# Машинное обучение

Лекция 1

# Давайте знакомиться

Мышлянов Алексей Владимирович

Старший преподаватель и аспирант 3 года кафедры АСУ НИТУ МИСиС

Аналитик больших данных в МегаФоне (Senior Data Scientist) + Lead DS Accelerator

Контакты:

tg: @l3lush

# Вводный тест

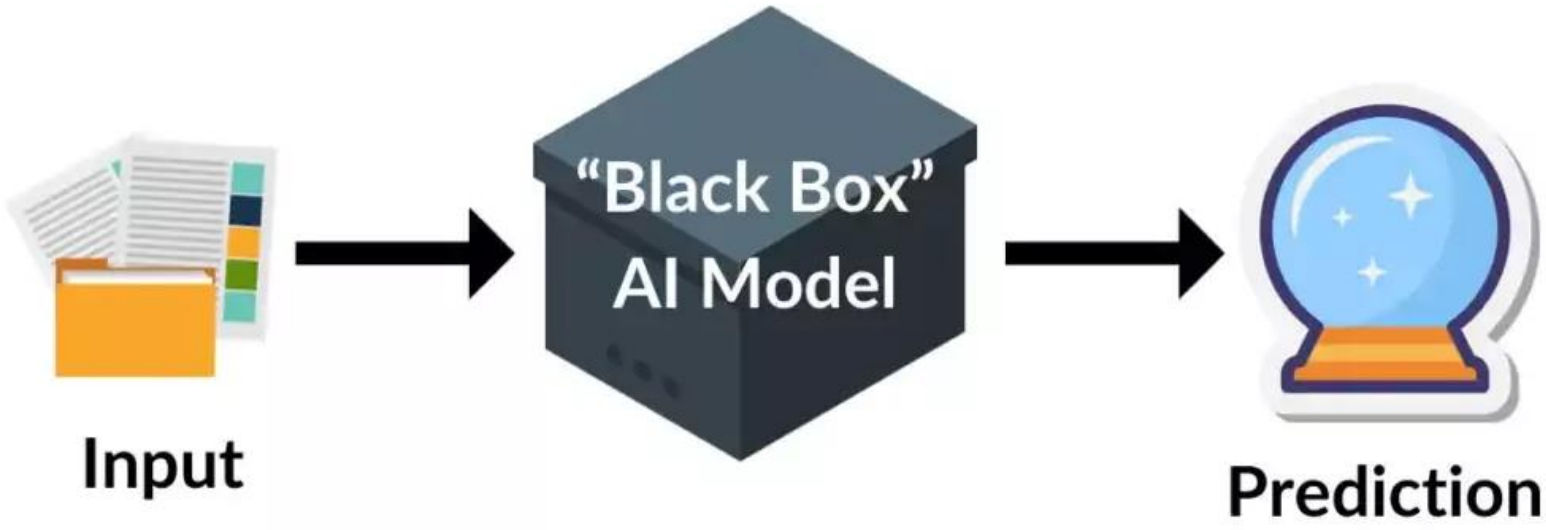
# База по МЛ. Цель машинного обучения

- **Целью машинного обучения** является частичная или полная **автоматизация** решения **сложных профессиональных задач** в самых разных областях человеческой деятельности (википедия)
- Машинное обучение — это **способ автоматического улучшения алгоритмов благодаря опыту**. (образовательная площадка №1)
- Цель машинного обучения — **предсказать результат по входным данным**. (образовательная площадка №2)
- Настоящая **цель машинного обучения** в том, чтобы дать программе возможность **научиться самостоятельно ставить условия и искать закономерности**. (ВК клауд)

