Анализ данных и Машинное обучение в гидрологии

...

Неделя 7

План

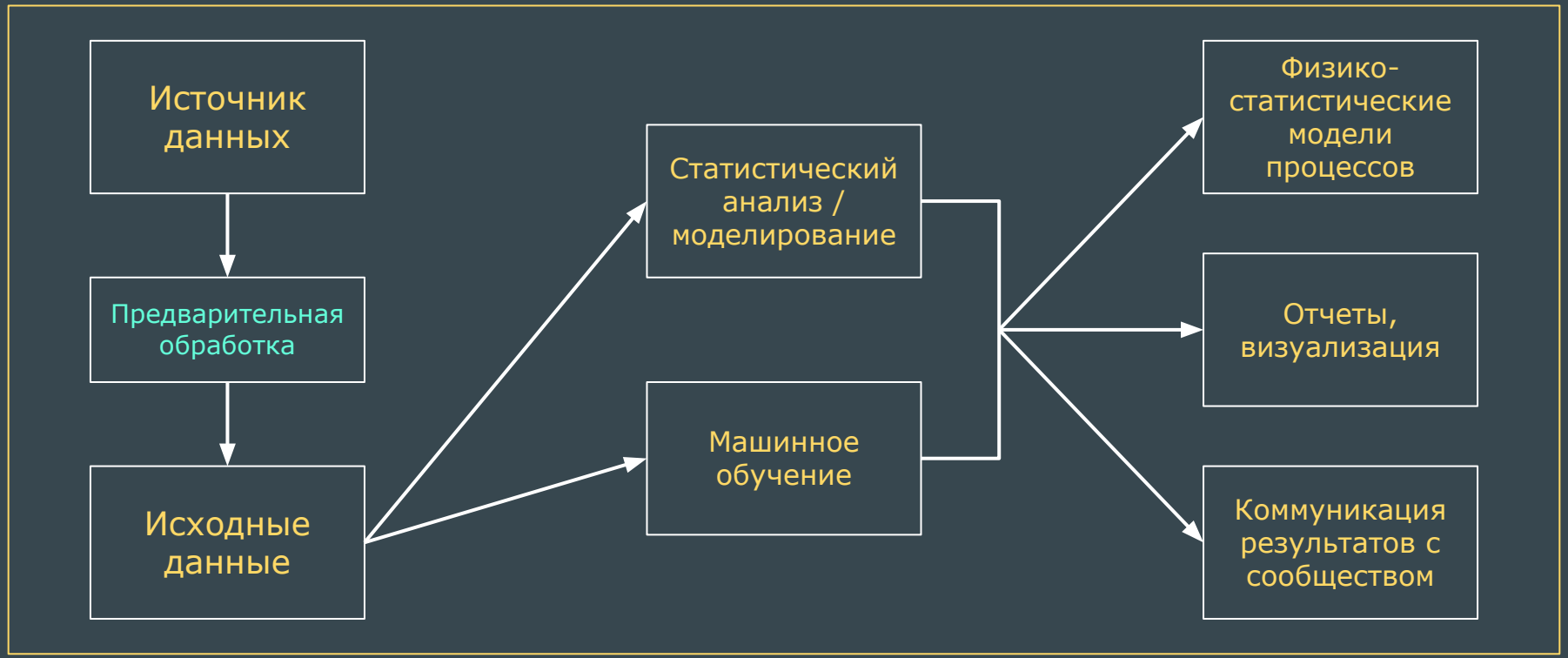
Лекция

- одиночное дерево решений
- параметры, свойства, обучение
- ансамбли деревьев решений
- обобщающая способность, интерпретируемость
- результаты использования

