

Подготовить стенд
с обезличенными
данными и сохранить
пользу для тестирования



Валентина Васильева
Лид тестирования





Валентина Васильева

 Лид тестирования

- 11 лет в тестировании
- 9 лет работаю с качеством данных
- Тестирую продукт по обезличиванию данных «**Маскировщик**»

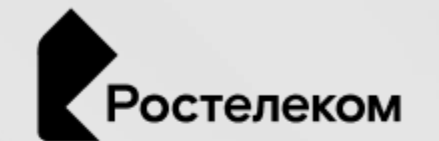
Почему тема качества данных нам так близка?

1

Тестируем свои продукты DQ **сами** и знаем, как важен их формат и качество

2

Без знания **специфики данных** сложно подготовить синтетику, покрывающую большинство кейсов





СКВАЙР ТРЕЛОНИ

**ТУП, ЖАДЕН, ПРОЖОРЛИВ, ЛЕНИВ,
ТРУСЛИВ, НАДМЕНЕН.
ХАРАКТЕР ОТСУТСТВУЕТ.
НЕ ЖЕНАТ.**

01
Что такое
персональные данные?

Зачем вообще нужны обезличенные персональные данные?

1 Для тестирования

Реальные системы требуют тщательного тестирования. Чтобы делать его качественно, необходимо проверять на очень похожих данных.



Зачем вообще нужны обезличенные персональные данные?

1 Для тестирования

Реальные системы требуют тщательного тестирования. Чтобы делать его качественно, необходимо проверять на очень похожих данных.

2 Для безопасности

Доступ к данным на боевых средах находится под строгим контролем, тогда как доступ к тестовым средам открыт бóльшему количеству людей.



Зачем вообще нужны обезличенные персональные данные?

1 Для тестирования

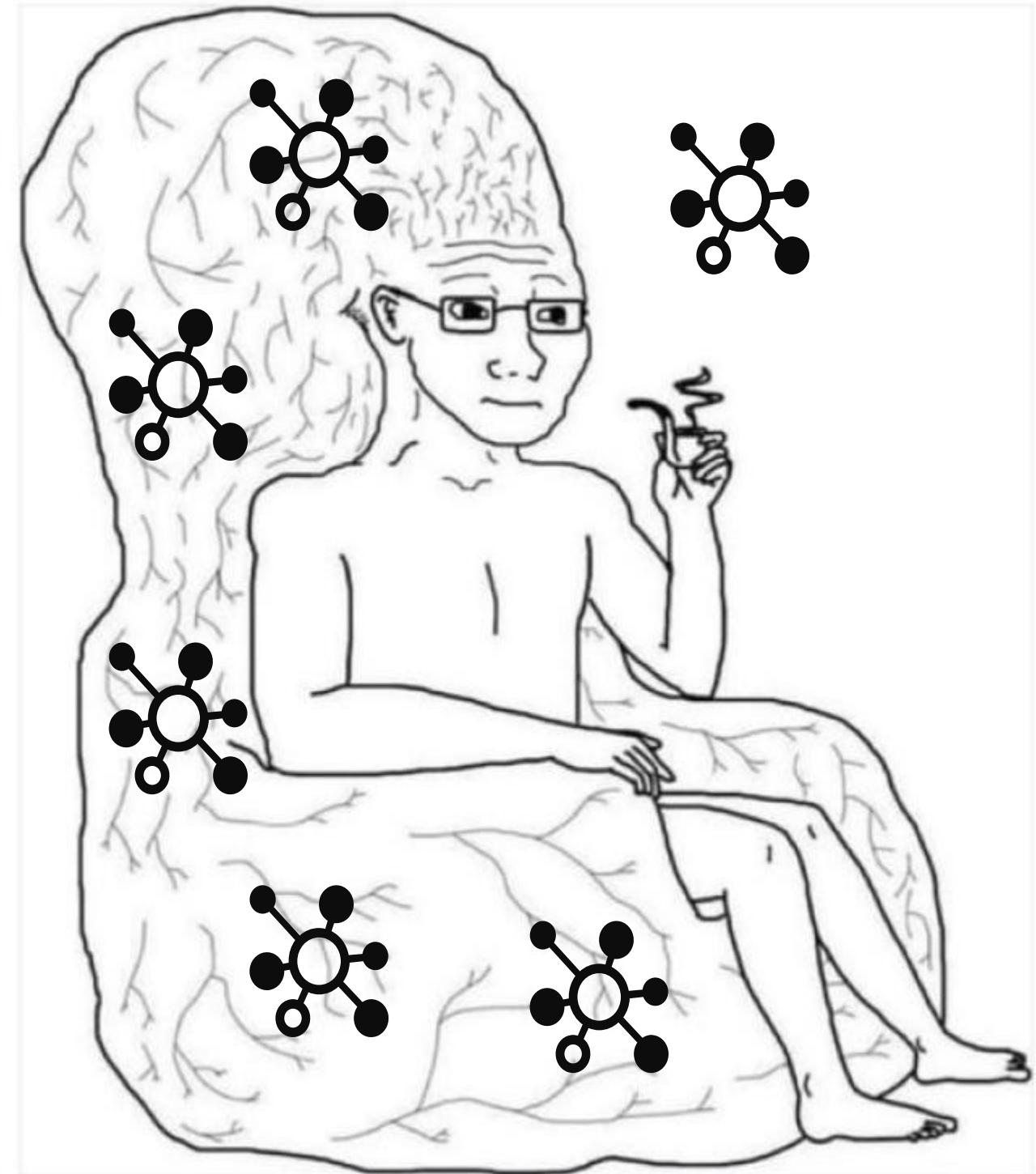
Реальные системы требуют тщательного тестирования. Чтобы делать его качественно, необходимо проверять на очень похожих данных.

