

# Введение в анализ данных

Лекция 1

Введение

Евгений Соколов

[esokolov@hse.ru](mailto:esokolov@hse.ru)

НИУ ВШЭ, 2019

# Как перевести часы в минуты?



# Как перевести часы в минуты?

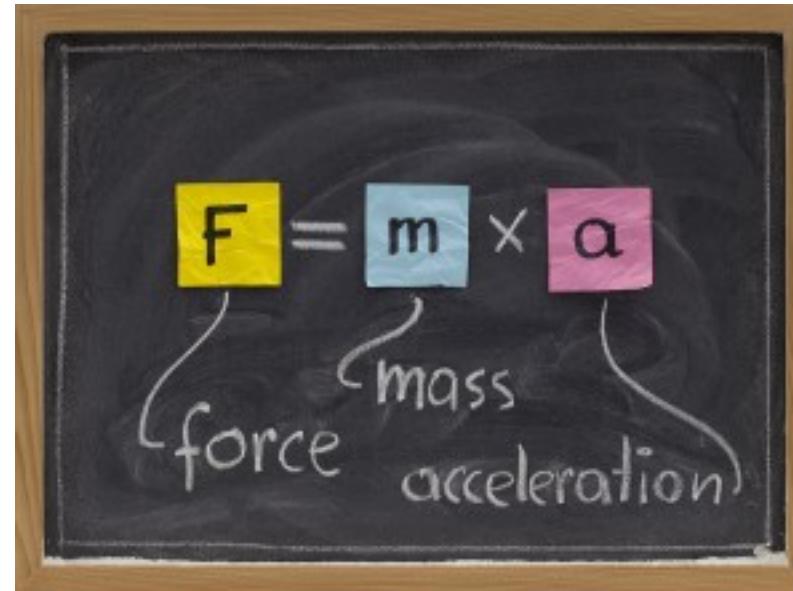
- $x$  — часы
- $f(x) = 60x$  — преобразование в минуты, функция

# 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

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

$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} +$  Позволяют найти скорость воздуха и давление в любой точке

$$\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} +$$
 Очень тяжело решать

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

# 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

— это про то, как восстановить сложные зависимости  
по конечному числу примеров

Организационное

# Про курс

- wiki:  
[http://wiki.cs.hse.ru/Майнор Интеллектуальный анализ данных/\\_Введение в анализ данных](http://wiki.cs.hse.ru/Майнор_Интеллектуальный_анализ_данных/_Введение_в_анализ_данных)
- [https://telegram.me/hse minor intro dm 2019](https://telegram.me/hse_minor_intro_dm_2019)
- Домашние задания
- Проверочные работы
- Контрольная работа (где-то в апреле)
- Письменный экзамен
- Автоматы — от 6 и выше при хорошей контрольной

# Про оценку

- $O_{\text{накопленная}} = 0.2 * O_{\text{самостоятельные}} + 0.6 * O_{\text{дз}} + 0.2 * O_{\text{контрольная}}$
- $O_{\text{итоговая}} = 0.7 * O_{\text{накопленная}} + 0.3 * O_{\text{экзамен}}$

# Про план курса

- Введение
- Метод k ближайших соседей
- Математика для анализа данных
- Линейные методы
- Решающие деревья и случайные леса
- Кластеризация
- Рекомендательные системы
- ...

# Про литературу

- Luis Pedro Coelho and Willi Richert. Building Machine Learning Systems with Python.
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning.
- Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms.

