Introduction to MLOps with ClearML

PROF. DR. – ING. MORRIS RIEDEL, UNIVERSITY OF ICELAND & JUELICH SUPERCOMPUTING CENTRE (GERMANY)

30TH SEPTEMBER, RAISE EC PROJECT SEMINAR SEPTEMBER, ONLINE

@ProfDrMorrisRiedel @Morris Riedel @MorrisRiedel @MorrisRiedel https://www.youtube.com/channel/UCWC4VKHmL4NZgFfKoHtANKg

IHPC National Competence Center for HPC & AI in Iceland

EuroHPC Joint Undertaking

EOSC NORDIC

RAISE Center of Excellence

ADMIRE malleable data solutions for HPC

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

EURO

HELMHOLTZAI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

DEEP Projects

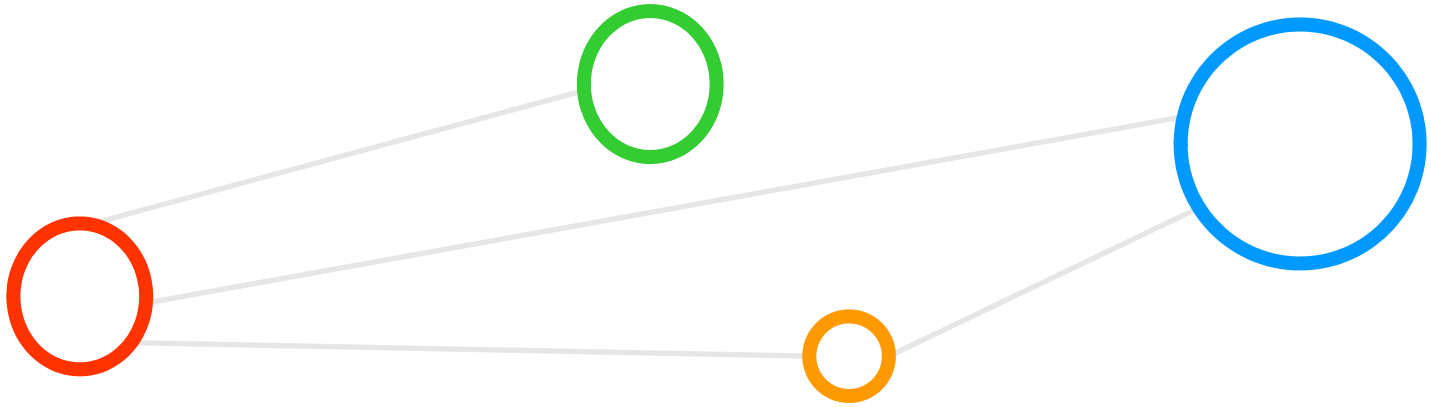
JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Outline

- Selected Foundations
  - Modus Operandi in Data Sciences & Limits
  - Understanding the complexity of AI Lifecycles
  - ‘Chaotic‘ Practice in Machine Learning vs. MLOps Approach
  - Understanding growing numbers of Machine Learning Models
  - Differences and Commonalities to DevOps & MLFlow Example
- MLOps with ClearML
  - Allegro AI Legacy & ClearML
  - ClearML Lean-Stack MLOps Stack Design Approach
  - Selected Clear ML Products & ClearML Experiment Example
  - RAISE Unique AI Framework & ClearML
- Selected References
- Acknowledgements

[1] Icelandic HPC Community Web page

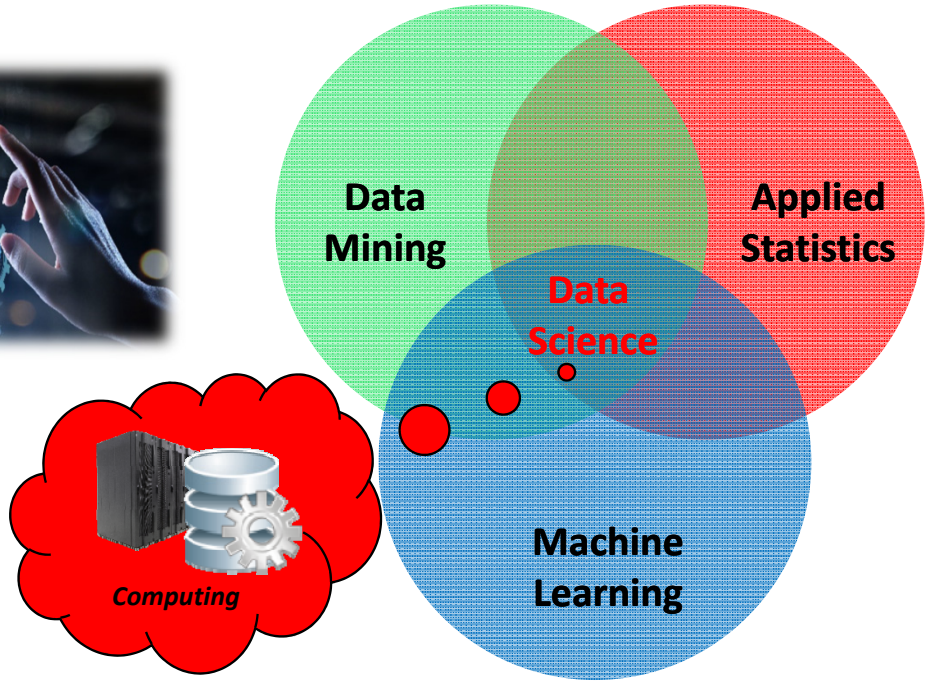
Selected Foundations



Machine Learning Prerequisites & Computing Challenges

1. Some pattern exists
2. No exact mathematical formula
3. Data exists

- Idea ‘Learning from Big Data‘
  - Shared with a wide variety of other disciplines
  - E.g. signal processing, big data data mining, etc.
- Challenges
  - Data is often complex
  - Requires ‘Big Data analytics‘
  - Learning from data requires processing time → Clouds or High Performance Compuing



- **Machine learning is a very broad subject and goes from very abstract theory to extreme practice (‘rules of thumb’)**
- **Training machine learning models needs processing time (clouds or high performance computing)**
- **While data analysis is more describing the process of analysin the data, the term data analytics also includes and the necessary scalable or parallel infrastructure to perform analysis of ‘big data’**

