# ЛЕКЦИЯ 1. ВВЕДЕНИЕ В ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ. KDD И CRISP-DM

ДИСЦИПЛИНА: «ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

# ЦЕЛЬ ЛЕКЦИИ

- Сформировать у обучающихся целостное представление об интеллектуальном анализе данных (Data Mining), его задачах, месте в процессе получения знаний из данных, а также познакомить с моделями **KDD** и **CRISP-DM** как основными подходами к организации аналитического процесса.

# ЧТО ТАКОЕ ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ?

- **Интеллектуальный анализ данных (Data Mining)** — это процесс выявления в больших и сложных наборах данных **закономерностей, зависимостей, взаимосвязей, групп, аномалий и других полезных знаний** с использованием методов статистики, машинного обучения, искусственного интеллекта и анализа данных.

# ПОЧЕМУ СЕГОДНЯ НЕОБХОДИМО ИЗУЧАТЬ DATA MINING?

## ИНТЕРНЕТ-МАГАЗИН

Каждый день он получает тысячи и даже миллионы записей:

- кто сделал покупку;
- какой товар был приобретен;
- когда была совершена покупка;
- сколько товаров куплено;
- сколько денег потратил клиент;
- с какого устройства был выполнен заказ;
- в каком регионе находится клиент.

| Клиент | Товар | Количество | Дата |
|---|---|---|---|
| C001 | Хлеб | 1 | 10.09 |
| C001 | Молоко | 1 | 10.09 |
| C001 | Сыр | 1 | 10.09 |
| C002 | Хлеб | 2 | 10.09 |
| C002 | Масло | 1 | 10.09 |

Получается огромное количество данных.

# ГЛАВНАЯ ИДЕЯ:

- **Data Mining превращает большие объемы данных в знания, которые можно использовать для принятия решений.**

# НАПРИМЕР

| Клиент | Товар | Количество | Дата |
|---|---|---|---|
| C001 | Хлеб | 1 | 10.09 |
| C001 | Молоко | 1 | 10.09 |
| C001 | Сыр | 1 | 10.09 |
| C002 | Хлеб | 2 | 10.09 |
| C002 | Масло | 1 | 10.09 |

# ДАННЫЕ → ИНФОРМАЦИЯ → ЗНАНИЯ
## ОДНО ИЗ ВАЖНЫХ РАЗЛИЧИЙ В DATA MINING:

**Данные**

- Это исходные факты и наблюдения.
- Например:
- 1000 клиентов совершили 5000 покупок.

**Информация**

- Данные после обработки и структурирования.
- Например:
- 65% клиентов совершают повторную покупку в течение месяца.

**Знания**

- Выявленная закономерность, которая имеет практический смысл.
- Например:
- Клиенты, которые покупают товар А, имеют высокую вероятность приобрести товар В.

**Решение**

- Практическое действие на основе полученного знания:
- Создать персональную рекомендацию товара B клиентам, купившим товар A.

Таким образом:

- **Данные → Обработка → Информация → Анализ → Знания → Решение**

# ЧТО ТАКОЕ DATA MINING?

Современные организации накапливают огромные объемы данных.

Например:

- банки — транзакции клиентов;
- магазины — покупки;
- университеты — данные студентов;
- больницы — медицинские записи;
- предприятия — данные датчиков;
- социальные сети — публикации и взаимодействия;
- телекоммуникационные компании — данные звонков и интернет-трафика.

Человек не способен эффективно анализировать миллионы записей вручную.

Поэтому используются алгоритмы Data Mining.

- **Data Mining** — это процесс выявления в больших объемах данных полезных закономерностей, взаимосвязей, групп, аномалий и других скрытых структур с применением методов статистики, машинного обучения, искусственного интеллекта и анализа данных.

# ОСНОВНЫЕ ЗАДАЧИ DATA MINING

| Задача | Что позволяет сделать | Пример |
|---|---|---|
| **Классификация** | Отнести объект к классу | Определить мошенническую транзакцию |
| **Кластеризация** | Найти группы похожих объектов | Сегментация клиентов |
| **Ассоциативный анализ** | Найти взаимосвязи | Какие товары покупают вместе |
| **Обнаружение аномалий** | Найти необычные объекты | Подозрительная транзакция |
| **Прогнозирование** | Предсказать значение | Прогноз продаж |
| **Снижение размерности** | Упростить данные | PCA для визуализации |
| **Text Mining** | Анализировать текст | Анализ отзывов |
| **Graph Mining** | Анализировать связи | Анализ социальных сетей |

# DATA MINING И MACHINE LEARNING

**Machine Learning**

- Основной акцент:
- построить модель, которая обучается на данных и делает прогнозы или принимает решения.

