R
A software environment for comprehensive statistical analysis of astronomical data

Eric Feigelson
Center for Astrostatistics
Penn State University

Data Intensive Astronomy, IAU General Assembly
Beijing 2012

Brief history of statistical computing

1960s – c2003: Statistical analysis developed by academic statisticians, but implementation relegated to commercial companies (SAS, BMDP, Statistica, Stata, Minitab, etc).

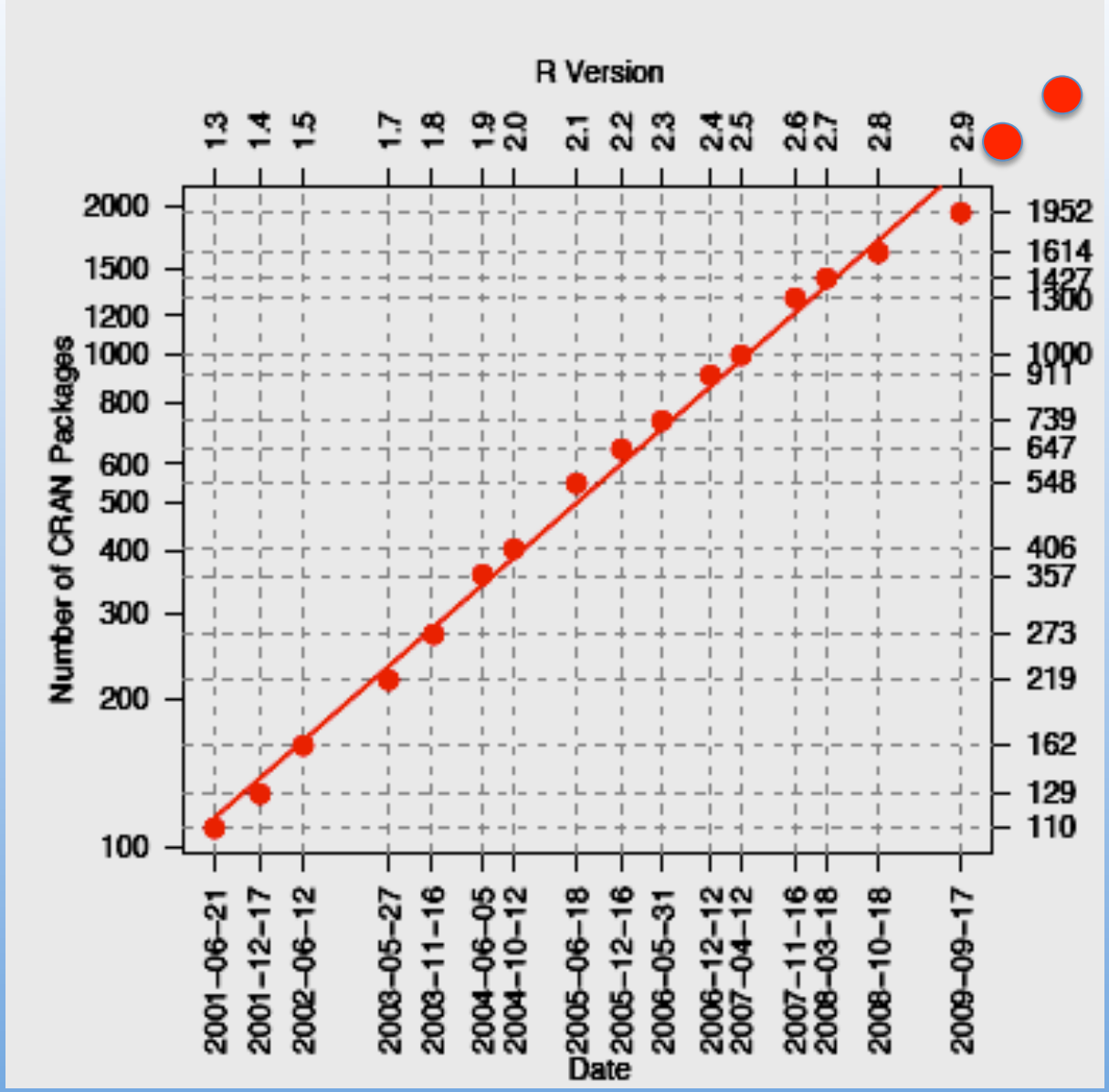
1980s: John Chambers (ATT) develops **S** system, C-like command line interface.

1990s: Ross Ihaka & Robert Gentleman (Univ Auckland NZ) mimic **S** in an open source system, **R**. Expands to ~15 Core Team members, GNU GPL release.

