



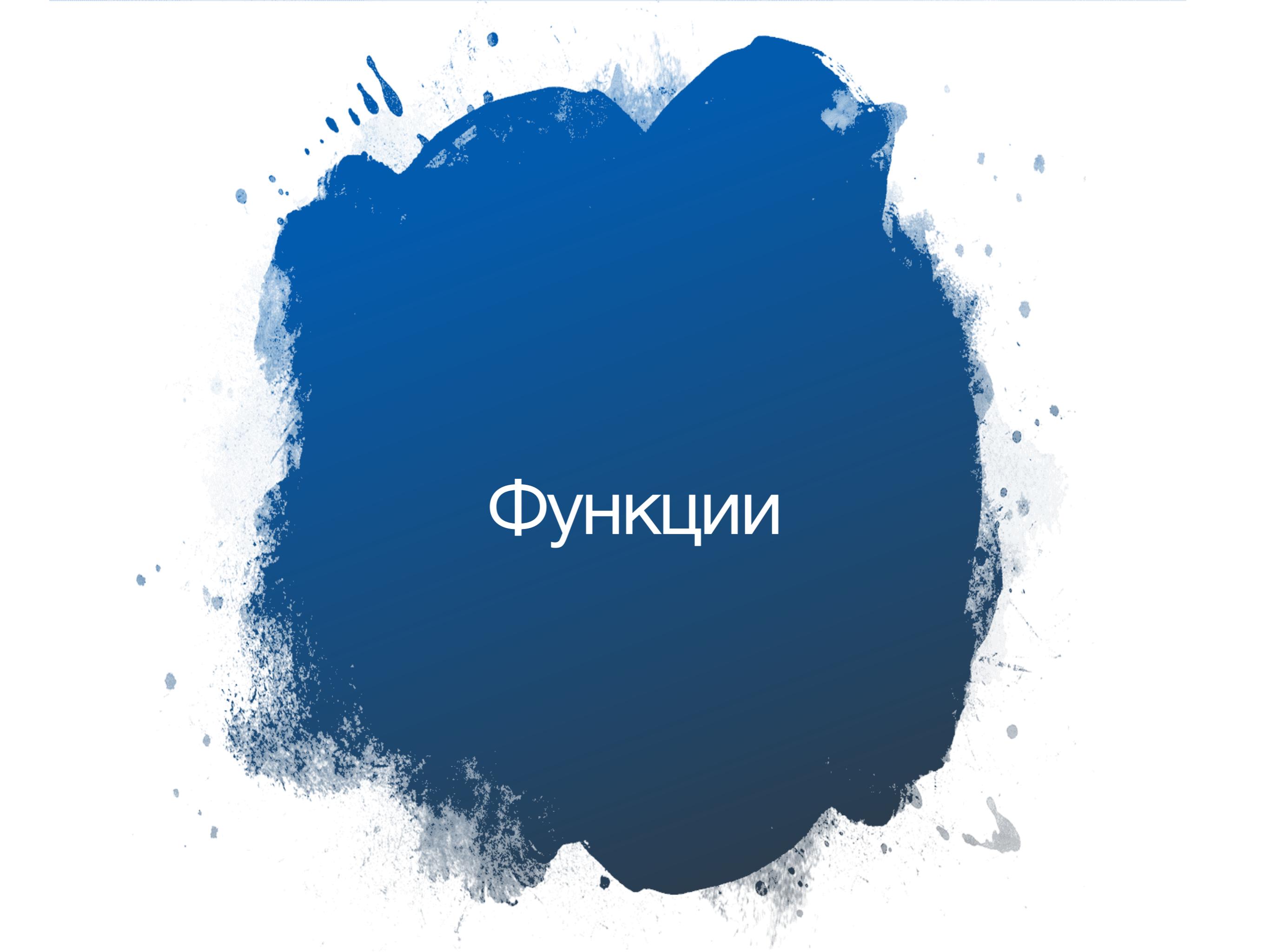
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

МАТЕМАТИКА ДЛЯ DATA SCIENCE: ВАЖНЫЕ ВОСПОМИНАНИЯ И НОВИНКИ

Теванян Элен

12.04.2019

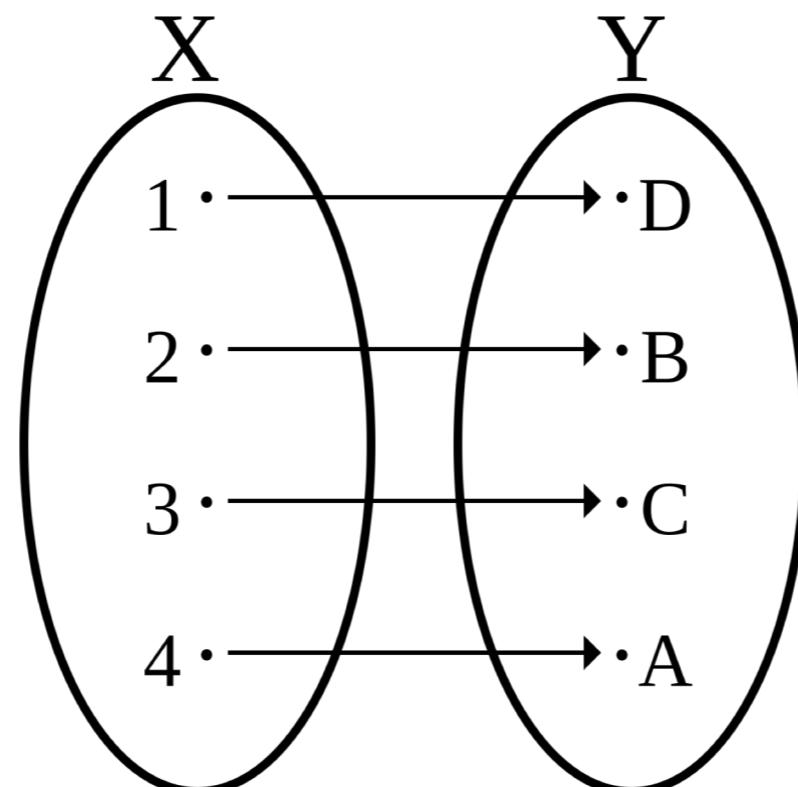
Москва 2019



ФУНКЦИИ

ФУНКЦИЯ

- Правило, по которому всякому значению из множества X ставим в соответствие какое-то значение из множества Y



ПРИМЕРЫ ФУНКЦИЙ

График функции $y = 2x + 2$

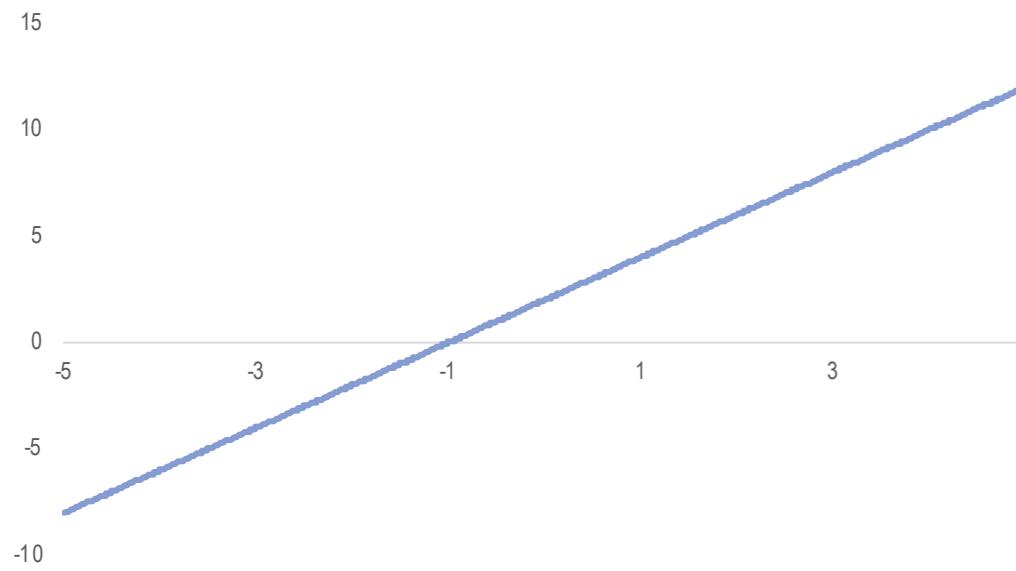
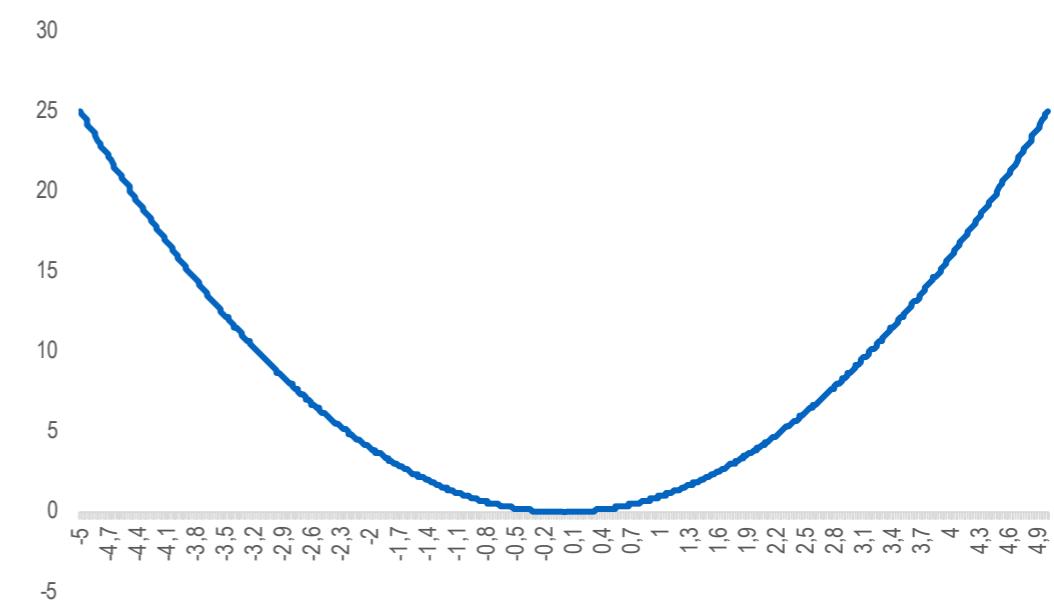


