Введение в машинное обучение

Надежда Чиркова
факультет компьютерных наук НИУ ВШЭ

Является ли сообщение спамом?

Пишите Всероссийские проверочные работы (4, 5, 10, 11 класс)? ⚠️
Тогда Вам к нам! 💙

📍 Наша группа - https://vk.com/club[redacted] 📍
🔥 У нас будут материалы (задания и ответы) на все классы. 🔥

Русский язык на 18 и 20 апреля для 4-х классов уже есть! 📄
Спеши купить! За день до проведения станет дороже! ⚠️

Является ли сообщение спамом?

Что подарить, чтобы не прогадать? 🧐

Стилизованные портреты на холсте - отличный выход из ситуации!
Это индивидуальный подарок, при этом подходящий всем без исключения. 😊👍

➡️ Группа - https://vk.com/[redacted]
👉 Оформите заказ у менеджера можно прямо сейчас - https://vk.com/[redacted]

Является ли сообщение спамом?

Как составляют прогноз погоды, почему он не всегда точен и где лучше всего смотреть погоду:
https://postnauka.ru/faq/75484

Спам и не спам

Как научить компьютер автоматически отличать рекламные сообщения и удалять их?

Первое, что приходит в голову

```
if num_likes > 100:
    return False
elif "недорого" in s or "поможем" in s or "в наличии" in s:
    return True
elif num_smiles > 20:
    if s.count("!") > 10:
        return True
    elif "http" not in s:
        return True
    else:
```

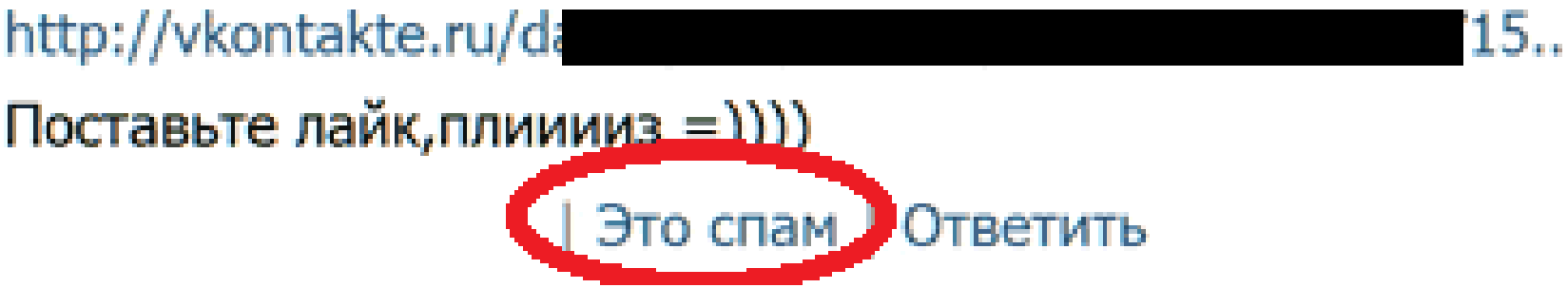
Второе, что приходит в голову

```
super_bad_words = ["лайк", "недорого", "уникальный", "жми"]
bad_words = ["новый", "лучший", "купите"]
good_words = ["считаю", "думаю", "нет"]
super_good_words = ["понравился", "худший"]
score = 0
score += sum([5 for word in s if word in super_bad_words])
score += sum([2 for word in s if word in bad_words])
score += sum([-2 for word in s if word in good_words])
score += sum([-5 for word in s if word in super_good_words])
if score > 0:
    return True
else:
    return False
```

В чем минус подхода?

- долго
- негибко
- неоптимально
- требует опыта

Мы уже накопили много опыта!



Альтернатива

Создавать алгоритм автоматически!

Данные = много примеров:

Пишите Всероссийские проверочные работы (4, 5, 10, 11 класс)? ⚠️ Тогда Вам к нам! 💙 📍 Наша группа - https://vk.com/club[redacted] 📍 🔥 У нас будут материалы (задания и ответы) на все классы. 🔥 Русский язык на 18 и 20 апреля для 4-х классов уже есть! 📔 Спеши купить! За день до проведения станет дороже! ⚠️	1
Что подарить, чтобы не прогадать? 😳 Стилизованные портреты на холсте - отличный выход из ситуации! Это индивидуальный подарок, при этом подходящий всем без исключения. 😊👍 ➡️ Группа - https://vk.com/[redacted] 👉 Оформите заказ у менеджера можно прямо сейчас - https://vk.com/[redacted]	1
http://vkontakte.ru/d[redacted]15.. Поставьте лайк,плиииииз =)))) Это спам Ответить	1
А вот и среда подходит к концу. Еще один успешно пройденный день на пути к пятнице. Кажется, даже сама погода радуется, подарив нам теплый и солнечный денек. Сон мудренее вечера, а это значит, что нужно отложить всю работу, все хвосты по учебе до утра, потому что ночь была создана для того, чтобы отбросить все заботы прошедшего дня.	0
Чтобы не пропустить интересные события культурной жизни школы, подписывайтесь на телеграм-канал: https://t.me/sch_int Будем писать о концертах, лекциях, встречах с интересными людьми. Там же будут новости науки и рецензии на книги от учителей и учеников «Интеллектуала».	0

Альтернатива

- ~~долго~~ быстро
- ~~негибко~~ один код на разные данные
- ~~неоптимально~~ лучше широкого круга решений
- ~~требует опыта~~ достаточно истории примеров

Альтернатива — машинное обучение

Цель: **автоматически** принимать решения

Используя свои знания, писать программу для конкретного случая

Используя накопленные **данные**, **компьютер** сам создает **оптимальную** программу

Альтернатива — машинное обучение

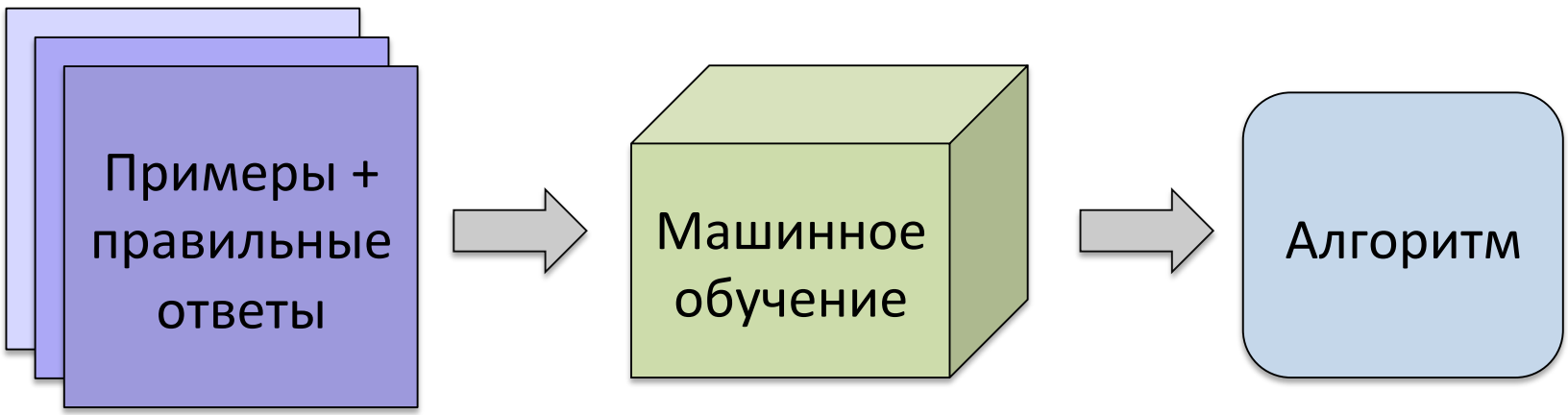
Цель: **автоматически** принимать решения

