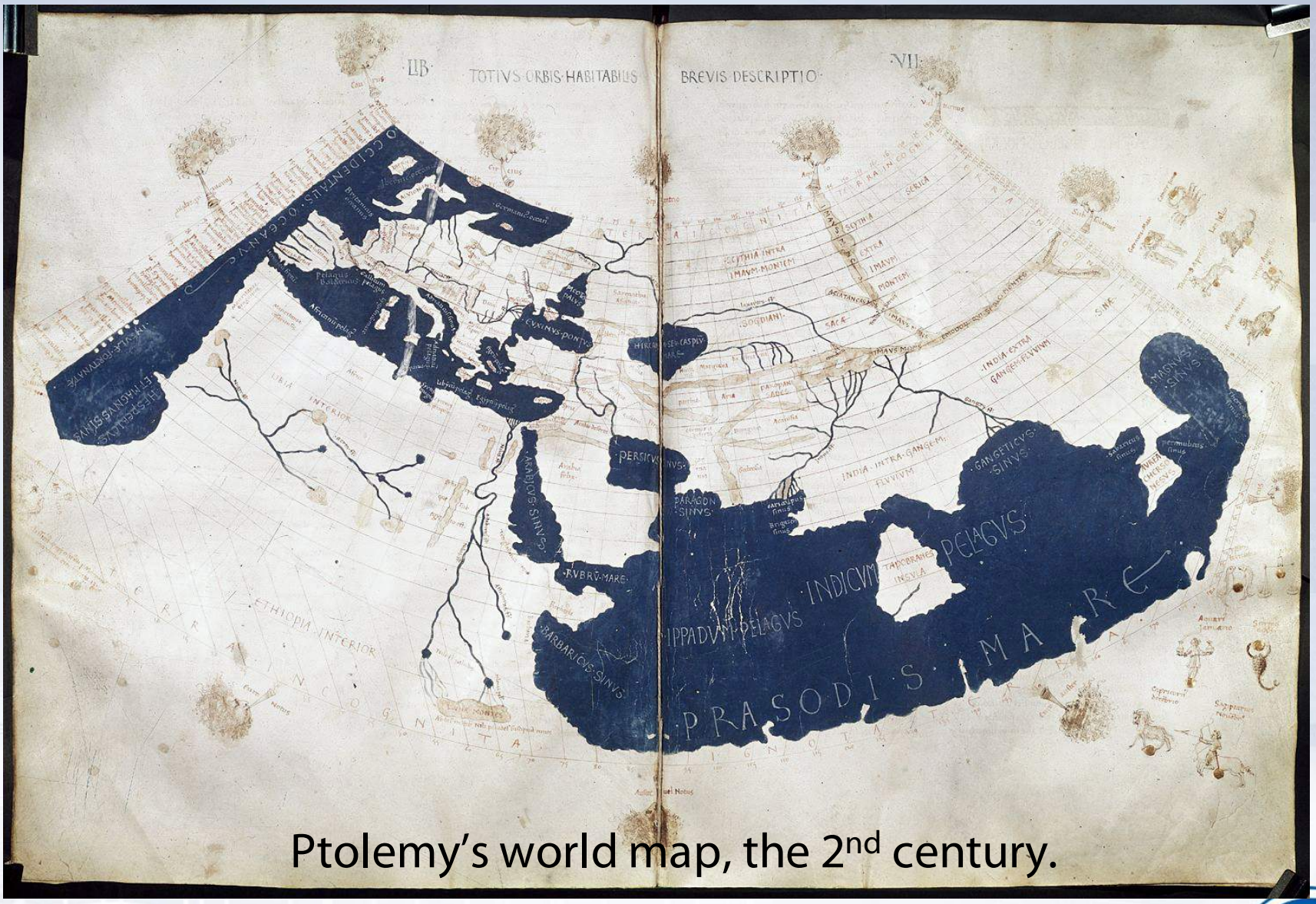
Search for Dark Matter Hints with Machine Learning at new CERN experiment

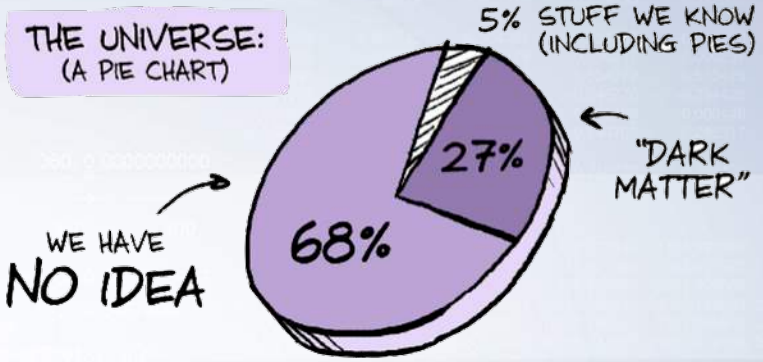


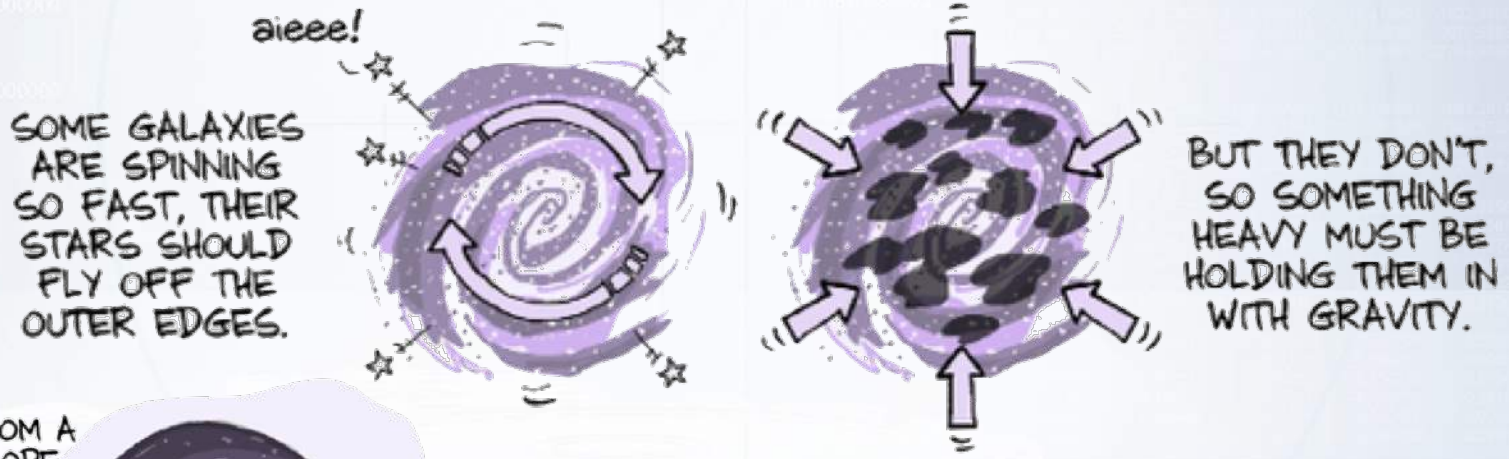
Ptolemy's world map, the 2nd century.

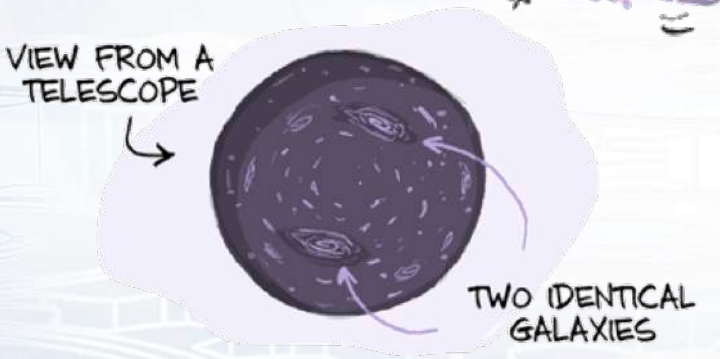
http://bit.ly/2vqSM19

Dark Matter Intro

Illustrations from J. Cham;
D. Whiteson.
"We Have No Idea"

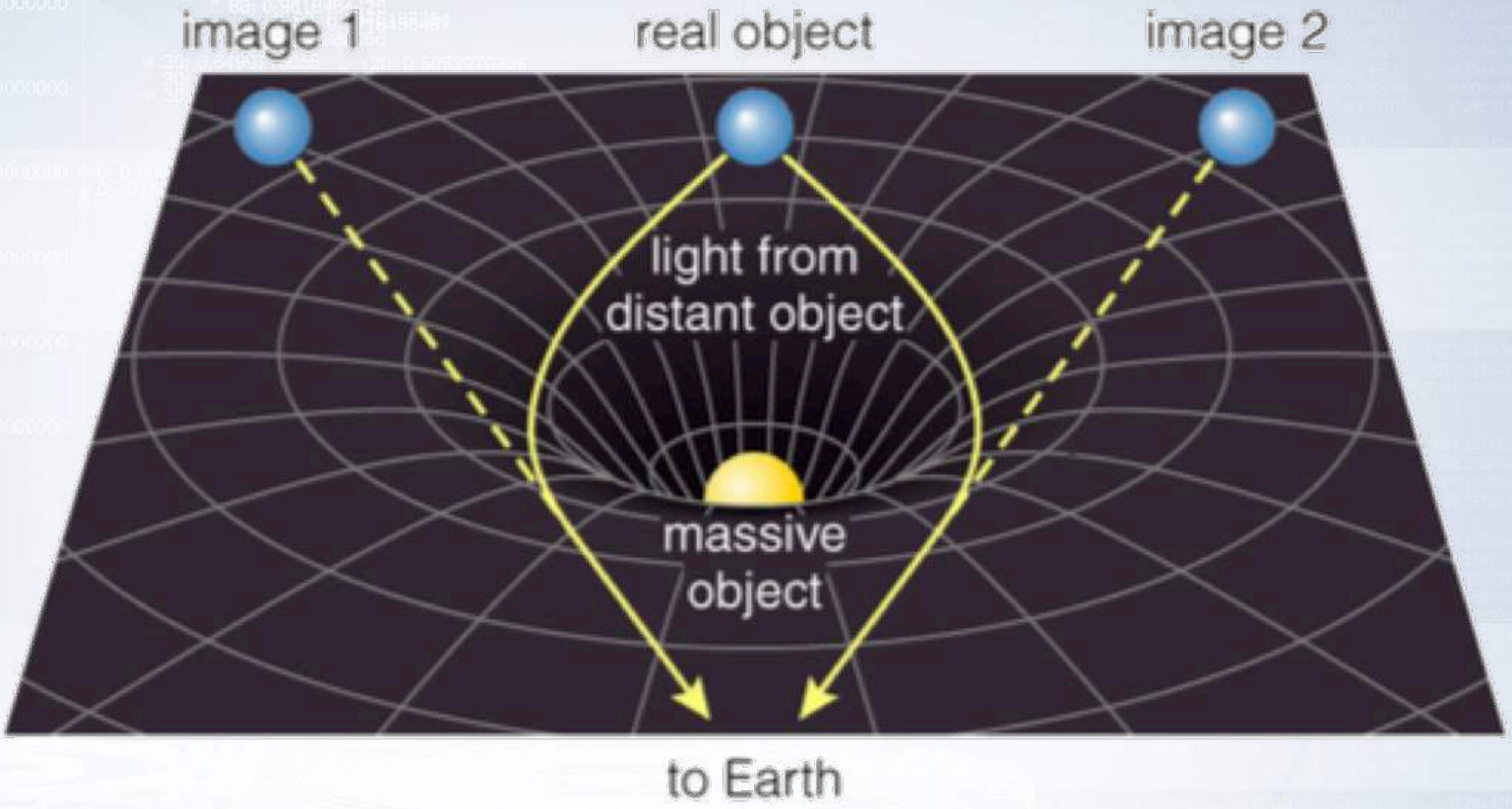






https://en.wikipedia.org/wiki/Dark_matter

Gravitational Lensing



Sometimes displacement effect is much stronger than displacement caused by visible mass in the center. Hence we get a hint for a bigger object in the middle.

https://en.wikipedia.org/wiki/Dark_matter

Why bother?

- What if Dark Matter (DM) is made of some new kind of particle that we are able to produce and study in high-energy colliders?
- What If we could find new fundamental laws of Nature?
- And what if this new discovery lets us manipulate regular matter in new ways? (e.g. new source of energy?)

