

Инструменты, источники и подготовка данных

Андрей Макеев

Бизнес-архитектор в Комус

Ex-ведущий аналитик в Glowbyte Consulting





Андрей Макеев

Бизнес-архитектор в **Комус**
Ex-ведущий аналитик в **Glowbyte Consulting**

Аккаунты в соцсетях



fb.com/andmkv



Содержание

1

Зачем визуализатору техника

2

Какие инструменты бывают и для каких задач используются

3

Примеры работы в инструментах



Задача: разобраться в том, какие инструменты нужно знать и для чего их применять



**Визуализация
данных**



**Творчество
или
набор инструментов
и навыков?**



Не только рисование

1

Собрать
данные

в том объёме, который
необходим для задачи,
и с должным качеством

2

Увидеть скрытую
в них информацию

объединить данные и найти
в них инсайты, т. о.
сформировав информацию,
или **знания**

3

Транслировать
информацию

правильно донести до
конечного потребителя —
например, через
визуализацию



Три класса инструментов

1

Загрузка и
подготовка данных



2

Исследование
данных



3

Визуализация
данных





Задача на сегодня

**Рассмотреть инструменты всех
трёх классов**

Некоторые — на примерах.

Владеть всеми не обязательно,
но знать о них нужно.



**Но прежде
важный вопрос:**

**где
брать
данные?**



Откуда берутся данные

Внутренние источники

- Корпоративные информационные системы
- Данные оборудования
- Данные соцопросов
- Внутренняя экспертиза компании

Внешние источники

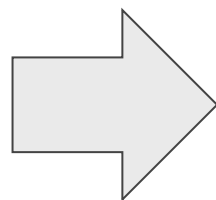
- Внешние информационные системы (веб-аналитика и т. д.)
- Порталы открытых данных
- Веб-скрейпинг
- Провайдеры платных данных и т. д.



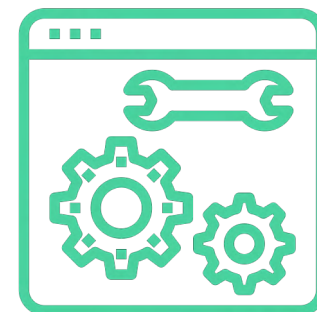
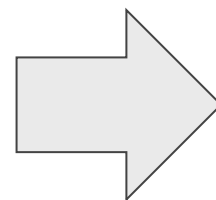
Внутренние источники



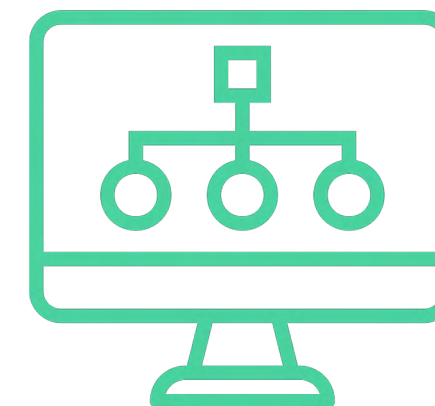
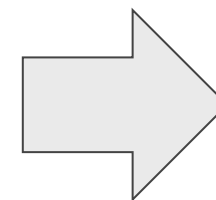
Корпоративные
информационные
системы



База данных,
хранилище
данных



Подготовка
данных



Анализ и
визуализация
данных



Внешние источники

- Открытые данные могут создать хороший **контекст** для вашего процесса анализа
- Это **данные об окружающем мире**, статистика, собираемая правительствами и частными компаниями
- Публикуется на **порталах открытых данных**



Где искать?

- Google Dataset Search — поисковик по открытым данным
<https://datasetsearch.research.google.com>
- World Bank Open Data — данные о мировой экономике <https://data.worldbank.org/>
- Росстат <https://rosstat.gov.ru/opendata/>
- Портал открытых данных
<https://data.gov.ru/>
- Портал открытых данных Правительства Москвы <https://data.mos.ru/>



Как пользоваться?

- Google Dataset Search — поисковик по открытым данным
<https://datasetsearch.research.google.com>
- World Bank Open Data — данные о мировой экономике <https://data.worldbank.org/>
- Росстат <https://rosstat.gov.ru/opendata/>
- Портал открытых данных
<https://data.gov.ru/>
- Портал открытых данных Правительства Москвы <https://data.mos.ru/>



Порталы открытых данных

Доступ через API

— для автоматизированной загрузки данных

API Портала открытых данных

Получить ключ

Документация

Условия использования

Вход

English version



Надежность каждого символа

Получите доступ к современной платформе для работы с актуальными данными Правительства Москвы.

Получить доступ »

С чего начать?

Для работы с API Вам потребуется уникальный ключ. Чтобы получить его — пройдите несложную процедуру регистрации или воспользуйтесь существующим аккаунтом одной из популярных социальных сетей.

Получить ключ »

Документация

Изучите все возможности платформы. Описание ресурсов API, входные параметры, варианты формирования запросов, а также примеры использования подробно описаны в соответствующем разделе.





Узнать больше »



Порталы открытых данных

Поиск и загрузка данных в виде файлов
— для ad hoc-анализа данных

Включает описание структуры данных и сам датасет.