2000s: **Comprehensive R Analysis Network (CRAN)** for user-provided specialized packages grows exponentially. ~20 early packages incorporated into base-R.

By ~2005, **R/CRAN** is the dominant public statistical software system for the development and promulgation of new statistical methodology. Used extensively by statisticians and many user communities (genomics, econometrics, ecology). Estimated 2M users (2010)

Growth of CRAN contributed packages



Aug 24, 2012 count:

4,001 packages

Aspects of the social organization and trajectory of the R project, J. Fox, ***The R Journal***, 1/2, 5 (2009)

The R statistical computing environment

- **R** integrates data manipulation, graphics and extensive statistical analysis. Uniform documentation and coding standards. Quality control is limited outside of base-**R**.
- Fully programmable C-like language (similar to IDL). Specializes in vector or matrix inputs.
- Easily downloaded from http://www.r-project.org for Windows, Mac or linux.
- Many resources: **R** help files (3500p for base **R**), on-line tutorials, >100 books, *Use R!* conferences, *The R Journal* & *J. Stat. Software*
- 4000+ user-provided add-on **CRAN** packages
- Difficulties: Finding what you want, and understanding what you find. Improved education in statistics addresses the latter problem.

Some broad topics covered by base-R

arithmetic & linear algebra
bootstrap resampling
empirical distribution tests
exploratory data analysis
generalized linear modeling
graphics
robust statistics
linear programming
local and ridge regression
maximum likelihood estimation
multivariate analysis
multivariate clustering
neural networks
smoothing
spatial point processes
statistical distributions
statistical tests
survival analysis
time series analysis

Selected methods in Comprehensive R Archive Network (CRAN)

Bayesian computation & MCMC, classification & regression trees, genetic algorithms, geostatistical modeling, hidden Markov models, irregular time series, kernel-based machine learning, least-angle & lasso regression, likelihood ratios, map projections, mixture models & model-based clustering, nonlinear least squares, multidimensional analysis, multimodality test, multivariate time series, multivariate outlier detection, neural networks, non-linear time series analysis, nonparametric multiple comparisons, omnibus tests for normality, orientation data, parallel coordinates plots, partial least squares, periodic autoregression analysis, principal curve fits, projection pursuit, quantile regression, random fields, random forest classification, ridge regression, robust regression, self-organizing maps, shape analysis, space-time ecological analysis, spatial analyisis & kriging, spline regressions (MARS, BRUTO), tessellations, three-dimensional visualization, wavelet toolbox

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

Task Views provide brief overviews of CRAN packages by topic & functionality. Maintained be expert volunteers, updated regularly

- Bayesian ~100 packages
- ChemPhys ~70 packages
- Cluster ~110 packages
- Graphics ~40 packages
- High Performance Computing ~60 packages
- Machine Learning ~60 packages
- Medical Imaging ~15 packages
- Robust ~20 packages
- Spatial ~110 packages
- Survival ~140 packages
- TimeSeries ~90 packages

Interfaces: BUGS, C, C++, Fortran, Java, Perl, Python, Xlisp, XML
(This is very important for astronomical programmers. R scripts can ingest subroutines from these languages. Two-way communication for C, Fortran, Python & Ruby: you can ingest R functions in your legacy codes.)

I/O: ASCII, binary, bitmap, cgi, FITS, ftp, gzip, HTML, SOAP, URL

Graphics & emulators: Grace, GRASS, Gtk, Matlab, OpenGL, Tcl/Tk, Xgobi

Math packages: GSL, Isoda, LAPACK, PVM

Text processor: LaTeX

Some features of R

- Designed for individual use on workstation, exploring data interactively with advanced methodology and graphics. ***But*** it can be used for automated pipeline analysis. Very similar experience to IDL.

- Designed for using one CRAN package at a time. ***But*** packages like *Rattle* (for data mining) and *parallel* (for multicore computing) combine related packages into an integrated environment.

- Designed for static canvas graphics. ***But*** many extensions to interactive, 3D, tree graphics, SVG, RGTK2, Java, and other GUIs & devices. See huge graphics gallery at http://www.oga-lab.net/RGM2.

- Uni- or bi-directional interfaces to other languages: BUGS, C, C++, Fortran, Java, JavaScript, Matlab, Python, Perl, Xlisp, Ruby.

- Only a few astronomy **CRAN** package to date, including FITS I/O.

VOStat (http://vostat.org)

A tiny subset of R/CRAN is implemented as a Virtual Observatory Web service with SAMP communication with other VO data acquisition & analysis tools. Outputs results, plots, R script.

