

Введение в машинное обучение



Как перевести часы в минуты?



Как перевести часы в минуты?

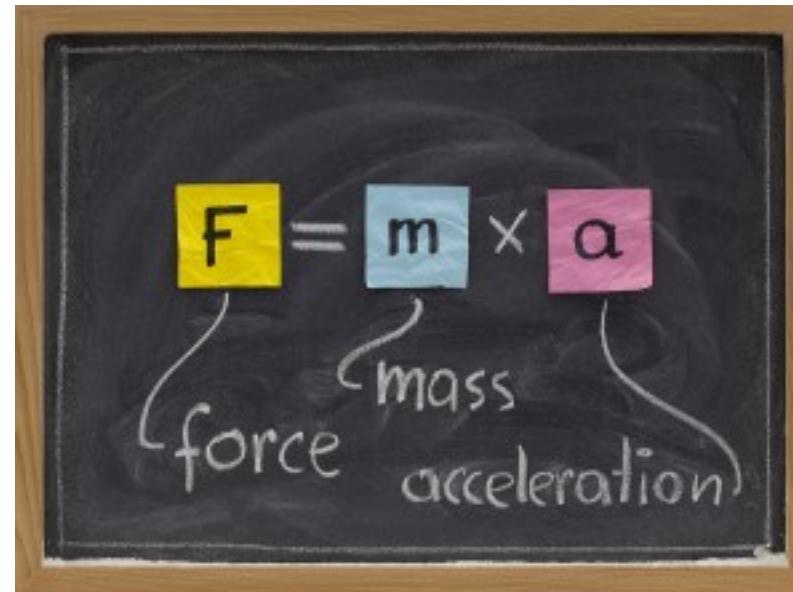
- x — часы
- $f(x) = 60x$ — преобразование в минуты, функция

Какая сила приложена к телу?

- Известны масса тела m и его ускорение a
- Чему равна сила F ?

Какая сила приложена к телу?

- Известны масса тела m и его ускорение a
- Чему равна сила F ?
- Второй закон Ньютона: $F = ma$



Как предсказать погоду?



Уравнения Навье-Стокса

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} = - \frac{\partial P}{\partial x} + Re \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right),$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} = - \frac{\partial P}{\partial y} + Re \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right),$$

$$\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} = - \frac{\partial P}{\partial z} + Re \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right),$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0.$$

Уравнения Навье-Стокса

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} = - \frac{\sigma_x}{\rho} + Re \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right),$$

$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} = - \frac{\sigma_y}{\rho} + Re \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right)$

$\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} = - \frac{\sigma_z}{\rho} + Re \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right)$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0.$$

Анализ тональности текста

- Какой эмоциональный окрас имеет текст?
- Варианты: позитивный, нейтральный, негативный
- Применение: автоматический анализ отзывов от пользователей

Анализ тональности текста

«Большое спасибо! Сюда по всему, это как раз то, чего не хватает всем зарубежным курсам по Machine Learning и Knowledge Discovery. Это теория, математика, объяснение того, как оно устроено “в кишках”.»

Какой окрас?

Анализ тональности текста

«Я вижу очень большой минус, что курс будет на готовой библиотеке sci-kit. Курс от Andrew лучше тем, что ученик сам пишет алгоритм и видит изнутри, как он работает.»

Какой окрас?

Анализ тональности текста

- x — текст на русском языке
 - $f(x)$ — его окрас (принимает значения -1, 0, 1)
 - Можно ли выписать формулу для $f(x)$?
-
- На входе — вовсе не числа
 - Точная зависимость может не существовать

Больше сложных задач!

- Какой будет спрос на товар в следующем месяце?
- Сколько денег заработает магазин за год?
- Вернет ли клиент кредит?
- Заболеет ли пациент раком?
- Сдаст ли студент следующую сессию?
- На фотографии гуманитарий или технарь?
- Кто выиграет битву в онлайн-игре?

Больше сложных задач!

- Везде — очень сложные неявные зависимости
- Нельзя выразить их формулой
- Но есть некоторое число примеров
 - Тексты с известным окрасом
- Будем приближать зависимости, используя примеры

Анализ данных и машинное обучение

—proto, как восстановить сложные зависимости
по конечному числу примеров

Основные термины

Пример задачи

- Сеть ресторанов
- Хотим открыть еще один
- Несколько вариантов размещения
- Какой из вариантов принесет максимальную прибыль?

* см. kaggle.com, TFI Restaurant Revenue Prediction

Обозначения

- x — объект, sample — для чего хотим делать предсказания
 - Конкретное расположение ресторана
- \mathbb{X} — пространство всех возможных объектов
 - Все возможные расположения ресторанов
- y — ответ, целевая переменная, target — что предсказываем
 - Прибыль в течение первого года работы
- \mathbb{Y} — пространство ответов — все возможные значения ответа
 - Все вещественные числа

Обучающая выборка

- Мы ничего не понимаем в экономике
- Зато имеем много объектов с известными ответами
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- ℓ — размер выборки

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание



Вектор

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание



Признаки

- Про демографию:
 - Средний возраст жителей ближайших кварталов
 - Динамика количества жителей
- Про недвижимость:
 - Средняя стоимость квадратного метра жилья поблизости
 - Количество школ, банков, магазинов, заправок
 - Расстояние до ближайшего конкурента
- Про дороги:
 - Среднее количество машин, проезжающих мимо за день

Алгоритм

- $a(x)$ — алгоритм, модель — функция, предсказывающая ответ для любого объекта
- Отображает \mathbb{X} в \mathbb{Y}
- Линейная модель: $a(x) = w_1x^1 + \dots + w_dx^d$

