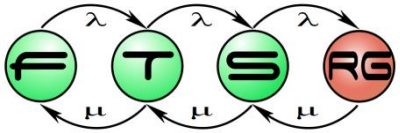
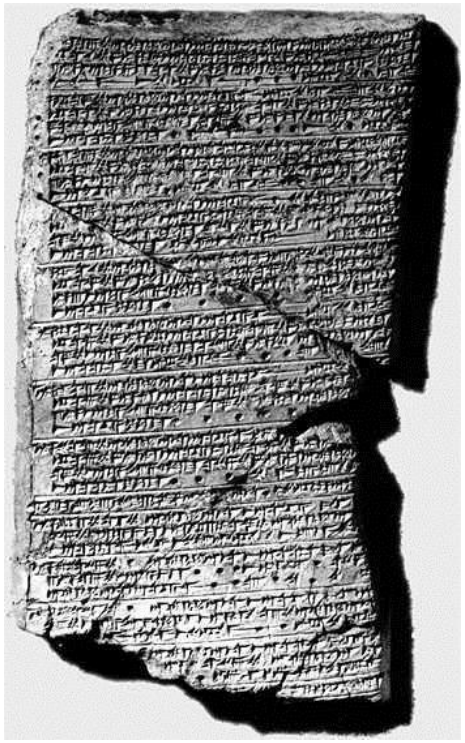
R

„Big Data” elemzési módszerek

Kocsis Imre
ikocsis@mit.bme.hu
2015.09.30.



Adatelemzés (a számítógépig)



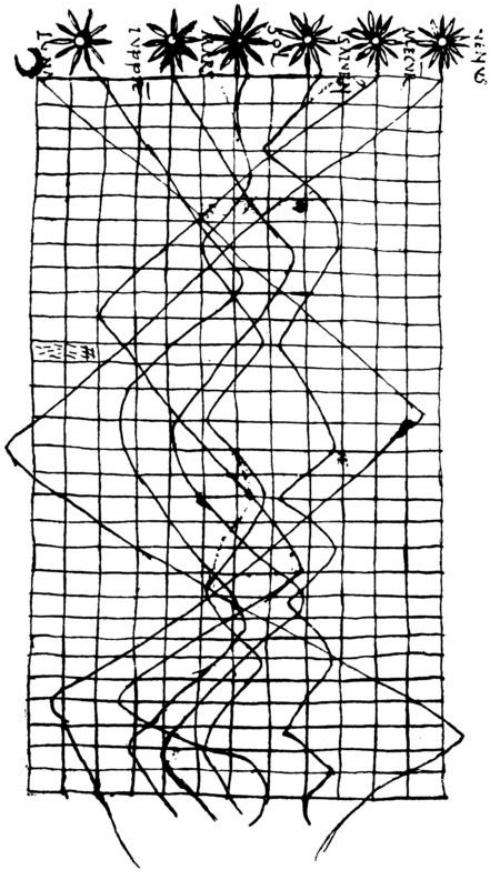
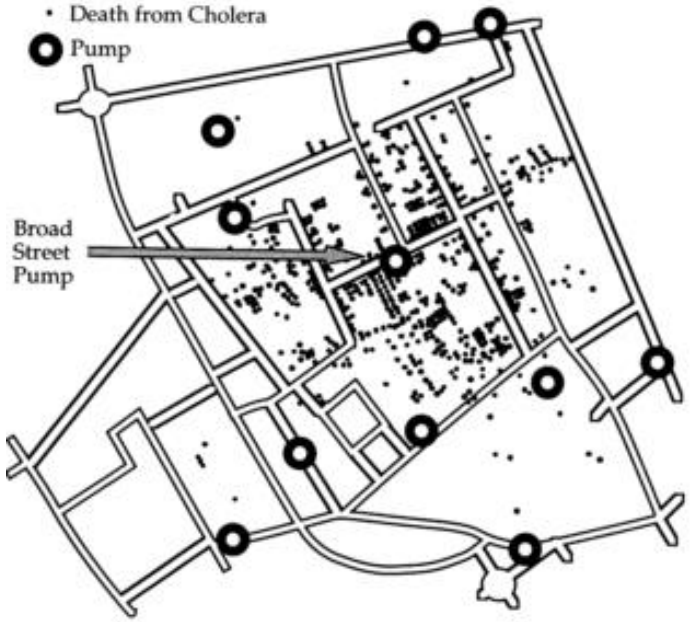


Table 1
Cholera deaths in Golden Square, Soho from August 31 to September 11, 1854 (adapted from Snow, 1855a, p. 49)

Date	No. of fatal attacks	Deaths
August 31	56	3
September 1	143	70
September 2	116	127
September 3	54	76
September 4	46	71
September 5	36	45
September 6	20	37
September 7	28	32
September 8	12	30
September 9	11	24
September 10	5	18
September 11	5	15

