Yandex SCHOOL OF DATA ANALYSIS

Dark matter search in OPERA and SHIP experiments

V. Belavin, A. Filatov, S. Shirobokov, Andrey Ustyuzhanin
Yandex School of Data Analysis
NRU Higher School of Economics

Why bother?

- What if dark matter is made of some new kind of particle that we are able to produce and harness in high-energy colliders?

- Or in discovering what it is, we figure out something about the laws of physics we didn't know about before, such as a new fundamental interaction or a new way that the existing interactions can work?

- And what if this new discovery lets us manipulate regular matter in new ways?

Intro to experiment and data

- OPERA experiment.
  Was looking for rare neutrino oscillations. To detect them used emulsions detectors – think as huge film photocamera with long exposure.
  Oscillations found ✔

- SHIP experiment looking for unknown particles(light dark matter) and utilize idea of OPERA with emulsion films. It could use emulsion films to capture electrons from LDM scattering on electron or nuclei. However, the origin of decay is not known now. Possibility for discovery ?

The above detector is a “brick” which consists of 57 films. Particles, that fly through the detector left tracks in this films. The track is described with 6 features: 3 spatial coordinates, 2 angles of direction, and χ^2.

OPERA brick (similar to SHiP)

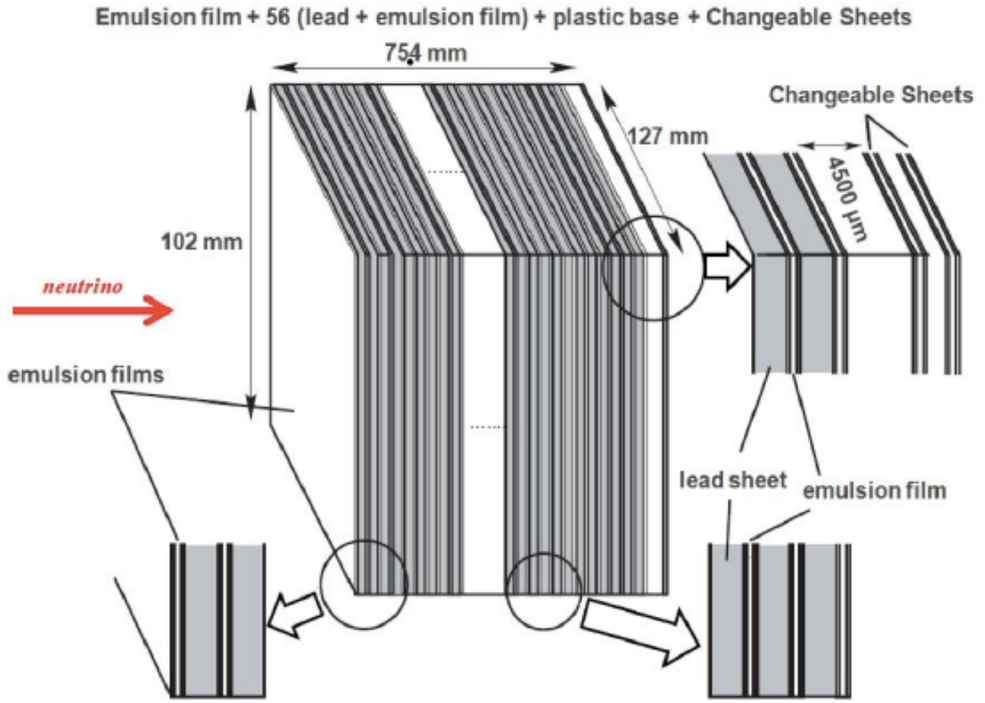
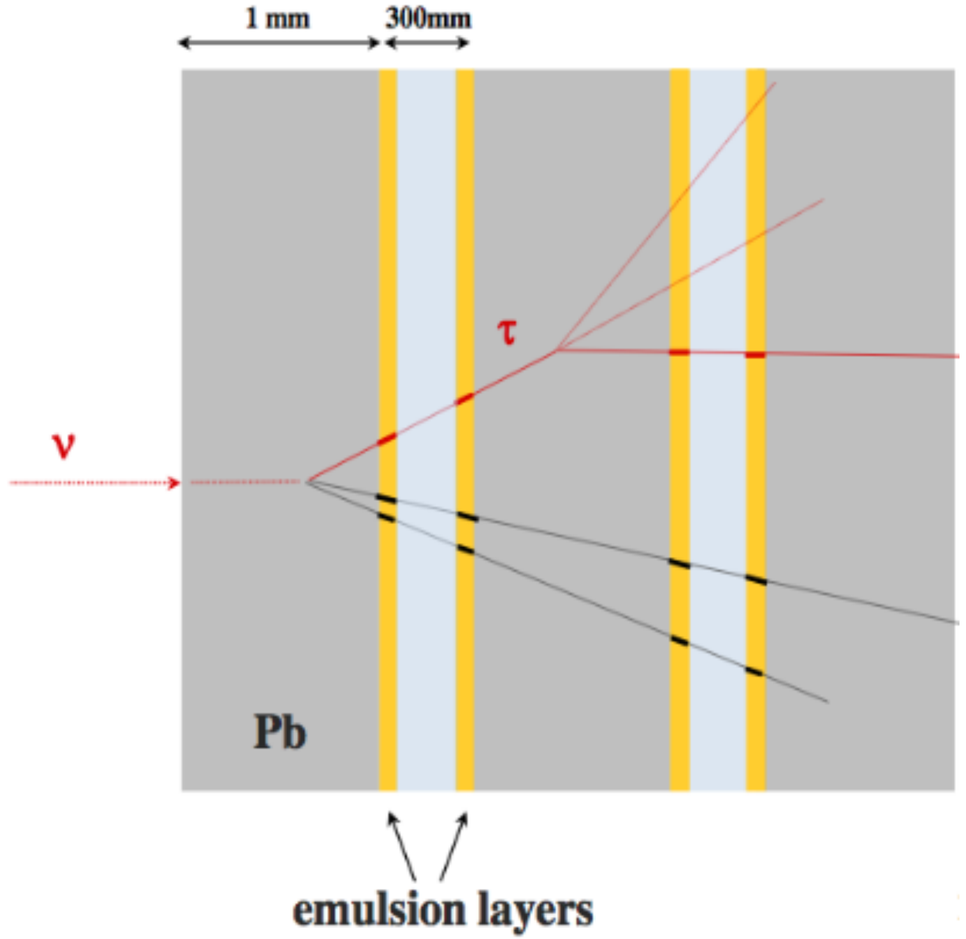
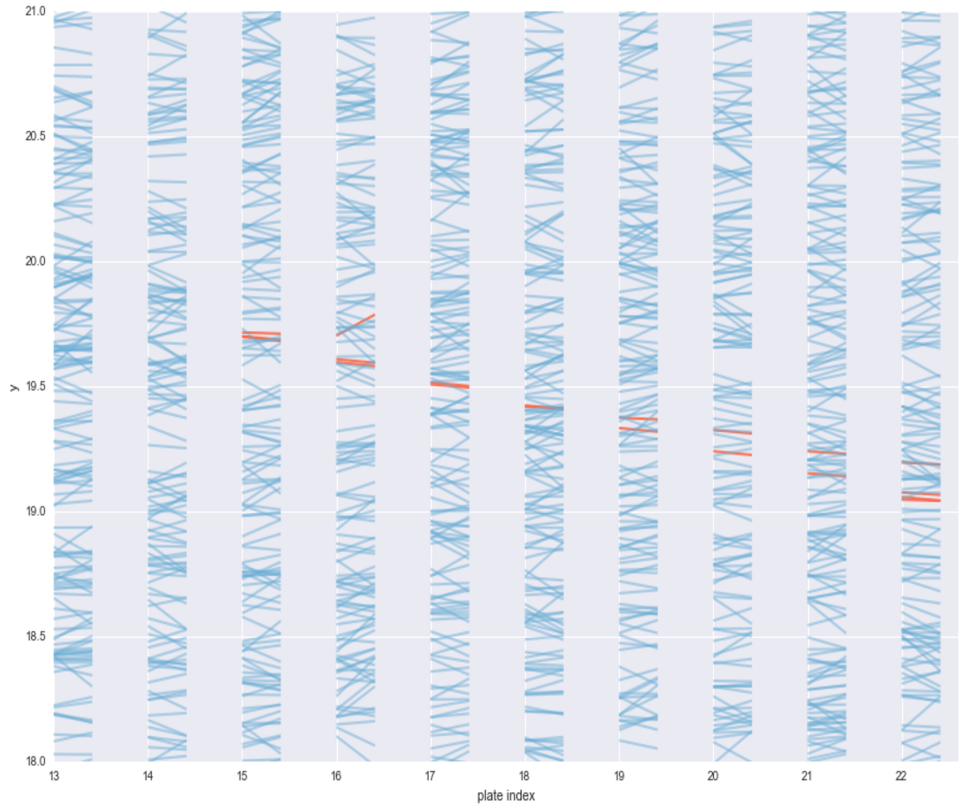
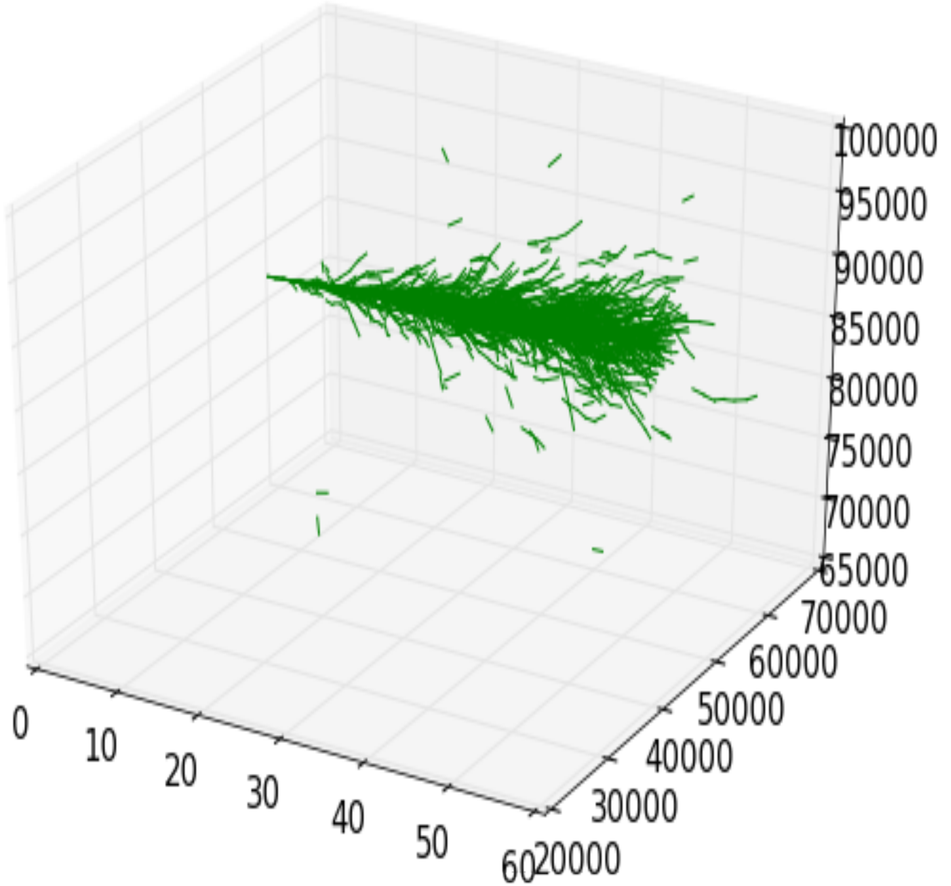


