# Astroinformatics

## Machine Learning and Applications in Space Explorations

Suryoday Basak, Rahul Aedula, Rohith Raj, Vaishnavi Sagar, Poulami Sarkar, Snehanshu Saha and Gajanan V Honnavar

Center for Mathematical Modeling and Simulation, Department of Computer Science and Engineering, PESIT Bangalore South Campus

# Data Scientific Applications in Astronomy: Advent and Opportunities

- Modern astronomical instruments record huge volumes of data in the form of images, catlogues, raw data and signals in different bandwidths.
- A new wave of pursuits in astronomy thus involve the use of **statistics**, **machine learning**, and **artificial intelligence** to solve problems. An emerging **interdisciplinary** area of study which calls for scientists to collaborate from the fields of:
  1. Astronomy and astrophysics
  2. Statistics
  3. Mathematics
  4. Computer and Information sciences.

  **Opportunities:**
  1. to attract students from astrophysics, statistics, and computer science backgrounds; to create awareness in the fast developing field
  2. Scientometric analysis can provide insights to the pattern of collaboration between scientists from different domains, the popularity of various sub domains, and the direction in which future pursuits can be made.
  3. to develop an evolving, inter-disciplinary field and create nationwide community on AstroInformatics.
  4. bring in International Astrostatistics Association (IAA) for cross collaboration and greater visibility.

# Talking points

1. Classifying habitability of exoplanets
2. Clustering and classification for Novae
3. Classification of Stars and Quasars
4. Galex Catalog
5. Gravitational Waves Simulation and classification
6. Fundamental methods in Machine Learning and applications in Astronomy

# Our Work: Cobb-Douglas Habitability Production Function

The CD-HPF is a novel approach to estimate the habitability of an exoplanet. It is formulated as a minimization problem:

$$Y = f(R, D, T_s, V_e) = K \cdot R^{\alpha} \cdot D^{\beta} \cdot T_s^{\gamma} \cdot V_e^{\delta}$$

Key aspects of this model:

1. Inspired by econometric models
2. Important factors of habitability can contribute differently for different planets
3. convergence can be proved
4. can be solved by a computer in the **log-linear** form
5. new result powered by the principle of mathematical induction ensuring global optima under additional parameters such as orbital velocity, eccentricity etc.

**Video link**

# Classification of Exoplanets

The rate at which exoplanets are being discovered is increasing. With the scheduled launch of the JWST in 2018, automated methods of classification need to be explored.

As a part of this endeavor, we plan:

1. Explore the efficacy of various classification algorithms
2. Propose the best methods of classification based on linear separability of data, etc.
3. Simulate the growth of data and tested the efficacy of ML algorithms on artificially augmented datasets
4. propose new methods to handle under-represented classes

# Eccentricity Estimation of Exoplanets

One of the aspects of the PHL-EC catalog is that it assigns an eccentricity of 0 if the eccentricity of the orbit of a planet cannot be estimated. This prevents eccentricity from directly being used as a feature in the CD-HPF as it is in a **multiplicative** form.

To tackle this, the following methods are being tried:

1. Modeling the eccentricity of planets using perturbation theory
2. Modeling the eccentricity of planets using quadratic and logarithmic regression

Having determined the eccentricities reliably, the CD-HPF can be extended as:

$$Y = k \cdot R^{\alpha} \cdot D^{\beta} \cdot T_s^{\gamma} \cdot V_e^{\delta} \cdot S_f^{\zeta} \cdot (E + \epsilon)^{\tau}$$

thus incorporating other important factors for habitability.

# Metallicity effects on habitability- Computational DoE Problem

Target variable, $y$, habitability; factorial analysis needed to determine importance of metals toward habitability; $n^K$ designs where $K$ is number of levels refers to designs with $K$ factors where each factor (metal) has $n$ levels; levels are continuous posing a huge design problem. Requires novel metaheuristic, Multi-stage Memetic Algorithm (MSMA) to solve the problem.

- **MSMA Operators:** Four operators are defined and used in two stages- used to accomplish. Each stage requires croosover and mutaion operators, $\Omega_1$ and $\Omega_2$ for stage 1 and, $\Omega_3$ and $\Omega_4$ for stage 2 respectively. We optimize over the support points in stage 1 and optimize over the distribution of the support points in stage 2. Hence the name, Multi stage memetic algorithm.
- iterations and resets(needed in case of degenerate population or premature convergence): maxit_stage1 and maxit_stage2, max-resets. **Parameters:** $x_i$ - randomly sampled from the design space [-1, +1]; specify nominal model parameters, $\beta_i$ in U[-3, +3]; design runs are varied between two fixed values, k and k+j

