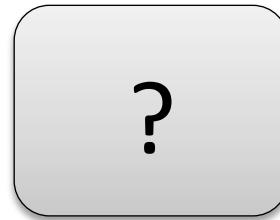


# Введение в машинное обучение

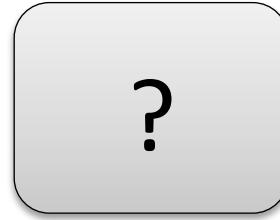
Илья Лысенков, Itseez3D

ННГУ, 2016

# Задача: на фотографии изображено лицо или нет?



“лицо”



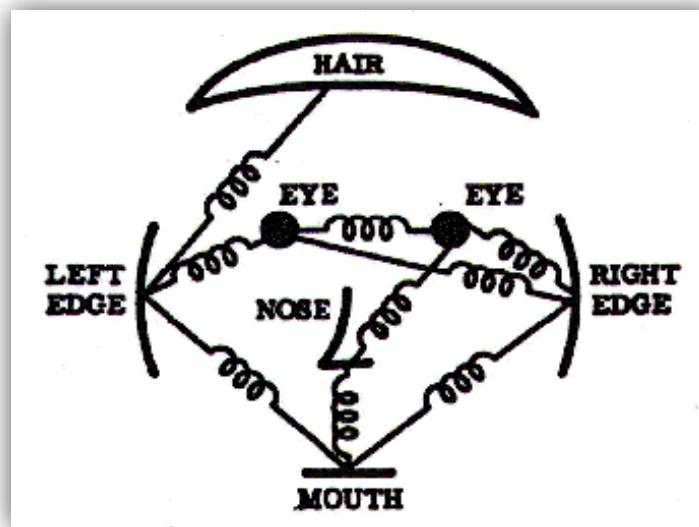
“нет”

Images credit:

[http://magazine.foxnews.com/sites/magazine.foxnews.com/files/Coffee\\_0.jpg](http://magazine.foxnews.com/sites/magazine.foxnews.com/files/Coffee_0.jpg)

[http://www.uni-regensburg.de/Fakultaeten/phil\\_Fak\\_II/Psychologie/Psy\\_II/beautycheck/english/durchschnittsgesichter/m\(01-32\).jpg](http://www.uni-regensburg.de/Fakultaeten/phil_Fak_II/Psychologie/Psy_II/beautycheck/english/durchschnittsgesichter/m(01-32).jpg)

# Прямой подход



“Лицо”

Images credit:

<https://people.csail.mit.edu/fergus/iccv2005/face.gif>

[http://www.uni-regensburg.de/Fakultaeten/phil\\_Fak\\_II/Psychologie/Psy\\_II/beautycheck/english/durchschnittsgesichter/m\(01-32\).jpg](http://www.uni-regensburg.de/Fakultaeten/phil_Fak_II/Psychologie/Psy_II/beautycheck/english/durchschnittsgesichter/m(01-32).jpg)

# Примеры реальных изображений



Image credit:

<http://vis-www.cs.umass.edu/lfw/>

# Машинное обучение

## Лица



## Всё остальное



Images credit:  
[http://stephencotterell.com/wp-content/uploads/2012/05/people-picture-pile-1024x768\(pp\\_w900\\_h675\).jpg](http://stephencotterell.com/wp-content/uploads/2012/05/people-picture-pile-1024x768(pp_w900_h675).jpg)  
<http://www.gratisfaction.co.uk/wp-content/uploads/2014/07/100-FREE-Photo-Prints-From-Jessops-Using-Code-MSEJP100-Gratisfaction-UK-Freebies.jpg>

# Машинное обучение

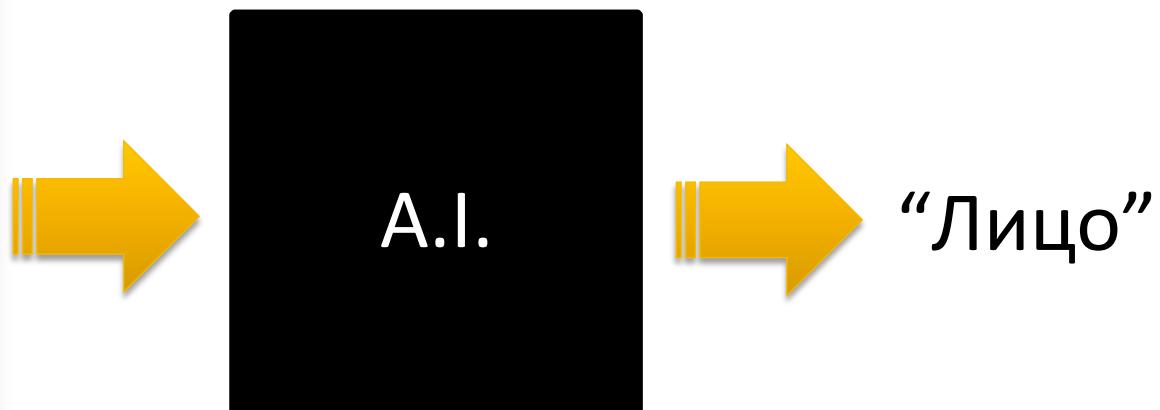
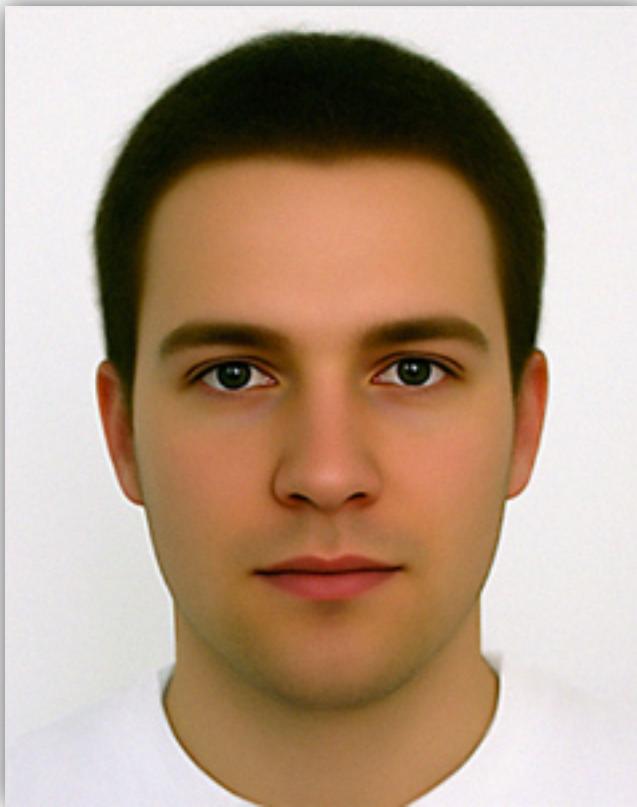
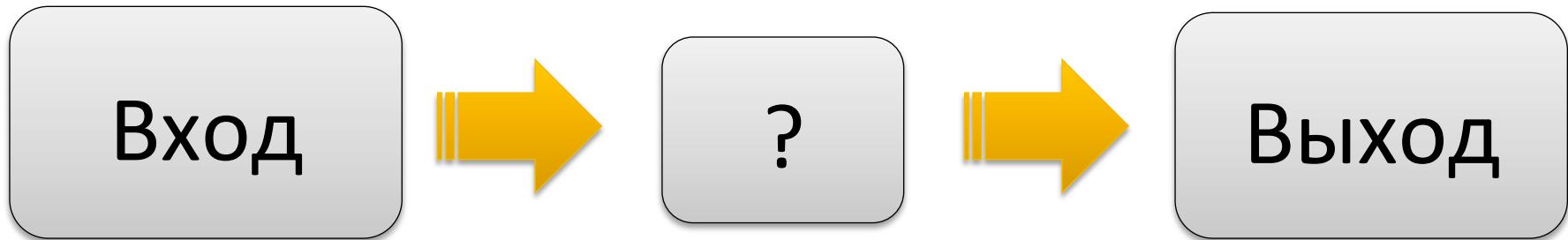


Image credit:

[http://www.uni-regensburg.de/Fakultaeten/phil\\_Fak\\_II/Psychologie/Psy\\_II/beautycheck/english/durchschnittsgesichter/m\(01-32\).jpg](http://www.uni-regensburg.de/Fakultaeten/phil_Fak_II/Psychologie/Psy_II/beautycheck/english/durchschnittsgesichter/m(01-32).jpg)

# Общий случай



# Формализация

- $x$  – вход,  $y$  – выход
- $y = f(x)$

Постановка задачи машинного обучения

Дана обучающая выборка:      Найти:

$$(x_1, y_1)$$

$$f(x)$$

$$(x_2, y_2)$$

...

$$(x_N, y_N)$$

# Формализация

$x$  – вектор, задаётся набором признаков:

$$x = (x^{(1)}, x^{(2)}, \dots, x^{(d)})$$

$x^{(j)} \in \mathbb{R}$  – количественный признак

$x^{(j)} \in \{0, 1, \dots, k - 1\}$  – категориальный признак

$y$  – число:

