Introduction to R

Setia Pramana
Politeknik Statistika STIS

Available Statistical Packages

Available Statistical Packages

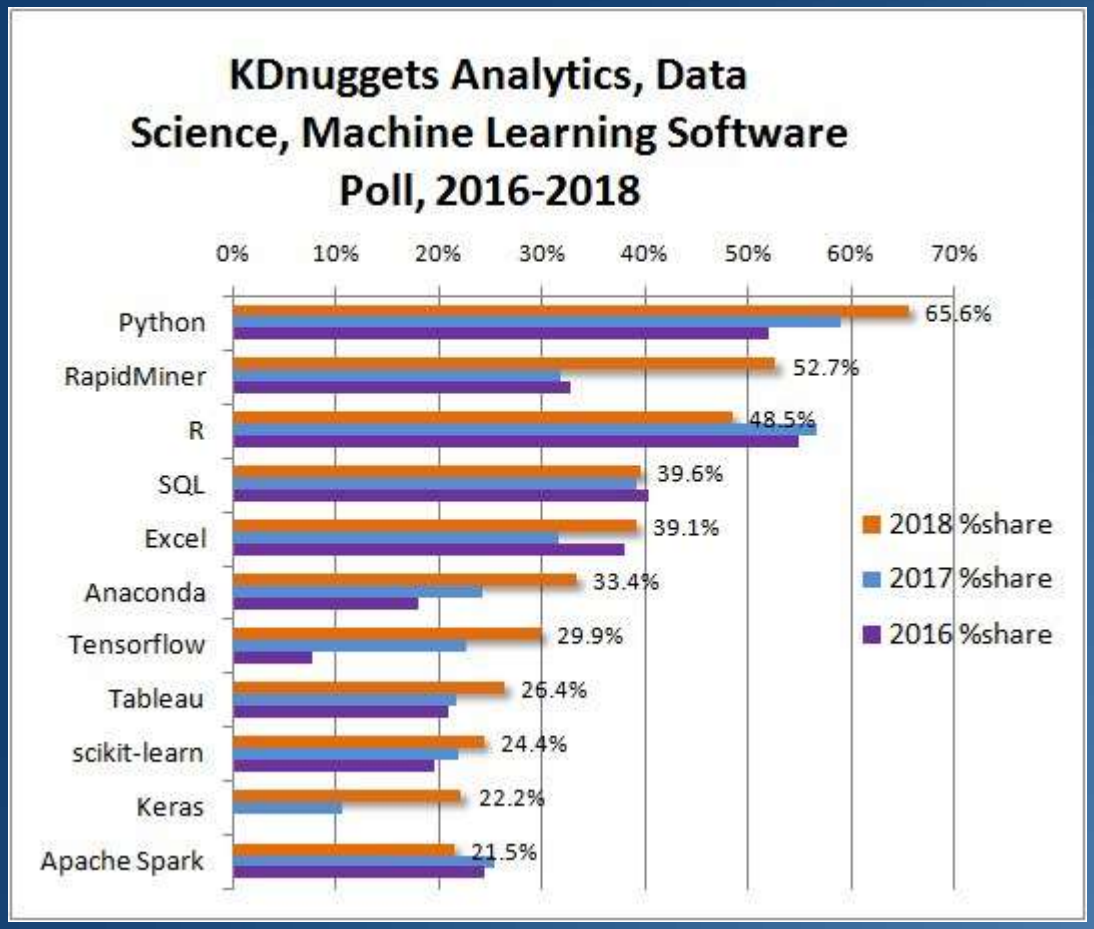
Proprietary

- Excel
- SPSS
- MINITAB
- SAS
- Stata
- Statistica
- Many more ……

Free Software

- LibreOffice Calc
- R
- CS Pro
- WinBugs
- EpiInfo
- Many more……..

Statistical Software Used



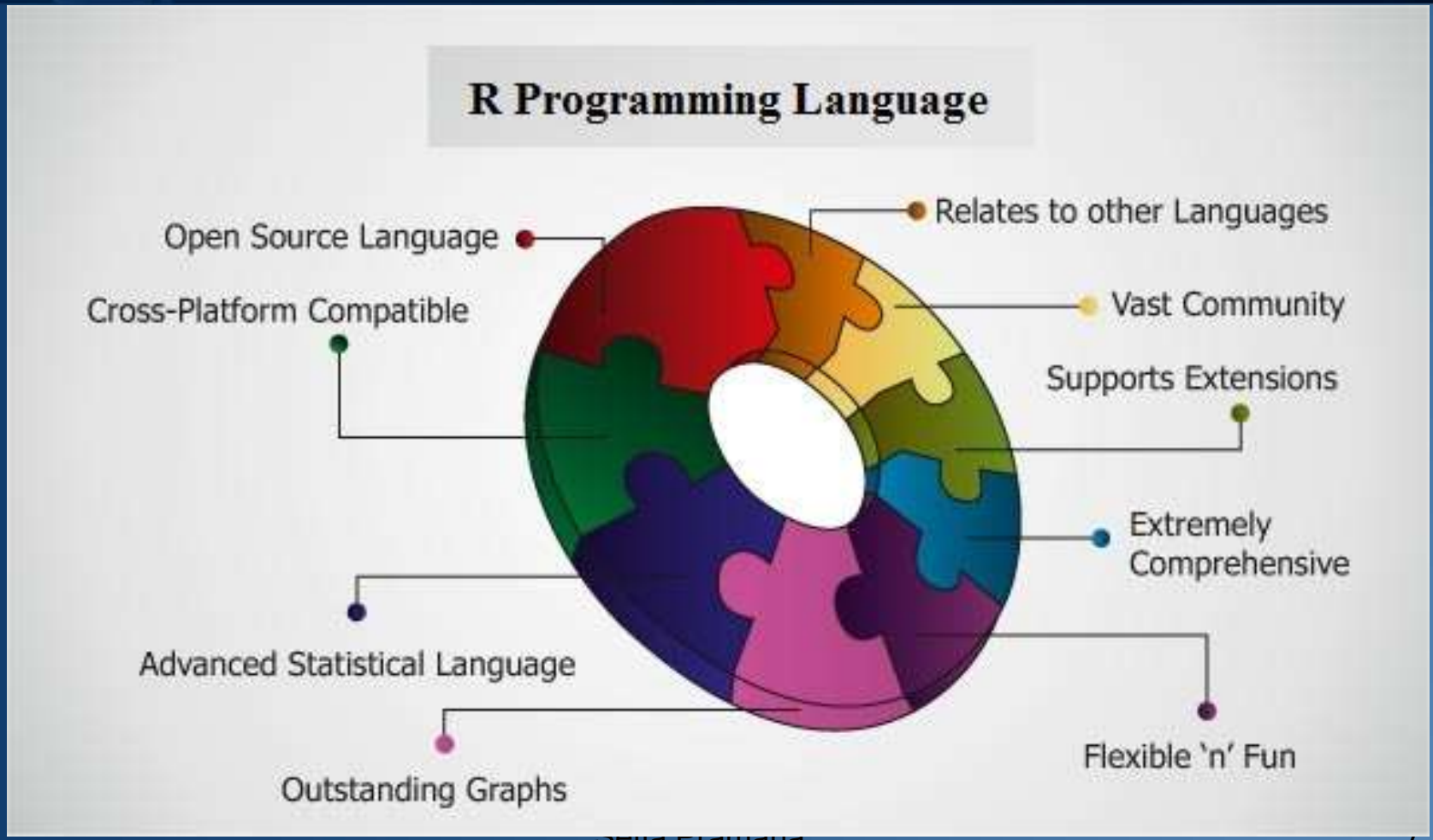
What is R?

- A language and environment for statistical computing and graphics.
- An integrated suite of software facilities for data manipulation, calculation and graphical display.
- First appeared in 1996 by Prof. Ross Ihaka and Robert Gentleman of the University of Auckland, NZ.
- GNU software -> Free. Similar like S language.
- Open source, maintained and developed by a community of developers.
- Works in Windows, Unix, MacOs

R Includes

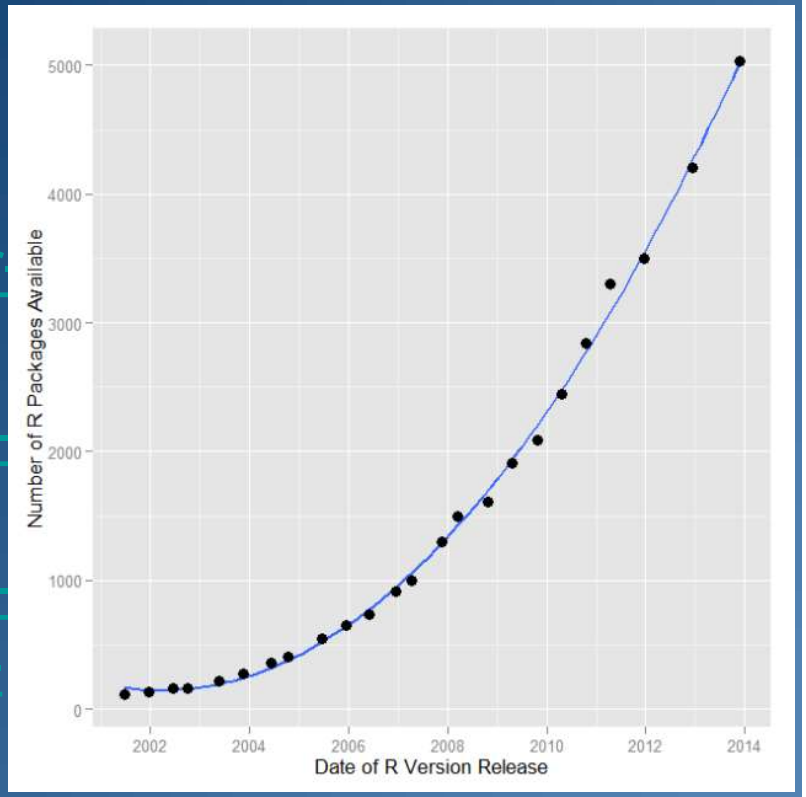
- Effective data handling and storage facility,
- A suite of operators for calculations on arrays, in particular matrices
- A large, coherent, integrated collection of intermediate tools for data analysis,
- Graphical facilities for data analysis and display either on-screen or on hardcopy
- Well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

