# Параллельная реализация метода поиска закономерностей в последовательностях событий

Вишневский В.В.

27 мая 2011 г.

# Задача исследования поведения

В поведении существуют закономерности. Повседневные церемонии: ритуалы приветствия, рабочие процессы, груминг у животных, состоят из поведенческих паттернов.

Пример. Принятие пищи: «подойти к столу», «отодвинуть стул», «сесть», «съесть главное блюдо», «съесть десерт», «выпить чай», «отодвинуть стул», «встать», «отойти от стола».

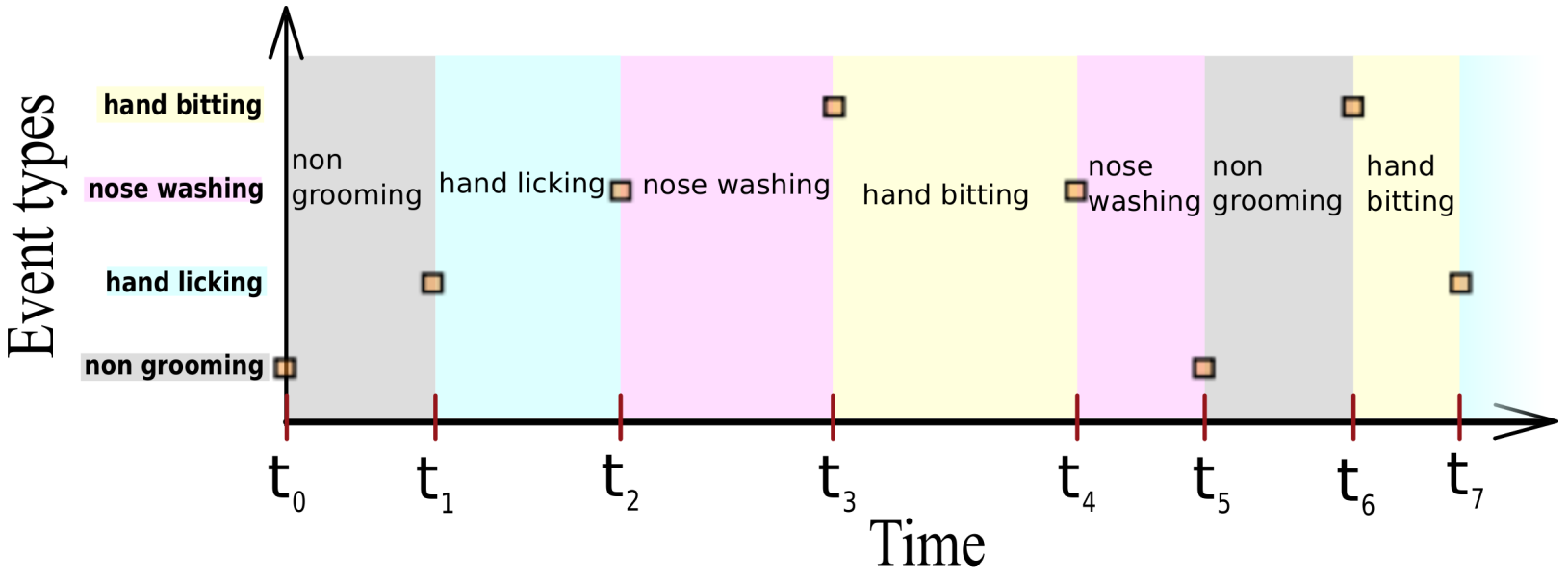
- События связанны временны́ми интервалами.
- Есть иерархия.

# Мотивация

Выделив такие поведенческие паттерны, становится возможно:

- делать выводы о сложности поведения,
- определять изменения в поведении наблюдаемых животных,
- *измерять* поведение — анализировать влияние различных факторов на поведение.

# Входные данные



# Недостатки существующих методов

- Стандартные методы поиска закономерностей не предназначены для поиска *поведенческих* паттернов (важна иерархия, упорядоченность, временны́е интервалы).
- Существуют методы поиска закономерностей, где учитывается только порядок актов.
- Широко используемый метод поиска T-Паттернов **очень** чувствителен к шуму, допускает слабую *вариабельность* паттернов. Закрытые исходные коды.

# Подход к поиску паттернов

Паттерн — это часто встречающаяся последовательность событий(поведенческих актов), возникающих один за другим через определенные промежутки времени.

