Integrating AI tools in data analysis frameworks: the Vera Rubin LSST and Euclid cases

Riccio G.[1], Cavuoti S.[1], Angora G.[1], Brescia M.[2,1]

1. INAF - Astronomical Observatory of Capodimonte, Napoli (Italy)
1. Department of Physics “E. Pancini”, University of Naples Federico II (Italy)

IVOA Interoperability Meeting -
15 - 17 November 2024 – La Valletta, Malta

Context

In the last two decades, Astronomy has been the scene of the realization of panchromatic surveys, with sophisticated instruments acquiring a huge amount of exceptional quality data.

- ESA Euclid : ~100 GB/day for 6 years → 200 TB
- Rubin/LSST : ~20 TB/night for 10 years → >60 PB
- JWST : ~30GB/day for 10 years (and more)
- GAIA : ~1 PB in 5 year
- SKA : 100 Pbytes – 3 EBytes/year
- Pan-STARRS, KiDS, DES, Herschel-ATLAS, Hi-GAL, E-ELT…

NEEDS

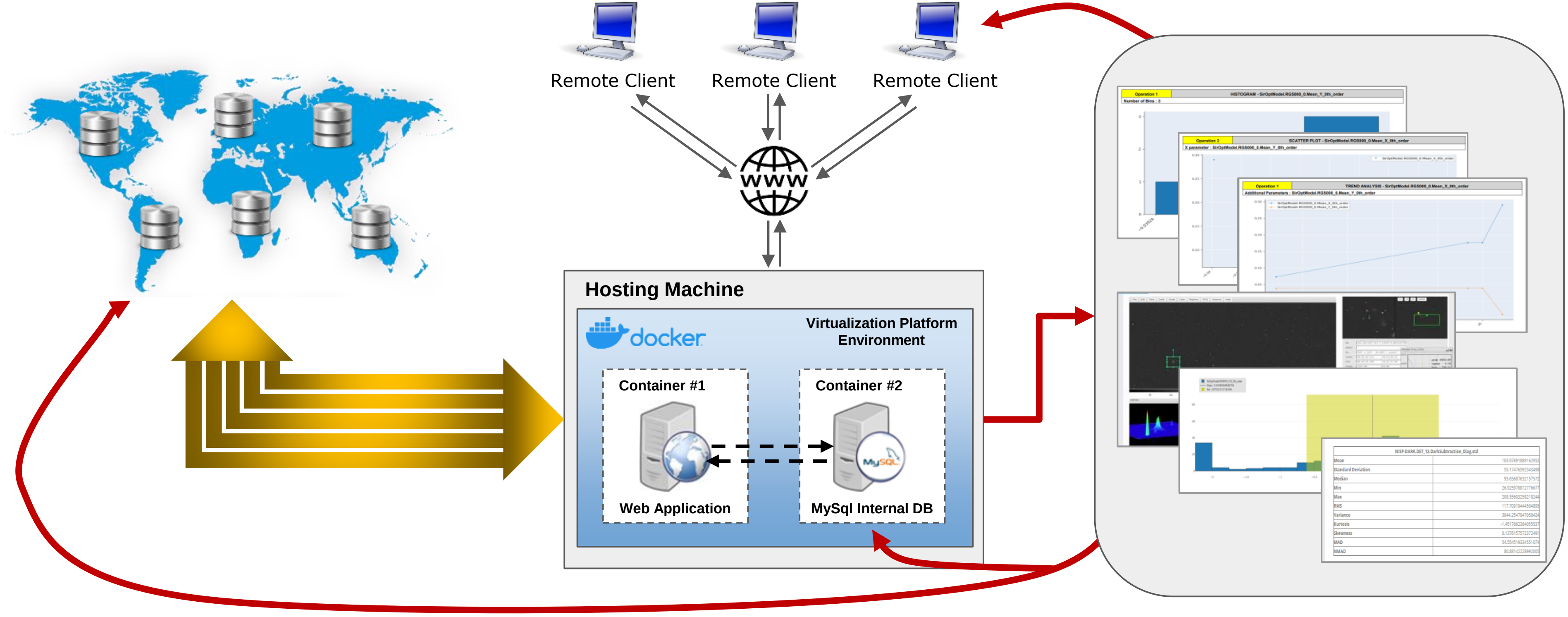
- to integrate advanced **data-driven** science methodologies for the **automatic exploration** of huge and multi-dimensional **data archives**
- efficient short- and long-term **monitoring** and **diagnostics** systems

GOALS

- To keep the **quality** of the observations under control
- To detect and circumscribe **anomalies** and **malfunctions**
- To facilitate rapid and effective **corrections**
- To ensure correct maintenance of all components and the **good health of scientific data** over time mainly crucial for space-borne observation systems, both in logistical and economic terms

AIDA - Advanced Infrastructure for Data Analysis

AIDA is a portable and modular web application, designed to provide an efficient and intuitive software infrastructure to support monitoring of data acquiring systems over time, diagnostics and both scientific and engineering data quality analysis, particularly suited for astronomical instruments



Operation 1 | HISTOGRAM - SirOptModel.RGS000_0.Mean_Y_0th_order

Number of Bins : 3

Operation 2 | SCATTER PLOT - SirOptModel.RGS000_0.Mean_X_0th_order

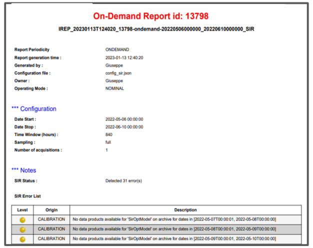
X parameter : SirOptModel.RGS000_0.Mean_Y_0th_order

