Microsoft

R Session

Julia Jauß
Technical Evangelist
julia.jauss@microsoft.com
@blaujule

April 15, 2016

Agenda

- What is R?
- Applications of R
- Revolution Analytics
- Microsoft and R(evolution Analytics)
- Demo: R and Azure Machine Learning
- Demo: R Tools for Visual Studio

WHAT IS R?

: What Is It?

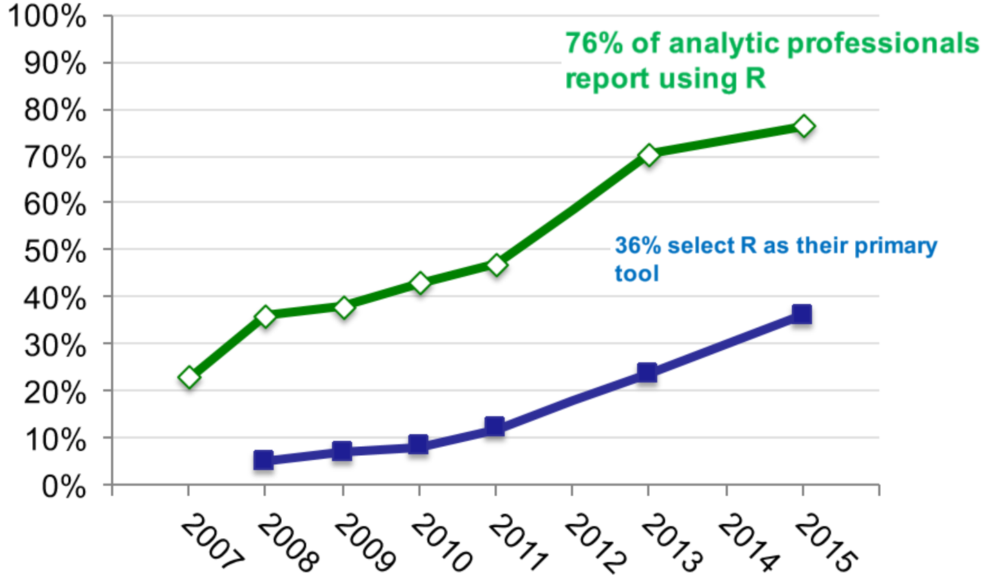
A Language Platform…

A Procedural Language optimized for Statistics and Data Science
A Data Visualization Framework
Provided as Open Source

R's popularity is growing rapidly

R Usage Growth

Rexer Data Miner Survey, 2007-2015

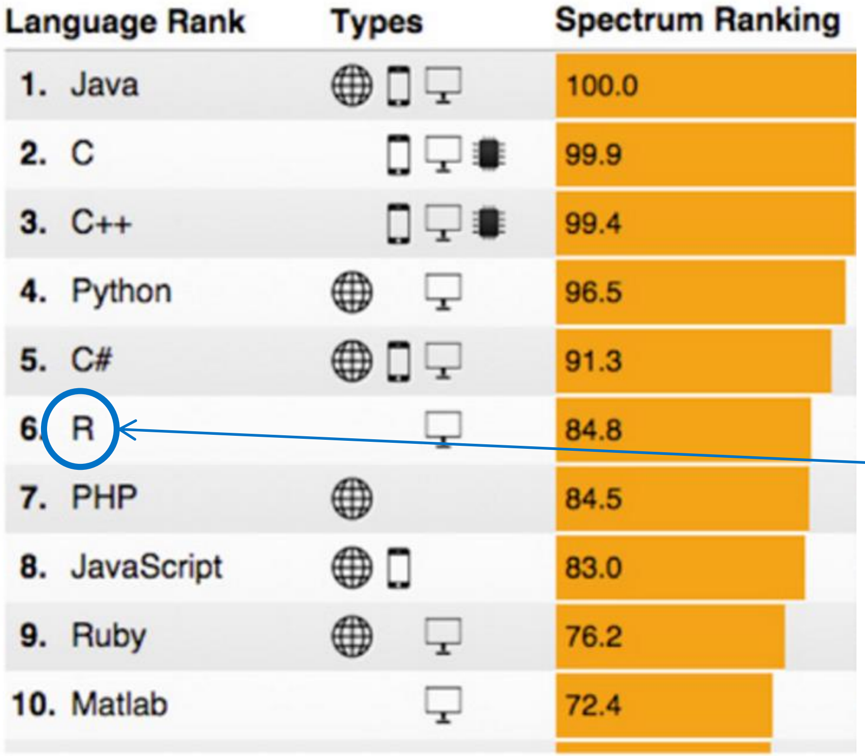


76% of analytic professionals report using R

36% select R as their primary tool

Language Popularity

IEEE Spectrum Top Programming Languages 2015



Language Rank	Types	Spectrum Ranking
1. Java	🌐 📱 🖥	100.0
2. C	📱 🖥 ▪	99.9
3. C++	📱 🖥 ▪	99.4
4. Python	🌐 🖥	96.5
5. C#	🌐 📱 🖥	91.3
6. R	🖥	84.8
7. PHP	🌐	84.5
8. JavaScript	🌐 📱	83.0
9. Ruby	🌐 🖥	76.2
10. Matlab	🖥	72.4

#6: R

R: What Is It?

A Language Platform…

A Procedural Language optimized for Statistics and Data Science

A Data Visualization Framework

Provided as Open Source

A Community…

2.5M+ Statistical Analysis and Machine Learning Users

- Taught in Most University Statistics Programs
- Active User Groups Across the World

Community Resources

R Project websites

www.r-project.org ; cran.r-project.org

Find the best R package to solve a problem:

- MRAN (mran.revolutionanalytics.com)

Get your R question answered:

- Stackoverflow (R tag)

Read R blogs

- Revolutions blog (blog.revolutionanalytics.com)
- R-bloggers (r-bloggers.org)

R user discussions

- #rstats hashtag on Twitter

R user groups and events

Revolution Analytics Supports the

Open Source Community

Supporting 51 Groups Worldwide

Revolution Supported Conferences

BioC2013
July 18-19, 2013 (Developer Day: July 17)
Fred Hutchinson Cancer Research Center
Seattle, WA

Deuxièmes Rencontres R
June 27-28, 2013
Lyon, Campus de la Doua
http://r2013-lyon.sciencesconf.org

useR! 2013
The R User Conference 2013
July 10-12 2013
University of Castilla-La Mancha
Albacete, Spain

JSM2013
August 3-8, 2013
Montréal, Québec

Meielisalp 2013
2013 Meielisalp Workshop & Summer School
June 30 - July 4, 2013
Lake Thune, Switzerland

IV JORNADAS
4th R User Conference
November 15-16, 2012
Barcelona, Spain

WZUR2012
Annual Polish R Users Rally

Contributed Software

Funding R Development

Revolutions Blog:
http://blog.revolutionanalytics.com/

Tracking Community Events:
http://blog.revolutionanalytics.com/calendar.html

Community Website:
http://www.inside-r.org/

R Evangelism in the Media:

Forbes

The New York Times

Revolution Analytics is proud to be a member of The R Foundation
www.revolutionanalytics.com

: What Is It?

A Language Platform…

- A Procedural Language optimized for Statistics and Data Science
- A Data Visualization Framework
- Provided as Open Source

