CERN openlab

AI model registry on cloud

A new Oracle project during Phase VII

ORACLE

Sofia Vallecorsa, Renato Cardoso and the CERN IT-DB team

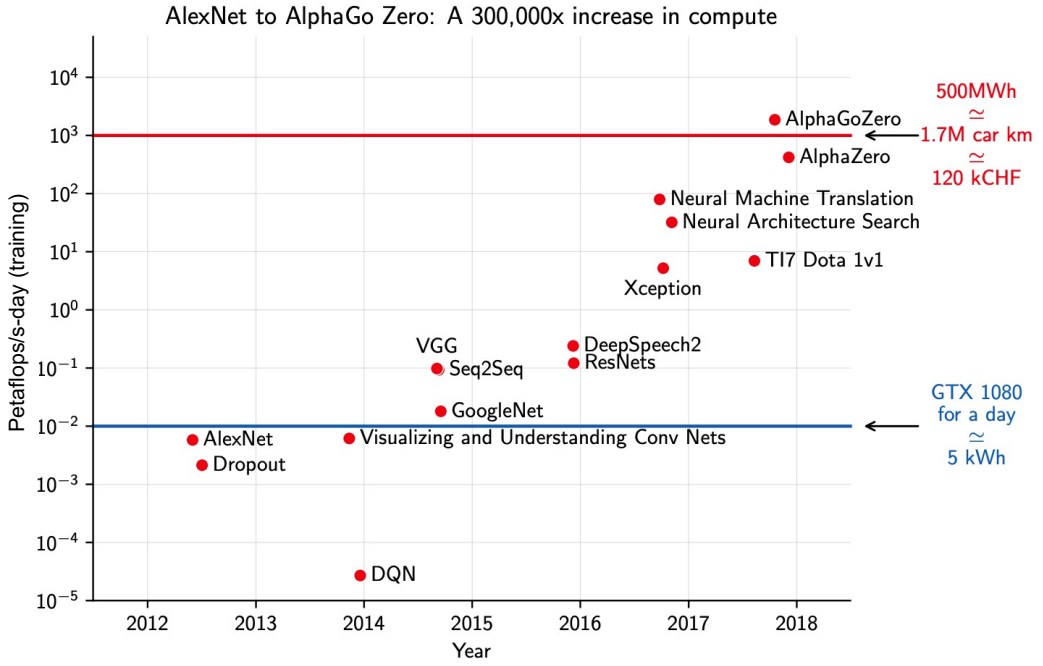
Openlab Technical Workshop - 21/03/2022

DL Applications in HEP

Classical **Machine Learning** used for many years, mostly during the final steps of data analysis for signal /background separation

Deep Learning is studied for different applications that have **different requirements**

- Real-time filtering
- Raw data processing
- Optimisation
- Analysis
- Simulation
- Monitoring and Control Systems



(Radford, 2018)

"The biggest lesson that can be read from 70 years of AI research is that **general methods that leverage computation** are ultimately the most effective, and by a large margin" (Richard Sutton, 2019)

Enabling Deep Learning

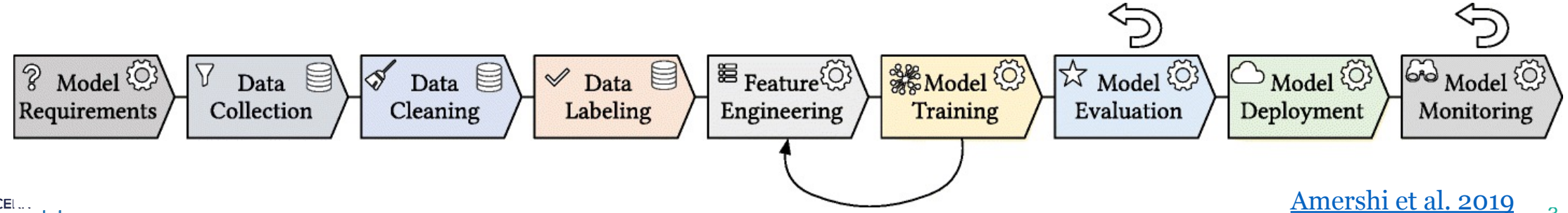
Efficient and **productive** R&D goes hand-in-hand with **smart infrastructure design**.

Developing AI systems is **different** than typical software engineering. In particular:

- **data** discovery, management, and versioning are **more complex**;
- **modular design is not trivial** since AI components exhibit **complex entanglement**.

However **integration** into “regular” software and services is critical

- Interest in **AI lifecycles** since the past 10 years - inspired by data mining standards (e.g. CRISP-DM, TDSP, etc..)