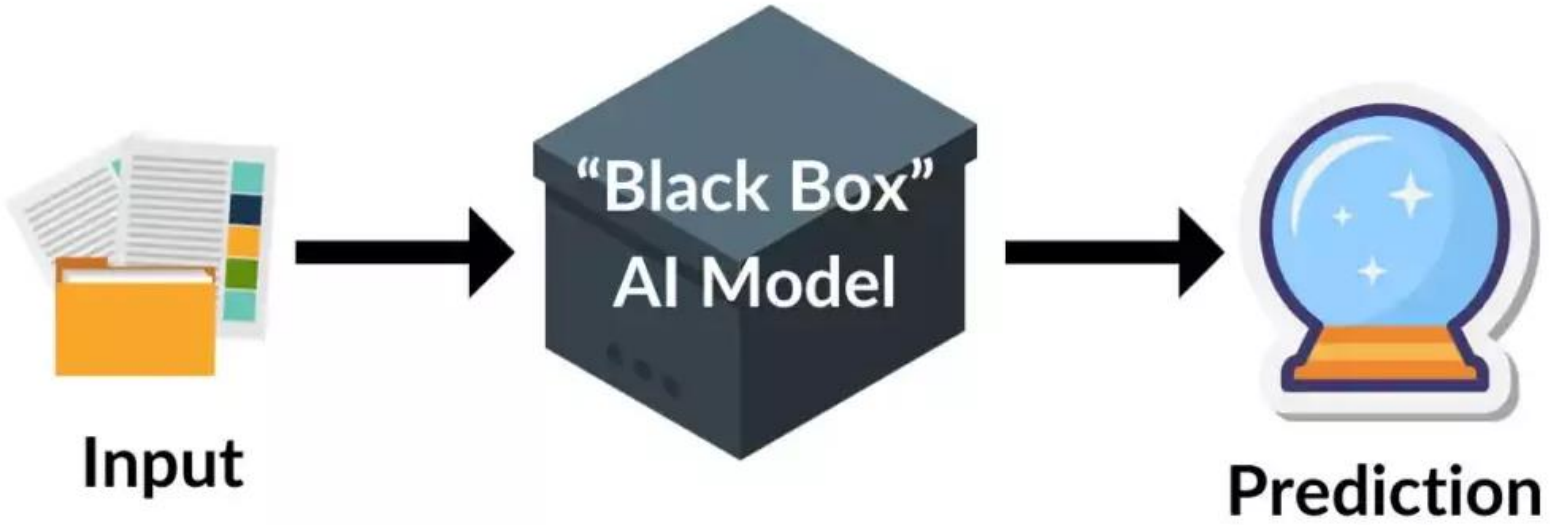
# База по МЛ. Черная коробка (пока)

На входе данные (табличные, картинки, видео, аудио, текст)

На выходе предсказания модели



# База по МЛ. Черная коробка (пока). Пример




| Имя | Фамилия | Доход | На какие нужды | На какой срок |
|---|---|---|---|---|
| Иван | Иванов | 100 000 | Машина | 10 лет |

