# Сотрудничество LHCb с Яндекс

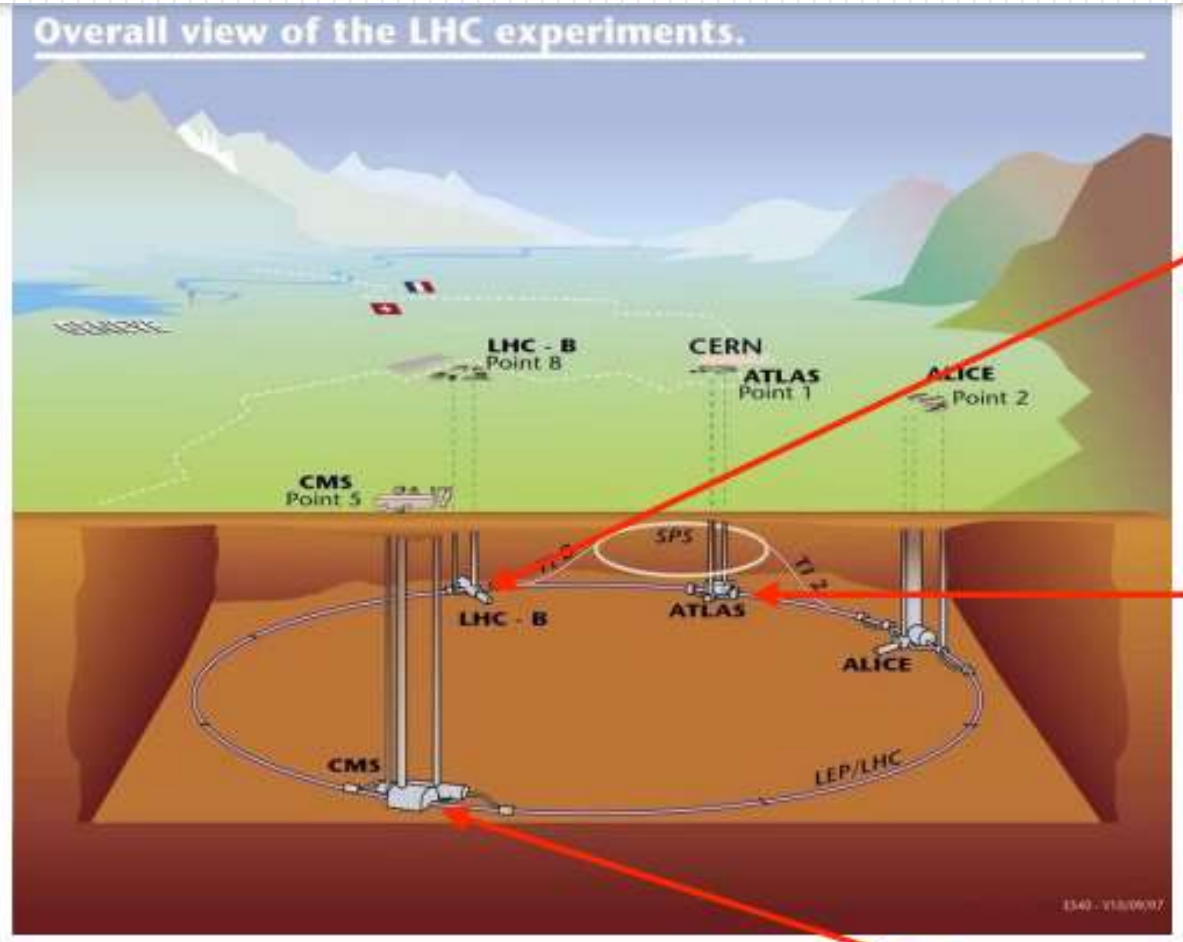
ПОДГОТОВИЛ:
БИКБАЕВ Т.Э.

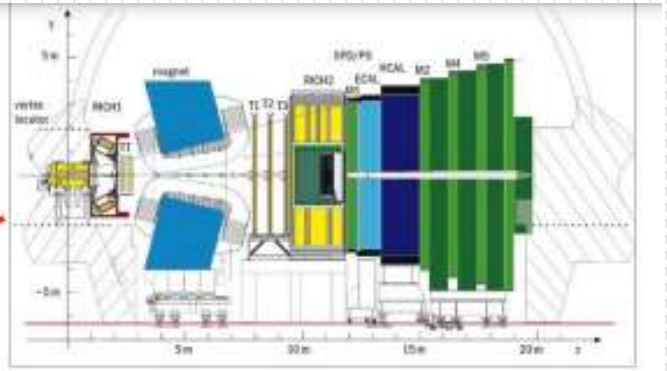

Яндекс

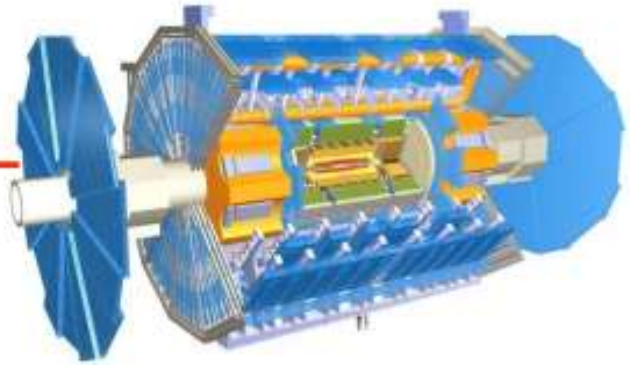
# Что такое Яндекс?

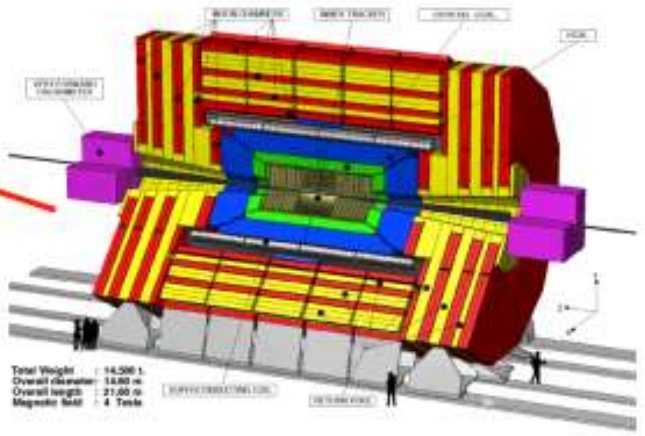

- **Технологии поиска**
- **Технологии машинного обучения**
- **Технологии работы с большими объемами данных**

# Большой Адронный Коллайдер (LHC)



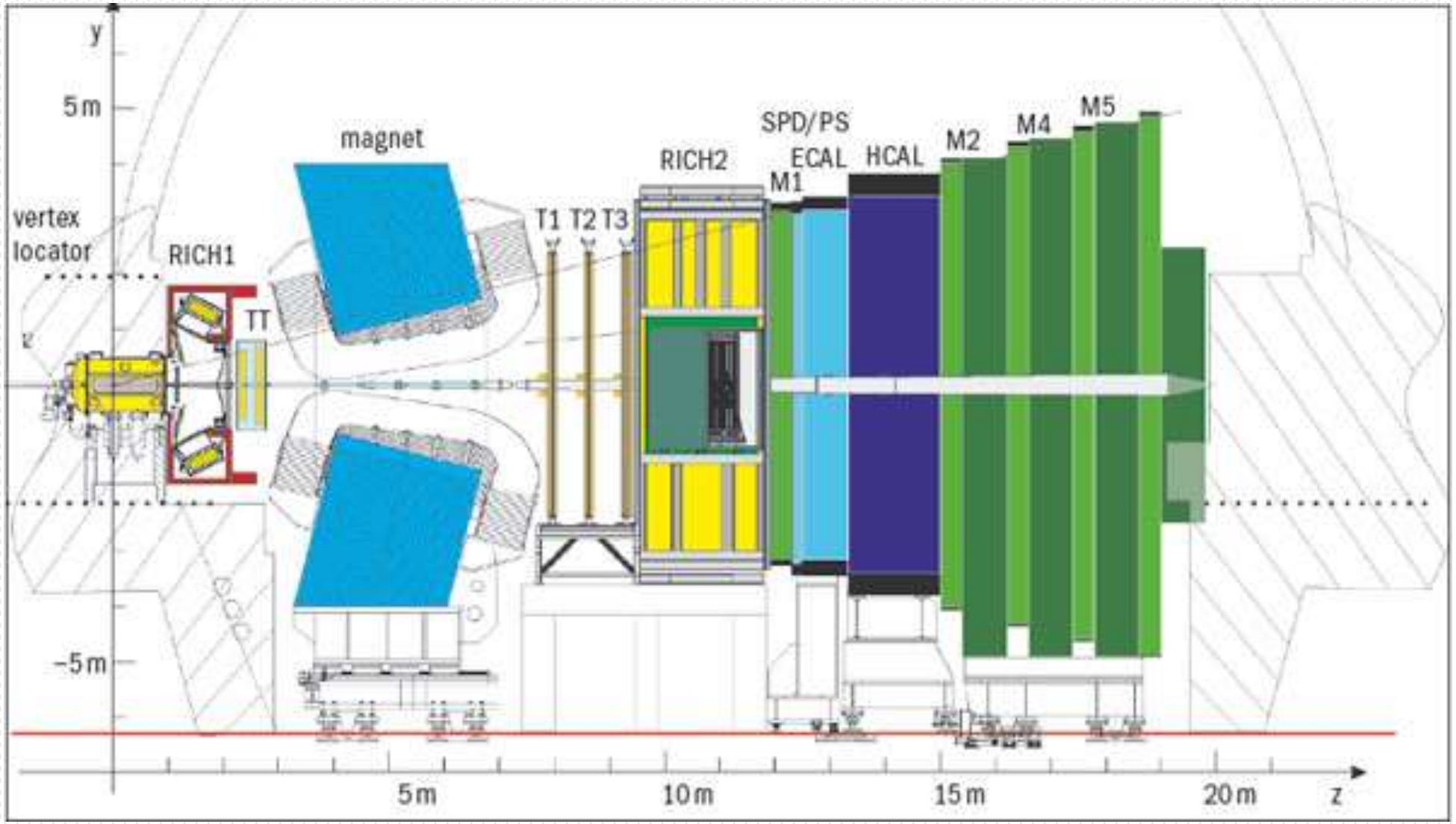







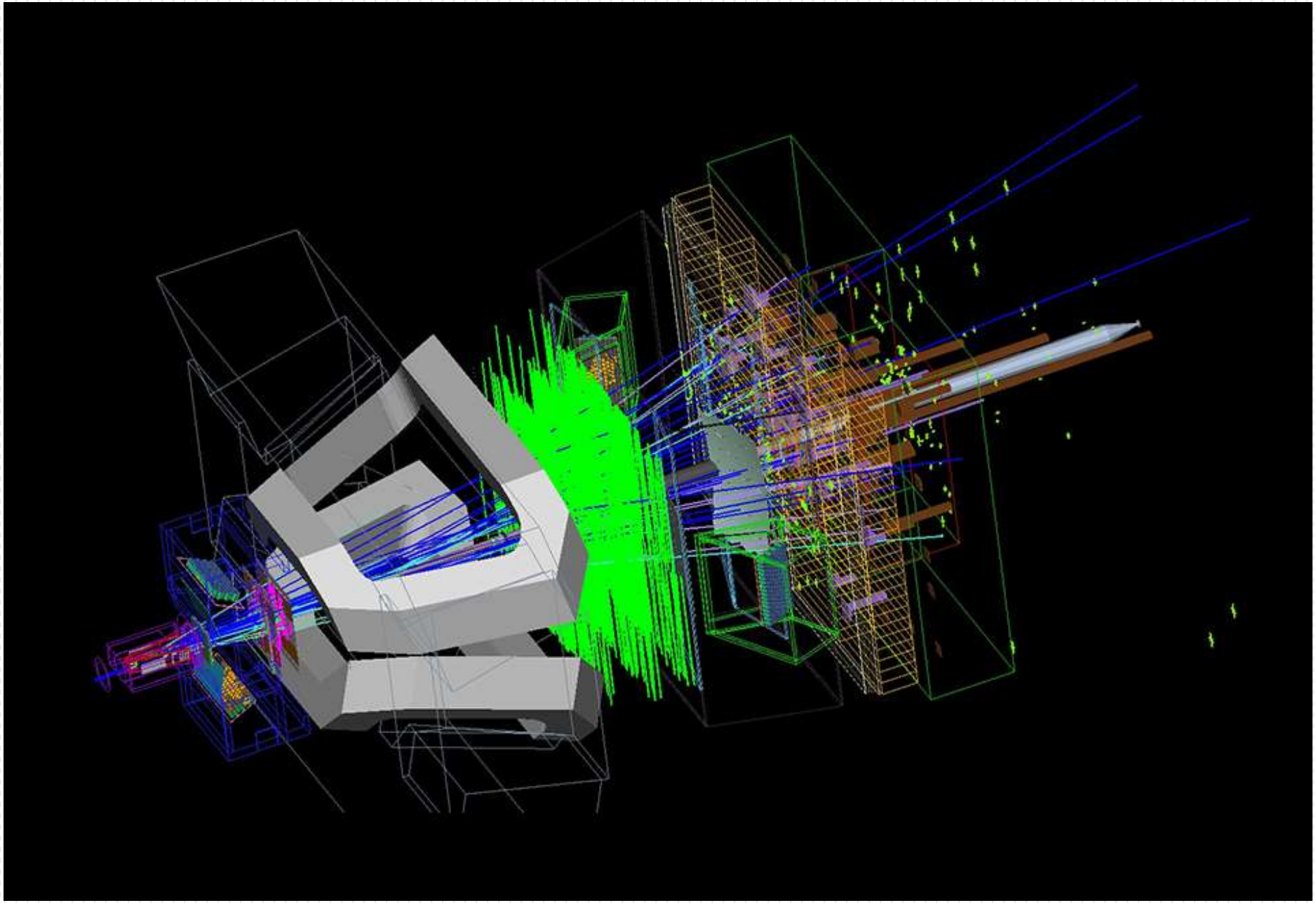
- 2 протонных пучка, скорость - 99.9999991% скорости света
- 4 точки пересечения пучков
- энергия в пучках ~ энергии Титаника на полном ходу

