



Доживем до визуализации: путь данных через тернии процессов

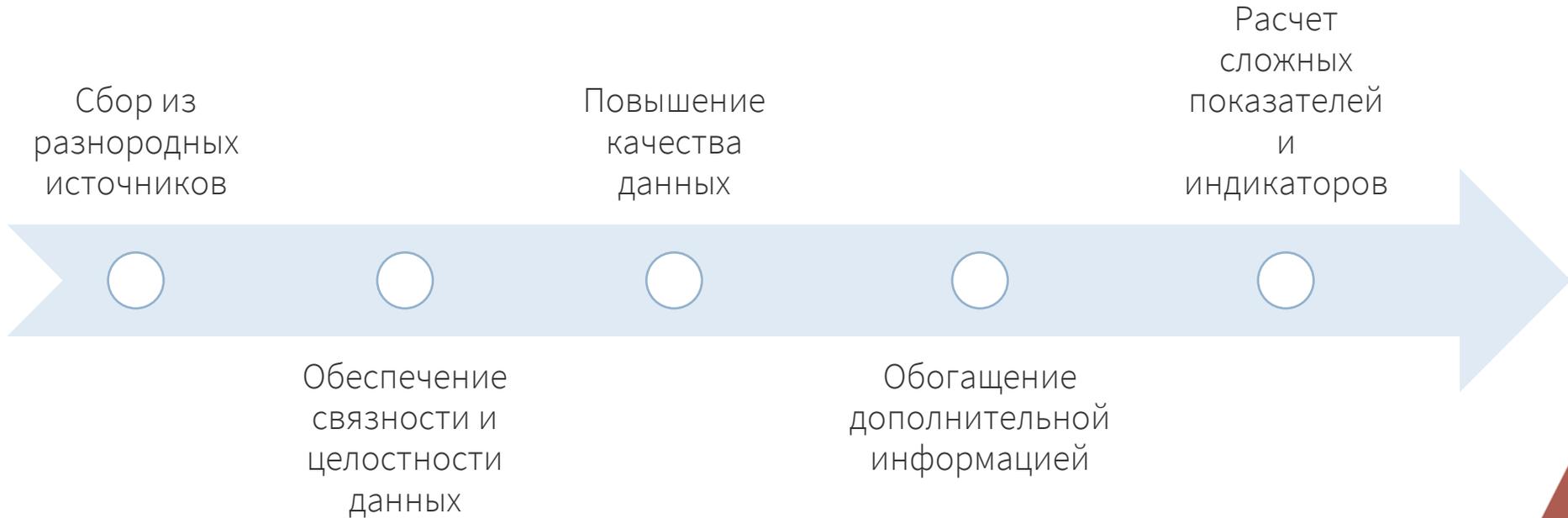
Алексей Арустамов

Loginom Company

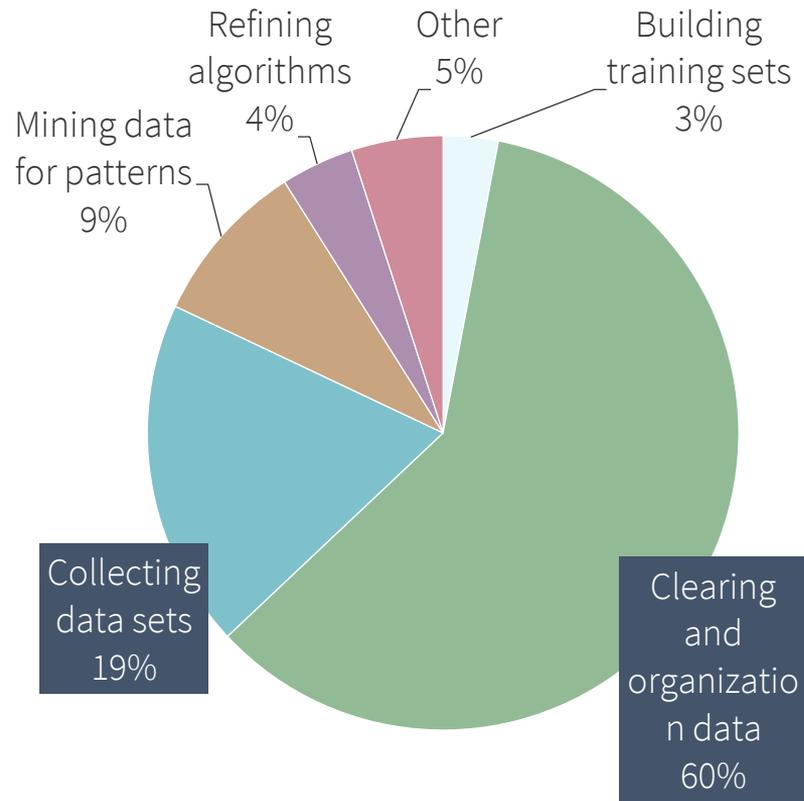
Data driven...
everything:
данные – основа
принятия всех
решений



От «сырых» данных до дашбордов



80% времени
аналитика – сбор,
организация и
очистка данных



Программист

- Знает языки программирования
- Разбирается в математических методах
- Может формализовать логику

VS

Бизнес-эксперт

- Знает предметную область
- Может поставить задачу
- Может интерпретировать результаты

Эксперт-предметник – ключевое лицо,
способное извлечь из данных знания.



