

КУРС «АНАЛИЗ ДАННЫХ В ВЕК ИИ»

Из хороших данных — хорошие модели

Preprocessing + Feature Engineering

Урок 15 · заключительное занятие



Модель видит не мир. Она видит наши признаки.

Почему подготовка данных иногда важнее выбора алгоритма?

СЫРЫЕ ДАННЫЕ

72 м² 3 комнаты
? минут до метро
дом 1999 года



ПОДГОТОВЛЕННЫЕ

пропуск обработан
масштабы сопоставлены
возраст = 27 лет

Одна и та же квартира. Разное представление для модели.

Почему данные приходят «грязными»?

Сырые данные почти никогда не бывают готовой таблицей для модели.

1

Пропуски

Человек не ответил на вопрос.

2

Ошибки

Измеритель или оператор ошибся.

3

Разные форматы

Системы по-разному хранят даты и категории.

Проблема начинается до обучения модели.

Знакомьтесь: таблица с проблемами

Учебные искусственные данные об аренде квартир.

Квартира	Площадь м ²	Комнат	До метро, мин	Аренда ₴(цель)
A	58	2	7	5 000
B	72	3	?	6 200
C	94	3	14	8 100
D	54	2	0	4 900

? — неизвестно • 0 — реальное значение • цены — целевая переменная

Что такое preprocessing — и зачем он нужен?

Предварительная обработка подготавливает данные к корректному анализу.

ПРЕПРОЦЕССИНГ = ПОДГОТОВКА ДАННЫХ

Сырые данные

Пропуски и ошибки

Разбираемся

Выбираем преобразования

Готовим признаки

Передаём данные модели

Цель — сохранить смысл данных, а не «улучшить» цифры.

Четыре типа подготовки

Preprocessing — не волшебная кнопка, а набор разных решений.

01

Пропуски

Что делать с
неизвестным?

02

Масштаб

Какие величины
сравнимы?

03

Категории

Как представить
текст и классы?

04

Признаки

Можно ли описать
объект лучше?

Выбираем то, что отвечает задаче и особенностям модели.

Пропуск — это не ноль

0 минут до метро и «неизвестно» — разные ситуации.

0

Значение известно

?

Значения нет

Подмена пропуска нулём может создать несуществующую закономерность.

Что делать с пропущенным значением?

Нет одного правильного способа для всех таблиц.

1

Сначала выяснить

Почему отсутствует значение?

2

Затем выбрать

Удалить, заполнить или оставить особый признак.

3

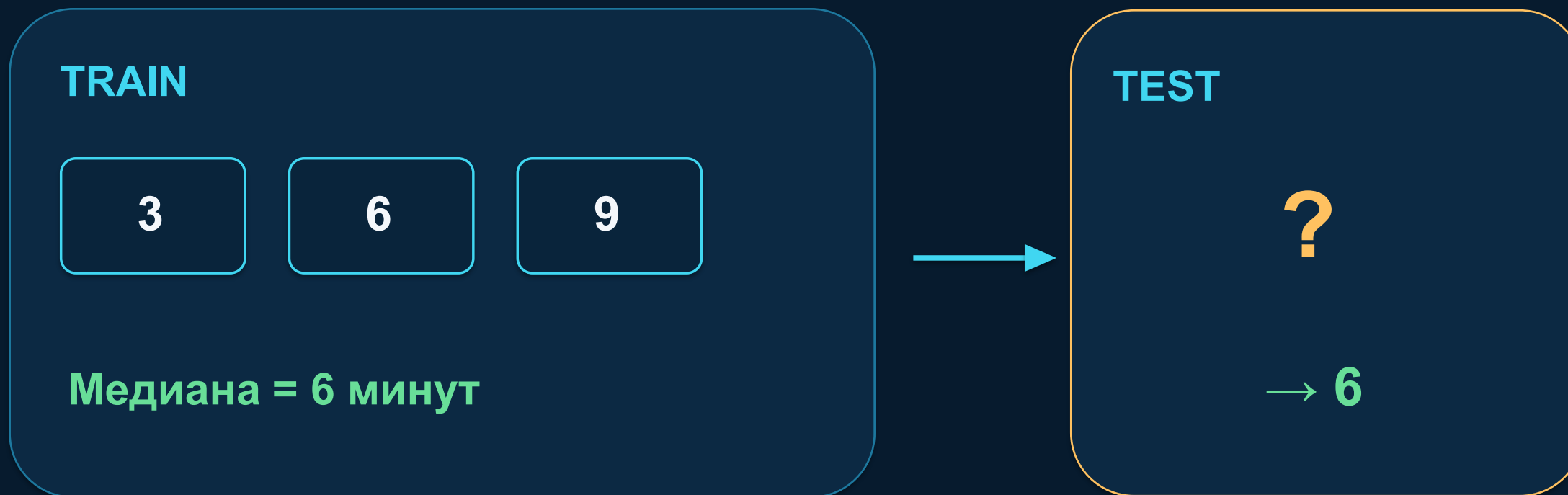
После проверить

Не исказили ли данные и оценку модели?

Не спешим заменять каждый «?» средним или нулём.

Чем заполнять пропуски?

Простой способ: медианой, посчитанной на обучающей части.



Тестовая часть не участвует в расчёте медианы.

Числа с разными масштабами

Большие числа могут «перекричать» маленькие в метрике расстояния.

Площадь, м²

120



Комнаты

3



Нормализация изменяет представление шкал, а не смысл квадратных метров.

Какой признак громче в расстоянии?

Для kNN числовой масштаб может незаметно решить исход сравнения.

РАЗНИЦА В ПЛОЩАДИ

30 м²

Большое число: 30

РАЗНИЦА В КОМНАТАХ

1 комната

Малое число: 1

Без нормализации единицы измерения влияют на расстояние.

Всем ли моделям нужна нормализация?

Вспоминаем старых знакомых: kNN, регрессию и деревья.

kNN

Часто да

Сравнивает расстояния

Регрессия

Зависит

Особенно при
регуляризации

Дерево / лес

Обычно нет

Ищут пороги по признакам

Предобработка зависит от выбранного алгоритма — это не универсальный рецепт.

Практика Orange: настройка Preprocess

Работаем с настоящим интерфейсом, не с нарисованным окном.

The screenshot displays the Orange3 data mining environment. On the left, a workflow is visible with the following components and connections:

- File** widget connected to **Data** widget.
- Data** widget connected to **Data Table** widget.
- Data** widget connected to **Preprocess** widget.
- Preprocess** widget connected to **Data Table (1)** widget.

On the right, the **Preprocess - Orange** widget settings window is open. It features a menu of preprocessors on the left and configuration options on the right.

Preprocessors list:

- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features** (highlighted)
- Randomize
- Remove Sparse Features
- Principal Component Analysis
- CUR Matrix Decomposition

Impute Missing Values section:

- ☒ Average/Most frequent
- ☐ Replace with random value
- ☐ Remove rows with missing values.

Normalize Features section:

- ☒ Standardize to $\mu=0, \sigma^2=1$
- ☐ Center to $\mu=0$
- ☐ Scale to $\sigma^2=1$
- ☐ Normalize to interval $[-1, 1]$
- ☐ Normalize to interval $[0, 1]$

At the bottom of the settings window, there is a checkbox for **Apply Automatically** (checked) and a status bar showing **303** rows and **303** features.



Statistical Summary

See the story first. Open the spreadsheet later.

98/100 - Healthy

303

Rows

14

Variables

0

Missing cells

10

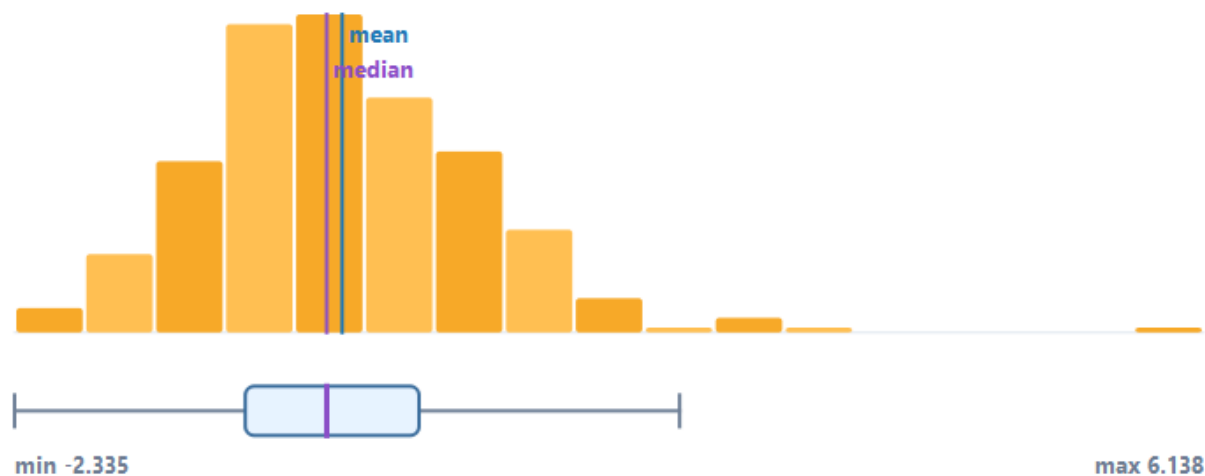
Quality notes

Explore variable:

cholesterol

feature · numeric

Distribution and box plot · cholesterol



Typical value

Mean 2.4183e-16

Median -0.1101 · Multiple modes: 3 values
-0.961341, -0.825922, -0.245555

Spread

SD 1.0017

IQR 1.2381 · MAD 0.6191

Range

-2.3349 → 6.1385

Middle 90%: -1.385 → 1.5516

Selected variable quality

0% missing

5 possible outliers · 152 unique

Dr Igor's quick reading: Moment skewness suggests a strong right tail. 5 values lie outside the selected IQR fences. 2 automatic note(s) for this variable. Unusual does not automatically mean wrong.



Statistical Summary

See the story first. Open the spreadsheet later.

98/100 - Healthy

303

Rows

14

Variables

0

Missing cells

10

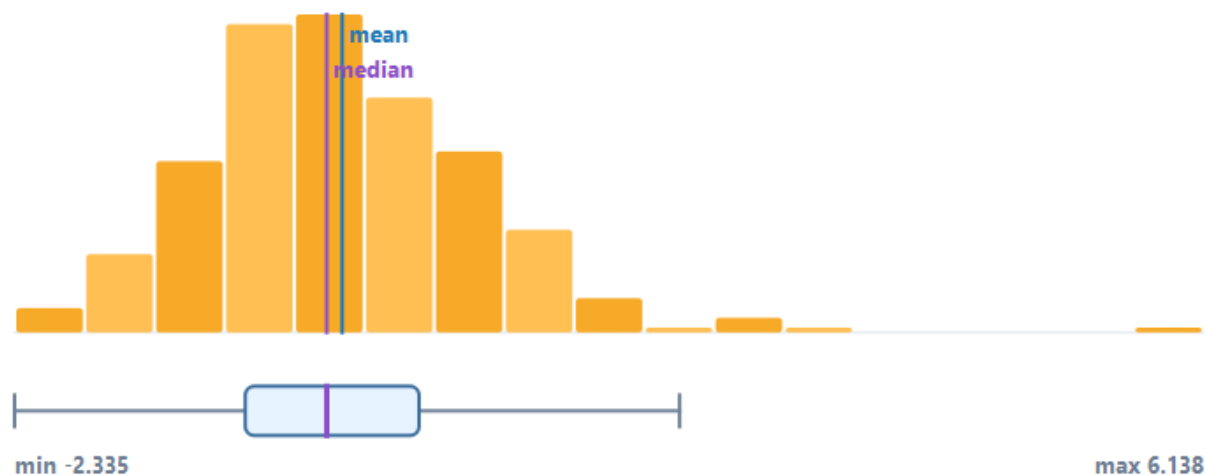
Quality notes

Explore variable:

cholesterol

feature · numeric

Distribution and box plot · cholesterol



Typical value

Mean $2.4183e-16$

Median -0.1101 · Multiple modes: 3 values
-0.961341, -0.825922, -0.245555

Spread

SD 1.0017

IQR 1.2381 · MAD 0.6191

Range

-2.3349 → 6.1385

Middle 90%: -1.385 → 1.5516

Selected variable quality

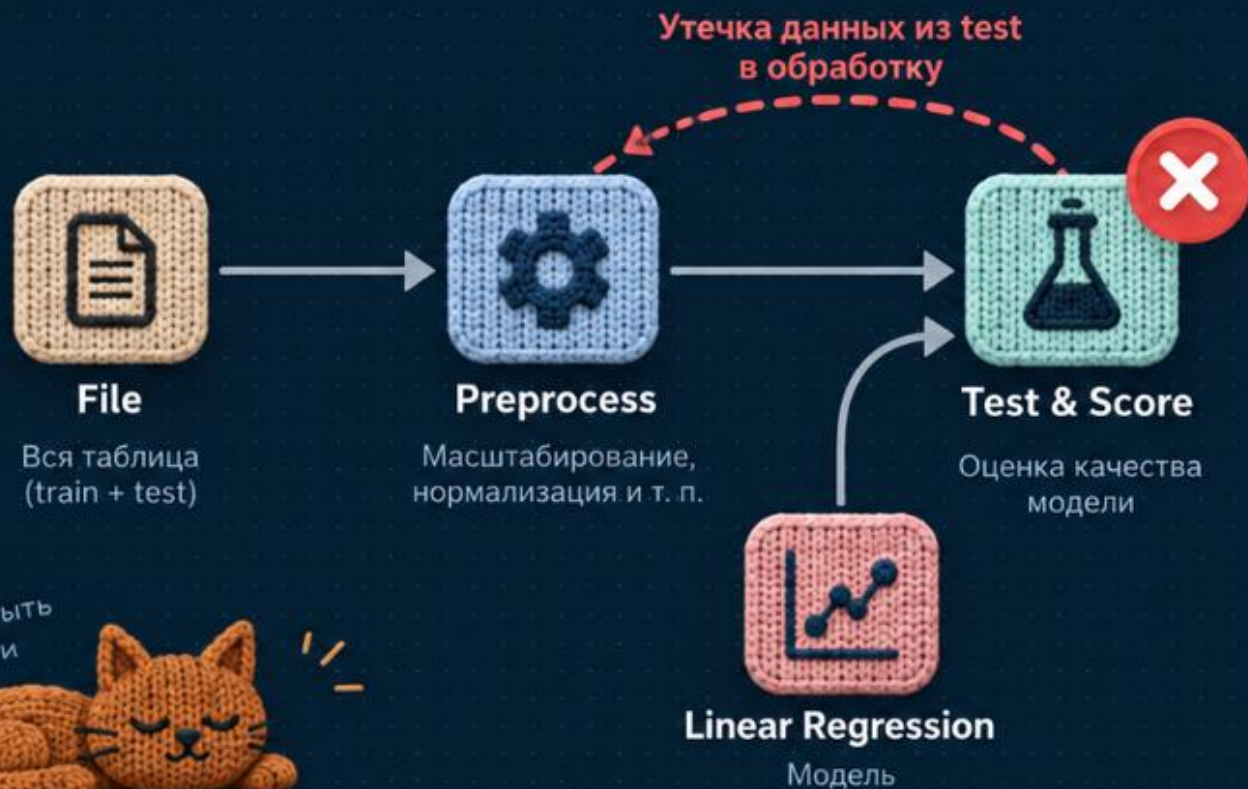
0% missing

5 possible outliers · 152 unique

Dr Igor's quick reading: Moment skewness suggests a strong right tail. 5 values lie outside the selected IQR fences. 2 automatic note(s) for this variable. Unusual does not automatically mean wrong.

