MAchinE Learning for Scalable meTeoROlogy and climate

D4.5 MAELSTROM Boot Camp

Bing Gong (FZJ)

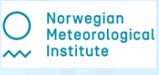
www.maelstrom-eurohpc.eu

D4.5 MAELSTROM Boot Camp

Author(s): Bing Gong (FZJ)

Dissemination Level: Public

Date: 21/10/2022
Version: 1.0
Contractual Delivery Date: 21/10/2022
Work Package/ Task: WP4/ T4.5
Document Owner: FZJ
Contributors: Bing Gong
Status: Final



MAELSTROM

Machine Learning for Scalable Meteorology and Climate

Research and Innovation Action (RIA)
H2020-JTI-EuroHPC-2019-1: Towards Extreme Scale Technologies and Applications

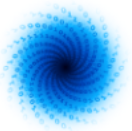
Project Coordinator: Dr Peter Dueben (ECMWF)
Project Start Date: 01/04/2021
Project Duration: 36 months

Published by the MAELSTROM Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX, United Kingdom
Peter.Dueben@ecmwf.int

The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway

EuroHPC
Joint Undertaking



Contents

1 Executive Summary **5**

2 Introduction **6**

2.1 About MAELSTROM 6

2.2 Scope of this deliverable 6

2.2.1 Objectives of this deliverable *6*

2.2.2 Work performed in this deliverable *6*

2.2.3 Deviations and counter measures *7*

3 MAELSTROM Boot Camp outline **8**

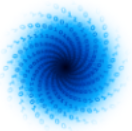
3.1 The agenda of Boot Camp 8

3.2 Training curriculum 8

3.3 5-min presentation and feedback 9

4 Participants, tutors and networking **11**

5 Women in Science **12**



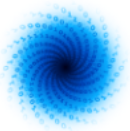
1 Executive Summary

MAELSTROM Boot Camp was organised and run in Juelich Supercomputing Center (JSC) between 27th September to 30th September 2022 to facilitate early dissemination of MAELSTROM outcomes to early-career scientists in the W&C prediction and high performance computing (HPC) communities.

The goal of this four-day Boot Camp was to provide training to the participants to use machine learning (ML) in weather and climate applications on HPC systems. The participants have explored the applications developed in the MAELSTROM Project after an introduction to the scientific background of meteorology, ML, and HPC.

The Boot Camp has covered the following topics:

1. An overview of the MAELSTROM project
2. Introductory lectures on ML methods, meteorology, and HPC systems
3. Lectures on bespoken ML approaches of the six meteorological applications in MAELSTROM
4. Comprehensive hands-on tutorials in the scope of the MAELSTROM applications



2 Introduction

2.1 About MAELSTROM

To develop Europe's computer architecture of the future, MAELSTROM will co-design bespoke compute system designs for optimal application performance and energy efficiency, a software framework to optimise usability and training efficiency for machine learning at scale, and large-scale machine learning applications for the domain of weather and climate science.

The MAELSTROM compute system designs will benchmark the applications across a range of computing systems regarding energy consumption, time-to-solution, numerical precision and solution accuracy. Customised compute systems will be designed that are optimised for application needs to strengthen Europe's high-performance computing portfolio and to pull recent hardware developments, driven by general machine learning applications, toward needs of weather and climate applications.

The MAELSTROM software framework will enable scientists to apply and compare machine learning tools and libraries efficiently across a wide range of computer systems. A user interface will link application developers with compute system designers, and automated benchmarking and error detection of machine learning solutions will be performed during the development phase. Tools will be published as open source.

The MAELSTROM machine learning applications will cover all important components of the workflow of weather and climate predictions including the processing of observations, the assimilation of observations to generate initial and reference conditions, model simulations, as well as post-processing of model data and the development of forecast products. For each application, benchmark datasets with up to 10 terabytes of data will be published online for training and machine learning tool-developments at the scale of the fastest supercomputers in the world. MAELSTROM machine learning solutions will serve as blueprint for a wide range of machine learning applications on supercomputers in the future.