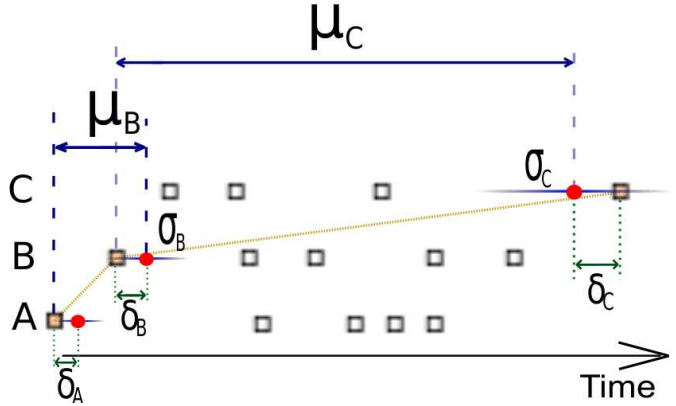
Инициализируем множество паттернов поведенческими актами. Потом итеративно повторяем:

- **Конструирование:**Для всех пар паттернов проверить, повторяется ли один за другим достаточно часто. Если да, то получаем новый паттерн.
- **Редукция:** Удалить одинаковые паттерны, которые были сконструированы по-разному.

# Plan

# Вероятностная модель P-Паттерна

- $P = A[\mu_A, \sigma_A]B[\mu_B, \sigma_B]C[\mu_C, \sigma_C]$




- Функция потерь:

$$f_{Loss}(x, N) = \begin{cases} \exp\left(-\frac{\lambda x}{N}\right), & x < N, \\ 0, & x = N. \end{cases}$$

- Правдоподобие паттерна:

$$L_P(\varepsilon) = f_{Loss}(N_-, N_P) \prod_{i=1}^{N_P} \left(\frac{1}{\sqrt{2\pi}\,\sigma_i}\right) \prod_{i \in \mathcal{N}_+} \exp\left(-\frac{\delta_i^2}{2\sigma_i^2}\right)$$

# Правдоподобие

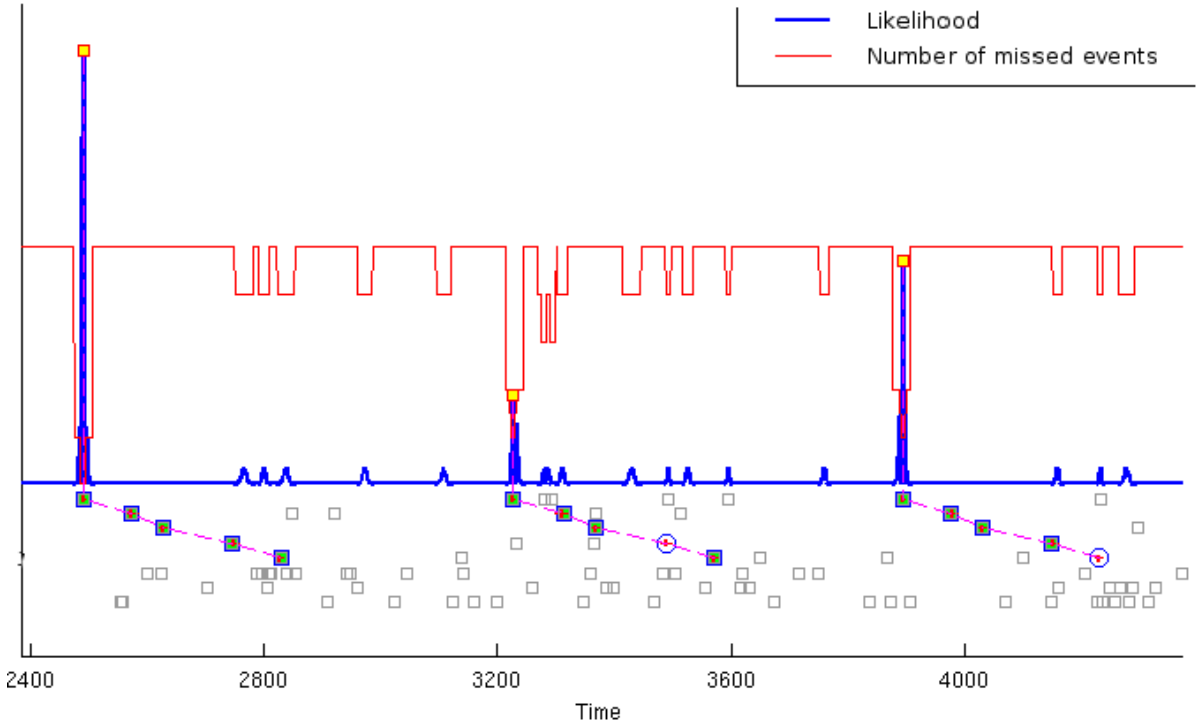



Рис.: Пример функции правдоподобия паттерна. Желтыми маркерами с красной границей изображены максимумы функции правдоподобия: моменты времени, когда мы считаем, что паттерн имеет место. В нижней части рисунка закрашенными квадратами показаны присутствующие события, полыми кружками — пропущенные события в паттерне. Полые серые квадраты соответствуют наблюдаемым поведенческим актам.

