Методологии и основные задачи анализа данных и машинного обучения

ИУ-5

1. История развития машинного обучения. Машинное обучение как часть направления искусственного интеллекта. Обзор основных обучающих ресурсов в области машинного обучения.

История развития машинного обучения

Искусственный интеллект, машинное обучение, наука о данных, Data Mining.

Как они соотносятся между собой?

Искусственный интеллект (википедия)

- Участники Российской ассоциации искусственного интеллекта дают следующие определения искусственного интеллекта:
  - Научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования тех видов человеческой деятельности, которые традиционно считаются интеллектуальными.
  - Свойство интеллектуальных систем выполнять функции (творческие), которые традиционно считаются прерогативой человека. При этом интеллектуальная система — это техническая или программная система, способная решать задачи, традиционно считающиеся творческими, принадлежащие конкретной предметной области, знания о которой хранятся в памяти такой системы. Структура интеллектуальной системы включает три основных блока — **базу знаний, решатель и интеллектуальный интерфейс**, позволяющий вести общение с ЭВМ без специальных программ для ввода данных.
  - Наука под названием «Искусственный интеллект» входит в комплекс компьютерных наук, а создаваемые на её основе технологии к информационным технологиям. Задачей этой науки является воссоздание с помощью вычислительных систем и иных искусственных устройств разумных рассуждений и действий.
- Как указывает председатель Петербургского отделения Российской ассоциации искусственного интеллекта Т. А. Гаврилова, в английском языке словосочетание *artificial intelligence* не имеет той слегка фантастической антропоморфной окраски, которую оно приобрело в довольно неудачном русском переводе. Слово *intelligence* означает «умение рассуждать разумно», а вовсе не «интеллект», для которого есть английский аналог *intellect*
- **Сильный и слабый искусственные интеллекты** — гипотеза в философии искусственного интеллекта, согласно которой некоторые формы искусственного интеллекта могут действительно обосновывать и решать проблемы.
  - теория **сильного** искусственного интеллекта предполагает, что компьютеры могут приобрести способность мыслить и осознавать себя, хотя и не обязательно их мыслительный процесс будет подобен человеческому.
  - теория **слабого** искусственного интеллекта отвергает такую возможность.
- Временная шкала развития искусственного интеллекта

Data Mining (википедия)

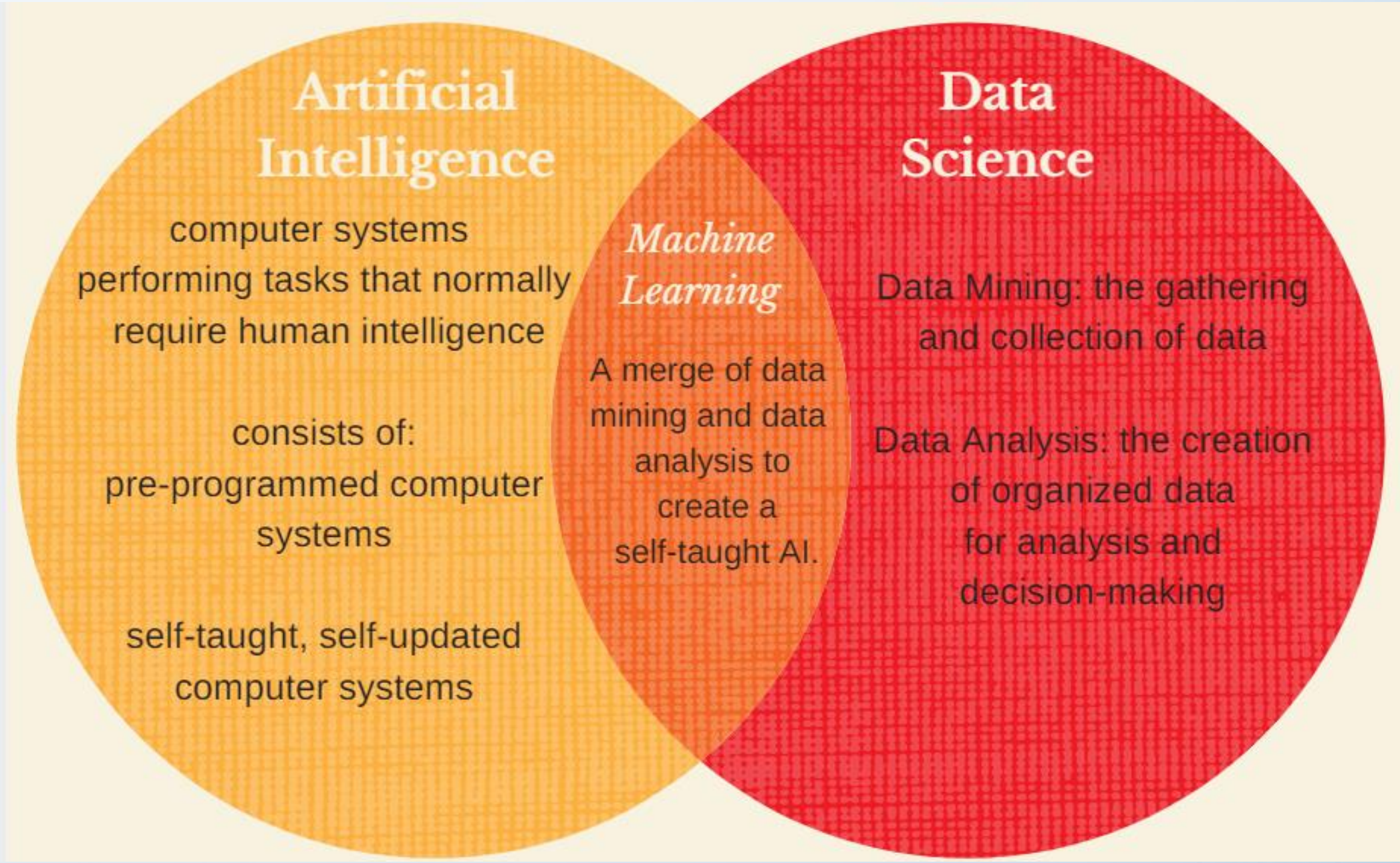
- Интеллектуальный анализ данных. Методы извлечения нетривиальных зависимостей из данных. Методы довольно разрозненные, аспект хранения данных не систематизирован (как правило используется термин «набор данных» - датасет).
- В целом данное направление близко к СППР (википедия) (ближе к слабому чем к сильному ИИ), хотя могут использоваться общие методы.
- Данное направление использует методы, наиболее близкие к машинному обучению.

