NATIONAL RESEARCH CENTER "KURCHATOV INSTITUTE"

Machine learning for dark matter search experiments

Alexey Grobov
NRC Kurchatov Institute
Cagliari, 24/02/2021
alexey.grobov@gmail.com
Grobov_AV@nrcki.ru

DARKSIDE

DEAP 3600

dmlc XGBoost

scikit learn

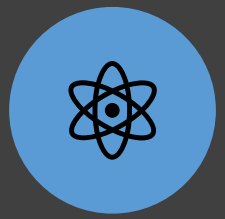
TensorFlow

PyTorch

Keras

ROOT
Data Analysis Framework

Machine learning for physicists



Nowadays machine learning used almost everywhere, and physicists too decided to take advantage of its techniques.

For physicists ML algorithms is a tool to solve problems.

When it is okay to use machine learning?

When you can do without it:

- You can solve the problem without ML and you're fine with the result – Great!

No need to hammer nails with a microscope.

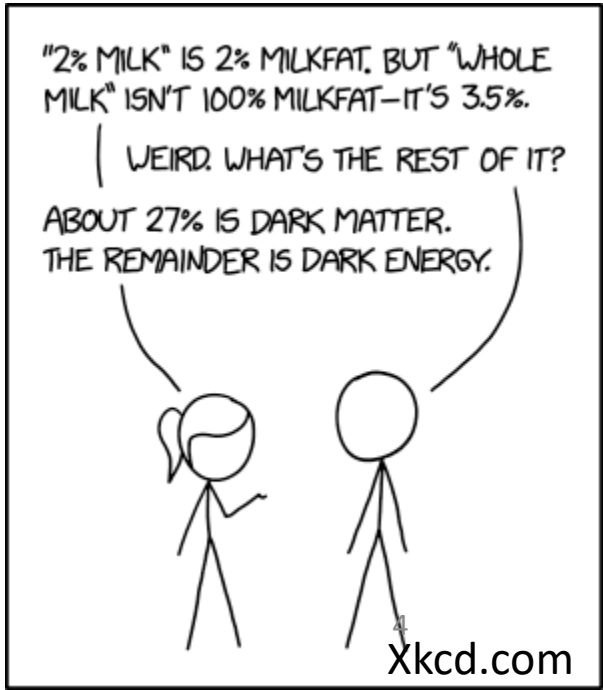
When your problem has many variables and you seek to improve existing analysis, and even a small increase in performance is important for you – it's okay.

Dark matter and where to find it

- No one knows!
- Dark matter particle is a theoretical construction we use to explain variety of observational facts.
- Why are we so confident that dark matter is real?

Objects	Distance (in kpc)	Luminosity (in sol. lum.)	Mass (in sol. mass)	Mass/Lum. f
Solar Neighborhood	—	—	—	4
Triangulum Nebula, M33	480	1.4×10^9	5×10^9	4
Large Magellanic Cloud	44	1.2×10^9	2×10^9	2
Andromeda Nebula	460	9×10^9	1.4×10^{11}	16
Globular Cluster, M92	11	1.7×10^5	$<8 \times 10^5$	<5
Elliptical Galaxy, NGC 3115	2100	9×10^8	9×10^{10}	100
Elliptical Galaxy, M32	460	1.1×10^8	2.5×10^{10}	200
Average S in Double Gal.	—	1.3×10^9	7×10^{10}	50
Average E in Double Gal.	—	8×10^8	2.6×10^{11}	300
Average in Coma Cluster	25000	5×10^8	4×10^{11}	800

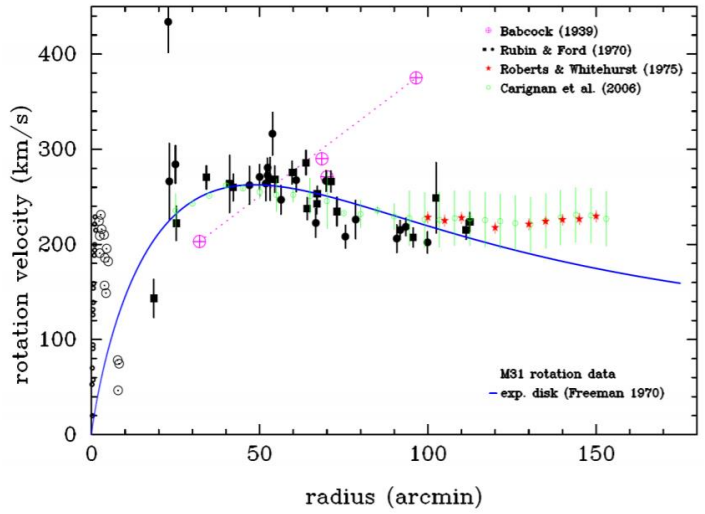
FIG. 1. A snapshot of the dark matter problem in the 1950s: the distance, mass, luminosity, and mass-to-light ratio of several galaxies and clusters of galaxies. From Schwarzschild, 1954.



