Lessons Learned from Applying Machine Learning to the Data Analysis Pipeline of the COSI Telescope

Andreas Zoglauer
Space Sciences Laboratory & Berkeley Institute for Data Science
(UC Berkeley)

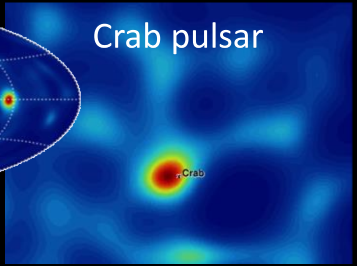
COSI - The Compton Spectrometer and Imager

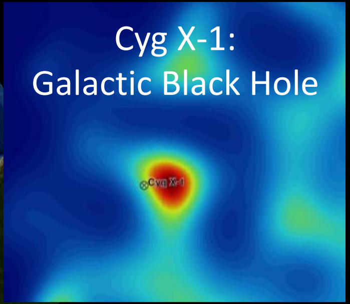
Telescope & Flight:

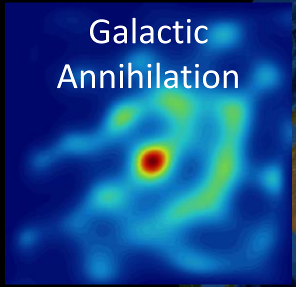
- Balloon-borne gamma-ray telescope
- Flight altitude: 110,000 feet
- 2016: Flight from New Zealand
- Next planned flight: 2019/20

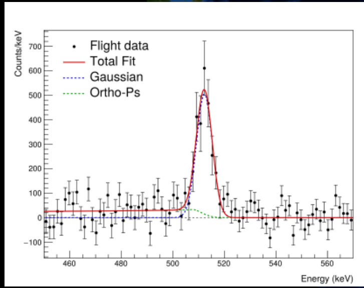
Science goals:

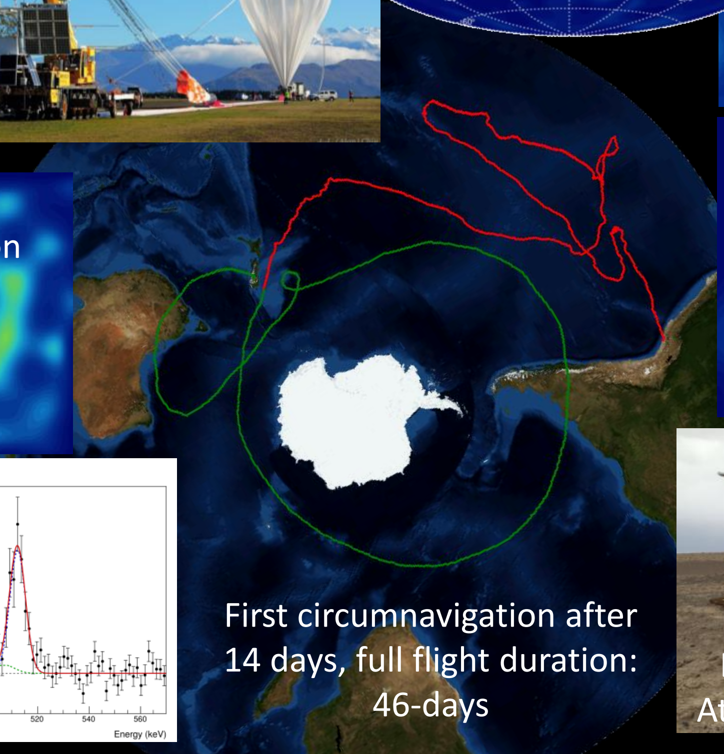
- Observe the most violent events (supernovae, neutron star mergers)
- Observe the most extreme environments (pulsars, black holes)
- Better understand the life cycle of (anti-) matter in our Galaxy









