# Perspectives in HPC and machine learning for large data driven applications

Claudio Gheller
Swiss Plasma Center
EPFL

Swiss SKA Days
Bern
June 19-20, 2019

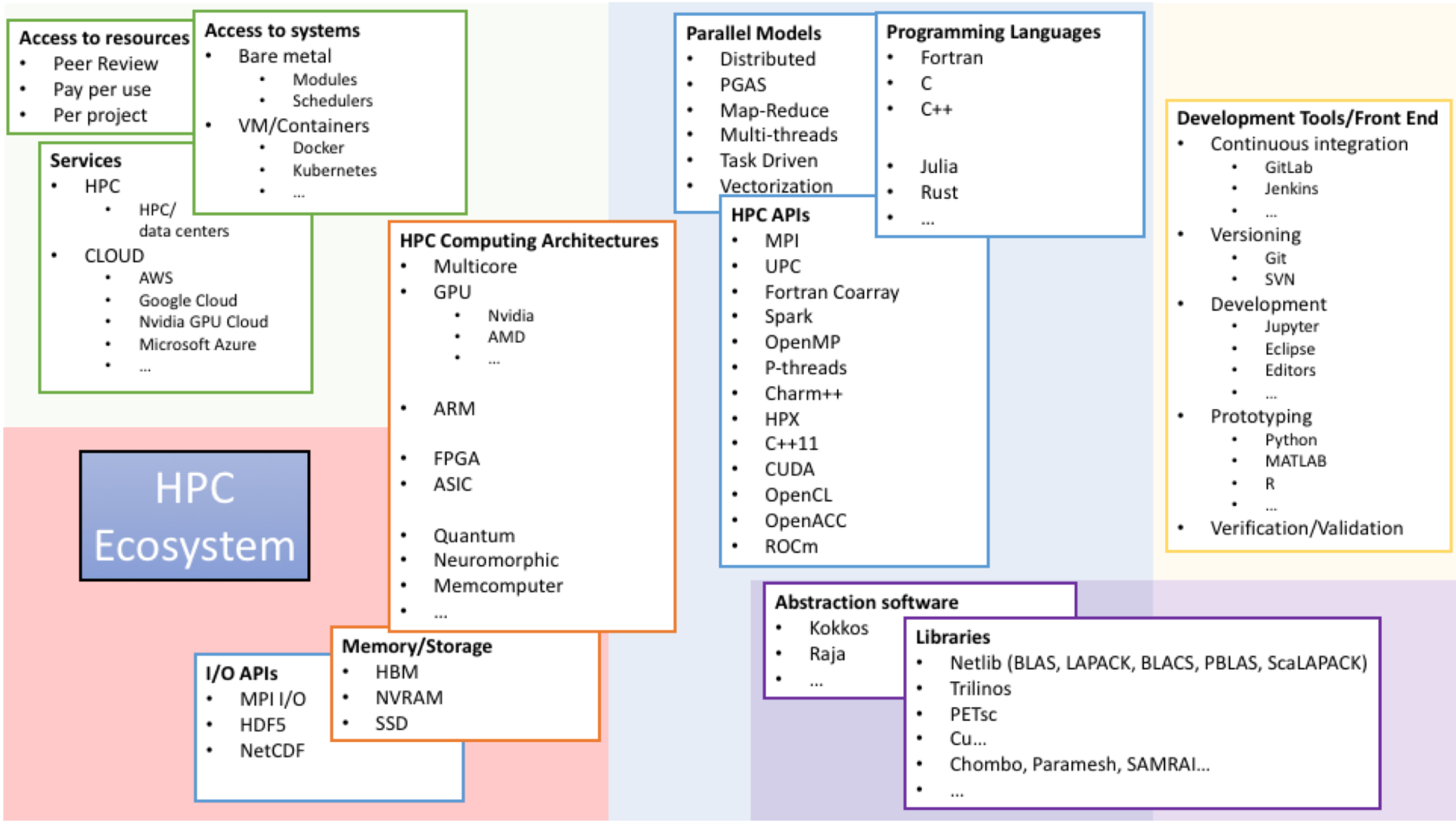
École
polytechnique
fédérale
de Lausanne

# Data driven science

Data tsunami

- Storage and management
- **HPC** for remote analysis/visualization
- **Data modelling**