# Large Hadron Collider beauty experiment (LHCb)



**B-мезоны**

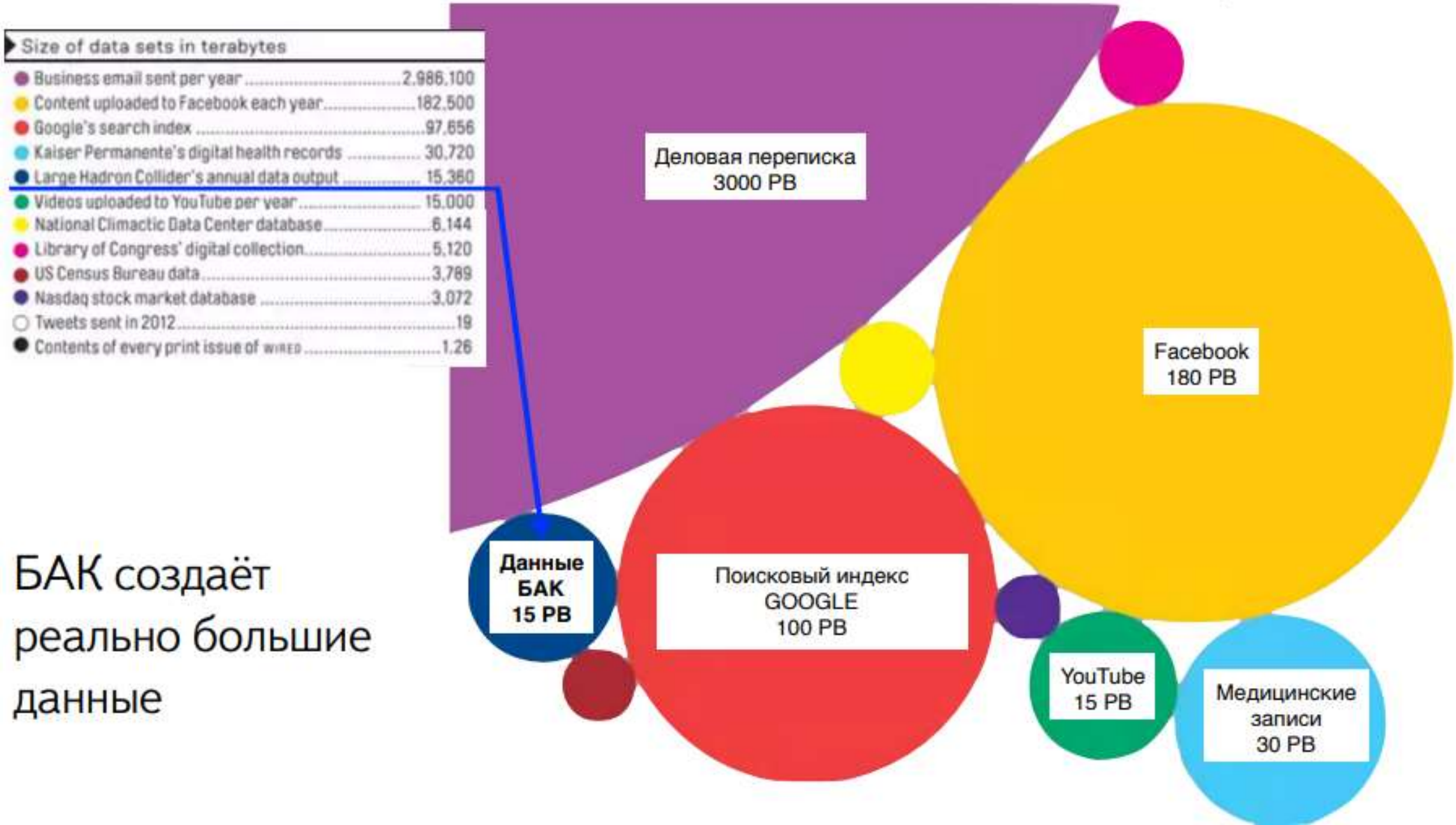
| Частица | Символ | Анти-частица | Кварковый состав | Заряд | Изоспин (I) | Спин и чётность ($J^P$) | Энергия покоя (МэВ/с²) | S | C | B' | Время жизни (с) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B-мезон | $B^+$ | $B^-$ | $u\bar{b}$ | +1 | ½ | $0^-$ | 5279,15 ± 0,31 | 0 | 0 | +1 | $(1{,}638 \pm 0{,}011)\cdot 10^{-12}$ |
| Нейтральный B-мезон | $B^0$ | $\bar{B}^0$ | $d\bar{b}$ | 0 | ½ | $0^-$ | 5279,53 ± 0,33 | 0 | 0 | +1 | $(1{,}530 \pm 0{,}009)\cdot 10^{-12}$ |
| Странный B-мезон | $B_s^0$ | $\bar{B}_s^0$ | $s\bar{b}$ | 0 | 0 | $0^-$ | 5366,3 ± 0,6 | −1 | 0 | +1 | $(1{,}470^{+0{,}027}_{-0{,}026})\cdot 10^{-12}$ |
| Очарованный B-мезон | $B_c^+$ | $B_c^-$ | $c\bar{b}$ | +1 | 0 | $0^-$ | 6276 ± 4 | 0 | +1 | +1 | $(0{,}46 \pm 0{,}07)\cdot 10^{-12}$ |



Столкновение двух протонов, зарегистрированное детектором LHCb

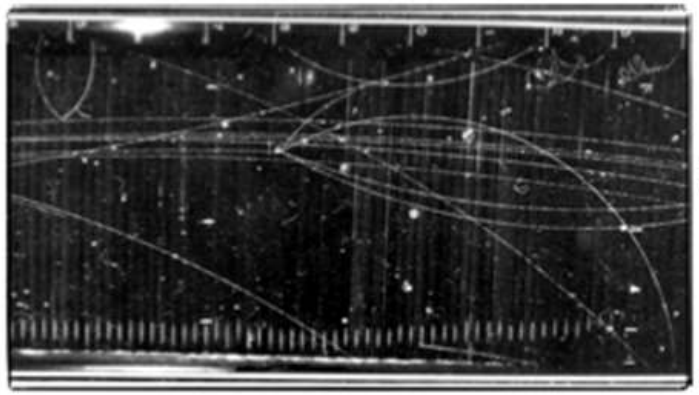
## данные, накопленные в 2012 году

адаптировано из http://www.wired.com/2013/04/bigdata/

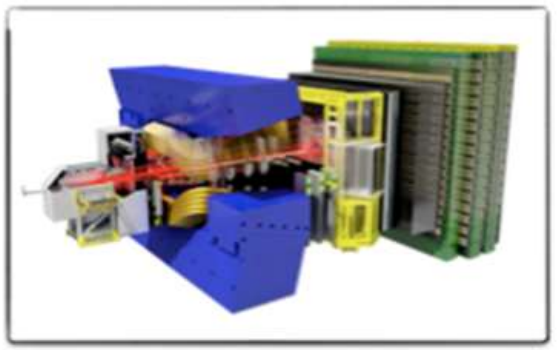



БАК создаёт реально большие данные

# What is Physics? Yesterday



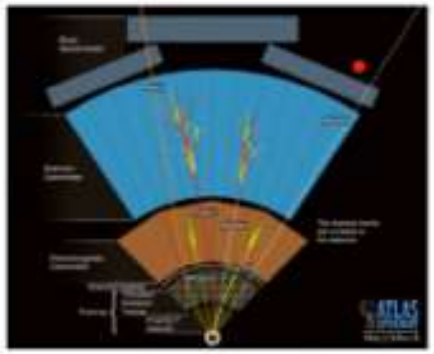
# What is Physics? Today



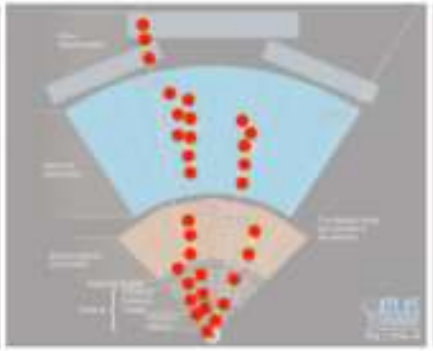
Сложные алгоритмы реконструкции событий и анализа данных

Агрессивное, многошаговое уменьшение размерности задачи

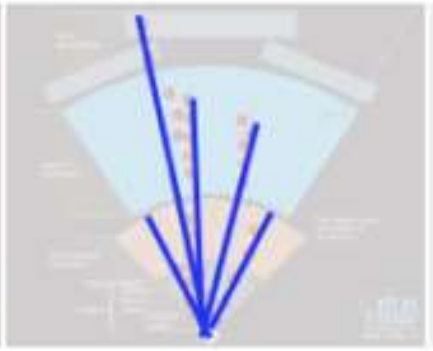
Широкий простор для применения машинного обучения и искусственного интеллекта