# T-Паттерны и P-Паттерны

- T-Паттерны распараллеливаем на SMP с помощью OpenMP. Тестирование на 4-х ядерном CPU. Примерная сложность: $O(n^2)$.
- P-Паттерны распараллеливаем на GPU с помощью CUDA. Тестирование на GF 8800GTX, 128 потоковых процессора. Примерная сложность: $O(n^3)$.

# Ускорение OpenMP

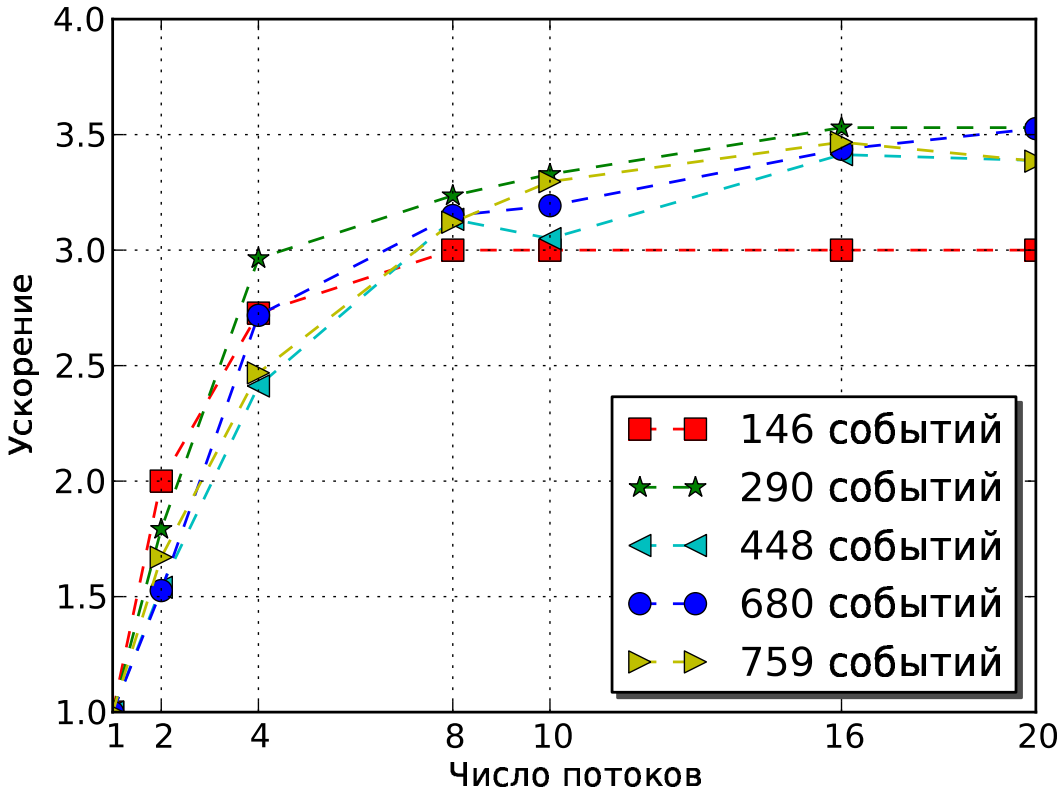



Рис.: Ускорение алгоритма поиска T-Паттернов на 4-х ядерном процессоре.

