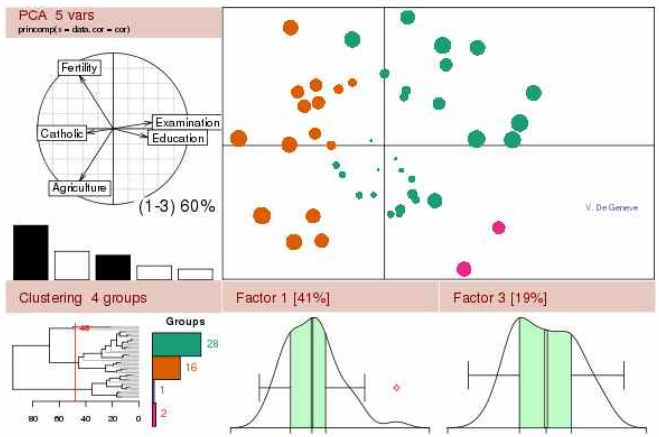
GTONE
Governance Technology



R, 그리고 빅데이터

제30회 Open Technet

-빅데이터 오픈소스 플랫폼 기술세미나

2012.07.26

이 동 우

지티원

R, 그리고 빅데이터

What is R?

R is a language and environment for statistical computing and graphics.

- Data analysis software
- A programming language
  - 통계학자들이 디자인하고 통계학자들을 위한 개발 플랫폼
- An environment
  - 데이터와 관련된 입출력, 핸들링, 관리, 분석, 그래픽 등
    최신의 알고리즘 및 라이브러리 제공
- An open-source software project
  - Free, open, and active
- A community
  - 수 천명의 contributors, 2백만이 넘는 사용자
  - 각 업무도메인과 관련된 리소스와 도움말 제공

R : Open source analytics for the enterprise

- Most advanced statistical analysis software available
- Half the cost of commercial alternatives
- 2M+ Users
- 2,500+ Applications

Statistics	Finance
	Life Sciences
Predictive Analytics	Manufacturing
	Retail
Data Mining	Telecom
	Social Media
Visualization	Government

Forbes

Power in the Numbers

Quentin Hardy, 05.06.10, 09:00 AM EDT

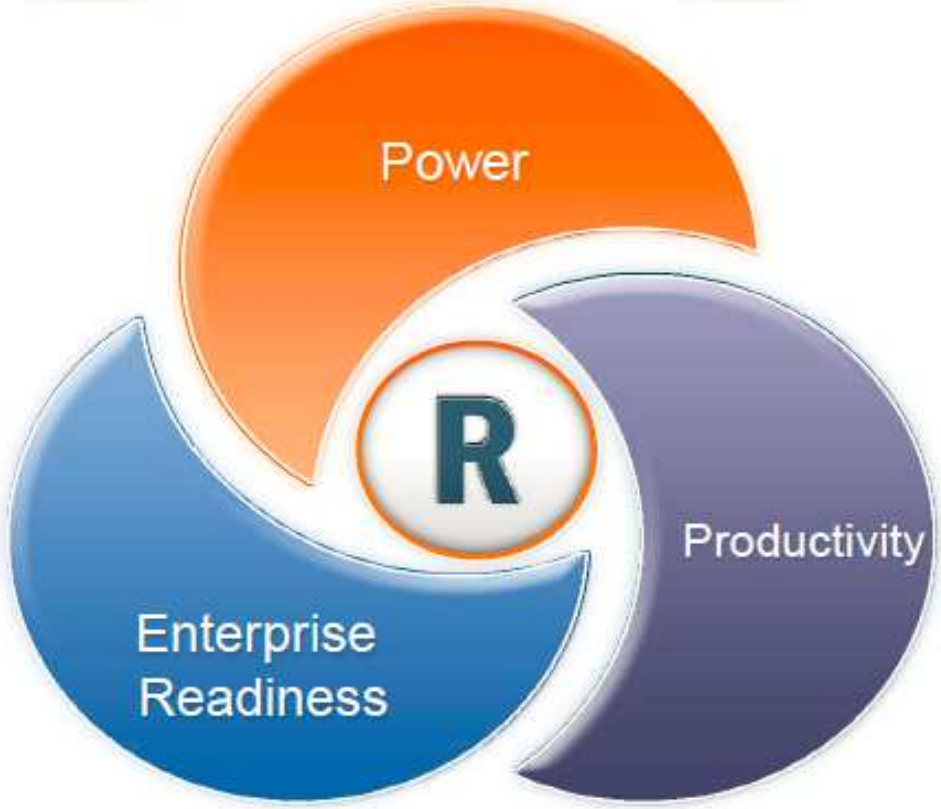
The professor who invented analytic software for the experts now wants to take it to the masses

The New York Times

Data Analysts Captivated by R's Power

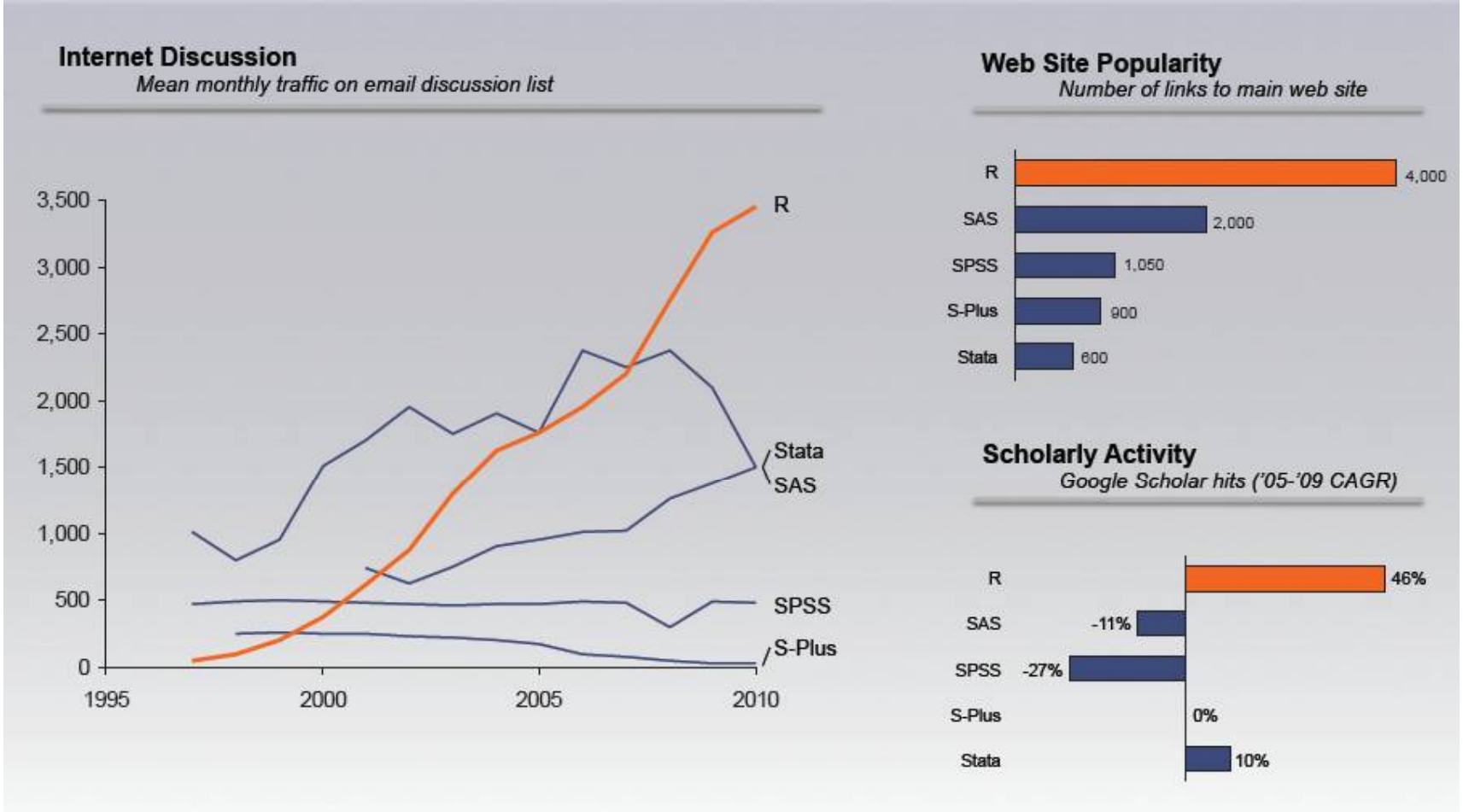
By ASHLEE VANCE

Published: January 6, 2009



R에 대한 관심의 증가

- 폭발적인 사용자 증가와 개발자의 확산으로, 대학교육의 표준 툴로 자리 잡음



출처 : Revolution Analytics

기업체에서의 R의 활용

- 빅 데이터 기업의 분석 플랫폼 엔진으로 사용 중이며, 유수기업에서 데이터 분석 tool로 사용 중임

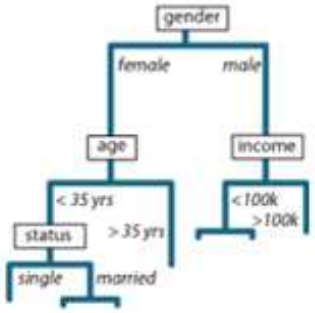
기업체에서의 R의 활용

- 특히, Google과 Facebook은 R을 자사의 주된 분석 플랫폼으로 활용하고 있음

Google's R Style Guide

How Google and Facebook are using R

by mike | February 19th, 2009



(March 26th Update: Video now available)
Last night, I moderated our Bay Area R Users Group kick-off event with a panel discussion entitled "The R and Science of Predictive Analytics", co-located with the Predictive Analytics World conference here in SF.

The panel comprised of four recognized R users from industry:

Bo Cowgill, Google
Itamar Rosenn, Facebook
David Smith, Revolution Computing
Jim Porzak, The Generations Network (and Co-Chair of our R Users Group)

The panelists were asked to explain how they use R for predictive analytics within their firms, its strengths and weaknesses as a tool, and provide a case study. What follows is my summary with comments.

Google uses R for data exploration and model prototyping, it is not typically used in production: in Bo's group, R is typically run in a desktop environment.

- Bo Cowgill, Google

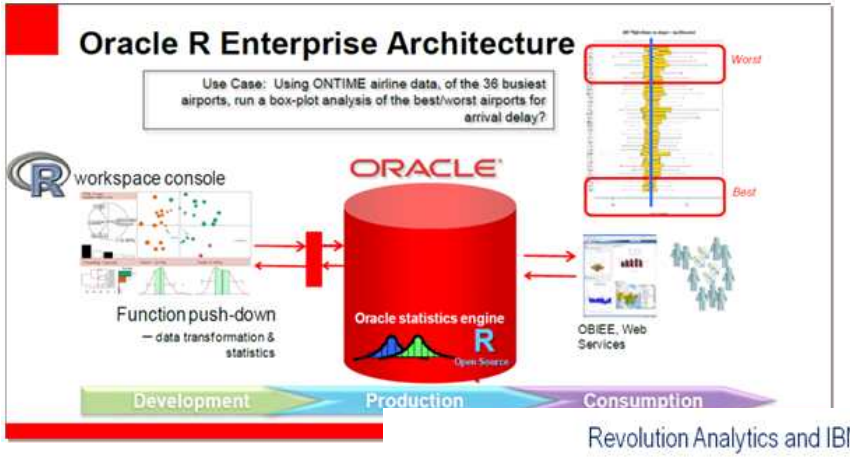
Itamar conveyed how Facebook's Data Team used R in 2007 to answer two questions about new users: (i) which data points predict whether a user will stay? and (ii) if they stay, which data points predict how active they'll be after three months?

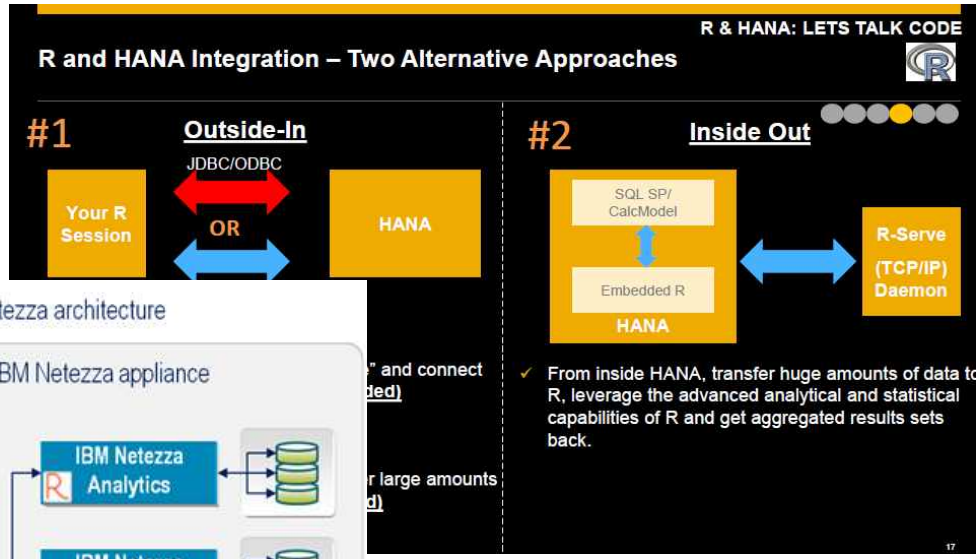
- Itamar Rosenn, Facebook

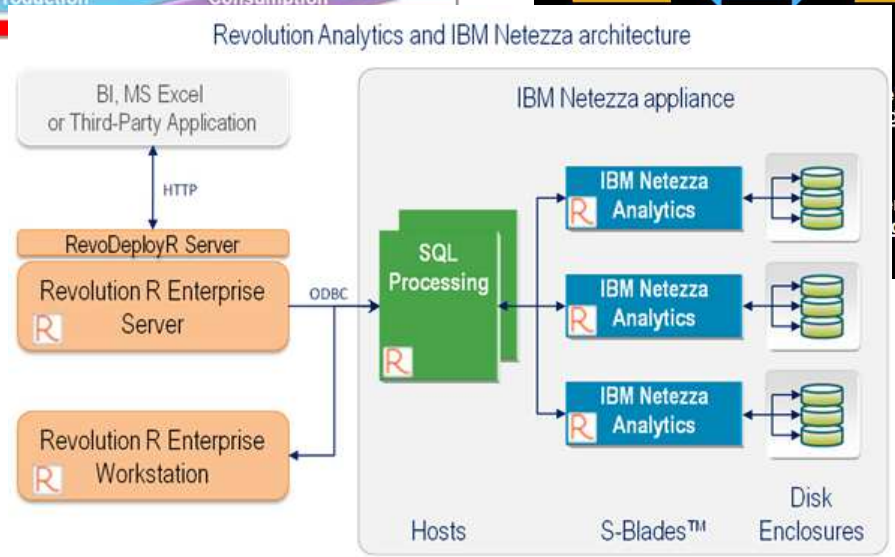
http://www.dataspora.com/2009/02/predictive-analytics-using-r/

소프트웨어 Vendor의 R 적용

- **Oracle, IBM의 Netezza, SAP의 HANA, Teradata 등에서 in-memory 혹은 in-database 분석 엔진으로 R을 적용함**

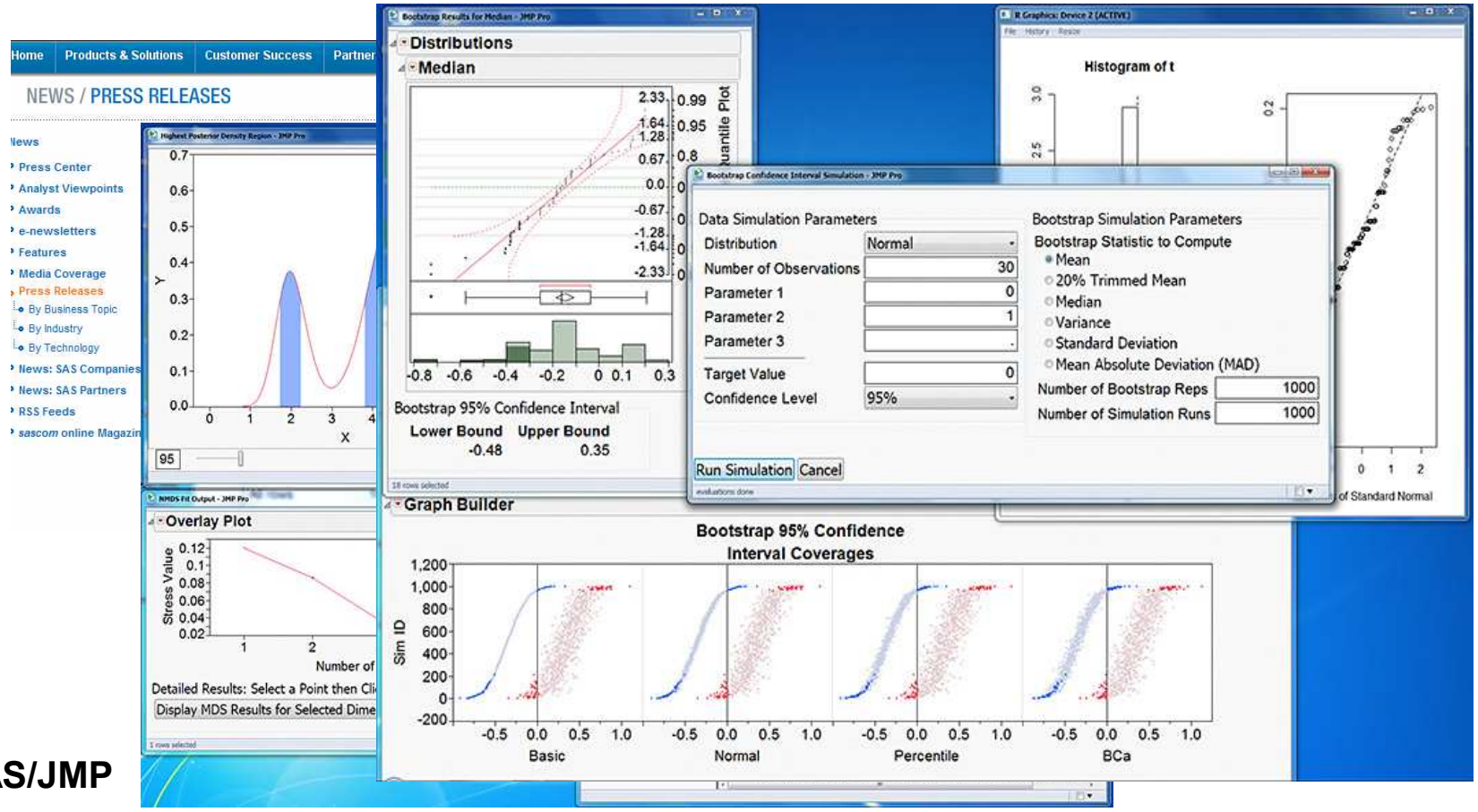






통계 소프트웨어 Vendor의 R 적용

- SAS나 SPSS 등의 통계 소프트웨어에서 R과의 연동을 통해 새로운 분석 방법을 제공



SAS/JMP

SAS understands why R

"A key benefit of R is that it provides near instant availability of new and experimental methods created by its user base — without waiting for the development/release cycle of commercial software. SAS recognizes the value of R to our customer base..."

— Michael Gilliland, Product Marketing Manager SAS Institute, Inc.

SPSS with R

- 왜 SPSS에서는 R이 필요한가?
  - 가장 최신의… 가장 많은 알고리즘 보유
  - 오픈 소스의 강점으로, 매우 희귀한 분석이나 다양한 최신의 알고리즘이 R을 통해서, 자유롭게 공급 배포되고 있음. (일반 상용 통계패키지가 접근이 어려운 점)
- SPSS Statistics 17버전(권장은 18 이후) 이후 R과의 결합으로 상호 Win-Win 달성
  - 더욱 더 강력해진 SPSS, 활용도가 높아진 R

"평소에 SPSS를 이용하여, 회귀분석이나, 각종 TEST를 수행을 하여, 논문 및 연구보고서를 만들었는데,…

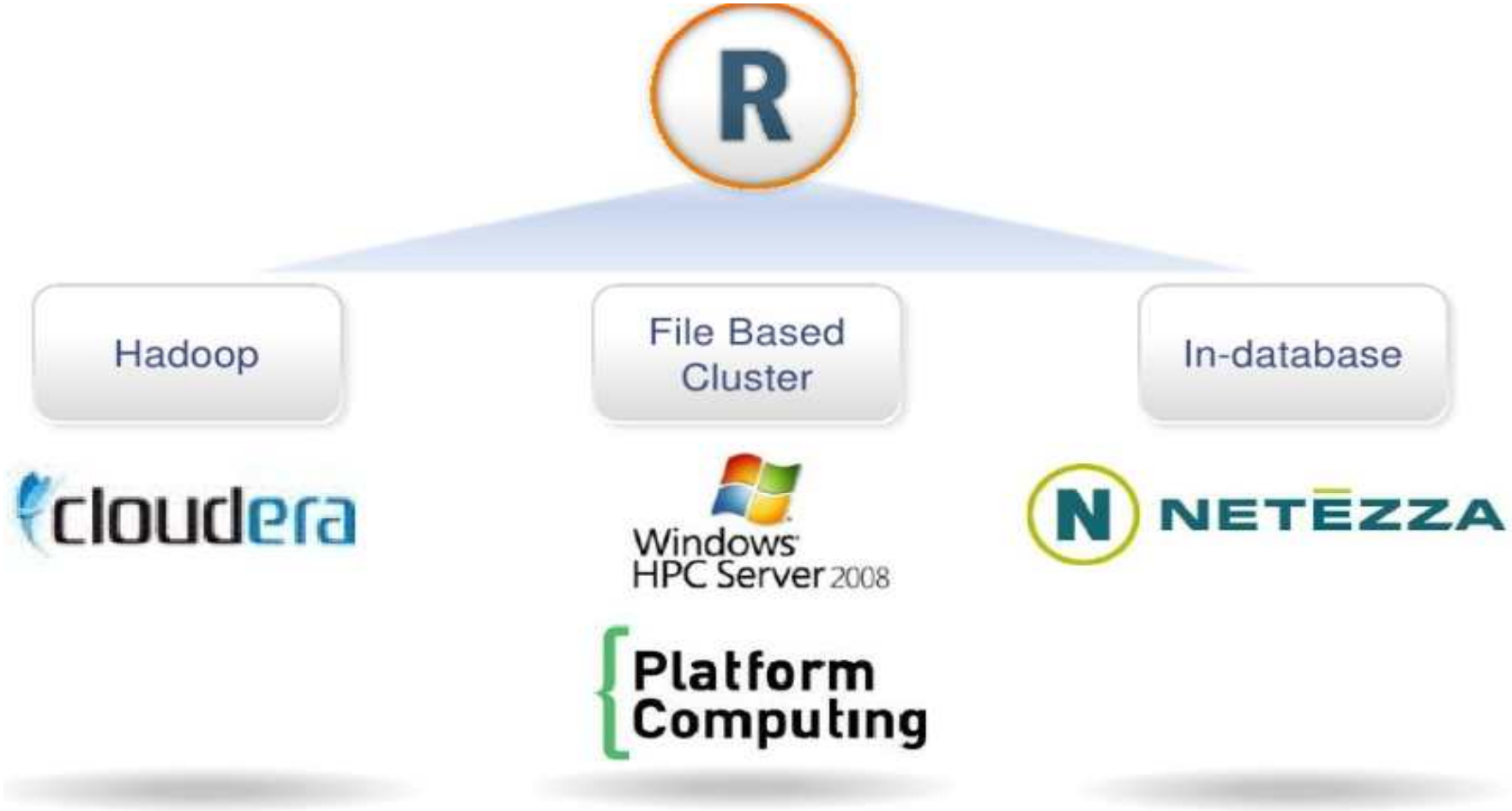
새로운 연구의 경우 Tobit 회귀를 써야 했어요… 그러나 SPSS에서 분석 기능이 지원이 되지 않아, 다른 통계 패키지를 배우고, 습득하는데, 고생이 많았습니다. SPSS쓰다 다른 패키지 갔다가… 헷갈리기도 하고…

그러나 이번 확장 모듈로 간단히 해결이 됐습니다!!!"

