AI-driven Materials Science

Martin Eriksen

Finançat per:

Flock of dinosaurs invading the ALBA synchrotron

Renaissance of neural networks

- 2012 Neurips paper
- 121014 Citations
- Important event in the history of AI.

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

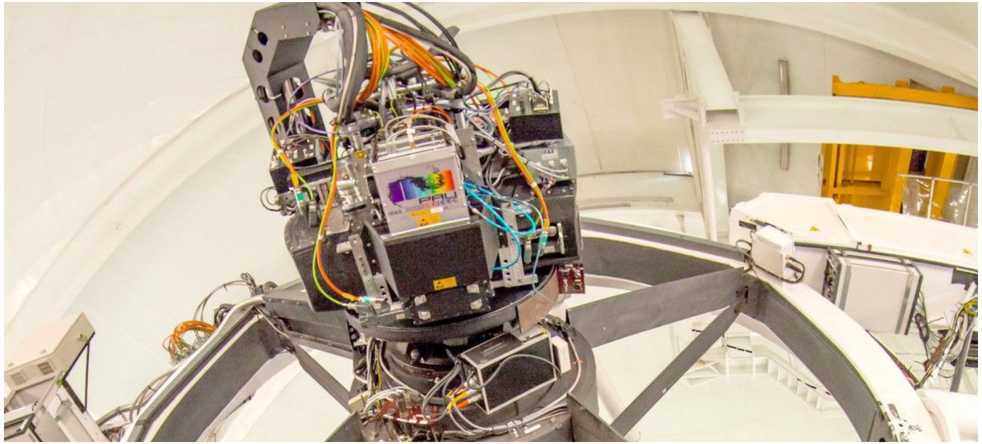
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

The PAU survey

PAU camera

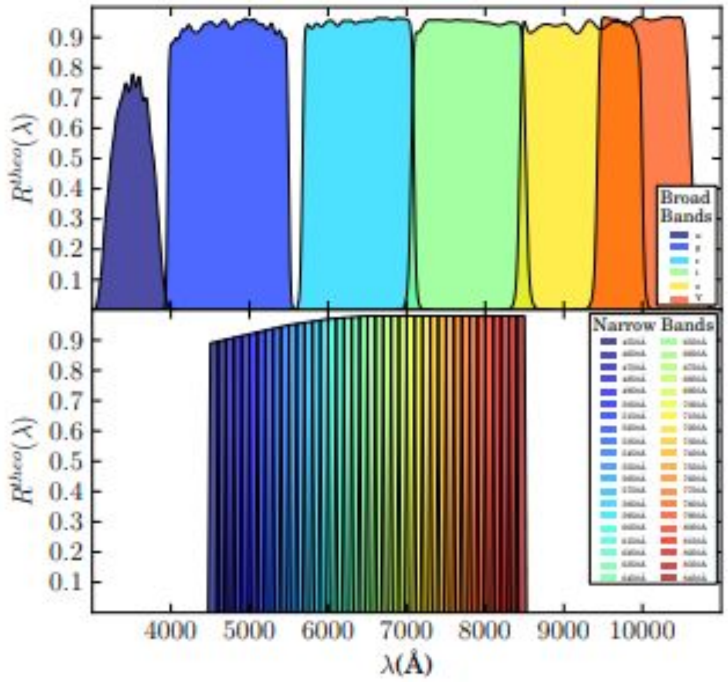


William Herschel Telescope

PAUS webpage: https://pausurvey.org/

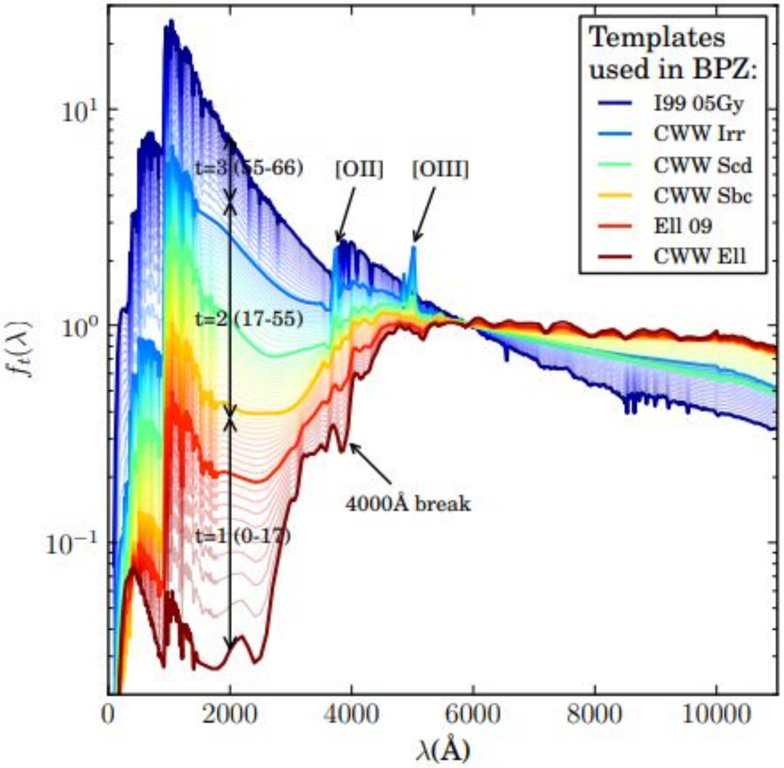
Narrow band filters

- Observing in the optical range.
- Normally a survey either take spectra of a galaxy or 5-6 broad band measurements.
- PAUS has in addition 40 narrow bands.



Template fitting with simulations

- Preparation for the instrument used simulations.
- Simulate galaxies with models on the right, generated some distributions and added noise.
- Fitted with a code using roughly the same models (SEDs).

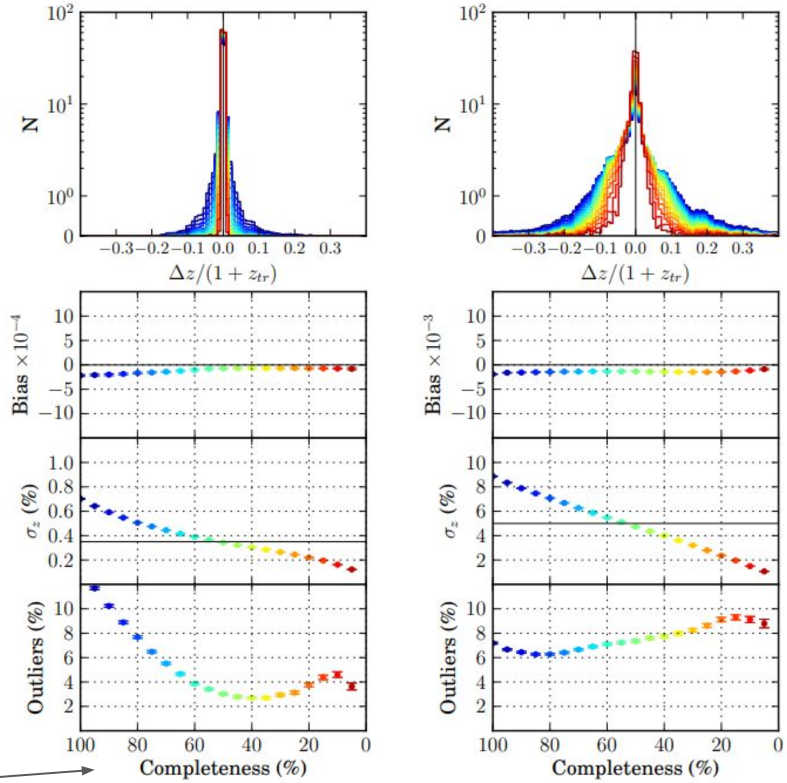


Results

- Simulations showed that we can achieve roughly what we promised with even simpler simulations.

Target precision

Fraction of galaxies after a "quality cut".



arxiv: 1402.3220

AnnZ - Neural network photo-z

arXiv
https://arxiv.org › astro-ph

[astro-ph/0311058] ANNz: estimating photometric redshifts ...

by AA Collister · 2003 · Cited by 632 — Abstract: We introduce **ANNz**, a freely available software package for photometric redshift estimation using Artificial Neural Networks.

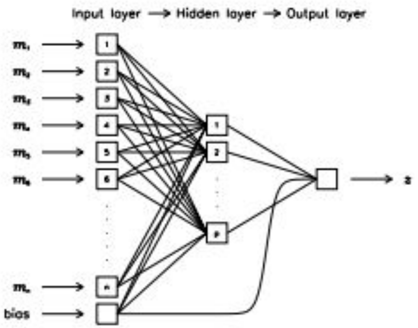


FIG. 1.— A schematic diagram of a multi-layer perceptron, as implemented by ANNz, with input nodes taking, for example, magnitudes $m_i = -2.5 \log_{10} f_i$ in various filters, a single hidden layer, and a single output node giving, for example, redshift z. The architecture is n:p:1 in the notation used in this paper. Each connecting line carries a weight w_{ij}. The bias node allows for an additive constant in the network function defined at each node. More complex networks can have additional hidden layers and/or outputs.