# Про литературу

- Курсы ПМИ ФКН:
  - [http://wiki.cs.hse.ru/Машинное обучение 1](http://wiki.cs.hse.ru/Машинное_обучение_1)
  - [http://wiki.cs.hse.ru/Машинное обучение 2](http://wiki.cs.hse.ru/Машинное_обучение_2)
- Онлайн-курсы:
  - <https://www.coursera.org/learn/machine-learning>
  - <https://www.coursera.org/learn/introduction-machine-learning>
  - <https://coursera.org/specializations/machine-learning-data-analysis>

# 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
- $\ell$  — 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

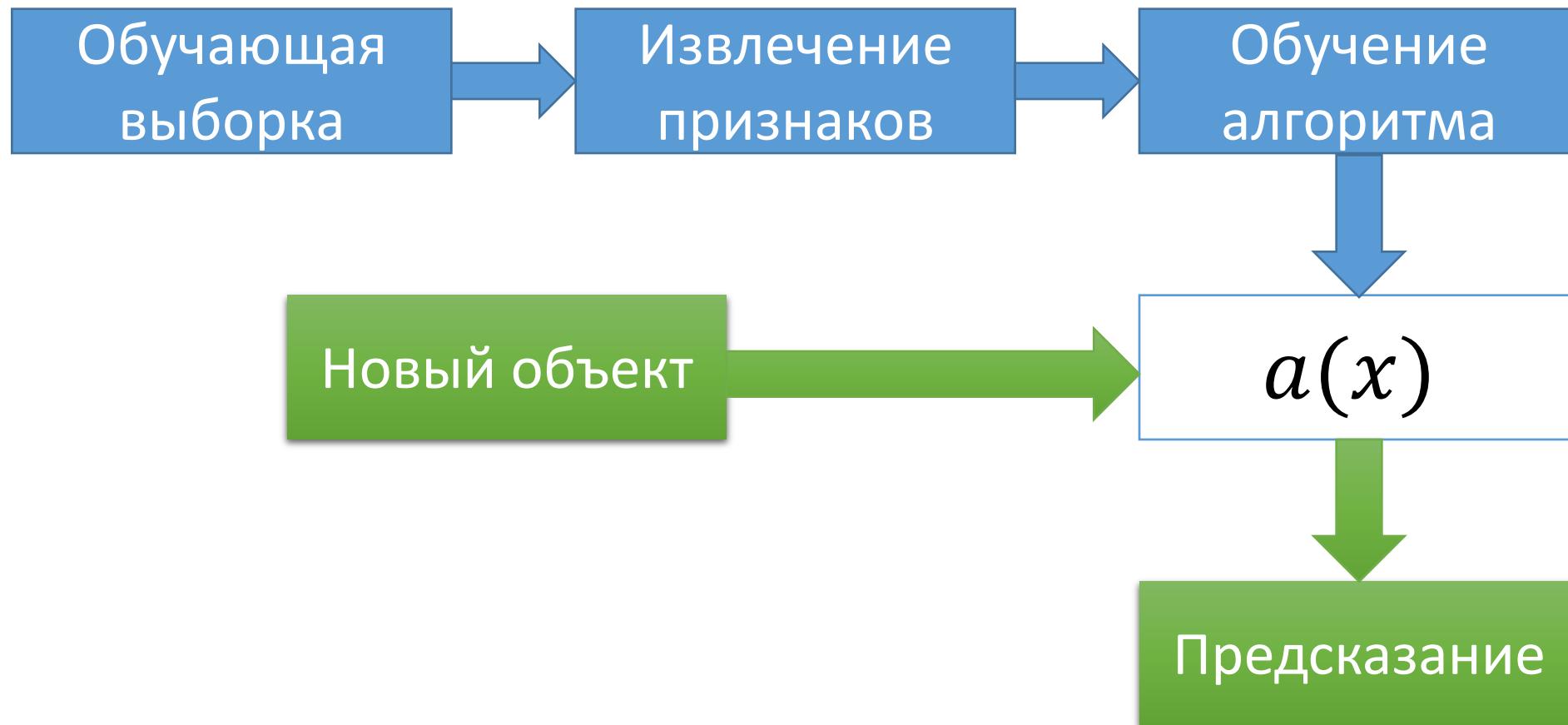
# Функционал качества

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

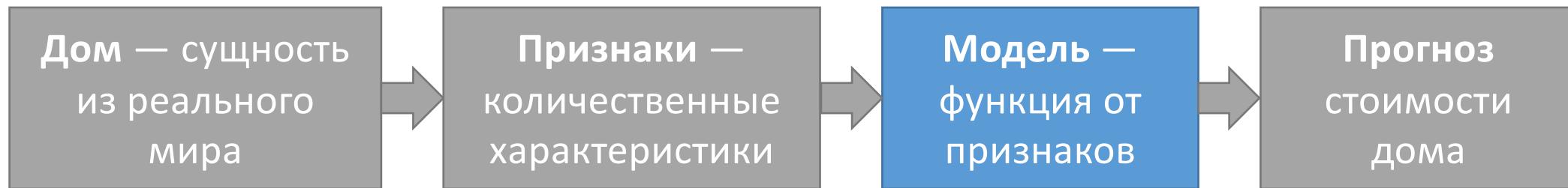
# Обучение алгоритма

- 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

# Машинное обучение



# Предсказание стоимости дома



# Предсказание стоимости дома

Обучающая выборка:

Площадь	Цена
50	250
