Введение в ML

Алексей Шаграев

На кого рассчитан этот мастер-класс

- На классных людей!
- Без большого опыта в приложениях ML
- С опытом на уровне «тулинга»

Как мы будем работать

- Smalltalk

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами
- Первый день – «семплинг» и данные

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами
- Первый день – «семплинг» и данные
- Второй день – методы оптимизации и метрики качества

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами
- Первый день – «семплинг» и данные
- Второй день – методы оптимизации и метрики качества
- Если что-то не получается – говорите, поможем

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами
- Первый день – «семплинг» и данные
- Второй день – методы оптимизации и метрики качества
- Если что-то не получается – говорите, поможем

Как мы будем работать

- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами
- Первый день – «семплинг» и данные
- Второй день – методы оптимизации и метрики качества
- Если что-то не получается – говорите, поможем
- Если не успели доделать – доделайте потом!

Как мы будем работать

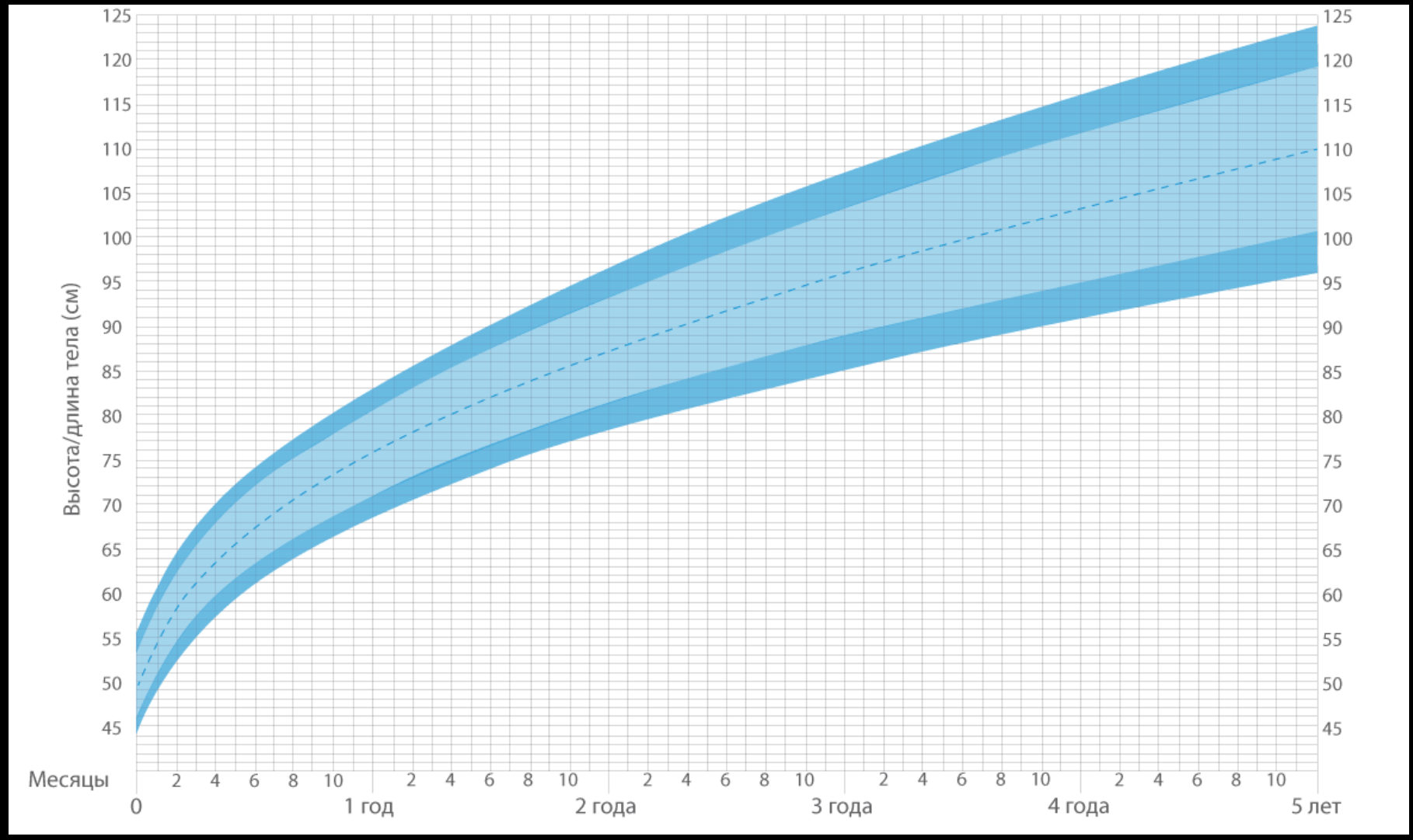
- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами
- Первый день – «семплинг» и данные
- Второй день – методы оптимизации и метрики качества
- Если что-то не получается – говорите, поможем
- Если не успели доделать – доделайте потом!
- Если неохота делать – можно не делать ☺

Как мы будем работать

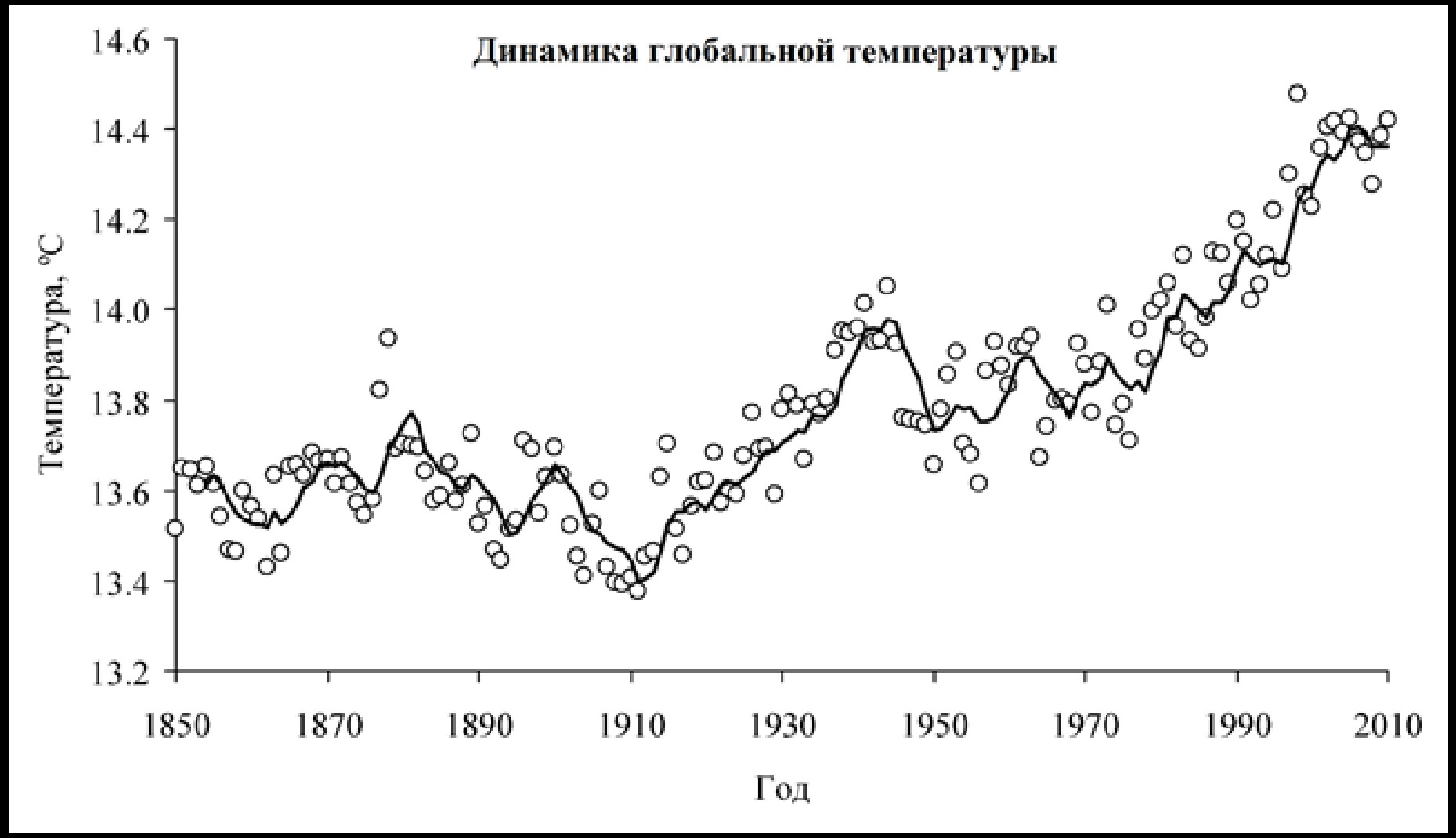
- Smalltalk
- Не очень много (4-6) задачек
- Потребуются python и matplotlib
- На каждую задачу около 15 минут, далее делимся результатами
- Первый день – «семплинг» и данные
- Второй день – методы оптимизации и метрики качества
- Если что-то не получается – говорите, поможем
- Если не успели доделать – доделайте потом!
- Если неохота делать – можно не делать ☺
- Задавайте вопросы! Прямо вслух!