$y \in \{0, 1, \dots, k - 1\}$  – задача классификации

$y \in \mathbb{R}$  – задача регрессии

# Пример: спам-фильтр

Вход: e-mail

Выход: спам или нет

Задача бинарной классификации:  $y \in \{0, 1\}$

0 – нормальное письмо

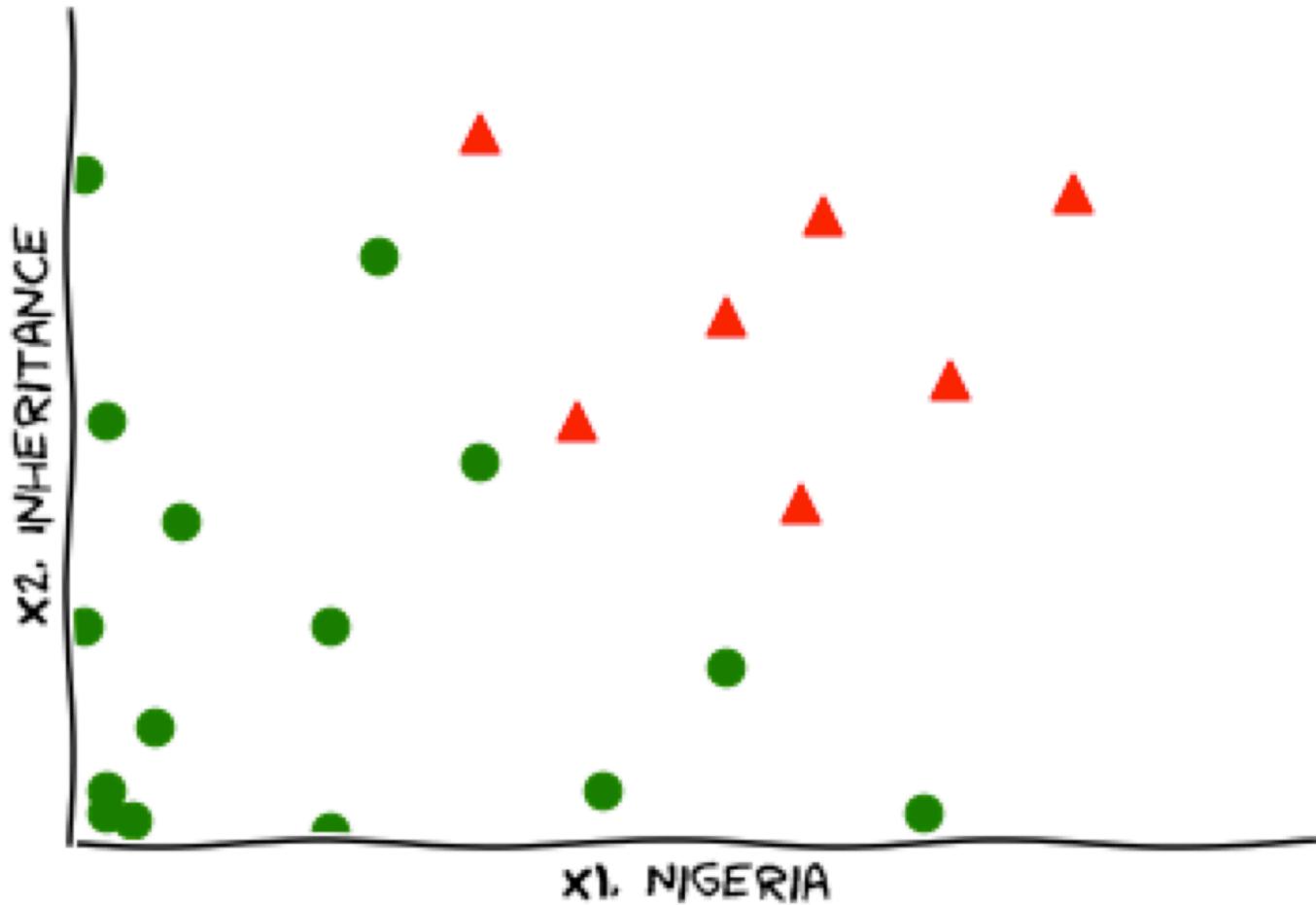
1 – спам

$x$  – частоты вхождения слов:

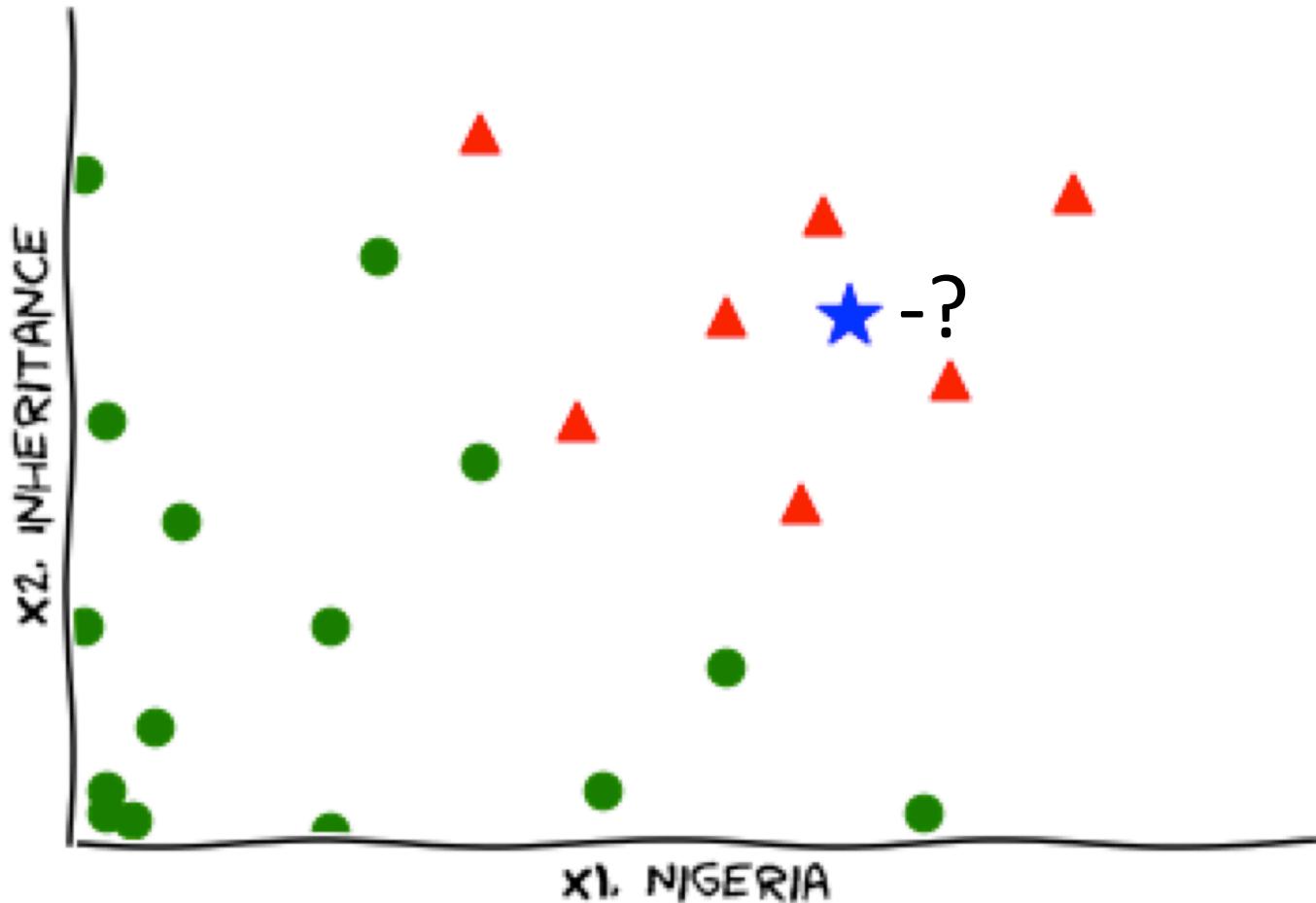
$x_0 = (0.001, 0.0003, 0.002, \dots, 0.0001)$