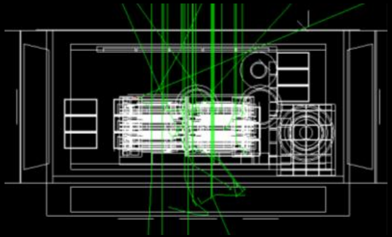


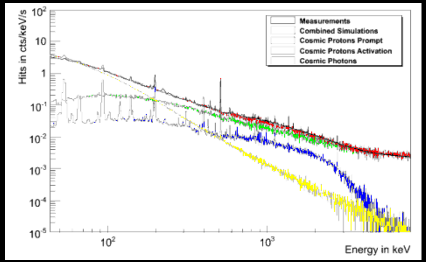
The Analysis Toolkit: MEGAlib

Medium-Energy Gamma-ray Astronomy library:

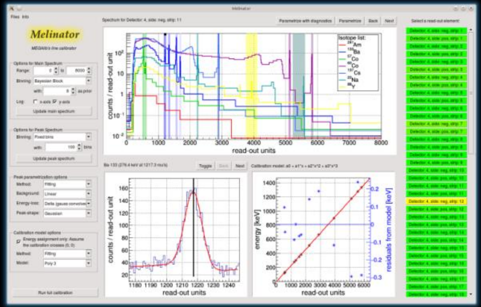
- Full data analysis chain for γ-ray instruments in space & on ground
- Free & open-source: http://github.com/zoglauer/megalib
- Generalized to be applied to arbitrary detector systems not only COSI