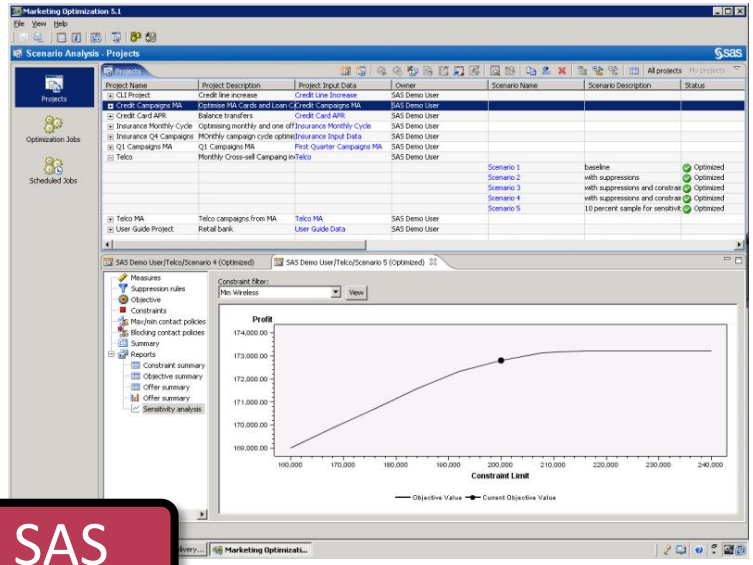


Mi *nem* statisztikai eszköz/csomag?

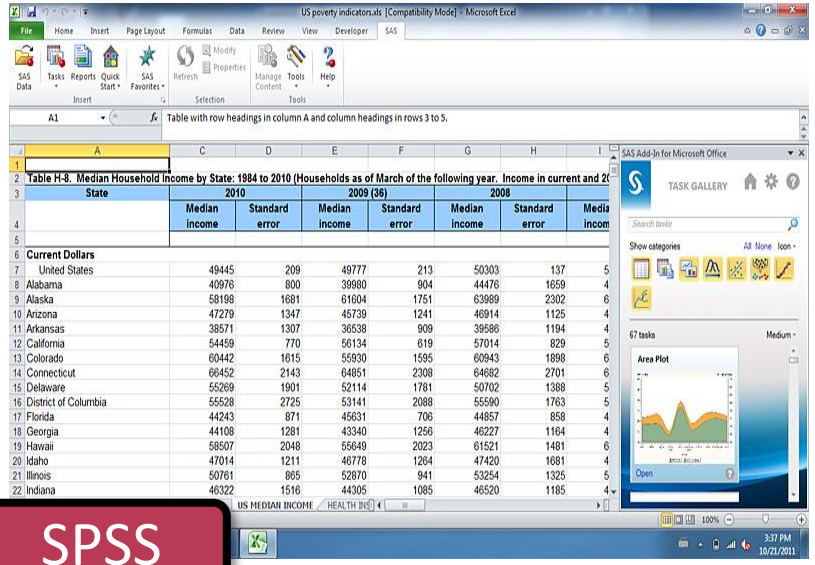
- Táblázatkezelő
  - Lásd pl. [4]
- Adatbáziskezelő
  - SQL
- Saját C/FORTRAN/Perl/Java…
  - EDA…?
  - Stat. függvények?

- Úgy értve, hogy klasszikusan
  - Mindhárom területen változik
- + adatelemzés != statisztika

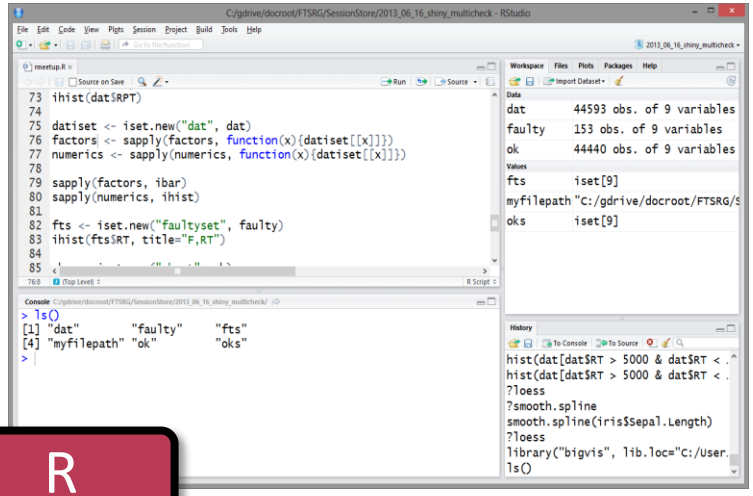
Mi az, ami igen



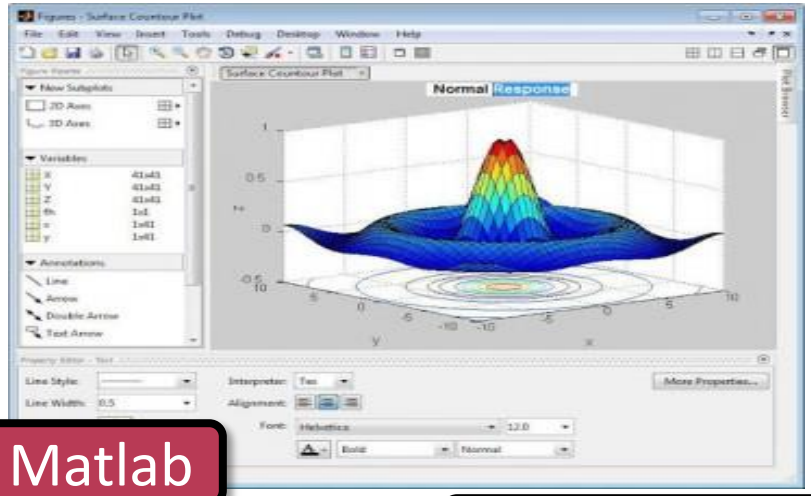
SAS



SPSS



R



Matlab

+ wikipedia [5]