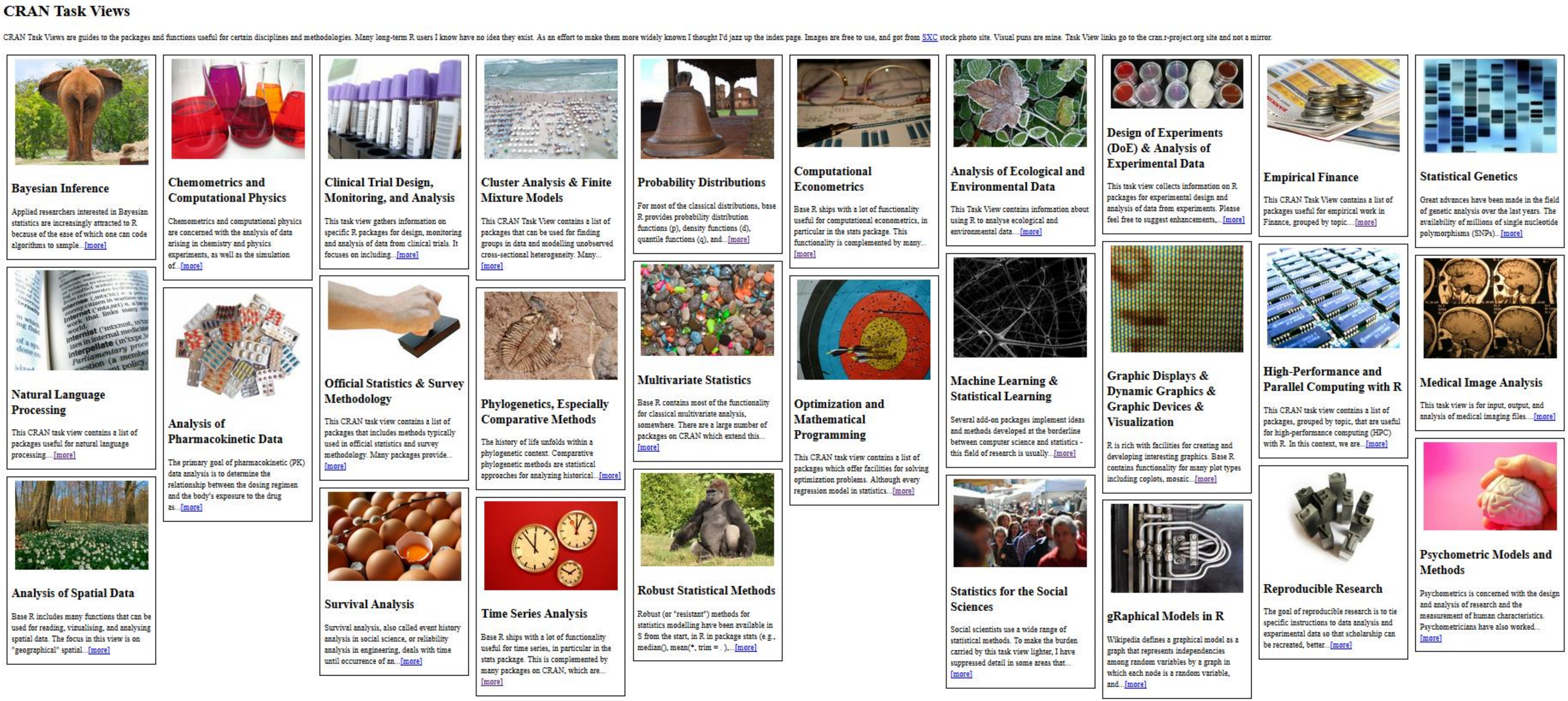
A Community…

- 2.5M+ Statistical Analysis and Machine Learning Users
  - Taught in Most University Statistics Programs
  - Active User Groups Across the World

An Ecosystem

- CRAN: 7000+ Freely Available Algorithms, Test Data and Evaluations
- Many Applicable to Big Data If Scaled

CRAN: Resources For All Fields of



CRAN Task Views

CRAN Task Views are guides to the packages and functions useful for certain disciplines and methodologies. Many long-term R users I know have no idea they exist. As an effort to make them more widely known I thought I'd jazz up the index page. Images are free to use, and got from SXC stock photo site. Visual puns are mine. Task View links go to the cran.r-project.org site and not a mirror.

Bayesian Inference

Applied researchers interested in Bayesian statistics are increasingly attracted to R because of the ease of which one can code algorithms to sample...[more]

Natural Language Processing

This CRAN task view contains a list of packages useful for natural language processing....[more]

Analysis of Spatial Data

Base R includes many functions that can be used for reading, vizualising, and analysing spatial data. The focus in this view is on "geographical" spatial...[more]

Chemometrics and Computational Physics

Chemometrics and computational physics are concerned with the analysis of data arising in chemistry and physics experiments, as well as the simulation of...[more]

Analysis of Pharmacokinetic Data

The primary goal of pharmacokinetic (PK) data analysis is to determine the relationship between the dosing regimen and the body's exposure to the drug as...[more]

Clinical Trial Design, Monitoring, and Analysis

This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including...[more]

Official Statistics & Survey Methodology

This CRAN task view contains a list of packages that includes methods typically used in official statistics and survey methodology. Many packages provide... [more]

Survival Analysis

Survival analysis, also called event history analysis in social science, or reliability analysis in engineering, deals with time until occurrence of an...[more]

Cluster Analysis & Finite Mixture Models

This CRAN Task View contains a list of packages that can be used for finding groups in data and modelling unobserved cross-sectional heterogeneity. Many... [more]

Phylogenetics, Especially Comparative Methods

The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical...[more]

Time Series Analysis

Base R ships with a lot of functionality useful for time series, in particular in the stats package. This is complemented by many packages on CRAN, which are... [more]

Probability Distributions

For most of the classical distributions, base R provides probability distribution functions (p), density functions (d), quantile functions (q), and...[more]

Multivariate Statistics

Base R contains most of the functionality for classical multivariate analysis, somewhere. There are a large number of packages on CRAN which extend this... [more]

Robust Statistical Methods

Robust (or "resistant") methods for statistics modelling have been available in S from the start, in R in package stats (e.g., median(), mean(*, trim = .),...[more]

Computational Econometrics

Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many... [more]

Optimization and Mathematical Programming

This CRAN task view contains a list of packages which offer facilities for solving optimization problems. Although every regression model in statistics...[more]

Analysis of Ecological and Environmental Data

This Task View contains information about using R to analyse ecological and environmental data....[more]

Machine Learning & Statistical Learning

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually...[more]

Statistics for the Social Sciences

Social scientists use a wide range of statistical methods. To make the burden carried by this task view lighter, I have suppressed detail in some areas that... [more]

Design of Experiments (DoE) & Analysis of Experimental Data

This task view collects information on R packages for experimental design and analysis of data from experiments. Please feel free to suggest enhancements,...[more]

Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic...[more]

gRaphical Models in R

Wikipedia defines a graphical model as a graph that represents independencies among random variables by a graph in which each node is a random variable, and...[more]

Empirical Finance

This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic....[more]

High-Performance and Parallel Computing with R

This CRAN task view contains a list of packages, grouped by topic, that are useful for high-performance computing (HPC) with R. In this context, we are...[more]

Reproducible Research

The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better...[more]

Statistical Genetics

Great advances have been made in the field of genetic analysis over the last years. The availability of millions of single nucleotide polymorphisms (SNPs)...[more]

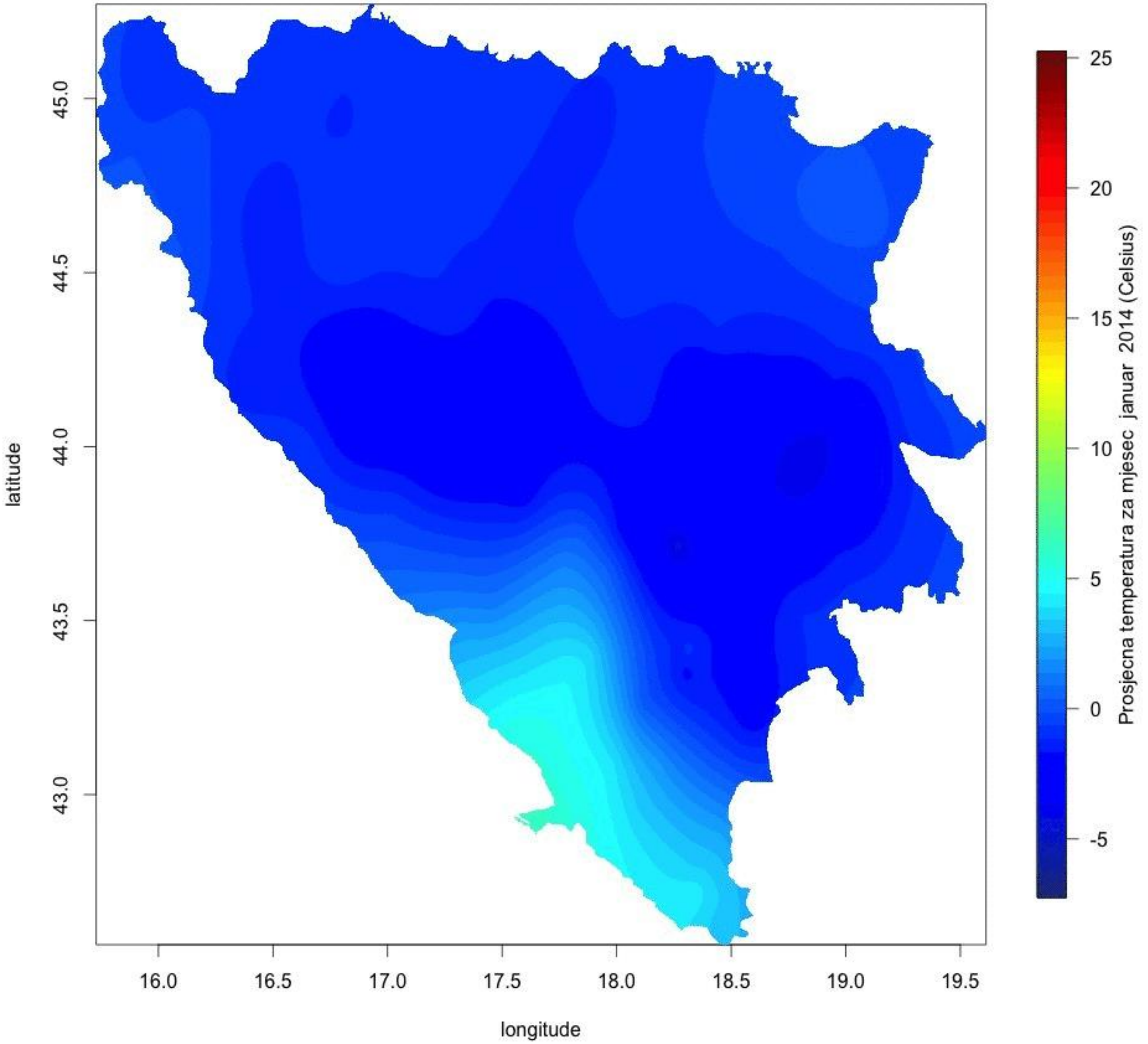
Medical Image Analysis

This task view is for input, output, and analysis of medical imaging files....[more]