Ловушка: сначала обработали всю таблицу

Так в проверку незаметно просачивается информация из test.



Почему это плохо?

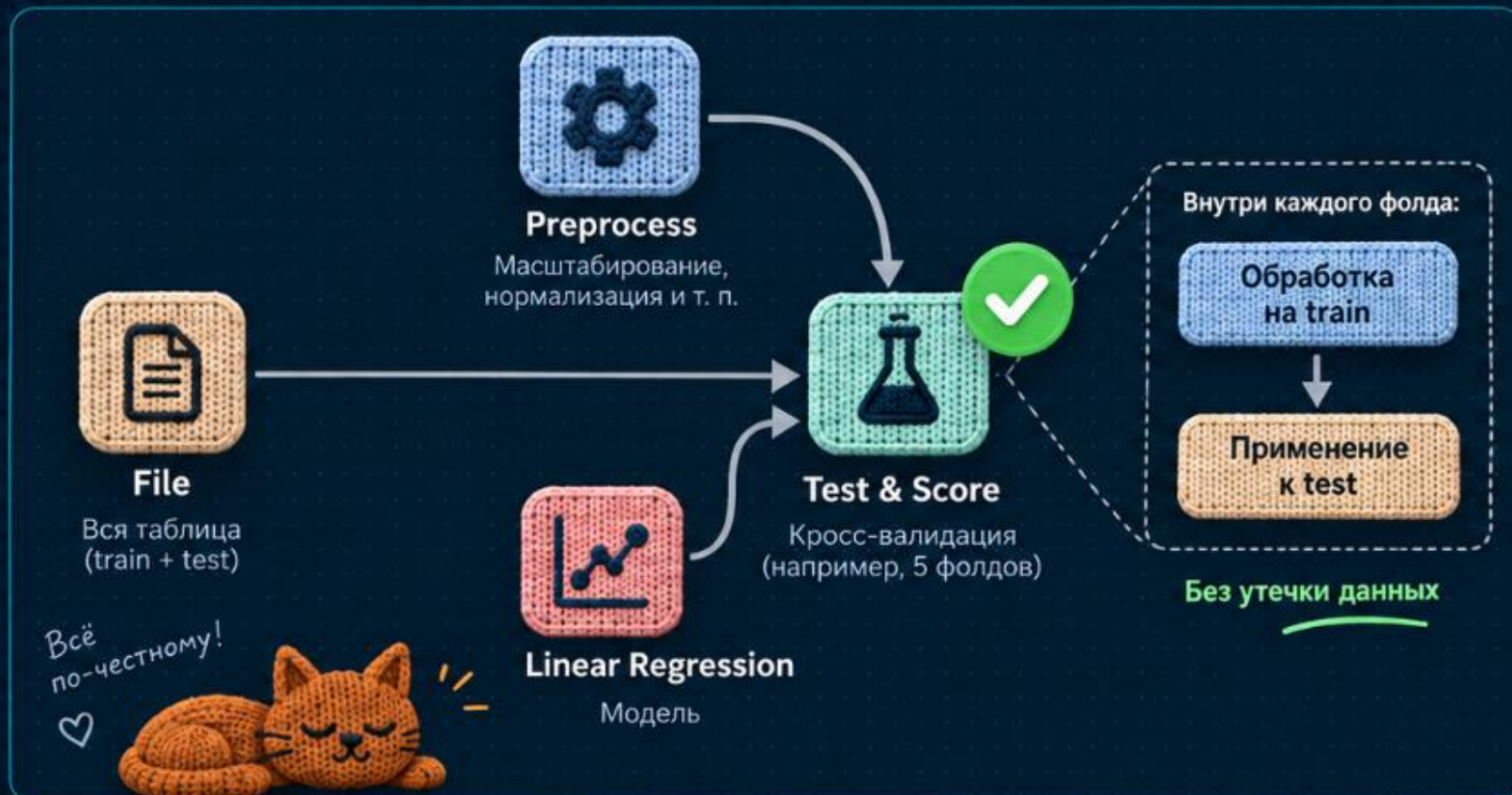
- средние и масштабы считаются по всей таблице
- test уже повлиял на обработку
- метрика становится слишком оптимистичной



Сначала обработали всё — потом устроили экзамен. Это уже нечестно.

Правильно: передаём Preprocess в Test & Score

Тогда обработка обучается только на train внутри каждого фолда.



Что происходит внутри?

- train и test сначала разделяются
- обработка учится только на train
- к test применяются уже готовые настройки



Честная схема: сначала split, потом preprocessing внутри обучения.

Как подготовка может подсмотреть экзамен?

Неверный порядок действий иногда делает метрику слишком оптимистичной.

НЕЧЕСТНЫЙ ПУТЬ

Сначала считаем медиану по ВСЕМ данным.

Потом делим на train/test.

ЧЕСТНЫЙ ПУТЬ

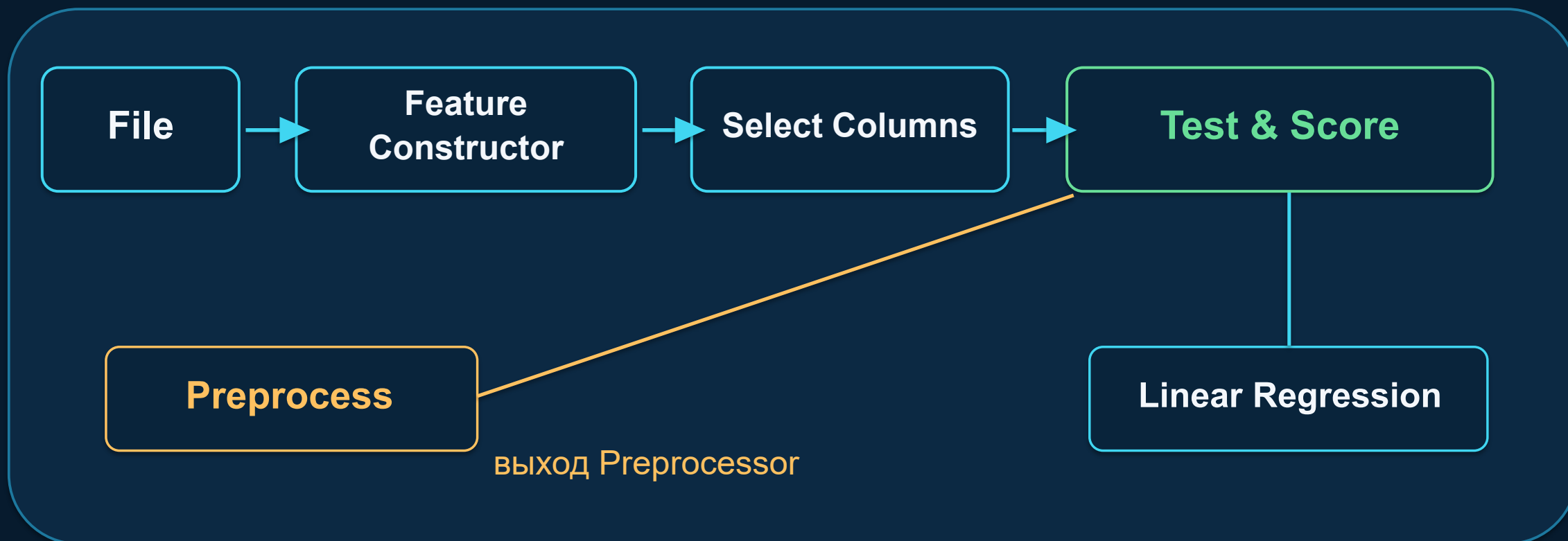
Сначала отделяем test.

Медиану считаем только на train.

Тестовые наблюдения не должны обучать preprocessing.

Честное сравнение: preprocessing внутри CV

В Orange подгоняемые преобразования применяются в каждом обучающем фолде.



Назначьте `rent_ils` целью. Обработку подгоняем внутри обучения.

Практика: сравним «до» и «после»

Запускаем обе версии на одинаковой процедуре оценки.

БЕЗ ДОПОЛНИТЕЛЬНОЙ ОБРАБОТКИ

СКРИНШОТ Test & Score

вставить реальные RMSE / MAE

**С PREPROCESS + НОВЫМИ
ПРИЗНАКАМИ**

СКРИНШОТ Test & Score

вставить реальные RMSE / MAE

Цифры появятся только после настоящего эксперимента.

Подводный камень: утечка данных

Даже правильный алгоритм с неправильной схемой даст слишком красивую оценку.

НЕПРАВИЛЬНО

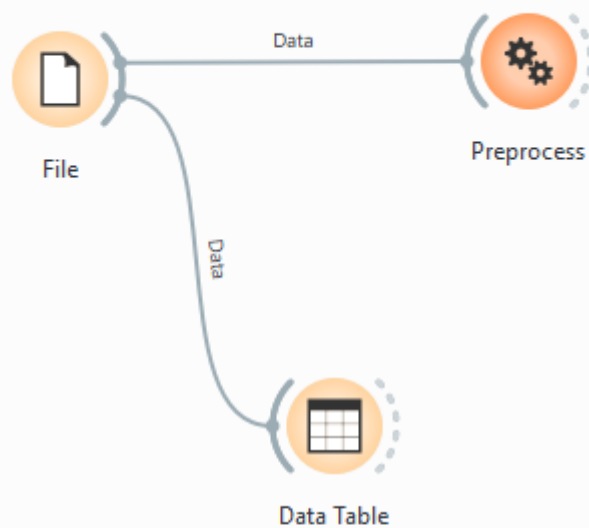
**File → Preprocess Data
→ Test & Score**

Статистики посчитаны на всех данных

ПРАВИЛЬНО

**File → Test & Score
Preprocess → Preprocessor**

Обработка обучается внутри фолда



Preprocess - Orange

File View Window Help

Preprocessors

- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Remove Sparse Features
- Principal Component Analysis
- CUR Matrix Decomposition



Apply Automatically



303



303

А если чистых данных всё равно недостаточно?

Иногда важная информация спрятана в сочетании двух столбцов.

КВАРТИРА А

72 м² • 3 комнаты

24 м² на комнату

КВАРТИРА Б

72 м² • 1 комната

72 м² на комнату

Новый признак = новый взгляд на те же данные.

А теперь создаём новые признаки

Feature Engineering — превращаем знания о задаче в новые столбцы.

Площадь на комнату

area_m2 / rooms

$$72 / 3 = 24 \text{ м}^2$$

Возраст здания

2026 - year_built

$$2026 - 1999 = 27 \text{ лет}$$

Новый признак — новая гипотеза

Он может помочь, но хорошим становится только после проверки.

1

Смысл

Что этот признак
измеряет?

2

Доступность

Знаем ли его при
прогнозе?

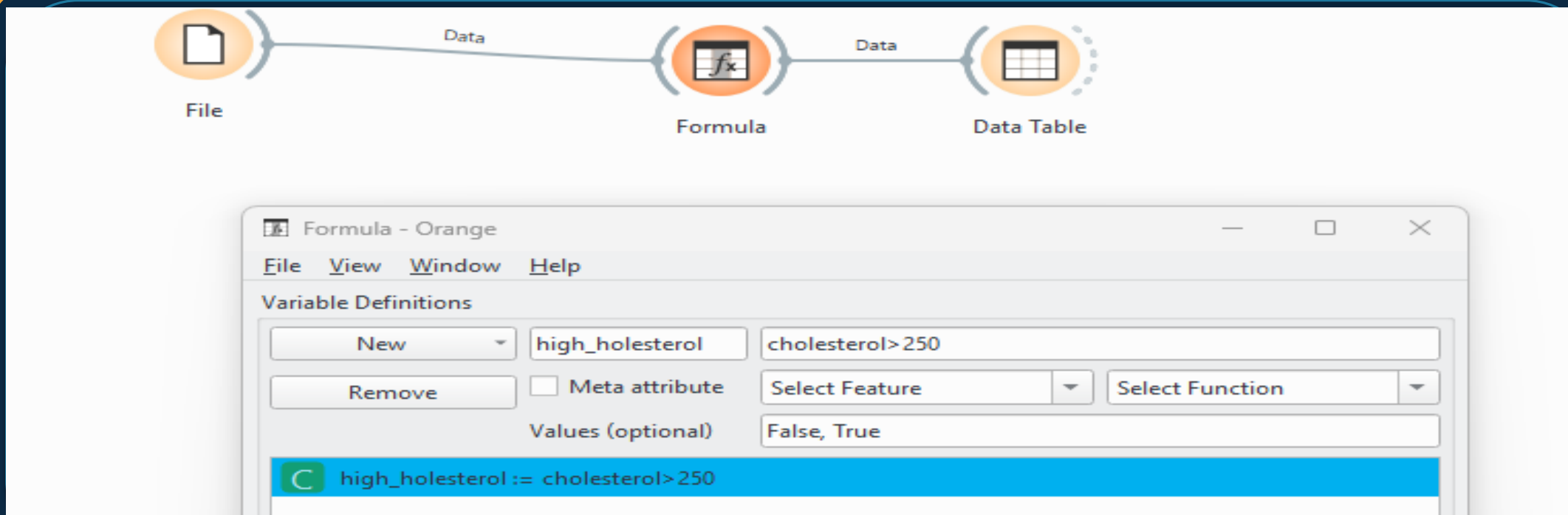
3

Проверка

Помогает ли на новых
данных?

Практика Orange: Feature Constructor

Создадим два столбца и увидим их в Data Table.



Особенно опасный «прекрасный признак»

Предсказываем аренду квартиры. Можно ли взять цену за квадратный метр?

rent_ils / area_m2

Этот признак уже содержит неизвестную аренду!

