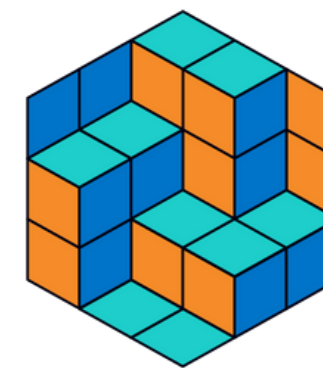


# **Предсказание временных рядов**

**Лекция 1**

**Летняя школа  
МКН СПбГУ 2023**



**Понизяйкин Владислав**

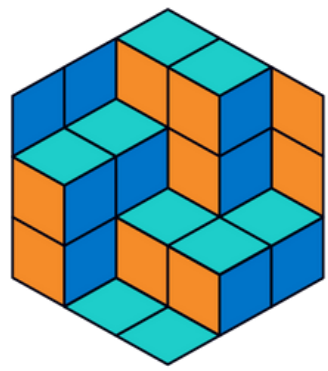
tg: @ArChanDD



**Ушаков Михаил**

tg: @MuwecTb

# Главное - не теряйтесь!



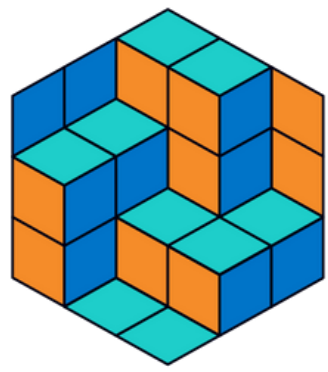
тг-канал курса



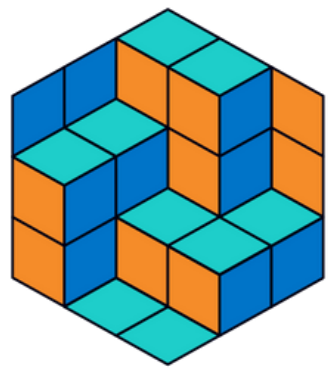
github-репозиторий



# Что вас ждет



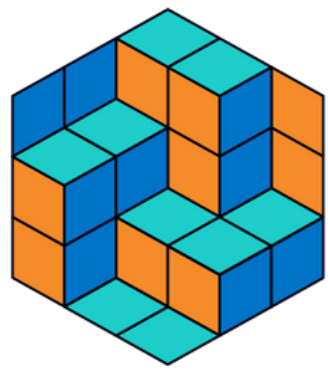
- 2 Лекции
- 2 Практики
- 1 Kaggle-соревнование