# Ускорение алгоритма поиска P-Паттернов. CUDA

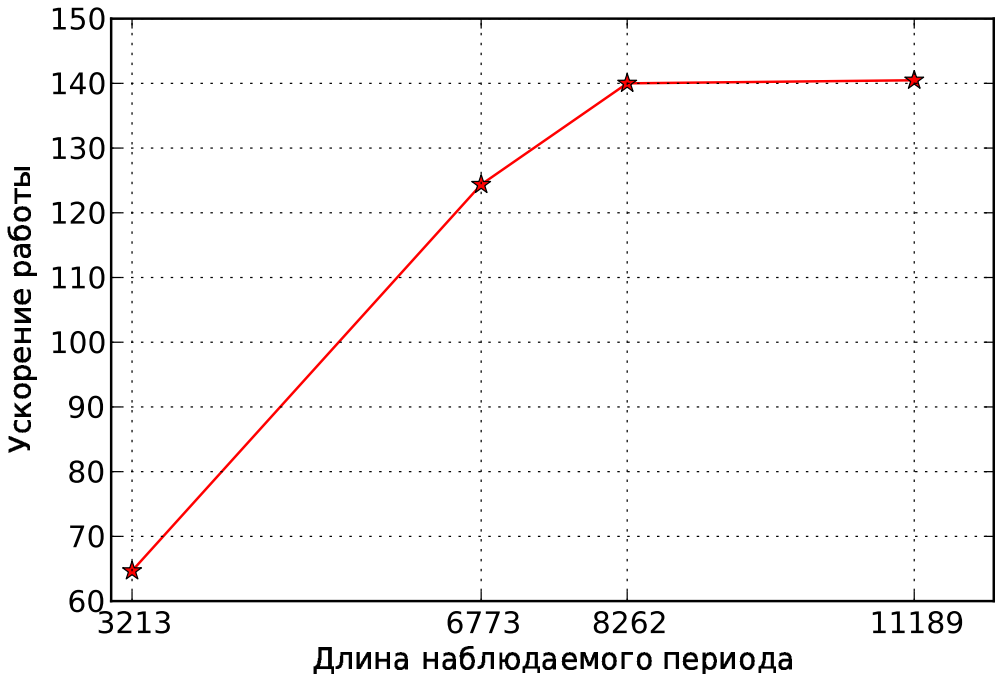



Рис.: Ускорение стадии подсчета правдоподобия паттернов в зависимости от размера входных данных.

# Ускорение алгоритма поиска P-Паттернов. CUDA

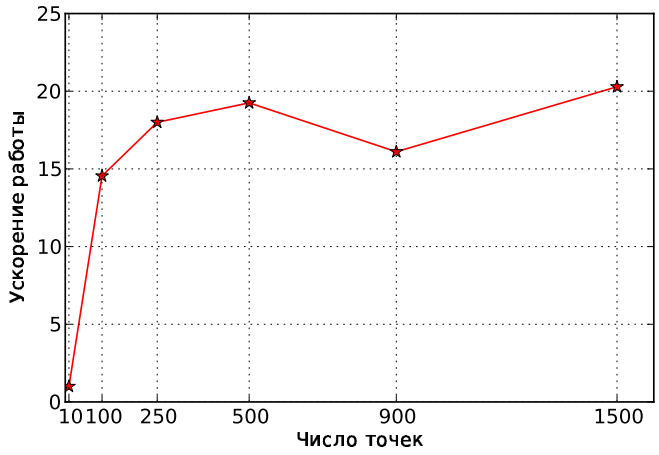
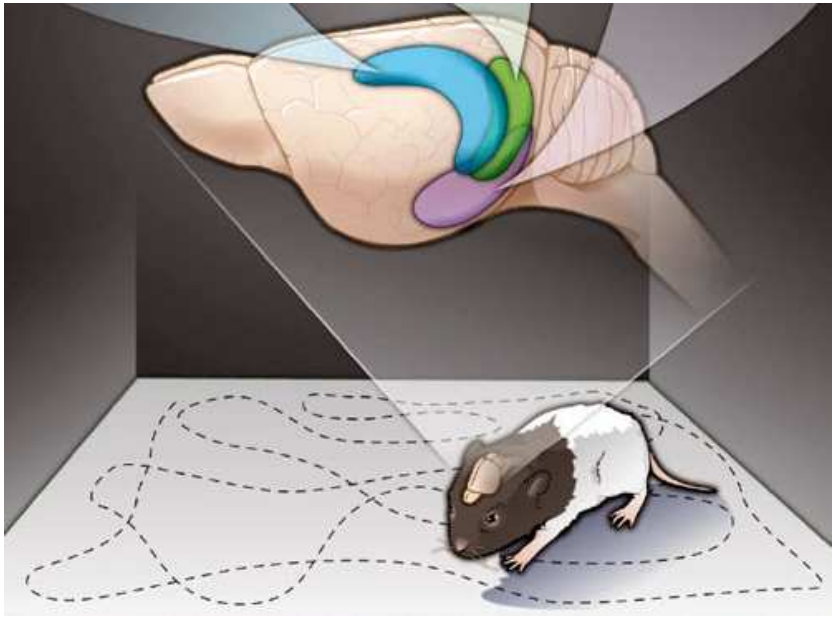



Рис.: Ускорение стадии конструирования паттернов в зависимости от размера входных данных.

