ML applications at PETRA

Ilya Agapov, LER workshop CERN 15 Feb 2024

Contents

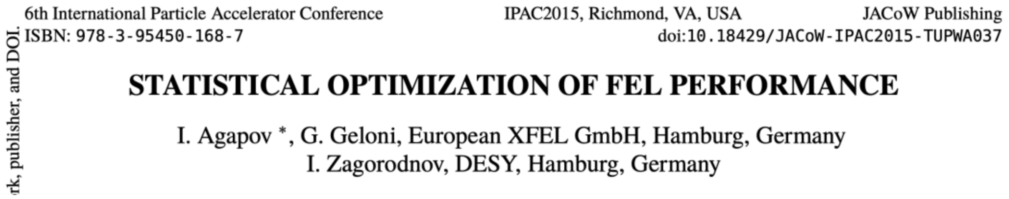
- Historical remarks
- On-line optimization
- Deep learning
- Computer vision
- ML/AI engineering
- Outlook

Historical remarks

- My personal experience with ML applications dates back to 2005 (HERA)
- Applications included beam dynamics modelling and background data prediction (now known as “generative models” or “surrogate models”)
- Difficulties along the way:
  - Had to program NN libraries and backpropagation (tensorflow or pytorch not available)
  - Computers were slow
  - Data was extremely sparse (in hindsight)
  - Neural Networks (aka deep learning) was not widely believed to be a suitable ML tool (Bayesian, decision trees, various regressions being state of the art in applications)
  - No interest in ML from management (at best)
- Success was extremely limited and did not go beyond proof of principle for toy problems, which in hindsight is no surprise, since proper solutions for complex facilities (such as colliders) mostly do not exist up to now even with modern tools

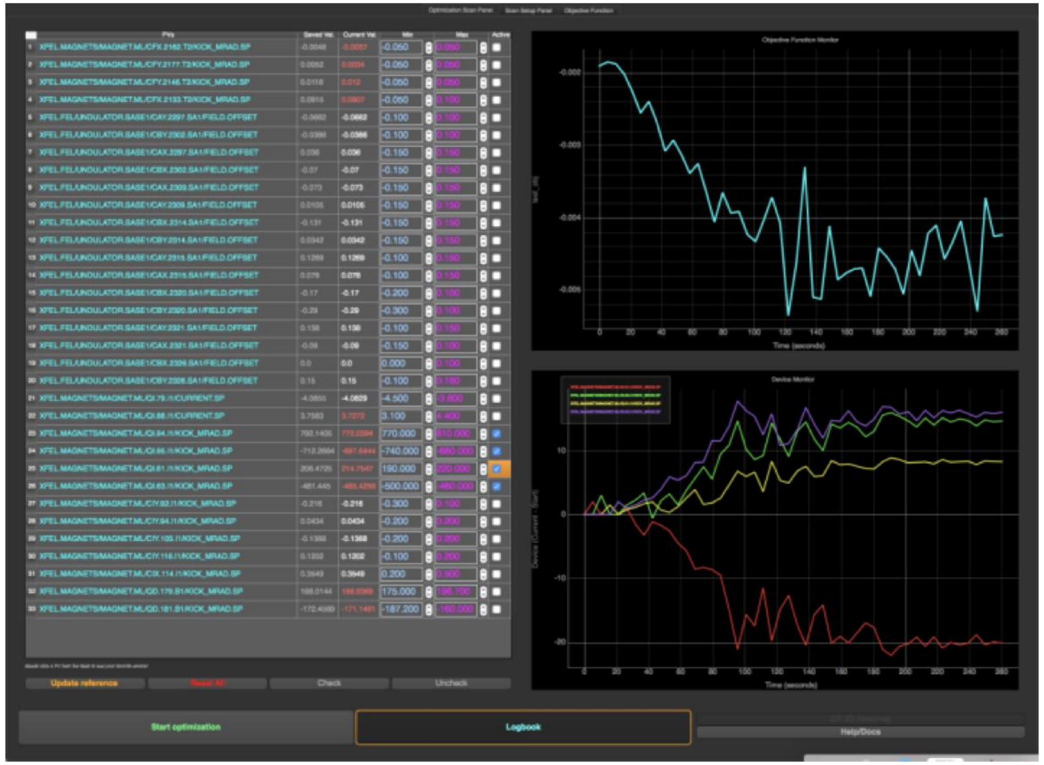
Optimizers: a decade of FEL optimization

- SASE tuning at FLASH required many hours a week
- First successful demonstration in 2015 for FEL SASE optimization
- Deployed at XFEL.EU and widely used in commissioning and operation
- Developed in collaboration with SLAC, changed to BADGER to distinguish from the simulation code


6th International Particle Accelerator Conference IPAC2015, Richmond, VA, USA JACoW Publishing
ISBN: 978-3-95450-168-7 doi:10.18429/JACoW-IPAC2015-TUPWA037

STATISTICAL OPTIMIZATION OF FEL PERFORMANCE

I. Agapov *, G. Geloni, European XFEL GmbH, Hamburg, Germany
I. Zagorodnov, DESY, Hamburg, Germany




OCELOT: A software framework for synchrotron light source and FEL studies

I. Agapov [a], G. Geloni [a], S. Tomin [b], I. Zagorodnov [c]

Show more

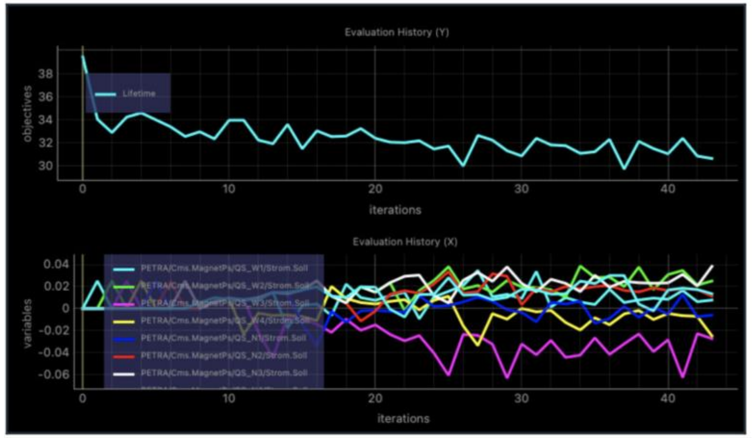
+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.nima.2014.09.057 Get rights and content

Optimizers at PETRA

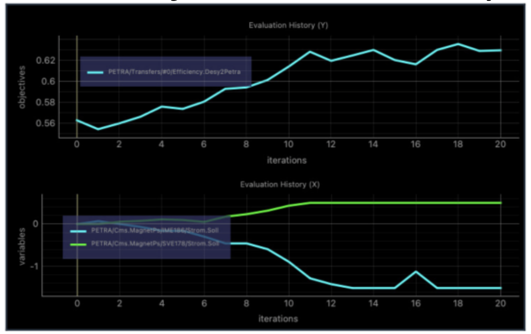
- Several optimizations were tested with OCELOT/BADGER at PETRA, but not used much so far. MATLAB RCDS is also available.
- At EBS more sextupole families allow for efficient lifetime optimization, see talk of Simone Liuzzo