출처 : RUserConf2011 - SPSS를 이용한 R 연동 기능소개와 분석기능의 시너지 효과, 허준, SPSS Korea, 2011.10.28

Big Data 분석을 위한 R의 활용

- 빅 데이터 분석을 위한 아키텍처 전반에 걸쳐 공통적인 분석 플랫폼으로 자리잡음

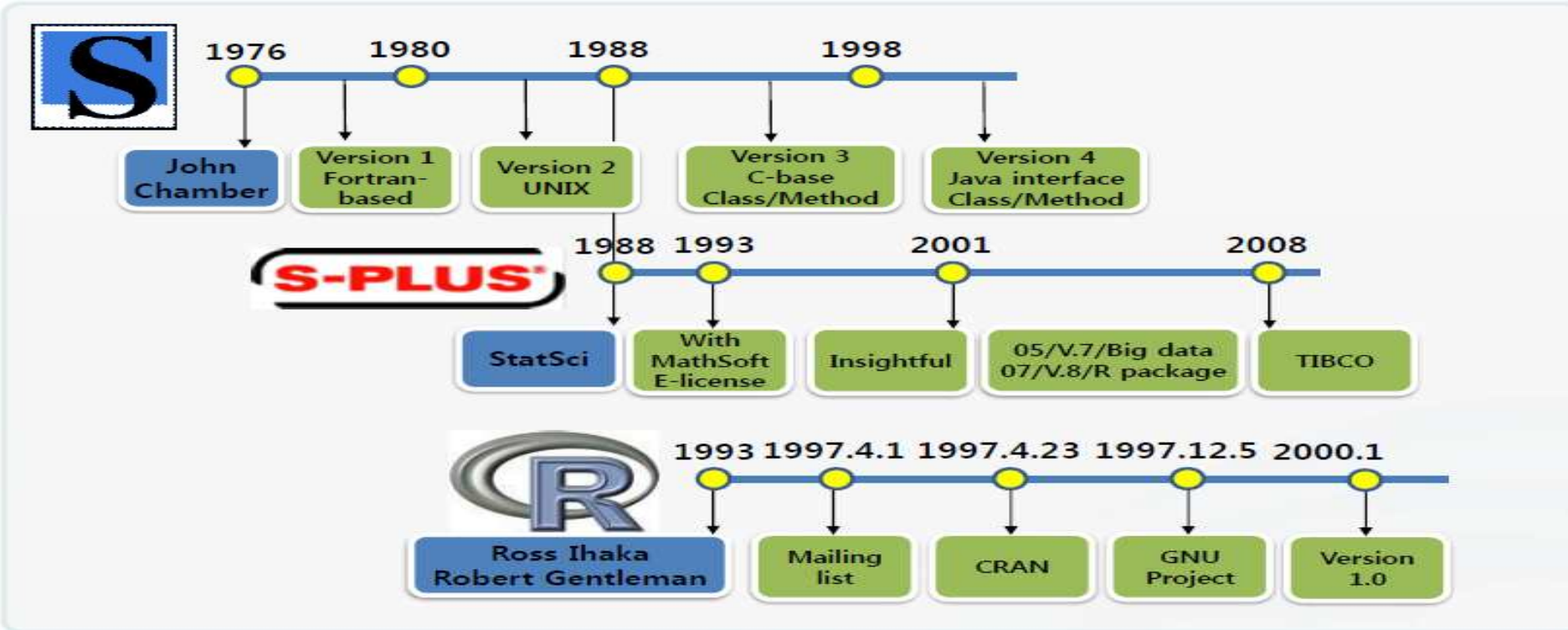


출처 : Revolution Analytics

R의 특징

- In-Memory Computing
  - 빠른 처리 속도, H/W 메모리 크기에 영향을 받음
- Object-oriented programming
  - 데이터, 함수가 object로 관리되어 짐
  - 클래스(class) & 메소드(method)
- Package
  - 최신의 알고리즘 및 방법론을 적용
  - 다양한 함수 및 데이터 내장, Help의 Examples 바로 사용 가능
- Visualization
  - 분석에 통찰을 부여할 수 있는 그래픽에 대한 강력한 지원
  - Chart, Plot, MotionChart, Map 연계 등을 R에서 바로 사용 가능

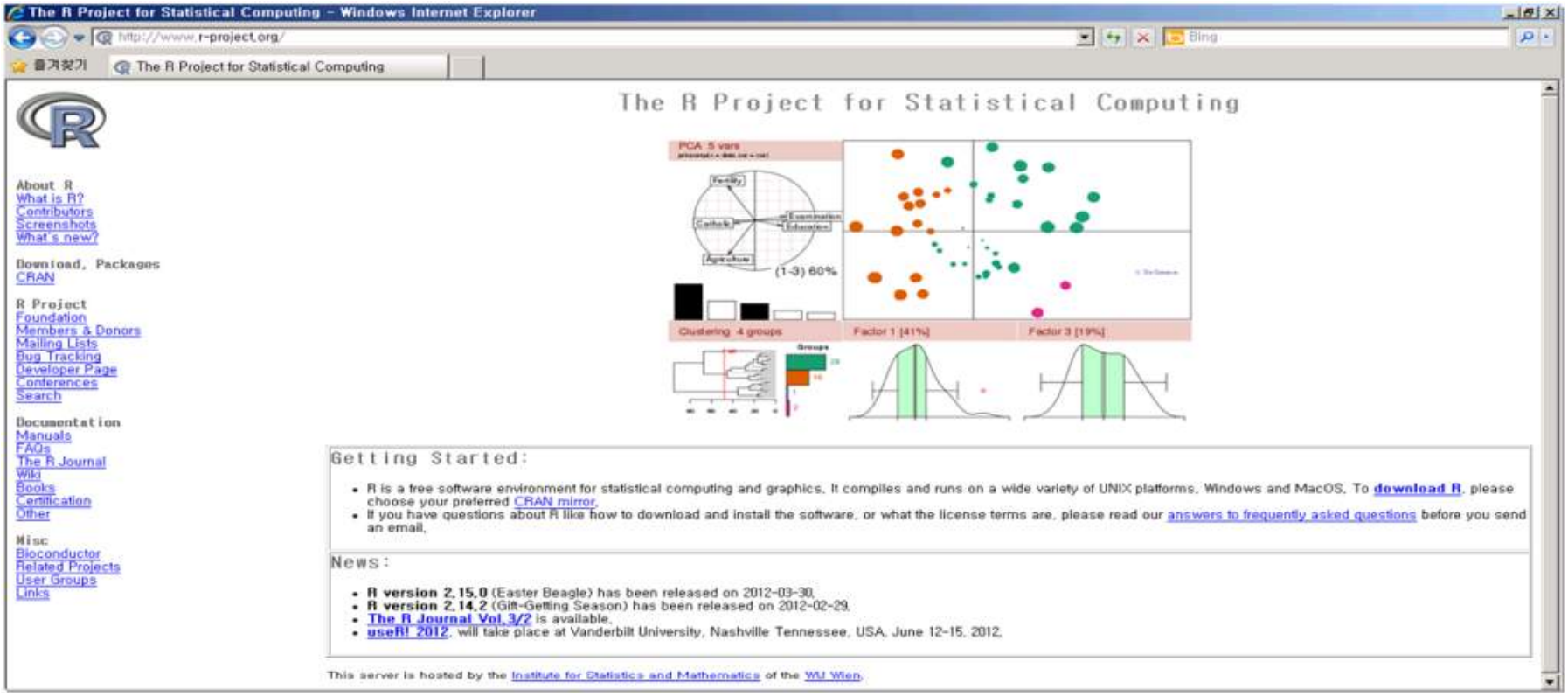
History of R



출처: 유충현 (넥스알, 빅데이터 애널리틱스 인사이드 2011)

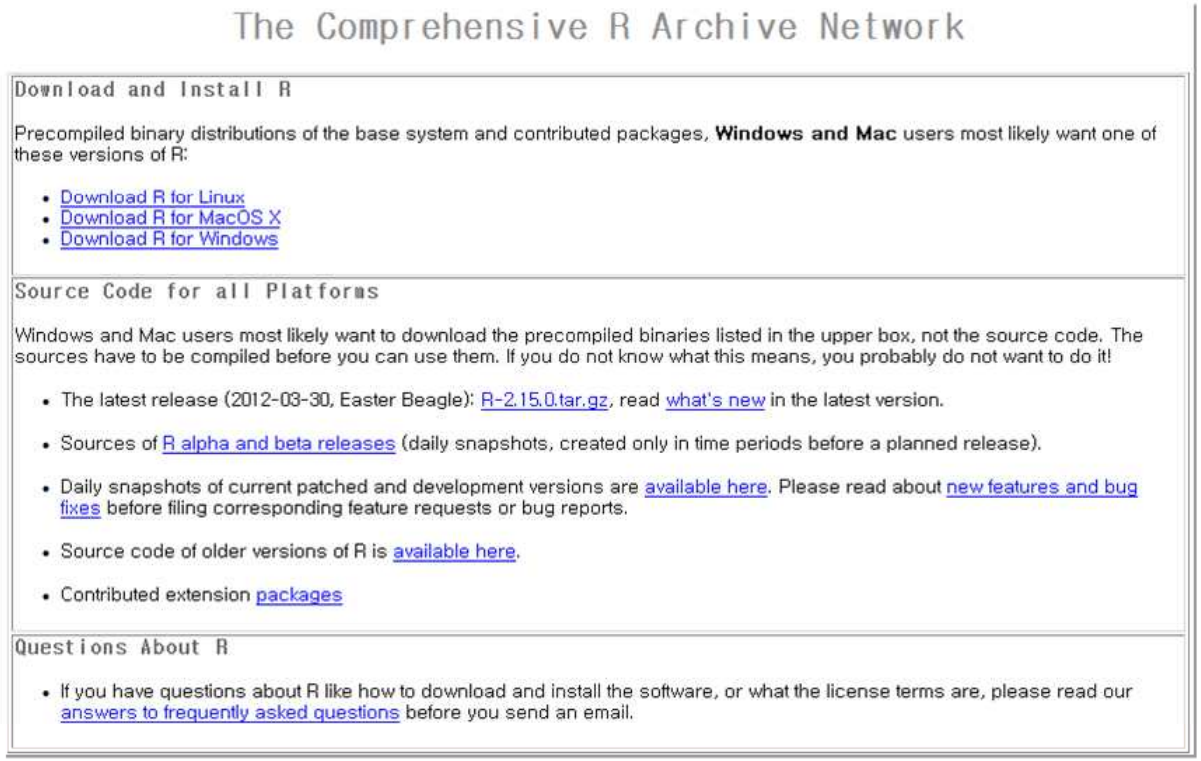
- R은 1993년 뉴질랜드 오클랜드대학의 통계학과 교수 2명(Ross Ihaka, Robert Gentleman)에 의하여 개발
- 1976년 Bell Lab의 John Chambers, Rick Becker, Allan Wilks에 의하여 개발된 S Language에 그 뿌리를 두고 있음

The R Foundation (http://www.r-project.org)



- R Development Core Team 멤버들에 의하여 설립된 비영리 단체
- R의 배포와 수정은 R Development Core Team과 많은 기여자들에 의하여 이루어 지고 있음

CRAN (The Comprehensive R Archive Network)



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for MacOS X
- Download R for Windows

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2012-03-30, Easter Beagle): R-2.15.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

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

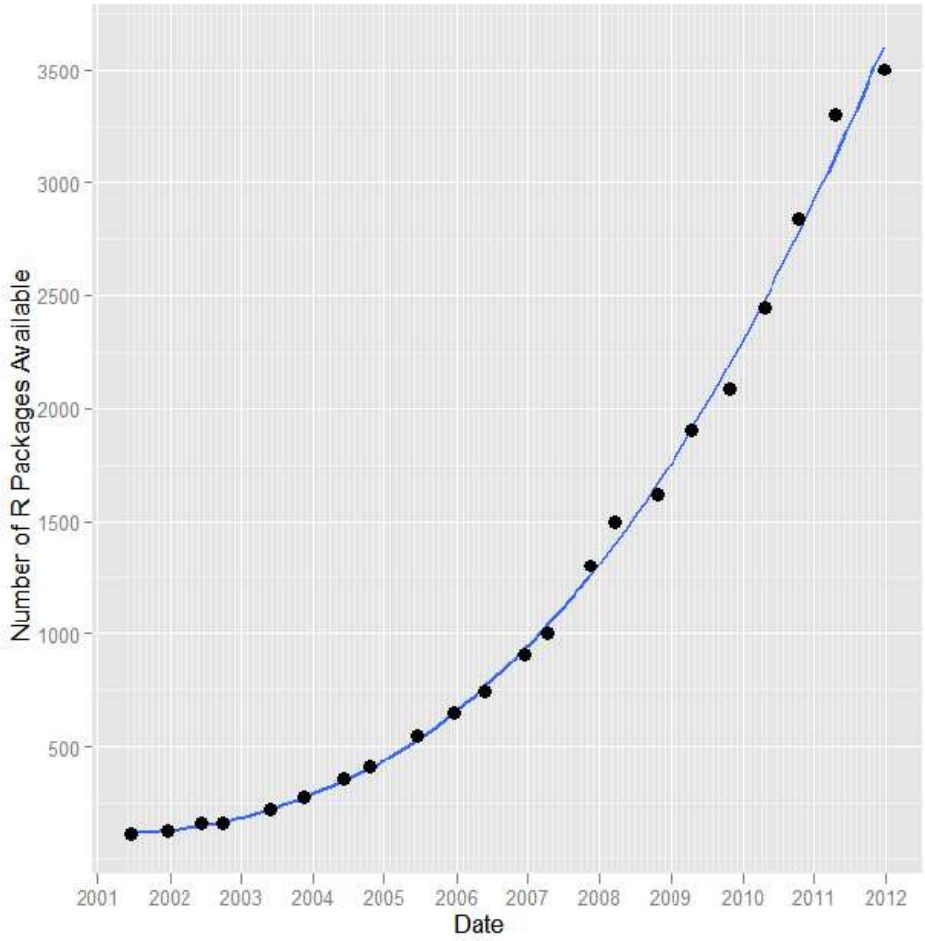
Korea :

http://cran.nexr.com/

- R은 CRAN Site를 통하여 자유롭게 다운로드 받아 설치할 수 있음
- 현재 39개국 87개 Mirror 사이트 운영 중

R Package (http://cran.r-project.org/web/packages/)

R 패키지 수

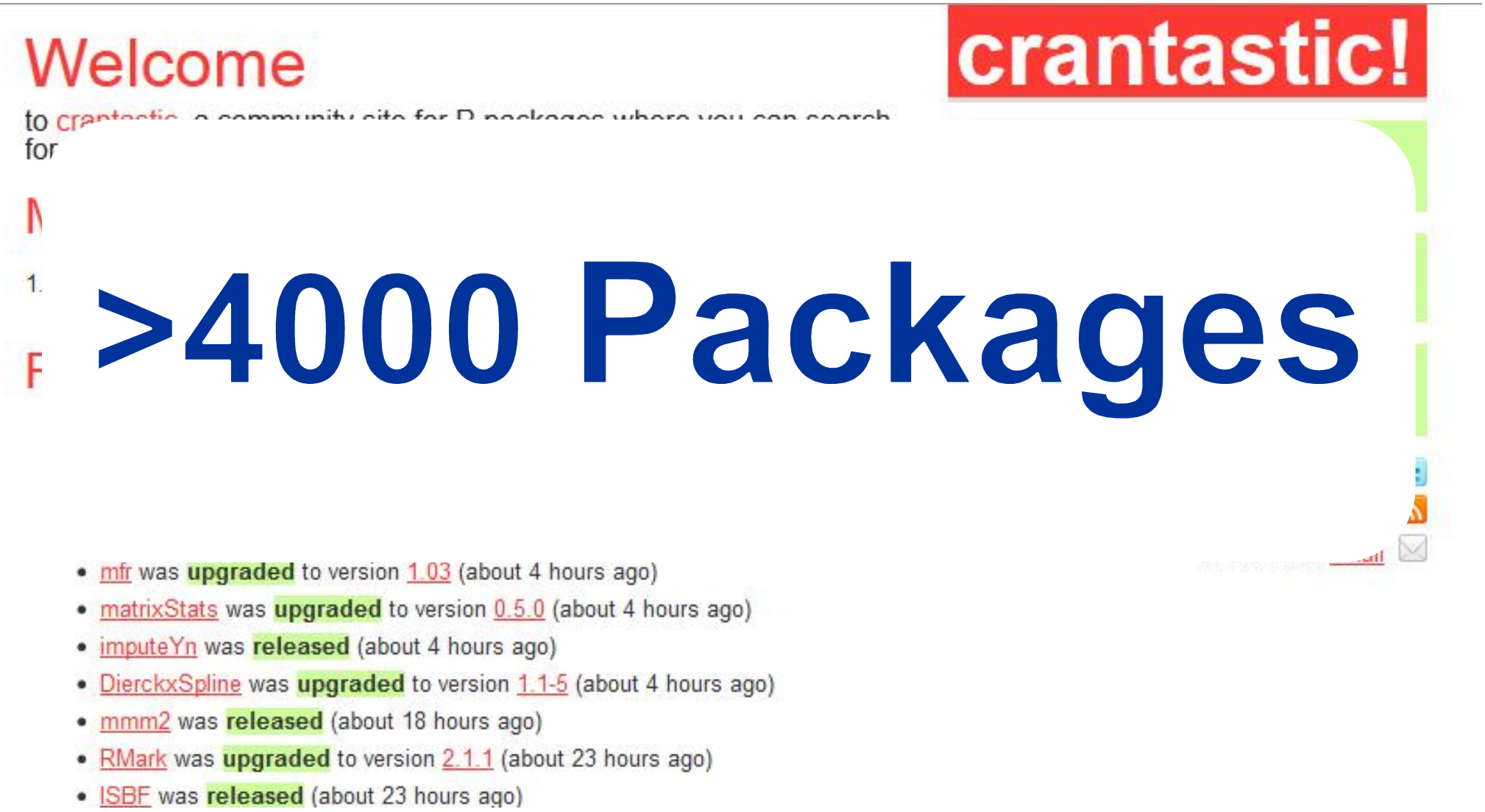


- CRAN Site에 3,938개 등록됨
  (2012년 7월 20일 기준)
- 이러한 패키지들은 새로운 통계분석 알고리즘이나 새로운 IT 기술의 응용에 관한 것을 포함
- Software Vendor에 의하여 Version Up이 되지 않는다는 것이 다른 통계분석 소프트웨어와의 차이임

CRAN Task View (http://cran.r-project.org/web/views/)

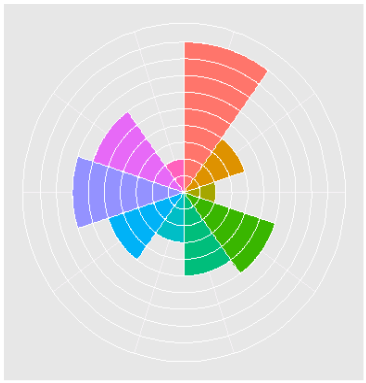
Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Computational Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis
gR	gRaphical Models in R

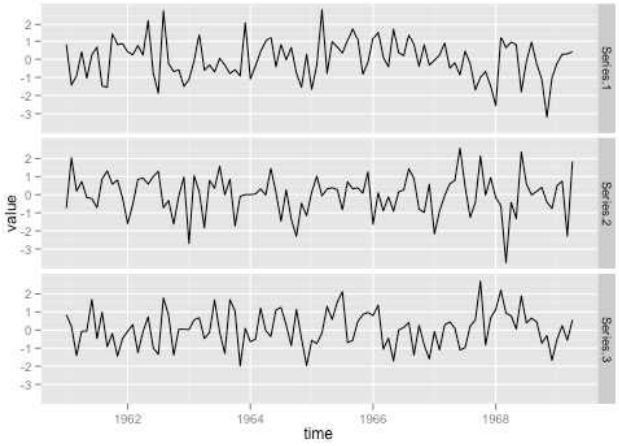
Crantastic! (http://crantastic.org)

