IRIS-HEP Fellowship Proposal: Developing a Pythonic tool for smoothing histograms
Applicant: Roman Petrov
Mentors: Mohamed Aly, Alexander Held, Matthew Feickert

Project Details:

Histograms are fundamental objects in statistical modeling frameworks used in Large Hadron Collider analyses. They serve as the primary representation of both experimental data and theoretical predictions, with systematic uncertainties often modeled as variations on these histograms. Smoothing refers to procedures used to reduce statistical noise in histograms, enhancing the stability of statistical models and ultimately leading to more reliable results in physics analyses.

While histogram smoothing is essential for high-energy physics (HEP) analyses, existing tools like ATLAS SystematicSmoothingTool[1] lack accessible interfaces, comprehensive documentation, and integration into the Scikit-HEP ecosystem. Given that data analysis in HEP is uniquely reliant on histograms compared to other scientific fields, a user-friendly and performant smoothing tool is an integral missing component. The challenge lies in determining the "right way" to smooth a histogram. This is particularly difficult because the optimal smoothing approach depends on the underlying physical process and the specific characteristics of the histogram. Without sufficient statistics, it's hard to know which smoothing method will best preserve the true signal while reducing noise.

This project aims to address these gaps by developing a dedicated smoothing tool in Python based on existing algorithms, enabling physicists to apply smoothing methods more efficiently and consistently in their workflows. The tool will implement specialized procedures that have proven effective specifically for HEP applications, such as those implemented in the ATLAS SystematicSmoothingTool. The core functionality will be to take one histogram as input, apply a selected smoothing algorithm, and return a new histogram containing the smoothed result.

The project will also develop methods to evaluate algorithm accuracy by drawing samples from known distributions, histogramming them, applying smoothing, and then comparing the results to the true distributions. This will allow for quantitative assessment of how well different algorithms can recreate the underlying distributions, providing guidance to users on algorithm selection.

While the primary focus will be on algorithm implementation rather than visualization, the tool will be designed to integrate smoothly with existing visualization tools like mplhep or Cabinetry[2], which is used to steer HEP statistical analyses, providing a cohesive workflow for analysts. As a stretch goal, if time permits, the project could explore the development of novel smoothing algorithms for

scenarios where existing methods prove inadequate, particularly for multi-dimensional histograms. Current smoothing algorithms used in ATLAS don't cover 2D histograms.

Timeline:

- Weeks 1-2: Research existing histogram smoothing algorithms both in general statistics and those specific to HEP applications. Set up GitHub repository and development environment. Begin literature review of smoothing methods.
- Weeks 3-4: Implement basic infrastructure for the package, including data structures, interfaces, and testing framework. Implement the first basic smoothing algorithm.
- Weeks 5-6: Develop initial validation framework to compare algorithm accuracy using known distributions. Once this infrastructure is in place, begin implementing additional smoothing algorithms with immediate comparative evaluation.
- Weeks 7-8: Complete implementation of core algorithms. Develop a comprehensive validation suite to quantitatively assess algorithm performance across different scenarios.
- Weeks 9-10: Focus on integration with the Scikit-HEP ecosystem, ensuring compatibility with existing tools and frameworks. Implement proper error handling and edge case management.
- Weeks 11-12: Complete documentation including API references, usage examples, and tutorials. Finalize package structure for distribution via PyPI.

Deliverables:

1. A Python package implementing multiple histogram smoothing algorithms
2. Integration with the Scikit-HEP ecosystem
3. Complete documentation including API references and usage examples
4. Publication of the package on PyPI for easy installation
5. Comprehensive validation framework for assessing algorithm performance

References:

1. ATLAS SystematicSmoothingTool: https://gitlab.cern.ch/atlas/statistics/SystematicSmoothingTool
2. Cabinetry: https://cabinetry.readthedocs.io/