

# **Statistics and the Astronomical Enterprise**

**Eric Feigelson  
Center for Astrostatistics  
Penn State University**

**2016 Sagan Exoplanet Summer Workshop**

# ***Everything* is statistical with exoplanets!**

## Extremely low signal-to-noise

- <1 m/s Doppler effect in radial velocity
- >15 mag sensitivity in direct imaging
- transmission spectroscopy of atmospheres

## Rare events in enormous samples

- $10^5$  stars monitored for transit events
- $10^8$  stars monitored for microlensing events

## Tiny fraction of true population is sampled

- true  $\eta_{\oplus} \sim 10^5$  x observed fraction

*Therefore, more so than most scientists, exoplanetary astronomers need to be savvy about statistical methodology.*

*This talk gives a broad conceptual & historical overview of the role of statistics in astronomy, describes the current situation, and provides advice.*

*At the end, we introduce R, the statisticians' principal software system that essentially implements all of modern statistics in an integrated environment.*

# What is astronomy?

**Astronomy** is the observational study of matter beyond Earth: planets in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations.

**Astrophysics** is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an indirect, inferential intellectual effort based on the assumption that physics – gravity, electromagnetism, quantum mechanics, etc – apply universally to distant cosmic phenomena.

# What is statistics? *(No consensus !!)*

- “... briefly, and in its most concrete form, the object of statistical methods is the reduction of data”

(R. A. Fisher, 1922)

- “Statistics is the mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.”

(Wikipedia, 2014)

- “Statistics is the study of the collection, analysis, interpretation, presentation and organization of data.”

(Wikipedia, 2015)

- “A statistical inference carries us from observations to conclusions about the populations sampled”

(D. R. Cox, 1958)

# *Does statistics relate to scientific models?*

## *The pessimists ...*

“Essentially, all models are wrong, but some are useful.”

(Box & Draper 1987)

“There is no need for these hypotheses to be true, or even to be at all like the truth; rather ... they should yield calculations which agree with observations” (Osiander’s Preface to Copernicus’ *De Revolutionibus*, quoted by C. R. Rao in *Statistics and Truth*)

"The object [of *statistical* inference] is to provide ideas and methods for the critical analysis and, as far as feasible, the interpretation of empirical data ... The extremely challenging issues of *scientific* inference may be regarded as those of synthesising very different kinds of conclusions if possible into a coherent whole or theory ... The use, if any, in the process of simple *quantitative* notions of probability and their numerical assessment is unclear."

(D. R. Cox, 2006)

## ***The positivists ...***

“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, *Bayesian Logical Data Analysis for the Physical Sciences*, 2005)



# Recommended steps in the statistical analysis of scientific data

The application of statistics can reliably quantify information embedded in scientific data and help adjudicate the relevance of theoretical models. But this is not a straightforward, mechanical enterprise. It requires:

- exploration of the data
- careful statement of the scientific problem
- model formulation in mathematical form
- choice of statistical method(s)
- calculation of statistical quantities ← *easiest step with R*
- judicious scientific evaluation of the results

***Astronomers often do not adequately pursue each step***



- Modern statistics is vast in its scope and methodology. It is difficult to find what may be useful (jargon problem!), and there are usually several ways to proceed. Very confusing.
- Some statistical procedures are based on mathematical proofs which determine the applicability of established results. It is perilous to violate mathematical truths! Some issues are debated among statisticians, or have no known solution.
- Scientific inferences should not depend on arbitrary choices in methodology & variable scale. Prefer nonparametric & scale-invariant methods. Try multiple methods.
- It can be difficult to interpret the meaning of a statistical result with respect to the scientific goal. Statistics is only a tool towards understanding nature from incomplete information.

***We should be knowledgeable in our use of statistics  
and judicious in its interpretation***

# Astronomy & Statistics: A glorious past

*For most of western history,  
the astronomers were the statisticians!*

## Ancient Greeks to 18<sup>th</sup> century

Best estimate of the length of a year from discrepant data?

