Astrostatistics and Astroinformatics

NASA Big Data Task Force

29 June 2016

Eric Feigelson

Penn State University

Why astrostatistics?

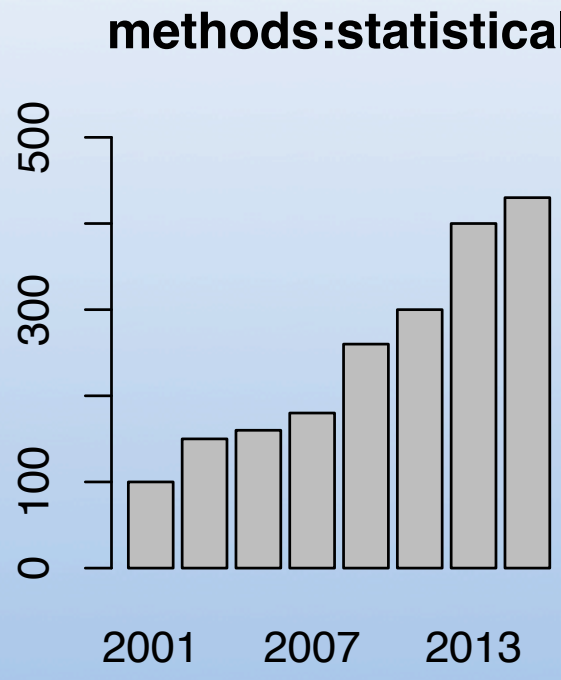
Space scientists are expert at satellite instrumentation and scientific goals … but not necessarily expert on methods for linking data and interpretation. Statistical modeling can be tricky and traditional procedures may not be effective. Improved methodology is often needed.

> *"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)*

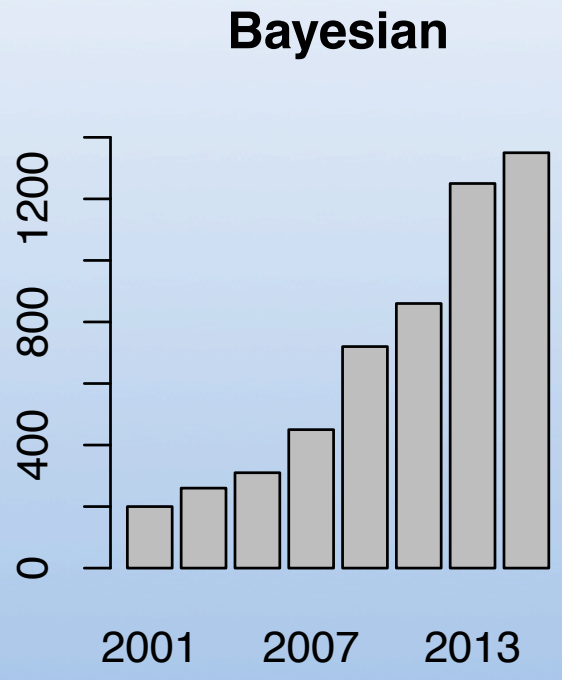
An astrostatistics lexicon …

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
ΛCDM parameters	⟷	Bayesian inference & model selection
Comparing data & simulation	⟷	*under development*

Rapid growth of statistical & Bayesian methodology ….