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


R Programming Language
Open Source Language
Relates to other Languages
Cross-Platform Compatible
Vast Community
Supports Extensions
Extremely Comprehensive
Advanced Statistical Language
Outstanding Graphs
Flexible 'n' Fun

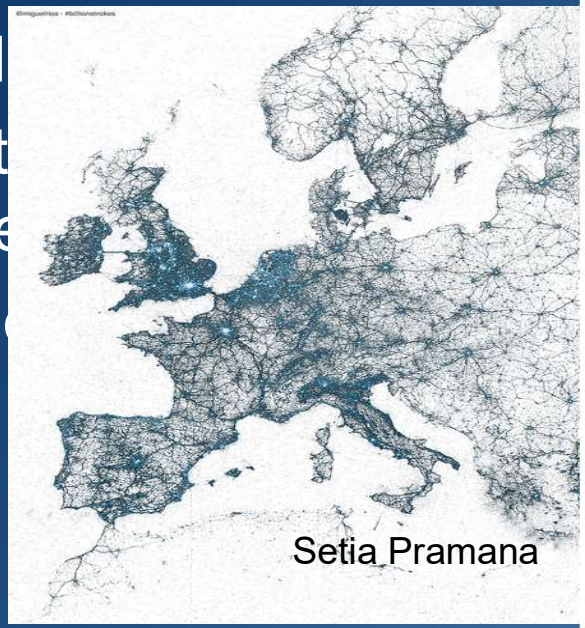
Why R?

- It is not only statistical software but also a language
- 10.000 add-on packages → lots of pre-prepared packages (http://cran.r-project.org/web/packages/)
- With many applications http://cran.r-project.org/web/views/, http://www.revolutionanalytics.com/r-language-features-applications-and-extensions#thirdparty .
- Access to powerful, cutting-edge analytics

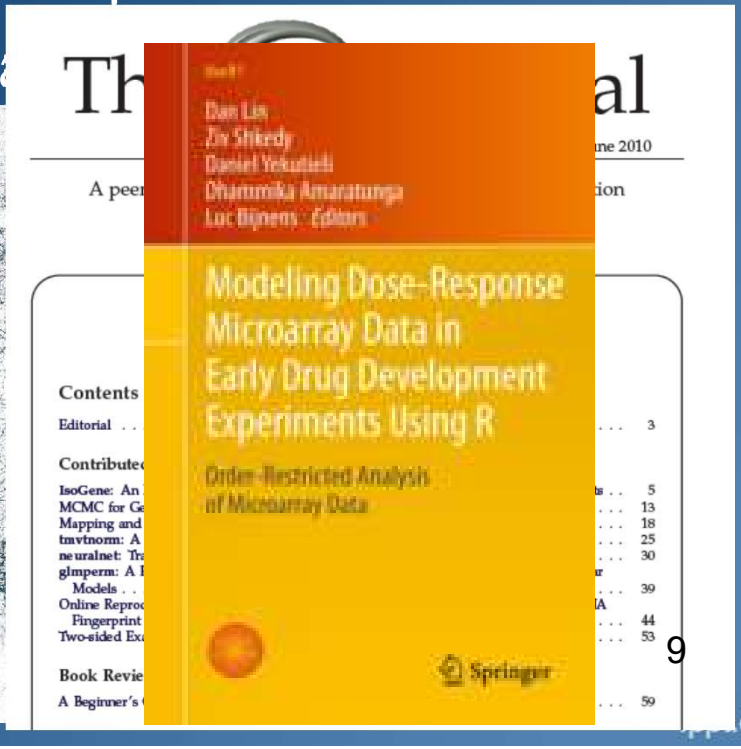


Why R?

- Flexible (complex or standard statistical practices, bayesian modelling, GIS map building, building interactive web applications, building interactive tests, etc.)
- We can make our own package and publish it
- Great Graphics and data visualiz
- Can be used
- Well Support
  resources-we
- And many m



Setia Pramana



Why R?

- Can be integrated with other languages (C/C++, Java).
- R can interact with many data sources and other statistical packages (SAS, Stata, SPSS, and Minitab).
- For the high performance computing task → multiple cores, either on a single machine or across a network.

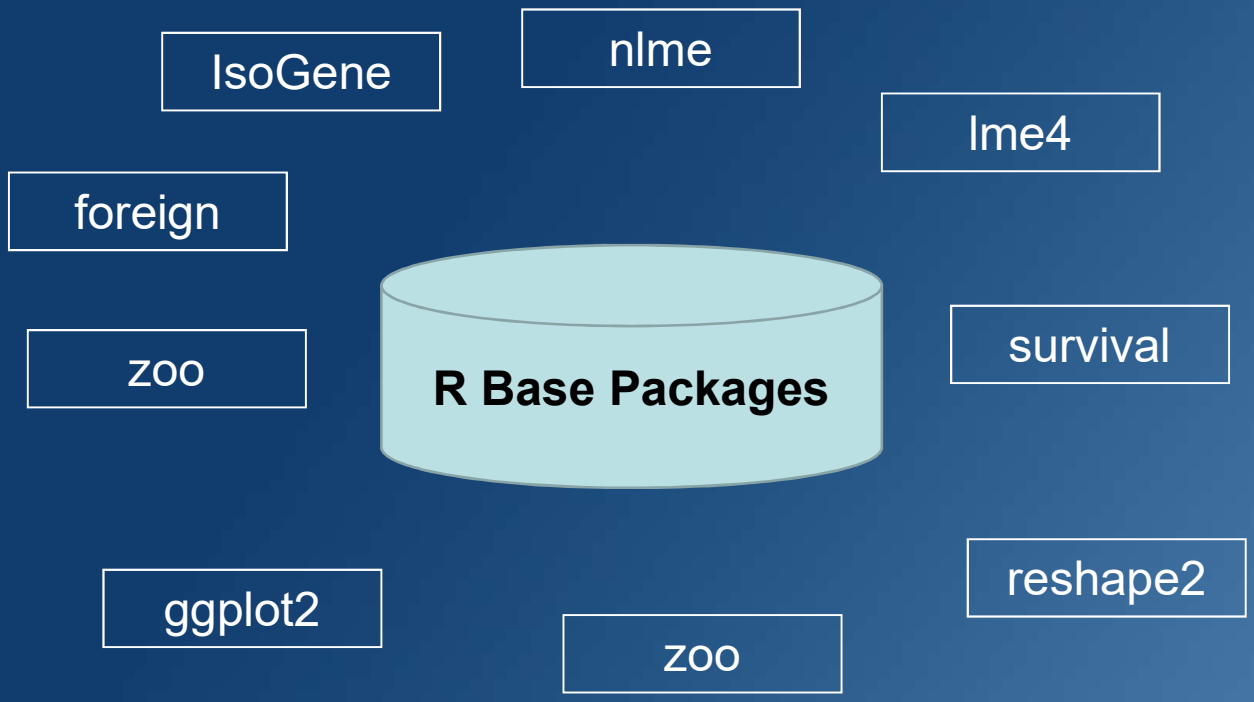
But…..

- R has no warranty
- Command Line Interface : difficult for some users.
- Users must learn a new way of thinking about data and data analysis sequence
- That's all ..... I guess

Learning R

Companies that use R for Analytics

R Library/packages



My R Packages

- IsoGene
- IsoGeneGUI
- nea
- neaGUI
- biclustGUI
- OCRME
- More detail: http://setiopramono.wordpress.com/r-programming/

Cutting Edge Technologies

Data Mining

- Decision trees: rpart, party
- Random forest: randomForest, party
- SVM: e1071, kernlab
- Neural networks: nnet, neuralnet, RSNNS
- Performance evaluation: ROCR
- Data Mining GUI rattle
- etc..
- http://www.rdatamining.com/

Social Media

- Text mining: tm
- Topic modelling: topicmodels, lda
- Word cloud: wordcloud
- Facebook: RFacebook
- Twitter data access: twitteR
- Social Network: sna, igraph, RSiena

  (http://www.jstatsoft.org/v24/i06/paper)

  http://www.r-bloggers.com/an-example-of-social-network-analysis-with-r-using-package-igraph/

Parallel Computing

- snow (Simple Network of Workstations) & snowfall for development of parallel R programs.
- multicore parallel processing of R code on machines with multiple cores or CPUs
- More: http://cran.r-project.org/web/views/HighPerformanceComputing.html