Operation 1 | TREND ANALYSIS - SirOptModel.RGS000_0.Mean_X_0th_order

Additional Parameters : SirOptModel.RGS000_0.Mean_Y_0th_order

NISP-DARK.DET_12.DarkSubtraction_Diag.std	
Mean	103.97691889162952
Standard Deviation	55.17476592343498
Median	93.65687632157572
Min	26.925978812776677
Max	208.55603258218244
RMS	117.70919444504895
Variance	3044.2547947058424
Kurtosis	-1.4517662364055537
Skewness	0.1376157572372491
MAD	54.554519334551074
NMAD	80.88142228992005

AIDA - Main Features



On-Demand Report id: 13798

IREP_20230113T124020_13798-ondemand-20220506000000_20220610000000_SIR

Report Periodicity	ONDEMAND
Report generation time :	2023-01-13 12:40:20
Generated by :	Giuseppe
Configuration file :	config_sir.json
Owner :	Giuseppe
Operating Mode :	NOMINAL

*** Configuration

Date Start :	2022-05-06 00:00:00
Date Stop :	2022-06-10 00:00:00
Time Window (hours) :	840
Sampling :	full
Number of acquisitions :	1

*** Notes

SIR Status :	Detected 31 error(s)

SIR Error List

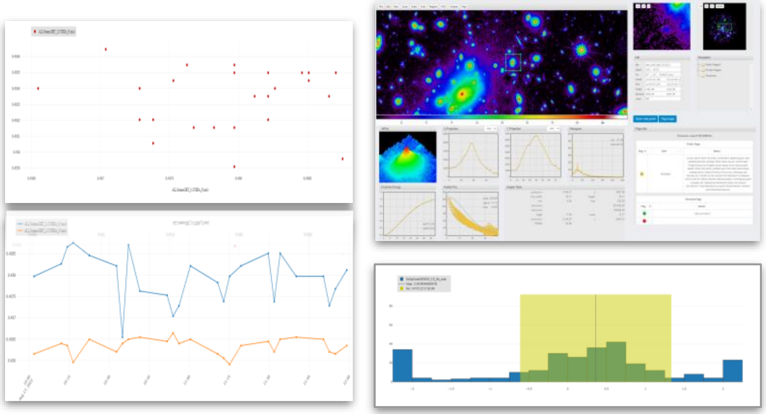
Level	Origin	Description
●	CALIBRATION	No data products available for 'SirOptModel' on archive for dates in [2022-05-07T00:00:01, 2022-05-08T00:00:00]
●	CALIBRATION	No data products available for 'SirOptModel' on archive for dates in [2022-05-08T00:00:01, 2022-05-09T00:00:00]
●	CALIBRATION	No data products available for 'SirOptModel' on archive for dates in [2022-05-09T00:00:01, 2022-05-10T00:00:00]

Instrument monitoring, report generation and delivery

- ✓ **periodic report** generation on a user-defined parameters list and delivery to remote archive
- ✓ **on demand customised report** generation on a user selected parameter list, locally stored

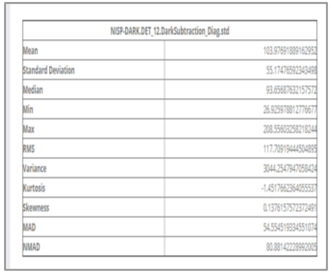
Visualization/Exploration

- ✓ **series of plots** on user selected parameters/data products and ranges
- ✓ **pre-generated histograms** stored into remote archives
- ✓ **observed images** (static view, dynamic windowing, statistical characterization)



Statistics

- ✓ **standard** (default) estimators
- ✓ **special** estimations (tables/images)
- ✓ statistical analysis on **image pixels**



NISP-DARK.DET_12.DarkSubtraction_Diag.std	
Mean	103.97691889162952
Standard Deviation	55.17476592343498
Median	93.65687632157572
Min	26.92597881277667
Max	208.55603258218244
RMS	117.70919444504895
Variance	3044.254794705842
Kurtosis	-1.4517662364055537
Skewness	0.1376157572372491
MAD	54.55451933455107
NMAD	80.88142228992005

