

Моделирование и обработка данных для экспериментов по дифракции лазерного излучения на отдельных макромолекулах

А. Б. Теслюк, С. А. Бобков, В. А. Ильин

Научный семинар
«Методы суперкомпьютерного моделирования»

2015 Таруса

Об экспериментах XFEL

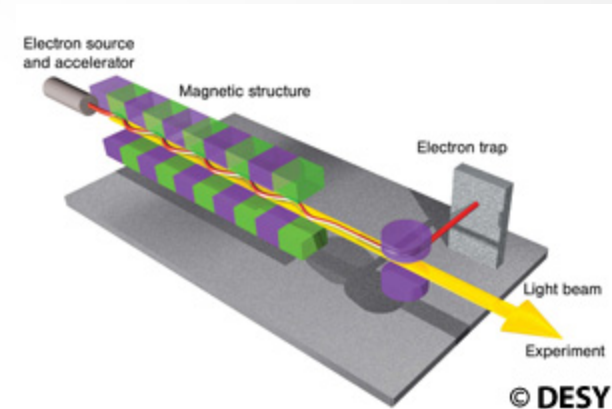
- Новое поколение лазеров на свободных электронах – новые возможности для исследований:

- структура
- динамика
- процессы

- Дифракция на отдельных молекулах

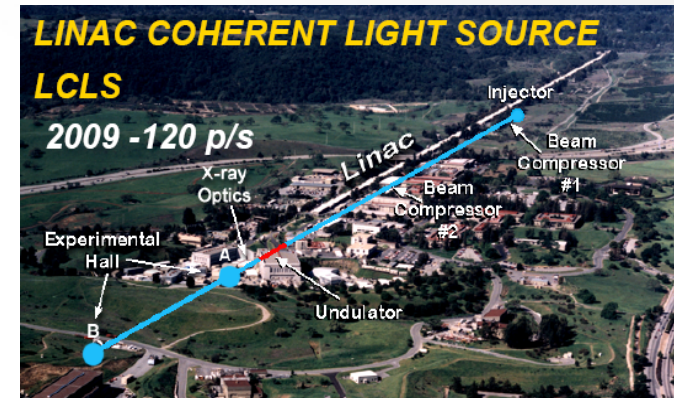
- Мега-установки:

- LCLS (Stanford), 2009
- XFEL/Spring-8 (Japan), 2011
- Европейский XFEL, Гамбург, 2016



Мега-установки

	<i>LCLS (USA)</i>	<i>SCSS (JAPAN)</i>	<i>EUROPEAN XFEL (SASE1)</i>
<i>Max. Electron Energy (GeV)</i>	14.3	8.0	17.5
<i>Minimum Wavelength(nm)</i>	0.15	0.13	0.10
<i>Peak Brilliance</i>	$1.5 \cdot 10^{33}$	$1 \cdot 10^{33}$	$5 \cdot 10^{33}$
<i>Average Brilliance</i>	$4.5 \cdot 10^{22}$	$1.5 \cdot 10^{23}$	$1.6 \cdot 10^{25}$
<i>Pulses/s</i>	120	3000	30 000
<i>Photons/pulse</i>	10^{12}	$2 \cdot 10^{11}$	10^{12}
<i>First Beam</i>	2009	2011	2014



