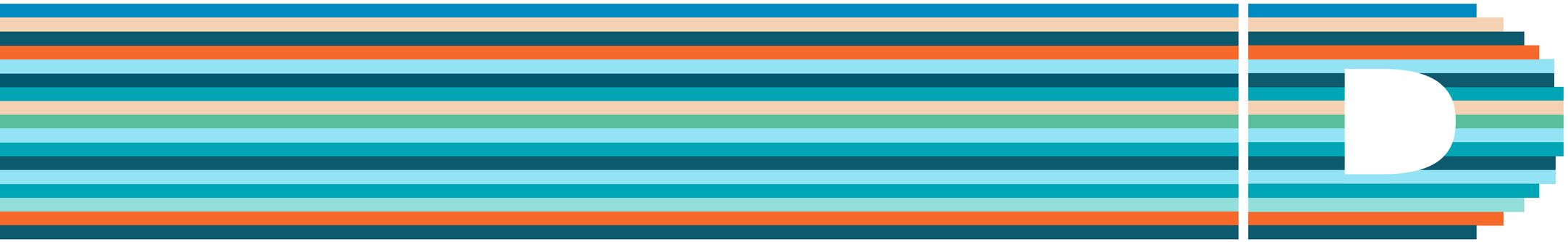
Тема 1. Методологии Data Mining

Полина Басина
Аналитик Центра анализа больших данных ТГУ



data-diving

академия аналитики данных при Томском государственном университете

Национальный исследовательский
Томский государственный университет

Наш кейс

Клиент N много лет занимается строительством домов

- Стоимость услуг: средняя по рынку
- Аудитория: люди со средним уровнем дохода в городе Т — месте присутствия клиента

Клиент собирается развивать новую нишу бизнеса — строительство домов премиум-сегмента, цена которых в разы превосходит его имеющийся ассортимент.

Заказчик знает, как работать с товарами среднего класса, но не понимает, кто может заинтересоваться товаром в новой нише.

О чем поговорим?

- С чего начать?
- Из каких этапов будет состоять наша работа?
- С какими данными мы будем иметь дело?
- Что нужно обязательно учесть и какой результат мы хотим получить?

Выбор методологии

Основные задачи

1. **Определить цель анализа данных**
2. **Определить требования к результату исследования** — то есть тот результат, который хотим получим мы и наш заказчик
3. **Понять — какие типы данных** нам нужны
4. **Определиться с источниками и способами их получения**
5. **Выбрать методы и инструменты** анализа данных
6. **Оценить риски и условия реализации** проекта

ПЛАН ИССЛЕДОВАНИЯ

Технологии интеллектуального анализа данных или Data Mining

Data Mining — процесс обнаружения в данных ранее **неизвестных, нетривиальных, практически полезных и доступных интерпретации** знаний, необходимых для принятия решений в различных сферах человеческой деятельности