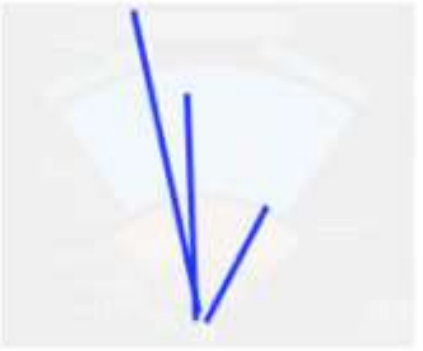
данные с детектора ~$10^7$



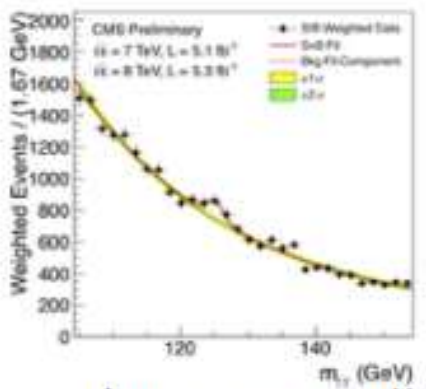
кластеризация ~$10^4$



реконструкция ~50



отбор ~10




физический результат ~1

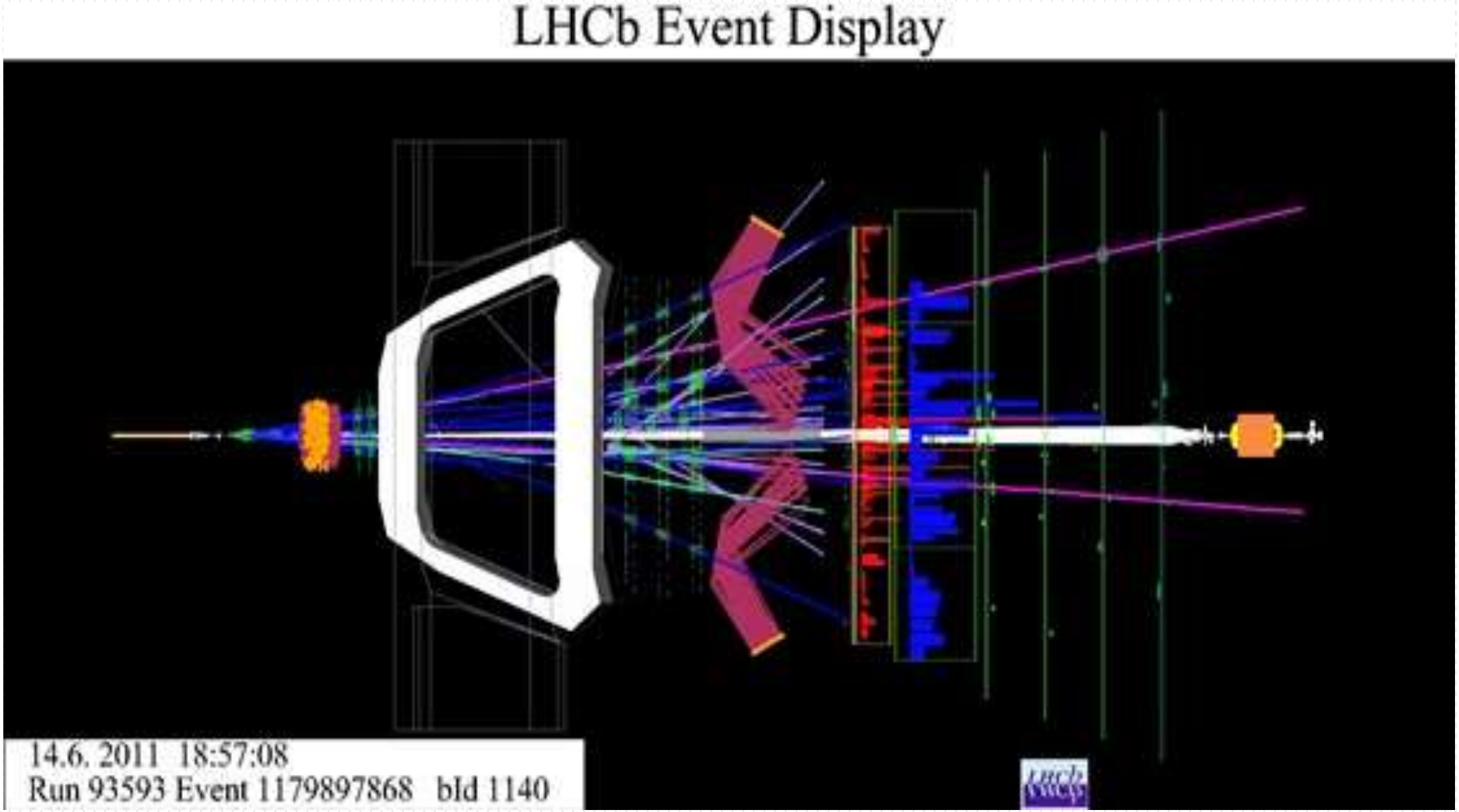
# Yandex for CERN

CERN openlab

Datacenter

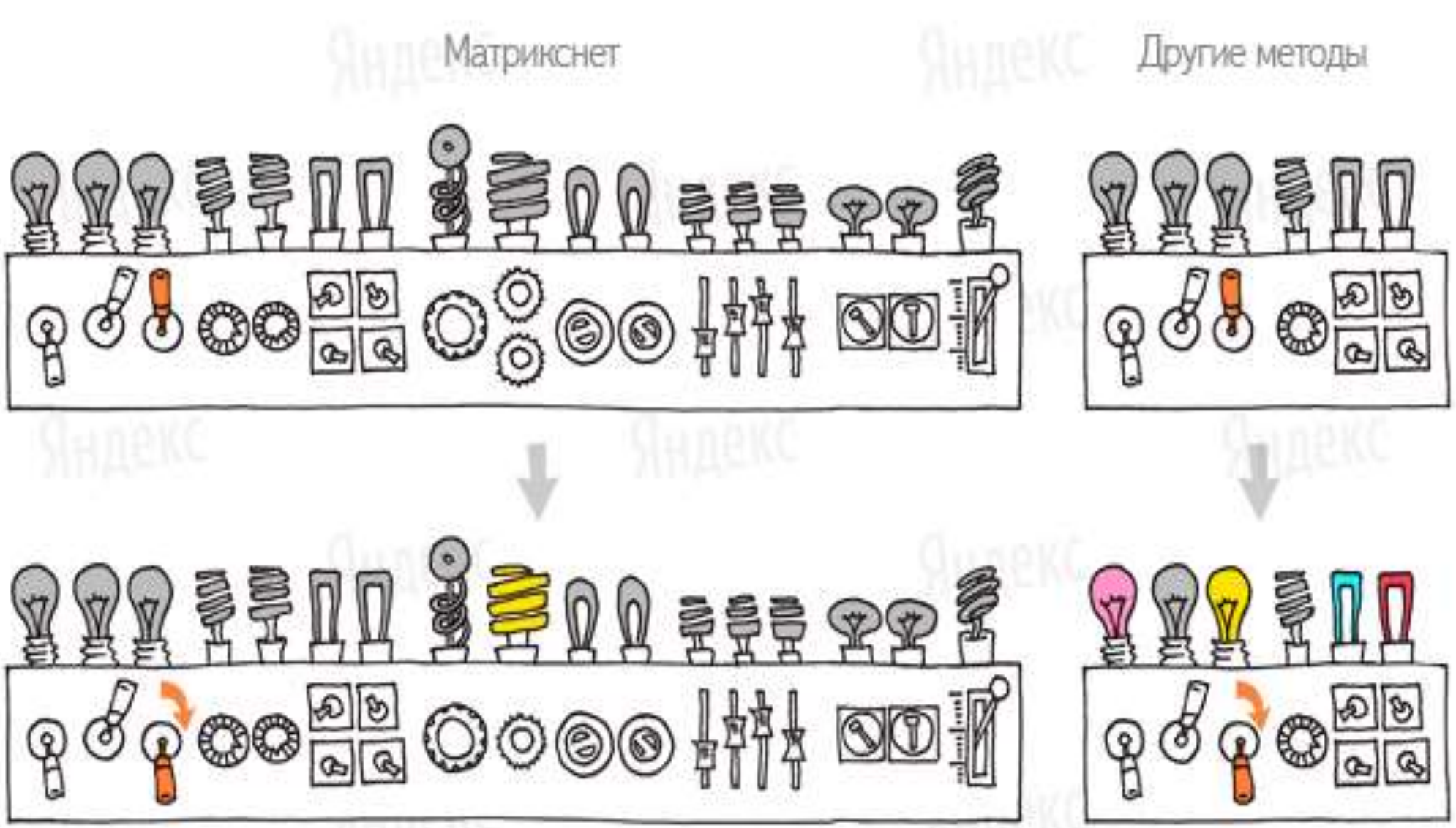
Event search

Matrix Net

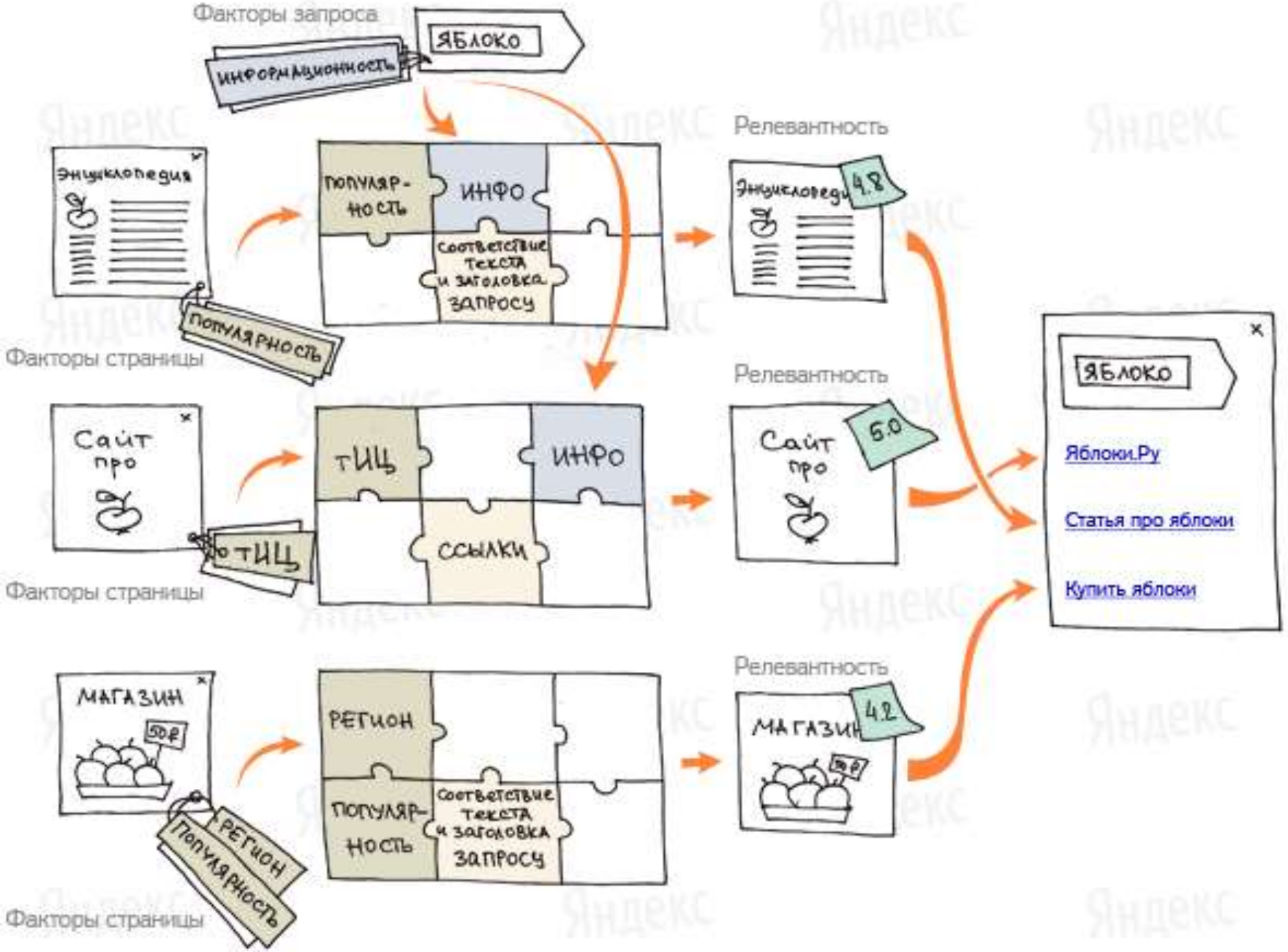


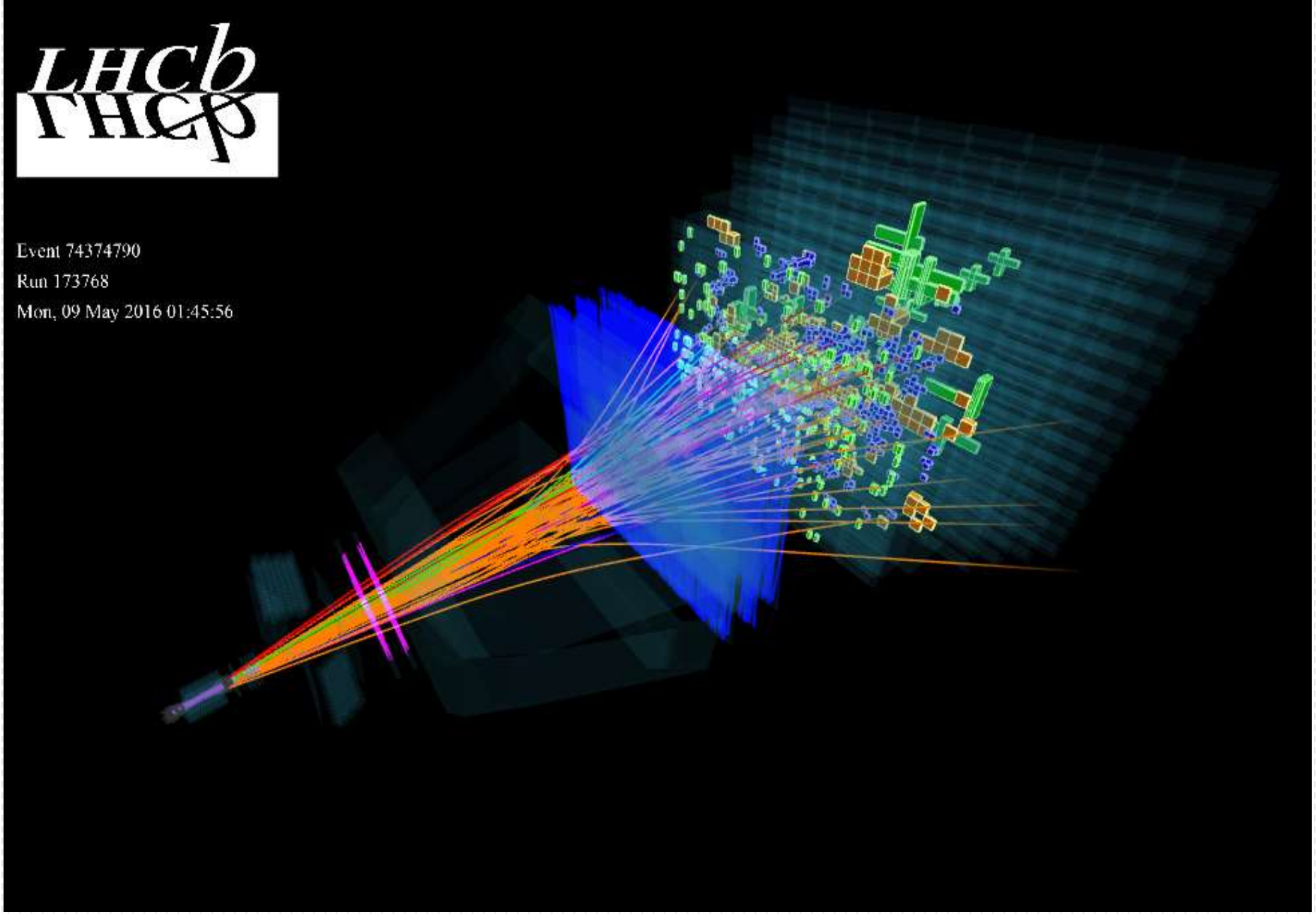

Кандидат на распад Bs0-> mumu, наблюдаемый в эксперименте LHCb (фото ЦЕРН)

# MatrixNet

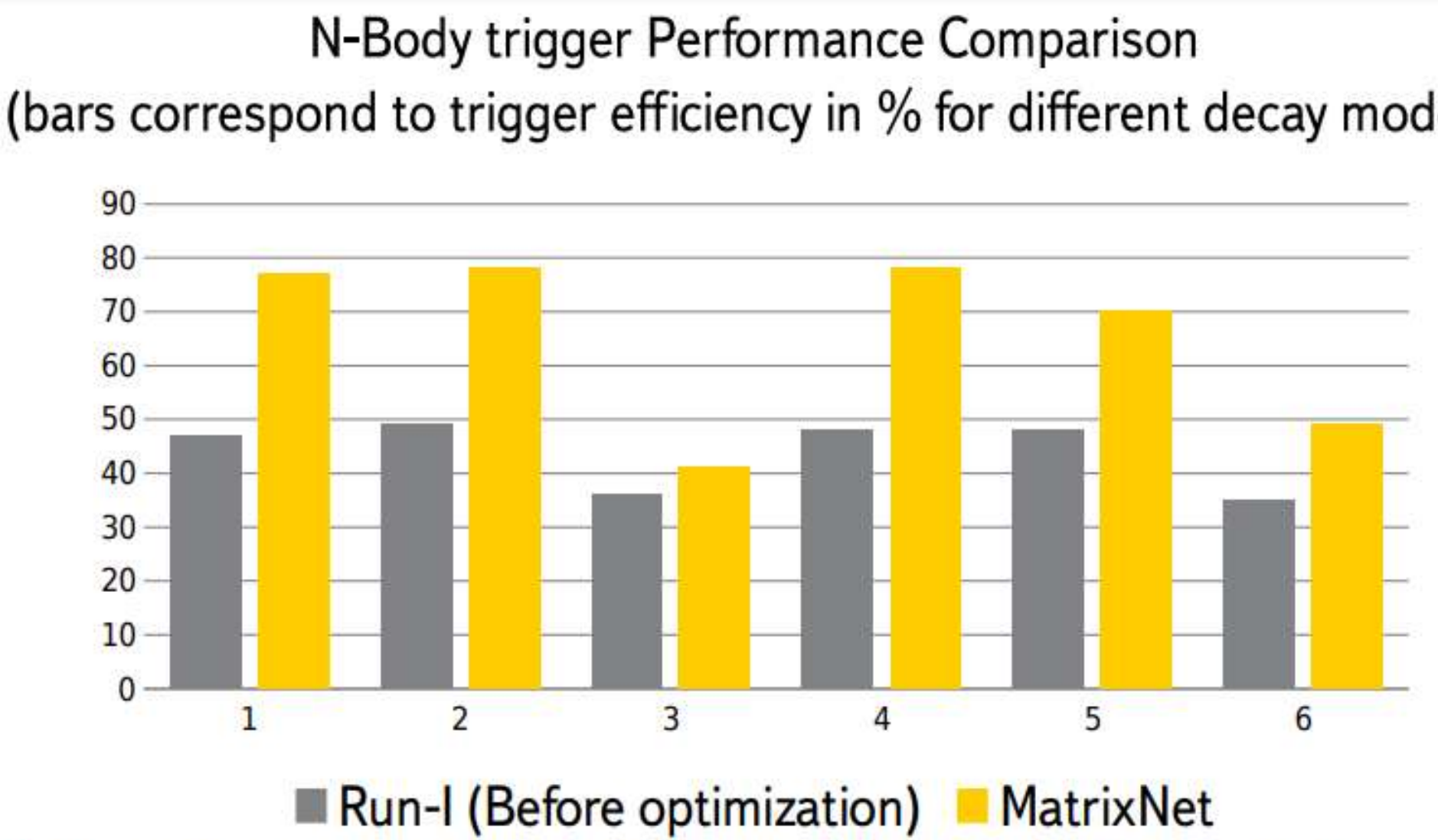


# Как устроено ранжирование






***Типичное событие на LHCb. При столкновении протонов возник пучок частиц, которые пролетели через детектор, определивший их тип и энергию. LHCb Experiment***




- Использование MatrixNet (поисковый движок Яндекса) позволило существенно улучшить эффективность тригерного отбора событий


- Данные хранятся на дисках (быстро, но дорого) и/или на лентах (дёшево, но медленно)
- С использованием машинного обучения удалось предсказать востребованность различных файлов и соответственно реорганизовать их на диске/ленте
- Результат: экономия 40% данных LHCb при ошибке предсказания ~1%