ФУНКЦИЯ ПОТЕРЬ

- Не все алгоритмы полезны
- $a(x) = 0$ — не принесет никакой выгоды
- Функция потерь — мера корректности ответа алгоритма
- Предсказали \$10000 прибыли, на самом деле \$5000 — хорошо или плохо?
- Квадратичное отклонение: $(a(x) - y)^2$

Функционал качества

- Функционал качества, метрика качества — мера качества работы алгоритма на выборке
- Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

- Чем меньше, тем лучше

Функционал качества

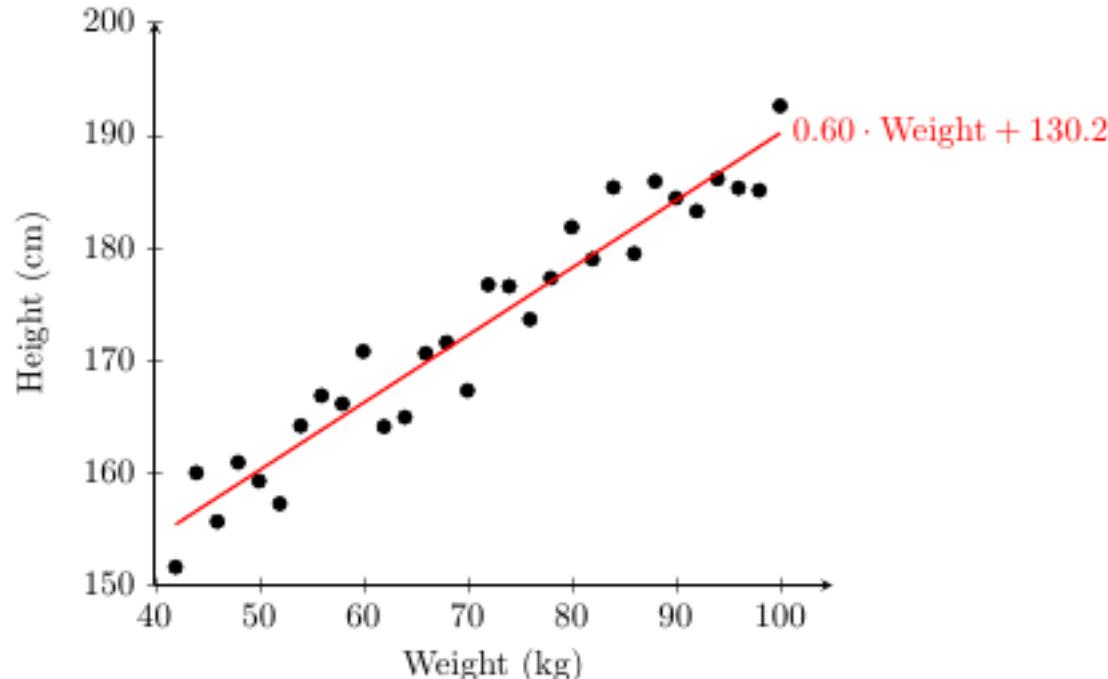
- Должен соответствовать бизнес-требованиям
- Одна из самых важных составляющих анализа данных

Обучение алгоритма

- Есть обучающая выборка и функционал качества
- Семейство алгоритмов \mathcal{A}
 - Из чего выбираем алгоритм
 - Пример: все линейные модели
 - $\mathcal{A} = \{w_1x^1 + \dots + w_dx^d \mid w_1, \dots, w_d \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения функционала качества