Идеальная метрика может оказаться обычной подсказкой ответа.

Совет от доктора Игоря

Новая колонка — не достижение. Полезная и честная колонка — достижение.

Три проверки любого нового признака

СМЫСЛ

что измеряет?

ЧЕСТНОСТЬ

можно знать до ответа?

ПОЛЬЗА

стало лучше на новых
данных?

Вопрос

Когда правильно вычислять медиану для заполнения пропусков?

?

- A.** Один раз по всей таблице перед проверкой
- B.** Отдельно на обучающей части каждого разбиения
- C.** Только после того, как увидели тестовую ошибку

Ответ

Медиана — тоже параметр, который мы оцениваем по данным.



В. Только по обучающей части.

Вопрос

Что нельзя использовать для предсказания аренды?

?

A. Площадь квартиры, делённую на число комнат

B. Возраст здания по году постройки

C. Аренду, делённую на площадь квартиры

Ответ

Самый «предсказательный» признак иногда оказывается утечкой.



С. `rent_ils / area_m2` уже содержит целевую цену.

Карта части 15.1

Четыре правила, которые стоит унести с собой.

01 Проверить данные

ноль, пропуск, категории, аномалии

02 Подготовить к модели

масштаб и заполнение — не для всех одинаково

03 Создать признаки

только со смыслом и без подсказанного ответа

04 Сравнить честно

обработка и подбор — внутри обучения

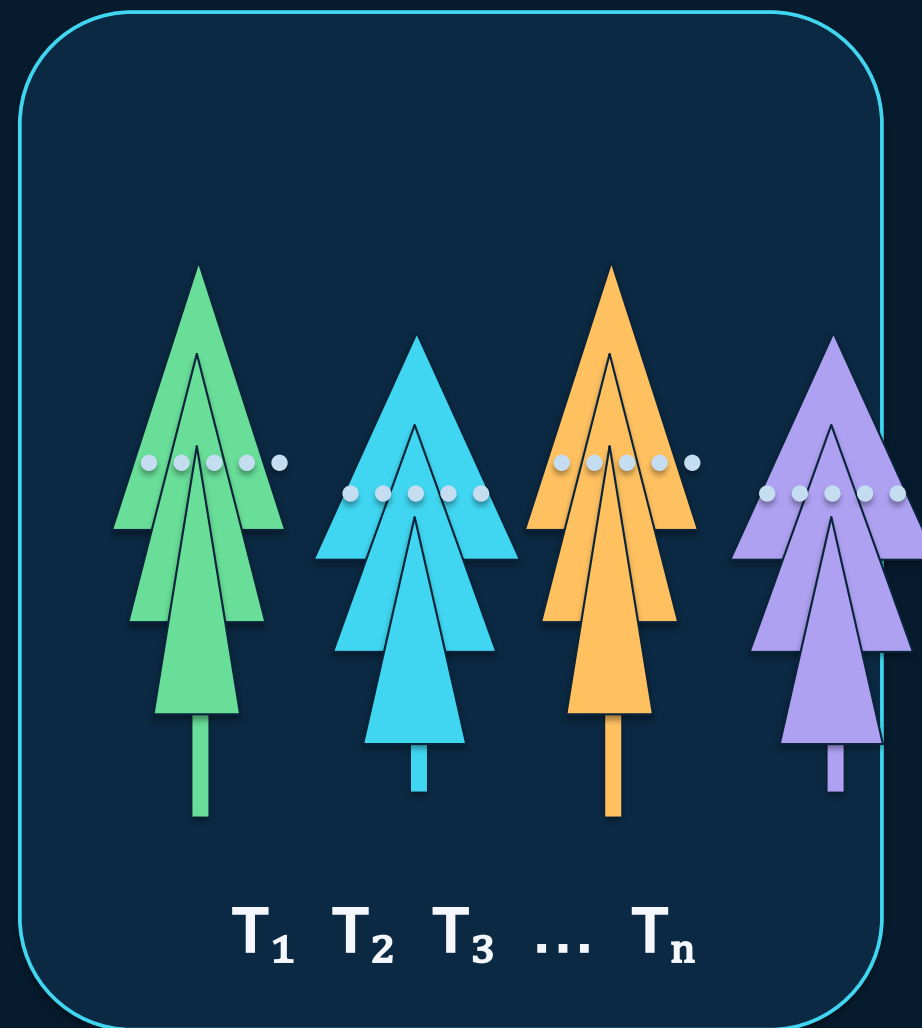
Следующий блок: **Random Forest** — когда один алгоритм превращается в команду.

КУРС «АНАЛИЗ ДАННЫХ В ВЕК ИИ»

ЧАСТЬ 15.2

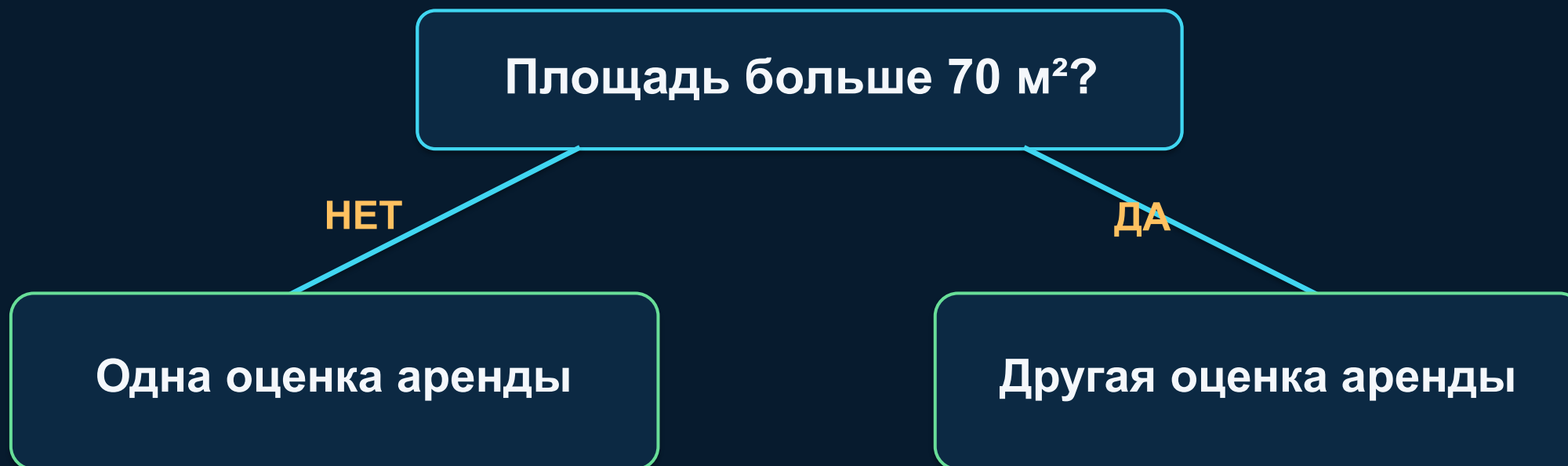
Одно дерево хорошо. А целый лес?

Random Forest: мудрость ансамбля без
магических обещаний.



Напомним: дерево задаёт вопросы

Одно дерево строит цепочку правил, ведущую к прогнозу.



Дерево понятно. Но его правила зависят от обучающих данных.

Один эксперт может ошибиться

Одно дерево чувствительно к случайным особенностям выборки.

Дерево иногда слишком хорошо запоминает учебные примеры.

Почему одно дерево может колебаться?

Небольшая замена наблюдений — и лучший вопрос уже другой.

ВЫБОРКА № 1

Важнее оказалось расстояние до метро.

Первый вопрос дерева — о метро.

ВЫБОРКА № 2

Важнее оказалась площадь.

Первый вопрос дерева — о площади.

Одно дерево может сильно зависеть от деталей выборки.

Три дерева — три мнения

Три разных обученных дерева могут предсказывать по-разному.

Одно дерево

Клиент → «кредит одобрить»

Одно решение может быть
неустойчивым.

Три дерева

Да · Да · Нет

Голосование → «Да»

Сто одинаковых деревьев — не мудрость толпы

Чтобы ансамбль помог, деревья должны отличаться друг от друга.

КОПИИ

100 одинаковых деревьев
→ 100 одинаковых ошибок

РАЗНЫЕ ДЕРЕВЬЯ

Разные обучающие строки и
кандидаты-признаки
→ ошибки могут различаться

Разнообразие деревьев помогает усреднять шум.

Почему деревья оказываются разными?

Random Forest добавляет случайность двумя способами.

Наблюдения

Bootstrap: случайные строки с возвращением

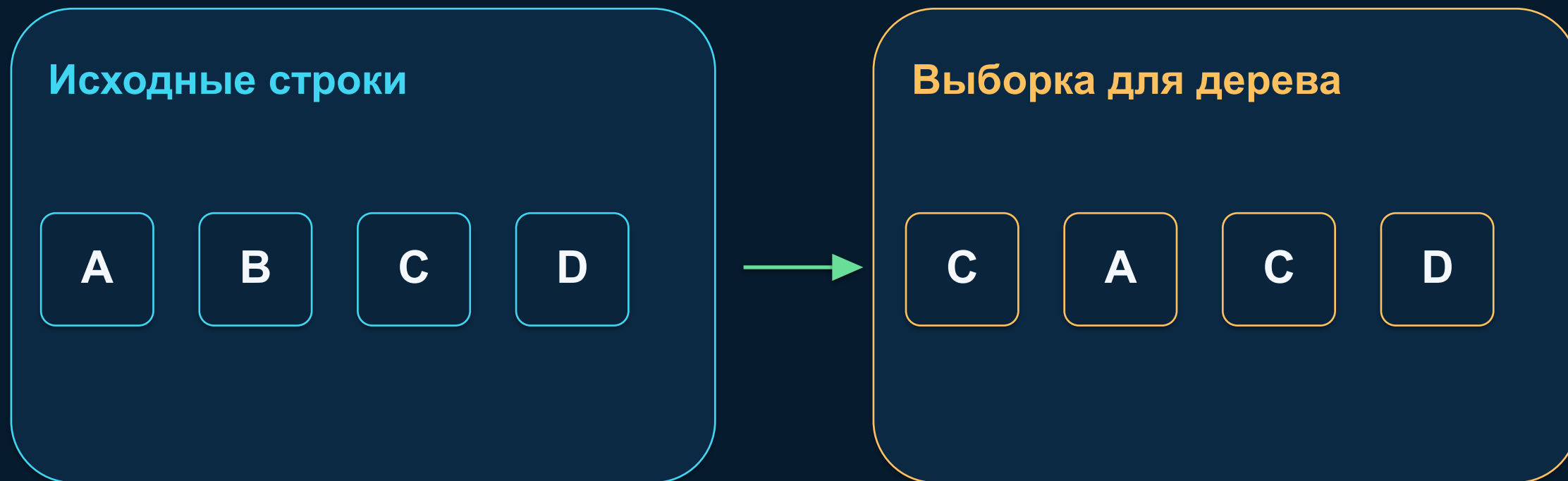
Признаки

Случайное подмножество кандидатов на разбиение

Так мы снижаем сходство деревьев, а затем объединяем их ответы.

Bootstrap: берём строки с возвращением

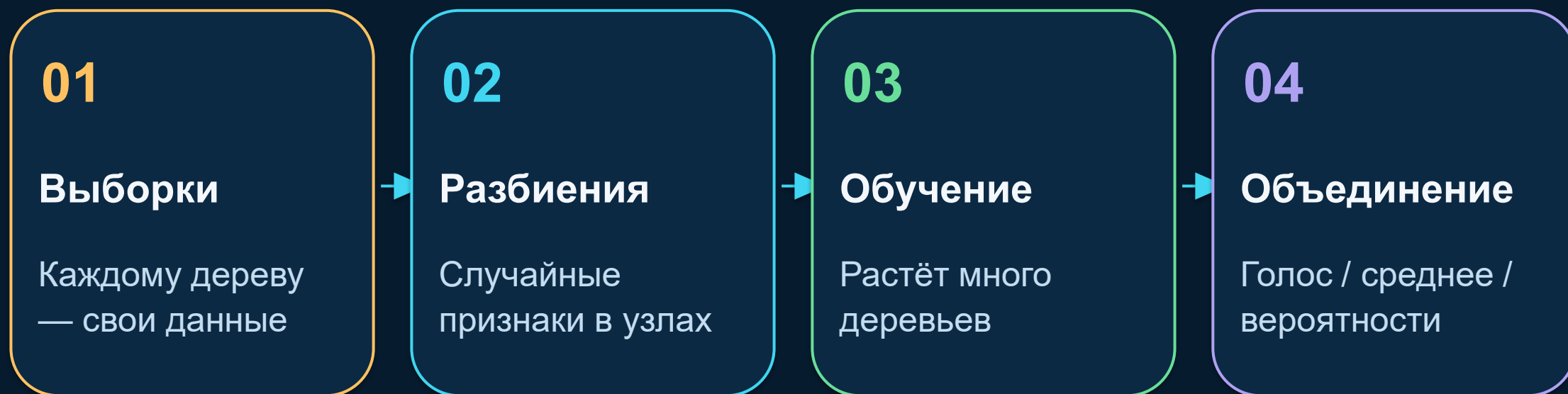
Одна строка может попасть дважды, другая — ни разу.



В не попала, С повторилась — это нормально.

Как лес принимает решение?

Четыре этапа вместо одного гигантского дерева.



Классификация и регрессия

Лес умеет отвечать на два знакомых типа вопросов.

Классификация

Класс А, В или С?

Собираем решения деревьев или усредняем оценки вероятностей.

