



# Жизненный цикл DS проектов

NewProLab, осень 2018 г.  
Олег Хомюк



## **Олег Хомюк**

oleg.khomyuk@gmail.com

telegram: @khomyuk

<https://www.linkedin.com/in/olegkhomyuk>

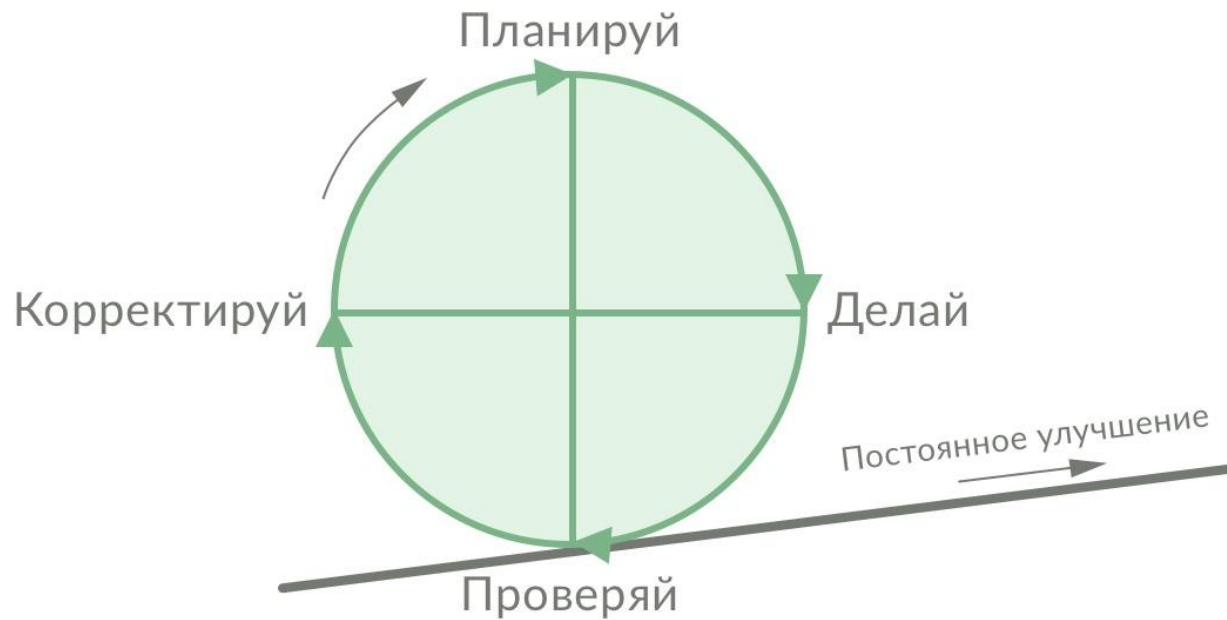
- Yandex, Consultant Plus, Ezhome
- Lamoda, Head of Research & Development



# CRISP-DM

Cross Industry Standard Process for Data Mining

- Бизнес-анализ (Business understanding)
- Анализ данных (Data understanding)
- Подготовка данных (Data preparation)
- Моделирование (Modeling)
- Оценка результата (Evaluation)
- Внедрение (Deployment)



