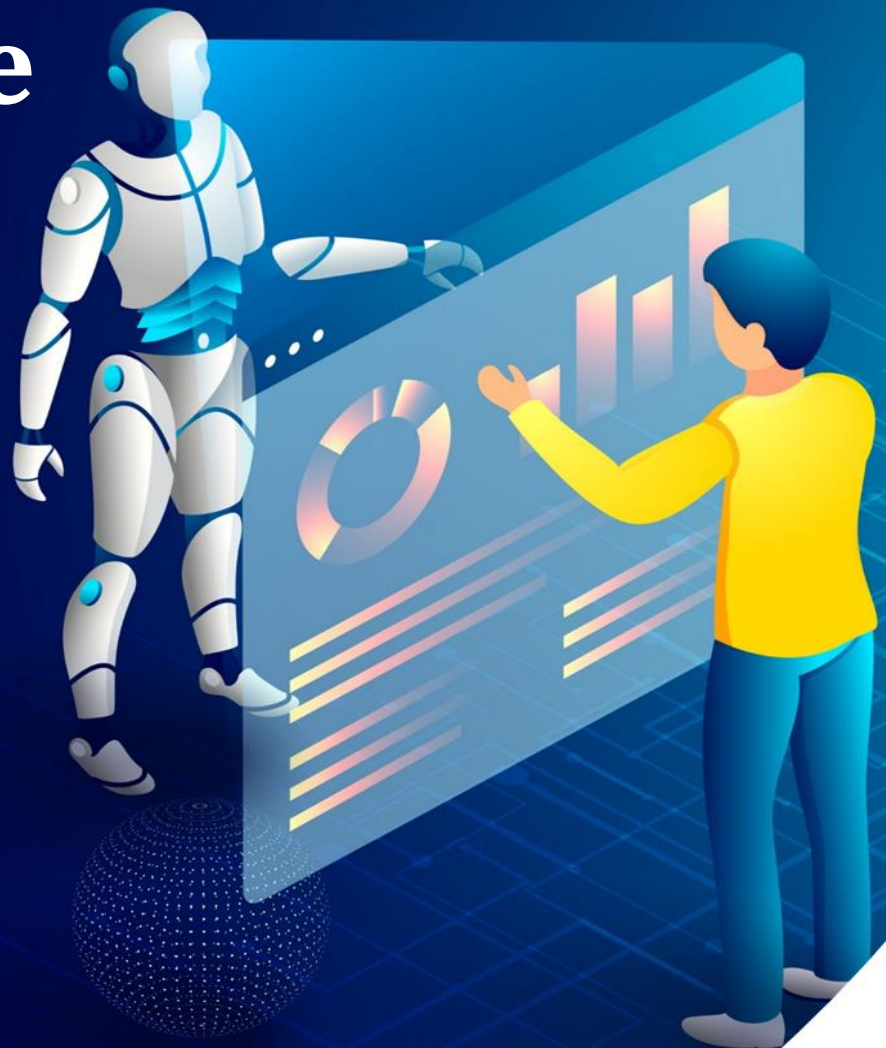
Введение в машинное обучение



Дмитрий Перец

Почта: Perets.Dmitry@gmail.com

Telegram: @Train_Brain

Познакомимся?

ДМИТРИЙ ПЕРЕЦ

РУКОВОДИТЕЛЬ ПО ML И КЛИЕНТСКОЙ АНАЛИТИКЕ

@Train_Brain

2021-по н.в.	Yota **Архитектор машинного обучения**
2020-2021	Научно-технический центр Газпромнефти **Руководитель направления**
2016-2020	Научно-технический центр Газпромнефти **Главный специалист**
2014-2016	ИЭФБ РАН имени Сеченова **Ведущий инженер**

если есть желание узнать чуть больше, то:

Интервью «10 вопросов data scientist'у»

Мои публикации в Scopus

Science Slam на тему машинного обучения в нефтянке

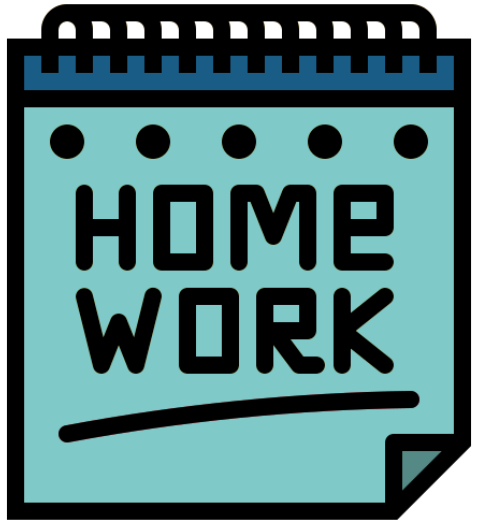
Правила игры

Правила игры

Дискутируйте, задавайте вопросы! Самый глупый вопрос – это …?

Правила игры

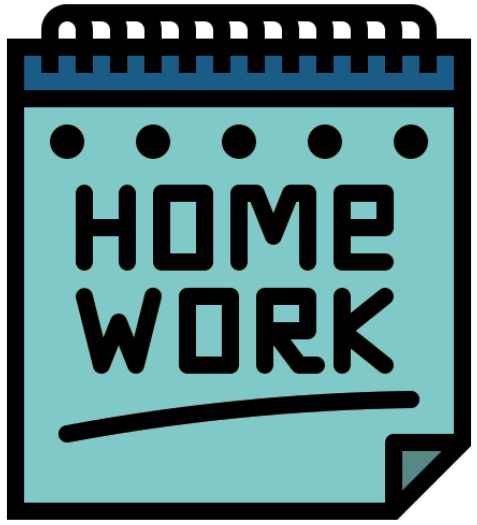
Дискутируйте, задавайте вопросы! Самый глупый вопрос – это …?



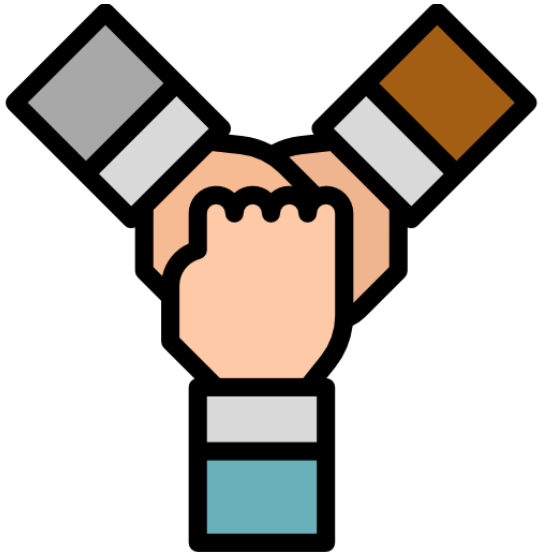
Выполняйте задания, это полезно не только в рамках курса

Правила игры

Дискутируйте, задавайте вопросы! Самый глупый вопрос – это …?



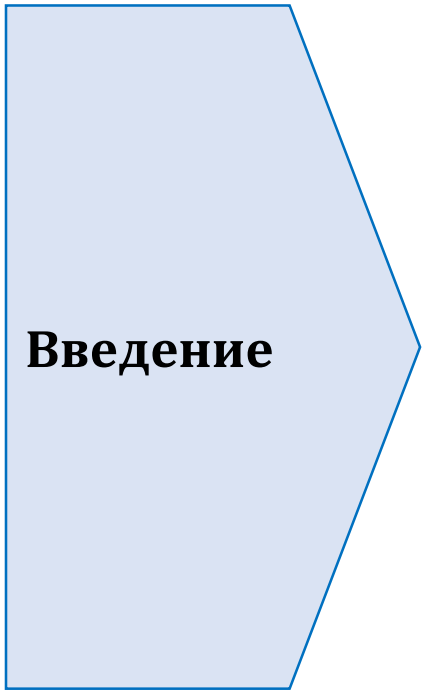
Выполняйте задания, это полезно не только в рамках курса