Machine Learning

- ✓ **Regression/classification** experiments on available data

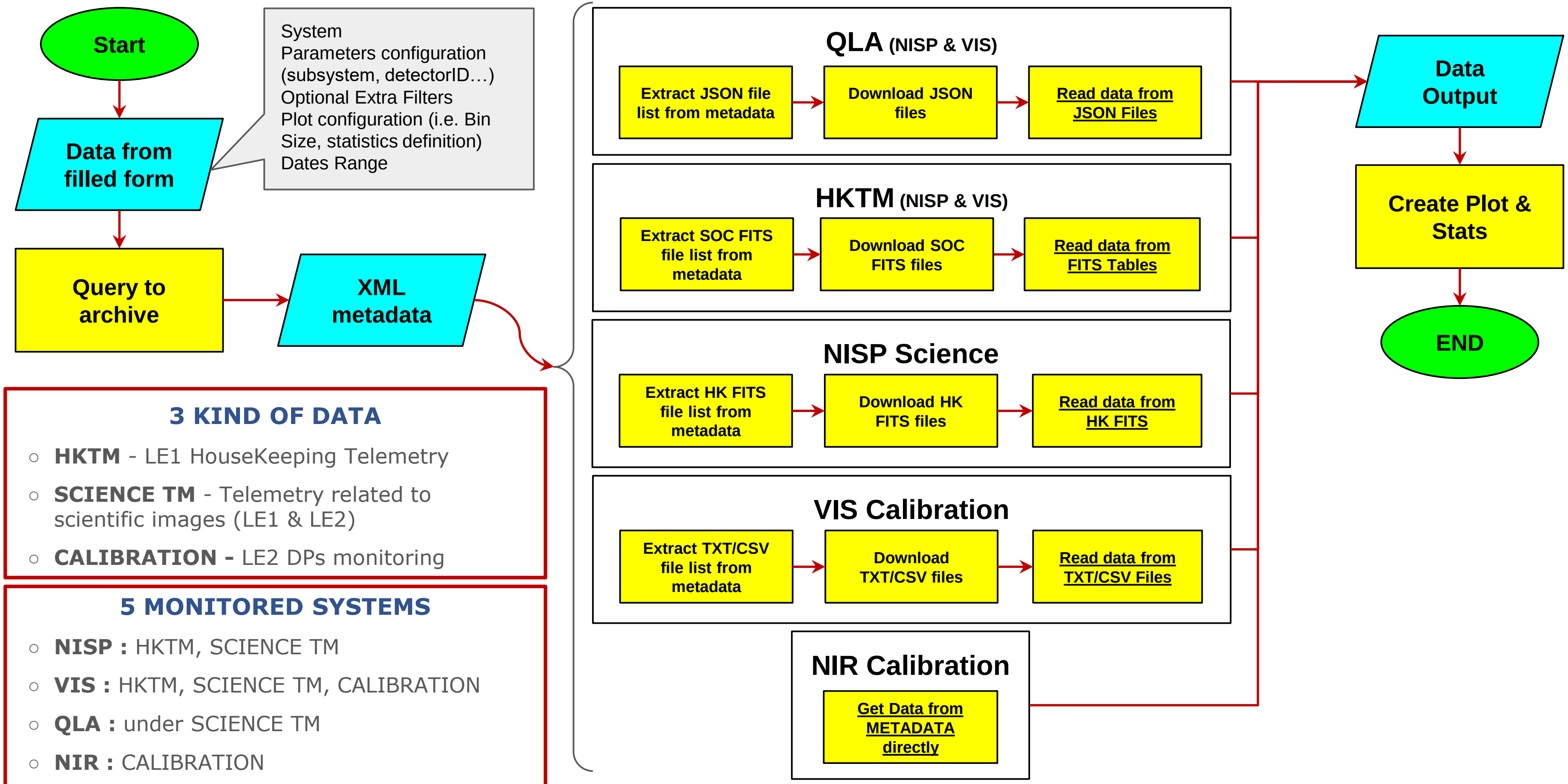
Code Name : Flexibility

AIDA has been designed as a **modular system**, based on **Object-Oriented Programming** and specific information on DB, so it is possible to **extend its functionalities**, by integrating and customizing monitoring and diagnostics systems, as well as scientific data analysis solutions, including machine/deep learning and data mining methods

- **Available plots and statistics are defined as classes/functions** linked to a specific table in DB. To add a new operation, it is sufficient to implement the related class/function and add it to the local DB;
- **A JSON configuration file is associated to every system monitored by AIDA**. It includes info about the instrument and connection to the related data and metadata archives;
- **Repositories and systems have a dedicated classes** which implement methods for interfacing AIDA with the data repository. To add a new system/repository, it is sufficient to create its own configuration file (only for systems), implement the related class and methods, and fill DB with required information.

A specialized version of AIDA is already the **official monitoring** and analysis tool for the **ESA Euclid space mission** and another one is going to be used for the commissioning of the **V. Rubin Telescope, suitable also for LSST survey data**

AIDA/IODA for Euclid Data



Start

Data from filled form

System
Parameters configuration (subsystem, detectorID…)
Optional Extra Filters
Plot configuration (i.e. Bin Size, statistics definition)
Dates Range

Query to archive

XML metadata

QLA (NISP & VIS)

Extract JSON file list from metadata → Download JSON files → Read data from JSON Files

HKTM (NISP & VIS)

Extract SOC FITS file list from metadata → Download SOC FITS files → Read data from FITS Tables

NISP Science

Extract HK FITS file list from metadata → Download HK FITS files → Read data from HK FITS

VIS Calibration

Extract TXT/CSV file list from metadata → Download TXT/CSV files → Read data from TXT/CSV Files