BADGER coupling 2023

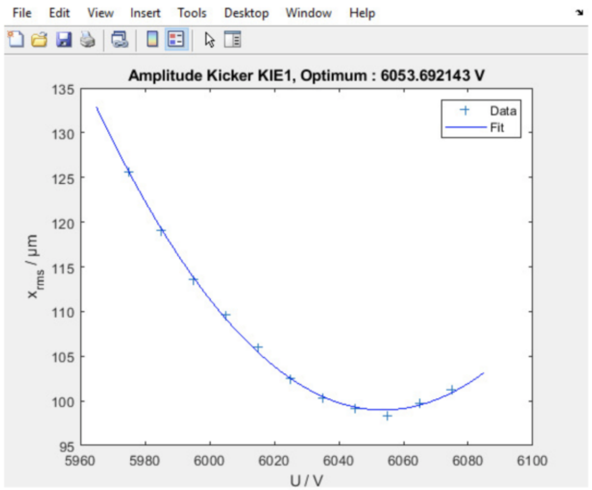


OCELOT tune match 2018

BADGER injection efficiency 2023



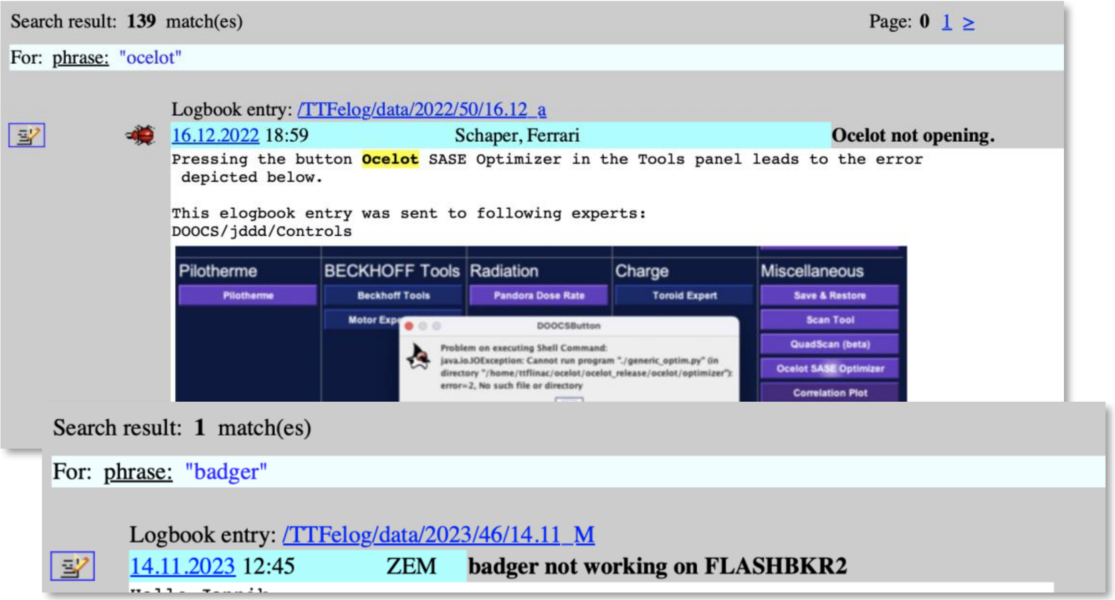
RCDS kicker bump 2023



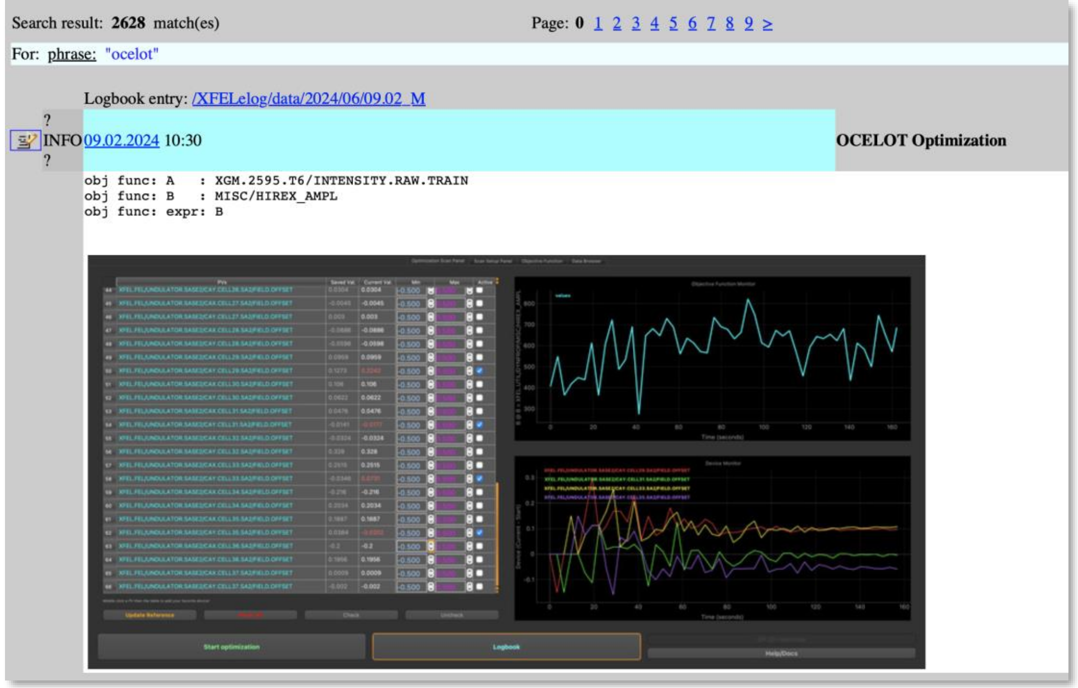
Experience with optimizers

- Many factors (amount of maintenance, personal preferences, general operations planning, time-saving potential) play a role in how widely such tools are used. Software usability and UI often come before algorithm

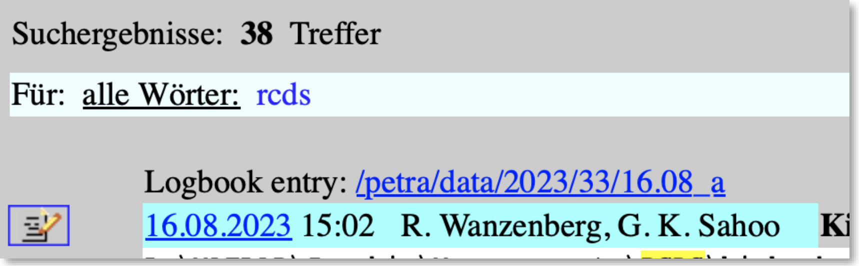
FLASH: something went wrong and not yet fixed