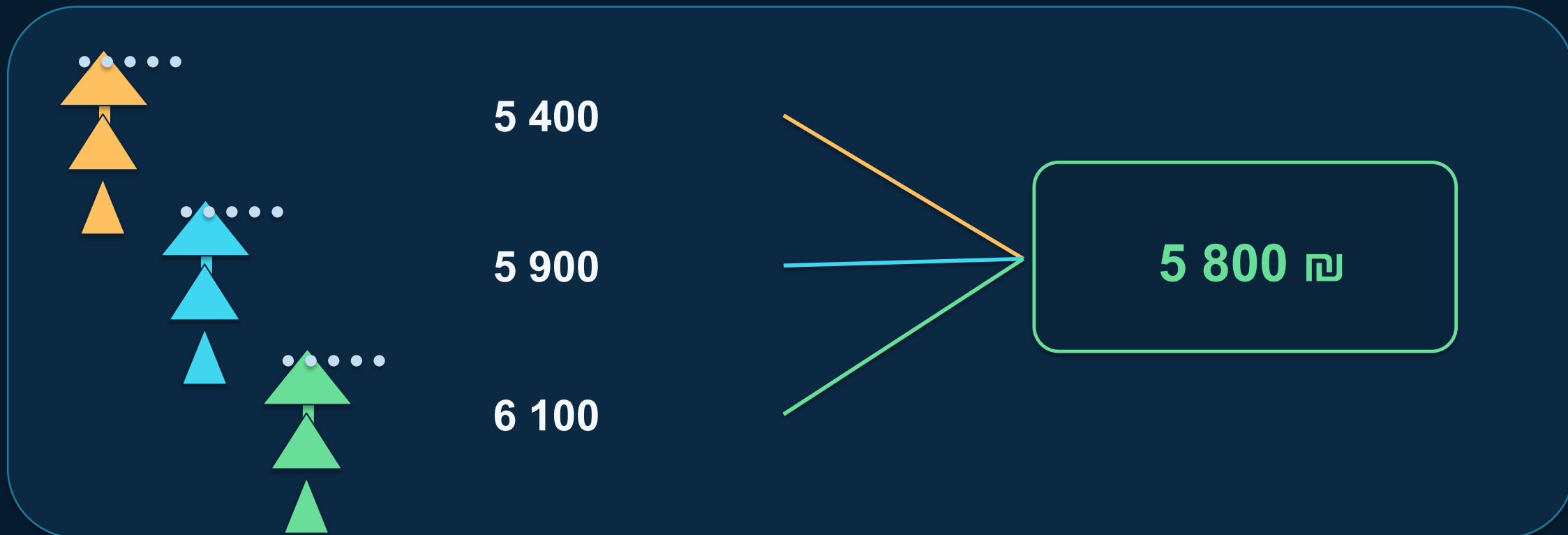
Регрессия

Аренда: 5400, 5900, 6100 ₺

Среднее = 5800 ₺

Почему лес часто устойчивее?

Ошибки отдельных деревьев могут частично компенсироваться.



За что мы платим качеством ансамбля?

Лес гибок, но не так прозрачен, как отдельное дерево.

Одиночное дерево

Можно пройти от корня к листу и увидеть правило решения.

Часто легче объяснить.

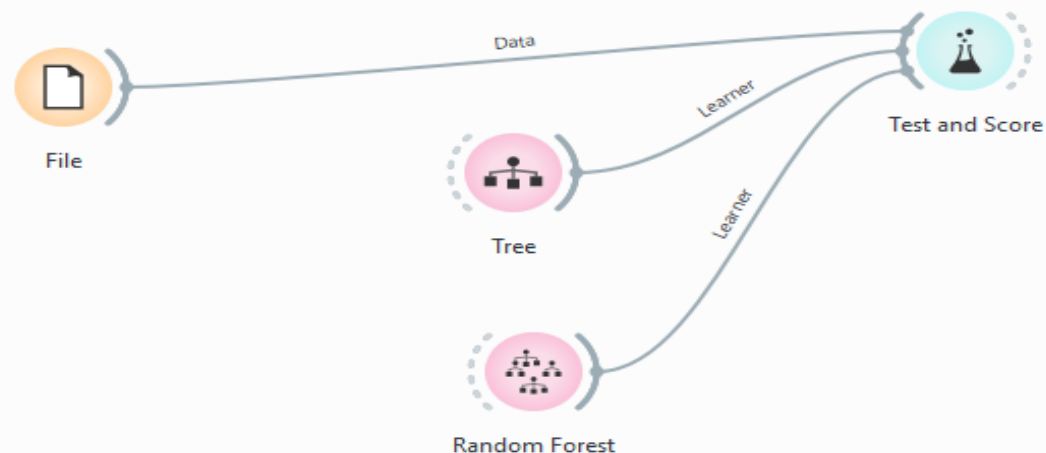
Случайный лес

Много разных правил, итог агрегируется.

Часто точнее, но объяснять сложнее.

Orange: дерево против леса

Проведём честный эксперимент на одном наборе данных.



Test and Score - Orange

File Edit View Window Help

☒ Cross validation
Number of folds: 5
☐ Stratified

☐ Cross validation by feature
[]

☐ Random sampling
Repeat train/test: 10
Training set size: 66 %
☒ Stratified

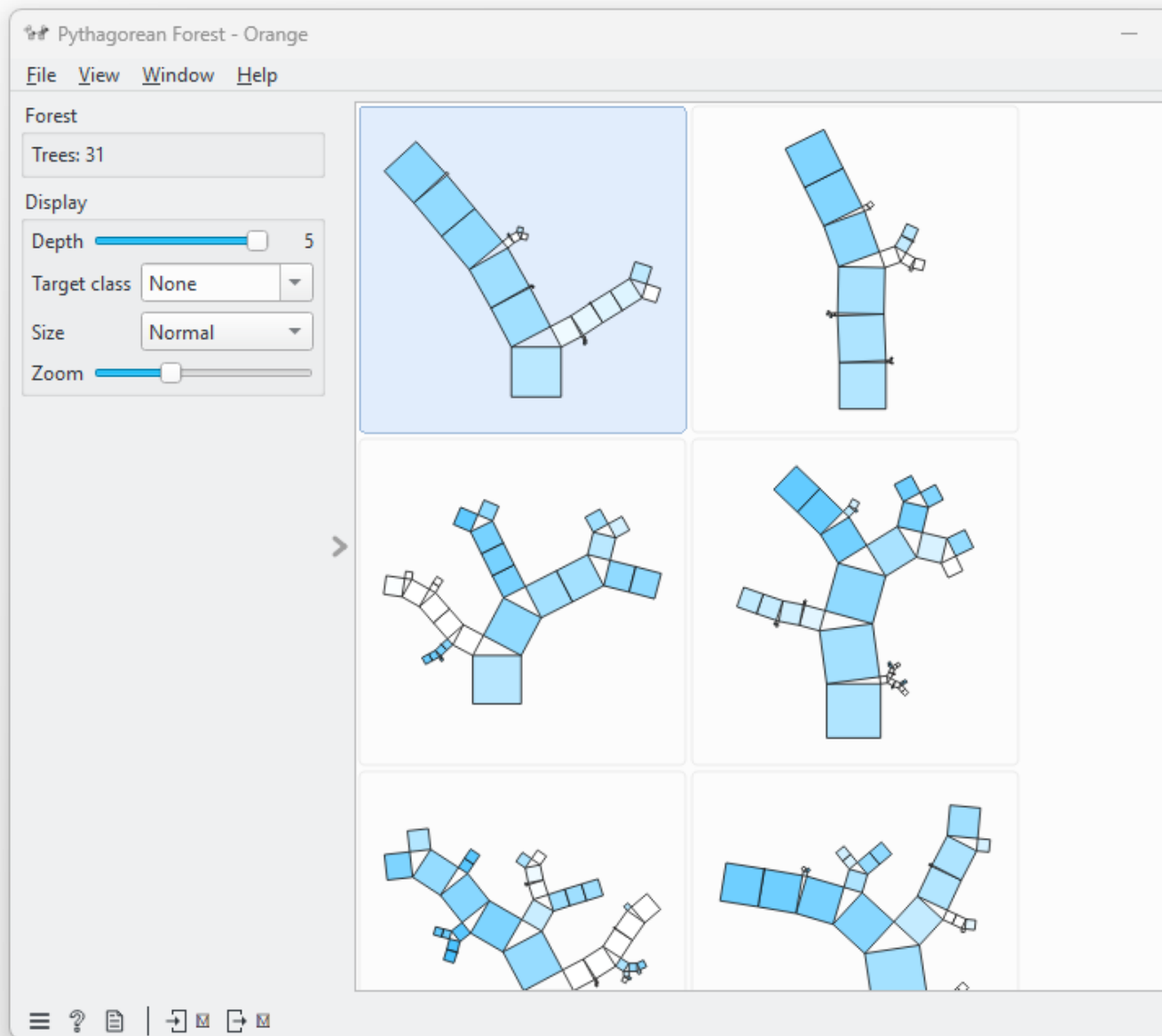
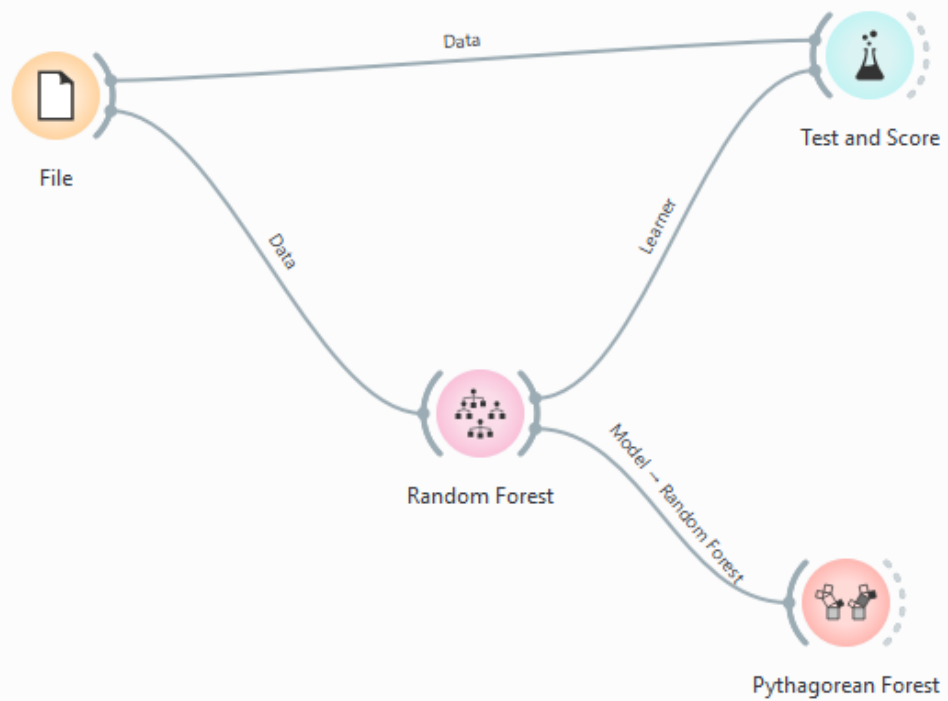
☐ Leave one out
☐ Test on train data
☐ Test on test data

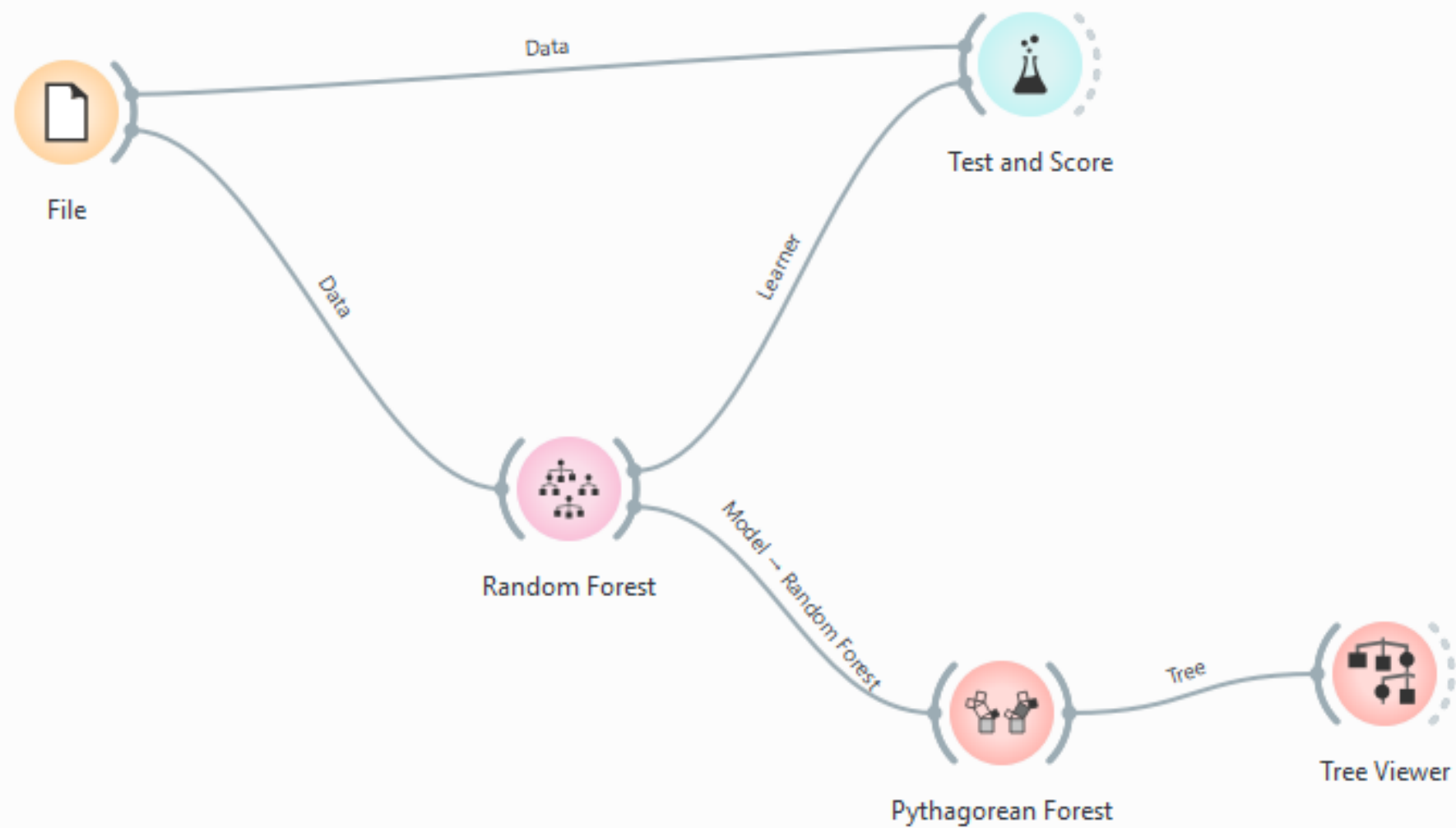
Evaluation results for target (None, show a)

Model	AUC
Tree	0.686
Random Forest	0.712

Compare models by: Area und []

Tree Random





Три настройки, которые важно понимать

В интерфейсе Orange начинаем с простых параметров.

1

Trees

Сколько деревьев
построить?

2

Depth

Насколько сложное
каждое дерево?

3

Random seed

Как повторить
эксперимент?

Подводные камни

Лес — не гарантия победы.

ЛОВУШКА

«1000 деревьев всегда лучше 100».

ПРОВЕРКА

Сравнить качество, время и стабильность на CV.

Совет преподавателя

Сначала простая модель, затем усложнение.

1 Baseline

Что получится без леса?

2 CV

Лучше ли лес на новых данных?