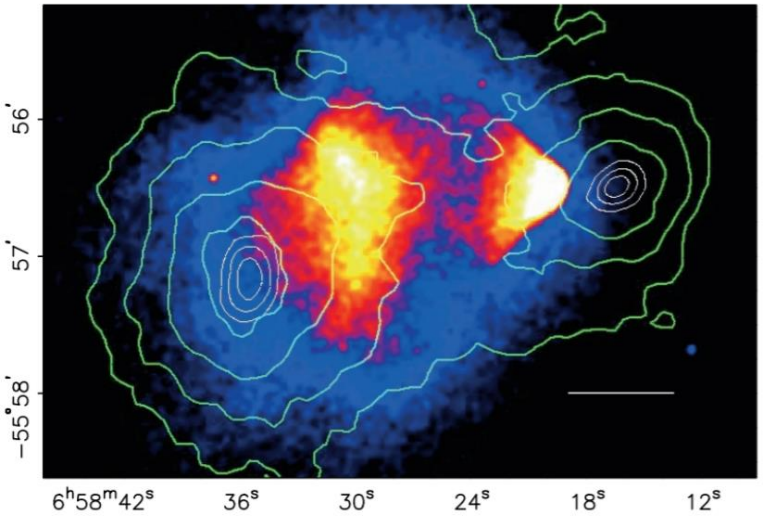
Xkcd.com

D. Hooper and G. Bertone “A history of dark matter” arxiv:1605.04909

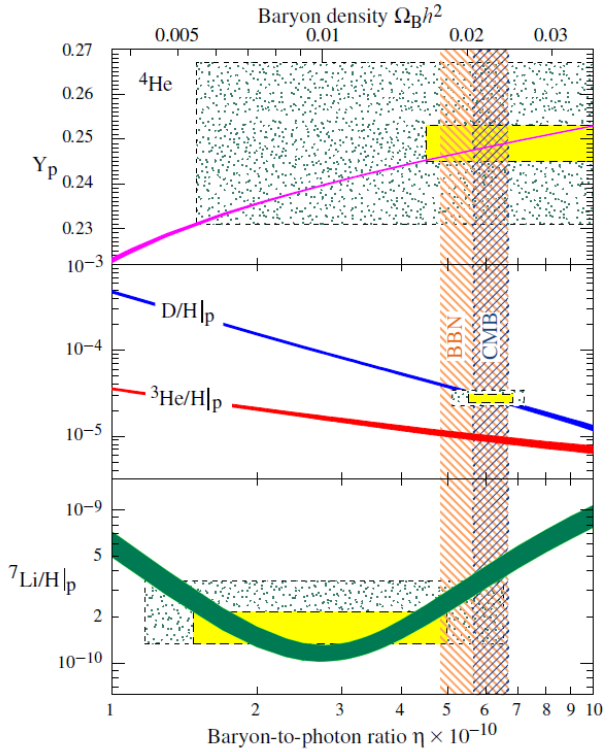
Dark matter



Looks like Galaxies are immersed into Halo of dark matter



Bullet cluster



$$\frac{D}{H} \Rightarrow \Omega_B h^2 = 0.0214 \pm 0.0020$$

Detection idea

PHYSICAL REVIEW D VOLUME 31, NUMBER 12 15 JUNE 1985

Detectability of certain dark-matter candidates

Mark W. Goodman and Edward Witten
Joseph Henry Laboratories, Princeton University, Princeton, New Jersey 08544
(Received 7 January 1985)

Astrophysics model

$$\frac{dR}{dE_R} = \frac{\rho_T}{m_T} \cdot \frac{\rho_\chi}{m_\chi} \epsilon(E_R) \int_{v_{\{min\}}}^{\infty} v f_\chi^{\oplus}(\vec{v}) \frac{d\sigma}{dE_R} d^3\vec{v}$$

