

Менеджмент Data Science проектов: методология CRISP-DM

Сергей Зотов
Technical Product / Platform Owner





ВЕЩАЕТ

Сергей Зотов

Управляю IT продуктами с 2015 года, руководжу Data Science проектами с 2017 года. Руководитель направления Research & Development в FunBox

СВЯЗАТЬСЯ СО МНОЙ



@s.zotov



linkedin.com/in/szotov



facebook.com/zottttttt

```
>> import pandas as pd  
>> from sklearn.tree import DecisionTreeRegressor
```



**— You know, I'm something
of a scientist myself**

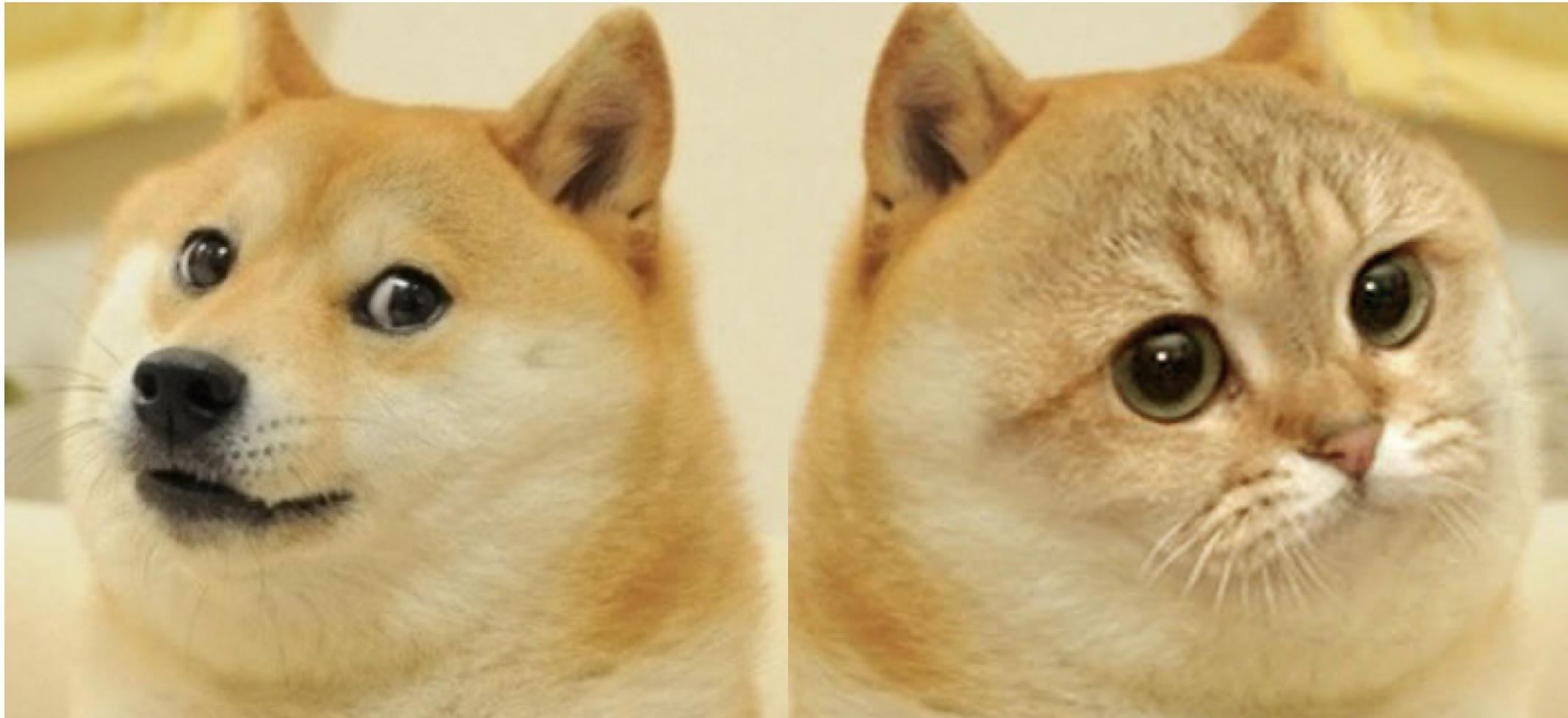
Основные проблемы

Постановка задачи

Я хочу классифицировать фото кошечек и доге



Я хочу классифицировать фото кошечек и доге



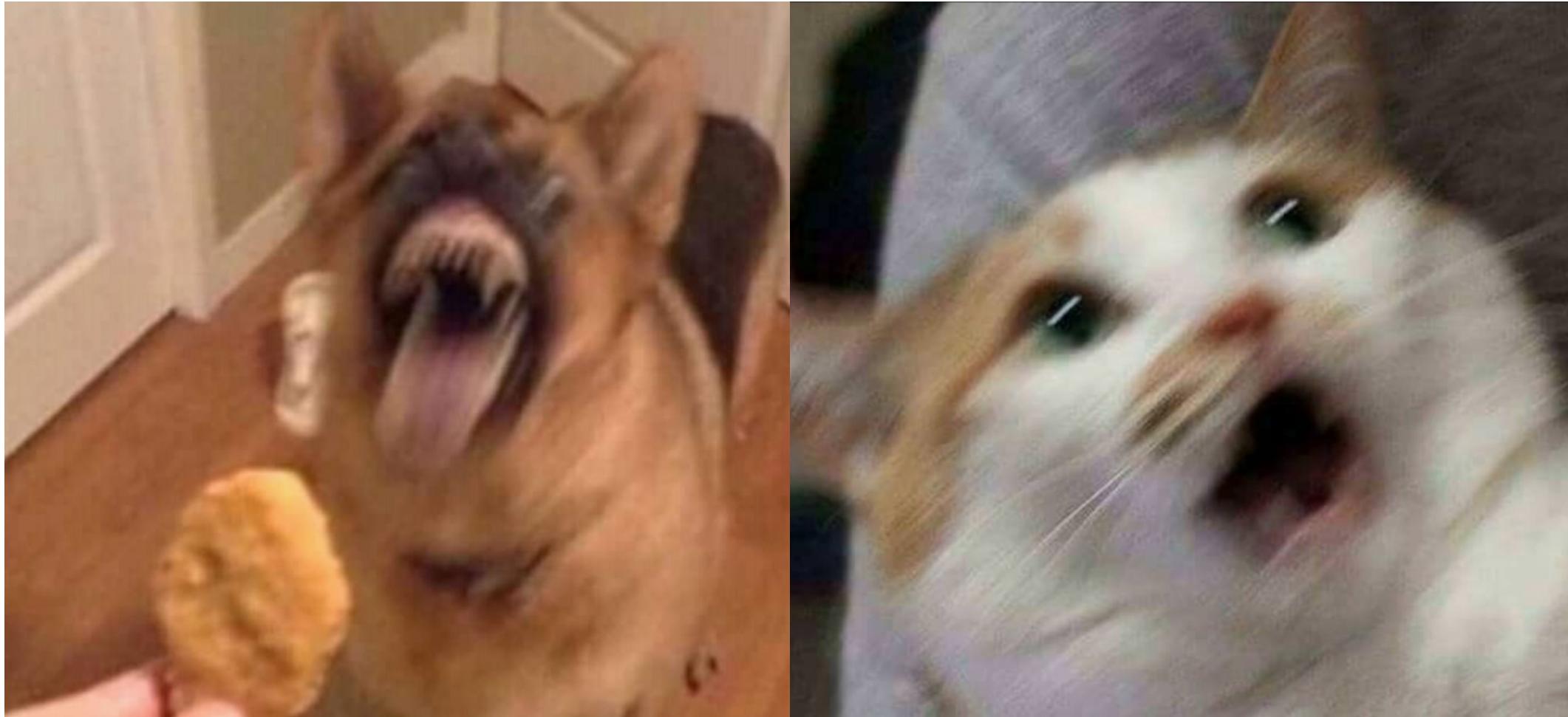
Я хочу классифицировать фото кошечек и доге