Используя свои знания, писать программу для конкретного случая

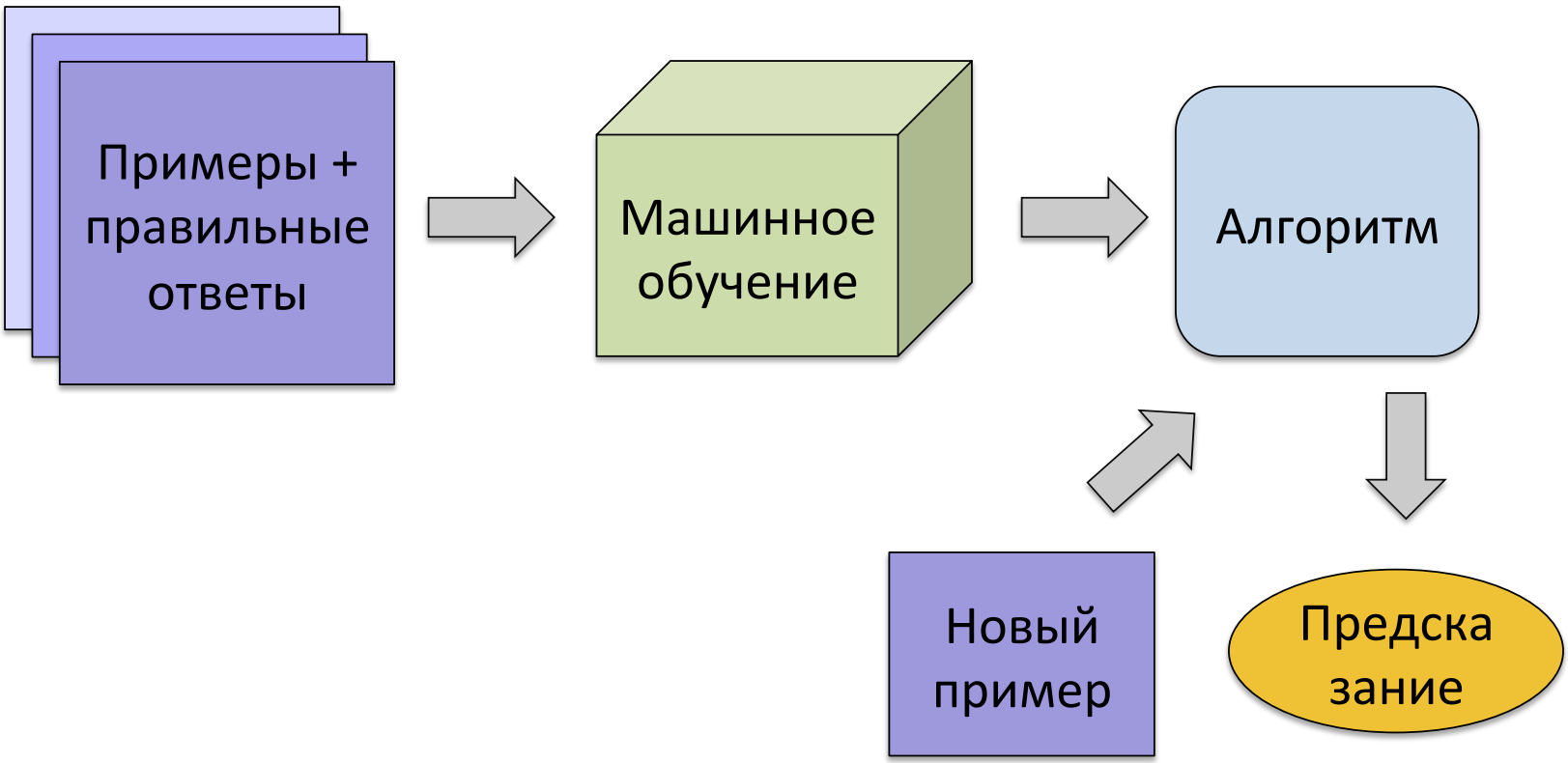
Используя накопленные **данные**, **компьютер** сам создает **оптимальную** программу

Тоже программа!

Задача машинного обучения



Задача машинного обучения



Откуда данные?

Вариант A: разметка, выполняемая людьми
(машинное обучение для автоматизации того, что умеет делать человек)

- распознавание объектов на изображении, слов в речи
- перевод с одного языка на другой
- ответы на вопросы
- постановка медицинского диагноза
- интеллектуальные игры
- ...

Откуда данные?

Вариант Б: из жизни, науки, производства, бизнеса, ...
(машинное обучение для поиска зависимостей в данных, для предсказаний)

- медицина: болен ли пациент гриппом
- банки: вернет ли заемщик кредит
- торговля: купит ли покупатель товар
- Интернет: кликнет ли посетитель сайта ссылку
- экономика: обанкротится ли предприятие
- финансы: каким будет курс доллара
- ...

Медицинская диагностика

Данные — таблица с признаками и **правильными ответами**

Пол	Давление	Температура	Лейкоциты	...	**Диагноз**
жен	140/90	38.2	$15 \cdot 10^9$	...	**Грипп**
муж	130/90	36.6	$7 \cdot 10^9$	...	**Здоров**
муж	120/80	39.1	$11 \cdot 10^9$	...	**Ангина**
		...		...	

Медицинская диагностика

Данные — таблица с признаками и **правильными ответами**

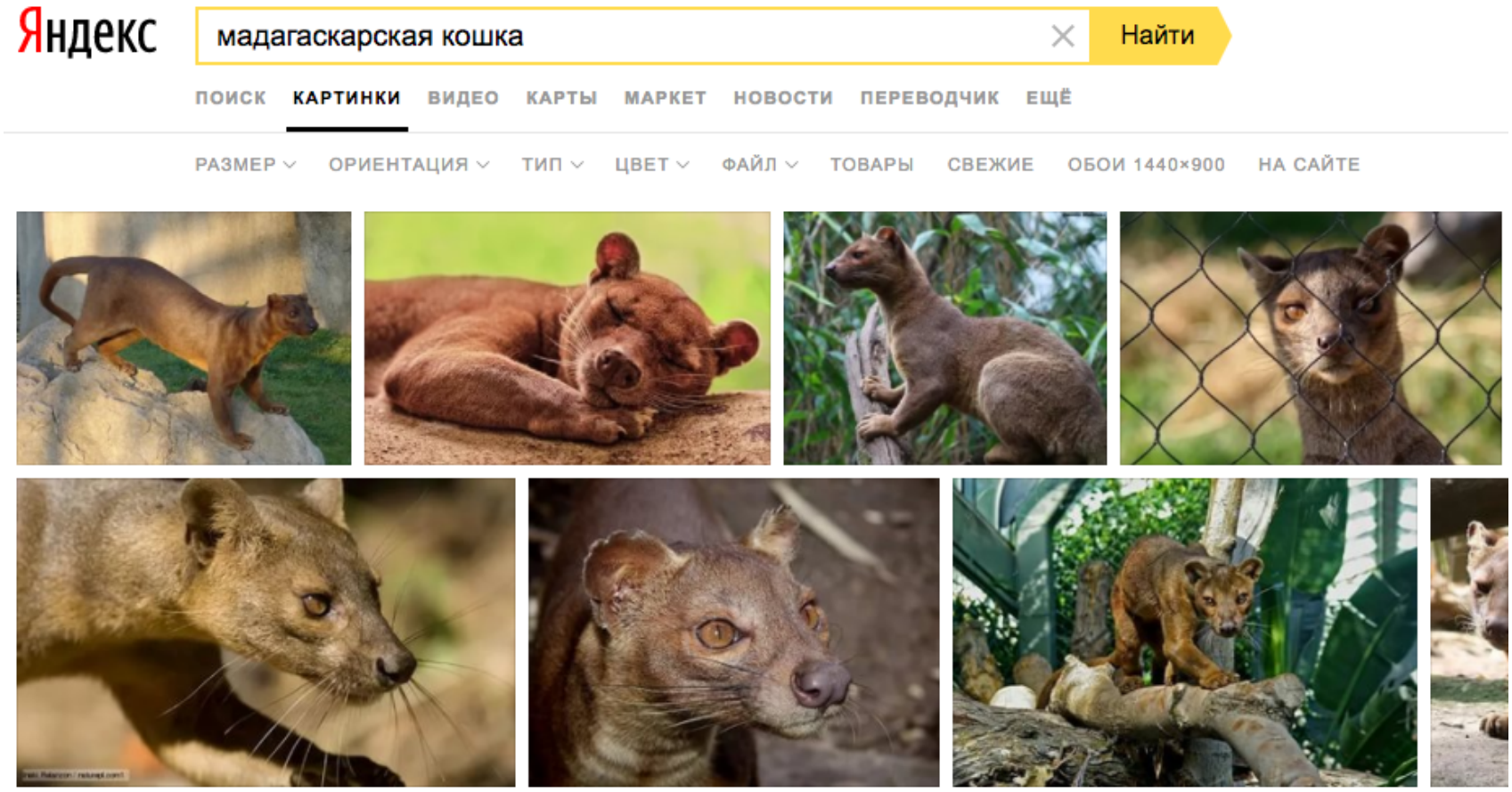
Пол	Давление	Температура	Лейкоциты	...	**Диагноз**
жен	140/90	38.2	$15 \cdot 10^9$	...	**Грипп**
муж	130/90	36.6	$7 \cdot 10^9$	...	**Здоров**
муж	120/80	39.1	$11 \cdot 10^9$	...	**Ангина**
		...		...	
жен	120/80	36.7	$8 \cdot 10^9$	...	?

Распознавание изображений: Imagenet

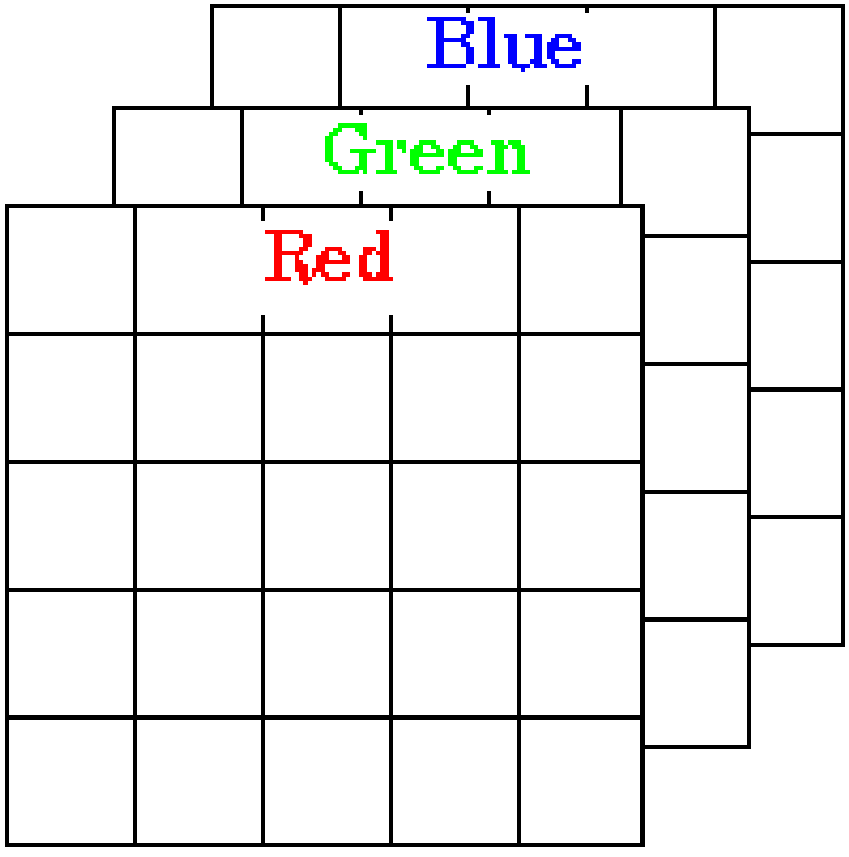
10 млн изображений с ручной разметкой классов (виды животных, объекты мебели, машины, здания, ...)