График функции $y=(x - 5)^2$



ПРИМЕРЫ ФУНКЦИЙ



$\sin(x)$



$\cos(x)$



$\tan(x)$



$\cot(x)$



$|x|$



x



x^2



$x^2 + y^2$



\sqrt{x}



$\sqrt{-x}$



$\frac{1}{x}$

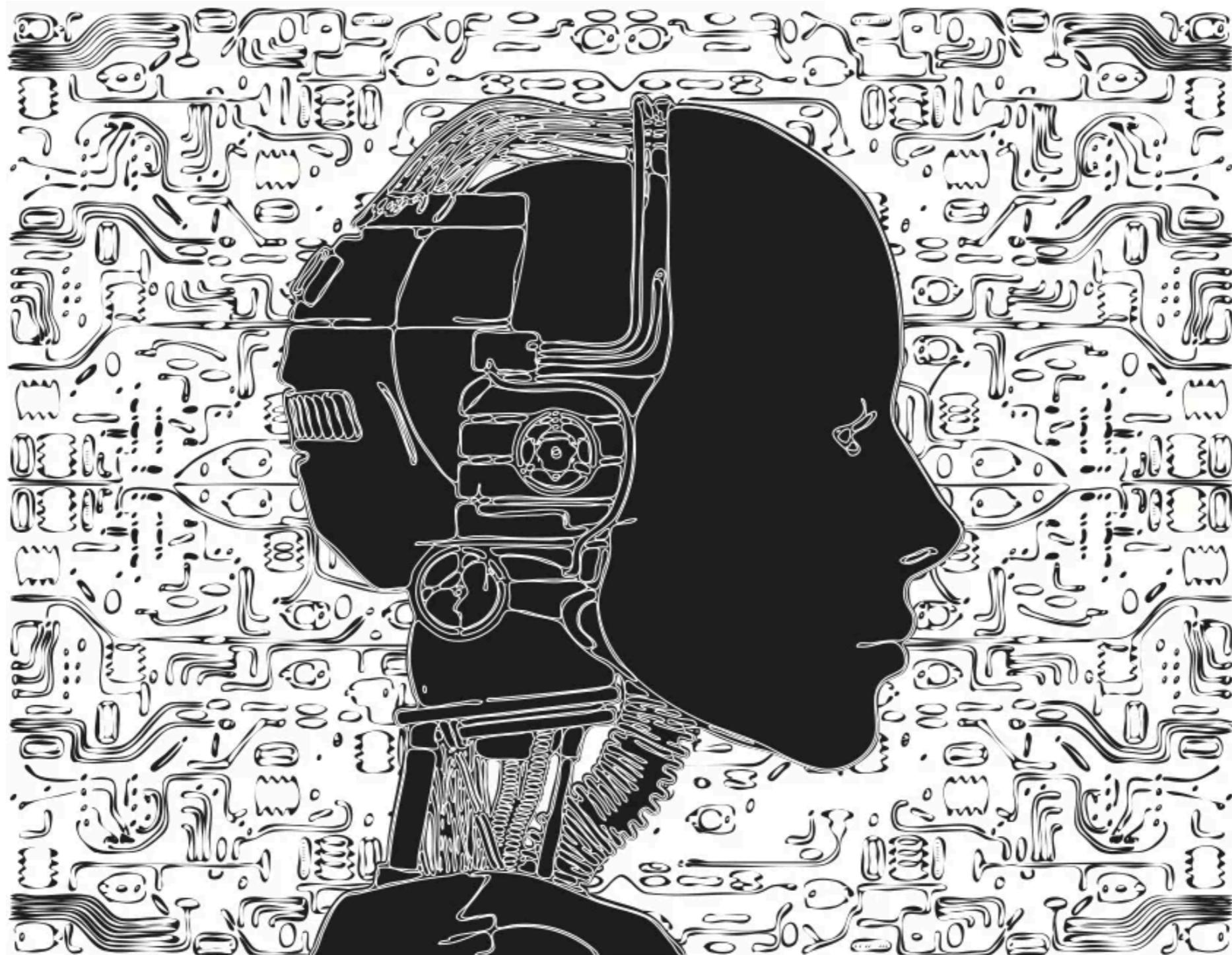


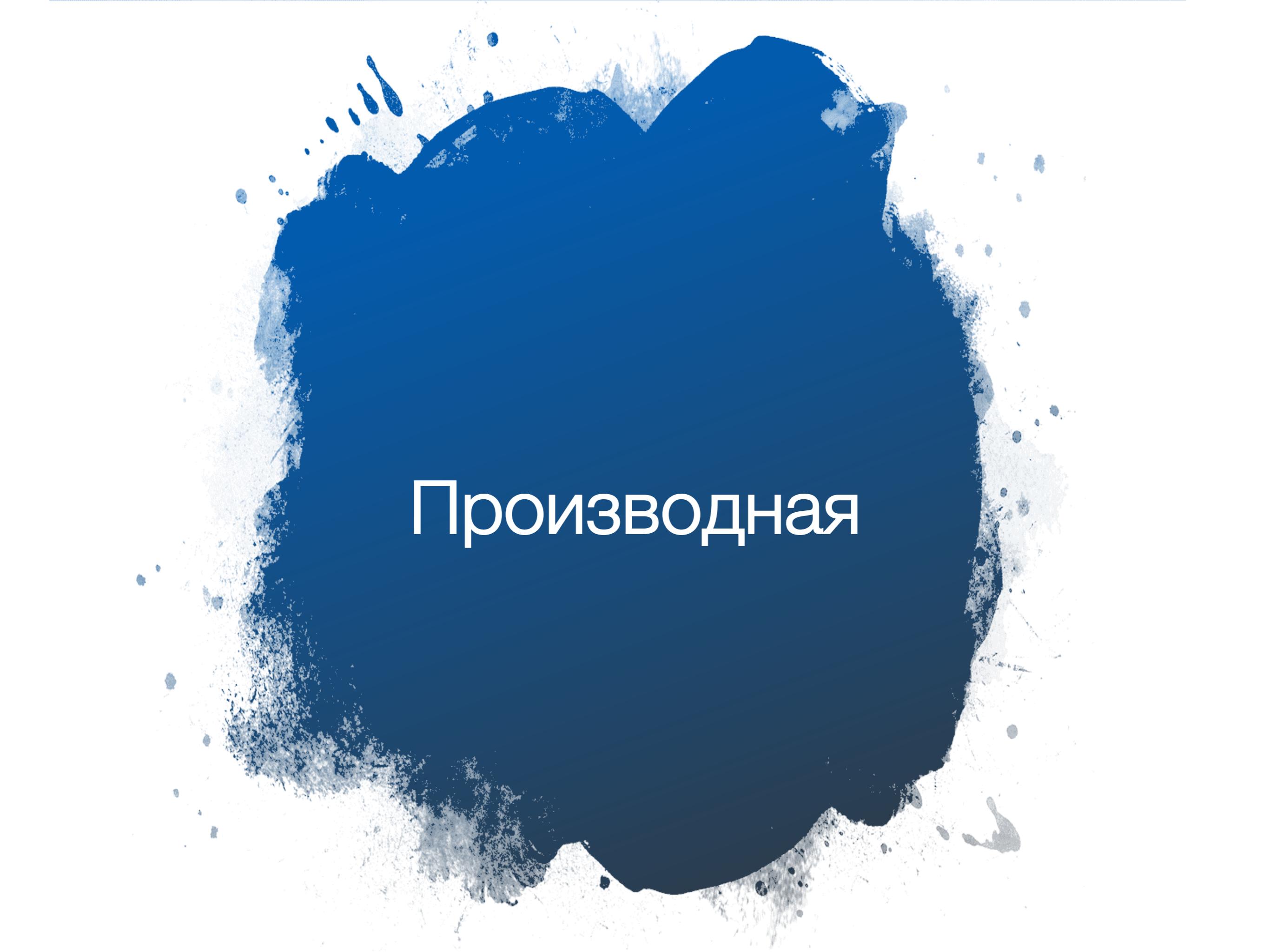
ЗАЧЕМ ЛЮДИ ПРИДУМАЛИ ФУНКЦИИ?

- Сколько денег нужно на путешествие?
- На сколько 100 новых сотрудников смогут увеличить прибыль компании?
- Сколько нужно привлечь блогеров, чтобы получить 1000 новых клиентов?

ГДЕ НАМ ПОНАДОБЯТСЯ ФУНКЦИИ?

- Алгоритм машинного обучения $a(x)$ – это функция

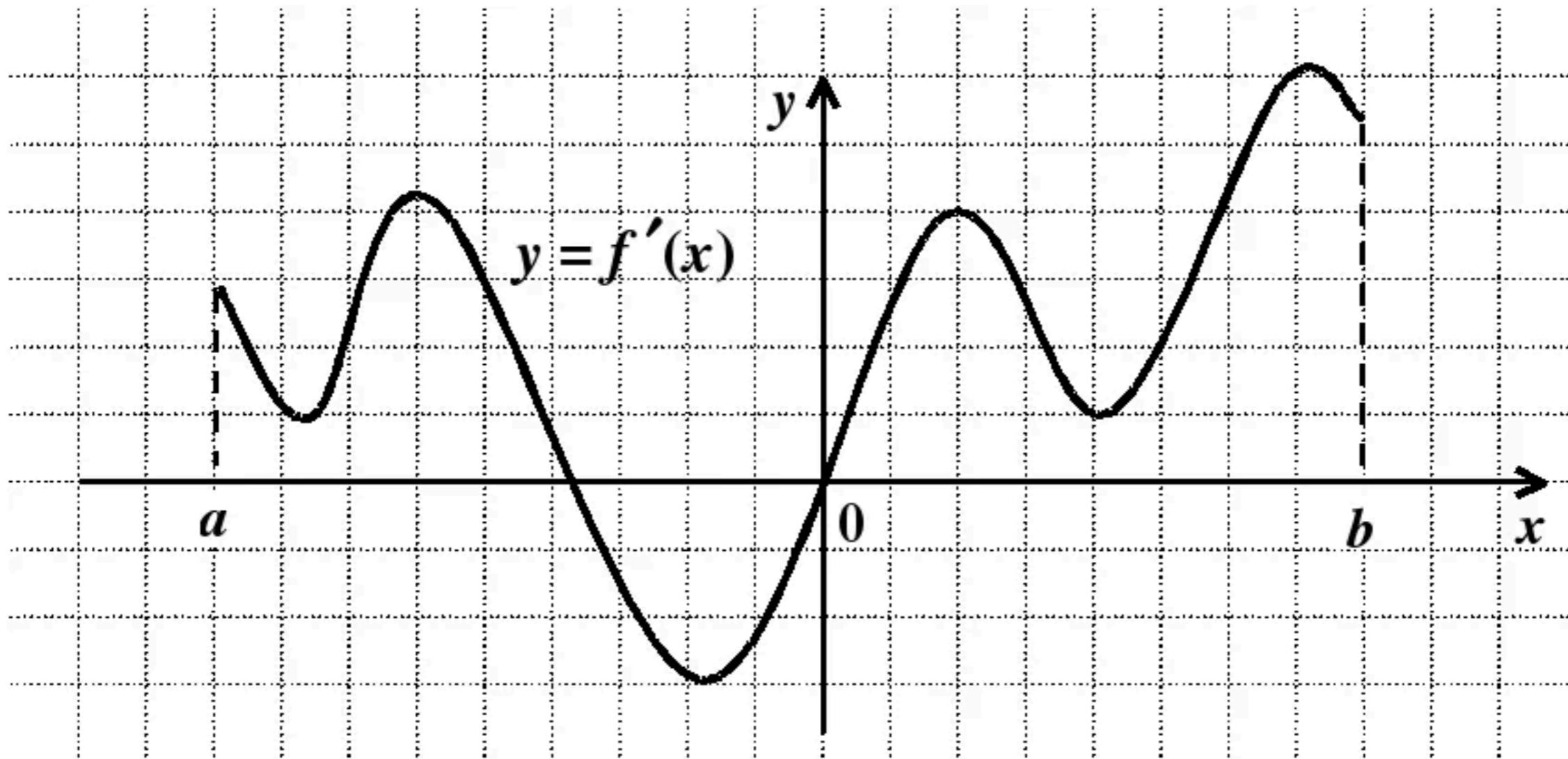




Производная

ПРОИЗВОДНАЯ

- Физически: скорость изменения функции в точке
- Геометрически: тангенс угла касательной к функции в точке

