Российский государственный университет нефти и газа (НИУ) имени И.М. Губкина
ФАКУЛЬТЕТ КОМПЛЕКСНОЙ БЕЗОПАСНОСТИ ТЭК

Интеллектуальный анализ данных (Data Mining)

Введение

Савченко Наталья Александровна
ст.преподаватель

Определение Data Mining (короткое)

Data Mining это –

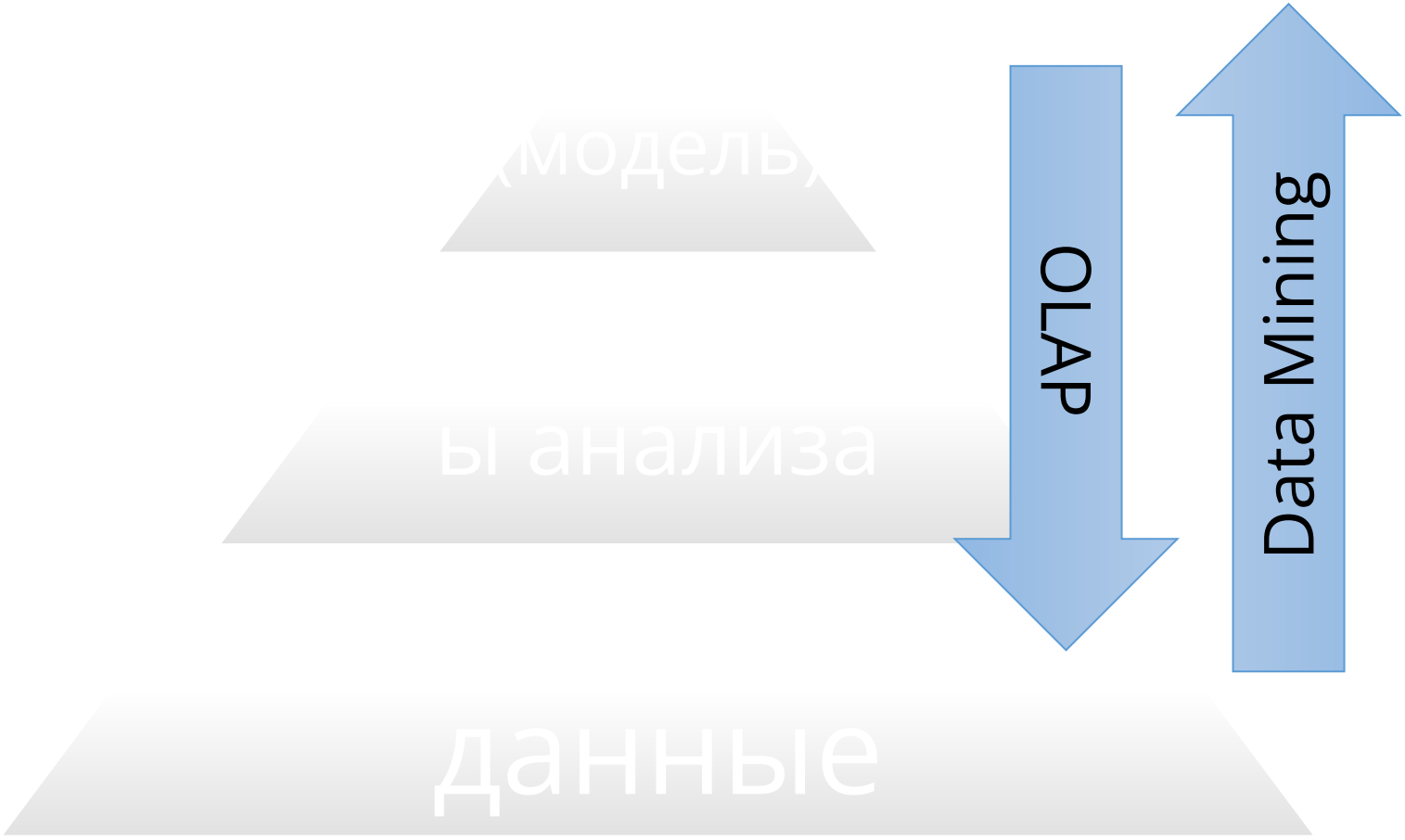
процесс «обнаружения знаний в базах данных».

