Next Revolution
Toward Open Platform

R의 이해와 응용

2012-05-29

유충현

넥스알 – Data Science Team (antony.ryu@nexr.com)
KRUG – (bdboy@r-project.kr)

목차

- 업계에서의 R의 관심
- Data Analytics을 위한 R의 소개
- R을 이용한 데이터 분석의 비교
  - Small Data Analytics using Native R
  - Large Data Analytics using R
  - Big Data Analytics using R
- Visualization

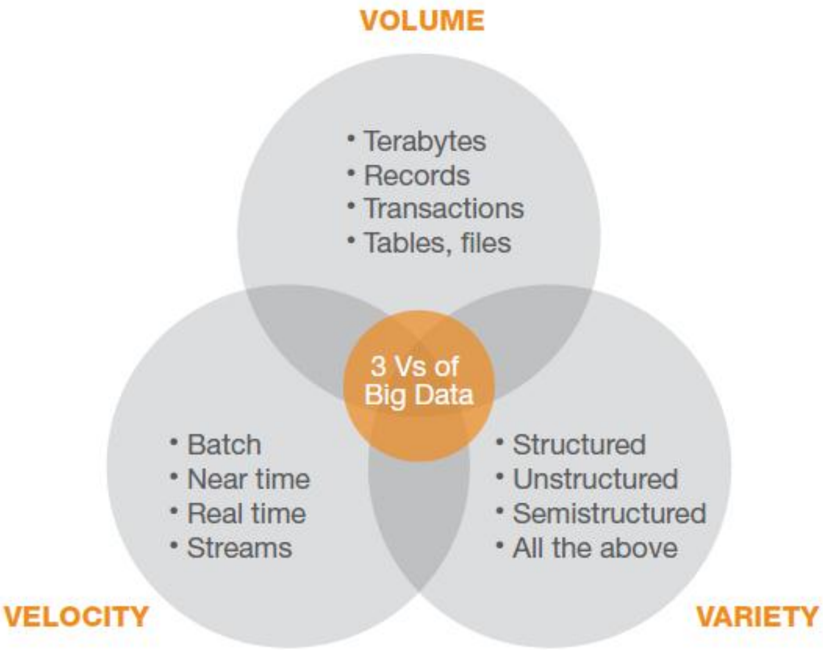
업계에서의 R의 관심

- Big Data Analytics
- BioConductor
- Clone of S System

업계에서의 R의 관심 – Big Data Analytics

Big Data를 있는 그대로 탐색하여 숨어 있는 Detail한 비즈니스 기회를 찾아내는 기술

Big Data 정의 (1)



Big Data Analytics(2) 정의

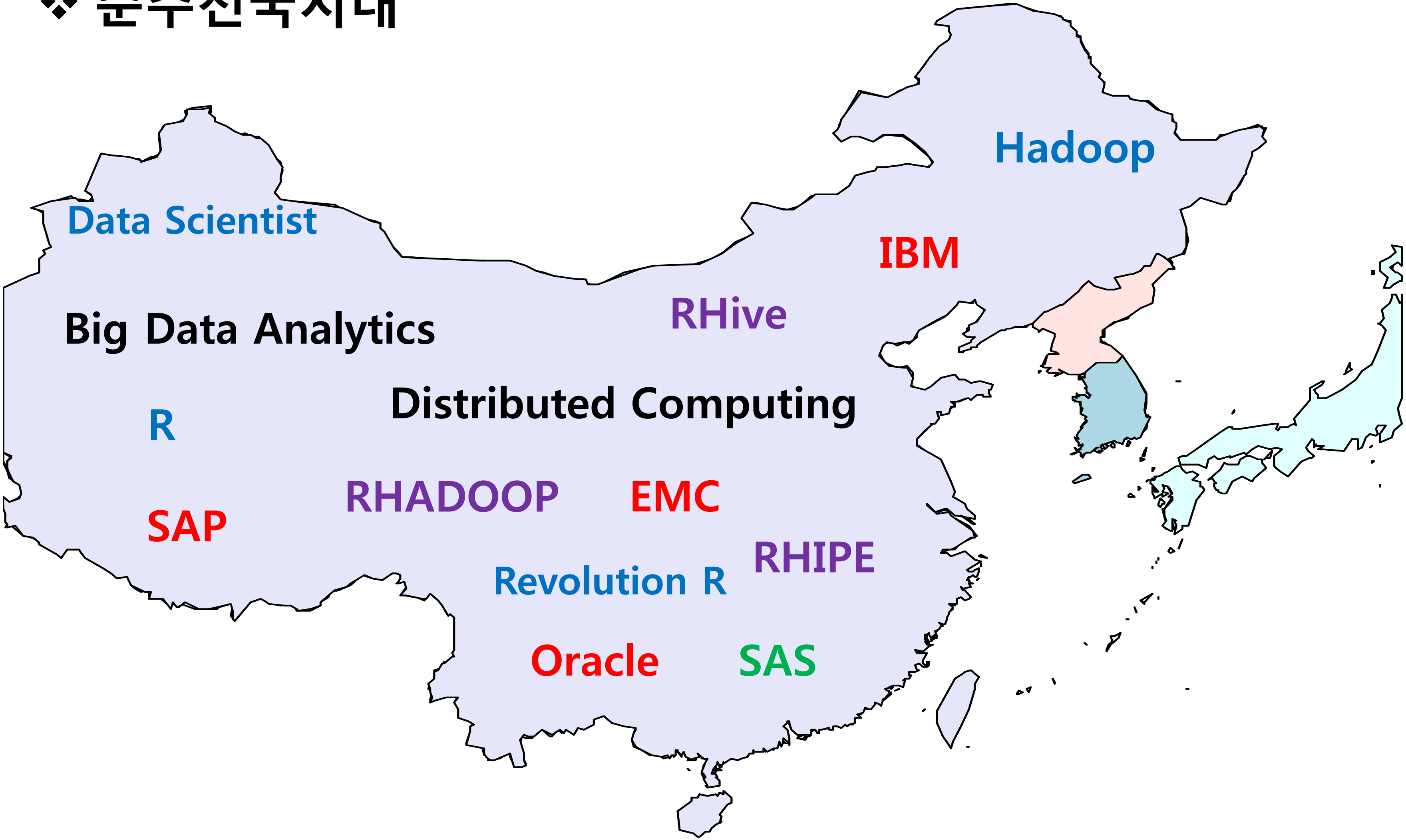
- Advanced Analytics, Discovery Analytics
  - Discovery of new business facts with plenty of detail (Big Data)
- Pareto's Law → Long-Tail Law
  - 데이터를 가공 (Sampling, Summary)하면 Long-tail(Detail)이 사라지거나 오차발생

Detail한 정보의 손실 없이 Big Data를 분석하는 기술

※ (1),(2) TDWI RESEARCH 2011 4Q : Big Data Analytics. http://tdwi.org

Big Data Analytics의 기술 및 시장환경

❖ 춘추전국시대



Big Data Analytics의 기술 및 시장환경 – R의 사용

● Appliance DBMS for Big Data Analytics

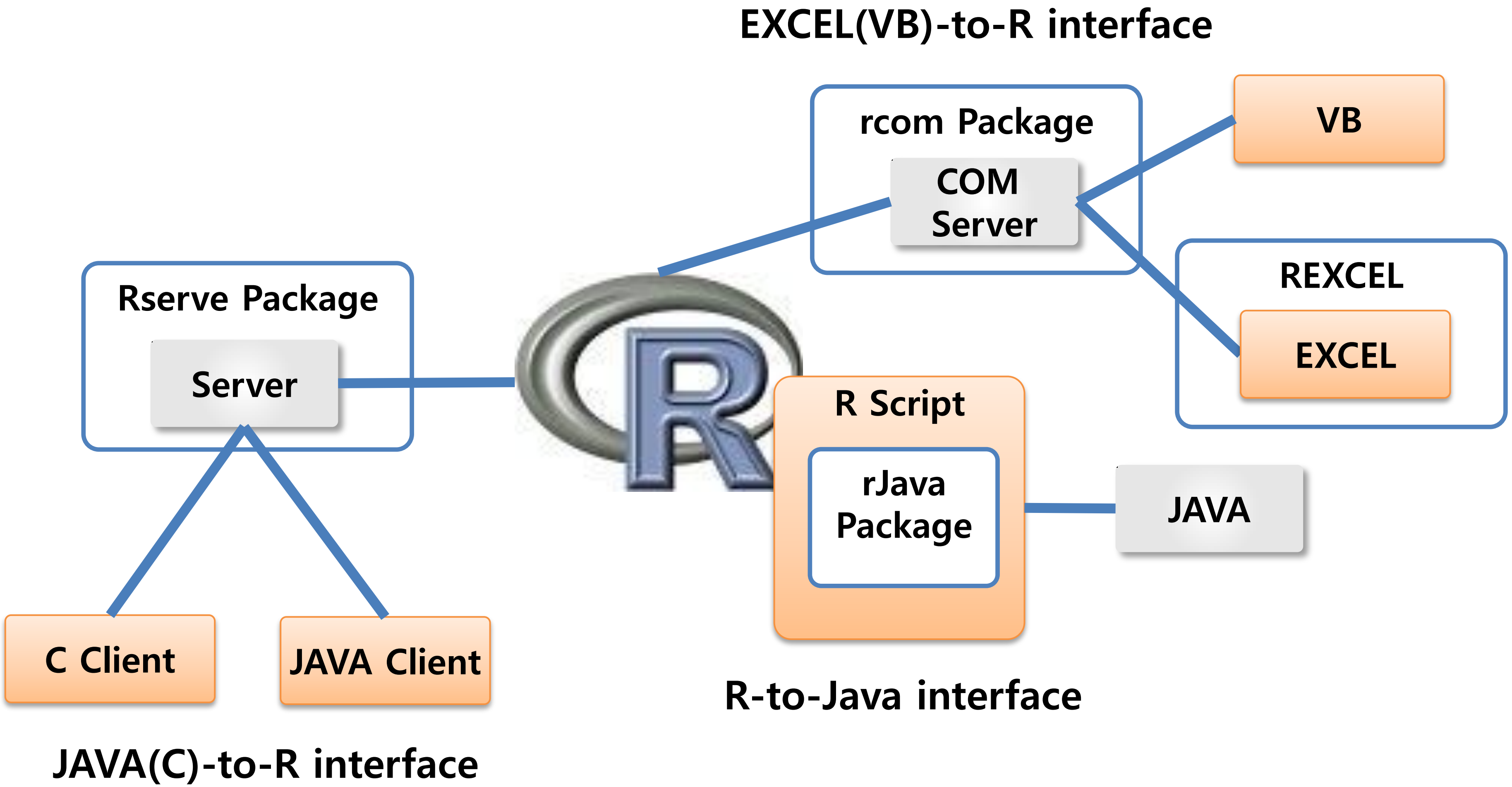
벤더	제품	Analytics Engine
Oracle	• Big Data Appliance • Exadata	Oracle R Enterprise (R)
IBM	• InfoSphere BigInsights • Netezza Appliance	Netezza & Revolution R 연동 사례
Teradata	Aster Discovery Platform	SQL-Map/Reduce, SAS, R
EMC	Greenplum Data Computing Appliance	Java, R
SAP	HANA (In memory Appliance) – Not Big Data	R 연동 사례

[특징]

- ❖ Appliance DBMS & Hadoop
  - ❖ Hadoop보다는 Appliance DBMS에 치중
- ❖ Analytics
  - ❖ Analytics Product을 DBMS Product 내부에 포함 시키고 있음
  - ❖ Analytics Engine은 공통적으로 R을 사용

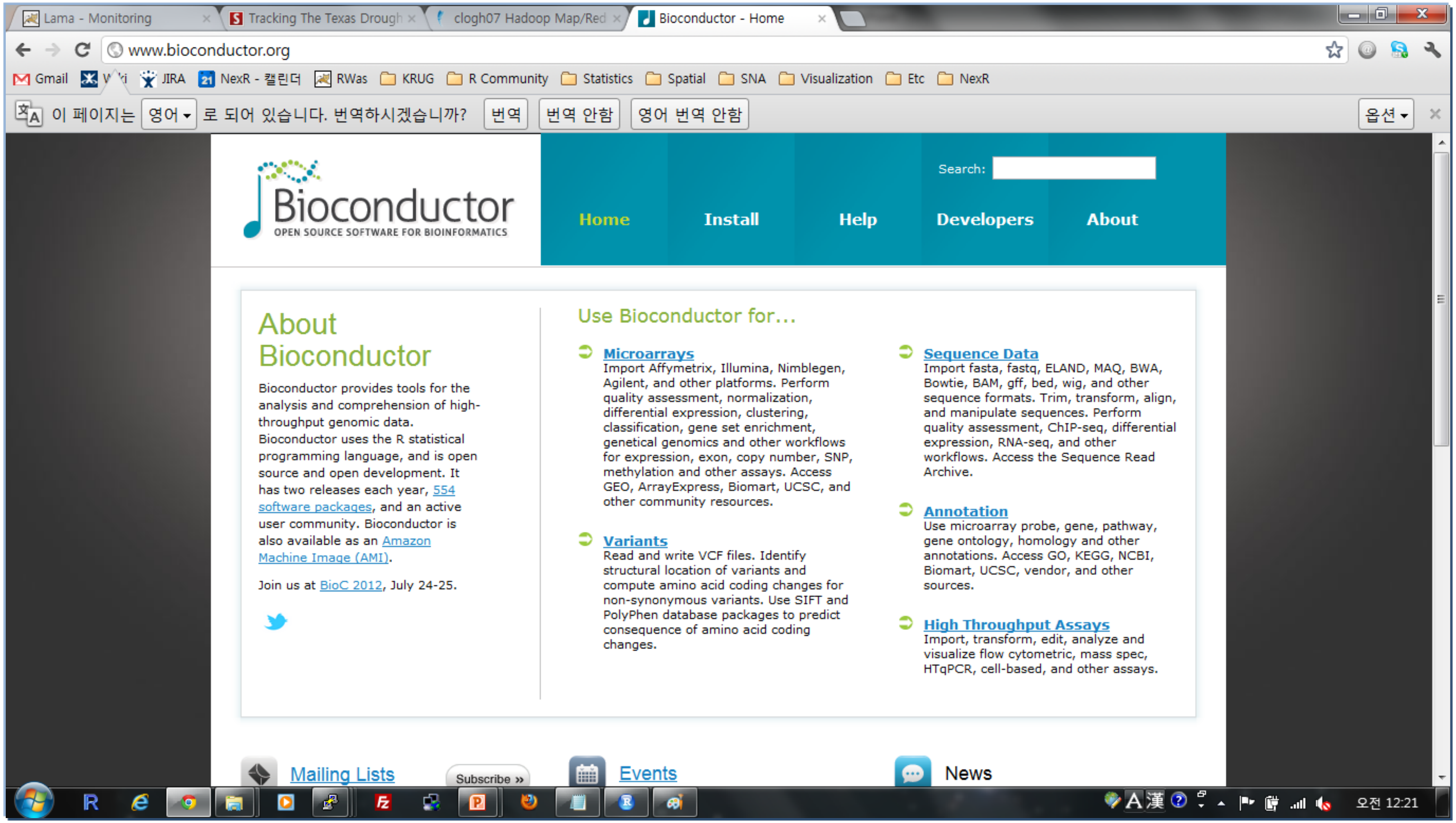
R의 Connectivity – 시스템 통합을 위한 요인

R의 System Integration 예시



Open Sources – 새로운 분석 방법론의 수용을 위한 요인

Bio Analytics의 표준(Bioconductor)



About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 554 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).

Join us at BioC 2012, July 24-25.

Use Bioconductor for...

Microarrays
Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.

Variants
Read and write VCF files. Identify structural location of variants and compute amino acid coding changes for non-synonymous variants. Use SIFT and PolyPhen database packages to predict consequence of amino acid coding changes.

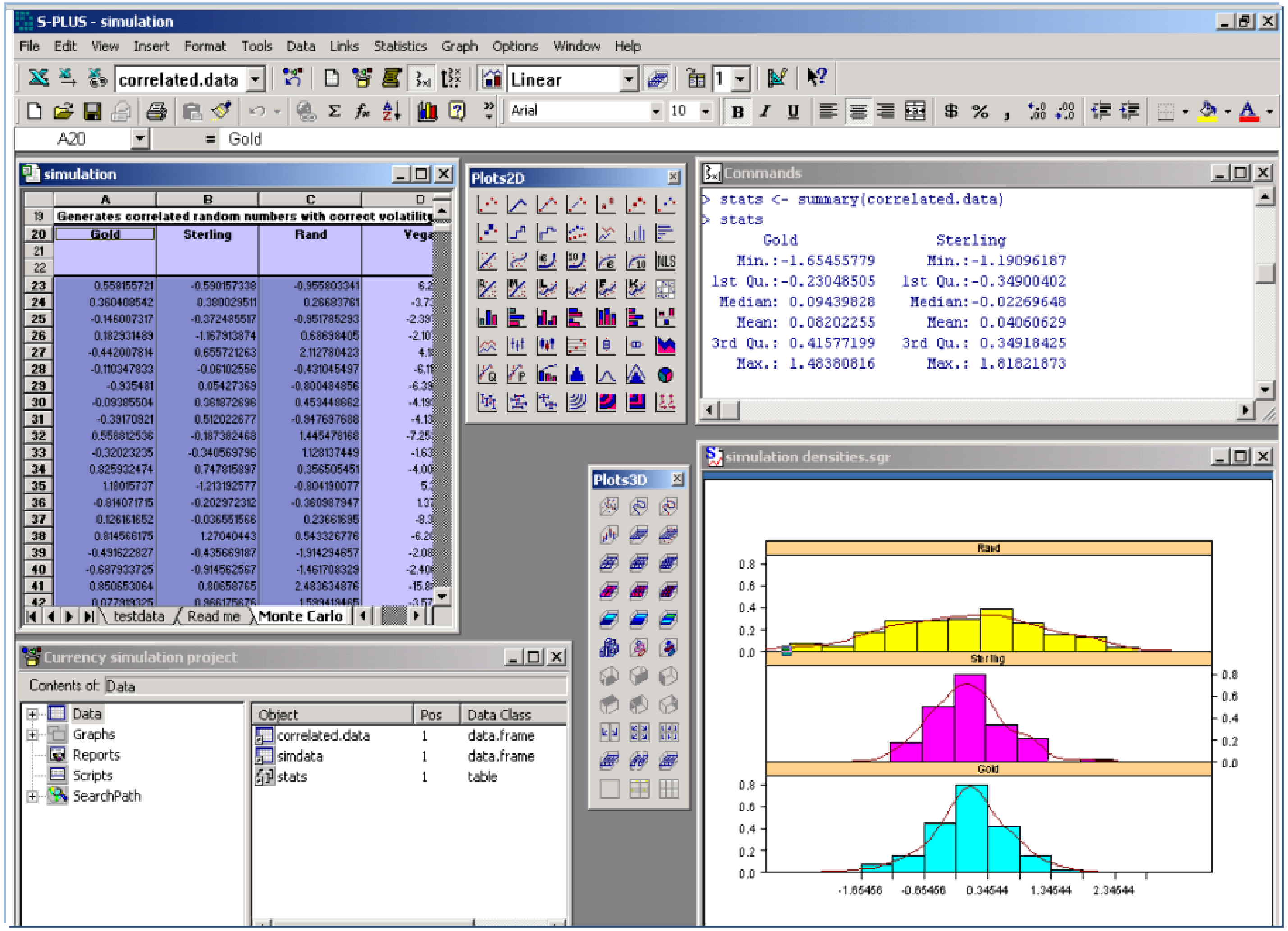
Sequence Data
Import fasta, fastq, ELAND, MAQ, BWA, Bowtie, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.

