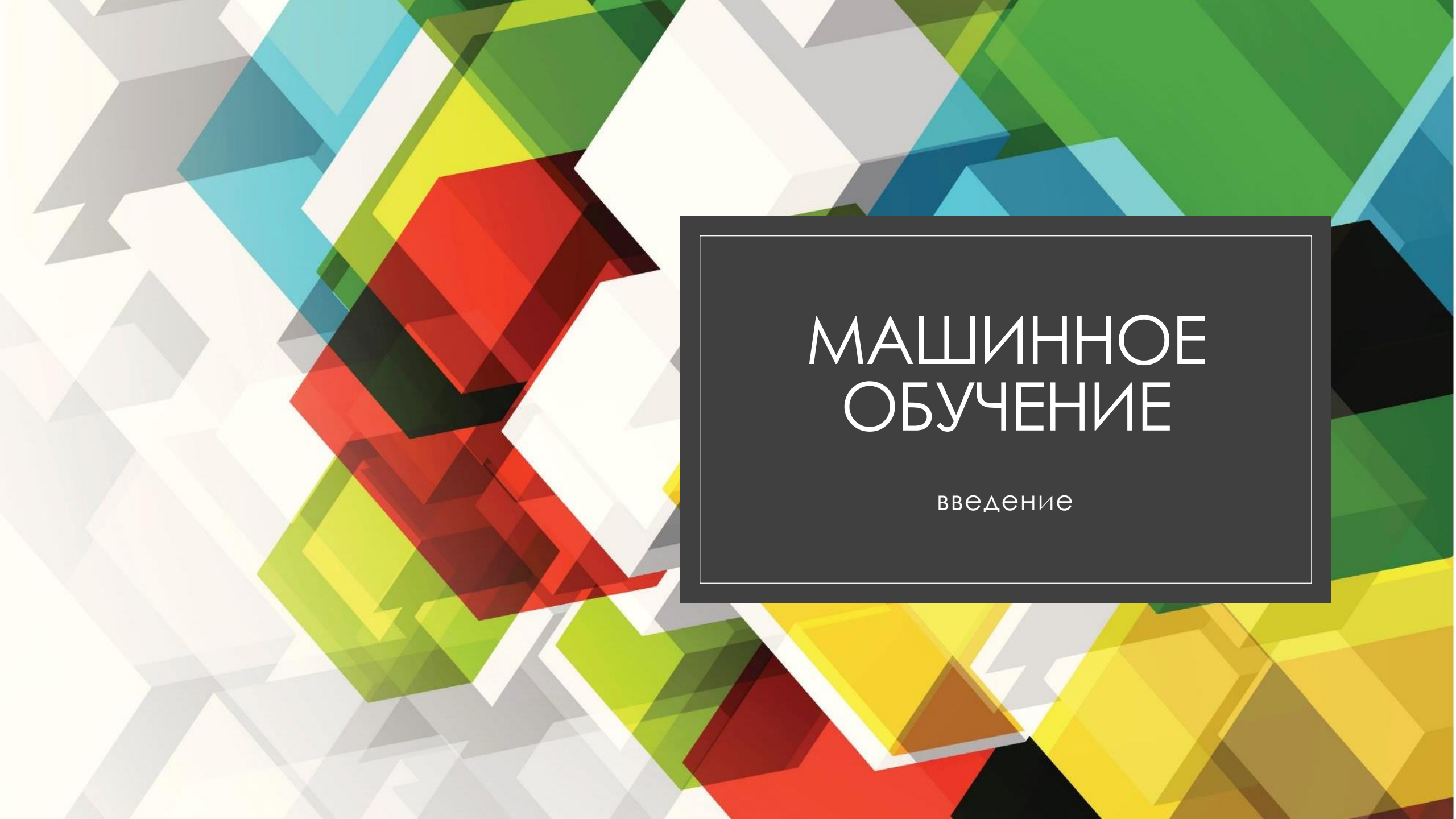

МАШИННОЕ ОБУЧЕНИЕ
введение

Понятие машинного обучения

- Машины **не учатся**
- Классическое машинное обучение = **поиск математических формул**, которые при применении к набору входных данных дают желаемые результаты
- Артур Сэмюэл (Arthur Samuel), американский пионер в области компьютерных игр и искусственного интеллекта, придумал термин «Машинное обучение» в 1959 году, когда работал в IBM.

