

# Основы pandas



**Константин Башевой**

Аналитик-разработчик, Яндекс



**Константин Башевой**  
Аналитик-разработчик  
Яндекс

Помогаю аналитикам с инфраструктурой  
Собираю инструменты обработки данных  
Рассказываю как это весело

Последние 10 лет:

Rambler&Co

Ростелеком

Яндекс

# Что сегодня будет

## Программа на сегодня

4

- ✓ Обзор возможностей библиотеки pandas
- ✓ Немного про презентации в jupyter notebook
- ✓ Базовые операции в pandas

# Еще одна библиотека ура

**Pandas = panel + data**

6



## Понимает кучу источников данных

7

- CSV- и Excel-файлы
- буфер обмена (Ctrl+c)
- HTML-страницы (тэг <table>)
- JSON-файлы
- формат parquet
- SQL-запросы

## Вычисления в оперативной памяти

8

DataFrame – хранится в оперативной памяти

Пример:

- есть ноутбук с 4Gb RAM
- 2Gb свободно
- с данными до 1.5Gb можно работать



## Удобно смотреть данные

9

```
data = pd.read_csv('power.csv')  
data.head()
```

	country	year	quantity	category
0	Austria	1996	5.0	1
1	Austria	1995	17.0	1
2	Belgium	2014	0.0	1
3	Belgium	2013	0.0	1
4	Belgium	2012	35.0	1

- GROUP BY – groupby с любыми функциями
- JOIN – join и merge
- ORDER BY – sort\_values

## Цепочки вычислений

11

```
(data.groupby('country').count()  
  .sort_values('quantity')  
  .query('category > 16')  
  .reset_index()[['country', 'quantity']]  
)
```

	country	quantity
0	Yemen Arab Rep. (former)	45
1	Yemen, Dem. (former)	61
2	Pacific Islands (former)	68
3	Antarctic Fisheries	90
4	German Dem. R. (former)	106

# Вычисления любой сложности

12

```
def baltic(country):  
    """Объединение стран Прибалтики"""  
  
    if country in ['Lithuania', 'Latvia', 'Estonia']:  
        return 'Прибалтика'  
  
    return 'Other'
```

```
data['baltic'] = data.country.apply(baltic)  
data.baltic.value_counts()
```

```
Other          1162176  
Прибалтика      27306  
Name: baltic, dtype: int64
```