Категориальный признак: домен отправителя

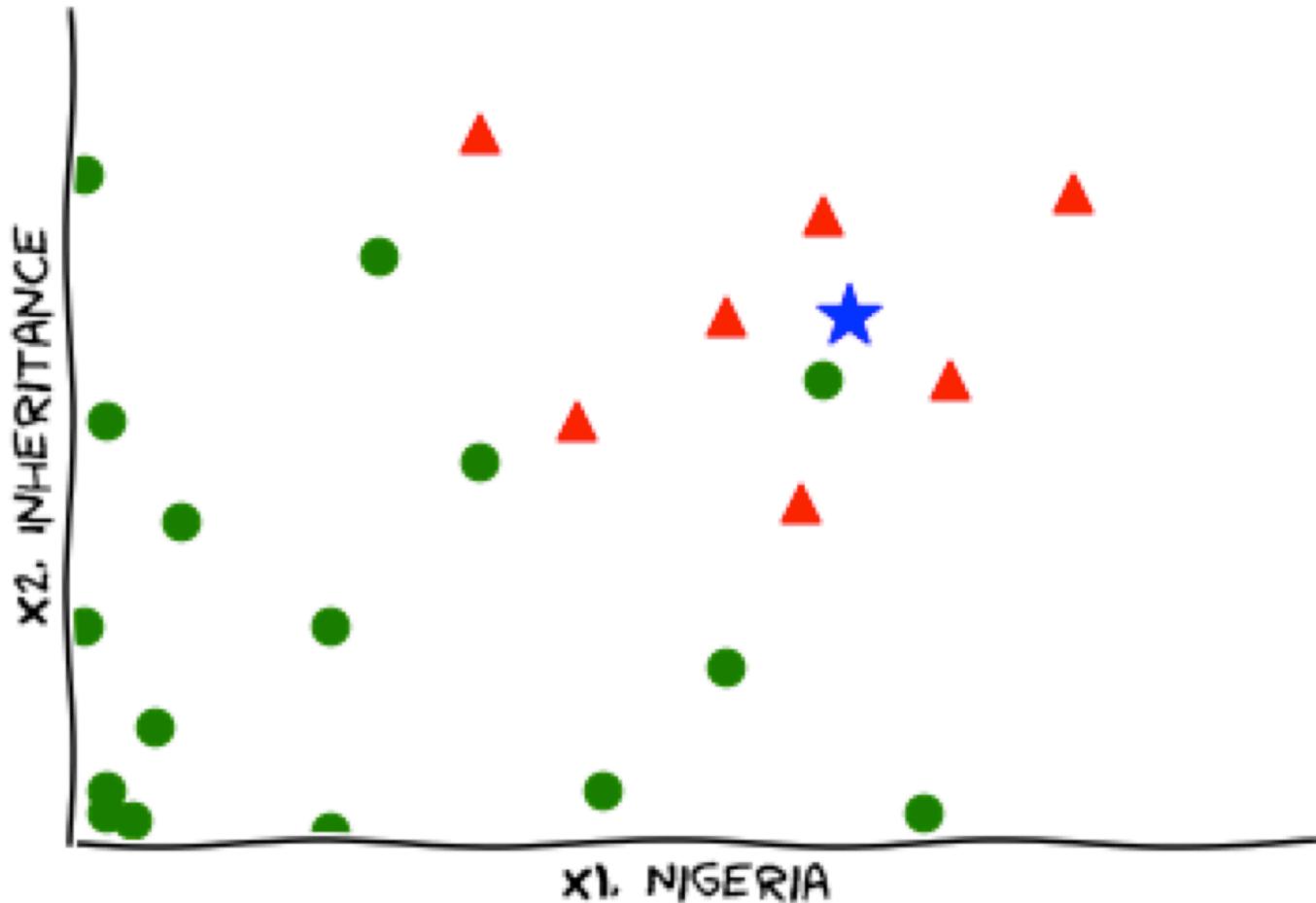
# Пример: спам-фильтр



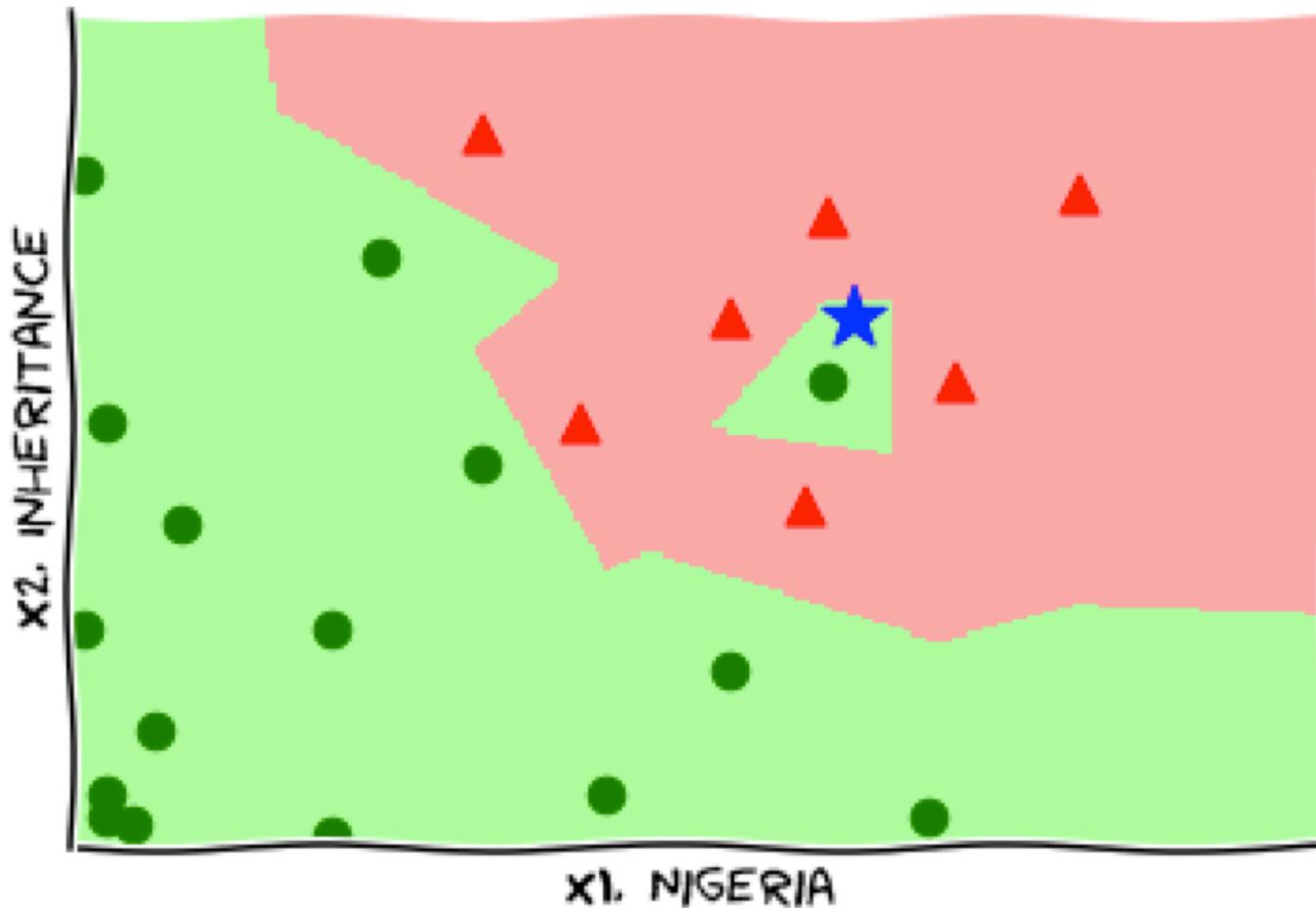
# Пример: спам-фильтр



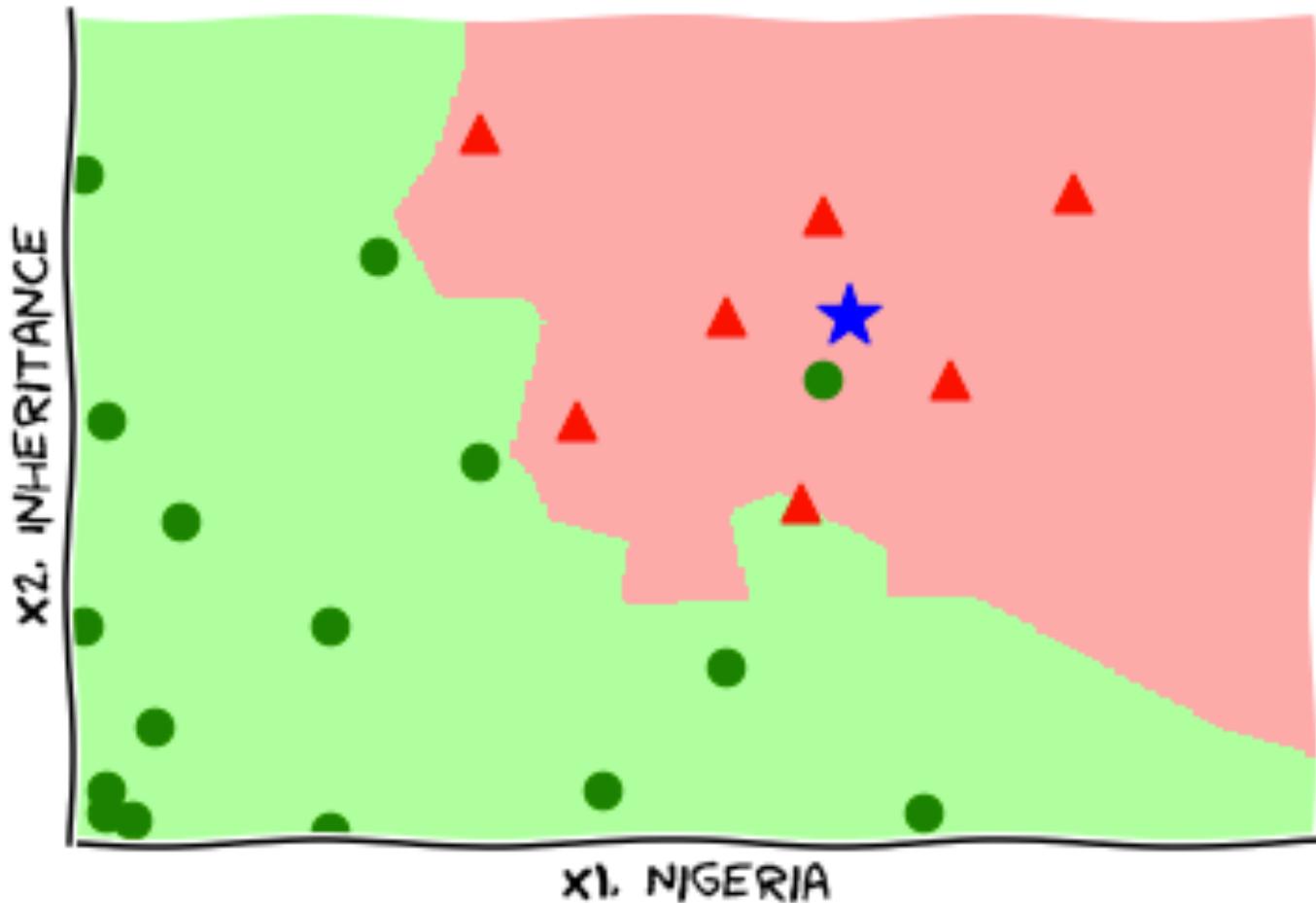
# Пример: спам-фильтр



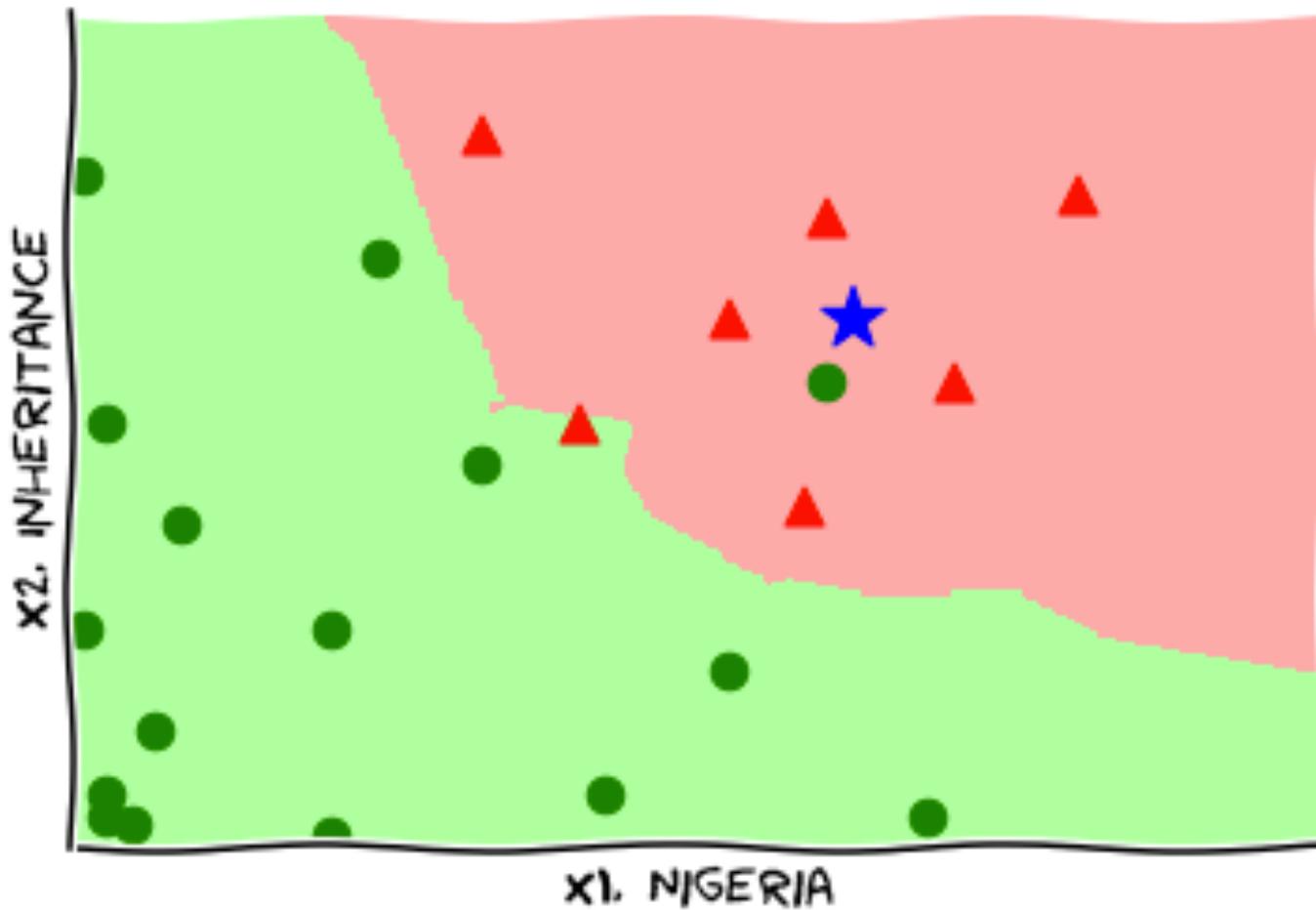
# 1 ближайший сосед



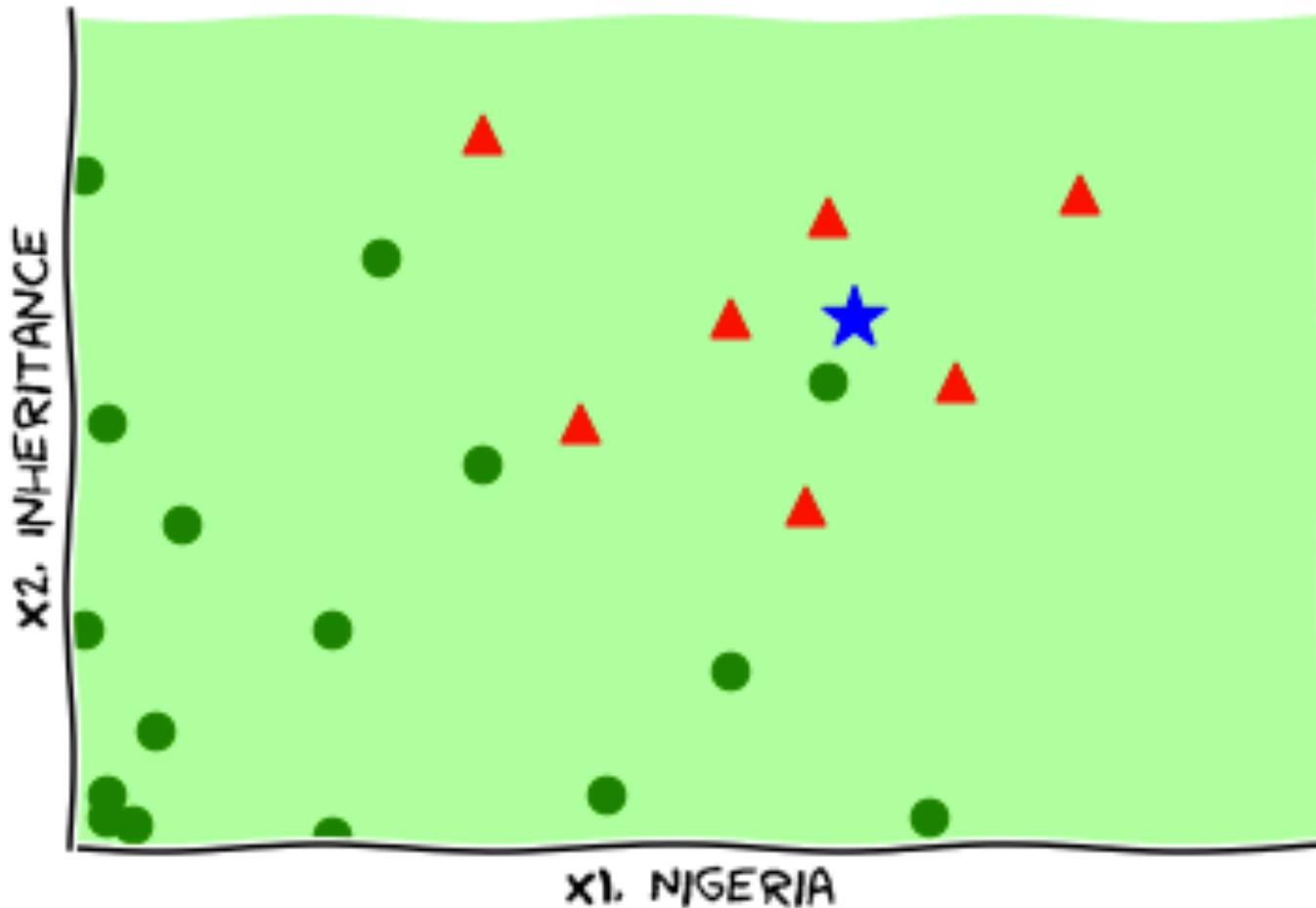
# 3 ближайших соседа



# 7 ближайших соседей



# 23 ближайших соседа

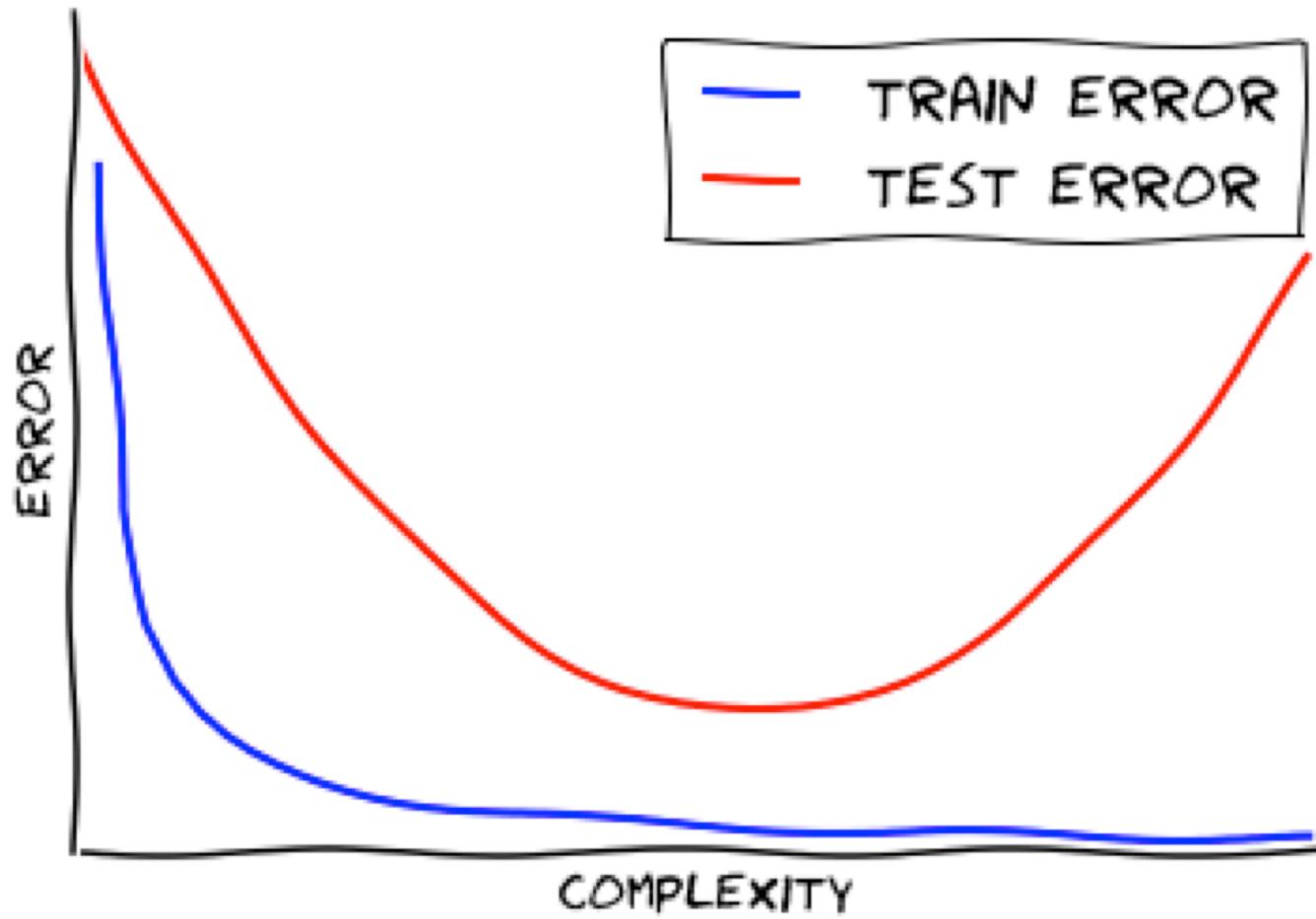


# Как выбрать наилучший вариант?

Оценка качества работы алгоритма:

- Ошибка на обучающей выборке
- Ошибка на новых данных (*тестовой выборке*)
- Кросс-валидация (скользящий контроль)

# Переобучение



# Наиболее популярные алгоритмы

- Машина опорных векторов (Support Vector Machine, SVM)
- Дерево решений (Decision Tree)
- Ансамбль деревьев решений (Random Forest)
- Бустинг (AdaBoost, Gradient Boosting)
- Нейронная сеть
- Байесовский классификатор

