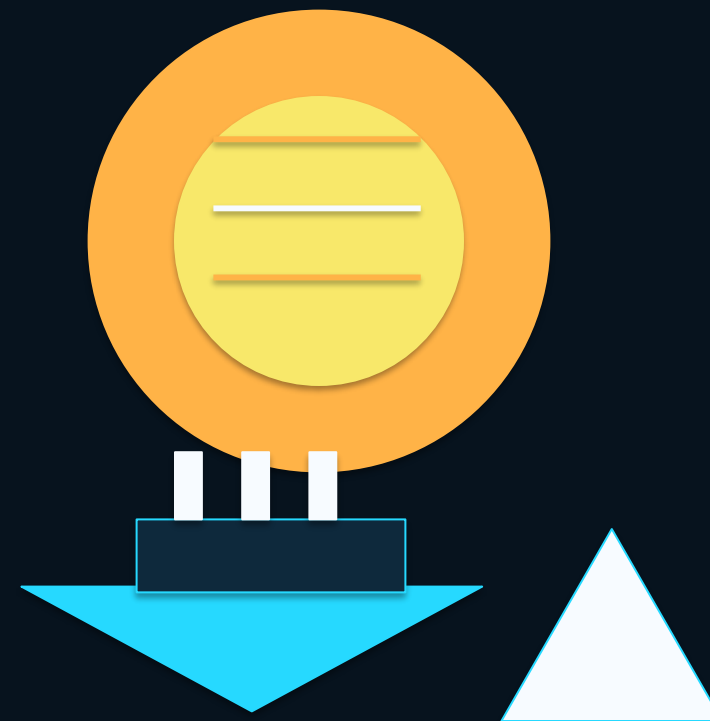


Анализ данных в век ИИ
**Просто, доступно,
полезно**
Встреча 1 часть 1

Добро пожаловать



АНАЛИЗ ДАННЫХ В ВЕК ИИ

Нейросеть подсказывает.
Человек проверяет.

СТОП. ПРОВЕРЯЕМ!



**Доктор
Игорь
Клейнер**

курс для начинающих
и любопытных

О курсе

12-14 встреч

По субботам 19:00-22:00 (по Израилю)

Живые лекции в Зум или по записи

Адрес курса <https://dsall.netlify.app/>



О курсе

Практические задания

Квизы

Игры

Пасхалки

О курсе

Сертификат - победителям после прохождения испытаний

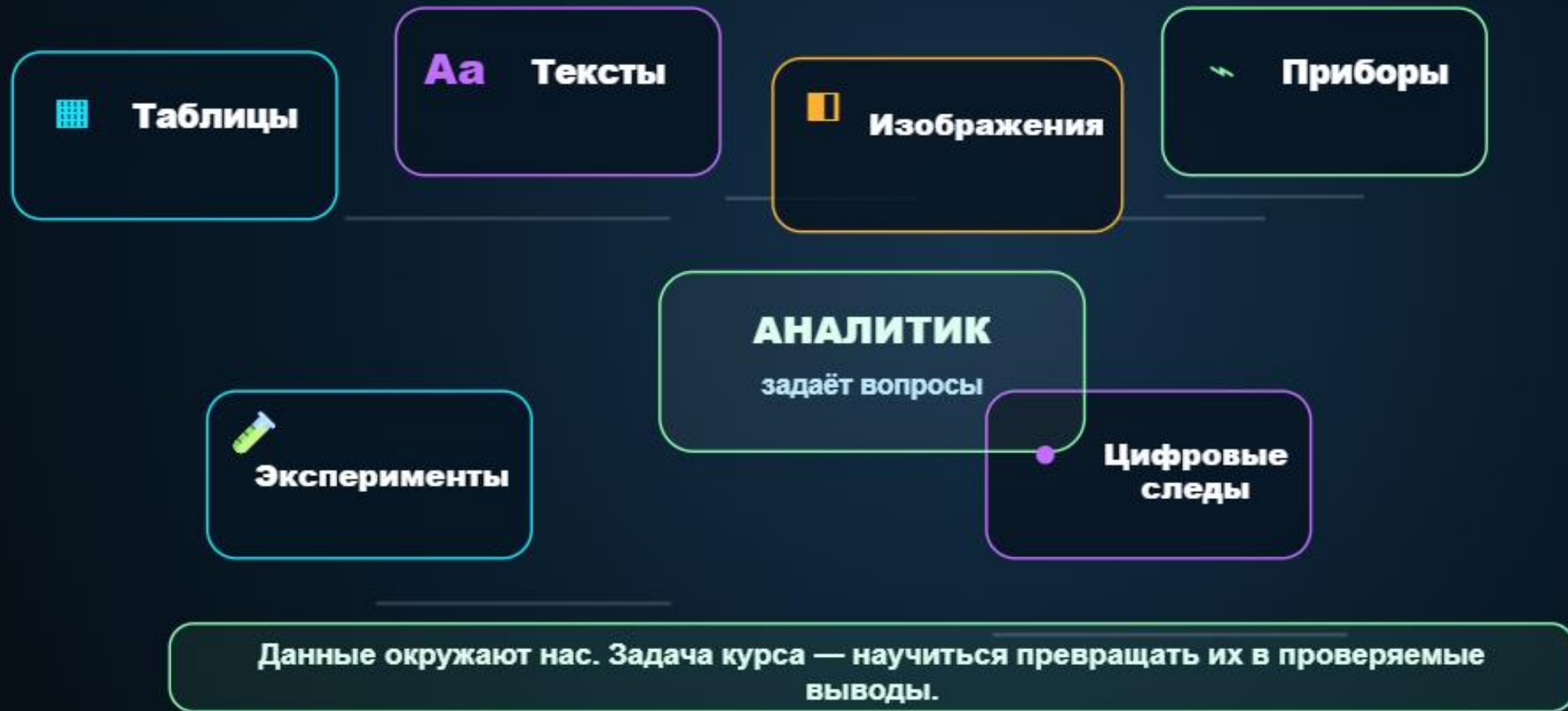
Требования к участникам: желание изучать новое и не бояться испачкаться в данных

Orange Data Mining

AI tools

Почему курс появился сейчас: мир данных

Данных становится всё больше, но сами они ничего не объясняют.



Почему курс появился сейчас: новые возможности

Сделать анализ стало легче. Понять и проверить его всё ещё должен человек.



Эффективные алгоритмы

быстрее находят закономерности



Визуальное программирование

собираем анализ из блоков



ИИ-помощник

объясняет, предлагает, проверяет



Дешёвые ресурсы

компьютеры и облака доступнее



Аренда мощностей

платим только за время работы

Инструменты стали сильнее. Ответственность за выводы осталась у человека.

Для кого этот курс

Для начинающих и тех, кто хочет понимать данные без тяжёлого программирования.



**понимать,
что говорят
данные**



**строить графики
и модели**



**использовать
нейросети
осознанно**



**начать без
тяжёлого
программирования**

**Можно
начать с нуля**

Предварительное
знание
машинного обучения
не требуется.

Нужно любопытство и
готовность проверять.

С какими данными мы будем работать

Будем не только считать, но и исследовать данные разного типа.



**Таблицы
и числа**



**Тексты
и сообщения**



Изображения



Наблюдения



Эксперименты



Симуляции

Один курс — разные данные, разные вопросы и разные способы проверки.

Чему мы научимся

Главный навык — понимать, что показывает результат и можно ли ему доверять.



Видеть данные

графики, визуализация, статистика без страха



Строить модели

регрессия, прогнозирование, классификация



Работать с разными форматами

тексты, изображения, кластеры, симуляции



Проверять выводы

качество моделей, ошибки, критическое мышление, ИИ-помощник

Не просто “получить ответ”, а понять путь от данных к выводу.

Как проходит курс

Всё обучение собрано в понятный маршрут.



Живые занятия

в Zoom



Записи занятий

на YouTube



Материалы и задания

на панели курса



Практика

на занятиях и
самостоятельно

**Лекция → практика → задание → проверка
понимания**



карта курса

Как устроена одна неделя

Сначала понимаем. Потом пробуем сами. Ошибаемся. Исправляем. Учимся



Лекция

объясняем идею



Практика в Orange

работаем руками



Квиз или игра

закрепляем



Самостоятельная миссия

пробуем сами



Дополнительные материалы

идём глубже

Понимаем

Пробуем

Проверяем

Продолжаем

Orange: анализ данных без тяжёлого программирования

Не пишем длинный код. Но думаем как аналитики.

**Загружаем
таблицу**

File

данные

Data Table

таблица

Строим графики

Scatter Plot

график

Обучаем модели

Model

модель

**Проверяем
результат**

Test & Score

проверка

Orange — наша лаборатория: загружаем таблицы, соединяем блоки, строим графики, обучаем модели и проверяем результат.

ИИ как помощник



Нейросеть может помочь:

- предложить гипотезу;
- объяснить сложный результат;
- найти возможную ошибку;
- предложить следующий шаг.

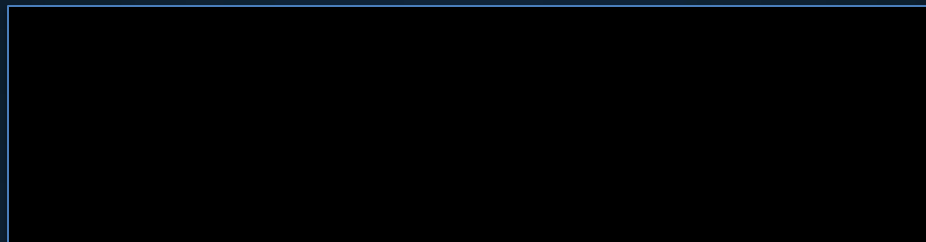


Но есть опасность:

Она может уверенно выдумать факт, число, источник или объяснение. Поэтому нельзя незаметно подменять данные красивыми словами.

**Человек задаёт вопросы и
проверяет ответ.**

Главное правило курса



**Откуда
взялись
данные?**

**СТОП
ПРОВЕРЯЕМ**

**Правильно ли
построен график?**

**Подходит ли
модель?**

**Следует ли вывод
из результата?**

**Имеет ли
результат
смысл?**

Где находятся материалы

Все материалы собраны на панели курса.

 Карта недель

 Слайды

 Записи
занятий

 Задания и
миссии

 Файлы для
практики

 Дополни́тельные
материалы

<https://dsall.netlify.app/>



карта недель и панель курса

Один сайт — все ссылки

Сайт курса — главный вход. Не нужно хранить десять разных адресов.

ai-teaching-lab-igor.netlify.app/ru#program



Панель курса



YouTube



Telegram



Stepik



LinkedIn



Email

LinkedIn: bigdatagile Email: igkleiner@gmail.com

Кто ведёт курс



**Доктор
Игорь Клейнер**

преподаватель ·
исследователь
статистика и анализ данных

**Моя задача — не просто показать
кнопки,**

Практический опыт работы

Microsoft

Intel

Huawei

Skype



Преподавание

университетские
курсы и сложные
идеи



Исследования

модели,
вероятности,
проверка гипотез



**Data
Science**

реальные данные
и прикладные
задачи

Нейросеть помогает, но человек отвечает за понимание и проверку.

Мой опыт и проекты



Преподаю

анализ данных,
статистику, теорию
вероятностей и машинное
обучение.



Работал в Data Science

с реальными данными,
моделями и прикладными
задачами.

Создаю образовательные проекты



YouTube-канал
онлайн-курсы
книга по
теории
вероятностей

Сочетаю университетское
преподавание, практику и
популяризацию науки.

Не просто теория и не просто инструменты — курс построен на сочетании опыта,
практики и объяснения.

История этого курса

Курс, который постоянно развивается.

Более 15 запусков на нескольких

Каждый запуск добавлял новые
вопросы, примеры и улучшения.

Это не просто новый набор
слайдов,
а следующая версия многолетнего
курса.

Кто учился на разных версиях курса



школьники



студенты



пенсионеры



профессора



географы



исследователи
древности



специалисты
из бизнеса



любопытные
начинающие

Обратная связь постоянно улучшает объяснения, примеры, задания и темп курса.

Почему я так преподаю

Понять

Попробовать

Ошибиться

Проверить

Объяснить



реальные истории



эксперименты



Orange



симуляции



нейросети



проверка выводов

Следующий блок: «Данные вокруг нас»

ВСТРЕЧА 1 · ЧАСТЬ 2

ДАННЫЕ ПОВСЮДУ

Почему анализ данных стал одним из главных навыков эпохи ИИ

Данные → вопросы → графики → модели →
проверка



**Анализ данных
в век ИИ**

блок 2

О чём будет этот блок

Данные повсюду: от цифровых следов до решений, которые меняют мир.

1

**Где
появляются
данные**

телефон, бизнес, наука,
медицина

2

**Почему их
стало так
много**

рост, интернет,
устройства, сервисы

3

**Зачем
нужны
аналитики**

данные молчат, человек
задаёт вопросы

4

**Почему
анализ
популярен
сейчас**

алгоритмы, Orange,
облака, ИИ

**Финальная идея: анализ данных — способ разговаривать с миром через
вопросы, графики, модели и проверку.**

Что может быть данными

Данные — это не только таблица с числами.

123 Числа
и таблицы

Aa Тексты
и
сообщения

 Изображения
и видео

 Звук
и речь

 Координаты
и маршруты

 Действия
пользователей

 Измерения
приборов

 Симуляции
и
эксперименты

Смысл анализа зависит не от “магии формул”, а от качества вопроса и данных.

Объём данных растёт почти взрывным образом

Экспонента здесь — образ ускорения, а не обещание точной формулы.

- Согласно некоторым оценкам размер данных созданных человеком к 2020 году примерно составит 40 зетабайт

- 1 зетабайт $\approx 1000000000000000000000$

General Big Data Statistics

Now that you know the latest data and how big data affects the industry, let's dive deeper.

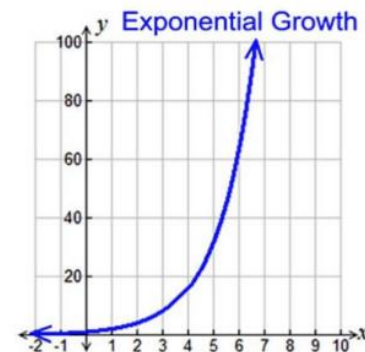
9. In 2020, there will be around 40 trillion gigabytes of data (40 zettabytes).

(Source: EMC)

Measuring the amount of data we have today is not an exact science. When going through the numbers related to big data, we found numerous predictions and estimates, but very few numbers could hold on to. The number of big data growth statistics was also very high. One of the many predictions about the amount of big data came from IDC's study – "The Digital Universe in 2020". According to the source, next year, we should have around 40 trillion gigabytes of data, i.e., 40 zettabytes.

! Почему это важно?

Когда данных мало — можно смотреть вручную. Когда много — нужны методы, графики, модели и проверка.



Что происходит в интернете за одну минуту

Масштаб проще почувствовать через “интернет за 60 секунд”.

138,9 млн

Reels смотрят
за минуту

3,4 млн+

просмотров YouTube
за минуту

251,1 млн

email-сообщений
за минуту

1,04 млн

сообщений Slack
за минуту

362 тыс. часов

Netflix стримят
за минуту

5,52 млрд

интернет-пользователей
в мире

Это не “вся правда о данных”, а хороший способ увидеть масштаб.

Данных много. Но они молчат

Таблица сама не скажет, что именно в ней важно.

данные

Нужны вопросы:

? Что измеряли?

? Чего не хватает?

? Где ошибка?

? Какая связь?

? Какой вывод допустим?

Аналитик превращает молчаливые данные в проверяемый рассказ.

Почему нужны аналитики данных

Данные есть у многих. Умение разговаривать с ними есть не у всех.



**задать
правильный
вопрос**



**увидеть
закономерность**



**проверить
модель**



**объяснить
вывод**

В этом блоке появляется главная метафора: мы будем учиться говорить с данными.

Финал: мы будем разговаривать с данными

Русская версия метафоры из старого курса.

Parseltongue is the language of serpents and those who can converse with them. An individual who can speak Parseltongue is known as a Parselmouth. It is a very uncommon skill, and may be hereditary. Nearly all known Parselmouths are descended from Salazar Slytherin.



<http://harrypotter.wikia.com/wiki/Parseltongue>

Анализ данных — это современная способность разговаривать с данными и понимать мир через них.

И хорошая новость: этому можно научиться.

Финал: мы будем разговаривать с данными

Русская версия метафоры из старого курса.

Анализ данных это способность людей разговаривать с данными. Люди которые могут разговаривать с данными известны как аналитики данных. Умение анализировать данные – редкая способность, возможно наследуемая ☹, но возможно и приобретаемая ☺



<http://harrypotter.wikia.com/wiki/Parseltongue>

Анализ данных — это современная способность разговаривать с данными и понимать мир через них.

И хорошая новость: этому можно научиться.

Данные в научных исследованиях

Современная наука всё чаще начинается с измерений и наблюдений.



эксперименты



наблюдения



география



археология



биология



медицина



физика



история

Данные помогают проверять гипотезы, искать закономерности и строить объясняющие модели.

Данные для бизнеса

Данные помогают принимать решения быстрее и точнее.



продажи



клиенты



логистика



реклама



спрос



персонализация



риски



качество

Бизнесу нужны не просто графики, а решения: что изменить, где риск, кому предложить.

Данные, которые спасают жизнь

Анализ данных — это не только прибыль, но и помощь людям.

**эпидемии
и коронавирус**

**ранняя
диагностика**

**данные
спасают**

**болезни
и риски**

**климат
и потепление**

**Главное: данные помогают увидеть проблему раньше и
действовать точнее.**

Почему анализ данных особенно популярен сегодня

Сошлись несколько факторов одновременно.



**данных
больше**



**алгоритмы
эффективнее**



**инструменты
проще**



**ресурсы
доступнее**

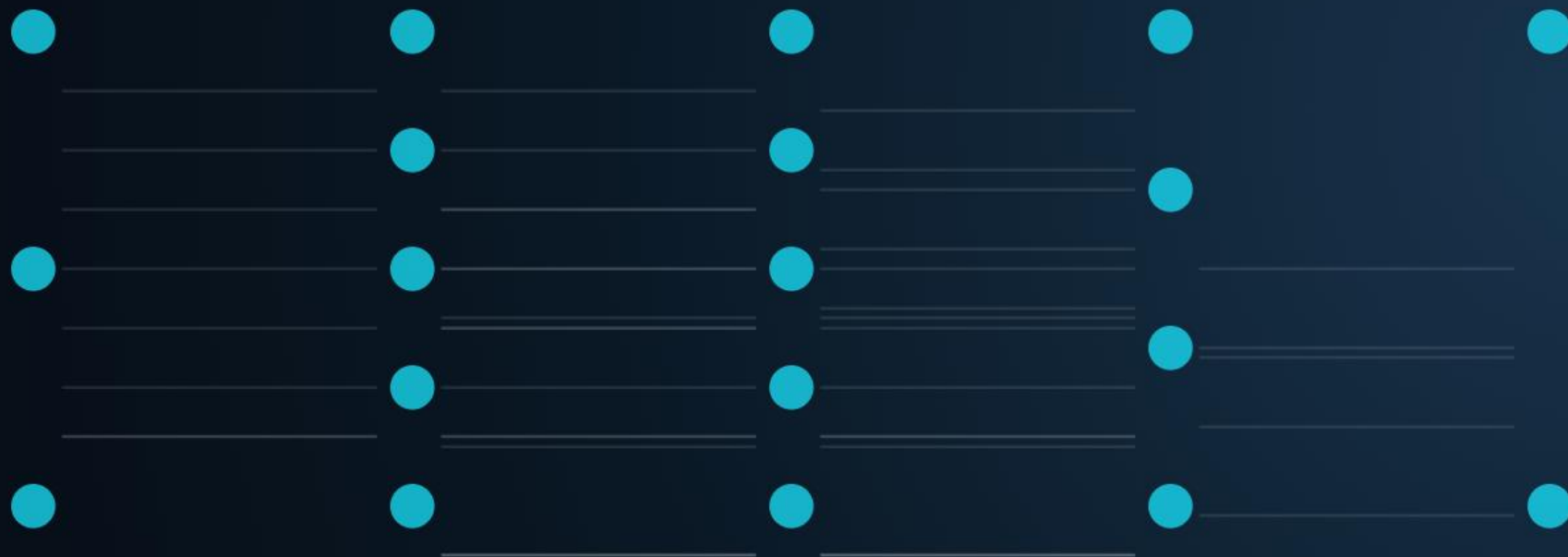


**ИИ
помогает**

Но доступность инструментов не отменяет ответственности: результат всё равно нужно понимать и проверять.

Алгоритмы стали эффективнее

Модели умеют искать закономерности там, где вручную всё слишком сложно.



