

# Artificial intelligence and modern physics: a two-way connection



## Report of Contributions

Contribution ID: 1

Type: **not specified**

## Welcome and introduction

*Monday, 30 September 2024 08:45 (30 minutes)*

Contribution ID: 2

Type: **not specified**

## Data access and preparation

*Monday, 30 September 2024 11:20 (1h 40m)*

**Presenter:** BESCHI, Andrea

**Session Classification:** Lecture

Contribution ID: 3

Type: **not specified**

## Data access and preparation

**Presenter:** Dr BESCHI, Andrea (McKinsey)

**Session Classification:** Lecture

Contribution ID: 4

Type: **not specified**

## **Data access and preparation**

**Session Classification:** Lecture

Contribution ID: 5

Type: **not specified**

## **Data access and preparation**

**Session Classification:** Lecture

Contribution ID: 6

Type: **not specified**

## Networking event

Contribution ID: 7

Type: **not specified**

## Bayesian statistics

*Tuesday, 1 October 2024 09:20 (1h 40m)*

**Presenter:** MOORE, Christopher J.

**Session Classification:** Lecture



Contribution ID: 8

Type: **not specified**

## Data access and preparation

*Monday, 30 September 2024 09:20 (1h 40m)*

**Presenter:** BESCHI, Andrea

**Session Classification:** Lecture

Contribution ID: 9

Type: **not specified**

## Social event

Contribution ID: **10**

Type: **not specified**

## **Social dinner**

*Thursday, 3 October 2024 20:00 (2 hours)*

Terrace above the reception building

**Project proposal: general context**

**Project proposal: description of the problem**

**Machine learning methods**

**Input dataset**

**Goal and FOM**

Contribution ID: 11

Type: **not specified**

## **ML elements**

*Tuesday, 1 October 2024 11:20 (1h 40m)*

**Project proposal: general context**

**Project proposal: description of the problem**

**Machine learning methods**

**Input dataset**

**Presenter:** Dr FINOTELLO, Riccardo (CEA Paris-Saclay)

**Session Classification:** Lecture

Contribution ID: 12

Type: **not specified**

## **ML elements**

*Wednesday, 2 October 2024 09:20 (1h 40m)*

**Presenter:** FINOTELLO, Riccardo (CEA LIST)

**Session Classification:** Lecture

Contribution ID: 13

Type: **not specified**

## ML tools

*Wednesday, 2 October 2024 11:20 (1h 40m)*

**Presenter:** GIAGU, Stefano (Sapienza Università di Roma and Istituto Nazionale di Fisica Nucleare)

**Session Classification:** Lecture

Contribution ID: 14

Type: **not specified**

## ML tools

*Thursday, 3 October 2024 09:20 (1h 40m)*

**Presenter:** GIAGU, Stefano (Sapienza Università di Roma and Istituto Nazionale di Fisica Nucleare)

**Session Classification:** Lecture

Contribution ID: 15

Type: **not specified**

## Bayesian statistics

*Thursday, 3 October 2024 11:20 (1h 40m)*

**Presenter:** MOORE, Christopher J.

**Session Classification:** Lecture



Contribution ID: **16**

Type: **not specified**

## **Hands-on session 7 (ML tools)**

*Friday, 4 October 2024 11:20 (1h 40m)*

**Session Classification:** Hands-on session

Contribution ID: 17

Type: **not specified**

## **ML applications to biomedical data**

*Thursday, 3 October 2024 14:20 (1h 40m)*

**Presenter:** SANGUINETTI, Guido (SISSA)

**Session Classification:** Lecture

Contribution ID: 18

Type: **Hackathon project proposal**

# STARFINDER - Machine Learning Techniques for Evaluating Globular Cluster's Stars Membership Probabilities

*Friday, 4 October 2024 10:40 (20 minutes)*

## Abstract

Globular clusters (GCs), spheroidal conglomeration of stars tightly bound together by means of gravitational force, are among the oldest objects that live within our galaxy. A key characteristic of these objects is their high density, significantly greater than the average galactic star density (between  $\sim 10^4$  to  $\sim 10^6$  stars within a spheroid of radius up to  $\sim 100 pc$ , in stark contrast to the local average stellar density of about  $\sim 1 - 2 \frac{\text{stars}}{pc^3}$ ), so that they can be considered collisional systems. The ESA's Gaia (Global Astrometric Interferometer for Astrophysics) mission, which has mapped nearly 2 billion stars in our galaxy up to its third data release, provides the largest set of high-resolution data available, enabling the detailed study of GCs' internal dynamics.