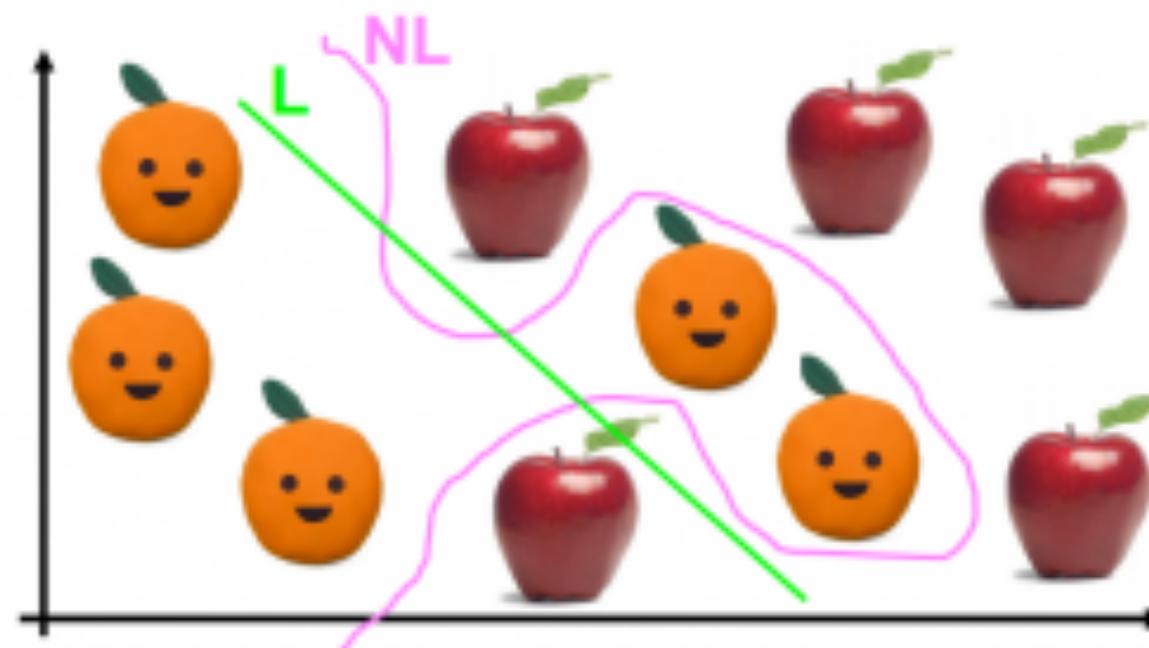
Регрессия

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



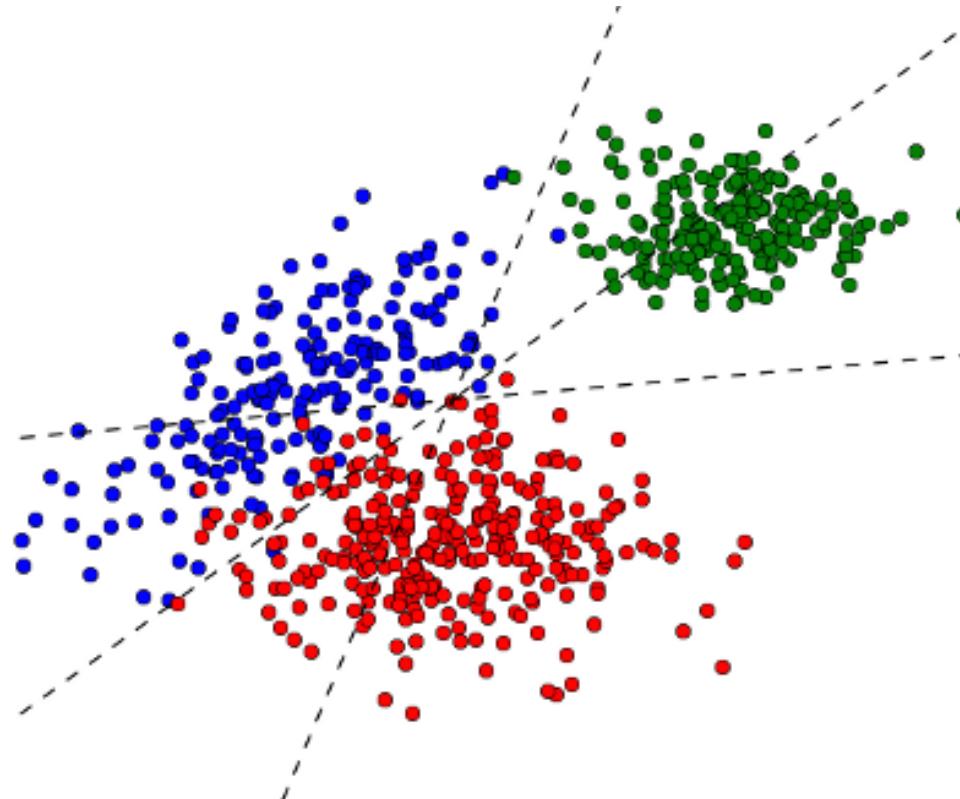
Классификация

- Конечное число ответов: $|\mathbb{Y}| < \infty$
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$