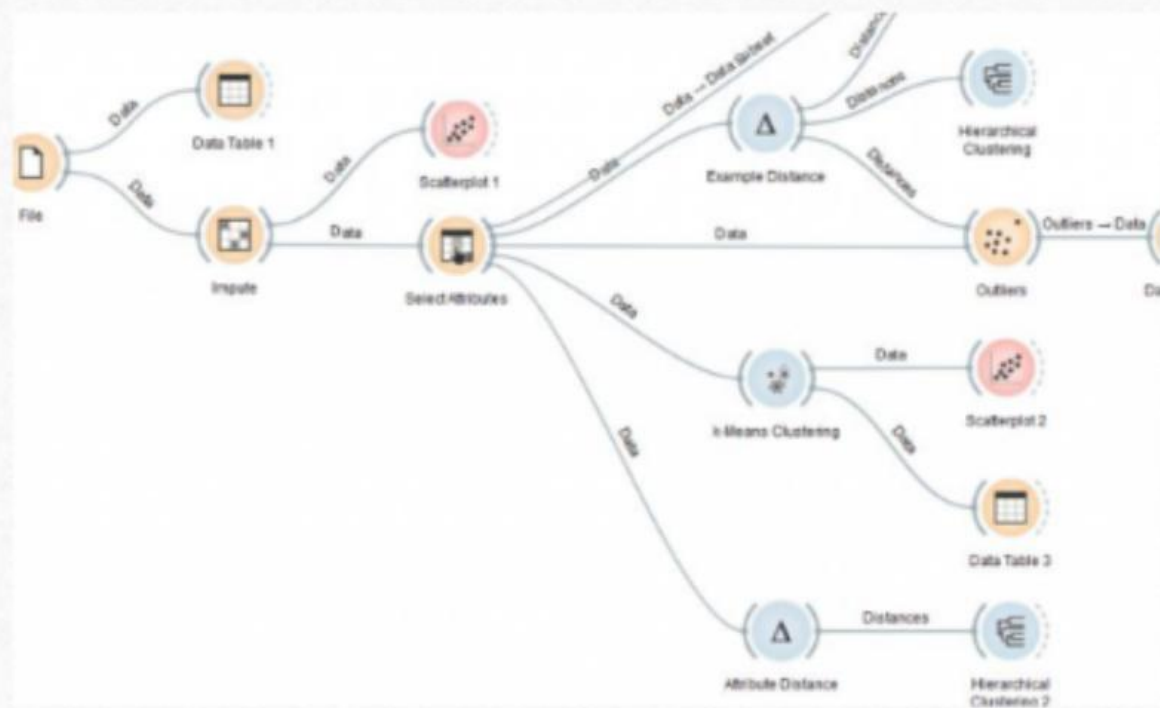
**Что это
даёт?**

Прогнозы,
классификация,
поиск групп и
аномалий.

**Сильный алгоритм не спасает плохие данные и неправильный
вопрос.**

Orange: визуальная разработка

Можно анализировать данные как схему: блок за блоком.



Что видно в Orange?

Данные идут от блока к блоку: таблица → график → модель → проверка. Такой путь легче обсуждать и исправлять.

Визуальный интерфейс снижает порог входа, но не отменяет мышление аналитика.

Вычисления стали доступнее

Раньше нужны были дорогие компьютеры. Теперь ресурсы можно взять на время.



Дешёвые устройства

Маленькие компьютеры стали символом доступности технологий.



Облачные мощности

Можно арендовать вычисления только на время эксперимента.



Ниже барьер входа

Можно учиться, пробовать и запускать модели без собственной серверной.



Это не значит “всё стало легко”. Это значит: практиковаться стало намного доступнее.

ИИ стал помощником аналитика

Он ускоряет работу, но не берёт ответственность за вывод.



ИИ помогает

предложить гипотезу, написать код, объяснить результат, найти возможную ошибку.



ИИ ошибается

может уверенно выдумать факт, источник, число или объяснение.



Человек проверяет

данные, график, модель, смысл и итоговый вывод.



Именно поэтому в эпоху ИИ аналитическое мышление становится ещё важнее.

От “самой сексапильной профессии” к профессии эпохи ИИ

В 2012 HBR ярко сформулировал роль data scientist. Сегодня смысл стал шире.

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

Summary Save Share 17 Comment Print \$8.95 Buy Copies

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives

2012 **2012**

Data Scientist называли “the sexiest job of the 21st century”. Тогда бизнес только массово осознал ценность данных.

2026 **Сегодня**

В эпоху ИИ особенно ценится человек, который отличает убедительный ответ от обоснованного.

Итоги блока

Что важно запомнить перед переходом к следующей теме.

1

Данные есть почти везде

числа, тексты, изображения, координаты, действия и измерения

2

Данных стало очень много

интернет, устройства, сервисы и эксперименты создают постоянный поток

3

Анализ нужен разным областям

наука, бизнес, медицина, климат и общество используют данные для решений

4

Главное остаётся за человеком

Orange, дешёвые ресурсы и ИИ помогают, но человек понимает смысл и проверяет вывод

Следующий шаг: смотрим на данные и задаём первые аналитические вопросы.

Встреча 1 - часть 2.5

ЧТО ТАКОЕ DATA SCIENCE?

Наука о данных, её логика, инструменты и роль человека

**Данные сами по себе молчат.
Нужно уметь задавать им правильные
вопросы.**

данные

вопрос

модель

проверка

математика
программирование
смысл

Что такое Data Science?

Наука о данных — это умение превращать данные в проверяемые выводы и решения.

**Data Science = данные + вопрос + метод + вычисления +
проверка + смысл**

данные

что мы наблюдаем

вопрос

что именно мы хотим
узнать

метод

как будем исследовать

проверка

можно ли доверять
выводу

Поэтому Data Science — это не просто таблицы и не просто код.

Цитата Дональда Кнута

Donald Ervin Knuth > Quotes > Quotable Quote



"Science is what we understand well enough to explain to a computer; art is everything else."

— Donald E. Knuth, Things a Computer Scientist Rarely Talks About

tags: art, computer, science

**«Наука — это то, что мы понимаем настолько хорошо,
что можем объяснить компьютеру.
Искусство — всё остальное, что мы делаем».**

— Дональд Кнут

Идея для курса: если мы умеем достаточно точно описать задачу, компьютер может помочь её исследовать.

**чётко
сформулировать
вопрос**

**определить
данные**

**описать шаги
анализа**

**проверить
результат**

От идеи к объяснению для компьютера

Компьютер полезен там, где задача описана достаточно чётко.

1

Идея

что нас интересует

2

Вопрос

что нужно узнать

3

Данные

что у нас есть

4

Алгоритм

какие шаги
выполняем

5

Проверка

насколько ответ
надёжен

**Если мы не можем объяснить себе шаги анализа, то и компьютеру
доверять рано.**

Три опоры Data Science

Наука о данных держится не на одном таланте, а на сочетании трёх вещей.

Математика и статистика

чтобы считать, сравнивать,
измерять неопределённость

Программирование и инструменты

чтобы работать с данными,
графиками и моделями

Знание предметной области

чтобы понимать смысл задачи
и не делать абсурдных
выводов

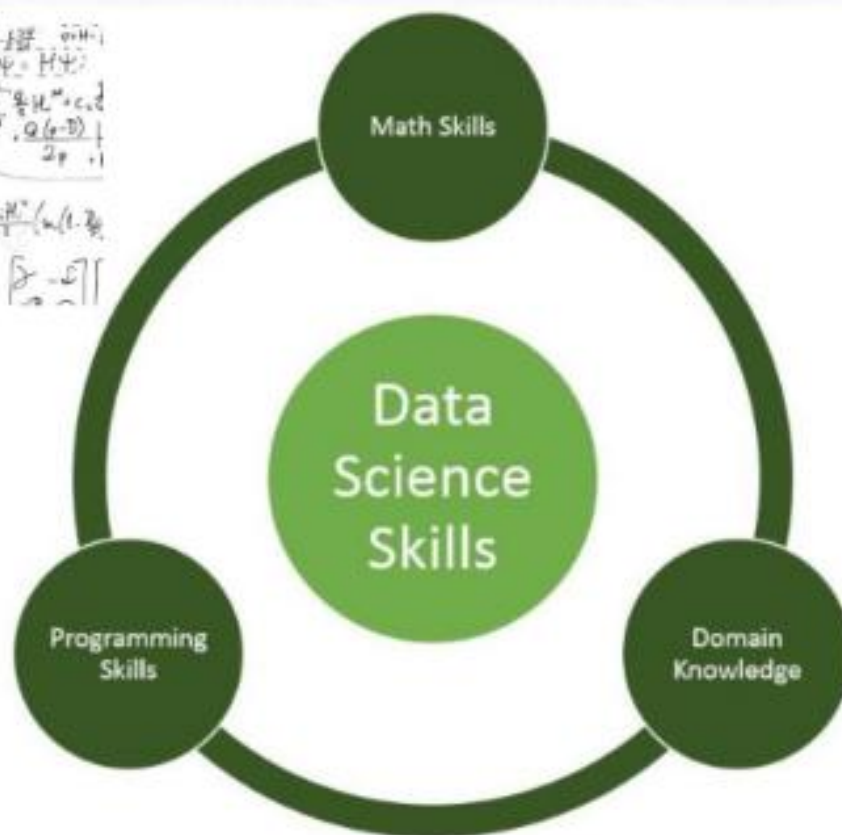
Сильный аналитик — не тот, кто знает только код или только формулы, а тот, кто умеет собрать всё вместе.

Три опоры Data Science

Наука о данных держится не на одном таланте, а на сочетании трёх вещей.

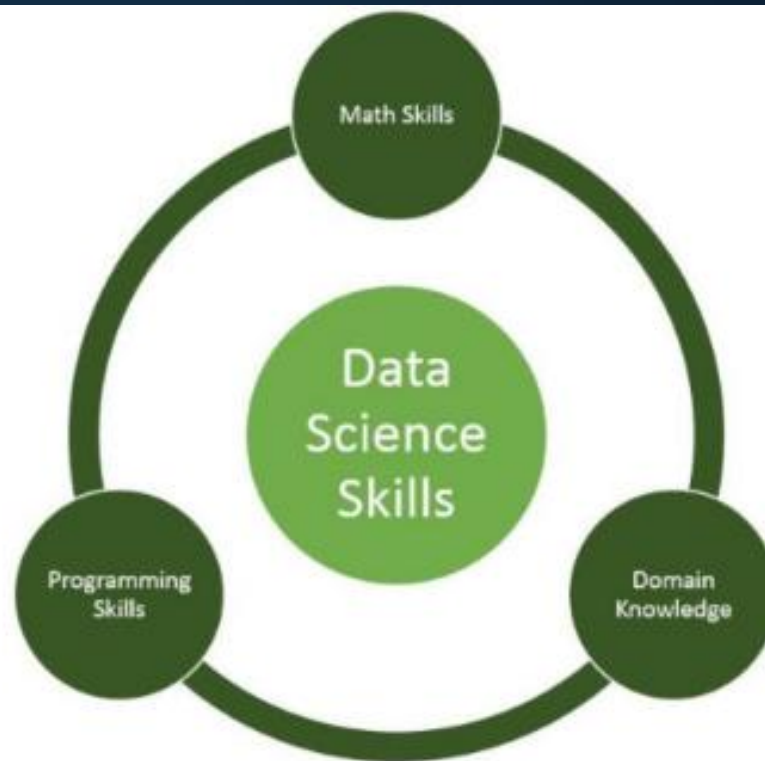
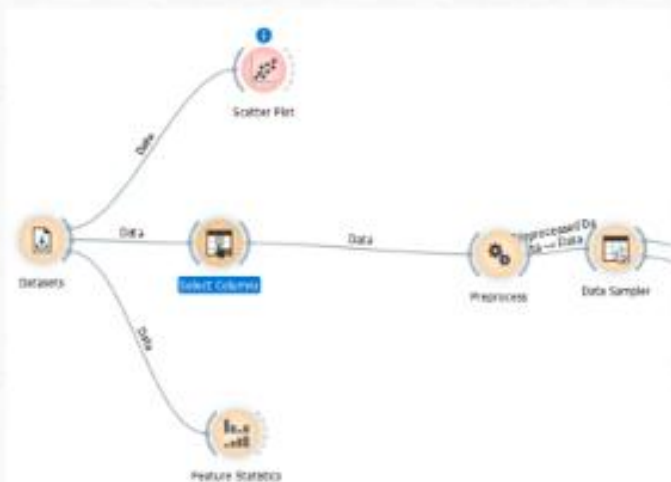
$$\begin{aligned} \mu &= \int x f(x) dx \\ \sigma^2 &= \int (x - \mu)^2 f(x) dx \\ H &= - \sum p(x) \log p(x) \\ Q, \varphi, \mu &= \sum_{i=1}^n \left[\frac{1}{\sigma^2} S_i + c_i T_i + \frac{1}{2} \log \left(\frac{1}{\sigma^2} \right) \right] \end{aligned}$$

```
1 #!/usr/bin/env python
2 import sys
3 import os
4 import simpleknn
5 from bigfile import BigFile
6
7 if __name__ == "__main__":
8     trainCollection = 'toydata'
9     nimages = 2
10     feature = 'f1'
11     dim = 3
12
13     testCollection = trainCollection
14     testset = testCollection
15
16     featureDir = os.path.join(rootpath, trainCollect
```



Три опоры Data Science

Наука о данных держится не на одном таланте, а на сочетании трёх вещей.



	Name	tributi	Center	Dispersion
1	nr.employed		5166.48	0.01
2	euribor3m		3.62	0.48
3	cons.conf.idx		-40.50	-0.11
4	cons.price.idx		93.58	0.01
5	emp.var.rate		0.08	18.39
6	previous		0.19	2.85
7	pdays		5.86	0.67

Data Scientist как переводчик

Одна из главных ролей аналитика — переводить реальную задачу на язык данных и обратно.



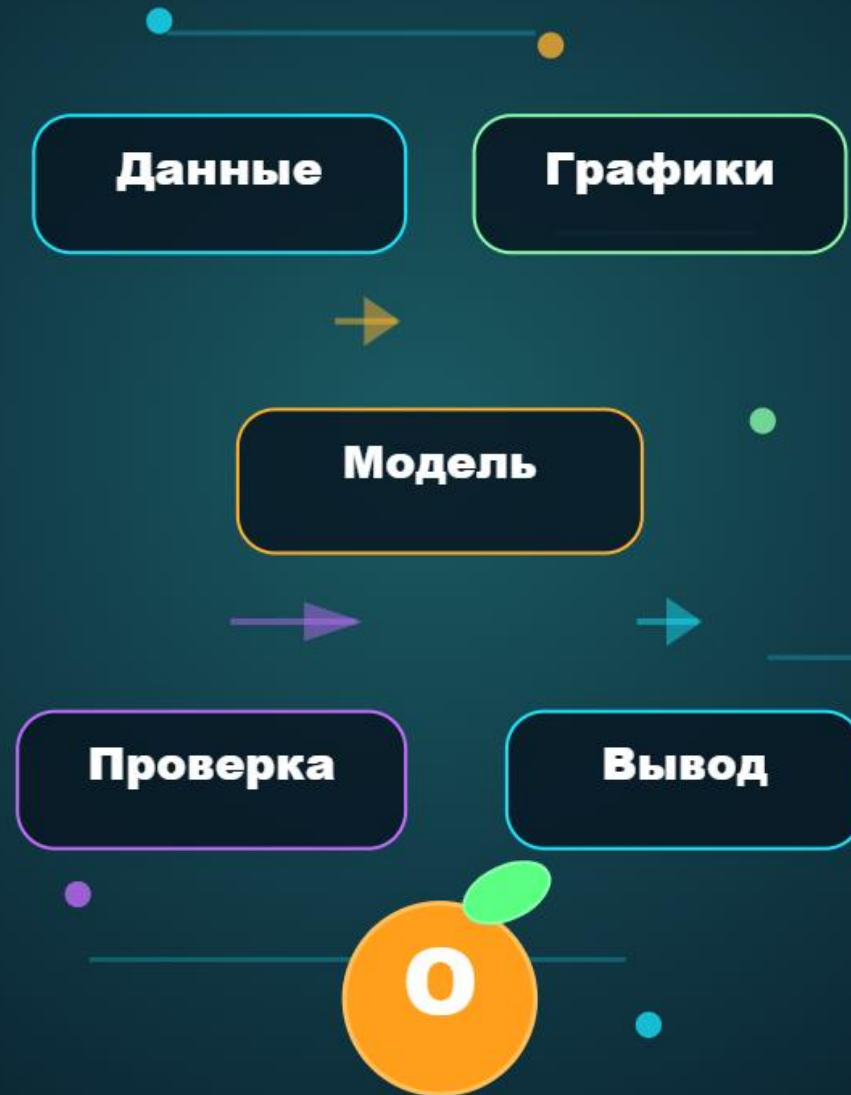
Хороший аналитик не просто считает — он связывает данные, вопрос и смысл.

ORANGE

от выбора инструмента
до первого запуска

Сначала выбираем правильный инструмент.
Потом устанавливаем и открываем первую схему.

Доктор Игорь Клейнер



Что мы изучим

В этой части мы не просто ставим программу, а понимаем, зачем она нужна.

1

**зачем нужен
Orange**

2

**как сравнить
Orange, R и
ChatGPT**

3

**как выбрать
инструмент
под задачу**

4

**как скачать
и установить
Orange**

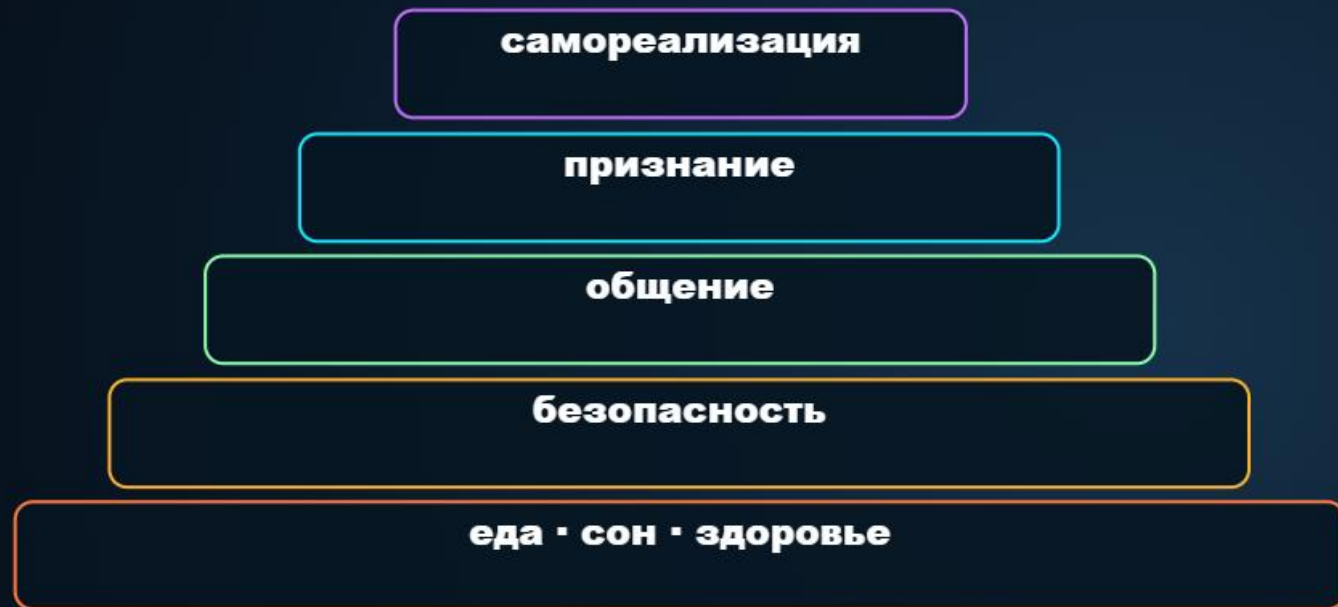
5

**как открыть
первую схему**

Итог: Orange открыт, рабочее поле готово, можно добавлять первый виджет.

Пирамида потребностей человека

Нам нужна не психология, а удобная аналогия.



Зачем здесь пирамида?

У каждой деятельности есть нижний этаж. Без него вершина не держится.



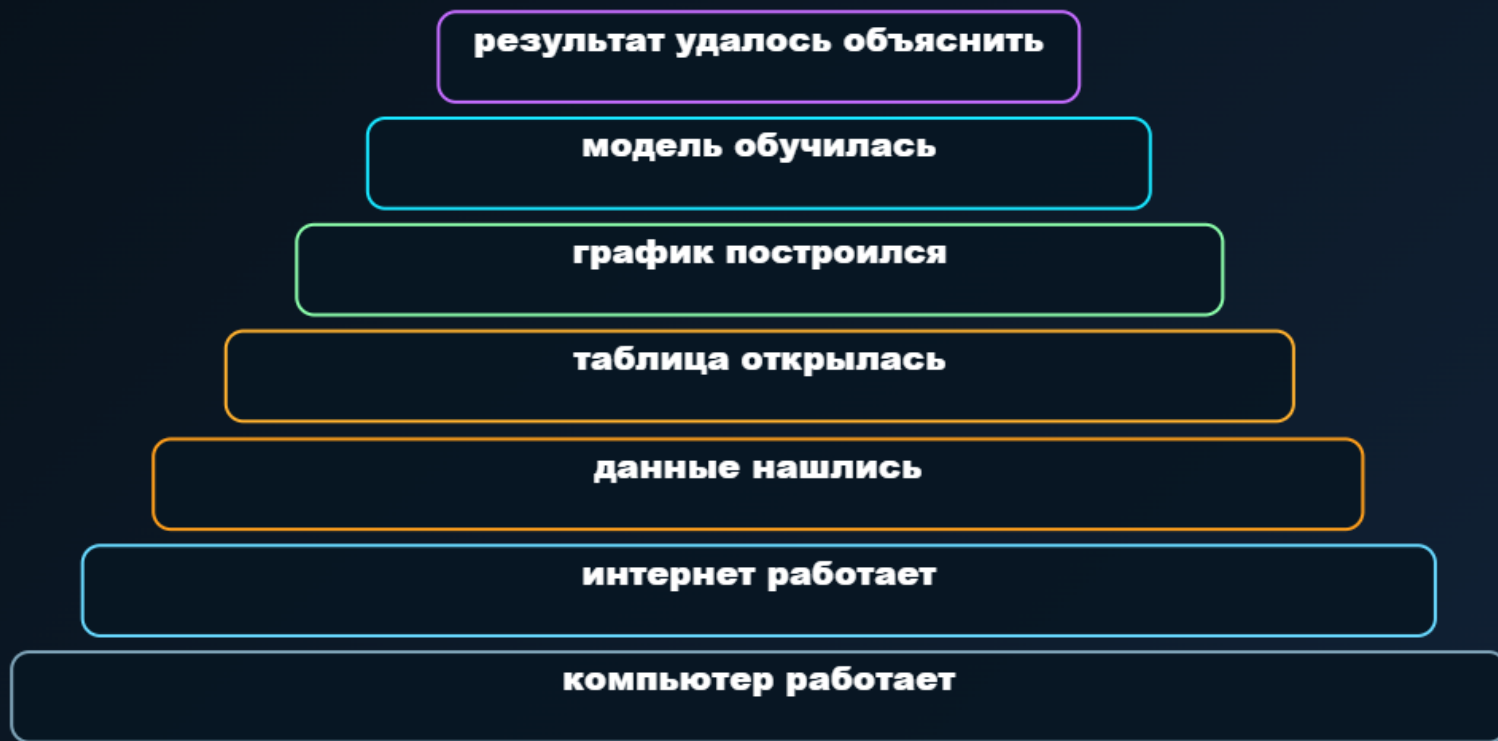
Переход к данным

У аналитика тоже есть базовые потребности: компьютер, данные, графики и проверка.