Поиск по названию		Всего 660 наборов	Фильтр по наборам данных	Экспорт реестра
18.	 Автовокзалы и автостанции Москвы	?	Экспорт	Паспорт
19.	 Автозаправочные станции, реализующие топливо, несоответствующее установленным экологическим требованиям	?	Паспорт	
20.	 Автозаправочные станции, реализующие топливо, соответствующее установленным экологическим требованиям	?	Паспорт	
21.	 Велосипедные парковки	?	Экспорт	Паспорт



Внешние информационные системы



Системы веб-аналитики



Провайдеры внешних данных

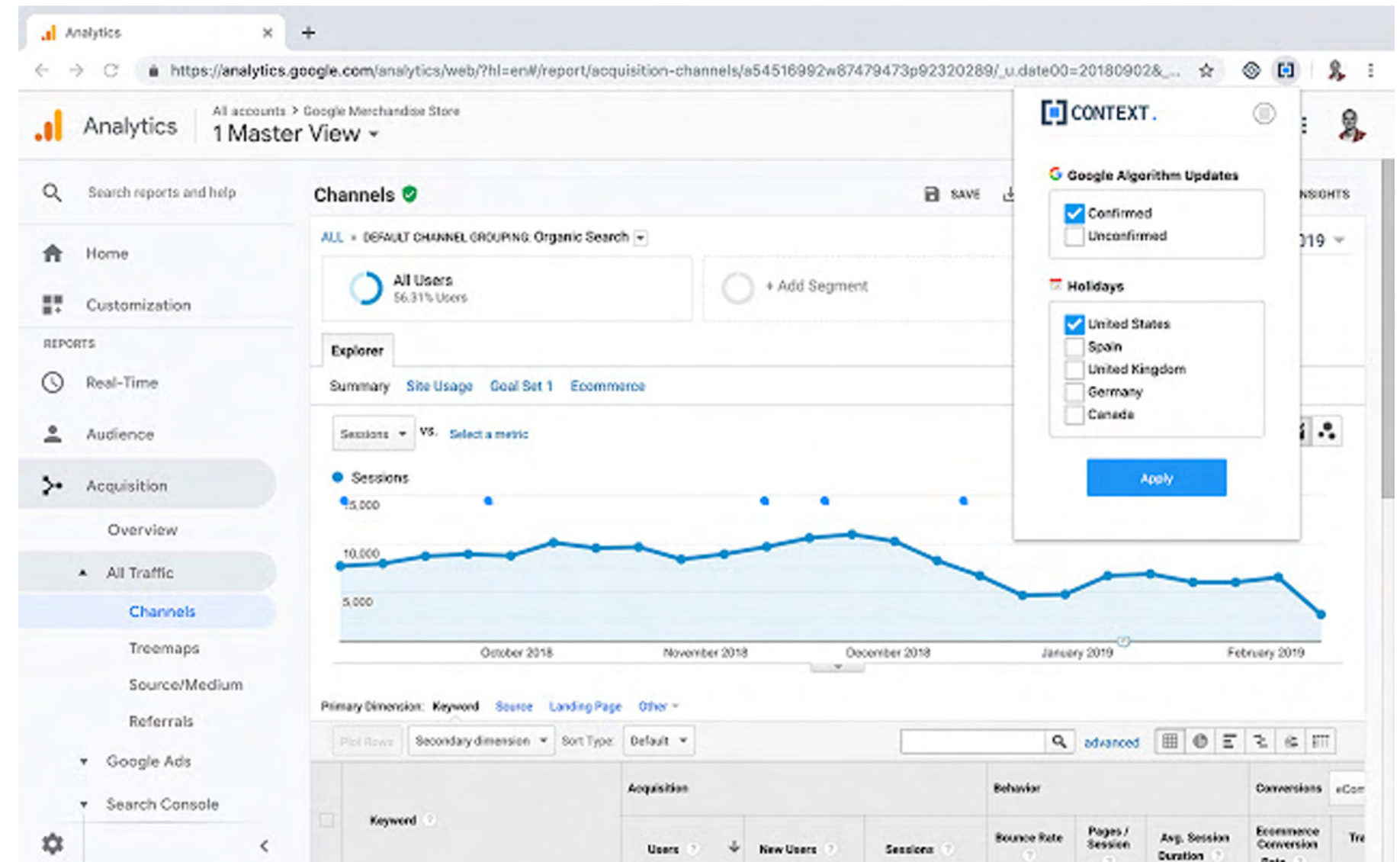


SaaS CRM-системы

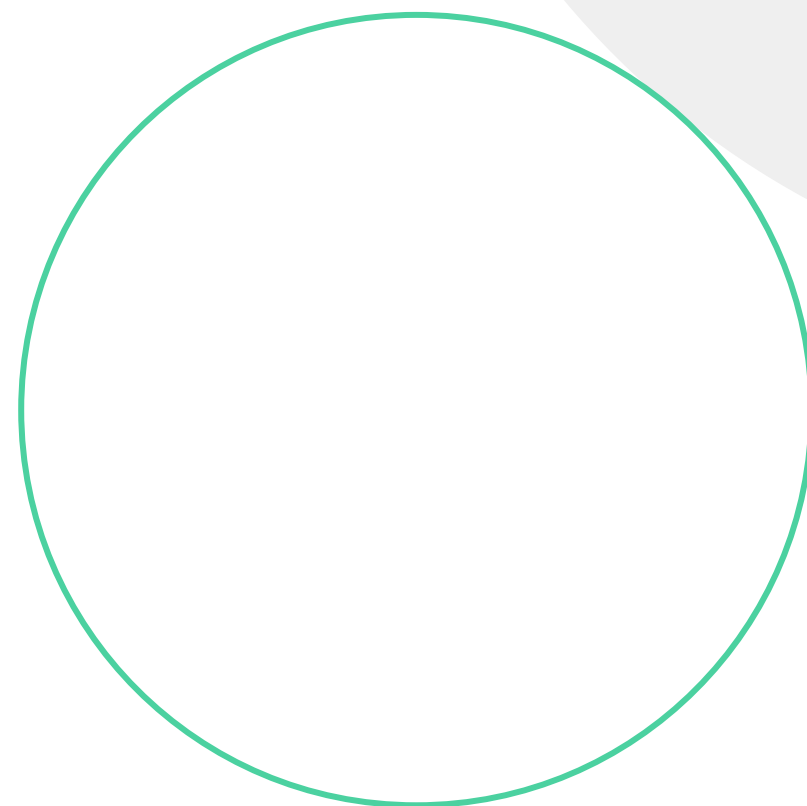


и другое

Получать данные так же: либо скачивая датасеты в формате CSV или Excel, либо подключаясь по API.