2 Для безопасности

Доступ к данным на боевых средах находится под строгим контролем, тогда как доступ к тестовым средам открыт бóльшему количеству людей.

3 Для построения моделей

Для построения аналитических моделей (в том числе, с привлечением подрядчиков) также необходим доступ к данным.



Методы обезличивания от РКН



- Введение идентификаторов

Методы обезличивания от РКН



- Введение идентификаторов
- Изменение состава или семантики

Методы обезличивания от РКН



- Введение идентификаторов
- Изменение состава или семантики
- Декомпозиция

Методы обезличивания от РКН



- Введение идентификаторов
- Изменение состава или семантики
- Декомпозиция
- Перемешивание

Методы обезличивания от РКН



- Введение идентификаторов
- Изменение состава или семантики
- Декомпозиция
- Перемешивание
- **Хеширование**



Роскомнадзор **не дает**
готового решения!

02

Подходы
к обезличиванию



Подход 1

Замена на «звездочки»

— Меняет тип данных

Поэтому алгоритм не годится для обезличивания чисел и дат

— Слабо защищает популярные имена и адреса

Даже если сократить число «звездочек» до одной, популярные имена, отчества и города легко расшифровывать

— Убивает смысл данных

Обезличенные данные теряют семантику, валидность, социально-демографические характеристики и связи по домохозяйствам

Фамилия, имя, отчество

Абдюшев Павел
Рашитович



А*****В П***л
Р*****ч

Дата рождения

21.01.1979



2*.*1.1**9

Паспорт

6806 108711



6**61****1

Телефон

8 926 118-12-12



89*****12

Адрес

Москва,
Турчанинов пер., 6,
стр. 2



М****а,
Т*****в пер., 6,
стр. *

Замена букв на буквы, цифр на цифры

⊖ Поддается расшифровке

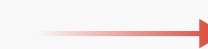
Если алгоритм замены простой, то обезличенные данные можно восстановить

⊖ Портит качество и полноту данных

Заменяя осмысленные буквы и цифры случайными, мы теряем семантику, валидность, социально-демографические характеристики, связи по домохозяйствам

