курс | data analysis | 8 недель

Математика и статистика для анализа данных

r_d

Теория множеств и линейная алгебра

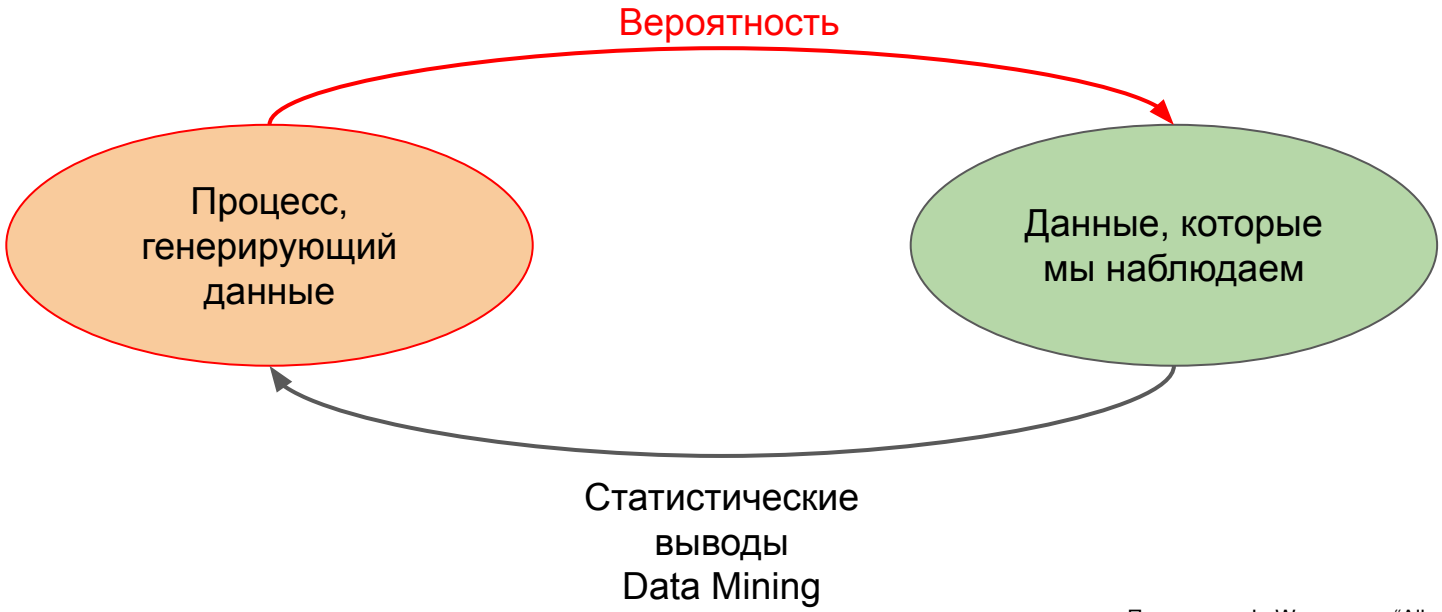
- Основные понятия теории множеств
- Множества, как основа математики
- Диаграммы Венна
- Операции над множествами
- Повторение дифференциального и интегрального исчисления
- Матрицы и векторы

r_d

Не Паникуй!

Дуглас Адамс – Автостопом по Галактике

План курса



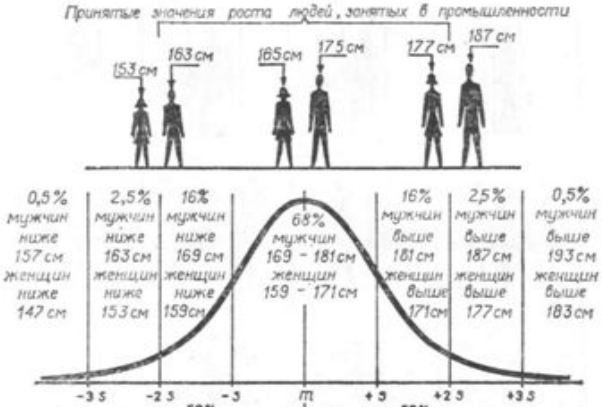
По мотивам L. Wasserman “All of Statistics”

Повторение математики

Зачем?

Картинку с кубиками можно заменить. Например, на что-то в таком ключе https://www.shutterstock.com/ru/image-vector/dice-cubes-on-white-background-vector-1034612269

- *Нам нужны ответы на такие вопросы:*
  - *Допустим, мы решим, что каждая сторона кости выпадает с вероятностью ⅙ . Как понять, с какой вероятностью цифра 5 выпадет на первом броске, если мы кидаем кости 2 раза?*
  - *Допустим, мы знаем, как распределена высота населения, то есть какой процент населения имеет какой рост в определенном интервале. Какая вероятность встретить случайно человека с ростс больше 195 см?*
  - *Допустим, мы знаем вероятности, что в очереди за IPhone мы будем ждать меньше 10, 20, 30, 40, 50 мин Какая вероятность, что мы будем ждать 27 ± 2 мину*



Было бы очень здорово перерисовать эту картинку. Можно только людей с ростом и кривую без текста с процентами.

Зачем?

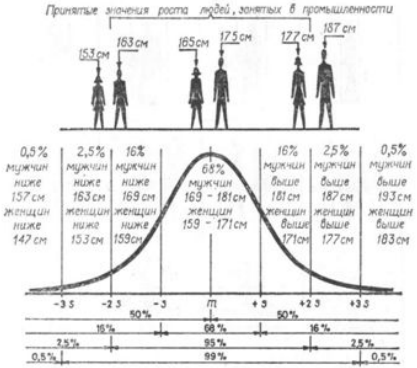
- *Нам нужны ответы на такие вопросы:*
  - *Допустим, мы решим, что каждая сторона кости выпадает с вероятностью ⅙ . Как решить, с какой вероятностью цифра 5 выпадет на первом броске, если мы кидаем кости 2 раза?*

Зачем?

- *Нам нужны ответы на такие вопросы:*
  - *По каким законам мы можем оперировать с вероятностью?*

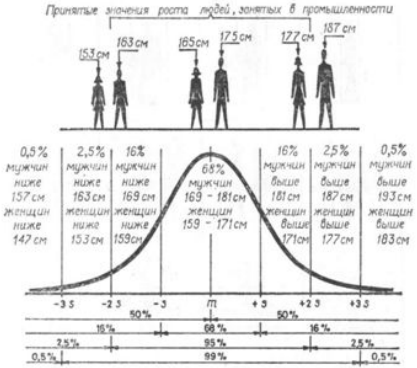
Зачем?

- *Нам нужны ответы на такие вопросы:*
  - *По каким законам мы можем оперировать с вероятностью?*
  - *Допустим, мы знаем, как распределена высота населения, то есть какой процент населения имеет какой рост в определенном интервале. Какая вероятность встретить случайно человека с ростом больше 195 см?*



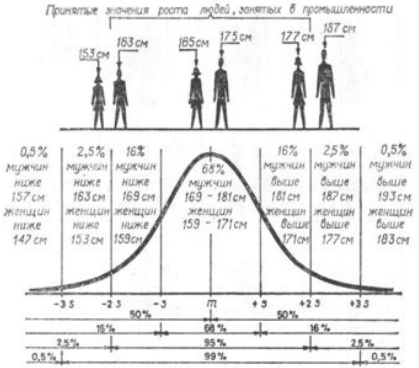
Зачем?

- *Нам нужны ответы на такие вопросы:*
  - *По каким законам мы можем оперировать с вероятностью?*
  - *Как найти площадь под кривой распределения?*



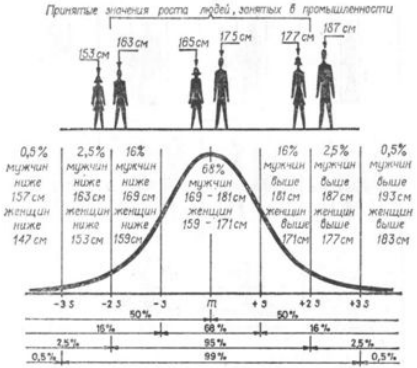
Зачем?

- *Нам нужны ответы на такие вопросы:*
  - *По каким законам мы можем оперировать с вероятностью?*
  - *Как найти площадь под кривой распределения?*
  - *Допустим, мы знаем вероятности, что в очереди за IPhone мы будем ждать меньше 10, 20, 30, 40, 50 минут. Какая вероятность, что мы будем ждать 27 ± 2 минуты?*



Зачем?

- *Нам нужны ответы на такие вопросы:*
  - *По каким законам мы можем оперировать с вероятностью?*
  - *Как найти площадь под кривой распределения?*
  - *Как найти распределение, если мы знаем площади под его кривой для различных интервалов*



Ответы

- *Мы получим ответы на эти вопросы с помощью математических инструментов.*
  - *По каким законам мы можем оперировать с вероятностью?*
  - *Как найти площадь под кривой распределения?*
  - *Как найти распределение, если мы знаем площади под его кривой для различных интервалов*

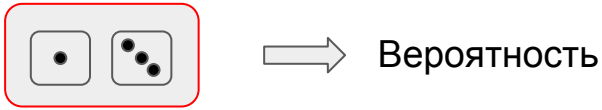
Ответы

- *Мы получим ответы на эти вопросы с помощью математических инструментов.*
  - *Математическая вероятность на основе Теории множеств*
  - *Интегральное исчисление*
  - *Дифференциальное исчисление*

Теория Множеств

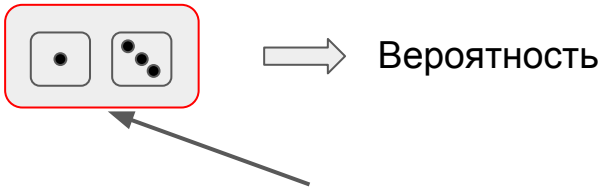
Что это и зачем?

- Наша цель – говорить о результатах случайных процессов (событиях), которые нас интересуют и присваивать им вероятности.
  - Как упадут кости на следующем броске?
  - Кто выиграет на выборах?
  - Какой мой шанс выиграть в лото?
- Сейчас вероятность для нас все еще слова, которые мы наивно используем…
- Фокус этого урока – сами результаты и их описание.



Что это и зачем?

- Наша цель – говорить о результатах случайных процессов (событиях), которые нас интересуют и присваивать им вероятности.
  - Как упадут кости на следующем броске
  - Кто выиграет на выборах
  - Какой мой шанс выиграть в лото
- Сейчас вероятность для нас все еще слова, которые мы наивно используем…
- Фокус этого урока – сами результаты и их описание



Один из возможных результатов броска 2 костей

Случайные процессы

- *Процессы, результат которых невозможно точно предсказать:*
  - *Бросок монетки или кости*
  - *Движение атмосферы и как результат - погода завтра*
  - *Выборы президента (мы не знаем точное мнение каждого человека)*
  - *Физика станка и качество каждой детали*
- *Возможные результаты процесса можно описать заранее и они взаимно исключают друг друга:*
  - *Орел или решка, 1 или 6*
  - *Четное число при броске кости*
  - *Дождик или сухой день*
  - *Клинтон или Буш*
  - *Брак или не брак*

Множества

- Мы будем использовать **множества** для описания всех возможных результатов одного случайного процесса.

Множество: *набор, **совокупность**, собрание каких-либо объектов, которые называются **элементами** этого множества и обладают общим для всех их характеристическим свойством.*

- Множество: все члены одного шахматного клуба.
- Характеристическое свойство: быть членом этого клуба, иметь членскую книжку.
- *Быть членом одного клуба не мешает быть также членом другого клуба.*

Примеры множеств

- Все результаты броска монетки:

$$S = \{O, P\}$$

- Все результаты броска кости:

$$S = \{1, 2, 3, 4, 5, 6\}$$

- Результаты броска кости:
  - Число 6

$$\mathrm{A} = \{6\}$$

  - Четное число

$$B = \{2, 4, 6\}$$

  - Число больше 2

$$C = \{3, 4, 5, 6\}$$

Примеры множеств

- Результаты выборов президента $D = \{\text{Буш}, \text{Клинтон}\}$
- Множество натуральных чисел $\mathbb{N} = \{1, 2, 3, 4, \ldots\}$
- Пустое множество $\emptyset = \{\}$
- Негативное число как результат броска кости $\emptyset$

Бесконечные множества

- Сколько элементов в множестве натуральных чисел?
- Множество вещественных чисел тоже бесконечно, но оно даже не счетное (не обязательное знание для этого курса).
- Если есть дальнейший интерес: почитайте о парадоксе Рассела.