Diff. scat. rate

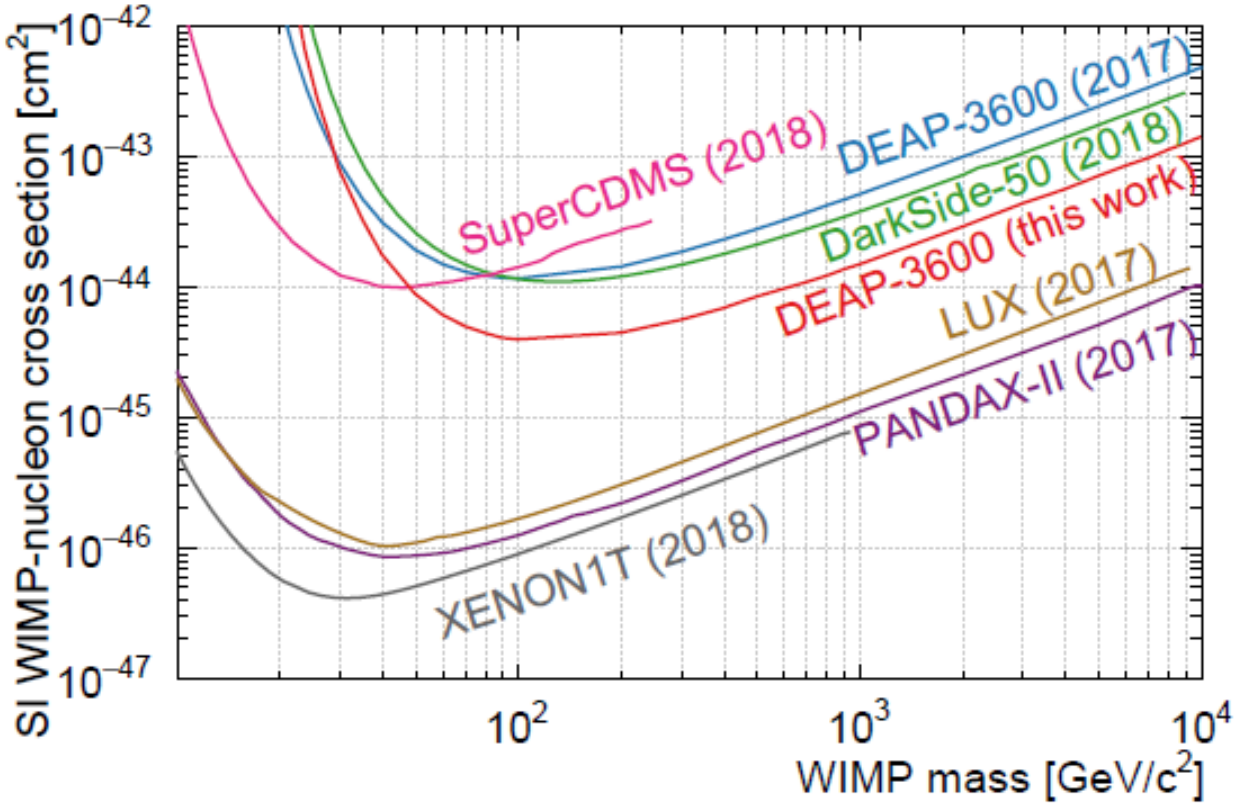
Detector model

Particle physics model

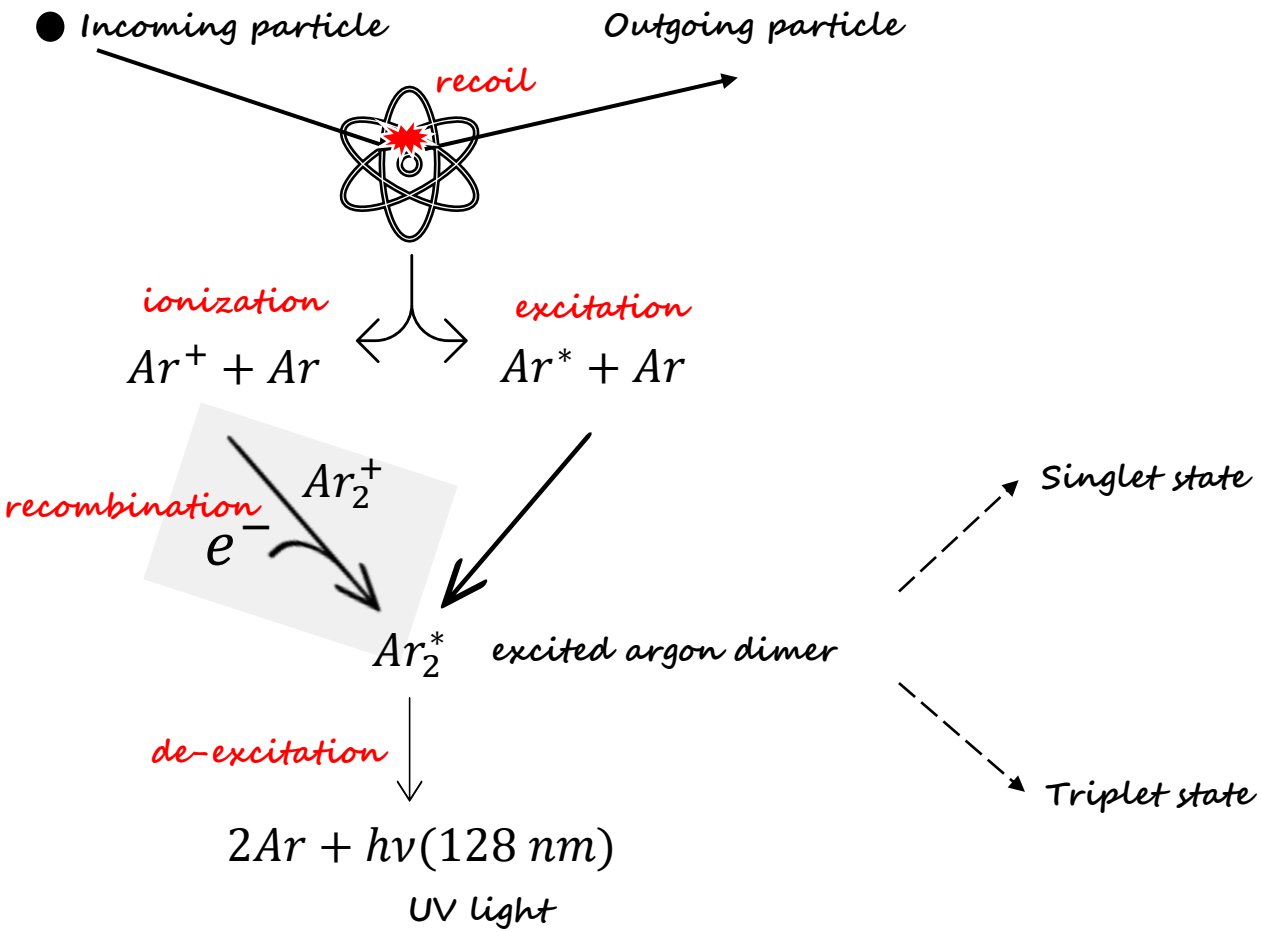
SNOLAB
@SNOLABscience

Hunt for the unseen
A whir, bubble, flash of light
Dark matter escapes

📷 Aric Guite, Robert Provencher, Douglas Takayesu



Bright and Scalable

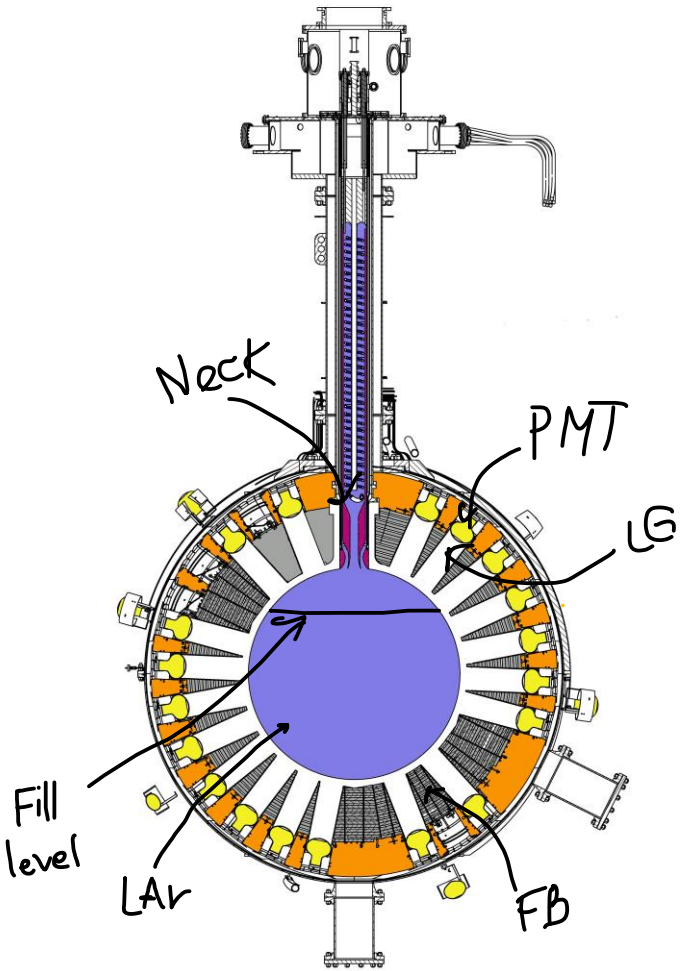


$$RATIO = \frac{N_{\{singlet\ states\}}}{N_{\{triplet\ states\}}}$$

NUCLEAR RECOILS	ELECTRON RECOILS
$RATIO \simeq 3$	$RATIO \simeq 0.3$

This is the core property of liquid argon that allows Pulse Shape Discrimination

DEAP-3600 detector



DEAP-3600 is a single-phase liquid argon (LAr) direct-detection dark matter experiment.

Location: 2km underground at SNOLAB (Sudbury, Canada).

Target: 3279 kg of LAr (30 cm of GAr on top) in a spherical acrylic vessel (AV)

Light detection: 255 PMTs connected to AV by 45 cm light guides (LGs).

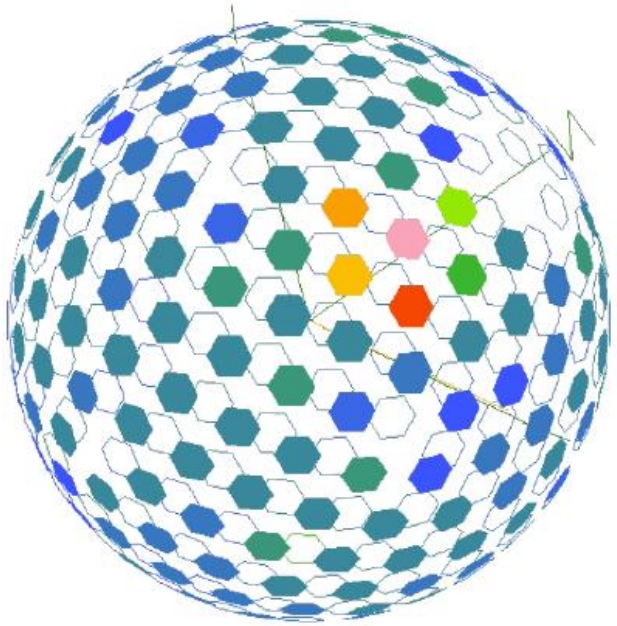
Construction: Filling of the detector done through the neck with LN2 cooling coil. AV and PMTs enclosed in stainless steel shell.