Фамилия, имя, отчество

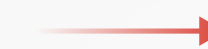
Еременко Наталья
Сергеевна



Нонингел Гпъпдбч
Мношнагп

Дата рождения

21.07.1961



11.12.2835

Паспорт

6806 108711



7187 315818

Телефон

8 926 118-12-12



9 037 229-23-23

Адрес

Тамбов,
ул. Советская, 11



Нуктаз,
ец. Зукогвяу, 57

Подход 3

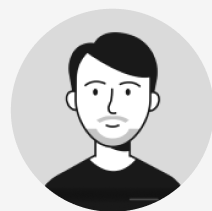
Маскирование с сохранением формата данных и аналитических характеристик

H F Labs



Сохранение аналитических характеристик

Еременко Петр
Сергеевич



21 июля 1960

Ванина, 1, Тамбов

6806 108771

8 926 118-12-12
mario@gmail.com

Антонов Сергей
Андреевич



Сохраняется гендерный баланс

11 февраля 1961

Дорожная 5, Тамбов

6807 203771 8 926 311-89-84 elf@mail.ru

Сохранение аналитических характеристик

Еременко Петр
Сергеевич



21 июля 1960

Ванина, 1, Тамбов

6806 108771

8 926 118-12-12
mario@gmail.com

Сохраняются родственные связи.
Если Петр превратился в Сергея,
то Петровичи – в Сергеевичей

Антонов Сергей
Андреевич



Сохраняется гендерный
баланс

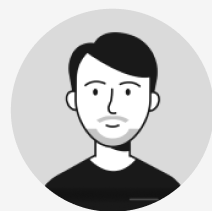
11 февраля 1961

Дорожная 5, Тамбов

6807 203771 8 926 311-89-84 elf@mail.ru

Сохранение аналитических характеристик

Еременко Петр
Сергеевич



21 июля 1960

Ванина, 1, Тамбов

6806 108771

8 926 118-12-12

mario@gmail.com

Сохраняются родственные связи.
Если Петр превратился в Сергея,
то Петровичи – в Сергеевичей

Антонов Сергей
Андреевич



11 февраля 1961

Дорожная 5, Тамбов


6807 203771 8 926 311-89-84 elf@mail.ru

Сохраняется гендерный
баланс

Сохраняются социально-
демографические группы

Сохранение аналитических характеристик

Еременко Петр Сергеевич



21 июля 1960

Ванина, 1, Тамбов


6806 108771

8 926 118-12-12

mario@gmail.com

Сохраняются родственные связи. Если Петр превратился в Сергея, то Петровичи – в Сергеевичей

Антонов Сергей Андреевич



11 февраля 1961

Дорожная 5, Тамбов

6807 203771 8 926 311-89-84 elf@mail.ru

Сохраняется гендерный баланс

Сохраняются социально-демографические группы

В адресе сохраняется страна, город и район

Сохранение аналитических характеристик

Еременко Петр
Сергеевич



21 июля 1960

Ванина, 1, Тамбов

6806 108771

8 926 118-12-12

mario@gmail.com

Сохраняются родственные связи.
Если Петр превратился в Сергея,
то Петровичи — в Сергеевичей