Практикум

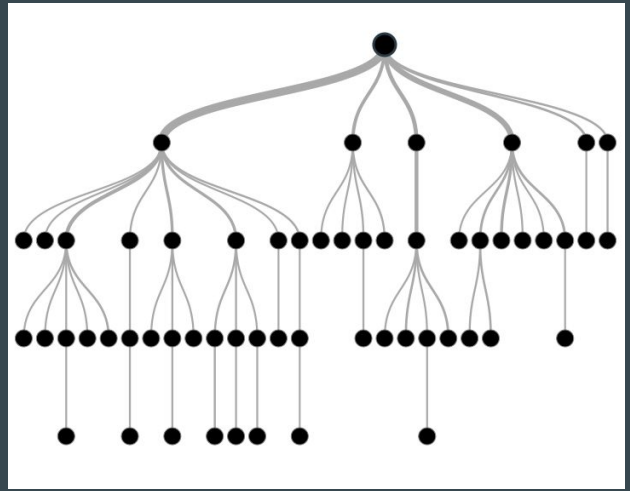
- подготовка данных
- исследование одиночного решающего дерева
- модель случайного леса
- настройка гиперпараметров
- проверка устойчивости

Research workflow



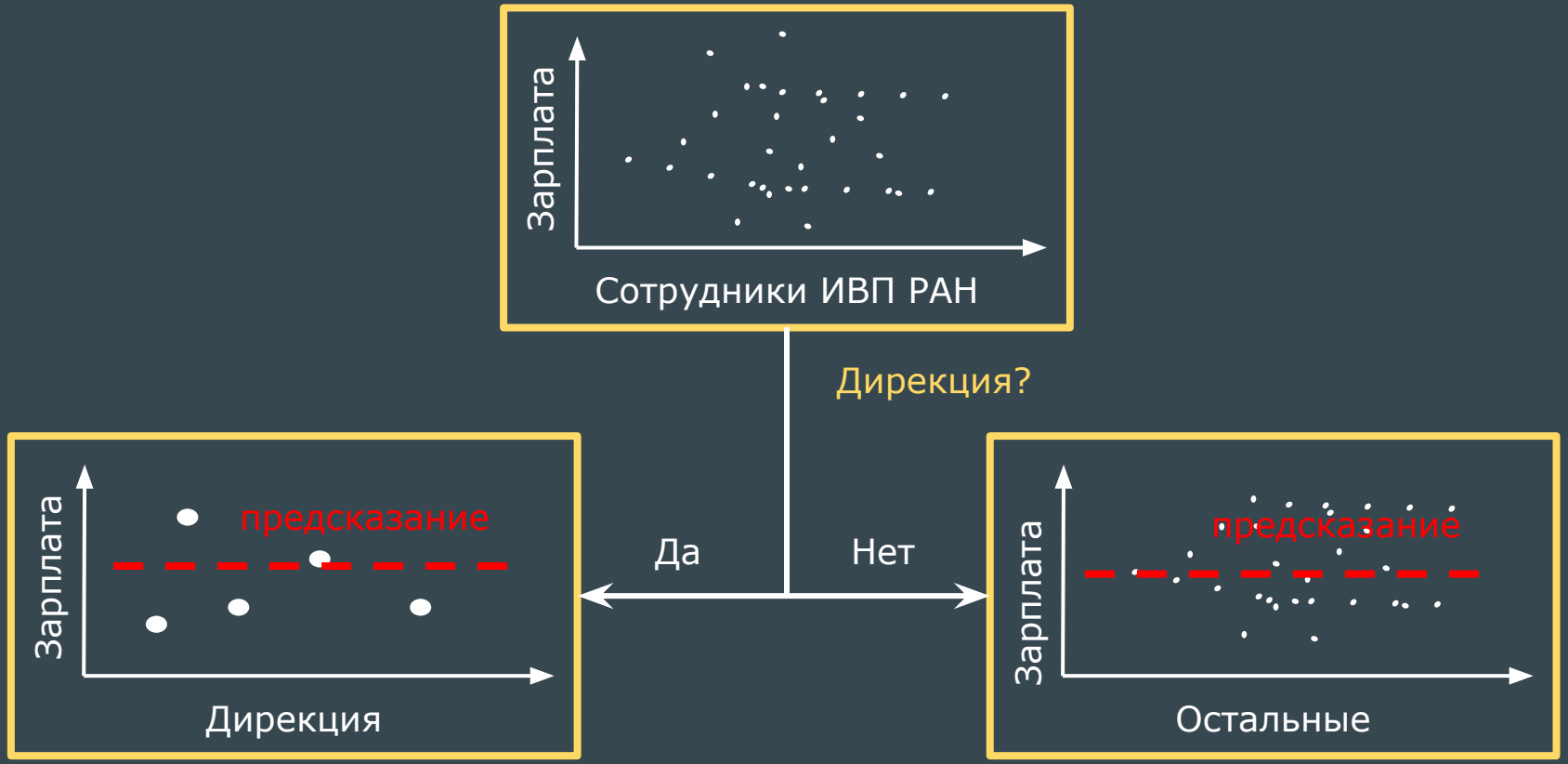
Одиночное решающее дерево

+ Breiman et al., 1984