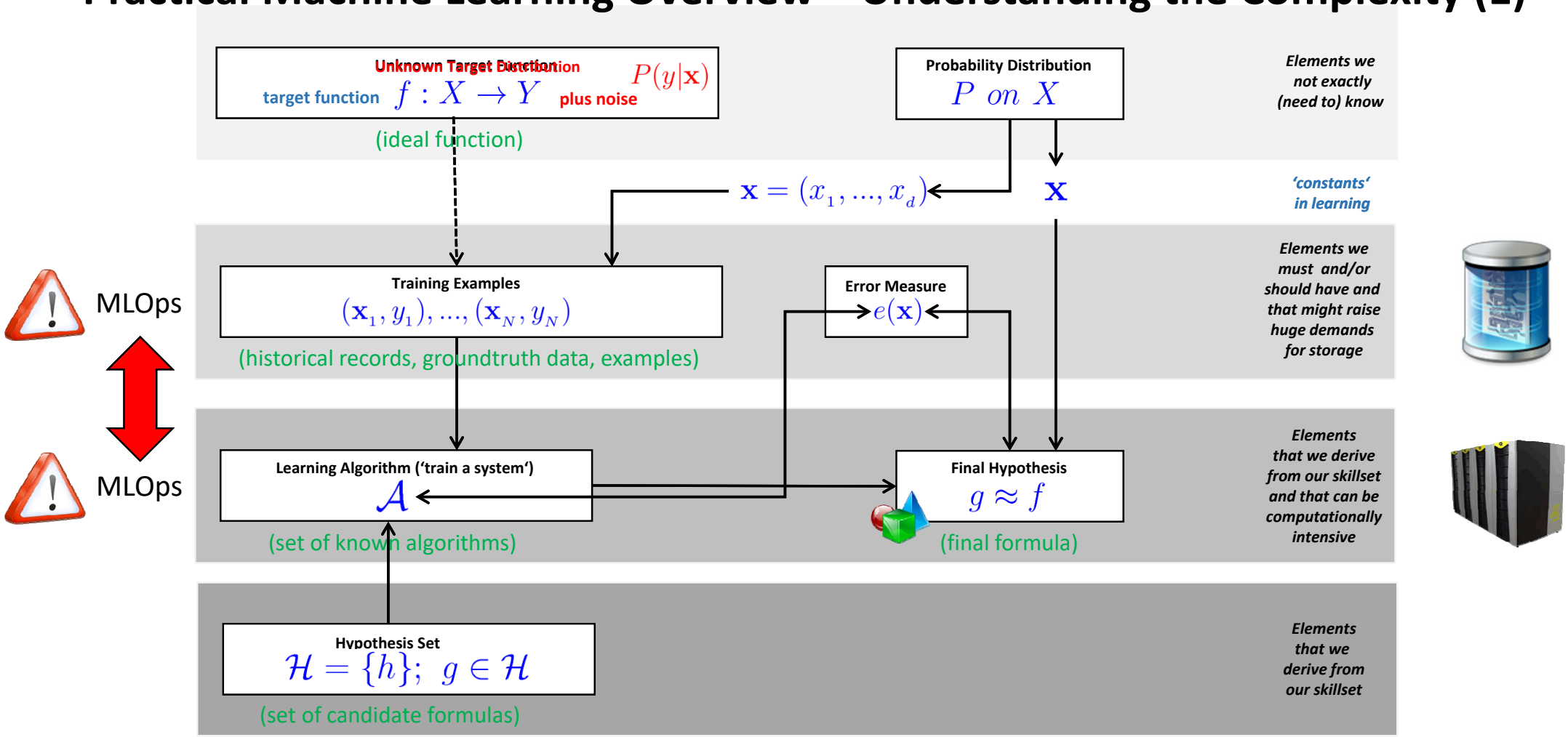
‘Modus Operandi‘ and Limits in Data Sciences

- ‘Data exists‘ addressed
  - Performing real data science is challenging
  - Many activities to perform data cleaning, data understanding, etc.
- Challenge
  - Traditional DevOps for machine/deep learning & data sciences does not work
  - ‘Not one clear application to deploy and operate in increasing versioning‘
- Common ‘Iterative Modus Operandi‘ in Data Science
  - Many different approaches for modeling & very iterative model design
  - Fast paced & solution-oriented (not just for the sake to find all solutions)
  - Different environments: laptop, small-scale HPC, and large-scale HPC
  - Too many experiments with models and too many parameters to choose from (i.e., community drive to always find the best model no matter what!)
  - Challenge on how environment moves ‘from experiments to deployments‘
  - E.g., the CRISP-DM model different steps & iterative nature never beeing done

[3] CRISP-DM Refrence Model

[2] MLOps

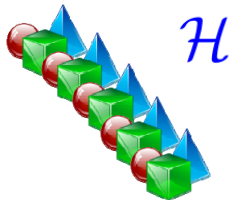
Practical Machine Learning Overview – Understanding the Complexity (1)



Practical Machine Learning Overview – Understanding the Complexity (2)

Hypothesis Set

$$\mathcal{H} = \{h\};\ g \in \mathcal{H}$$



$$\mathcal{H} = \{h_1, \ldots, h_m\};$$

(all candidate functions derived from models and their parameters)

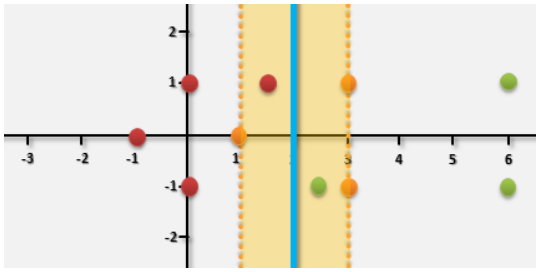
- **Choosing from various model approaches h_1, …, h_m is a different hypothesis**
- **Additionally a change in model parameters of h_1, …, h_m means a different hypothesis too**

'select one function' that best approximates

Final Hypothesis

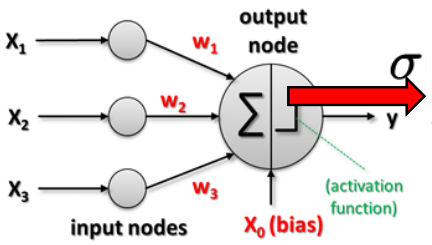
$$g \approx f$$

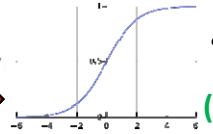
h_1



(e.g. support vector machine model)

h_2





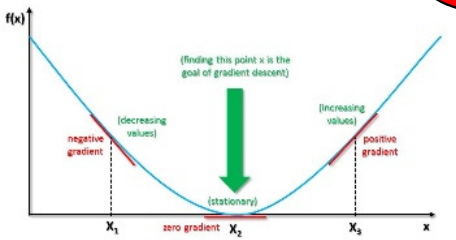
$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