However, the high density of these regions presents a challenge for Gaia's 1.45-meter primary mirror, often resulting in compromised data quality and insufficient resolution. Consequently, accurately associating stars with clusters becomes difficult due to poor estimates and high errors in the parameters.

Machine Learning (ML) algorithms offer a promising solution to this problem. As demonstrated in referenced paper [1], techniques inspired by ML such as Mixture Modelling, which uses Markov-Chain Monte Carlo, Extreme Deconvolution and Maximum Likelihood Estimation, can be employed to infer the general distribution properties of the cluster, distinguishing them from field star distributions. Enhancing these methodologies with neural networks such as Generative Adversarial Networks, which could be used to simulate stellar populations based on observational data, would allow for the assignment of membership probabilities to each source in the sample, significantly increasing the number of sources available, up to a factor of  $10^2$ , and thereby enhancing the statistical robustness of subsequent astrophysical analyses.

## References

[1] Vasiliev, Baumgardt (2021). *Gaia EDR3 view on Galactic globular clusters*; MNRAS 505, 5978–6002

## Project proposal: general context

Globular clusters (GCs) are among the oldest and most densely populated stellar systems in our galaxy, offering unique opportunities to study stellar dynamics and galactic evolution. The European Space Agency's Gaia mission has provided extensive high-resolution data on nearly 2 billion stars, enabling detailed investigation of these clusters. However, the high density of stars within GCs presents significant observational challenges, mainly for the quality of the data.

## Project proposal: description of the problem

The primary challenge in studying GCs using Gaia data is the high stellar density, which often results in compromised data quality and insufficient resolution. This limitation makes it difficult to accurately associate stars with their respective clusters, leading to poor parameter estimates and

high error margins. Addressing these issues is crucial for enhancing the reliability of astrophysical analyses and improving our understanding of stellar dynamics within GCs.

## Machine learning methods

To overcome these challenges, the project proposes the application of advanced Machine Learning (ML) techniques. Specifically, Generative Adversarial Networks simulations, Markov-Chain Monte Carlo (MCMC) simulations within Mixture modelling, together, with Extreme Deconvolution and Maximum Likelihood Estimation, will be employed to analyze Gaia's data. These methods will help infer the distribution properties of clusters and distinguish cluster stars from field stars.

## Input dataset

The necessary data can be easily accessed through the 'astroquery' Python library, which can set up for Gaia data. Alternatively, a Python package that I am currently developing, which will be ready or in the final stages of completion by the time of the event, can be used. This package will be specialized in acquiring and analyzing Gaia's Globular Cluster Data.

The number of sources can vary from  $10^3$  to  $10^5$  entries, based on the search parameters and the chosen cluster, and all the data about these are packed and stored in 'astropy tables'.

