

Методологии и основные задачи анализа данных и машинного обучения

ИУ-5



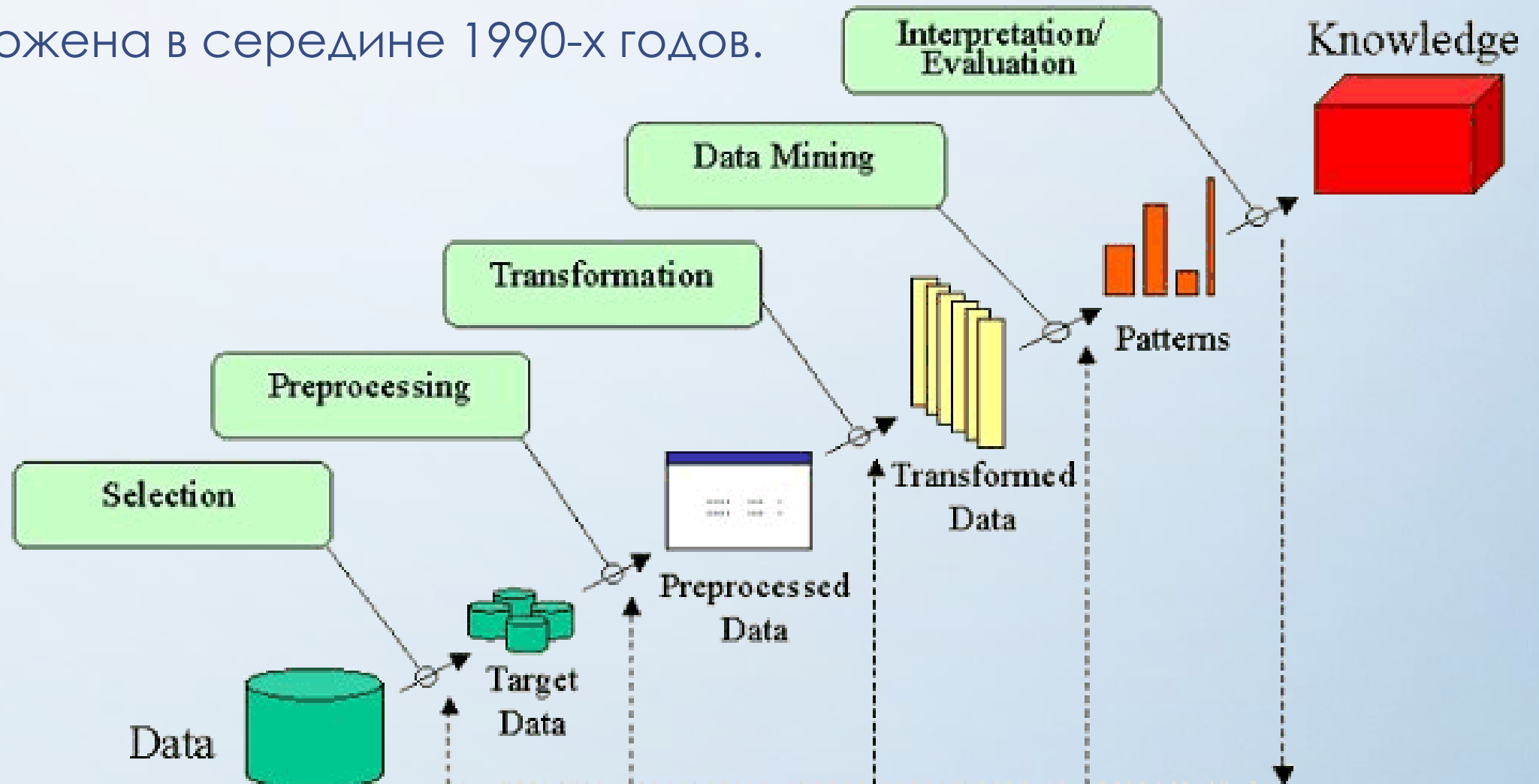
1. Методологии анализа данных

Методология

- В рамках **практического аспекта** (ориентированного на решение практических задач) **методология понимается как совокупность систематизированных определённым образом приёмов и способов организации деятельности**, применяемых в какой-либо области научного или практического знания. **Организовать деятельность означает упорядочить её в целостную систему с чётко определёнными характеристиками, логической структурой и процессом её осуществления (временной структурой)**. В границах обслуживания типовых программ деятельности практически ориентированная методология сводится к обеспечению их нормативно-рационального построения — алгоритмизируется.
 - Источник: <https://gtmarket.ru/concepts/6870>