Через минуту построим версию для аналитика.

Пирамида потребностей аналитика данных

Шутка шуткой, но порядок очень важен.



Почему Orange?

Он помогает быстро пройти путь:
данные → график → модель →
проверка

Пока не начинаем
с тяжёлого программирования.

Wi-Fi — фундамент пирамиды,
о котором Маслоу не успел написать.

Пирамида потребностей аналитика

версия 2026: от помощника к просветлению

ПРОСВЕТЛЕНИЕ

модель работает,
и мы понимаем почему

СТАТИСТИКА

не страшная, а полезная

ПРОГРАММИРОВАНИЕ

ещё одна библиотека —
и всё точно заработает

KNIME

ещё одна визуальная лаборатория

ORANGE

мы здесь: быстро видим анализ

CHATGPT

помощник: идеи, объяснения, вопросы

Выше поднимемся

постепенно

KNIME, программирование и статистика
не исчезают. Мы просто не начинаем с
самого тяжёлого этажа.

Потом — Orange

Видим схему анализа: данные, графики,
модели и проверка.

Начинаем с помощника

ChatGPT помогает формулировать вопросы,
но не заменяет проверку.

Wi-Fi — фундамент, но ChatGPT теперь живёт ещё ниже в подвале.

Одна задача — три пути

Проверим инструменты на простой и понятной задаче.

Задача

**Определить, является
ли
сообщение спамом**

Вы выиграли приз! Нажмите ссылку

Ваш код подтверждения: 4821

Скидка 90% только сегодня



Orange

визуальная
лаборатория

R

R

код и полный
контроль



ChatGPT

помощник и
объяснения

**Вопрос аудитории: какой путь вы выбрали бы
сейчас?**

Что должна сделать модель

Классификация: из примеров учимся относить новые объекты к классам.

1

**есть
примеры**
спам / не спам

2

**модель
учится**
ищет признаки

3

**приходит
новое SMS**
ещё без ответа

4

**получаем
прогноз**
класс +
вероятность

!

Правила могут помочь

Слова “выигрыш”, “приз”, “бесплатно” часто подозрительны, но ручных правил мало.

✓

Нужна проверка качества

Модель должна проверяться на данных, которые она не видела при обучении.

Не просто “угадать”, а научиться проверять качество

Классификация сообщений: спам \ не спам

	text	type
1	Hope you are having a good week. Just checking in	ham
2	K..give back my thanks.	ham
3	Am also doing in cbe only. But have to pay.	ham
4	complimentary 4 STAR Ibiza Holiday or £10,000 cash needs your URGENT collection. 09066364349 NOW from Landline not to lose out! Box434SK38WP150PPM18+	spam
5	okmail: Dear Dave this is your final notice to collect your 4* Tenerife Holiday or #5000 CASH award! Call 09061743806 from landline. TCs SAE Box326 CW25WX 150ppm	spam
5	Aiya we discuss later lar... Pick u up at 4 is it?	ham
7	Are you this much buzy	ham
3	Please ask mummy to call father	ham
3	Marvel Mobile Play the official Ultimate Spider-man game (£4.50) on ur mobile right now. Text SPIDER to 83338 for the game & we ll send u a FREE 8Ball wallpaper	spam
10	fyi I'm at usf now, swing by the room whenever	ham
11	Sure thing big man. i have hockey elections at 6, shouldn€t go on longer than an hour though	ham
12	I anything lor...	ham
13	By march ending, i should be ready. But will call you for sure. The problem is that my capital never complete. How far with you. How's work and the ladies	ham
14	Hmm well, night night	ham
15	K I'll be sure to get up before noon and see what's what	ham
16	Ha ha cool cool chikku chikku:-):-DB-)	ham
17	Darren was saying dat if u meeting da ge den we dun meet 4 dinner. Cos later u leave xy will feel awkward. Den u meet him 4 lunch lor.	ham
18	He dint tell anything. He is angry on me that why you told to abi.	ham
19	Up to u... u wan come then come lor... But i din c any stripes skirt...	ham
20	U can WIN £100 of Music Gift Vouchers every week starting NOW! Txt the word DRAW to 87066 ToC:www.idon.com SkillGame 1Winaweek 1pp15 150ppnormanSubscription	spam

Решение в Orange

Визуальная схема показывает весь путь анализа.



Что видно

Данные, обработка, модель, проверка и прогноз лежат на одном рабочем поле.



Что проверяем

Открываем виджет и смотрим, какие данные пришли и что получилось.

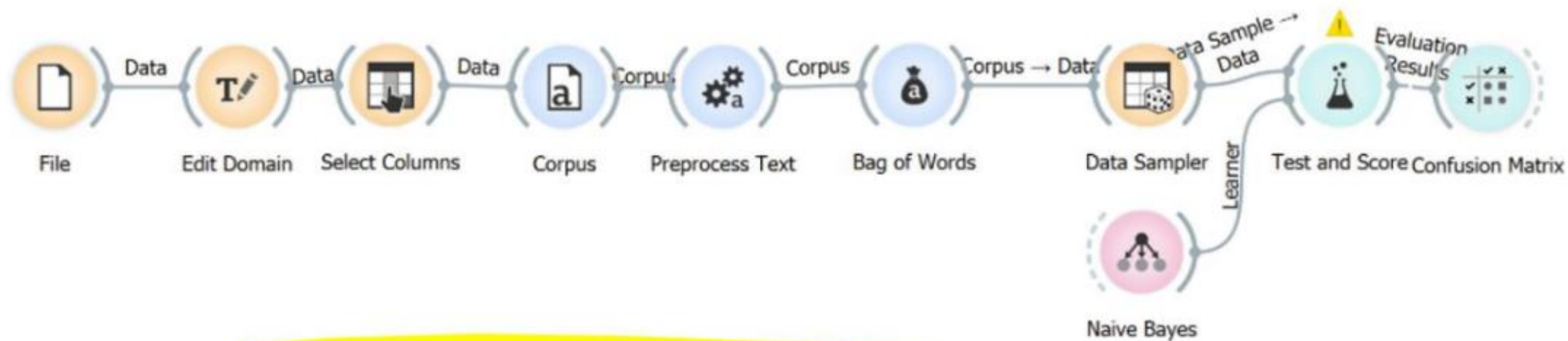


Почему удобно

Сначала понимаем логику анализа, а не боремся с синтаксисом.

Главная идея: процесс виден, поэтому каждый этап можно обсудить и проверить.

Магия анализа данных: решение на Orange



Решение на языке R

R даёт полный контроль, но требует программирования.

Магия анализа данных: решение на R

```
# read the sms data into the sms data frame
sms_raw <- read.csv("sms_spam.csv", stringsAsFactors = FALSE)

# examine the structure of the sms data
str(sms_raw)

# convert spam/not to factor
sms_raw$type <- factor(sms_raw$type)

# examine the type variable more carefully
str(sms_raw$type)
table(sms_raw$type)

# build a corpus using the text mining (tm) package
library(tm)
sms_corpus <- VCorpus(VectorSource(sms_raw$text))

# examine the sms corpus
print(sms_corpus)
inspect(sms_corpus[1:2])

as.character(sms_corpus[[1]])
inspect(sms_corpus[1:2], as.character)

# clean up the corpus using tm_map()
sms_corpus_clean <- tm_map(sms_corpus, content_transformer(tolower))

# show the difference between sms_corpus and corpus_clean
as.character(sms_corpus[[1]])
as.character(sms_corpus_clean[[1]])

sms_corpus_clean <- tm_map(sms_corpus_clean, removeNumbers) # remove numbers
sms_corpus_clean <- tm_map(sms_corpus_clean, removeWords(stopwords())) # remove stop words
```

```
sms_corpus_clean <- tm_map(sms_corpus_clean, removePunctuation) # remove punctuation

# tip: create a custom function to replace (rather than remove) punctuation
removePunctuation("hello...world")
replacePunctuation <- function(x) { gsub("[:punct:]", "", x) }
replacePunctuation("hello...world")

# illustration of word stemming
library(SnowballC)
wordStem(c("learn", "learned", "learning", "learns"))

sms_corpus_clean <- tm_map(sms_corpus_clean, stemDocument())

sms_corpus_clean <- tm_map(sms_corpus_clean, stripWhitespace) # eliminate unneeded whitespace

# examine the final clean corpus
inspect(sms_corpus_clean[1:3], as.character)
inspect(sms_corpus_clean[1:3], as.character)

# create a document-term sparse matrix
sms_dtm <- DocumentTermMatrix(sms_corpus_clean)

# alternative solution: create a document-term sparse matrix directly from the sms corpus
sms_dtm2 <- DocumentTermMatrix(sms_corpus, control = list(
  tolower = TRUE,
  removeNumbers = TRUE,
  stopwords = TRUE,
  removePunctuation = TRUE,
  stemming = TRUE
))

# alternative solution: using custom stop words function ensures identical result
sms_dtm3 <- DocumentTermMatrix(sms_corpus, control = list(
```

R

Что делает код

Загрузка данных, обработка текста, обучение модели и оценка качества.



Сильная сторона

Можно управлять каждой деталью и повторять анализ автоматически.

Шутка: “Ещё один пакет — и всё точно заработает”.

```

# read the sms data into the sms data frame
sms_raw <- read.csv("sms_spam.csv", stringsAsFactors = FALSE)

# examine the structure of the sms data
str(sms_raw)

# convert spam/ham to factor.
sms_raw$type <- factor(sms_raw$type)

# examine the type variable more carefully
str(sms_raw$type)
table(sms_raw$type)

# build a corpus using the text mining (tm) package
library(tm)
sms_corpus <- VCorpus(VectorSource(sms_raw$text))

# examine the sms corpus
print(sms_corpus)
inspect(sms_corpus[1:2])

as.character(sms_corpus[[1]])
lapply(sms_corpus[1:2], as.character)

# clean up the corpus using tm_map()
sms_corpus_clean <- tm_map(sms_corpus, content_transformer(tolower))

# show the difference between sms_corpus and corpus_clean
as.character(sms_corpus[[1]])
as.character(sms_corpus_clean[[1]])

sms_corpus_clean <- tm_map(sms_corpus_clean, removeNumbers) # remove numbers
sms_corpus_clean <- tm_map(sms_corpus_clean, removeWords, stopwords()) # remove stop words

sms_corpus_clean <- tm_map(sms_corpus_clean, removePunctuation) # remove punctuation

# tip: create a custom function to replace (rather than remove) punctuation
removePunctuation("hello...world")
replacePunctuation <- function(x) { gsub("[[:punct:]]+", " ", x) }
replacePunctuation("hello...world")

# illustration of word stemming
library(SnowballC)
wordStem(c("learn", "learned", "learning", "learns"))

sms_corpus_clean <- tm_map(sms_corpus_clean, stemDocument)

sms_corpus_clean <- tm_map(sms_corpus_clean, stripWhitespace) # eliminate unneeded whitespace

# examine the final clean corpus
lapply(sms_corpus[1:3], as.character)
lapply(sms_corpus_clean[1:3], as.character)

# create a document-term sparse matrix
sms_dtm <- DocumentTermMatrix(sms_corpus_clean)

# alternative solution: create a document-term sparse matrix directly from the SMS corpus
sms_dtm2 <- DocumentTermMatrix(sms_corpus, control = list(
  tolower = TRUE,
  removeNumbers = TRUE,
  stopwords = TRUE,
  removePunctuation = TRUE,
  stemming = TRUE
))

# alternative solution: using custom stop words function ensures identical result
sms_dtm3 <- DocumentTermMatrix(sms_corpus, control = list(

```

```

sms_dtm_freq_train <- removeSparseTerms(sms_dtm_train, 0.999)
sms_dtm_freq_train

# indicator features for frequent words
findFreqTerms(sms_dtm_train, 5)

# save frequently-appearing terms to a character vector
sms_freq_words <- findFreqTerms(sms_dtm_train, 5)
str(sms_freq_words)

# create DTMs with only the frequent terms
sms_dtm_freq_train <- sms_dtm_train[, sms_freq_words]
sms_dtm_freq_test <- sms_dtm_test[, sms_freq_words]

# convert counts to a factor
convert_counts <- function(x) {
  x <- ifelse(x > 0, "Yes", "No")
}

# apply() convert_counts() to columns of train/test data
sms_train <- apply(sms_dtm_freq_train, MARGIN = 2, convert_counts)
sms_test <- apply(sms_dtm_freq_test, MARGIN = 2, convert_counts)

## Step 3: Training a model on the data ----
library(e1071)
sms_classifier <- naiveBayes(sms_train, sms_train_labels)

## Step 4: Evaluating model performance ----
sms_test_pred <- predict(sms_classifier, sms_test)

library(gmodels)
CrossTable(sms_test_pred, sms_test_labels,
  prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
  dnn = c('predicted', 'actual'))

tolower = TRUE,
removeNumbers = TRUE,
stopwords = function(x) { removeWords(x, stopwords()) },
removePunctuation = TRUE,
stemming = TRUE
))

# compare the result
sms_dtm
sms_dtm2
sms_dtm3

# creating training and test datasets
sms_dtm_train <- sms_dtm[1:4169, ]
sms_dtm_test <- sms_dtm[4170:5559, ]

# also save the labels
sms_train_labels <- sms_raw[1:4169, ]$type
sms_test_labels <- sms_raw[4170:5559, ]$type

# check that the proportion of spam is similar
prop.table(table(sms_train_labels))
prop.table(table(sms_test_labels))

# word cloud visualization
library(wordcloud)
wordcloud(sms_corpus_clean, min.freq = 50, random.order = FALSE)

# subset the training data into spam and ham groups
spam <- subset(sms_raw, type == "spam")
ham <- subset(sms_raw, type == "ham")

wordcloud(spam$text, max.words = 40, scale = c(3, 0.5))
wordcloud(ham$text, max.words = 40, scale = c(3, 0.5))

```

Решение с ChatGPT

Нейросеть помогает, но не становится источником истины.

Диалог

Помоги построить классификатор спама
и объясни каждый шаг.

Могу предложить план, написать код,
объяснить метрики и помочь найти ошибку.

Но данные и результат всё равно проверяет
человек.

+

Может помочь

- составить план
- написать код
- объяснить результат
- найти ошибку

!

Может ошибиться

- придумать столбцы
- неверно понять данные
- написать код с ошибкой
- уверенно объяснить не то

Правило курса остаётся тем же: **“Стоп.
Проверяем!”**

«Я уверен на 97%.
Правда, не уверен, откуда взял 97%.»

**СТОП
ПРОВЕРЯЕМ**

Orange или R?

Это не война инструментов. Это разные роли.

Orange

- быстрый старт
- визуальный процесс
- меньше технических препятствий
- удобно изучать идеи

R

- полный контроль
- сложная обработка
- автоматизация
- воспроизводимый код

Вывод: Orange — быстрее понять. R — глубже управлять.

Orange или ChatGPT?

Лаборатория и помощник рядом с ней.

Orange

- работает с загруженными данными
- показывает графики и модели
- считает качество
- каждый шаг виден

ChatGPT

- объясняет
- предлагает идеи
- помогает исправлять ошибки
- пишет инструкции и код

Вывод: Orange — лаборатория. ChatGPT — помощник

R или ChatGPT?

Код выполняет анализ. Нейросеть помогает с мышлением и текстом.

R

- выполняет точные команды
- повторяет анализ
- работает в больших проектах
- не придумывает результат

ChatGPT

- ускоряет написание кода
- переводит ошибку на язык людей
- предлагает варианты
- иногда пишет уверенно неверно

Вывод: ChatGPT помогает писать на R, но R выполняет анализ.

Миссия: выберите инструмент

Три ситуации. Выбираем не “лучший”, а подходящий инструмент.

МИНИ-КВИЗ

Быстро понять принцип модели

Orange

Сложный автоматизированный

R

Объяснить ошибку или предложить

ChatGPT

**В реальной
инструменты часто
используются
вместе.**

Orange + R + ChatGPT
может быть сильнее,
чем каждый отдельно.

Почему курс начинается с Orange

Сначала понимаем анализ. Потом усложняем инструменты.



**начинаем
сразу**



**не тратим
старт
на синтаксис**



**видим
структуру
анализа**



**быстро
экспериментируем**



**проверяем
модели**

Orange не умеет всё. Но он отлично подходит, чтобы увидеть механику анализа данных.

Перед установкой

Сначала проверяем базовые вещи, чтобы не тратить занятие на хаос.

операционная система

Windows, macOS или Linux

место на диске

оставьте запас

права на установку

или portable-версия

официальный сайт

никаких случайных загрузок

открытые документы

сохраните перед установкой

Не скачиваем “Orange Pro Ultra 2026” с сайта, где десять кнопок Download.

Где скачать и как запустить

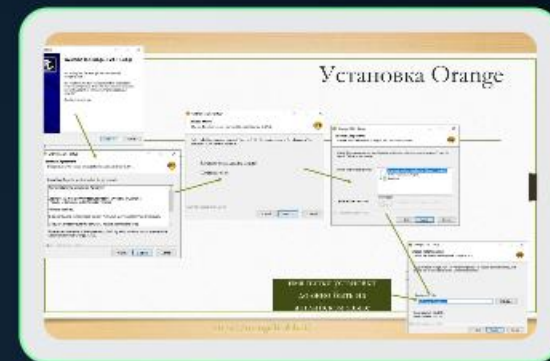
Используем официальный сайт Orange Data Mining.

1 открыть официальный сайт
orangedatamining.com/download

2 выбрать систему
Windows / macOS / Linux

3 скачать установщик
не со случайного сайта

4 запустить Orange
создать новый workflow



Источник: официальный сайт Orange Data Mining

УРОВЕНЬ ПРОЙДЕН

Orange готов к первой практике

- ✓ **Orange установлен**
- ✓ **Программа открылась**
- ✓ **Создан новый workflow**

Награда:
доступ в
лабораторию
данных

Честное уточнение: это только начало.
Пока мы ещё не построили ни одной модели.

Что мы узнали

Финальная проверка перед первой практической схемой.

Orange

визуальная лаборатория анализа данных

R

гибкий язык программирования

ChatGPT

ИИ-помощник, не источник истины

Вместе

инструменты дополняют друг друга

Старт курса

начинаем с Orange

Критерий успеха

Orange открыт и готов к работе

Следующая часть: первая схема в Orange

Какой виджет добавим первым?

СЕКРЕТНАЯ КАРТА №3

ПОЗВОЛЬТЕ ПОДЕЛИТЬСЯ С ВАМИ МАГИЕЙ ORANGE

Позвольте
поделиться
с вами магией
Orange!



1

ЗАГРУЗКА
ДАННЫХ



Загрузи файл
и начни путь.

2

ВИЗУАЛИЗАЦИЯ



Смотри на данные
глазами
исследователя.

3

МОДУЛИ



Соединяй виджеты
и собирай анализ.

4

МОДЕЛИ



Пробуй прогнозы
без сложного
кода.

5

ВЫВОДЫ



Из данных
рождаются
идеи.

Orange – это не только фрукт.
Это портал в анализ данных.

Да,
магия
настоящая!

УЧИСЬ. ИССЛЕДУЙ. СОЕДИНЯЙ ВИДЖЕТЫ.

ВСТРЕЧА 1 · ЧАСТЬ 4А

ИСТОРИИ ДАННЫХ

Watson / Jeopardy и eHarmony

Что происходит, когда данные становятся продуктом и игроком?

Доктор Игорь Клейнер



Что изучим в этой части

Две истории: интеллект в викторине и алгоритм совместимости

1

Jeopardy

почему игра трудна для
компьютера

2

Watson

как система выбирает ответ
и оценивает уверенность

3

eHarmony

как данные превращаются
в продукт

4

Проверка

где заканчивается “магия
алгоритма”

Главный вопрос: когда данные помогают, а когда красивая история
требует проверки?

Снаружи анализ данных выглядит как

Данные вошли — чудо вышло? Очень заманчиво, но опасно

Мифическая

данные



чёрный ящик



деньги, подарки, победы,
“идеальная
совместимость”

Если бы всё было так просто, Excel уже
правил бы миром.



