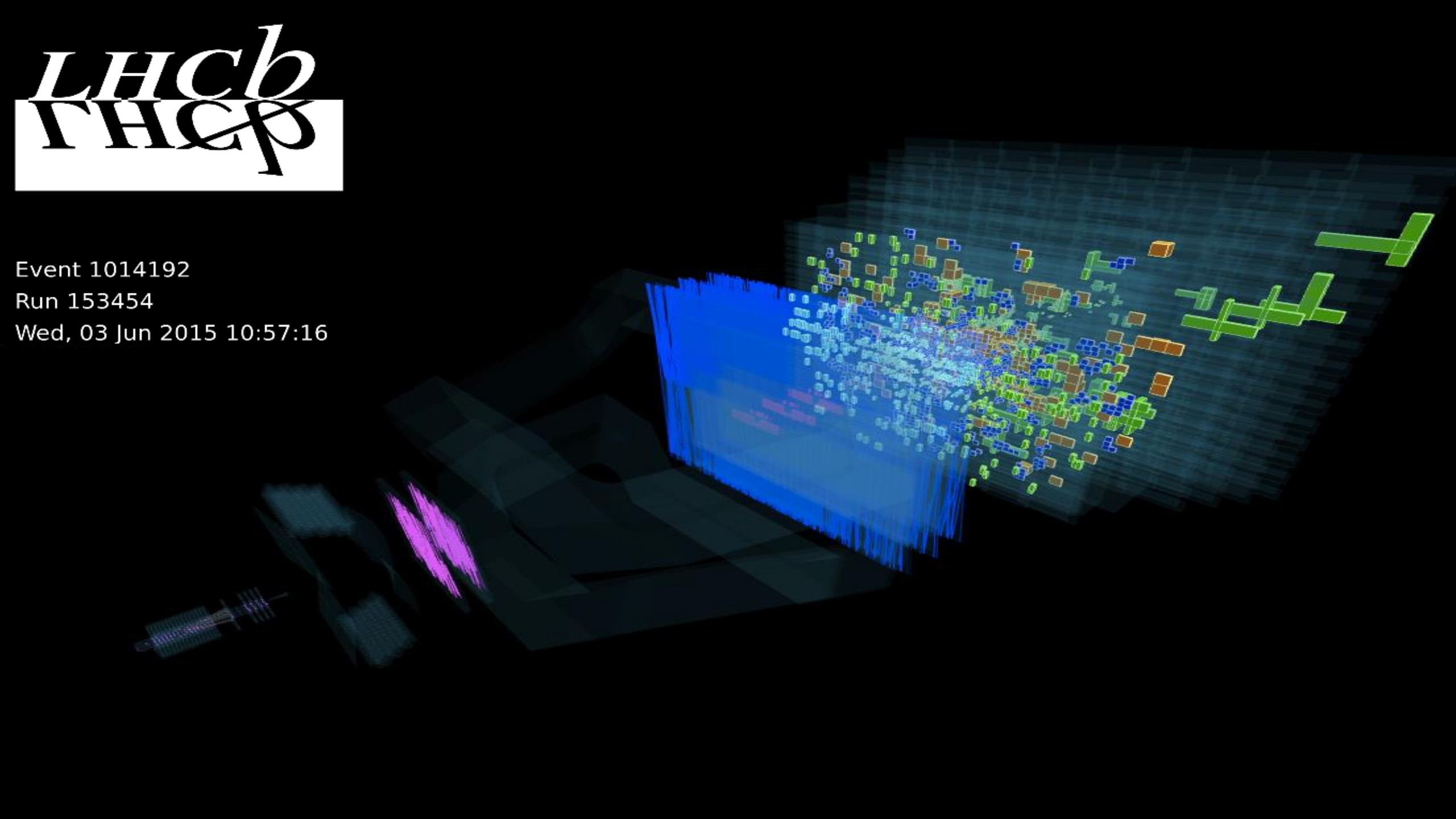
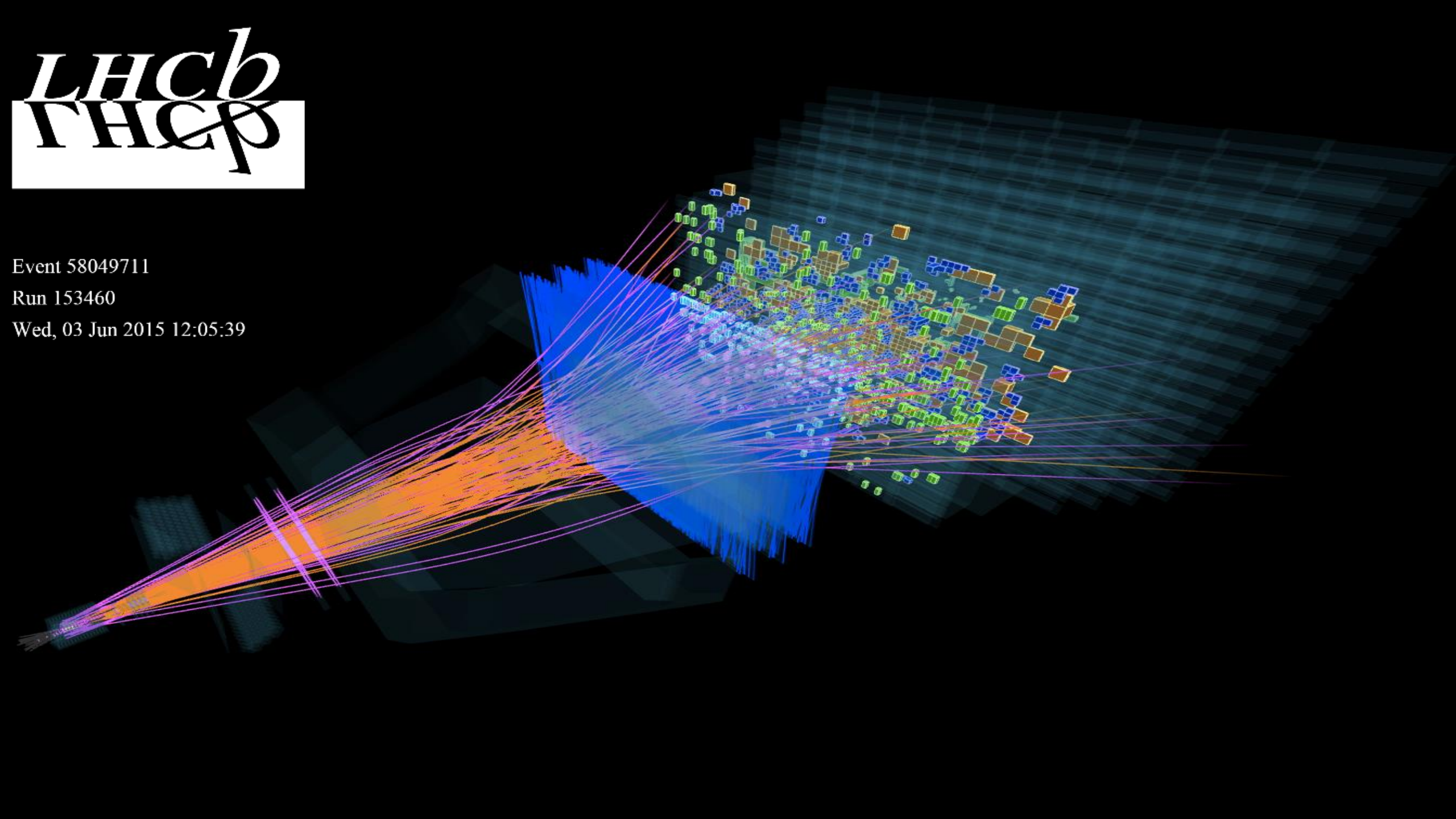
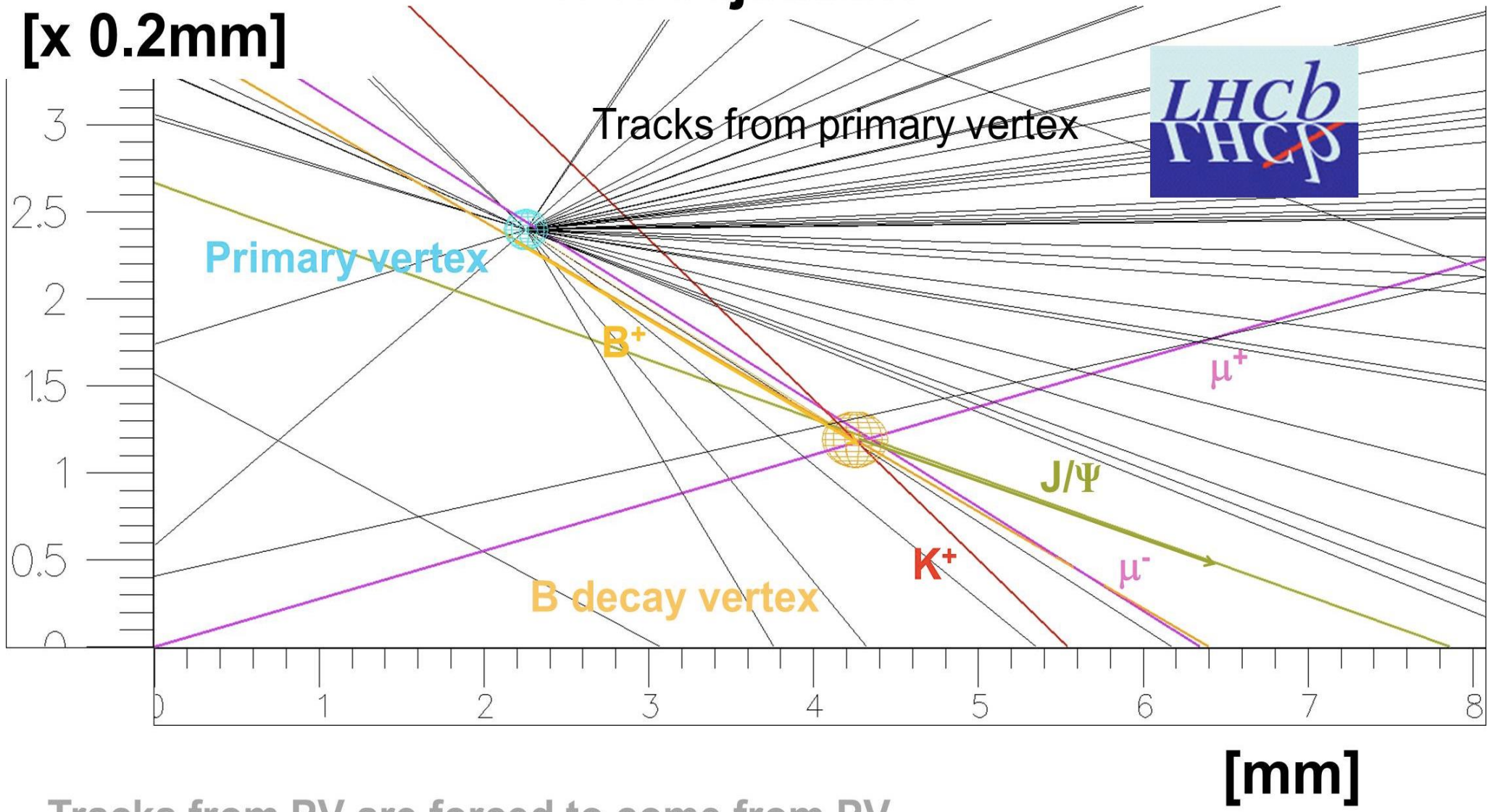
Увеличение эффективности алгоритмов реального времени отбора событий LHCb

