

Интерактивная лекция и
научно-практический
семинар

7 ноября 2019 года



**Практическое введение
в машинное обучение**

Гапанюк Юрий Евгеньевич,
доцент кафедры ИУ-5

План

1. Интерактивная лекция

- Часть 1. История развития машинного обучения. Машинное обучение как часть направления искусственного интеллекта. Обзор основных обучающих ресурсов в области машинного обучения.
- Часть 2. Основные задачи машинного обучения и способы их решения с использованием современных программных средств на основе языка программирования Python.

2. Научно-практический семинар

- Часть 1. Реализация метода ближайших соседей (на основе языка программирования Python) как одного из наиболее простых методов машинного обучения.
- Часть 2. Решение задачи анализа тональности текста методами машинного обучения на основе языка программирования Python.

Часть 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

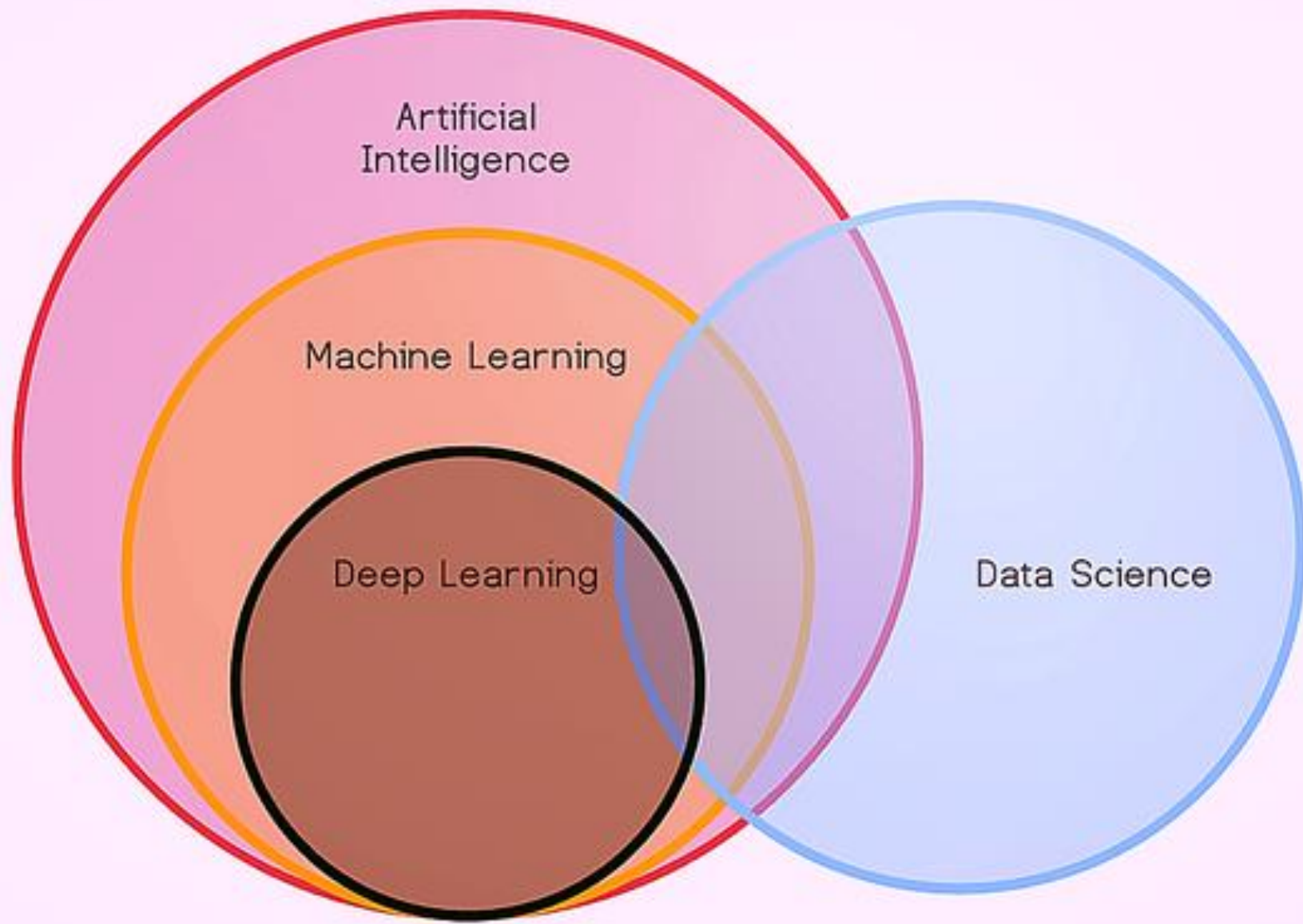
Machine Learning

A merge of data
mining and data
analysis to
create a
self-taught AI.