**Business  
Understanding/  
Бизнес-анализ**

Determine Business  
Objectives/  
Определение  
бизнес-целей

Assess Situation/  
Оценка текущей  
ситуации

Determine Data  
Mining Goals/  
Определение целей  
аналитики

Produce Project Plan/  
Подготовка плана  
проекта

**Data  
Understanding/  
Анализ данных**

Collect Initial Data/  
Сбор данных

Describe Data/  
Описание данных

Explore Data/  
Изучение данных

Verify Data Quality/  
Проверка качества  
данных

**Data Preparation/  
Подготовка  
данных**

Select Data/  
Выборка данных

Clean Data/  
Очистка данных

Construct Data/  
Генерация данных

Integrate Data/  
Интеграция данных

Format Data/  
Форматирование  
данных

**Modeling/  
Моделирование**

Select Modeling  
Techniques/  
Выбор алгоритмов

Generate Test  
Design/  
Подготовка плана  
тестирования

Build Model/  
Обучение моделей

Assess Model/  
Оценка качества  
моделей

**Evaluation/  
Оценка решения**

Evaluate Results/  
Оценка результатов

Review Process/  
Оценка процесса

Determine Next  
Steps/  
Определение  
следующих шагов

**Deployment/  
Внедрение**

Plan Deployment/  
Внедрение

Plan Monitoring and  
Maintenance/  
Планирование  
мониторинга и  
поддержки

Produce Final Report/  
Подготовка отчета

Review Project/  
Ревью проекта

# 1. Бизнес-анализ / Business understanding

- Бизнес-цель проекта  
(заказчик, бюджет, бизнес-цель, чем не устраивает текущее решение)
- Аудит текущей ситуации  
(ресурсы - железо, инфраструктура, доступность данных, эксперты по предметной области, анализ текущего решения, риски)
- Цели по аналитике  
(метрики качества, критерии приемки / успешности)
- План проекта  
(оценка всех этапов, сроки, роли, команда, ответственные)

# В чем сложность этапа постановки задачи

## Необходимо:

- собрать полную информацию о бизнес задаче
- корректно конвертировать ее в математическую постановку

## Ошибки и неточности на этом этапе

- могут весьма драматическим образом сказаться на результате
- к сожалению, не редкость.

## Трудности перевода:

В реальности существует колоссальный разрыв между тем, что нужно бизнесу, и тем, что привыкли делать аналитики, data scientist-ы и математики.

# В чем сложность этапа постановки задачи

## **Бизнес-задача:**

- Сформулированная задача, позволяющая достигать цели компании
- Требуется экспертных знаний в предметной области
- Во многих случаях успех измеряется в деньгах

## **Математическая постановка:**

- Постановка в терминах анализа данных
- Требуется экспертизы в математике и машинном обучении
- Успех измеряется численно (точность, полнота)



Что же делать

Работать над постановкой задачи в формате  
кросс-функциональной команды и делиться  
экспертизой!

# Чек-лист по постановке задачи

## Вводные:

- Какой процесс хотим оптимизировать? Как он работает?
- Где мы видим точки роста / уязвимости? Что хотим улучшить? Какие есть идеи?

## Проработка:

- Потенциал в плане экономического эффекта
- Как и где будет использоваться модель?
- Как оценить экономический эффект в случае внедрения? Как будем понимать, что проект успешен?

Финальное решение о старте проекта  
рекомендуется принимать после **полной**  
**проработки** постановочной части!

# Постановка задачи. Кейс

На входе:

- Нужно сделать модель, прогнозирующую продажи товаров на следующую неделю

Какую метрику взять?

- MAE, MSE, RMSE, MAPE, sMAPE

Разные последствия для бизнеса от:

- Недопрогноза
- Перепрогноза

А стоит ли вообще браться за этот проект?

Перед тем, как приступить к следующим этапам надо оценить экономический потенциал проекта! \*

# Кейс по оттоку

Что хотим оптимизировать?

Входные данные:

- получаем с пользователя  $X$  рублей за все его «время жизни» ( $X$  обычно называют LTV)
- добиться того, чтобы он был с нами нам стоит  $Y$  рублей

Вывод:

- с каждого потраченного рубля мы получили в  $X/Y$  раз больше и логично это отношение максимизировать.

# Кейс по оттоку

Что хотим оптимизировать?

*Хотим оптимизировать стоимость для нас денег,  
заработанных на пользователе*

# Кейс по оттоку

Экономический эффект на одного пользователя

$$MQ * Z * ARPU - COST$$

**MQ** - качество модели (доля правильно угаданных отточников)

**Z** - успешность удержания

**ARPU** - средняя выручка на пользователя

**COST** - стоимость удержания одного пользователя



## 2. Анализ данных / Data understanding

- Сбор данных  
(собственные / сторонние / потенциальные)
- Описание данных  
(ключи, объемы, доступность, возможные значения, статистики)
- Исследование данных  
(основные статистики, гипотезы, какие данные помогут решить задачу)
- Качество данных  
(пропущенные значения, опечатки / ошибки, противоречия)

# Оценка доступных данных

- Какие данные доступны?
- Есть ли историчность? За какой период? (для выявления сезонности нужно >2 года)
- Есть ли возможность использовать данные совместно (ключи)
- Есть ли нужный для задачи сигнал в данных
- Будет ли модель потом работать на live данных в production

