Уральский федеральный университет
имени первого Президента России Б.Н.Ельцина

Курс «Сбор и верификация данных»

Введение

Тема 1.1. Введение в сбор и верификацию данных.

Уральский федеральный университет
имени первого Президента России Б.Н.Ельцина

Лекцию читает

Станислав Поляков

Старший преподаватель
Учебно-научный центр «Искусственный интеллект»
Институт радиоэлектроники и информационных технологий-РТФ
Уральский федеральный университет имени первого Президента России Б.Н. Ельцина

Сбор и верификация данных

О чем этот курс?

Мы научимся превращать информационный хаос в качественное сырье для анализа и машинного обучения.

Сбор и верификация данных

Технологический стек

pandas

Requests

BeautifulSoup

Selenium

matplotlib

seaborn

plotly

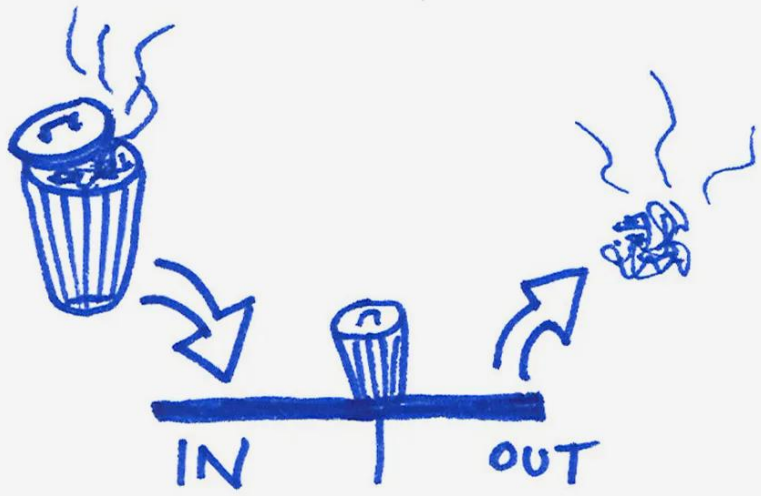
Streamlit

NumPy

Сбор и верификация данных

Garbage In → Garbage Out (GIGO)

Если на вход подать «мусор», на выходе получится «мусор», какой бы крутой ни была ваша модель.



Сбор и верификация данных

Чем на самом деле занят Data Scientist?

80% — сбор и очистка данных.
20% — построение моделей и красивые графики.

Сбор и верификация данных

Цена ошибки: Данные vs Реальность

Алгоритм определял наличие объектов не по фактической информации, а по погоде на фото.

Сбор и верификация данных

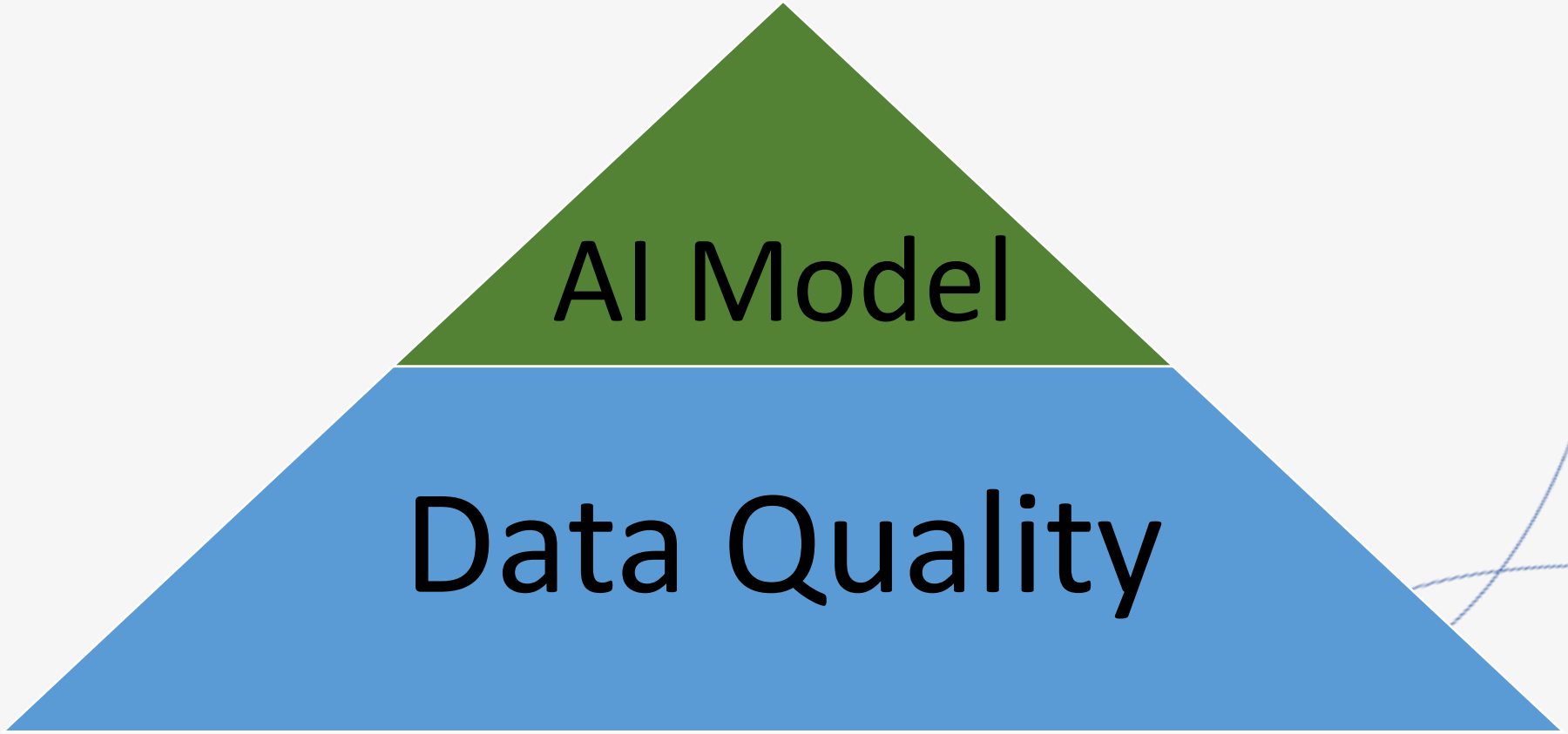
Дискриминация в данных: Amazon Case

Алгоритм найма отклонял резюме женщин, потому что «учился» на данных за 10 лет, когда нанимали в основном мужчин.

Сбор и верификация данных

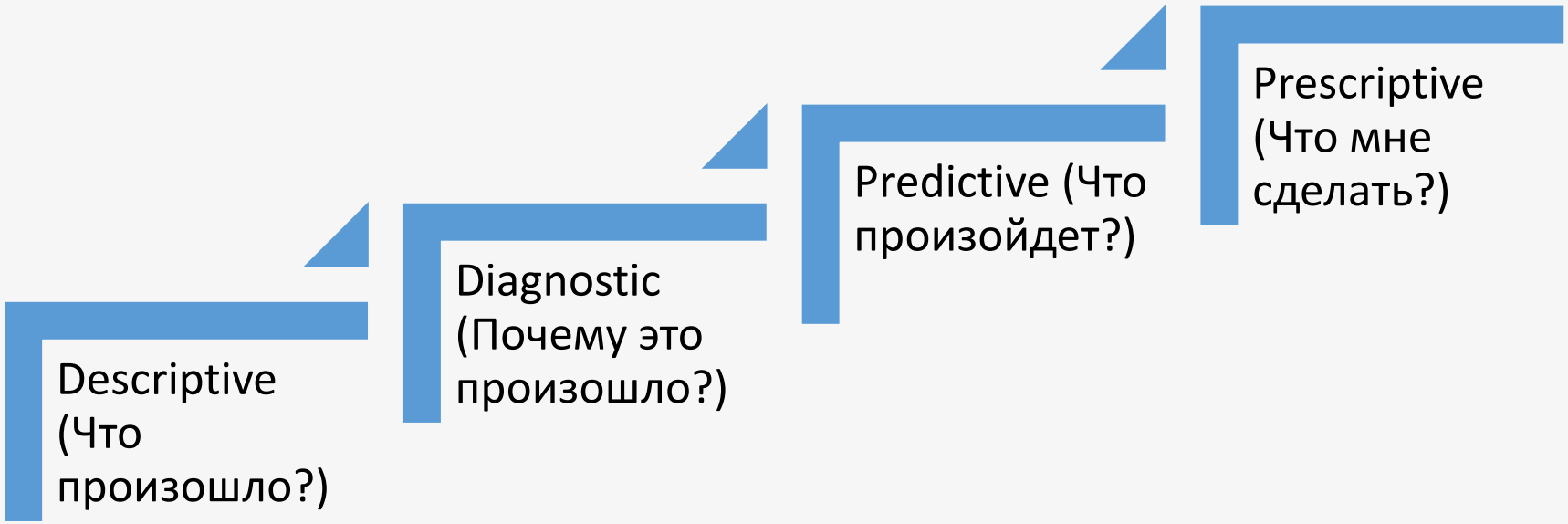
Данные — это фундамент

Ошибка на этапе сбора стоит в 10 раз дороже, чем ошибка в коде модели.



Сбор и верификация данных

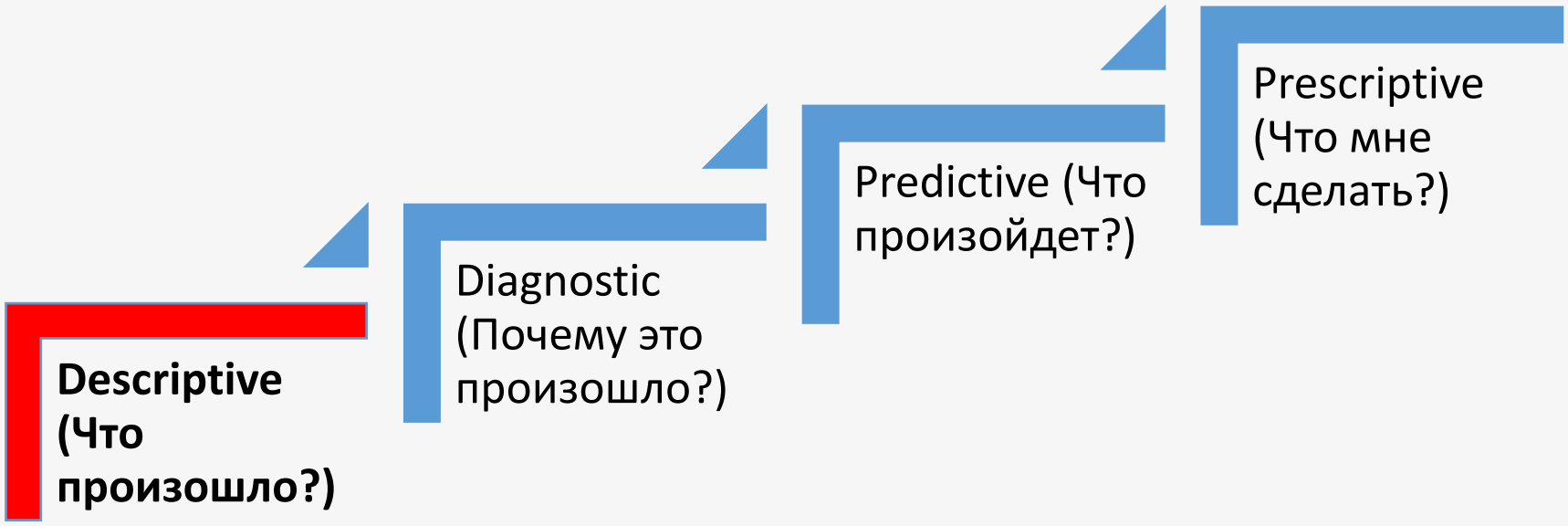
Лестница ценности данных



Сбор и верификация данных

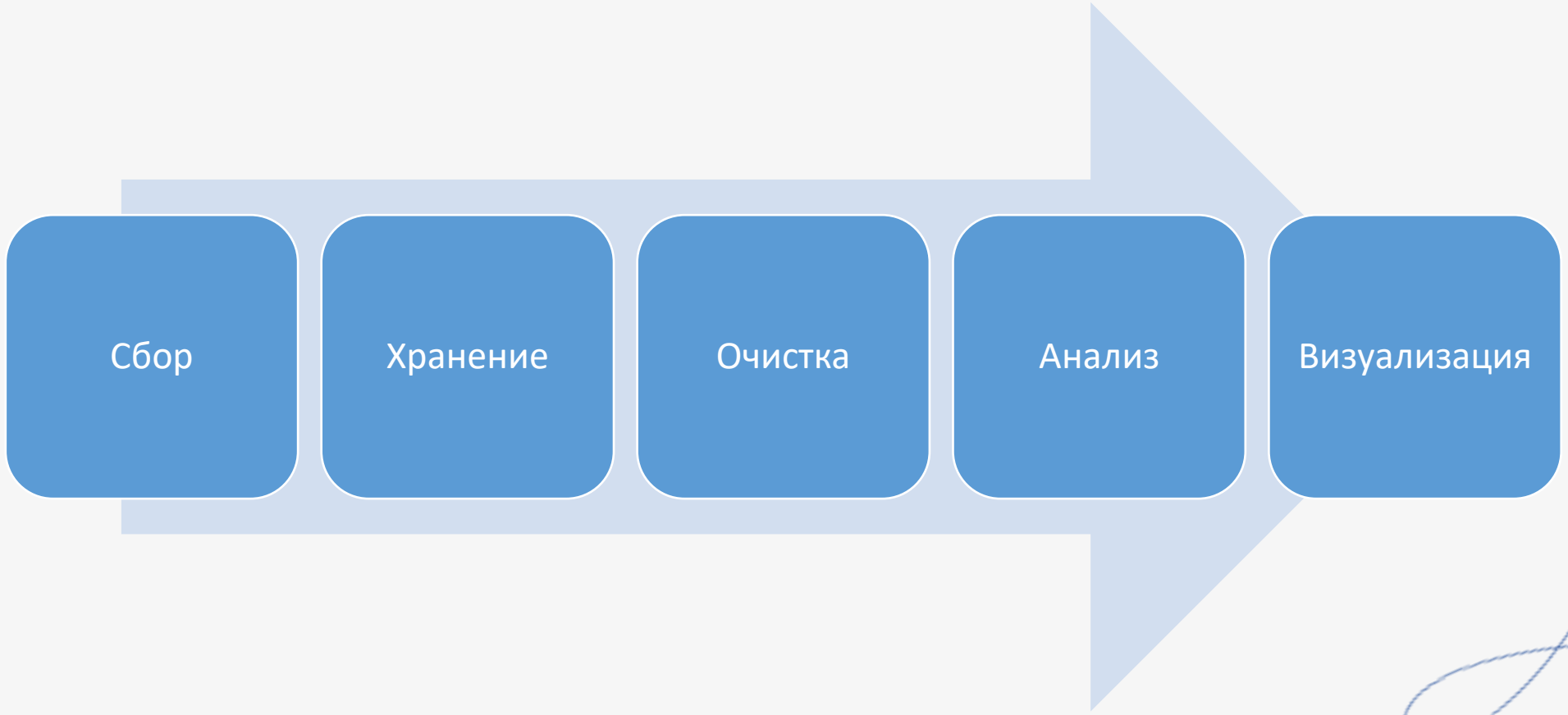
Где находимся мы?