Néhány általános jellemző

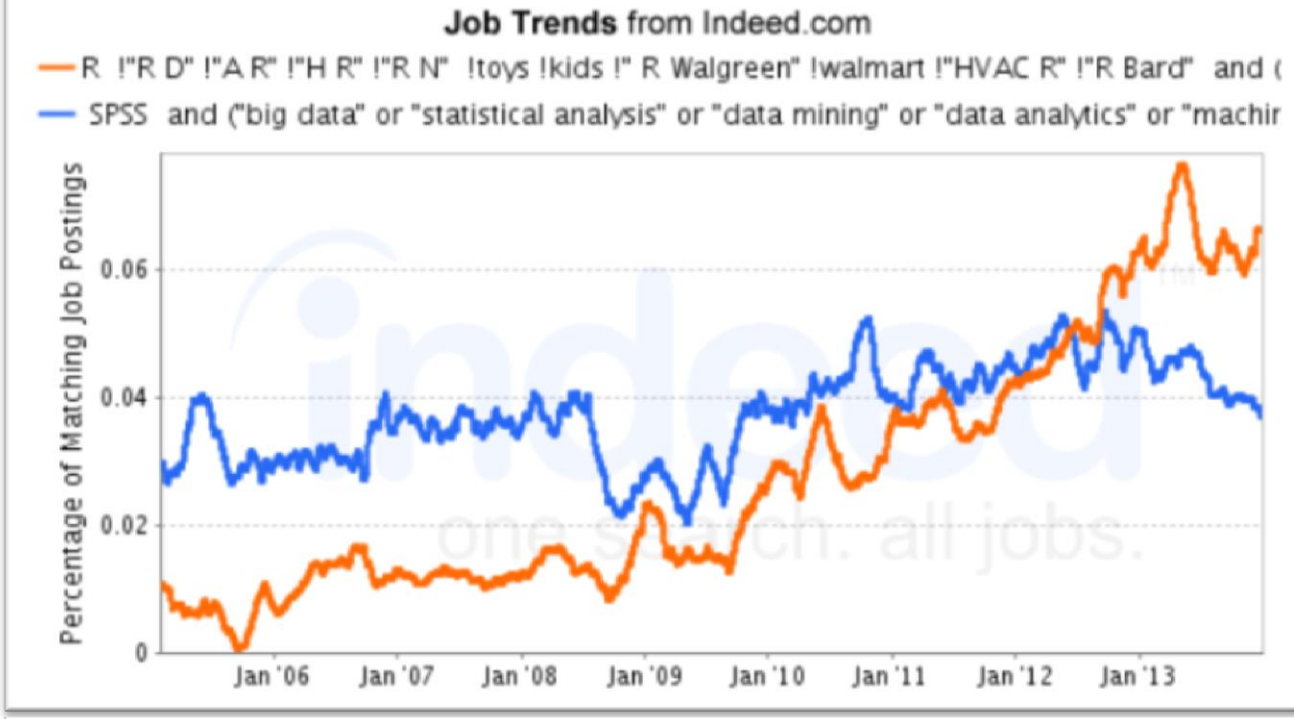
- Saját szkriptnyelv
  - Interaktív futtatással is
- *Validált* stat. eljárások széles köre
  - As in: „clinical trial data for FDA submissions”
- „Workspace” modell
  - Jellemzően in-memory (vs. „out-of-memory” elemzés)
- Erős vizualizációs képességek
- Kapcsolódó funkciók
  - jelentések, adatbázis-kapcsolat, GUI-szkriptelés, webalkalmazások, munkafolyamatok, etc.
- Gyökerek: 70-es évektől →…
  - SAS Institute: 1976, az egyetemmel szemben
  - Szoftvertechnológiailag erősen látszik; az új generáció már más
  - http://julialang.org/blog/2012/02/why-we-created-julia/