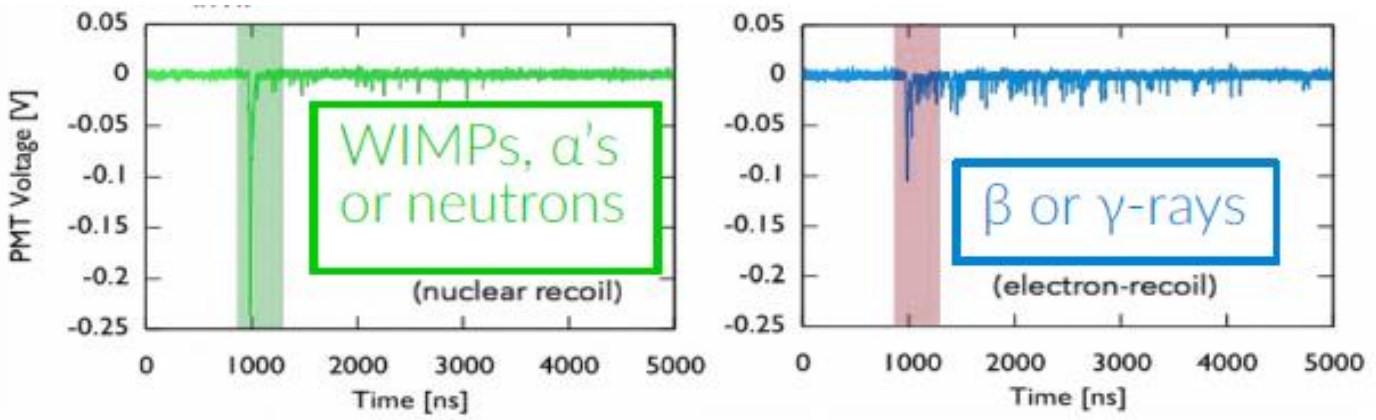
Shielding: Filler blocks (FB) between LGs used for thermal insulation and neutron shielding. Steel shell is immersed in 300 tons of H_2O, viewed by 48 veto PMTs. Neck of the detector has 4 Neck veto PMTs.

The DEAP Collaboration, *Design and Construction of the DEAP-3600 Dark Matter Detector*, Astropart. Phys. 108, 1 (2019).

Data

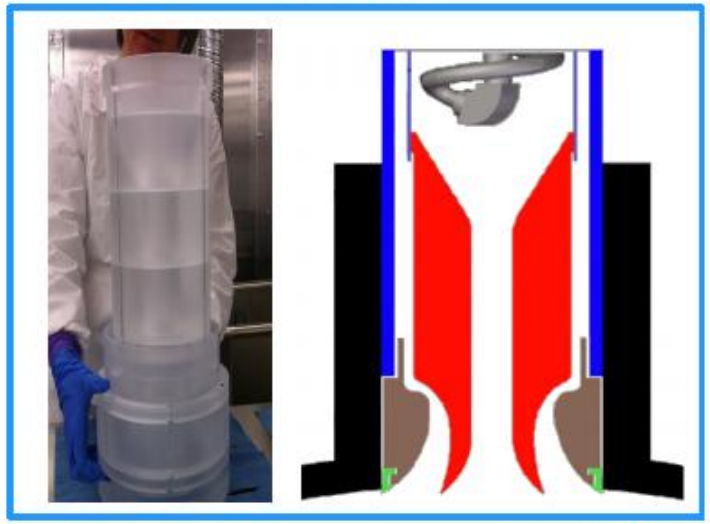
- When event triggers the detector we get a lot of information coded into variables:
  - Total amount of collected light – recoil energy
  - Light and time patterns – position of the event
  - Prompt fraction of light - pulse shape of the event
  - Maximum fraction of light collected by PMTs
  - Which PMTs were first to detect light
  - And about 100 others
- Based on this events are identified as alphas, Cherenkov, cosmogenics etc.
- We can check how good we understand physics underneath with MC

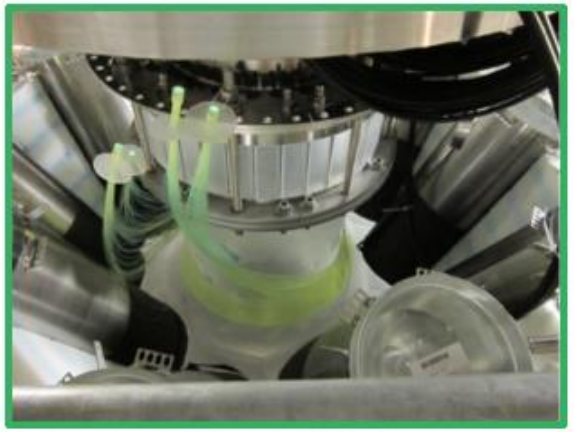




The Neck

Flow guides



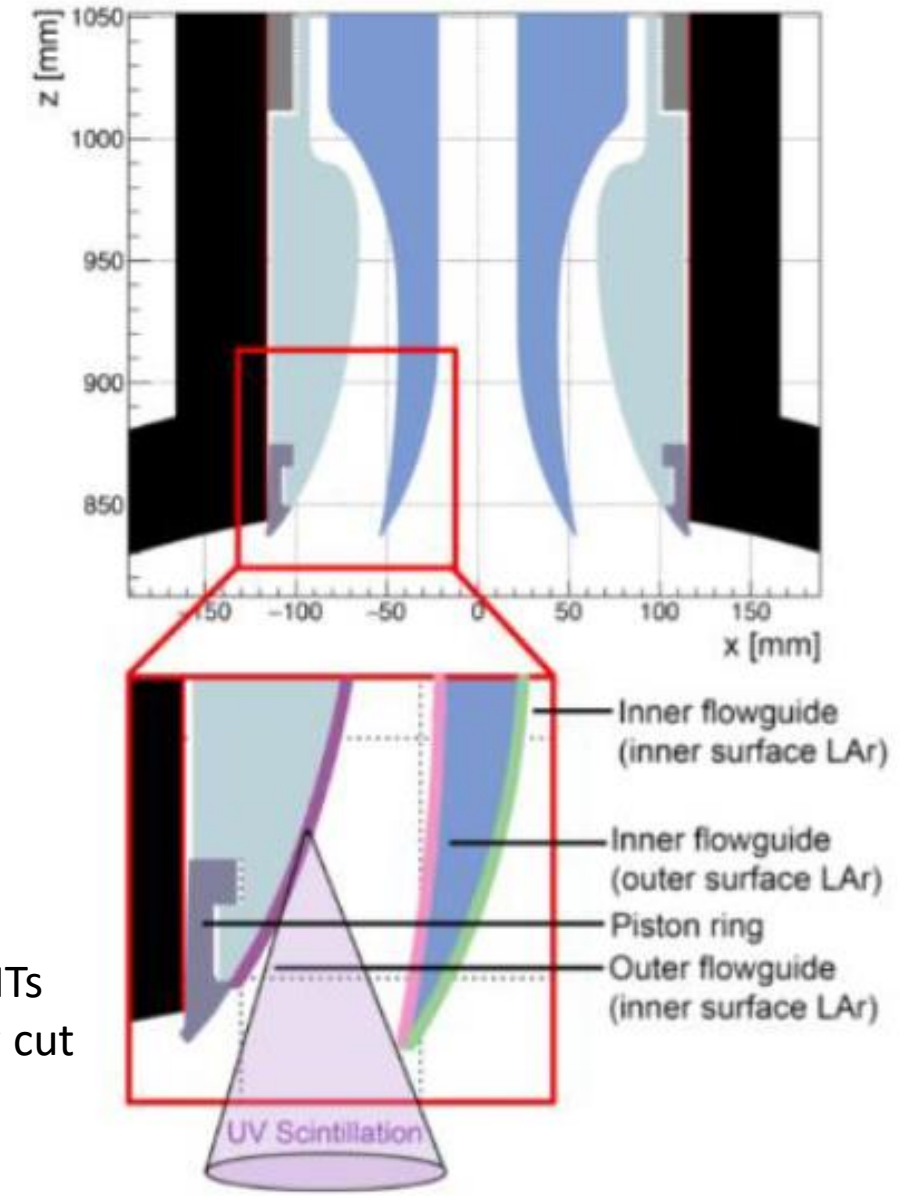


Neck Veto

- Largest contribution to the background rate
- FGs are not coated with TPB
- Acrylic absorbs UV light
- “Shadowed” event topologies