Элементы

- Объекты, из которых состоит множество, называют **элементами** множества.

$$S = \{1, 2, 3, 4, 5, 6\}$$

  - 6-ой элемент множества результатов броска кости

$$6 \in S$$

  - 7-ой не является результатом броска кости

$$7 \notin S$$

Подмножества

- Одно множество A может быть **подмножеством** другого множества B, если все его элементы также элементы другого множества.

  - Множество результатов броска кости больше числа 2 подмножества всех результатов

$$S = \{1, 2, 3, 4, 5, 6\}$$

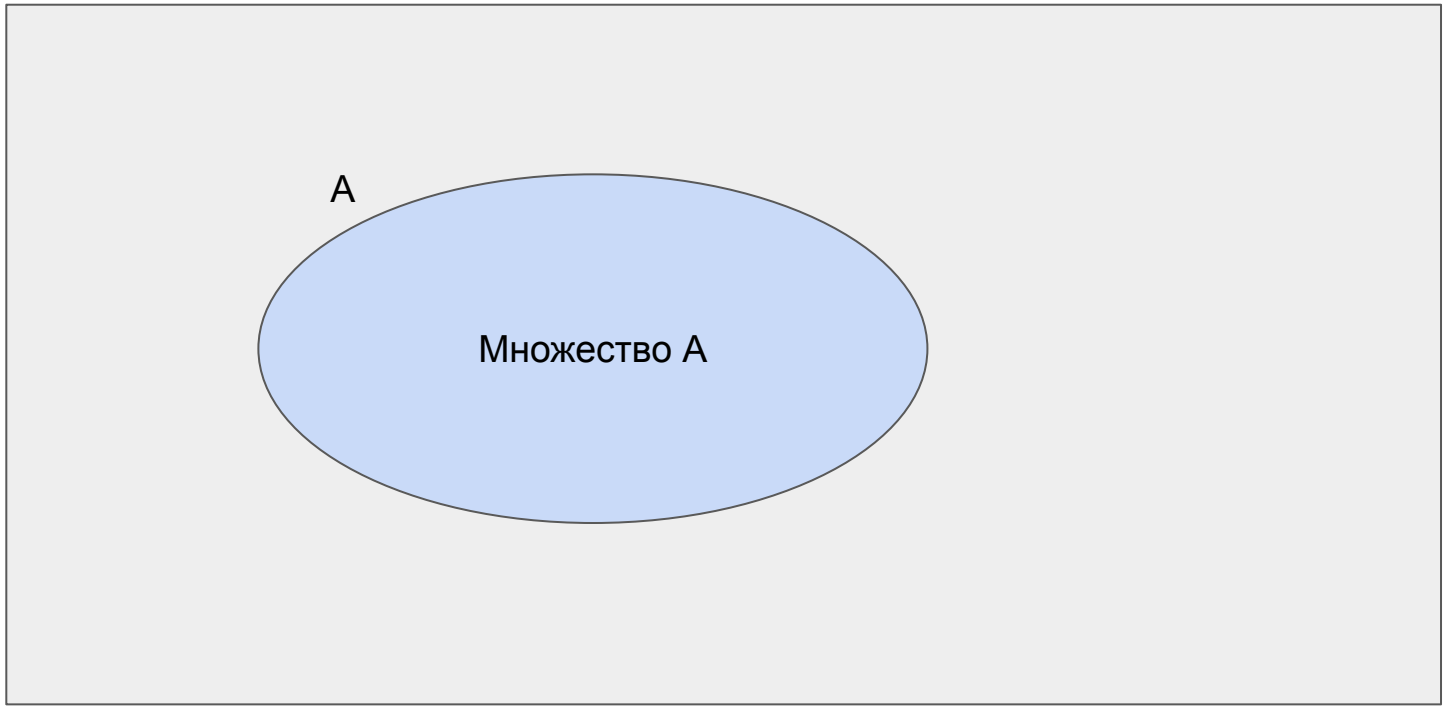
$$C = \{3, 4, 5, 6\}$$

$$\boxed{C \subset S}$$

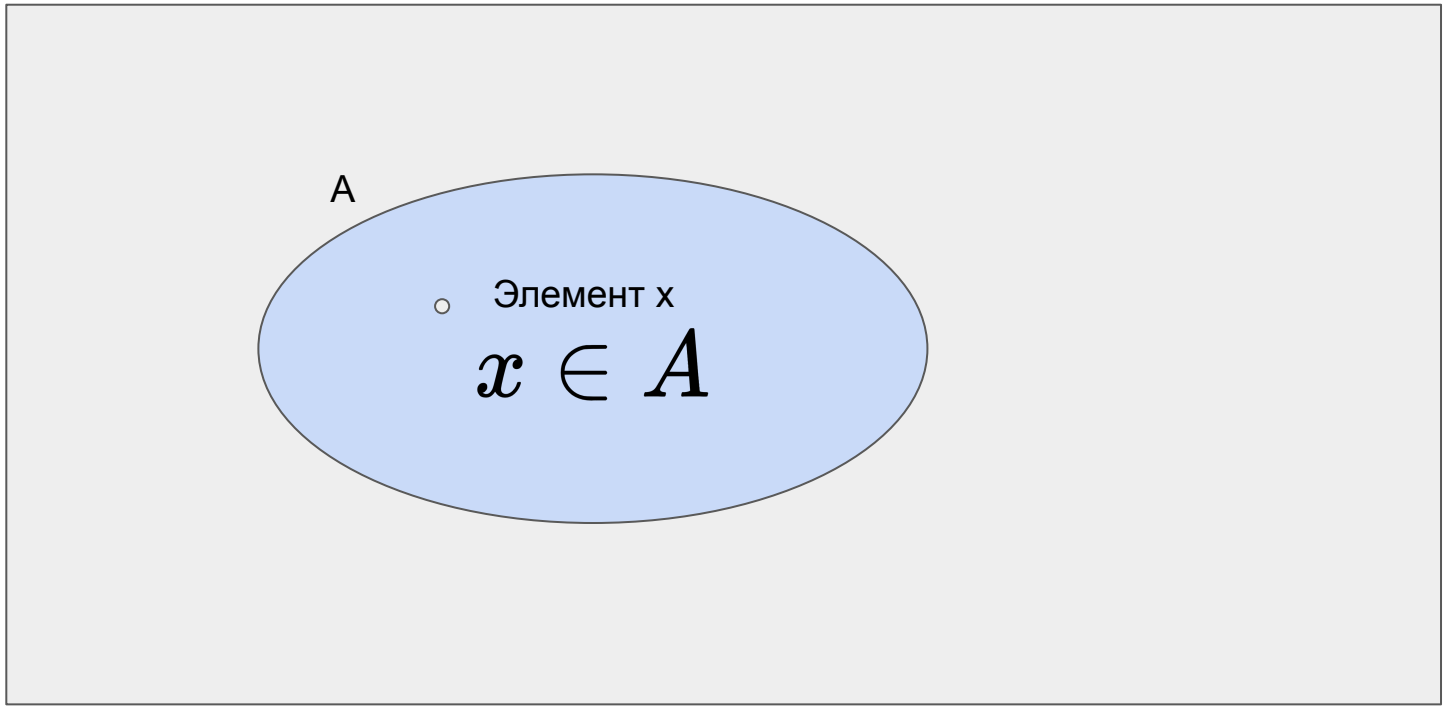
Диаграммы Венна

Настоящее или фиктивное множество, все множества, которые нас интересуют, будут считаться подмножеством этого множества