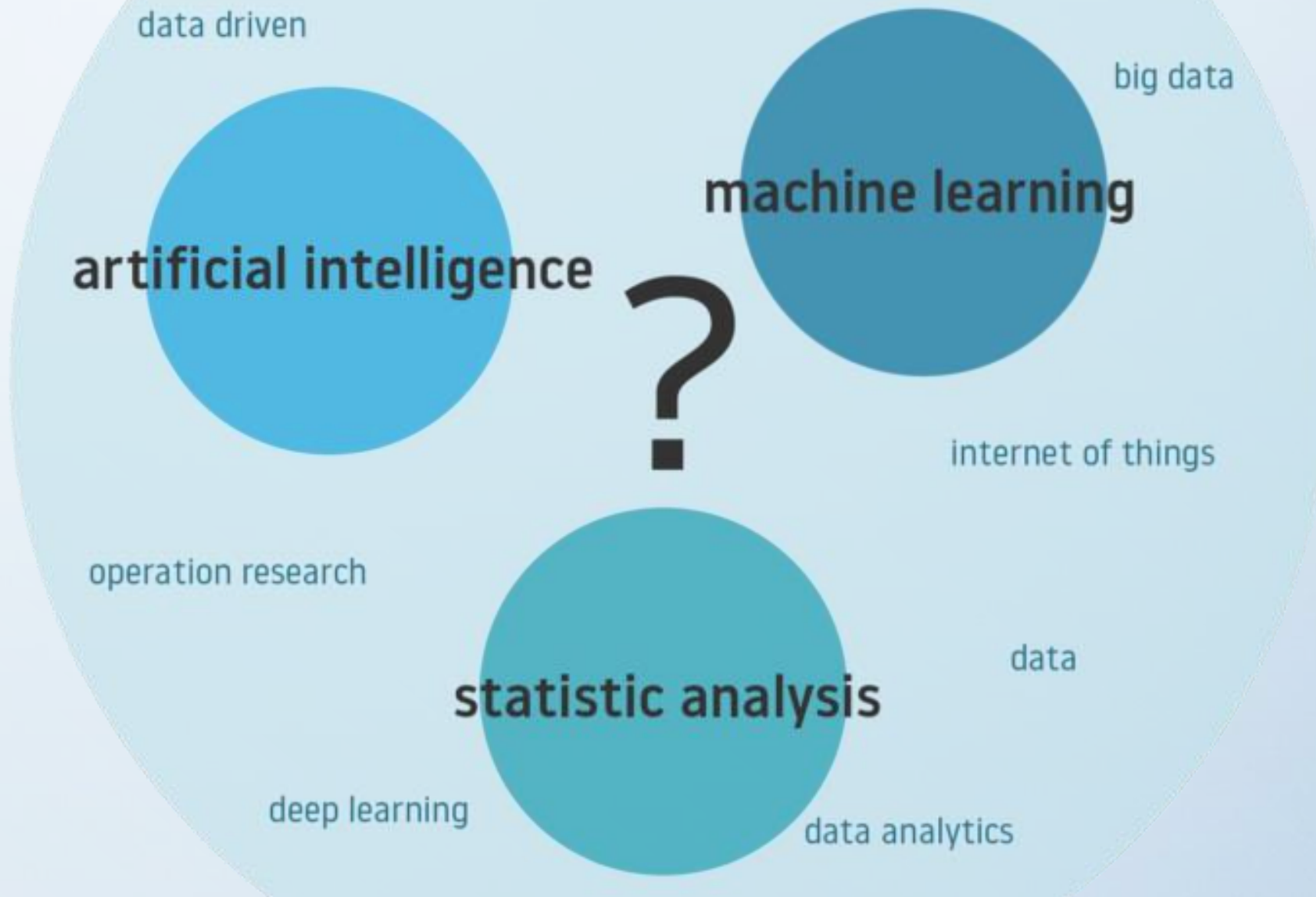
Data Science

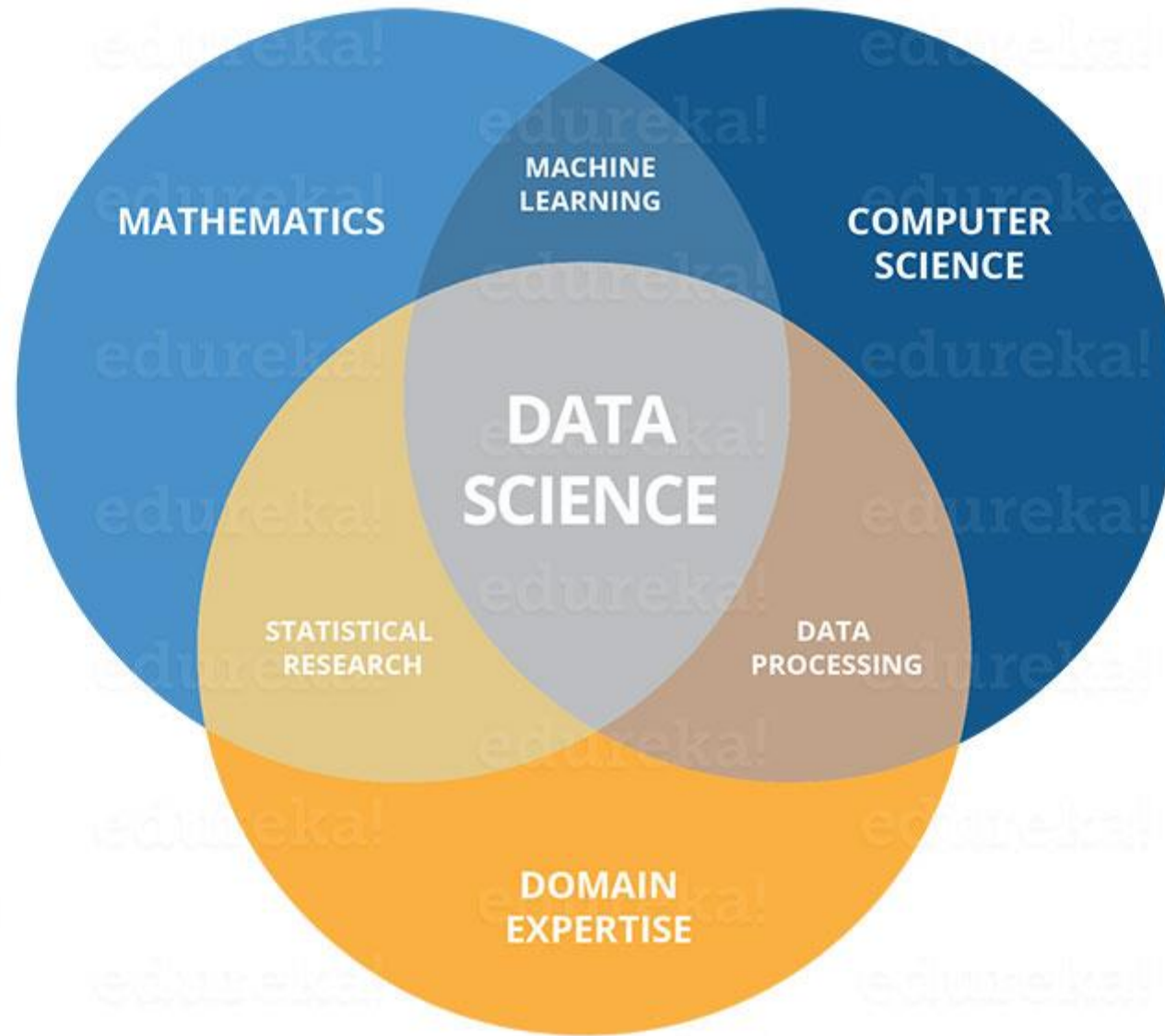
Data Mining: the gathering
and collection of data

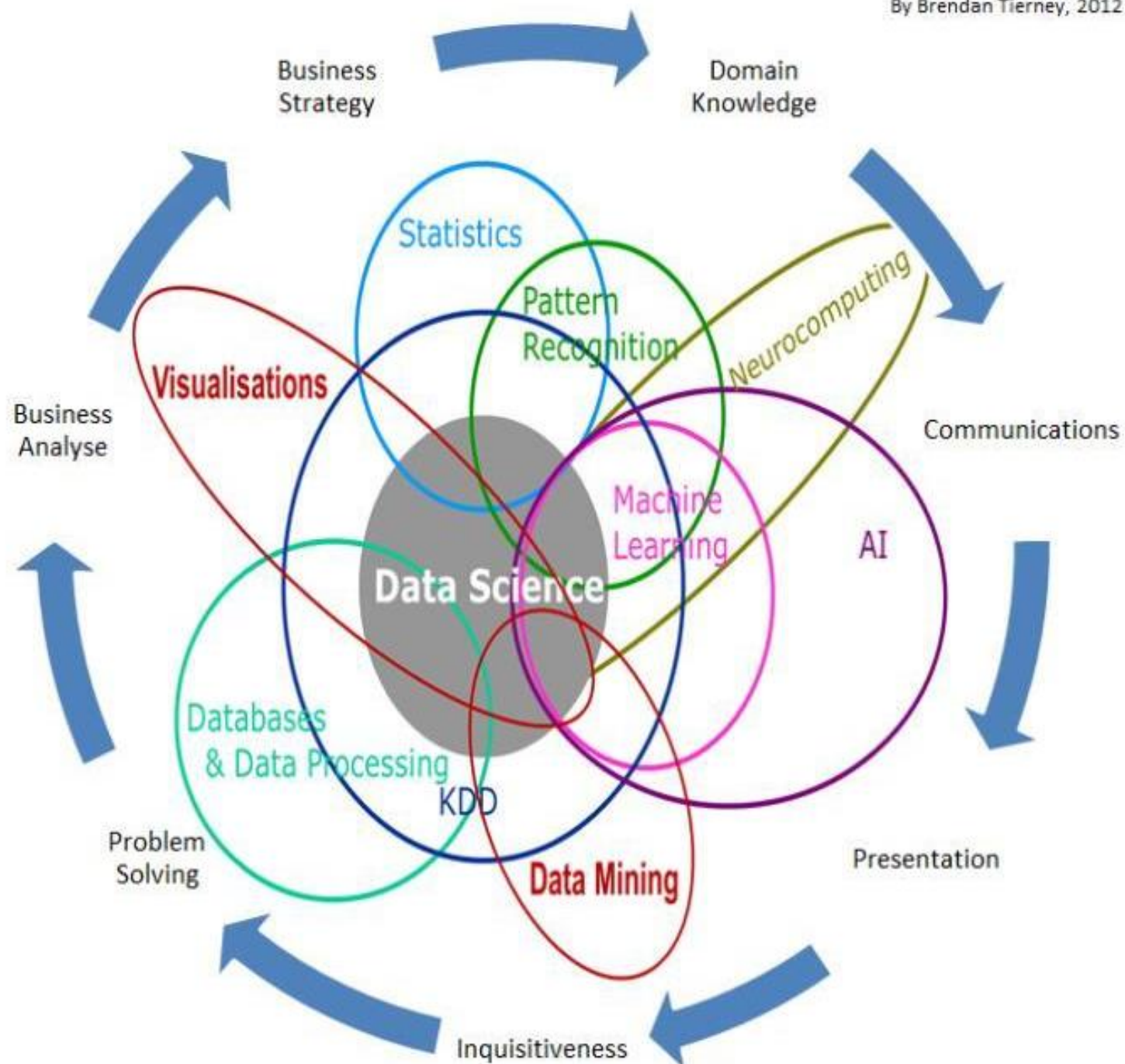
Data Analysis: the creation
of organized data
for analysis and
decision-making