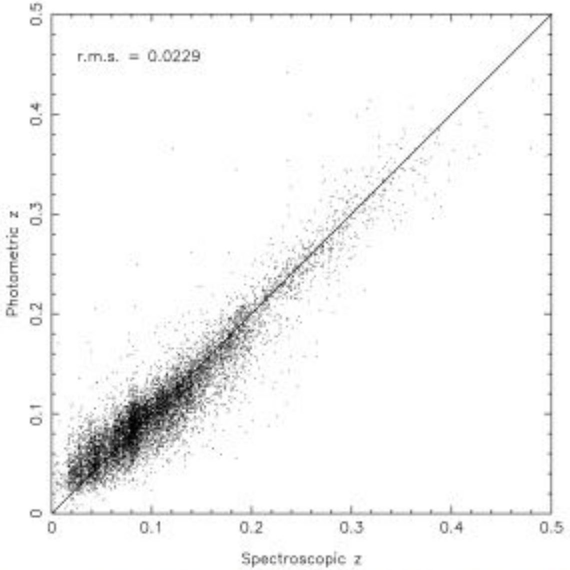
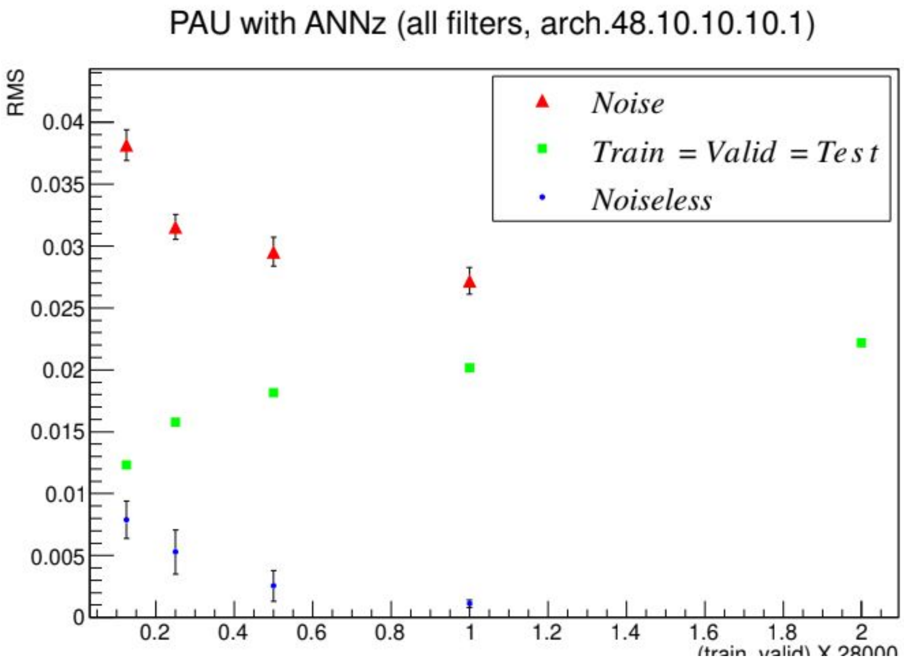


FIG. 2.— Spectroscopic *vs.* photometric redshifts for ANNz applied to 10,000 galaxies randomly selected from the SDSS EDR.

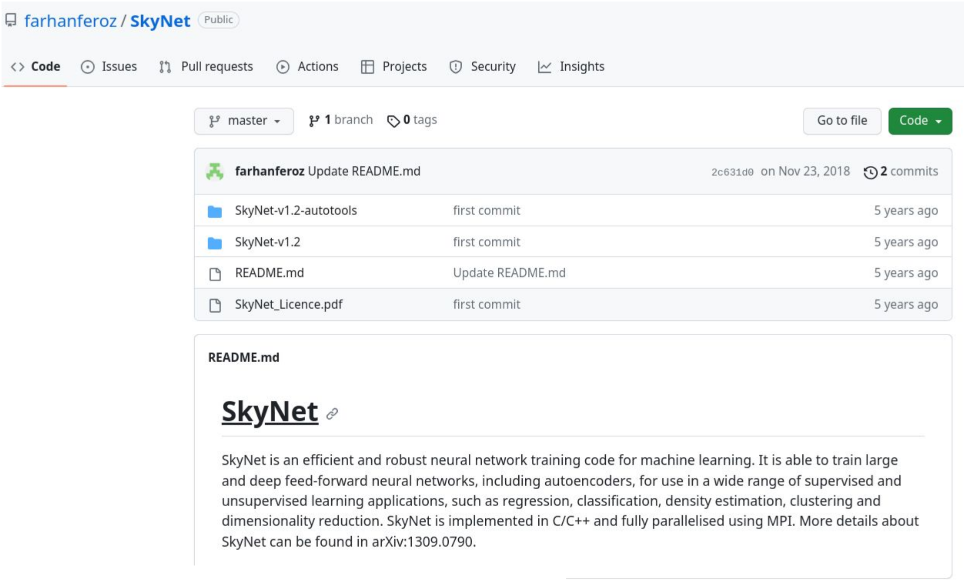
Results on data

- Result in email from Carlos Sanchez Alonso on 4 May **2012**, 09:56.
- Results were far from what we needed.



SkyNet applied to photo-z

- Example of paper applying neural networks to estimate redshifts.
- Used the SkyNet package for neural network training.



Redshift distributions of galaxies in the Dark Energy Survey Science Verification shear catalogue and implications for weak lensing

C. Bonnett *et al.* (Dark Energy Survey Collaboration)
Phys. Rev. D **94**, 042005 – Published 30 August 2016

Solid deep learning frameworks

PyTorch (Initial release in 2016)

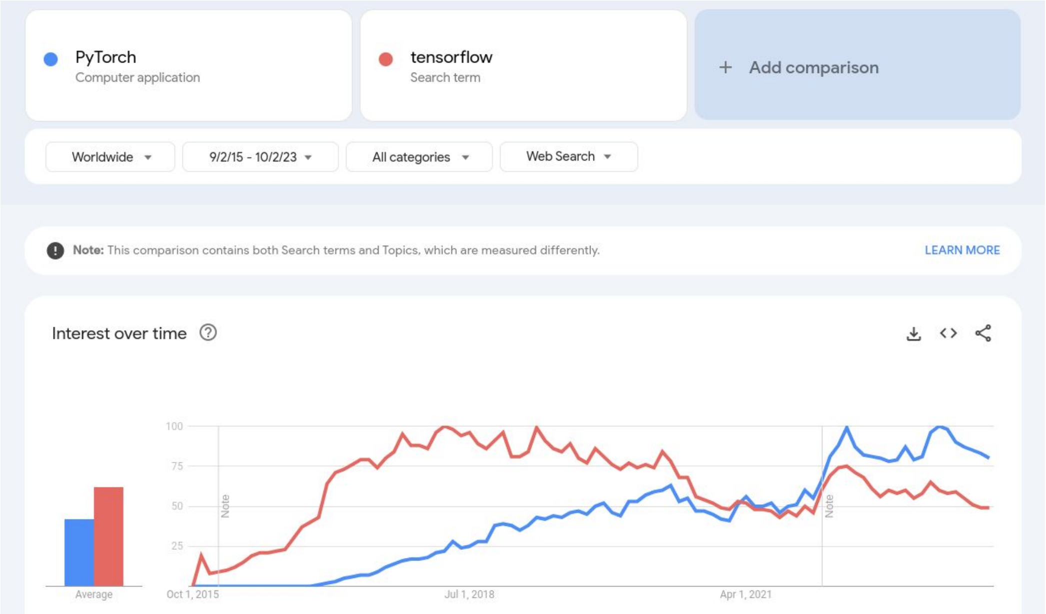
Tensorflow (Initial release 2015)

Keras, userfriendly interface to Tensorflow

Frameworks helps you a lot

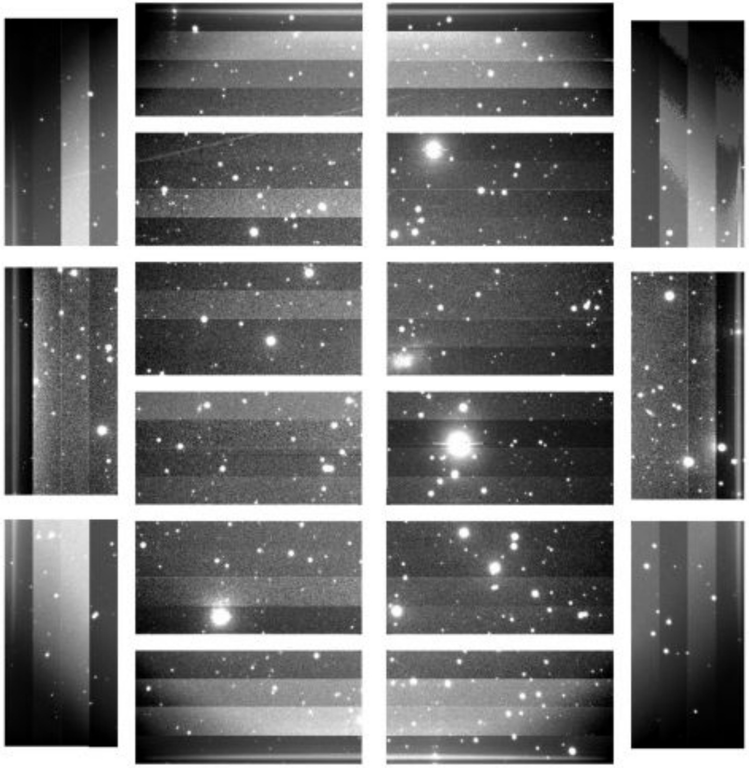
Which to choose

- Pytorch is currently more used especially in research literature.
- More flexible and easier to implement new ideas.
- Easier to debug.
- A bit harder to start.
- I prefer PyTorch.