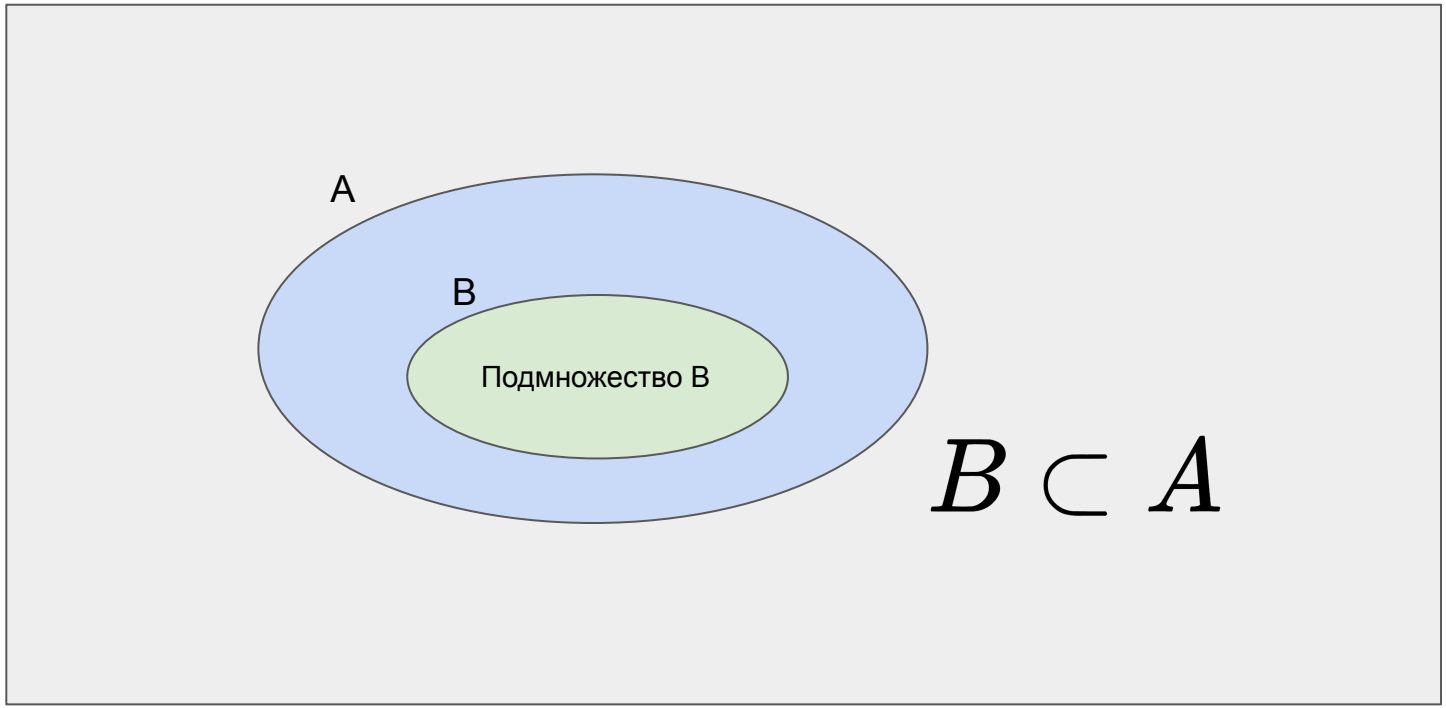
Диаграммы Венна



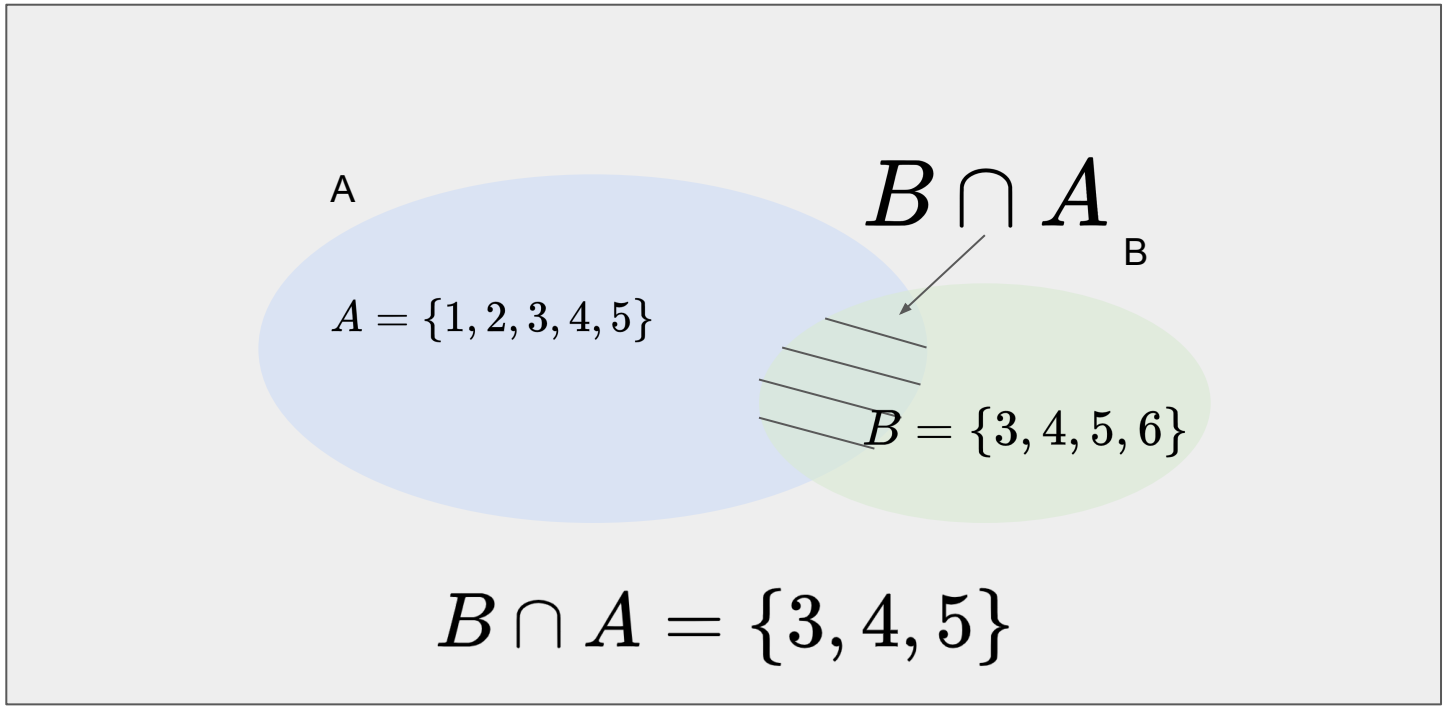
Диаграммы Венна



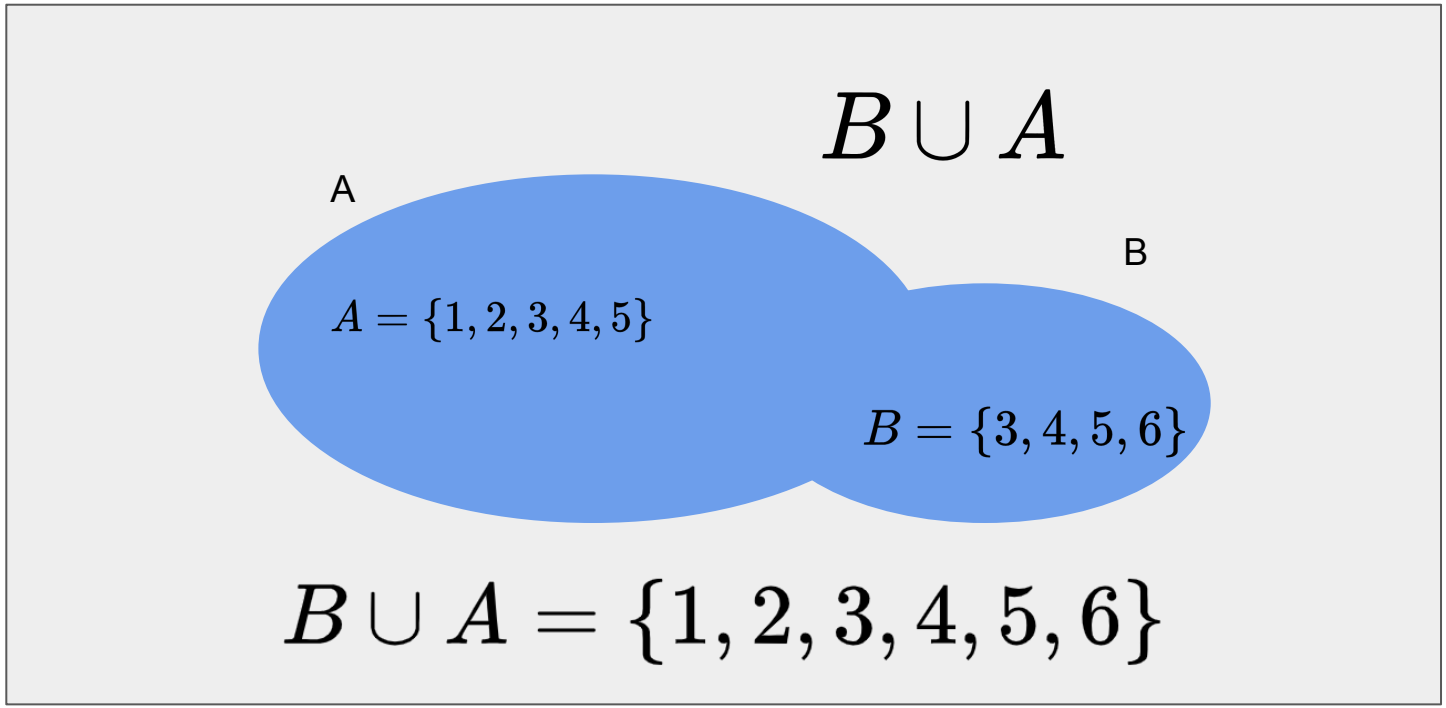
Диаграммы Венна



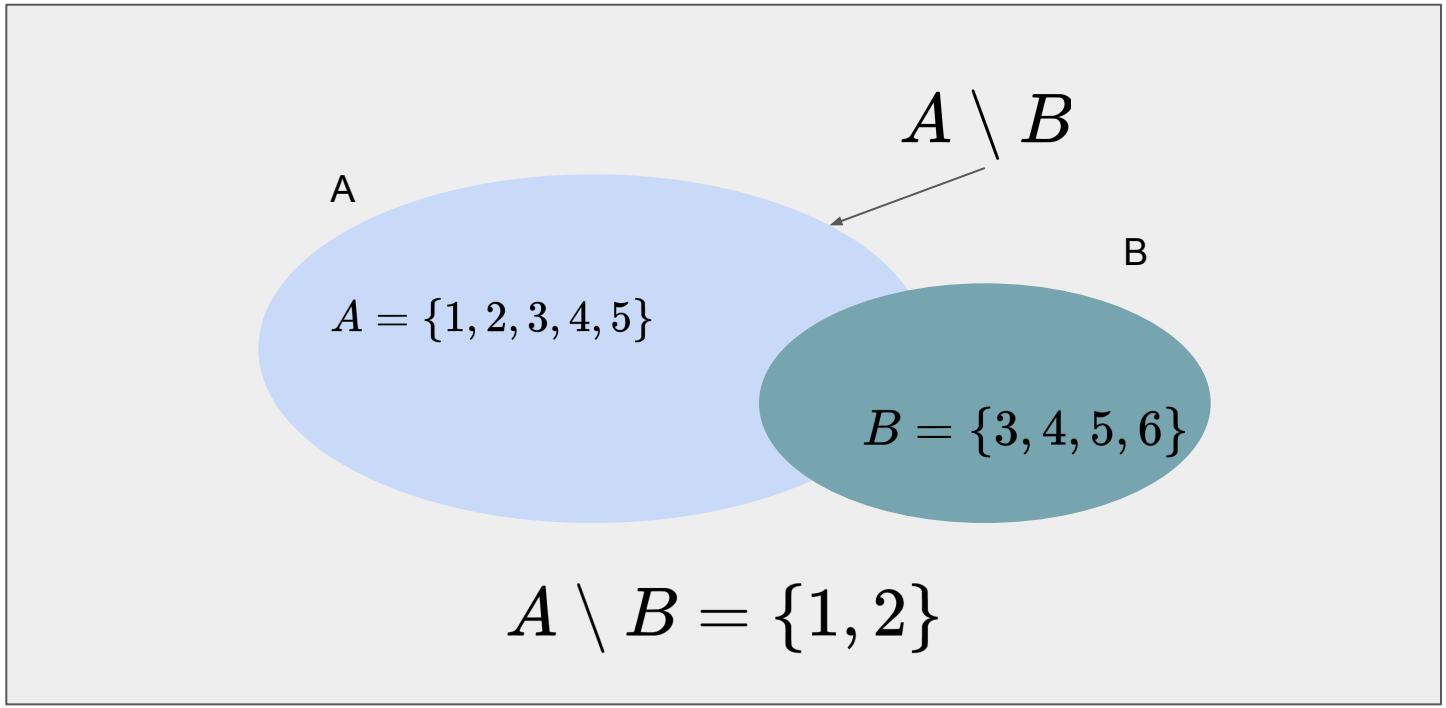
Пересечение множеств



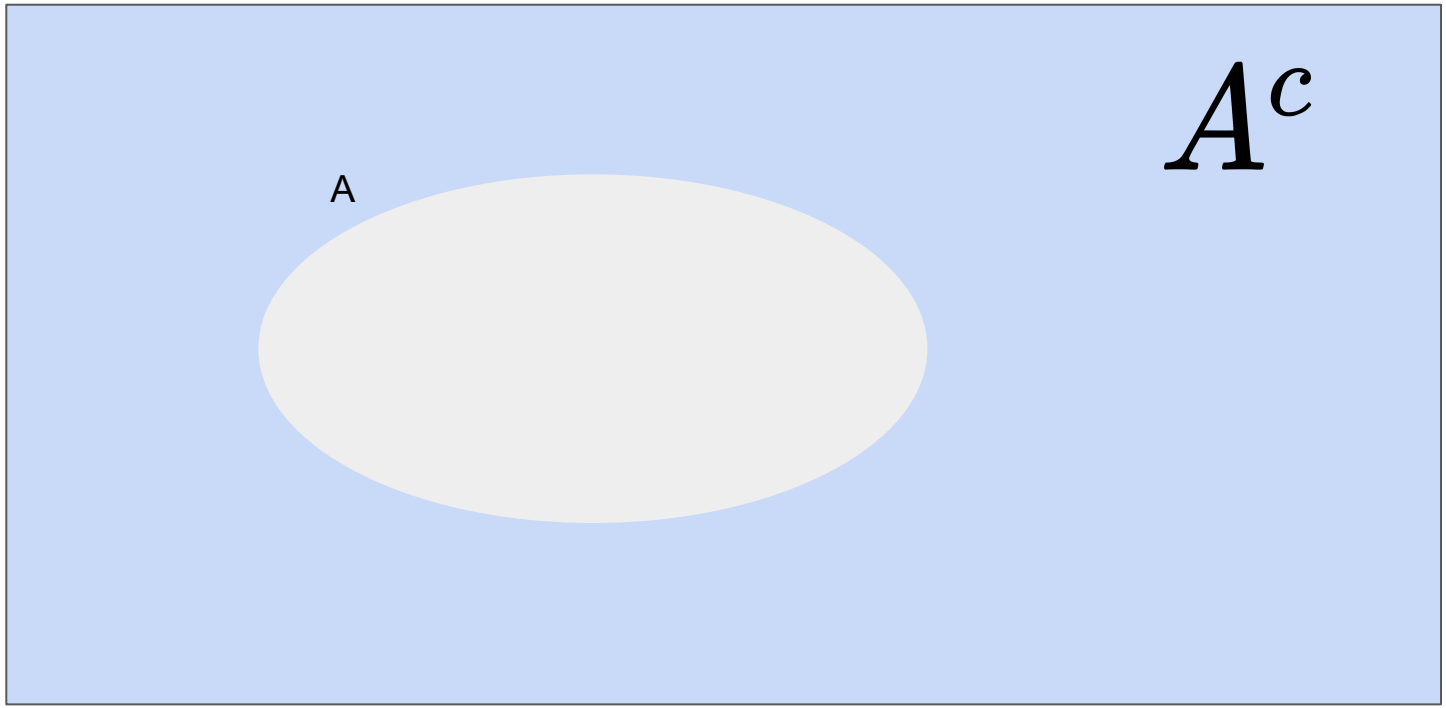
Объединение множеств



Разность множеств



Дополнение множеств



Булеан – множество всех подмножеств

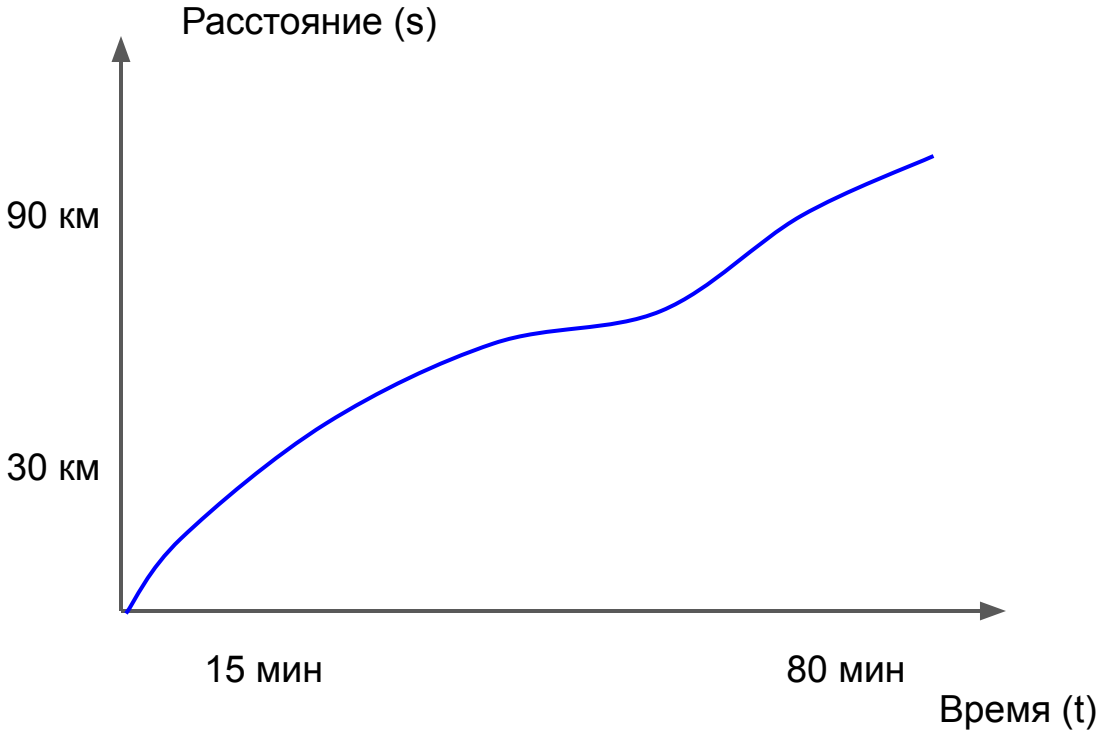
$$A = \{\mathrm{O}, \mathrm{P}\}$$

$$\mathcal{P}(A)$$

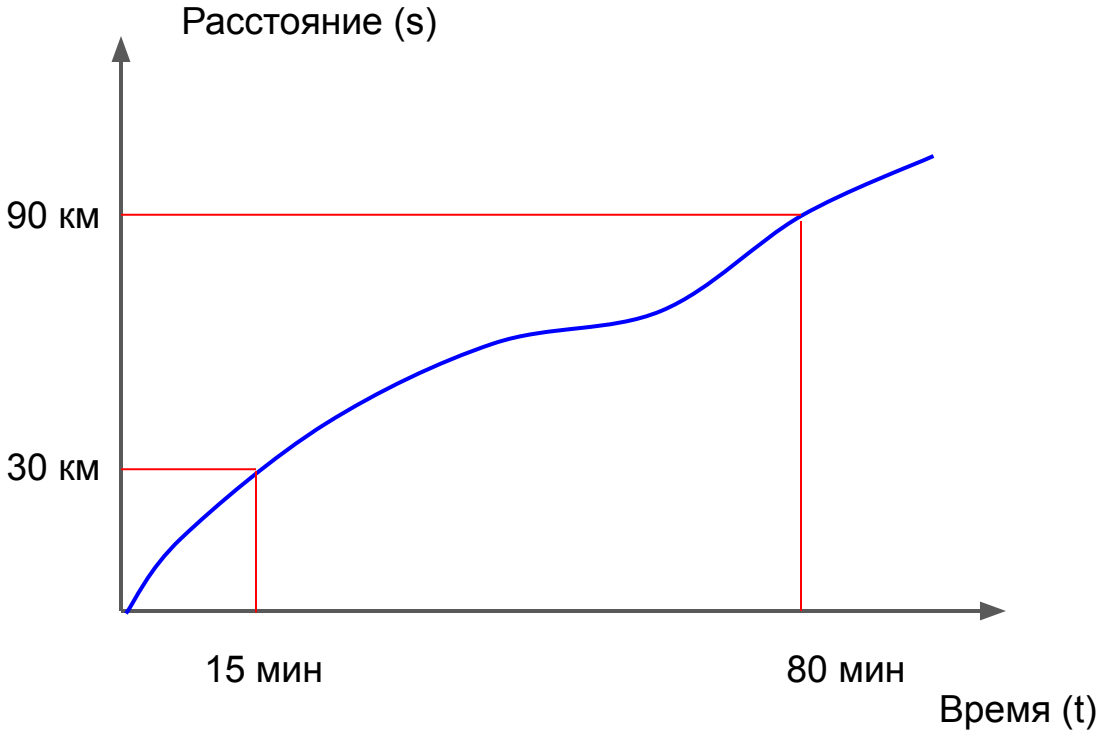
