Introduction to R

Eric Feigelson (Penn State)

Summer School in Statistics for Astronomers
June 2022

Prelude to R ….

A brief history of statistical computing

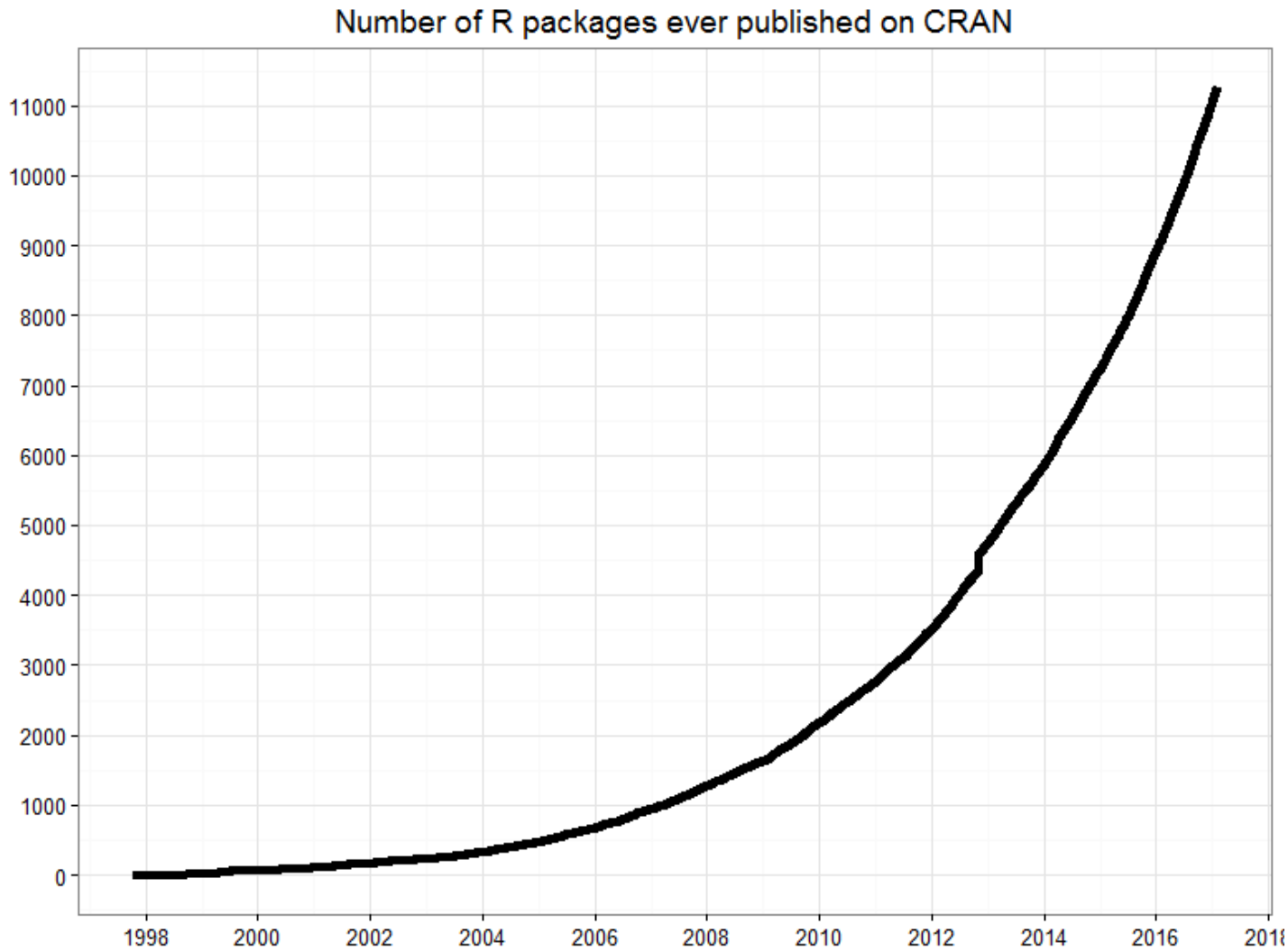
1960s – c2000: Statistical analysis developed by academic statisticians, but implementation relegated to commercial companies (SAS, BMDP, Statistica, Stata, Minitab, etc).