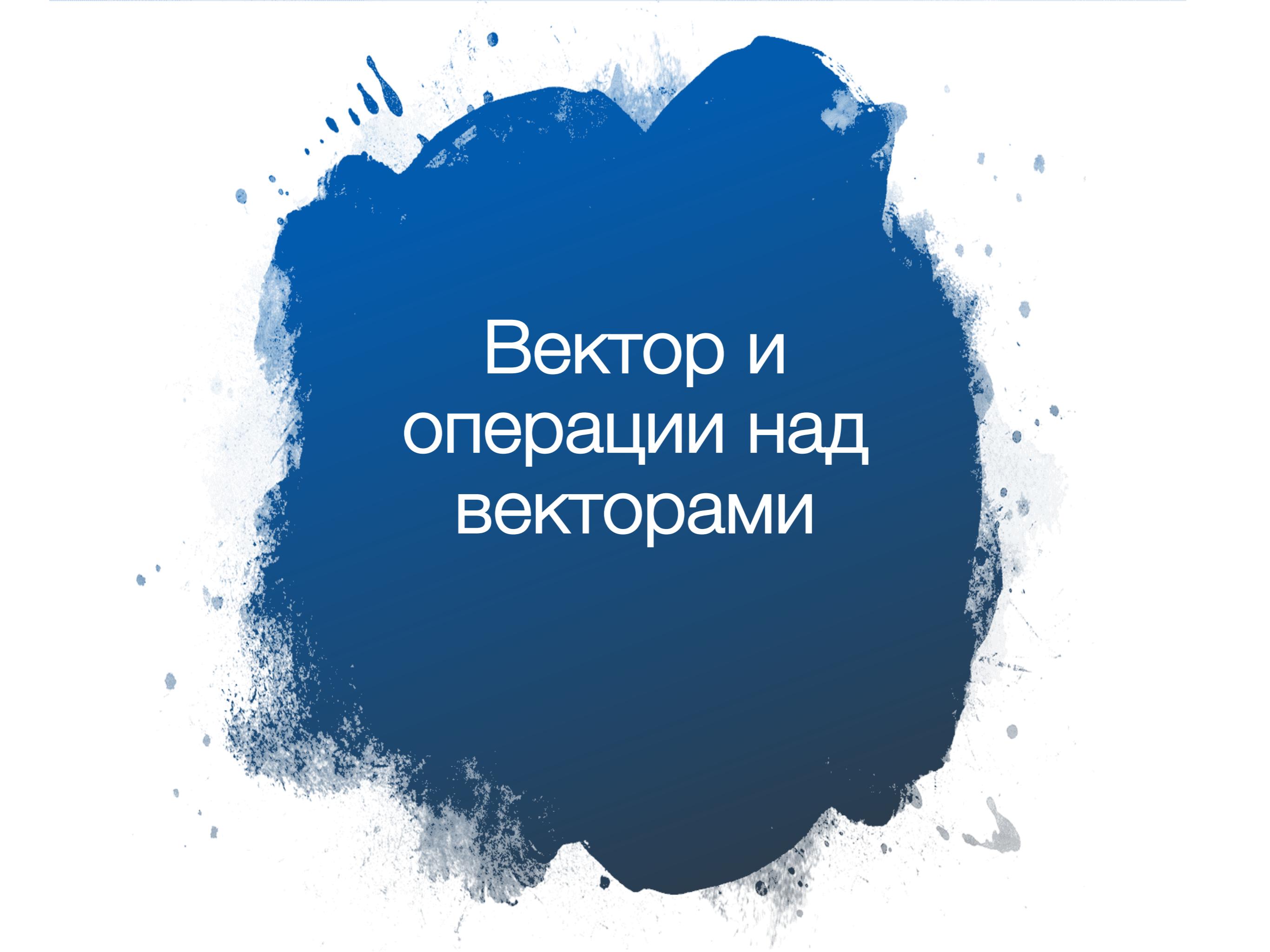


ЗАЧЕМ ЛЮДИ ПРИДУМАЛИ ПРОИЗВОДНУЮ?

- Как минимизировать траты на отдых?
- Как максимизировать прибыль с помощью 100 новых сотрудников?
- Как максимизировать число новых клиентов при работе с блогерами?

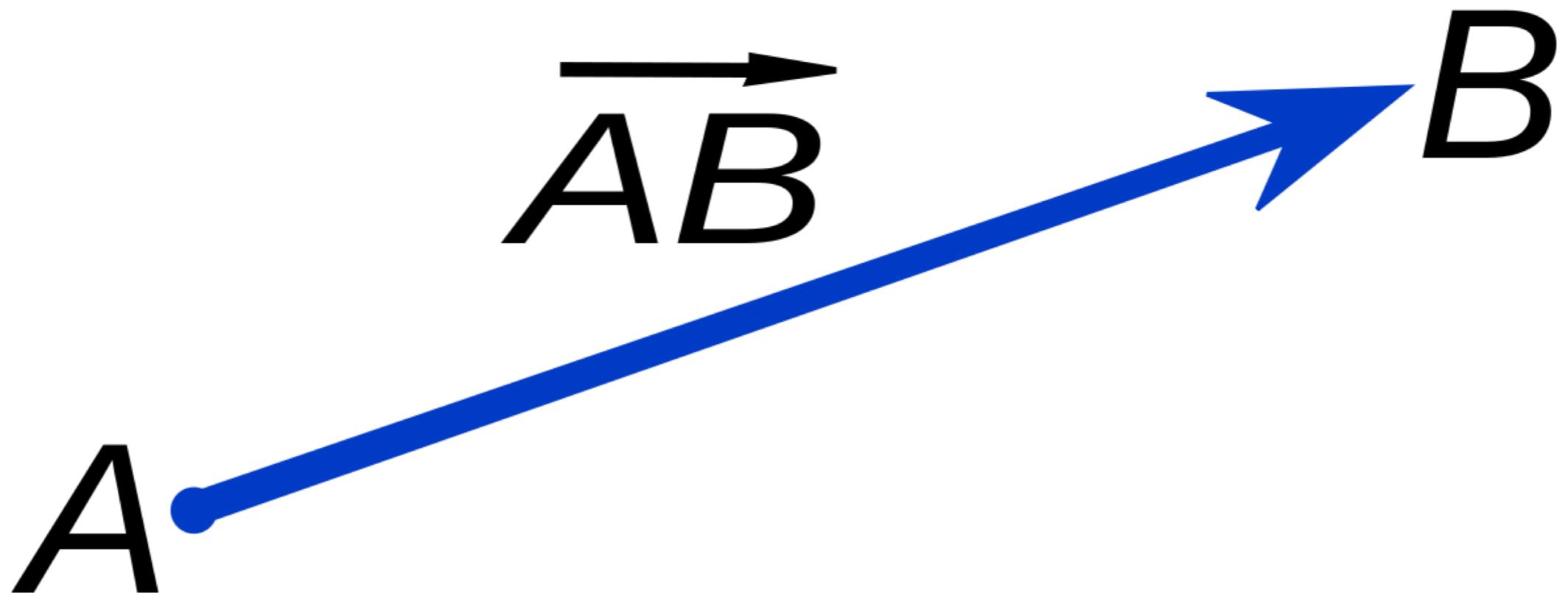
ГДЕ НАМ ПОНАДОБЯТСЯ ПРОИЗВОДНЫЕ?

- В чистом и явном виде в этом модуле не увидимся



Вектор и операции над векторами

ВЕКТОР



ВЕКТОР

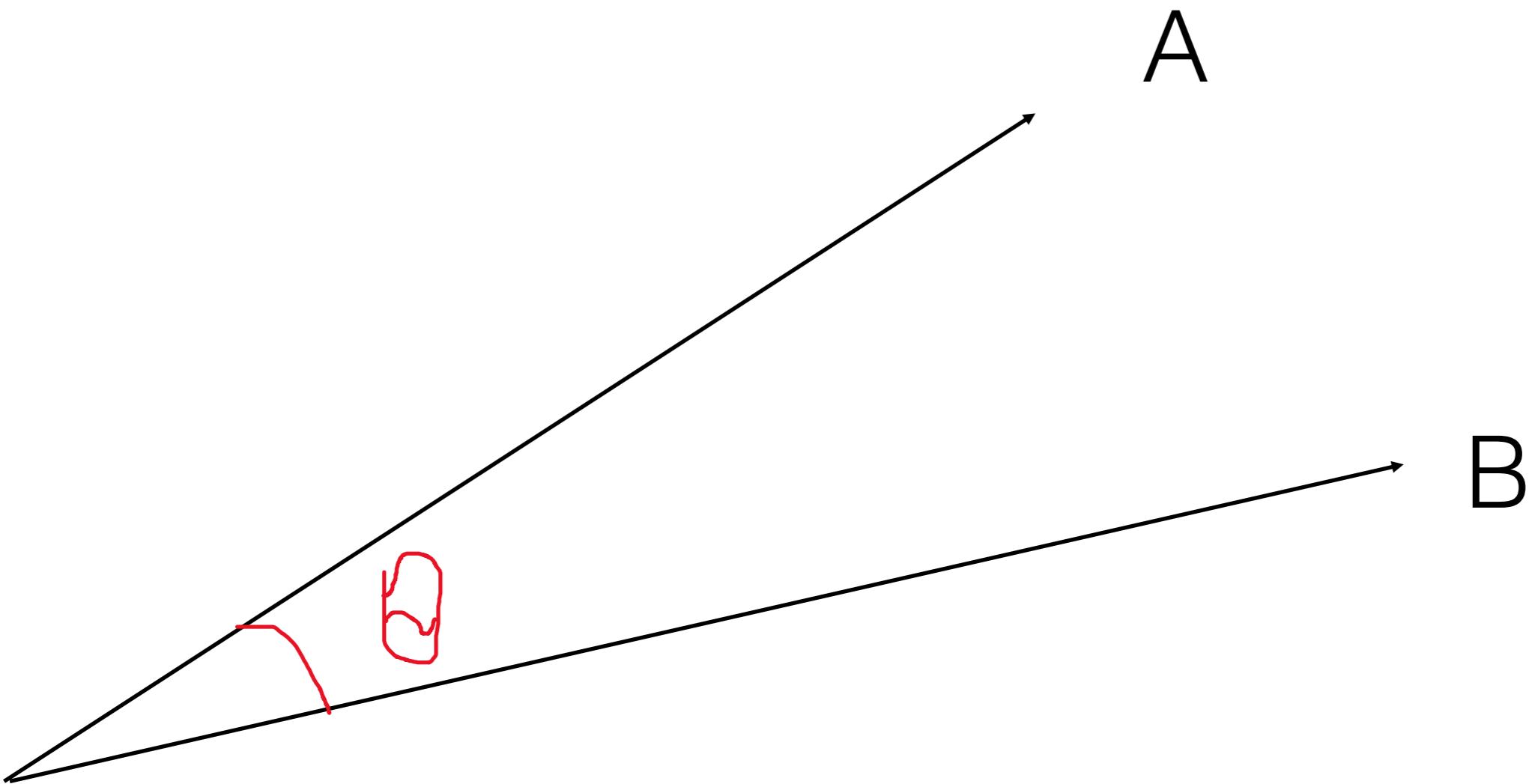
- $\vec{x} = (x_1, x_2, \dots, x_n)^T$ – обозначение вектора
- $\vec{x} = (x_1, x_2, x_3)^T$ – трехмерный вектор

МОДУЛЬ ВЕКТОРА

$$|\vec{a}| = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

УГОЛ МЕЖДУ ВЕКТОРАМИ

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$



ГДЕ НАМ НУЖНЫ ВЕКТОРЫ

- Каждый столбец данных – вектор 😊

Расстояния между точками (и векторами)

КАК НАЙТИ РАССТОЯНИЕ МЕЖДУ ТОЧКАМИ?