Annotation
Use microarray probe, gene, pathway, gene ontology, homology and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.

High Throughput Assays
Import, transform, edit, analyze and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.

Clone of S System – 검증된 시스템

S-PLUS의 Open Source



A20 = Gold

	A	B	C	D
19	Generates correlated random numbers with correct volatility			
20	Gold	Sterling	Rand	Vega
21				
22				
23	0.558155721	-0.590157338	-0.955803341	6.2
24	0.360408542	0.380029511	0.26683761	-3.73
25	-0.146007317	-0.372485517	-0.951785293	-2.39
26	0.182931489	-1.167913874	0.68698405	-2.10
27	-0.442007814	0.655721263	2.112780423	4.1
28	-0.110347833	-0.06102556	-0.431045497	-6.1
29	-0.935481	0.05427369	-0.800484856	-6.39
30	-0.09385504	0.361872696	0.453448862	-4.19
31	-0.39170921	0.512022677	-0.947697688	-4.13
32	0.558812536	-0.187382468	1.445478168	-7.25
33	-0.32023235	-0.340569796	1.128137449	-1.63
34	0.825932474	0.747815897	0.356505451	-4.00
35	1.18015737	-1.213192577	-0.804190077	5.
36	-0.814071715	-0.202972312	-0.360987947	1.3
37	0.126161652	-0.036551566	0.23661695	-8.3
38	0.814566175	1.27040443	0.543326776	-6.2
39	-0.491622827	-0.435669187	-1.914294657	-2.08
40	-0.687933725	-0.914562567	-1.461708329	-2.40
41	0.850653064	0.80658765	2.483634876	-15.8
42	0.077919325	0.966175676	1.599419465	-3.57

```
> stats <- summary(correlated.data)
> stats
      Gold                Sterling
  Min.:-1.65455779     Min.:-1.19096187
1st Qu.:-0.23048505  1st Qu.:-0.34900402
 Median: 0.09439828   Median:-0.02269648
   Mean: 0.08202255     Mean: 0.04060629
3rd Qu.: 0.41577199  3rd Qu.: 0.34918425
   Max.: 1.48380816     Max.: 1.81821873
```

Currency simulation project

Contents of: Data

- Data
- Graphs
- Reports
- Scripts
- SearchPath

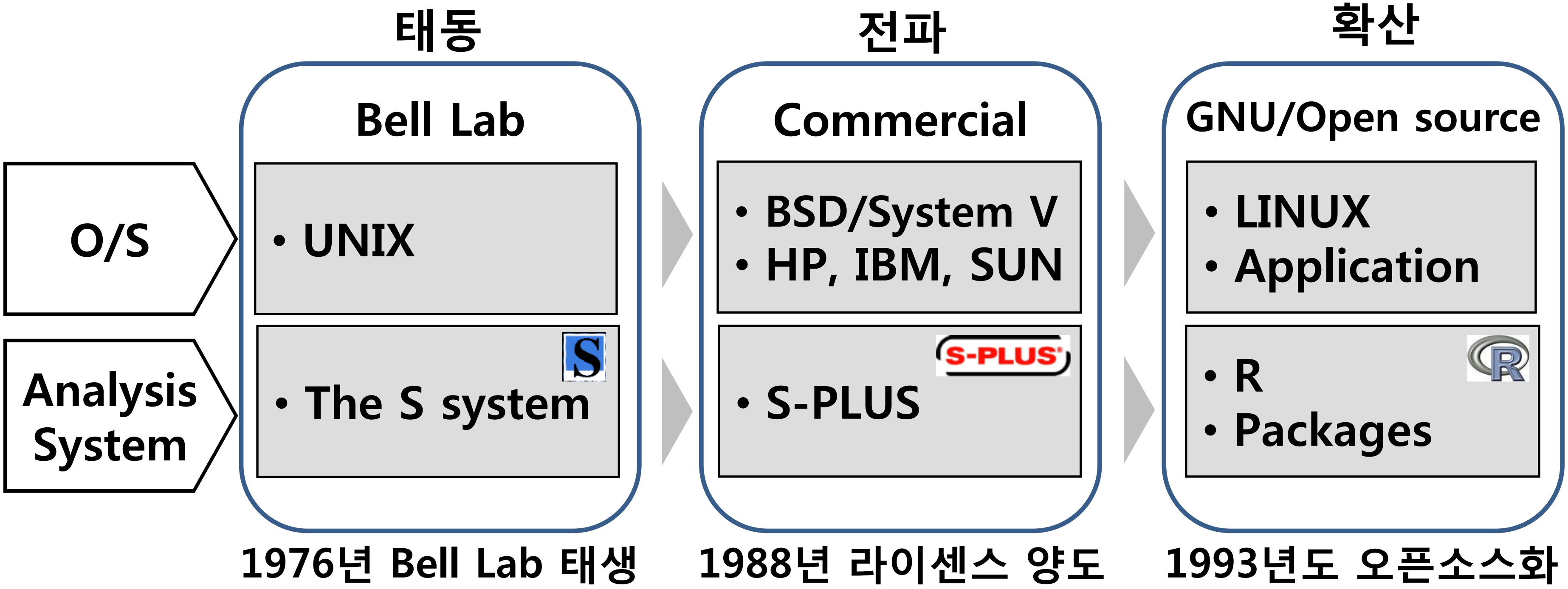
Object	Pos	Data Class
correlated.data	1	data.frame
simdata	1	data.frame
stats	1	table

Data Analytics를 위한 R의 소개

- R의 소개
- R 활용 툴

R의 소개 – R이란

R is a **language** and environment for **statistical computing** and **graphics**. It is a **GNU project** which is **similar to the S language** and environment which was developed at **Bell Laboratories** (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for **S runs unaltered under R**.



	태동	전파	확산
	Bell Lab	Commercial	GNU/Open source
O/S	• UNIX	• BSD/System V • HP, IBM, SUN	• LINUX • Application
Analysis System	• The S system	• S-PLUS	• R • Packages
	1976년 Bell Lab 태생	1988년 라이센스 양도	1993년도 오픈소스화

R의 소개 – 장단점

R ?

- **Free** Analytics Software

Free ?

- **분석의 자유**
  - 생각하는 분석 기법은 모두 지원한다. (4,000여 개 이상의 패키지)
  - 최신 분석 기법의 제공 및 자유로운 분석 환경 제공
- **배포의 자유**
  - 자유로운 "실행, 복사, 수정, 배포 " 의 권리를 갖는 사용 허가권
- **비용의 자유**
  - 무료 소프트웨어 (단, 소프트웨어 업체의 R을 이용한 저작물은 비용 발생 가능)

[R의 장점 및 단점]

구분	장점	단점	비고
In-Memory 구조	연산 수행 속도 빠름	대용량 데이터 분석 불가	상용 R 시스템
Open Source	• 저렴한 비용 • 시스템 통합 용이	교육, 기술지원 지원 부족	시장 형성기
Language 구조	• 알고리즘 구현 용이 • Detail 분석 가능	프로그램 능력이 필요함	S3, S4 Spec

R의 소개 – statistical computing

주요 통계계산 기능	통계량/기초통계	• EDA(Exploratory Data Analysis) • Summary
	통계분석	• 전통적인 통계분석 방법론 • 최신 통계분석 방법론, Spatial, Bayesian 통계 등
	마이닝 분석	• Decision Tree, SVM, Clustering, … • WEKA interface
	시뮬레이션	• 모형 시뮬레이션 • Operation Research
	수치해석	• 미분, 적분, 행렬대수 • 근사값 계산, Optimization
교육	대학/대학원 교육	• 대학 및 대학원에서의 통계 교육의 표준으로 사용
업계의 활용	분석업무 활용	• Google : Google Analytics(SaaS)에 R을 사용 • Facebook, Yahoo 등 회사에서 내부 분석용 도구로 활용
	제품 개발	• Oracle, Teradata, EMC 등 업체의 DBMS 내 분석툴로 제공
활용 프로젝트	Bioinformatics 프로젝트	• BioConductor Project – 460 이상의 Packages • 게놈, Bio, 신약연구 등 • Bioinformatics의 표준 통계분석 언어
	Finmatrics 프로젝트	• 금융 예측분석에 사용, 여러 가지 금융 예측모형 구현

R의 소개 – statistical computing

통계계산 최적화 사례 - 회귀분석

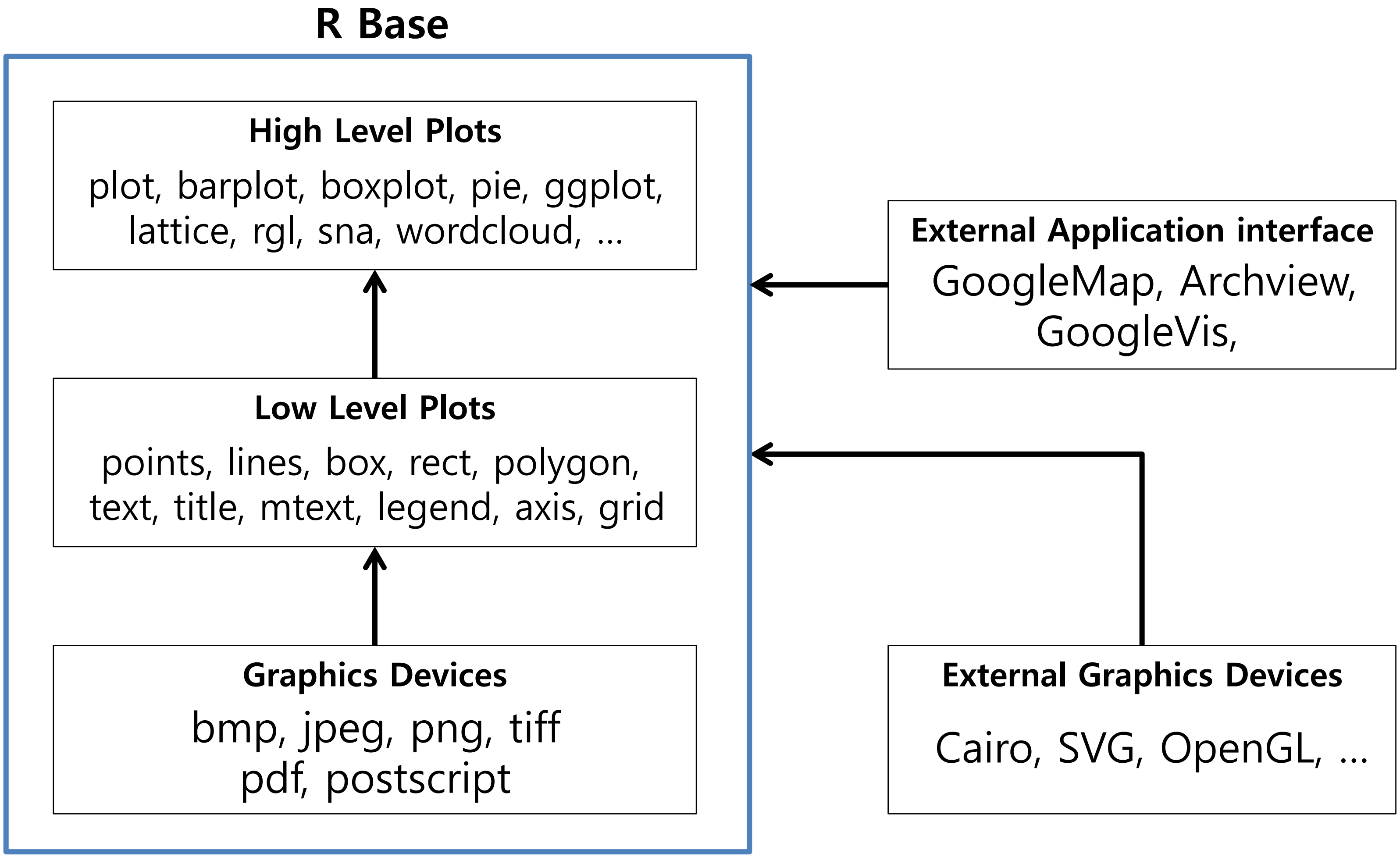
```
> stack.loss[1:6]
[1] 42 37 37 28 18 18
> X <- cbind(1,stack.x)
> head(X)
       Air.Flow Water.Temp Acid.Conc.
[1,] 1      80         27         89
[2,] 1      80         27         88
[3,] 1      75         25         90
[4,] 1      62         24         87
[5,] 1      62         22         87
[6,] 1      62         23         87
> solve(t(X) %*% X) %*% t(X) %*% stack.loss
                         [,1]
                  -39.9196744
Air.Flow            0.7156402
Water.Temp         1.2952861
Acid.Conc.        -0.1521225
> lm(stack.loss ~ stack.x)

Call:
lm(formula = stack.loss ~ stack.x)

Coefficients:
    (Intercept)    stack.xAir.Flow   stack.xWater.Temp   stack.xAcid.Conc.
       -39.9197             0.7156              1.2953             -0.1521
```

"행렬/벡터 데이터 타입 지원" 과
"행렬 연산 지원" 으로
"복잡한 구조의 반복문 제거"
"코드를 이해가 쉬움"

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

R의 소개 – graphics

Hierarchical architecture



R Base

High Level Plots

plot, barplot, boxplot, pie, ggplot, lattice, rgl, sna, wordcloud, …

Low Level Plots

points, lines, box, rect, polygon, text, title, mtext, legend, axis, grid

Graphics Devices

bmp, jpeg, png, tiff pdf, postscript

External Application interface

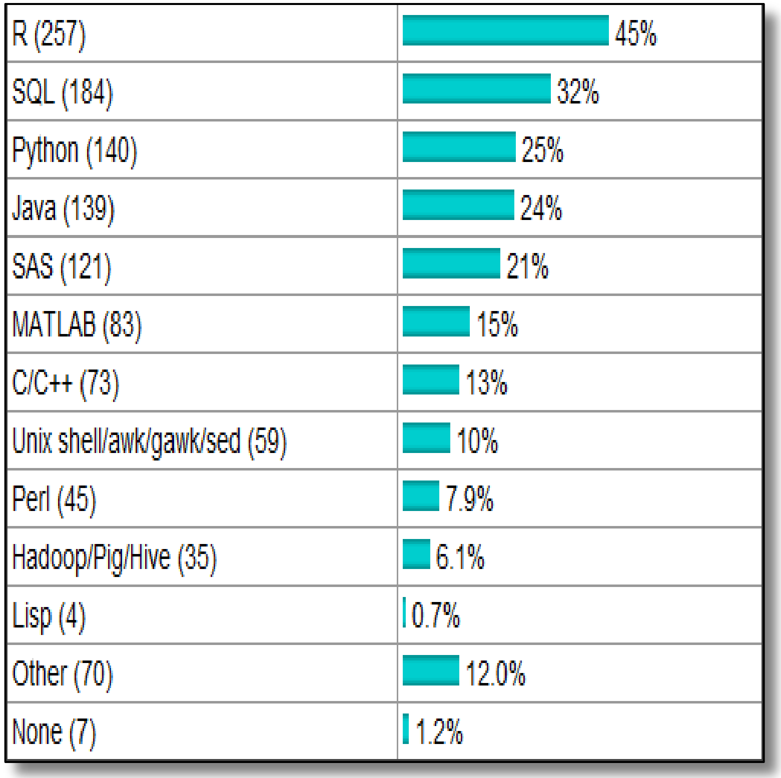
GoogleMap, Archview, GoogleVis,

External Graphics Devices

Cairo, SVG, OpenGL, …

R의 소개 – Populations

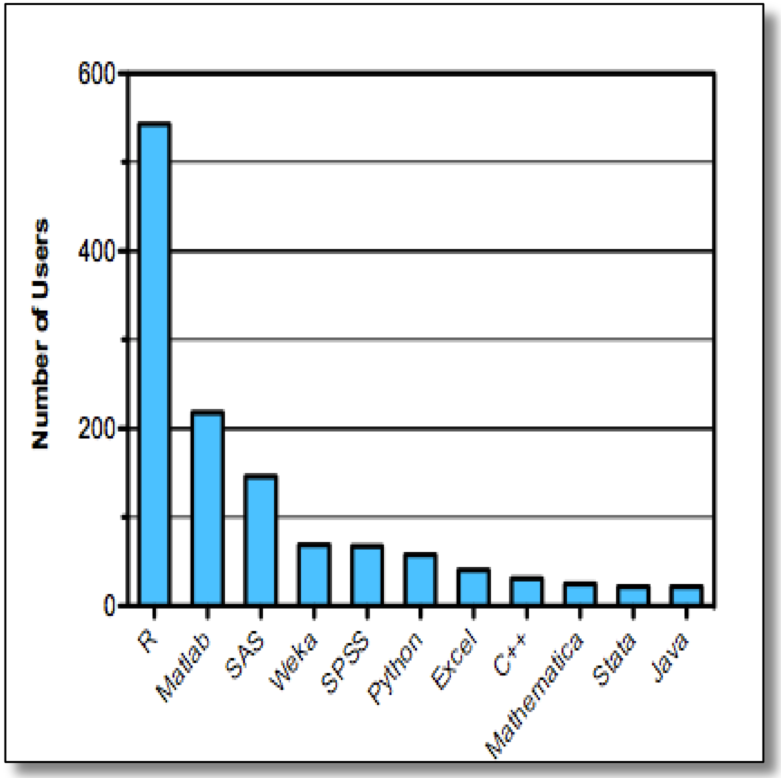
Kdnugget Poll : Language for DM



R (257)	45%
SQL (184)	32%
Python (140)	25%
Java (139)	24%
SAS (121)	21%
MATLAB (83)	15%
C/C++ (73)	13%
Unix shell/awk/gawk/sed (59)	10%
Perl (45)	7.9%
Hadoop/Pig/Hive (35)	6.1%
Lisp (4)	0.7%
Other (70)	12.0%
None (7)	1.2%

http://www.kdnuggets.com/2011/08/poll-languages-for-data-mining-analytics.html

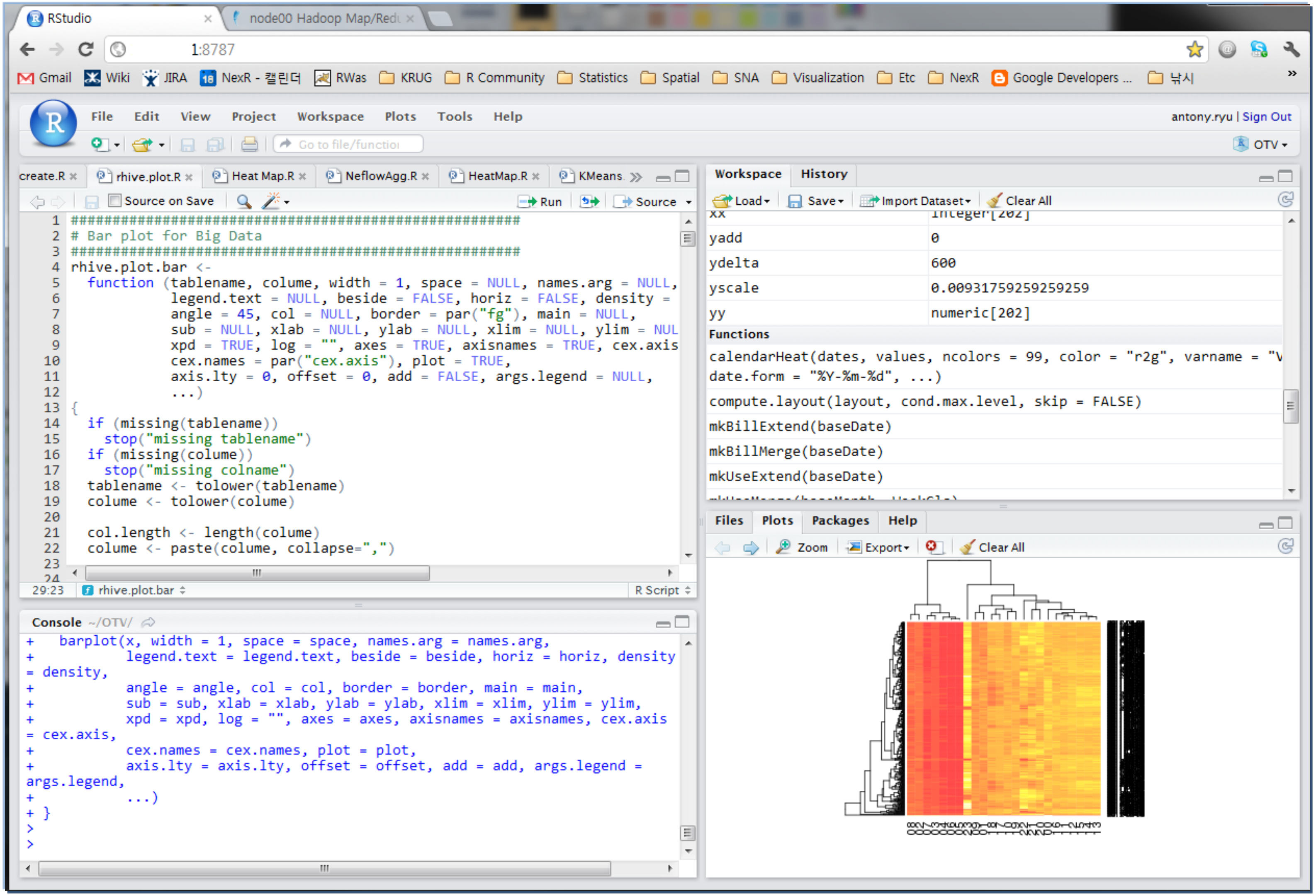
Kaggle : Tool of competitors



http://blog.revolutionanalytics.com/2011/11/r-still-the-preferred-tool-of-predictive-modelers-competing-at-kaggle.html