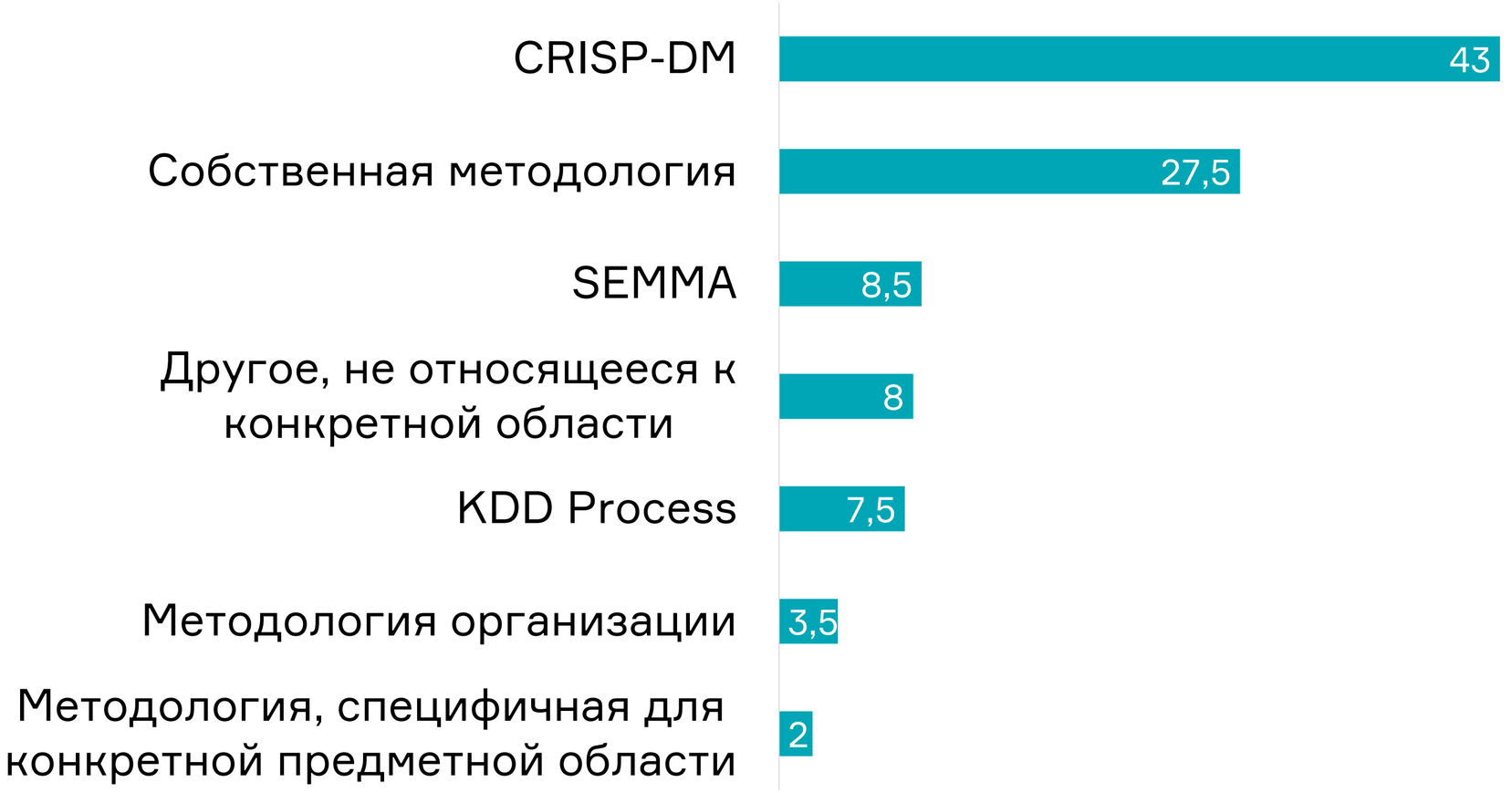
Как выглядят данные?

Строки — наблюдения, индивиды

Столбцы — переменные, признаки

Возврат кредита в срок	Возраст	Пол	Средний уровень дохода в тыс. рублей
Да	40	М	45
Нет	50	М	20
Да	60	Ж	31
Да	48	Ж	165
Нет	65	М	41

Популярные методологии Data Mining



Источник рисунка: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Как выбрать?

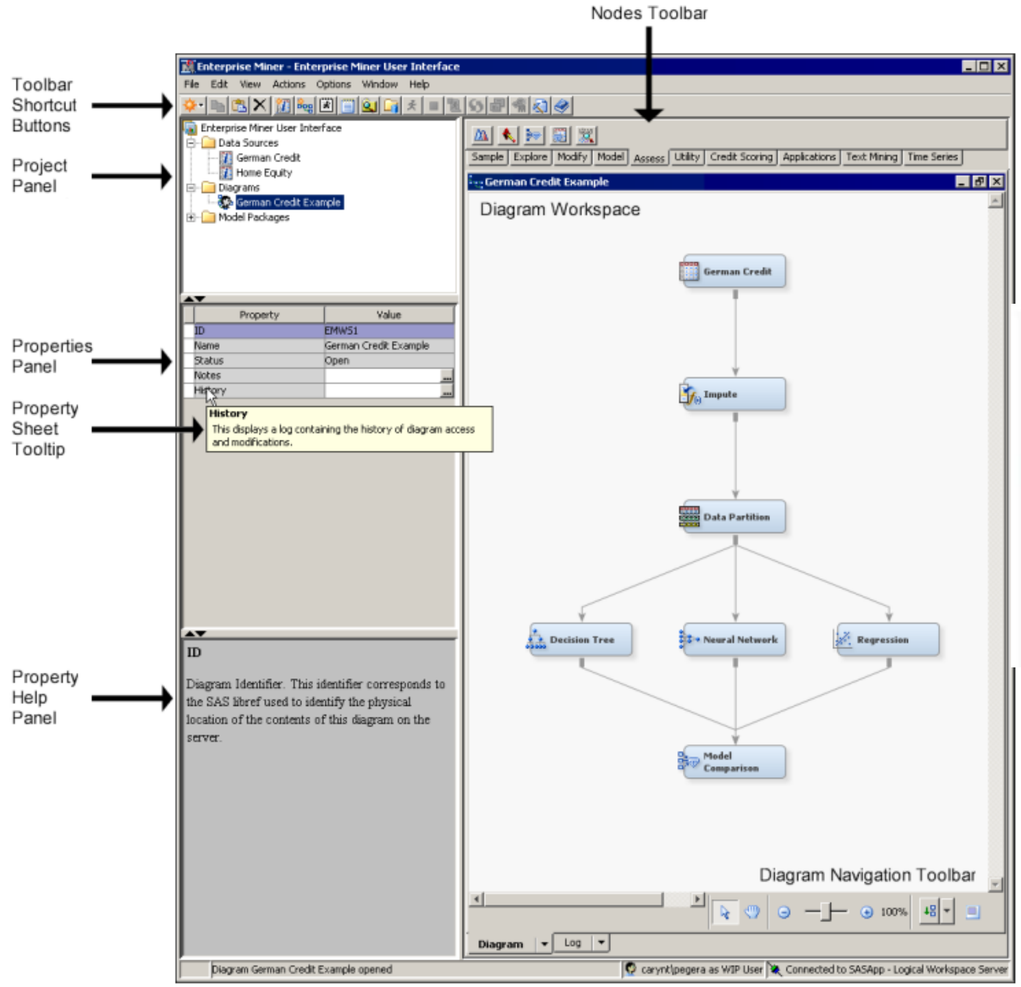
- Условия работы
- Зона ответственности
- Требуемый результат