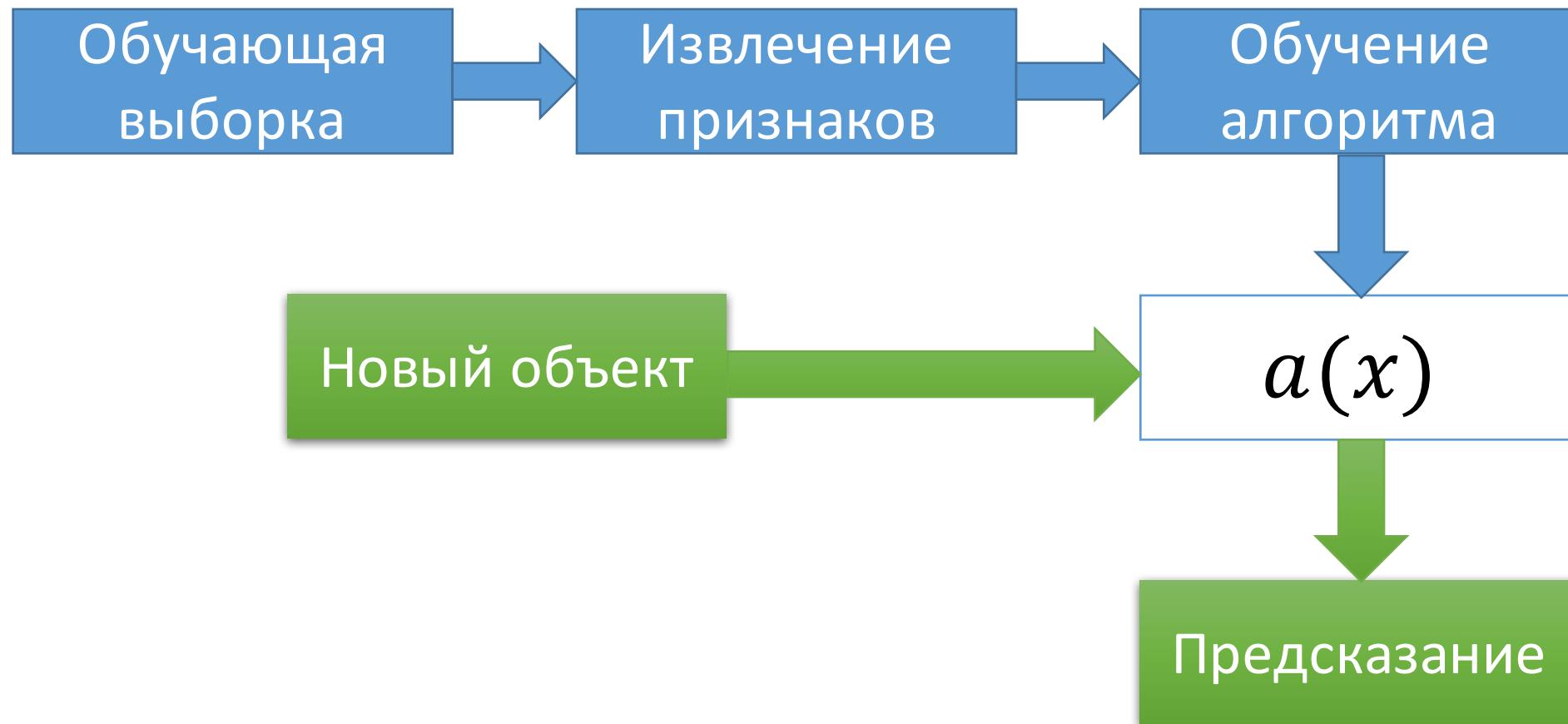


Классификация

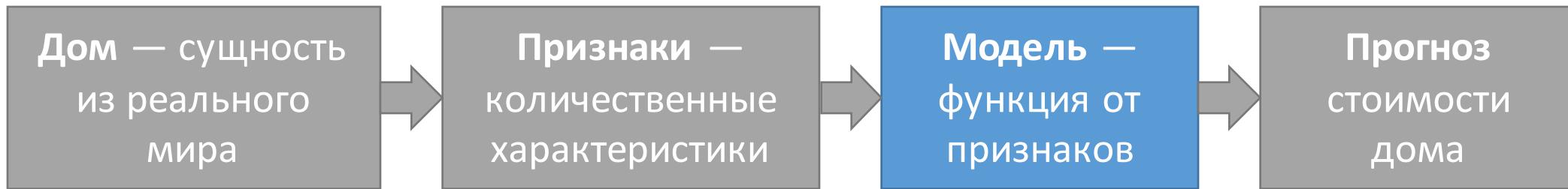
- Многоклассовая классификация: $\mathbb{Y} = \{1, 2, \dots, K\}$



Машинное обучение



Предсказание стоимости дома



Предсказание стоимости дома

Обучающая выборка:

Площадь	Цена
50	250
60	340
10	20
90	800

Возможные признаки:

- площадь
- площадь²
- площадь³
- $\sin(\text{площадь})$
- $\sqrt{\text{площадь}}$
- и так далее

Возможные модели:

- $w_1 * \text{площадь}$
- $w_1 * \text{площадь}^2$
- $w_1 * \text{площадь} + w_2 * \text{площадь}^2$
- и так далее

Вид модели — работа эксперта либо полный перебор.

Выбор весов w_1, w_2 — автоматический процесс (на основе данных)

Предсказание стоимости дома

Модель $a(x) = 5 * \text{площадь}$

Площадь	Прогноз	Цена	$(a - y)^2$
50	250	250	0
60	300	340	1600
10	50	20	900
90	450	800	122500

MSE: 31 250

RMSE: 176,78

Модель $a(x) = 0.1 * \text{площадь}^2$

Площадь	Прогноз	Цена	$(a - y)^2$
50	250	250	0
60	360	340	400
10	10	20	100
90	810	800	100

MSE: 150

RMSE: 12,25

Предсказание стоимости дома

Признаков может быть
больше:

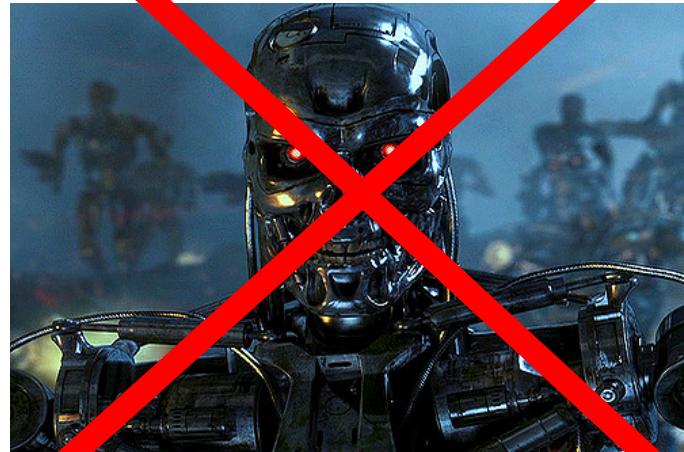
- Площадь
- Год постройки
- Наличие бассейна
- Число комнат
- Удалённость от центра
- Рейтинг полицейского участка
- И так далее

Возможные модели:

- Линейная: $w_1 * \text{площадь} + w_2 * \text{год} + w_3 * \text{бассейн} + w_4 * \text{комнаты} + w_5 * \text{удалённость} + w_6 * \text{полиция}$
- Решающие деревья
- Нейронные сети
- Метод k ближайших соседей
- И так далее

Зачем это нужно?