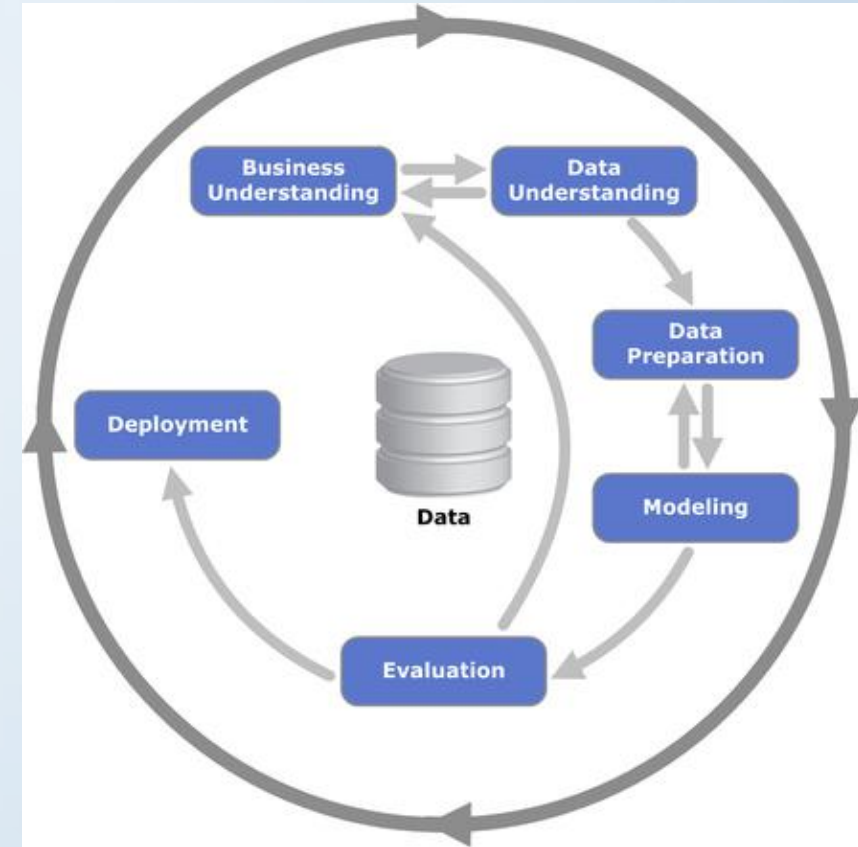
KDD Process

- Методология KDD (Knowledge Discovery in Databases) Process - http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
- Предложена в середине 1990-х годов.



CRISP-DM

- CRISP-DM (Cross-Industry Standard Process for Data Mining – межотраслевой стандартный процесс для исследования данных) – проверенная в промышленности и наиболее распространённая методология по исследованию данных.
- Первые версии предложены в конце 1990-х годов.
- Модель жизненного цикла исследования данных состоит из шести фаз, а стрелки обозначают наиболее важные и частые зависимости между фазами. Последовательность этих фаз строго не определена. Как правило в большинстве проектов приходится возвращаться к предыдущим этапам, а затем снова двигаться вперед. Описание фаз:
 1. Понимание бизнес-целей (Business Understanding)
 2. Начальное изучение данных (Data Understanding)
 3. Подготовка данных (Data Preparation)
 4. Моделирование (Modeling)
 5. Оценка качества модели (Evaluation)
 6. Внедрение (Deployment)