Figure 2.4 – Schematic structure of an ECC brick.



Visualised signal and background





Data samples from experiment

Background consists from tracks randomly scattered around brick. In real brick there should be ~ 2.3 * 10^7 tracks.

Data Background: Real tracks from last 10 layers of different bricks ~ 10^6 tracks.

- One needs to subsample couple of bricks data to obtain correct background ratio.

Signal consists from tracks forming a cone like shape . There are about 300 – 400 tracks in each signal event.

Signal Data: Simulation of pure EM showers (~6000 events) ~ 3 * 10^6 base tracks in total.

Research goal

Track features: X,Y,Z, angle_XZ, angle_YZ, chi2

Given signal(1) and background(0), described by 6 features mentioned above, develop a classification algorithm that can

- detect e-m shower tracks within a brick tracks
- identify shower origin

Kaggle competitions

There were two competitions

1. Background reduced to $\mathbf{10^6}$ tracks, with 100 signal events, with initial point of shower given.

2. There are 100 bricks with $\mathbf{10^5}$ background tracks, and 1- 5 signal events in each brick, no initial point given.

Scoring metric: ROC AUC

- More stable than PR/Recall curve AUC, no need to choose threshold as for F-score . However, even close to 0.9 value does not mean good classifier.

Data preparation

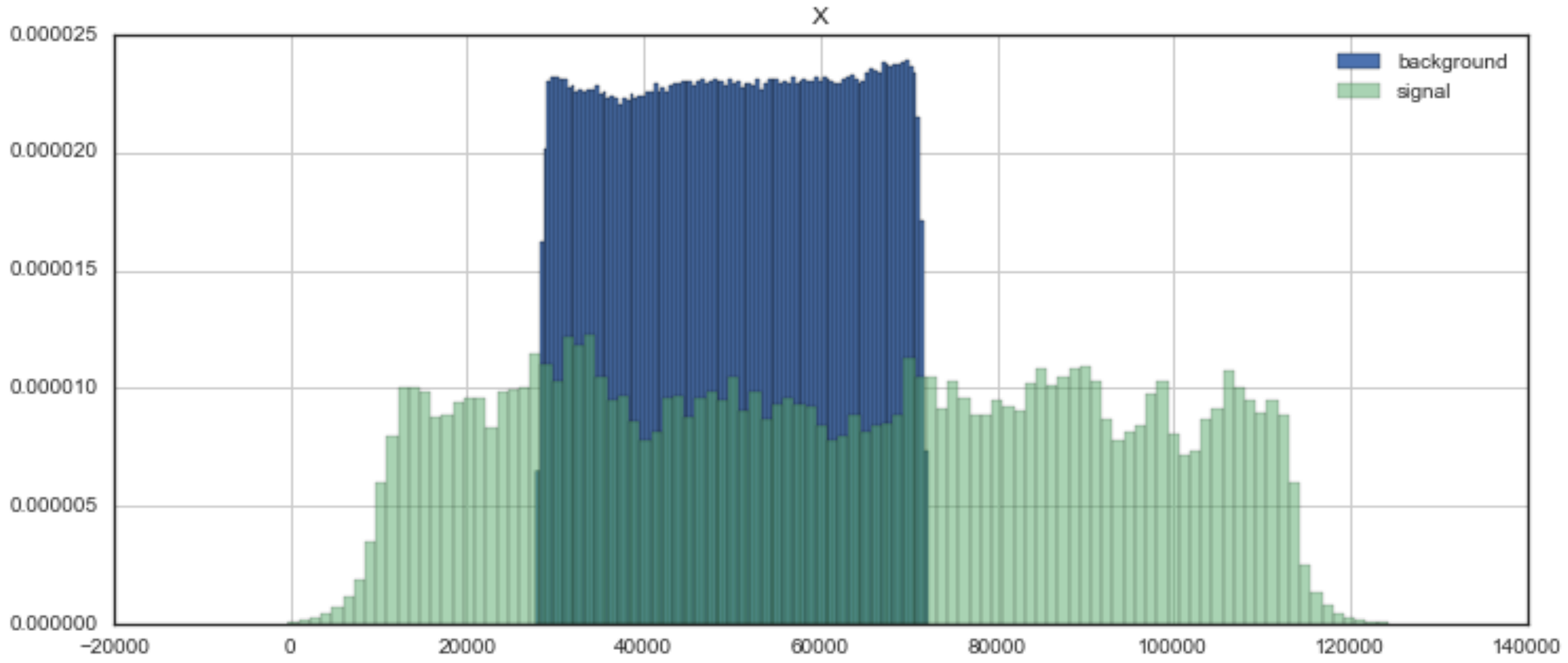
Problems with data from real world: All background is real, whether all signal is simulated.

1. The background is obtained by physically scanning films in bricks. Different bricks scanned with different settings. This results in different coordinates scanned as well as different distributions of tracks.