Антонов Сергей
Андреевич



11 февраля 1961

Дорожная 5, Тамбов

6807 203771 8 926 311-89-84 elf@mail.ru

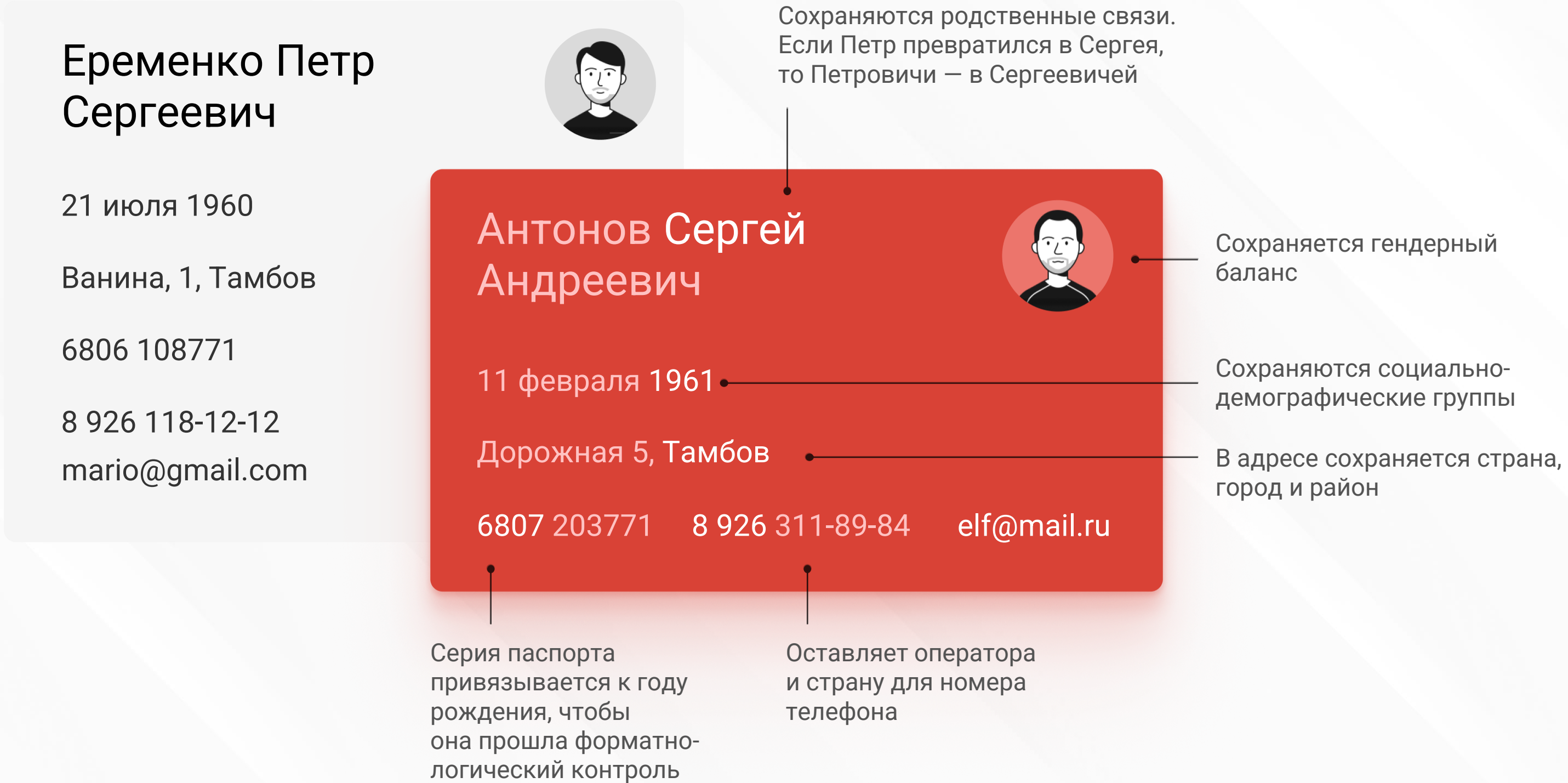
Сохраняется гендерный
баланс

Сохраняются социально-
демографические группы

В адресе сохраняется страна,
город и район

Серия паспорта
привязывается к году
рождения, чтобы
она прошла форматно-
логический контроль