PAUS images

A sky exposure of the full 18-detector PAUCam mosaic. As a raw image, all instrumental signatures are present, and the 72 amplifiers can be identified as well as the vignetting from the WHT prime focus corrector. (arxiv:2206.14022)



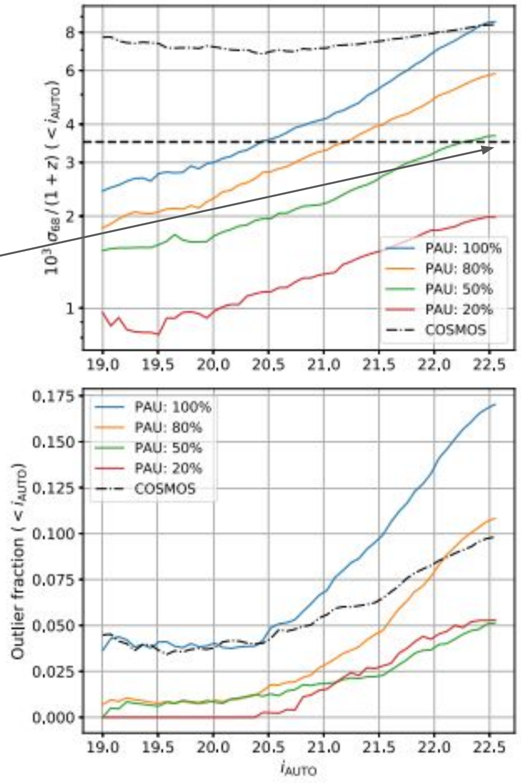
M101

- Discovered in 1781 by Pierre Mechain.
- In one of our science fields.
- Took the pictures just because we could.
- (arxiv: 2206.14022)

First PAUS results

- Template fitting.
- New code BCNz2 for properly modelling emission lines and including calibration factors.
- Became much slower because of running with 35 different configurations.
- Extension the modelling became increasingly difficult..

Green line (50%) below the target for the faint galaxies. Success!

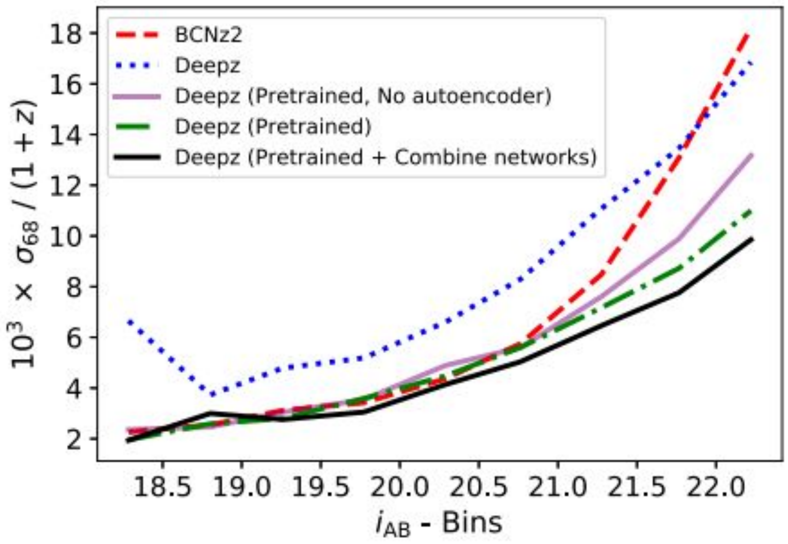


$$\chi^2[z,\boldsymbol{\alpha}] = \sum_{i,NB}\left(\frac{\tilde{f}_i - l_i k f_i^{\mathrm{Model}}}{\sigma_i}\right)^2 + \sum_{i,BB}\left(\frac{\tilde{f}_i - l_i f_i^{\mathrm{Model}}}{\sigma_i}\right)^2$$

MNRAS 484, 4200–4215 (2019) doi:10.1093/mnras/stz204

The PAU Survey: Photometric redshifts using transfer learning from simulations

- High i_{AB} are fainter galaxies, these are more interesting.
- Same performance for the bright galaxies.
- Better results with the neural network.



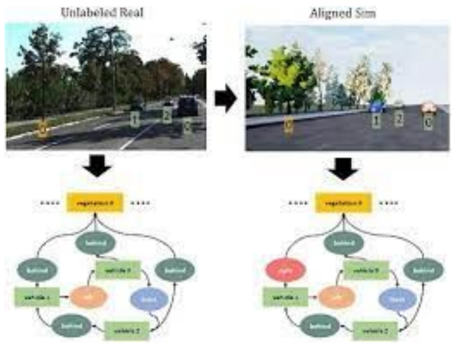
MNRAS 497, 4565–4579 (2020)

Pretrain with simulations

- Simulate 1 million galaxies with a theory code.
- Data: <10.000 galaxies
- Train the network on simulations, then continue training on data.
- There are important unknown shifts between the simulations and data.

Parameter	Range	Unit
zred	[0, 1.2]	Redshift
logzsol	[−0.5, 0.2]	$Z/Z_\odot$
tage	[0, 14]	Gyr
tau (τ)	[0.1, 12]	Gyr
const (k)	[0, 0.25]	Fraction
sf_start (t_i)	[0, 14]	Gyr
dust2 ($E(B-V)$)	[0, 0.6]	Colour
log_gasu	[−4, 1]	Dimensionless

[Self-Supervised Real-to-Sim Scene Generation]

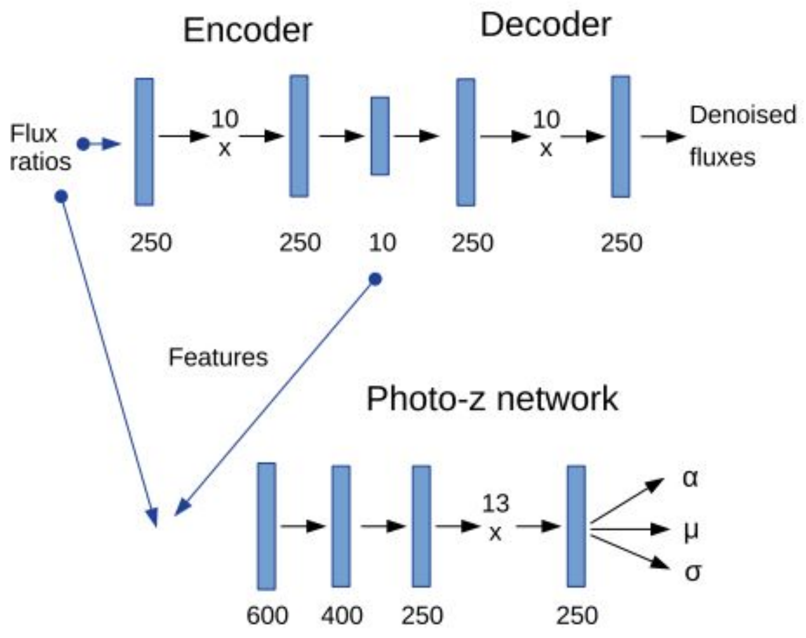


Extensive literature on the combination of simulations and data. **Still have potential to improve the procedure.**

Auto-encoder + MDN

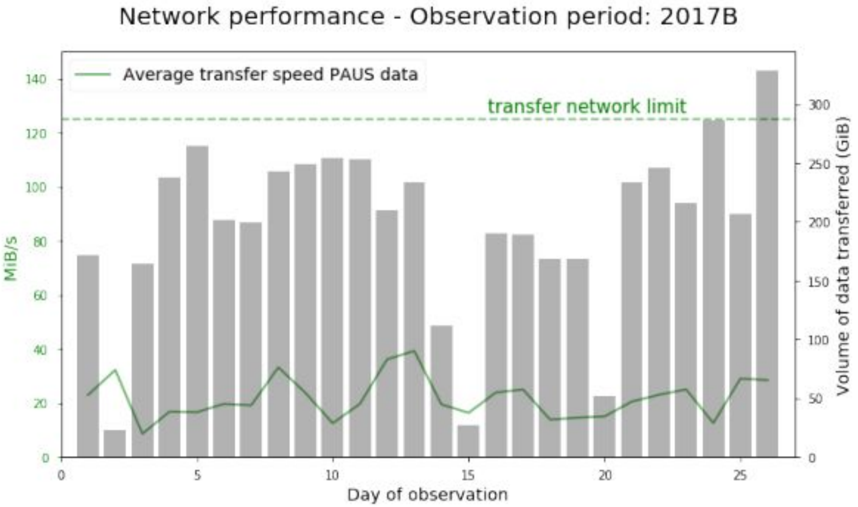
- Auto-encoder extract features of the galaxies, which can be trained **without label**.
- Mixture density network that predicts the probability distribution in redshift.
- Not sufficient to simply randomly vary the input.

$$p(z) \propto \sum_{i=1}^{M} \beta_i N(\mu_i, \sigma_i),$$

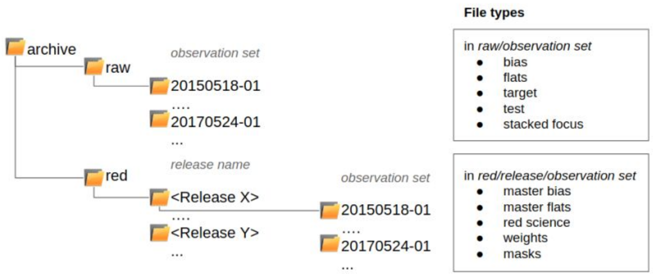


PAUS data management

- Observed at La Palma, transferred to PIC.