(model function)

```
1 / (1 + exp(-w^T x))
```

$$h(\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x})$$

(e.g. logistic regression model)



Error Measure

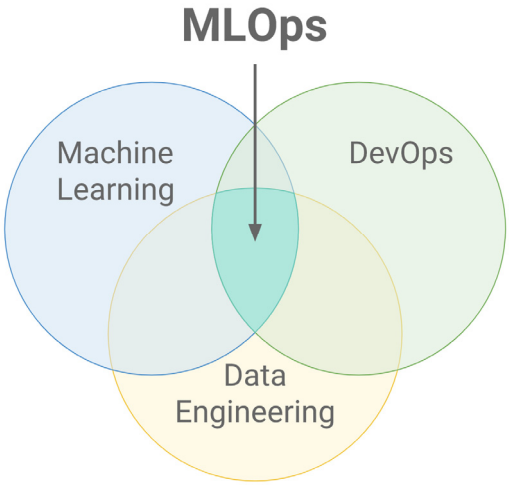
$$e(\mathbf{x})$$

```
log(1 + exp(-yw^T x))
```

(loss function, logistic loss)

MLOps Approach to support the whole AI Lifecycle

- Selected Facts
  - Work against ‘de-facto chaos and ad-hoc approaches in data sciences‘
  - Includes automation of machine/deep learning pipelines
  - Enables orchestration of all the tools needed for the pipelines in a good order
  - Offers reproducability of the solution to be used in other use cases & same results
  - Being realistic to overcome technical challenges to deploy solutions
- Overlaps with DevOps & Data Engineering
  - DevOps: Delivery of high quality machine/deep learning software pipelines
  - Data Engineering: software engineering design for data-intensive workfloads

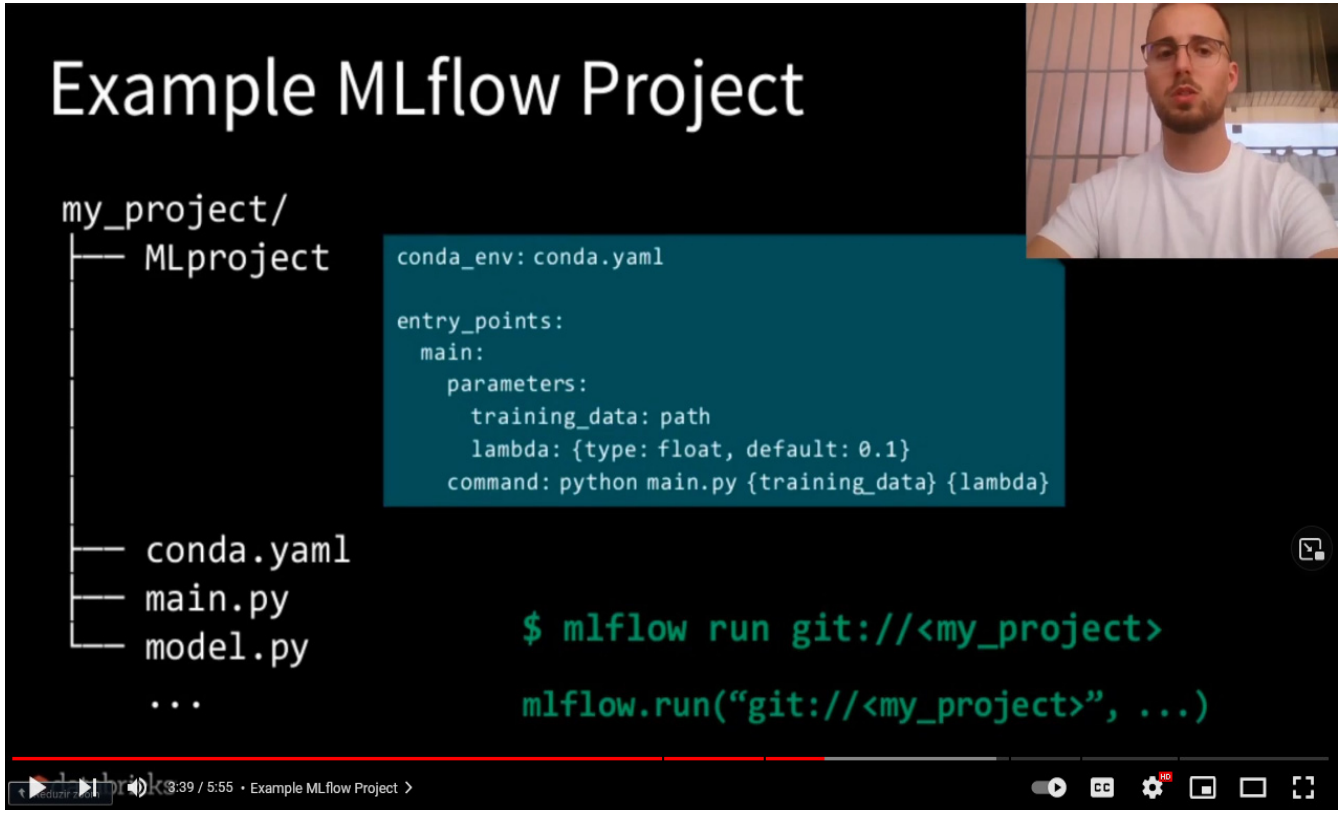


- **DevOps is a set of pratices that combines software development (aka ‘dev’) with IT operations (aka ‘ops’) to shorten the system development life cycle and to provide continuous delivery of high software quality**
- **Data Engineering is a set of methods to gather & prepare the data for machine/deep learning models**
- **MLOps is a set of practices that aims to deploy & maintain machine/deep learning models in production reliably → When an algorithm is ready to be launched, MLOps is practiced between data scientists, DevOps, and machine learning engineers to transition the algorithm to production system deployments**
- **Production system means to move from local laptop or prototype HPC systems to 24/7 usage of models**