От 28.2% ошибок (2010) к 3.57% (2015)

Распознавание изображений: применение



Представление изображений в компьютере



Предсказание возраста

У многих пользователей «В контакте» не указан возраст (или указан неверно :), хотя многим приложениям и рекламодателям эта информация нужна.

друзей	подписок	like'ов на аватарке	...	**Возраст**
110	20	75	...	**18**
7	11	3	...	**45**
30	4	10	...	**24**
		...		

Предсказание возраста

Или проще — моложе/старше 30

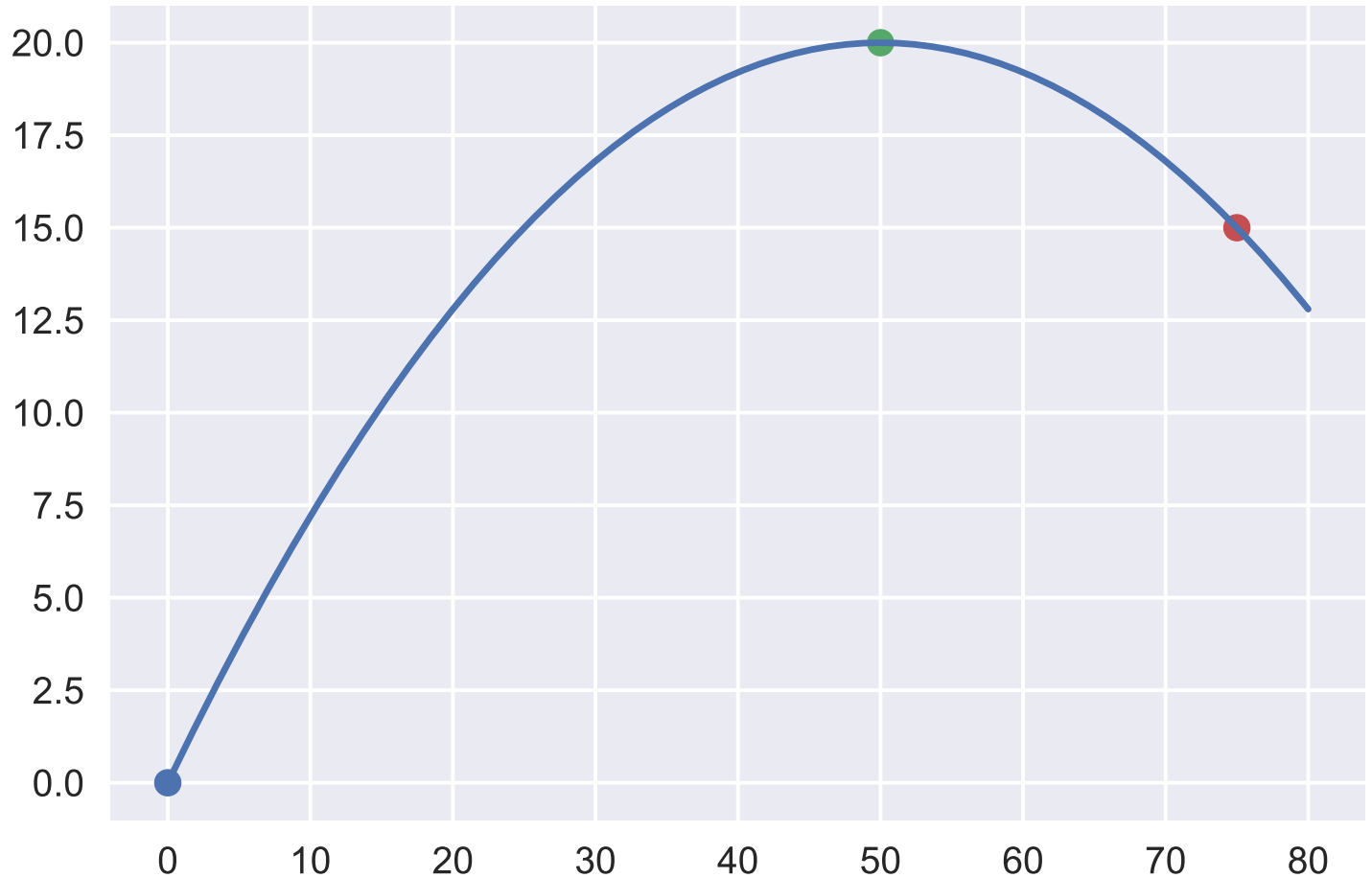
друзей	подписок	Like'ов на аватарке	...	**Моложе 30?**
110	20	75	...	**да**
7	11	3	...	**нет**
30	4	10	...	**да**
		...		

Два самых популярных вида задач

Классификация	Регрессия
предсказываем класс из конечного множества	предсказываем вещественное (или целое, или натуральное) число
предсказание моложе/старше 30	предсказание возраста
постановка диагноза	предсказание курса валют
распознавание объектов на изображении	предсказание температуры завтра

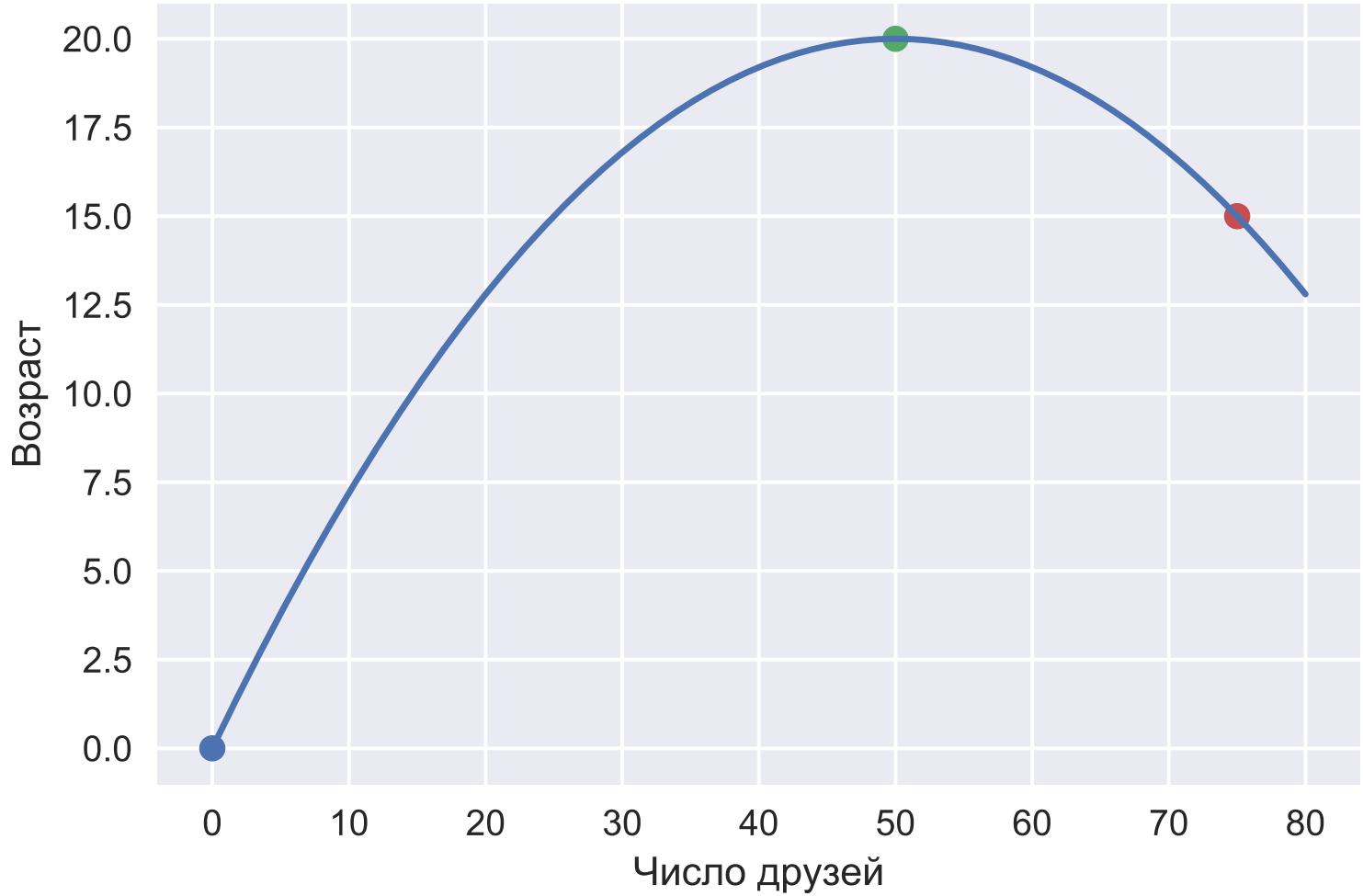
Задача про параболу

Парабола, проходящая через три точки:
(0, 0), (50, 20), (75, 15)



