ML and AI in Particle Physics and Astronomy: Some Recent Examples and Next Steps

Sascha Caron (RU and Nikhef)

disclaimer

- This is not all we do:

→Great paper by Jochem Kip and Zhongyi on “Neutrino signals from DM” and Neutrino-DM interplay → next seminar?

→Work in ATLAS on 4-top and triggering…

Scientific discovery: The 5th paradigm ?

- **First paradigm : Empiric / Observation**
- **2nd paradigm : Theoretical Models (analytically solvable?)**

Scientific discovery: The 5th paradigm ?

- **First paradigm : Empiric / Observation**
- **2nd paradigm : Theoretical Models (analytically solvable?)**
- **3rd paradigm (1970s): Simulation / Numerical computation**

Scientific discovery: The 5th paradigm ?

- **First paradigm : Empiric / Observation**
- **2nd paradigm : Theoretical Models (analytically solvable?)**
- **3rd paradigm (1970s): Simulation / Numerical computation**
- **2018 The Fourth Paradigm: Data-Intensive Scientific Discovery (Jim Gray), Data and Machine Learning -→ but also hype**
- https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery

Scientific discovery: The 5th paradigm ?

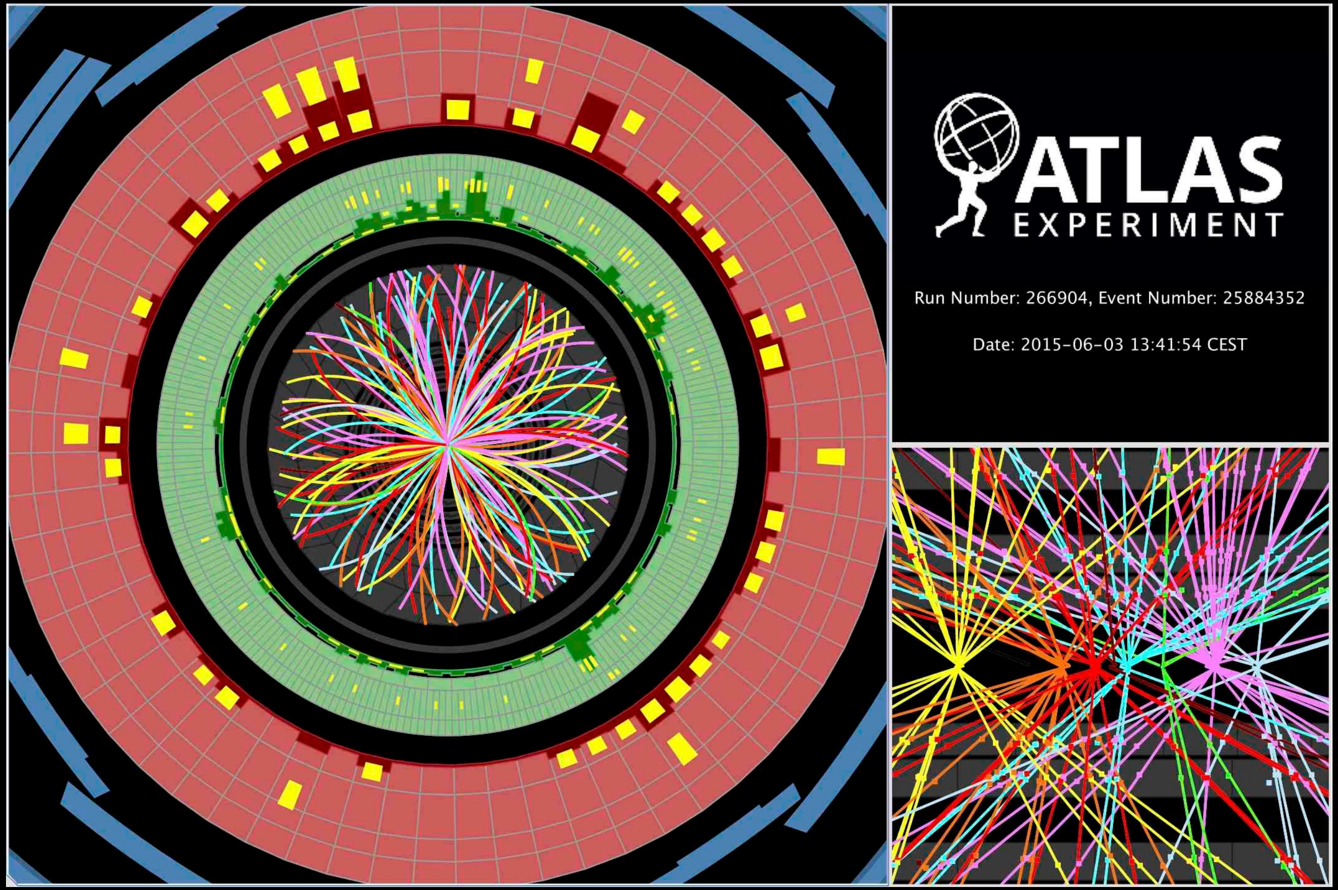
- **First paradigm : Empiric / Observation**
- **2nd paradigm : Theoretical Models (analytically solvable?)**
- **3rd paradigm (1970s): Simulation / Numerical computation**
- **2018 The Fourth Paradigm: Data-Intensive Scientific Discovery (Jim Gray), Data and Machine Learning -→ but also hype**
- https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery
- **June 2022: AI4Science to empower the fifth paradigm of scientific discovery (Christopher Bishop, QFT thesis), AI trained on scientific simulators (**machine learning, quantum physics, computational chemistry, molecular biology, fluid dynamics, software engineering, and other disciplines)
- https://www.microsoft.com/en-us/research/blog/ai4science-to-empower-the-fifth-paradigm-of-scientific-discovery/

Some of examples of our work in 2021-2022

LHC searches - New approach: Hyperclass of models

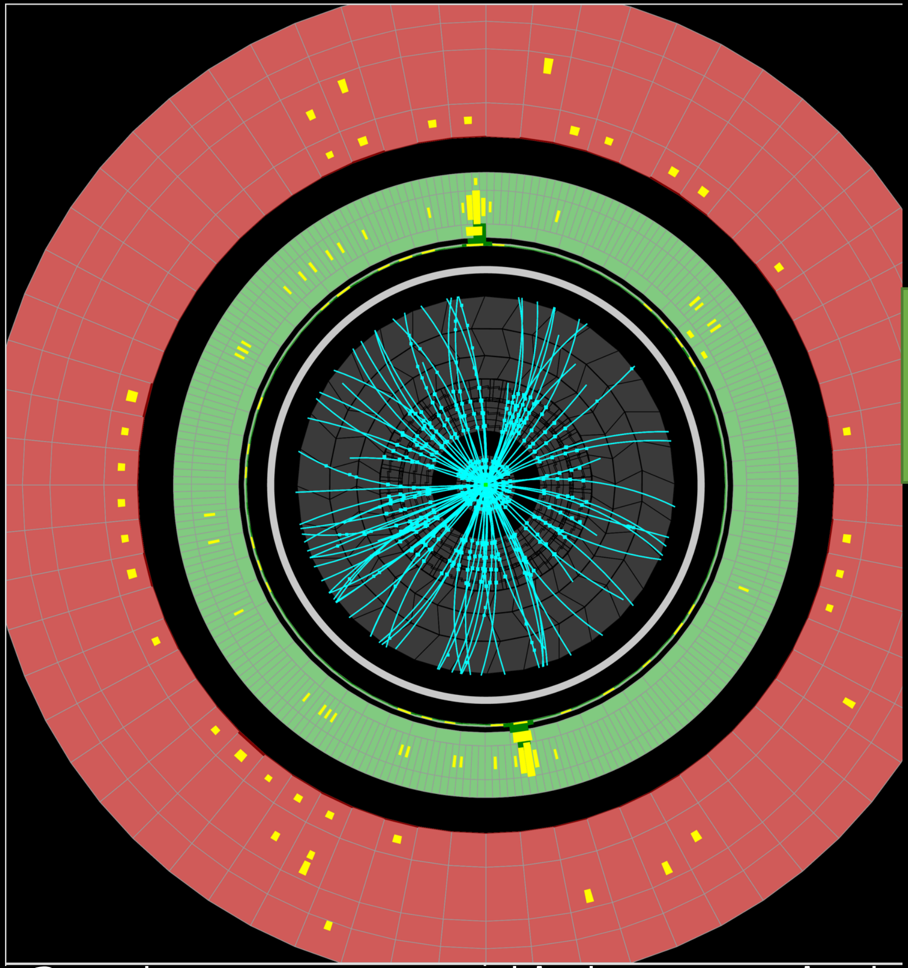
→ **Learn the simulated density function of our combined prejudice (or prior) of new physics**

Most events look like this…



Event from LHC run-2

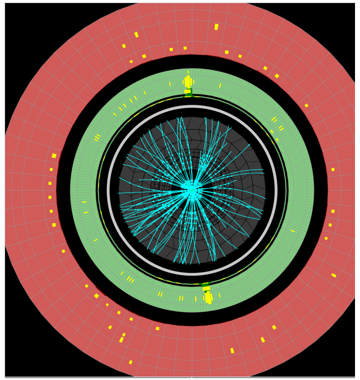
1 in >1000 billion events looks like this



Mass of the Higgs is reconstructed with photon energies

Higgs to 2 photon candidate with mass of 125 GeV

Traditional approach Model driven



1. Pick a model of new physics
2. Simplify
3. Pick a likely (?) set of parameters
4. Make a prediction ➔ p_BSM(x)
5. Train classifier (p_BSM(x) vs p_SM(x)) to test the prediction
6. Hypothesis test with data|old model vs data|new model on classifier output
7. Exclude the model parameter point ?
8. Go to 3 or 1

Idea: **Extend model-by-model supervised search for new physics**

What can we change / improve ?

Found 3 more directions (are there more?):

→Look systematically in all data for new physics (<u>brute force</u>) -> 2018 Jeroen in ATLAS

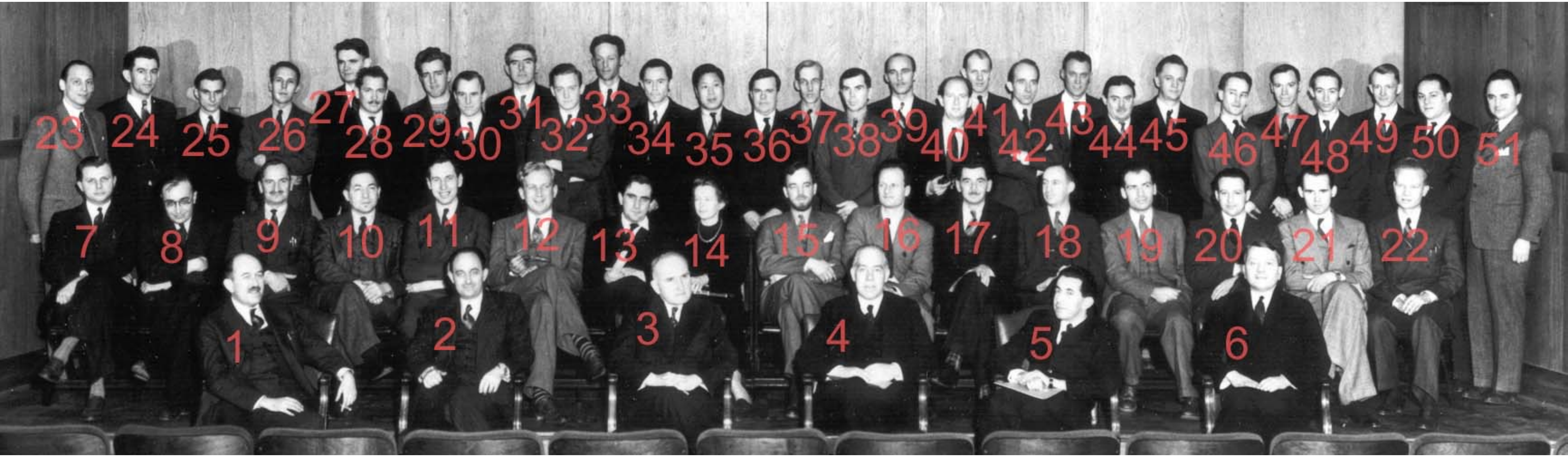
→Hyper-class augmentation: Train a ML classifier on <u>many simulated models</u> of new physics → 2022 Zhongy, Roberto

→Anomaly detection: Train ML classifier <u>only on known physics -> </u>5 pheno since 2019, work in ATLAS with Polina and Clara

Search via “Hyperclass: Mixture of theories”

Assume the model/parameter set is not the correct one, but includes some knowledge about the new phenomenon we expect in the data..

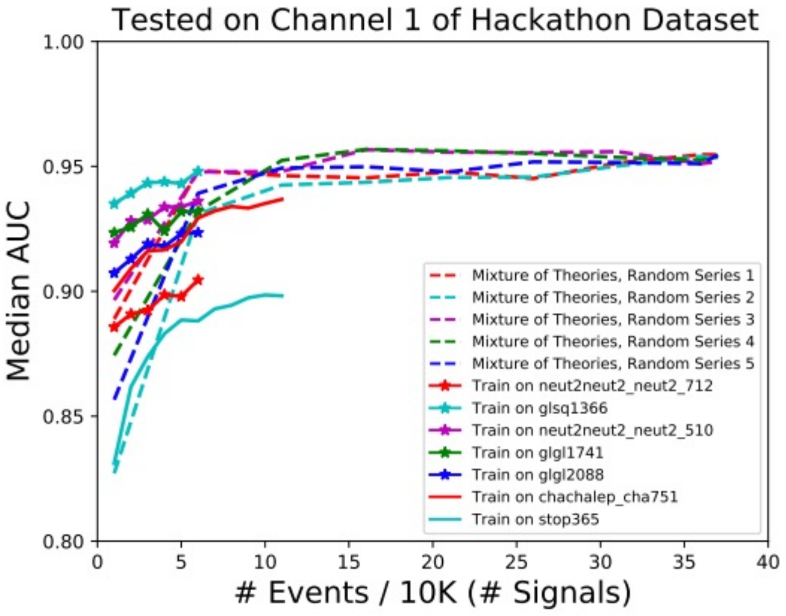
Maybe we should **mix the knowledge of the theory community**.