R 활용 툴 – IDE

- RStudio



```
################################################################
# Bar plot for Big Data
################################################################
rhive.plot.bar <-
  function (tablename, colume, width = 1, space = NULL, names.arg = NULL,
            legend.text = NULL, beside = FALSE, horiz = FALSE, density =
            angle = 45, col = NULL, border = par("fg"), main = NULL,
            sub = NULL, xlab = NULL, ylab = NULL, xlim = NULL, ylim = NUL
            xpd = TRUE, log = "", axes = TRUE, axisnames = TRUE, cex.axis
            cex.names = par("cex.axis"), plot = TRUE,
            axis.lty = 0, offset = 0, add = FALSE, args.legend = NULL,
            ...)
{
  if (missing(tablename))
    stop("missing tablename")
  if (missing(colume))
    stop("missing colname")
  tablename <- tolower(tablename)
  colume <- tolower(colume)

  col.length <- length(colume)
  colume <- paste(colume, collapse=",")

```

```
+   barplot(x, width = 1, space = space, names.arg = names.arg,
+           legend.text = legend.text, beside = beside, horiz = horiz, density
= density,
+           angle = angle, col = col, border = border, main = main,
+           sub = sub, xlab = xlab, ylab = ylab, xlim = xlim, ylim = ylim,
+           xpd = xpd, log = "", axes = axes, axisnames = axisnames, cex.axis
= cex.axis,
+           cex.names = cex.names, plot = plot,
+           axis.lty = axis.lty, offset = offset, add = add, args.legend =
args.legend,
+           ...)
+ }
>
>
```

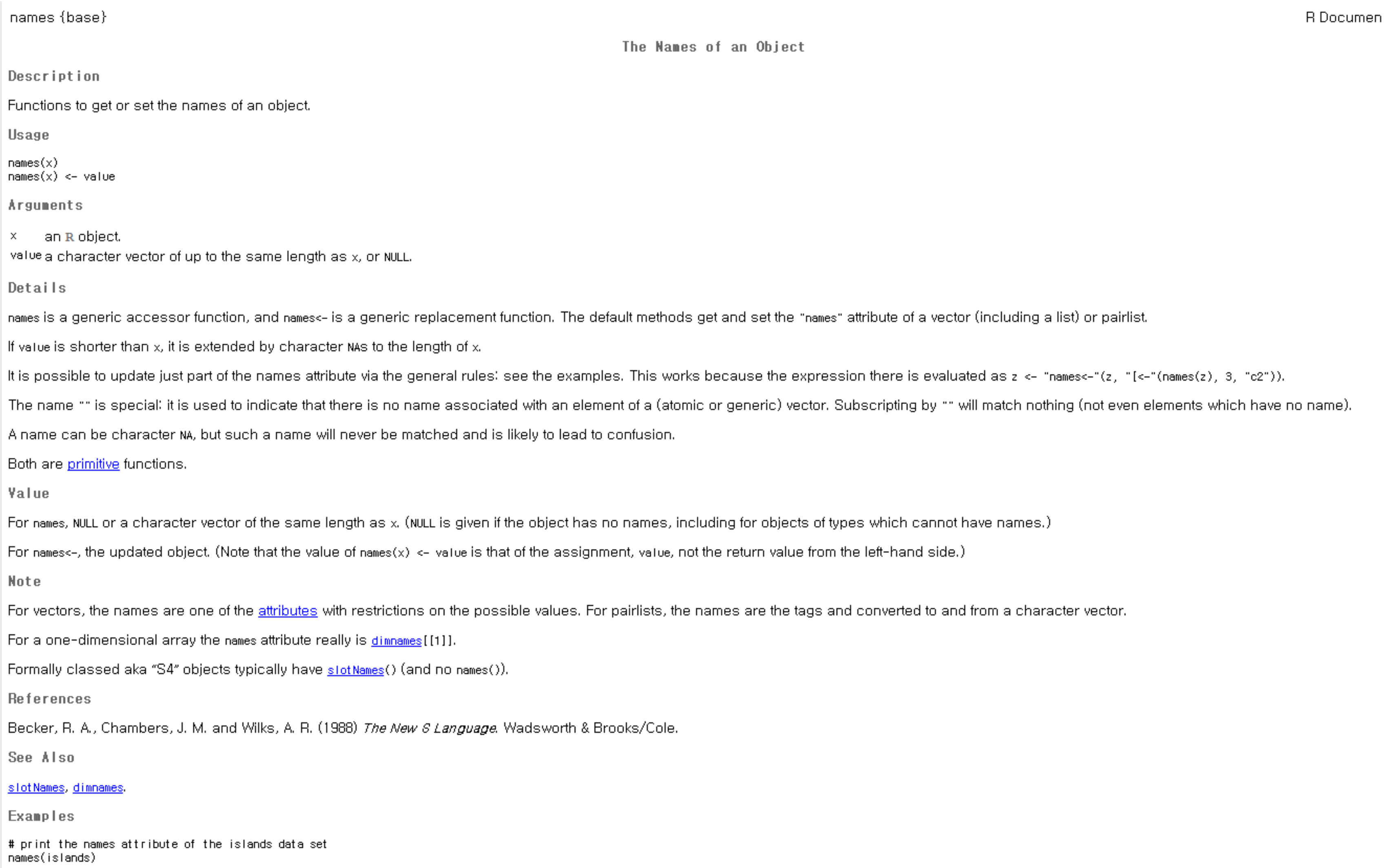
xx	integer[202]
yadd	0
ydelta	600
yscale	0.00931759259259259
yy	numeric[202]

Functions

```
calendarHeat(dates, values, ncolors = 99, color = "r2g", varname = "V
date.form = "%Y-%m-%d", ...)
compute.layout(layout, cond.max.level, skip = FALSE)
mkBillExtend(baseDate)
mkBillMerge(baseDate)
mkUseExtend(baseDate)
```

R 활용 툴 – Help

● Help Documentations



names {base} R Documen

The Names of an Object

Description

Functions to get or set the names of an object.

Usage

```
names(x)
names(x) <- value
```

Arguments

x an R object.

value a character vector of up to the same length as x, or NULL.

Details

names is a generic accessor function, and names<- is a generic replacement function. The default methods get and set the "names" attribute of a vector (including a list) or pairlist.

If value is shorter than x, it is extended by character NAs to the length of x.

It is possible to update just part of the names attribute via the general rules: see the examples. This works because the expression there is evaluated as z <- "names<-"(z, "[<-"(names(z), 3, "c2")).

The name "" is special: it is used to indicate that there is no name associated with an element of a (atomic or generic) vector. Subscripting by "" will match nothing (not even elements which have no name).

A name can be character NA, but such a name will never be matched and is likely to lead to confusion.

Both are primitive functions.

Value

For names, NULL or a character vector of the same length as x. (NULL is given if the object has no names, including for objects of types which cannot have names.)

For names<-, the updated object. (Note that the value of names(x) <- value is that of the assignment, value, not the return value from the left-hand side.)

Note

For vectors, the names are one of the attributes with restrictions on the possible values. For pairlists, the names are the tags and converted to and from a character vector.

For a one-dimensional array the names attribute really is dimnames[[1]].

Formally classed aka "S4" objects typically have slotNames() (and no names()).

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

See Also

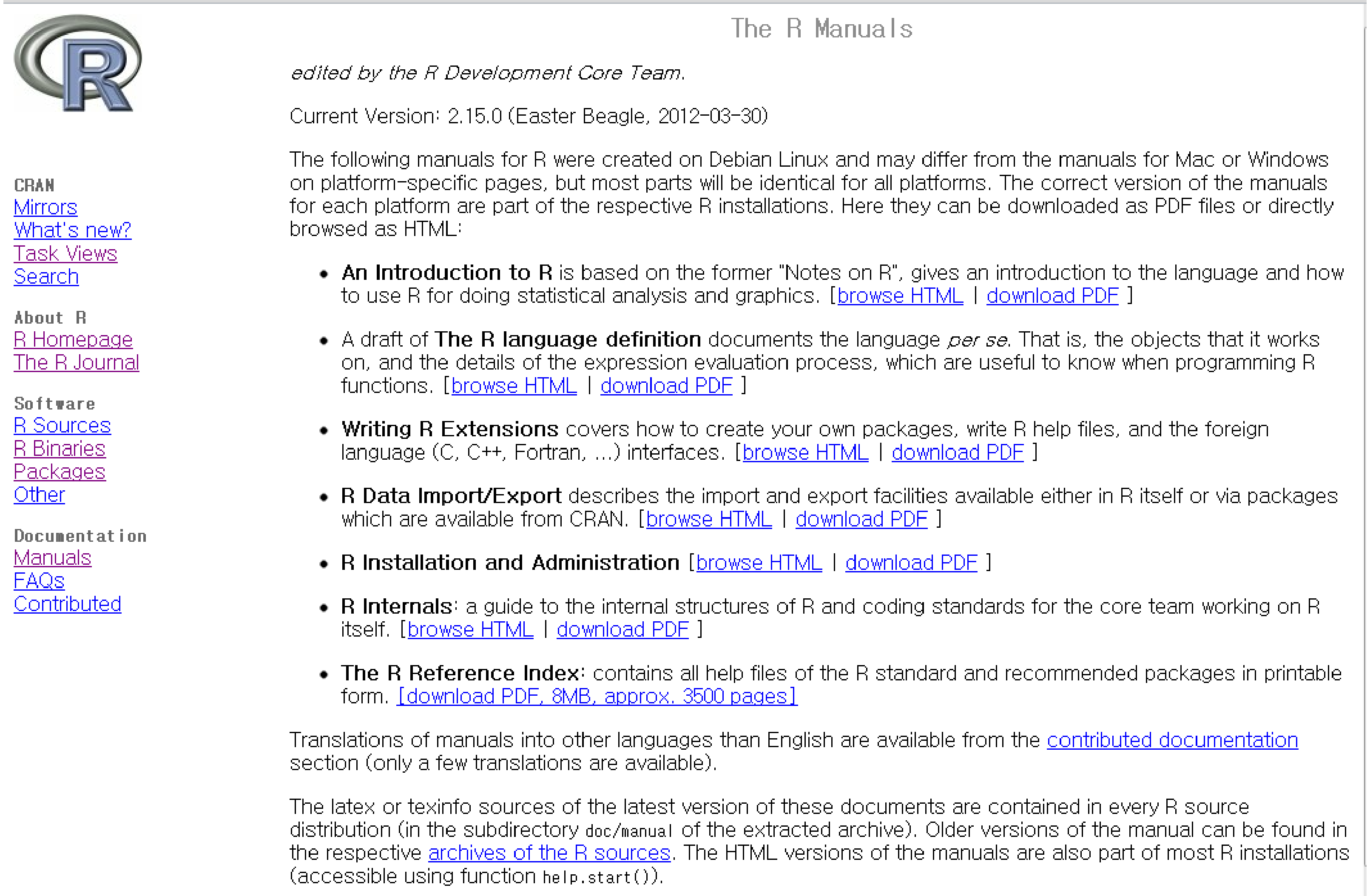
slotNames, dimnames.

Examples

```
# print the names attribute of the islands data set
names(islands)
```

R 활용 툴 – Manuals

- R Manuals (http://cran.nexr.com/manuals.html)



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The R Manuals

edited by the R Development Core Team.

Current Version: 2.15.0 (Easter Beagle, 2012-03-30)

The following manuals for R were created on Debian Linux and may differ from the manuals for Mac or Windows on platform-specific pages, but most parts will be identical for all platforms. The correct version of the manuals for each platform are part of the respective R installations. Here they can be downloaded as PDF files or directly browsed as HTML:

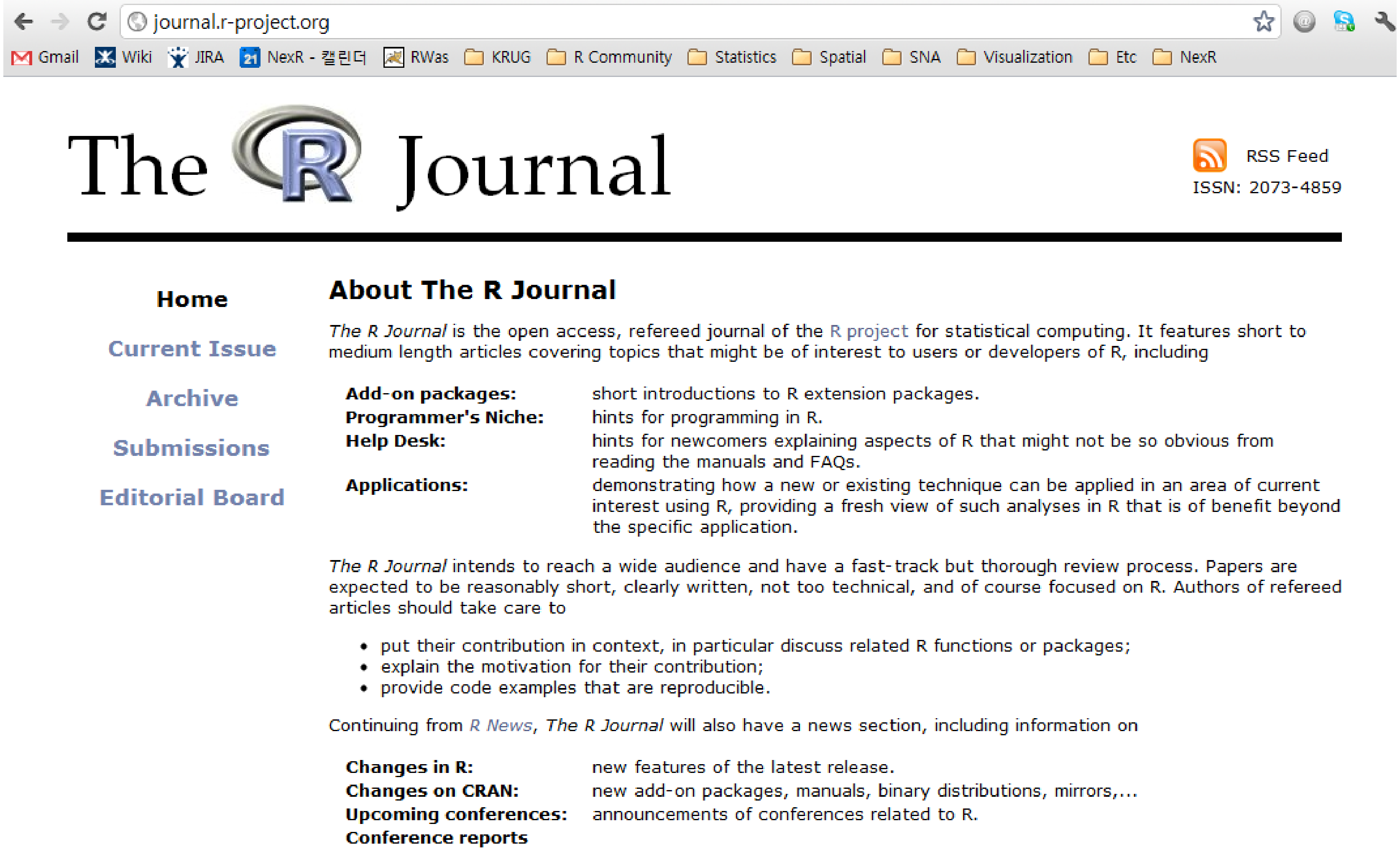
- **An Introduction to R** is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics. [browse HTML | download PDF]
- A draft of **The R language definition** documents the language *per se*. That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions. [browse HTML | download PDF]
- **Writing R Extensions** covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces. [browse HTML | download PDF]
- **R Data Import/Export** describes the import and export facilities available either in R itself or via packages which are available from CRAN. [browse HTML | download PDF]
- **R Installation and Administration** [browse HTML | download PDF]
- **R Internals**: a guide to the internal structures of R and coding standards for the core team working on R itself. [browse HTML | download PDF]
- **The R Reference Index**: contains all help files of the R standard and recommended packages in printable form. [download PDF, 8MB, approx. 3500 pages]

Translations of manuals into other languages than English are available from the contributed documentation section (only a few translations are available).

The latex or texinfo sources of the latest version of these documents are contained in every R source distribution (in the subdirectory `doc/manual` of the extracted archive). Older versions of the manual can be found in the respective archives of the R sources. The HTML versions of the manuals are also part of most R installations (accessible using function `help.start()`).

R 활용 툴 – Journal

- The R Journal (http://journal.r-project.org/)



The R Journal

RSS Feed

ISSN: 2073-4859

Home

Current Issue

Archive

Submissions

Editorial Board

About The R Journal

The R Journal is the open access, refereed journal of the R project for statistical computing. It features short to medium length articles covering topics that might be of interest to users or developers of R, including

Add-on packages:	short introductions to R extension packages.
Programmer's Niche:	hints for programming in R.
Help Desk:	hints for newcomers explaining aspects of R that might not be so obvious from reading the manuals and FAQs.
Applications:	demonstrating how a new or existing technique can be applied in an area of current interest using R, providing a fresh view of such analyses in R that is of benefit beyond the specific application.