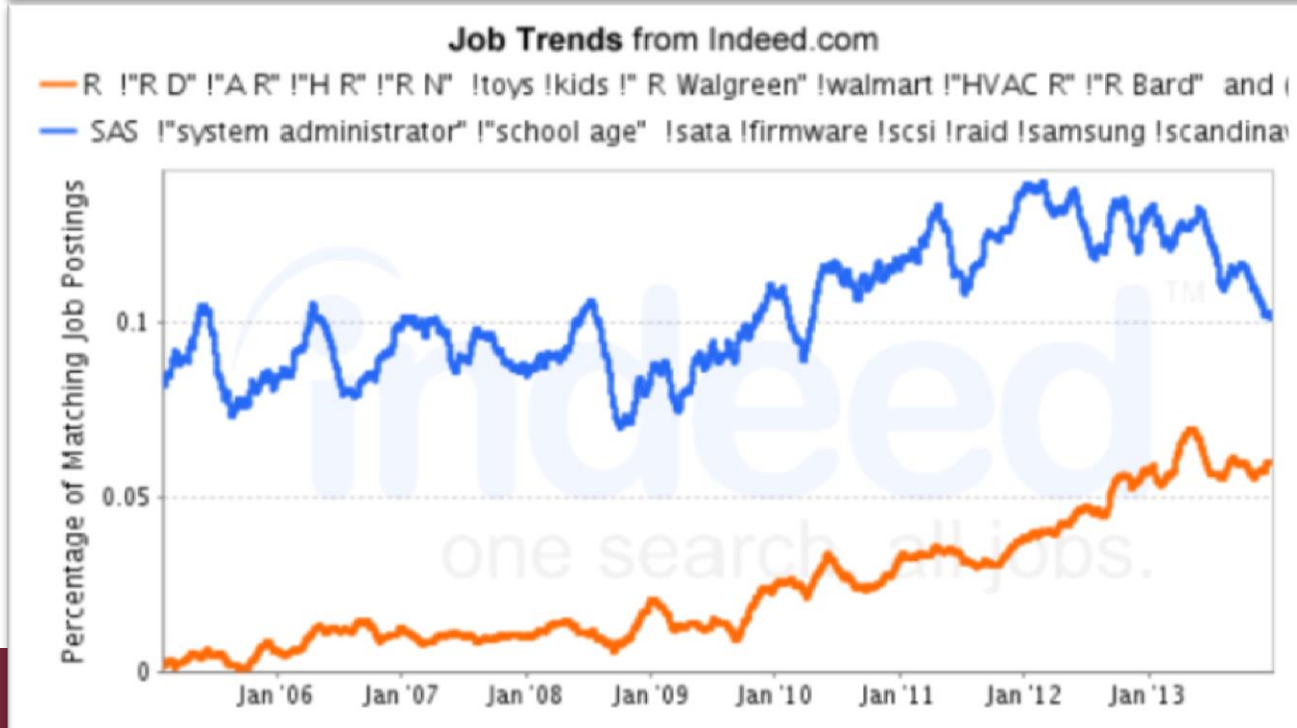
R

- Az S nyelv „GNU verziója”
  - Környezet és nyelv egyben

- Statisztikai számítások és grafika
- Nem csak ingyenes; *nyílt* is
- Hatékonyság: „kihívás” C/C++/FORTRAN-ba
- Egyre inkább „lingua franca”, ha adatot kell elemezni
  - + Python

Miért R? (r4stats.com)





Miért R? (r4stats.com)