DATA SCIENCE





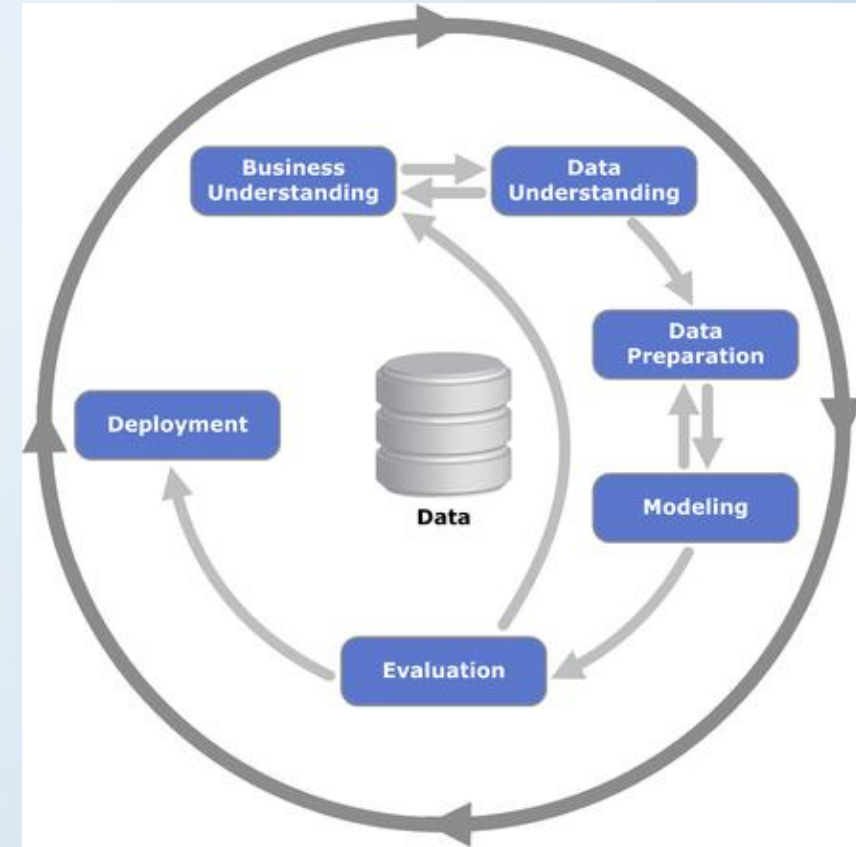


[Ссылка на источник](#)

Методологии анализа данных

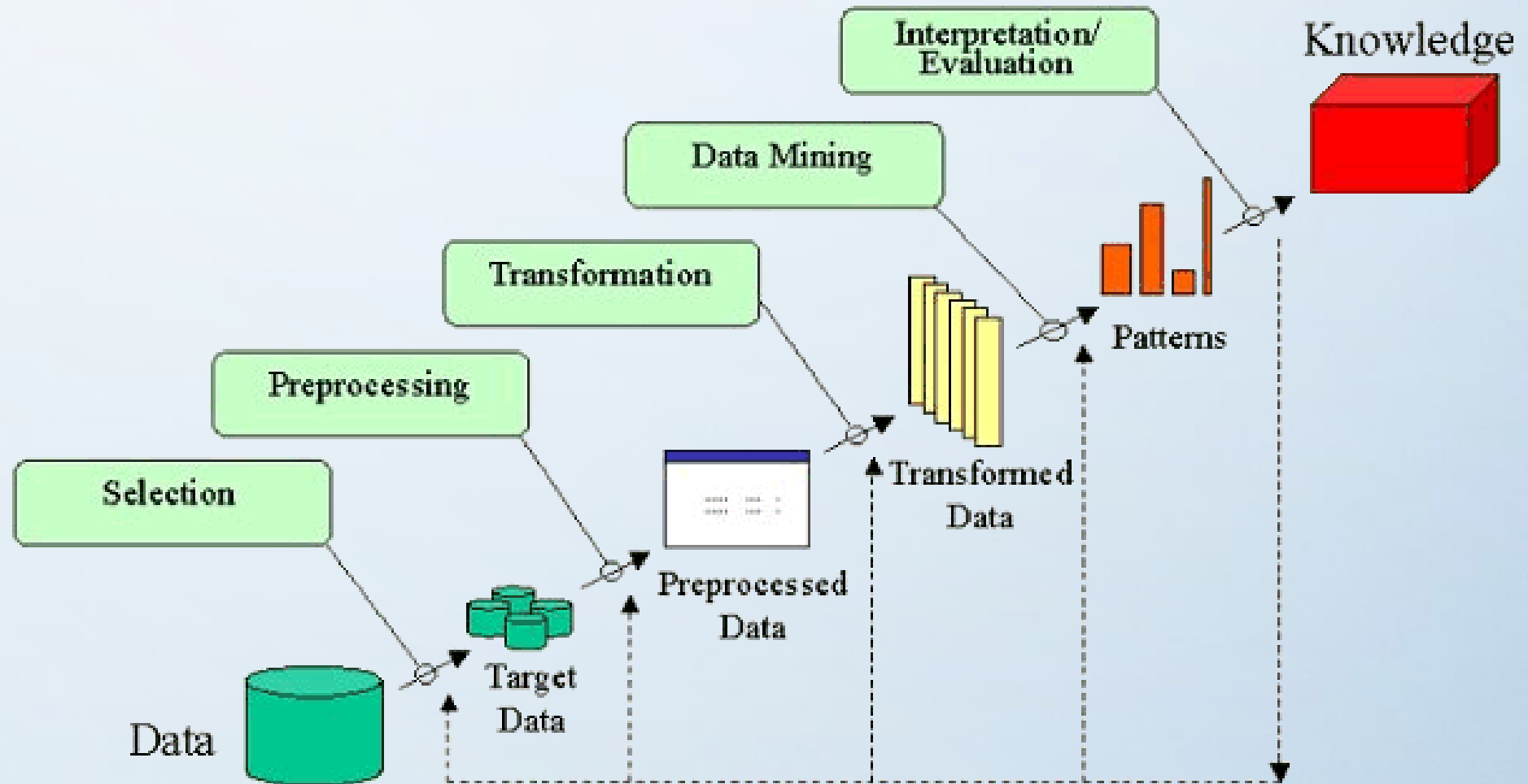
CRISP-DM

- 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.
- 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 (Evaluation)
 6. Внедрение (Deployment)



KDD Process

- Методология KDD (Knowledge Discovery in Databases) Process является аналогом CRISP-DM - http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html