XFEL: constantly used and maintained

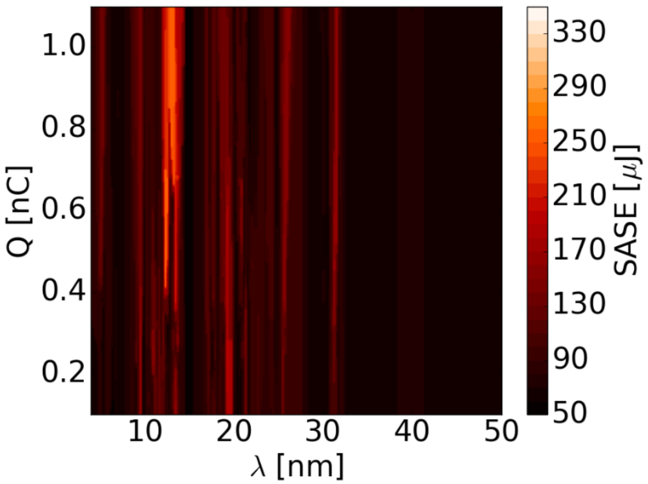


PETRA: RCDS most used, but not too often



Optimizers - ML features

- GP Optimizers are now commonly available (e.g. through standard python libraries) and used in simulations and on-line optimization. They usually require dedicated setup to work efficiently and offer speedup in optimization rather than enhanced capabilities.
- We explored several classical ML/data analysis tools (clustering, regressions etc.) to build models and explore data from early FLASH optimization. Practical applications were (and still are) unclear. Data quality and data processing was non-trivial.



Regression model of FLASH radiation power, arXiv:1704.02335v1

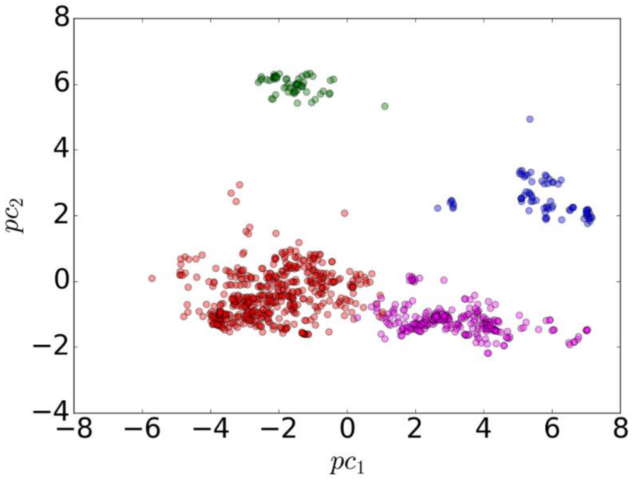


FIG. 13. Clustering of principle components of electron orbit data. 4 clusters are picrured in different colours.

PCA analysis of FLASH orbit data, unpublished (2016)

…at the same time AI revolution has happened

- “Deep Learning” 2015
- DeepMind founded 2010
- OpenAI founded 2015
- AlphaGo vs Lee Sedol 2016
- “Attention is all you need” 2017
- GPT 1 2017
- ChatGPT 2022

nature

Explore content | About the journal | Publish with us | Subscribe

nature > review articles > article

Review Article | Published: 27 May 2015

Deep learning

Yann LeCun, Yoshua Bengio & Geoffrey Hinton

Nature **521**, 436–444 (2015) | Cite this article

917k Accesses | **37k** Citations | **1270** Altmetric | Metrics

Riding the AI wave

- Increased understanding of potential of AI for accelerators
- White paper in 2018
- ICFA ML Workshops since 2018
- National platforms (e.g Helmholtz.AI since 2020)
- At DESY dedicated funding to work on AI from 2019

https://visitlofoten.com/en/activity/surfing/

Physics > Accelerator Physics

[Submitted on 7 Nov 2018]

Opportunities in Machine Learning for Particle Accelerators

Auralee Edelen, Christopher Mayes, Daniel Bowring, Daniel Ratner, Andreas Adelmann, Rasmus Ischebeck, Jochem Snuverink, Ilya Agapov, Raimund Kammering, Jonathan Edelen, Ivan Bazarov, Gianluca Valentino, Jorg Wenninger

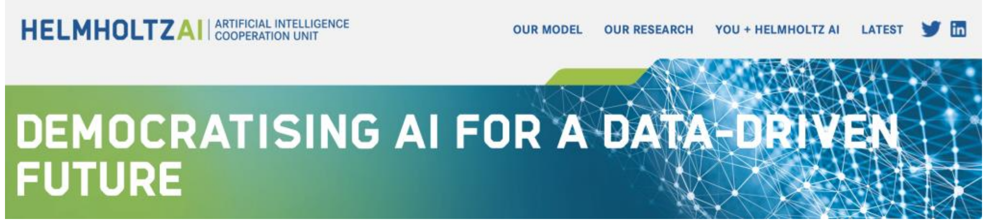
MaLAPA Conferences

4TH MACHINE LEARNING APPLICATIONS FOR PARTICLE ACCELERATORS (2024), GYEONGJU, SOUTH KOREA. HOSTED BY PAL

3RD ICFA BEAM DYNAMICS MINI-WORKSHOP ON MACHINE LEARNING APPLICATIONS FOR PARTICLE ACCELERATORS (2022), CHICAGO, USA. HOSTED BY BNL.

2ND ICFA MINI-WORKSHOP ON MACHINE LEARNING FOR CHARGED PARTICLE ACCELERATORS (2019), VILLIGEN, SWITZERLAND. HOSTED BY PSI.

1ST MACHINE LEARNING APPLICATIONS FOR PARTICLE ACCELERATORS (2018), MENLO PARK, USA. HOSTED BY SLAC.



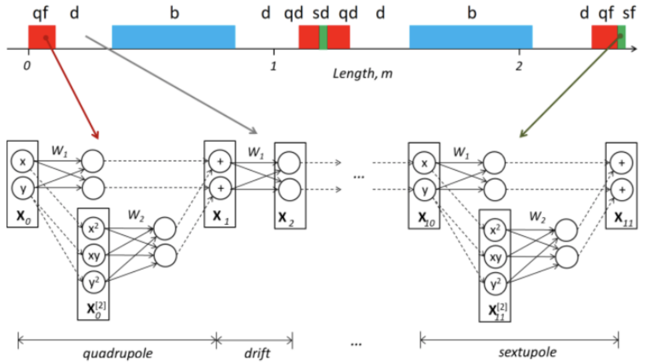
Landscape@DESY

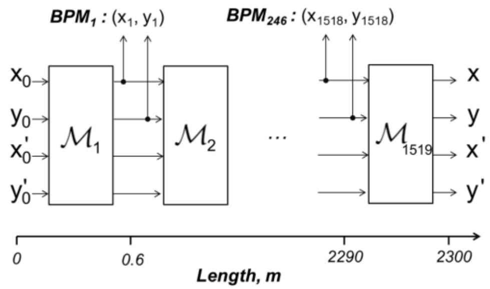
- Goals are broadly set to automation (accelerators, experiments), improved data analysis (experiments)
- More specific problems
  - Predictive maintenance
  - Optimizers
  - Control algorithms
  - Reinforcement learning
  - Robotics
  - Data processing
  - Triggers
  - Generative models for simulations
- The majority of ML developments at DESY are in particle physics/photon science
- In accelerator division ML is mostly developed at ARES (test accelerator) and XFEL
- Some examples explored at PETRA machine are shown next

