



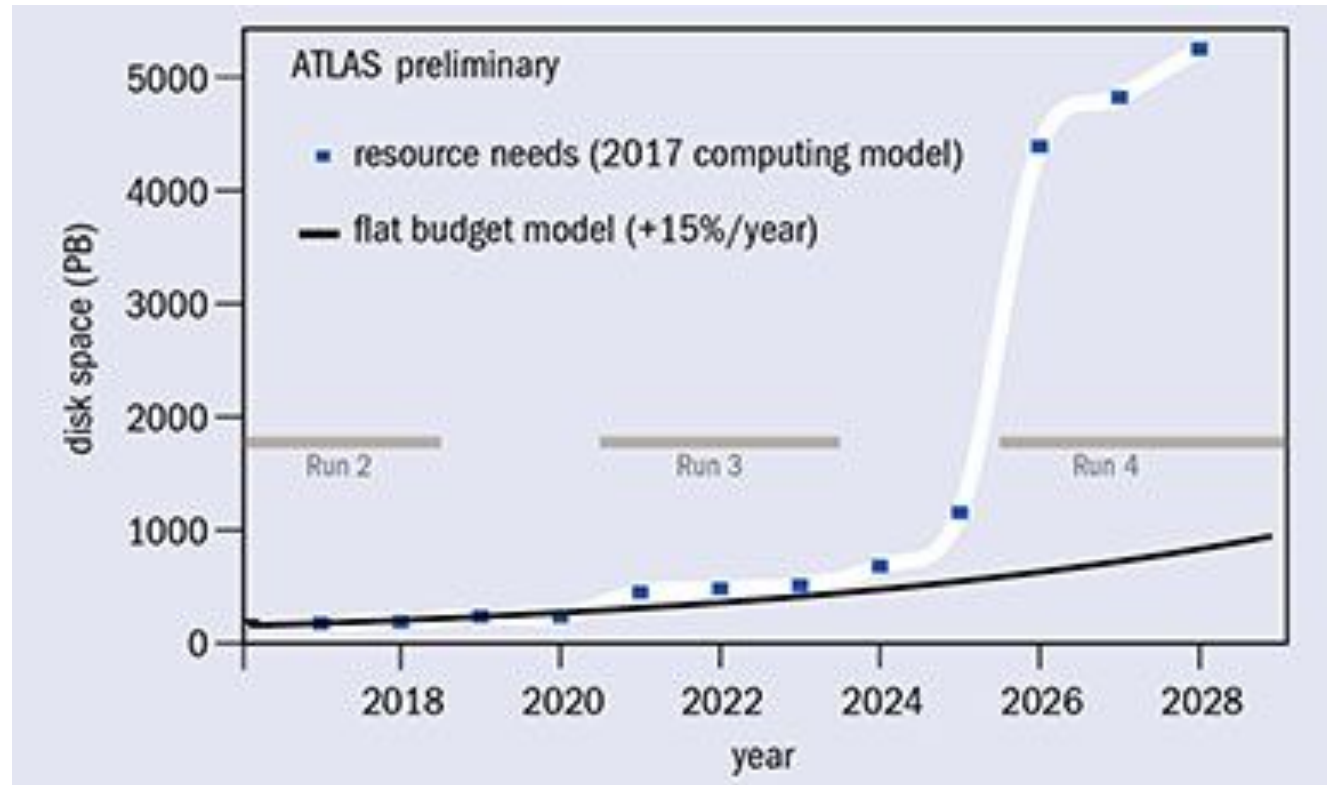
Machine learning based lossy compression

