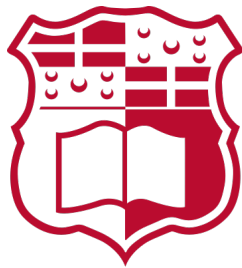


# Machine learning for data mining and performance optimization at the CERN Large Hadron Collider

## Progress Report

Marc Ferriggi

Supervisor: Dr. Gianluca Valentino



**L-Università  
ta' Malta**

**Faculty of Science  
University of Malta  
December 2018**

*Submitted in partial fulfillment of the requirements for the degree of B.Sc. (Hons.)  
Computing Science AND Statistics and Operations Research*

# Contents

1	Abstract . . . . .	3
2	Introduction & Motivation . . . . .	3
3	Why is the Problem non-Trivial . . . . .	3
4	Background Research and Literature Review . . . . .	4
5	Aims and Objectives . . . . .	4
6	Methods and Techniques Used or Planned . . . . .	4
7	The Evaluation Strategy and Technique being Proposed . . . . .	4
8	Deliverables . . . . .	4
9	Progress . . . . .	4

## **Abstract**

The CERN particle accelerator complex generates around 2 TB of data per week from almost 1 million signals. In this dissertation, unsupervised machine learning techniques for applications such as clustering and anomaly detection shall be used to analyse past LHC data in order to visualize correlations, determine data driven models and identify opportunities for improving the LHC machine availability and performance reach.

## **Introduction & Motivation**

The LHC is filled to flat top intensity by injecting each beam with kicker waves 12 times [1]. This is a challenging task given the high energy of the beam, the very small apertures and the delivery precision's tight tolerances, thus multiple sensors are installed around the CERN particle accelerator complex [2] which gather readings and data that can be used to check the quality of the injected beam. This data is stored using CERN's LS (Logging Service). The LS is heavily used and in 2013, it was noted that close to 1000 users relied on it [3]. While many studies have been made using this logged data and lots of statistical tests have been done with regards to injection quality checks for the LHC (such as [4] and [1]), no studies can be found on the CERN Document Server [5] where researchers used unsupervised machine learning methods to analyse this data.

The purpose of this thesis is to go through the entire process in a data science project and learn to use unsupervised machine learning techniques for applications such as clustering and anomaly detection. These techniques will then be used to analyse past LHC data in different machine configurations to visualise correlations, determine data-driven models and identify opportunities for improving the LHC machine availability and performance reach in terms of beam lifetime, beam stability and luminosity.

## **Why is the Problem non-Trivial**

The LS currently generates around 2 TB of data per week. This data is generated from multiple machines each measuring different features of the waves at different points in the accelerator complex. Thus the data must be thoroughly analysed and normalised in order to be able to apply the machine learning analysis techniques properly. Identifying opportunities to improve the beam lifetime, beam stability and luminosity is also a non-trivial problem to tackle.

**Background Research and Literature Review**

**Aims and Objectives**

**Methods and Techniques Used or Planned**

**The Evaluation Strategy and Technique being Proposed**

**Deliverables**

**Progress**

# Bibliography

- [1] V. Kain *et. al*, “Injection beam loss and beam quality checks for the lhc.” in *Proc. IPAC*, 2010, pp. 1671-1673.
- [2] C. Lefevre. “The cern accelerator complex.” Technical report, 2008.
- [3] C. Roderick, L. Burdzanowski and G. Kruk. “The cern accelerator logging service- 10 years in operation: a look at the past, present and future,” presented at the 14<sup>th</sup> Int. Conf. Accelerator & Large Experimental Physics Control Systems, USA, 2013.
- [4] L. N. Drosdal *et. al.*, “Automatic injection quality checks for the lhc.” in *Proc. ICALEPCS*, 2011, pp. 1077-1080.
- [5] “Cern document server” Internet: `cds.cern.ch`, [Nov. 11, 2018].