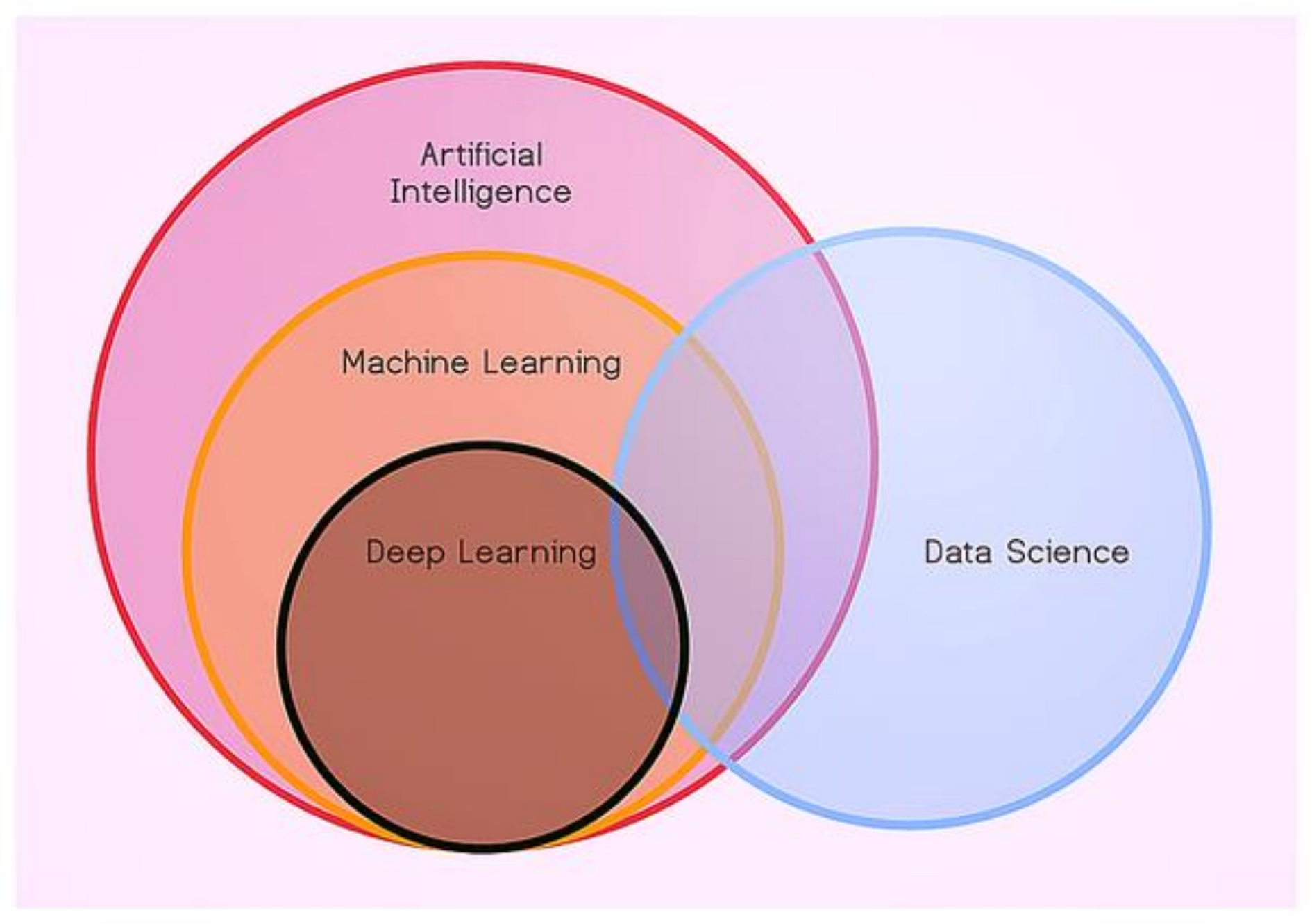
Машинное обучение (википедия)

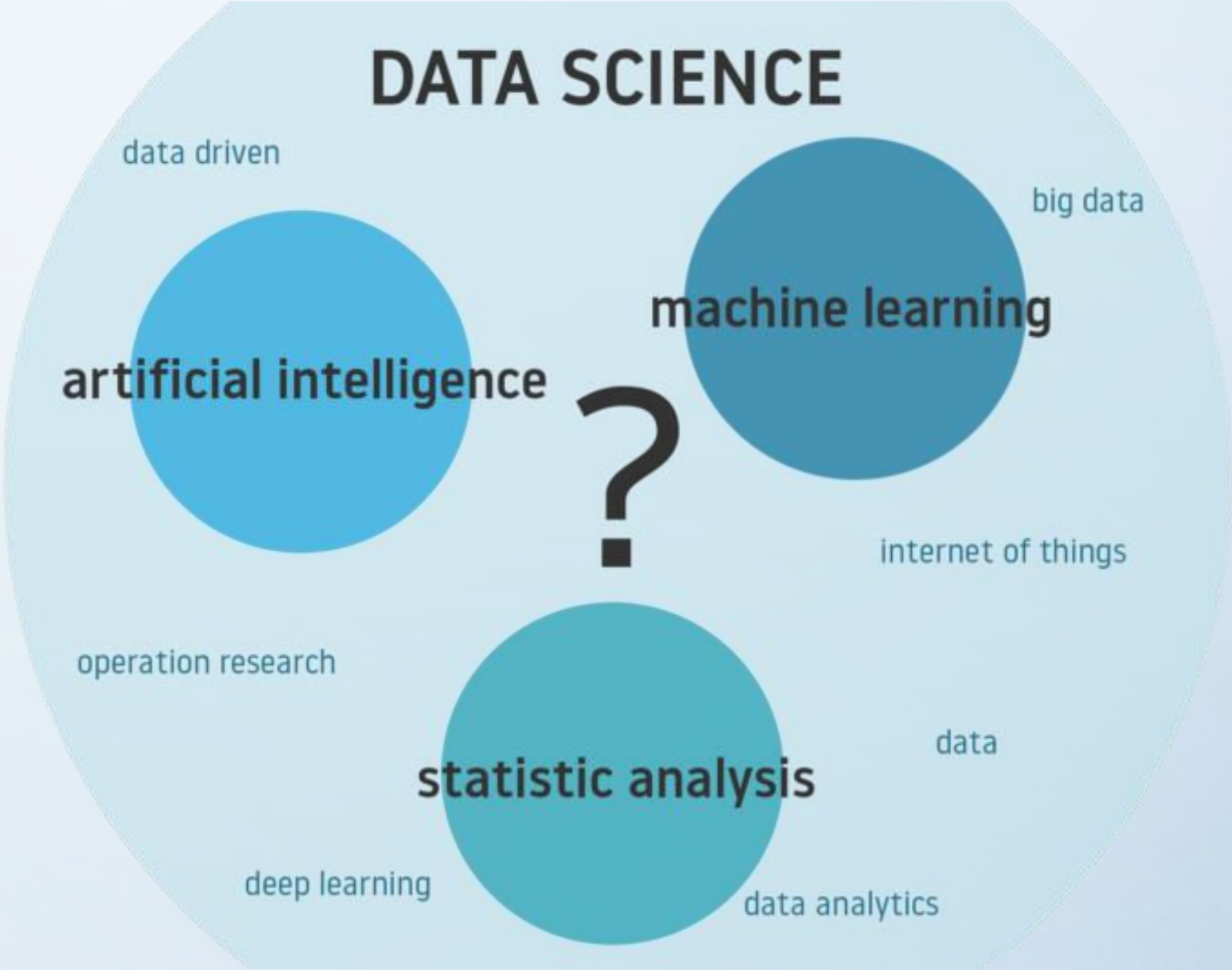
- Machine Learning, ML.
- Фактически является аналогом термина «обучение по прецедентам», который использовался в Data Mining.
- Основная задача – предсказание результата на основе предыдущих накопленных данных. Накопление данных называют обучением, поэтому используется термин «машинное обучение».
- Данные могут быть не упорядочены по времени (задачи классификации, регрессии) или упорядочены (прогнозирование временного ряда).
- Является набором наиболее низкоуровневых методов в ИИ. Применяемые алгоритмы очень сильно зависят от набора данных, на разных наборах данных разные алгоритмы могут показывать очень разное качество.
- Фактически основная задача – подобрать алгоритм, который покажет приемлемое качество предсказания на заданном наборе данных и не будет переобучаться.
- Появился лозунг «Data is the new science», то есть накопленные массивы данных определяют характер методов их обработки.