Psychometric Models and Methods

Psychometrics is concerned with the design and analysis of research and the measurement of human characteristics. Psychometricians have also worked... [more]

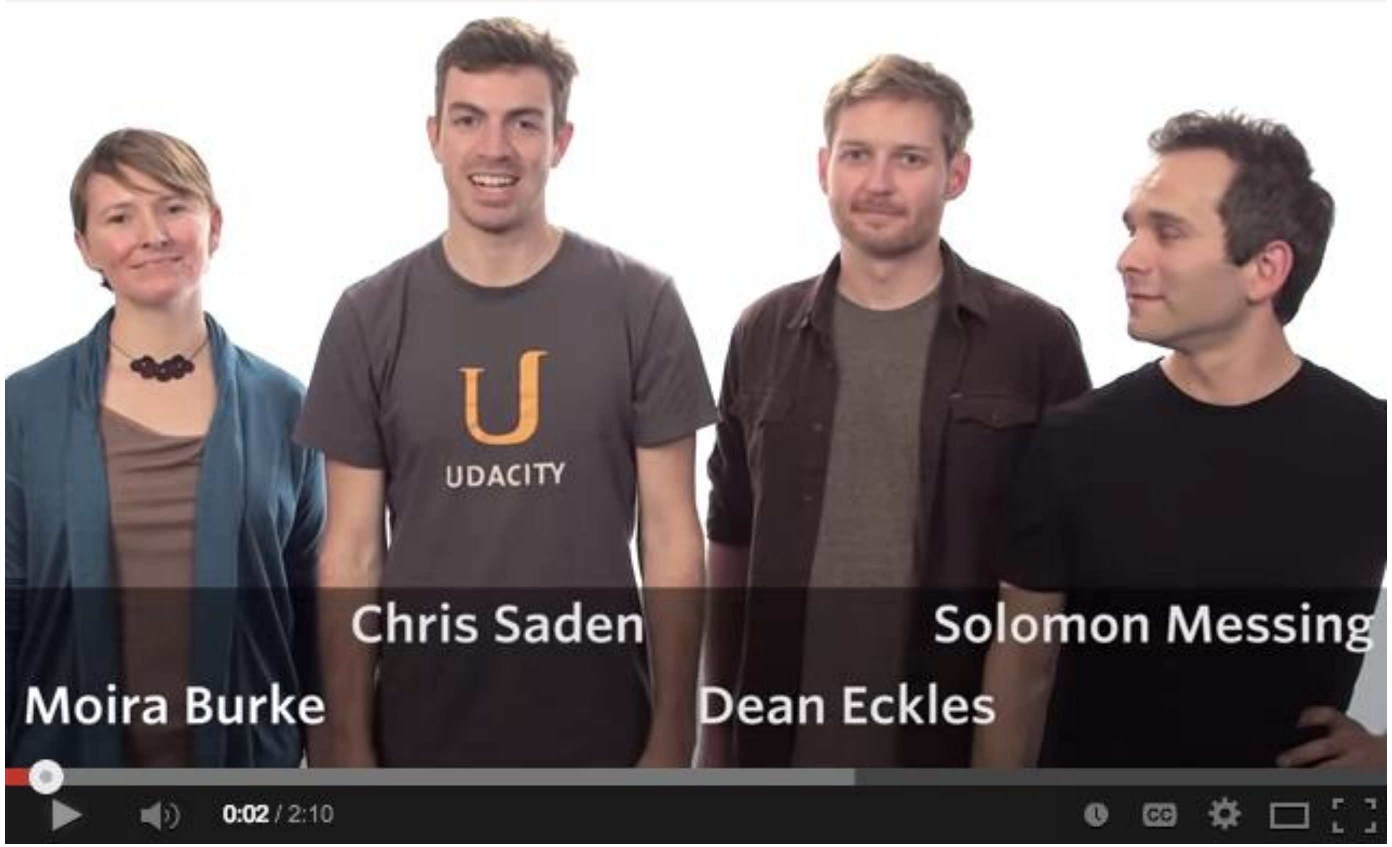
APPLICATION OF R?



MakeAGIF.com

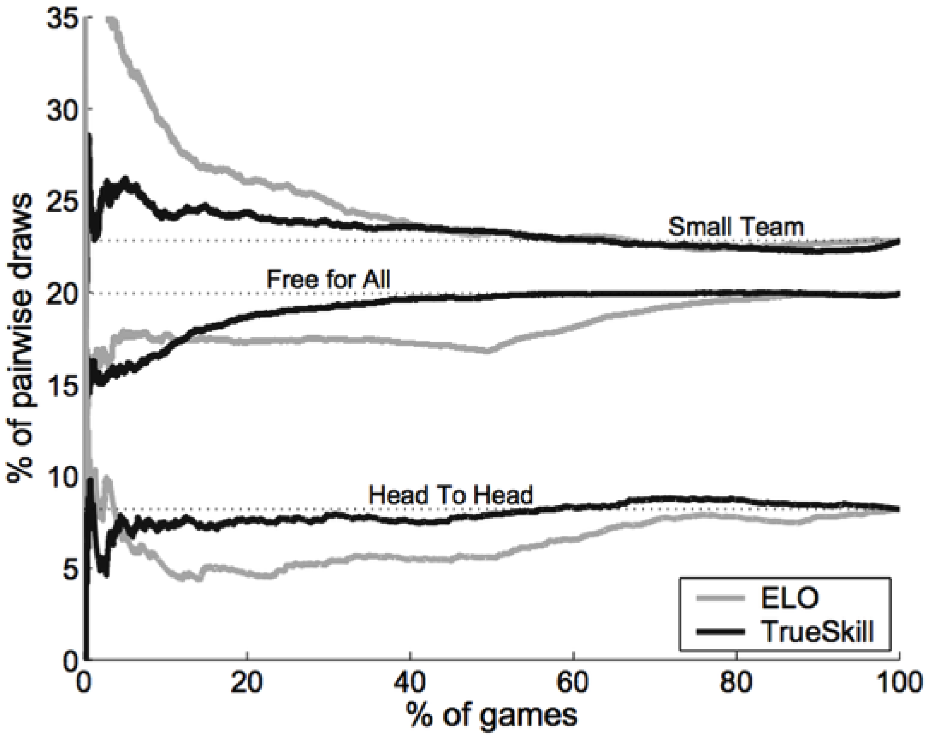
Create beautiful data visualizations

facebook



- Exploratory Data Analysis
- Experimental Analysis

“Generally, we use R to move fast when we get a new data set. With R, we don’t need to develop custom tools or write a bunch of code. Instead, we can just go about cleaning and exploring the data.” — **Solomon Messing, data scientist at Facebook**



- Player Churn
- Game design
  - Difficulty curve
  - Level trouble-spots
- In-game purchase optimization
- Fraud detection
- Player communities

- Multiplayer Matchmaking
- Game Analysis

ABOUT REVOLUTION ANALYTICS

COMPANY

REVOLUTION ANALYTICS

MOUNTAIN VIEW ▪ LONDON ▪ SINGAPORE

The leading provider of **advanced analytics software and services** based on open source R, since 2007

PRODUCT

REVOLUTION R: The enterprise-grade predictive analytics application platform based on the R language