Я хочу классифицировать фото кошечек и доге



Я хочу классифицировать фото кошечек и доге



Я хочу классифицировать фото кошечек и доге



Я хочу классифицировать фото кошечек и дого



Я хочу классифицировать фото кошечек и доге



Я хочу классифицировать фото кошечек и доге



Процесс работы

JIRA



Менеджер

Я хочу драг'н'дропать фоточку в аппку, чтобы она давала ответ: кошечка это или доге.

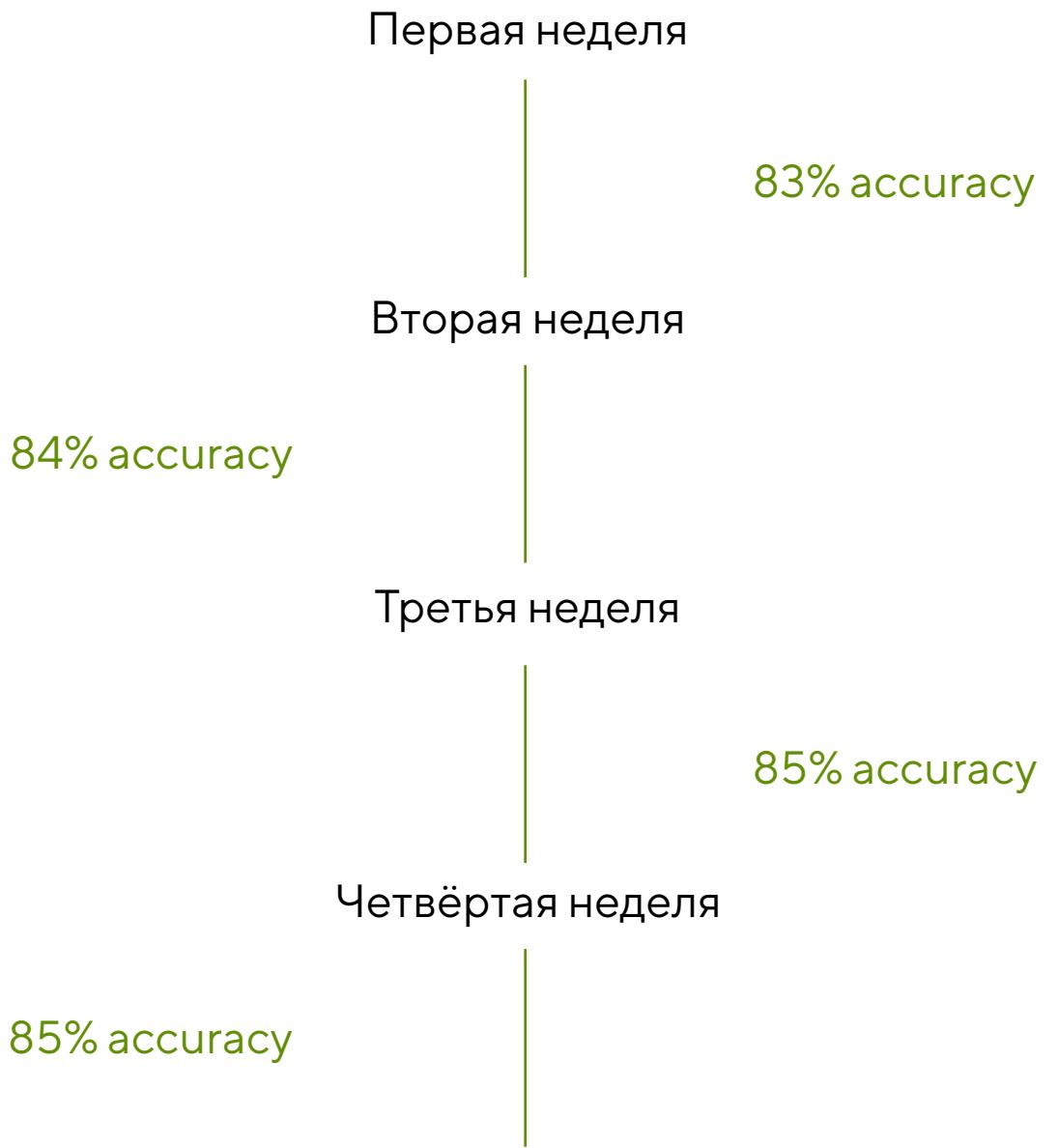
Данные в приложении.

???