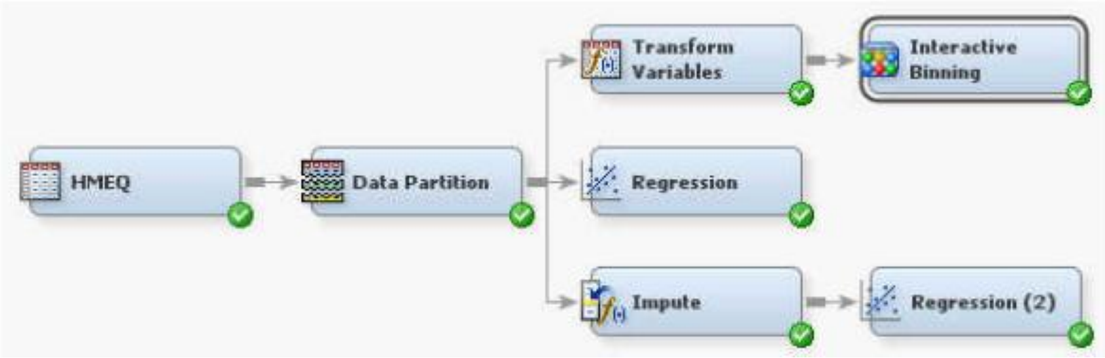
Для чего?

- Сохранение опыта и воспроизводимость проектов
- Упрощение процесса планирования и управления
- Простота включения новых членов команды
- Уменьшение зависимости от «лидеров»

Методология SEMMA

Figure 1.1 *SAS Enterprise Miner User Interface*





Источник рисунка: https://go.documentation.sas.com/api/docsets/emcs/14.3/content/emcs.pdf?locale=en

Методология SEMMA

1. **Sample** (отбор данных, т. е. создание выборки)
2. **Explore** (исследование отношений в данных)
3. **Modify** (модификация данных)
4. **Model** (моделирование взаимозависимостей)
5. **Assess** (оценка полученных моделей и результатов)