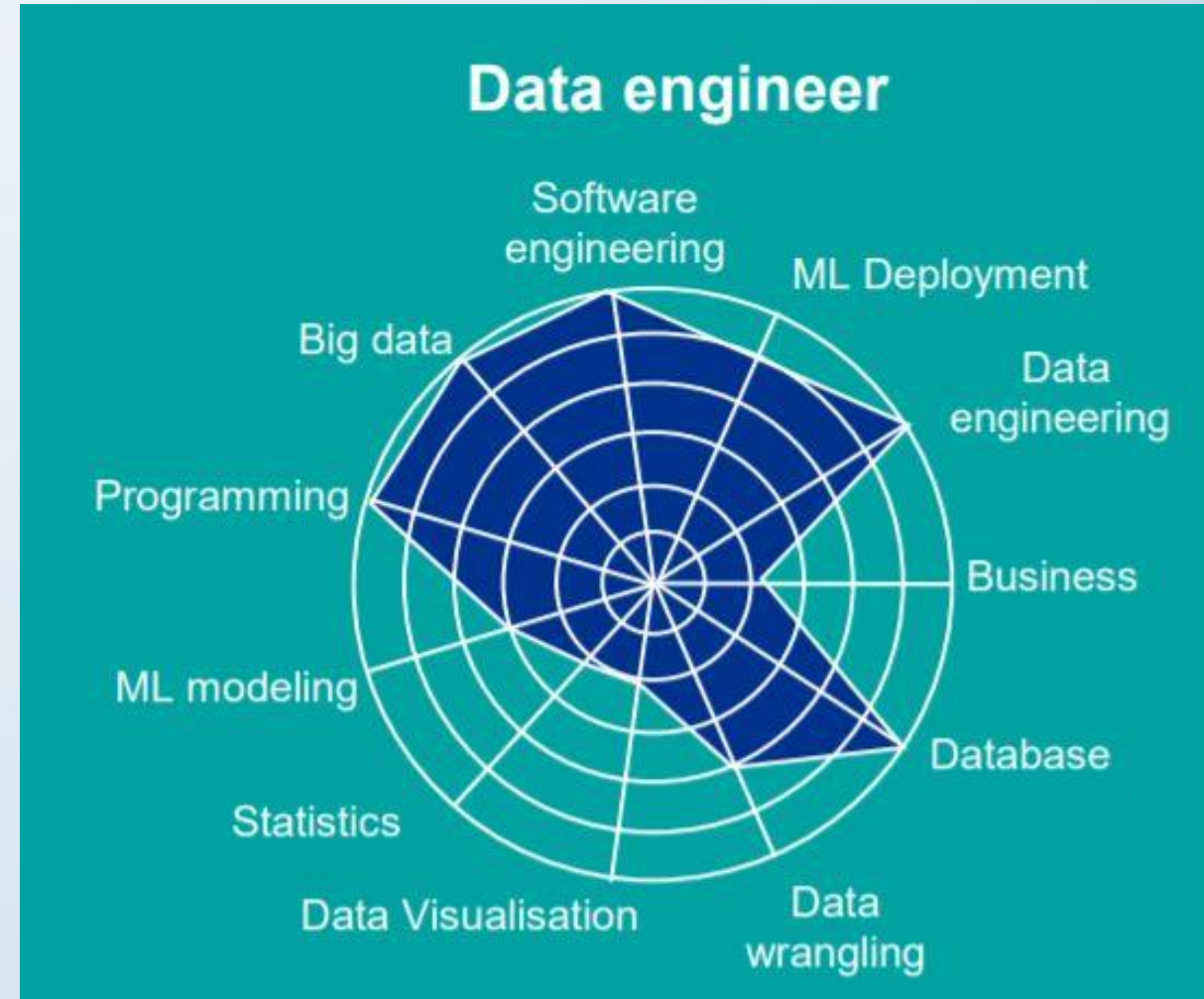
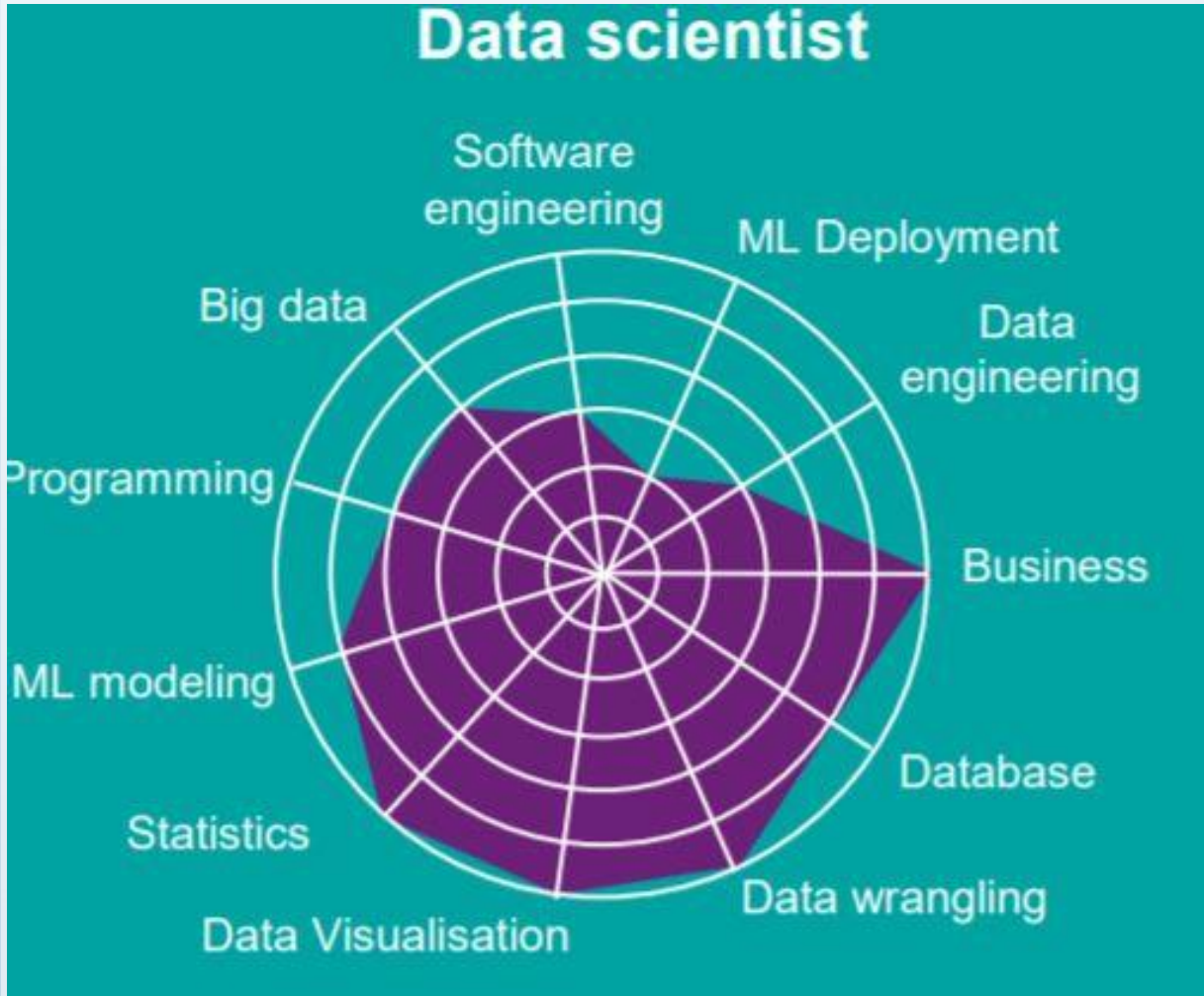
Профили специалистов в области анализа данных

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>).