NIR Calibration

Get Data from METADATA directly

Data Output

Create Plot & Stats

END

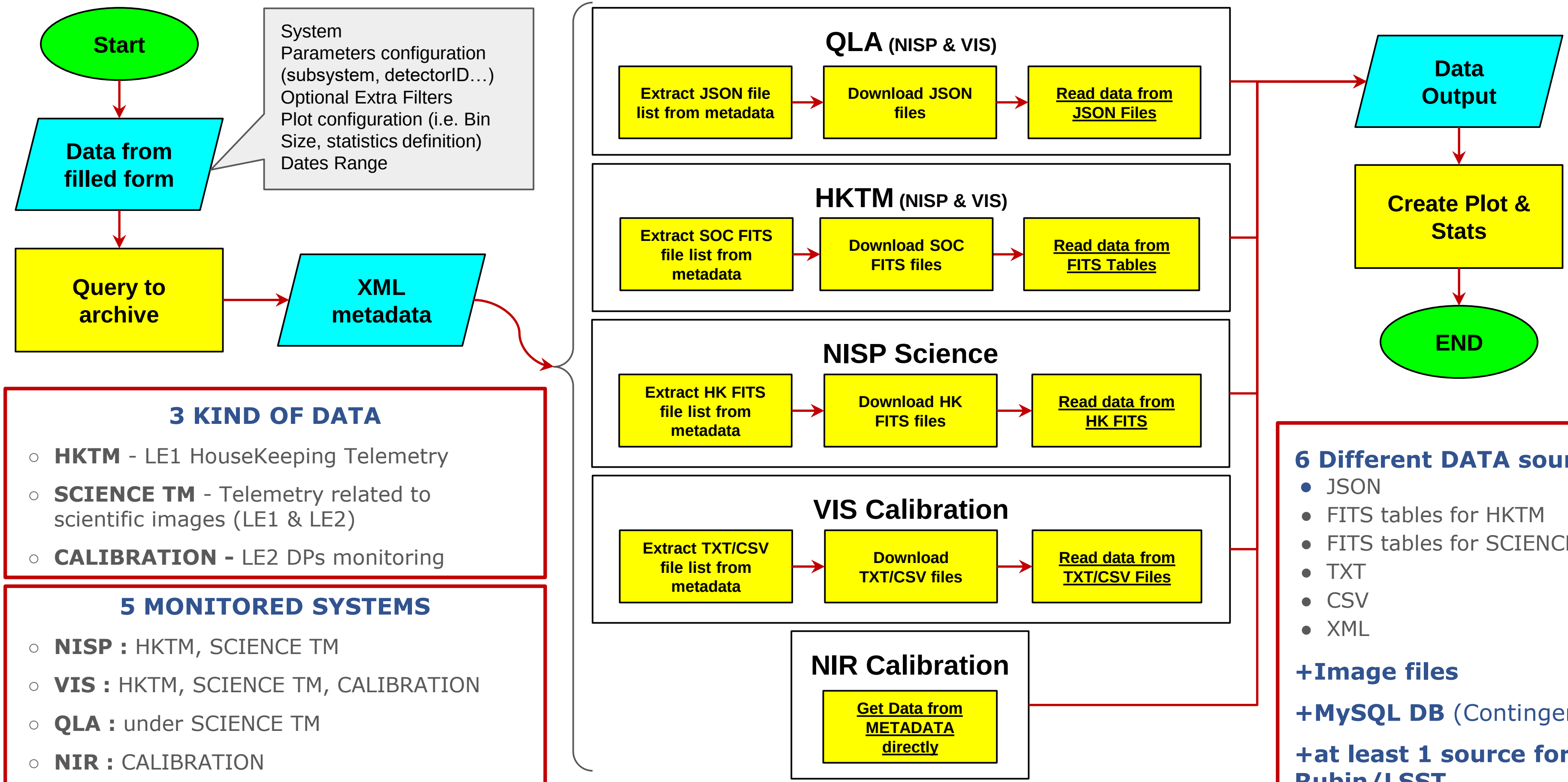
3 KIND OF DATA

- **HKTM** - LE1 HouseKeeping Telemetry
- **SCIENCE TM** - Telemetry related to scientific images (LE1 & LE2)
- **CALIBRATION -** LE2 DPs monitoring

5 MONITORED SYSTEMS

- **NISP :** HKTM, SCIENCE TM
- **VIS :** HKTM, SCIENCE TM, CALIBRATION
- **QLA :** under SCIENCE TM
- **NIR :** CALIBRATION
- **SIR :** CALIBRATION

AIDA/IODA for Euclid Data



Start

Data from filled form

System
Parameters configuration (subsystem, detectorID…)
Optional Extra Filters
Plot configuration (i.e. Bin Size, statistics definition)
Dates Range

Query to archive

XML metadata

QLA (NISP & VIS)
- Extract JSON file list from metadata → Download JSON files → Read data from JSON Files

HKTM (NISP & VIS)
- Extract SOC FITS file list from metadata → Download SOC FITS files → Read data from FITS Tables

NISP Science
- Extract HK FITS file list from metadata → Download HK FITS files → Read data from HK FITS

VIS Calibration
- Extract TXT/CSV file list from metadata → Download TXT/CSV files → Read data from TXT/CSV Files

NIR Calibration
- Get Data from METADATA directly

Data Output

Create Plot & Stats

END

3 KIND OF DATA

- **HKTM** - LE1 HouseKeeping Telemetry
- **SCIENCE TM** - Telemetry related to scientific images (LE1 & LE2)
- **CALIBRATION -** LE2 DPs monitoring

5 MONITORED SYSTEMS

- **NISP :** HKTM, SCIENCE TM
- **VIS :** HKTM, SCIENCE TM, CALIBRATION
- **QLA :** under SCIENCE TM
- **NIR :** CALIBRATION
- **SIR :** CALIBRATION

6 Different DATA sources

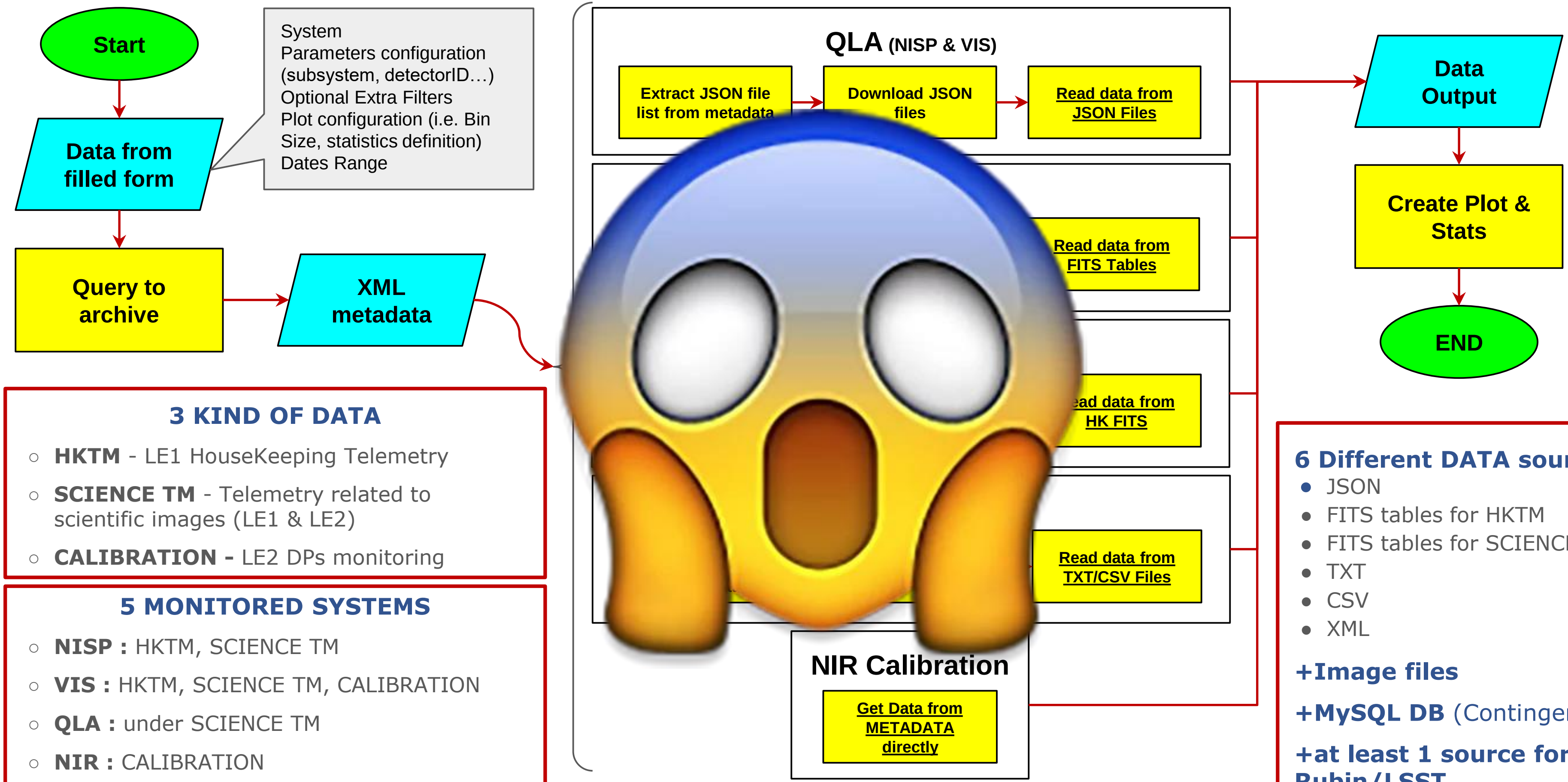
- JSON
- FITS tables for HKTM
- FITS tables for SCIENCE
- TXT
- CSV
- XML