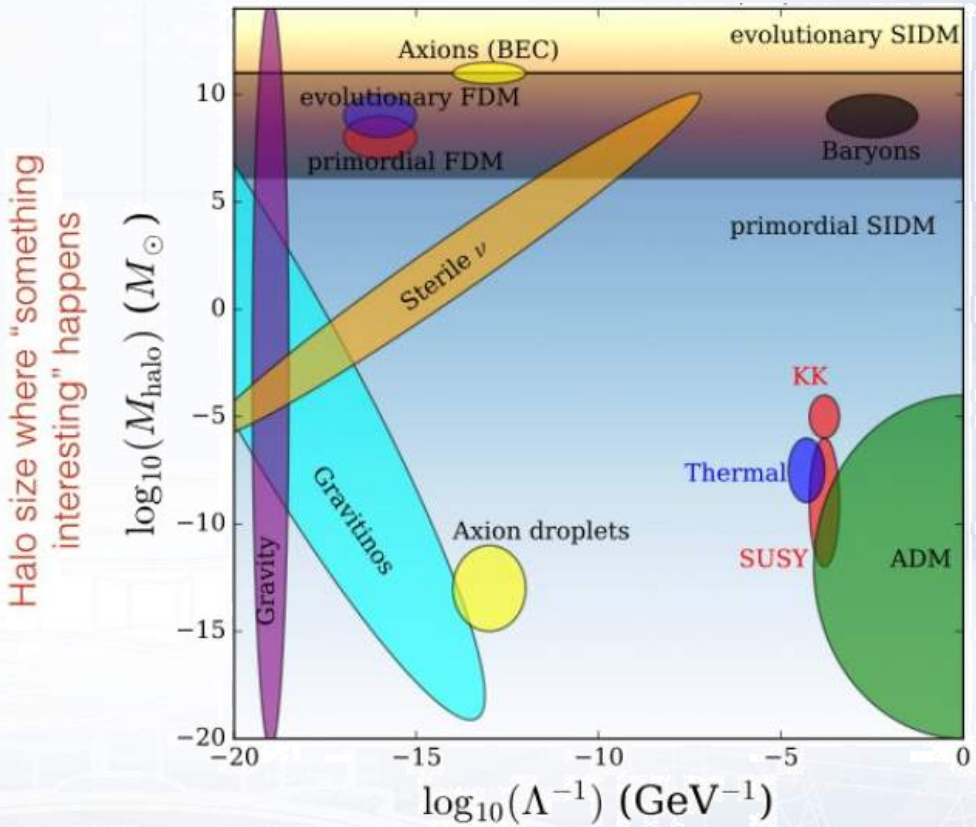
Two big challenges

1. We do not know what we search for

Unknown fundamental (microscopic) nature impacts the strength of potential signal and implies also macroscopic uncertainties;

2. There is no (totally) clean observable

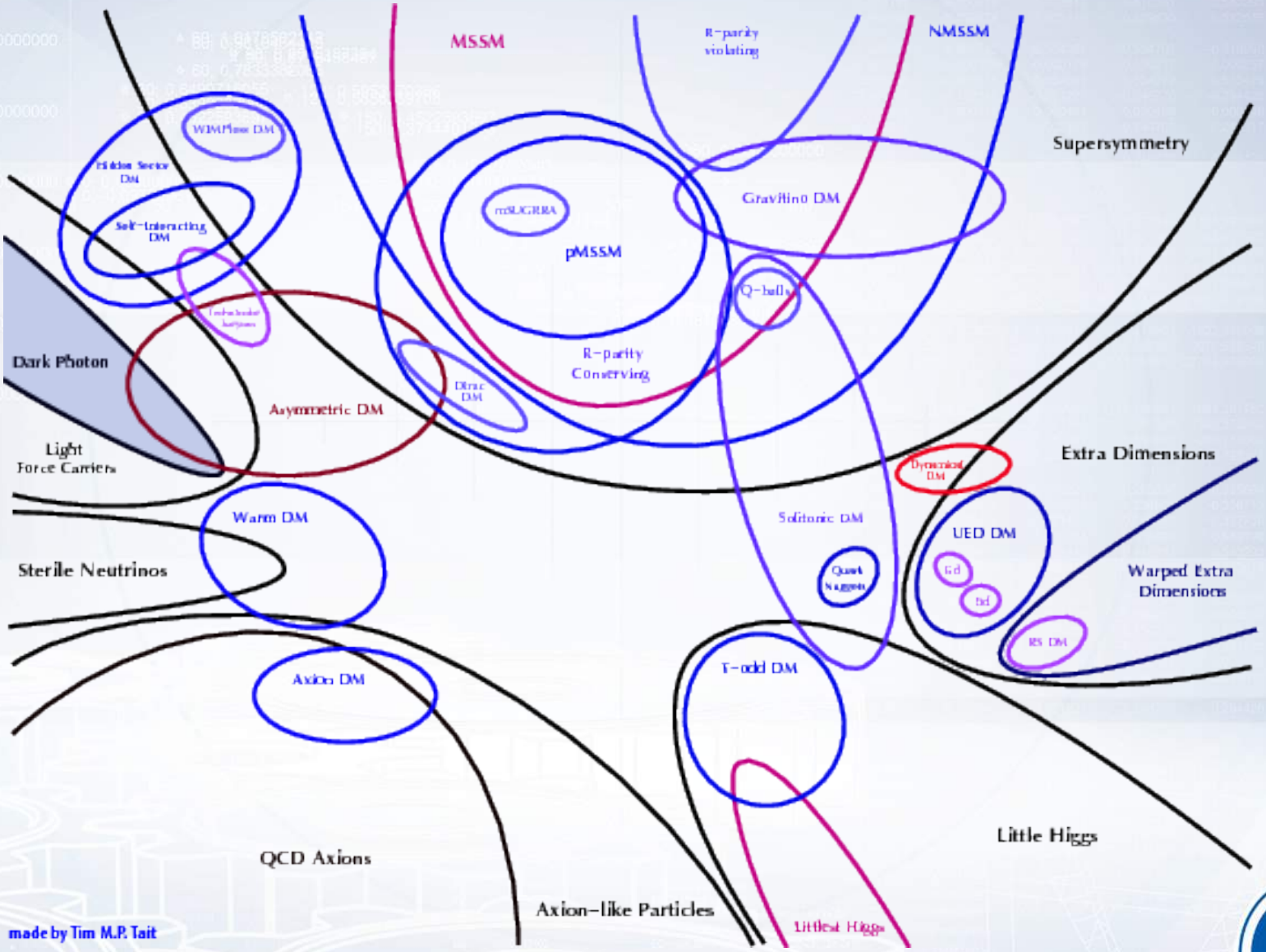
Direct/indirect detection targets cosmo signals, where there are many other players, “backgrounds” are typically poorly known.



http://bit.ly/2IRPINV

How much dark matter interacts with us

Sea of Models

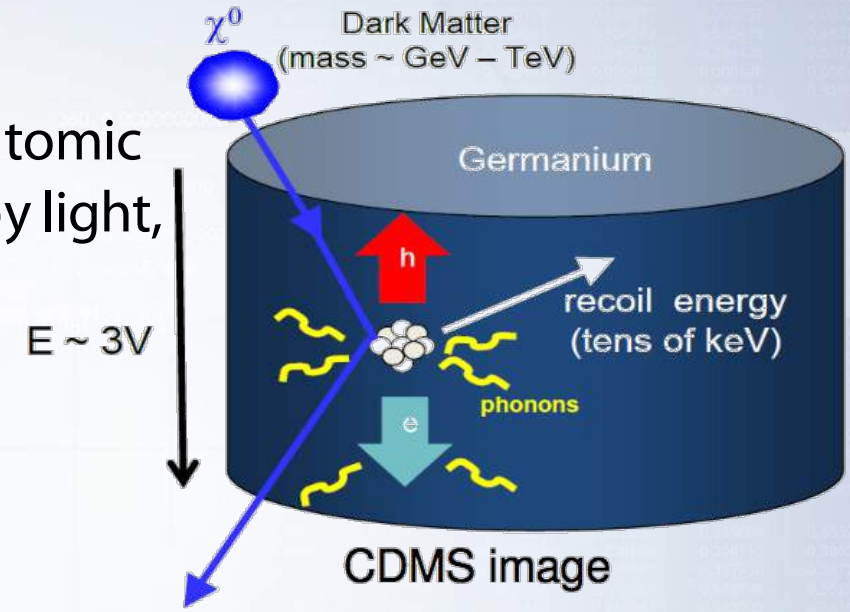


http://bit.ly/2uq3OmT

Two strategies

1. Direct

Scattering of DM particle target atomic nuclei. Recoil energy measured by light, charge or phonons.

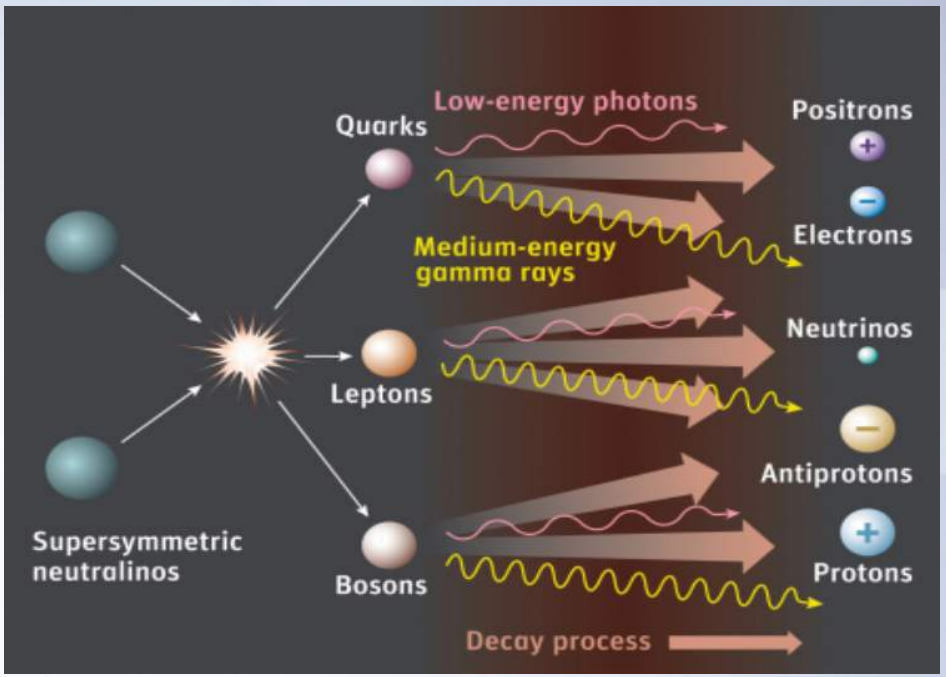


Experiments (several examples):

DAMA/LIBRA, ANAIS,
KIMS, DM-Ice, PICO-LON, SABRE,
Nuclear emulsion (NEWS),
Anysotropic crystals (ADAMO),
Liquid Ar TPC, Negative Ion Time Expansion Chamber (NITEC), Carbon nanotubes, DRIFT, MIMAC, DMTPC, NEWAGE, D3

http://bit.ly/2IRPINV

Two strategies



2. Indirect

Annihilations (or decays) of DM particles in astrophysical objects generate fluxes of "standard" detectable particles.

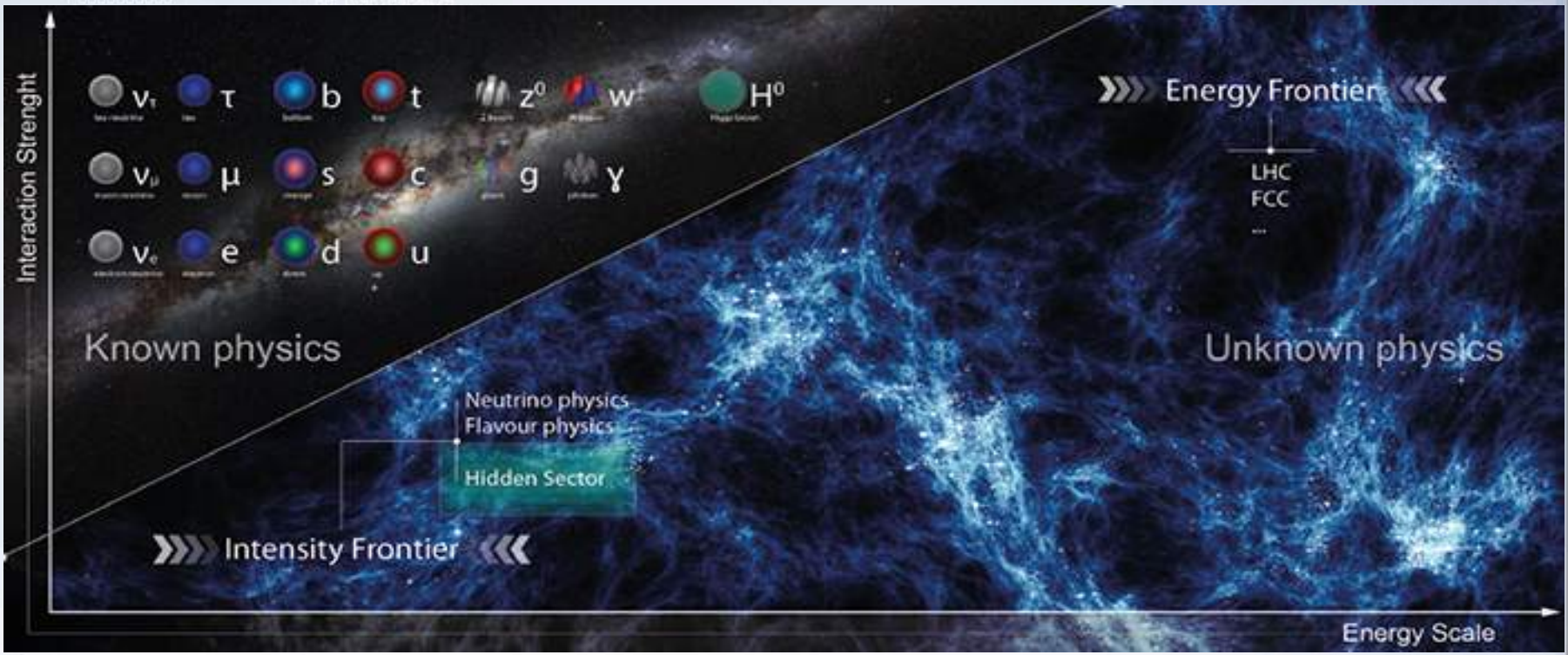
Non trivial to discriminate from the background.

