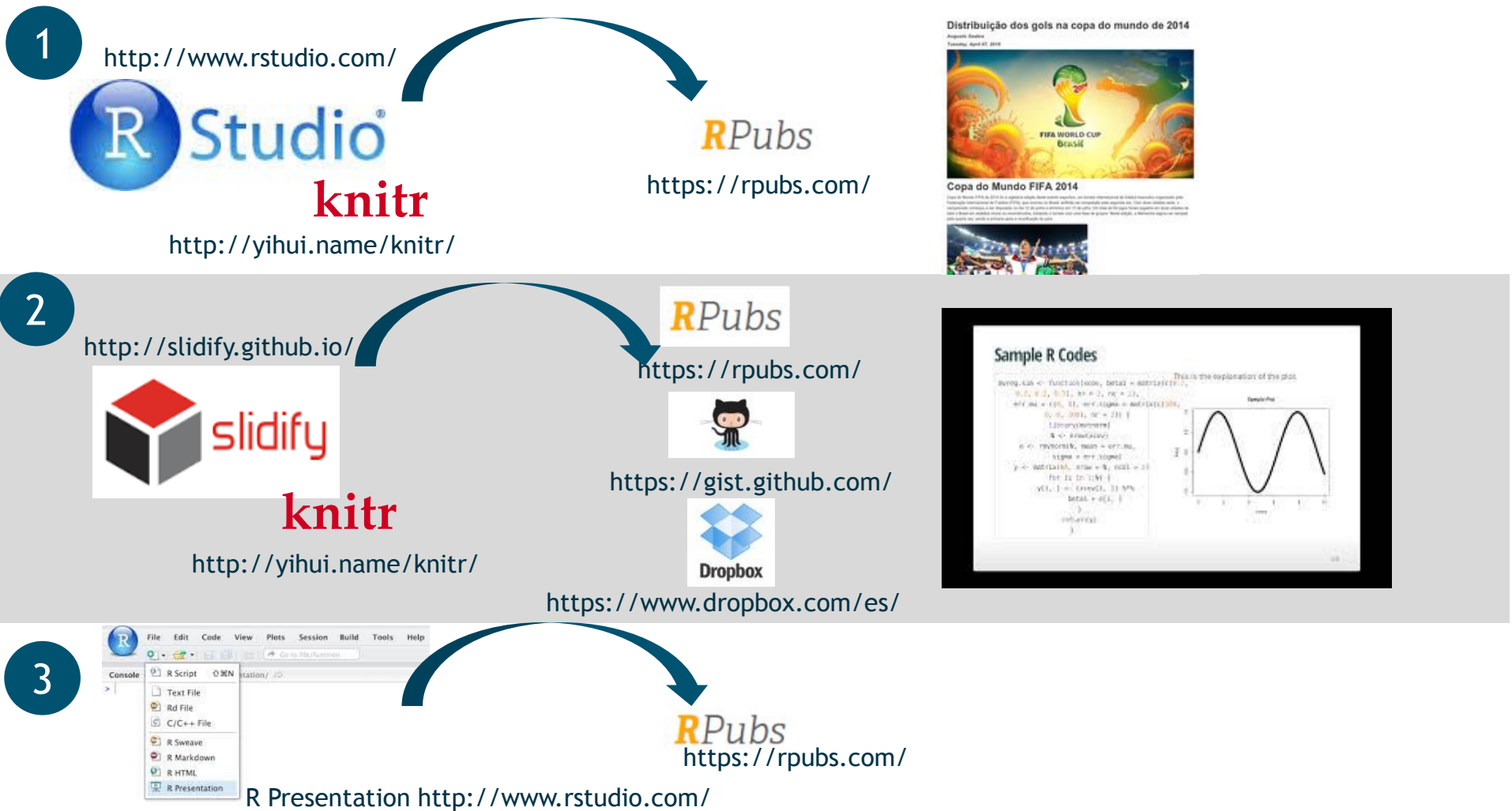
Traditional BI: Reports & Dashboards with

- Reports in Html, MS Word and Pdf with r markdown and knitr
- Very easy way to create reports from r markdown files with RStudio