Искусственный интеллект



Сильный ИИ

через 20-100 лет

Яндекс

фильм где астронавту протыкают скафандр



X



Найти

ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ



Марсианин

The Martian, 2015 (16+)

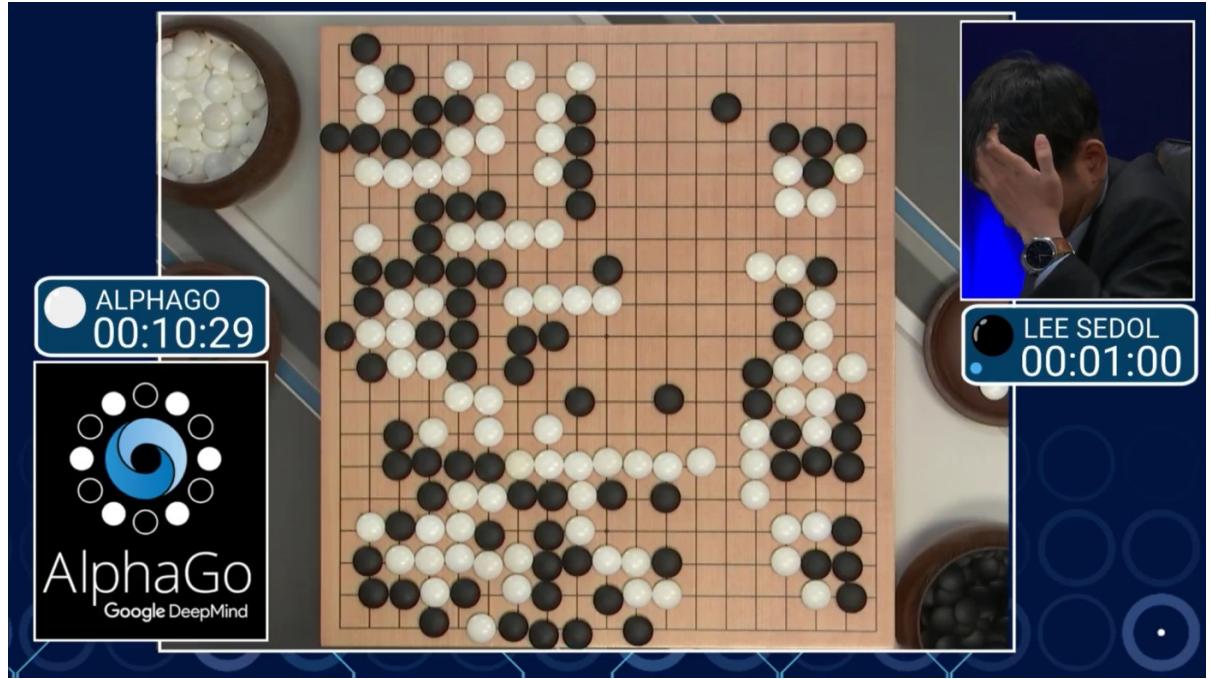
Марсианская миссия «Арес-3» в процессе работы была вынуждена экстренно покинуть планету из-за надвигающейся песчаной бури. Инженер и биолог Марк Уотни получил повреждение скафандра во время песчаной бури. Сотрудники миссии, посчитав его погибшим,...
[Читать дальше](#)

Специализированный ИИ

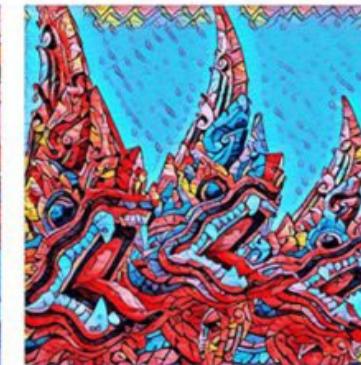
уже сейчас

AlphaGo

- Модель для игры в Го
- Оценивает успешность хода
- Обучалась путём игры с собой
- Победила чемпиона мира в 2016 году
- Долгое время игра в Го считалась невозможной задачей для компьютера



Перенос стиля

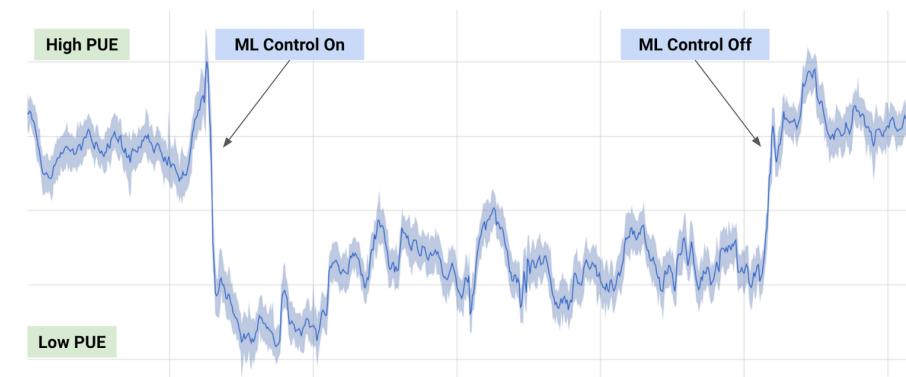


Машинное обучение в HR

- Поиск кандидатов и предсказание исхода собеседования
- Помощь при ротации
- Предсказание ухода сотрудника
- Анализ внутренних форумов, выделение жалоб

Автоматизация системы охлаждения

- Одна из главных компонент дата-центра — система охлаждения
- Результат работы системы сложным образом зависит от её параметров
- Необходимо быстро адаптироваться под изменение условий (нагрузка на серверы, погода)
- Все дата-центры разные — эвристические правила одного центра не работают в другом
- Машинное обучение позволило сократить затраты электричества на охлаждение на 40%



Рекомендательные системы

- Полки рекомендаций на Amazon генерируют 35% от всех покупок
- Рекомендации на основе машинного обучения и анализа больших объёмов данных

Frequently Bought Together

Price For All Three: \$86.01

Add all three to Cart Add all three to Wish List

Show availability and shipping details

This item: Machine Learning for Hackers by Drew Conway Paperback \$33.87

Machine Learning in Action by Peter Harrington Paperback \$25.75

Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback \$26.39

Customers Who Bought This Item Also Bought

Page 1 of 17

Item	Author	Type	Price
Programming Collective Intelligence: Building Smart Web 2.0 Applications	Toby Segaran	Paperback	\$26.39
Machine Learning in Action	Peter Harrington	Paperback	\$25.75
Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and More Social Networks	Matthew A. Russell	Paperback	\$26.36
Data Analysis with Open Source Tools	Philipp K. Janert	Paperback	\$24.05
R Cookbook (O'Reilly Cookbooks)	Paul Teator	Paperback	\$32.43
The Art of R Programming: Tour of Statistical Analysis, Visualization, and Data Munging with R	Norman Matloff	Paperback	\$25.06

Are any of these items inappropriate for this page? [Let us know](#)

Зачем это нужно?