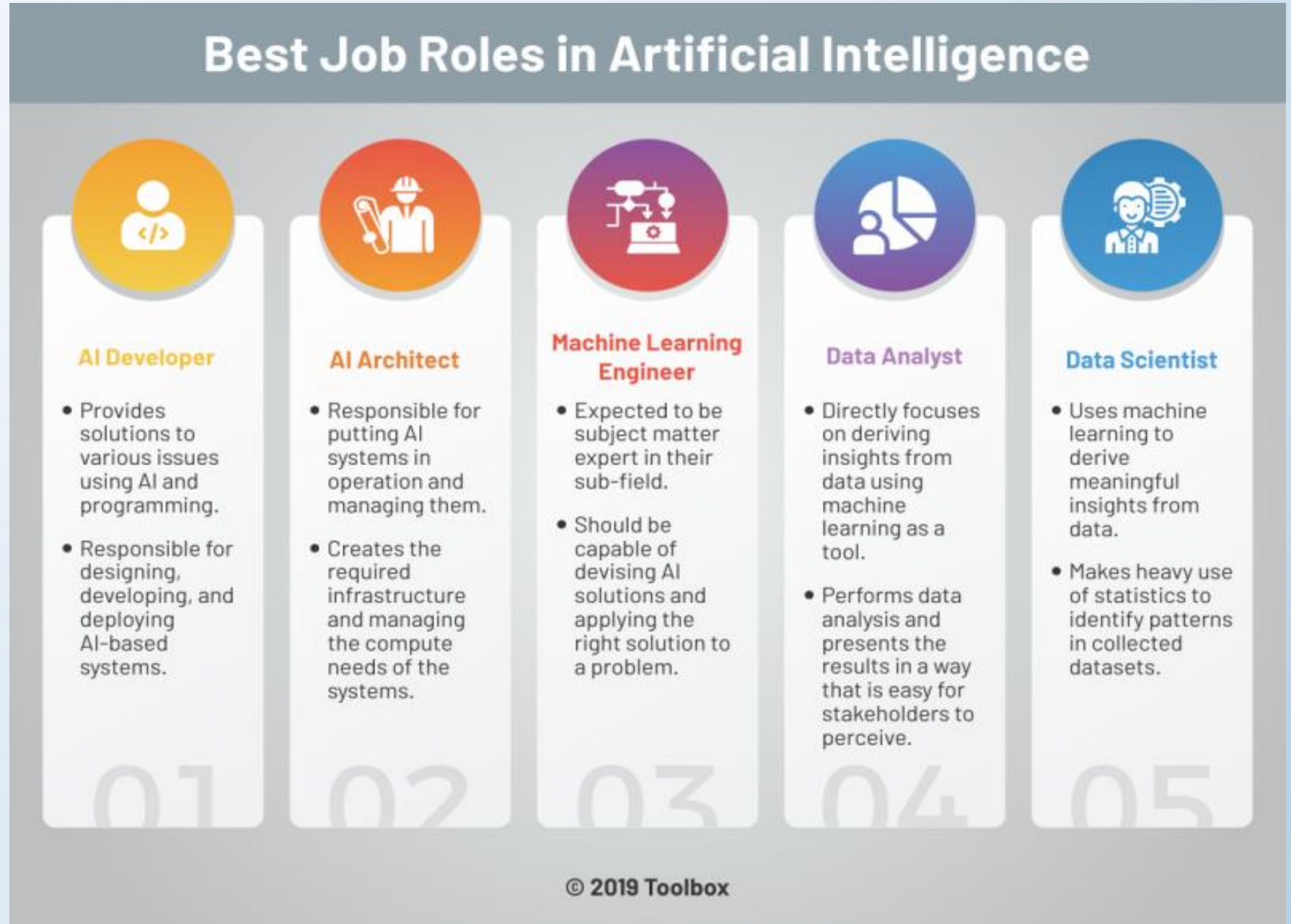
Data Scientist и Data Engineer - 2

- [Ссылка на статью](#)



Профили специалистов в области анализа данных

- [Описание 1](#)
- [Описание 2](#)
- [Описание 3:](#)



Обзор основных обучающих ресурсов в области машинного обучения

Курсы по машинному обучению

- Курсы на stepik.org:
 - <https://stepik.org/course/50305/promo>
 - <https://stepik.org/course/8057/promo>
 - <https://stepik.org/course/57839/promo>
 - <https://stepik.org/course/50352/promo>
 - <https://stepik.org/course/54098/promo>
- Видеолекции курса в ШАД - <https://yandexdataschool.ru/edu-process/courses/machine-learning>
- <https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie> - краткий курс
- <https://www.coursera.org/specializations/machine-learning-data-analysis> - специализация из 6 курсов
- <https://mlcourse.ai/> - курс по машинному обучению от [OpenDataScience](https://opendata-science.org/). Статьи на хабре: <https://habr.com/ru/company/ods/blog/322626/>
- Открытый курс «Deep Learning на пальцах» - <https://habr.com/ru/company/ods/blog/438940/>
- Курсы «Data Mining in Action» - <http://dmia.space/>

Интересные ссылки

- <http://www.machinelearning.ru> - основной ресурс по машинному обучению на русском
- <https://neurohive.io/ru/> - статьи на русском по ИИ
- <https://www.kaggle.com/>
- <https://towardsdatascience.com/>
- <https://machinelearningmastery.com/>
- <https://paperswithcode.com/sota> - публикации, сгруппированные по основным направлениям ИИ
- <https://arxiv.org/> - открытая библиотека научных публикаций Корнеллского университета

Книги

- Хорошая вводная книга непосредственно по Python.



Книги

- Хорошая практическая книга с описанием библиотек Python. Меньше примеров задач, больше описания библиотек.

Дж. Вандер Плас

Python

для сложных задач
наука о данных:
и машинное обучение



Санкт-Петербург • Москва • Екатеринбург • Воронеж
Нижний Новгород • Ростов-на-Дону • Самара • Минск

2018

Книги

- Хорошая практическая книга. Исторически была переведена первой.
- Содержит в основном примеры решения задач. В меньшей степени содержит описание библиотек.



Книги

- Рассматриваются как задачи обучения с учителем, так и задачи обучения без учителя.
- Рассматривается весь жизненный цикл анализа данных – выделение признаков, оценка качества моделей.

Андреас Мюллер, Сара Гвидо

Введение в машинное обучение с помощью Python

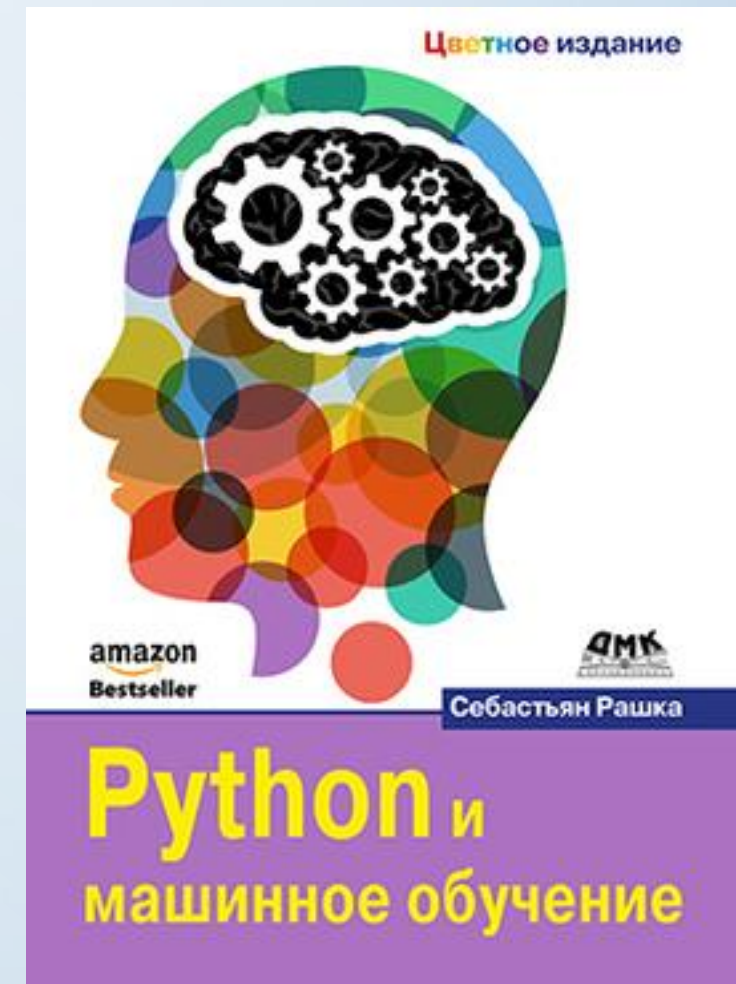
Руководство для специалистов по работе с данными



Москва
2016-2017

Книги

- Рассматривается весь жизненный цикл анализа данных – выделение признаков, оценка качества моделей.
- Рассматриваются много различных задач машинного обучения, в том числе довольно специфических.