Без качественного сбора (нижняя ступень) верхние уровни недостижимы.



Сбор и верификация данных

Жизненный цикл данных (Data Lifecycle)

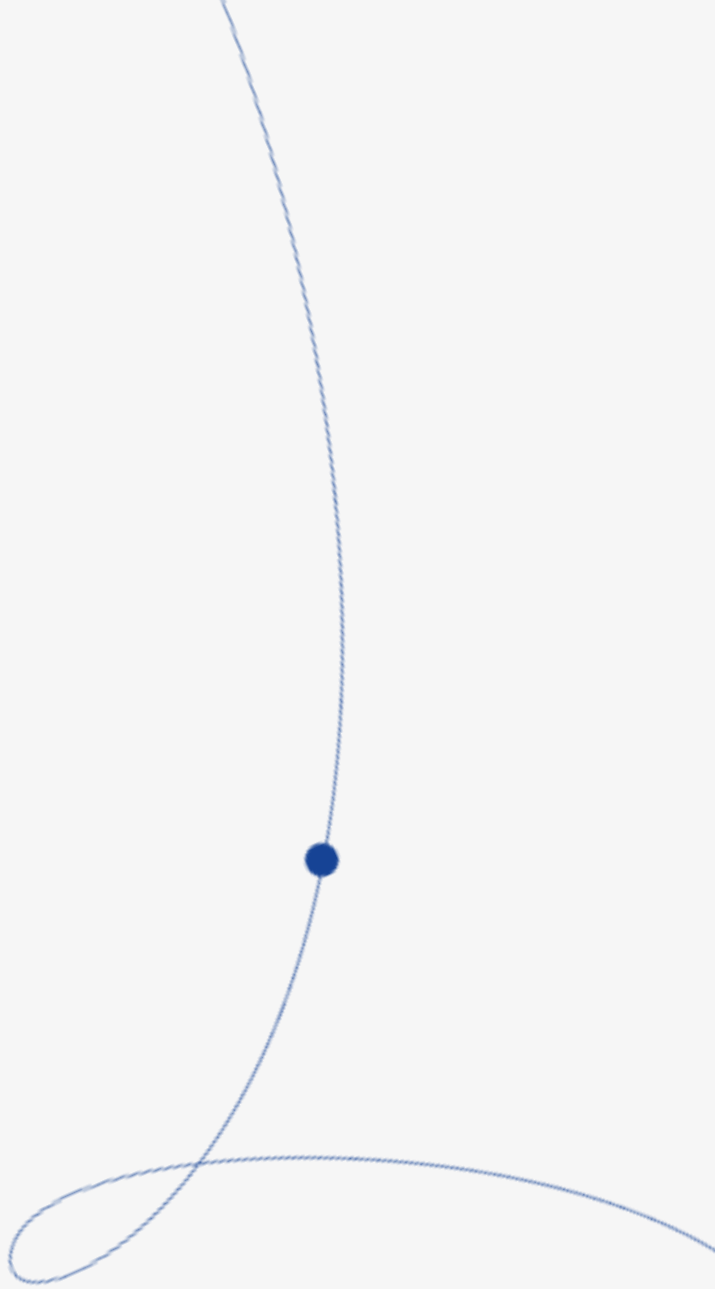


Сбор и верификация данных

Data Engineer vs Data Scientist

Инженер: готовит данные (сбор, верификация).

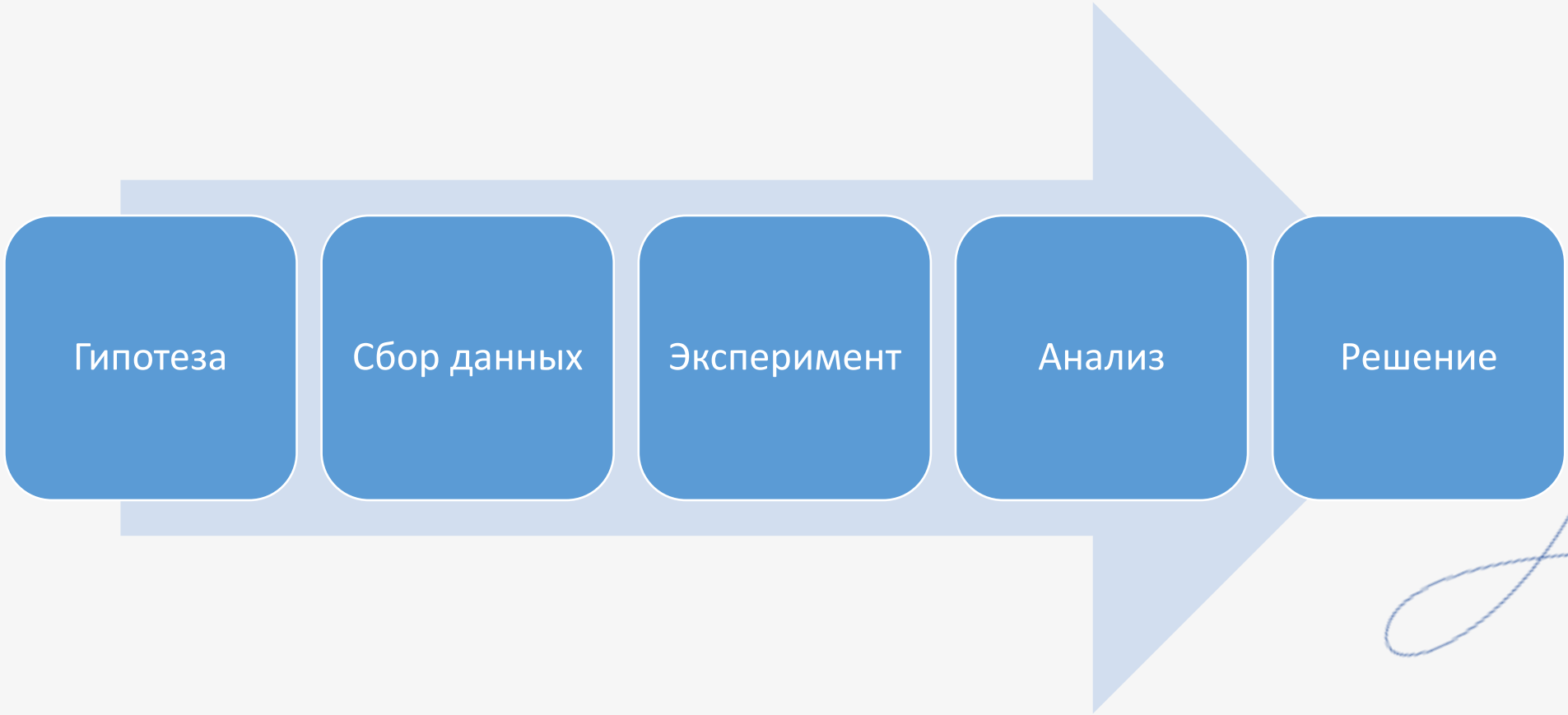
Сайентист: работает с тем, что подготовил инженер.



Сбор и верификация данных

Принцип Data-Driven

Решения принимаются на основе фактов, а не интуиции.



Сбор и верификация данных

Типы данных

Теперь нужно понять, с какими данными предстоит работать

Сбор и верификация данных

1.Структурированные (Structured)

Реляционные базы данных, Excel/CSV.
Строгая схема, фиксированные поля.

	Laptop	Status	Brand	Model	CPU	RAM	Storage	Storage type	GPU	Screen	Touch	Final Price
0	ASUS ExpertBook B1 B1502CBA-EJ0436X Intel Core...	New	Asus	ExpertBook	Intel Core i5	8	512	SSD	NaN	15.6	No	1009.00
1	Alurin Go Start Intel Celeron N4020/8GB/256GB ...	New	Alurin	Go	Intel Celeron	8	256	SSD	NaN	15.6	No	299.00
2	ASUS ExpertBook B1 B1502CBA-EJ0424X Intel Core...	New	Asus	ExpertBook	Intel Core i3	8	256	SSD	NaN	15.6	No	789.00
3	MSI Katana GF66 12UC-082XES Intel Core i7-1270...	New	MSI	Katana	Intel Core i7	16	1000	SSD	RTX 3050	15.6	No	1199.00
4	HP 15S-FQ5085NS Intel Core i5-1235U/16GB/512GB...	New	HP	15S	Intel Core i5	16	512	SSD	NaN	15.6	No	669.01
...	...	...	...	...	...	...	...	...	...	...	...	...
2155	Razer Blade 17 FHD 360Hz Intel Core i7-11800H/...	Refurbished	Razer	Blade	Intel Core i7	16	1000	SSD	RTX 3060	17.3	No	2699.99
2156	Razer Blade 17 FHD 360Hz Intel Core i7-11800H/...	Refurbished	Razer	Blade	Intel Core i7	16	1000	SSD	RTX 3070	17.3	No	2899.99
2157	Razer Blade 17 FHD 360Hz Intel Core i7-11800H/...	Refurbished	Razer	Blade	Intel Core i7	32	1000	SSD	RTX 3080	17.3	No	3399.99
2158	Razer Book 13 Intel Evo Core i7-1165G7/16GB/1T...	Refurbished	Razer	Book	Intel Evo Core i7	16	1000	SSD	NaN	13.4	Yes	1899.99
2159	Razer Book FHD+ Intel Evo Core i7-1165G7/16GB/...	Refurbished	Razer	Book	Intel Evo Core i7	16	256	SSD	NaN	13.4	Yes	1699.99

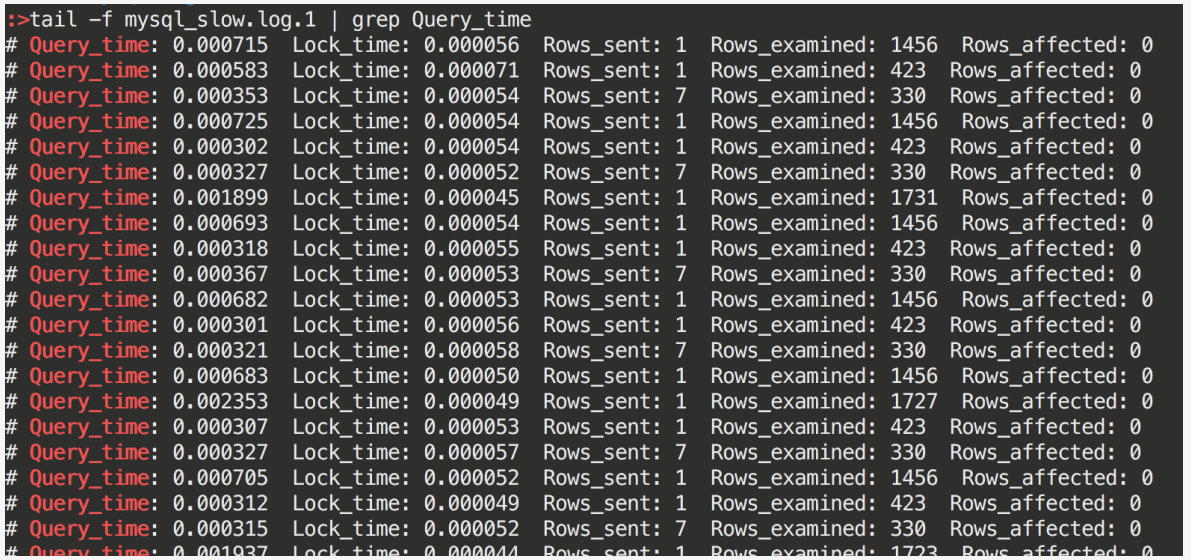
2160 rows × 12 columns

Сбор и верификация данных

2.Полуструктурированные

JSON, XML, HTML, логи серверов. Схемы нет, но есть теги или ключи.

```
{
  "people" : [
    {
        "firstName": "Joe",
        "lastName": "Jackson",
        "gender": "male",
        "age": 28,
        "number": "7349282382"
    },
    {
        "firstName": "James",
        "lastName": "Smith",
        "gender": "male",
        "age": 32,
        "number": "5678568567"
    },
    {
        "firstName": "Emily",
        "lastName": "Jones",
        "gender": "female",
        "age": 24,
        "number": "456754675"
    }
  ]
}
```