The R Journal intends to reach a wide audience and have a fast-track but thorough review process. Papers are expected to be reasonably short, clearly written, not too technical, and of course focused on R. Authors of refereed articles should take care to

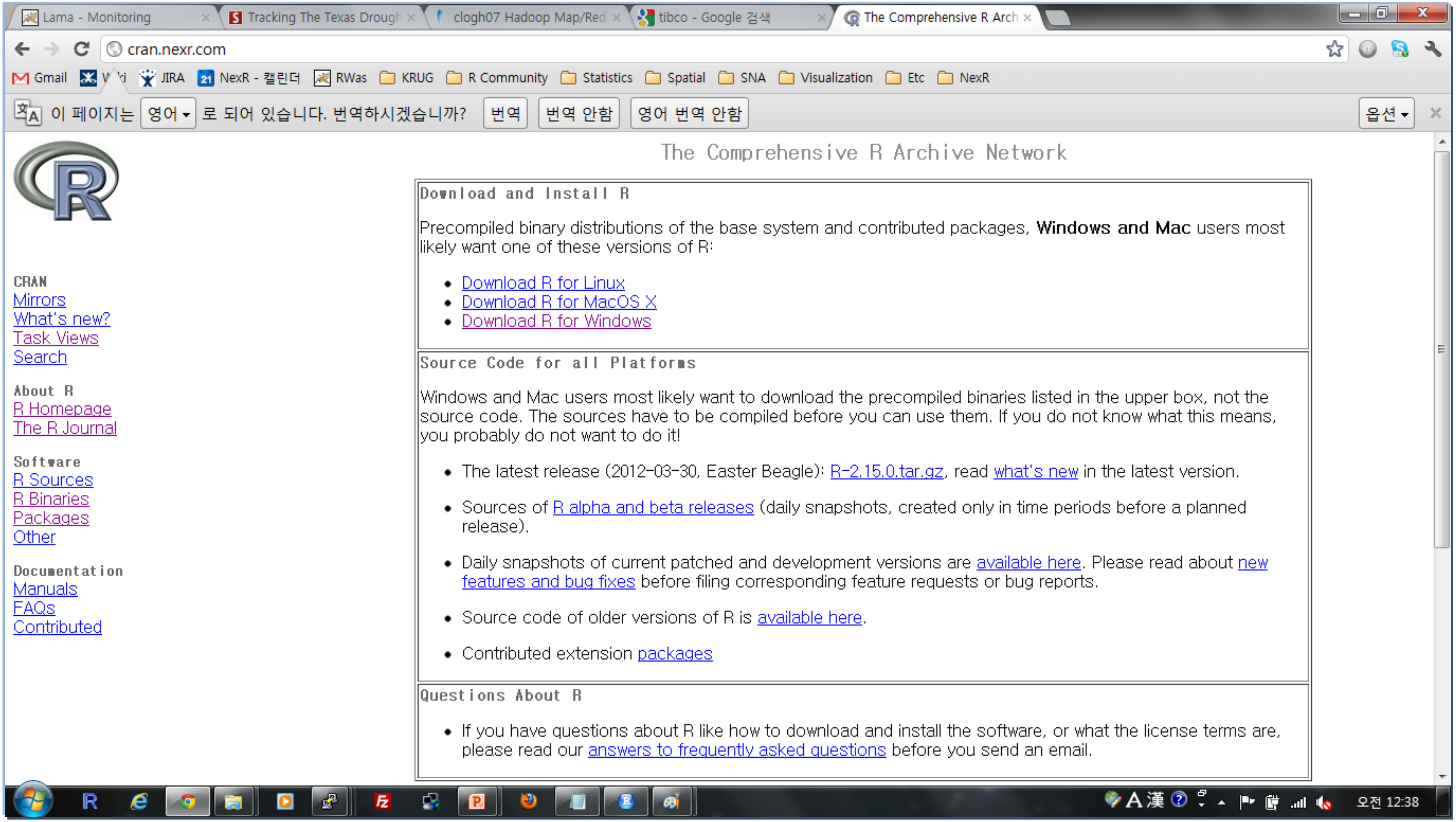
- put their contribution in context, in particular discuss related R functions or packages;
- explain the motivation for their contribution;
- provide code examples that are reproducible.

Continuing from *R News*, *The R Journal* will also have a news section, including information on

Changes in R:	new features of the latest release.
Changes on CRAN:	new add-on packages, manuals, binary distributions, mirrors,...
Upcoming conferences:	announcements of conferences related to R.
Conference reports	

R 활용 툴 – 검색

- CRAN (The Comprehensive R Archive Network)



The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

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!

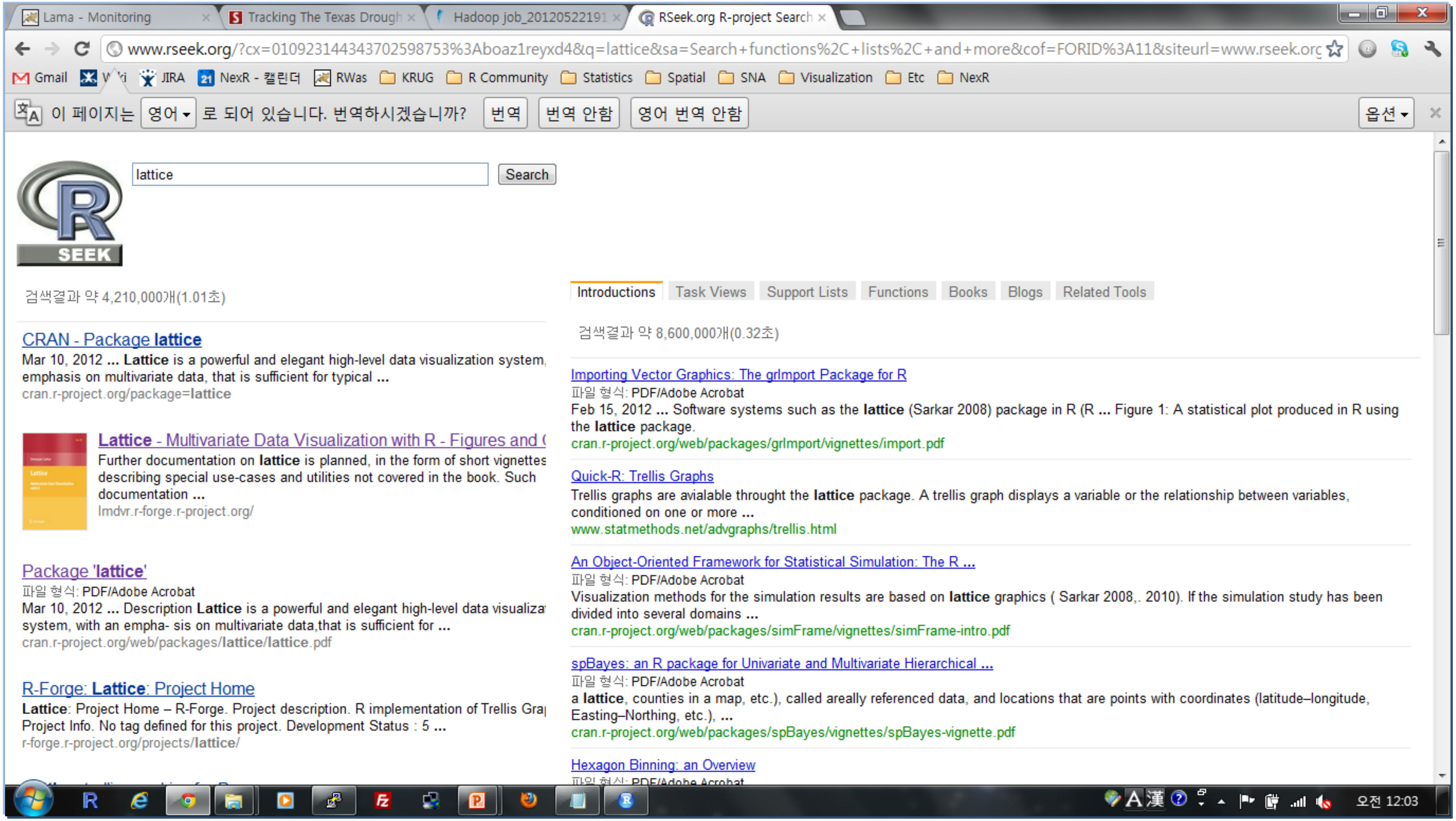
- The latest release (2012-03-30, Easter Beagle): R-2.15.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. 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.

R 활용 툴 – 검색

- Rseek (www.rseek.org)



검색결과 약 4,210,000개(1.01초)

CRAN - Package **lattice**
Mar 10, 2012 ... **Lattice** is a powerful and elegant high-level data visualization system, emphasis on multivariate data, that is sufficient for typical ...
cran.r-project.org/package=**lattice**

Lattice - Multivariate Data Visualization with R - Figures and C
Further documentation on **lattice** is planned, in the form of short vignettes describing special use-cases and utilities not covered in the book. Such documentation ...
lmdvr.r-forge.r-project.org/

Package '**lattice**'
파일 형식: PDF/Adobe Acrobat
Mar 10, 2012 ... Description **Lattice** is a powerful and elegant high-level data visualiza system, with an empha- sis on multivariate data,that is sufficient for ...
cran.r-project.org/web/packages/**lattice**/**lattice**.pdf

R-Forge: **Lattice**: Project Home
Lattice: Project Home – R-Forge. Project description. R implementation of Trellis Gra
Project Info. No tag defined for this project. Development Status : 5 ...
r-forge.r-project.org/projects/**lattice**/

검색결과 약 8,600,000개(0.32초)

Importing Vector Graphics: The grImport Package for R
파일 형식: PDF/Adobe Acrobat
Feb 15, 2012 ... Software systems such as the **lattice** (Sarkar 2008) package in R (R ... Figure 1: A statistical plot produced in R using the **lattice** package.
cran.r-project.org/web/packages/grImport/vignettes/import.pdf

Quick-R: Trellis Graphs
Trellis graphs are avialable throught the **lattice** package. A trellis graph displays a variable or the relationship between variables, conditioned on one or more ...
www.statmethods.net/advgraphs/trellis.html

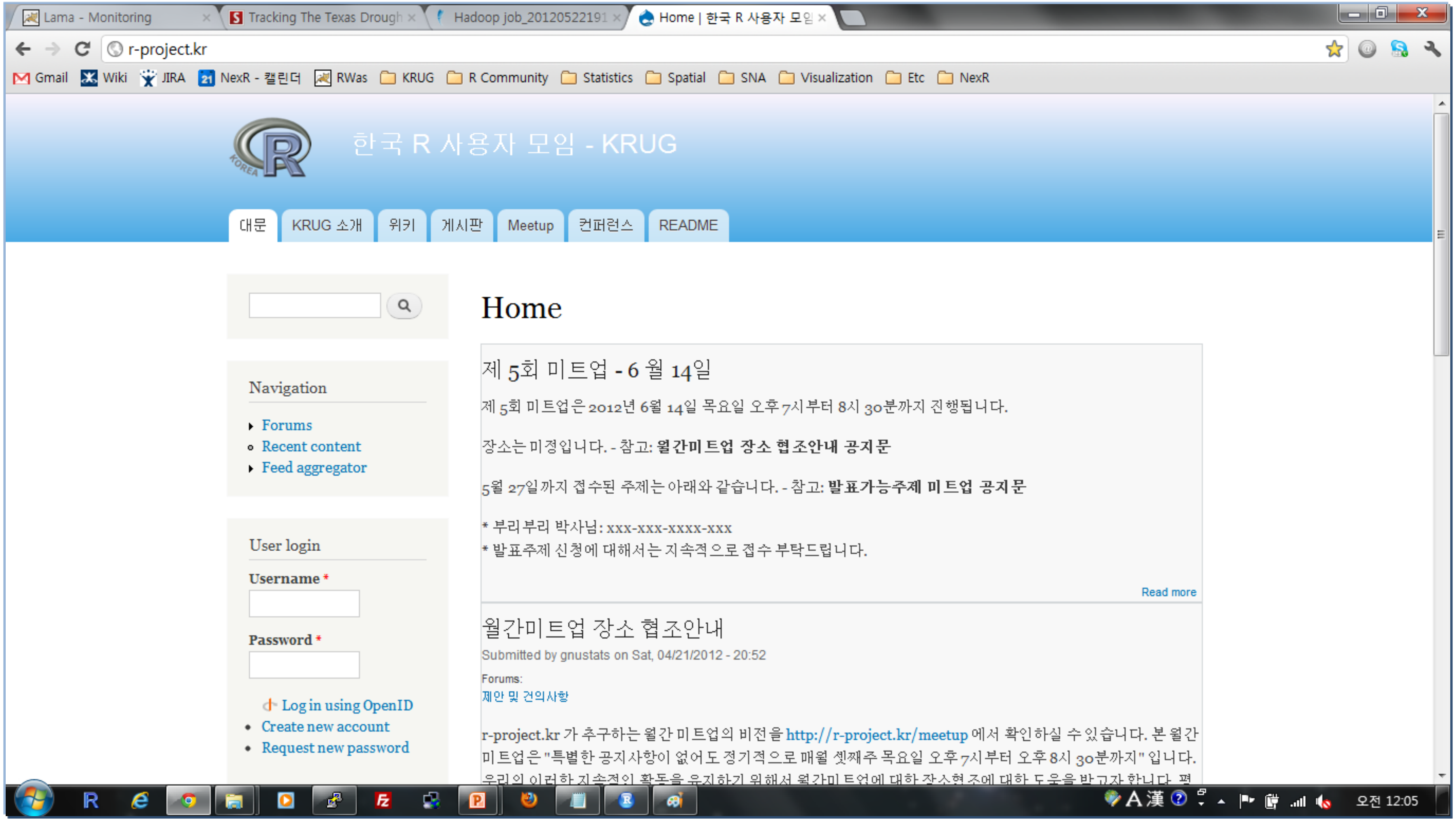
An Object-Oriented Framework for Statistical Simulation: The R ...
파일 형식: PDF/Adobe Acrobat
Visualization methods for the simulation results are based on **lattice** graphics (Sarkar 2008,. 2010). If the simulation study has been divided into several domains ...
cran.r-project.org/web/packages/simFrame/vignettes/simFrame-intro.pdf

spBayes: an R package for Univariate and Multivariate Hierarchical ...
파일 형식: PDF/Adobe Acrobat
a **lattice**, counties in a map, etc.), called areally referenced data, and locations that are points with coordinates (latitude–longitude, Easting–Northing, etc.), ...
cran.r-project.org/web/packages/spBayes/vignettes/spBayes-vignette.pdf

Hexagon Binning: an Overview
파일 형식: PDF/Adobe Acrobat

R 활용 툴 – User Group

- KRUG (www.r-project.kr)



한국 R 사용자 모임 - KRUG

Home

Navigation

- Forums
- Recent content
- Feed aggregator

User login

Username *

Password *

Log in using OpenID

- Create new account
- Request new password

제 5회 미트업 - 6 월 14일

제 5회 미트업은 2012년 6월 14일 목요일 오후7시부터 8시 30분까지 진행됩니다.

장소는 미정입니다. - 참고: **월간미트업 장소 협조안내 공지문**

5월 27일까지 접수된 주제는 아래와 같습니다. - 참고: **발표가능주제 미트업 공지문**

* 부리부리 박사님: xxx-xxx-xxxx-xxx
* 발표주제 신청에 대해서는 지속적으로 접수 부탁드립니다.

Read more

월간미트업 장소 협조안내

Submitted by gnustats on Sat, 04/21/2012 - 20:52

Forums:
제안 및 건의사항

r-project.kr 가 추구하는 월간 미트업의 비전을 http://r-project.kr/meetup 에서 확인하실 수 있습니다. 본 월간 미트업은 "특별한 공지사항이 없어도 정기적으로 매월 셋째주 목요일 오후7시부터 오후8시 30분까지" 입니다. 우리의 이러한 지속적인 활동을 유지하기 위해서 월간미트업에 대한 장소협조에 대한 도움을 받고자 합니다. 편

R을 이용한 데이터 분석의 비교

- Small Data Analytics using Native R
- Large Data Analytics using R
- Big Data Analytics using R

Small Data Analytics using Native R

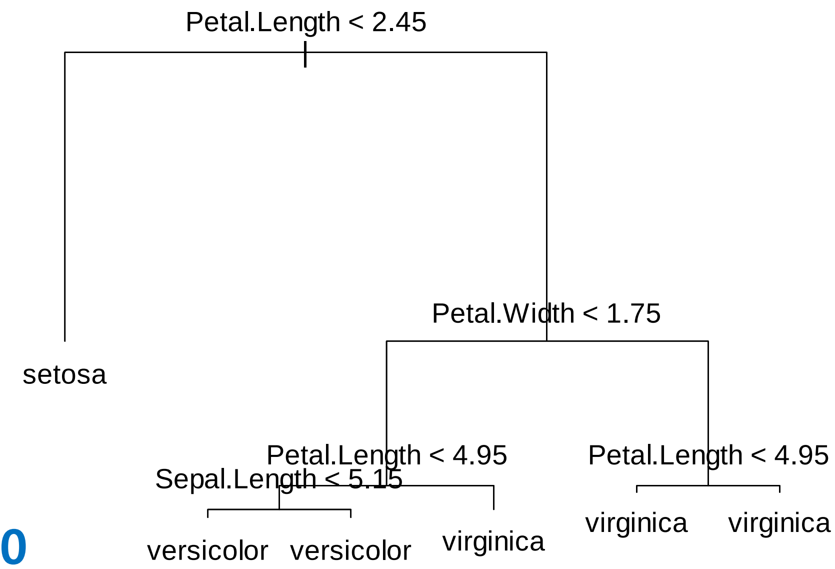
In-Memory

- Classification Tree Model
- iris : 150건, 5개 변수

- R Script

```
> library(tree)
> ir.tr <- tree(Species ~., iris)
> summary(ir.tr)
Classification tree:
tree(formula = Species ~ ., data = iris)
Variables actually used in tree construction:
[1] "Petal.Length" "Petal.Width"  "Sepal.Length"
Number of terminal nodes:  6
Residual mean deviance:  0.1253 = 18.05 / 144
Misclassification error rate: 0.02667 = 4 / 150
> plot(ir.tr)
> text(ir.tr)
```