XFEL в РФ

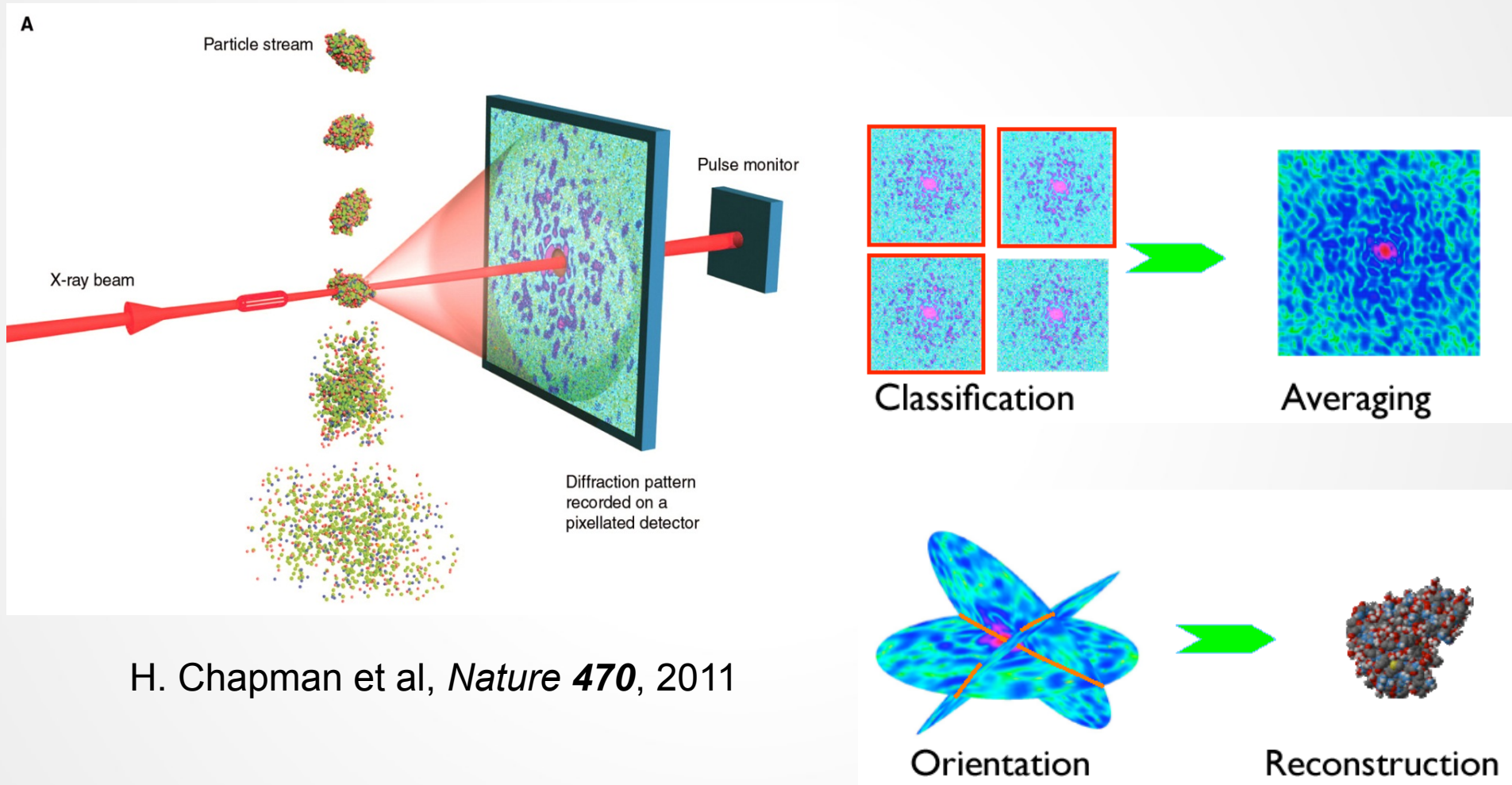
Основные участники проекта XFEL:

- Германия (50%)
- Россия (27%)

Направления исследований в России:

- Выбор подходящих объектов для исследований. Разработка методов выделения и подготовки образцов
- Разработка элементов экспериментальной установки
- Математические методы обработки результатов и ИТ-инфраструктура

Эксперимент



H. Chapman et al, *Nature* **470**, 2011

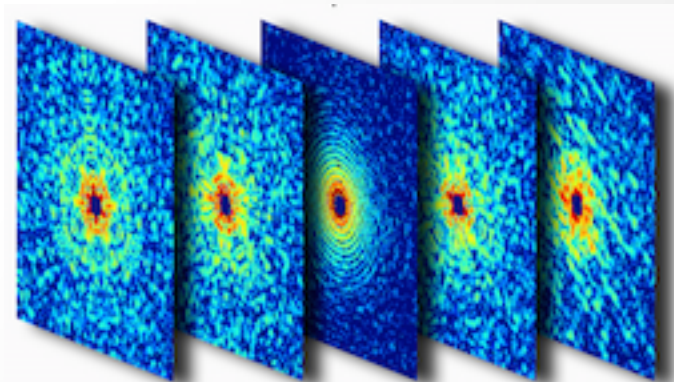
Проблемы

- ИТ-инфраструктура

- 2.3 млрд. изображений в сутки
- Big data: хранения, передача, индексирование, публикация

- Новые методы анализа

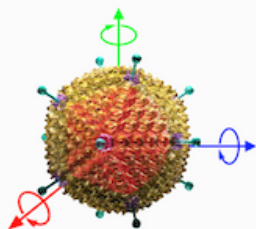
- изображения невоспроизводимы
- пространственная ориентация случайна
- внутренняя динамика частиц