SOME KUDOS

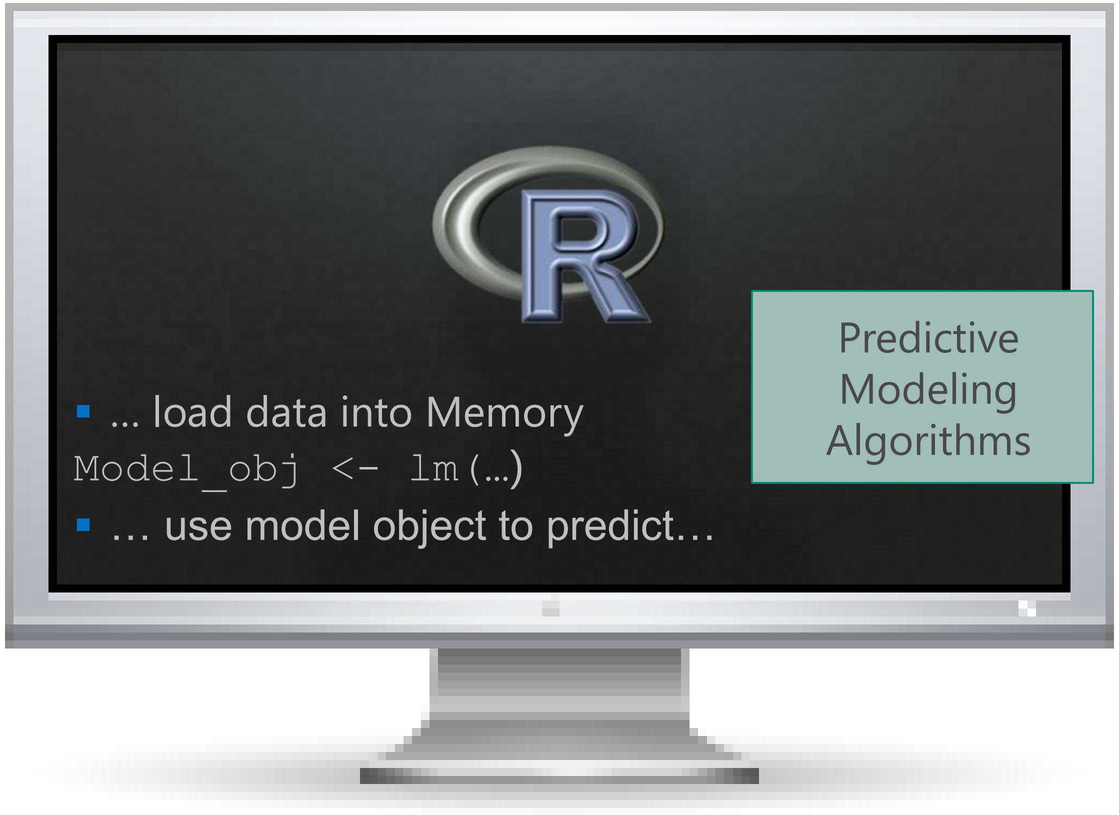
"This acquisition will help customers use advanced analytics within Microsoft data platforms"

-- Joseph Sirosh, CVP C+E

Microsoft

R Analytics: Simple, Easy, Powerful.

In Open Source R:



- … load data into Memory

```
Model_obj <- lm(…)
```

- … use model object to predict…

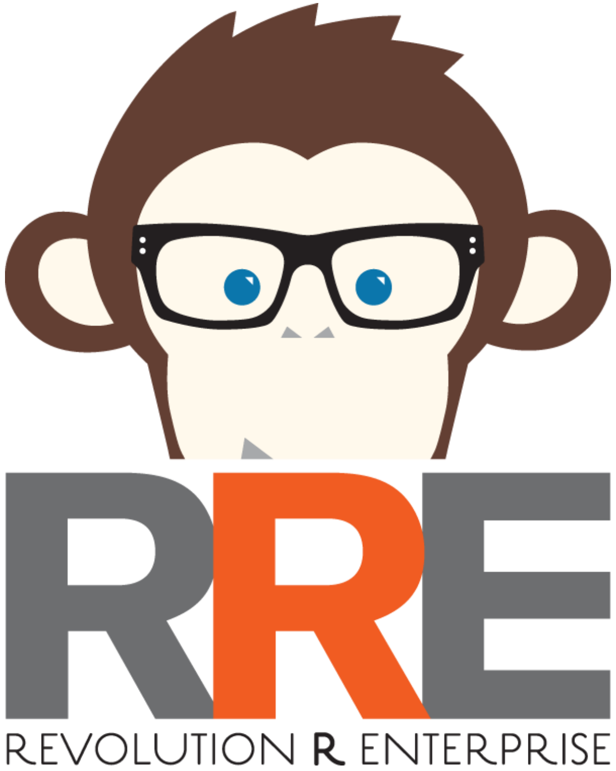
Predictive Modeling Algorithms

… but limited…

- In-Memory
- Single Threaded
- Requires Data Movement
- Not Supported Commercially

Manifestations:

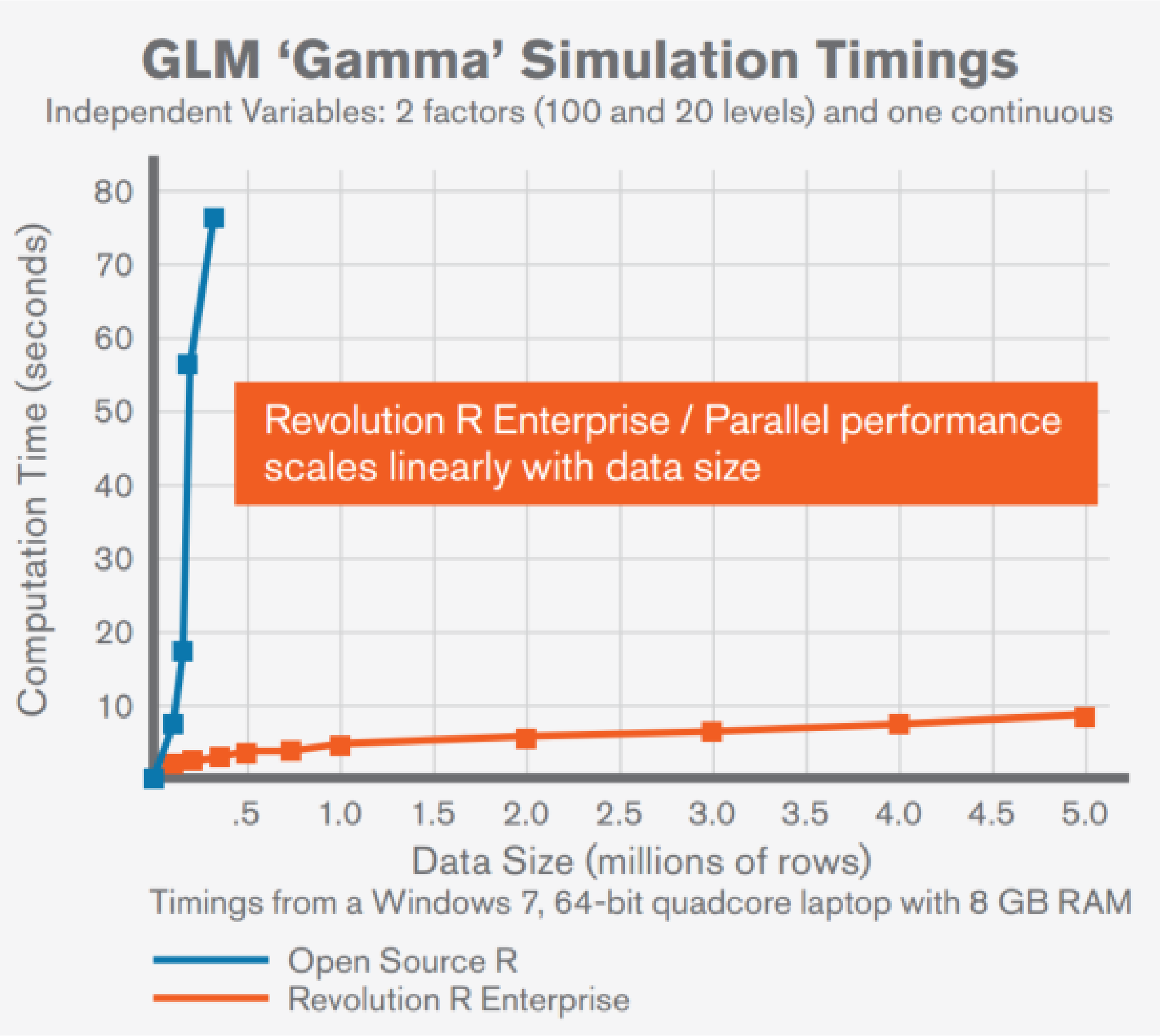
- Out of memory on large data sets
- Restricted to sampled data
- Slow computation
- Data movement slower yet
- Poor productivity
- Non-production use only
- Very complex manual parallelization



RRE is....

the only big data analytics platform based on open source R

- High Performance, Scalable Analytics
- Portable Across Enterprise Platforms
- Easier to Build & Deploy Analytics