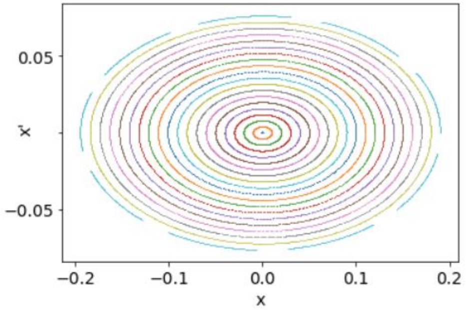
A. Ivanov

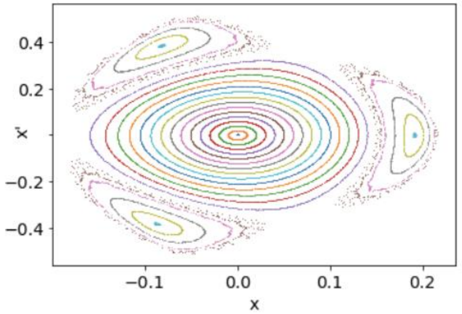
Deep Learning explored - Physics Based NN

- Observation: computational graphs in tensor manipulation codes developed for deep learning can be used to compute accelerator maps (Taylor maps)
- Accelerator model can be implemented as a deep neural network
- Weights have direct correspondence to magnet strengths
- Exact beam dynamics calculations are possible
- The calculations are differentiable and training algorithms can be used for model fitting
- Preserving symplecticity at training is solvable by regularization



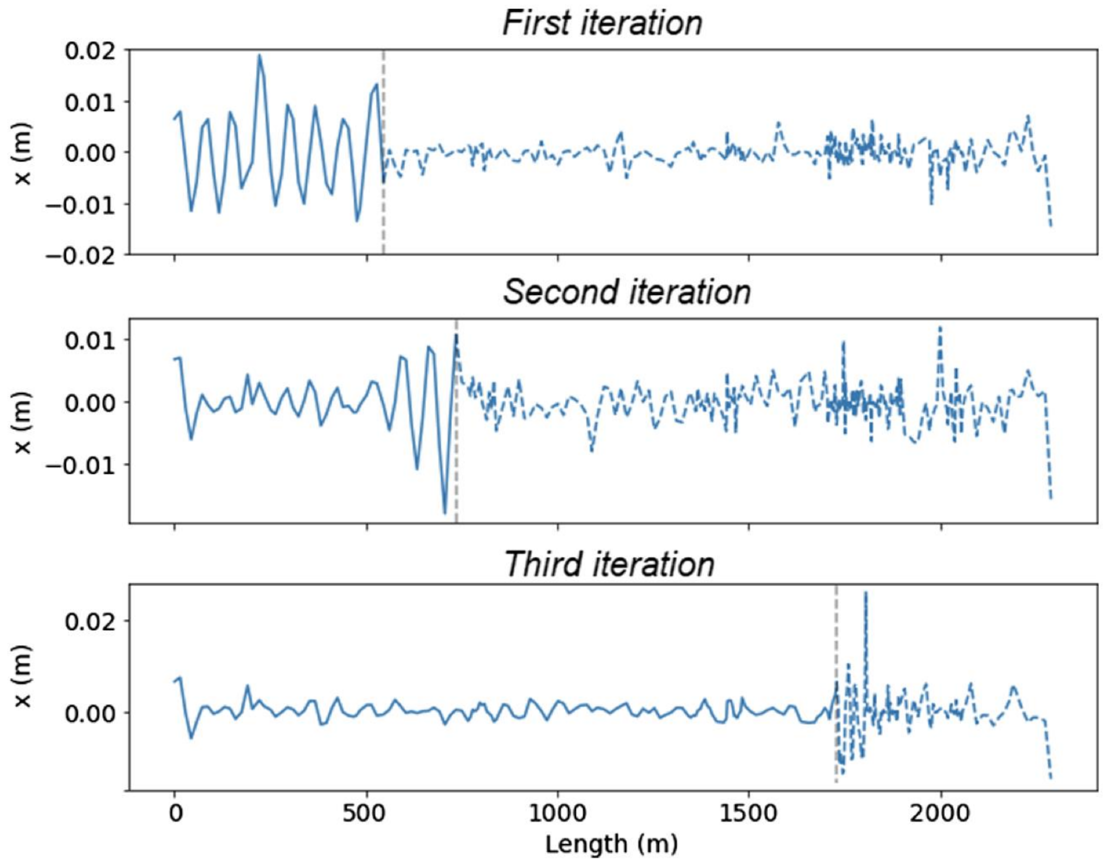






Physics Based NN applications

- Developed as an add-on to ocelot
- Benchmarked with elegant on PETRA III and IV lattices
- Tested first-turn steering based on th code with one-shot learning at PETRA III
- Developed a physics-informed RL engine for a test problem
- The approach of using tensor manipulation engines (tensorflow, pytorch) has been adopted for several (student) beam optics code projects



PHYSICAL REVIEW ACCELERATORS AND BEAMS **23,** 074601 (2020)

Physics-based deep neural networks for beam dynamics in charged particle accelerators

Andrei Ivanov* and Ilya Agapov
Deutsches Elektronen Synchrotron DESY, Notkestrasse 85, 22607, Hamburg, Germany

