# Modern Analysis Techniques for Large Data Sets
# Phys 576, Spring 2021
# Instructor: Miguel Morales

While analyzing large datasets is nothing new for physicists, in the last few years there have been major advancements in the tools and techniques available. Team taught by Miguel Morales (Physics) and Bryna Hazelton (eScience), the goal of this class is to introduce students to current techniques and best practices in the statistically rigorous analysis of large data sets. The class is organized around four themes: practical statistics, advanced data visualization, building collaborative analysis code, and advanced data analysis practices (see below for details).

The class is open to graduate students, postdocs, research groups, and seniors with permission. Evaluation will be based on homework and projects, and students are encouraged to use their own data for the projects to enhance their current research.

Prof. Miguel Morales has experience in particle physics, astrophysics, and cosmology data analysis, and is considered an international expert in the analysis of 21 cm cosmology data. Senior Research Scientist Bryna Hazelton has worked on everything from cosmology to botany to homelessness as part of the eScience Institute. She is a co-author of the open source and peer reviewed pyuvdata software package, and has developed the reference analysis pipeline for analyzing Epoch of Reionization radio cosmology data.

Topic list (not in syllabus order):

Advanced practical statistics
Foundations
non-Gaussian and non-analytic statistics
Maximum likelihood
Feldman-Cousins and extensions
Issues with large data sets and trials
Practical considerations
Determining background distributions from data
Systematic errors
End-to-end error propagation (including non-Gaussian extensions)
Parameters, covariance, Fischer Matrices, non-linear effects, and the art of parametrization
Asking statistically valid questions
How to mathematically formulate your question(s)
Case studies of mistakes in the literature
Jackknife and null tests
Data visualization
Features of high quality visualizations
Data density
Classes of plots, and their pros and cons
Meta information and drillability
Scaling & color
Animations and movies
Developing a consistent visual language
Accessibility considerations (e.g. colorblind, pattern recognition, etc.)
Visualizations for data exploration and hunting systematics
Turning statistical questions into plots

Developing plots for data rampages
Visualizations for instrument and data monitoring
Sparklines, comparisons with nominal performance
Notebooks and dashboards
Visualizations for presentations and publications
Developing plots as a teaching tool
Specific concerns for presentations and publication plots
Case studies of valuable visualization techniques
Tools and best practices for building collaborative analysis pipelines
Using GitHub to your advantage
Branching and merging for collaborative data analysis
Unit testing
Git hashes, metadata, and analysis provenance
Collaborative development of analysis tools
Issue tracking
Issue assignment and managing releases
Pull requests
Shared libraries for enhanced communication
Publishing peer reviewed code
Advanced data analysis practices
Making sure your analysis is right
Analysis level unit tests
Designing a thicket of tests
Tracing your analysis as it evolves
The golden master development pattern
Analysis jackknifes, and testing below the thermal noise
Tiered testing with data as part of the development cycle
Improving your analysis (hunting systematics, biases, calibration errors, and subtle analysis mistakes)
Turning questions into tests
Newtons method of isolating issues
Interrogating your data for systematics and biases (question driven data rampages)

FAQs:

**What constitutes a large data set?** The short answer is if it is large for you, it counts. What is big data varies wildly by field, but the statistical and analysis issues are effectively the same whether you have 1,000 data points or $10^{15}$.

**What format will the projects take?** If you have your own data, the projects will be applying the techniques we learn to your data. And for the final project you will propose what you plan to do, so it should be directly applicable to your research.

**Is prior knowledge of any particular coding language expected?** We are carefully language agnostic. Many examples will be in python, but we frequently use Matlab, IDL, C, and have experience in a variety of other languages. Do the projects and homework in whatever you are comfortable in.

**Does this count as a graduate distribution requirement?** Yes, it should count regardless of your area of study.