Игошин Антон


LHCb
Event 1014192
Run 153454
Wed, 03 Jun 2015 10:57:16


LHCb
Event 58049711
Run 153460
Wed, 03 Jun 2015 12:05:39

YZ Projection



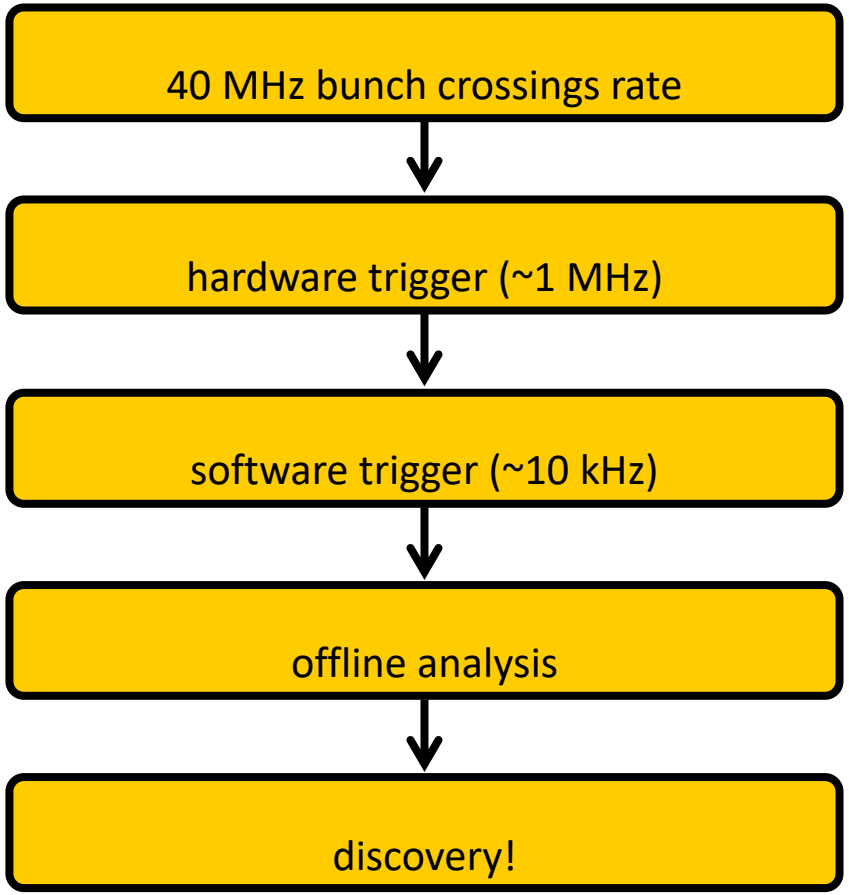
Tracks from PV are forced to come from PV