Traditional BI: Reports & Dashboards with

- The three most known and easiest options to publish reports in R



Discover Analytics with

- Interactive reports

1 
<http://www.rstudio.com/>

2 
<http://www.rstudio.com/>

knitr

<http://yihui.name/knitr/>

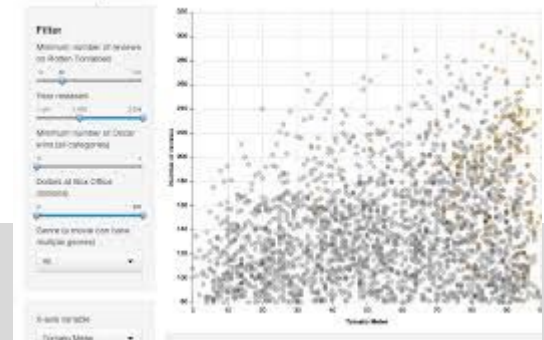


On-premise Shiny Server -
<http://shiny.rstudio.com/>



Cloud Shinyapps.io -
<https://www.shinyapps.io/>

Movie explorer

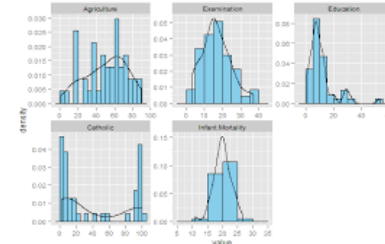


3  **Intuitics**
<https://www.intuitics.com>

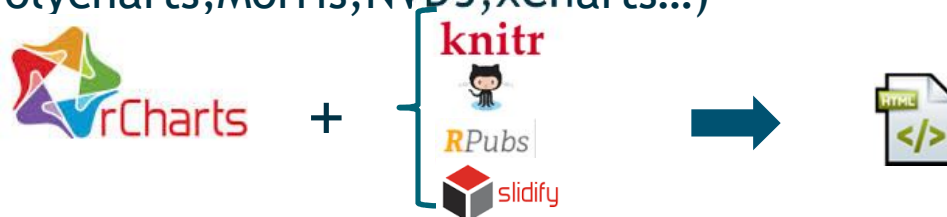


Data Visualizations with

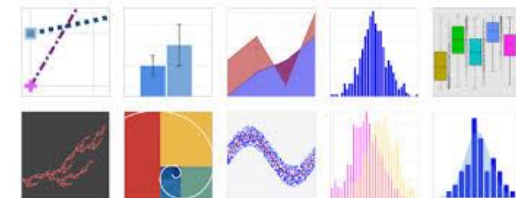
- **ggplot2** (<http://ggplot2.org/>) contains a very complete catalog of visualization widgets (PieChart, BarCharts, Directed/Undirected Graphs, CloudWords, Gauges, Tree Map, Scatter charts...)



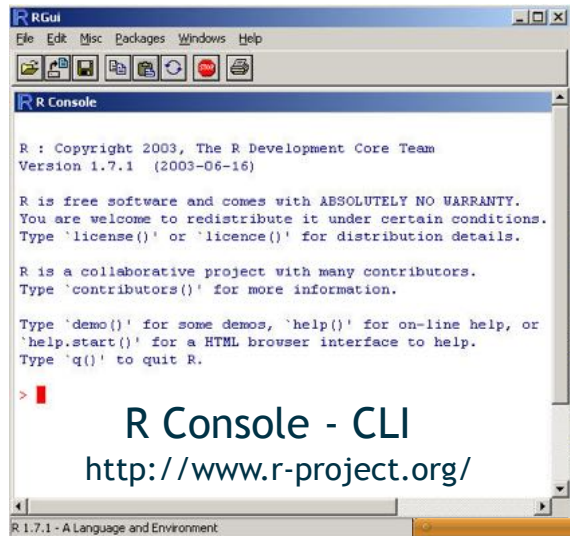
- **Rcharts** (<http://rcharts.io/>) use R to create graphs in html5 by leveraging the most advanced javascript libraries for visualizations (d3js, Polycharts, Morris, NVD3, xCharts...)



- **Plotly** (<https://plot.ly/>) is a platform to create and publish html5 graphs from several programming languages: R, python, matlab, excel...