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

b^1

b^2

b^3

ЕВКЛИДОВО РАССТОЯНИЕ

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

b^1

b^2

b^3

- $d(b^1, b^2) = \sqrt{\sum_{i=1}^n (b_i^1 - b_i^2)^2}$

ЕВКЛИДОВО РАССТОЯНИЕ

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

b^1

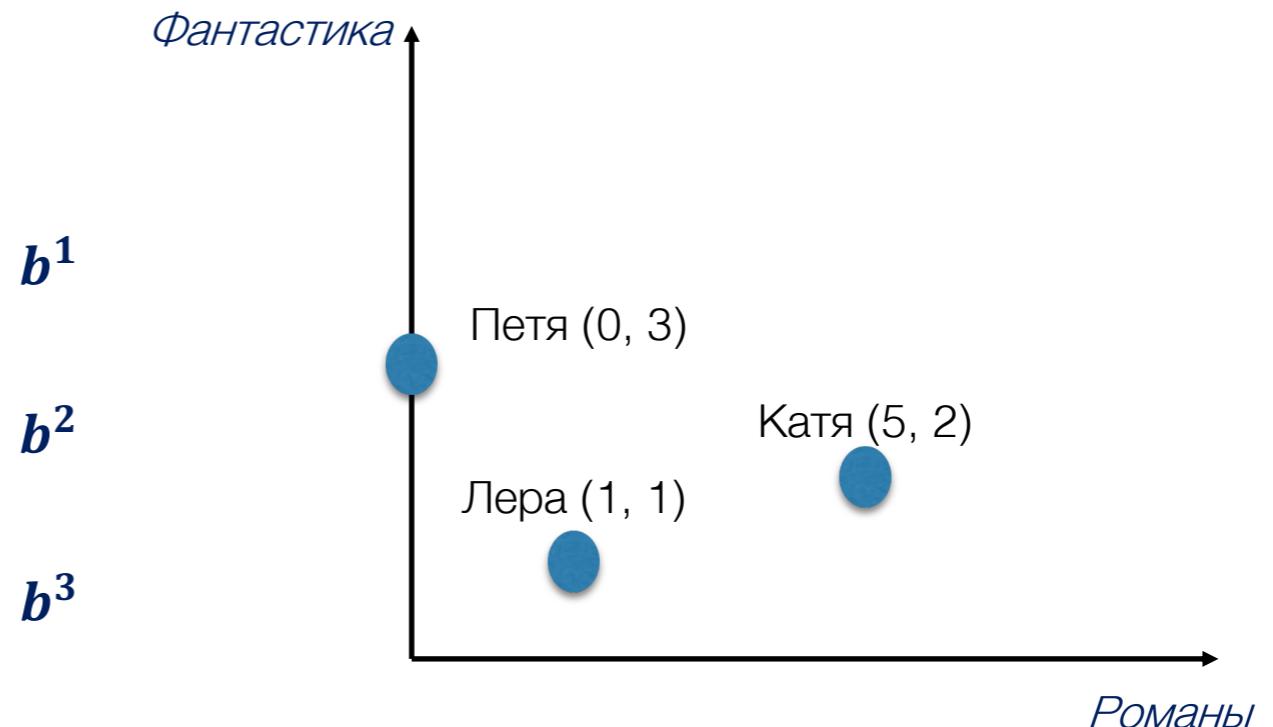
b^2

b^3

- $d(b^1, b^2) = \sqrt{\sum_{i=1}^n (b_i^1 - b_i^2)^2}$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

ЕВКЛИДОВО РАССТОЯНИЕ

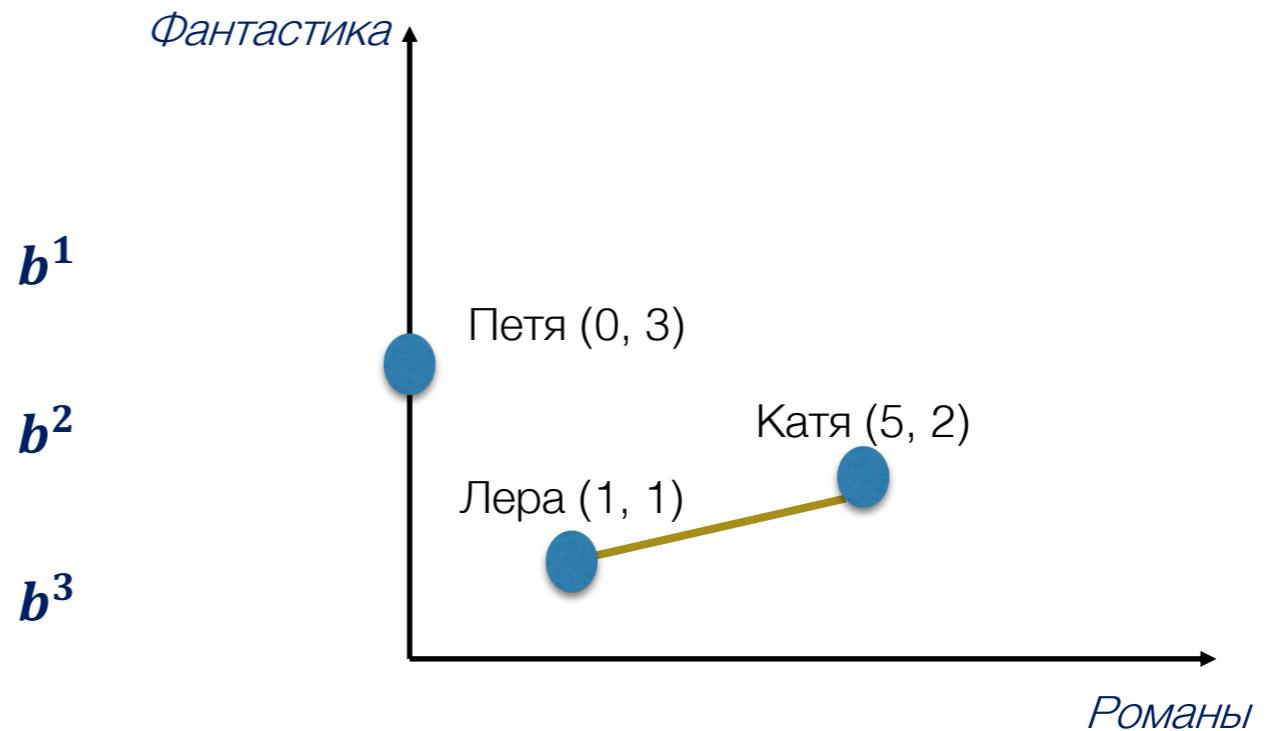
	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3



- $d(b^1, b^2) = \sqrt{\sum_{i=1}^n (b_i^1 - b_i^2)^2}$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

ЕВКЛИДОВО РАССТОЯНИЕ

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

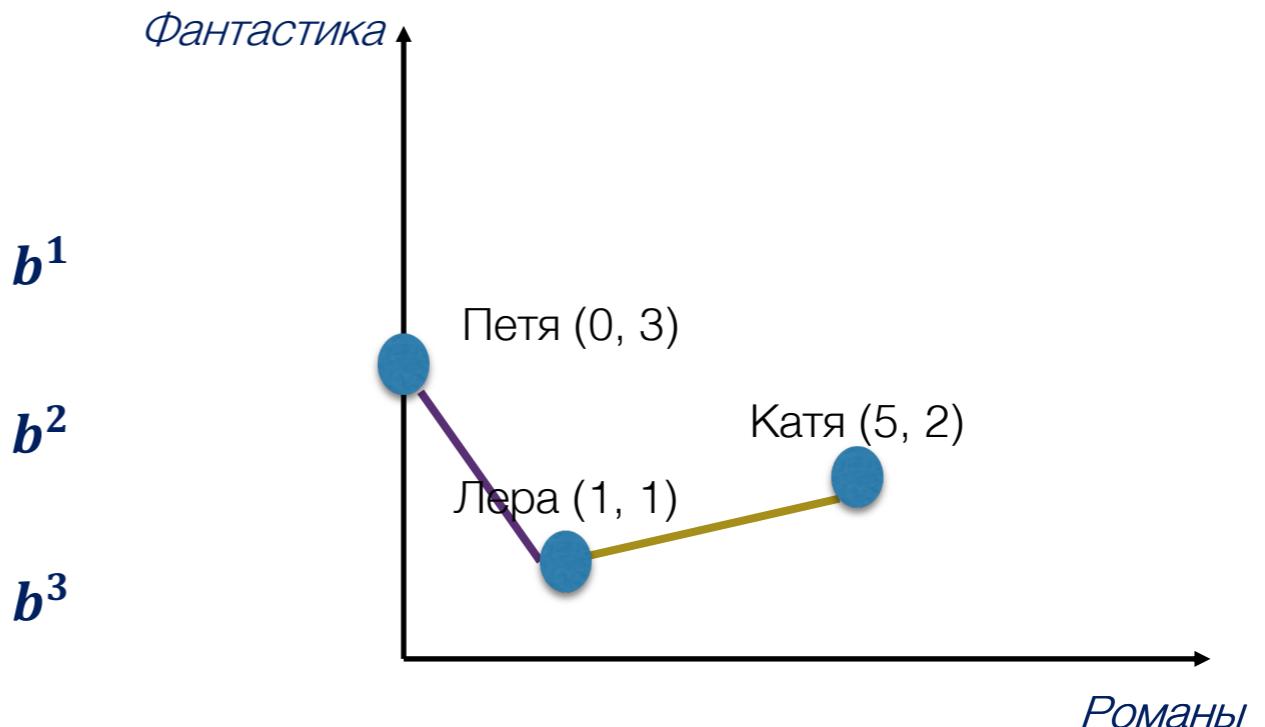


- $d(b^1, b^2) = \sqrt{\sum_{i=1}^n (b_i^1 - b_i^2)^2}$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