Данные



Информация



Визуализация



Знания

Требования к инструменту

1. Визуальное проектирование
2. Повторное использование наработок
3. Интеграция без кодирования
4. Высокая производительность



Loginom

Сделать продвинутую аналитику массовой

Визуальное проектирование



```
diamond-sizes.Rmd x
---
title: "Diamond sizes"
date: 2016-08-25
output: html_document
---
```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)

smaller <- diamonds %>%
 filter(carat <= 2.5)
```

We have data about `r nrow(diamonds)` diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are larger than
2.5 carats. The distribution of the remainder is shown
below:

```{r, echo = FALSE}
smaller %>%
 ggplot(aes(carat)) +
 geom_freqpoly(binwidth = 0.01)
```
```

8:17 Chunk 1: setup R Markdown

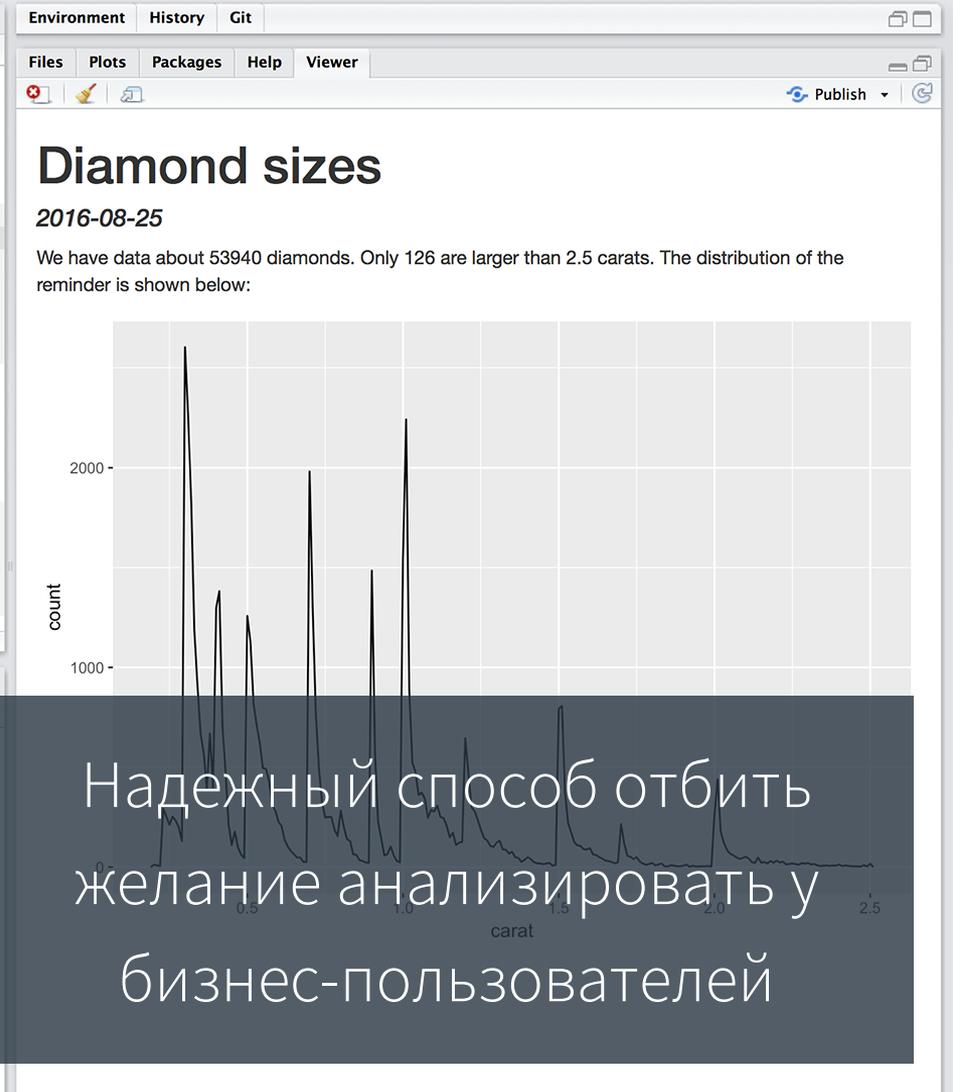
Console R Markdown x
~/Documents/r4ds/r4ds/rmarkdown/ ↗
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