2. One more not so obvious problem.

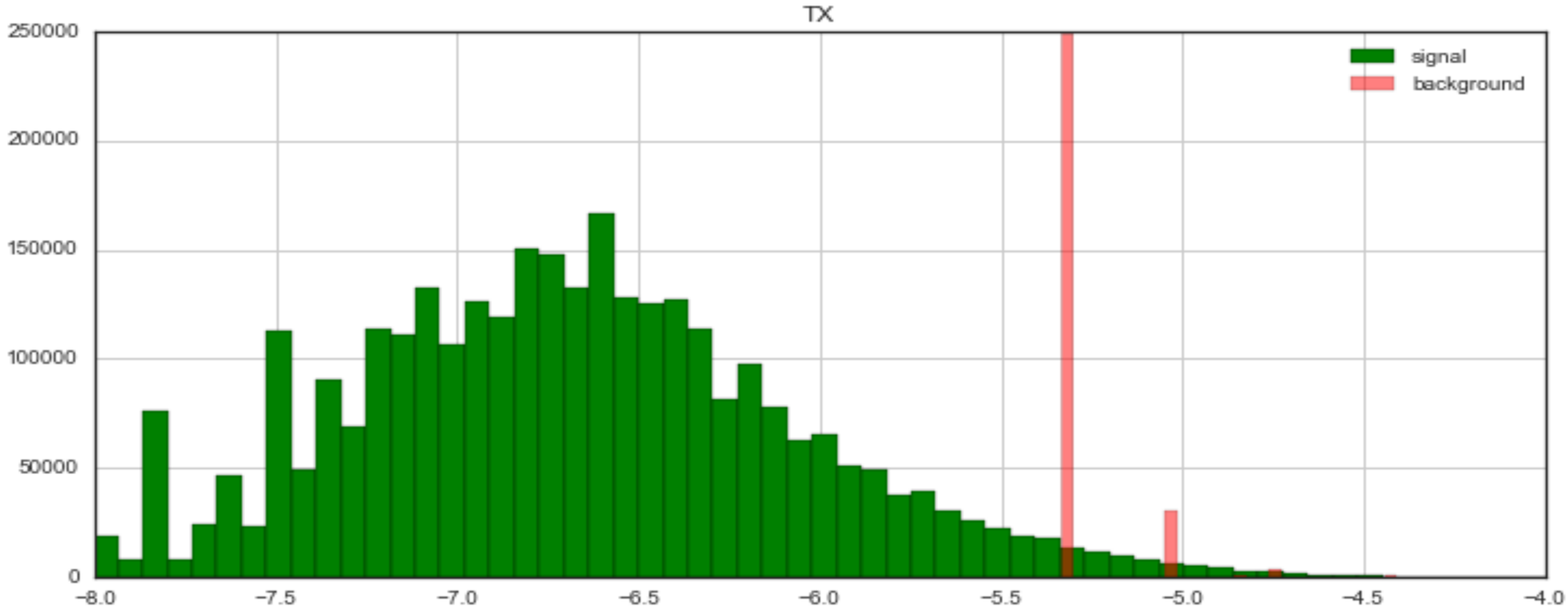
Data preparation

Distribution of X coordinate in real data



Data preparation

Could you guess, what is the problem with data here?



Data preparation

1. Have to squeeze, shift data to fit in brick. Afterwards, have to shift signal to fit in brick. You can not simply squeeze it - the cone structure will be destroyed. Make sure shifted signal does not stands out of brick.

2. Have to binirize signal features manually, otherwise, competition cheater can detect discrepancy in the data, build ”microscope” grid and find all background events.

Our solutions: generic approach

1. Describe every track by set of constructed features
2. Train classifier X to discriminate signal track from background track
3. Apply a cut on the classifier output
4. Topology filter Y to identify shower among selected tracks
5. Estimate quality.

Solution 1

Features:

1. Impact parameter
2. Euclidean distance between tracks
3. Angle
4. Angle difference
5. Chi2