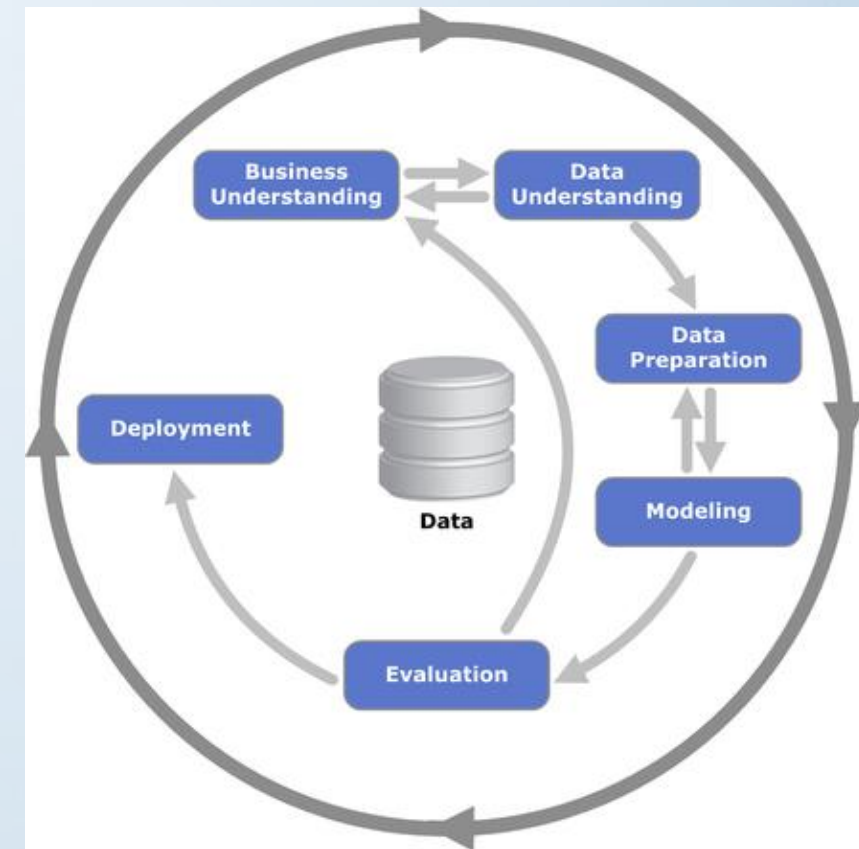


Анализ данных и АСОИУ

- На первый взгляд может показаться что анализ данных и «традиционные» информационные системы являются различными подходами. Так ли это?
- Проектирование АСОИУ (существуют различные модели проектирования: каскадная, спиральная):
 1. Определение целей автоматизации, постановка задач.
 2. Изучение предметной области.
 3. Построение модели (схемы) базы данных (с учетом целей автоматизации) – выделение сущностей, связей, атрибутов.
 4. Разработка информационной системы (автоматизация бизнес-процессов, создание форм, отчетов и т.д.)
 5. Оценка качества разработанной системы (тестирование, проверка работоспособности, моделирование нагрузки).
- Постановка и решение задачи анализа данных:
 1. Понимание бизнес-целей. Определение целей анализа данных.
 2. Начальное изучение данных (первичное изучение набора данных, первичная визуализация данных).
 3. Подготовка данных. Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи:
 - feature extraction – «технический» процесс выделения признаков, например из текстов или изображений.
 - feature engineering – «смысловое» выделение и синтез признаков, которые позволят получить наилучшее качество решения задачи.
 - Кодирование признаков (прежде всего категориальных).
 4. Моделирование. Разработка модели в терминах алгоритмов машинного обучения (применение одного или нескольких алгоритмов).
 5. Оценка. Оценка качества разработанной модели (с помощью методов оценки качества, используемых в машинном обучении).
- При проектировании АСОИУ акцент делается на «накопленные» пользователем бизнес процессы (в каком порядке и какие данные вводятся в формы ввода и сохраняются в БД, какие формируются отчеты и т.д.)
- При решении задачи анализа данных акцент делается на «накопленные» пользователем данные. Как помочь пользователю извлечь пользу из накопленных им данных. Какие нетривиальные зависимости можно найти. Какие решения можно помочь принять. Задачу анализа данных нужно рассматривать как элемент СППР.
- Решение задачи машинного обучения можно рассматривать как частный случай АСОИУ, где мы помогаем пользователю в решении задач, на основе накопленных им данных. Здесь работают как Data Scientist, так и Data Engineer.