Alexander Ekman and Axel Gallén



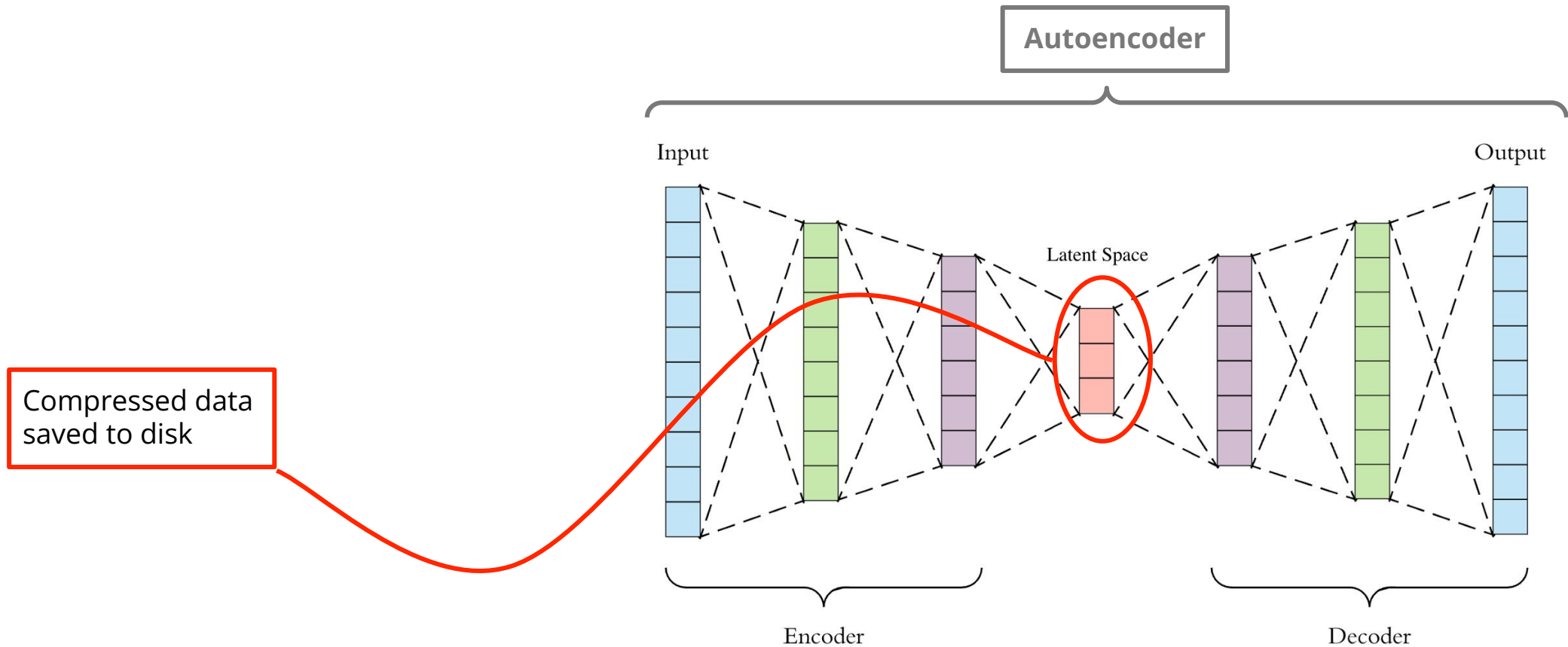
The problem

- Problem: Too much data, too little storage
 - High demand for compression



The solution

- Lossy compression needs to be tailored
 - Solution: Lossy Machine Learning based compression



The catch

- Lossy compression comes with a price:
 - Decompressed data is not equal to original data
 - Does this mean that lossy compression is bad?
- Works well in cases where more data is better
 - For example: Particle physics, where more data compensate for the loss
- Works extremely well in cases where data would be thrown away
 - Trigger level analysis