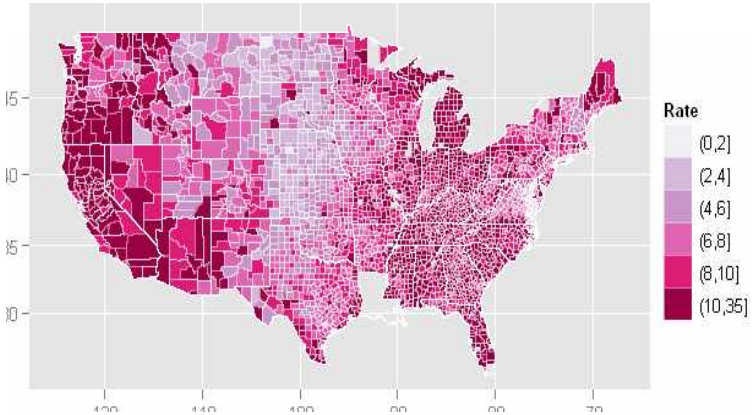


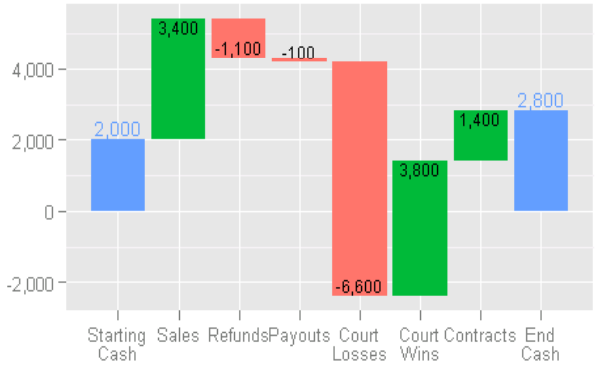
ggplot2 패키지

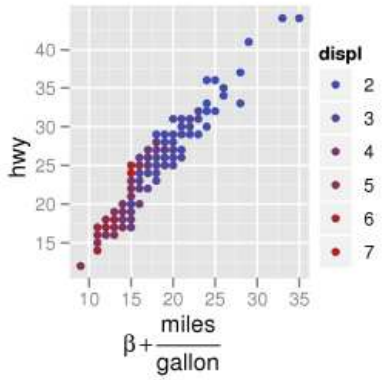
- ggplot2 is a flexible, colorful, and dynamic graphics package:

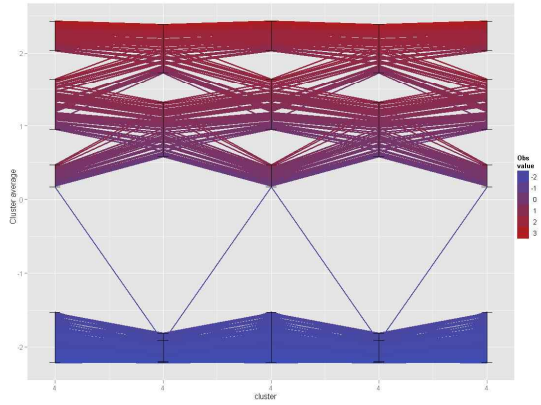








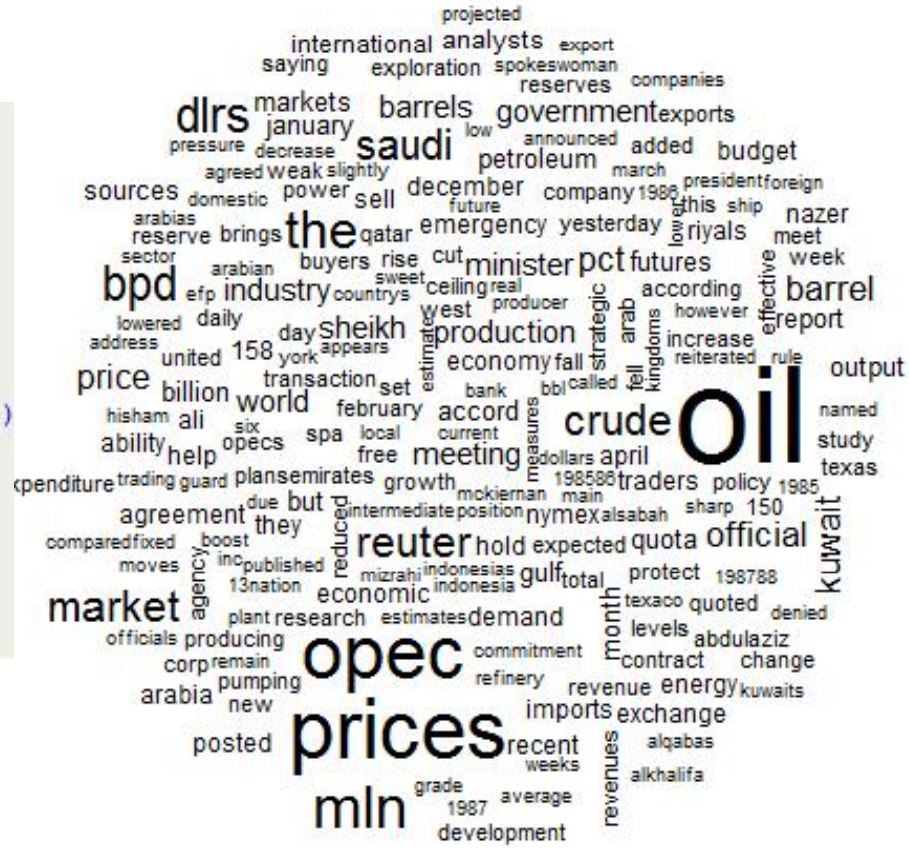




Word Cloud

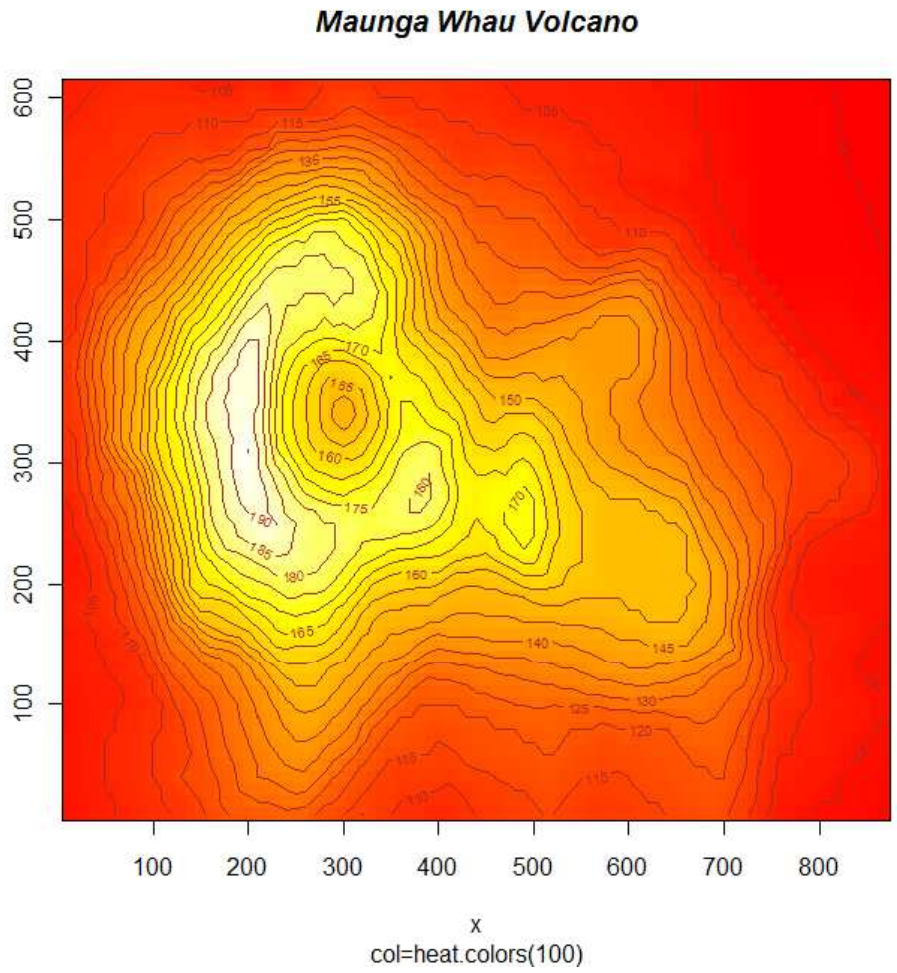
- R에서 text mining 시 많이 사용하는 Word Cloud 결과를 바로 출력

```
> require(tm)
Loading required package: tm
> require(wordcloud)
Loading required package: wordcloud
Loading required package: Rcpp
>
> data(crude)
> crude <- tm_map(crude, removePunctuation)
> crude <- tm_map(crude, function(x)removeWords(x,stopwords()))
> tdm <- TermDocumentMatrix(crude)
> m <- as.matrix(tdm)
> v <- sort(rowSums(m),decreasing=TRUE)
> d <- data.frame(word = names(v),freq=v)
>
> wordcloud(d$word,d$freq)
```



2-D contour map of Maunga Whau volcano

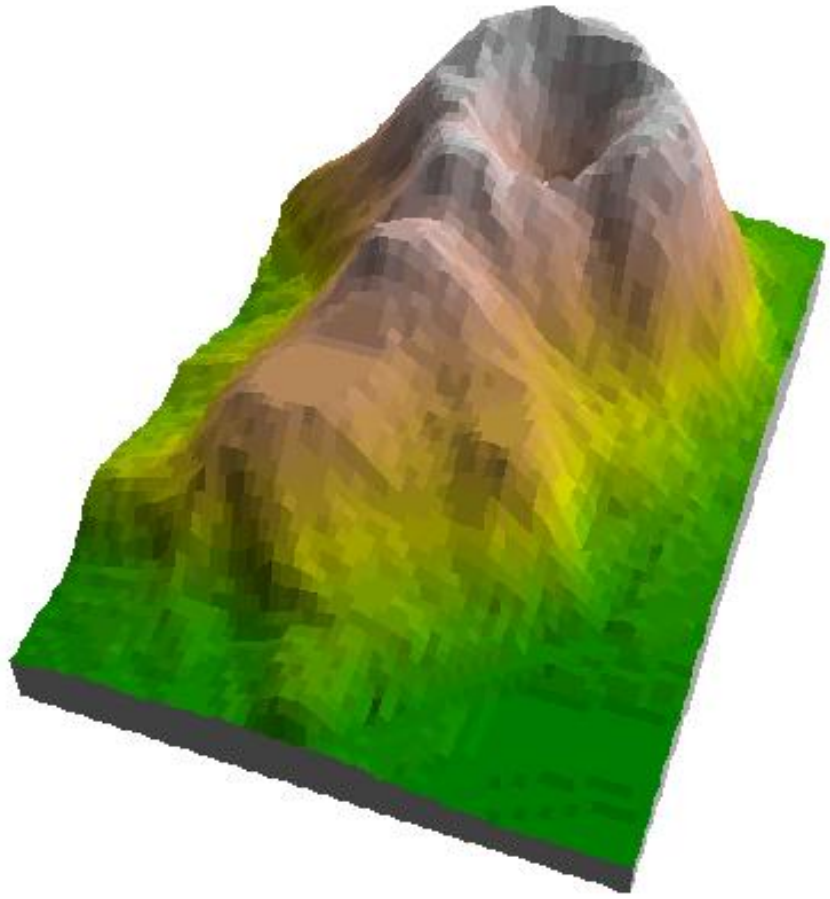
```
> x <- 10*(1:nrow(volcano))
> x.at <- seq(100, 800, by=100)
> y <- 10*(1:ncol(volcano))
> y.at <- seq(100, 600, by=100)
>
> image(x, y, volcano, col=heat.colors(100), axes=FALSE)
> contour(x, y, volcano,
+   levels=seq(90, 200, by=5),
+   add=TRUE,
+   col="brown")
> axis(1, at=x.at)
> axis(2, at=y.at)
> box()
> title(main="Maunga Whau Volcano",
+   sub = "col=heat.colors(100)",
+   font.main=4)
```



3-D contour map of Maunga Whau volcano

- Make it 3-d:

```
> z <- 2 * volcano
> x <- 10 * (1:nrow(z))
> y <- 10 * (1:ncol(z))
>
> z0 <- min(z) - 20
> z <- rbind(z0, cbind(z0, z, z0), z0)
> x <- c(min(x) - 1e-10, x, max(x) + 1e-10)
> y <- c(min(y) - 1e-10, y, max(y) + 1e-10)
>
> fill <- matrix("green3",
+   nr = nrow(z)-1,
+   nc = ncol(z)-1)
> fill[ , i2 <- c(1,ncol(fill))] <- "gray"
> fill[i1 <- c(1,nrow(fill)) , ] <- "gray"
>
> fcol <- fill
> zi <- volcano[ -1,-1] + volcano[ -1,-61] +
+         volcano[-87,-1] + volcano[-87,-61]
> fcol[-i1,-i2] <-
+     terrain.colors(20)[cut(zi, quantile(zi, seq(0,1, len = 21))
+                            include.lowest = TRUE)]
> par(mar=rep(.5,4))
> persp(x, y, 2*z, theta = 110, phi = 40, col = fcol,
+   scale = FALSE,ltheta = -120, shade = 0.4, border = NA,
+   box = FALSE)
```

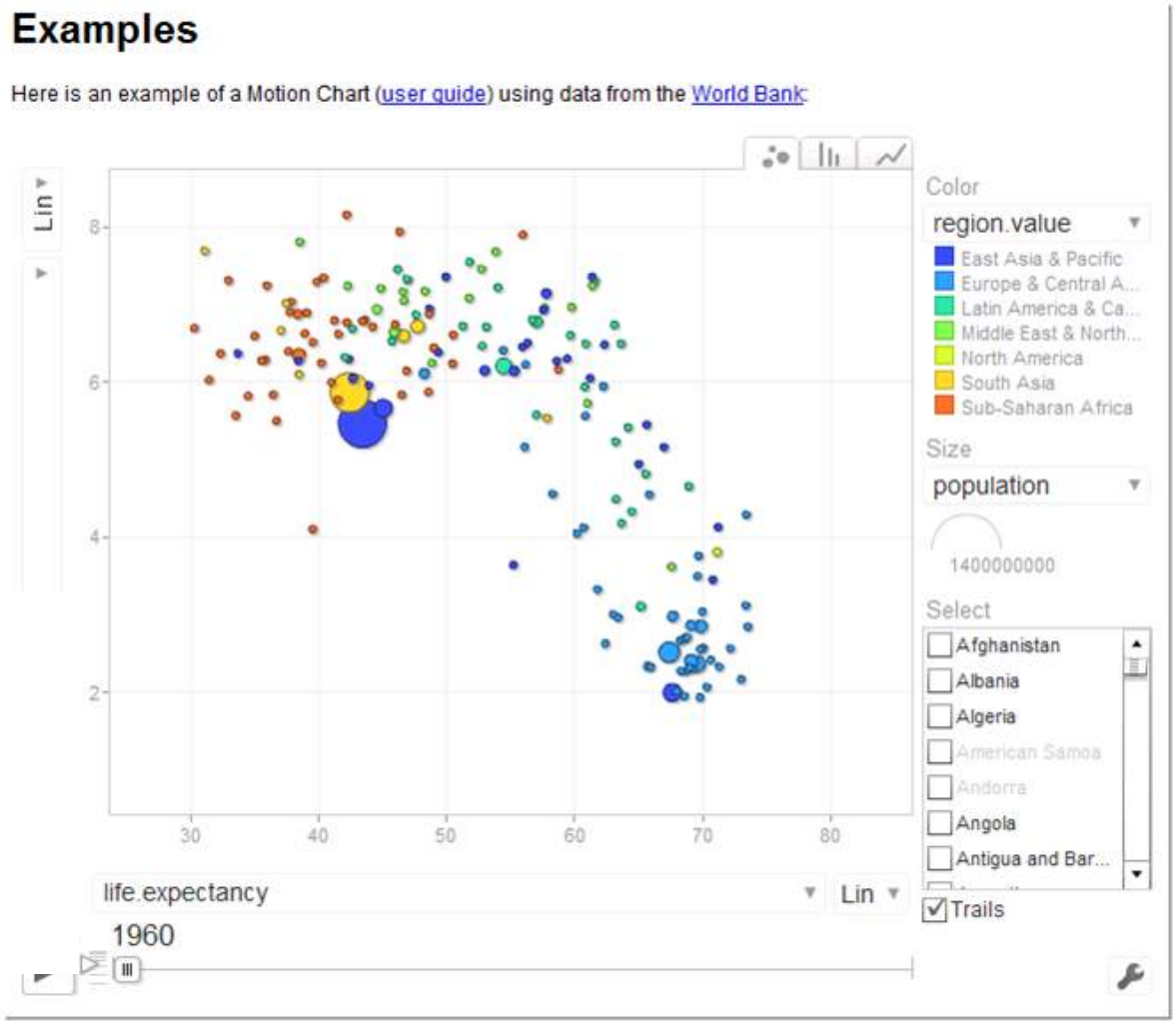


googleVis 패키지

- **R에서 구글 데이터 시각화 API를 이용**

MotionChart 기술은 Gap Miner를
googleVis 에 붙여서 R에서 바로 사용이 가능

Hans Rosling: No more boring data: TEDTalks

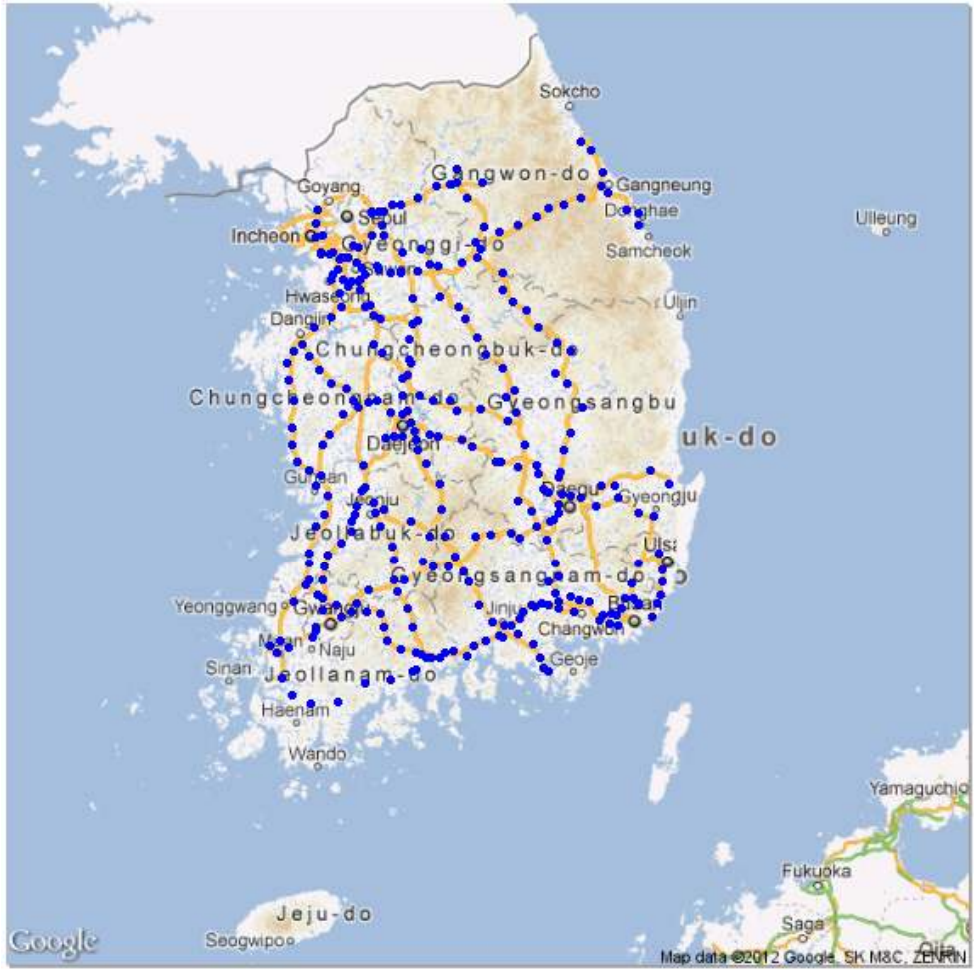


http://code.google.com/p/google-motion-charts-with-r/

RgoogleMaps 패키지

- 구글 지도 상에 다양한 정보를 표출

고속도로 영업소 위치 표출

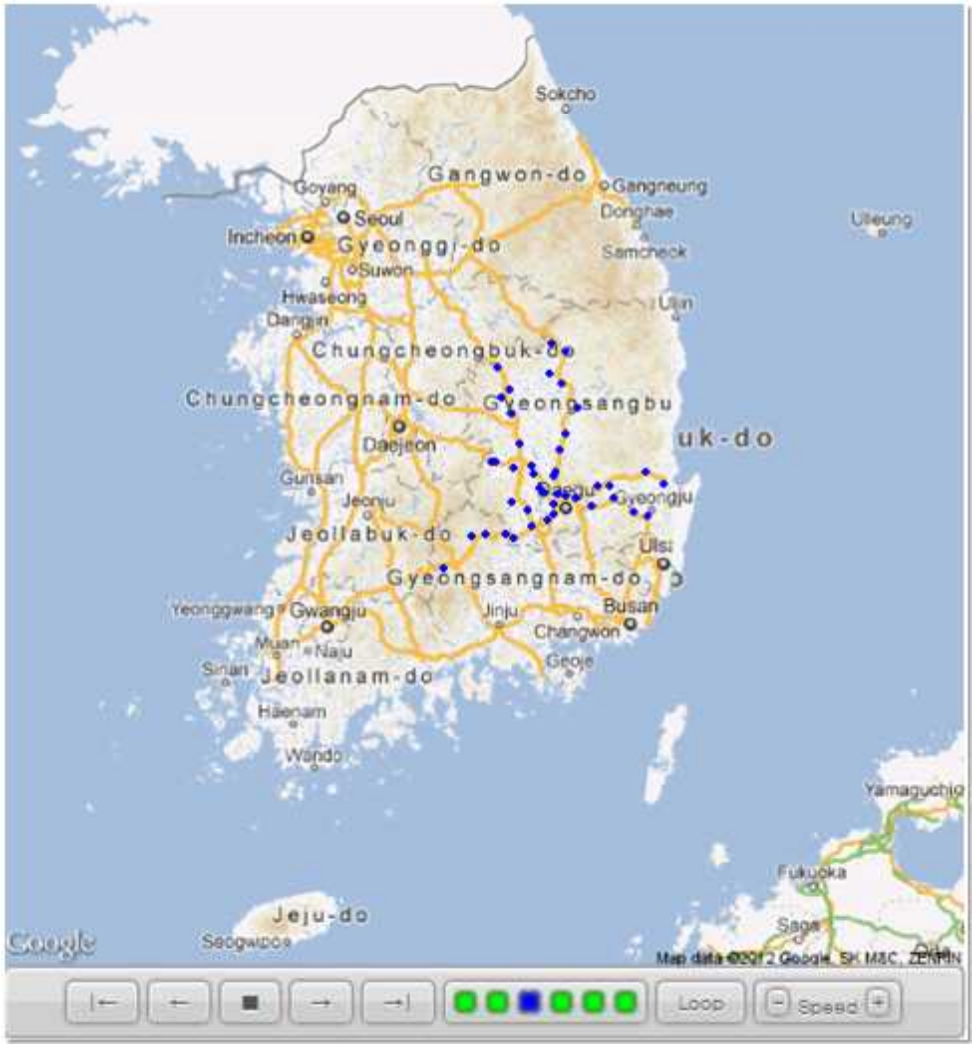



```
# 패키지 로딩...
library(RgoogleMaps)

# 데이터 읽어들이기
tollgate_info <- read.csv("영업소정보.csv")

# 지도 중심 위치 설정
map.center.loc <- c(36, 128)

# 지도 레벨
input.zoom <- 7
map_data <- tollgate_info

# 고속도로 영업소 표시
win.graph()
mymap <- GetMap(center = map.center.loc,
zoom = input.zoom, maptype = "road", format
= "roadmap", destfile = "mymap.png")
PlotOnStaticMap(mymap, lat =
map_data$Y좌표, lon = map_data$X좌표,
destfile = "mymap.point.png", cex = 1, pch
=20, col="blue“)
```