- Middle of range: Hipparcos (4<sup>th</sup> century B.C.)
- Observe only once! (medieval)
- Mean: Brahe (16<sup>th</sup> c), Galileo (17<sup>th</sup> c), Simpson (18<sup>th</sup> c)
- Median w/ bootstrap (21<sup>th</sup> c)

## 19<sup>th</sup> century

Discrepant observations of planets/moons/comets used to estimate orbital parameters using Newtonian celestial mechanics

- Legendre, Laplace & Gauss develop least-squares regression and normal error theory (~1800-1820)
- Prominent astronomers contribute to least-squares theory (~1850-1900)

## ***The lost century of astrostatistics....***

In the late-19th and 20th centuries, statistics moved towards human sciences (demography, economics, psychology, medicine, politics) and industrial applications (agriculture, mining, manufacturing).

During this time, astronomy recognized the power of modern physics: electromagnetism, thermodynamics, quantum mechanics, relativity. Astronomy & physics were wedded into astrophysics.

Thus, astronomers and statisticians substantially broke contact; e.g. the curriculum of astronomers heavily involved physics but little statistics. Statisticians today know little modern astronomy.

# The state of astrostatistics today

*(not good!)*

Many astronomical studies are confined to a narrow suite of familiar statistical methods:

- Fourier transform for temporal analysis (Fourier 1807)
- Least squares regression (Legendre 1805, Pearson 1901)
- Kolmogorov-Smirnov goodness-of-fit test (Kolmogorov, 1933)
- Principal components analysis for tables (Hotelling 1936)

Even traditional methods are sometimes misused!

*see Beware the Kolmogorov-Smirnov test! page on ASAIP*

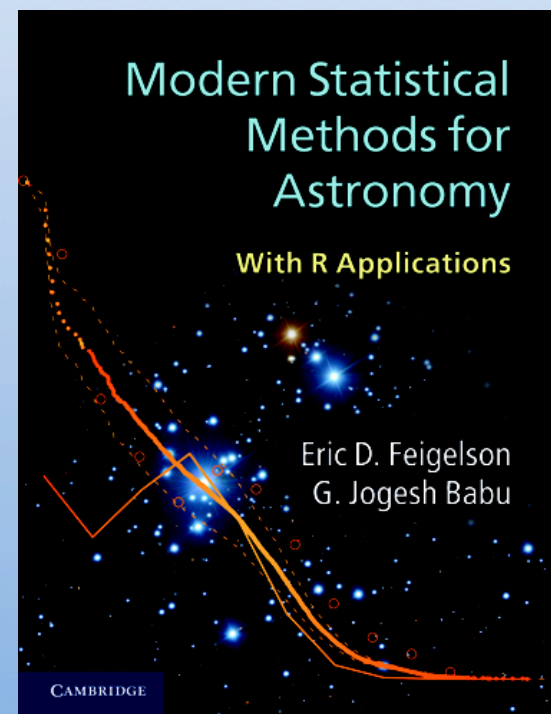
## ***Under-utilized methodology:***

- modeling (MLE, EM Algorithm, BIC, bootstrap)
- multivariate classification (LDA, SVM, CART, RFs)
- time series (autoregressive models, state space models)
- spatial point processes (Ripley's K, kriging)
- nondetections (survival analysis)
- image analysis (computer vision methods, False Detection Rate)
- statistical computing (R)

*Advertisement ...*

### **Modern Statistical Methods for Astronomy with R Applications**

E. D. Feigelson & G. J. Babu,  
Cambridge Univ Press, 2012



*Winner 2012 PROSE Award for  
Best Astronomy & Cosmology Book*

# *An astrostatistics lexicon ...*

## **Cosmology**



## **Statistics**

Galaxy clustering		Spatial point processes, clustering
Galaxy morphology		Regression, mixture models
Galaxy luminosity fn		Gamma distribution
Power law relationships		Pareto distribution
Weak lensing morphology		Geostatistics, density estimation
Strong lensing morphology		Shape statistics
Strong lensing timing		Time series with lag
Faint source detection		False Discovery Rate
Multiepoch survey lightcurves		Multivariate classification
CMB spatial analysis		Markov fields, ICA, etc
$\Lambda$ CDM parameters		Bayesian inference & model selection
Comparing data & simulation		<i>under development</i>