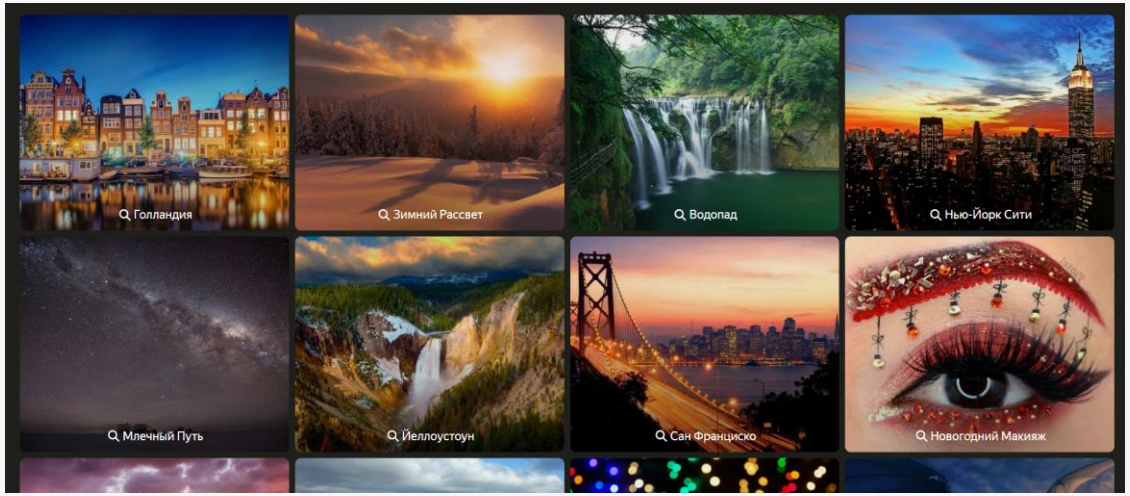


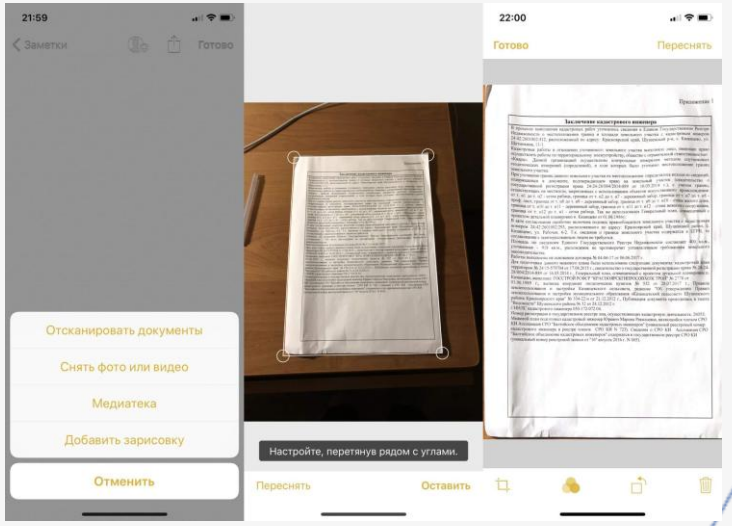
```
:>tail -f mysql_slow.log.1 | grep Query_time
# Query_time: 0.000715  Lock_time: 0.000056  Rows_sent: 1  Rows_examined: 1456  Rows_affected: 0
# Query_time: 0.000583  Lock_time: 0.000071  Rows_sent: 1  Rows_examined: 423  Rows_affected: 0
# Query_time: 0.000353  Lock_time: 0.000054  Rows_sent: 7  Rows_examined: 330  Rows_affected: 0
# Query_time: 0.000725  Lock_time: 0.000054  Rows_sent: 1  Rows_examined: 1456  Rows_affected: 0
# Query_time: 0.000302  Lock_time: 0.000054  Rows_sent: 1  Rows_examined: 423  Rows_affected: 0
# Query_time: 0.000327  Lock_time: 0.000052  Rows_sent: 7  Rows_examined: 330  Rows_affected: 0
# Query_time: 0.001899  Lock_time: 0.000045  Rows_sent: 1  Rows_examined: 1731  Rows_affected: 0
# Query_time: 0.000693  Lock_time: 0.000054  Rows_sent: 1  Rows_examined: 1456  Rows_affected: 0
# Query_time: 0.000318  Lock_time: 0.000055  Rows_sent: 1  Rows_examined: 423  Rows_affected: 0
# Query_time: 0.000367  Lock_time: 0.000053  Rows_sent: 7  Rows_examined: 330  Rows_affected: 0
# Query_time: 0.000682  Lock_time: 0.000053  Rows_sent: 1  Rows_examined: 1456  Rows_affected: 0
# Query_time: 0.000301  Lock_time: 0.000056  Rows_sent: 1  Rows_examined: 423  Rows_affected: 0
# Query_time: 0.000321  Lock_time: 0.000058  Rows_sent: 7  Rows_examined: 330  Rows_affected: 0
# Query_time: 0.000683  Lock_time: 0.000050  Rows_sent: 1  Rows_examined: 1456  Rows_affected: 0
# Query_time: 0.002353  Lock_time: 0.000049  Rows_sent: 1  Rows_examined: 1727  Rows_affected: 0
# Query_time: 0.000307  Lock_time: 0.000053  Rows_sent: 1  Rows_examined: 423  Rows_affected: 0
# Query_time: 0.000327  Lock_time: 0.000057  Rows_sent: 7  Rows_examined: 330  Rows_affected: 0
# Query_time: 0.000705  Lock_time: 0.000052  Rows_sent: 1  Rows_examined: 1456  Rows_affected: 0
# Query_time: 0.000312  Lock_time: 0.000049  Rows_sent: 1  Rows_examined: 423  Rows_affected: 0
# Query_time: 0.000315  Lock_time: 0.000052  Rows_sent: 7  Rows_examined: 330  Rows_affected: 0
# Query_time: 0.001937  Lock_time: 0.000044  Rows_sent: 1  Rows_examined: 1723  Rows_affected: 0
```

Сбор и верификация данных

3.Неструктурированные данные

Самый большой объем (80% мировых данных). Тексты, аудио, видео, картинки, PDF.



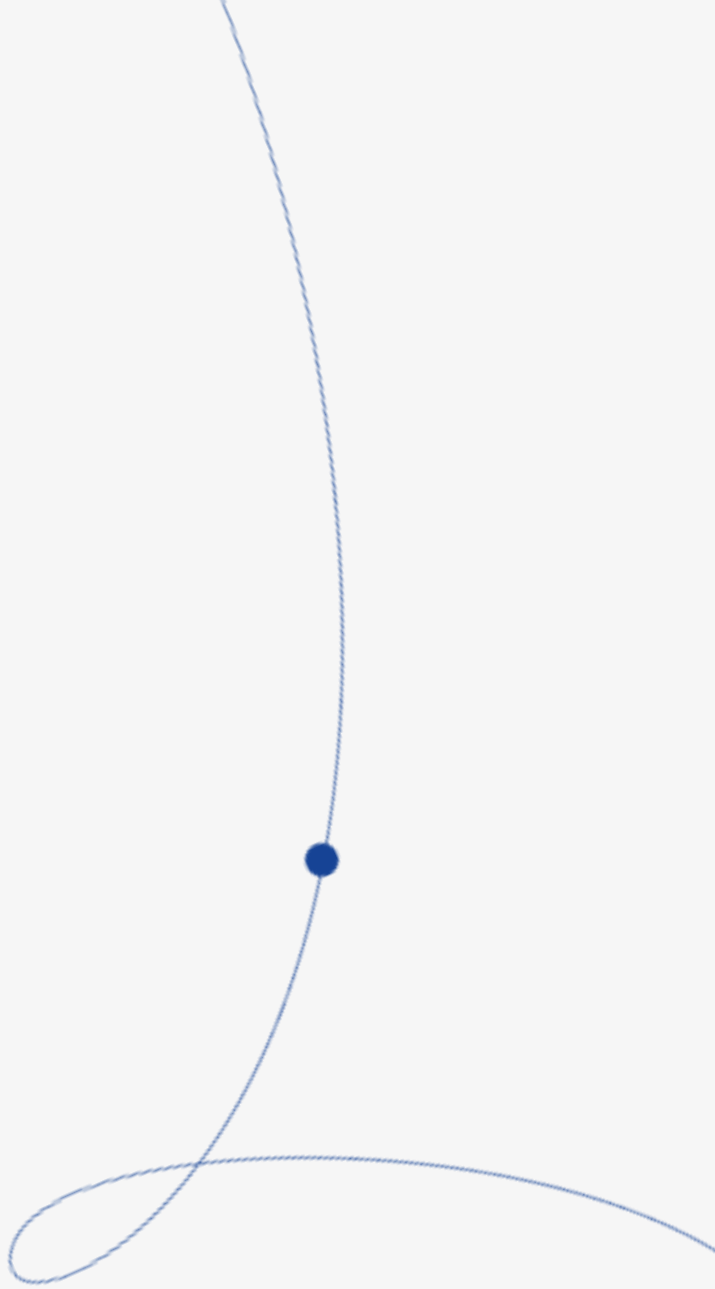


Сбор и верификация данных

Сложность vs Гибкость

Структурированные: Легко считать — трудно менять.

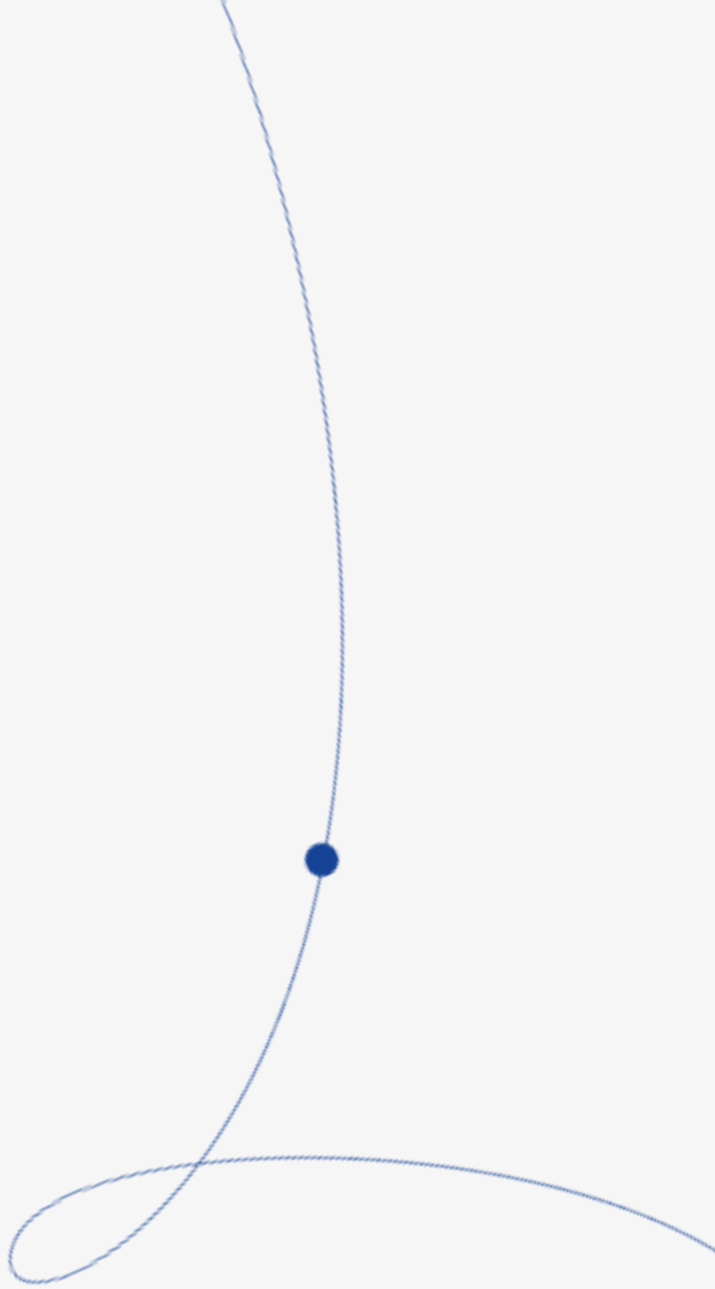
Неструктурированные: Тяжело парсить — бесконечная гибкость.



Сбор и верификация данных

Наша главная цель в Python

Превратить хаос (Scraping, API, Logs) в структуру (Pandas DataFrame).



Сбор и верификация данных

Где рождаются данные?

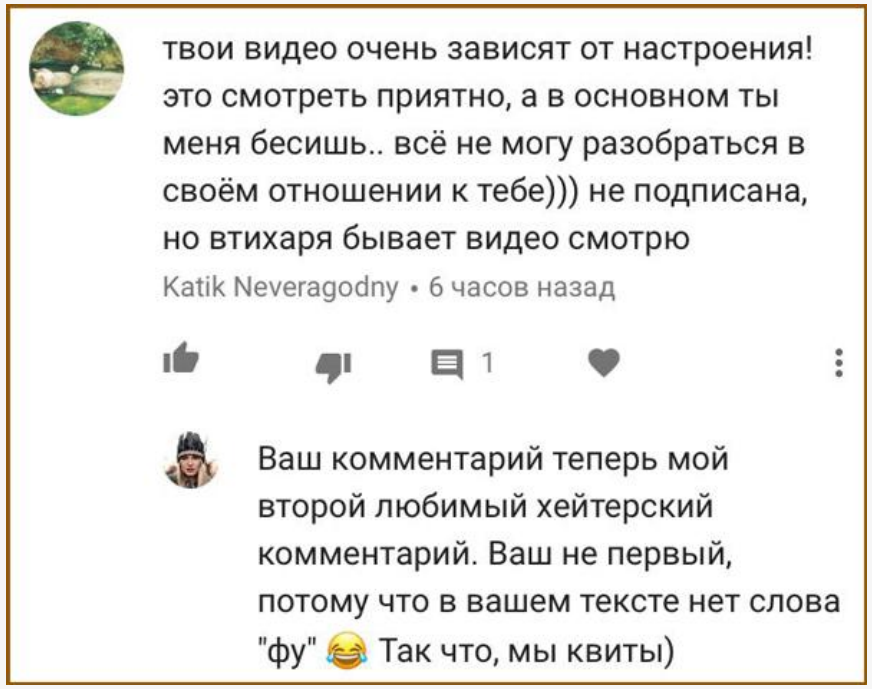
Данные не возникают из ниоткуда. Источник определяет методы сбора и уровень доверия.

Сбор и верификация данных

Человек как источник (User Generated)

Соцсети, отзывы, анкеты, поисковые запросы.

Особенности: Опечатки, сленг, сарказм, субъективность.

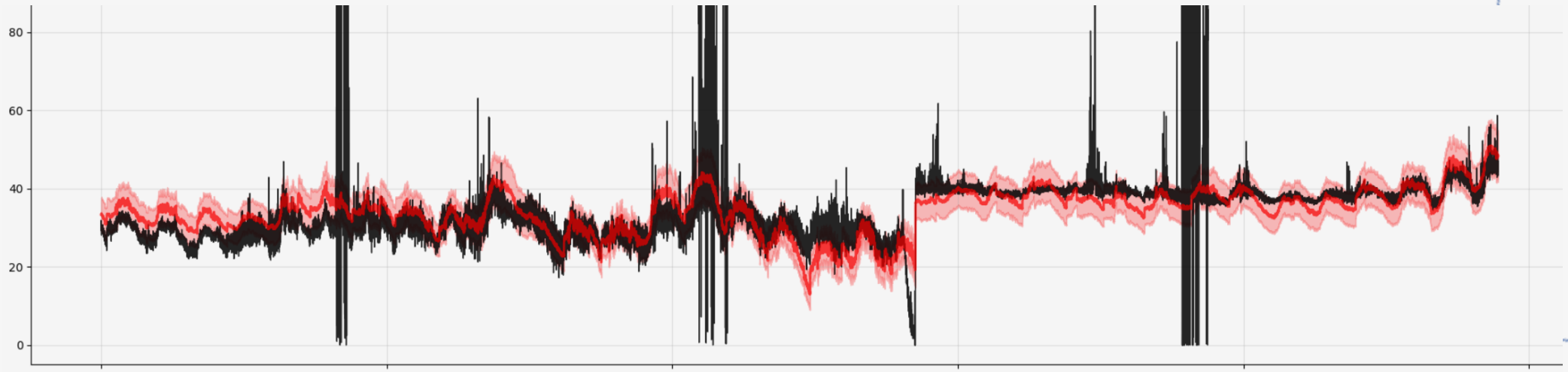


Сбор и верификация данных

Машины и датчики (Machine Generated)

Логи серверов, телеметрия авто, датчики температуры, умные часы.

Особенности: Огромный объем, строгий формат, но частые артефакты (пропуски) из-за сбоев связи или других особенностей оборудования.

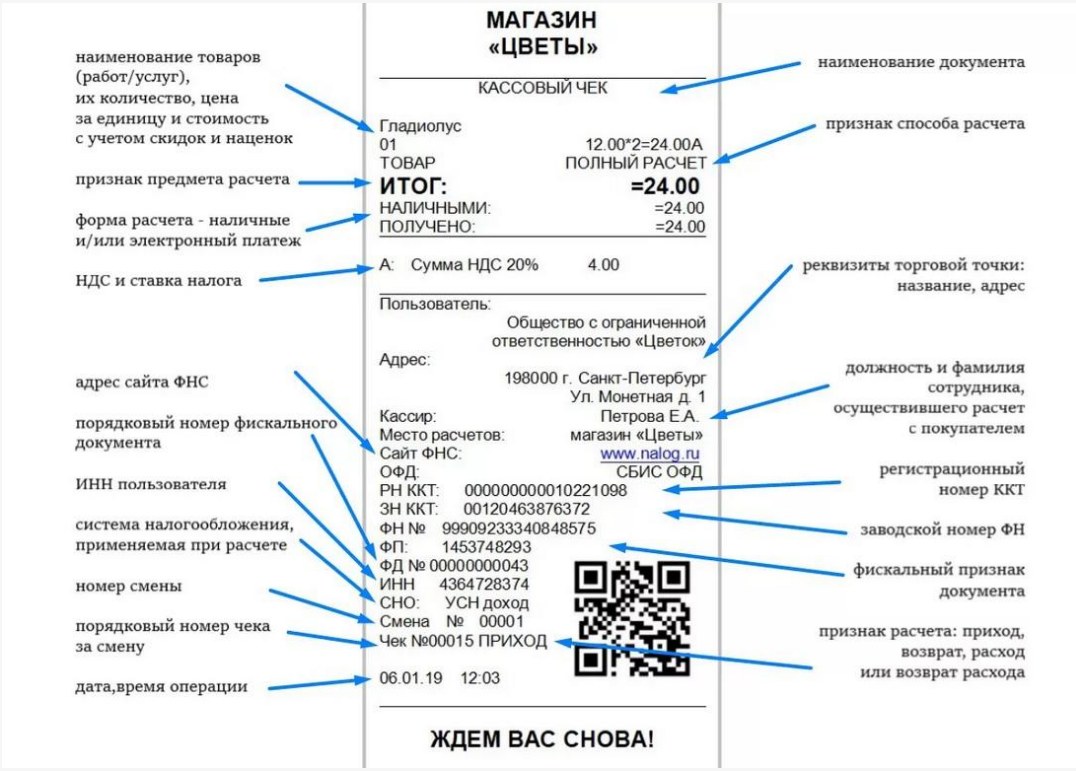


Сбор и верификация данных

Транзакционные данные

Банковские переводы, чеки в супермаркетах, бронирования отелей.