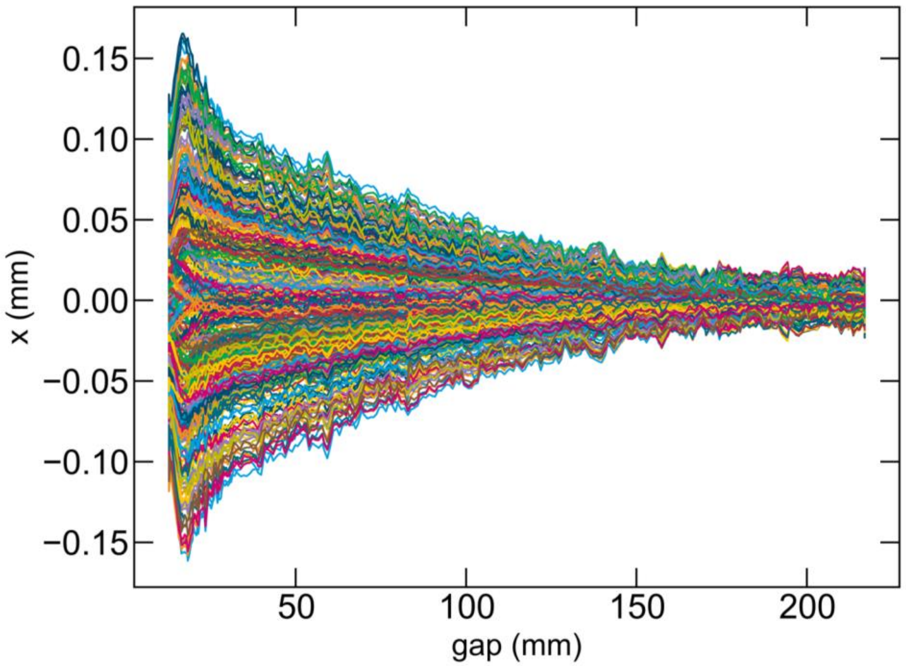
Physics Based NN - outlook

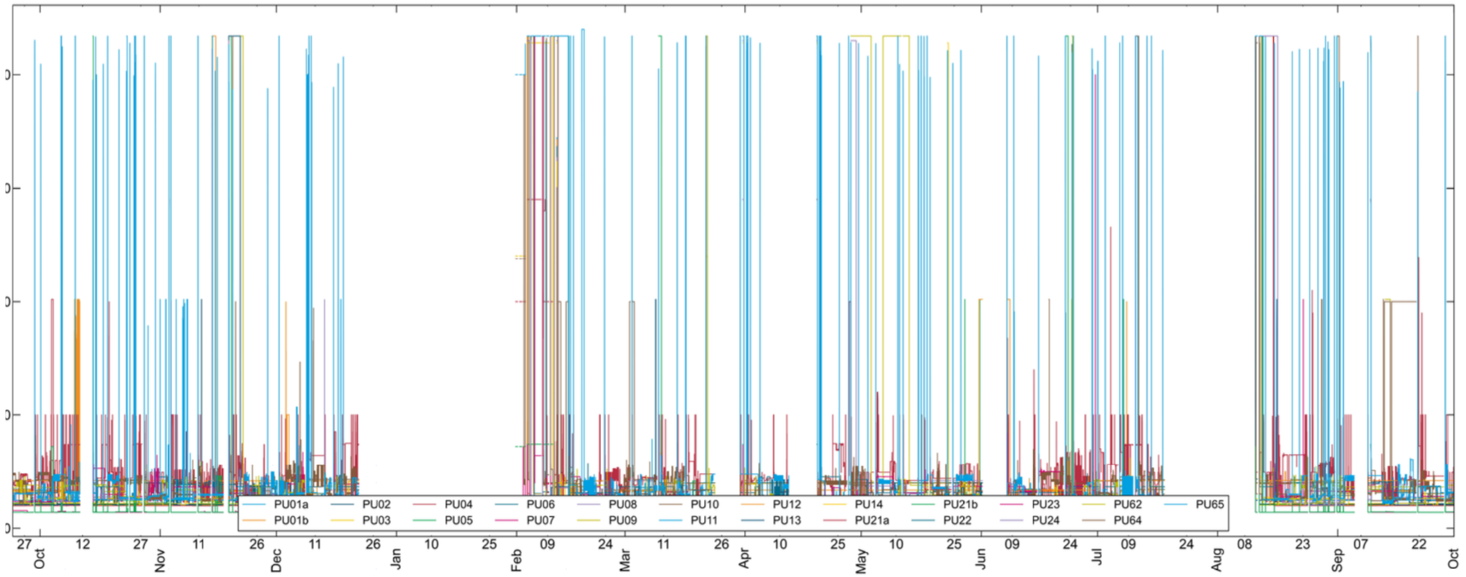
- Training requires substantial fine-tuning as with all NN models
- Accounting for higher nonlinearity is problematic. There is little speed benefit if octupole and higher order maps are used directly, and training stability is unclear. Map truncation can be done, but leads to the ancient problem of symplectification.
- Extending to account for self-consistent collective effects is non-trivial
- Practical application might include linear optics matching, and real-time simulations for hardware tests such as of the fast orbit feedback if implemented on FPGA

B. Veglia

Deep Learning explored - NN-based control

- Moving undulator gaps creates orbit distortions, that are compensated by correctors, settings are from feed-forward tables (and additionally orbit feedback)
- The gap dependency is nonlinear, and can be approximated by a NN

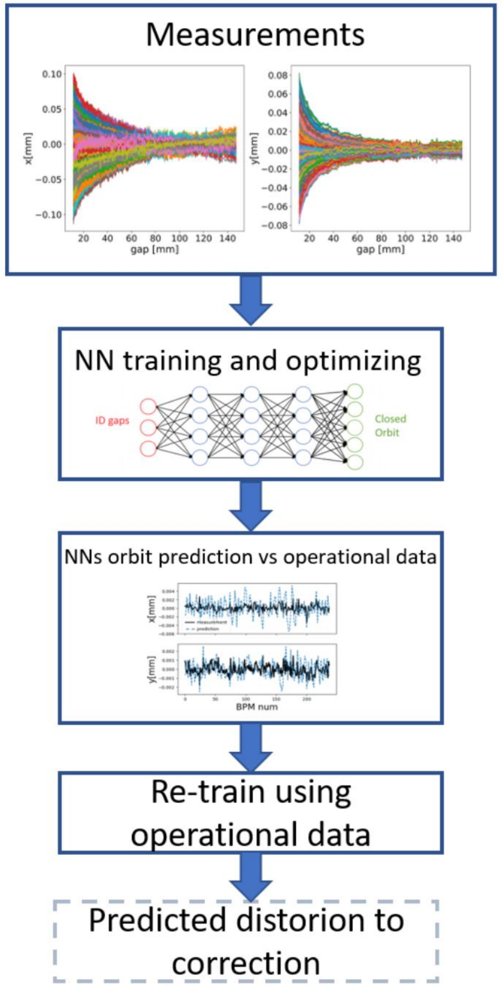




B. Veglia

Deep Learning explored - NN-based control

- Several NN architectures trained on measured data
- Model can be stably re-trained in operation (albeit not in user run)
- Predictions are accurate and can be used in a control loop



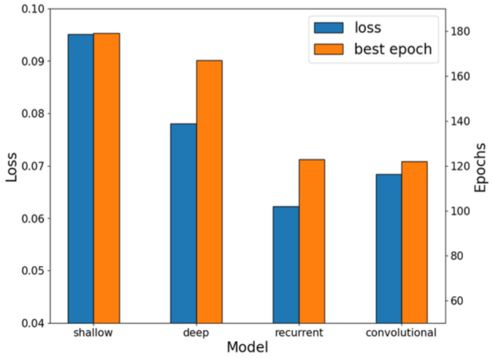
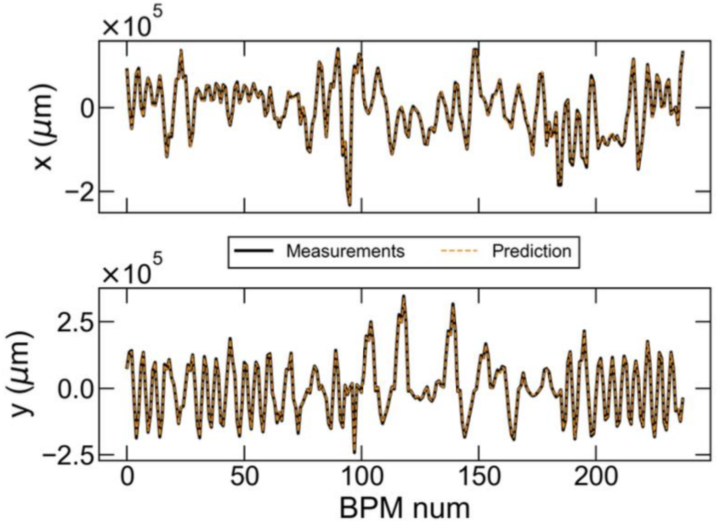


FIG. 11: Comparison of the performance of the four different model architectures.