3 Смысл

Стоит ли усложнение выигрыша?

Вопрос

Что делает случайный лес отличным от одного дерева?

?

A. Просто повторяет одно и то же дерево

B. Обучает разные деревья и объединяет их ответы

C. Всегда использует все признаки в каждом узле

Ответ

Лес работает благодаря разным деревьям, а не их простому количеству.



В. Разнообразие + объединение.

Вопрос

Чему равен прогноз леса для регрессии?

?

A. Самому большому прогнозу дерева

B. Среднему прогнозу деревьев

C. Числу деревьев, умноженному на RMSE

Ответ

Для численной цели лес объединяет ответы деревьев средним.



В. Обычно усредняем прогнозы.

Карта части 15.2

Случайный лес в четырёх мыслях.

ИДЕЯ

Много разных деревьев вместо одного

СЛУЧАЙНОСТЬ

Bootstrap и случайные кандидаты признаков

ОТВЕТ

Классификация: агрегация; регрессия: среднее

ПРОВЕРКА

Качество и сложность оцениваем на новых данных

КУРС «АНАЛИЗ ДАННЫХ В ВЕК ИИ»

ЧАСТЬ 15.3

ИИ строит НОВЫЙ инструмент

От подсказки в чате — к работающему коду в Orange.

```
from Orange.data import  
Table  
out_data = transform(...)  
assert valid(out_data)
```

ИИ предлагает → человек
проверяет

Проблема: нужного виджета нет

А если нужна операция, которой нет в меню?

Мы можем описать её словами и попросить ИИ написать код.

Почему нам вообще нужен свой инструмент?

Одну и ту же операцию неудобно повторять вручную.

1

Есть задача

Посчитать новые
признаки для сотен
квартир.

2

Нет готового блока

Нужной операции нет
среди виджетов.

3

Помогает ИИ

Пишет код — мы
проверяем на тестах.

Важен не код сам по себе, а воспроизводимый результат.

Три роли в новой команде

Человек + ИИ + Orange.

1

Аналитик

Определяет смысл и проверки.

2

ИИ

Пишет и объясняет код.

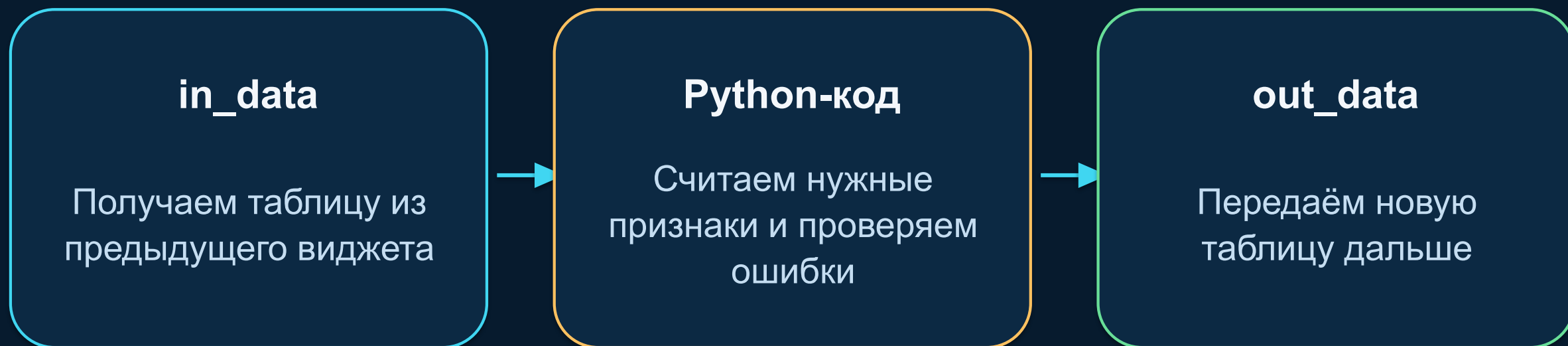
3

Orange

Выполняет и показывает результат.

Что делает Python Script внутри Orange?

Это обычный узел рабочего процесса: вход → вычисление → выход.



Связь с Orange важнее красивой программы.

Два пути расширения Orange

Не путать скрипт внутри рабочего процесса с новым виджетом.

Python Script

Открыли готовый виджет →
вставили код → получили новую
таблицу.

Быстрый учебный путь.

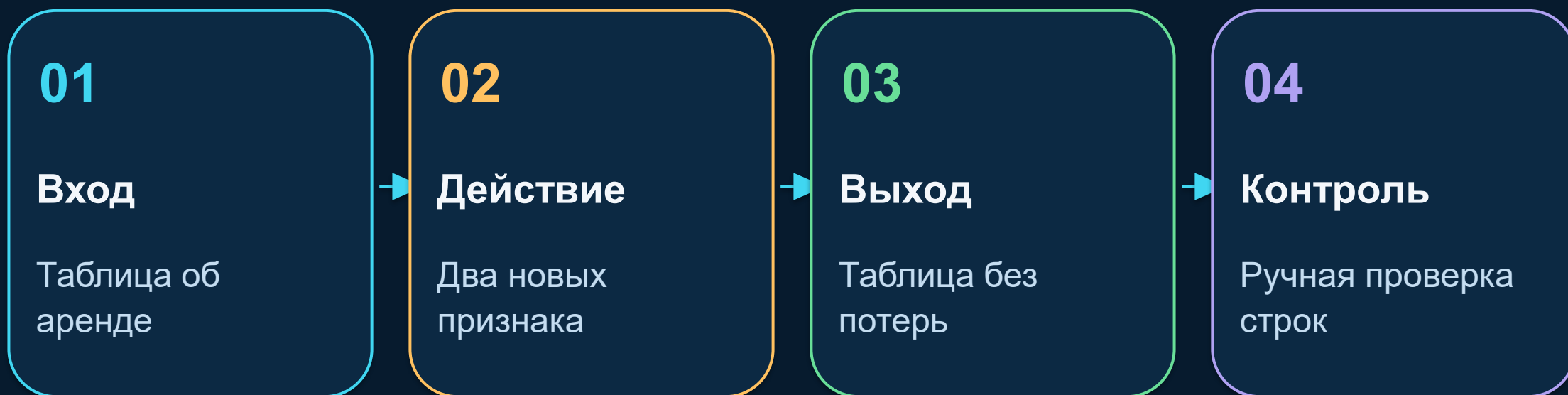
Собственный Widget

Python-класс + входы/выходы +
интерфейс + установка add-on.

Более сложный путь.

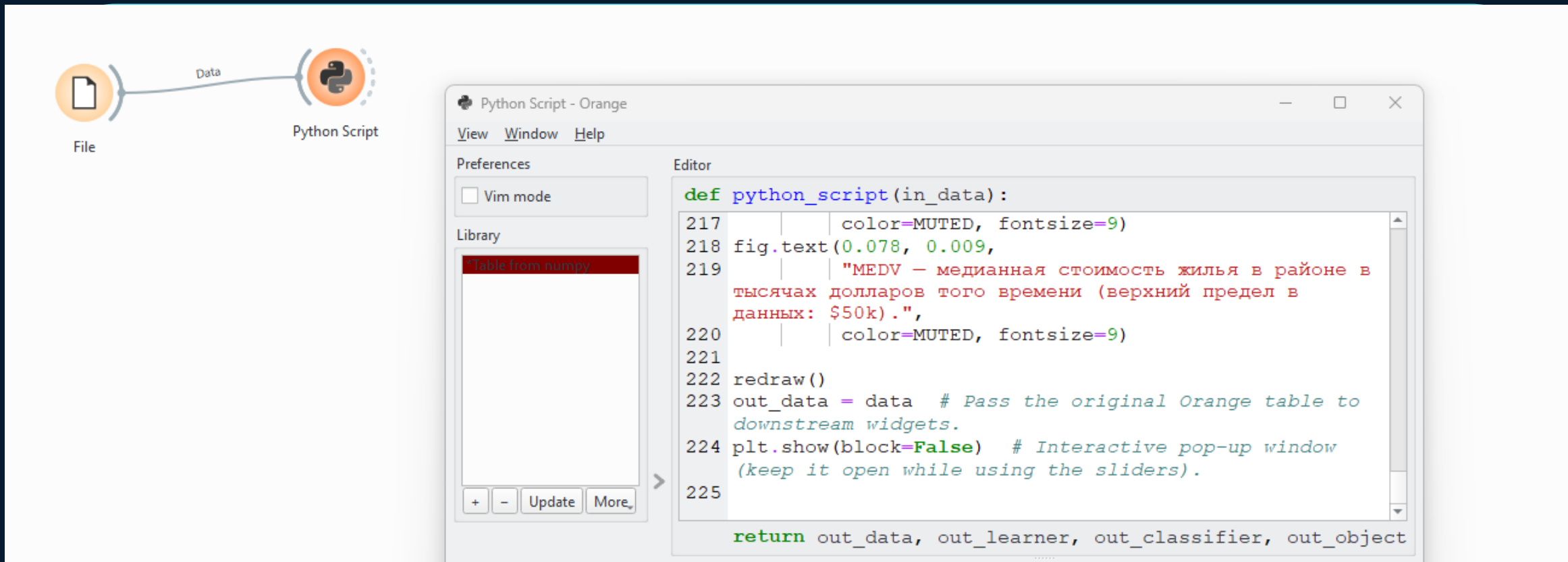
Что поручим ИИ?

Учебный пример продолжает проект с квартирами.



Хороший запрос к ИИ — половина работы

Не «напиши что-нибудь умное», а проверяемая спецификация.

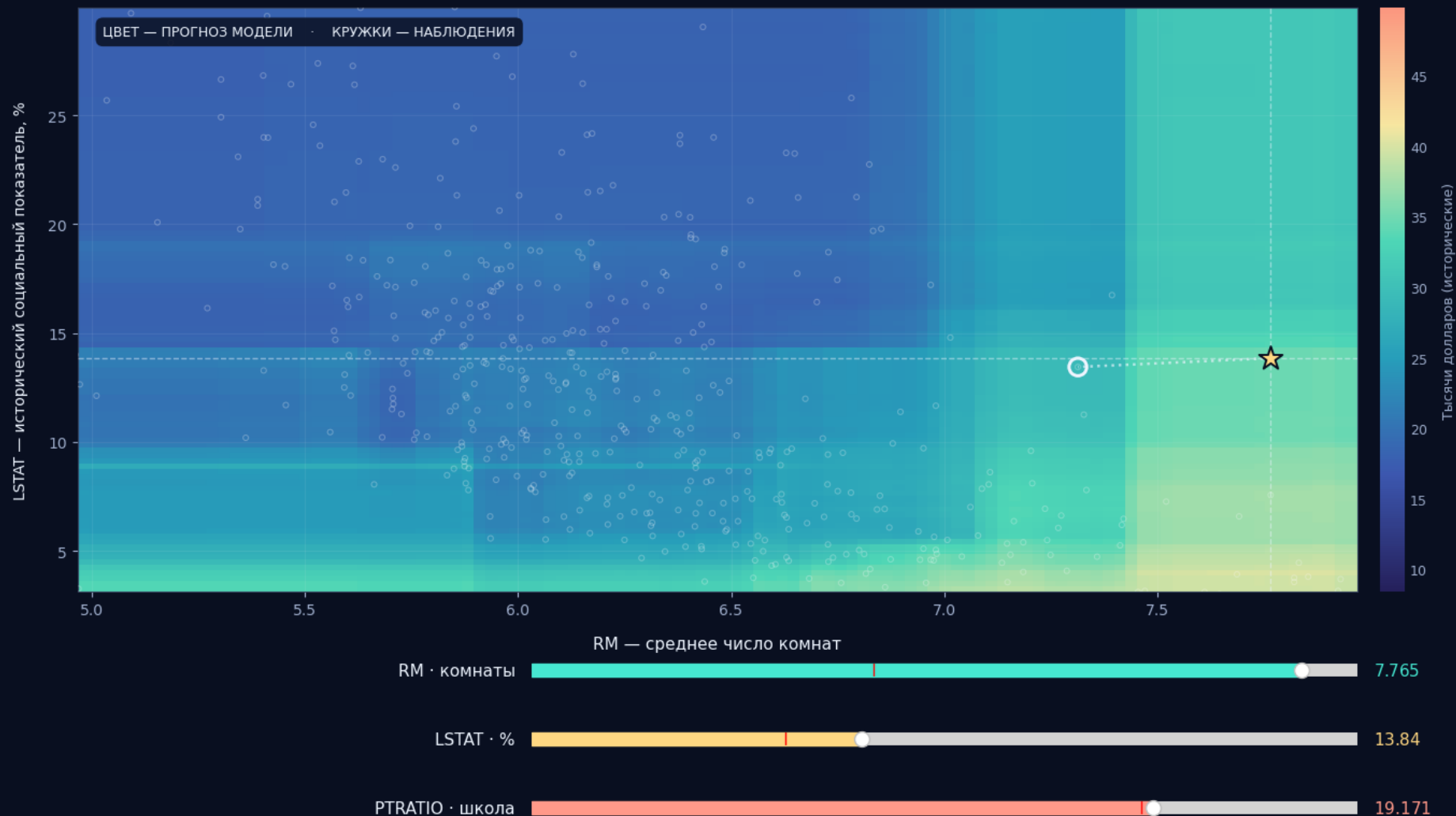


```
def python_script(in_data):
    217     color=MUTED, fontsize=9)
    218     fig.text(0.078, 0.009,
    219             "MEDV — медианная стоимость жилья в районе в
тысячах долларов того времени (верхний предел в
данных: $50k).",
    220             color=MUTED, fontsize=9)
    221
    222     redraw()
    223     out_data = data # Pass the original Orange table to
downstream widgets.
    224     plt.show(block=False) # Interactive pop-up window
(keep it open while using the sliders).
    225

    return out_data, out_learner, out_classifier, out_object
```

BOSTON / ЛАБОРАТОРИЯ ЦЕН

Как модель оценивает медианную стоимость жилья? Меняйте ползунки и ищите неожиданные сочетания.



ВАШ ГИПОТЕТИЧЕСКИЙ РАЙОН

\$34.7k

+13.5k к медиане выборки (\$21.2k)

БЛИЖАЙШИЙ РЕАЛЬНЫЙ РАЙОН

\$15.0k

КАЧЕСТВО НА ТЕСТЕ (25%)

R² 0.75

MAE \$2.7k

Random Forest · 3 признака
Обучение: 379 · Тест: 127

СБРОС

Когда скрипта уже недостаточно?

Рабочий код можно превратить в собственный виджет для многократного использования.

PYTHON SCRIPT

Код внутри проекта.

Нужно открывать и менять текст скрипта.