Monte-Carlo simulations

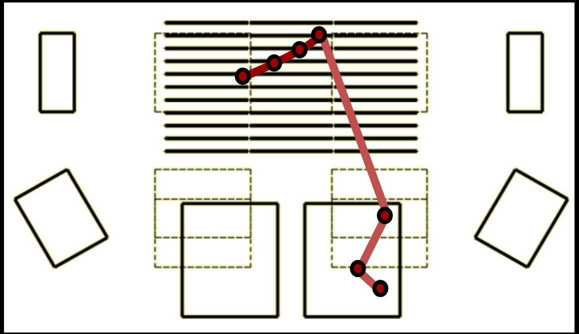


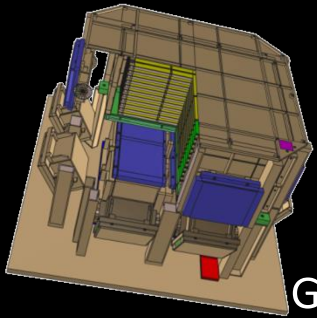


Detector calibrations



Event pattern classification





Geometry

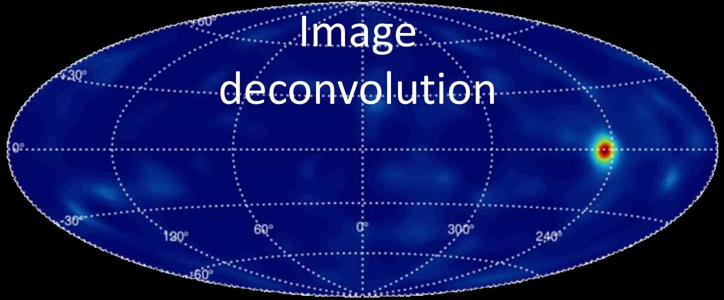
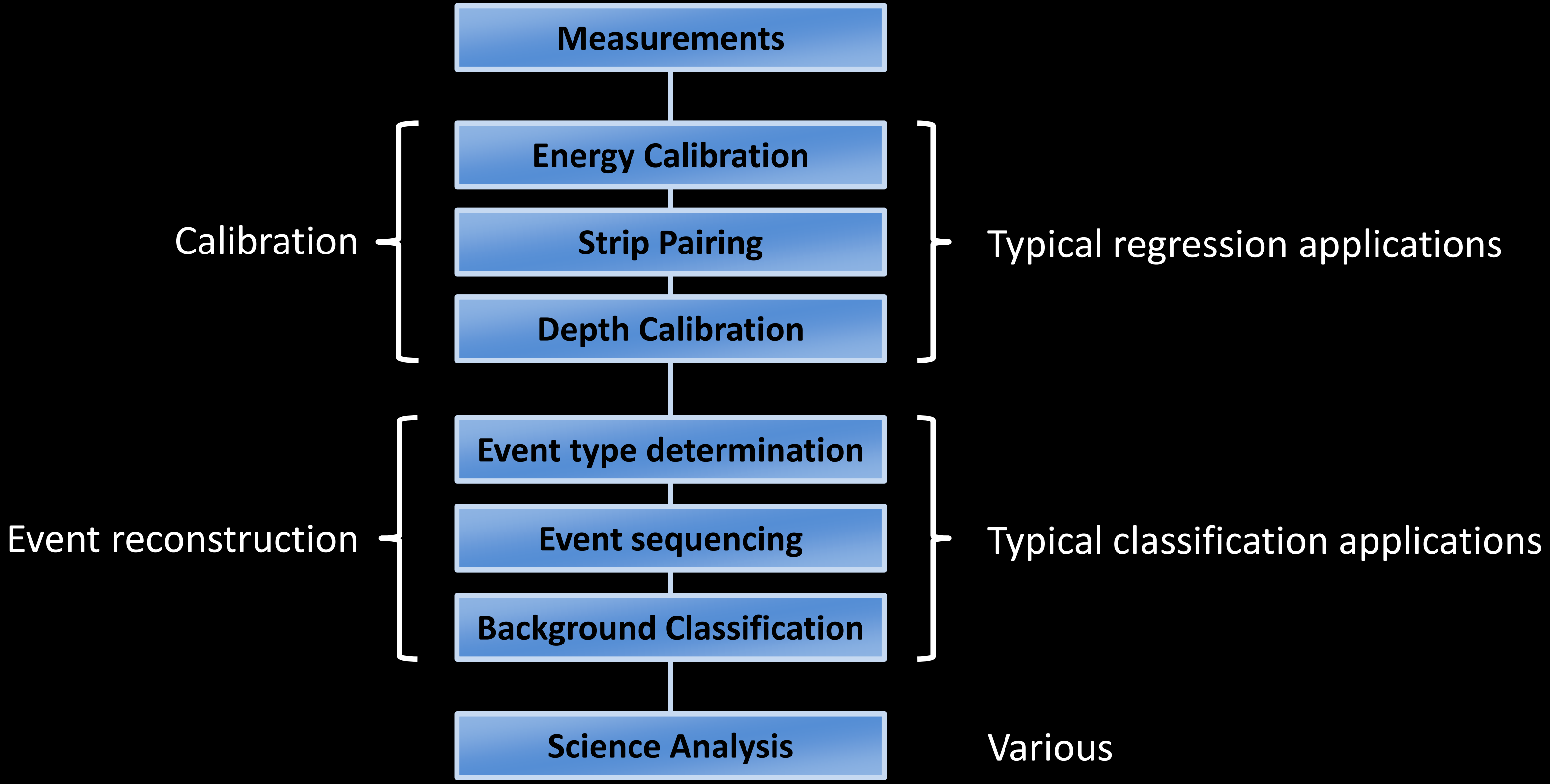


Image deconvolution

Enhancing the COSI Data Analysis Pipeline



The Software Libraries

MEGAlib
the Medium-Energy
Gamma-ray
Astronomy library

A. Zoglauer et al. 2006

ROOT
CERN's high-energy
physics data analysis
framework

R. Brun & F. Rademakers, 1997

TMVA
Toolkit for Multivariate Data
Analysis

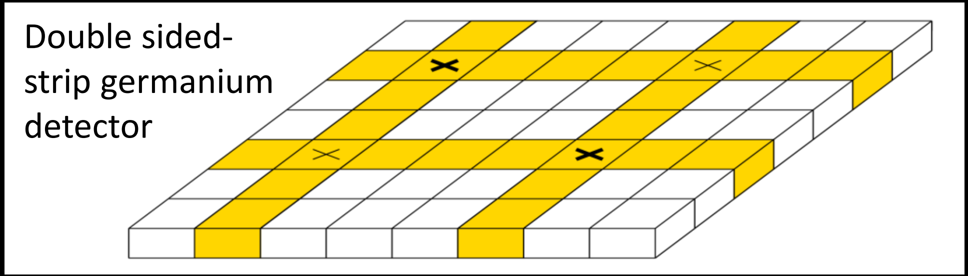
P. Speckmayer et al. 2010

Example: Strip Pairing

together with Devyn Donahue (2nd year data-science undergraduate)