[2] MLOps

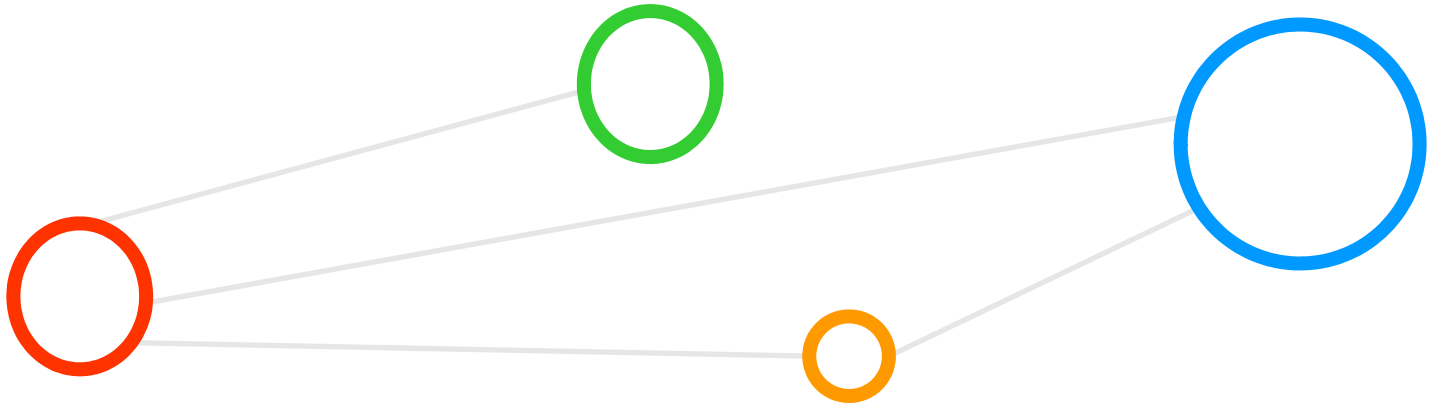
[6] Wikipedia on DevOps *[5] Wikipedia on MLOps*

[Video] MLOps using MLFlow



[7] ML Lifecycle | Why bother to start using MLflow?

MLOps with ClearML



Allegro AI Legacy & ClearML

- Legacy
  - Allegro AI founded in 2016 (founding team initially from Google)
  - Hold 47 patents in AI & Computer Vision
  - ‘Building up AI stacks from a lot of different tools on its own do not work‘
- Allegro Trains → Clear ML-Ops Platform
  - Released by Allegro as open-source
  - Used by 1000 organizations in >50 countries
- ClearML
  - Rebrand from Allegro AI (01/2021)
  - Partners like Nvidia, Intel, IBM, NetApp, Arrow, Hpe, academia, and mony others

allegro AI

[2] MLOps

CLEAR | ML

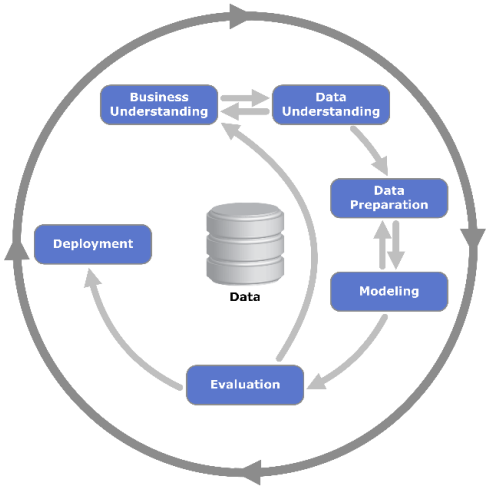
[3] ClearML Web page

BOSCH	Ford	HYUNDAI	PHILIPS	SAMSUNG	SONY
Agilent	Alibaba Group	facebook	Hewlett Packard Enterprise	Tencent 腾讯	TOSHIBA
AMD	IBM	Microsoft	NVIDIA	SoftBank	YAHOO! JAPAN
Deutsche Telekom	NYU	Stanford University	TECHNION Israel Institute of Technology	UNIVERSITY OF TORONTO	USC University of Southern California

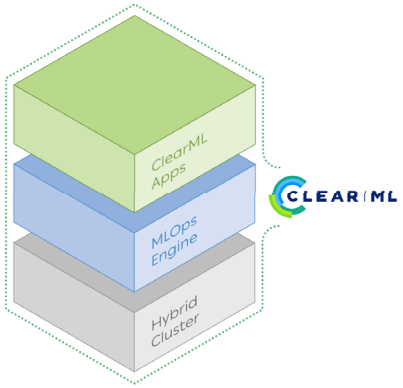
- **ClearML is an open-source platform that automates & simplifies developing & managing machine learning solutions.**
- **The goal of integrating the ClearML platform into machine learning pipelines and workflows is to enable reproducibility and automation.**

ClearML – Key Design as Lean-Stack MLOps Stack

- ‘Lean-Stack‘ solution approach
  - E.g., experiment management, workload orchestration, data management, one-click model deployment, etc.
- Design bottom-up & practice oriented
  - Easy integration with machine/deep learning codes
  - Logging and keeping the status (machine learning is a process!)
  - Enable comparisons of pipelines & workflow versioning
  - Dataset and machine/deep learning model management



[3] CRISP-DM Refrence Model



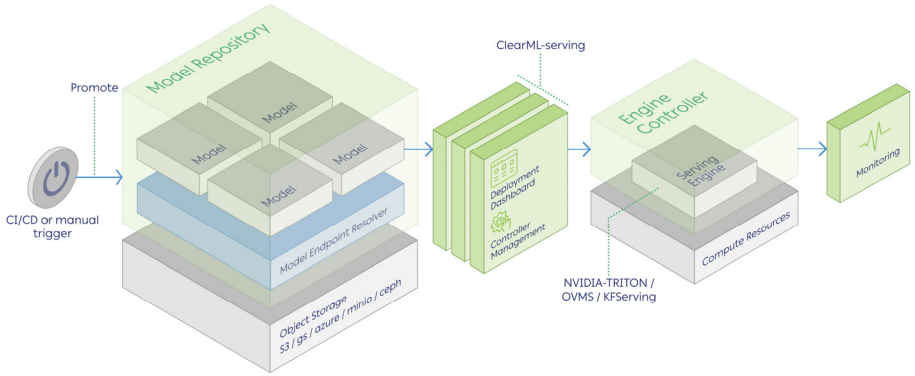
- **ClearML as a whole is an open-source stack architecture that consists of three different key layers**
- **ClearML Apps is a flexible, user-friendly MLOps tools for data science teams**
- **MLOps Engine represents the core functionality connecting ClearML apps to your compute resources**
- **Hybrid cluster stands for any compute resource available to the software stack like High-Performance Computing (HPC), Cloud Computing, or other computational resources**