$$\bullet \quad d(b^1, b^2) = \sqrt{(5 - 1)^2 + (2 - 1)^2} = 4.21$$

ЕВКЛИДОВО РАССТОЯНИЕ

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

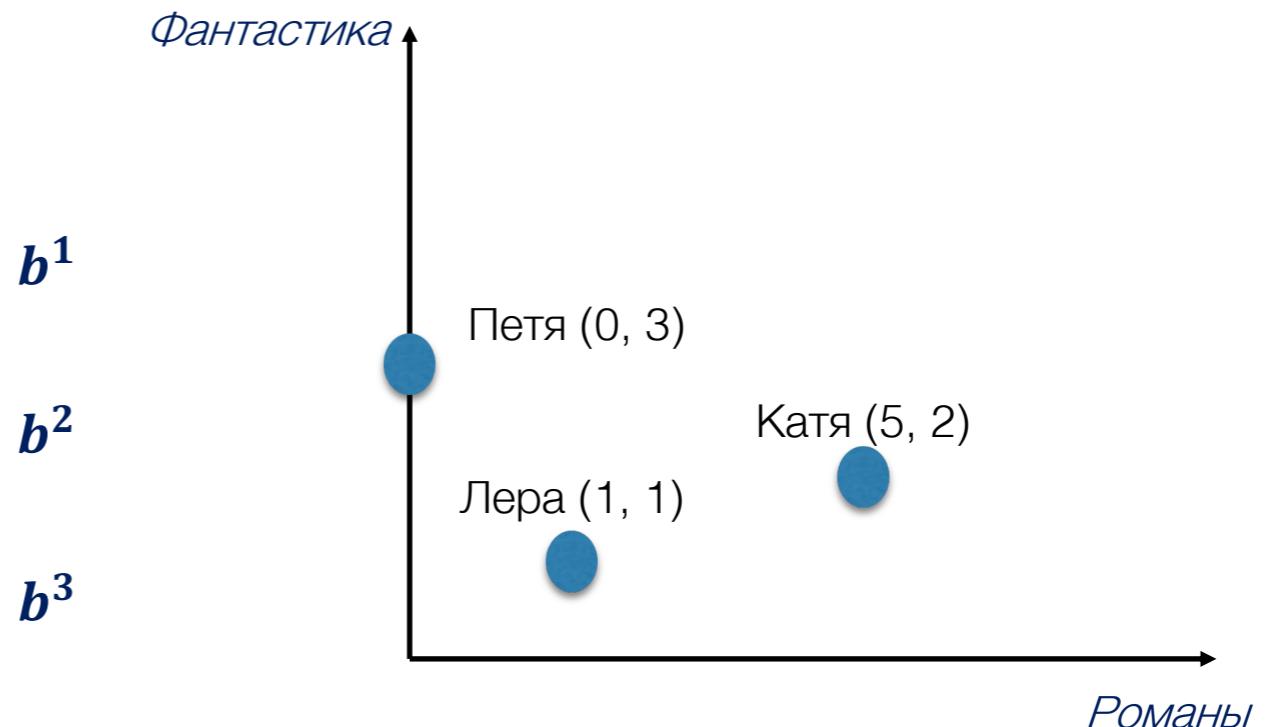


- $d(b^1, b^2) = \sqrt{\sum_{i=1}^n (b_i^1 - b_i^2)^2}$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

- $d(b^1, b^2) = \sqrt{(5 - 1)^2 + (2 - 1)^2} = 4.21$
- $d(b^1, b^2) = \sqrt{(1 - 0)^2 + (1 - 3)^2} = 2.24$

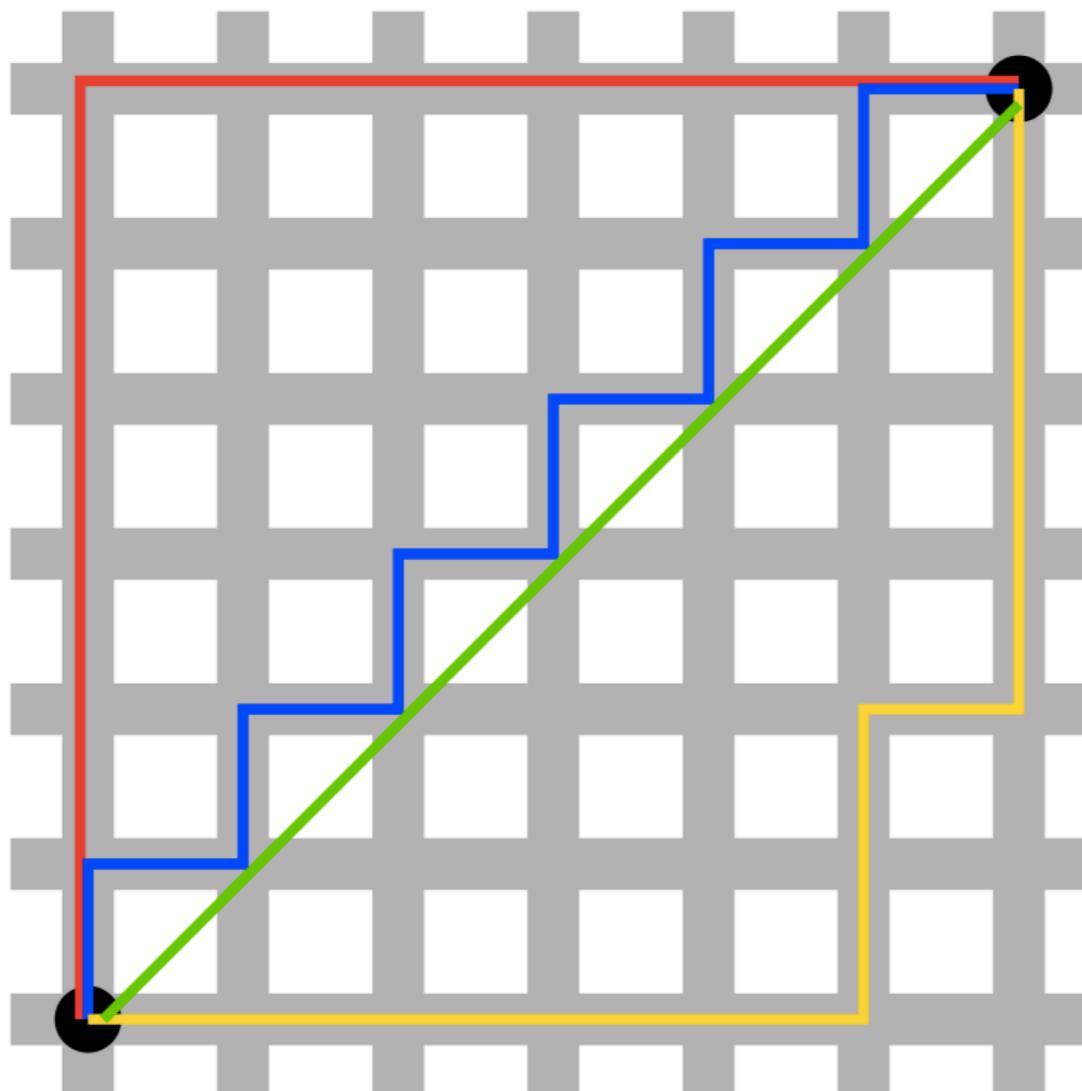
РАССТОЯНИЕ ХЭММИНГА (МАНХЕТТЕН)

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3



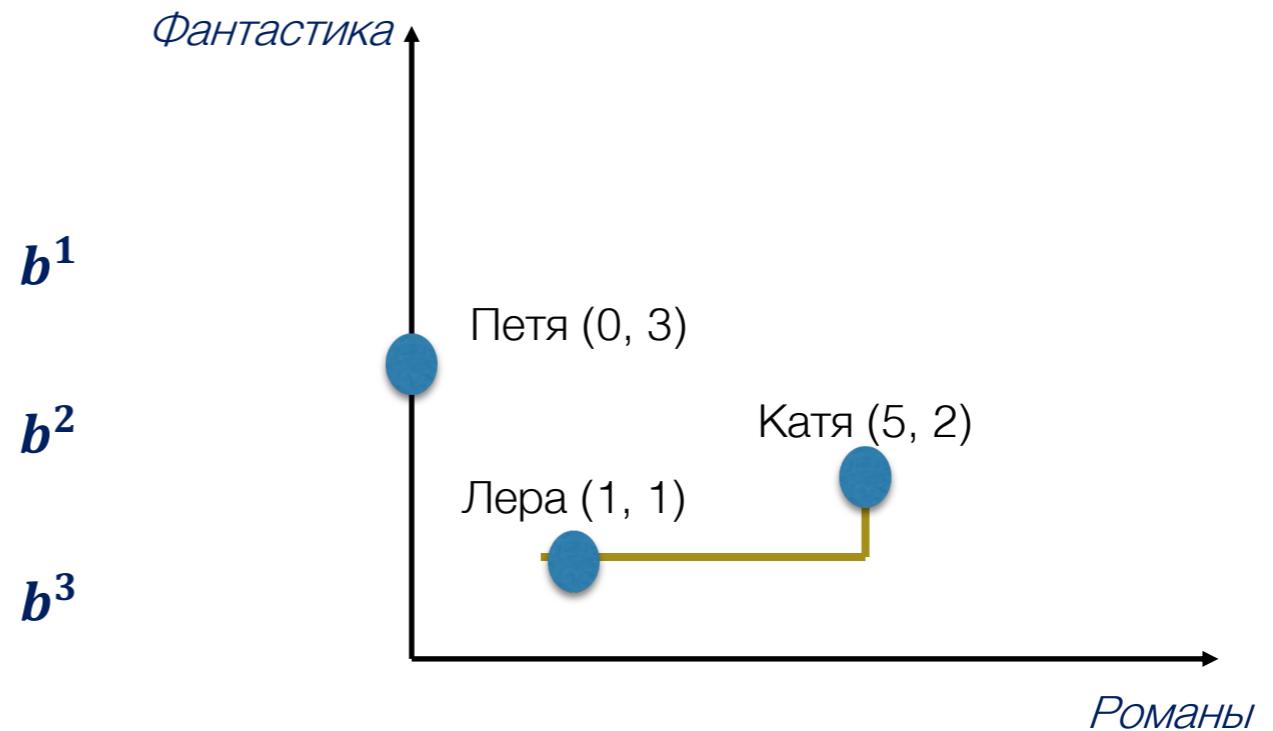
- $d(b^1, b^2) = \sum_{i=1}^n |b_i^1 - b_i^2|$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

РАССТОЯНИЕ ГОРОДСКИХ КВАРТАЛОВ



РАССТОЯНИЕ ХЭММИНГА (МАНХЕТТЕН)

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

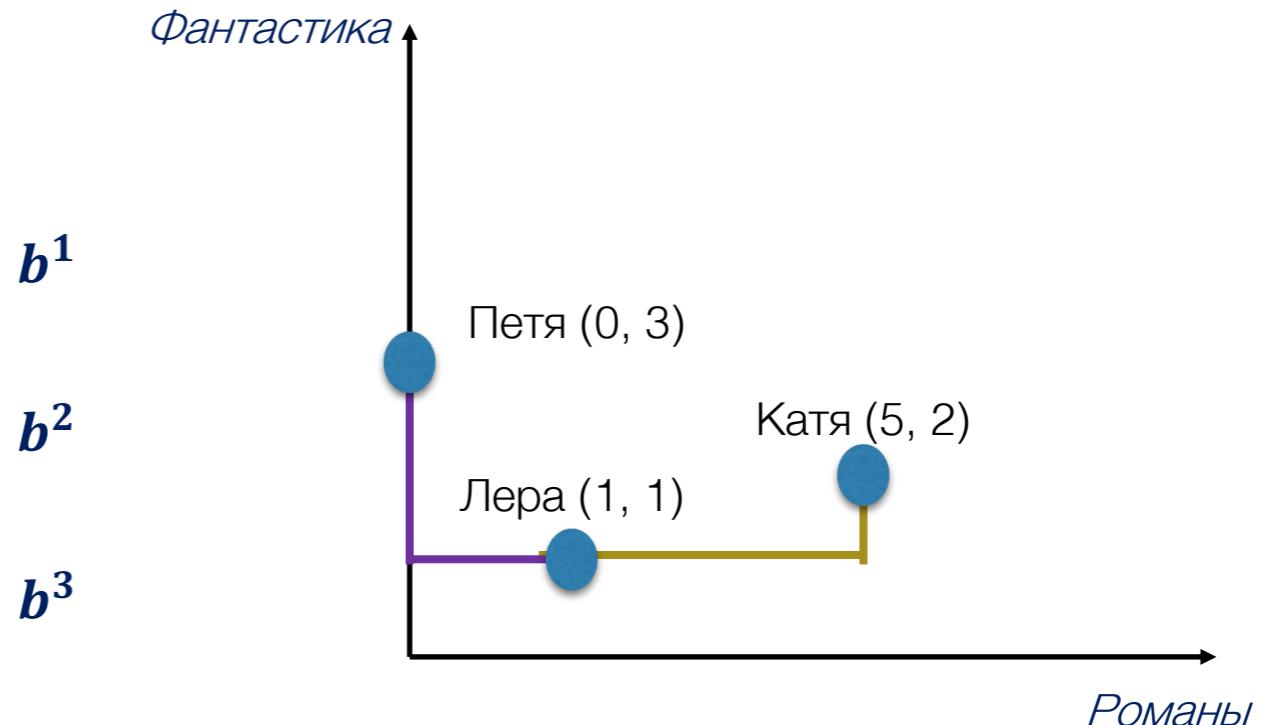


- $d(b^1, b^2) = \sum_{i=1}^n |b_i^1 - b_i^2|$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