# Remarks: Metallicity effects on habitability- Computational DoE Problem

- The proposed design is to be optimized in multiple stages, where in one stage we optimize over the design points and in the other stage, we optimize over the distribution of the support(design) points.
- $\Omega_1$ is defined as the transposition(crossover) operator in breeding strategy while $\Omega_2$ is the mutation operator, used to choose and create mutants.
- generate random population with $K$ design runs, where design points are sampled randomly from the design space and all weights are assigned the same.
- Information Matrix is used to evaluate fitness of population.
- The problem of convergence in the proposed metaheuristic is tackled by proving the **theorem**:
  **Under suitable conditions of local refinement, the sequence of solutions(design points) converge to optimal solutions i.e. a subsequence of the set of design points.**

# Quasar-Star Classification

The **Sloan Digital Sky Survey** (SDSS) captures deep field multi-band images of stellar objects and makes the data available to the public in the form of catalogues and images.

The problem of quasar-star classification, based on SDSS data, has been addressed before, albeit incorrectly.

1. Results of classifiers have been reported without any analysis of the trends (linear separability, etc.) in the data
2. No evidence of handling imbalance: the number of star samples are much greater than the number of quasar samples
3. No statistical measures such as TPR, FPR, F-score, etc, have been reported

- We propose a previously unexplored classification method which takes into account the imbalance in data: **Asymmetric AdaBoost**.

# Nova Classification

The challenge in pursuing classification of novae is the lack of data. For this, data is being compiled from various sources and all derived parameters are being estimated based on a small set of independent observables.

1. Quadratic and logarithmic regression is performed to estimate missing feature values
2. Derived features can be estimated from independent features, once gathered or estimated.
3. too many classes and too few samples
4. We propose novel clustering approach based on MSMA
5. artificial sample augmentation by hybrid SVM-KNN labeling method

# GALEX Catalogue

**Galaxy Evolution Explorer**(GALEX) is an ultraviolet space telescope orbiting earth. Originally launched in 2003 by the Pegasus Rocket, its main purpose was to study the evolution and change of galaxies.

The challenges while dealing with the GALEX data are:

1. The present existing catalogue has multiple redundant records of the same object over multiple observations.
2. GALEX has a 5" resolution, which makes it difficult to isolate and distinguish objects under that range.
3. There might be spurious entries in the data such as cosmic rays etc which are also recorded as objects that have to be eliminated.

Our proposed approach to tackle these challenges is to implement **Equivalence classes** along with **Cluster Based Filtering** methods to remove redundancies and any spurious records that might exist within the data.

# Novelty

1. Knowledge creation in **Machine Learning** and **computational Design of Experiments** (DoE)
2. Novel applications in highly sporadic and imbalanced data arising in exoplanet studies, classification, star quasar classification, nova classification
3. "small-data treatment": The treatment for classification of exoplanets, and novae requires reasonable **under-sampling** and **artificial augmentation**. This is the opposite big-data pursuit and **requires novel interventions**!
4. New **theorems and proofs** solidifying the analytical and statistical foundation
5. Finding hidden correlations in the complex astronomical Big Data
6. Standardization of meta-data for better science

In the realm of computing and astrophysics, these approaches are rather new and forthcoming.

# Novel contributions at a glance

- Novel *convex optimization* based model for *habitability score*, more *powerful* and *general* compared to the existing model, ESI
- propose improved habitability models under constraints and greater parameters by introducing *stochastic fronter analysis* to tackle the problem of curvature violation
- *Modeling Eccentricity*, a very important feature in habitability, ignored so far by using calculus of variations; propose a **novel additive model** to handle eccentricity
- New *habitability catalogue* and *GALEX catalogue*
- Novel **computational method-MSMA** to determine relevance and importance of factors in metallicity of exoplanets
- *New theorem* to prove *convergence* of MSMA and propose a *MSMA based novel clustering approach* for the NOVA problem
- *Equivalence class partitioning* to handle redundancies in GALEX; a new technique

# Intended Outcome/Deliverables

1. Top-tier journal publications such as MNRAS, Nature Astronomy, ASCOM and JML; publication record is established already
2. A software called ExoPlanet built for a general ML audience
3. Relevant conference presentations: Astroinformatics: From Big Data to Understanding the Universe at Large EWASS 2017, Prague, Czech Republic; *Annual Astronomical Data Analysis Software and Systems (ADASS) conference; Astroinformatics 2017, Cape Town, South Africa; Conference on Big Data from Space (BiDS'17) Toulouse, France*
4. Scientometric study on Astoinformatics-growth of internationality of the topic, knowledge network analysis, community detection and collaboration patterns–study on the growth of inter-disciplinary knowledge network
5. E-book: disseminate knowledge. Link (Work in Progress): **https://drive.google.com/file/d/0B1yuFIvTaFbhMzdkcDJPbF9mUVk/view?usp=sharing**
6. contribution to the code repository: ascl, git etc.