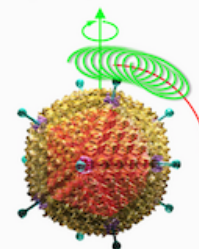
Моделирование

- Дефицит ПО
- Дефицит экспериментальных данных
- Доработка пакета SPSIM

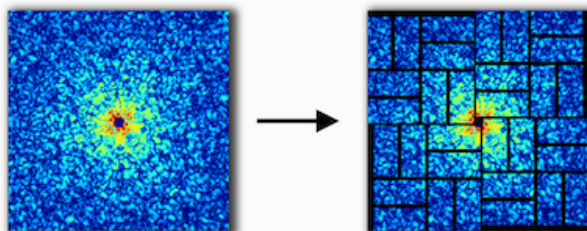
Случайная ориентация частиц



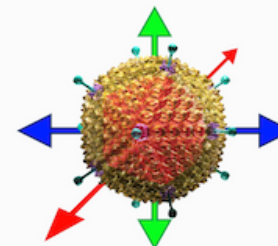
Равномерное вращение



Структура детектора



Случайное положение частицы



SPSIM + GPU

- Технология CUDA
- Каждый пиксель изображения – отдельный поток GPU
- Ускорение в ~200 раз

Профайлер

compute_pattern_on_list (Расчет дифракции)
98.21% 
calculate_pixel_solid_angle (Размер пикселя)
0.73% 
get_HKL_list_for_detector (Положение пикселей)
0.33% 
calculate_thomson_correction (Учет поляризации)
0.13% 
write_2D_array_to_vtk (Вывод результата)
0.13% 
...

Расчет 1 изображения (1000x1000)

CPU (1 поток, Intel Q8300)

4м 53с  100%

GPU (GeForce GTX 460)

0м 7с  2.5%

Расчет 10 изображений (1000x1000)

CPU (1 поток, Intel Q8300)

48м 11с  100%

GPU (GeForce GTX 460)

1м 9с  2.4%

Классификация изображений

- Новые метод для классификации дифракционных картинок
- Задачи:
 - Фильтрация «плохих» картинок
 - Отбор «хороших» картинок
 - Классификация дифракционных картин для разных молекул
 - Индексирование и база данных

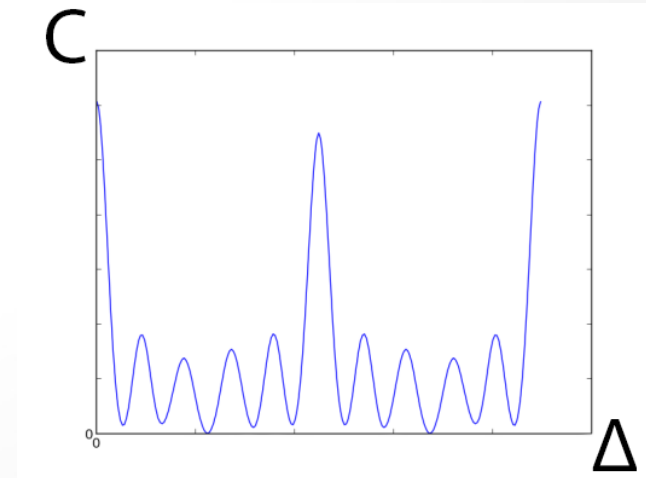
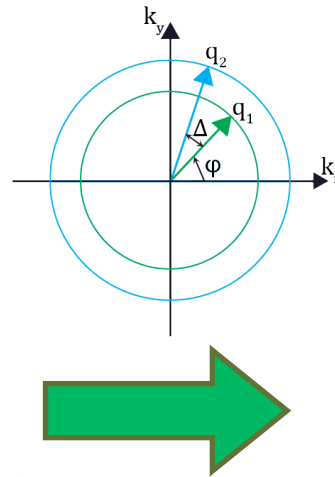
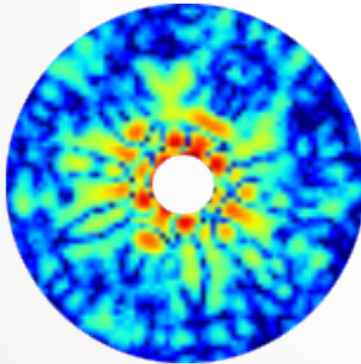
Идея метода

- Выбор небольшого количества признаков, определяемых структурой объекта – характеристический вектор
- Применение алгоритмов компьютерного зрения для классификации характеристических векторов