Thus we have to include **accelerator searches** for Dark Matter (Hidden Particles).

More details on DM search at http://bit.ly/2uhwfDk

Search for Dark Matter at Accelerator Experiment

Accelerator Search Frontiers



- Energy Frontier
  - Heavy particles → high energy collisions
- Intensity Frontier:
  - Weakly interacting particles → high intensity beam

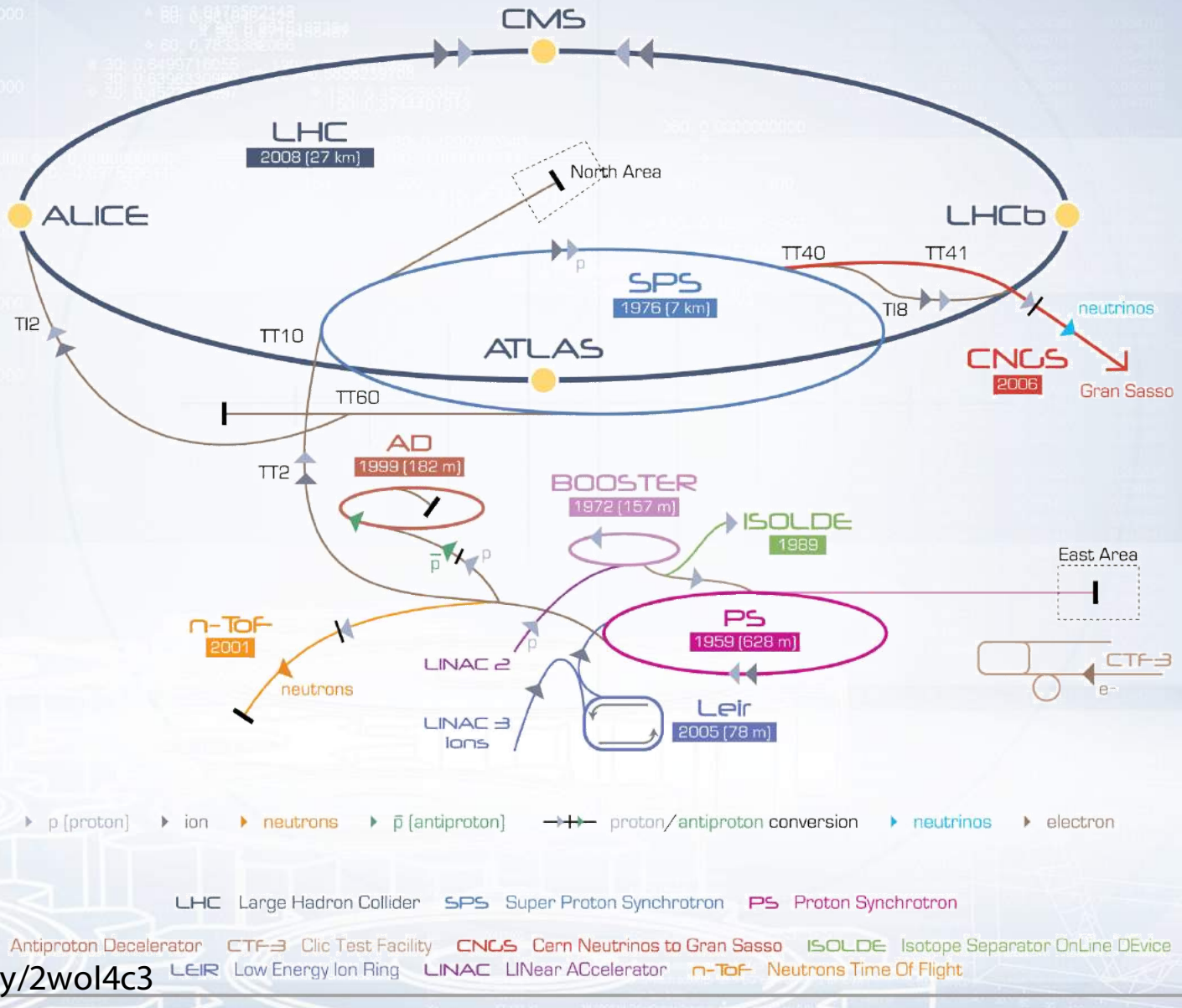
Dark Matter Candidate

Many theoretical models (in particular portal models) predict new light **very Weakly Interacting Massive Particles** (vWIMP) that can be mediators to DM, or even DM particles.

References:

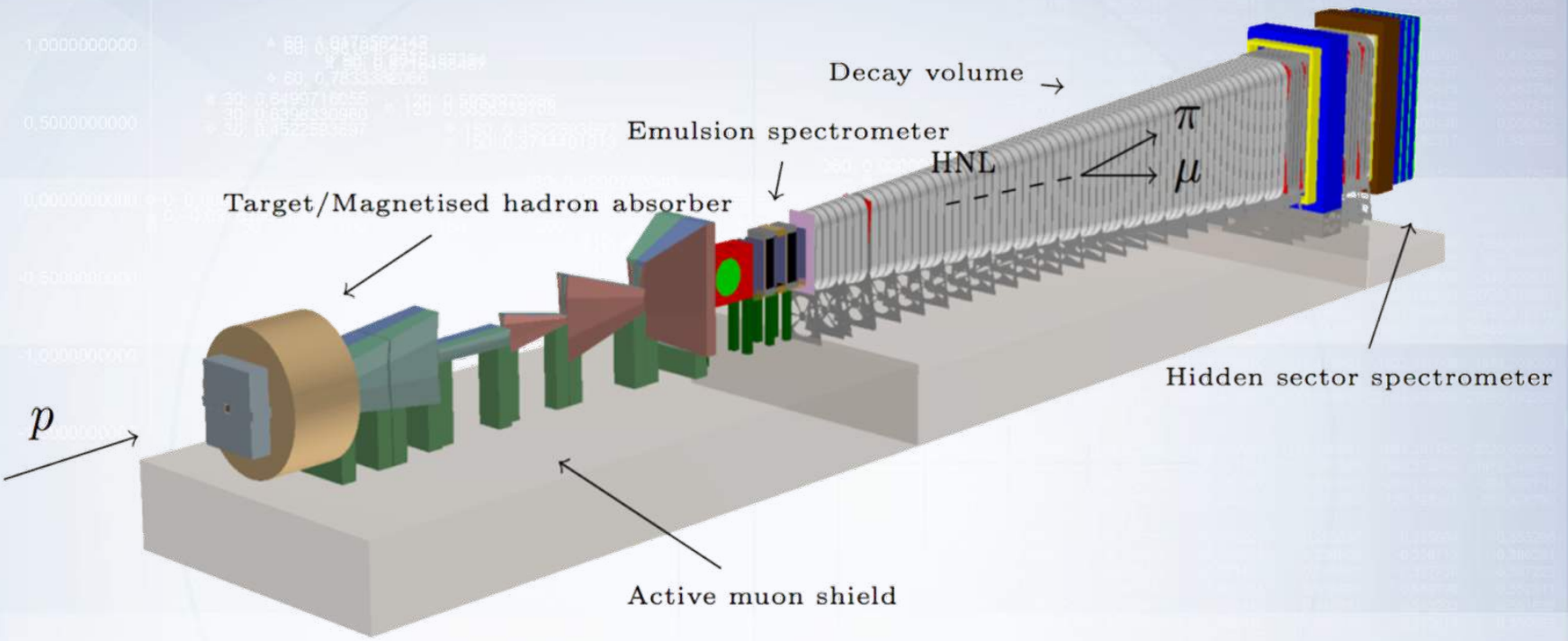
- SHiP Physics Paper: Rep.Progr.Phys.79(2016) 124201 (137pp),
- Dark Sector Workshop 2016: Community Report – arXiv: 1608.08632.

CERN Accelerator Complex



http://bit.ly/2woI4c3

SHiP: Search for Hidden Particles



Light Dark Matter (LDM) Search Experiment

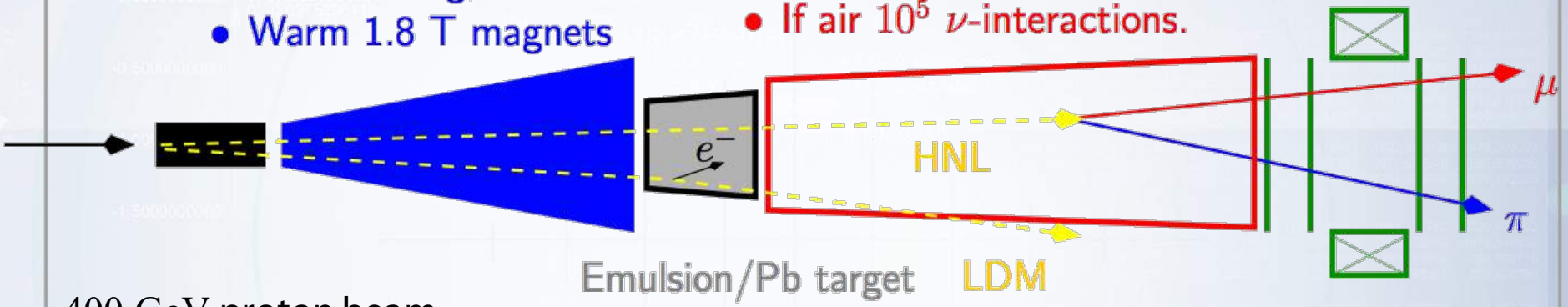
Data taking is expected at 2025+

http://ship.web.cern.ch/ship/

SHiP in a Nutshell

Magnetic μ-shield (10^{11} μ/s)
- $\sim$ 35 m long, 2 kton Fe
- Warm 1.8 T magnets

Vacuum decay-vessel surrounded by scintillators
- "nothing" in, but HS-decay products out.
- If air 10^5 ν-interactions.



Emulsion/Pb target LDM

LDM particles scatter on e⁻

400 GeV proton beam
$2x10^{20}$ protons in 5 years
Produces variety of exotic
Particles (LDM candidates)

Tracker
Muon detector

SHiP challenges

Physics challenges:

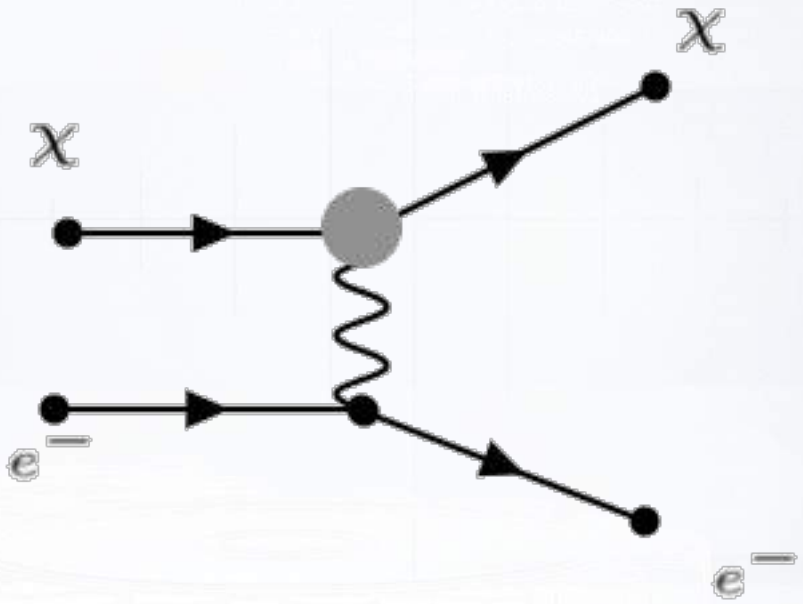
- Variety of Hidden Sector portals exploration
- Tau neutrino physics
- Light Dark Matter (LDM) Search