# The HPC Ecosystem

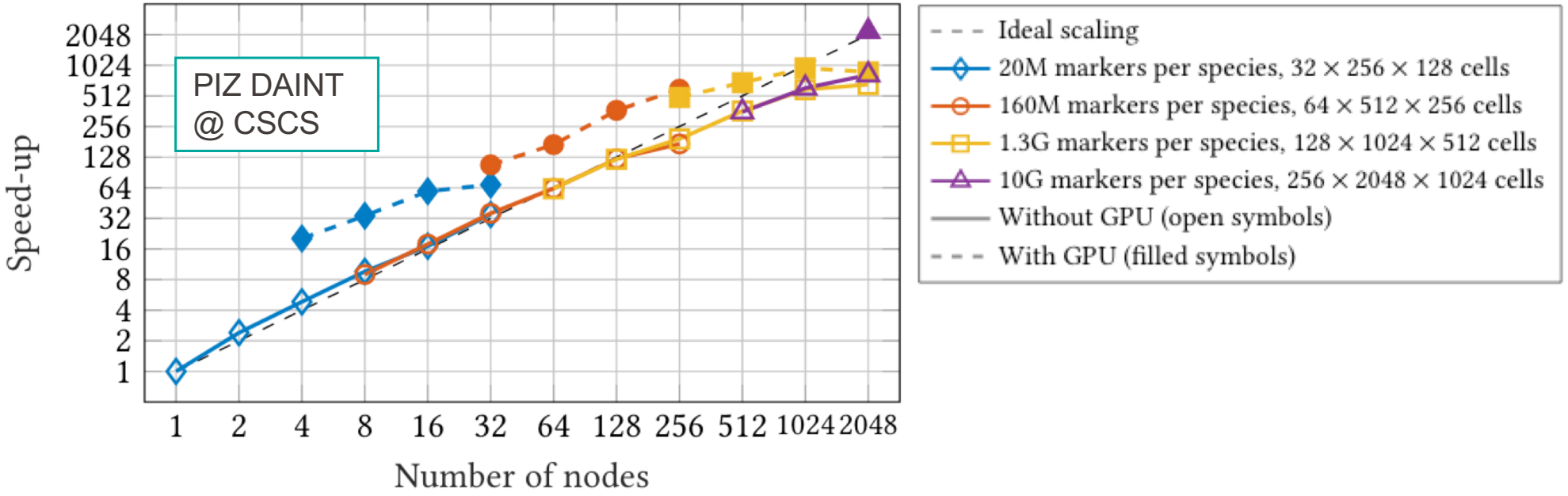


Claudio Gheller

# Can we exploit hybrid architectures at scale?

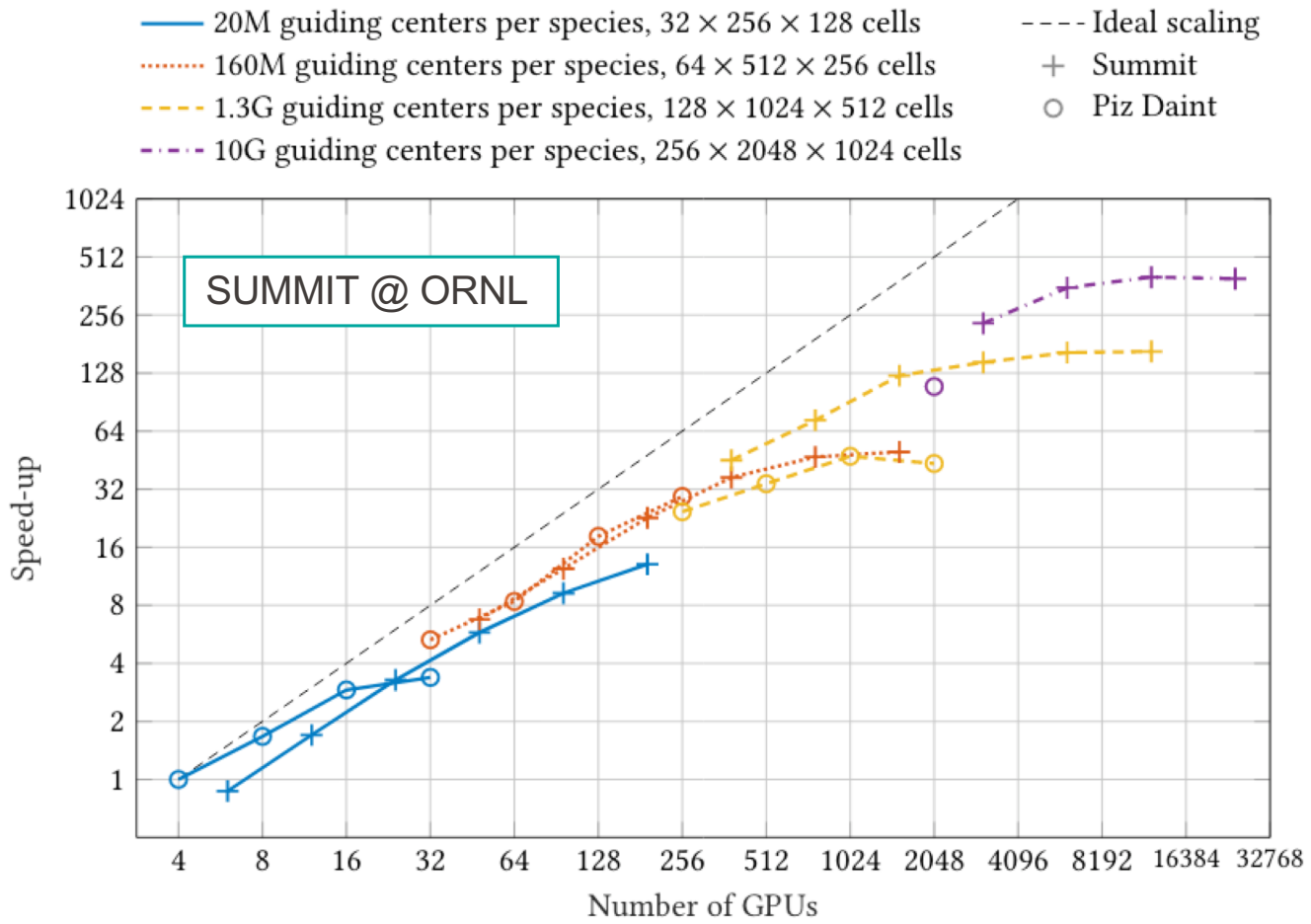
ORB5 Gyrokinetic PIC code for plasma physics simulations