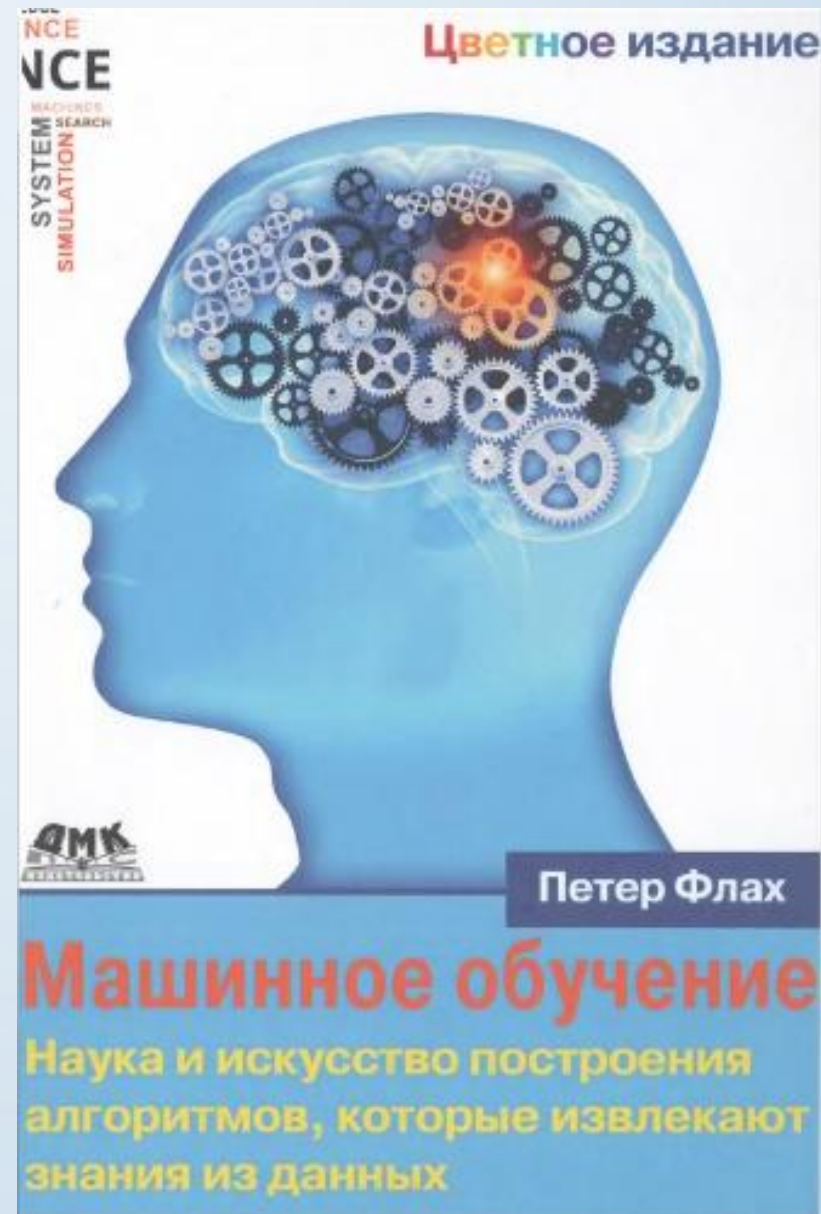
Книги

- Хорошая вводная книга в основном по методам машинного обучения. Методы разбираются достаточно детально.
- Примеров кода относительно немного.



Книги

- Теоретический учебник. Разбираются теоретические основы машинного обучения на основе большого количества примеров.
- Не привязан к конкретному языку программирования.



Часть 2. Основные задачи машинного обучения и способы их решения с использованием современных программных средств на основе языка программирования Python.

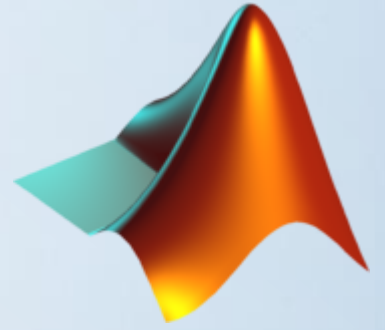
Языки и инструменты

Языки и инструменты

- 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.

Инструменты для Python

- Основной инструмент – Jupyter notebooks (IPython notebooks). Напоминает одноколоночную таблицу Excel. В ячейках можно писать код, который сразу выполняется:
 - [28 Jupyter Notebook tips, tricks, and shortcuts](#)
 - Для документации используется упрощенный язык разметки [Markdown](#). [Краткая справка](#) по использованию в ноутбуках.
- Технология Jupyter notebooks (в различных вариациях) применяется в нескольких дистрибутивах:
 - [Conda](#) (оффлайнный дистрибутив, устанавливается локально на компьютер).
 - <https://datalore.io/> – онлайнный сервис от JetBrains.
 - <https://colab.research.google.com> - онлайнный сервис от Google.

Наборы данных (датасеты)

Некоторые типы шкал измерений

- 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.
 - Частный случай – бинарная шкала $\{0, 1\}$, $\{\text{False}, \text{True}\}$.
- 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 ил. 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.
- **Количественный (действительный) признак, который является действительным числом. Основной вид шкалы, к которому пытаются свести все остальные.**



Пример датасета

Атрибуты (свойства, поля данных)

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

Классификация и постановки основных задач машинного обучения

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)

Background

Types of machine learning

Reinforcement Learning:
*feedback to algorithm
when it does something
right or wrong*