Event selection:

- Upper Fprompt cut
- Early pulses in GAr PMTs
- Charge fraction in top 2 rows of PMTs
- Position reconstruction consistency cut



ML Routine

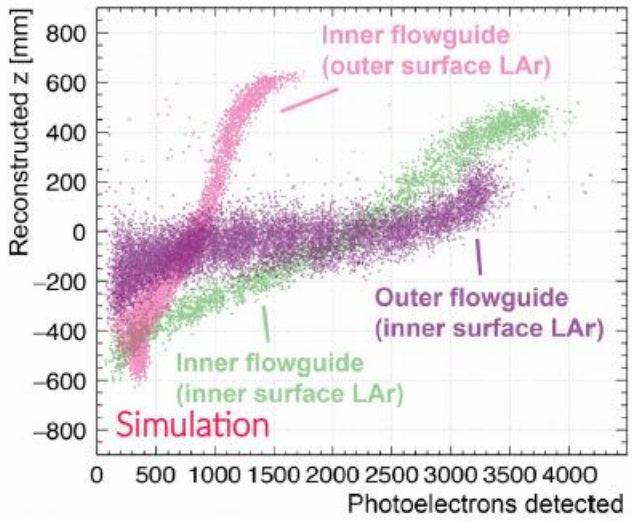
Need to classify something? You can never go wrong with Boosted Decision Trees.

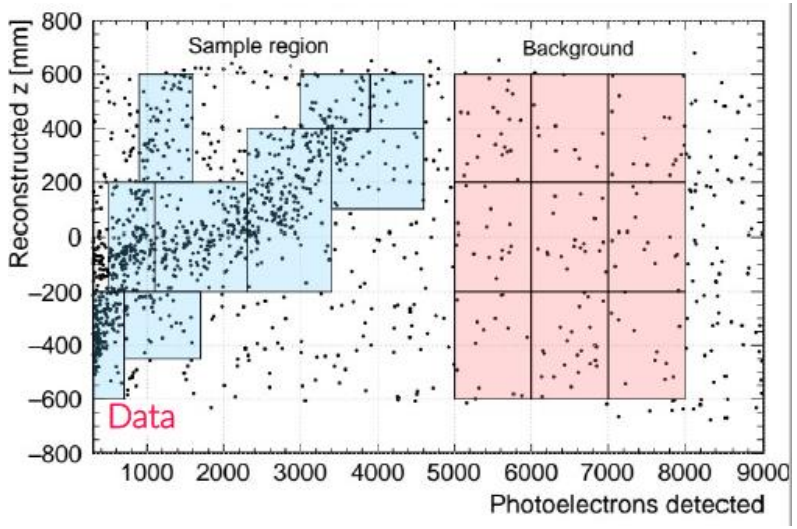
- Establish problem

- Decide on metrics (accuracy, signal acceptance, precision, number of electrons per apple)

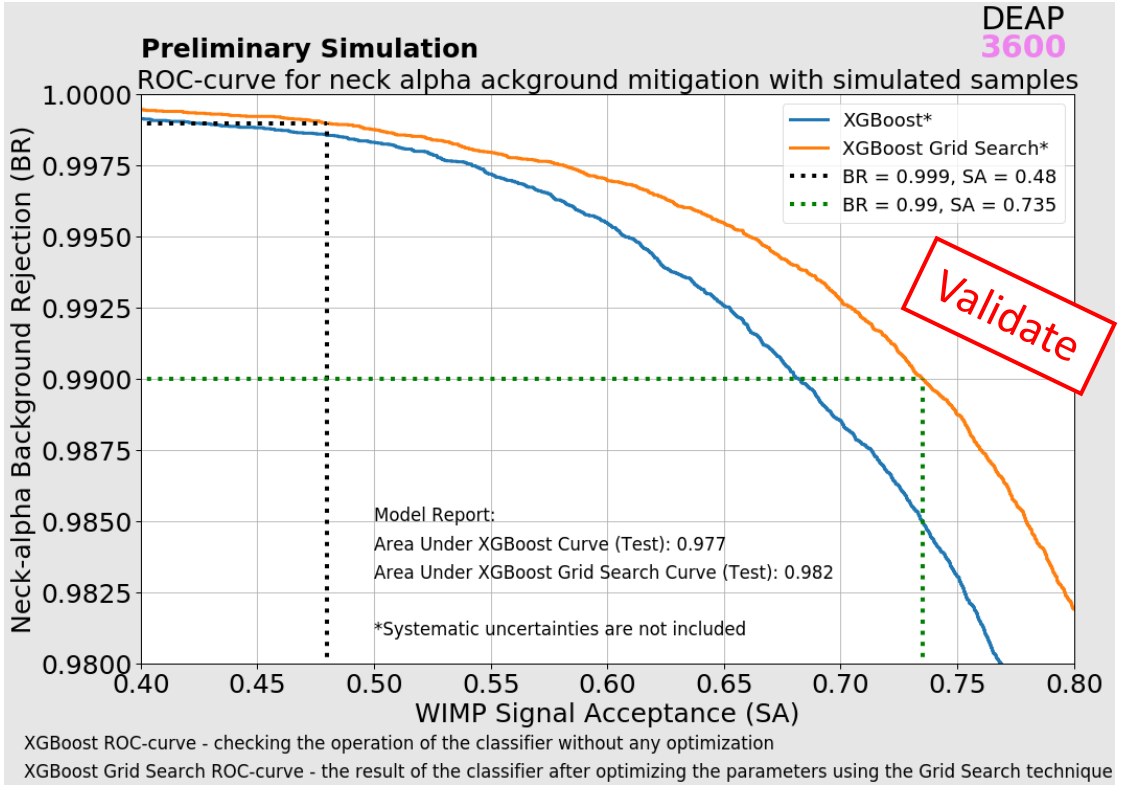
- Prepare dataset (in physics it is often relies on MC and takes up to 70-80% of time)
- Select features
- Choose ML model (try not to overkill)
- Train\validate\test -> tune model (you can gain about 10% of performance)
- Evaluate performance (calculate metrics and decide on threshold)
- Make sure that it works as intended (e.g. opening black box)
- Deploy for large-scale analysis (this is where you test a lot)

Mitigating neck alphas





(i) excluded energy estimators
(ii) We excluded features defining pulse-shape, Ar 40 and neck-alpha events share this signature
(iii) included position parameters (reconstructed X, Y, Z coordinates): given the geometry of the detector it is more likely to neck alpha event to reconstruct near the bottom of the acrylic vessel
(iv) (iv) included charge distribution patterns and number of hits in different PMTs. We are aware of some correlation between those and position features
(v) (v) PMT, which sees the first light in the gaseous phase



XGBoost ROC-curve - checking the operation of the classifier without any optimization
XGBoost Grid Search ROC-curve - the result of the classifier after optimizing the parameters using the Grid Search technique

Inference and deployment

- What to do after you trained and tested you ML model? Put it to operation.