CRISP-DM и машинное обучение (анализ датасетов)

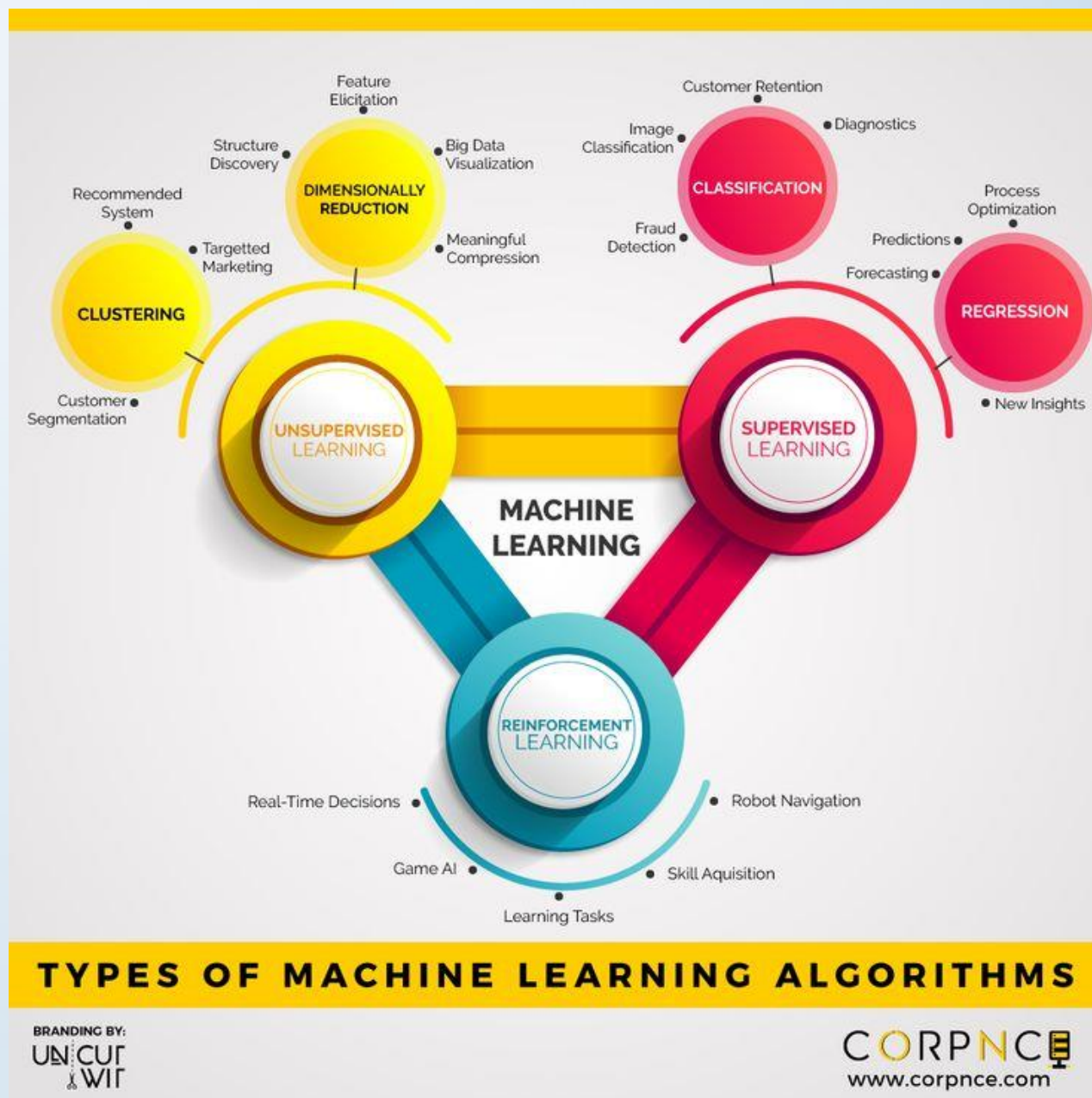
1. **Понимание бизнес-целей (Business Understanding)** – НЕТ. Как правило, на этапе решения задачи машинного обучения цель уже задана.
2. Начальное изучение данных (Data Understanding) – ДА. Первичное изучение набора данных, первичная визуализация данных.
3. Подготовка данных (Data Preparation) – ДА. Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи.
4. Моделирование (Modeling) – ДА. Разработка модели в терминах алгоритмов машинного обучения.
5. Оценка (Evaluation) – ДА. Оценка качества разработанной модели с помощью методов оценки качества, используемых в машинном обучении.
6. **Внедрение (Deployment)** – НЕТ.