Подготовка данных



Задачи

1

Загрузка
данных



2

Очистка
данных

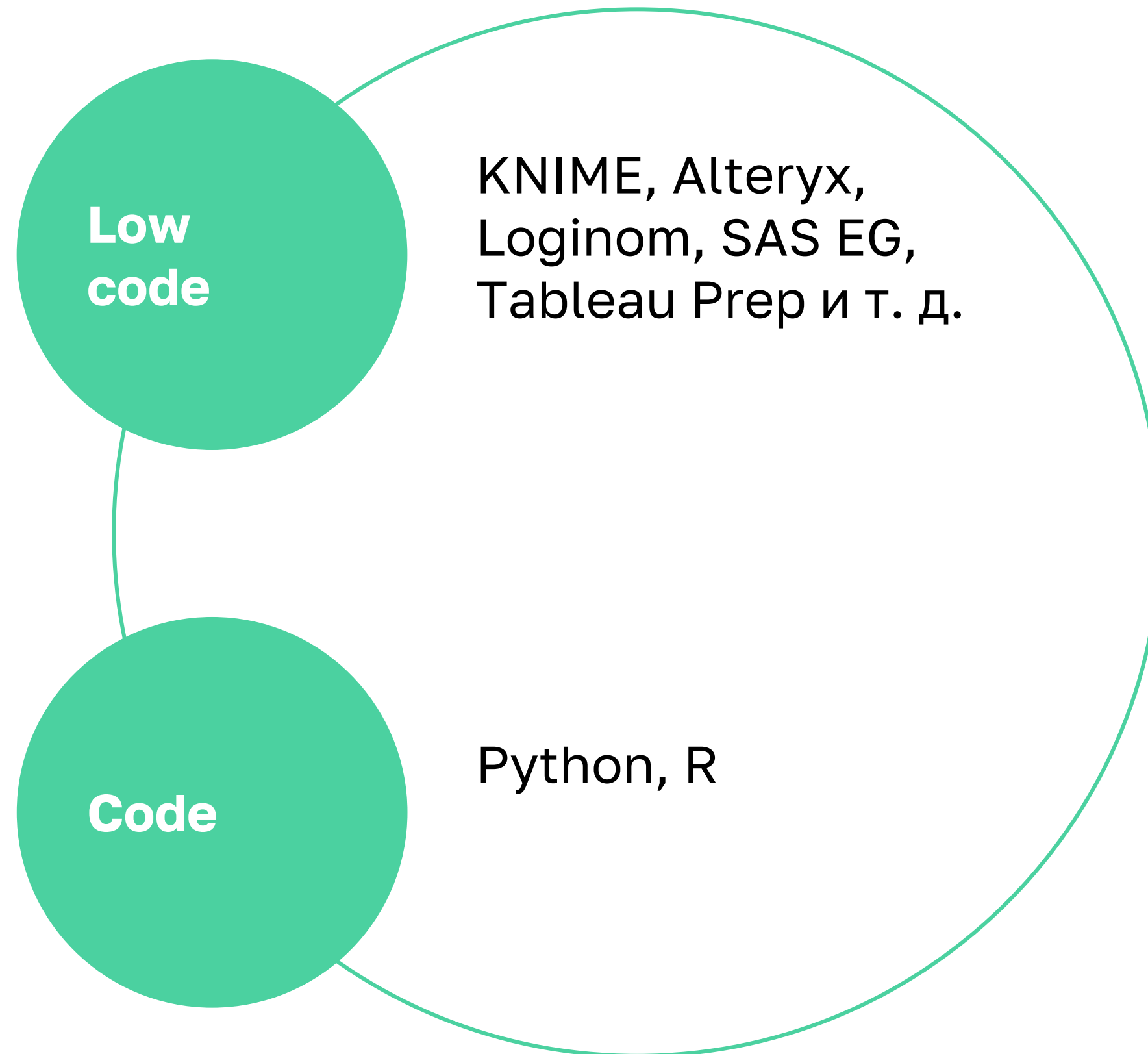


3

Объединение
данных



Инструменты подготовки данных



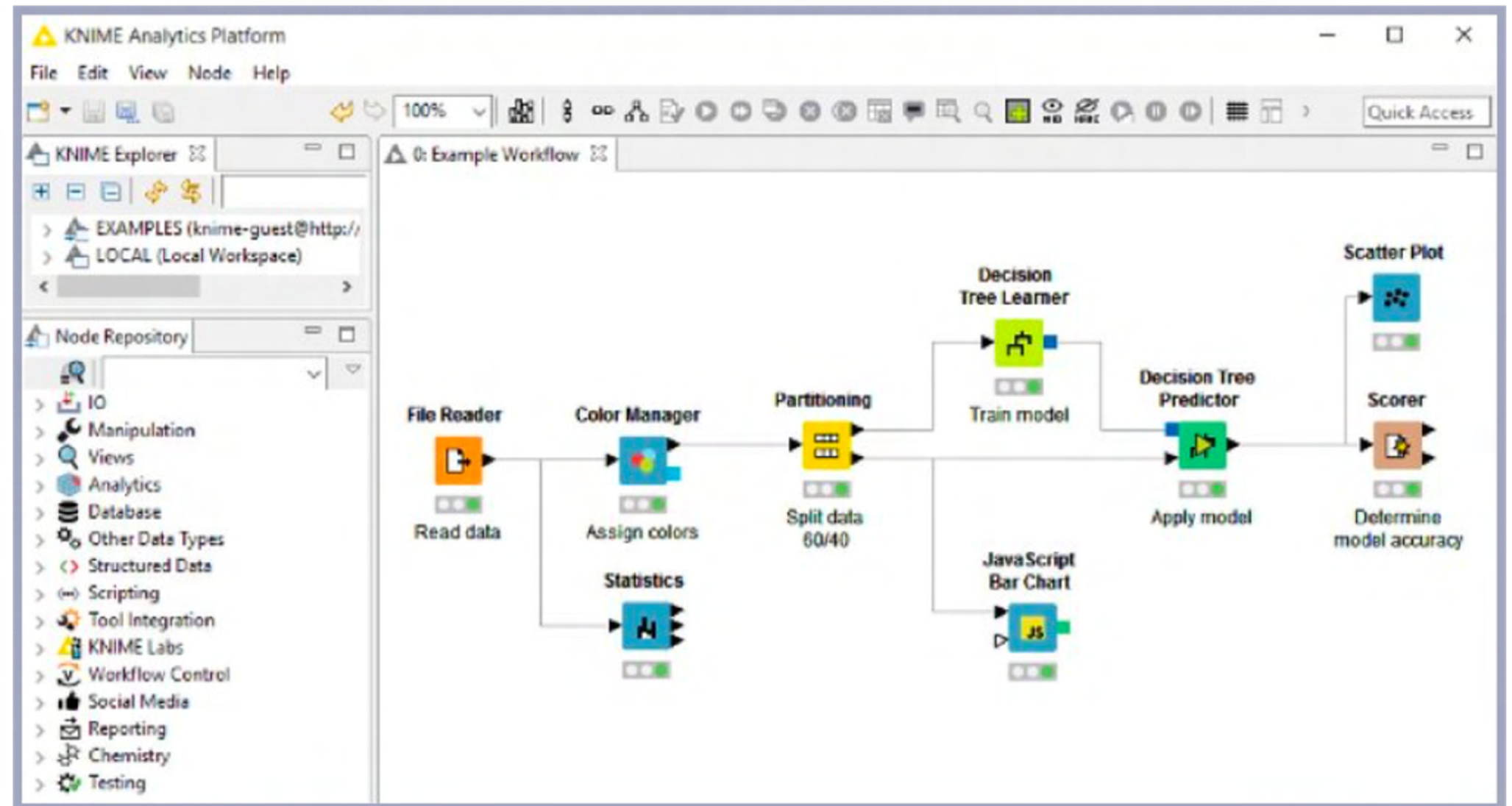