# Регрессия

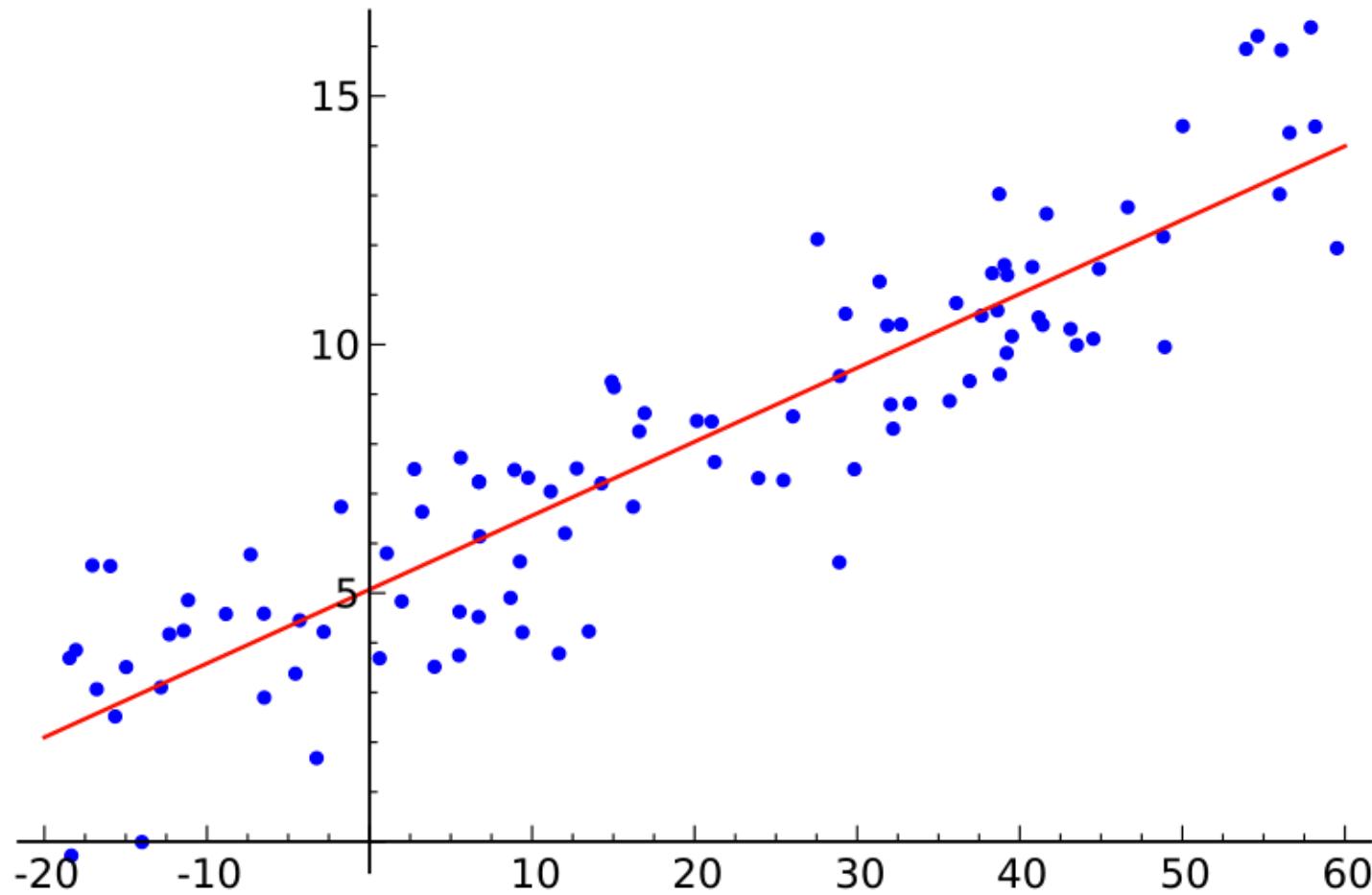


Image credit:

[https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear\\_regression.svg/800px-Linear\\_regression.svg.png](https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear_regression.svg/800px-Linear_regression.svg.png)

# Кластеризация

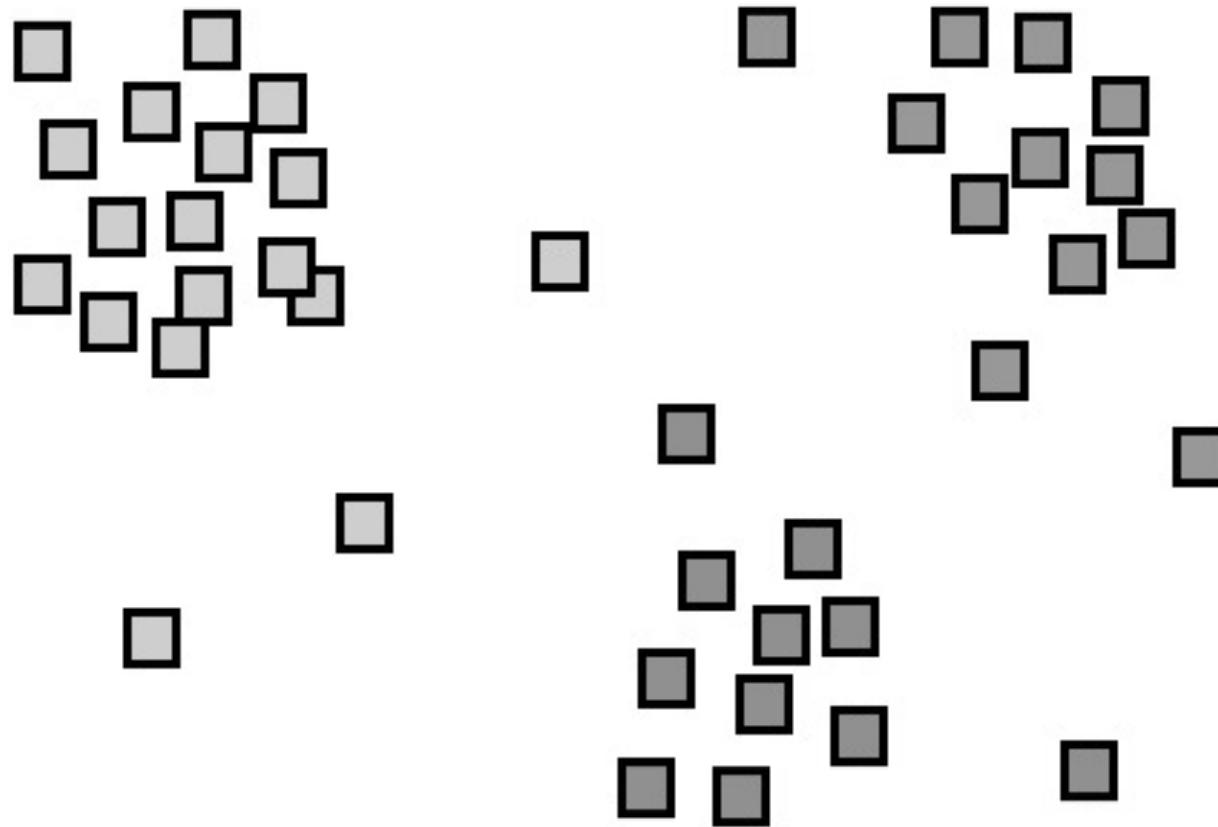


Image credit:  
<https://upload.wikimedia.org/wikipedia/commons/4/41/Cluster-2.png>

# Кластеризация

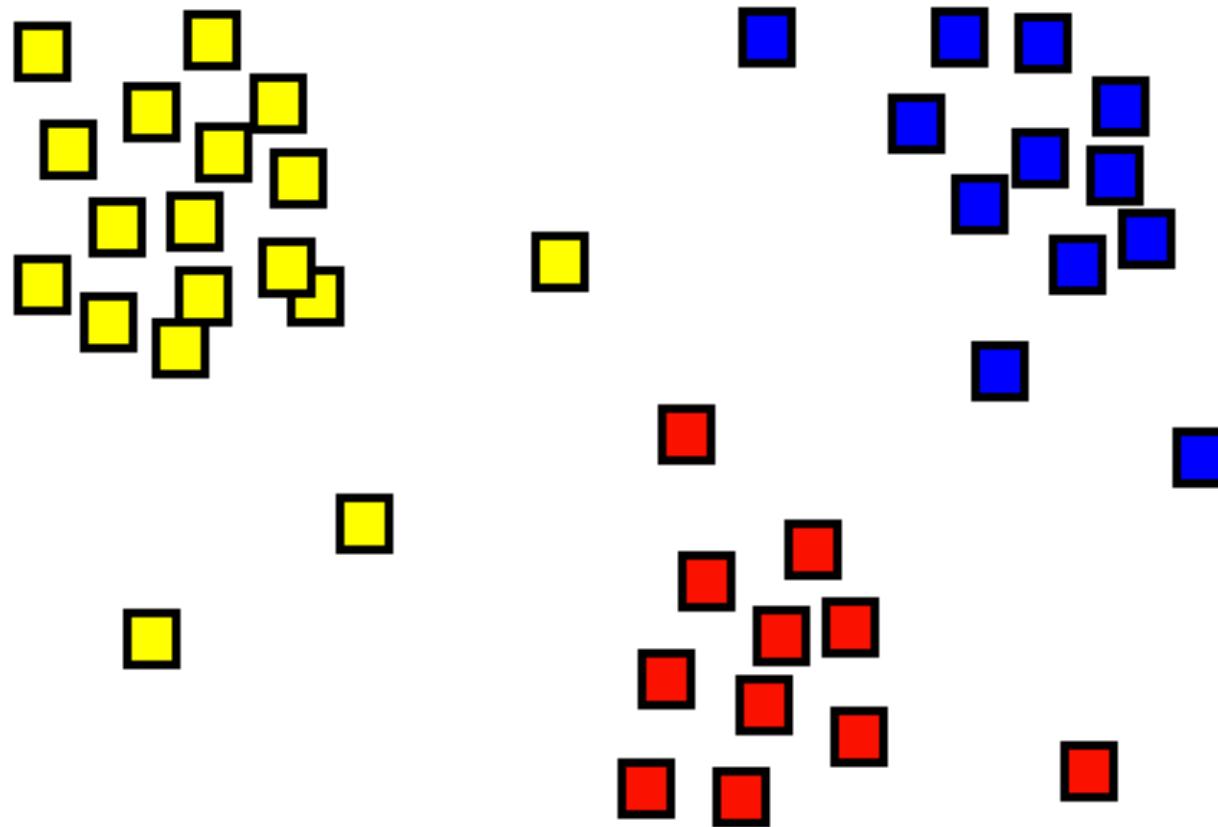


Image credit:  
<https://upload.wikimedia.org/wikipedia/commons/4/41/Cluster-2.png>