Машинное обучение

Использование математических моделей данных, которые помогают компьютеру обучаться без непосредственных инструкций.

С помощью алгоритмов выявляются закономерности в данных. На основе этих закономерностей создается модель данных.

Способ обучать компьютеры без программирования и явных инструкций, используя только шаблоны и логические выводы.

Преимущества машинного обучения

Получение аналитических сведений

- помогает определять закономерности или структуры на основе как структурированных, так и неструктурированных данных

Расширение возможностей пользователей

- Адаптивные интерфейсы, целевое содержимое, чат-боты и виртуальные помощники с поддержкой голоса

Прогнозирование поведения клиентов

- можно выполнять интеллектуальный анализ данных, связанных с клиентами

Повышение уровня целостности данных

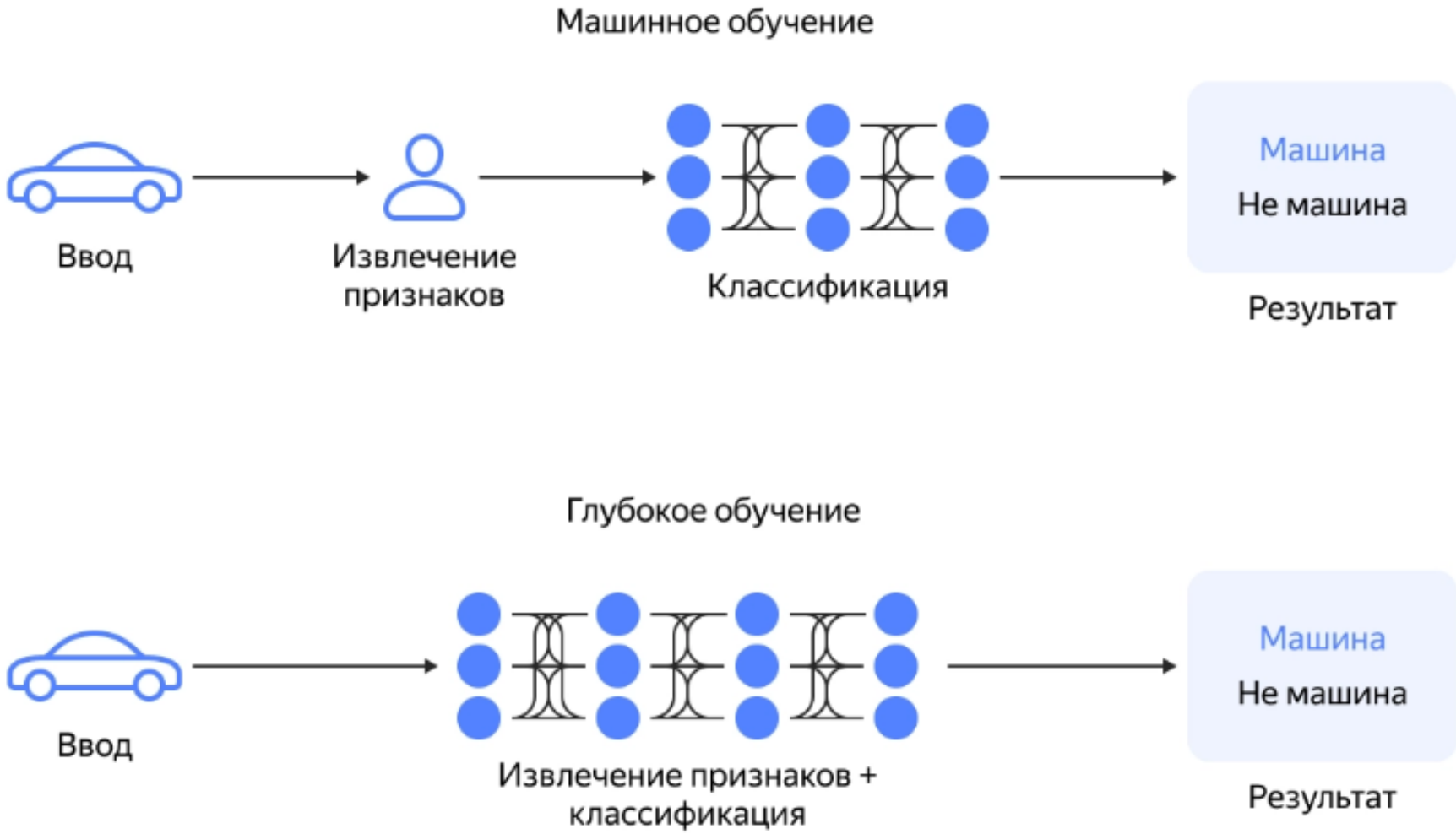
- идеальный вариант для интеллектуального анализа данных, повышающий его точность и расширяющий его возможности с течением времени