+ решение задач классификации
+ обобщение для регрессии

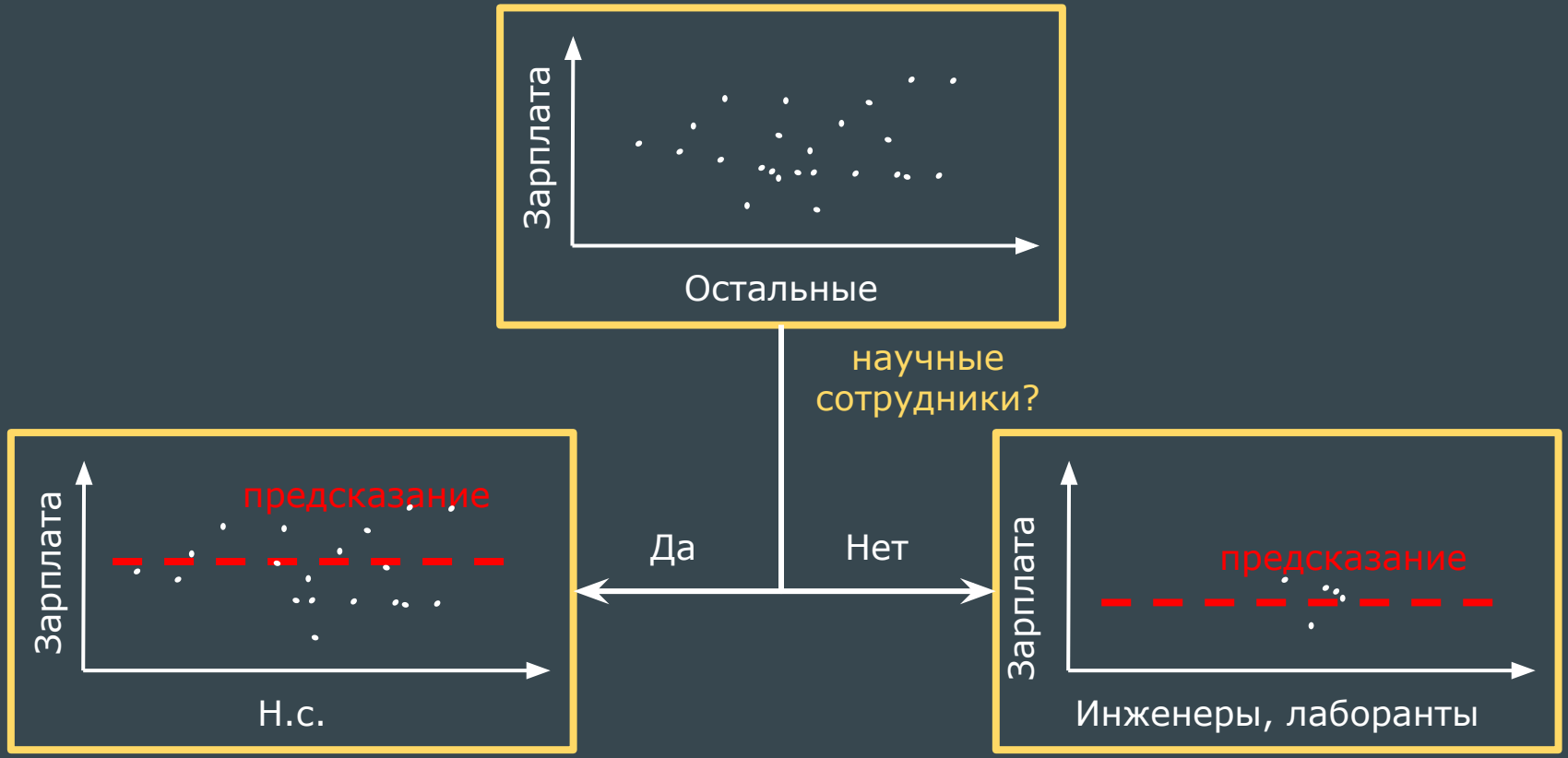


+ интерпретируемость
+ быстрота обучения
+ толерантность к неполным данным

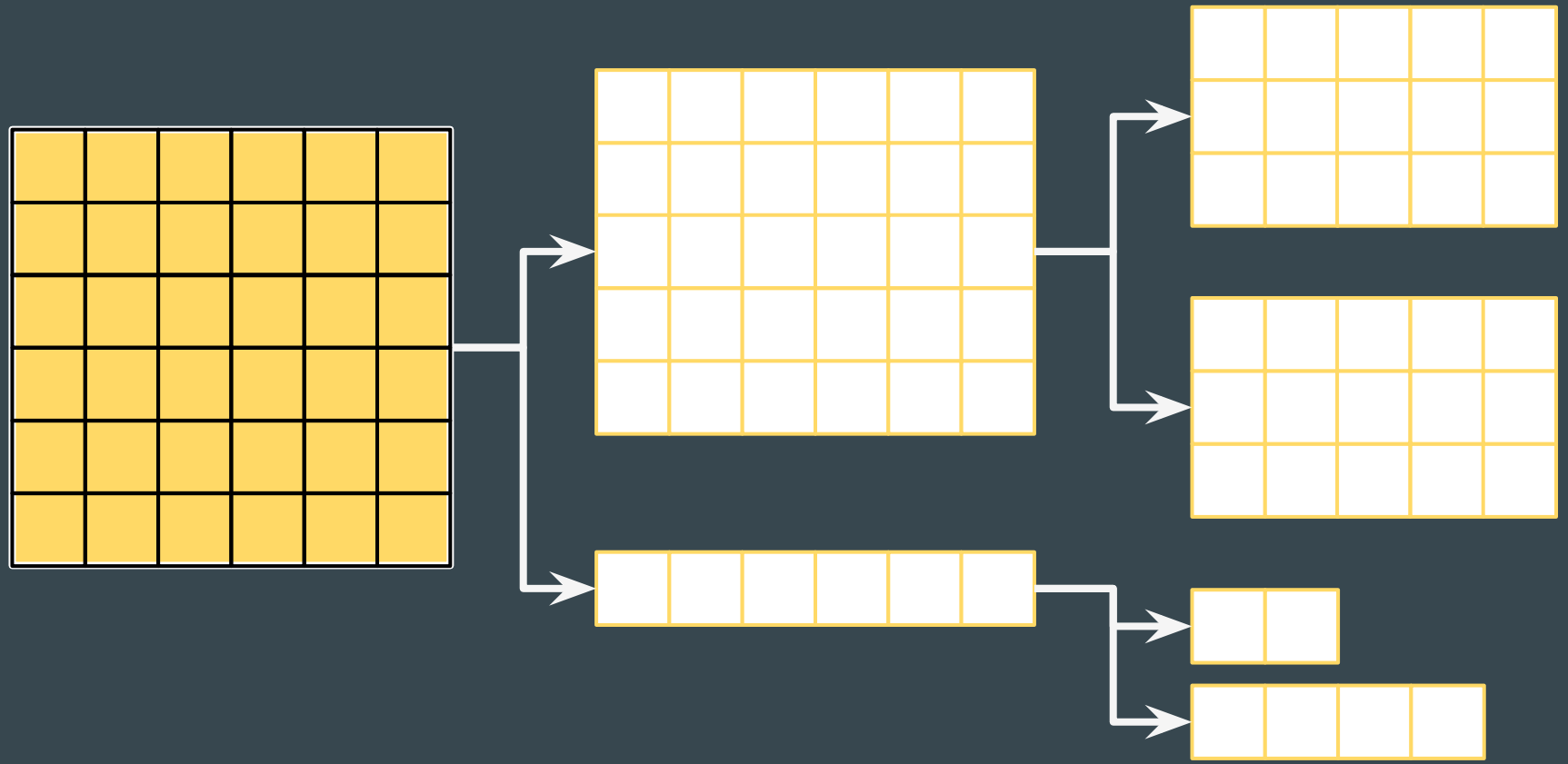
Одиночное решающее дерево. Рост



Одиночное решающее дерево. Рост



Одиночное решающее дерево. Рост