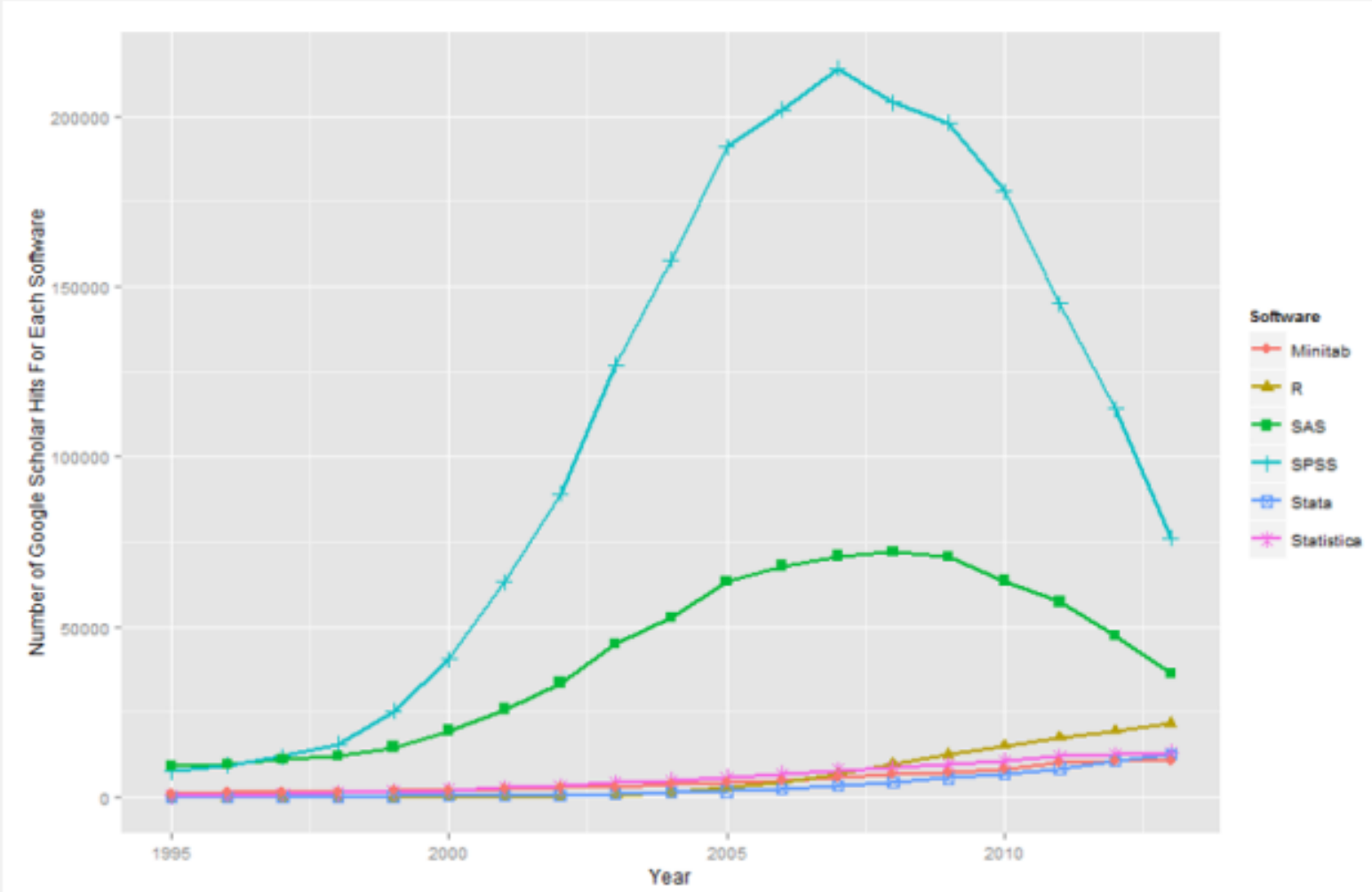


Figure 2b. Number of scholarly articles found for the top five classic statistics packages.

Miért R? (r4stats.com)

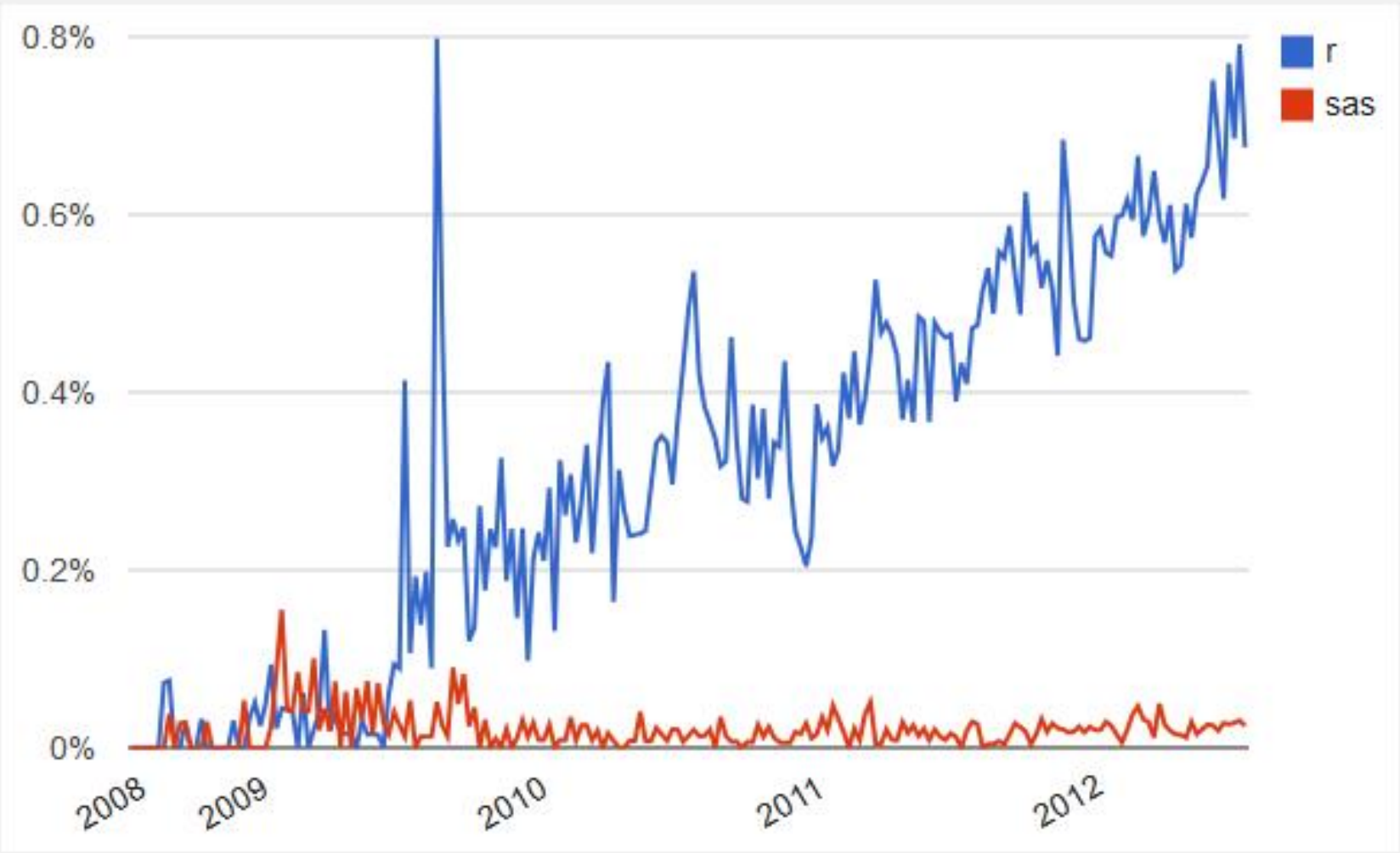


Figure 1c. Number of R- or SAS-related posts to Stack Overflow by week.

Forrás: [1]

Miért R? (r4stats.com)

- Mert HF

Forrás: [1]

cran.r-project.org

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 (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

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 (2013-09-25, Frisbee Sailing) R-3.0.2.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.