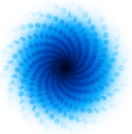
2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

To provide an overview of the MAELSTROM project and the six applications, as well as the opportunities and challenges for the use of ML in weather and climate predictions; To provide hands-on training sessions for scientists to learn and use the ML benchmark datasets and ML solutions developed in WP1; To collect further information about the requirements of ML use cases within the community; To get feedback on the improvement for the next MAELSTROM Hackathon.

2.2.2 Work performed in this deliverable

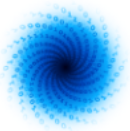
MAELSTROM Boot Camp Organising committee (Bing Gong, Amirpasha Mozaffari, Michael Langguth from JSC, and supervised by Martin Schultz) was formed at the beginning of 2022. The organising committee had regular (monthly) meetings to discuss, plan the agenda and coordinate the



preparation of all the relevant activities for the Boot Camp. The organising committee invited experts to give lectures for the first Day of the Boot Camp. The developers of the six applications designed and developed the hands-on training materials. A four-day Boot Camp was successfully held in Juelich Supercomputing Center (JSC), Forschungszentrum Jülich, Germany, from 27th to 30th September 2022. More than 30 participants and 16 tutors joined this event.

2.2.3 Deviations and counter measures

There are no significant deviations from the planned contributions of the deliverable.



3 MAELSTROM Boot Camp outline

3.1 The agenda of Boot Camp

A four-day Boot Camp was held at the Juelich Supercomputing Center (JSC), Forschungszentrum Jülich, Germany from 27th to 30th September, 2022. The agenda can be found on Gitlab[1]. On Day 1, we had a series of lectures to introduce the MAELSTROM Project, machine learning in weather and climate, HPC systems, weather and climate predictions, and scalable deep learning theories.

From Day 2 on, the participants could choose between six groups corresponding to six applications based. The application developers provided a hands-on training section on using ML benchmark datasets and ML solutions for the application. The six teams joined together at the main venue of the Boot Camp and gave a five minute pitch by participants/tutors in front of all teams to share their progress, and what they learned at the end of the day.

The same schedule holds for Day 3. On Day 4, the hands-on training sections were provided by the tutors in the morning. In the afternoon, each team again gave 5-minute of pitch followed by two minutes of feedback. The details of the feedback will be given in Section 5 of this report.

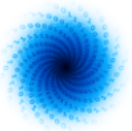
By the end of the Boot Camp, we organised a tour to visit three institutes of FZJ (IBG2, IBG3 and IKP4) to provide insight to all the participants about the research that is currently carried out in FZJ. Starting with an introductory lecture 'Forschungszentrum Jülich', Dr. Christopher Wolf gave an overview on the history and the research landscape of the entire FZJ. Hendrik Böhmer from IGB-2, Prof. Dr. NIcolas Brüggemann from IBG-3, and Yury Valdau from IKP-4 shortly introduced their institutes the participants could visit. The participants finally chose the institutes of interest and spent 1.5 hours on the tour before leaving FZJ.

3.2 Training curriculum

The training materials were developed by the lecturers and the six application developers. The materials were published and can be accessed through Gitlab[2]. The goal of this four-day Boot Camp was to provide training for the participants to use ML in weather and climate applications on HPC systems. Therefore, several lectures were designed on the first day to provide an introduction to the theoretical background of (scalable) Machine Learning, Weather and Climate domain, and HPC systems to the participants. Each lecture was given for 1-1.5 hours. The lectures include:

- MAELSTROM introduction and overview (Peter Dueben)
- Introduction to Machine Learning in weather and climate (Martin Schultz)
- Introduction to HPC system (Andreas Herten)
- Introduction to Scalable Deep Learning (Jenia Jitsev)
- Introduction to weather and climate (Michael Langguth and Matthew Chantry)

[1] https://gitlab.jsc.fz-juelich.de/esde/training/maelstrom_bootcamp/-/blob/master/Agenda.md
[2] https://gitlab.jsc.fz-juelich.de/esde/training/maelstrom_bootcamp/-/blob/master/



Regarding the hands-on training section, the application developers had flexibility to design the training materials including the lectures and exercises. The scope of the training is limited to introduction to the applications, the Tier 1/Tier 2 datasets and ML solutions in the Deliverable 1.1, Deliverable 1.2, and Deliverable 1.3. For application 6, the ML workflow tool - MLflow was explored by the participants and used for ML solution comparisons.