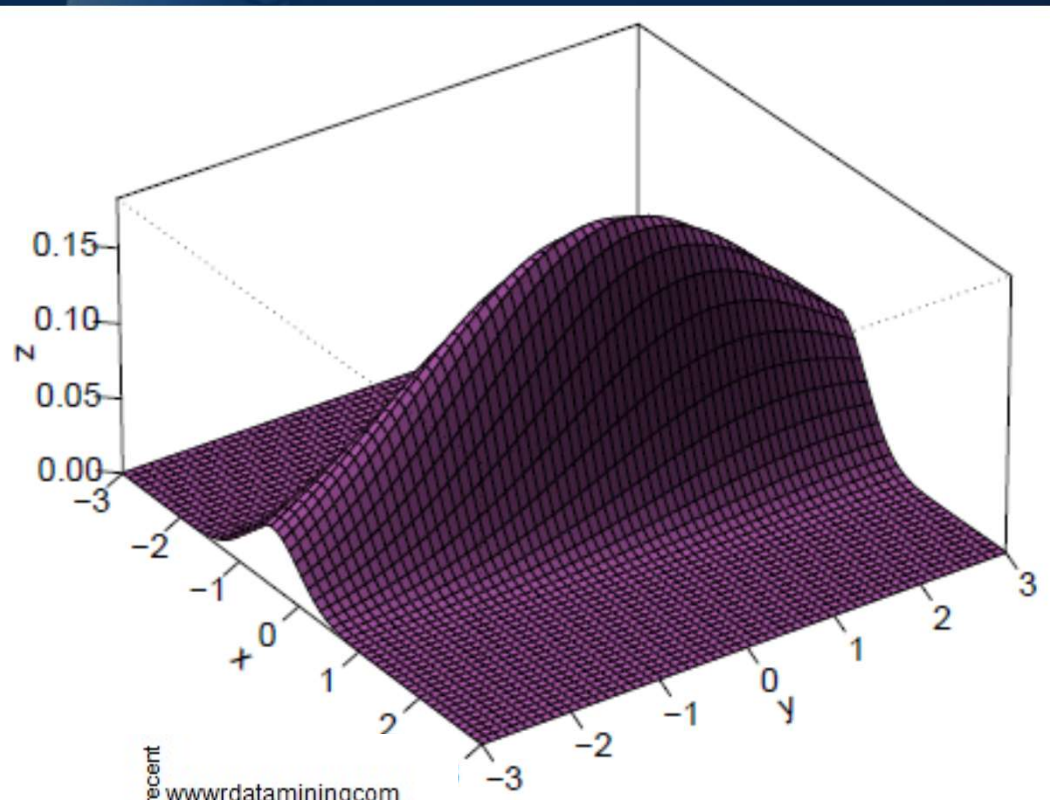
Big Data

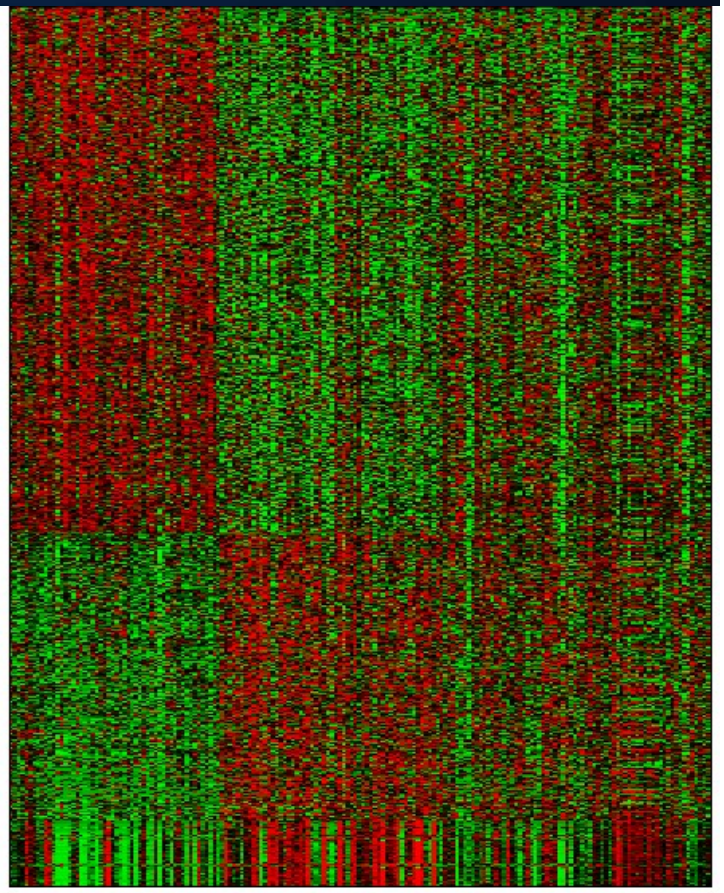
- RHadoop - a collection of five R packages that allow users to manage and analyze data with Hadoop, developed by Revolution Analytics
- RHIPE - an R and Hadoop Integrated Programming Environment
- More…...

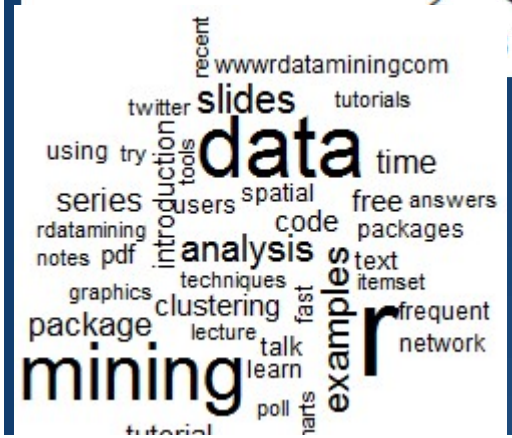
R Graphics and Visualization

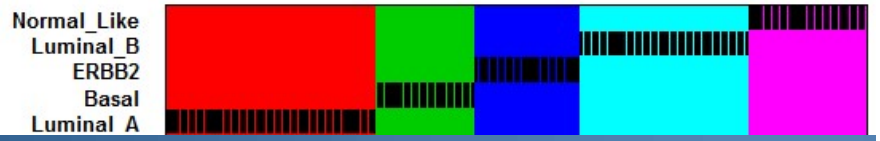
- R provides wide range graphics and visualizations
- Basic Plots: bar plots, basic 3D plots, heatmap.,etc
- Geographic Maps
- Projection Maps
- Social Network Graphs
- Animated graphics and movies (animation)
- Motion Charts (GoogleViz)
- Interactive Graphics (rggobi)
- Image format: BMP, JPEG, PDF, PNG etc…
- More: https://www.r-graph-gallery.com/

R Graphics

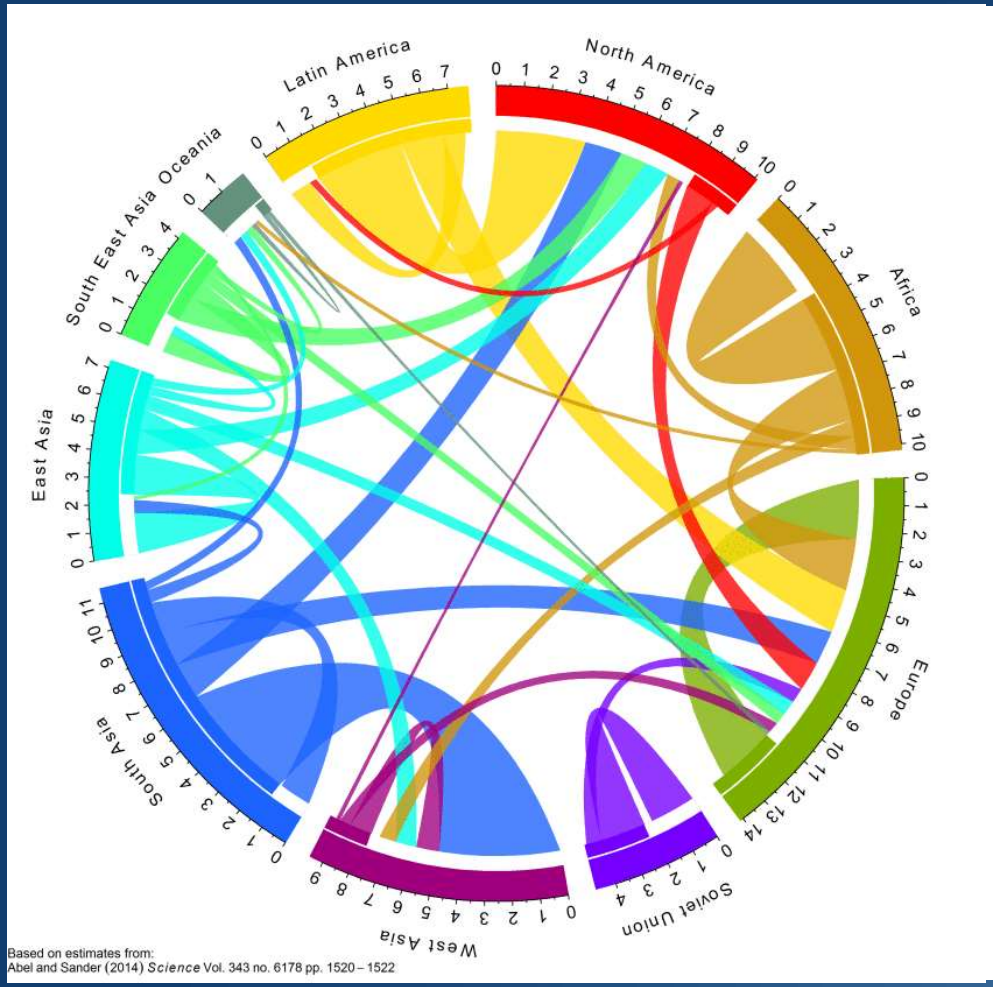


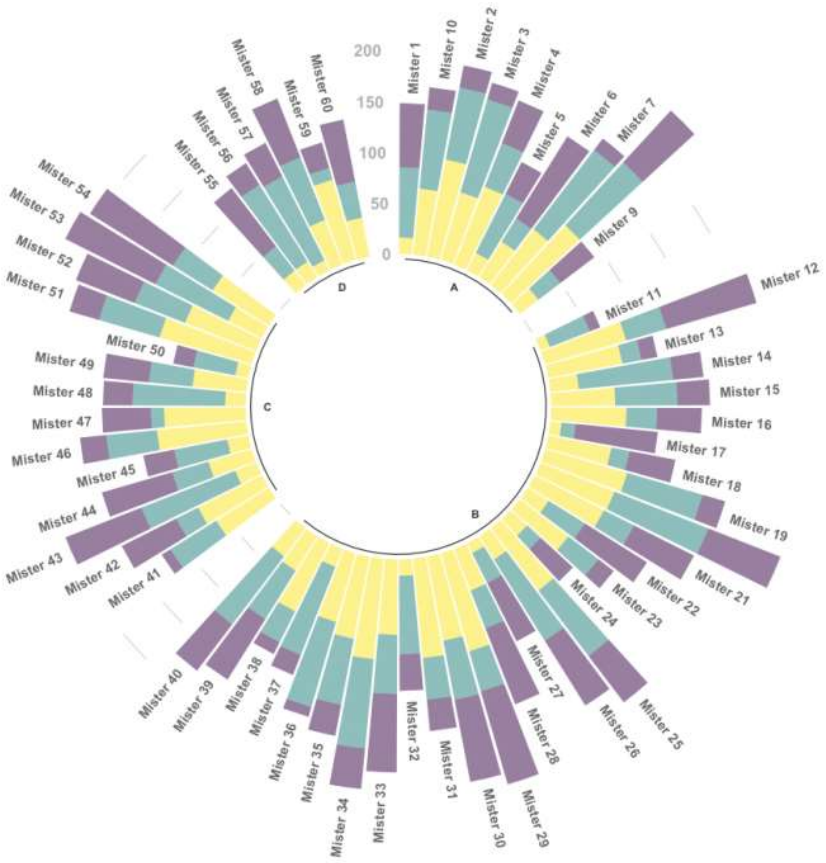






R Graphics





R Graphical User Interfaces

- R uses Command line interface and it is preferred for advanced users → allows direct control, more accurate, flexible and the analysis is reproducible.
- Requires good knowledge of the language → difficult for beginners or less frequent users.
- R provides tools for building GUIs → RGUI