# Recent resurgence in astrostatistics

- Improved access to statistical software. R/CRAN public-domain statistical software environment with thousands of functions. Increasing capability in Python.
- Papers in astronomical literature doubled to ~500/yr in past decade (“Methods: statistical” papers in *NASA-Smithsonian Astrophysics Data System*)
- Short training courses (Penn State, India, Brazil, Greece, China, Italy, France, Germany, Spain, Sweden, IAU/AAS/CASCA/... meetings)
- Cross-disciplinary research collaborations (Harvard/ICHASC, Carnegie-Mellon, Penn State, NASA-Ames/Stanford, CEA-Saclay/Stanford, Cornell, UC-Berkeley, Michigan, Imperial College London, Swinburne, ...)
- Cross-disciplinary conferences (*Statistical Challenges in Modern Astronomy, Astronomical Data Analysis 1991-2016, PhysStat, SAMSI 2006/2012, Astroinformatics 2012-16*)
- Scholarly society working groups and a new integrated Web portal <http://asaip.psu.edu> serving: Int’l Stat Institute’s Int’l Astrostatistical Assn, Int’l Astro Union Working Group (Commission), Amer Astro Soc Working Group, Amer Stat Assn Interest Group, LSST Science Collaboration, IEEE Astro Data Miner Task Force)



# A new imperative: Large-scale surveys & megadatasets

Huge imaging, spectroscopic & multivariate datasets are emerging from specialized survey projects & telescopes:

- $10^9$ - $10^{10}$ -object photometric catalogs x  $10^0$ - $10^3$  epochs from 2MASS, SDSS, VISTA, CRTS, Pan-STARRS, DES, LSST ...
- $10^6$ - $10^8$ - galaxy redshift catalogs from SDSS, LAMOST, ...
- Spectral-image datacubes (VLA, ALMA, IFUs)
- Radio interferometer data streams (e.g. 30 Tflops processor for LOFAR)

*The Virtual Observatory is an international effort to federate many distributed on-line astronomical databases.*

**Powerful statistical tools are needed to derive scientific insights from TBy-PBy-EBy databases**

*To treat massive data streams and databases ...*

# **Rapid rise of astroinformatics**

*Statistics guides the scientist on what to compute*

*Informatics helps the scientist perform the computation*

Methodology: Computationally intensive astronomy, data mining, multivariate regression & classification, machine learning, Monte Carlo methods, NlogN algorithms, etc.

Software & hardware: Parallel processing on multi-processors machines, cloud computing, CUDA & GPU computing, database management & promulgation, etc.

Workshops & training schools emerging. IAU Symposium #325 Astroinformatics in Sorrento IT, October 2016. Growing perception that more community training is needed.

# New resources in astrostatistics

## Textbooks

*Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support*

Gregory, 2005

*Practical Statistics for Astronomers*

Wall & Jenkins, 2<sup>nd</sup> ed, 2012

*Modern Statistical Methods for Astronomy with R Application,*

Feigelson & Babu, 2012

*Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data,*

Ivecic, Connolly, VanderPlas & Gray, 2014

## Societies (join one!)

Intl Astrostatistics Assn affiliated with ISI (2010)

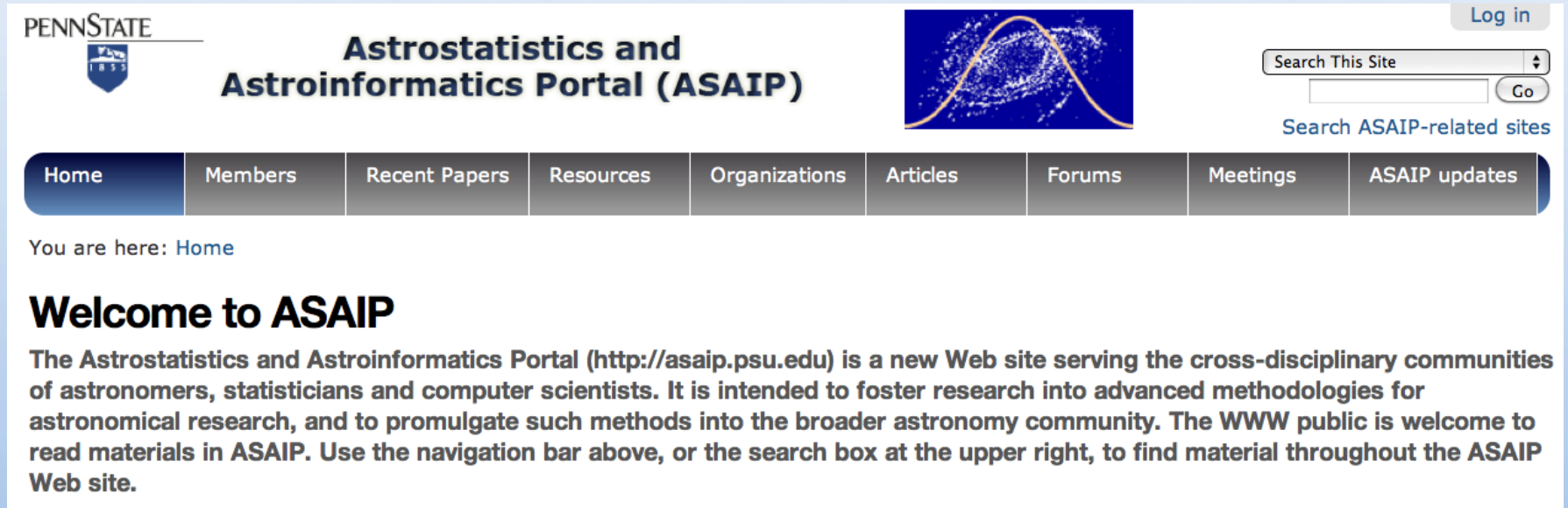
AAS Working Group in Astrominformatics & Astrostatistics (2013)