- Trigger system
- Data structure
- Quality estimation
- Approaches
- Real-time

Trigger system

Две составляющие отбора событий:

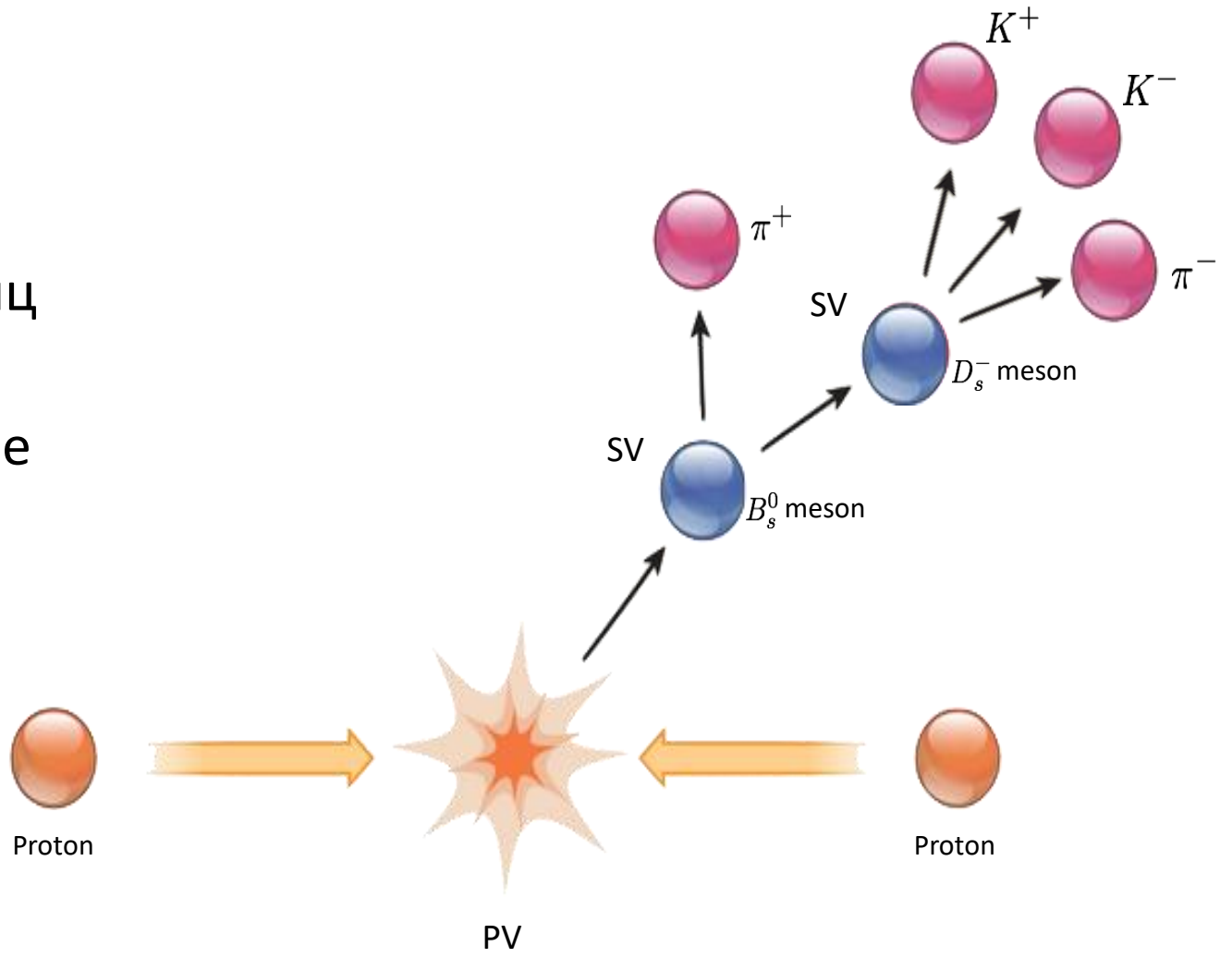
- Аппаратная часть
- Программная часть



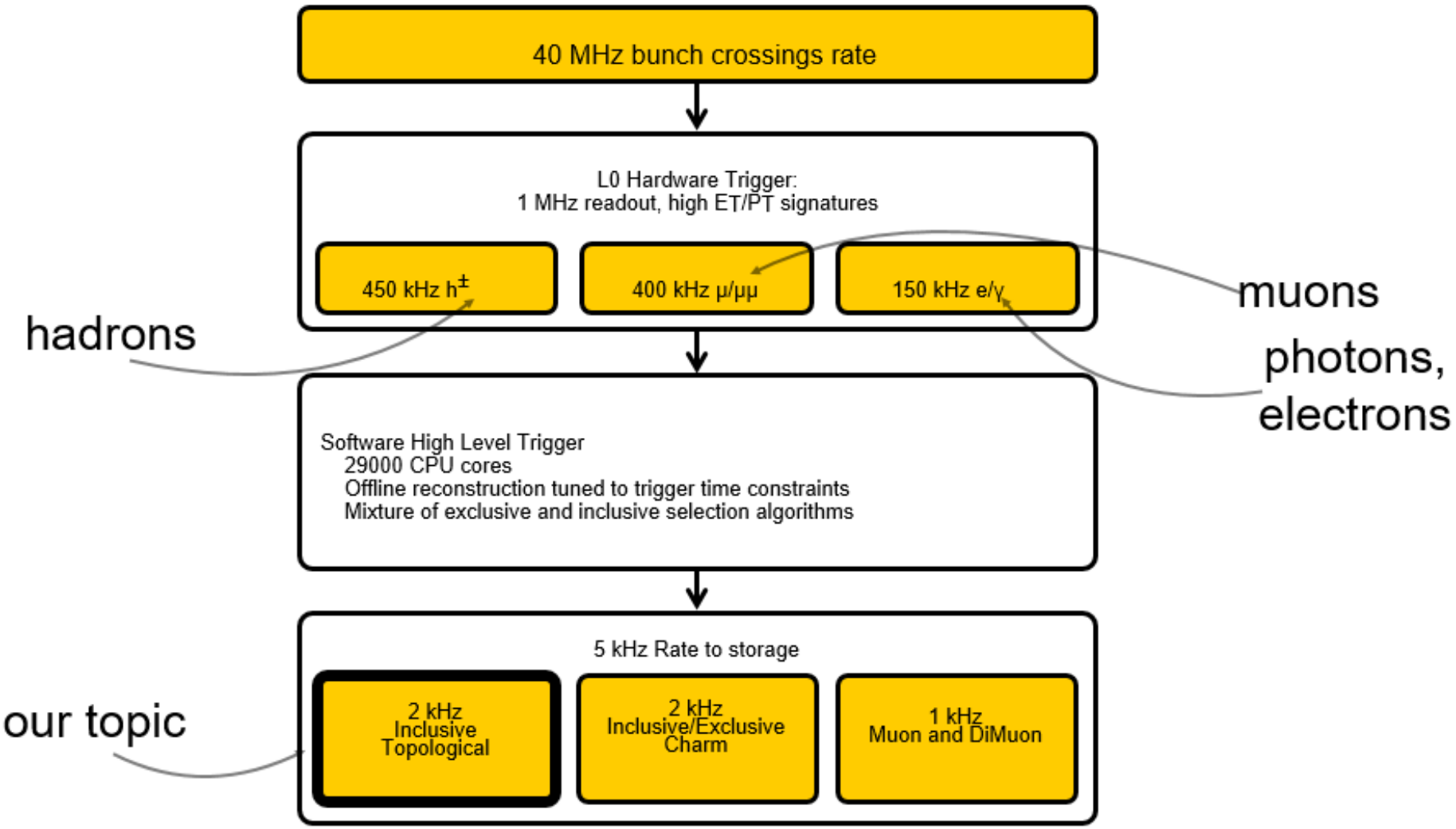
- Программная часть (High Level Trigger, HLT):
  - HLT выполняет реконструкцию событий и записывает данные для немедленного использования
  - HLT гарантирует высокой уровень распознавания искомых сигналов
  - Порядок количества событий падает с нескольких сотен тысяч до нескольких тысяч или даже сотен
  - Машинное обучение используется в топологическом триггере