### 3. Подготовка данных / Data preparation

- Отбор данных  
(отбор релевантных данных, полезных для решения задачи)
- Очистка данных  
(удаление / обработка пропусков, ошибок, кодировки, шумов)
- Генерация новых данных  
(построение новых признаков из имеющихся данных)
- Интеграция данных и форматирование  
(объединение данных из разных источников)

## 4. Моделирование / Modeling

- Выбор алгоритмов  
(сложные / простые, учет специфики задачи)
- Планирование тестирования  
(кросс-валидация, train/test/validation, подбор гипер-параметров)
- Обучение моделей  
(непосредственное написание программного кода для обучения и валидации и его запуск)
- Оценка результатов обучения  
(выбрать лучшие модели, провести анализ качества, принять решение о готовности к внедрению)

## 5. Оценка результата / Evaluation

- Оценка результатов моделирования  
(насколько хорошо модель решает бизнес-задачу)
- Ретроспектива по проекту  
(разбор полетов, возникшие проблемы, можно ли было что-нибудь сделать лучше / быстрее / эффективнее?)
- Определение следующих шагов  
(внедряем или нет, если да, то какую модель и куда. Надо ли строить новый сервис?)

# Отличие модели от сервиса

- Оффлайн моделям могут быть доступны любые данные, которые вы подготовите, даже те, что сложно получать в реальном времени
- Качество работы сервиса естественнее измерять в бизнес показателях, моделей - в ML метриках
- Сервис реализует действие, которое рекомендует модель - например, сервис автоматических рассылок

# Мониторинг качества решения

За чем надо следить?

- Изменилось ли качество модели?
- Изменилось ли распределение во входящих данных?
- Триггеры для поддержки качества (нужно отличать случайные изменения качества и “протухание”)

Автоматизация:

- Расчет триггеров
- Регулярное обновление моделей (расписание / триггеры)

## 6. Внедрение / Deployment

- Развертывание  
(определение вида конечного решения / сервиса, внедрение)
- Настройка мониторинга модели  
(мониторинг качества модели, протухание, частота переобучения)
- Подготовка отчета  
(отчет по проекту)
- Ревью проекта  
(финальный отчет по результатам внедрения)



