Checking Plausibility in Exploratory Data Analysis

@ VLDB PhD Workshop – 16. August 2021

Hermann Stolte

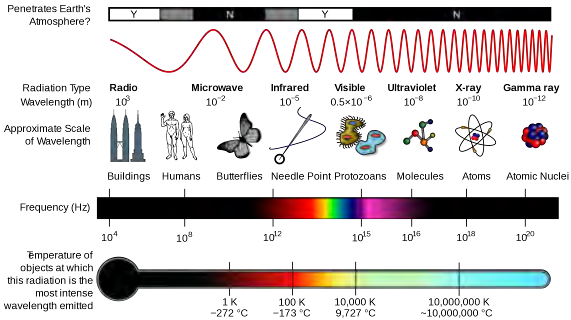
Humboldt-Universität zu Berlin

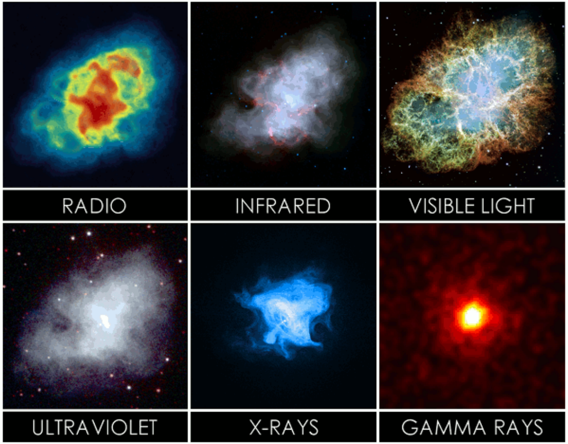
supervised by

Prof. Matthias Weidlich (HU)
Dr. Elisa Pueschel (DESY)

Gamma-Ray and Multi-Wavelength (MW) Astronomy

On the electromagnetic spectrum and multiple perspectives of reality





Credit: EarthSky (NASA, Wikipedia) CC BY-SA 3.0

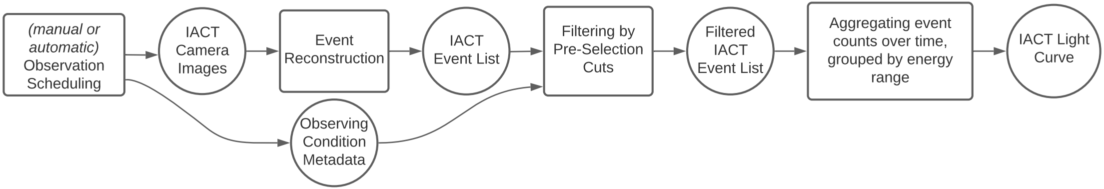
Credit: http://qdl.scs-inc.us/2ndParty/Pages/16754.html

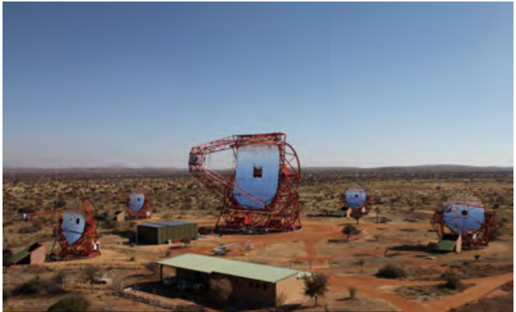
How to observe gamma rays?

Background: Cherenkov Light


Credit: DESY, Science Communication Lab

Imaging Atmospheric Cherenkov Technique (IACT)



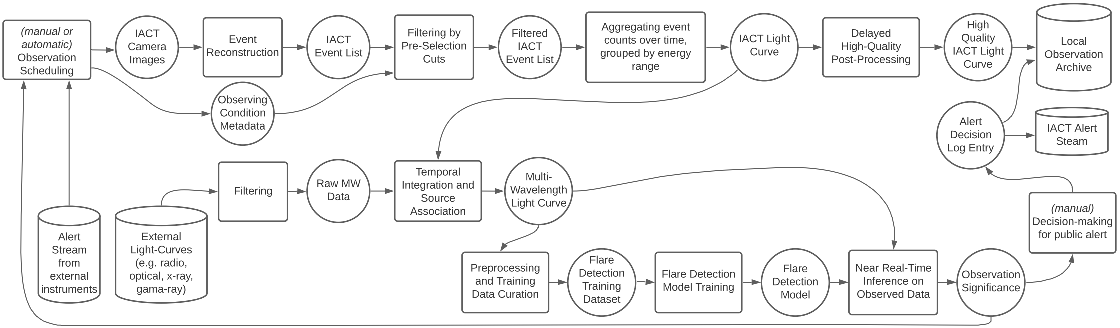


Credit: Christian99, CC BY-SA 3.0 via Wikimedia Commons

Challenges in the IACT data analysis:

- background noise
- data sparsity due to limited observing conditions
- bias in the data due to challenges in the event reconstruction

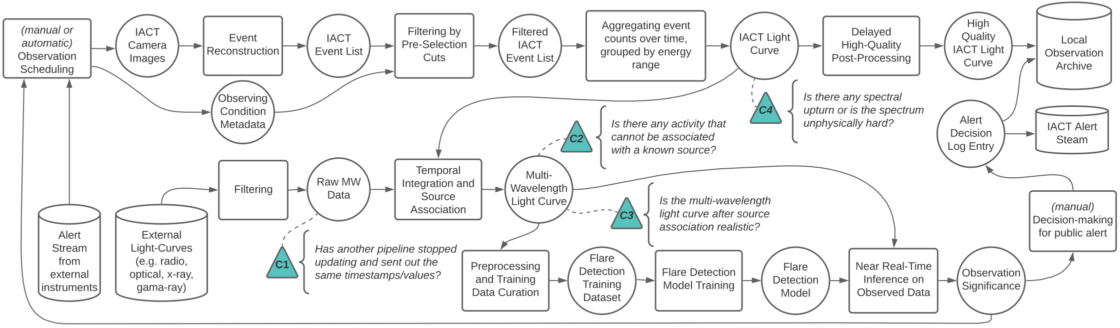
Pipeline for Exploratory Analysis of IACT and Multi-Wavelength Data



- working on this pipeline is error-prone, due to complex data and analysis steps
- how to avoid errors and have trust in results?