Одиночное решающее дерево. Рост

$$\mathbf{S} = \square (y_i - m_c)^2 \rightarrow 0$$

$i \in$ cluster

$c \in$ Tree

Решающее дерево. Критерии останова

Подходы:

1. Пороговый. Сравниваем критерий с порогом.
2. Строим до единственного значения в листе (кластере), затем обрезаем (prunning).

Критерии:

1. Глубина дерева.
2. Количество наблюдений в вершине.
3. Разделимость (толерантность).

Решающее дерево. Подводим итоги

Преимущества:

1. Простота, интерпретируемость.
2. Встроенный отбор признаков.
3. Работа с непрерывными/категориальными признаками.
4. Не нужна нормализация.

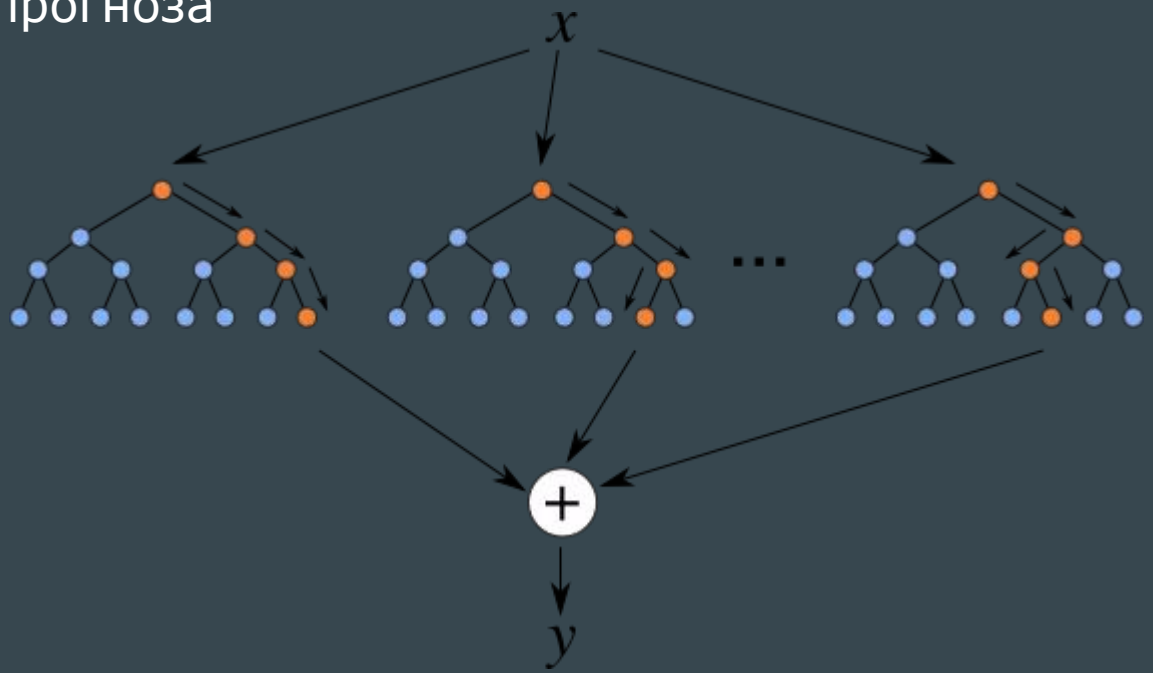
Недостатки:

1. Глубокое дерево для линейной зависимости (разделимости).
2. Переобучение.
3. Нет онлайновости (полное перестроение при новых данных).

Ансамбли решающих деревьев

Эвристика:

Осреднение по большому количеству разных моделей снизит дисперсию прогноза



Ансамбли решающих деревьев

+ Bagging (bootstrap aggregating) on decision trees
+ AdaBoost (adaptive boosting)
+ Random Forest
+ ExtraTrees
+ XGboost (gradient boosting)

Random Forest

+ универсальный алгоритм
+ сильный алгоритм
+ серебряная пуля машинного обучения
+ высокая скорость обучения
+ поддерживает параллельные вычисления

+ мало гиперпараметров
+ поиск оптимума перебором по сетке

Random Forest

формирование моделей

Sample bootstrap

каждая новая модель обучается на индивидуальном подмножестве наблюдений (с возвращением)

Features bootstrap

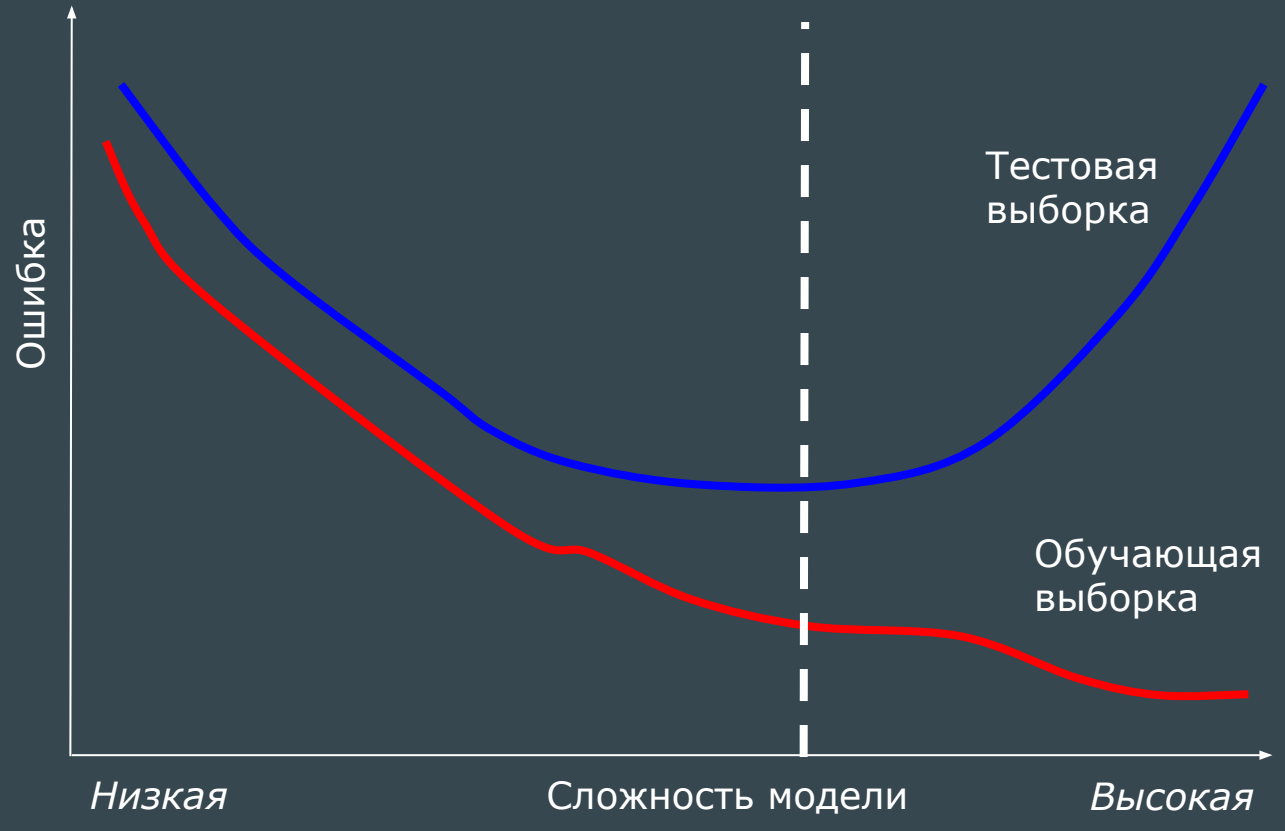
каждая новая модель обучается на случайном наборе признаков