На самом деле это не магия

Полезный результат рождается из процесса

Данные

Вопрос

Человек

ИИ-помощник

Компьютер

Проверка

Данные + вопрос + человек + ИИ + вычисления +
проверка → полезный результат

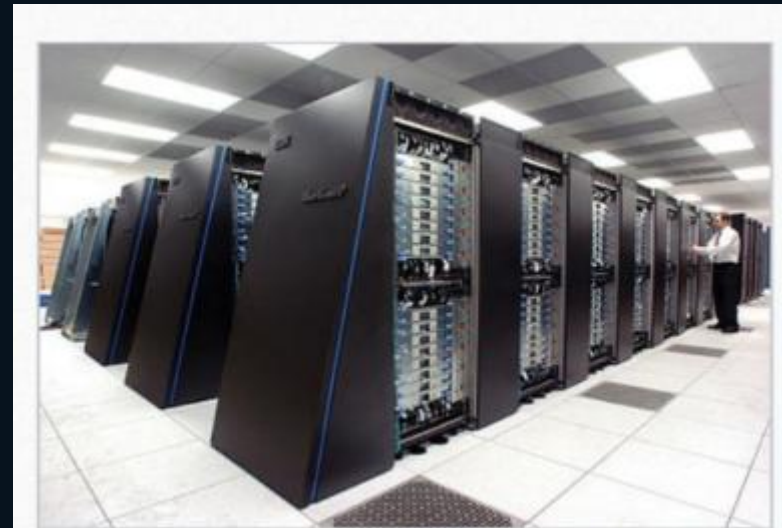


IBM Watson



IBM Watson

- IBM вкладывает немало средств в науку
- Компьютер Deep Blue победил чемпиона мира по шахматам
- Компьютер Blue Gene помогает ученые анализировать геном человека



Jeopardy

Jeopardy – американская телевизионная игра

Участники должны придумать вопрос на показанный им ответ

THE DINOSAURS	NOTABLE WOMEN	OXFORD ENGLISH DICTIONARY	NAME THAT INSTRUMENT	BELGIUM	COMPOSERS BY COUNTRY
\$200	\$200	\$200	\$200	\$200	\$200
\$400	\$400	\$400	\$400	\$400	\$400
\$600	\$600	\$600	\$600	\$600	\$600
\$800	\$800	\$800	\$800	\$800	\$800
\$1000	\$1000	\$1000	\$1000	\$1000	\$1000

Что такое Jeopardy

Игра выглядит как викторина, но устроена необычно

?

Правило наоборот

Игрок видит формулировку-ответ и должен произнести правильный вопрос.

Aa

Нужен язык

Подсказки содержат игру слов, намёки, категории и культурный контекст.



Нужна скорость

Нужно понять подсказку и успеть нажать кнопку раньше соперников.



Нужна уверенность

Если ответ неверный, игрок теряет деньги. Молчание иногда лучше риска.



Jeopardy

В 2005 решила IBM создать компьютер который научиться и сможет играть в эту игру

В феврале 2011, компьютер Ватсон победил в игре 2х лучших игроков

https://www.youtube.com/watch?v=WFR3lOm_xhE



Watson выходит на сцену

IBM поставила компьютеру задачу играть с лучшими людьми



IBM

Проект IBM

Система Watson была создана для игры Jeopardy.

vs

Соперники

Watson играл против Ken Jennings и Brad Rutter.

2011

Год

Публичные матчи Jeopardy прошли в 2011 году.

Аа

Смысл

Символ: машина работает с естественным языком.

Jeopardy

Ватсон – супер компьютер с 3000 процессоров
Ватсон «прочитал» множеств книг, журналов, газет, энциклопедий
С помощью алгоритмов для анализа текстов Ватсон победил



Watson побеждает чемпионов

Но победа не означает “машина всё понимает как человек”

Watson

\$77,147

Ken Jennings

\$24,000

Brad Rutter

\$21,600

Важно: Watson тоже ошибался. В одном Final Jeopardy он ответил “Toronto?????” вместо Chicago.

Как Watson выбирает ответ

Упрощённая схема: от подсказки до решения

Подсказка

текст из игры

Кандидаты

возможные ответы

Доказательства

факты и совпадения

Уверенность

оценка вероятности

Buzz?

отвечать или молчать

Чем больше независимых алгоритмов приходили к одному ответу, тем выше была уверенность Watson.

Уверенность: отвечать или молчать

В анализе данных часто важен не только ответ, но и уверенность

Ответ А



достаточно уверенности → можно
отвечать

Ответ В



низкая уверенность → лучше
промолчать

В курсе мы тоже будем спрашивать: модель уверена или просто красиво звучит?

Урок Watson

Не “магия ответа”, а система проверки кандидатов



1. Данные

Большая база текстов и фактов не превращается автоматически в понимание.



2. Алгоритмы

Нужны методы, которые ищут кандидатов и собирают доказательства.



3. Уверенность

Решение зависит от оценки качества ответа, а не только от самого ответа.



4. Проверка

Ошибки возможны даже при сильной системе. Поэтому проверяем.

Фраза курса: если ответ красивый, но непонятно откуда он взялся — это не конец анализа, а начало проверки.

От викторины к людям

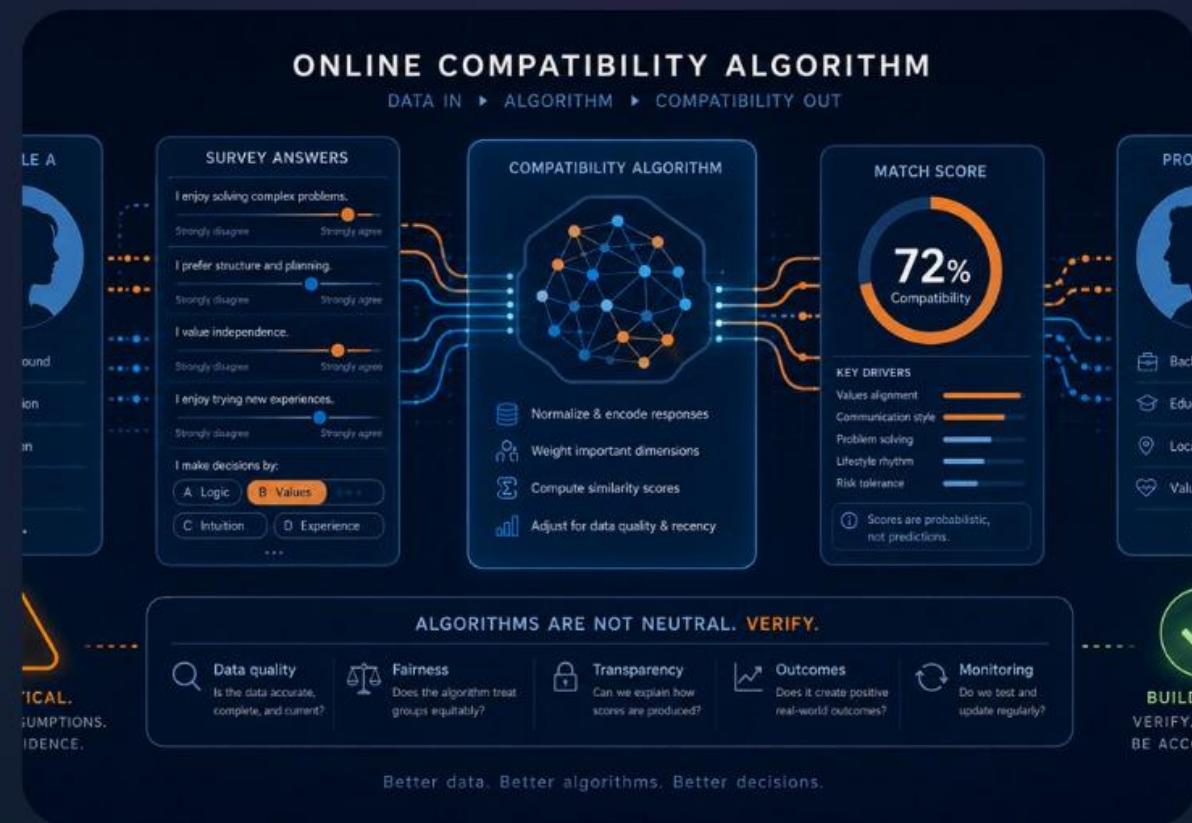
Теперь данные пытаются предсказывать не ответ, а совместимость

Вопрос меняется

Watson спрашивает:
“Какой ответ правильный?”

eHarmony обещает:
“Кто вам подходит?”

Чем ближе к людям,
тем важнее осторожность.



От викторины к людям

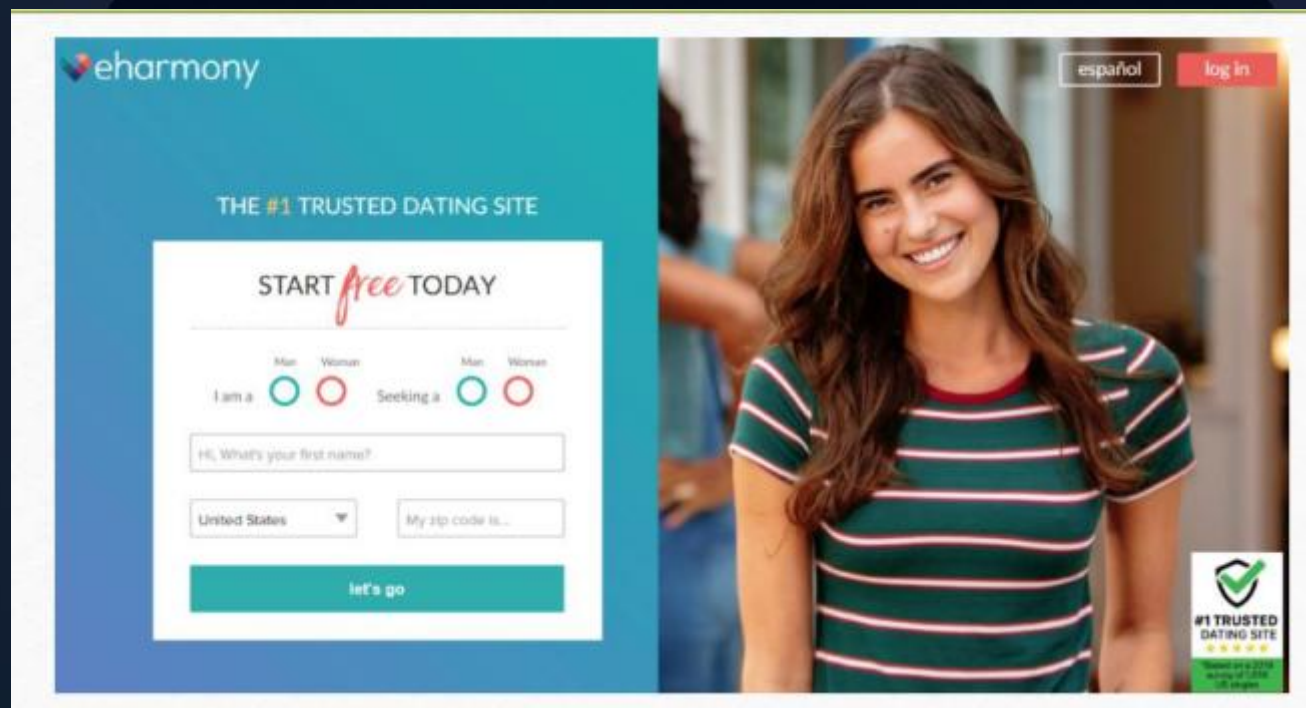
Теперь данные пытаются предсказывать не ответ, а совместимость

Вопрос меняется

Watson спрашивает:
“Какой ответ правильный?”

eHarmony обещает:
“Кто вам подходит?”

Чем ближе к людям,
тем важнее осторожность.



Better data. Better algorithms. Better decisions.

От викторины к людям

Теперь данные пытаются предсказывать не ответ, а совместимость

- «Научный» подход к отношениям и любви
- Сайт с прибылью более миллиарда долларов
- 4-5% браков в США при участии eHarmony

eHarmony is More than Traditional Dating Sites

At eHarmony, we deliver more than just dates. Our patented Compatibility Matching System® narrows the field from thousands of single men or single women to match with a highly select group of compatible singles -- singles who have been prescreened on 29 Dimensions® of Compatibility: scientific predictors of long-term relationship success. Why spend hours browsing through personal ads and searching the Internet when you don't have to? It's our goal at eHarmony to match you with the best possible person for you and we are eager to provide you with the best dating experience that we possibly can.



eHarmony: данные как продукт

Алгоритм становится частью обещания сервиса

1 Анкета

Пользователь отвечает на вопросы о себе, ценностях и отношениях.

2 Профиль

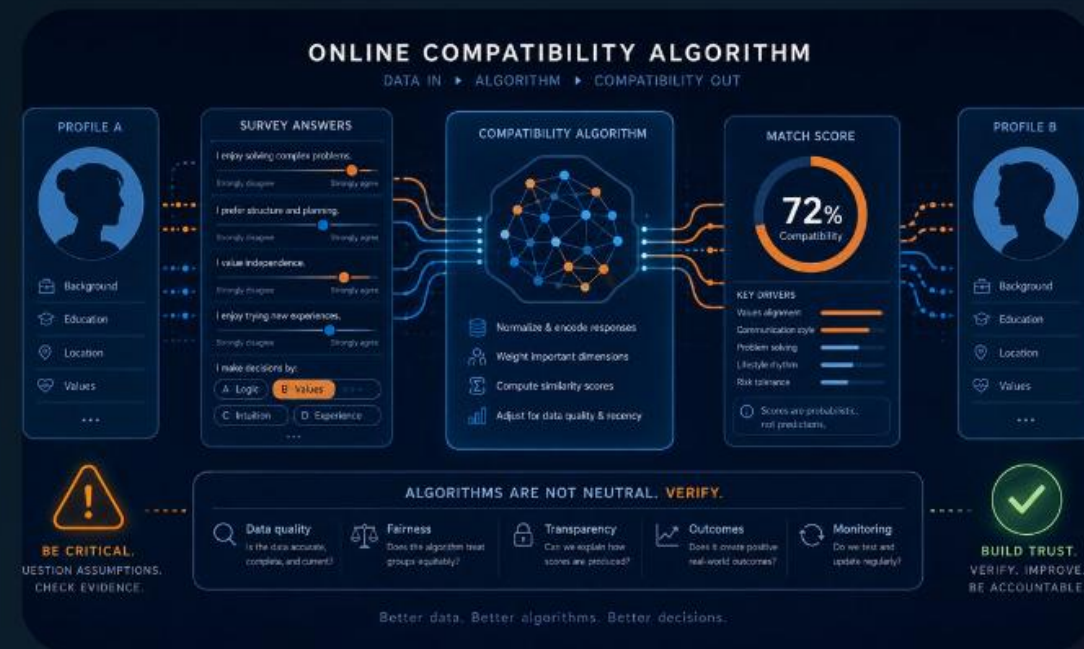
Ответы превращаются в признаки и описание личности.

3 Сопоставление

Сервис сравнивает профили пользователей.

4 Оценка

Пользователь видит совместимость с конкретным человеком.



Где нужна осторожность

Алгоритм может быть бизнесом — но это не автоматически наука



Какие признаки?

**Что именно
измеряет анкета:
ценности, привычки,
стиль общения,
ожидания?**



Какая проверка?

Есть ли независимые доказательства, что подбор улучшает реальные отношения?



Какая метрика?

**Что считается
успехом: сообщение,
свидание, долгие
отношения, брак?**



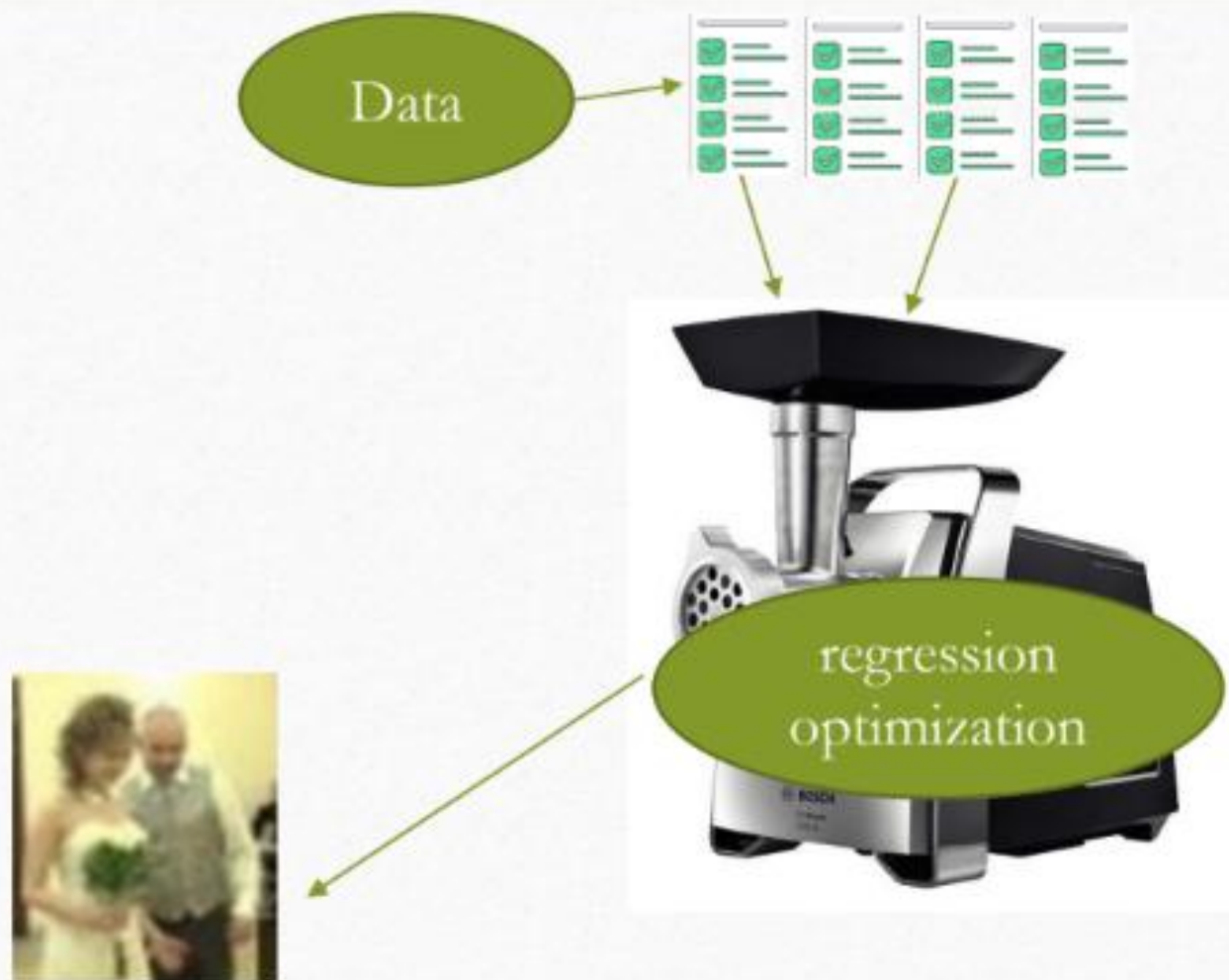
Какая честность?

**Пользователь
понимает, что score
— подсказка, а не
судьба?**

Шутка: совместимость может считать алгоритм, но посуду всё равно моет не модель.

От викторины к людям

Теперь данные пытаются предсказывать не ответ, а совместимость



Watson и eHarmony: что общего?

Две разные истории, но одна логика анализа

Watson

подсказка → кандидаты → уверенность →
ответ

цель: выиграть игру
риск: ответить неверно

eHarmony

анкета → признаки → score →
рекомендация

цель: подобрать людей
риск: продать красивую уверенность

В обоих случаях нам нужны данные, модель, оценка уверенности и проверка результата.

ПОЗДРАВЛЯЮ!

Вы нашли секретную

Watson показал: важна не только версия ответа,
но и уверенность в нём.

eHarmony показал: красивое обещание алгоритма
нужно проверять вопросами к данным и результатам.

**Главное сокровище — не карта,
а умение не теряться в данных.**



Встреча 1 · Часть 4В

CHALLENGER И MONEYBALL

Истории данных: ошибка и стратегия

Когда данные нужно не только иметь,
но и правильно использовать.

Что изучим в этой части

Две истории — два очень разных урока анализа данных



Challenger

Как данные, инженерные опасения и коммуникация решений связаны с риском. Мы разберём историю осторожно и без упрощения до “одного графика”.



Moneyball

Как статистический взгляд помог пересмотреть привычные решения в спорте и искать недооценённые возможности.

Главный вопрос: как данные меняют решение, если их правильно понять?

Данные могут предупреждать и менять правила

Но они не спасают автоматически



Предупреждать

показать риск раньше, чем он станет катастрофой



Ломать привычки

поставить под сомнение “так всегда делали”



Искать ценность

найти то, что рынок или экспертный глаз недооценил

Данные нужно правильно показать, понять и защитить перед решением.

Challenger: данные были, но их не услышали

Январь 1986 года: запуск шаттла при необычно низкой температуре

- Ужасное событие в истории НАСА
- Шаттл взорвался спустя 73 секунды от взлета



Challenger: данные были, но их не услышали

Январь 1986 года: запуск шаттла при необычно низкой температуре

- Первые 24 полета прошли успешно
- Катастрофа может произойти если все 5 колец-О ломаются
- Были полеты в которых ни сломалось ни одно кольцо
- Были полеты где сломалось от 1 до 3 колец



Challenger: данные были, но их не услышали

Январь 1986 года: запуск шаттла при необычно низкой температуре

- В день 25 полета было очень холодно – минус 1 градус по Цельсию (-30 по Фаренгейту)
- Не было данных о полетах при такой низкой температуре ☹



Challenger: данные были, но их не услышали

Январь 1986 года: запуск шаттла при необычно низкой температуре

1 Что произошло

Комиссия связала потерю Challenger с отказом уплотнений в правом твердотопливном ускорителе.

2 Почему это пример для нас

Были данные, опасения инженеров и решение. Почему сигнал риска не остановил запуск?