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

1990s: Ross Ihaka & Robert Gentleman (Univ Auckland NZ) mimic S in an open source system, R. R Core Development Team expands, GNU GPL release.

Early-2000s: Comprehensive R Analysis Network (CRAN) for user-provided specialized packages grows exponentially. Important packages incorporated into base-R.

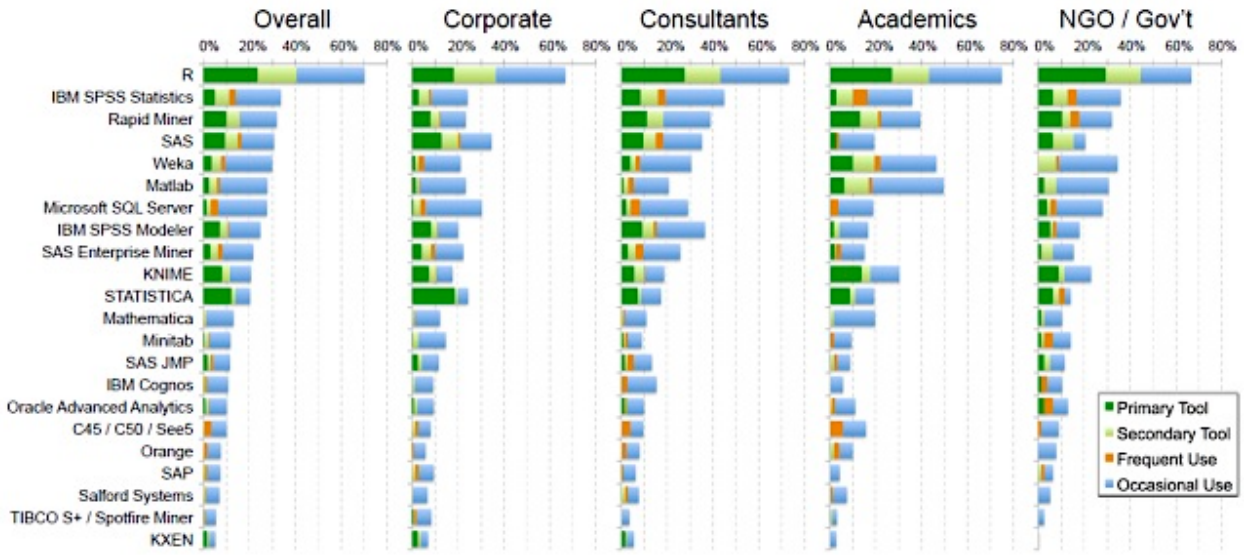
Growth of CRAN contributed packages



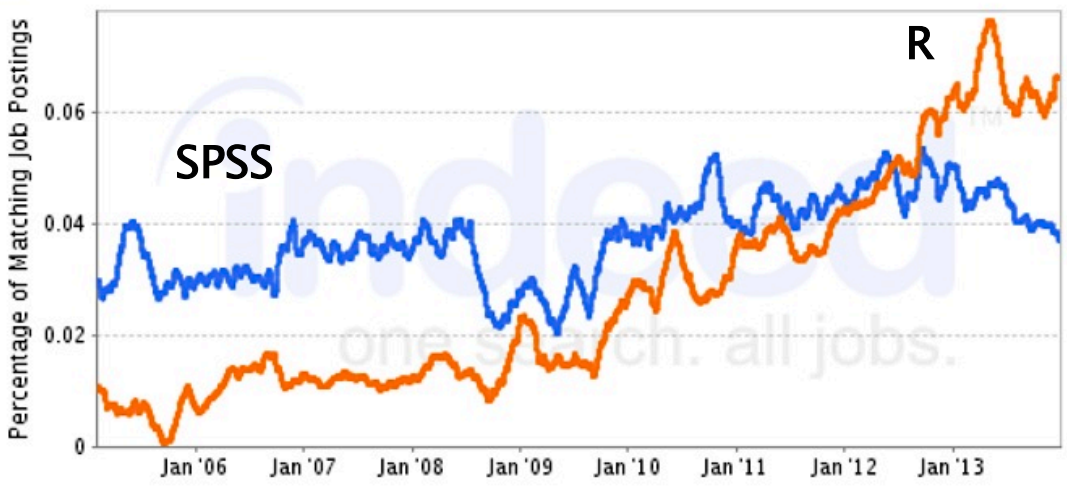
May 20 2022:
18,662 packages
(+5/day)

~150,000
functions ?