Характеристический вектор

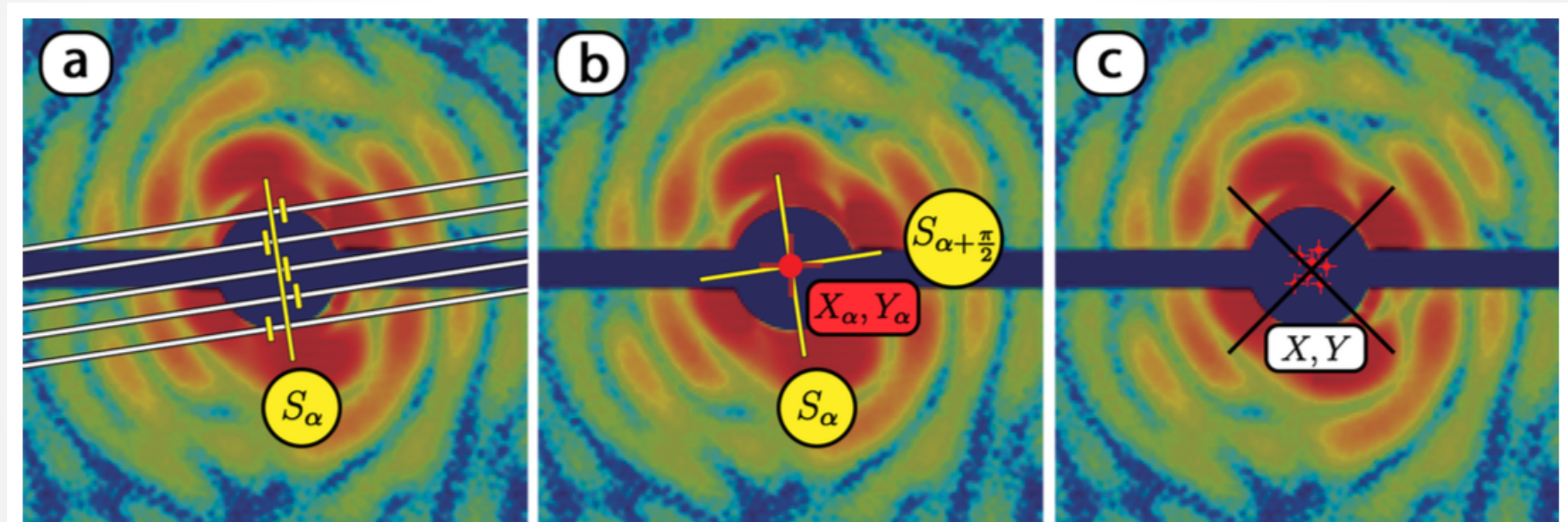
Кросс-корреляционная функция

- $C(q_1, q_2, \Delta) = \langle I(q_1, \varphi) I(q_2, \varphi + \Delta) \rangle_\varphi$
- Autocorrelation, $q_1 = q_2$



Особенности метода

- Точное определение центра дифракционной картины

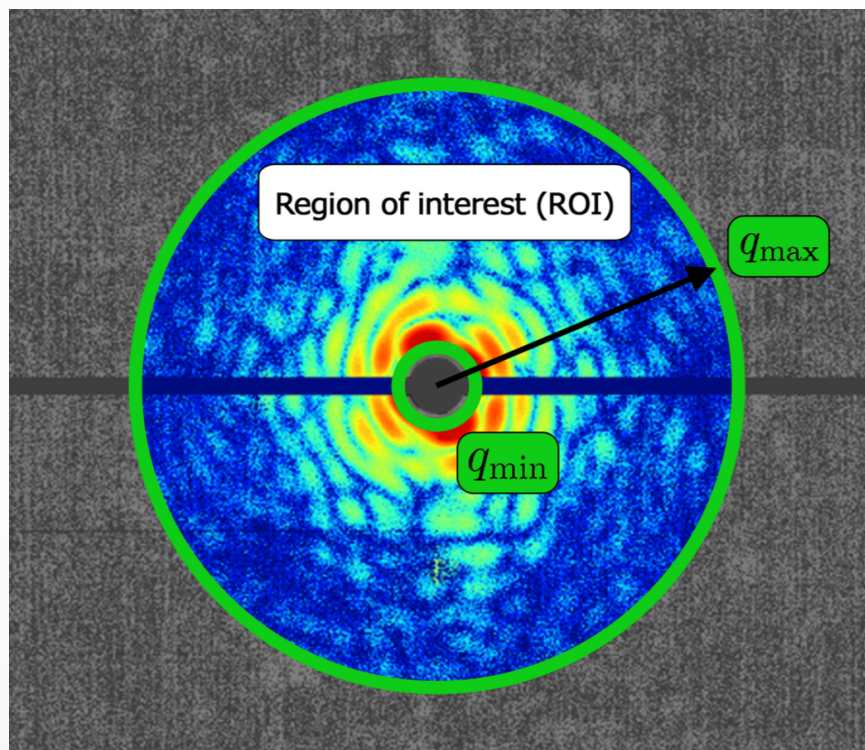


$$D_{i,\alpha}(x) = \frac{\int_0^N |I_{i,\alpha}(x+l) + I_{i,\alpha}(x-l)|^2 dl}{\int_0^N |I_{i,\alpha}(x+l) - I_{i,\alpha}(x-l)|^2 dl}$$