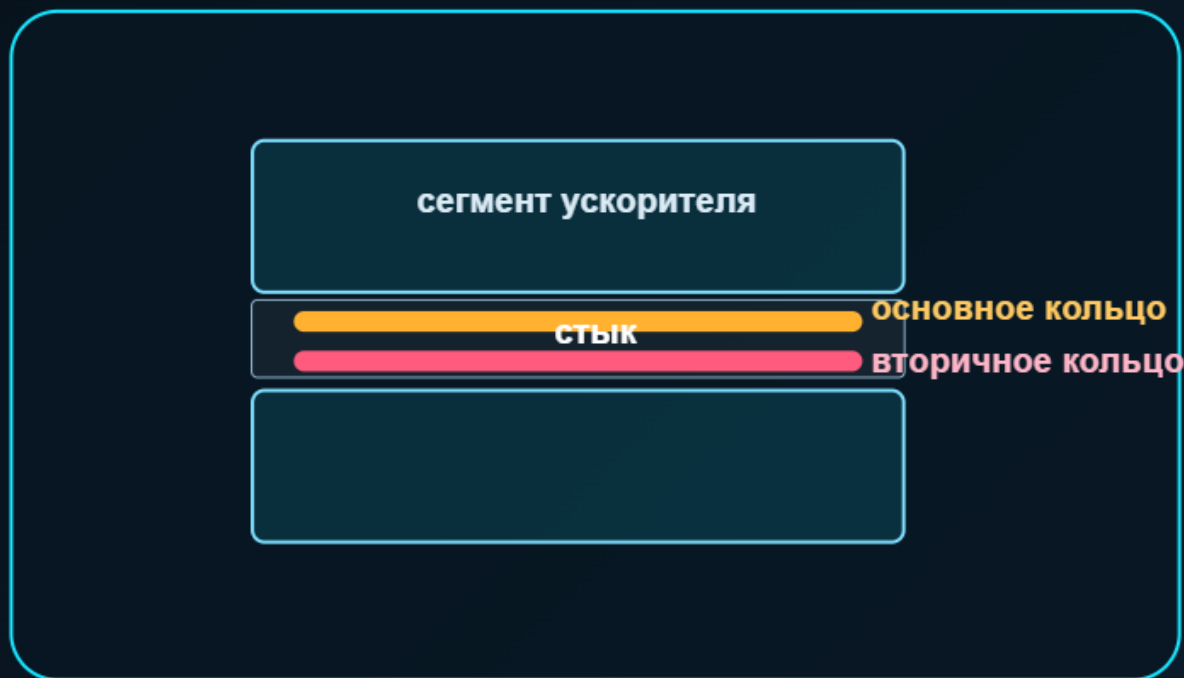
3 Как мы будем разбирать

Разбираем как учебный пример о данных, визуализации, коммуникации и ответственности.



Что такое O-ring

Упрощённая схема: не инженерный чертёж, а учебная картинка



Зачем нужны кольца

Они должны удерживать горячие газы внутри ускорителя. Если уплотнение не сработало, появляется путь для утечки.



Почему важна температура

Холод ухудшает способность резины быстро возвращаться к форме и следовать за раскрытием зазора.

Простая таблица: температура и повреждённые

24 предыдущих полёта: для каждого полёта показаны температура запуска и число повреждённых колец.

Полёт	Темп. °F	Поврежд.
1	53	3
2	57	1
3	58	1
4	63	1
5	66	0
6	67	0
7	67	0
8	67	0
9	68	0
10	69	0
11	70	1
12	70	0

Полёт	Темп. °F	Поврежд.
13	70	1
14	70	0
15	72	0
16	73	0
17	75	0
18	75	2
19	76	0
20	76	0
21	78	0
22	79	0
23	80	0
24	81	0

Что здесь важно?

В таблице оставлены только три столбца: номер полёта, температура запуска и число повреждённых колец.

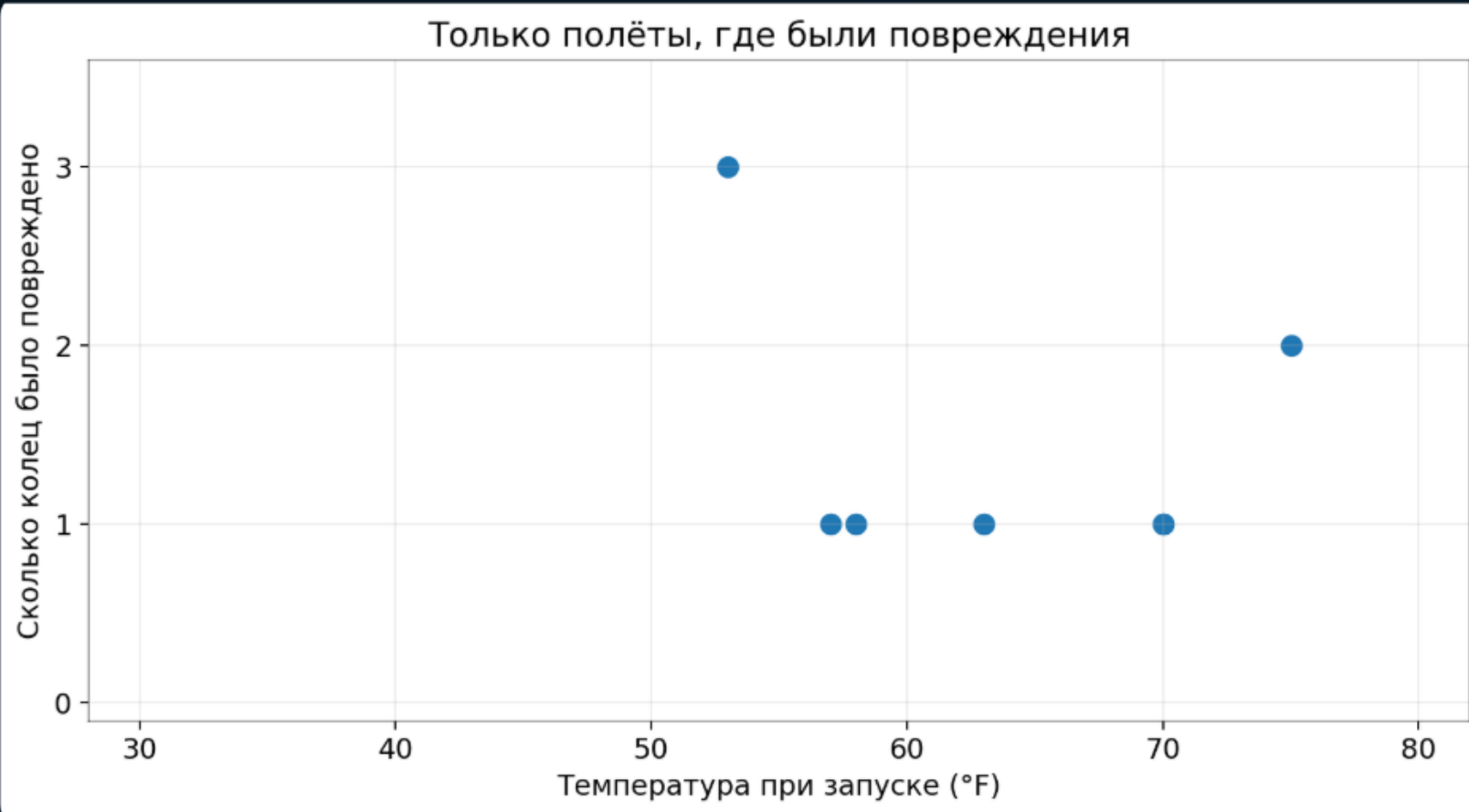
Педагогическая мысль

Даже простая таблица позволяет задать важный вопрос: что происходит, когда температура становится ниже обычного диапазона?

Источник: учебная таблица Challenger O-ring, используемая для демонстрации. В разных источниках способы подсчёта повреждений могут отличаться.

8. Искажённая статистика: видим только поломки

Сначала покажем ошибочную картину: на графике оставлены только полёты, где повреждения уже были. Случаи без повреждений скрыты.



Что видно?

- на графике только 7 полётов с повреждениями;
- нет ни одного полёта с 0 повреждений;
- по такой картинке кажется,

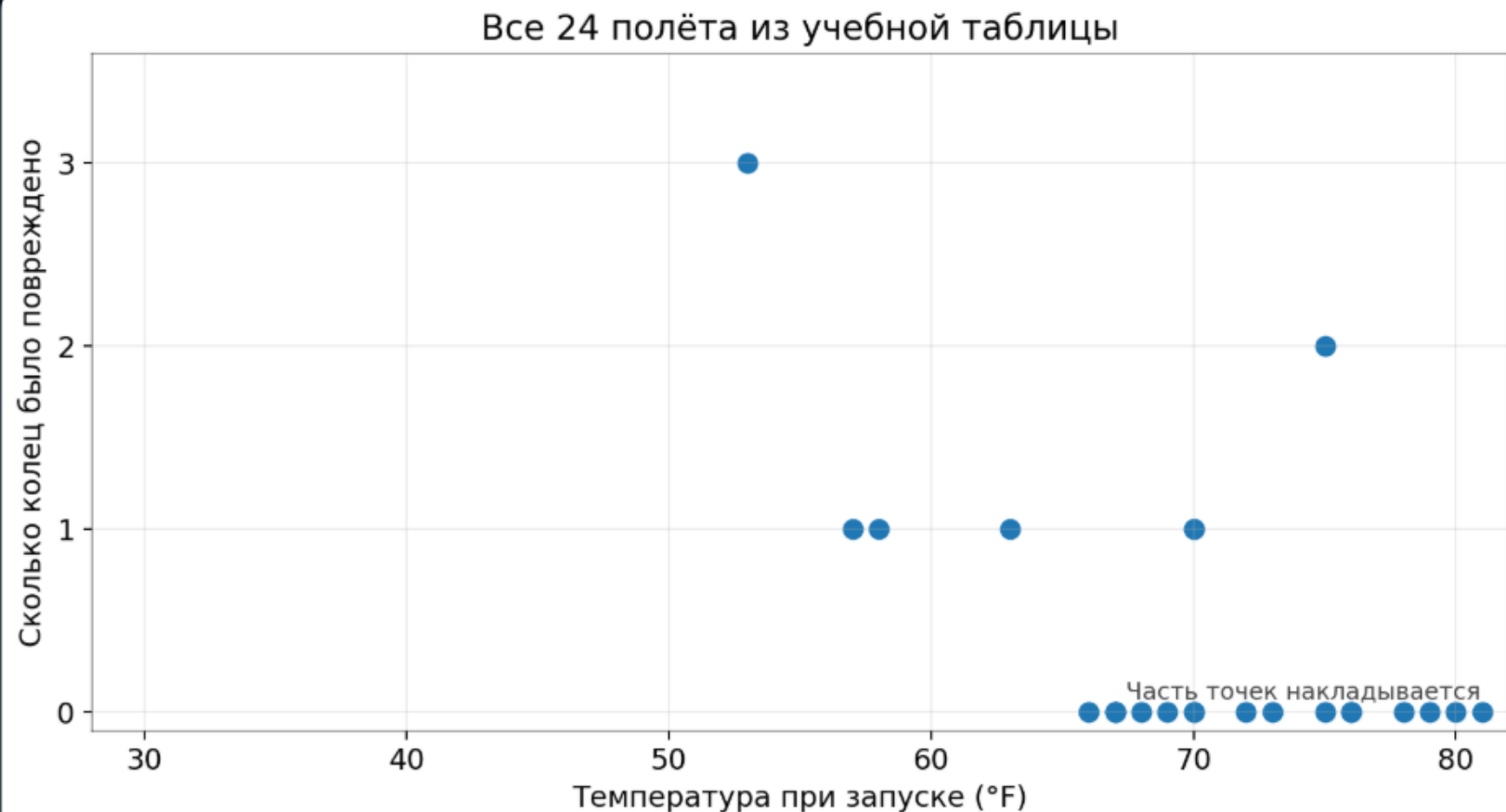
Почему это искажение?

Мы выбросили все нулевые случаи. Когда отрицательные примеры исчезают, статистика становится однобокой и может привести к плохому выводу.

Построено по таблице с предыдущего слайда: оставлены только строки, где число повреждений больше 0.

9. Полная картина: добавляем все 24 полёта

Теперь возвращаем на график все строки из таблицы: и с повреждениями, и без повреждений. Только так видна структура данных.



Что меняется?

- появляются полёты с 0 повреждений;
- видно, что при более тёплых температурах нулей много;
- холодные запуски

Педагогическая мысль

Именно добавление «обычных» случаев делает вывод сильнее. Анализ данных почти всегда требует не только примеров проблемы, но и примеров без проблемы.

Построено по полной таблице из 24 полётов. Часть точек накладывается, потому что некоторые полёты имеют одинаковые значения.

Что пошло не так

Не “данных не было”, а “сигнал риска не стал решением”

Неполная картина

не все данные были одинаково понятны и сопоставимы

Слабая визуализация

опасный сигнал можно спрятать плохой подачей

Проблема коммуникации

инженерные опасения не были достаточно защищены

Давление решения

расписание и ожидания могут давить на осторожность

Плохая подача данных может спрятать опасный сигнал.

Урок Challenger

Если цена ошибки велика, проверка должна быть особенно строгой

СТОП. ПРОВЕРЯЕМ!

**Что именно мы
считаем?**

**Что считается
событием?**

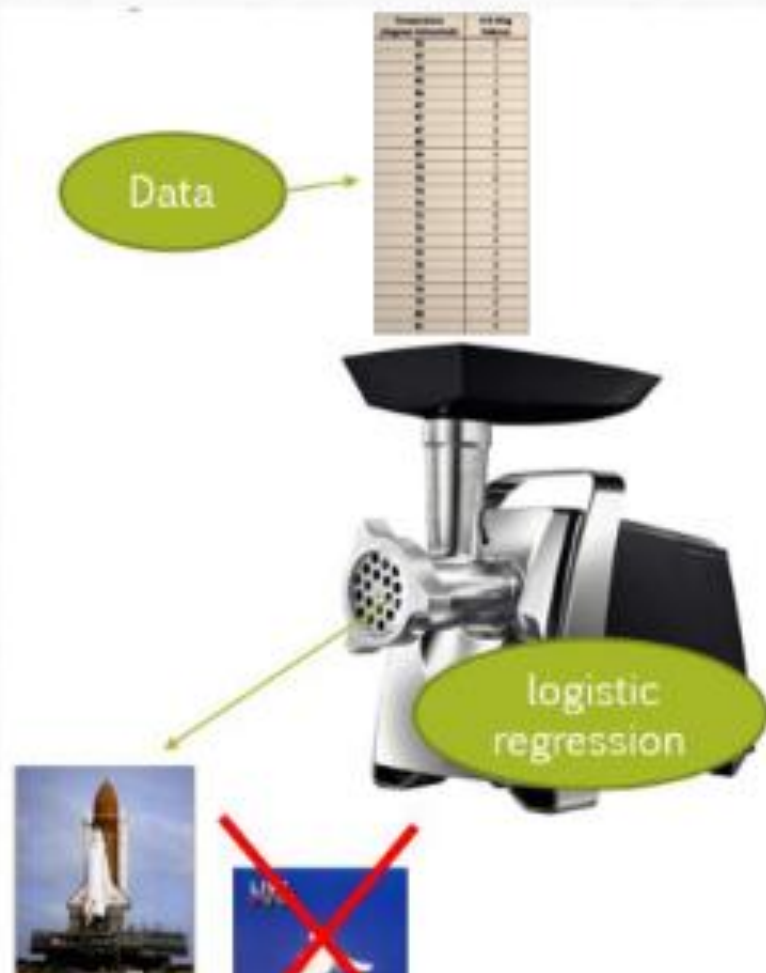
Как показан риск?

**Кто принимает
решение?**

**Что будет, если мы
ошиблись?**

Урок Challenger

Если цена ошибки велика, проверка должна быть особенно строгой



$$P(\text{Catastrophe} \mid X = 36) = 0.56$$

От предупреждения к стратегии

Данные нужны не только чтобы остановить риск, но и чтобы найти новое решение



Challenger

данные должны были усилить осторожность



Moneyball

данные помогли пересмотреть привычный способ выбирать игроков

Обе истории — про качество



Moneyball: данные против дорогой привычки

Oakland Athletics искали способ конкурировать при ограниченном бюджете

1 Контекст

В книге Michael Lewis описал подход Oakland Athletics и Billy Beane к поиску недооценённых игроков.

2 Идея

Смотреть не только на привычное впечатление, а на метрики, которые связаны с вкладом в победы.

3 Осторожность

Это не сказка “таблица победила всех”. Это пример того, как правильный вопрос меняет стратегию.



Старый подход: опыт и интуиция

Опыт эксперта важен, но он может видеть не всё



Скаут смотрит

как игрок выглядит,
двигается, бросает,
держится на поле



Рынок оценивает

кого считают
перспективным,
дорогим, “звёздным”



Риск

некоторые полезные
качества могут быть
недооценены

Шутка: “Выглядит как звезда” — это ещё не столбец в таблице.

Новый подход: статистика бейсбола

Sabermetrics — это попытка оценивать игру через данные и метрики

Цель

побеждать чаще

Метрики

связанные с
вкладом

Рынок

кто недооценён

Решение

кого выбрать



Главная идея

Данные помогают увидеть то, что привычный взгляд пропускает. Но метрику всё равно выбирает человек.

Урок Moneyball

Данные меняют стратегию, когда мы правильно выбираем вопрос

Эксперты могут ошибаться

особенно если рынок привык смотреть на одни и те же признаки

Данные дают новый угол

но не отменяют необходимость понимать игру

Метрика должна вести к цели

иначе модель будет хорошо решать не ту задачу

Вопрос не “верить людям или данным”, а “какие данные помогают

Что мы выучили

Второй фрагмент секретной карты найден



Challenger

Данные без правильной коммуникации могут не остановить плохое решение. Если цена ошибки высока, проверка должна быть особенно строгой.



Moneyball

Данные могут изменить стратегию, если метрика связана с целью и помогает увидеть недооценённые возможности.

**Данные не заменяют ответственность.
Они усиливают того, кто умеет задавать правильные вопросы.**

КУРС ПО АНАЛИЗУ ДАННЫХ

Встреча 1 · следующий пример

СИМУЛЯЦИЯ: данные из кубиков

Не “из воздуха”, а по правилам + случайность.

СИМУЛЯЦИЯ ОТДЕЛЕНИЯ НЕОТЛОЖНОЙ ПОМОЩИ
И МЕТОД МОНТЕ-КАРЛО



Высшие силы сегодня —

Однозвездочная бо

Вылечить не мене
пациентов с диагно
«Злокачественный
просветит»

0/3

Кабинеты

Предметы

Норм

\$ 161 513

Репутация

ОК
06

Год



Симуляция и моделирование -
просто и доступно



Данные “из воздуха”? Почти. Но честно.

Симуляция создаёт искусственные наблюдения не магией, а по заданным правилам.



Правила

кто приходит,
сколько ждёт,
сколько коек



Кубики

случайность: один
день спокойный,
другой перегружен



Виртуальные

много искусственных
“дней” для анализа

Важная мысль: симуляция не доказывает сама себя. Её нужно

История: отделение неотложной помощи

MIT thesis: модель работы Emergency Department в Beth Israel Deaconess Medical Center

Department at Beth Israel Deaconess Medical
Center via Simulation

by

Clay W. Noyes

B.S. Operations Research

United States Air Force Academy, 2006

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

© 2008 Clay W. Noyes. All rights reserved.

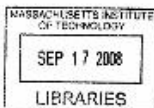
The author hereby grants to MIT permission to reproduce and distribute publicly paper
and electronic copies of this thesis document in whole or in part in any medium now
known or hereafter created.

Author Sloan School of Management
Interdepartmental Program in Operations Research
May 15, 2008

Certified by Francis Carr
Charles Stark Draper Laboratory
Technical Supervisor

Certified by Dimitris Bertsimas
Boeing Professor of Operations Research
Thesis Supervisor

Accepted by



1 год данных

анонимизированные данные Dashboard



47 537 визитов

все посещения ED за период



5 процессов

прибытие, триаж, ожидание, лечение, выход



Monte Carlo

много случайных прогонов модели

Почему это важно для нашего курса

Симуляция — это мост между “у нас нет готовой таблицы” и “мы можем исследовать идею”.



Создаём мир

задаём правила: кто приходит, как долго обслуживается, где возникает очередь



Кидаем кубики

генератор случайных чисел создаёт много



Анализируем данные

строим графики, считаем метрики,

Ближе к концу курса

мы сами создадим учебные данные с помощью симуляции, загрузим их в Orange и проверим, какие решения действительно улучшают систему.

Данные не падают с неба.

..

Встреча 1 · Часть 5

МНОГОЛИКИЕ ДАННЫЕ

Таблицы, тексты, картинки, сети, события и
следы цифровой жизни

**Первый вопрос аналитика:
«Покажи данные».**



Что изучим в этой части

Зачем сначала смотреть на данные и почему они бывают такими разными

1 главный вопрос проекта

Покажи данные, а не только идею

2 качество данных

плохие данные → плохой вывод

3 формы данных

таблицы, тексты, графы, гео, медиа

4 типы внутри

числа, категории, даты, пропуски

5 мини-проверка

определяем тип данных и риск

ПЛОХИЕ ДАННЫЕ

→ плохое исследование

Что может пойти не так?

- пропуски и дубли
- ошибки ввода
- разные форматы
- не те объекты
- смещённая выборка

СТОП. ПРОВЕРЯЕМ!

Сначала качество данных, потом красивые модели.

Табличные данные

Строки — объекты, столбцы — признаки

студент	возраст	город	балл	сдал?
Аня	21	Хайфа	87	да
Дима	24	Тель-Авив	64	да
Лена	19	Иерусалим	48	нет
Рами	22	Хайфа	91	да

Почему таблицы

- легко фильтровать
- легко строить графики
- легко обучать модели
- хорошо подходят для Orange

Но таблица — это не “просто Excel”: внутри столбцов живут разные типы

Табличные данные

Строки — объекты, столбцы — признаки

- Табличные данные — ABT — analytical base table

The diagram illustrates an Analytical Base Table (ABT) with 10 rows and 5 columns. A bracket on the left labeled 'Objects' spans the rows, and a bracket at the top labeled 'Attributes' spans the columns. The table contains the following data:

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Analytical base table - Wikipedia

In database theory, the **Analytical Base Table** (ABT) is a flat table that is used for building analytical models and scoring (predicting) the future behavior of a subject. A single record in this table represents the subject of the prediction (e.g. a customer) and stores all data (variables) describing this subject.