2. Постановки задач машинного обучения

Типы («Классификация») задач ML

- Обучение с учителем (supervised learning)
 - Классификация
 - Регрессия
 - Прогнозирование временных рядов
- Обучение без учителя (unsupervised learning)
 - Кластеризация
 - Методы понижения размерности
- Обучение с подкреплением (reinforcement learning)
- [Карта методов scikit-learn](#)



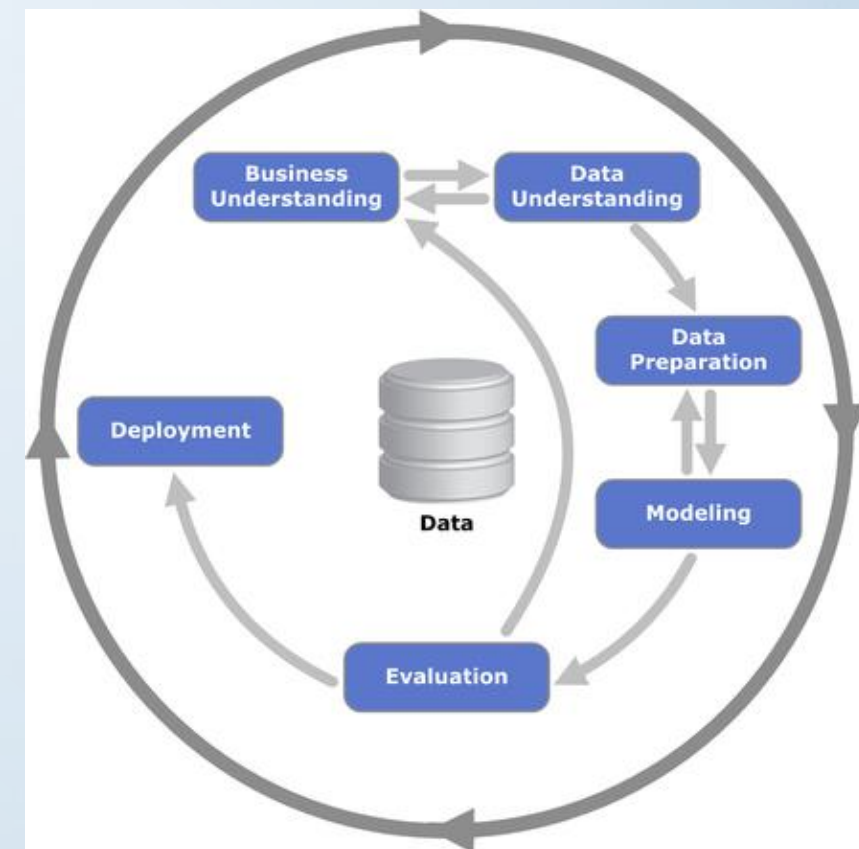
Некоторые типы шкал измерений

- Количественный (действительный) признак, который является действительным числом. Основной вид шкалы, к которому пытаются свести все остальные.
- Качественный (категориальный) признак.
 - Номинальная шкала (шкала наименований, классификационная шкала), по которой объектам дается некоторый признак (производится классификация объектов по этому признаку). Название «номинальный» объясняется тем, что такой признак дает лишь ничем не связанные имена объектам. Примерами измерений в номинальном типе шкал могут служить номера автомашин, телефонов, коды городов, объектов и т.д. **(Nominal variable). Способ кодирования - One-Hot Encoding.**
 - Частный случай – бинарная шкала $\{0, 1\}$, $\{\text{False}, \text{True}\}$. **(Dichotomous variable)**
 - Шкала называется ранговой (шкала порядка), если множество ее значений состоит из монотонно возрастающих чисел. При этом нет метрики, по которой можно сказать насколько одно значение больше или меньше другого. Примером шкалы порядка может служить шкала твердости минералов (предложенная в 1811 г. немецким ученым Ф. Моосом), шкала силы ветра, сортности товаров в торговле, различные социологические шкалы и т.д. **(Ordinal variable). Способ кодирования - Label Encoding.**



CRISP-DM и машинное обучение (анализ датасетов)

1. **Понимание бизнес-целей (Business Understanding)** – НЕТ. Как правило, на этапе решения задачи машинного обучения цель уже задана.
2. Начальное изучение данных (Data Understanding) – ДА. Первичное изучение набора данных, первичная визуализация данных.
3. **Подготовка данных (Data Preparation)** – ДА. Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи.
4. Моделирование (Modeling) – ДА. Разработка модели в терминах алгоритмов машинного обучения.
5. Оценка (Evaluation) – ДА. Оценка качества разработанной модели с помощью методов оценки качества, используемых в машинном обучении.
6. **Внедрение (Deployment)** – НЕТ. Оставим эту задачу дата-инженерам.