Инструменты low code

Конструирование пошагового процесса загрузки и трансформации данных в виде визуальных блоков.

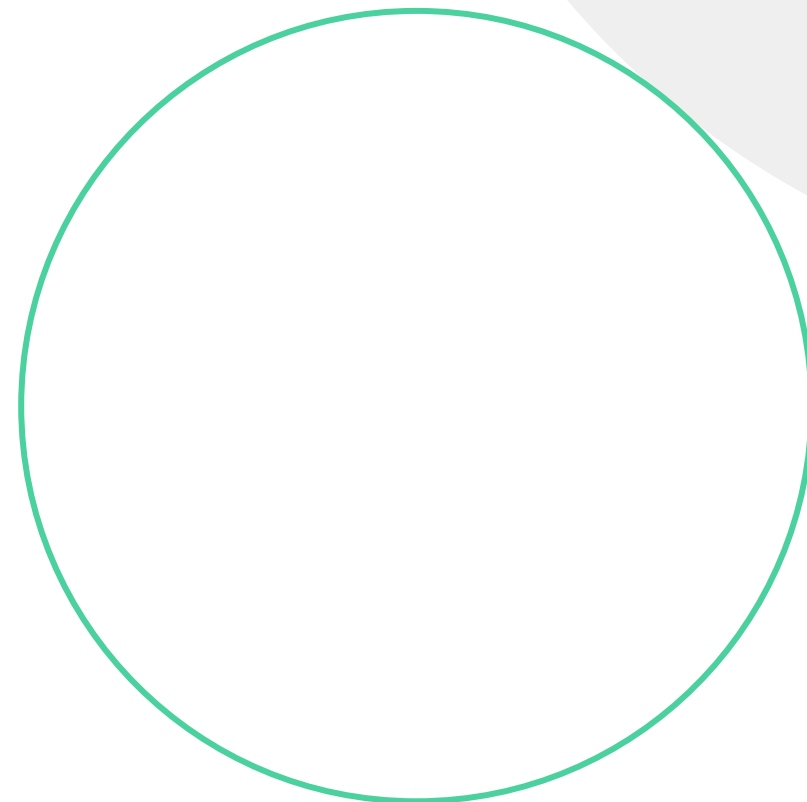
Программирование не требуется или требуется в редких случаях.