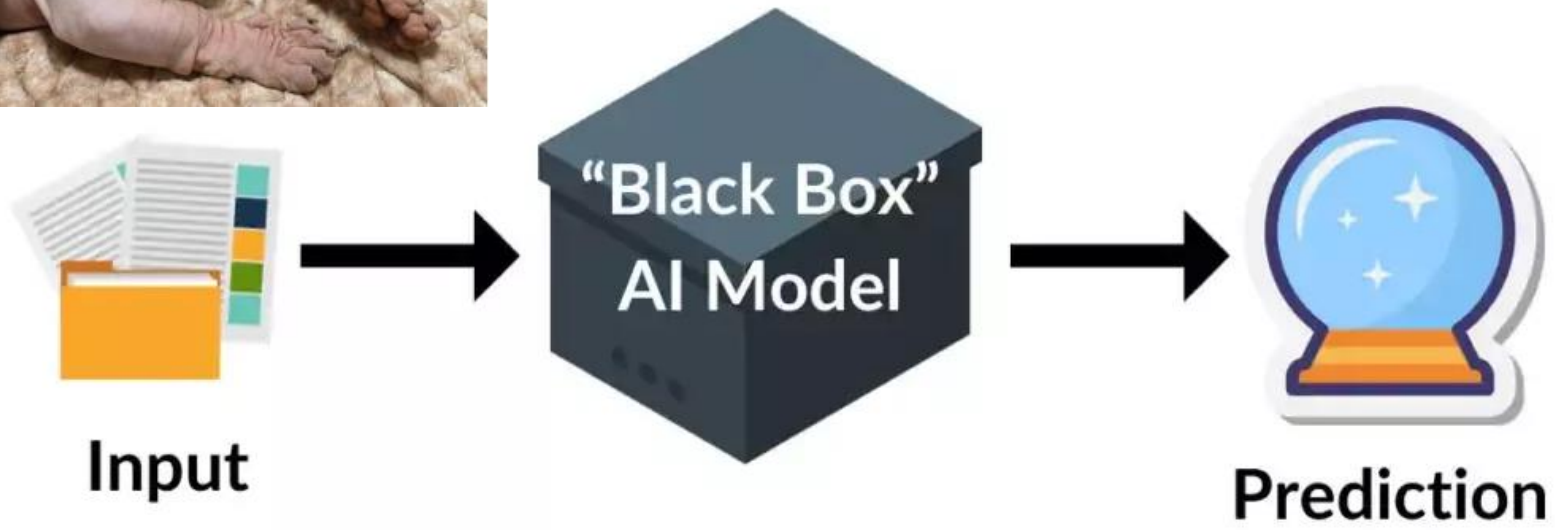
**Одобрить кредит**

да

# База по МЛ. Черная коробка (пока). Пример

**Что изображено?**

кот



# База по МЛ. Как выглядят данные

## Исторические данные

| Имя | Фамилия | Доход | На какие нужды | Сумма кредита | На какой срок | Вернул деньги |
|---|---|---|---|---|---|---|
| Петр | Зимов | 10 000 | Квартира | 10 000 000 | 30 лет | нет |
| Василий | Петров | 50 000 | Не уточнил | 1 000 000 | 1 год | да |
| Ирина | Васильева | 40 000 | Ремонт | 500 000 | 3 года | да |
| Диана | Якова | 30 000 | Путешествие | 300 000 | 2 года | нет |

## Новые данные

| Имя | Фамилия | Доход | На какие нужды | Сумма кредита | На какой срок | Вернул деньги |
|---|---|---|---|---|---|---|
| Иван | Иванов | 100 000 | Машина | 1 000 000 | 10 лет | ??? |

# База по МЛ. Основные понятия

$X$ – множество объектов (признаковое описание объектов, предикторы, независимые переменные, признаки, фичи, features)

$Y$ – множество ответов (метки ответов, зависимая переменная, целевая переменная, target, responses)

$X$ $Y$

| Имя | Фамилия | Доход | На какие нужды | Сумма кредита | На какой срок | Вернул деньги |
|---|---|---|---|---|---|---|
| Петр | Зимов | 10 000 | Квартира | 10 000 000 | 30 лет | нет |
| Василий | Петров | 50 000 | Не уточнил | 1 000 000 | 1 год | да |
| Ирина | Васильева | 40 000 | Ремонт | 500 000 | 3 года | да |
| Диана | Якова | 30 000 | Путешествие | 300 000 | 2 года | нет |

# База по МЛ. Основные понятия

$\{x_1, \dots, x_m\} \subset X$ (либо $\mathbb{X}$)
$\{y_1, \dots, y_m\} \subset Y$ (либо $\mathbb{Y}$)

$y: X \to Y$ – неизвестная зависимость (target function)

$a: X \to Y$ – алгоритм машинного обучения (decision function)
$a \sim y$

| | Имя | Фамилия | Доход | На какие нужды | Сумма кредита | На какой срок | Вернул деньги | |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | Петр | Зимов | 10 000 | Квартира | 10 000 000 | 30 лет | нет | $y_1$ |
| $x_2$ | Василий | Петров | 50 000 | Не уточнил | 1 000 000 | 1 год | да | $y_2$ |
| $x_3$ | Ирина | Васильева | 40 000 | Ремонт | 500 000 | 3 года | да | $y_3$ |
| $x_4$ | Диана | Якова | 30 000 | Путешествие | 300 000 | 2 года | нет | $y_4$ |

# База по МЛ. Какие бывают признаки

Типы признаков:

1. Бинарные $\in \{0, 1\}$
2. Числовые $\in \mathbb{R}$
3. Категориальные $\in \{0, \dots, M\}$
4. Порядковые $\in \{0, \dots, M\}$

# База по МЛ. Вникаем в X

Один объект $x = (x_1, \dots, x_m)$ – это вектор признаков в размере $m$ штук

Учитывая, что объектов может быть много, запишем $x_i = (x_{i1}, \dots, x_{im})$

В итоге $X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{l1} & \dots & x_{lm} \end{pmatrix}$

