

R, 그리고 빅데이터

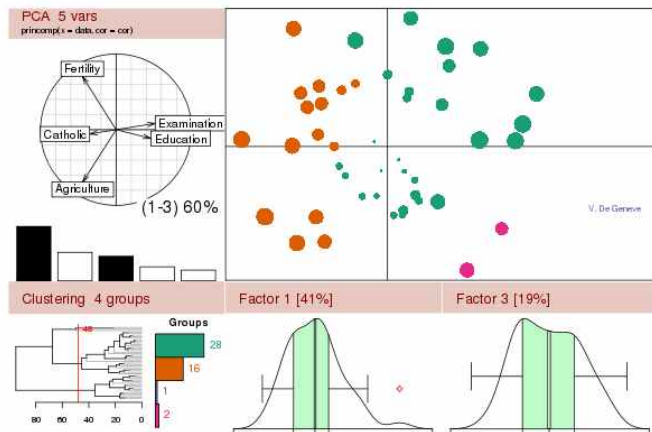


제30회 Open Technet

-빅데이터 오픈소스 플랫폼 기술세미나

2012.07.26

이동우



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

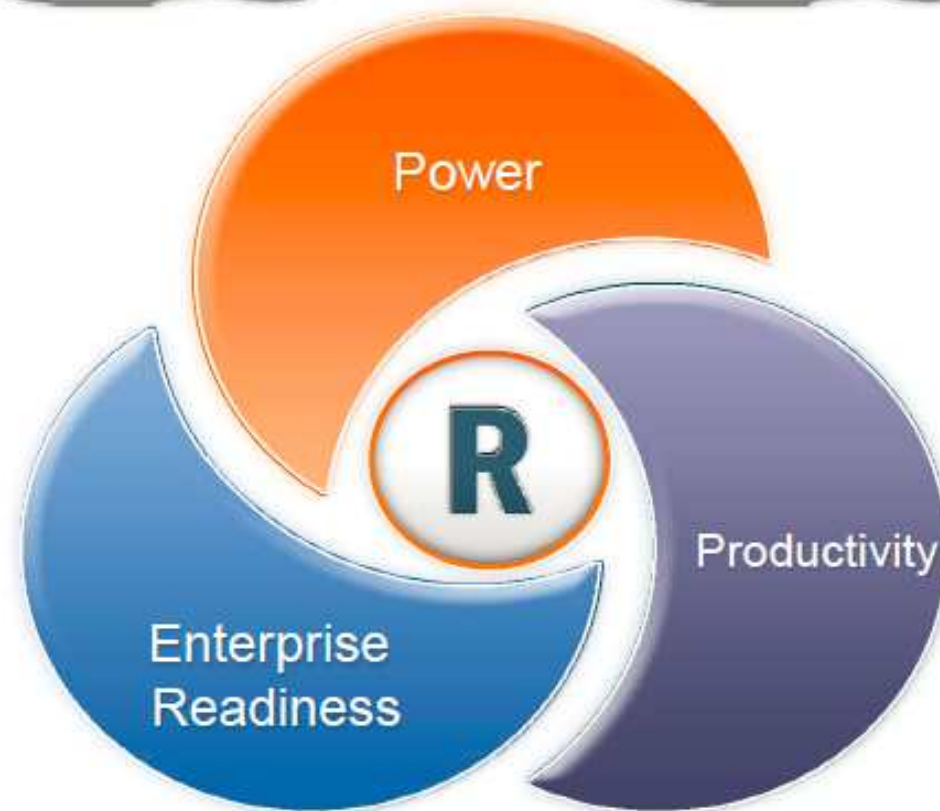
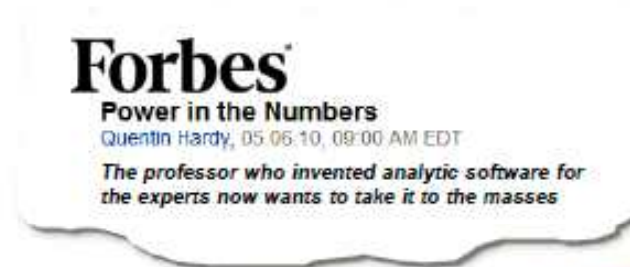
R : Open source analytics for the enterprise

→ Most advanced statistical analysis software available

→ Half the cost of commercial alternatives

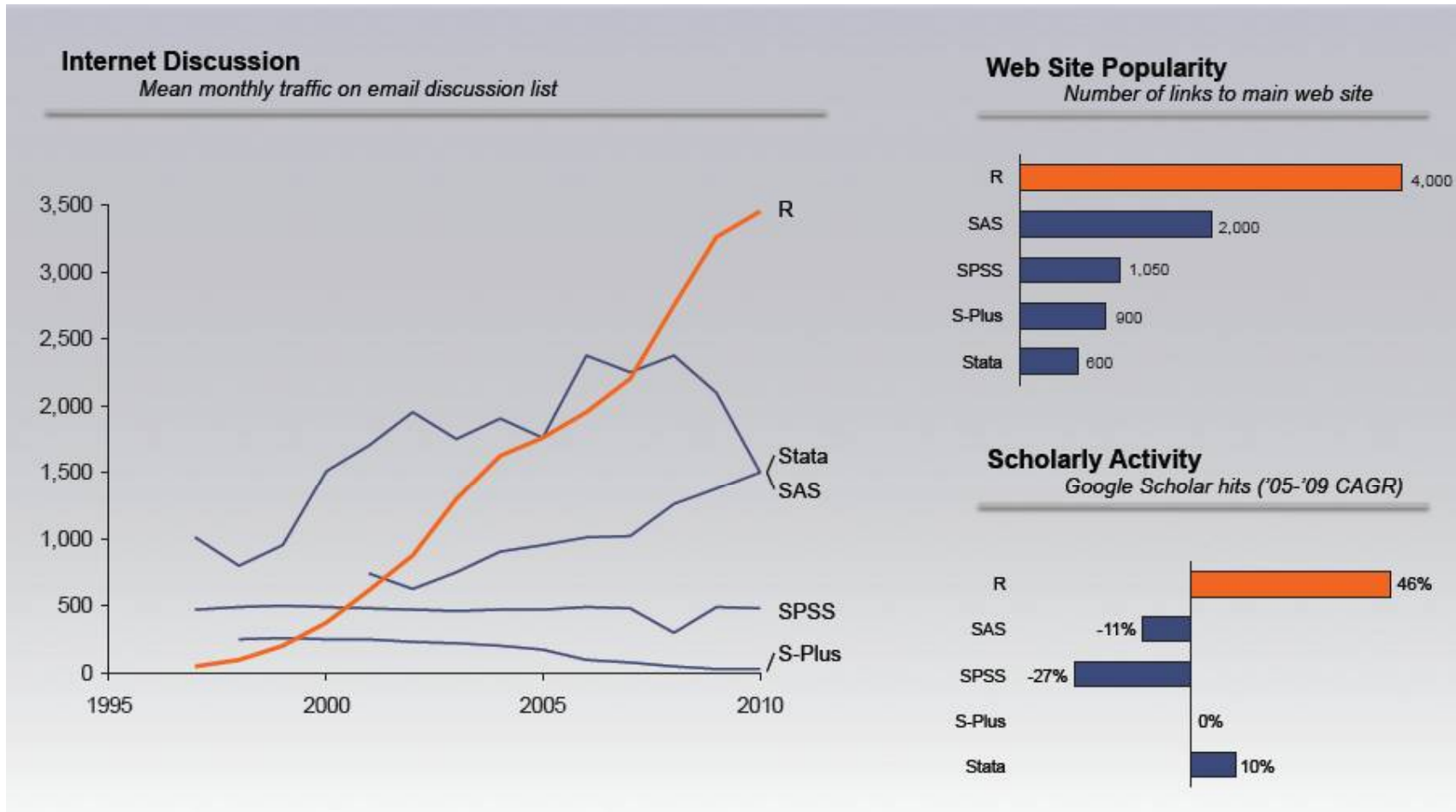
→ 2M+ Users

→ 2,500+ Applications



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, 우수기업에서 데이터 분석 tool로 사용 중임



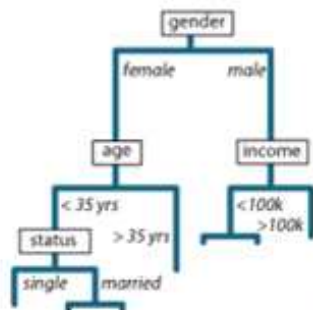
기업체에서의 R의 활용

- 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

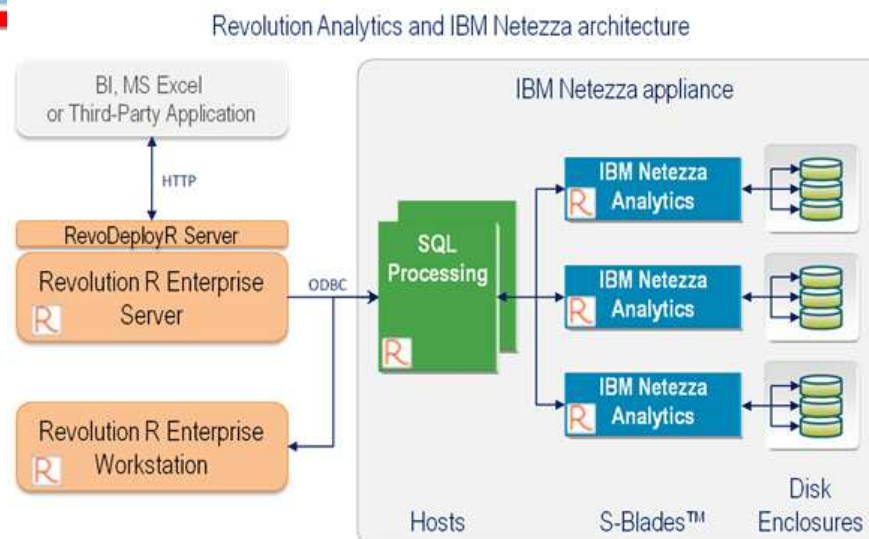
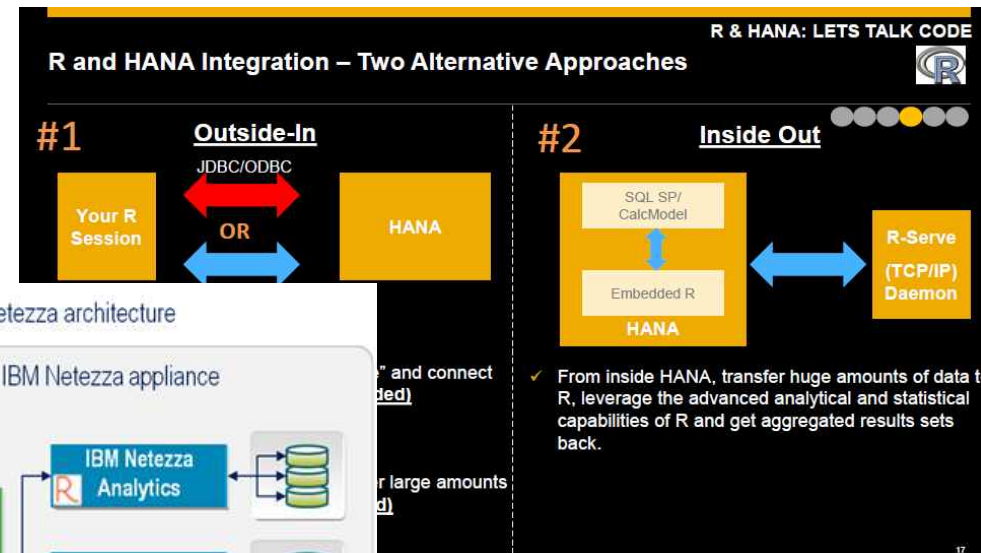
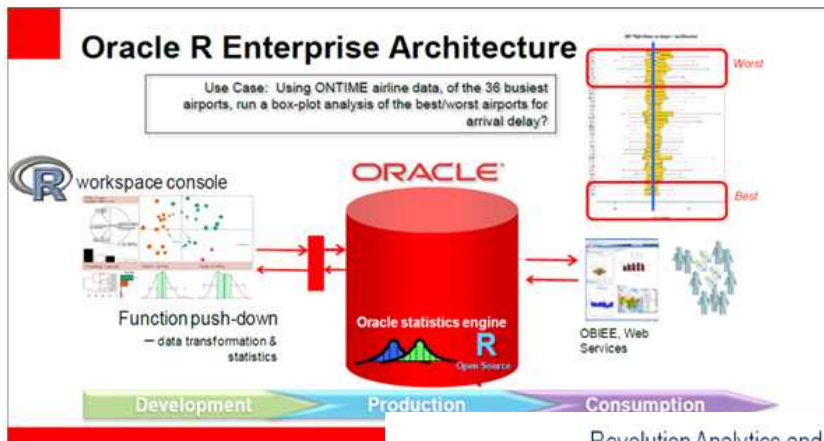
Itamar conveyed how Facebook's Data Team used R in 2007 to answer two questions about new users: (i) which data points predict whether a user will stay? and (ii) if they stay, which data points predict how active they'll be after three months?

- Itamar Rosenn, Facebook

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

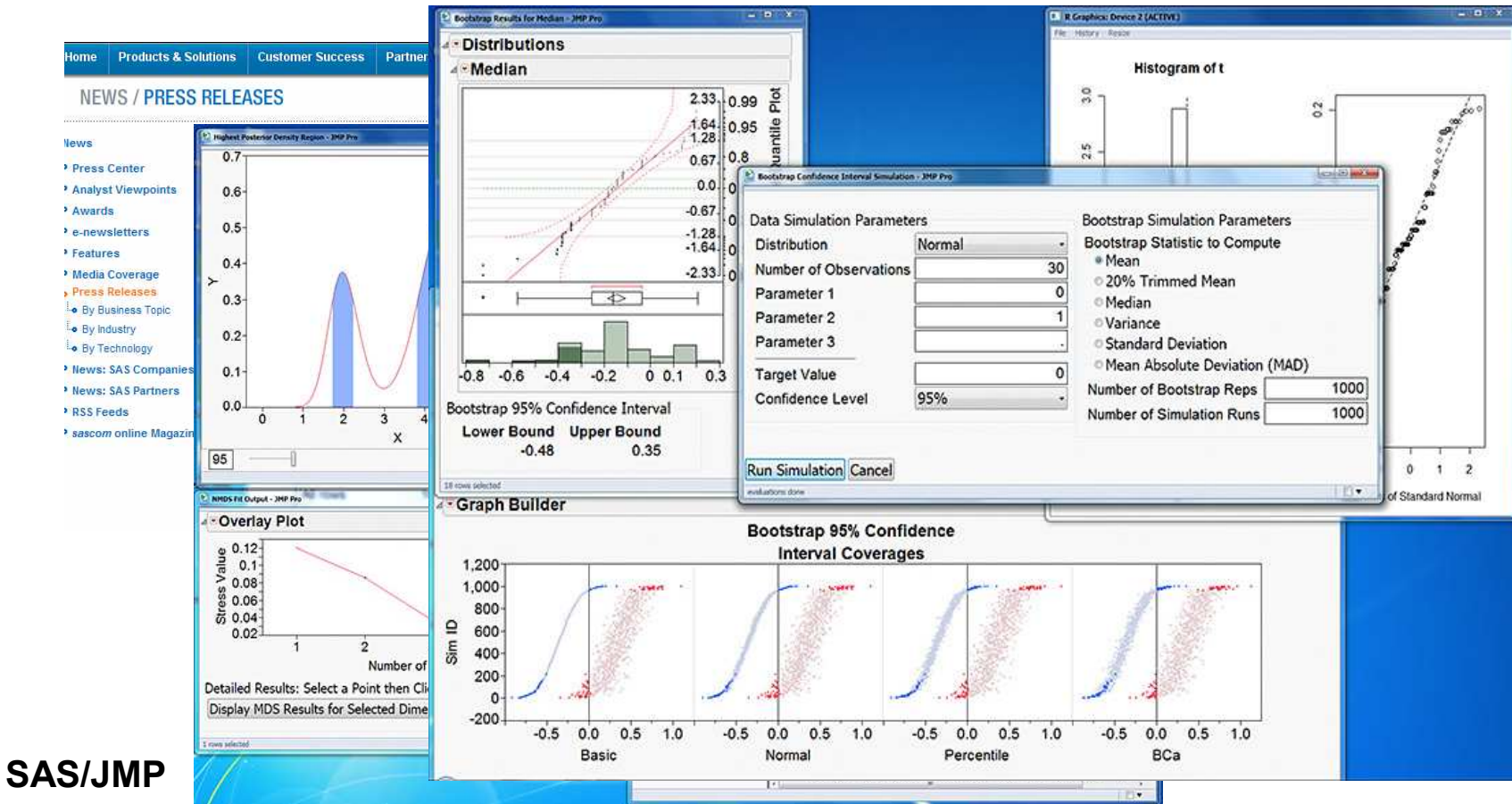
소프트웨어 Vendor의 R 적용

- 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,..."