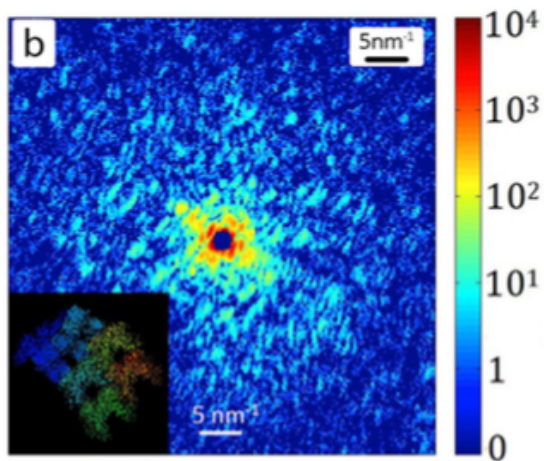
$$\bar{x}_{i,\alpha} = \operatorname{argmax}_x [D_i(x)],$$

Особенности метода

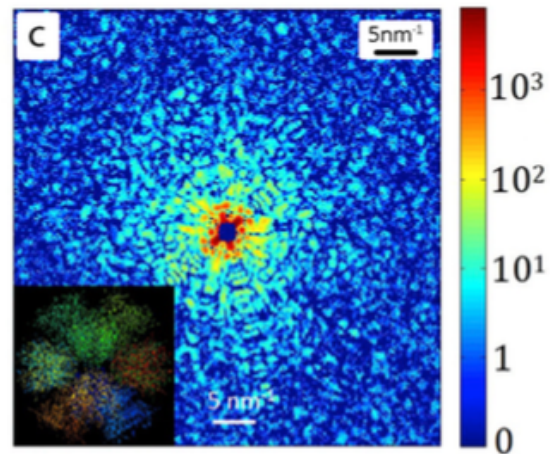
- Определение зоны «сигнала»



