# Incorporating Open Data and Aspects of Modern Physics Research in a Course on Scientific Computing


PURDUE UNIVERSITY NORTHWEST

**Dr. James Dolen**

**Assistant Professor of Physics at Purdue University Northwest**

PICUP

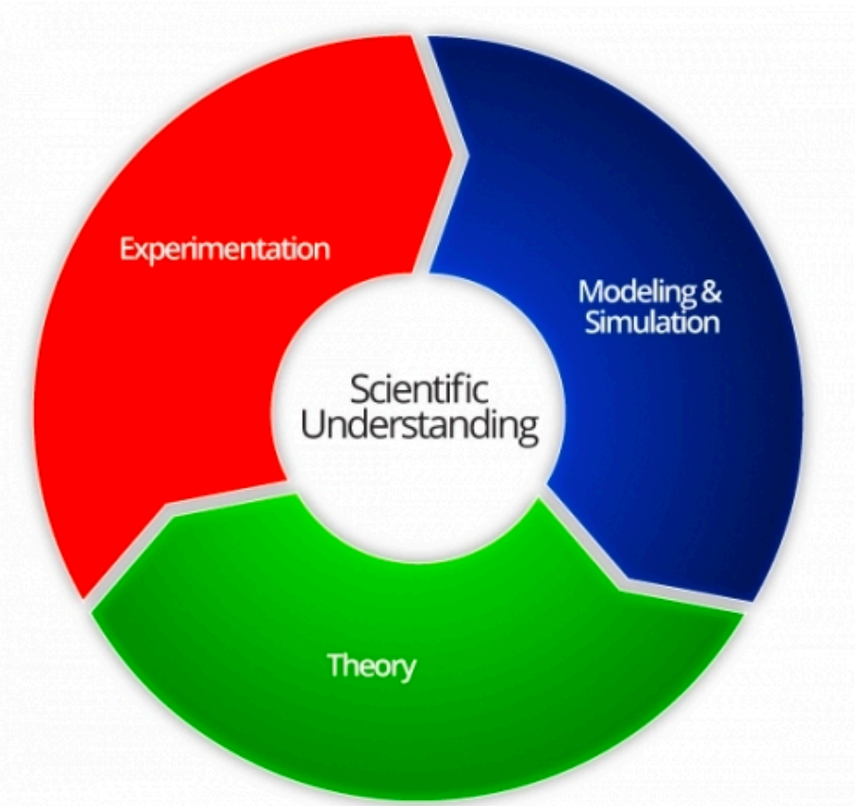

PICUP Capstone Conference - August 2021

# *Scientific Computing*

- Upper division physics course taught within the Chemistry and Physics Department at Purdue University Northwest (PNW)
  - Textbook: "Computational Physics" by Mark Newman
- Course advertised to both physics and non-physics majors
  - Students enrolled from Physics, Chemistry, Engineering, and from the "Modeling, Simulation, and Visualization" graduate program at PNW
- Computational techniques are interdisciplinary
  - Prepare students for senior research projects and graduate school
  - Applicable to careers in industry: Finance, engineering, data science

# *Course Aims*

- Demonstrate that computation is a fundamental component of all modern scientific research
  - Build a base in the foundational concepts of the field and introduce students to commonly used tools (Python, Jupyter Notebooks, etc.)
  - Exposure to more advanced methods




- Demonstrate that algorithms govern our modern world
  - We must understand the accuracy and uncertainty of our calculations!
  - Prepare students for research and careers

# *Scientific Computing: Example Topics*

- Numerical differentiation
- Numerical integration
- Ordinary Differential Equations
- Partial Differential Equations
- Fitting a function to data
- Interpolation
- Solving linear equations
- Solving non-linear equations
- Root finding
- Monte Carlo simulation
- Fourier analysis
- Computational fluid dynamics
- Heat transfer
- Structural analysis
- Electromagnetic field behavior
- Linux/Bash
- Common tools (python scientific libraries etc.)
- Classification
- Data visualization