ROOT

Data Analysis Framework

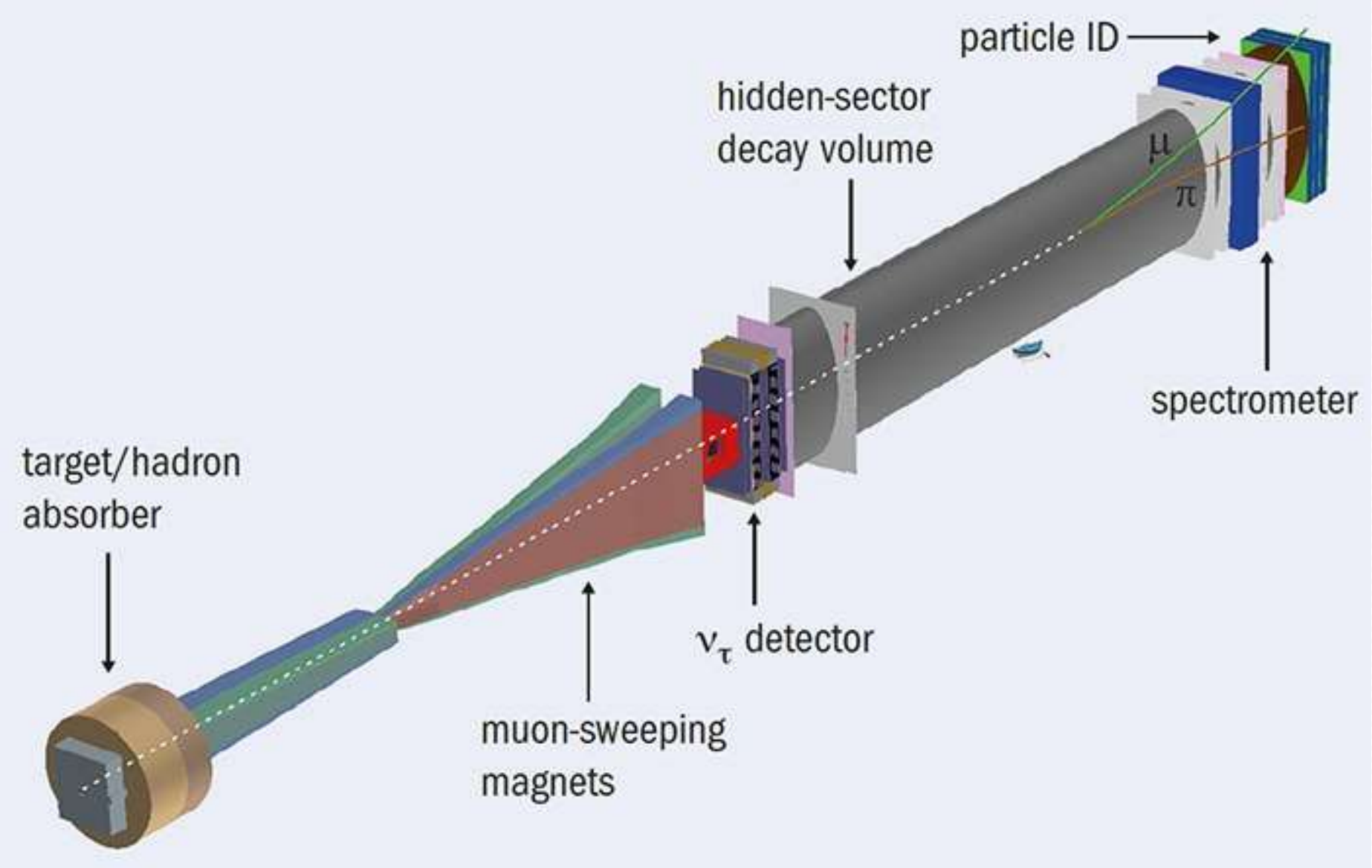

http://tmva.sourceforge.net/

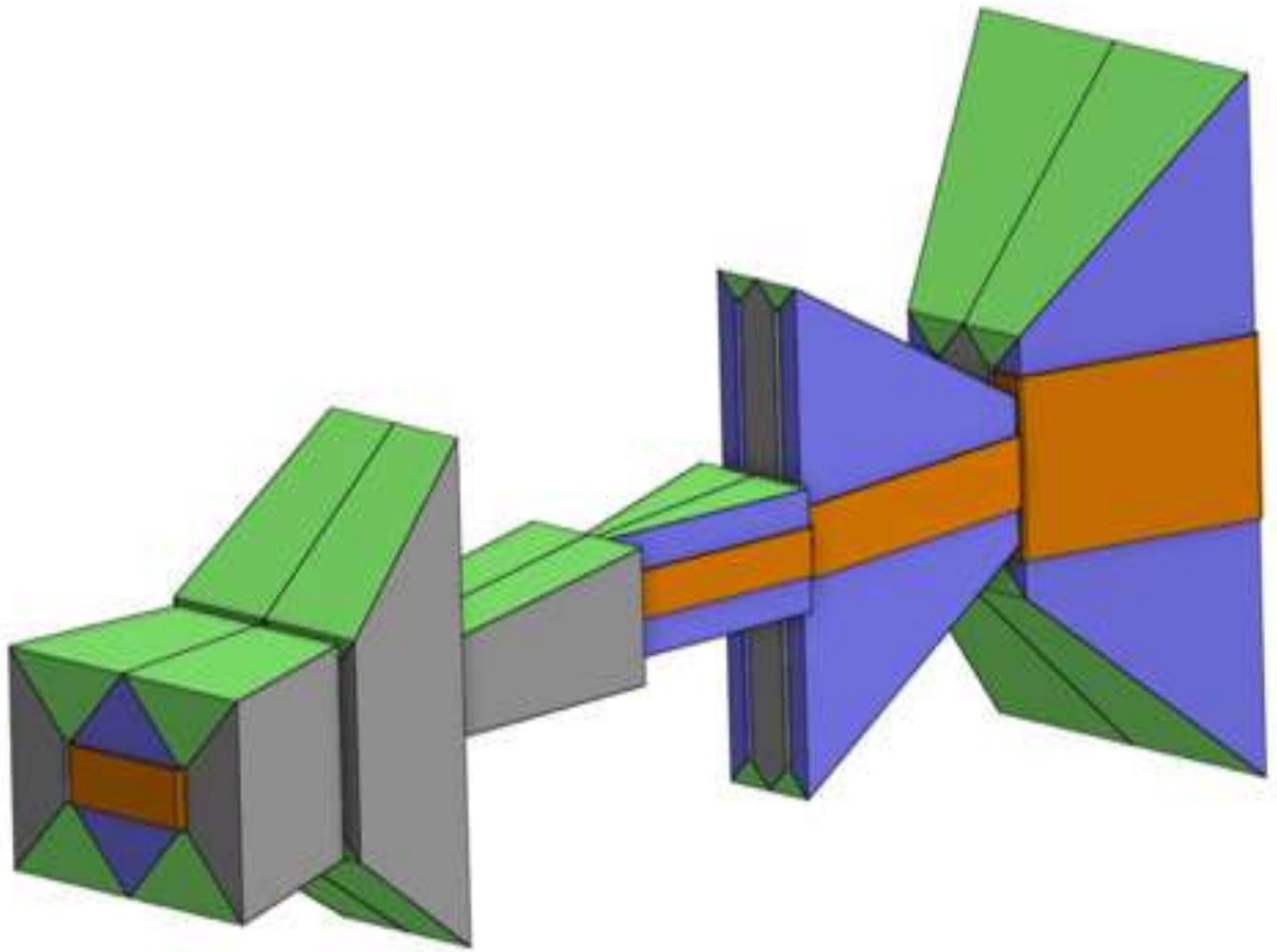
LAMBDA • HSE

ШКОЛА АНАЛИЗА ДАННЫХ

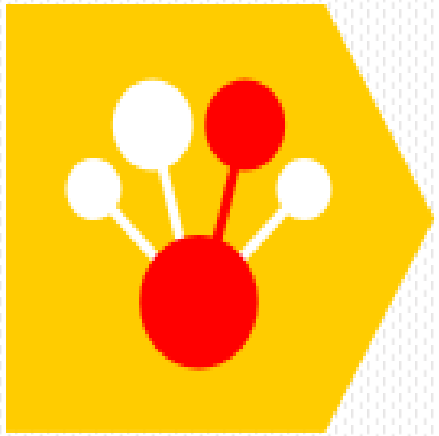
Научно-учебная лаборатория методов анализа больших данных

# Общая схема эксперимента SHiP(Search for Hidden Particles)



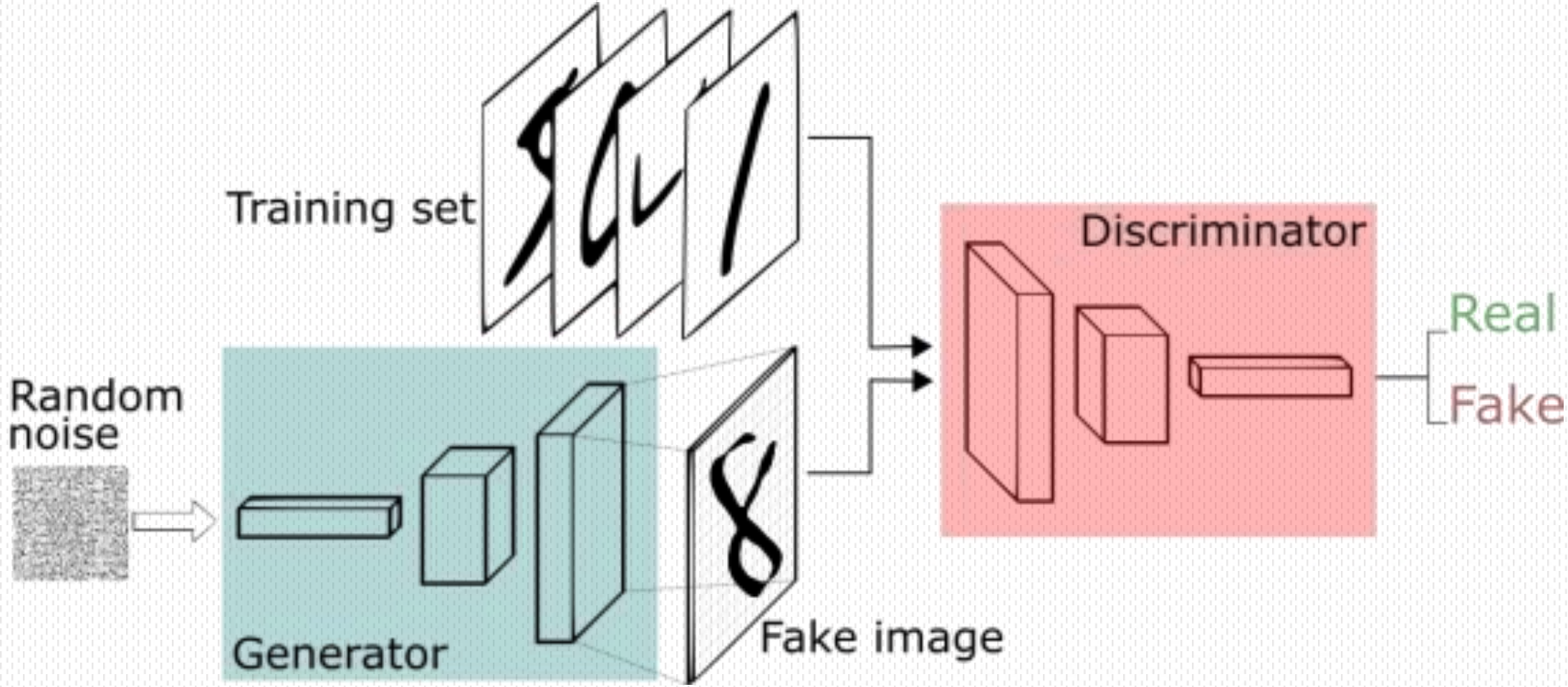


Конфигурация магнитов для эксперимента SHiP

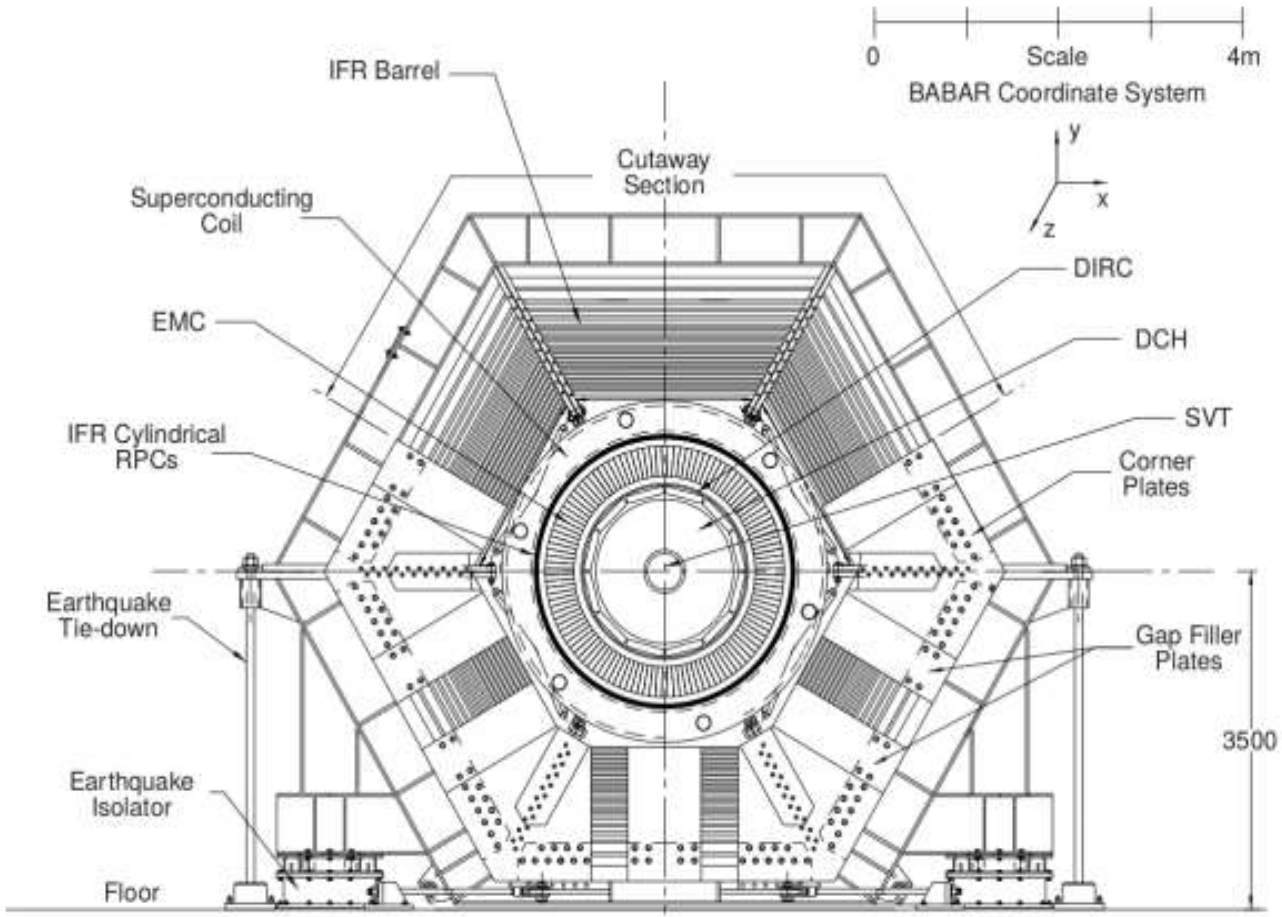


Yandex
CatBoost

**CatBoost** — открытая программная библиотека, разработанная компанией Яндекс и реализующая уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга.