**Data Mining**

- Основной акцент:
- обнаружить в данных полезные, ранее неизвестные закономерности и знания.
- На практике эти направления тесно связаны.

Например:

**Задача:** интернет-магазин хочет понять поведение клиентов.

Data Mining может использовать:

- кластеризацию → сегментация клиентов;
- ассоциативные правила → товары, покупаемые вместе;
- обнаружение аномалий → необычные покупки;
- методы машинного обучения → прогноз оттока клиента.

# ЧТО ТАКОЕ KDD?

**KDD — Knowledge Discovery in Databases**, то есть:

- **обнаружение знаний в базах данных.**

KDD представляет собой **полный процесс поиска полезных знаний в данных**.

Важно:

- **KDD ≠ только Data Mining.**

Data Mining является одним из этапов более широкого процесса KDD.

Общая схема:

- **Данные → Предварительная обработка → Преобразование → Data Mining → Оценка → Знания**

# ЭТАПЫ KDD

Классический процесс KDD можно представить следующим образом:

**1. Selection — выбор данных**

Из общего хранилища выбираются данные, необходимые для конкретной задачи.

**Пример:**

Интернет-магазин имеет 10 млн записей, но для анализа поведения покупателей выбираются:

- ID клиента;
- дата покупки;
- товар;
- количество;
- стоимость.

# PREPROCESSING — ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА

На этом этапе устраняются проблемы качества данных:

- пропущенные значения;
- дубликаты;
- ошибки;
- шум;
- некорректные значения;
- выбросы.

**Пример:**

| Студент | Возраст | Балл |
|---|---|---|
| A | 20 | 85 |
| B | 21 | 90 |
| C | — | 78 |
| D | 20 | 105 |

# TRANSFORMATION — ПРЕОБРАЗОВАНИЕ ДАННЫХ

Данные приводятся к форме, удобной для анализа.

Например:

- нормализация;
- кодирование категориальных признаков;
- агрегация;
- создание новых признаков;
- изменение структуры данных.

**Пример:**

Было:

Дата рождения = 15.03.1998

Можно получить новый признак:

Возраст = 28 лет.

# DATA MINING — ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ

На этом этапе применяются алгоритмы поиска закономерностей.

Например:

- Apriori;
- FP-Growth;
- K-means;
- DBSCAN;
- Isolation Forest;
- PCA;
- алгоритмы классификации.

# INTERPRETATION / EVALUATION — ИНТЕРПРЕТАЦИЯ И ОЦЕНКА

Полученные результаты необходимо проверить.

Например, алгоритм обнаружил правило:

- Хлеб → Молоко

Но возникает вопрос:

**Действительно ли это полезная закономерность?**

Для ассоциативных правил используются, например:

- Support;
- Confidence;
- Lift.

Если **lift ≈ 1**, взаимосвязь может быть несущественной.

# KNOWLEDGE — ПОЛУЧЕНИЕ ЗНАНИЙ

После анализа формулируются понятные человеку выводы.

Например:

Покупатели, приобретающие кофе, значительно чаще приобретают печенье.

Это уже может стать основой для бизнес-решения.

# ЧТО ТАКОЕ CRISP-DM?

**CRISP-DM — Cross-Industry Standard Process for Data Mining.**

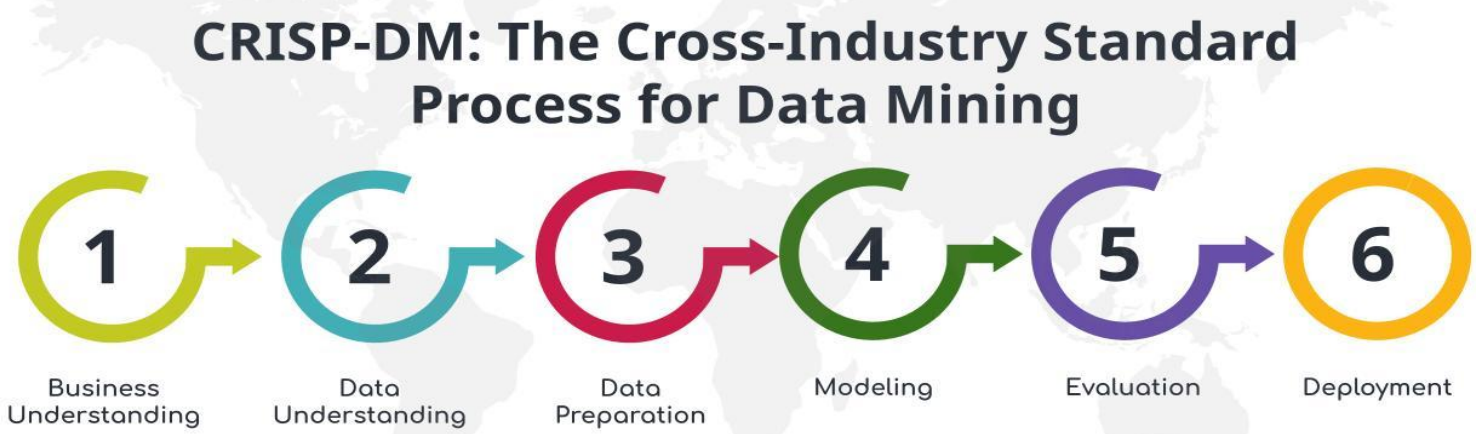
Это методология организации проектов Data Mining.

