

Данные в машинном обучении

Схема работы машинного обучения

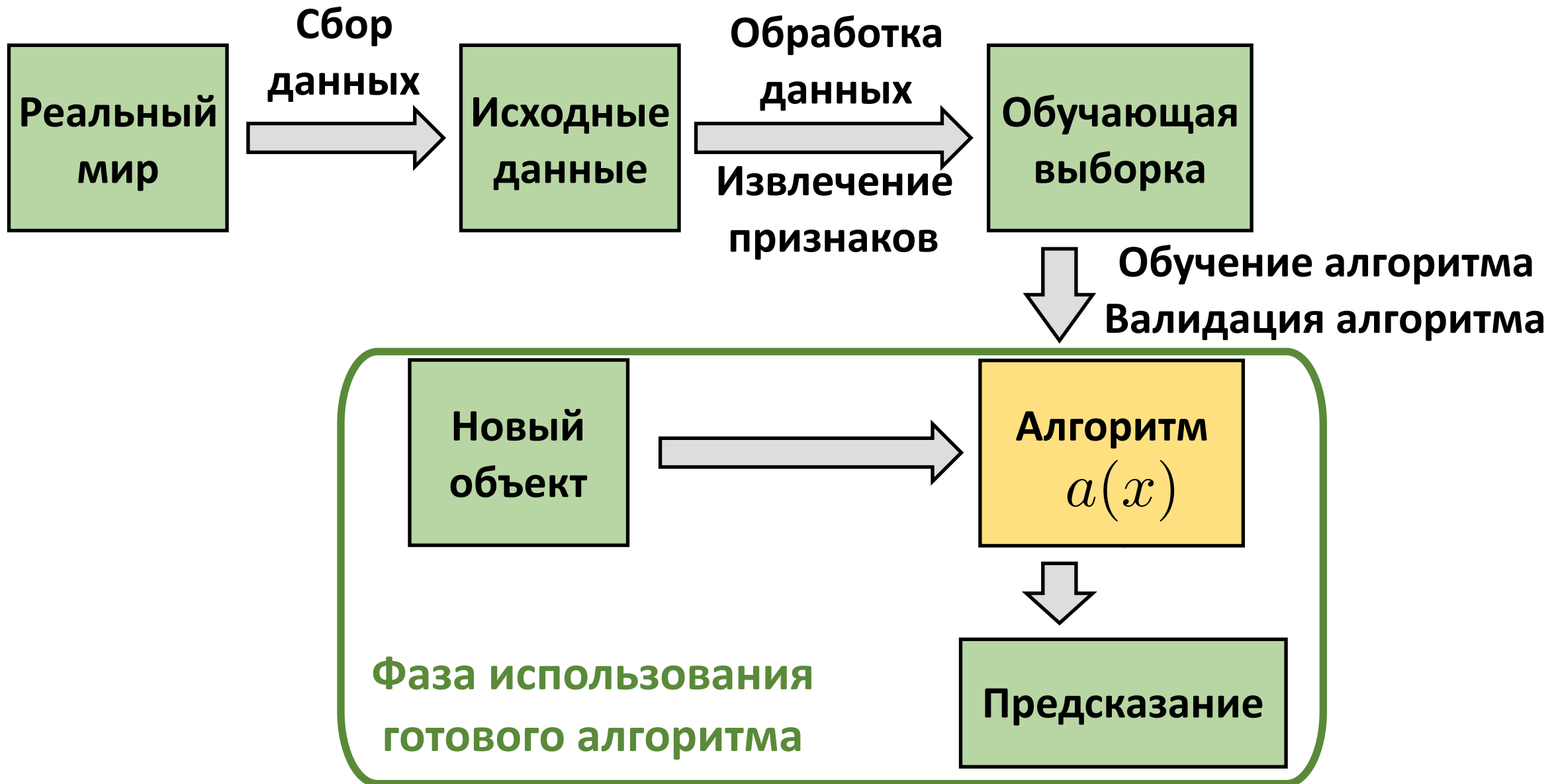
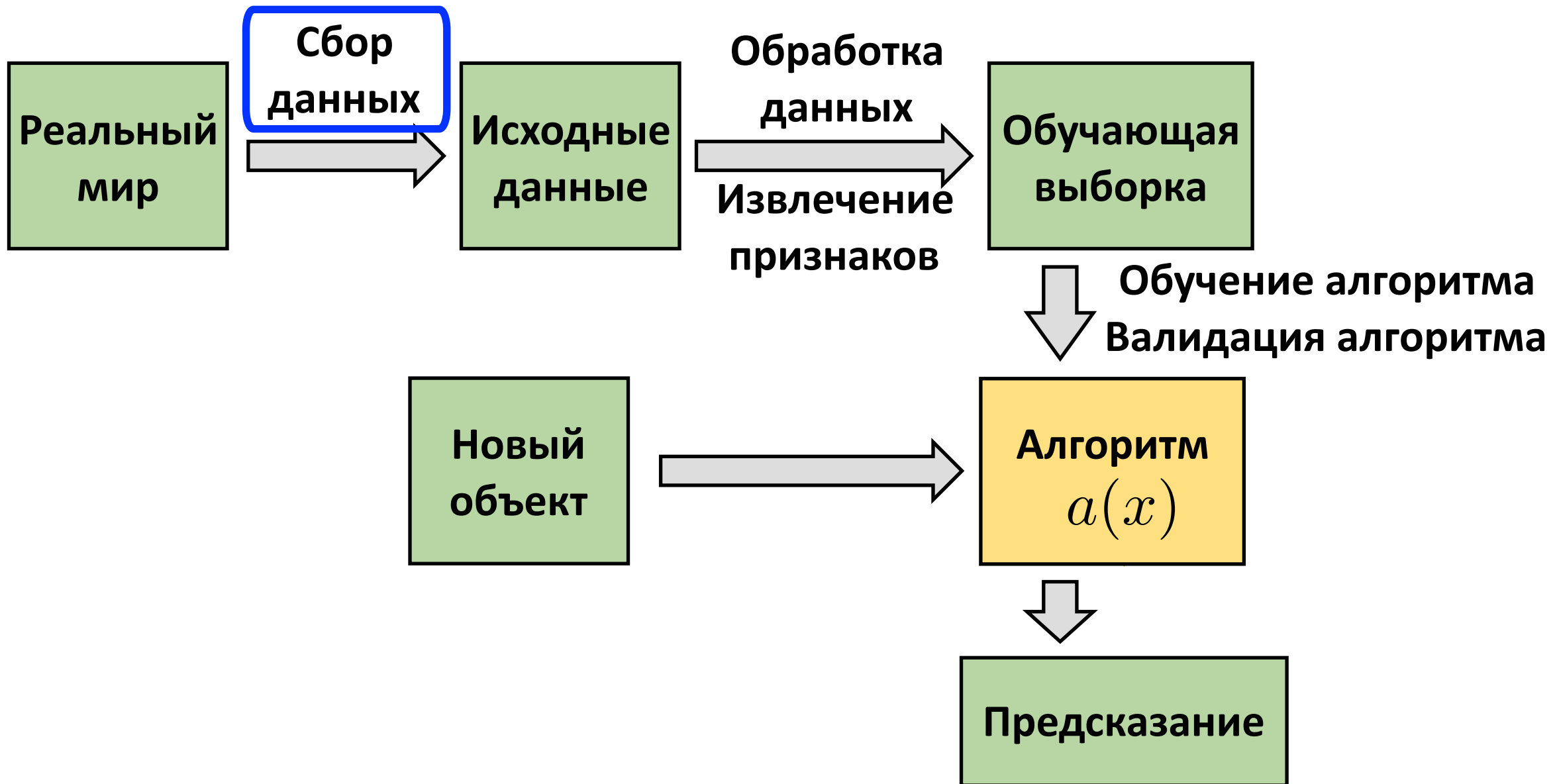