Уменьшение риска

- позволяет отслеживать и определять новые приемы злоумышленников

Алгоритм применения машинного обучения

Шаг 1	Сбор и подготовка данных
Шаг 2	Обучение модели (настройка модели)
Шаг 3	Проверка модели
Шаг 4	Интерпретация результатов


Машинное обучение
Ввод
Извлечение признаков
Классификация
Машина
Не машина
Результат
Глубокое обучение
Ввод
Извлечение признаков + классификация
Машина
Не машина
Результат

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

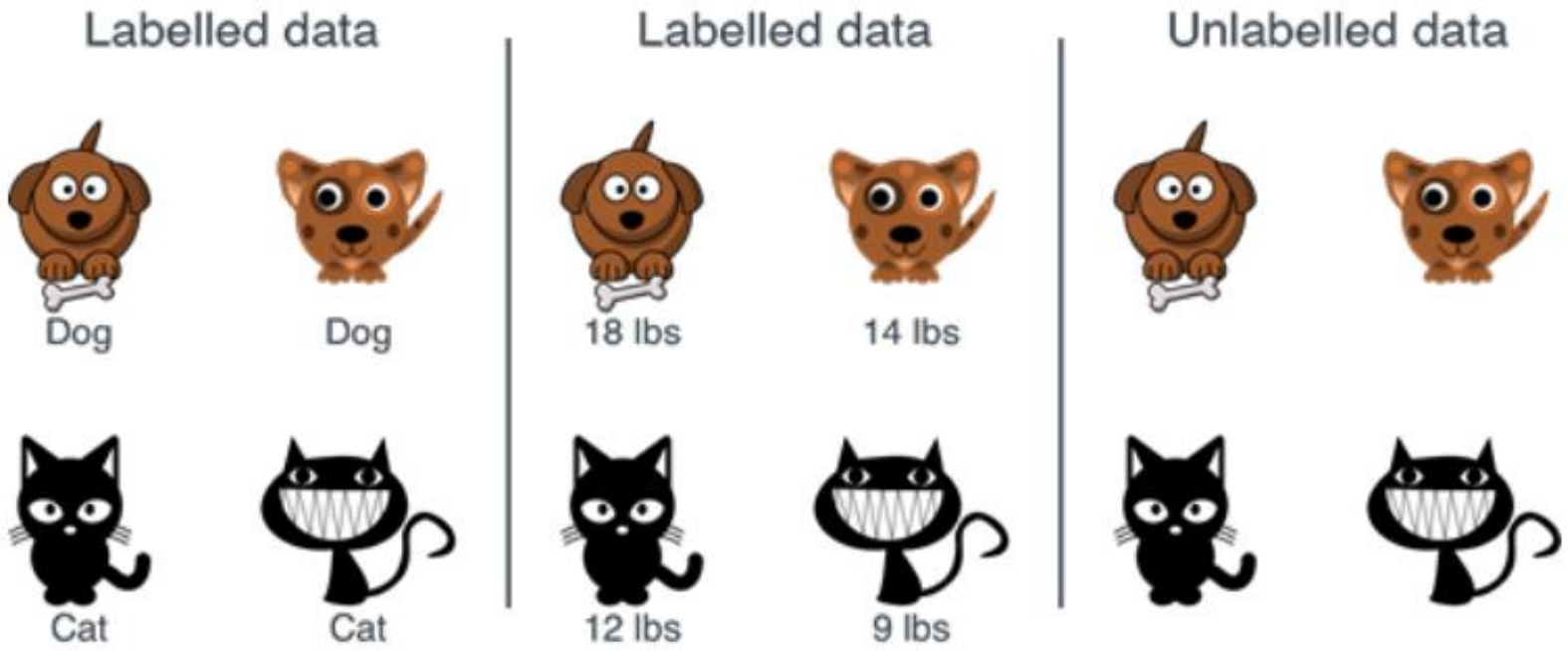
«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Источник: https://vas3k.ru/blog/machine_learning