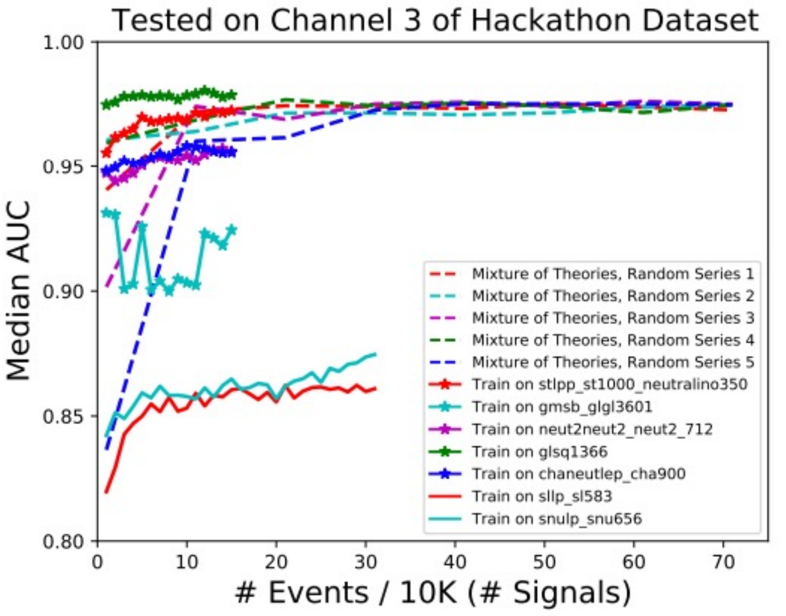


https://home.gwu.edu/~kargaltsev/HEA/washington-conferences.html

Our approach Model driven

1. Pick many "model of new physics"
2. Pick many likely (?) sets of parameters!
3. Make many predictions
4. Mix them
5. Train a classifier (NN, BDT) on $\sum_i^N w_i p_{S,i}(x)$ vs p_SM(x)
6. Hypothesis test in signal region data|SM





Mixture theories outperforms “on average” compared to single theory training

→ See later for comparison with other approaches

With Zhongyi Zhang, Roberto di Austri

Search for new physics at LHC/ATLAS

Anomaly detection: Out of distribution

→ Learn the density function of the physical simulator

Anomaly detection

1. Pick **no** “new physics model"
2. Learn the background model
3. Train ML classifier to test the prediction (is event background or not?)
4. Hypothesis test with data|background model on classifier output
5. Exclude the background model?

In which variable should you search?
Need a variable to "flag" an outlier

Is the data in the simulation ?

- Autoencoder:

data → Simulation^-1 → code → Simulation → data'

→Is data = data' or distance in latent space from target

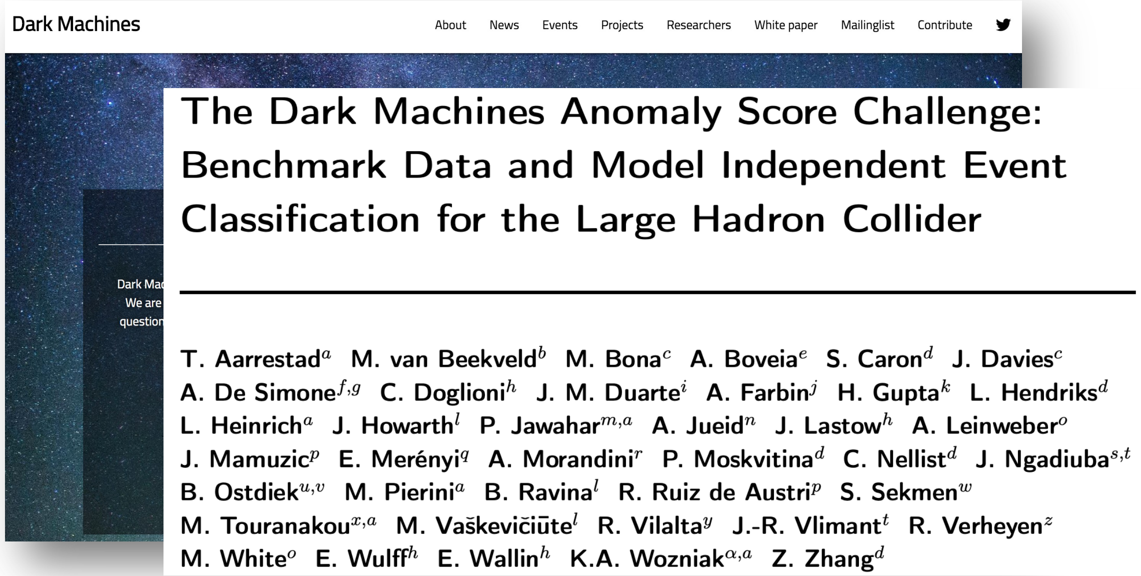
→Is this a good question ?

→Is this the best approach ?

→Comparison

Comparisons of approaches

Darkmachines (www.darkmachines.org) anomaly score challenge:
Objective → compare different approaches to define an "event- by-event" anomaly score



Event data:
4-vectors, jets, leptons, charge, photons

Different to
LHC Olympics (full signal and bump hunting / density comparisons with a few signals + background expectation)

Results (on arxiv https://arxiv.org/abs/2105.14027)

Contact persons: Comparisons: B. Ostdiek (bostdiek@g.harvard.edu)
Datasets: M. van Beekveld (melissa.vanbeekveld@physics.ox.ac.uk)

Compared performance of **>20 methods** to define anomalies
With > 1000 hyperparameter settings (i.e. algorithms to define anomalies)
Using
>20 signals
Using
> **1 Billion LHC events**
Using
A secret dataset (labels are still blind, *only Melissa van Beekveld (Oxford) knows*)

Task: Classify 100000s of events as SM or not by assigning **a score between 0 and 1…**

Figure of merit: **By how much can we improve the significance for that signal i.e. Significance Improvement SI per signal**

Organizers:
C. Doglioni, M. Pierini, S.C

$$\sigma'_S = \frac{S'}{\sqrt{B'}} = \frac{\epsilon_S S}{\sqrt{\epsilon_B B}} = \frac{\epsilon_S}{\sqrt{\epsilon_B}}\sigma_S \quad \Rightarrow \quad \mathrm{SI} \equiv \frac{\epsilon_S}{\sqrt{\epsilon_B}};$$

Many signals
many algorithms
many channels

Charginos

Gluinos

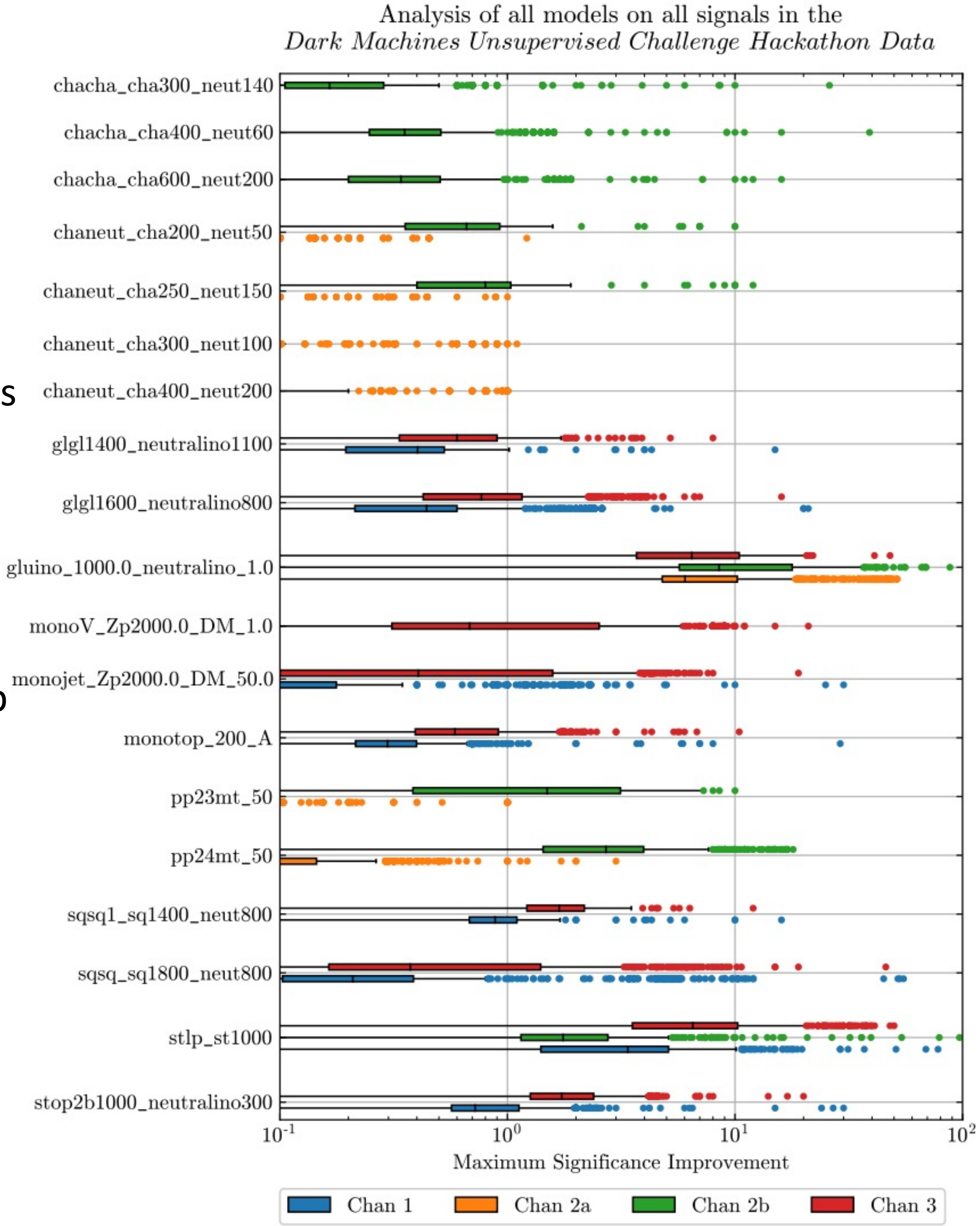
Zprime

Monotop

Squarks

Stops

....

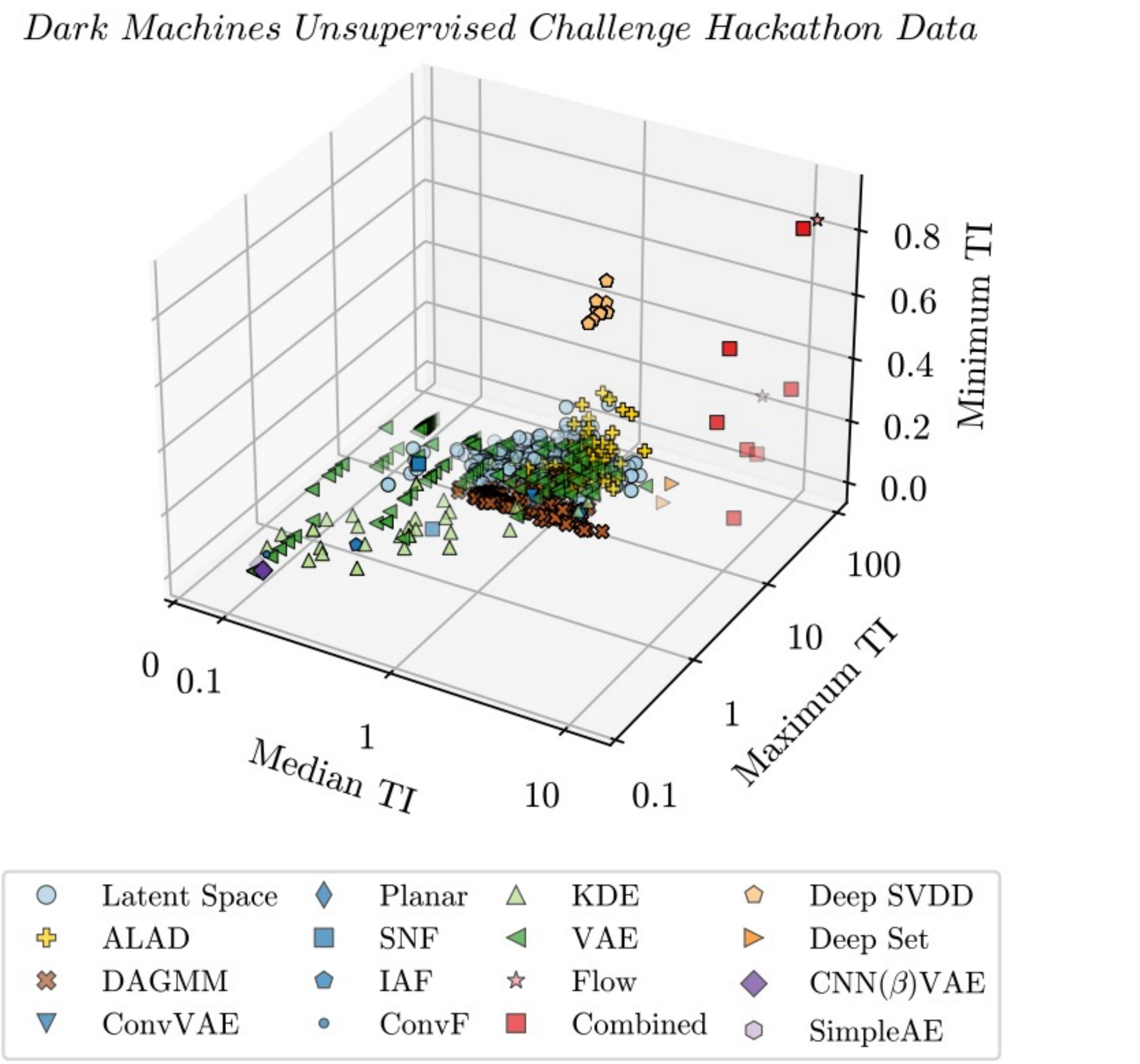


Summary plot

TI =
Total Improvement. (over many signals)

(median, max and min Improvement of many toy signals)

