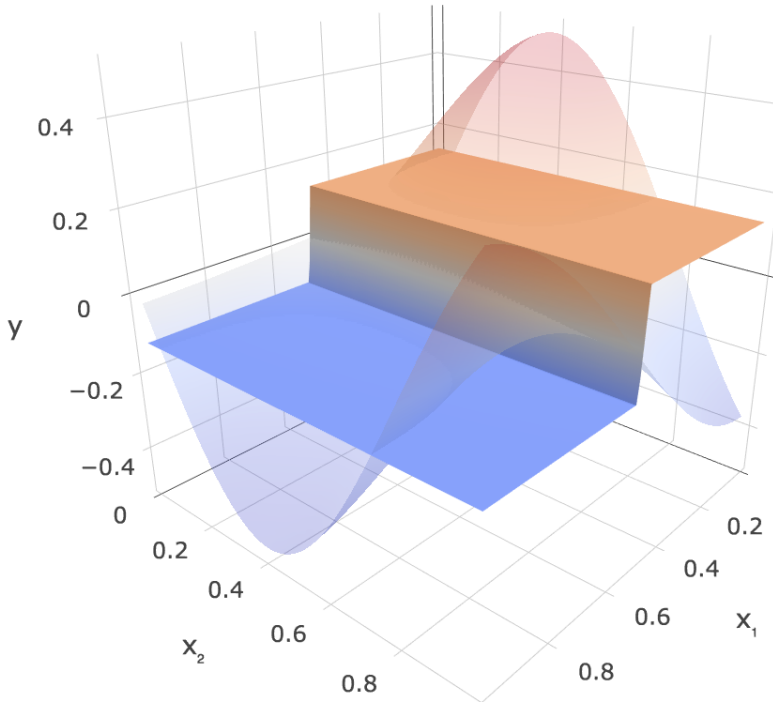
LSML #4

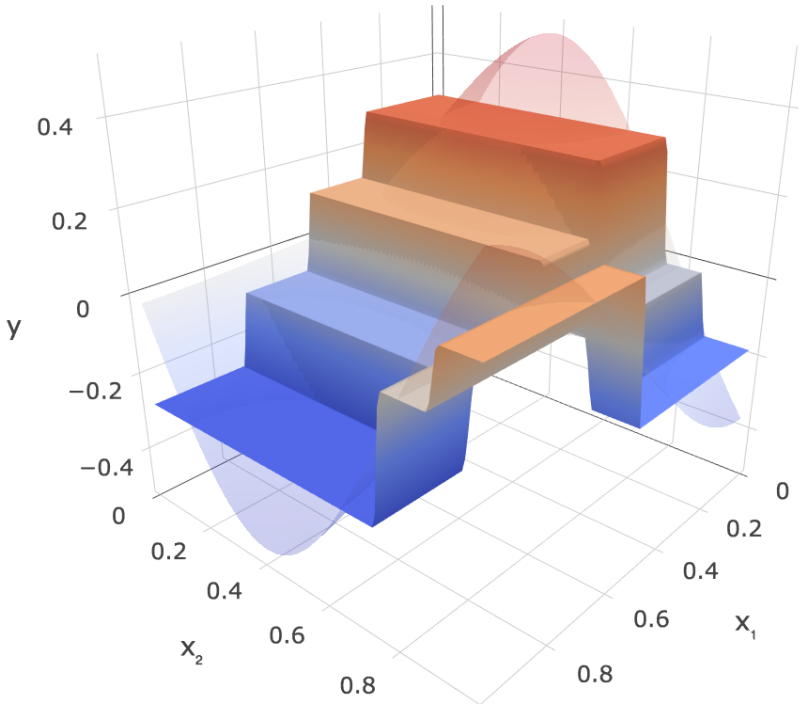
Градиентный бустинг

Одно дерево решений

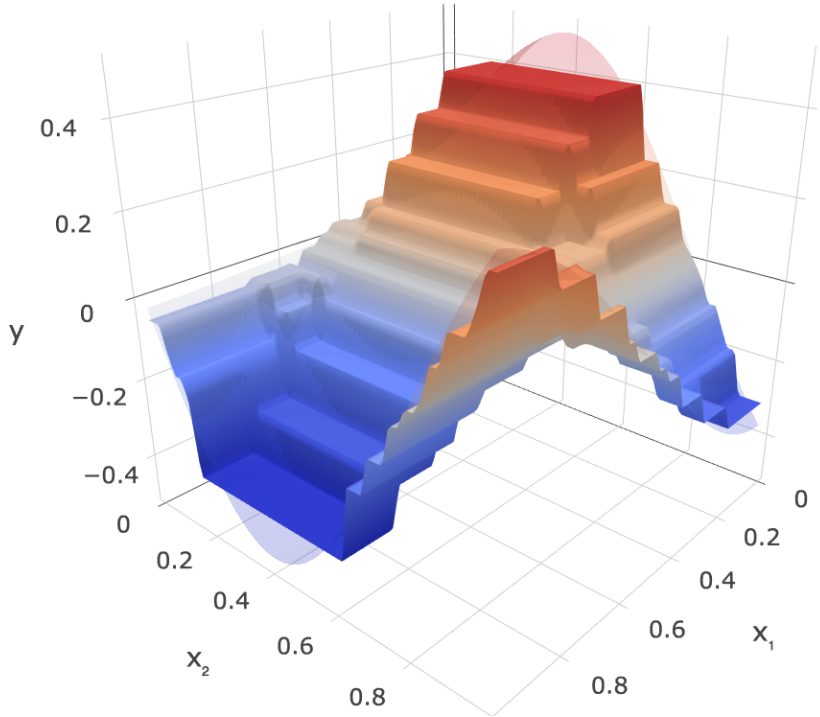