Hidden Technical Debt in Machine Learning Systems

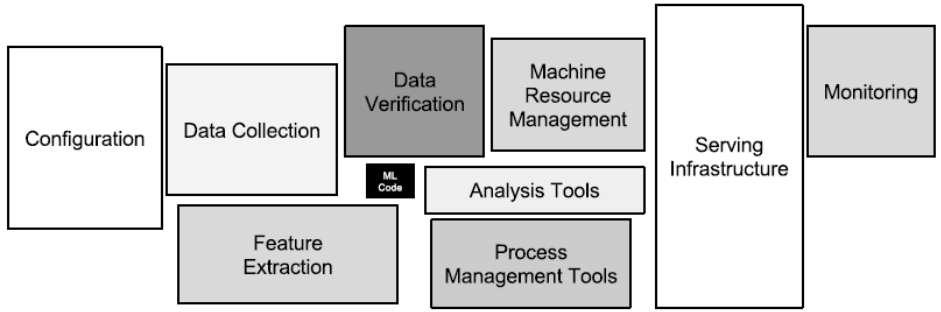
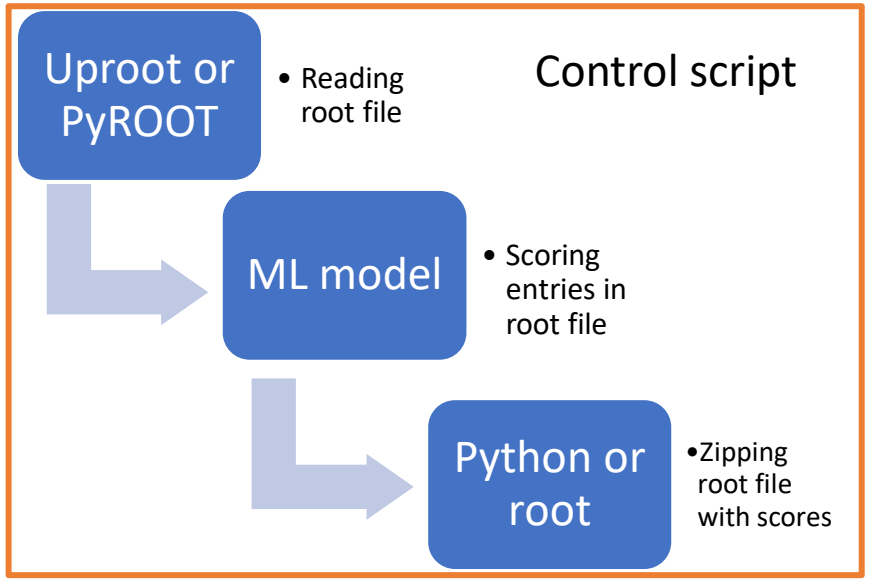
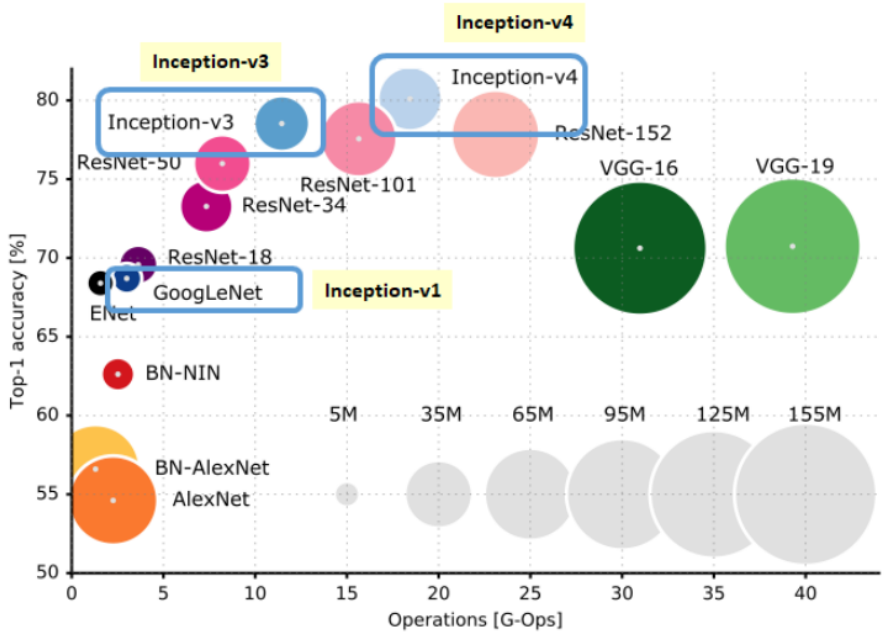


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

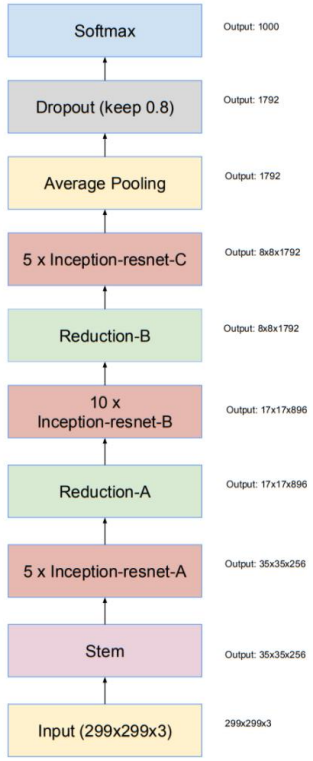


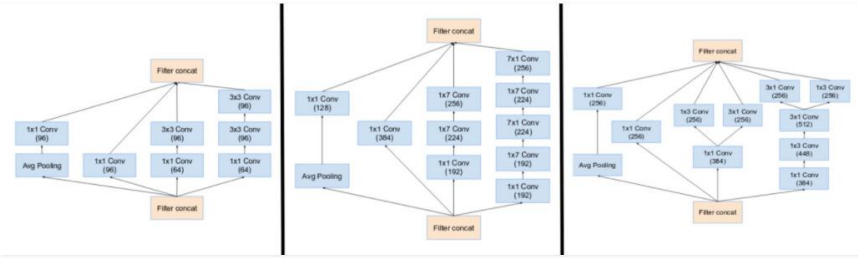
Another application for ML – Position reconstruction

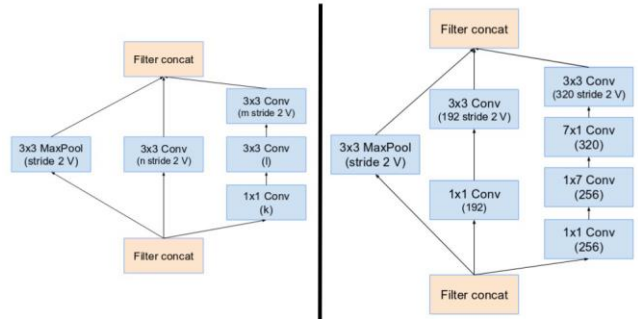
- Position reconstruction is a classical regression problem
- It can be regarded as one of the Computer Vision problems – we can use Convolutional Neural Networks!