The limit is the connection out of La Palma.



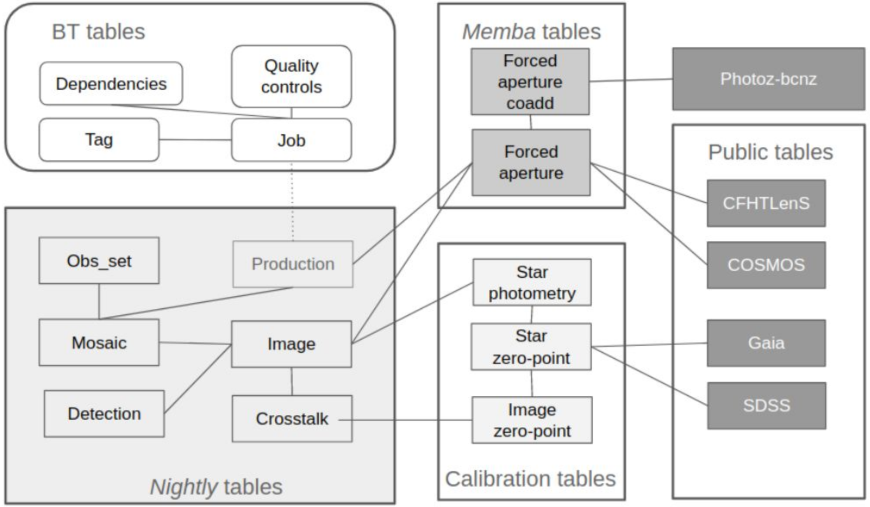
Stored at PIC. Defined structure at disc and backup to tape.

Database tables

- PostgreSQL server storing the reduced catalogs.
- Common and central data reduction for the whole collaboration.

Production Table

Production	Input Production	Pipeline	Release	Comments	Created
1075	1068	photoz	20A-R11-CMK-03-BCNz03		Oct 3, 2023 9:29:14 PM
1074	1068	photoz	20A-R11-CMK-03-BCNz02	BCNz over KIDZ_COSMOS 1068 until mag_i 23.1. The calibration is done using the spectroscopic redshifts from the validation sample of KiDZ-COSMOS. The photo-z are computed using new models that arrive up to zb_max = 2.0. Without negative fluxes cut. Objects with NB coverage equal or greater than 30 are included. Fluxes corrected by extinction. New filters with atmospheric absorption.	Sep 28, 2023 11:13:25 PM
1073	1057	photoz	20A-R11-W2-05-BCNz04	BCNz over W2-KiDS DR4 1057 until mag_i 23.1. The calibration is done using the spectroscopic redshifts from the validation sample of W2, mainly SDSS and GAMA. The photo-z are computed using new models that arrive up to zb_max = 2.0. Without negative fluxes cut. Objects with NB coverage equal or greater than 30 are included. Fluxes corrected by extinction. New filters with atmospheric absorption.	Sep 28, 2023 4:59:51 PM
1072	943	memba	20A-R11-CM-15	FA over COSMOS until 20A without crosstalk correction (NightlyR11). Fixed circular apertures of diameter 3 arcsec, constant annulus from 30 to 45 pixels, Scatterlight correction regular (32-5), flagging scatterlight to 10x ratio. Star apertures fixed. With Mask production. Coadd fluxes and errors with variance weighted average and with MBE2.1_XSL.	Sep 27, 2023 6:09:30 PM
1071	943	memba	20A-R11-W2-07	FA over W2-KiDS until 20A without crosstalk correction (NightlyR11). Fixed circular apertures of diameter 3 arcsec, constant annulus from 30 to 45 pixels, Scatterlight correction regular (32-5), flagging scatterlight to 10x ratio. Star apertures fixed. With Mask production. Coadd fluxes and errors with variance weighted average and with MBE2.1_XSL.	Sep 24, 2023 11:00:35 AM



Continuum between data reduction and science pipelines.

- Proper data management can accelerate science.
- Engineering and science are two different cultures. This has implications on tools.

Big data in astronomy

Surveys, Projects	Short	Range	Information Volume
Digitized First Byurakan Survey	DFBS	opt	400 GB
Digitized Sky Survey (based on POSS)	DSS	opt	3 TB
Two Micron All-Sky Survey	2MASS	NIR	10 TB
Galaxy Evolution Explorer	GALEX	UV	30 TB
Sloan Digital Sky Survey	SDSS	opt	40 TB
SkyMapper Southern Sky Survey	SkyMapper	opt	500 TB
Panoramic Survey Telescope and Rapid Response System	PanSTARRS	opt	~40 PB
Large Synoptic Survey Telescope, *expected*	LSST	opt	~200 PB
Square Kilometer Array, *expected*	SKA	radio	~4.6 EB

https://combao.bao.am/2020/159-180.pdf

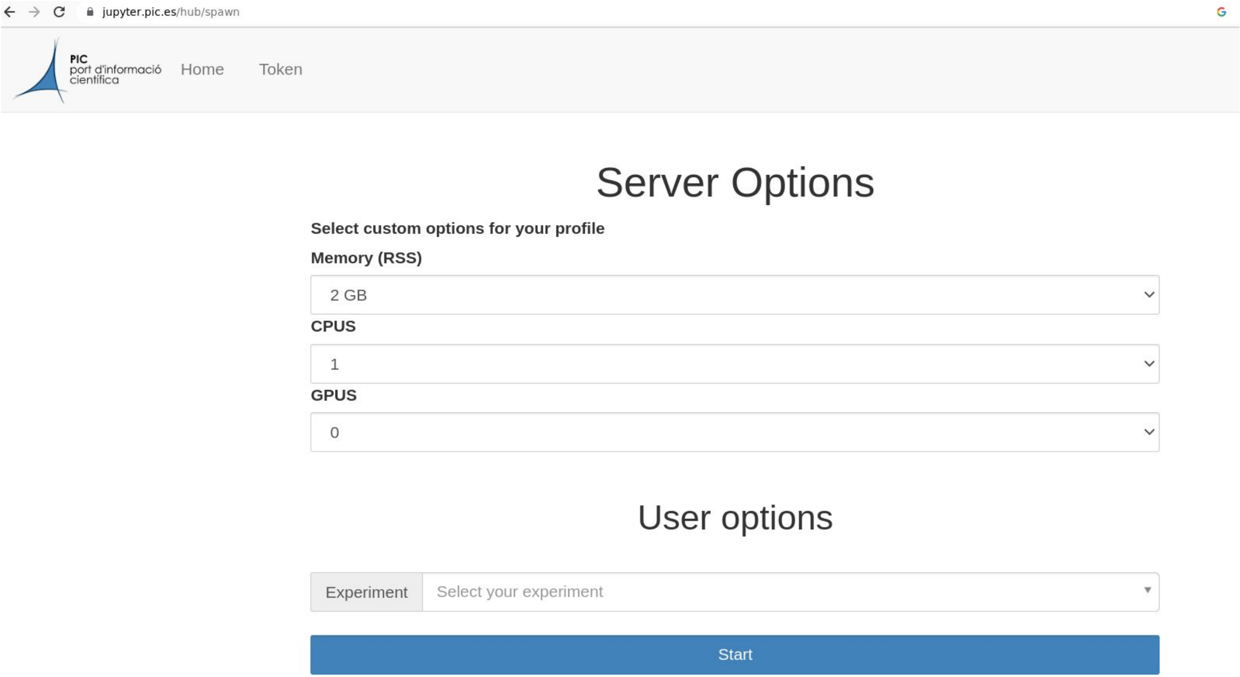
Astronomy has transitioned

Edwin Hubble

Euclid satellite. ESA mission, launched July 1st 2023, ~2500 scientists, 15+ countries. PIC is a science data center.