+Image files

+MySQL DB (Contingency)

+at least 1 source for Rubin/LSST

AIDA/IODA for Euclid Data


Start
Data from filled form
System
Parameters configuration (subsystem, detectorID…)
Optional Extra Filters
Plot configuration (i.e. Bin Size, statistics definition)
Dates Range
Query to archive
XML metadata
QLA (NISP & VIS)
Extract JSON file list from metadata
Download JSON files
Read data from JSON Files
Read data from FITS Tables
ead data from HK FITS
Read data from TXT/CSV Files
NIR Calibration
Get Data from METADATA directly
Data Output
Create Plot & Stats
END

3 KIND OF DATA

- **HKTM** - LE1 HouseKeeping Telemetry
- **SCIENCE TM** - Telemetry related to scientific images (LE1 & LE2)
- **CALIBRATION -** LE2 DPs monitoring

5 MONITORED SYSTEMS

- **NISP :** HKTM, SCIENCE TM
- **VIS :** HKTM, SCIENCE TM, CALIBRATION
- **QLA :** under SCIENCE TM
- **NIR :** CALIBRATION
- **SIR :** CALIBRATION

6 Different DATA sources

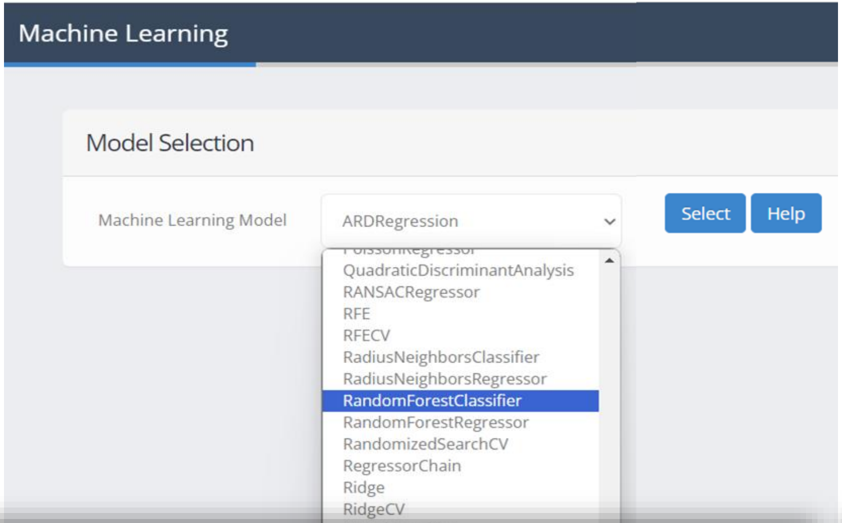
- JSON
- FITS tables for HKTM
- FITS tables for SCIENCE
- TXT
- CSV
- XML

+Image files

+MySQL DB (Contingency)

+at least 1 source for Rubin/LSST

Machine Learning Tools



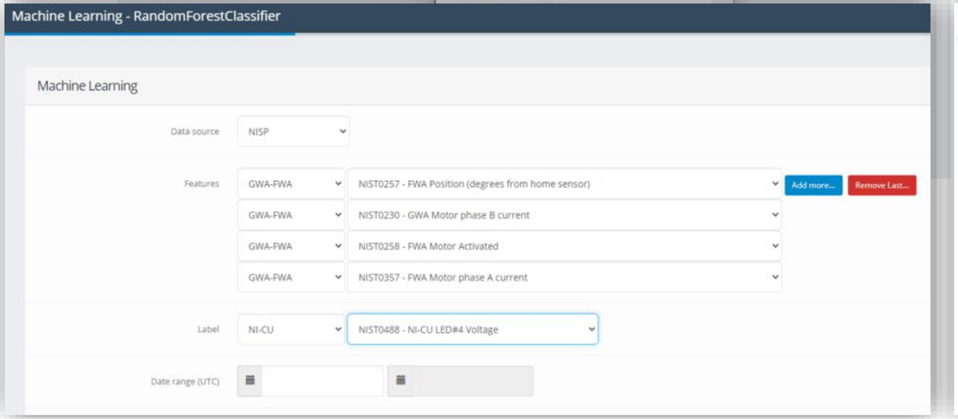
Machine Learning

Model Selection

Machine Learning Model: ARDRegression

The tool includes **more than 100 prediction, classification and regression models** based on **Machine Learning** to apply on available tabular data, useful in this case to identify operating anomalies or correlations between instrumental information.