60	340
10	20
90	800

Возможные признаки:

- площадь
- площадь<sup>2</sup>
- площадь<sup>3</sup>
- $\sin(\text{площадь})$
- $\sqrt{\text{площадь}}$
- 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:

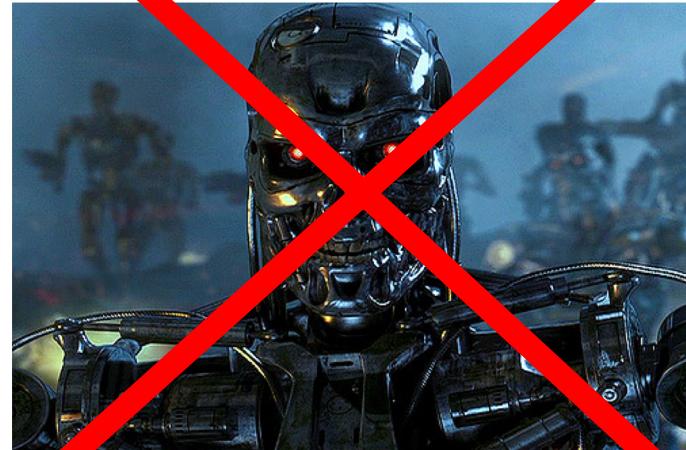
- Линейная:  $w_1 * \text{площадь} + w_2 * \text{год} + w_3 * \text{бассейн} + w_4 * \text{комнаты} + w_5 * \text{удалённость} + w_6 * \text{полиция}$
- Решающие деревья
- Нейронные сети
- Метод k ближайших соседей
- И так далее

# Что нужно знать

1. Как сформулировать задачу?
2. Какие признаки использовать?
3. Откуда взять обучающую выборку?
4. Как выбрать метрику качества?
5. Как обучить алгоритм?
6. Как оценить качество алгоритма?

Зачем это нужно?

# Искусственный интеллект



Сильный ИИ

через 20-100 лет

Яндекс

фильм где астронавту протыкают скафандр



Найти

ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ



Марсианин

The Martian, 2015 (16+)