# Глава 1

**Кто такие временные ряды  
и что с ними делать?**

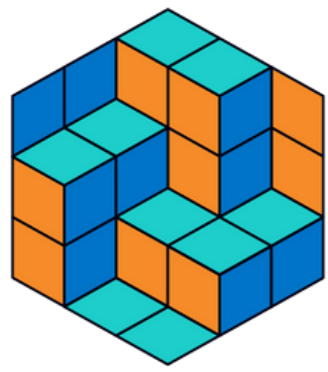
# Примеры



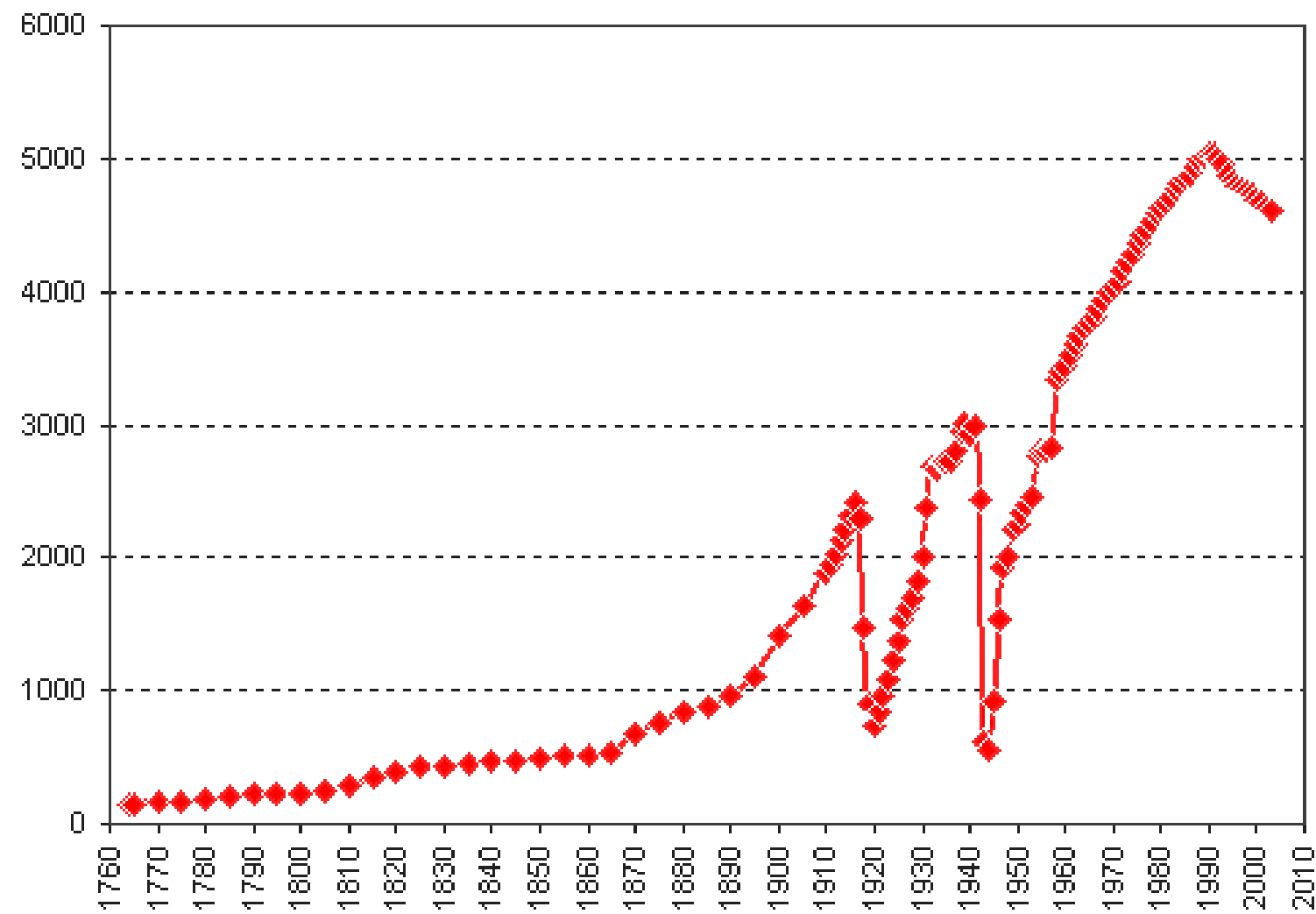
## 1. Котировки Акций



# Примеры

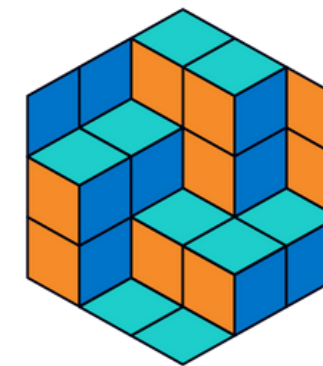


## 2. Численность населения Санкт-Петербурга

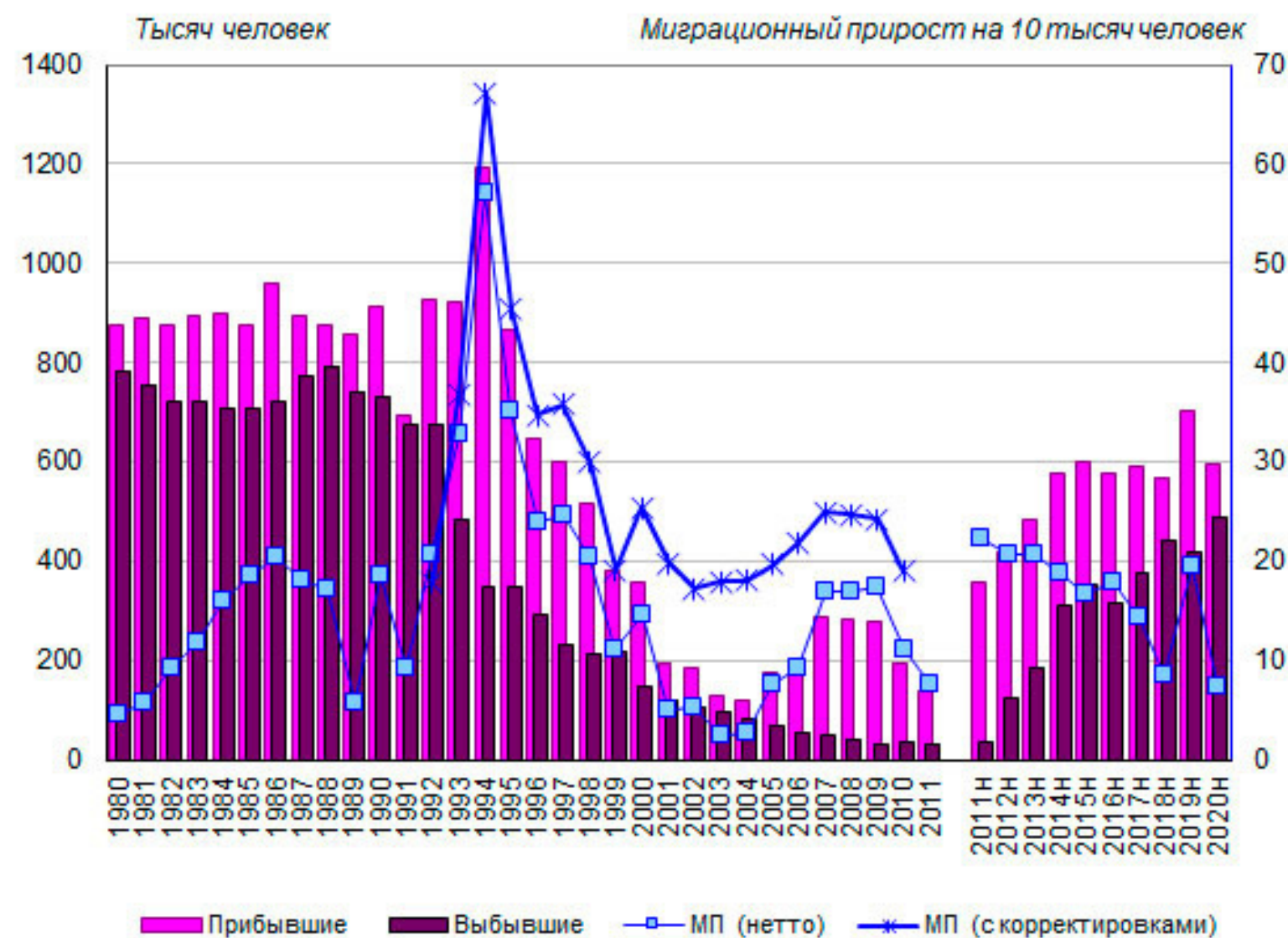




# Примеры

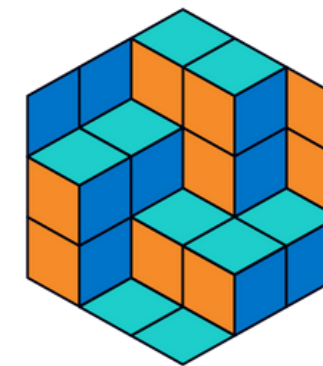


## 3. Миграционный прирост

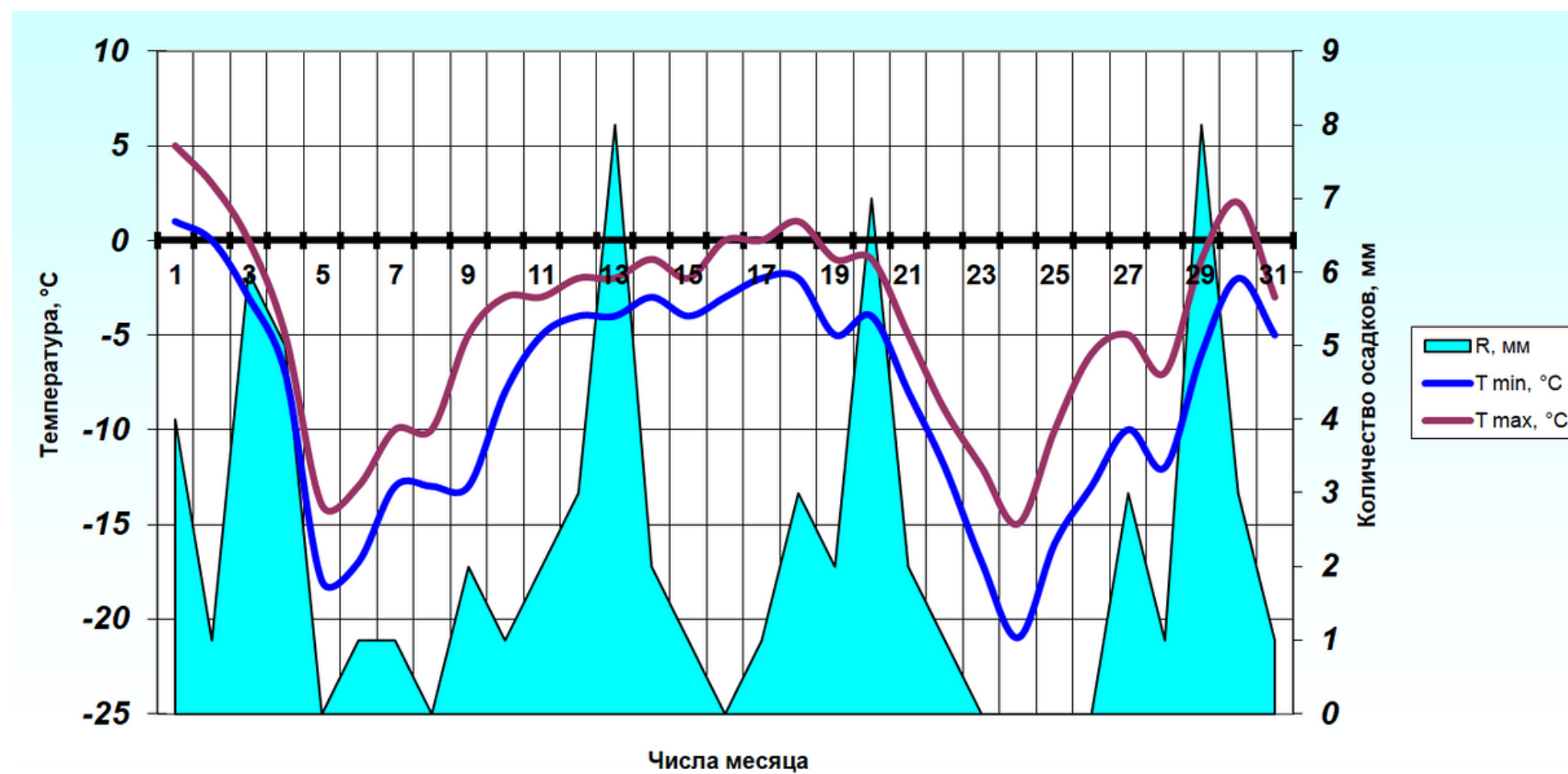




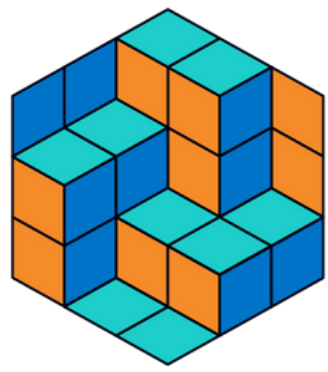
# Примеры



## 4. Изменение температуры в Москве

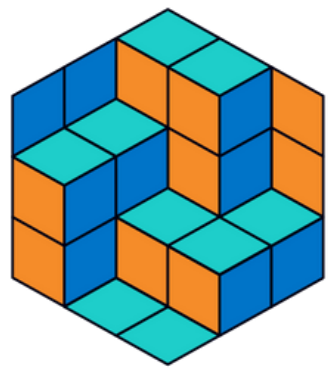


# Определение



**Временной ряд** - это данные, последовательно собранные в регулярные промежутки времени.

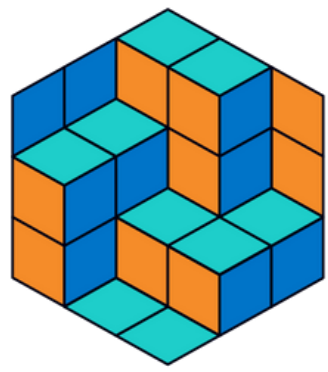
# Определение



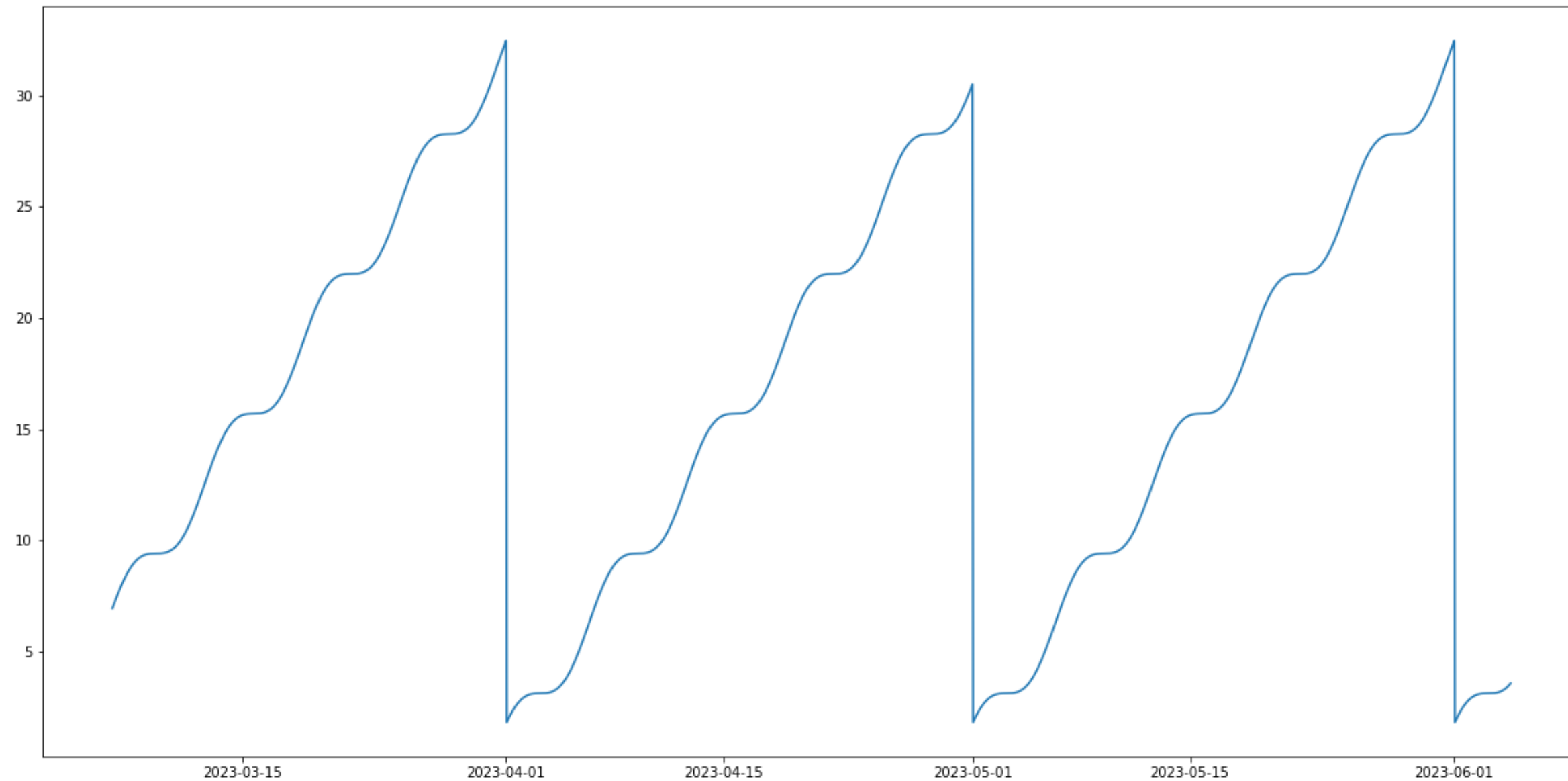
**Временной ряд** - это данные, последовательно собранные в регулярные промежутки времени.

**Задача:** продолжить имеющийся временной ряд.  
Фактически, необходимо угадать функцию, которая его задает.

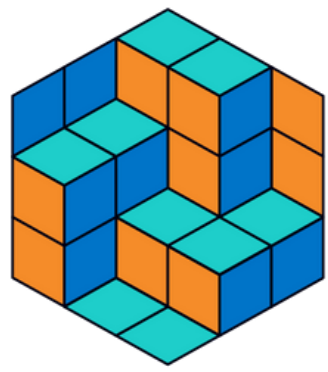
# Пример (2)



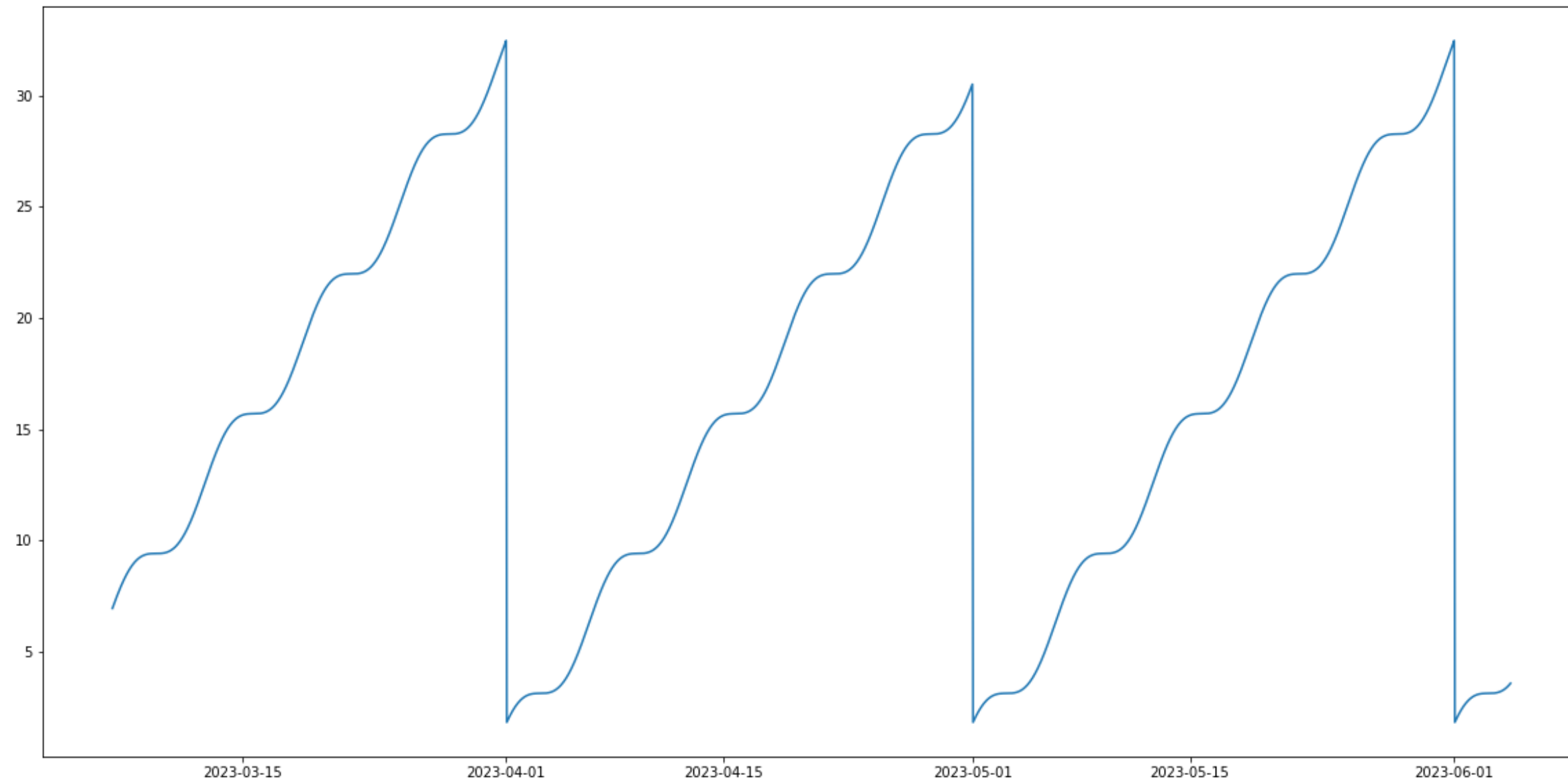
Какая функция задает этот ряд?



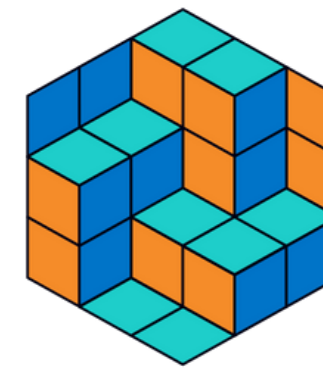
# Пример (2)



Какая функция задает этот ряд? Это  $y = x + \sin(x)$



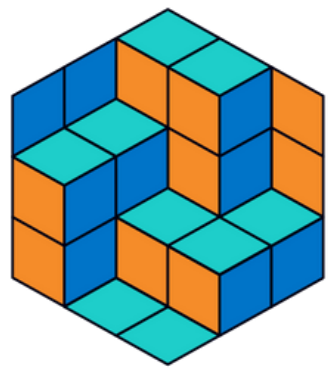
# Пример с шумом



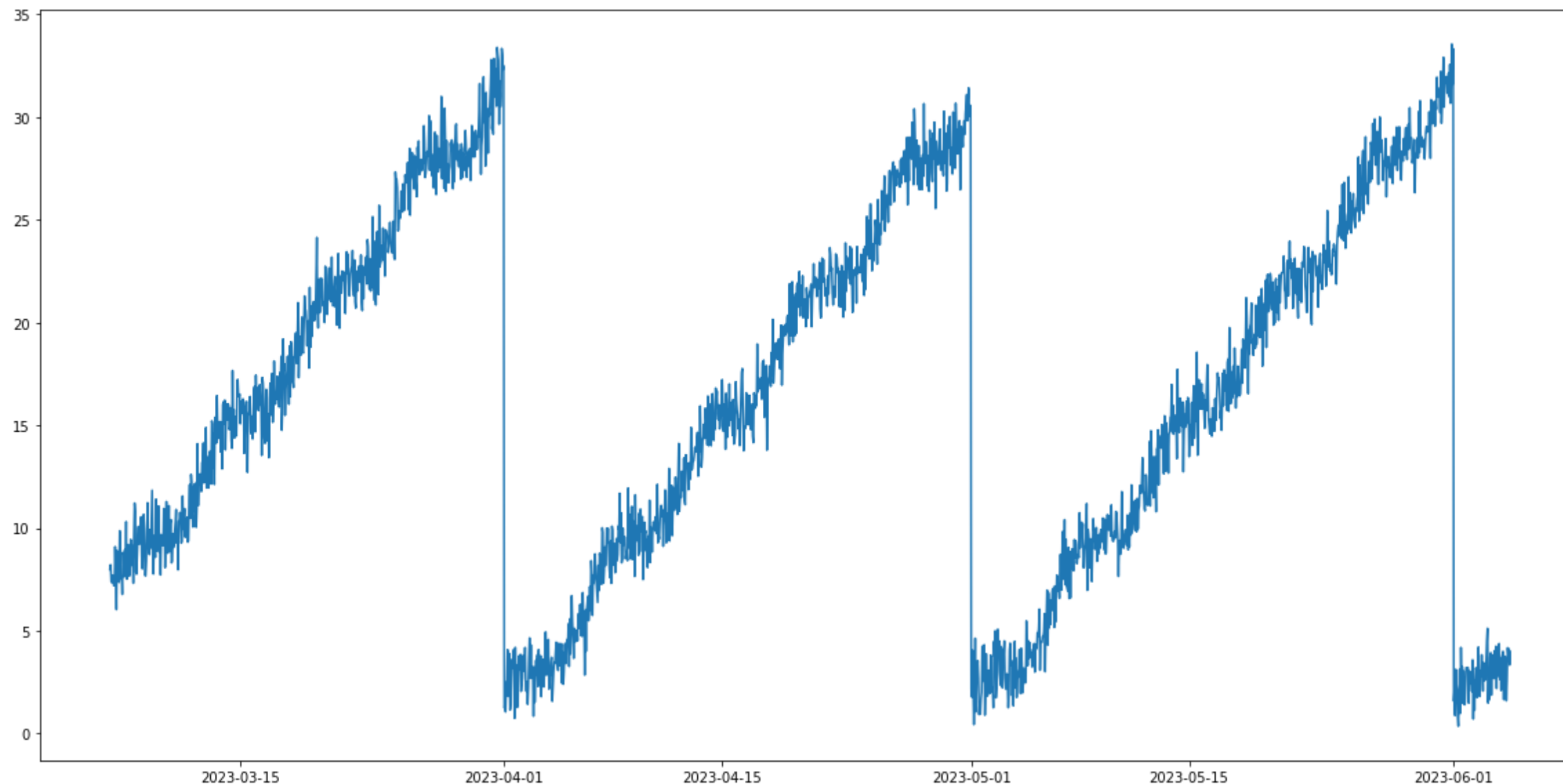
А если построить график  $y = x + \sin(x) + U$ ,  
где  $U$  имеет нормальное распределение  $N(0, 1)$



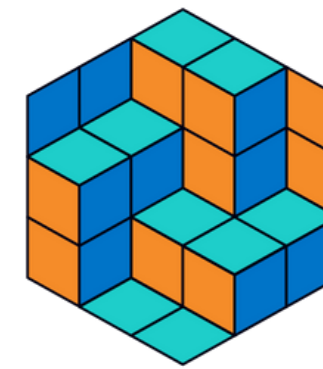
# Пример с шумом



А если построить график  $y = x + \sin(x) + U$ ,  
где  $U$  имеет нормальное распределение  $N(0, 1)$



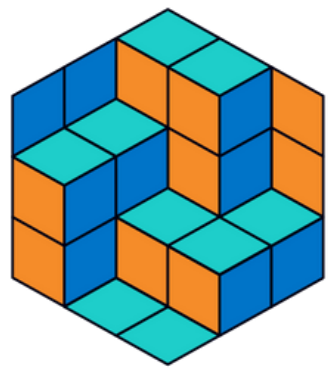
# Что делать?