3.3 5-min presentation and feedback

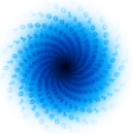
To encourage all the participants to work as a group and share what they had learned and experienced in the application hands-on training section, the tutors lead their team to prepare and give a 5-min presentation at the end of Day 2, 3 and 4. The presentations can be accessed through Gitlab[3].

It is worth to mention that the Boot Camp participants even managed to beat the benchmark for the machine learning solution of the ensemble post-processing application[4].

In the end, we also received valuable feedback from each team, including the pros and cons of the Boot Camp and the expectations for the next MAELSTROM Hackathon. The feedback is summarised in Table 1. It is noticed that three of the teams suggested that the training time was too short and that one more day would have been useful. Instability of JSC-Jupyter Notebook has been reported and this feedback will be passed to the JSC support team to further improve the service on JSC-Jupyter Notebooks.

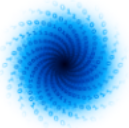
Team	Pros (+)	Cons (-)	Expectation for next MAELSTROM Hackathon
AP1	-	-	-Would be a good with a introductory to TensorFlow/PyTorch -One more day will be good
AP2	-Notebooks interesting, learned something -Topic was interesting, - -Nice balance: Some presentations + some tutorials -Focus on Deep Learning (and HPC)	-Stability of JSC-jupyter notebook	-Give options (depending on skill level) -Intro to ML+Deep Learning a bit longer and focus on big model ideas like FNNs, CNNs, RNNs, Transformers (maybe skip some detailed models), better motivation for different models -State-of-art ML (more approachable, more examples) -Use slurm to submit jobscript (write jobscript yourself) Hackathon: scrum -More bootcamp than hackathon because of training (lectures) provided -Talk about metrics in tutorials or talks -Record lectures and put them on youtube

[3] https://gitlab.jsc.fz-juelich.de/esde/training/maelstrom_bootcamp/-/tree/master/Pitch
[4] https://arxiv.org/abs/2206.14786



AP3	-	-	-Adding (some) materials maybe a little beforehand of the event starts -Collect all the slides from different applications to share Less jargon without explanation -Lectures a bit more spread in time -Cheat sheet
AP4	-	-Stability of JSC-jupyter notebook	-
AP5	-Very heterogeneous in terms of knowledge levels	-HPC systems are difficult -We learned a lot, but quite hectic due to limited time	-More differentiation what tasks are handled by whom, such that the team can overall progress more -Improved structured setting that gives more room for experimentation
AP6	-	-Sometimes, too technical words, especially on HPC	-Related topics (maybe some links to beginner courses will help) -More hands-on Sessions Machine/Deep Learning applied to Climate Data -Is 4 days not too short? -Recording lectures for easier digestion later

Table 1. Summary of the feedback and the expectation for the next MAELSTROM Hackathon from the participants



4 Participants, tutors and networking

There are 43 registers and more than 30 participants join Boot Camp on site. 16 supervisors (including the lecturers and application developers) provided lectures and training on the use of the MAELSTROM applications. The supervisors include:

- John Bjørnar Bremnes, Lonneke van Bijsterveldt, Thomas Nipen (MetNor)
- Fabian Emmerich, Kristian Ehlert (4cast)
- Matthew Chantry, Peter Dueben (ECMWF)
- Martin Schultz, Bing Gong, Michael Langguth, Yan JI, Andreas Herten, Stepan Nassyr, Jenia Jitsev (FZJ)
- Saleh Ashkboos, Langwen Huang (ETH)

The Helmholtz AI team at JSC led by Stefen Kesselheim also provided technical support for this event. MAELSTROM provided 2600 Euros travel support for two participants (1300 Euros for each) – selected from 26 applications – to enable them to join the meeting from Africa.