$$\mathcal{P}(A) = \{\emptyset, \{\mathrm{O}\}, \{\mathrm{P}\}, \{\mathrm{O}, \mathrm{P}\}\}$$

Дифференциальное исчисление

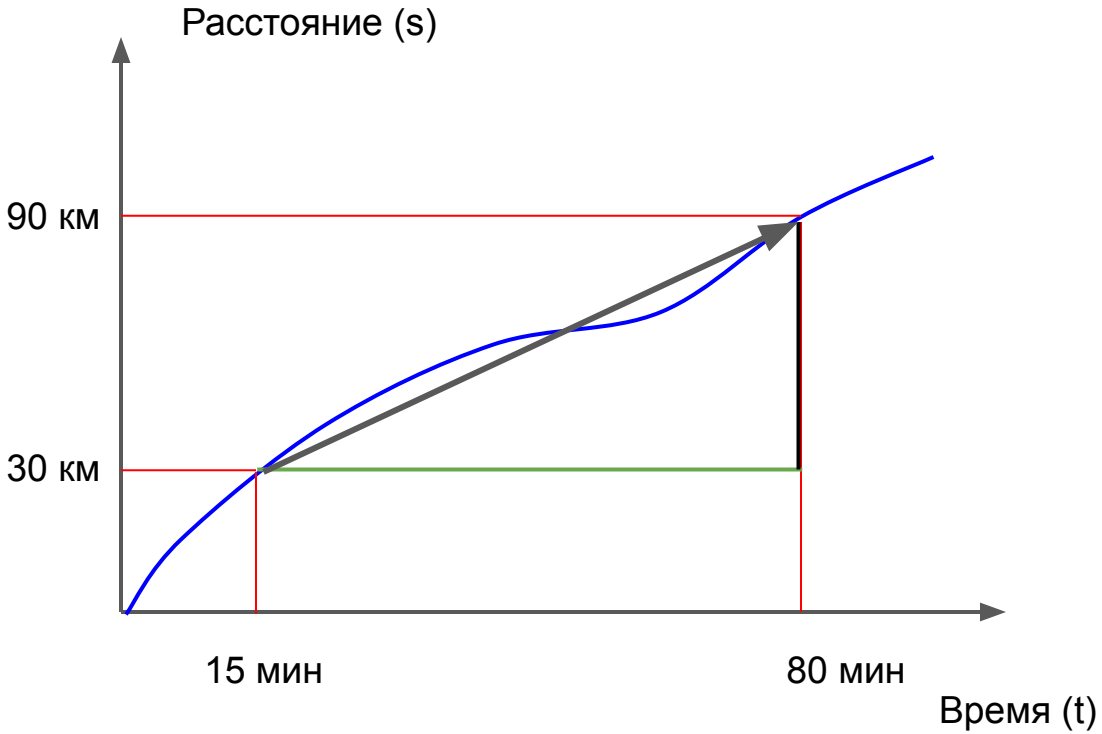
Что такое скорость?



Что такое скорость?



Что такое скорость?

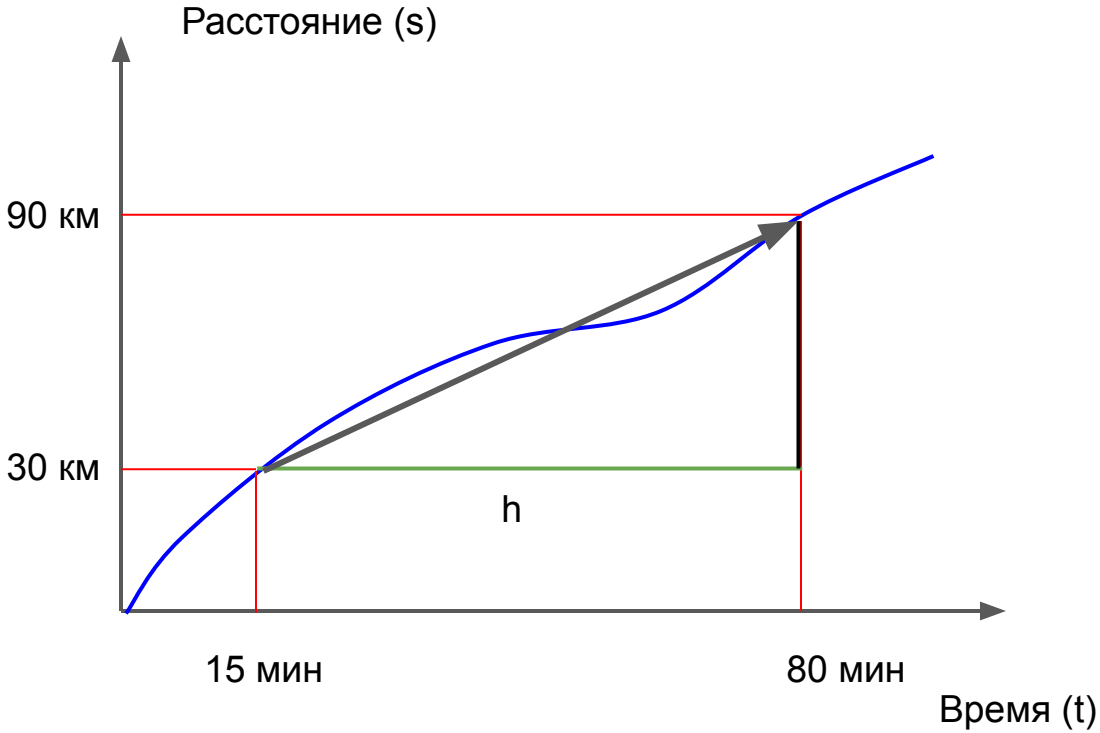


Средняя скорость

$$v = \frac{\text{Расстояние}}{\text{время}}$$

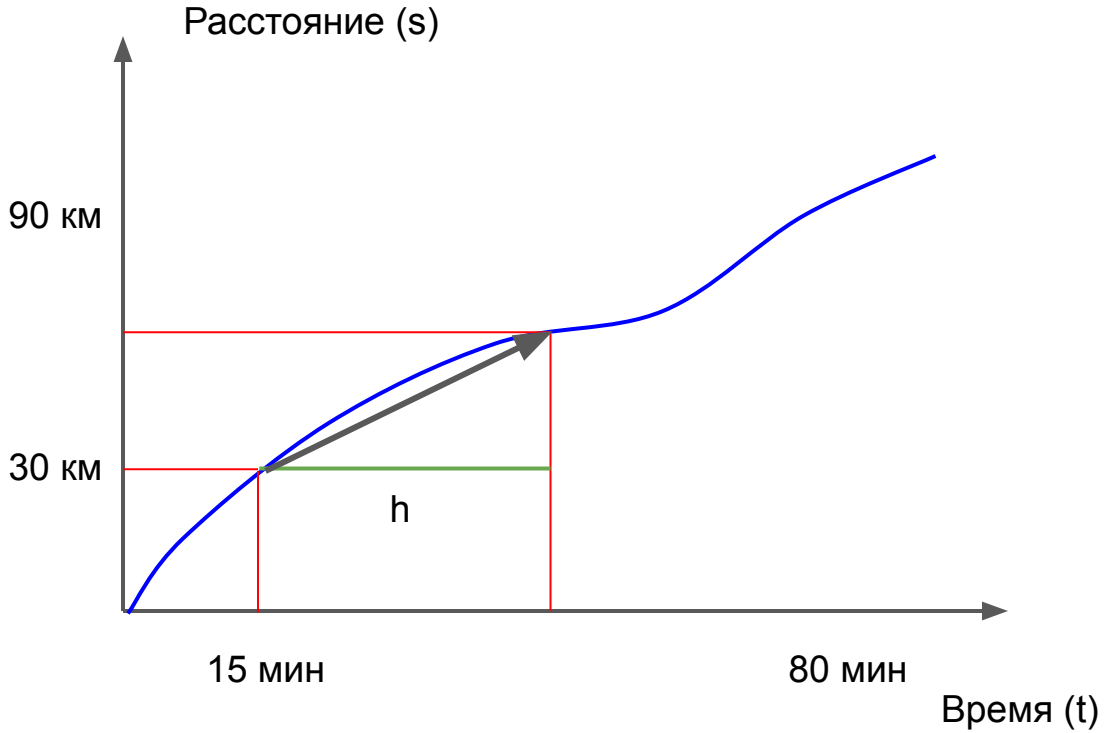
$$v = \frac{60\text{км}}{65\text{мин}} = 55\text{км}/\text{час}$$

Какая скорость машины в каждый момент времени?