# *Scientific Computing: Example Topics*

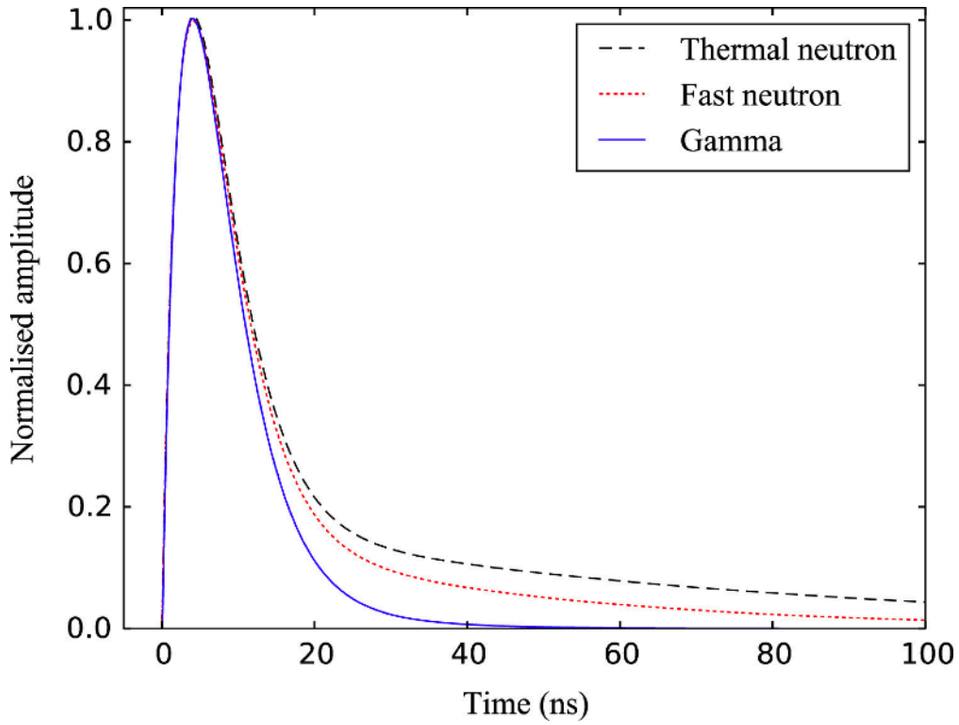
- Numerical differentiation
- Numerical integration
- Ordinary Differential Equations
- Partial Differential Equations
- Fitting a function to data
- Interpolation
- Solving linear equations
- Solving non-linear equations
- Root finding
- Monte Carlo simulation
- Fourier analysis
- Computational fluid dynamics
- Heat transfer
- Structural analysis
- Electromagnetic field

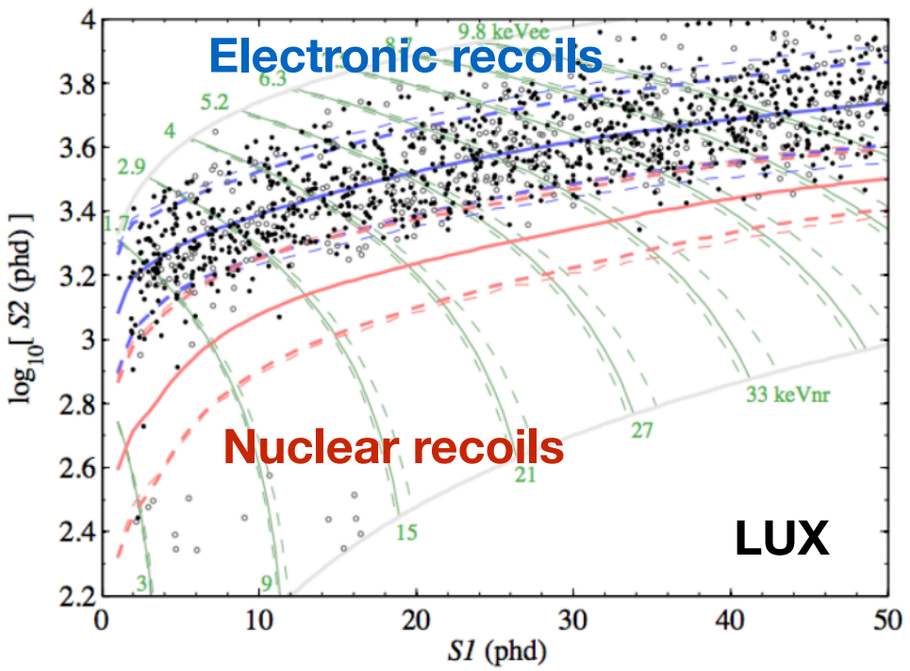
**Critical topics in research and industry**