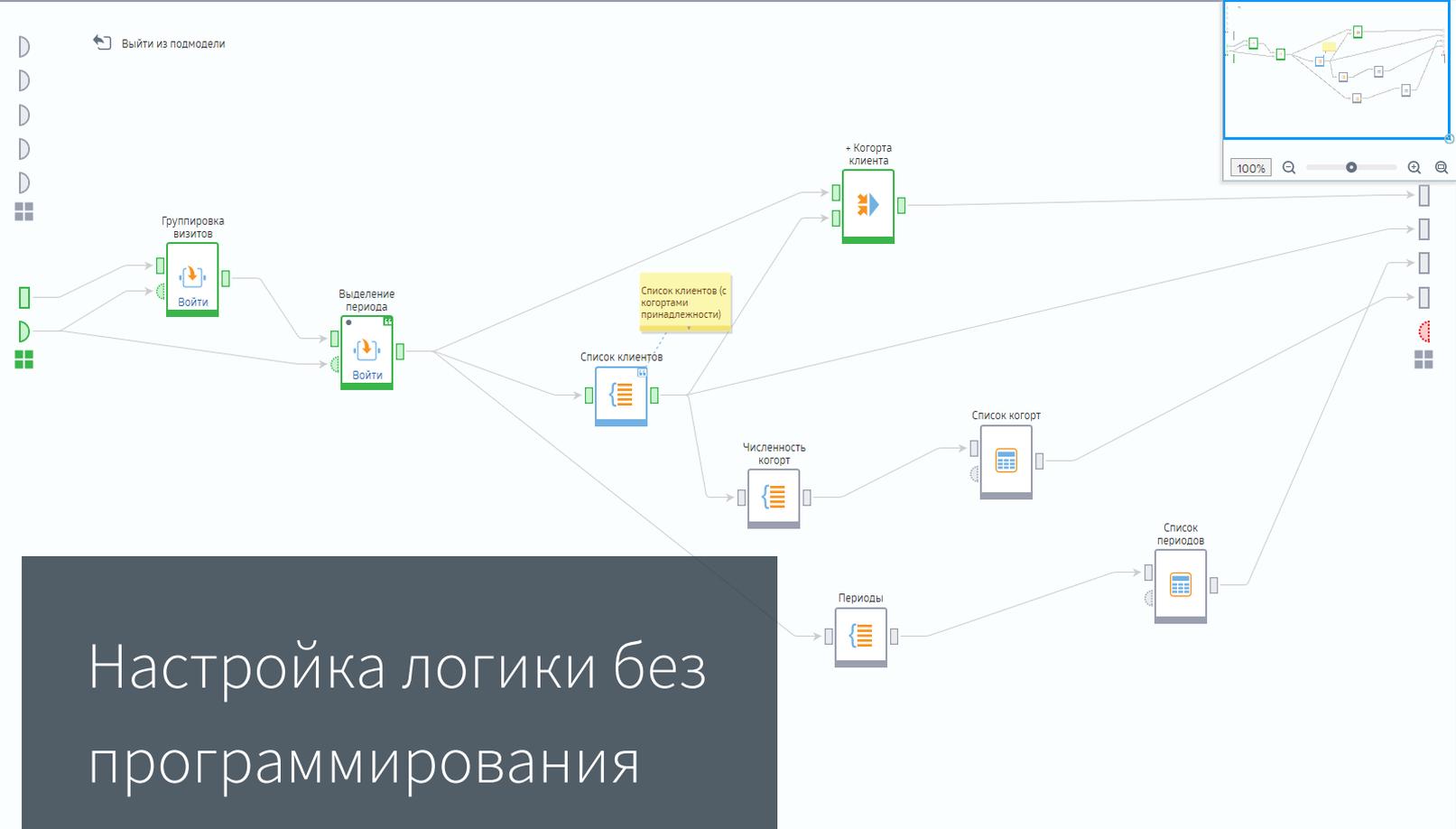
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.



- Дата и время
- Дополнение данных
- Замена
- Калькулятор
- Кросс-таблица
- Объединение
- Параметры полей
- Разгруппировка
- Свертка столбцов
- Скользящее окно
- Слияние
- Соединение
- Сортировка
- Фильтр строк
- Управление
 - Выполнение узла
 - Подмодель
 - Узел-ссылка
 - Условие
 - Цикл
- Исследование
 - Автокорреляция
 - Качество данных
 - Корреляционный анализ
 - Факторный анализ
- Предобработка
 - Заполнение пропусков
 - Квантование
 - Конечные классы
 - Разбиение на множества
 - Редактирование выбросов
 - Сглаживание
 - Сэмплинг
- Data Mining
 - Ассоциативные правила
 - Кластеризация
 - Кластеризация транзакций
 - Самоорганизующиеся сети
 - FM Кластеризация



Настройка логики без программирования

Настройки компонента квантования

Состояние входа:

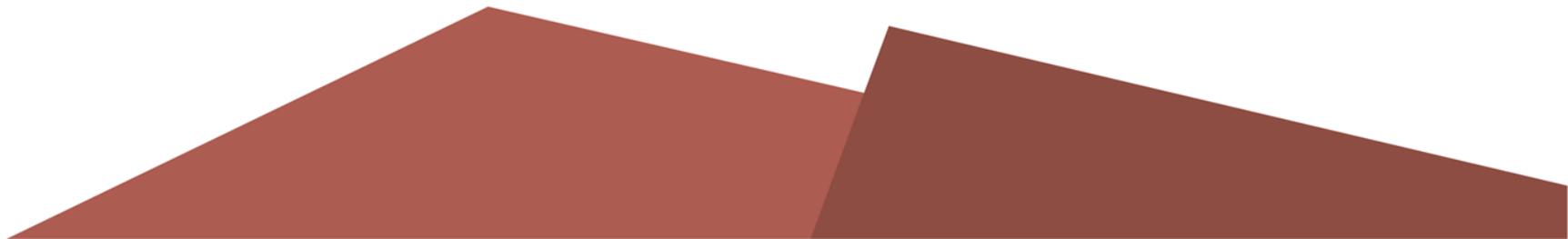
| Поле | Метод | Автоматичес... | Интер... | Минимум | Максимум | |
|----------------|----------------|-------------------------------------|----------|------------------|------------------|--|
| Дата | <Не определен> | <input type="checkbox"/> | 0 | 21.03.2005, 0.00 | 06.05.2005, 0.00 | |
| Цена | <Не определен> | <input type="checkbox"/> | 0 | 1,55 | 128,68 | |
| 9.0 Вес | Количество | <input type="checkbox"/> | 18 | 0,04 | 1,27 | |
| 9.0 Объем | Плитка | <input checked="" type="checkbox"/> | 4 | 0,00 | 0,00 | |
| 9.0 Количество | Плитка | <input checked="" type="checkbox"/> | 6 | 1,00 | 190,00 | |
| 9.0 Сумма | <Не определен> | <input type="checkbox"/> | 0 | 35,77 | 49 041,05 | |
| 9.0 Скидка | Количество | <input type="checkbox"/> | 3 | 0,00 | 0,00 | |

