AAS 235
HAWAI'I, 2020 JANUARY 5

Overview of WGAA Science White Paper

Matthew J. Graham
Center for Data-Driven Discovery/ZTF, Caltech
mjg@caltech.edu

CENTER FOR DATA-DRIVEN DISCOVERY

ZTF

Astro2020 Science White Paper (arXiv:1903.06796)

The Next Decade of Astroinformatics and Astrostatistics

Aneta Siemiginowska + 51 co-authors

Relevant to:

- Planetary Systems
- Star and Planet Formation
- Formation and Evolution of Compact Objects
- Cosmology and Fundamental Physics
- Stars and Stellar Evolution
- Resolved Stellar Populations and their Environments
- Galaxy Evolution
- Multi-Messenger Astronomy and Astrophysics

1. What is the role of astrostatistics and astroinformatics research?

"To develop modern methods for extracting scientific information from astronomical data"

Astrostatistics: foundation for robust algorithms and principled methods that are applied to a variety of problems in astronomy

Astroinformatics: the systematic and disciplined development of code, data management and dissemination techniques, high-performance code, and machine learning-based inference

Astro data science lives at the intersection of astronomy (observation and theory), statistics, algorithm development, and applied computer science

In response to Astro2010 Decadal Survey:

- Several white papers on both astrostatistics and astroinformatics were submitted

- Some recommendations have been implemented:
  - Formation of the Working Group on Astroinformatics & Astrostatistics within the AAS (June 2012)
  - Formation of the Astrostatistics Interest Group within the American Statistical Association

- Underdeveloped areas:
  - Formal recognition of and financial commitment to the efforts needed to make necessary progress

Astro data science is a growth area

Publications :

- Number with keyword `Methods: Statistical' has increased by a factor of 2.5 in the past decade
- Number with `machine learning' has increased by a factor of 4 in the past five years
- Number with `deep learning' has tripled every year since 2015

Teaching:

- Rise in summer schools, e,g., La Serena Data Science School
- Special sessions at AAS meetings

Recognized as a necessity for nontrivial scientific inference with astronomy data in the next decade: volume, velocity, variety, and veracity

2. How do modern astrostatistics and astroinformatics methods impact astronomy?

"They overcome challenges with data and improve scientific inference"

Astrostatistics and astroinformatics research does not fit traditional thematic boundaries

Recent advances and expected challenges

Impact of emerging methods in thematic areas of astronomy

Distance measurements

Examples: stars, dust, quasars

Traditionally: invert parallaxes, sample truncation, astrometry-based luminosity, template fitting, stellar variability, photo-z

Limitations & challenges: biases, need bias corrections, uncertainties ignored

Emergent methodologies: Bayesian inference for distances from parallaxes and for proper motions from astrometric data, machine learning methods, photometric redshifts

Mass estimates

Examples: the Milky Way, dwarf galaxies, supermassive black holes, galaxy groups and clusters

Traditionally: kinematic tracers, timing argument, hyper velocity stars, reverberation, mass-σ relation, power-law scaling

Limitations & challenges: data incompleteness, large incertainties, extrapolation to larger distances, over-simplified models, biases

Emergent methodologies: Bayesian hierarchical models, Approximate Bayesian Computation (ABC), CARMA models, machine learning, Bayesian model averaging

Stellar properties and evolution

Examples: stellar type, temperature, composition, metallicity, coronal composition, density, stellar evolution

Traditionally: forward fitting physics-based models (with chemical networks, MHD and planets for protoplanetary disks), stellar evolution models, spectral line fitting, isochrone fitting, catalog matching and membership classification

Limitations & challenges: degenerate models and parameters, model selection difficulty, uncertainty quantification, correlated measurement uncertainties, inefficient sampling methods (e.g., MCMC) and simplifications to forward fitting models

Emergent methodologies: Gaussian processes, machine learning methods, Bayesian inference, model independent data-driven approaches for rotation curves, matched filter for line searches

Population studies

Examples: source detection, structures of diffuse regions, classifying galaxies and stars, identifying moving groups, stellar populations, globular cluster populations

Traditionally: spectral line fitting, photometry, color-magnitude diagrams, two-point correlation function

Limitations & challenges: overlapping sources, faint structures, non-Gaussian uncertainties, unknown populations, complex morphology

Emergent methodologies: probabilistic catalogs, machine learning for open clusters, identifying members of stellar groups, spatial point patterns, accounting for uncertainties, wavelet-based clustering methods

Stars and stellar evolution: magnetic activity, population evolution, environment

Advances: stellar cluster catalog matching and membership classification; structure in diffuse X-ray background; solar feature classification and properties; flare modeling, energy release and evolution; thermal segmentation of the corona; stellar coronal thermal density structure via Emission Measure distributions; sources of coronal heating (e.g., nanoflares); effect of stellar activity on exoplanets