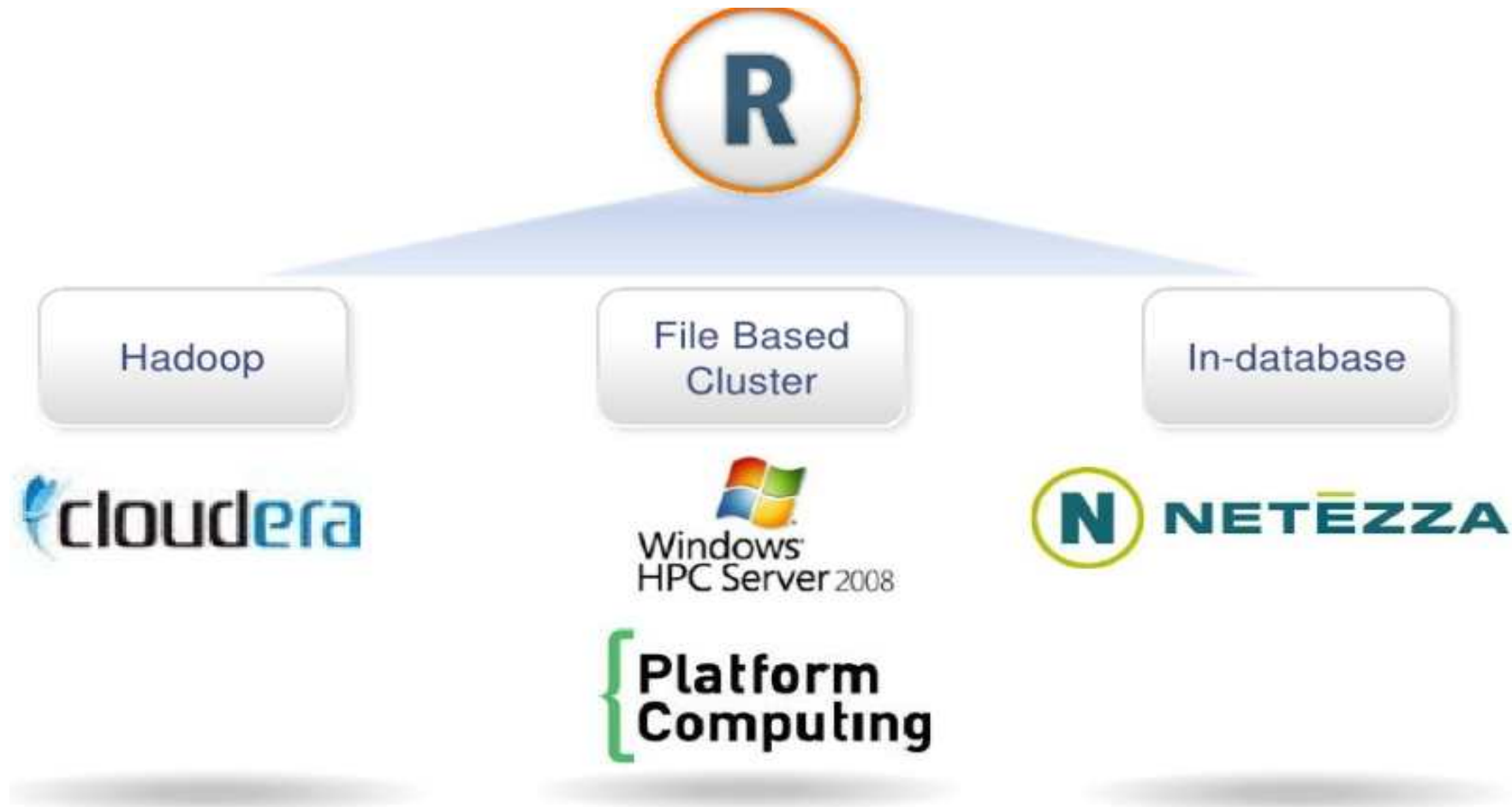
새로운 연구의 경우 Tobit 회귀를 써야 했어요... 그러나 SPSS에서 분석 기능이 지원이 되지 않아, 다른 통계 패키지를 배우고, 습득하는데, 고생이 많았습니다. SPSS쓰다 다른 패키지 갔다가... 헛갈리기도 하고...

그러나 이번 확장 모듈로 간단히 해결이 되었습니다!!!"

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

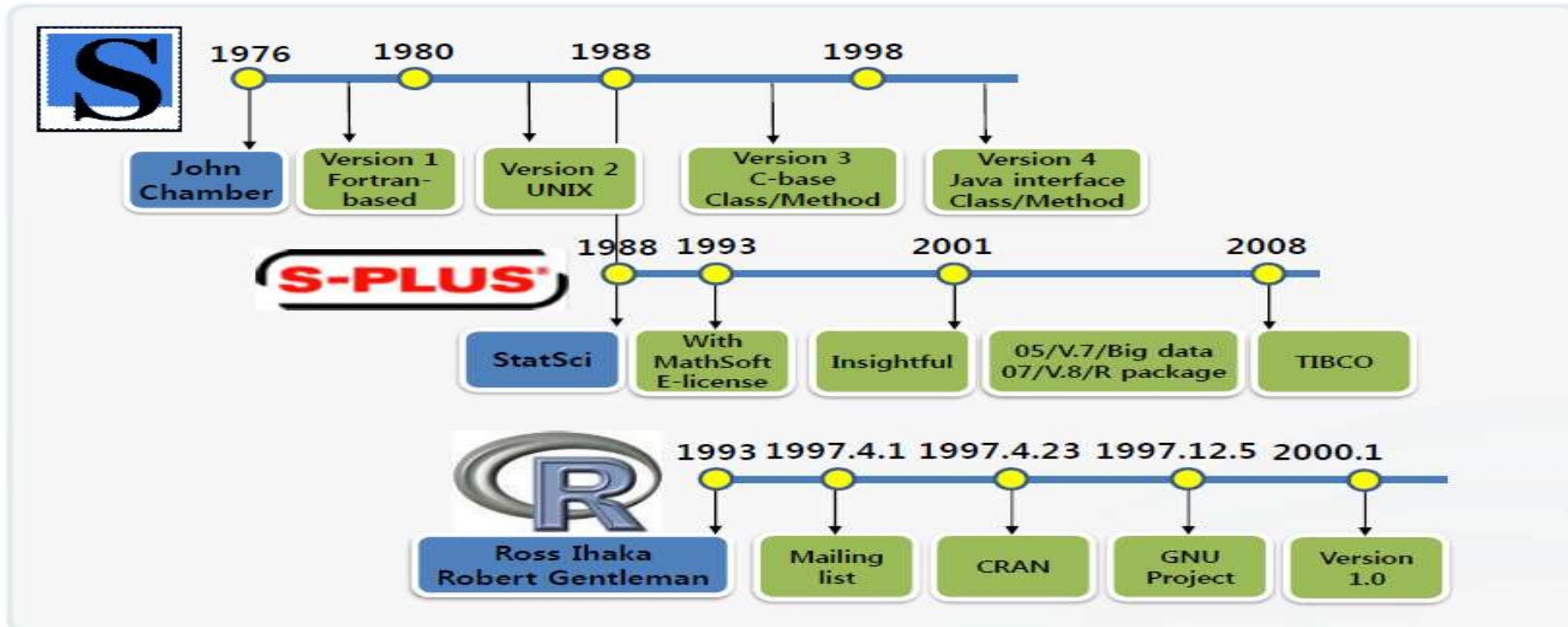


출처 : Revolution Analytics

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 인사이트 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>)

The R Project for Statistical Computing

PCA 5 vars
[Pie chart showing categories: Faculty, Education, Economics, Agriculture, and others]

Clustering 4 groups
[Dendrogram showing hierarchical clustering]

Factor 1 (41%)
Factor 3 (19%)
[Histograms showing distributions for factors]

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).
- 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.

News:

- **R version 2.15.0** (Easter Beagle) has been released on 2012-03-30.
- **R version 2.14.2** (Gift-Getting Season) has been released on 2012-02-29.
- **The R Journal Vol 3/2** is available.
- [useR! 2012](#), will take place at Vanderbilt University, Nashville Tennessee, USA, June 12-15, 2012.

This server is hosted by the [Institute for Statistics and Mathematics](#) of the [WU Wien](#).

- 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/>

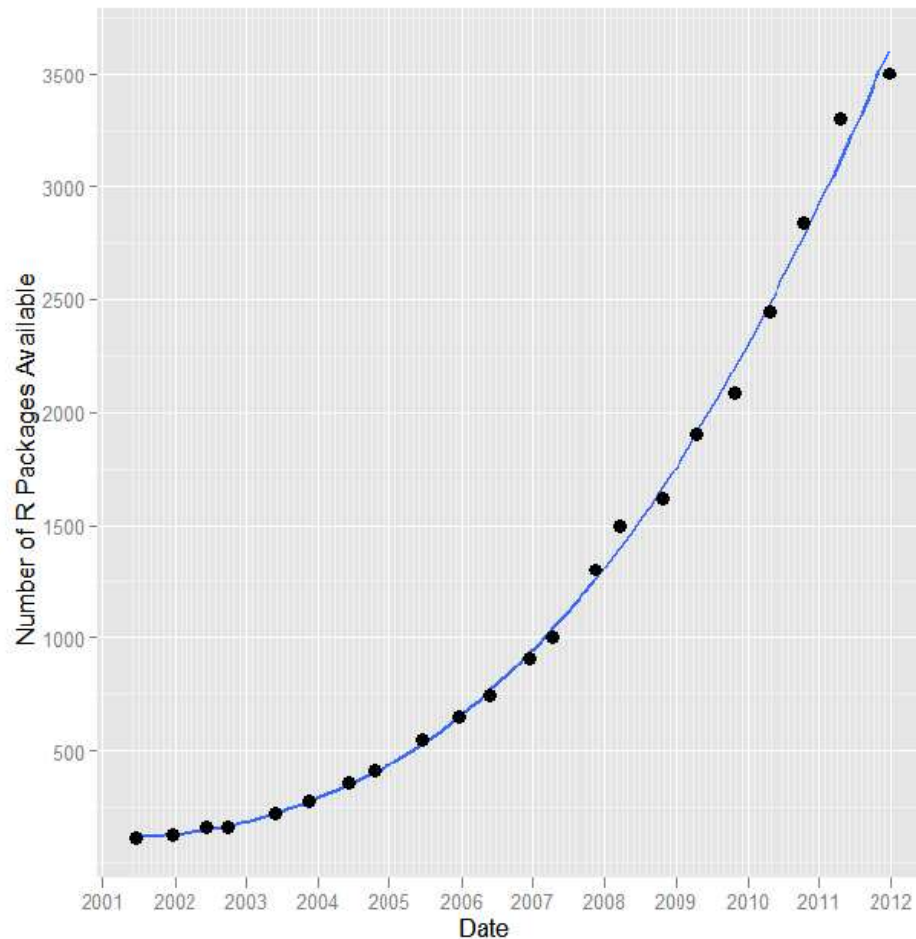
Korea :

<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 [\(http://cran.r-project.org/web/views/\)](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>)

Welcome

to crantastic, a community site for R packages where you can search for

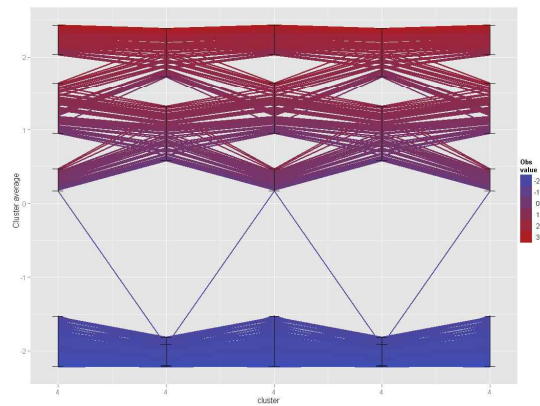
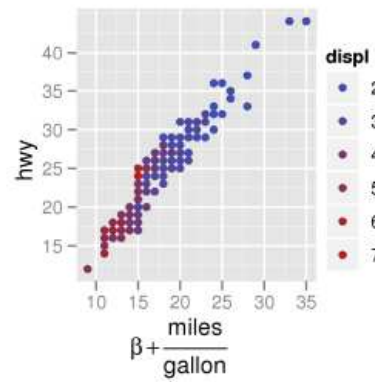
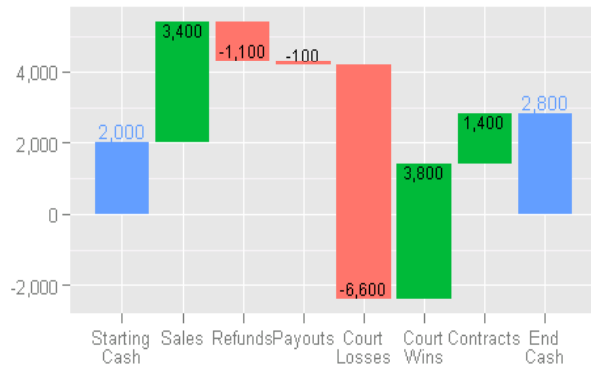
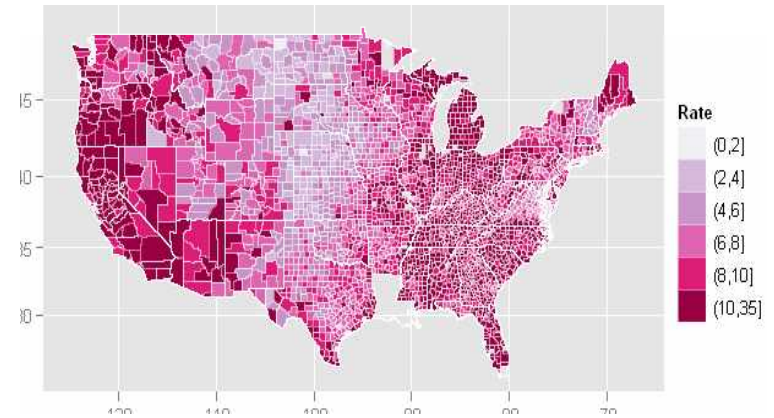
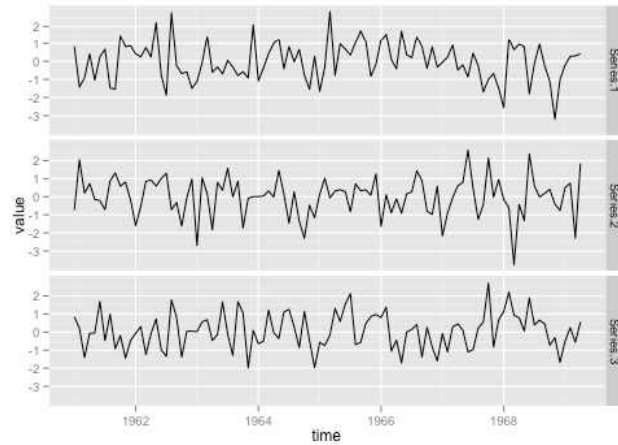
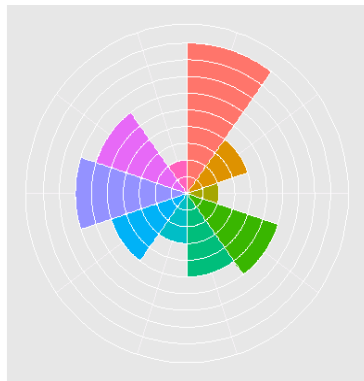
crantastic!

