CERN openlab

Quantitative Workflow characterization and modeling

Supervisor: Andrea Sciaba & Markus Schulz

Alexia Topalidou

National and Kapodistrian
UNIVERSITY OF ATHENS

Project Description

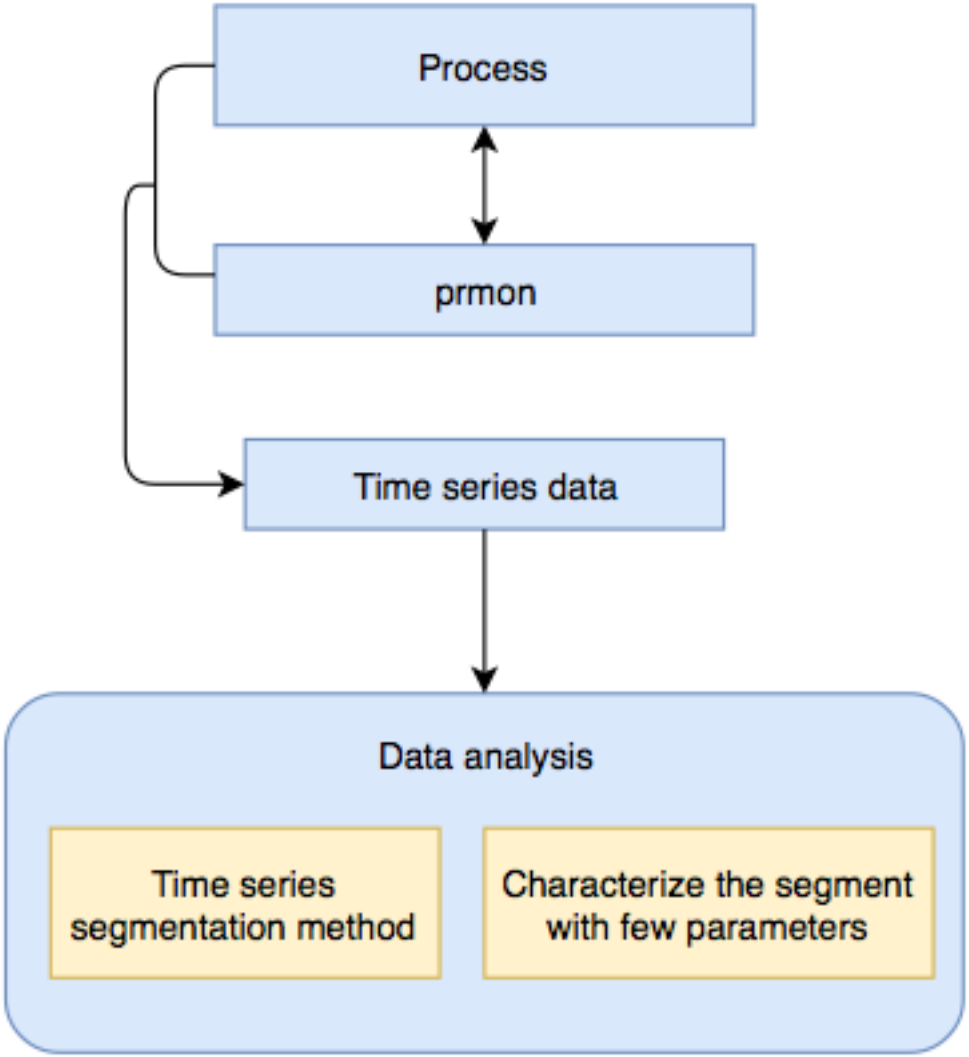
This project is about building a predictive model that can be used to quantitatively understand the impact of architecture, technology and data analysis strategy decisions for HL-LHC (High Luminosity Large Hadron Collider which is an upgrade to the LHC).

This is critical, because the computing demands at the HL-LHC will massively exceed the available resources unless large efficiency gains can be made.

How can we understand the computing demands?

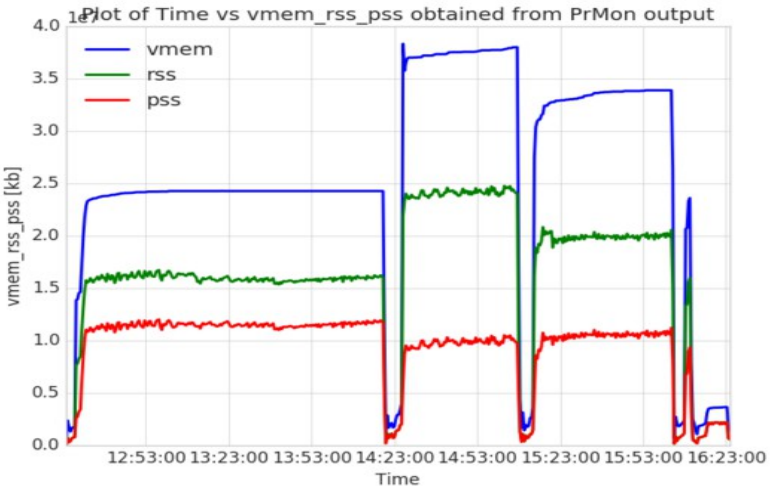
The first step to understand the behavior of the different workloads of the experiments is to measure, in a controlled environment, critical resource usage parameters and their dependency on running conditions.