출처 : 베가스 R 소개자료, 김준기

animation 패키지

- **R Graph 결과를 animation으로 생성**

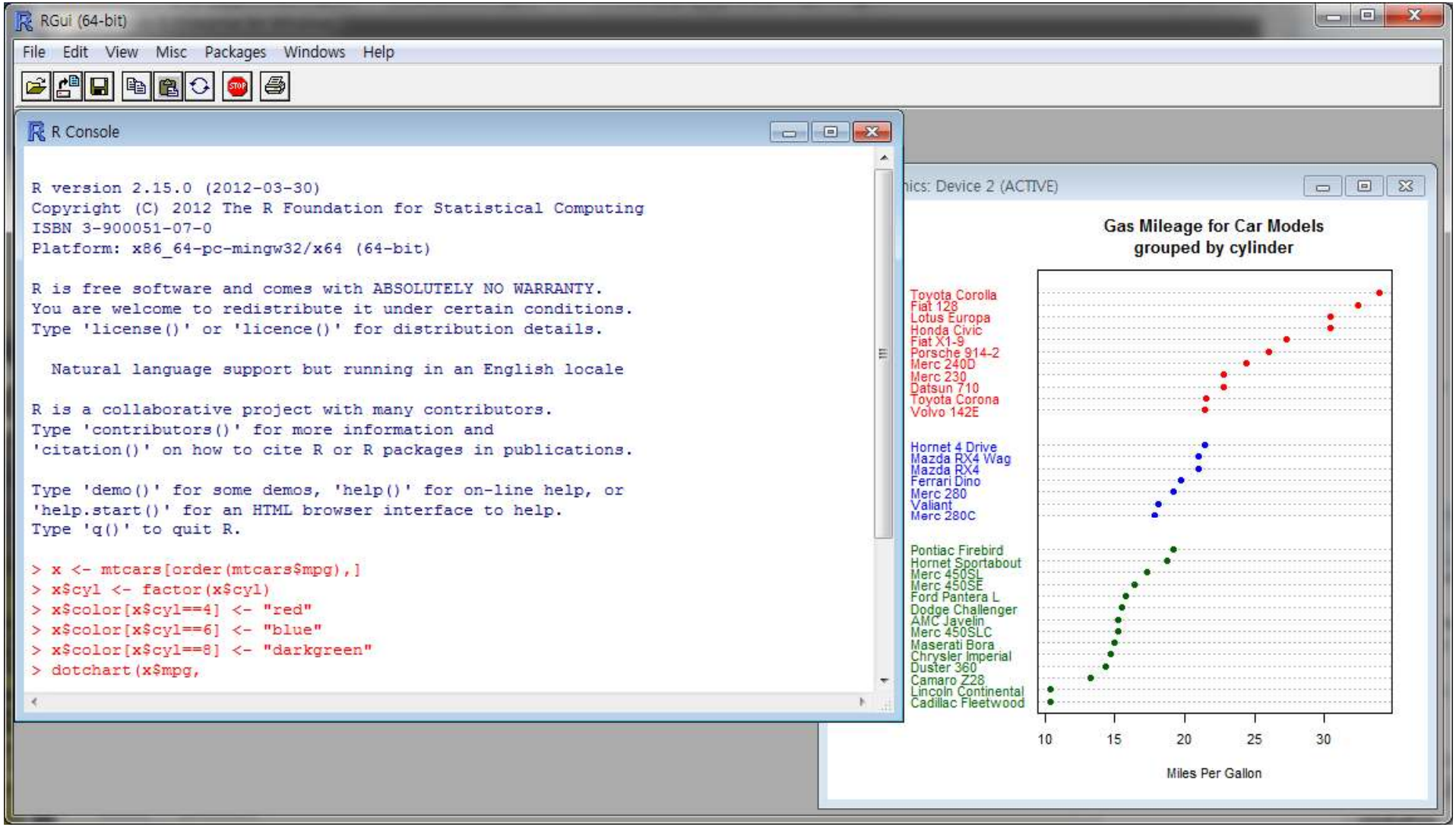



```
# animation으로 생성

saveHTML({
  for(map.i in 1:length(unique.name)) {
    mymap <- GetMap(center =
map.center.loc, zoom = input.zoom, maptype
= "road", format = "roadmap", destfile =
"mymap.png")
    PlotOnStaticMap(mymap, lat =
map_data[map_data$지역본부 ==
unique.name[map.i], ]$Y좌표, lon =
map_data[map_data$지역본부 ==
unique.name[map.i], ]$X좌표, destfile =
"mymap.point.png", cex = 1, pch =20,
col="blue")
  }
}, img.name = "unif_plot", imgdir = "unif_dir",
htmlfile = "random.html",
    autobrowse = FALSE, title = "고속도로
영영소",
    description = c("RgoogleMaps 패키지를
활용한 데모.\n\n"))
```

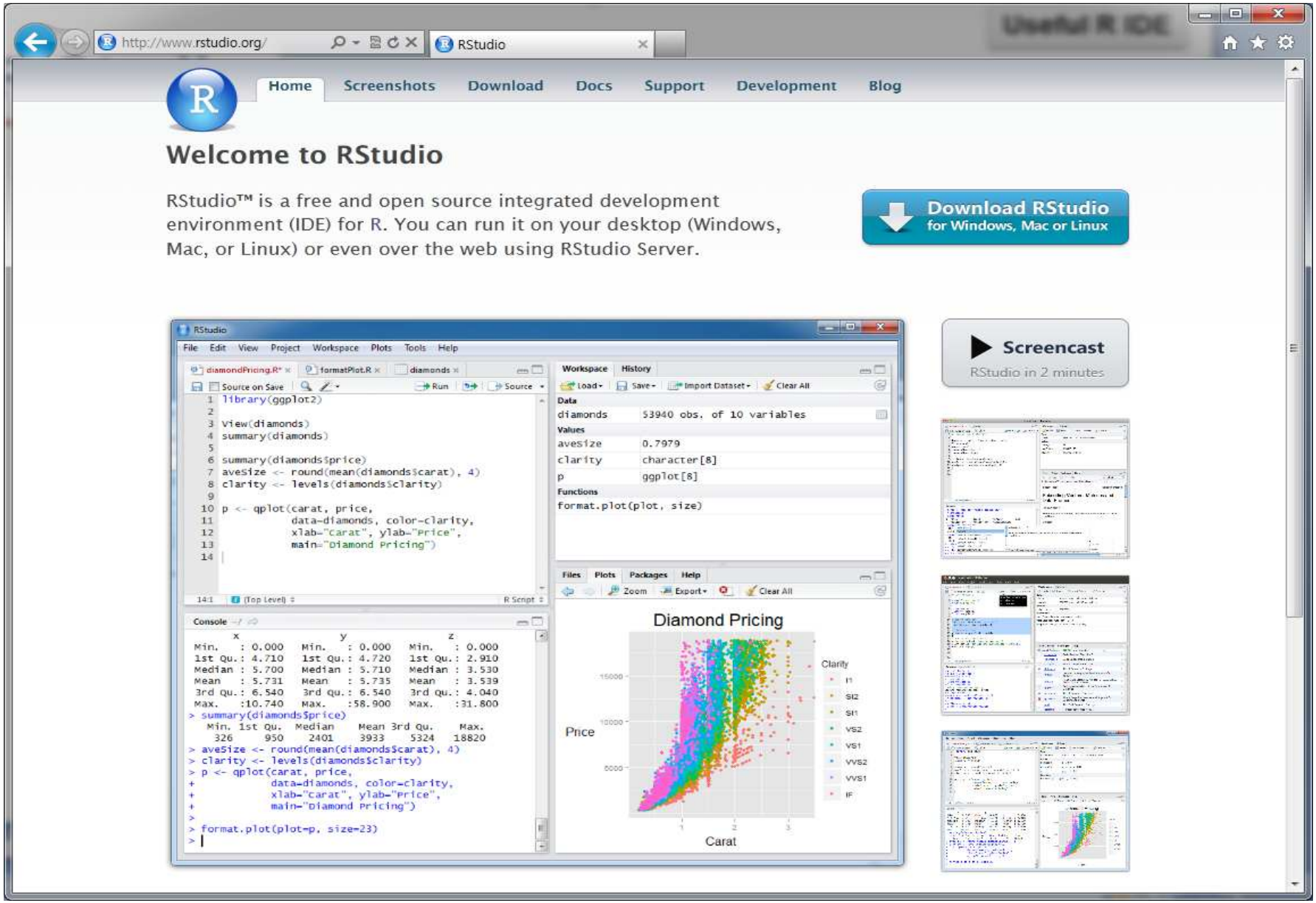
출처 : 베가스 R 소개자료, 김준기

RGUI

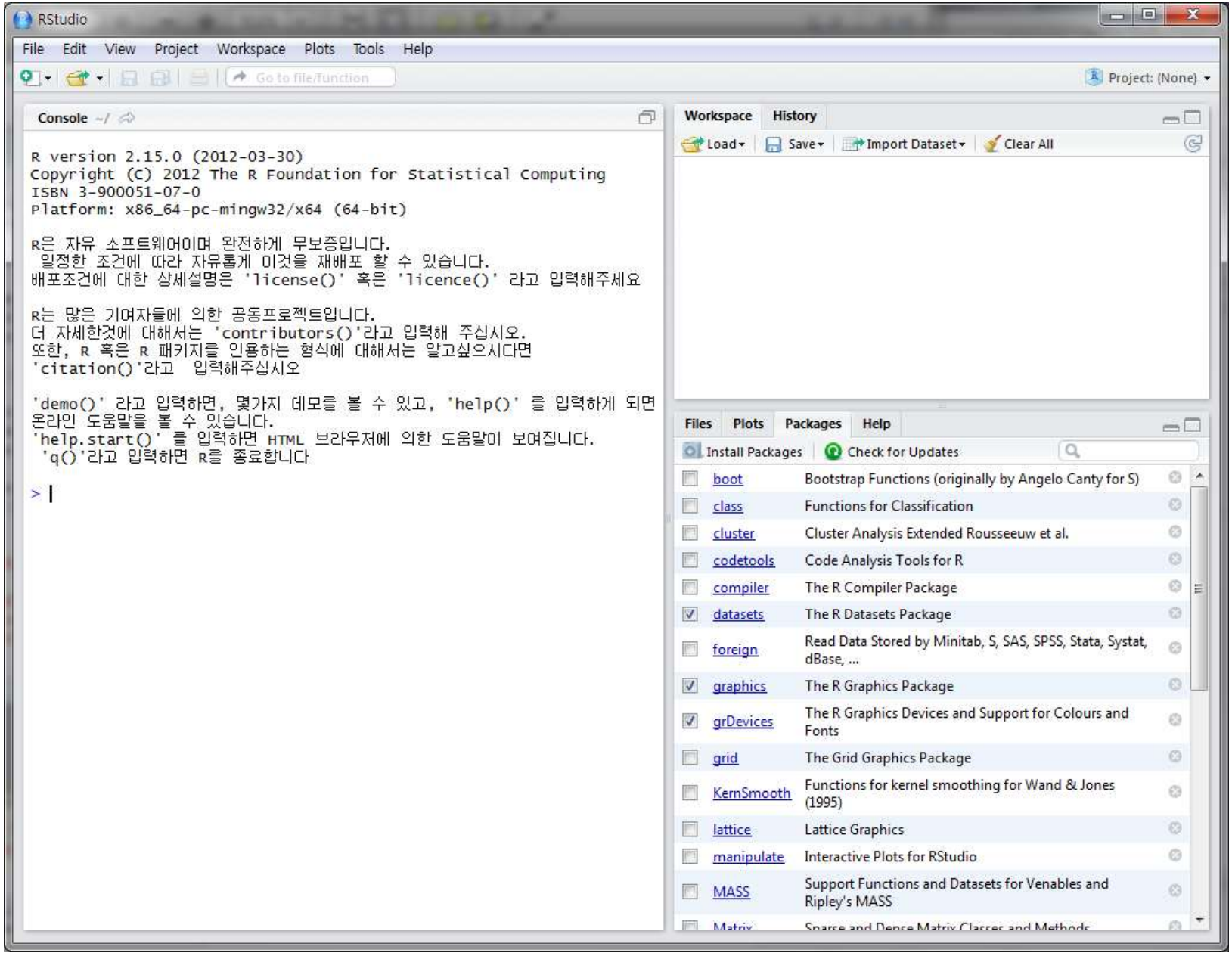
- RGui 실행 기본 화면은 메뉴, 단축아이콘, 콘솔 창으로 구성



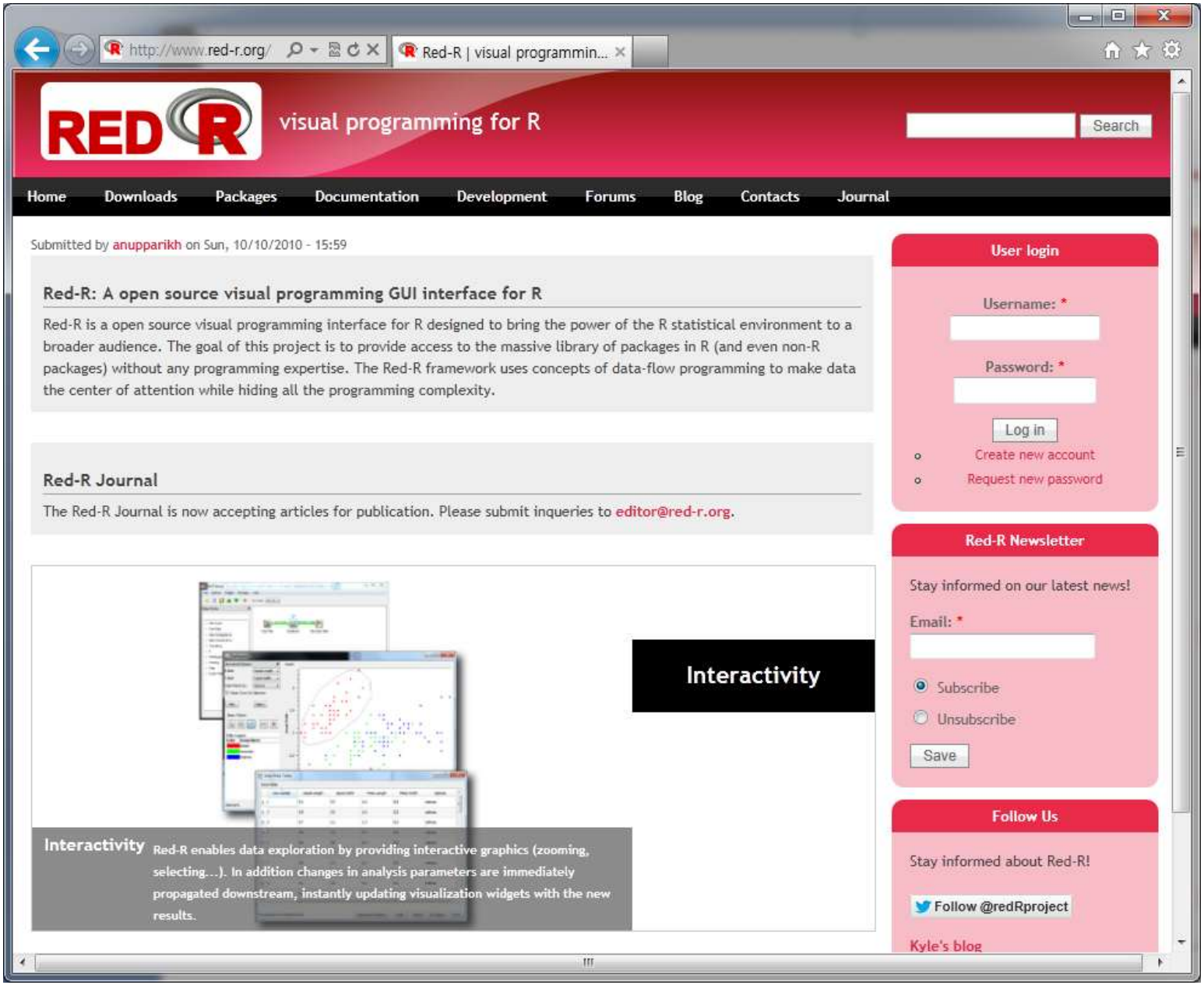
R Studio (http://www.rstudio.org/)



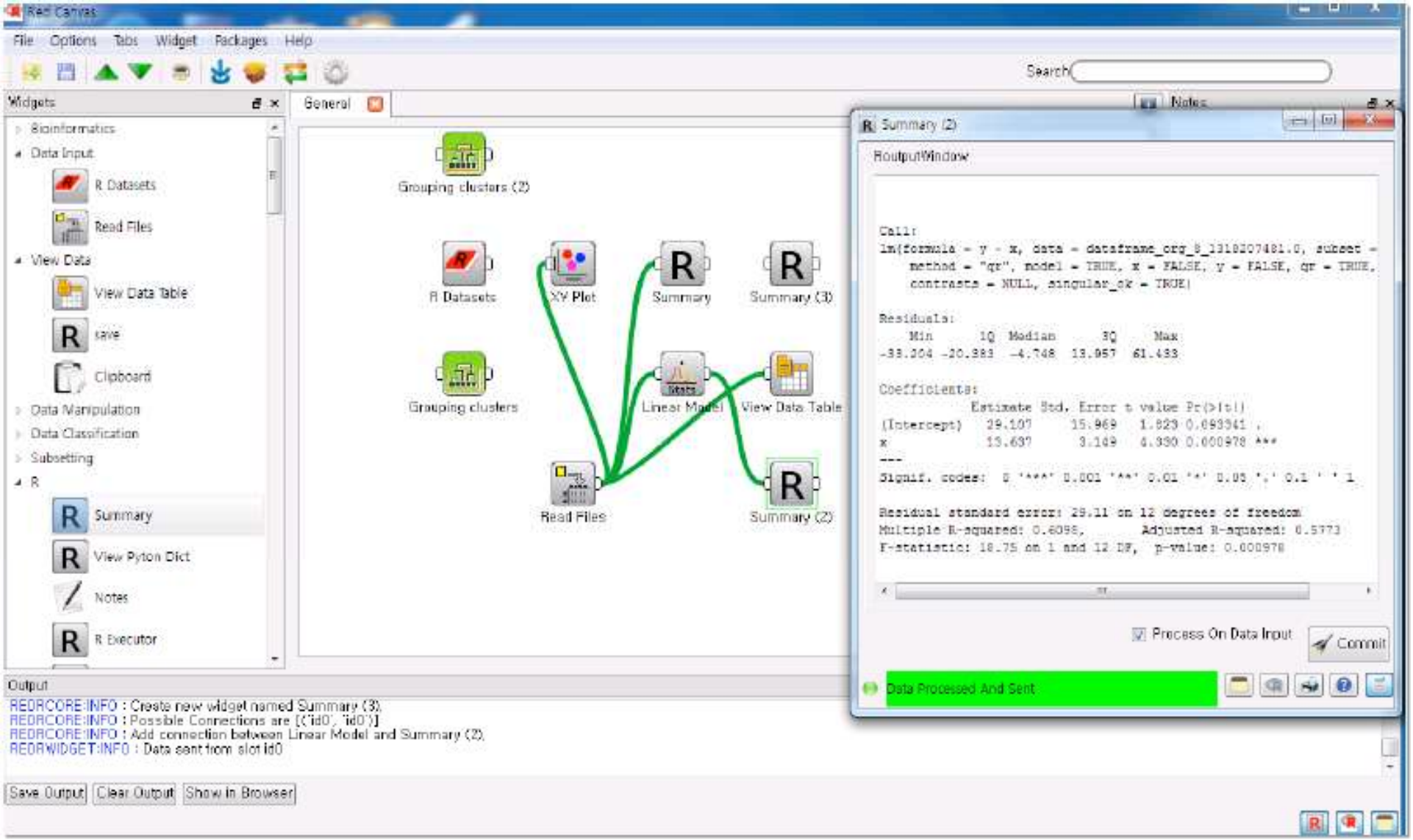
R Studio


RStudio
File Edit View Project Workspace Plots Tools Help
Go to file/function
Project: (None)
Console ~/
R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)
R은 자유 소프트웨어이며 완전하게 무보증입니다.
일정한 조건에 따라 자유롭게 이것을 재배포 할 수 있습니다.
배포조건에 대한 상세설명은 'license()' 혹은 'licence()' 라고 입력해주세요
R는 많은 기여자들에 의한 공동프로젝트입니다.
더 자세한것에 대해서는 'contributors()'라고 입력해 주십시오.
또한, R 혹은 R 패키지를 인용하는 형식에 대해서는 알고싶으시다면
'citation()'라고 입력해주십시오
'demo()' 라고 입력하면, 몇가지 데모를 볼 수 있고, 'help()' 를 입력하게 되면
온라인 도움말을 볼 수 있습니다.
'help.start()' 를 입력하면 HTML 브라우저에 의한 도움말이 보여집니다.
'q()'라고 입력하면 R를 종료합니다
>
Workspace History
Load Save Import Dataset Clear All
Files Plots Packages Help
Install Packages Check for Updates
boot Bootstrap Functions (originally by Angelo Canty for S)
class Functions for Classification
cluster Cluster Analysis Extended Rousseeuw et al.
codetools Code Analysis Tools for R
compiler The R Compiler Package
datasets The R Datasets Package
foreign Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...
graphics The R Graphics Package
grDevices The R Graphics Devices and Support for Colours and Fonts
grid The Grid Graphics Package
KernSmooth Functions for kernel smoothing for Wand & Jones (1995)
lattice Lattice Graphics
manipulate Interactive Plots for RStudio
MASS Support Functions and Datasets for Venables and Ripley's MASS