Размеченные (labelled) и неразмеченные (unlabelled) данные



Размеченные (labelled) и неразмеченные (unlabelled) данные

Размеченные

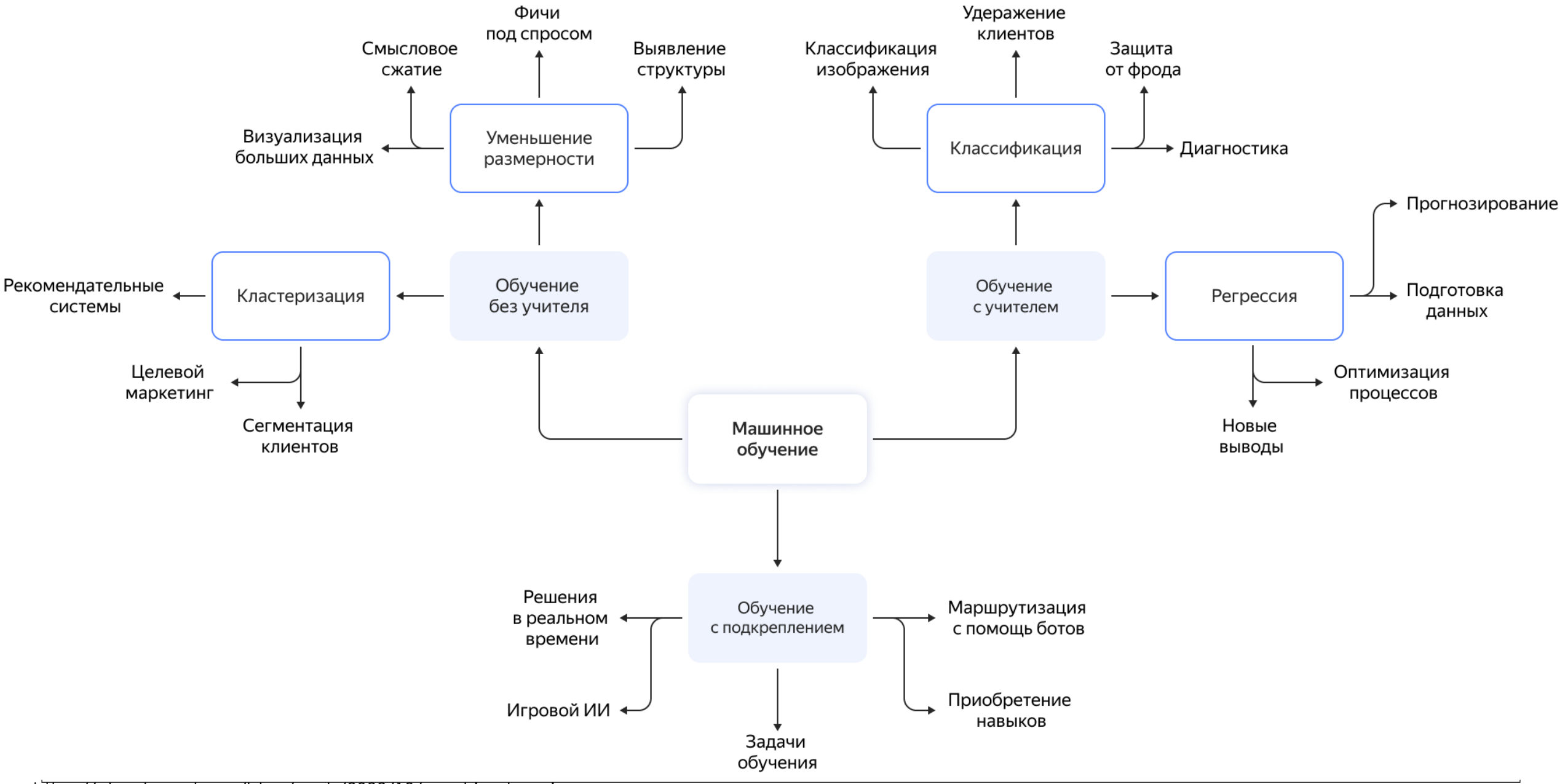
ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