- Common tools (python scientific libraries etc.)
- Classification
- Data visualization

# Examples: Classification

| Topic | Classification |
|---|---|
| Nuclear/Particle Physics | Pulse shape discrimination<br>Particle ID<br>Signal Event topology vs. Background |
| Cosmology | Supernova type classification (light curves) |
| Industry | Spam emails<br>Fraudulent transactions<br>Handwriting ID<br>Image recognition |
| Military Radar | Incoming airplane vs Radio Noise and other sources |
| Medicine | Tumor (observe using x-rays) ID vs benign tumors, organs, etc. |



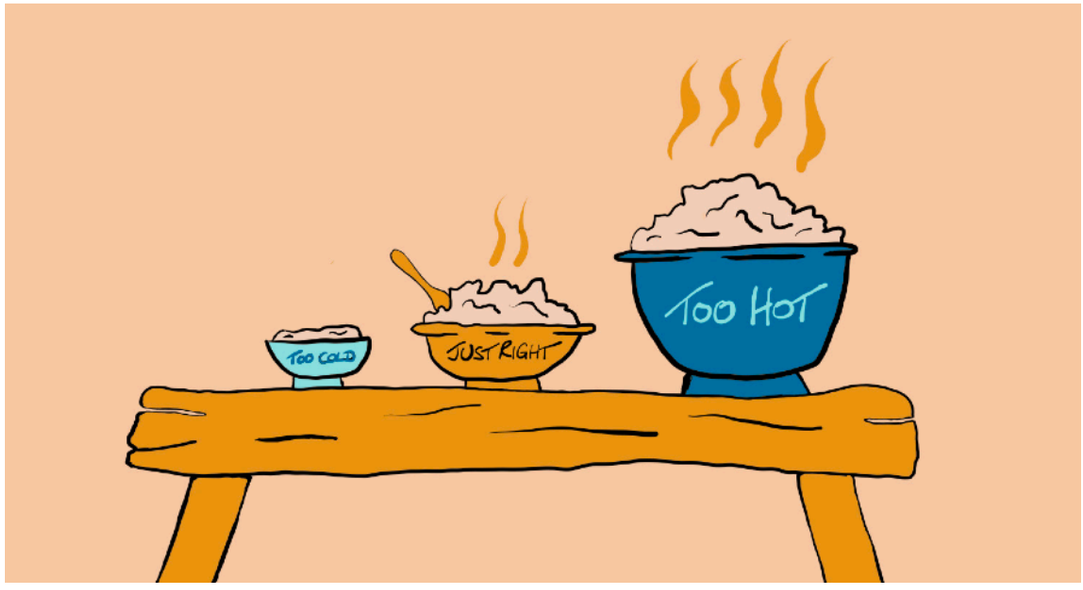



# *Open Data*

- Data made publicly available by scientific collaborations
- Excellent training datasets for students
- Examples:
  - NASA Open Data: https://data.nasa.gov/
  - CERN Open Data: http://opendata.cern.ch/
  - Machine Learning Datasets: https://archive.ics.uci.edu/ml/index.php
  - Microwave Background legacy data: https://lambda.gsfc.nasa.gov/
  - Liquid Argon TPC (neutrino) data: http://deeplearnphysics.org/DataChallenge/
  - NIST Data: https://www.nist.gov/pml/productsservices/physical-reference-data