Speed-up = N x T(1 CPU) / T(N)



# Can we exploit hybrid architectures at scale?

ORB5 Gyrokinetic PIC code for plasma physics simulations



# Machine Learning

- Broad family of algorithms in the domain of Artificial Intelligence acting on data:
  - Regression
  - Classification
  - Generation
  - Segmentation
  - …

Many different solutions:

- CNN, RNN
- K-means
- Decision trees, Random forests
- Generative networks
- Reinforcement learning
- Autoencoders
- …

# Machine Learning in literature

Preprint with:

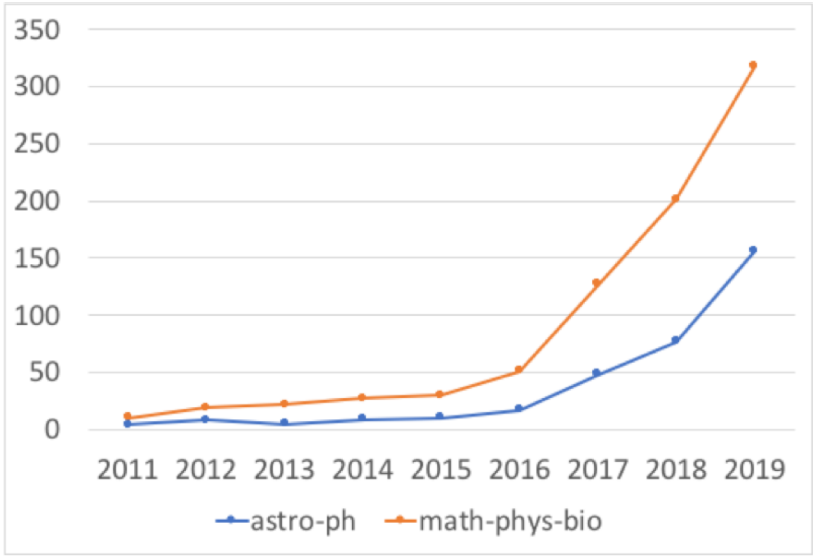
- *Deep Learning or*
- *Machine Learning or*
- *Neural Network*

in the title

Preprint in:

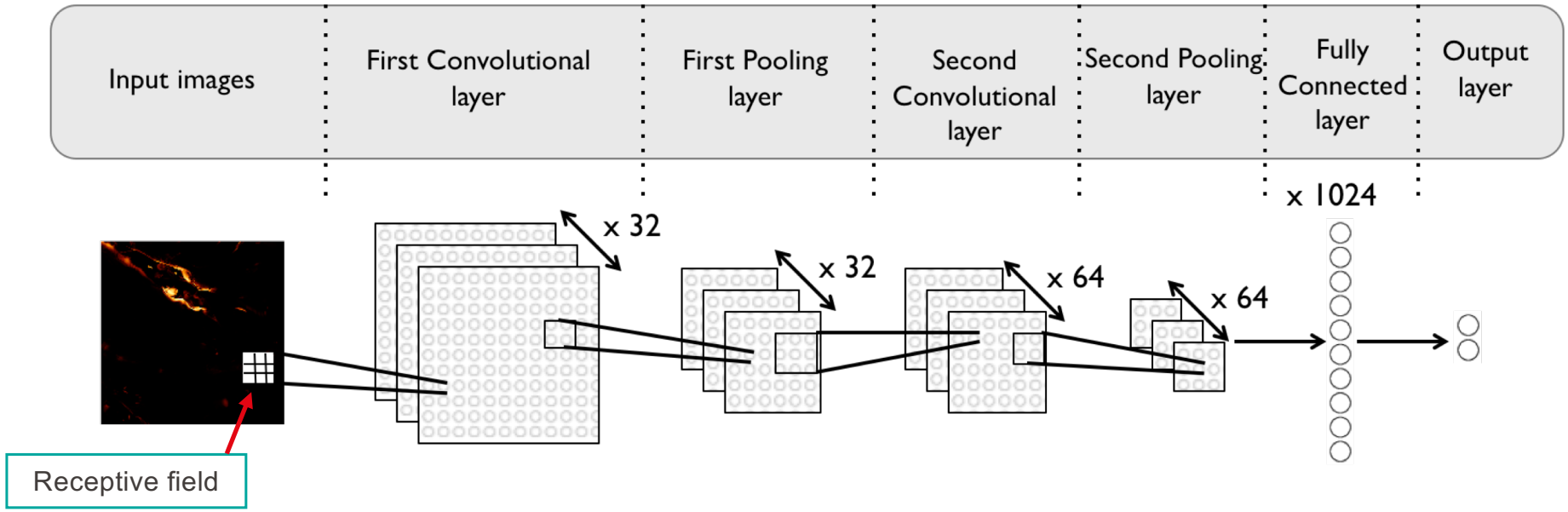
- *Math or*
- *Physics or*
- *Bio*

| Year | astro-ph | Math-Phys-Bio |
|---|---|---|
| 2000-2010 | 51 | 103 |
| 2011 | 4 | 10 |
| 2012 | 8 | 19 |
| 2013 | 5 | 22 |
| 2014 | 9 | 27 |
| 2015 | 10 | 30 |
| 2016 | 17 | 51 |
| 2017 | 48 | 127 |
| 2018 | 77 | 201 |
| 2019 (jan-feb) | 26 (proj. 156) | 53 (318) |



Claudio Gheller

# Example 1: Convolutional Neural Networks




Activation function: **ReLU**

$\sigma$ = max(0,a)

$$a = \left( b + \sum_{l=0}^{2} \sum_{m=0}^{2} w_{l,m} a_{j+l,k+m} \right)$$

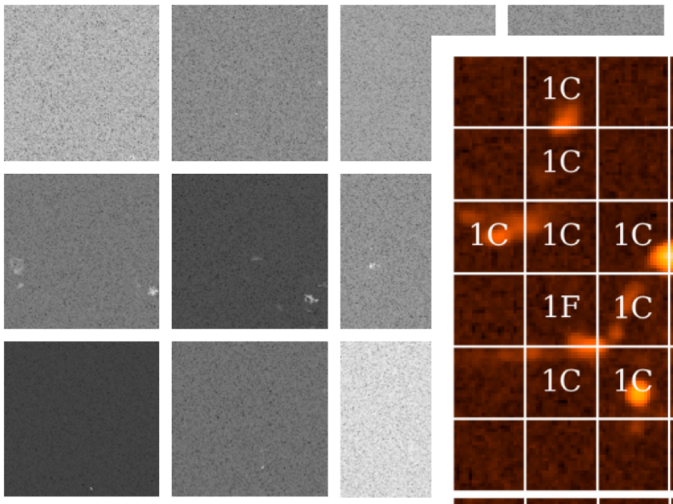
Pooling function:
**Max Pool** → Max of the values in a 2x2 window