Engineering challenges

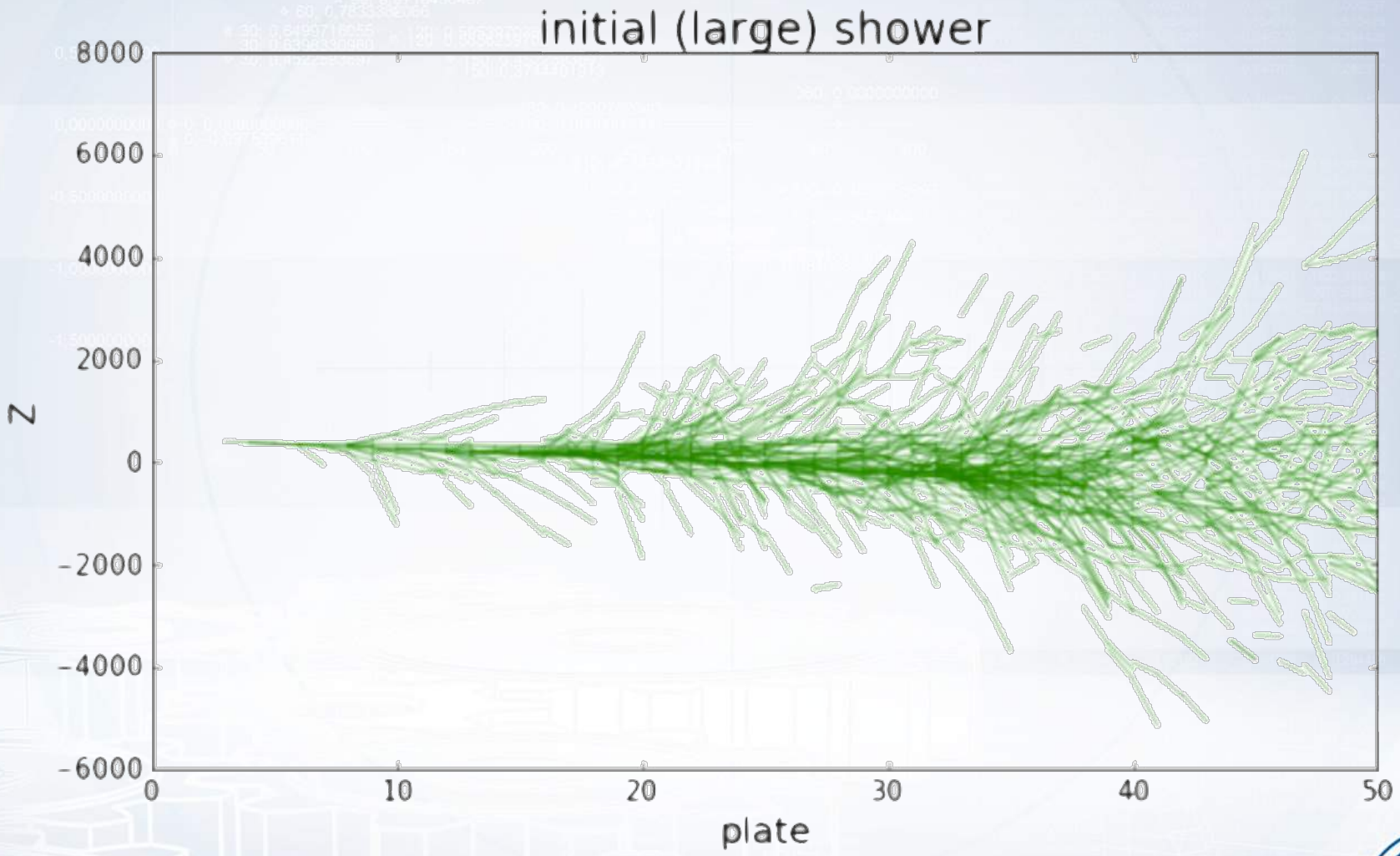
ML challenges:

- Experiment design (shield, emulsion optimization, tracker)
- Fast simulation
- Speed-up data processing
- Signal/background separation in emulsion

Light Dark Matter Signal



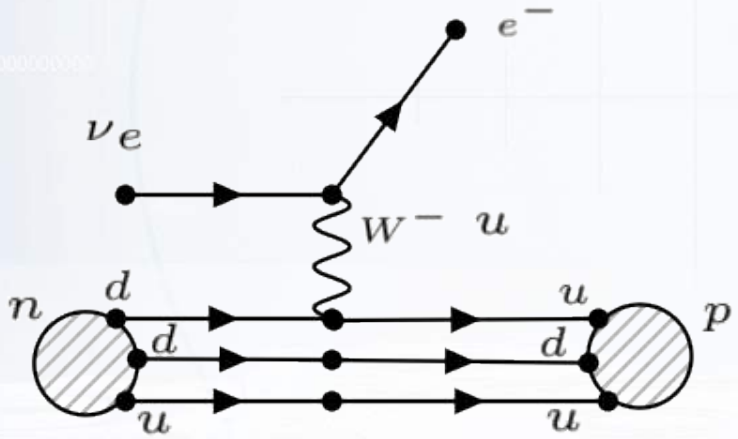
Electromagnetic shower



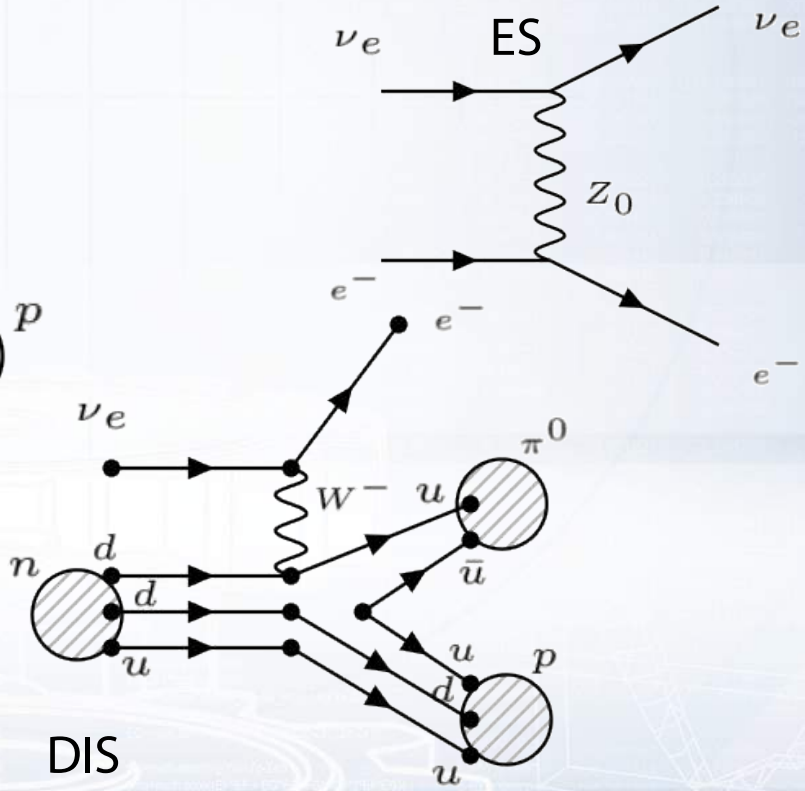
Background

Dominant background comes from **neutrino** interactions:

- Elastic scattering on electrons (ES)(topologically irreducible);
- Quasi-elastic scattering (QE) (nuclei neutron);
- Deep inelastic scattering (DIS).



QE



DIS

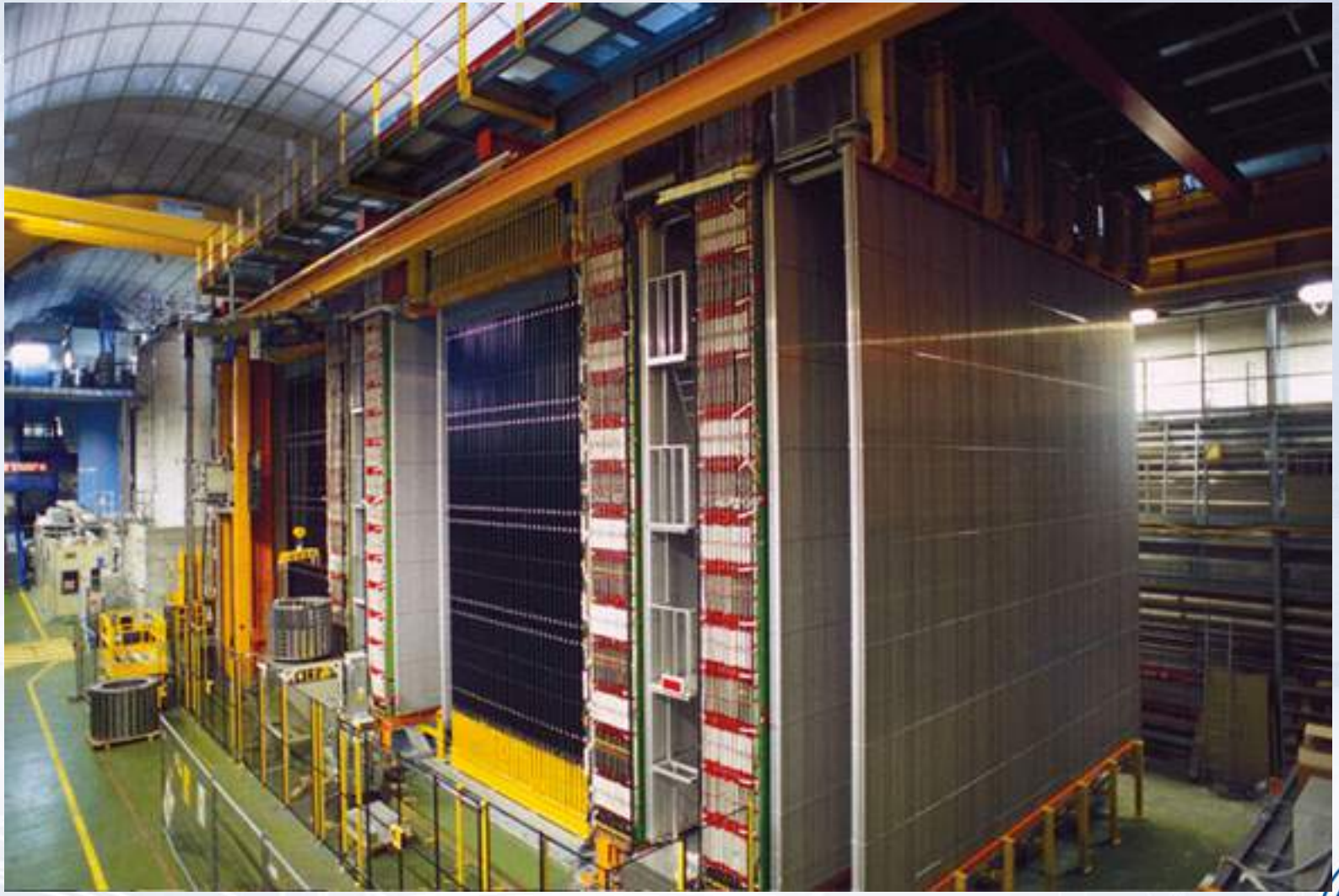
Signal/Background separation

- Find electromagnetic shower;
- Final state is different (QE produces **proton,** DIS produces **hadron jet**), so have to be able to identify protons and jets;
- Use energy-angle correlation of the detected electron to discriminate vWIMP against neutrino. Those differences are due to the mass differences between vWIMP and neutrino;

Emulsion has superior sensitivity to identify those processes and such technology has been developed for search for neutrino oscillation at OPERA experiment.

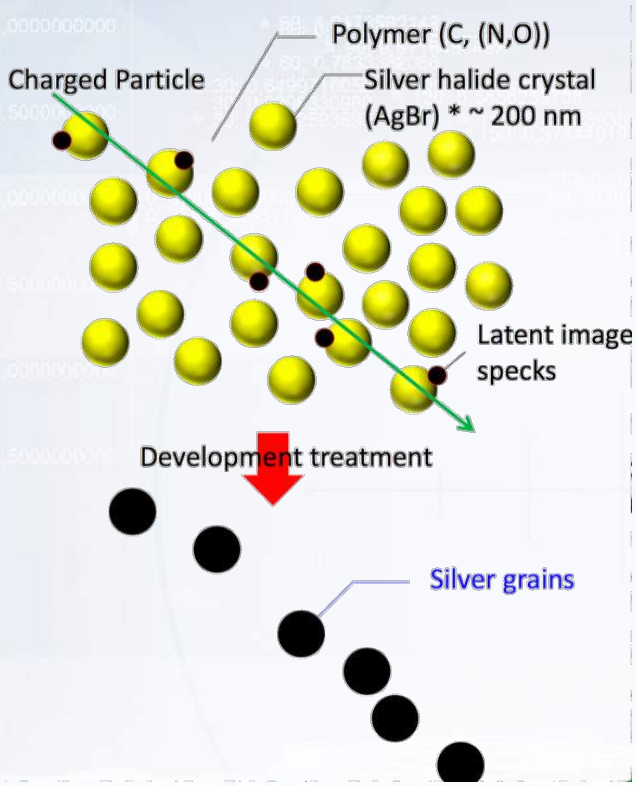
Getting Data Before Experiment is built

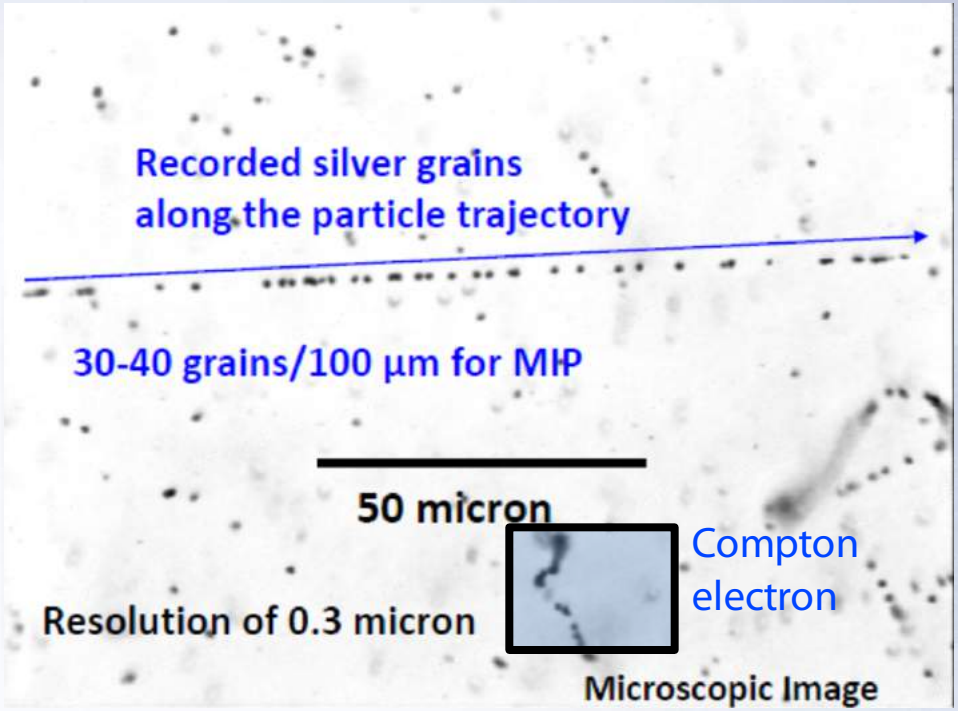
Developed for OPERA experiment (2008-2012)