Deep Learning methods coding is on going



Machine Learning - RandomForestClassifier

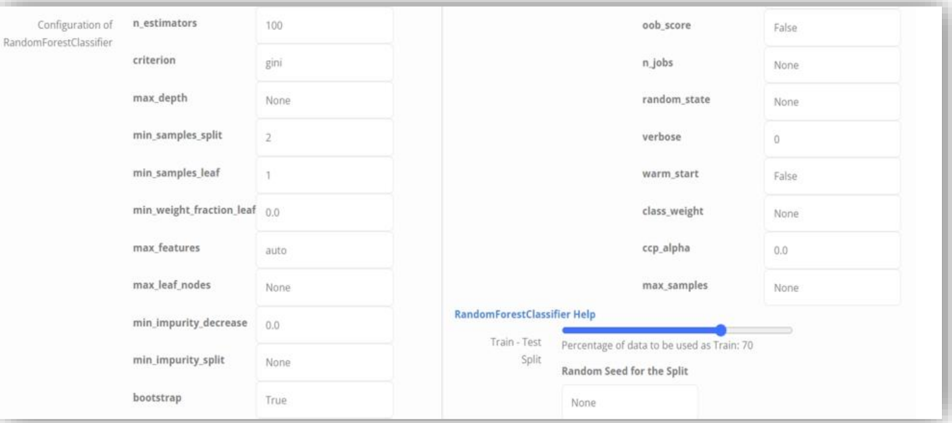
Machine Learning

Data source: NISP

Features:
- GWA-FWA — NIST0257 - FWA Position (degrees from home sensor)
- GWA-FWA — NIST0230 - GWA Motor phase B current
- GWA-FWA — NIST0258 - FWA Motor Activated
- GWA-FWA — NIST0357 - FWA Motor phase A current

Label: NI-CU — NIST0488 - NI-CU LED#4 Voltage

Date range (UTC)



Configuration of RandomForestClassifier

n_estimators	100
criterion	gini
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	auto
max_leaf_nodes	None
min_impurity_decrease	0.0
min_impurity_split	None
bootstrap	True

oob_score	False
n_jobs	None
random_state	None
verbose	0
warm_start	False
class_weight	None
ccp_alpha	0.0
max_samples	None

RandomForestClassifier Help

Train - Test Split: Percentage of data to be used as Train: 70

Random Seed for the Split: None

Summary

- ✓ The **AIDA web application** has been designed to provide an **efficient and intuitive software infrastructure** to support **monitoring** of data acquisition systems over time, **diagnostics** and both scientific and engineering **data quality analysis**, in particular for astronomical instruments

- ✓ It provides **a number of tools** for data analysis & system diagnostics

  - ❑ **Instrument monitoring, report generation and delivery**
  - ❑ **Visualization Exploration**
  - ❑ **Statistics**
  - ❑ **Machine/Deep Learning**

- ✓ a specific version of AIDA is already the **official monitoring** and analysis tool for the **ESA Euclid space mission** and another one is going to be used for the commissioning of the **V. Rubin Telescope, suitable also for LSST survey data**

- ✓ Being designed as a modular system, **it is possible to integrate and customize** monitoring and diagnostics systems, as well as scientific data analysis solutions

An **high level of standardization** for data and tools is crucial to easily customize AIDA to have a **general infrastructure** for as many astronomical projects as possible

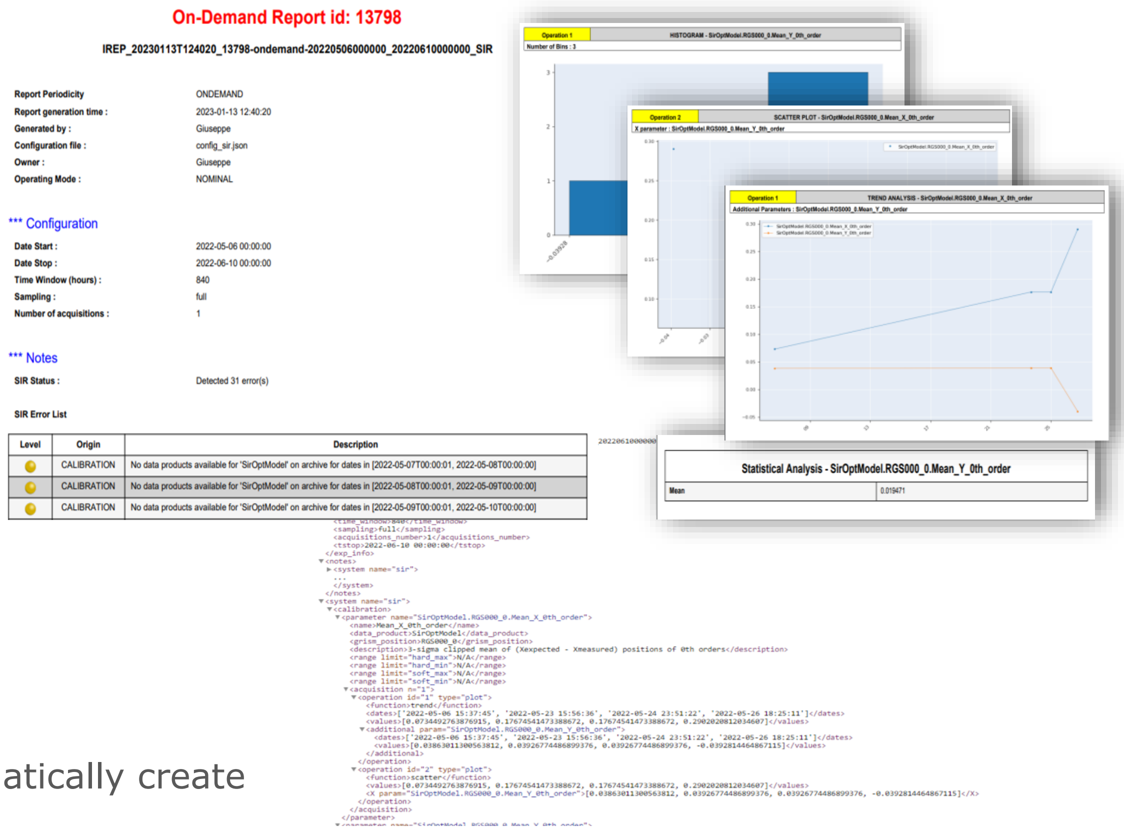
Ideas for next AI tools (1)