Red-R (http://www.red-r.org/)

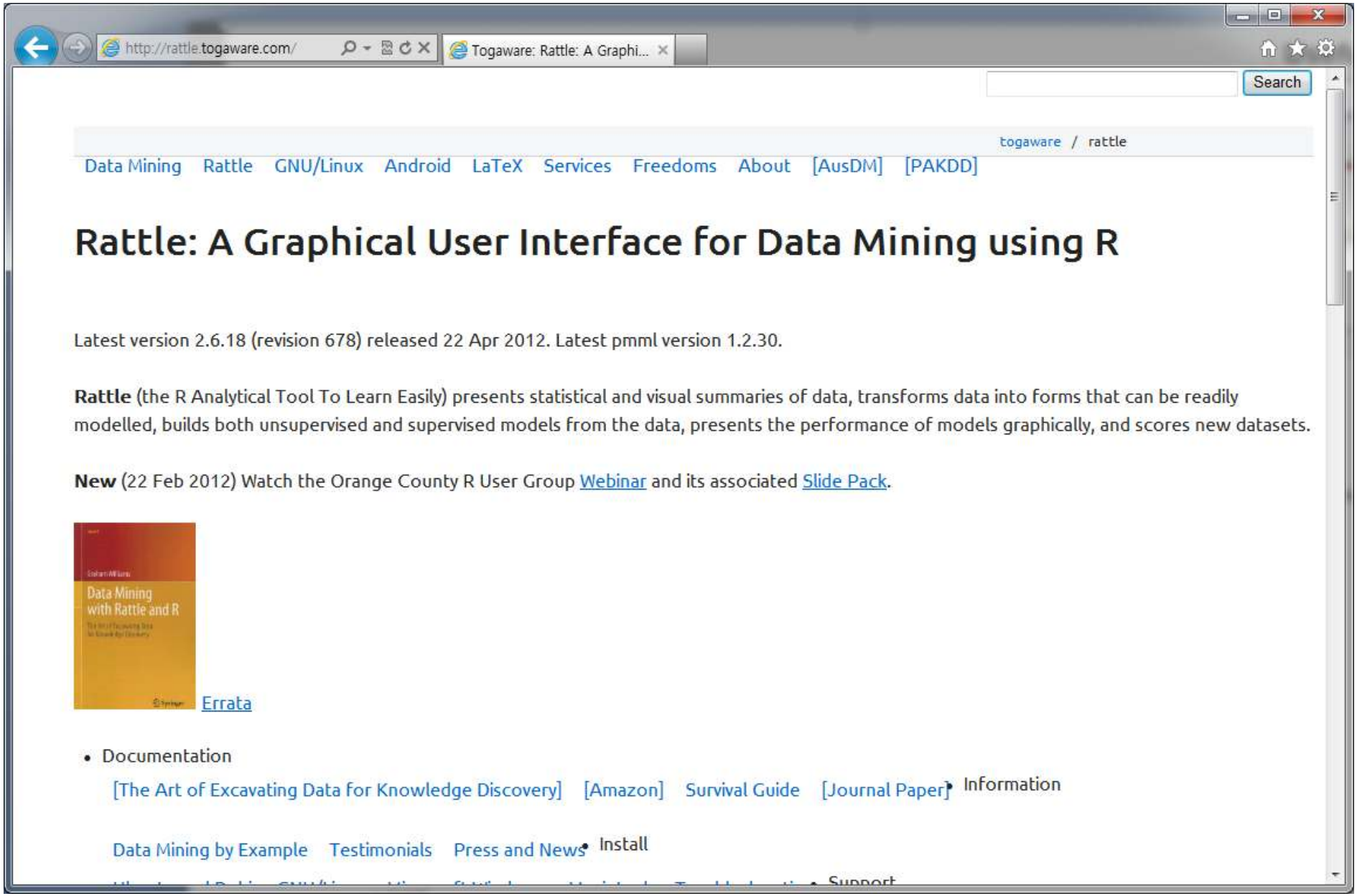


Red-R

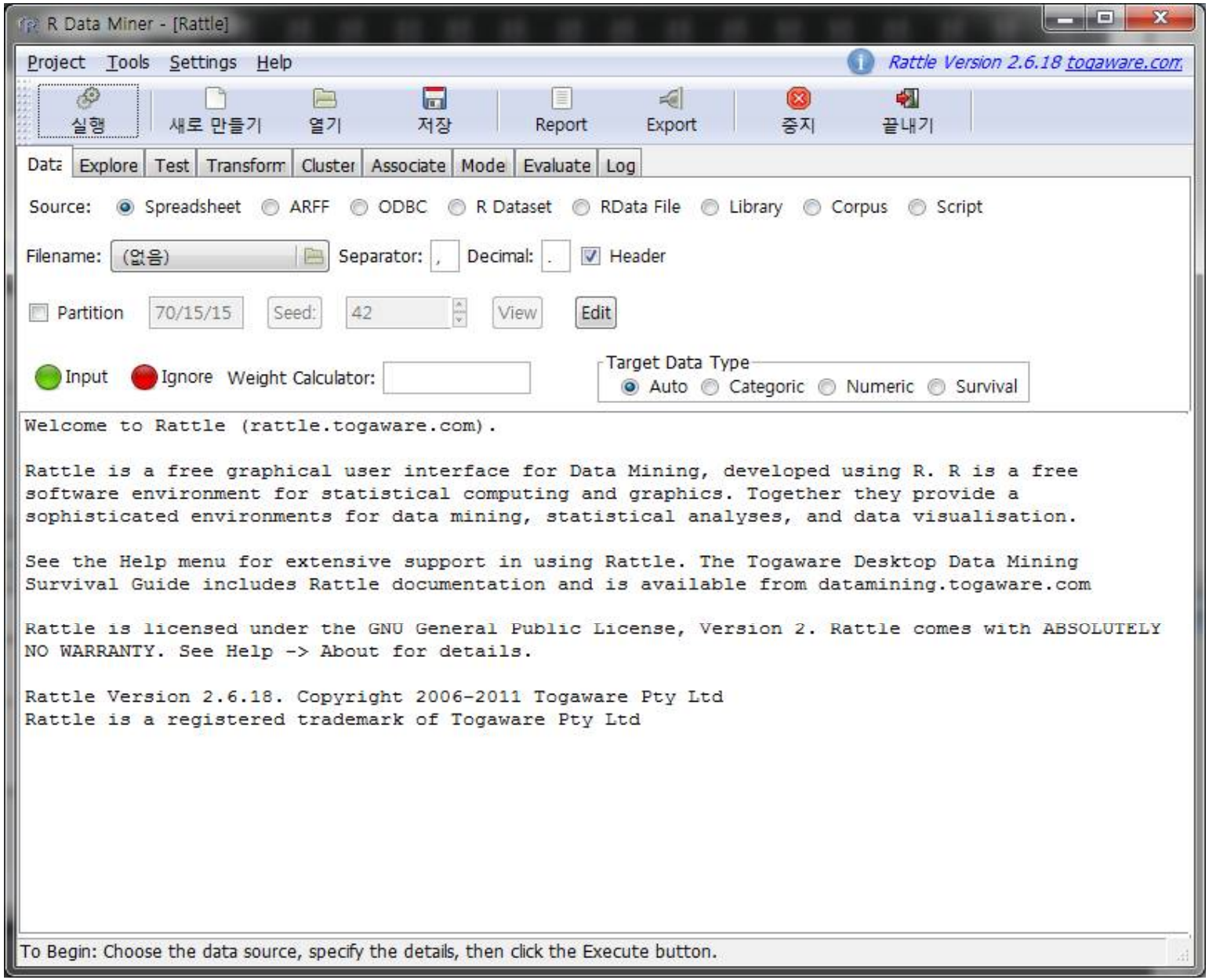

Red Canvas
File Options Tabs Widget Packages Help
Search
Widgets
Bioinformatics
Data Input
R Datasets
Read Files
View Data
View Data Table
save
Clipboard
Data Manipulation
Data Classification
Subsetting
R
Summary
View Pyton Dict
Notes
R Executor
General
Grouping clusters (2)
R Datasets
XY Plot
Summary
Summary (3)
Grouping clusters
Linear Model
View Data Table
Read Files
Summary (2)
Notes
Summary (2)
RoutputWindow
Call:
lm(formula = y ~ x, data = dataframe_org_8_1310207481.0, subset =
method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
contrasts = NULL, singular.ok = TRUE)
Residuals:
Min 1Q Median 3Q Max
-33.304 -20.383 -4.748 13.957 61.433
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.107 15.969 1.823 0.093941 .
x 13.637 3.149 4.330 0.000978 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 29.11 on 12 degrees of freedom
Multiple R-squared: 0.6098, Adjusted R-squared: 0.5773
F-statistic: 18.75 on 1 and 12 DF, p-value: 0.000978
Process On Data Input
Commit
Data Processed And Sent
Output
REDRCORE:INFO : Create new widget named Summary (3)
REDRCORE:INFO : Possible Connections are [('id0', 'id0')]
REDRCORE:INFO : Add connection between Linear Model and Summary (2)
REDRWIDGET:INFO : Data sent from slot id0
Save Output
Clear Output
Show in Browser

Rattle (http://rattle.togaware.com/), (install.packages("rattle"))

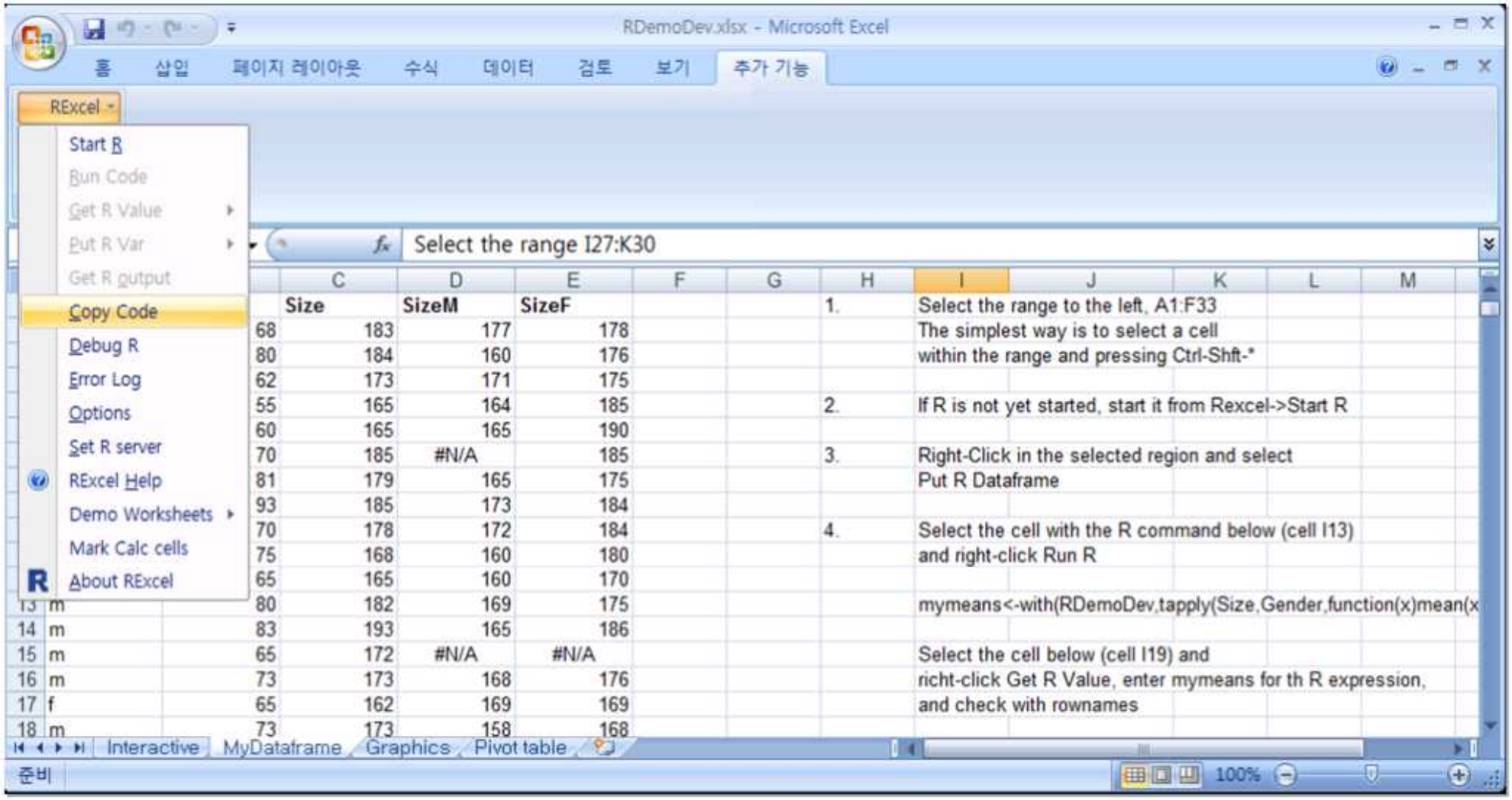


Rattle (library(rattle); rattle())

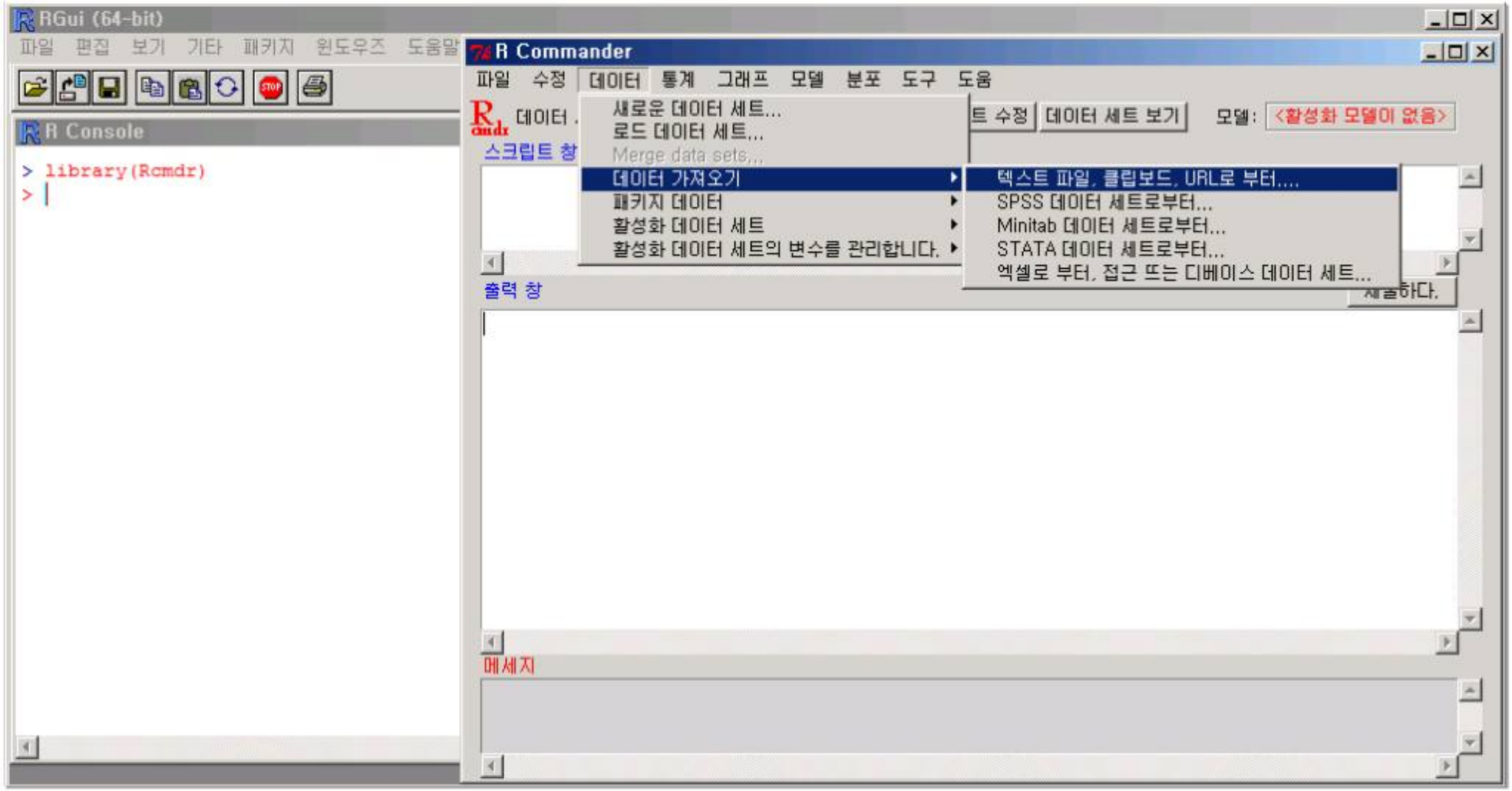


RExcel (http://rcom.univie.ac.at/)

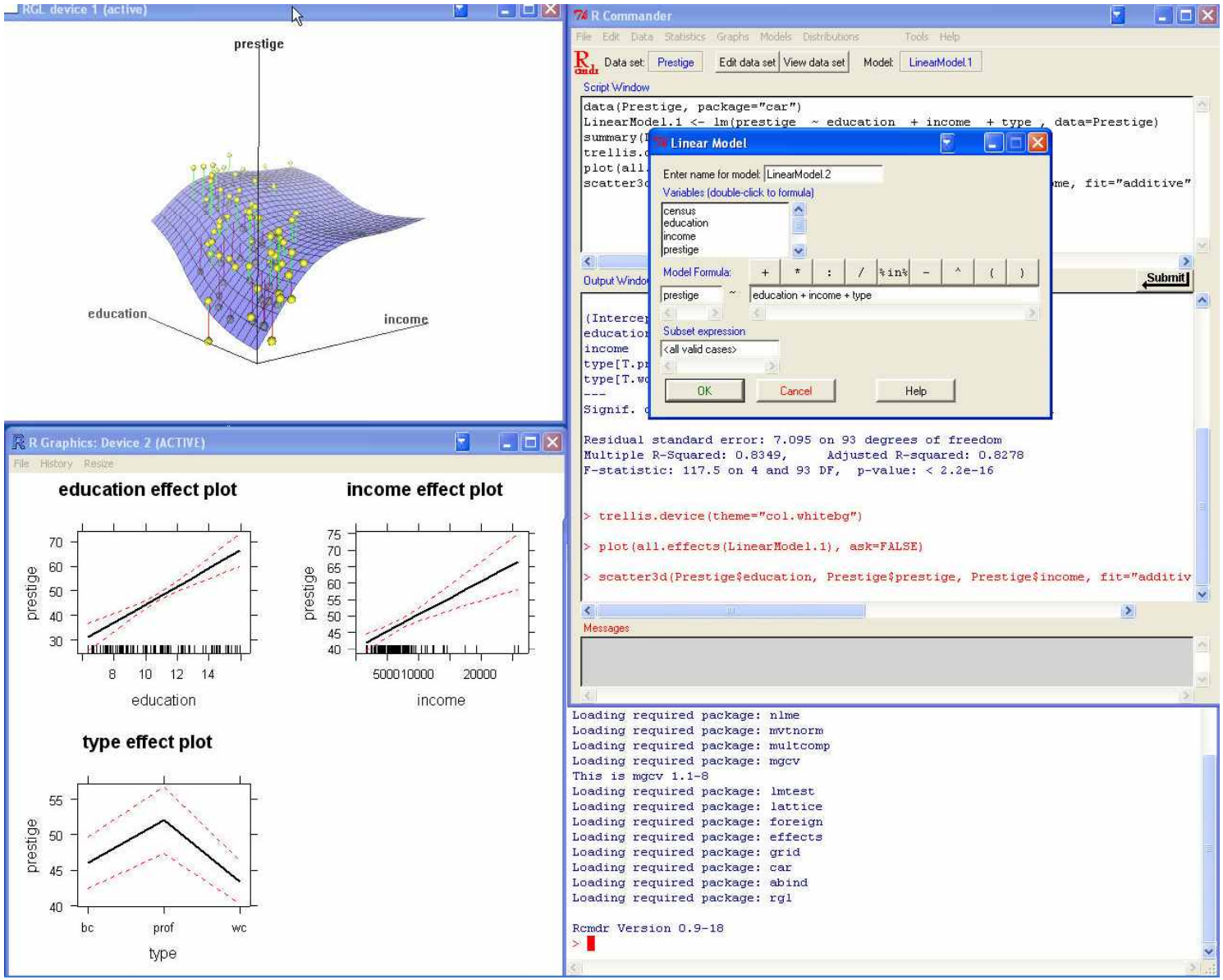
RExcel


RDemoDev.xlsx - Microsoft Excel
RExcel
Start R
Run Code
Get R Value
Put R Var
Get R output
Copy Code
Debug R
Error Log
Options
Set R server
RExcel Help
Demo Worksheets
Mark Calc cells
About RExcel
Select the range I27:K30
Size SizeM SizeF
1. Select the range to the left, A1:F33
The simplest way is to select a cell
within the range and pressing Ctrl-Shft-*
2. If R is not yet started, start it from Rexcel->Start R
3. Right-Click in the selected region and select
Put R Dataframe
4. Select the cell with the R command below (cell I13)
and right-click Run R
mymeans<-with(RDemoDev,tapply(Size,Gender,function(x)mean(x
Select the cell below (cell I19) and
richt-click Get R Value, enter mymeans for th R expression,
and check with rownames
Interactive
MyDataframe
Graphics
Pivot table
준비
100%