http://bit.ly/2Gce7wa

Scattering and Nuclear Emulsion



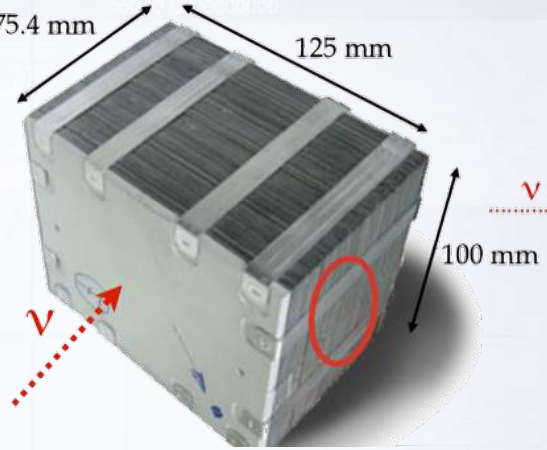


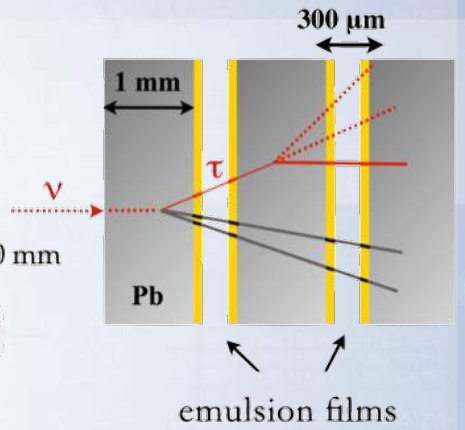
- After the passage of charged particles through the emulsion, a latent image is produced;
- The emulsion chemical development makes silver grains visible with an optical microscope.

Emulsion Cloud Chamber

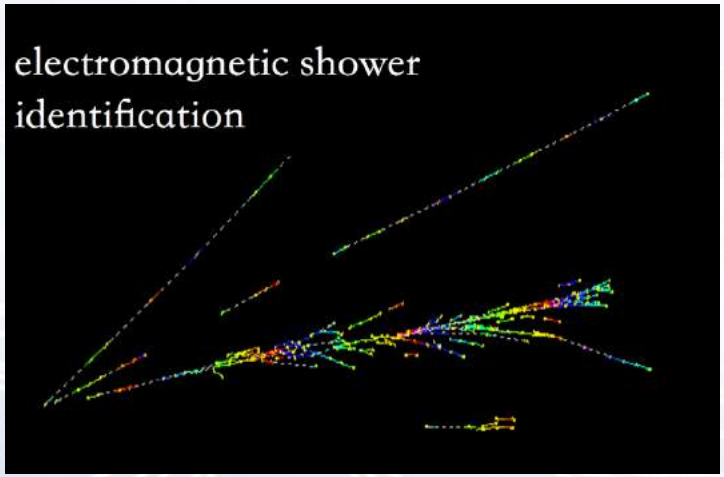
Sandwich-like structure: lead (massive target) + emulsion

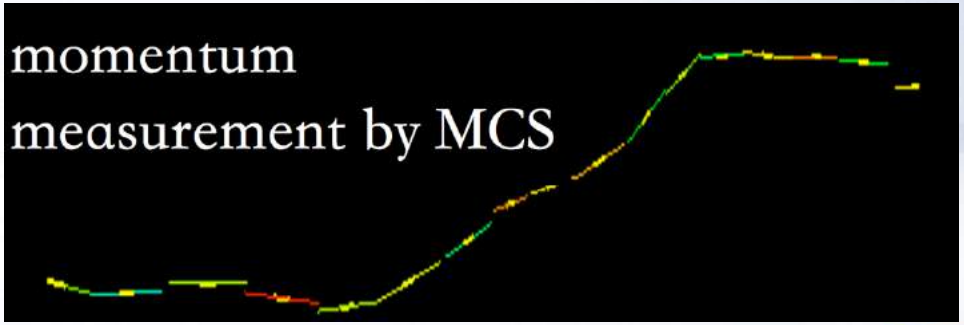
Size of the shower mostly depends on target density (Molière radius)



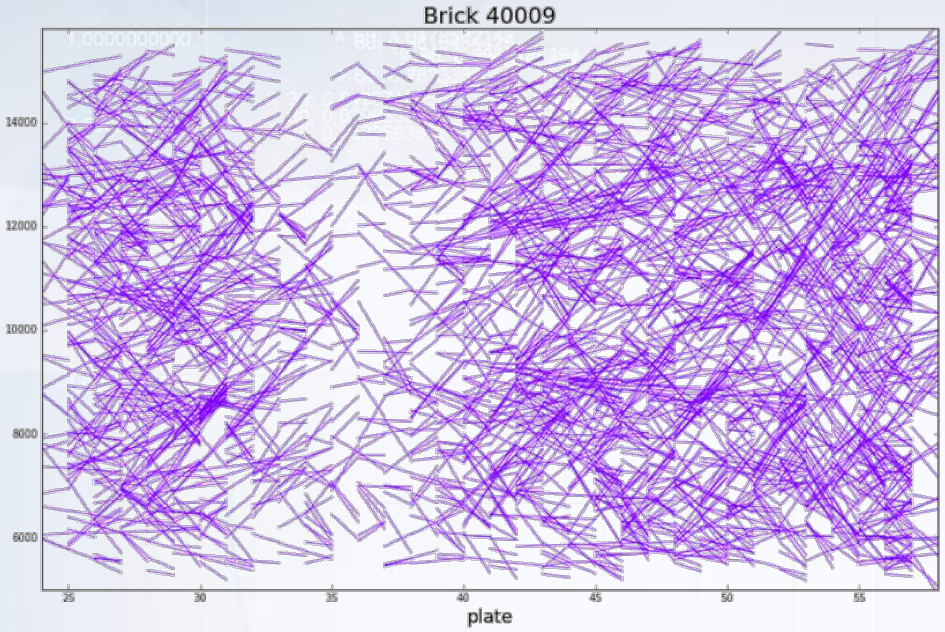


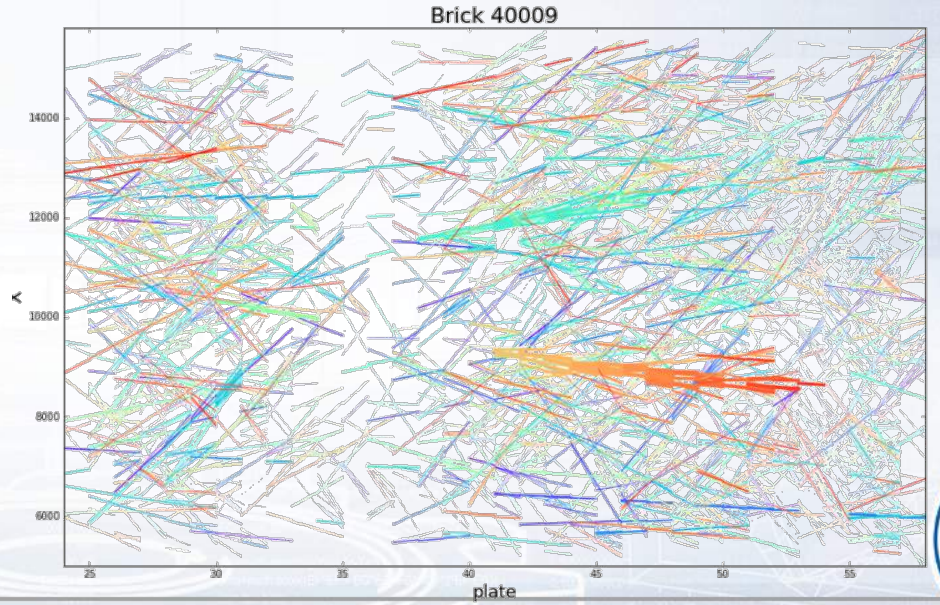
Reconstructed trajectories:



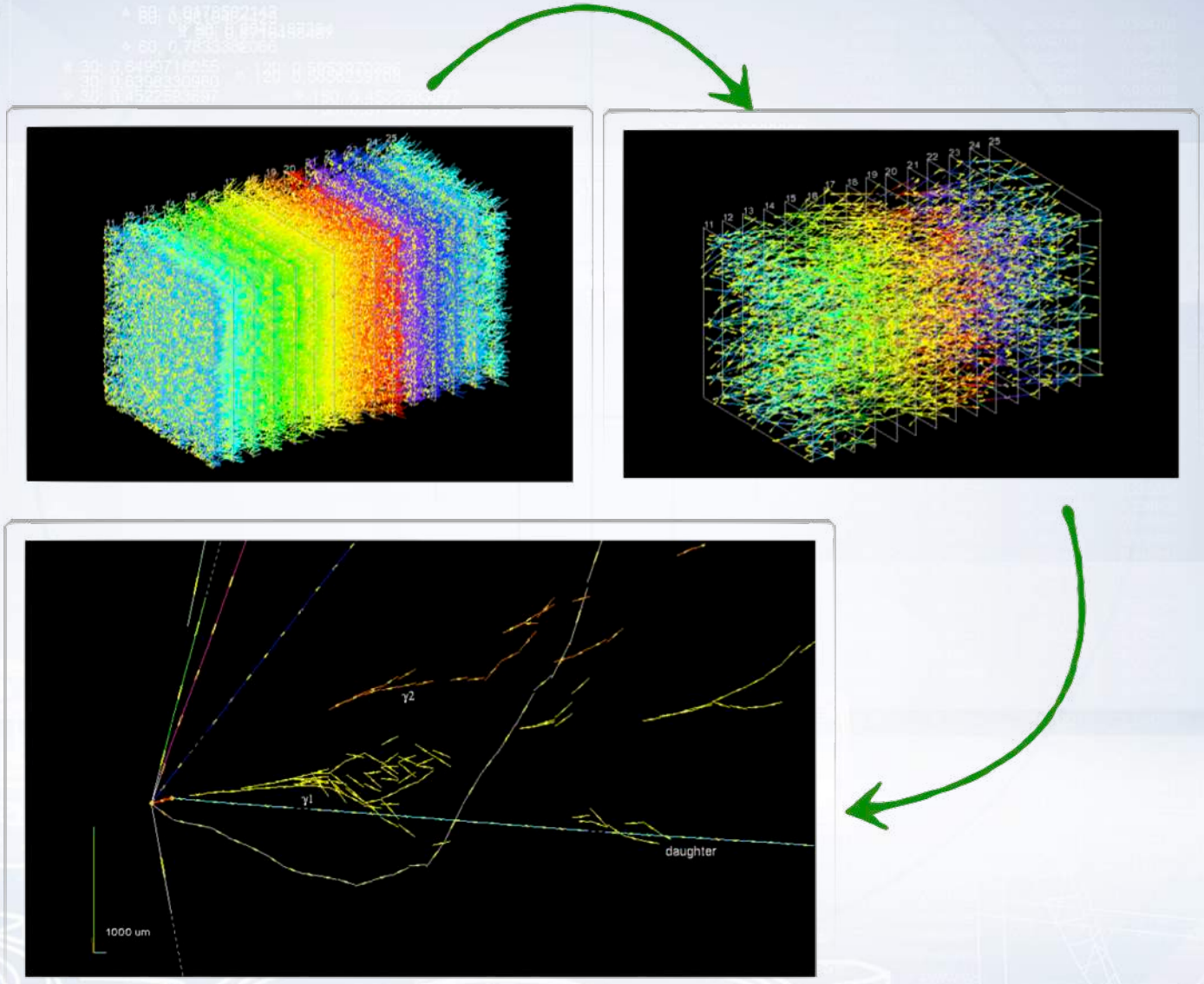


Electromagnetic shower + noise





Electromagnetic shower + noise, realistic



Machine Learning Challenges for

- Tracking in high density environment (both for single tracks and showers)
- Vertex reconstruction
- Particle identification