GLM ‘Gamma’ Simulation Timings

Independent Variables: 2 factors (100 and 20 levels) and one continuous

Revolution R Enterprise / Parallel performance scales linearly with data size

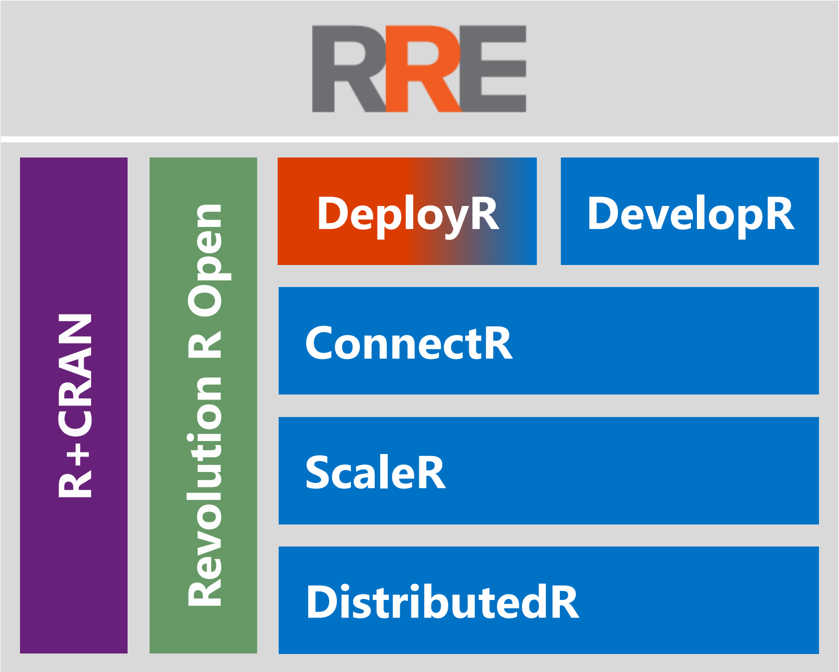
Timings from a Windows 7, 64-bit quadcore laptop with 8 GB RAM

RRE

REVOLUTION R ENTERPRISE

Revolution R Enterprise (RRE)

The All-Inclusive Big Data Big Analytics Platform



RRE

R+CRAN

Revolution R Open

DeployR

DevelopR

ConnectR

ScaleR

DistributedR

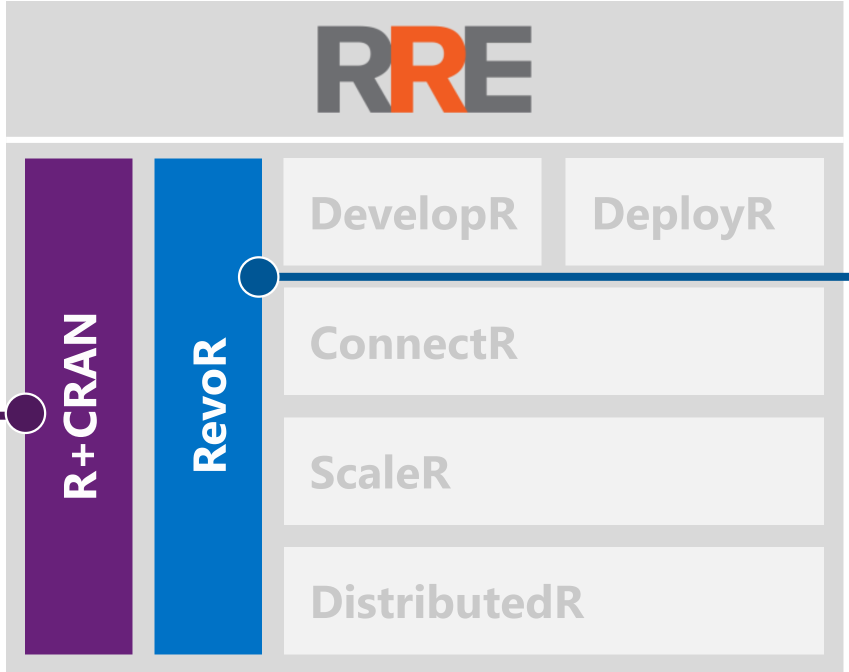
High-performance open source R **plus**:

- Data source connectivity to big-data objects
- Multi-platform environment support
- In-Hadoop and in-Teradata predictive modeling
- Secure, Scalable R Deployment
- Technical support, training and services
  - 24x7 support option

The Platform Step by Step:
R Capabilities

R+CRAN

- Open source R interpreter
  - UPDATED R 3.1.1
- Freely-available R algorithms
- Algorithms callable by RevoR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages



RevoR

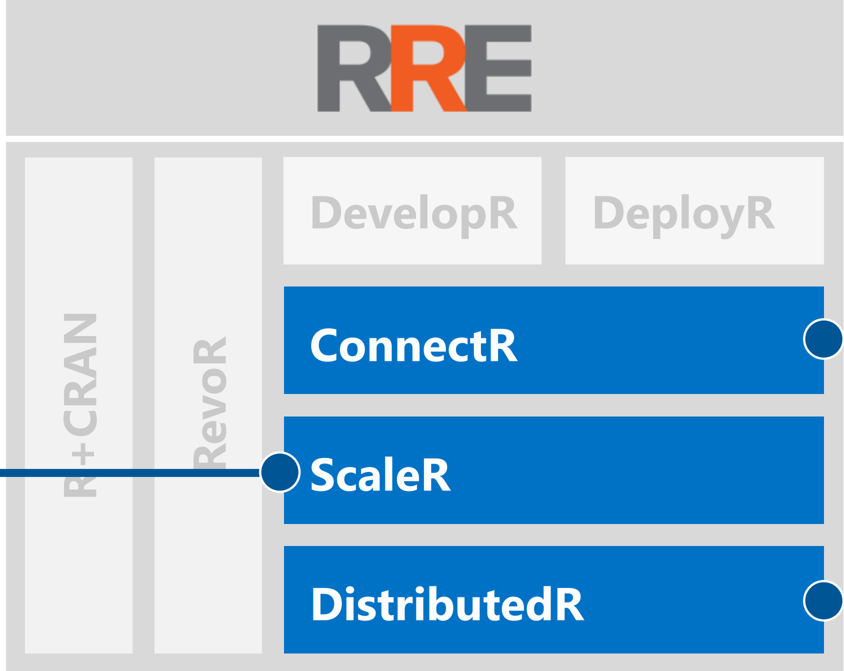
- Performance enhanced R interpreter
- Based on open source R
- Adds high-performance math

Available On:

- Platform™ LSF™ Linux®
- Microsoft® HPC Clusters
- Windows® & Linux Servers
- Windows & Linux Workstations
- Cloudera Hadoop®
- Hortonworks Hadoop
- MapR Hadoop
- Teradata® Database

The Platform Step by Step

Parallelization & Data Sourcing



ScaleR

- Ready-to-Use high-performance big data big analytics
- Fully-parallelized analytics
- Data prep & data distillation
- Descriptive statistics & statistical tests
- Correlation & covariance matrices
- Predictive Models – linear, logistic, GLM
- **NEW** Stochastic Gradient Boosted Decision Trees
- Machine learning
- Monte Carlo simulation
- **NEW** Tools for distributing customized algorithms across nodes

ConnectR

- High-speed & direct connectors

Available for:

- High-performance XDF
- SAS, SPSS, delimited & fixed format text data files
- Hadoop HDFS (text & XDF)
- Teradata Database & Aster
- EDWs and ADWs
- ODBC

DistributedR

- Distributed computing framework
- Delivers portability across platforms

Available on:

- Windows Servers
- Red Hat and SuSE Linux Servers
- IBM Platform LSF Linux
- Microsoft HPC Clusters
- Teradata Database
- Cloudera Hadoop
- Hortonworks Hadoop
- MapR Hadoop