Example:
Child gets
feedback 'on the
job' when it does
something right
or wrong

**Supervised
Learning**

Supervised Learning:
*pre-labelled data trains a
model to predict new
outcomes*

Example: Sorting
LEGO blocks by
matching them with
the colour of the bags



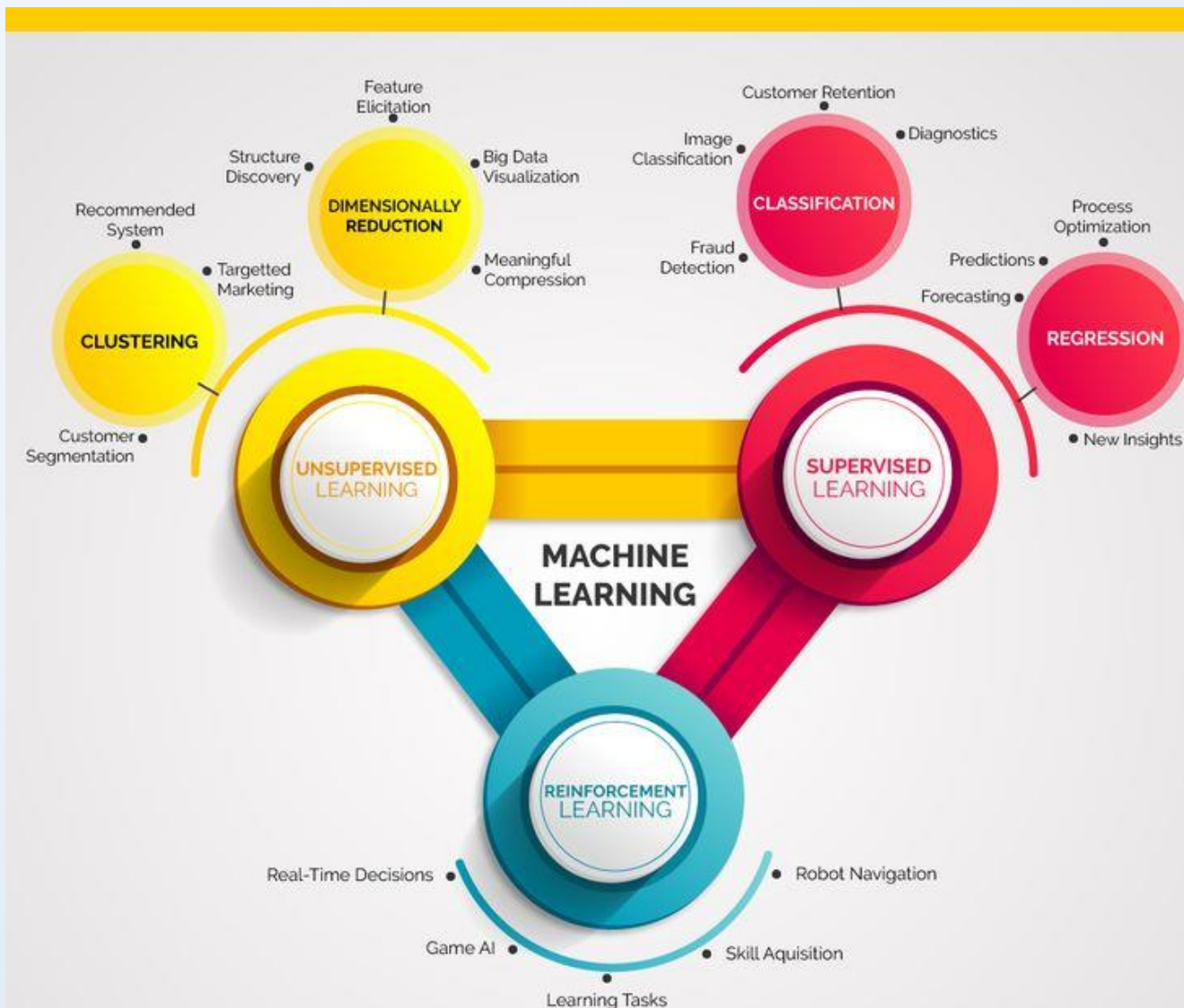
**Machine
Learning**

**Reinforcement
Learning**

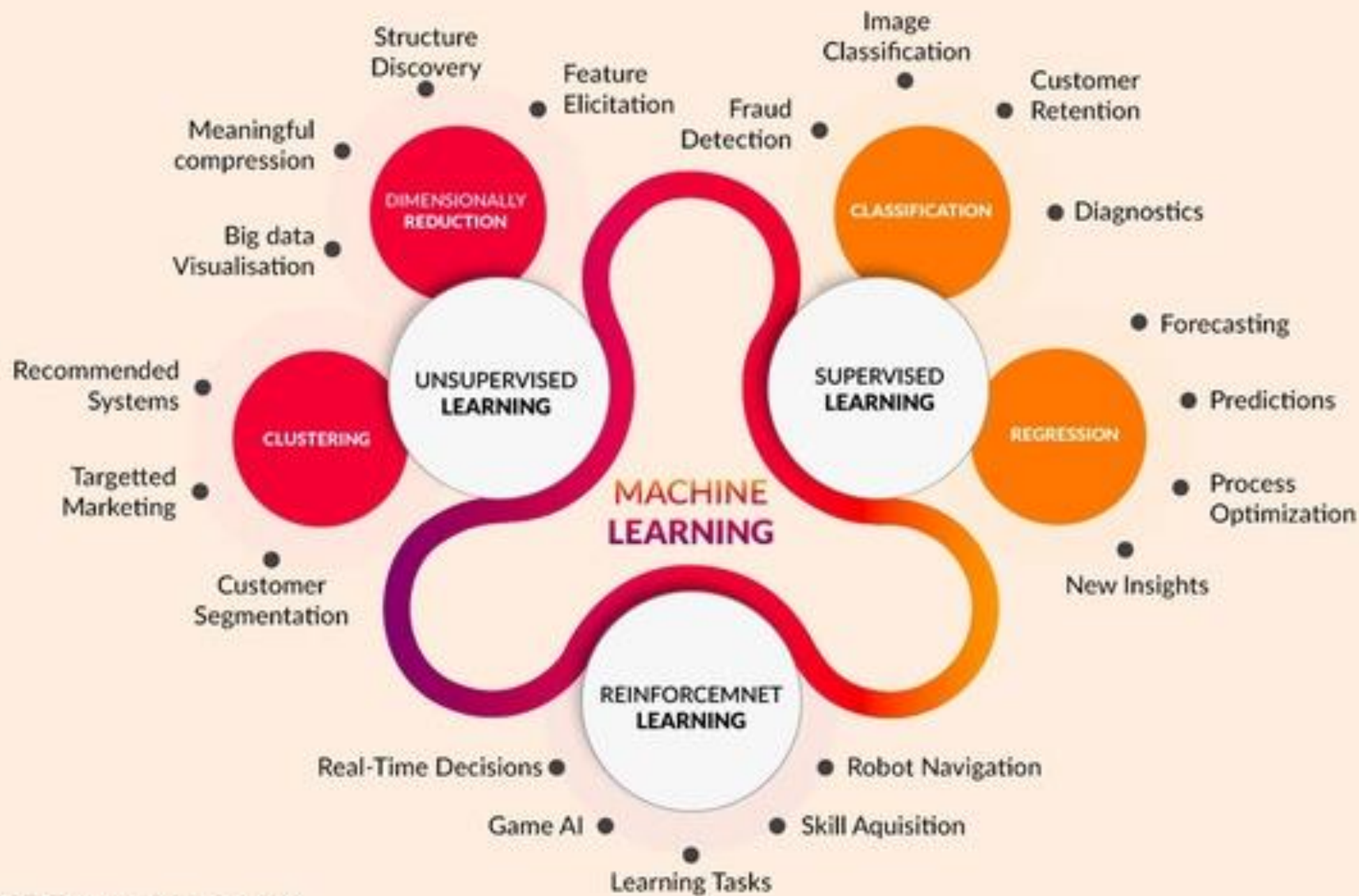
**Unsupervised
Learning**



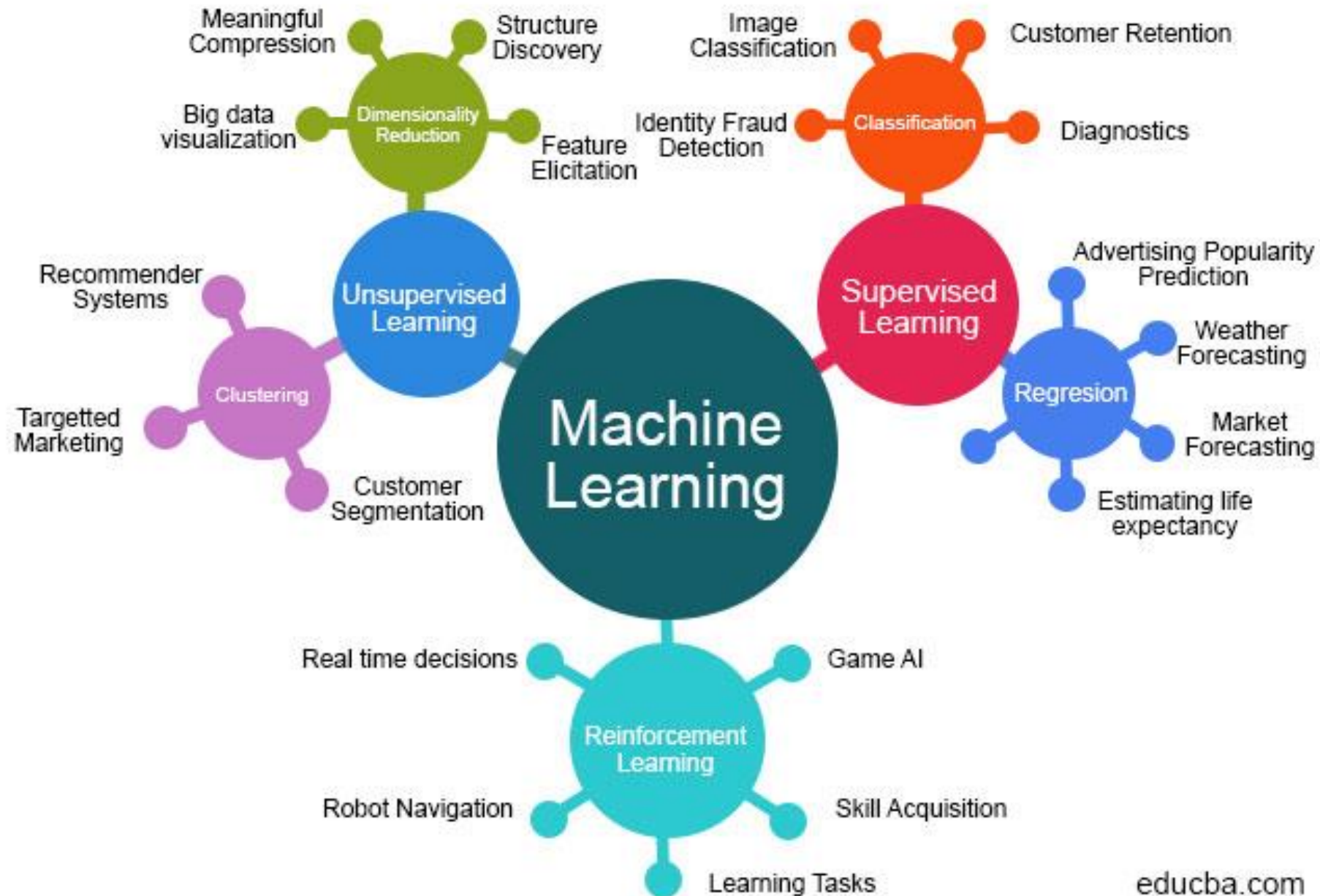
Unsupervised Learning:
*Non-labelled data self
organises to predict new
outcomes (e.g. clustering)*