ASA Interest Group in Astrostatistics (2014)

IAU Commissions B1–B2–B3 & WG/TDA (2015)

# *Astrostatistics and Astroinformatics Portal*

*http://asaip.psu.edu*



The screenshot shows the homepage of the Astrostatistics and Astroinformatics Portal (ASAIP). At the top left is the Penn State logo. The main title is "Astrostatistics and Astroinformatics Portal (ASAIP)". To the right is a blue graphic of a galaxy with a yellow curve overlaid. Further right is a search box labeled "Search This Site" with a "Go" button and a "Log in" link. Below the search box is a link to "Search ASAIP-related sites". A navigation bar contains links for Home, Members, Recent Papers, Resources, Organizations, Articles, Forums, Meetings, and ASAIP updates. Below the navigation bar, it says "You are here: Home". The main heading is "Welcome to ASAIP". The introductory text states: "The Astrostatistics and Astroinformatics Portal (http://asaip.psu.edu) is a new Web site serving the cross-disciplinary communities of astronomers, statisticians and computer scientists. It is intended to foster research into advanced methodologies for astronomical research, and to promulgate such methods into the broader astronomy community. The WWW public is welcome to read materials in ASAIP. Use the navigation bar above, or the search box at the upper right, to find material throughout the ASAIP Web site."

Recent papers, meetings, jobs, blogs, courses, forums, ...

## ***A vision of astrostatistics by 2025 ...***

- Astronomy graduate curriculum has 1 year of statistical and computational methodology
- Some astronomers have M.S. in statistics and computer science
- Astrostatistics and astroinformatics is a well-funded, cross-disciplinary research field involving a few percent of astronomers (cf. astrochemists) pushing the frontiers of methodology.
- Astronomers regularly use many methods coded in R.
- *Statistical Challenges in Modern Astronomy* meetings are held annually with ~400 participants