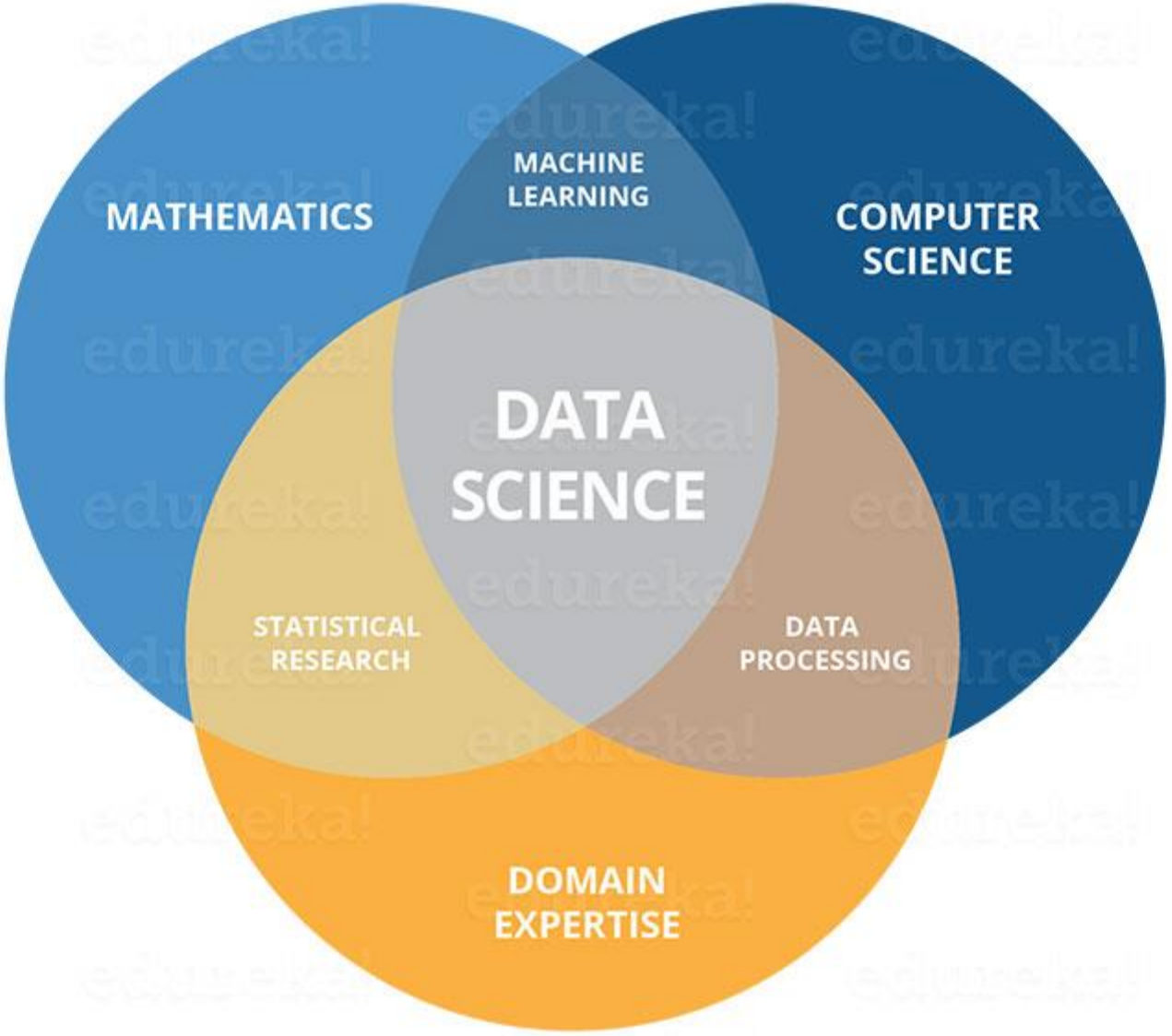
Наука о данных (data science)