Почему что-либо можно предсказать?

Почему что-либо можно предсказать?



Почему что-либо можно предсказать?



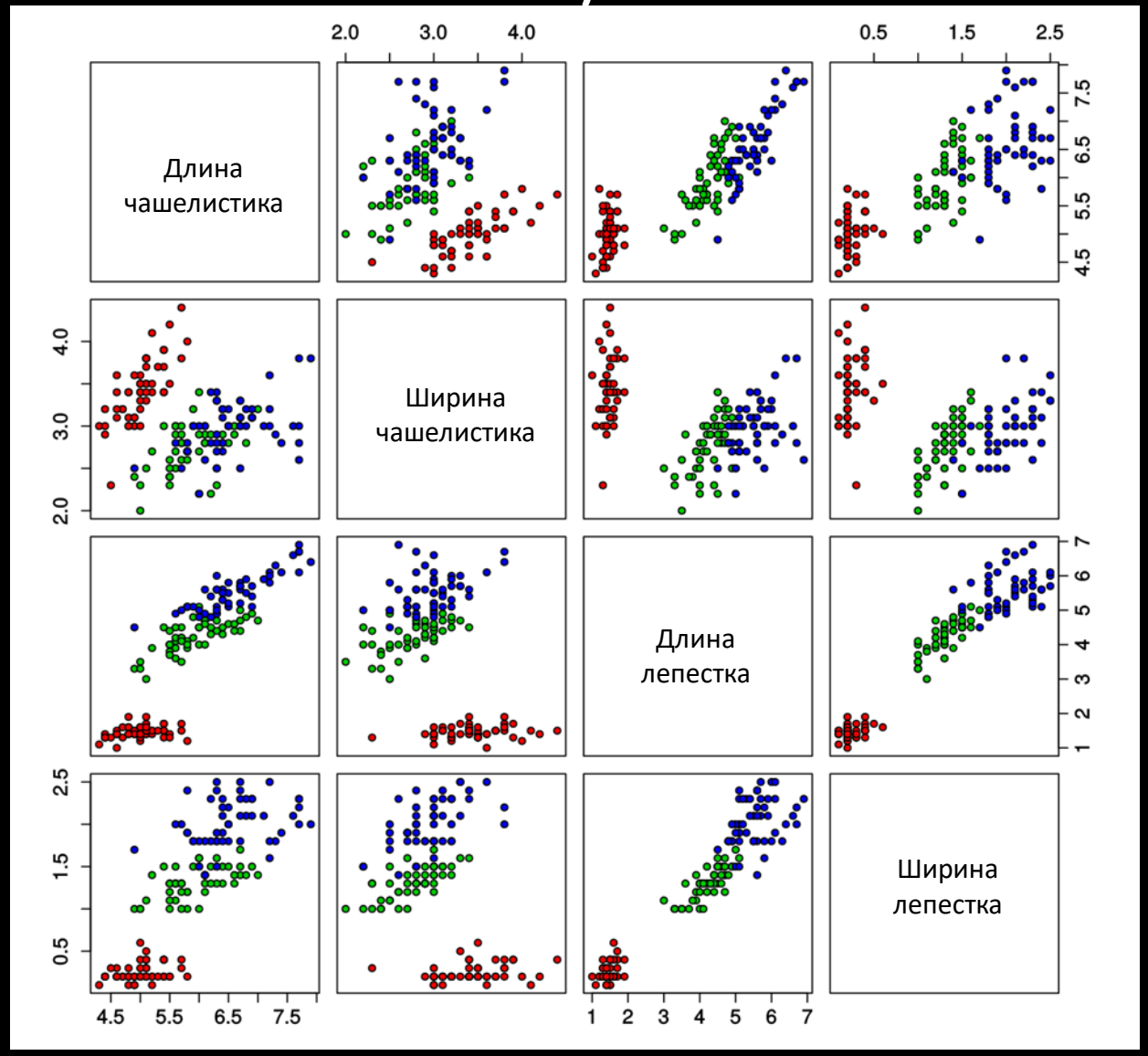
Почему что-либо можно предсказать?