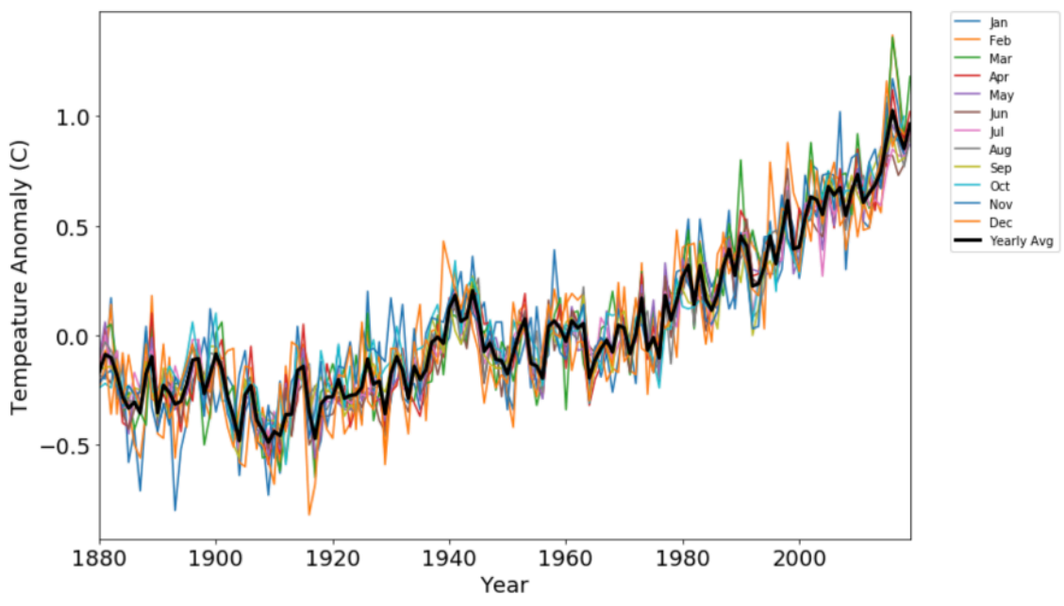
# *Open Data Processing*

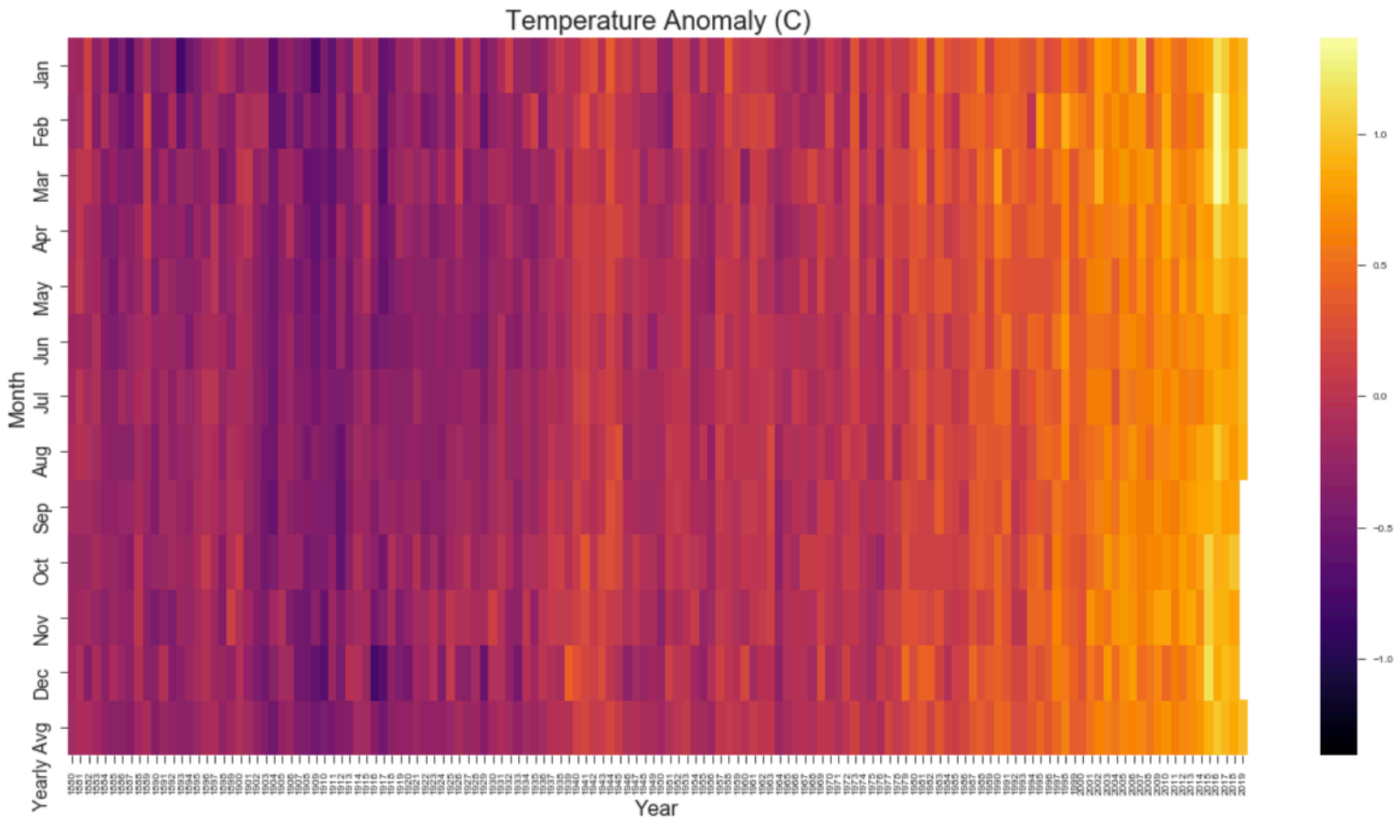
- Some open data is unprocessed raw data
  - Designed for experts (uncommon file formats, un-calibrated data). Useful in a student research setting but instructors must significantly pre-process this data for use in undergraduate course projects.
- Some open data is very processed
  - Suitable for high school students doing data analysis with Excel. Over-processed for undergraduates.
- Some open data is "just right"