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
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
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
SpatioTemporal	Handling and Analyzing Spatio-Temporal Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

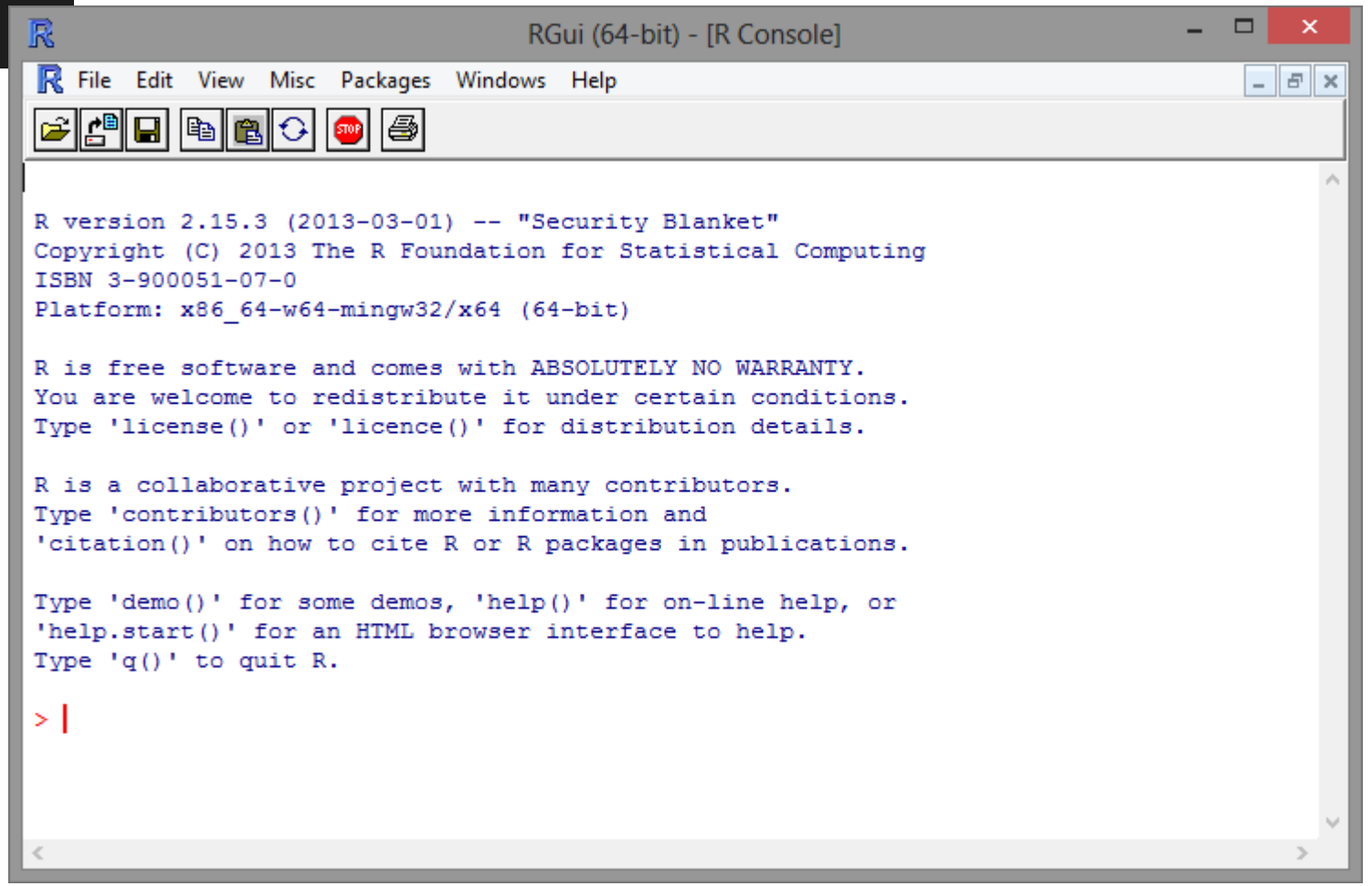
Available Packages

Currently, the CRAN package repository features 5898 available packages.