- Это круто
 - Сложные задачи
 - Движение к искусственному интеллекту
- Это полезно
 - Извлечение прибыли из данных
 - Data-driven companies

Как можно заниматься анализом данных?

- Data scientist
 - Работа с данными
 - Знание инструментов и методов
 - Опыт решения задач
- Менеджер
 - Понимание, как работает машинное обучение
 - Понимание узких мест, оценивание сроков
- Заказчик
 - Метрики качества
 - Требования к данным
 - Ограничения современных подходов

Машинное обучение в искусстве и дизайне



Машинное обучение в искусстве и дизайне

- Генерация музыки и видео
- Генерация текстов и поэзии

VIOLA:

Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:

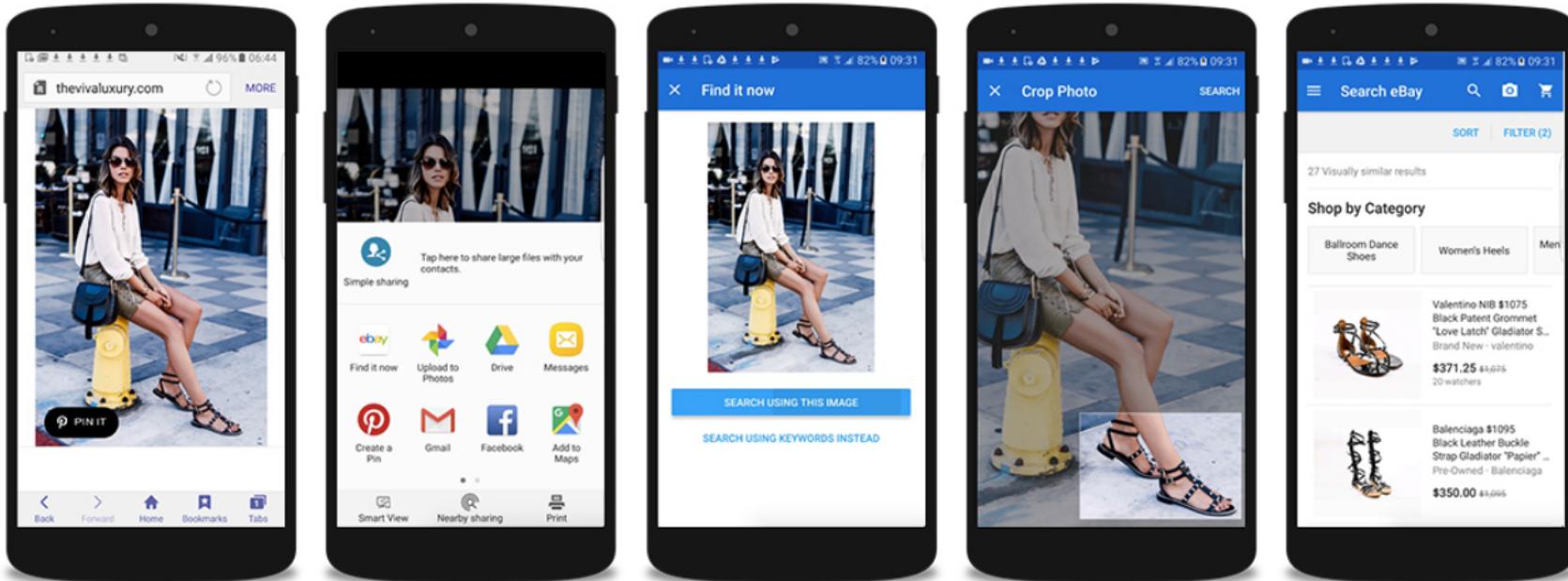
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.





Машинное обучение в искусстве и дизайне

- Генерация новых стилей одежды и классификация стиля
- Поиск товаров по фотографии



ebay

Машинное обучение в искусстве и дизайне

- Новые инструменты для творчества
- Новые возможности в fashion
- Искусственный интеллект не может придумывать принципиально новое в искусстве
- Задача автора — придумывать новые концепции, а не заниматься рутинной работой

Интересные предложения

РЕКОМЕНДУЕМ



Кулеры и системы охлаждения д...
[Deepcool NEPTWIN V2](#)

от 2 548 ₽

ХИТ ПРОДАЖ



Защитные пленки и стекла для т...
[Защитное стекло Xiaomi](#)

от 149 ₽

ТОВАР НЕДЕЛИ · ДО -20%



Электрообогреватели и тепловы...
[Ballu BOH/CL-07](#)

от 1 630 ₽

СКИДКИ ДО -10%



Цифровые бытовые метеостанции
[Oregon Scientific LW301](#)

от 8 990 ₽

РЕКОМЕНДУЕМ



Жесткие диски, SSD и сетевые н...
[Samsung MZ-7KE512BW](#)

Территория Grohe



ХИТ ПРОДАЖ



Дрели, шуруповерты, гайковерты
[Интерскол DA-12EP-01 ...](#)

РЕКОМЕНДУЕМ



Жесткие диски, SSD и сетевые н...
[Samsung MZ-7KE1T0BW](#)