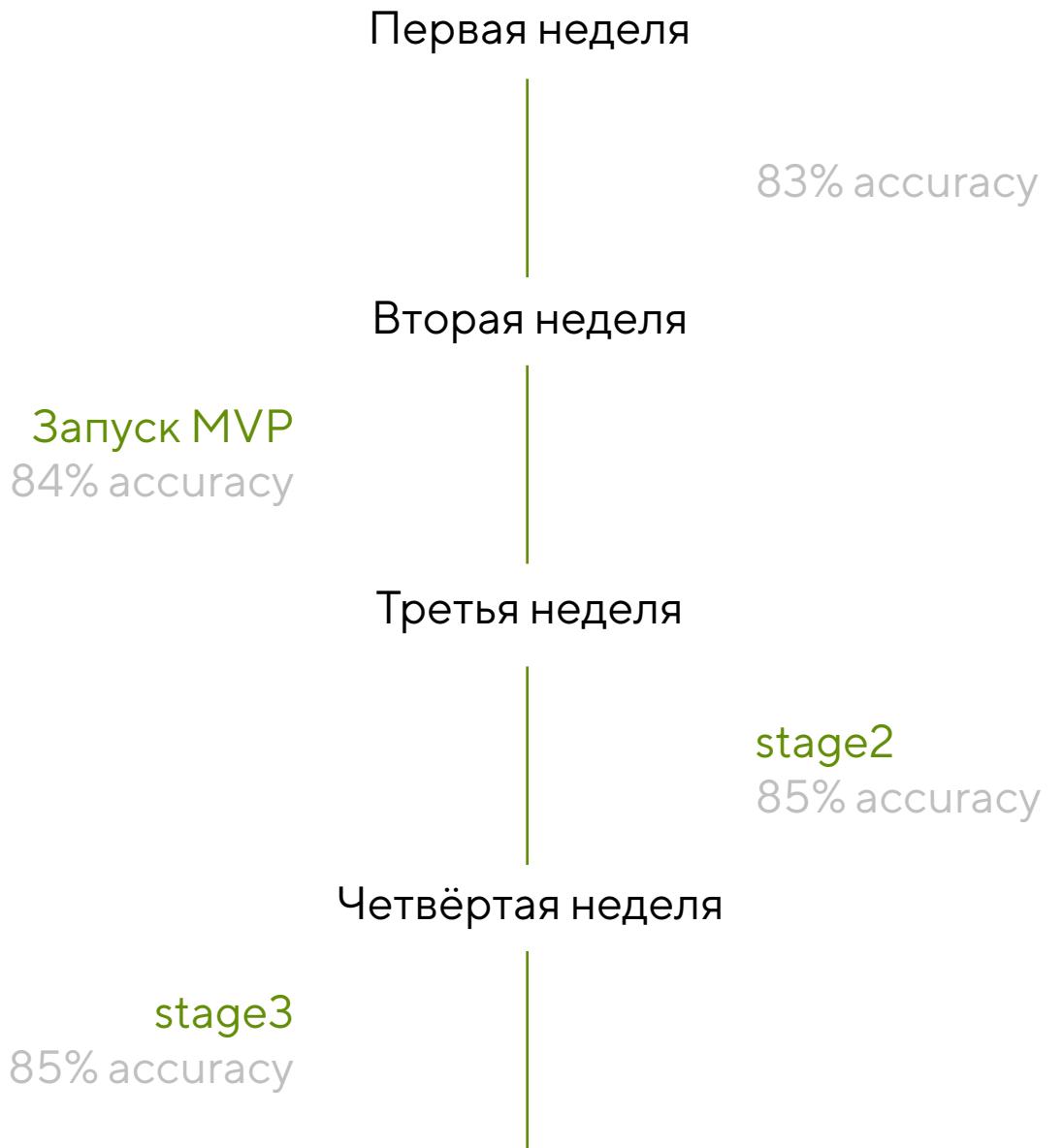


Data Scientist

Roadmap



Roadmap



Результат

Результат

Запрос	Prediction
1.jpg	0
2.jpg	1
3.jpeg	0
4.jpg	1

Запрос	Prediction
1.jpg	0,121
2.jpg	0,898
3.jpeg	0,283
4.jpg	0,581

Контроль версий

Контроль версий

1. Как был получен предыдущий датасет?

Контроль версий

1. Как был получен предыдущий датасет?
2. Jupyter тетрадка может находиться в полном хаосе.

Контроль версий

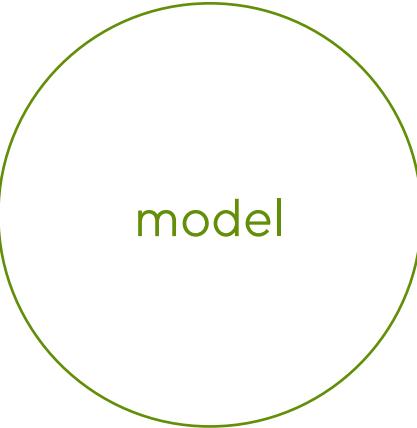
1. Как был получен предыдущий датасет?
2. Jupyter тетрадка может находиться в полном хаосе.
3. Какая метрика была получена с помощью этой версии тетрадки?

Контроль версий

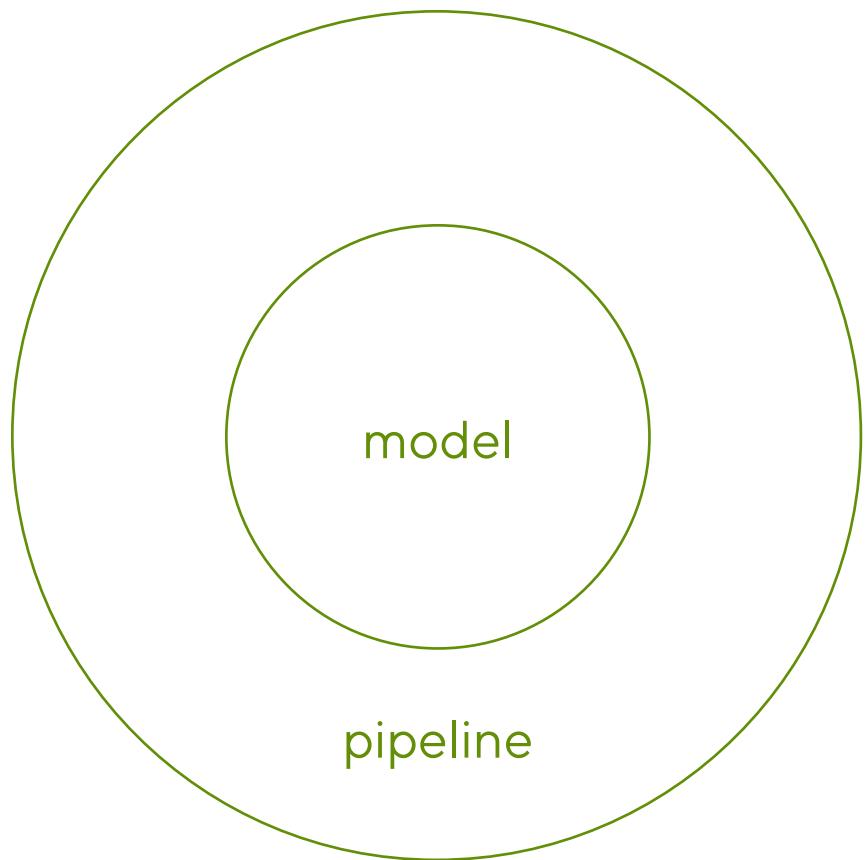
1. Как был получен предыдущий датасет?
2. Jupyter тетрадка может находиться в полном хаосе.
3. Какая метрика была получена с помощью этой версии тетрадки?
4. Можем ли мы использовать git?