OPERA Data Example

Data:

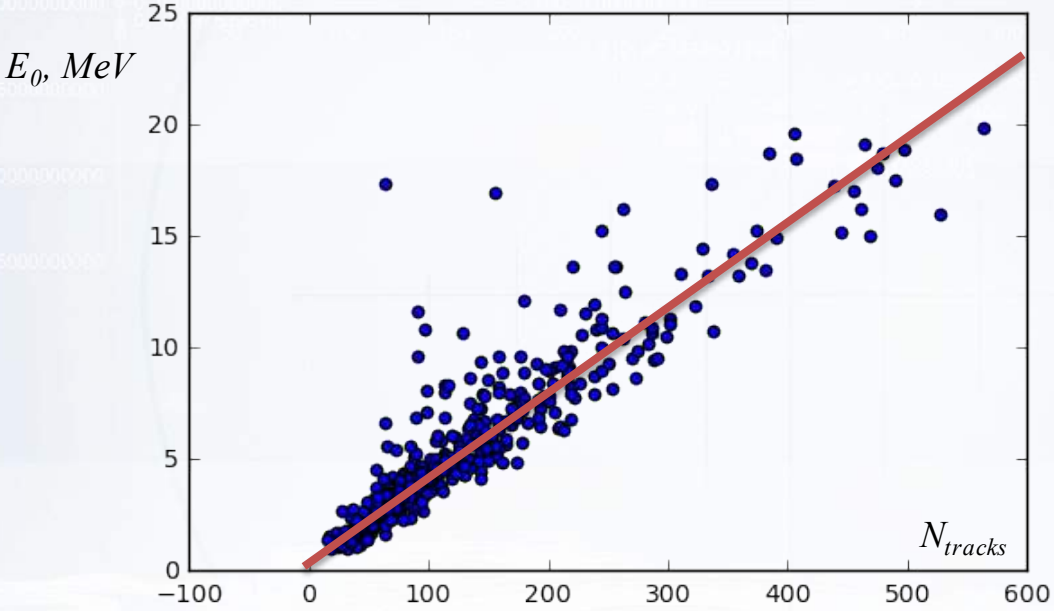
- Background consists of tracks randomly scattered around brick. In real brick there are ~ 10^7 tracks.
- Signal consists of tracks forming a cone-like shape. There are about 10^3 tracks per shower.
- Origin (coordinates and angles of the initial particle) of each shower is known.

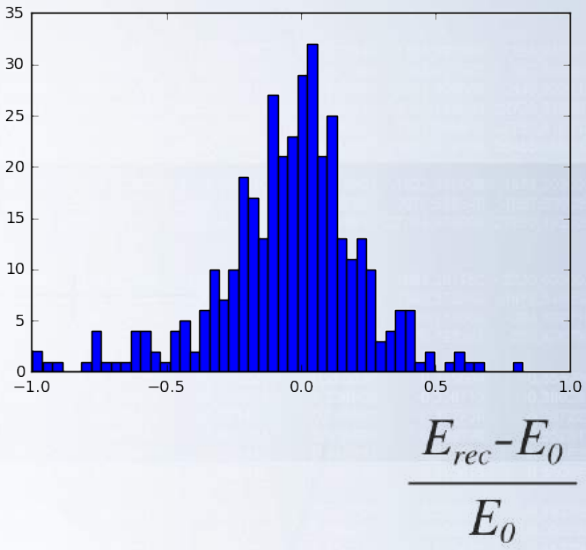
OPERA Example

- Each BaseTrack (BT) is described by:
  - Coordinates: (X, Y, Z)
  - Angles in the brick-frame: (TX, TY)
  - Goodness of fit of Ag crystals to the BT: χ
- Background consists of basetracks (BT) randomly scattered around brick. In real brick there are ~ 10^7 tracks, `label=0`
- Signal consists of BTs forming a cone-like shape. There are about 10^3 BTs per shower. `label=1`
- Origin and Energy of the shower is known $(X_0, Y_0, Z_0, TX_0, TY_0, E_0)$

Figure of Merit: Energy Resolution

- For every shower of energy E_0 we reconstruct number of base tracks (N) that approximate it's energy





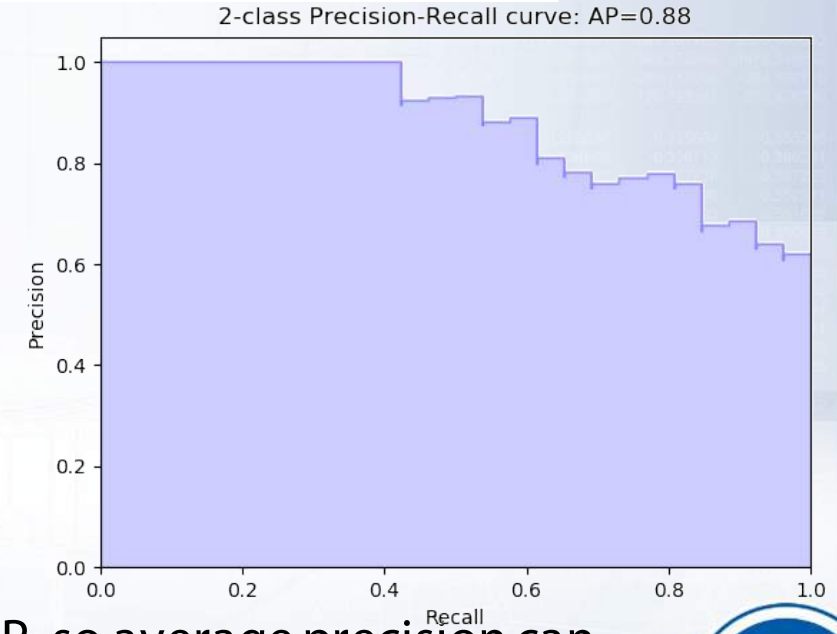
- So $E_{rec} = a\,N + b,\ \ (a,\ b)$ can be approximated by linear regression (left);
- Energy resolution E_{res} is a standard deviation of relative residuals (right);

Figure of Merit Proxy

- In terms of ML, we can estimate the following simple metrics for every algorithm, giving predictions for a BT to belong to `label=1`

	True label = 0	True label = 1
Predicted label = 0	True negative (TN)	False negative
Predicted label = 1	False positive	True positive

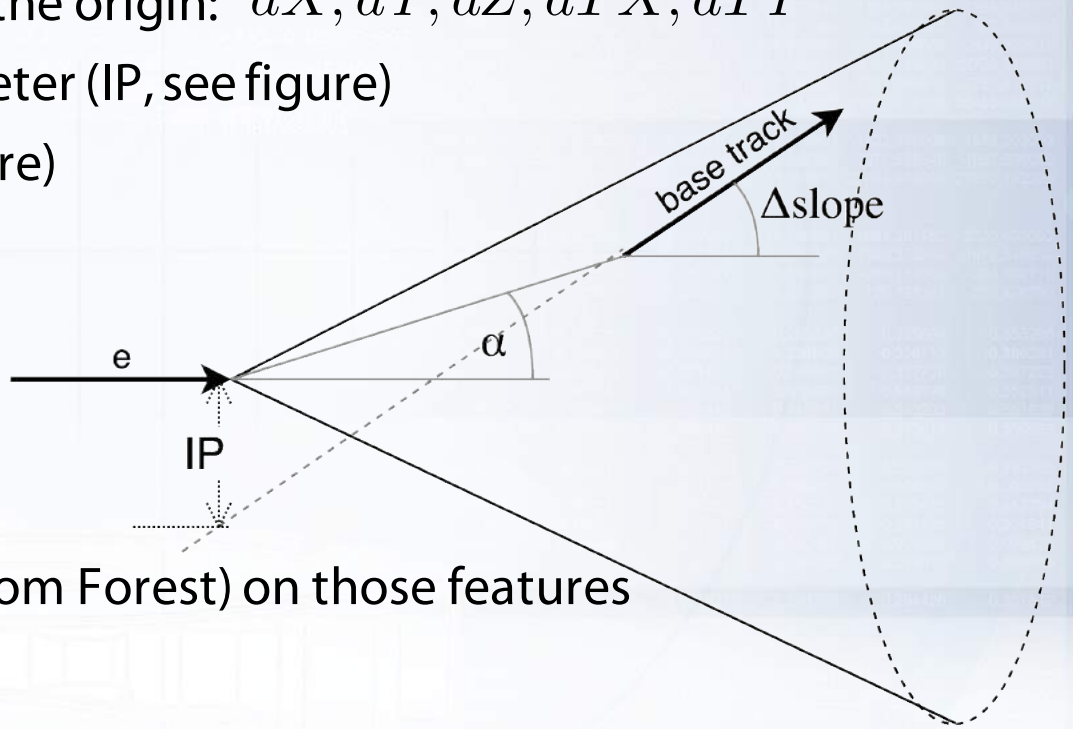
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- If algorithm gives predictions as a float-point number [0, 1], we make plot Precision/Recalcurvel



- Number of BT correspond to TP + FP, so average precision can serve as a proxy of classifier quality. Or similarly - ROC AUC can.

Baseline solution, given origin

- Consider only tracks within cone volume (50 mrad)
- Iterate through all BTs in the cone volume:
  - Compute distance from the origin: dX, dY, dZ, dTX, dTY
  - Compute Impact Parameter (IP, see figure)
  - Compute alpha (see figure)



- Train classifier (e.g. Random Forest) on those features
- Metrics:
  - ROC AUC
- Baseline result: ~0.96, precision ~1.0 at 0.5 recall

Going Deeper

Tougher Challenge

- No origin information is known apriori;

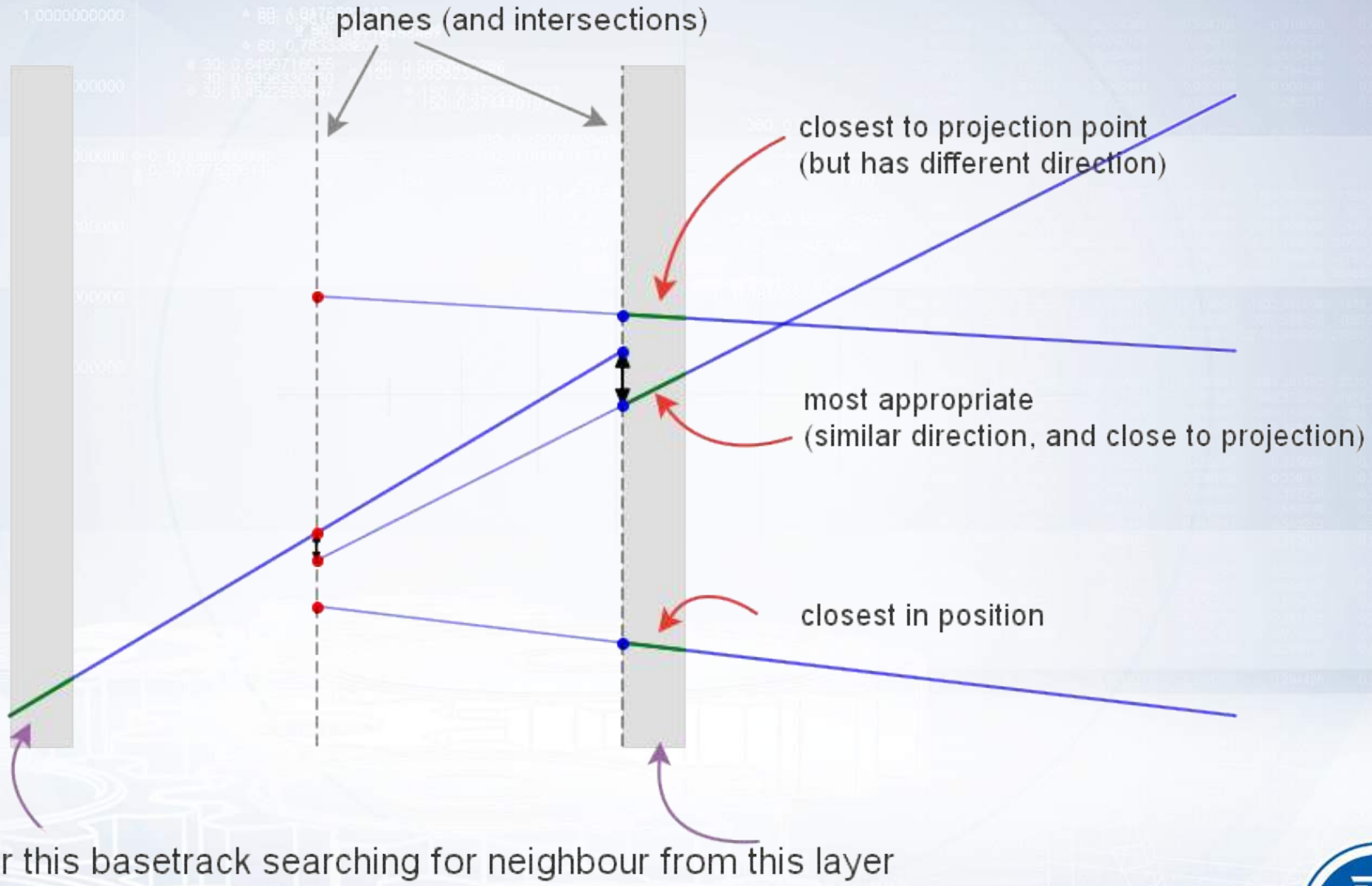
Methods to apply:

- Clustering;
- Conditional Random Field;
- Message Passing Neural Networks.