Получается, угадать функцию не выйдет?



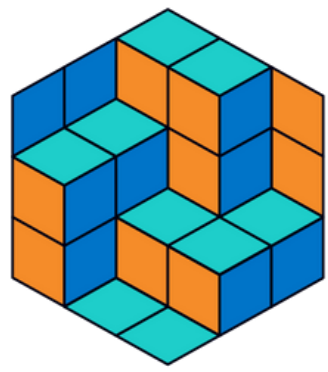
# Что делать?



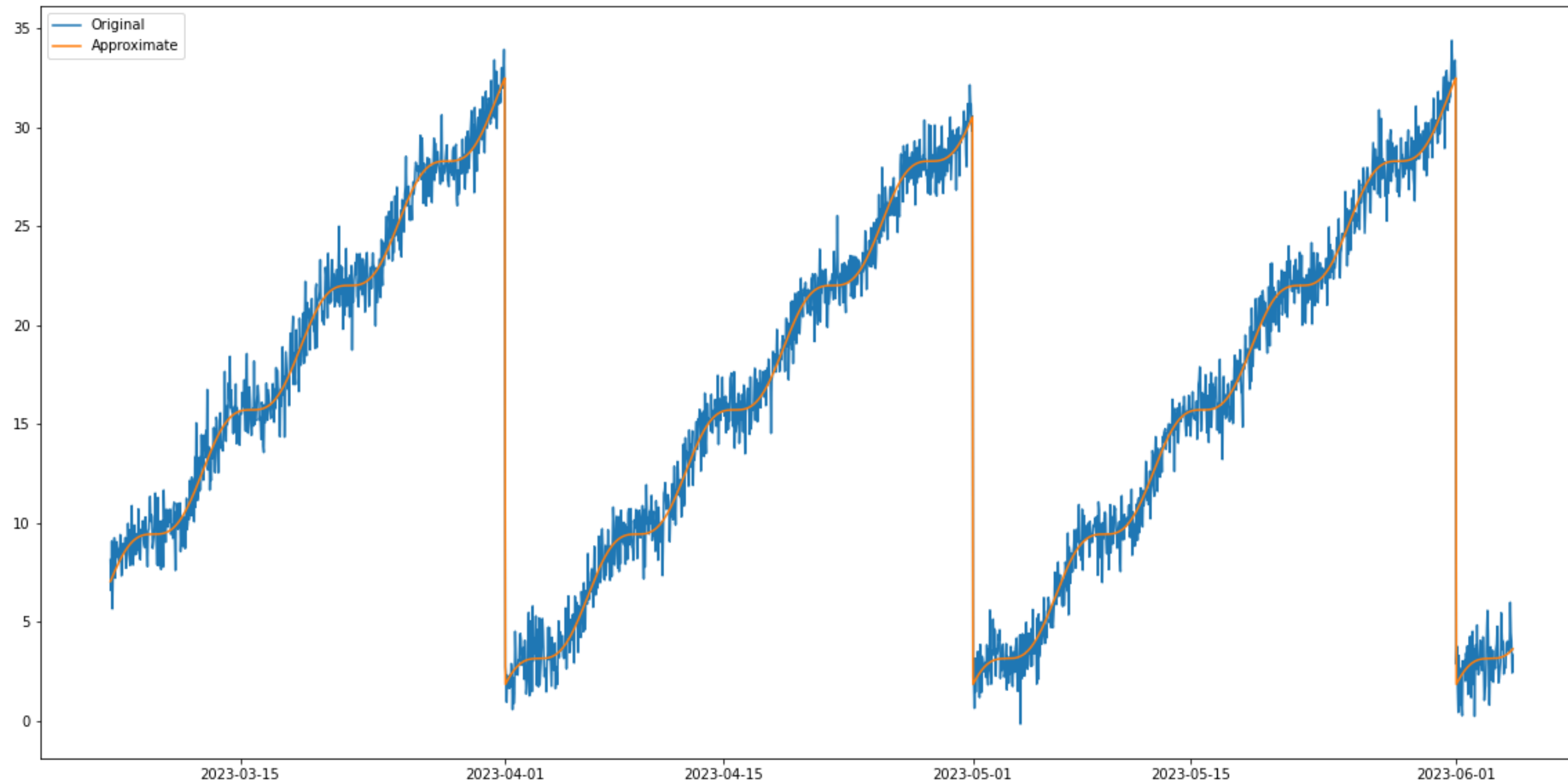
Получается, угадать функцию не выйдет?  
Да, но ведь можно ее приблизить!

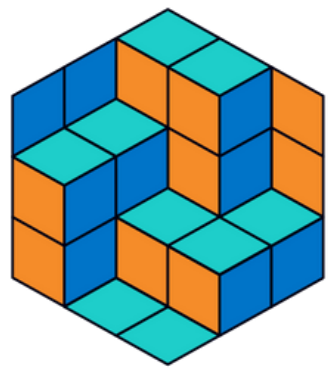


# Что делать?



Очевидно, что  $y = x + \sin(x) + U$  приближается  $y = x + \sin(x)$

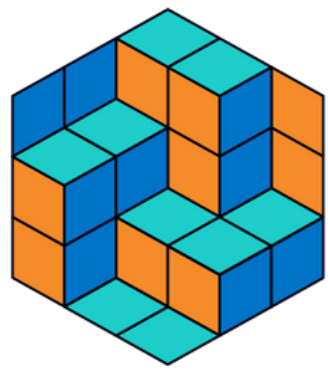




# **Глава 2**

## **Простые способы предсказать временной ряд**

# Наивная модель

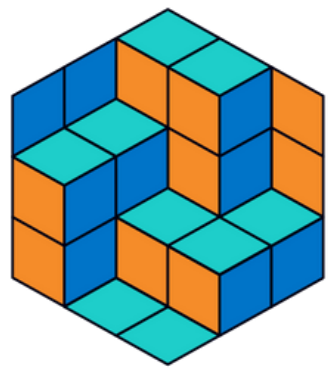


Какая модель тут будет считаться "наивной"?\*

\* "Naive Model" в словаре ML-щика - простая, банальная, почти всегда бесполезная, но почему-то существующая модель



# Наивная модель



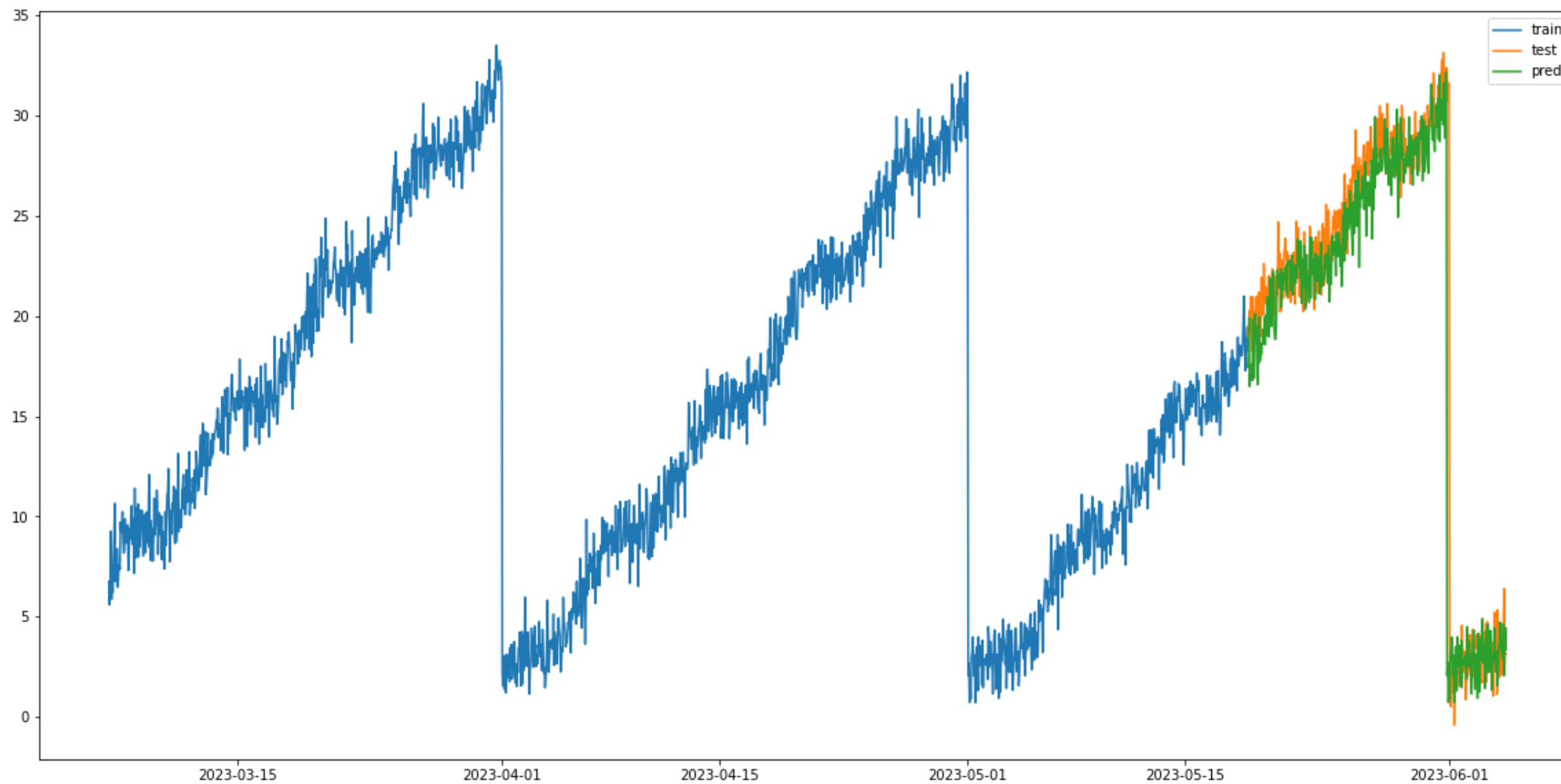
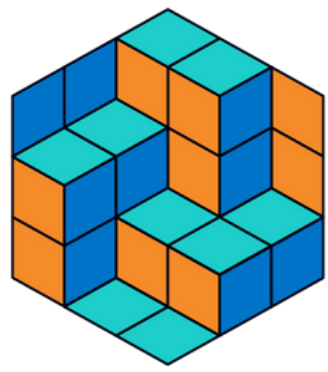
Какая модель тут будет считаться "наивной"?\*

Наивной моделью считаем ту, которая:  $y[i] = y[i - lag]$

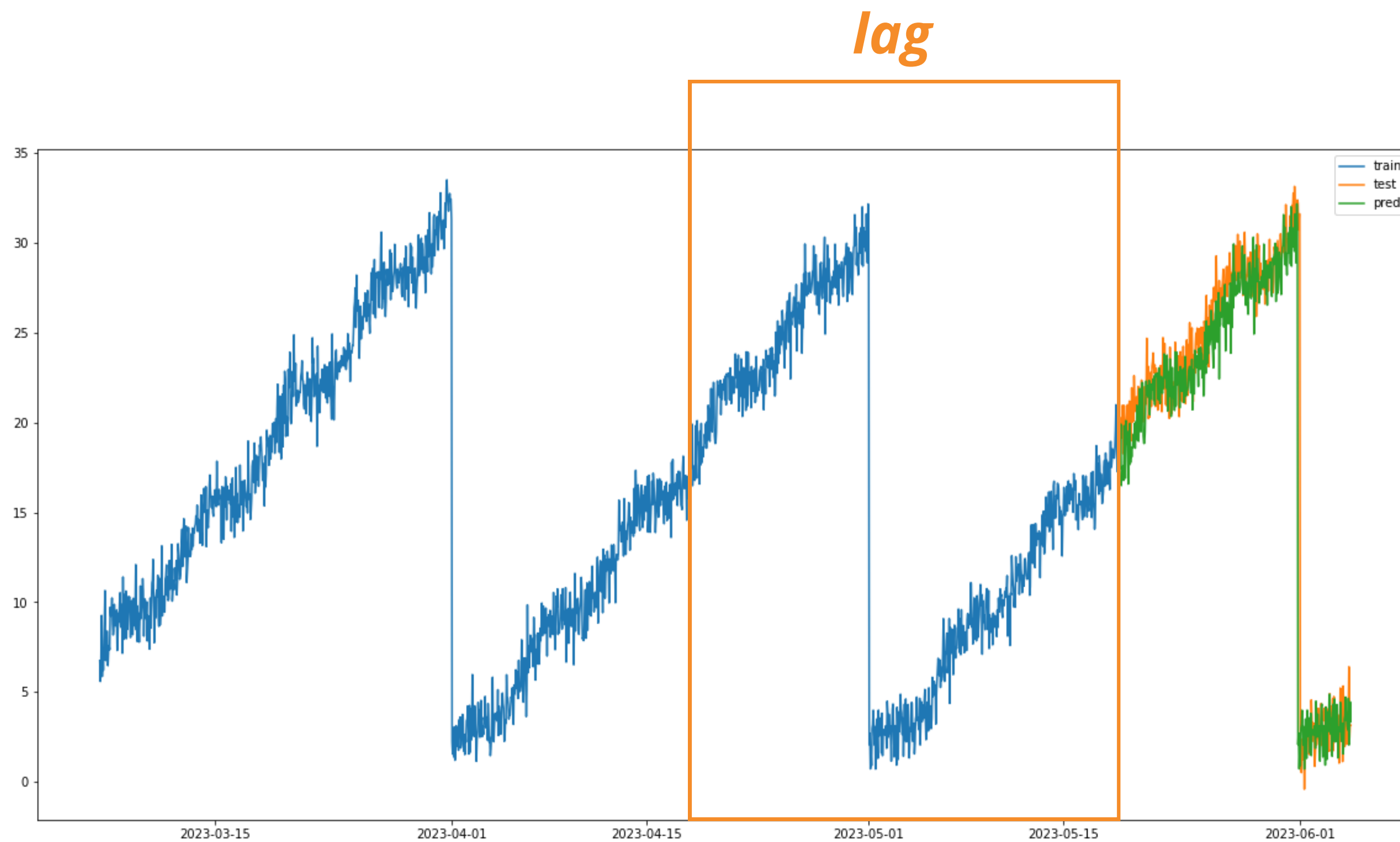
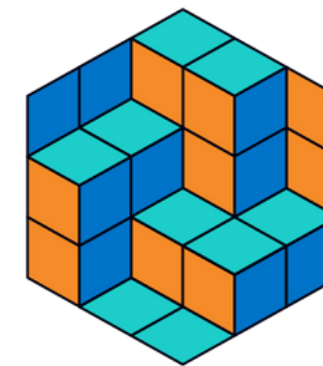
То есть, в качестве предсказанного нового значения она будет брать то, которое было  $lag$  позиций назад.