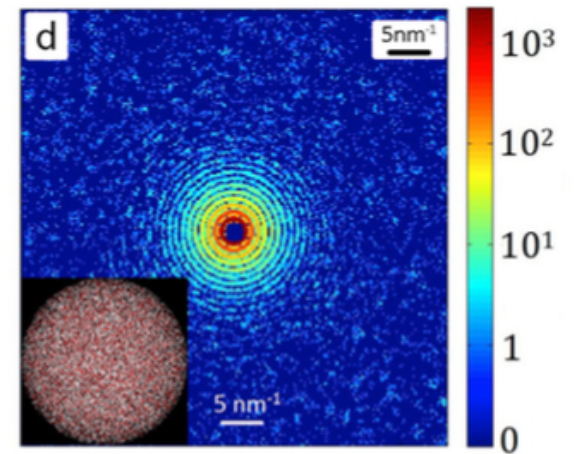
Модельные данные



Bluetongue virus core



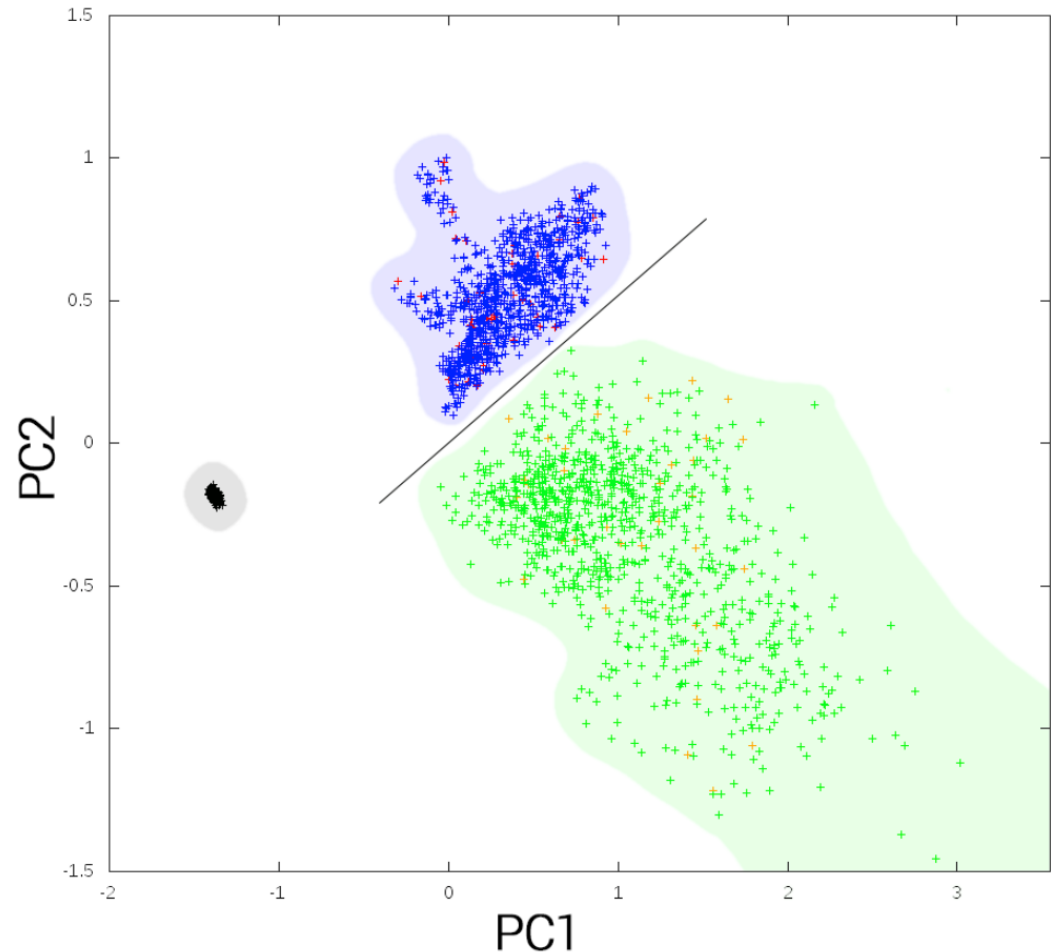
Adenovirus capsid



Water

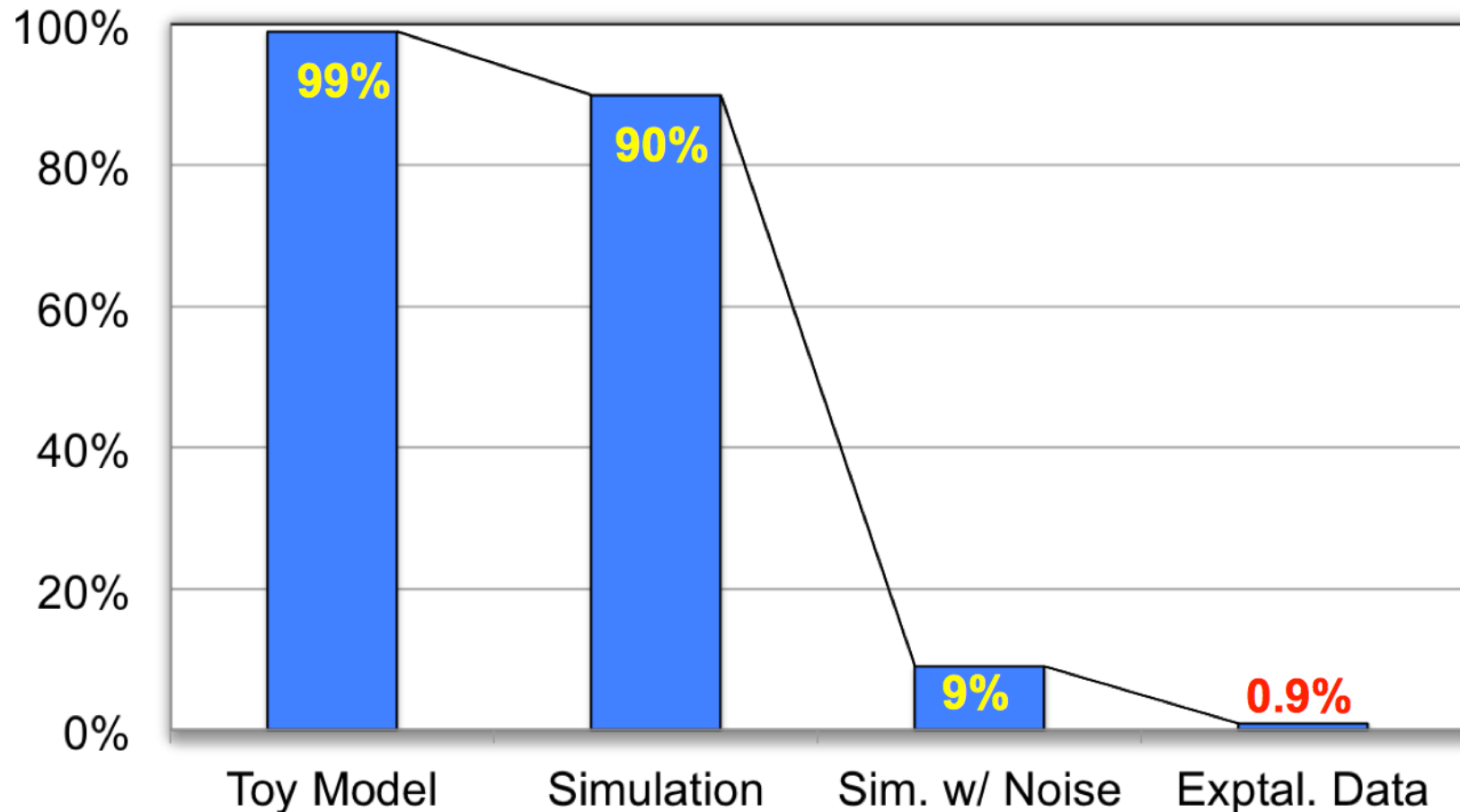
Метод главных компонент

- 100% точность разделения
- В базисе PC1-PC2 ~30% вариабельности данных