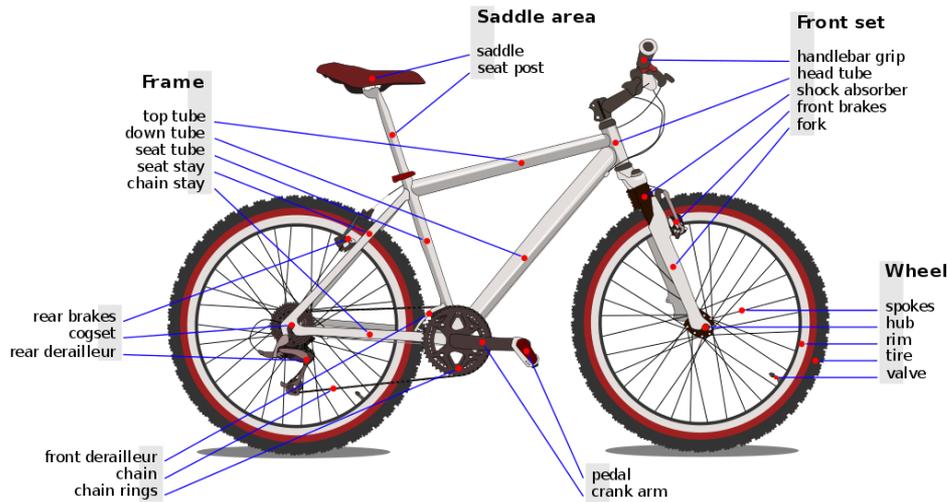
Настройка мастерами: меньше элементов, больше возможностей

Нижняя граница открыта Верхняя граница открыта Шаблон: %FD %OP%MIN.%MAX%CP

| № | Нижняя | Тип | Верхняя | Метка | Объем |
|---|--------|---------|---------|---------------------------|-------|
| 0 | 0,00 | <= X <= | 0,08 | Вес - [0,003706; 0,07629] | 2% |
| 0 | 0,08 | < X <= | 0,15 | Вес - (0,07629; 0,1489] | 9% |
| 0 | 0,15 | < X <= | 0,22 | Вес - (0,1489; 0,2215] | 14% |
| 0 | 0,22 | < X <= | 0,29 | Вес - (0,2215; 0,2941] | 23% |
| 0 | 0,29 | < X <= | 0,37 | Вес - (0,2941; 0,3666] | 20% |
| 0 | 0,37 | < X <= | 0,44 | Вес - (0,3666; 0,4392] | 13% |
| 0 | 0,44 | < X <= | 0,51 | Вес - (0,4392; 0,5118] | 9% |
| 0 | 0,51 | < X <= | 0,58 | Вес - (0,5118; 0,5844] | 4% |
| 0 | 0,58 | < X <= | 0,66 | Вес - (0,5844; 0,657] | 2% |
| 0 | 0,66 | < X <= | 0,73 | Вес - (0,657; 0,7296] | 3% |
| 0 | 0,73 | < X <= | 0,80 | Вес - (0,7296; 0,8022] | 1% |
| 0 | 0,80 | < X <= | 0,87 | Вес - (0,8022; 0,8748] | 1% |
| 0 | 0,87 | < X <= | 0,95 | Вес - (0,8748; 0,9474] | 0% |
| 0 | 0,95 | < X <= | 1,02 | Вес - (0,9474; 1,02] | 0% |
| 0 | 1,02 | < X <= | 1,09 | Вес - (1,02; 1,093] | 0% |

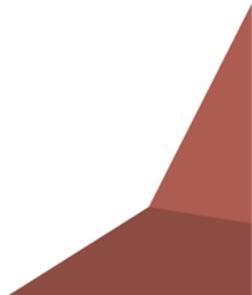
Повторное использование



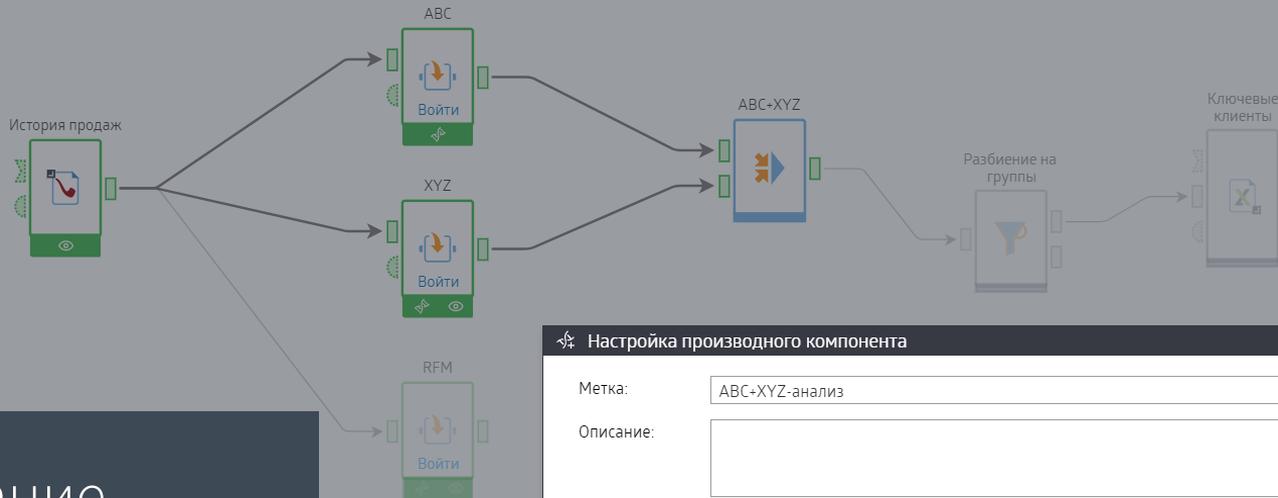


В любом проекте есть типовые модули. Зачем изобретать велосипед при наличии готовых компонентов?