Ускорение метода в целом $\sim$ 40 раз.
Типичные экспериментальные данные: 12 секунд на GPU, 470 секунд на CPU(1 поток).
Утилизация GPU $\sim$ 230 GFLOPS (Заявленная производительность 518 GFLOPS)

# Эксперимент с грызунами без гиппокампа

- Гиппокамп — отдел головного мозга. Его функции связывают с механизмами работы памяти, обучением, пространственной навигацией.
- Две группы: контрольная(12 особей) и грызуны без гиппокампа(12 особей).
- Определить по поведению к какой группе относится особь.
- Как меняется поведение после воздействия на определенный участок мозга?



# Результаты экспериментов

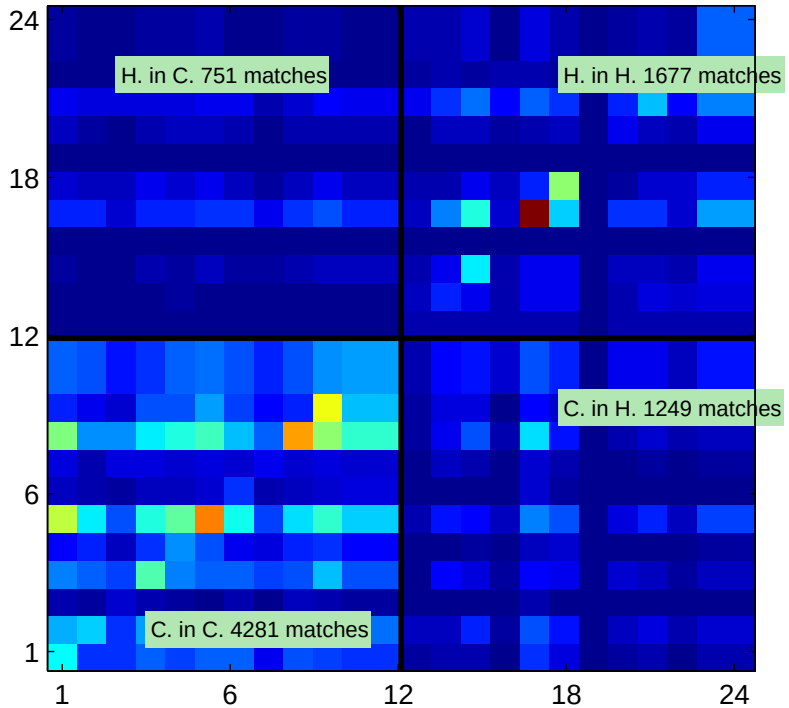



Рис.: Таблица соответствий паттернов. Неформально: по вертикали *откуда* берутся паттерны, по горизонтали — *где* ищутся вхождения этих паттернов; например, в ячейке $(3, 10)$ записано число соответствий паттернов третей особи в поведении десятой.