Экспериментальные данные

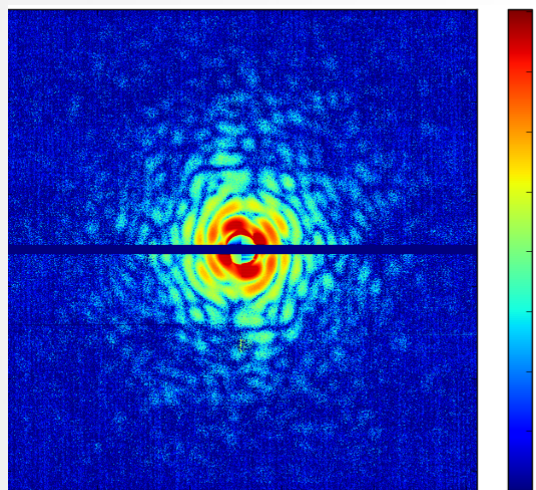
Fraction of Successful Theories



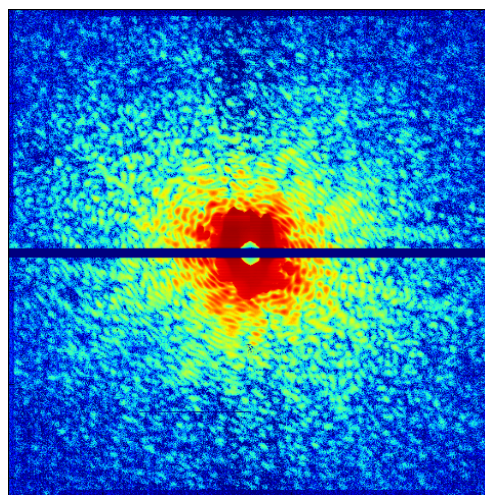
Abbas Ourmazd

Экспериментальные данные

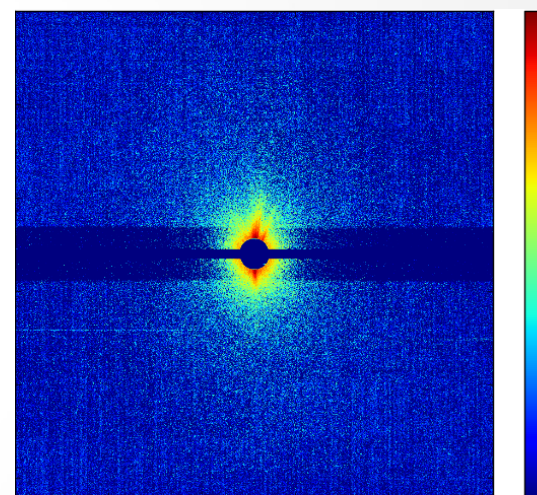
Набор данных с эксперимента LCLS (Stanford), два типа молекул