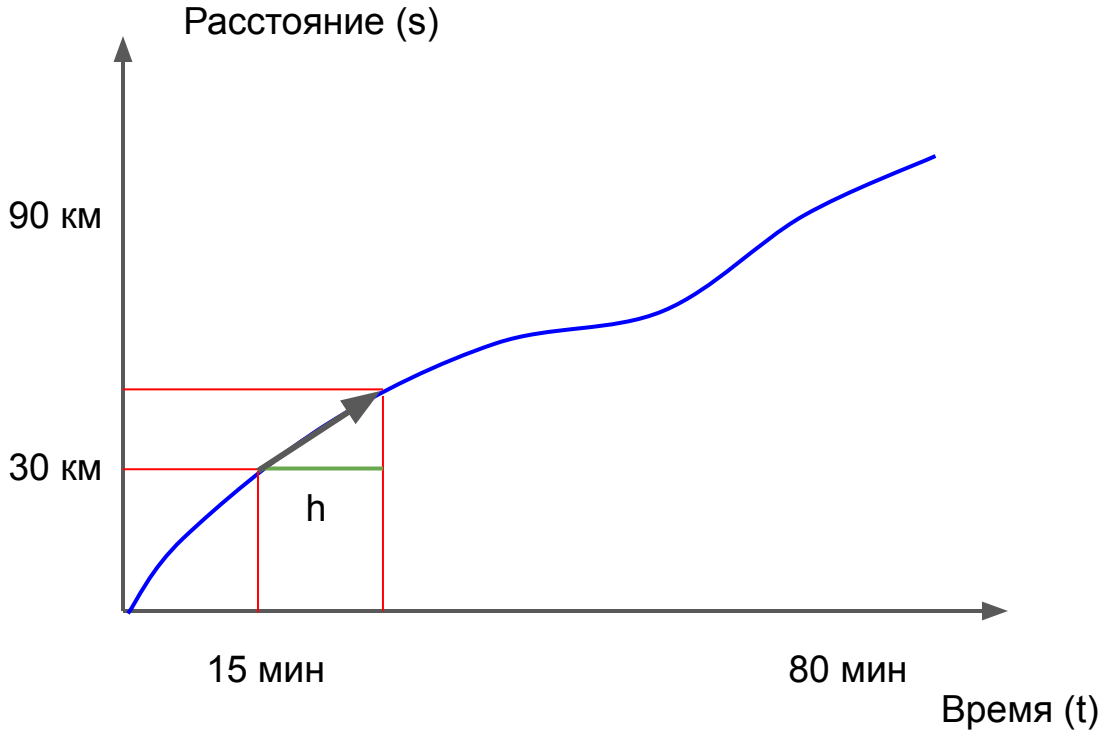
$$\lim_{h \to 0}$$

Какая скорость машины в каждый момент времени?



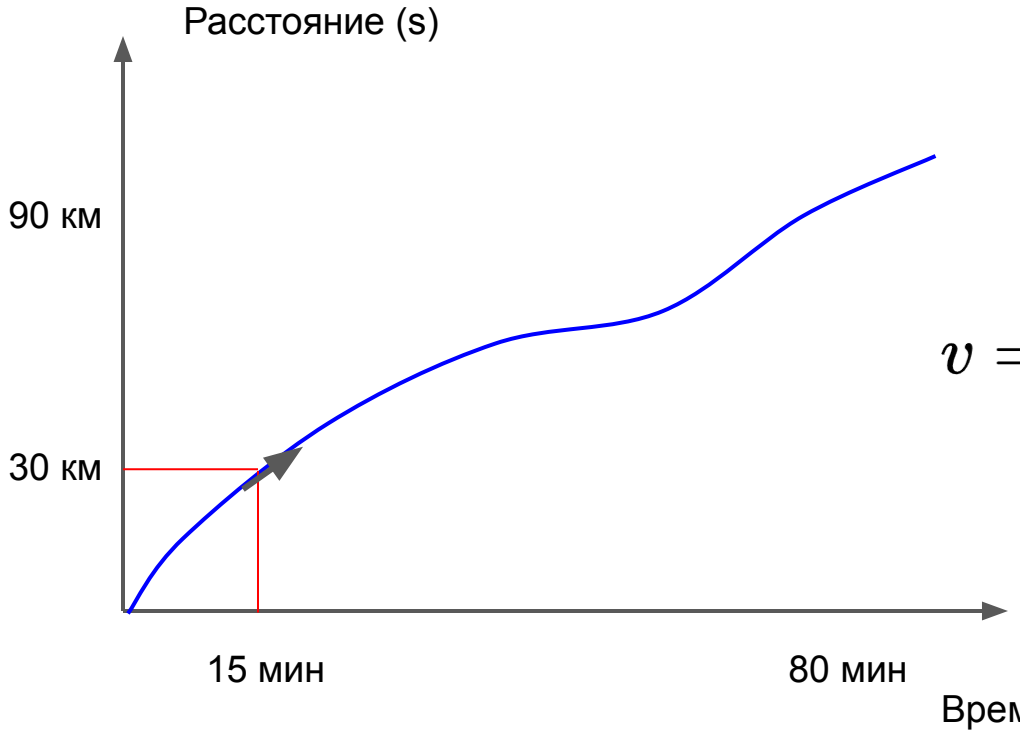
$$\lim_{h \to 0}$$

Какая скорость машины в каждый момент времени?



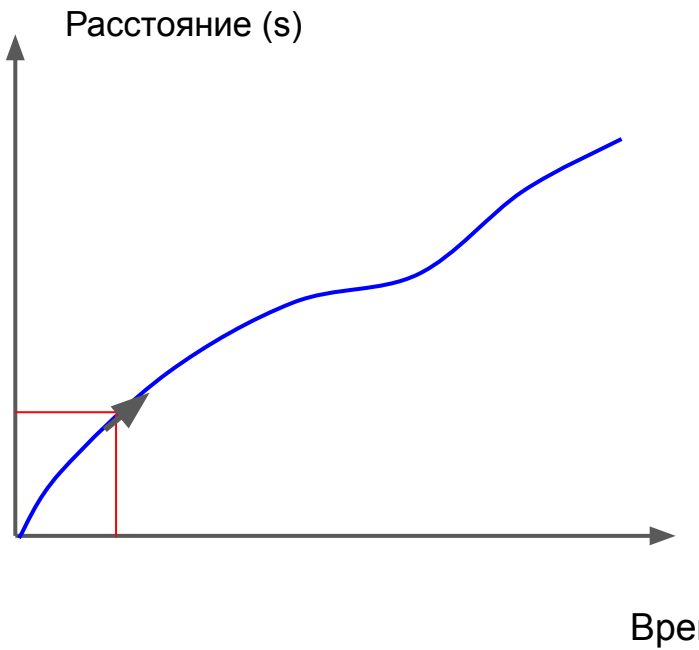
$$\lim_{h \to 0}$$

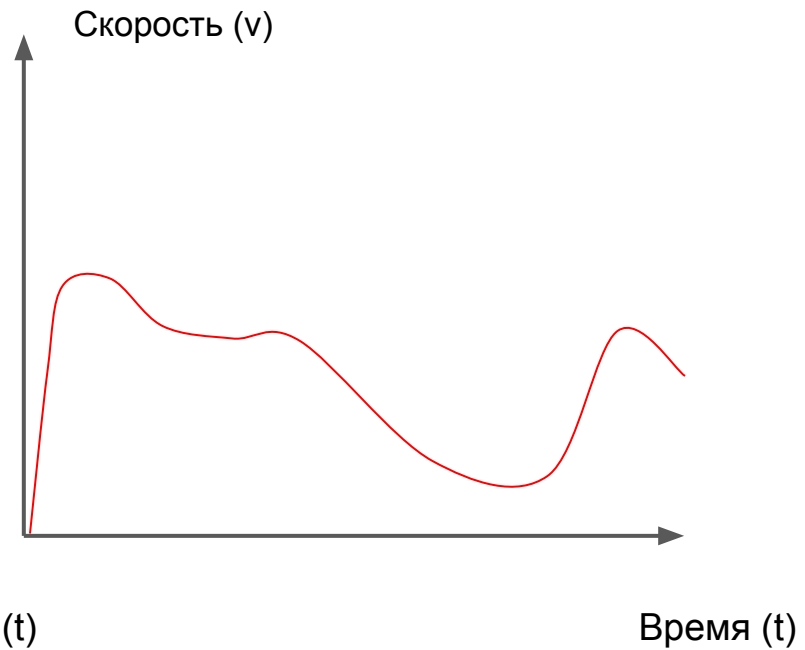
Какая скорость машины в каждый момент времени?



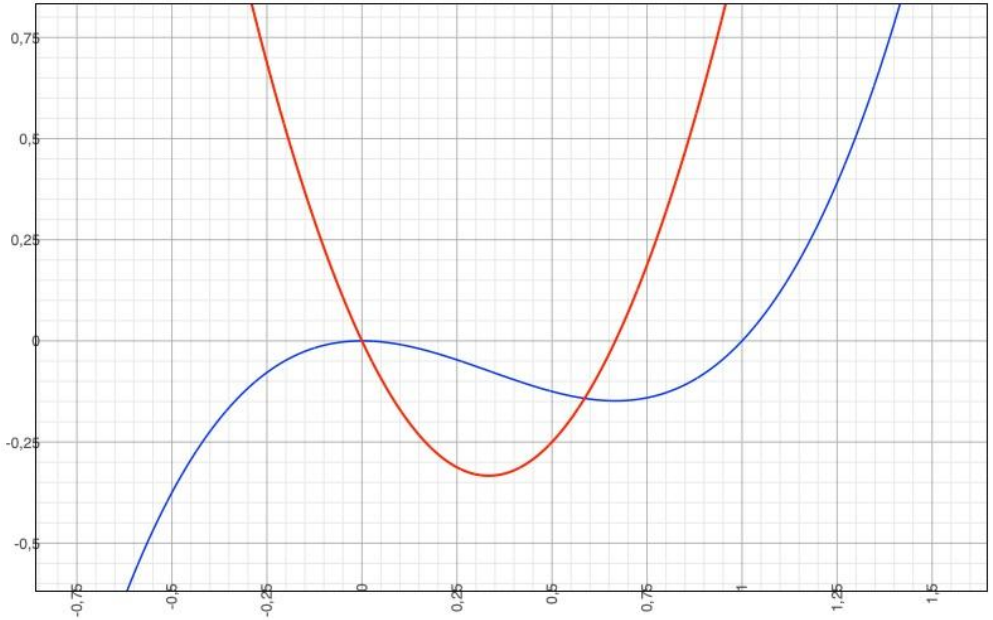
$$v = \frac{ds}{dt} = \lim_{h \to 0} \frac{s(t+h) - s(t)}{h}$$

Какая скорость машины в каждый момент времени?