Возможно рассылать результат по почте, ставить процесс на расписание.

Направлены на быстрое освоение и автоматизацию процессов.



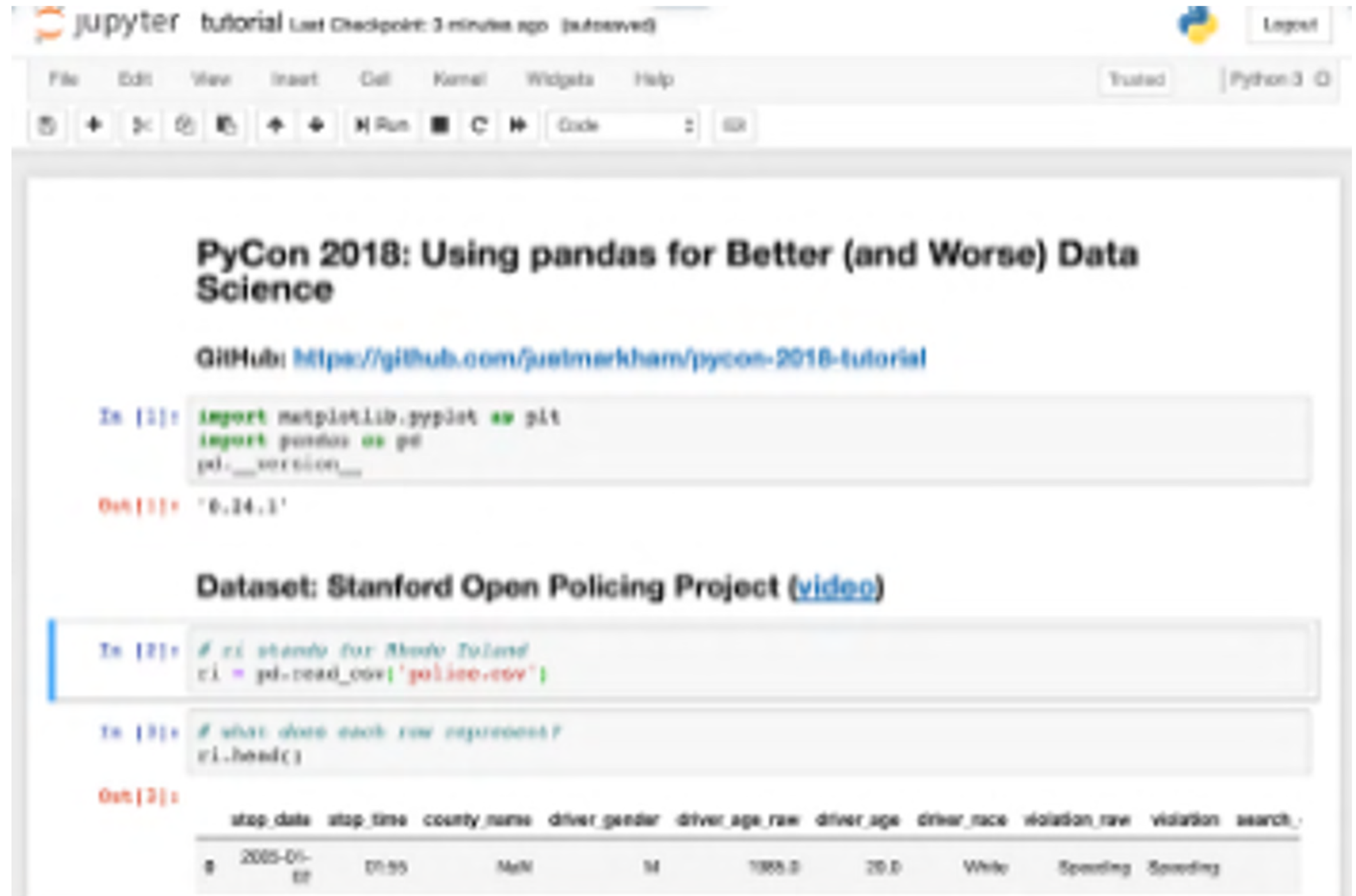
Инструменты low code – пример KNIME



Code-инструменты

Исследование данных
и автоматизация процессов
при помощи
программирования
на языке R или Python.