+ GitHub, BioC, R-Forge, saját, …

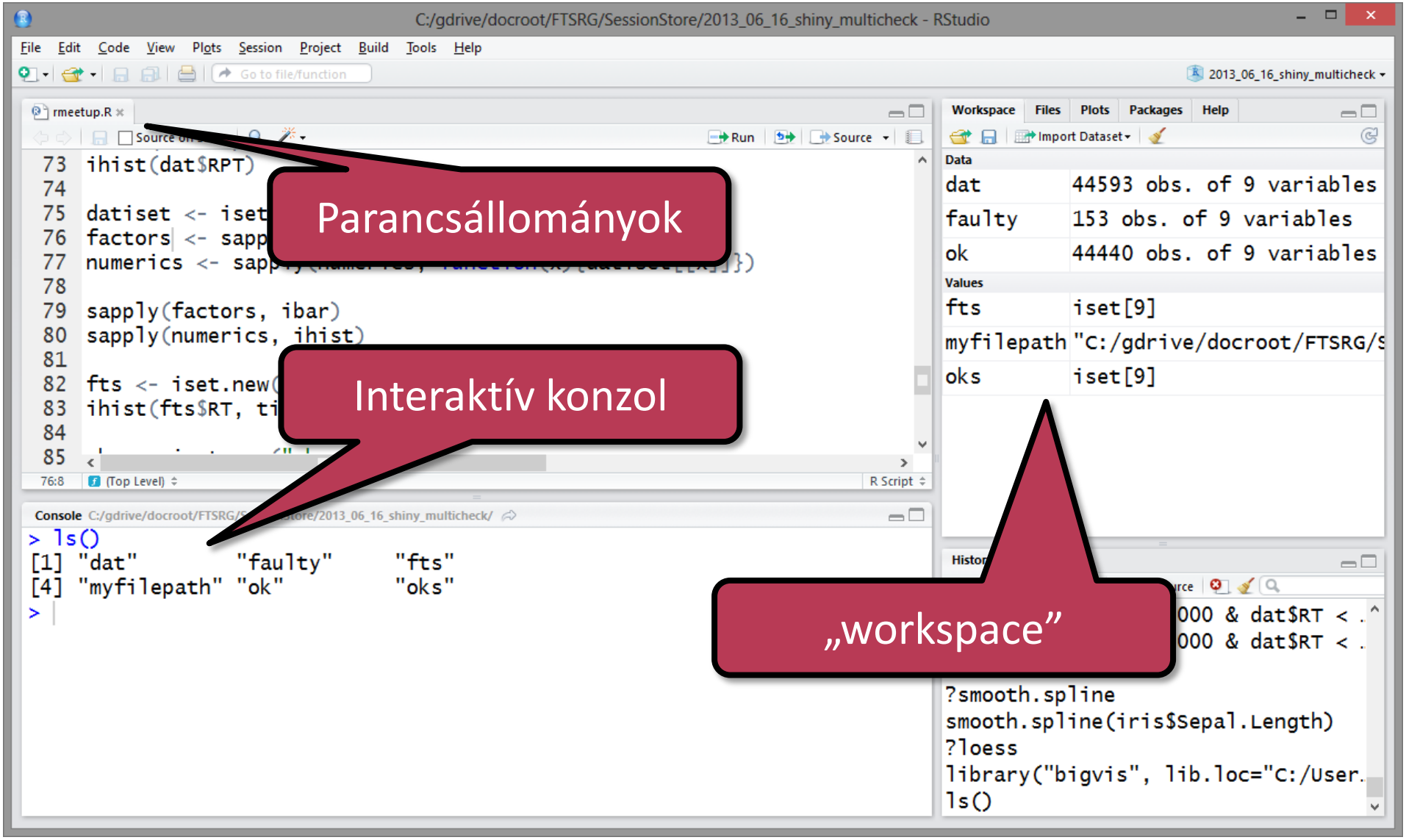
R konzol


R i386 2.15.3
R x64 2.15.3


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>

...

RStudio


C:/gdrive/docroot/FTSRG/SessionStore/2013_06_16_shiny_multicheck - RStudio
File Edit Code View Plots Session Project Build Tools Help
2013_06_16_shiny_multicheck
rmeetup.R
Run
Source
ihist(dat$RPT)
datiset <- iset
factors <- sapp
numerics <- sapp
sapply(factors, ibar)
sapply(numerics, ihist)
fts <- iset.new(
ihist(fts$RT, ti
76:8
(Top Level)
R Script
Parancsállományok
Interaktív konzol
Console C:/gdrive/docroot/FTSRG/
/2013_06_16_shiny_multicheck/
> ls()
[1] "dat" "faulty" "fts"
[4] "myfilepath" "ok" "oks"
>
Workspace Files Plots Packages Help
Import Dataset
Data
dat 44593 obs. of 9 variables
faulty 153 obs. of 9 variables
ok 44440 obs. of 9 variables
Values
fts iset[9]
myfilepath "C:/gdrive/docroot/FTSRG/
oks iset[9]
„workspace”
?smooth.spline
smooth.spline(iris$Sepal.Length)
?loess
library("bigvis", lib.loc="C:/User
ls()

Ismerkedés az R-rel

- Interaktív bevezetés az R nyelvbe és környezetbe példákon keresztül
- Rintro.R
- Induláshoz javasolt:
  - FTSRG tech cheat sheet [6]
  - Magyarul: [2] és [3]
- N.B.: nem kell hozzá informatikusnak lenni
  - Előny és hátrány is

typeof

1. táblázat. Fontosabb `typeof` visszatérési értékek

érték	jelentése
NULL	Null
symbol	változó neve
closure	függvény
logical	logikai értékekből álló vektor
integer	egész számokból álló vektor
double	lebegőpontos számokból álló vektor
complex	komplex adatokból álló vektor
character	karaktervektor
list	lista
raw	bináris vektor

Forrás: [2]

www.meetup.com/Budapest-Users-of-R-Network

Find a Meetup Group Start a Meetup Group

Budapest Users of R Network

Home | Members | Sponsors | Photos | Discussions | More | My profile

Budapest, Hungary

Founded Aug 11, 2013

About us...

useRs 216

Group reviews 2

> print('Hello, Hungary!')