Solution 1:

1. SVM Stage
   On given features, use SVM on pair of tracks to find probability that the next track is continuation of the previous track. If there is no tracks with probability greater than threshold – delete track. Apply this stage N times. Results in chains of selected tracks.

2. CRF Stage
   Build a graphical model over sequence of edges between tracks and train it on signal sequences and pseudo background sequences. The output is the classification of chain as signal or background.

Result: 0.85 precision at 0.92 recall for 100 signal events per brick

Solution 2

Features:

1. Impact parameter projection
2. Angle
3. Impact parameter to angle
4. Chi2

Solution 2:

1. XGBoost Stage
   On given features, apply XGBoost to classify tracks(not pairs of tracks, as in solution 1).

2. Topology filter. Utilize the fact of longitudial structure of the shower.

3. Origin search.

Result: 0.7 precision at recall 0.75 for 1 signal event per brick

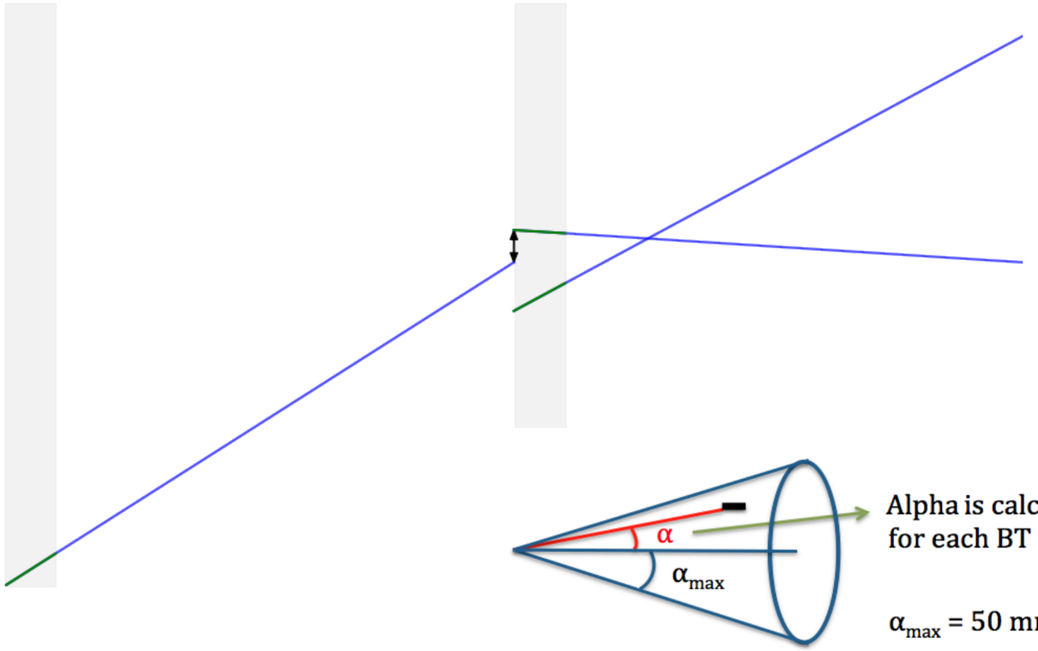
Questions?