# **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).

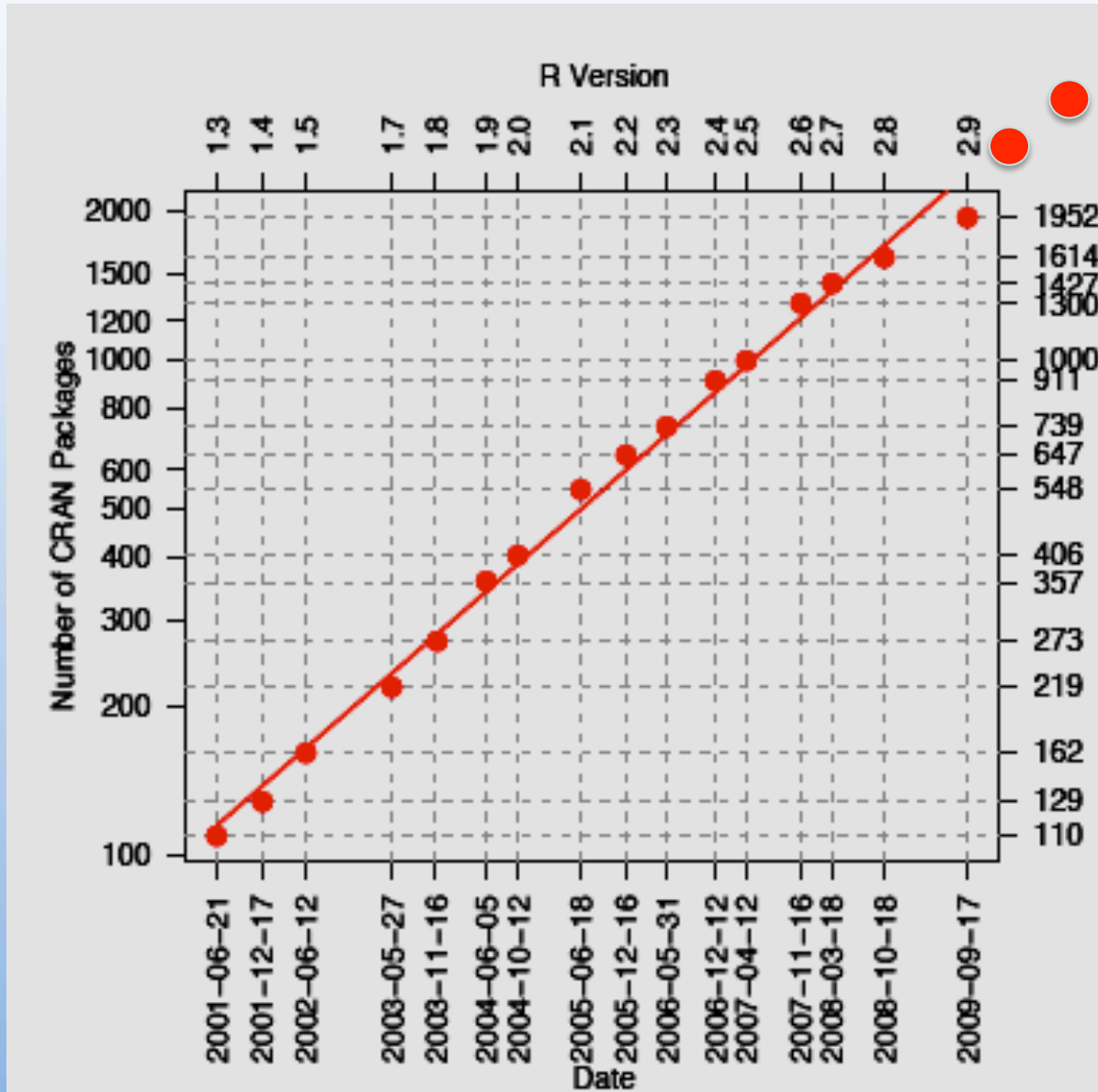
1980s: John Chambers (ATT, USA) develops S system, C-like command line interface.

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

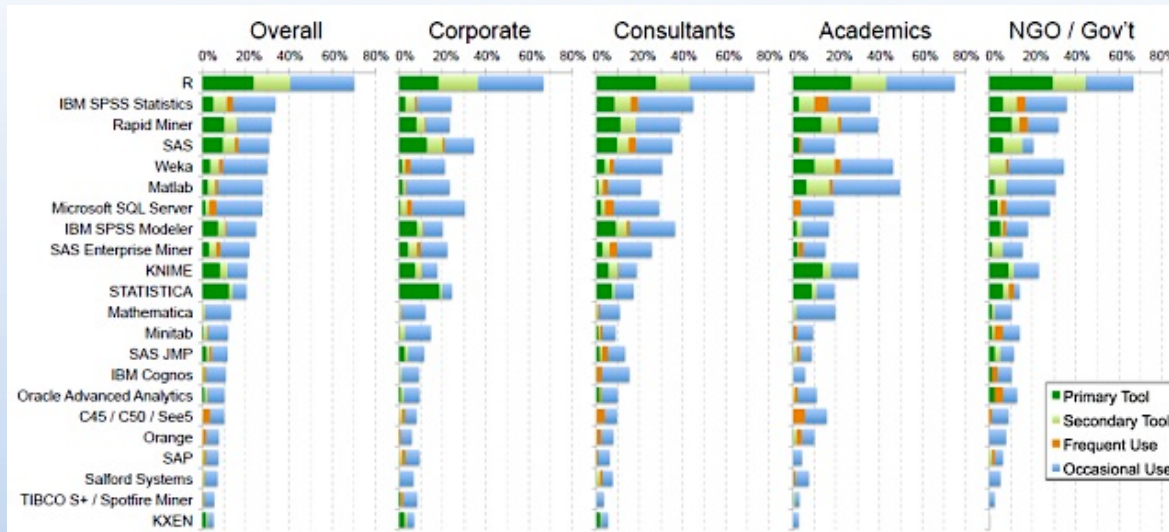


July 16 2016:  
8772 packages  
(~4/day)