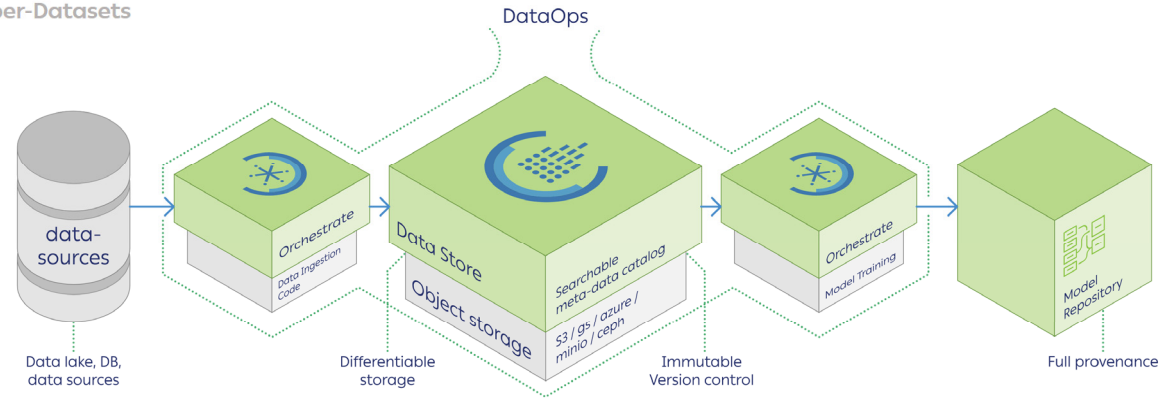
[2] MLOps

[3] ClearML Web page

ClearML Products work with other Tools – Not re-inventing the ‘wheel‘

- **The ClearML products consists of essentially six different products but highly intertwined**
- **ClearML Orchestrate integrates seamlessly with ClearML Experiment and ClearML Deploy: thus leveraging an end-to-end cross-department visibility in research, development, and production**

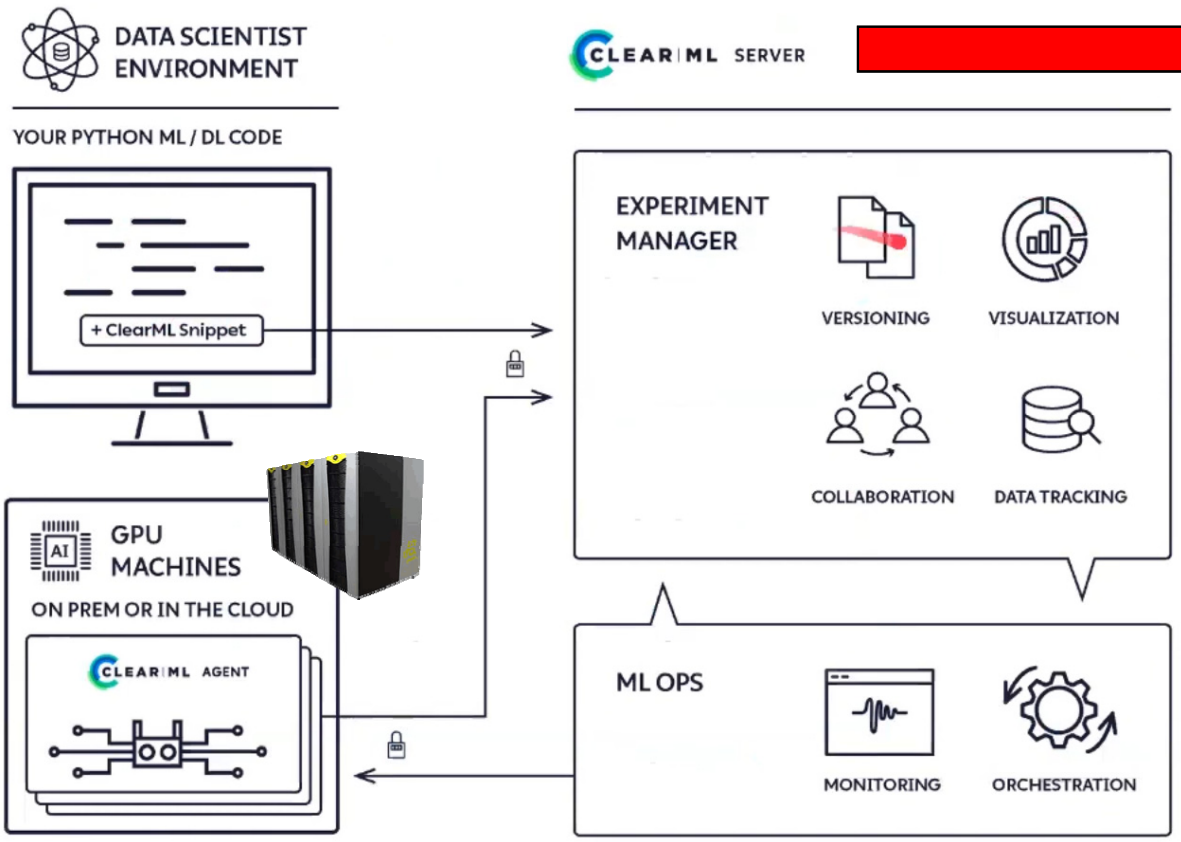




CLEAR | ML

[3] ClearML Web page

ClearML Overview & Benefits



(not for training purposes of models but rather for preparing the whole workflow/pipeline for production and deployment)

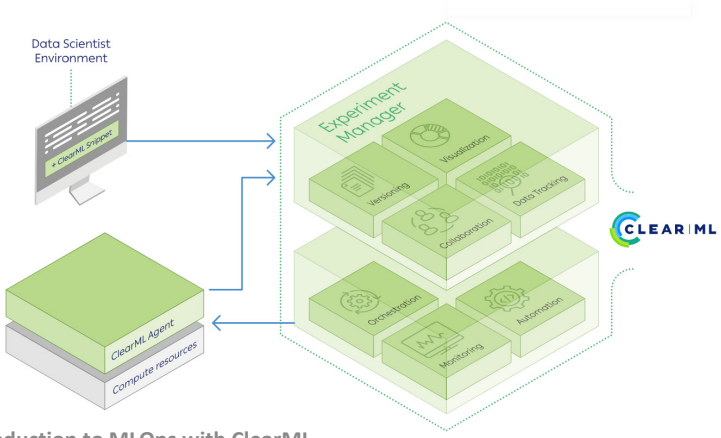
- **Your Python machine/deep learning (ML/DL) code: application logic & modeling of concrete algorithsm**
- **ClearML Snippets are integrated code in the python codes for ML/DL that links to the ClearML server**
- **Benefits of using the ClearML server are versioning, visualization, collaboration, and data tracking, etc.**
- **That also includes the MLOps layer with monitoring and orchestration capabilities linking to ClearML Agents**
- **ClearML Agents connect to GPU resources on premise, in the cloud or in HPC data centers to pull experiments from the Experiment manager to real computing resources**
- **ClearML is open-source is available to enable own operations and deployments of the software**