Но таблица — это не “просто Excel”: внутри столбцов живут разные типы

Табличные данные

Строки — объекты, столбцы — признаки

Data Table

Info

264 instances

92 features (no missing values)

No target variable.

2 meta attributes (34.5% missing values)

Variables

☒ Show variable labels (if present)

☐ Visualize numeric values

☒ Color by instance classes

Selection

☒ Select full rows

Restore Original Order

	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20
128	Holy See	41.9029	12.4534	0	0	0	0
129	Honduras	15.2	-86.2419	0	0	0	0
130	Hungary	47.1625	19.5033	0	0	0	0
131	Iceland	64.9631	-19.0208	0	0	0	0
132	India	21	78	0	0	0	0
133	Indonesia	-0.7893	113.921	0	0	0	0
134	Iran	32	53	0	0	0	0
135	Iraq	33	44	0	0	0	0
136	Ireland	53.1424	-7.6921	0	0	0	0
137	Israel	31	35	0	0	0	0
138	Italy	43	12	0	0	0	0
139	Jamaica	18.1096	-77.2975	0	0	0	0
140	Japan	36	138	2	2	2	2
141	Jordan	31.24	36.51	0	0	0	0
142	Kazakhstan	48.0196	66.9237	0	0	0	0
143	Kenya	-0.0236	37.9062	0	0	0	0
144	Korea, South	36	128	1	1	2	2
145	Kuwait	29.5	47.75	0	0	0	0

Текстовые данные

Отзывы, письма, новости, комментарии и медицинские записи

Пример текста

«Доставка была быстрой, но упаковка повреждена.
Служба поддержки ответила вежливо.»

Что можно извлечь?

тональность · темы · ключевые слова · жалобы · намерения

Главная идея

Компьютер не “понимает роман” как человек.
Он превращает слова в признаки, а потом ищет закономерности.

Шутка: модель видит “вкусный борщ” и радостно ставит плюс.

Текстовые данные

Отзывы, письма, новости, комментарии и медицинские записи

Пример текста

«Доставка была быстрой, но упаковка повреждена.
Служба поддержки ответила вежливо.»

Что можно извлечь?

тональность · темы · ключевые слова · жалобы · намерения

Главная идея

Компьютер не “понимает роман” как человек.
Он превращает слова в признаки, а потом ищет закономерности.

Шутка: модель видит “вкусный борщ” и радостно ставит плюс.

События и транзакции

Данные как следы действий во времени

Лог — это не просто таблица. Важен порядок событий: что было

Примеры: покупки, клики, банковские операции, входы в приложение, вызовы сервера.



09:01

зашёл



09:03

поиск



09:07

клик



09:12

корзина

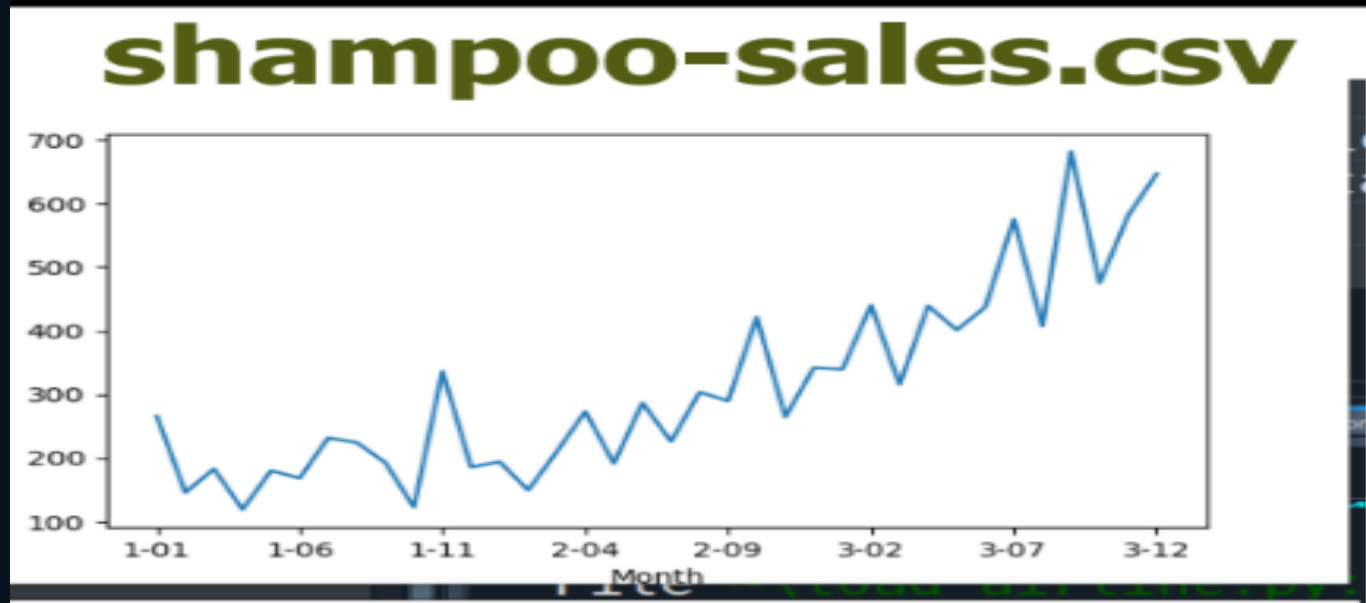


09:20

покупка

Временные ряды

Когда одно и то же измеряется снова и снова



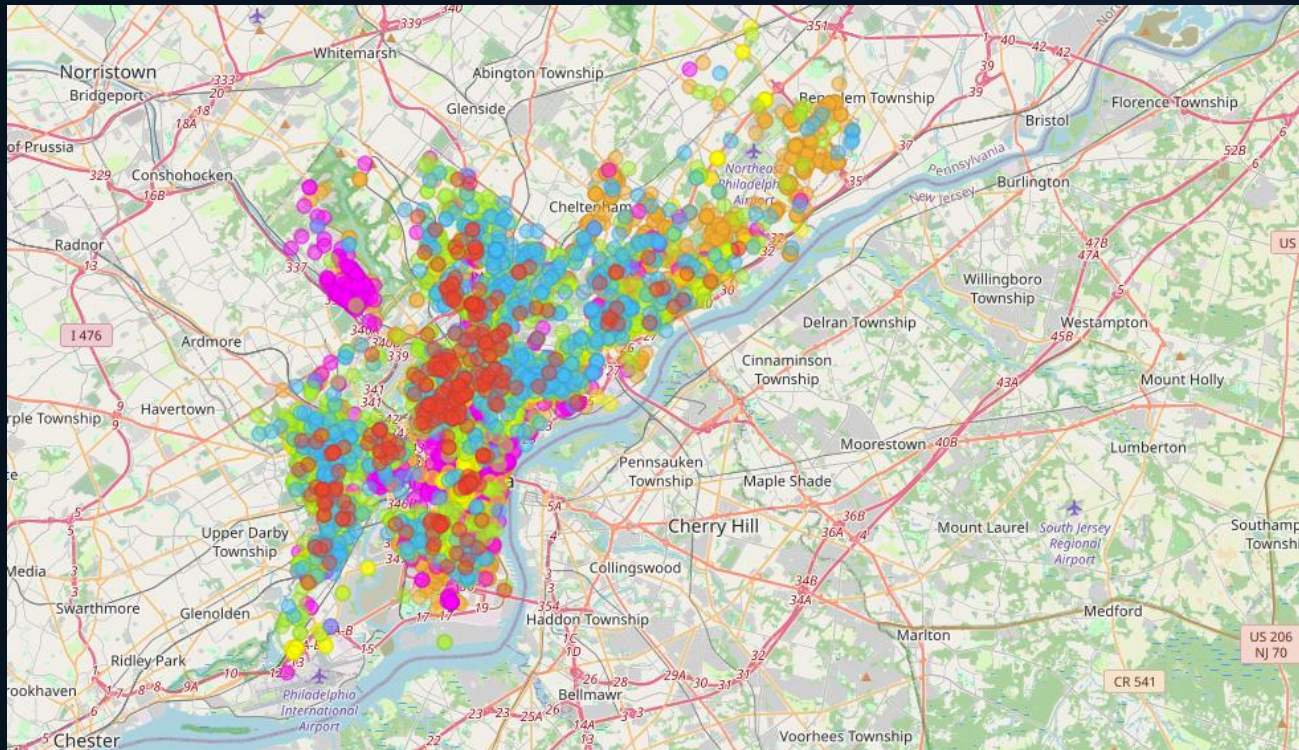
Что тут особенного?

- важен порядок
- есть тренды и сезонность
- старые значения помогают предсказывать новые
- данные могут идти потоком

Датчики превращают мир в поток чисел.

Географические данные

Координаты, маршруты, районы, расстояния



Что можно анализировать?

- где больше спрос
- какие маршруты длиннее
- где находятся кластеры
- как расстояние влияет на результат

Гео данные часто соединяются с таблицами: район + цена + расстояние + время.

Медиа-данные

Картинка — это тоже таблица чисел, только очень большая



Изображения

пиксели, объекты,
медицинские снимки

модель ищет



Звук

волна, речь, музыка,
шумы

модель ищет



Видео

последовательность
кадров + звук

модель ищет

Важная мысль: “неструктурированные” данные всё равно превращаются в признаки.

Последовательности

Иногда данные похожи на фразу, геном или маршрут действий

ДНК

A C G T G C A A T G

текст

СЛОВО → СЛОВО → СЛОВО

поведение

поиск → клик → корзина → покупка

Если перемешать элементы, смысл может исчезнуть.

Один человек оставляет много типов данных

Профиль, отзывы, покупки, клики, геолокация, фото, звук, время

Пример

Покупатель в интернет-магазине:

- возраст и город — таблица
- отзыв — текст
- покупки — транзакции
- маршрут кликов — последовательность
- адрес магазина — геоданные

Вывод: реальная задача редко живёт в одном “чистом” типе данных.

Что мы выучили

И маленькая проверка перед следующим блоком

Summary

1. Первый вопрос проекта: покажите данные.
2. Плохие данные портят исследование.
3. Данные бывают в разных формах.
4. В таблицах тоже есть разные типы.
5. Один объект может создавать много видов данных.

Проверочный вопрос

К вам пришёл друг и говорит:
«У меня есть данные из приложения:
клики, время, город, отзывы и
фотографии».

Какие типы данных здесь есть?

Шутка: «Данные есть?» — «Есть!» — «А они
настоящие?» — «Почти...»

Ответ: события/логи, время, геоданные, тексты, изображения и
таблицы.

Прошлое

Дайте мне точку опоры —
и я переверну Землю



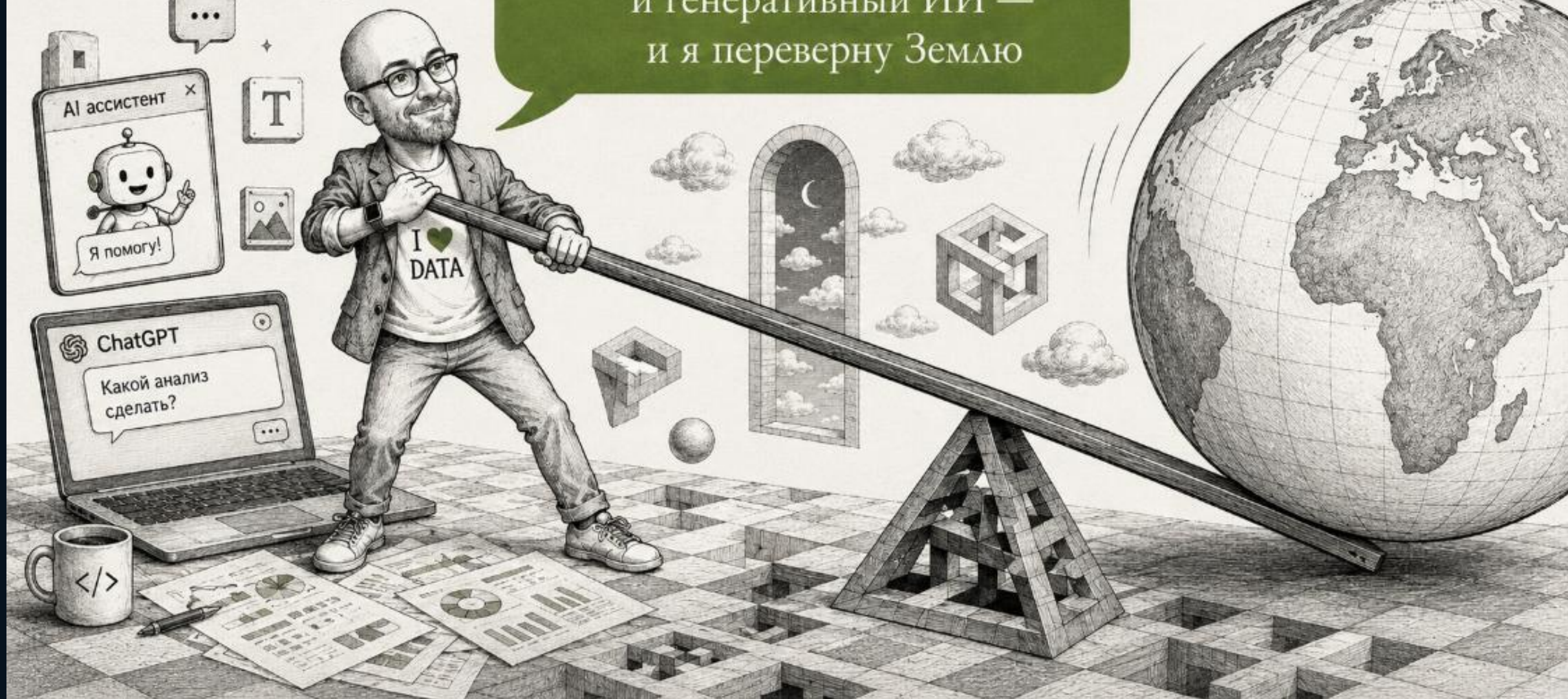
5 лет назад

Дайте мне данные, компьютер
и аналитику данных —
и я переверну Землю



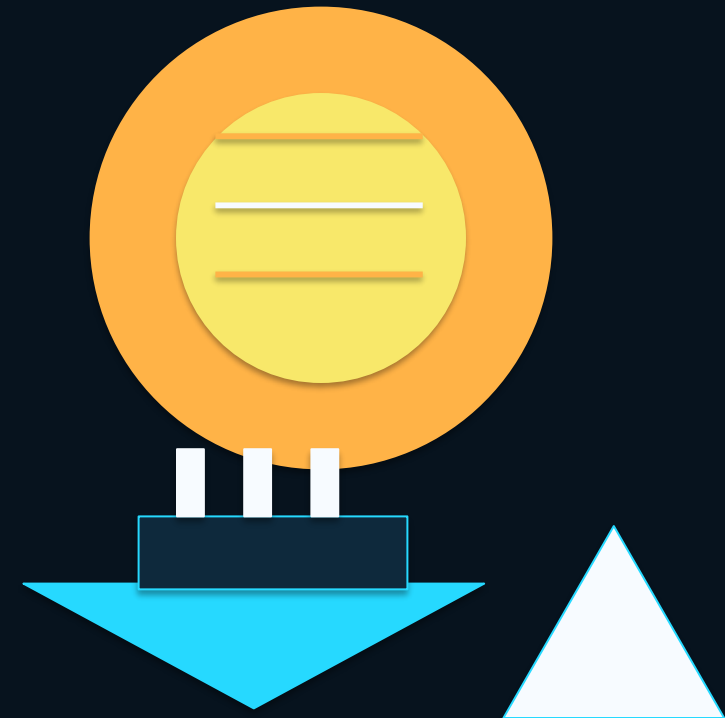
Сегодня

Дайте мне данные, компьютер,
аналитику данных, ChatGPT
и генеративный ИИ —
и я переверну Землю



ORANGE // ПЕРВЫЕ ШАГИ TITANIC

Orange ready to sail



План блока

1. Orange

2. Titanic

3. Первые шаги

4. Выводы

Поплыли!

Что будем изучать

Интерфейс

Загрузка

Таблица

Признаки

Графики

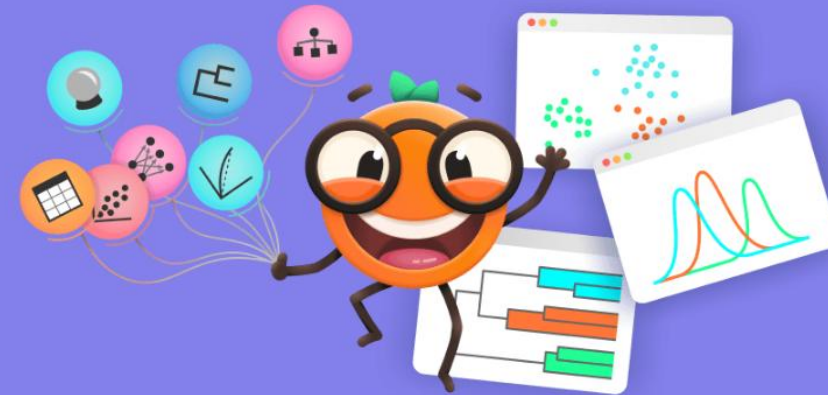
Orange бесплатная программа для визуальной работы с данными

<https://orangedatamining.com/>

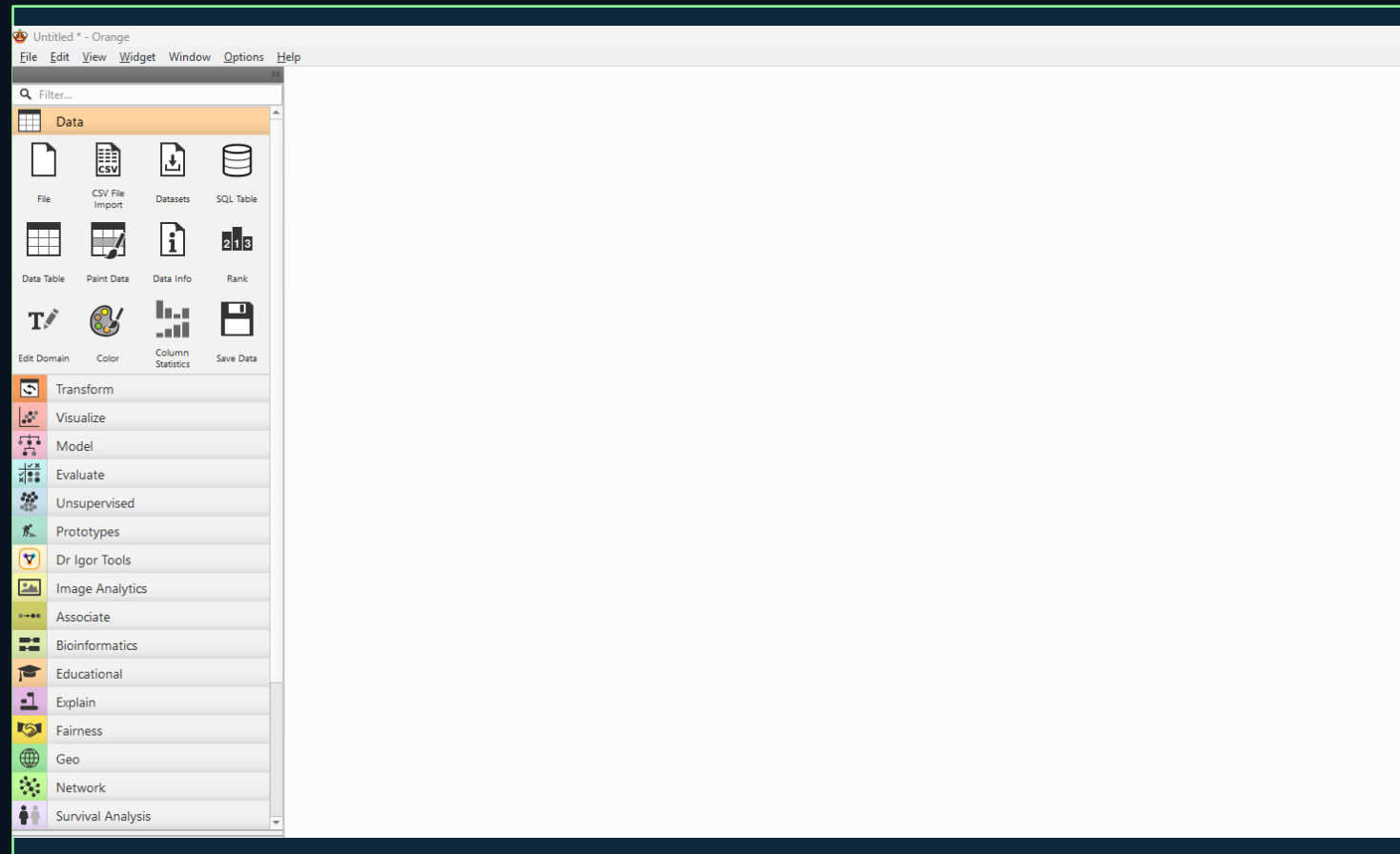
Data Mining Fruitful and Fun

Open source machine learning and data
visualization.