TYPES OF MACHINE LEARNING ALGORITHMS

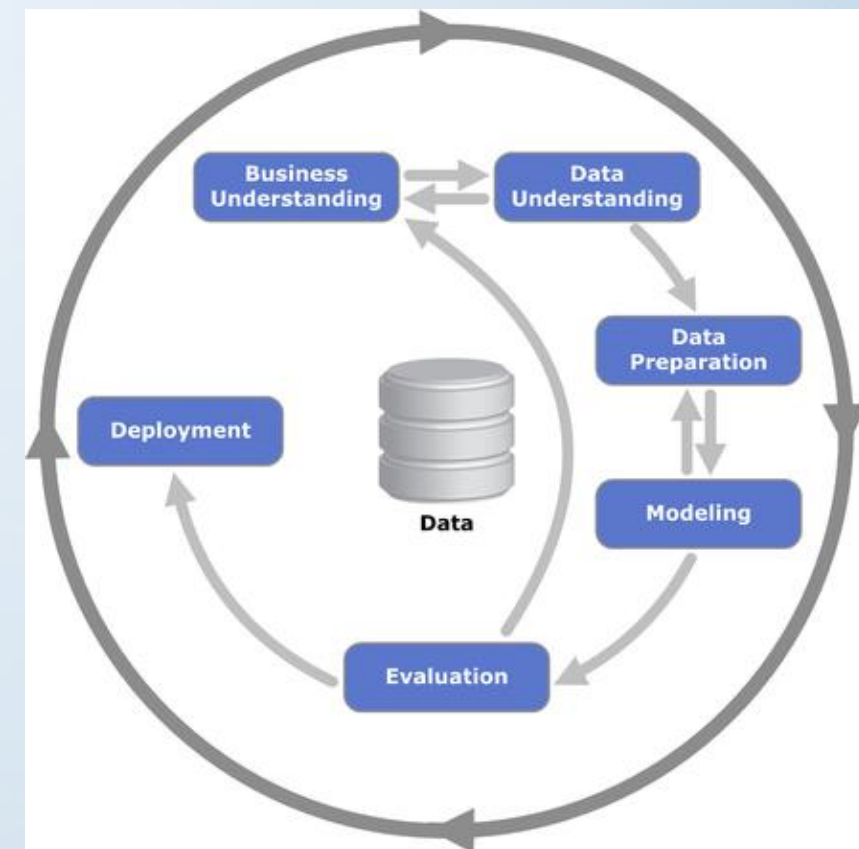


Machine Learning Algorithms



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)



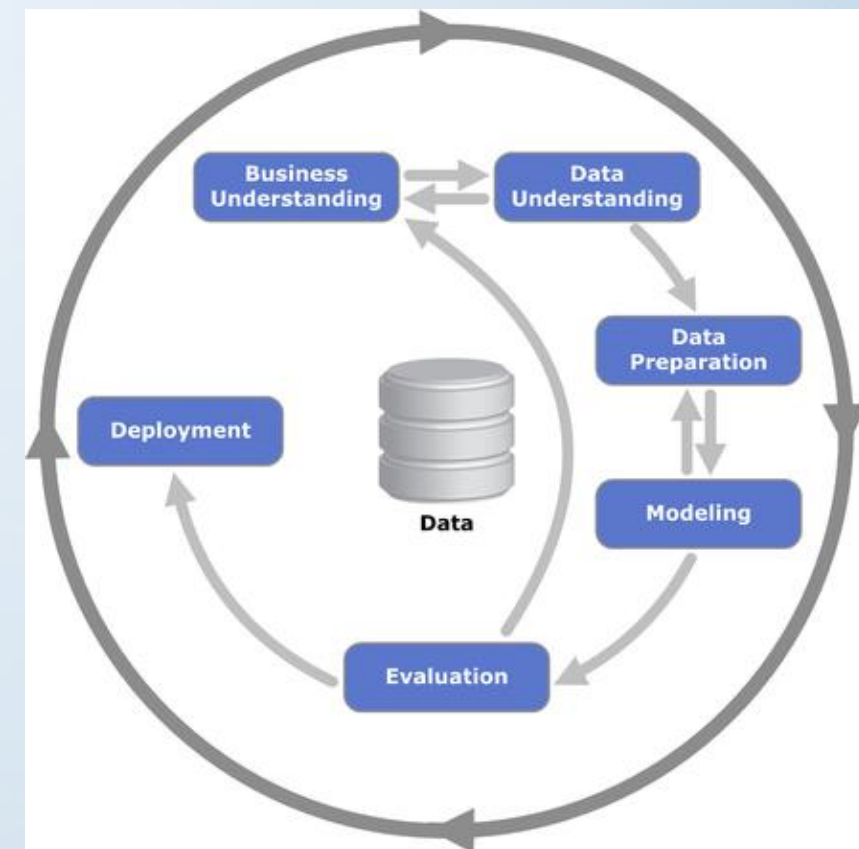
Изображения (изображения-пиксели)



- 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 (Y-значение по номинальной шкале)

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 (Y-множество значений по номинальной шкале)

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						

- Пример метрики 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.



Обучение с подкреплением

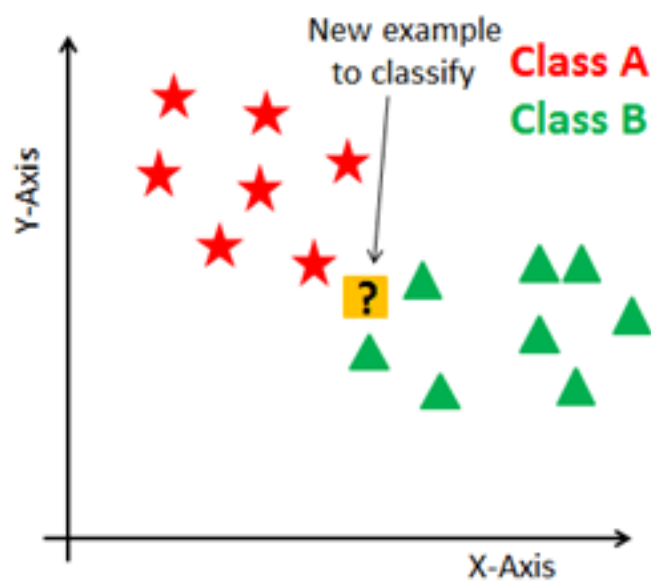
- Обучение с обратной связью, с опосредованным учителем.
- Алгоритм обучается, взаимодействуя с некоторой средой. Откликом среды являются сигналы подкрепления, поэтому такое обучение является частным случаем обучения с учителем, но учителем является среда или её модель.
- Частным случаем обучения с подкреплением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



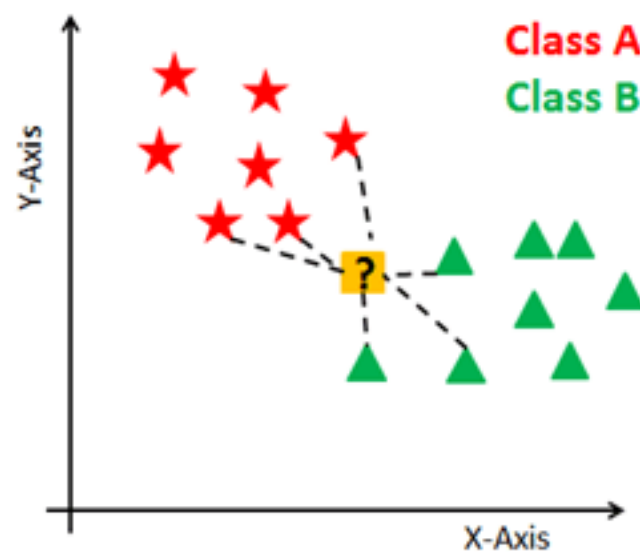
Научно-практический семинар

Часть 1. Реализация метода ближайших соседей (на основе языка программирования Python) как одного из наиболее простых методов машинного обучения.

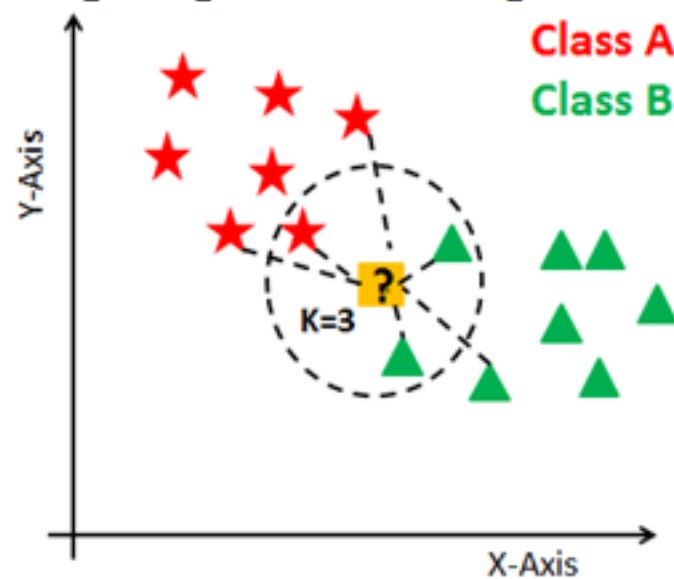
Initial Data



Calculate Distance



Finding Neighbors & Voting for Labels



Часть 2. Решение задачи анализа
тональности текста методами
машинного обучения на основе языка
программирования Python.