Top-1 Accuracy against Number of Operations (Size is the number of parameters)

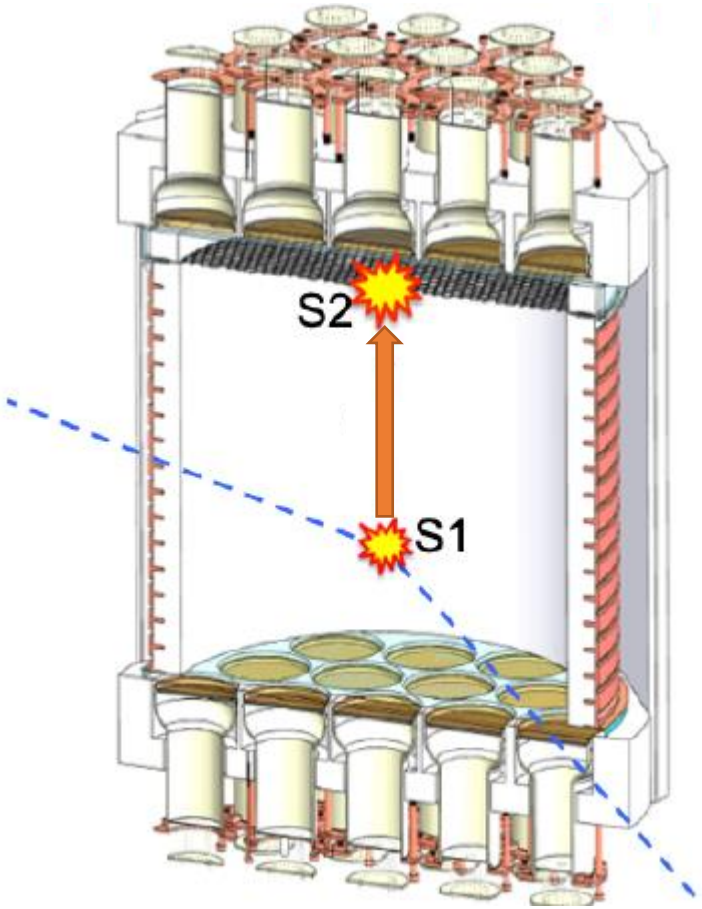




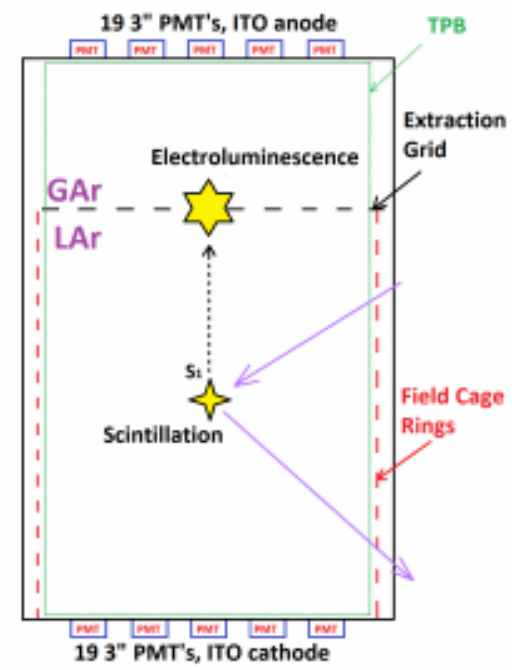


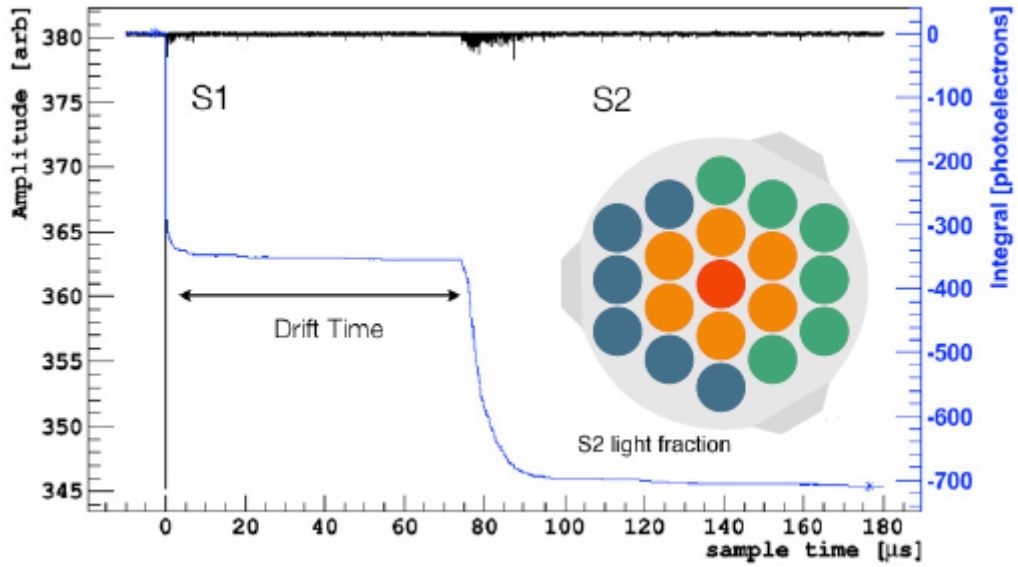
Etc…

DarkSide-50



In the double phase TPCs x-y position reconstruction relies on S2 signal



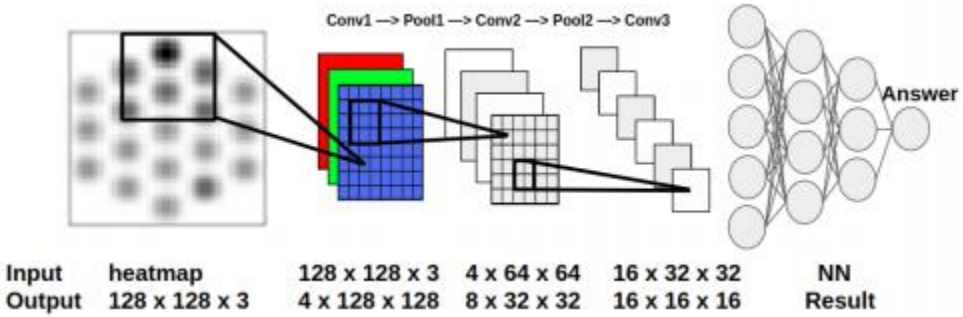


courtesy of Claudio Savarese

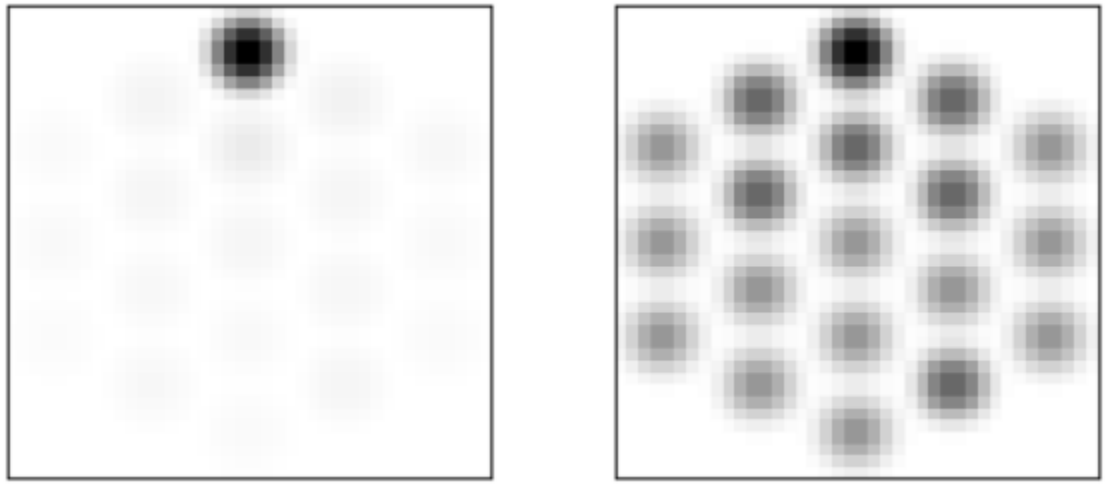
It becomes very important when people do S2-only analysis

It’s all convoluted

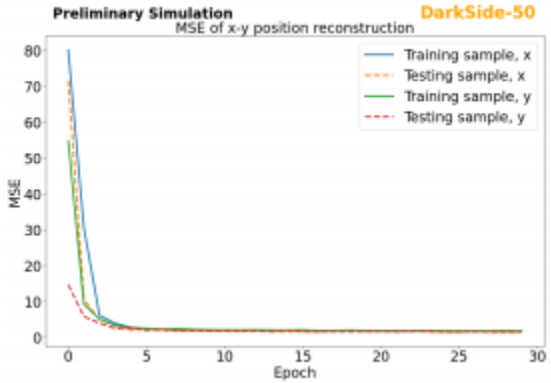