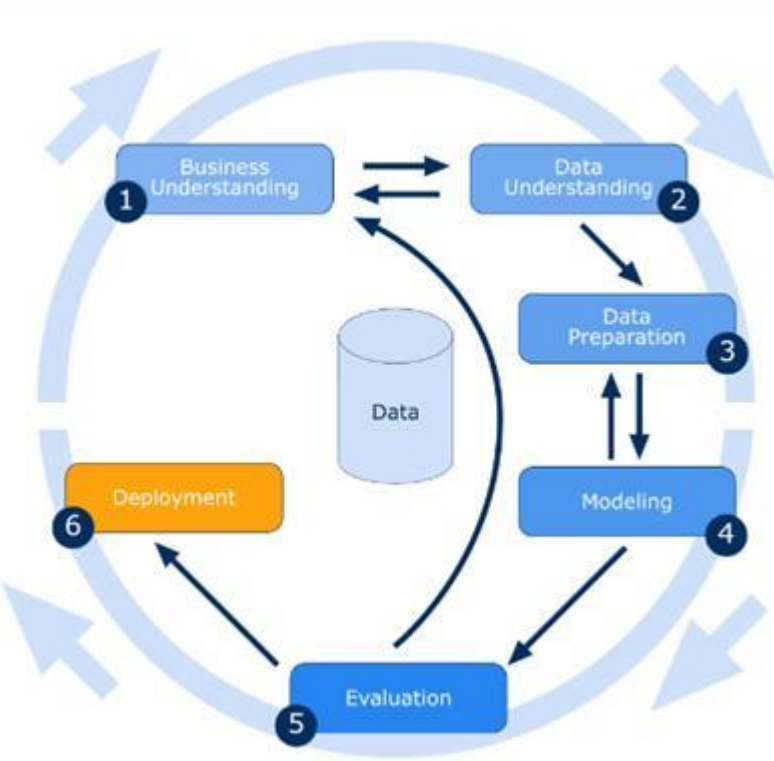
Она была разработана как универсальный подход, который можно применять в различных областях:

- бизнес;
- промышленность;
- финансы;
- образование;
- медицина;
- маркетинг;
- телекоммуникации.

CRISP-DM состоит из **шести основных этапов**.

# ЭТАПЫ CRISP-DM



# BUSINESS UNDERSTANDING

**Понимание бизнес-задачи**

Необходимо определить:

- какую проблему решаем;
- зачем нужен анализ;
- какой результат ожидается;
- как измеряется успех.

**Пример:**

- Банк хочет уменьшить количество мошеннических транзакций.

Цель:

- Снизить количество финансовых потерь от мошенничества.

# DATA UNDERSTANDING

**Понимание данных**

Исследуются:

- источники данных;
- структура;
- типы признаков;
- распределения;
- пропущенные значения;
- взаимосвязи.

| Признак | Тип |
|---|---|
| **Transaction_ID** | категориальный |
| **Amount** | числовой |
| **Time** | временной |
| **Location** | категориальный |
| **Device** | категориальный |
| **Fraud** | целевая переменная |

# DATA PREPARATION

**Подготовка данных**

Это один из наиболее трудоемких этапов.

Выполняются:

- очистка;
- обработка пропусков;
- удаление дубликатов;
- обработка выбросов;
- преобразование признаков;
- объединение источников;
- формирование итогового набора данных.

# MODELING

Выбираются и применяются алгоритмы.

Например:

- Decision Tree;
- Random Forest;
- K-means;
- DBSCAN;
- Apriori;
- FP-Growth;
- Isolation Forest.

**Пример:**

Для обнаружения подозрительных транзакций можно использовать:

- Isolation Forest.