→ **Good algorithms have large max, min and mean TI**



https://arxiv.org/abs/2105.14027

Summary plot

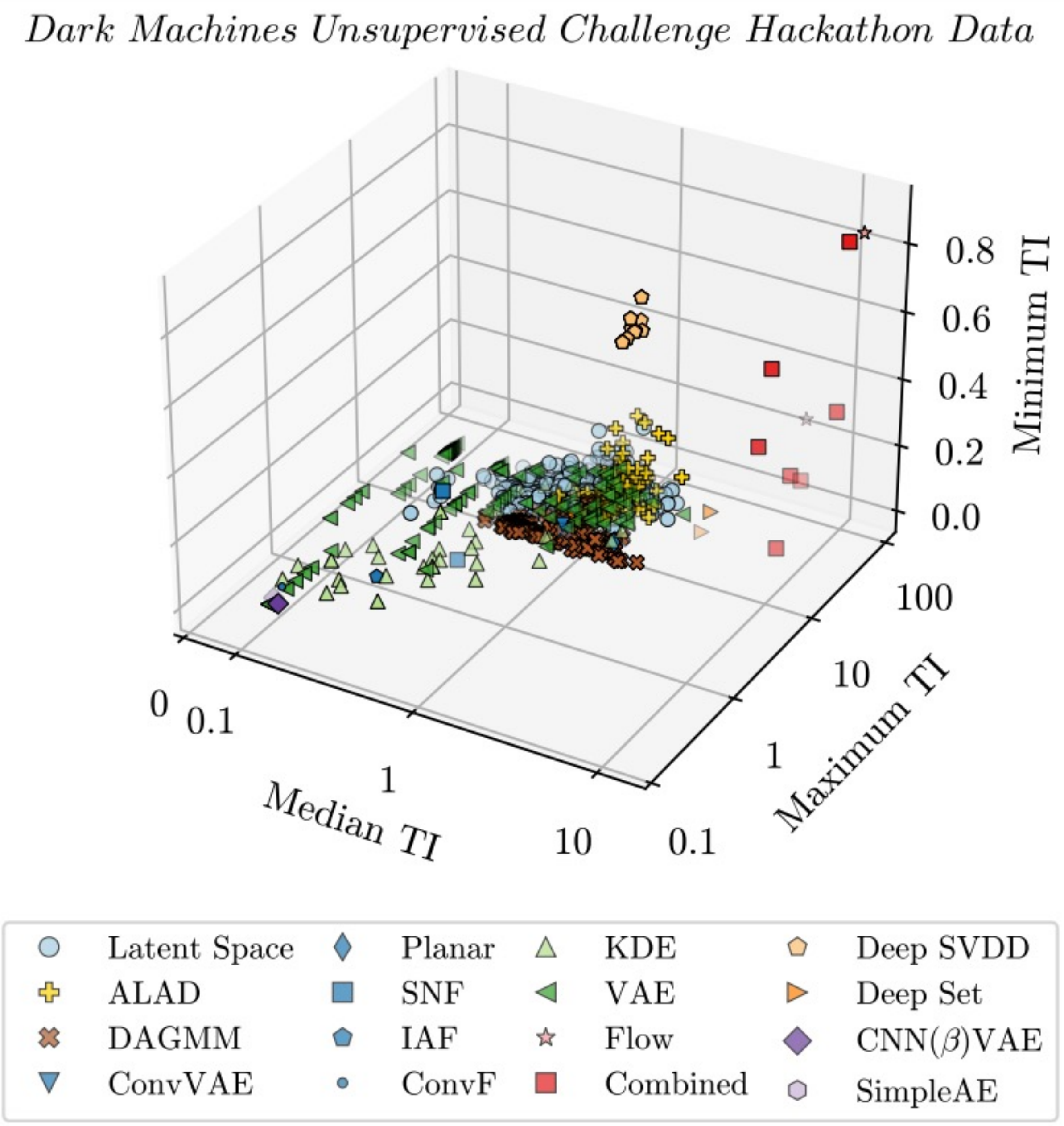
TI =
Total Improvement. (over many signals)

(median, max and min Improvement of many toy signals)

→ **Good algorithms have large max, min and mean TI**

→ ***DeepSVDD, Flow , Combined, DeepSets***
largely outperform traditional approaches (e.g. KDE), but also all autoencoder and VAEs !!

Why ? --> decoder seems not to be needed!



https://arxiv.org/abs/2105.14027

Rare and Different

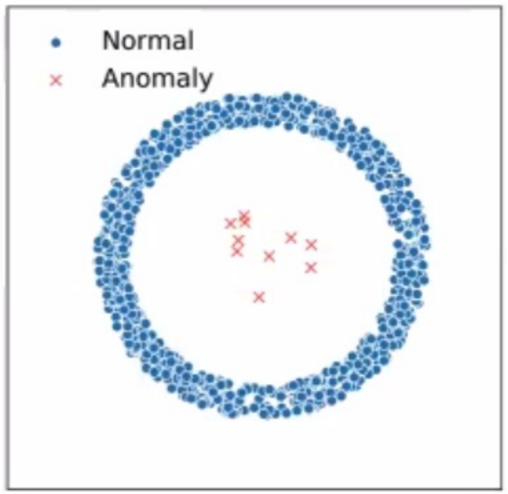
Idea:

Anomalies can be either rare, meaning that these events are a minority in the normal dataset, or different, meaning they have values that are not inside the dataset.

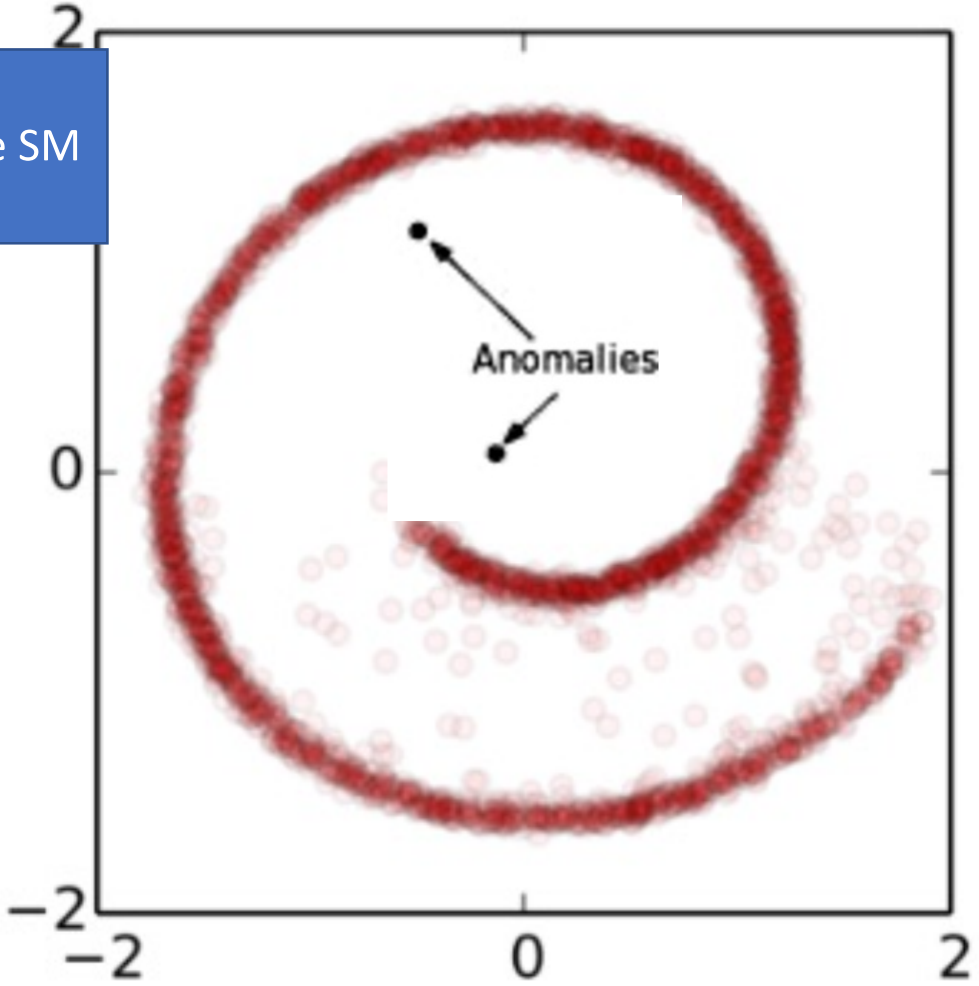
We **quantify and combine these two properties/objectives**

https://arxiv.org/abs/2106.10164

Rare → Density estimation



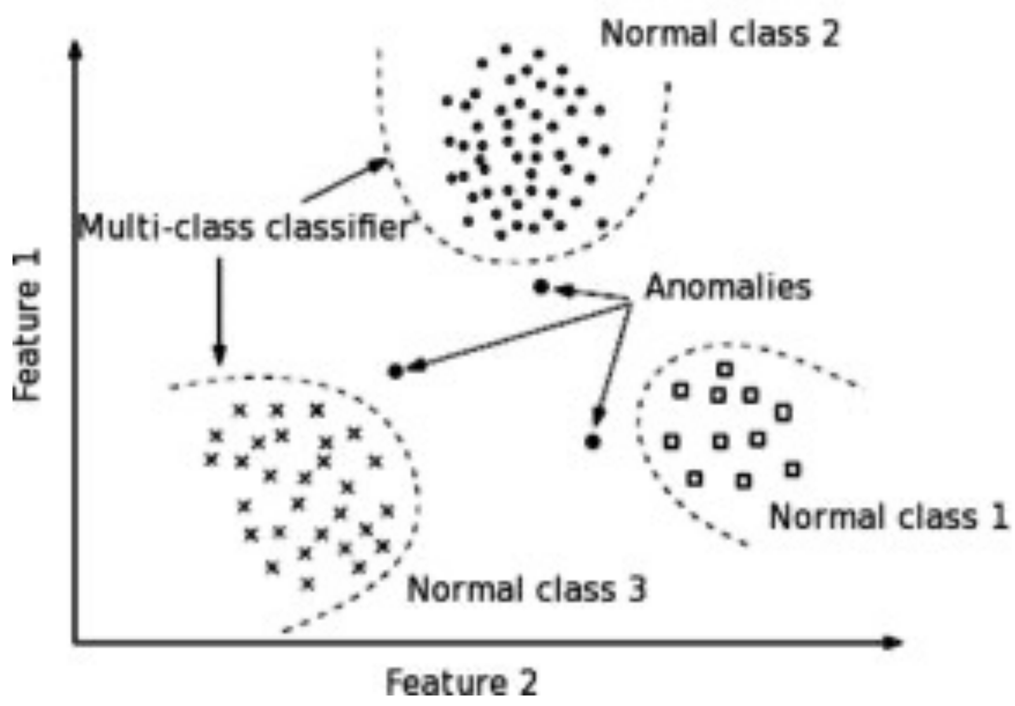
Idea:
Signal region is region outside the SM /simulation



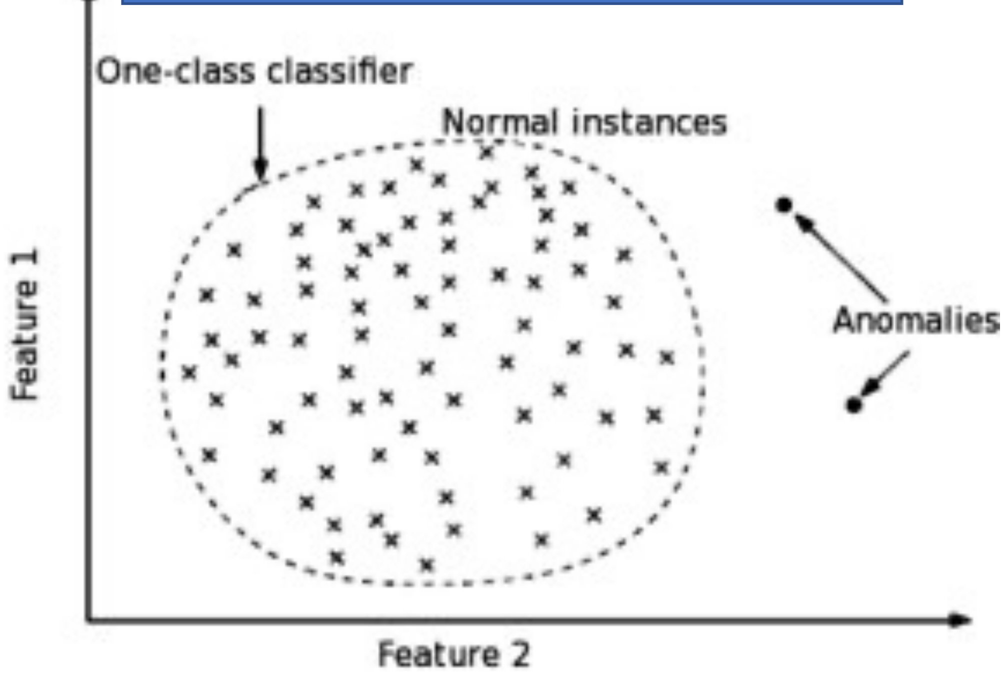
Series of paper on flow models from RU :
- Bob and Rob Verheyen 2021
- Luc Hendris, SC, Rob Verheyen, 2021
- Rob Verheyen : Surjective normalizing flows work even,

better as anomaly detectors…
→ https://inspirehep.net/literature/2077178

Different ? One class classification



Idea:
Signal region is region outside the SM



Different? Deep SVDD

Alternatively one could try to pass the events through a trained "filter" that only allows events to pass if they belong to the training data

Here: Deep SVDD

X → Network → 42

Anomaly score:

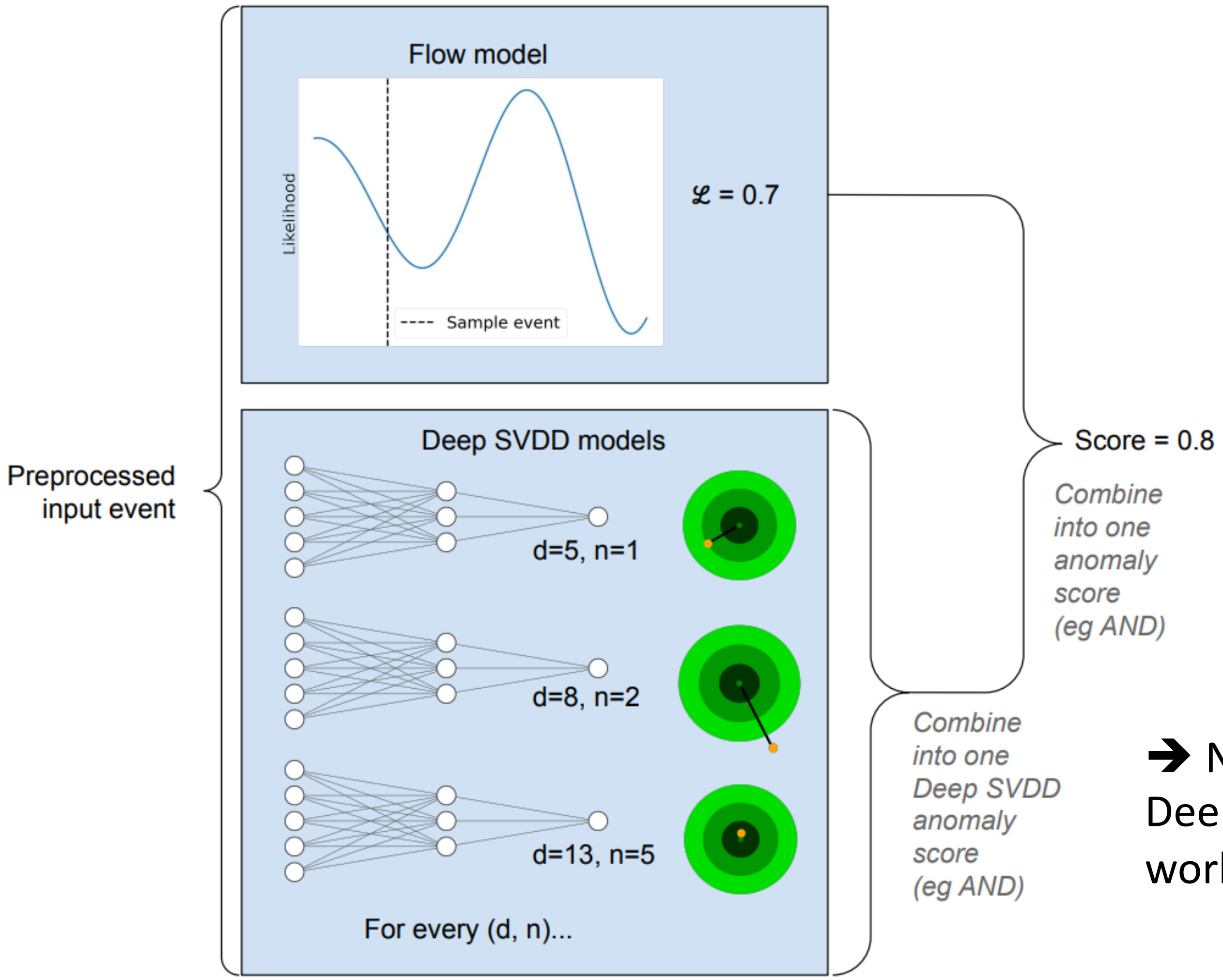
Difference from 42 !