Predictive Analytics with : Open Source Tools



```
R GUI
File Edit Misc Packages Windows Help
R Console
R : Copyright 2003, The R Development Core Team
Version 1.7.1 (2003-06-16)

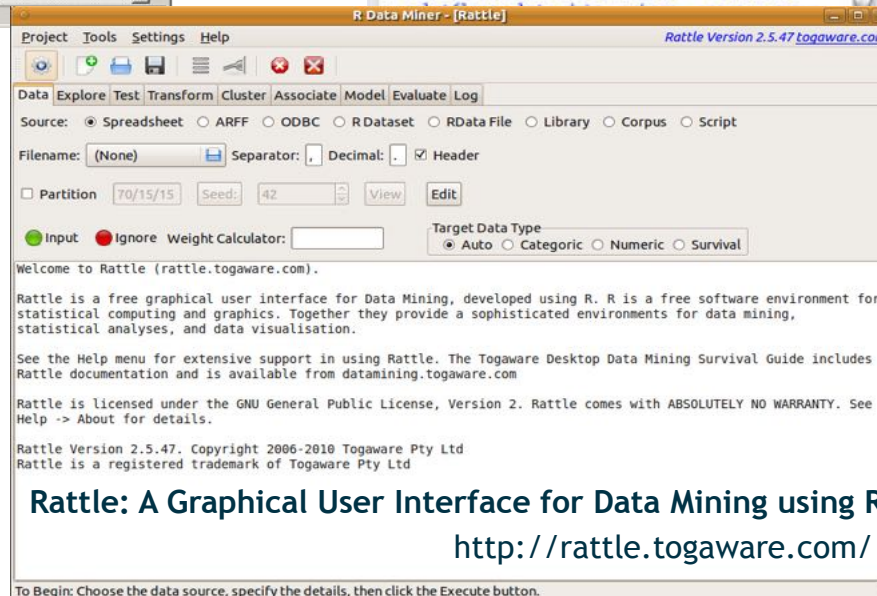
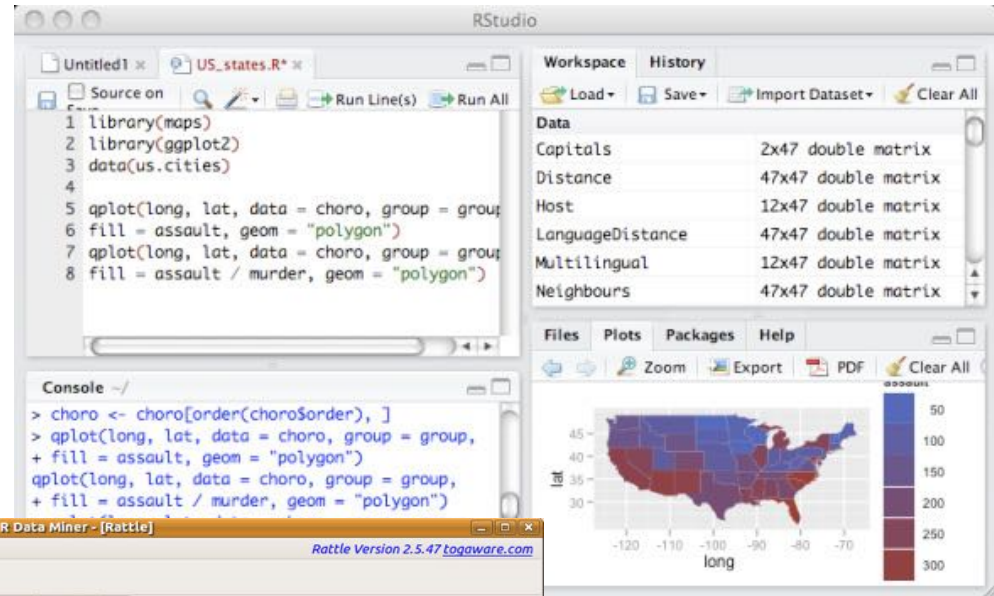
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

>
```

R Console - CLI
<http://www.r-project.org/>



R Data Miner - [Rattle] Rattle Version 2.5.47 togaware.com

Project Tools Settings Help

Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: Spreadsheet ARFF ODBC RDataset RData File Library Corpus Script

Filename: (None) Separator: . Decimal: . Header

Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator: Target Data Type Auto Categorical Numeric Survival

Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environments for data mining, statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The Togaware Desktop Data Mining Survival Guide includes Rattle documentation and is available from datamining.togaware.com

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See Help -> About for details.

Rattle Version 2.5.47. Copyright 2006-2010 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd

Rattle: A Graphical User Interface for Data Mining using R
<http://rattle.togaware.com/>

To Begin: Choose the data source, specify the details, then click the Execute button.