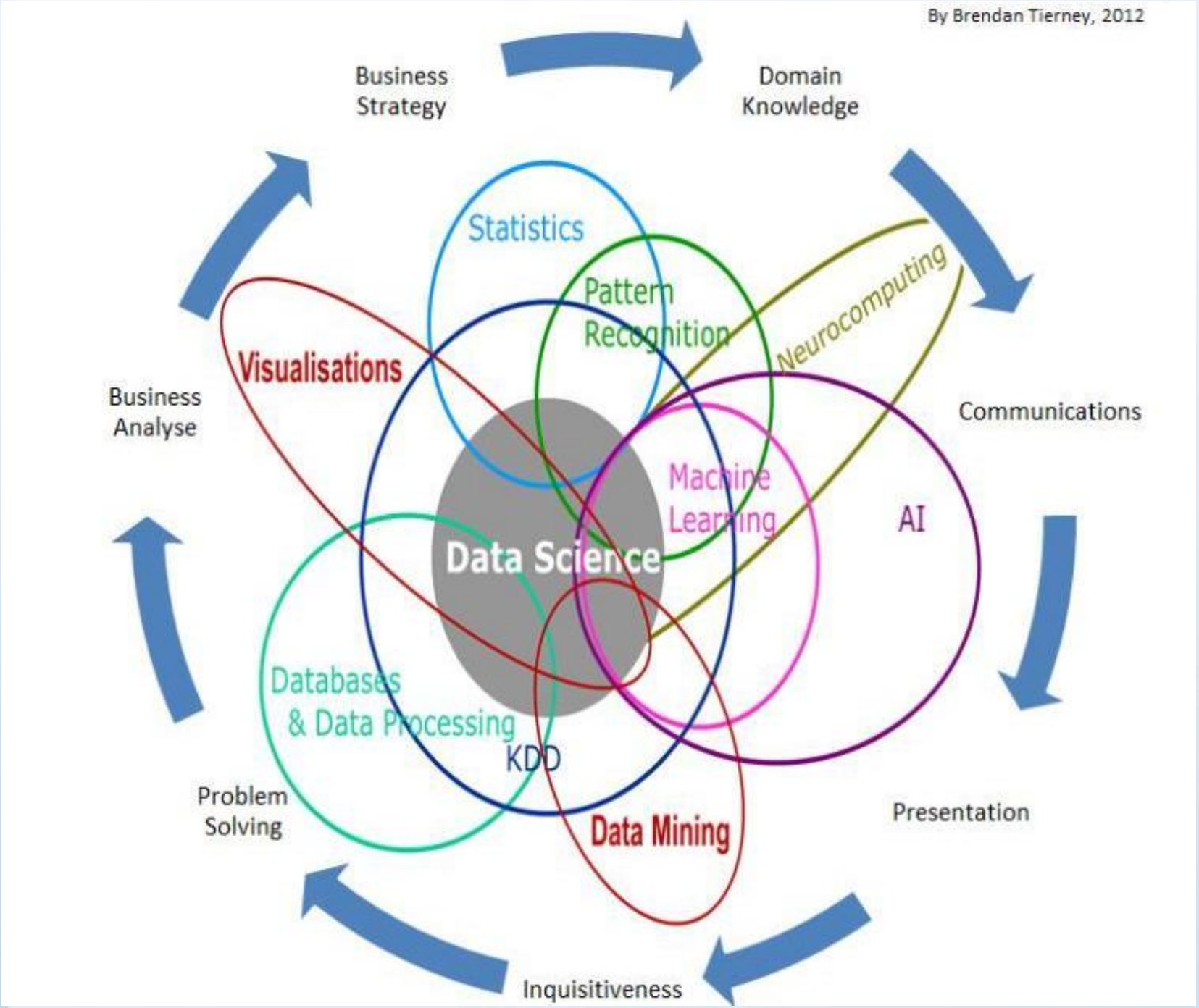
- В соответствии с (википедия):
  - Наука о данных (иногда даталогия – datalogy) – раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме.
  - Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных.


Artificial Intelligence
computer systems performing tasks that normally require human intelligence
consists of: pre-programmed computer systems
self-taught, self-updated computer systems
Machine Learning
A merge of data mining and data analysis to create a self-taught AI.
Data Science
Data Mining: the gathering and collection of data
Data Analysis: the creation of organized data for analysis and decision-making


Artificial
Intelligence
Machine Learning
Deep Learning
Data Science


DATA SCIENCE
data driven
big data
machine learning
artificial intelligence
?
internet of things
operation research
data
statistic analysis
deep learning
data analytics


MATHEMATICS
MACHINE LEARNING
COMPUTER SCIENCE
DATA SCIENCE
STATISTICAL RESEARCH
DATA PROCESSING
DOMAIN EXPERTISE



[Ссылка на источник]

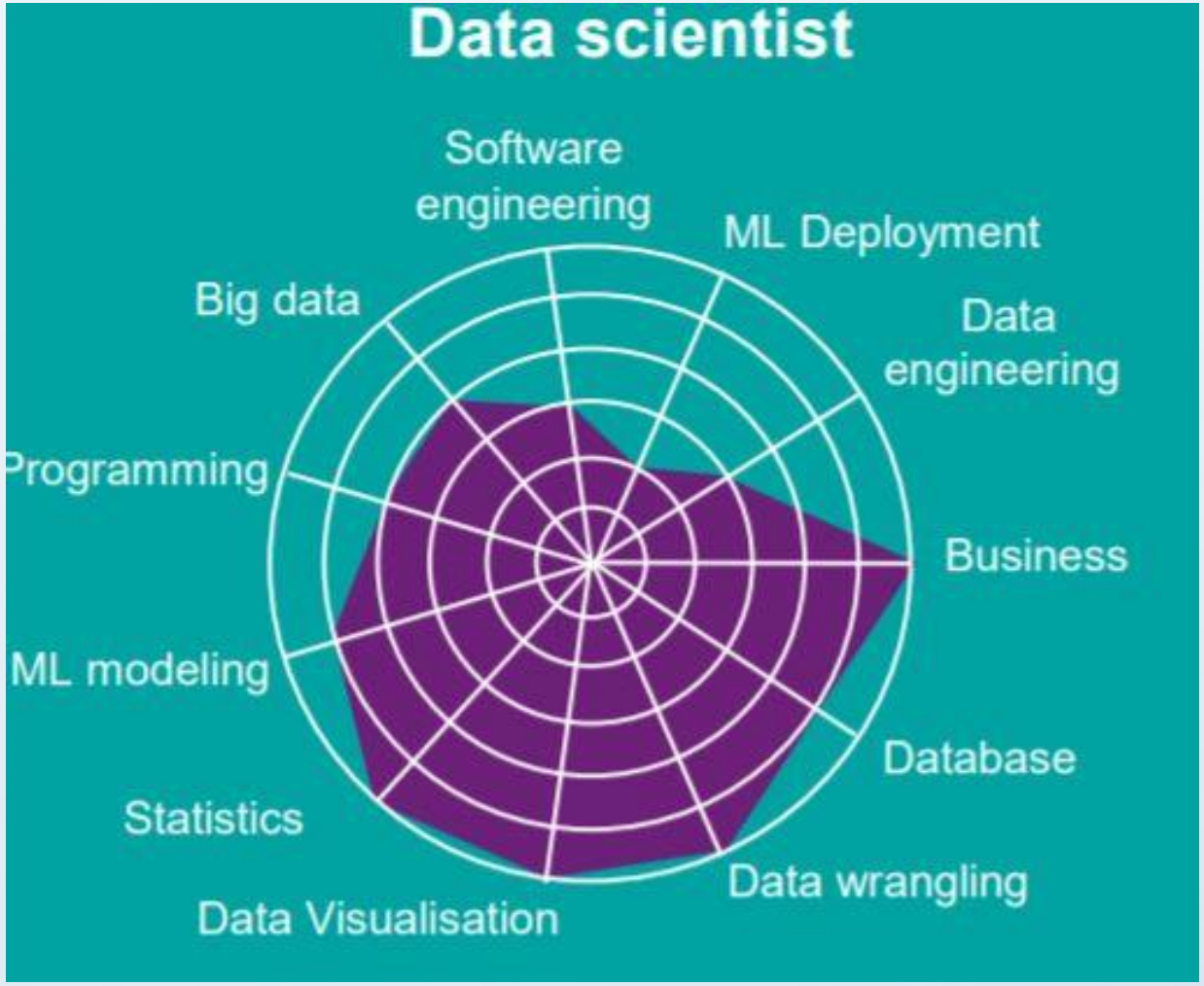
2. Профили специалистов в области анализа данных