Бактериофаг Т4

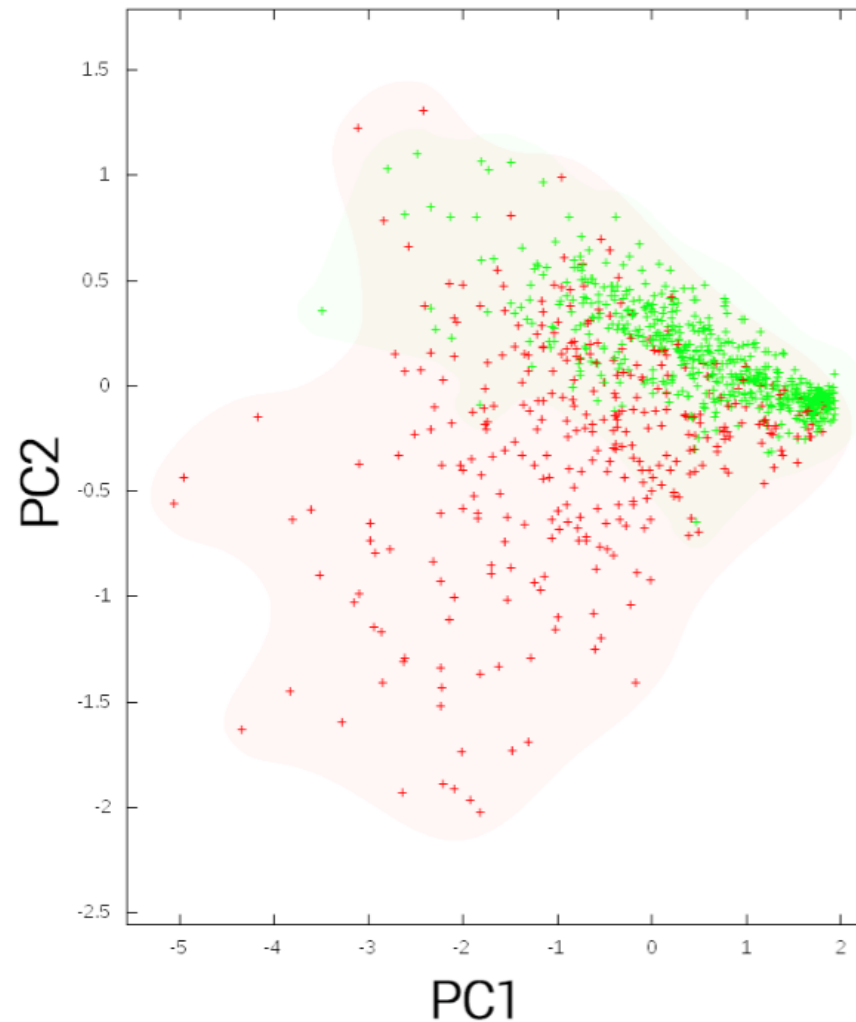


Paramecium
bursarium Chlorella
virus

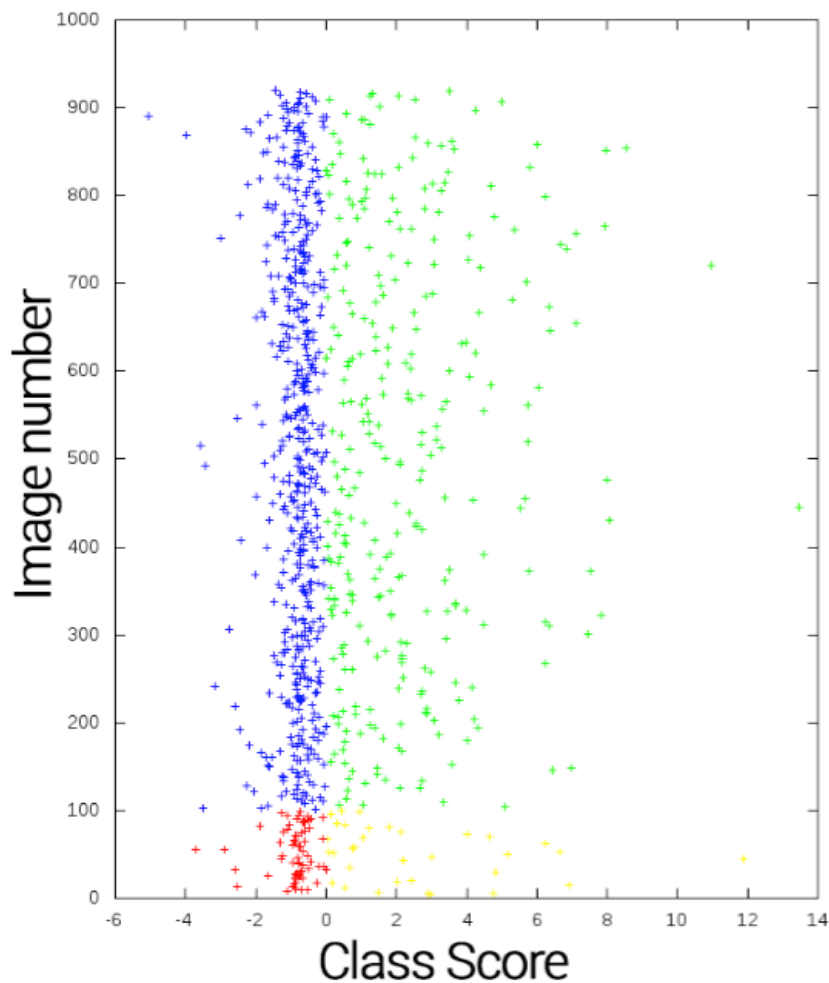


Пустая картинка

Экспериментальные данные - PCA



Метод опорных векторов



- Точность классификации ~86.4%
- 100% фильтрация «пустых» изображений

База данных и индексирование

- Координаты в базисе главных компонент – индексы изображения в базе данных
- Поиск «похожих» изображений – SQL запрос в базу данных
- Реализация PostgreSQL + Python



ИТ-реализация

- Основной код Python + Numpy + Intel MKL
- OpenMP параллелизация
- Моделирование с помощью GPU + CUDA
- 24 Core server – обработка изображений в real-time
- Курчатовский суперкомпьютер

Основные результаты и планы

- Моделирование эксперимента
- Метод классификации
- Индексируемая база данных изображений

Спасибо за внимание