Plausibility Checking

In the Pipeline for IACT and MW Blazar Flare Detection



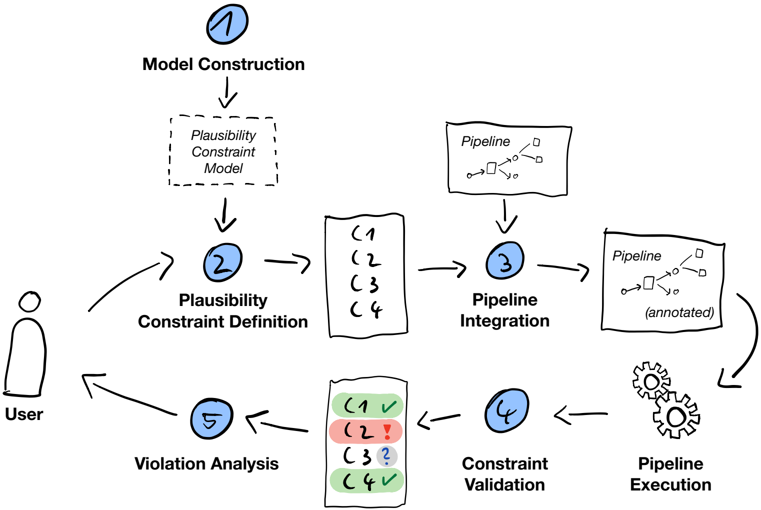
Two causes for implausible data:

- an error in the pipeline
- unexpected, interesting phenomena in the input data

Related Work

- Scientific Workflows, Workflow Engines
  - Survey – *Liew et al., 2016*
- Data Provenance Management
  - Survey – *Herschel et al., 2017*
  - ProvOne – *Cuevas-Vicenttın et al., 2016*
  - VisTrails – *Freire et al., 2012*
  - BugDoc – *Lourenço et al., 2020*
- Program Verification
  - Automatic Test Case Generation – *Anand et al., 2013*
  - Invariant Mining – *Lou et al., 2010*

Plausibility Checking: 5-Step Approach



Plausibility Checking: (1) Meta-Model for Plausibility Constraints

- Plausibility constraints are statements about data items of a pipeline
- The foundation: provenance graphs
  - mapping relations between data items
  - linking computational steps of a pipeline to data items

Figure: ProvONE: Data Model for Workflow Provenance

- A constraint function mapping from the data item domain to a plausibility distribution, expressing plausibility

Plausibility Checking: (2) Constraint Definition

How to minimize the user effort for defining plausibility constraints?

Idea:

- Constraints could be found and suggested to a user for review automatically
- For MW and IACT flare detection, constraints are often related to physical models
    - Can technical instrument specifications be leveraged to mine constraints?

Plausibility Checking: (3) Pipeline Integration

Idea: Create wrappers or hooks for common software components:

Before:

```
iact_light_curve = derive_light_curve(iact_event_list)
ctools.ctbutterfly(iact_light_curve)
```

After:

```
def plauscheck_ctbutterfly(iact_light_curve):
    entity = plauscheck.entities.IACT_LIGHT_CURVE
    constraintTypes = plauscheck.getConstraintsFor(entity)
    for constraintType in constaintTypes:
        constraint = constraintType(iact_light_curve)
        plauscheck.validate(constraint)
    ctools.ctbutterfly(iact_light_curve)

iact_light_curve = derive_light_curve(iact_event_list)
plauscheck_ctbutterfly(iact_light_curve)
```

Plausibility Checking: (4) Constraint Validation

How to check a constraint on real data?

Challenges:

- data sparsity
- data uncertainty
- multi-resolution data

Direction: Probabilistic Constraint Validation:

- real data may not be suitable for validation a constraint in extreme cases
- confidence may change with more data being available over time

Plausibility Checking: (5) Violation Analysis

How to support the user in finding the root cause for a violation?

A violation could be caused by...

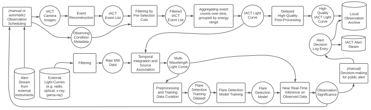
- an error in the pipeline
- unexpected, interesting phenomena in the input data

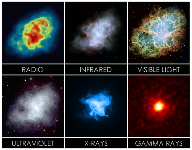
Ideas:

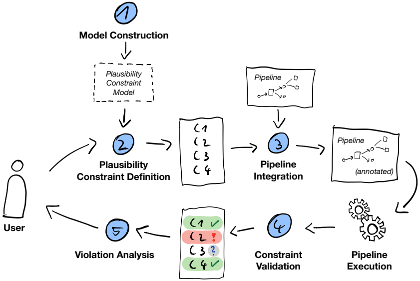
- Identify data items that correlate with a constraint violation
- Outlier detection on upstream data items, possibly showing abnormal trends closer to the root cause

Summary and Next steps

Pipeline for detecting novel trends in IACT and MW data based using unsupervised and probabilistic machine learning







Plausibility Checking for Exploratory Data Analysis:

- Create extended taxonomy of plausibility constraints, narrow down the focus
- How to infer plausibility constraints, e.g. from physical models underlying the blazar flare detection pipeline?