- $d(b^1, b^2) = |5 - 1| + |2 - 1| = 5$

РАССТОЯНИЕ ХЭММИНГА (МАНХЕТТЕН)

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

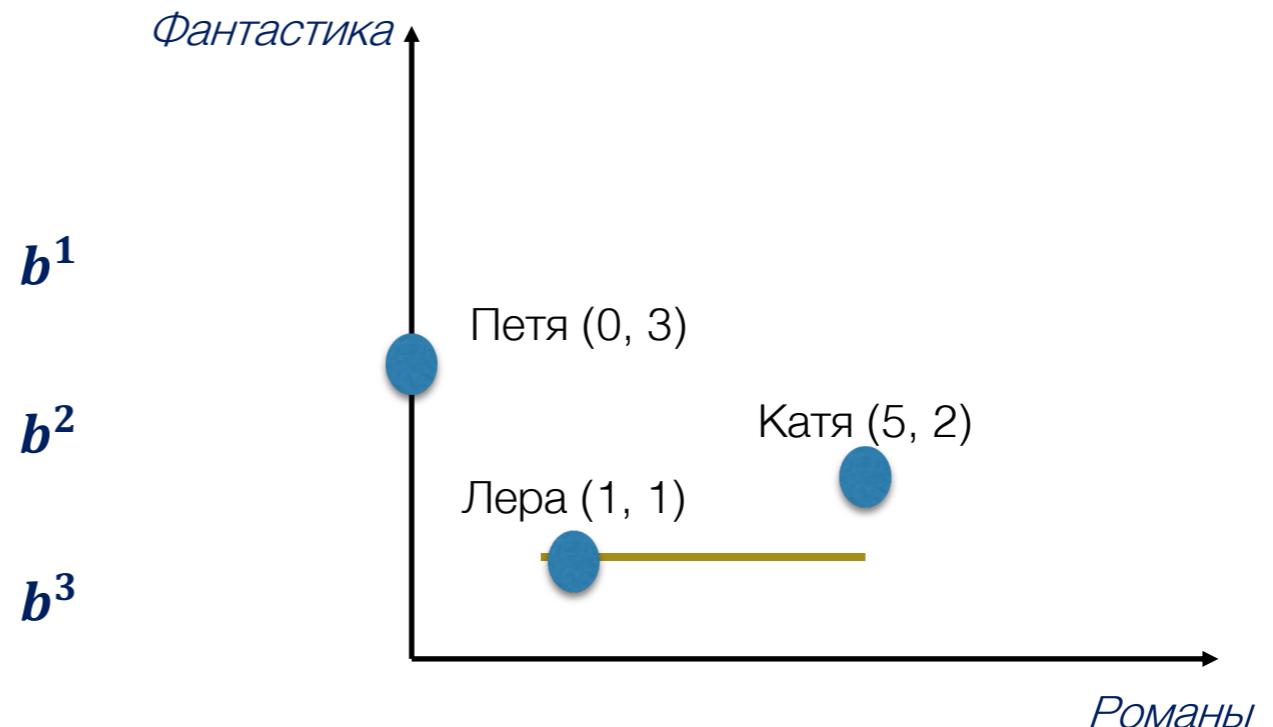


- $d(b^1, b^2) = \sum_{i=1}^n |b_i^1 - b_i^2|$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

- $d(b^1, b^2) = |5 - 1| + |2 - 1| = 5$
- $d(b^2, b^3) = |1 - 0| + |1 - 3| = 3$

РАССТОЯНИЕ ЧЕБЫШЕВА

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3

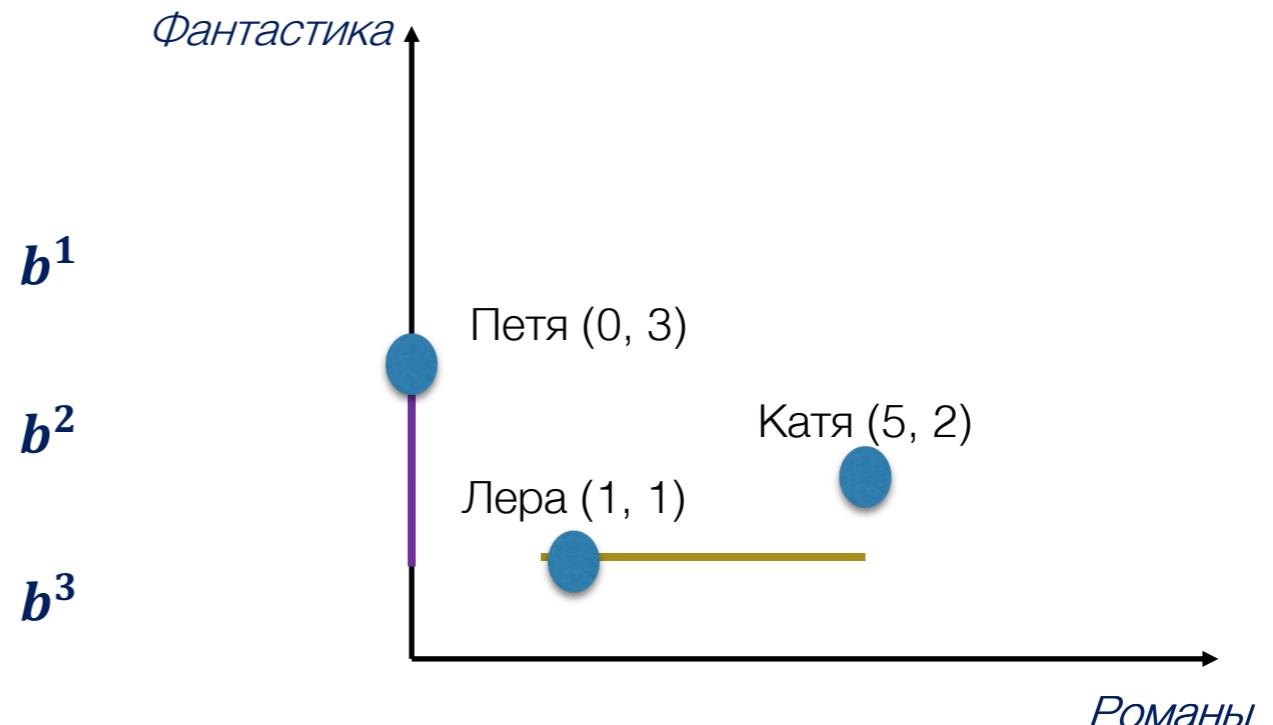


- $d(b^1, b^2) = \max\{|b_i^1 - b_i^2|\}$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

- $d(b^1, b^2) = \max\{|5 - 1|, |2 - 1|\} = 4$

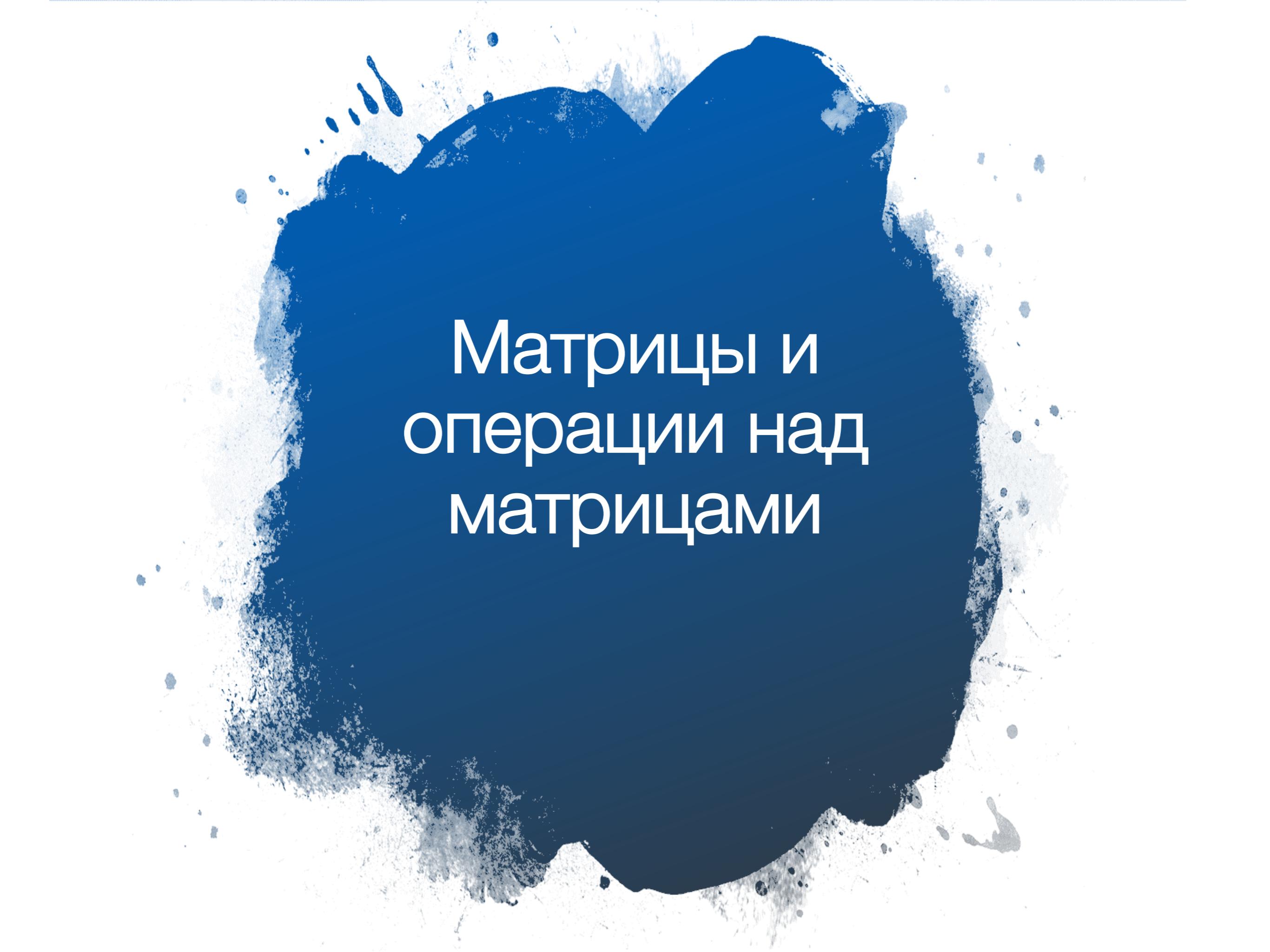
РАССТОЯНИЕ ЧЕБЫШЕВА

	Романы	Фантастика
Катя	5	2
Лера	1	1
Петя	0	3



- $d(b^1, b^2) = \max\{|b_i^1 - b_i^2|\}$
- d – расстояние
- b^1 - все покупки человека под №1
- b_i^1 - количество книг жанра i человека под №1