The Deep SVDD network is similar to the encoder component of an autoencoder. The loss is defined as

$$s(x) = O_n^d - \mathrm{Model}(x), \tag{3}$$

where the model maps the input x to the same tensor shape as the manifold O. In our case, O is a vector of identical scalar values, with the subscript n defining the scalar value and superscript d the number of elements in the vector. For example, O_3^4 identifies the vector $(3, 3, 3, 3)$. The optimisation of the Deep SVDD model is fundamentally very simple: it is a NN that receives some input x and transforms it to some output O_n^d.

Rare and Different



➔ Need an ensemble of Deep SVDDs to make it work

[Luc Hendriks, Rob Verheyen, SC: https://arxiv.org/abs/2106.10164](https://arxiv.org/abs/2106.10164)

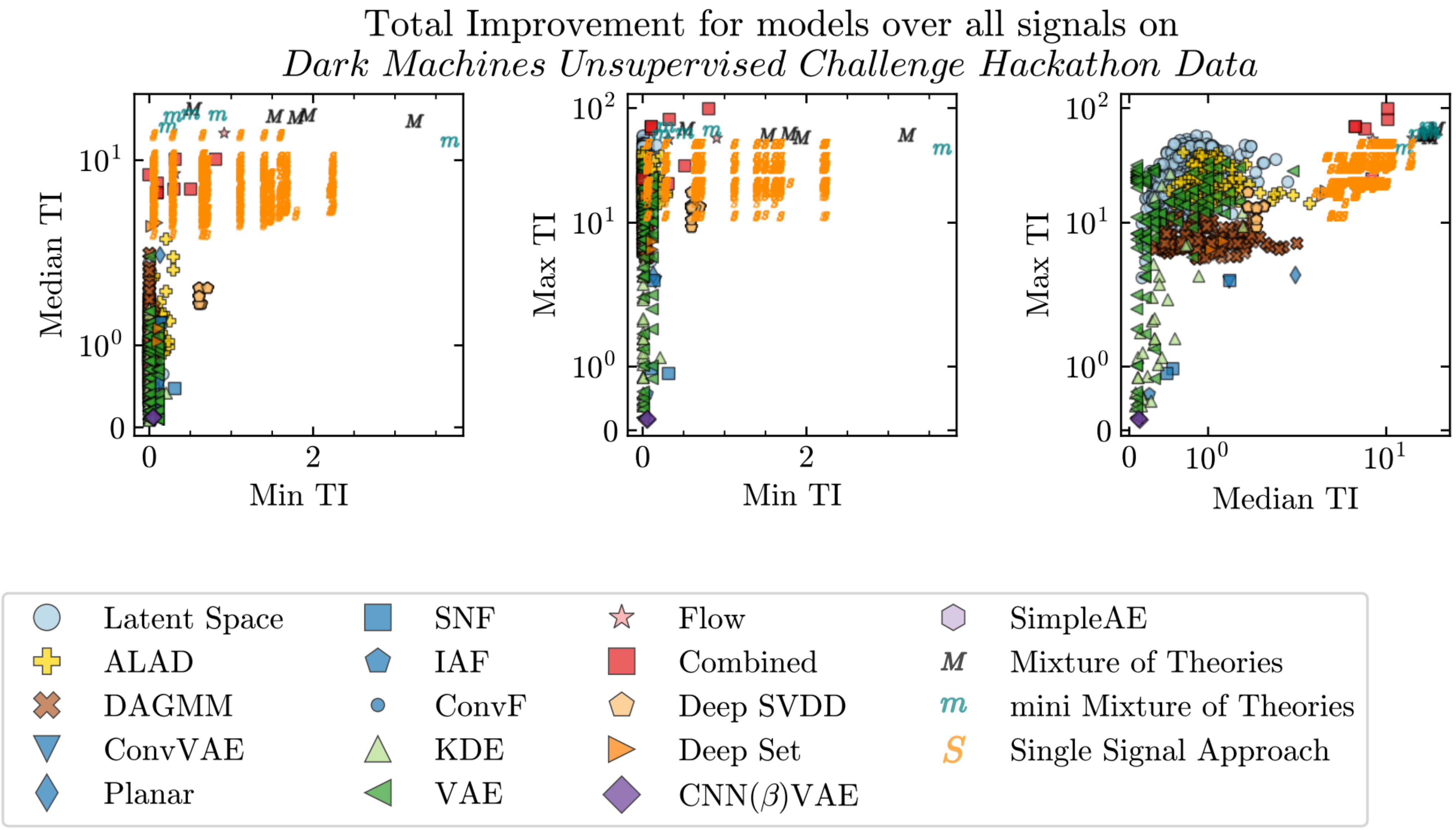
Compare them all (besides brute force)

Compare them all

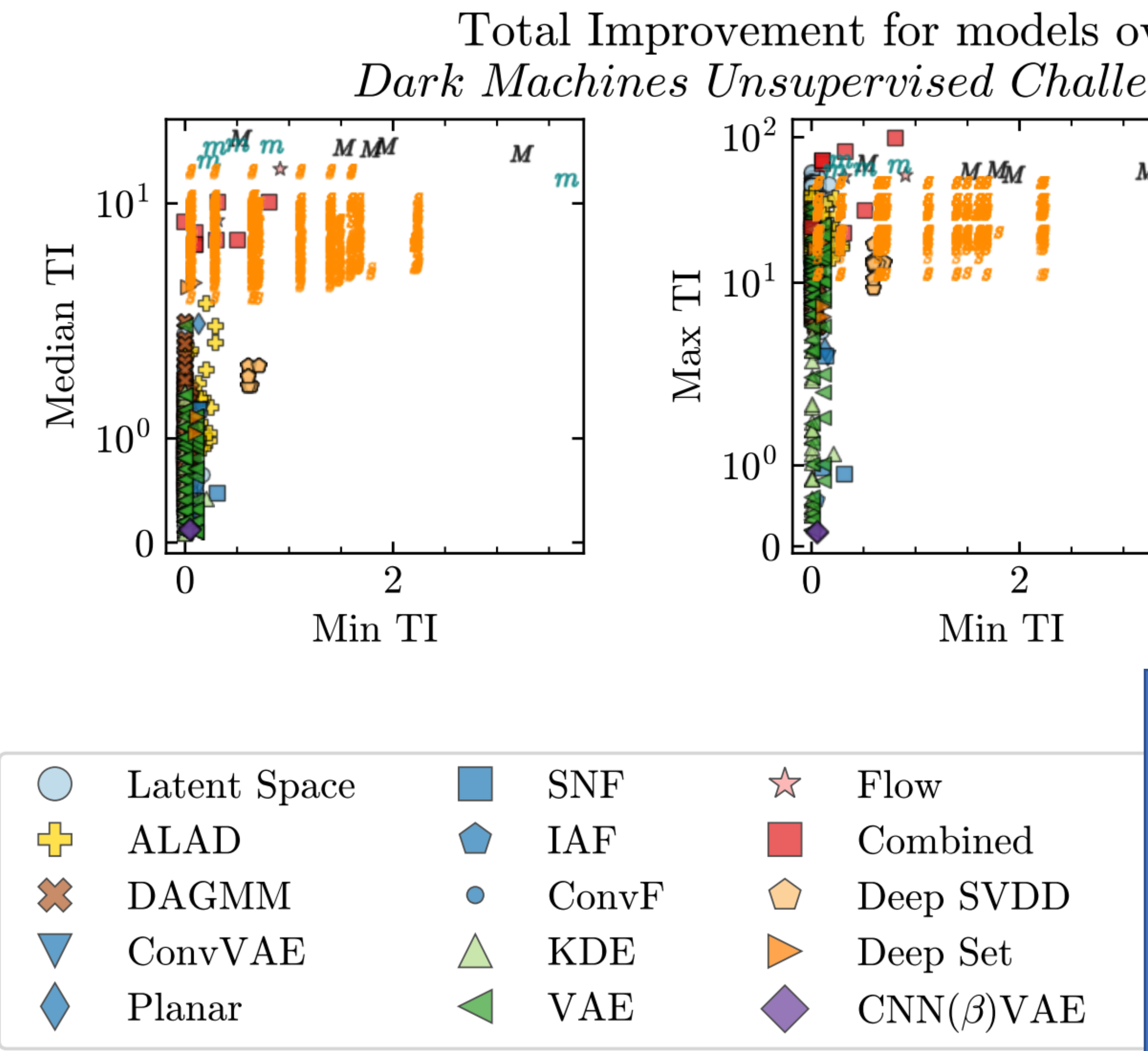
Compared:

- Supervise approaches (100s trained on different “single” signals)
- Mixture of Theory approach
- Unsupervised approaches

Who wins?

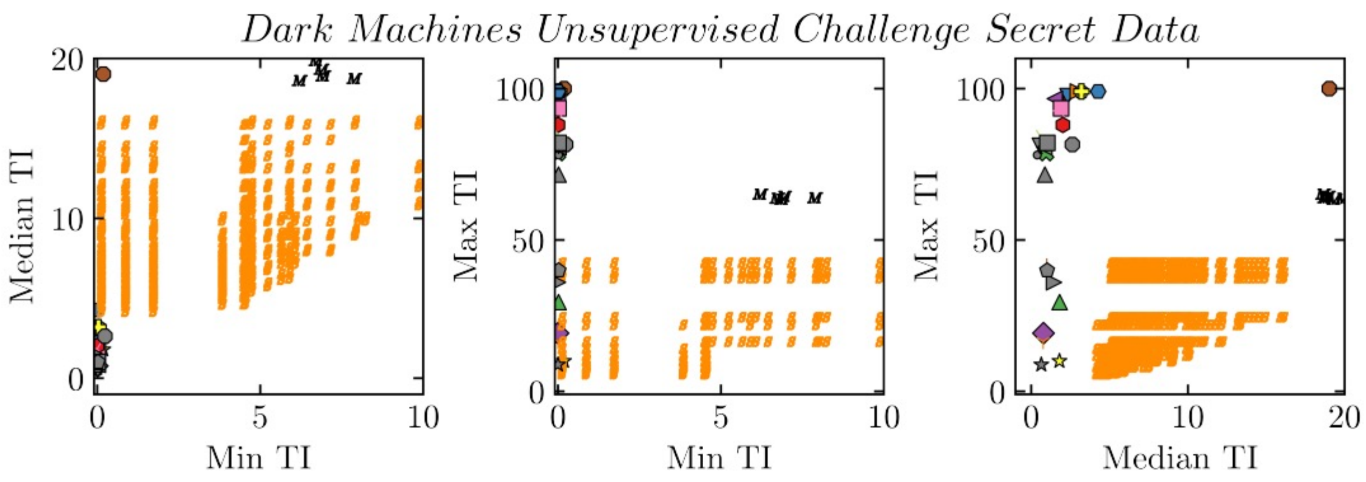


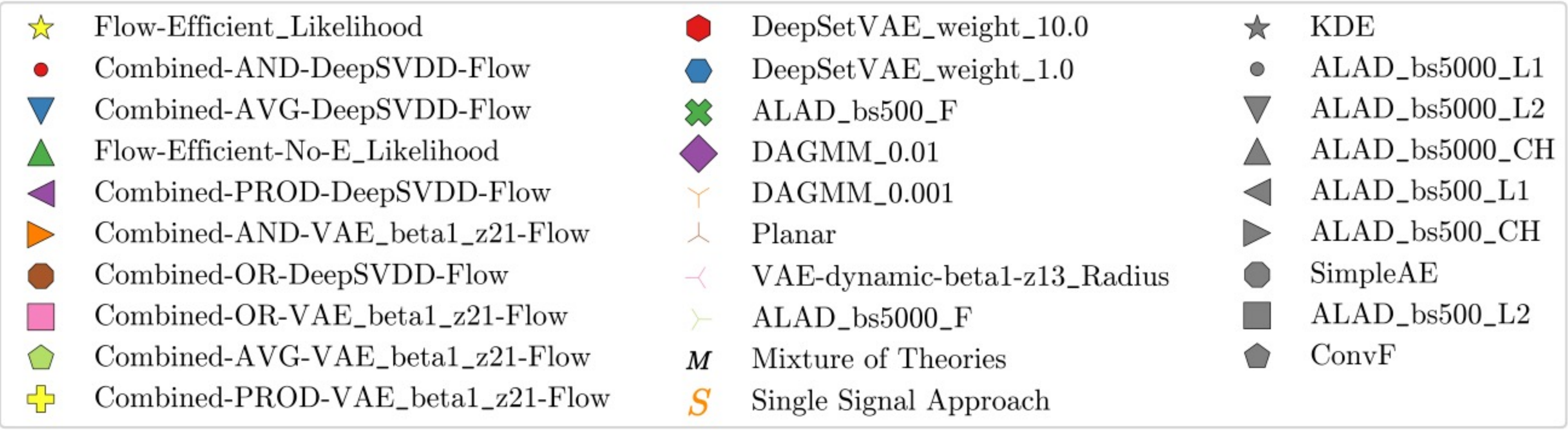
Total Improvement for models over all signals on
Dark Machines Unsupervised Challenge Hackathon Data
Median TI
Min TI
Max TI
Min TI
Max TI
Median TI
Latent Space
ALAD
DAGMM
ConvVAE
Planar
SNF
IAF
ConvF
KDE
VAE
Flow
Combined
Deep SVDD
Deep Set
CNN(β)VAE
SimpleAE
Mixture of Theories
mini Mixture of Theories
Single Signal Approach



- Modern DL outperforms traditional techniques
- AE not the optimal tools (no decoder needed)
- Flow models work very good
- Combined (rare+different) works good
- Supervised approaches outperform many AE's etc.
- Mixed signal approach outperform all supervised approaches

Secret dataset!!