Интересное событие

- Primary vertex (PV) - точка столкновения протонов
- Secondary Vertex (SV) - точка распада нестабильных частиц
- SV называется интересной, если распадается на искомые частицы
- Событие называется интересным, если содержит хотя бы одну интересную вторичную вершину (SV)



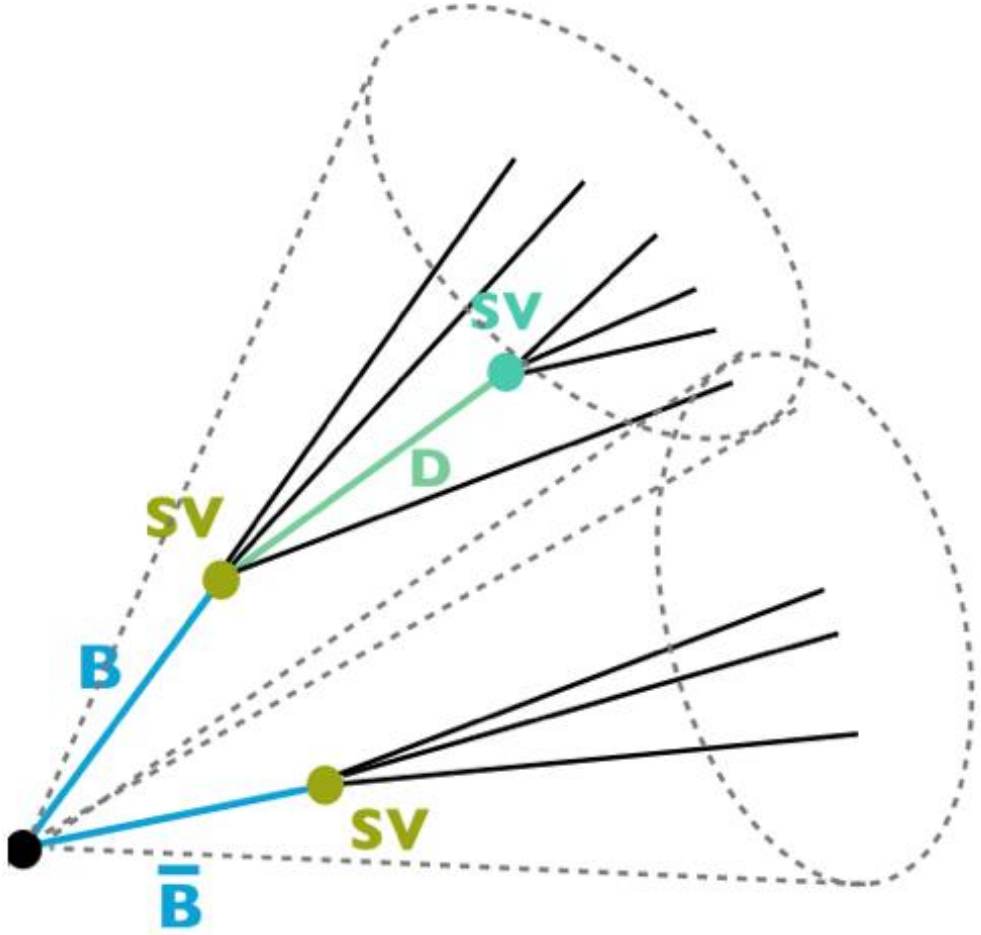
LHCb trigger



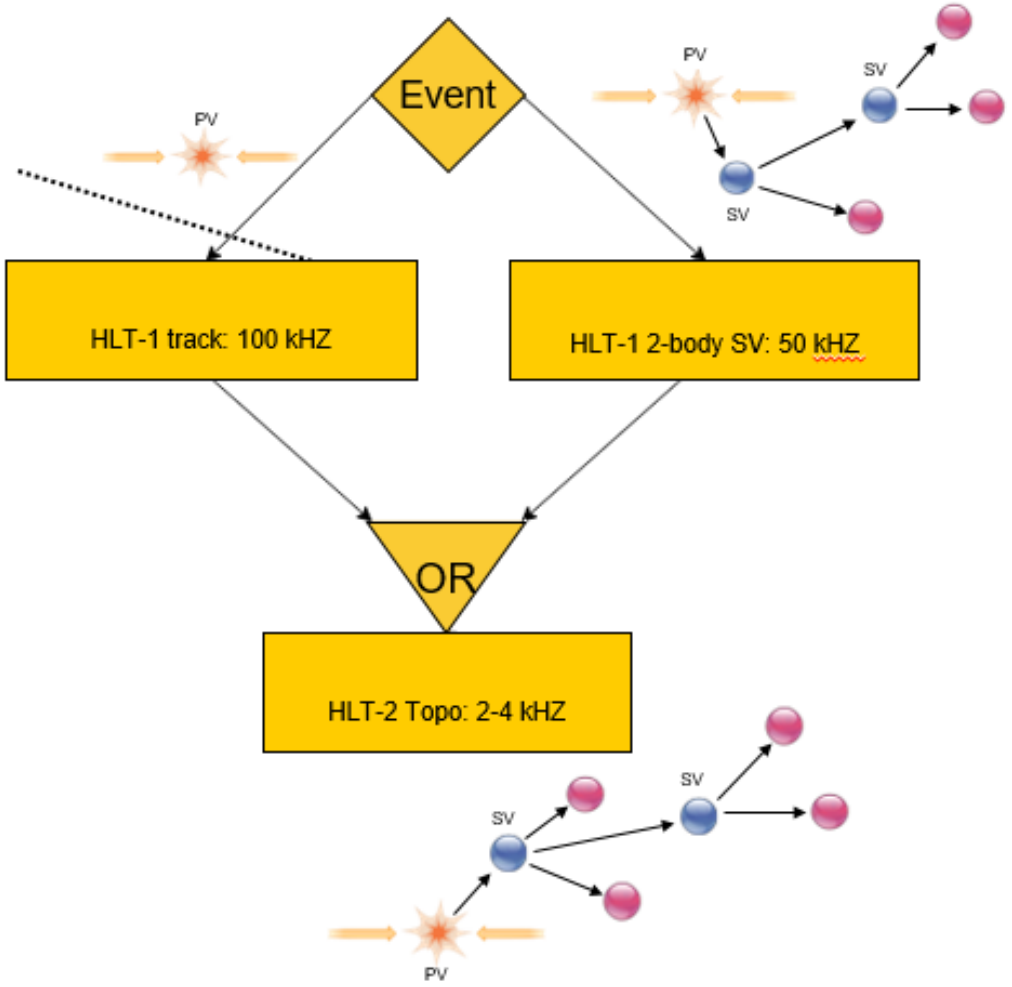
Muons, hadrons, photons are particle species