Very simple model already achieves reasonable results



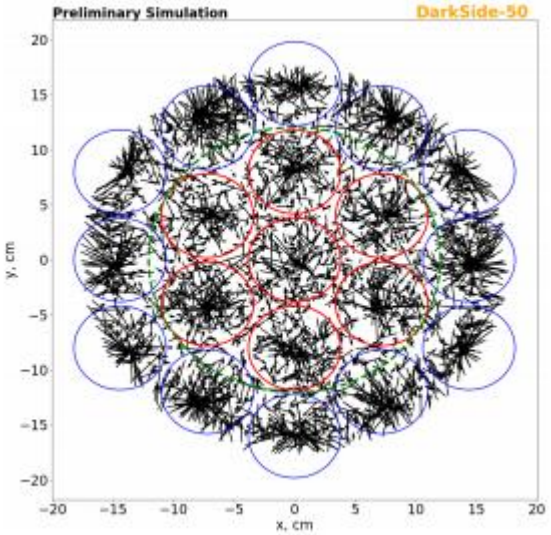
Scaling



Model converges after several epochs

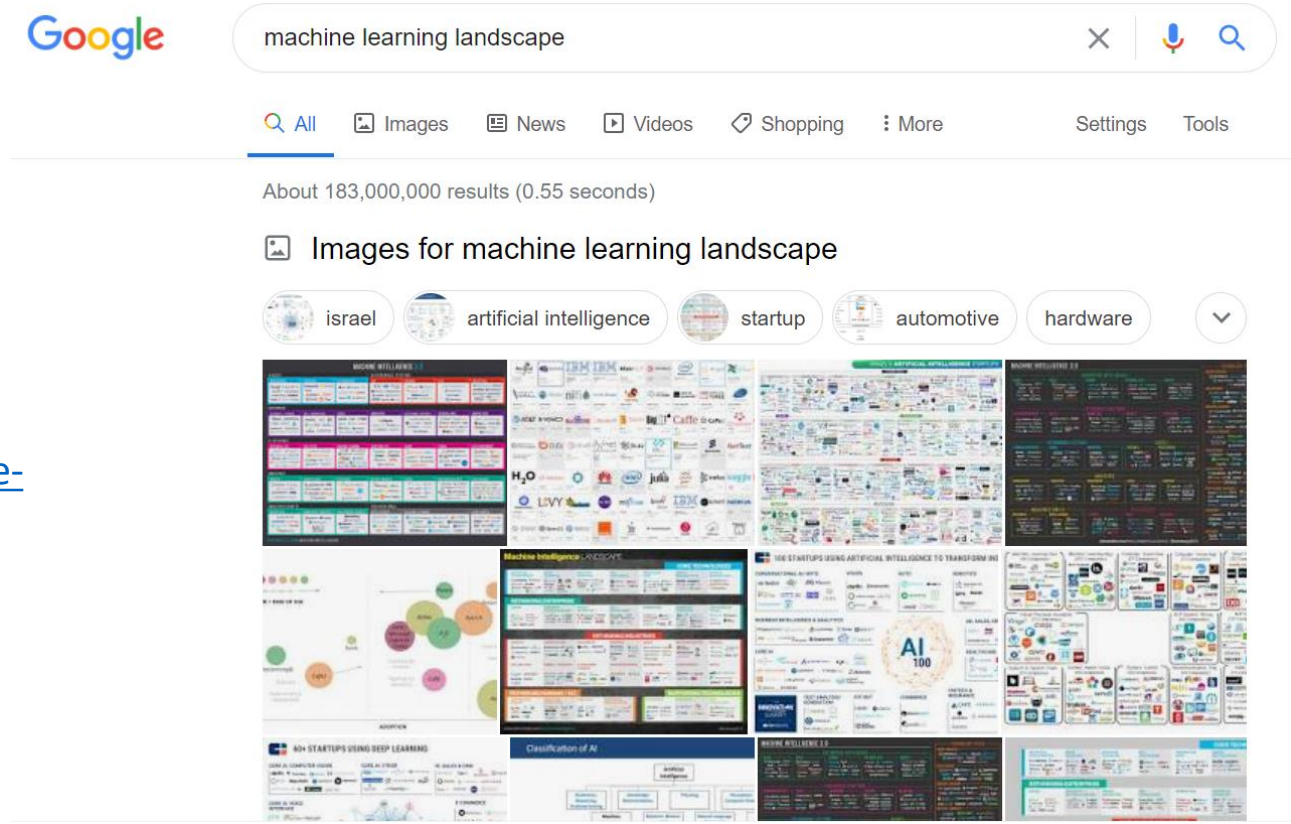


Results are yet moderate and there is definitely room for improvement: All events tend to reconstruct towards PMT centers



A little bit information

- Kaggle.com
- https://www.kdnuggets.com/2017/04/top-20-papers-machine-learning.html
- https://github.com/kjw0612/awesome-deep-vision



Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century