\* "Naive Model" в словаре ML-щика - простая, банальная, почти всегда бесполезная, но почему-то существующая модель

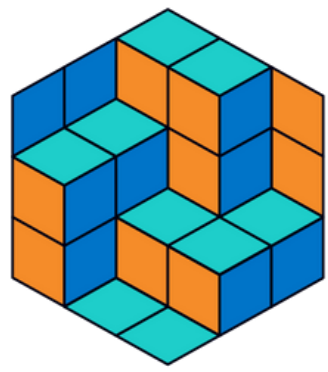
# Наивная модель



# Наивная модель



# Наивная модель



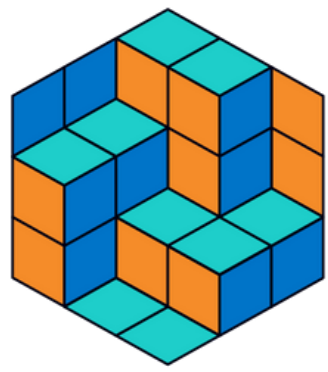
Плюсы:

- Работает очень просто
- По сути ничего обучать не нужно

Минусы:

- Модель не способна предсказать что-то новое
- Ей не важен контекст - ей важно только то, что было *lag* шагов назад

# Пример

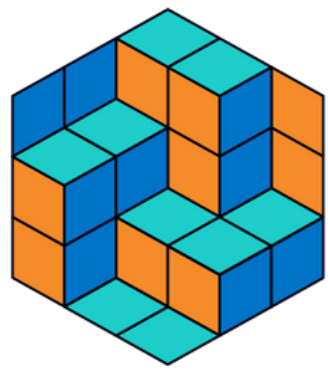


Посмотрим на изменение цены акций Visa  
Наивная модель не сработает  
Как нам определиться с трендом\*? (хотя бы)

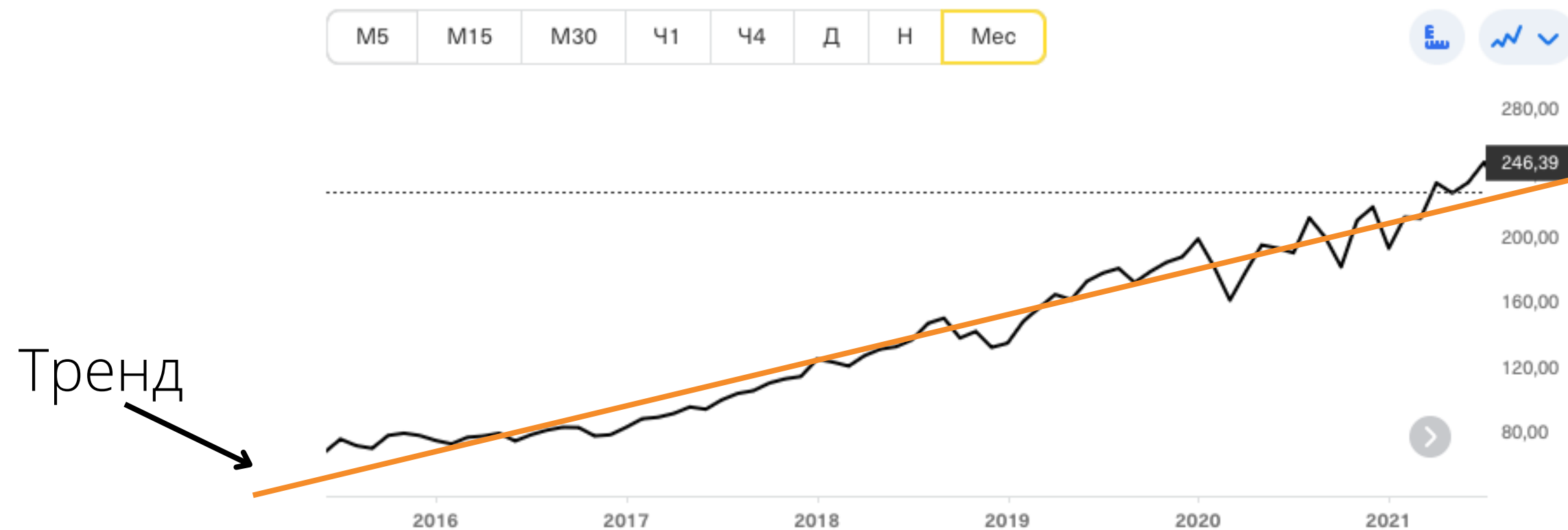


\* тренд - основная тенденция изменения чего-либо (в нашем случае целевого значения - цены)

# Пример



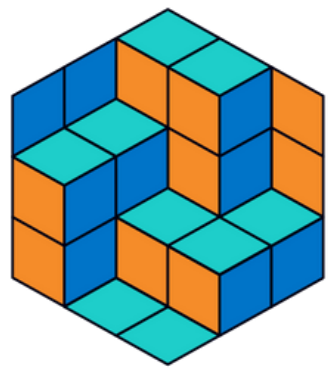
Посмотрим на изменение цены акций Visa  
Наивная модель не сработает  
Как нам определиться с трендом\*? (хотя бы)



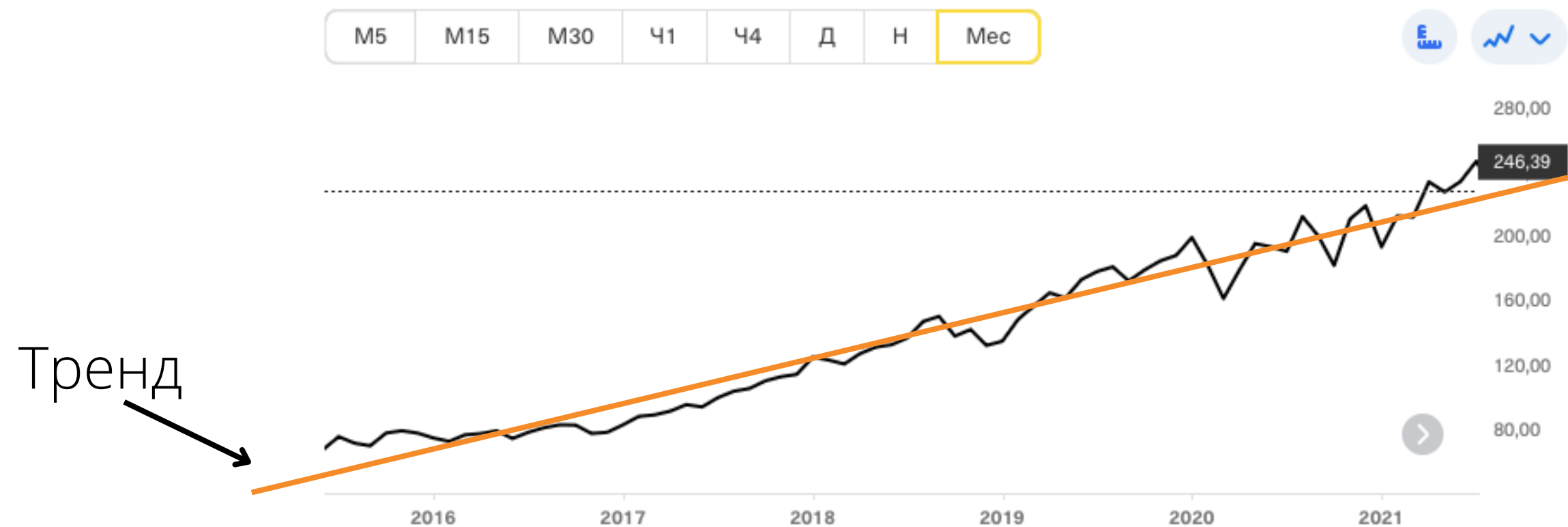
\* тренд - основная тенденция изменения чего-либо (в нашем случае целевого значения - цены)



# Пример

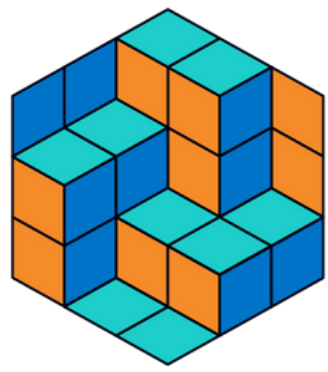


Посмотрим на изменение цены акций Visa  
Вопрос 2: как быстро растет цена?

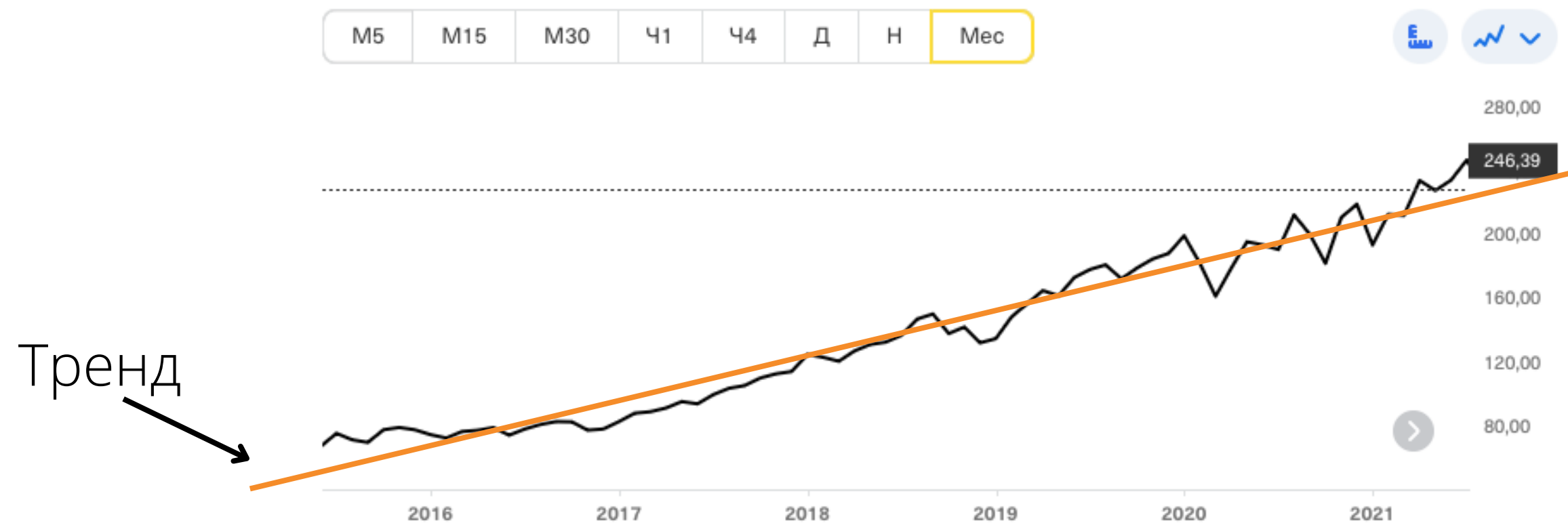


\* тренд - основная тенденция изменения чего-либо (в нашем случае целевого значения - цены)

# Пример



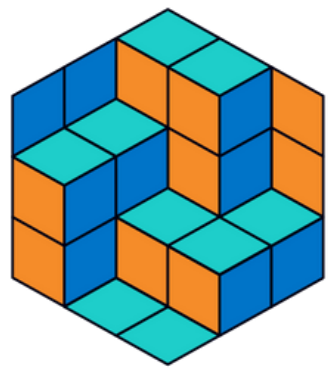
Посмотрим на изменение цены акций Visa  
Вопрос 2: как быстро растет цена?