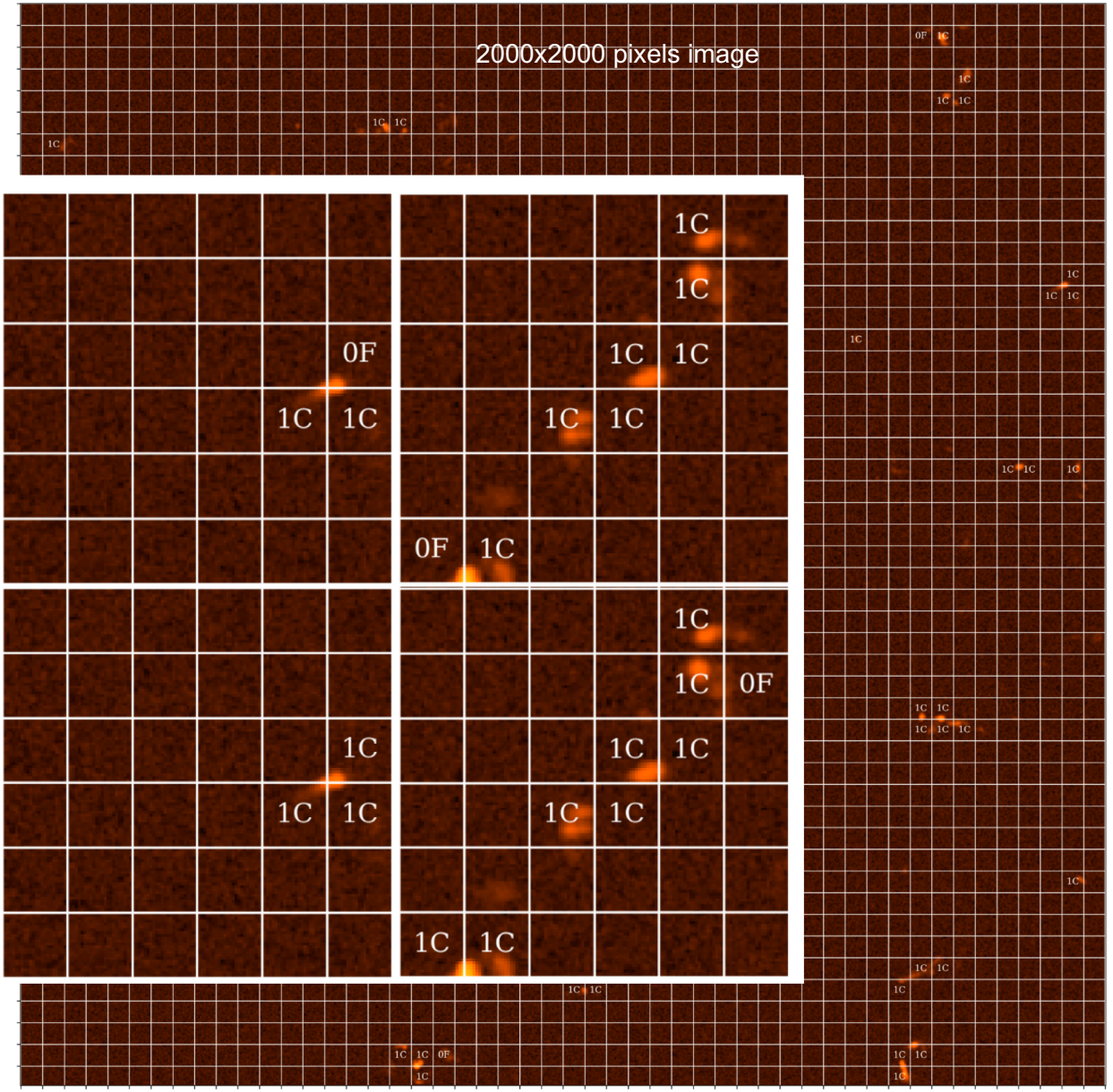
Cost function:
**Cross Entropy**

$$C = -\frac{1}{n} \sum_{x} [y \ln a + (1-y) \ln(1-a)]$$

Gradient descendent optimiser:
**ADAM**

EPFL






| Task (P100 GPU) |