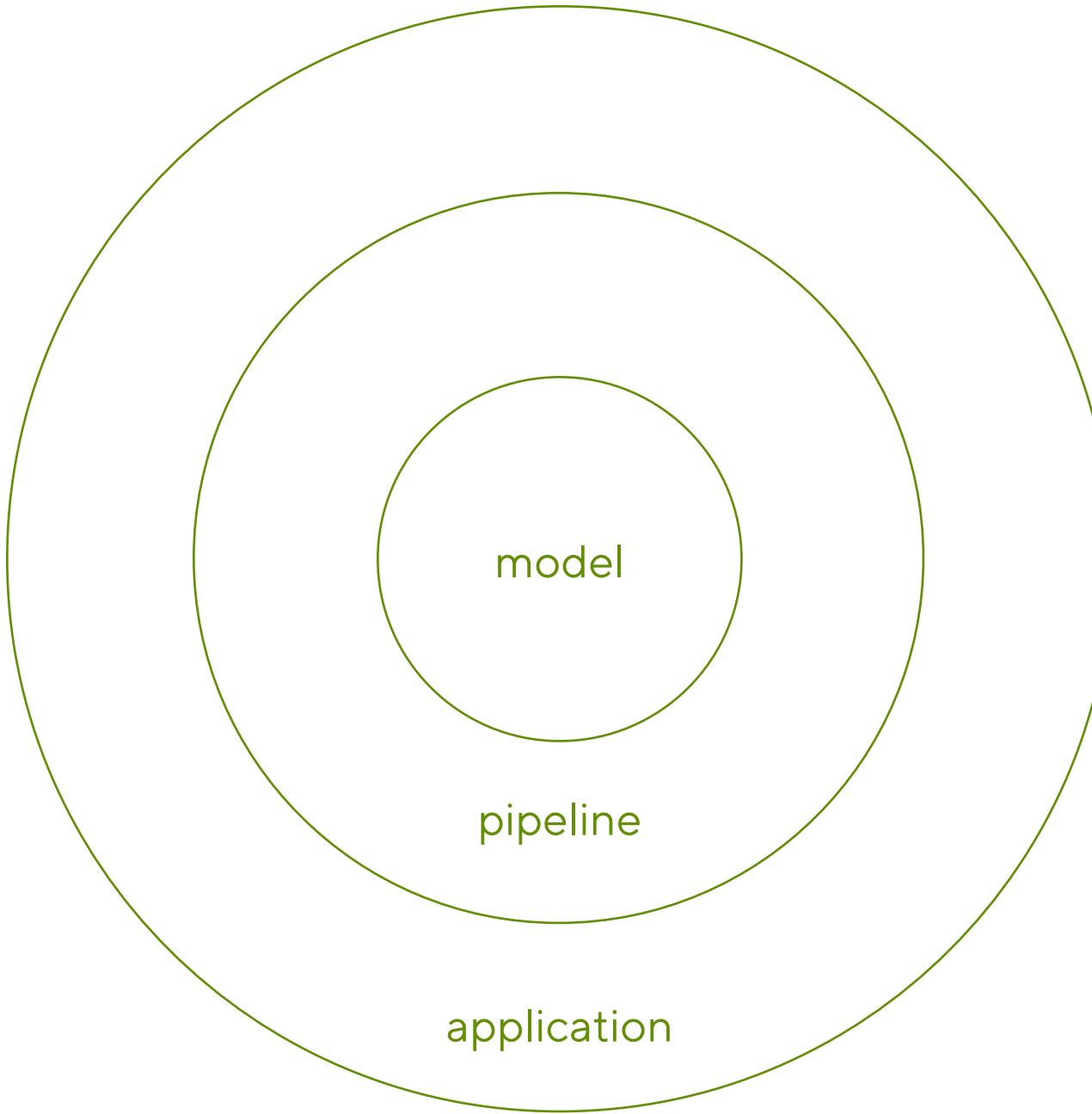
Решение

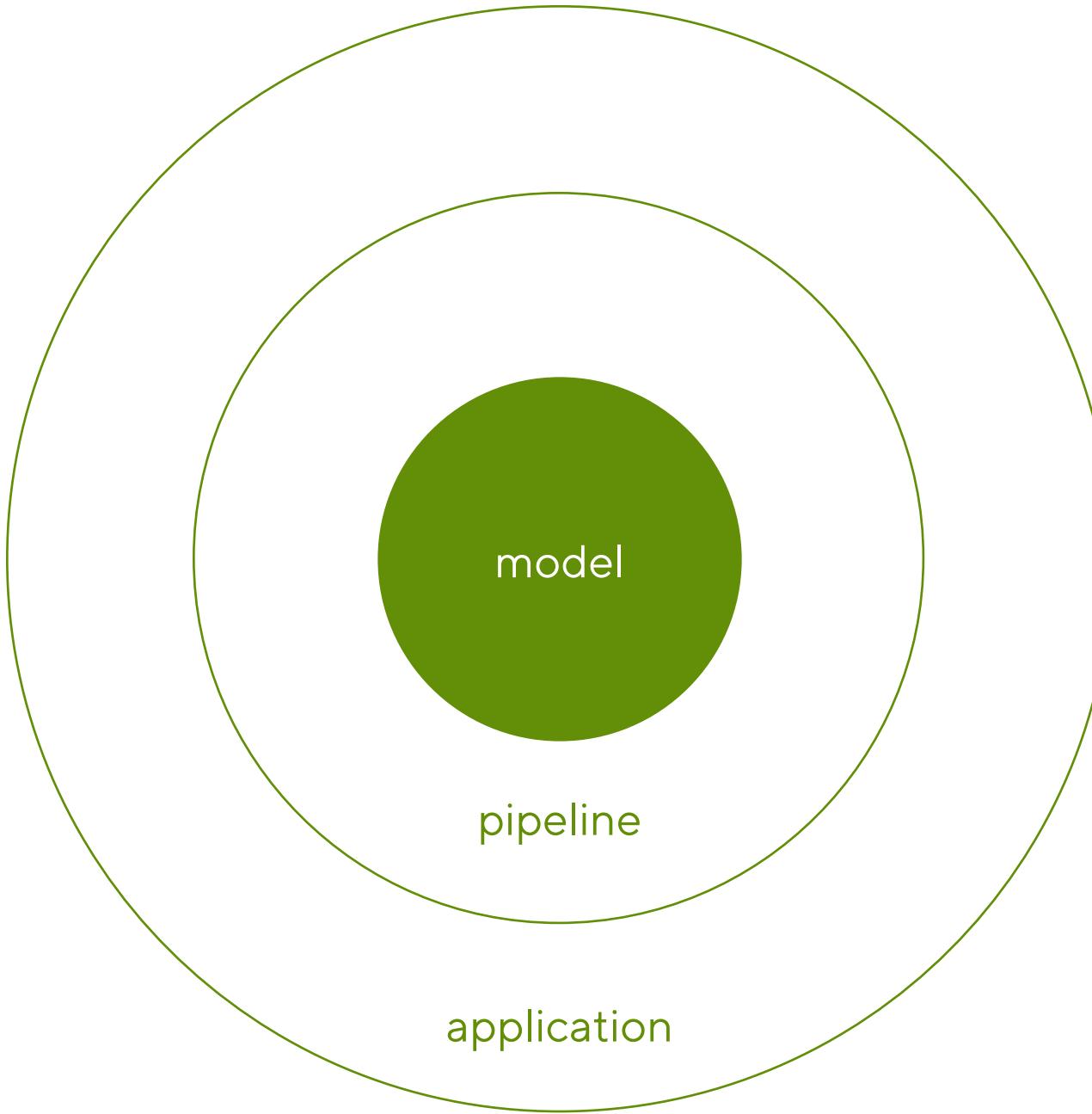
Понимание ожиданий

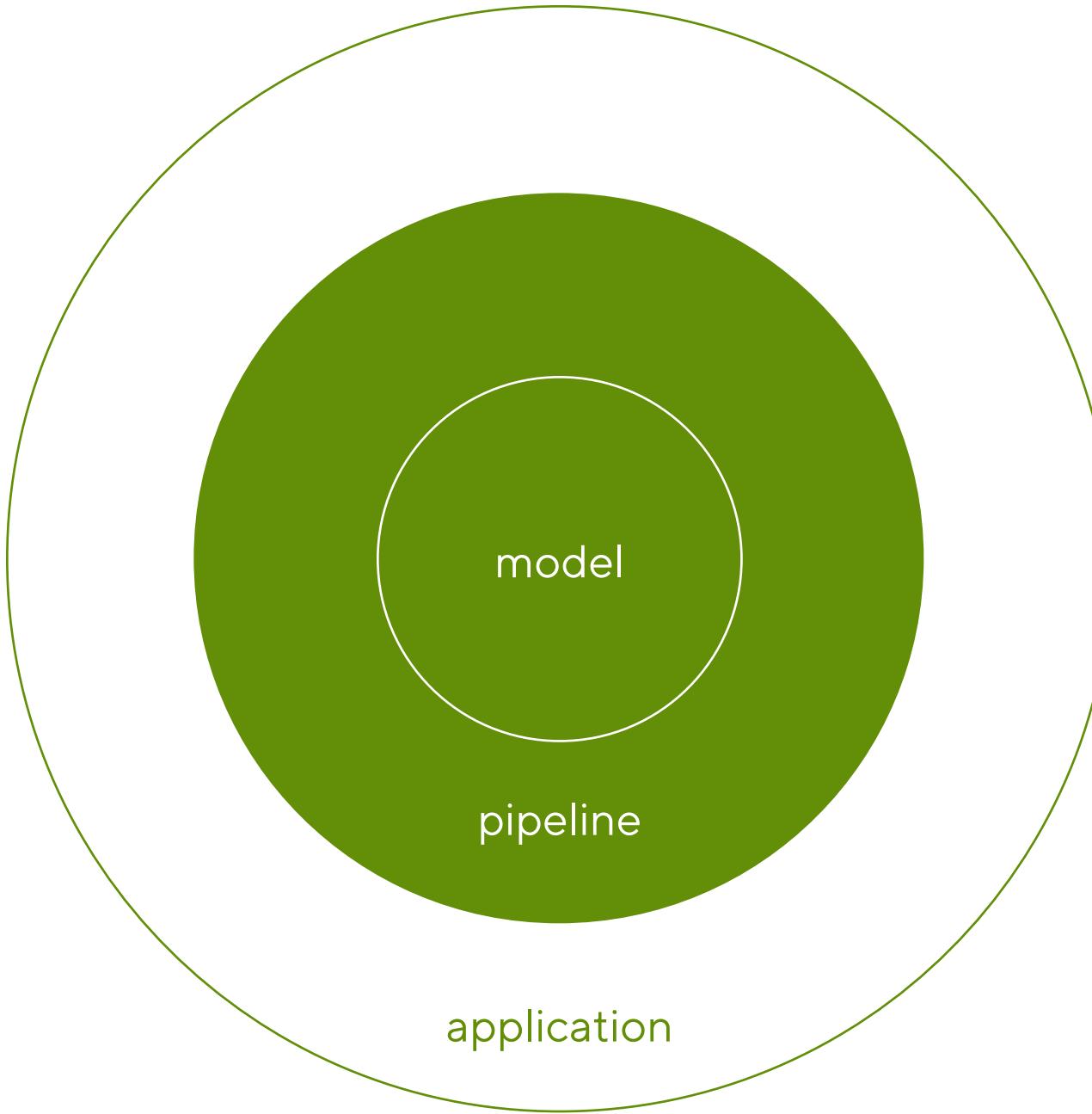


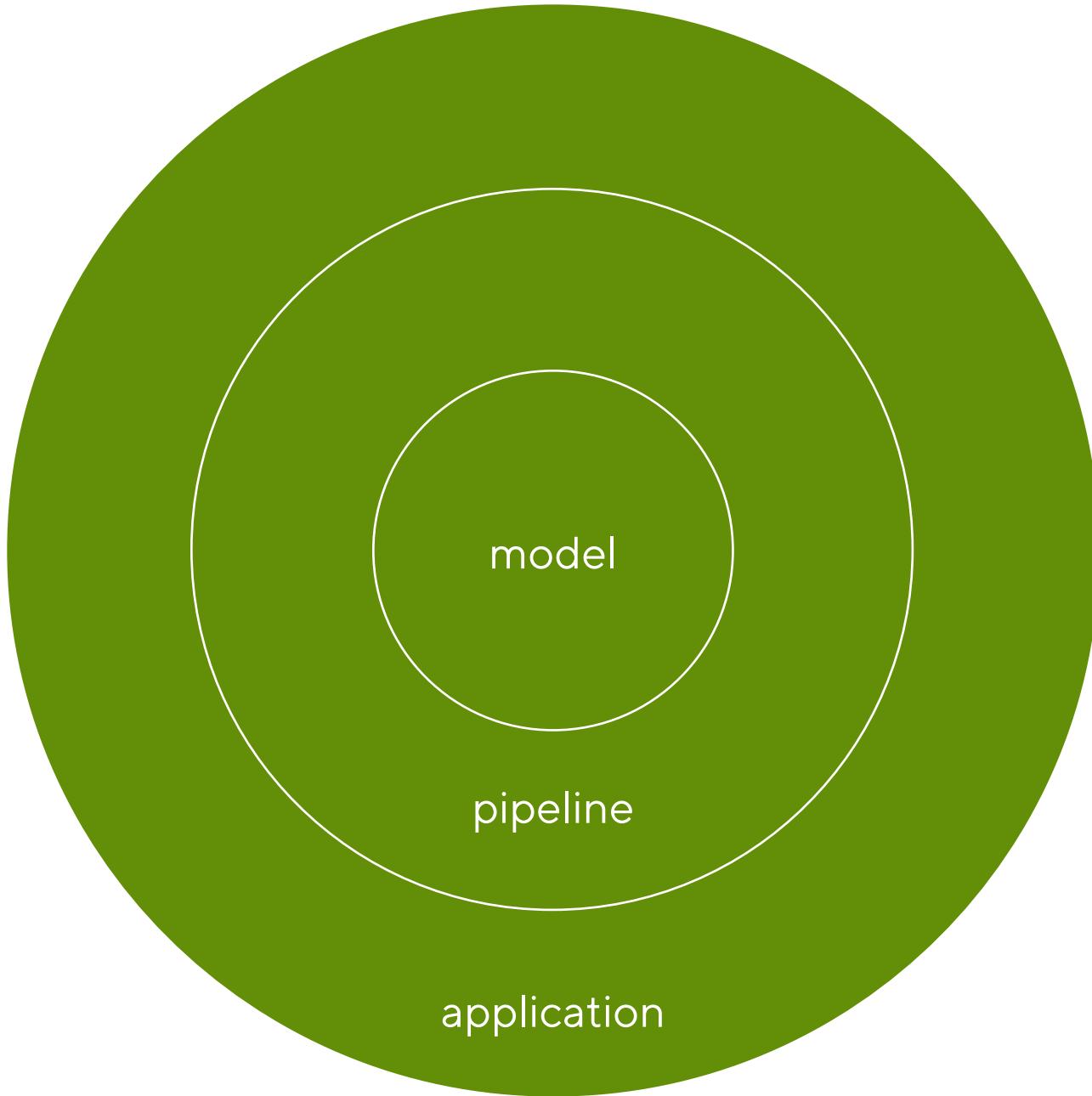
model











Воркфлоу

Воркфлоу

1. Основан на CRISP-DM (Cross Industry Standard for Data Mining).

Воркфлоу

1. Основан на CRISP-DM (Cross Industry Standard for Data Mining).
2. CRISP-DM + стандартный воркфлоу разработки FunBox очень похожи на ASUM-DM.

Воркфлоу

1. Основан на CRISP-DM (Cross Industry Standard for Data Mining).
2. CRISP-DM + стандартный воркфлоу разработки FunBox очень похожи на ASUM-DM.
3. Это не waterfall, это всё равно agile.

CRISP-DM



Понимание бизнес-целей

1. Опиши задачу.
-

Мы хотим автоматизировать процесс классификации кошечек и доге.



Понимание бизнес-целей

2. Опиши как и кем задача выполняется сейчас, если решается.
-

У нас есть контент-менеджер, который классифицирует вручную 500 фотографий в день и распределяет по двум папкам.



× 500

Понимание бизнес-целей