ISBN: 978-3-95450-231-8 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2023-WEPA095

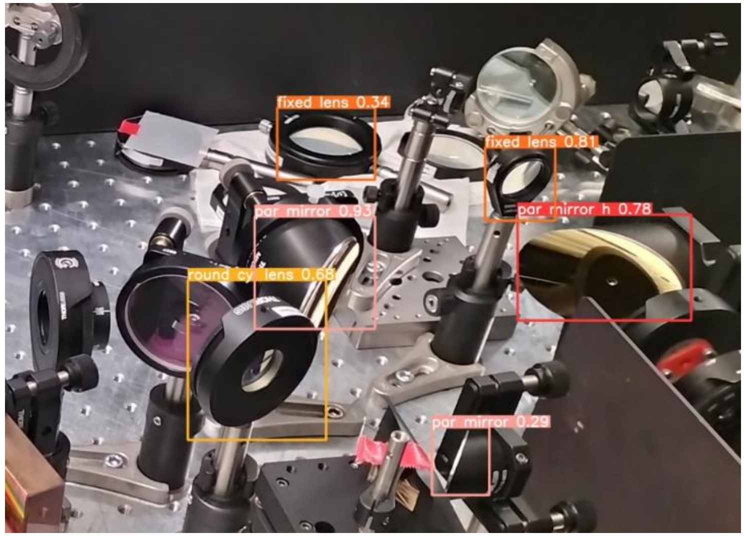
NEURAL NETWORKS FOR ID GAP ORBIT DISTORTION COMPENSATION IN PETRA III*

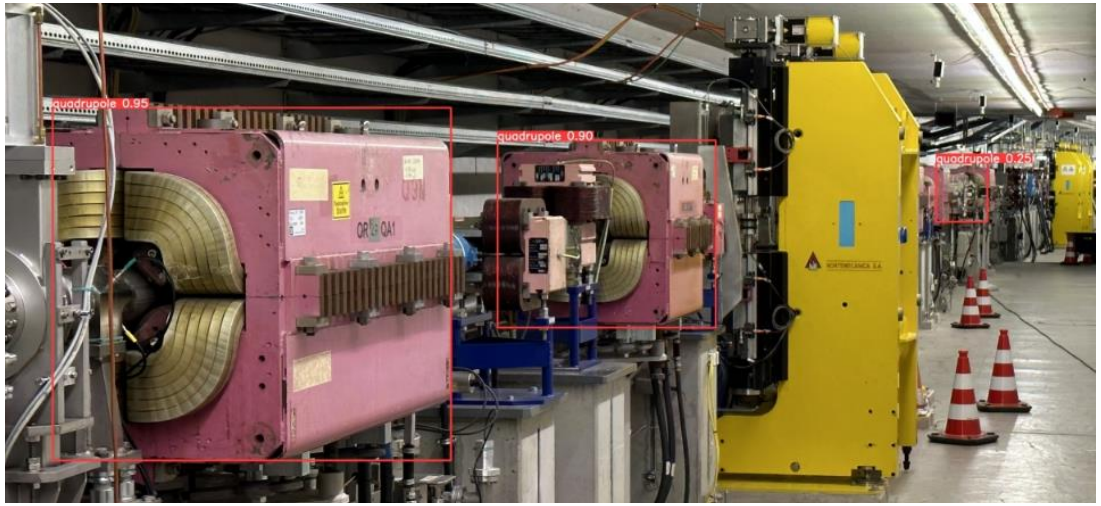
B. Veglia†, I. Agapov, J. Keil, Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany

B. Veglia

Deep learning explored - computer vision

- Trained DNNs to perform identification of several types of equipment in visible and IR
- Possible first step towards autonomous navigation in a tunnel or laboratory
- To be presented at the ICFA ML workshop in 3 weeks



