

Introduction to Data Science (for physics)

Outline of the course (labs):

1. Statistics and Data Analysis
2. Multivariate Techniques and Machine Learning
3. Physics Modeling, Simulation and Monte Carlo Methods
4. Regression, Classification, Clustering and Retrieval

First three parts will focus on applications in physics (mostly in High Energy Physics)

The last part will discuss more typical „Data Science” problems and solutions.

LABS: how we will operate it

- **Normal times: run in person:**
 - Time for you to write your code and (for me) to discuss with each student her/his progress with assignments.
- **COVID-19 times: run on-line:**
 - We will go through content of assignments, present (mostly you) results of analyses and observations, have short oral presentations (please be active)
 - Occasion to share with everybody problems, exchange snippets of the code or interesting observations.

Getting your ETCs for labs

This is not a course of programming, but you will be expected to write programs.

- Basic choices are: C++/Root or Python + Anaconda libraries.
- You can use also R or other Data Science specific programming language/library

I will not be teaching you programming or helping to debug your code, you are on your own ...

For labs you will be graded with:

- completed assignments:
- personalised project
- short topical presentations

Graded will be not (necessarily) quality of the code, but maturity of how you analyse and interpret the data.

To pass the course you need to collect at least **60 scores.**

Assignments, Projects, Short presentations

PEGAZ system:

This system we will use to collect your assignments/projects/short presentations

- I will be sending you back comments
- **You will see your grades there**

Please don't use email to send me your scripts!

MSTeams:

This system we will use for on-line classes

- you can use it also for communication among yourself, eg. setting up chats/meetings within a team.
- for communication with me preferably use emails

Introduction to Data Science (for physicists)

Wydział Fizyki, Astronomii i Informatyki Stosowanej,
Uniwersytet Jagielloński w Krakowie

Rok akademicki 2020/2021

Konsultacje: wtorek, godz. 15:00 - 16:00; pokój G-0-10.

Lectures

Recommended books/articles for reading:

- => [G. Cowan, "Statistical Data Analysis"](#)
- => [F. James, "Statistical Methods in Experimental Physics"](#)
- => [J. Narsky, F. Porter, "Statistical Analysis Techniques in Particle Physics"](#)
- => [J.A.Rice, "Mathematical Statistics and Data Analysis"](#)
- => [K. Cranmer, "Practical statistics for LHC"](#)

Date	Lecture slides	Additional material
	Statistics and Data Analysis	Statistical methods for LHC (advanced)
13.10.2020	Introduction , StatAnal-lecture-1	
20.10.2020	StatAnal-lecture-2 , StatAnal-lecture-3 ,	
27.10.2020	StatAnal-lecture-4 ,	
3.11.2020	LHCStatAnal-lecture-1 , LHCStatAnal-lecture-2	
10.11.2020	LHCStatAnal-lecture-3	
17.11.2020	LHCStatAnal-lecture-4	
	Multivariate Techniques and Machine Learning	
24.11.2020	MVandML-lecture-1	
1.12.2020	MVandML-lecture-2	
8.12.2020	MVandML-lecture-3	
	Physics Modeling, Simulation and Monte Carlo Methods	
15.12.2020	PhysModelAndMC-lecture	
	Regression, Classification, Clustering and Retrieval	
12.01.2021	DataScience-lecture-1	
19.01.2021	DataScience-lecture-2	
26.01.2021	DataScience-lecture-3	

Assignments:

Date	Topic	Suggested for reading	Root/C++ or use PyRoot	Datasets/Tutorials	Python + Anaconda	Datasets/Tutorials
13.10.2020	Getting organised with the framework Data exploration		Select few examples from this link: histograms eg.: h1draw fibonacci filrandom ratioplot1 or from PyRoot examples from this link: PyRoot	HowToStart	assignment-0-python assignment-0-numpy assignment-0-numpy-matplotlib assignment-0-pandas	HowToStart DS-cheatsheet_numpy.pdf DS_cheatsheet_matplotlib.pdf DS_cheatsheet_jupyter_notebook.pdf DS_cheatsheet_pandas.pdf numpy tutorial matrix algebra tutorial
	Statistics and Data Analysis				scripts in K. Cranmer, Statistics and Data Science	PYHF: python based fitting/limit-setting/interval estimation
20.10.2020	StatAnal labs-lecture-1.txt StatAnal labs-lecture-2.txt	Efficiency uncertainties				
27.10.2020	StatAnal labs-lecture-3.txt StatAnal labs-lecture-4.txt					
3.11.2020						
10.11.2020						
17.11.2020			LHCStatAnal-labs-4-Root			
	Multivariate techniques and Machine Learning					
24.11.2020						
1.12.2020						
8.12.2020						
	Physics Modeling, Simulation and Monte Carlo Methods					