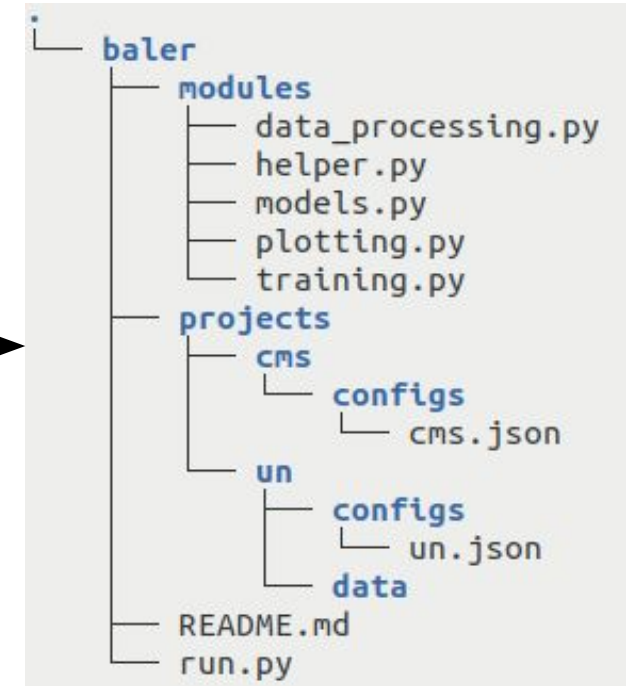
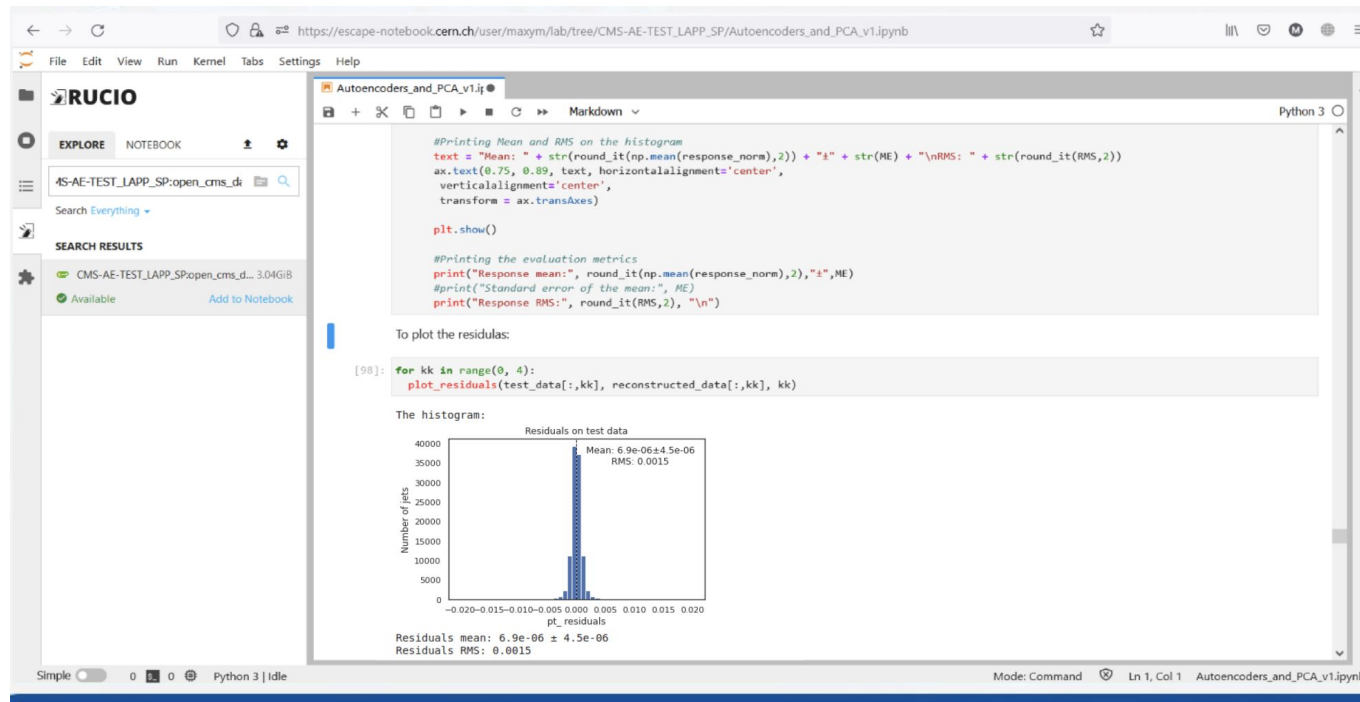
The prologue