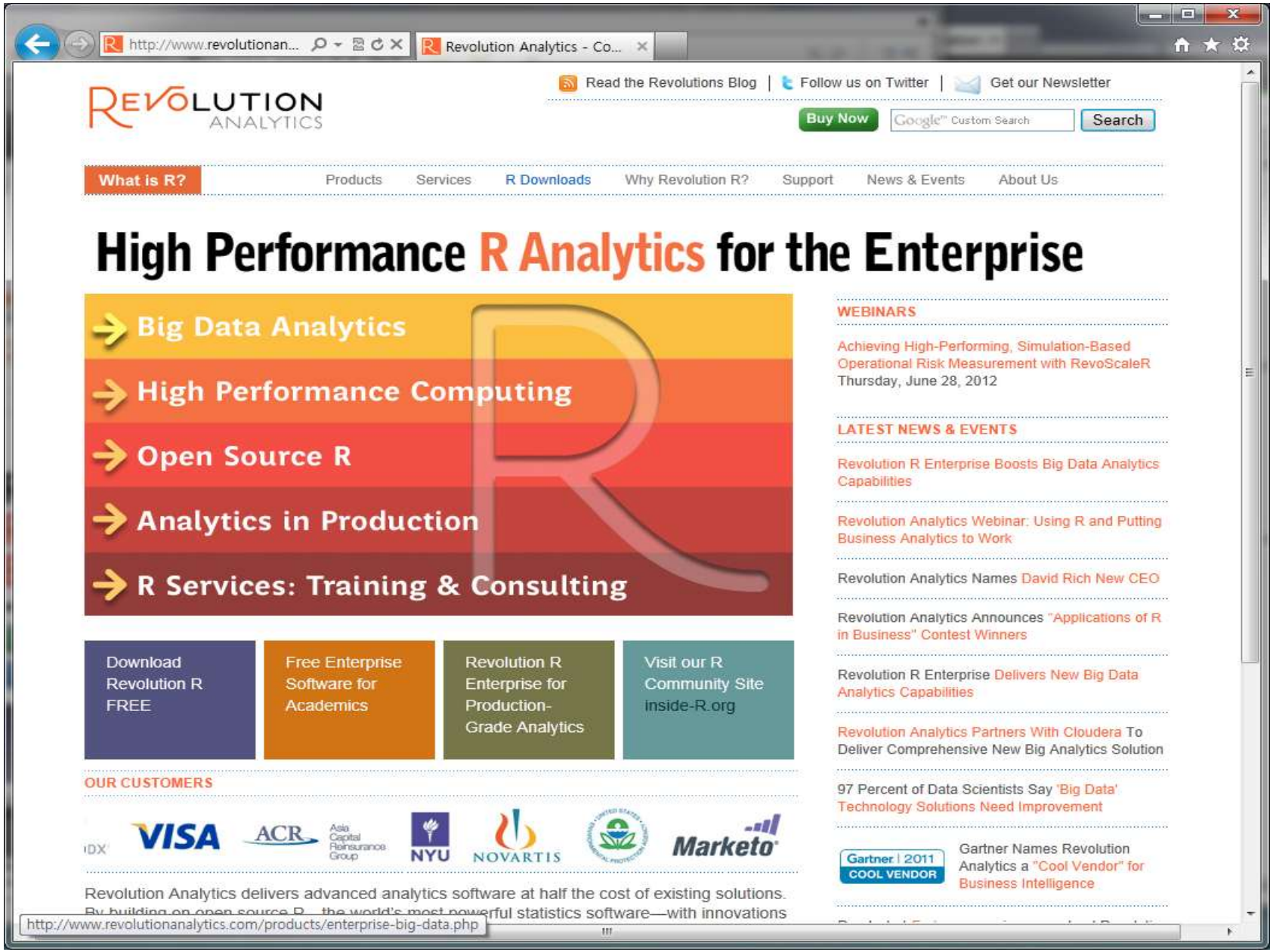
R Commander (install.packages("Rcmdr"), library(Rcmdr))



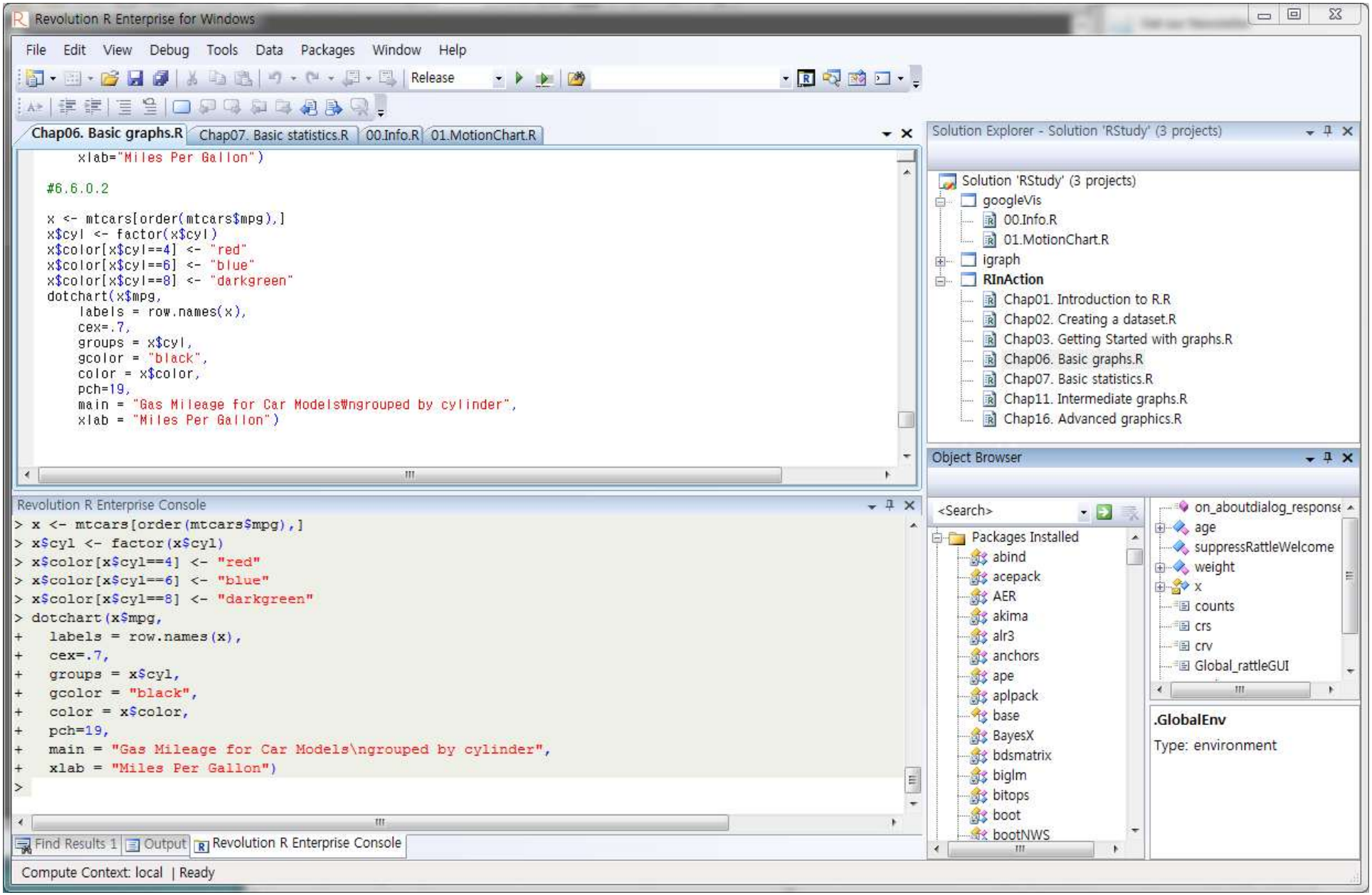
R Commander


RGL device 1 (active)
prestige
education
income
R Graphics: Device 2 (ACTIVE)
education effect plot
income effect plot
type effect plot
prestige
education
income
type
R Commander
Data set: Prestige
Edit data set
View data set
Model: LinearModel.1
Script Window
Linear Model
Enter name for model: LinearModel.2
Variables (double-click to formula)
census
education
income
prestige
Model Formula:
prestige ~ education + income + type
Subset expression
<all valid cases>
OK
Cancel
Help
Submit
Residual standard error: 7.095 on 93 degrees of freedom
Multiple R-Squared: 0.8349, Adjusted R-squared: 0.8278
F-statistic: 117.5 on 4 and 93 DF, p-value: < 2.2e-16
> trellis.device(theme="col.whitebg")
> plot(all.effects(LinearModel.1), ask=FALSE)
> scatter3d(Prestige$education, Prestige$prestige, Prestige$income, fit="additiv
Messages
Loading required package: nlme
Loading required package: mvtnorm
Loading required package: multcomp
Loading required package: mgcv
This is mgcv 1.1-8
Loading required package: lmtest
Loading required package: lattice
Loading required package: foreign
Loading required package: effects
Loading required package: grid
Loading required package: car
Loading required package: abind
Loading required package: rgl
Rcmdr Version 0.9-18

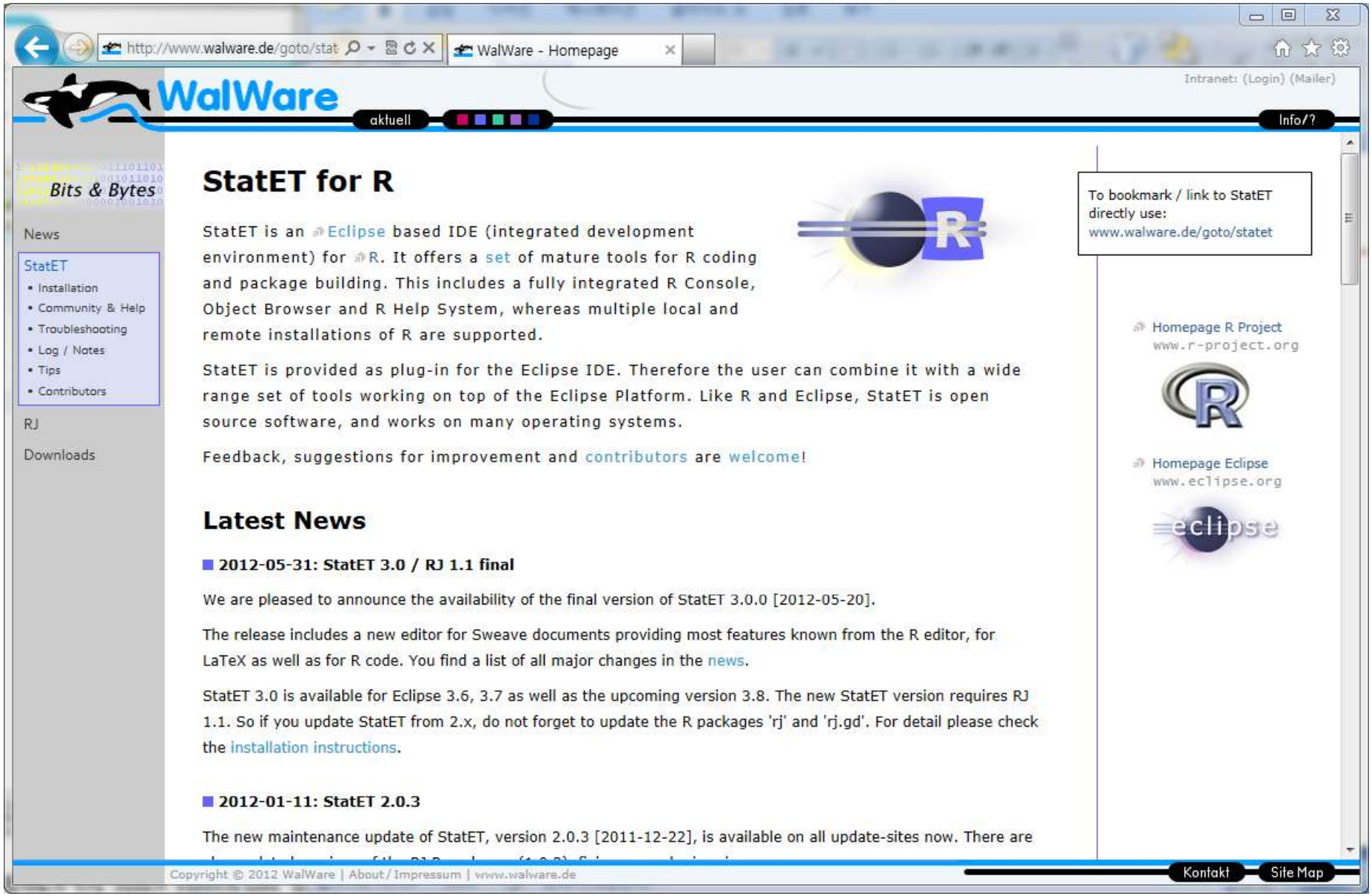
Revolution R (http://www.revolutionanalytics.com/)



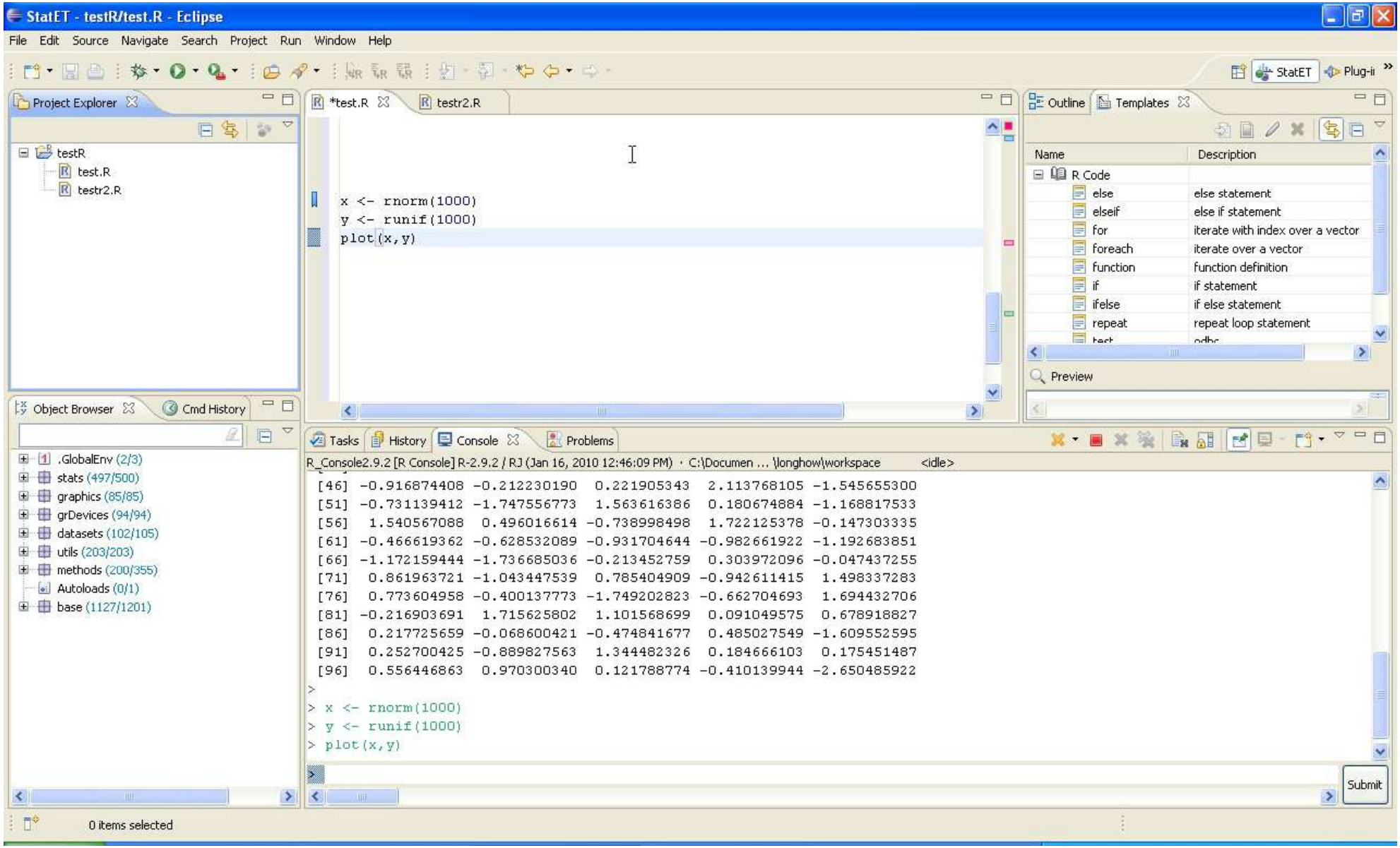
Revolution R



StatET for R (http://www.walware.de/goto/statet)



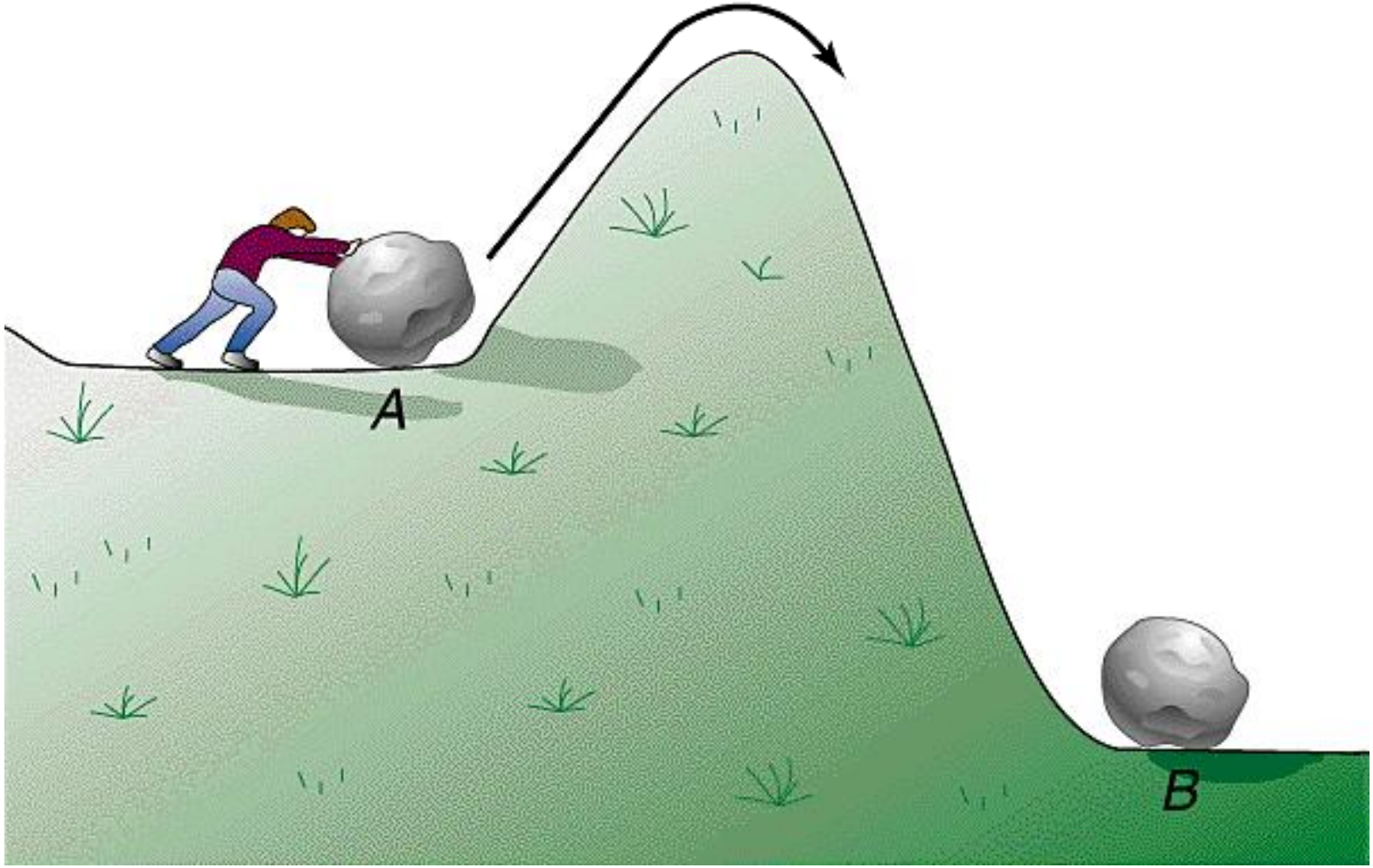
StatET for R


StatET - testR/test.R - Eclipse
File Edit Source Navigate Search Project Run Window Help
Project Explorer
testR
test.R
testr2.R
*test.R
testr2.R
x <- rnorm(1000)
y <- runif(1000)
plot(x,y)
Outline
Templates
Name
Description
R Code
else
else statement
elseif
else if statement
for
iterate with index over a vector
foreach
iterate over a vector
function
function definition
if
if statement
ifelse
if else statement
repeat
repeat loop statement
Preview
Object Browser
Cmd History
.GlobalEnv (2/3)
stats (497/500)
graphics (85/85)
grDevices (94/94)
datasets (102/105)
utils (203/203)
methods (200/355)
Autoloads (0/1)
base (1127/1201)
Tasks
History
Console
Problems
R_Console2.9.2 [R Console] R-2.9.2 / RJ (Jan 16, 2010 12:46:09 PM) - C:\Documen ... \longhow\workspace
<idle>
[46] -0.916874408 -0.212230190 0.221905343 2.113768105 -1.545655300
[51] -0.731139412 -1.747556773 1.563616386 0.180674884 -1.168817533
[56] 1.540567088 0.496016614 -0.738998498 1.722125378 -0.147303335
[61] -0.466619362 -0.628532089 -0.931704644 -0.982661922 -1.192683851
[66] -1.172159444 -1.736685036 -0.213452759 0.303972096 -0.047437255
[71] 0.861963721 -1.043447539 0.785404909 -0.942611415 1.498337283
[76] 0.773604958 -0.400137773 -1.749202823 -0.662704693 1.694432706
[81] -0.216903691 1.715625802 1.101568699 0.091049575 0.678918827
[86] 0.217725659 -0.068600421 -0.474841677 0.485027549 -1.609552595
[91] 0.252700425 -0.889827563 1.344482326 0.184666103 0.175451487
[96] 0.556446863 0.970300340 0.121788774 -0.410139944 -2.650485922
> x <- rnorm(1000)
> y <- runif(1000)
> plot(x,y)
Submit
0 items selected