Assignments:

Date	Topic	Suggested for reading	Root/C++ or use PyRoot	Datasets/Tutorials	Python + Anaconda	Datasets/Tutorials
13.10.2020	Getting organised with the framework Data exploration		Select few examples from this link: histograms eg.: h1draw fibonacci filrandom ratioplot1 or from PyRoot examples from this link: PyRoot	HowToStart	assignment-0-python assignment-0-numpy assignment-0-numpy-matplotlib assignment-0-pandas	HowToStart DS-cheatsheet_numpy.pdf DS_cheatsheet_matplotlib.pdf DS_cheatsheet_jupyter_notebook.pdf DS_cheatsheet_pandas.pdf numpy tutorial matrix algebra tutorial
	Statistics and Data Analysis				scripts in K. Cranmer, Statistics and Data Science	PYHF: python based fitting/limit-setting/interval estimation
20.10.2020	StatAnal labs-lecture-1.txt StatAnal labs-lecture-2.txt	Efficiency uncertainties				
27.10.2020	StatAnal labs-lecture-3.txt StatAnal labs-lecture-4.txt					
3.11.2020						
10.11.2020						
17.11.2020			LHCStatAnal-labs-4-Root			
	Multivariate techniques and Machine Learning					
24.11.2020						
1.12.2020						
8.12.2020						
	Physics Modeling, Simulation and Monte Carlo Methods					
15.12.2020						
	Regression, Classification, Clustering and Retrieval					
12.01.2021						
19.01.2021						
26.01.2021						

Statistical Analysis in HEP Physics

N. Beger, Foundation of Statistics, Lectures at CERN Summer School 2019

[link1](#), [link2](#), [link3](#)

Statistics and Data Science

K. Cranmer, Course at NYU Physics, Fall 2020, [link](#)

Machine learning applications in HEP physics:

B. Nachman,

["Advanced Machine Learning for Classification, Regression, and Generation in Jet Physics"](#)

M. Stoye,

["ML applications in CMS"](#)

ML techniques in HEP, Workshop, Berkeley Laboratory, 11 - 13 December 2018

<https://indico.physics.lbl.gov/indico/event/546/>

A. Castaneda, LHCP conference, Puebla, Mexico, 2019

[ML and Big data tools at HEP](#)

I will keep adding links as I progress in preparing lectures and assignments.

Data exploration

 [Assignment-0: due 25.10.2020, final deadline 1.11.2020](#)

Statistics and Data Analysis

 [StatAnal_labs-1: due 25.10.2020, final deadline 1.11.2020](#)

 [StatAnal_labs-2: due 1.11.2020, final deadline 8.11.2020](#)

 [StatAnal_labs-3: due 8.11.2020, final deadline 15.11.2020](#)

 [StatAnal_labs-4: due 15.11.2020, final deadline 22.11.2020](#)

 [StatAnal project: due 22.11.2020, final deadline 29.11.2020](#)


Multivariate Techniques and Machine Learning

 [MTandML-labs-1: due 29.11.2020, final deadline 6.12.2020](#)

 [MTandML-labs-2: due 6.12.2020, final deadline 13.12.2020](#)

 [MTandML project: due 10.01.2021, final deadline 17.01.2021](#)

Physics Modeling, Simulation and Monte Carlo Methods

 [PhysModel-and-MC-labs: due 17.01.2021, final deadline 24.01.2021](#)

Regression, Classification, Clustering and Retrieval

 [Data Science project: due 24.01.2021, final deadline 31.01.2021](#)