Генеративно-состязательная сеть (англ. Generative adversarial network, GAN)

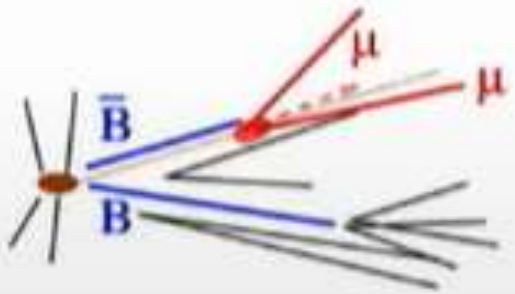




Поперечное сечение детектора $BABAR$ в поперечном направлении

# Предварительный отбор (триггеры)

общее количество событий, наблюдаемых детектором за 2011-2012 гг: $\sim 10^{15}$
размерность каждого события $\sim 10^{7}$

- частота событий на входе триггера ~ $10^{7}$ событий/секунду
- определение траекторий по хитам различных элеметов детектора, размерность $10^{4}$
- определение типов частиц (траекторий): мюон, каон, пион, протон, ...
- реконструкция, восстановление структуры распадов:
  - 2 трека сходятся в одну точку - *вторичная вершина*
  - *вторичная вершина* находится на расстоянии от *первичной вершины* (точки соударений протонов)
- частота событий на выходе триггера ~ $10^{4}$ событий/секунду
- сохраняются в распределенной система хранения и обработки данных



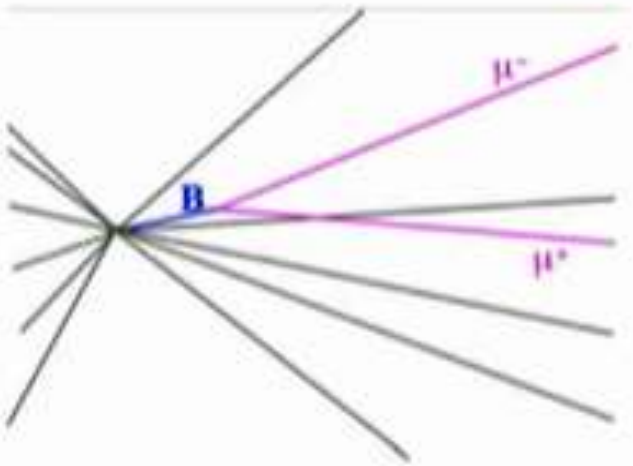
# Отбор событий-кандидатов

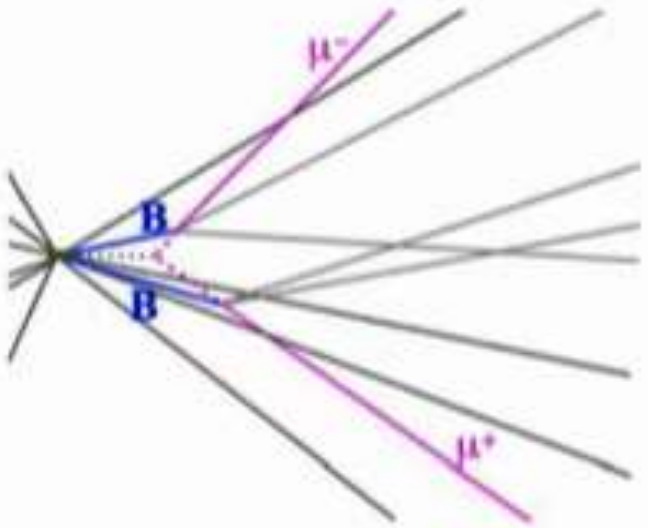
событие-кандидат - распад родительской частицы заданной массы на интересующие дочерние частицы. внутри одного события могут присутствовать несколько распадов

- 2 трека мюонов сходятся $\sim$ в одну точку (вторичная вершина)
- вторичная вершина находится на расстоянии от первичной вершины
- размерность "события-кандидата" - $10^2$
- инвариантная масса родительской частицы *, восстановленной из 2 мюонов, оказывается близкой к массам $B_s$ или $B_0$ ([4.9-6] ГэВ$/c^2$)
- каждое событие-кандидат представляем вектором признаков, размерность $\sim 10$

* вторичная вершина - точка распада частицы, породившей мюоны. поэтому эту частицу называют *родительской частицей*

# Сигнал, фон



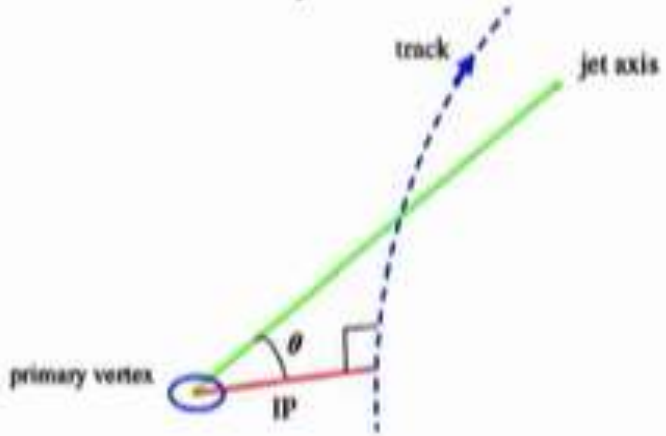

Сигнал:

- два мюона
- инвариантная масса в районе массы $B_s$, $B^0$
- хорошо реконструированная вторичная вершина на заметном (~ 2мм) расстоянии от первичной вершины, импульс которой сонаправлен с направлением полета

Фон:

- два распада $B$ на $\mu$ и $X$
- распад $B$ на $\mu$ и ошибочно идентифицированный мюон
  - $B_s \to K^- \mu^+ \nu$
- Одиночные распады B:
  - $B_s \to K^+ K^-$
- ...

# Признаки событий, B-мезон

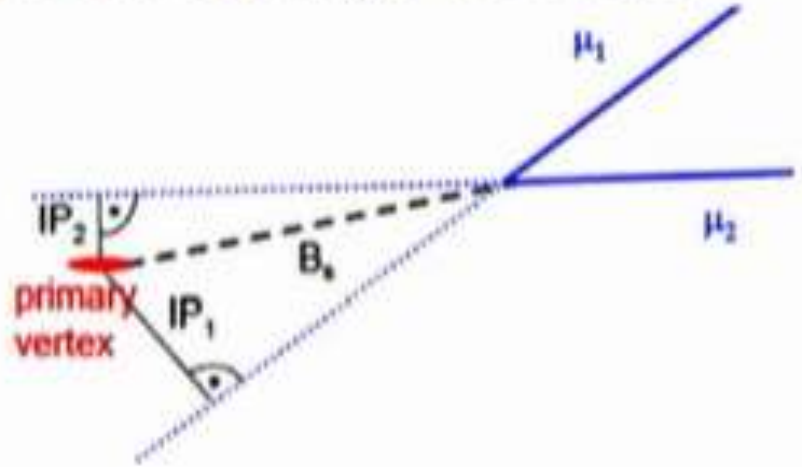



точка распада B - вторичная вершина распада

1. время жизни B
2. IP - расстояние от траектории B до первичной вершины
3. pT - поперечный импульс в системе координат детектора
4. изоляция B
5. угол между импульсом B и $P_{thrust}$
6. угол между импульсом $\mu^+$ в системе покоя B и $P_{thrust}$ в системе покоя B

$P_{thust}$ - сумма импульсов всех треков, начинающихся из точки распада другого b-адрона

# Признаки событий, мюоны

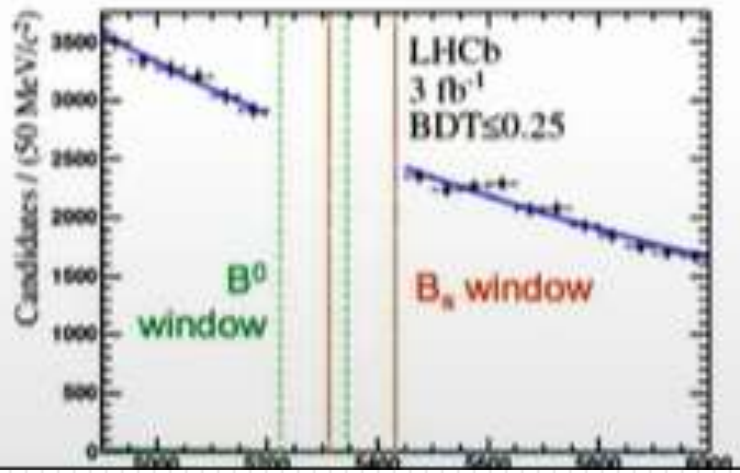



1. минимальная значимость прицельного параметра
2. минимальное расстояние между 2 мюонами
3. изоляция трека
4. угол поляризации (the cosine of the angle between the muon momentum in the dimuon rest frame and the vector perpendicular to both the B candidate momentum and the beam axis)
5. модуль разницы между углами $\phi$ мюонов в сферической системе координат
6. модуль разницы между псевдобыстротой мюонов

# Сигнальные регион, боковой регион, слепой анализ

- ожидаем сигнал в районе инвариантной массы родительской частицы
  - комбинаторный фон
  - сигнал (*hopefully*)
  - "пиковый" фон с двойной ошибкой идентификации
- вырезаем сигнальный регион и прячем до лучших времен
- боковые регионы - содержат только фон (различный)

**ВАЖНО: ожидаем, что распределение комбинаторного фона по массе должно принадлежать одному классу ($e^{-m}$) в сигнальном и боковом регионах**



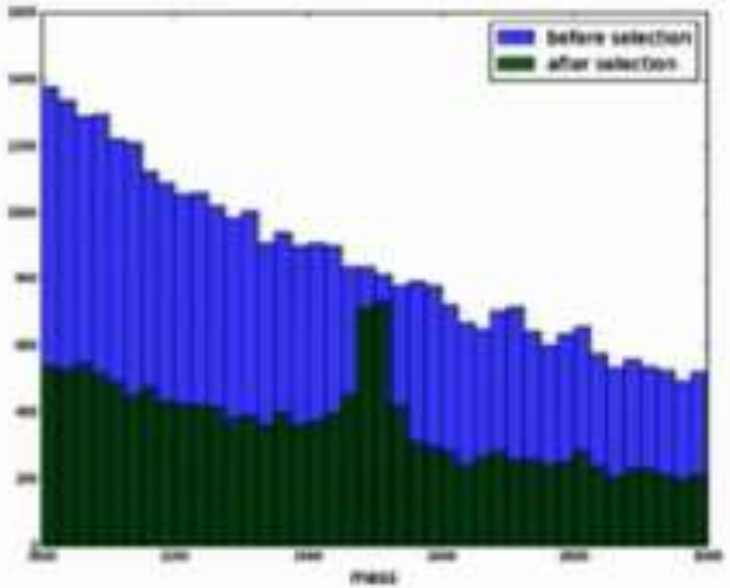
# Использование машинного обучения

- *не можем разметить сигнальные события в реальных данных!*
- сгенерировать распады для сигнальных каналов
  - конфигурация распада элементарных частиц
  - отклик детектора на распады
- сгенерировать распады для фона
- провести сгенерированные данные через те же этапы отбора событий как и реальные события
  - триггеры
  - идентификация частиц
  - реконструкция
- привести к признакам описания событий-кандидатов
- датасет из 12 колонок (признаков) с разметкой:
  - 1 - сигнал
  - 0 - фон

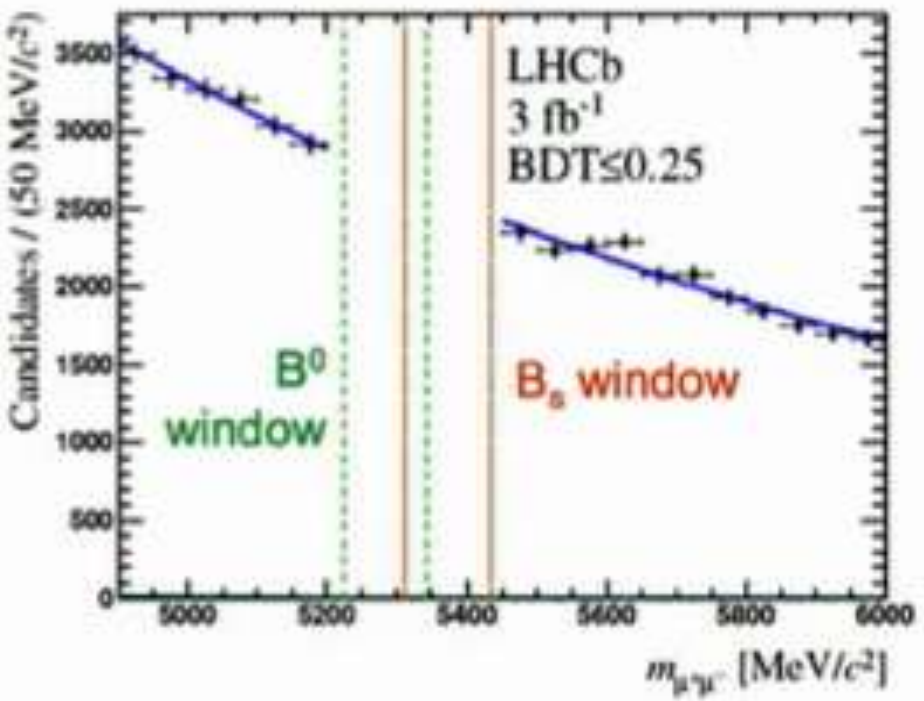
# Особенности обучения

- результат работы классификатора - оценка *количества* сигнальных событий, а не вероятности принадлежности отдельного события выбранному каналу
- другими словами, с помощью классификатора $g$ мы хотим найти область $G : g(x) > p$, в которых находится как можно больше сигнальных (true positive) событий и как можно меньше фоновых (false positive)

- **распределение фоновых событий, отобранных классификатором в сигнальном и боковых регионах, должно принадлежать одному классу, чтобы не создавать "искусственных пиков":**