Possible solution for no-origin case

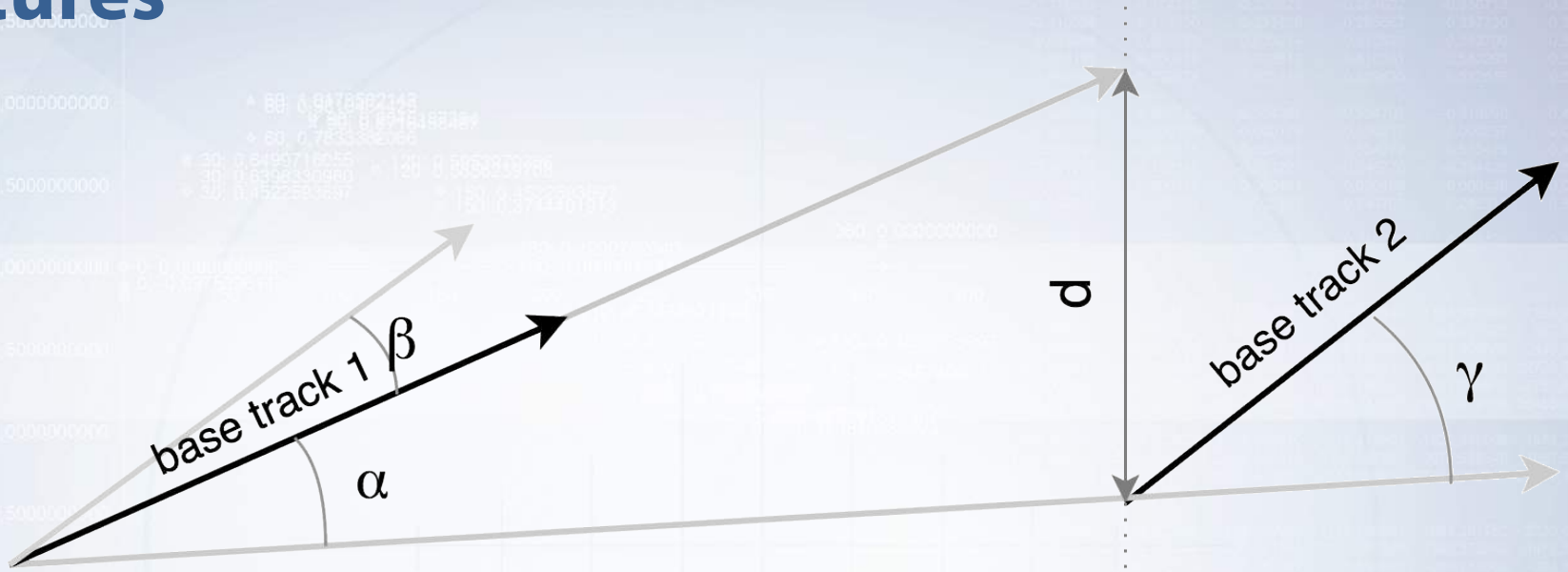
- Find Neighbors for selected tracks
- Build chains of 5-track candidates
- Train classification algorithm dealing with 5-tracks
- Cluster showers using DB-SCAN algorithm

Find Nearest Neighbor



http://bit.ly/2DW7LyO

Features

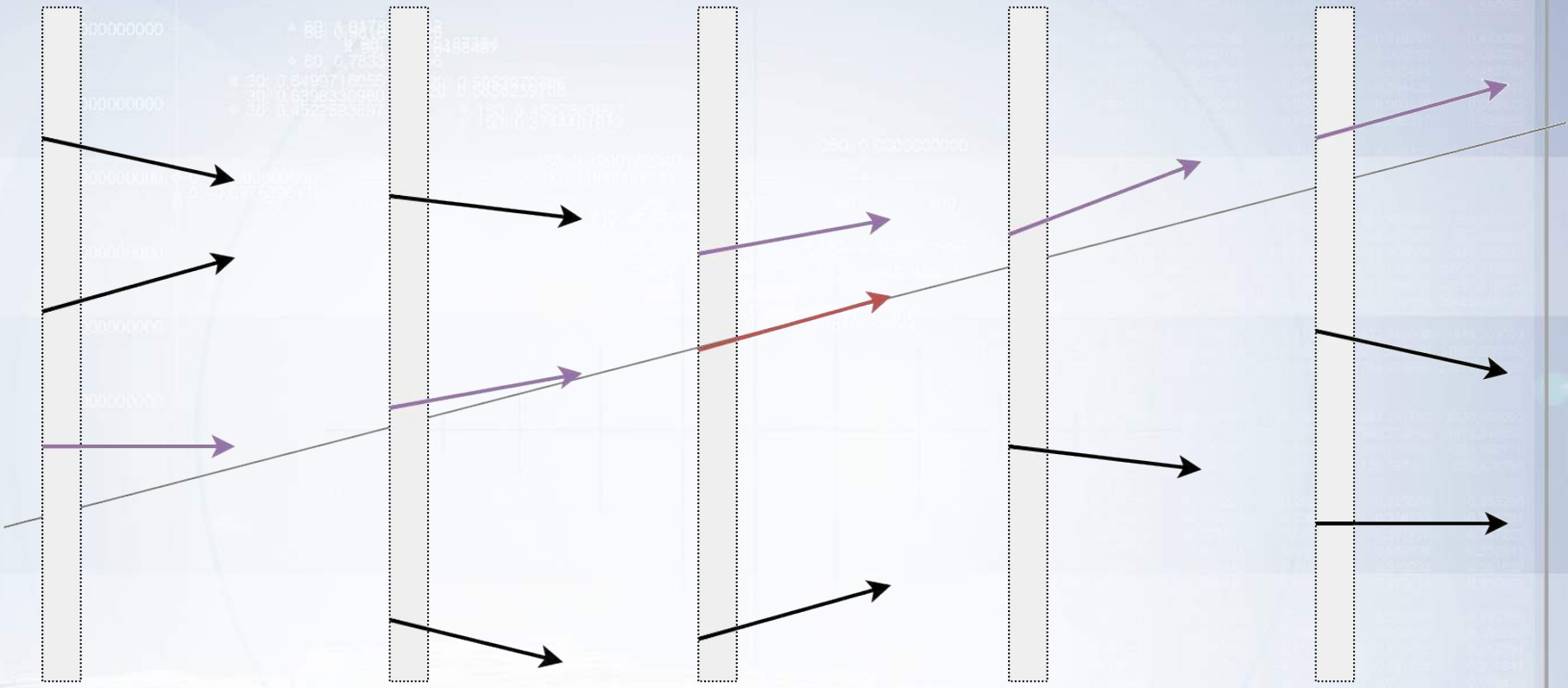


Features:

- angle between directions,
- impact parameters,
- mixed product of two directions and
- vector connecting positions of basetracks,
- some projections.

http://bit.ly/2DW7LyO

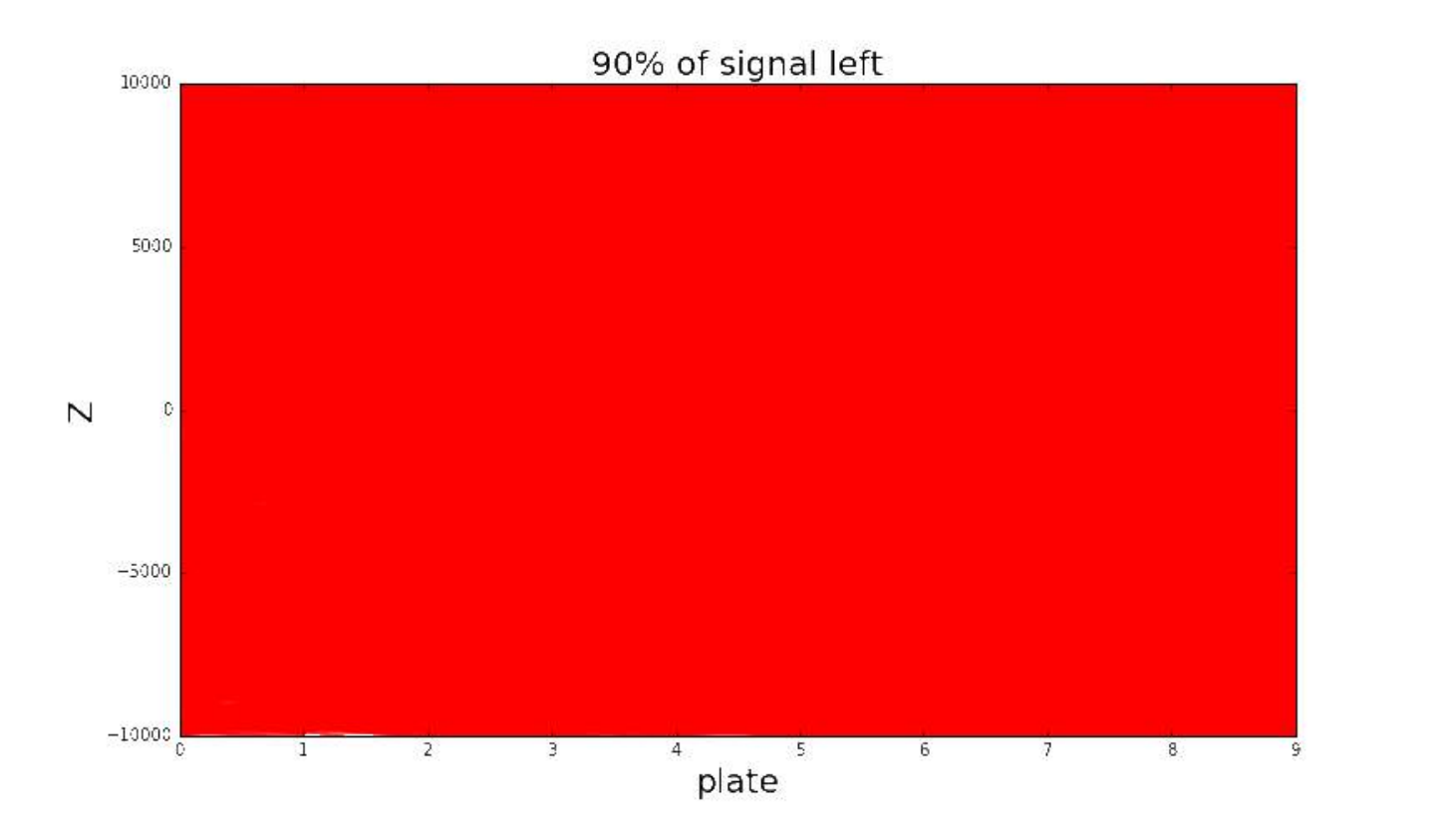
Going deeper



Building classifier model (e.g. Random Forest) over features of 5-tracks gives Precision ~1 at recall 0.5.
No origin information is used!

http://bit.ly/2DW7LyO

Going deeper

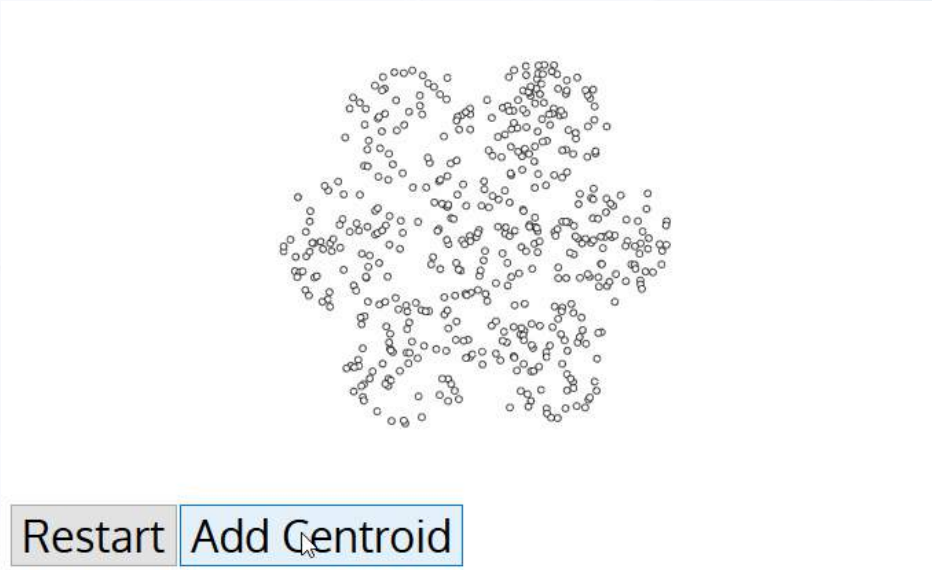


This gives Precision ~1 at Recall 0.5.