REPORT

Periodic automatic or on-demand generation

```
{
   "General Info": {
      "Start Time": "2022-05-06T00:00:00",
      "Time Window": 840,
      "Number of acquisitions": 1,
      "Sampling": "full"
   },
   "SIR"   :   {
      "CALIBRATION"   :   {
         "SirOptModel" : {
            "RGS000_0" : {
               "Mean_X_0th_order" :{
                  "Operation_1" : {
                     "Type" : "trend",
                     "Additional Parameters" : ["SirOptModel.RGS000_0.Mean_Y_0th_order"]
                  },
                  "Operation_2" : {
                     "Type" : "scatter",
                     "X" : "SirOptModel.RGS000_0.Mean_Y_0th_order"
                  }
               },
               "Mean_Y_0th_order" : {
                  "Operation_1" : {
                     "Type" : "histogram",
                     "Number of Bins" : 3
                  },
                  "Operation_2" : {
                     "Type" : "mean"
                  }
               }
            }
         }

      }
   }
}
```



On-Demand Report id: 13798

IREP_20230113T124020_13798-ondemand-20220506000000_20220610000000_SIR

Report Periodicity	ONDEMAND
Report generation time :	2023-01-13 12:40:20
Generated by :	Giuseppe
Configuration file :	config_sir.json
Owner :	Giuseppe
Operating Mode :	NOMINAL

*** Configuration

Date Start :	2022-05-06 00:00:00
Date Stop :	2022-06-10 00:00:00
Time Window (hours) :	840
Sampling :	full
Number of acquisitions :	1

*** Notes

SIR Status :	Detected 31 error(s)

SIR Error List

Level	Origin	Description
●	CALIBRATION	No data products available for 'SirOptModel' on archive for dates in [2022-05-07T00:00:01, 2022-05-08T00:00:00]
●	CALIBRATION	No data products available for 'SirOptModel' on archive for dates in [2022-05-08T00:00:01, 2022-05-09T00:00:00]
●	CALIBRATION	No data products available for 'SirOptModel' on archive for dates in [2022-05-09T00:00:01, 2022-05-10T00:00:00]

2022061000000

Operation 1	HISTOGRAM - SirOptModel.RGS000_0.Mean_Y_0th_order
Number of Bins : 3	

Operation 2	SCATTER PLOT - SirOptModel.RGS000_0.Mean_X_0th_order
X parameter : SirOptModel.RGS000_0.Mean_Y_0th_order	

Operation 1	TREND ANALYSIS - SirOptModel.RGS000_0.Mean_X_0th_order
Additional Parameters : SirOptModel.RGS000_0.Mean_Y_0th_order	

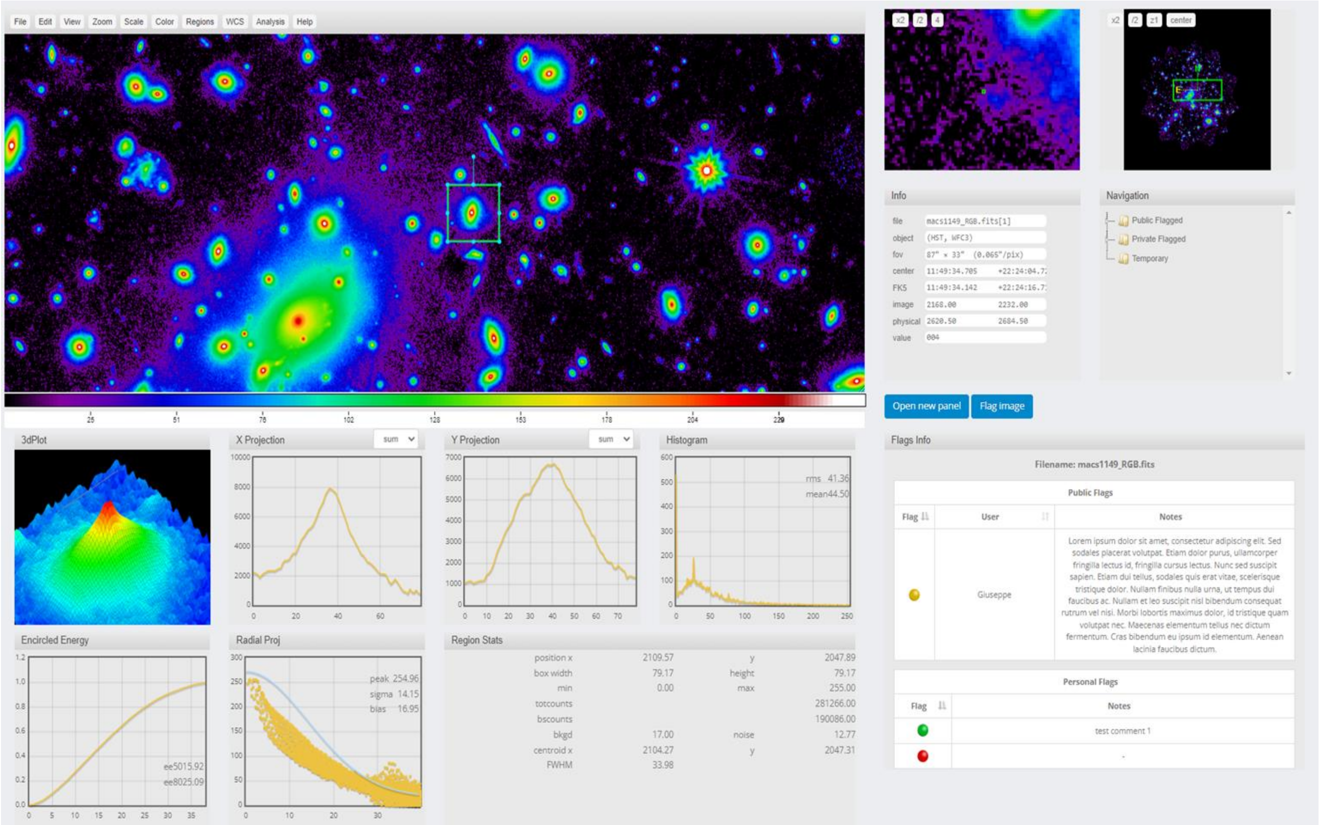
3
2
1
0
-0.03928
0.30
0.25
0.20
0.15
0.10
-0.04
-0.03
SirOptModel.RGS000_0.Mean_X_0th_order
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
SirOptModel.RGS000_0.Mean_X_0th_order
SirOptModel.RGS000_0.Mean_Y_0th_order
09
13
17
21
25