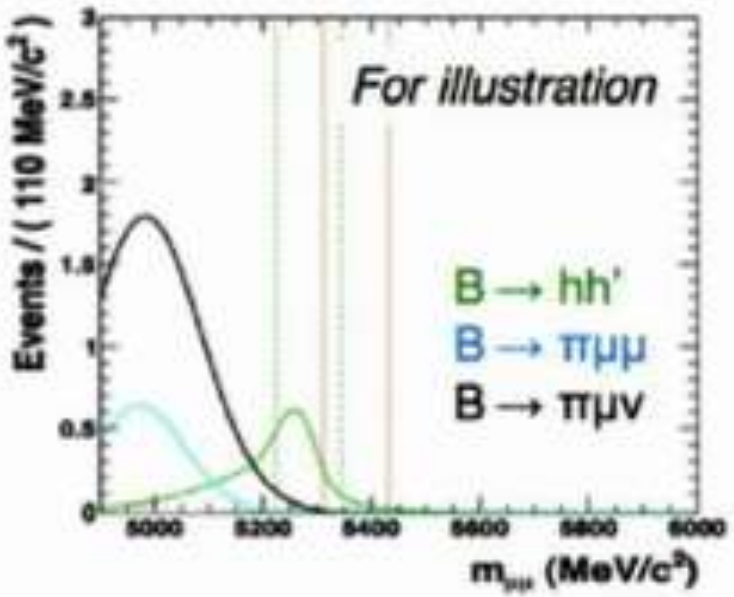


# Комбинаторный фон



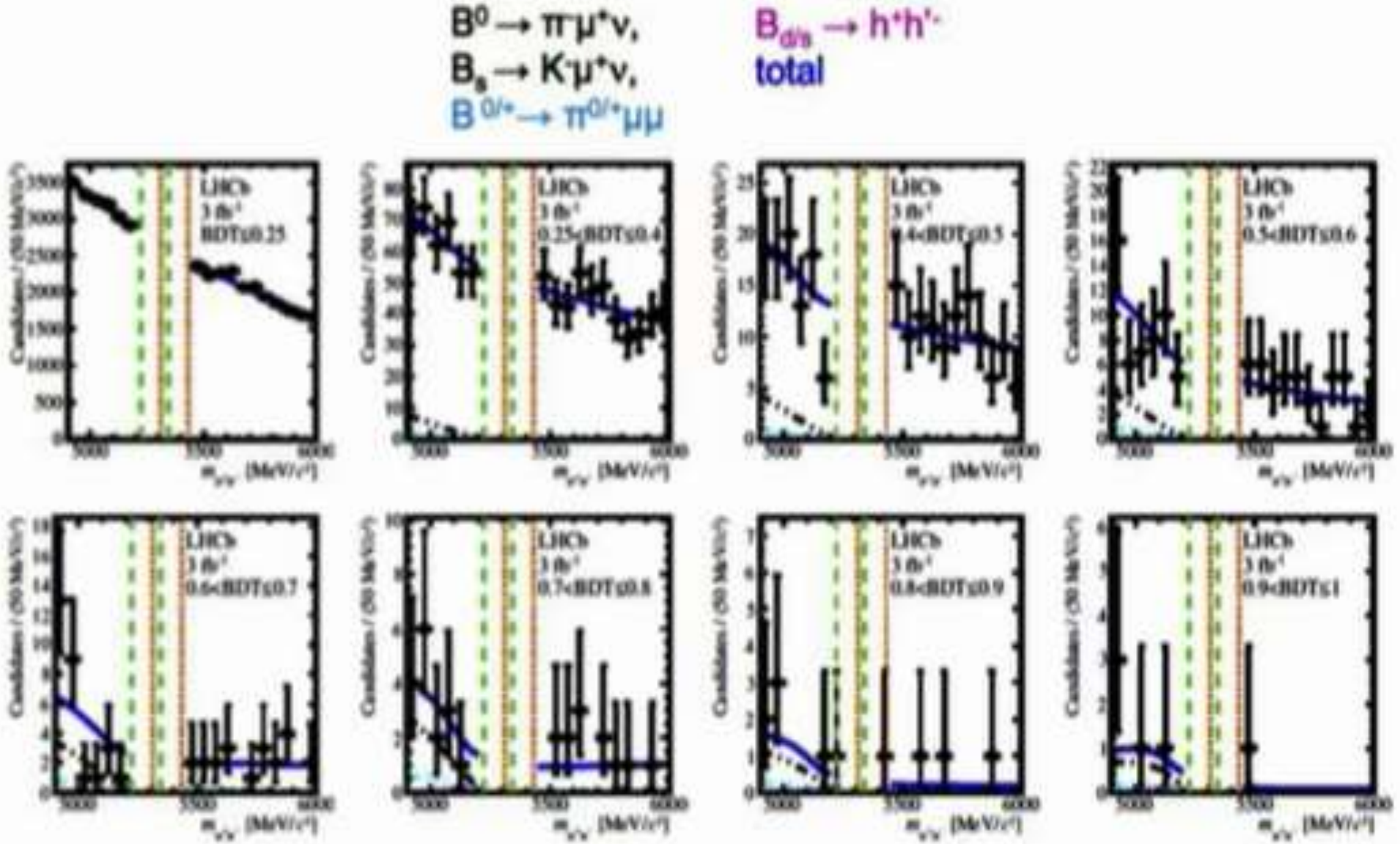

- основной источник фона распадов вида $bb \to \mu\mu X$
- оцениваем ожидаемое количество фоновых событий в сигнальном регионе, аппроксимируя экспоненциальное распределение комбинаторного фона, на сигнальный регион для каждого интервала на значения порогов классификатора
- при высоких значениях на порог классификатора практически исчезает

# Пиковые фоны

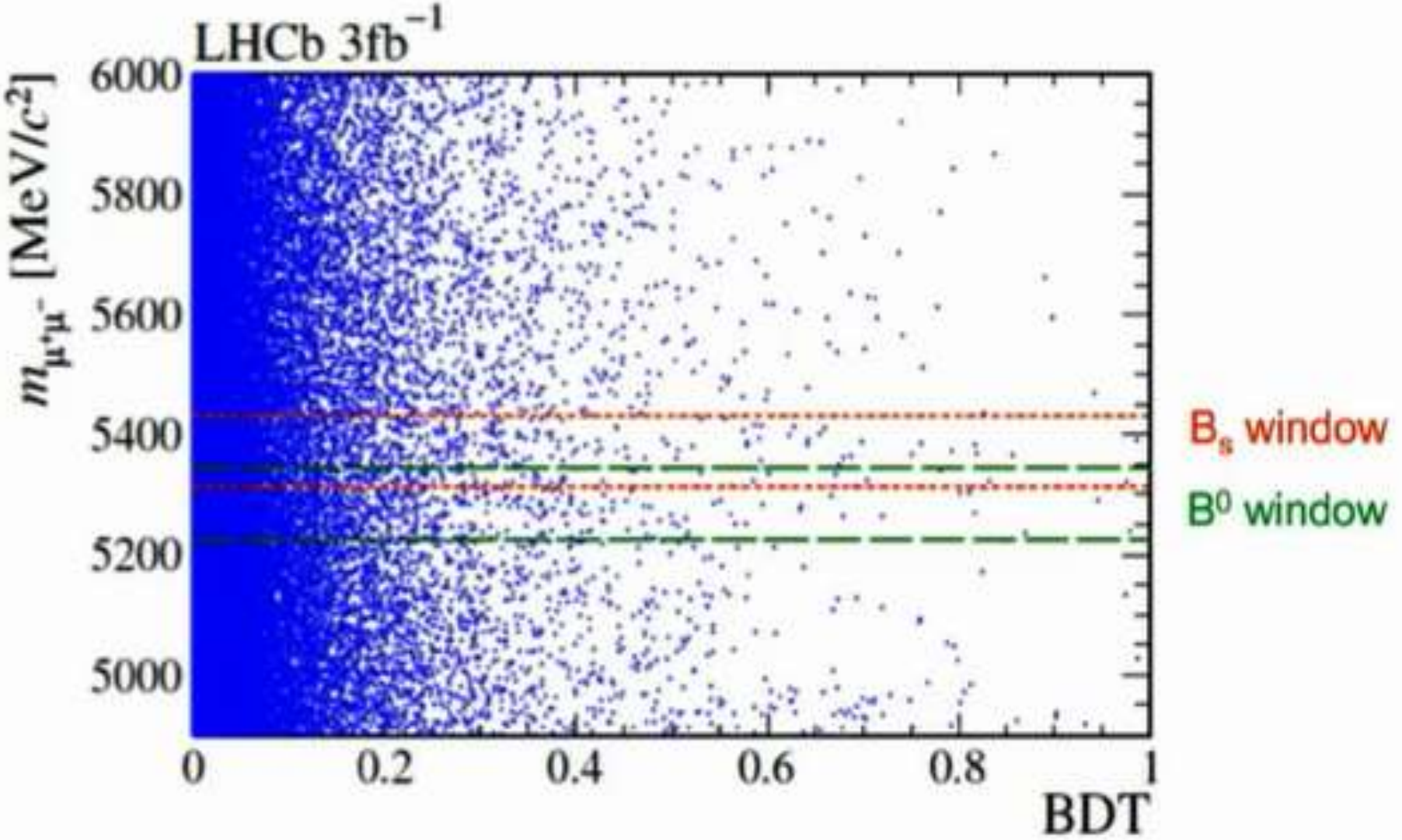



- в сигнальном регионе играет роль только $B \to h^+ h'^-$ (двойная мисидентификация мюонов)
- в боковых регионах, распады с одиночной ошибкой идентификации или 2 мюонами исходящими из одной вершины:
- $B^0 \to \pi^- \mu^+ \nu, B^{0/+} \to \pi^{0/+} \mu\mu$
- $B_s \to K^- \mu^+ \nu, B_c \to \pi^{0/+} J/\psi(\mu\mu)\mu\nu$
- $\Lambda_b \to p\mu\nu$
- прочие каналы вносят несущественный вклад