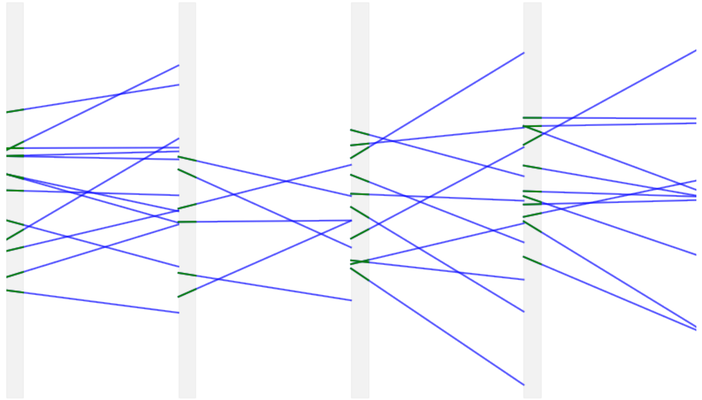
Backup

Basetrack features examples

For every track select a cone 50mrad, and compute

- \Delta - distance between tracks
- \Alpha (see figure on the right)
- \Theta (angle between basetracks)
- SX, SY (slope difference)
- IP– Impact Parameter
- \Chi2
- …
- Use distance/angles computed from/to plate-after-the-next





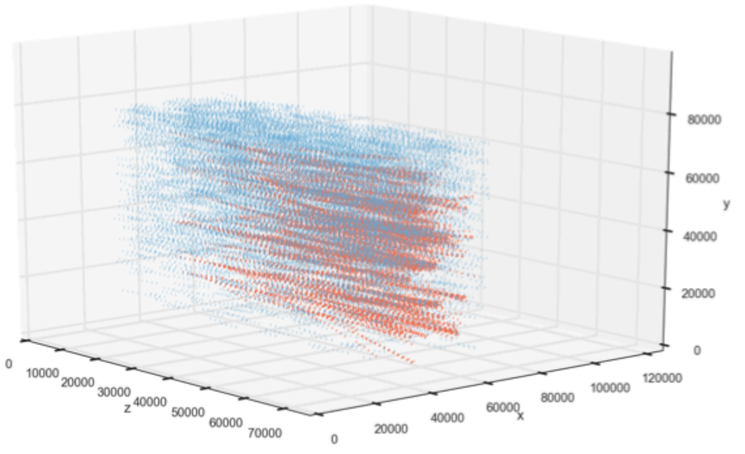


Figure 5: Brick after SVM step

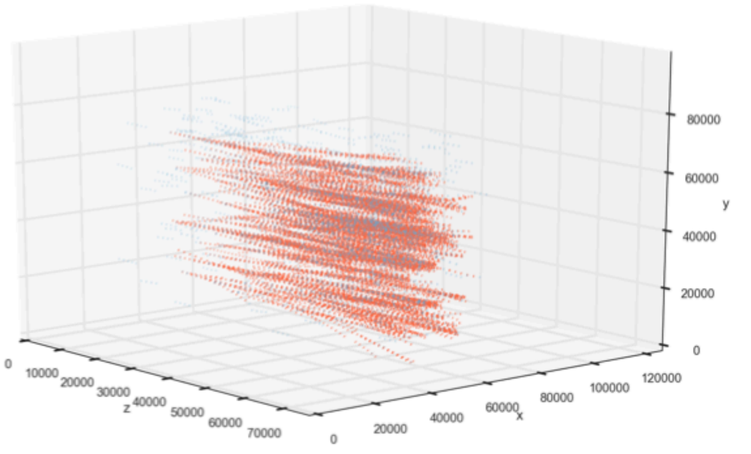
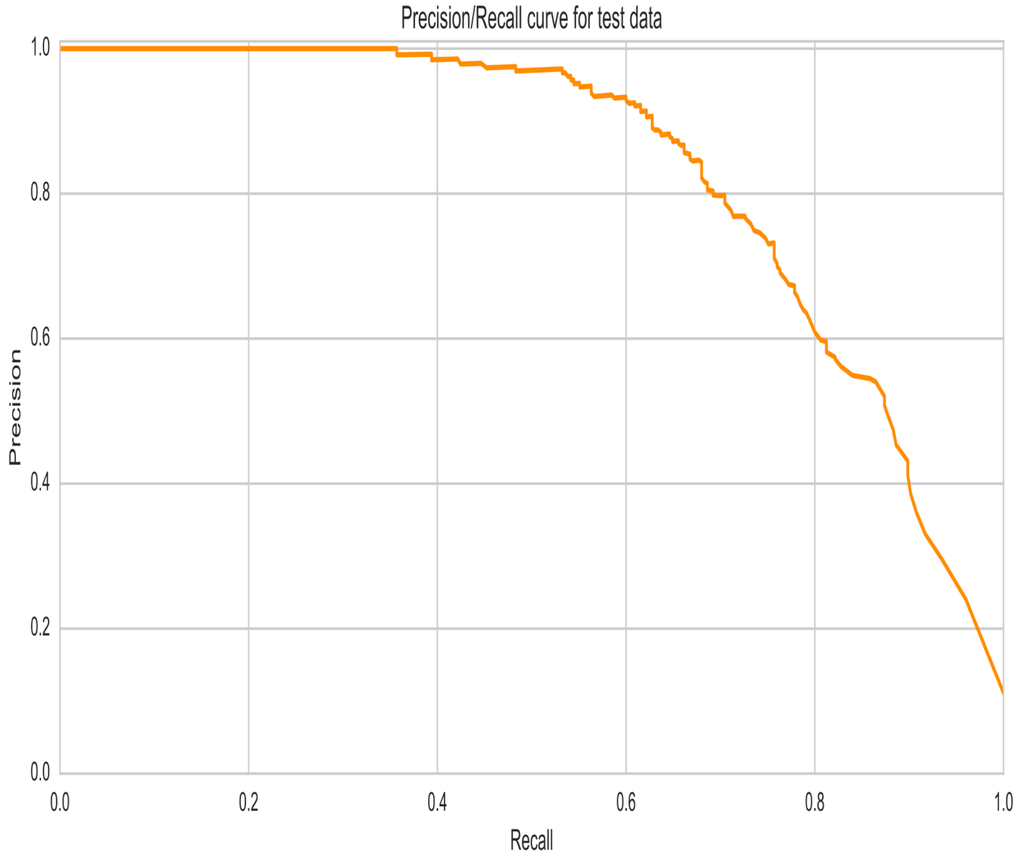