Best:

Best supervised Mixed-model

outperform supervised and simple unsupervised

Classify LHC events with new methods? (for ATLAS)

→ Train NNs on simulation of LHC events

ATLAS physics: Classification

Comparison of different NN structures
To classify events as 4top or background
→Quite significant improvement expected to current baseline with dedicated ”multi-head self-attention”+physics based NNs

(guess that 4 sigma can become 6 sigma expected significance, just by better NN…)

Next steps: finish paper, implement in ATLAS, switch loss function to anomaly detection

Preliminary numbers

Classifier	AUC	Accuracy
ParT (pair int.)	0.870	0.786
LightGBM (pair int.)	0.840	0.760
ParT (no pair int.)	0.840	0.759
LightGBM (no pair int.)	0.831	0.750
1D CNN	0.831	0.750
FCN	0.829	0.744
PN	0.825	0.747
Unoptimised FCN	0.822	0.746

Figure: AUC and accuracy of all tested classifiers

Working here with
Luc Bultjes, Polina, Clara, Rob, Roberto, Zhonghy

Transformers is all you need

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

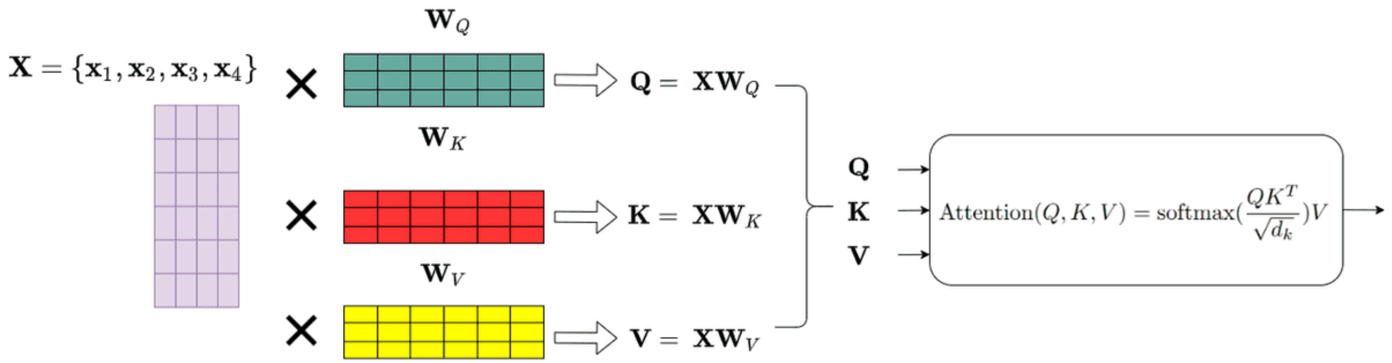
Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

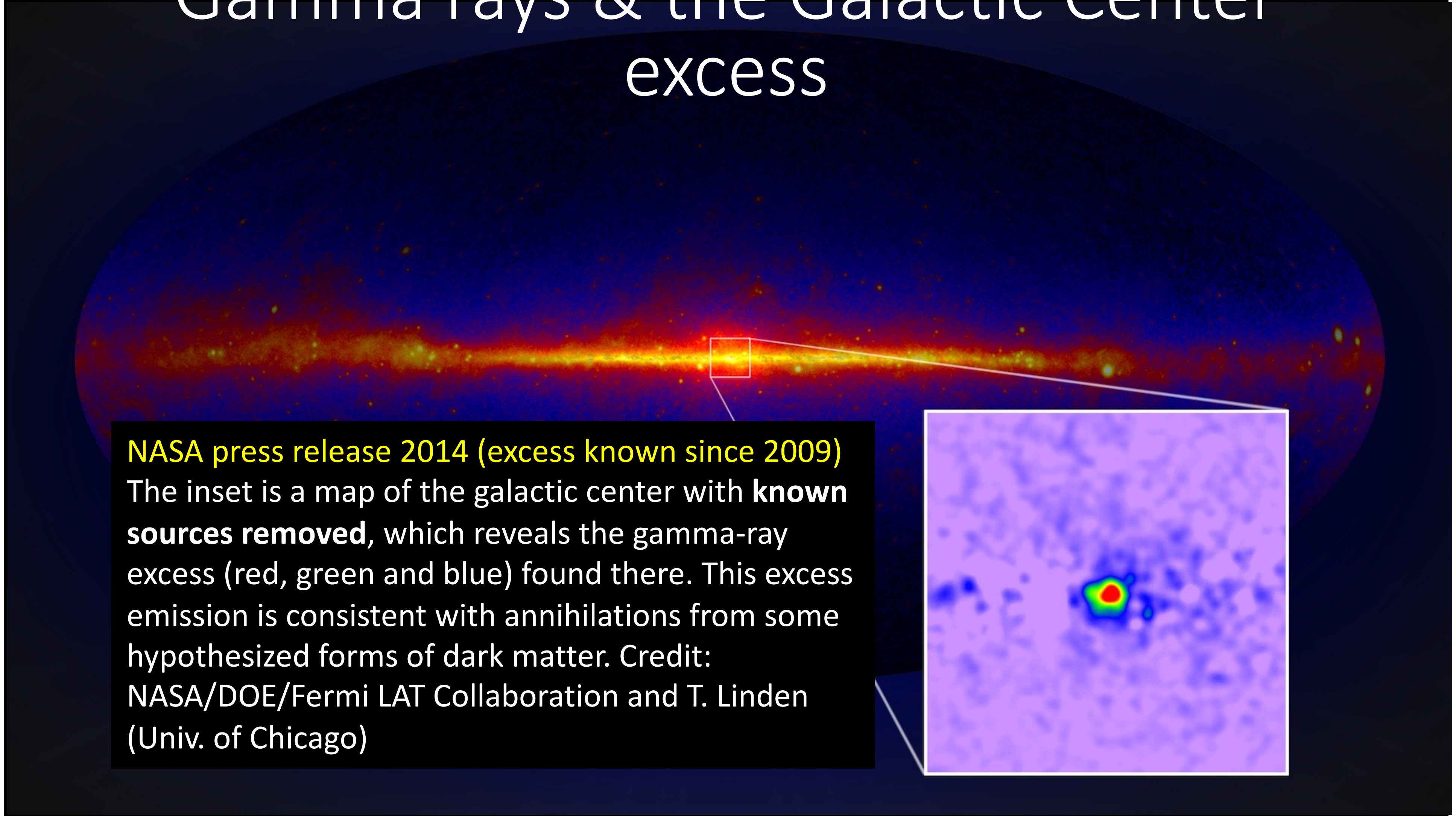
Abstract



Gamma rays: Galacitic Center and the reality gap

→ **Do we see Dark Matter in the center of our galaxy ?**

Gamma rays & the Galactic Center excess



NASA press release 2014 (excess known since 2009)
The inset is a map of the galactic center with **known sources removed**, which reveals the gamma-ray excess (red, green and blue) found there. This excess emission is consistent with annihilations from some hypothesized forms of dark matter. Credit: NASA/DOE/Fermi LAT Collaboration and T. Linden (Univ. of Chicago)

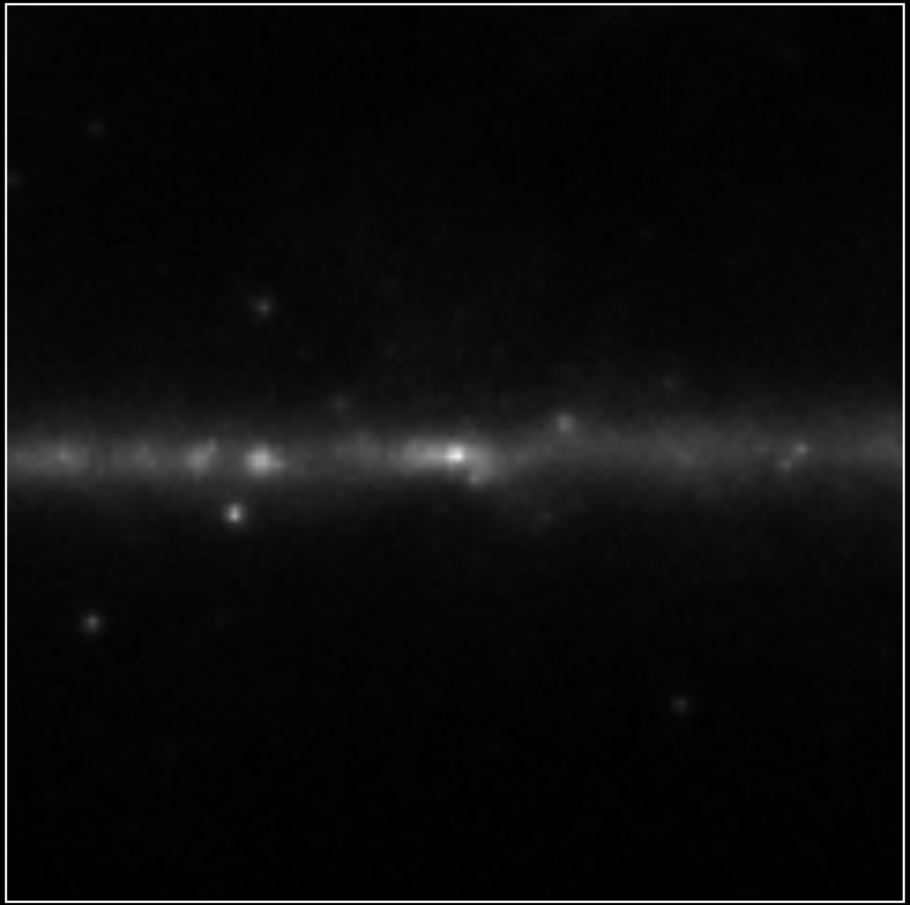
Official paper in 2015

Fermi-LAT Observations of High-Energy Gamma-Ray Emission Toward the Galactic Center
Fermi-LAT Collaboration (M. Ajello (Clemson U.) *et al.*). Nov 9, 2015. 29 pp.
e-Print: **arXiv:1511.02938 [astro-ph.HE]** | PDF

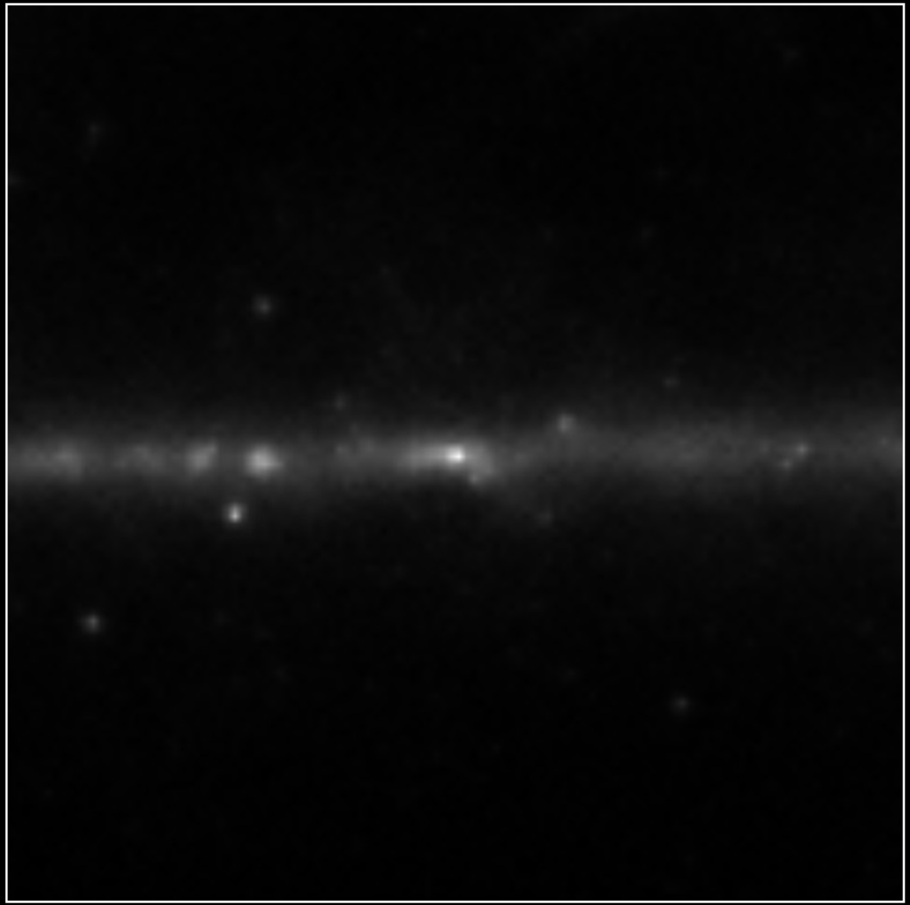
Guess the fraction of point sources

www.mydarkmachine.org

What is this fraction?



This is 0.5



Your prediction:

Invert image:

Guess

Simulation with Parameters → Pictures

First idea: Train Conv Network for

Pictures → Parameters

Does this work ? Does it work better than conventional methods ? Why ?