http://bit.ly/2DW7LyO

Clustering and Finding Several Showers. K-means

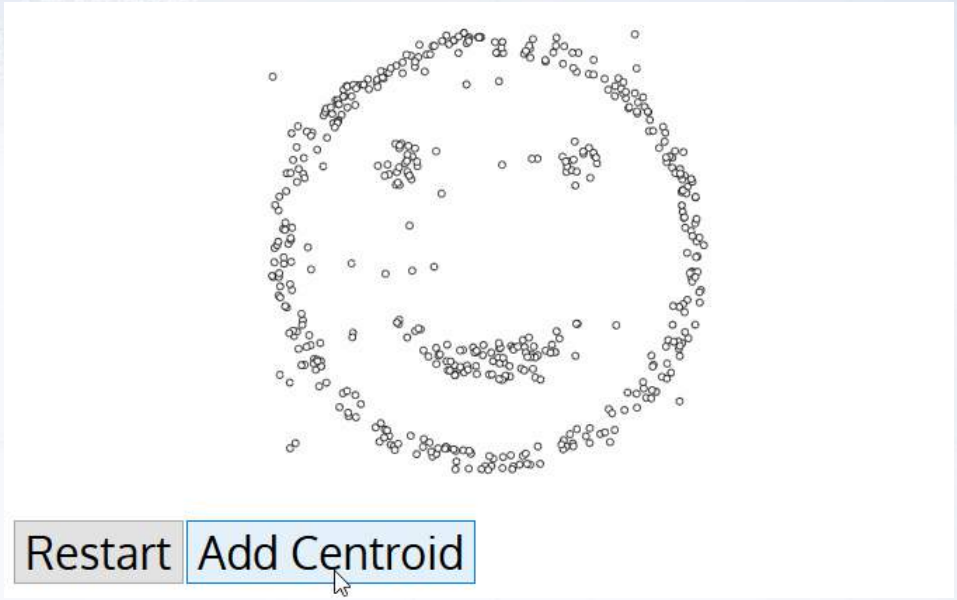
The simpliest approach: it captures the idea that each point in cluster should be near to the center of that cluster.



- Chose number of clusers (k) and iterate:
  - Update centroids
  - Update cluster members

http://bit.ly/2DW7LyO

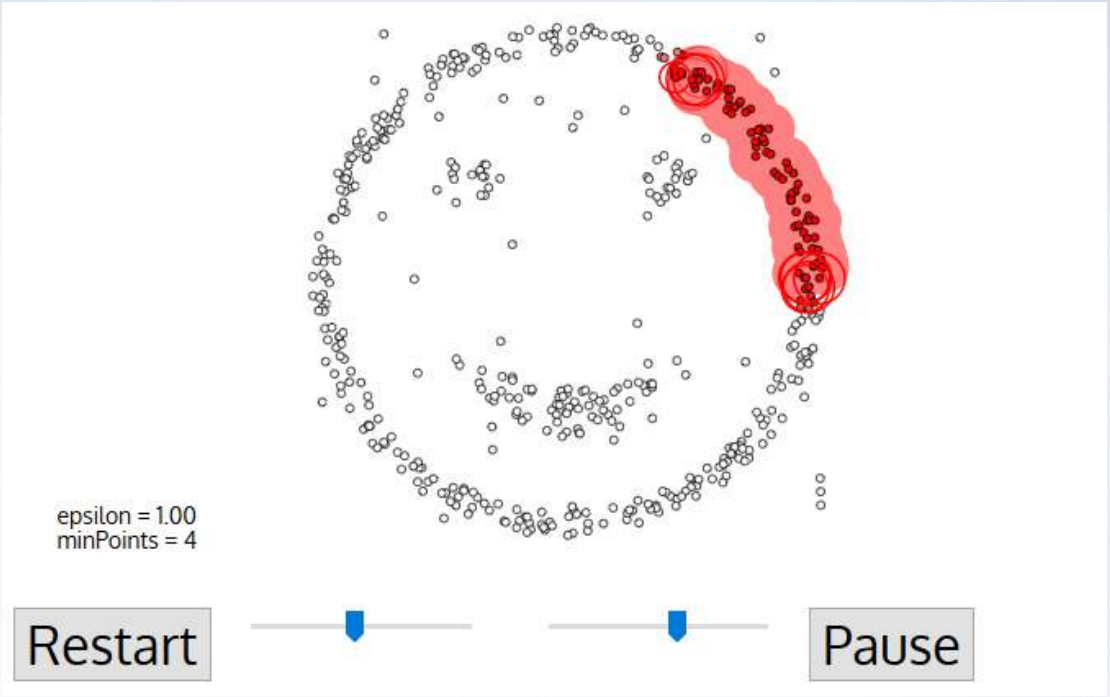
K-means shortcomings



- Works well in case Euclidean metric, but cannot capture more complex dependencies
- Initial choice of K and centroid is annoying

http://bit.ly/2DW7LyO

Density-Based Spatial Clustering of Applications with Noise



- Starts with 2 params: ε (maximum distance to neighbors) and minPoints (to form a cluster);
- Add all points within ε distance to the current cluster recursively;
- Pick a new arbitrary point and repeat the process;
- If a point has fewer than minPoints neighbors (in ε-ball) – drop it;
- Repeat until no points left.

http://bit.ly/2DVgwcz

DBSCAN for OPERA data

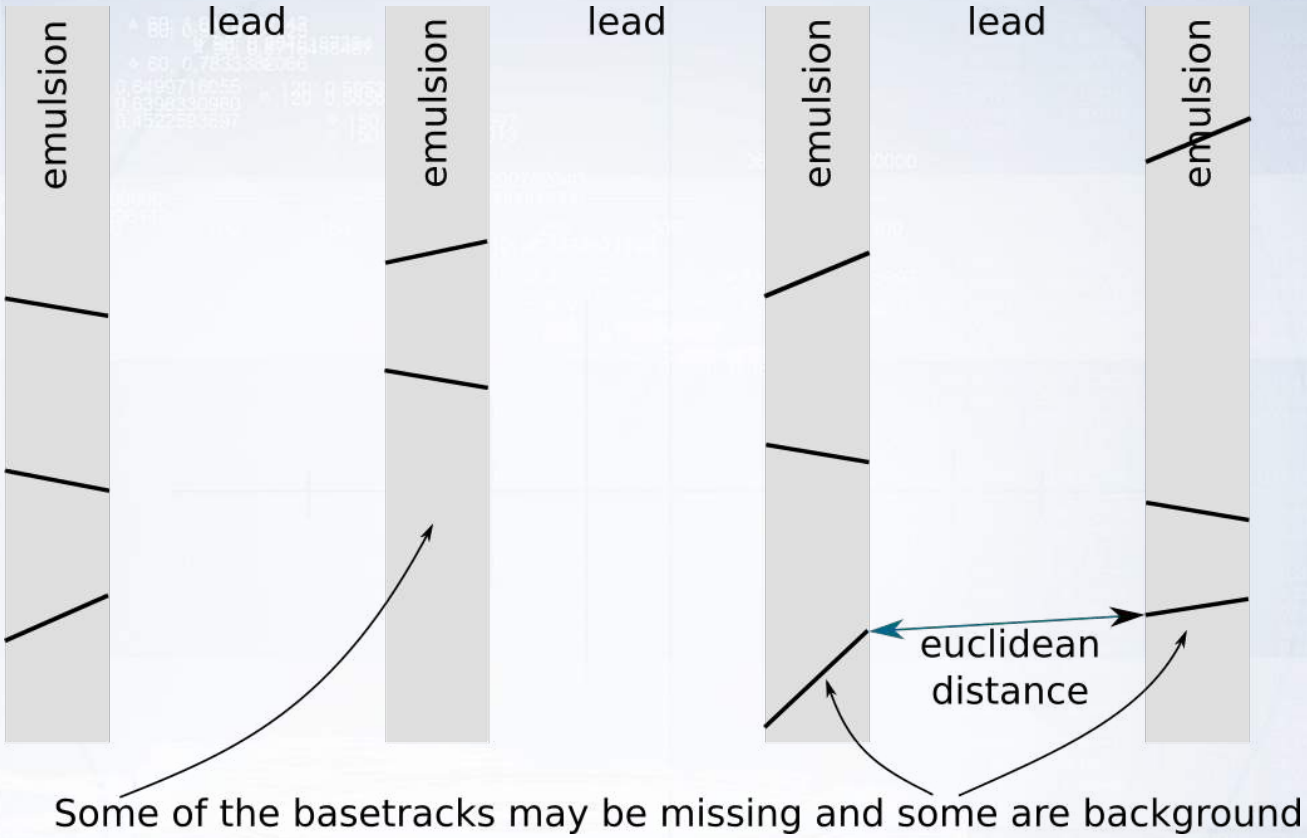
Straighforward application doesn't work very well, because Euclidean (default) metric

$$\rho(\mathrm{track}_1, \mathrm{track}_2)^2 = ||\mathbf{x}_1 - \mathbf{x}_2||^2$$

is not very relevant to basetrack alignment

http://bit.ly/2G9Pujv

Adjusting metric



Improved metric:

$$\rho(\mathrm{track}_1, \mathrm{track}_2)^2 = ||\mathbf{x}_1 - \mathbf{x}_2||^2 + \alpha||e_1 - e_2||^2$$

http://bit.ly/2G9Pujv

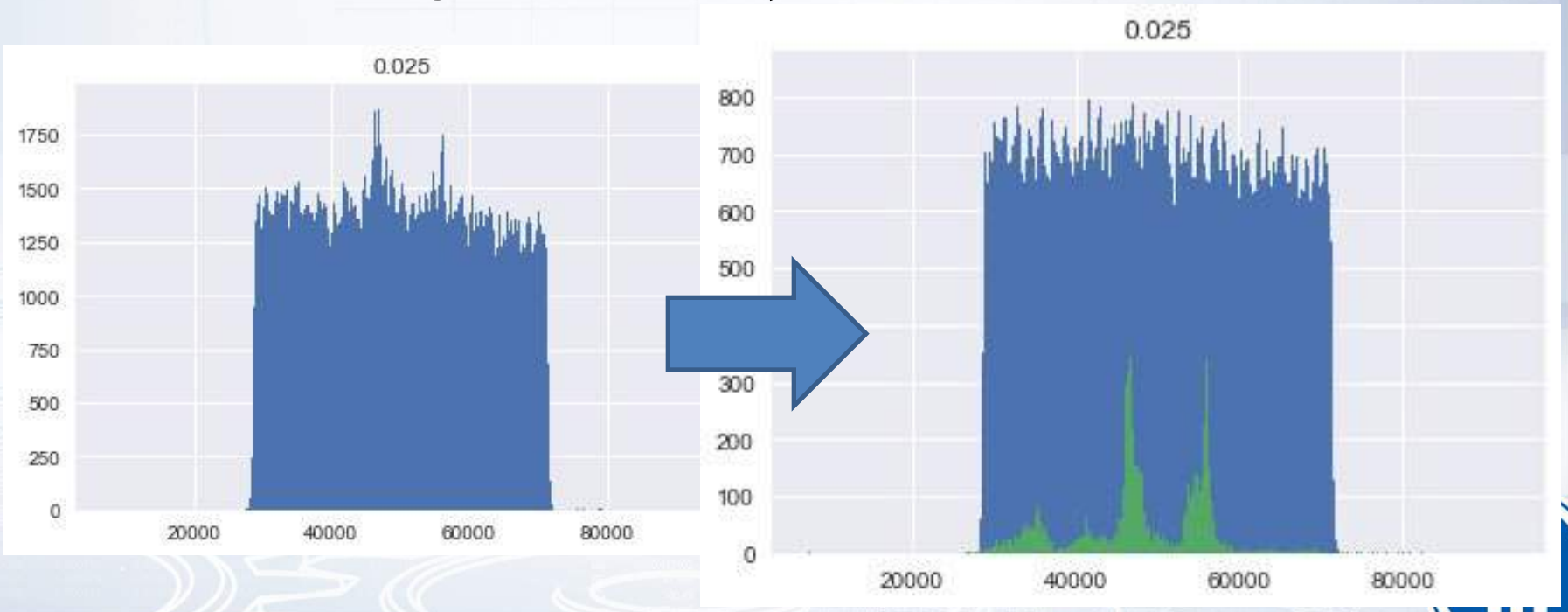
Result

Showers are better visible, although there is room for improvement (at some plates some tracks may be missing and one have to account for direction alignment)

http://bit.ly/2G9Pujv

Possible Improvement Ideas

- Conditional Random Fields
- Recurrent Neural Networks
- Message Passing Neural Networks
- Estimate origin positions by basetrack densities



Looking Ahead

Even More Tough Challenge

- There are O(100) showers in the volume with significant overlapping probability, no origin is known;
- Background is not random, but mixture of other tracks passing through;
- Add timing information.

Conclusion

- Dark Matter is one of the most challenging Physics topic:
  - many questions, many hypothesis, many approaches.
- SHiP – proposed experiment at CERN with rich DM program;
- Emulsion plays important role due to high sensitivity:
  - Electromagnetic shower reconstruction tasks.
- In reality there are many more ML challenges:
  - Emulsion detector design + timing;
  - Optimization of experiment design;
  - Particle identification.

From D.Whiteson, J Cham book "We have no idea"