**Business Understanding/  
Бизнес-анализ**

Determine Business Objectives/  
Определение бизнес-целей

Assess Situation/  
Оценка текущей ситуации

Determine Data Mining Goals/  
Определение целей аналитики

Produce Project Plan/  
Подготовка плана проекта

**Data Understanding/  
Анализ данных**

Collect Initial Data/  
Сбор данных

Describe Data/  
Описание данных

Explore Data/  
Изучение данных

Verify Data Quality/  
Проверка качества данных

**Data Preparation/  
Подготовка данных**

Select Data/  
Выборка данных

Clean Data/  
Очистка данных

Construct Data/  
Генерация данных

Integrate Data/  
Интеграция данных

Format Data/  
Форматирование данных

**Modeling/  
Моделирование**

Select Modeling Techniques/  
Выбор алгоритмов

Generate Test Design/  
Подготовка плана тестирования

Build Model/  
Обучение моделей

Assess Model/  
Оценка качества моделей

**Evaluation/  
Оценка решения**

Evaluate Results/  
Оценка результатов

Review Process/  
Оценка процесса

Determine Next Steps/  
Определение следующих шагов

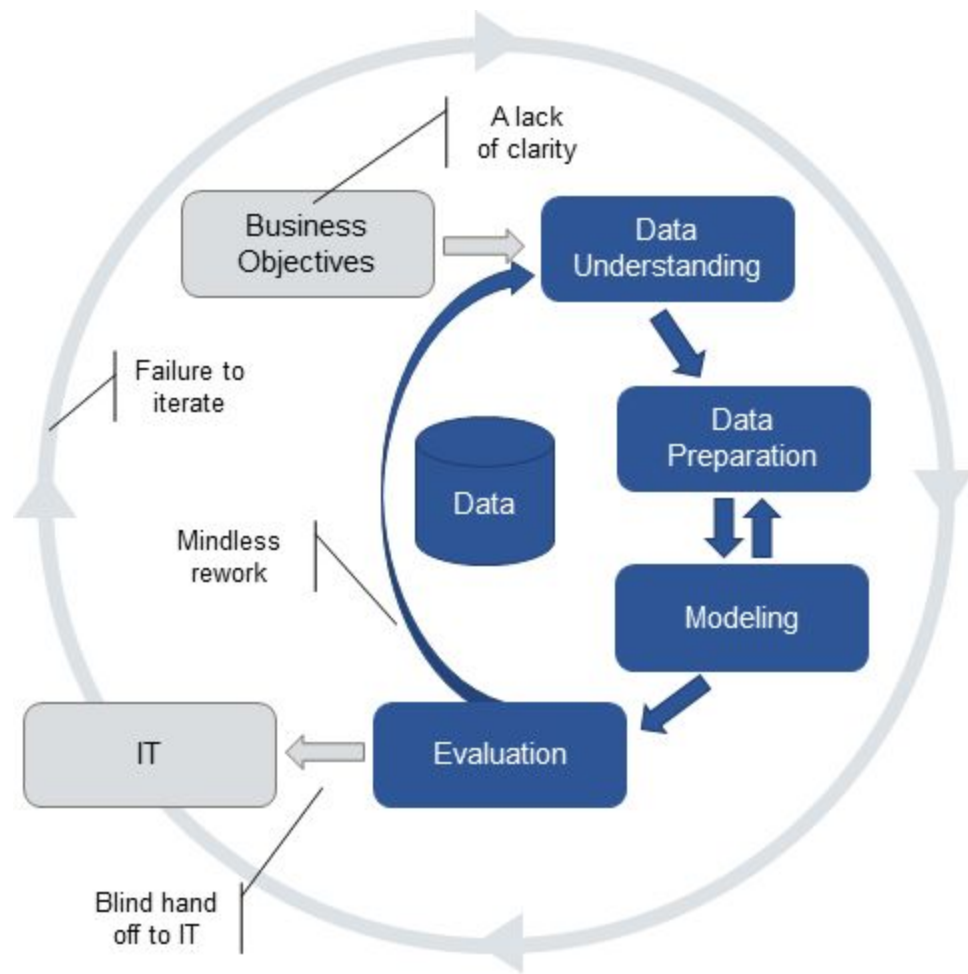
**Deployment/  
Внедрение**

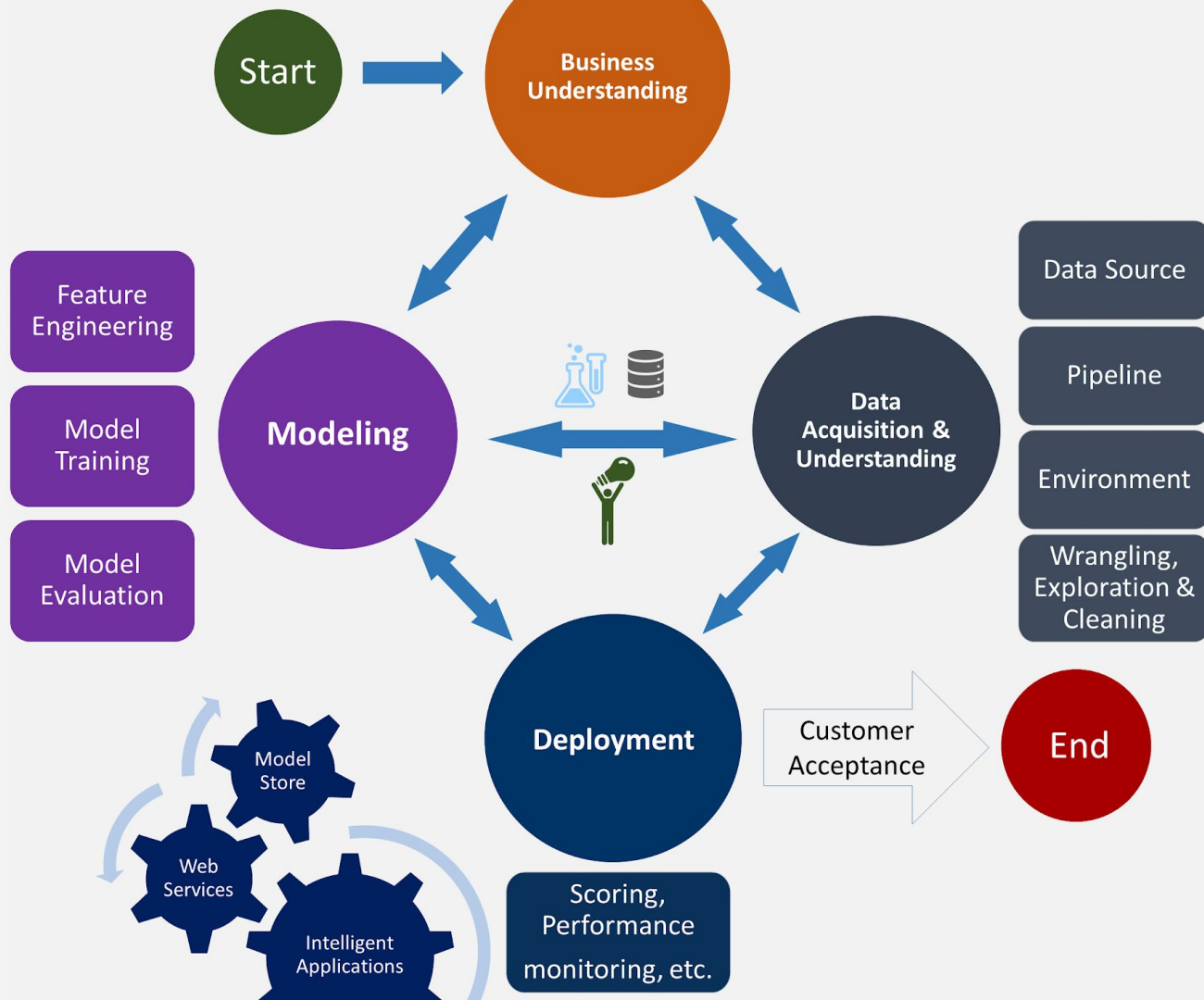
Plan Deployment/  
Внедрение

Plan Monitoring and Maintenance/  
Планирование мониторинга и поддержки

Produce Final Report/  
Подготовка отчета

Review Project/  
Ревью проекта





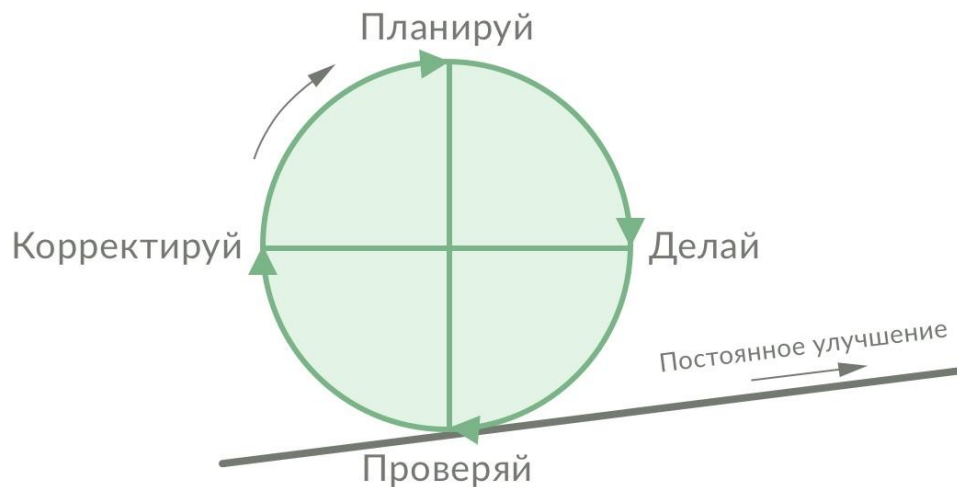
# Какие компетенции могут понадобиться

- Product / Project Manager
- Бизнес аналитик
- Data Scientist
- Data Engineer / Software Developer
- Server administrator / DevOps

Кроме этого:

- Эксперты в предметной области
- Команды сервисов и IT-систем, с которыми необходима интеграция

# Дальнейшая поддержка решения



Спасибо за внимание!