$$y=kx+b$$

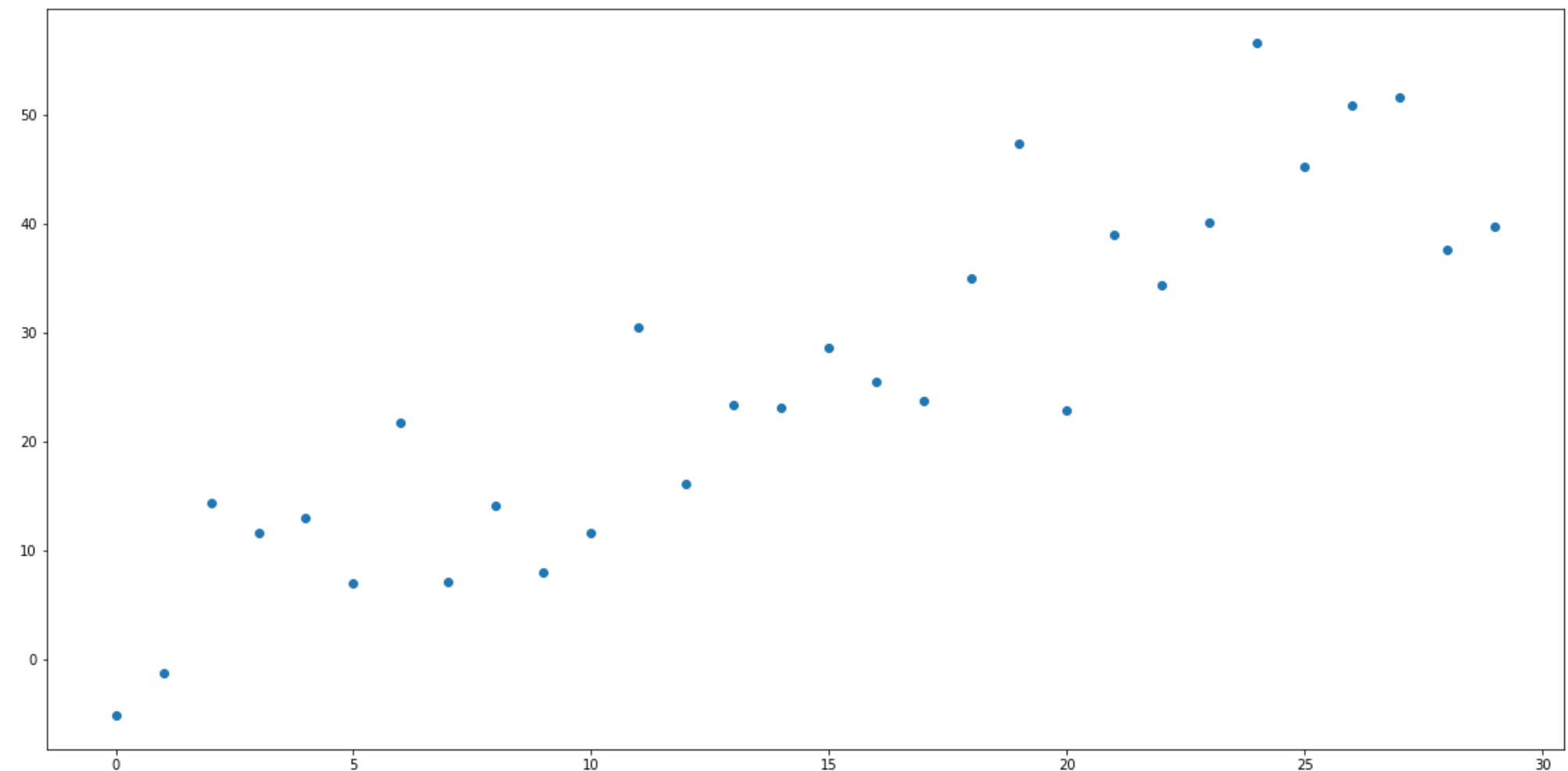
\* тренд - основная тенденция изменения чего-либо (в нашем случае целевого значения - цены)

# Линейная регрессия

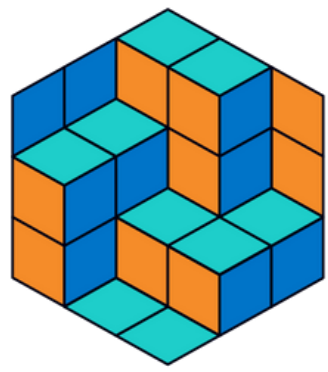


**Дано:** набор точек вида  $(x, y)$

**Задача:** построить наиболее близкую к ним прямую вида  $y = kx + b$



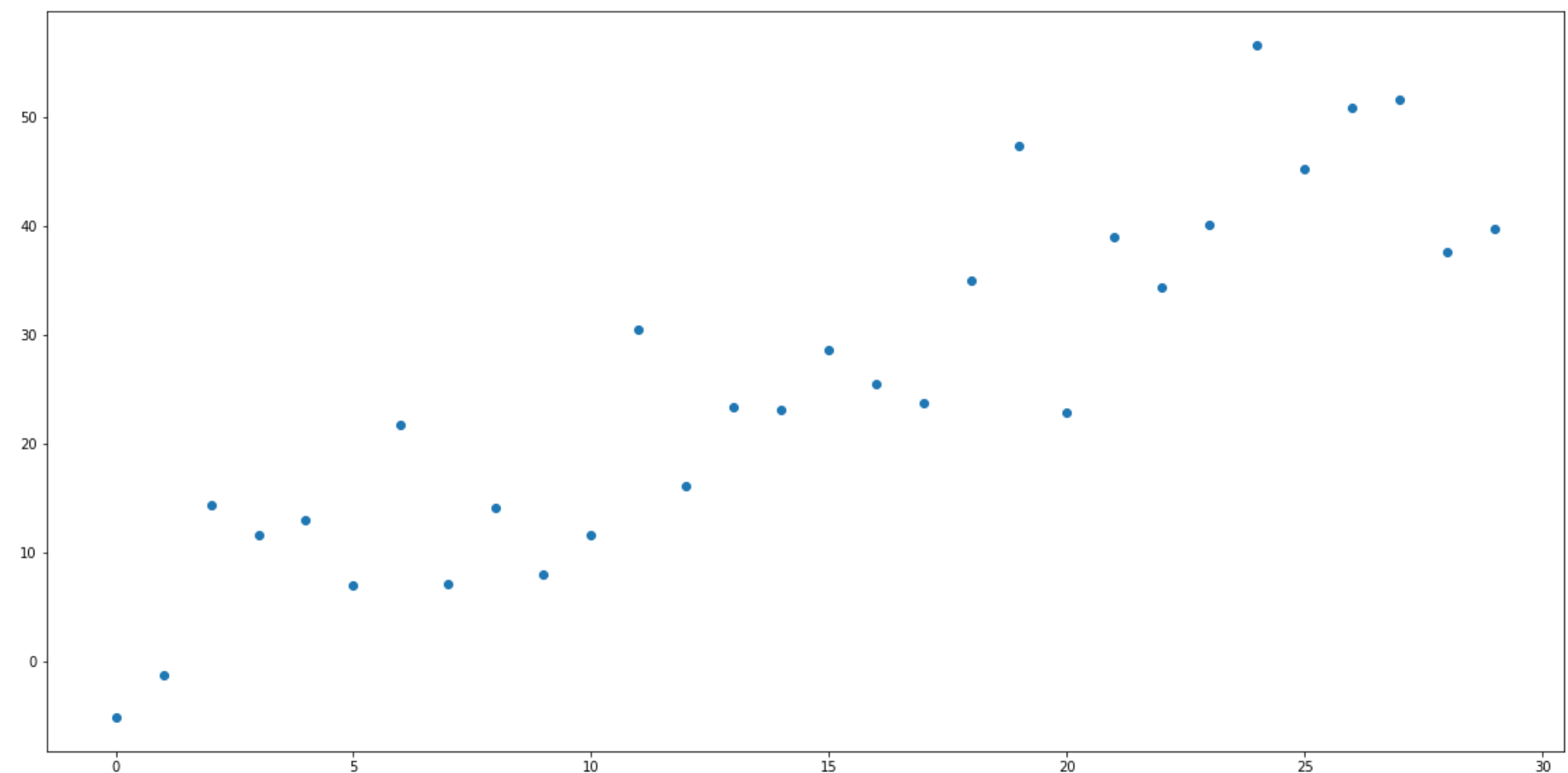
# Линейная регрессия



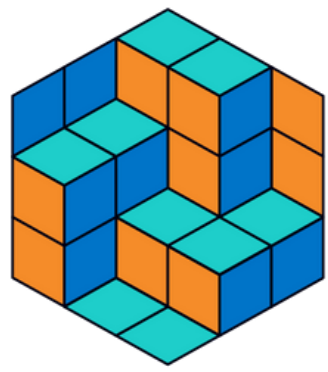
**Дано:** набор точек вида  $(x, y)$

**Задача:** построить наиболее близкую к ним прямую вида  $y = kx + b$

**Оптимизация:** MSE (МНК)



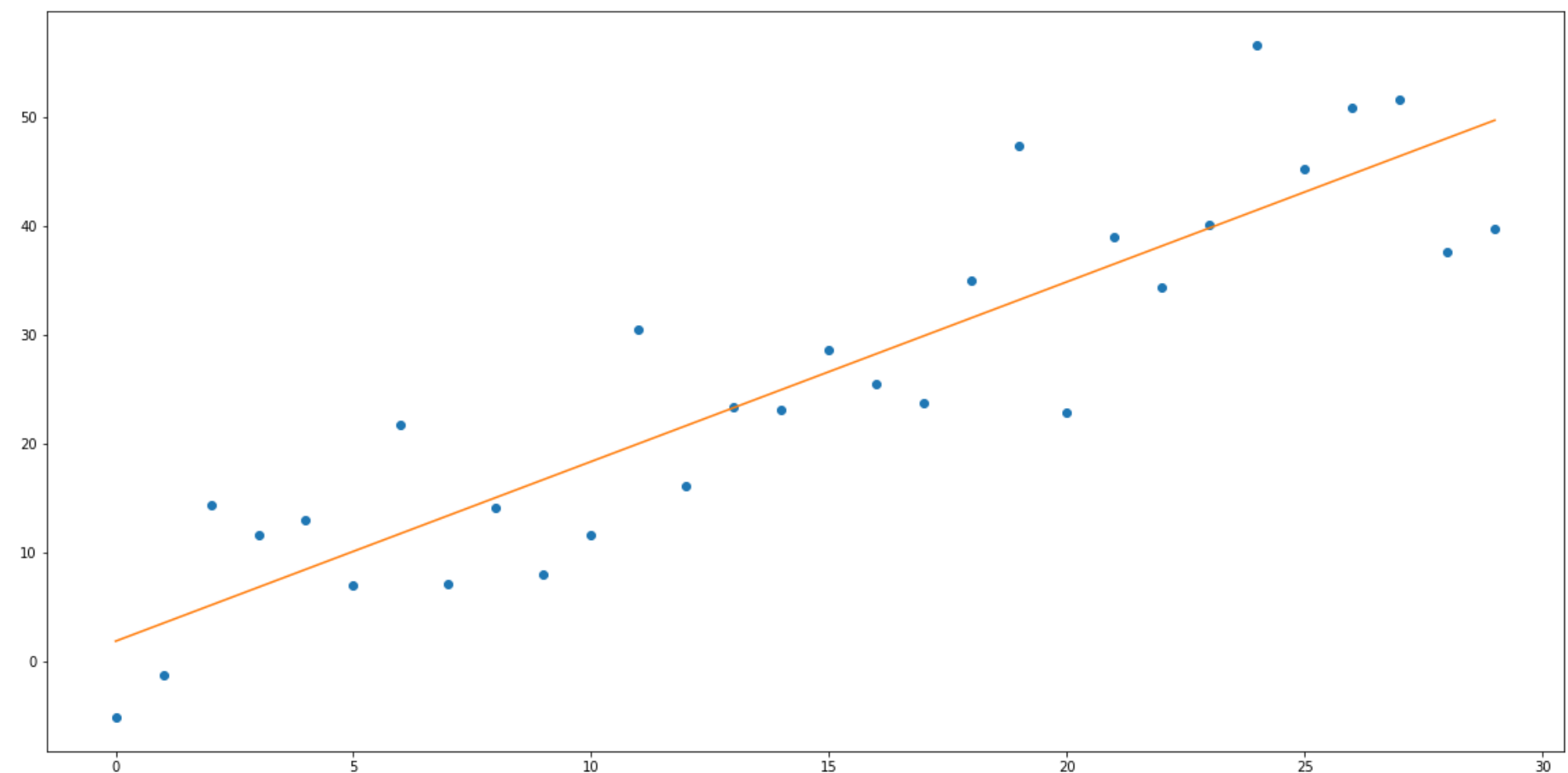
# Линейная регрессия



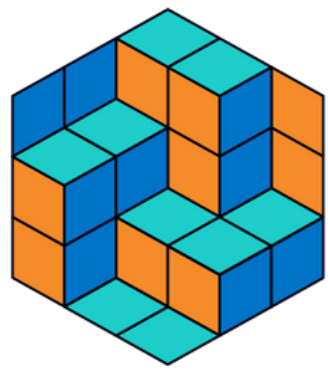
**Дано:** набор точек вида  $(x, y)$

**Задача:** построить наиболее близкую к ним прямую вида  $y = kx + b$

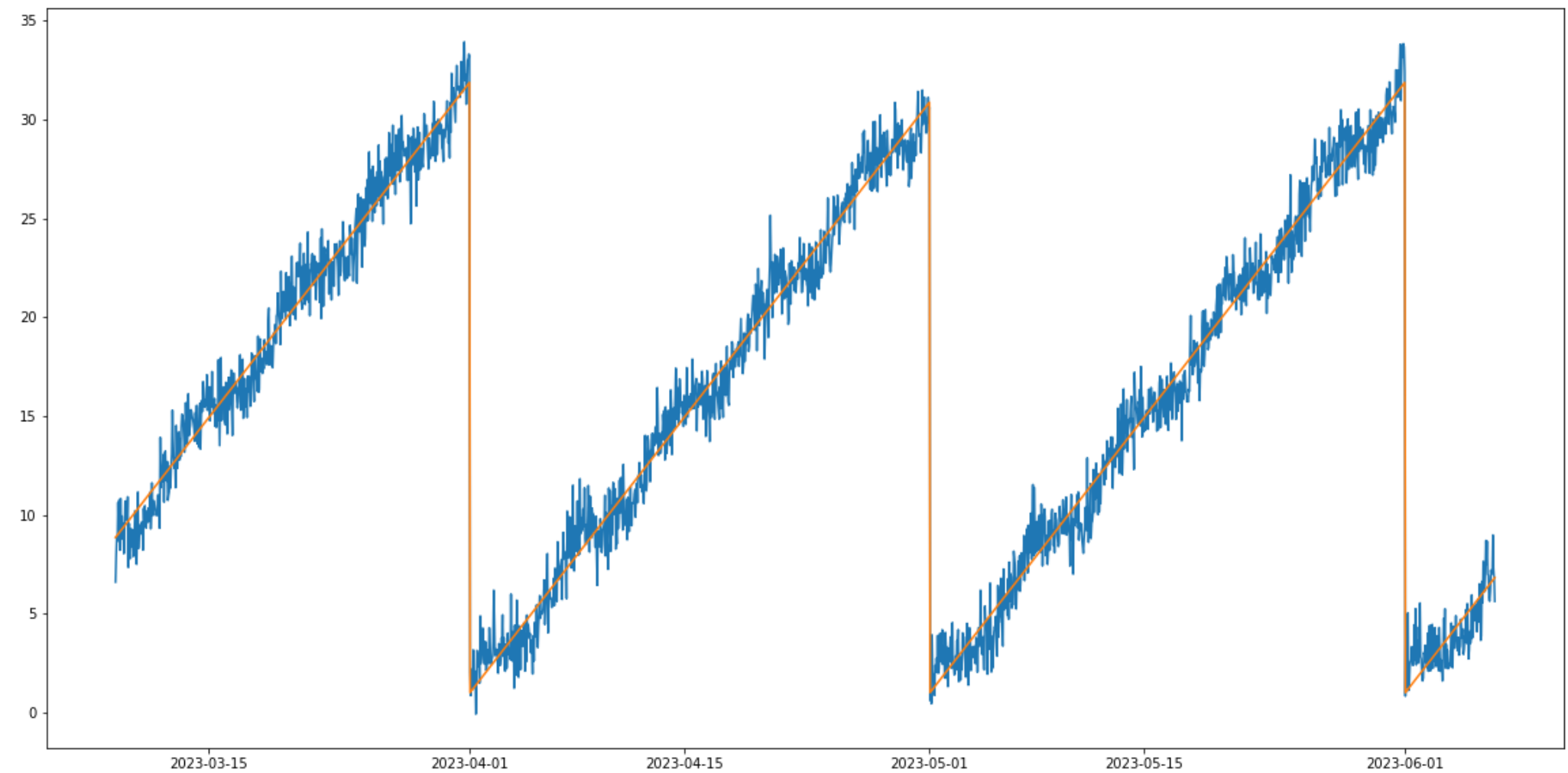
**Оптимизация:** MSE (МНК)



# Линейная регрессия

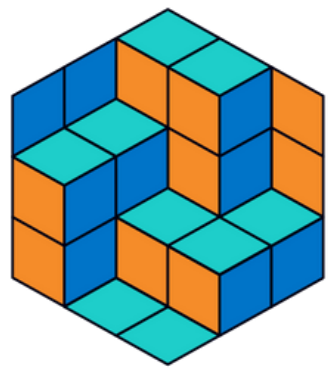


Удивительно, но для предыдущего примера все так же может работать хорошо!