We also tried to create and maintain the networking through this Boot Camp. For example, in between the training section we organised the coffee break, and joint dinner activities to let participants and tutors have a chance to further communicate among each other. The tutors and participants shared their contacts (email, affiliation etc) among each other and the MAELSTROM Boot Camp group was created and established in LinkedIn[5].

Figure 1: Group photo of the participants of the MAELSTROM Bootcamp in front of the JSC building.

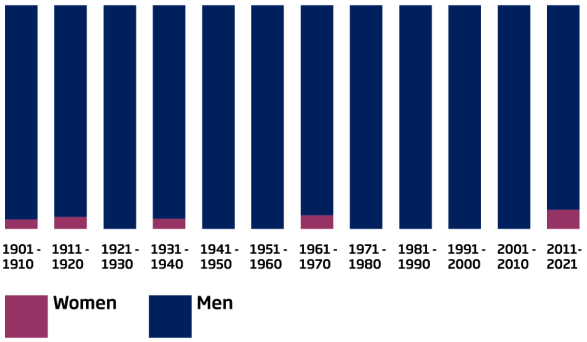
[5] https://www.linkedin.com/groups/12717048

5 Women in Science

In preparation of the web page on this topic and to generate first-hand content beyond desk research, we took the opportunity to have a diverse mix of people from different countries and backgrounds in one room and presented questions to the audience in a live poll. (Live poll means, questions are displayed on a large screen, each one in the audience can answer the question via mobile phone or own computer, and the aggregated results are immediately visible on the screen.)

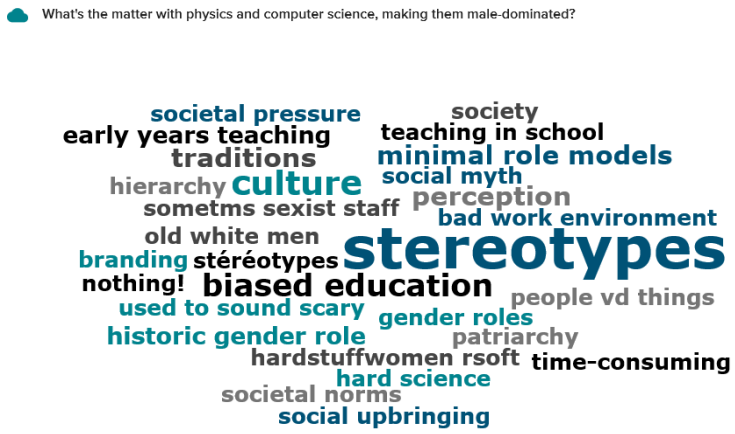
Questions were either open-ended, resulting in a word cloud, or multiple choice, resulting in bar charts. We sharpened the topic "women in science" to women in physics and computer science, for two reasons: On the one hand, those two disciplines are at the core of MAELSTROM's requirements; on the other hand, they seem to be even more male-skewed than e.g., biology.

To set the mind, we opened the poll with a quite striking chart showing Nobel Prize winners in physics and chemistry since the first price in 1901:



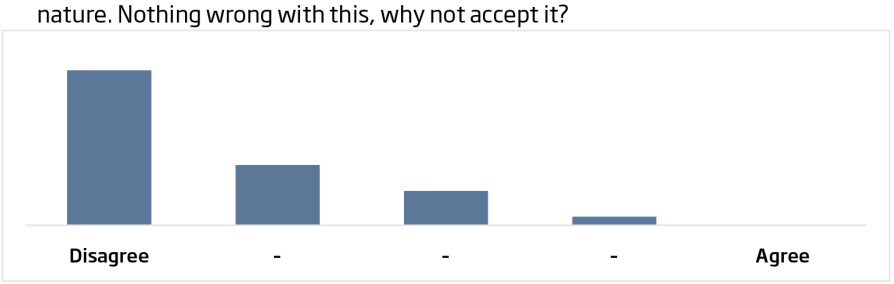
The subsequent questions and results were as follows:

Q1 Please help us understand: What's the matter with physics and computer science, making them so male-dominated?