Predictive Analytics with : Packages

- More than 5,000 packages for statistical, predictive analytics and data visualization

Descriptive Statistics



- Min / Max
- Mean
- Median
- Quantiles
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data
- Marginal Summaries of Cross Tabulations

Sampling



- Subsample (observations & variables)
- Random Sampling

Variable Selection



- Stepwise Regression
 - Linear
 - Logistic
 - GLM

Predictive & Classification



- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM)
 - All exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions including: cauchy, identity, log, logit, probit
- Covariance Matrix
- Correlation Matrix
- Logistic Regression
- Classification & Regression Trees
- Residuals for all models
- Decision Trees
- Decision Forests
- Boosted Decision Trees

Text and figures from



Cluster Analysis



- K-Means
- Hierarchical
- Model Based

Deployment



- Prediction (scoring)
- PMML Export

As a Service

In Cloud

- <https://www.elasticr.com>
- <http://www.ebi.ac.uk/Tools/rcloud/>
- AWS http://www.louisaslett.com/RStudio_AMI
- <http://azure.microsoft.com/en-us/documentation/articles/machine-learning-r-csharp-web-service-examples>
- <https://api.blockspring.com/docs/r-quickstart-rur>



On Premise

- <http://www.openanalytics.eu/architect-server>
- <https://www.opencpu.org> (*)
- <http://www.rforge.net/Rserve>
- <http://www.rforge.net/FastRWeb>
- <http://sysbio.mrc-bsu.cam.ac.uk/Rwui>
- <http://www.math.montana.edu/Rweb>



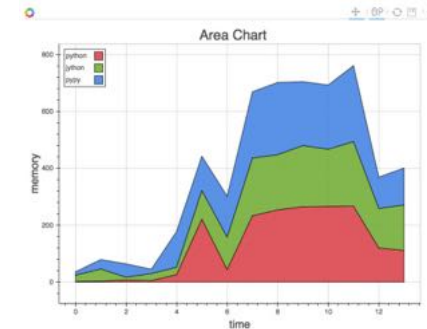
(*) It could be run in Amazon EC2 too

Data Visualizations with

- Rbokeh (<http://hafen.github.io/rbokeh>) use R to create graphs in html5/d3js



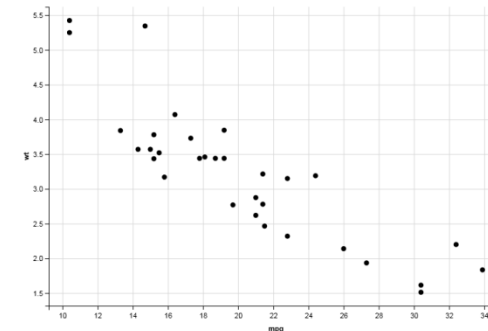
+ knitr



- ggvis (<http://ggvis.rstudio.com/>) is a data visualization package for R using Vega, a javascript html5 library

ggvis

+



R & BIG DATA



Limitations of for enterprises

- Big Data → In-memory bound for many use cases
- Speed of Analysis → Single threaded by design
- Enterprise Readiness → Community support
- Analytic Breadth & Depth → 5700+ innovative analytic packages
- Commercial Viability → Risk of deployment of open source