  - Upper division undergraduates can access and analyze the data within a reasonable timeframe.



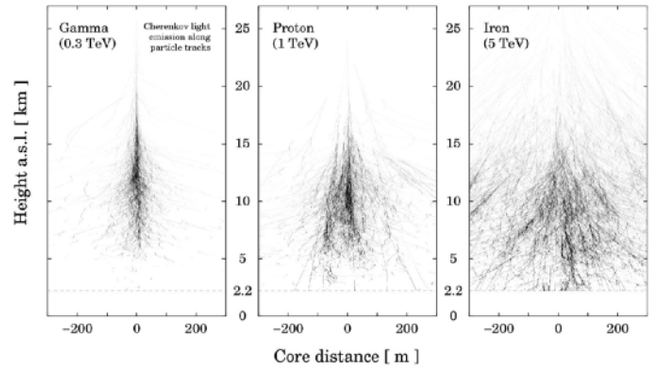
# *Example Projects*

- Planet Earth's Global Temperature
  - Open data from NASA’s Goddard Institute for Space Studies.
    - https://data.giss.nasa.gov/gistemp/
  - Mean monthly temperature 1880-present (combined Earth land-surface air and sea-surface water temperature).
  - Excellent opportunity to practice data visualization with Python, Matplotlib, Numpy, Pandas.