Данные

и

признаки

Табличные данные (объекты-атрибуты)

Атрибуты (свойства, поля данных)

Город	Год рождения	Доход	Пол
Москва	1990	100,00	Ж
Курск	1975	85,3	М
Москва	1983	40,5	Ж
Брянск	1960	90,5	М

Объекты

номинальная
шкала

шкала
порядка

действительный
признак

бинарная
шкала



Признаки

Город	Год рождения	Доход	Пол
1	1990	100,00	0
2	1975	85,3	1
1	1983	40,5	0
3	1960	90,5	1

Объекты

Текстовые данные (тексты-слова)



Изображения (изображения-пиксели)

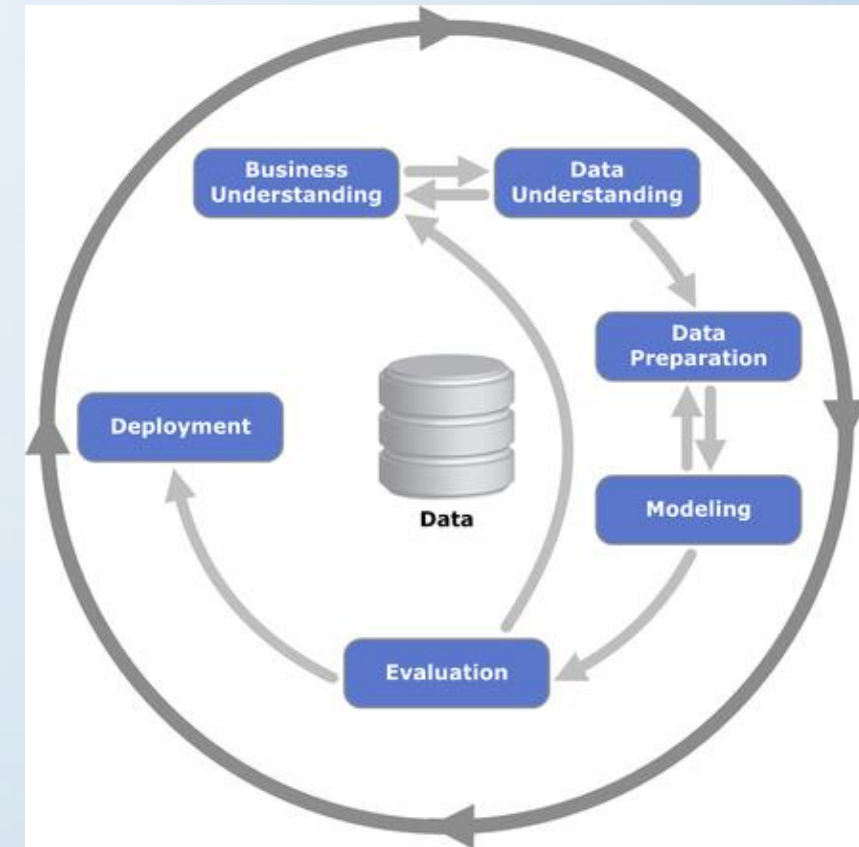


Формирование признаков
(feature engineering)

- Матрица объекты-признаки (feature data)
- Эту матрицу традиционно обозначают буквой X.

CRISP-DM и машинное обучение (анализ датасетов)

1. **Понимание бизнес-целей (Business Understanding)** – НЕТ. Как правило, на этапе решения задачи машинного обучения цель уже задана.
2. Начальное изучение данных (Data Understanding) – ДА. Первичное изучение набора данных, первичная визуализация данных.
3. Подготовка данных (Data Preparation) – ДА. Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи.
4. **Моделирование (Modeling)** – ДА. **Разработка модели в терминах алгоритмов машинного обучения.**
5. **Оценка (Evaluation)** – ДА. **Оценка качества разработанной модели с помощью методов оценки качества, используемых в машинном обучении.**
6. **Внедрение (Deployment)** – НЕТ. Оставим эту задачу дата-инженерам.



Обучение с учителем (на примере регрессии)

- Каждой строке матрицы X ставится в соответствие значение столбца ответов Y .
 Y -действительный признак.