Our 2017 convolutional network

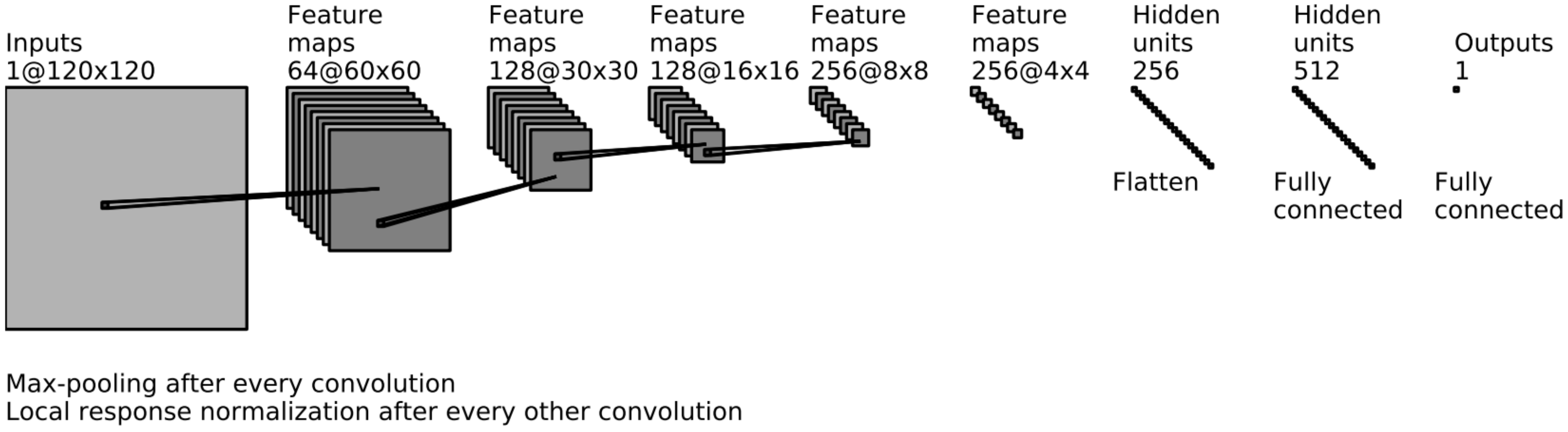


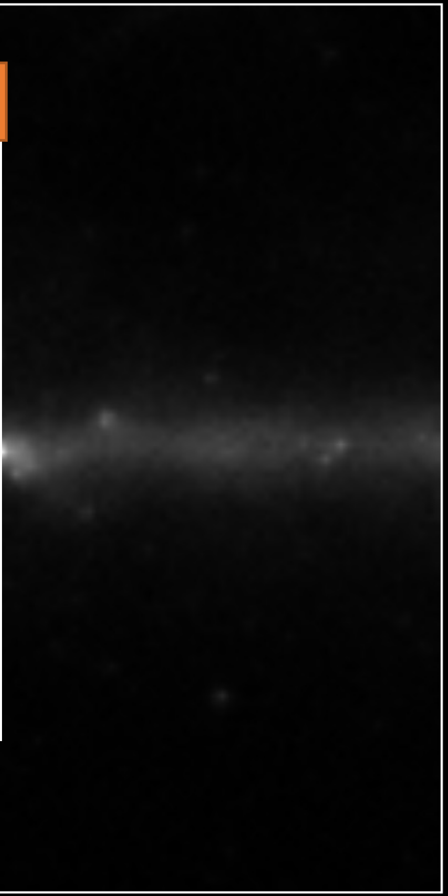
Figure 6: Visualization of the convolutional neural network. The network consists of an input layer, 5 convolutional + pooling layers, 2 fully connected layers and finally an output layer.

Can a NN determine the number of unresolved point sources relative to isotropic radiation ?

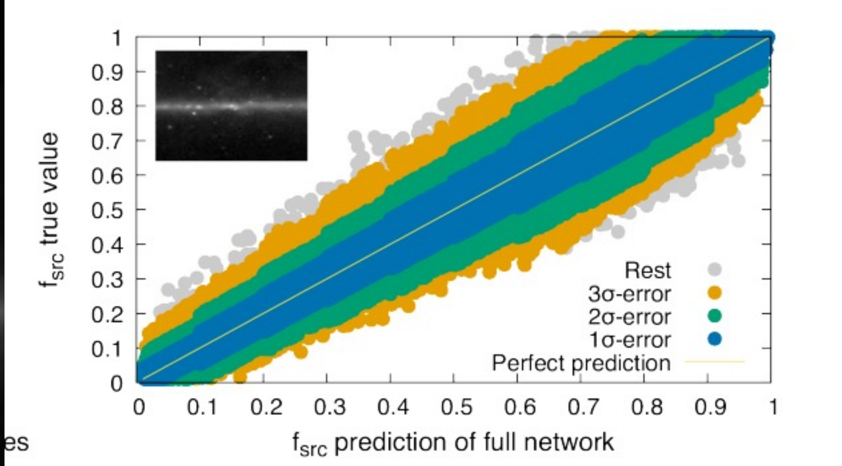
- Published in: *JCAP* 05 (2018) 058, e-Print: 1708.06706 [astro-ph.HE]

What is this fraction?

This is 0.5



Network can generalize over randomness



(b) Prediction of the full network versus true values.

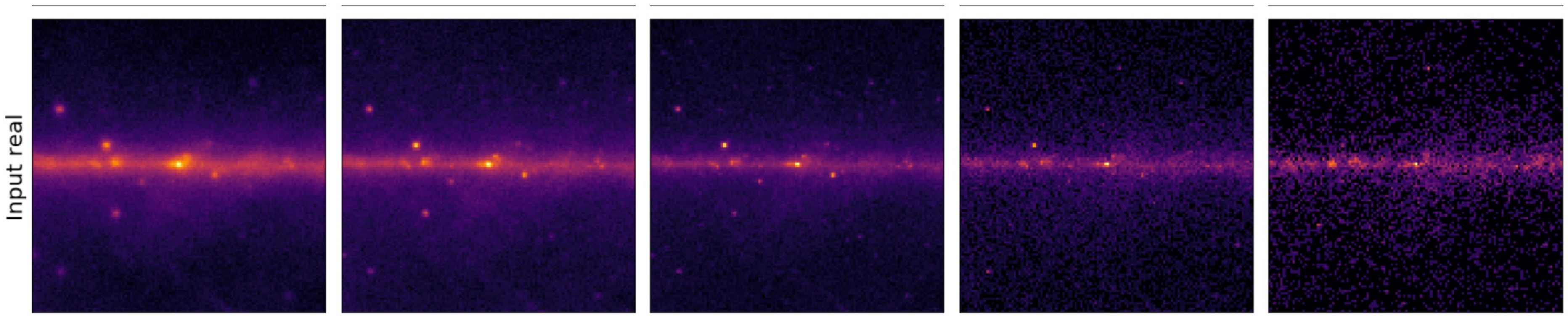
Your prediction:
Invert image:

Truth:	0.052
Network:	0.1230
Your guess:	0.5

Who is better? The network

Interpretation here is
frequentists and relies on the model
to be correct (uncertainties
from toy experiments, no p-value yet)

Today: More wavelenghts→ Bayesian determination of 25 parameters

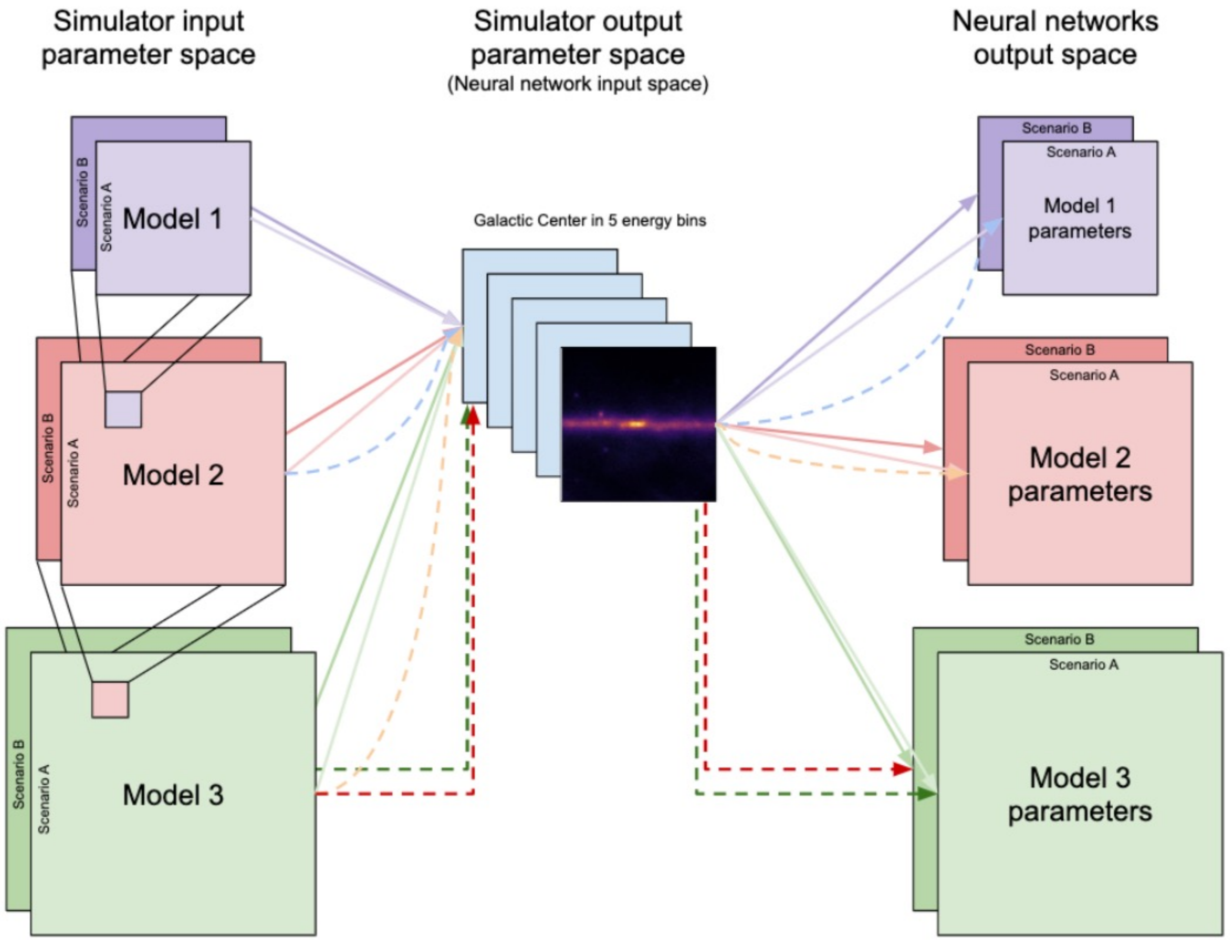


LET NETWORK OUTPUT all parameters AND all uncertainties

New paper after 5 years !

RU Internal / in Fermi-LAT review:

Idea: Test more complex simulations, learn the best simulations from data, for the first time include all uncertainties (also "out of simulation")



Gamma rays: Galacitic Center and the reality gap

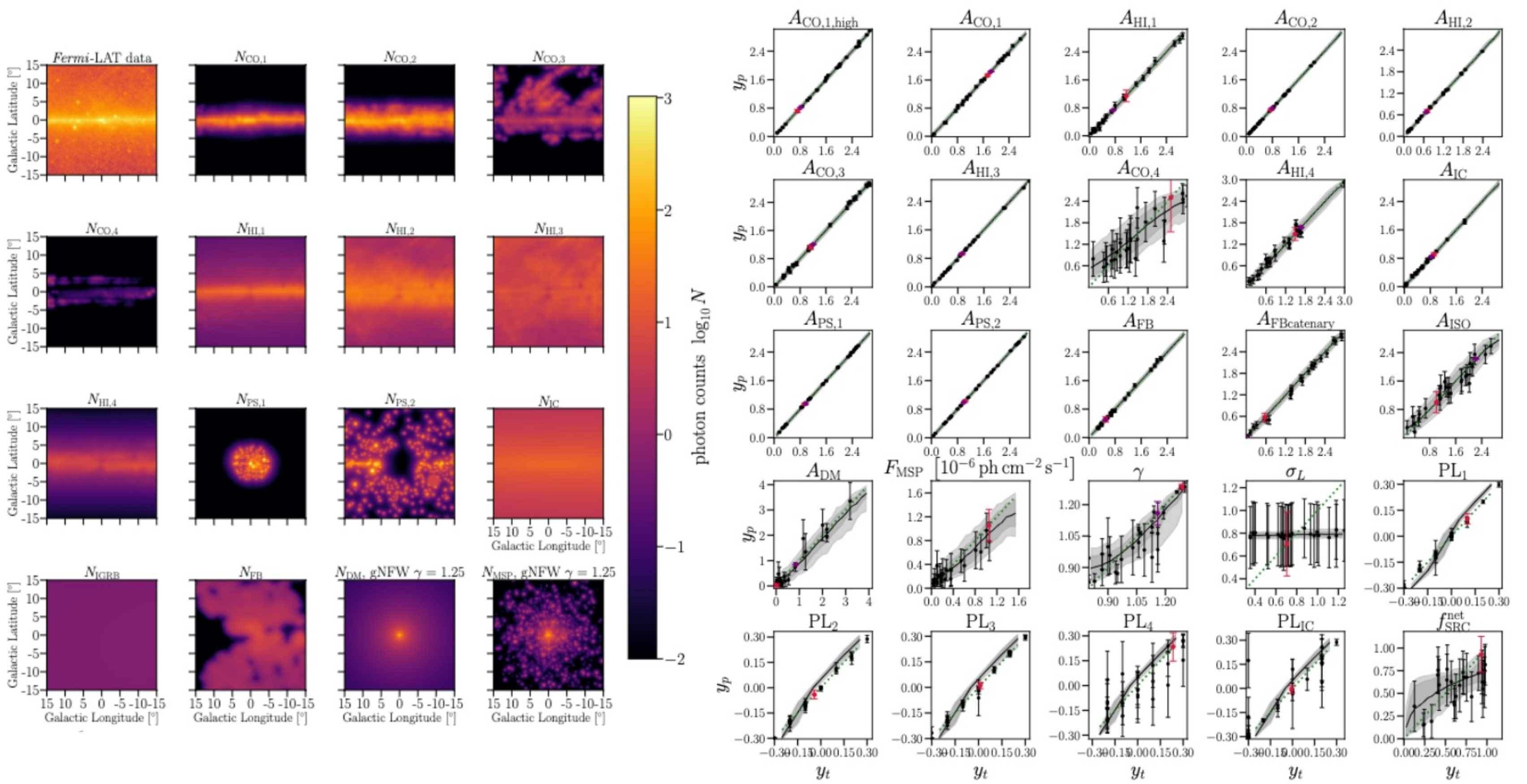


Figure 8: The same as Fig. 4, but in the context of Model 3 (Scenario A).