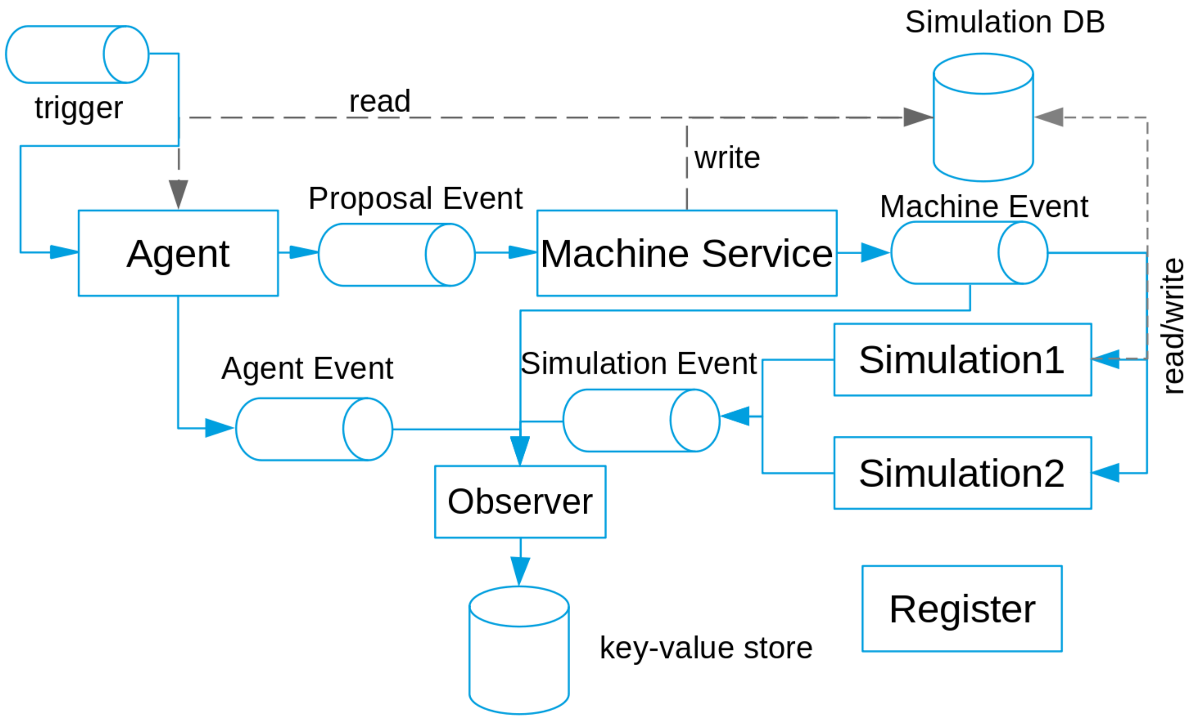


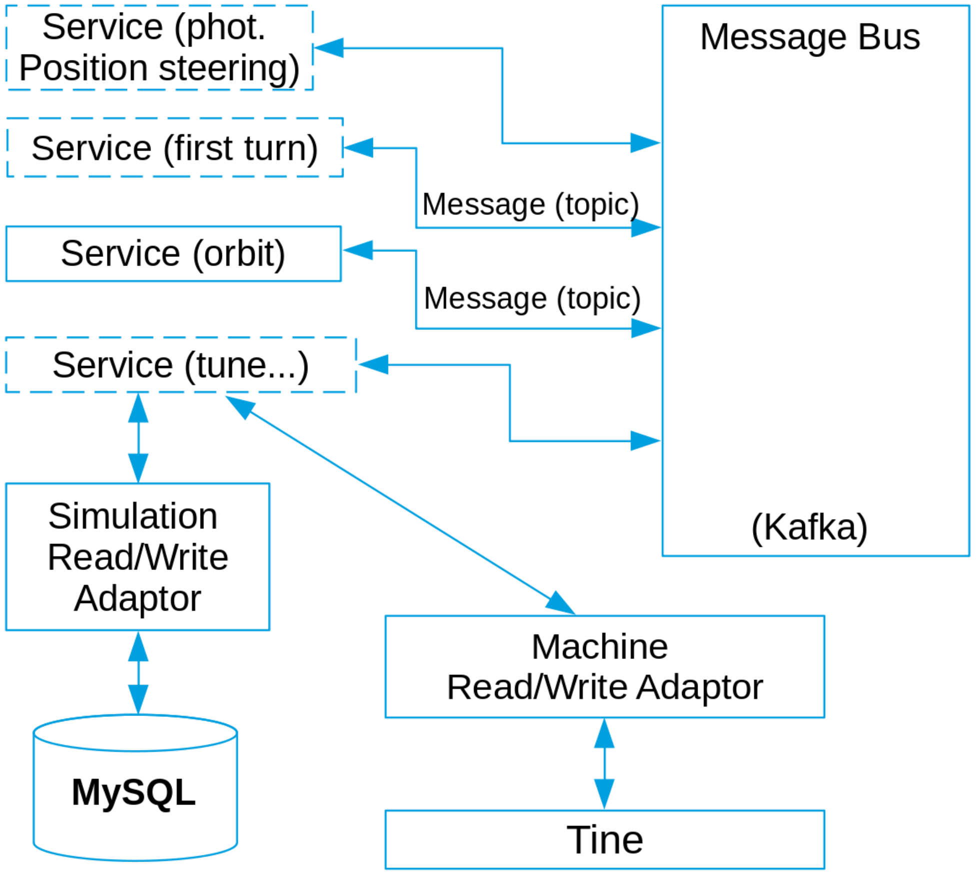
ML engineering - concept

- Can we create a system that is capable of decision-making and controlling a facility?
- Created software infrastructure that allows AI agents/services to communicate
- Following features were foreseen:
  - Protocol of communication between various agents over a network
  - Execution of control sequences by certain agents
  - Triggering of AI models to retrain on HPC cluster
  - Operation in simulation mode
  - Mimicking a communication procedure between human operators

M. Boese

ML engineering implementation - too complex






```
service_objs = pipe.get_service_objects()
services = {service.name: service for service in service_objs}
orbit_obs = services['orbit_corr_observer']
orbit_ctrl = services['orbit_corr_controller']
orbit_agent = services['orbit_corr_agent']
set_cors_service = services['cors_sim_ctrl']
```

```
config_data={'n_sv_x':20, 'n_sv_y': 10, 'x_factor':-1000.0, 'y_factor':0.0 }
orbit_agent.reconfig(config_data)

config_data={'exclude_cor': ['PKPDA', 'PKPDD']}
for service in [orbit_agent, orbit_ctrl, orbit_obs]:
    service.reconfig(config_data)
```

```
orbit_agent.trigger()
```


13th Int. Particle Acc. Conf. IPAC2022, Bangkok, Thailand JACoW Publishing
ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOMS016