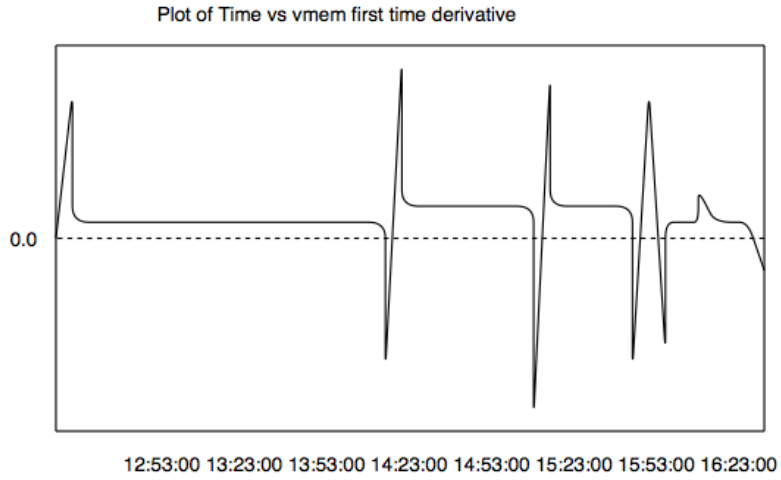
Project Description



Project Description

- Metrics may drastically change during an application's lifetime
- We first identify the changes in the time series data of the metric.
- Based on these changes we split the metric data into unique segments.
- Each of the segments will then be parametrized (e.g constant, linear increase, logarithmic increase etc.)





How do we measure the resource consumption of a process?

PRocess MONitor (prmon)

- Small stand alone program that can monitor the resource consumption of a process and its children.
- Helps to understand the resource requirements.

How it works

- Each process has a unique PID (Process IDentification number).
- Prmon extracts CPU, memory and IO usage as a time series for the specified process.
- The results are then processed using an analysis tool chain.

What metrics do we extract using prmon to monitor the resource consumption of a process?

	Time	wtime	stime	utime	pss	rss	swap	vmem	rchar	read_bytes	wchar	write_bytes	rx_bytes	rx_packets	tx_bytes	tx_packets
0	0	1	0	0	8355	10604	0	86740	1089812	5120000	187	4096	269006	236	15938	160
1	31	32	3	19	686313	772100	0	2245216	28163857	781160448	4467033	3231744	378991	421	66987	281
2	62	63	5	41	819294	827944	0	1865076	45590729	876908544	7304611	3526656	601644	749	114774	469
3	93	94	9	60	688570	697012	0	1783072	71873304	1456459776	10925151	4739072	935657	1095	176367	628
4	124	125	9	62	751303	759848	0	1863820	73205149	1478549504	11171864	4931584	2291285	3356	634949	2426
5	155	156	10	72	1028461	1037020	0	2248820	148087397	1785049088	12281044	5783552	6005457	7306	1289368	5019
6	186	187	10	79	1023721	1032280	0	2287576	151378506	1815986176	12475443	5955584	9644590	11048	1792282	7453
7	218	218	22	191	6849715	16767088	0	27269048	446807475	2292613120	17593217	11128832	15679938	17171	2982985	12236
8	248	248	32	398	9260842	18541320	0	29067584	804513375	2447769600	21542712	15249408	15694416	17285	3020147	12319
9	280	280	40	647	9365656	18636840	0	29133332	2050731597	2752233472	37359079	31178752	15713899	17418	3081112	12405

What metrics do we extract using prmon to monitor the resource consumption of a process?

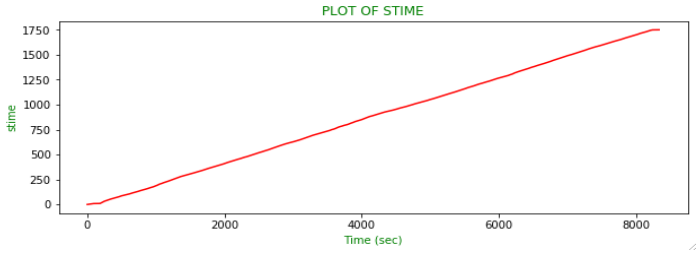
Metric types:

Cumulative metrics:
- Current value + previous value, grows over time e.g. stime etc.

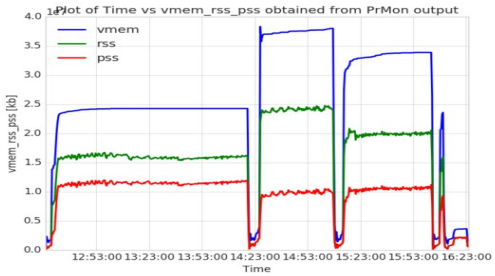
Load metrics:
- Current value ,they increase or decrease e.g. vmem, RSS etc.