CLEARML

[3] ClearML Web page

MLOps Product Example – ClearML Experiment

- Tracks everything in the AI lifecycle and automates the process
  - Log, share, and version all machine learning experiments
  - Instantly orchestrate pipelines
  - Integrates with the orchestration workflow allowing for full visibility into any running process
  - Outputs : Console/Matplotlib etc. & development environment (Git/Uncommitted changes/Python packages/Args etc.) are automatically logged
  - The experiment manager GUI enables then tasks to be cloned, modified, and placed in an execution queue for a remote ClearML agent to pull



Install

```
pip install clearml
```

Copy paste into your code

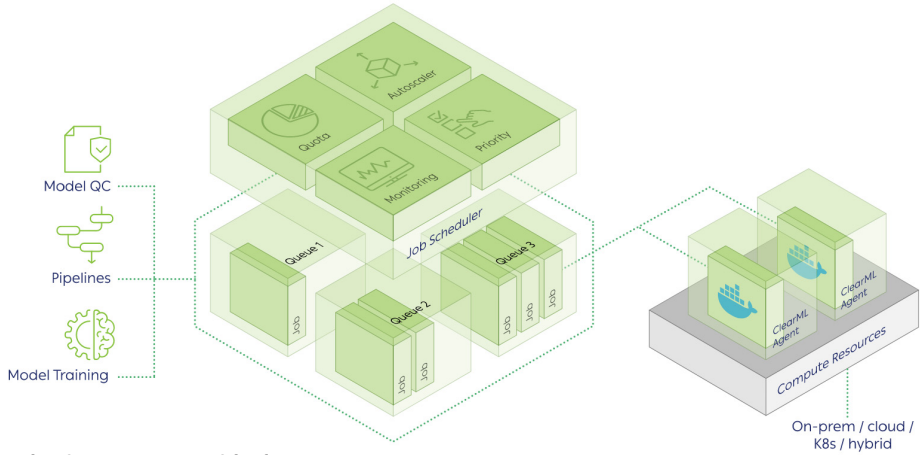
```
from clearml import Task
task = Task.init(
    project_name="your project",
    task_name="task name",
)
```

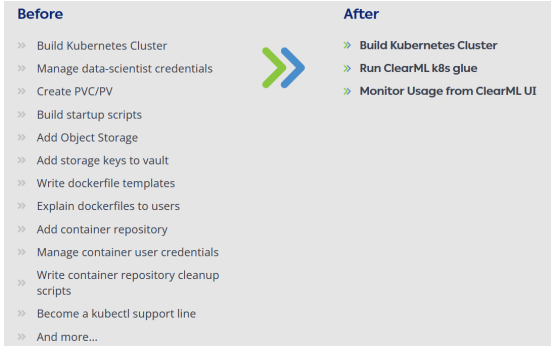
CLEARML

[3] ClearML Web page

MLOps Product Example – ClearML Orchestrate

- Enables Orchestration to DevOps & Automation to Data Scientists
  - Provides data scientists with autonomy and control over compute resources
  - Optimizes utilization, scale, and cost: Manage cloud bursting and autoscale maintaining high utilization rates and saving money with no vendor lock-ins
  - Access, manage, and control compute resources: Dynamically pool all compute resources from any environment, while managing priorities and scheduling from a unified interface across Kubernetes, on-prem, or the cloud

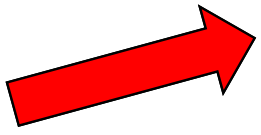




[3] ClearML Web page

Using ClearML Experiment – Simple Example (1)

- Setup
  - Essentially another Python package (i.e., from clearml import Task)
  - But not completely straightforward like other packages (e.g., NumPy)
- Integration
  1. 2 lines of code method
  2. 0 lines of code method



```
from sklearn import datasets
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
import numpy as np
import matplotlib.pyplot as plt

1  from clearml import Task


2  # Connecting ClearML with the current process,
   # from here on everything is logged automatically
   task = Task.init(project_name="My Workshop Examples", task_name="scikit-learn joblib example")

iris = datasets.load_iris()
X = iris.data
y = iris.target
```

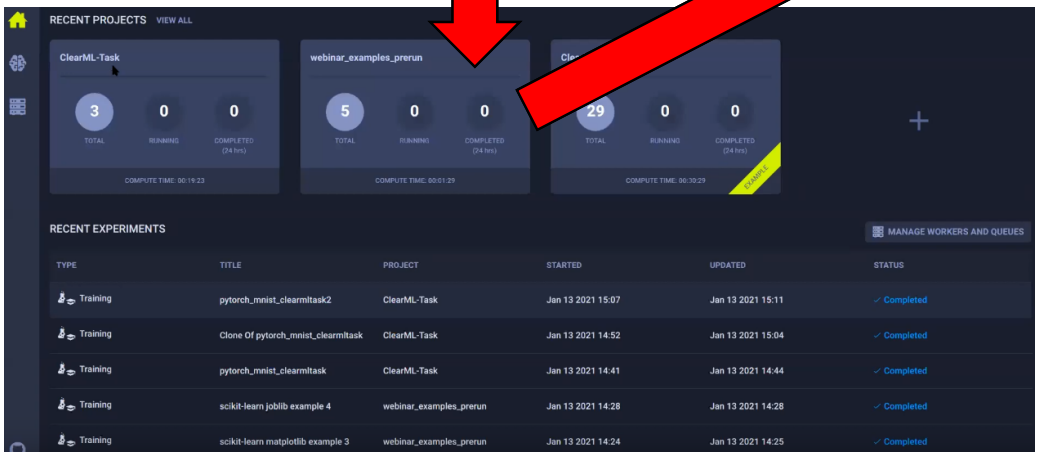
```
ariel@tafsik:~$ pip3 install clearml
Collecting clearml
  Using cached clearml-0.17.4-py2.py3-none-any.whl (873 kB)
Requirement already satisfied: python-dateutil>=2.6.1 in ./.local/lib/python3.8/site-packages (from clearml) (2.8.1)
Requirement already satisfied: jsonschema>=2.6.0 in ./.local/lib/python3.8/site-packages (from clearml) (3.2.0)
Requirement already satisfied: attrs>=18.0 in ./.local/lib/python3.8/site-packages (from clearml) (20.3.0)
```