ScaleR Functions & Algorithms

Data Step

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

Descriptive Statistics

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

Sampling

- Subsample (observations & variables)
- Random Sampling

Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

Variable Selection

- Stepwise Regression

Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

Cluster Analysis

- K-Means

Classification

New in v7.3

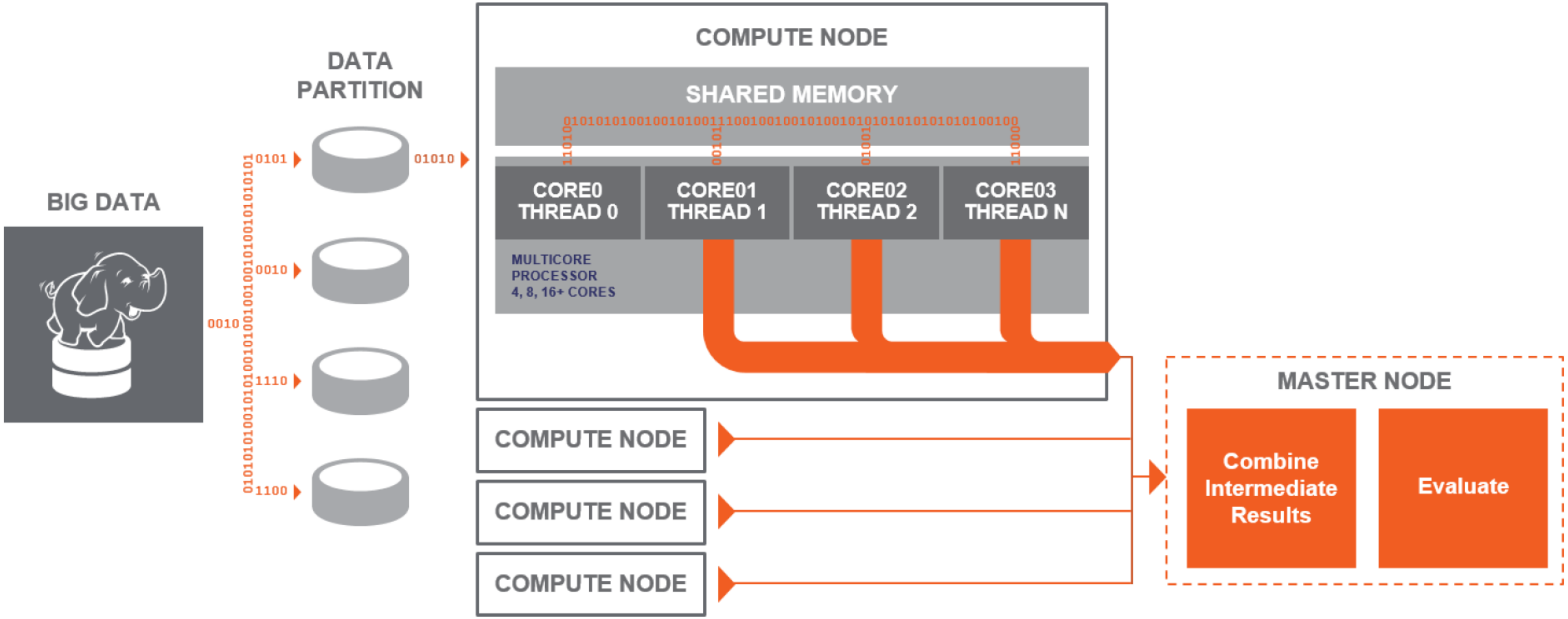
- Decision Trees
- Decision Forests
- **Gradient Boosted Decision Trees**
- **Naïve Bayes**

Combination

Coming in v7.4

- **PEMA-R API**
- rxDataStep
- rxExec

Revolution R Enterprise - ScaleR



The Revolution R Product Suite

Revolution R Open

- Free and open source R distribution
- Enhanced and distributed by Revolution Analytics

OPEN
REVOLUTION R

Revolution R Enterprise

- Secure, Scalable and Supported Distribution of R
- Includes proprietary components for Big Data analytics, integration and developer IDE

RRE
REVOLUTION R ENTERPRISE

Revolution R Open

- Enhanced Open Source R distribution
- Compatible with all R-related software
- Multi-threaded for performance
- Focus on reproducibility
- Open source (GPLv2 license)
- Available for Windows, Mac OS X, Ubuntu, Red Hat and OpenSUSE
- Free download at `mran.revolutionanalytics.com`

	Revolution R Open	Revolution R Enterprise
R Language Engine with multi-core processing	Included	Supported
R Reproducibility Toolkit & MRAN	Included	Supported
ParallelR: Parallel Programming Toolkit		Supported
RHadoop: R interface to Hadoop MapReduce		Supported
DeployR Open: Web Services API		Supported
RRE DeployR – Multi-server, enterprise authentication		Licensed & Supported
RRE ScaleR – Big Data toolkit and PEMAs for R		Licensed & Supported
RRE DistributedR – EDW, Grids, Hadoop		Licensed & Supported
AdviseR Technical Support		**Included**
Open Source Assurance		**Included**
Revolution Analytics Services (Consulting / Training)	**Available**	**Available**

MICROSOFT AND R(EVOLUTION ANALYTICS)

Microsoft and Revolution Analytics

Strategic Rationale

Help more companies use the power of R

Opens new opportunities for our existing customers

Enables us to provide cross-platform, in-db analytics

Compatibility with Azure => more cost efficiently

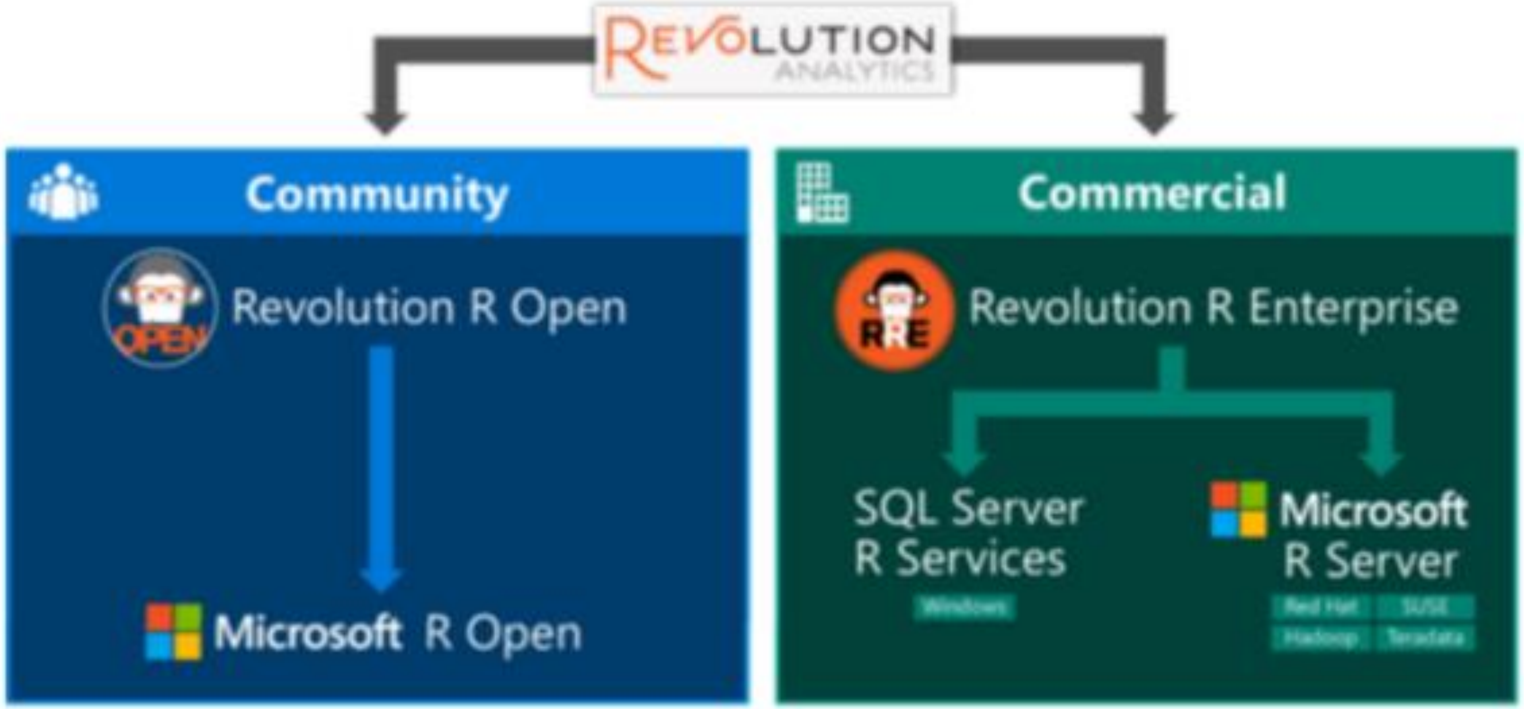
What’s new?