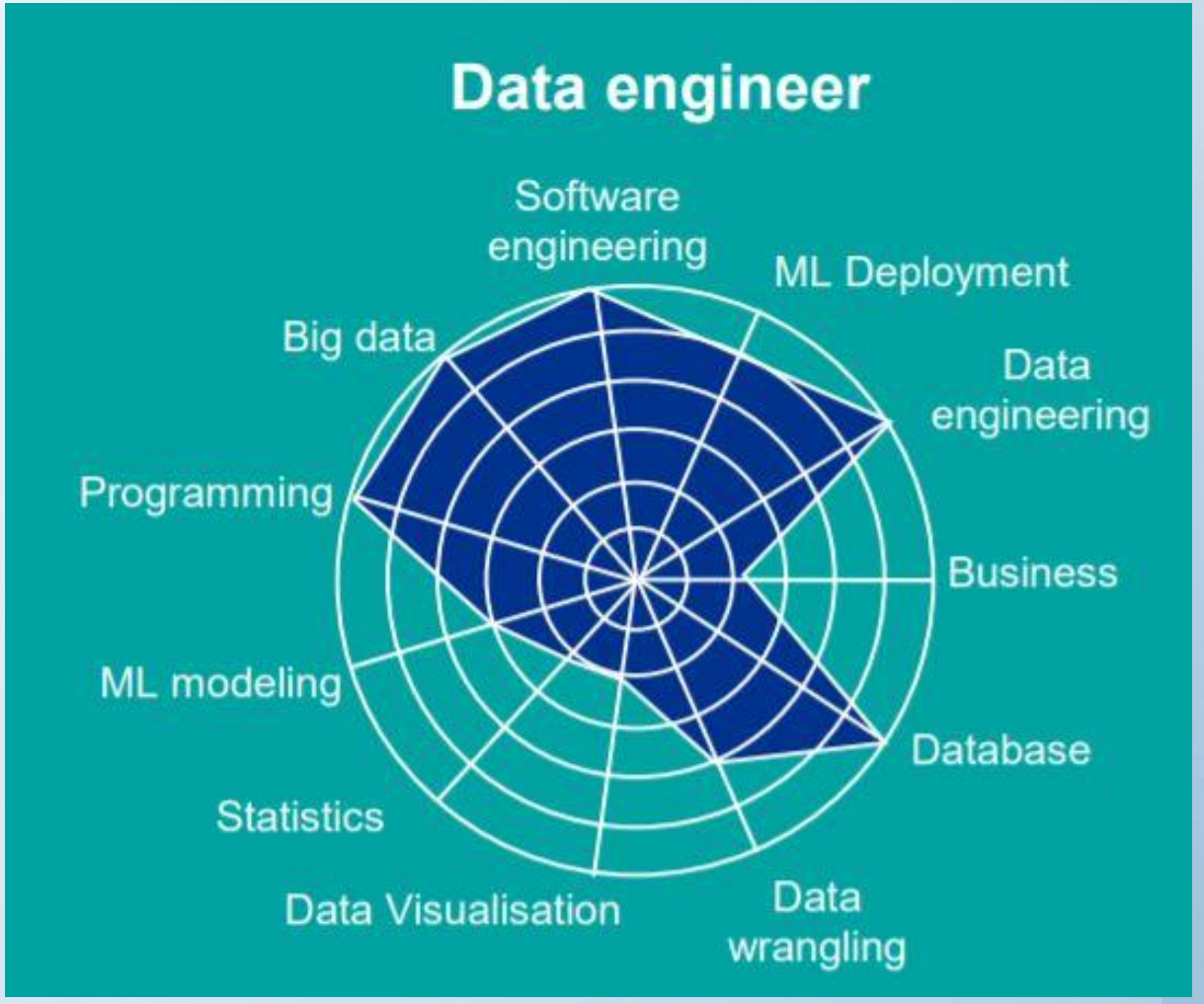
Data Scientist и Data Engineer

- Традиционно роль IT-специалиста, выполняющего анализ данных, называют «Data Scientist», то есть специалист по изучению данных и построению моделей, аналитик данных.
- Но данные для анализа нужно где-то хранить, передавать, обрабатывать и т.д. Роль IT-специалиста, которые это обеспечивает называют «Data Engineer».
- В основном роль IT-специалиста «Data Engineer» сейчас связана именно с обработкой больших данных.
- Data Engineer это специалист по базам данных, системный администратор (с хорошим знанием виртуализации и Big Data фреймворков), разработчик ETL-процессов (https://ru.wikipedia.org/wiki/ETL).