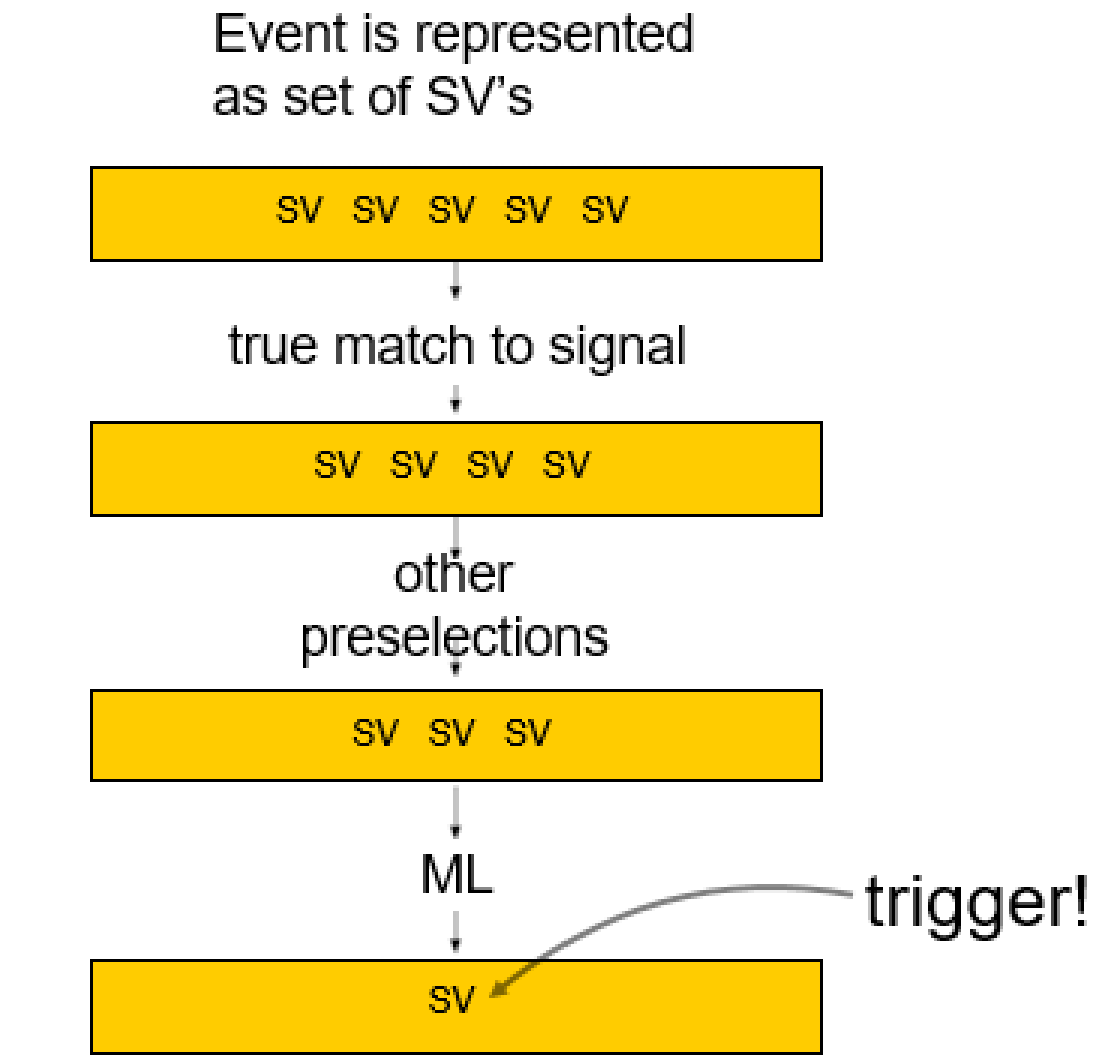
Trigger system в терминах машинного обучения

- События (столкновения протонов)
- Задача бинарной классификации
- События состоят из:
  - Описания треков
  - Описания вторичных вершин



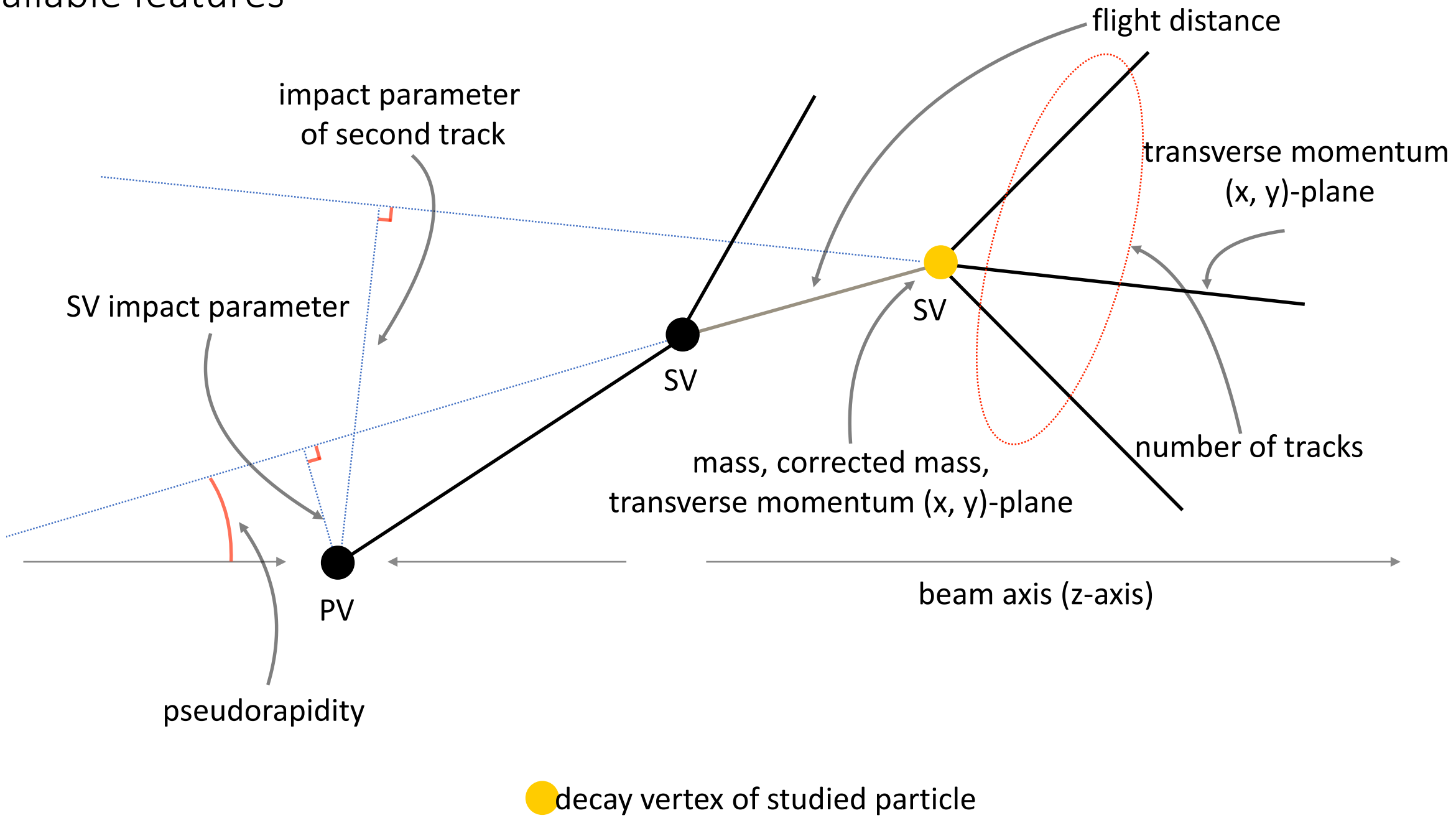
- HLT-1 track ищет интересные треки (по отдельности)
- HLT-1 2-body SV classifier ищет интересные вершины путем сведения 2 треков во вторичную вершину
- HLT-2 improved topological classifier проводит полную реконструкцию, отбирает интересные вторичные вершины





