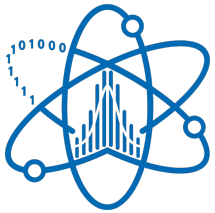
National Research
Tomsk
State
University



**Лаборатория
анализа данных
физики высоких энергий**

Томского
государственного
университета

Физический анализ данных

Томский Государственный Университет

Мария Диденко от лица группы анализа данных

18.12.2025

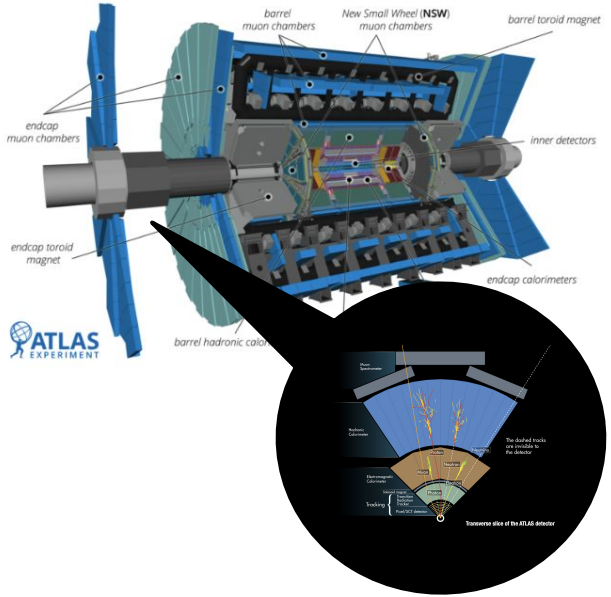
Анализ данных физики высоких энергий

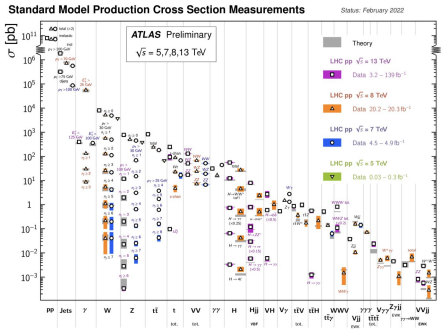
Что изучаем: фундаментальные частицы и их взаимодействия при экстремальных энергиях.

Роль анализа данных: выделение редких физических сигналов, измерение параметров Стандартной модели и проверка теоретических предсказаний.

Данные:

- **ATLAS Open Data 2015–2016 (**формат DAODPhysLite***)**, **36 fb^{-1}**
- Реальные данные, симуляции основных физических процессов и систематические вариации (электрослабые бозоны, Хиггс, QCD-джеты, топ-кварки и тд)
- Масштаб: **~9 млрд событий, ~100 ТБ данных**





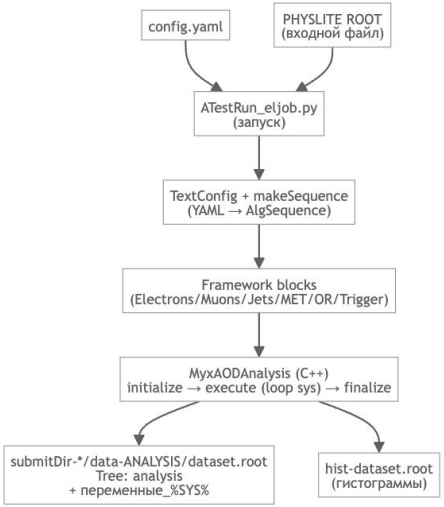
Подход и фокус работы:
Методы машинного обучения, статистический анализ и распределённые вычисления для воспроизведения результатов ATLAS и тестирования новых методов анализа на открытых данных CERN, то что можно использовать и для **будущих экспериментов таких как NICA**

* **DAOD** (Derived Analysis Object Data) — это «производные» данные, получаемые из исходных форматов ATLAS после реконструкции событий.
***PhysLite** — «облегчённая» версия DAOD (стандартный стартовый формат).

Инфраструктура и и воспроизведение анализа

Создание и валидация вычислительной среды:

- Развернута и протестирована вычислительная среда анализа данных на кластере ТГУ.
- Обеспечен **доступ к инфраструктуре ATLAS** с вычислительных узлов кластера ТГУ.
- Настроены распределённые вычисления на базе **HTCondor.**
- Реализован пайплайн обработки **DAOD PhysLight → ROOT ntuples** (фреймворк **Athena AnalysisBase v25.2.45**)

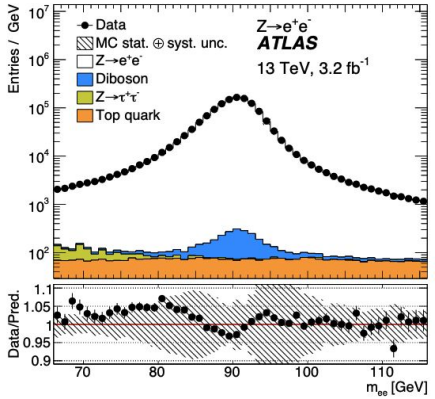


Воспроизведение физического анализа:

- **На финальной стадии** воспроизведение анализа - *“Measurements of top-quark pair to Z-boson cross-section ratios”*.