Hadoop processing modes with

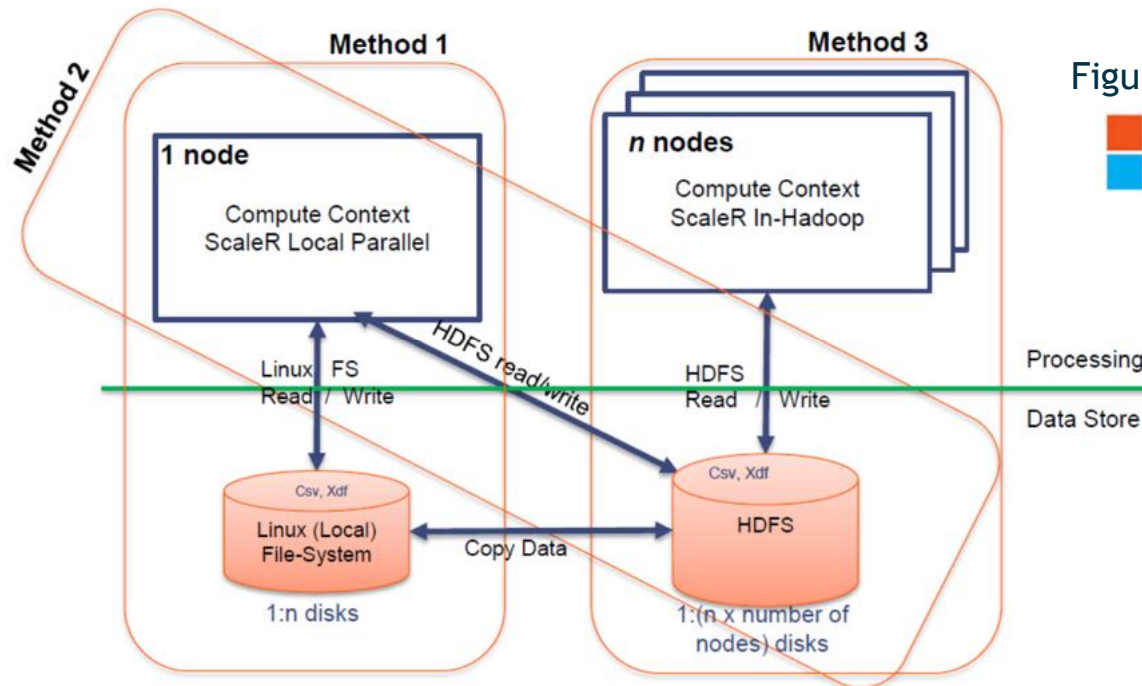


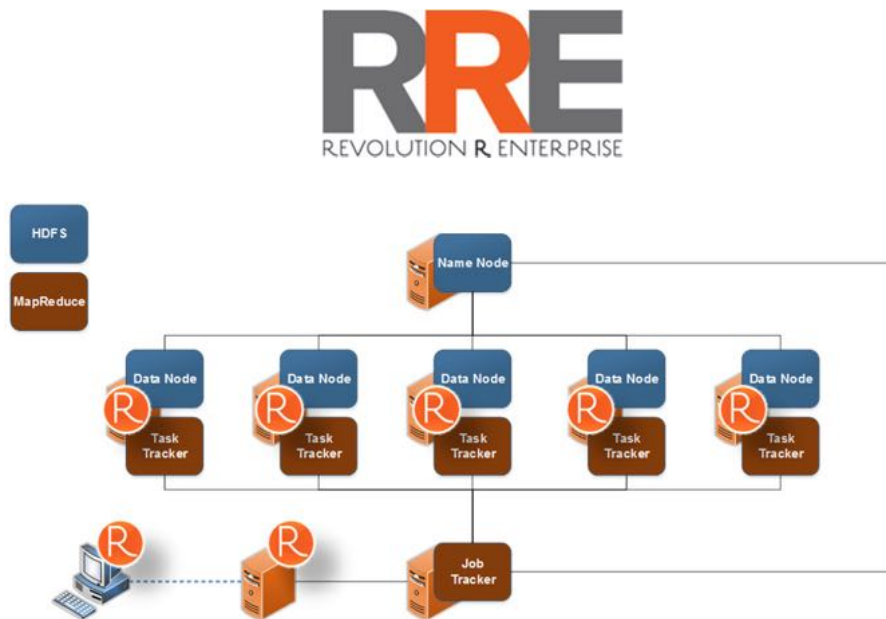
Figure from



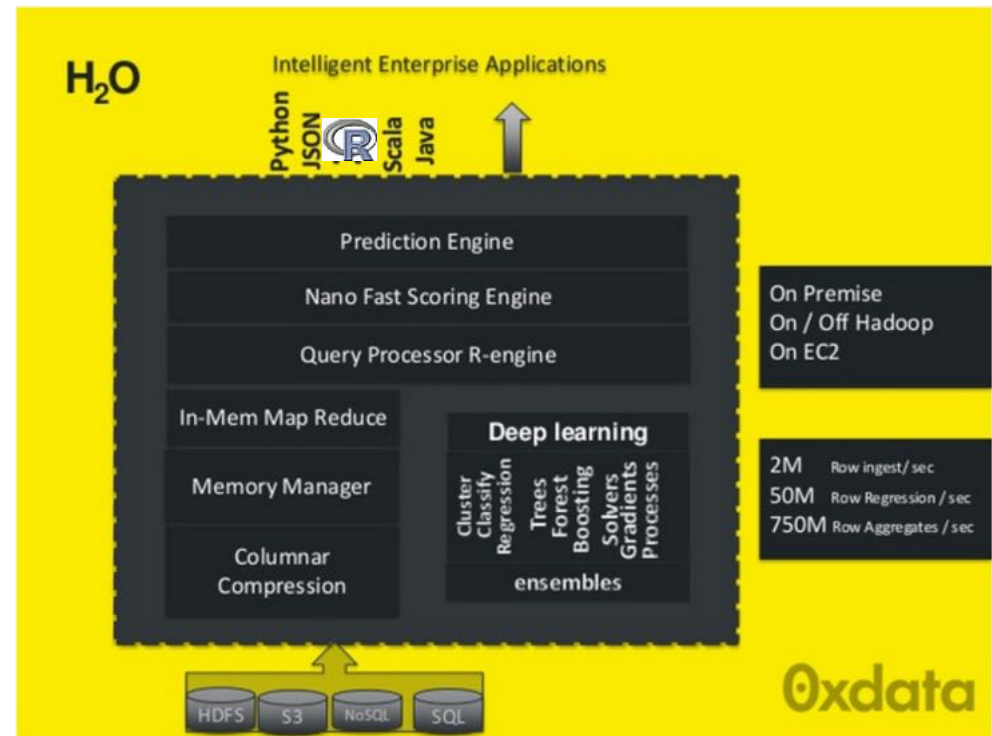
- **Method 1:** Local parallel processing using all cores on one node, using local linuxfile-system data
 - Revolution Analytics parallelR (<http://projects.revolutionanalytics.com/documents/parallelr/parallerrpks/>)
- **Method 2:** Local parallel processing using all cores on one node, reading from / to HDFS data
 - Revolution Rhadoop (<https://github.com/RevolutionAnalytics/RHadoop/wiki>), RHIPE (<https://www.datadr.org/>), ORAAH (Oracle R Advanced Analytics for Hadoop) or package RHIVE (<http://cran.r-project.org/web/packages/RHive/RHive.pdf>)
 - Revolution Analytics parallelR (<http://projects.revolutionanalytics.com/documents/parallelr/parallerrpks/>)

Hadoop processing modes with

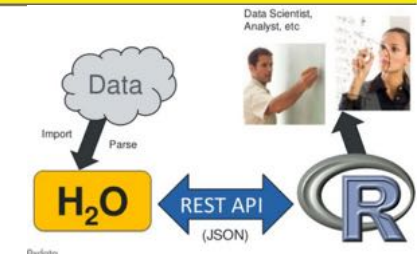
- Method 3: Hadoop (Map-Reduce) parallel processing using all cores on n nodes, using HDFS data in-situ



Commercial Tool



Open Source Tool



BD Analytic Tools

Strenghts

- Most widely used data analysis and predictive software in the world
- A lot of packages (5000+) to do almost everything you want, kept by a huge developers community
- Completely free
- Integration with a great amount of tools (free and commercial)
- Multiple connectors to get a lot of type of data