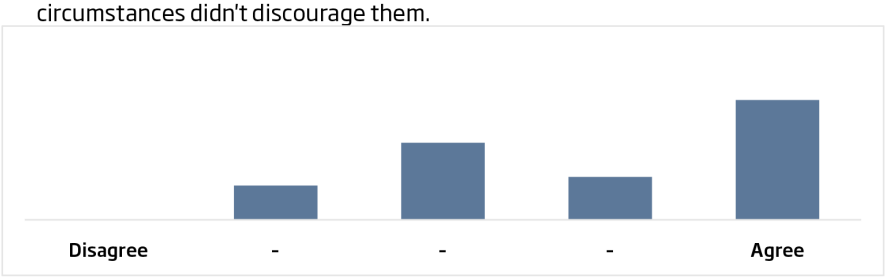
Q2 How far do you agree or disagree with the following statement?

Women are less interested in physics and computer science than men - it's by nature. Nothing wrong with this, why not accept it?

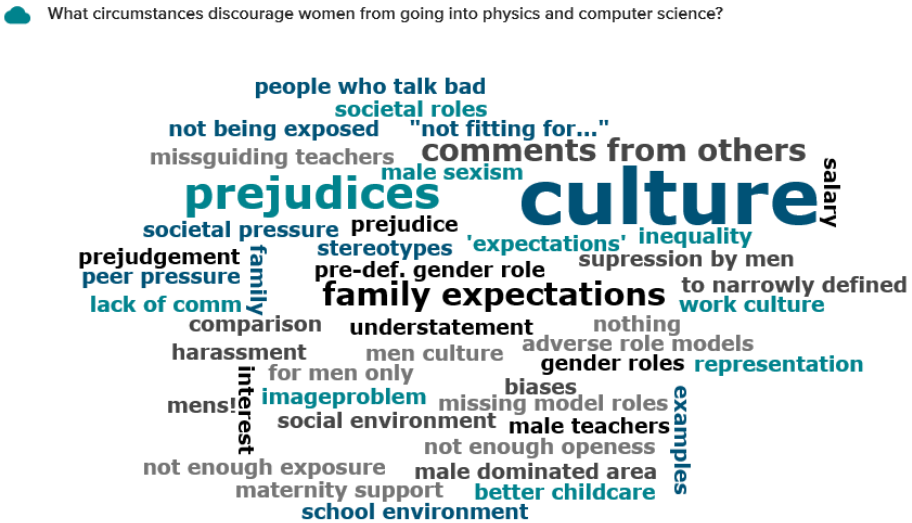


Q3 How far do you agree or disagree with the following statement?

Women would be equally interested in physics and computer science if only the circumstances didn't discourage them.

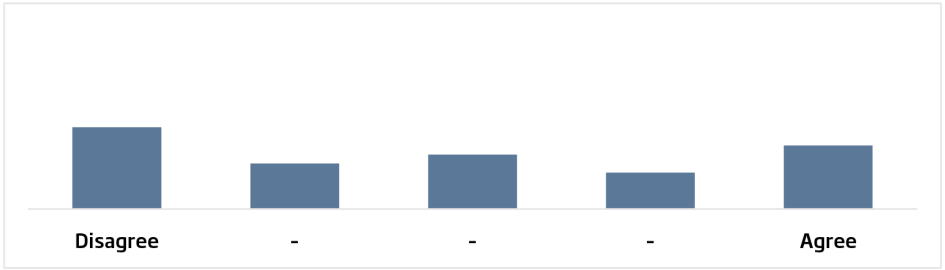


Q4 What circumstances discourage women from going into physics and computer science?



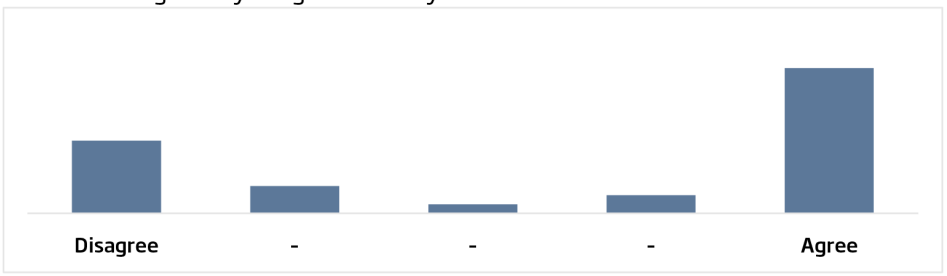
Q5 How far do you agree or disagree with the following statement?