Data Scientist и Data Engineer - 2

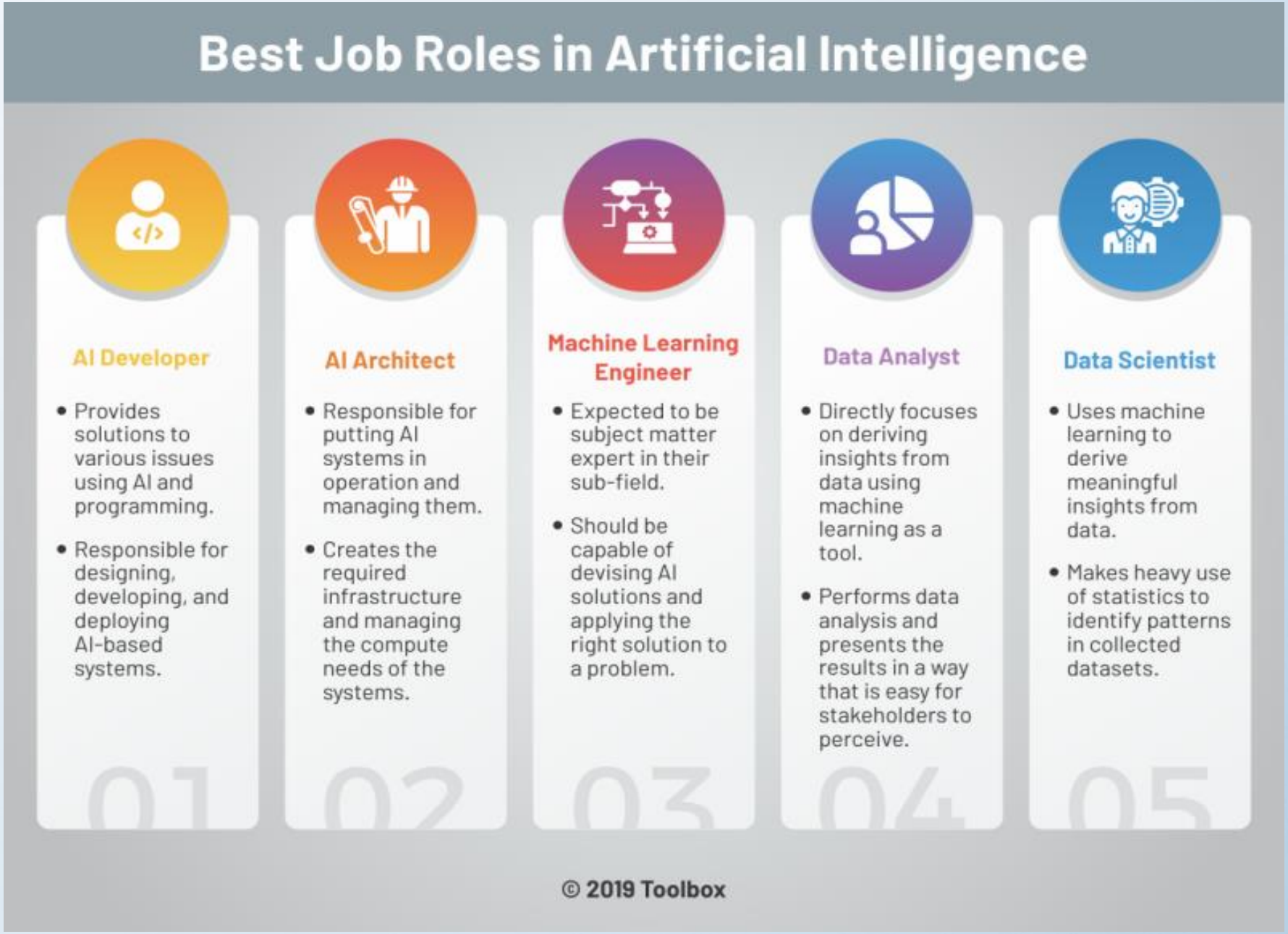
- Ссылка на статью





Профили специалистов в области анализа данных

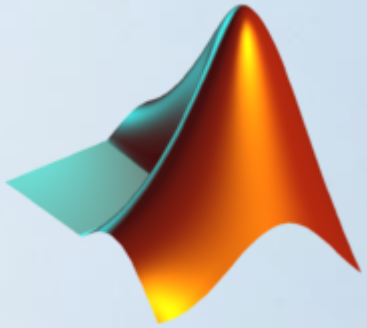
- Описание 1
- Описание 2
- Описание 3:



3. Языки анализа данных

- Общая особенность всех языков, применяемых для машинного обучения – использование векторизации вычислений. Компиляторы (интерпретаторы) реализуют высокую производительность для векторизованных вычислений.
- Алгоритмы работают не с отдельными ячейками данных, а с многомерными массивами, что увеличивает их производительность.
- В некоторых языках векторизация встроена непосредственно в язык, в некоторых реализована с помощью библиотек. В частности в Python векторизация реализована с помощью библиотеки NumPy.
- В Python также для повышения производительности используют элементы функционального программирования.

MATLAB



- Matrix Laboratory - пакет прикладных программ для решения задач технических вычислений и одноименный язык программирования, используемый в этом пакете. Пакет используют более миллиона инженерных и научных работников, он работает на большинстве современных операционных систем, включая Windows, Linux, Mac OS.
- De facto является пакетом №1 для анализа данных и машинного обучения.
- DSL-язык пакета MATLAB ориентирован на математиков. Не предназначен для разработки полнофункциональных программных систем.
- Пакет является проприетарным и платным.
- Существует свободно-распространяемый аналог - GNU Octave, язык которого в целом совместим с MATLAB, но который содержит меньше библиотек и отличается менее высокой производительностью.

R

- R — язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом.
- Изначально был ориентирован на задачи математической статистики, но в настоящее время содержит большое количество пакетов для анализа данных и машинного обучения.
- DSL-язык, ориентированный на математиков. Не предназначен для разработки полнофункциональных программных систем.

Julia

- Высокоуровневый высокопроизводительный свободный язык программирования с динамической типизацией, созданный для математических вычислений. Эффективен также и для написания программ общего назначения.
- Синтаксис языка схож с синтаксисом других математических языков (например, MATLAB и Octave), однако имеет некоторые существенные отличия. Julia написана на Си, C++ и Scheme.
- В стандартный комплект входит JIT-компилятор на основе LLVM, благодаря чему, по утверждению авторов языка, приложения, полностью написанные на языке, практически не уступают в производительности приложениям, написанным на статически компилируемых языках вроде Си или C++. Большая часть стандартной библиотеки языка написана на нём же.
- Также язык имеет встроенную поддержку большого числа команд для распределенных вычислений.
- Преимущества:
  - Производительность;
  - Ориентирован на параллельные вычисления.
- Недостатки:
  - Активно развивается, но пока находится в экспериментальной фазе.

Python

- Высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода. Синтаксис ядра Python минималистичен. В то же время стандартная библиотека включает большой объём полезных функций.
- Python поддерживает несколько парадигм программирования, в том числе структурное, объектно-ориентированное, функциональное, императивное и аспектно-ориентированное.
- Как и C++ поддерживает множественное наследование.
- Большинство библиотек является обертками над библиотеками, написанными на C/C++, что обеспечивает хорошую производительность работы библиотек.
- **Как правило, вызов библиотечной функции (написанной на C/C++) намного производительнее аналогичного кода написанного прикладным программистом на Python (особенно если это ML-алгоритм написанный без использования векторизации).**
- Основное преимущество Python состоит в том, что может использоваться и как язык для обработки данных и как язык для разработки приложений (веб-приложений). Это очень облегчает встраивание ML-решений в веб-приложения.