конечный результат: простое осреднение предсказаний

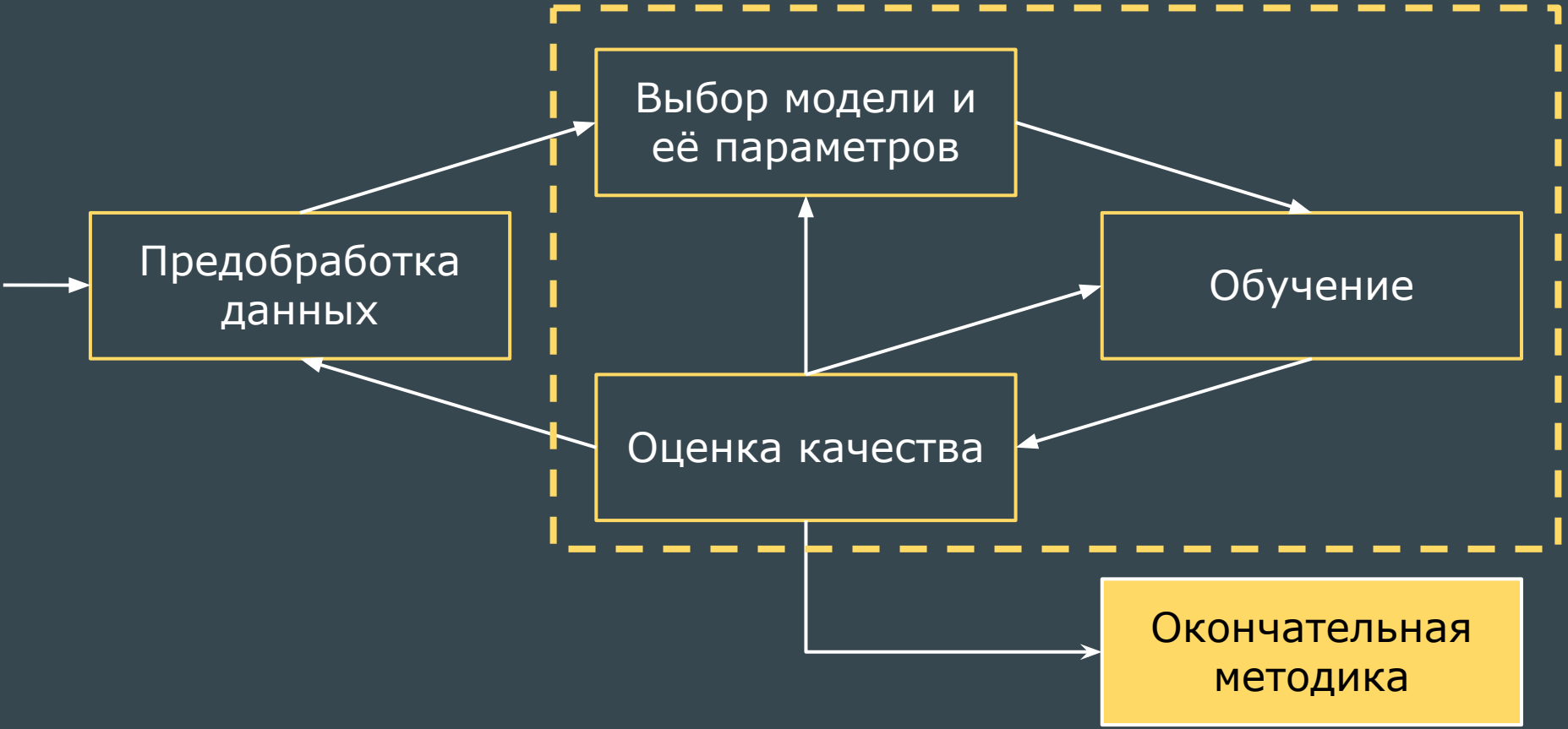
Random Forest. Параметры

+ количество деревьев в лесу
+ критерий расщепления
+ максимальное количество предикторов
+ параметры останова (глубина дерева, размер листа и т.д.)
+ параметр воспроизводимости

Поиск баланса



Workflow


Выбор модели и её параметров
Предобработка данных
Обучение
Оценка качества
Окончательная методика

Подведем итоги

+ решающее дерево -- отличная модель для порогового моделирования

+ ансамбли решающих деревьев работают хорошо и устойчиво

+ Random Forest -- серебряная пуля машинного обучения, с параметрами по умолчанию хорошо подходит для базового

+ Простота и быстрота реализации позволяют использовать полный поиск по сетке для оптимизации гиперпараметров

Примеры решения реальных задач