# Результаты экспериментов

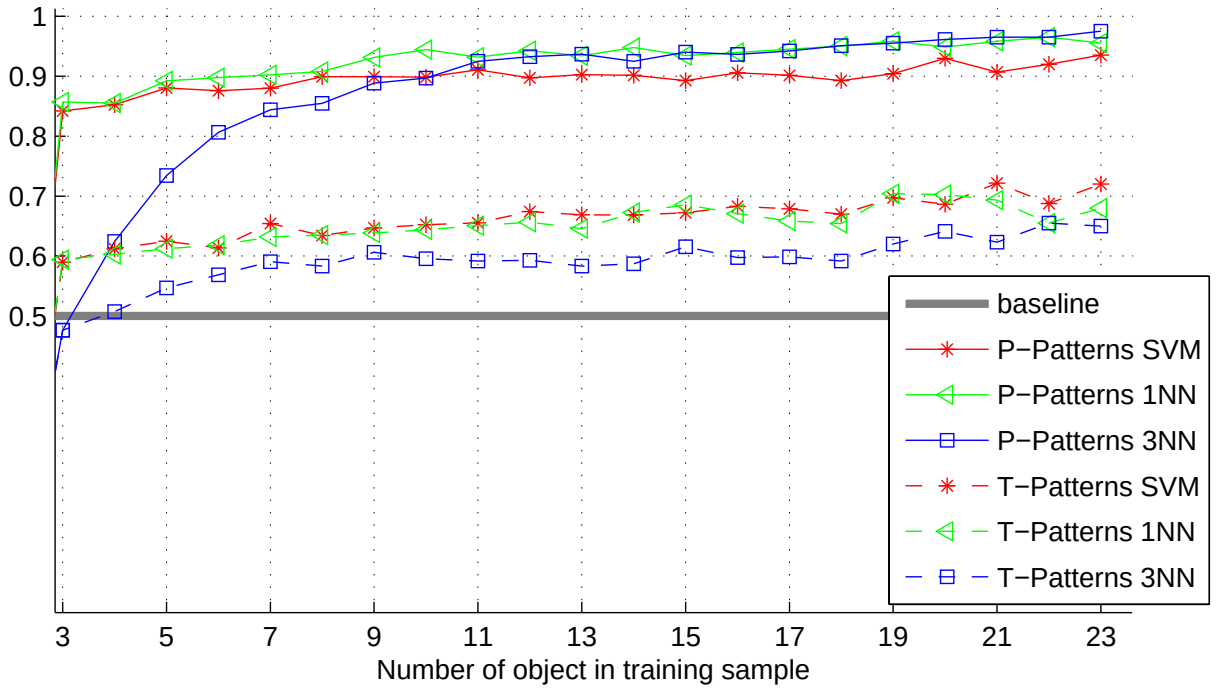



Рис.: Средняя доля правильных ответов классификации методами SVM и $k$NN с разными способами поиска паттернов, в зависимости от размера обучающей выборки. Средняя доля правильных классификации: P-Паттерны — 92%, T-Паттерны —68%.

# Характерные классу животных паттерны

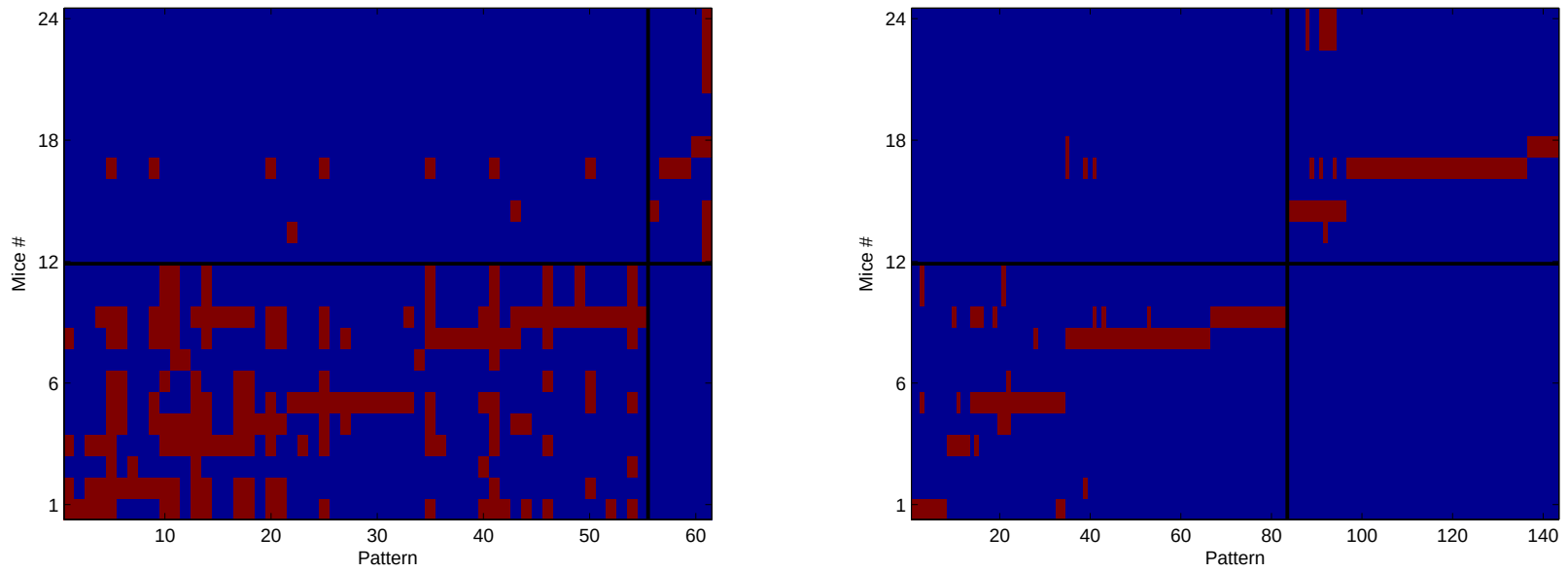



Рис.: Встречаемость найденных паттернов в разных группах мышей. Слева изображены P-Паттерны, справа — T-Паттерны. Ячейка закрашена красным цветом, если данный паттерн встретился в поведении определенной особи, иначе ячейка закрашена синим.