If at least one SV in the event passed all stages, the whole event passes trigger

Available features

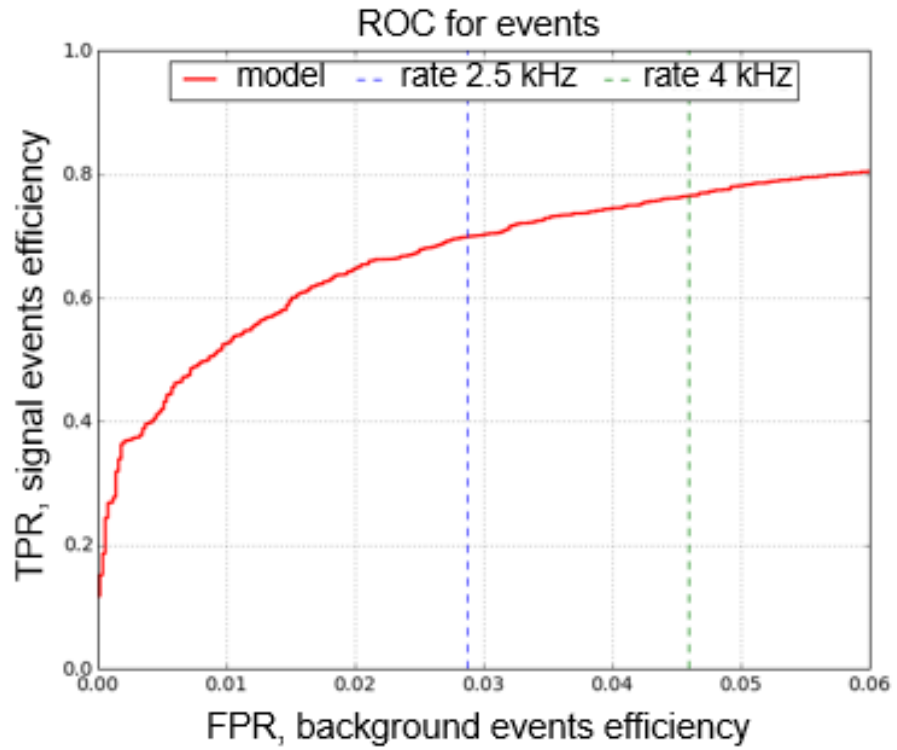


Machine learning problem

- "Signal":
  - Monte Carlo sample is simulated for various types of interesting events (different decays)
- "Background":
  - generic proton-proton collisions are simulated during a small period of time
- Imposed restriction:
  - output rate is fixed (2.5 kHz), thus, false positive rate (FPR) for events is fixed
- Goal:
  - get the highest efficiency for each type of signal events at given FPR

ROC curve, computed for events

- Output rate = false positive rate (FPR) for events
- Optimize true positive rate (TPR) for fixed FPR for events
- Weight signal events in such a way that decays have the same sum of weights
- Optimize ROC curve in a region with small FPR



ROC curve interpretation

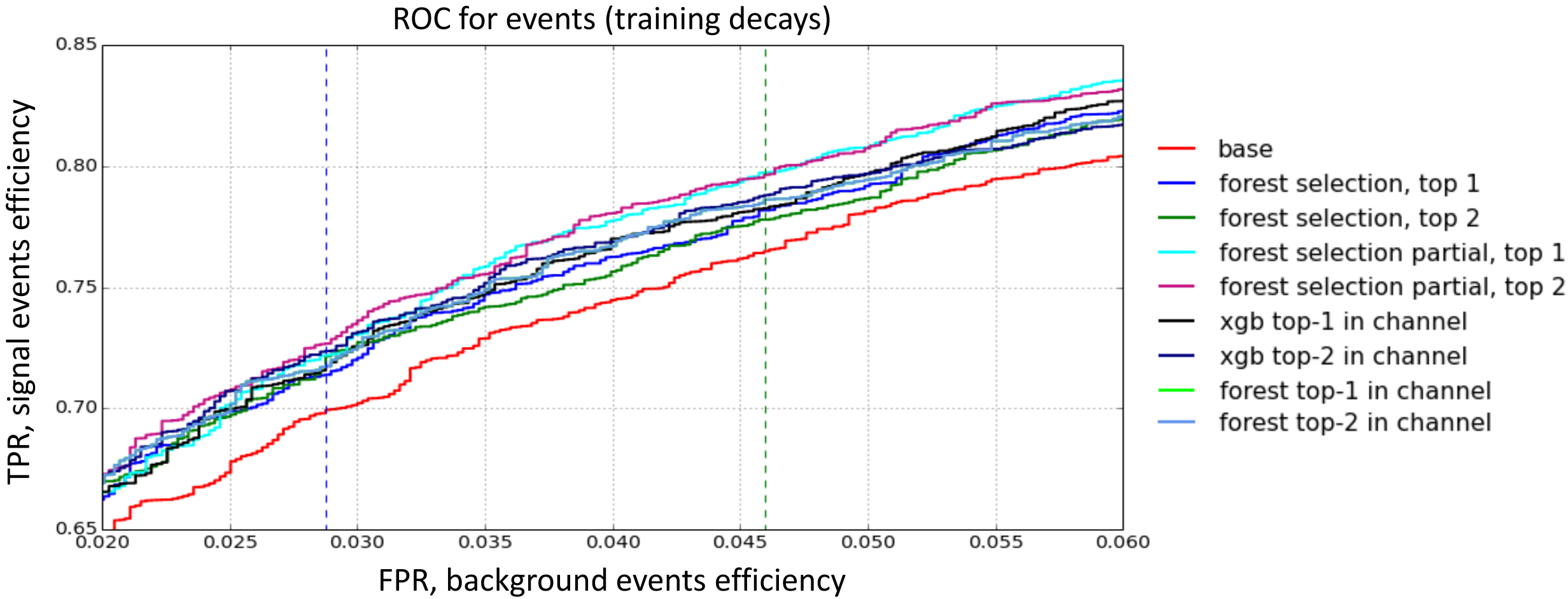
Approaches

- initial selection of well-defined SVs
- max/mean to recompute event probability from SV probability
- different models, classes balancing