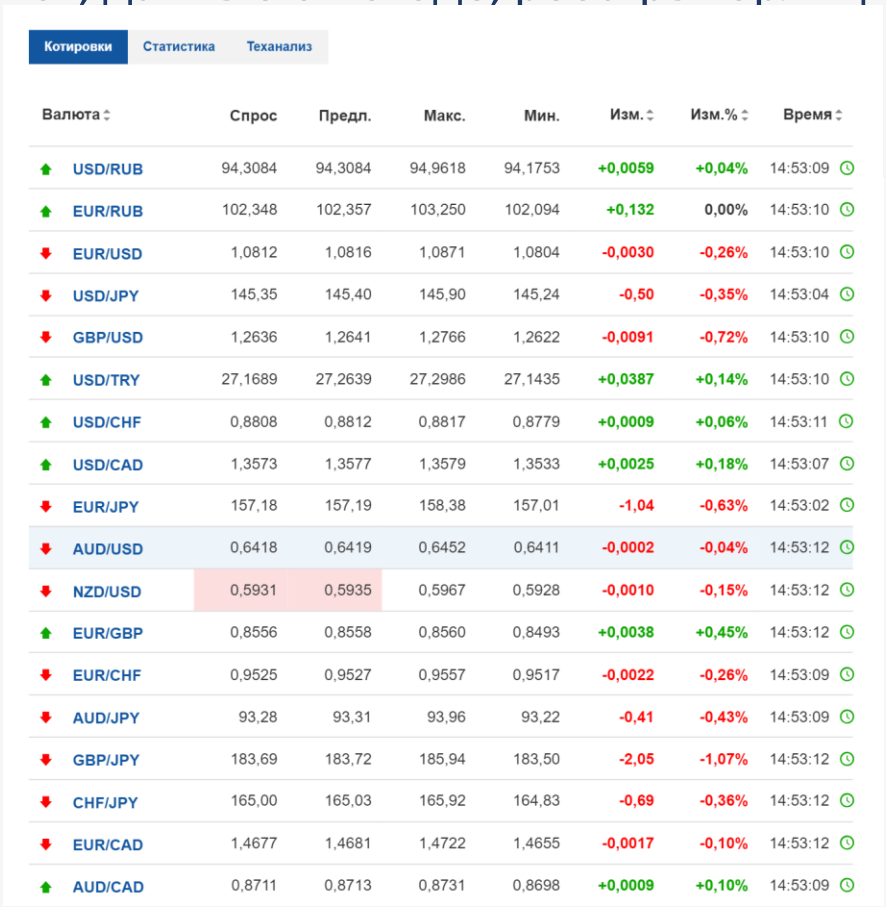
Особенности: Самые «чистые» и ценные данные. Высокая цена ошибки.



Сбор и верификация данных

Открытые данные (Open Data)

Демография, курсы валют, данные о погоде, реестры юрлиц.



Котировки | Статистика | Теханализ

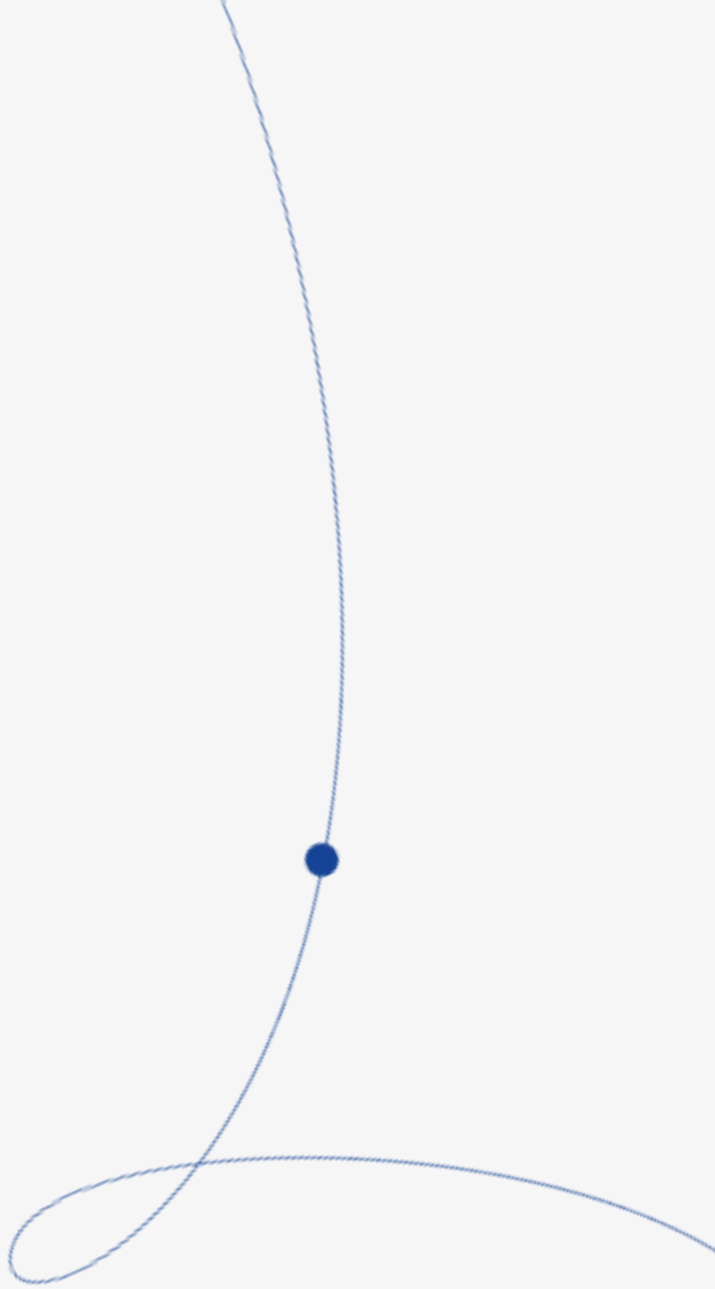
Валюта	Спрос	Предл.	Макс.	Мин.	Изм.	Изм.%	Время
USD/RUB	94,3084	94,3084	94,9618	94,1753	+0,0059	+0,04%	14:53:09
EUR/RUB	102,348	102,357	103,250	102,094	+0,132	0,00%	14:53:10
EUR/USD	1,0812	1,0816	1,0871	1,0804	-0,0030	-0,26%	14:53:10
USD/JPY	145,35	145,40	145,90	145,24	-0,50	-0,35%	14:53:04
GBP/USD	1,2636	1,2641	1,2766	1,2622	-0,0091	-0,72%	14:53:10
USD/TRY	27,1689	27,2639	27,2986	27,1435	+0,0387	+0,14%	14:53:10
USD/CHF	0,8808	0,8812	0,8817	0,8779	+0,0009	+0,06%	14:53:11
USD/CAD	1,3573	1,3577	1,3579	1,3533	+0,0025	+0,18%	14:53:07
EUR/JPY	157,18	157,19	158,38	157,01	-1,04	-0,63%	14:53:02
AUD/USD	0,6418	0,6419	0,6452	0,6411	-0,0002	-0,04%	14:53:12
NZD/USD	0,5931	0,5935	0,5967	0,5928	-0,0010	-0,15%	14:53:12
EUR/GBP	0,8556	0,8558	0,8560	0,8493	+0,0038	+0,45%	14:53:12
EUR/CHF	0,9525	0,9527	0,9557	0,9517	-0,0022	-0,26%	14:53:09
AUD/JPY	93,28	93,31	93,96	93,22	-0,41	-0,43%	14:53:09
GBP/JPY	183,69	183,72	185,94	183,50	-2,05	-1,07%	14:53:12
CHF/JPY	165,00	165,03	165,92	164,83	-0,69	-0,36%	14:53:12
EUR/CAD	1,4677	1,4681	1,4722	1,4655	-0,0017	-0,10%	14:53:12
AUD/CAD	0,8711	0,8713	0,8731	0,8698	+0,0009	+0,10%	14:53:09

Сбор и верификация данных

Какой источник выбрать?

Стоимость сбора vs Качество vs Объем.

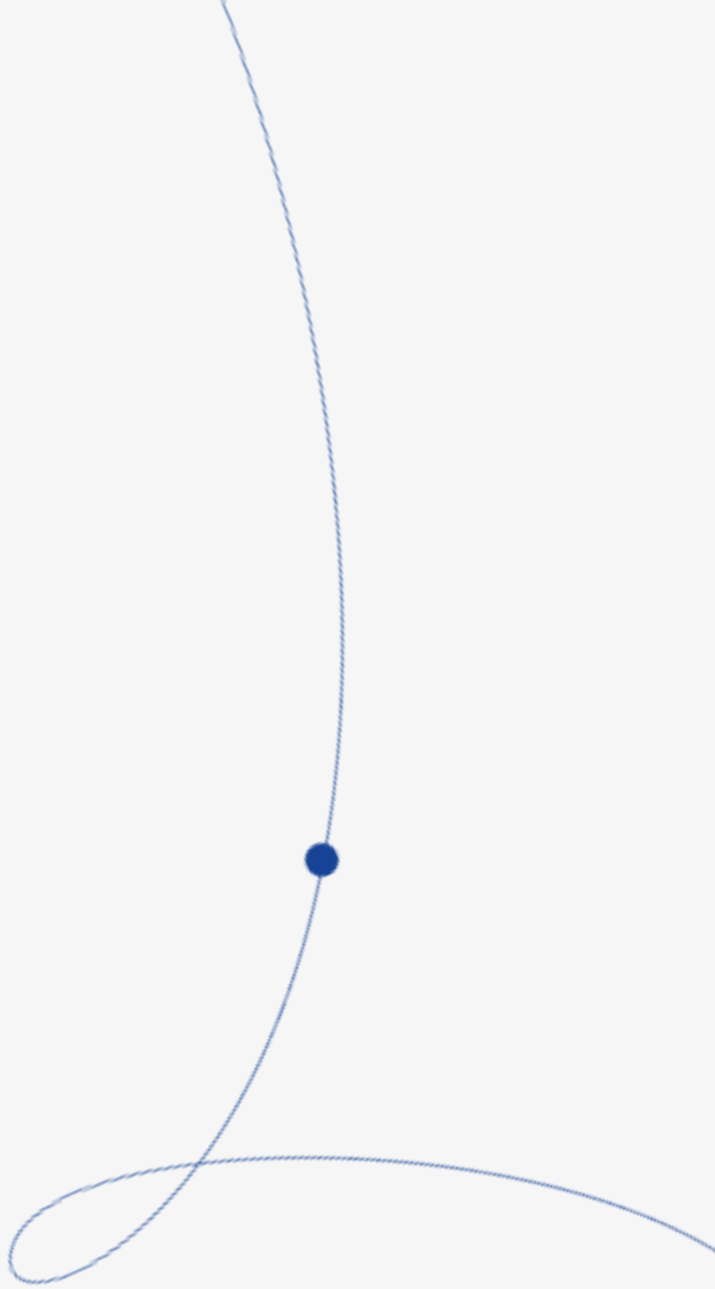
И ТЗ, конечно же.



Сбор и верификация данных

Форматы хранения: Текст и таблицы

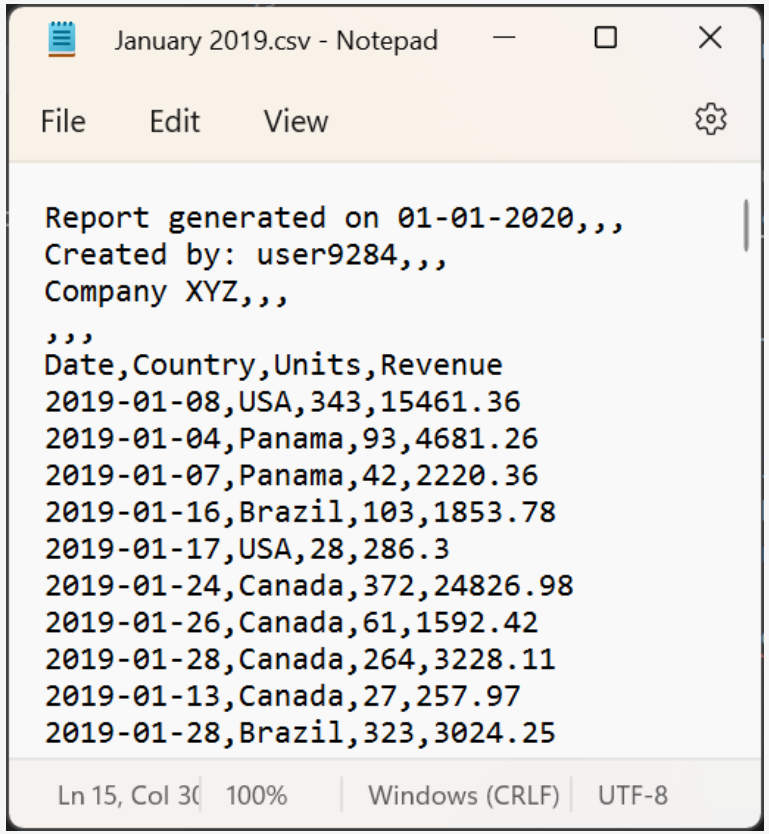
Самые популярные варианты хранения данных



Сбор и верификация данных

CSV (Comma Separated Values)

Простейший текстовый формат. Одна строка — одна запись.