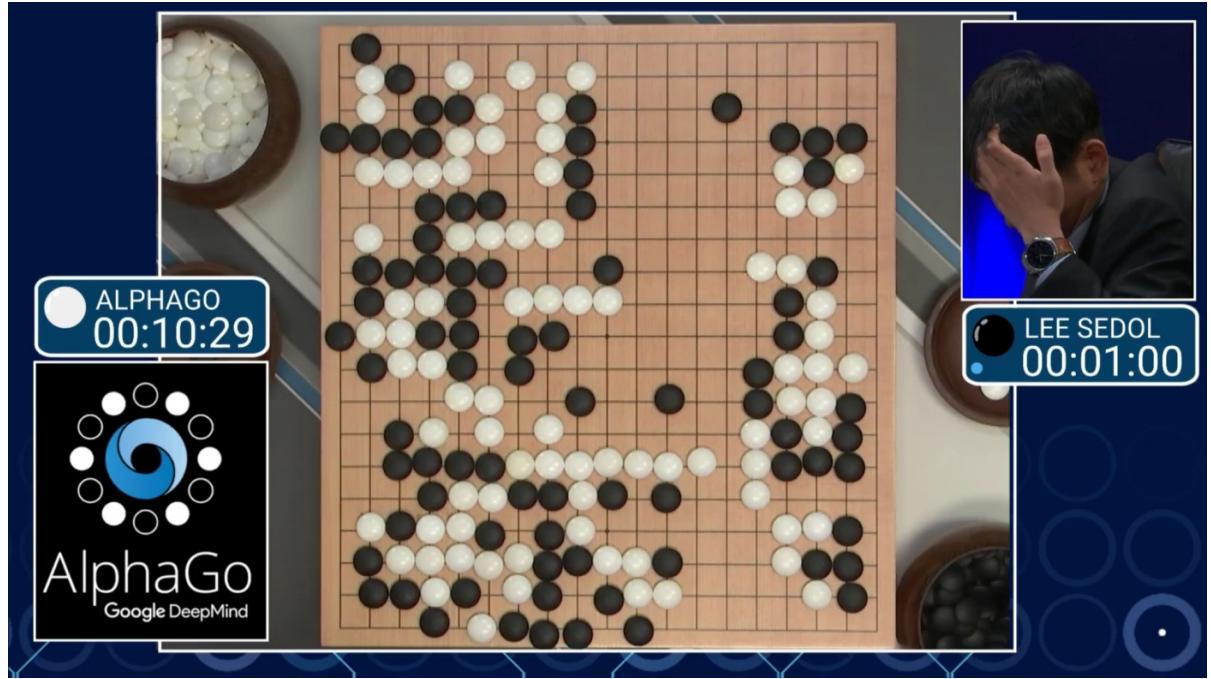
Марсианская миссия «Арес-3» в процессе работы была вынуждена экстренно покинуть планету из-за надвигающейся песчаной бури. Инженер и биолог Марк Уотни получил повреждение скафандра во время песчаной бури. Сотрудники миссии, посчитав его погибшим,...  
[Читать дальше](#)

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

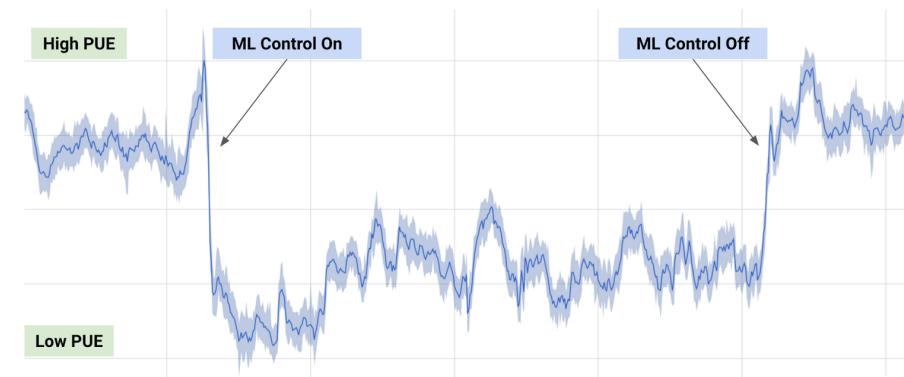


# Машинное обучение в HR

- 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 Using Python	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: A Tour of Statistical Analysis and Computation Using 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

# На следующей лекции

- Типы задач в машинном обучении
- Типы признаков
- Примеры прикладных задач