Работайте в группах

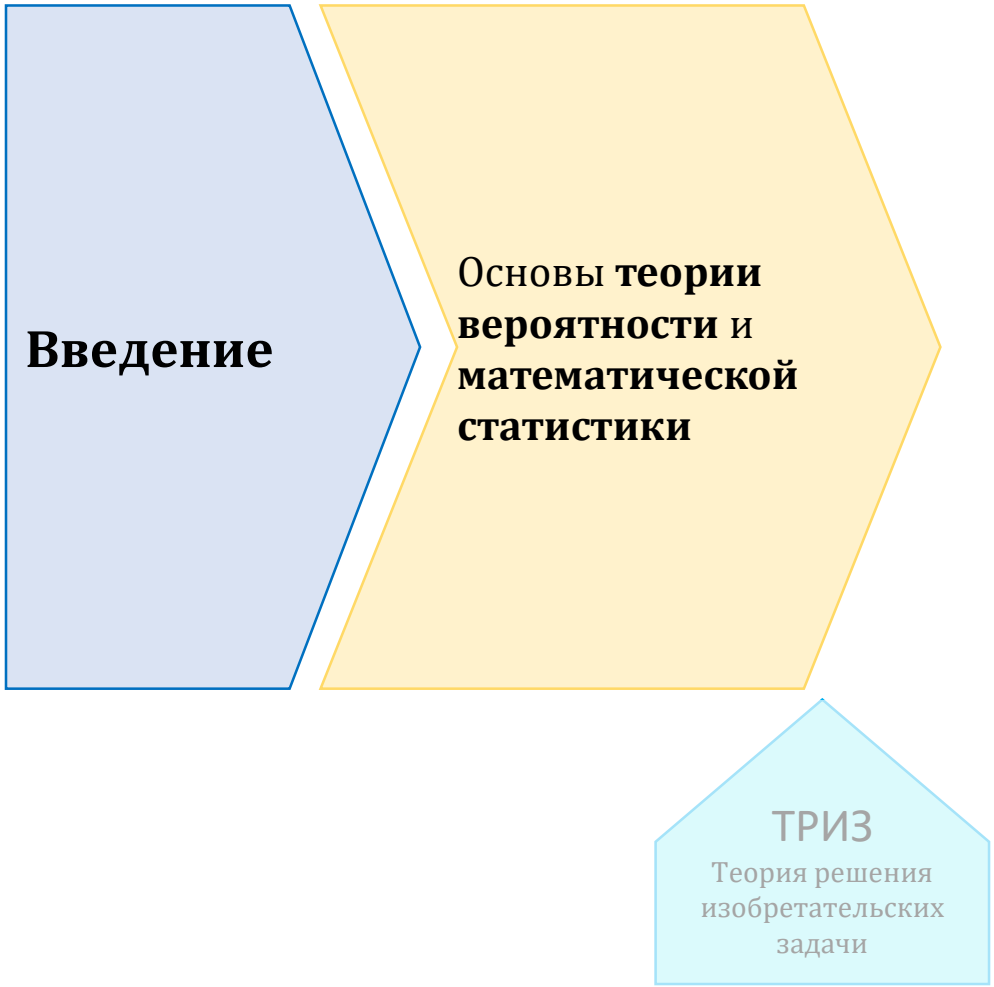
Краткая структура курса

Краткая структура курса



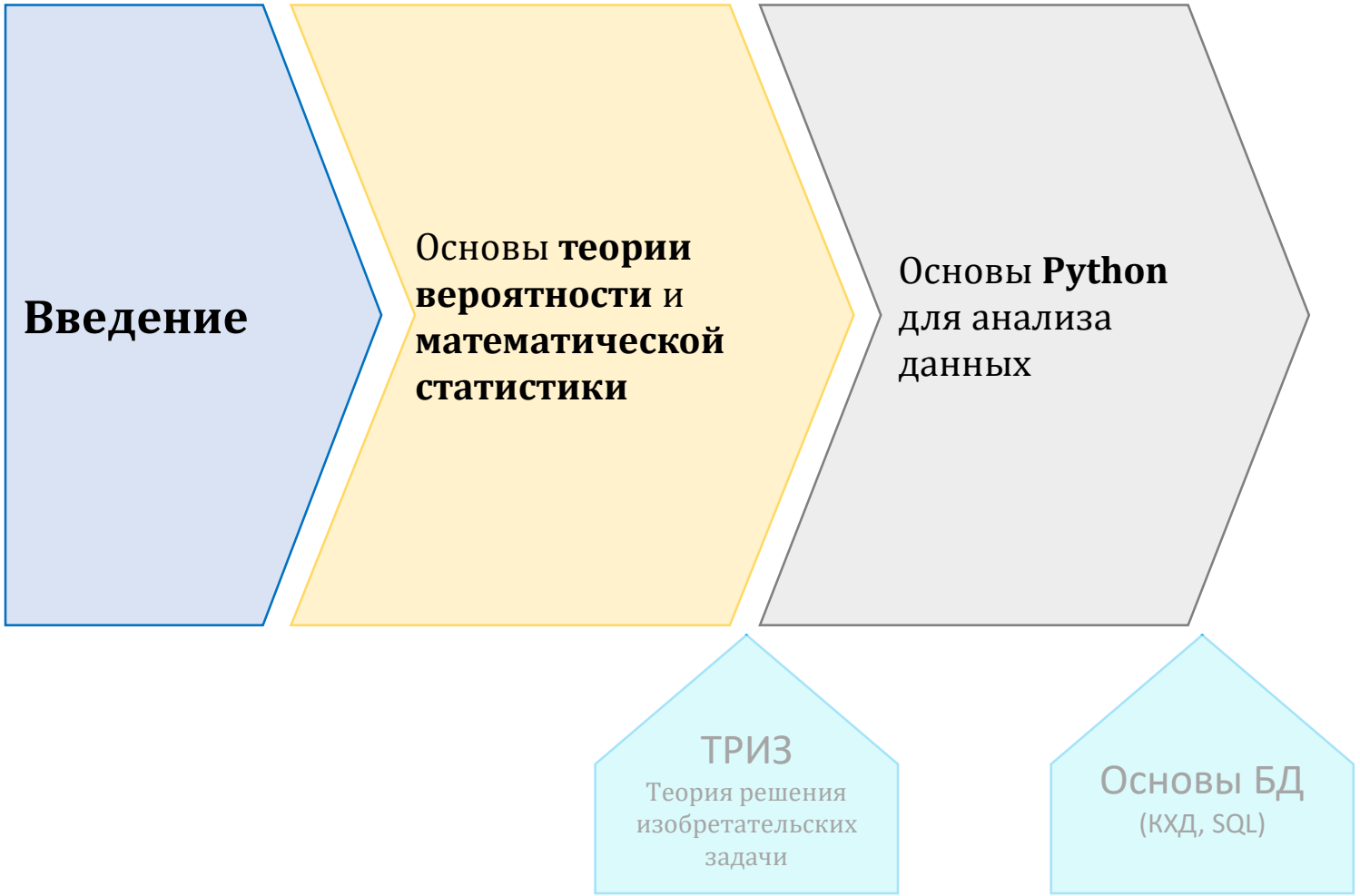
Краткая структура курса

Краткая структура курса



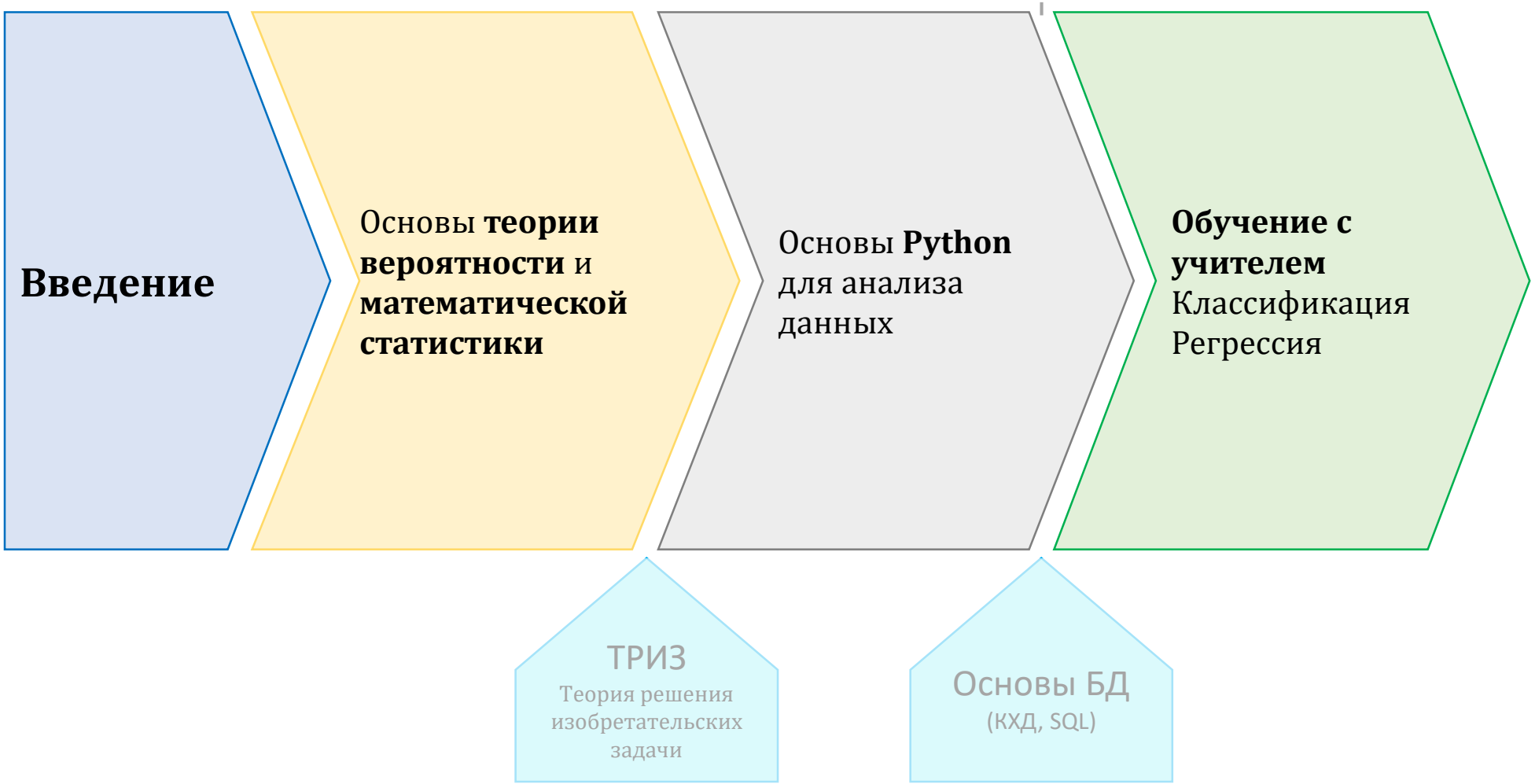
Краткая структура курса

Краткая структура курса



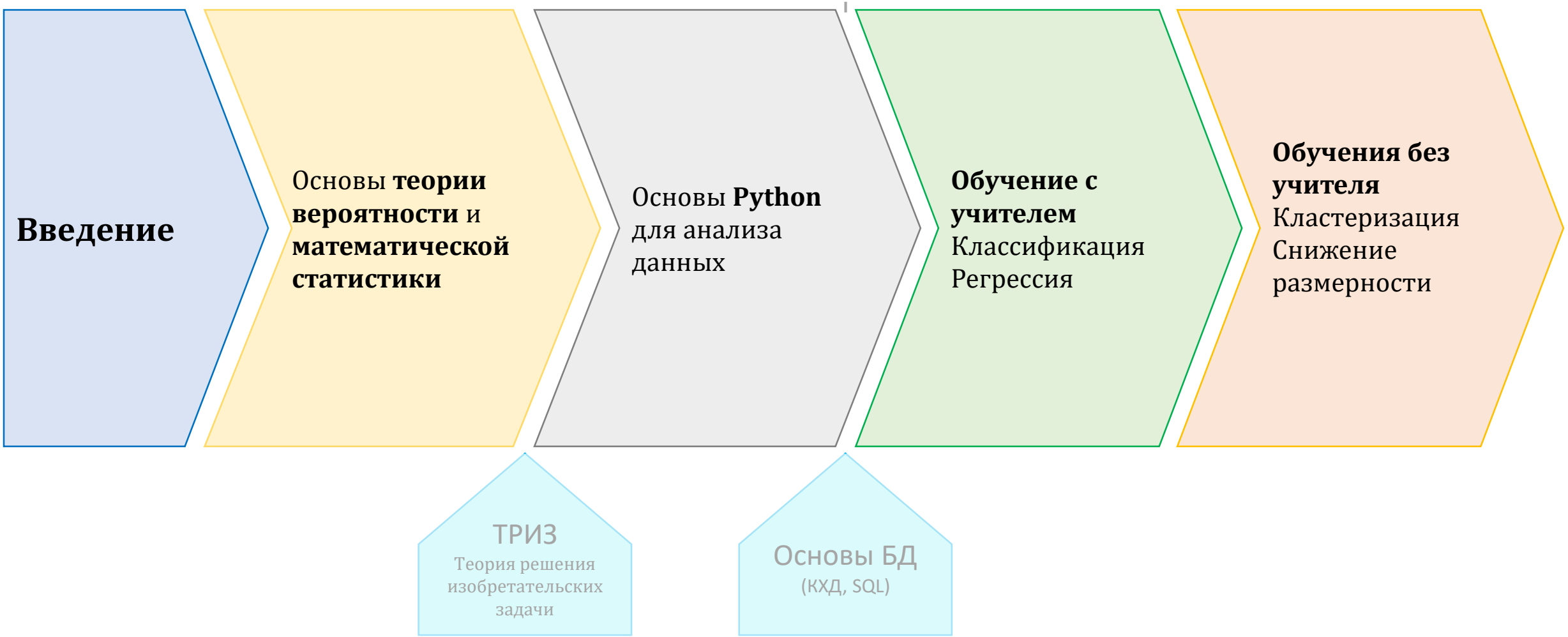
Краткая структура курса

Машинное обучение



Краткая структура курса