Flexible & Agile

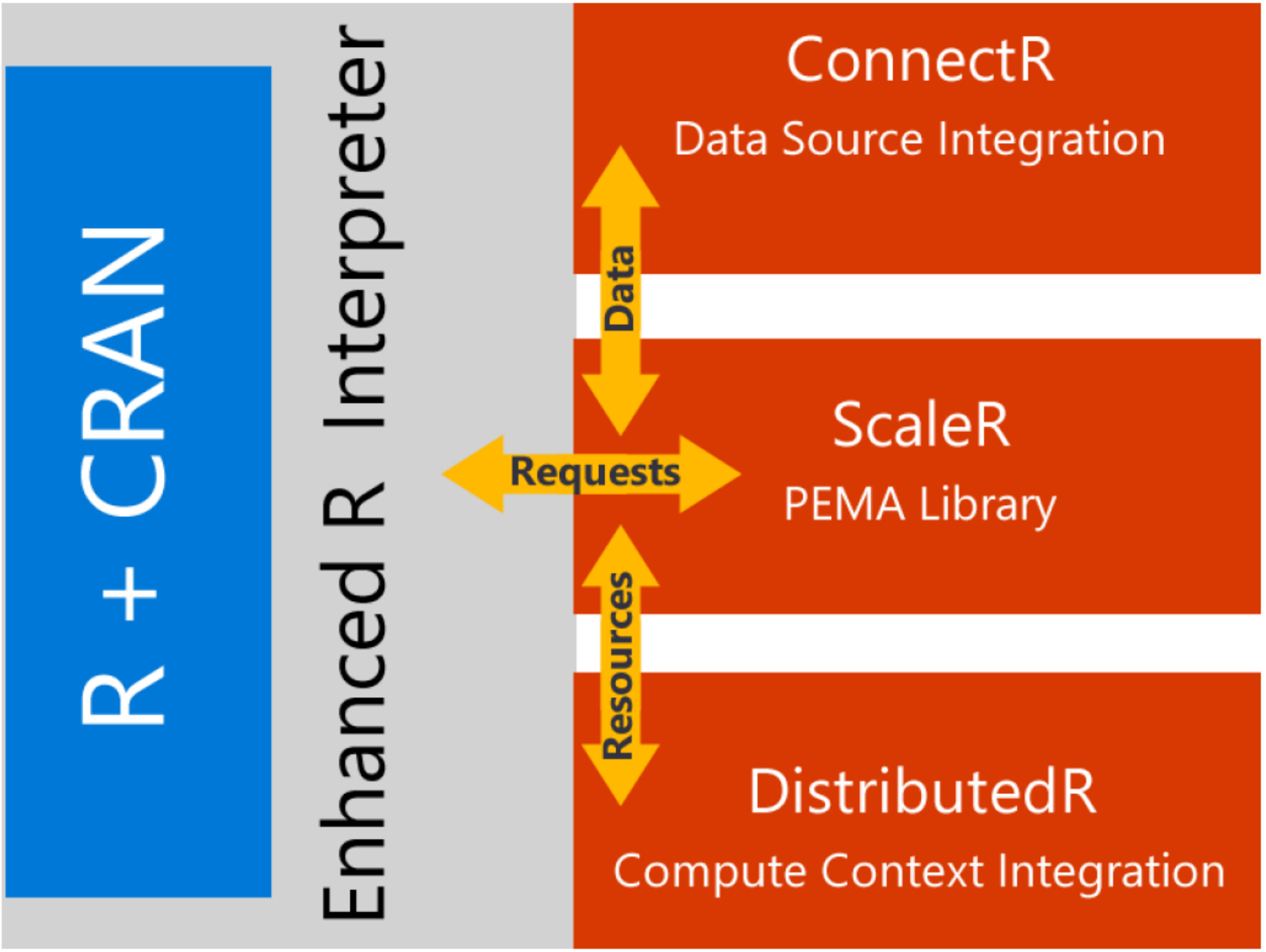
Open benefits, enterprise support

R for SQL Server

Microsoft and R(evolution Analytics)



Microsoft R Server



R + CRAN

Enhanced R Interpreter

ConnectR

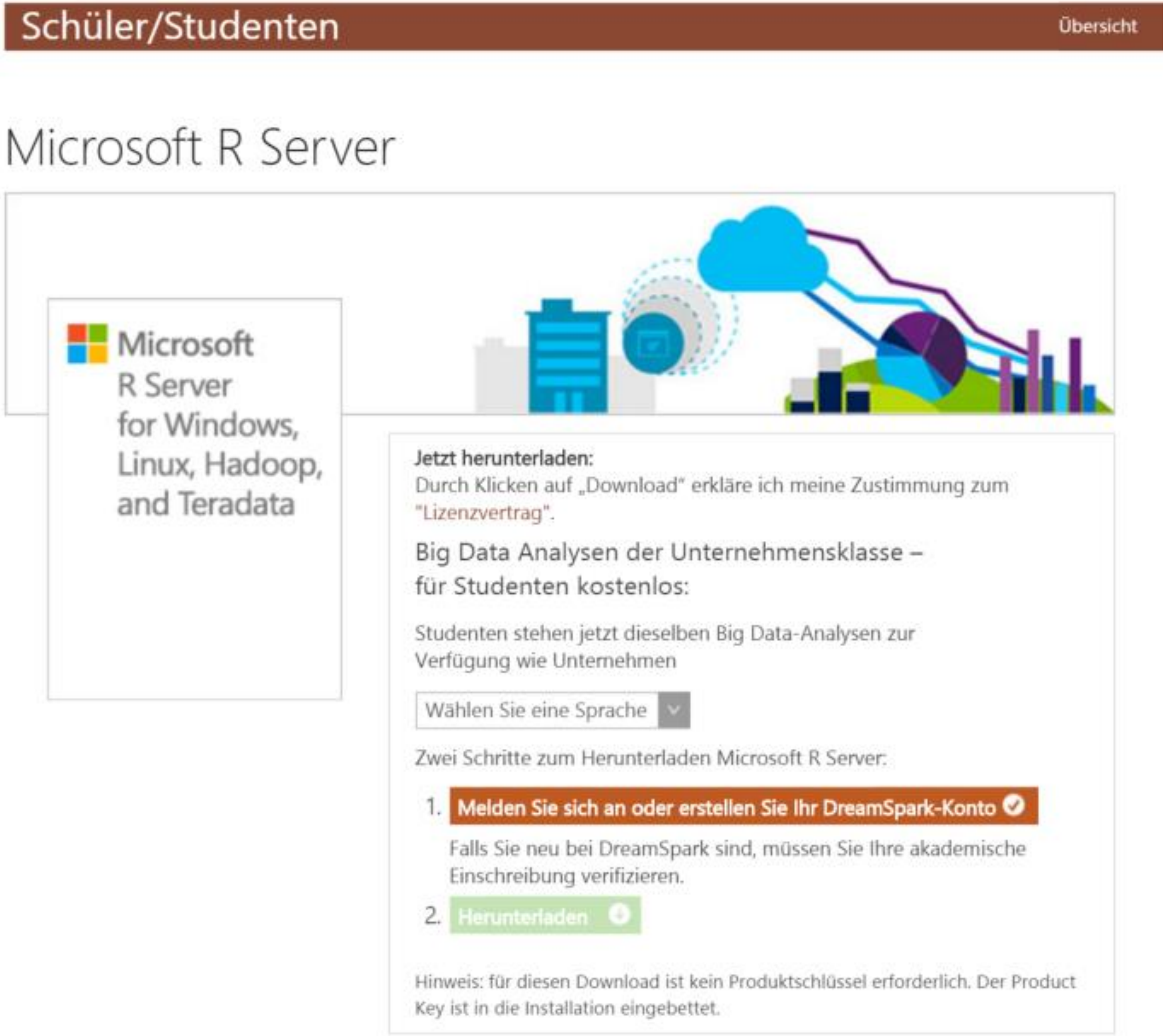
Data Source Integration

ScaleR

PEMA Library

DistributedR

Compute Context Integration



Schüler/Studenten

Übersicht

Microsoft R Server

Jetzt herunterladen:
Durch Klicken auf „Download“ erkläre ich meine Zustimmung zum "Lizenzvertrag".

Big Data Analysen der Unternehmensklasse –
für Studenten kostenlos:

Studenten stehen jetzt dieselben Big Data-Analysen zur Verfügung wie Unternehmen

Wählen Sie eine Sprache

Zwei Schritte zum Herunterladen Microsoft R Server:

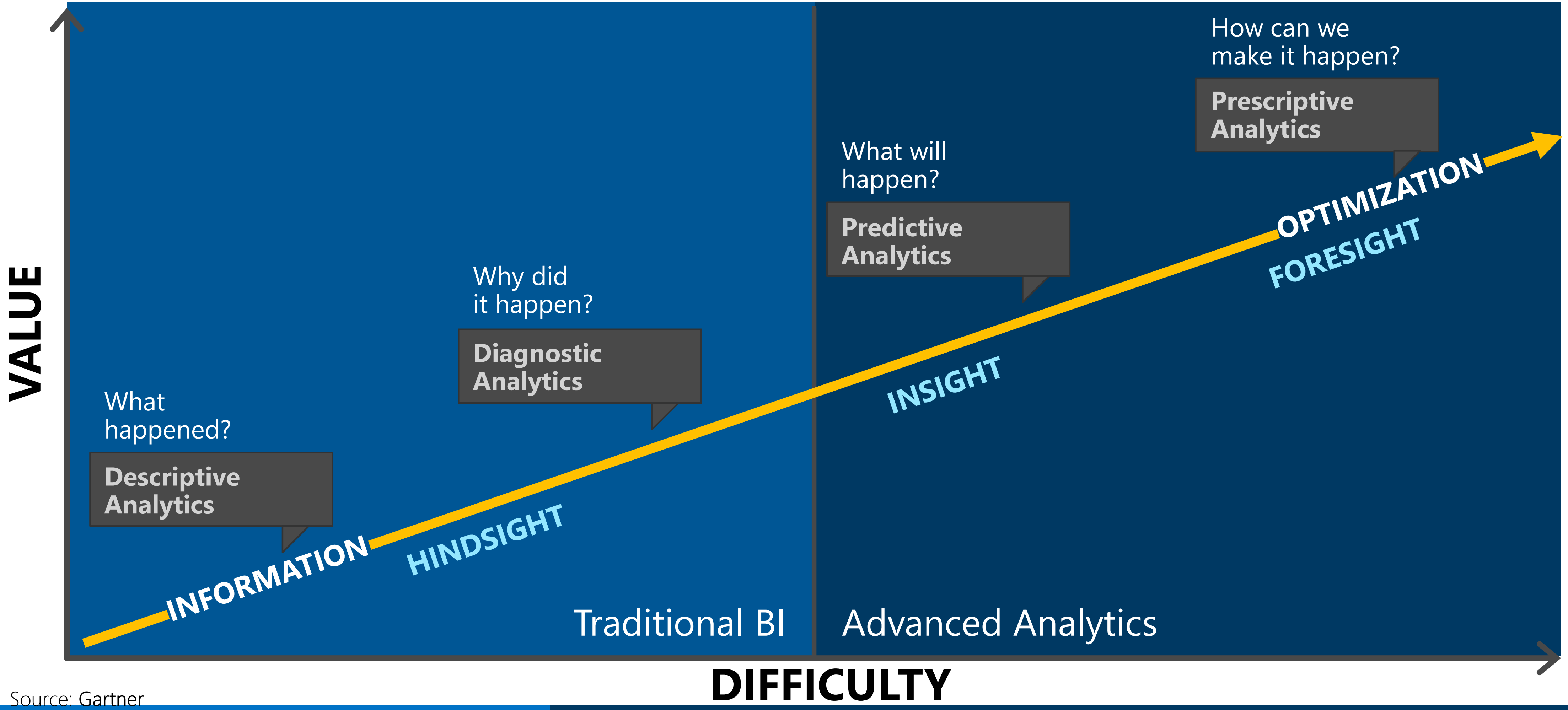
1. Melden Sie sich an oder erstellen Sie Ihr DreamSpark-Konto