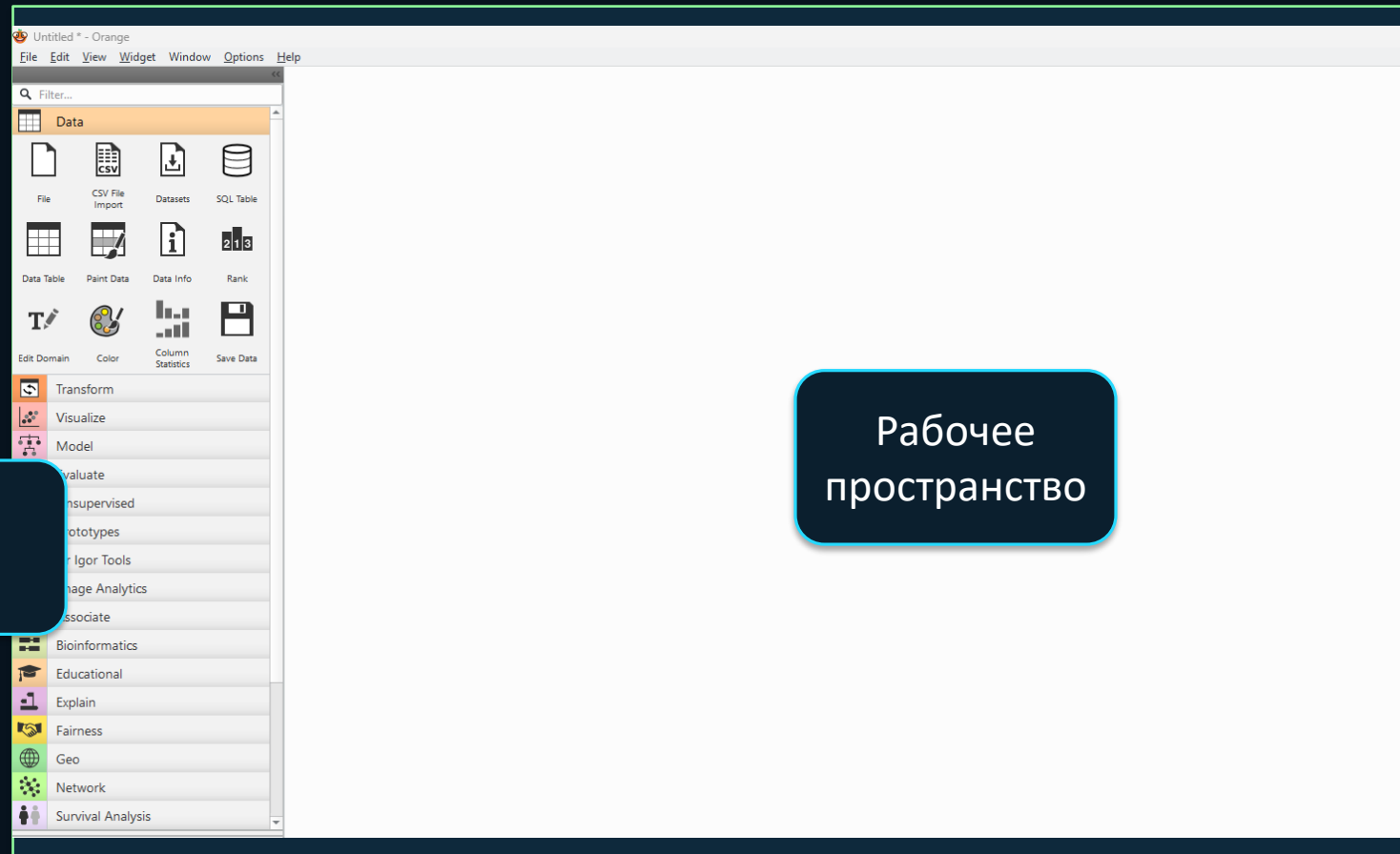
Download Orange 3.40.0



Рабочее пространство



Рабочее пространство



Модули
widgets

Рабочее
пространство

Рабочее пространство

The screenshot displays the Orange data mining software interface. On the left, the 'Data' widget category is selected, showing various widgets like File, CSV File Import, Datasets, SQL Table, Data Table, Paint Data, Data Info, Rank, Edit Domain, Color, Column Statistics, and Save Data. On the right, a workflow diagram is shown with a 'File' widget connected to both a 'Data Table' widget and a 'Scatter Plot' widget. A dark blue box with the text 'Тут творим' (Here we create) is overlaid on the workflow area.

Тут творим

Первый Пример

Titanic: первый понятный вопрос

Учебные данные хороши тем, что история понятна: есть пассажиры, признаки и исход.

Вопрос исследования

Какие признаки связаны с шансом выжить?

sex	пол
age	ребёнок или взрослый
class / status	класс или роль
survived	цель: выжил или нет



Спойлер: Orange поможет не “угадать”, а увидеть связи в данных.

Но: связь в данных ещё не равна причине.

Первый Пример

Первый путь в Orange: открыть и посмотреть

Спойлер-практика: мы не доказываем всё сразу. Мы учимся задавать первый вопрос к данным.

1 Открыть Datasets

найти учебный набор Titanic

2 Подключить Data Table

посмотреть строки и столбцы

3 Подключить визуализацию

увидеть, что связано с выживанием

Важно

В этом блоке мы показываем принцип. Настоящая практика будет позже: там студенты сами соберут workflow, нажмут на виджеты и сделают первый вывод.

Первый Пример

The screenshot displays the Orange data mining software interface. On the left is a sidebar with a 'Filter...' search bar and a list of widget categories: Data, Transform, Visualize, Model, Evaluate, Unsupervised, Prototypes, Dr Igor Tools, Image Analytics, Associate, Bioinformatics, Educational, Explain, and Fairness. The 'Data' category is selected, showing icons for File, CSV File Import, Datasets, SQL Table, Data Table, Paint Data, Data Info, Rank, Domain, Color, Column Statistics, and Save Data. A 'File' widget icon is highlighted in the center workspace. On the right, the 'File - Orange' configuration window is open. It has a menu bar (File, View, Window, Help) and a 'Source' section with 'File' selected and the path 'datasets/titanic.tab'. The 'File Type' section has a button 'Determine type from the file extension'. The 'Info' section provides details about the 'Titanic dataset': 'Data from the real Titanic passengers and their survival.', '2201 instances', '3 features (no missing values)', 'Classification; categorical class with 2 values (no missing values)', and '0 meta attributes'. The 'Columns (Double click to edit)' section contains a table:

	Name	Type	Role	Values
1	status	categorical	feature	crew, first, second, third
2	age	categorical	feature	adult, child
3	sex	categorical	feature	female, male
4	survived	categorical	target	no, yes

At the bottom of the window are a 'Reset' button, a 'Browse documentation datasets' button, and a status bar showing icons for help, file, and a count of 2201 instances.

Первый Пример

The screenshot displays the Orange data mining software interface. On the left, a widget palette is visible with various data-related widgets. The main workspace shows a workflow with a 'File' widget connected to a 'Data Table' widget. A 'Data Table - Orange' window is open, showing the loaded data.

Widget Palette (Data category):

- File
- CSV File Import
- Datasets
- SQL Table
- Data Table
- Paint Data
- Data Info
- Rank
- Edit Domain
- Color
- Column Statistics
- Save Data

Workflow:

```
graph LR; File[File] -- Data --> DataTable[Data Table]
```

Data Table - Orange Window:

Info: 2201 instances (no missing data), 3 features, Target with 2 values, No meta attributes.

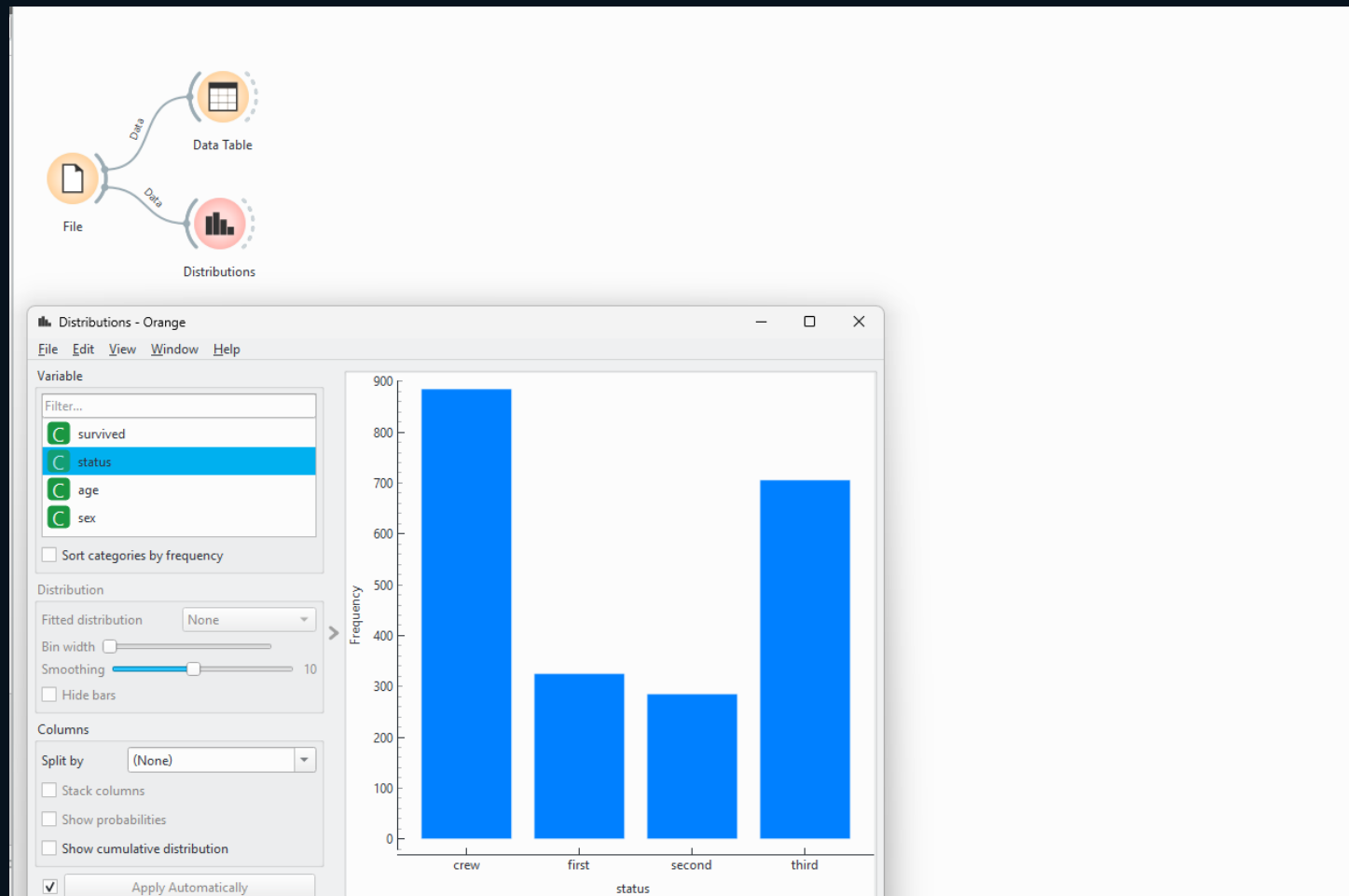
Variables:

- ☒ Show variable labels (if present)
- ☐ Visualize numeric values
- ☒ Color by instance classes

Selection:

	survived	status	age	sex
1	yes	first	adult	male
2	yes	first	adult	male
3	yes	first	adult	male
4	yes	first	adult	male
5	yes	first	adult	male
6	yes	first	adult	male
7	yes	first	adult	male
8	yes	first	adult	male
9	yes	first	adult	male
10	yes	first	adult	male
11	yes	first	adult	male
12	yes	first	adult	male

Первый Пример



Первый Пример

- Титаник на Кагл <https://www.kaggle.com/c/titanic>



The Challenge

The sinking of the Titanic is one of the most infamous shipwrecks in history.

Titanic: первый понятный вопрос

Учебные данные хороши тем, что история понятна: есть пассажиры, признаки и исход.

Вопрос исследования

Какие признаки связаны с шансом выжить?

sex	пол
age	ребёнок или взрослый
class / status	класс или роль
survived	цель: выжил или нет



Спойлер: Orange поможет не “угадать”, а увидеть связи в данных.

Но: связь в данных ещё не равна причине.

СТОП. ПРОВЕРЯЕМ!

Первое правило практики: не превращать красивую картинку в слишком сильный вывод.

Можно сказать

В этих данных пол, возраст или класс могут быть связаны с выживанием.

Нужно проверить

Сколько данных? Нет ли ошибок? Как именно закодированы признаки?

Нельзя сразу сказать

“Мы доказали причину” или “модель всё поняла”.

Мини-диалог

Друг: “Сделай мне проект по анализу данных!”

Аналитик: “Отлично. Покажи данные.”

Друг: “А если данных нет?”

Аналитик: “Тогда у нас пока не проект, а мечта.”

Это шутка, но почти профессиональный стандарт.

Что запомнить

Этот спойлер нужен, чтобы студенты поняли язык курса до первой полноценной практики.

Данные

это объекты, признаки и часто целевая переменная

Orange

это визуальный конструктор анализа

Titanic

это первый учебный пример с понятным вопросом

Проверка

связь $\neq$ причинность, картинка $\neq$ доказательство

Проверочный вопрос

Что нужно сделать первым, когда вам предлагают “проект по анализу данных”?

Ответ: попросить показать данные и сформулировать вопрос исследования.

Встреча 1 · спойлер-практика

ORANGE: HEART DISEASE

Первое медицинское расследование данных

Сегодня мы не ставим диагноз. Мы учимся смотреть на данные, признаки и первый исследовательский вопрос.

Мини-история: что данные о пациентах могут рассказать о риске



Что изучим

Этот блок — короткий трейлер практики: как в Orange выглядит медицинская таблица и как из неё рождается вопрос.

Orange

визуальный конструктор анализа

Данные

пациенты, признаки, значения

Heart Disease

пример медицинского набора данных

Проверка

связь $\neq$ диагноз и не причина

Главная мысль: в медицине особенно важно сначала понять данные, а уже потом строить модели.

Orange на экране: рабочее поле анализа

Мы собираем исследование как цепочку блоков: источник данных, таблица, визуализация, модель и проверка.

Главная идея

Orange похож на конструктор: каждый блок выполняет один понятный шаг.

Данные переходят от одного блока к другому по линиям.

Workflow читается как история исследования: от данных к выводу.



ORANGE CANVAS

Datasets → Data Table → Visualize

Рабочее поле Orange

Логика Orange: виджеты и связи

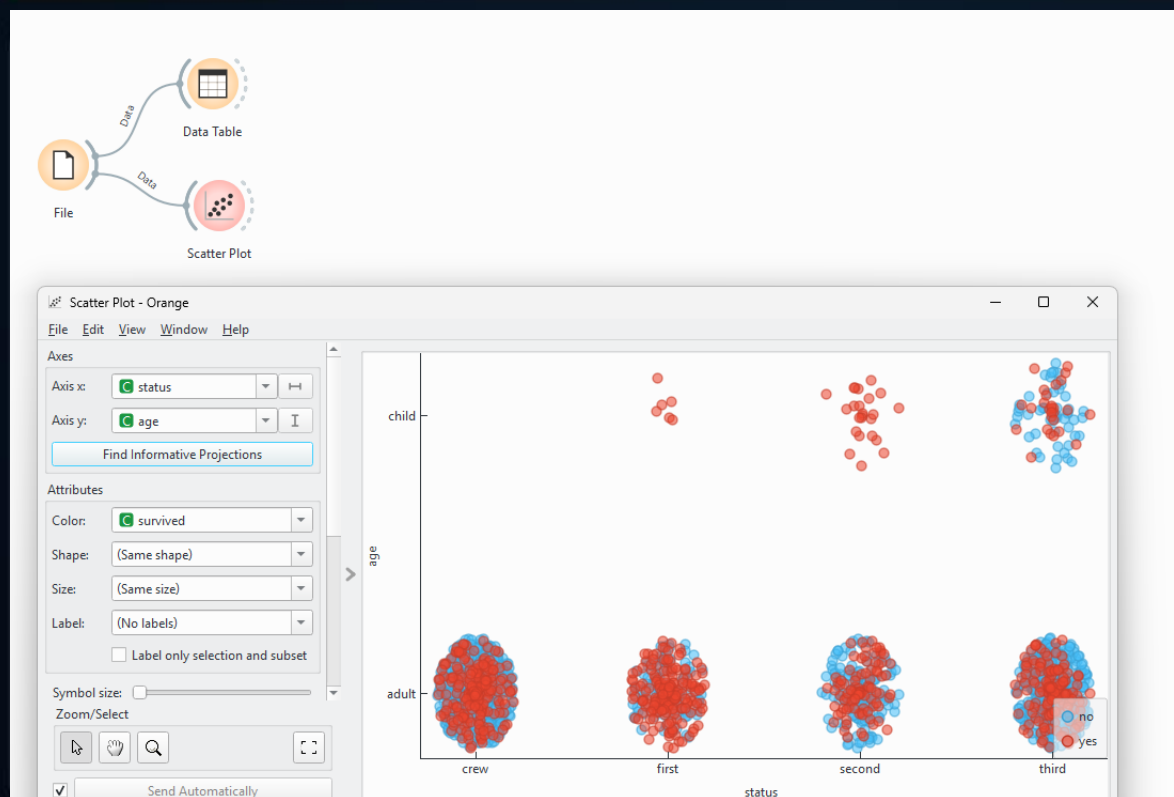
Виджет — это маленькая операция над данными. Связь показывает, куда данные идут дальше.

Три слова

Виджет — открыть, посмотреть, нарисовать или обучить.

Связь — передать данные дальше.

Workflow — весь путь анализа от вопроса к выводу.



Шаг 1: выбираем набор данных

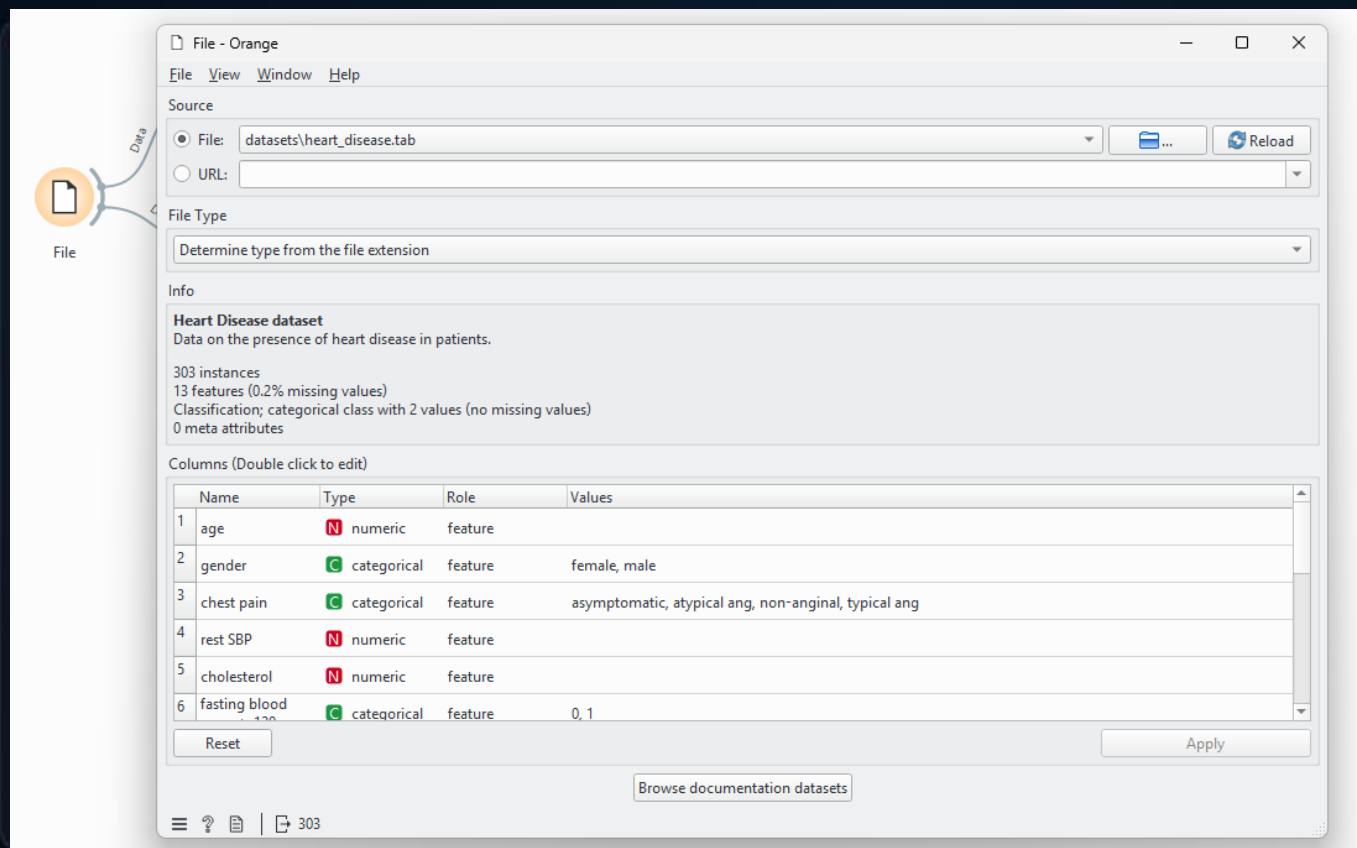
Начинаем с источника данных. В этом примере объектами будут пациенты, а столбцами — их измерения и признаки.

Что происходит

Открываем набор Heart Disease.

Переходим от названия набора к реальной таблице.

Не делаем медицинских выводов до просмотра данных.