# Классификация изображений

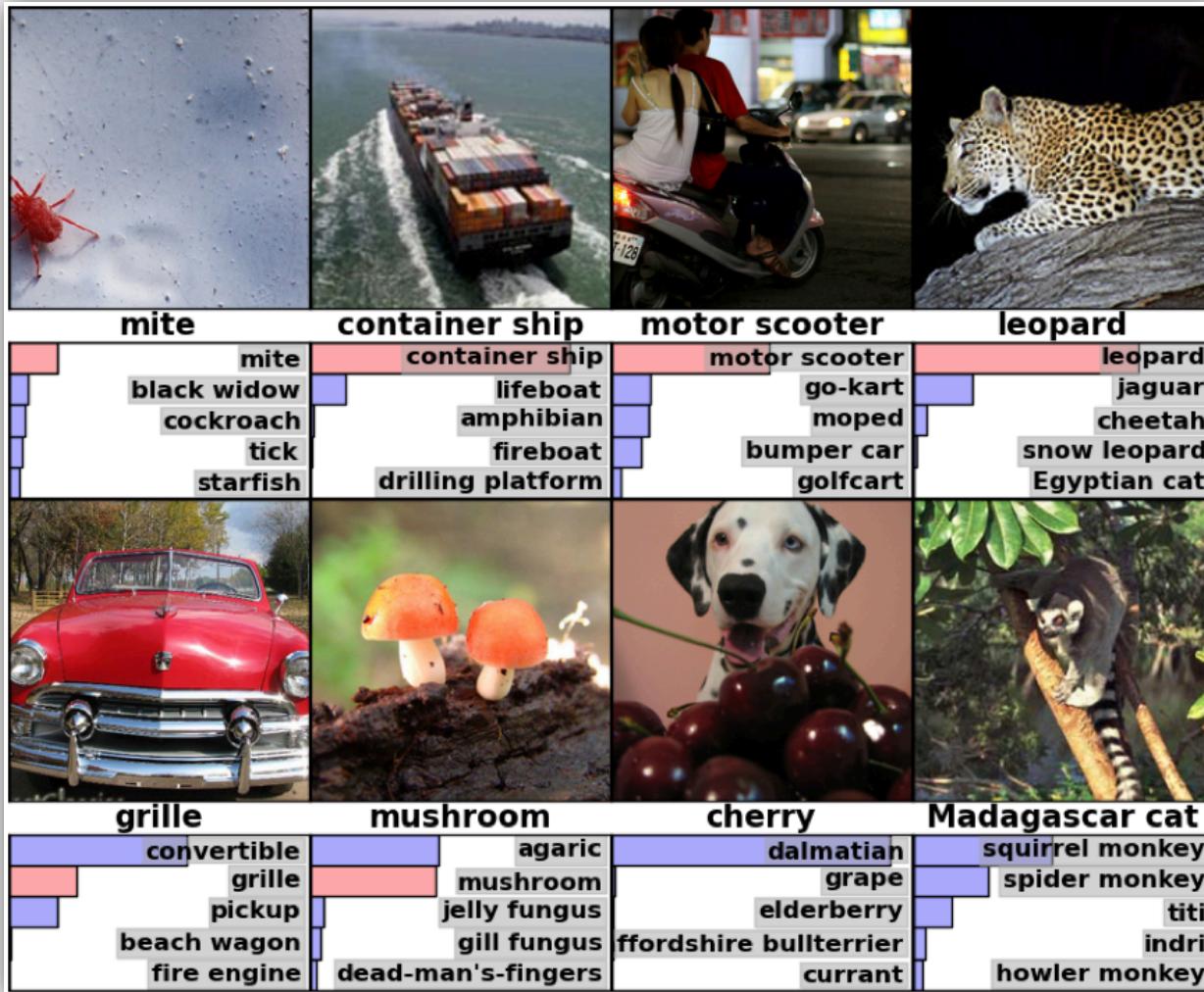


Image credit:

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, NIPS 2012

# Классификация рукописных цифр



Image credit:

[http://neuralnetworksanddeeplearning.com/images/mnist\\_100\\_digits.png](http://neuralnetworksanddeeplearning.com/images/mnist_100_digits.png)

# Детектирование лиц

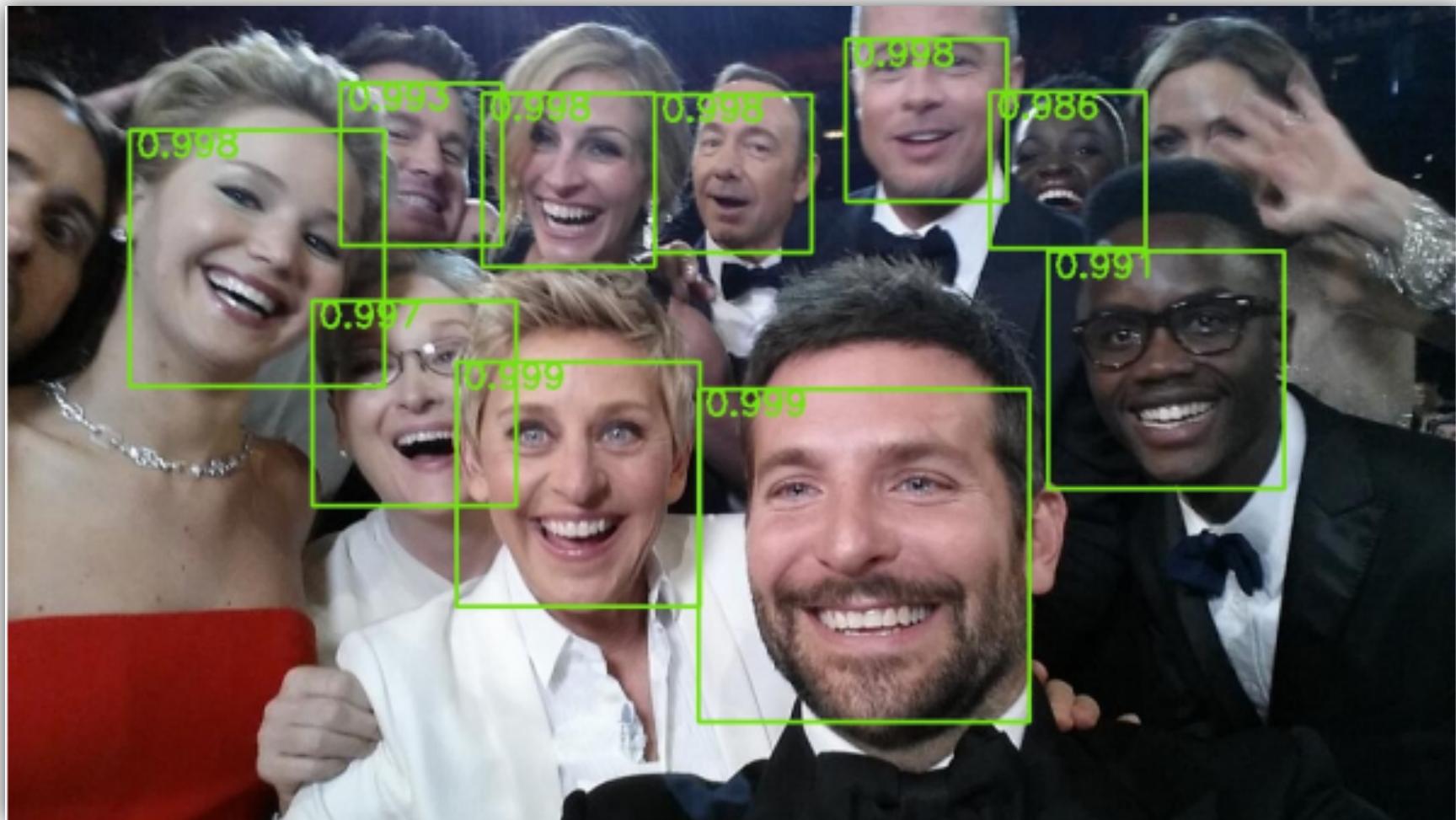


Image credit:

<http://www.technologyreview.com/sites/default/files/images/Face%20detection.png>

# Детектирование пешеходов

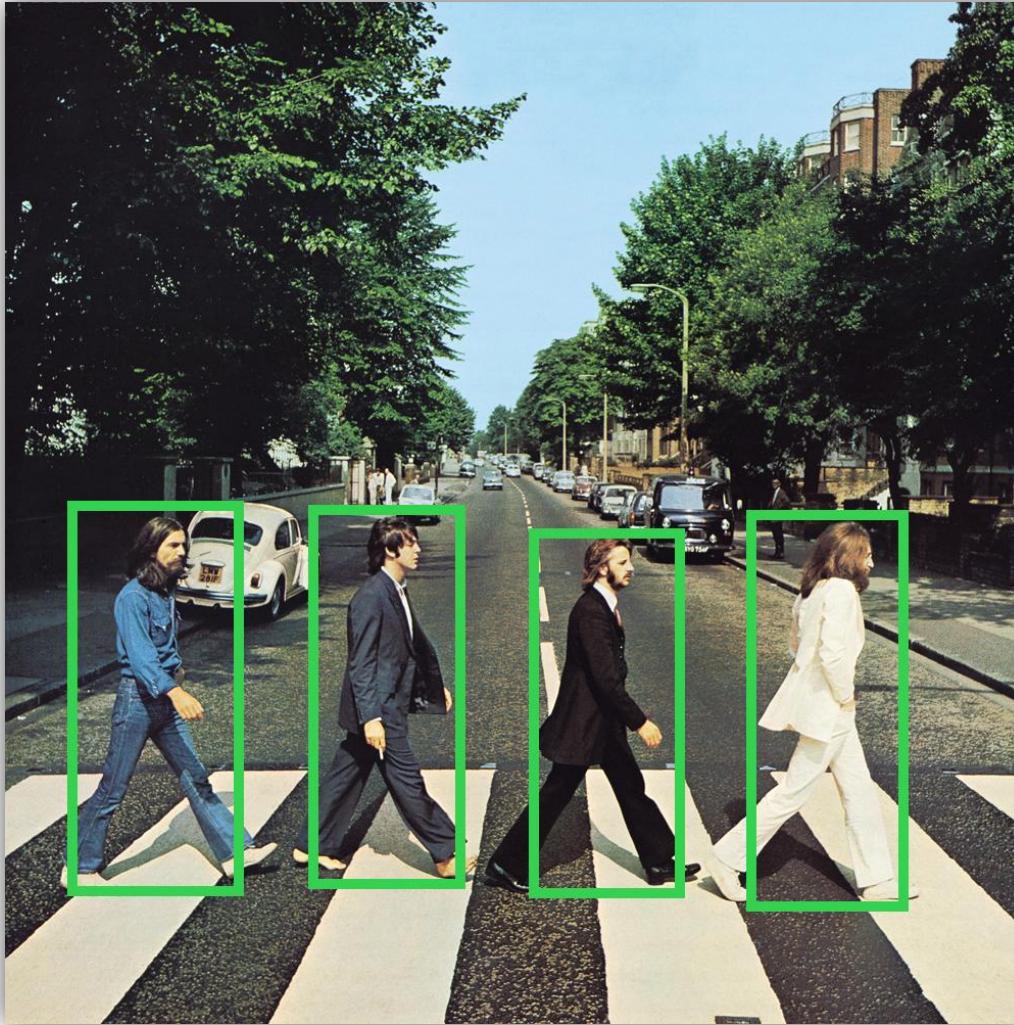


Image credit:

[https://d817ypd61vbww.cloudfront.net/sites/default/files/styles/media\\_responsive\\_widest/public/tile/image/original\\_593.jpg?itok=lhwAxXPe](https://d817ypd61vbww.cloudfront.net/sites/default/files/styles/media_responsive_widest/public/tile/image/original_593.jpg?itok=lhwAxXPe)