- Tree Chart

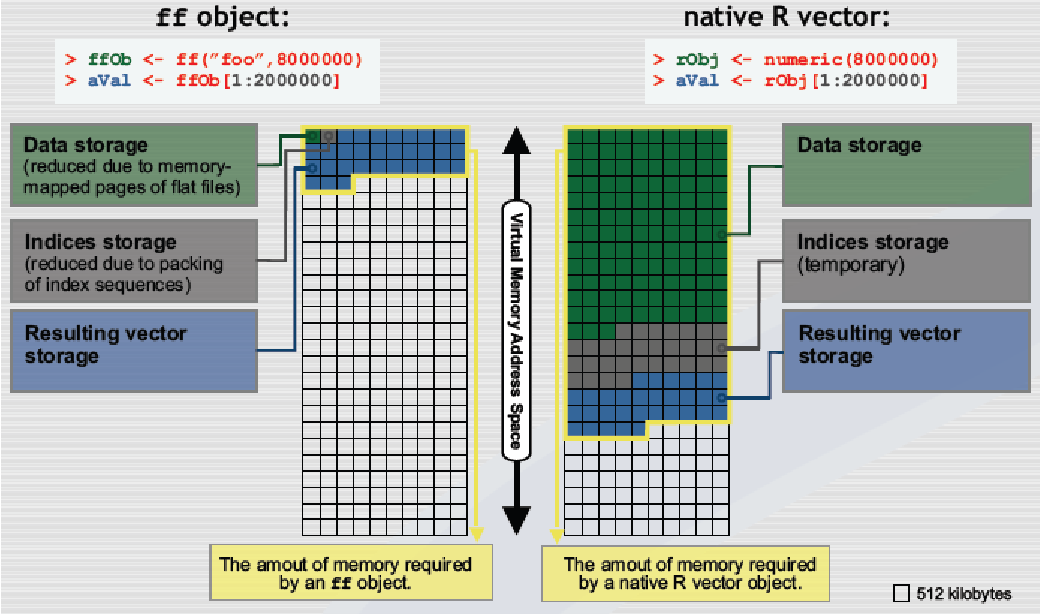


Large Data Analytics using R

In-Disk / Memory Index

- Data를 Disk에 Load
- 메모리에는 Disk의 Data영역 Index 정보가 올라감
- Data를 Loading하는 작업 필요, 별도의 분석 라이브러리 개발 필요

[개념도 (ff Package 예시)]


ff object:
native R vector:
Data storage
(reduced due to memory-mapped pages of flat files)
Indices storage
(reduced due to packing of index sequences)
Resulting vector storage
Virtual Memory Address Space
Data storage
Indices storage
(temporary)
Resulting vector storage
The amout of memory required by an ff object.
The amout of memory required by a native R vector object.
□ 512 kilobytes

```
> ffOb <- ff("foo",8000000)
> aVal <- ffOb[1:2000000]
```

```
> rObj <- numeric(8000000)
> aVal <- rObj[1:2000000]
```

[대표적인 Packages]

Package 명	비고
bigmemory	분석용 Package (biganalytics)
ff	분석용 Package (ffbase)
RevoScaleR	상용 (Revolution Analytics 사)

Large Data Analytics using R

bigmemory Example Script

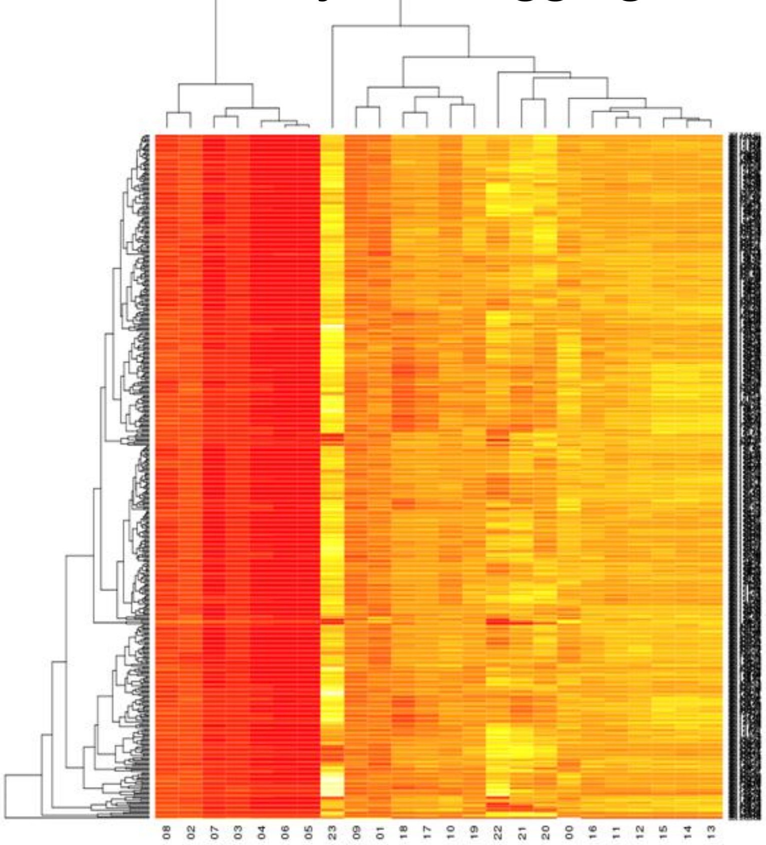
- airline : 123,534,959건, 29개 변수, 11GB
- 29개 변수의 산술평균 구하기
- Ubuntu linux 64Bit/ i7(dual) / 8G (Notebook)

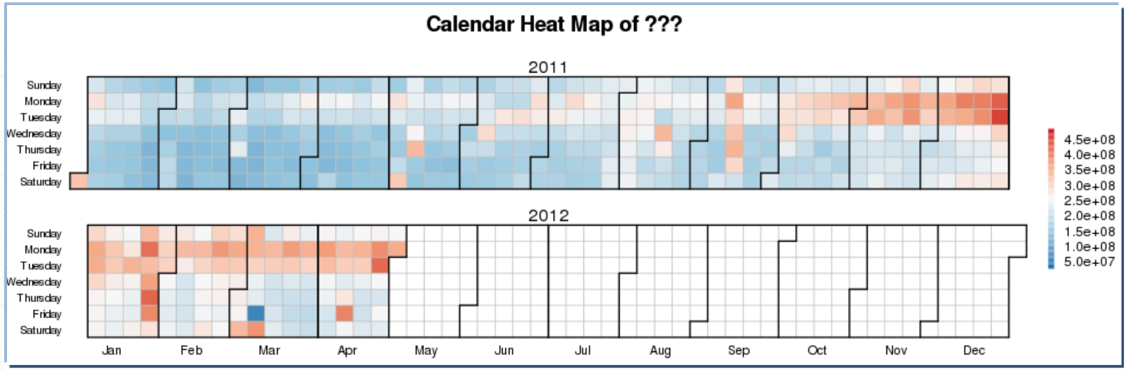
```
> library(bigmemory)
> airline <- read.big.matrix("/home/antony/anal/airline.csv", header=T,
+ backingfile="airline.bin",  descriptorfile="airline.desc", type="integer",
+ backingpath="/home/antony/anal/back/")

> library(biganalytics)
> colmean(airline, na.rm=T)
     …
  WeatherDelay          NASDelay     SecurityDelay   LateAircraftDelay
  7.883406e-01      4.103548e+00      2.670679e-02        4.756176e+00
```

작업	수행속도
데이터 로드	33m 17s
산술평균	2m 38s

RHive - Visualization

- Visualization으로 Long-Tail 파악 한다.
- XX 데이터 (2011-01-01~2012-04-30, 16개월 로그데이터)
- hiveQuery 함수, aggregate 함수, heatmap 함수 이용

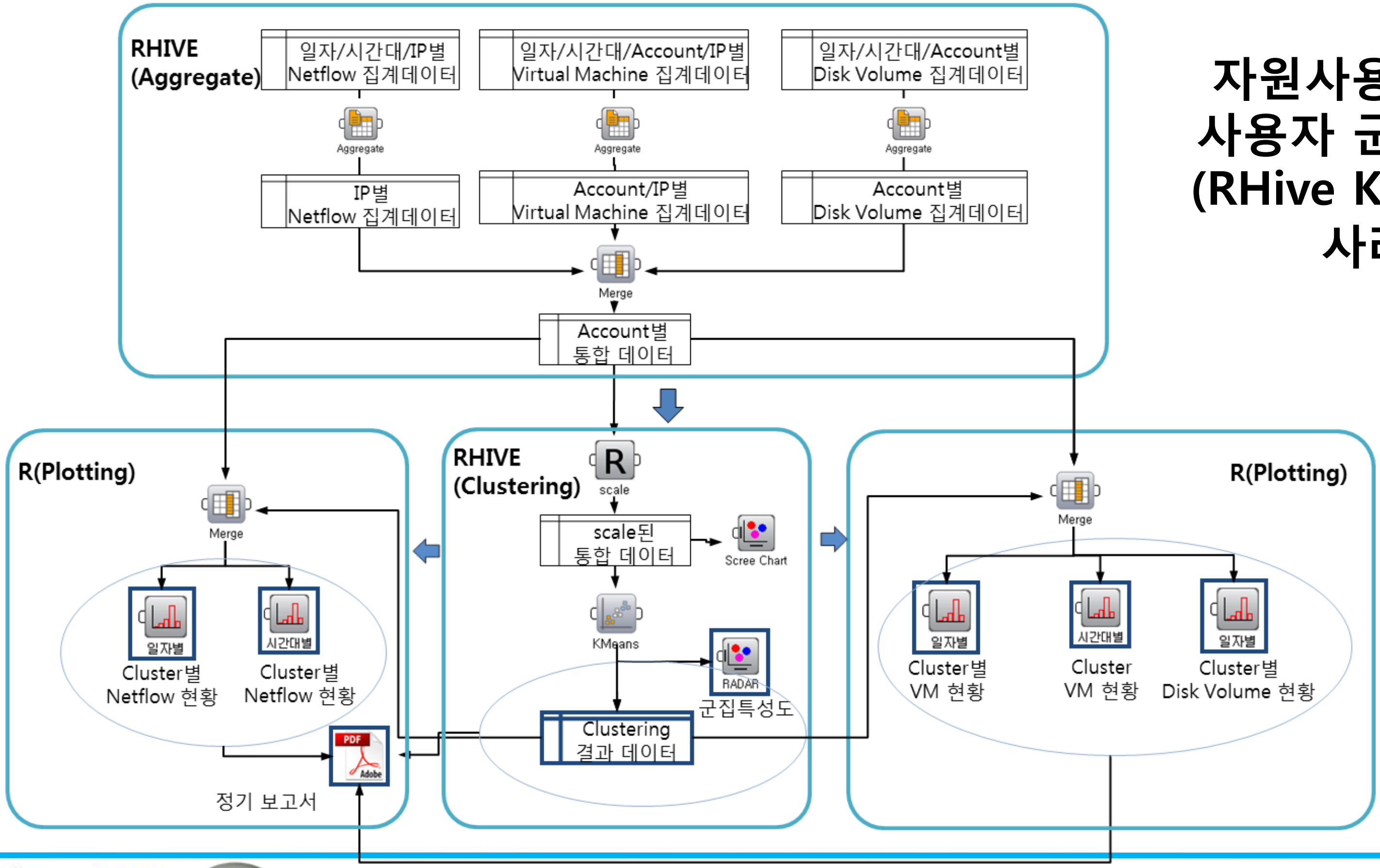




Big Data Analytics using R

RHive – Enterprise Analytics

- RHive를 이용한 KT Cloud 로그분석의 사례
- Cloud 시스템 운영에 필요한 모니터링 정보 제공



RHIVE (Aggregate)

일자/시간대/IP별 Netflow 집계데이터

일자/시간대/Account/IP별 Virtual Machine 집계데이터

일자/시간대/Account별 Disk Volume 집계데이터

Aggregate

IP별 Netflow 집계데이터

Account/IP별 Virtual Machine 집계데이터

Account별 Disk Volume 집계데이터

Merge

Account별 통합 데이터

R(Plotting)

Merge

일자별 Cluster별 Netflow 현황

시간대별 Cluster별 Netflow 현황

PDF Adobe

정기 보고서

RHIVE (Clustering)

R scale

scale된 통합 데이터

Scree Chart

KMeans

RADAR 군집특성도

Clustering 결과 데이터

R(Plotting)

Merge

일자별 Cluster별 VM 현황

시간대별 Cluster VM 현황

일자별 Cluster별 Disk Volume 현황

자원사용 기반 사용자 군집분석 (RHive KMeans) 사례

Visualization

- Visualization의 필요성
- EDA
- Special Chart
- Big Data Analytics

Visualization의 필요성 Anscombe – regression

원천 데이터

Anscombe, Francis J. - American Statistician - **“Graphs in statistical analysis”** - 1973

관측수	I		II		III		IV	
	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Visualization의 필요성 Anscombe – regression

통계량 및 단순회귀분석

통계량

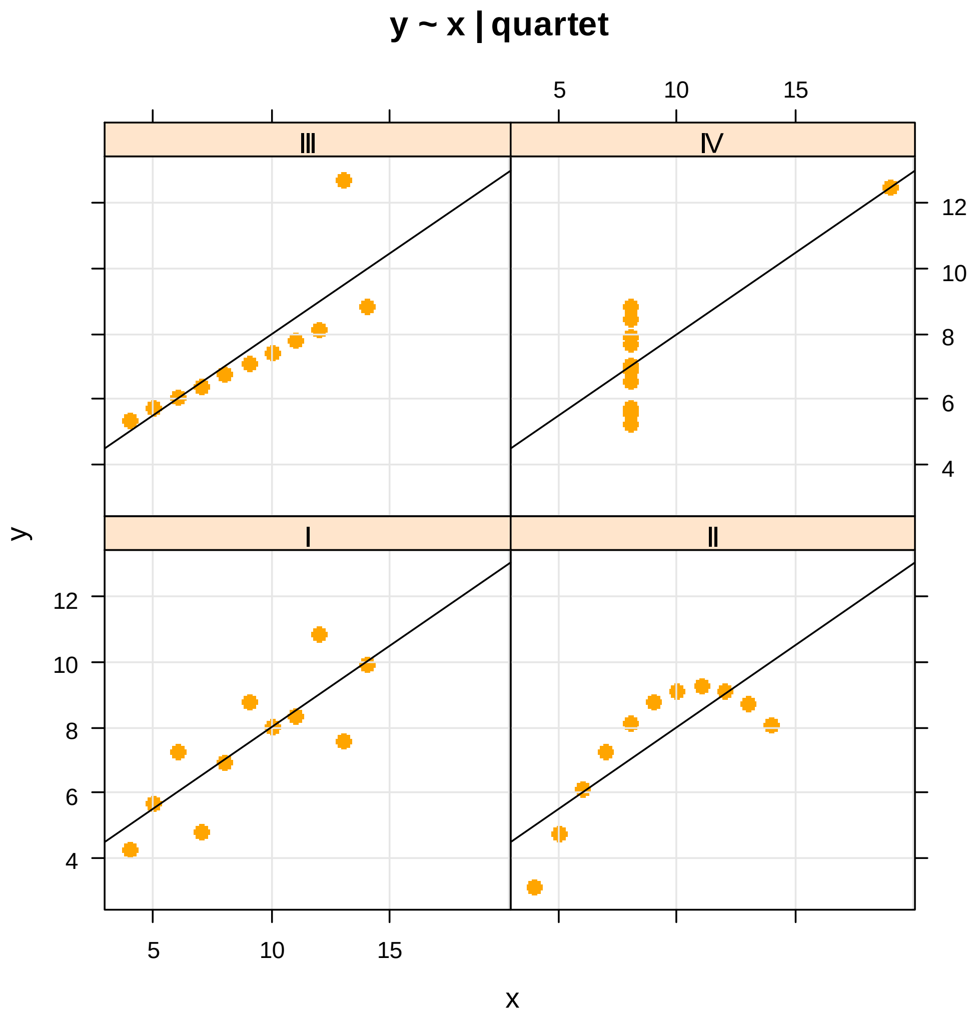
지표	Ⅰ		Ⅱ		Ⅲ		Ⅳ	
	x1	y1	x2	y2	x3	y3	x4	y4
평균	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
분산	11	4.1273	11	4.1276	11	4.1226	11	4.1232
상관계수	0.8164205		0.8162365		0.8162867		0.8165214	

단순회귀분석

지표	Ⅰ	Ⅱ	Ⅲ	Ⅳ
Coefficient Intercept	3.0	3.0	3.0	3.0
Coefficient x	0.5	0.5	0.5	0.5
Regression sum of squares	27.51	27.50	27.47	27.49
Residuals sum of squares	13.76	13.78	13.76	13.74
Estimated standard error of b1	0.12	0.12	0.12	0.12
Multiple R-square	0.67	0.67	0.67	0.67