Машинное обучение



Убираем телефоны

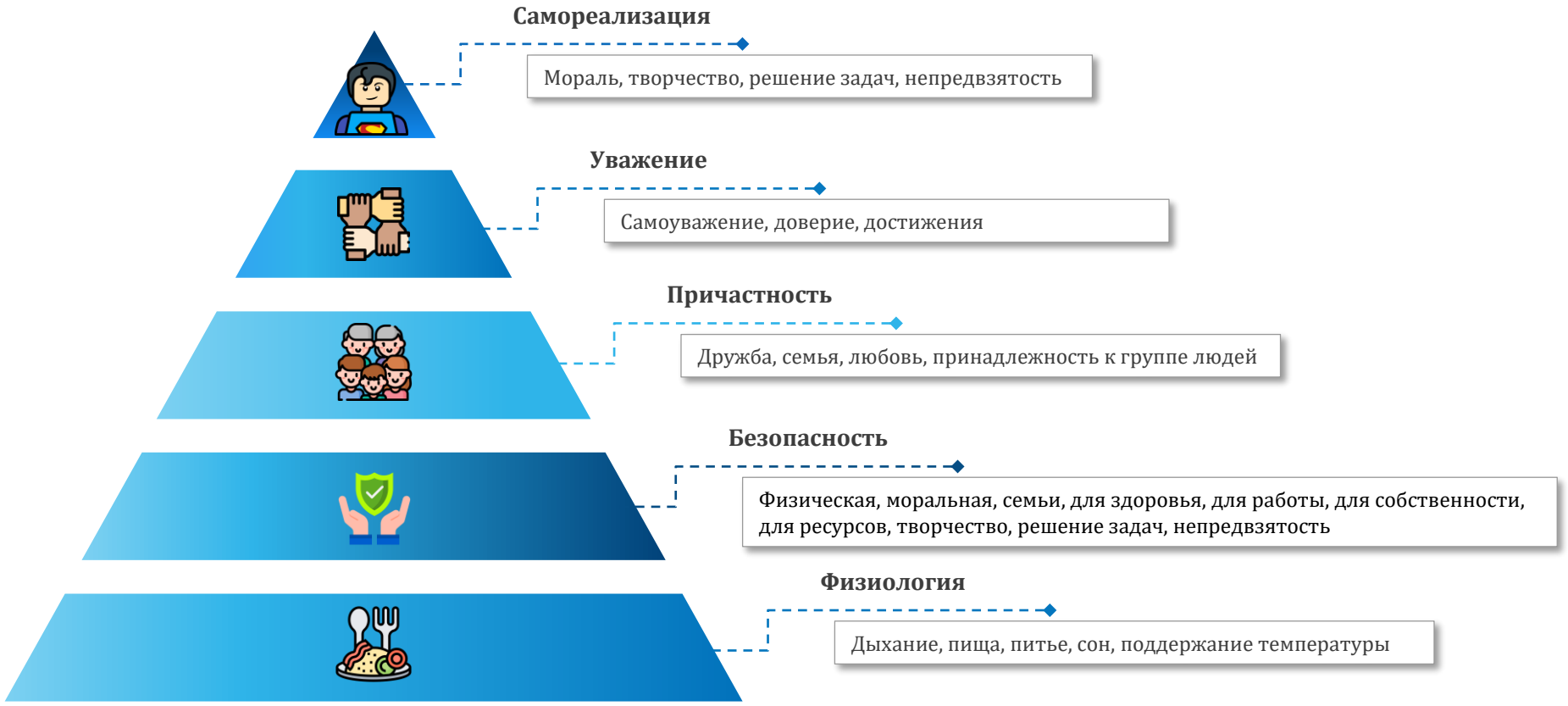
Достаем двойные листочки

Достаем телефоны

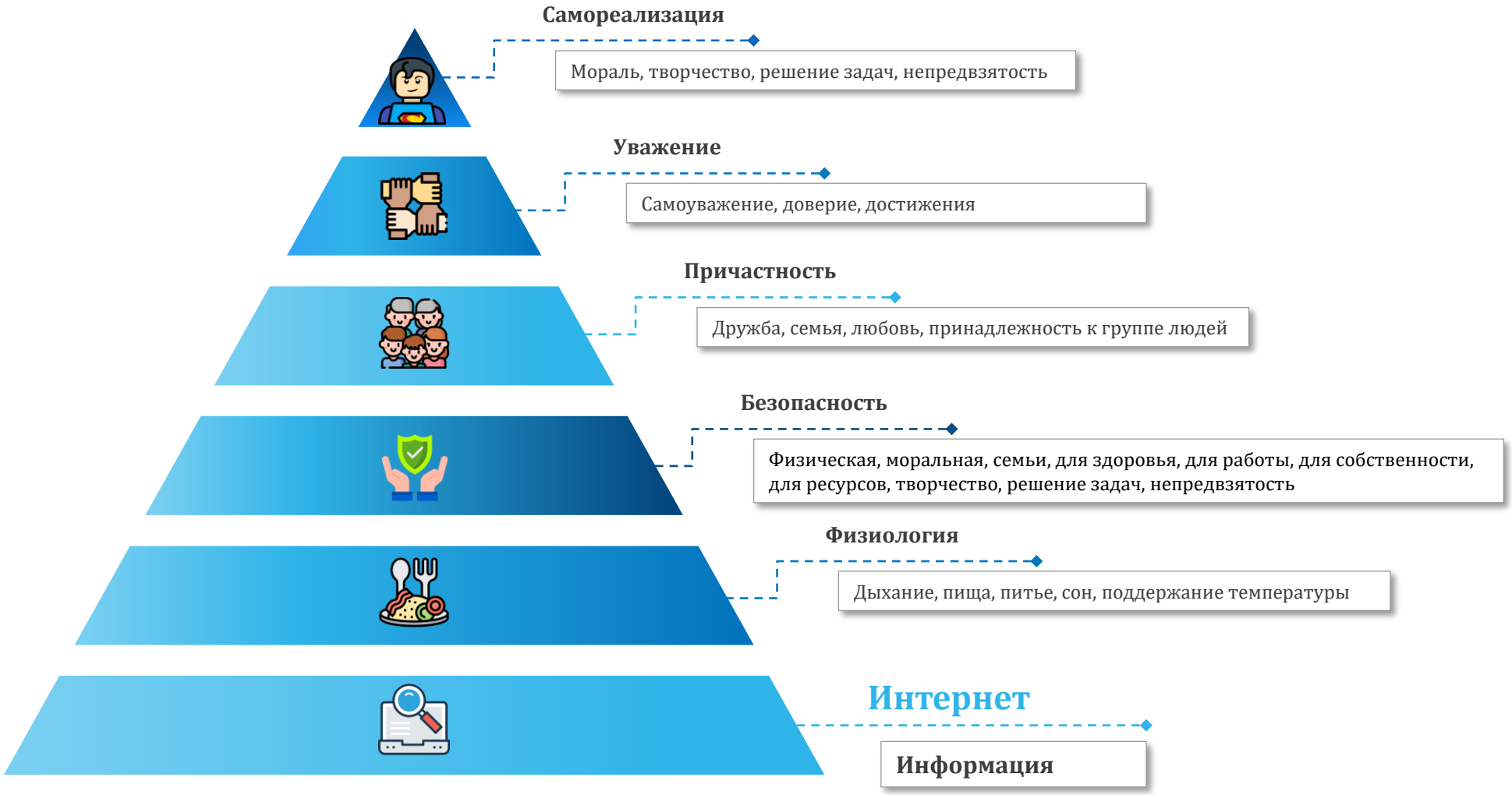
Что для вас ИИ, машинное обучение, примеры?

Немного отвлечемся

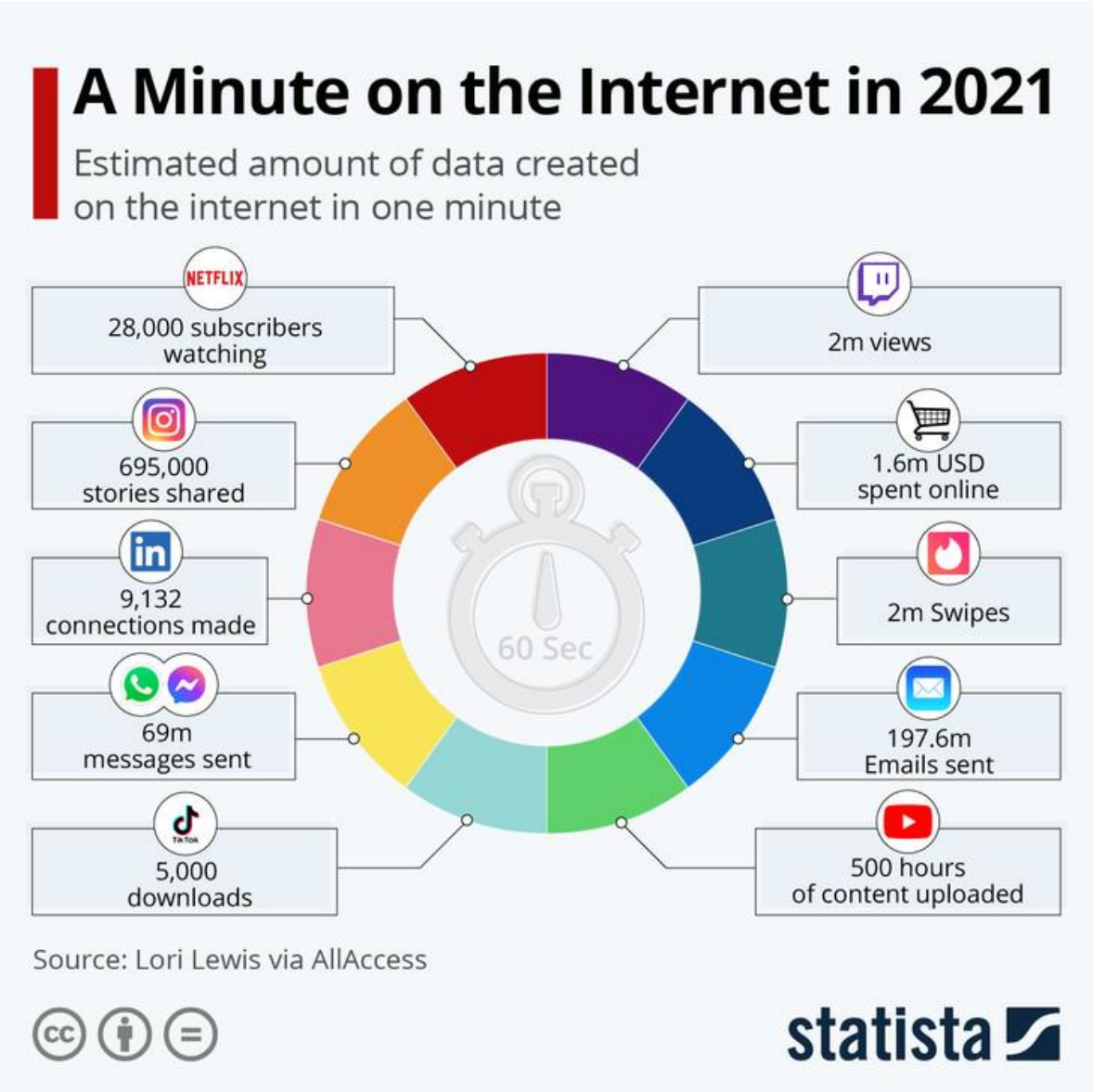
Пирамида потребностей по Маслоу



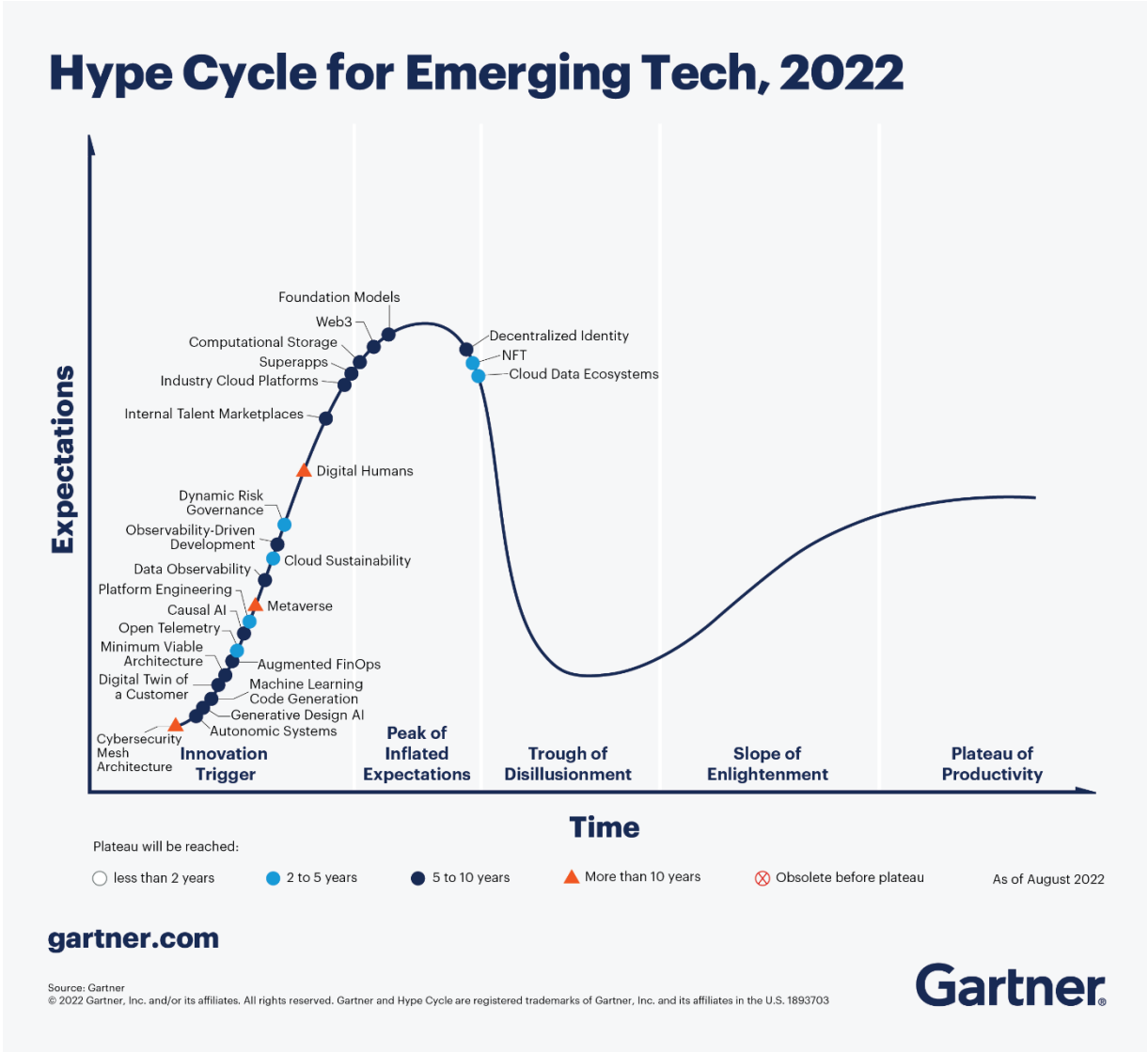
Пирамида потребностей по Маслоу: версия 2.0