Пример с функциями



$$s(t) = t^3 - t^2$$

$$v(t) = 3t^2 - 2t$$

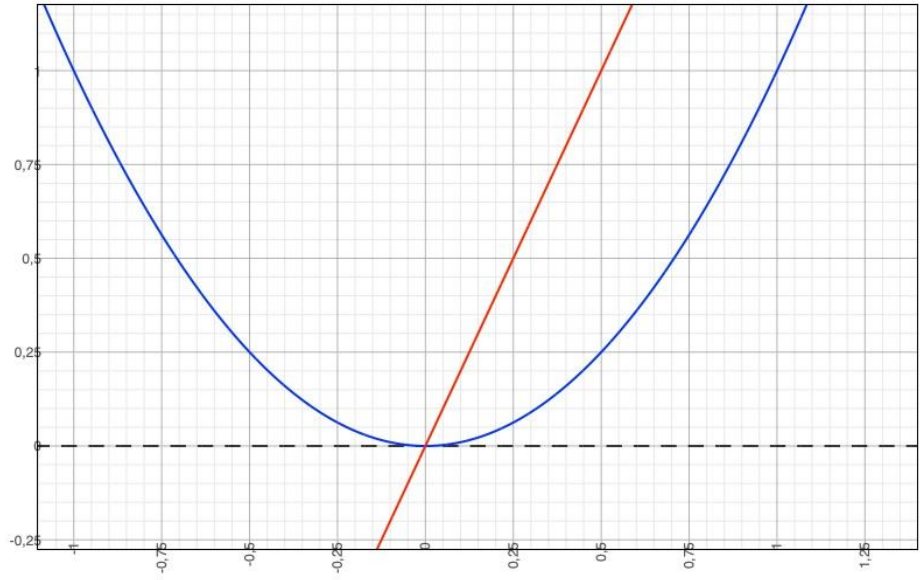
Примеры дифференциалов

$$s(t) = t^2$$

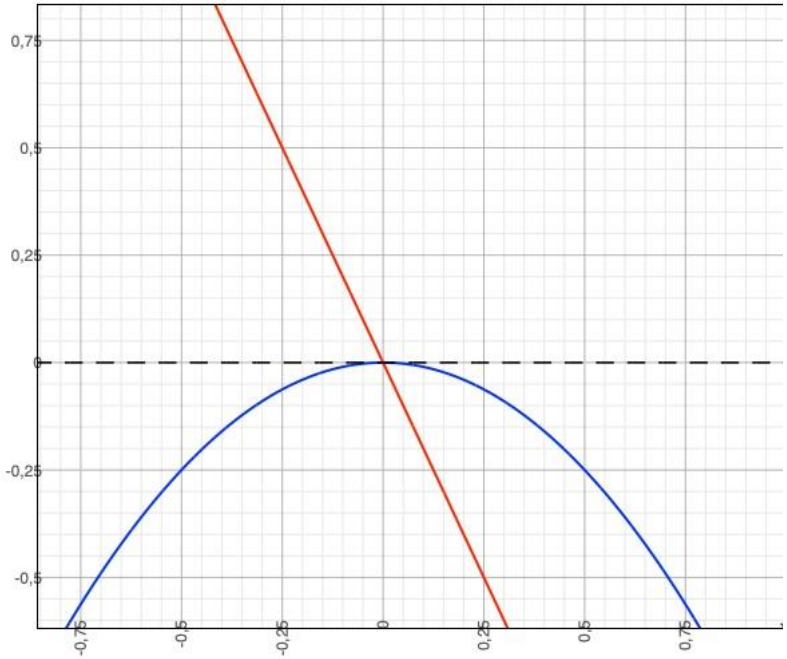
$$v(t) = \frac{ds}{dt}$$

$$= \lim_{h \to 0} \frac{s(t+h) - s(t)}{h}$$

$$= \frac{(t+h)^2}{h} = \frac{t^2 + 2th + h^2}{h} = 2t$$



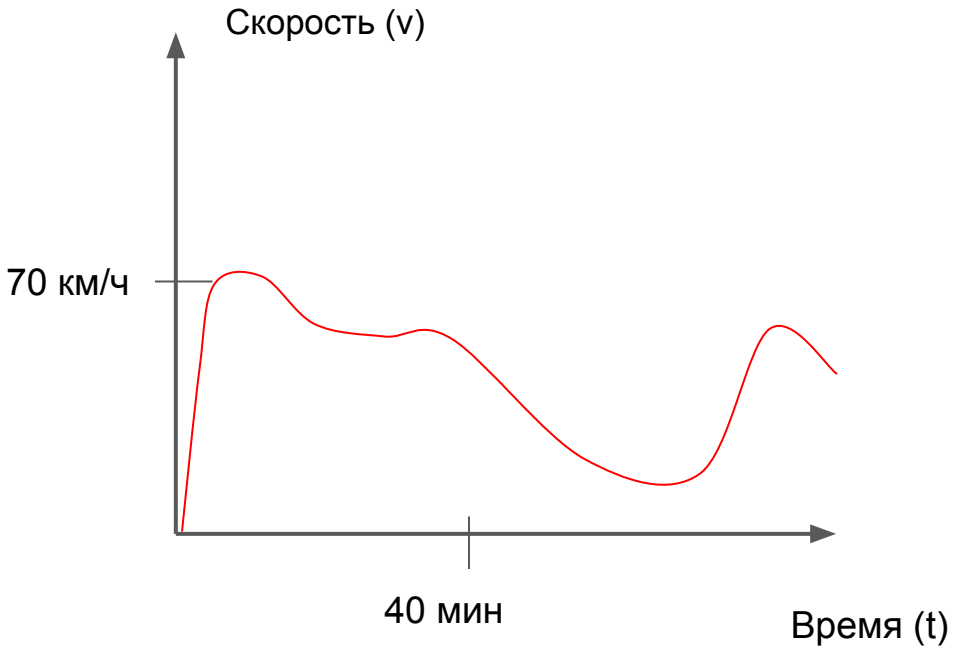
Зачем нам это?



- Нам нужно будет знать, где функции принимают свое максимальное или минимальное значение.
- Мы будем использовать, что дифференциал принимает в этом месте значение 0!

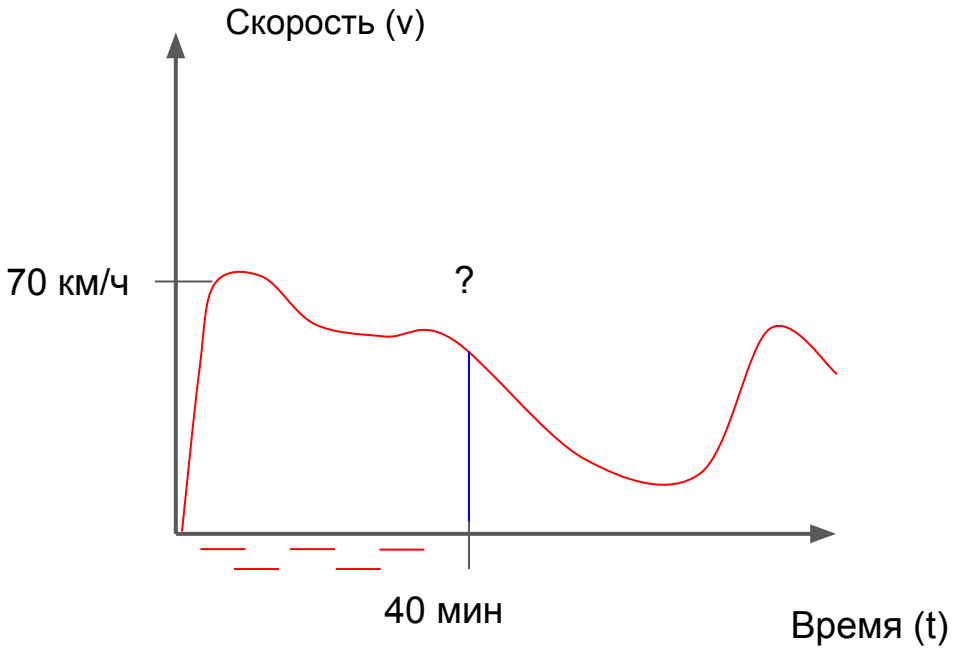
Интегральное исчисление

Что, если мы знаем только скорость?



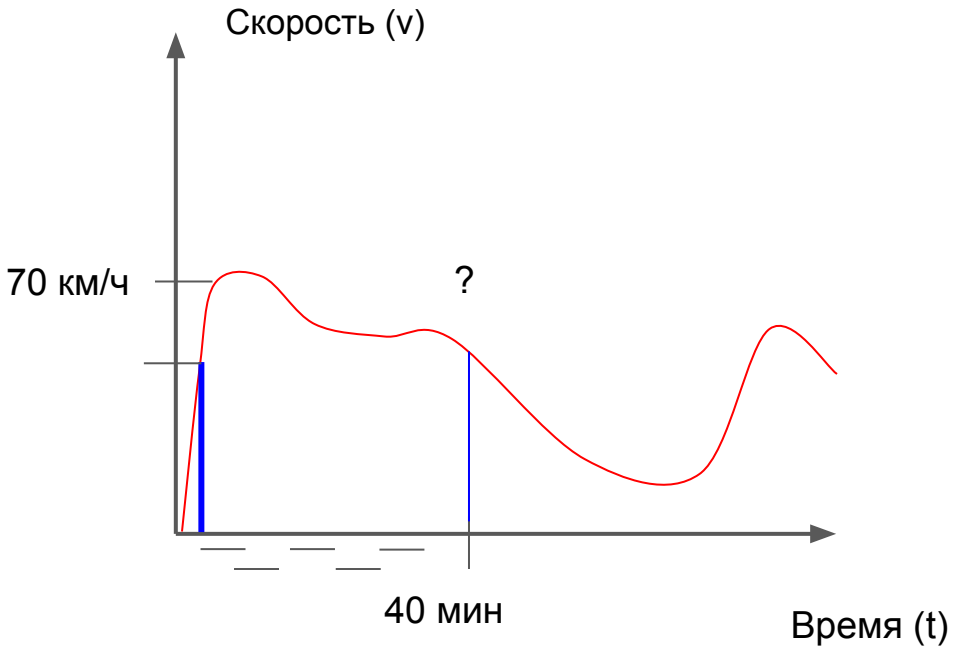
- Какое расстояние проехала эта машина в течение 40 мин?

Что, если мы знаем только скорость?



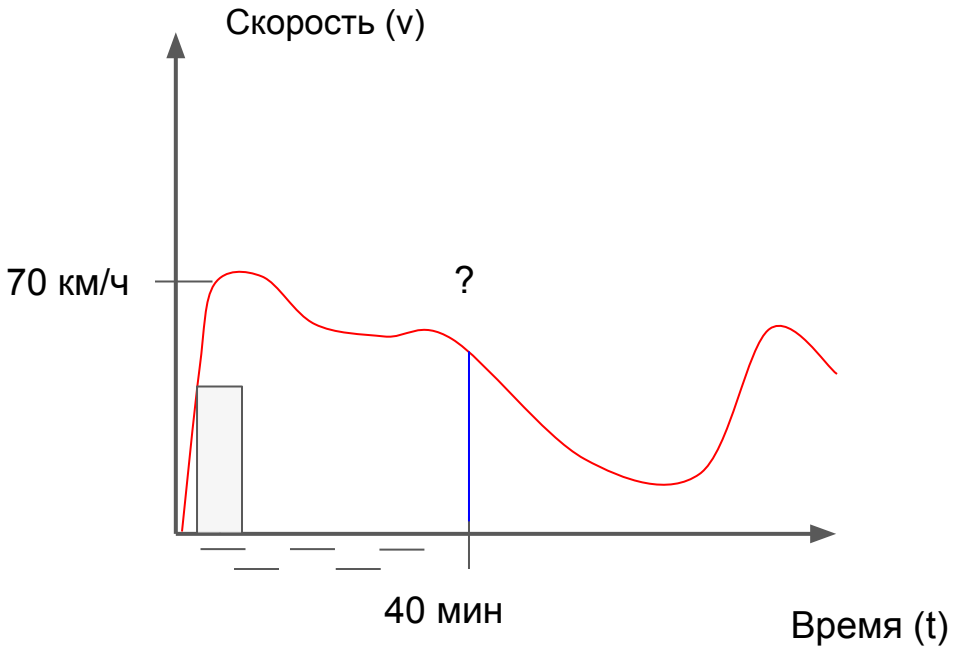
- Какое расстояние проехала эта машина в течение 40 мин?
- Расстояние – это
  - *скорость* время*
- **Идея:**
  - *Делим время на маленькие интервалы*

Что, если мы знаем только скорость?