Visualization의 필요성 Anscombe – regression

Visualization



통계량 및 회귀계수 등의 수치는 동일

$$\hat{y} = 3.0 + 0.5x$$

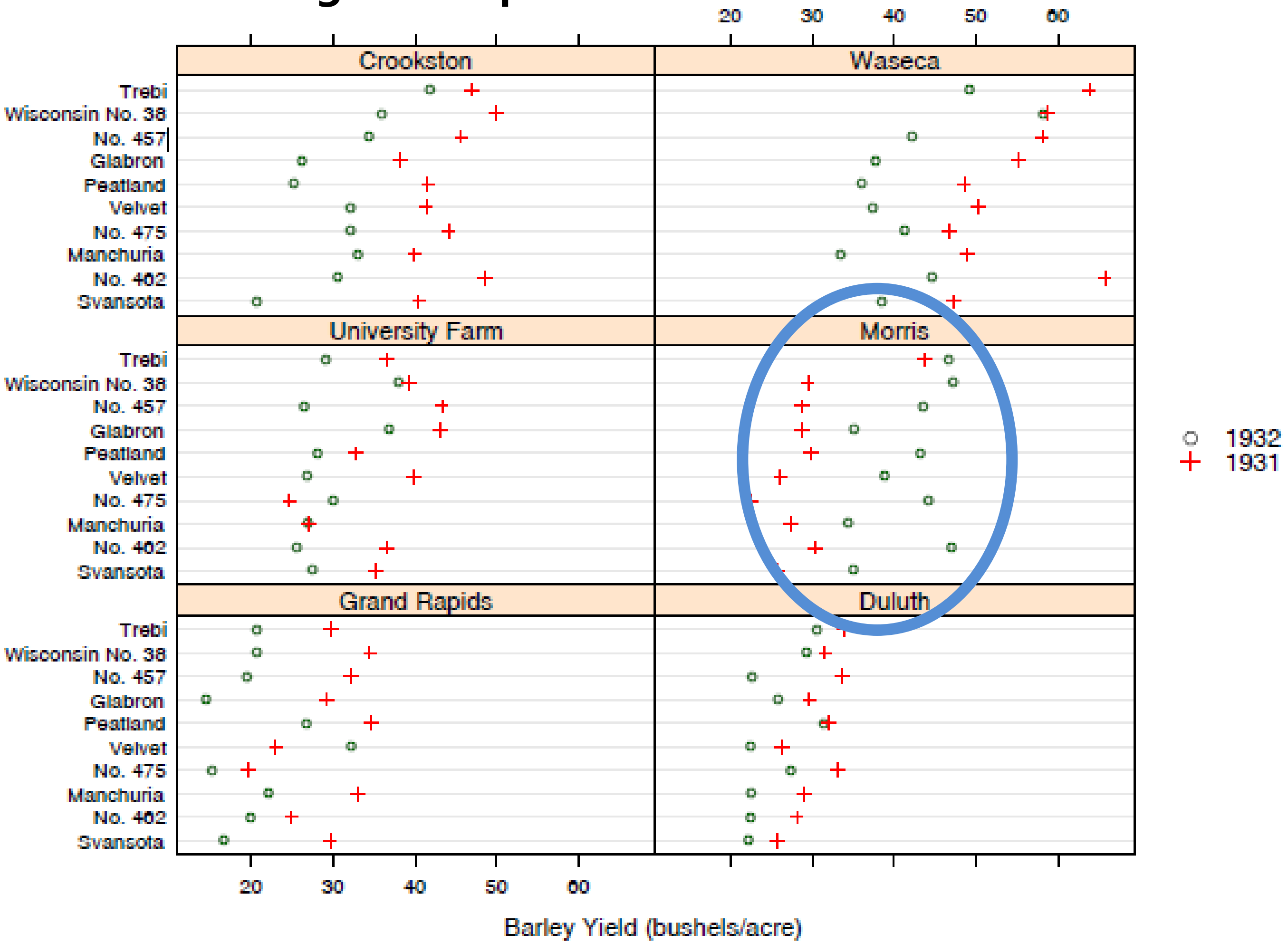
산점도 상의 네 데이터 분포는 상이

百數以不如一畵

Visualization의 필요성 Barley Yields

Multivariate Visualization

R. A. Fisher's "The design of experiments" 1930s ~ 1990s 인용된 자료



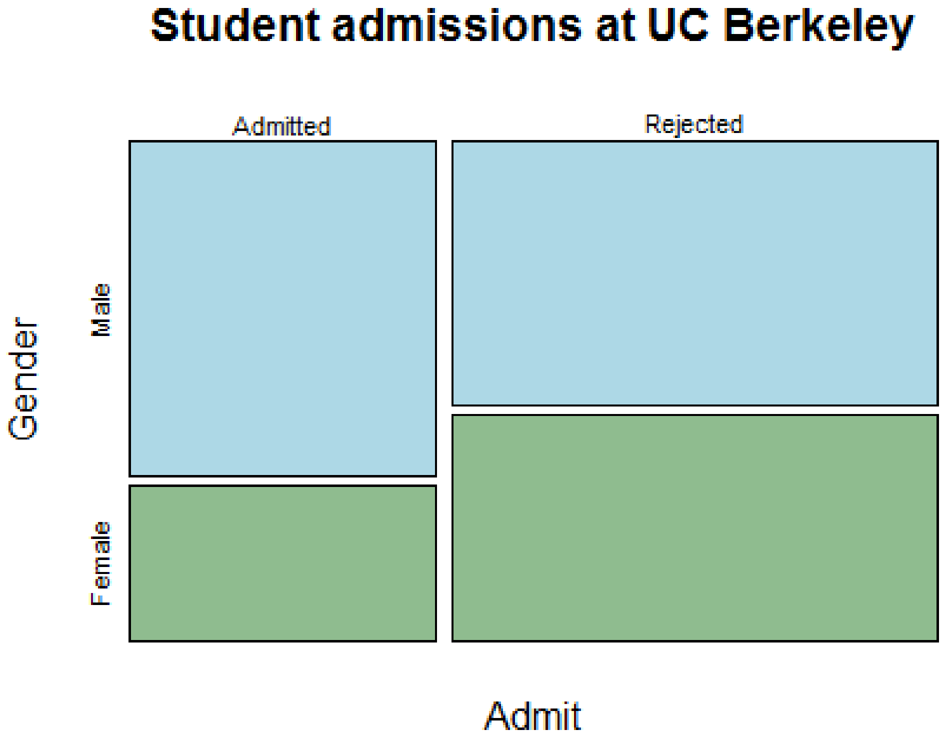
Visualization의 필요성

성별 합격율

1973년도 버클리 대학원의 6개 단과대학별 성별 합격여부 데이터

```
> apply(UCBAdmissions, c(1, 2), sum)
          Gender
Admit      Male  Female
Admitted   1198     557
Rejected   1493    1278
```

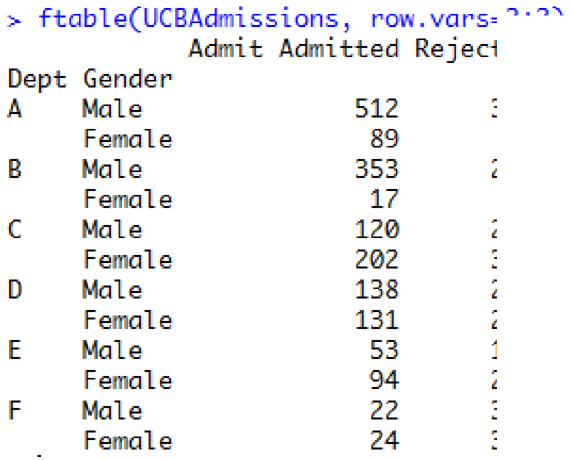
```
> prop.table(apply(UCBAdmissions, c
          Gender
Admit      Male      Female
Admitted   0.4451877 0.3035422
Rejected   0.5548123 0.6964578
```



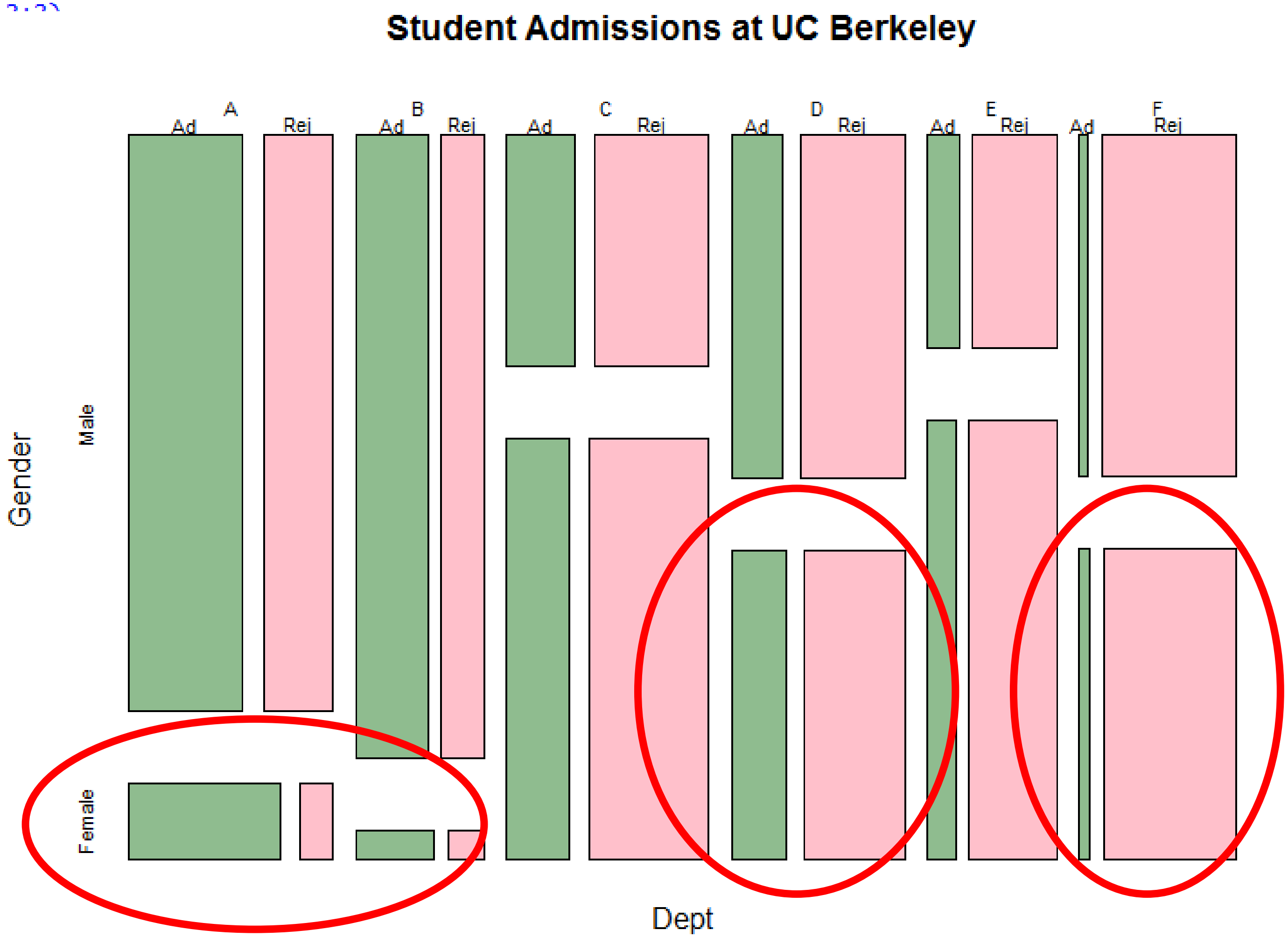
Visualization의 필요성 Student Admissions at UC Berkeley

성별 단과대학별 합격율

Simpson's Paradox

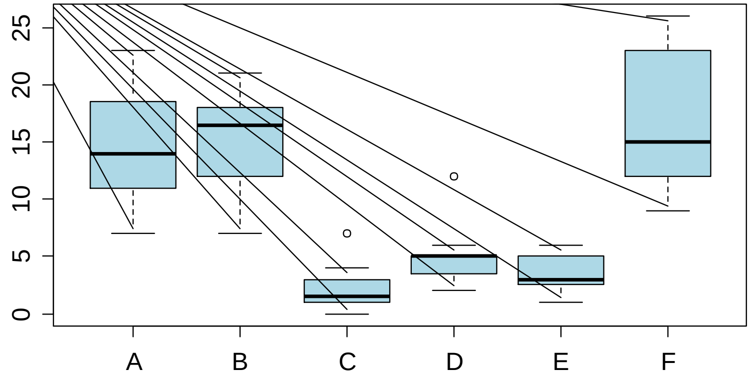


```
> ftable(UCBAdmissions, row.vars=
             Admit Admitted Reject
Dept Gender
A    Male            512
     Female           89
B    Male            353
     Female           17
C    Male            120
     Female          202
D    Male            138
     Female          131
E    Male             53
     Female           94
F    Male             22
     Female           24
```

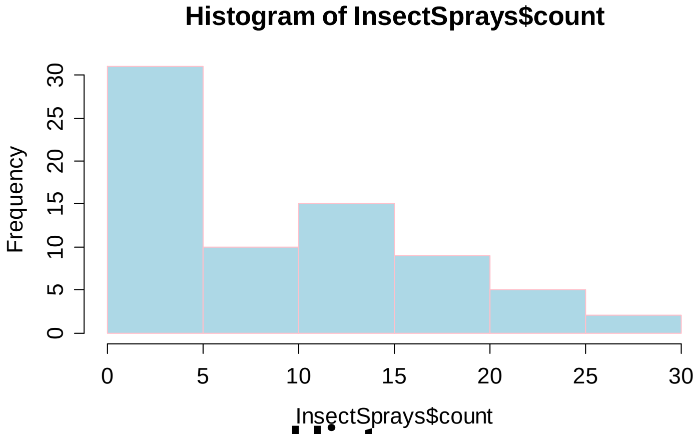


EDA EDA(Exploratory Data Analysis)

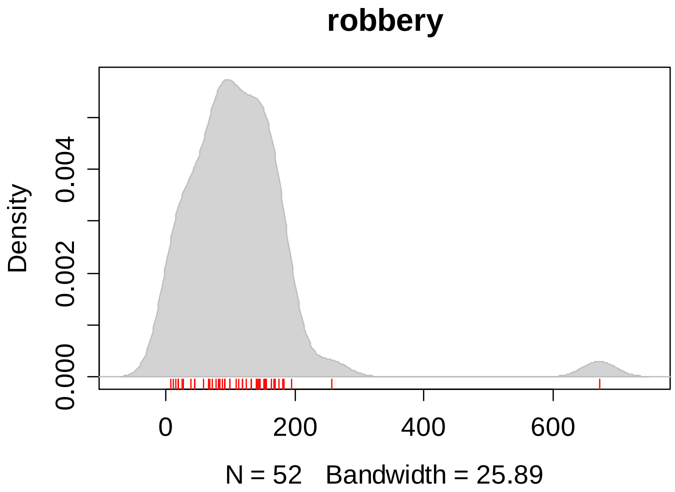
Chart for EDA



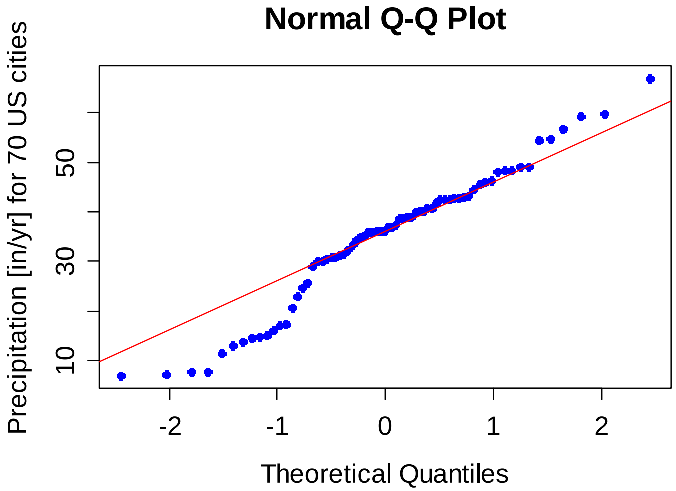
Box Plot



Histogram



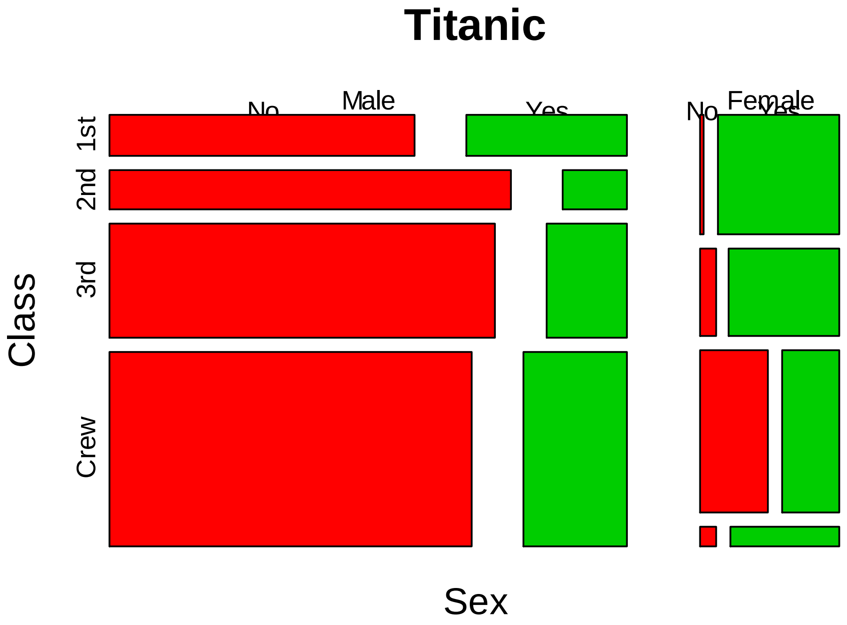
Density Plot



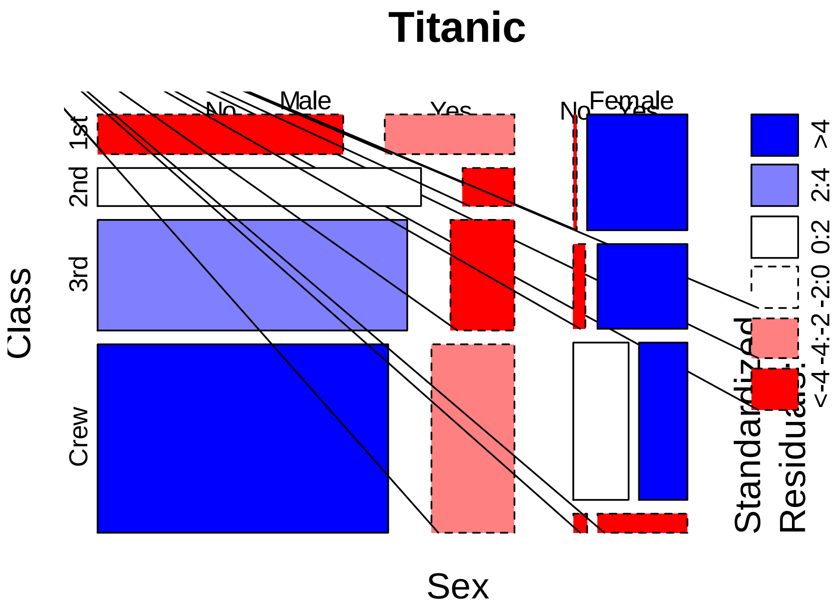
Q-Q Plot

EDA Categorical Data Plot

Mosaics Plot



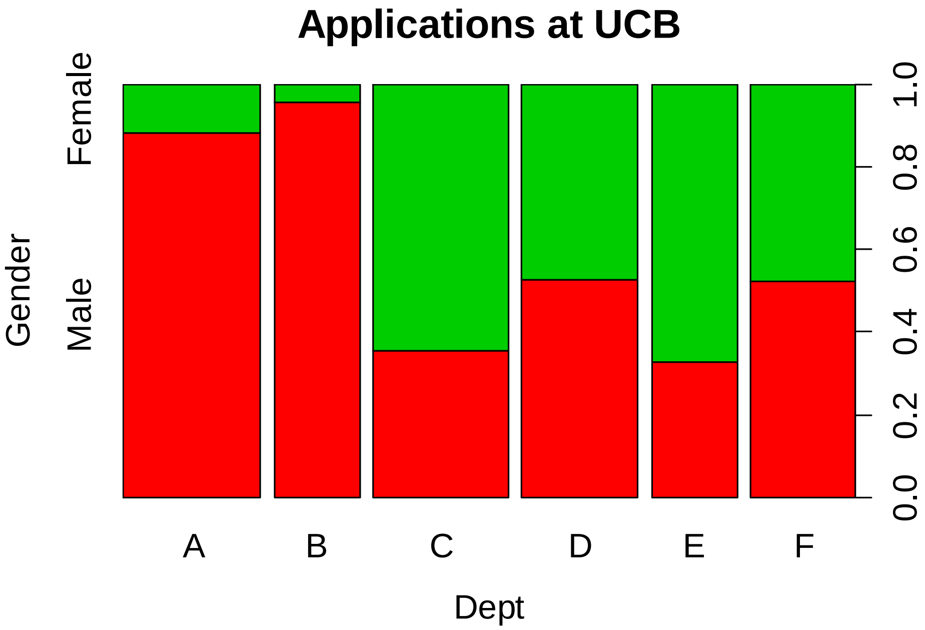
Mosaics Plot



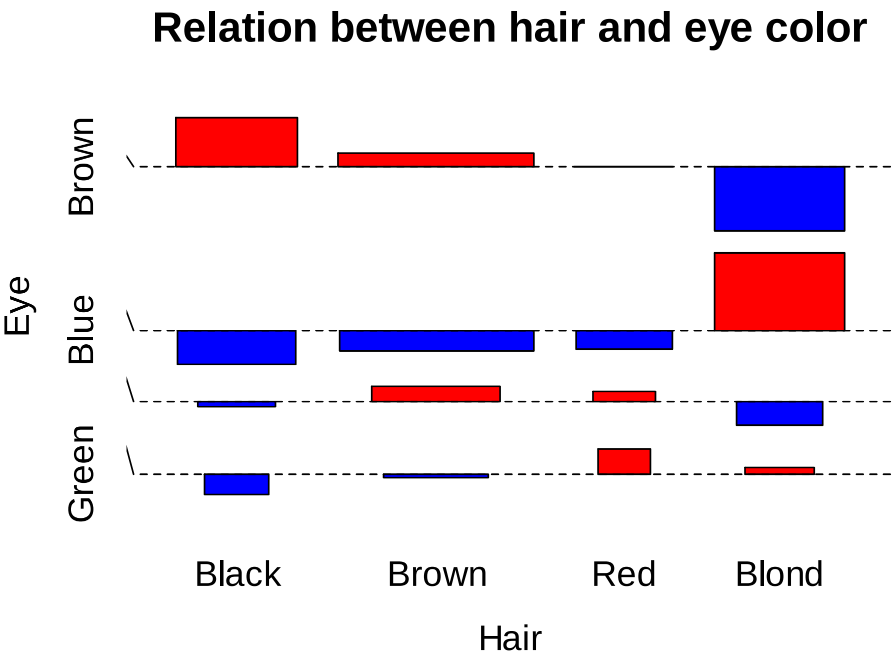
Mosaics Plot (Shade)