Глубины 1



Глубины 3



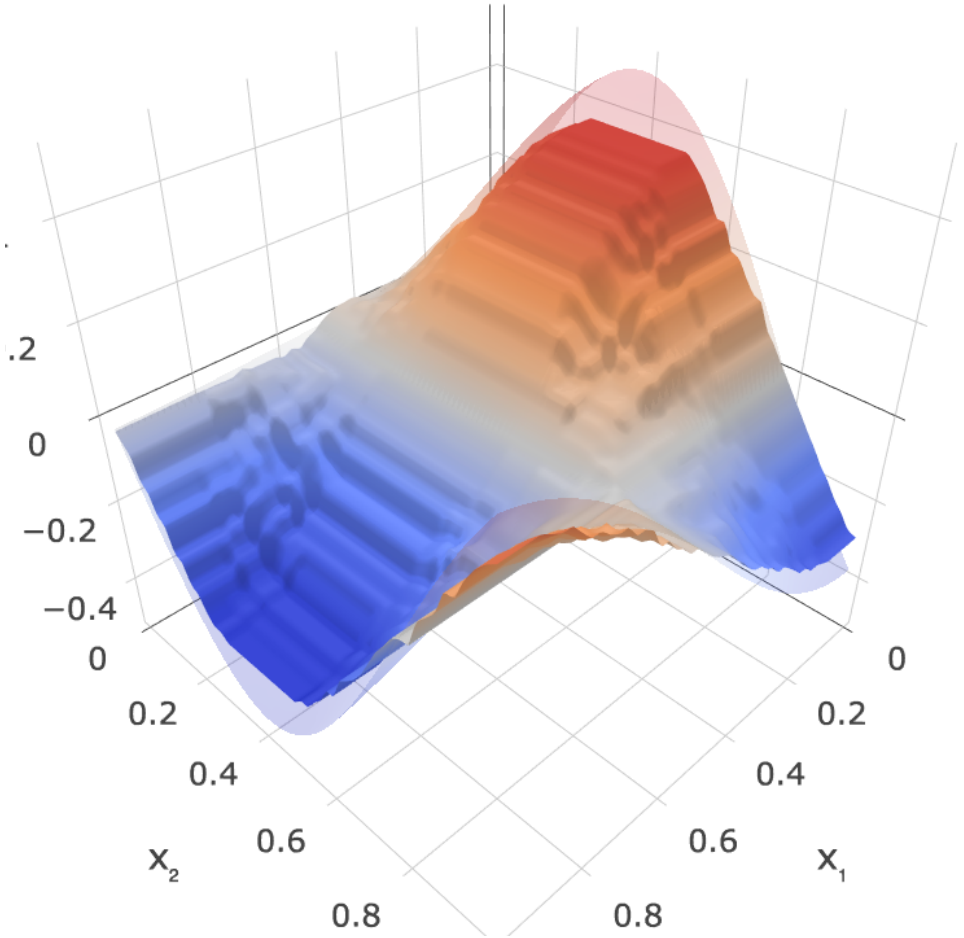
Глубины 6



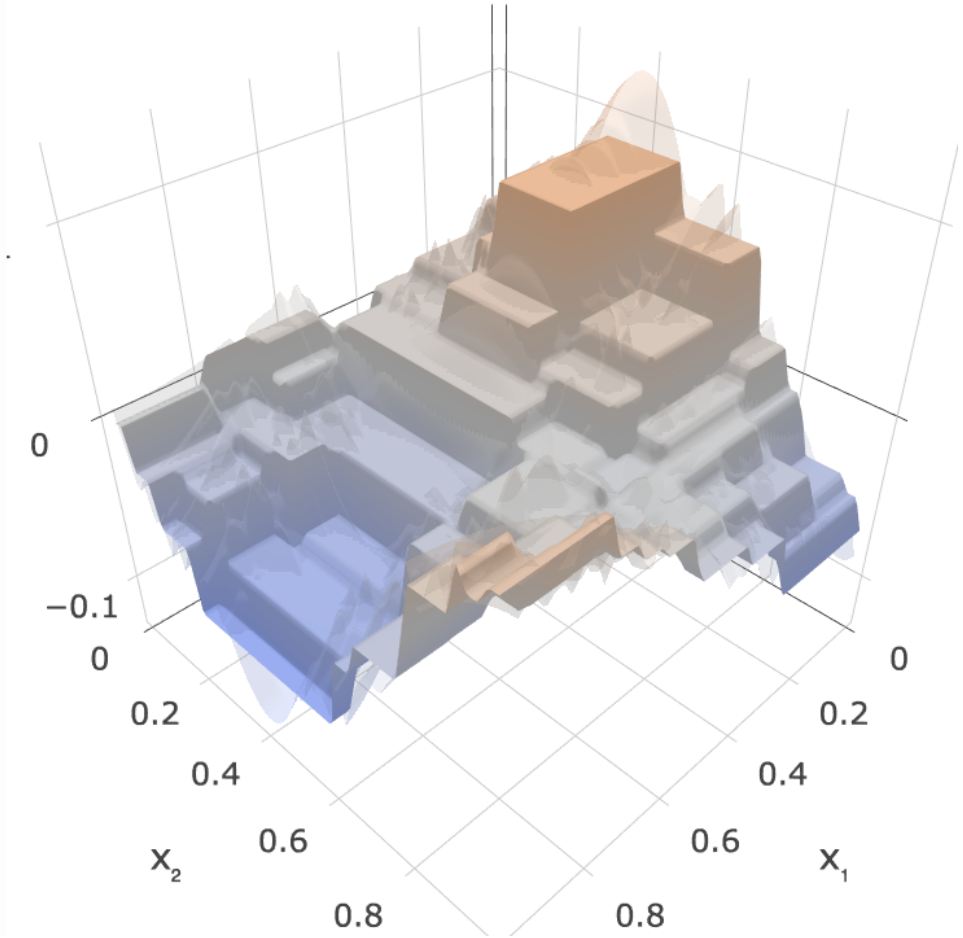
http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html