| --- |
| Training |
| Processing |

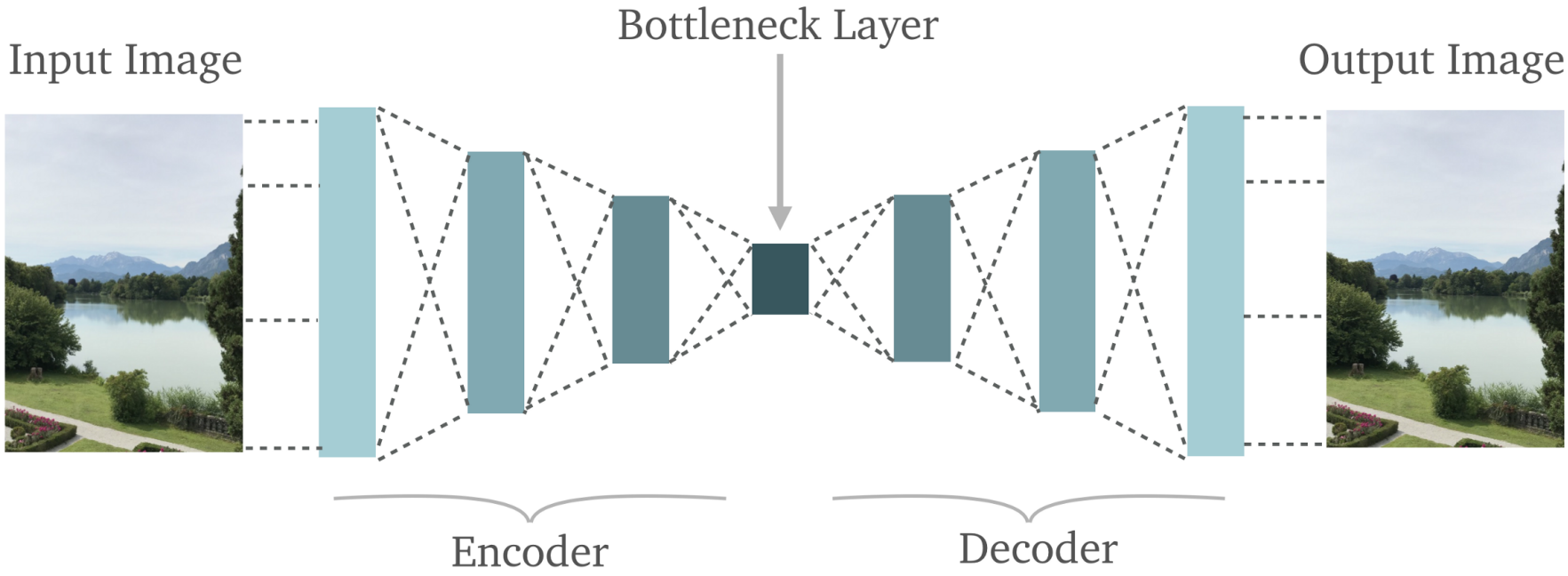
**Accuracy: 0.88 – 0.**

CPU performance: ~ 10 times slower

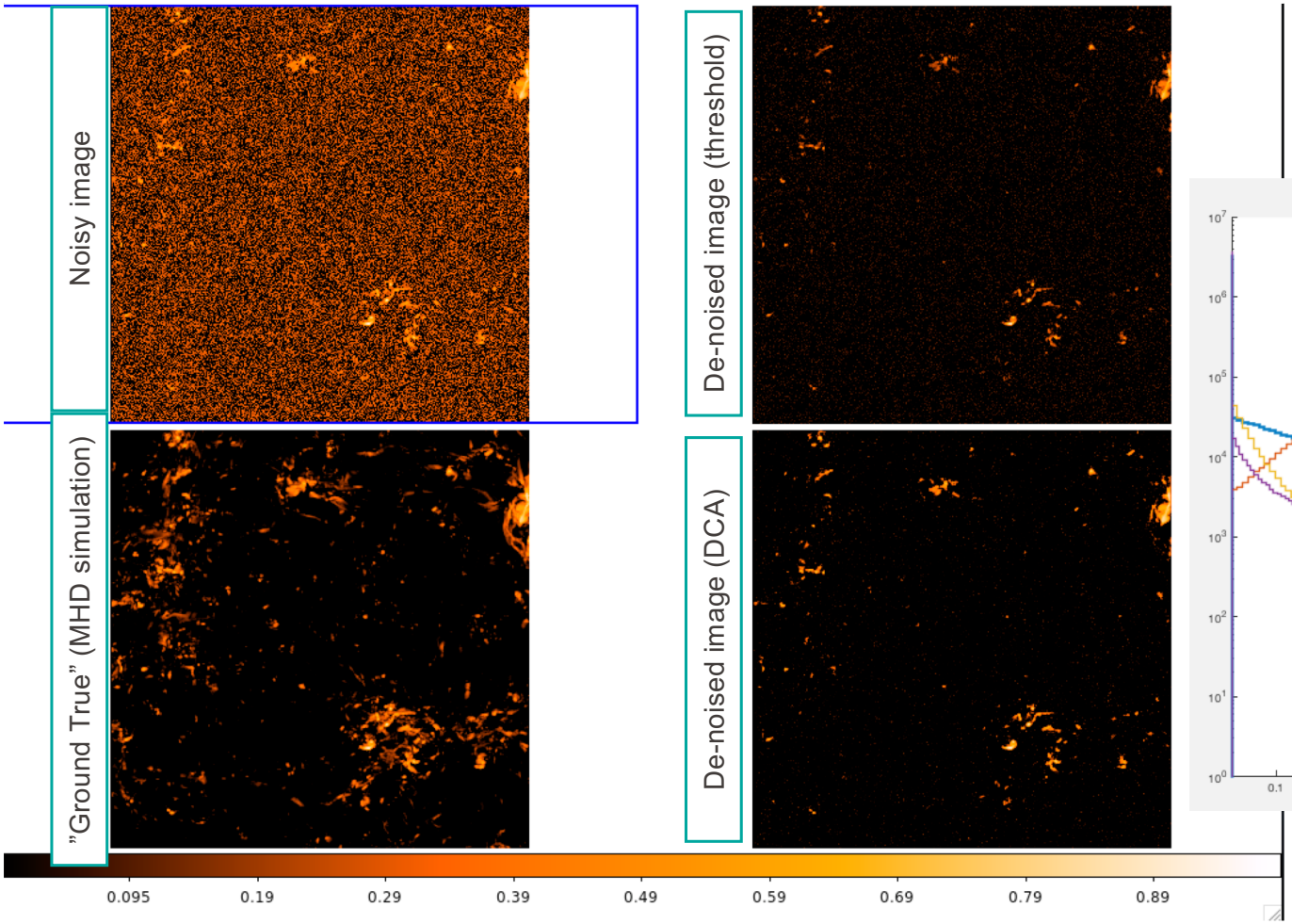
(Tensorflow implementation)