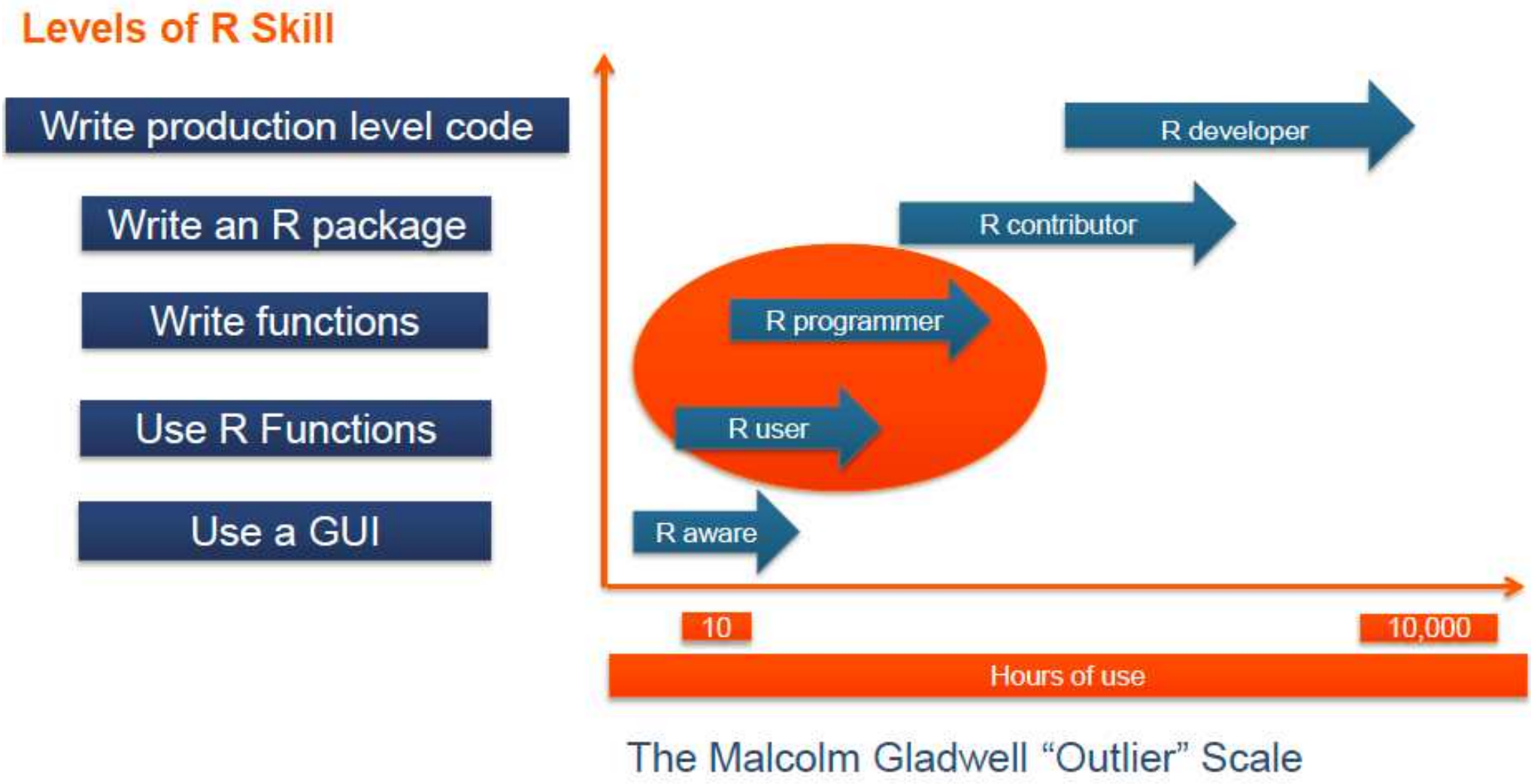
R을 배우는데 도움이 될 만한 곳

- http://www.r-project.org
- http://www.r-project.kr (한국 R 사용자 모임 - KRUG)
- http://www.r-bloggers.com
- http://stackoverflow.com
- http://stats.stackexchange.com
- http://www.inside-r.org/
- http://www.r-statistics.com/
- http://support.rstudio.org/
- http://quora.com

R을 배워 업무에 적용하려면



Learning R



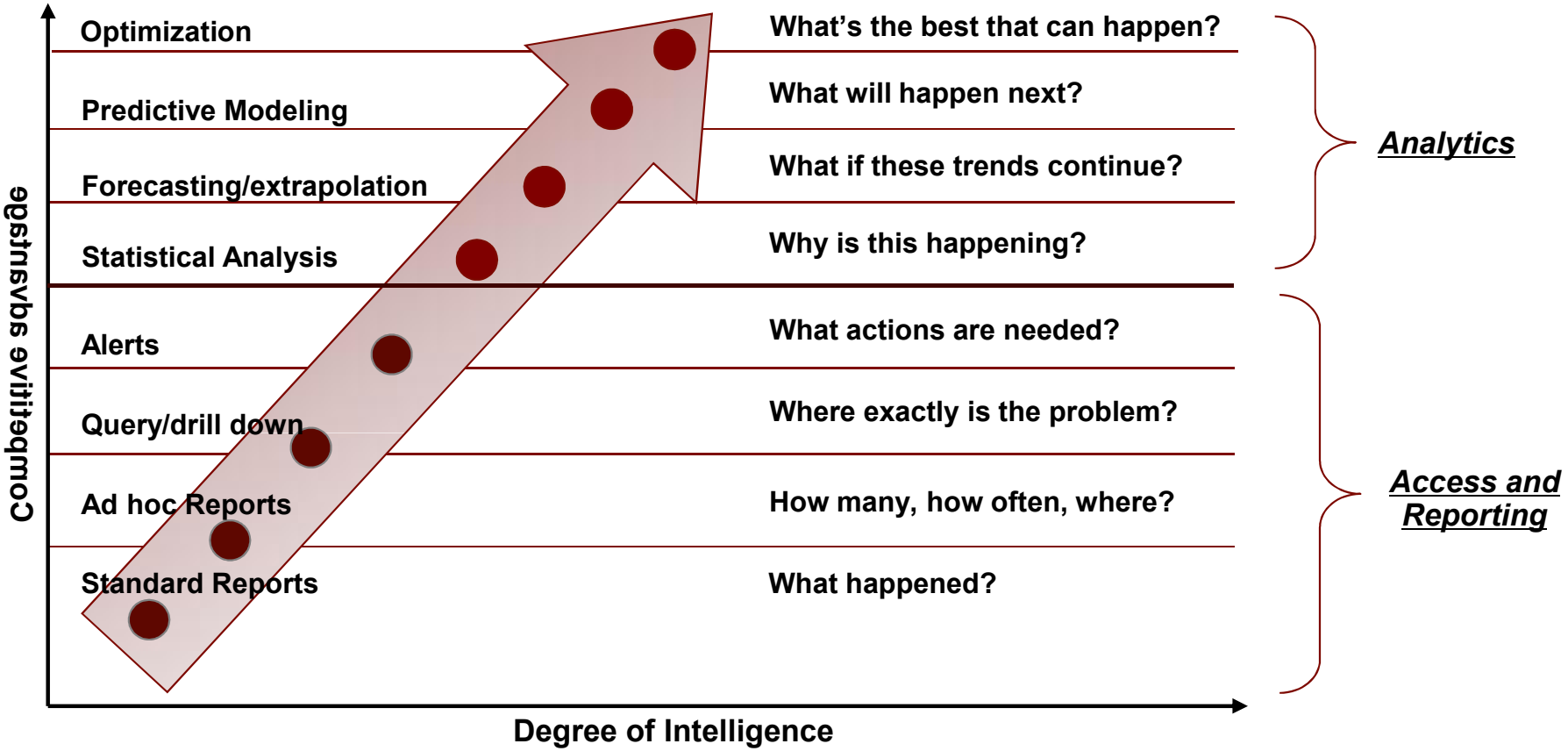
출처 : Revolution Analytics

R, 그리고 빅데이터

Advanced Analytics

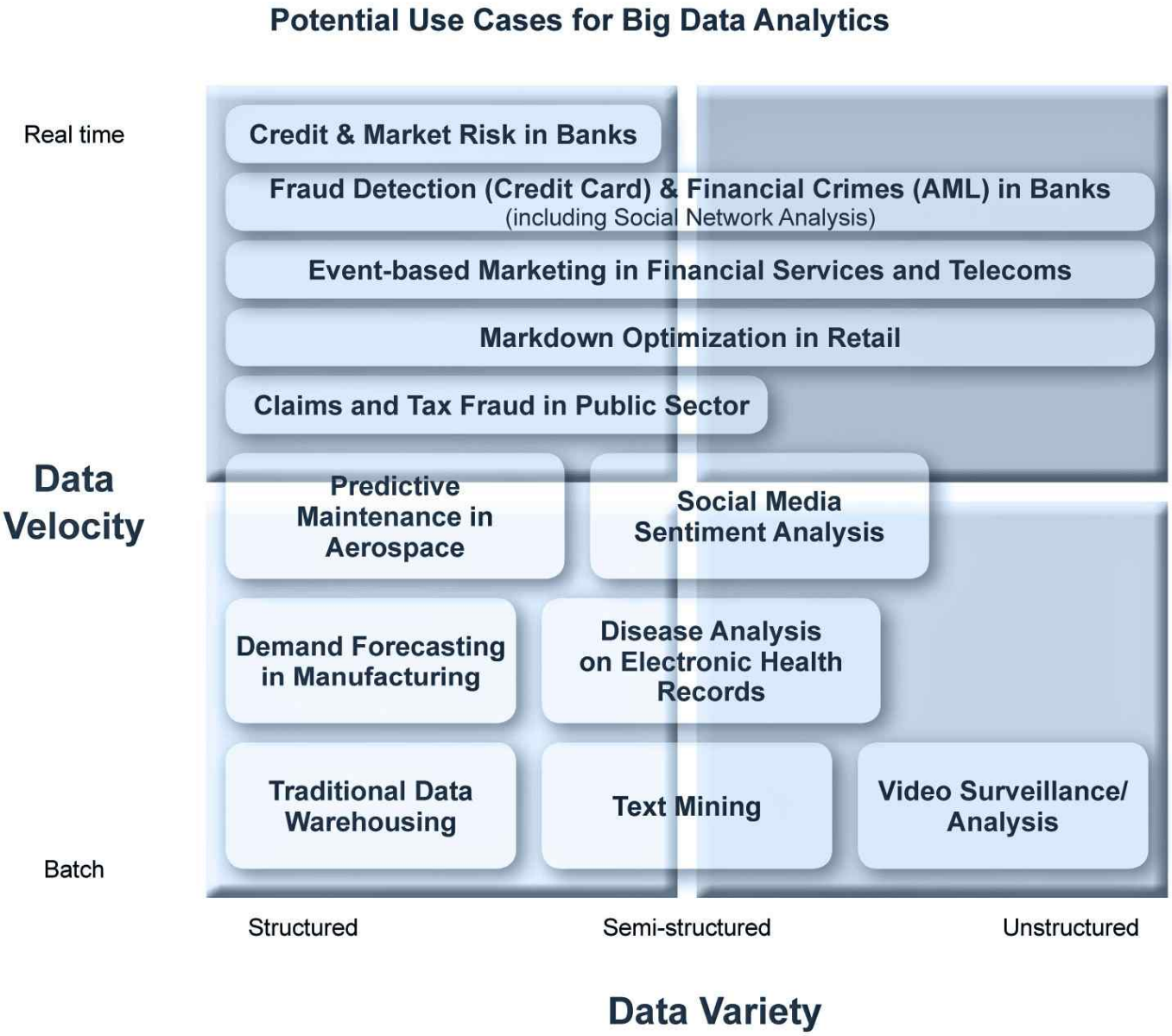
- 요즘 화두가 되는 고급분석(Advanced Analytics)은 통계분석, 예측(스코어) 모형, 시계열 분석과 최적화 등을 의미합니다

대부분의 기업들은 고급 분석이 가능한 분석 전문가들을 리포팅 작성에 활용하고 있습니다



출처 : Davenport and Harris, "Competing on Analytics", Harvard Business School Press (2007) 참조

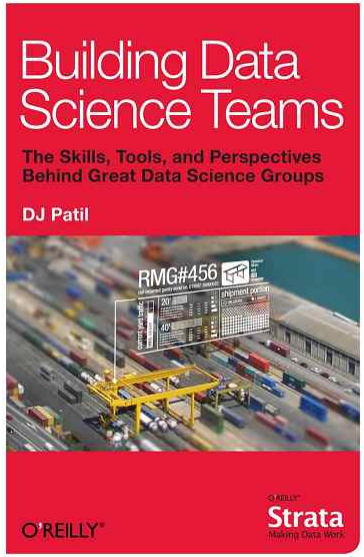
What is Big Data Analysis?



출처 : SAS and IDC

What makes a good data scientist?

- **Technical expertise**(기술적인 전문성): the best data scientists typically have deep expertise in some scientific discipline
- **Curiosity**(호기심): a desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.
- **Storytelling**(스토리텔링): the ability to use data to tell a story and to be able to communicate it effectively.
- **Cleverness**(영리함): the ability to look at a problem in different, creative ways.



** Patil, DJ (2011-09-22). Building Data Science Teams. O'Reilly Media*

빅 데이터 분석에서의 R의 문제점

- Single Core 연산
  - 멀티코어 CPU에서 1코어만 사용한다
  - R 2.14 부터 parallel 패키지 기본 탑재
- In-Memory 연산의 특징상 메모리 한계 문제
  - 모든 데이터를 메모리에 로딩 후 처리하는 작업 방식
  - 불필요한 데이터 저장으로 인한 메모리 부족 현상

➔ Open Source 에서의 대응방안

- Snow, multicore, parallel, bigmemory 등의 패키지들이 Multi-core 사용 및 논리적으로 메모리한계를 극복한 패키지들을 제공 하고 있음
- 하지만 위의 방법들 모두 로컬머신으로 데이터를 가져온다는 문제로 인하여 다른 방법의 해결책이 필요

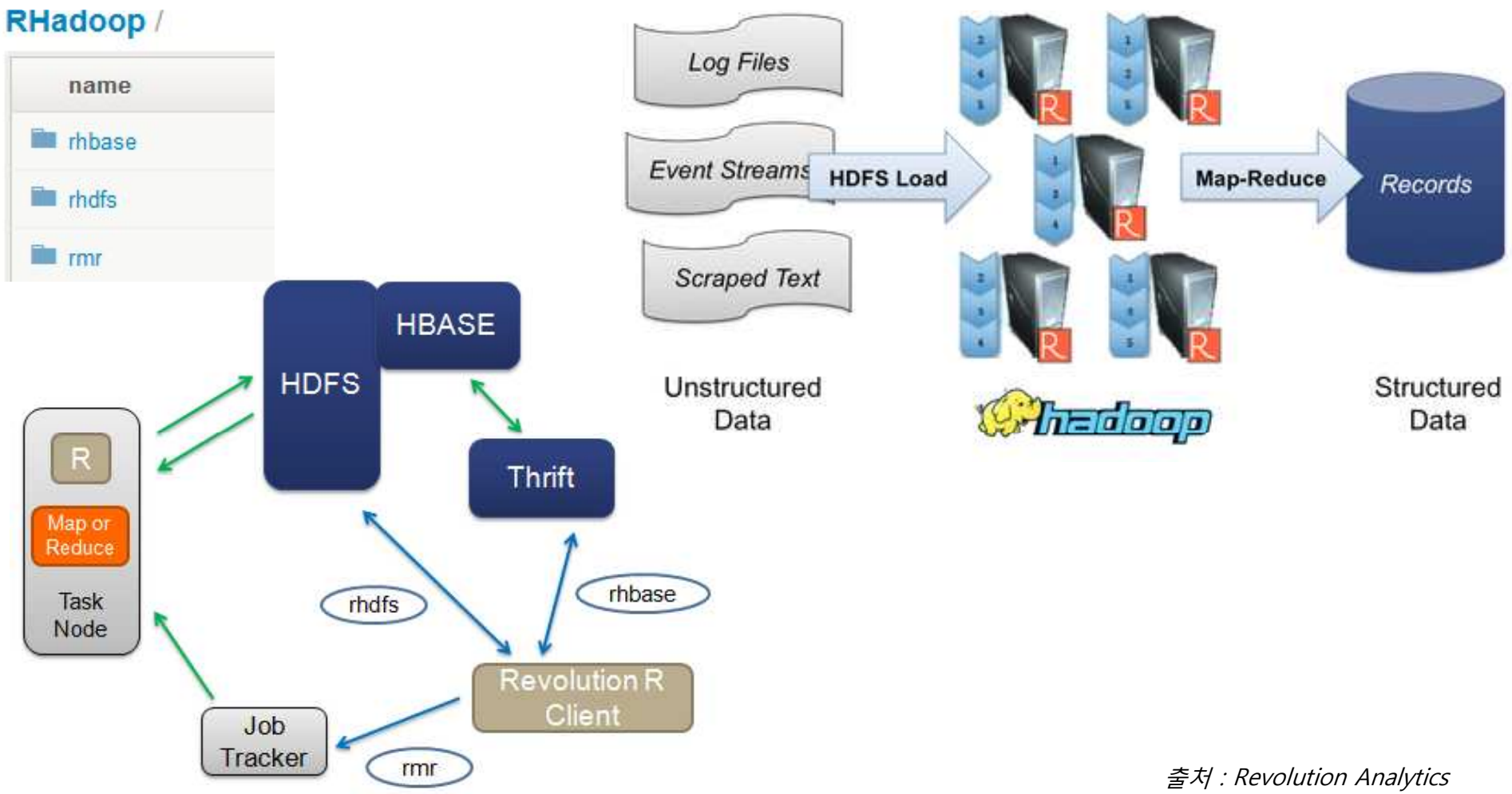
RHIPE (http://www.datadr.org/index.html)

- RHIPE(R and Hadoop Integrated Programming Environment)는 Purdue Univ.의 통계학 박사과정 학생이었던 Saptarshi Guha에 의해 개발된 R 패키지
- R을 Hadoop 환경에서 MapReduce 개념의 분산처리가 가능하게 해 줌
- 이후 Revolution Analytics 사에서 RHadoop 패키지를 공개함.
- Divide & Recombine (D&R) 기법, not parallel processing.

Facebook 에서 RHIPE 소개하는 Video (2010.03.09)
http://www.lecturemaker.com/2011/02/rhipe/

RHadoop (https://github.com/RevolutionAnalytics/RHadoop)

- **Hadoop 기반의 Map-Reduce 기능 구현을 R Code로 작성할 수 있어 쉽게 Hadoop 기반의 분석이 가능함**

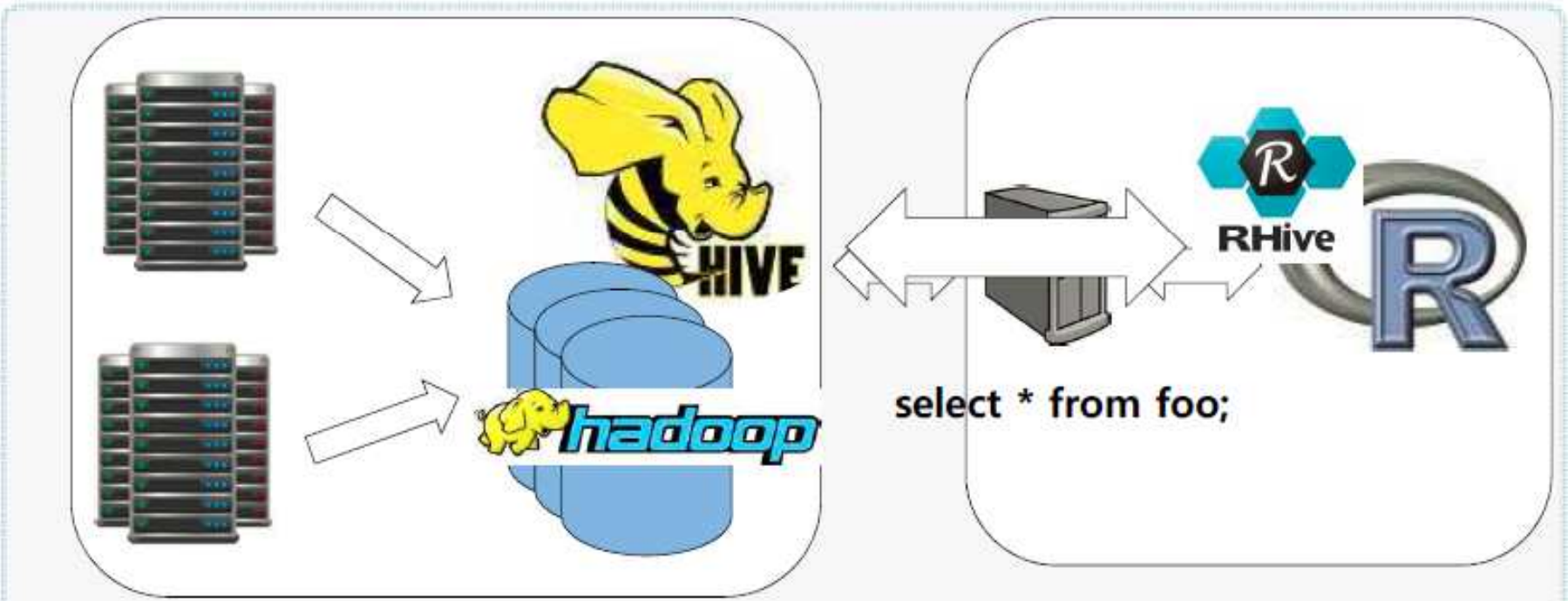


출처 : Revolution Analytics

RHadoop (Word Count Example)

출처 : 베가스 R 소개자료, 김준기

RHive – Motivation of Rhive (NexR)