**Направлены на продвинутых
аналитиков, им свойственны
более широкие возможности
и гибкость.**



jupyter tutorial Last Checkpoint: 3 minutes ago (autosaved) Python 3

File Edit View Insert Cell Kernel Widgets Help

PyCon 2018: Using pandas for Better (and Worse) Data Science

GitHub: <https://github.com/justmarkham/pycon-2018-tutorial>

```
In [1]: import matplotlib.pyplot as plt
import pandas as pd
pd.__version__
```

```
Out[1]: '0.14.1'
```

Dataset: Stanford Open Policing Project ([video](#))

```
In [2]: # ci stands for Rhode Island
ci = pd.read_csv('police.csv')
```

```
In [3]: # what does each row represent?
ci.head()
```

```
Out[3]:
```

	stop_date	stop_time	county_name	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_
0	2005-06-01	07:55	Maple	M	1985.0	20.0	White	Speeding	Speeding	





Jupyter












- Open source-среда для разработки на Python (в основном)
- Исследование, подготовка и визуализация данных в одном месте
- Формат ноутбуков — деление скриптов на последовательные действия
- Можно делать автоматизацию (ставить на расписание)



Jupyter — пример ноутбука

 jupyter tutorial Last Checkpoint: 3 minutes ago (autosaved)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

       Run    Code 

PyCon 2018: Using pandas for Better (and Worse) Data Science

GitHub: <https://github.com/justmarkham/pycon-2018-tutorial>

```
In [1]: import matplotlib.pyplot as plt
import pandas as pd
pd.__version__
```

```
Out[1]: '0.24.1'
```

Dataset: Stanford Open Policing Project ([video](#))

```
In [2]: # ri stands for Rhode Island
ri = pd.read_csv('police.csv')
```

```
In [3]: # what does each row represent?
ri.head()
```

```
Out[3]:
```

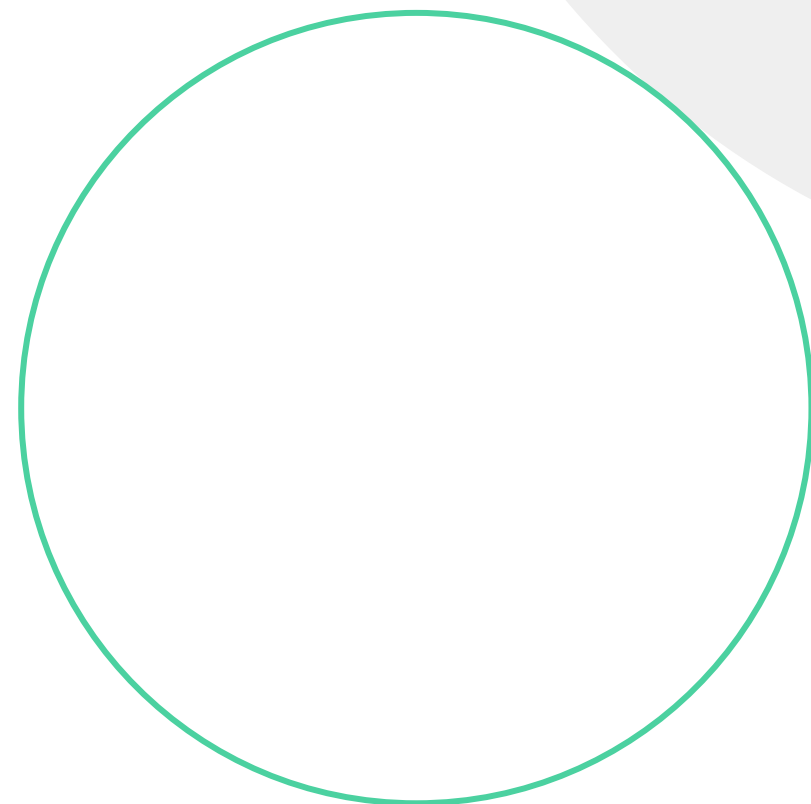
	stop_date	stop_time	county_name	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_
0	2005-01-02	01:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	
1	2005-01-18	08:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	
2	2005-01-23	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	
3	2005-02-20	17:15	NaN	M	1986.0	19.0	White	Call for Service	Other	