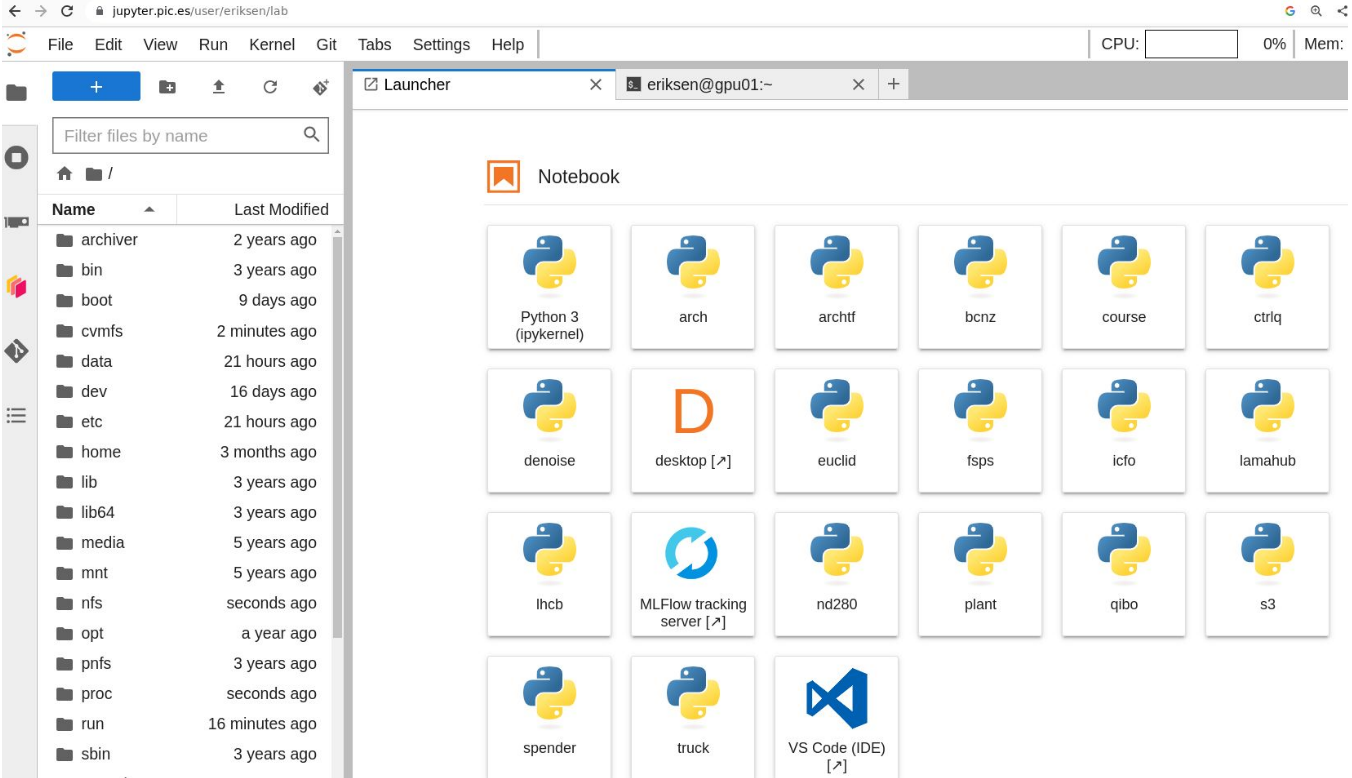
How to access large data volumes

Directly work in the datacenter. Usability is critical. We are using the service ourself.



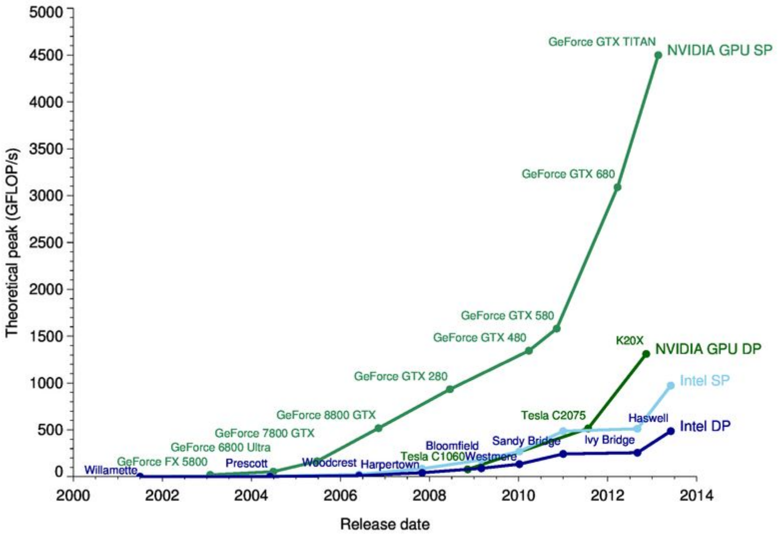
JupyterLab interface

Directly running at PIC. Free to install software yourself. I no longer have codes installed on my laptop.



GPUs for training

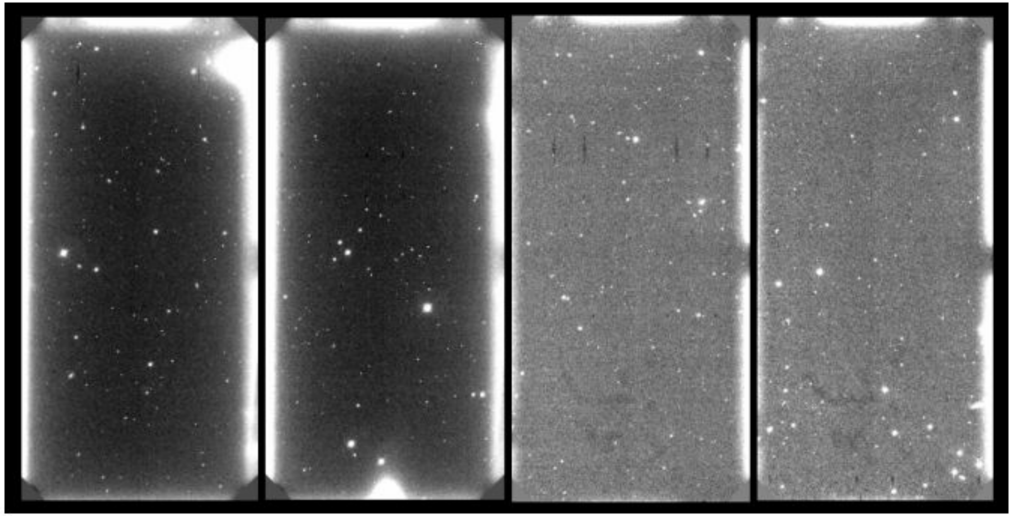
- GPUs are essential for training large networks.
- CUDA allows for using GPUs for general purpose calculations.
- At PIC we have 19 GPUs, 8 which are generally available (NVIDIA 2080Ti).
- Evolution in the robustness of the platform.
- GPUs can be expensive. Resource suitable for sharing.
- All users can request GPUs. No special permission need. Allow students and post.docs. to easily experiment with deep learning.



With some custom calculations, I had ~40 times better performance on a GPU than a CPU.

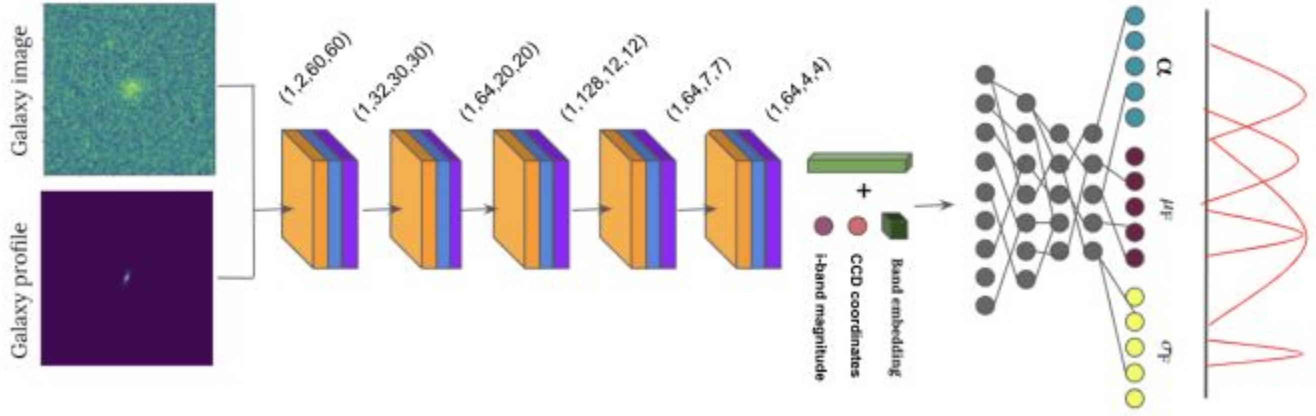
The PAU Survey: Background light estimation with deep learning techniques

Scattered light is an image artefact. Light has been leaking into the camera.