Методология SEMMA

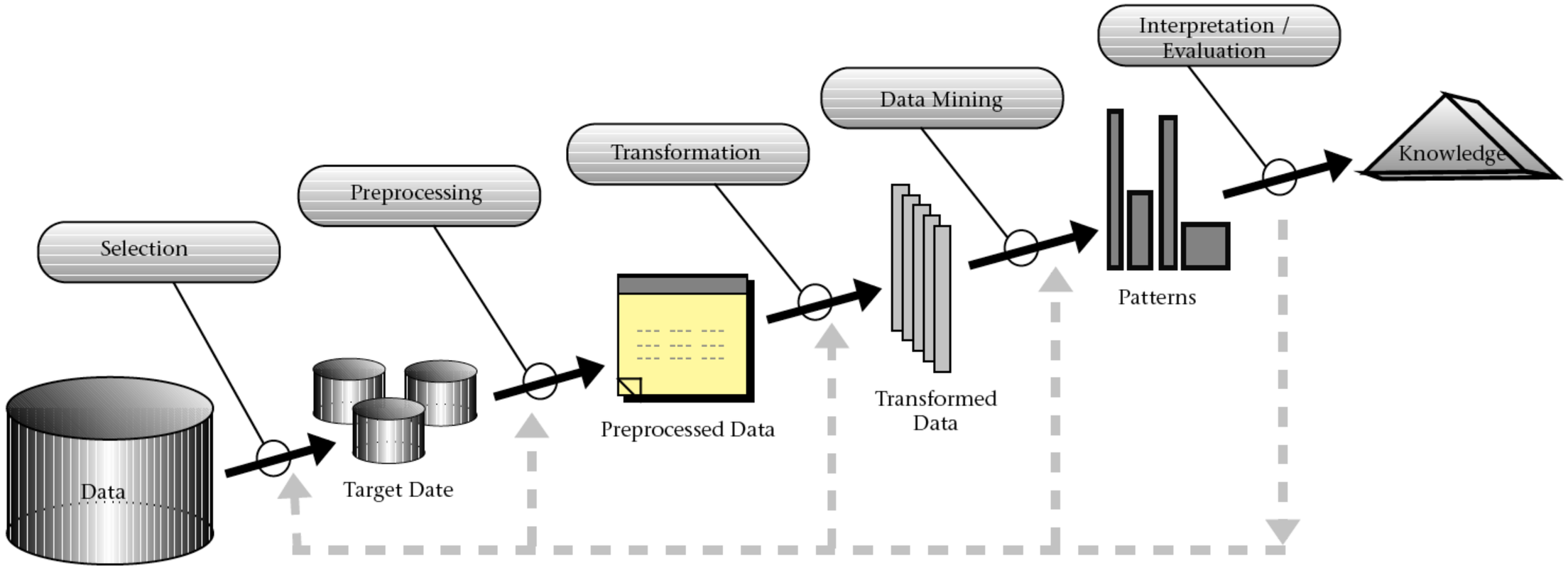
Источник рисунка: https://www.bigdataschool.ru/wiki/semma

Методология KDD (Knowledge Discovery in Databases)

KDD — обнаружение знаний в базах данных

1. **Отбор** (Selection)
2. **Предварительная обработка** (Preprocessing)
3. **Преобразование** (Transformation)
4. **Data Mining**
5. **Интерпретация** (Intepretation | Evaluation)

Методология KDD (Knowledge Discovery in Databases)



Источник рисунка: https://nelsonism.wordpress.com/2014/01/25/manajemen-pertambangan-pada-industri/

Методология Crisp-DM

Этапы жизненного цикла CRISP-DM

1. **Понимание бизнеса** (business understanding)
2. **Понимание данных** (data understanding)
3. **Подготовка данных** (data preparation)
4. **Моделирование** (modeling)
5. **Оценка результатов** (evaluation)
6. **Внедрение** (deployment)
7. **Контроль**

Методология Crisp-DM

Business Understanding / **Бизнес-анализ**	Data Understanding / **Анализ данных**	Data Preparation / **Подготовка данных**	Modeling / **Моделирование**	Evaluation / **Оценка решения**	Deployment / **Внедрение**
Determine Business Objectives / **Определение бизнес-целей** Assess Situation / **Оценка текущей ситуации** Determine Data Mining Goals / **Определение целей аналитики** Product Project Plan / **Подготовка плана проекта**	Collect Initial Data / **Сбор данных** Describe Data / **Описание данных** Explore Data / **Изучение данных** Verify Data Quality / **Проверка качества данных**	Select Data / **Выборка данных** Clean Data / **Очистка данных** Construct Data / **Генерация данных** Integrate Data / **Интеграция данных** Format Data / **Форматирование данных**	Select Modeling Techniques / **Выбор алгоритмов** Generate Test Design / **Подготовка плана тестирования** Build Model / **Обучение моделей** Assess Model / **Оценка качества моделей**	Evaluate Results / **Оценка результатов** Review Process / **Оценка процесса** Determine Next Steps / **Определение следующих шагов**	Plan Deployment / **Внедрение** Plan Monitoring and Maintenance / **Планирование мониторинга и поддержки** Produce Final Report / **Подготовка отчета** Review Project / **Ревью проекта**