R GUI Projects

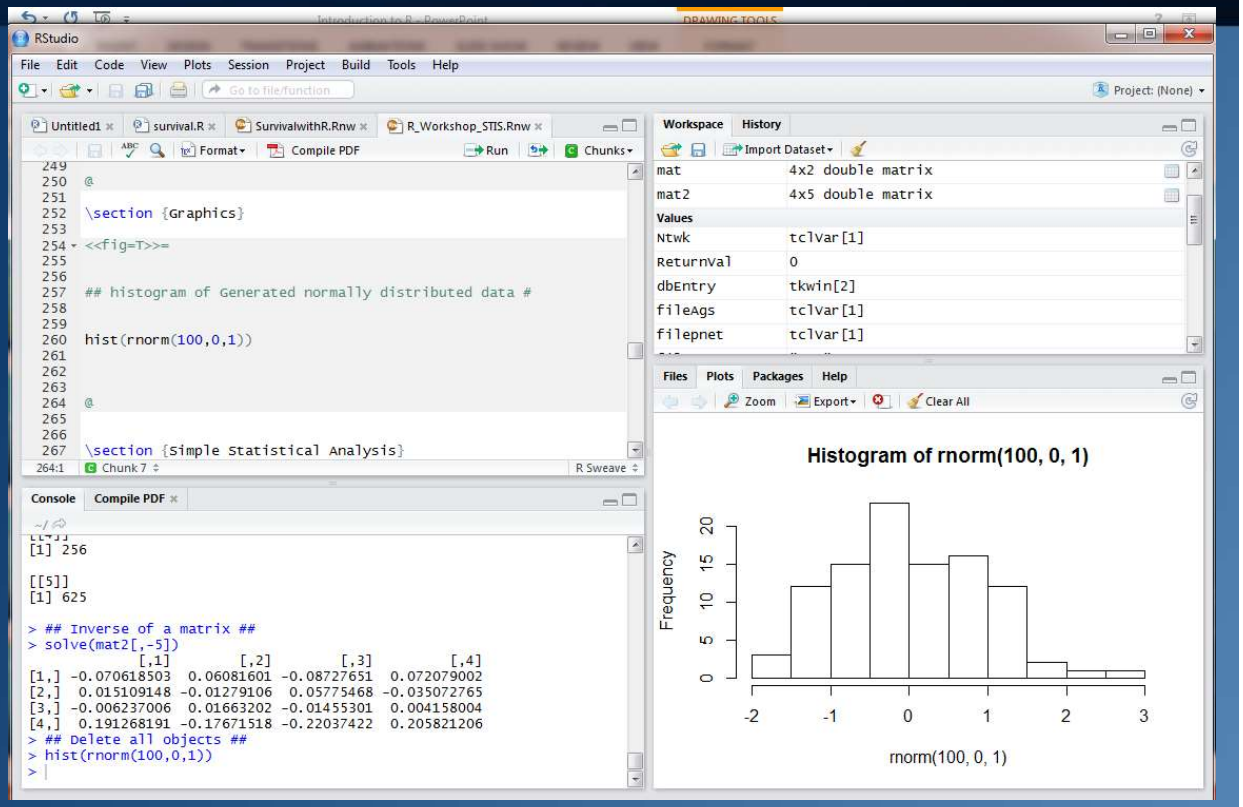
- Integrated development environment (IDE)/Script Editors aimed to provide feature-rich environments to edit R scripts and code: Rstudio (www.rstudio.com), and architect (www.Openanalytics.eu)
- Web based application: the Rweb (Banfield, 1999), R.Net (www.u.arizona.edu/~ryckman/Net.php), or gWidgetsWWW (Verzani, 2012).

R GUI Projects

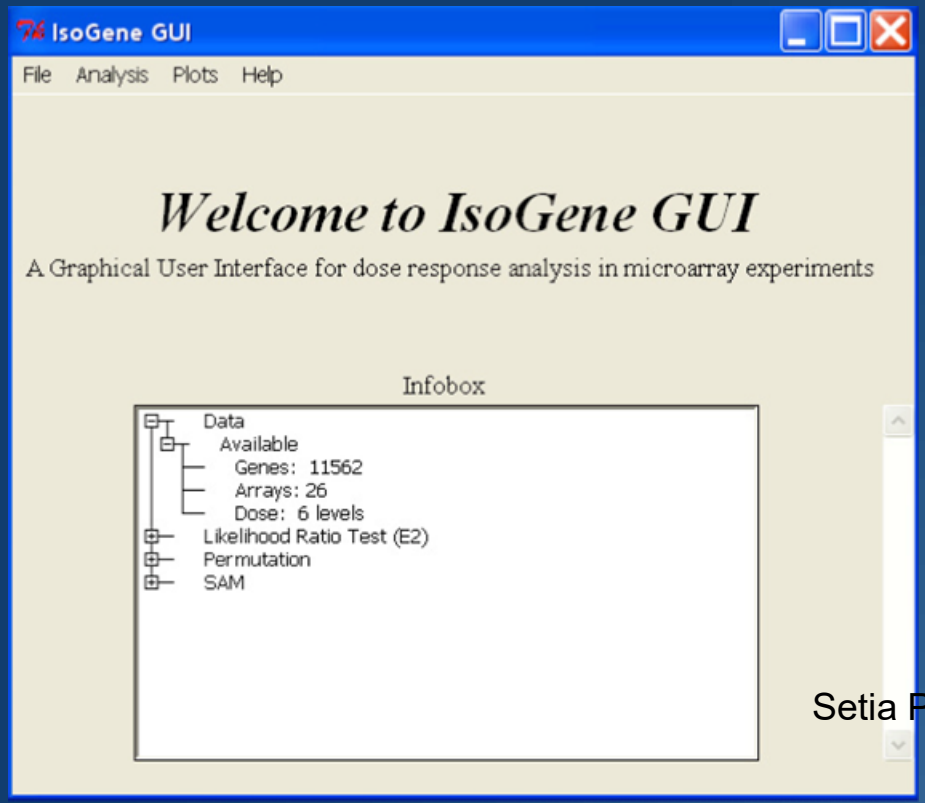
- Python: **OpenMeta-Analyst** (Wallace et al, 2012)
- Java: **JGR** (Java GUI for R), **Deducer** (Fellows, 2012), and **Glotaran** (Snellenburg, 2012).
- Php: R-php (http://dssm.unipa.it/R-php/)
- Other extensions connect R to graphical toolboxes for developing menus and dialog boxes: **Tcltk**, **Gtk**.

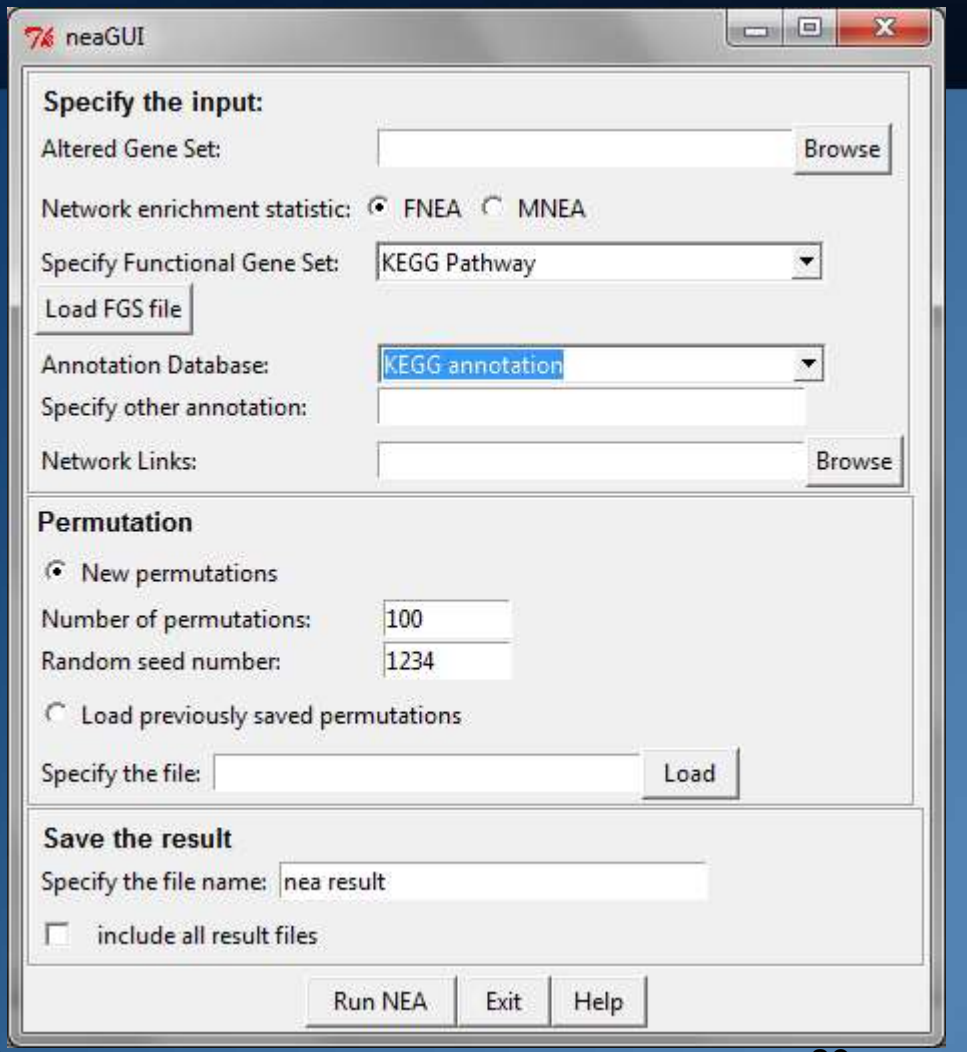
R Studio

- Download from Rstudio.com
- Powerfull IDE (Integrated Development Environment) for R.