This paper introduces BKGnet, a deep neural network to predict the background and its associated error… On average,the use of BKGnet improves the photometric flux measurements by 7% and up to 20% at the bright end. (arxiv: 1910.02075)

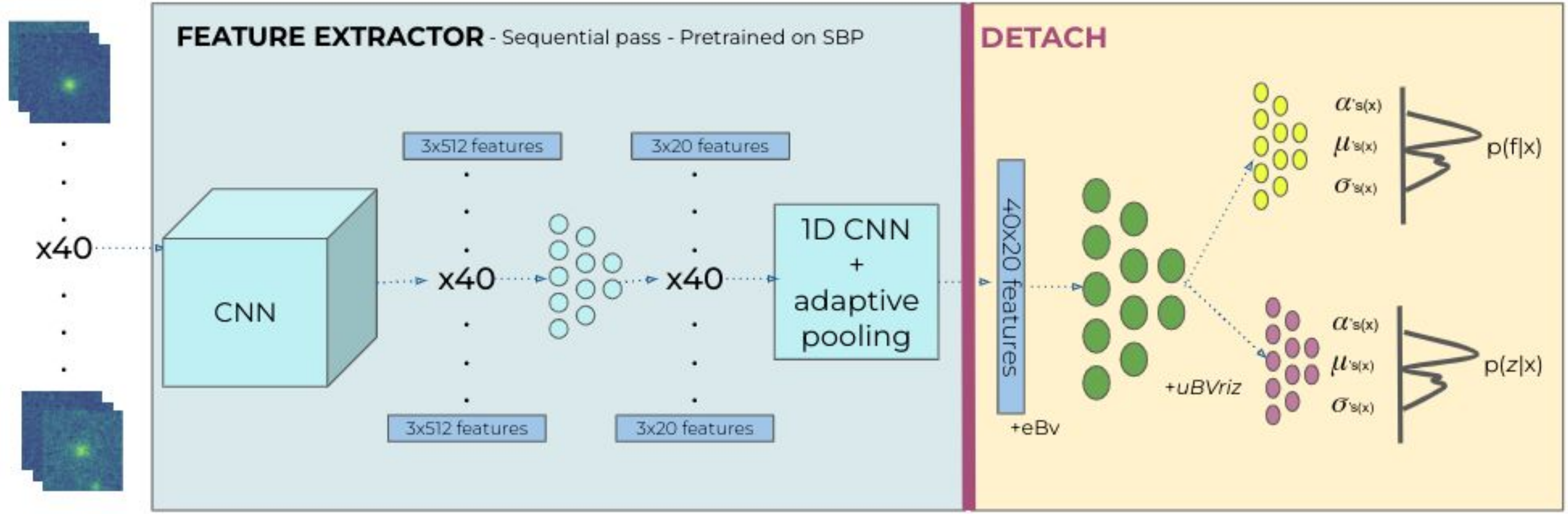
The PAU survey: Estimating galaxy photometry with deep learning



On average, Lumos **increases the SNR** of the observations by a factor of 2 compared to an aperture photometry algorithm. It also incorporates other advantages like **robustness towards distorting artefacts**, e.g. cosmic rays or scattered light, the ability of deblending and less sensitivity to uncertainties in the galaxy profile parameters used to infer the photometry (arxiv: 2104.02778)

End-to-end pipeline

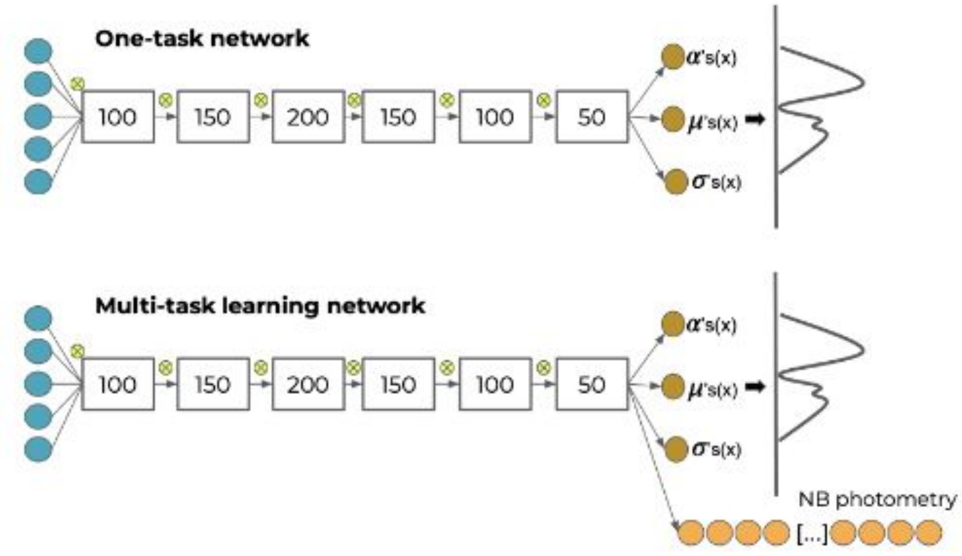
Background estimation -> Flux estimation -> Distance termination (photo-z)



Combined architecture including images from all narrow-bands. Directly produce the photo-z and the photometry. Almost there, but there is some technical issue with the correlation between the bands?

The PAU Survey and Euclid: Improving broadband photometric redshifts with multi-task learning

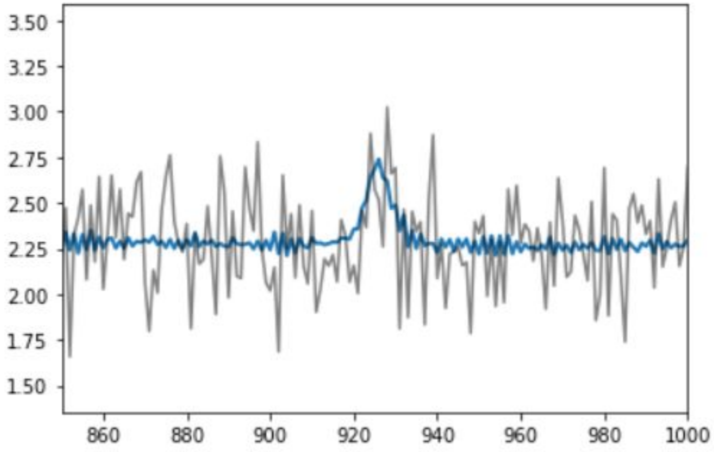
- Euclid is 15000 sq.deg. deep BB survey.
- PAUS is 50 sq.deg. narrow band survey.
- Can we use the overlap in certain regions to improve *Euclid* photo-*zs*.
- YES. Using multi-task learning. Joint publication.

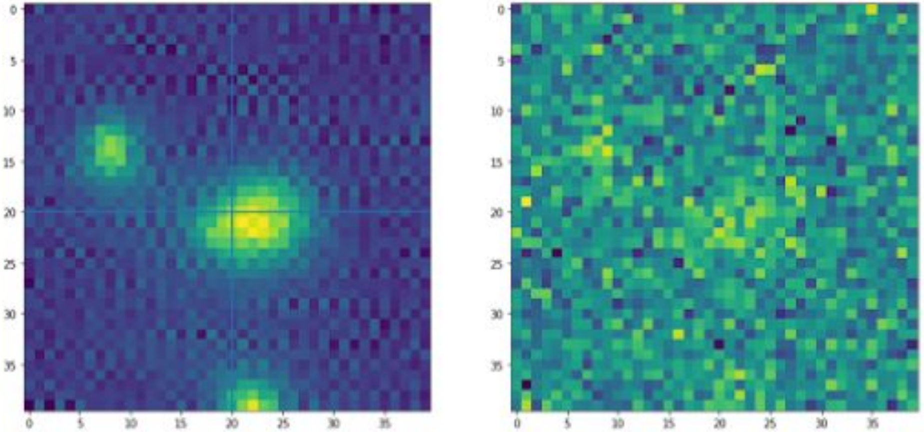


A&A October 4, 2023

Unsupervised denoising

- Allows for denoising without knowing the ground truth.





Galaxies after and before denoising.

PyData ecosystem

NumPy

Arrays

```
[2]: # The standard way to import NumPy:
     import numpy as np

     # Create a 2-D array, set every second element in
     # some rows and find max per row:

     x = np.arange(15, dtype=np.int64).reshape(3, 5)
     x[1:, ::2] = -99
     x
     # array([[  0,   1,   2,   3,   4],
     #        [-99,   6, -99,   8, -99],
     #        [-99,  11, -99,  13, -99]])

     x.max(axis=1)
     # array([ 4,  8, 13])

     # Generate normally distributed random numbers:
     rng = np.random.default_rng()
     samples = rng.normal(size=2500)
     samples

[2]: array([-1.26017669,  1.08057382, -0.67548957, ...,  0.89614109,
            -0.07415828,  0.40363923])
```

pandas

Tables

```
[4]: import numpy as np
     import pandas as pd

[6]: df2 = pd.DataFrame(
         {
             "A": 1.0,
             "B": pd.Timestamp("20130102"),
             "C": pd.Series(1, index=list(range(4)), dtype="float32"),
             "D": np.array([3] * 4, dtype="int32"),
             "E": pd.Categorical(["test", "train", "test", "train"]),
             "F": "foo",
         }
     )

     df2
```

[6]:

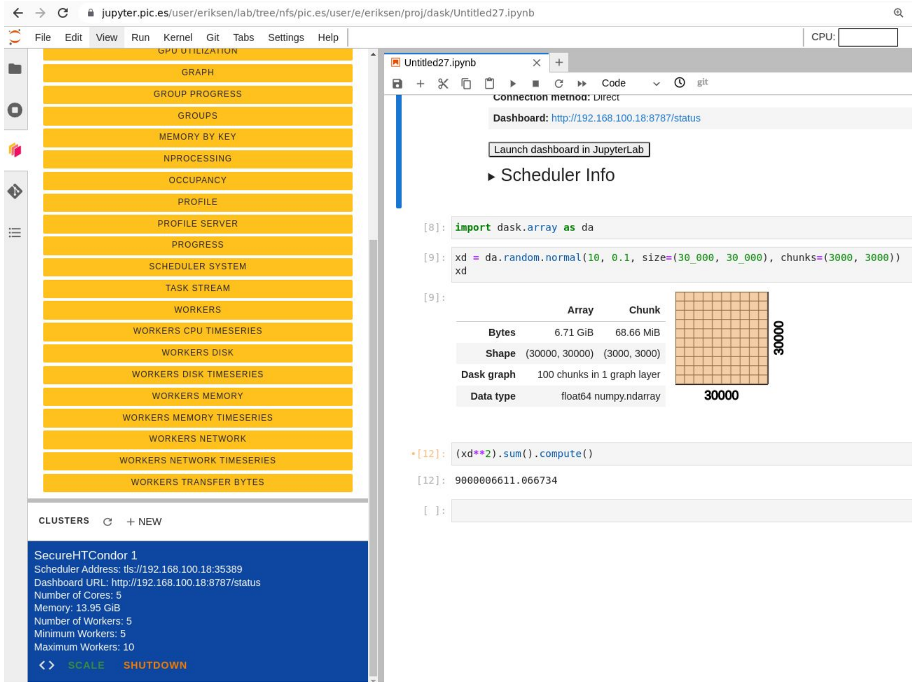
	A	B	C	D	E	F
0	1.0	2013-01-02	1.0	3	test	foo
1	1.0	2013-01-02	1.0	3	train	foo
2	1.0	2013-01-02	1.0	3	test	foo
3	1.0	2013-01-02	1.0	3	train	foo

Why Dask?