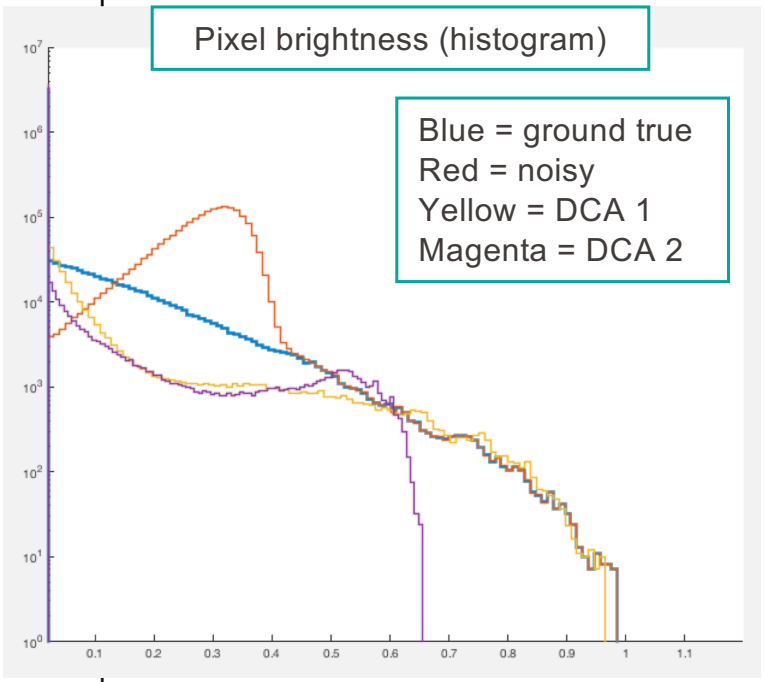
Claudio Gheller

# Deep Convolutional Autoencoders



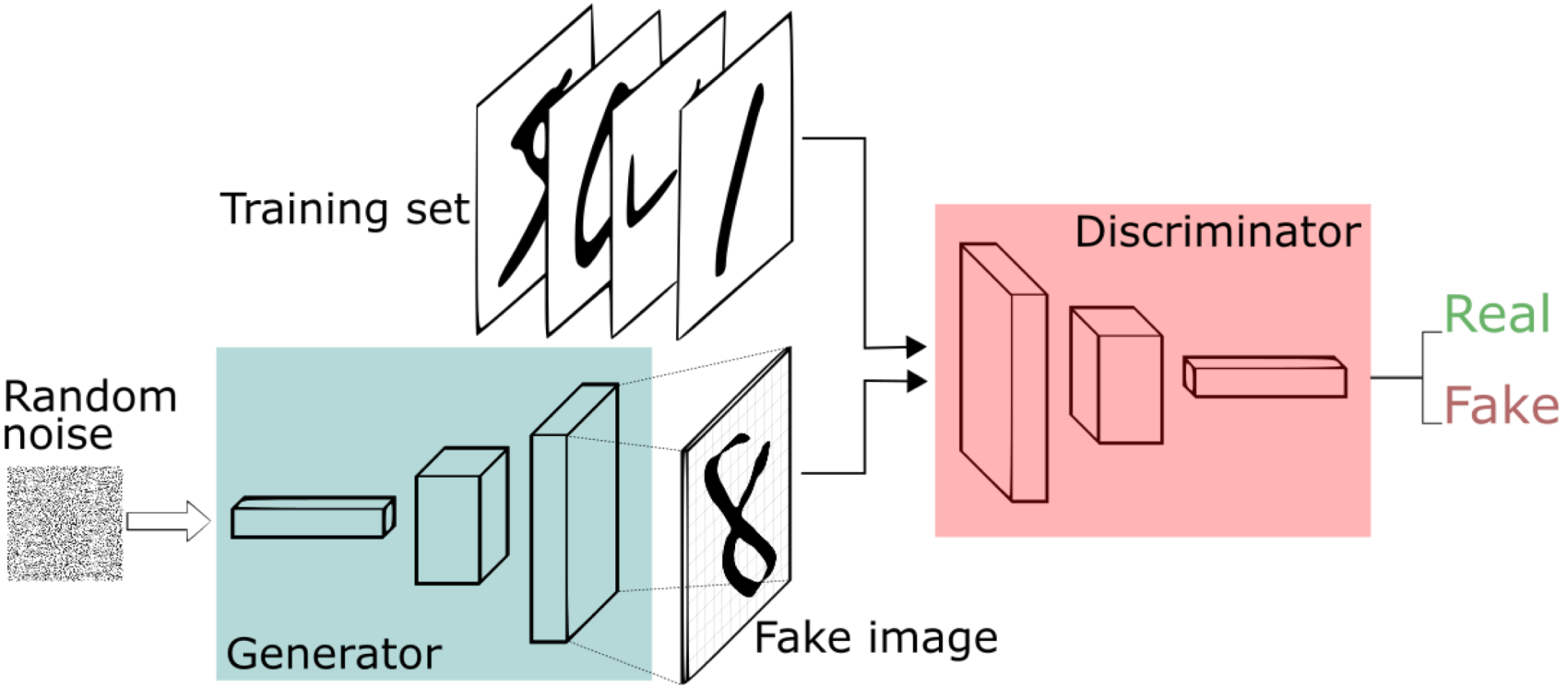
Claudio Gheller



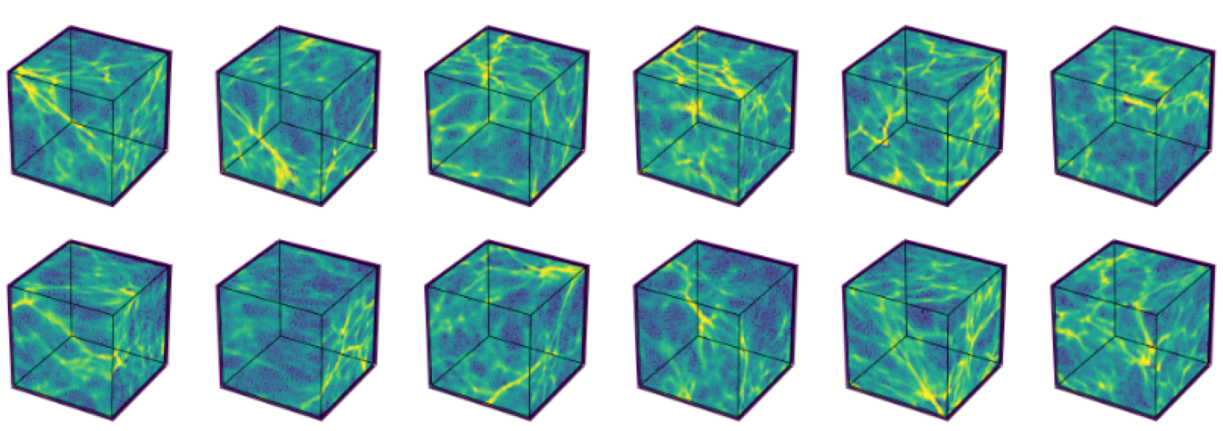
Claudio Gheller

# Generative Adversarial Networks

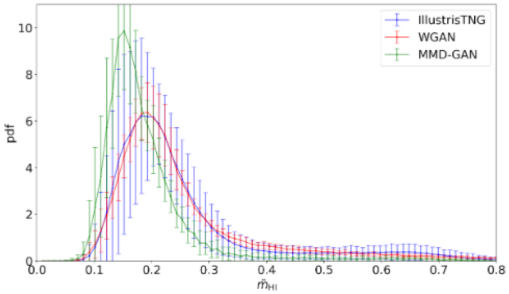


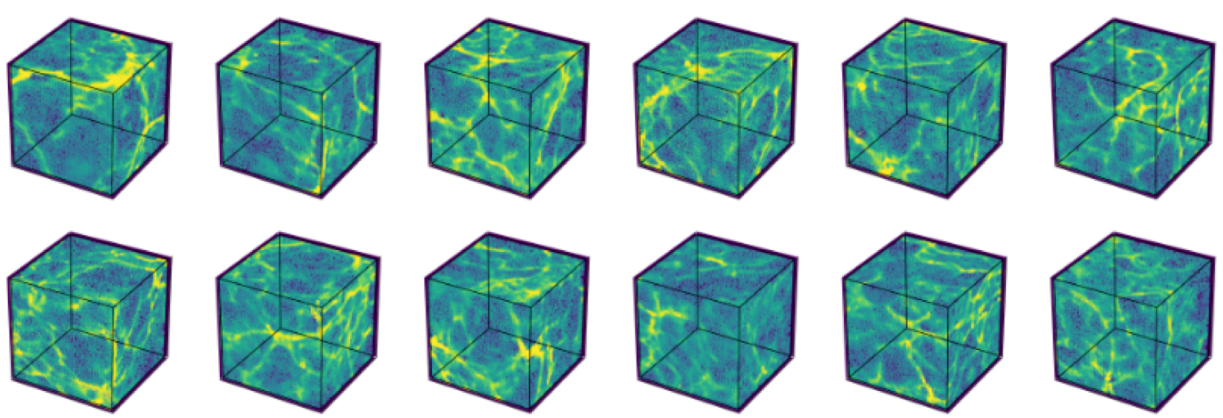
EPFL

# HI distribution



IllustrisTNG simulation



100000 times faster!

GAN

J. Zamudio-Fernandez et al., arXiv:1904.12846, 2019

Claudio Gheller

# Machine Learning and HPC

## A controversial relationship

- Plus (forward propagation):
  - Lots of data → plenty of computation
  - Lots of data → large memory needs
  - Lots of linear algebra
- Minus (backward propagation):
  - Full connection → Lots of communication
  - Layers synchronization
  - Work balance at different layers
- Precision and accuracy
- Reproducibility
- Flexibility

# Summary





- We (will) have a lot of data
- HPC is necessary
- The scenario is complex, but it must be exploited
- Machine learning can do a lot and exploit it, but…
- …it is not the “one fits all” solution