Набор данных (обучающая выборка):

$$S = \{(x_i, y_i)\}_{i=1}^{\ell}$$

# База по МЛ. Вникаем в Y. Тип задачи

Возможные варианты по типам задач:

1. Регрессия $Y = \mathbb{R}$
2. Классификация $Y = \{1, \dots, K\}$
3. Бинарная классификация $Y = \{-1, +1\}$
4. Многоклассовая (multilabel) классификация $Y = \{0, 1\}^K$

Все это – **обучение с учителем**, в обучающей выборке есть ответы (supervised learning)

# База по МЛ. Вникаем в Y. Тип задачи

Возможные варианты по типам задач:

1. Кластеризация
2. Понижение размерности
3. Визуализация данных
4. Задача обнаружения аномалий

В этих задачах у нас нет $Y$, то есть выборка $S = \{(x_i)\}_{i=1}^{\ell}$. Такое называется **обучением без учителя** (unsupervised learning).

# Как решать задачи

Предполагаем, что модель дает предсказания, запишем это в виде

$$\hat{Y} = f(X).$$

Причем $\hat{Y}$ должно быть максимально приближено к $Y$.

Как оценить, насколько $\hat{Y}$ приближено к $Y$?

# Как решать задачи. Функции потерь

Для задачи регрессии (MSE loss, метод наименьших квадратов):

$$Q(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} (y - \hat{y})^2$$

Для задачи классификации (Cross Entropy Loss):

$$Q(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} y \cdot \log(\hat{y})$$