Признаки (X)				Объекты	Ответы (Y)	
Город	Год рождения	Доход	Пол		Доход в будущем периоде	
1	1990	100,00	0		120,05	
2	1975	85,3	1		87,30	
1	1983	40,5	0		55,20	
3	1960	90,5	1	87,40		
обучающая выборка						
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ	
тестовая выборка						

- Ответы на тестовой выборке могут быть известны, но аналитику данных их не дают, заказчик может использовать их для итогового тестирования.
- Признаки на обучающей и тестовой выборке должны быть одинаково закодированы.**
- Обучение с учителем происходит в две фазы:
 - Собственно обучение. $M = \text{Alg.fit}(X_{\text{обуч}}, Y_{\text{обуч}}, H)$. Используемый нами алгоритм Alg строит модель соответствия M между $X_{\text{обуч}}$ и $Y_{\text{обуч}}$ с учетом гиперпараметров алгоритма H .
 - Предсказание. $Y_{\text{тест}} = \text{Alg.predict}(M, X_{\text{тест}})$.
- Гиперпараметры алгоритма – параметры, значение которых задается до начала обучения (значение остальных параметров настраивается в процессе обучения). У каждого алгоритма гиперпараметры свои, для их правильной настройки используются специальные методы, в частности перебор по сетке (grid search).
- Модель соответствия M можно рассматривать как функцию $f: Y=f(X)$. Но в более общем виде стоит рассматривать M как морфизм из теории категорий (введение в теорию категорий).

Оценка качества (на примере регрессии)

- Идея всех методов оценки качества состоит в том, чтобы понять насколько велика ошибка предсказания алгоритма, насколько хорошо или плохо он предсказывает. Разница только в используемых метриках.
- $M = \text{Alg.fit}(X_{\text{обуч}}, Y_{\text{обуч}}, H)$. $\hat{Y}_{\text{обуч}} = \text{Alg.predict}(M, X_{\text{обуч}})$. $\hat{Y}_{\text{обуч}}$ – результат работы алгоритма на обучающей выборке.
- При оценке качества стараются учесть возможное переобучение модели.
- Наиболее простая метрика – среднеквадратичная ошибка:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Классификация

- Бинарная классификация (Y-значение по бинарной шкале)

Признаки (X)				Объекты	Ответы (Y)	
Город	Год рождения	Доход	Пол		Переедет в другой город?	
1	1990	100,00	0		1 (Да)	
2	1975	85,3	1		0 (Нет)	
1	1983	40,5	0		0 (Нет)	
3	1960	90,5	1		1 (Да)	
обучающая выборка						
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ	
тестовая выборка						

- Многоклассовая классификация ([Multiclass classification](#)). Y-значение по номинальной или ранговой шкале, Label Encoding.

Признаки (X)				Объекты	Ответы (Y)	
Город	Год рождения	Доход	Пол		В какой город переедет?	
1	1990	100,00	0		2	
2	1975	85,3	1		2	
1	1983	40,5	0		1	
3	1960	90,5	1		1	
обучающая выборка						
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ	
тестовая выборка						

- Многометочная классификация ([Multi-label classification](#)) Y-множество значений по номинальной шкале, Label Encoding. Предсказывается несколько значений классов.

Признаки (X)				Объекты	Ответы (Y)	
Город	Год рождения	Доход	Пол		В какой город переедет?	
1	1990	100,00	0		2, 3	
2	1975	85,3	1		2	
1	1983	40,5	0		1	
3	1960	90,5	1		1, 3	
обучающая выборка						
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ	
тестовая выборка						

- Метрика качества – точность ([accuracy](#)) – доля правильно предсказанных меток классов.

Обучение без учителя (на примере кластеризации)

- Обучающей выборки нет.
- Для каждой строки матрицы X алгоритм пытается предсказать значение метки (номера) кластера Y .
- $Y = \text{Alg.fit_predict}(X, H)$. Используется алгоритм Alg с набором гиперпараметров H .
- Метрики оценки качества базируются на оценке расстояний между получившимися кластерами.
- Одним из наиболее сложных и интересных методов обучения без учителя являются самоорганизующиеся карты Кохонена.
- Другой важной задачей обучения без учителя является задача снижения (понижения) размерности данных.



Обучение с подкреплением

- Обучение с обратной связью, с опосредованным учителем.
- Алгоритм обучается, взаимодействуя с некоторой средой. Откликом среды являются сигналы подкрепления, поэтому такое обучение является частным случаем обучения с учителем, но учителем является среда или её модель.
- Частным случаем обучения с подкреплением является Q-обучение.