# Детектирование объектов

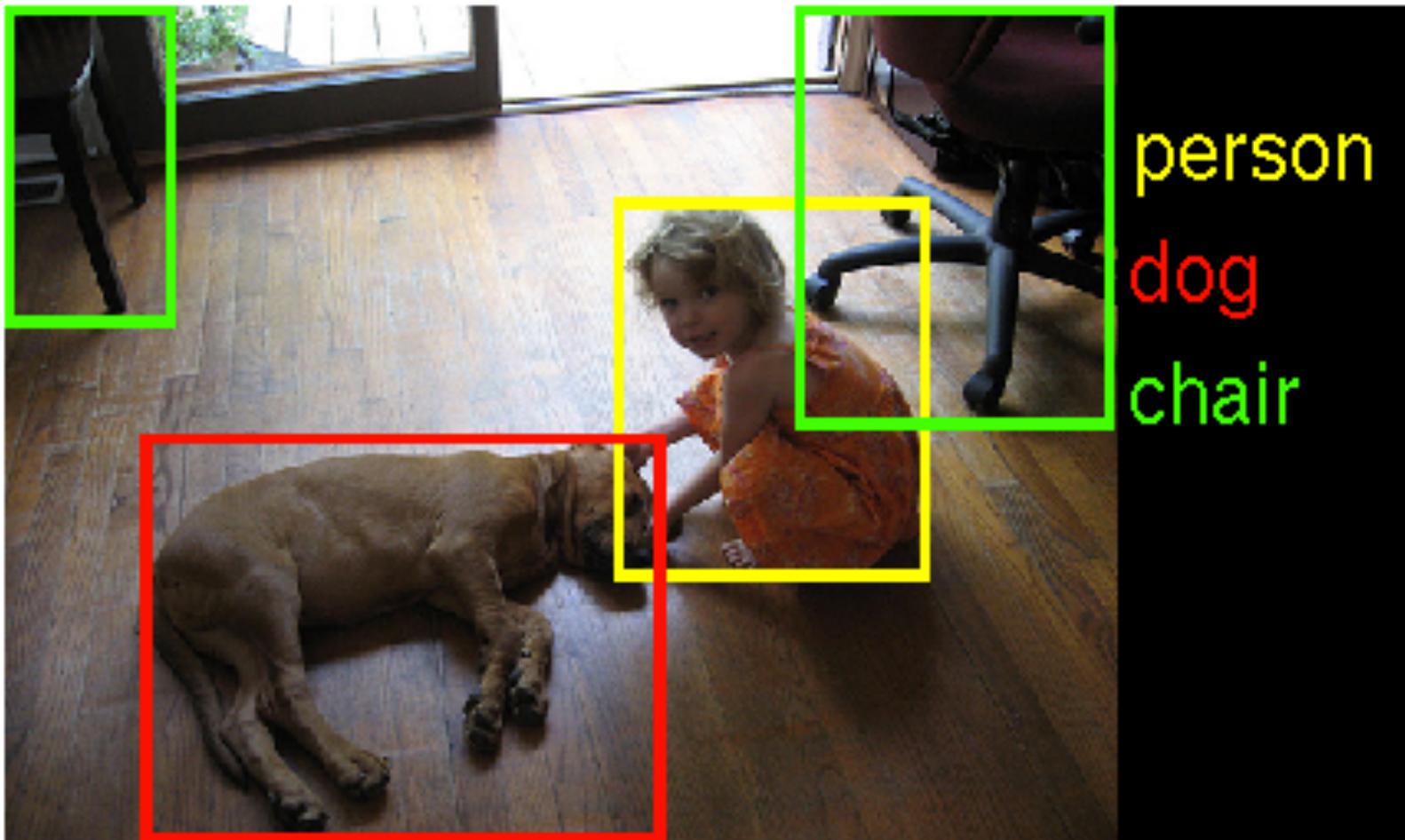


Image credit:

[http://www.image-net.org/challenges/LSVRC/2014/ILSVRC2012\\_val\\_00042692.png](http://www.image-net.org/challenges/LSVRC/2014/ILSVRC2012_val_00042692.png)

# Определение пола и возраста ([www.how-old.net](http://www.how-old.net))

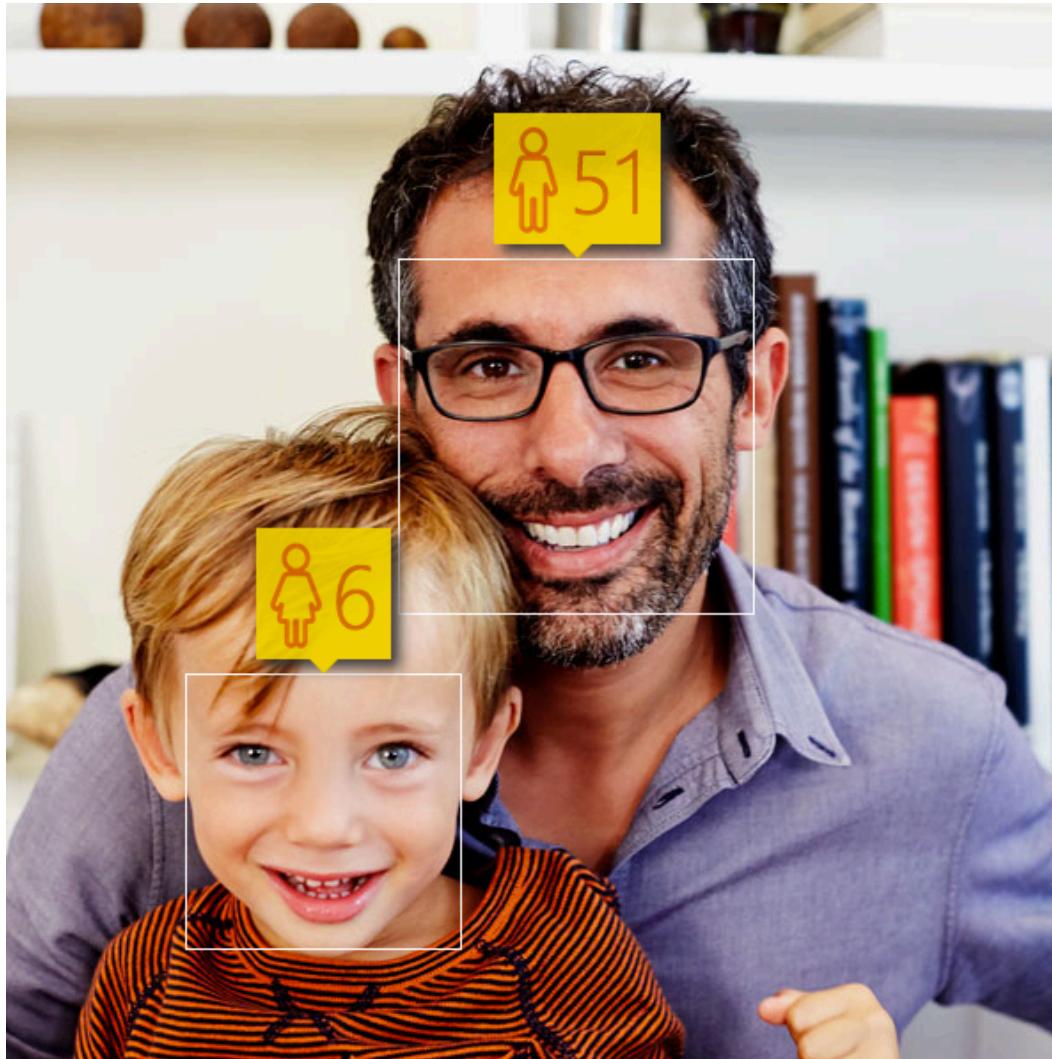
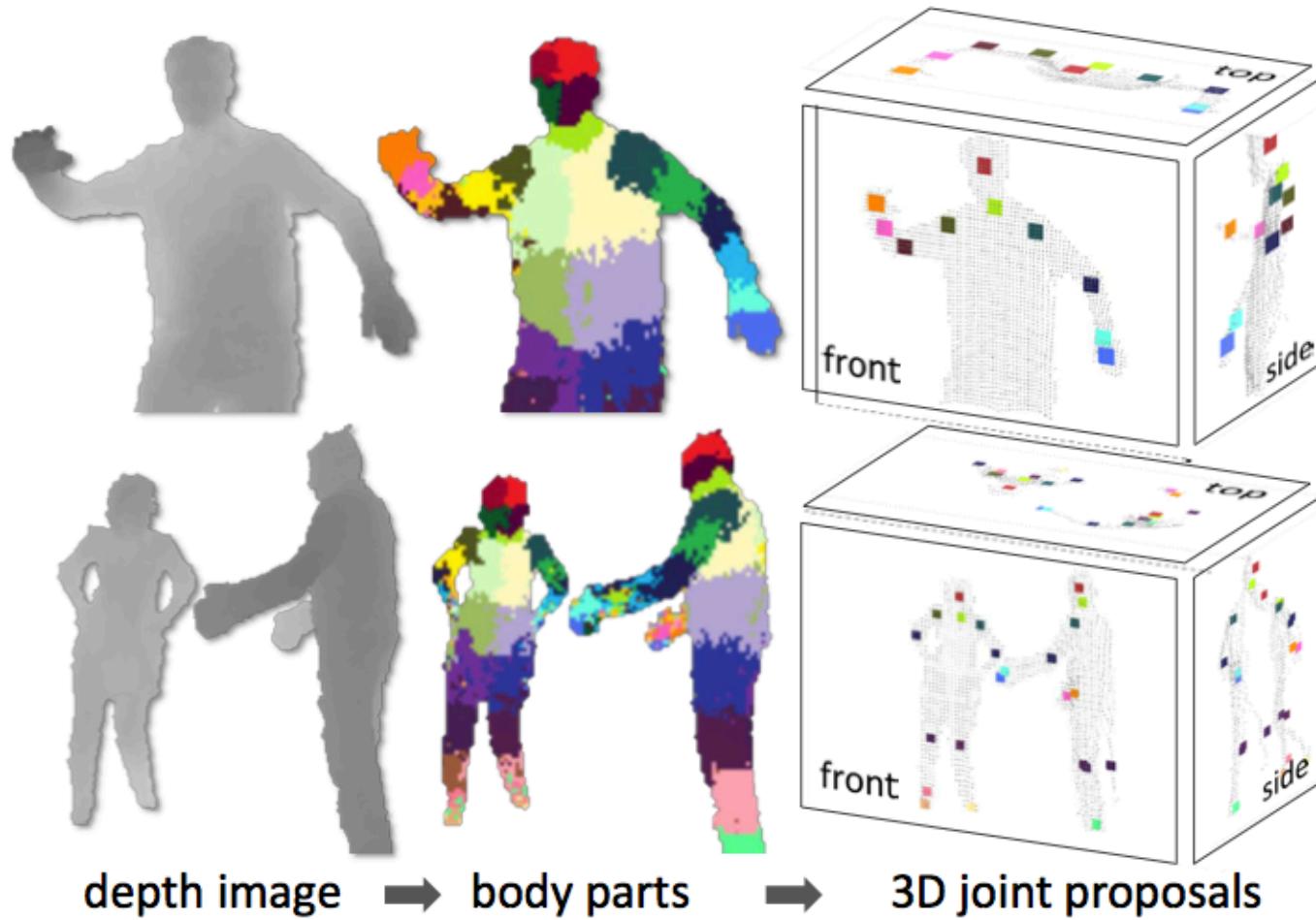


Image credit:  
<https://how-old.net/>

# Кинект: оценка позы человека



Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., ... & Moore, R. (2013).