# Пример характерного P-Паттерна

P-Паттерн в формате: $<\ldots$ Событие$_i[\,\mu_i;\sigma_i]\ldots>$ Смещения даны в секундах .

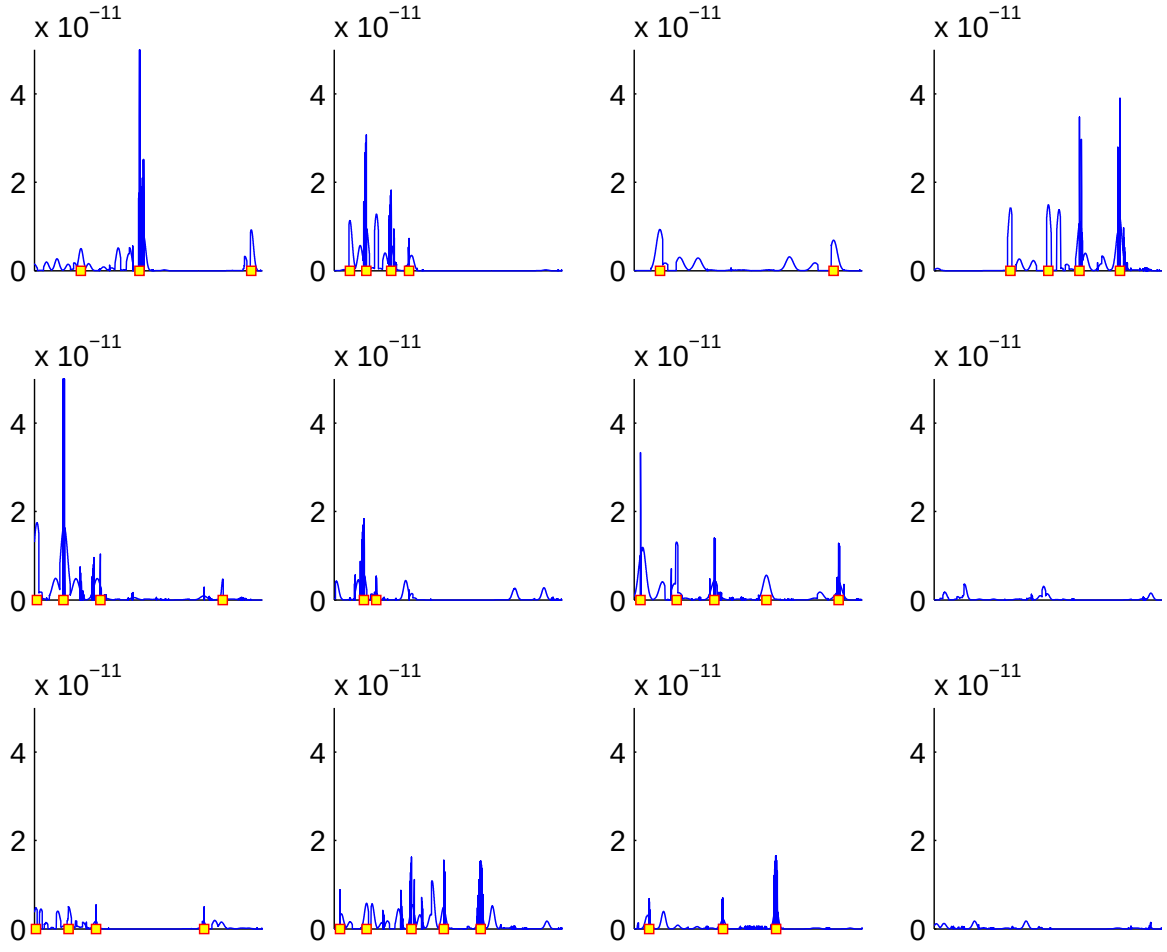
---

Вычес-е задн. конечностями [22.9; 7.9] Вылиз-е ладоней [1.1; 2.7] Быстр. умыв-е носа [0.4; 0.5] Умыв-е головы с ушами [3.2; 7.9] Умыв-е носа [17.0; 7.9] Вылиз-е задн. конечностей