>4000 Packages

- [mfr](#) was **upgraded** to version [1.03](#) (about 4 hours ago)
- [matrixStats](#) was **upgraded** to version [0.5.0](#) (about 4 hours ago)
- [imputeYn](#) was **released** (about 4 hours ago)
- [DierckxSpline](#) was **upgraded** to version [1.1-5](#) (about 4 hours ago)
- [mmm2](#) was **released** (about 18 hours ago)
- [RMark](#) was **upgraded** to version [2.1.1](#) (about 23 hours ago)
- [ISBF](#) was **released** (about 23 hours ago)

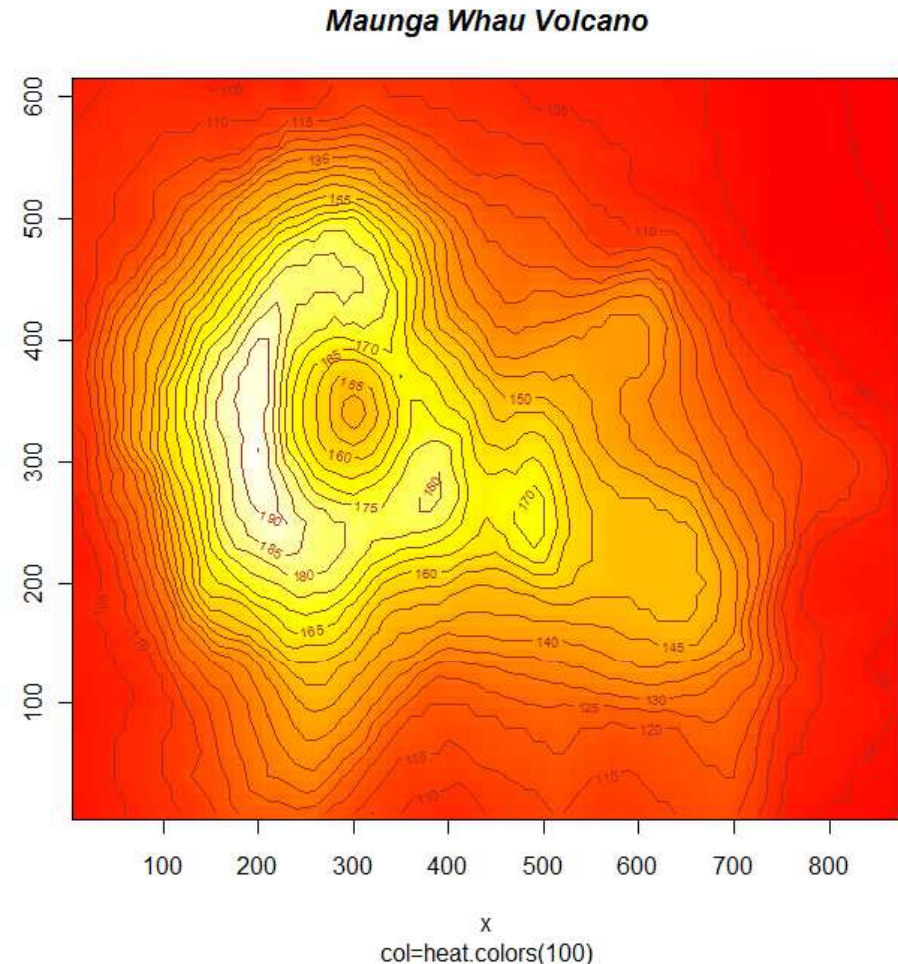
ggplot2 패키지

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



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)
```



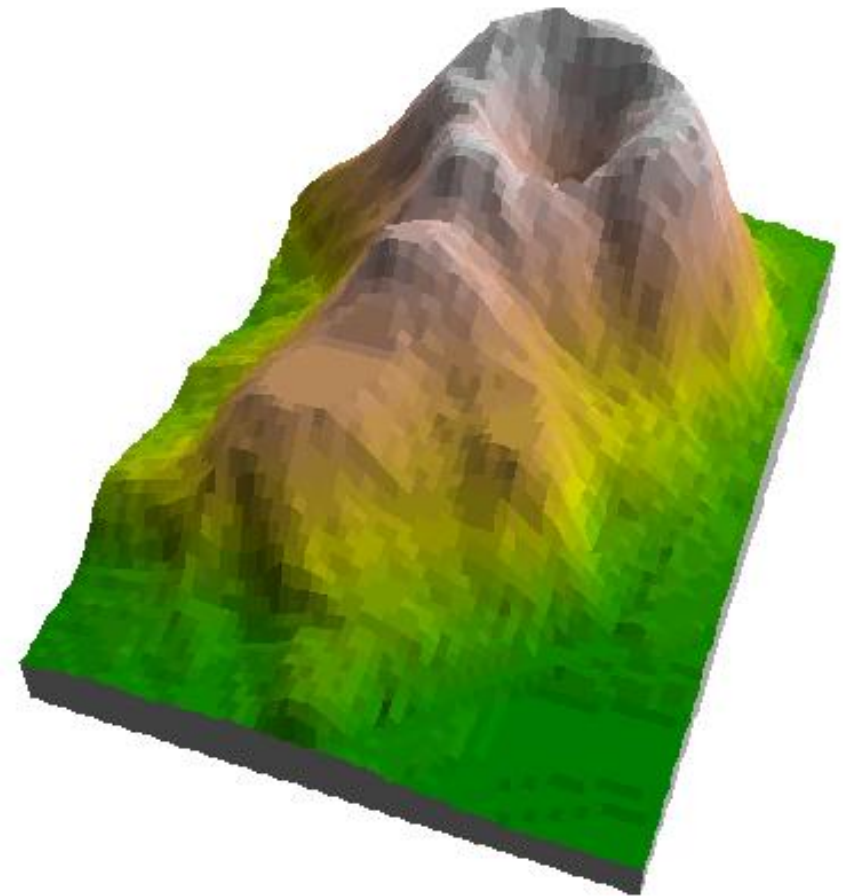
3-D contour map of Maunga Whau volcano

- 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)

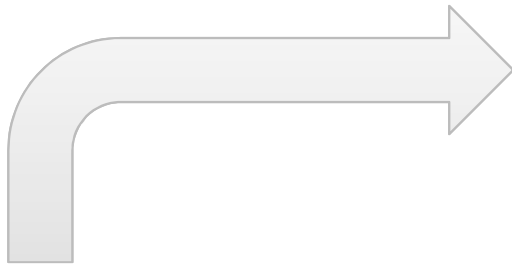
```



googleVis 패키지

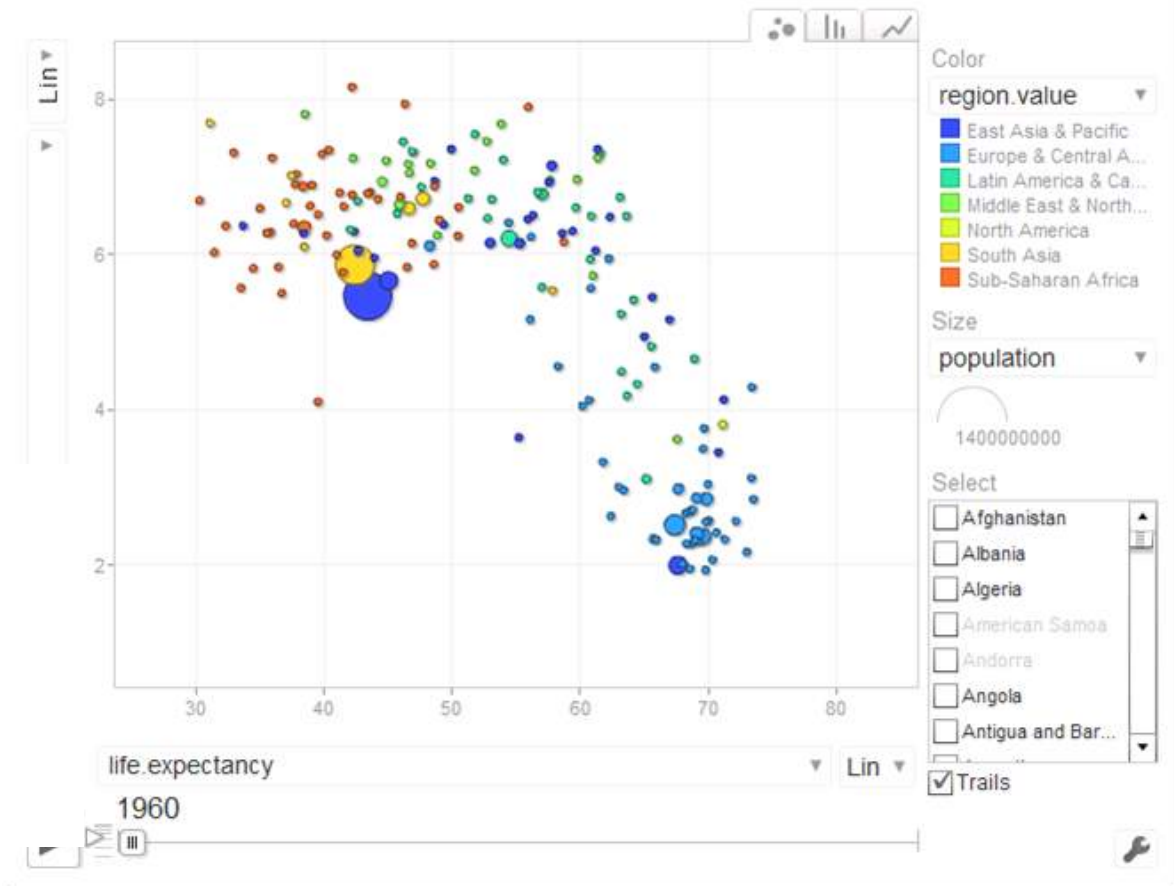
- R에서 구글 데이터 시각화 API를 이용

MotionChart 기술은 Gap Miner를 googleVis 에 붙여서 R에서 바로 사용이 가능



Examples

Here is an example of a Motion Chart ([user guide](#)) using data from the [World Bank](#):



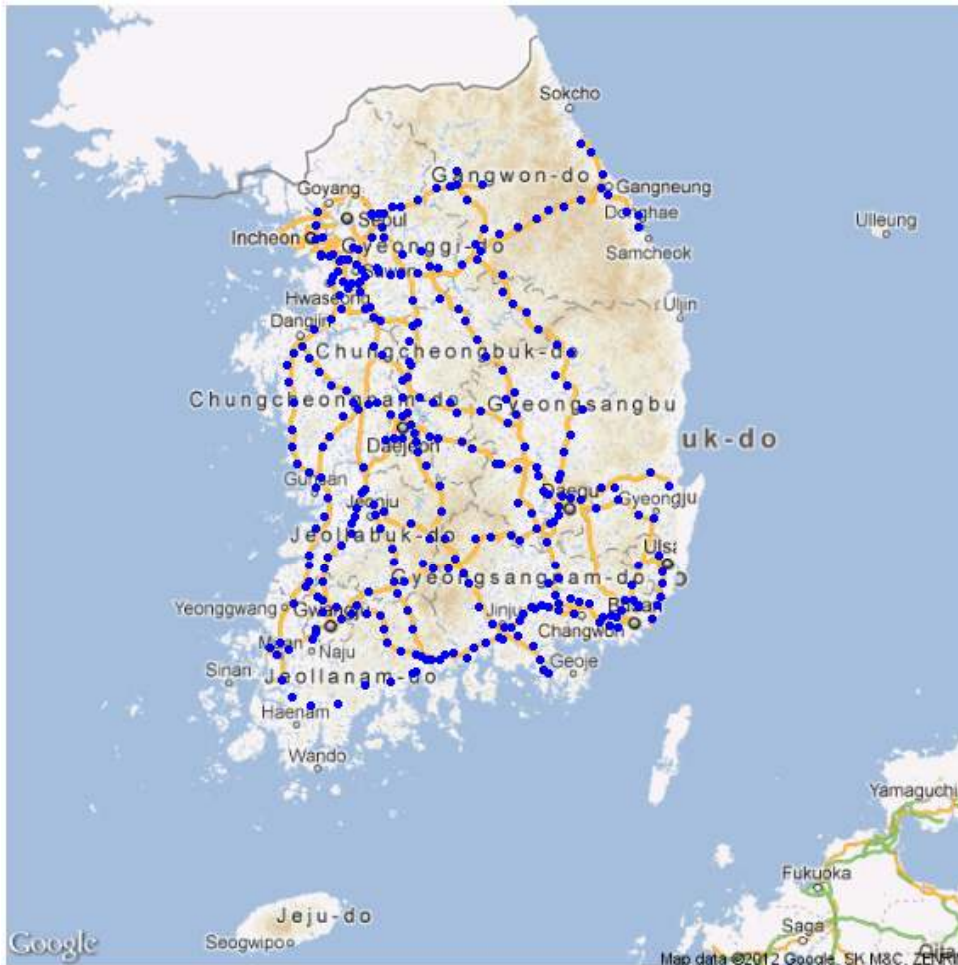
Hans Rosling: No more boring data: TEDTalks

<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)

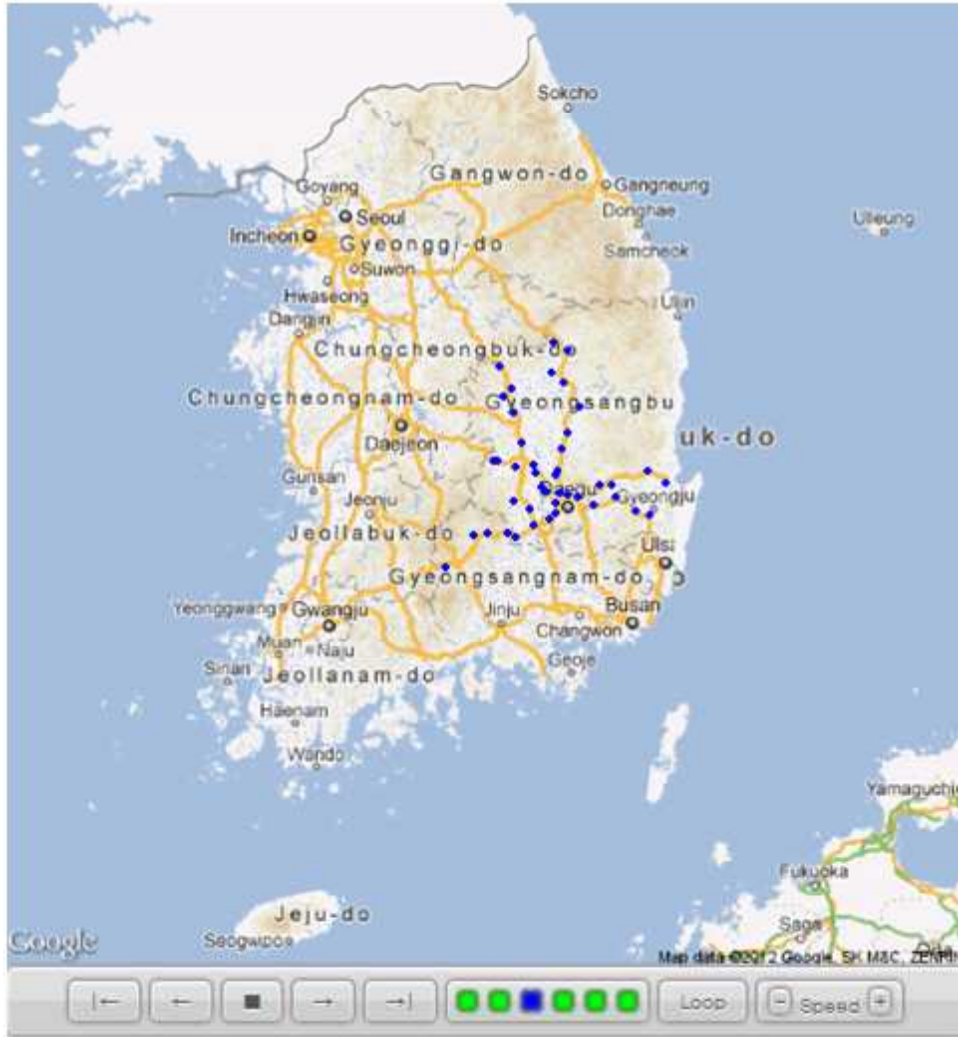
# 지도 레벨
input.zoom <- 7
map_data <- tollgate_info