# Линейная регрессия



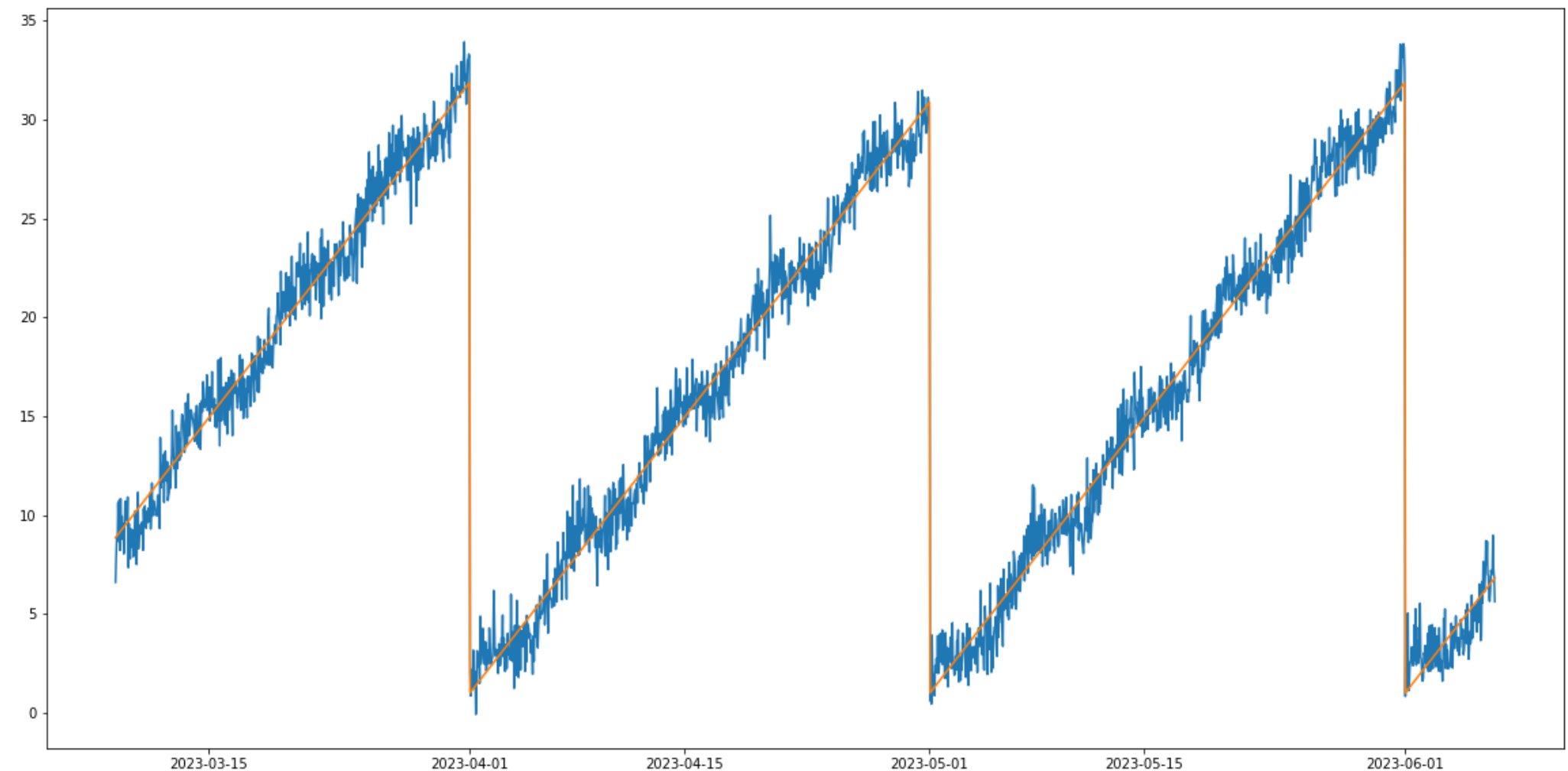
Удивительно, но для предыдущего примера все так же может работать хорошо!

## Почему получилось?

Преобразование координаты

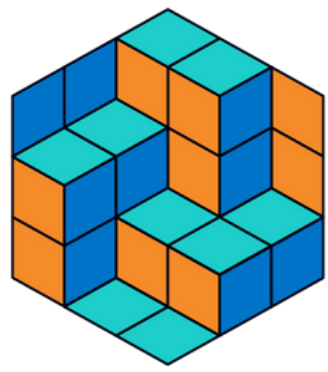
## Почему так вообще можно делать?

Потому что есть *сезонность*\*

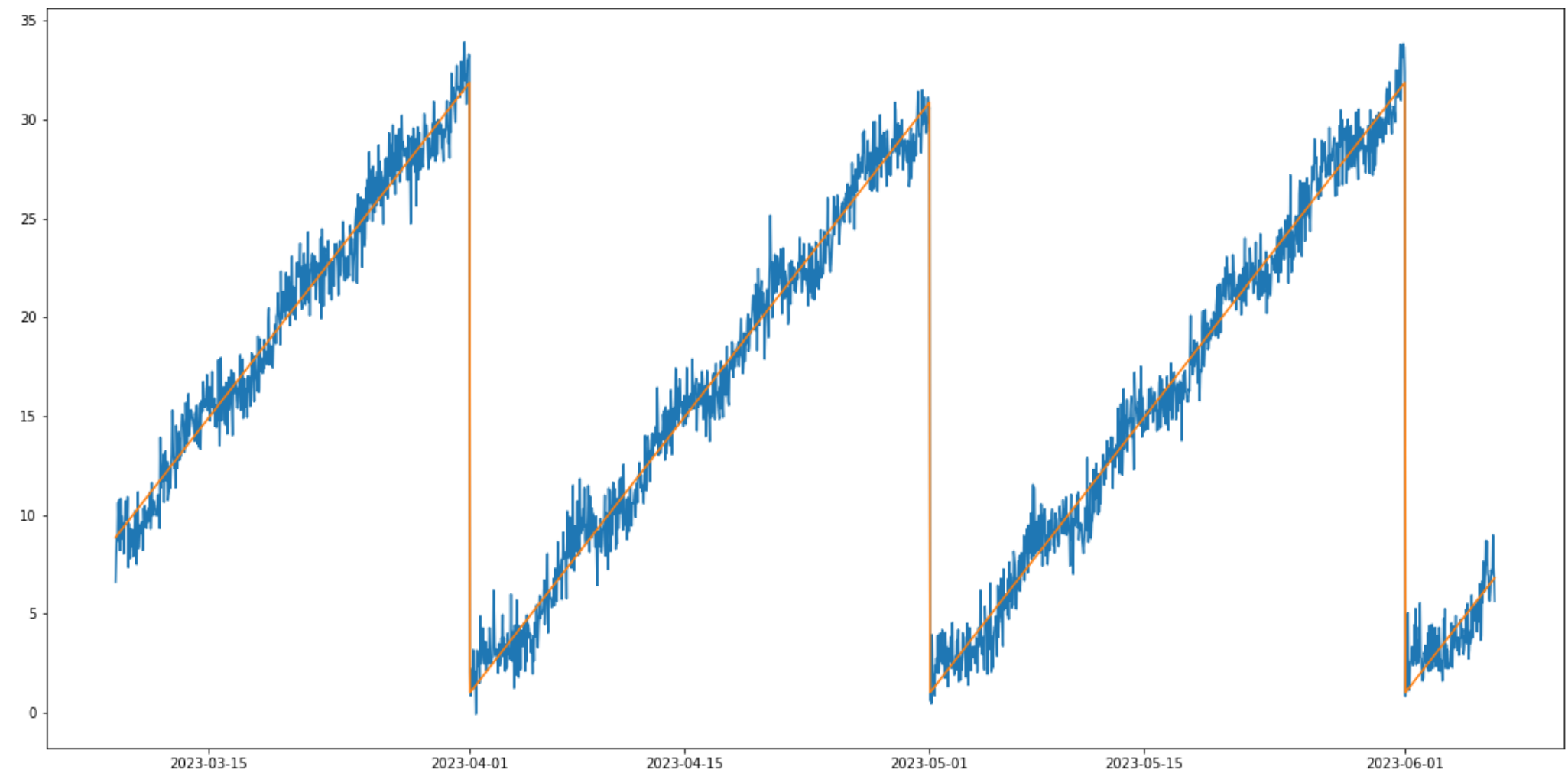


\* сезонность - явление повторения значения через определенный промежуток времени

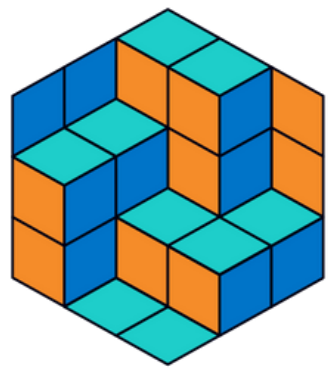
# Линейная регрессия 2



А можно сделать лучше?



# Линейная регрессия 2

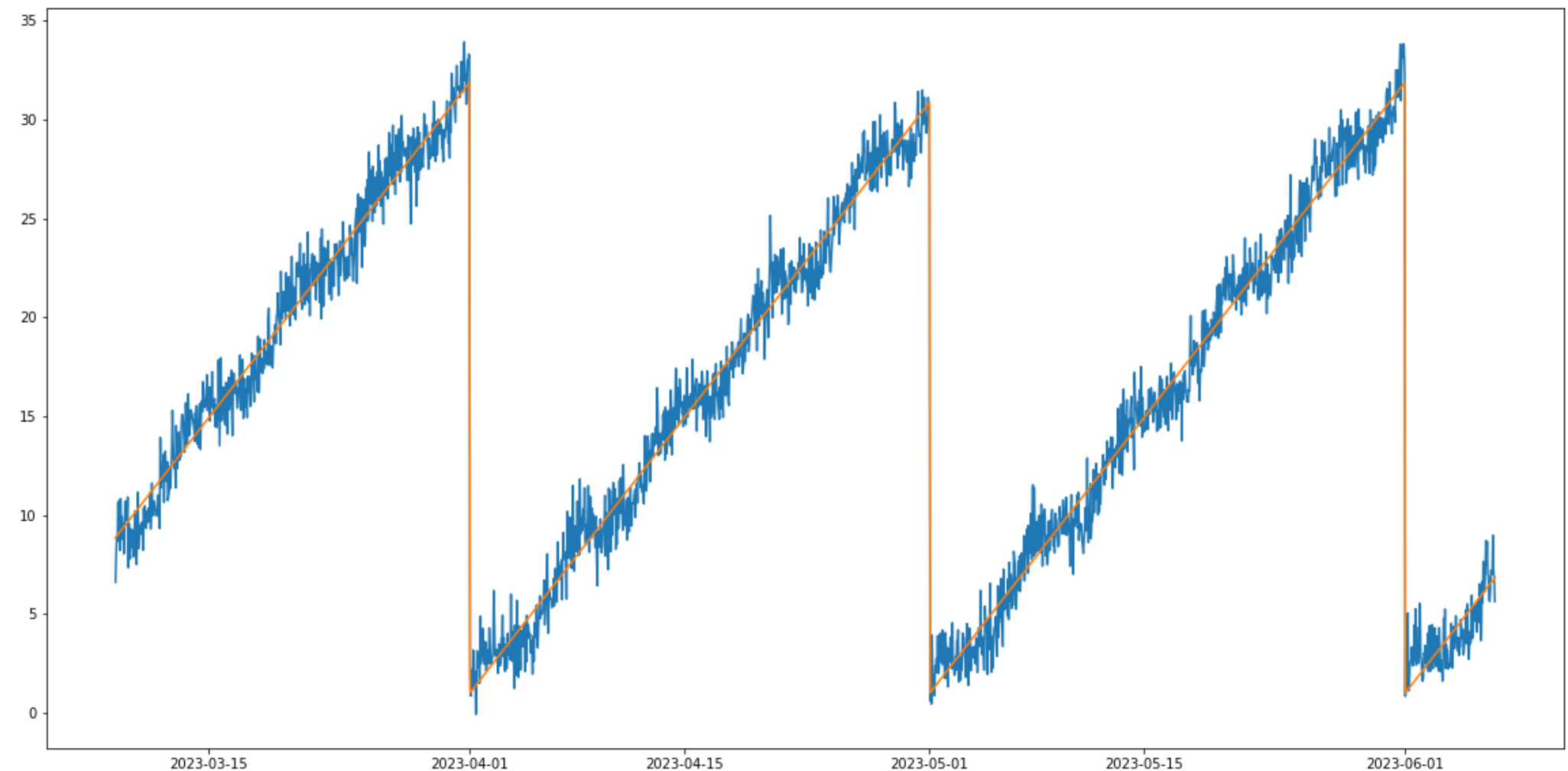


А можно сделать лучше?