3. Определи критерии успешности.
-

Мы хотим увеличить количество классифицируемых фотографий до 50 000 в день при точности более 85%.

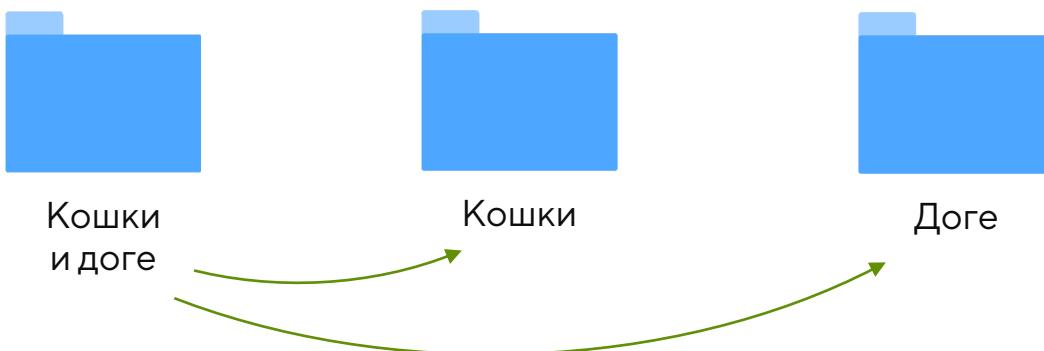


× 50 000

Понимание бизнес-целей

- Укажи формат результатов эксперимента.
-

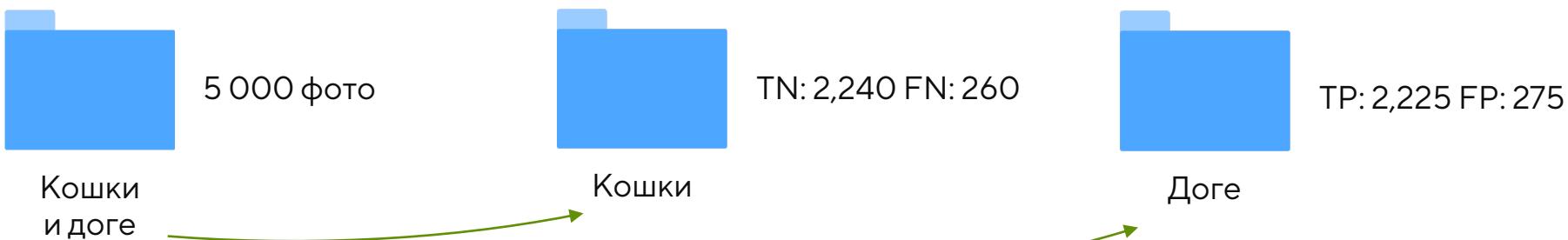
Нам нужен скрипт на Python 3, который в качестве аргумента принимает папку с фотографиями формата .jpg и .jpeg, классифицирует их и перемещает в две папки: кошки и дого.



Понимание бизнес-целей

5. Опиши алгоритм проверки результатов.
-

Менеджер запустит скрипт по одной из папок, а дальше контент-менеджер посмотрит и вручную посчитает ошибки.



Понимание бизнес-целей. Итог

Работа менеджера:

1. Опиши бизнес-задачу.
2. Опиши как и кем задача выполняется сейчас, если решается.

Работа менеджера совместно с командой:

1. Определи критерии успешности.
2. Укажи формат результатов эксперимента.
3. Опиши алгоритм проверки результатов.

CRISP-DM



Исследование данных

1. Получи изначальные данные.

У нас есть 50 000 фотографий, которые:

- похожи на те, что мы будем классифицировать в будущем;
- уже проklassифицированы вручную;
- их можно забрать с ftp.

Обычно нам ещё нужно указать базы данных, таблицы, столбцы, откуда можно забрать датасет, а также указать его период

Исследование данных

2. Проверь качество данных и подготовь нужные преобразования.

Все наши фотографии:

- имеют только один объект на фотографии;
- отображают либо кошечку, либо дого;
- не обрабатывались специфически, кроме кадрирования;
- имеют расширения .jpg или .jpeg.
- имеют разрешение 500×500 пикселей.

CRISP-DM



Подготовка данных

1. Подготовь трансформации в данных.

С текстовыми или табличными данными обычно нужно провести достаточно много трансформаций:

- заполнить пропуски (NaN);
- обработать выбросы;
- дискретизировать признаки (binning);
- использовать one-hot encoding и т. д.

Подготовка данных

2. Раздели датасет на train / test.

Можно применить много разных техник:

- train / test split;
- кросс-валидация (k-folds, LOOCV).