СВОЙ ВИДЖЕТ

Удобные настройки и входы/выходы.

Можно использовать повторно.

Сначала доказываем работу скрипта — потом строим интерфейс.

Как попросить ИИ создать рабочий виджет Orange?

От хорошего промта — к рабочему add-on и установке в Orange.



1

Чётко ставим задачу

Описываем, что должен делать виджет, примеры и требования



2

Просим полный код add-on

Запрашиваем структуру проекта, код виджета и метаданные



3

Проверяем локально

Устанавливаем в отдельном окружении, исправляем ошибки



4

Устанавливаем в Orange

Через Extensions или pip (инструкция от ИИ)



5

Тестируем на данных

Проверяем, что виджет работает и даёт ожидаемый результат

От идеи к работающему виджету!



Что хотим получить?

- виджет с GUI
- входы и выходы Orange
- инструкцию по установке



Без чёткой постановки ИИ часто пишет красивый, но нерабочий код.



Промт для ИИ: просим не идею, а рабочий результат

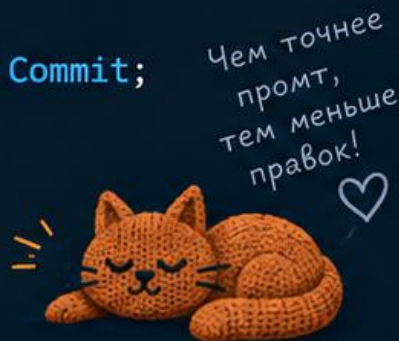
Чем конкретнее требования, тем выше шанс получить рабочий виджет.

Пример промта

Создай рабочий add-on для Orange3 с одним виджетом OWSmartFeatures.

Требования:

- 1) пакет `orangecontrib.smartfeatures`;
- 2) виджет наследует `OWidget`;
- 3) вход: `Data (Orange.data.Table)`;
- 4) выход: `Transformed Data (Orange.data.Table)`;
- 5) GUI: чекбоксы `rooms_per_area` и `age_of_building`, кнопка `Commit`;
- 6) файлы: `setup.py`, `__init__.py`, `widgets/__init__.py`, `OWSmartFeatures.py`, `icons/smart.svg`;
- 7) регистрация через entry_points `orange.widgets`;
- 8) добавь пример запуска через `WidgetPreview`;
- 9) дай инструкцию установки и запуска.



Почему так лучше?

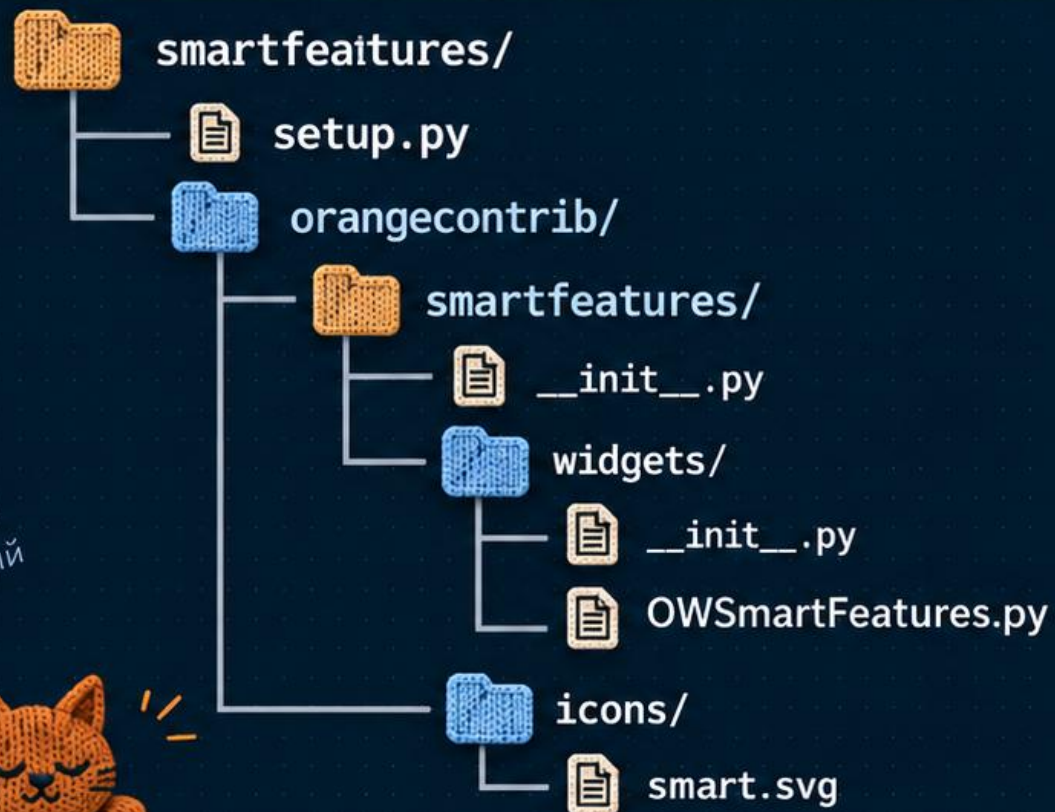
- ЯСНЫ ВХОДЫ И ВЫХОДЫ
- ЯСНА СТРУКТУРА ФАЙЛОВ
- СРАЗУ ПРОСИМ УСТАНОВКУ И ТЕСТ



Хороший промт описывает интерфейс, сигналы, файлы и способ проверки.

Что должен сгенерировать ИИ?

Рабочий виджет — это не один файл, а маленький add-on.



Порядок
в файлах =
работающий
виджет!



Ключевые элементы

- класс OWWidget
- Inputs / Outputs
- Setting и gui
- entry_points для orange.widgets



Минимум внутри виджета

- name, description, icon
- обработка входных данных
- commit/send на выход
- понятный GUI



Если нет структуры add-on, Orange просто не увидит ваш виджет.



Проверка: виджет должен запускаться и честно работать

Сначала локальная проверка, потом уже установка в Orange.



1
Запустить
через
WidgetPreview



2
Подать
тестовые
данные



3
Проверить
выходной
Table



4
Убедиться,
что нет
ошибок GUI



```
if __name__ == '__main__':  
    WidgetPreview(OWSmartFeatures).run(Orange.data.Table('iris'))
```



Чек-лист проверки

- виджет открывается
- категория видна
- сигналы идут без ошибок
- результат появляется в Data Table



Проверяем не только код, но и поведение виджета на реальных данных.

Деплой и подводные камни

Установить виджет мало — нужно ещё не наступить на типичные грабли.



1
`pip install -e .`

Установить
виджет в режим
разработки



2
Перезапустить
Orange

Применить
новую установку



3
Проверить
новую категорию

Найти ваш виджет
в списке



4
Соединить
File → Smart Features
→ Data Table

Проверить работу
на реальных данных

Деплой
тоже может
быть уютным!



Подводные камни

- забыли `entry_points`
- неверный тип сигнала
- виджет не отправляет `output`
- тяжёлая логика блокирует GUI
- нет нормальной проверки после установки



Финальная проверка

- виджет виден в Canvas
- обрабатывает данные
- не падает на простом примере



Рабочий виджет = код + установка + проверка после деплоя.



Как понять, что скрипт работает?

Один красивый результат ничего не доказывает.

Было

Площадь 72 м²
Комнат 3
Построено в 1999

Стало

Площадь на комнату = 24
Возраст здания = 27
Аренда сохранена

Проверяем крайние случаи

Не только обычные квартиры — ещё и неприятные входы.

area_m2	rooms	Результат	Смысл
72	3	24	нормально
72	0	missing	защита от деления
?	2	missing	не выдумываем число

Три строки важнее десяти комплиментов ИИ о собственной программе.

Подводные камни

ИИ может написать уверенный, но неверный код.

ЛОВУШКА

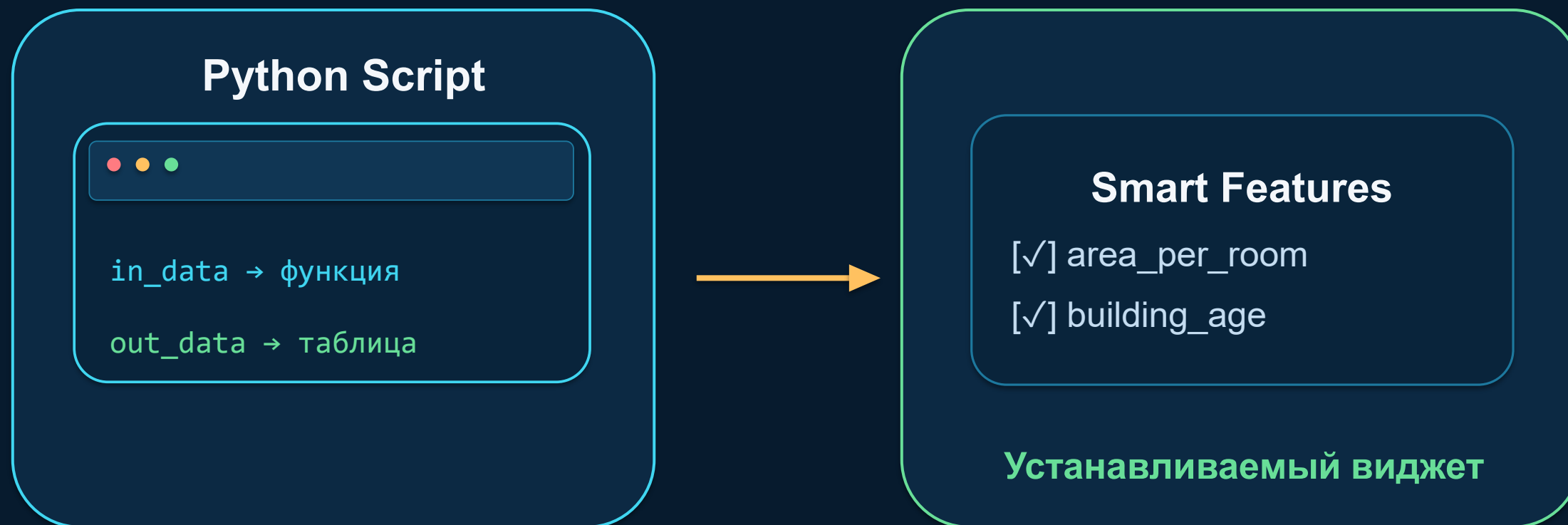
«Ошибок нет, значит результат правильный».

ПРОВЕРКА

Сравнить строки, столбцы, пропуски и целевую переменную.

От Python Script к настоящему виджету

Следующий шаг: интерфейс с входами, выходами и настройками.



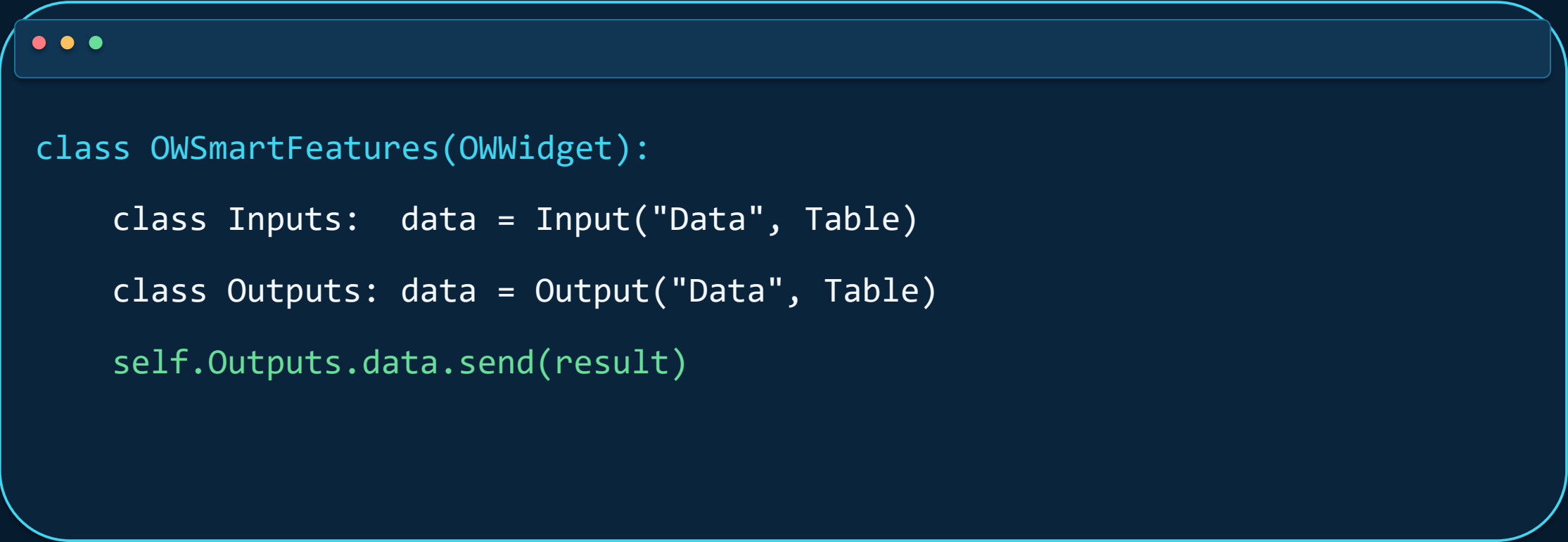
Как устроен настоящий виджет

Четыре строительных блока программирования.



Фрагмент структуры виджета

Имена важнее украшений.



```
class OWSmartFeatures(OWWidget):  
    class Inputs: data = Input("Data", Table)  
    class Outputs: data = Output("Data", Table)  
    self.Outputs.data.send(result)
```

Полный пример исходника и инструкция установки — в комплекте урока.

Подводные камни

Сгенерированный код получает доступ к вашему компьютеру.

ЛОВУШКА

«Модель написала, значит можно запускать».

ПРОВЕРКА

Просмотреть imports, работу с файлами, сетью и конфиденциальными данными.

Вопрос

Что возвращает Python Script, чтобы следующая часть workflow получила таблицу?

?

A. Переменную `out_data`

B. Название файла `screen.png`

C. Только красивый комментарий в коде

Ответ

Это выходной объект типа `Orange.data.Table`.



A. out_data

Вопрос

После запуска код добавил столбцы, но исчезла цель `rent_ils`. Можно использовать результат?

?

A. Да, ведь скрипт не упал

B. Нет, нужно исправить сохранение цели

C. Да, если новый столбец называется `AI_score`

Ответ

Правильный результат — не просто отсутствие ошибок.



В. Сначала исправляем схему таблицы.

Карта части 15.3

Мы умеем расширять анализ данных собственными действиями.