   Falls Sie neu bei DreamSpark sind, müssen Sie Ihre akademische Einschreibung verifizieren.

2. Herunterladen

Hinweis: für diesen Download ist kein Produktschlüssel erforderlich. Der Product Key ist in die Installation eingebettet.

Advanced Analytics
Beyond business intelligence



Source: Gartner

Scenarios for SQL Server 2016

Exploration

Analyze large datasets and build predictive models with the compute happening **on the SQL Server machine**

Operationalization

Developer can operationalize R script/model over SQL Server data by using T-SQL constructs

DBA can manage, secure & govern the R runtime execution in SQL Server

R Script Usage from SQL Server

- Original R Script:

```
IrisPredict <- function(data, model){
library(e1071)
predicted_species <- predict(model, data)
return(predicted_species)
}


library(RODBC)
conn <- odbcConnect("MySqlAzure", uid = myUser, pwd = myPassword);
Iris_data <-sqlFetch(conn, "Iris_Data");
Iris_model <-sqlQuery(conn, "select model from my_iris_model");
IrisPredict (Iris_data, model);
```

- Calling R Script from SQL Server:

```
/* Input table schema */
create table Iris_Data (name varchar(100), length int, width int);
/* Model table schema */
create table my_iris_model (model varbinary(max));

declare @iris_model varbinary(max) = (select model from my_iris_model);
exec sp_execute_external_script
  @language = 'R'
, @script = '
IrisPredict <- function(data, model){
library(e1071)
predicted_species <- predict(model, data)
return(predicted_species)
}
IrisPredict(input_data_1, model);
'
, @parallel = default
, @input_data_1 = N'select * from Iris_Data'
, @params = N'@model varbinary(max)'
, @model = @iris_model
with result sets ((name varchar(100), length int, width int
, species varchar(30)));
```

yellow

aqua

demo

Text mining with R in Azure ML

demo

R in Visual Studio

https://github.com/Microsoft/RTVS

http://microsoft.github.io/RTVS-docs/

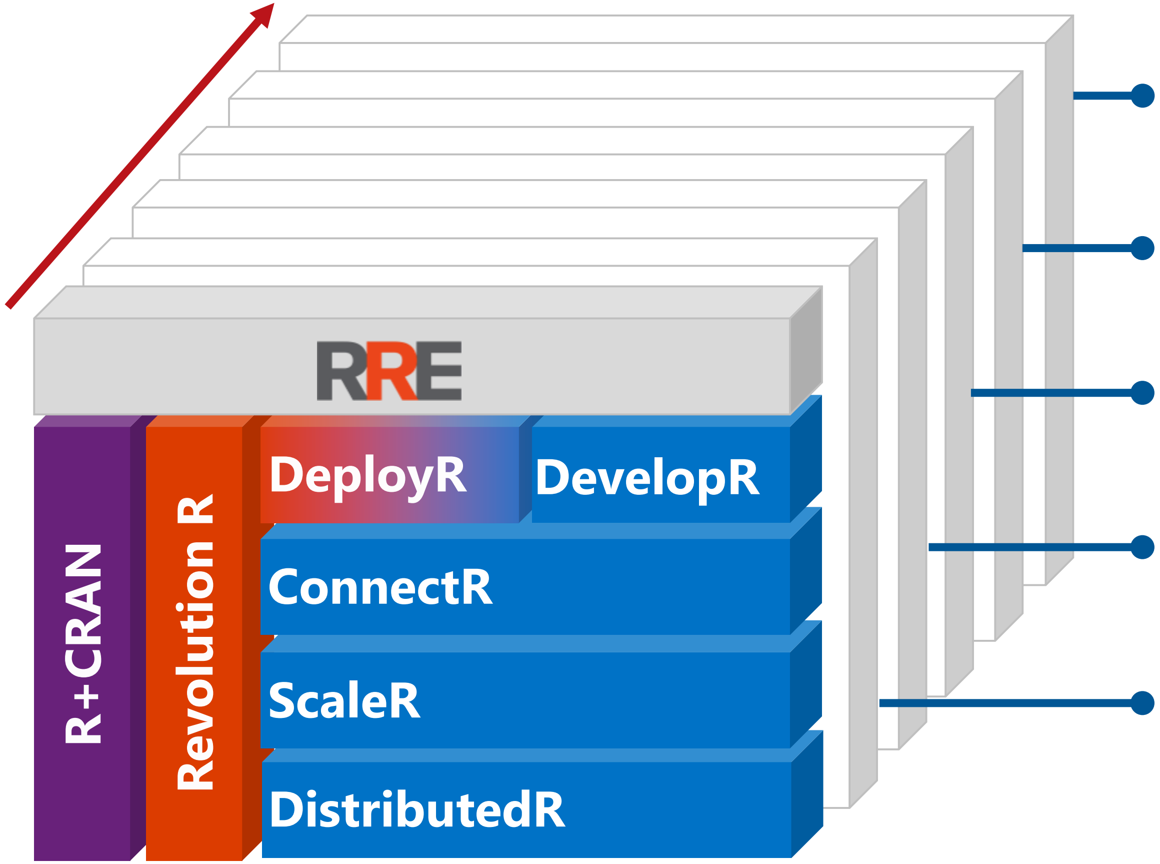
https://github.com/Microsoft/R-Host

ANY
QUESTIONS
?

Get Started

- Shiny
- R-Tutor
- Cheat Sheet
- More Cheat Sheets

Write Once. Deploy Anywhere.



Hadoop → Hortonworks
Cloudera
MapR (soon HDInsights)

EDW → Teradata

Clustered Systems → Linux Clusters (LSF For Now)
Microsoft HPC

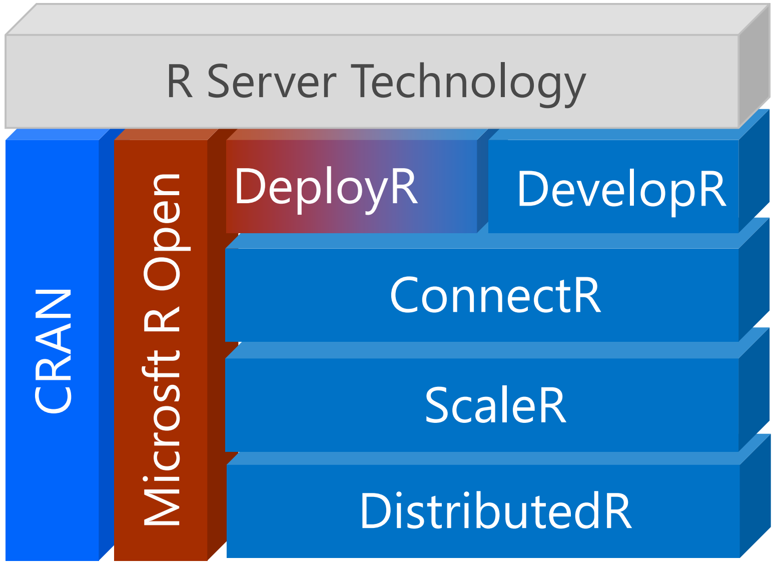
Workstations & Servers → Linux
Windows

In the Cloud → Coming: Azure

DESIGNED FOR SCALE, PORTABILITY & PERFORMANCE

Microsoft R Server & SQL Server R Services

- Open Source Compatible
- 100% R Compatible
- Runs R Scripts Unchanged
- Runs CRAN, Bioconductor & GitHub R Packages



- Enterprise Scalability, Stability, Support & Productivity
- Big Data Advanced Analytics
- Fast R IDE
- Web Services Deployment
- Freedom from Memory Constraints
- Flexible Data Integration