EDA Categorical Data Plot

Chart for Categorical Data



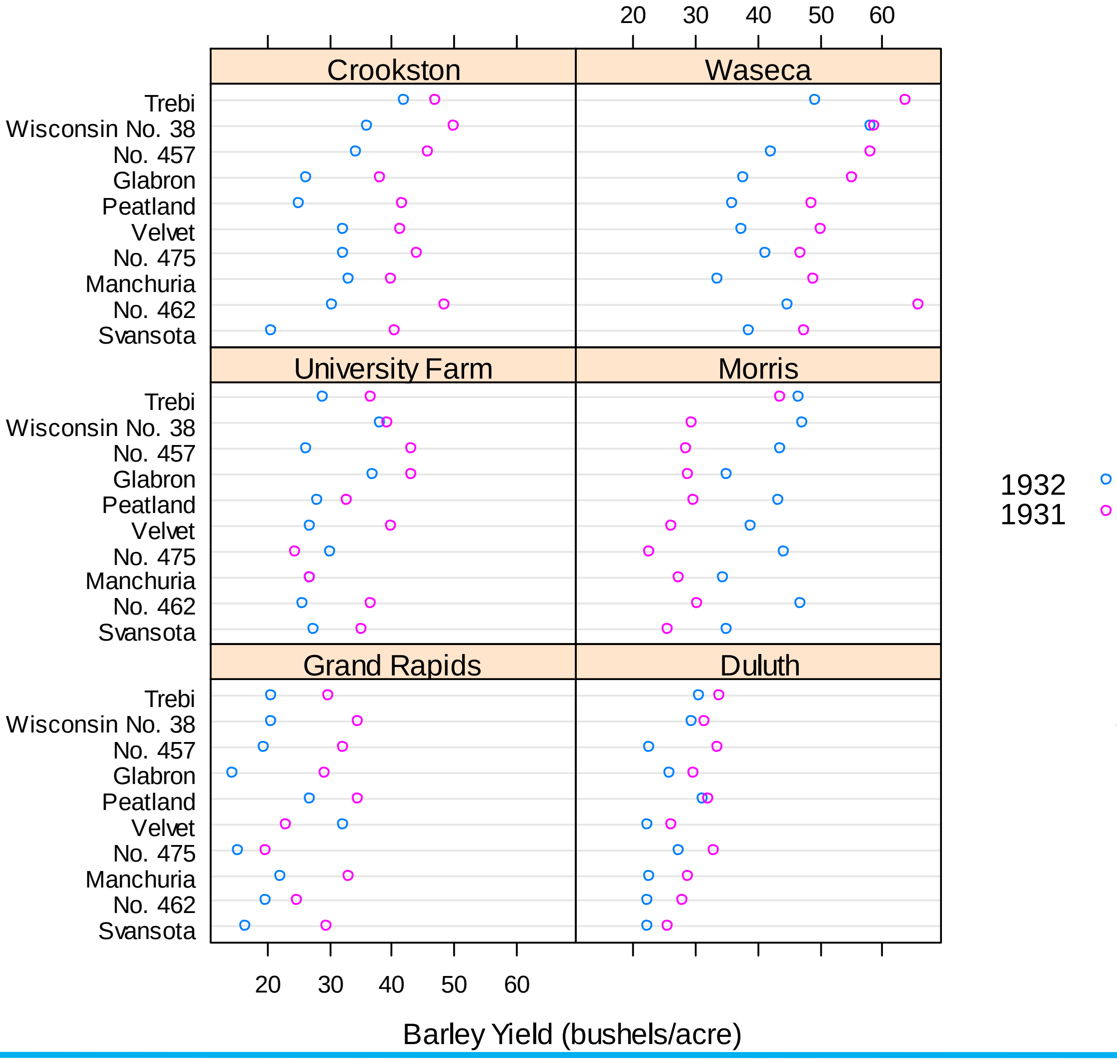
Spine Plot

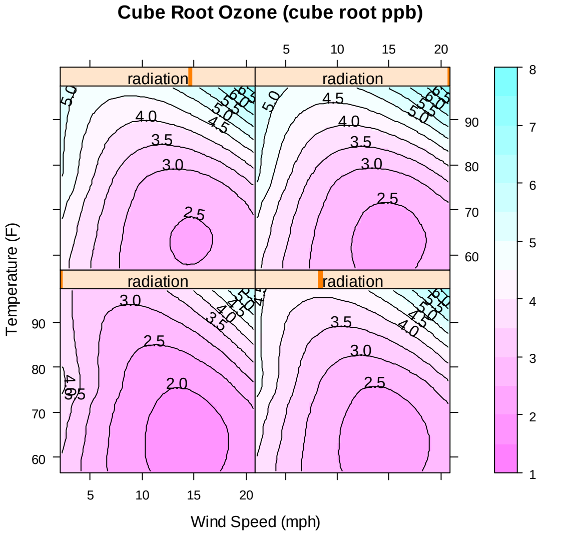


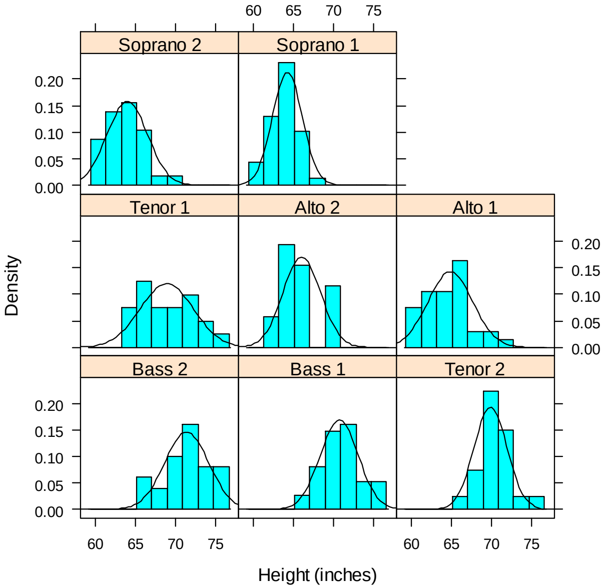
Association Chart

EDA Mosaics Plot

Trellis(Lattice)