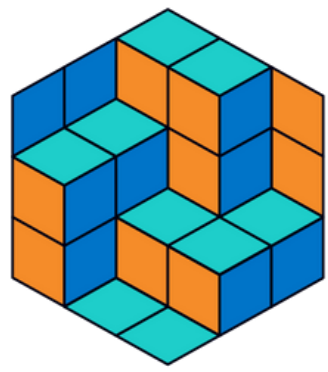
Добавим *экзогенные* переменные

Сделаем из  $y = kx$

$$y = \sum_{i=1}^n k_i x_i$$



# Линейная регрессия 2

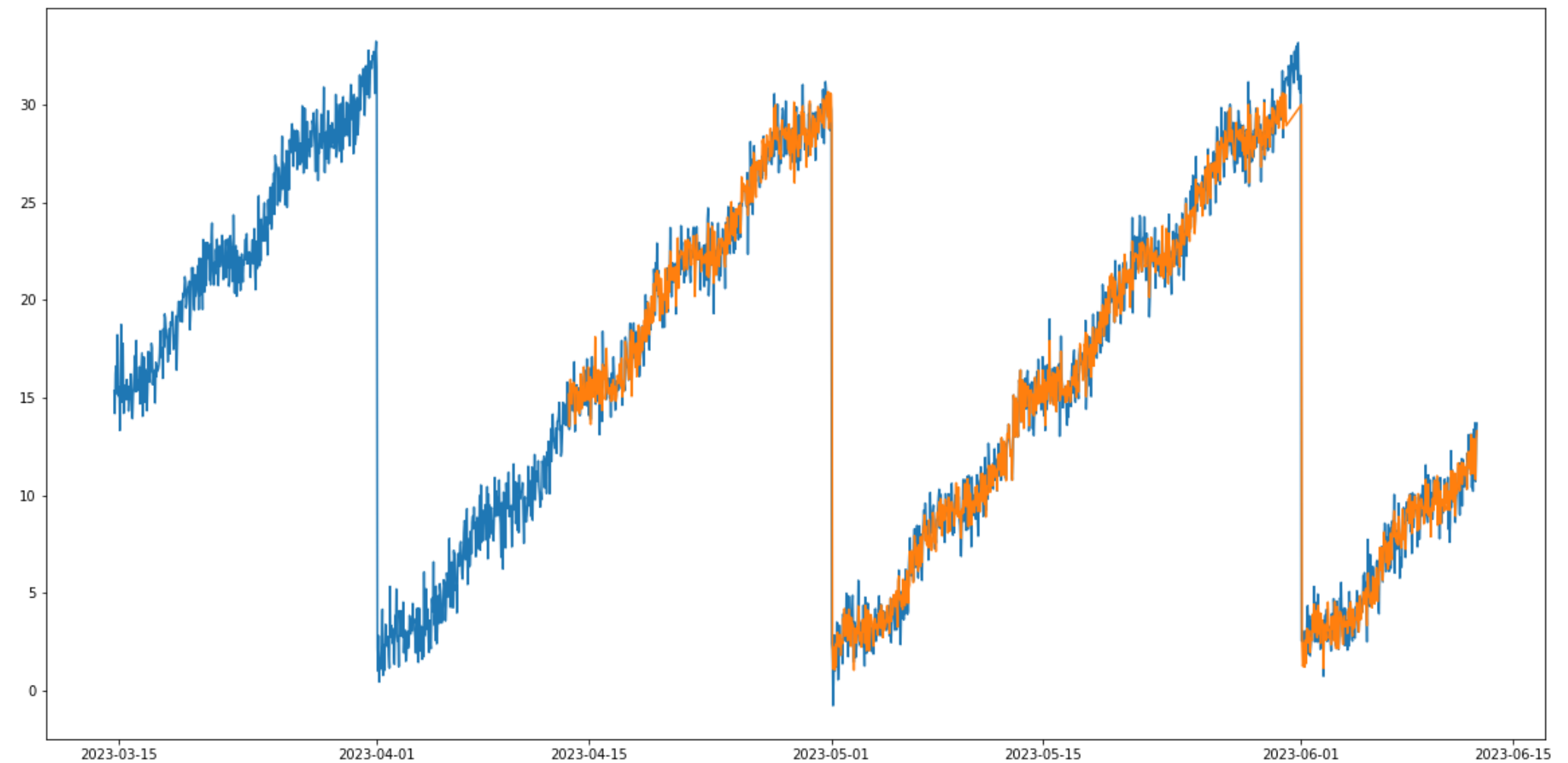


А можно сделать лучше?

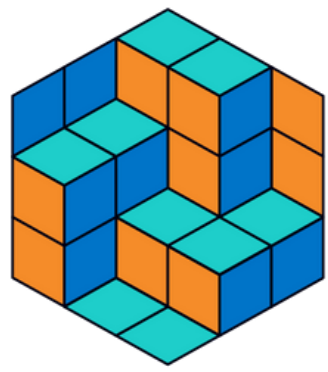
Добавим *экзогенные* переменные

Сделаем из  $y = kx$

$$y = \sum_{i=1}^n k_i x_i$$



# Линейная регрессия 2

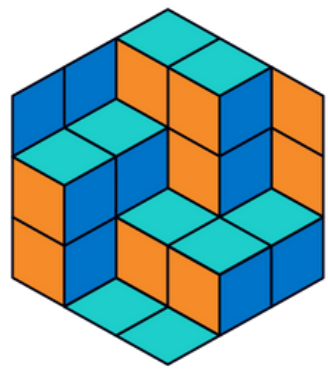


Плюсы:

- более точно

Минусы:

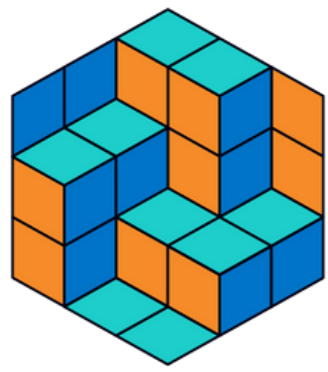
- очень много параметров
- делает только то, что видела



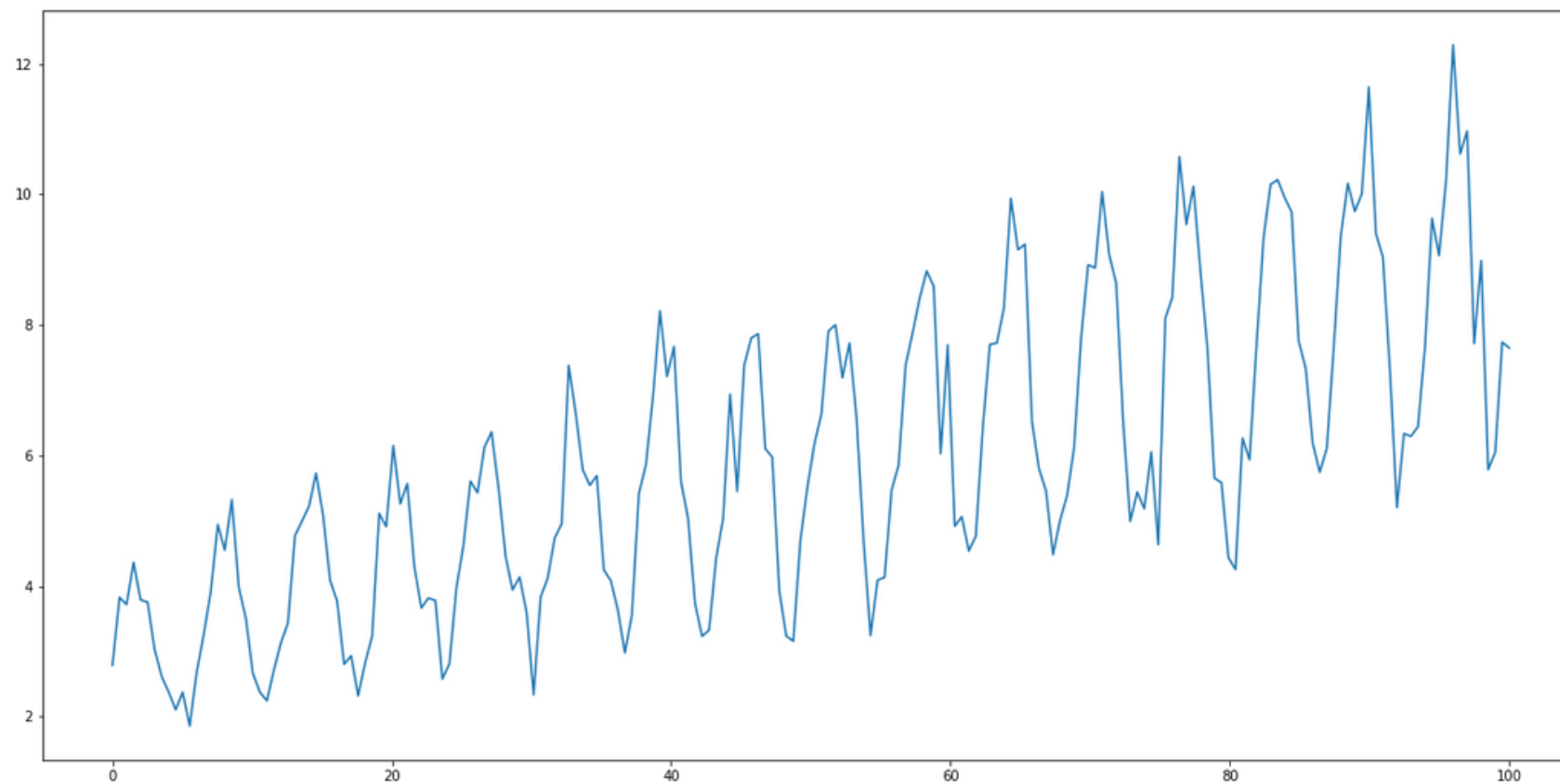
# Глава 3

## Moving Average & ARIMA

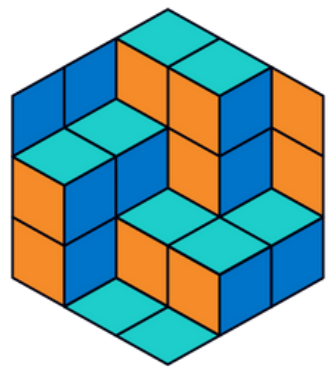
# Стационарность



Что не так с этим графиком?

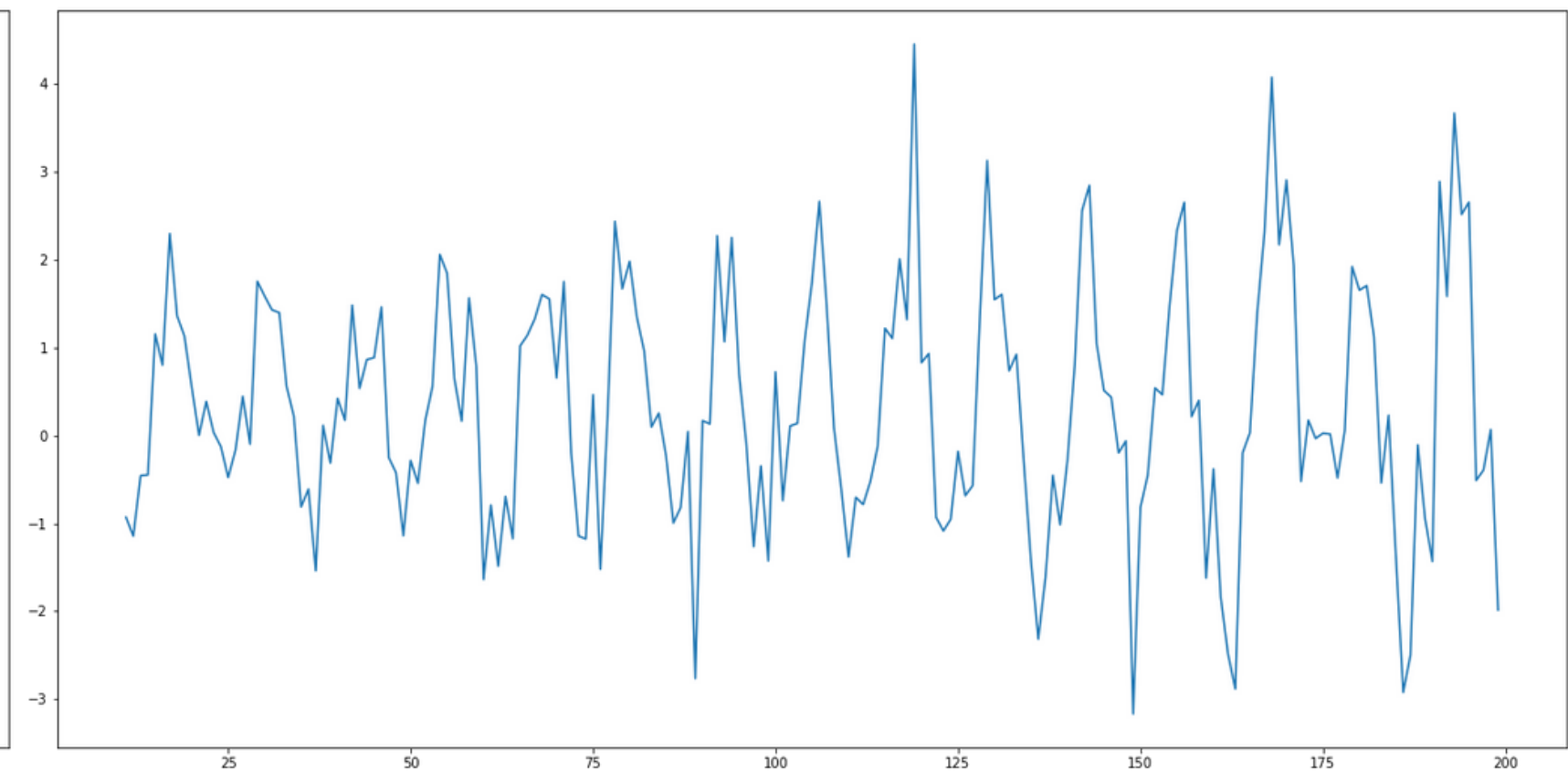
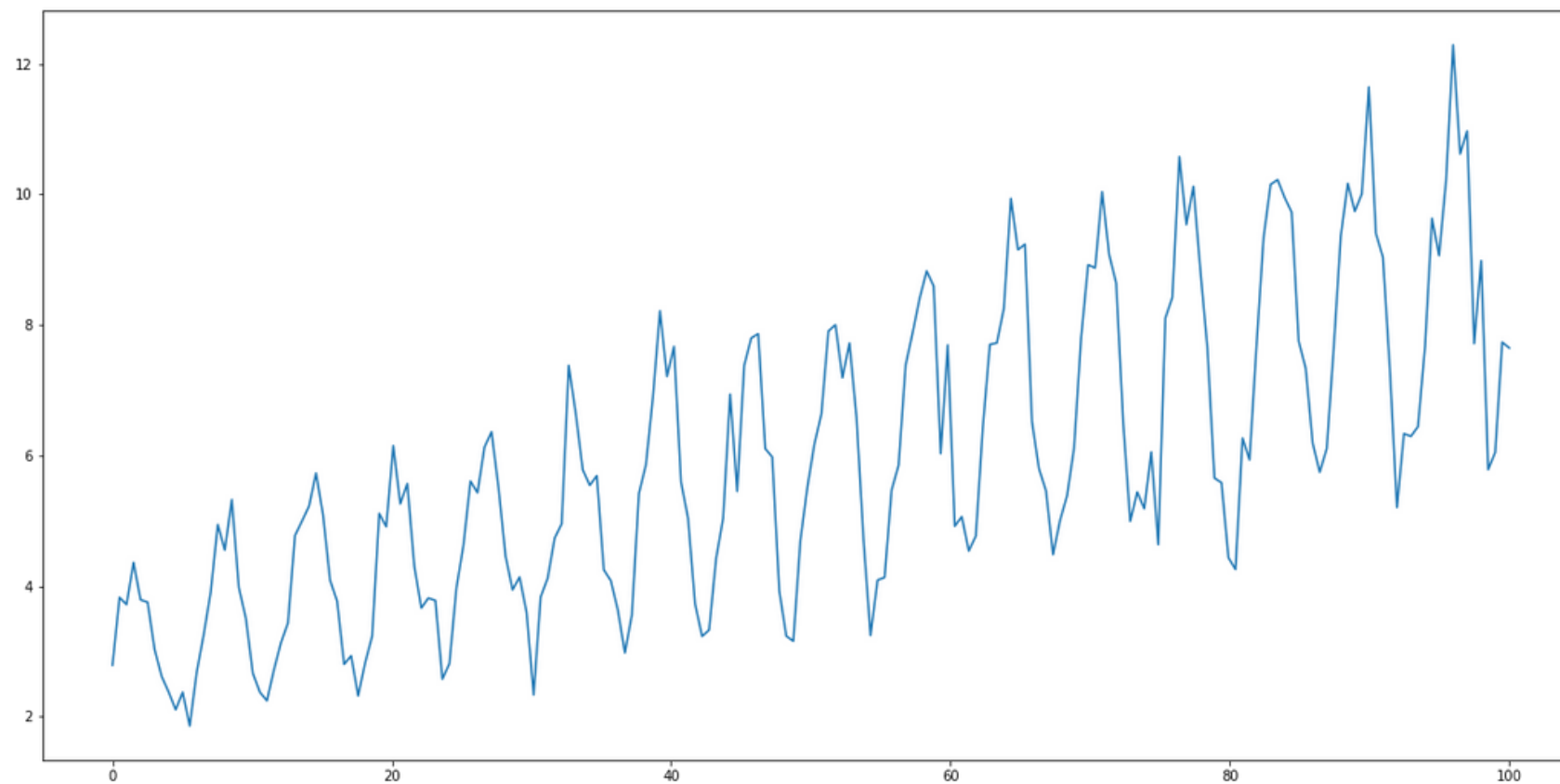