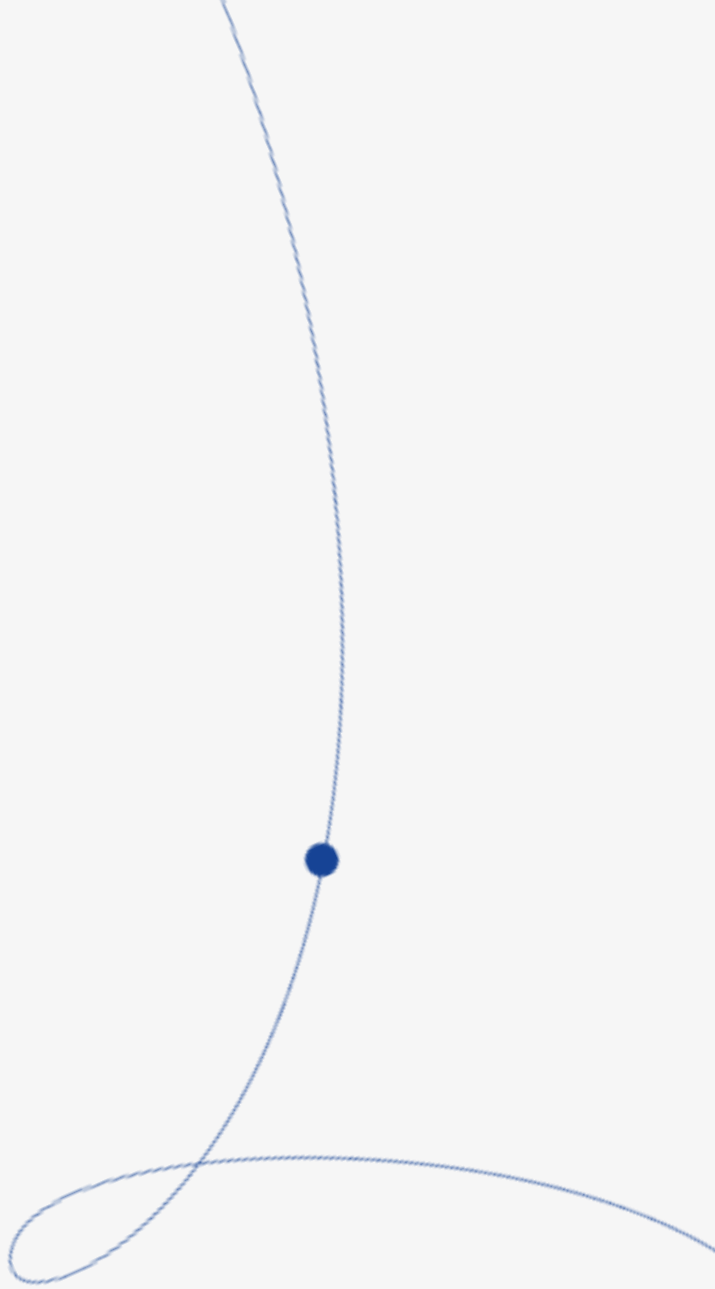
Сбор и верификация данных

Когда CSV ломается?

Запятая внутри текста: "Иванов, Иван".

Разные разделители: запятая (US) vs точка с запятой (EU/RU).

Отсутствие типов: число 001 может превратиться в 1.



Сбор и верификация данных

JSON (JavaScript Object Notation)

Иерархический формат. Ключ — Значение. Идеален для API.

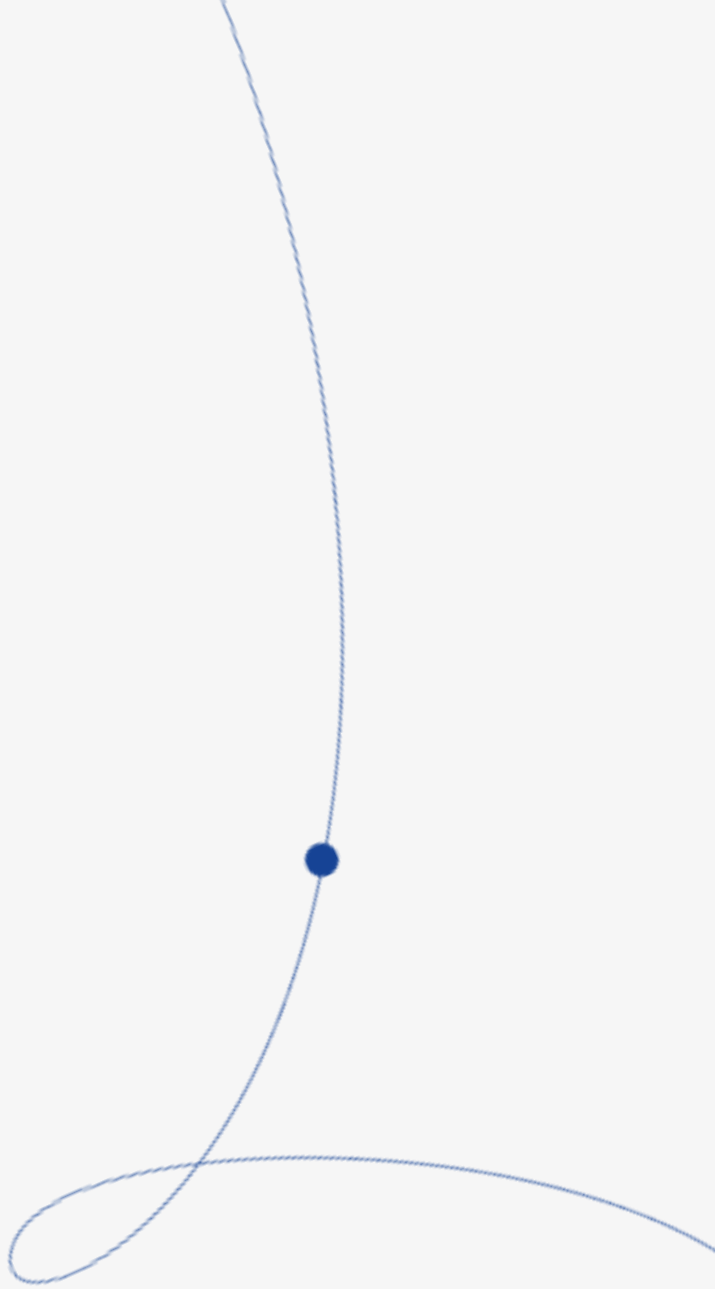
Сбор и верификация данных

Почему JSON вытесняет всё?

Легко читать человеку.

Поддержка вложенности.

Типы данных (число, строка, булевые данные).



Сбор и верификация данных

XML — Старая школа

Формат на основе тегов. Строгий, громоздкий, но надежный. Часто в госструктурах и банках.

```
<?xml version="1.0"?>
- <Dataset xmlns="http://www.safe.com">
  - <Building id="Surrey Head Office">
      <Address>"7445 132 St."</Address>
      <City>Surrey</City>
      <Province>BC</Province>
      <Country>Canada</Country>
    - <Location>
        <Longitude>-122.860</Longitude>
        <Latitude>49.138</Latitude>
      </Location>
      <Reference>https://www.google.ca/maps/
        3m1!4b1!4m5!3m4!1s0x5485dbd520cc
        122.8574636?hl=en</Reference>
    - <Room id="Admin_100">
        <Name>Reception</Name>
        <Category>Admin</Category>
        <Area units="m2">12</Area>
      - <Wall id="Admin_100_WallA">
          <material>wood</material>
        </Wall>
```

Сбор и верификация данных

Какой формат выбрать?

CSV — для простых таблиц и ML.

JSON — для обмена данными в вебе.

XML — когда нужна жесткая валидация схемы.

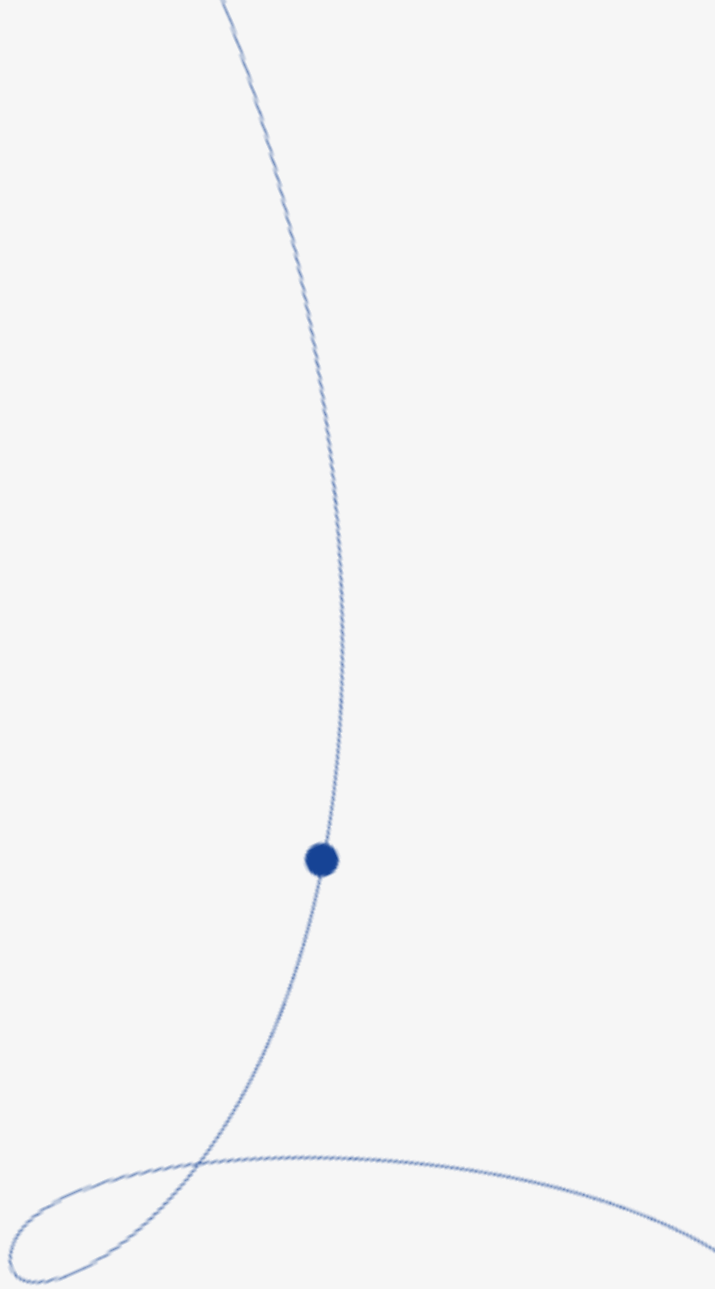
Сбор и верификация данных

Когда CSV и JSON начинают тормозить

Текст занимает много места на диске.

Медленное чтение (нужно парсить каждую запятую).

Типы данных (числа/даты) нужно восстанавливать заново.



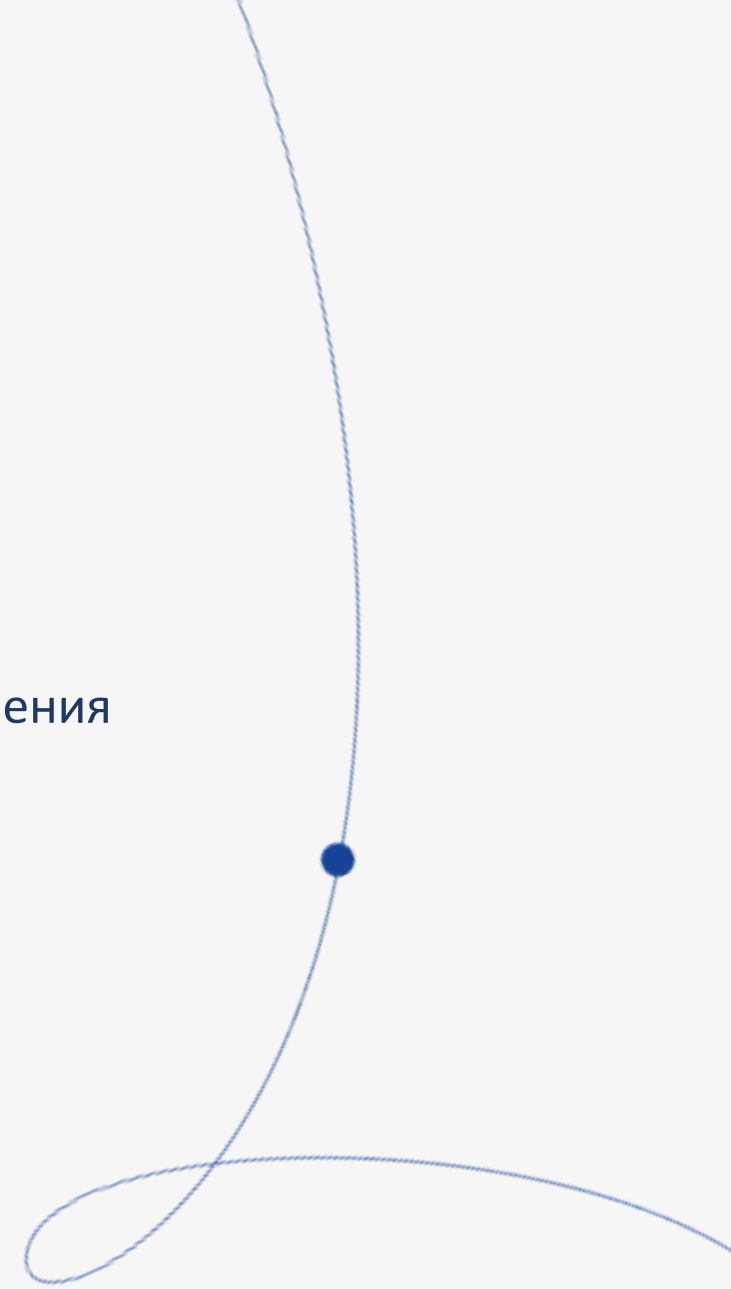
Сбор и верификация данных

Pickle: Замораживаем объекты

Сохраняет любой объект Python (даже сложную модель).

Очень быстро читается.

Внимание: Не открывайте чужие файлы `.pickle` (риск выполнения вредоносного кода!).



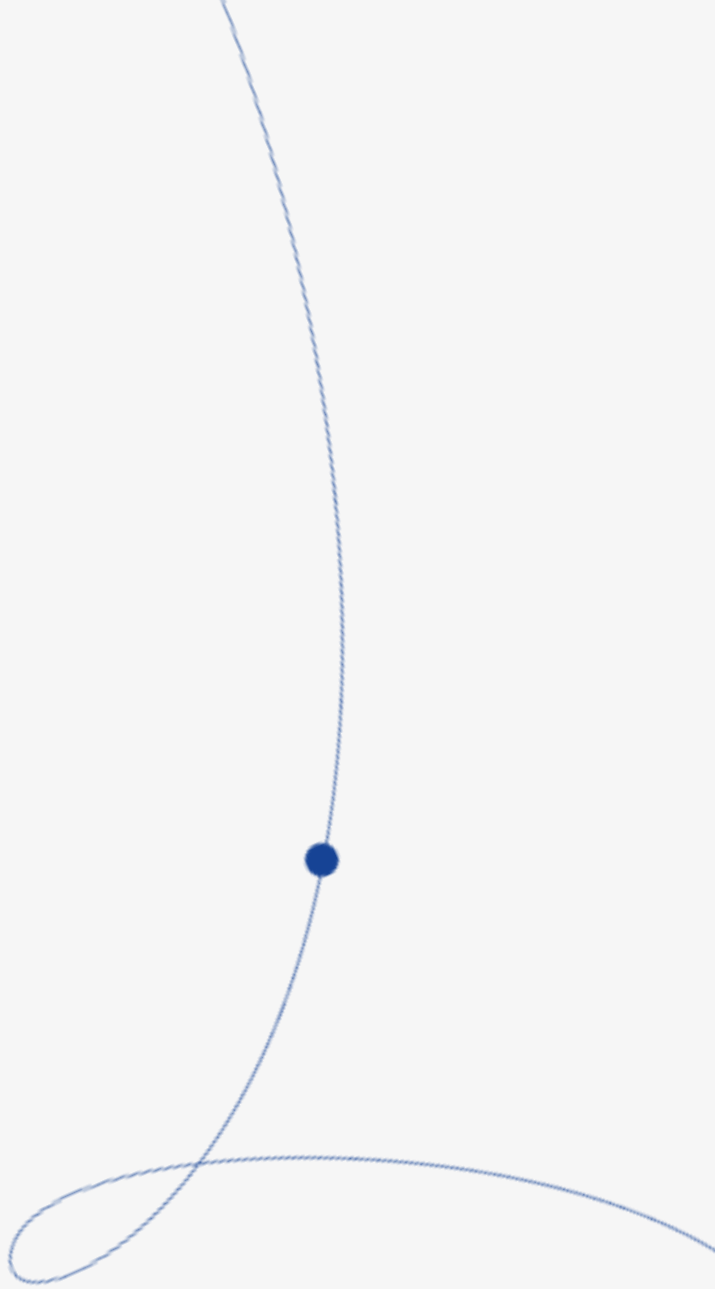
Сбор и верификация данных

Parquet: Колоночное хранение

Данные хранятся не по строкам, а по столбцам.