Challenges: solar and stellar flare onsets and distributions; characterizations of stellar activity to reveal hidden signals of exoplanets; determining the nature and magnitude of the Star-Planet Interaction effect; isochrone fitting ti determine ages, metalliicity, and star formation history of star clusters; completeness and limitations of Heliophysics Event Knowledgebase

Emergent methodologies: Solar dispersed image spectral decomposition; incorporating atomic data uncertainties; disambiguating photons from overlapping close binaries in confused fields for spectral and timing analysis; loop recognition in solar coronal images; morphological analysis to recognize diffuse structure

Formation and evolution of compact objects

Advances: X-ray spectral timing analysis; Bayesian inference and evolutionary algorithms for neutron star equation of state; merging systems; transient detections; accretion states

Challenges: use of spectral, spatial, time, and polarimetry domain; period detection; “needle in a haystack” searches; state transitions; localization

Emergent methodologies: New models, computational power; Gaussian processes in time domain for Poisson X-ray and gamma ray; use of higher order Fourier product and nonlinear signal processing; machine learning methods

Galactic astronomy and galaxy evolution

Advances: Gaia data; machine learning methods and Bayesian inference

Challenges: account for uncertainties, incompleteness, and biases

Emergent methodologies: Machine learning to discover new stellar open clusters; photometric redshifts; Bayesian inference for distances from parallaxes and for proper motions from astrometric data; the mass of the Milky Way; identifying members of stellar groups

Multi-messenger astronomy and astrophysics

Advances: Detecting transients, multi-band identifications

Challenges: Localization, nanohertz GW detection, GW-EM coincidence

Emergent methodologies: Bayesian hierarchical models with efficient samplers, Gaussian mixture models

Planetary systems

Advances: >100,000 target stars in Kepler, characterization of planetary systems, analysis of transit timing variations to characterize the exoplanet mass-radius relationship

Challenges: Inadequate statistical estimators used in early analysis; Bayesian hierarchical models to combine large measurement uncertainties and intrinsic astrophysical variability; TESS + ground-based Doppler surveys

Emergent methodologies: characterizing stellar variability, machine learning and data-driven models for analyzing high resolution spectroscopic time series, quantifying the evidence of low-mass planets, Gaussian processes for RV time series

Stars and planet formation

Advances: Extract information from high resolution optical and infrared spectra; ALMA images; use molecular lines to probe high-dimensional space with dynamic and chemical information; multiwavelength studies of spatial and kinematic structures in clusters

Challenges: Gaussian processes to deal with correlated residuals from model systematics, and to construct physics-based forward models; data-driven approaches for accurate spectral models on a pixel-by-pixel bases; oversimplified models used with Bayesian inference; inadequate statistics for understanding highly non-homogeneous distributions of young stellar objects

Emergent methodologies: Bayesian inference for sophisticated models in computationally tractable implementation; data driven approaches to model disk rotation curves and search for planetary influence; Fourier-based matched filters for molecular line searches; new statistics for inhomogeneous point processes; advanced cluster analysis methods including mixture models and density clustering

Cosmology and fundamental physics

Advances: Harness subtle signals to reduce scatter; quickly generate mock data; discriminate models to quantify statistical and systematic uncertainties; likelihood-free cosmological inference and classify objects

Challenges: Apply new advances in ML interpretability, including saliency maps and the deep k-nearest neighbors approach

Emergent methodologies: A brand new Bayesian machine learning framework suitable for many layers of statistical error structures: selection bias, errors-in-measurements, systematics, etc.

3. How can the state-of-the-art methods be best applied in astronomy?

"Through astronomy involvement in active methodology research"

Explanatory models rather than task-specific predictive models:

- *Scalable probabilistic machine learning (including DL):* most ML algorithms aim to make one set of predictions but we need methods that quantify uncertainty and provide results that enable uncertainty propagation – we should be driving this research

- *Interpretable machine learning (especially DL):* complex black box methods are coming to astronomy but generalizable understanding of a phenomenon requires an interpretable model – we should be actively participating in this research

4. Recommendations

Funding:

- Astrostatistics and astroinformatics must be recognized as a subfield of astronomical research that affects all its thematic areas
- Proposals in this field must be evaluated by appropriately cross-disciplinary panels

Communication:

- Astronomy conferences must make room for methodological discussion, both for new advances and non-experts
- Funding for tutorials and other means of communication should be encouraged

Sustainability:

- Grants and fellowships to support graduate students and postdocs for several years
- Astronomy departments should be encouraged to have more tenure-track positions focused on data science research

Infrastructure:

- Support for maintaining data archives and training data sets, publicly available and support software, and efficient computing