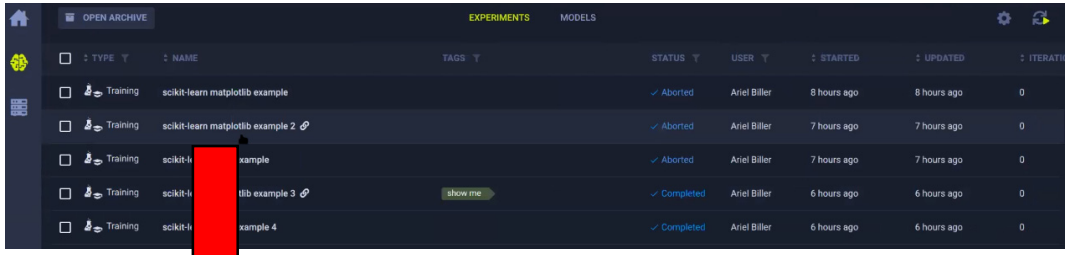
```
ClearML Hosts configuration:
Web App: https://app.community.clear.ml
API: https://api.community.clear.ml
File Store: https://files.community.clear.ml

Verifying credentials ...
Credentials verified!

New configuration stored in /home/ariel/clearml.conf
ClearML setup completed successfully.
ariel@tafsik:~$
```

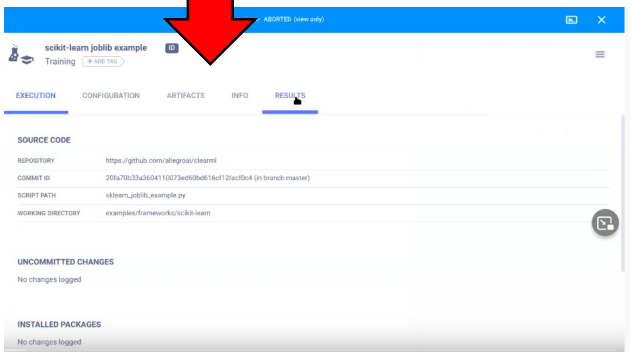
CLEAR | ML

[2] MLOps [3] ClearML Web page

Using ClearML – Simple Example (2)

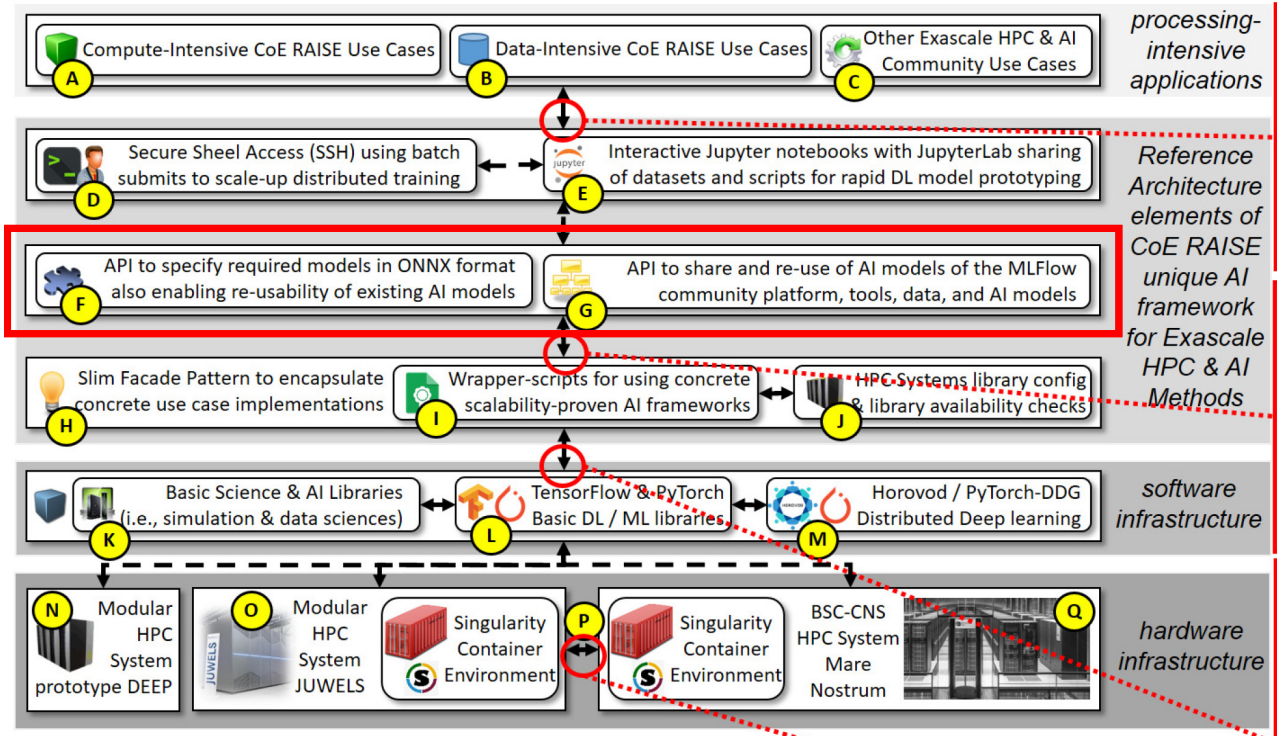






[2] MLOps [3] ClearML Web page

Summary and RAISE Unique AI Framework Context



✓ RQ1, RQ2, RQ4, RQ5

❖ Parts of the framework layout plan is to provide Kernels for Jupyter notebooks with correct version setups of modules for specific HPC Systems

✓ RQ3, RQ6

❖ Parts of the framework layout plan is to provide a lightweight and abstract Python API building on ONNX enabling also exchanges via MLFlow/ClearML

✓ RQ1, RQ2, RQ8, RQ9

❖ Parts of the framework layout plan is to provide a lightweight Python API that abstracts from low level versioning of AI packages (with proven scalability) and is harmonized with different available HPC system module versions

✓ RQ6, RQ7, RQ8, RQ9

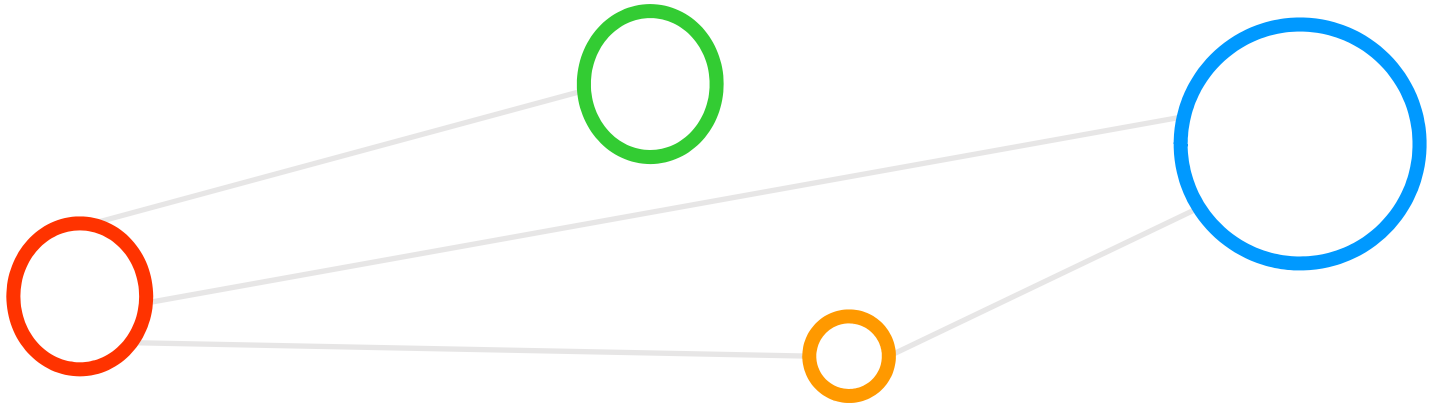
❖ Part of the framework layout plan is to provide containers in Singularity with prepackaged datasets & software stacks needed for AI agnostic to hardware & good I/O performance