Сжатие до 10 раз эффективнее CSV.

Идеально для Pandas и больших таблиц.



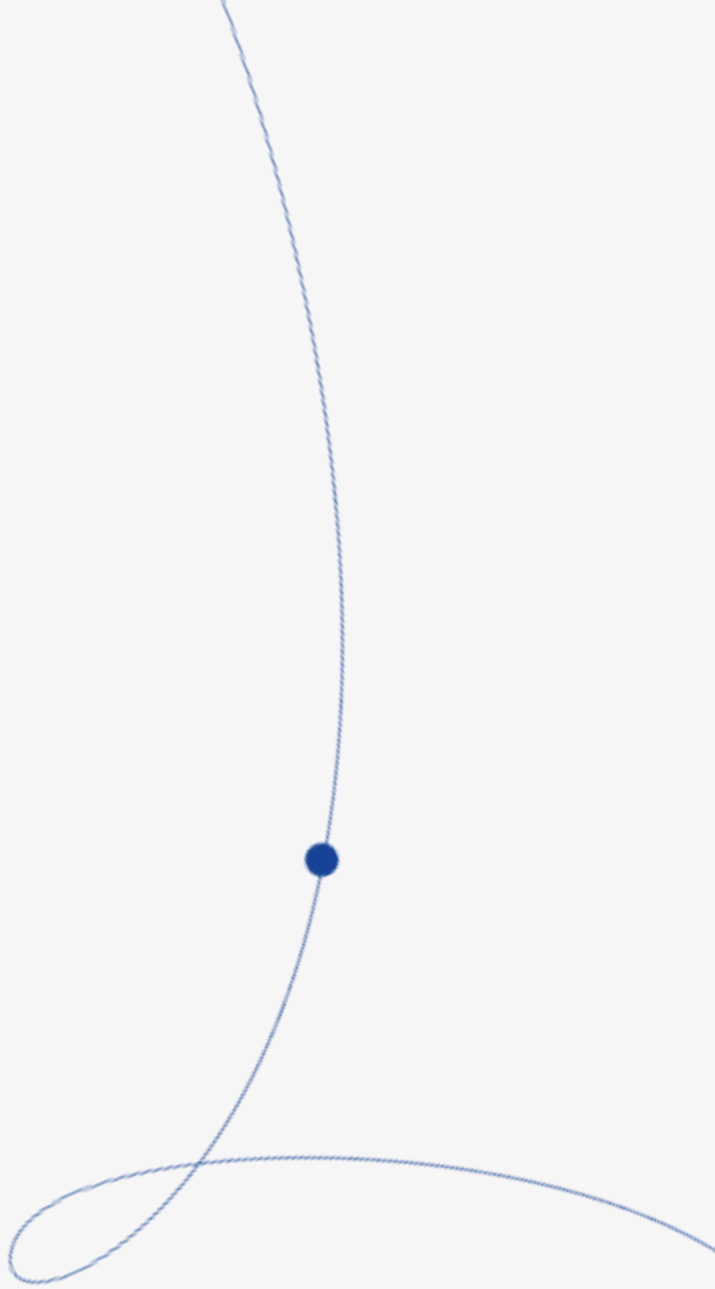
Сбор и верификация данных

XLSX — это не просто таблица

Это ZIP-архив с кучей XML-файлов внутри.

Огромный объем метаданных (шрифты, цвета, стили).

Медленная обработка программными методами.



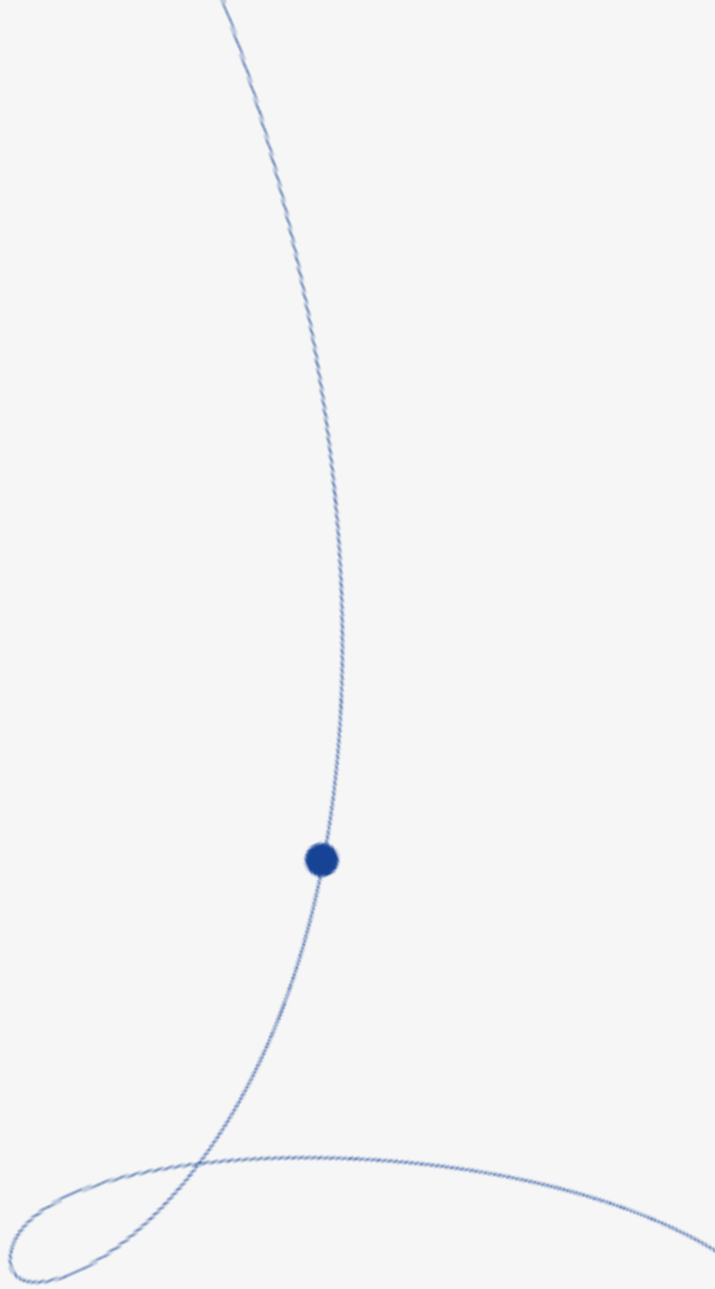
Сбор и верификация данных

Выбираем формат под задачу

Обмен данными: JSON.

Быстрый старт в ML: CSV.

Профессиональная работа: Parquet.



Сбор и верификация данных

Что такое кодировка (Encoding)?

Компьютер понимает только числа.

Кодировка — это таблица перевода:

Число ↔ Буква

Бит (01000001) → Число (65) → Буква ("A").

Сбор и верификация данных

ASCII: С чего всё начиналось

Всего 128 символов (7 бит).

Только английский алфавит, цифры и спецсимволы.

Для других языков места не осталось.

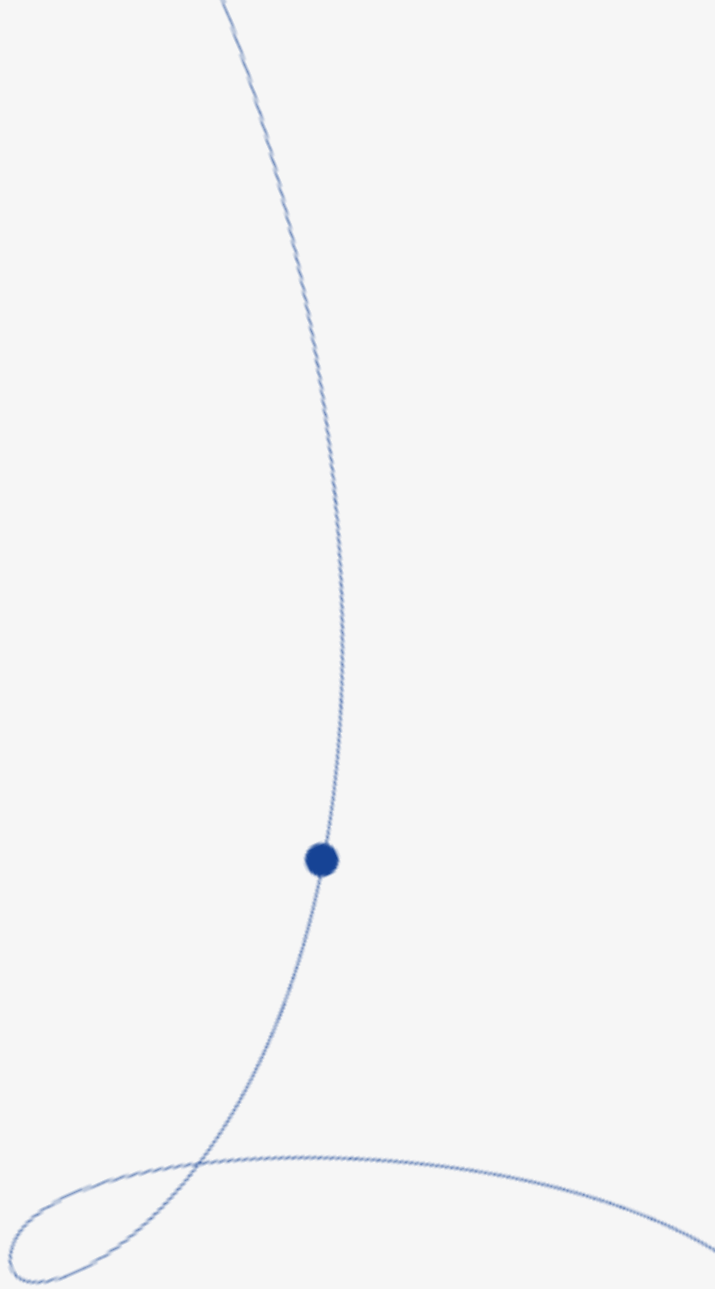
Сбор и верификация данных

Зоопарк региональных кодировок (CP1251, KOI8-R)

Каждая страна придумала свой стандарт для 8-го бита.

Windows использовала CP1251.

Linux использовал KOI8-R.



Сбор и верификация данных

Unicode и UTF-8: Один стандарт для всех

Вмещает все языки мира, эмодзи и даже мертвые языки.

Переменная длина (латиница — 1 байт, кириллица — 2 байта).

Стандарт де-факто в современном мире.

Сбор и верификация данных

Что такое «Кракозябры» (Mojibake)?

Почему мы видим ÐŸÑ€Ð¸Ð²ÐµÑ‚?

Проблема возникает, когда файл записан в одной кодировке, а открыт в другой.

Сбор и верификация данных

Рецепт выживания в Python

Всегда старайся сохранять в utf-8.

В Pandas используй параметр: `encoding='utf-8'` или `'cp1251'`.

Библиотека **`chardet`** — для автоматического определения кодировки.

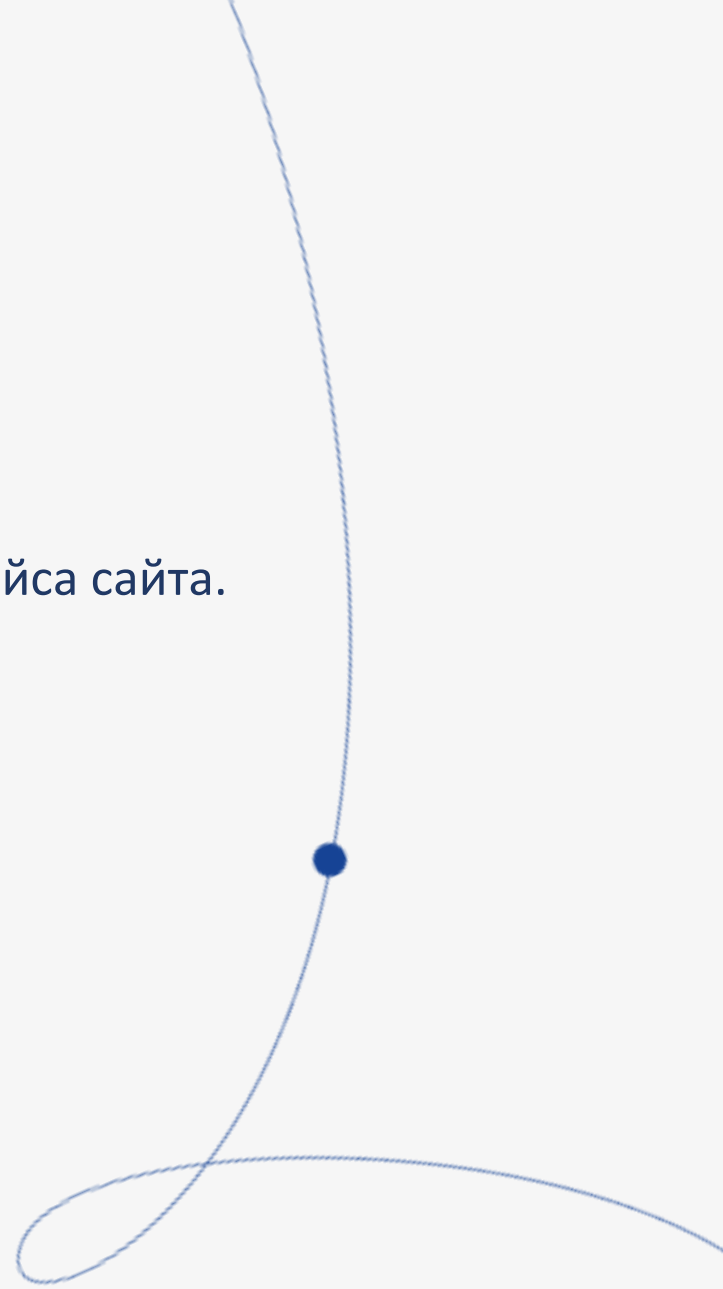
```
df = pd.read_csv('data.csv', encoding='utf-8')
```

Сбор и верификация данных

Как «добыть» данные из интернета?

API: Официальный, структурированный интерфейс.

Парсинг (**Scraping**): Извлечение данных прямо из визуального интерфейса сайта.