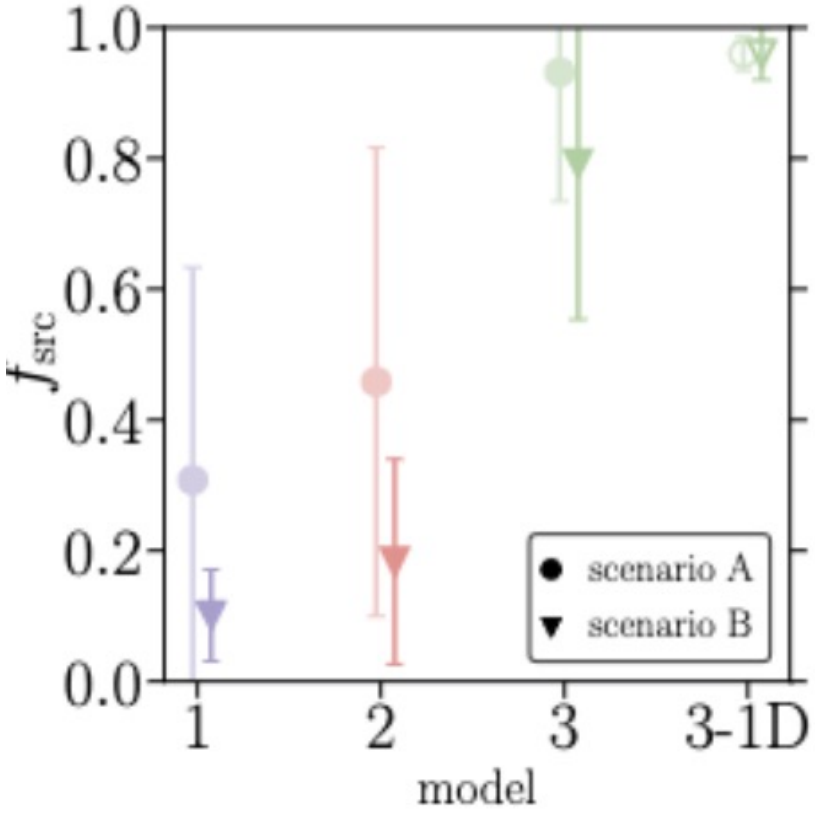
Our new 2022 NN can regress here 25 parameters at once

And NN has learned its uncertainty with deep ensemble networks

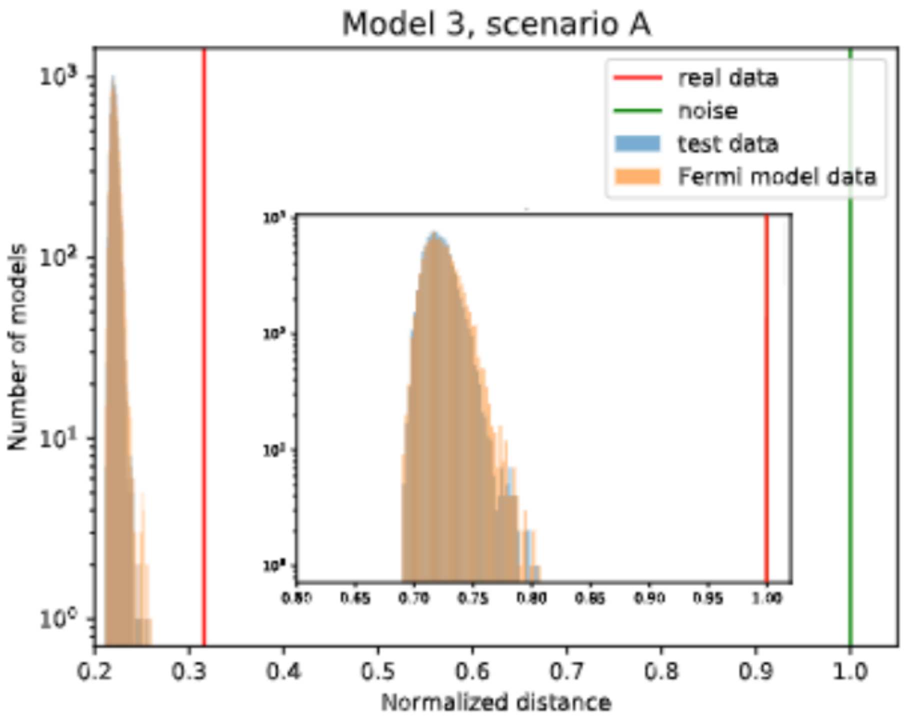
(discussion also with Laurens Sluijterman, E. Cator)

The finding

What is the fraction of point sources for this simulation?



Is the data in the simulation ?



Can we also determine the point sources directly?

Yes, other project (shown already some time ago)… show you next astrosourceid applied to astronomy data

Astronomy and gamma rays: autosourceID

→ **Automatically identification of astronomical objects**

Automatic ID of astrophysical objects: AutosourceID, slides by Fiorenzo Stoppa

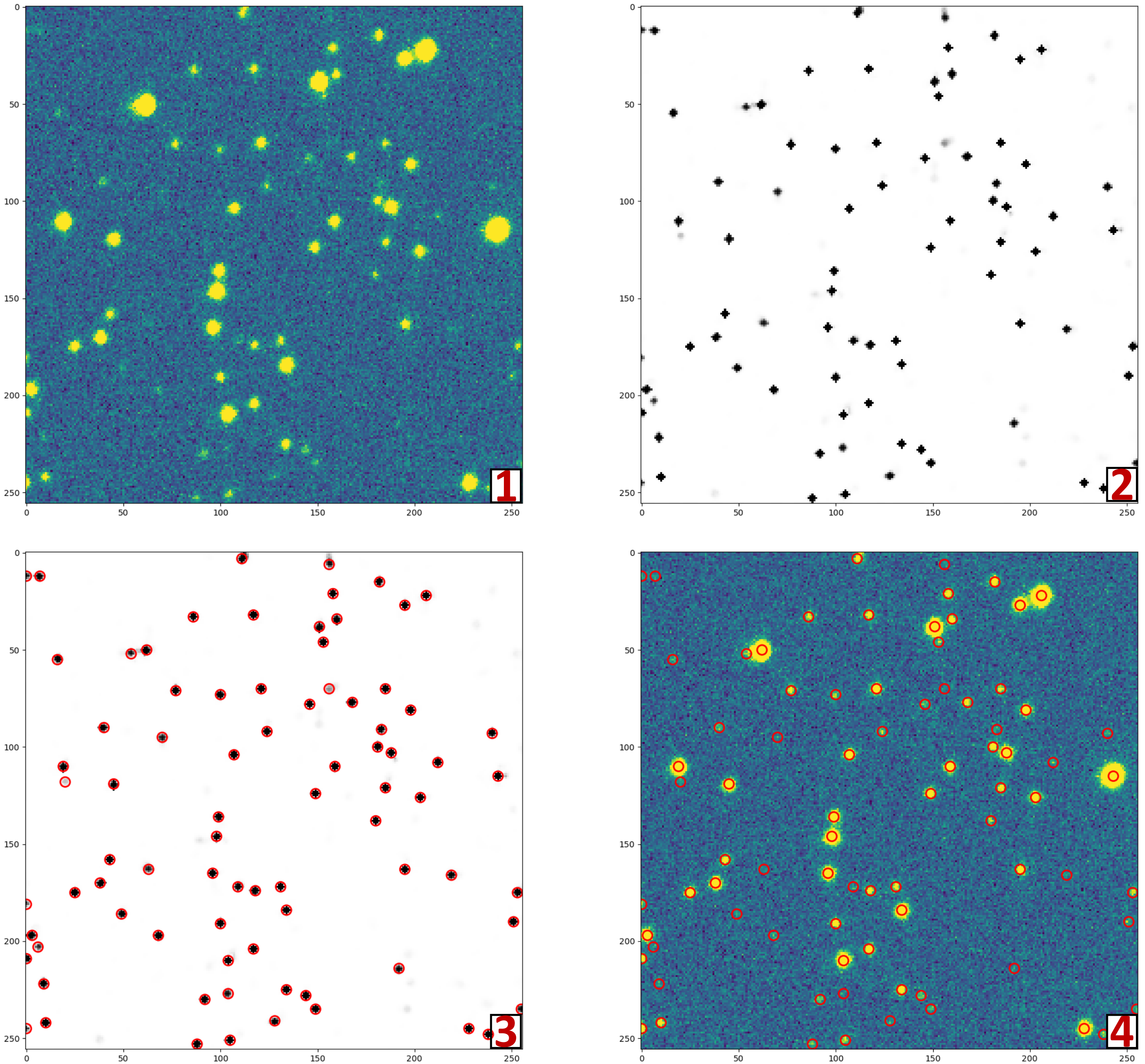
IDEA: FASTER / REALTIME ID OF ASTRONOMIC SOURCES

Full field (10.5k x 10.5k pixels) is 3.7 seconds for AutosourceID and 120 for SExtractor.

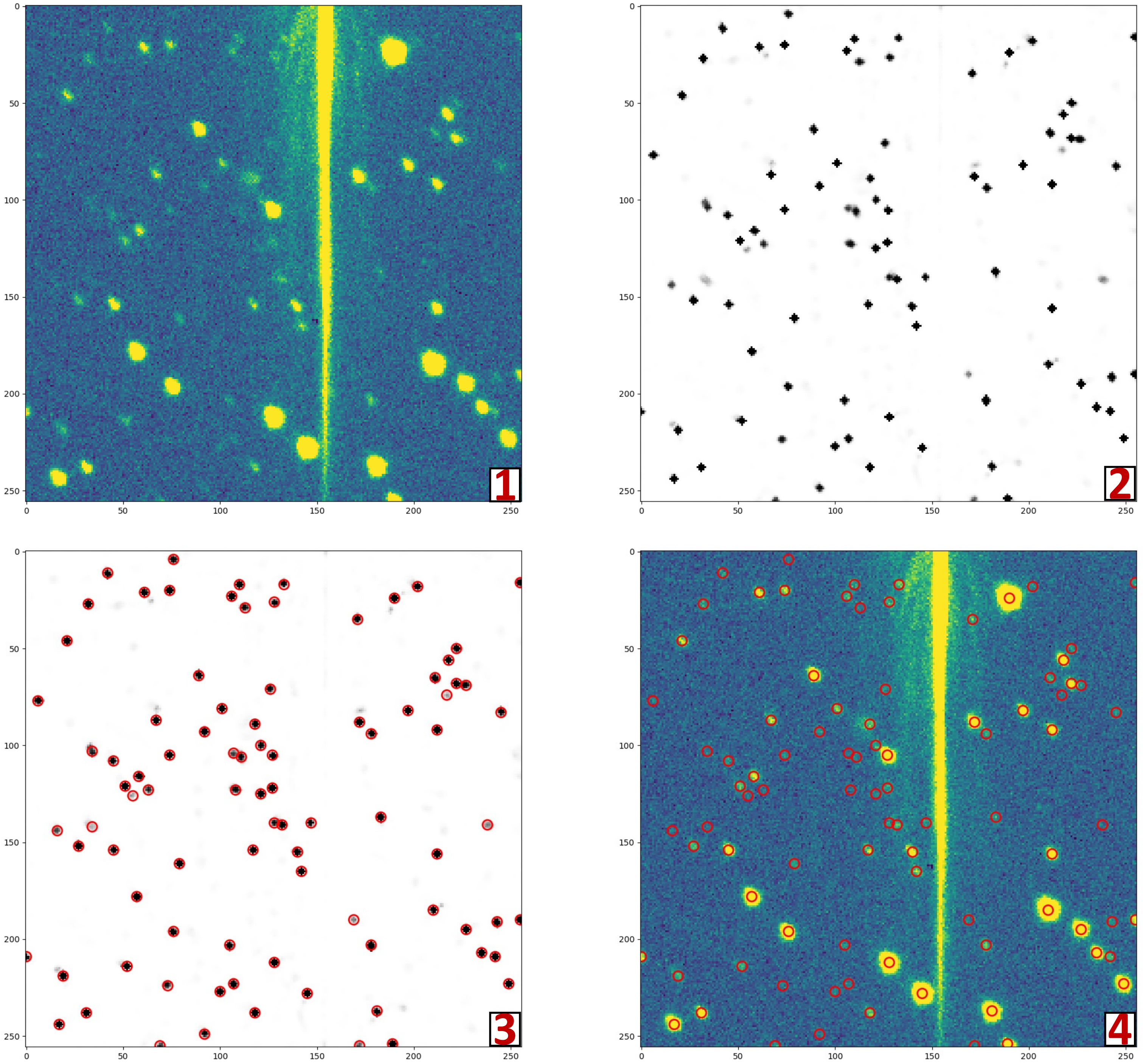
Train the ML on the simulation and/or the astronomer.