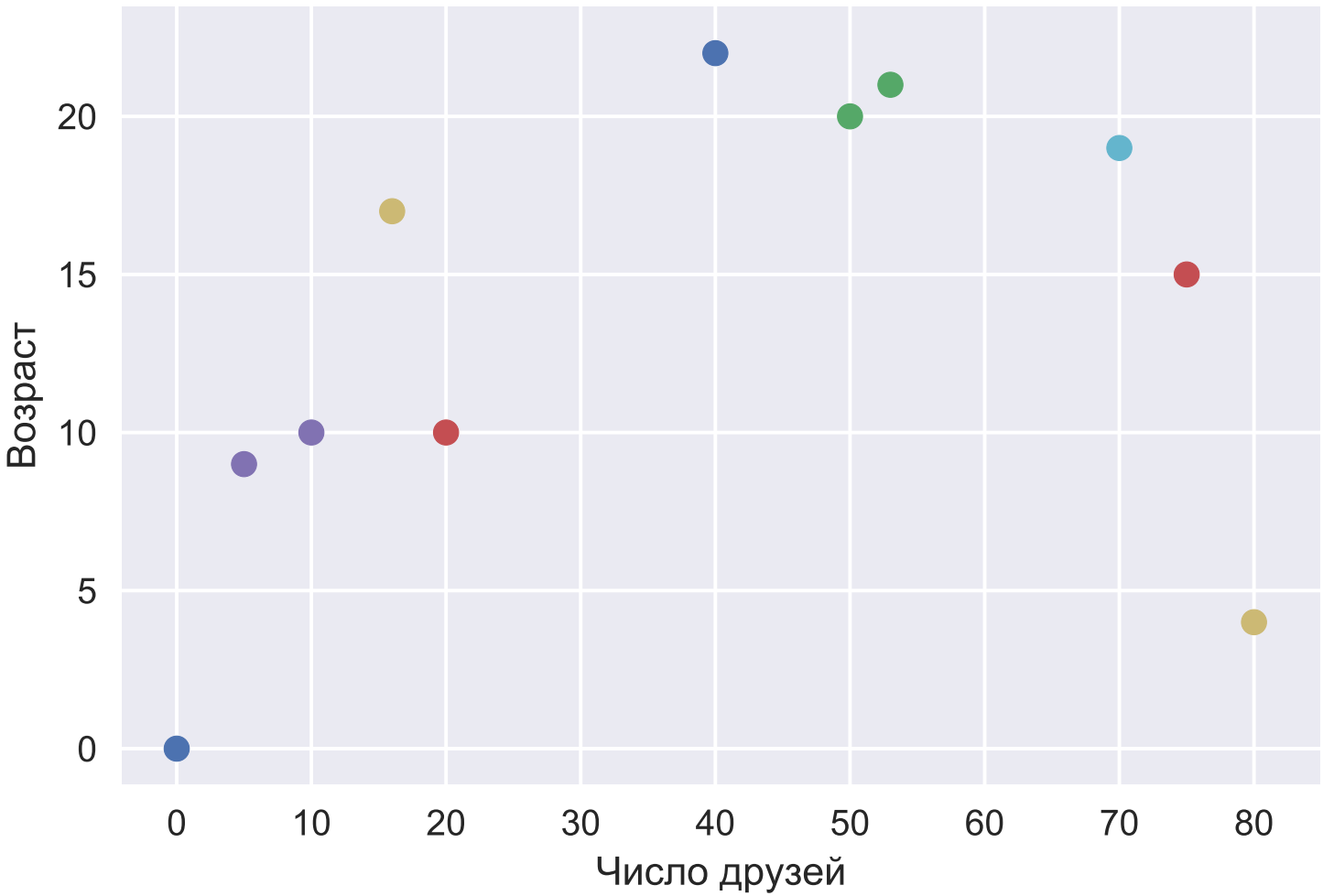
Предсказание возраста по числу друзей

Это решение задачи предсказания возраста по числу друзей, если объектов всего три!



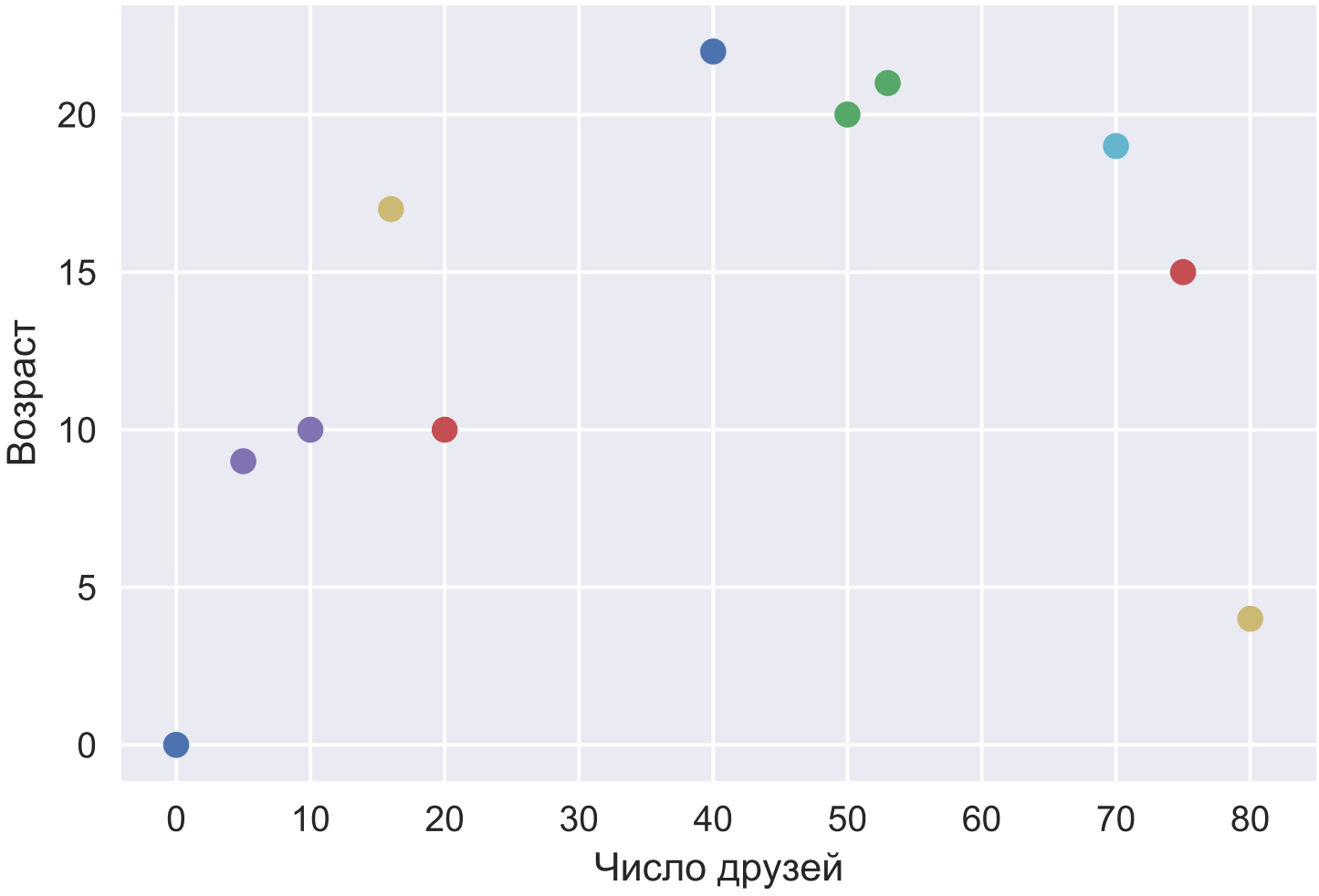
Предсказание возраста по числу друзей

А если объектов больше?