Rate metrics:
- metrics that represent the time derivative of another metric, metric/sec.

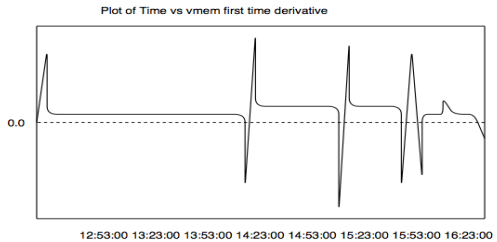
Cumulative metric



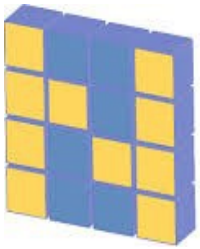
Load metric



Rate metrics



How do we analyze these metrics?



NumPy

Pandas

matplotlib

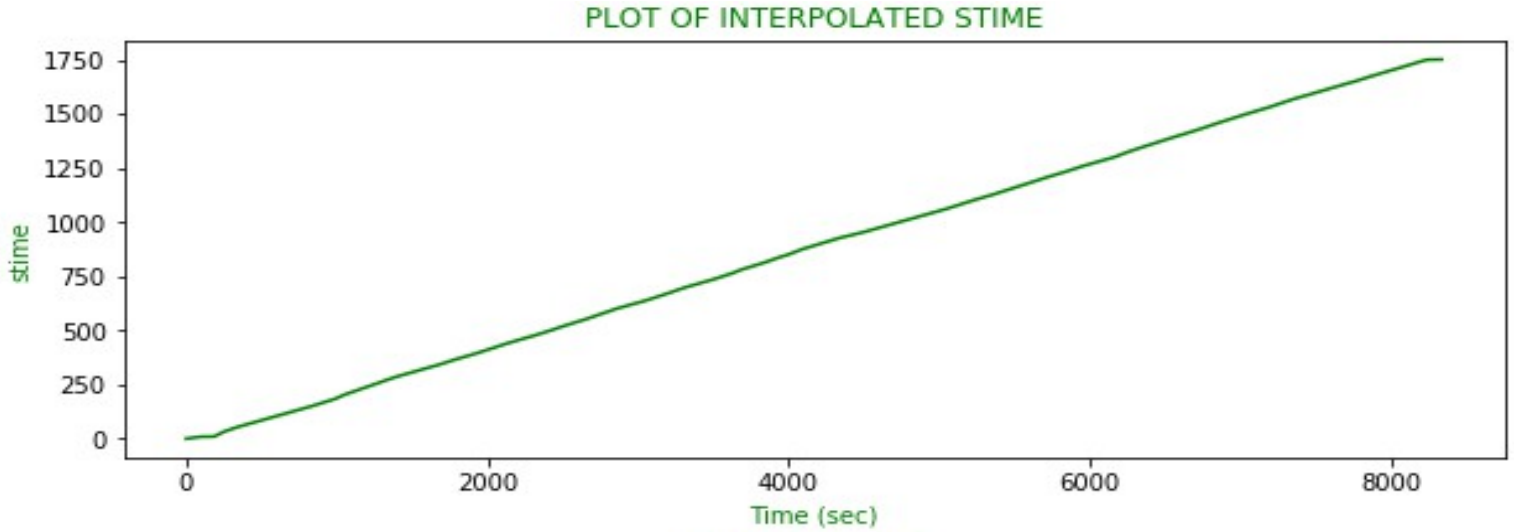
How do we analyze these metrics?

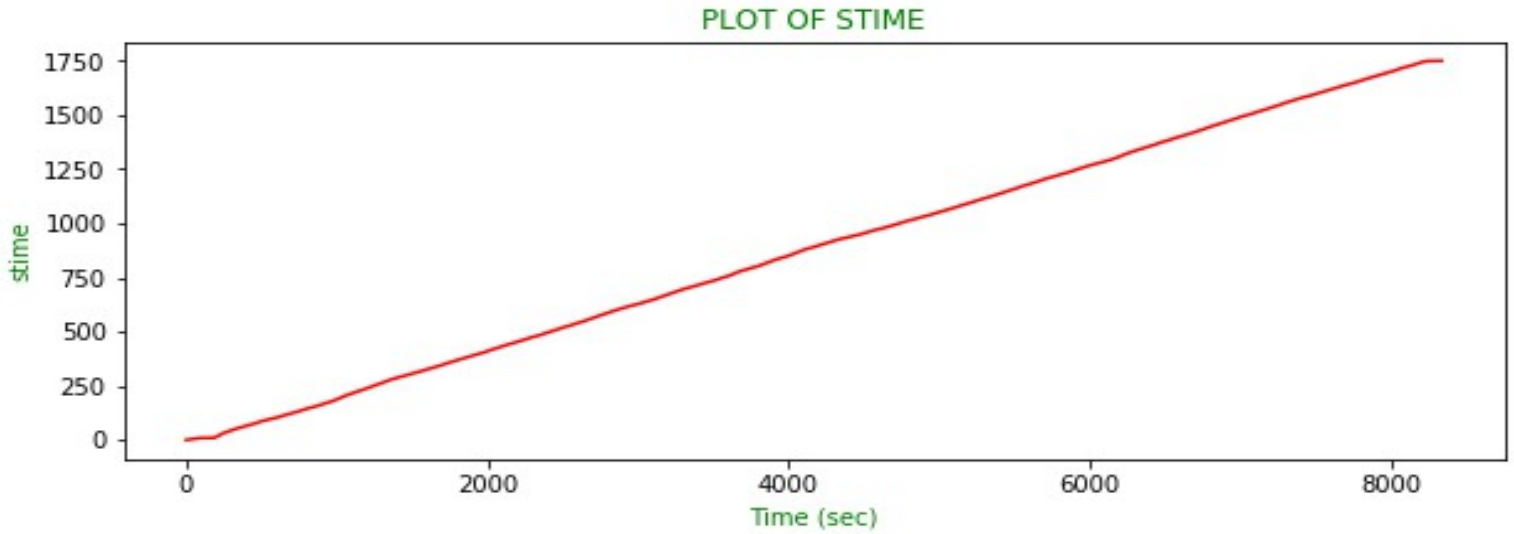
SWAN (Service for Web based Analysis): is a platform to perform interactive data analysis in the cloud.