**Primary author:** FERRAIUOLO, Pietro (INAF - Osservatorio Astrofisico di Arcetri)

**Co-author:** Mr MENESSINI, Matteo (INAF - Osservatorio Astrofisico di Arcetri)

**Session Classification:** Hackathon

Contribution ID: 19

Type: **not specified**

## **Hands-on session 1 (data access and preparation)**

*Monday, 30 September 2024 14:00 (2 hours)*

**Session Classification:** Hands-on session

Contribution ID: 20

Type: **not specified**

## **Hands-on session 2 (data access and preparation)**

*Monday, 30 September 2024 16:00 (2 hours)*

**Session Classification:** Hands-on session

Contribution ID: 21

Type: **not specified**

## **Hands-on session 3 (Bayesian statistics)**

*Tuesday, 1 October 2024 14:20 (1h 40m)*

**Session Classification:** Hands-on session

Contribution ID: 22

Type: **not specified**

## **Hands-on session 4 (ML elements)**

*Tuesday, 1 October 2024 16:00 (1h 40m)*

**Session Classification:** Hands-on session



Contribution ID: 23

Type: **not specified**

## **Hands-on session 5 (ML elements)**

*Wednesday, 2 October 2024 14:20 (1h 45m)*

**Session Classification:** Hands-on session

Contribution ID: 24

Type: **not specified**

## **Hands-on session 6 (ML tools)**

*Wednesday, 2 October 2024 16:20 (1h 40m)*

**Project proposal: general context**

**Project proposal: description of the problem**

**Machine learning methods**

**Input dataset**

**Session Classification:** Hands-on session

Contribution ID: 25

Type: **not specified**

## **Hackathon session 1**

*Thursday, 3 October 2024 16:20 (1h 45m)*

**Session Classification:** Hackathon

Contribution ID: 26

Type: **not specified**

## **Hackathon session 1**

**Session Classification:** Hackathon

Contribution ID: 27

Type: **not specified**

## Hackathon session 3

*Friday, 4 October 2024 14:20 (1h 40m)*

**Session Classification:** Hackathon

Contribution ID: 28

Type: **not specified**

## Hackathon session 4

*Friday, 4 October 2024 16:20 (1h 10m)*

**Session Classification:** Hackathon

Contribution ID: 29

Type: **Hackathon project proposal**

# Classification of Order and Species of Mosquitoes

*Friday, 4 October 2024 10:40 (20 minutes)*

Conventional manual counting methods for the monitoring of mosquito species and populations can hinder the accurate determination of the optimal timing for pest control in the field. In this exercise is required to train a deep learning-based automated image analysis algorithm, for a two-fold task: the classification of different species and order of mosquito, based on a professionally made dataset of mosquitos photographs from multiple species.

## Project proposal: general context

Deep Neural Network base don CNN architectures, multi-classification taks, application on medical physics

## Project proposal: description of the problem

Analysis and preprocessing of the dataset (highly unbalanced among species), identification of the correct deep learning architecture, two tasks - two loss training, analysis of performances

## Machine learning methods

CNN

## Input dataset

input data in image format

## Goal and FOM

confusion matrix, accuracy, precision, recall, F1 score

**Primary authors:** CIARDIELLO, Andrea (Istituto Nazionale di Fisica Nucleare); GIAGU, Stefano (Sapienza Università di Roma and Istituto Nazionale di Fisica Nucleare)

**Session Classification:** Hackathon

Contribution ID: 30

Type: **Hackathon project proposal**

# Anomaly detection for new physics searches in HEP

*Friday, 4 October 2024 10:40 (20 minutes)*

Identify rare new physics process through an anomaly detection technique based on deep neural network (Graph Neural Network architecture).

Material for the exercise i.e. datasets and examples have been copied to the leonardo cluster and are available at:

/leonardo/home/usertrain/a08trb55/anomalyDetection/LHCO

## Project proposal: general context

High Energy Physics NP searches, Graph Neural Networks, Anomaly Detection

## Project proposal: description of the problem

Given a (pre-processed) dataset from fast simulation of a generic HEP detector containing a large number of events from Standard Model background processes and a test dataset containing both background and new physics signal events, design and train an anomaly detection model for anomaly detection of the NP processes.

## Machine learning methods

Graph Neural Networks and Auto-Encoder architectures

## Input dataset

Preprocessed data fro LHC OLYMPIC benchmark dataset, provided as numpy arrays

## Goal and FOM

ROC curves, AUC

**Primary authors:** CIARDIELLO, Andrea (Istituto Nazionale di Fisica Nucleare); GIAGU, Stefano (Sapienza Università di Roma and Istituto Nazionale di Fisica Nucleare)