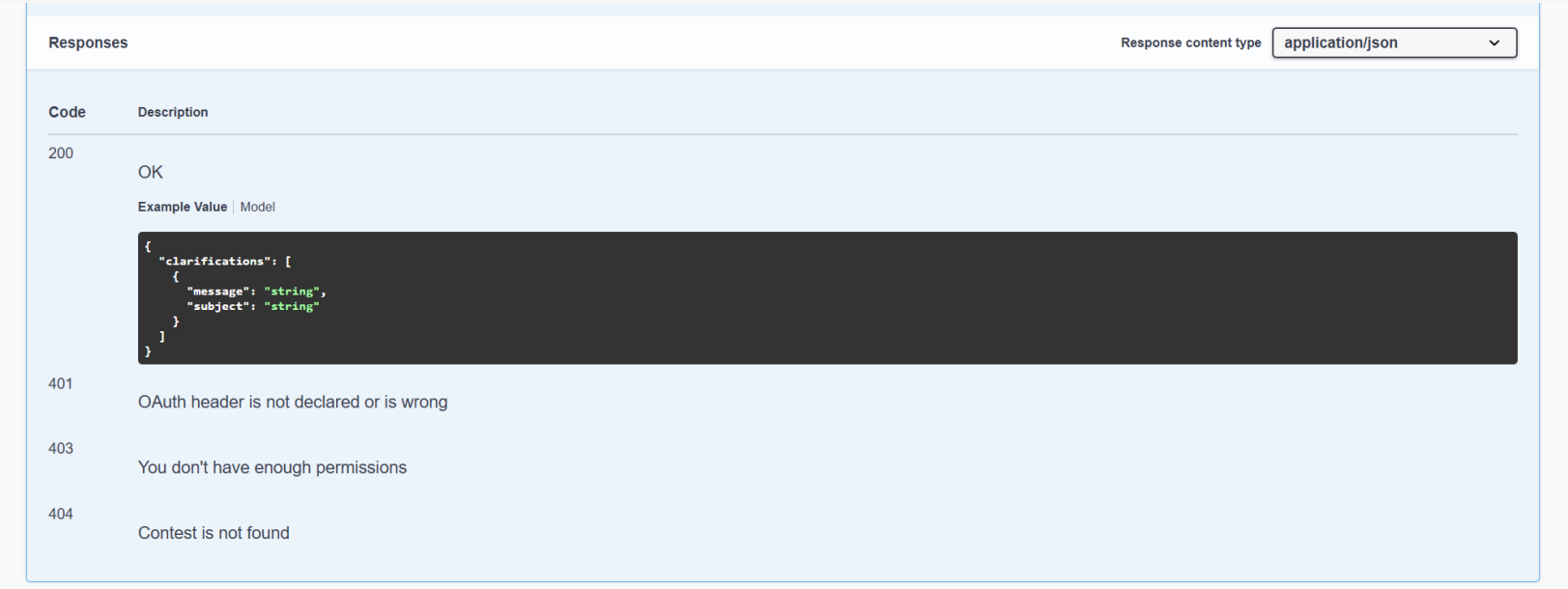
Сбор и верификация данных

API: Контракт между системами

Application Programming Interface.

Запрос (Request) → Ответ (Response).

Обычно в формате JSON.



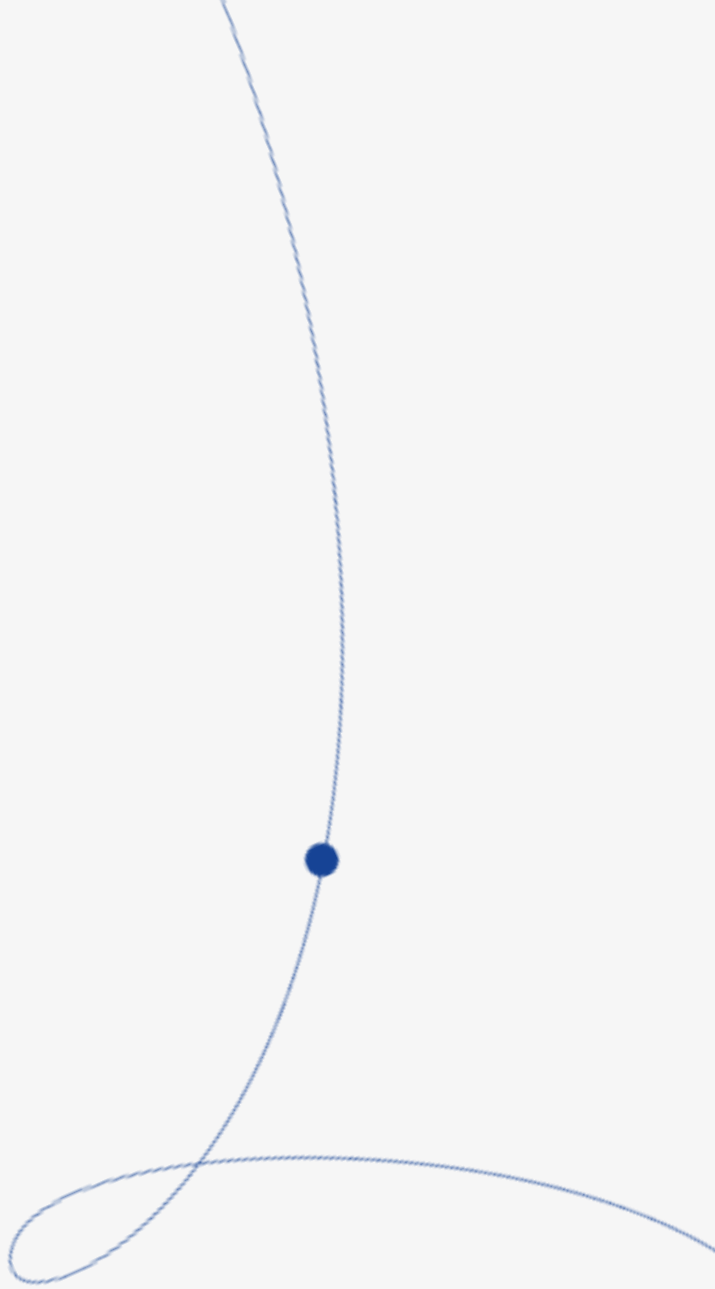
Сбор и верификация данных

Почему API — это лучший выбор?

Стабильность (формат редко меняется).

Скорость (не нужно скачивать картинки и стили сайта).

Легальность (вы используете разрешенный канал).



Сбор и верификация данных

Web Scraping — когда двери закрыты

Парсинг: Читаем HTML как компьютер

Программа имитирует браузер, открывает страницу и «выкусывает» из неё нужный текст по тегам.

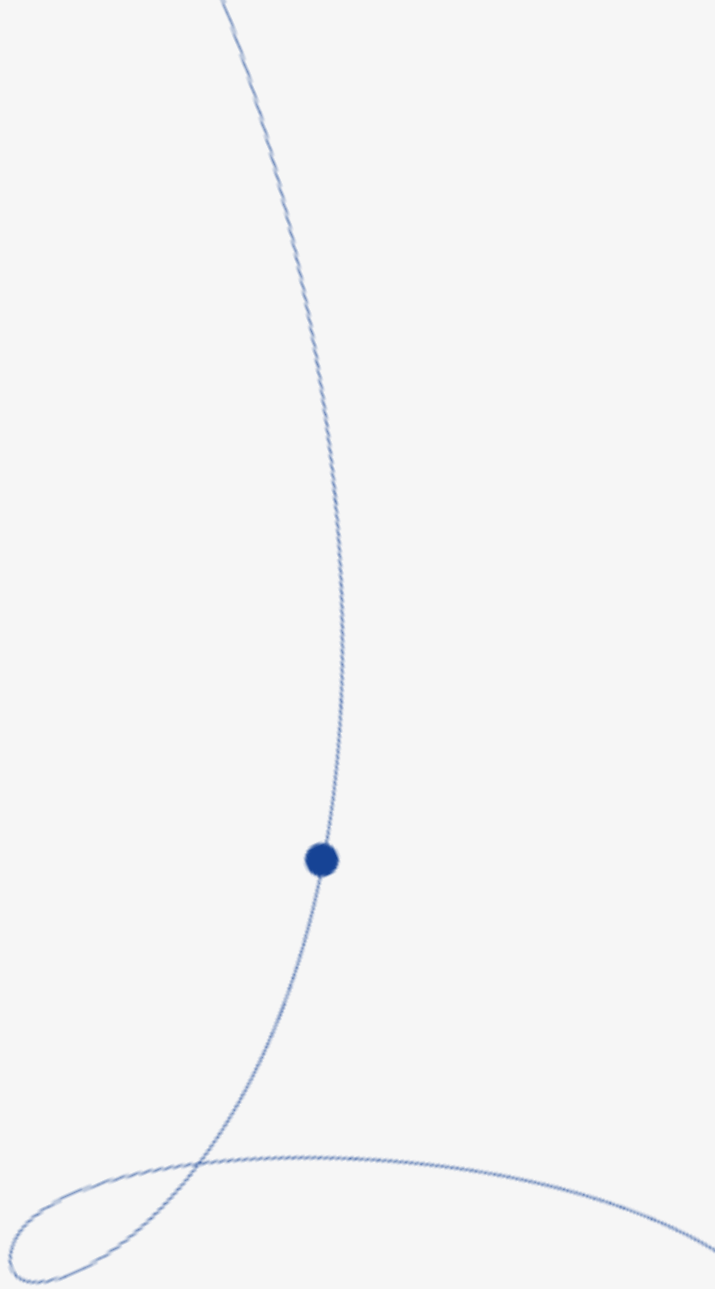
Сбор и верификация данных

Почему парсинг — это больно?

Сайт обновил дизайн = ваш код сломался.

Блокировки по IP (защита от ботов).

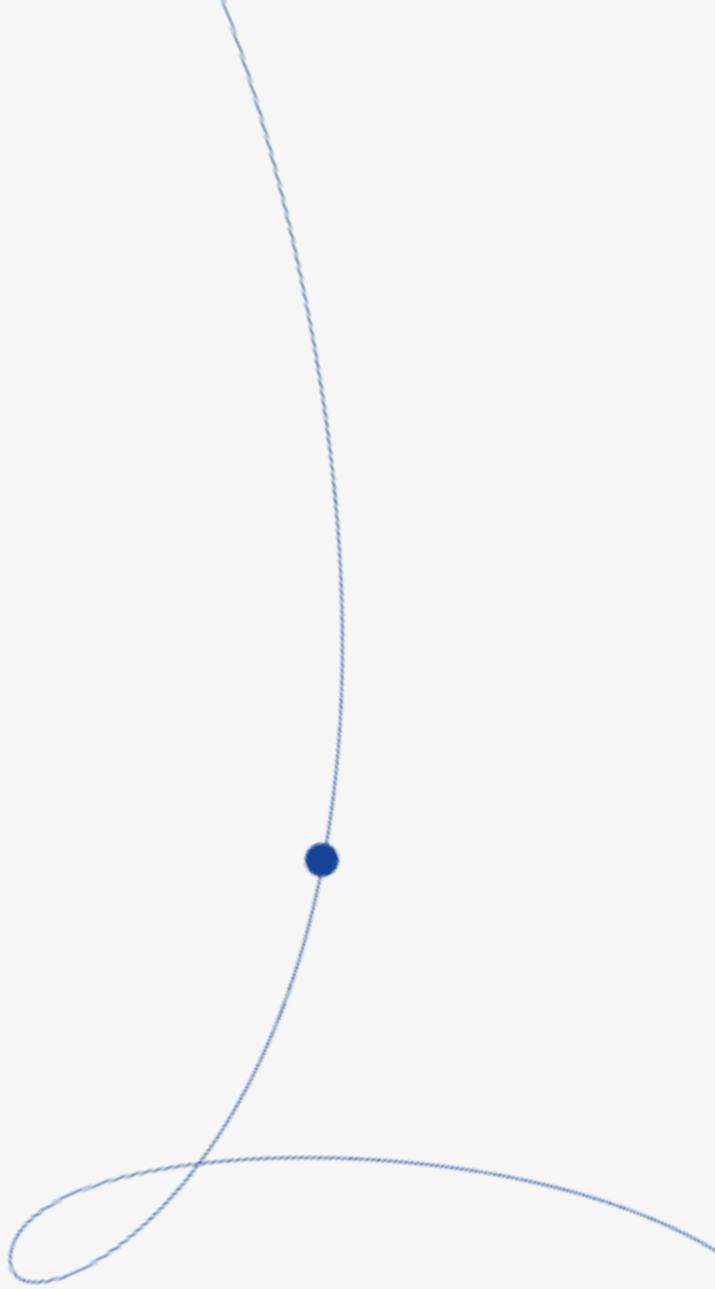
Этика и `robots.txt` (просьба сайта не парсить его).



Сбор и верификация данных

Сравнение: Что выбрать?

Если есть API — используй API.



Сбор и верификация данных

4 кита качества данных

Качество — это не отсутствие ошибок, а пригодность данных для решения вашей задачи.

Полнота, Точность, Согласованность, Актуальность

Сбор и верификация данных

1. Полнота: А все ли на месте?

Проблема: пропущенные значения (NaN / Null).

Если у 50% пользователей не указан возраст, можно ли считать средний возраст?

Сбор и верификация данных

2. Точность: Данные vs Реальность

Соответствуют ли цифры фактам?

Пример: Рост человека 3 метра. Технически — число, фактически — ошибка.

Сбор и верификация данных

3. Согласованность: Логика связей

Противоречия внутри одной записи.

Пол: Мужской, Беременность: Да.

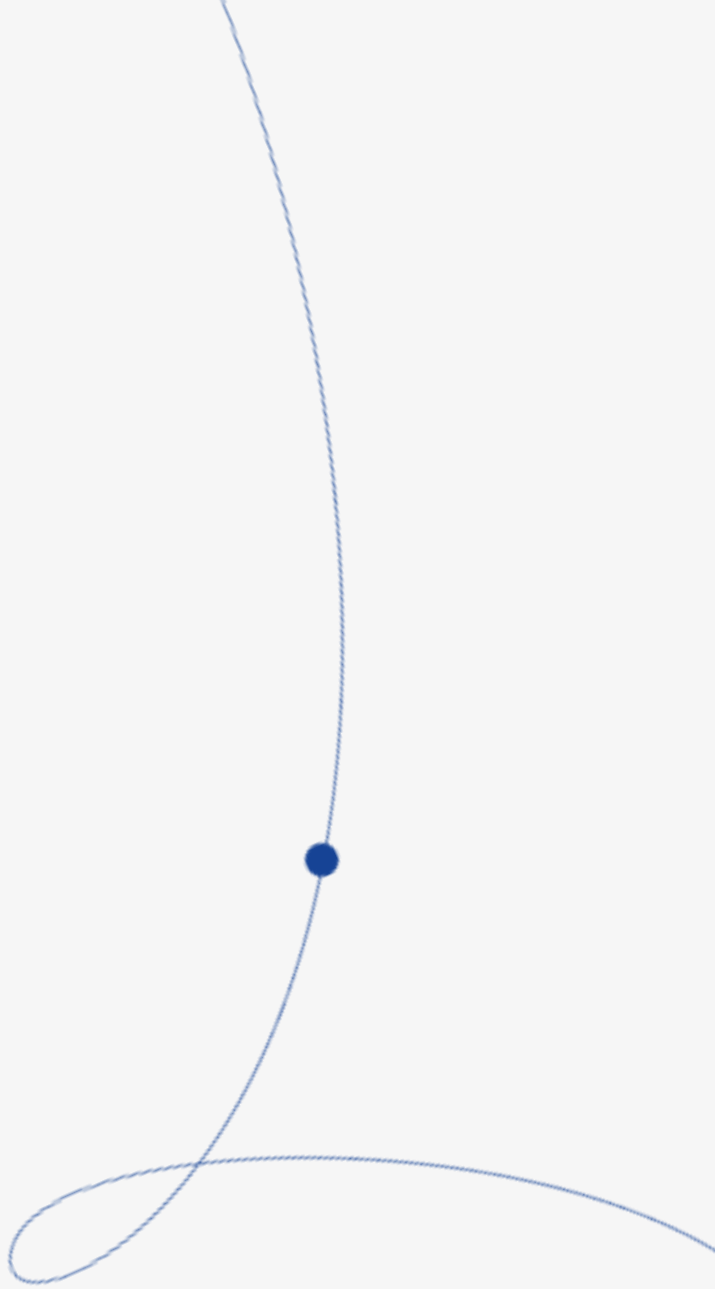
Город: Москва, Страна: Франция.