Machine Learning is not necessarily the problem here,
Machine learning can be the solution ! → Less Energy

1 Input optical image
2 Predicted mask
3
4 Laplacian of Gaussian Results



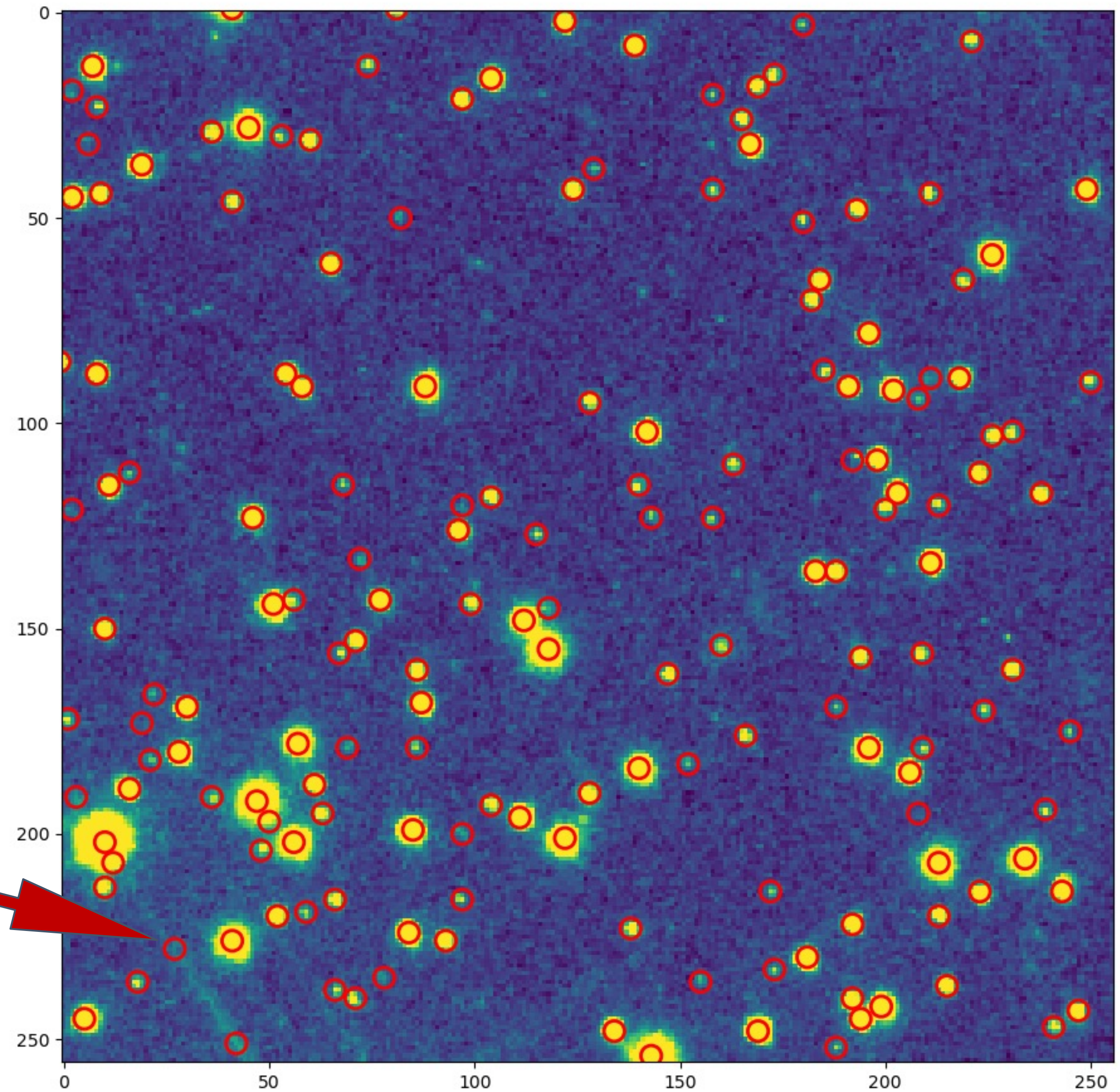
1. Input optical image
2. Predicted mask
3. Laplacian of Gaussian
4. Results



Hubble HD images

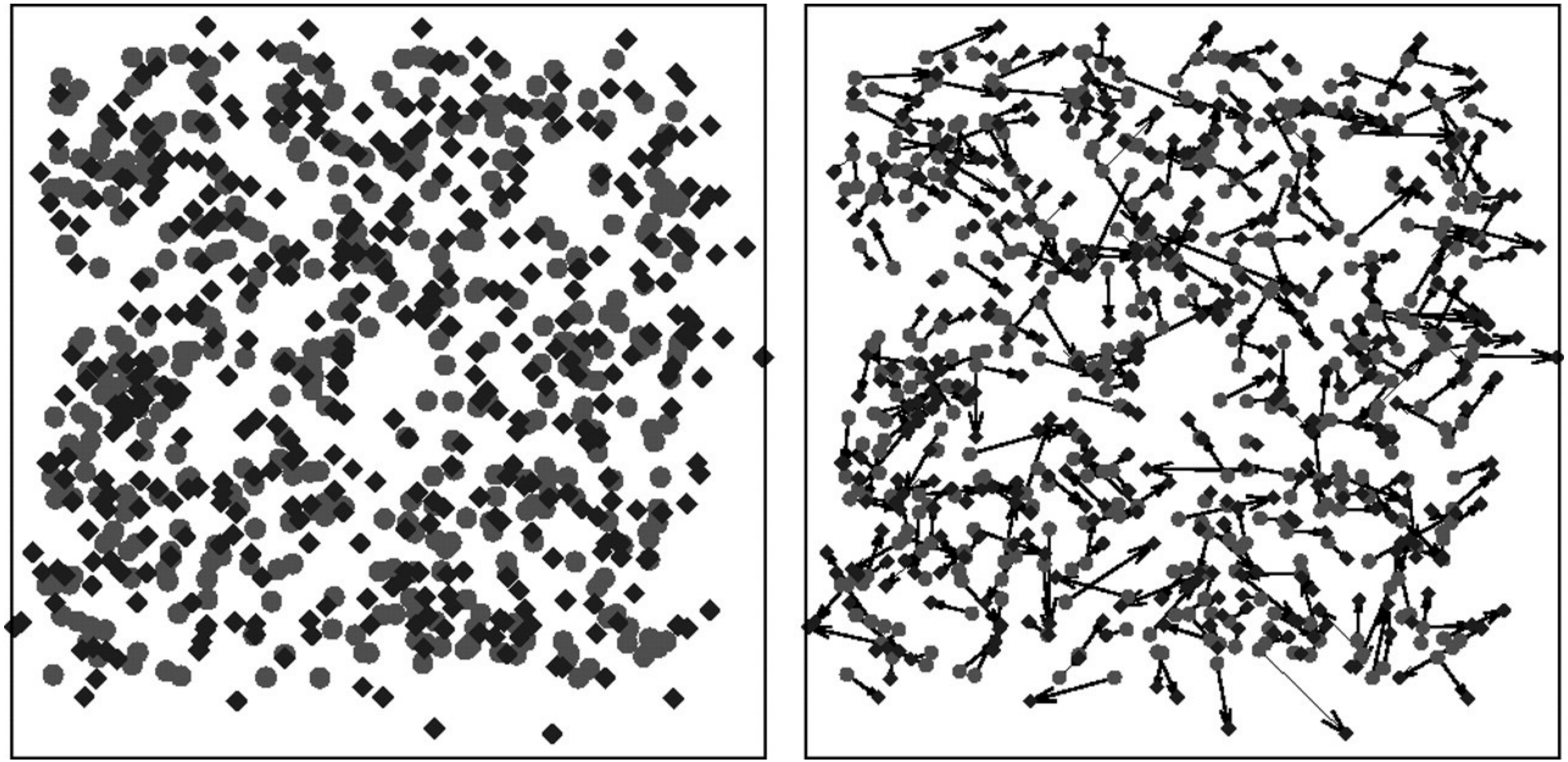
Most of the visible sources correctly localized.

Small problems with a diffraction spike



•*Astron.Astrophys.* 662 (2022) A109, With Astro department (mainly Fiorenzo Stoppa)

ATLAS: tracking → inference at 40 MHz ?

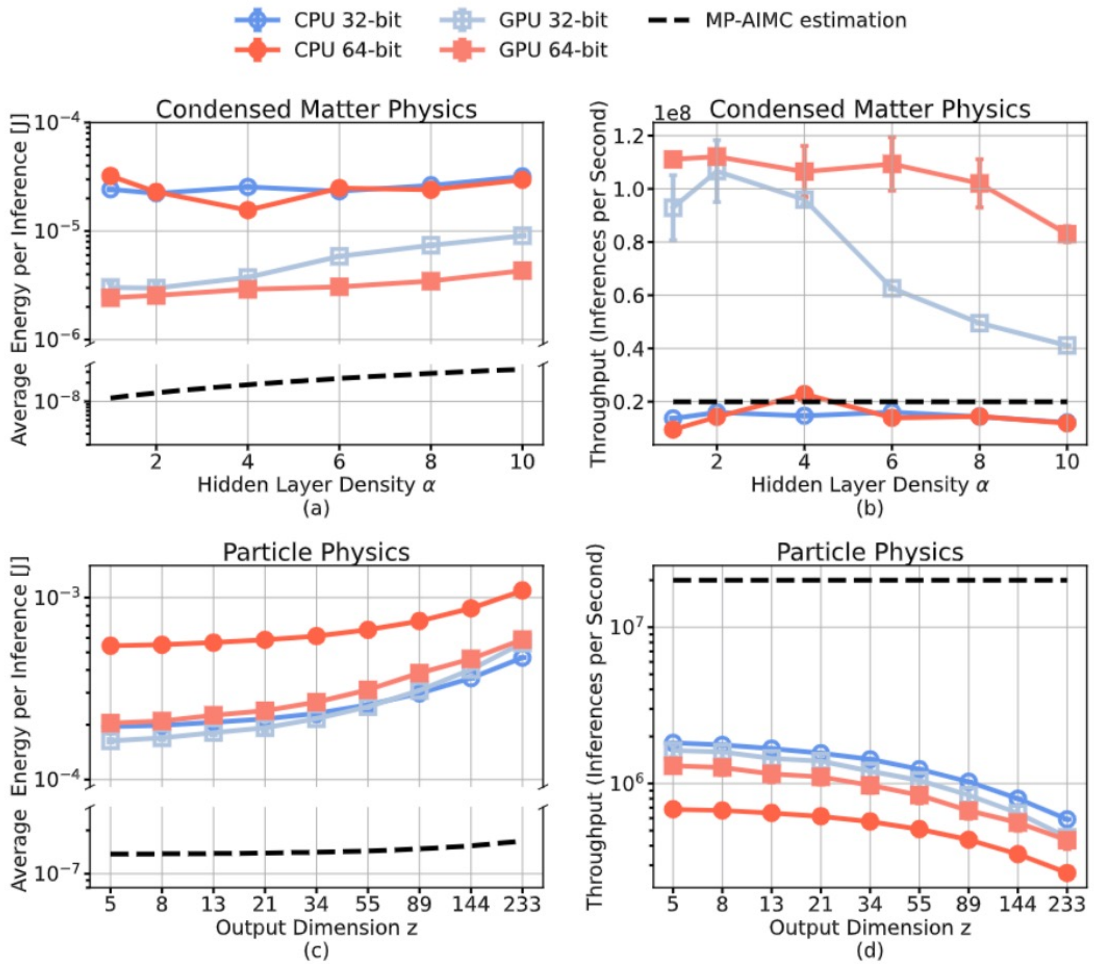


Idea : Train u-nets to go from **(almost) ALL pixels → (almost) ALL tracks** (including ALL uncertainties) in one step
Do this on dedicated hardware accelerators (GPUs, FPGAs, neuromorphic, future quantum ?) → CODE?

Neuromorphic computing

→ **The next step: Print scientific Neural Networks on computer chips**

LHC etc. : Neuromorphic Computing on AIMC architecture with IBM and IMM



How fast can neuromorphic chips process scientific data?
→ ATLAS trigger
How much energy do they consume ?

(also compare to quantum hardware, maybe enourmous gain!)

Benchmarking energy consumption and latency for neuromorphic computing in condensed matter and particle physics

Dominique J. Kösters,[1,2,3] Bryan A. Kortman,[1,4] Irem Boybat,[3] Elena Ferro,[3,5] Sagar Dolas,[6] Roberto Ruiz de Austri,[7] Johan Kwisthout,[8] Hans Hilgenkamp,[1,9] Theo Rasing,[10] Heike Riel,[3] Abu Sebastian,[3] Sascha Caron,[4,11] and Johan H. Mentink[10]

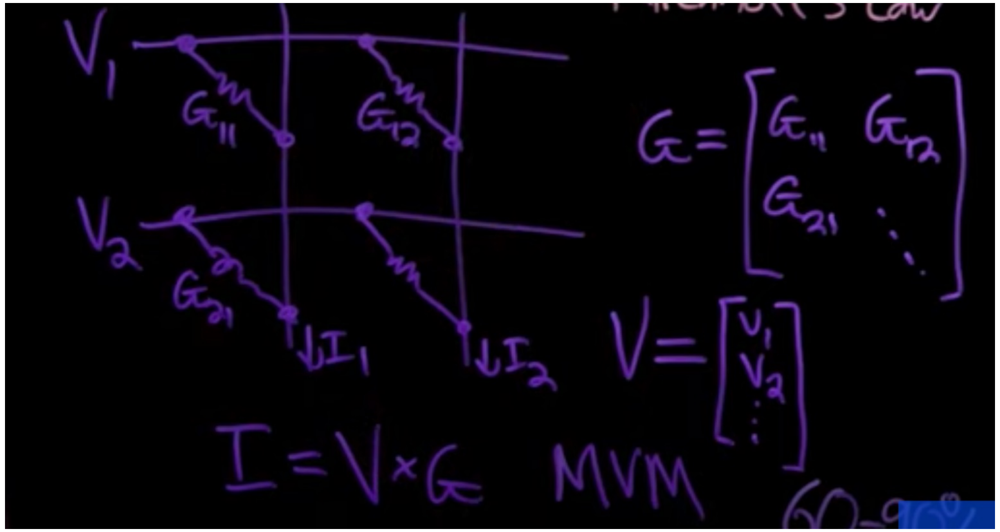
[1] *University of Twente, Faculty of Science and Technology, P.O. Box 217 7500 AE, Enschede, The Netherlands*
[2] *Radboud University, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 AJ, Nijmegen, The Netherlands*
[3] *IBM Research Europe - Zürich, 8803 Rüschlikon, Säumerstrasse 4, Switzerland*
[4] *Nikhef, P.O.Box 41882 1098 XG Amsterdam, The Netherlands*
[5] *Eidgenössische Technische Hochschule Zürich, Department of Information Technology and Electrical Engineering, Gloriastrasse 35, 8092 Zürich, Switzerland*
[6] *SURF Cooperation, Innovation Team, Moreelespark 48, 3511 EP, Utrecht, The Netherlands*
[7] *University of Valencia-CSIC, Instituto de Física Corpuscular, Parc Científic UV, c/ Catedrático José Beltrán 2, E-46980 Paterna, Spain*
[8] *Radboud University, Donders Institute for Brain, Cognition and Behaviour, P.O. Box 9104 6500 HE, Nijmegen, The Netherlands*

Answering referee comments

→ published in a few weeks

Neuromorphic computing

I = current
V= voltage
G= resistances

NN = 

(source IBM video)

Various approaches raning from classifical FPGA, ASICs to In memory computing (previous slide), spiking NN on chips (Inter Loihi) or even photonic !
Etc.

→ Need dedicated study, will likely become highly important for compuational science
→ Main topic of our NWA proposal “datascope” (was nextgraspp before)
→ RU could become a leader here ?

Next steps: philosophy and chatbots

➔**Let’s look 10 years ahead: What is the future of AI in physics ?**

Next steps: philosophy and chatbots

With Henk de Regt (philosophy of science), Kristian Gonzalez and Tom Claasen (causal discovery, ICIS)

Questions:

- Science Bot → Assume (build?) SIRI/ALEXA (Bert/GPT3 etc) that can ask scientific questions:

How to make causal relations, how to trust this machine, what do scientist want, do we like that this is a “Googlebot (made by google) or HEPbot (made by HEP community)”

- Sustainability of computational science

Summary

- Actually these are not all projects (really a lot of interesing things to work on)
- Hope to convince you (a bit) that ML is interesting for HEP, astronomy, sustainability , FNWI and Nikhef….