- **Scaling Python data analytics to multiple machines.**
- Numpy and Pandas was not intended to scale.
- Started at Anaconda in 2015.
- Can be used on your laptop.
- Partly as a reaction to Spark and the usage on JVM.
- Actually started as a single machine project.

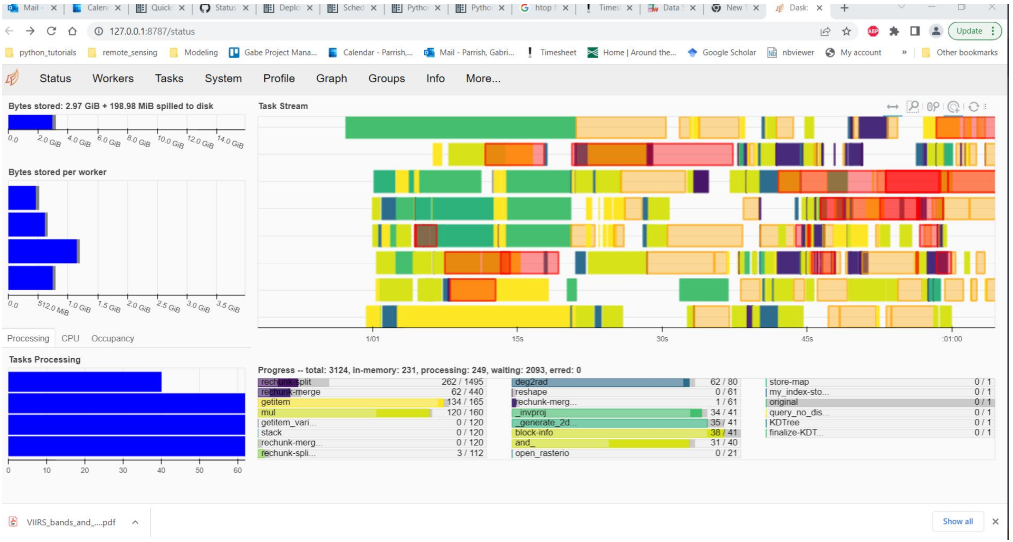
Dask integration

- Dask scales data science libraries like Numpy and Pandas to multiple machines.
- Based on low level task parallelism, allowing parallelization of custom codes.
- Integrated into Jupyter so the user can start and monitor a cluster from the GUI.
- The started cluster will request resources through HTCondor.
- The adaptive cluster size can scale up and down based on the workload.



Preprocessing of data

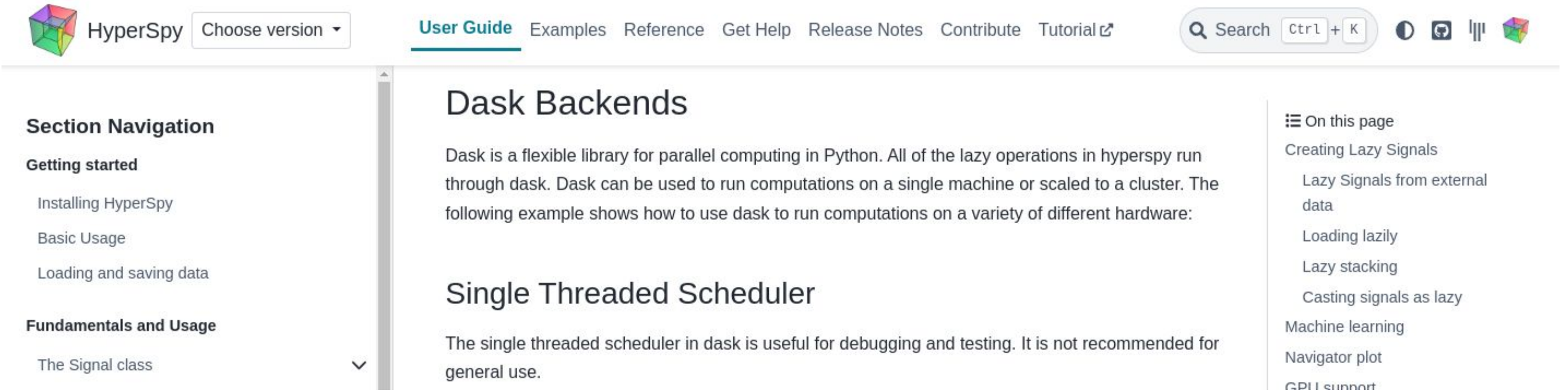
Nice monitoring of the progress of your job.



[Source]

Dask as a building block

- Extremely fun when it works.
- Can have a bit of a learning curve.
- Tools are built upon Dask.
- HyperSpy is having support for Dask.



New applied AI group

New scientific group at PIC from autumn 2022

Works on **deep learning** in different fields, aiming to **developing synergies**

Ongoing work in **cosmology**, **material science**, **bio imaging** and **quantum computing**

Collaboration or interactions with **theory, GW and neutrino groups**

Teaching of deep learning methods

Involved in **developing infrastructure**, like the **Dask** integration

Martin Eriksen
[Group leader]

Laura Cabayol
[Post. Doc.]

Antoni Alou
[Quantum technician]

+ 2 new PhD students [CSC grants]

Chen Jiefeng

Hanyue Guo

Arriving these days. Will work on deep learning for astronomy.

Chat-GPT hallucinations

M

Martin Børstad Eriksen

Fri, Sep 22, 12:09 AM (11 days ago)

Dear Ana and Lucia, What about: Title: "AI-Driven Material Science: InCAEM and the Port d'Informació Científica (PIC) Data Center" Abstract: Explore the fascina

A

Ana Belén Martínez Bonillo

Fri, Sep 22, 12:25 PM (11 days ago)

to me, Lucia

Spanish > English Translate message

Turn off for: Spanish

Thanks a lot Martin,

It's already published: https://indico.cells.es/event/1431/

You will see that I've corrected the name of InCAEM project (In Situ Correlative Facility for Advanced Energy Materials instead of Innovations in Computational Approaches for Emerging Materials).

Best wishes,

Ana

How to get a PIC account?

https://pic.es/register

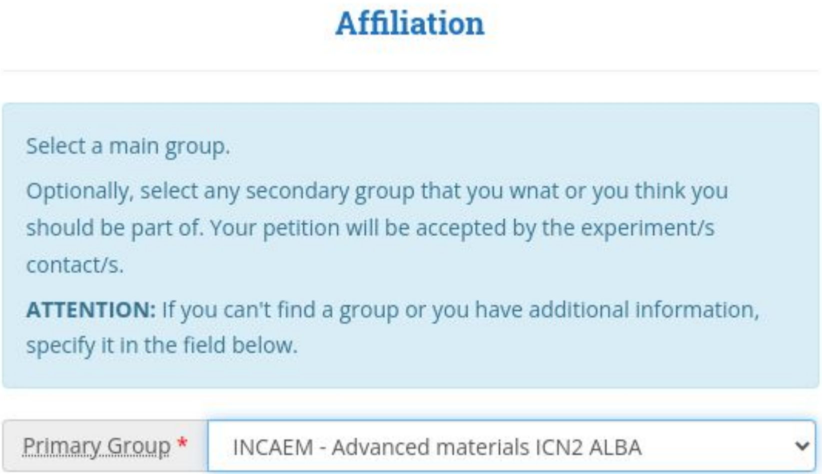

Sponsor Information

Name of Legal Entity *

Name of Contact Person *

Email of Contact Person *

Relationship with Entity * Employee


Affiliation

Select a main group.

Optionally, select any secondary group that you wnat or you think you should be part of. Your petition will be accepted by the experiment/s contact/s.

ATTENTION: If you can't find a group or you have additional information, specify it in the field below.

Primary Group * INCAEM - Advanced materials ICN2 ALBA

Normally add someone higher up on the food chain.