- **Data input** (ASCII/URL/SAMP, select cases & variables)
- **Plots/summaries** (boxplot, histograms, 2D/3D scatter plots)
- **Density estimation** (histogram, ASH, kernel smoothing, parametric fits)
- **Goodness-of-fit** (tests of normality, KS & AD tests, chi-square test)
- **Hypothesis tests** (t test, z test, paired t test, signed rank test)
- **Regression** (linear fitting, NW and LOESS local regression, quantile regression, robust regression, semi-parametric regression)
- **Multivariate analysis** (mult regression, PCA, hierarchical & normal mixture clustering)
- **Spatial analysis** (autocorrelation variogram, k-NN, Ripley's K, Voronoi tess)
- **Directional data** (plots, kernel smooth, von Mises distribution fit & tests)
- **Survival analysis** (Kaplan-Meier estimator, 2-sample tests, LBW estimator)
- **Time series analysis** (autocorrelation, autoregressive models, periodogram)

Computational aspects of R

Vector/matrix functionalities are fast (like C)

e.g. a million random numbers generated in 0.1 sec, a million-element FFT in 0.3 sec on MacBook pro.

Some **R** functions are much slower

e.g. *for (i in 2:1000000) x[i] = x[i-1] + 1*

The **R** compiler is now being rewritten from `parse tree' to `byte code' (similar to Java & Python) leading to several-fold speedup.

While designed for an individual exploring small datasets, R can be pipelined and can treat megadatsets

Some high-performance computing packages

- Many packages for **parallel computing communication**: Message Passing Interface, netWorkSpaces, SOCK, PVM, Open MP, …

```
#  Example: generating 5 random numbers in parallel
library(doParallel)  ;  parallelcl <- makeCluster(5) ;   registerDoParallel(cl)
x <- foreach(i = 1:5) %dopar% {i + runif(1)}
unlist(x)
```

Result: 1.539 2.357 3.499 4.583 5.172

Can treat nested and/or conditional loops. *foreach* is a new looping construct for parallel execution from the company RevolutionAnalytics. Similar to Python's *list comprehensions* including filtering some evaluations.

- Process management on **multicore computers & clusters**: *multicore, batch, condor*:

- Distributed computing of **multiple MCMC chains**: *bugsparallel*

- Several packages for **cloud & grid processing** …
  - *cloudRmpi:* parallel processing using MPI on Amazon's EC2 cloud
  - *GridR*: executes R functions on remote hosts, clusters or grids
  - *rmr* and *RHIPE*: executes R functions via MapReduce on a Hadoop cluster

- Several packages treat **large memory & out-of-memory data** …
  - *bigmemory*: points to very large objects in memory
  - *ff:* fast access to large data on disk
  - *HadoopStreaming*: R operations on streaming tabular or string data

- Several packages provide **GPU computing** …
  - *gputools*: CUDA implementation of linear modeling, clustering, SVM, ICA, …
  - *magma*: matrix algebra for multicore CPU & GPU architectures
  - *OpenCL*: R interface to OpenCL for heterogeneous CPU/GPU systems

See CRAN Task View on High Performance Computing

http://cran.r-project.org/web/views/HighPerformanceComputing.html

R training for astronomers

Summer Schools:

- Annual school by Center for Astrostatistics, Penn State University
- Biennial summer school by Indian Institute of Astrophysics, Bangalore
- Occasional training sessions in U.S., Brazil, Greece, China, …

Books:

- **Modern Statistical Methods for Astronomy with R Applications**
  E. Feigelson & G. J. Babu Cambridge Univ Press, August 2012
- **An Introduction to R** (http://www.r-project.org, one of many free tutorials)
- **>100 other books on R (none yet specializing on HPC)**

Statistical computing Discussion forum, Q&A, astronomical CRAN packages:

- **Astrostatistics & Astroinformatics Portal (http://asaip.psu.edu)**
- **IAU Working Group in Astrostatistics & Astroinformatics**
  (accepted Monday by Commission 5)

Conclusions

Astronomers need no longer be frustrated by unavailability of code for advanced statistical analysis of complex data.

R/CRAN implements a huge range of methods in a friendly, capable, integrated software environment.

R can be used interactively or called from Python or C.

R/CRAN can operate in a HPC environment.

The next challenge is intellectual: What methods best address the scientific issues under study?

The goal of science is to unlock nature's secrets. ... Our understanding comes through the development of theoretical models which are capable of explaining the existing observations as well as making testable predictions. ... Fortunately, a variety of sophisticated mathematical and computational approaches have been developed to help us through this interface, these go under the general heading of statistical Inference. (P.C. Gregory, 2005)