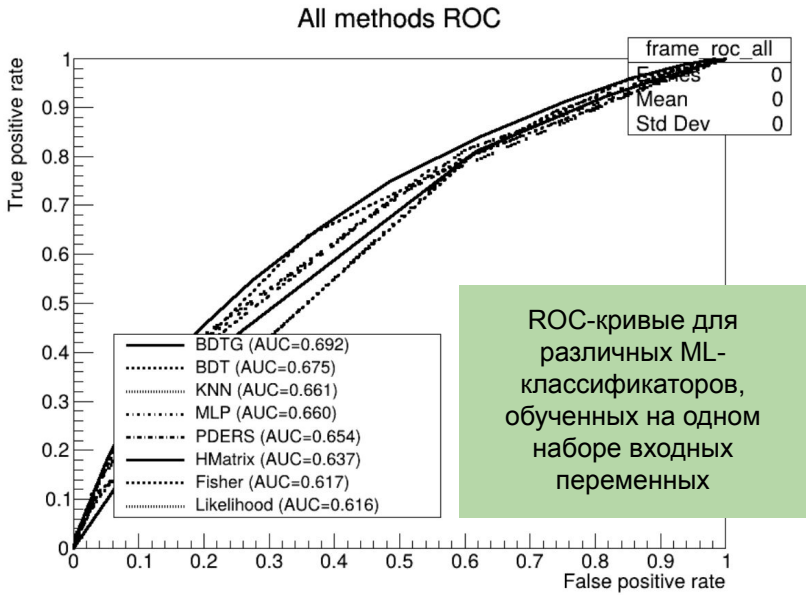
- Анализ используется как **тест корректности** программного обеспечения, сервисов и производительности вычислительного кластера.

- Выполнен полный статистический анализ на базе **TRExFitter**, реализующий **profile likelihood fit** с учетом вклада сигнала и фонов



Сравнение ML-алгоритмов

- Классификация сигнал / фон - **ключевая задача анализов БАК**
- Традиционно используются **BDT и NN**
- Альтернативные ML-подходы **систематически не сравниваются**
- **Цель:** оценить эффективность современных алгоритмов ML для задач классификации в физике частиц



ROC-кривые для различных ML-классификаторов, обученных на одном наборе входных переменных

Данные и физический процесс

- ATLAS Open Data (2015–2016) и HEPData
- Одиночное рождение топ-кварка + W-бозон (**ATLAS tW**)

Реализованный подход и результаты

- Воспроизведены ключевые результаты анализа **ATLAS tW** → **референсная точка**
- Проведено **сравнение** различных **ML-классификаторов** (*BDT, BDTG, MLP, SVM, PCA*) в рамках единого анализа
- Получено согласие результатов между разными программными средами **TMVA/Python/R** → **воспроизводимость**
- **Методы на основе деревьев решений (BDT/BDTG)** → **эффективность** использования входных переменных и **устойчивость** при ограниченной статистике
- **Нейросетевые (MLP) и вероятностные методы (SVM, Naive Bayes)** → нет значительного выигрыша по качеству классификации для данного набора переменных и доступной статистики

Better Bootstrap Confidence Intervals

$$\mu_{BC}[\alpha] = \hat{G}^{-1}\left(\Phi\left(2z_0 + z^{(\alpha)}\right)\right)$$

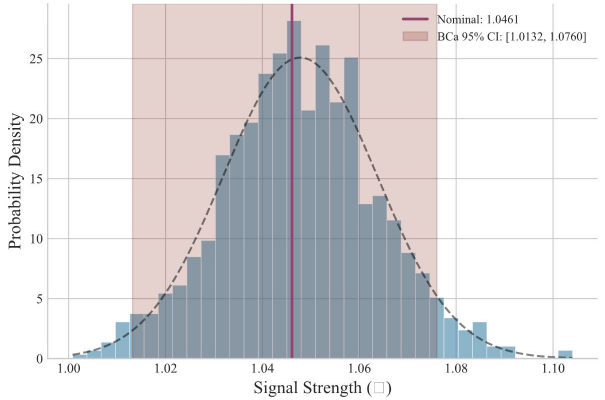
- Неопределённости в HEP → **не всегда корректно описываются асимптотикой**
- Параметр POI (μ) → **смещение и асимметричное распределение**

BCa bootstrap:

- коррекция смещений и асимметрии
- устойчивые доверительные интервалы

Подход и результаты:

- **Анализ WVZ:** первичная валидация bootstrap → корректное распределение μ
- **Метод на открытых данных HEPData -** *"Measurement of t-channel production of single top quarks and antiquarks"*
  - **1000 bootstrap-реплик** (Poisson), 100% сходимость
  - Реализация в **pyhf**, кросс-проверка с **TRExFitter**
  - **Доверительные интервалы (μ):** построены **BC/BCa** интервалы, **BCa** (коррекция смещения и асимметрии) → устойчивые CI при отсутствии асимптотических ошибок



Bootstrap-распределение силы сигнала μ с BC 95% CI

```
Antitop quark (t̄q):
  z0 = -0.0954, a = 0.0112
  Standard    : [1.0166, 1.0790]  (width=0.0624)
  Percentile  : [1.0153, 1.0797]  (width=0.0644)
  BC          : [1.0132, 1.0760]  (width=0.0628)
```

Будущие планы

Машинное обучение

- Переход к более современным ML-методам в средах ***R и Python*** *(DNN, ensemble methods, transformers)*
- Сравнение с BDT/BDTG при увеличенной статистике
- Анализ устойчивости и переобучения ML-классификаторов

Статистика и bootstrap

- Расширенная валидация **BCa bootstrap** *(coverage, closure tests)*
- Применение bootstrap к параметрам интереса и систематикам
- Сравнение с асимптотическими подходами

Методологическое развитие

- Объединение **ML и bootstrap** для оценки устойчивости физических измерений
- Подготовка воспроизводимого workflow *(HEPData → ML → статистика)*
- Формирование рекомендаций по применению **bootstrap / BCa** в HEP-анализах

Спасибо за внимание!