Научно-учебная лаборатория методов анализа больших данных (Lambda)

Фёдор Ратников

Учёный совет
факультета компьютерных наук
27 апреля 2017 года

LAMBDA

Состав лаборатории

- руководитель (Устюжанин А. Е.);
- менеджер (Глазистов А. В.);
- 4 исследователя;
- 5 стажёров;

- Учёные степени:
  - 2 к. ф.-м. н.;
  - 1 PhD.

- 2 аспиранта.
  - 1 в совместной международной аспирантуре

- в этом году планируется написание
  - 4 магистерских работ
  - 1 бакалаврскую работу
  - 2 курсовых работы

Устюжанин Андрей Евгеньевич

Заведующий лабораторией

Глазистов Артём Владимирович

Менеджер

Арзыматов Кененбек

Стажер-исследователь

Борисяк Максим Александрович

Стажер-исследователь

Деркач Денис Александрович

Старший научный сотрудник

Казеев Никита Александрович

Стажер-исследователь

Ратников Федор Дмитриевич

Старший научный сотрудник

Умнов Алексей Витальевич

Младший научный сотрудник

Шахуро Владислав Игоревич

Стажер-исследователь

Широбоков Сергей Константинович

Стажер-исследователь

Научная работа

Наша научная деятельность условно разделена несколько категорий:

- разработка методов машинного обучения для решения практических задач
  - и их применение к соответствующим задач
    - преимущественно сложные задачи в естественных науках

- обобщение методов

- публикация этих методов в виде программных продуктов.

Прикладные исследования в Европейском центре ядерных исследований (CERN)

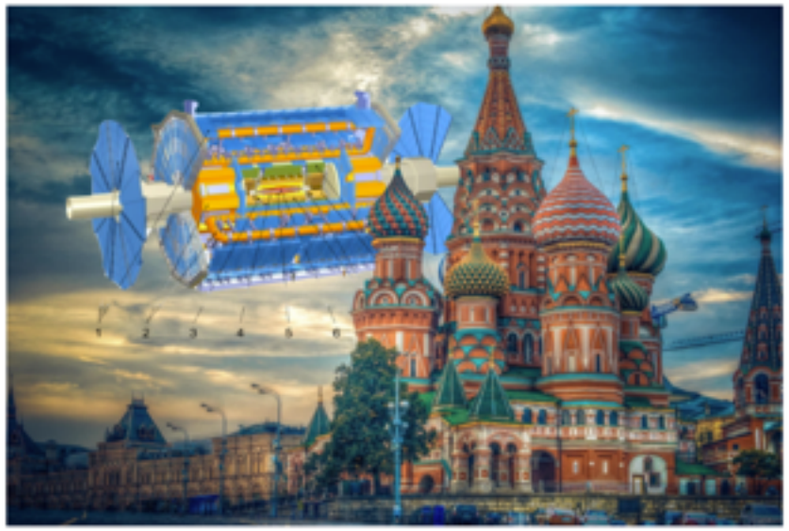
Примеры компьютерных задач:

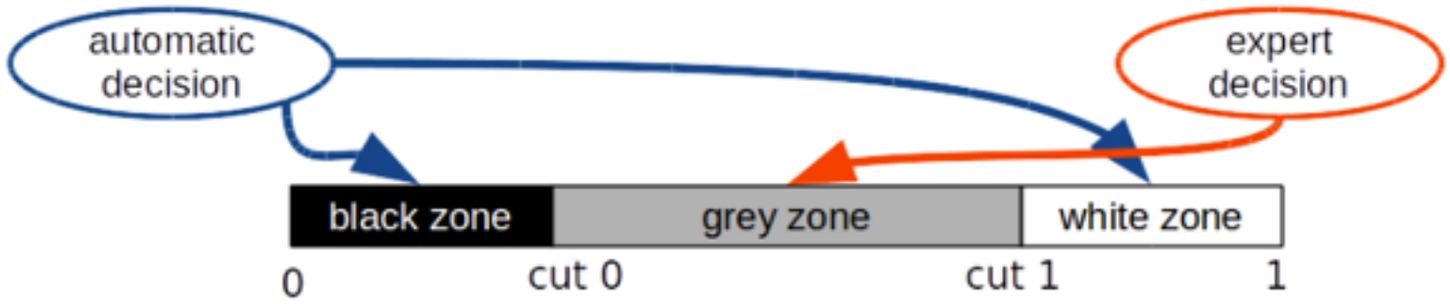
- Автоматизированный поиск аномалий в работе детектора (совместно с Церном):
  Построена принципиальная схема, позволяющая в онлайн режиме предсказывать аномалии с помощью простых классификаторов. Алгоритм протестирован, доложен на CHEP, просидинги приняты, идёт выкатывание для реальной работы.

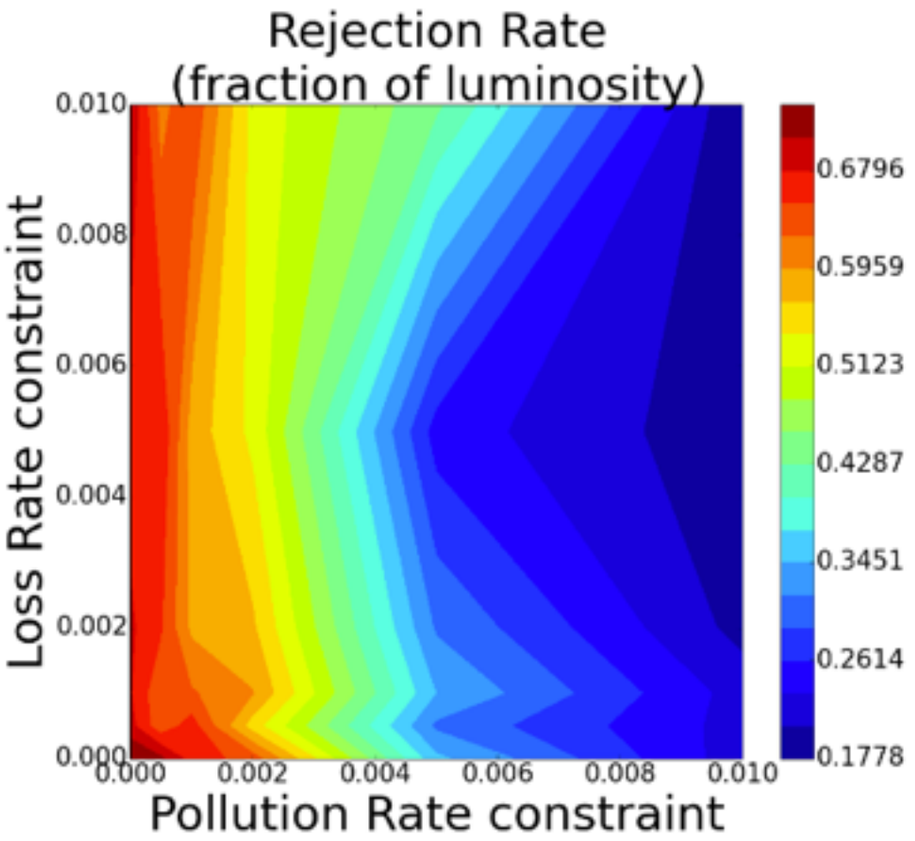
- Предсказание популярности данных с помощью методов машинного обучения (совместно с Церном).
  Алгоритм запущен в виде сервиса в Церне. Сделано несколько докладов на конференциях (KDIR, CHEP, Московский суперкомпьютерный форум, МФТИ), опубликованы просидинги. На данный момент ведётся работа над патентом.