Statistical Analysis - SirOptModel.RGS000_0.Mean_Y_0th_order	
Mean	0.019471

```
    <time_window>840</time_window>
    <sampling>full</sampling>
    <acquisitions_number>1</acquisitions_number>
    <tstop>2022-06-10 00:00:00</tstop>
  </exp_info>
▼<notes>
  ▶<system name="sir">
    ...
    </system>
  </notes>
▼<system name="sir">
  ▼<calibration>
    ▼<parameter name="SirOptModel.RGS000_0.Mean_X_0th_order">
        <name>Mean_X_0th_order</name>
        <data_product>SirOptModel</data_product>
        <grism_position>RGS000_0</grism_position>
        <description>3-sigma clipped mean of (Xexpected - Xmeasured) positions of 0th orders</description>
        <range limit="hard_max">N/A</range>
        <range limit="hard_min">N/A</range>
        <range limit="soft_max">N/A</range>
        <range limit="soft_min">N/A</range>
      ▼<acquisition n="1">
        ▼<operation id="1" type="plot">
            <function>trend</function>
            <dates>['2022-05-06 15:37:45', '2022-05-23 15:56:36', '2022-05-24 23:51:22', '2022-05-26 18:25:11']</dates>
            <values>[0.0734492763876915, 0.17674541473388672, 0.17674541473388672, 0.2902020812034607]</values>
          ▼<additional param="SirOptModel.RGS000_0.Mean_Y_0th_order">
              <dates>['2022-05-06 15:37:45', '2022-05-23 15:56:36', '2022-05-24 23:51:22', '2022-05-26 18:25:11']</dates>
              <values>[0.03863011300563812, 0.03926774486899376, 0.03926774486899376, -0.0392814464867115]</values>
            </additional>
          </operation>
        ▼<operation id="2" type="plot">
            <function>scatter</function>
            <values>[0.0734492763876915, 0.17674541473388672, 0.17674541473388672, 0.2902020812034607]</values>
            <X param="SirOptModel.RGS000_0.Mean_Y_0th_order">[0.03863011300563812, 0.03926774486899376, 0.03926774486899376, -0.0392814464867115]</X>
          </operation>
        </acquisition>
      </parameter>
    ▼<parameter name="SirOptModel.RGS000_0.Mean_Y_0th_order">
```

Why not **a pre-trained LLM** to automatically create configuration files?

Ideas for next AI tools (2)



Info

file	macs1149_RGB.fits[1]	
object	(HST, WFC3)	
fov	87" × 33" (0.065"/pix)	
center	11:49:34.705	+22:24:04.7:
FK5	11:49:34.142	+22:24:16.7:
image	2168.00	2232.00
physical	2620.50	2684.50
value	004	

Region Stats

position x	2109.57	y	2047.89
box width	79.17	height	79.17
min	0.00	max	255.00
totcounts			281266.00
bscounts			190086.00
bkgd	17.00	noise	12.77
centroid x	2104.27	y	2047.31
FWHM	33.98		

Flags Info

Filename: macs1149_RGB.fits

Public Flags

Flag	User	Notes
●	Giuseppe	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed sodales placerat volutpat. Etiam dolor purus, ullamcorper fringilla lectus id, fringilla cursus lectus. Nunc sed suscipit sapien. Etiam dui tellus, sodales quis erat vitae, scelerisque tristique dolor. Nullam finibus nulla urna, ut tempus dui faucibus ac. Nullam et leo suscipit nisl bibendum consequat rutrum vel nisi. Morbi lobortis maximus dolor, id tristique quam volutpat nec. Maecenas elementum tellus nec dictum fermentum. Cras bibendum eu ipsum id elementum. Aenean lacinia faucibus dictum.

Personal Flags

Flag	Notes
🟢	test comment 1
🔴	-

A very useful tool could be a function, runnable from the Image Explorer panel, to **automatically generate thumbnails** from images, to be used by Deep Learning methods

A standard and automatic thumbnail extractor would be very useful for astronomical community in general

Based on JS9 Library : https://js9.si.edu/

May
the plots
be with you…

AIDA

IODA

Thank you!