- Multiple previous studies have on ATLAS data by Caterina and her students
- Me and Axel inherited a jupyter notebook, hosted on the European open science cloud
- We got a Manchester based grant of 100 000 kr to make the tool available for other researchers
- We were invited to present this at a conference in Brussels
- Interest from the general physic community was huge
- Got a phone call LTH computer science lecturer concerning a new master student
 - But then she asked for the **documentation.**

The prologue

- Multiple previous studies have on ATLAS data by Caterina and her students
- Me and Axel inherited a jupyter notebook, hosted on the European open science cloud
- We got a Manchester based grant of 100 000 kr to make the tool available for other researchers
- We were invited to present this at a conference in Brussels
- Interest from the general physic community was huge
- Got a phone call LTH computer science lecturer concerning a new master student
 - But then she asked for the **documentation.**
- Panic ensued

Refactoring



- Refactored jupyter notebook to python library

Project management

Projects / baler

Backlog

AG NS P +2

Epic Type

Insights

BALER Sprint 3 29 Nov - 9 Dec (12 issues) 1w 1d 1h 2d 0m Complete sprint

Obtain HEP physics results and start running on CFD data

- BALER-78 Reproduce previous response performance PHASE 1 2d PENDING MERGE AG
- BALER-83 Save as .ROOT HEP PUBLICATION 5h TO DO AG
- BALER-84 mji plots / Higgs analysis HEP PUBLICATION 3d TO DO AG
- BALER-51 Fix decompressed negative values PHASE 1 TO DO P
- BALER-55 Create HDF5 to Pandas dataframe function PHASE 1 3h TO DO MS
- BALER-53 Publishable plotting PHASE 1 3h TO DO MS
- BALER-70 Share HDF5 datafiles CFD PUBLICATION 3h TO DO MS
- BALER-56 Set up cluster accounts GPU/CLUSTER 1d TO DO NS
- BALER-58 Update use case diagram PHASE 1 1h TO DO MS
- BALER-81 Update PRD 2h TO DO MS
- BALER-82 implement one unit test for reading root file TO DO NS
- BALER-86 Write thesis goal document OPEN SOURCING TO DO FB

+ Create issue

Backlog (14 issues) 2d 3h 0m 0m Create sprint

- BALER-64 Contact Lucas Heinrich HEP PUBLICATION IN PROGRESS MS
- BALER-57 Implement regression test PHASE 1 3h TO DO MS
- BALER-36 Upload to the VRE 2d TO DO MS
- BALER-23 Windows support TO DO MS
- BALER-63 Add variables to exclude from compression TO DO MS
- BALER-65 Plots of latent space TO DO MS
- BALER-66 Does latent space reproduce physical laws? CFD PUBLICATION TO DO MS
- BALER-68 Run Baler on Manchester cluster TO DO MS
- BALER-73 Bug in the normalization. Not reconstructing correctly TO DO MS
- BALER-74 Try energy mover's distance as additional way to give weight to nodes (Ben N's paper) TO DO MS
- BALER-75 Implement χ^2 alongside all plots TO DO MS
- BALER-76 Fix poor performance TO DO MS
- BALER-77 Contact authors of IEEE paper TO DO MS
- BALER-87 Training on CMS data crashes on saving output TO DO MS

Projects / baler

Roadmap

Give feedback Share Export

FB NS MS +3

Status category Epic Type

NOV DEC

Sprints

BALER Sprint 1 BALER Sprint 2 BALER Sprint 3

BALER-7 Phase 1

- BALER-5 Make root read function DONE MS
- BALER-6 Implement normalization DONE MS
- BALER-43 Learn new modularized code DONE MS
- BALER-3 Save compressed data to .csv DONE MS
- BALER-18 Implement Argparser DONE MS
- BALER-25 Plot response DONE MS
- BALER-44 Separate compression and decompressi... DONE MS
- BALER-34 Save compression models DONE MS
- BALER-57 Implement regression test TO DO MS
- BALER-38 Function to only compress data DONE MS
- BALER-37 Function to only decompress data DONE MS
- BALER-60 Loss plots DONE MS
- BALER-78 Reproduce previous response pe... PENDING ME... MS
- BALER-51 Fix decompressed negative values TO DO P
- BALER-55 Create HDF5 to Pandas dataframe funct... TO DO MS
- BALER-53 Publishable plotting TO DO MS
- BALER-58 Update use case diagram TO DO MS