With SWAN you can:

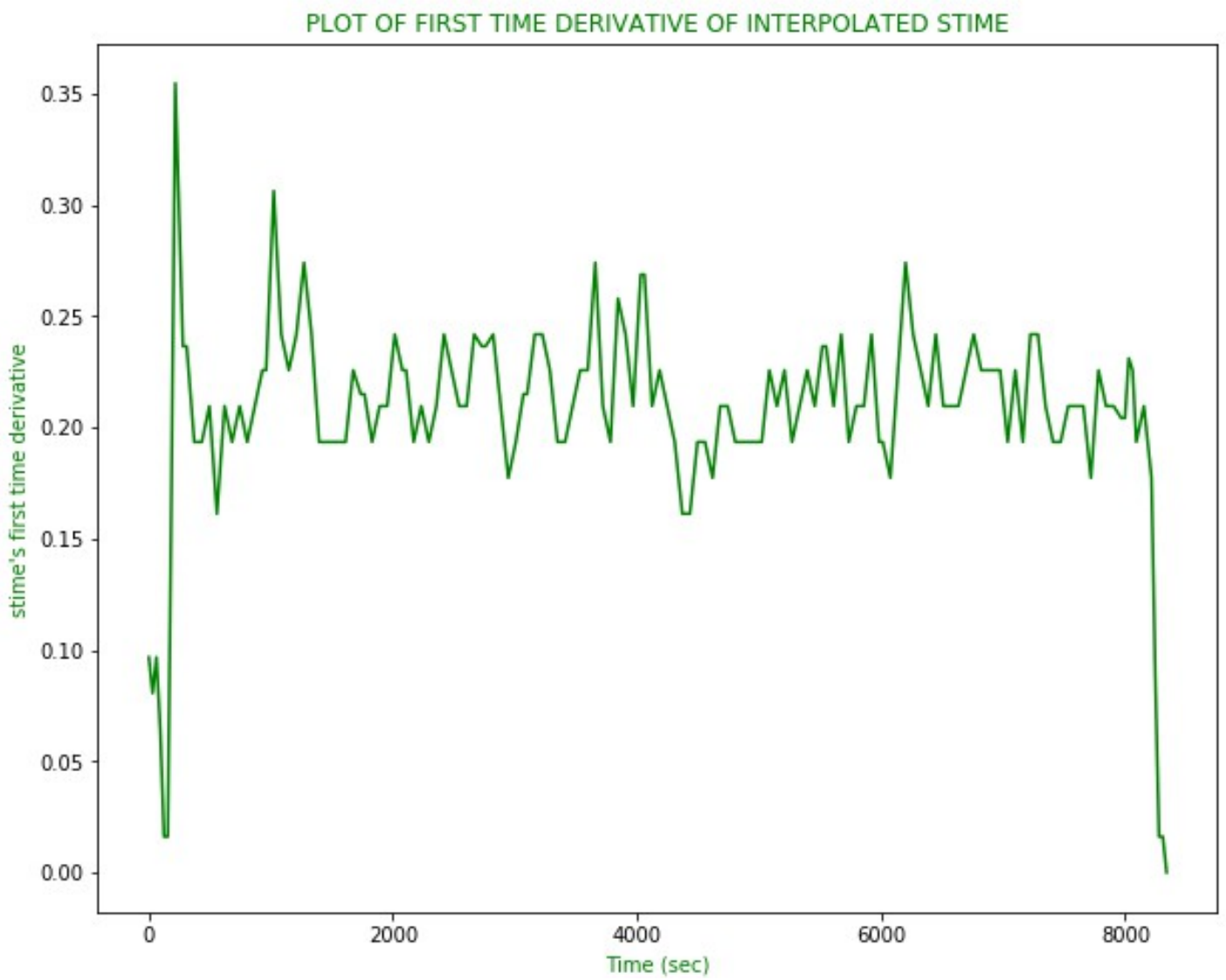
- Analyze data without the need to install any software.
- Have Jupyter notebook interface as well as shell access from browser.
- Use CERNBox as your home directory and synchronize your local user storage with the cloud.