Сохранение аналитических характеристик



Сохранение домохозяйств

Иванов Петр Ильясович

21.07.1961

Тамбов,
Советская, 11

Петров Андрей Фатихович

11.02.1961

Тамбов, Астраханская, 5

Иванов Ильяс Петрович

05.03.1988

Тамбов,
Советская, 11

Петров Фатих Андреевич

11.08.1987

Тамбов, Астраханская, 5

03

Выбрать подход —
недостаточно



01. Качество и смысл данных для тестирования

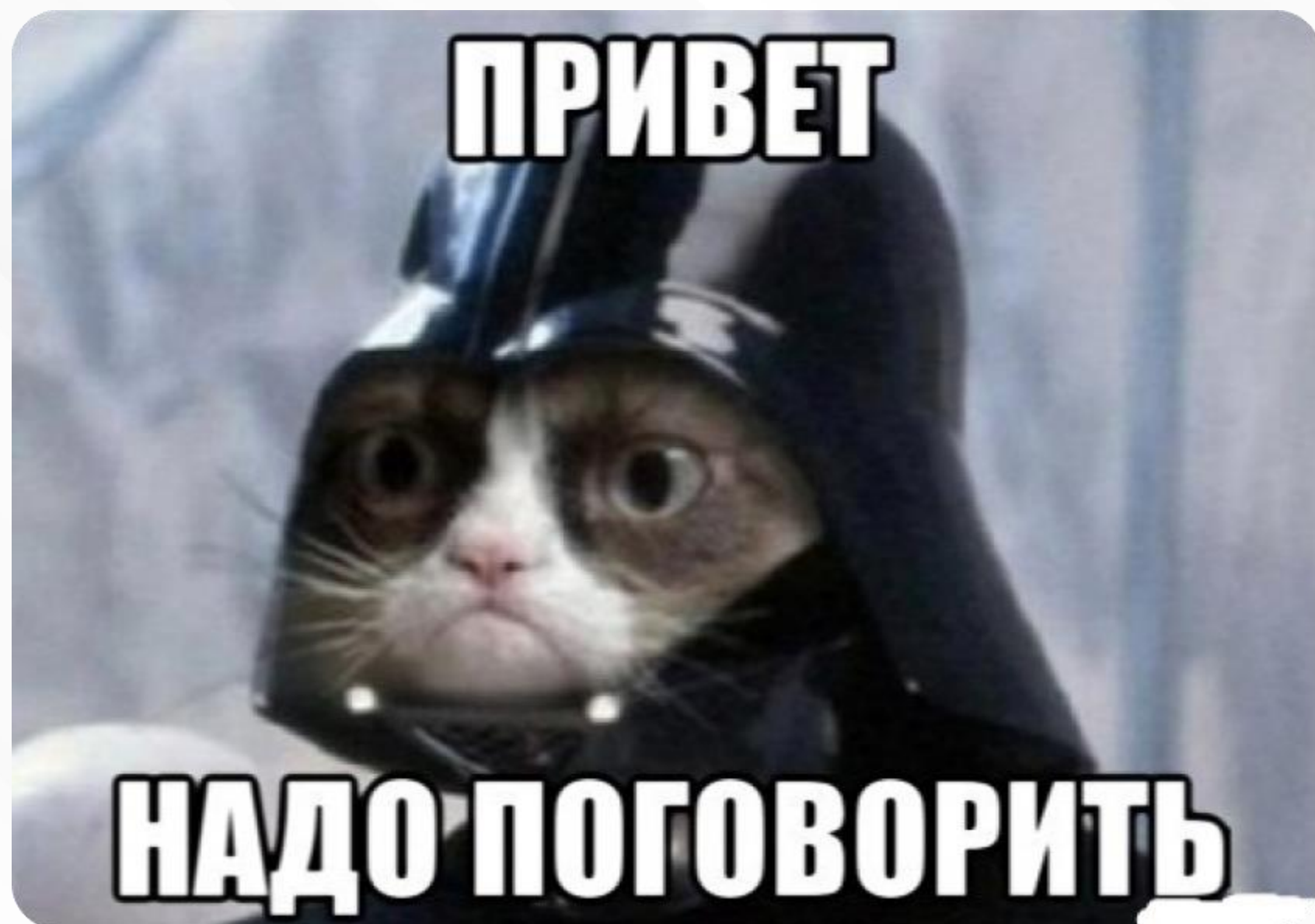


- Не сохранены смысловые характеристики и метрики качества данных

- Часть тестовых сценариев не воспроизводимы
- Теряются бизнес-правила обработки данных

01. Качество и смысл данных для тестирования

Что учесть?