Iris setosa

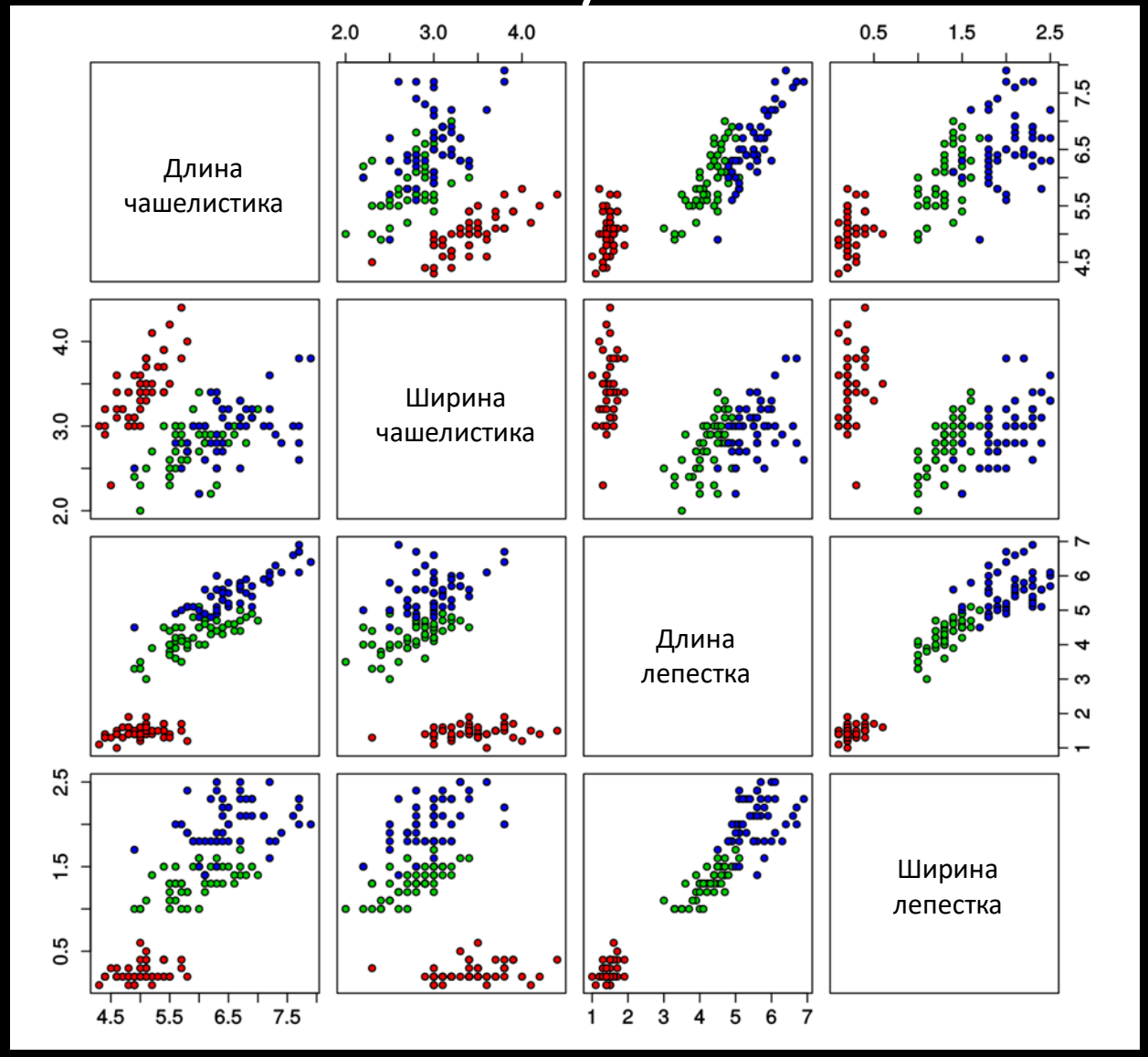
Iris versicolor

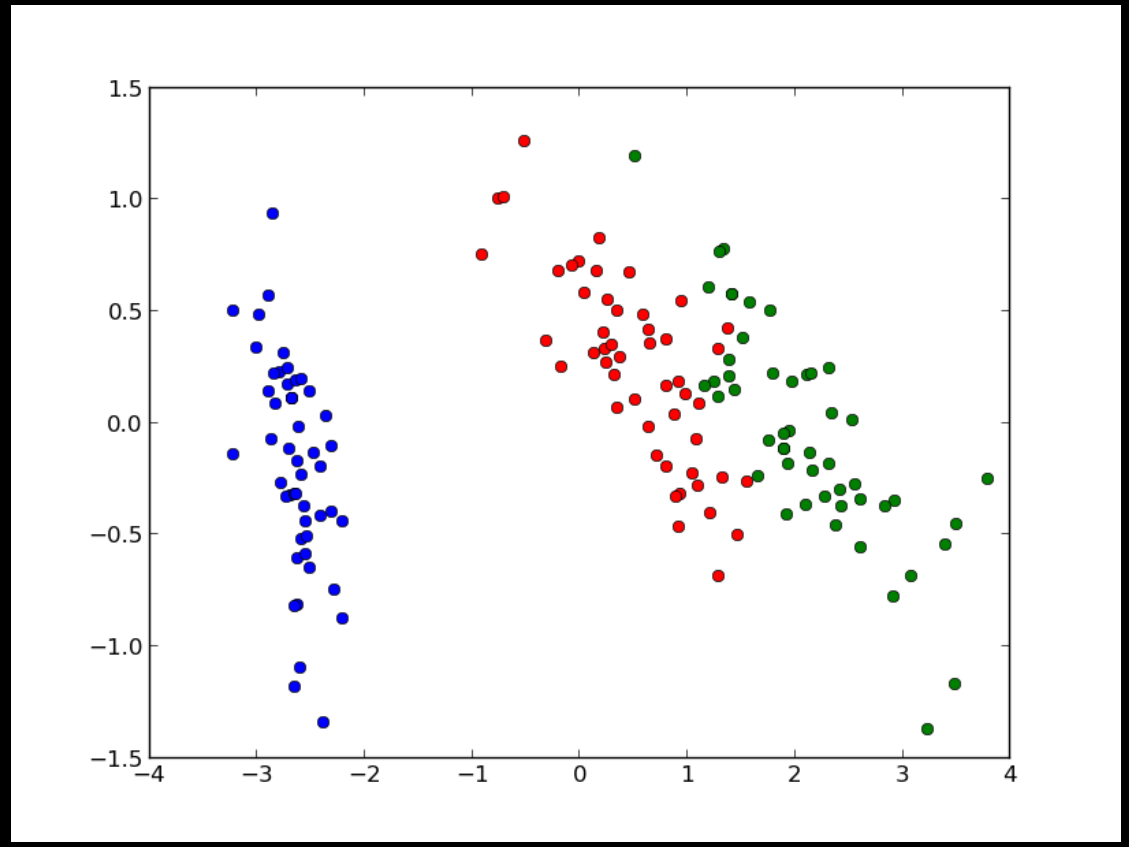
Iris virginica

Почему что-либо можно предсказать?