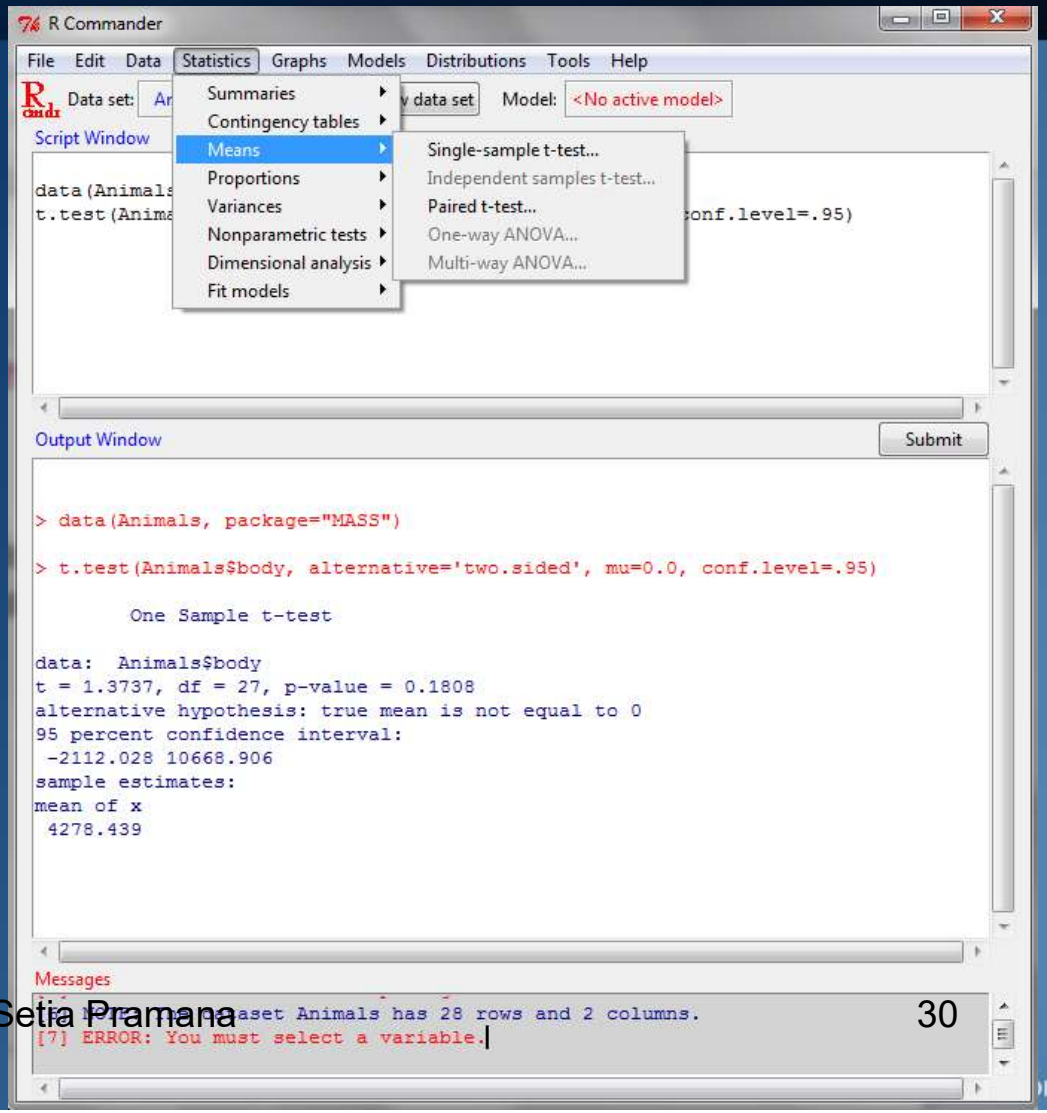
RGUI Developed using tcltk





RGUI: RCommander

- Rcommander.com
- Helpful for R beginner
- Install inside R

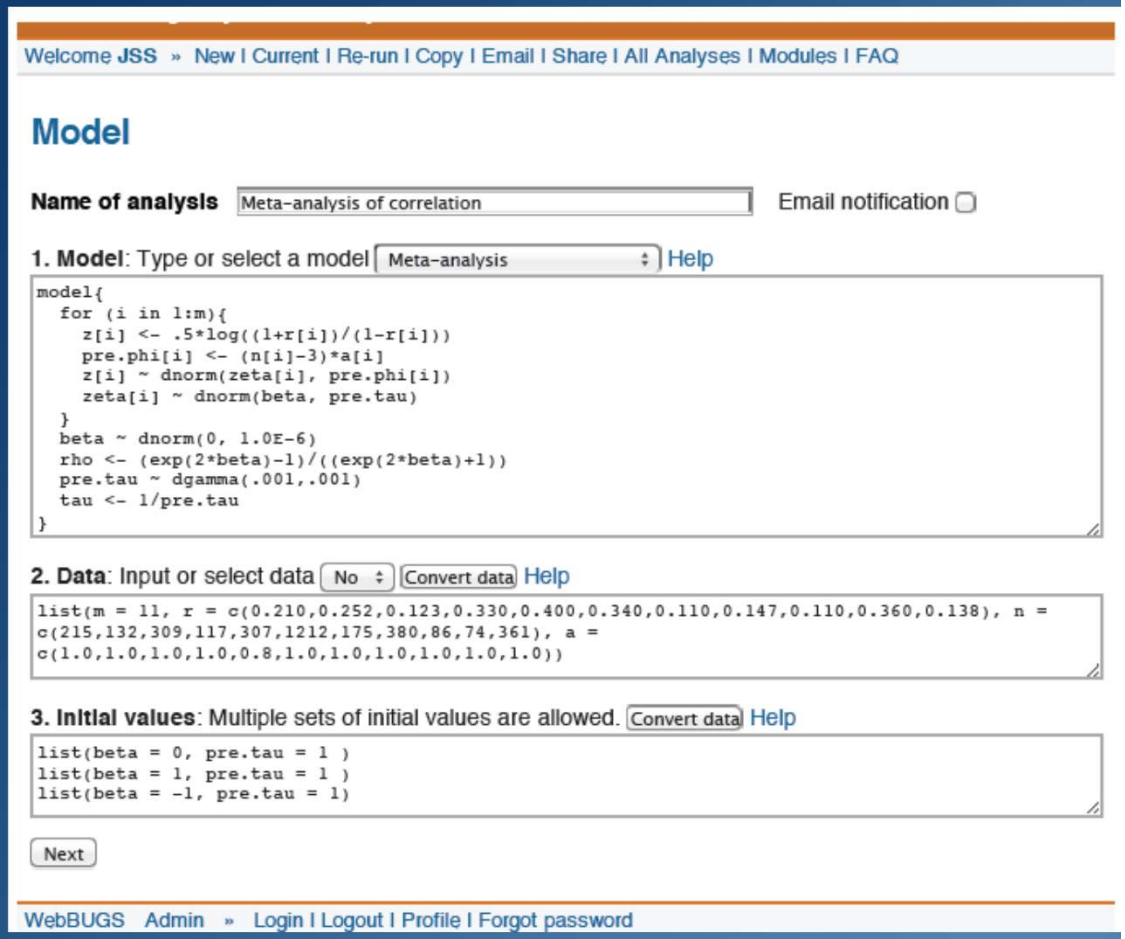


RGUI: Web Based App

WebBUGS

- Conducting Bayesian Statistical Analysis Online
- Combines OpenBUGS and R

www.webbugs.psychstat.org



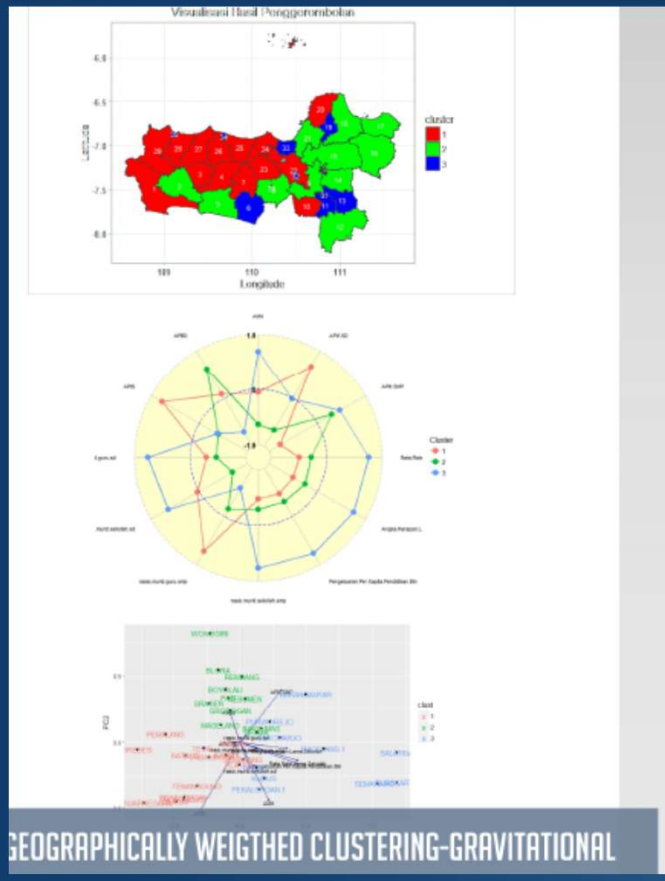
RGUI: Shiny

- A new package from Rstudio to build interactive web applications with R.
- Really Easy!
- Build useful web applications with only a few lines of code—no JavaScript required.
- Self learning: http://shiny.rstudio.com/
- http://www.showmeshiny.com/