labels

Неразмеченные

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

unlabeled



https://cloud.yandex.ru/blog/posts/2022/10/machine-learning

Типы задач машинного обучения

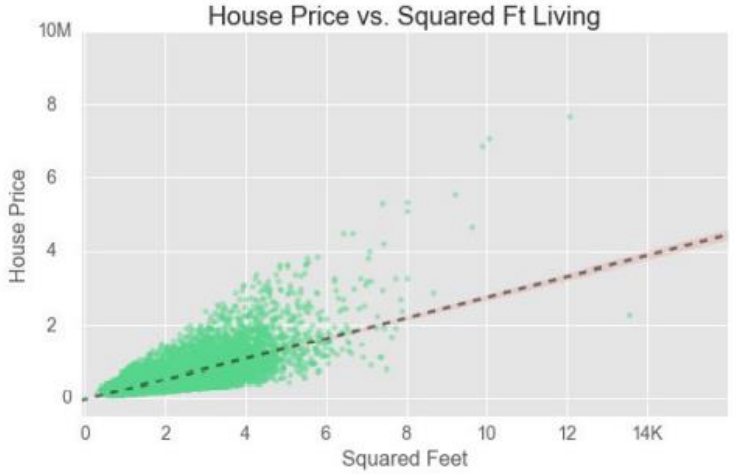
Классификация — это задача автоматического определения метки для неразмеченного образца.

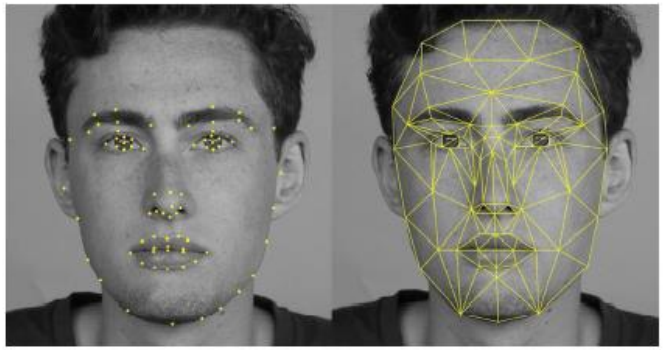
Регрессия — это задача прогнозирования метки с действительным значением (часто называют также целевым значением) для образца без метки.

Кластеризация - задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Примеры задач регрессии

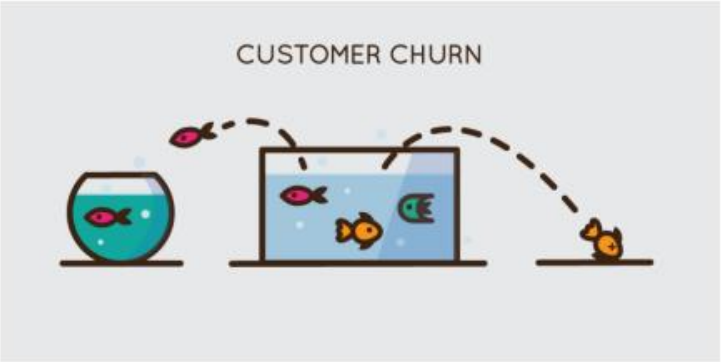
- **Предсказание стоимости жилья** для риэлторской компании
- **Предсказание времени доставки**
- **Предсказание спроса на такси** в конкретном районе в конкретный час завтрашнего дня.
- **Предсказание ключевых точек лица — т.н. дескрипторов**