+ Исследование динамики уровня воды Аральского моря по данным дистанционного зондирования и климатического реанализа (ссылка)
+ Runoff calculations for ungauged river basins of the Russian Arctic region (ссылка)
+ Применение методов машинного обучения для моделирования толщины снежного покрова (ссылка)

Чем занять себя две недели?

Подготовка к #OpenDataHack от ECMWF

https://ecmwf-opendatahack.devpost.com/

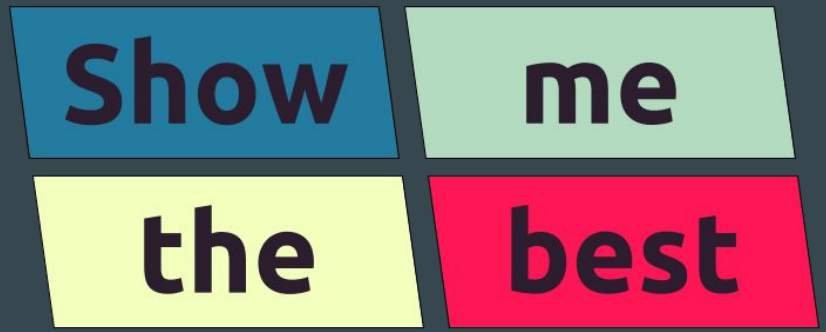
+ "GET OUT"
+ "GET CREATIVE"
+ "GET GEEKY"
+ "GET FUN"

На правах рекламы

Plovcast

новости науки глазами молодых ученых ИВП и ИО РАН

Show me the best



все яркие события из мира машинного обучения, python и науки

Важно

Вы можете помочь существенно улучшить этот курс!

- ayzelgv@gmail.com, hydrogo@yandex.ru
- vk.com/ayzelgv, facebook.com/ayzelgv
- ИВП РАН, кабинет 617