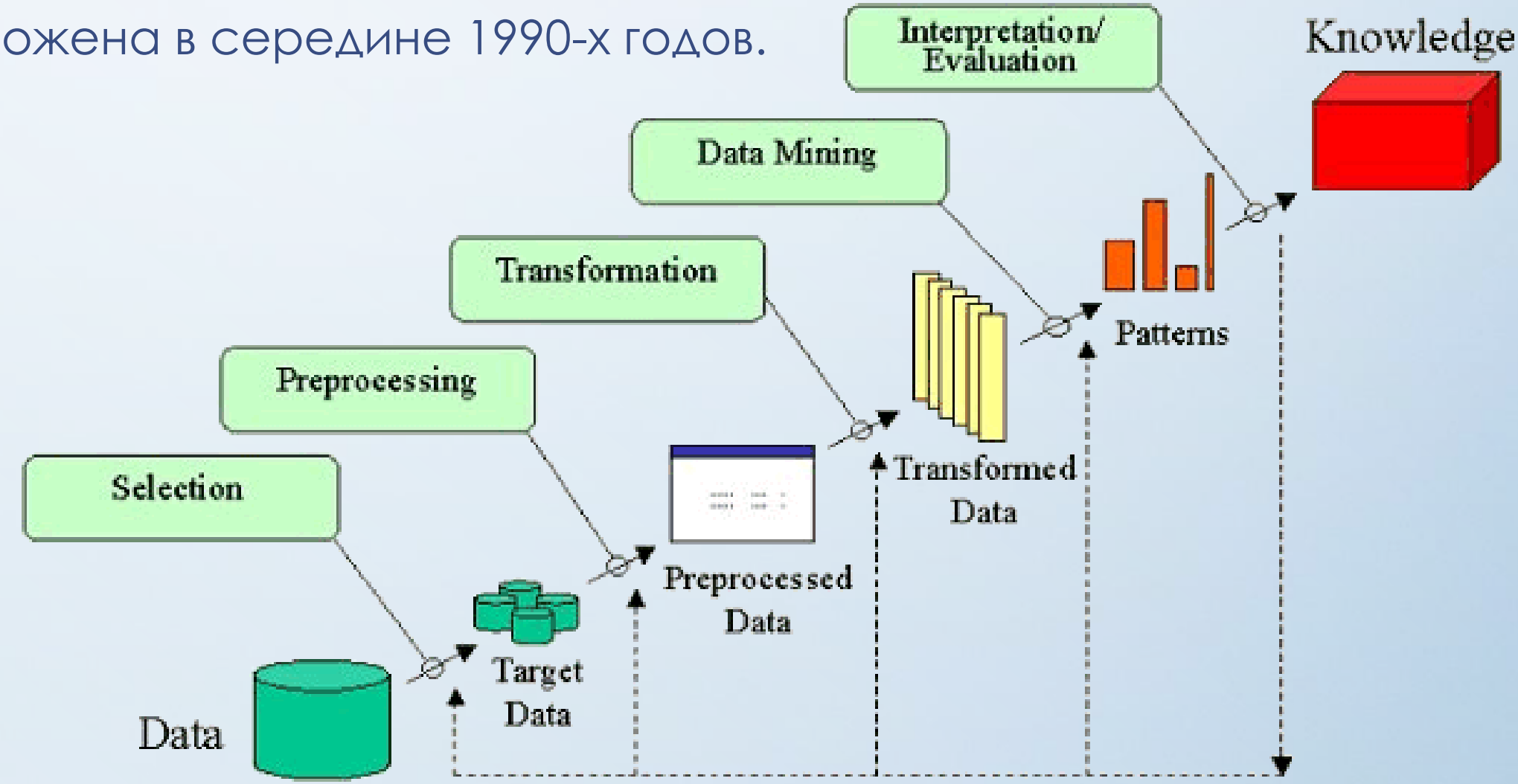
4. Методологии анализа данных

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 (временно́й структурой).** В границах обслуживания типовых программ деятельности практически ориентированная методология сводится к обеспечению их нормативно-рационального построения — алгоритмизируется.
  - Источник: https://gtmarket.ru/concepts/6870

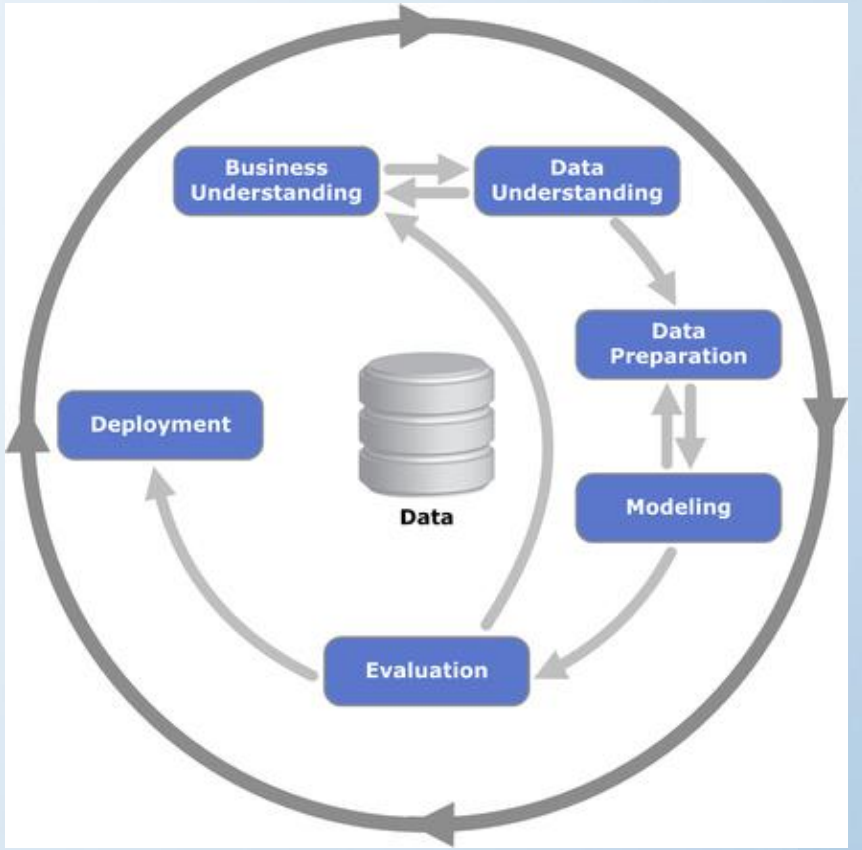
KDD Process

- 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.