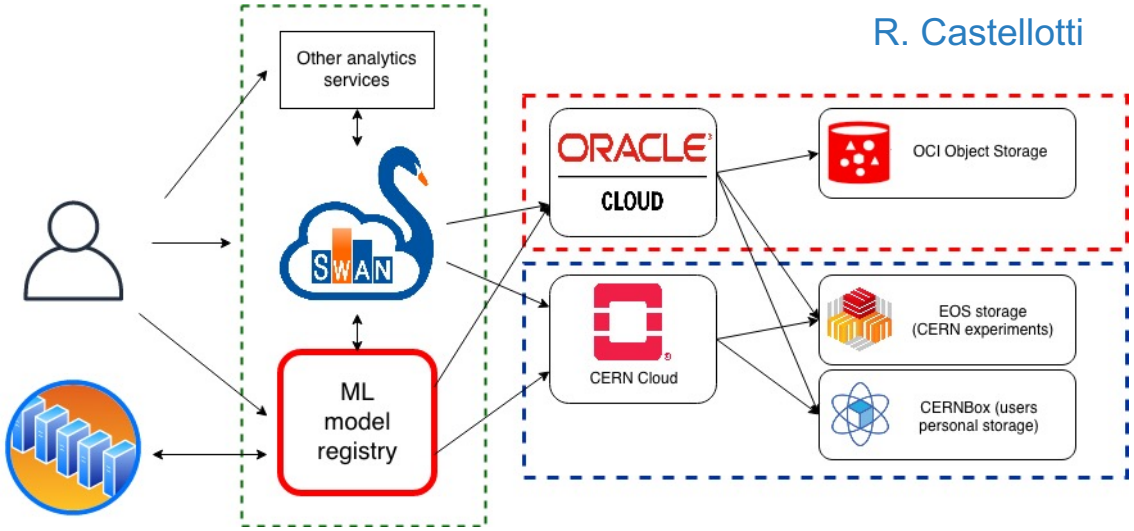
Amershi et al. 2019

AI models catalog

R. Castellotti

Increasingly popular components in AI ecosystems

- **Help introduce** ML/DL in areas where expertise is limited or still building
- **Accelerate** new model development
- Improve **reproducibility** and **quality monitoring**
- **Simplify** deployment



We propose the creation of a **AI model registry at CERN**:

- Deployment **on CERN cloud and OCI**, following a "hybrid cloud" model
- Test and prototype solutions for production usage of **AI tools and platforms**
- Focus on **HEP- related applications**

Challenges

- Models are mostly implemented **without standard** APIs, input format, or hyper-parameters notation
- Might need **further optimization before deployment** because of runtime constraints (hardware, data feeding lines, latency, etc..)

- What is the scope of a model catalog?
- How much we can **realistically re-use a pretrained model** on different datasets?

Model sharing across different use cases

- How much we can **realistically re-use a pretrained model** on different datasets?

Possibility to directly deploy models out of existing catalogs is limited by the nature of the **input data** and the **model generalization capabilities**.

In HEP directly applying a model trained for analising the output of a specific detector to a different use case is usually **unfeasible**, **inefficient** or **too costly** in terms of the data pre-processing step.

Toward “smart” catalogs

DL model generalization is a lively field of research[1,2]

Advanced training techniques extend the **concept of transfer learning** [3-8]

Evaluate to which extent a model catalog could be used in our field:

- Choose a specific task and a pre-trained DL model and study how its performance changes while varying the nature of the input dataset
- Understand challenges and advantages related to the use of **generalizable models** to High Energy Physics use cases.

- Towards improved "smart" catalogs, containing **adaptative/generalizable models**
- a seed to community-based DL model development at CERN ?

Thanks!

Sofia.Vallecorsa@cern.ch

https://openlab.cern/

References

[1] Kawaguchi, Kenji, Leslie Pack Kaelbling, and Yoshua Bengio. "Generalization in deep learning." arXiv preprint arXiv:1710.05468 (2017).

[2] Russin, Jacob, et al. "Compositional generalization by factorizing alignment and translation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2020.

[3] Shen, Zheyan, et al. "Towards out-of-distribution generalization: A survey." *arXiv preprint arXiv:2108.13624* (2021)

[4] Arjovsky, Martin. *Out of distribution generalization in machine learning*. Diss. New York University, 2020

[5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[6]Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

[7] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. Behavioral and Brain Sciences, pages 1–101, 2016

[8] Sebastian Thrun and Lorien Pratt. Learning to learn. Springer Science & Business Media, 2012