Строим следующее дерево на остатки

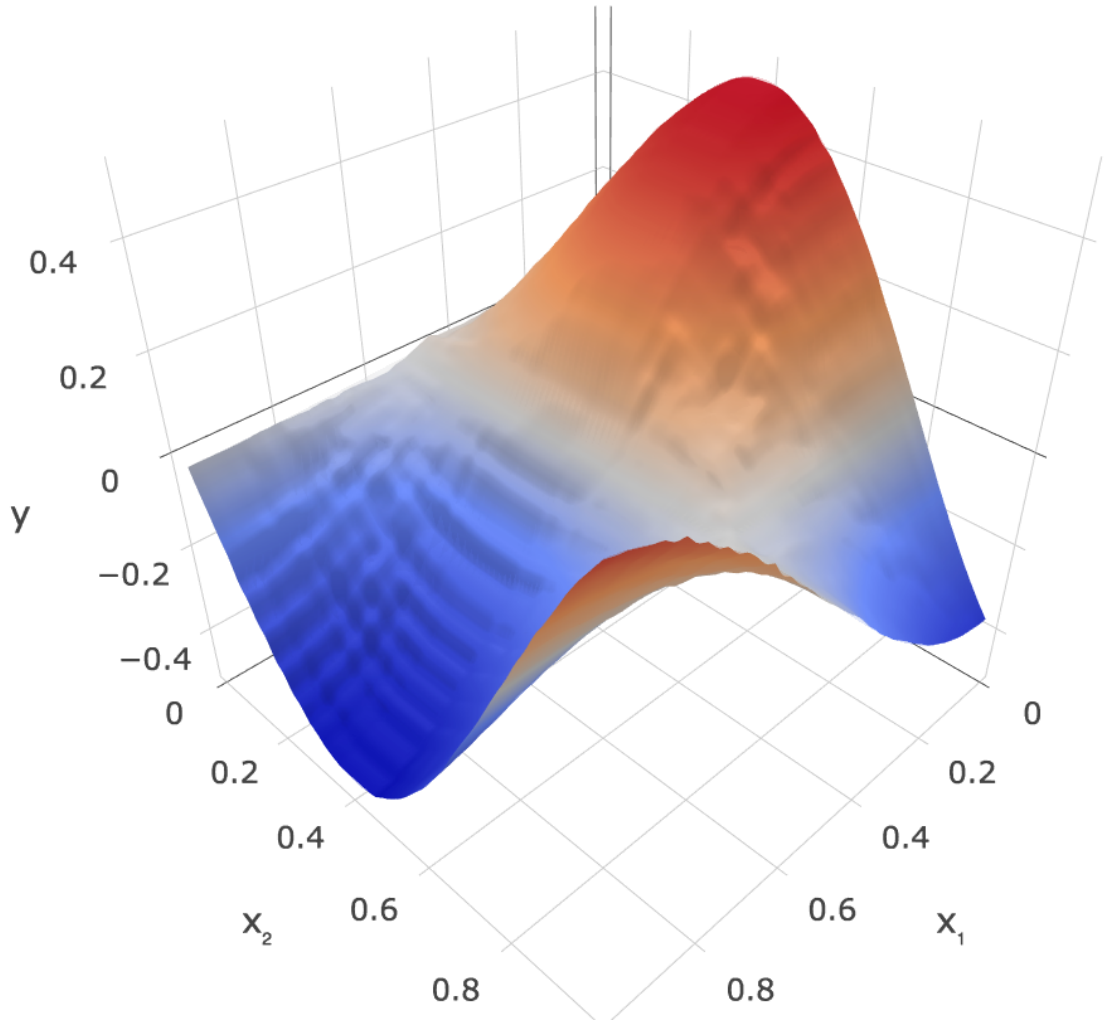
Текущая комбинация



Остатки

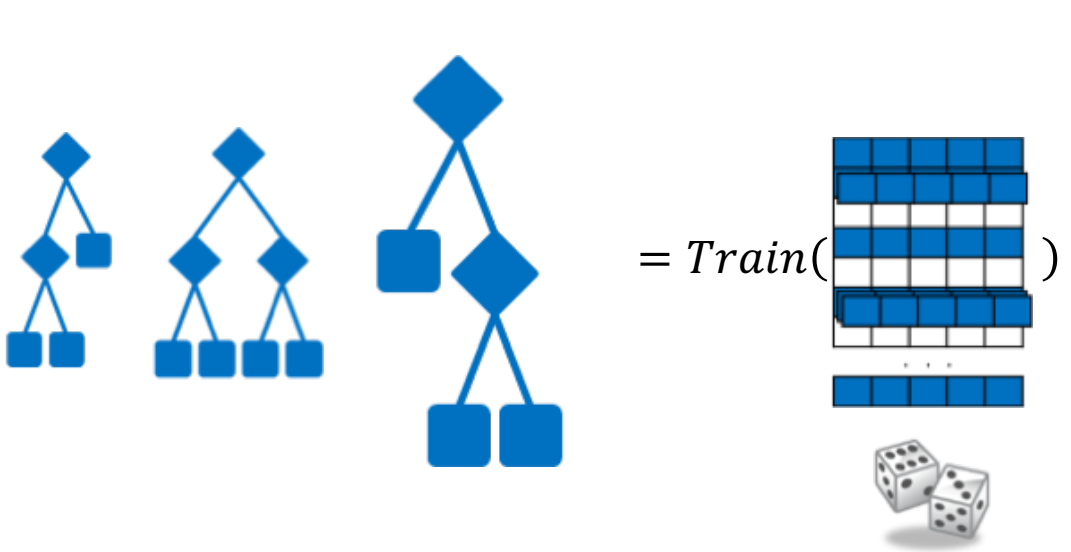


Градиентный бустинг: 100 деревьев

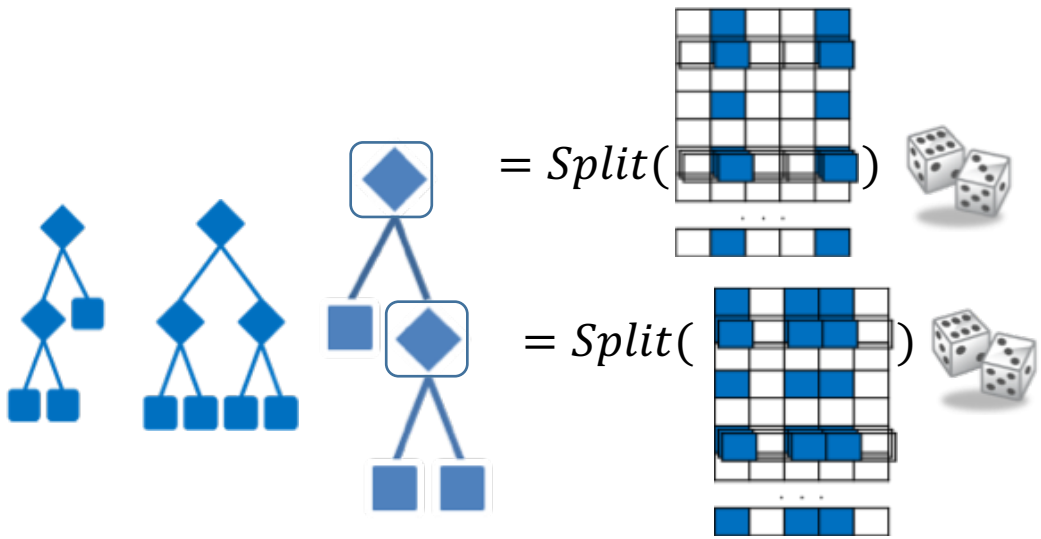


А как же Bagging и Random Forest?

- **Bagging** – учим каждое дерево на bootstrap выборке
- **RF** – bootstrap + семплирование признаков при каждом разбиении
- Легко параллелятся по деревьям
- Локальные для машины данные могут заменять bootstrap



Bagging



RF

http://hunch.net/~large_scale_survey/

Критерий расщепления для регрессии

$$\sum_{j \in N} (y_j - c)^2 - \sum_{j \in N_1} (y_j - c_1)^2 - \sum_{j \in N_2} (y_j - c_2)^2 \to max$$

$$c_1 = \frac{Y_1}{N_1}$$

среднее

$$\sum_{j \in N_1} (y_j - c_1)^2 = \sum_{j \in N_1} y_j^2 - 2c_1 \sum_{j \in N_1} y_j + \sum_{j \in N_1} c_1^2$$