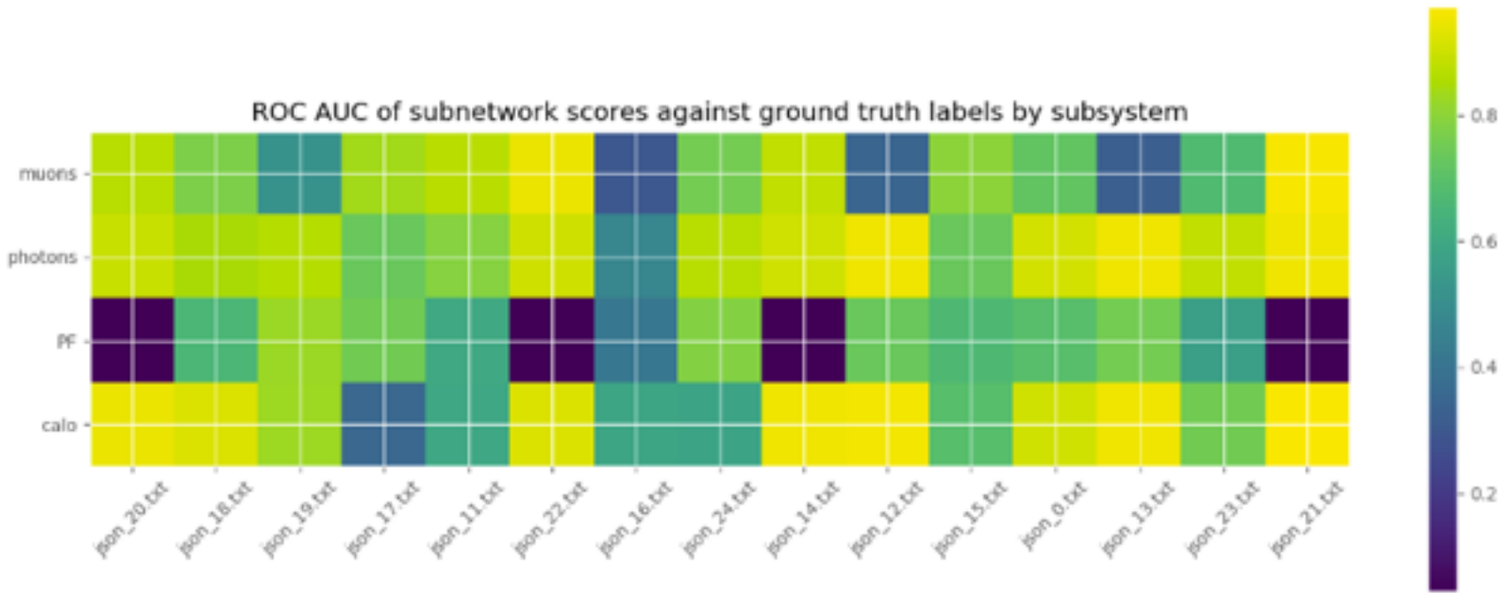
- Оптимизация деления поточных данных (с университетом Оксфорда, Yandex Data Factory и Церном).
  Алгоритм проходит тестирование в Церне. Продолжается обсуждение возможность установки его на онлайн системы набора данных. Принцип алгоритма доложен на ACAT, просидинги опубликованы.

Поиск аномалий в данных детектора









Ф.Д. Ратников

LAMBDA

Прикладные исследования в Европейском центре ядерных исследований (CERN)

Примеры физических задач:

- Разделение различных типов фотонов на основе первичной информации с использованием ML:
  Метод показал хорошие результаты на Монте-Карло. Сейчас измеряем качество на реальных данных

- Предсказание аромата сигнальных частиц (совместно с MIT):
  Доклад опубликован. Алгоритм проходит тестирование.

- Классификация заряженных частиц по отклику детектора с использованием глубоких нейросетей (с университетом Кэмбриджа, МФТИ)
  Готовится статья.