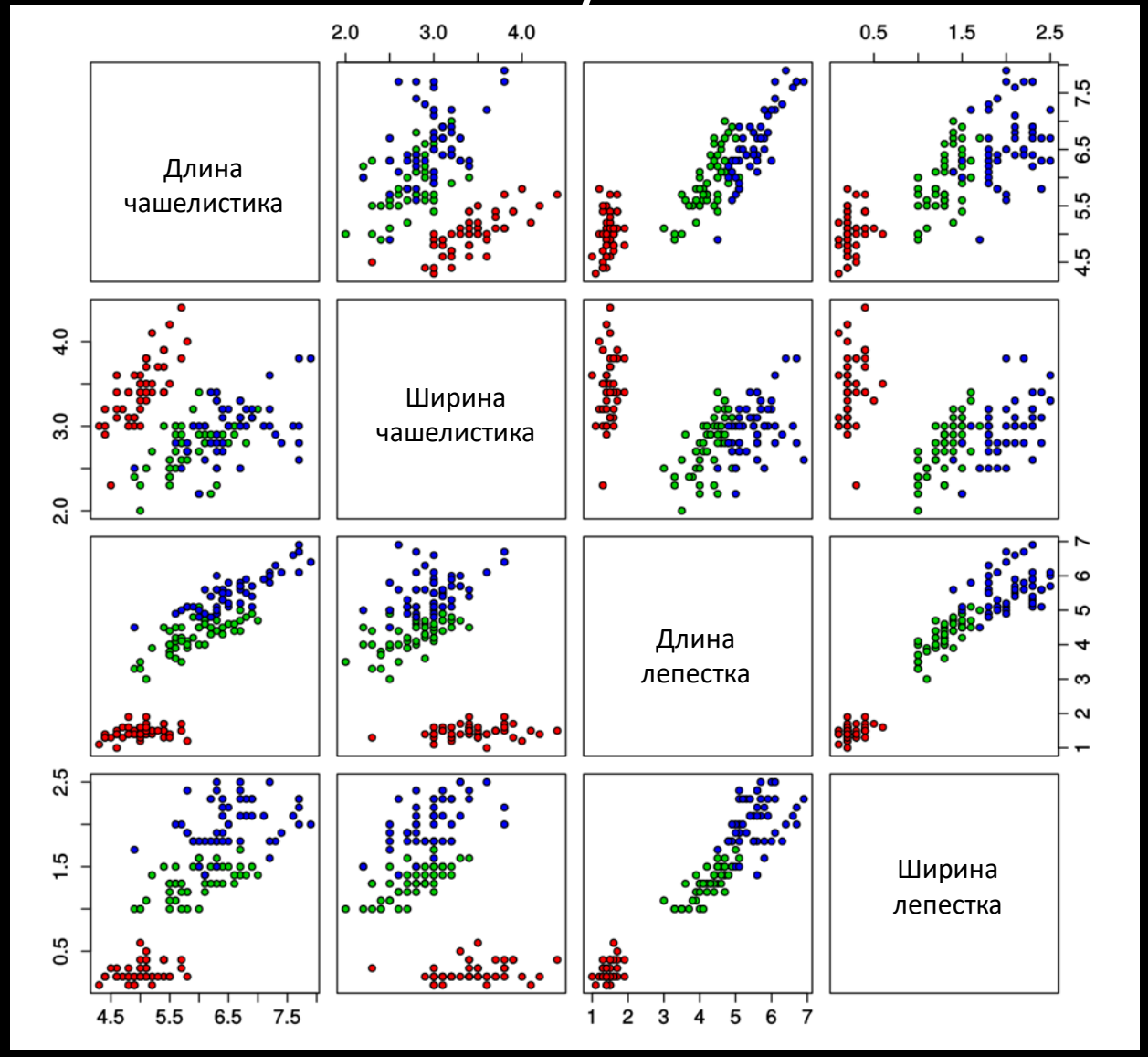


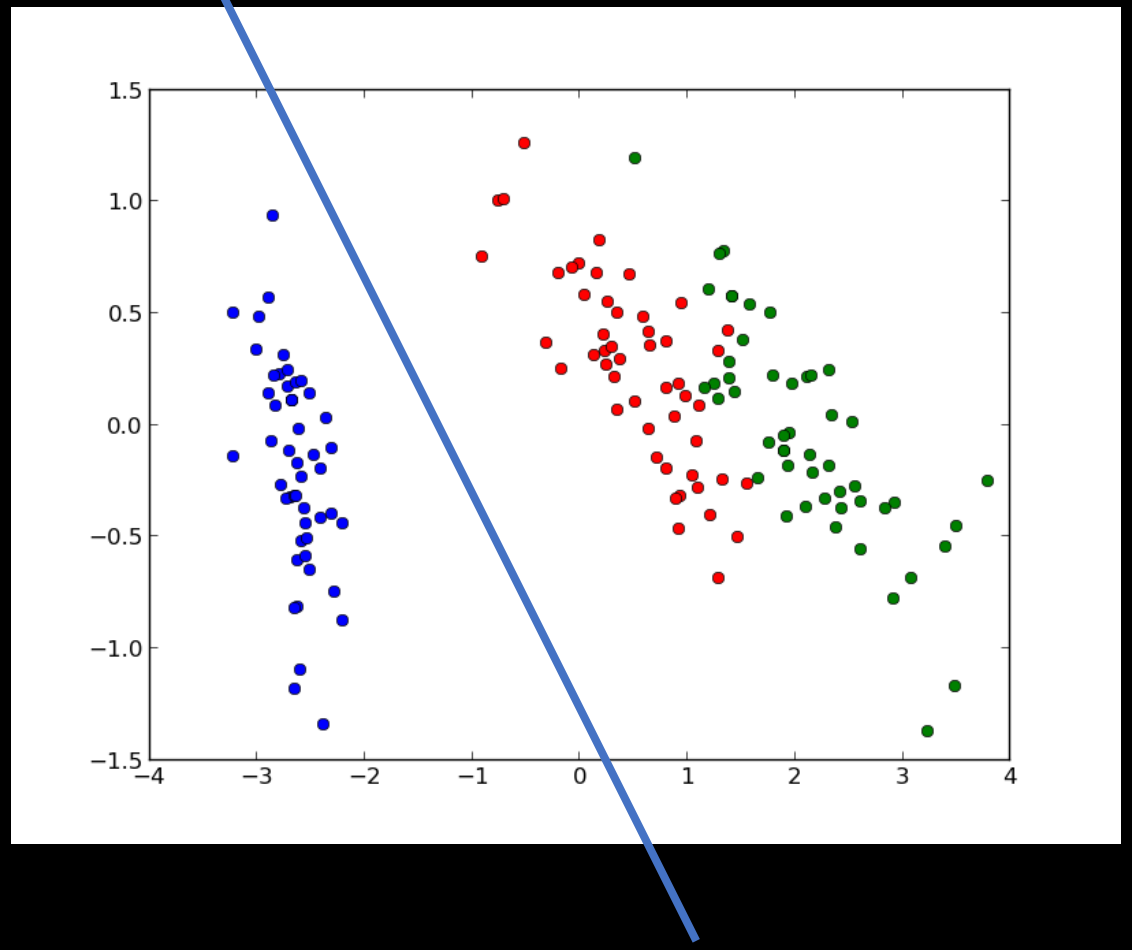
Почему что-либо можно предсказать?



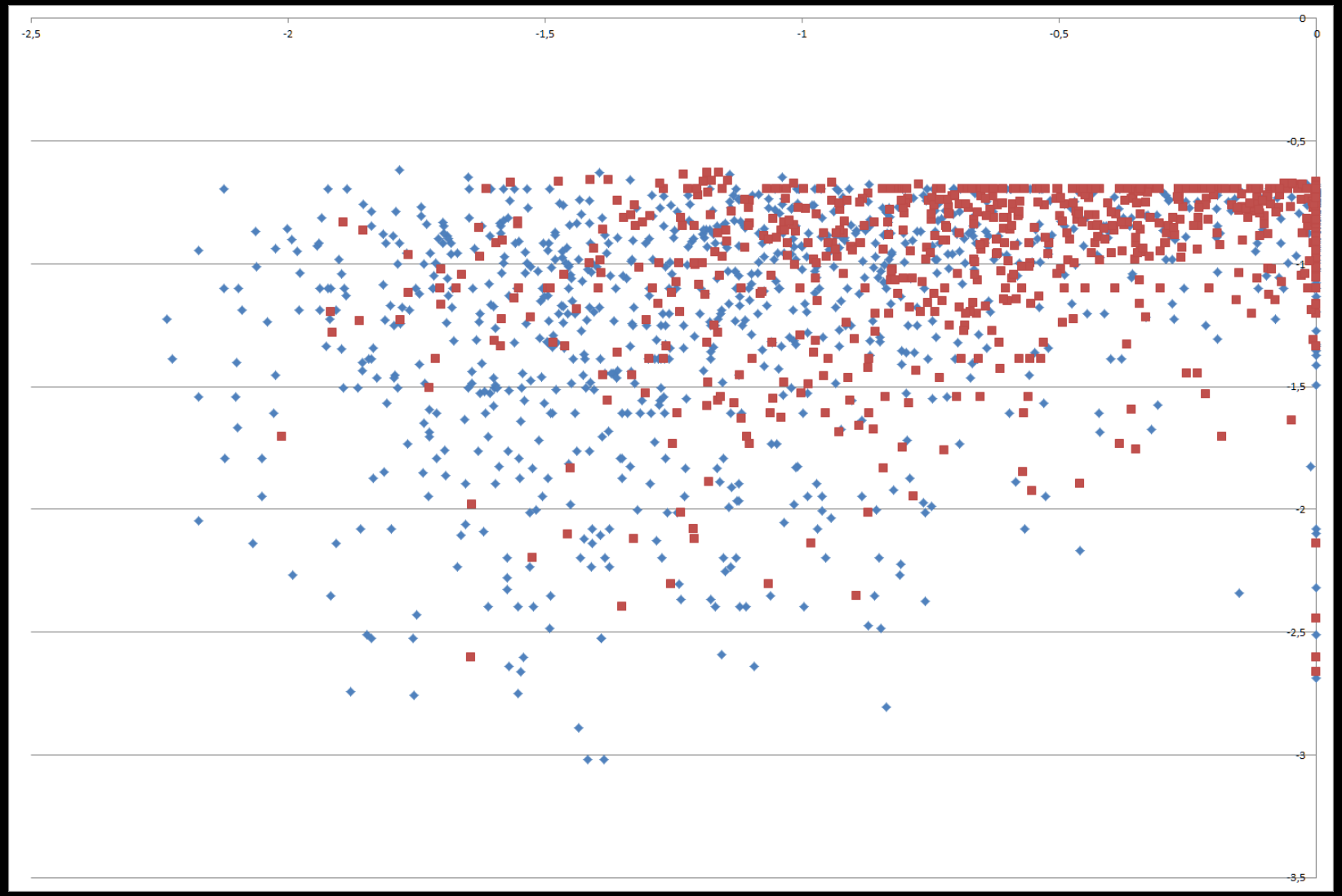


Почему что-либо можно предсказать?