Повторное
использование –
способ
сокращения
времени проекта

1. Готовые
компоненты
 2. Подготовленные
данные
 3. Веб-сервисы
- 

Использование готовых и создание собственных компонентов



Настройка производного компонента

Метка:

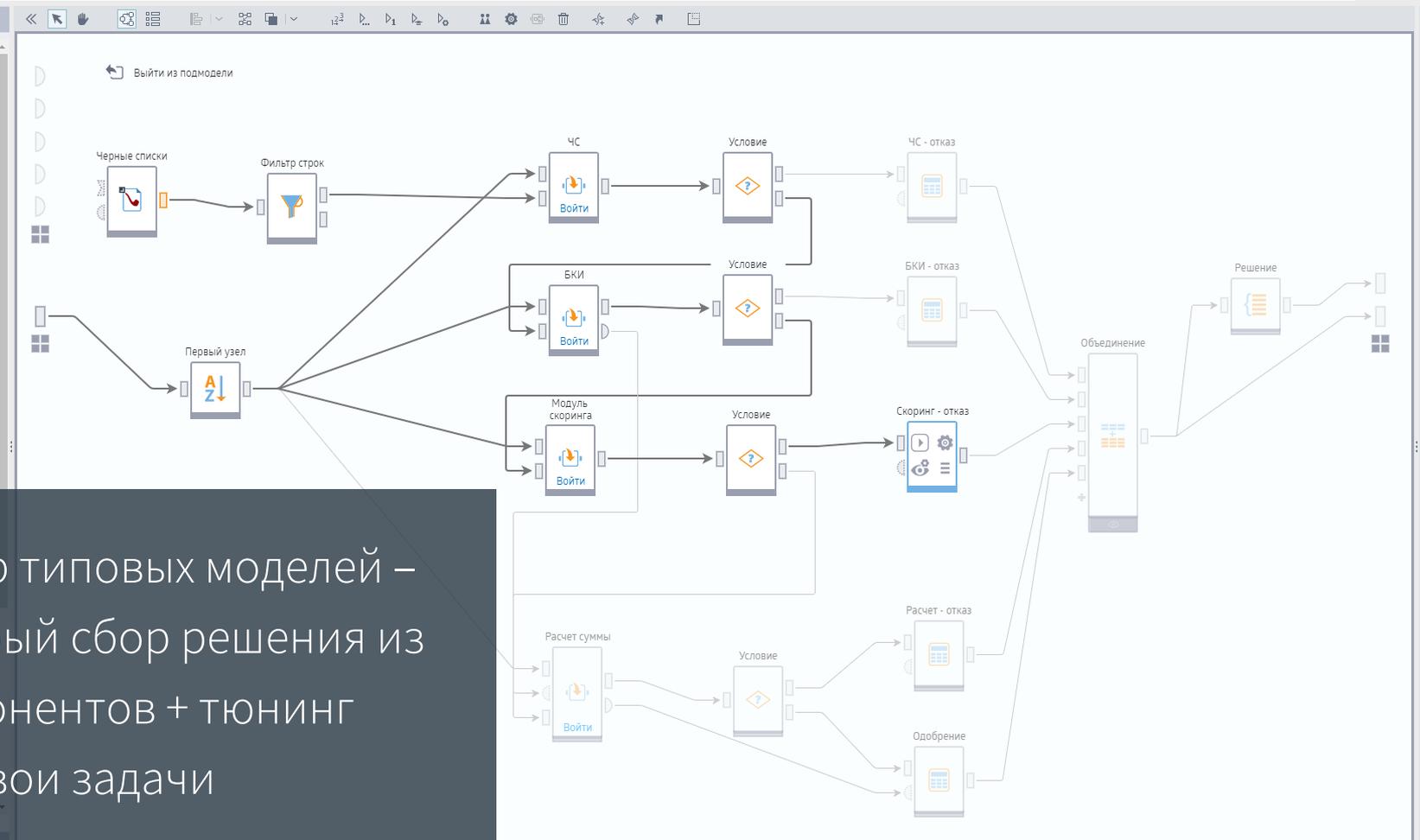
Описание:

Область видимости

- Закрытый (доступен только внутри текущего модуля)
- Частично закрытый (доступен только внутри текущего модуля и его подмоделей)
- Внутренний (доступен только внутри текущего пакета)
- Открытый (доступен во всех пакетах)

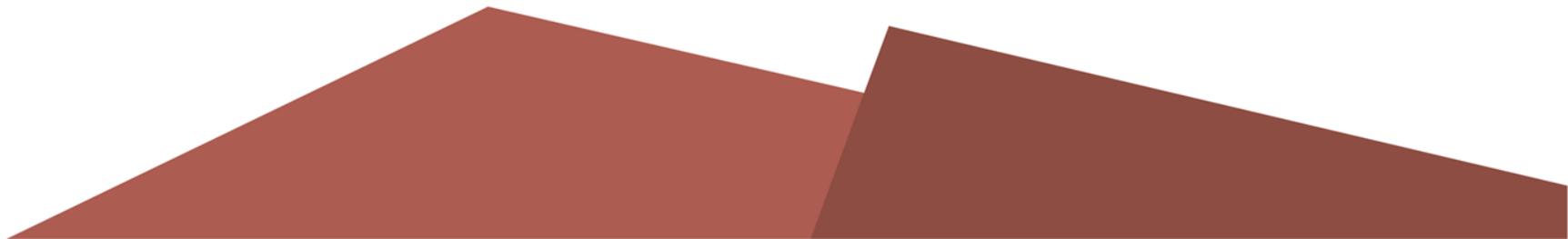
[Отмена](#)