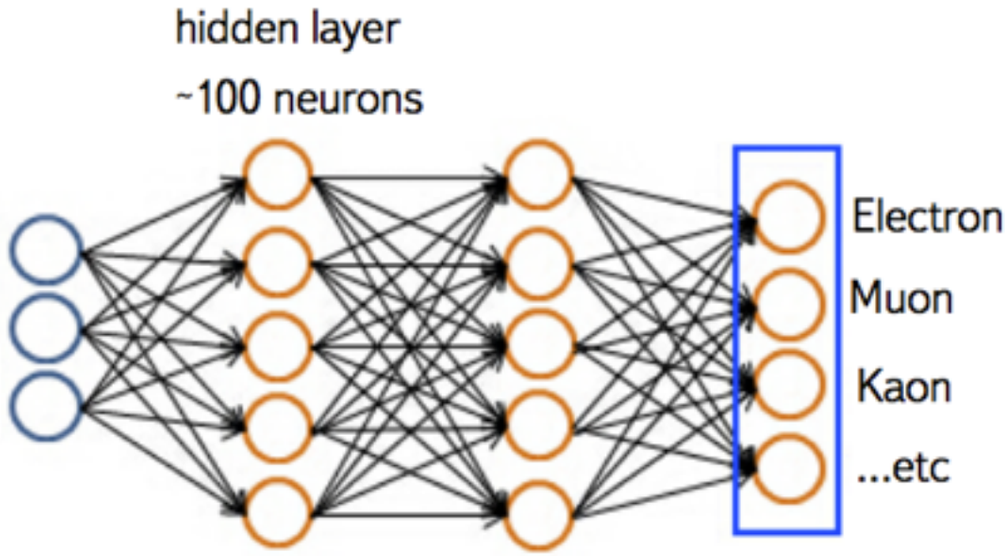
x

Мультиклассовая идентификация частиц в LHCb.

Muon

RICH

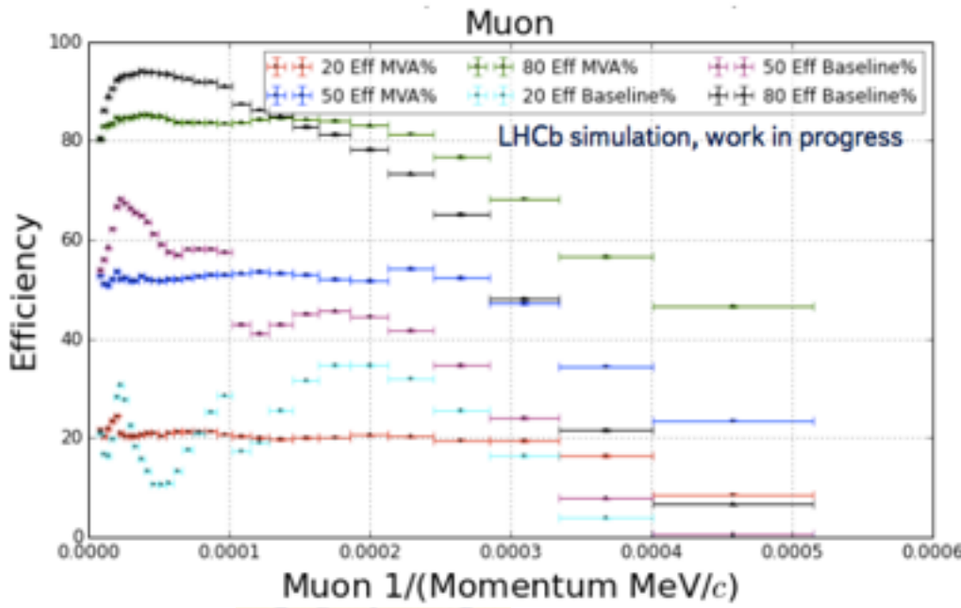
CALO



one vs rest

multiclass

	Ghost	Electron	Muon	Pion	Kaon	Proton
baseline	0.9484	0.9854	0.9844	0.9345	0.9147	0.9178
keras DL	0.9632	0.9914	0.9925	0.9587	0.9319	0.9320



LAMBDA

Другие приложения

- Астрофизика CrayFis (совместно с CalTech и NYU).
  Уникальный CrowdScience эксперимент, сотрудники лаборатории участвуют в разработке алгоритмов выделения сигнала. Проведено тестирование телефонов. Построены первые алгоритмы машинного обучения для поиска заряженных частиц.