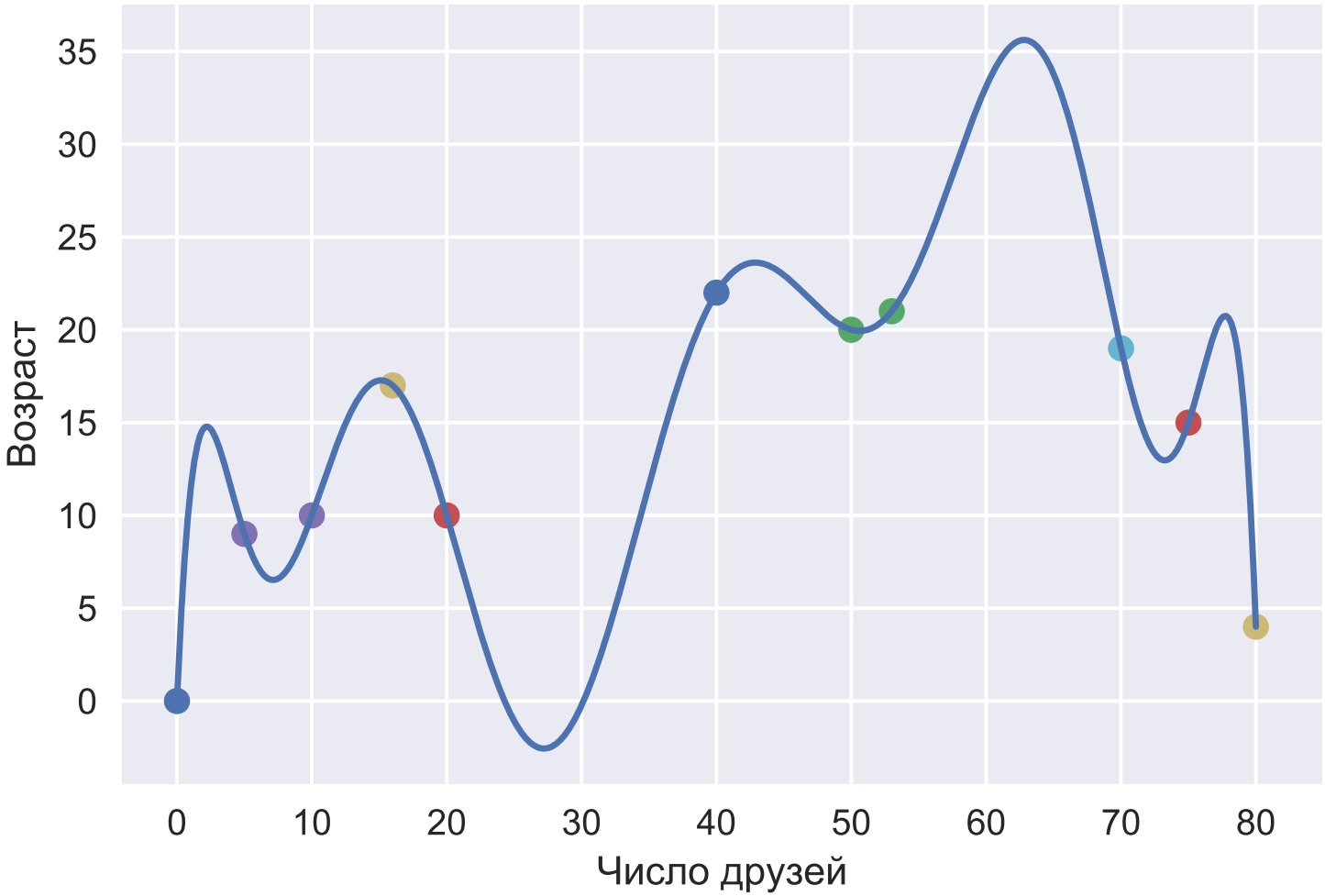
Предсказание возраста по числу друзей

А если объектов больше? Какой степени многочен пройдет через все точки?



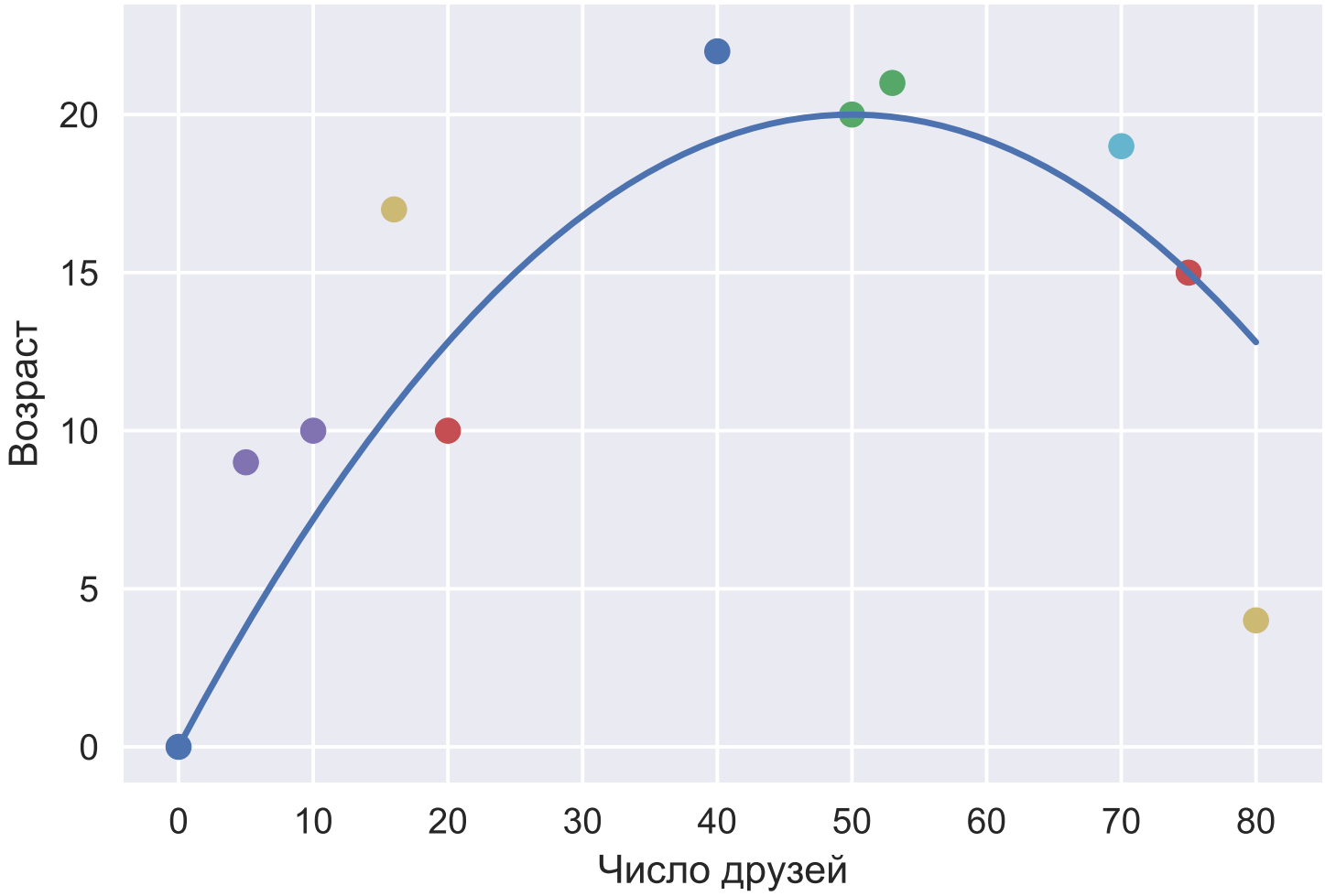
Предсказание возраста по числу друзей

Можно провести многочлен 10 степени через все точки, но если их еще больше?



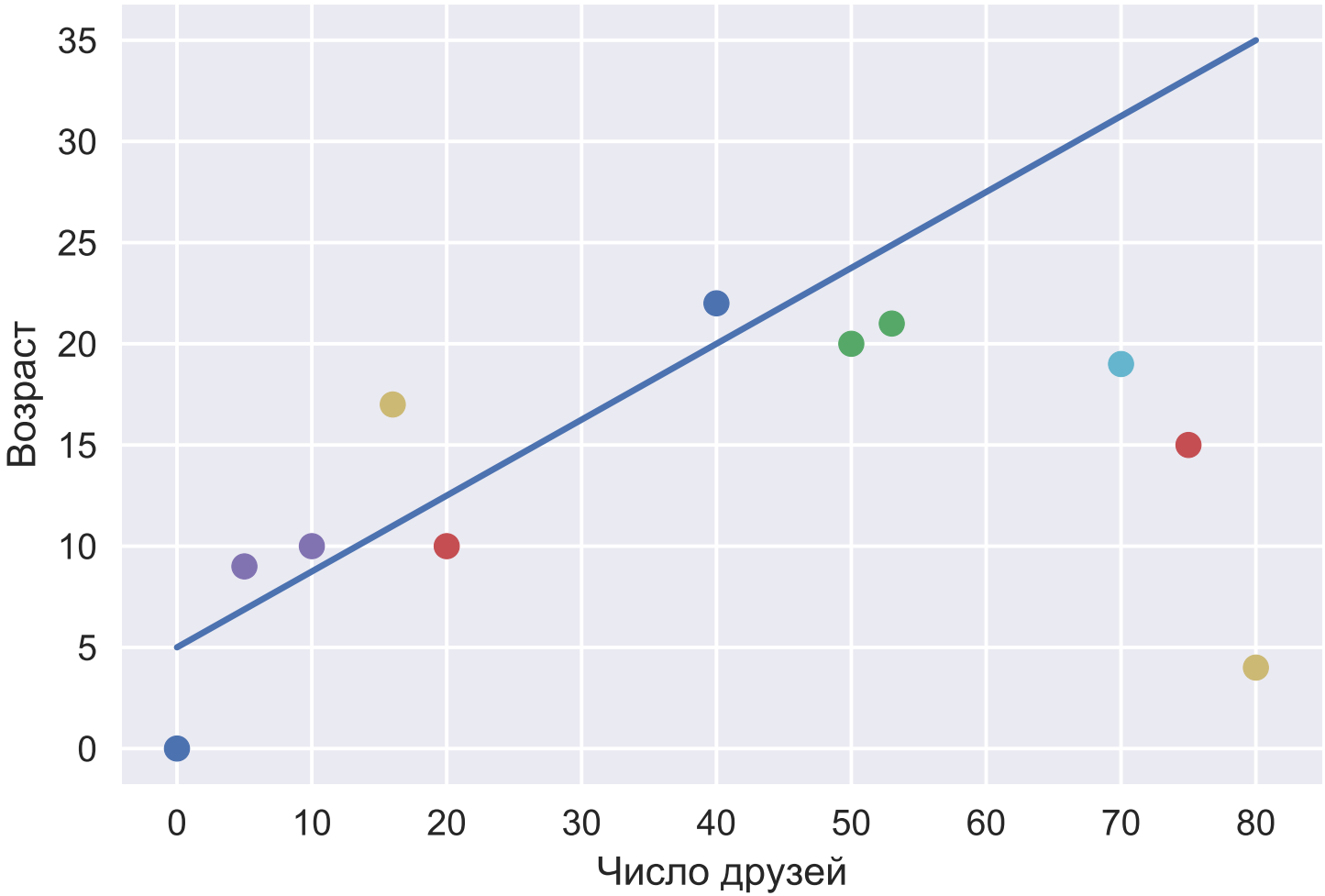
Предсказание возраста по числу друзей

Лучше приближенно, но проще, чем точно и сложно.