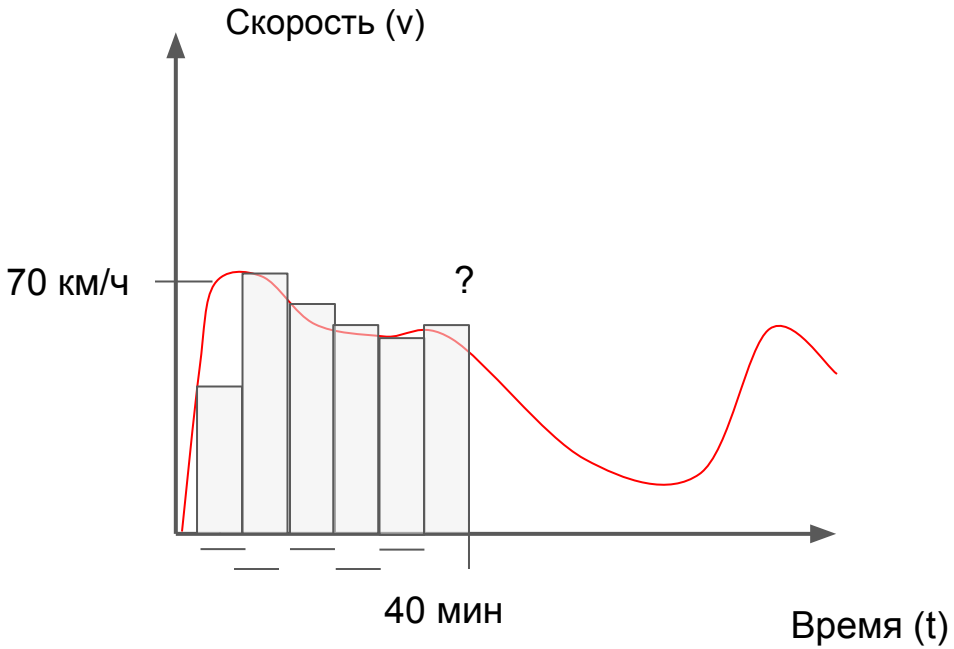
- Какое расстояние проехала эта машина в течение 40 мин?
- Расстояние – это
  - *скорость* время*
- **Идея:**
  - *Делим время на маленькие интервалы*
  - *Берем скорость в момент начала интервала*

Что, если мы знаем только скорость?



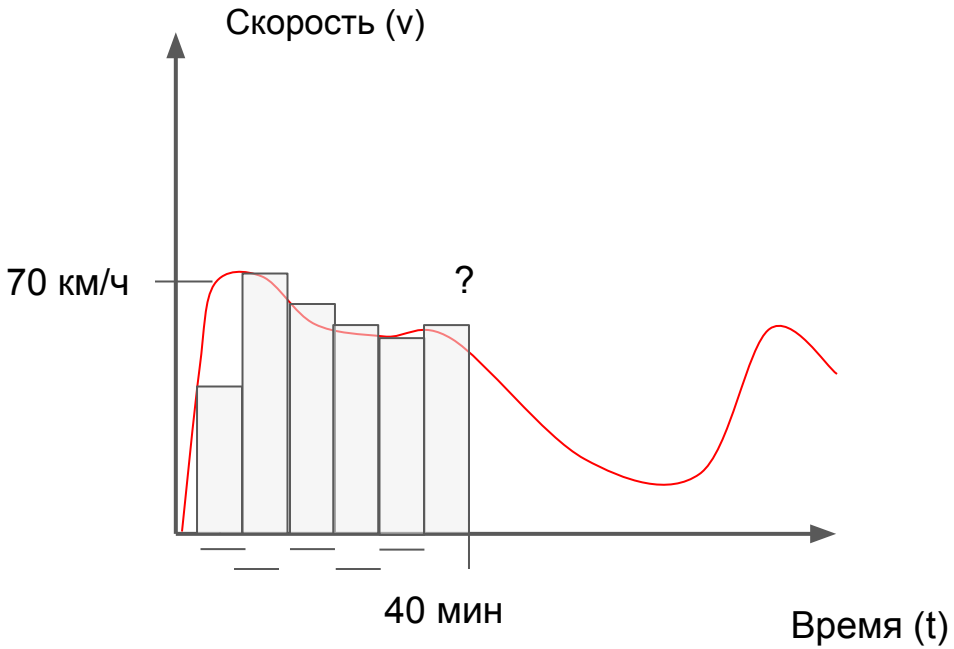
- Какое расстояние проехала эта машина в течение 40 мин?
- Расстояние – это
  - *скорость* время*
- **Идея:**
  - *Делим время на маленькие интервалы*
  - *Берем скорость в момент начала интервала*
  - *Умножаем скорость на длину интервала и получаем расстояние, пройденное в этом интервале*

Что, если мы знаем только скорость?



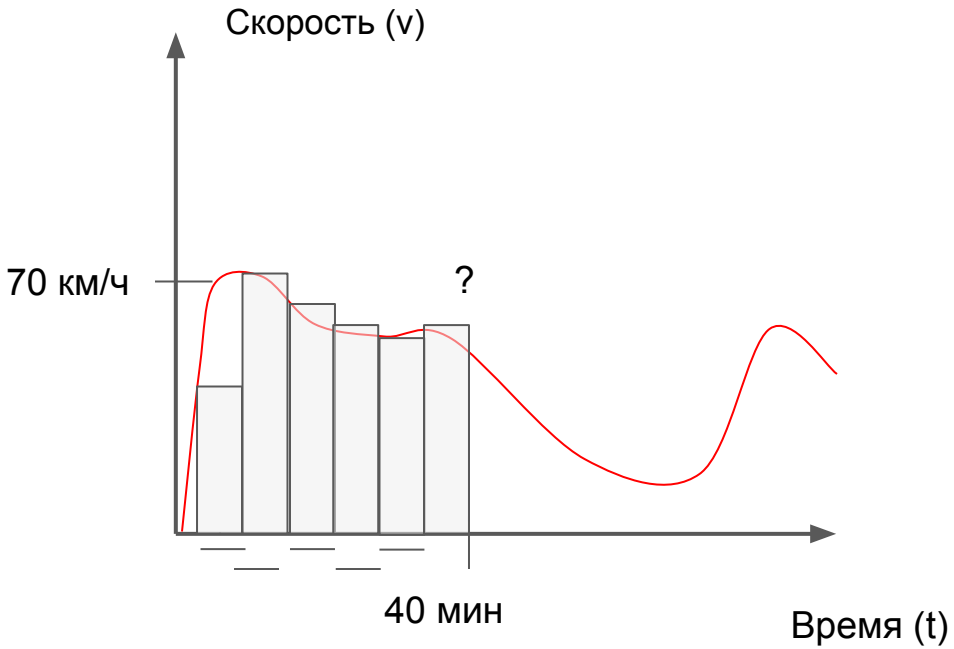
- Какое расстояние проехала эта машина в течение 40 мин?
- Расстояние – это
  - *скорость* время*
- **Идея:**
  - *Делим время на маленькие интервалы*
  - *Берем скорость в момент начала интервала*
  - *Умножаем скорость на длину интервала и получаем расстояние, пройденное в этом интервале*
  - *Суммируем все расстояния*

Что, если мы знаем только скорость?



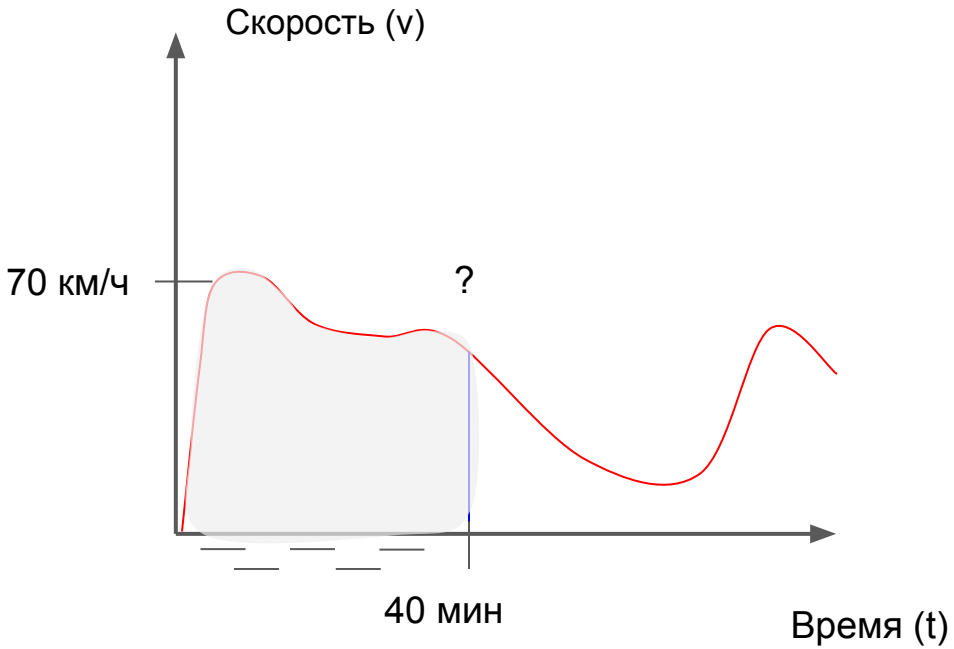
- Какое расстояние проехала эта машина в течение 40 мин?
- Расстояние – это
  - *скорость* время*
- **Наша проблема превратилась в проблему нахождения площади под кривой → интегральное исчисление**

Что, если мы знаем только скорость?



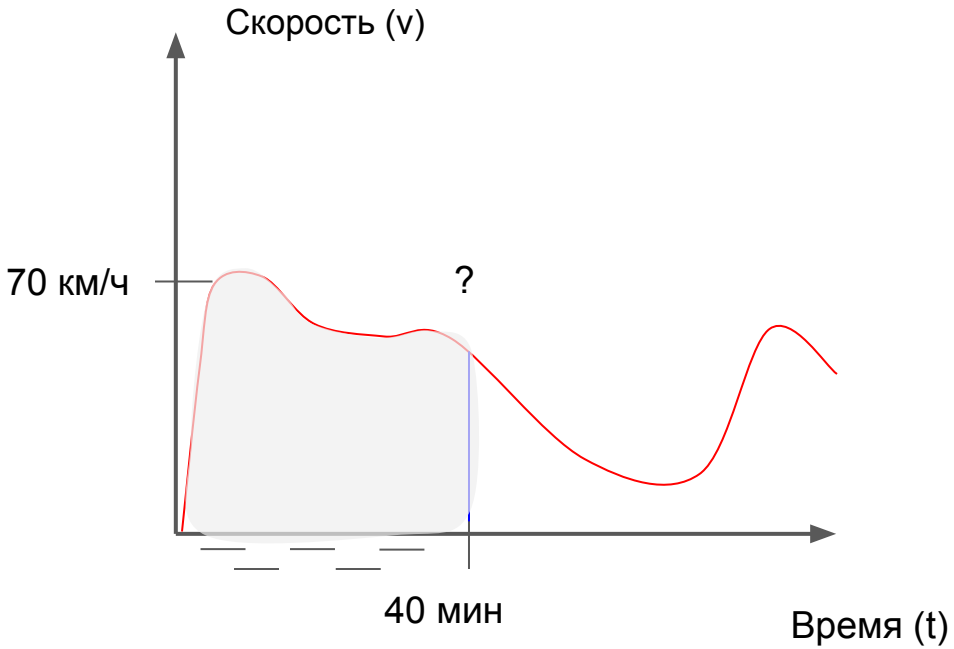
- Какое расстояние проехала эта машина в течение 40 мин?
- Расстояние – это
  - *скорость* время*
- **В следующем шаге нам надо уменьшить ошибку, делая наши отрезки все меньше и меньше**

Что, если мы знаем только скорость?



- Какое расстояние проехала эта машина в течение 40 мин?
- Расстояние – это
  - *скорость* время*
- **В следующем шаге нам надо уменьшить ошибку, делая наши отрезки все меньше и меньше**

Что, если мы знаем только скорость?