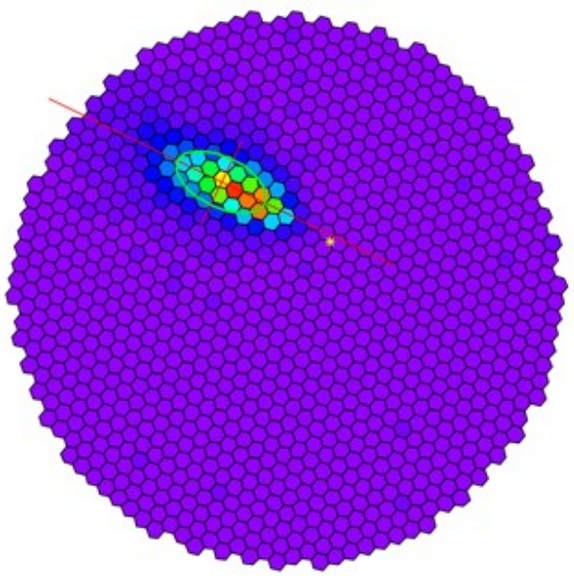


# *Example Projects*



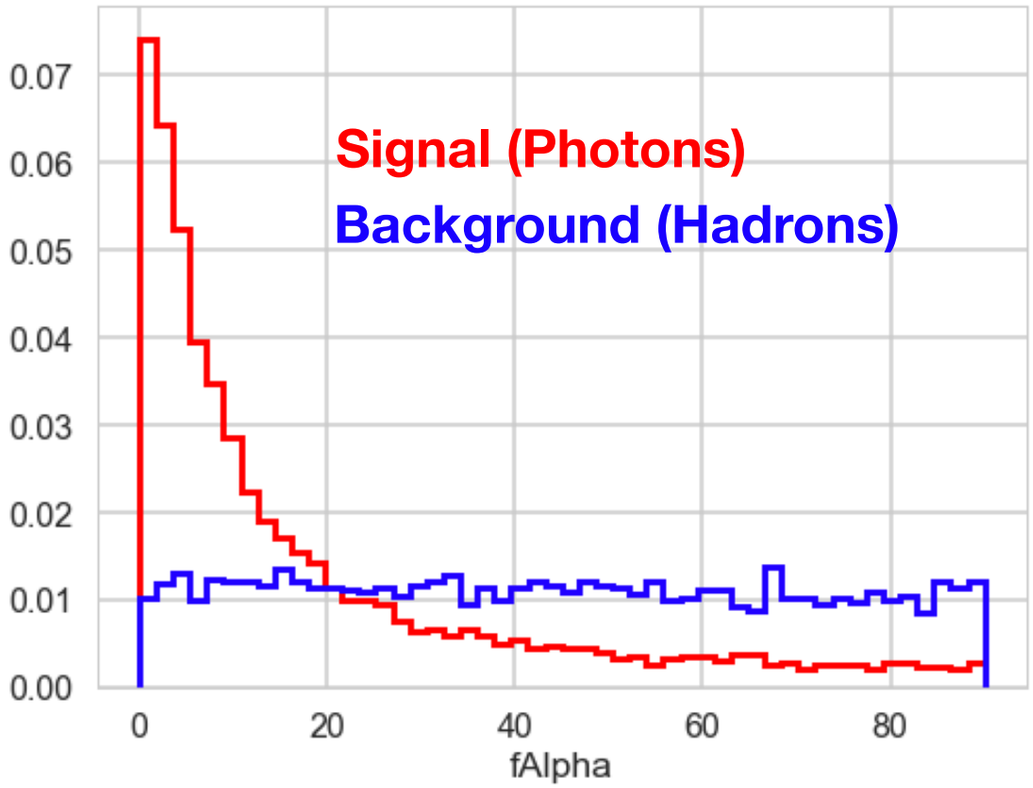

- Classification of cosmic ray showers
  - Open data from MAGIC (Major Atmospheric Gamma Imaging Cherenkov Telescopes).
    - https://archive.ics.uci.edu/ml/datasets/magic+gamma+telescope
  - Dataset contains simulated cosmic ray events reconstructed by MAGIC.
    - Ten shower shape variables
    - Truth information (γ or hadron)
  - Use to teach aspects of **classification** through project-based learning.