# EVALUATION

**Оценка**

Полученную модель или закономерность необходимо проверить.

Например, для классификации используются:

- Accuracy;
- Precision;
- Recall;
- F1-score;
- ROC-AUC.

Но важно оценивать не только технические метрики.

Необходимо спросить:

Решает ли полученный результат исходную практическую задачу?

# DEPLOYMENT

**Внедрение**

Результаты анализа внедряются в реальную систему.

Например:

Модель обнаружения мошенничества интегрируется в банковскую систему.

При поступлении новой транзакции:

**Транзакция → Модель → Оценка риска → Уведомление / блокировка / дополнительная проверка**

# ПРИМЕР CRISP-DM: ПРОГНОЗ ОТТОКА СТУДЕНТОВ

Рассмотрим университет.

**Проблема**

- Университет хочет выявлять студентов с высоким риском отчисления.

**1. Business Understanding**

- Цель:
- Своевременно выявлять студентов группы риска.

**2. Data Understanding**

- Используются:
- посещаемость;
- оценки;
- количество академических задолженностей;
- активность в LMS;
- результаты промежуточного контроля.

**3. Data Preparation**

Выполняется:

- обработка пропусков;
- удаление дубликатов;
- кодирование категориальных признаков;
- формирование аналитического набора.

**4. Modeling**

Можно применить: **Decision Tree / Random Forest / Logistic Regression.**

**5. Evaluation**

Например: Recall = 0,85 Это означает, что модель смогла обнаружить 85% студентов из группы риска.

**6. Deployment**

Результаты интегрируются в информационную систему университета.

Система формирует:

**Студент → высокий риск → рекомендуется академическое сопровождение**

# KDD И CRISP-DM: В ЧЕМ РАЗЛИЧИЕ?

| KDD | CRISP-DM |
|---|---|
| **Knowledge Discovery in Databases** | Cross-Industry Standard Process for Data Mining |
| **Описывает процесс обнаружения знаний** | Описывает методологию управления Data Mining-проектом |
| **Основной акцент — получение знаний** | Основной акцент — решение практической задачи |
| **Включает этап Data Mining** | Включает этап Modeling |
| **Начинается с выбора и подготовки данных** | Начинается с понимания бизнес-задачи |
| **Завершается получением знаний** | Завершается внедрением результата |

# СОПОСТАВЛЕНИЕ ЭТАПОВ KDD И CRISP-DM

| KDD | CRISP-DM |
|---|---|
| **Selection** | Data Understanding |
| **Preprocessing** | Data Preparation |
| **Transformation** | Data Preparation |
| **Data Mining** | Modeling |
| **Interpretation/Evaluation** | Evaluation |
| **Knowledge** | Evaluation + Deployment |

# ВАЖНАЯ ОСОБЕННОСТЬ CRISP-DM — ИТЕРАТИВНОСТЬ

CRISP-DM **не является строго линейным процессом**.

Например:

- **Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation**

После оценки может оказаться, что модель работает недостаточно хорошо.

Тогда мы возвращаемся:

- **Evaluation → Data Preparation**

и изменяем подготовку данных.

Или:

- **Evaluation → Modeling**

и выбираем другой алгоритм.

Поэтому процесс выглядит скорее как цикл:

- **Задача → Данные → Подготовка → Модель → Оценка → Улучшение → ...**

# ГДЕ ПРИМЕНЯЕТСЯ DATA MINING?

**Финансы**

- обнаружение мошенничества;
- кредитный скоринг;
- анализ клиентов;
- выявление необычных операций.

**Маркетинг**

- сегментация клиентов;
- анализ покупательского поведения;
- рекомендации;
- прогноз оттока.

**Образование**

- анализ успеваемости;
- прогноз академических рисков;
- анализ поведения студентов;
- Learning Analytics.

**Медицина**

- анализ медицинских данных;
- выявление факторов риска;
- классификация пациентов;
- поиск закономерностей.

**Промышленность**

- прогнозирование отказов оборудования;
- анализ датчиков;
- обнаружение аномалий;
- Predictive Maintenance.

**Кибербезопасность**

- обнаружение атак;
- анализ сетевого трафика;
- поиск аномального поведения;

выявление подозрительной активности.

# ВОПРОСЫ

- Как сформулировать **Business Understanding**?
- Какие данные необходимо исследовать?
- Какие этапы preprocessing потребуются?
- Какой метод Data Mining можно применить?
- Как проверить качество результата?
- Как использовать полученные знания на практике?



# СПАСИБО ЗА ВНИМАНИЕ!