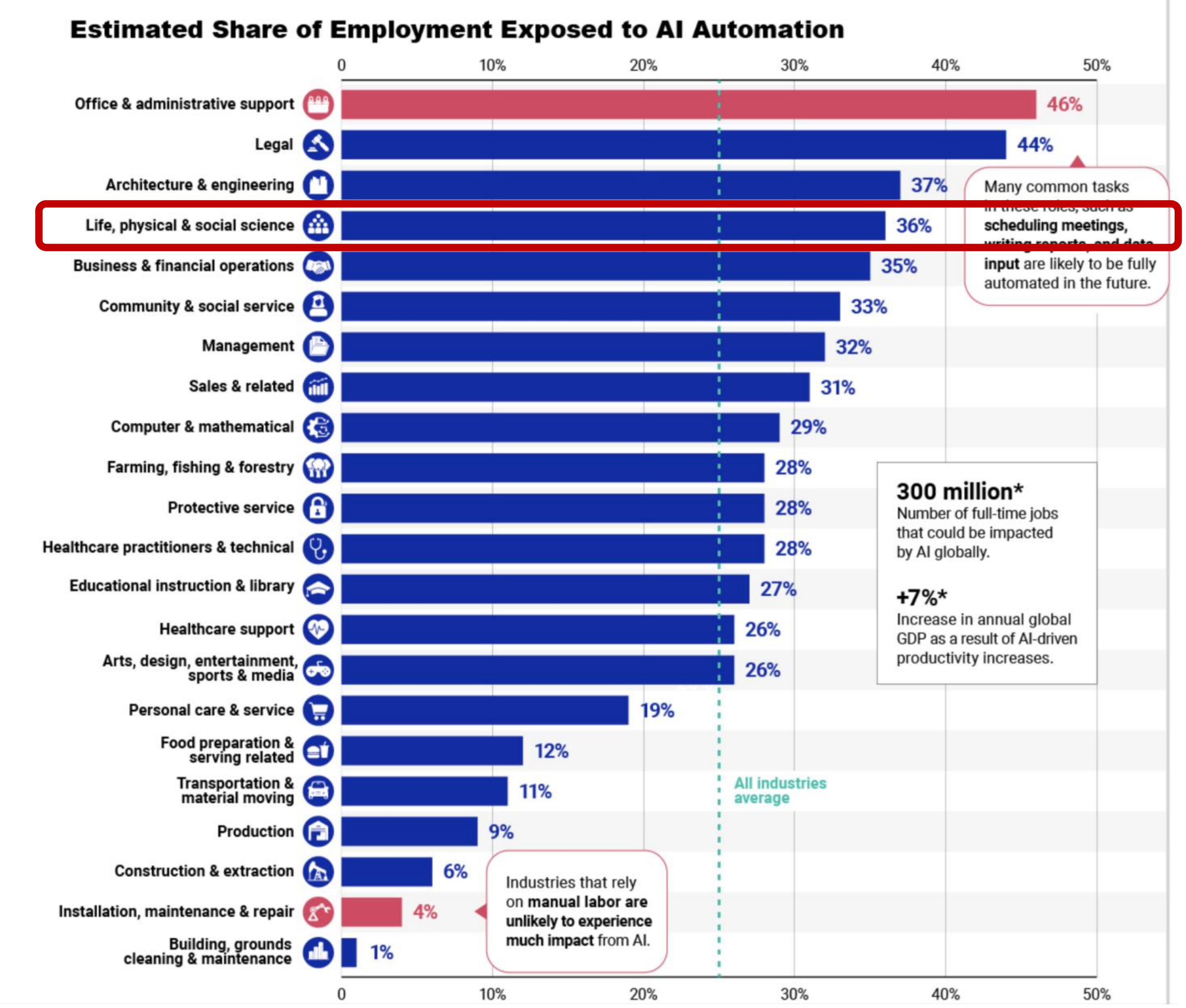
A PIPELINE FOR ORCHESTRATING MACHINE LEARNING AND CONTROLS APPLICATIONS

I. Agapov*, M. Böse, L. Malina
Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

ML engineering lessons learned and challenges

- Even with very reduced functionality a complex software stack was emerging
- The software stack would have been unmaintainable in operation with current workforce
- The effort to bring an ML “jupyter notebook” research project to production has been becoming clear
- Typical number of employees in AI-centric industries:
  - DeepMind: 2000
  - OpenAI: 770
  - TESLA Autopilot: > 1000
- Typical AI effort in accelerator laboratories: 2-5 per laboratory (1 for PETRA), mostly funded via grants with research objectives. With that, goals such as “autonomous facility” are unrealistic.
- To reach future operation goals, we reverted to gradual improvement of legacy tools (“taskomat”, “save/restore”, “startup script”, etc.)

Is accelerator physics/operation easily "automatable"?

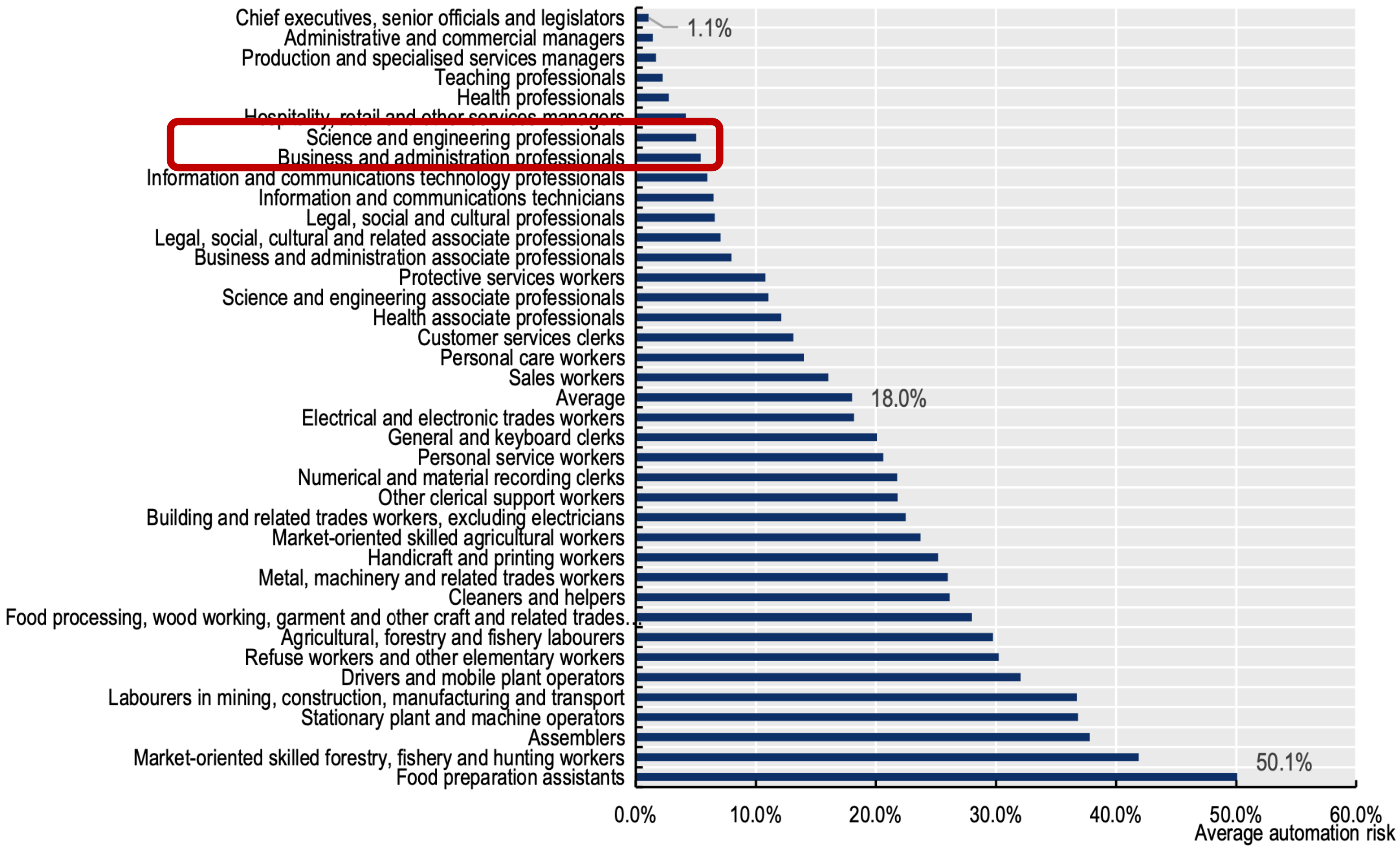


Stop to reflect:
maybe the activity is inherently too complex to automate completely

https://writerupdated.com/2023/06/28/industries-highest-potential-for-ai-infographic/

Physics has a high potential for automation

Is accelerator physics/operation easily "automatable"?



OECD Future of work:

Physics is not at risk of automation!

What are we to expect?

Way forward?

- We can run all present facilities (even newest ones such as the EBS and APS-U) the old way ("matlab middle layer")
- ML/AI capabilities are nice to have but not essential
- Economic and social impacts of AI in accelerator facilities should be better understood (cost of implementing AI system vs. running cost of a facility is high)
- We are facing increasing danger of a large body of legacy operation and simulation software in danger of being unmaintainable in the future
- Investing in modern computing capabilities is essential
- AI/ML deployment will automatically follow
- Challenge: incremental advance while keeping the effort relatively low and aligning with short-term facility goals
- Side note: ChatGPT became standard tool in research (writing papers, presentation, code, information retrieval). Consequences might be profound

Summary and outlook

- A number of applications of ML to accelerator-related problems have been demonstrated
- Deployment beyond proof-of-principle has been difficult: there are no running ML/AI applications at PETRA III and no ML techniques have been used in PETRA IV design
- There are no plans as of now to integrate AI into PETRA IV operation
- AI techniques are now commonplace in consumer products
- We need to bring state of the art software engineering to SR facilities as a first step towards successful AI deployment
- But economics of that is challenging
- Most promising avenue seems to be aligning with **large infrastructure projects such as the FCC**
- Need to explore ways to incrementally incorporate advancing software technologies - e.g. via the **pyML collaboration**