- Not only for analytics, good to data discover and reporting too

Weaknesses

- More difficult to learn than other software
- Help files are written for relatively advanced users
- R holds all its data in your computer's main memory. There are free and commercial tools to parallelize R but not too many alternatives
- Because the great amount of packages it is often difficult finding and choosing the better ones
- R core is quite stable, but sometimes some package changes and dependencies are not updated
- Integration with web apps is not mature

Packages & Projects Reference (<http://crantastic.org/> or <http://cran.r-project.org/web/packages/>)

Data Access RForcecom github.com/rfsp/r
 RJDBC₍₁₀₎ RCurl₍₂₆₎ RSAP₍₂₉₎
 RODBC yhatr₍₂₆₎ XML₍₂₎
 sqldf₍₉₎ RHive₍₁₉₎ twitterR₍₂₇₎ rjson₍₃₎
 ROracle₍₁₁₎ foreign Rfacebook rmongodb dplyr
 RSQLServer₍₁₃₎ RCassandra tidyr
 xlsx github.com/nicolewhite/RNeo4j
 RMySQL₍₁₄₎ Hmisc rPython datadr.org
 RPostgresSQL rJava
 github.com/RevolutionAnalytics/RHadoop/wiki
 amplab-extras.github.io/SparkR-pkg

Reporting & Discover
 rpubs.com manipulate
 rstudio.com shinyapps.io
 intuitics.com
 slidify.github.io rcharts.io
 ggvis.rstudio.com plot.ly/r
 ggplot2.org yihui.name/knitr
 github.com/Bart6114/scheduleR
 maps sp mapdata mapproj

Predictive
 rstudio.com
 care rattle.togaware.com
 caret topepo.github.io/caret
 pvclust mclust yhatr
 neuralnet opencpu
 ga tm
 maps sp mapdata mapproj

Telefonica