Сбор и верификация данных

4. Актуальность: Срок годности данных

Данные имеют свойство «протухать».

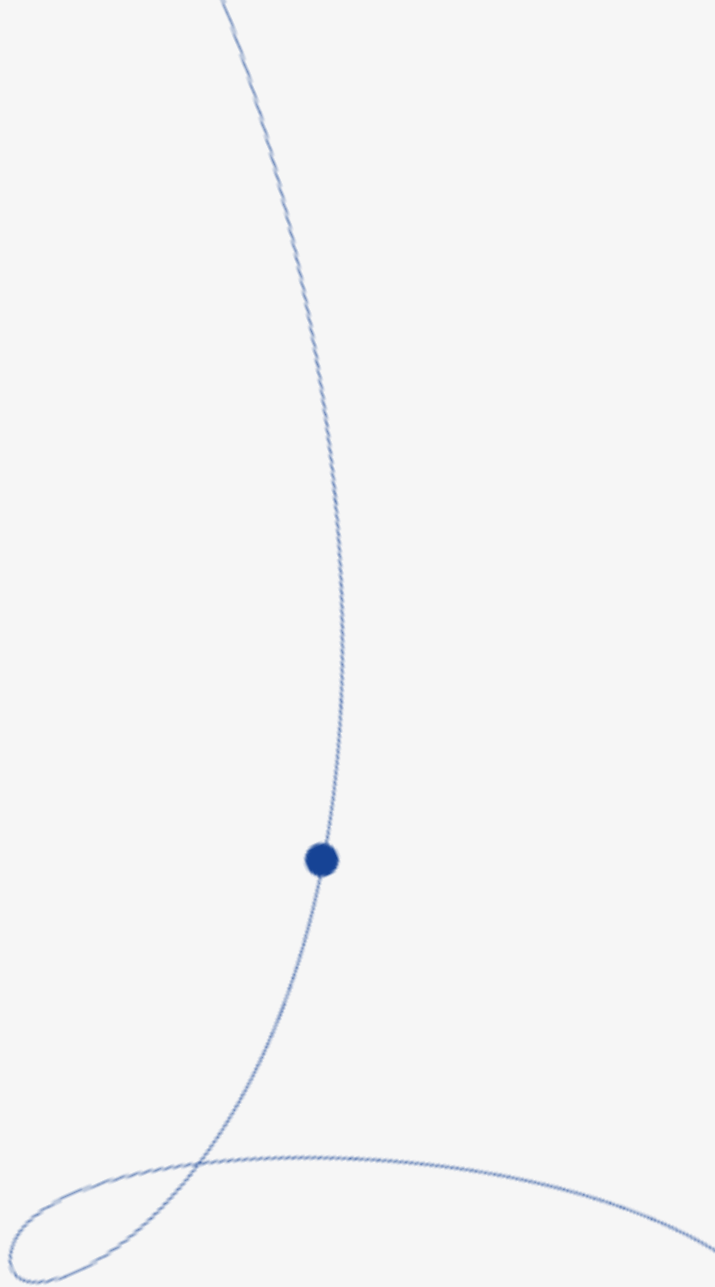
Курс доллара за вчера не поможет торговать сегодня.



Сбор и верификация данных

Технический чеклист

1. Сколько пропусков в каждой колонке?
2. Есть ли дубликаты строк?
3. Попадают ли значения в разумные границы? (мин/макс).



Сбор и верификация данных

Что такое верификация?

Верификация vs Валидация

Валидация: Правильные ли данные мы собрали? (бизнес-логика).

Верификация: Правильно ли мы их собрали технически? (отсутствие багов сбора).

Сбор и верификация данных

Два уровня ошибок

Синтаксис: Число записано как текст ("100"), дата в формате 12/31/2023 вместо 2023-12-31.

Семантика: Отрицательная цена, возраст 200 лет, город «Париж» в колонке «Фамилия».

Сбор и верификация данных

Аномалии: Ошибка или инсайт?

Выброс — это значение, которое резко выделяется из общей массы.

Может быть ошибкой датчика, а может — признаком мошенничества (фрод).

Сбор и верификация данных

Легендарные ошибки в данных

Когда верификация спасает миллионы денег

Mars Climate Orbiter (1999).

Причина: Одна команда считала в ньютонах, другая — в фунт-силах.

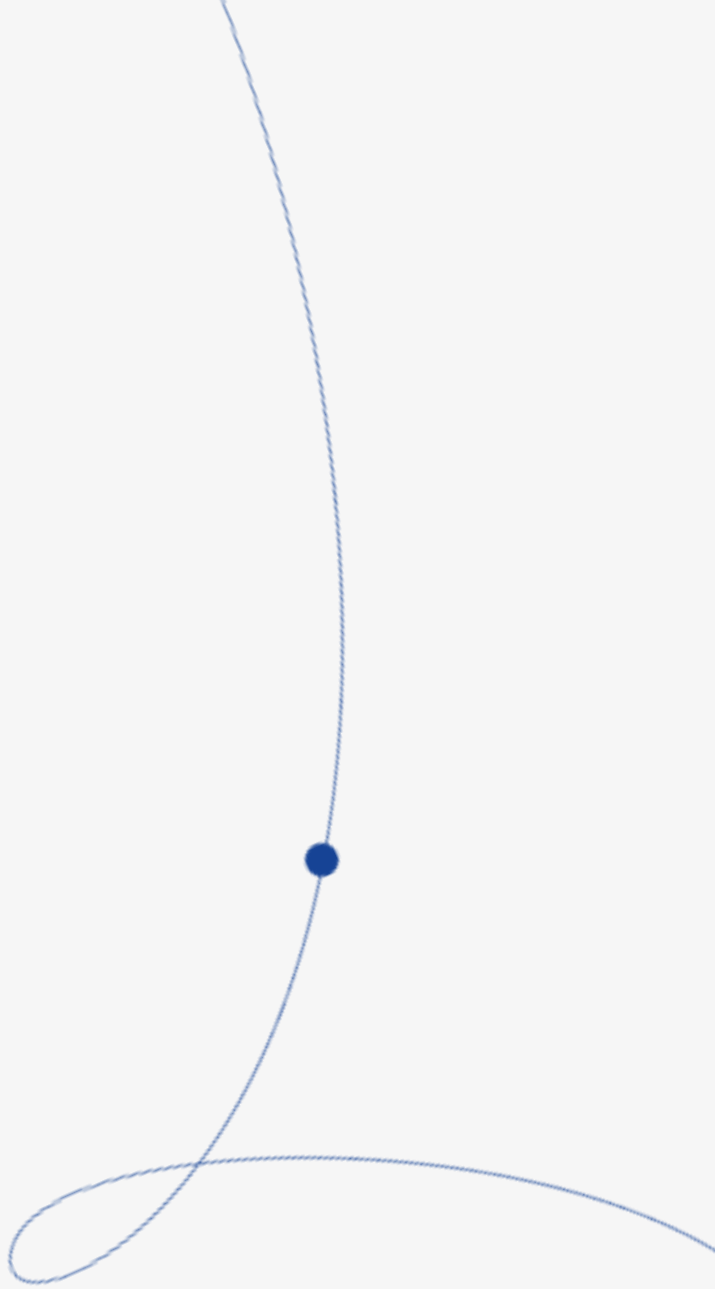
Итог: Потеря аппарата за $125 млн.

Сбор и верификация данных

Пустота в таблице (NaN)

Три стратегии:

1. Игнорировать (удалить строку).
2. Заполнить (средним, медианой, константой).
3. Исследовать (почему данных нет?).

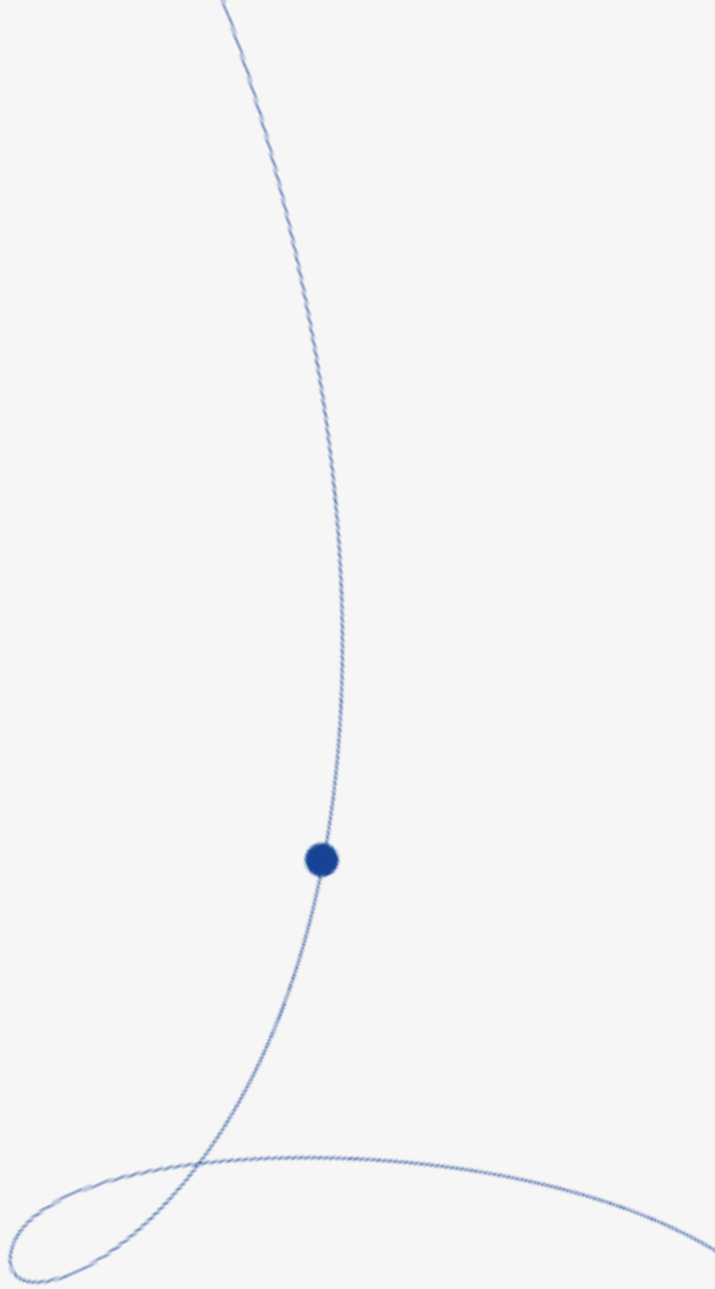


Сбор и верификация данных

Что мы пройдем за семестр?

Краткий список модулей:

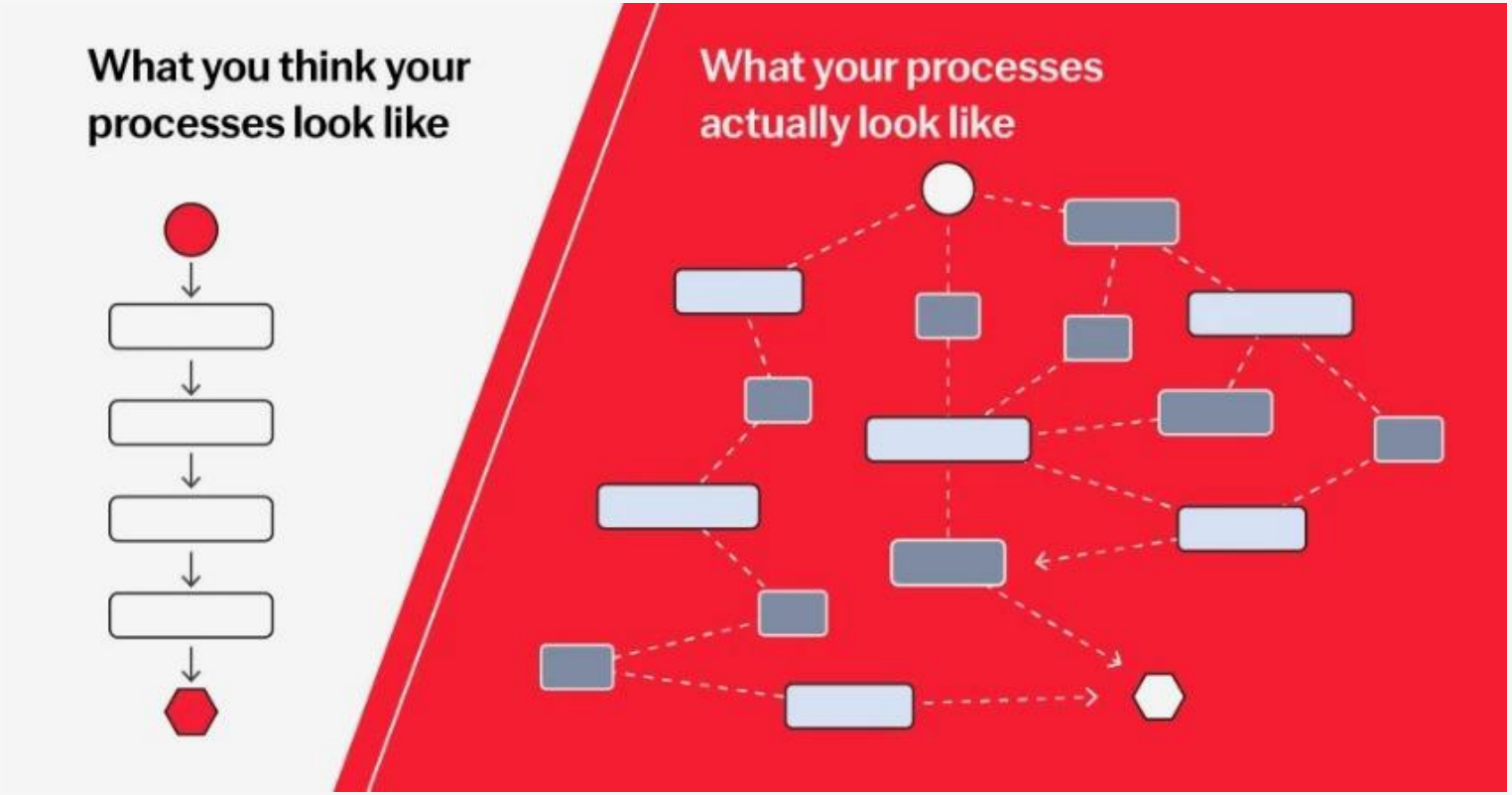
1. Сбор (Парсинг/API).
2. Чистка и Pandas.
3. Статистика.
4. Визуализация и BI.



Сбор и верификация данных

Финал

Данные — это новая нефть. Но только если они очищены.



Уральский федеральный университет
имени первого Президента России Б.Н.Ельцина

Спасибо за внимание!

Курс «Сбор и верификация данных»
Поляков Станислав Олегович