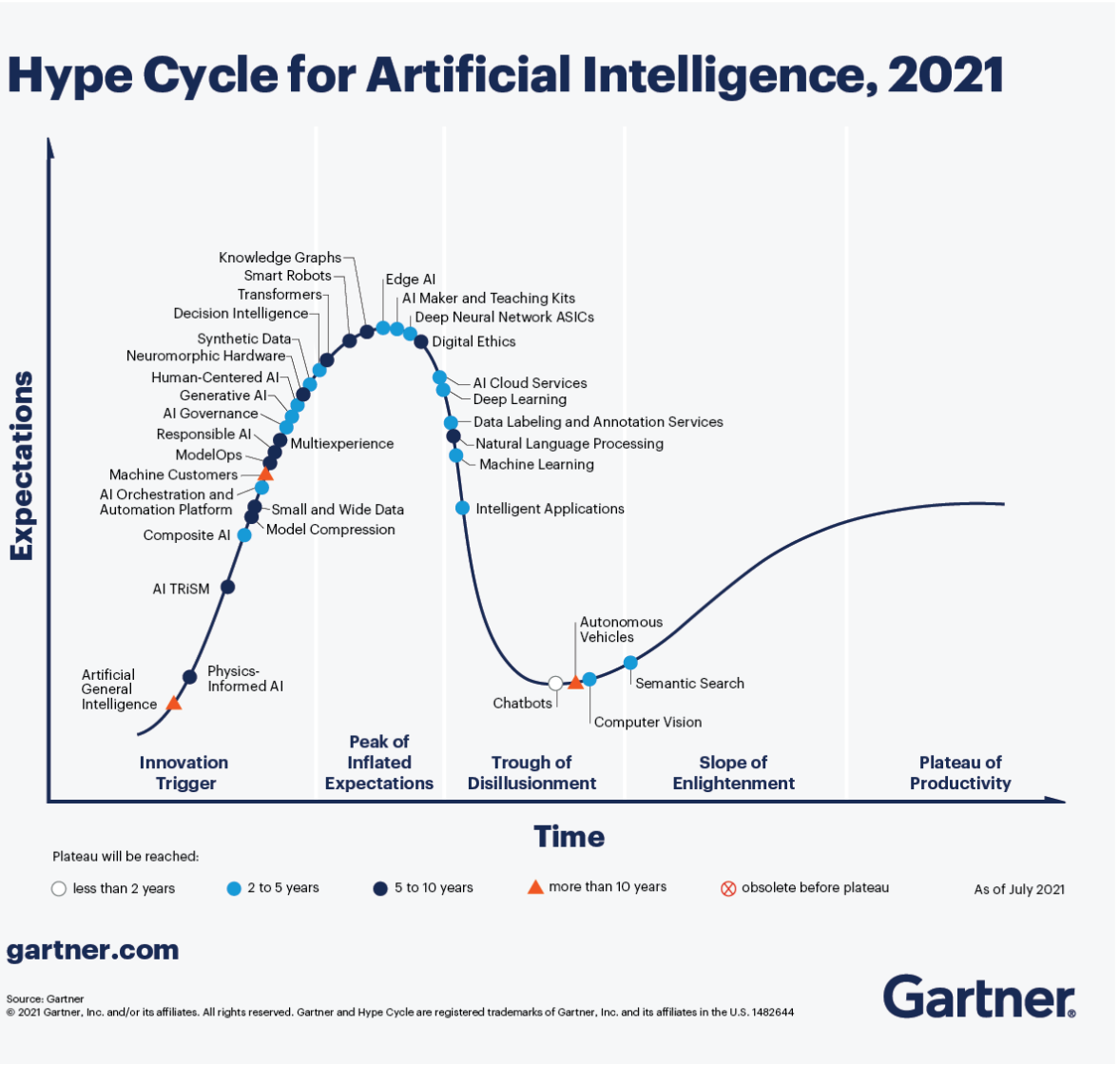


Почему машинное обучение?



Кривая Гартнера





А что на рынке труда?

Позиция	Средняя зарплата, руб.	Медианная зарплата, руб.	Вакансий с зарплатой	Всего вакансий
Senior Data Scientist	308333	251000	15	106
Middle Data Scientist	235800	131000	5	27
Junior Data Scientist	93400	71000	5	15

Терминология

Немного терминологии 1/5

Искусственный интеллект наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ.

Немного терминологии 2/5

Искусственный интеллект наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ.

Машинное обучение
класс методов искусственного интеллекта, основной чертой которых является не прямое решение задачи, а обучение на исторических данных.

Немного терминологии 3/5

Искусственный интеллект
наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ.

Машинное обучение
класс методов искусственного интеллекта, основной чертой которых является не прямое решение задачи, а обучение на исторических данных.

Нейронные сети

математическая модель,
построенная по принципу
организации и функционирования
биологических нейронных сетей.

Немного терминологии 4/5

Искусственный интеллект
наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ.

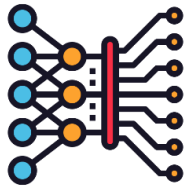
Машинное обучение
класс методов искусственного интеллекта, основной чертой которых является не прямое решение задачи, а обучение на исторических данных.

Нейронные сети

математическая модель, построенная по принципу организации и функционирования биологических нейронных сетей.

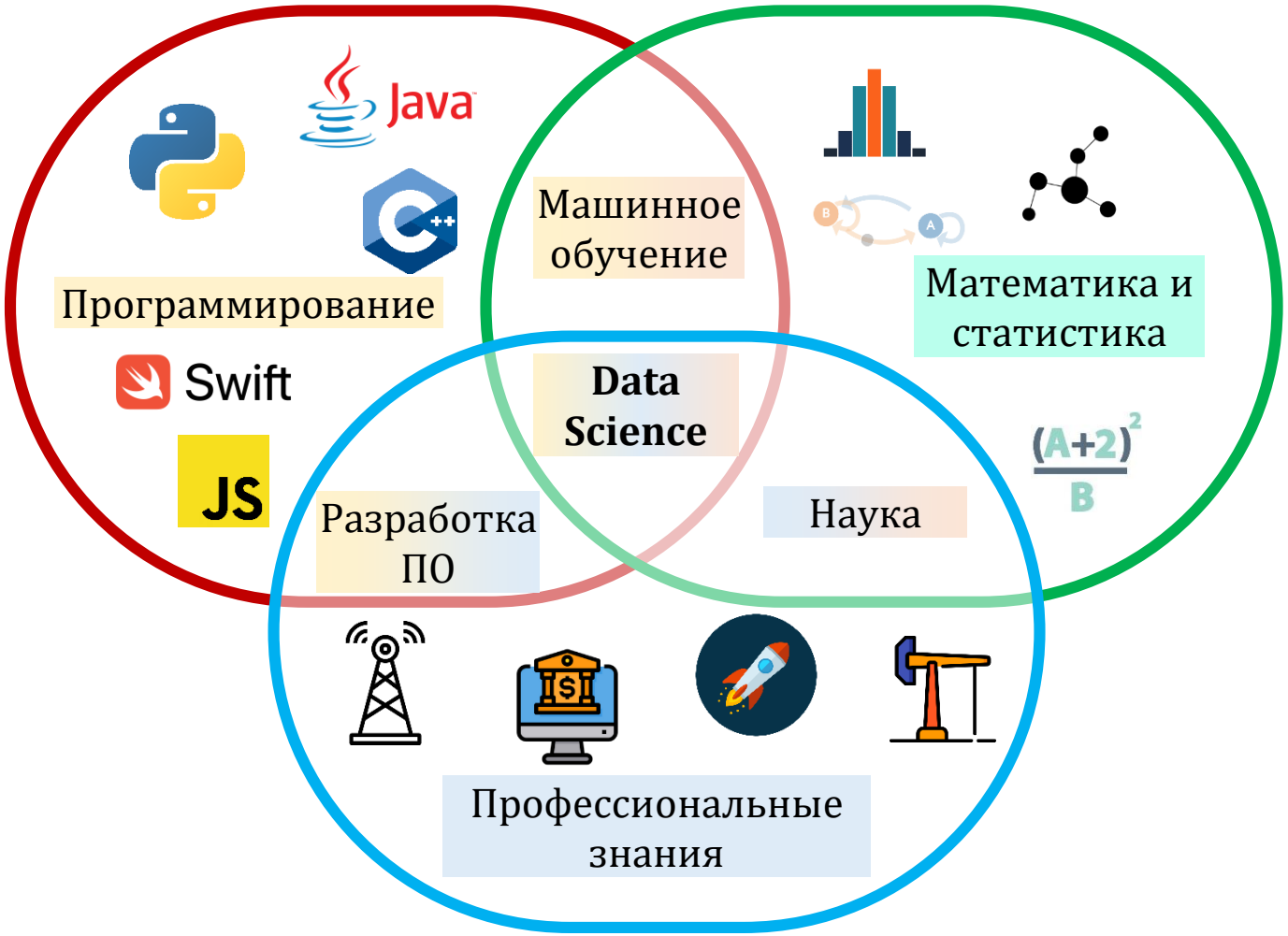
Глубинное обучение

архитектура нейросетей, один из подходов к их построению и обучению. На практике сегодня мало кто отличает, где глубокие нейросети, а где не очень.