Women have a different approach to doing research or solving scientific problems. We should address and embrace these differences.



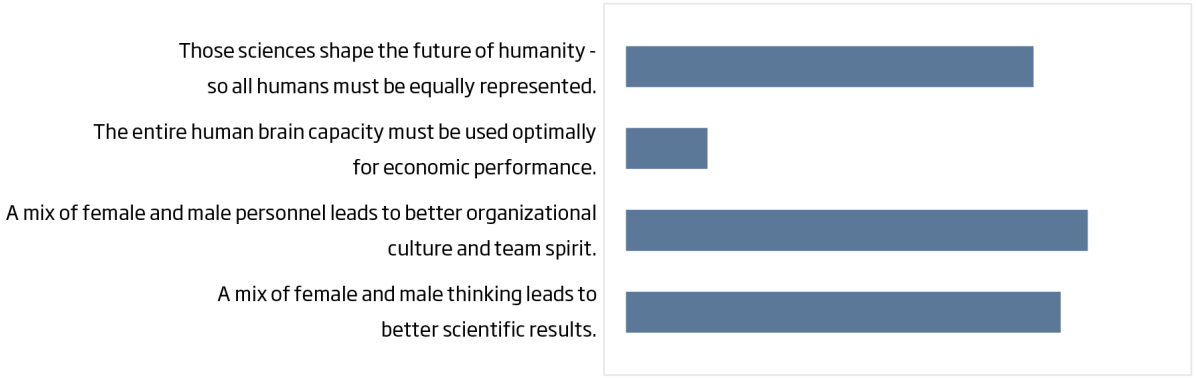
Q6 How far do you agree or disagree with the following statement?

When we try to address "female scientific approaches", we end up stereotyping but won't gain anything. Better stay clear.



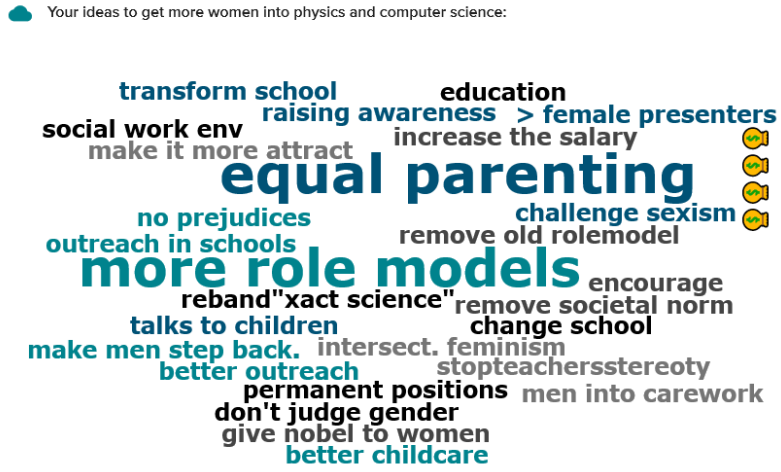
Q7 What do you think is **currently** motivating companies or institutes to increase the ratio of female science professionals?

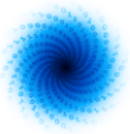
Q8 What **should** in fact be seen as the motives to to increase the ratio of female science professionals?



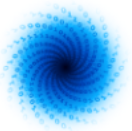
Q9 Can you think of other motives to get more women into physics and computer science?

Q10 What are your ideas to get more women into physics and computer science?





The topic “women in science” seems to have a vast resonance, as the lively participation in our poll and the demonstrated willingness for further discussion showed. Results will flow into our material collection and analysis for our “women in science” web page.



Document History

Version	Author(s)	Date	Changes
1.0	Bing Gong (FZJ)	19/10/2022	

Internal Review History

Internal Reviewers	Date	Comments
Peter Dueben (ECMWF)	19/10/2022	Revised with minor comments

Estimated Effort Contribution per Partner

Partner	Effort
Total	0