# 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



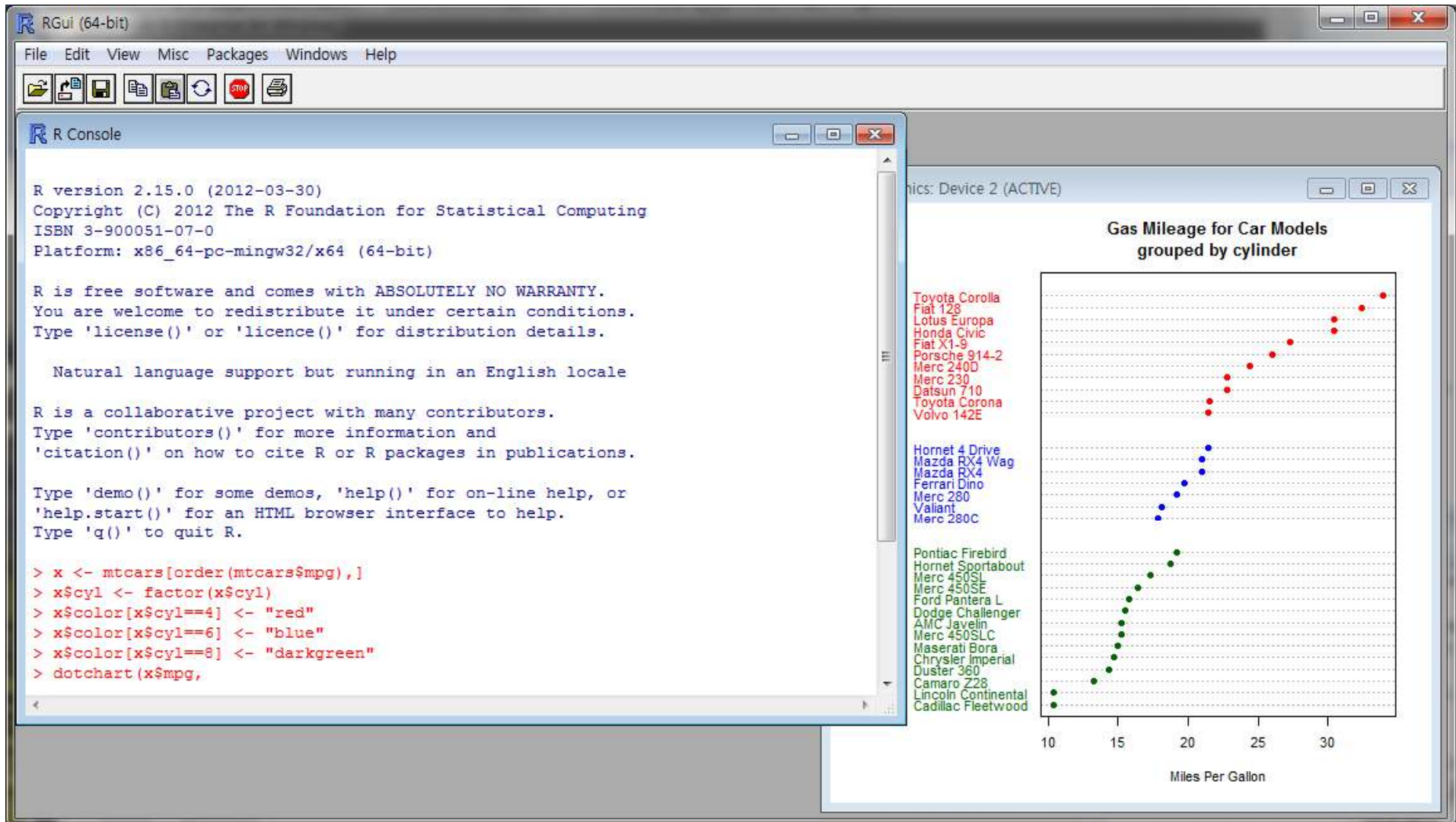
animation으로 생성

```
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
영역소",
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 [\(http://www.rstudio.org/\)](http://www.rstudio.org/)

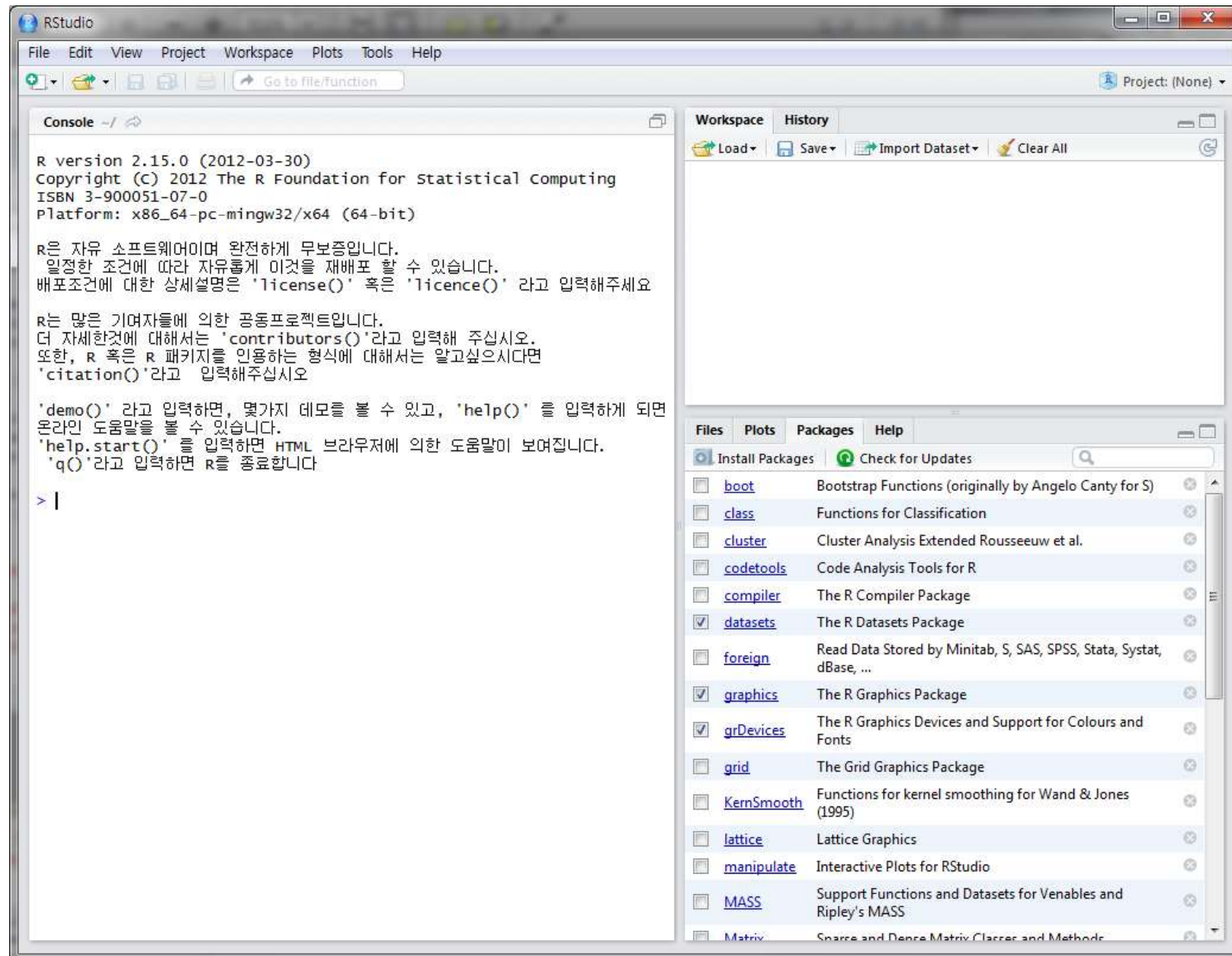
The image shows the RStudio website in a browser window. The website has a navigation menu with links for Home, Screenshots, Download, Docs, Support, Development, and Blog. Below the navigation is the R logo and the heading "Welcome to RStudio". A paragraph describes RStudio as a free and open source integrated development environment (IDE) for R, available for Windows, Mac, or Linux, or even over the web using RStudio Server. A prominent blue button with a download icon says "Download RStudio for Windows, Mac or Linux".

To the right of the main text is a "Screencast" button with a play icon and the text "RStudio in 2 minutes". Below this are four small thumbnail images showing different views of the RStudio interface.

The main focus is a large window showing the RStudio IDE interface. The window title is "RStudio" and it has a menu bar with File, Edit, View, Project, Workspace, Plots, Tools, and Help. The interface is divided into several panes:

- Source Editor:** Contains R code for analyzing diamonds. The code includes loading ggplot2, summarizing the diamonds data, calculating average carat size, and creating a faceted scatter plot of Price vs. Carat, colored by Clarity.
- Console:** Shows the output of the R code, including summary statistics for the diamonds data and the execution of the plotting commands.
- Workspace:** Shows the loaded data object "diamonds" with 53940 observations and 10 variables. It also lists the loaded packages: ggplot2 and format.plot.
- Plots:** Displays a faceted scatter plot titled "Diamond Pricing". The y-axis is "Price" (ranging from 0 to 15000) and the x-axis is "Carat" (ranging from 1 to 3). The plot is faceted by "Clarity" into 16 panels, each showing a different clarity level. The legend indicates the clarity levels: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

R Studio



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

The screenshot shows a web browser window displaying the Red-R website. The browser's address bar shows the URL <http://www.red-r.org/>. The website has a red header with the Red-R logo and the text "visual programming for R". A navigation menu includes links for Home, Downloads, Packages, Documentation, Development, Forums, Blog, Contacts, and Journal. A search bar is located in the top right corner. The main content area features a submission notice from 'anupparikh' dated 10/10/2010. Below this is a section titled "Red-R: A open source visual programming GUI interface for R" with a descriptive paragraph. Further down is a "Red-R Journal" section. A large graphic titled "Interactivity" shows several overlapping windows from the Red-R interface, including a scatter plot and a data table. To the right of the main content are three sidebar widgets: "User login" with fields for Username and Password, a "Log in" button, and links for "Create new account" and "Request new password"; "Red-R Newsletter" with an email input field, "Subscribe" and "Unsubscribe" radio buttons, and a "Save" button; and "Follow Us" with a "Follow @redRproject" button and a link to "Kyle's blog".

Red-R

The screenshot shows the Red-R IDE interface. On the left is a 'Widgets' panel with categories like 'Data Input', 'View Data', and 'R'. The main canvas contains a workflow with the following widgets: 'Read Files', 'Grouping clusters', 'XY Plot', 'R Datasets', 'Linear Model', 'View Data Table', and two 'Summary' widgets. Green arrows indicate data flow between these widgets. An output window titled 'R Summary (2)' is open, showing the following R output:

```

Call:
lm(formula = y ~ x, data = dataframe_org_8_1318207481.0, subset =
  method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
  contrasts = NULL, singular.ok = TRUE)

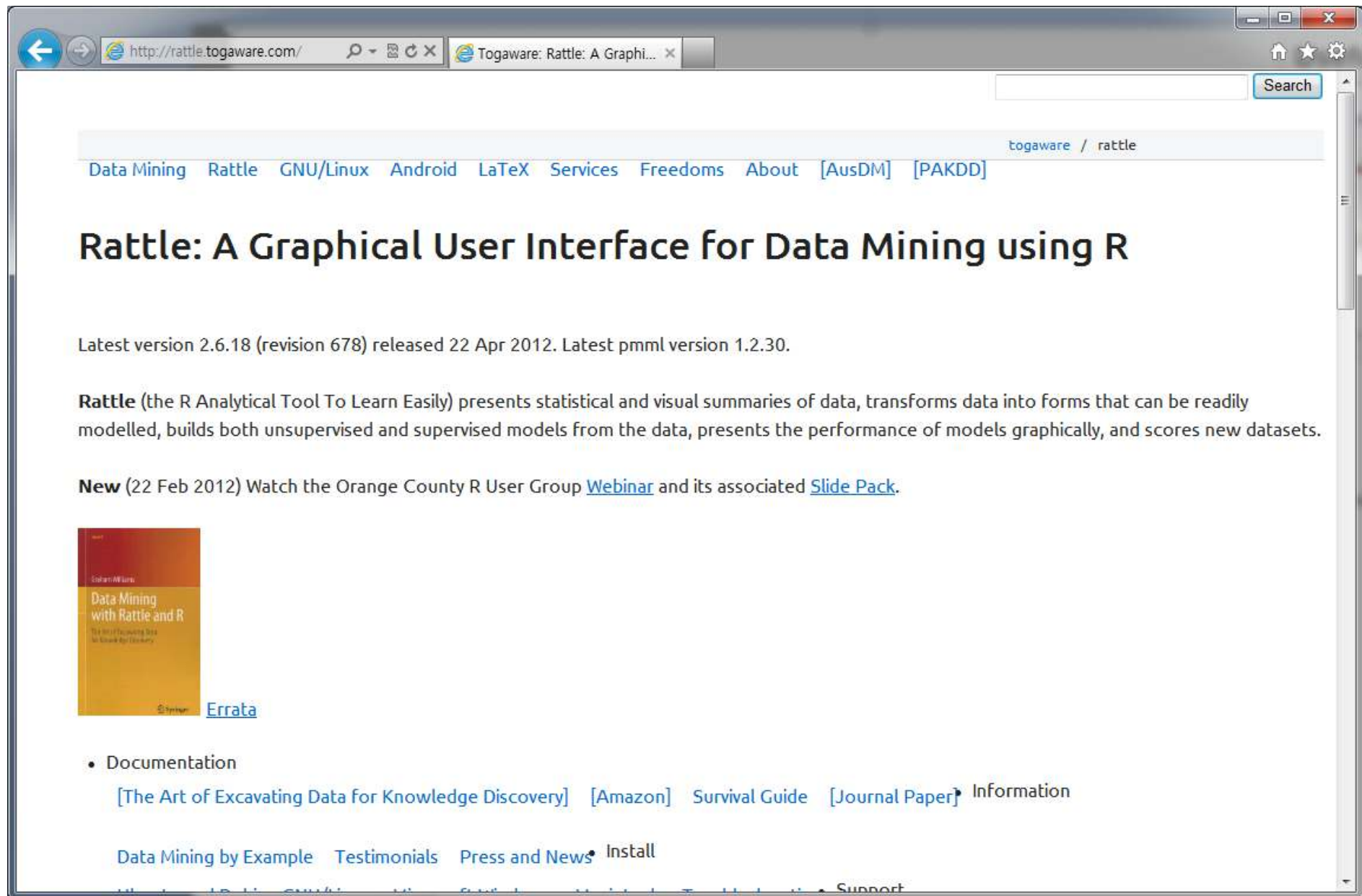
Residuals:
    Min     1Q   Median     3Q    Max
-33.204 -20.983  -4.748   13.957   61.433