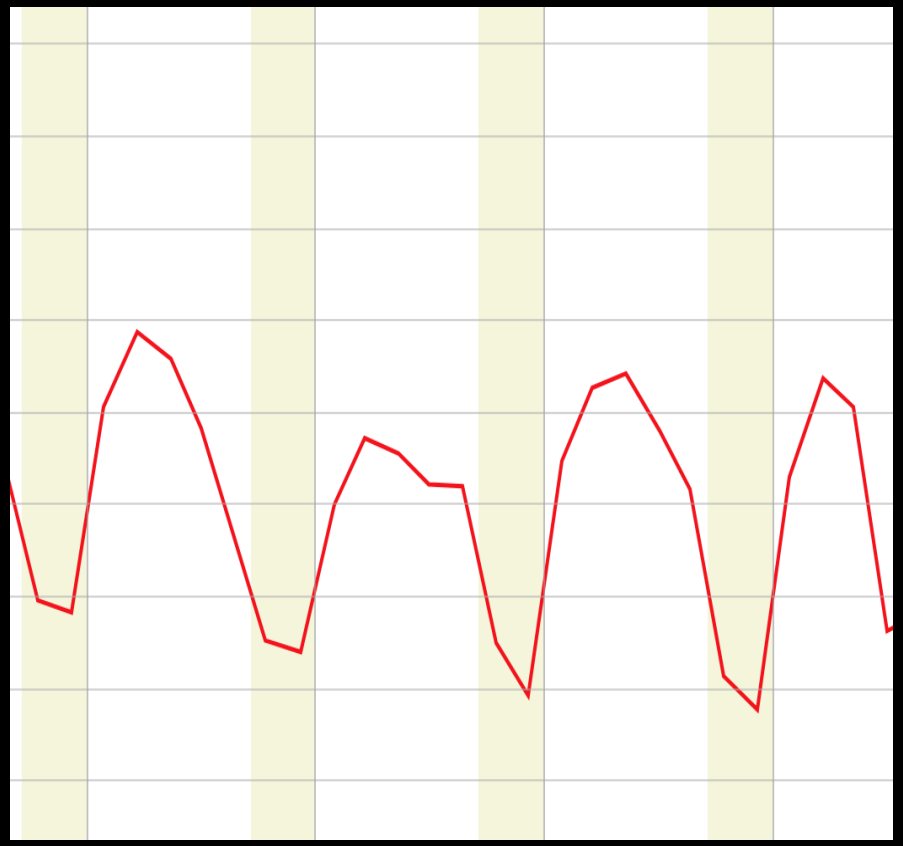




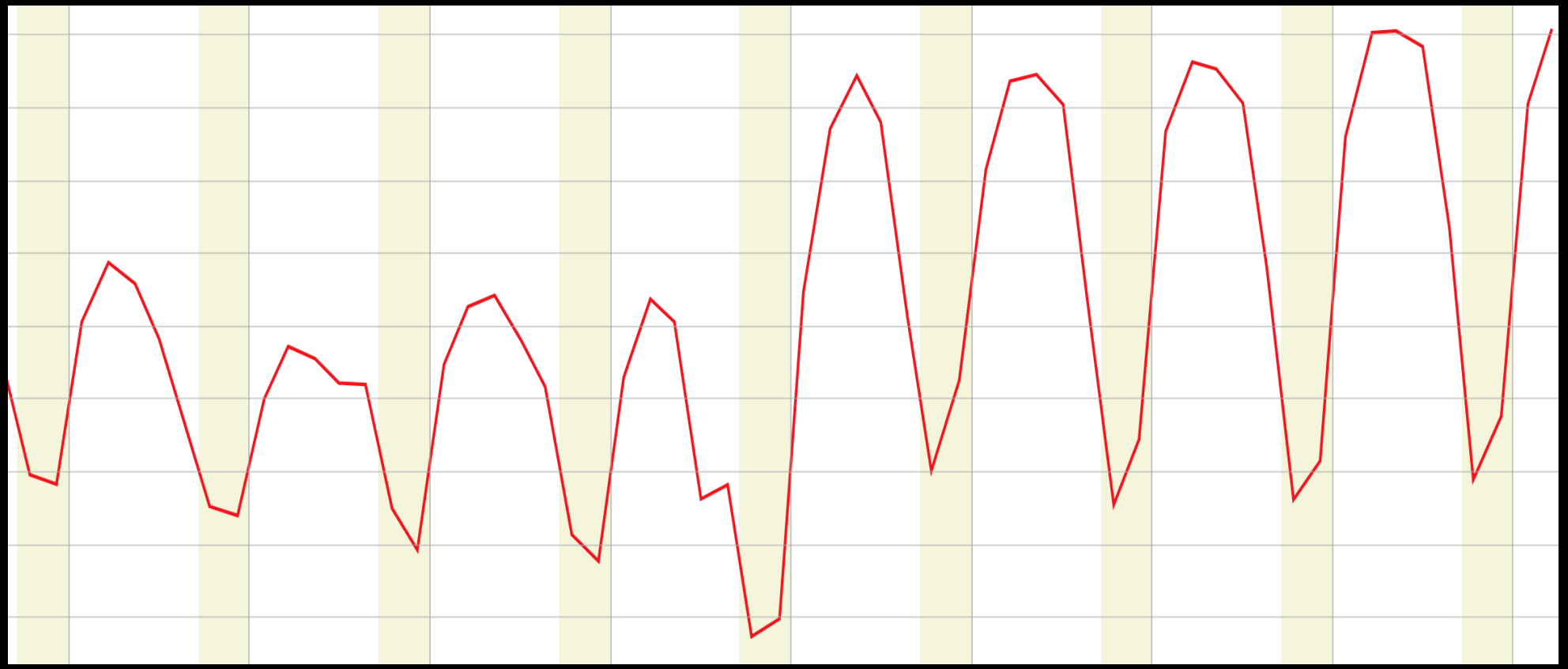
Почему что-либо можно предсказать?



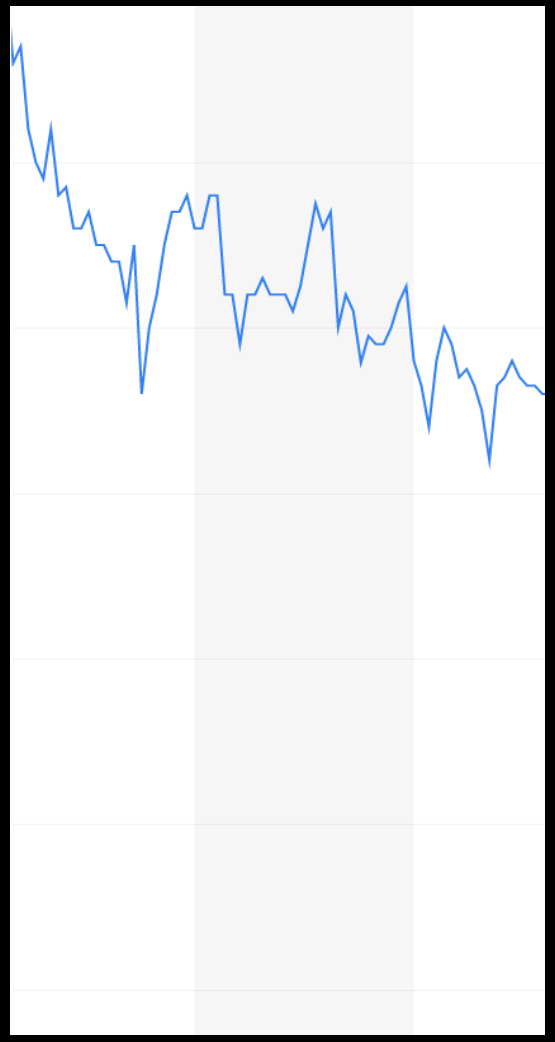
Почему что-либо можно предсказать?