Определение Data Mining (полное)

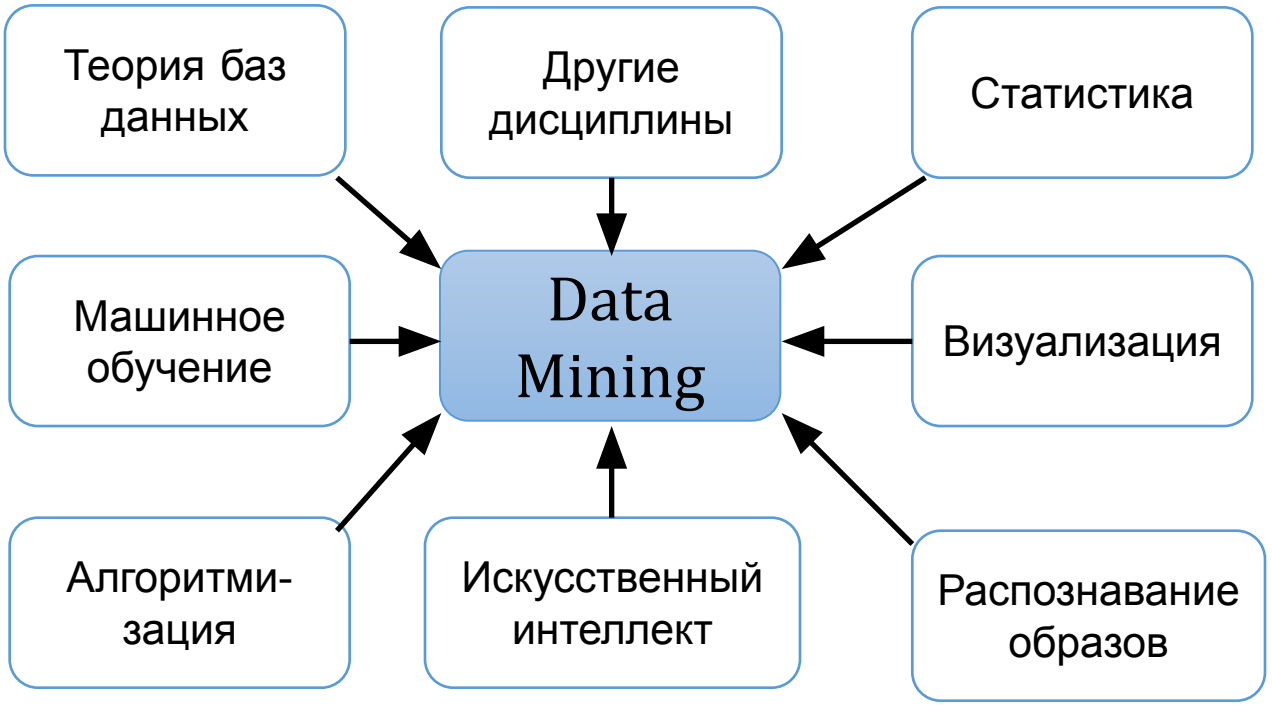
Data Mining это

процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Процесс анализа данных



Междисциплинарность интеллектуального анализа данных



Что позволяет сделать Data Mining:

1. Найти закономерности в накопленных данных;
2. Построить модели и правила, описывающих выявленные закономерности ;
3. Построить модели и правила, прогнозирующих дальнейшее развитие некоторых процессов.