- Какое расстояние проехала эта машина в течение 40 мин?

$$s(40) = \int_0^{40} v(t)dt$$

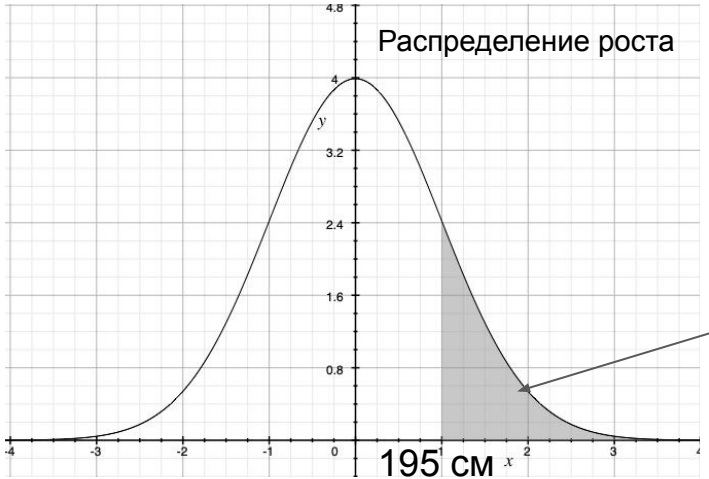
Фундаментальная теорема интегрального и дифференциального исчисления

$$\int_a^b f(x)dx = F(b) - F(a)$$

$$\frac{dF}{dx} = f(x)$$

Зачем нам это?

- *Допустим, мы знаем, как распределена высота населения, то есть какой процент населения имеет какой рост в определенном интервале. Какая вероятность встретить случайно человека с ростом больше 195 см?*

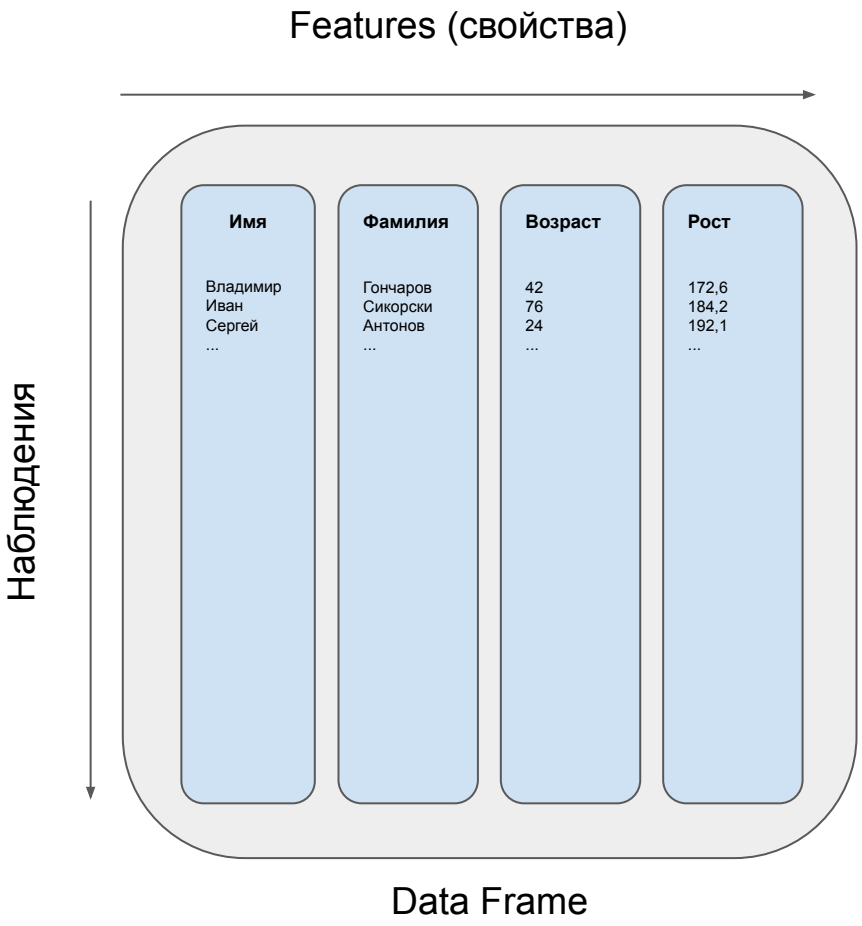


Вероятность встретить человека с ростом больше 195 см

Повторение:
Матрицы и векторы

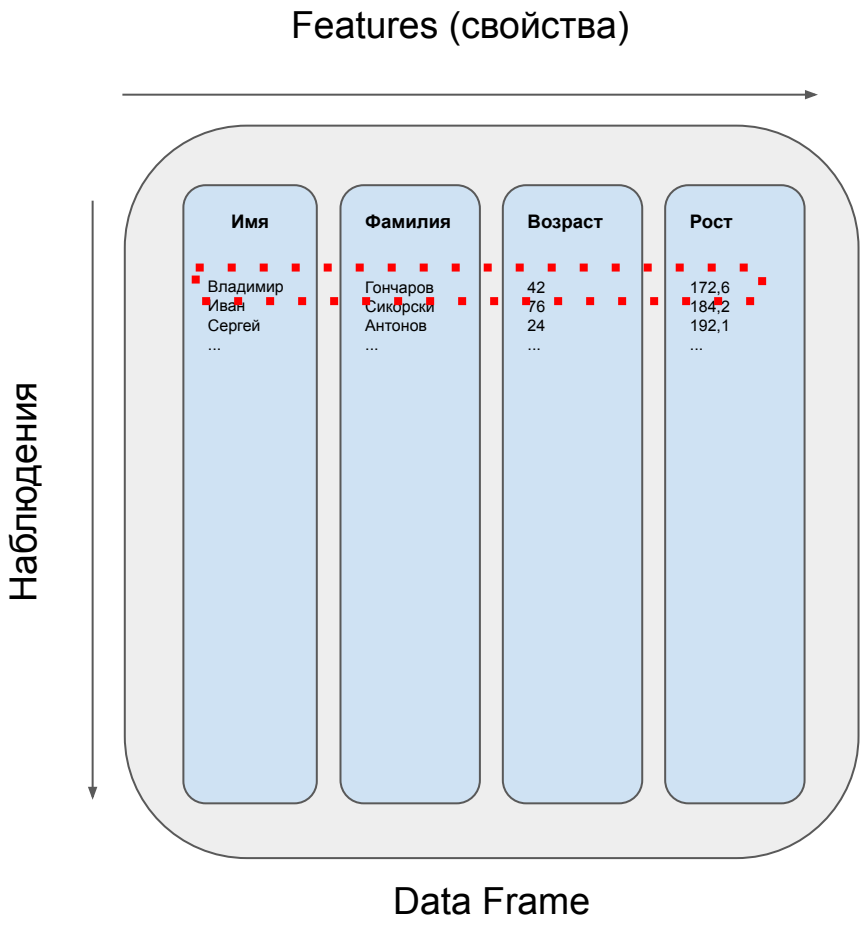
Векторы

- Для анализа данных мы используем понятия из линейной алгебры
- Мы не будем часто использовать эти методы в этом курсе, но они лежат в основании всей математики для машинного обучения
- На прошлом курсе мы видели структуры для описания данных



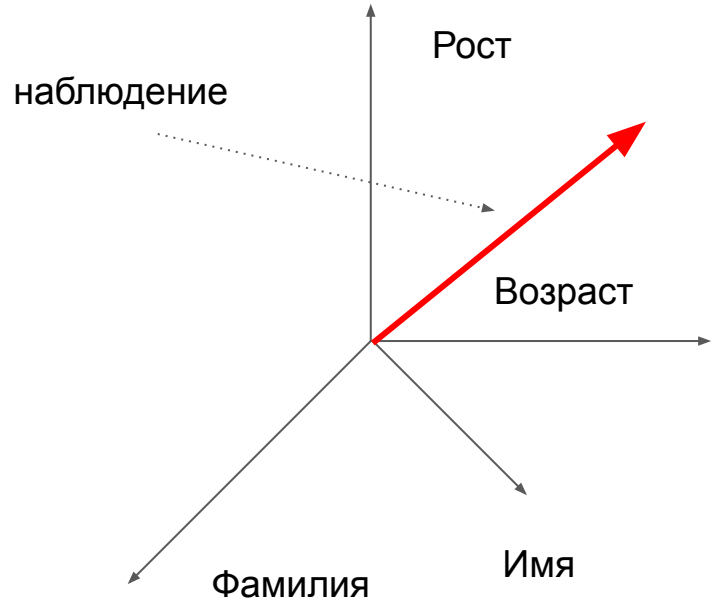
Векторы

- Одно наблюдение имеет, например здесь, 4 свойства:
  - Имя
  - Фамилия
  - Возраст
  - Рост



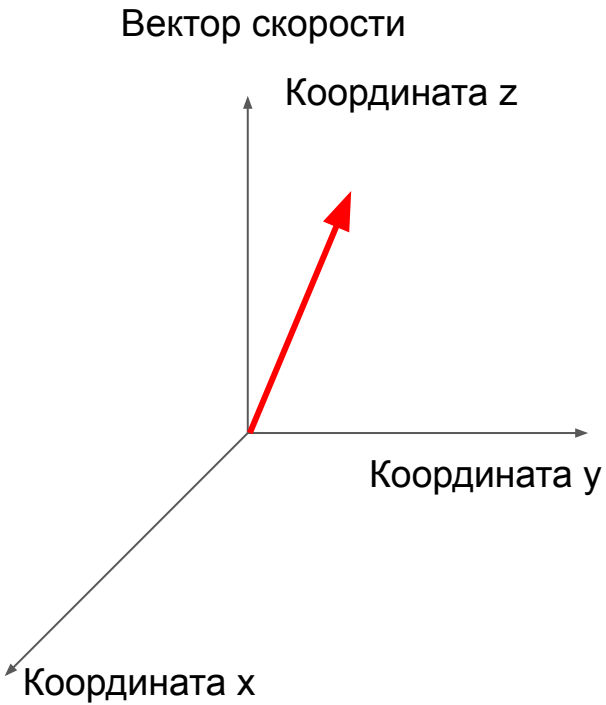
Векторы

- Одно наблюдение имеет, например здесь, 4 свойства:
  - Имя
  - Фамилия
  - Возраст
  - Рост
- Мы можем представить, что наблюдения живут в четырехмерном пространстве

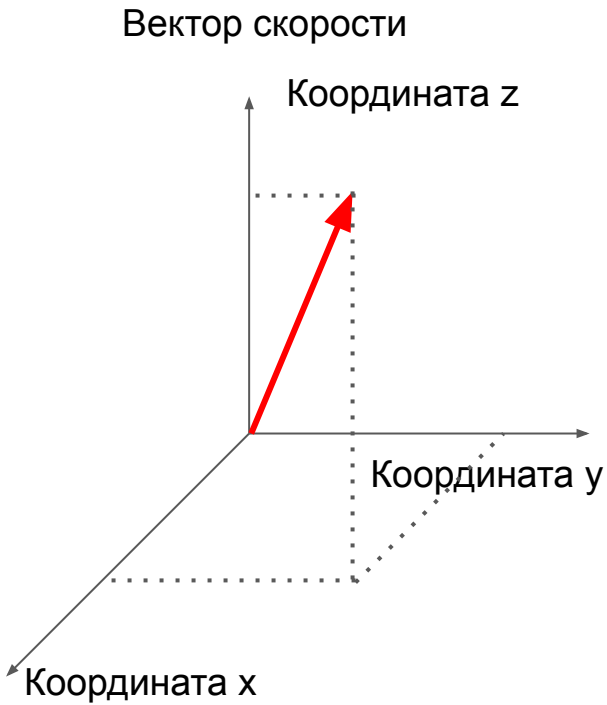


Векторы

- Математически более правильно и проще представлять себе это, если каждое свойство число.
- **Вектор – это математический объект, характеризующийся величиной и направлением.**
- Типичный пример – это вектор скорости.
- Скорость имеет направление и саму величину.



Векторы



$$\vec{v} = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}$$

Матрицы

Матрицы будут иметь для нас два значения:

- Матрицы описывают операции над векторами
  - Пример: вращение в 2 или 3 измерениях
- Мы будем использовать матрицы для сведения данных:
  - Матрицы – это массивы NumPy

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$