Источник рисунка: https://cdto.wiki/Цикл_работы_с_данными

Методология Crisp-DM

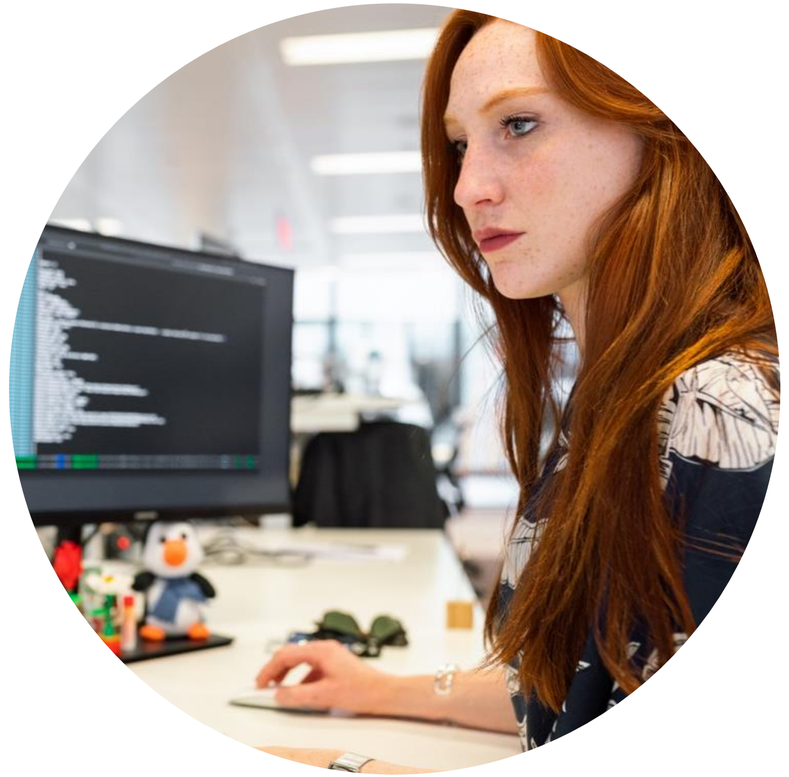
Crisp-DM —мягкая методология: не строгий переход между этапами.

В зависимости от результата этапа принимается решение на какой этапе переходить дальше.

Источник рисунка: https://lafayettefirefighters.com/tr/difference-between-data-mining-and-vs-machine-learning

Преимущество методологии Crisp-DM

Внимание к бизнес-целям компании !



Сравним методологии

KDD	SEMMA	CRISP-DM
—	—	Business Understanding / **Бизнес-анализ**
Selection / **Отбор**	Sample / **Отбор**	Data Understanding / **Анализ данных**
Preprocessing / **Предварительная обработка**	Explore / **Исследование**	
Transformation / **Преобразование**	Modify / **Модификация**	Data Preparation / **Подготовка данных**
Data Mining	Model / **Моделирование**	Modeling / **Моделирование**
Interpretation /Evaluation / **Интерпретация**	Assess / **Оценка**	Evaluation / **Оценка решения**
—	—	Deployment / **Внедрение**

Основные процессы в методологиях

Шаг 1. Работа с заказчиком

Шаг 2. Работа с данными

Шаг 3. Разработка аналитического решения

Практика: выбираем методологию для нашего кейса

1. **Какая наша роль в проекте?** Мы работаем с клиентом от начала и до конца

2. **Это новая отрасль для клиента.** Он еще не понимает и не имеет наработок по интересующему нас вопросу

3. **В качестве решения в первую очередь нам нужен аналитический отчет.** Исследование проводится впервые по данному направлению, соответственно, разрабатывать какие-то модели еще рано

Выбираем методологию для нашего кейса

Методологии SEMMA и KDD не подходят

Используем методологию CRISP-DM
Полный цикл анализа данных, начиная с понимания бизнеса — чего хочет заказчик, и какими аналитическими средствами мы реализуем бизнес-цель

В CRISP-DM мы не зациклены на определенном результате: им может быть как технологическое решение, так и аналитический отчет

Подведем итоги

Эффективные методологии анализа данных — CRISP-DM, SEMMA, KDD.

Методология CRISP-DM полный цикл анализа данных, начиная с понимания бизнеса и внедрения решения.

SEMMA и KDD опора на внутренние аналитические процессы работы с данными.

Методологии очень похожи друг на друга.

Основное отличие в точке «старта» — на каком этапе начинаем работу по анализу данных, **и точке «выхода» —** какой результат получаем.

Общая стратегия анализа — три основных процесса:

1. работа с заказчиком
2. работа с данными
3. разработка аналитического решения

О чем поговорим далее?

- Рассмотрим этапы анализа данных
- Продолжим работу над планированием исследования для нашего заказчика
- Начнем с определения бизнес-цели заказчика

Спасибо за внимание!

Пикабу

Полина Басина
Аналитик Центра анализа больших данных ТГУ