$$= \sum_{j \in N_1} y_j^2 - 2c_1^2 N_1 + c_1^2 N_1$$

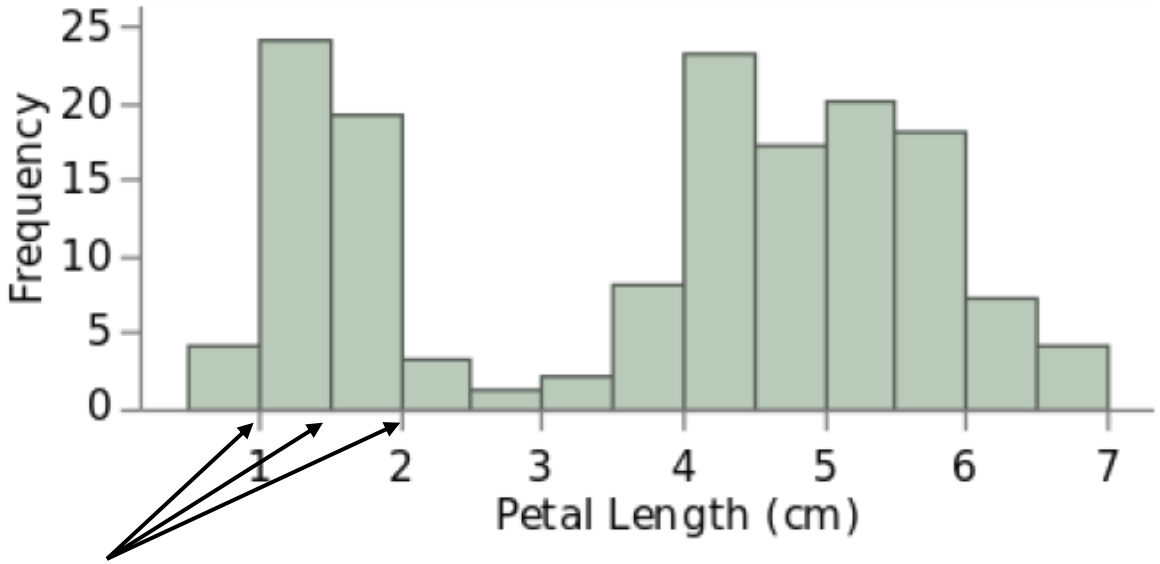
$$= \sum_{j \in N_1} y_j^2 - \left(\frac{Y_1}{N_1}\right)^2 N_1$$

$$= \sum_{j \in N_1} y_j^2 - \frac{Y_1^2}{N_1}$$

$$-\frac{Y_1^2}{N_1} - \frac{Y_2^2}{N_2} \to min$$

Gradient Boosting

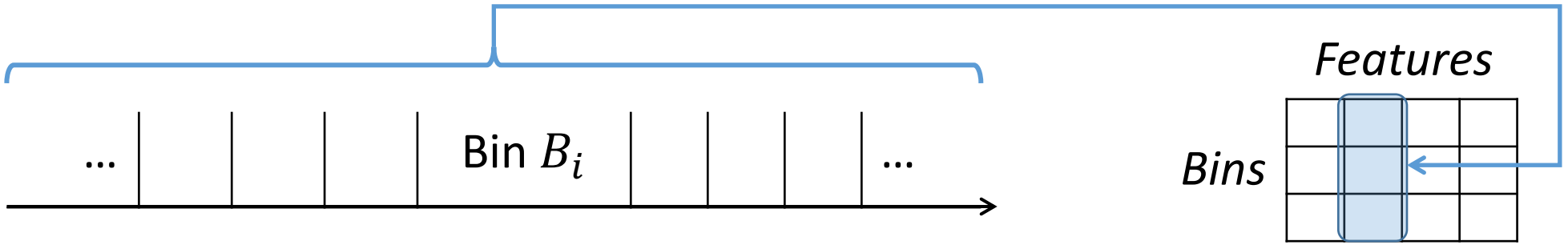
- Бустинг итерационный – надо параллелить создание дерева
- Для каждой вершины перебираются все признаки и пороги
  - Вещественные признаки дискретизируем по корзинкам (binning)
  - Порогами будут являться границы корзинок



Пороги

Feature Binning

- На примере задачи **регрессии**, разбили признак на k корзинок



- **Первый шаг**: собираем статистики

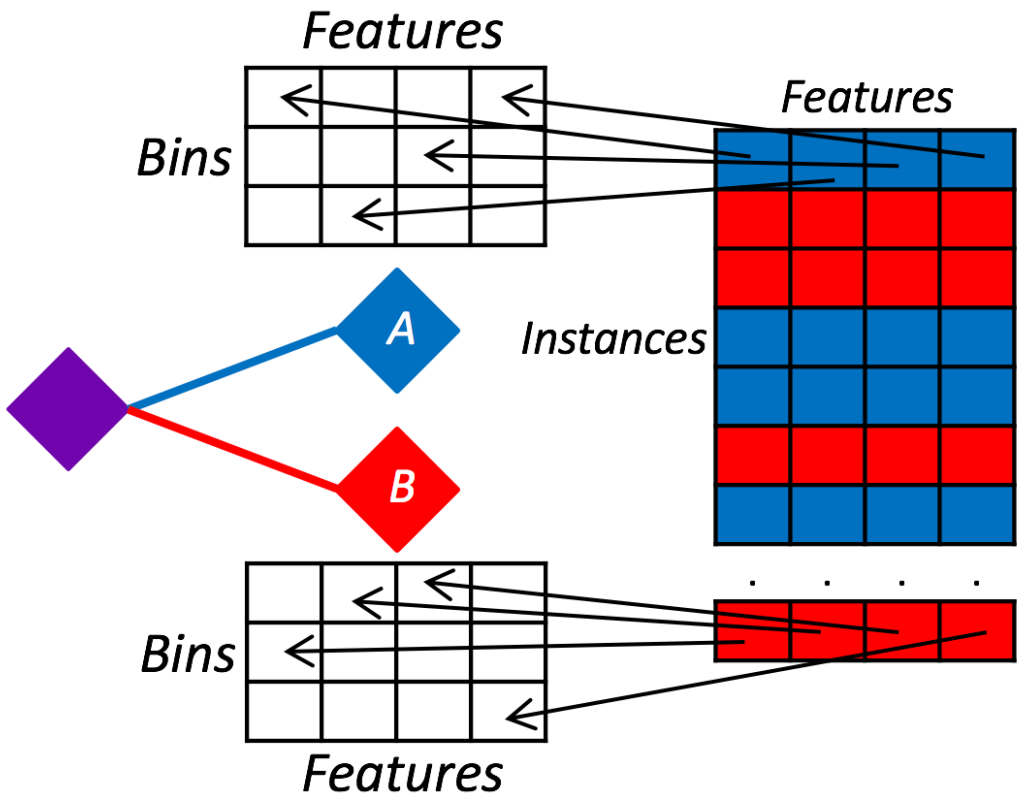
$$N_i = \sum 1 \qquad Y_i = \sum y \qquad N = \sum N_i \qquad Y = \sum Y_i$$