Основные ограничения при использовании Data Mining

1. **Качество данных**

Около 75% работы над Data Mining состоит в сборе данных, который совершается еще до того, как запускаются сами инструменты интеллектуального анализа.

2. **Data Mining не может заменить аналитика**

Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов, которые обнаружены. Построенные модели должны быть грамотно интегрированы в бизнес-процессы для возможности их оценки и обновления.

Основные стадии Data Mining

Свободный поиск (выявление закономерностей)

На стадии свободного поиска осуществляется исследование набора данных с целью поиска скрытых закономерностей.

Закономерность (law) - существенная и постоянно повторяющаяся взаимосвязь, определяющая этапы и формы процесса становления, развития различных явлений или процессов.

Основные действия на этапе свободного поиска

выявление закономерностей условной логики (conditional logic);

выявление закономерностей ассоциативной логики (associations and affinities);

выявление трендов и колебаний (trends and variations);

а также валидация (тестирование, проверка) выявленных закономерностей.

«Прозрачность» выявленных закономерностей

Полученные закономерности, а точнее, их конструкции, могут быть:

прозрачными, т.е. допускающими толкование аналитика;

непрозрачными, так называемыми "черными ящиками".

Прогностическое моделирование (Predictive Modeling)

Выявленные закономерности используются для предсказания неизвестных значений.

Прогностическое моделирование включает такие **действия**:
предсказание неизвестных значений (outcome prediction);
прогнозирование развития процессов (forecasting).

Прогностическое моделирование (Predictive Modeling)

Выявленные закономерности используются для предсказания неизвестных значений.

Прогностическое моделирование включает такие **действия**:
предсказание неизвестных значений (outcome prediction);
прогнозирование развития процессов (forecasting).

Этапы подготовка к проведению Data mining

Понятие предметной области

Предметная область - это мысленно ограниченная область реальной действительности, подлежащая описанию или моделированию и исследованию.

Предметная область состоит из объектов, различаемых по свойствам и находящихся в определенных отношениях между собой или взаимодействующих каким-либо образом.

Анализ предметной области

В процессе изучения предметной области должна быть создана ее модель.

Модель предметной области описывает процессы, происходящие в предметной области, и данные, которые в этих процессах используются.

Постановка задачи Data Mining

Включает следующие шаги:

формулировка задачи;

формализация задачи.

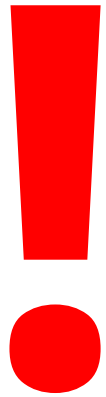
Постановка задачи включает также описание статического и динамического поведения исследуемых объектов.

Постановка задачи Data Mining

Описание статики:
описание объектов и их свойств.

Описании динамики:
описывается поведение объектов и те причины, которые влияют на их поведение.

Постановка задачи Data Mining



Технология Data Mining не может заменить аналитика и ответить на те вопросы, которые не были заданы.

Подготовка данных

Определение и анализ требований к данным

Определение и анализ требований к данным, которые необходимы для осуществления Data Mining.

Включая вопросы:

распределения пользователей;

вопросы доступа к данным, которые необходимы для анализа;

аналитические характеристики системы.

Сбор данных

Определение необходимого количества данных

Для определения оптимального объема данных необходимо ответить на следующие вопросы:

- Упорядочены ли данные?
- Включает ли набор данных сезонную/цикличную компоненту?
- Есть ли в наборе устаревшие данные или описывающие какую-то нетипичную ситуацию?
- Каково соотношение количества записей в наборе и кодичества входных переменных?
- Репрезентативен ли используемый набор данных?

Предварительная обработка данных

Предварительная обработка данных

Качество данных (Data quality) в данном случае является параметром, который характеризует прежде всего возможность их интерпретации.

Задачи Data-Mining:

- Классификация;
- Кластеризация;
- Поиск ассоциативных правил;
- Прогнозирование;
- Анализ отклонений.

Спасибо за внимание!

www.fdo.gubkin.ru