Немного терминологии 5/5

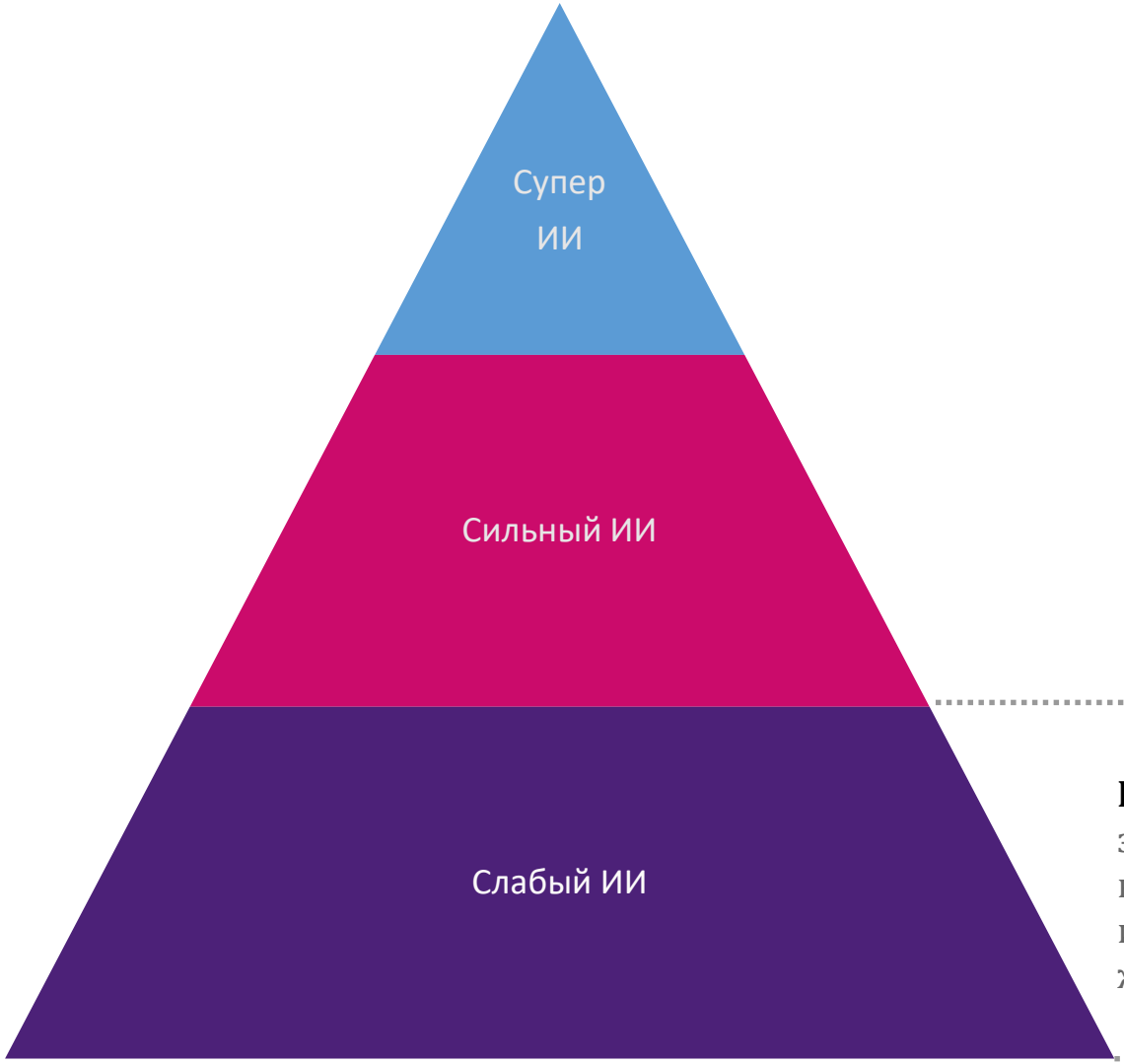
Data Science (Интеллектуальный анализ данных) - это направление информационных технологий, охватывающее всю область проблем, связанных с извлечением знаний из массивов данных.



Почему появилось такое направление:

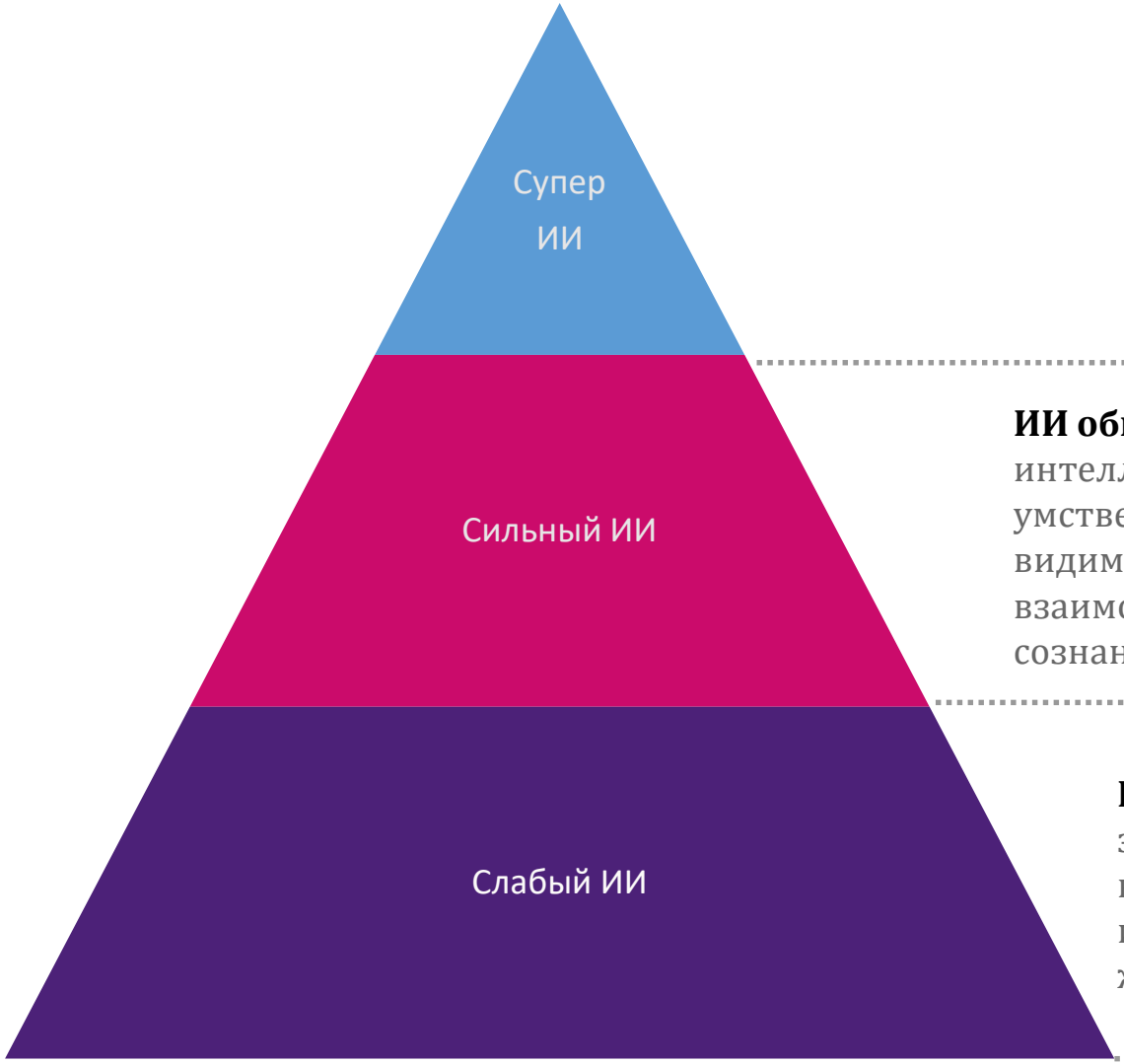
- данные могут быть **неточными, неполными (содержать пропуски)**, противоречивыми, разнородными, косвенными.
- процессы переработки сырых данных в информацию, а информации в знания уже не могут быть выполнены по старинке вручную, и требуют **нетривиальной автоматизации.**

Какой бывает ИИ?



ИИ узкого назначения, также известный как **слабый**, — это ИИ в сегодняшнем понимании. Он запрограммирован на выполнение одной задачи — будь то мониторинг погоды, игра в шахматы или анализ данных для написания журналистских репортажей.

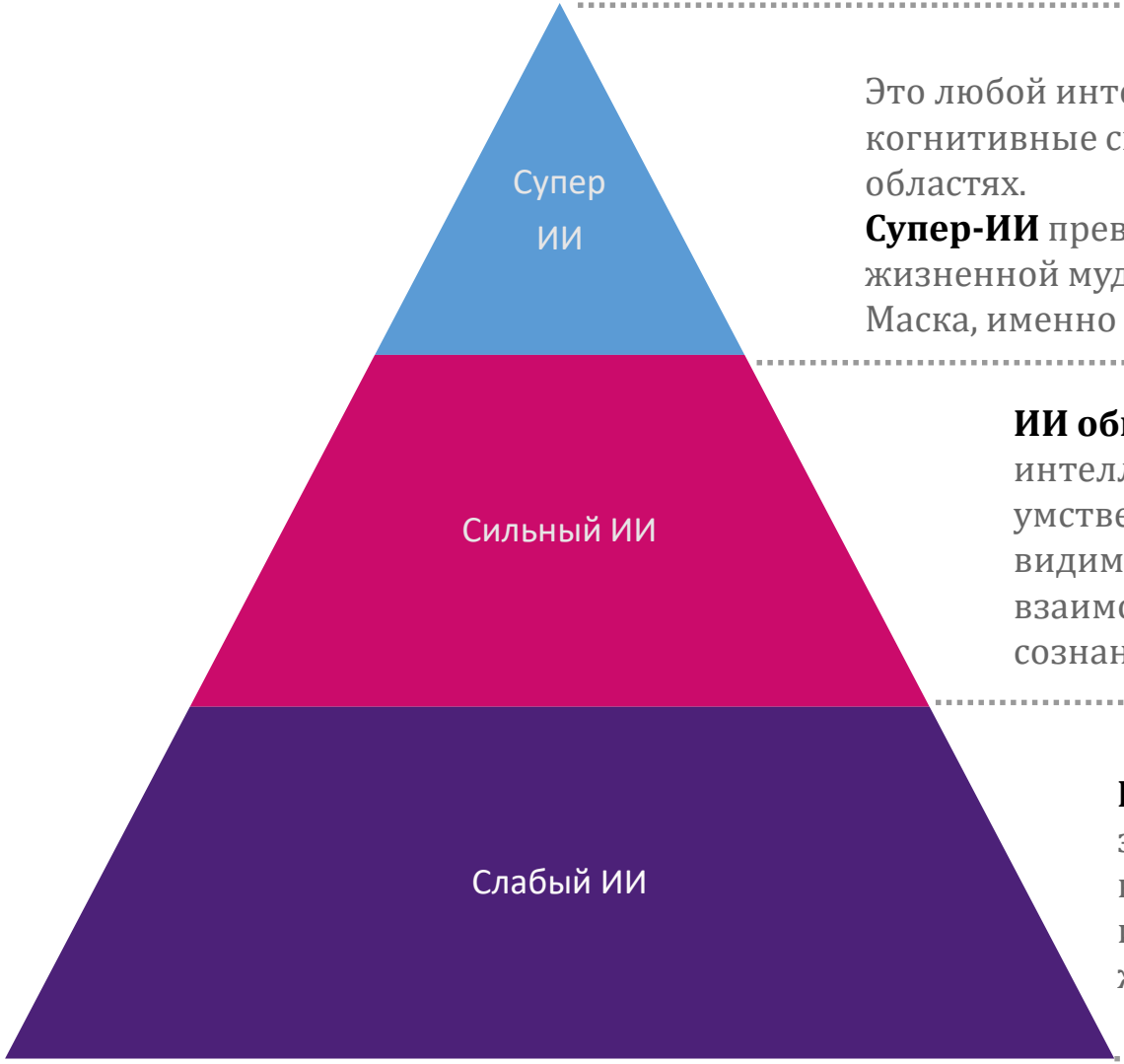
Какой бывает ИИ?



ИИ общего назначения, или **сильный ИИ**, схож с человеческим интеллектом. Иными словами, он может успешно выполнять любые умственные задачи, которые под силу людям. Именно такие системы мы видим в научно-фантастических фильмах, посвященных взаимодействию человека с машинами, обладающими чувствами и сознанием.

ИИ узкого назначения, также известный как **слабый**, — это ИИ в сегодняшнем понимании. Он запрограммирован на выполнение одной задачи — будь то мониторинг погоды, игра в шахматы или анализ данных для написания журналистских репортажей.

Какой бывает ИИ?



Это любой интеллект, который значительно превосходит когнитивные способности человека практически во всех областях.
Супер-ИИ превзойдет людей во всех аспектах — от творчества до жизненной мудрости и решения проблем. По мнению Илона Маска, именно он приведет к вымиранию людей как вида.

ИИ общего назначения, или **сильный ИИ**, схож с человеческим интеллектом. Иными словами, он может успешно выполнять любые умственные задачи, которые под силу людям. Именно такие системы мы видим в научно-фантастических фильмах, посвященных взаимодействию человека с машинами, обладающими чувствами и сознанием.

ИИ узкого назначения, также известный как **слабый**, — это ИИ в сегодняшнем понимании. Он запрограммирован на выполнение одной задачи — будь то мониторинг погоды, игра в шахматы или анализ данных для написания журналистских репортажей.