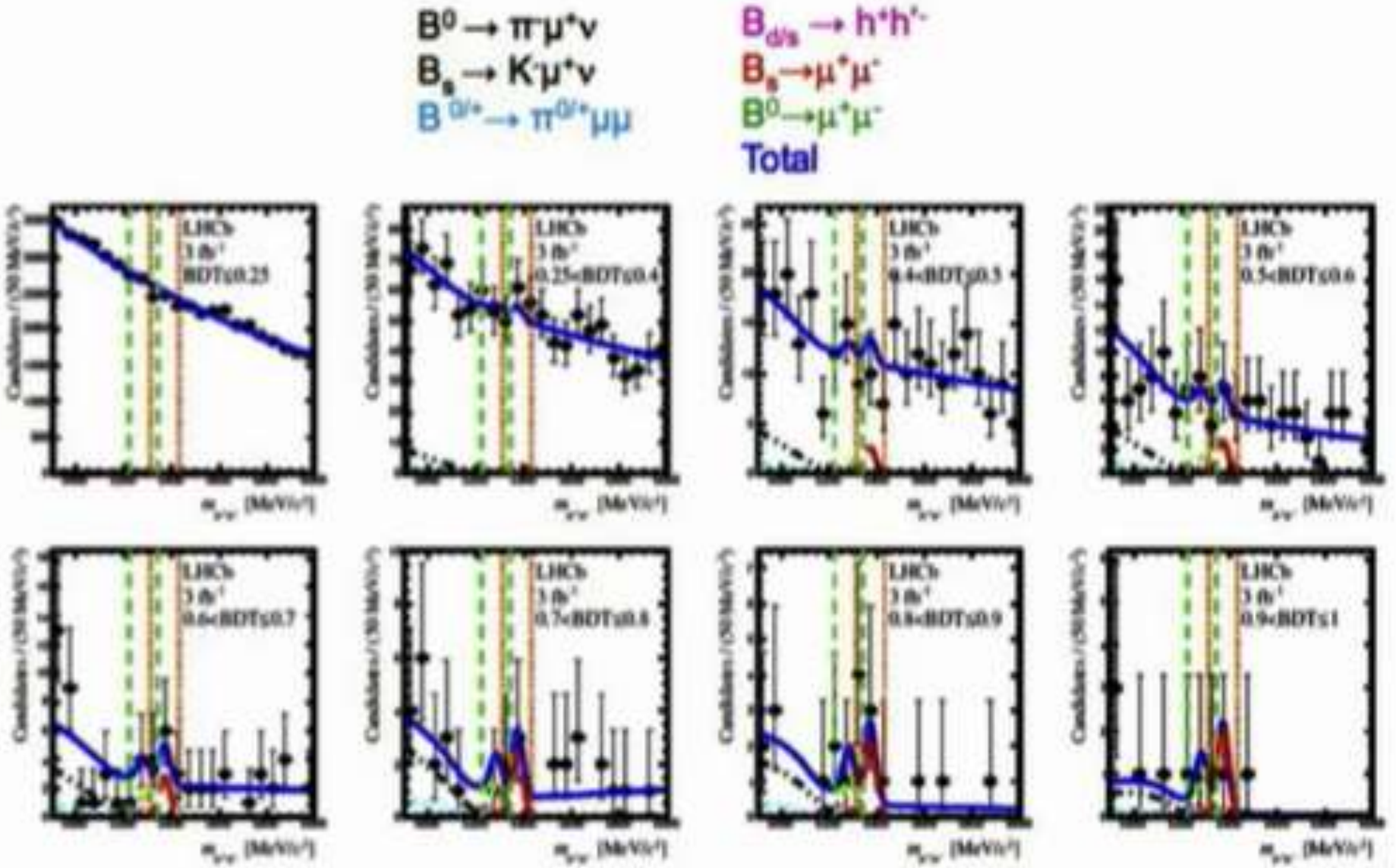
# Аппроксимация фоновых каналов



# Результаты применения классификатора к реальным данным

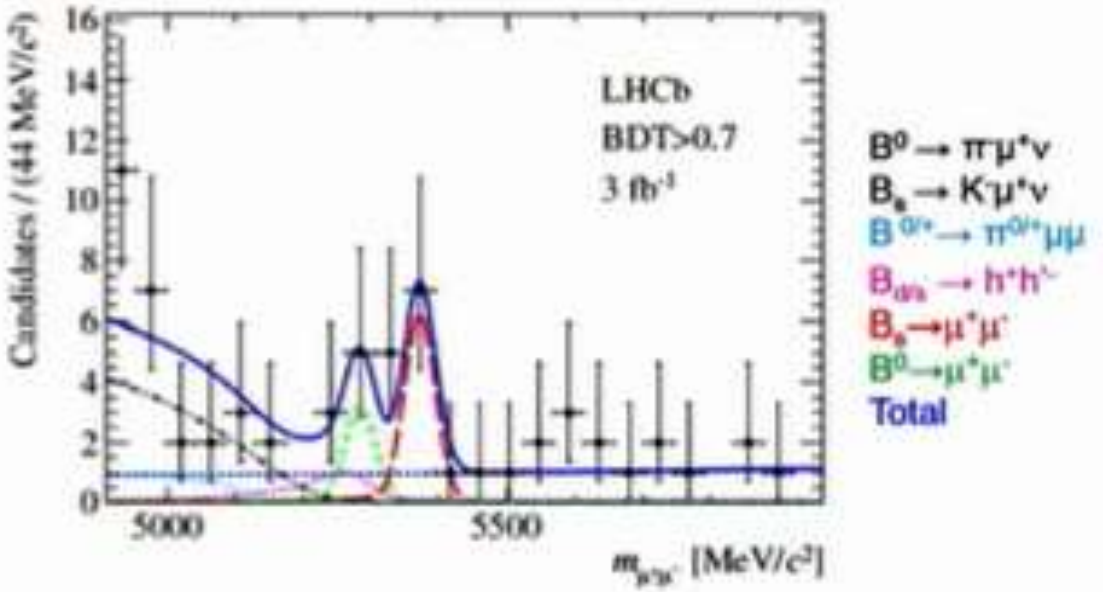


# Аппроксимация сигнального региона



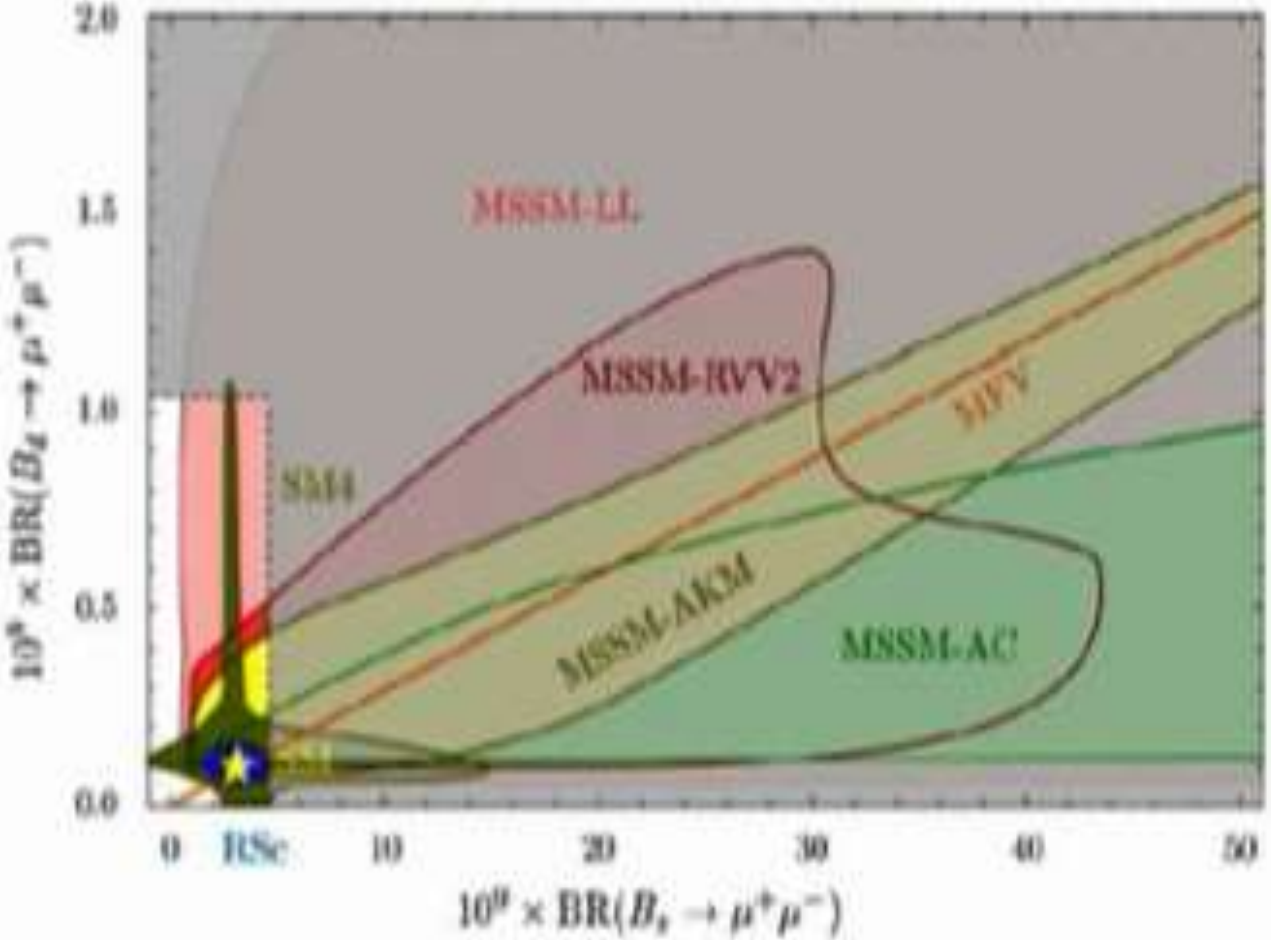
# Подсчет событий

подсчет событий для области $G : g(x) > 0.7$

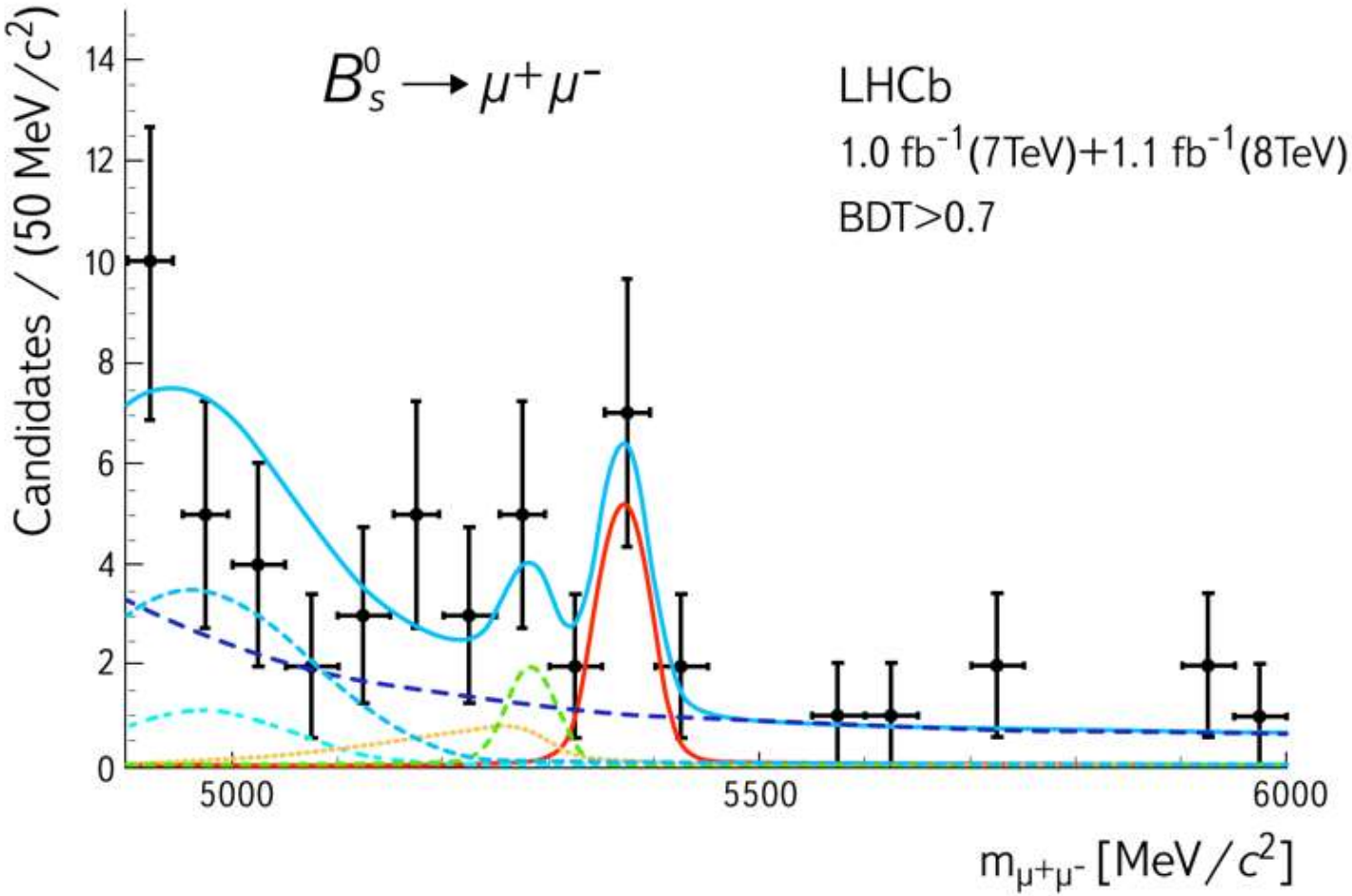



- $BF(B_s \to \mu\mu) = (2.9^{+1.4}_{-1.1})10^{-9}$ $(4\sigma)$
- $BF(B^0 \to \mu\mu) = (3.7^{+2.4}_{-2.1})10^{-10}$ $(2\sigma)$
- Предсказания стандартной модели:
  - $BF(B_s \to \mu\mu) = (3.25 \pm 0.17)10^{-9}$
  - $BF(B^0 \to \mu\mu) = (1.07 \pm 0.10)10^{-10}$

# Влияние на *новую физику* (2012)



Спасибо за внимание!


$B_s^0 \to \mu^+\mu^-$
LHCb
1.0 fb$^{-1}$(7TeV)+1.1 fb$^{-1}$(8TeV)
BDT>0.7
Candidates / (50 MeV/$c^2$)
0
2
4
6
8
10
12
14
5000
5500
6000
$m_{\mu^+\mu^-}$ [MeV/$c^2$]