- Компоненты
- Импорт
 - Deductor Data файл
 - База данных
 - Текстовый файл
 - Файл Excel
 - Трансформация
 - Замена (поля)
 - Самплинг
 - Объединение
 - Фильтр строк
 - Разбиение на множества
 - Скользящее окно
 - Конечные классы
 - Автокорреляция
 - Сортировка
 - Сглаживание
 - Факторный анализ
 - Слияние
 - Вейвлет-сглаживание
 - Свертка столбцов
 - Группировка
 - Соединение
 - Дата и время
 - Корреляционный анализ
 - Квантование
 - Калькулятор (доп.)
 - Дополнение
 - Параметры полей
 - Data Mining
 - Классификация
 - EM Кластеризация
 - Ассоциативные правила
 - Самоорганизующиеся сети
 - Кластеризация
 - Управление
 - Выполнение узла
 - Цикл
 - Узел
 - Управление
 - Производные компоненты
 - Подключения

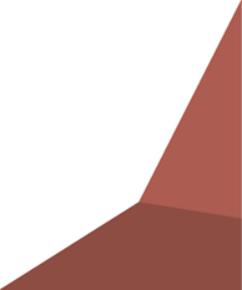


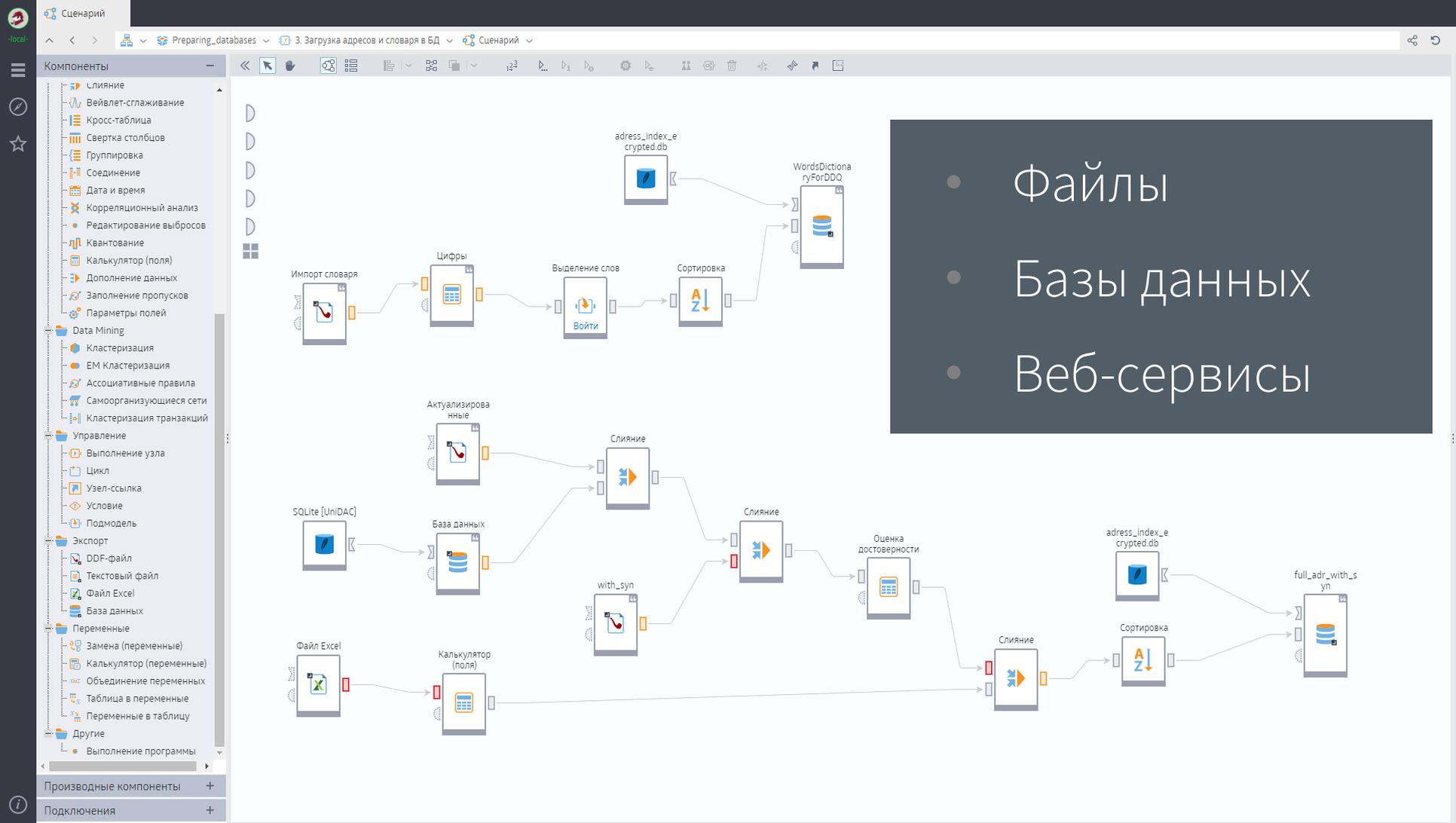
Набор типовых моделей –
быстрый сбор решения из
компонентов + тюнинг
под свои задачи