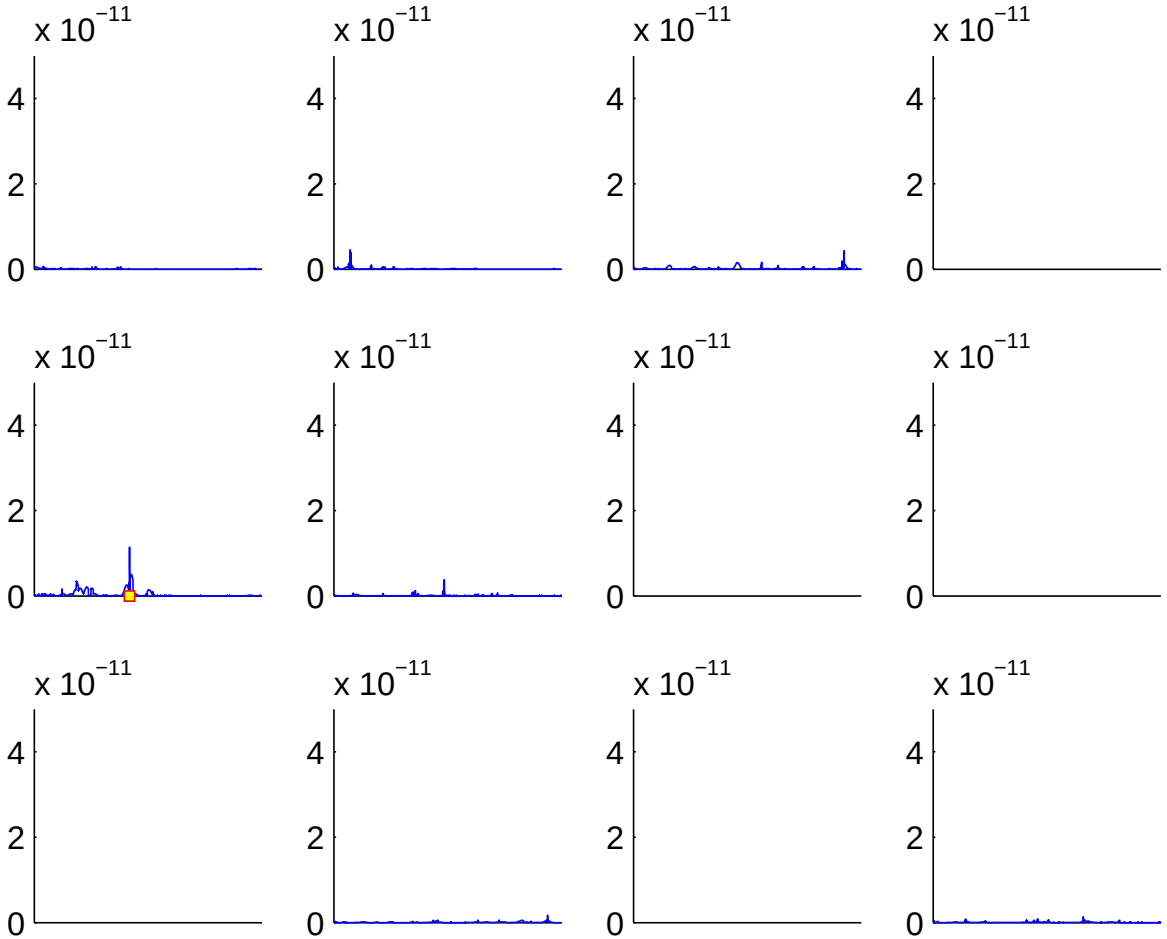
---

Найден у 8 из 12 особей контрольной группы и ни разу не найден в гиппокампальной группе. Имеет биологический смысл.

# Отклик на P-Паттерн в контрольной группе



# Отклик на P-Паттерн в гиппокампальной группе



# Выводы

- Предложенный метод расширяет существующий подход к поиску паттернов.
- Устойчивость к шуму.
- Более вариабельные паттерны, новый подход к описанию поведения с помощью откликов на множество P-Паттернов(мешок слов).
- Достигнуто ускорение параллельной версии на GPU в 40 раз.
- Качество классификации на экспериментальных данных $\sim 92\%$.
- Предложенный метод применим не только для анализа поведения животных(структура ДНК, спайковая активность нейронов, рынки, новостные тренды).

– Сложности на очень маленьких объемах данных.

– Долгое время работы на очень больших объемах данных.

# Выносится на защиту:

- Разработан новый метод поиска поведенческих закономерностей.
- Потенциально новый подход к исследованию поведения.
- Создана свободная, документированная, параллельная(ускорение порядка 40 раз) реализация метода.
- На реальных данных получен биологический результат: классификация мышей по поведению(качество классификации порядка 92%).