- **Второй шаг**: считаем пользу сплита по правой границе корзинки B_i

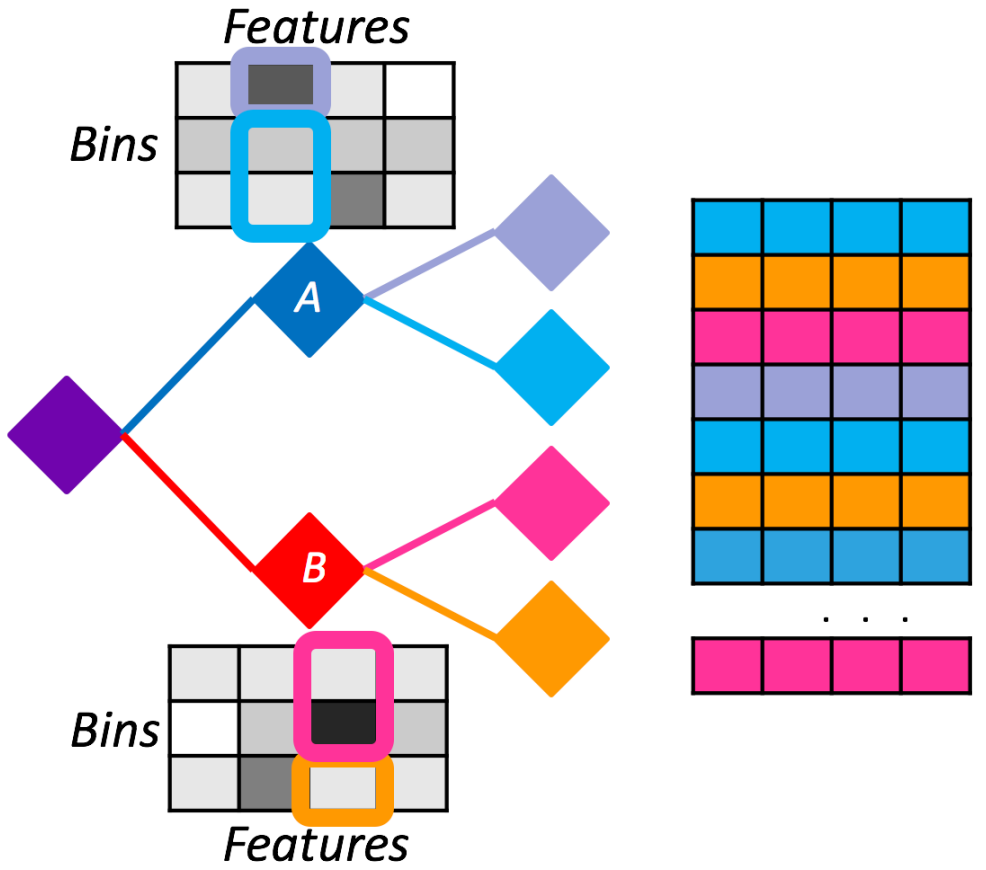
$$-\frac{Y_{0:i}^2}{N_{0:i}} - \frac{(Y - Y_{0:i})^2}{N - N_{0:i}}$$

$0{:}\,i$ – нотация суммирования

Feature Binning



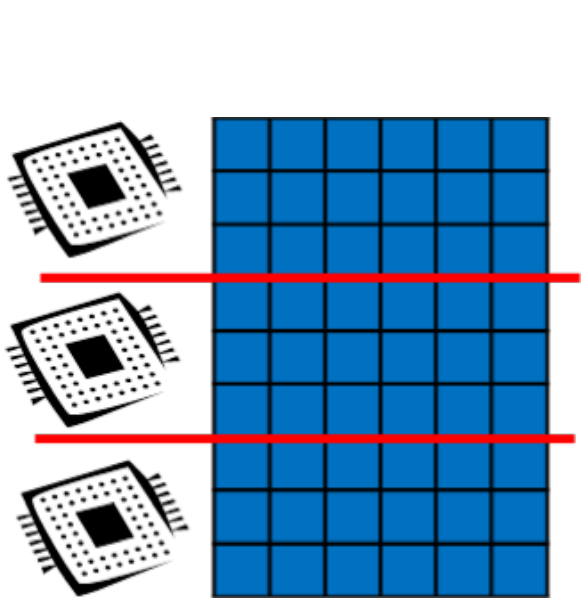
Первый шаг: считаем статистики

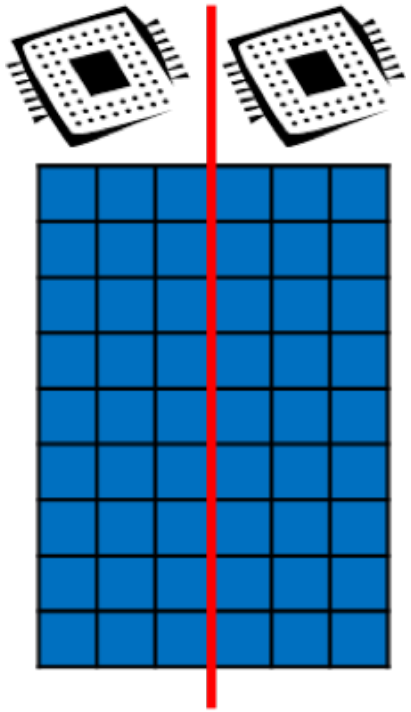


Второй шаг: выбираем сплит

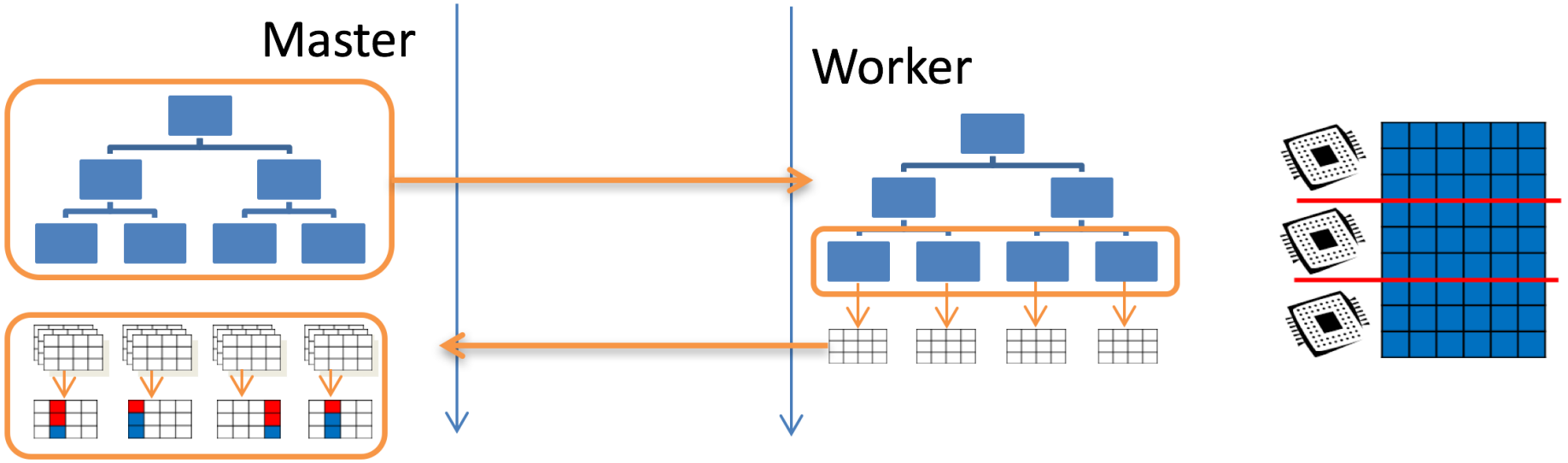
Feature Binning

- Один проход по данным для каждого уровня дерева
- Проход по данным можно распределять по объектам или признакам