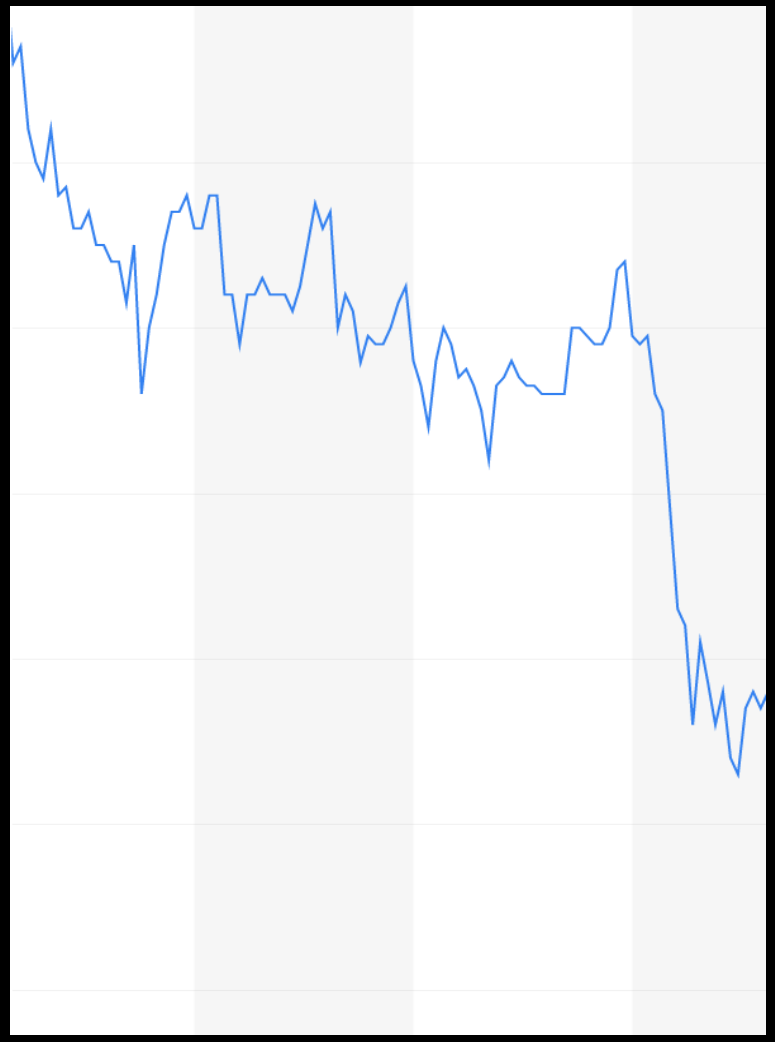
Почему что-либо можно предсказать?



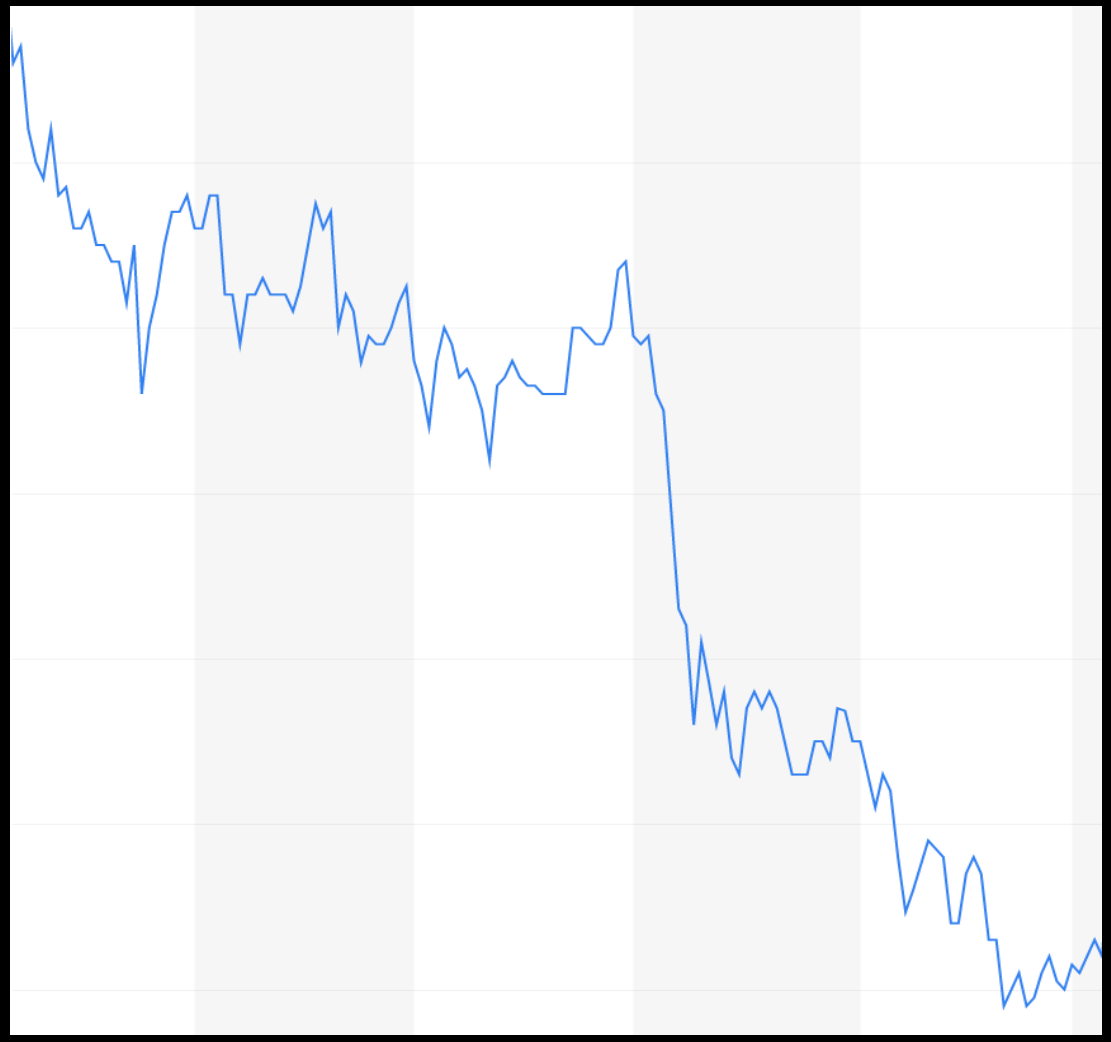
Почему что-либо можно предсказать?



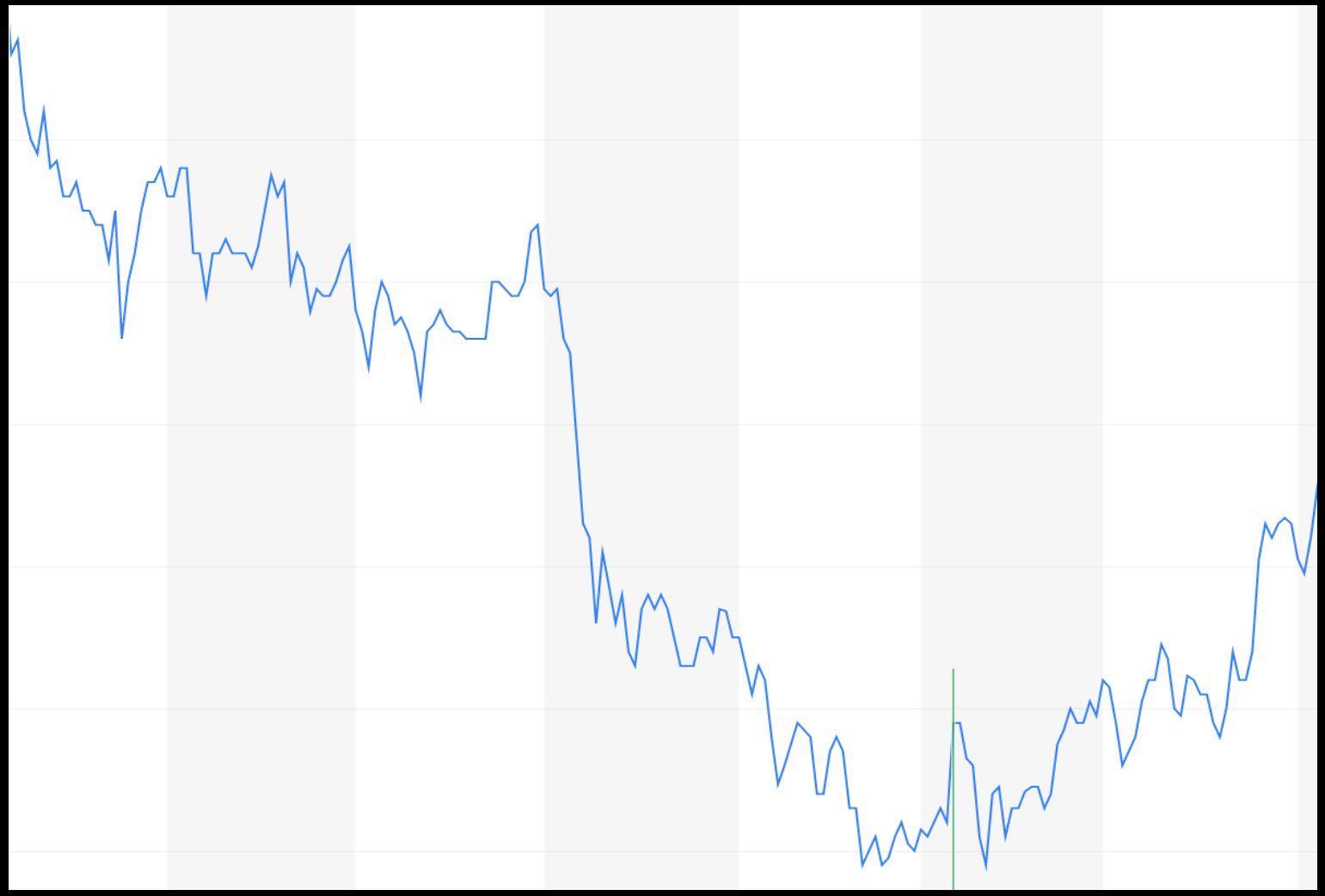
Почему что-либо можно предсказать?