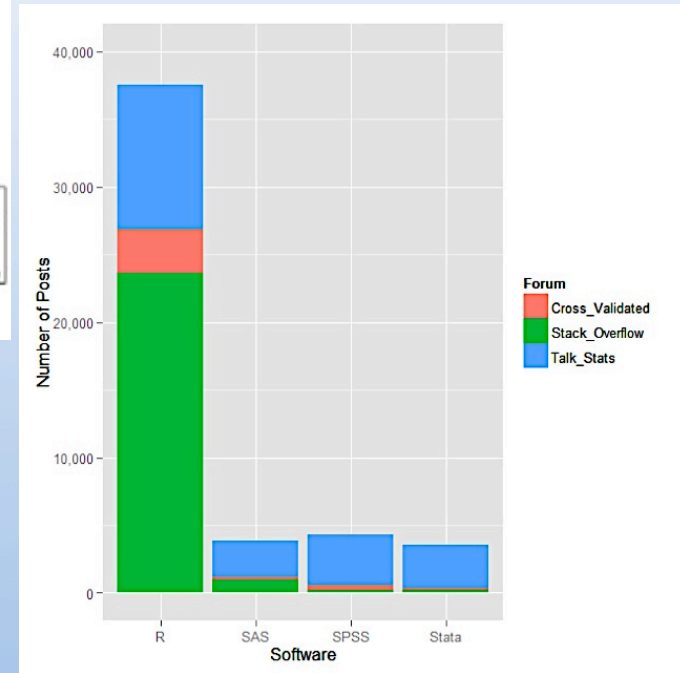
~150,000  
functions



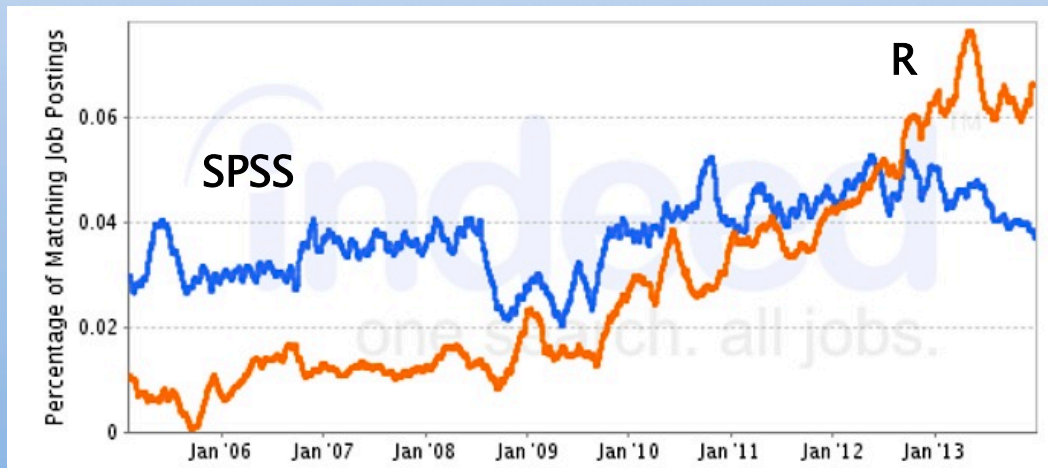
# R's growing importance in data science



Rexer Analytics Data Miner Survey 2013



Posts on software forums 2013



Job trends from Indeed.com

*See R vs. Python debates on ASAIP Software Forum*

# 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.
- >8700 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, >150 books, >400 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 analysis & 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 by expert volunteers. Partial list:

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



***Since c.2005, R has been the  
world's premier  
public-domain  
statistical computing package***

**Data scientists recommend both Python and R  
Usage of both is growing rapidly  
(<https://asaip.psu.edu/forums/software-forum/195790576>)**