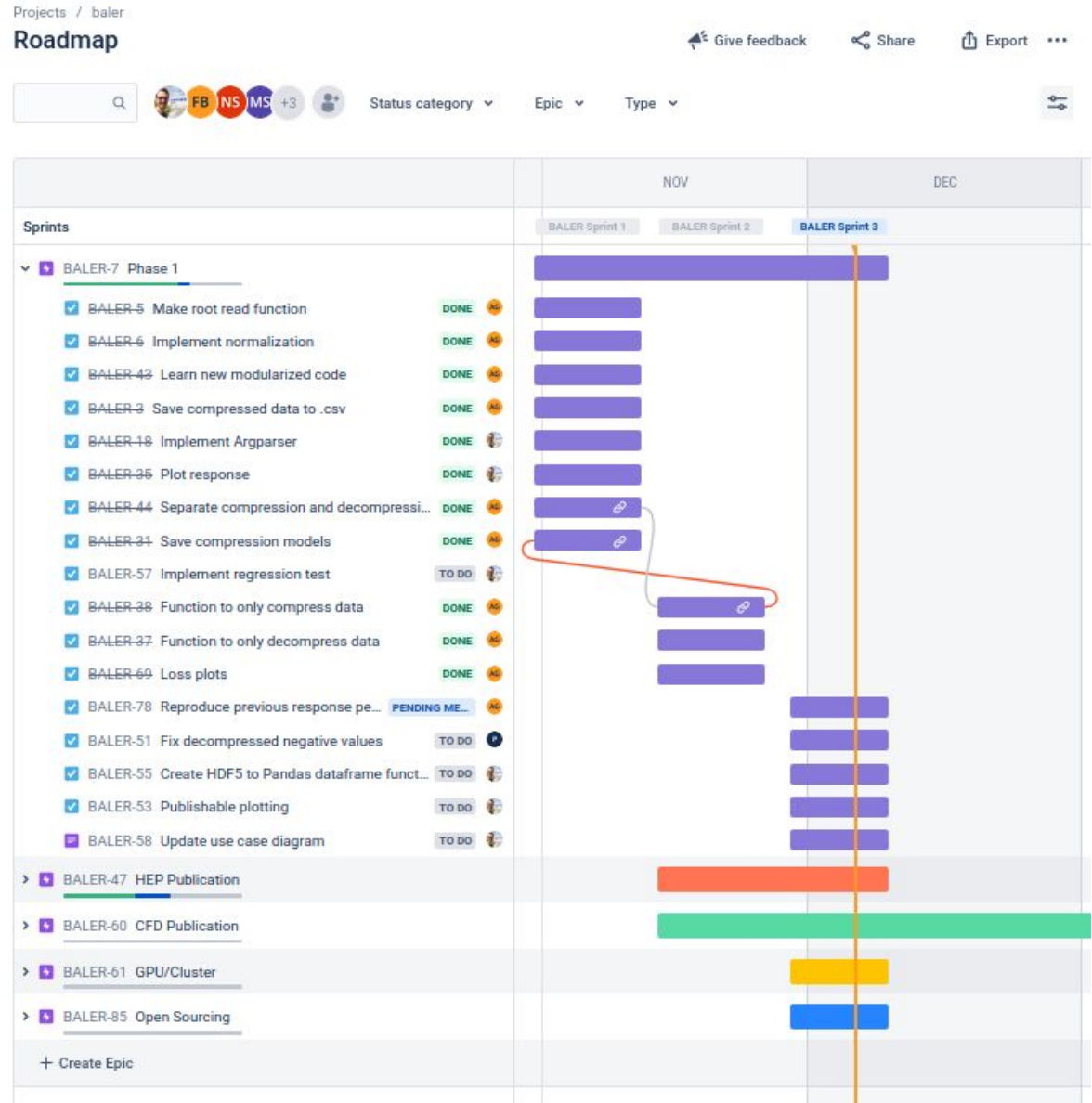
BALER-47 HEP Publication

BALER-60 CFD Publication

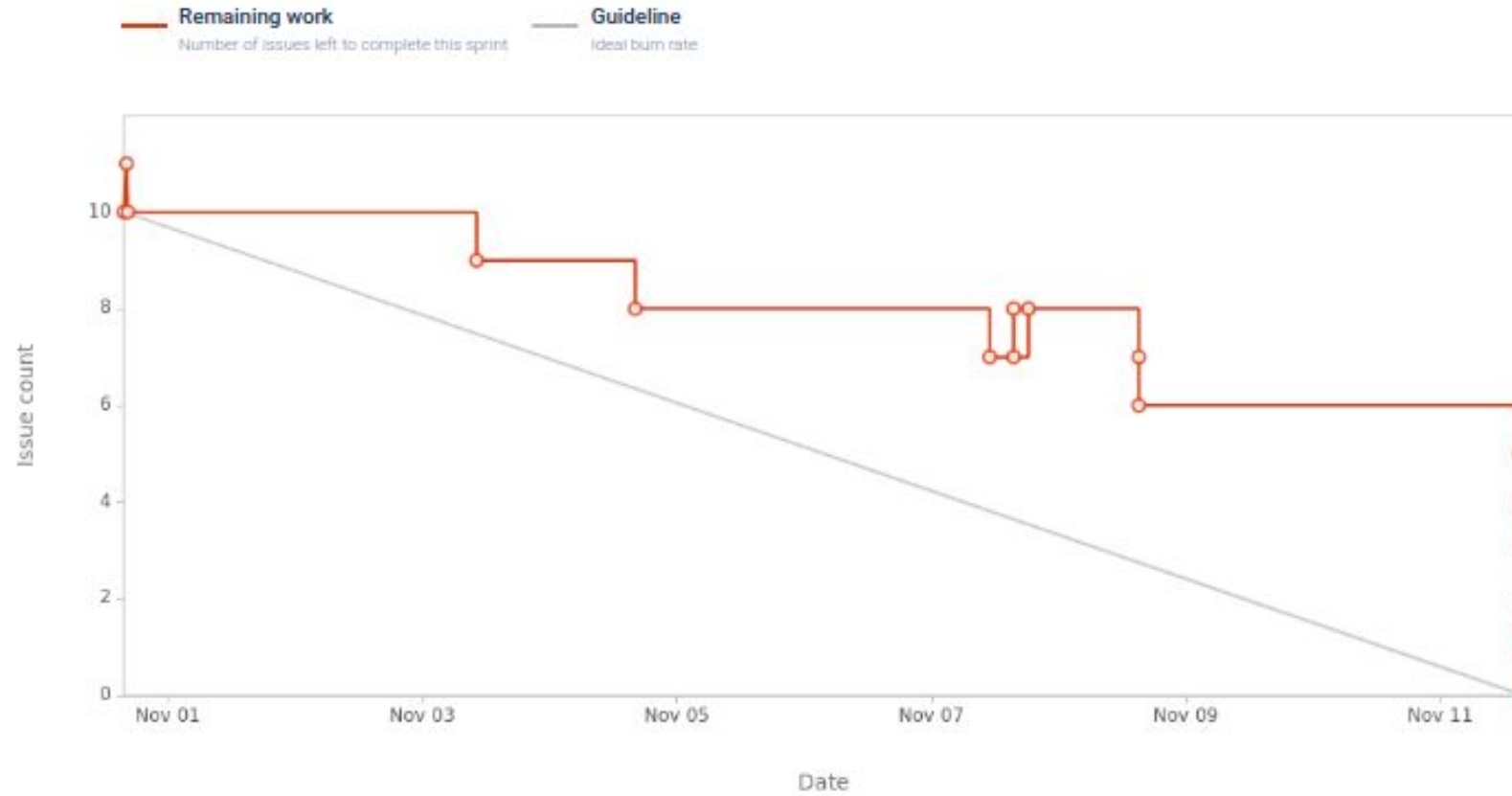
BALER-61 GPU/Cluster

BALER-85 Open Sourcing

+ Create Epic



Sprints

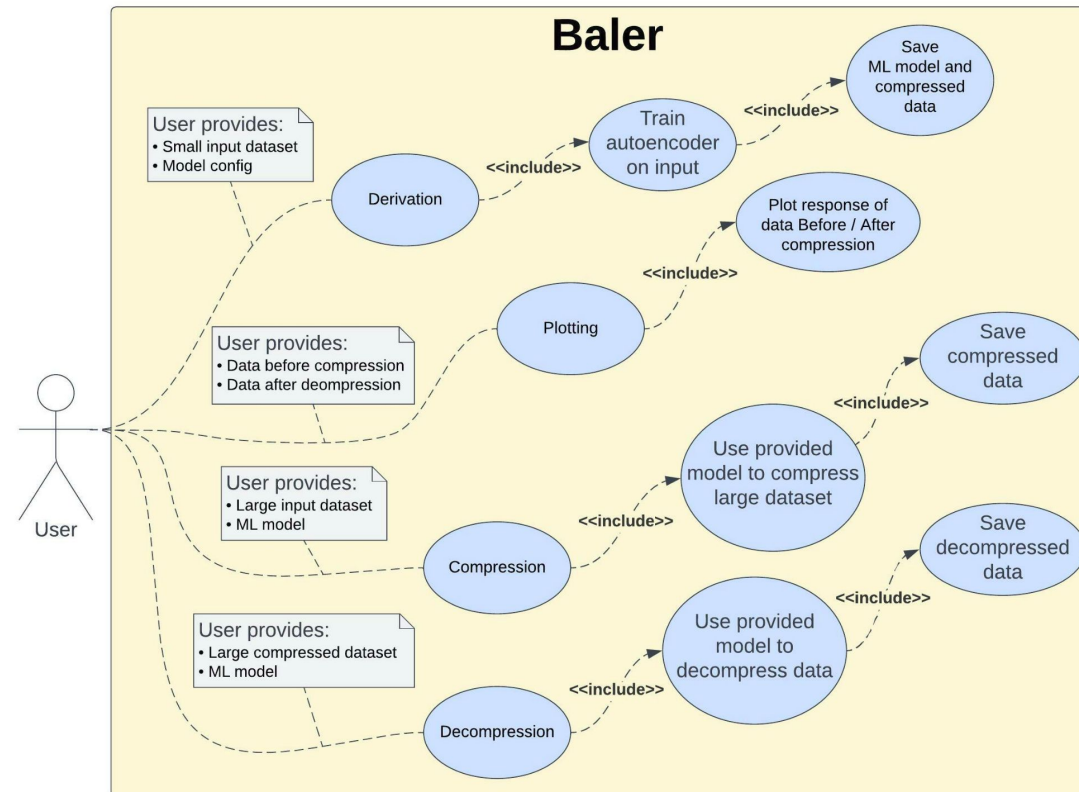


Baler



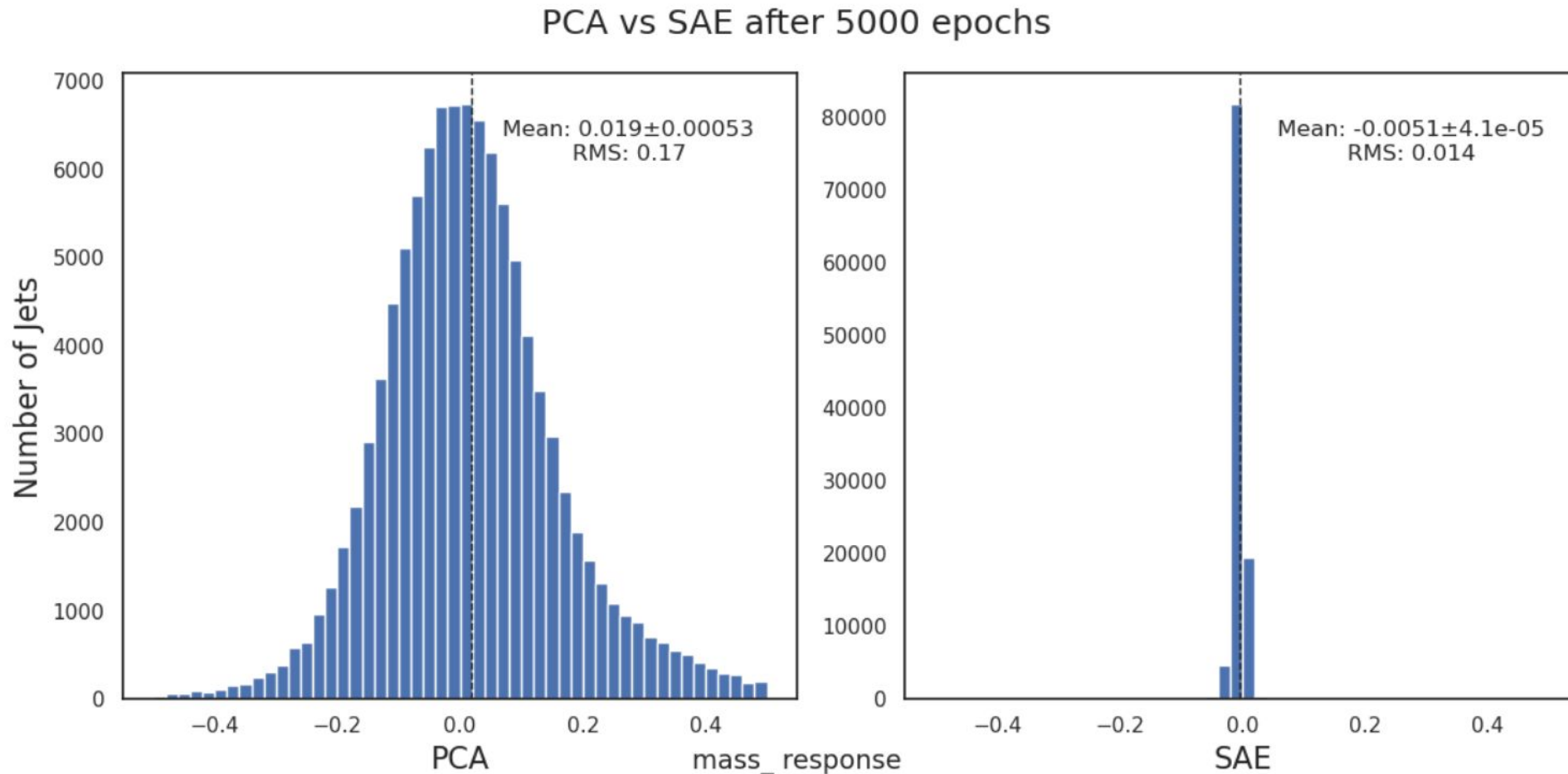
What is Baler

- It is a tool which lets you
 - Derive a machine learning model to optimally compress your dataset
 - Plot performance
 - Compress the dataset
 - De-compress the dataset



Does it work?

- Yes, and the work now is to determine how well it works
 - We can reconstruct most jet data with maximum response of 4%
 - A lot better than simpler methods like PCA



Future work

- Use baler to systematically reproduce the previous findings and publish results for HEP data
- Work together with team of 8 + RSE to create a publicly available open source tool to be used by researchers in vastly different fields

Takeaway points

- Jupyter notebooks are great as a persuasion tool, but not for tool development
- Basic organization of code and work backlogs get you very far
 - Adding collaborators is much simpler and needs much less meetings
- I would highly recommend Alma's course in large scale software development
 - [ETSN05](#)
- Try something new!



LUND
UNIVERSITY