Похожие товары



Canon EOS 200D Kit

от 32 390 ₽

1 отзыв 109 предложений

Любительская зеркальная
фотокамера

Байонет Canon EF/EF-S

Объектив в комплекте, модель
уточняйте у продавца

Матрица 25.8 МП (APS-C)

Цвет:

Цены 109



4.0

Canon EOS 750D Kit

от 32 650 ₽

5 отзывов 133 предложения

Любительская зеркальная
фотокамера

Байонет Canon EF/EF-S

Объектив в комплекте, модель
уточняйте у продавца

Матрица 24.7 МП (APS-C)

Цвет:

Цены 133



Canon EOS M10 Kit

от 21 250 ₽

4 отзыва 89 предложений

Фотокамера с поддержкой
сменных объективов

Байонет Canon EF-M

Объектив в комплекте, модель
уточняйте у продавца

Матрица 18.5 МП (APS-C)

Цвет:

Цены 89



до -30%



Canon EOS M6 Kit

от 38 300 ₽

до 84 940 ₽

2 отзыва 105 предложений

Фотокамера с поддержкой
сменных объективов

Байонет Canon EF-M

Объектив в комплекте, модель
уточняйте у продавца

Матрица 25.8 МП (APS-C)

Цвет:

Цены 105



С этим товаром часто покупают



4.0

HIPER MP15000

Универсальные внешние аккумулято...

от 1 453 ₽

30 отзывов 33 предложения

Цвет: ●

[Цены 33](#)[Цены 66](#)[Цены 16](#)[Цены 4](#)

2.5

Apple EarPods (Lightning)

Наушники и Bluetooth-гарнитуры

от 1 590 ₽

6 отзывов 66 предложений

Цвет: ○



Сетевая зарядка Apple

Зарядные устройства и адаптеры

от 642 ₽

16 предложений

Производитель: Apple

Тип: сетевая зарядка

Разъем подключения: for Apple
(Lightning)



Smart cover Apple

Чехлы для планшетов

от 3 080 ₽

4 предложения

Производитель: Apple

Совместимость: Apple

Тип: smart cover

Функция подставки: Да

Цвет: ● ●

Рекомендательные системы

- Персональные предложения для каждого пользователя на основе большого числа факторов
- Учёт статистики по всем пользователям сервиса
- Новый взгляд на маркетинг
- Работа маркетолога — правильное использование рекомендательных механизмов, выбор каналов для взаимодействия ИИ и пользователей

Рекомендации контента



Что мешает городским железным дорогам
Железная дорога – это не только рельсы и шпалы для перевозки...
[VARLAMOV.RU](#)

13 сюрпризов неайтишной компании
Отработав много лет в компаниях, занимающихся разработкой программного обеспечения на заказ, невольно начинаешь задумываться — а как всё-таки выглядит вся эта суэта с...
[ХАБРАХАБР](#)



Почему настоящие лидеры должны уметь думать медленно
Есть известная (но, возможно, апокрифическая) история о том, как теория относительности пришла в голову Альберту Эйнштейну, когда он ехал на велосипеде. Говорят, что Уоррен Баффетт читает не меньше шести часо...
[HBR-RUSSIA.RU](#)



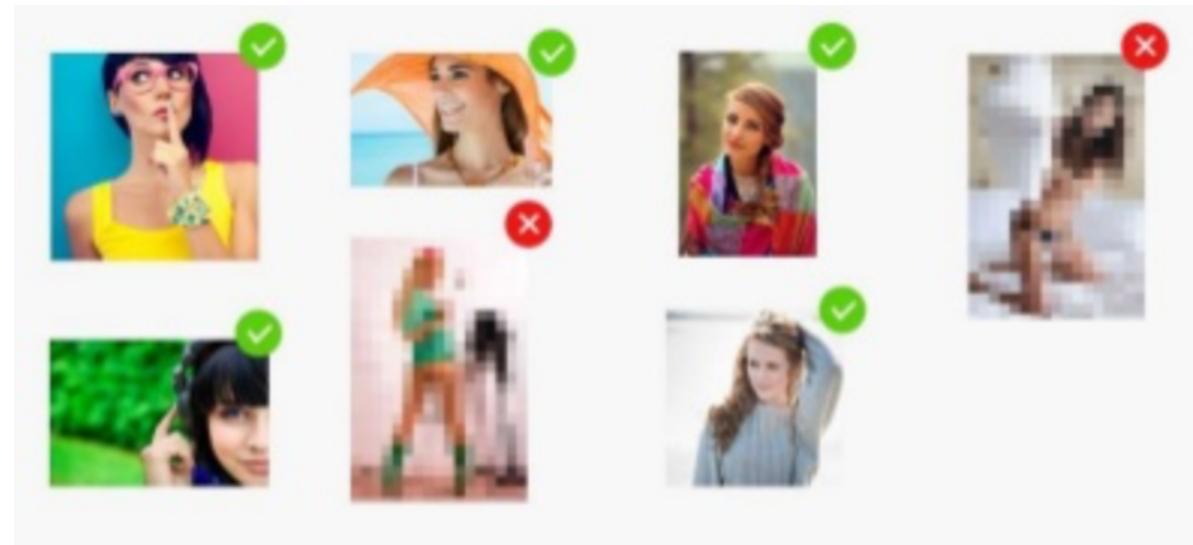
Математики обыграли букмекеров
Букмекеров можно обыгрывать, если учесть, что они иногда выставляют не самые выгодные для себя коэффициенты, чтобы снизить риски. Международная груп...
[N+1](#)