Что самое главное?

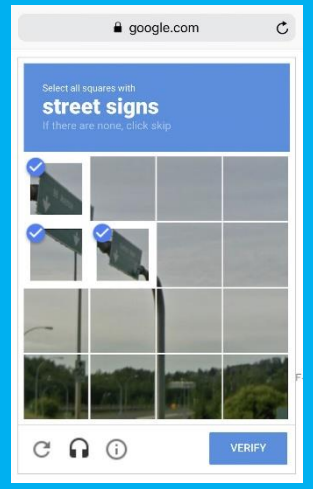
Данные

Данные собирают как могут.

Вручную — получается дольше, меньше, зато без ошибок.

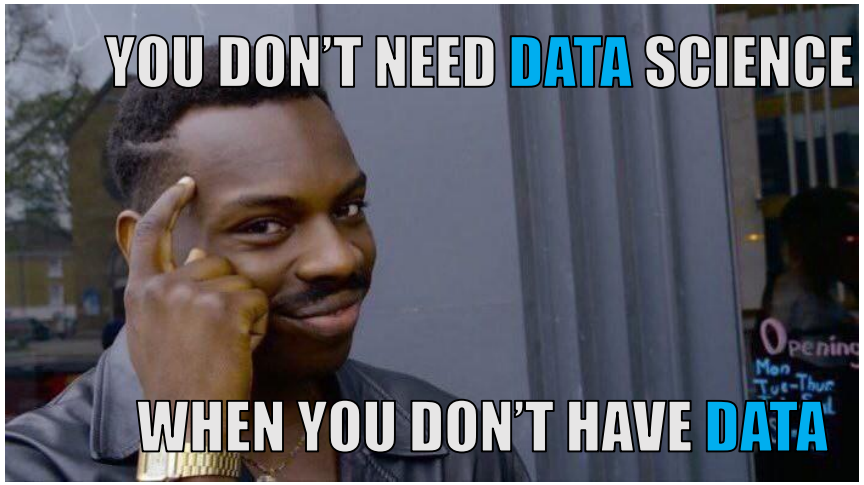
Автоматически — просто сливают машине всё, что нашлось, и верят в лучшее. Самые хитрые, типа гугла, используют своих же пользователей для бесплатной разметки.

Крупные компании, бывает, раскрывают свои алгоритмы, но **датасеты — крайне редко**.



Алгоритм

Одну задачу можно решить разными методами **всегда**. От выбора метода зависит точность, скорость работы и размер готовой модели. Но есть один нюанс: если данные плохие, даже самый лучший алгоритм не поможет. Не стоит бросать все 100% усилий на точность алгоритма, лучше собрать побольше данных и детально проанализировать.

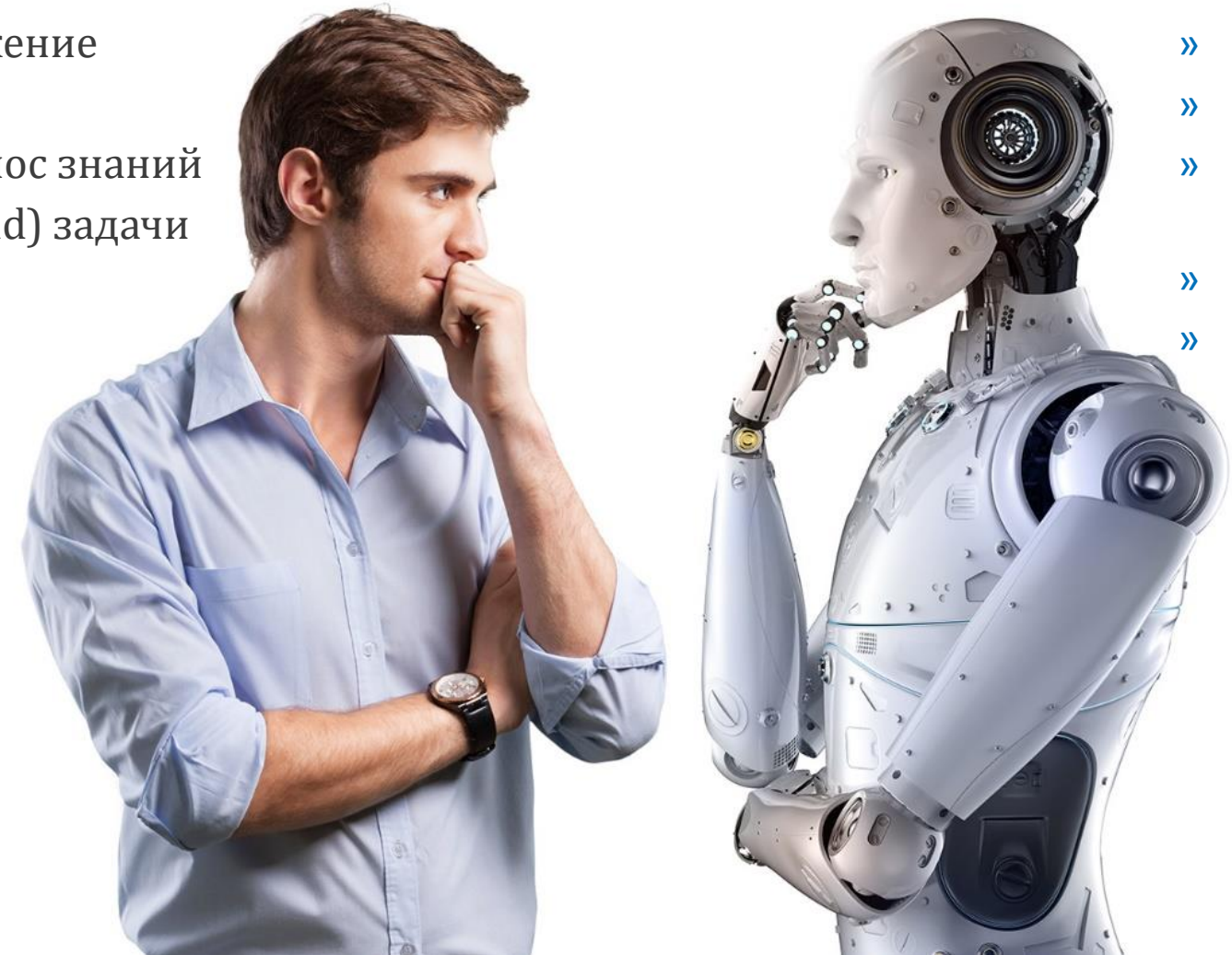


Заменим человека?!?

Заменим человека?!?

Заменим человека?!?

- Интуиция, воображение
- Рассуждения
- Обобщение и перенос знаний
- Открытые (open-end) задачи
- Целеполагание и планирование



- Скорость вычислений
- Объем памяти
- Выявление скрытых закономерностей
- Оперативный процессинг
- Первичная обработка сенсорных данных

Какое программное обеспечение?

Для полноценной работы нам понадобится установленное ПО

Anaconda

Или самостоятельно установите окружение Python, jupyter notebook, библиотеки scipy, numpy, scikit-learn, pandas, matplotlib, plotly, seaborn, …

Тест на знание математической статистики

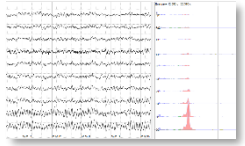
30 минут

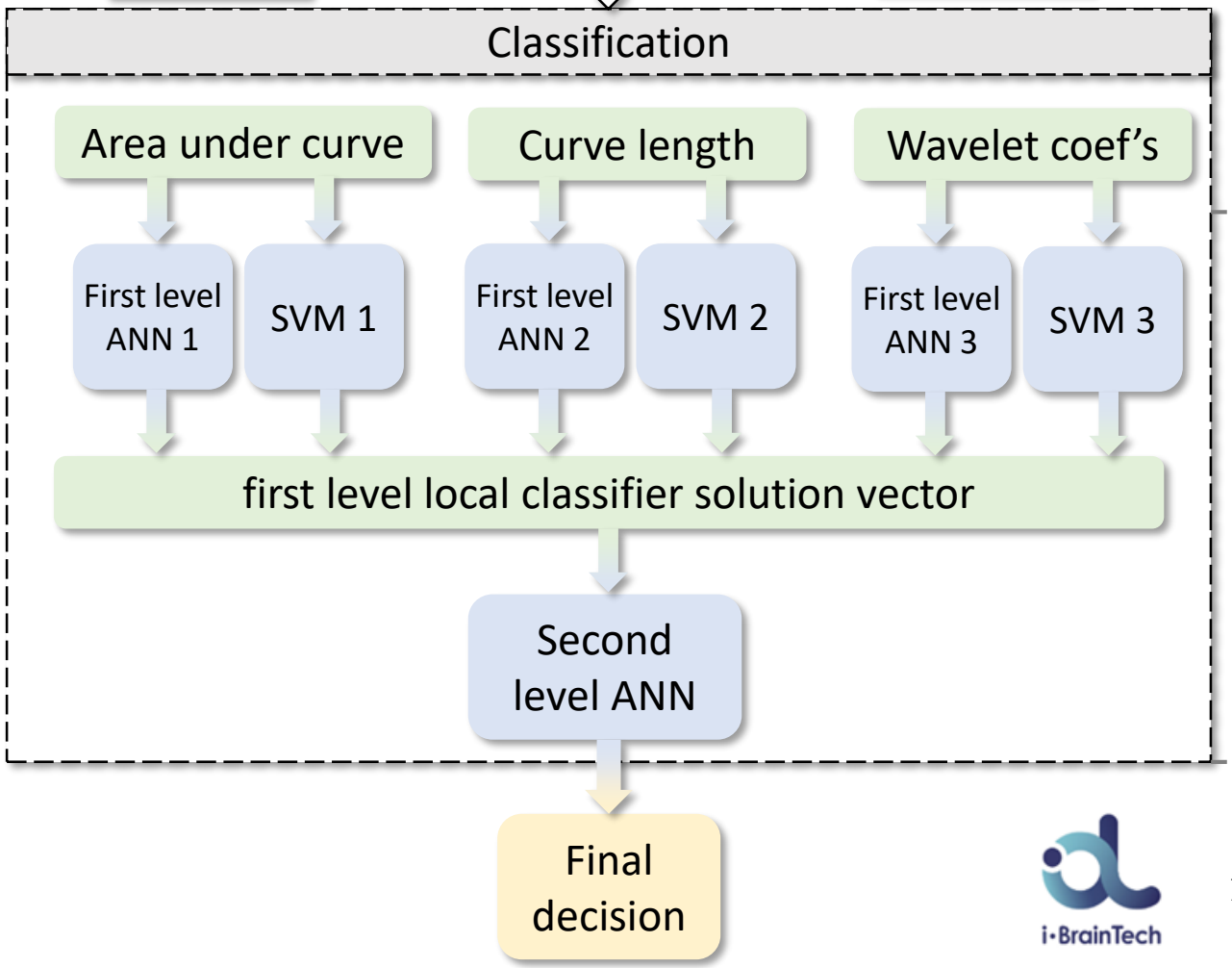
Что для вас искусственный интеллект?

7 1 1 38 0/7

Для решения каких задач хотелось бы попробовать применить знания в области машинного обучения?

Real-time brain-computer interface based on neurological committee of eeg signal classifiers.