Интеграция без кодирования

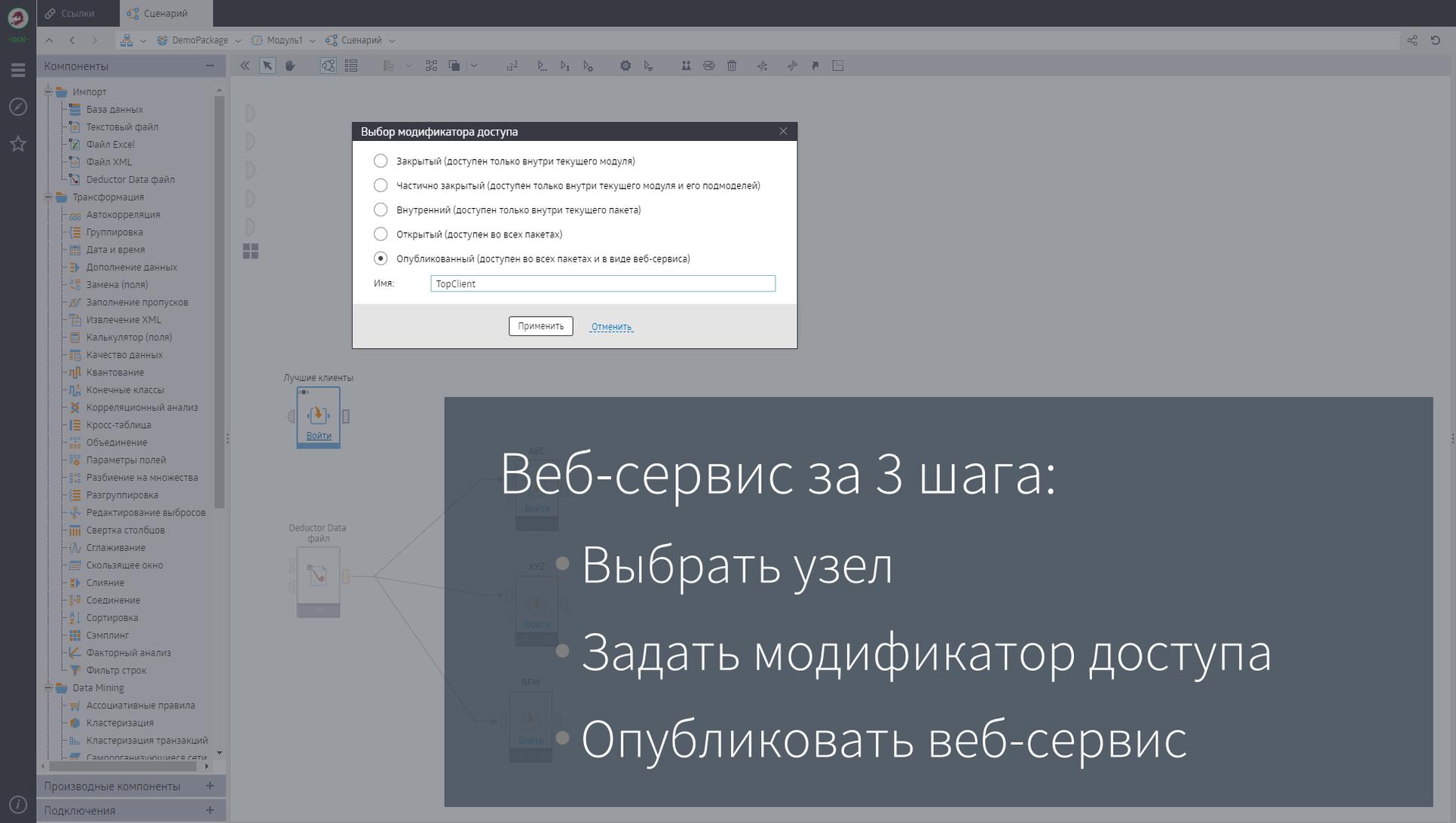


Задачи интеграции

- Взаимодействие со множеством источников
 - Необходимость очистки и предобработки
 - Требование реализации бизнес-логики
 - Необходимость обогащения данных
- 



- Файлы
- Базы данных
- Веб-сервисы



Выбор модификатора доступа

Закрытый (доступен только внутри текущего модуля)

Частично закрытый (доступен только внутри текущего модуля и его подмоделей)

Внутренний (доступен только внутри текущего пакета)

Открытый (доступен во всех пакетах)

Опубликованный (доступен во всех пакетах и в виде веб-сервиса)

Имя:

Лучшие клиенты

Войти

Веб-сервис за 3 шага:

- Выбрать узел
- Задать модификатор доступа
- Опубликовать веб-сервис

ABC
Войти

XYZ
Войти

RFM
Войти

Deductor Data файл

- Компоненты
- Импорт
 - База данных
 - Текстовый файл
 - Файл Excel
 - Файл XML
 - Deductor Data файл
 - Трансформация
 - Автокоррекция
 - Группировка
 - Дата и время
 - Дополнение данных
 - Замена (поля)
 - Заполнение пропусков
 - Извлечение XML
 - Калькулятор (поля)
 - Качество данных
 - Квантование
 - Конечные классы
 - Корреляционный анализ
 - Кросс-таблица
 - Объединение
 - Параметры полей
 - Разбиение на множества
 - Разгруппировка
 - Редактирование выбросов
 - Свертка столбцов
 - Сглаживание
 - Скользящее окно
 - Слияние
 - Соединение
 - Сортировка
 - Самплинг
 - Факторный анализ
 - Фильтр строк
 - Data Mining
 - Ассоциативные правила
 - Кластеризация
 - Кластеризация транзакций
 - Сампоорганизующие сети
- Производные компоненты +
- Подключения +