Task:

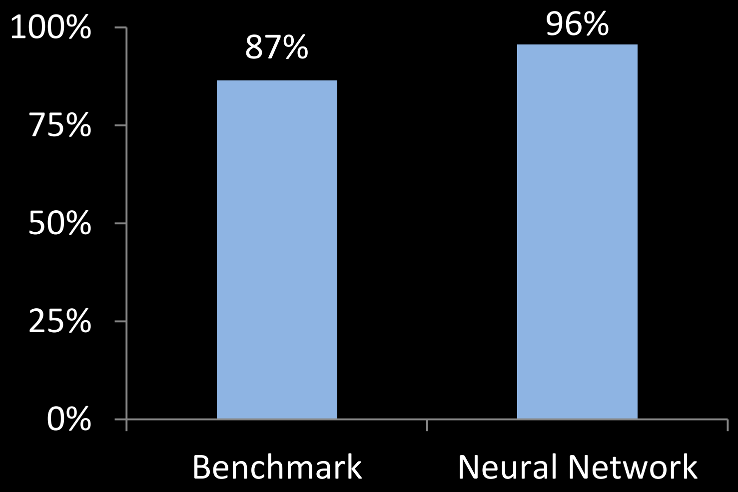
Find interaction locations in the (double-sided strip) detectors from the triggered strips



Yellow: Hit strips
x: Possible interaction locations
X: Real interaction locations

Results:

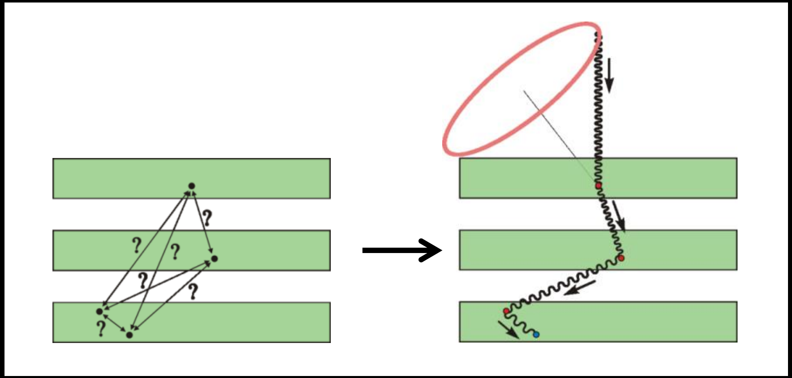
Benchmark (chi-square approach) vs. 4-layer fully connected neural network:



Example: Event Sequencing

Task:

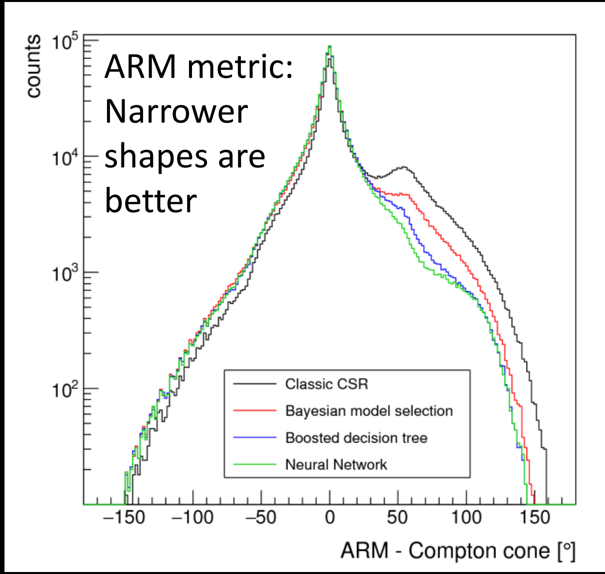
- Detectors just measure hits
- Find the path of the gamma ray in the detector using machine learning