RAISE Center of Excellence

Continously Updating

[8] RAISE CoE Web Page

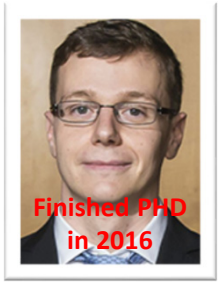
Selected References



Selected References

- [1] Icelandic HPC Community Web Page, Online: ihpc.is/community
- MLOps: From Zero to Hero in Two Lines of Code with ClearML, Online: https://www.youtube.com/watch?v=Y5tPfUm9Ghg
- [3] ClearML Web Page, Online: https://clear.ml/
- [4] Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22.
- [5] Wikipedia on MLOps, Online: https://en.wikipedia.org/wiki/MLOps
- [6] Wikipedia on DevOps, Online: https://en.wikipedia.org/wiki/DevOps
- [7] ML Lifecycle | Why bother to start using MLflow?, Online: https://www.youtube.com/watch?v=7TPHJUW9xFo
- [8] CoE RAISE Web page, Online: https://www.coe-raise.eu/

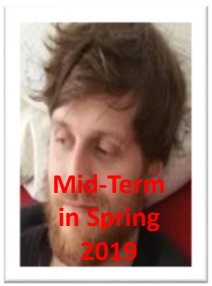
Acknowledgements – High Productivity Data Processing Research Group



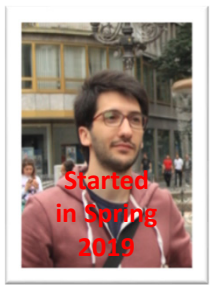
PD Dr.
G. Cavallaro

PD Dr.
A.S. Memon

PD Dr.
M.S. Memon

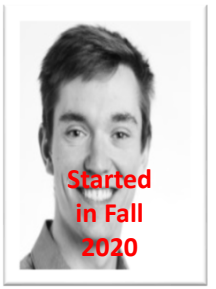


PhD Student
E. Erlingsson

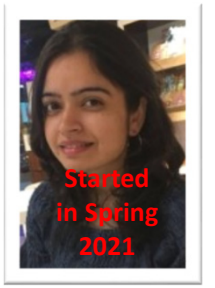


PhD Student
S. Bakarat

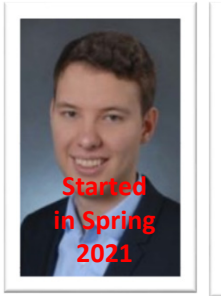
PhD Student
R. Sedona



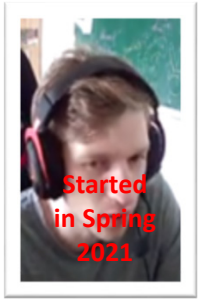
PhD Student
P. H. Einarsson



PhD Student
S. Sharma

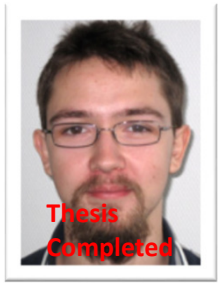


PhD Student
M. Aach

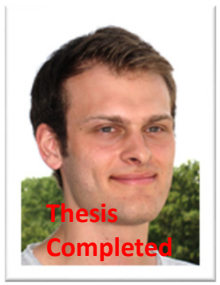


PhD Student
D. Helmrich

Dr. M. Goetz
(now KIT)

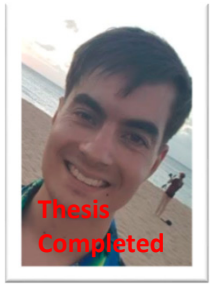


MSc M.
Richerzhagen
(now other division)



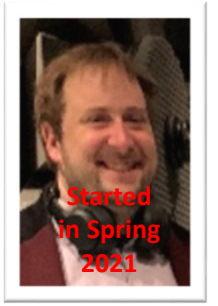
MSc
P. Glock
(now INM-1)

MSc
C. Bodenstein
(now
Soccerwatch.tv)



MSc G.S.
Guðmundsson
(Landsverkjun)

PhD Student
Reza



PhD Student
E. Sumner

This research group has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 763558 (DEEP-EST EU Project) and grant agreement No 951740 (EuroCC EU Project) & 951733 (RAISE EU Project)


services
measurement
within
fusion
new
funding
Policy-based
concepts
forms
cross-disciplinary
resources
device
Analysis
Centre
climate
Computer
expertise
year
like
Computational
Test
E.g
Computing
disciplines
results
Approaches
construction
Fusion
Enable
different
Cross-Disciplinary
methods
High
big
nodes
PB
community
Environment
Modelling
EU
Science
centers
Key
scientific
important
files
etc
DKRZ
Use
storage
basis
Infrastructure
simulations
ENES
technologies
SMAQ
international
model
engineering
References
formats
brain
Project
Data
increasing
via
EUDAT
physical
IPCC
processing
Common
TB
Big
MapReduce
create
Methods
computing
Using
European
databases
codes
GB
Supercomputing
Work
images
analysis
stored
Resources
analyze
Large
Summary
data-intensive
Service
Scientific
levels
often
JSC
use
DLCL
Level
Research
Matters
Concrete
external
performance
computational
research
e.g
science
HPC
Juelich
large-scale
system
datasets
acquisition
compute
Landscape
management
System
Climate
modelling
AR5
Hardware
cores
access
hundreds
Nutshell
Services
DLCLs
Understanding
structures
Technologies
Earth
distributed
Structure
current
network
Germany
Simulation
directory
project
General
Provide
NASA
Energy
systems
day
Processing
Health