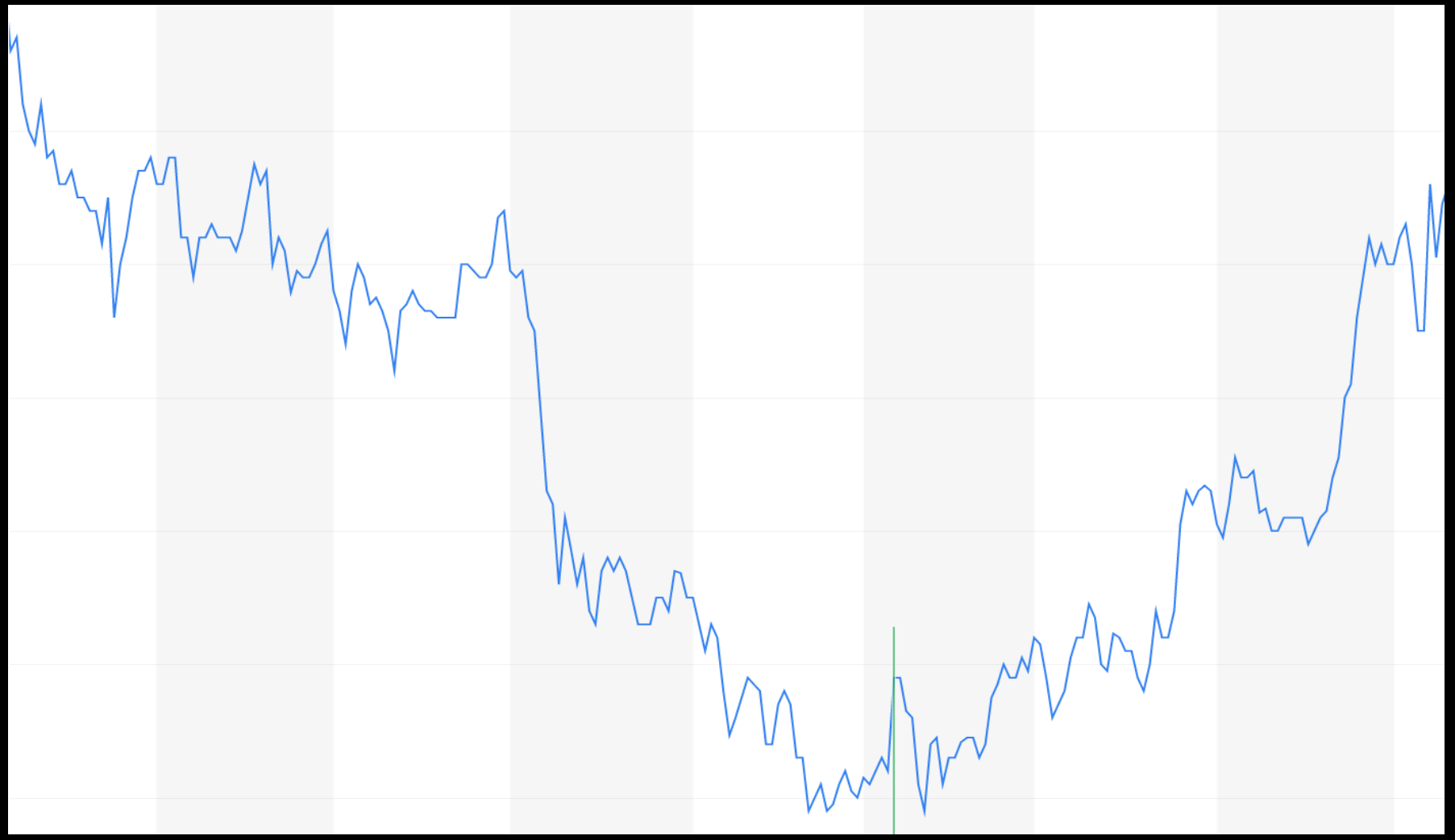
Почему что-либо можно предсказать?



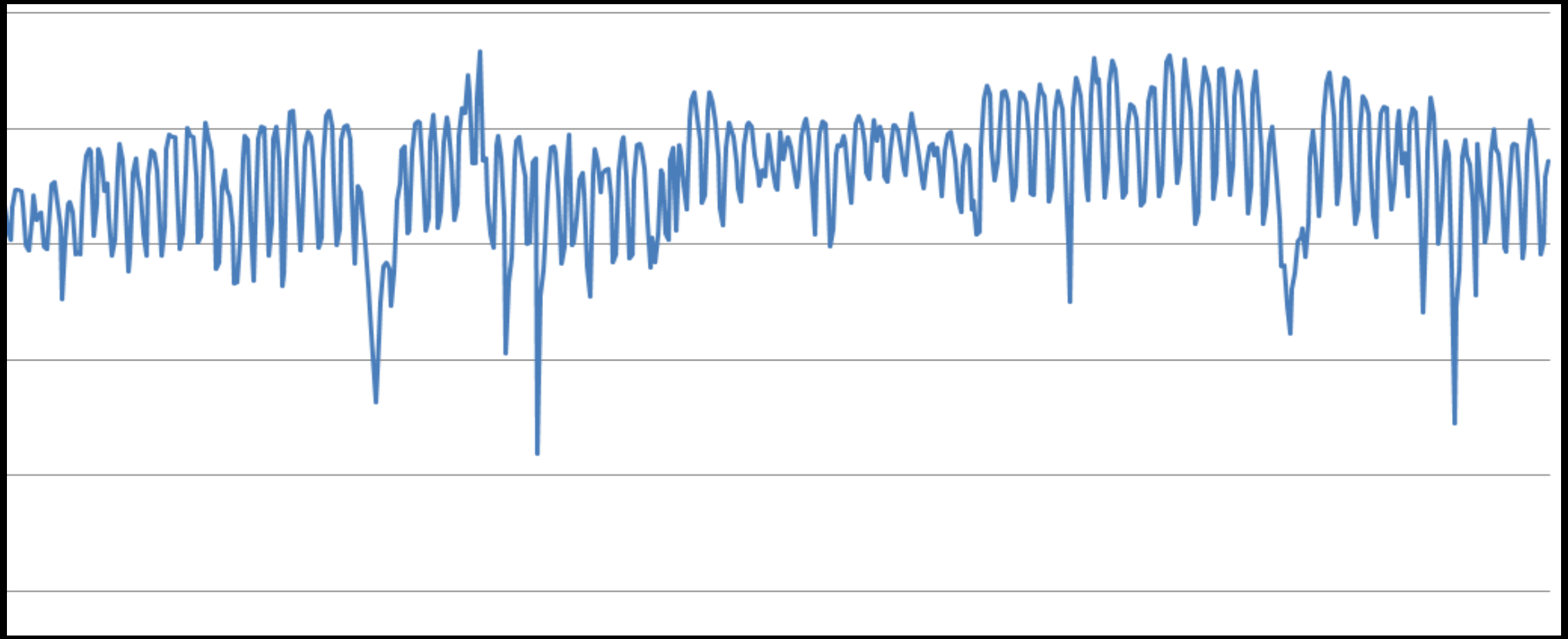
Почему что-либо можно предсказать?