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
    
```

At the bottom of the output window, there is a green bar that says 'Data Processed And Sent' and a 'Commit' button.

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



The screenshot shows a web browser window displaying the Rattle website. The browser's address bar shows the URL <http://rattle.togaware.com/>. The website has a navigation menu with links for Data Mining, Rattle, GNU/Linux, Android, LaTeX, Services, Freedoms, About, [AusDM], and [PAKDD]. The main heading is "Rattle: A Graphical User Interface for Data Mining using R". Below this, it states "Latest version 2.6.18 (revision 678) released 22 Apr 2012. Latest pmml version 1.2.30." A paragraph describes Rattle as "the R Analytical Tool To Learn Easily" that presents statistical and visual summaries of data, transforms data into forms that can be readily modelled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets. A "New" section (dated 22 Feb 2012) mentions a webinar by the Orange County R User Group and a slide pack. There is a book cover for "Data Mining with Rattle and R" by John Fox, with a link to "Errata". A "Documentation" section lists links for "[The Art of Excavating Data for Knowledge Discovery]", "[Amazon]", "Survival Guide", "[Journal Paper]", and "Information". At the bottom, there are links for "Data Mining by Example", "Testimonials", "Press and News", "Install", and "Support".

http://rattle.togaware.com/ Togaware: Rattle: A Graphi... x

togaware / rattle


Data Mining Rattle GNU/Linux Android LaTeX Services Freedoms About [AusDM] [PAKDD]

Rattle: A Graphical User Interface for Data Mining using R

Latest version 2.6.18 (revision 678) released 22 Apr 2012. Latest pmml version 1.2.30.

Rattle (the R Analytical Tool To Learn Easily) presents statistical and visual summaries of data, transforms data into forms that can be readily modelled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets.

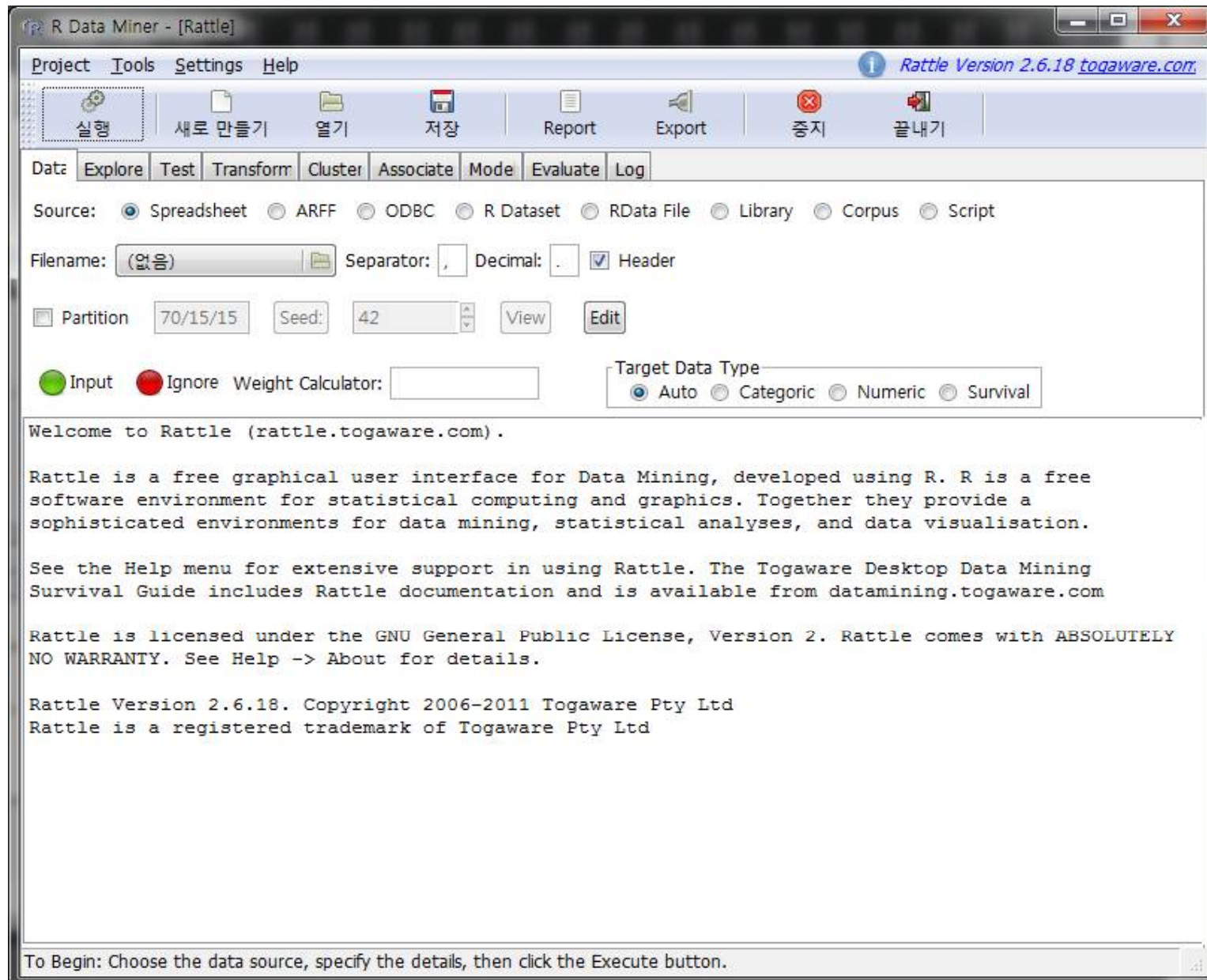
New (22 Feb 2012) Watch the Orange County R User Group [Webinar](#) and its associated [Slide Pack](#).



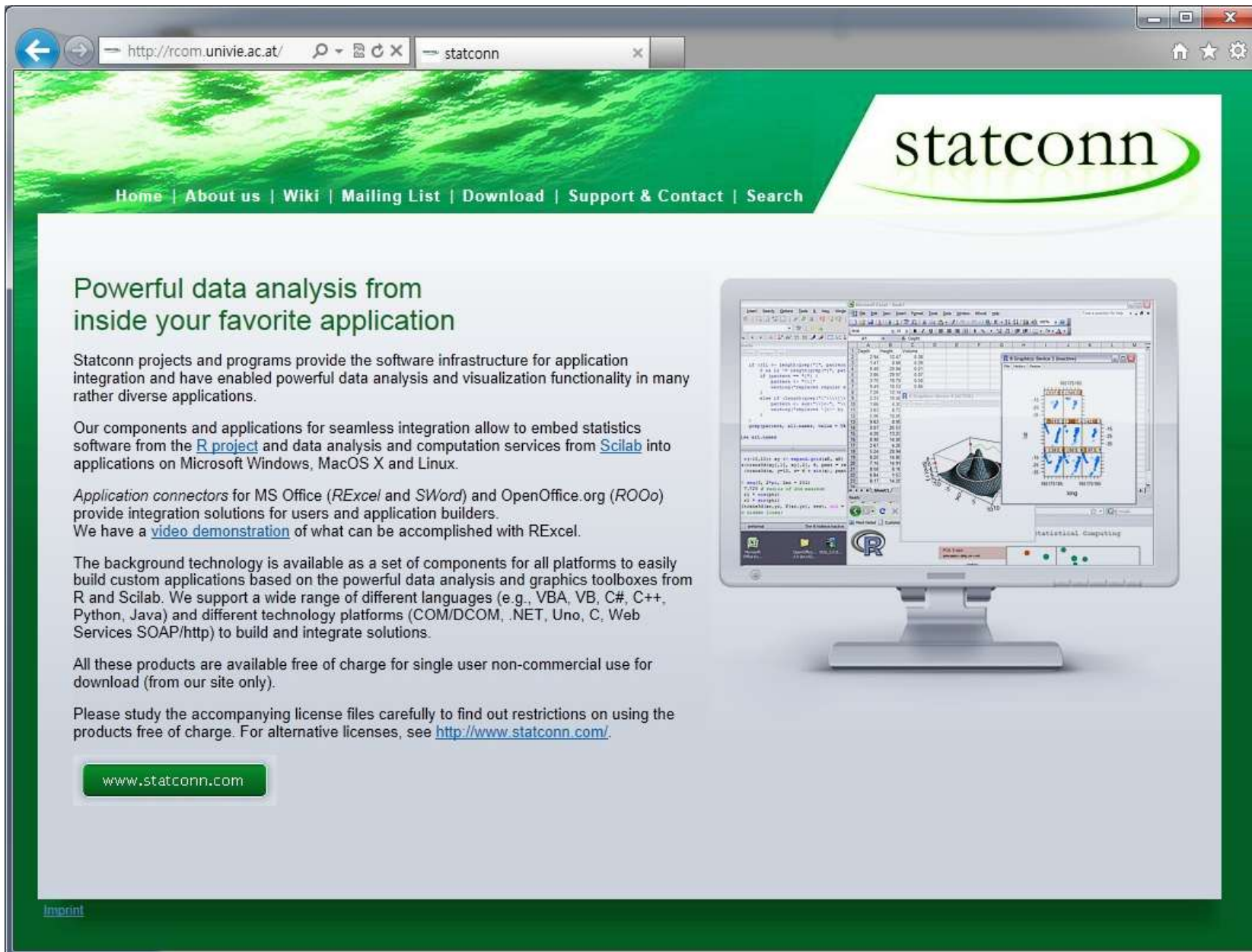
[Errata](#)

- Documentation
 - [\[The Art of Excavating Data for Knowledge Discovery\]](#) [\[Amazon\]](#) [Survival Guide](#) [\[Journal Paper\]](#) [Information](#)
 - [Data Mining by Example](#) [Testimonials](#) [Press and News](#) [Install](#)
 - [Support](#)

Rattle (library(rattle); rattle())



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



Home | About us | Wiki | Mailing List | Download | Support & Contact | Search

Powerful data analysis from inside your favorite application

Statconn projects and programs provide the software infrastructure for application integration and have enabled powerful data analysis and visualization functionality in many rather diverse applications.

Our components and applications for seamless integration allow to embed statistics software from the [R project](#) and data analysis and computation services from [Scilab](#) into applications on Microsoft Windows, MacOS X and Linux.

Application connectors for MS Office (*RExcel* and *SWord*) and OpenOffice.org (*ROOo*) provide integration solutions for users and application builders. We have a [video demonstration](#) of what can be accomplished with RExcel.


The background technology is available as a set of components for all platforms to easily build custom applications based on the powerful data analysis and graphics toolboxes from R and Scilab. We support a wide range of different languages (e.g., VBA, VB, C#, C++, Python, Java) and different technology platforms (COM/DCOM, .NET, Uno, C, Web Services SOAP/http) to build and integrate solutions.

All these products are available free of charge for single user non-commercial use for download (from our site only).

Please study the accompanying license files carefully to find out restrictions on using the products free of charge. For alternative licenses, see <http://www.statconn.com/>.

www.statconn.com

[Imprint](#)



RExcel

The screenshot shows the Microsoft Excel interface with the RExcel menu open. The menu items include: 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, and About RExcel. The worksheet contains a table with columns C, D, E, F, G, H, I, J, K, L, M. The data in the table is as follows:

	C	D	E	F	G	H	I	J	K	L	M
	Size	SizeM	SizeF								
	68	183	177	178							
	80	184	160	176							
	62	173	171	175							
	55	165	164	185							
	60	165	165	190							
	70	185	#N/A	185							
	81	179	165	175							
	93	185	173	184							
	70	178	172	184							
	75	168	160	180							
	65	165	160	170							
13	m	80	182	169	175						
14	m	83	193	165	186						
15	m	65	172	#N/A	#N/A						
16	m	73	173	168	176						
17	f	65	162	169	169						
18	m	73	173	158	168						

Instructions in the worksheet:

- Select the range to the left, A1:F33
The simplest way is to select a cell within the range and pressing Ctrl-Shift-*
- If R is not yet started, start it from RExcel->Start R
- Right-Click in the selected region and select Put R Dataframe
- Select the cell with the R command below (cell I13) and right-click Run R

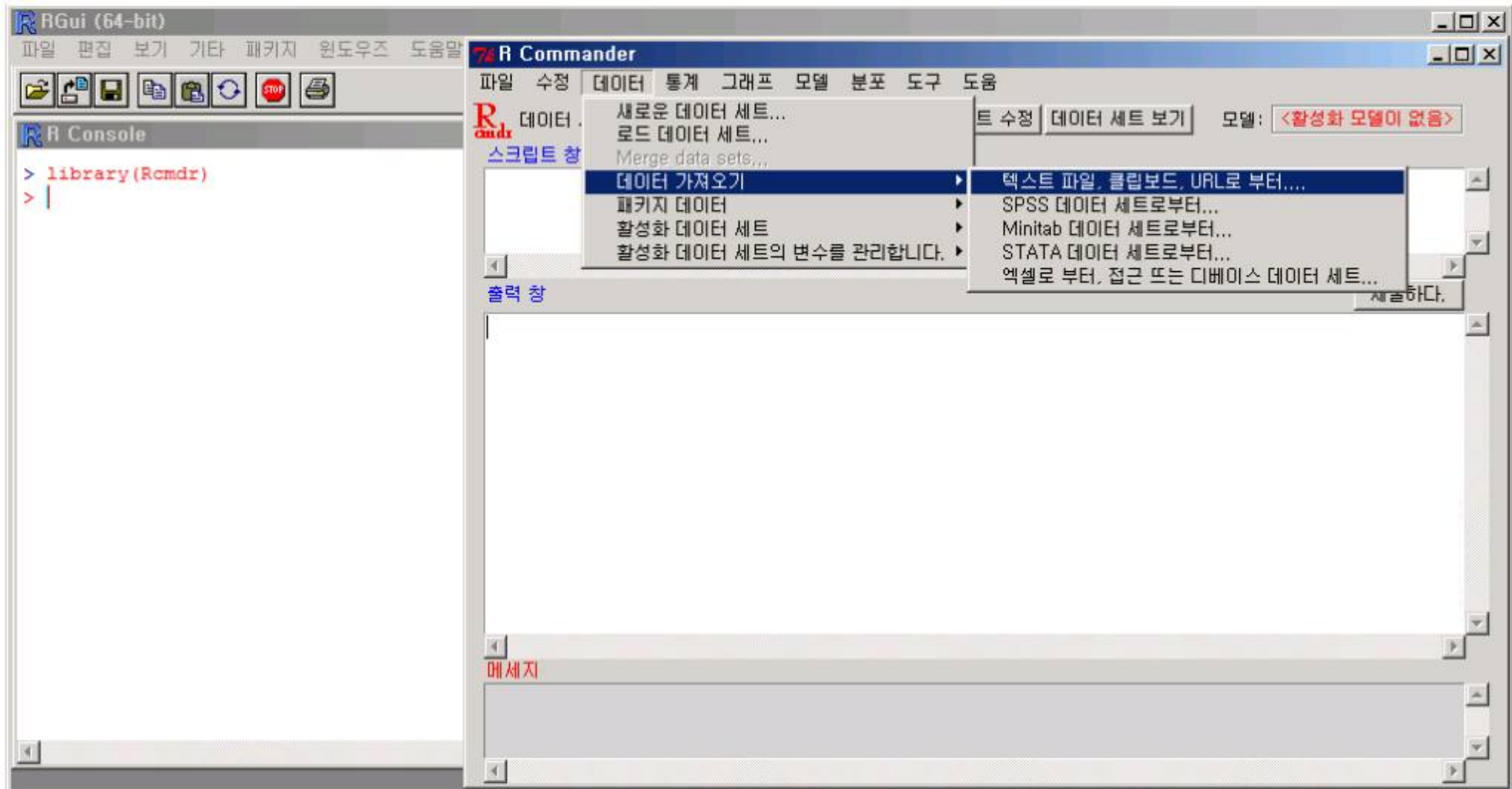
R command in cell I13:

```
mymmeans<-with(RDemoDev,tapply(Size,Gender,function(x)mean(x
```

Additional instruction:

Select the cell below (cell I19) and right-click Get R Value, enter mymmeans for th R expression, and check with rownames

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



R Commander

The screenshot displays the R Commander interface with the following components:

- RGL device 1 (active):** A 3D scatter plot showing the relationship between prestige (z-axis), education (x-axis), and income (y-axis). The data points are plotted on a blue mesh surface.
- R Graphics: Device 2 (ACTIVE):** Three 2D plots showing the effect of each variable on prestige:
 - education effect plot:** A line plot of prestige vs education, showing a positive linear relationship with a confidence interval.
 - income effect plot:** A line plot of prestige vs income, showing a positive linear relationship with a confidence interval.
 - type effect plot:** A line plot of prestige vs type (bc, prof, wc), showing a non-linear relationship with a confidence interval.
- R Commander Script Window:** Contains the following R code:


```
data(Prestige, package="car")
LinearModel.1 <- lm(prestige ~ education + income + type, data=Prestige)
summary(LM)
trellis.device(theme="col.whitebg")
plot(all.effects(LinearModel.1), ask=FALSE)
scatter3d(Prestige$education, Prestige$prestige, Prestige$income, fit="additiv
```
- Linear Model Dialog:** A dialog box for defining a new linear model. The model formula is set to `prestige ~ education + income + type`.
- Output Window:** Displays the summary statistics for the linear model:


```
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
```
- Messages:** A list of loaded packages including nlme, mvtnorm, multcomp, mgcv, lme4, lattice, foreign, effects, grid, car, abind, and rgl.

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

The screenshot shows the Revolution Analytics website with the following content:

- Navigation:** "What is R?", "Products", "Services", "R Downloads", "Why Revolution R?", "Support", "News & Events", "About Us".
- Header:** "REVOLUTION ANALYTICS", "Read the Revolutions Blog", "Follow us on Twitter", "Get our Newsletter", "Buy Now", "Google Custom Search", "Search".
- Main Heading:** "High Performance R Analytics for the Enterprise".
- Feature List:**
 - Big Data Analytics
 - High Performance Computing
 - Open Source R
 - Analytics in Production
 - R Services: Training & Consulting
- Call-to-Action Boxes:**
 - Download Revolution R FREE
 - Free Enterprise Software for Academics
 - Revolution R Enterprise for Production-Grade Analytics
 - Visit our R Community Site inside-R.org
- OUR CUSTOMERS:** Logos for VISA, ACR, NYU, NOVARTIS, and Marketo.
- WEBINARS:** "Achieving High-Performing, Simulation-Based Operational Risk Measurement with RevoScaleR Thursday, June 28, 2012".
- LATEST NEWS & EVENTS:** "Revolution R Enterprise Boosts Big Data Analytics Capabilities", "Revolution Analytics Webinar: Using R and Putting Business Analytics to Work", "Revolution Analytics Names David Rich New CEO", "Revolution Analytics Announces 'Applications of R in Business' Contest Winners", "Revolution R Enterprise Delivers New Big Data Analytics Capabilities", "Revolution Analytics Partners With Cloudera To Deliver Comprehensive New Big Analytics Solution", "97 Percent of Data Scientists Say 'Big Data' Technology Solutions Need Improvement".
- Awards:** "Gartner | 2011 COOL VENDOR" badge and "Gartner Names Revolution Analytics a 'Cool Vendor' for Business Intelligence".
- Footer:** "Revolution Analytics delivers advanced analytics software at half the cost of existing solutions. By building on open source R, the world's most powerful statistics software—with innovations".

Revolution R

The screenshot displays the Revolution R Enterprise IDE for Windows. The main window is titled "Chap06. Basic graphs.R" and contains the following R code:

```
xlab="Miles Per Gallon")

#6.6.0.2

x <- mtcars[order(mtcars$mpg),]
x$cyl <- factor(x$cyl)
x$color[x$cyl==4] <- "red"
x$color[x$cyl==6] <- "blue"
x$color[x$cyl==8] <- "darkgreen"
dotchart(x$mpg,
  labels = row.names(x),
  cex=.7,
  groups = x$cyl,
  gcolor = "black",
  color = x$color,
  pch=19,
  main = "Gas Mileage for Car Models\ngrouped by cylinder",
  xlab = "Miles Per Gallon")
```

The console window shows the execution of the code:

```
Revolution R Enterprise Console
> x <- mtcars[order(mtcars$mpg),]
> x$cyl <- factor(x$cyl)
> x$color[x$cyl==4] <- "red"
> x$color[x$cyl==6] <- "blue"
> x$color[x$cyl==8] <- "darkgreen"
> dotchart(x$mpg,
+ labels = row.names(x),
+ cex=.7,
+ groups = x$cyl,
+ gcolor = "black",
+ color = x$color,
+ pch=19,
+ main = "Gas Mileage for Car Models\ngrouped by cylinder",
+ xlab = "Miles Per Gallon")
>
```

The Solution Explorer on the right shows a project structure for "Solution 'RStudy' (3 projects)":

- Solution 'RStudy' (3 projects)
 - googleVis
 - 00.Info.R
 - 01.MotionChart.R
 - Igraph
 - RInAction
 - Chap01. Introduction to R.R
 - Chap02. Creating a dataset.R
 - Chap03. Getting Started with graphs.R
 - Chap06. Basic graphs.R
 - Chap07. Basic statistics.R
 - Chap11. Intermediate graphs.R
 - Chap16. Advanced graphics.R

The Object Browser on the right shows the search results for "on_aboutdialog_response" and the Global Environment:

Search: on_aboutdialog_response

- age
- suppressRattleWelcome
- weight
- x
 - counts
 - crs
 - crv
 - Global_rattleGUI

Global Environment: Type: environment

At the bottom, the status bar indicates "Compute Context: local | Ready".

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

WalWare

aktuell Info/?

Intranet: (Login) (Mailer)

Bits & Bytes

News

StatET

- Installation
- Community & Help
- Troubleshooting
- Log / Notes
- Tips
- Contributors

RJ

Downloads

StatET for R

StatET is an [Eclipse](#) based IDE (integrated development environment) for [R](#). It offers a [set](#) of mature tools for R coding and package building. This includes a fully integrated R Console, Object Browser and R Help System, whereas multiple local and remote installations of R are supported.

StatET is provided as plug-in for the Eclipse IDE. Therefore the user can combine it with a wide range set of tools working on top of the Eclipse Platform. Like R and Eclipse, StatET is open source software, and works on many operating systems.

Feedback, suggestions for improvement and [contributors](#) are [welcome!](#)

Latest News

- **2012-05-31: StatET 3.0 / RJ 1.1 final**

We are pleased to announce the availability of the final version of StatET 3.0.0 [2012-05-20].

The release includes a new editor for Sweave documents providing most features known from the R editor, for LaTeX as well as for R code. You find a list of all major changes in the [news](#).

StatET 3.0 is available for Eclipse 3.6, 3.7 as well as the upcoming version 3.8. The new StatET version requires RJ 1.1. So if you update StatET from 2.x, do not forget to update the R packages 'rj' and 'rj.gd'. For detail please check the [installation instructions](#).

- **2012-01-11: StatET 2.0.3**

The new maintenance update of StatET, version 2.0.3 [2011-12-22], is available on all update-sites now. There are

To bookmark / link to StatET directly use:
www.walware.de/goto/statet

Homepage R Project
www.r-project.org

Homepage Eclipse
www.eclipse.org

Kontakt Site Map

Copyright © 2012 WalWare | [About/Impressum](#) | www.walware.de

StatET for R

The screenshot displays the StatET for R IDE interface. The main editor window shows the following R code:

```
x <- rnorm(1000)
y <- runif(1000)
plot(x, y)
```

The console window at the bottom shows the execution output:

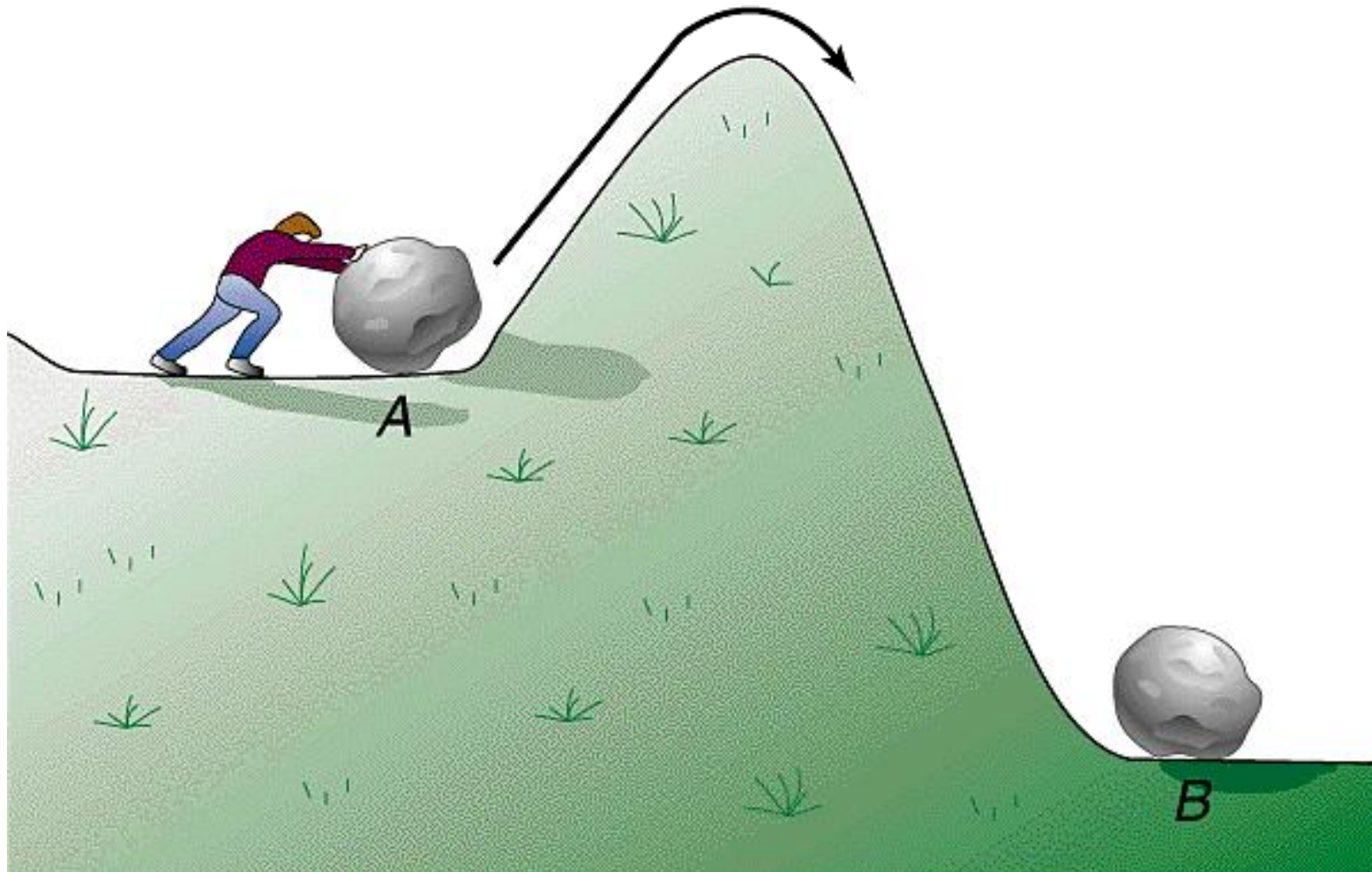
```
R_Console2.9.2 [R Console] R-2.9.2 / RJ (Jan 16, 2010 12:46:09 PM) · C:\Documen ... \longhow\workspace <idle>
[46] -0.916874408 -0.212230190  0.221905343  2.113768105 -1.545655300
[51] -0.731139412 -1.747556773  1.563616386  0.180674884 -1.168817533
[56]  1.540567088  0.496016614 -0.738998498  1.722125378 -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)
```

The interface includes a Project Explorer on the left showing a project named 'testR' with files 'test.R' and 'testR2.R'. The Object Browser at the bottom left shows installed packages like 'stats', 'graphics', and 'base'. The Outline and Templates panels on the right provide a list of R code constructs such as 'else', 'for', and 'function'.

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>

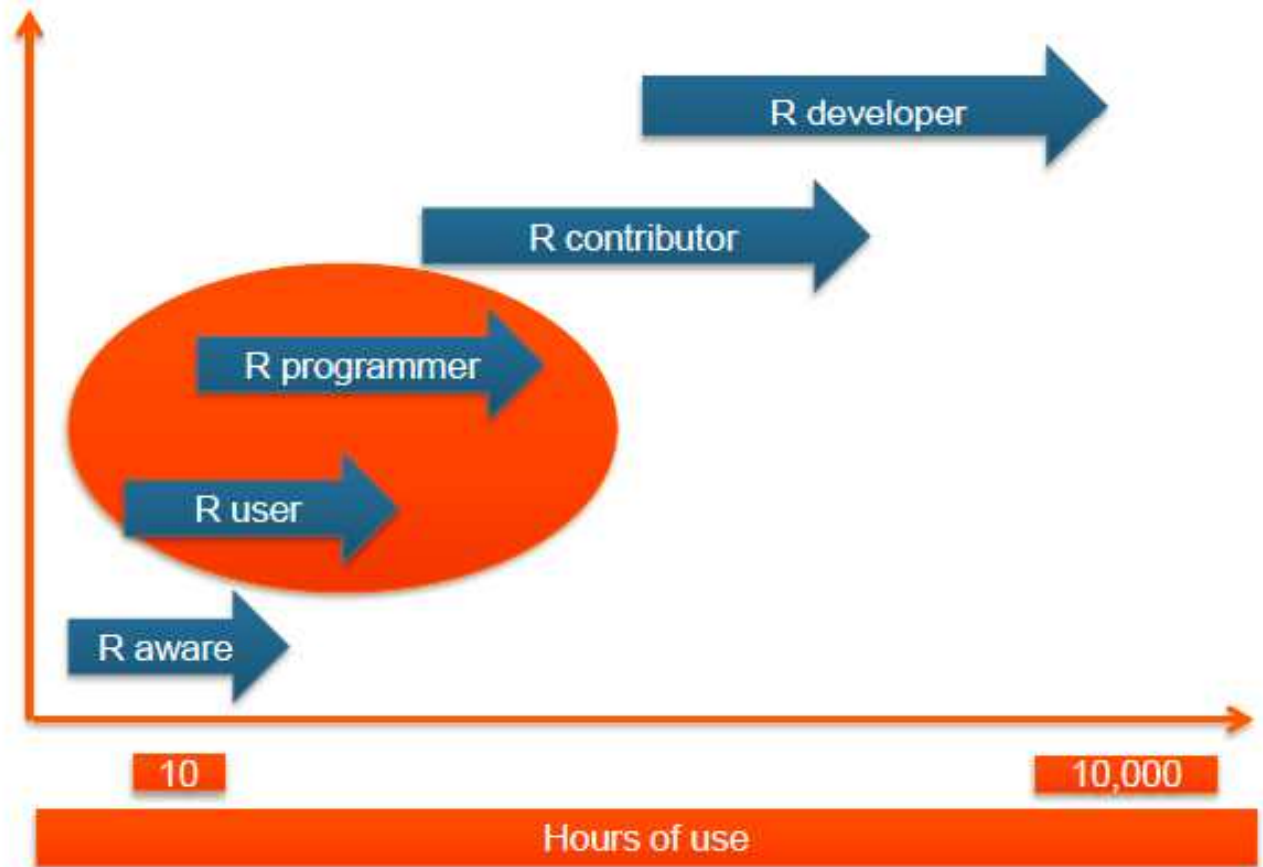
R을 배워 업무에 적용하려면



Learning R

Levels of R Skill

- Write production level code
- Write an R package
- Write functions
- Use R Functions
- Use a GUI



The Malcolm Gladwell "Outlier" Scale

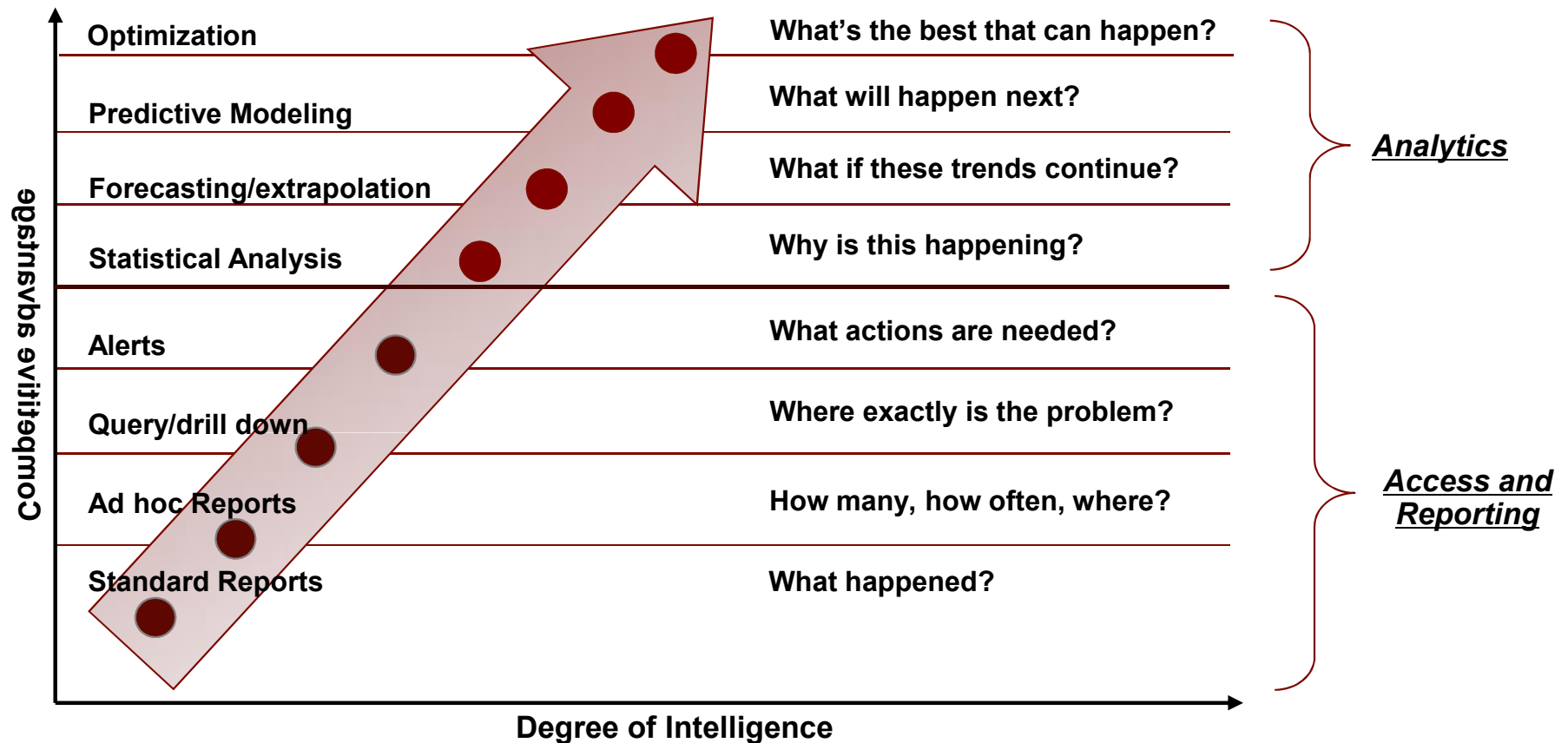
출처 : Revolution Analytics

R, 그리고 빅데이터

Advanced Analytics

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

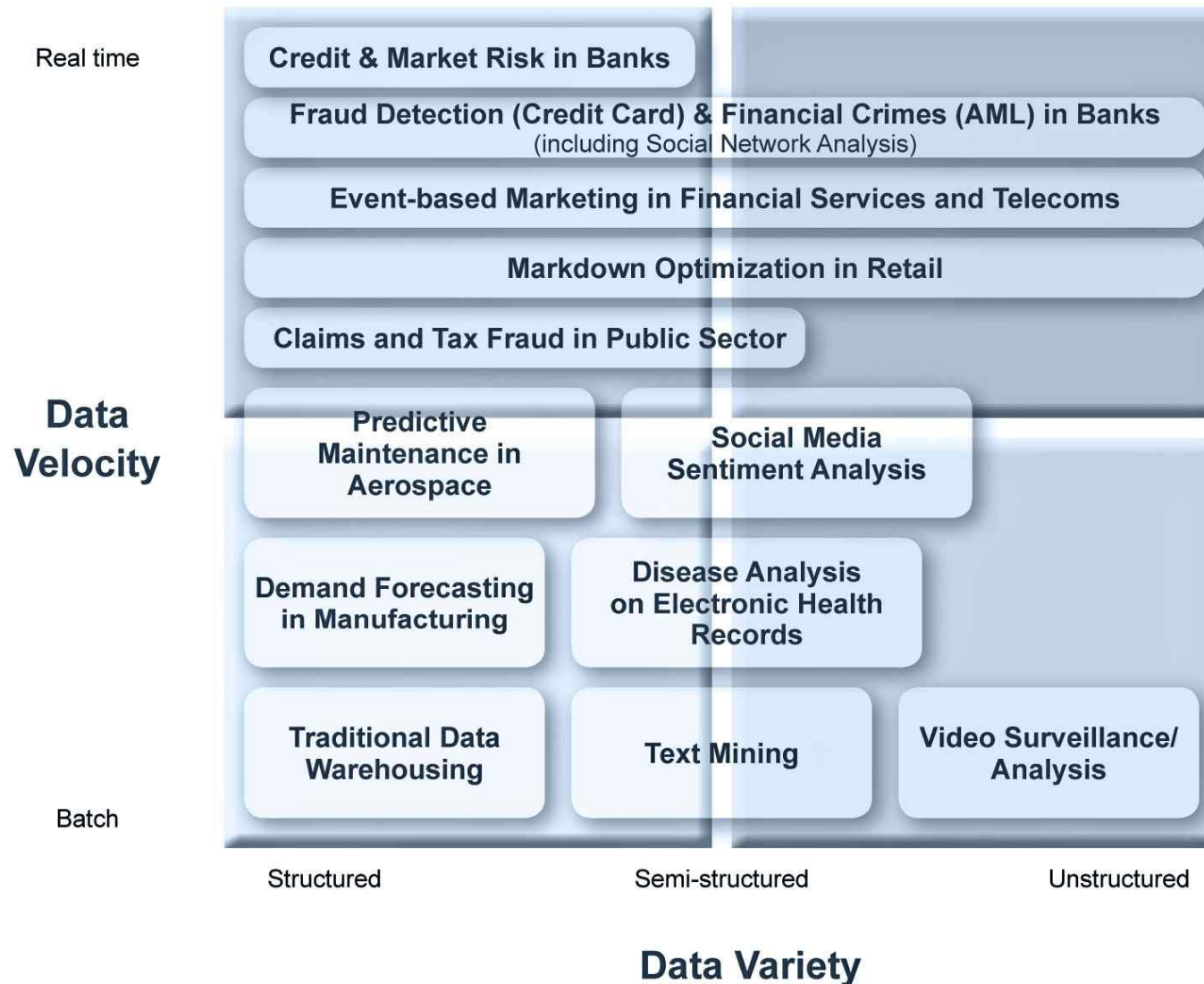
대부분의 기업들은 고급 분석이 가능한 분석 전문가들을 리포팅 작성에 활용하고 있습니다



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

What is Big Data Analysis?

Potential Use Cases for Big Data Analytics

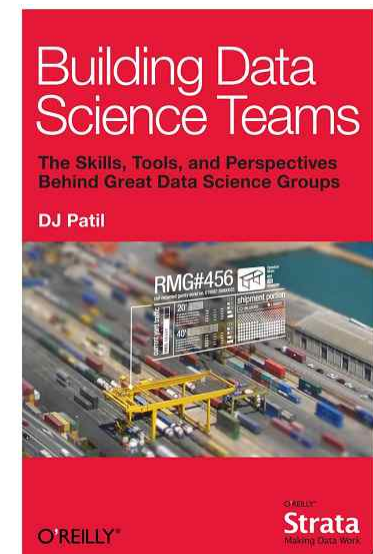


출처 : SAS and IDC

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.



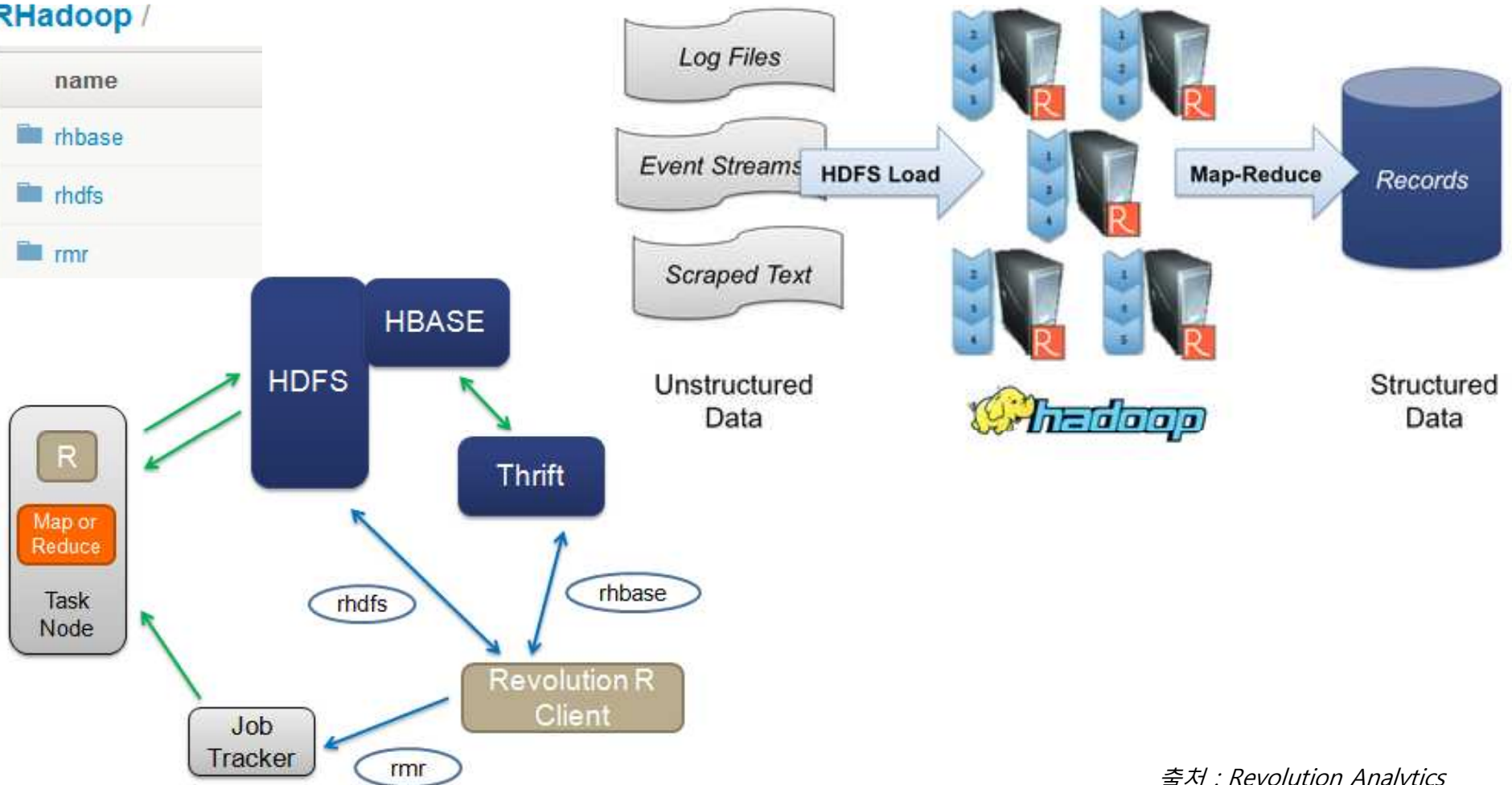
Facebook 에서 RHIPE 소개하는 Video (2010.03.09)
<http://www.lecturemaker.com/2011/02/rhipe/>

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

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

RHadoop /

name
rhbase
rhdfs
rmr



출처 : Revolution Analytics

RHadoop (Word Count Example)

Example: Word Count

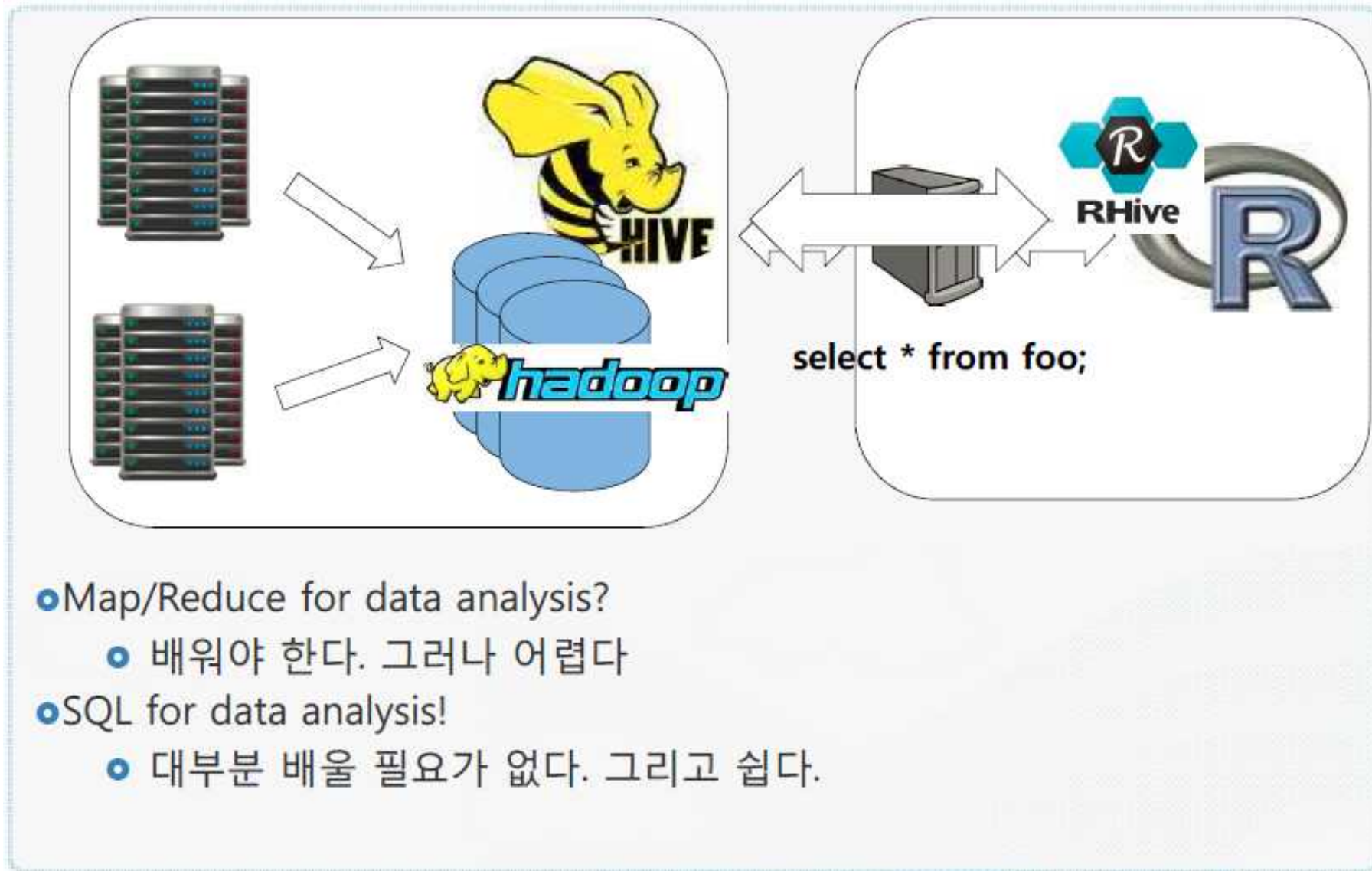
Data preparation:

```
1 > library("hive")
2 Loading required package: rJava
3 Loading required binaries:
4 > hive_start()
5 > hive_is_available()
6 [1] TRUE
7 > DFS_put("~/tmp/Reuters.txt", "/tmp/Reuters.txt")
8 > DFS_list("/tmp/Reuters.txt")
9 [1] "reut-000000"
10 [4] "reut-000000"
11 [7] "reut-000000"
12 > head(DFS_read_lines("/tmp/Reuters.txt"))
13 [1] "<?xml version='1.0' encoding='UTF-8'>"
14 [2] "<REUTERS>"
15 [3] "<DATE>2008-08-28T14:00:00Z"
16 [4] "<TOPICS>"
17 [5] "<PLACES>"
18 [6] "<D>usa"
19 }
```