CRISP-DM



Создание моделей

1. Создай / переиспользуй модели или используй другой подход.

- попробуй не ML алгоритмы: regexp, if-else, математика и т. д.
- натренируй несколько моделей (можно стартовать с совсем базовых) со стандартными гиперпараметрами;
- настрой гиперпараметры;
- выяви наиболее важные признаки в данных;
- do other ML stuff.

Создание моделей

2. Не забудь загружать промежуточные модели на платформу для трекинга результатов экспериментов.

- MLFlow;
- DVC;
- Sacred;
- ModelDB;
- Neptune.

CRISP-DM



Оценка

1. Оцени результаты по выбранным метрикам.
-

Выбрать метрику весьма сложно, нужно отталкиваться от задачи:

- метрики в задачах классификации: accuracy, precision, recall, F1-score, ROC, AUC, log loss etc.
- метрики в задачах регрессии: MSE, MAE, r2.

Оценка



5 000 фото

Кошки
и догое



TN: 2,240 FN: 260

Кошки



TP: 2,225 FP: 275

Догое



Оценка



5 000 foto

Кошки
и догое



TN: 2,240 FN: 260

Кошки



TP: 2,225 FP: 275

Доге

ACCURACY: **89,3%**



CRISP-DM



Внедрение

1. Прими решение об успешности эксперимента и о том, должна ли модель быть внедрена в продакшн.

Внедрение

1. Прими решение об успешности эксперимента и о том, должна ли модель быть внедрена в продакшн.
2. Заполни репорт об эксперименте с результатами по каждому этапу.

Внедрение

1. Прими решение об успешности эксперимента и о том, должна ли модель быть внедрена в продакшн.
2. Заполни репорт об эксперименте с результатами по каждому этапу.
3. Запланируй дальнейшее внедрение: создание API или другого интерфейса для обращения к пайплайну.