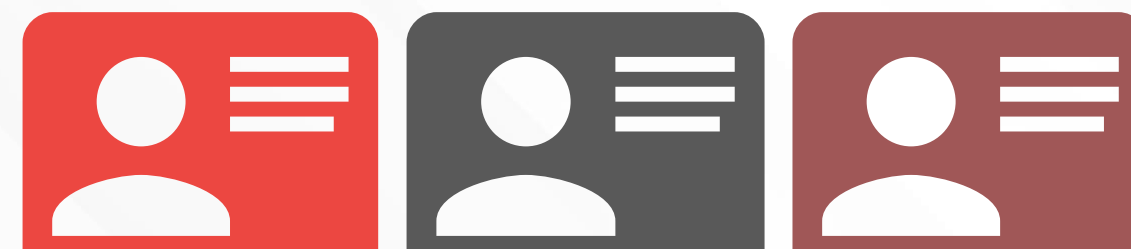


- Важные параметры данных, проработанные с бизнесом и аналитиками
- Логику систем, которые поднимаются на данных
- Возможность замены данных с учетом заданных параметров размытия
- Валидность контрольных сумм
- Наличие в справочниках и реестрах
- Консистентность данных между собой

02. Консистентность маскирования нескольких сред



- Разные источники маскируют данные по-разному



- Невозможность интеграционных тестов и агрегированной аналитики

02. Консистентность маскирования нескольких сред. **Что учесть?**

Одинаковое маскирование всех источников важно
для тестирования интеграций между несколькими базами данных



Маскировать **консистентно**
разные стенды и типы СУБД,
файлы, текст



Маскировать **инкрементально**:
домаскировать изменения в уже
маскированную ранее базу