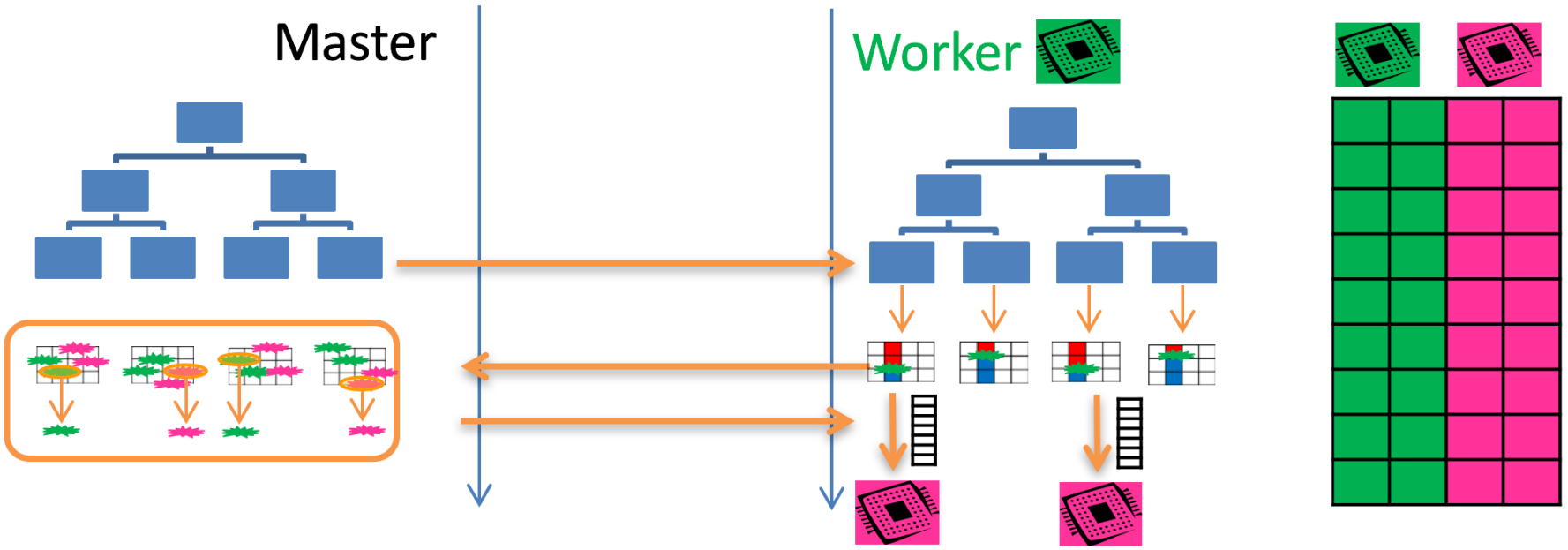


Распределяем объекты



- Мастер
  - Посылает воркерам текущую модель
  - Агрегирует локальные гистограммы (Features-Bins) воркеров и выбирает лучший сплит
- Воркеры
  - Делают проход по своим данным и заполняют локальные гистограммы

Распределяем признаки



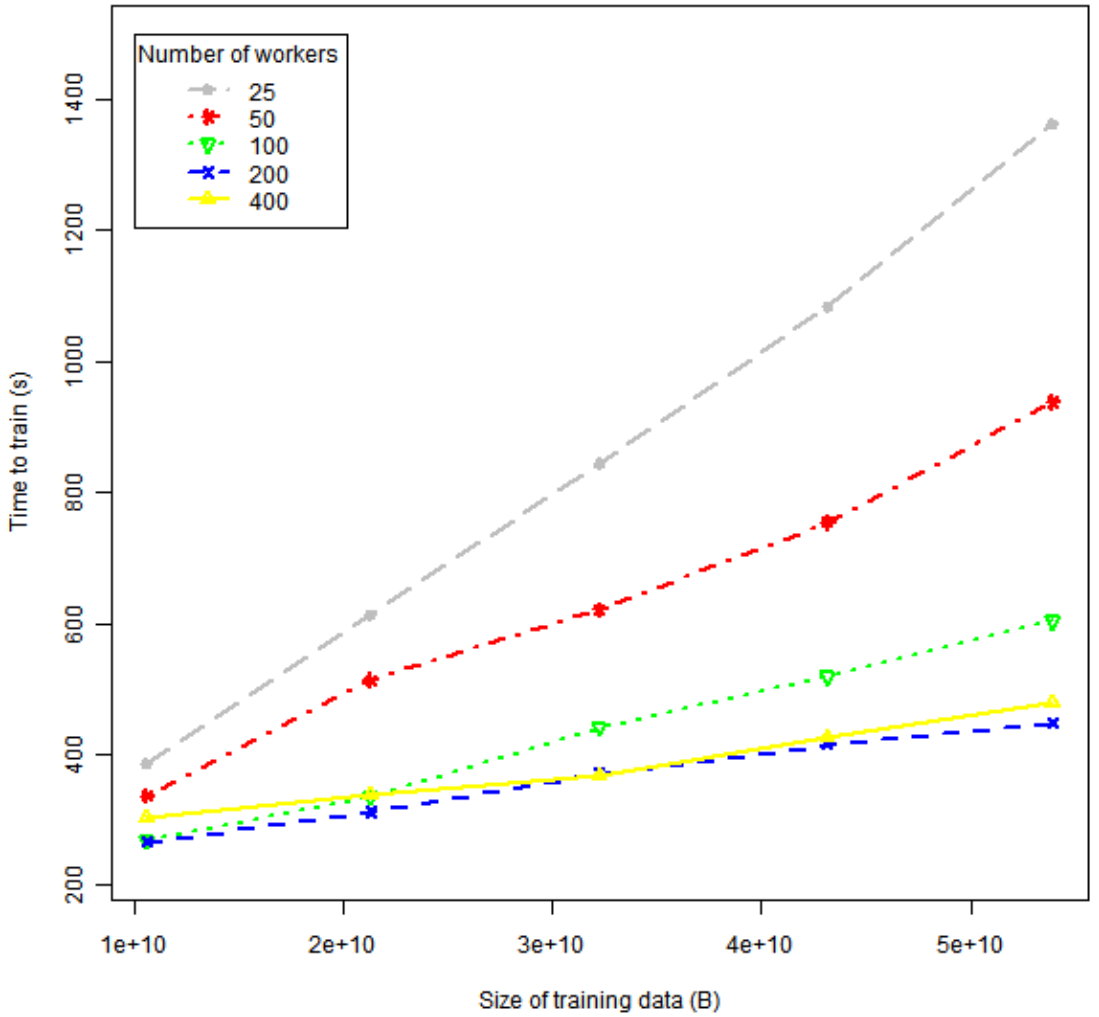
- Мастер
  - Получает лучшие сплиты по признакам от воркеров и выбирает лучший среди них
  - **Просит лучших разослать всем остальным информацию про новый выбранный сплит**
- Воркеры
  - **Всех признаков нет**, помнят в какой лист попадает каждый объект
  - Делают проход по своим признакам и выбирают лучший сплит