ОПИСАЛИ

Дали ИИ точную спецификацию

НАПИСАЛИ

Получили код под Python Script

ПРОВЕРИЛИ

Контрольные строки и сохранность данных

РАСШИРИЛИ

Поняли путь до настоящего виджета

КУРС «АНАЛИЗ ДАННЫХ В ВЕК ИИ»

ЧАСТЬ 15.4

Большие данные. Большая ответственность.

Этика — это часть качества анализа, а не необязательная лекция.



За каждой строкой может быть человек

Ошибка модели иногда меняет не график, а чью-то жизнь.

1

Кредит

Кому банк отказал в займе?

2

Поступление

Кого алгоритм не рекомендовал вузу?

3

Здоровье

Кому нужна дополнительная проверка?

В этих задачах недостаточно одной высокой метрики.

Хорошая метрика не равна хорошему решению

**Модель может быть точной — и всё
равно причинять вред.**

Техническое качество и последствия решения — разные вопросы.

Три уровня ответственности аналитика

Недостаточно только нажать Train.

1

ДАННЫЕ

Где взяты и кому принадлежат?

2

МОДЕЛЬ

Как проверена, для кого ошибается?

3

РЕШЕНИЕ

Кого затрагивает и кто отвечает?

«95% точности!» — и что дальше?

Общая метрика может скрывать неравномерные ошибки.

95%

общая ассурасу
(условный пример)

Группа А

98%

Группа В

78%

Числа придуманы для обсуждения; это не результаты исследования.

95% точности — а кто в оставшихся 5%?

Хороший средний результат ещё ничего не говорит о распределении ошибок.

95

верных решений из 100

5

ошибок — но у кого?

Нужны метрики по группам и по типам ошибок.

Несправедливость может начаться до модели

Не все группы одинаково представлены в данных.

Обучающая таблица

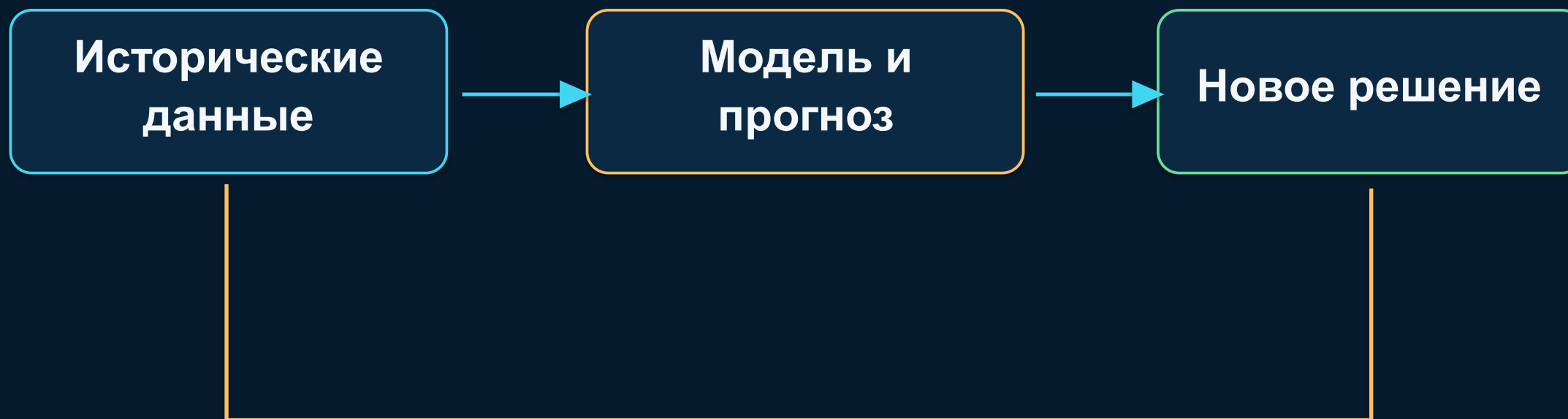
Исторические решения
Неполные записи
Старая практика отбора

Реальное применение

Новые люди
Другие условия
Старые перекосы могут повториться

Петля обратной связи

Предсказания могут менять будущие данные.

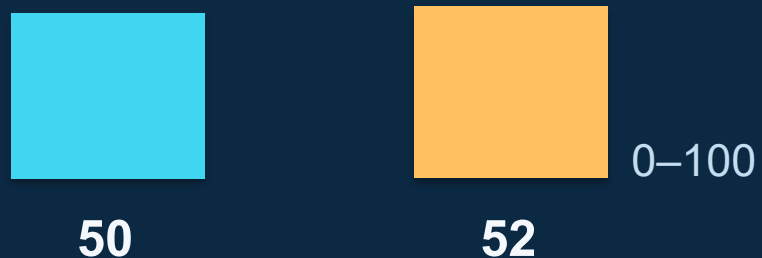


Новое решение формирует завтрашнюю обучающую таблицу

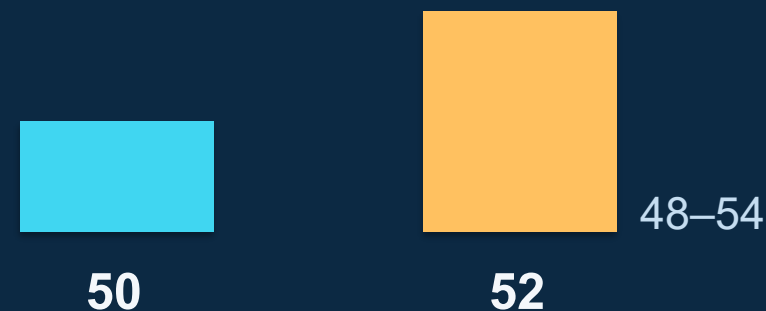
Одинаковые числа — разные впечатления

Масштаб оси может превратить небольшой сдвиг в «катастрофу».

Шкала от нуля



Шкала 48–54



Значения 50 и 52 — одинаковые на двух графиках.

Удалили имена — анонимно?

Иногда человека можно узнать по комбинации других признаков.

До удаления

Имя, возраст, индекс,
редкая профессия, дата...

После удаления

Возраст, индекс,
редкая профессия, дата...

Риск идентификации остаётся

Не отправляйте личные данные в чат «на всякий случай».

Удобство ИИ не отменяет правила обработки данных.

Сначала разрешение, минимизация и защита.

Три проверки перед публикацией результата

Этика не заменяет науку — она требует честной науки.

1

ПРОИСХОЖДЕНИЕ

Откуда взялись данные и есть ли право использовать?

2

ПОВТОРЯЕМОСТЬ

Другой аналитик сможет проверить шаги?

3

ОГРАНИЧЕНИЯ

Для кого и при каких условиях модель работает?

ИИ может придумать источник с идеальным названием

«В исследовании доказали...» — не доказательство.

Journal of Extremely Convincing Data

«Авторы: A. Someone, B. Nobody»

DOI: 10.0000/imaginary

**НУЖНО
ПРОВЕРИТЬ**

Текст примера вымышленный. Нельзя цитировать как настоящую статью.

Кто примет последнее решение?

Особенно когда от ответа зависит человек.

Модель

Предлагает оценку
риска

Человек

Проверяет условия и
контекст

Процедура

Даёт возможность
оспорить

Подводные камни

Подтверждение гипотезы любой ценой.

ЛОВУШКА

«Покажу заказчику только успешные метрики».

ПРОВЕРКА

Сообщить ограничения, ошибки и неопределённость.

Совет от доктора Игоря

Если модель влияет на людей, проверяйте её для людей.

1 Кому вредно?

Оцените цену ошибок

2 Какие данные?

Проверьте доступ и перекосы

3 Кто отвечает?

Определите контроль и пересмотр

Вопрос

Можно ли считать таблицу анонимной после удаления имён?

?

A. Да, автоматически

B. Нет, другие признаки могут идентифицировать человека

C. Да, если удалить ещё одну колонку

Ответ

Удаление имени само по себе не гарантирует приватность.



В. Возможна реидентификация.

Вопрос

ИИ привёл убедительную научную ссылку. Что делать?

?

- A.** Использовать: ИИ не выдумывает DOI
- B.** Проверить существование работы и что она действительно утверждает
- C.** Считать источник верным, если название сложное

Ответ

Ссылка — это приглашение к проверке, а не доказательство.



В. Проверить первоисточник.

Карта части 15.4

Техническая точность и человеческая ответственность должны идти вместе.

ЧЕСТНОСТЬ

Не скрывать ошибки и неудобные результаты

СПРАВЕДЛИВОСТЬ

Проверять различия и последствия ошибок

ПРИВАТНОСТЬ

Доступ, минимизация, риск идентификации

ПРОВЕРЯЕМОСТЬ

Данные, код, источники, ограничения

КУРС «АНАЛИЗ ДАННЫХ В ВЕК ИИ»

ЧАСТЬ 15.5

Курс завершён. Что учить дальше?

Три маршрута и один реальный проект,
который можно закончить.

01 ПРАКТИКА

02 КОД

03 ТЕОРИЯ

Выбираем свой маршрут

Что произойдёт после последнего занятия?

Без самостоятельных задач теория быстро превращается в забытые названия.

1

Сегодня

Понимаю идею и
повторяю пример.

2

Через неделю

Пробую объяснить и
повторить без подсказки.

3

Через месяц

Применяю метод в
собственном проекте.

Знание закрепляется, когда начинаем его использовать.

Не собирайте сертификаты. Собирайте решения задач.

Лучший следующий курс не заменит законченный проект.

Один проверенный проект полезнее десяти незаконченных.

Выберите свой следующий маршрут

Можно идти разными дорогами и комбинировать их.

1

ORANGE / KNIME

Углубить практику без
большого кода.

2

PYTHON + ИИ

Автоматизация и
собственные
инструменты.

3

СТАТИСТИКА / ML

Теория, проверки,
понимание моделей.

Orange или Python?

Вопрос не «кто победил», а «какая задача?».

Orange

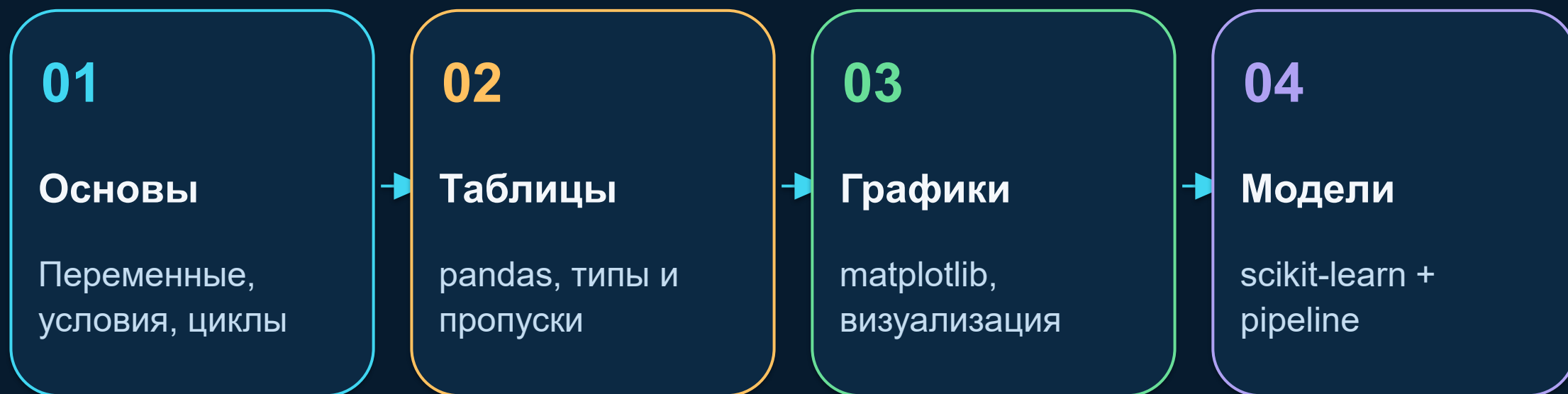
Быстрое исследование
Визуальные workflow
Учебные эксперименты

Python

Автоматизация
Скрипты и интеграции
Собственные инструменты

С чего начать Python?

Практический минимум для аналитика данных.



Как выбрать, что учить следующим?

Следующий шаг зависит от того, какую задачу вы хотите решать.

1

Хочу анализировать

Больше практики в
Orange и KNIME.

2

**Хочу
автоматизировать**

Python: таблицы,
графики, функции.

3

**Хочу понимать
глубже**

Статистика, вероятность
и проверка моделей.

Не обязательно начинать все три маршрута одновременно.

Что почитать после нашего курса?

Небольшой выбор проверяемых ресурсов — без списка на сто книг.

ISL with Python

Бесплатный учебник и практикум

Orange — Get Started

Документация и примеры workflow

KNIME Learning Center

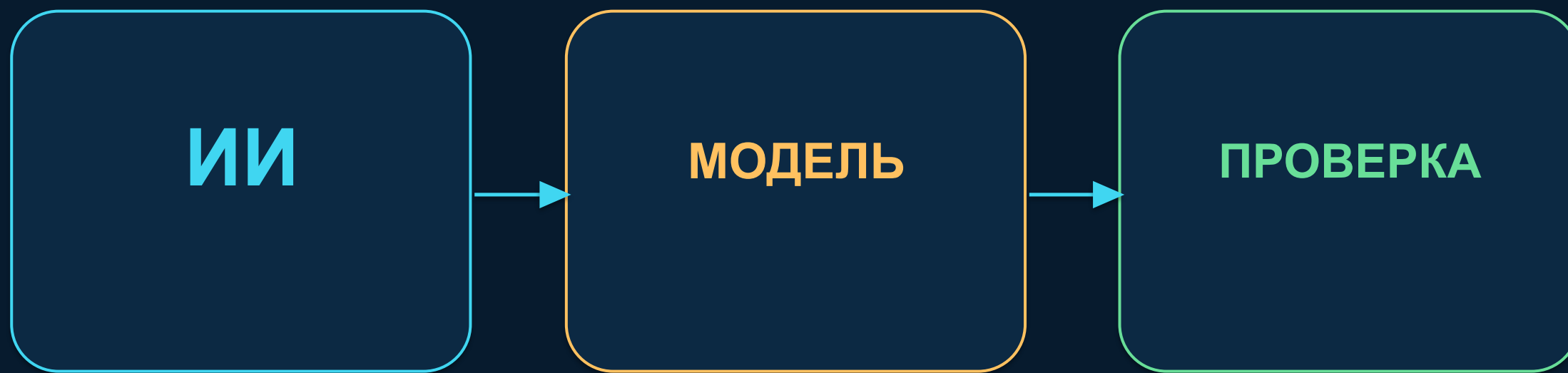
Бесплатные самостоятельные курсы

scikit-learn User Guide

Модели, CV, pipelines и ловушки

Статистика — не конкурент ИИ

Это язык, на котором проверяют уверенные предсказания.