+ SUGGEST A NEW MEETUP

Upcoming 1 | Suggested 0 | Past | Calendar

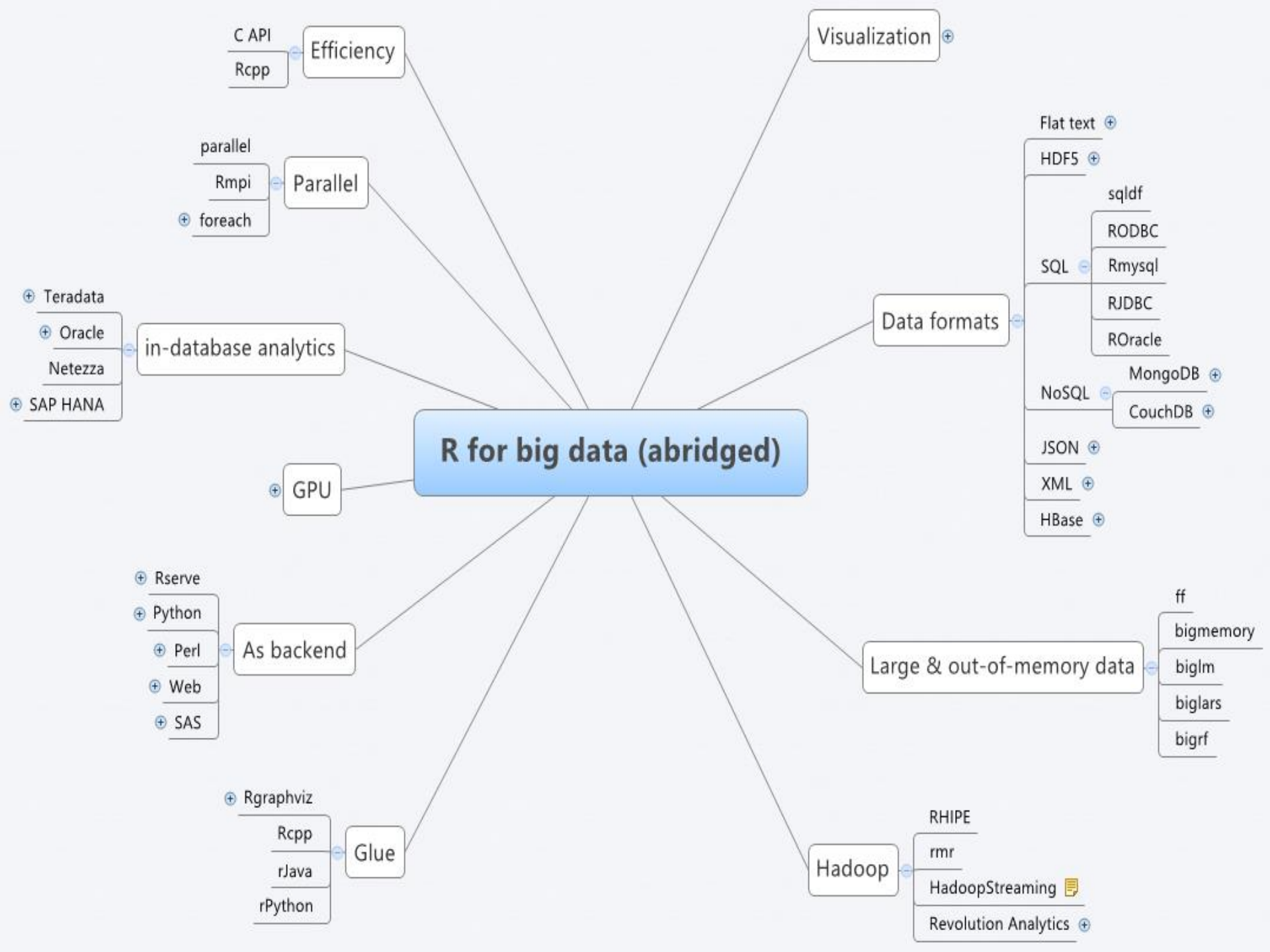
R integráció kereskedelmi és egyéb termékekben

BME I. épület, I.B 019. terem (földszinten a porta mögött)
BME I épület, Budapest, Magyar tudósok körútja, Hungary, Budapest (map)

Mon Oct 13
6:00 PM
✓ I'M GOING
39 going
1 comment

This will be a Hungarian speaking event,

What's new


R for big data (abridged)
Efficiency
C API
Rcpp
Visualization
Parallel
parallel
Rmpi
foreach
Data formats
Flat text
HDF5
SQL
sqldf
RODBC
Rmysql
RJDBC
ROracle
NoSQL
MongoDB
CouchDB
JSON
XML
HBase
in-database analytics
Teradata
Oracle
Netezza
SAP HANA
GPU
As backend
Rserve
Python
Perl
Web
SAS
Large & out-of-memory data
ff
bigmemory
biglm
biglars
bigrf
Glue
Rgraphviz
Rcpp
rJava
rPython
Hadoop
RHIPE
rmr
HadoopStreaming
Revolution Analytics

Hivatkozások

- [1] http://r4stats.com/articles/popularity/
- [2] http://cran.r-project.org/doc/contrib/Solymosi-Rjegyzet.pdf
- [3] http://www.inf.unideb.hu/~jeszy/R/
- [4] http://people.umass.edu/evagold/excel.html
- [5] http://en.wikipedia.org/wiki/List_of_statistical_packages
- [6] https://github.com/FTSRG/technology-cheat-sheets/wiki/R-programming-language