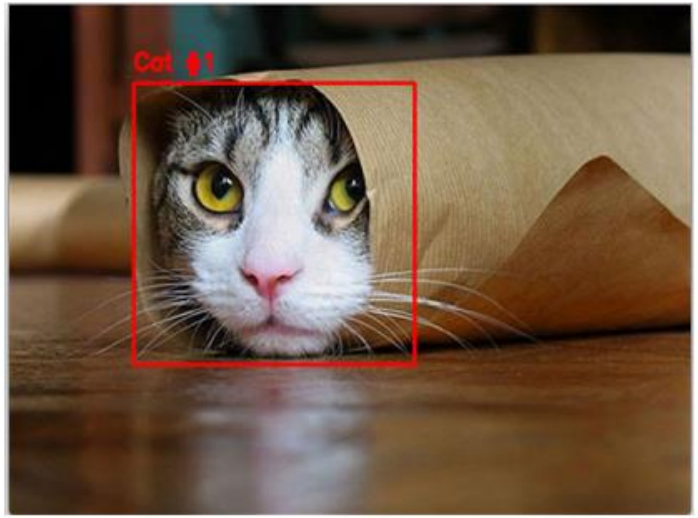




Примеры задач классификации

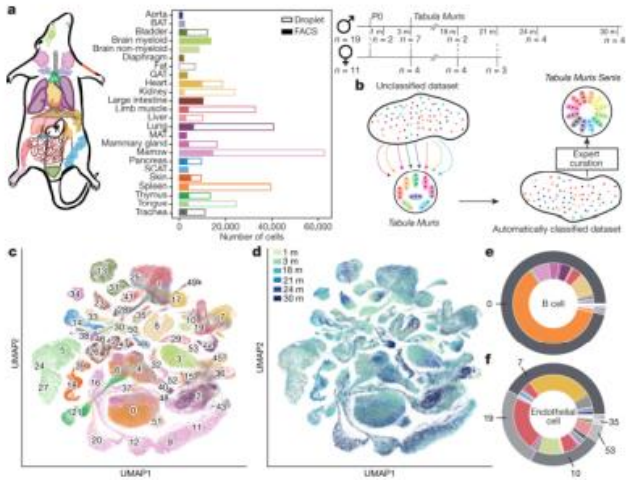
- **Предсказание оттока клиентов / сотрудников** на основе их поведения.
- **Классификация клеток ткани** на здоровые и опухолевые
- **Детекция объектов на фото**.

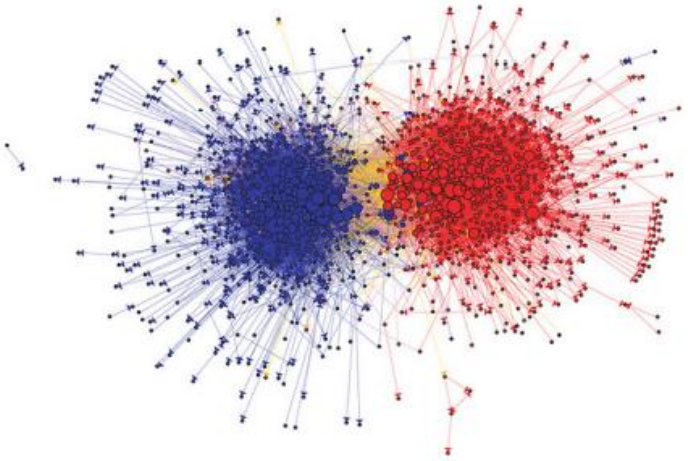




Примеры задач кластеризации

- **Сегментация аудитории** для таргетирования рекламы
- **Идентификация типов клеток** в образце данных секвенирования
- **Поиск сообществ в социальном графе** (из соцсети или из инсайдерской информации о структуре организации)
- **Задача разделения смеси распределений**





Недообучение

- это **неспособность модели** более или менее точно **предсказывать метки** данных, на которых она **обучалась**.

Причины:

- модель слишком проста для данных;
- спроектированные признаки недостаточно информативны.

Переобучение

- **Переобученная модель** слишком **хорошо** предсказывает **обучающие** данные, но **плохо** — данные из обеих **контрольных** выборок или хотя бы из одной из них.

Причины

- Модель слишком сложна для данных
- Слишком много признаков, но мало обучающих данных.

Борьба => Регуляризация – позволяет алгоритмам обучения строить менее сложные модели

Библиотеки для машинного обучения

Pandas

- Исследование и подготовка данных

NumPy

- Многомерные массивы

Scikit-learn

- Создание моделей машинного обучения

SciPy

- Предназначена для проведения научных исследований

Seaborn

- Визуализация

Библиотека Scikit-learn

Разработанна Дэвидом Курнапо в рамках программы Google Summer of Code в 2010 году.

- предварительная обработка данных
- уменьшение размерности
- выбор модели
- регрессия
- классификация
- кластерный анализ
- датасеты