Time series of metrics

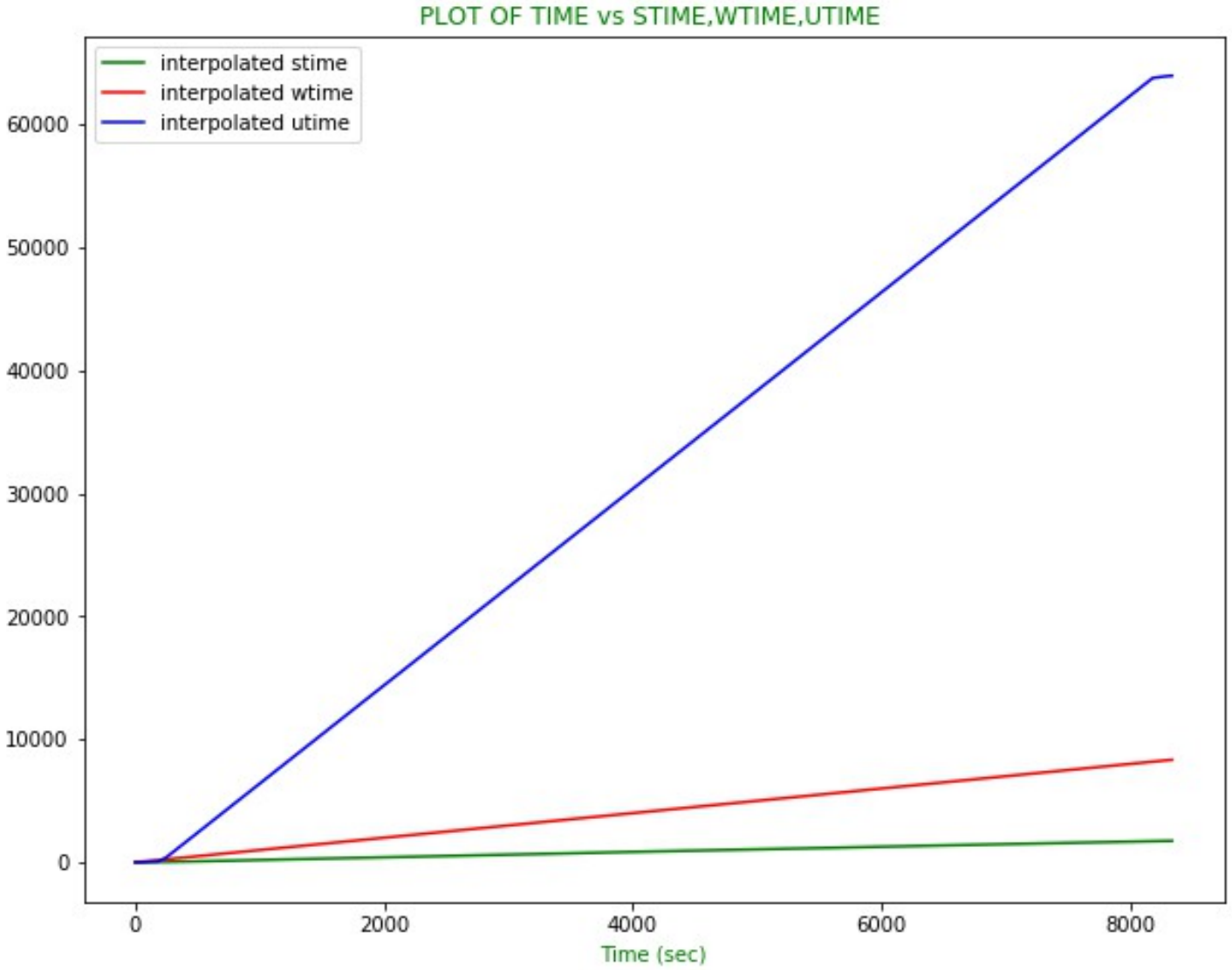




Time series of metrics

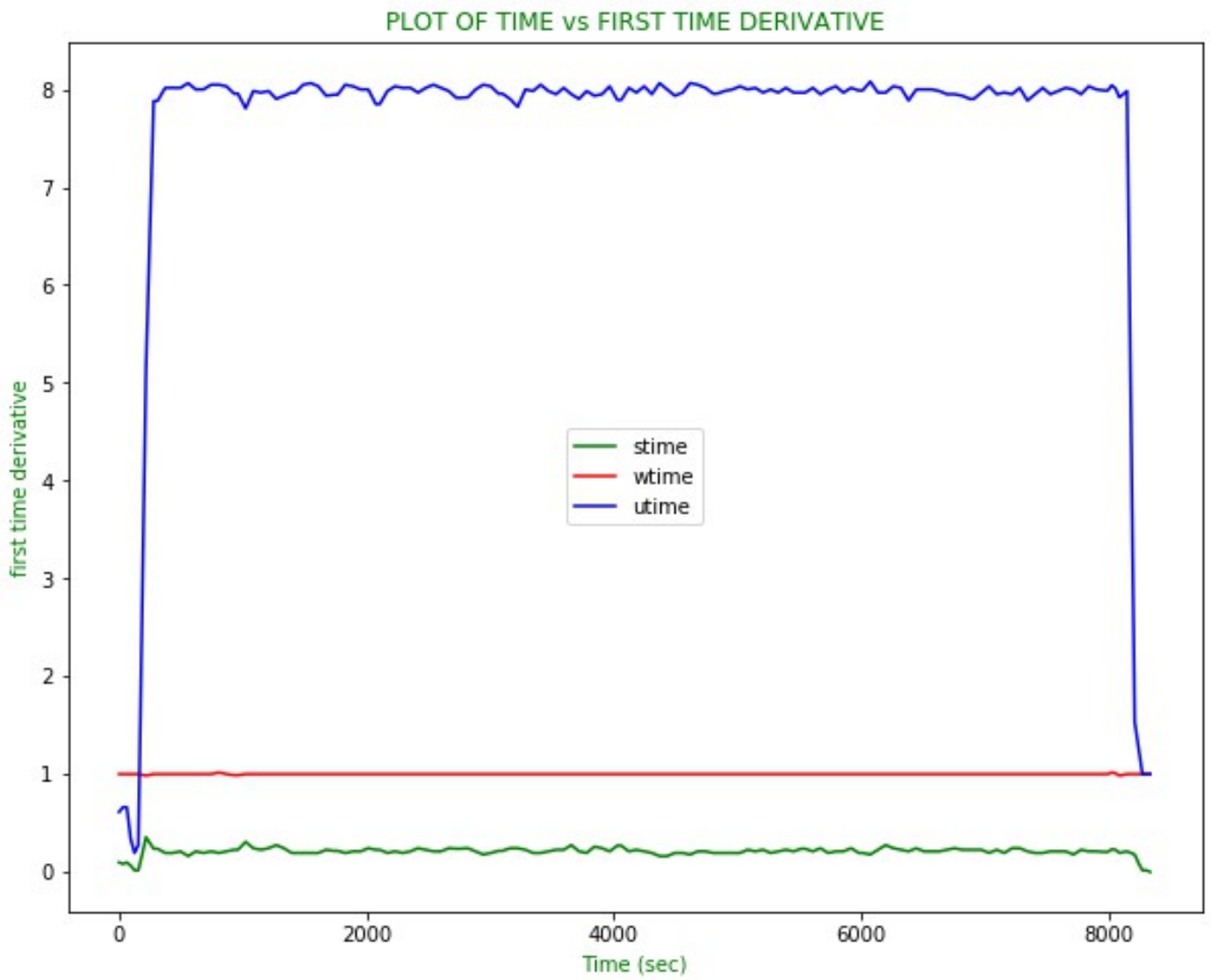


Time series of metrics

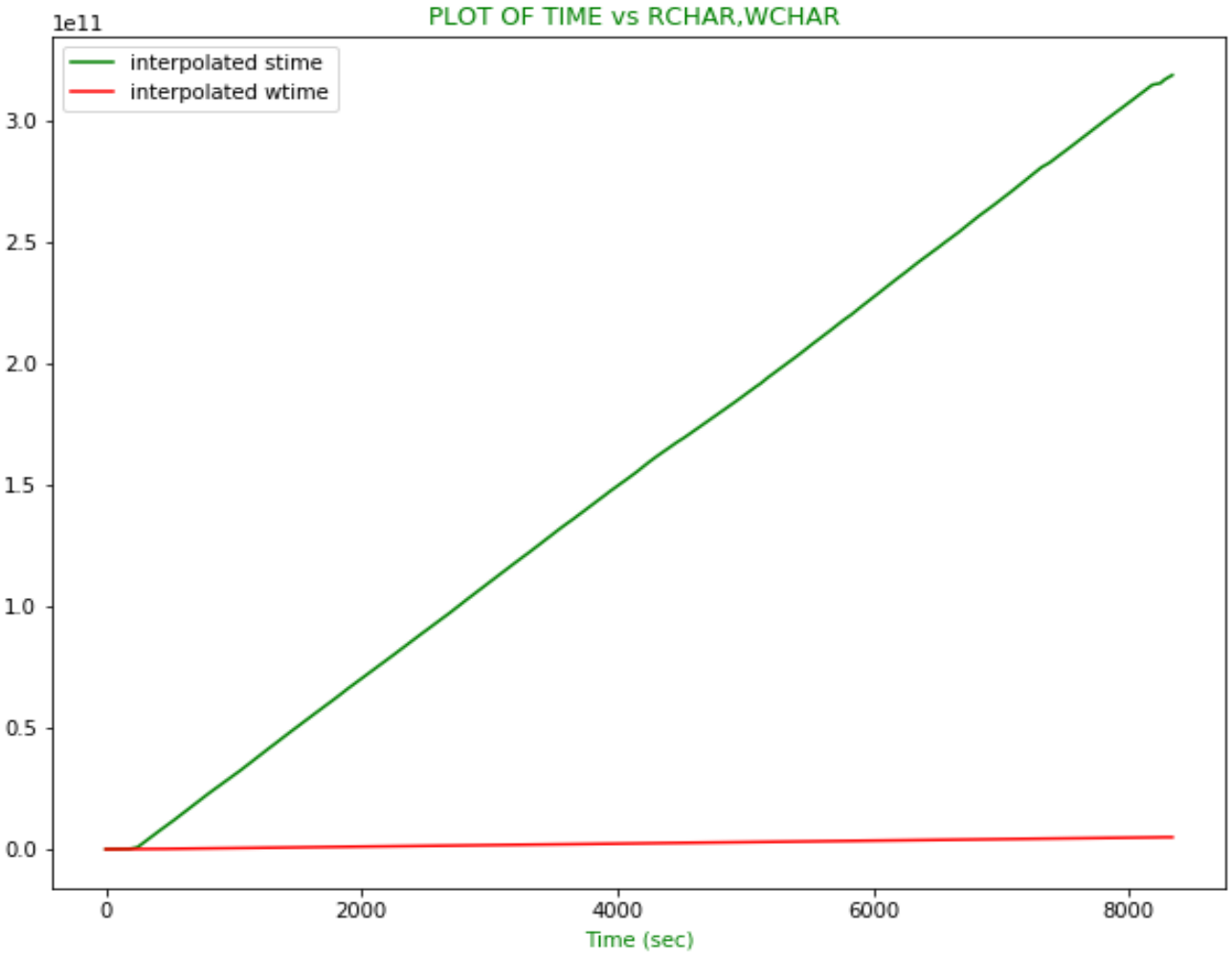