Multiple users from different groups

```
Last Priority Update: 10/2  23:23
Group                                          Config     Use    Effective   Priority  Wghted Total Usage  Time Since  Weighted  Submitter
  User Name                                     Quota   Surplus   Priority    Factor   In Use (wghted-hrs) Last Usage Requested   Ceiling
-------------------------------------------- --------- ------- ------------ --------- ------ ------------ ---------- ---------- ---------
group_HIGHPRIO                                    0.02 Regroup                1000.00     12    751473.25      <now>         14
  ops_score.ops002@pic.es                                              0.50      1.00      0      1833.63    0+00:04
  atsgm_score.atsgm001@pic.es                                          0.50      1.00      0      3956.87    0+00:10
  jupyter.ai_score.eriksen@pic.es                                      0.53      1.00      1       252.41    0+00:03
  jupyter.neutrinos_score.mrodrigu@pic.es                              0.55      1.00      1       445.63      <now>
  jupyter.agn_score.jalcolea@pic.es                                    0.56      1.00      1       432.35      <now>
  sgmcm_score.sgmcm001@pic.es                                          0.58      1.00      0     34060.25    0+00:06
  jupyter.herd_score.lfarinaa@pic.es                                   0.59      1.00      1      1090.37      <now>
  jupyter.desi_score.lcabayol@pic.es                                   0.61      1.00      1      1460.05      <now>
  jupyter.incaem_score.joton@pic.es                                    0.67      1.00      1        17.13      <now>
  jupyter.cms_mcore.pserrano@pic.es                                    0.74      1.00      4       269.81      <now>
  jupyter.ifaecovid_score.bosman@pic.es                                0.82      1.00      1      6019.63      <now>
  jupyter.ai_mcore.eriksen@pic.es                                      0.84      1.00      0       246.35    0+00:02
  jupyter.incaem_mcore.fsaiz@pic.es                                    2.11      1.00      2      1487.71      <now>
```

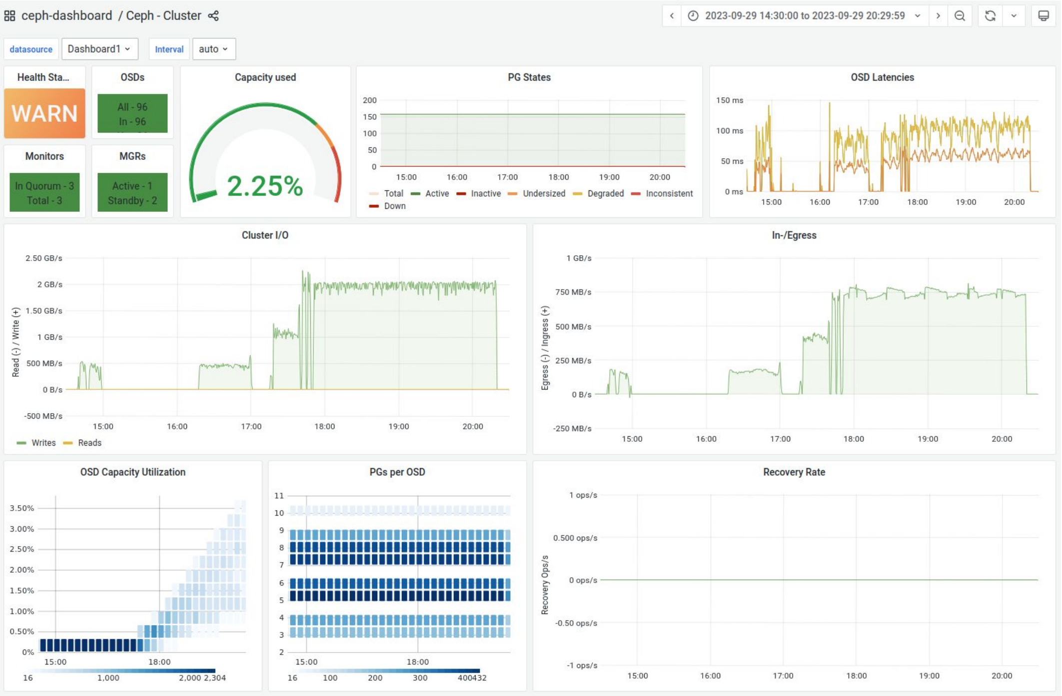
Some projects

- Neural network has theoretically been around for some time, but has returned.
- PAUS has unique data.
- Determined redshifts combining simulations and data.
- Remove background noise (CNNs).
- Estimate galaxy fluxes (CNNs).
- Combined flux and photo-z estimation.
- Multi-task learning for Euclid + PAUS.
- Denoising of images and spectra.
- + More not shown here

Any questions?

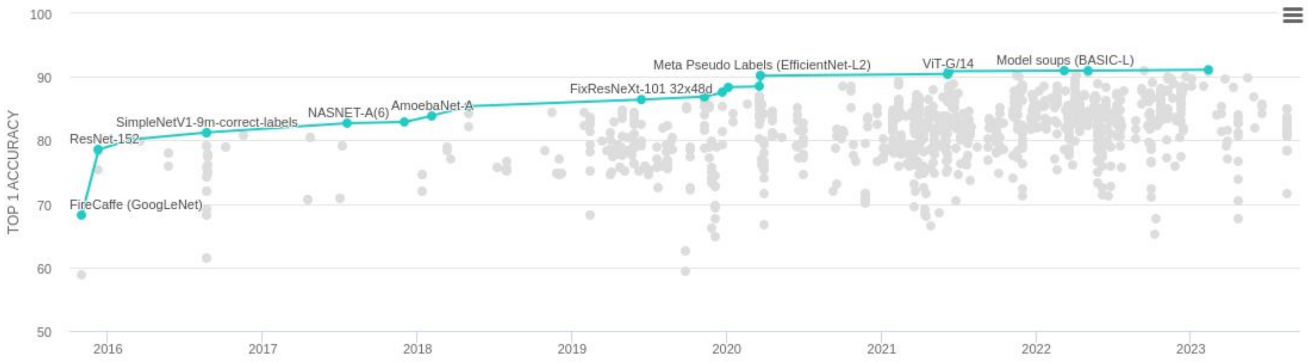

Clipdrop
by stability.ai

Performance tuning

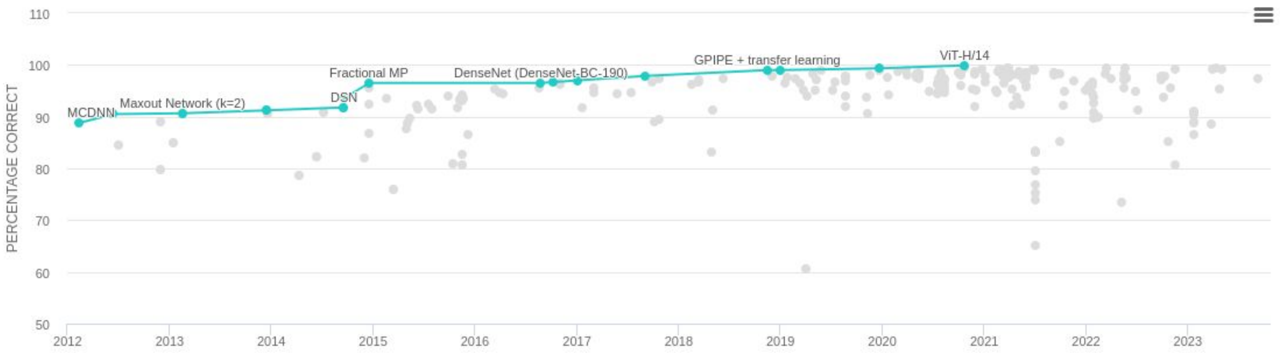


Improvements over time

Imagenet:



Cifar10



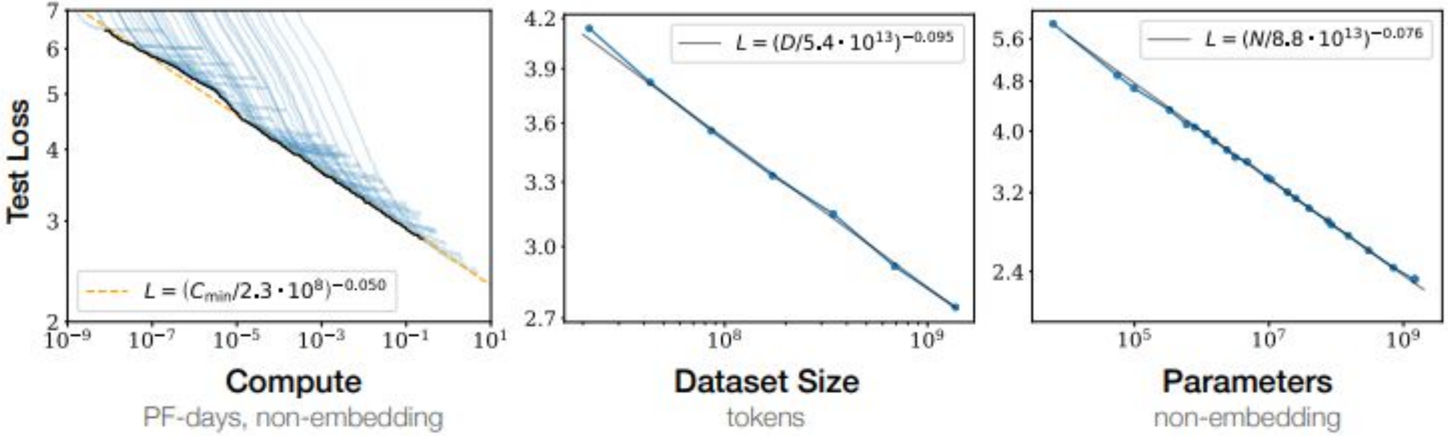


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.