Пример реализации: PLANET (2011)

- MapReduce, RPC
- Разбиение по объектам
- Binning во время инициализации
  - За один проход оценивают квантили

Пример реализации: PLANET (2011)

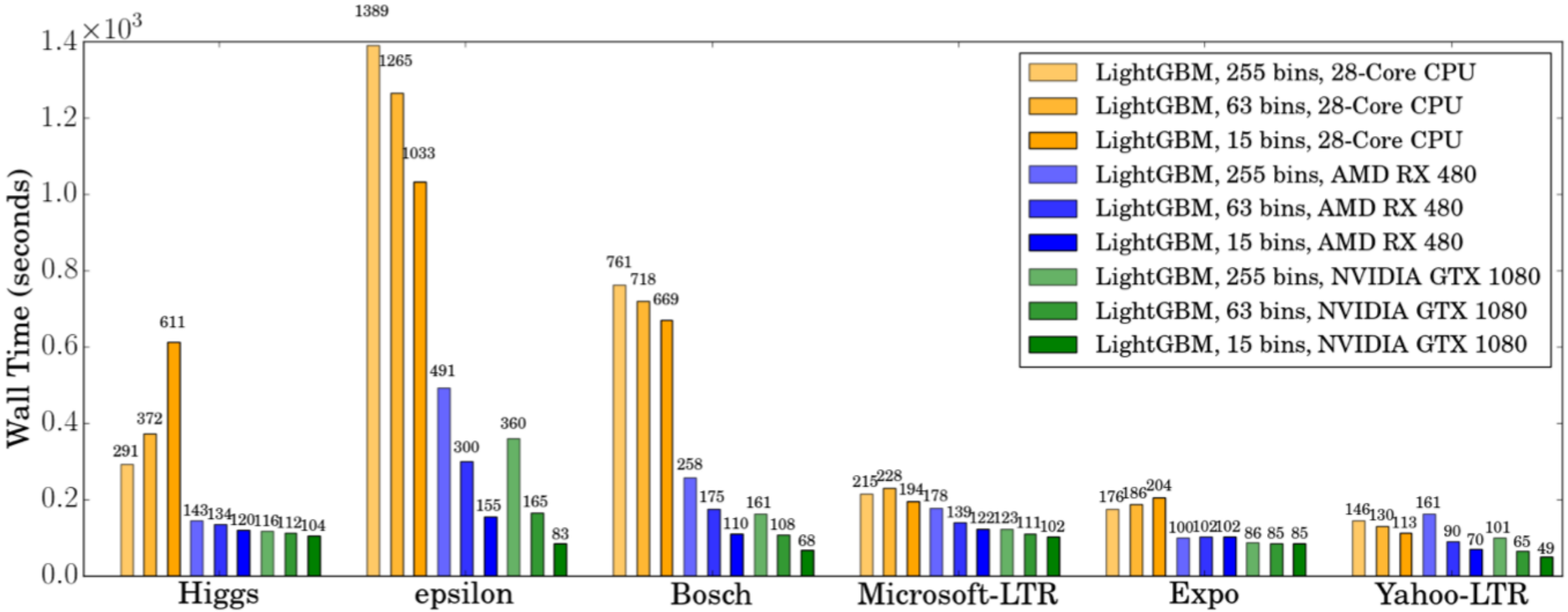
- Бинарная классификация
- Признаки:
  - 4 вещественных
  - 6 категориальных, $|C| \in$[2-500]
- 314 млн объектов
- Деревья глубины 3



Пример реализации: Yahoo! GBDT (2009)

- Hadoop, MPI
- Разбиение по объектам или признакам
- Результаты:
  - MapReduce Horizontal: 211 minutes x 2500 trees = 366 days (100 machines)
  - MapReduce Vertical: 28 seconds x 2500 trees = 19.4 hours (20 machines)
  - MPI: 5 seconds x 2500 trees = 3.4 hours (10 machines)

Пример реализации: LightGBM от Microsoft



https://github.com/Microsoft/LightGBM/blob/master/docs/GPU-Performance.md

Пример реализации: Catboost от Яндекс

- Разбиение по документам
- Ускорение обучения (CPU, GPU)
- Новые регуляризации и функции ошибок
- Хорошие модели без подбора параметров

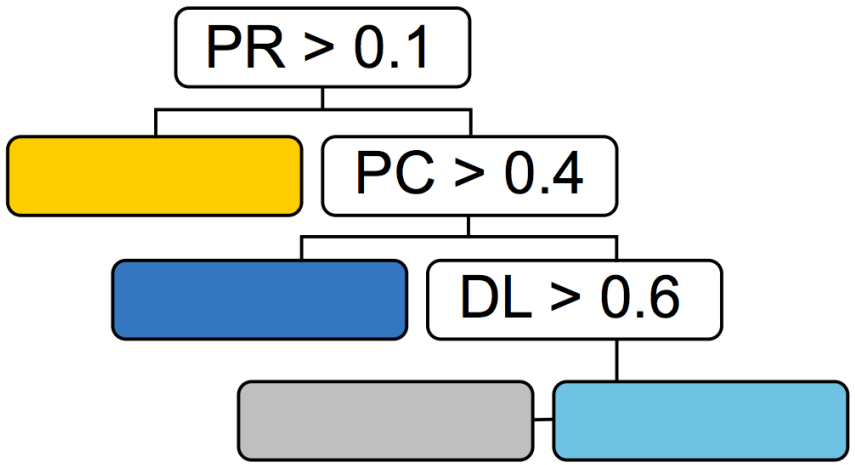
Пример реализации: Catboost от Яндекс

	Catboost	Azure Boosted DT	XGBoost	LightGBM
Pol	0,994	0,922 ↓0,14%	0,991 ↓0,23%	0,991 ↓0,23%
2dplanes	0,9476	0,9474 ↓0,02%	0,9474 ↓0,02%	0,9474 ↓0,01%
Elevator	0,915	0,909 ↓0,67%	0,9 ↓1,54%	0,908 ↓0,74%
Ailerons	0,86	0,856 ↓0,45%	0,837 ↓2,67%	0,856 ↓0,55%
Fried	0,957	0,955 ↓0,22%	0,954 ↓0,32%	0,955 ↓0,17%
House	0,677	0,68 ↑0,51%	0,658 ↓2,72%	0,661 ↓2,23%