Схема работы машинного обучения



Сбор данных в машинном обучении

Обучающая выборка

Обучающая выборка

объект
 x_i

Площадь	Год постройки	Число комнат	Цена
45	1995	1	7000000
60	2005	2	9900000
35	2010	1	5500000

ответ
 y_i

Источники данных

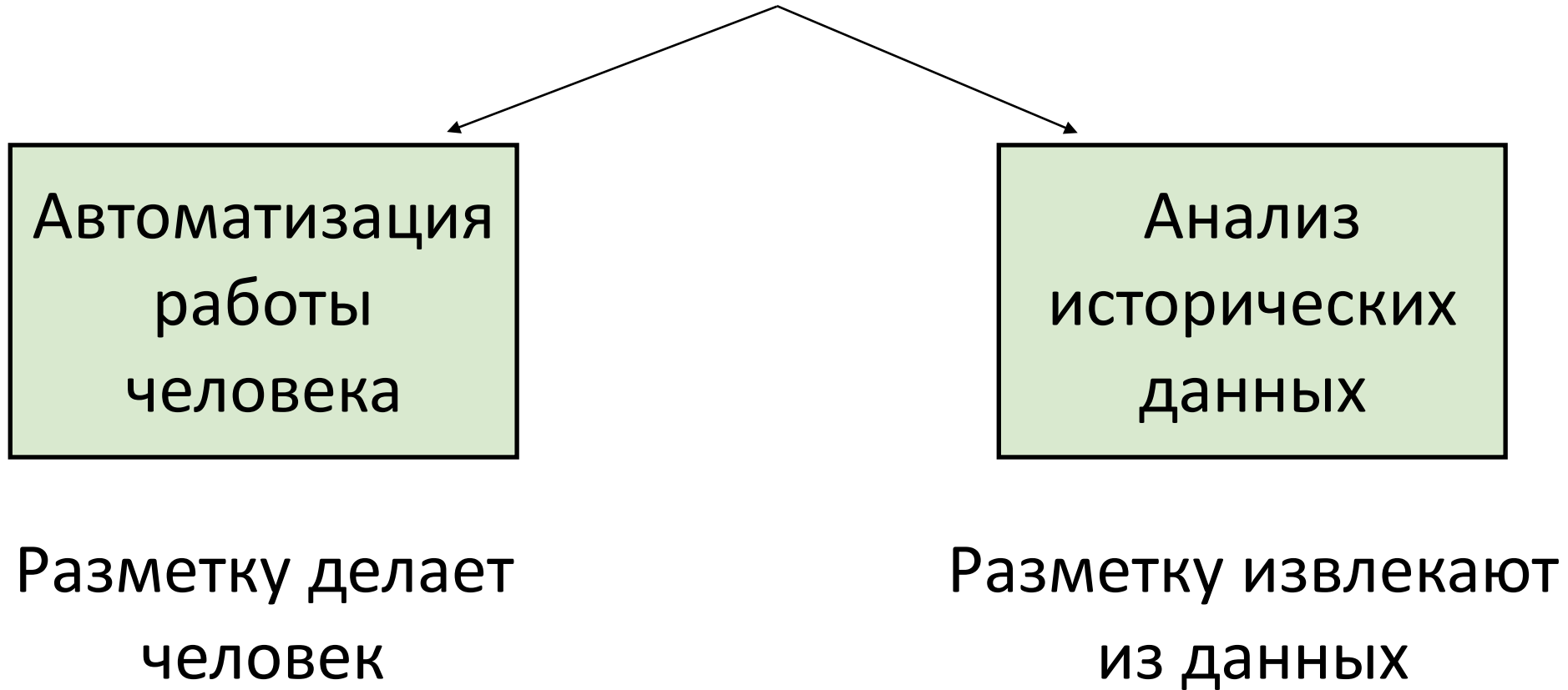
- API социальных сетей (Twitter, Facebook, VK), карт...
- Архивы крупных сайтов (Wikipedia dump)
- Открытые наборы данных (Amazon reviews, The Reuters Corpus ...)
- Соревнования по анализу данных (например, kaggle.com)
- Порталы открытых данных (например, data.gov.ru)
- Краулеры Интернета

Разметка в машинном обучении

- Обучение с учителем — построение моделей на основе **примеров**
- Пример — объект и ответ на нём
- Где брать ответы?