Вероятность, выборка, неопределённость и причинность остаются важны.

План на 30 дней

Четыре недели, один небольшой законченный проект.

1

ПРОБЛЕМА

Выберите вопрос и
данные

2

ИССЛЕДОВАНИЕ

EDA, очистка,
baseline

3

МОДЕЛЬ

CV, ошибки,
улучшения

4

ИТОГ

Отчёт +
ограничения +
вывод

Как выглядит достойный мини-проект?

Шесть компонентов, не обязательно огромный код.

ВОПРОС

зачем исследуем?

ДАННЫЕ

откуда и какие?

EDA

что заметили?

МОДЕЛЬ

честная проверка

ВЫВОД

что узнали?

ГРАНИЦЫ

чего не знаем?

Подводные камни

Ловушка бесконечного обучения.

ЛОВУШКА

«Ещё три курса — и тогда начну проект».

ПРОВЕРКА

Выбрать задачу сегодня и учиться по ходу.

Совет от преподавателя

Начните со своего вопроса, а не с модного алгоритма.

1 Вопрос

Что вы действительно хотите узнать?

2 Данные

Что есть в открытом доступе?

3 Результат

Что можно честно показать?

Вопрос

Чему лучше посвятить следующую неделю?

?

A. Ещё раз выучить названия всех алгоритмов

B. Взять маленькую задачу и довести исследование до вывода

C. Сразу строить самую сложную нейросеть

Ответ

Практика с честным выводом превращает знания в навык.



В. Маленький законченный проект.

Вопрос

ИИ написал работающий Python-скрипт. Что полезнее всего проверить?

?

A. Соответствует ли решение задаче и воспроизводится ли

B. Красивые ли комментарии

C. Длиннее ли код, чем у одноклассника

Ответ

Инструмент хорош тогда, когда решает нужную задачу корректно.



А. Соответствие и воспроизводимость.

Карта части 15.5

После курса есть три маршрута — и один общий принцип.

ПРАКТИКА

Orange и KNIME для новых задач

КОД

Python + библиотеки + проверка ИИ

ТЕОРИЯ

Статистика, эксперименты, ML

ПРОЕКТ

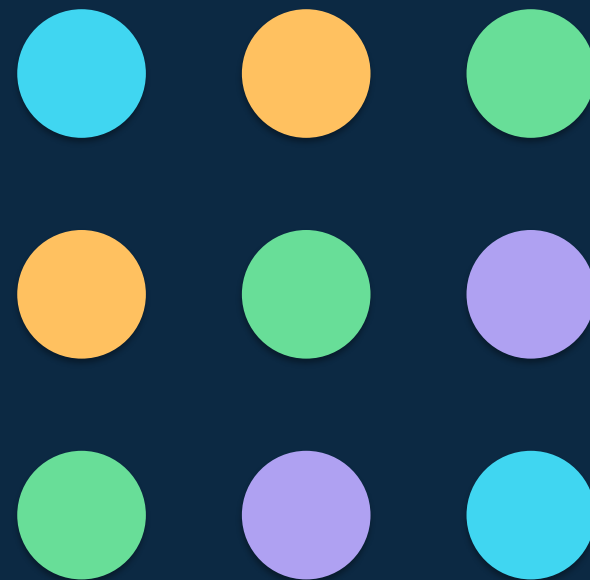
Один честный законченный анализ

КУРС «АНАЛИЗ ДАННЫХ В ВЕК ИИ»

ЧАСТЬ 15.6

Последний слайд? Нет. Первый проект.

Большой финал курса «Анализ данных в век ИИ».



СТОП, ПРОВЕРЯЕМ!

Мы учили не виджеты. Мы учили задавать вопросы.

**Любой инструмент — только часть
процесса.**

Стоп, проверяем!

Панорама знаний: что мы изучили

Одна исследовательская цепочка вместо пятнадцати изолированных тем.

ВОПРОС

Цель, тип задачи

ДАННЫЕ

EDA, качество, признаки

МОДЕЛИ

Регрессия, деревья, kNN, лес

СТРУКТУРА

Кластеры, тексты, сети, изображения

ВРЕМЯ

Тренд, сезонность, прогноз

ПРОВЕРКА

CV, ошибки, этика, ИИ

Какую модель выбирать теперь?

Нет одной волшебной кнопки «Анализировать всё».

ЧИСЛО

Регрессия, лес, ML

КЛАСС

Логистическая, kNN,
деревья

ГРУППЫ

Кластеризация

ТЕКСТ

Токены, мешок слов, сети

ВРЕМЯ

Ряды, лаги, периодичность

ИЗОБРАЖЕНИЕ

Признаки, эмбединги

Шесть шагов самостоятельного аналитика

Рецепт без магии — от идеи до честного вывода.

01 Вопрос

Что решаем?

02 Данные

Что реально знаем?

03 EDA

Что не так?

04 Модель

С чем сравним?

05 Проверка

Не подсмотрели?

06 Решение

Что утверждаем?

Новый уровень: «аналитик, который проверяет»

Финальное достижение нашего курса.



МОЖЕТ СКАЗАТЬ «НЕ ЗНАЮ»

и знает, как это проверить

Это важнее уверенного, но выдуманного ответа.

Вопрос

Модель показала фантастический результат на тесте. Что делаем первым?

?

A. Публикуем победу

B. Проверяем утечки и схему оценки

C. Добавляем ещё 300 деревьев

Ответ

Высокая метрика — повод исследовать, а не поклоняться.



В. Сначала проверяем эксперимент.

Совет преподавателя

Никогда не отдавайте ИИ право на последний вопрос.

1 СПРОСИТЬ

Что именно мы хотим узнать?

2 ПРОВЕРИТЬ

На каких данных основан ответ?

3 ОБЪЯСНИТЬ

Где вывод может оказаться неверным?

Вопрос

Что важнее унести из курса?

?

A. Список всех кнопок Orange

B. Привычку задавать вопросы и проверять выводы

C. Мнение, что ИИ всегда прав

Ответ

Вопрос → данные → проверка → ответственное решение.



В. Мышление исследователя.

Продолжение начинается сегодня

Возьмите вопрос из собственной жизни и попробуйте его исследовать.

1

Найдите данные

2

Проверьте их

3

**Сформулируйте
вывод**

И не забудьте: «Стоп, проверяем!»

Спасибо за пятнадцать занятий!

И спасибо за вопросы, сомнения и желание разбираться.

Анализ данных в век ИИ

Мы закончили курс. Исследование мира продолжается.

© 2026 доктор Игорь Клейнер

СТОП, ПРОВЕРЯЕМ!

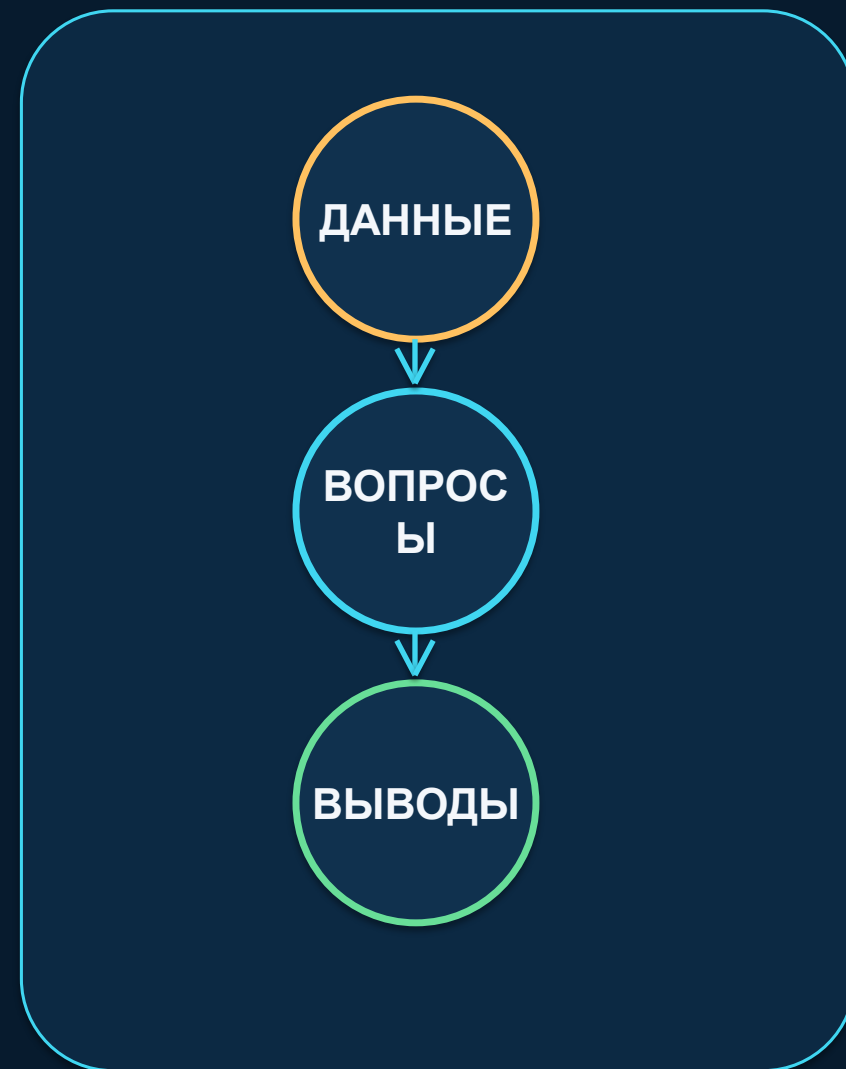
Главный принцип курса остаётся с вами.

**ИИ предлагает.
Аналитик проверяет.
Человек принимает решение.**

КУРС «АНАЛИЗ ДАННЫХ В ВЕК ИИ»

Финальный проект: ваши данные — ваши открытия

2 месяца • 2–3 вопроса
Orange + ИИ + Python



Теперь не мои данные. Ваши.

Проект начинается с живого вопроса, а не с поиска модной модели.

ГДЕ ВЗЯТЬ ДАННЫЕ?

- **Открытые наборы**
город, погода, спорт, экономика
- **Ваша работа**
только с разрешёнными данными
- **Собственные наблюдения**
если сбор безопасен и законен

ЗАЧЕМ ИССЛЕДОВАТЬ?

**Чтобы узнать что-то
полезное о мире —
и суметь это обосновать.**

Не обязательно находить сенсацию.

Настоящие данные → содержательные вопросы → честные выводы.

Два-три вопроса — одна история

Например, у вас есть реальные данные об аренде квартир.

01

СРАВНЕНИЕ

В каких районах аренда дороже?

02

СВЯЗЬ

Связана ли цена с площадью?

03

ПРОГНОЗ

Можно ли оценить аренду по признакам?

Связь — ещё не причина. А хороший вопрос можно проверить.

Два месяца: четыре этапа

Рабочий план на восемь недель, чтобы проект не превратился в ночь перед дедлайном.



Не гонка алгоритмов. Работа, которую можно объяснить и повторить.

Orange + ИИ + Python

Три инструмента, но главный исследователь в команде — человек.



ORANGE

Исследовать

Графики, таблицы,
модели, workflow .ows



ИИ

Помогать думать

Идеи, объяснения,
черновики кода, поиск
ошибок



PYTHON

Расширять

Сложные преобразования
и автоматизация, когда
нужны

ИИ предлагает. Вы проверяете. Данные решают спор.

Что отправить преподавателю?

Нужна работа, которую можно открыть, проверить и обсудить.

01 ДАННЫЕ

CSV или ссылка на источник; без чужих личных данных

02 ORANGE

Файл .ows с понятным маршрутом анализа

03 ВЫВОДЫ

2–3 вопроса, графики, ответы и ограничения

04 PYTHON / ИИ

Код, если применяли, и пояснение роли ИИ

Сдали → проверка преподавателя → сертификат

Сертификат — за выполненный и проверенный проект.

Подводные камни настоящего исследования

Коварная ошибка может выглядеть как отличная модель.

! УТЕЧКА

Обучили обработку
на всём наборе

Обрабатывать внутри
train

! ПРИВАТНОСТЬ

Опубликовали
личные данные

Обезличить или не
делиться

! ФАНТАЗИИ ИИ

Приняли ответ ИИ
без проверки

Сверять формулы и
факты

Перед отправкой откройте workflow заново и проверьте все выводы.

Совет от Игоря

Хорошая научная работа не обязана заканчиваться сенсацией.

**Честное «связи не нашли» лучше
красивого, но выдуманного открытия.**

01

Понятно

02

Повторяемо

03

Проверено

Нет подтверждения гипотезы? Это тоже результат исследования.

Вопрос

Когда двухмесячный исследовательский проект действительно готов?

А

Когда ИИ написал блестящий отчёт

Б

Когда выводы проверены и анализ можно повторить

В

Когда на графиках пять красивых цветов

Ответ: когда результат можно проверить

Б — важен не эффектный отчёт, а воспроизводимый путь от данных к выводу.

Б. Выводы проверены, шаги анализа воспроизводимы.



Теперь это исследование, а не набор красивых картинок.

Спасибо, что прошли этот путь вместе со мной

Пятнадцать занятий, множество вопросов и настоящая работа с данными.



**Вы сделали курс
живым.**

Спасибо за ваши вопросы,
сомнения и открытия.

Настоящее обучение начинается там, где человек спрашивает «почему?».

Курс закончился. Наше общение — нет.

Пишите о проектах, новых вопросах и даже о том, что не получилось.

ВАШИ ПРОЕКТЫ

Что получилось исследовать?

ВАШИ ВОПРОСЫ

Где нужна подсказка?

ВАШИ ИДЕИ

Как улучшить курс?

ВАШИ ОШИБКИ

Что оказалось неожиданным?

Оставайтесь на связи: сайт курса • YouTube • Telegram.

Оставьте след в ленте — расскажите о курсе

Ваш честный отзыв поможет следующему человеку решиться начать.

ПРИМЕР ПОСТА ИЛИ ОТЗЫВА

«Завершился бесплатный курс „Анализ данных в век ИИ“.
Больше всего пригодилось: ...
Теперь хочу применить это к ...
Совет будущим студентам: ...»

ПОЧЕМУ ЭТО ВАЖНО?

- Новые участники узнают о курсе
- Отзывы помогают улучшать материал
- Можно поделиться своим проектом

Отзыв может быть и критическим: главное — конкретно и честно.

СПАСИБО!

**Курс закончился.
Ваш анализ данных — нет.**

Наблюдайте. Спрашивайте. Проверяйте.

СТОП, ПРОВЕРЯЕМ!