**Session Classification:** Hackathon



Contribution ID: **31**

Type: **not specified**

## **Hackathon session 2**

**Session Classification:** Hackathon

Contribution ID: **32**

Type: **not specified**

## **Hackathon session 1**

**Session Classification:** Hackathon

Contribution ID: 33

Type: **Hackathon project proposal**

# Integrating Spliced and Unspliced Gene Expression Data for Improved Cell Type Annotation in Single-Cell RNA Sequencing

*Friday, 4 October 2024 10:40 (20 minutes)*

Cell type annotation is one of the primary tasks in single-cell RNA sequencing analysis and it is of a significance importance and difficulty in computational biology. This difficulty arises from two primary factors. First, gene expression levels typically exist on a continuum rather than being discrete, making it harder to draw clear boundaries between different cell types. Second, variations in gene expression don't always correlate with functional differences at the cellular level, adding complexity to the classification process.

Given these challenges, a more precise and reliable method for cell type annotation is always welcomed. We propose an approach that integrates additional layers of biological data, mainly spliced and unspliced expression, to look at gene expression estimates from a different perspective. By combining these pseudo-multi-modal layers of information, we might achieve more accurate cell type classification and better capture the functional nuances of different cell populations.

## Project proposal: general context

This project proposal focuses on computational biology, particularly the primary analysis of single-cell RNA sequencing (scRNA-seq) data and cell type annotation. scRNA-seq is essential for exploring gene expression at the single-cell level, enabling the identification of distinct cellular populations and their functional roles.

## Project proposal: description of the problem

The primary challenges in cell type annotation using single-cell RNA sequencing (scRNA-seq) arise from two main factors. First, gene expression levels exist on a continuum rather than discrete categories, making it difficult to delineate clear boundaries between different cell types. Second, variations in gene expression do not always correlate with functional differences at the cellular level, adding complexity to the classification process.

These challenges lead to potential inaccuracies in identifying cell types, which can hinder our understanding of cellular functions and the underlying biological processes. Developing a more reliable method that integrates diverse data types, such as spliced and unspliced gene expression, is essential to address these issues effectively.

## Machine learning methods

We will use community detection algorithm such as Louvain or Leiden algorithm to do unsupervised clustering for cells. Then we run statistical tests to retrieve genes associated to each cluster and annotate the cell accordingly.

## Input dataset

We will provide dataset for the hackathon. The data type is gene expression matrices and metadata, all stored and encoded in a well optimized data structure.

### **Goal and FOM**

We will use an already annotated dataset and compare the accuracy of annotation to that

**Primary authors:** SANGUINETTI, Guido (SISSA); KOUADRI BOUDJELTHIA, Idris

**Session Classification:** Hackathon

Contribution ID: 34

Type: **Hackathon project proposal**

# Dual Intelligence: Tackling Classification and Regression Challenges

*Friday, 4 October 2024 10:40 (20 minutes)*

We propose an hackathon divided into two challenges: in the first one, participants will tackle a regression problem using the well-known 'Asteroid Dataset', where they need to estimate the diameter of various types of asteroids. In the second challenge, participants will face a classification problem aimed at reconstructing the diagnosis of diabetes for different types of patients based on survey responses.

## Project proposal: general context

Astronomy and Health Physics.

## Project proposal: description of the problem

We are organizing an hackathon featuring two distinct challenges. The first challenge focuses on a regression task, where participants will use the renowned "Asteroid Dataset" to predict the diameter of asteroids based on various characteristics such as composition and orbital parameters. This problem will test the participants' ability to handle numerical and categorical data and build accurate predictive models.

The second challenge is centered on classification, where the goal is to determine the likelihood of diabetes diagnosis for a diverse set of patients. Using responses from a comprehensive health survey, participants will need to analyze and classify the data, applying machine learning techniques to predict whether a patient has diabetes based on lifestyle factors, medical history and other relevant features.

## Machine learning methods

Repeated cross validation, feature selection algorithms (Boruta), classical statistical tools, data manipulation techniques, data augmentation techniques (SMOTE), Linear Regression, Logistic Regression, Random Forest Classifier/Regressor, XGBoost Classifier/Regressor.

## Input dataset

- Asteroid dataset: 130,000 samples, 40 features - csv file.
- Diabetes dataset: 400,000 samples, 30 features - csv file.