CERN openlab

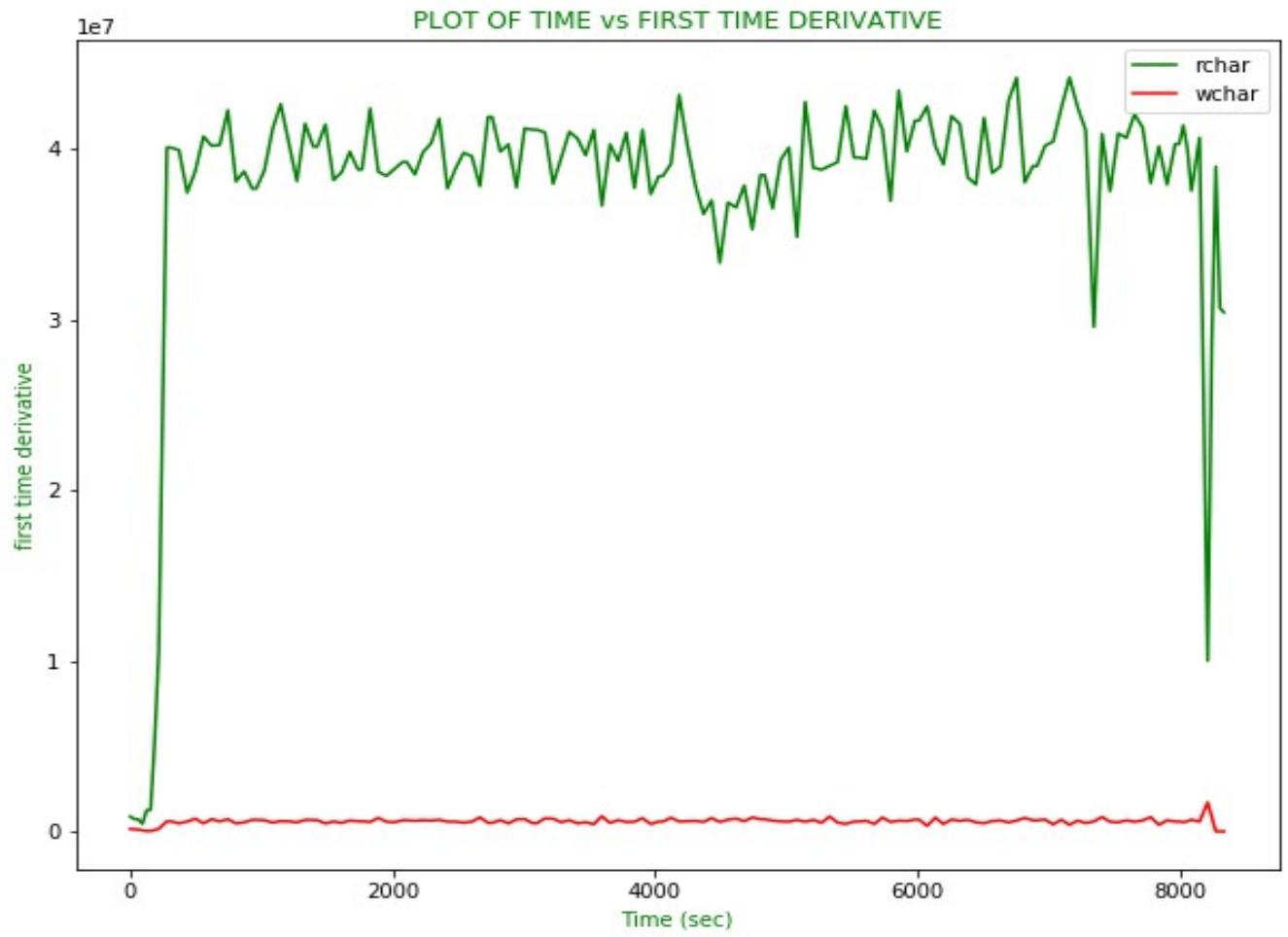
Time series of metrics



Time series of metrics

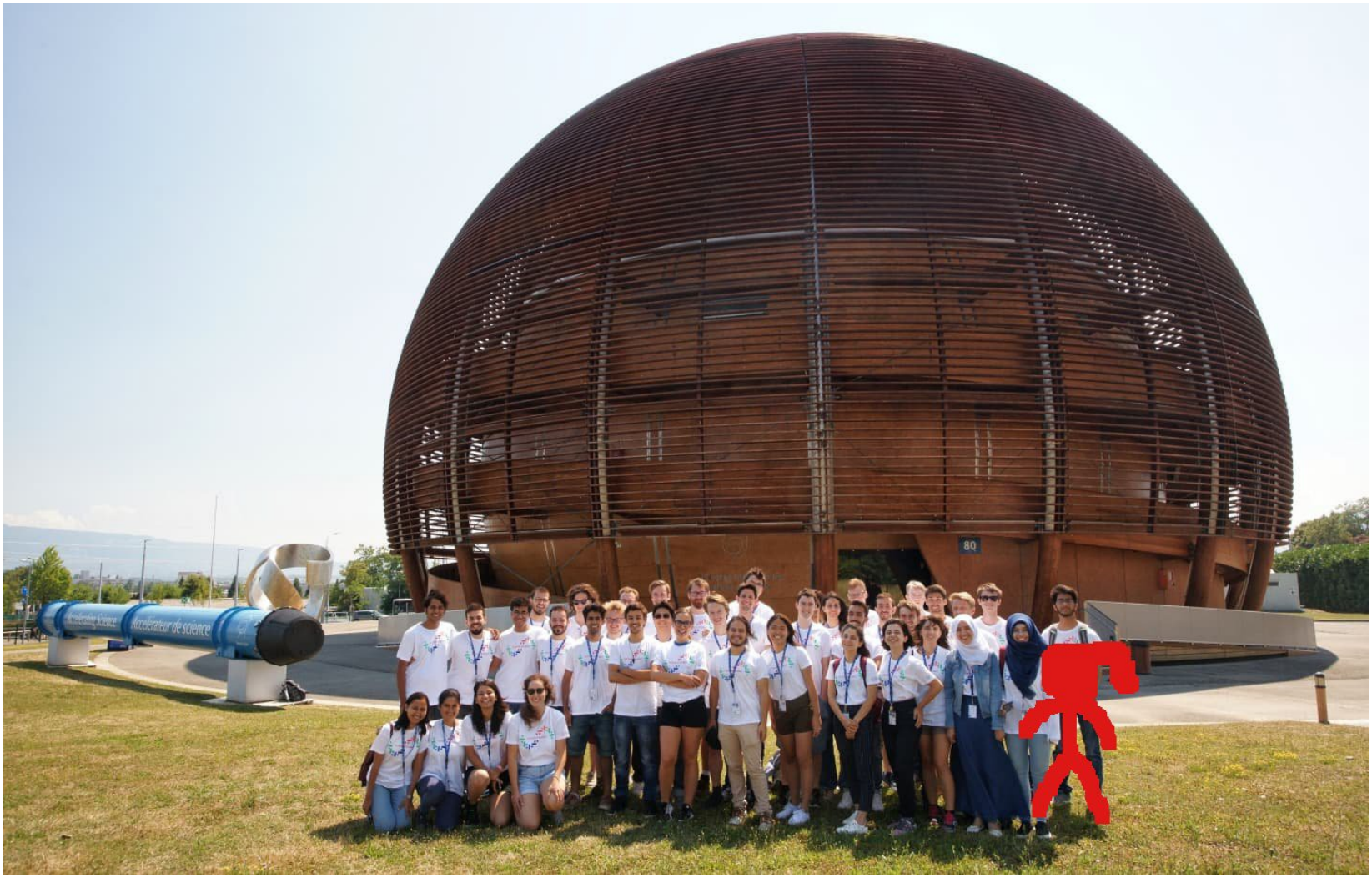


Time series of metrics



Next steps

- Implement a time series segmentation method based on first or second derivatives.
- For each time series segment, characterize the segment with the least possible number of parameters.
- demonstrate the full process that starts from a prmon time series and produces a series of change points separating segments and values describing the segments


80

THANK YOU!

alexia.topalidou@cern.ch