Почему что-либо можно предсказать?



Почему что-либо можно предсказать?



Сколько месяцев на этом графике?

Sampling

- Есть «много» объектов
- Рассмотреть так много не получится
- Хочется взять «репрезентативное» подмножество объектов и посмотреть на него

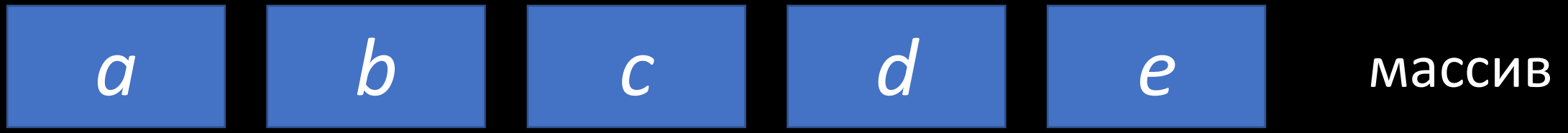
Алгоритм Random Shuffle

массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

Алгоритм Random Shuffle



массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

Алгоритм Random Shuffle

массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

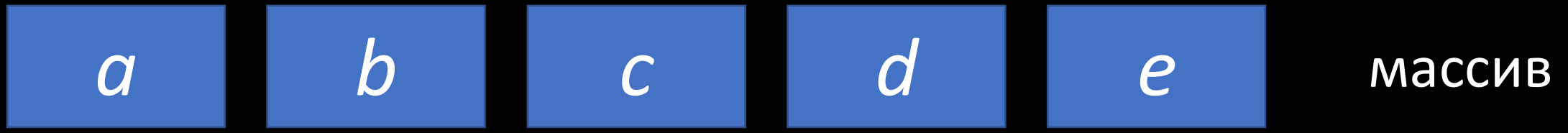
Алгоритм Random Shuffle

массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

Алгоритм Random Shuffle



массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

Алгоритм Random Shuffle

массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

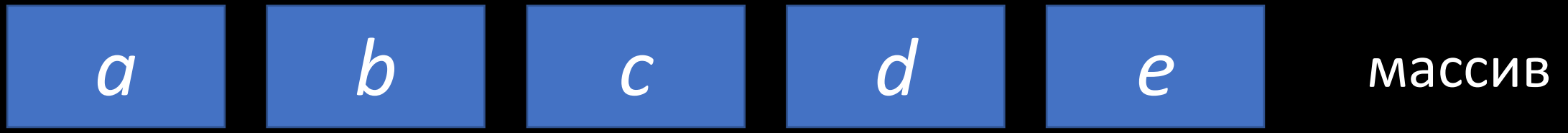
Алгоритм Random Shuffle

массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

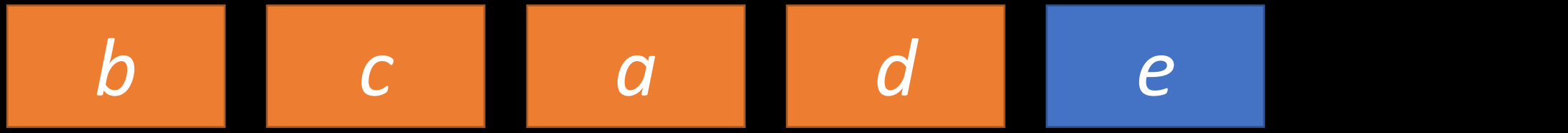
Алгоритм Random Shuffle



массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)



Алгоритм Random Shuffle

массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

Алгоритм Random Shuffle

массив

рассматриваем элементы последовательно

меняем местами очередной элемент и случайный элемент левее него (или меняем элемент с ним самим)

Задача №1

- Реализуйте функцию, которая возвращает k случайных элементов из массива длины n
- Сгенерируйте несколько раз раз 2 случайных элемента из массива [1,2,3,4,5]
- Посмотрите, что выводит функция
- Не пользуйтесь библиотечными функциями ☺

Задача №1

- Реализуйте функцию, которая возвращает k случайных элементов из массива длины n
- Сгенерируйте 10 раз 2 случайных элемента из массива [1,2,3,4,5]
- Подсчитайте, какое число сколько раз выпало

Задача №1

- Реализуйте функцию, которая возвращает k случайных элементов из массива длины n
- Сгенерируйте 10 раз 2 случайных элемента из массива [1,2,3,4,5]
- Подсчитайте, какое число сколько раз выпало
- Сгенерируйте 100 раз 2 случайных элемента из массива [1,2,3,4,5]
- Подсчитайте, какое число сколько раз выпало

Задача №1

- Реализуйте функцию, которая возвращает k случайных элементов из массива длины n
- Сгенерируйте 10 раз 2 случайных элемента из массива [1,2,3,4,5]
- Подсчитайте, какое число сколько раз выпало
- Сгенерируйте 100 раз 2 случайных элемента из массива [1,2,3,4,5]
- Подсчитайте, какое число сколько раз выпало
- Сгенерируйте 10000 раз 2 случайных элемента из массива [1,2,3,4,5]
- Подсчитайте, какое число сколько раз выпало

Задача №2

- Реализуйте функцию, которая выбирает k случайных элементов из массива длины n и вычисляет их среднее значение («выборочное среднее»)
- Реализуйте функцию, которая вычисляет ошибку выборочного среднего: истинное среднее минус выборочное среднее
- Сгенерируйте случайный массив из 1000 целых чисел от 0 до 10
- Несколько раз сгенерируйте выборочное среднее и ошибку выборочного среднего

Задача №3

- Реализуйте функцию, которая заданное количество раз генерирует выборочное среднее
- Нарисуйте гистограмму ошибок выборочного среднего для массива из 1000 целых чисел от 0 до 10 для выборок размера 10, 100, 200
- Используйте библиотеку matplotlib

Задача №4

- Реализуйте функцию, которая заданное количество раз генерирует выборочное среднее
- Нарисуйте гистограмму ошибок выборочного среднего для массива из 1000 вещественных чисел для выборок размера 10, 100, 200
- Случайные числа генерируйте функцией 1. / random.random()
- Используйте библиотеку matplotlib

Задача №5

- Скачайте оригинальный набор данных 20newsgroups с сайта http://qwone.com/~jason/20Newsgroups/
- Реализуйте функцию, которая загружает текст конкретного письма
- Загрузите тексты всех писем из темы alt.atheism
- Отсортируйте все слова всех текстов из этой темы по частоте
- Постройте график зависимости между частотой слова и его номером в списке слов, отсортированных по частоте

Задача №6

- Скачайте оригинальный набор данных 20newsgroups с сайта http://qwone.com/~jason/20Newsgroups/
- Реализуйте функцию, которая загружает текст конкретного письма
- Загрузите тексты всех писем из темы alt.atheism
- Отсортируйте все слова всех текстов из этой темы по частоте
- Постройте график зависимости между частотой слова и его номером в списке слов, отсортированных по частоте
- Определите, какие слова наиболее «контрастны» в этой теме, если сравнивать её с какими-нибудь другими темами? Почему? ☺

Спасибо!

https://github.com/ashagraev/ml_basics

https://habr.com/users/ashagraev/posts/

https://t.me/insilicio

https://www.facebook.com/ashagraev

https://vk.com/shagraev