Green: **Germanium detectors**
Dots: **Interaction locations**
Lines: **Possible paths**

Result:

Comparison of different machine learning approaches

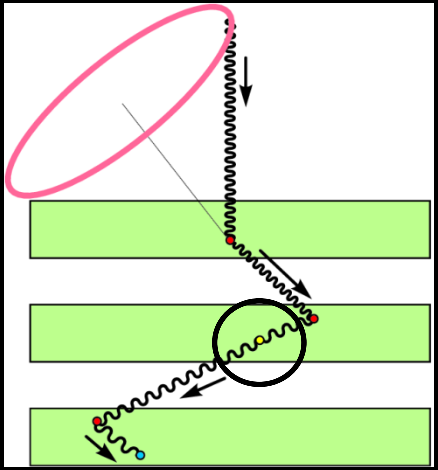


Neural networks perform best

Data Cleaning And Selection

“Selecting, cleaning & verifying the training & testing data can be the majority of the work.”

Data Cleaning: What to do with slightly non-conforming events during reconstruction?

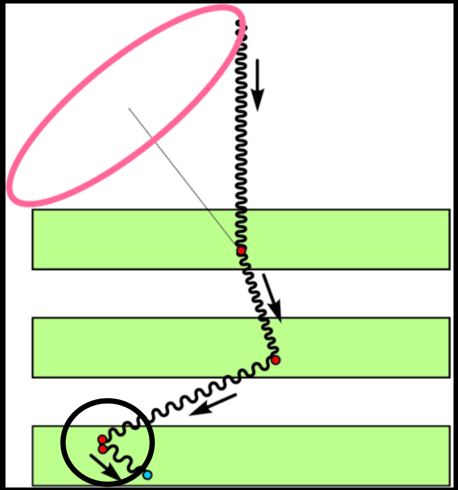


Rayleigh scattering within Compton sequence?

- ➢ OK since very small change!

Two Compton interactions in the same voxel?

- ➢ Only OK at end of sequence!



Data Cleaning And Selection

"Selecting, cleaning & verifying the training & testing data can be the majority of the work."

Always double check that you training data is correctly classified

Small errors can have large performance consequences

Eliminate Unknown Unknowns

"Try to make sure your AI cannot encounter something it is not trained for. If impossible, make sure it fails gracefully."

Example: Event Type Classification

Goal:

- Identify type: Singles, Compton events, Pair events, charged particle events, **others**

"Others" is catch-all for everything else that can happen in detector and what we are not interested in

Approach:

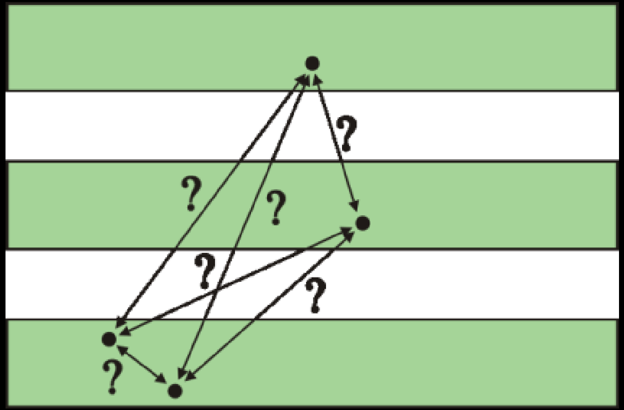
- Train with realistic simulations covering full energy range and all particle types
- Ensures that 99.999% of what we know is included, even in the "Others" category

Utilize all Available Information

"Avoid making your AI learn what you already know about your data. Provide this information as input."

Example: COSI event reconstruction

- All information is encoded in the measured positions and energies
- However, training with just position & energy does not yield good performance (with reasonable resources)
- Using all derivable, physical information (e.g. scatter angles, scatter & absorption probabilities), results in full performance
- Don't make you AI learn the physics, but provide the known physics as input!



Utilize all Available Information

"Avoid making your AI learn what you already know about your data."