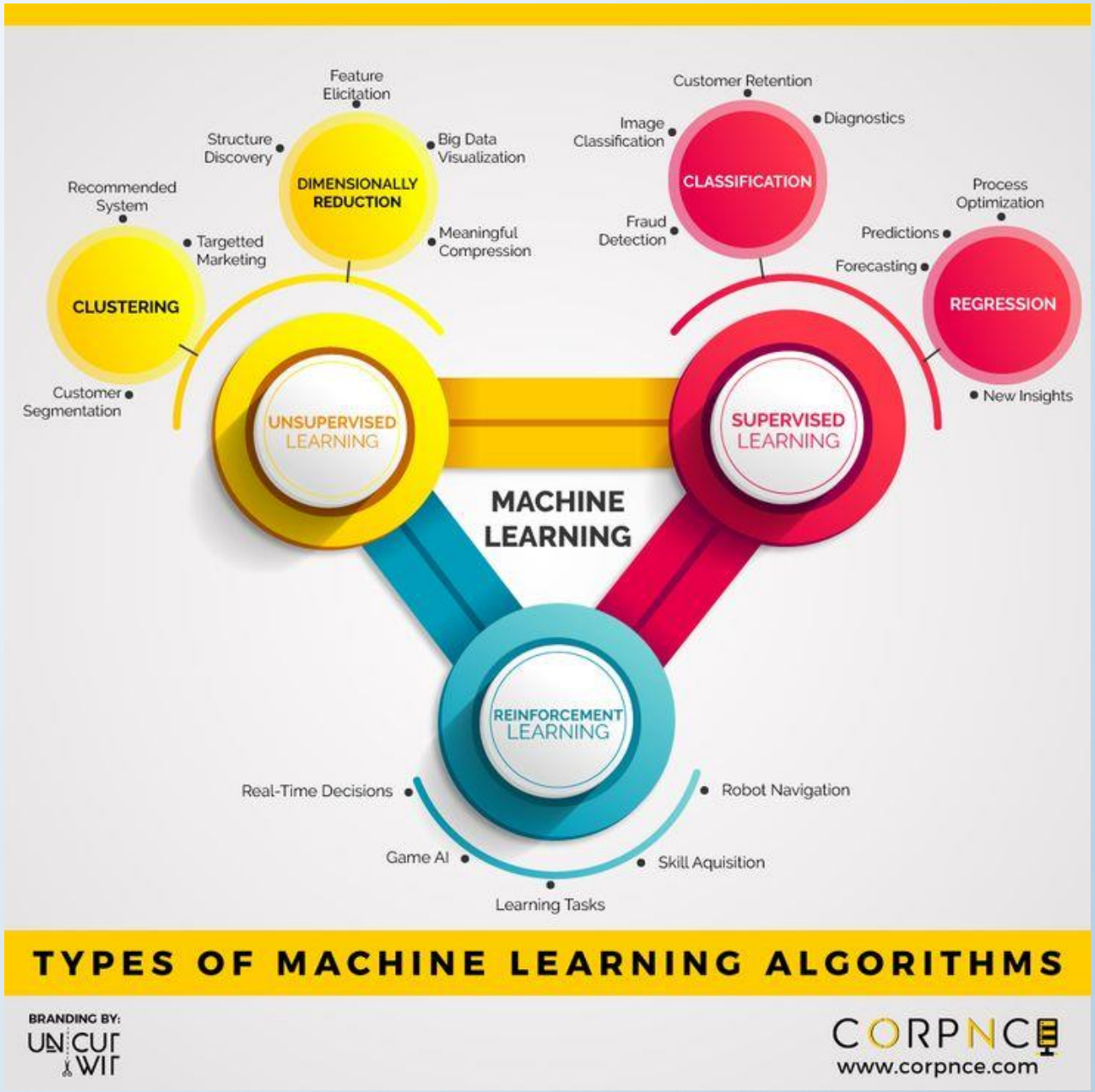
CRISP-DM и машинное обучение (анализ датасетов)

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.

5. 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}, {False, 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)

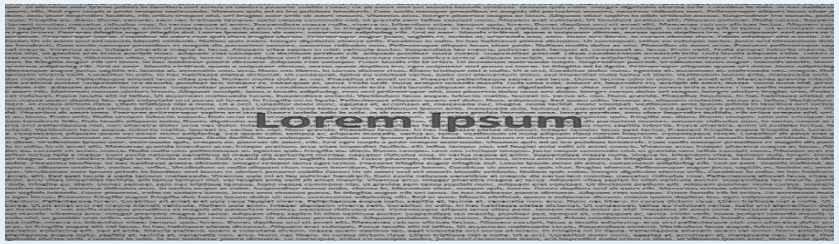
Город	Год рождения	Доход	Пол	
Москва	1990	100,00	Ж	*Объекты*
Курск	1975	85,3	М	
Москва	1983	40,5	Ж	
Брянск	1960	90,5	М	
номинальная шкала	шкала порядка	действительный признак	бинарная шкала	

Признаки

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

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

Текстовые данные (тексты-слова)



Изображения (изображения-пиксели)

Формирование признаков (feature engineering)

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)					*Ответы (Y)*
Город	**Год рождения**	**Доход**	**Пол**		**Доход в будущем периоде**
1	1990	100,00	0	*Объекты*	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:

1. Собственно обучение. $M=Alg.fit(X_{обуч}, Y_{обуч}, H)$. Используемый нами алгоритм Alg строит модель соответствия M между $X_{обуч}$ и $Y_{обуч}$ с учетом гиперпараметров алгоритма H.
2. Предсказание. $Y_{тест} = Alg.predict(M, X_{тест})$.

- 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=Alg.fit($X_{обуч}$, $Y_{обуч}$, H). $\hat{Y}_{обуч}$ = Alg.predict(M, $X_{обуч}$). $\hat{Y}_{обуч}$ – результат работы алгоритма на обучающей выборке.
- При оценке качества стараются учесть возможное переобучение модели.
- Наиболее простая метрика – среднеквадратичная ошибка:

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

Классификация

- Бинарная классификация (Y-значение по бинарной шкале)

Признаки (X)					*Ответы (Y)*
Город	Год рождения	Доход	Пол		Переедет в другой город?
1	1990	100,00	0	*Объекты*	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)					*Ответы (Y)*
Город	Год рождения	Доход	Пол		В какой город переедет?
1	1990	100,00	0	*Объекты*	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)					*Ответы (Y)*
Город	Год рождения	Доход	Пол		В какой город переедет?
1	1990	100,00	0	*Объекты*	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=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.