Currently, we are in doubt about what data storage to use but we'll solve this problem quickly.

## Goal and FOM

Regression task: coefficient of determination,  $R^2$ .

Classification task: F1 score.

**Primary authors:** Dr LORUSSO, Giovanni (Università degli Studi di Bari Aldo Moro); CARUSO, Mario (Università degli Studi di Bari Aldo Moro)

**Session Classification:** Hackathon

Contribution ID: 35

Type: **Hackathon project proposal**

# Exploitation of jets features to tag VBF like events in high energy particle physics

*Friday, 4 October 2024 10:40 (20 minutes)*

The goal of the project is to develop a tagger for distinguishing VBF-like signal events from background ones using properties of the two leading jets (tag jets) and eventually the third one. The key idea is to exploit the color flow patterns in signal events, where the QCD activity between jets is greater between each jet and the beam spot with respect to the region between the two tag jets. Relevant features for the tagger may include jet kinematic properties, jet substructure variables, and color flow observables. Signal (VBF) and background (Drell Yan) samples will be provided for training and validation, the study will be done at parton level.

## Project proposal: general context

The Large Hadron Collider (LHC) at CERN is the world's most powerful particle accelerator, designed to collide protons (and heavy ions) at extremely high energies, reaching up to 13.6 TeV and operating at a collision rate of 40 MHz. However, most of these proton-proton collisions result in low-energy "soft" interactions, which do not produce particles of interest for high-energy physics studies. To manage the high event rate, the LHC experiments employ a trigger system that selects and saves only the most promising, "hard scattering" events for further analysis.

Among the four main experiments at the LHC, ATLAS and CMS have a broad physics program aimed at exploring both the Standard Model (SM) and potential new physics. One process of particular interest to study the electroweak sector of the standard model is Vector Boson Fusion (VBF). In VBF a quark from each of the incoming LHC protons radiates off a heavy vector boson. These bosons interact to produce another boson. The initial quarks that first radiated the vector bosons are deflected only slightly and travel roughly along their initial directions, producing two forward jets (tag jets) with minimal activity between them

## Project proposal: description of the problem

While the trigger system of the experiments effectively filters out many soft events, background processes that mimic the signal of interest can still pass through. In the case of VBF, distinguishing this signal from background events like Drell-Yan remains challenging. In a Drell-Yan process, a quark of one proton and an antiquark of another proton annihilate, creating a virtual photon or Z boson which then decays into a pair of oppositely charged leptons.

Both processes can produce similar final-state particles, making it difficult to differentiate between signal and background. Therefore, it is necessary to develop a more sophisticated classifier that can accurately discriminate VBF events from background events based on jet kinematic properties, substructure, and color flow patterns.

## Machine learning methods

The task is classification/tagging. A classifier would work, but a different approach with a CNN that exploits the tag jets images could be interesting to explore.

## Input dataset

the notebooks provided here:

[https://github.com/AuroraPerego/VBS\\_tagger/tree/tagger/notebooks](https://github.com/AuroraPerego/VBS_tagger/tree/tagger/notebooks)

can be used to read and plot the data

two files, one for the signal (70MB) and one for the background (80 MB), each of them containing 10k events, saved as ROOT TTrees.

## Goal and FOM

the target can be the ROC and the AUC, the goal is to have high purity for the tagging.

**Primary author:** PEREGO, Aurora (Istituto Nazionale di Fisica Nucleare)

**Session Classification:** Hackathon



Contribution ID: 36

Type: **not specified**

## **Hackathon: final reports**

*Friday, 4 October 2024 17:30 (20 minutes)*

Each group should prepare 3 slides:

- 1) Problem statement
- 2) Chosen architecture and strategy
- 3) Results (showing the FOM indicated in the challenge)

Send us the slides by email at [aiphy@unimib.it](mailto:aiphy@unimib.it)

### **Project proposal: general context**

### **Project proposal: description of the problem**

### **Machine learning methods**

### **Input dataset**

### **Goal and FOM**

**Session Classification:** Hackathon