Real-time human pose recognition in parts from single depth images., CVPR 2013

Image credit:

# Медицинская диагностика

- Вход: пациент
- Выход: диагноз или курс лечения  
(задача классификации)
- Признаки:
  - Качественные:  
температура, давление, ...
  - Категориальные:  
желтушность, сыпь, группа крови, ...

# Кредитный скоринг

- Вход: клиент
- Выход: давать кредит или нет  
(задача бинарной классификации)
- Признаки:
  - Количественные:  
зарплата, возраст, сумма кредита, ...
  - Категориальные:  
пол, образование, ...

# Рекомендательные системы

- Вход: пары пользователь-фильм
- Выход: оценка пользователя за фильм  
(задача регрессии)

Пример:

Конкурс Netflix

Октябрь 2006 – Сентябрь 2009

Данные:

~500 000 пользователей

~20 000 фильмов

~100 000 000 рейтингов

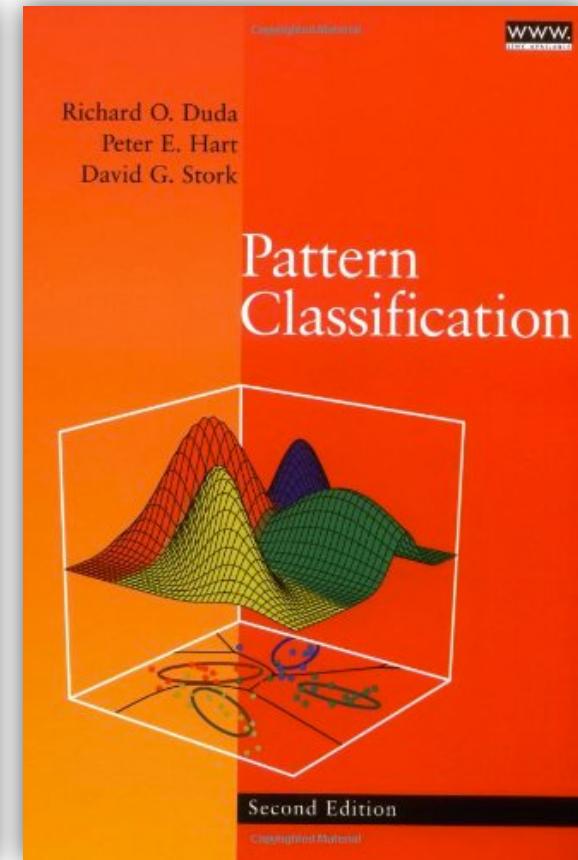
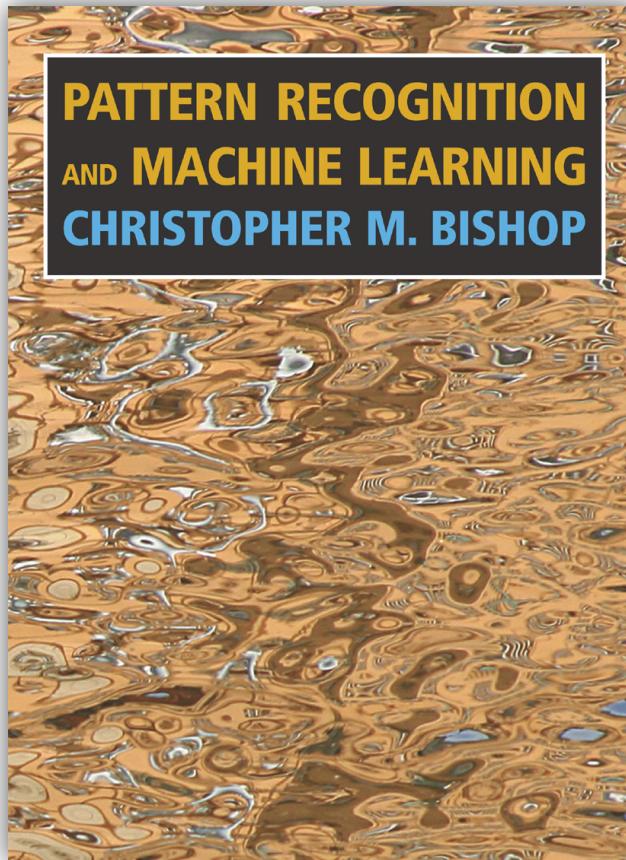
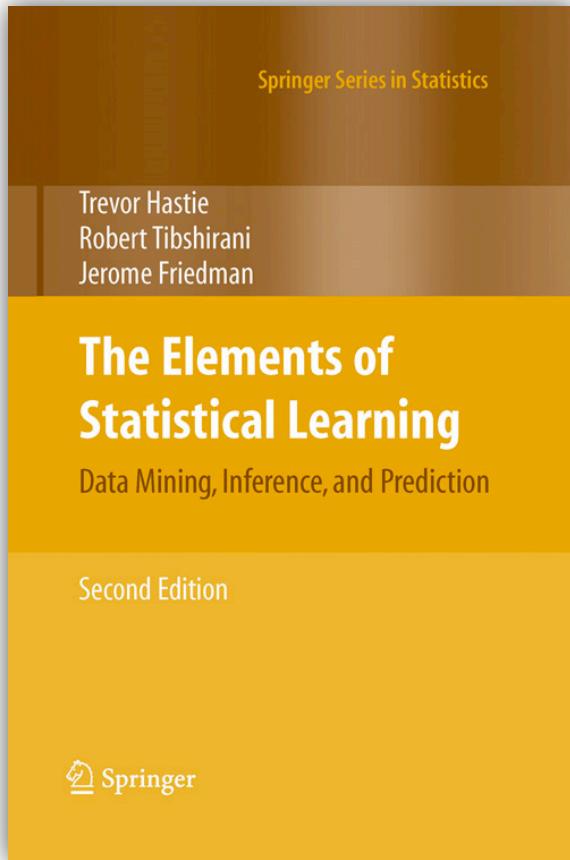
# Классификация текстов

- Вход: текст
- Выход: категория текста  
(задача классификации)
- Признаки:
  - Качественные:  
Частоты слов в тексте, заголовке, ...
  - Категориальные:  
Автор, издание, ...

# Определение стоимости недвижимости

- Вход: квартира
- Выход: стоимость  
(задача регрессии)
- Признаки:
  - Качественные:  
Площадь, этаж, площадь кухни, ...
  - Категориальные:  
Наличие балкона, тип дома, ...

# КНИГИ



[http://statweb.stanford.edu/  
~tibs/ElemStatLearn/](http://statweb.stanford.edu/~tibs/ElemStatLearn/)

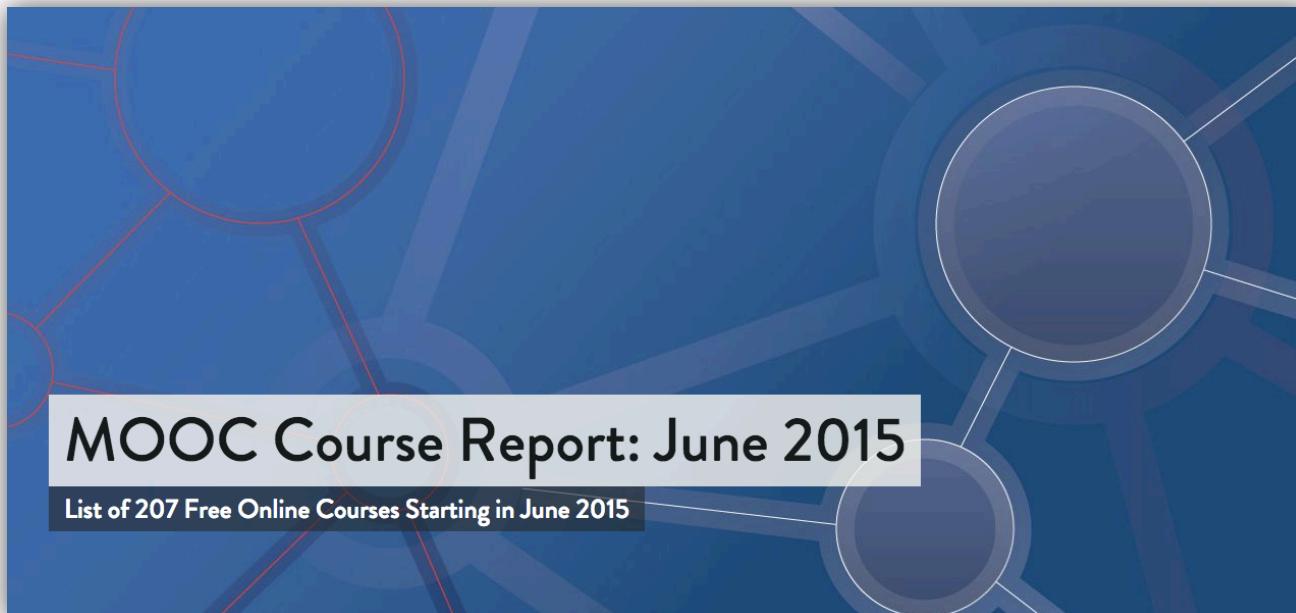
# Курсы по машинному обучению

- Н.Ю. Золотых, д.ф.-м.н.
  - ННГУ, ВМК, магистратура
  - ВШЭ, ПМИ, 4 курс
  - ВШЭ, ПМИ, 1 курс магистратуры



# Дистанционные курсы

- [www.coursera.org](http://www.coursera.org)
  - Andrew Ng's course
- [www.class-central.com](http://www.class-central.com)



# Школа Анализа Данных

Яндекс



[www.yandexdataschool.ru](http://www.yandexdataschool.ru)

# Контакты

Спасибо за внимание!

[ilya.lysenkov@itseez3d.com](mailto:ilya.lysenkov@itseez3d.com)