Практика

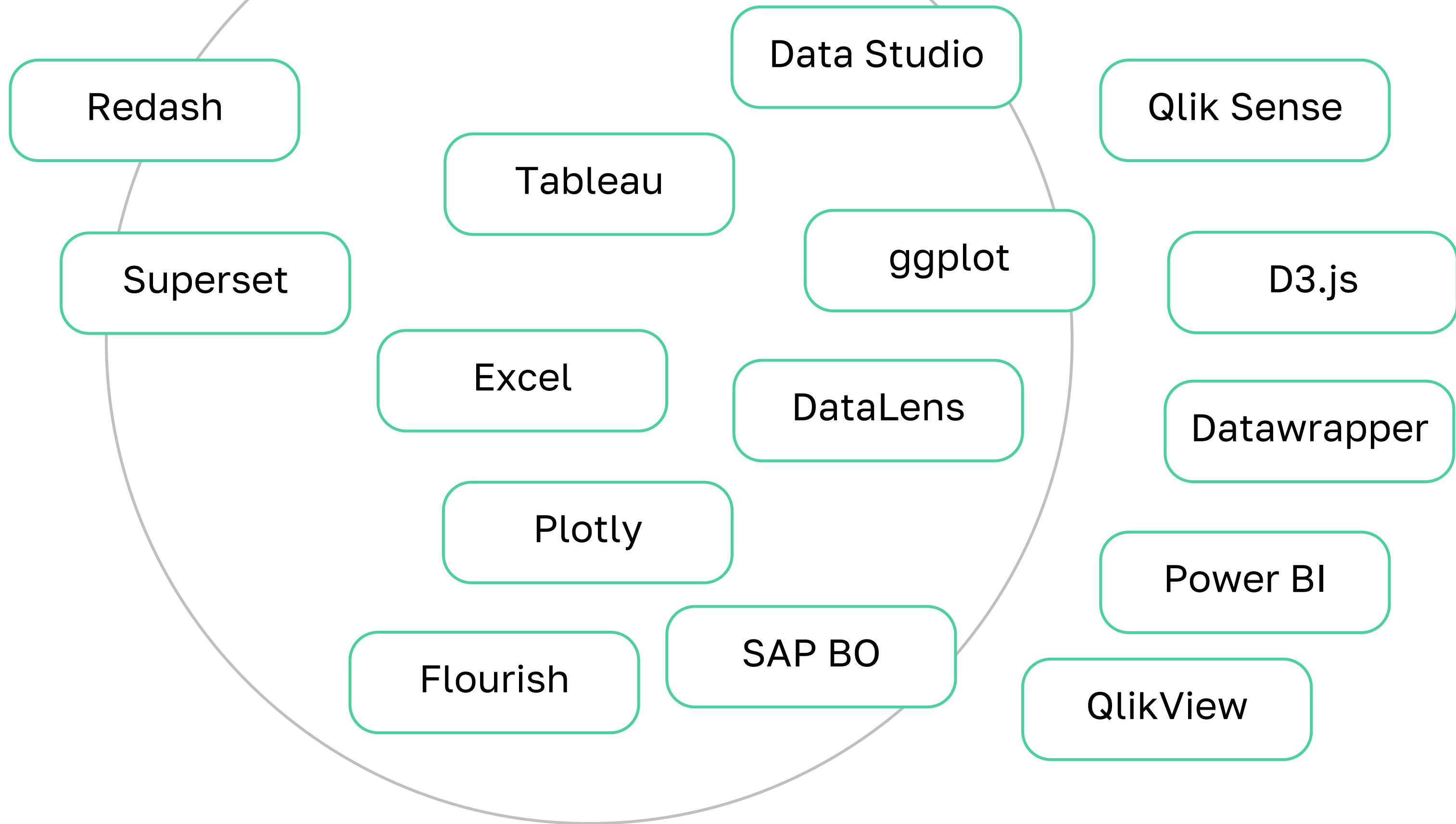
Загружаем данные с портала
открытых данных
Правительства Москвы



Инструменты визуализации



Какой выбрать?



Какой выбрать?