- $d(b^1, b^2) = \max\{|5 - 1|, |2 - 1|\} = 4$
- $d(b^2, b^3) = \max\{|1 - 0|, |1 - 3|\} = 2$



Матрицы и операции над матрицами

МАТРИЦА

- Таблица с данным размером m на n

$$A = \begin{pmatrix} 2 & 2 & -3 \\ 1 & 3 & 3 \\ 3 & -1 & 1 \end{pmatrix}$$

ТРАНСПОНИРОВАНИЕ МАТРИЦЫ

$$A^T = \begin{pmatrix} 3 & -2 & -1 \\ 1 & 3 & 2 \\ 5 & -2 & 4 \end{pmatrix}^T = \begin{pmatrix} 3 & 1 & 5 \\ -2 & 3 & -2 \\ -1 & 2 & 4 \end{pmatrix}$$

СЛОЖЕНИЕ И ВЫЧИТАНИЕ МАТРИЦ

$$\mathbf{A} := \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \qquad \mathbf{B} := \begin{pmatrix} 1 & 0 & 0 \\ -1 & -3 & -4 \end{pmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 2 & 2 & 3 \\ 3 & 2 & 2 \end{pmatrix}$$

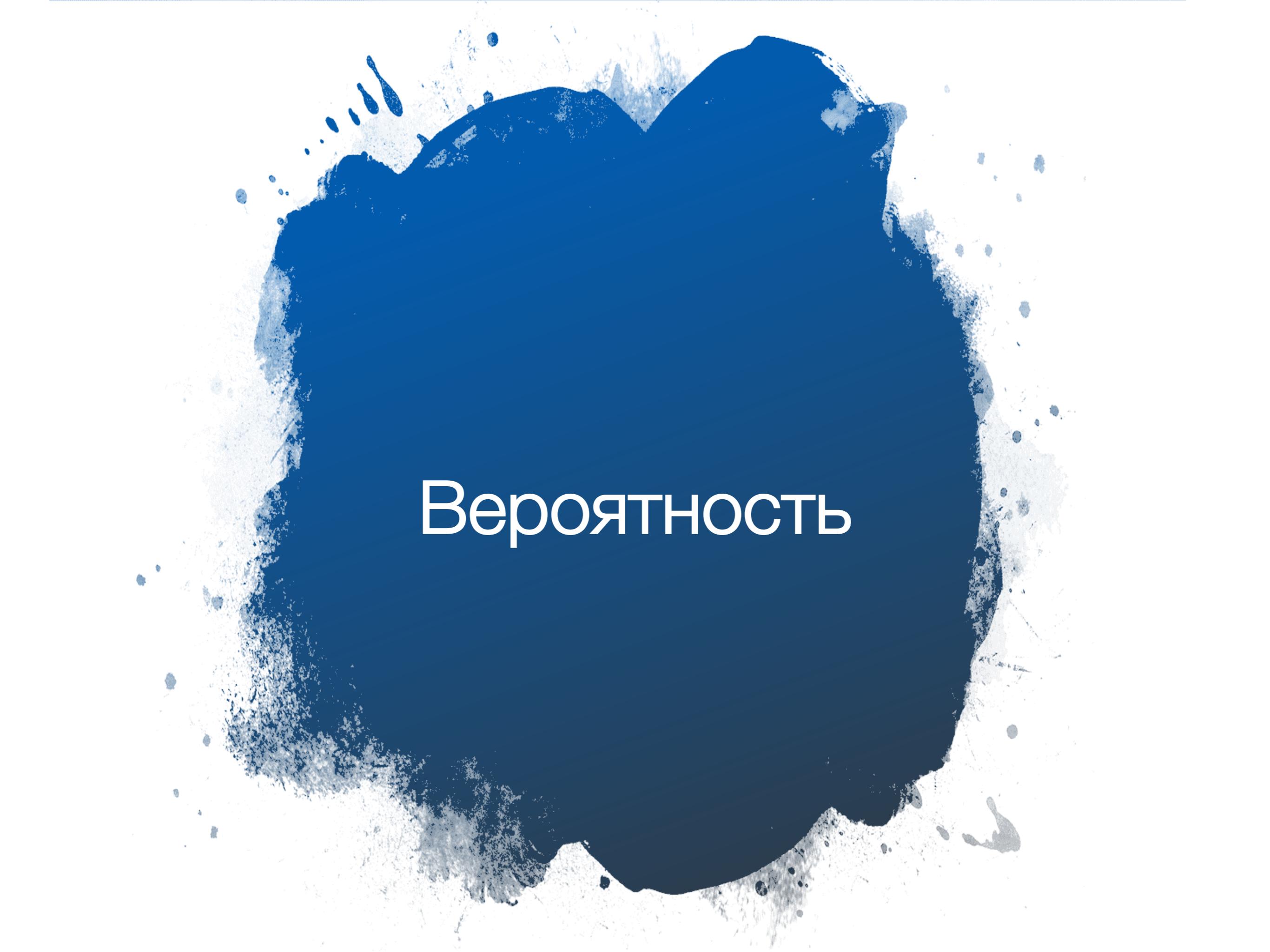
$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} 0 & 2 & 3 \\ 5 & 8 & 10 \end{pmatrix}$$

УМНОЖЕНИЕ МАТРИЦ

$$AB = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3 & 4 & 5 \\ 6 & 0 & -2 \\ 7 & 1 & 8 \end{pmatrix} = \\ = \begin{pmatrix} 36 & 7 & 25 \\ -4 & 3 & -3 \end{pmatrix}$$

ГДЕ НАМ НУЖНЫ МАТРИЦЫ

- Таблица с данными – это матрица



Вероятность

ВЕРОЯТНОСТЬ

- Научное определение сложное, отложим до лучших времен
- Интуитивно – мера уверенности
- Держим в уме классическое определение:

Вероятность случайного события А – отношение числа элементарных событий, составляющих А, к числу всех возможных событий в N

$$P(A) = \frac{n}{N}$$

ПРИМЕР

- Какова вероятность выпадения пяти очков при подбрасывании кости?

СВОЙСТВА ВЕРОЯТНОСТИ

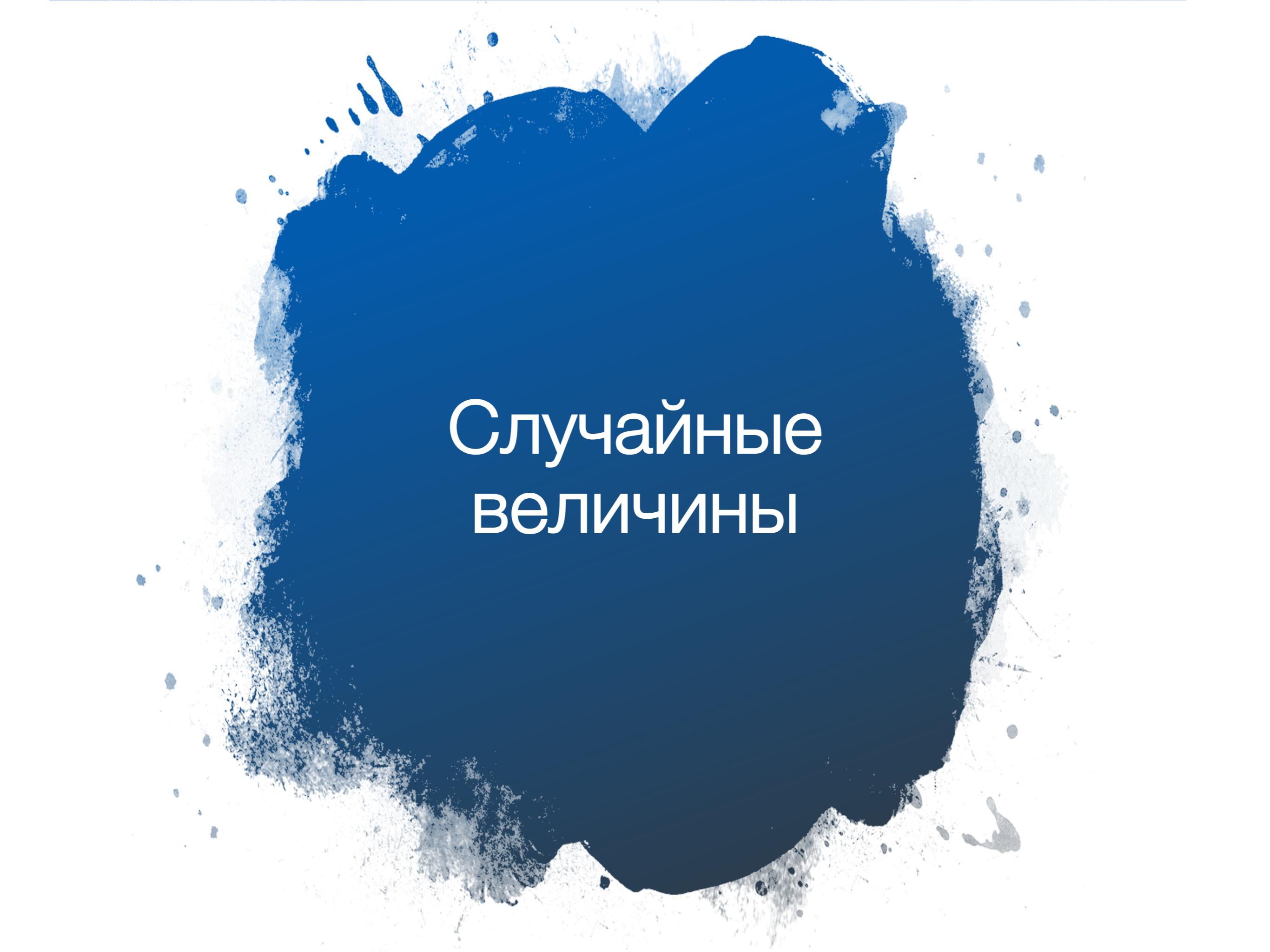
- $0 \leq P(A) \leq 1$
- $P(\bar{A}) = 1 - P(A)$, где \bar{A} - обратное событие
- $P(A + B) = P(A) + P(B) - P(AB)$
- $P(B \setminus A) = P(B) - P(AB)$
- $P(A \setminus B) = P(A) - P(AB)$
- События А и В называются независимыми, если
$$P(AB) = P(A)P(B)$$

ЗАЧЕМ НУЖНА ВЕРОЯТНОСТЬ

- Теория вероятностей и матстатистика помогают описать сложный мир простым языком
- Предсказать точный исход сложных процессов в мире сложно, а то и невозможно, но можно оценить, каковы шансы