Figure 6: Brick after CRF step

Algorithm 1: SVM step

```
for every layer in the brick do
    for every track in a layer do
        Find neighbors on the next layer;
        Find probability of each neighbor to be a continuation of the track using SVM;
        if highest probability is bigger than threshold or track has more than h ancestors
         then
            Leave track;
        else
            Delete track;
```


Precision/Recall curve for test data
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

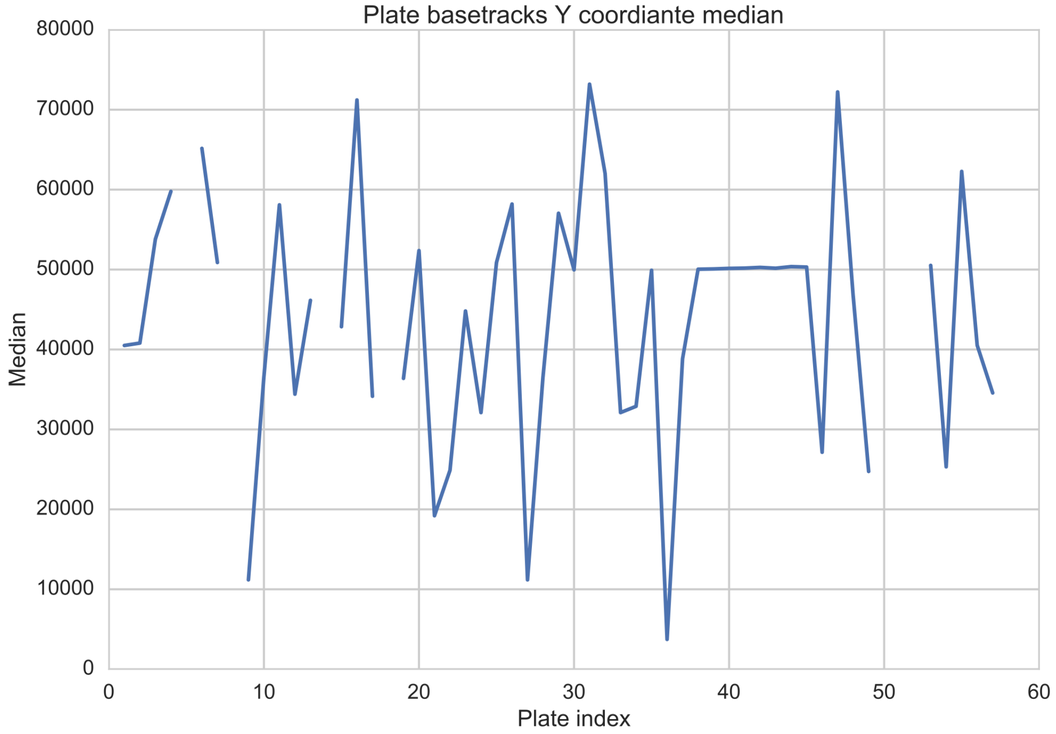
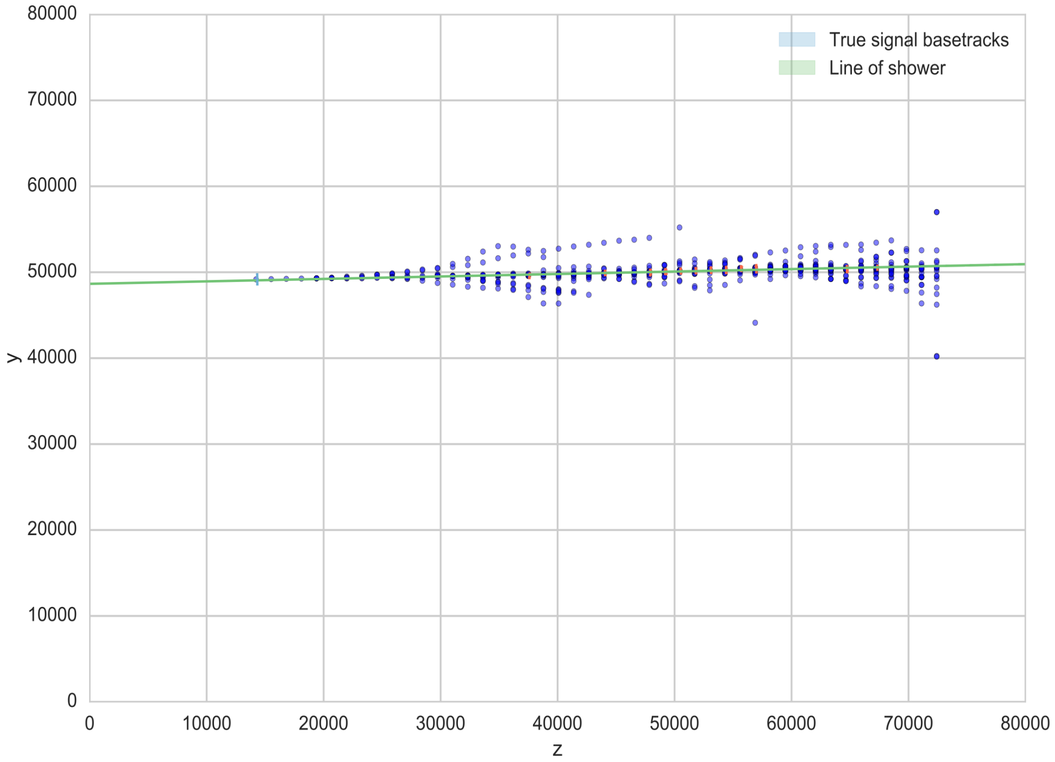
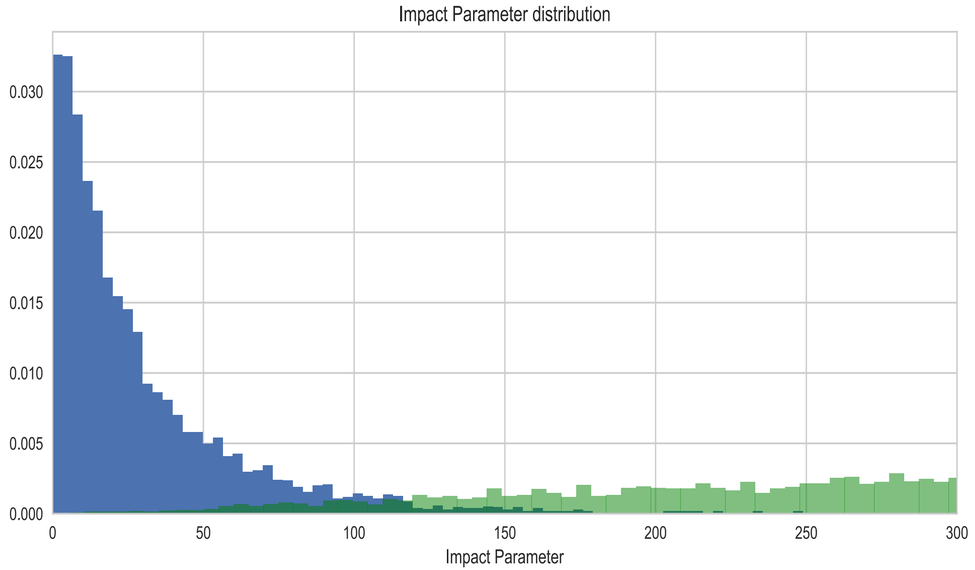


Plate basetracks Y coordiante median
Median
Plate index
80000
70000
60000
50000
40000
30000
20000
10000
0
0
10
20
30
40
50
60


True signal basetracks
Line of shower
80000
70000
60000
50000
40000
30000
20000
10000
0
y
0
10000
20000
30000
40000
50000
60000
70000
80000
z


Impact Parameter distribution
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
Impact Parameter