03. Специфика БД, систем и интеграций

- Формат и маска



- Падение интеграций

03. Специфика БД, систем и интеграций

- Формат и маска



- Падение интеграций

- Валидность объектов БД



- Сломанные процедуры и функции БД

03. Специфика БД, систем и интеграций

- Формат и маска



- Падение интеграций

- Валидность объектов БД



- Сломанные процедуры и функции БД

- Ключи



- А вдруг персональные данные – РК

03. Специфика БД, систем и интеграций

- Формат и маска



- Падение интеграций

- Валидность объектов БД



- Сломанные процедуры и функции БД

- Ключи



- А вдруг персональные данные – РК

- Ссылочная целостность



- Падение по FK

03. Специфика БД, систем и интеграций

- Формат и маска



- Падение интеграций

- Валидность объектов БД



- Сломанные процедуры и функции БД

- Ключи



- А вдруг персональные данные – РК

- Ссылочная целостность



- Падение по FK

- Спецсимволы



- Падения при особых символах

03. Специфика БД, систем и интеграций

- Формат и маска



- Падение интеграций

- Валидность объектов БД



- Сломанные процедуры и функции БД

- Ключи



- А вдруг персональные данные – РК

- Ссылочная целостность



- Падение по FK

- Спецсимволы



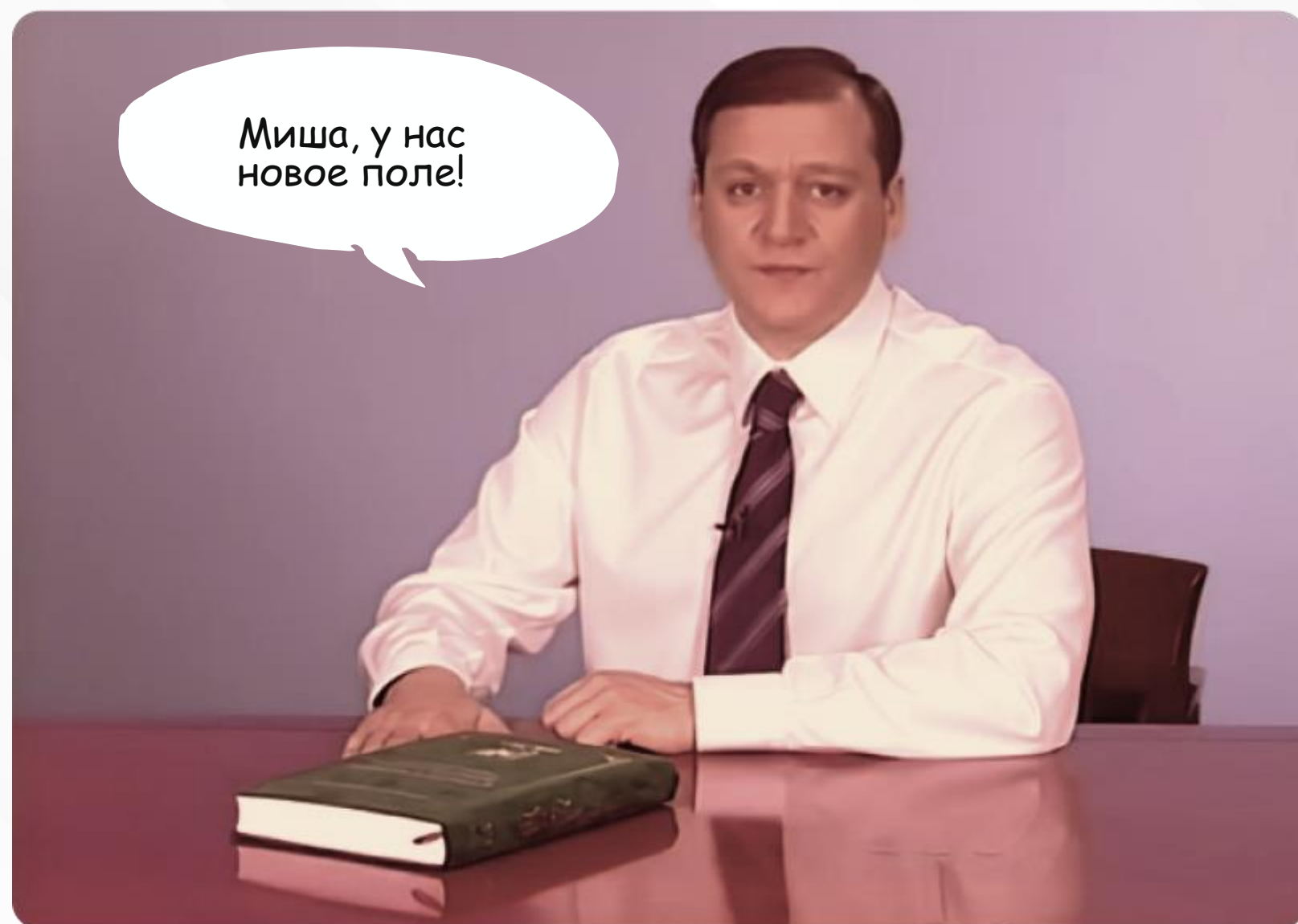
- Падения при особых символах

- Исключения



- Невозможность аутентификации для интеграционного тестирования

04. Процессы внутри компании



- Система живет, появляются новые поля и таблицы, ранее пустые заполняются новыми данными



- Появляется поле с ПДн, но аналитик забыл передать указания оператору маскирования



- Выходят новые разъяснения РКН



- Хеширование — уже не метод обезличивания

- В тестовой среде оказываются необезличенные данные

- В случае утечки — риски для репутации и оборотные штрафы

04. Процессы внутри компании

Что учесть на длинном пути?



● Принуждать аналитиков рассказывать об изменениях

● Настроить автоматический мониторинг наполнения полей

05. Безопасность маскирования



Безопасность

Полезность

Чек-лист для самостоятельной организации процесса маскирования

- Мониторинг требований РКН
- Взаимодействие со службой ИБ
- Контакт с бизнесом и аналитиками, ML
- Учет сценариев тестирования
- Баланс между безопасностью и пользой (с учетом сценариев и контекста)
- Мониторинг изменений бизнес-сценариев и модели данных

H

F

Labs

Спасибо за внимание!
Вопросы, cool stories?



Валентина Васильева
Лид тестирования
[@Valentine_vv](#)



Статья на Хабр