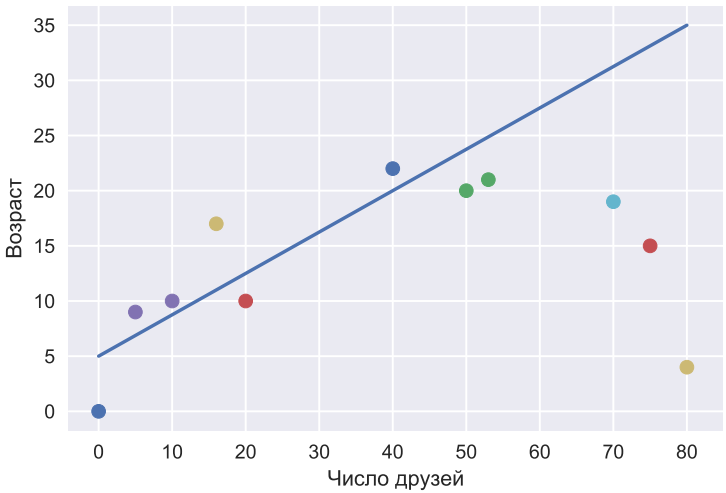
Предсказание возраста по числу друзей

Может еще проще?

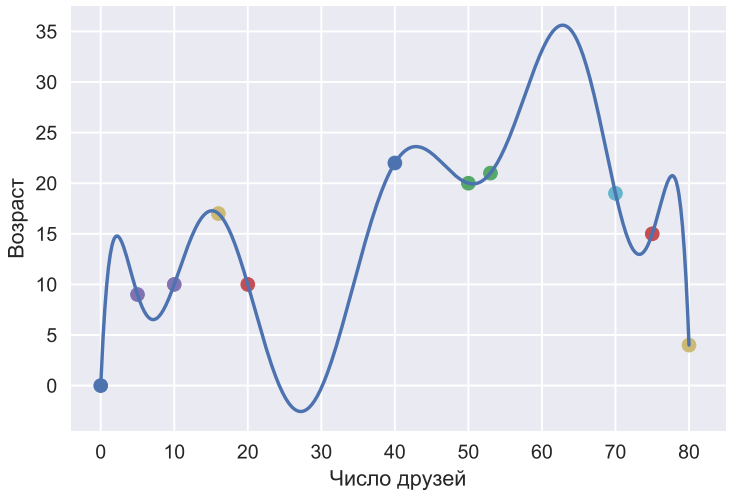


Баланс между точностью и обобщающей способностью

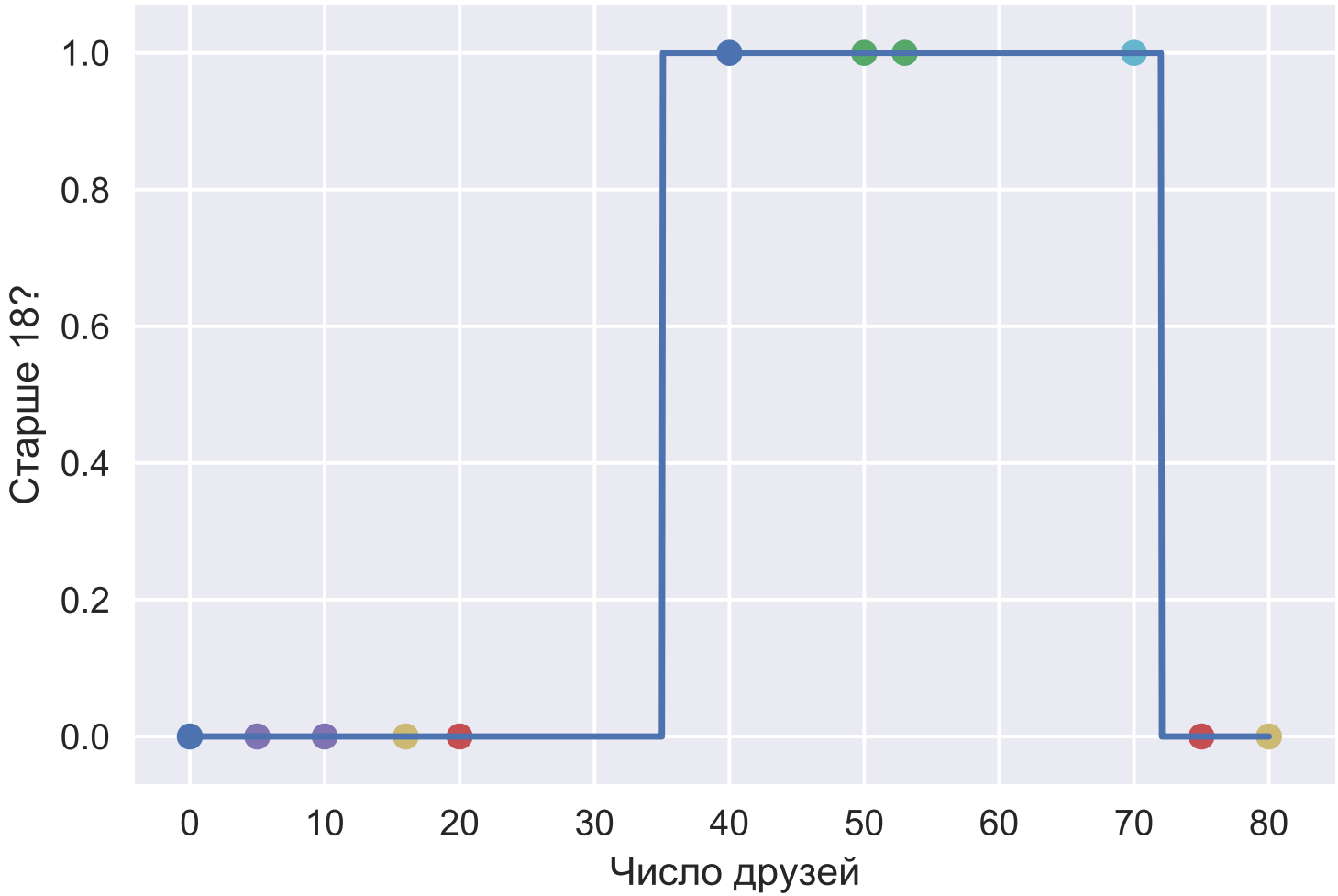
Недообучение



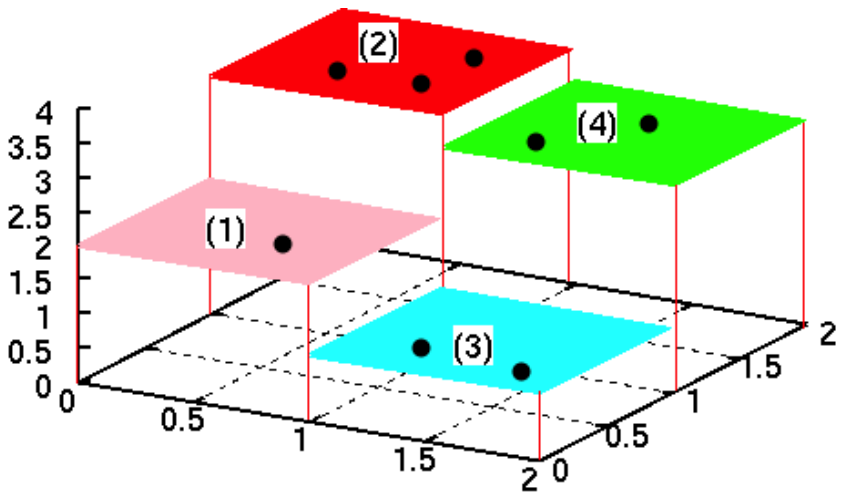
Переобучение

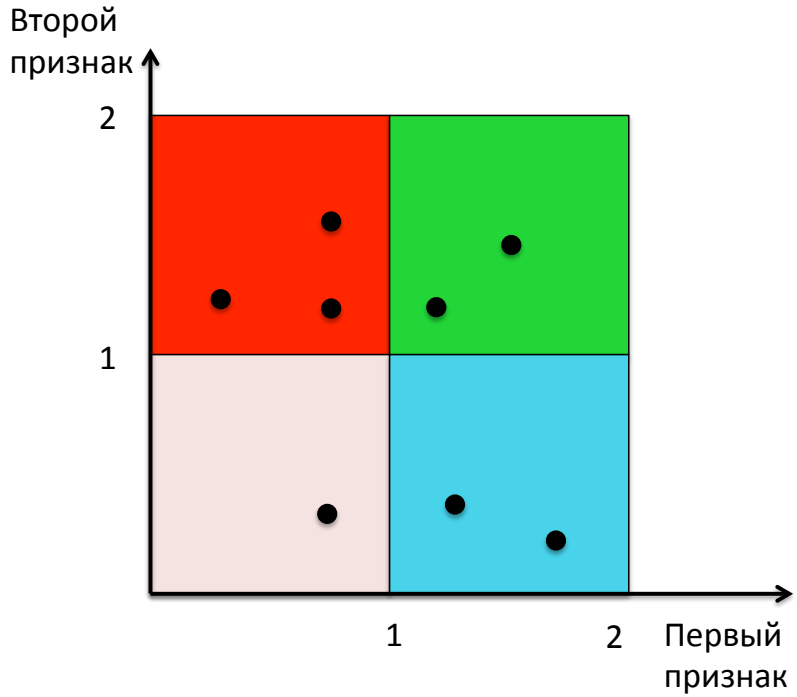


Задача приближения функции для классификации



Задача приближения функции для классификации



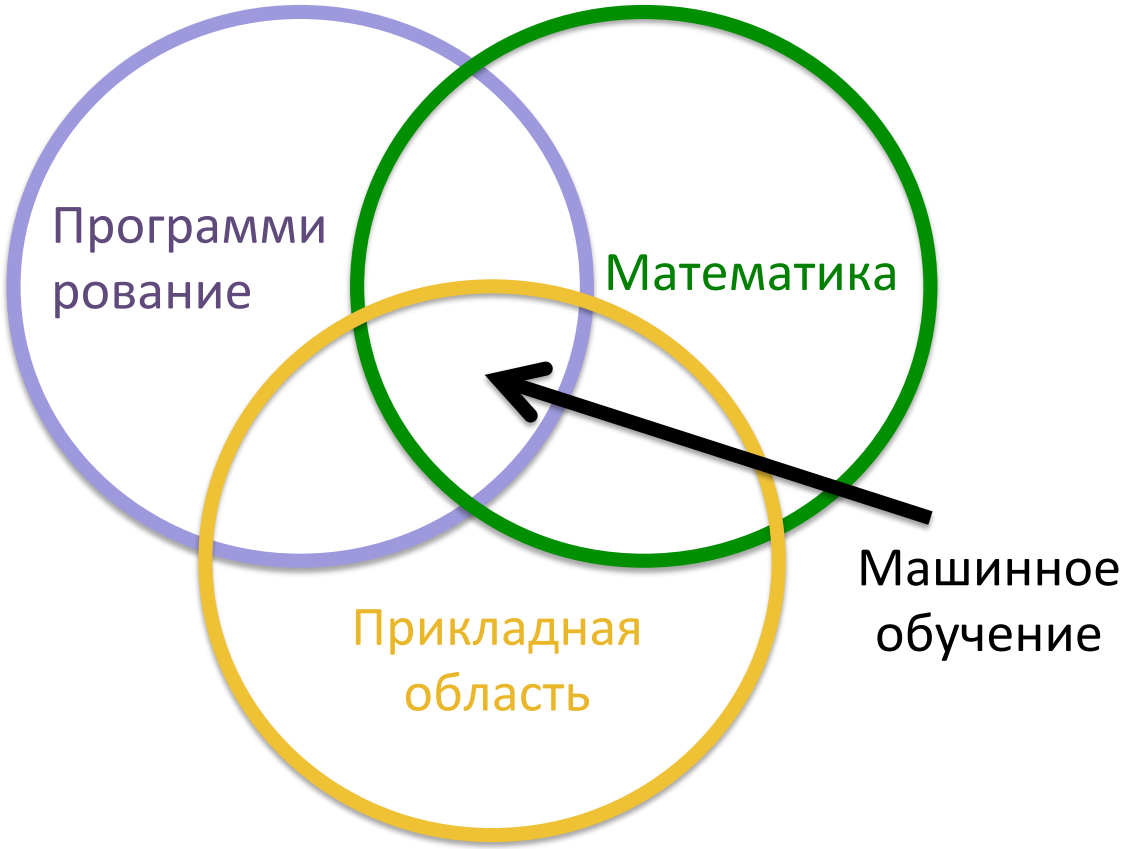


Итак,...

Машинное обучение:

- с точки зрения программирования: автоматическое создание программы, умеющей решать задачу предсказания, по данным
- с точки зрения математики: приближение функции по точкам
- находит применение во многих прикладных областях

Машинное обучение на стыке трех областей



Задачи машинного обучения

Задачи обучения с учителем: данные $X \in \mathbb{R}^{\ell \times d}$, $Y \in \mathbb{R}^{\ell}$
ℓ – число объектов, d – число признаков

- Регрессия: $Y \in \mathbb{R}^{\ell}$
- Классификация:
  - бинарная $Y \in \{0, 1\}^{\ell}$
  - многоклассовая с непересекающимися классами $Y \in \{1, \ldots, k\}^{\ell}$
  - многоклассовая с пересекающимися классами $Y \in \{0, 1\}^{\ell \times k}$
- Рекомендательные системы

Задачи обучения без учителя: данные $X \in \mathbb{R}^{\ell \times d}$

- Кластеризация: $X \in \mathbb{R}^{\ell \times d} \to Z \in \{1, \ldots, c\}^{\ell}$
- Понижение размерности: $X \in \mathbb{R}^{\ell \times d} \to Z \in \mathbb{R}^{\ell \times d'}$, $d' < d$
- Визуализация: $X \in \mathbb{R}^{\ell \times d} \to Z \in \mathbb{R}^{\ell \times 2}$

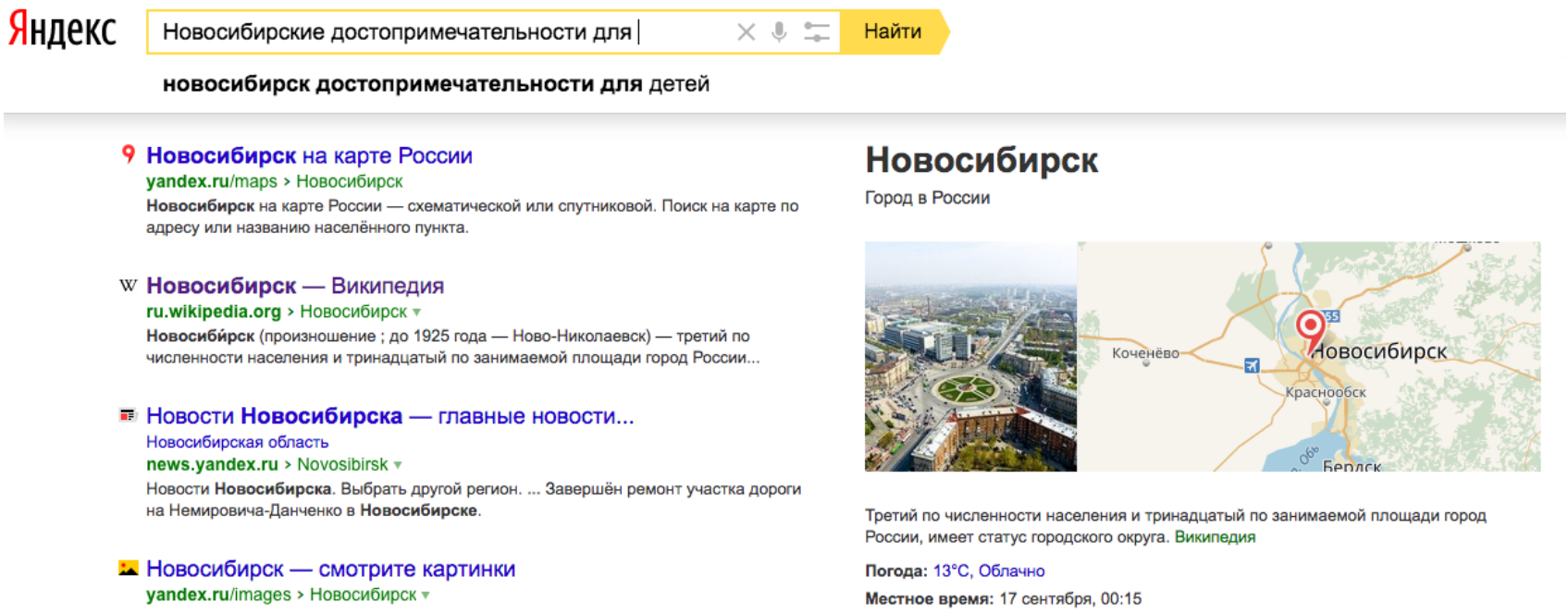
Чего не умеет машинное обучение

Разумеется, есть много задач, которые решаются без данных и их анализа:

- построение оптимального маршрута на карте
- сложные физические расчеты
- рендеринг трехмерных изображений
- ...

Для их решения нужно хорошо изучать информатику, алгоритмы и средства программирования!

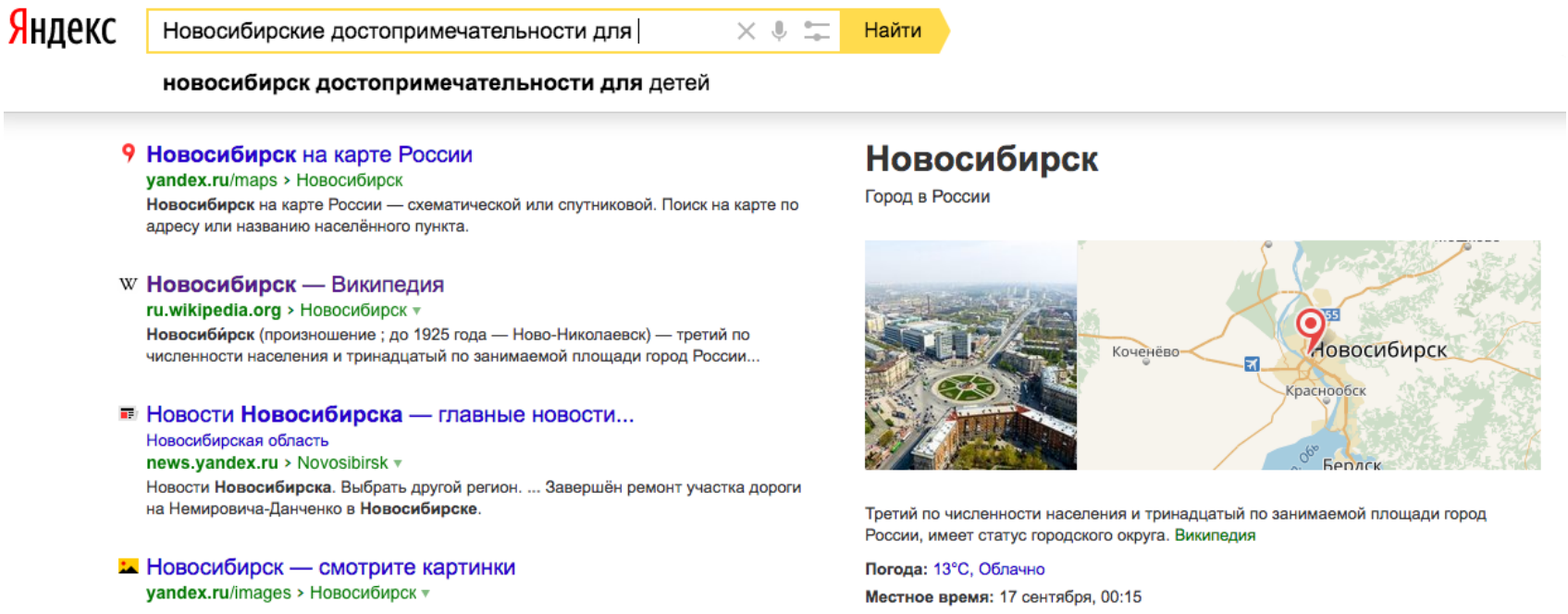
Поисковые сервисы



Задачи машинного обучения:

- предсказание релевантности страницы запросу
- предсказание релевантности страницы интересам пользователя
- распознавание именованных сущностей

Поисковые сервисы



Задачи программирования:

- быстрый отбор документов из огромной базы
- быстрое автозаполнение запроса
- выбор похожих запросов (с похожим множеством документов)

Self-driving cars

Задачи машинного обучения:

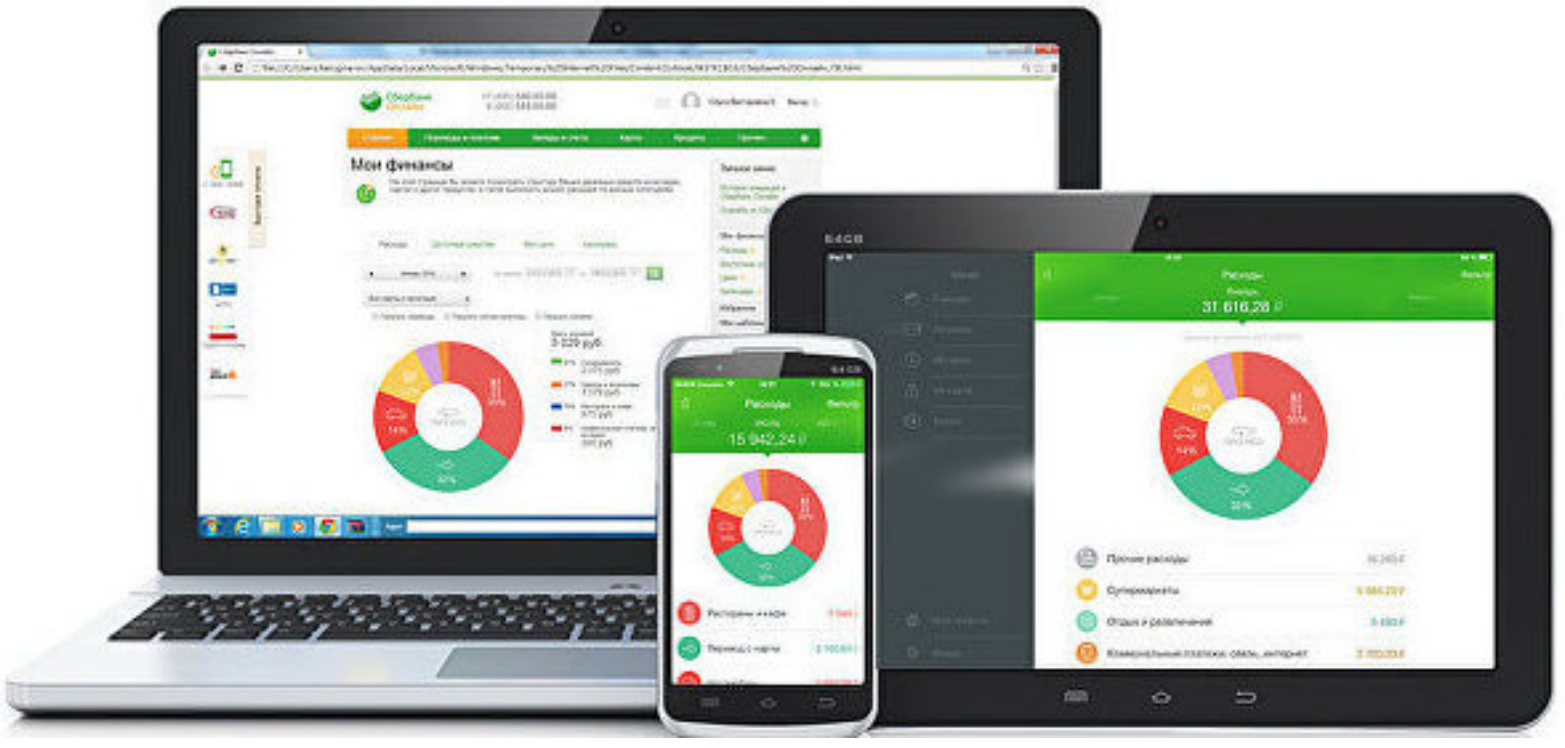
- распознавание знаков, сигналов светофора
- распознавание текстов на изображениях
- предсказание следующих действий пешеходов и других объектов
- планирование маневров
- прогнозирование времени прибытия в точку

Self-driving cars

Задачи программирования:

- поиск оптимального маршрута
- физические расчеты, моделирование движения

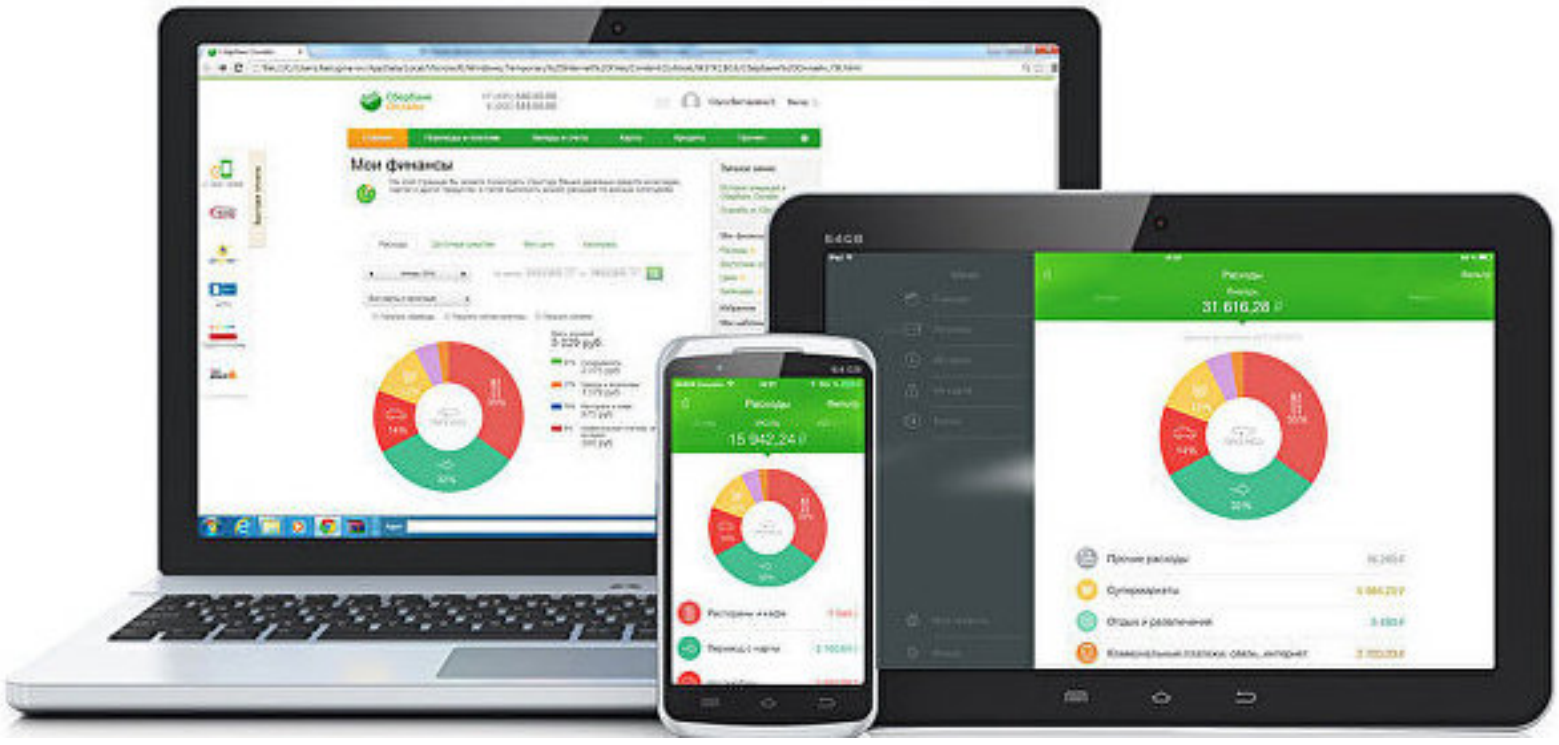
Машинное обучение в банках



Задачи машинного обучения:

- кредитный скоринг
- предсказание следующих транзакций пользователей
- кластеризация пользователей
- предсказание спроса на наличные деньги в банкоматах

Машинное обучение в банках



Задачи программирования:

- поддержка работы с базой клиентов банка
- написание программного обеспечения для банкоматов, онлайн-банков, создание приложений
- обеспечение информационной безопсности