Выбираем
инструмент,
ИСХОДЯ
из задачи



Сделать визуализацию для презентации



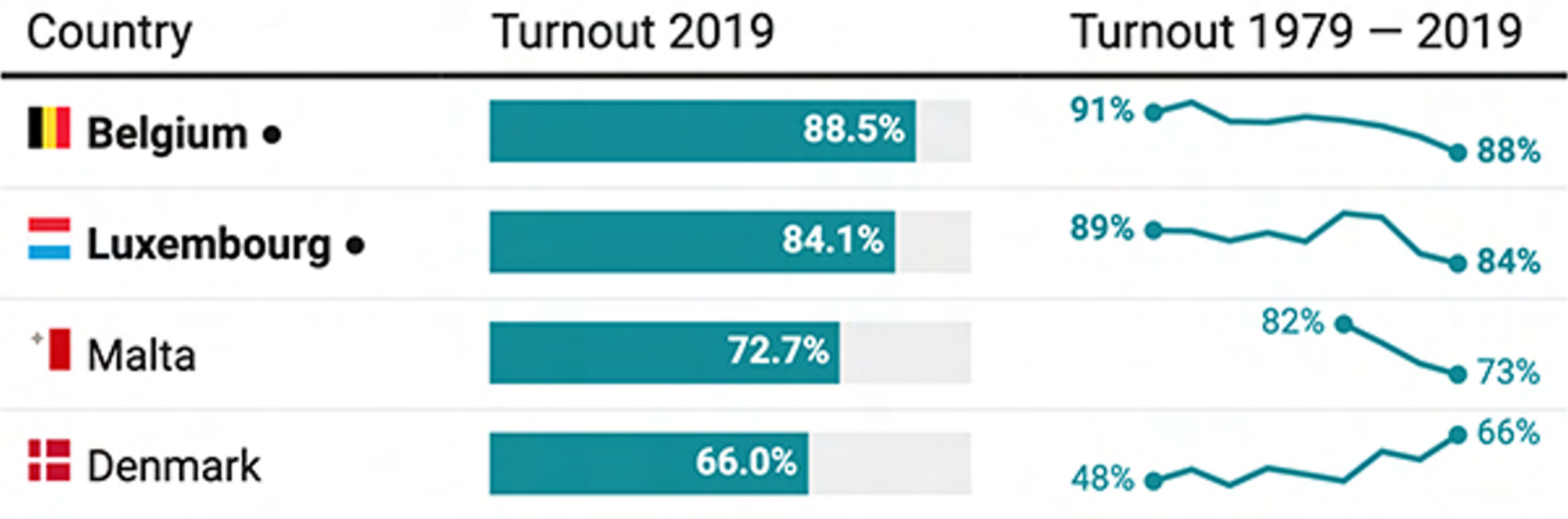
Помните: **в них нет ничего плохого**



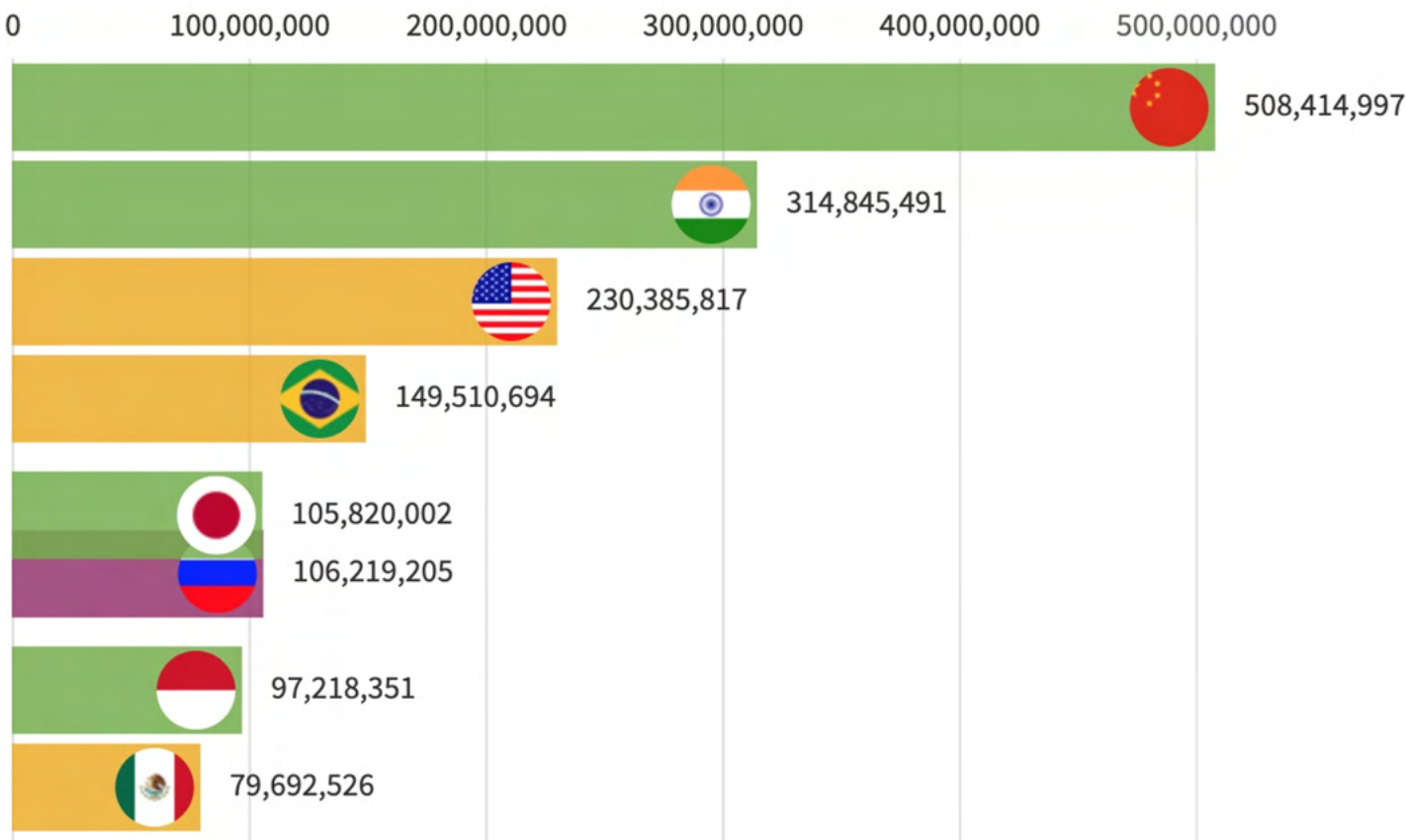
Сделать визуализацию для презентации или сайта

- Countries with compulsory voting (in 2019)
- Countries with compulsory voting (in past elections)

Search in table



Datawrapper



Flourish Studio

Простые, но профессиональные инструменты



Визуально исследовать данные



«Большая тройка» по версии Gartner



Сделать продукт корпоративной аналитики (дашборд, отчёт и т. д.)



Выбор больше, но, скорее всего, будет тем же, что и для исследования



Нет ограничениям: визуализация с помощью кода



Если нужно настроить график под себя





Практика

Делаем простой
дашборд в Tableau



Итого

- 1 Визуализация — это не только рисование графиков, но ещё и серьёзная подготовка, аналитика
- 2 Для разных этапов аналитики есть разные инструменты
- 3 Лучше всего выбрать свой набор инструментов, которые вы будете хорошо знать
- 4 Важно понимать, какими инструментами какие задачи решают, быть готовыми их быстро освоить



Спасибо за внимание!



fb.com/andmkv

Андрей Макеев
Бизнес-архитектор в Комус