Источники разметки

Для чего используют машинное обучение?



Предсказание дефолта

- Задача: оценить шансы дефолта у клиента
- Объект: клиент банка
- Ответ: вернёт ли кредит?

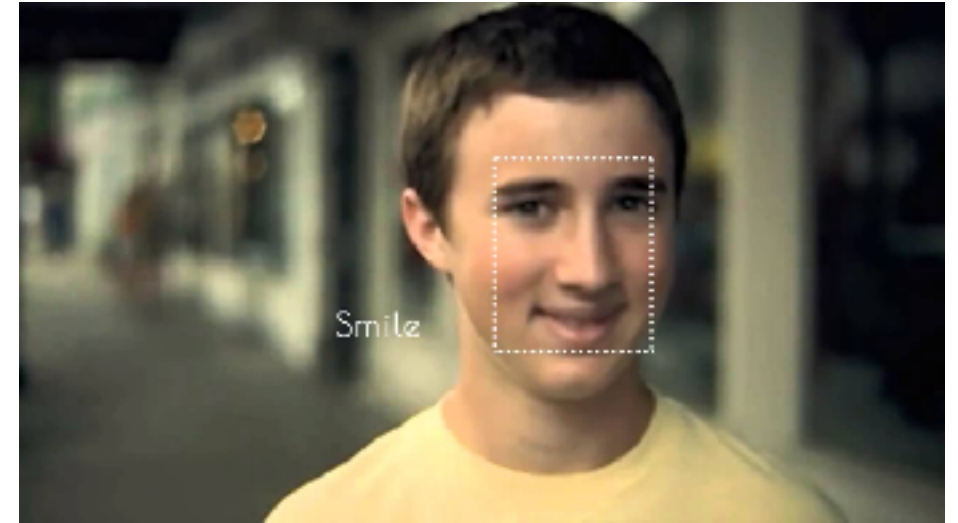
- Источники разметки?
- Потенциальные проблемы в разметке?

Рекомендательные системы

- Задача: выбрать для пользователя музыку, которая ему понравится
- Объект: пара «пользователь-композиция»
- Что является ответом?
- Источники разметки?

Определение эмоции

- Задача: определить, улыбается ли человек на фото
- Объект: фотография
- Ответ: есть ли улыбка?
- Источники разметки?



Краудсорсинг

- Примеры: Amazon Mechanical Turk, Яндекс.Толока

<p>Модерация новостных комментариев</p> <p>★★★★★ задания с обучением</p> <p>Скажите, соответствует ли комментарий правилам модерации.</p> <p>Инструкция</p>	<p>0,01 \$ за задание ≈ 0,84 \$ максимум</p> <p>Обучение</p>
<p>Расстановка ударений в выделенных словах</p> <p>★★★★★ задания с обучением • 18+</p> <p>Расставьте ударения и точки над "ё"</p> <p>Инструкция</p>	<p>0,02 \$ за задание ≈ 0,02 \$ максимум</p> <p>Обучение</p>
<p>DoD: попарное сравнение мобильных сайтов</p> <p>★★★★★ задания с обучением • 18+</p> <p>Попарное сравнение документов с высокой оплатой за качественные оценки. После прохождения экзамена...</p> <p>Инструкция</p>	<p>0,05 \$ за задание ≈ 8,24 \$ максимум</p> <p>0,01–1,00 \$ зависит от навыка «Оплата в проектах DoD»</p> <p>Обучение</p>

Краудсорсинг

- Примеры: Amazon Mechanical Turk, Яндекс.Толока
- Простые задания
- Потенциальные проблемы:
 - Неверные ответы из-за сложности заданий и низкой цены
 - Вредительство
 - Использование ботов для разметки

Задача ImageNet (2012)

- Задача: картинка → один из 1000 классов



- Обучили глубокую нейросеть на **1.2 млн** картинок:
 - Уменьшили ошибку с **25%** до **12%** (в 2012)
 - В 2018 ошибка **3%** (один ученый прошел сам, у него **5%**)

Схема работы машинного обучения