Our Recent R Packages

Name	Title	Brief Description	Author	Repository
spatialClust	Spatial Clustering	Clustering analysis with pay attentation on membership via spatial effects	Imam Habib Pamungkas, Setia Pramana	CRAN
advclust	S4 Object Oriented for Advanced Clustering(Fuzzy Clustering and Cluster Ensemble)	Advance on clustering with fuzzy clustering for overlapping cluster and objects on gray area. Cluster Ensemble performs combining several result as one robust and stable result.	Achmad Fauzi Bagus F, Setia Pramana	CRAN
RcmdrPlugin.FuzzyClust	R commander plugin for fuzzy clustering	Graphical User interface via Rcmdr Plugin for fuzzy clustering analysis	Achmad Fauzi Bagus F, Setia pramana	CRAN
MetaheuristicFPA	Metaheuristic with Flower Pollinantion Algorithm	Optimization of function objectives to get global optimum of parameter by using Flower Pollination Algorithm	Amanda Pratama Putra, Margaretha Ari Anggorowati	CRAN
Multiplier	Social Accounting Matrix and Finansial Social Accounting Matrix	Graphical User Interface for performing SAM (Social Accounting Matrix) and FSAM (Financial Social Accounting Matrix)	Tiara Ratna Dewi, Aisyah Fitri Yuniarshi	R-Forge
RcmdrPlugin.PCARobust	Robust PCA plugin for Rcmdr	Graphical User Interface for Robust Principal Component Analysis (PCA) with Hubert Algorithm for Dimension Reduction	Monalisa Sipahutar, Setia Pramana	CRAN

Our Recent R Packages



spatialClust

An R Package for Cluster Spatial Data

Available on CRAN:

September 3rd 2016

"spatialClust" - an R package

by Imam Habib Pamungkas, S.S.T and Setia Pramana, Ph.D

Our Recent R Packages

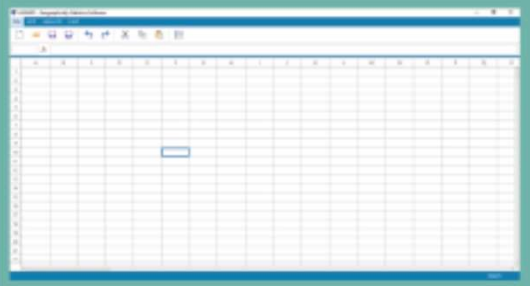
- Kalingga
- Muria
- C++

Asgard Alpha Version

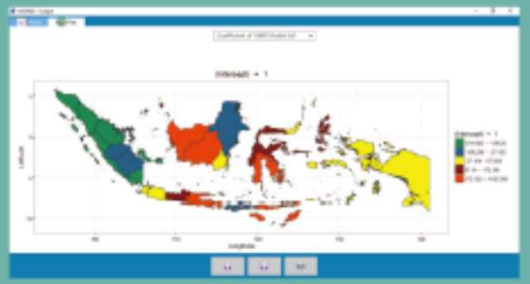
ASGARD is a statistics software used to perform geographically weighted regression (GWR). This software was made in 2016 and currently contains some basic GWR functions like GWR, Geographically Weighted Poisson Regression (GWPR), Geographically Weighted Logistic Regression (GWLR), Geographically Weighted Negative-Binomial Regression (GWNBR) and some Assumption Test related to GWR. In addition, ASGARD is also integrated with the map that make it easier for users to performs analysis.

MAIN FEATURES

Spreadsheet

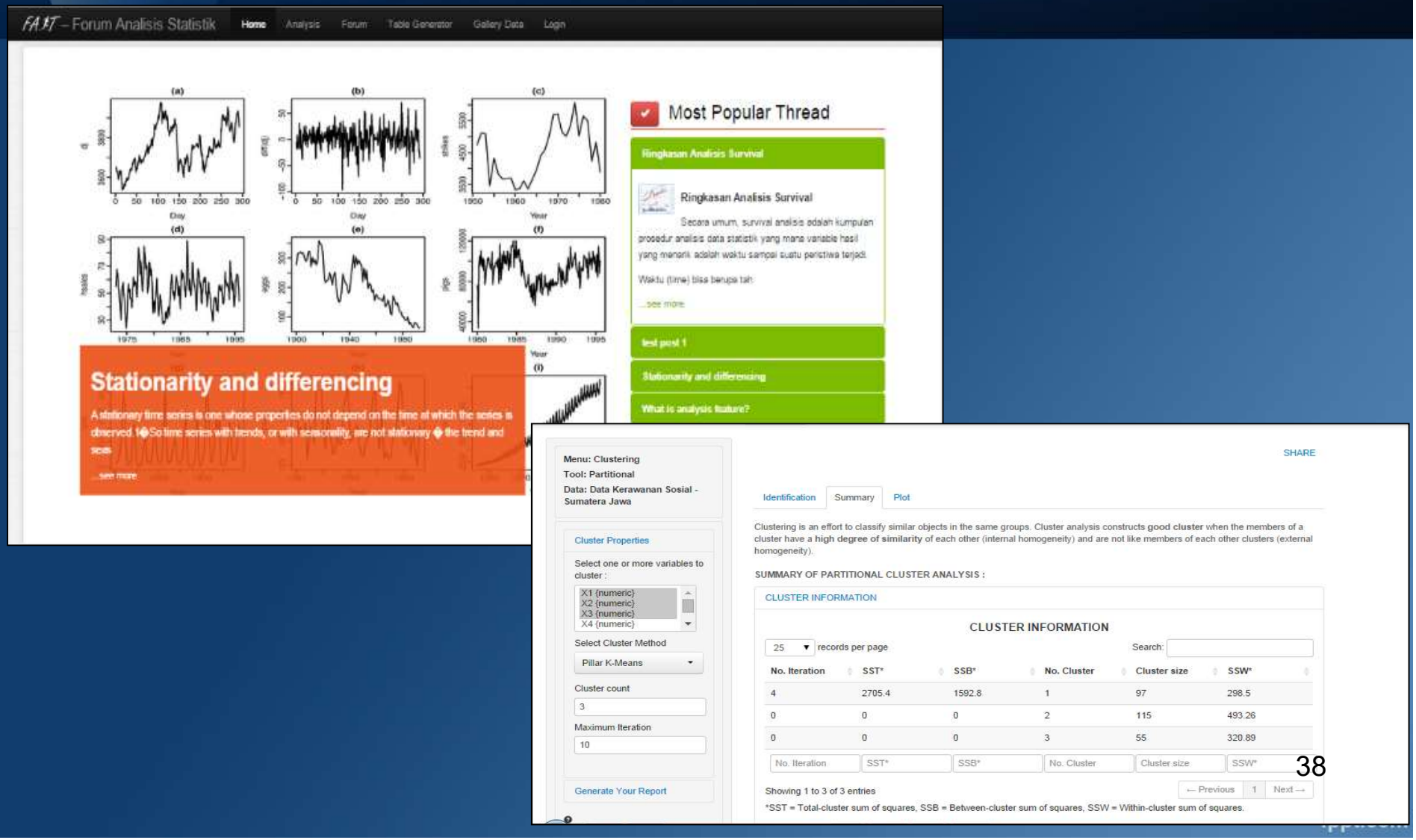


Fairly complete functions
- GWR
- GWPR
- GWLR
- GWNBR
- Variance Inflation Factor
- Breusch-Pagan Test



Map Visualization
Map Visualization can help users to understand the circumstances of the observation area.

RGUI using Shiny: FAST



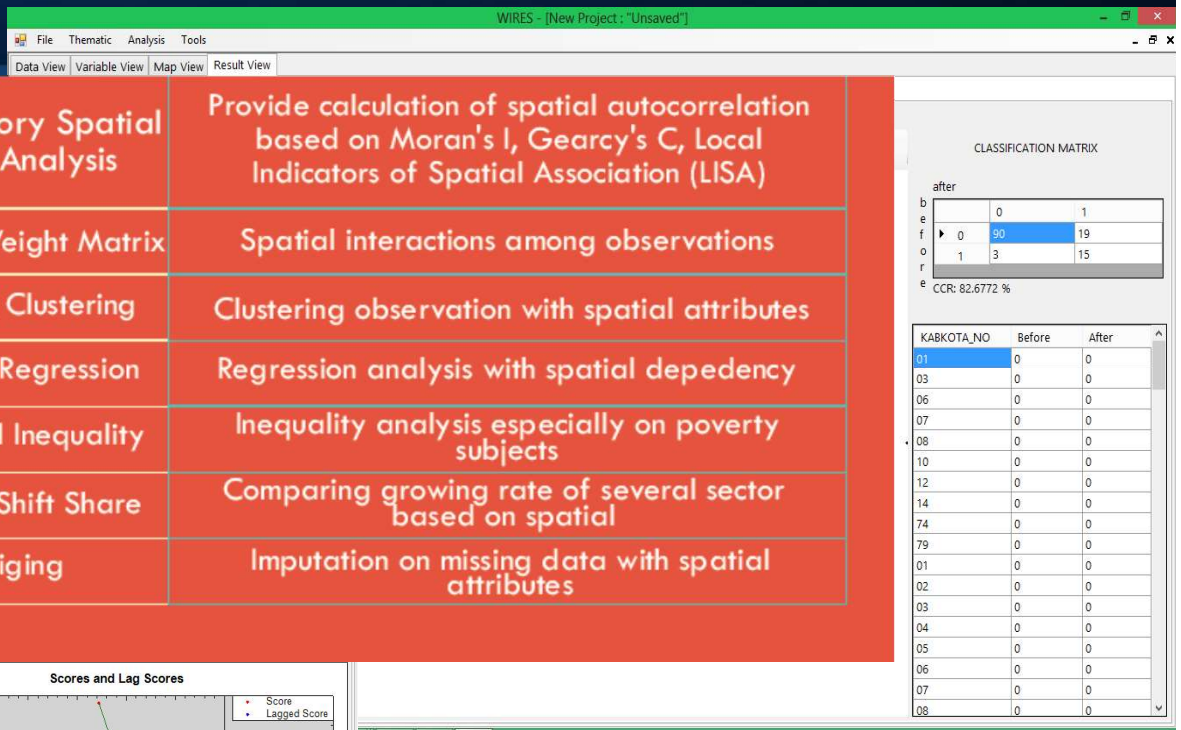
RGUI using C#: Wires

- For Spatial Data Analysis

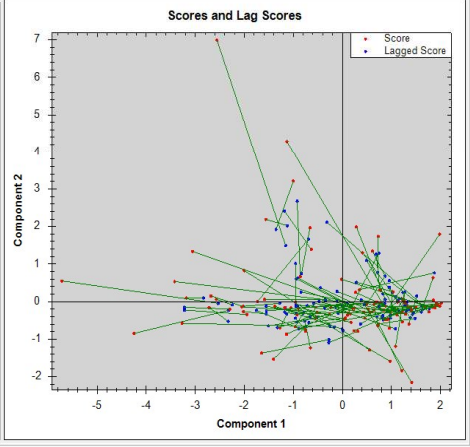
RGUI using C#: Wires

Features

Exploratory Spatial Data Analysis	Provide calculation of spatial autocorrelation based on Moran's I, Gearcy's C, Local Indicators of Spatial Association (LISA)
Spatial Weight Matrix	Spatial interactions among observations
Spatial Clustering	Clustering observation with spatial attributes
Spatial Regression	Regression analysis with spatial depedency
Regional Inequality	Inequality analysis especially on poverty subjects
Spatial Shift Share	Comparing growing rate of several sector based on spatial
Kriging	Imputation on missing data with spatial attributes