➡ Событие называется интересным, если ХОТЯ БЫ одна вторичная вершина является интересной

Random forest for SVs selection

- Train random forest (RF) on SVs
- Select top-1, top-2 SVs by RF predictions for each signal event
- Classifier training on selected SVs provides an increase in TPR of 2-3%

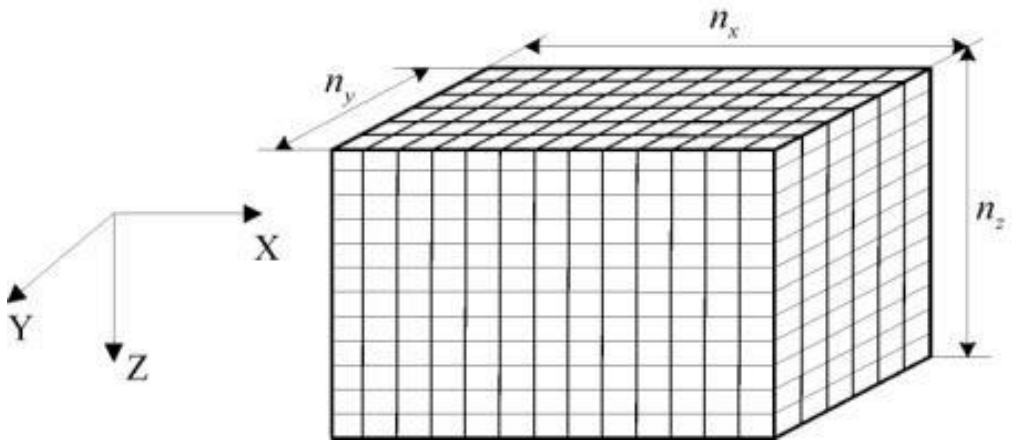


Online processing

- Final model – Yandex technology (MatrixNet) includes several thousand trees
- Prediction operation should take several microseconds
- To speed up prediction operation two possibilities are tested:
  - Bonsai boosted decision tree format (BBDT)
  - Post-pruning

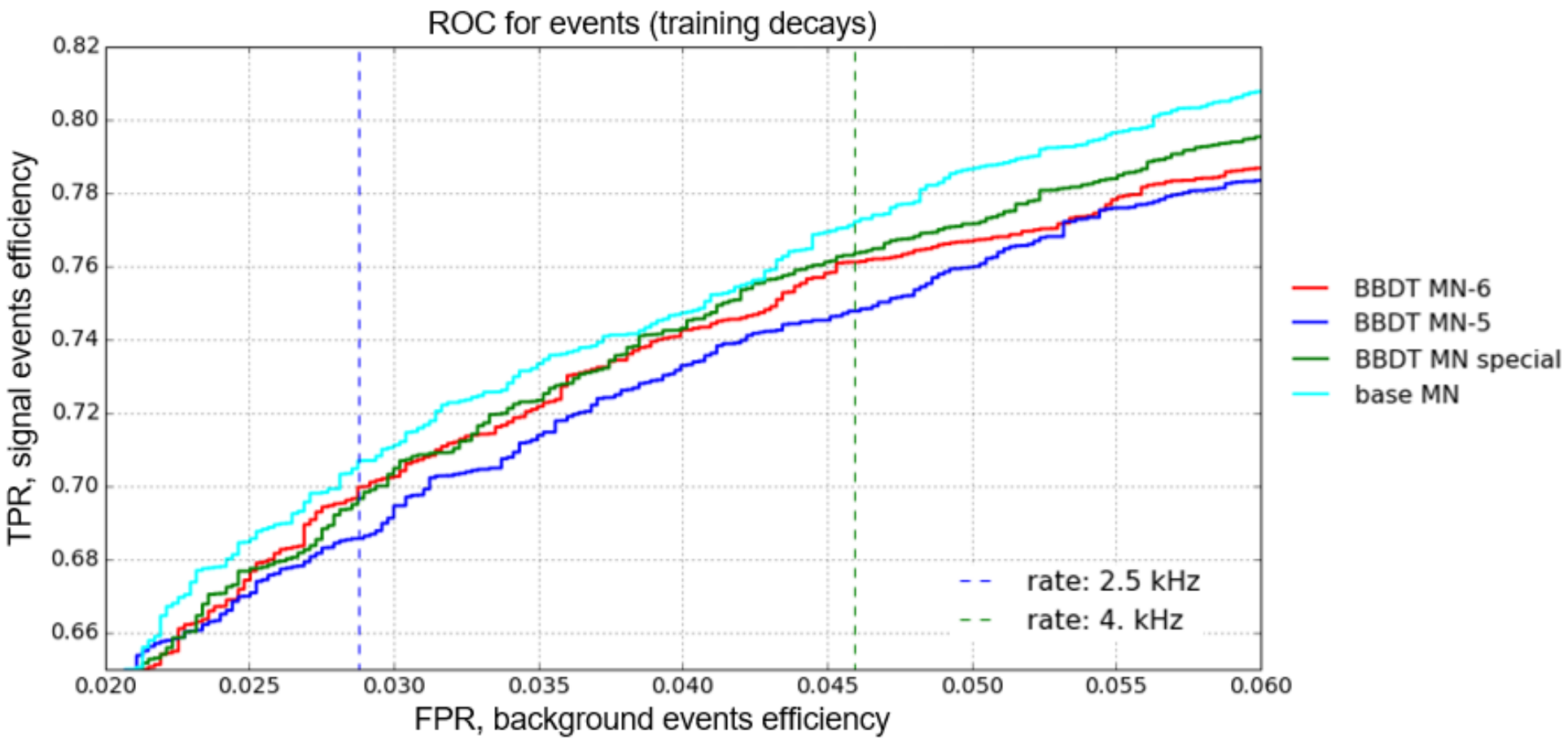
BBDT

- Hash features using bins before training
- Convert decision trees to n-dimensional table (lookup table)
- Table size is limited in RAM (1 GB), thus count of bins for each features should be small