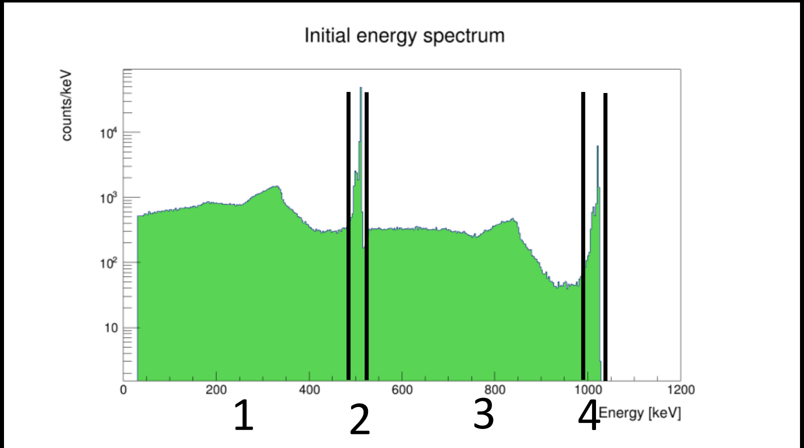
It is better to start with too many features.

You can always perform a feature ranking later, and eliminate the features which are not useful.

Changes in Behavior

"If your data shows significant change in behavior along one or more dimensions, consider to split the data along these dimensions and train individual AI's."

Example: Detection of 511-keV positron annihilation gamma rays



4 regions – 4 networks:

1. Mostly one incompletely absorbed gamma ray
2. Mostly one fully absorbed gamma ray:
3. Mostly one fully and one partially absorbed gamma ray
4. Only 2 fully absorbed gamma rays

Divide & Conquer

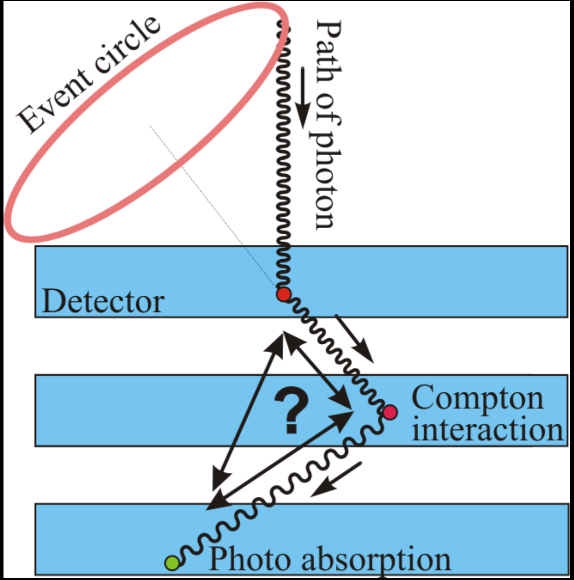
"Unless you have unlimited resources, it might be better to split a big question into several smaller ones."

Instead of asking:

- Originates this sequence from a completely contained, correctly reconstructed, Compton-scattered, astrophysical gamma ray and is not from any background source?

Ask this:

- Is it a Compton event (or pair, or charged particle, or …)?
- Is this the correct sequence for a Compton event?
- Is the Compton event completely contained?
- Does the event exhibit background signature A?
- Does the event exhibit background signature B, etc.



Blind Spots

"Always check that your AI has similar performance in all regions of the data space: Test it from all angles!"

Potential causes for performance variations:

- It simply doesn't work in this region of the data space
- Data too complex or changes too rapidly
- Network/Decision tree size is too small
- Not enough training data
- Wrong features selected

Trivial & Miscellaneous Lessons

- Make sure the data can answer your question
- Always test multiple machine learning approaches with the same data
- Don’t assume building identical neural networks from two different implementations/libraries will result in similar performance
- There is lot of trial and error involved in finding the best input data representation and the right network layout (number of nodes, hidden layers, etc.)

Thank You!

COSI US is supported through NASA Grant NNX14AC81G

COSI-X Phase-A study is supported by NASA

COSI imaging developments are supported through NASA grant NXX17AC84G

COSI machine learning is sponsored by BIDS/LLNL

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

Compton Spectrometer and Imager (COSI) @ Wanaka, New Zealand