	Score Component 1	Lag Score Component 1	Score Component 2	Lag Score Component 2	Inequality
SELAPARANG	-2.5570	-1.1572	6.9862	1.4920	5.6697
SUMBAWA	-1.1325	0.8284	4.2860	-0.0731	4.7798
DOMPU	-3.0534	0.1945	1.3361	-0.1566	3.5745
WOJA	-3.4202	-0.1159	0.5342	0.0115	3.3454
PRAYA	-5.7190	-2.5356	0.5513	-0.0462	3.2390
AIKMEL	-3.2690	-0.2523	-0.5785	-0.6618	3.0178
MOYO UTARA	1.6801	-0.3151	-0.1405	2.1151	3.0114
RHEE	1.9859	0.8254	1.7949	-0.3738	2.4597
SUKAMULIA	0.9716	-0.9446	-1.5866	-0.1341	2.4045
SELONG	-2.0031	-0.1675	0.8245	-0.6772	2.3716
SEMBALUN	1.3885	-0.8893	-0.1614	-0.7276	2.3471
TAMBORA	1.4137	0.4499	-2.1472	-0.0816	2.2794
KURIPAN	0.6826	-1.5507	-0.0123	-0.2446	2.2453
SOROMANDI	1.2874	-0.8441	-0.3371	0.2568	2.2127
PUJUT	-4.2471	-2.3220	-0.8469	-0.2324	2.0208
RABA	0.2869	0.8763	1.9864	0.0585	2.0160
MATARAM	-0.8592	-0.9215	0.6684	2.6770	2.0096
ALAS	0.4113	1.3838	1.3021	-0.4249	1.9820
GERUNG	-2.0509	-0.0880	-0.1390	-0.3074	1.9701
GUNUNG SARI	-1.7310	-0.6951	-0.0028	1.6604	1.9594
KAYANGAN	0.9773	-0.9436	-0.2941	-0.3153	1.9210
TARANO	1.2161	-0.5865	-0.2837	-0.0498	1.8177
TALIWANG	-0.0186	1.6467	0.5946	-0.0904	1.8007



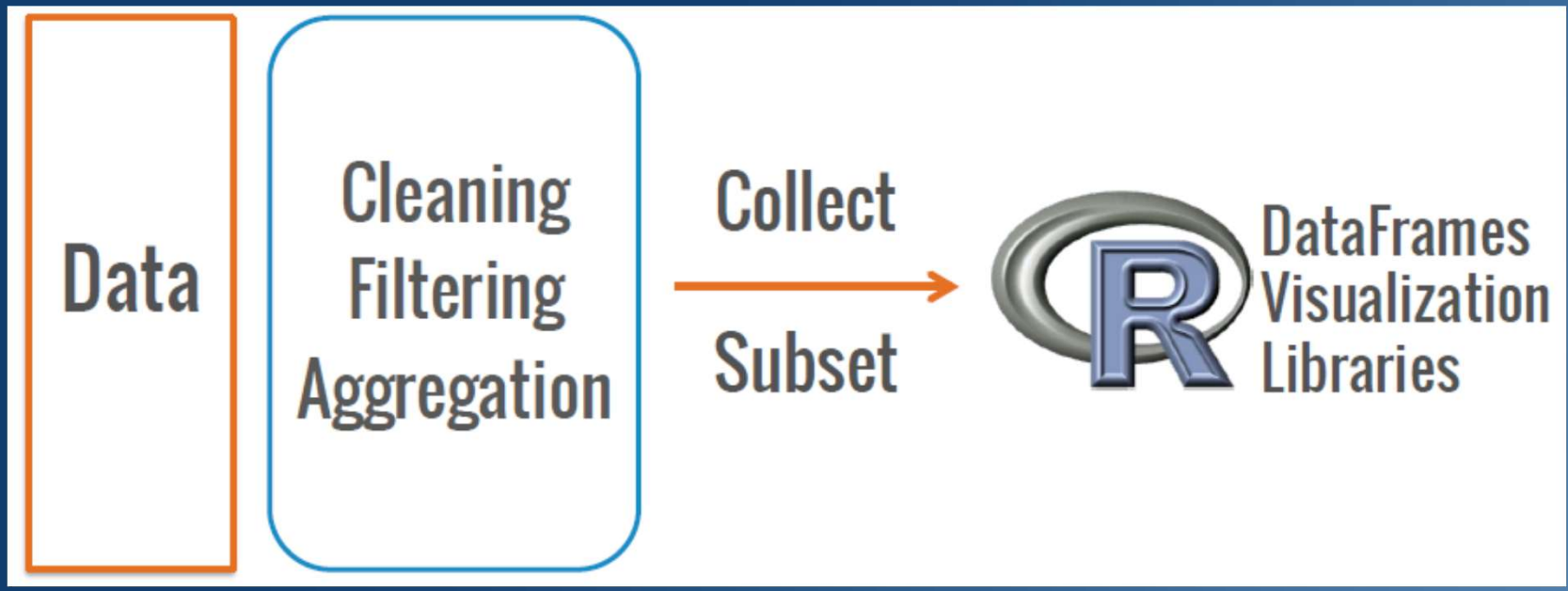
Dynamic Report Generation

Produce documents automatically: pdf, doc, html

Packages:

- Sweave
- knitr
- Markdown

Big Data and R

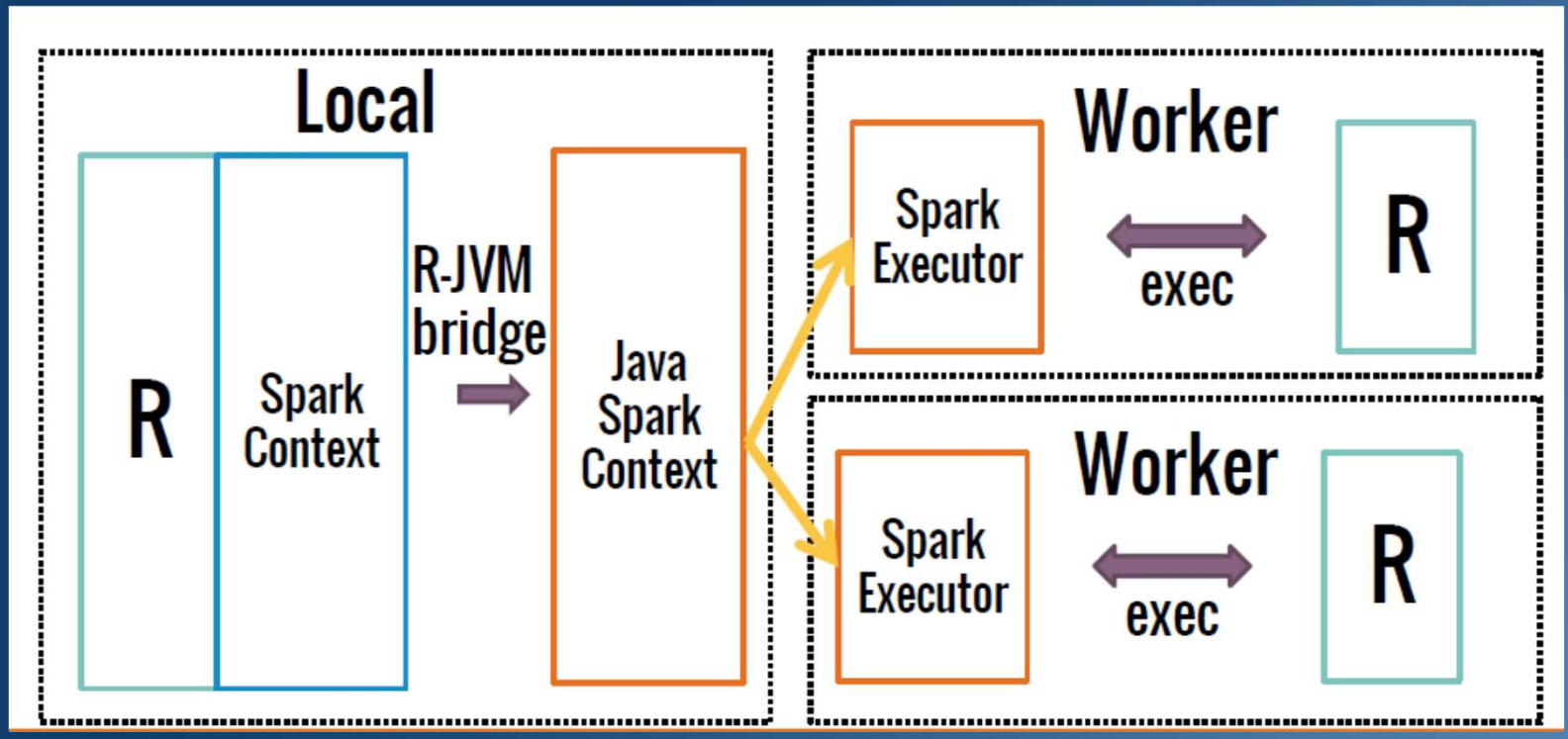


Big Data

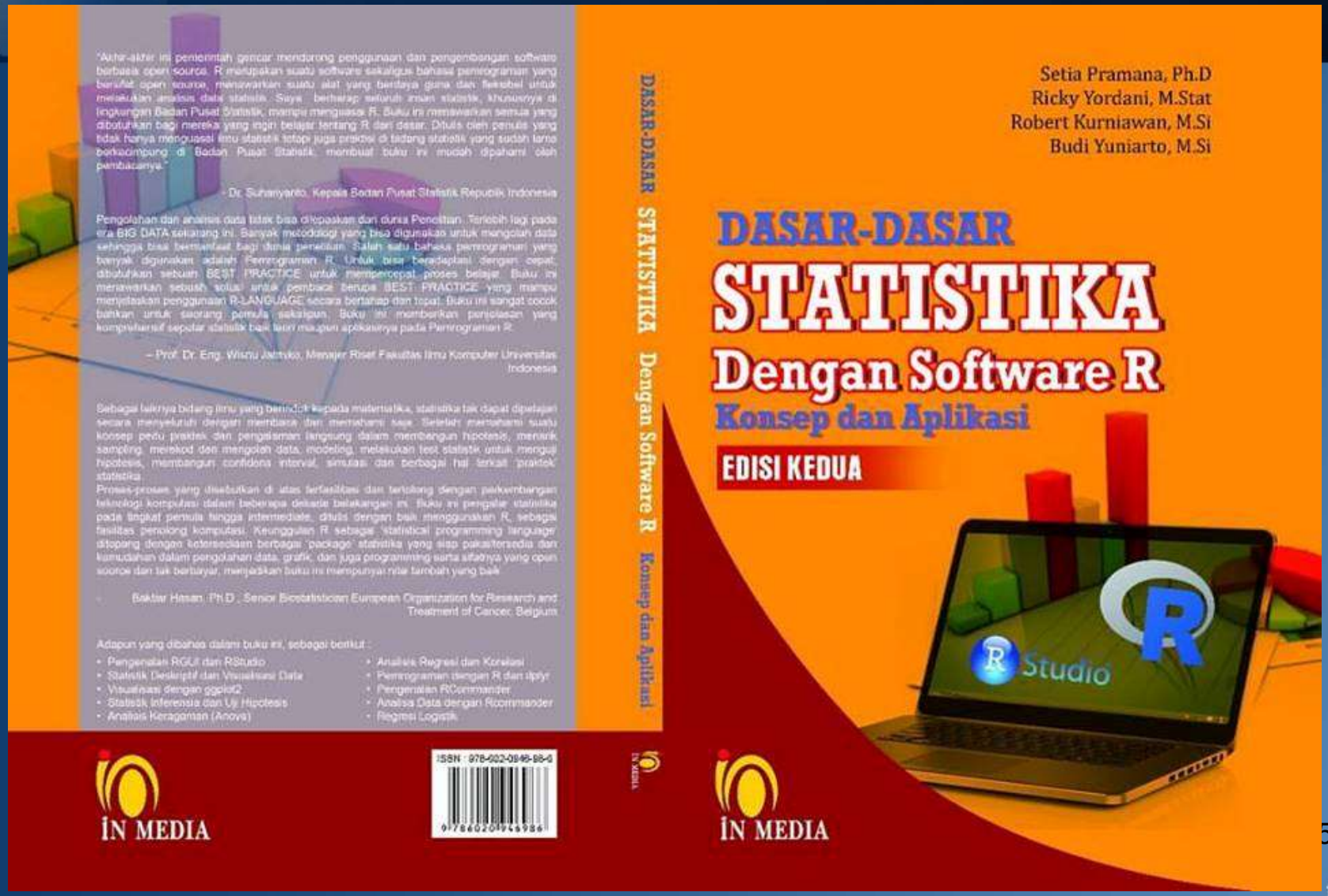
- **SparkR**

SparkR is a language binding that seamlessly integrates R with Spark, and enables native R programs to scale in a distributed setting

SparkR Architecture



A Start



DATA MINING dengan R

Konsep Serta Implementasi

Kebutuhan akan eksplorasi dan analisis data semakin meningkat beberapa tahun terakhir. Metode eksplorasi dan analisis data juga mulai bergeser ke arah penggunaan data mining dan beberapa algoritma machine learning. Hal ini mendorong perubahan kurikulum dan materi yang harus disampaikan dan dikuasai mahasiswa khususnya mahasiswa jurusan statistik. Buku ini sangat saya rekomendasikan baik kepada mahasiswa maupun para pengajar karena buku ini tidak hanya memberikan teori namun juga mengajarkan bagaimana mengaplikasikan teori tersebut dalam contoh-contoh praktis. Buku ini juga memberikan keberagaman aplikasi dari data mining dengan tipe data yang berbeda-beda yang dapat diaplikasikan dengan software R.

Dr. Erni Tri Astuti, M.Math - Direktur Politeknik Statistika STIS

R merupakan salah satu alat pengolahan data yang sangat ampuh. Dengan bahasa yang lugas dan "to-the-point", penulis berhasil menyajikan data mining dengan pendekatan praktis menggunakan R. Buku ini merupakan batu pijakan yang sangat berguna buat para aspiring data scientist yang ingin menggeluti bidang data science

Syafri Bahar S.Si., M.Sc., FRM - Vice President of Data Science GOJEK.

Bahasan buku ini mencakup:

1. Pengantar Data Mining
2. Eksplorasi dan Visualisasi Data
3. Regresi Linear dan Logistik
4. Analisis Komponen Utama
5. Multivariate Anova
6. Supervised Learning (KNN, Decision Tree, Random Forest, dll)
7. Unsupervised Learning (Cluster Analysis)
8. Text Mining
9. Analisis Sentimen
10. Data Mining dalam Bioinformatika

İN MEDIA

UMUM
ISBN 978-602-6469-

Harga P. Jawa Rp.

Setia Pramana, dkk

DATA MINING dengan R

İN MEDIA

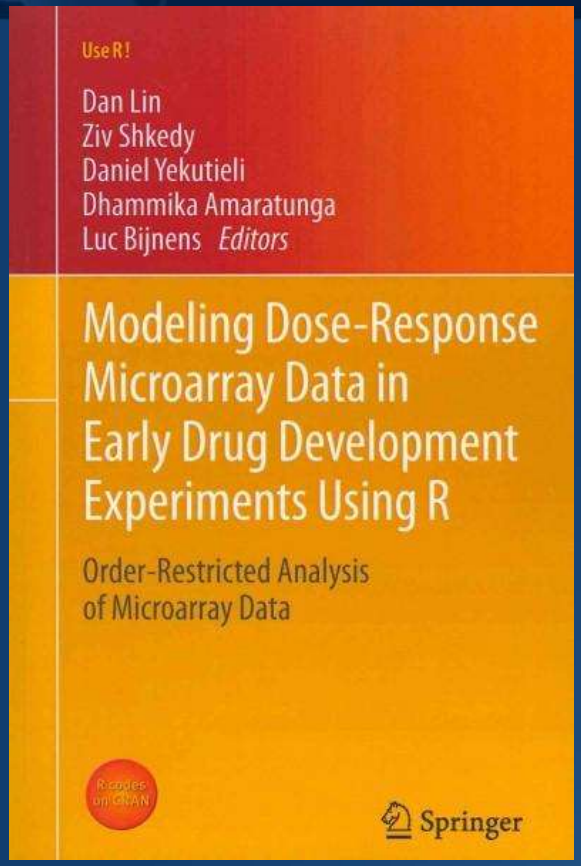
Setia Pramana
Budi Yuniarto
Siti Mariyah
Ibnu Santoso
Rani Nooraeni

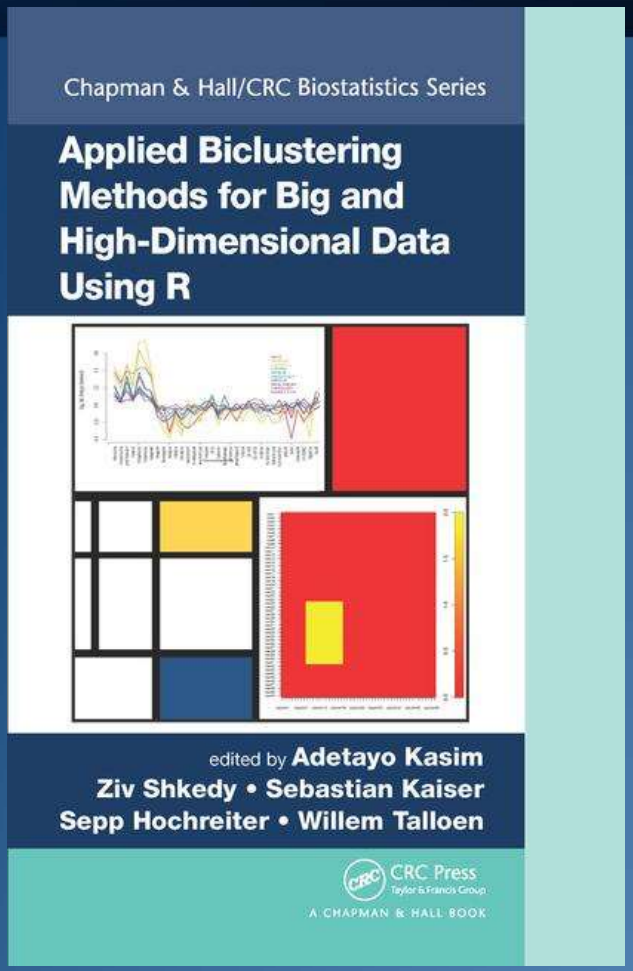
DATA MINING dengan R

Konsep Serta Implementasi

İN MEDIA

Book Chapters





Conclusion

If statistics programs/languages were cars...

Thank you for your attention!