```
1 mapper <- function(){
2   mapred_write_output <- function(key, value)
3     cat(sprintf("%s\t%s\n", key, value), sep = "\n")
4
5   trim_white_space <- function(line)
6     gsub("(^ +)|(+ $)", "", line)
7   split_into_words <- function(line)
8     unlist(strsplit(trim_white_space(line), " "))
9
10  reducer <- function(){
11    [...]
12    env <- new.env(hash = TRUE)
13    con <- file("stdin", open = "r")
14    while (length(con) > 0) {
15      warn <- FALSE
16      line <- trim_white_space(readLines(con, n = 1))
17      words <- split_into_words(line)
18      if (length(words) > 0) {
19        count <- split_into_words(words)
20        if (nchar(words) > 0) {
21          if (exists(words, envir = env)) {
22            oldcount <- env[[words]]
23            assign(oldcount + 1, words, envir = env)
24          } else assign(1, words, envir = env)
25        }
26      }
27      close(con)
28      for (w in ls(envir = env)) {
29        cat(w, "\t", env[[w]], "\n", sep = "")
30      }
31    }
32  }
33 }
```

```
1 > hive_stream(mapper = mapper,
2               reducer = reducer,
3               input = "/tmp/Reuters.txt",
4               output = "/tmp/Reuters_out")
5 > DFS_list("/tmp/Reuters_out")
6 [1] "_logs" "part-000000"
7 > results <- DFS_read_lines(
8   "/tmp/Reuters_out/part-000000")
9 > head(results)
10 [1] "-\t2" "--\t7"
11 [3] ":\t1" ".\t1"
12 [5] "0064</UNKNOWN>\t1" "0066</UNKNOWN>\t1"
13 > tmp <- strsplit(results, "\t")
14 > counts <- as.integer(unlist(lapply(tmp, function(x)
15   x[[2]])))
16 > names(counts) <- unlist(lapply(tmp, function(x)
17   x[[1]]))
18 > head(sort(counts, decreasing = TRUE))
19 the to and of at said
20 58 44 41 30 25 22
```

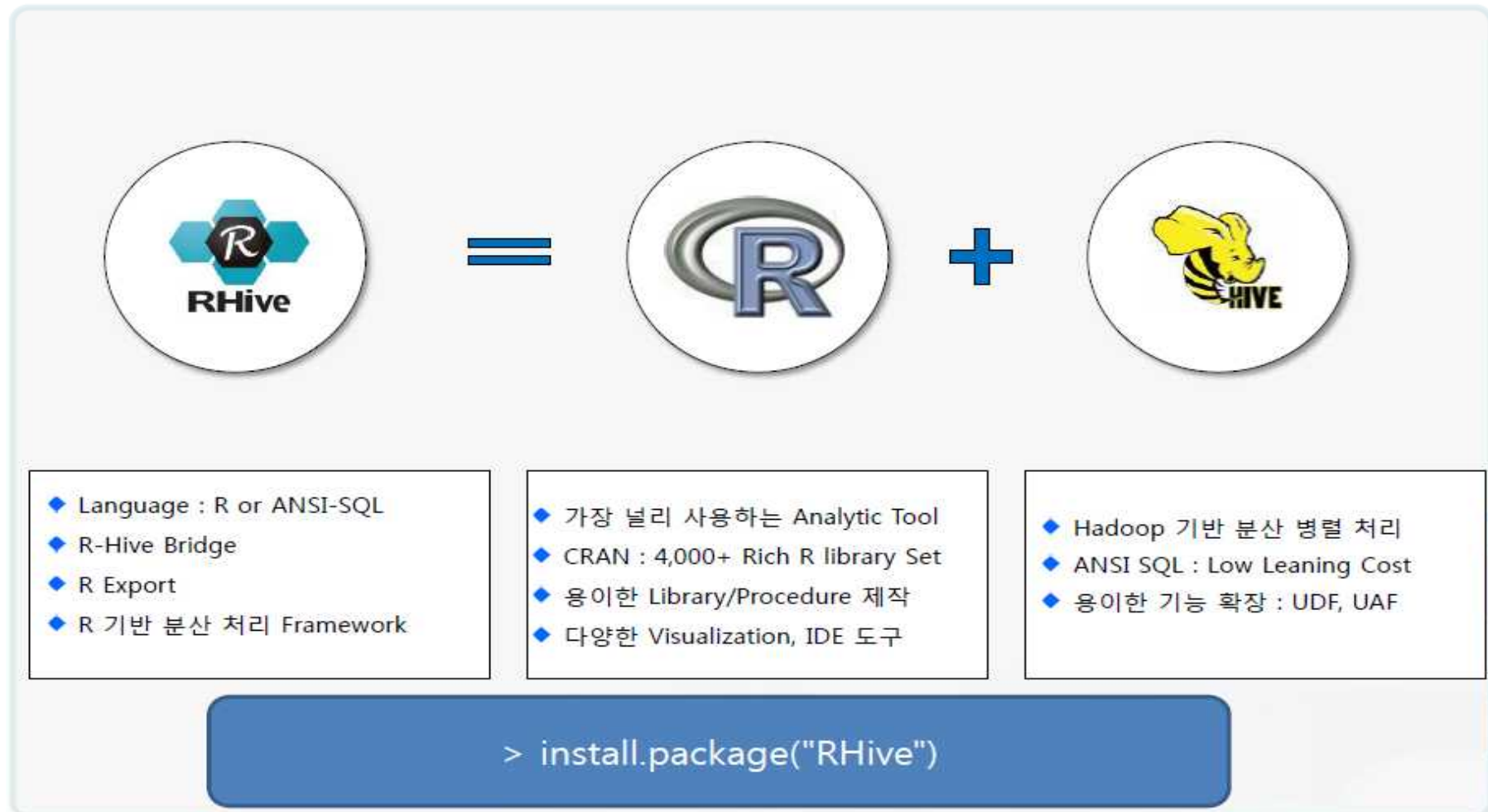
RHive – Motivation of Rhive (NexR)



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

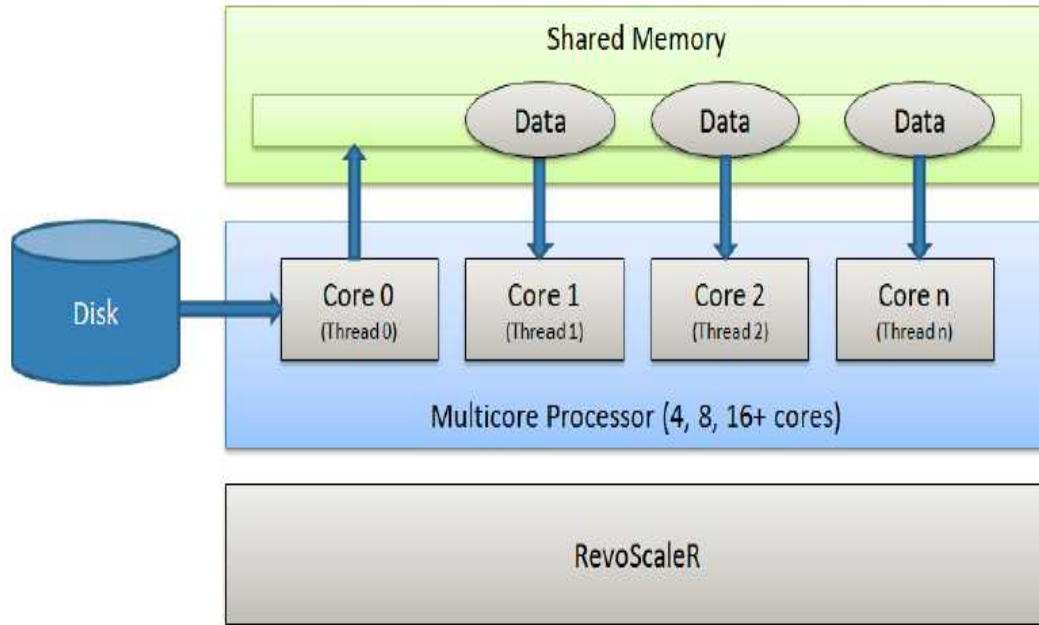
- The contents of the project: Collect KT Wireless CDR(Call Detail Record) to NDAP, NexR Big Data Platform and perform call quality analysis and reporting
- Analysis using RHive
 - Call Anomaly Detection
 - Location-based call quality analysis
 - Location-based subscriber clustering
 - SNA Analysis
-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



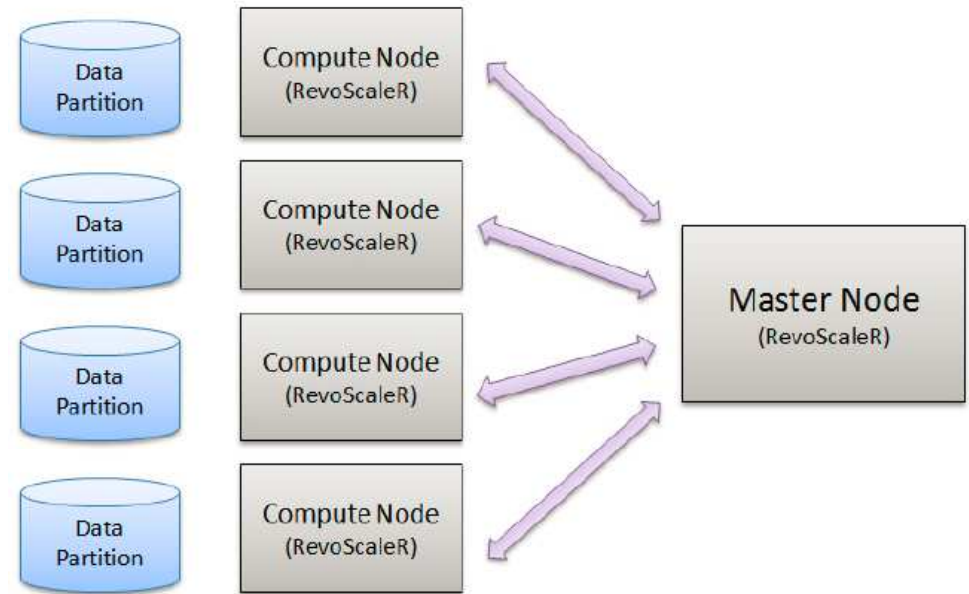
출처 : http://www.nexr.co.kr/nexrcorp_en/products_and_services/rhive_useCase.php

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

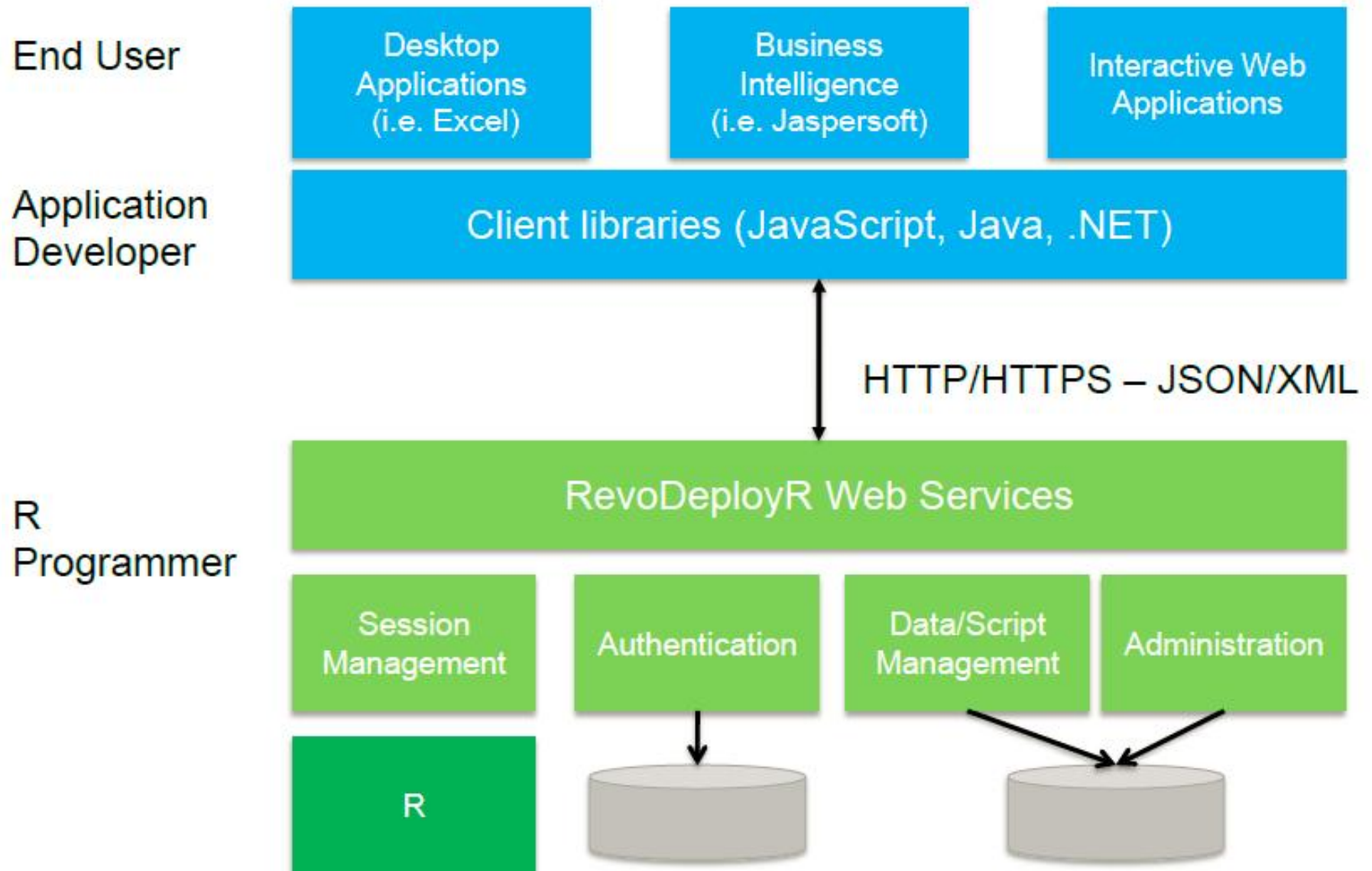
High-Performance Computing

(<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** → RHIFE, 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



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?

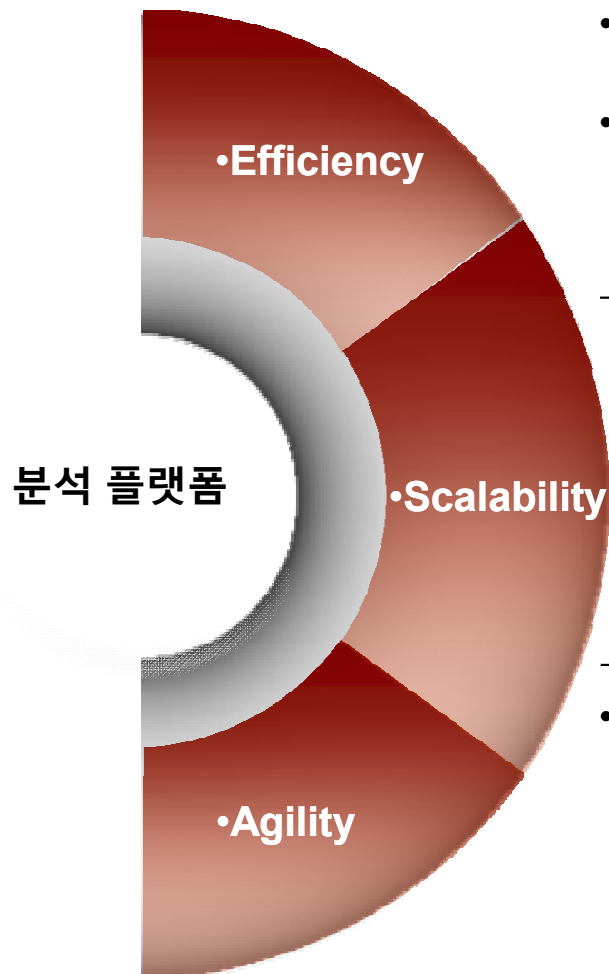
- 저비용으로 빅데이터 분석환경을 구축 할 수 있는지?

: 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