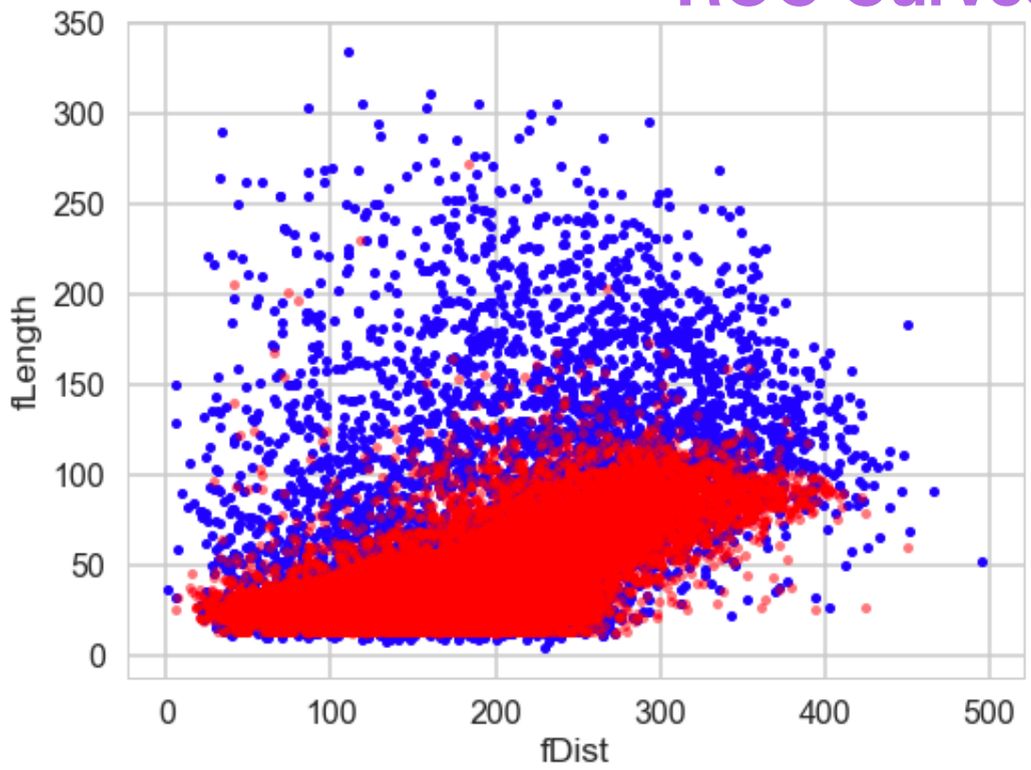


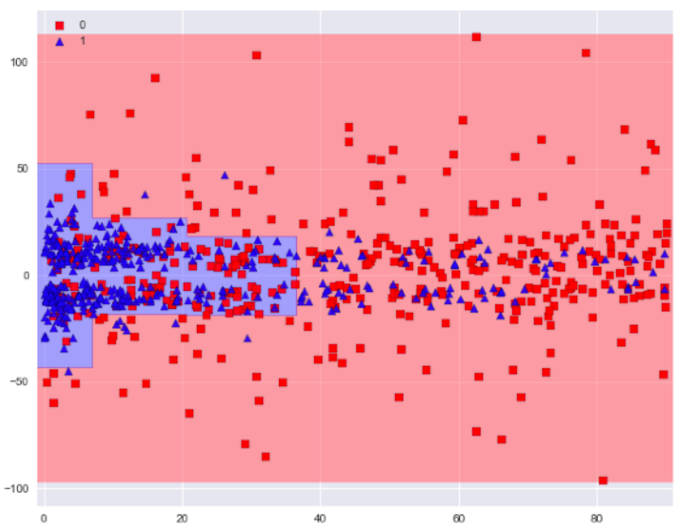
# Example Projects

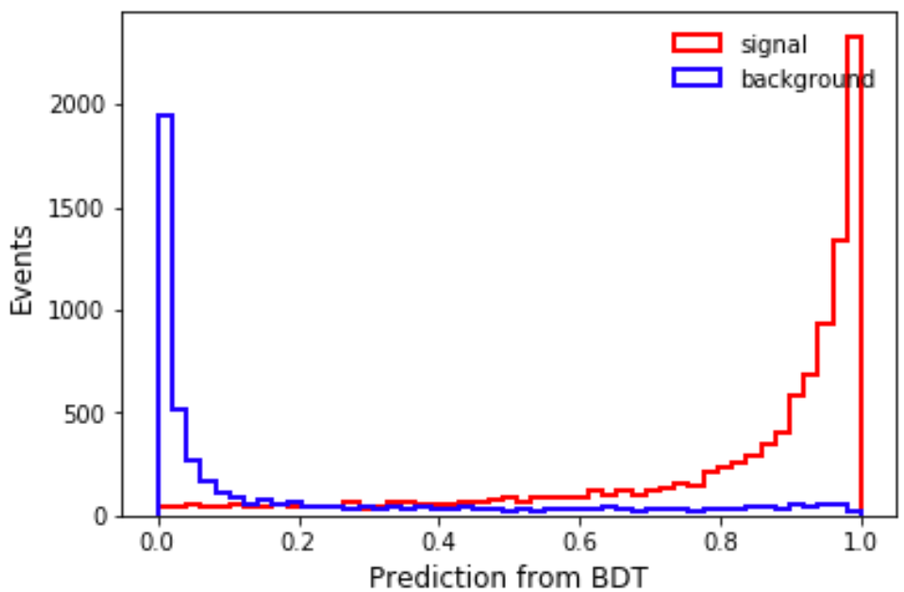
**Concepts:**
**Histograms**
**2D data visualization**
**Boosted decision trees**
**ROC Curves**

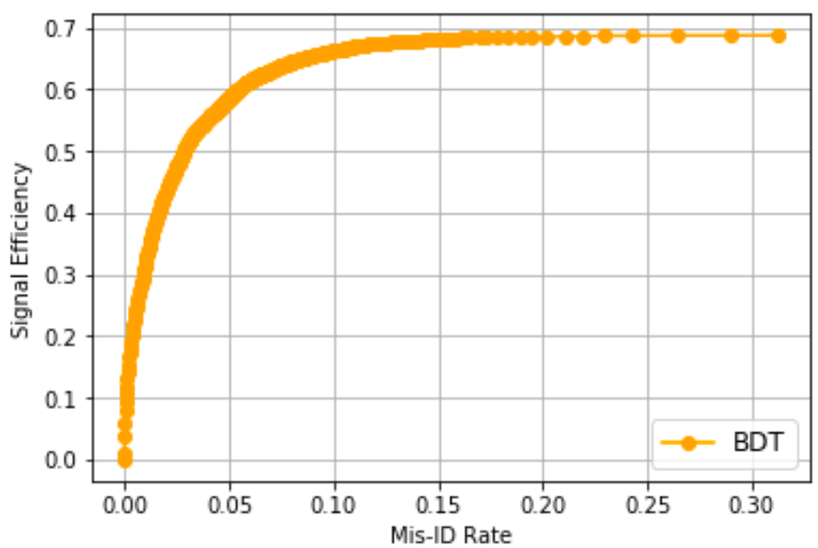
- Classification of cosmic ray showers





# *Conclusions*

- Numerical problem solving is common in most scientific computing courses and is critical knowledge
- It is useful to also include data analysis methods in such courses
  - Data analysis techniques used in research and in industry are increasingly advanced
  - **Open data** is an enormously useful tool in teaching foundational data analysis
- One may have to find or create pre-processed data in order to make a given open dataset useful within a one semester course.