# Стационарность



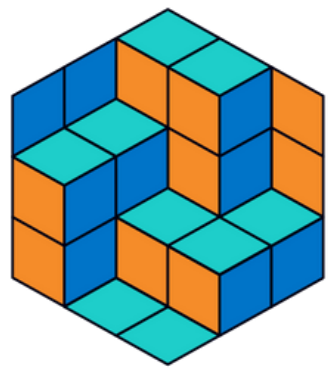
Что не так с этим графиком?

Приведем его к стационарному виду





# Теорема Волда



Каждый *слабо стационарный* временной ряд можно представить в виде *скользящего среднего* бесконечного порядка  $MA(\infty)$

Такое представление называют *представлением скользящим средним* для временных рядов.

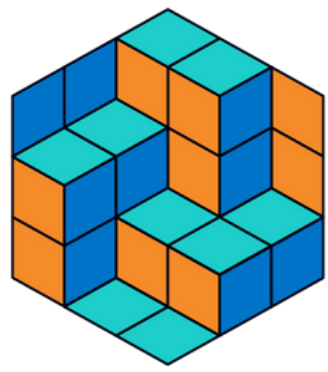
$$Y_t = \sum_{i=1}^{\infty} b_i \epsilon_{t-i} + \nu_t$$

- $Y_t$  - рассматриваемый временной ряд
- $\epsilon_{t-i}$  - белый шум
- $b_i$  - коэффициенты скользящего среднего
- $\nu_t$  - детерминированная компонента (равна 0 если нет трендов)

Коэффициенты  $b_i$  удовлетворяют условиям,

- ряд сходится абсолютно
- отсутствуют члены с  $j < 0$
- не зависят от  $t$

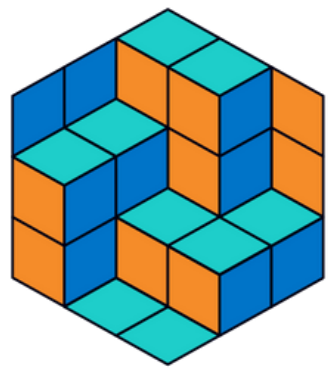
# MA(d)



Очевидно,  $MA(\infty)$  мы построить не можем, поэтому построим **MA(d)**

Но как тогда найти коэффициенты  $\epsilon_{t-i}$  ?

# AR(p)

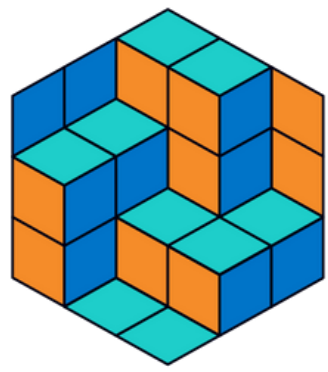


Для начала посмотрим на модель **AR** - *авторегрессивная модель*

$$Y_t = \sum_{j=1}^p Y_{t-j} \alpha_j + \epsilon_t$$


Эта модель полагается только на предыдущие значения - зная, что предыдущие значения были **2,3,4**, модель предскажет 5 (как пример)

# AR(p)



Для начала посмотрим на модель **AR** - *авторегрессивная модель*

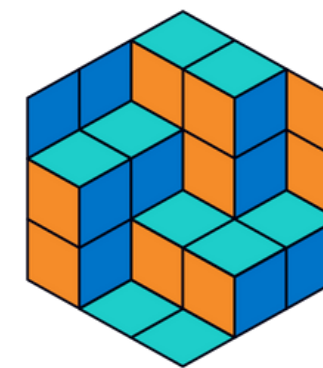
$$Y_t = \sum_{j=1}^p Y_{t-j} \alpha_j + \epsilon_t$$

Она нам и нужна! 

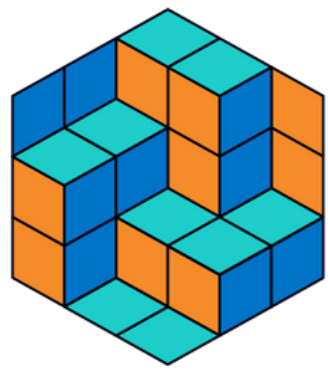
Эта модель полагается только на предыдущие значения - зная, что предыдущие значения были **2,3,4**, модель предскажет 5 (как пример)

?

Что делать дальше?

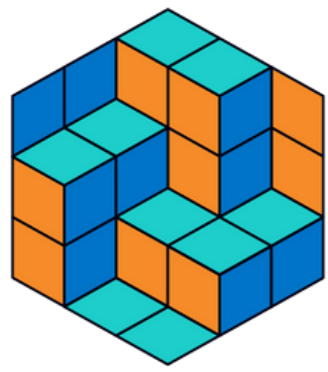


# ARMA(p,d)



$$Y_t = \sum_{i=1}^p Y_{t-i} \alpha_i + \sum_{j=1}^d b_j \epsilon_{t-j} + \mu_t + \epsilon_t$$

# ARMA(p,d)

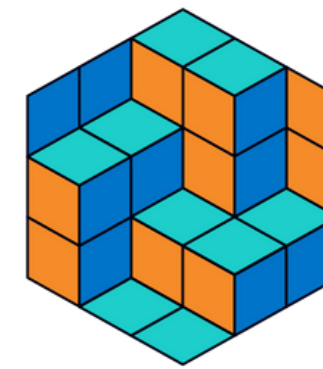


$$Y_t = \sum_{i=1}^p Y_{t-i} \alpha_i + \sum_{j=1}^d b_j \epsilon_{t-j} + \mu_t + \epsilon_t$$

Помните что это?

А это откуда?

# ARMA(p,d)

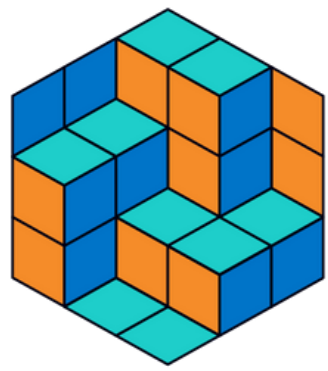


$$Y_t = \sum_{i=1}^p Y_{t-i} \alpha_i + \sum_{j=0}^d b_j \epsilon_{t-j}$$

Окей, параметры модель найдет  
А что делать с **p** и **d**?



# ARMA(p,d)



$$Y_t = \sum_{i=1}^p Y_{t-i} \alpha_i + \sum_{j=0}^d b_j \epsilon_{t-j}$$

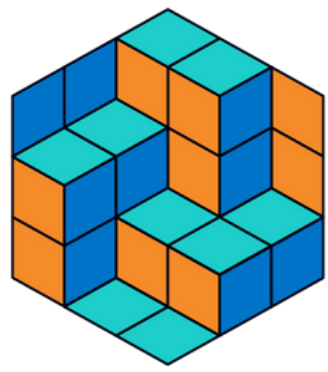
Окей, параметры модель найдет  
А что делать с **p** и **d**?

Лучший способ сейчас - угадать

Если интересен более корректные подход:

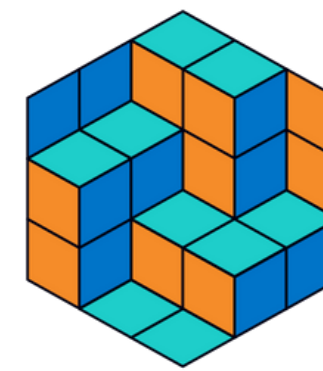
- <https://pythonpip.ru/examples/model-arima-v-python>

?



Но ARMA работает только для стационарных рядов, как быть в общем случае?

# ARIMA(p,q,d)

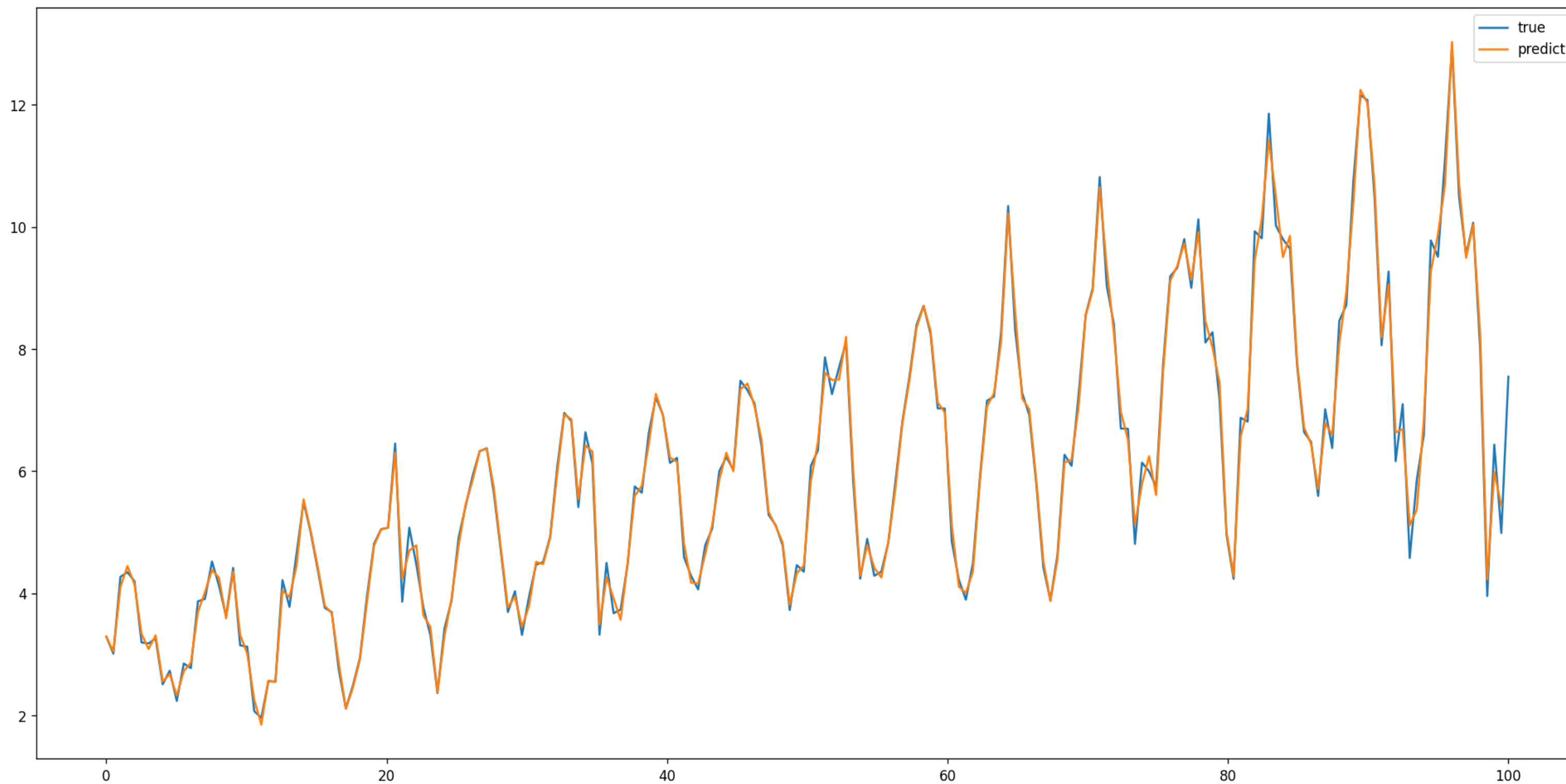
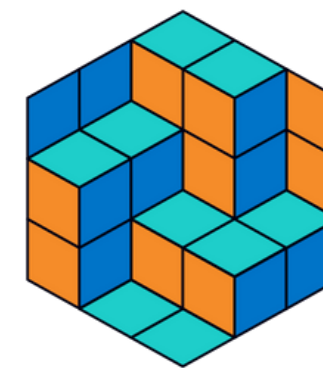


Надо просто сделать ряд стационарным и добавить что-то про тренд!

$$\Delta^q Y_t = c + \sum_{i=1}^p \Delta^q Y_{t-i} \alpha_i + \sum_{j=0}^d b_j \epsilon_{t-j}$$

$\Delta^q$  - операция взятия разности **q** раз подряд - сначала в самом ряду, потом в разностях, и тд.

# ARIMA(p,q,d)



# Вопросы?