Высокая производительность



Способы увеличения скорости:

1. In-Memory Computing – вычисления в оперативной памяти
2. Parallel Computing – параллельные вычисления
3. Экономное хранение больших данных
4. Быстрые вычислительные алгоритмы

Технология

Выгода

Асинхронная
обработка

Отсутствие блокировок
пользовательского интерфейса
при долгих операциях

Ленивые
вычисления

Экономия ресурсов за счет
расчетов только при
необходимости

LogiNot + инструмент
визуализации



От данных до принятия решений

Logiном

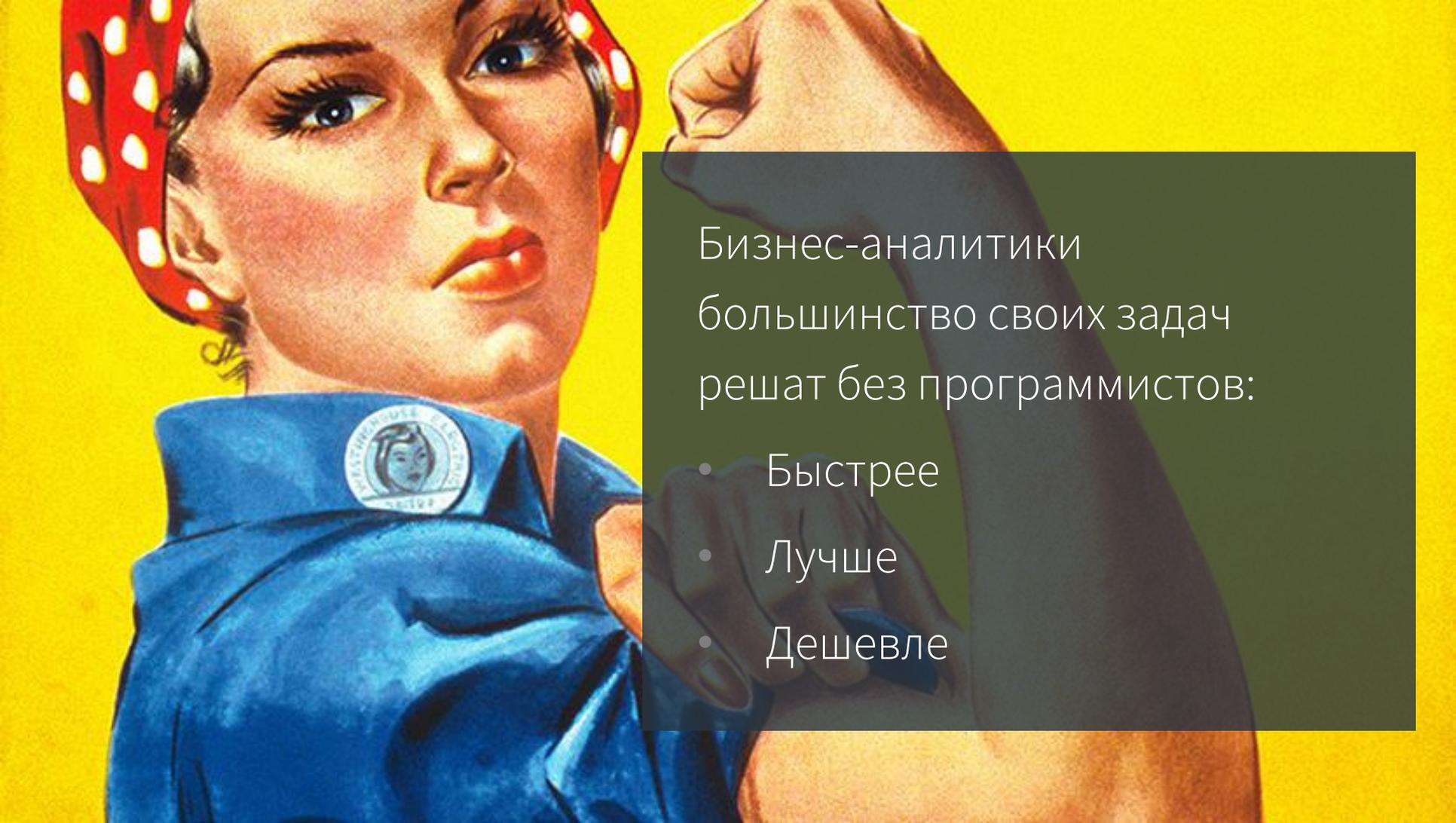
- Подготовка данных
- Обогащение данных
- Расчет показателей
- Моделирование

Средства визуализации

- Визуализация
- Дашборды
- Доставка данных
- Исследование

Эффект применения:

1. **Снижение трудозатрат** – весь цикл работы с данными в одной среде без кодирования
2. **Вовлеченность сотрудников** – самостоятельная реализация сложной логики анализа без привлечения IT
3. **Обогащение данных** – повышение информативности за счет использования данных из множества источников



Бизнес-аналитики
большинство своих задач
решат без программистов:

- Быстрее
- Лучше
- Дешевле

loginom.ru