Рекомендации контента

- Медийный бум приводит к взрывному росту объёмов информации в сети
- Рекомендательные системы помогают ориентироваться
- Для авторов — поиск целевой аудитории
- Пионеры в Китае — Toutiao (более 100 миллионов активных пользователей) и другие платформы

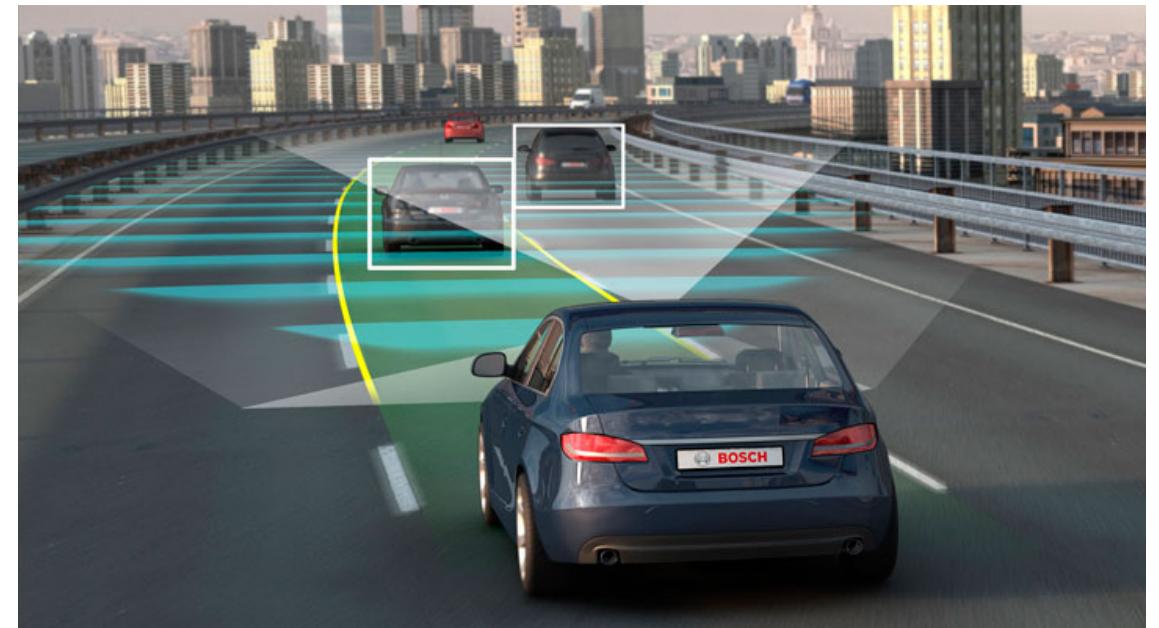
Автоматическая модерация

- Сервис онлайн-знакомств
- Фото в профиле не должно содержать эротику и не должно быть фотографией знаменитости
- Точность 88%
- Экономия порядка \$1,000,000 в год



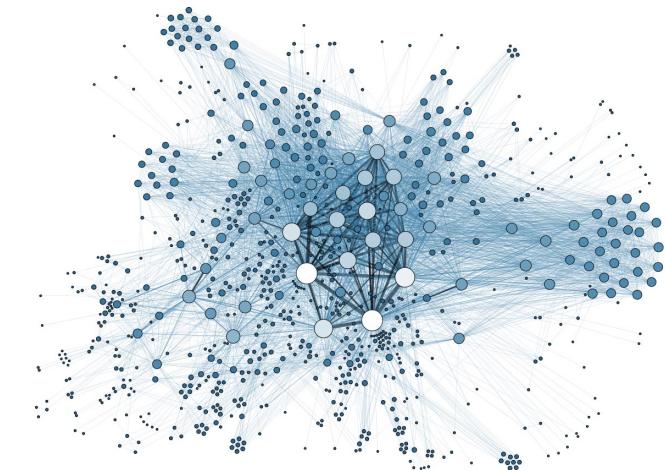
Экономические вызовы

- Создание новых отраслей экономики
 - Личные помощники
 - Персонализированная медицина
 - Гибкая сфера услуг
- Отмирание ряда профессий
 - Дальнобойщики
 - Водители такси
 - Пилоты самолётов
 - Операторы колл-центров
- Замена человека в творческих видах деятельности



Машинное обучение в экономике и социологии

- Финансы и финтех: анализ транзакционных и поведенческих данных, обнаружение фрода
- Страхование: предсказание вероятности аварии по данным с датчиков автомобиля
- Оценка мнений и трендов: анализ данных социальных сетей, форумов, новостных ресурсов
- Data-driven принятие решений



Машинное обучение в юриспруденции

- Поиск близких дел и материалов
- Проверка договоров, оптимизация правовых текстов
- Автоматизация принятия решений в суде
- Выявление проблем и трендов

- Правовые вопросы: обработка и передача больших данных, персональные данные, творчество, автопилоты

Машинное обучение в HR

- Поиск кандидатов и предсказание исхода собеседования
- Помощь при ротации
- Предсказание ухода сотрудника
- Анализ внутренних форумов, выделение жалоб

Откуда взять данные?

- **Социология:** лайки, комменты, посты в соцсетях и блогах
- **Международные отношения:** новости из разных источников, высказывания официальных лиц, курсы валют и акций
- **Экономика:** трансакции физических и юридических лиц, макроэкономические показатели, отчетность компаний
- **Лингвистика:** литературные произведения, общение в Интернете
- **Логистика:** данные по грузоперевозкам, загруженности складов, розничным продажам
- **Маркетинг:** поведение пользователей на сайте, онлайн-покупки и поисковые запросы, электронная переписка
- **Юриспруденция:** протоколы судов, данные о правоохранительной деятельности