ГДЕ НАМ ПОНАДОБИТСЯ ВЕРОЯТНОСТЬ

- Проверка гипотез
- Классификаторы



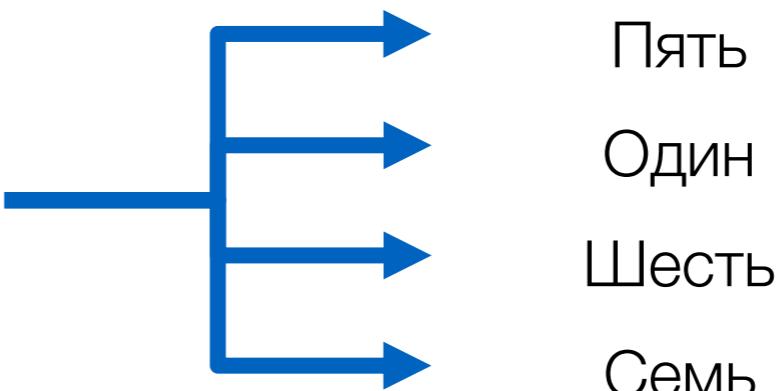
Случайные величины

СЛУЧАЙНАЯ ВЕЛИЧИНА

- Случайная величина - это множество исходов какого-то процесса
- Черный ящик, из которого можно вытащить разные варианты



Случайная
величина



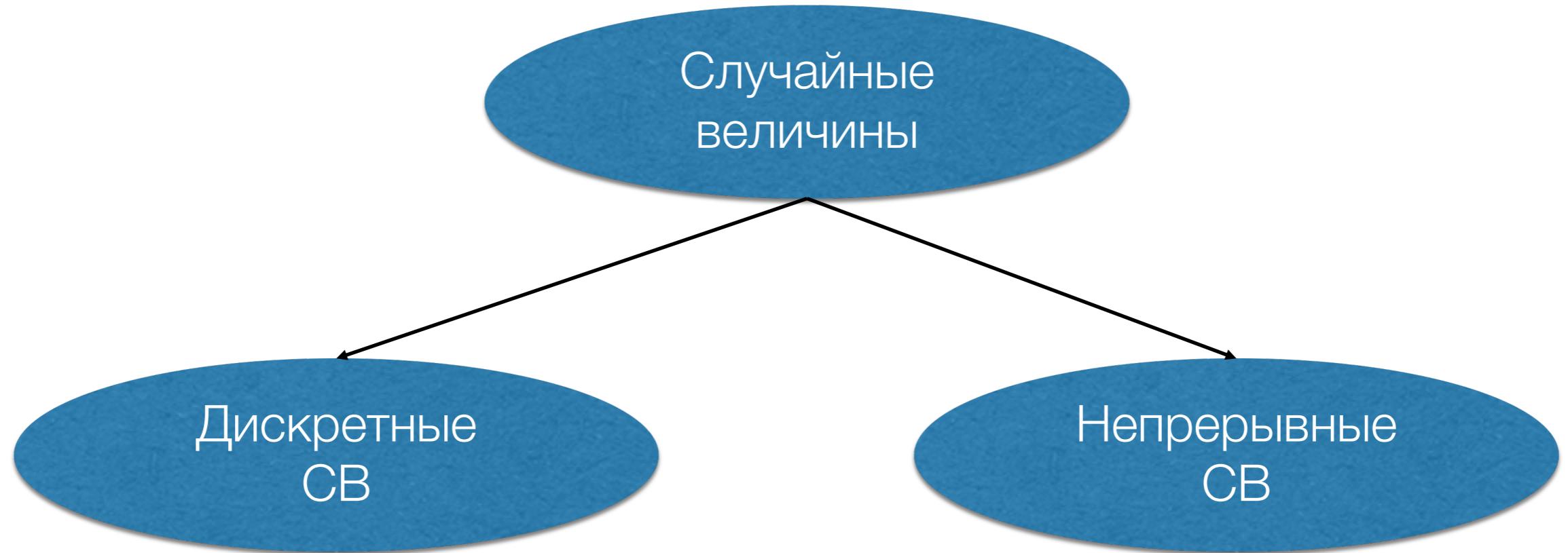
Реализация
случайной величины

РАСПРЕДЕЛЕНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН

- Правило, которое описывает, какие значения может принимать случайная величина и с какой вероятностью

Орел	Решка
0.3	0.7

ВИДЫ СВ



СВ можно принимать
конечное число
значений

СВ может принимать
любое значение из
интервала

РАВНОМЕРНОЕ РАСПРЕДЕЛЕНИЕ

- Все исходы равновероятны

$$X = (x_1, x_2, \dots, x_n)^T$$

$$P(x_i) = \frac{1}{n}$$

РАСПРЕДЕЛЕНИЕ БЕРНУЛИ

- Два исхода: успех или не успех

$$P(A) = p$$

$$P(B) = 1 - p = q$$

Орел	Решка
0.3	0.7

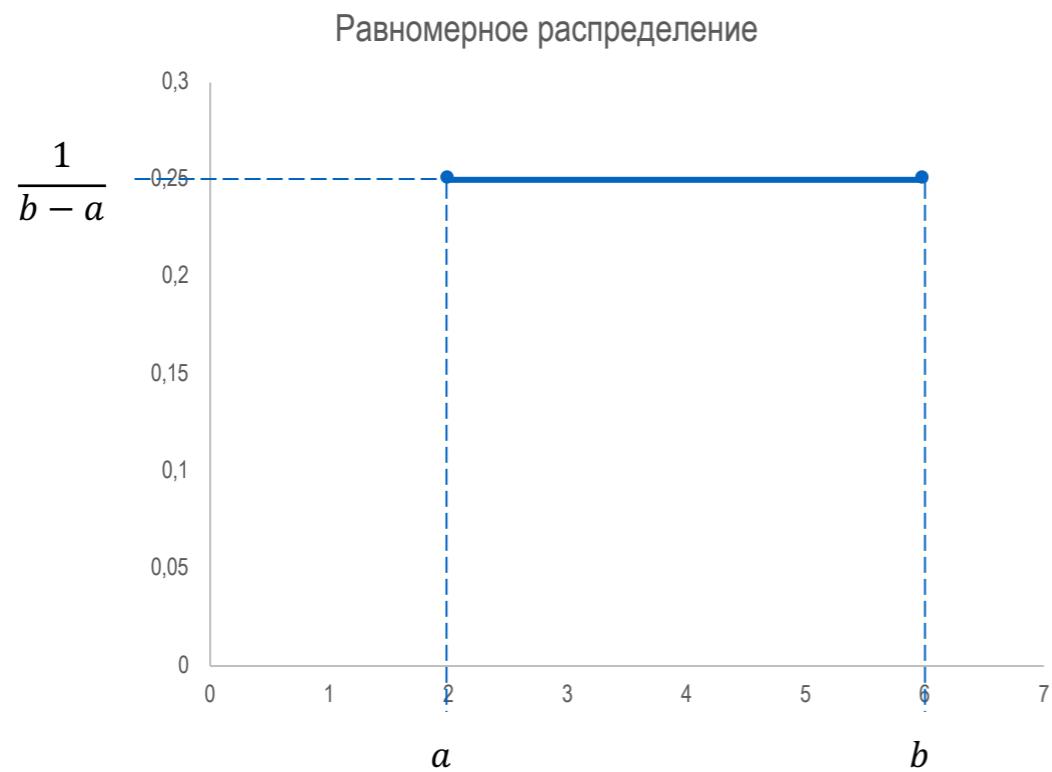
БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

- Распределение количества «успехов» в последовательности из n независимых экспериментов с вероятностью успеха в каждом p

$$P(X = k) = C_n^k p^k (1 - p)^{n-k}$$

$$C_n^k = \frac{n!}{k! (n - k)!}$$

НЕПРЕРЫВНОЕ РАВНОМЕРНОЕ РАСПРЕДЕЛЕНИЕ



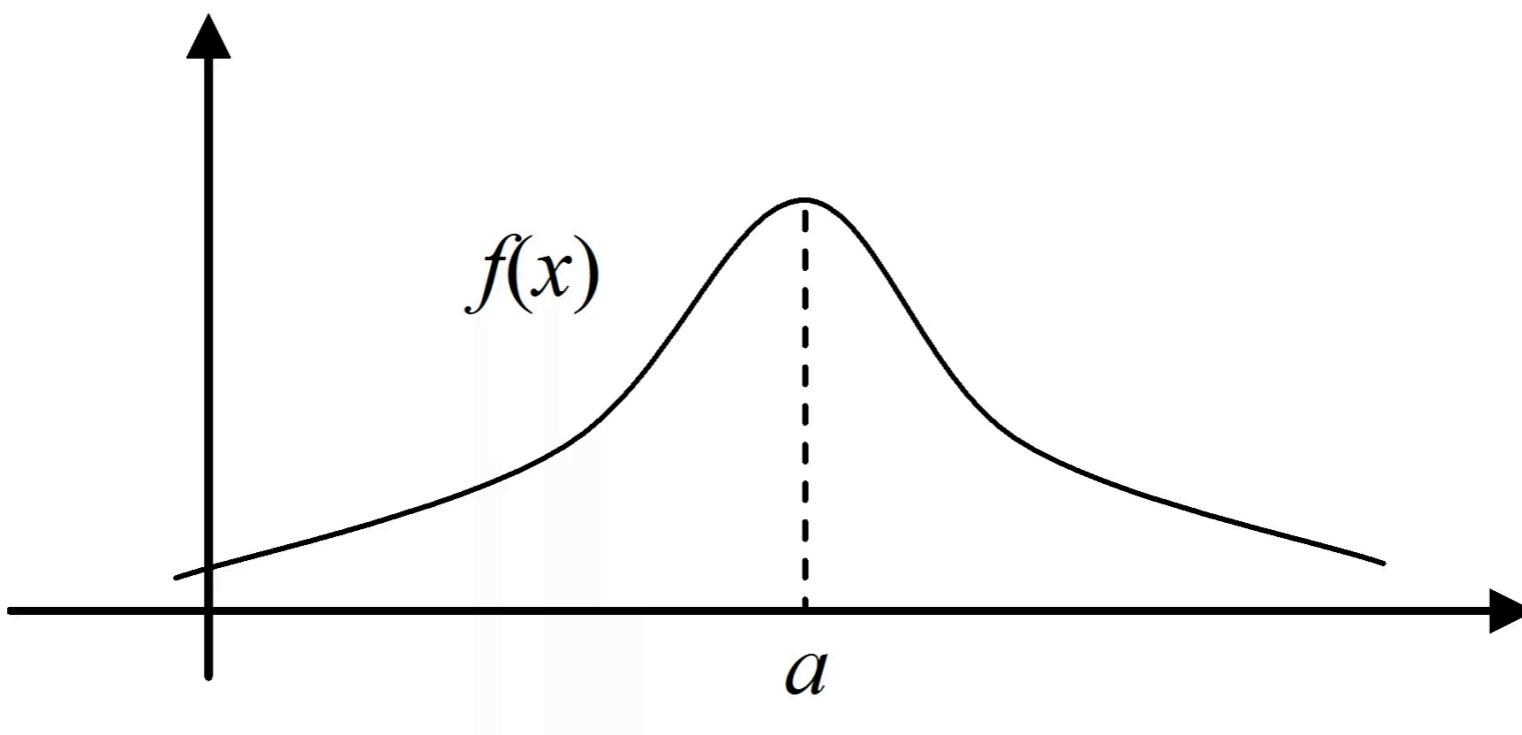
$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}$$

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ — среднее значение

σ^2 — дисперсия



МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ

Очень грубо – «среднее» значение

$$E[X] = \begin{cases} \sum_i a_i p_i, & X \text{ -- дискретная величина} \\ \int_{-\infty}^{+\infty} xf(x)dx, & X \text{ -- непрерывная величина} \end{cases}$$

ДИСПЕРСИЯ

Показатель «рассеивания» величины

$$D[X] = E[(X - E[X])^2]$$



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