Шаг 2: смотрим таблицу

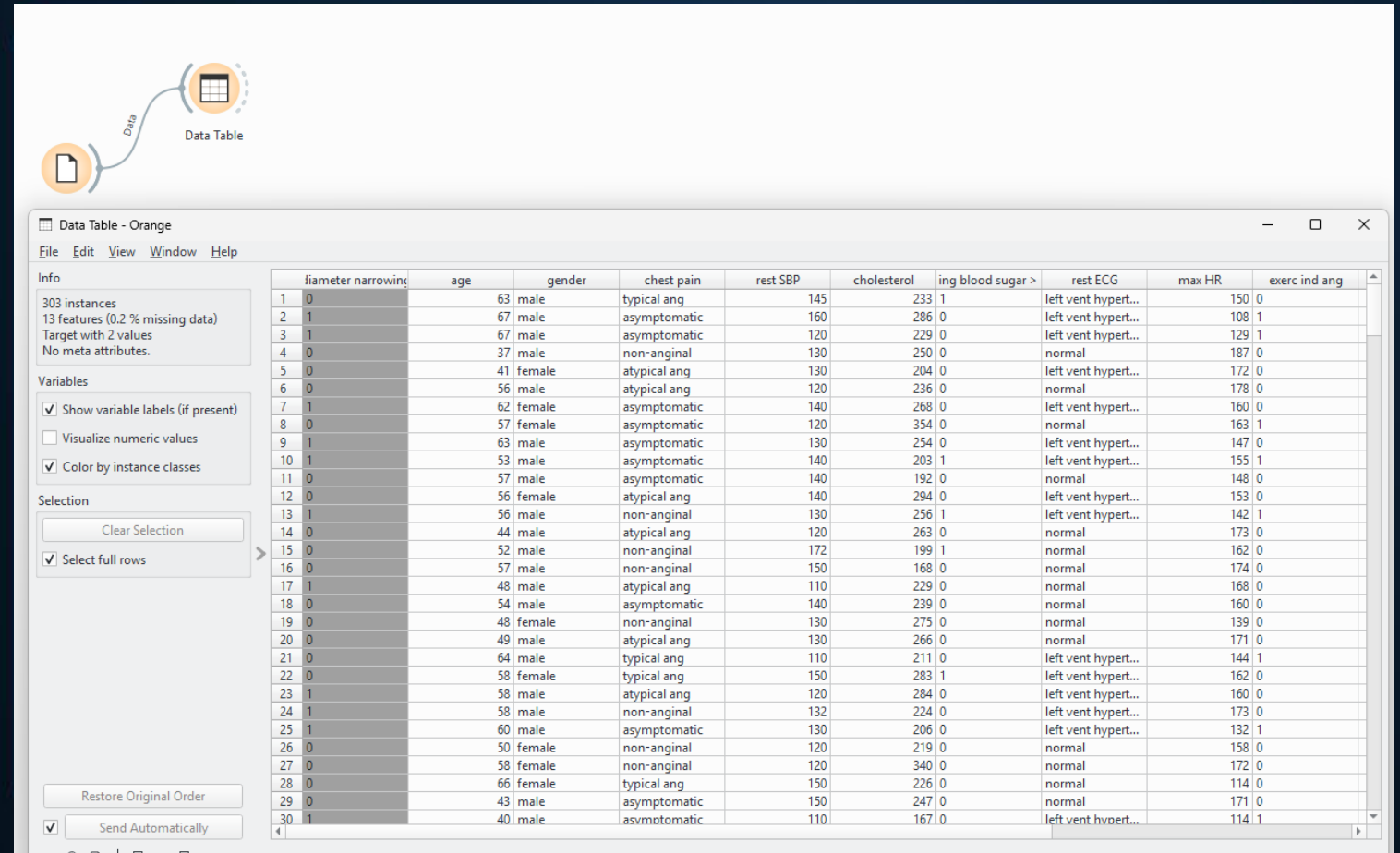
Data Table показывает реальные строки, столбцы и значения. Это обязательная остановка перед любым выводом.

Акцент

Одна строка — один пациент.

Столбец — признак, измерение или результат.

Непонятная таблица превращает красивую модель в гадание.



The screenshot shows the Orange Data Table widget. The sidebar on the left contains the following sections:

- Info:** 303 instances, 13 features (0.2 % missing data), Target with 2 values, No meta attributes.
- Variables:** ☒ Show variable labels (if present), ☐ Visualize numeric values, ☒ Color by instance classes.
- Selection:** Clear Selection button, ☒ Select full rows.
- Buttons:** Restore Original Order, ☒ Send Automatically.

The main table displays the following data (columns are truncated in the original image):

	liameter narrowing	age	gender	chest pain	rest SBP	cholesterol	ing blood sugar >	rest ECG	max HR	exerc ind ang
1	0	63	male	typical ang	145	233	1	left vent hypert...	150	0
2	1	67	male	asymptomatic	160	286	0	left vent hypert...	108	1
3	1	67	male	asymptomatic	120	229	0	left vent hypert...	129	1
4	0	37	male	non-anginal	130	250	0	normal	187	0
5	0	41	female	atypical ang	130	204	0	left vent hypert...	172	0
6	0	56	male	atypical ang	120	236	0	normal	178	0
7	1	62	female	asymptomatic	140	268	0	left vent hypert...	160	0
8	0	57	female	asymptomatic	120	354	0	normal	163	1
9	1	63	male	asymptomatic	130	254	0	left vent hypert...	147	0
10	1	53	male	asymptomatic	140	203	1	left vent hypert...	155	1
11	0	57	male	asymptomatic	140	192	0	normal	148	0
12	0	56	female	atypical ang	140	294	0	left vent hypert...	153	0
13	1	56	male	non-anginal	130	256	1	left vent hypert...	142	1
14	0	44	male	atypical ang	120	263	0	normal	173	0
15	0	52	male	non-anginal	172	199	1	normal	162	0
16	0	57	male	non-anginal	150	168	0	normal	174	0
17	1	48	male	atypical ang	110	229	0	normal	168	0
18	0	54	male	asymptomatic	140	239	0	normal	160	0
19	0	48	female	non-anginal	130	275	0	normal	139	0
20	0	49	male	atypical ang	130	266	0	normal	171	0
21	0	64	male	typical ang	110	211	0	left vent hypert...	144	1
22	0	58	female	typical ang	150	283	1	left vent hypert...	162	0
23	1	58	male	atypical ang	120	284	0	left vent hypert...	160	0
24	1	58	male	non-anginal	132	224	0	left vent hypert...	173	0
25	1	60	male	asymptomatic	130	206	0	left vent hypert...	132	1
26	0	50	female	non-anginal	120	219	0	normal	158	0
27	0	58	female	non-anginal	120	340	0	normal	172	0
28	0	66	female	typical ang	150	226	0	normal	114	0
29	0	43	male	asymptomatic	150	247	0	normal	171	0
30	1	40	male	asymptomatic	110	167	0	left vent hypert...	114	1

Что мы видим в таблице?

В медицинском примере важно сразу разделить объект, признаки и целевую переменную.

Heart Disease как данные

Объект	пациент
Признаки	возраст, пол, давление, холестерин, пульс, боль в груди
Цель	наличие или риск заболевания сердца
Вопрос	какие признаки связаны с результатом?

Heart Disease dataset

Data on the presence of heart disease in patients.

303 instances

13 features (0.2% missing values)

Classification; categorical class with 2 values (no missing values)

0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	age	N numeric	feature	
2	gender	C categorical	feature	female, male
3	chest pain	C categorical	feature	asymptomatic, atypical ang, non-anginal, typical ang
4	rest SBP	N numeric	feature	
5	cholesterol	N numeric	feature	
6	fasting blood sugar > 120	C categorical	feature	0, 1
7	rest ECG	C categorical	feature	normal, left vent hypertrophy, ST-T abnormal
8	max HR	N numeric	feature	
9	exerc ind ang	C categorical	feature	0, 1
10	ST by exercise	N numeric	feature	
11	slope peak exc ST	C categorical	feature	upsloping, flat, downsloping
12	major vessels colored	N numeric	feature	
13				

Цель — это ответ, который мы хотим объяснить или предсказать. Признаки — подсказки для анализа.

Шаг 3: собираем маленькую схему

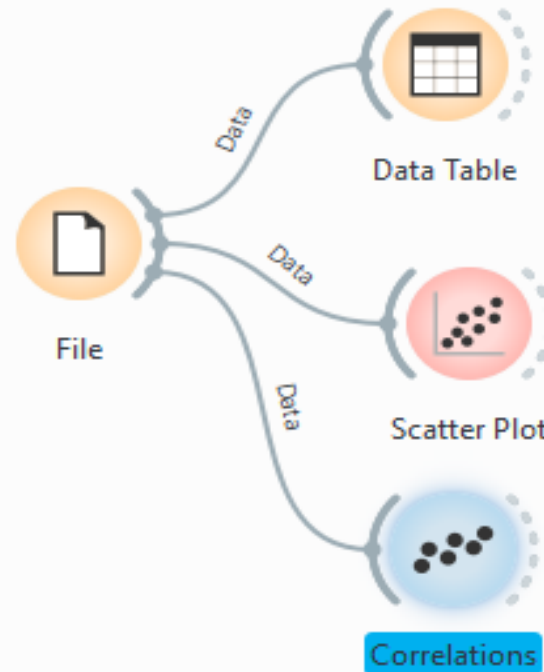
Весь анализ виден на одном экране: источник данных, таблица, визуализация и первый исследовательский шаг.

Почему это удобно

Вся схема анализа видна на экране.

Каждый виджет отвечает на один вопрос.

Линии показывают путь данных от шага к шагу.



Что мы выучили

Теперь видно, как медицинская задача превращается в данные: пациенты, признаки, цель, графики и осторожные выводы.

1

Orange

собирает анализ из виджетов

2

Данные

имеют объекты, признаки и цель

3

Heart Disease

даёт медицинский пример без
диагноза

4

Проверка

не даёт превратить связь в
опасный вывод

Проверочный вопрос

Что нужно сделать перед тем, как строить модель на медицинских данных?

Ответ: открыть таблицу, понять объект, признаки, цель и проверить смысл данных.

Filter...

Data Table Paint Data Data Info Rank



Edit Domain



Color



Column Statistics



Save Data



Transform



Visualize



Model



Evaluate



Unsupervised



Prototypes



Dr Igor Tools



Data Doctor



Image Doctor



Paradox Factory



Regression Tuner



Tree Tuner



Image Analytics



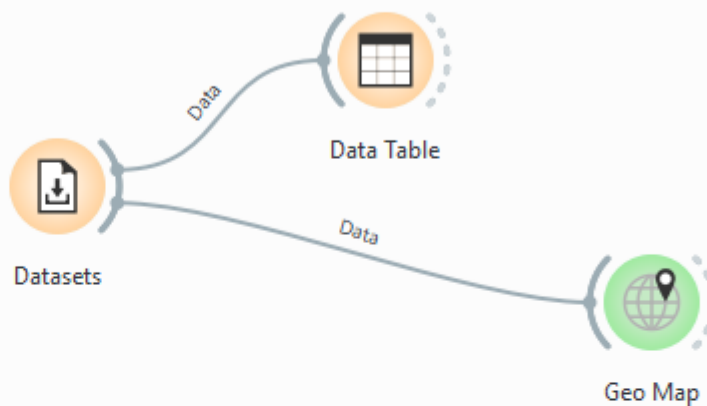
Associate



Bioinformatics



Educational



Paradox Factory



Data Lab

Regression Flow · Neon glass analytics

Run Flow

Load Demo

Save Flow

Export JSON

Final Document

MODULE LIBRARY

drag or click

Search modules...

DATA

- Data Source**
CSV / demo data
- Smart Summary**
types, missing, stats
- Select Target**
prediction column

VISUALS

- Scatter Plot**
x / y / color controls
- Histogram**
distribution and bins
- Correlation View**
target relations

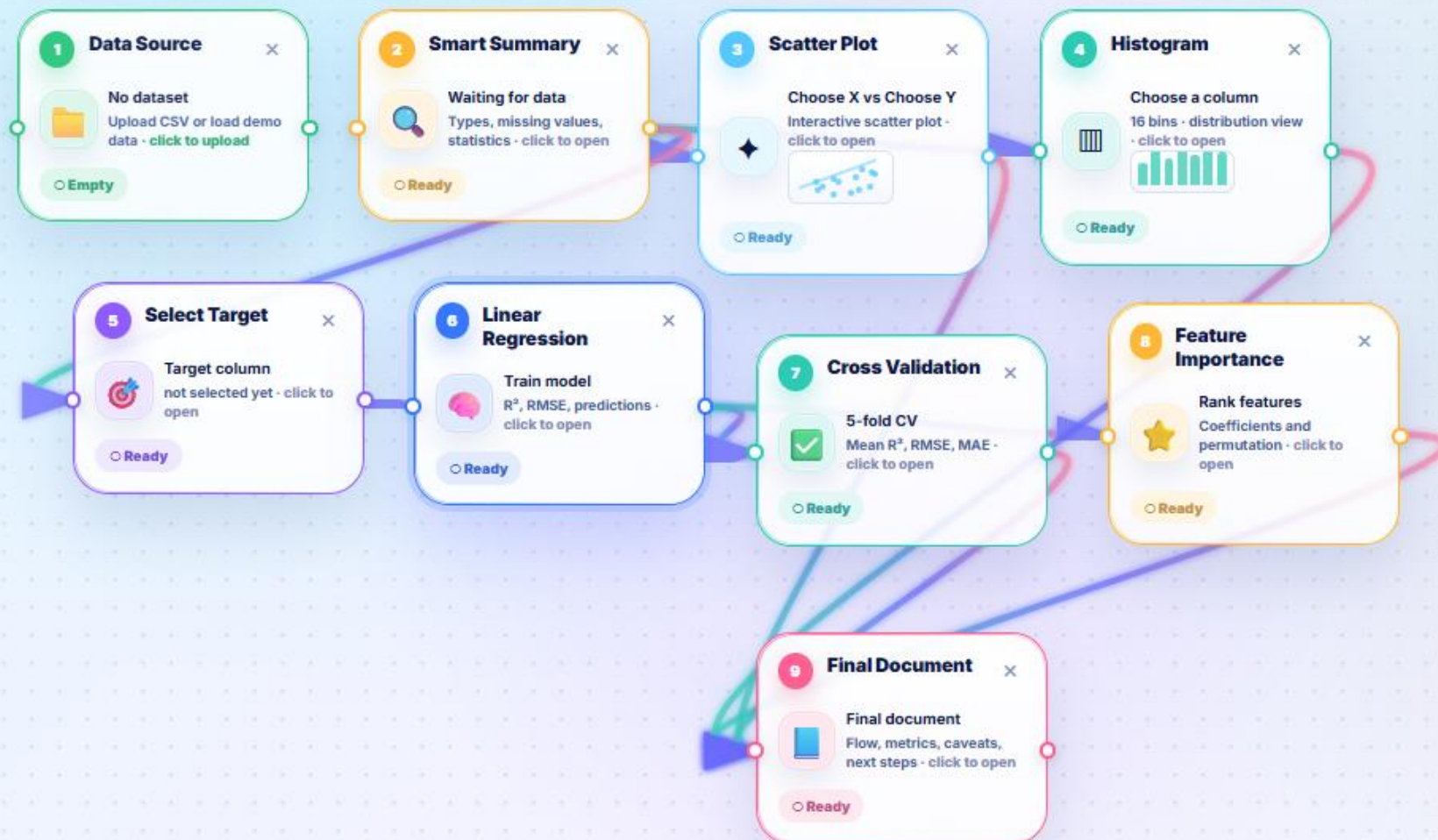
MODEL

- Linear Regression**
OLS / ridge-lite
- Cross Validation**
k-fold scores
- Feature Importance**
coefficients / permutation

OUTPUT

- Report**
beautiful summary

Ready for data No dataset loaded



Искусственный интеллект в анализе данных

Помощник, ускоритель и источник рисков

**ИИ — помощник аналитика.
Ответы принимает человек.**



2. Что изучим

ИИ уже рядом с анализом данных. Наша задача — использовать его сильно, честно и проверяемо.



**как ИИ помогает учиться и
начинать анализ**



**как он помогает строить workflow
в Orange**



как он ускоряет работу с кодом



**как он помогает объяснять модели
и графики**



**какие ошибки и ограничения
нельзя игнорировать**



**почему воспроизводимость
важнее красивого ответа**

3. ИИ помогает начать анализ

Самая частая польза ИИ — не “дать готовую истину”, а помочь быстро разложить задачу на понятные шаги.

1 Понять вопрос

что именно мы хотим
узнать?

2 Предложить гипотезы

какие связи стоит
проверить?

3 Составить план

какие данные, графики
и модели нужны?

4 Найти слабые места

что может быть
неверно или неполно?

Главная мысль: ИИ снижает страх перед первым шагом, но не отменяет проверку.

3. ИИ помогает начать анализ

Самая частая польза ИИ — не “дать готовую истину”, а помочь быстро разложить задачу на понятные шаги.

1 Понять вопрос

что именно мы хотим
узнать?

2 Предложить гипотезы

какие связи стоит
проверить?

3 Составить план

какие данные, графики
и модели нужны?

4 Найти слабые места

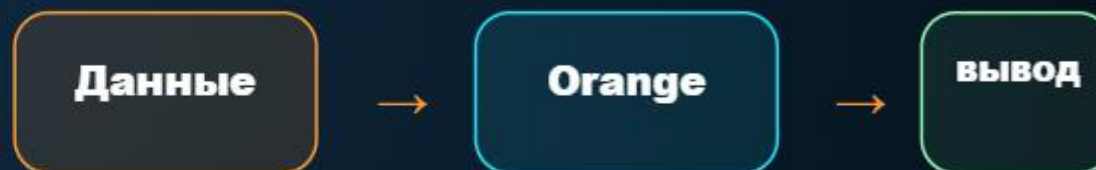
что может быть
неверно или неполно?

Главная мысль: ИИ снижает страх перед первым шагом, но не отменяет проверку.

4. ИИ + Orange: конструктор с подсказками

Orange даёт визуальные блоки. ИИ помогает выбрать путь, но результат всё равно проверяется на данных.

- ✓ какой виджет попробовать первым
- ✓ как соединить шаги анализа
- ✓ какую визуализацию выбрать
- ✓ как сформулировать вывод
- ✓ какой контрольный вопрос задать



Не “нейросеть сказала”, а “мы проверили”

ИИ может предложить workflow. Ответ становится учебным результатом только после просмотра данных, графиков и проверки смысла.

5. ИИ + код: быстрый черновик, не финальный ответ

ИИ хорошо помогает там, где нужно быстро написать или понять небольшой фрагмент кода.

пишет базовый Python / SQL

ускоряет рутинные шаги

объясняет чужой код

помогает читать, а не только копировать

ищет ошибку

предлагает гипотезу, где сломалось

код может быть неверным

он может работать, но считать не то

версии библиотек меняются

результат зависит от среды

нужен тест на маленьком примере

без проверки код остаётся черновиком

Код от ИИ — это приглашение проверить, а не повод закрыть глаза.

6. ИИ помогает объяснять модели

Модель даёт числа. ИИ помогает превратить их в понятный язык — но не доказывает, что модель права.

Ассурасу, AUC, Recall

перевести метрики на человеческий язык

Графики и ошибки

описать, что видно и что может быть подозрительным

Ограничения модели

сформулировать, где вывод может не работать

Сравнение вариантов

помочь увидеть плюсы и минусы моделей

Объяснение ИИ — это текст. Правильность проверяется данными, метриками и смыслом задачи.

7. Суперсилы ИИ в учебном курсе

Сильный ИИ в обучении — это не магия, а ускоритель вопросов, черновиков и обратной связи.



скорость

быстро получить черновик идеи



ориентация

понять, с чего начать



варианты

увидеть несколько путей анализа



объяснение

разобрать термин простым языком



проверка

искать слабые места в выводе



эксперимент

придумать новый тест на данных

ИИ особенно полезен, когда рядом есть человек, который включает режим: “Стоп, проверяем!”

8. Где ИИ опасен

Самая опасная ошибка — принять уверенный текст за проверенный результат.

Галлюцинации

ИИ может уверенно сообщить неверный факт или придумать деталь.

Разные ответы

один и тот же вопрос может получить разные варианты ответа.

Скрытая ошибка

ошибка в данных, коде или формуле может выглядеть красиво.

Лишняя уверенность

хороший стиль текста не означает хорошую аналитику.

Закрытые данные

персональные и конфиденциальные данные нельзя отправлять куда попало.

Копирование без понимания

готовый ответ не заменяет собственного вывода.

9. Воспроизводимость: ответ должен повторяться

В анализе данных ценится не самый красивый ответ, а такой результат, который можно проверить и получить снова.

данные + шаги анализа + параметры + проверка → воспроизводимый результат

Данные

какой файл, какая версия, какие строки

Workflow

какие виджеты Orange и в каком порядке

Код

какие библиотеки, параметры, seed

ИИ

какой запрос, какой ответ, что проверено

Вывод

какая метрика, какой график, какое ограничение

Если завтра результат нельзя повторить, сегодня он ещё не готов.

10. Правило работы с ИИ

ИИ можно использовать. Но каждый ответ проходит через контроль аналитика.

1 Сформулируй вопрос

что нужно узнать и на каких данных

2 Попроси вариант

план, объяснение, код или идею

3 Проверь на данных

таблица, график, метрика, смысл

4 Сохрани шаги

workflow, код, параметры, важные решения

5 Напиши свой вывод

не копия ответа, а проверенный результат

Финальный фильтр курса: “Стоп, проверяем!”

11. Что мы выучили

ИИ — сильный помощник, но ответственность за анализ остаётся у человека.

ИИ помогает

задавать вопросы, строить план, писать код, объяснять результаты

ИИ ошибается

может менять ответы, звучать уверенно и пропускать проблемы

Проверка обязательна

данные, графики, метрики и смысл важнее красивого текста

Воспроизводимость

результат должен быть понятен, сохранён и повторяем

Проверочный вопрос

Что важнее в анализе данных: быстрый ответ ИИ, красивый ответ или проверяемый результат?

Правильный акцент: проверяемый и воспроизводимый результат.

Доктор Игорь одобряет ИИ — после контрольного осмотра данных.