- Теоретическая физика (совместно с несколькими европейскими и американским университетом).
  Участие в феноменологической обработе экспериментальных и теоретических данных с помощью байесовской статистики. Оптимизация методов численного интегрирования с помощью Марковских цепей. Работа продолжается.

- Анализ сигналов эпилептической мозговой активности в магнитоэнцефалограммах
  Работа начата, идёт тестирование моделей

Обобщение опыта и разработка программных продуктов

Перевзвешивание распределений с использованием метода градиентного спуска.

Метод опубликован в просидингах конференции

Алгоритм для запуска распределённых вычислений для облачных и GRID технологий (SkyGrid).

Практически используется для распределённой обработки и генерации данных

Разработки среды для проведения совместных вычислительных экспериментов.

Постоянная разработка.

REP: https://github.com/yandex/rep

Everware: https://github.com/everware

Учебная работа внутри факультета

Максим Борисяк:
- Machine Learning and Data Mining (Магистратура; 2-й курс);

Умнов Алексей:
- Машинное обучение 1 (Бакалавриат);
- Машинное обучение на больших данных (Бакалавриат);
- Современные методы машинного обучения (Минор);
- Программирование (Бакалавриат).
- Алгоритмы и структуры данных (Бакалавриат)

Фёдор Ратников совместно с Александром Паниным
- Глубокое обучение (Магистратура, 2-й курс)

Учебная работа вне факультета

- Вторая летняя школа Machine Learning in High-Energy Physics под эгидой факультета прошла в Лунде, Швеция в июне 2016 года.

  Третья школа в июле 2017 в Ридинге (Великобритания)

  Для студентов факультета предусмотрены гранты на поездку на школу

- Совместно с университетом Манчестера, Падуи, Рио-де-Жанейро и Черном подготовлена первая лабораторная работа по анализу физических данных Большого адронного коллайдера для студентов-физиков (будет включена в курс некоторых университетов). Данные открыты с декабря 2015 года согласно политике открытых данных.
  Сделан доклад на конференции CHEP

- Один из сотрудников лаборатории является координатором группы Машинного обучения и статистики в эксперименте LHCb. Цель группы — выработать стандарты проведения современного анализа данных и контроль за соблюдением этих стандартов.

Конференции и публикации

С июня прошлого года сотрудники лаборатории сделали 9 докладов на конференциях и 10 докладов на различных воркшопах.

Соответствующие статьи были отправлено на публикацию в просидингах конференций.

Вышли 5 публикаций в просидингах предыдущих конференций

Внешнее финансирование

- Обсуждение проекта работ для Сколково
  - в заключительной стадии
- Обсуждение проекта работ для МИСИС
  - в начальной стадии

Популяризация науки

- Организация серии Data & Science на базе Яндекса
- Проведение международных мастер-классов для школьников
- Участие в ТВ программе “Великое в малом”
- Доклад и телемост на зимней школе в Вороново
- Выступление на “Дне Вышки”

Всё вместе

Наименование показателя	Ед. измерения	2015	2016	2017
Статьи на английском языке, индексируемые Web of Science и/или Scopus	ед.	3	5	5
Доходы от выполнения прикладных научно-исследовательских, консультационных и экспертных работ	млн. руб.			[5+5]
Прием стажеров	чел.	1	2	3
Чтение учебных курсов по тематике Центра	ед.		2	3