Collecting and transforming data

- Providing natural experiments to collect the data
- Preprocessing (finding anomalies, filling empty values)
- Wavelet transformation
- Create train and test samples

Training SVM and ANN

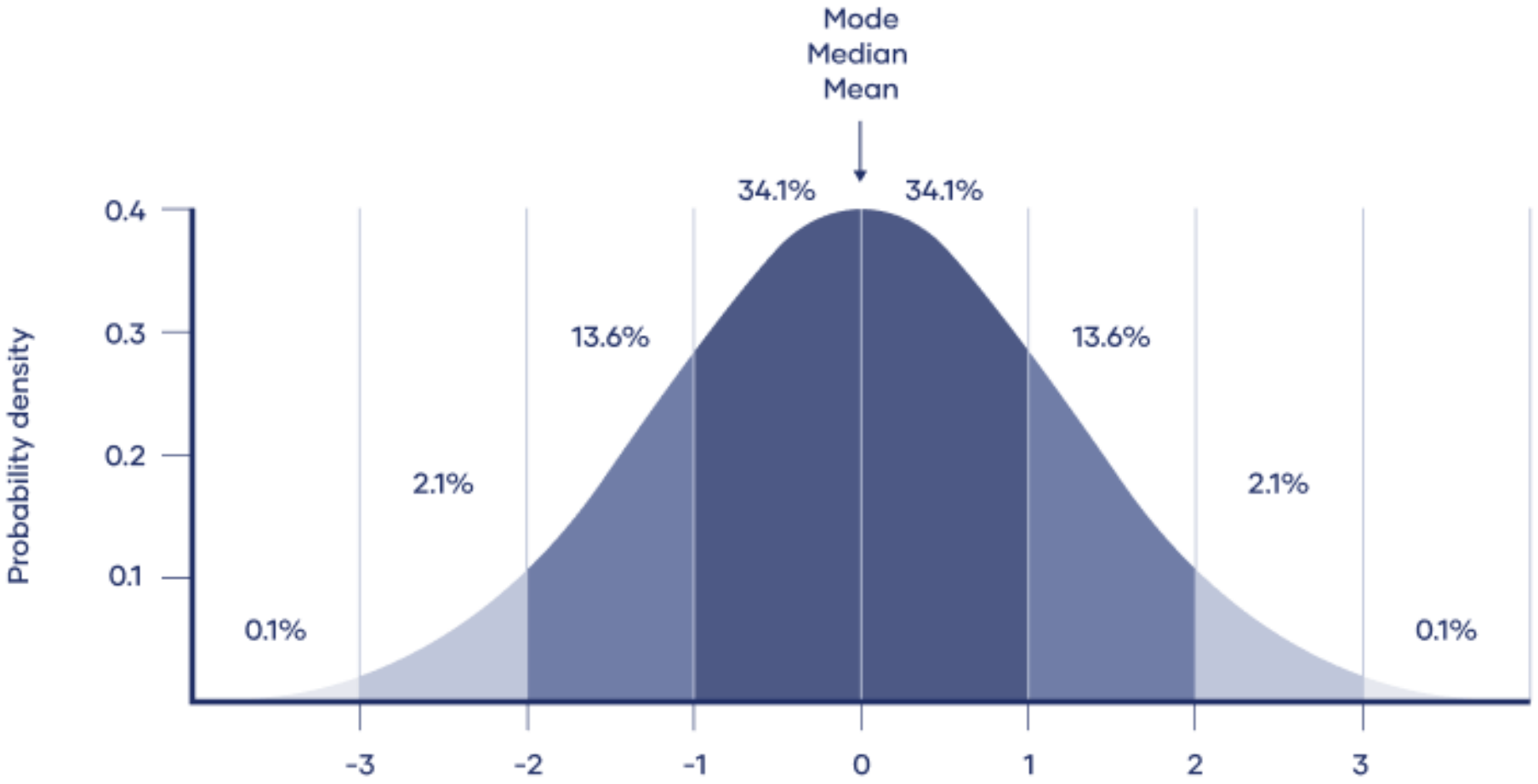
- Solve one vs rest task
- Solve multiclass task
- Compare results from different examinee

MATLAB®

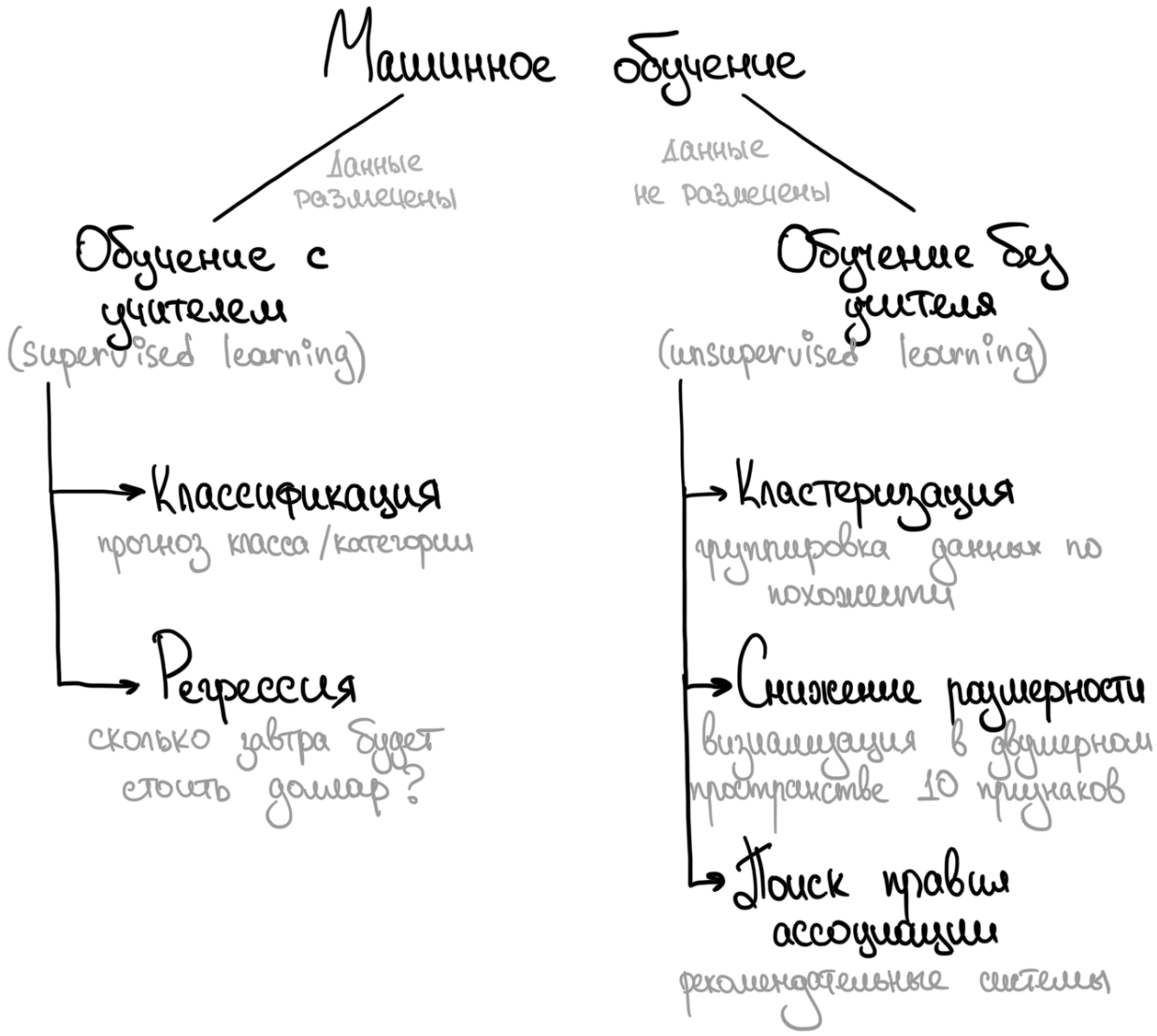
i·BrainTech

i-brain https://i-brain.tech/ru/

Какое самое распространенное распределение вероятности в мире?



Какие задачи можем решать?

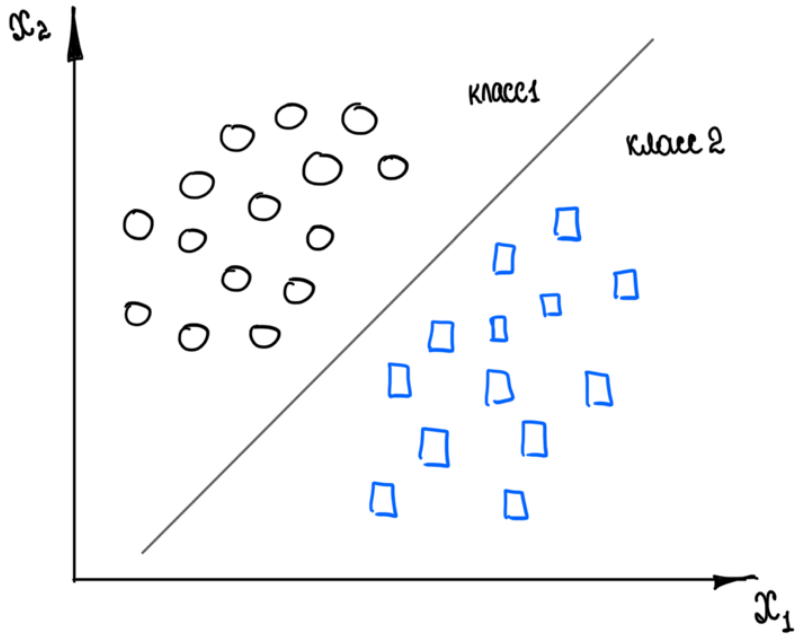


Обучение с учителем. Классификация

Постановка задачи

Пусть X— множество описаний объектов, Y - конечное множество номеров (имен, меток) классов. Существует некоторое отображение $f: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X_m = (x_1, y_1), \ldots, (x_m, y_m)$. Необходимо построить алгоритм $f^*: X \to Y$, способный классифицировать произвольный объект $x \in X$.

В математической статистике задачи классификации называются также задачами дискриминантного анализа.



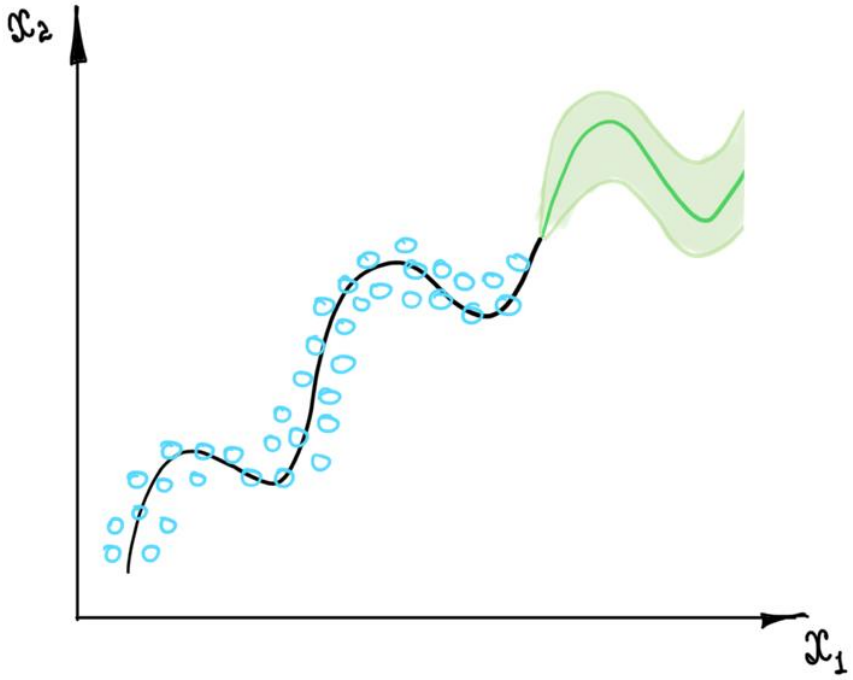
Примеры

- Кредитный скоринг
- Распознавание объектов на изображении
- Поиск мошеннических транзакий
- Спам-фильтры

Обучение с учителем. Регрессия

Постановка задачи

Пусть X— множество описаний объектов, Y – вещественная переменная. Существует некоторое отображение $f: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X_m = (x_1, y_1), \ldots, (x_m, y_m)$. Необходимо построить алгоритм $f^*: X \to Y$, способный спрогнозировать значение y, зная произвольный объект $x \in X$.