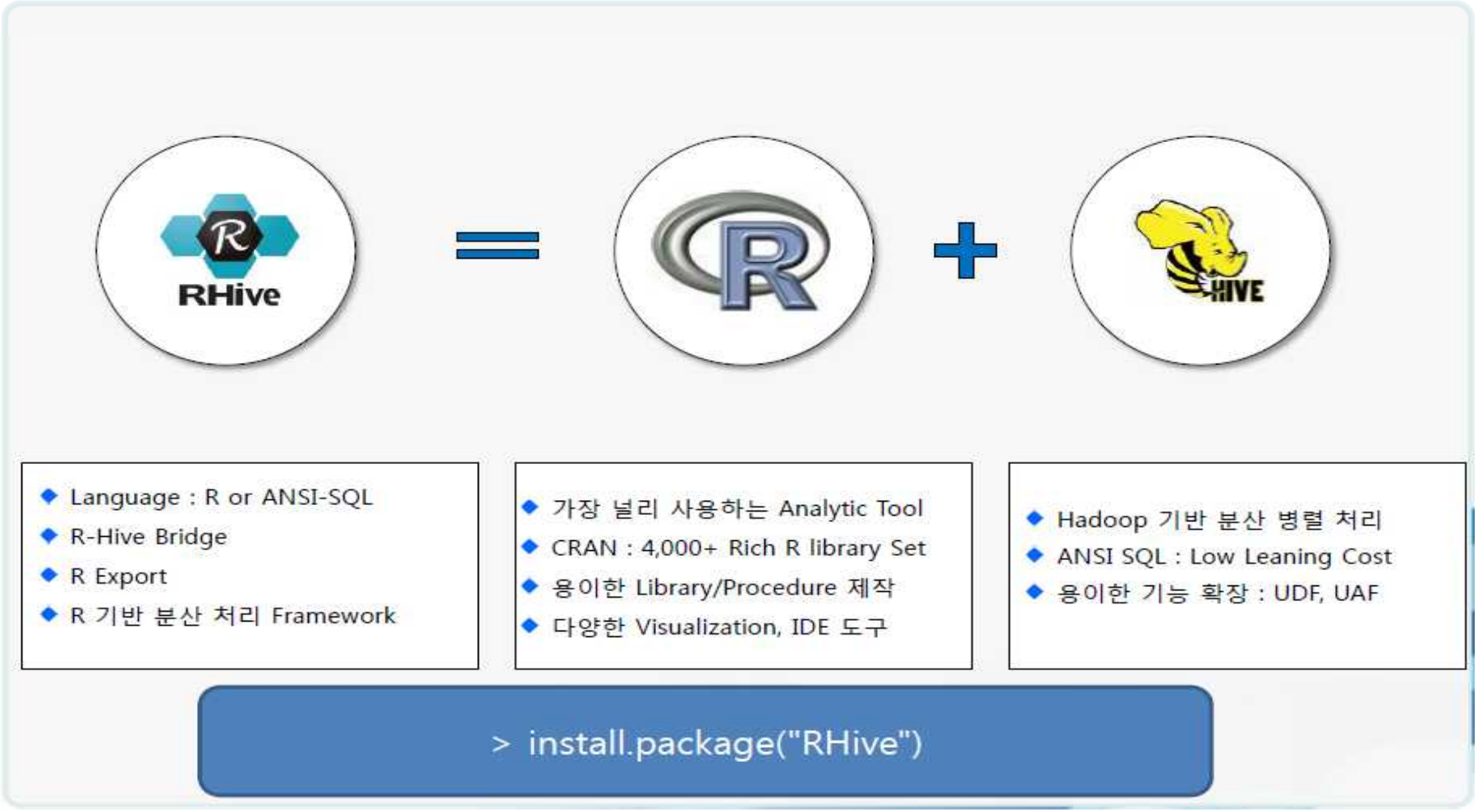


- Map/Reduce for data analysis?
  - 배워야 한다. 그러나 어렵다
- SQL for data analysis!
  - 대부분 배울 필요가 없다. 그리고 쉽다.

출처 : R and RHive in Data Scientists toolbox, 전희원(2011)

RHive (https://github.com/nexr/RHive)

- R과 Hive 기술을 접목하여, 작은 데이터는 R에서 바로처리하고 빅데이터는 Hive의 기술을 이용하여 Hadoop에서 처리가 가능



출처 : R and RHive in Data Scientists toolbox, 전희원(2011)

RHive Example

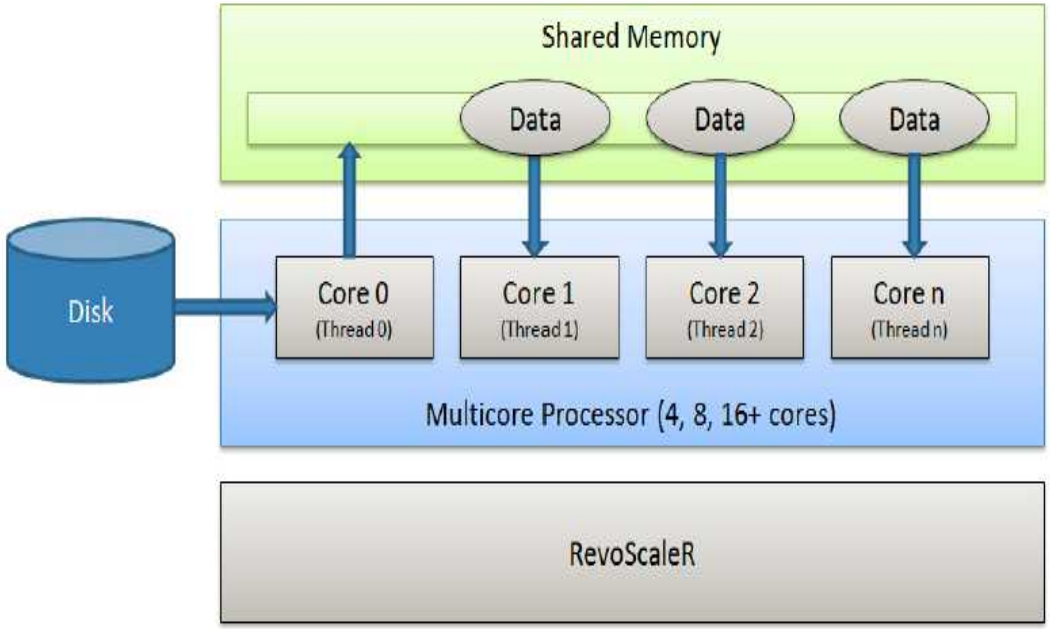
KT CDR analysis system project

- The contents of the project: Collect KT Wireless CDR(Call Detail Record) to NDAP, NexR Big Data Platform and perform call quality analysis and reporting
- Analysis using RHive
  - Call Anomaly Detection
  - Location-based call quality analysis
  - Location-based subscriber clustering
  - SNA Analysis
- Examples of location-based call quality analysis(Google Map Mash-up)

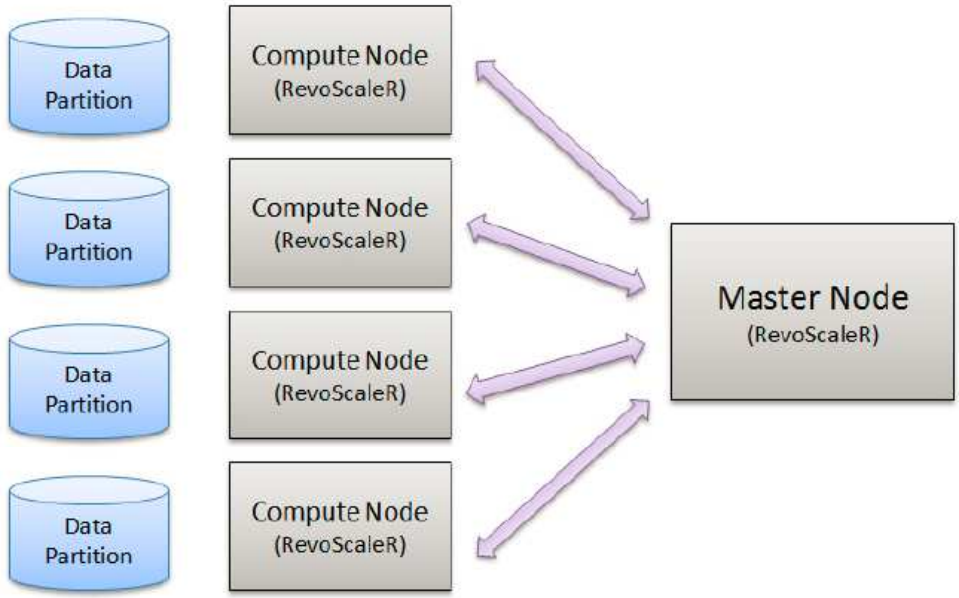
출처 : http://www.nexr.co.kr/nexrcorp_en/products_and_services/rhive_useCase.php

RevoScaleR

- RevoScaleR 을 통하여 Single Core 문제 및 메모리제한문제를 해결
- 또한 HPC Clusters/Cloud에 R 분산처리를 가능하도록 지원



RevoScaleR on Sing Computer



RevoScaleR on Multiple Computers

주) 현재는 HPC Clusters, IBM LSF Cluster에서 지원,
2012 하반기에 Linux Clusters도 지원예정

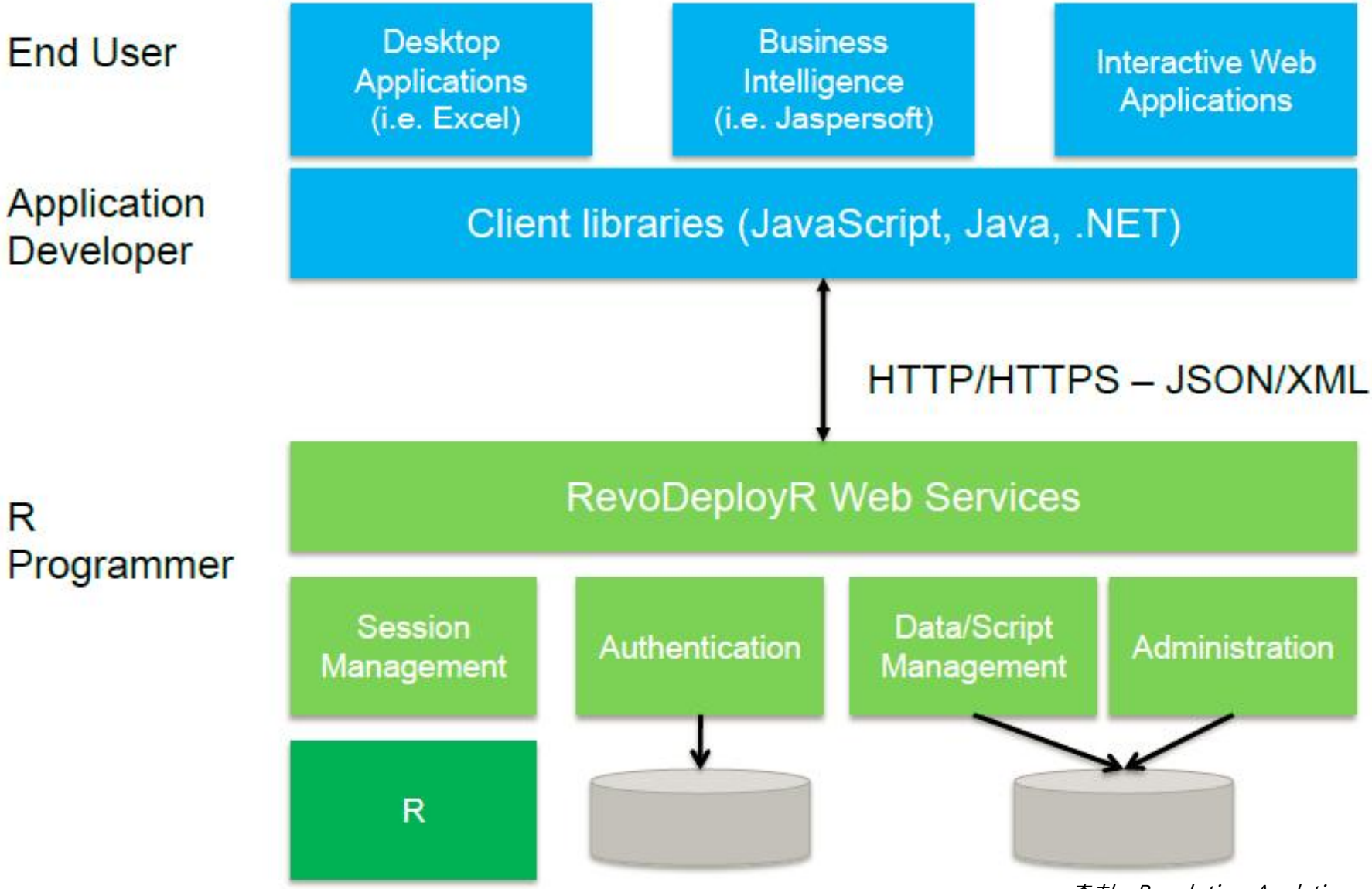
출처 : Revolution Analytics

High-Performance Computing
(http://cran.r-project.org/web/views/HighPerformanceComputing.html)

- **Parallel computing: Explicit parallelism** → Rmpi, snow, snowfall, foreach, doMPI
- **Parallel computing: Implicit parallelism** → fork, multicore
- **Parallel computing: Grid computing** → GridR, xgrid, biocep-distrib
- **Parallel computing: Hadoop** → RHIPE, Rhadoop, EMR(Elastic Map Reduce)
- **Parallel computing: Random numbers** → doRNG, rprng
- **Parallel computing: Resource managers and batch schedulers** → batch, BatchJobs
- **Parallel computing: Applications** → caret, maanova, tm, bcp
- **Parallel computing: GPUs** → gputools, cudaBayesreg, OpenCL, WideLM
- **Large memory and out-of-memory data** → bigmemory, biglm, ff, HaddpStreaming
- **Easier interfaces for Compiled code** → inline, Rcpp, rJava
- **Profiling tools** → profr, proftools

RevoDeployR (Architecture)

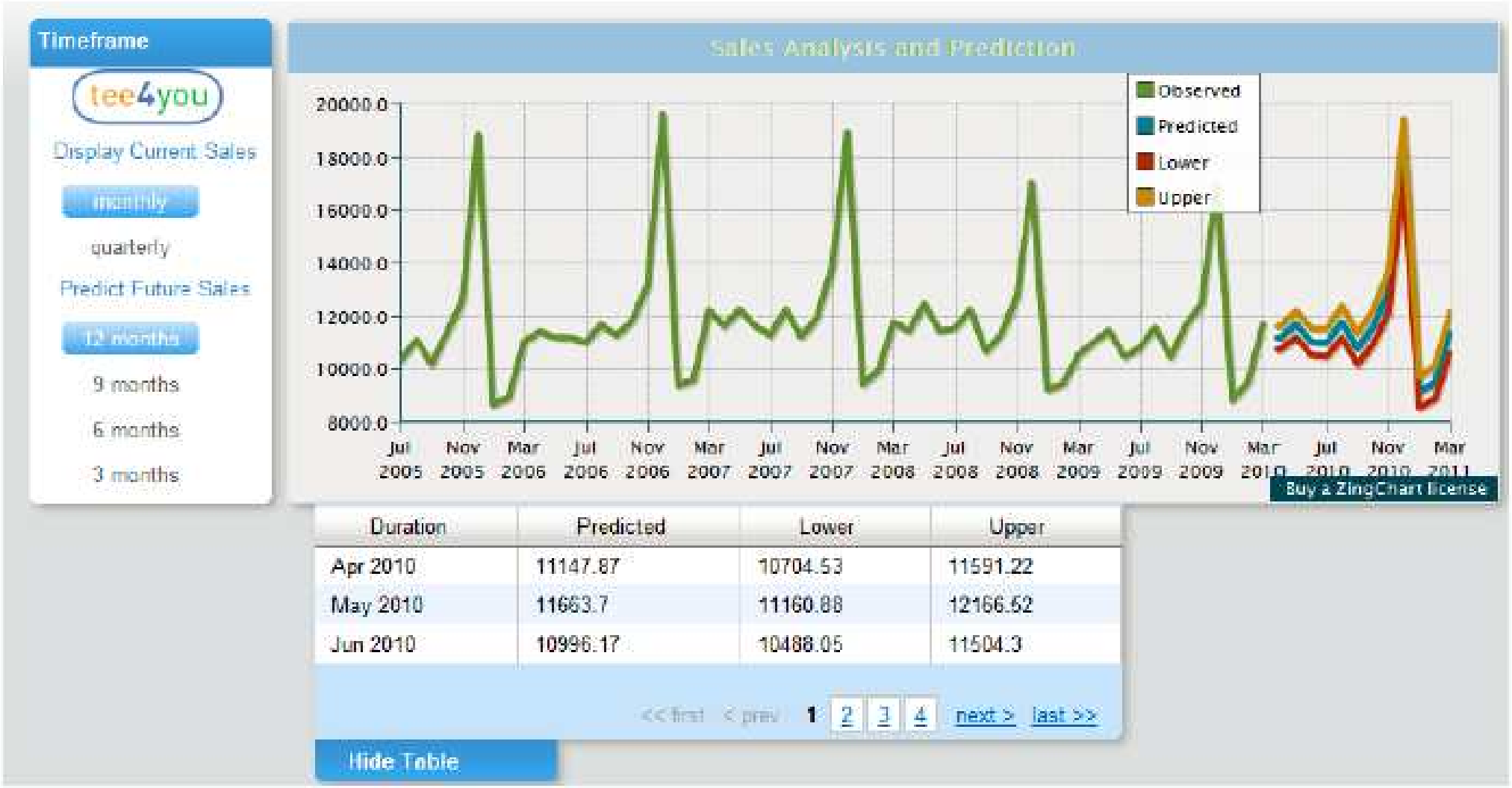
- **RevoDeployR 을 통하여 R 분석 환경을 웹 Application으로 구현가능**



출처 : Revolution Analytics

RevoDeployR (Sample App)

- **DeployR과 HTML5 plot library from JingCharts 를 이용한 웹 화면**



출처 : Revolution Analytics

고객의 니즈

- 빅데이터가 있는데 저장할 수 있는지?

  : 법적으로 보관해야 하는 데이터라서 무조건 일정기간 보관해야 함 그렇지만 상용 RDBMS (Scale-up)로는 비용이 너무 많이 들어가고, 한곳에서 다 처리를 못해서 분리해서 데이터를 저장함.

  ➔ 빅데이터 처리를 저비용으로 한곳에서 Scale-out 방식으로 할 수 있는지?

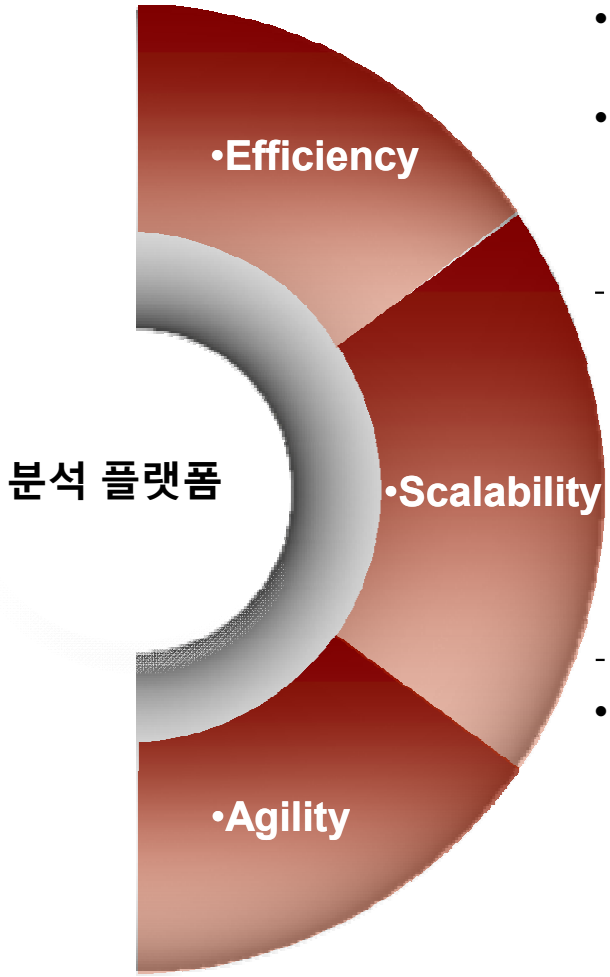
- 저비용으로 빅데이터 분석환경을 구축 할 수 있는지?

  : 빅데이터를 모우고 나면 분석을 해야하는데, 고성능의 SAS나 SPSS를 사용하려니 비용이 너무 비싸서 엄두를 못냄. 특히 해당 라이선스가 일회성이 아닌 매년 유지료를 내야하므로 비용감당이 어려움.

  ➔저비용으로 빅데이터 분석을 할 수 있는 기업용 고급분석 환경을 구축해 줄 수 있는지?

LCBEx (Low Cost But Excellent) 분석 플랫폼

- **바람직한 분석시스템의 구축은 분석엔진을 중심으로 마련된 저비용(Low Cost)이지만 고성능이며 확장성이나 인터페이스가 뛰어난 (Excellent) Analytic Platform(분석 플랫폼)을 중심으로 이루어져야 함**



• 저비용
  - 오픈 소스 소프트웨어 기반으로 구축해 최대한 도입비용을 낮춰야 함
• 고성능
  - 구현 사상을 고려하였을 때, 빠른 계산처리 및 새로운 알고리즘, 방법론이 제공되어야 함

• 확장 및 통합 용이성
  - 독립된 형태의 분석 시스템 구축 없이 분산 처리를 통한 처리가 가능하여야 함
  - Hadoop과 같은 오픈소스 기반의 솔루션을 활용할 수 있음

• 구현 신속성
  - 분석 방법이나 결과 등을 오브젝트로 관리하여 공유, 재활용이 가능하여야 함
  - 정형화된 분석 프로세스의 패키징이 용이하여 이관이나 재활용이 용이하여야 함

출처 : 베가스 R 소개자료, 김준기

결론

- 빅데이터는 있다. 하지만…
- 어떤 가치를 찾아야 하는 건가?

다양한 시도를 해 볼 수 있는 환경(비용 저렴, 기술 공개)이 만들어지고 있다.
➔ 결국 가치를 찾고자 하는 사람이 Data Scientist 가 되려고 노력(기술의 내재화)을 해야한다.

Slow and Steady Wins the Race.

Study, study, study, study……study.

감사합니다.

이동우

지티원
DG서비스사업부
Tel : 010-4801-6609
email : leewow@gtone.co.kr