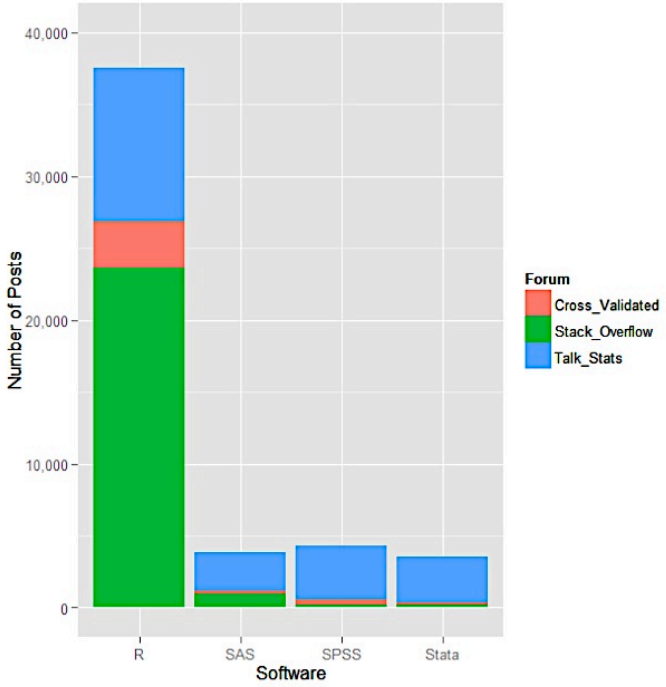
R's growing importance in data science



Rexer Analytics Data Miner Survey 2013



Job trends from Indeed.com



Posts on software forums 2013

c2017, Python overtook R as the preferred software environment by data scientists

The R statistical computing environment

- R integrates data manipulation, graphics and extensive statistical analysis. Uniform documentation and coding standards. But quality control is limited for community-provided CRAN packages.

- Fully programmable C-like language, similar to IDL. Specializes in vector/matrix inputs.

- Easy download from http://www.r-project.org for Windows, Mac or linux. On-the-fly installation of CRAN packages. Quick communication with C, Fortran, Python. Emulator of Matlab.

- ~19,000 user-provided add-on **CRAN** packages, ~150,000 statistical functions

- Many resources: R help files (3500p for base **R**), CRAN Task Views and vignette files, on-line tutorials, >170 books, >1000 blogs, *Use R!* conferences, galleries, companies, *The R Journal* & *J. Stat. Software,* etc.

Principal steps for using R in astronomical research:

- ***Knowing what you want*** **[education, consulting, thought]**
- ***Finding what you want*** **[Google, Rseek, Rdocumentation]**
- ***Writing R scripts*** **[R Help files, books]**
- ***Understanding what you find*** **[education, consulting, thought]**

Some functionalities of 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
max 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, tessellations, three-dimensional visualization, wavelet toolbox

CRAN Task Views

(http://cran.r-project.org/web/views)

CRAN Task Views provide brief overviews of CRAN packages by topic & functionality. Maintained be expert volunteers. Partial list:

- Bayesian ~110 packages
- Chem/Phys ~75packages (incl. 20 for astronomy)
- Cluster/Mixture ~100 packages
- Graphics ~40 packages
- HighPerfComp ~75 packages
- Machine Learning ~70 packages
- Medical imaging ~20 packages
- Robust ~50packages
- Spatial ~135packages
- Survival ~200 packages
- TimeSeries ~170 packages

Since c.2010, R has been the world's premier statistical computing package