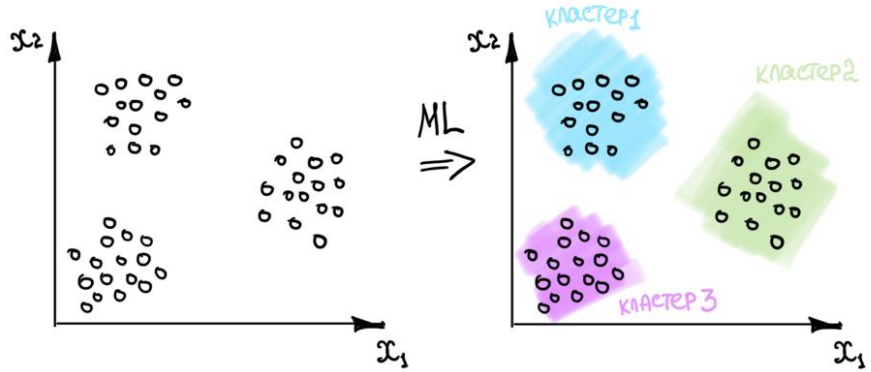
Примеры

- Прогноз стоимости ценных бумаг
- Анализ спроса, объема продаж
- Медицинские диагнозы
- Любые зависимости числа от времени

Обучение с учителем. Кластеризация

Постановка задачи

В алгоритмах обучения без учителя выходная ошибка модели на обучающем множестве не вычисляется. Вместо неё используется информация о текущем состоянии параметров модели и примеров обучающего множества.



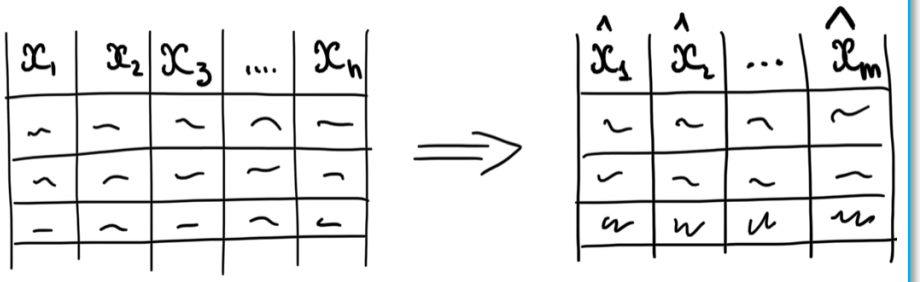
Примеры

- Сегментация рынка
- Сжатие изображений
- Поиск аномальных данных

Обучение с учителем. Снижение размерности

Постановка задачи

Очень часто датасеты (наборы данных) представляют из себя большие матрицы с точки зрения количества признаков, которыми описывается каждый объект. Но во-первых, далеко не все из этих признаков могут быть важны при решении задач машинного обучения, а во-вторых, визуализировать многомерные данные представляется нерешаемой задачей. В этом случае пробуем снизить размерность признакового пространства $X: R^n \to R^m$, где $m < n$.



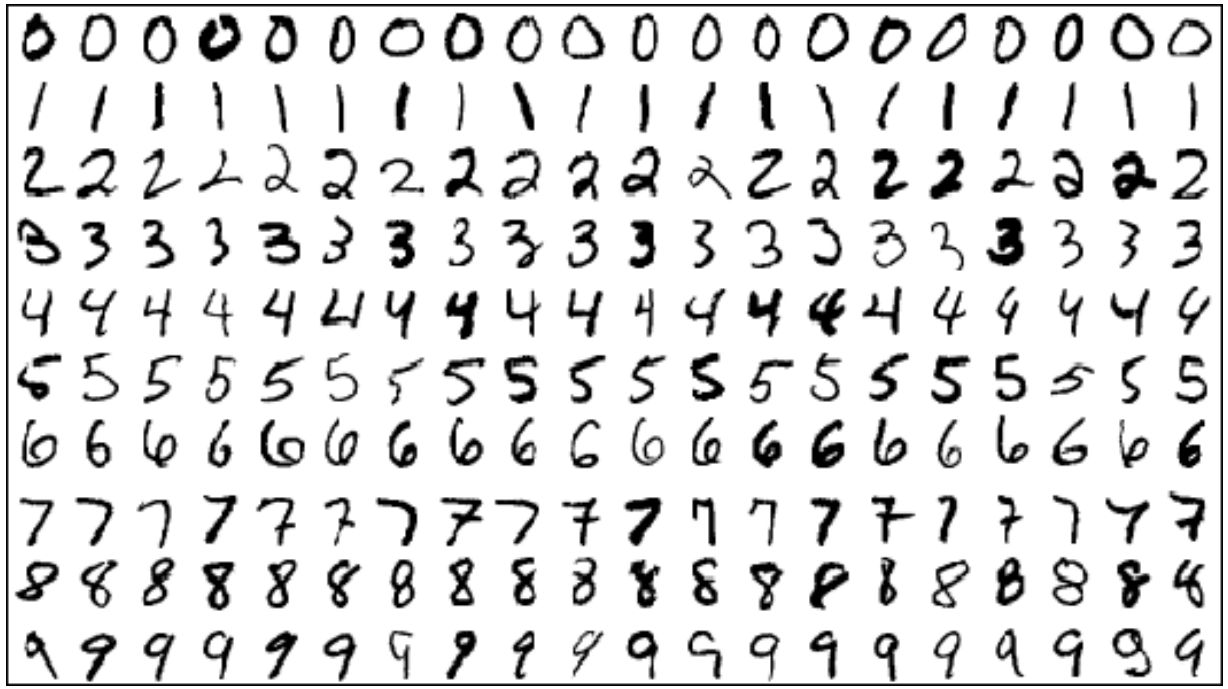
Примеры

- Визуализация многомерных данных
- Снижение размерности признакового пространства для моделей машинного обучения

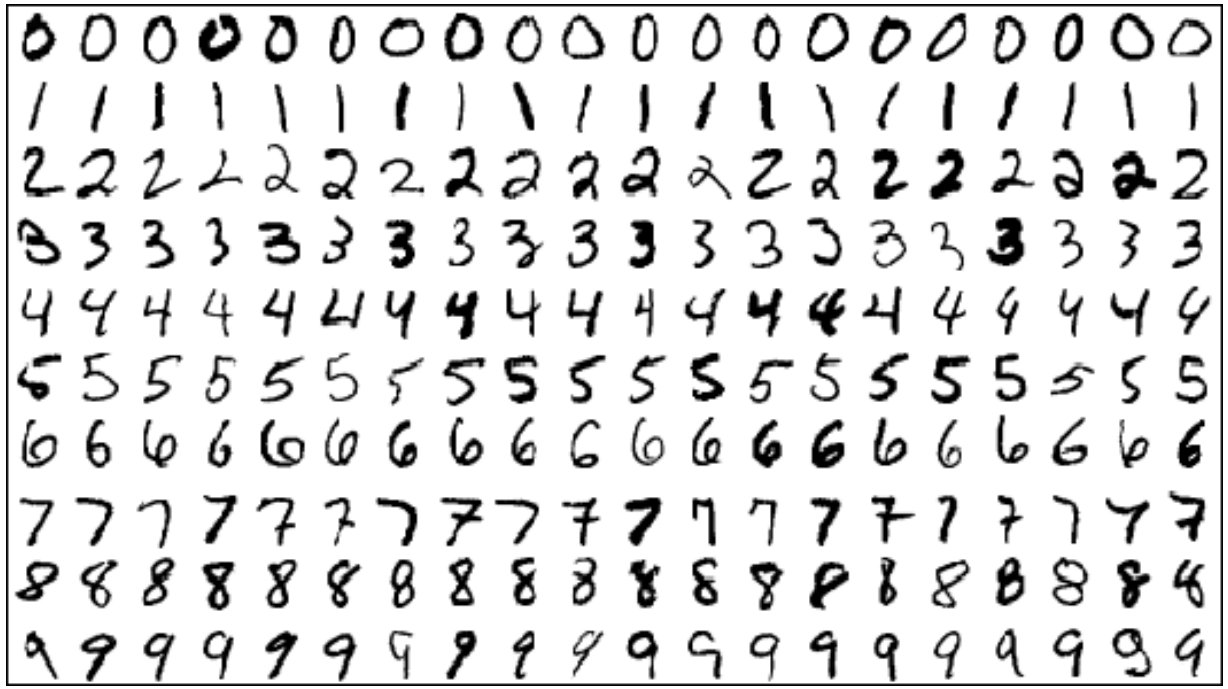
С какими данными происходит работа? 1/2

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S

С какими данными происходит работа? 2/2



С какими данными происходит работа? 2/2



Какие знания нужны?

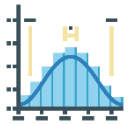
Какие знания нужны для курса?

Математика и статистика, технологии машинного и глубинного обучения

Голая статистика Чарльз Уиллан

Курс по статистике 1

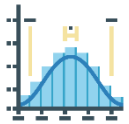
Курс по статистике 2



Какие знания нужны для курса?

Математика и статистика, технологии машинного и глубинного обучения

Голая статистика Чарльз Уиллан
[Курс по статистике 1]
[Курс по статистике 2]



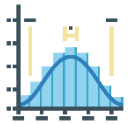
Навыки программирования

[Погружение в Python]
[Математика и Python]

Какие знания нужны для курса?

Математика и статистика, технологии машинного и глубинного обучения

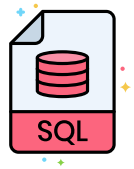
Голая статистика Чарльз Уиллан
Курс по статистике 1
Курс по статистике 2



Навыки программирования

Погружение в Python
Математика и Python

Работа с БД и знание SQL

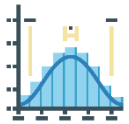


Курс по SQL

Какие знания нужны для курса?

Математика и статистика, технологии машинного и глубинного обучения

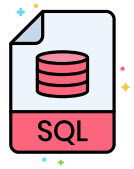
Голая статистика Чарльз Уиллан
[Курс по статистике 1]
[Курс по статистике 2]



Навыки программирования

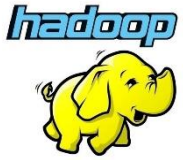
[Погружение в Python]
[Математика и Python]

Работа с БД и знание SQL



[Курс по SQL]

Навыки работы с большими данными: Hadoop, Spark, Hive, Kafka.

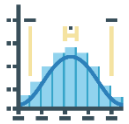


[Система обработки больших данных]

Какие знания нужны для курса?

Математика и статистика, технологии машинного и глубинного обучения

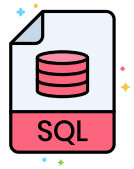
Голая статистика Чарльз Уиллан
[Курс по статистике 1]
[Курс по статистике 2]



Навыки программирования

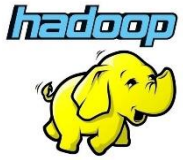
[Погружение в Python]
[Математика и Python]

Работа с БД и знание SQL



[Курс по SQL]

Навыки работы с большими данными: Hadoop, Spark, Hive, Kafka.



[Система обработки больших данных]

Навыки визуализации и презентации результатов работы: PowerPoint, Shiny/Dash, Power BI, Tableau, Qlik

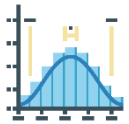
[Tableau]

[Beautiful Visualization]

Какие знания нужны для курса?

Математика и статистика, технологии машинного и глубинного обучения

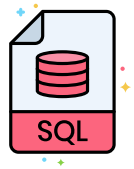
Голая статистика Чарльз Уиллан
[Курс по статистике 1]
[Курс по статистике 2]



Навыки программирования

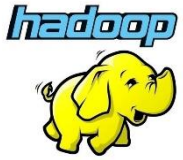
[Погружение в Python]
[Математика и Python]

Работа с БД и знание SQL



[Курс по SQL]

Навыки работы с большими данными: Hadoop, Spark, Hive, Kafka.



[Система обработки больших данных]

Навыки визуализации и презентации результатов работы: PowerPoint, Shiny/Dash, Power BI, Tableau, Qlik

[Tableau]

[Beautiful Visualization]

Отлаживать код и готовить к выкатке в промышленное пользование

[Kubernetes]

Ответы прошлого года

Go to **www.menti.com** and use the code **3147 9916**

Что для вас ИИ, машинное обучение?

Mentimeter

31

Для решения каких задач хотелось бы попробовать применить знания в области машинного обучения?