BBDT, results

- Discretization reduces quality by ~1%



Post-pruning

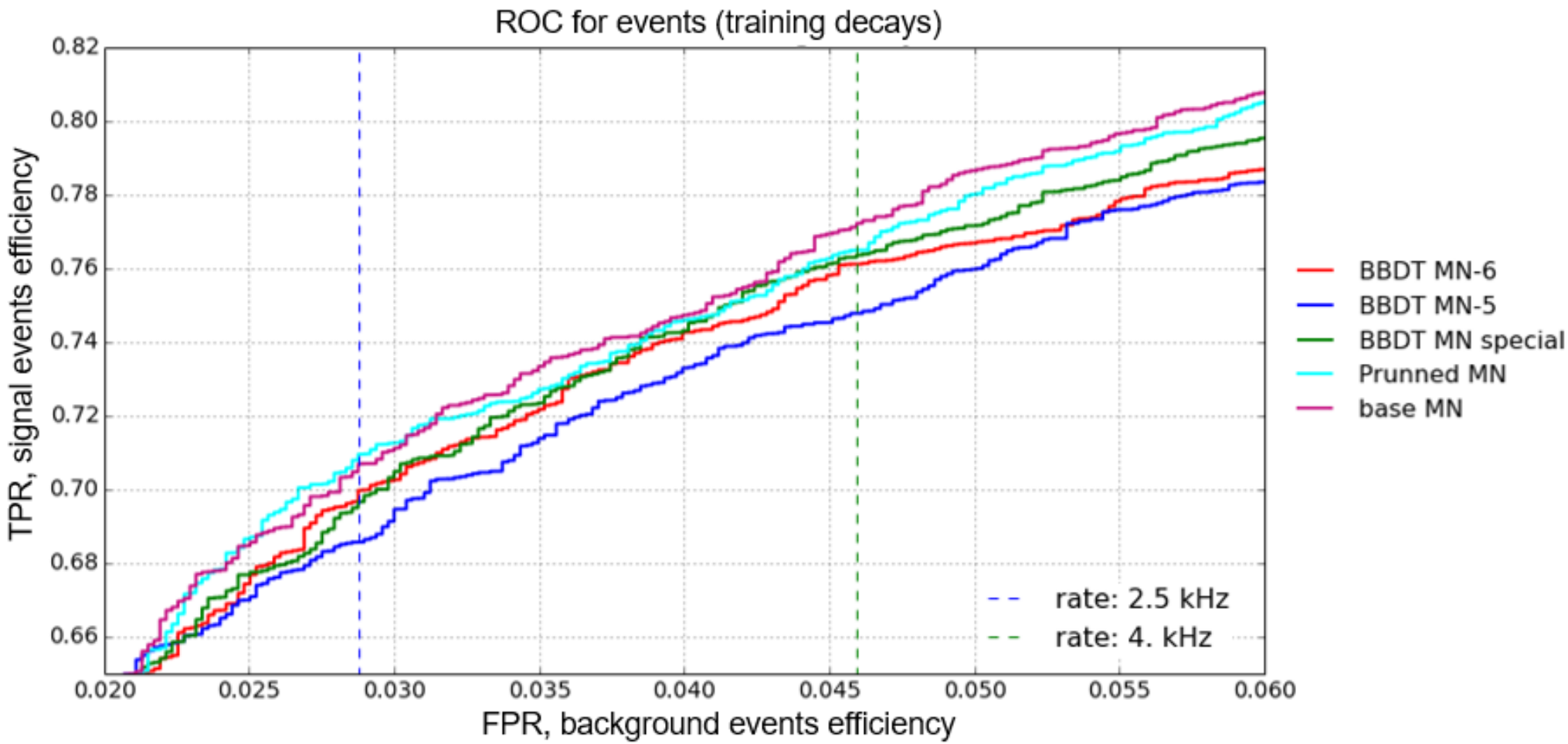
- Train gradient boosting over oblivious trees with several thousands trees
- Reduce this amount of trees to a hundred
- Greedily choose trees in a sequence from the initial ensemble to minimize a modified loss function:

$$\sum_{signal} \log\left(1 + e^{-F(x)}\right) + \sum_{background} e^{F(x)}$$

- At the same time change values in leaves (tree structure is preserved)

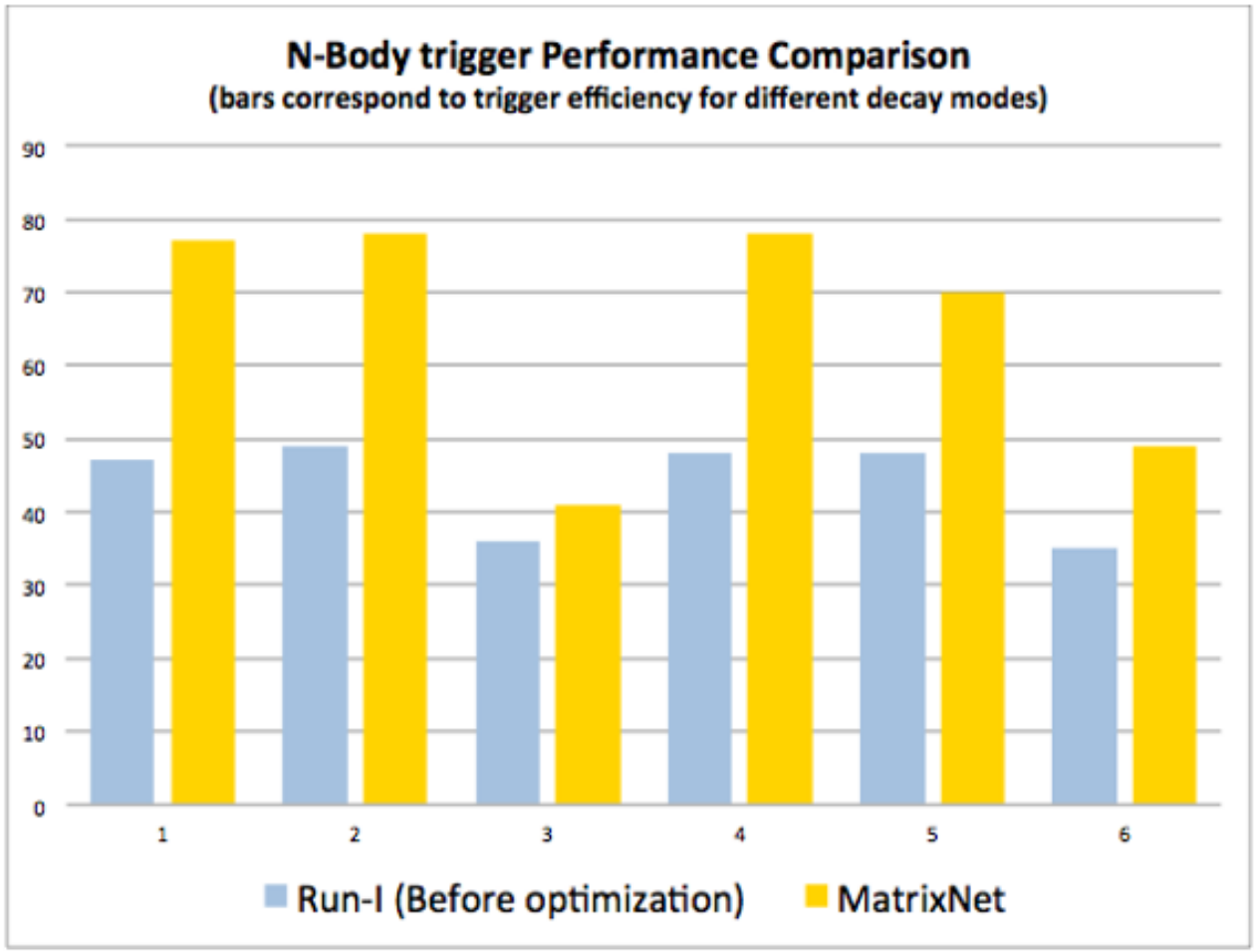
Post-pruning, results

- Pruned model is faster by several orders of magnitude comparing
- to the initial model and has the same quality for rates 2.5-3 kHz



Topological trigger results (without RF trick)

- Interpretation: 50% improvement means that the physics result, which otherwise would require 3 years of data taking, could be merely obtained just in 2 years.



When **High Energy Physics** meets **Machine Learning**