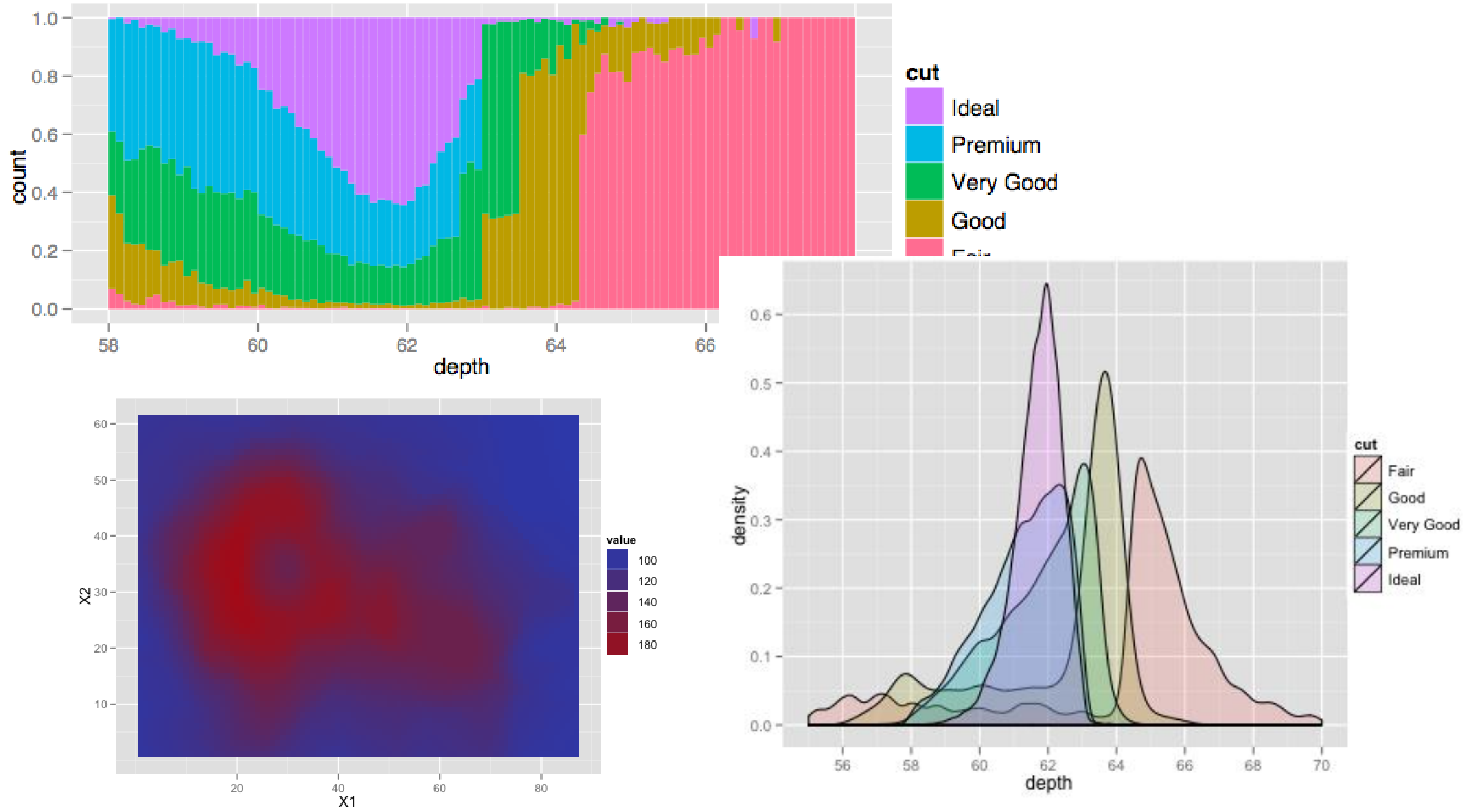






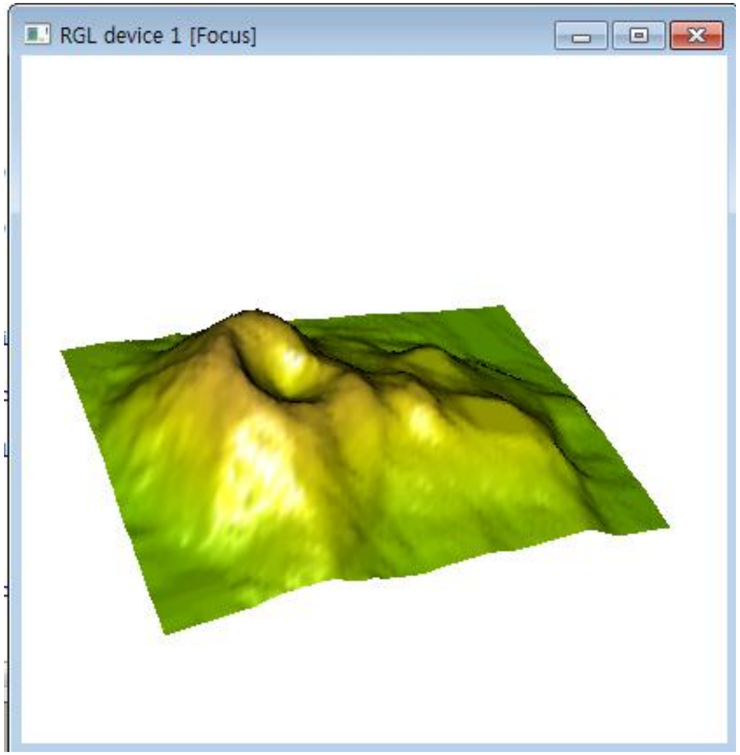
Special Chart ggplot

ggplot2



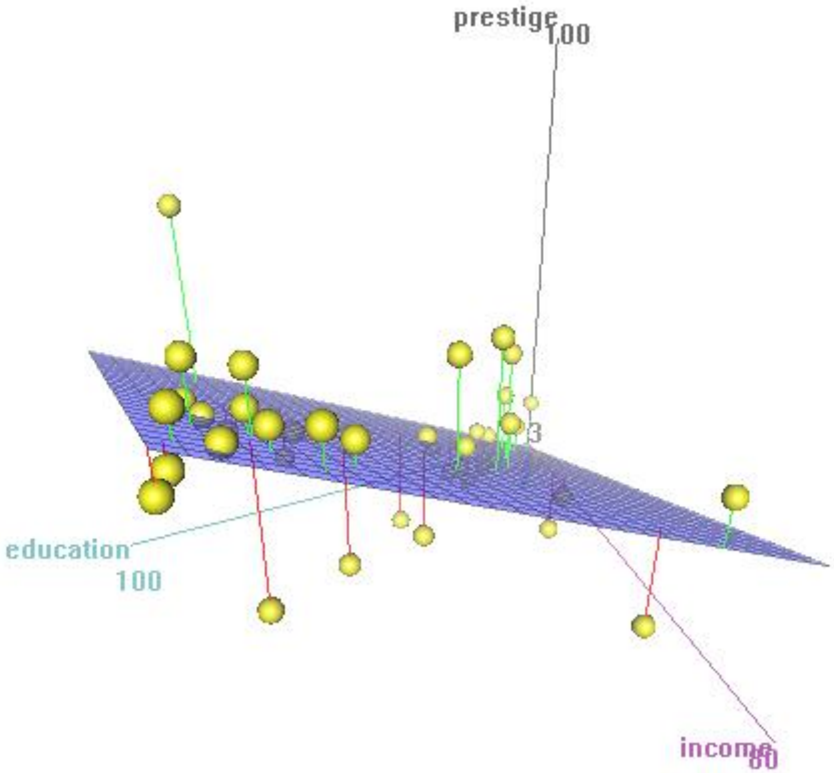
Special Chart Open GL Integration

rgl Package



rgl

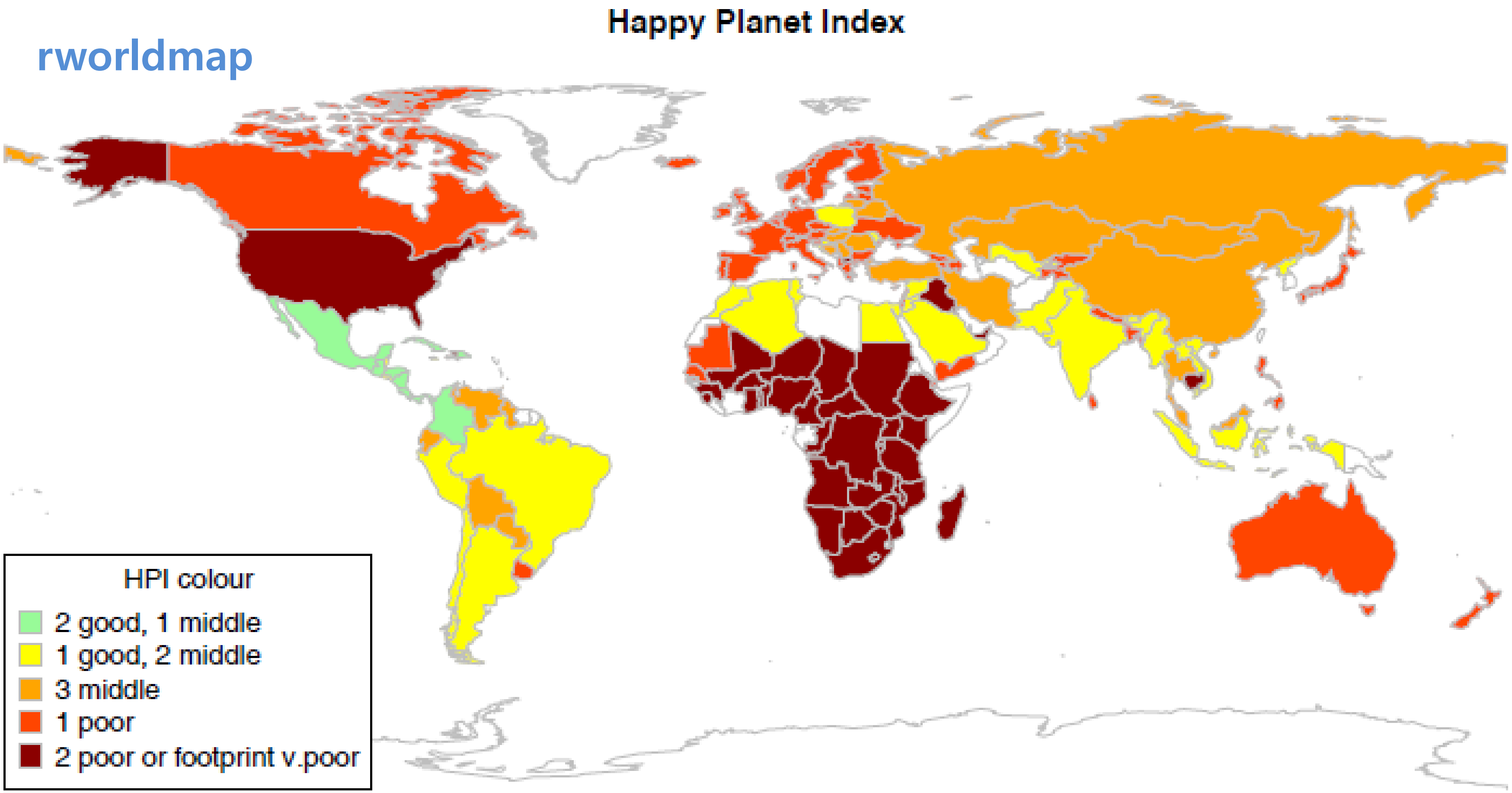
scatter3d



Special Chart Maps

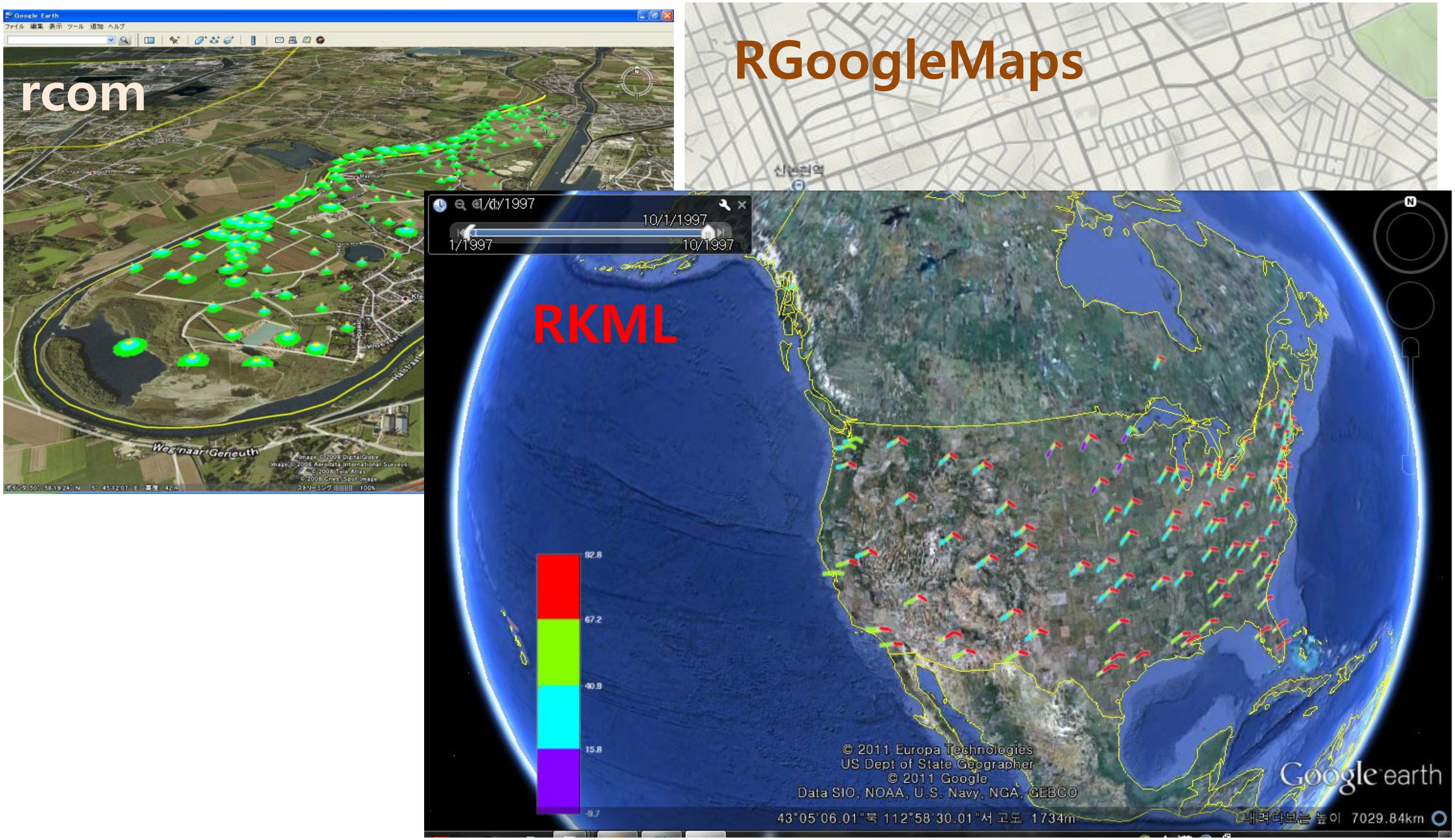
주제도

rworldmap



Special Chart Google Map Interface

Google Maps & Google Earth



rcom

RGoogleMaps

RKML

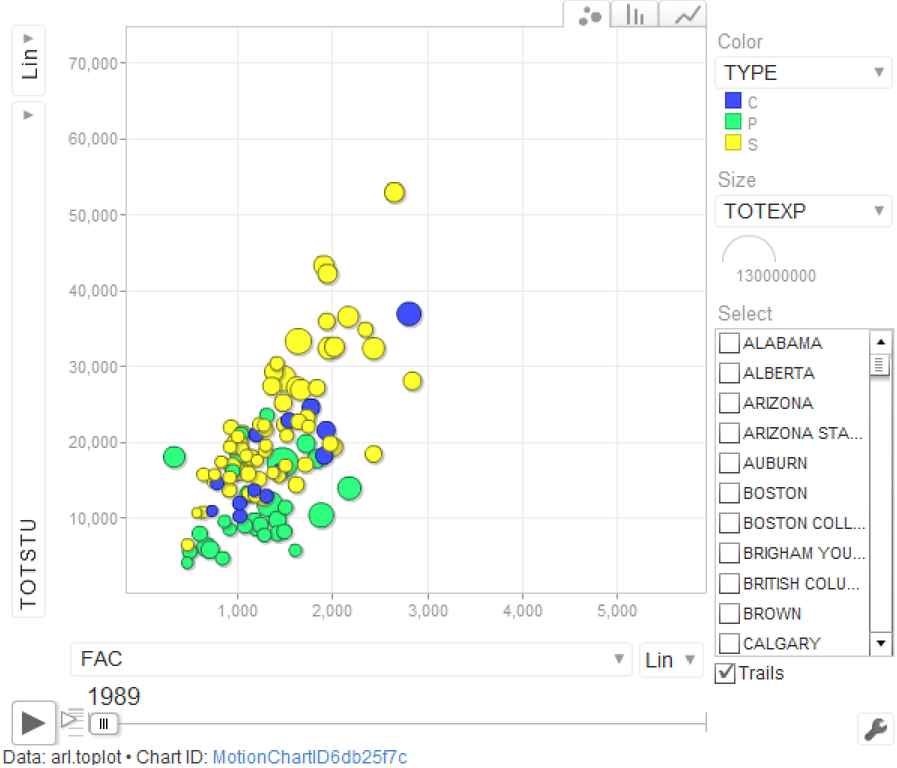
Google Earth
ファイル 編集 表示 ツール 追加 ヘルプ
Weg naar Geneuth
Image © 2008 DigitalGlobe
Image © 2008 Aerodata International Surveys
© 2008 Tele Atlas
© 2008 Cnes/Spot Image
신논현역
1/1997
10/1/1997
1/1997
10/1997
92.8
67.2
40.9
15.8
-9.7
© 2011 Europa Technologies
US Dept of State Geographer
© 2011 Google
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Google earth
43°05'06.01"북 112°58'30.01"서 고도 1734m
내려다보는 높이 7029.84km

Special Chart Google Visualization Interface

GoogleVis package

Speed kt
20
135
지도
위성
Google
지도 데이터 - 이용약관

gvisGeoMap



Color: TYPE ▼
🟦 C
🟩 P
🟨 S

Size: TOTEXP ▼
130000000

Select
□ ALABAMA
□ ALBERTA
□ ARIZONA
□ ARIZONA STA...
□ AUBURN
□ BOSTON
□ BOSTON COLL...
□ BRIGHAM YOU...
□ BRITISH COLU...
□ BROWN
□ CALGARY

☑ Trails

FAC ▼

1989

Data: arl.toplot • Chart ID: MotionChartID6db25f7c

gvisMotionChart

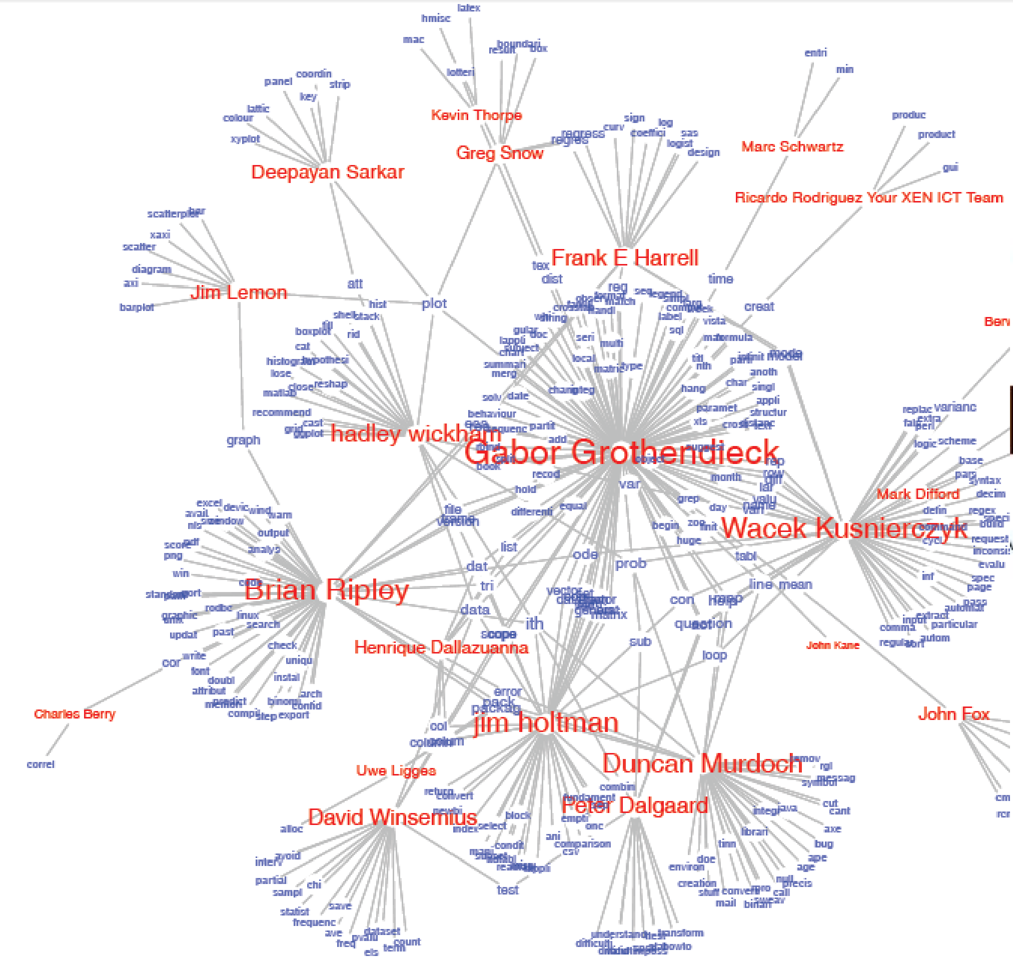
Big Data Analytics Social Networks

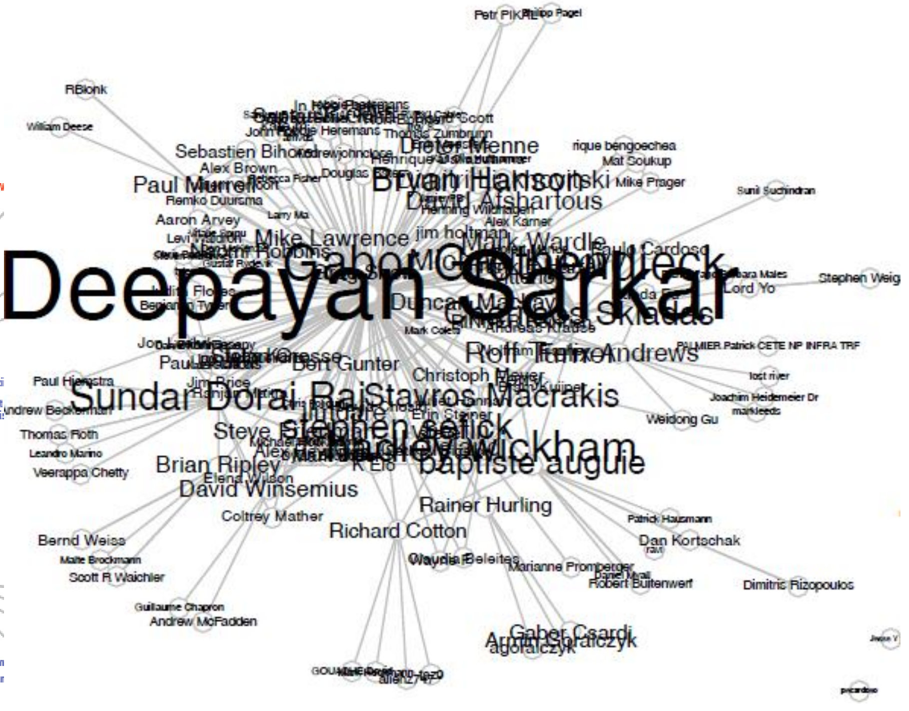
Insightful Visualization

facebook
December 2010

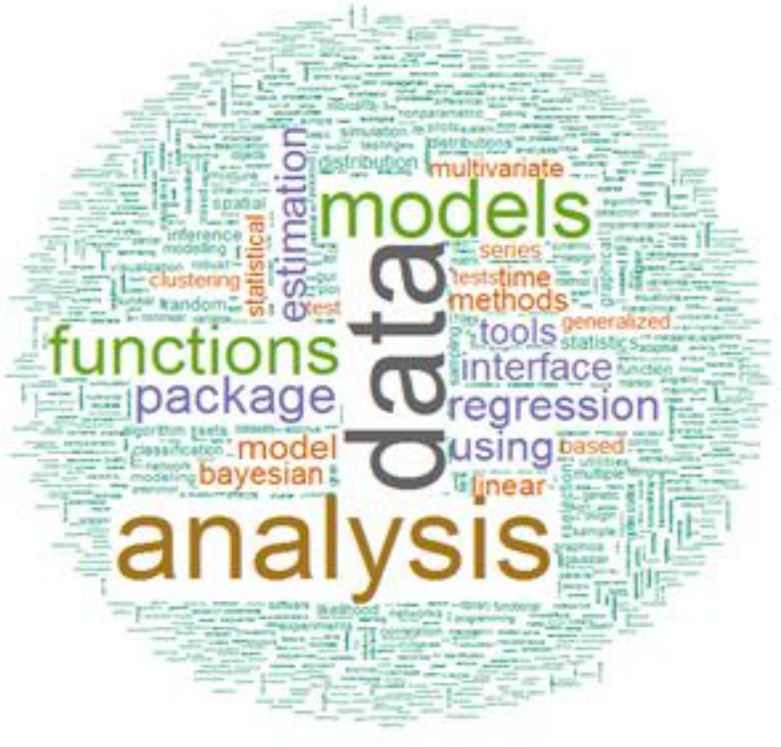
Social Network Analysis

R Journal 2011-1의 “Content-Based Social Network Analysis of Mailing Lists” 인용

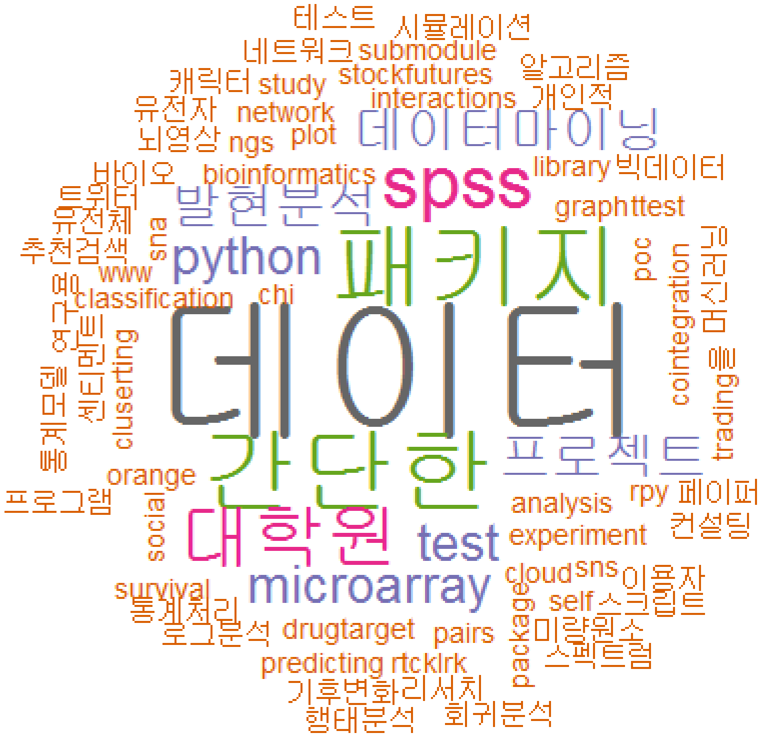




Word Cloud Chart



R Mailing List

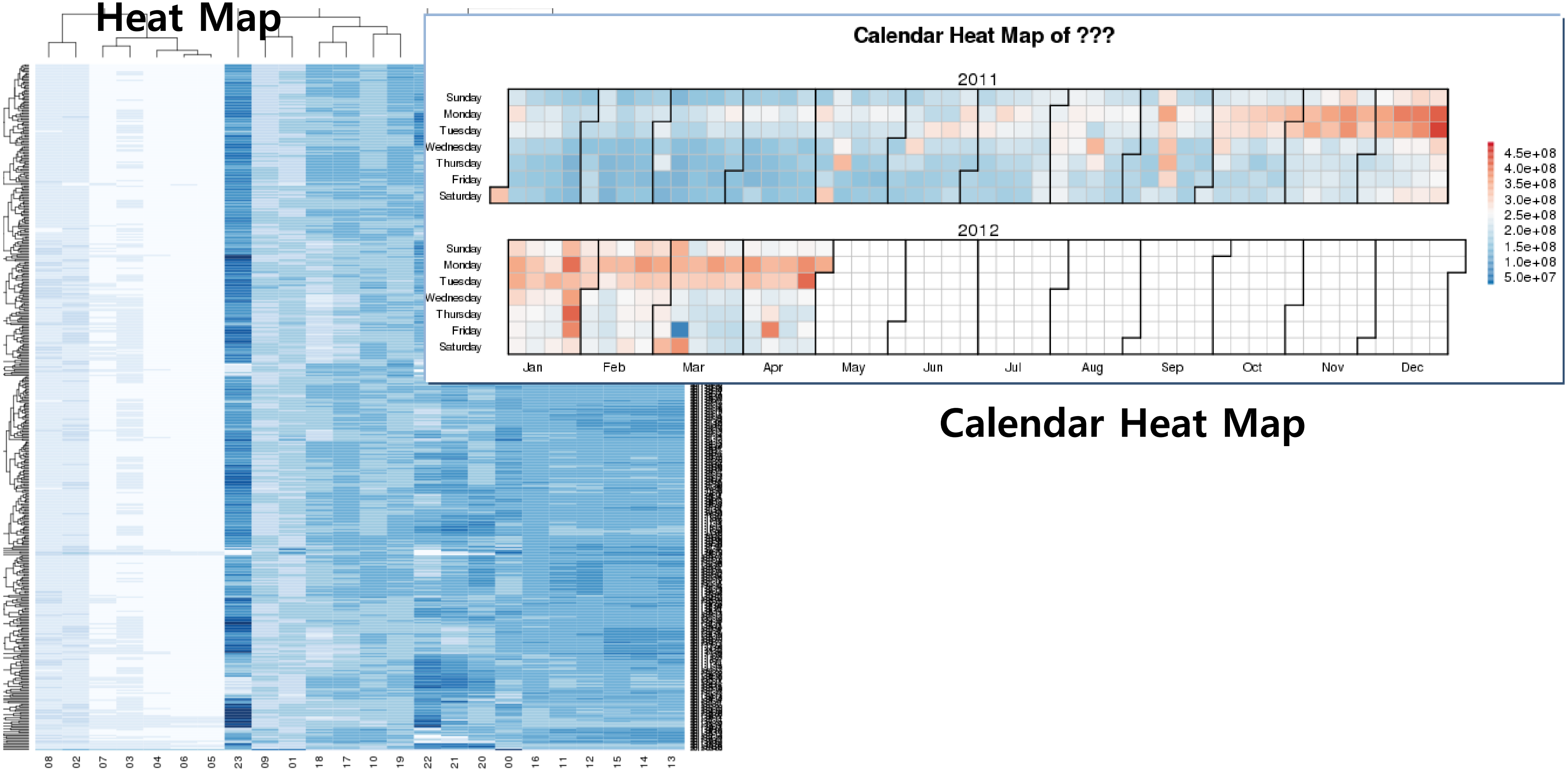


R User Conference Survey

Big Data Analytics ETC Chart

Heat Map

Heat Map



Calendar Heat Map

Q&A