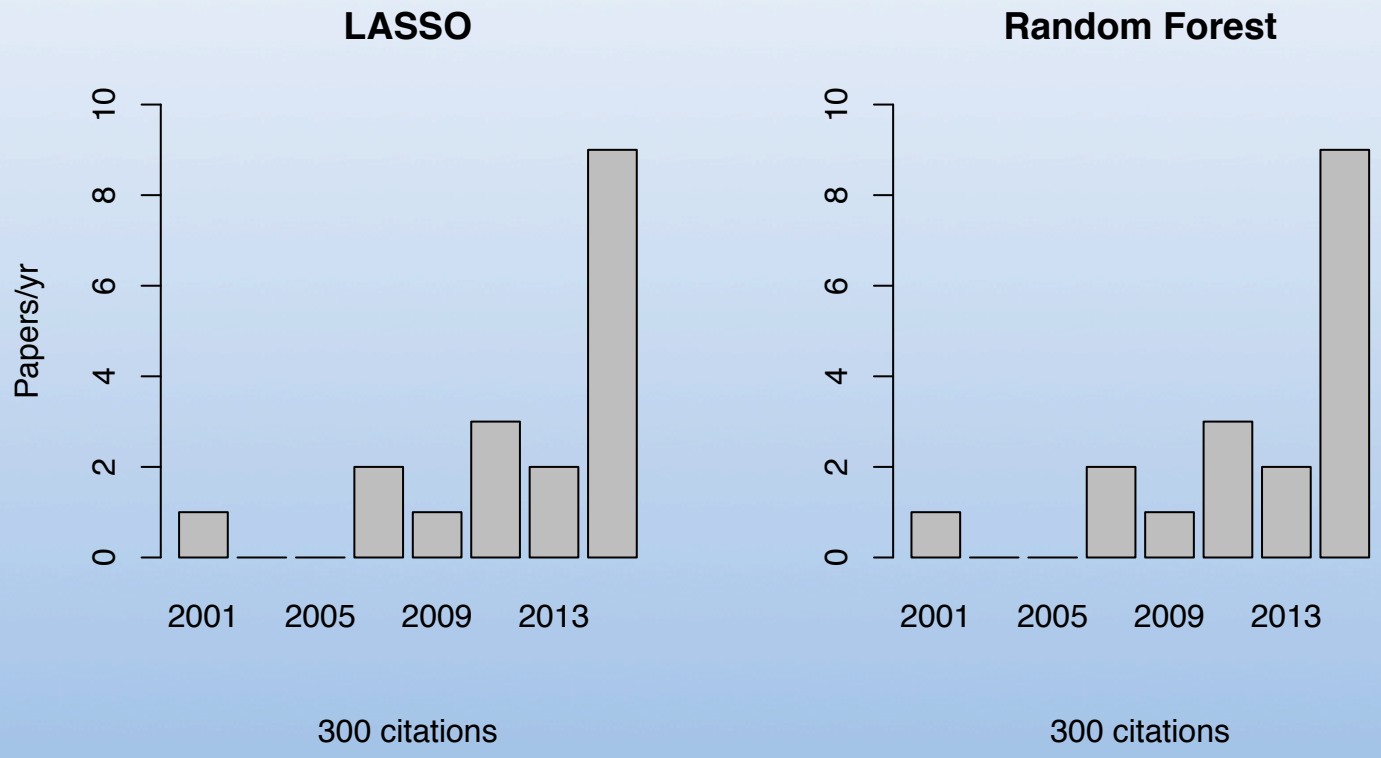
96,000 citations



370,000 citations

Full-text ADS bibliometry, refereed papers/yr in astronomy journals

… but still weak use of other modern methods



LASSO regression and CART/RF each have ~400K hits in Google

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

Some past difficulties

For its astronomy missions, NASA mission science centers and archive research centers provide software for reduction and modeling of data for scientific interpretation. Some limitations have included:

- Astrophysical models for reduced data are calculated with antiquated regression techniques and weak model selection tools
- Faint source detection is not based on modern treatments for multiple hypothesis testing
- Time series analysis software does not include methods from signal processing and econometrics
- Image processing does not include methods from computer vision
- Machine learning techniques for classification are rarely used
- NASA astronomy software teams often have few-to-no employees with training in statistics, applied mathematics or computer science

Why astroinformatics?

*Statistics guides the scientist on **what** to compute*

*Informatics guides the scientist on **how** to perform the computation**

Prior to c2010, NASA astronomy missions were limited by telemetry to relatively small datasets. But WFIRST (mid-2020s) will be in geosynchronous orbit to get continuous 1 Gbps telemetry … multi-petascale dataset.

Petascale datasets in astronomy mainly arise from ground-based wide-field surveys (e.g. LSST) and radio interferometers (e.g. SKA).

*** Informatics is rarely needed for small (<TBy) datasets**

Big Theory

NASA provides High Performance Computing for large-scale astrophysical calculations:

- cosmological ΛCDM (Big Bang) models; N-body + hydrodynamics
- black hole accretion simulations; magnetohydrodynamics with General Relativity

NASA has many other HPC needs relating to scientific simulations: aerodynamical & aerospace engineering, climate simulations, heliophysics modeling, computational fluid dynamics, quantum computing, oceanography, materials science, space weather, …

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)

To treat massive data streams and databases …

Rapid rise of astroinformatics

Methodology: Computationally intensive astronomy, data mining, multivariate regression & classification, machine learning, Monte Carlo methods, *O(NlogN)* vs. $O(N^2)$ 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 (Oct 2016), workshops, tutorials & Hack Days.

Growing perception that more community training is needed. NASA software products for data reduction and science analysis can improve with input from statisticians, applied mathematicians and computer scientists.

The tension today

National & international scholarly societies, National Academy of Sciences and NASA advisory reports, journal editorial boards, and community White Papers all call for improved education and quality in statistics and informatics for astronomical research.

But NASA closed its only SMD-wide program devoted to these issues in 2011: Applied Information Systems Research Program. Funds for methodology improvements fare poorly in highly competitive research programs designed for astronomy & astrophysics. Software development is not scoped to include methodology improvements.

Increased attention, funding and employment for
improved statistics & informatics in
NASA's astronomical space science enterprise

Prospects for improved methodology in astronomy

- Increased interest about Big Data throughout astronomy (university CyberScience institutes, organization Working Groups, new hires for `astronomical surveys', Time Domain Astronomy, philanthropy)
- Improved access to modern methods via Python & R languages
- Astrostatistics is maturing, though underfunded (conferences, collaborations, societies, ASAIP Web portal, textbooks)
- Although university curricula are slow to respond, informal training schools in astroinformatics proliferate (coding practices, parallel processing, machine learning)
- Promulgation & scientific reproducibility (github & ASCL software repositories; software & methods in AAS & A&C journals)

Closing thoughts

Change in mindset needed among (senior) astronomers: statistical & computational methodology is not a mechanical, peripheral enterprise to astronomical research, not just `software' for `data reduction'.

Methodology is an essential element of the process from the telescope to astronomical discovery and astrophysical understanding, with and without Big Data. . Quality methods can be as important as quality hardware.

Astronomers need improved education, resources and cross-disciplinary collaboration to improve the scientific enterprise.