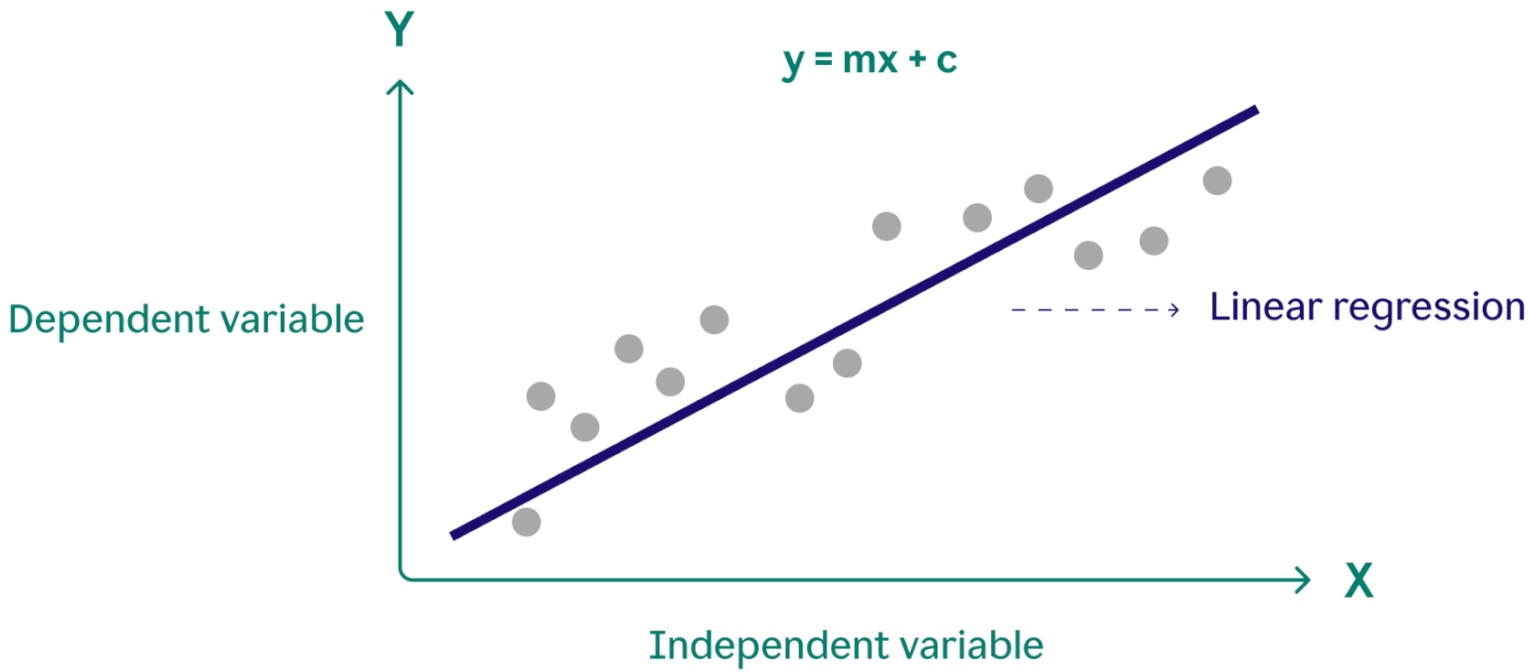
# Графическая интерпретация

Линейные модели:

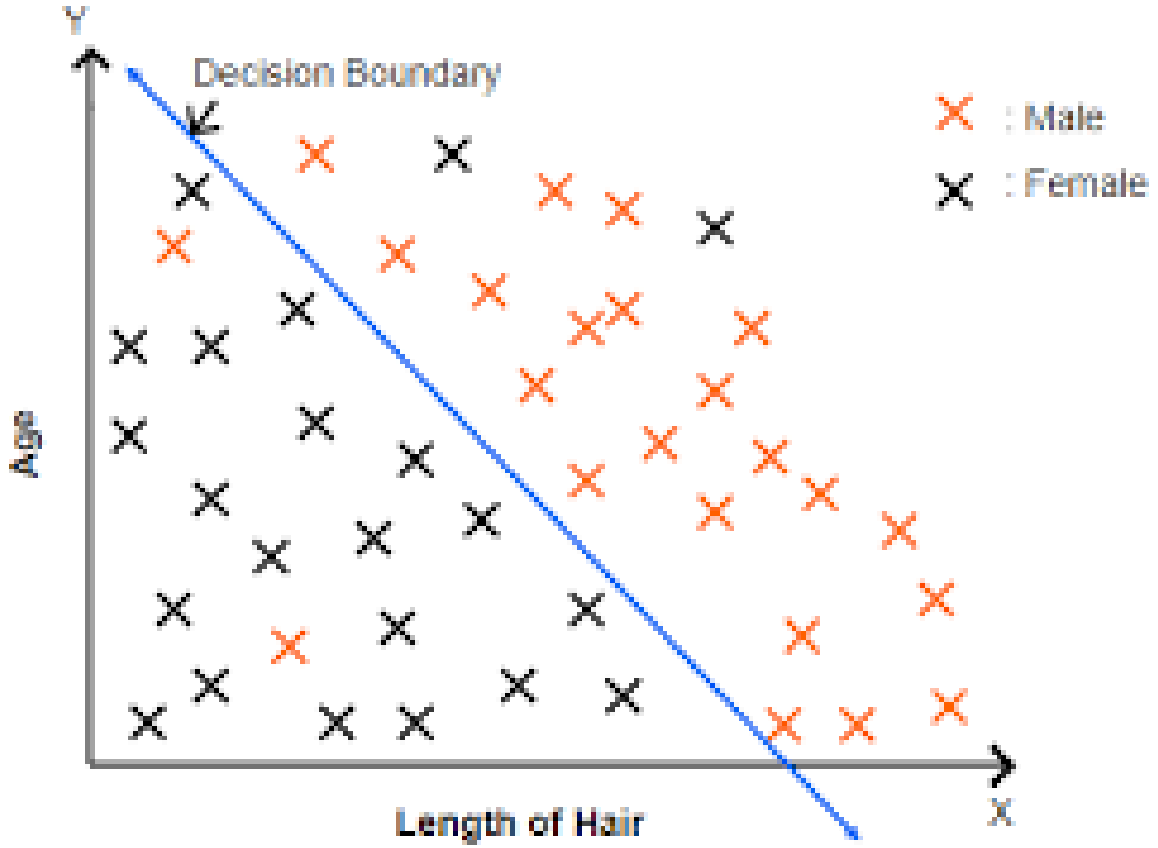
$$\hat{y} = \underset{M}{kX} + b$$

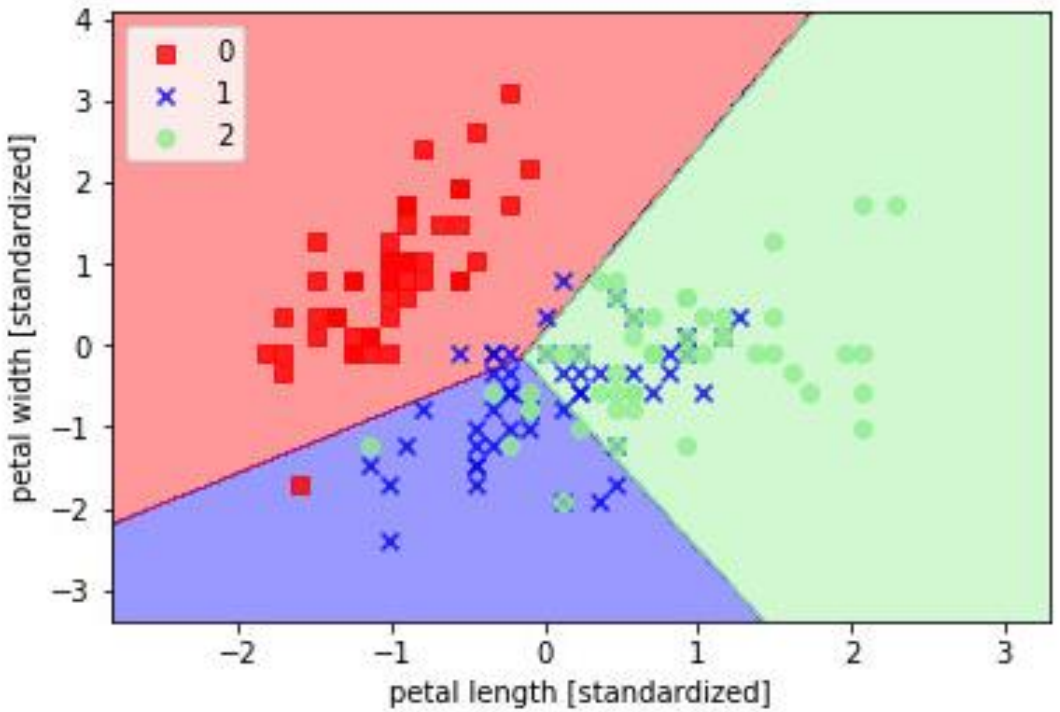
$$\hat{y} = \sum_{j=1} k_j X_j + b$$

# Графическая интерпретация. Регрессия



# Графическая интерпретация. Классификация



# Где можно увидеть ML?

Абсолютно везде и практически в любой сфере жизни

Давайте посмотрим на примерах

# Кредитный скоринг

$X$ – данные о клиенте

$Y$ – вернет ли кредитные деньги?

**Признаки:**

*Клиент:* пол, возраст, кредитная история, доход, место работы, должность

# Предложение тарифа

$X$ – пара (абонент, тариф)

$Y$ – подключить ли данный тариф

**Признаки:**

*Абонент:* возраст, пол, пользование сервисами, интернетом, звонками и пр.

*Тариф:* наполнение в Гб, минутах, дополнительные опции, стоимость

# Оценка квартиры

$X$ – данные по квартире

$Y$ – стоимость

Признаки:

*Квартира:* площадь, местоположение, количество комнат, тип ремонта, материалы дома

# Рекомендации фильмов/музыки/ленты

$X$ – пара (пользователь, фильм)

$Y$ – посмотрит / понравится фильм

Признаки:

*Пользователь:* возраст, пол, история просмотров, оценки посмотренных фильмов

*Фильм:* жанр, год выпуска, продолжительность, актерский состав

# Поисковое ранжирование

$X$ – запрос пользователя

$Y$ – ранжированный список документов (сайтов)

Признаки:

*Запрос пользователя:* непосредственно запрос вроде «какой сегодня год?»

# Определение объектов на изображении

$X$ – изображение (матрица пикселей)

$Y$ – класс того, что изображено на картинке (например, кот)

Признаки:

*Изображение:* яркость каждого пикселя по шкале от 0 до 255