CRISP-DM – регулярный процесс



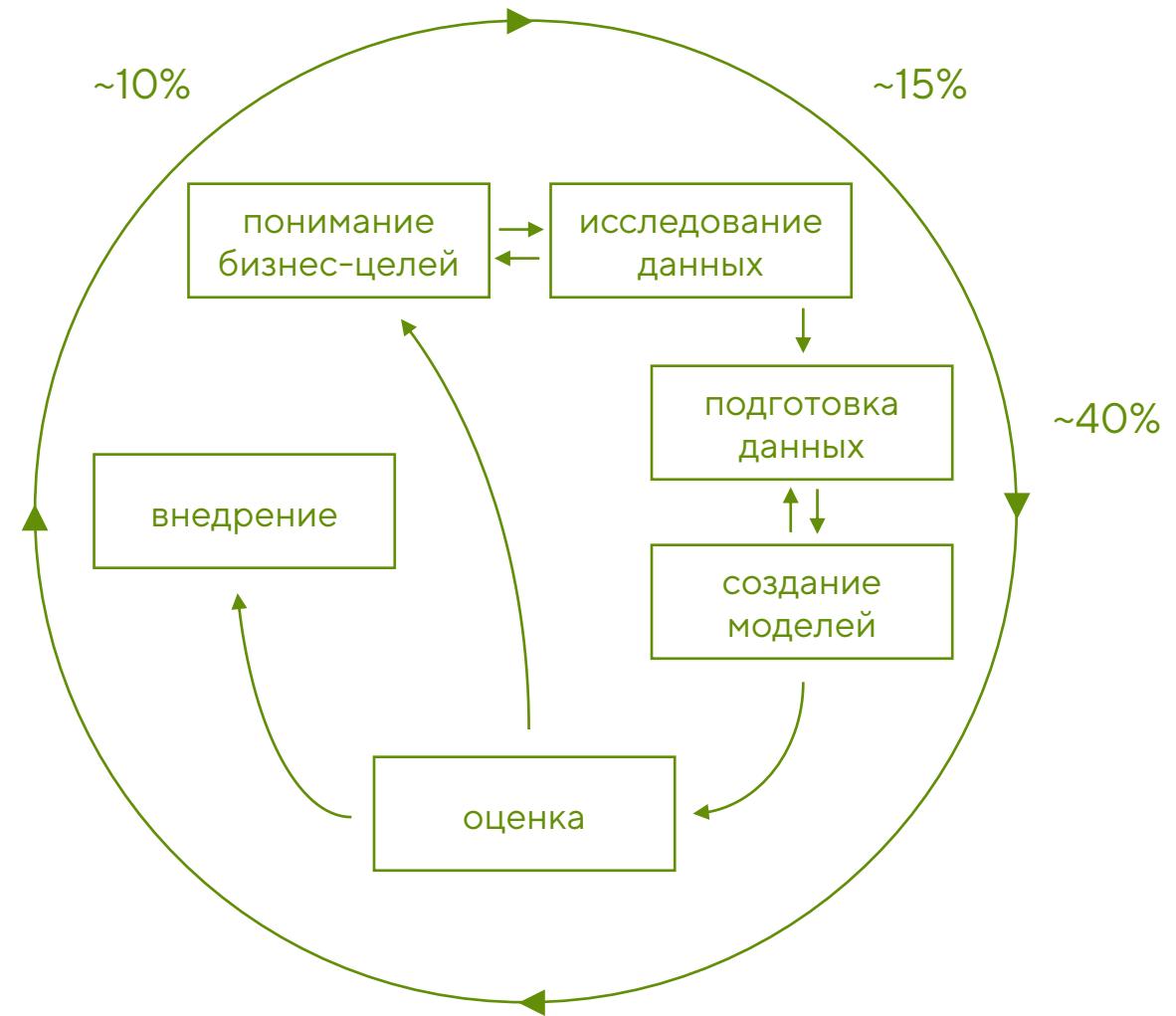
Временные траты



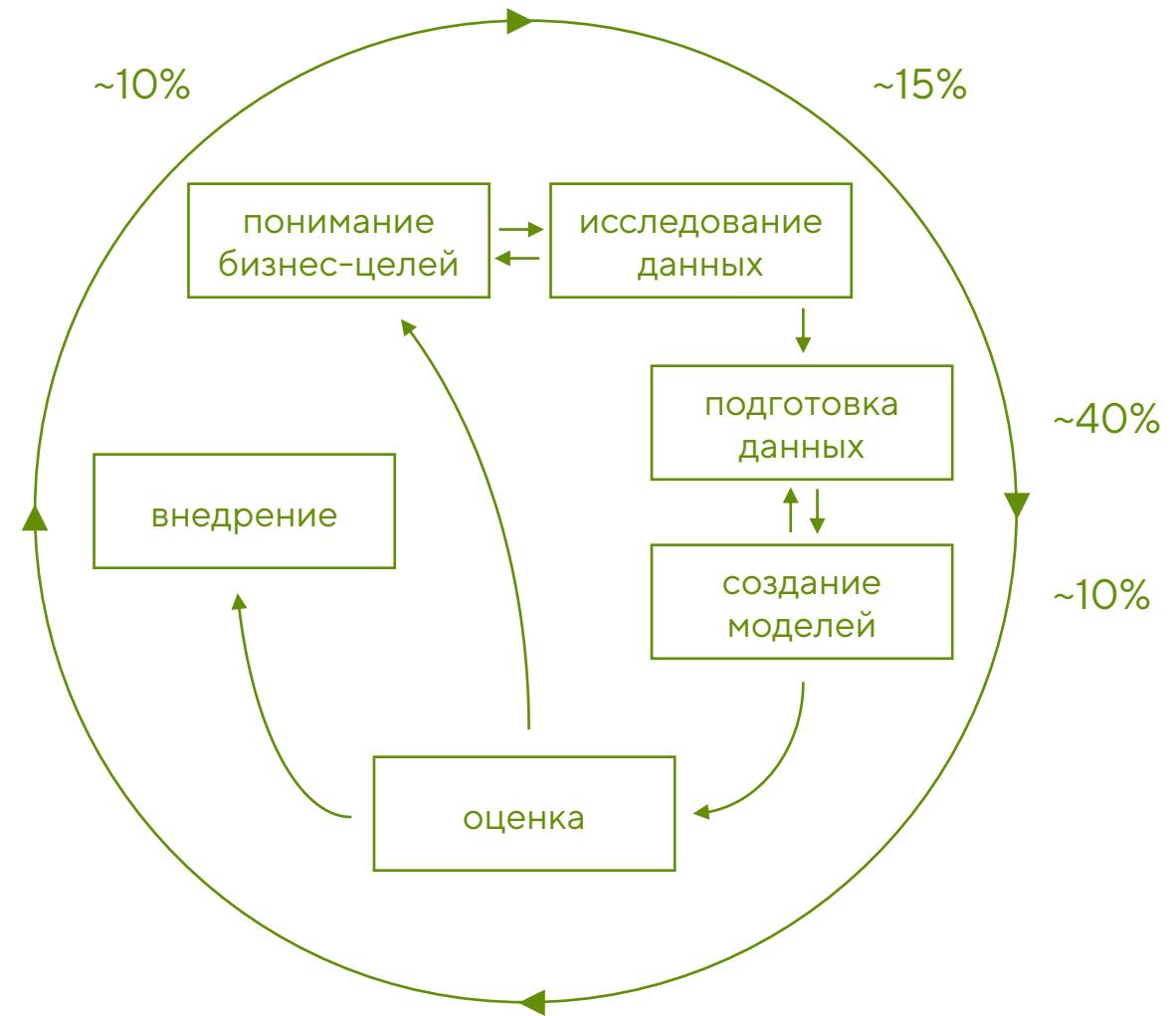
Временные траты



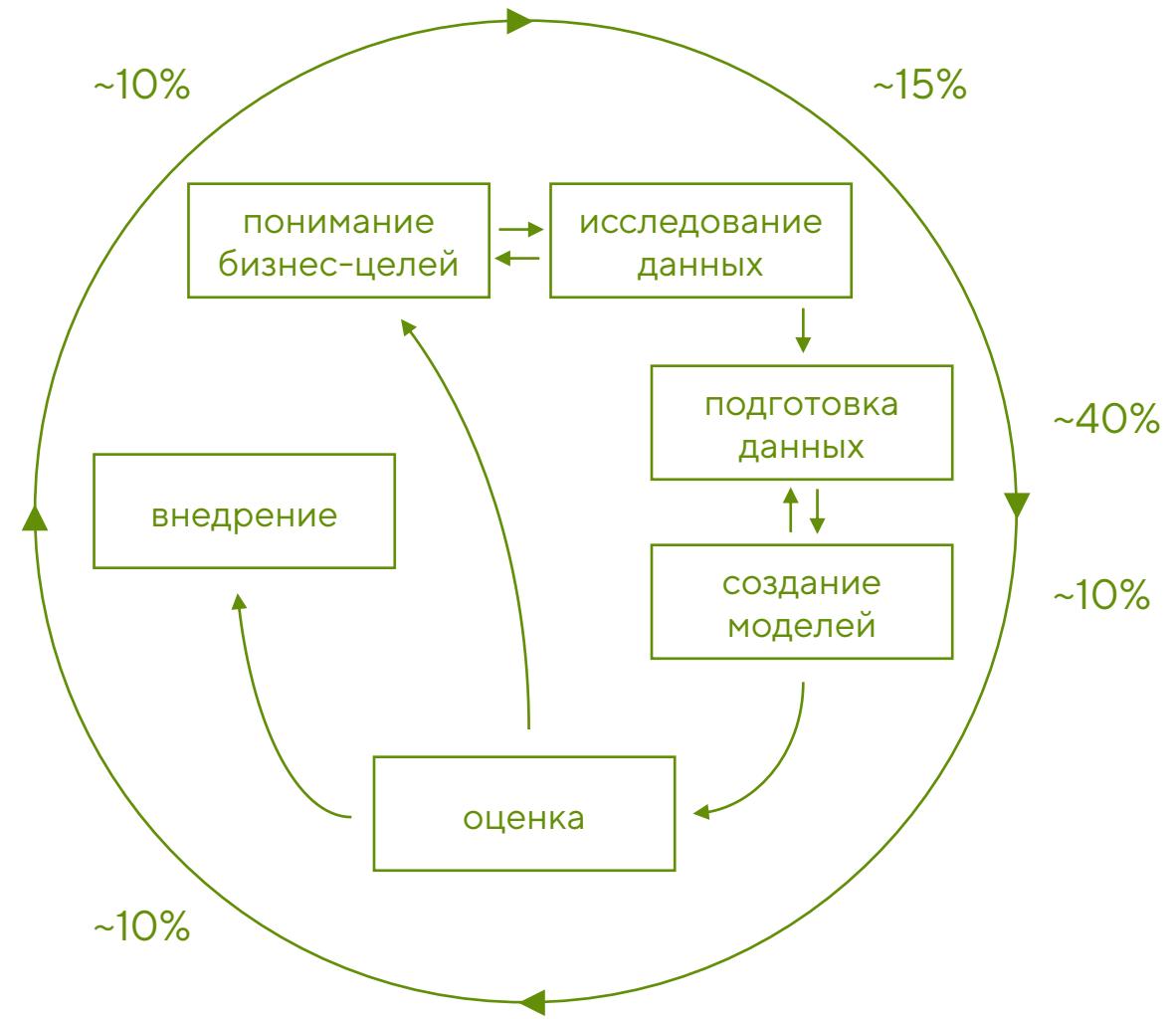
Временные траты



Временные траты



Временные траты



Временные траты



Благодарности



Максим Самсонов
Machine learning,
data science, R&D



Дина Ефремова
Data Scientist
& Data Engineer



Антон Никитин
Head of A2P & DMP

Вопросы?



Сергей Зотов

Technical Product /
Platform Owner