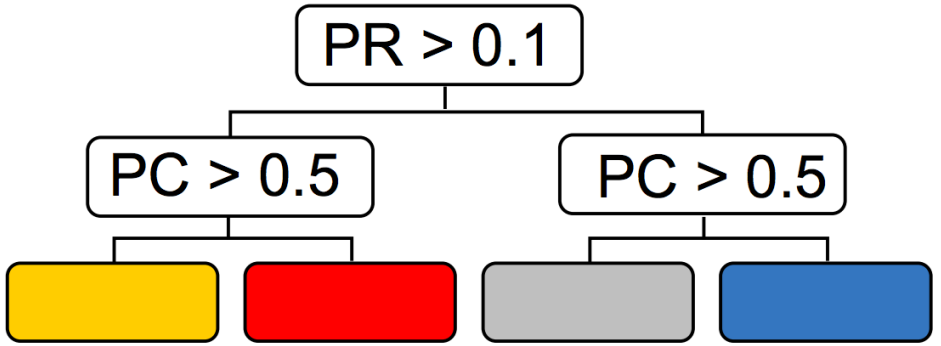
Сравнение на открытых наборах данных
с сайта www.openml.org
Метрика – R2-коэффициент детерминации

Oblivious trees в Catboost

Дерево решений



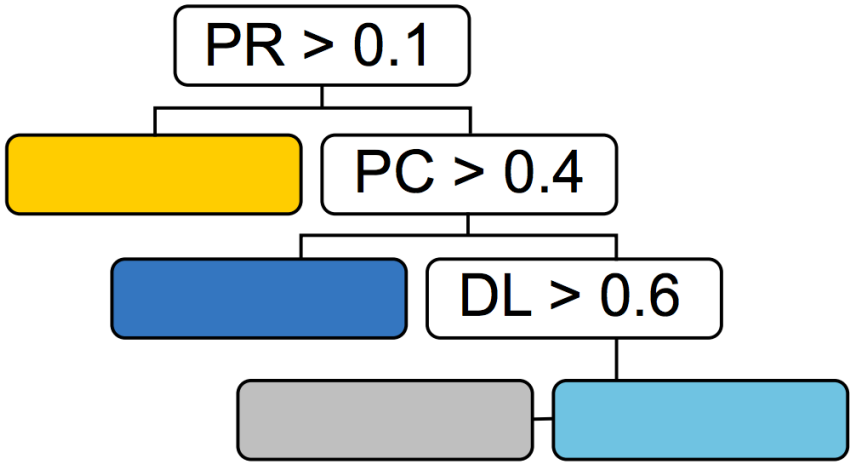
Oblivious дерево



https://yadi.sk/i/3nzyqx063EAYmi

Oblivious trees в Catboost

Дерево решений



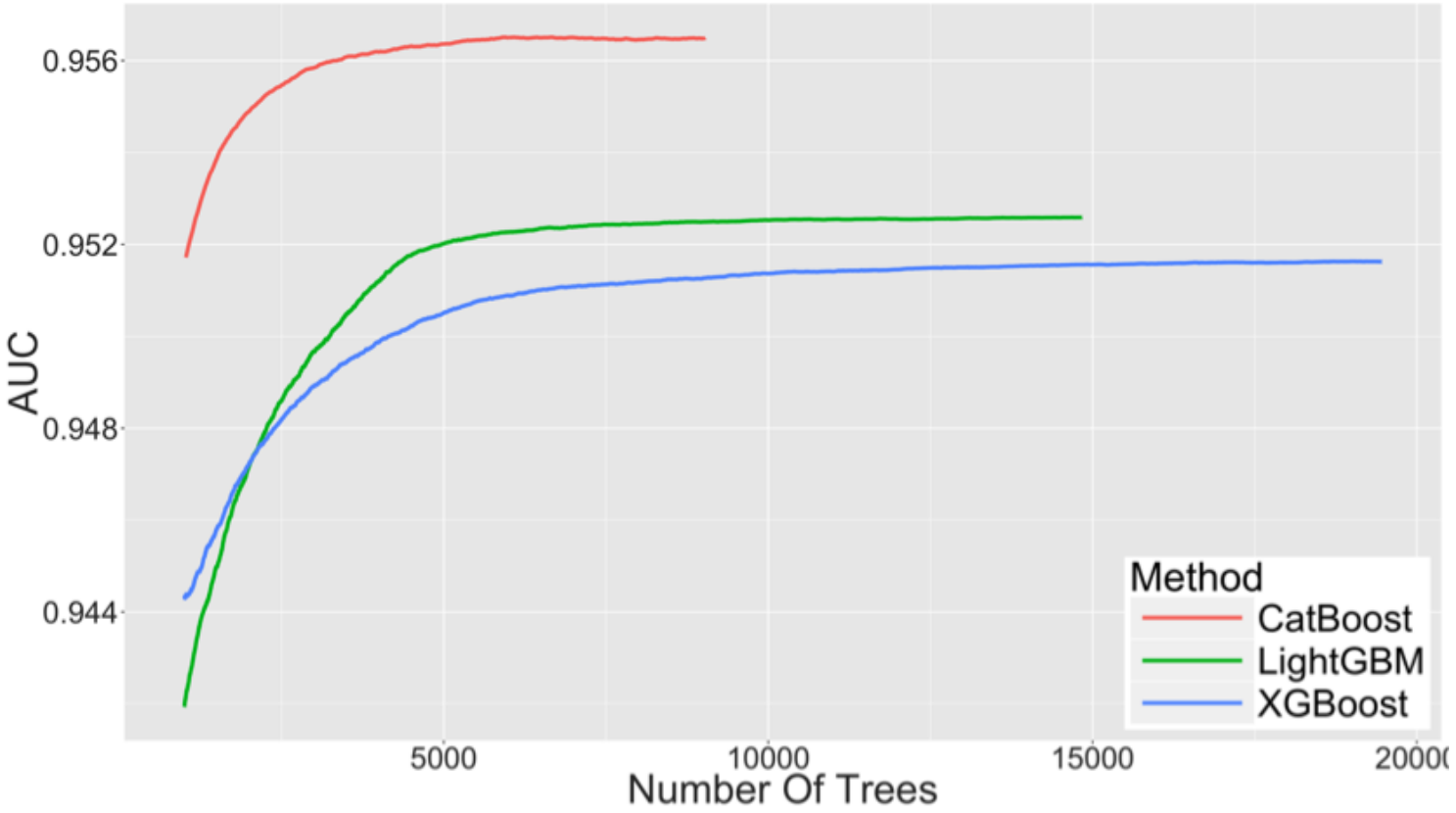
Oblivious дерево

	PR ≤ 0.1	PR > 0.1
PC > 0.5		
PC ≤ 0.5		

Catboost на GPU

- https://www.kaggle.com/c/criteo-display-ad-challenge
- first 36M samples, 26 categorical, 13 numerical features

	128 bins
CPU 32 cores	1060 (1.0)
K40	373 (2.84)
GTX 1080	285 (3.7)
P40	123 (8.6)
GTX 1080Ti	301 (3.5)
P100-PCI	82 (12.9)
V100-PCI	69.8 (15)



https://github.com/catboost/benchmarks/tree/master/gpu_training

Ссылки

- Как устроен xgboost http://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
- xgboost: https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf
- Catboost и обучение на GPU: https://github.com/catboost/benchmarks/tree/master/gpu_training
- LightGBM и обучение на GPU https://github.com/Microsoft/LightGBM/blob/master/docs/GPU-Performance.md