

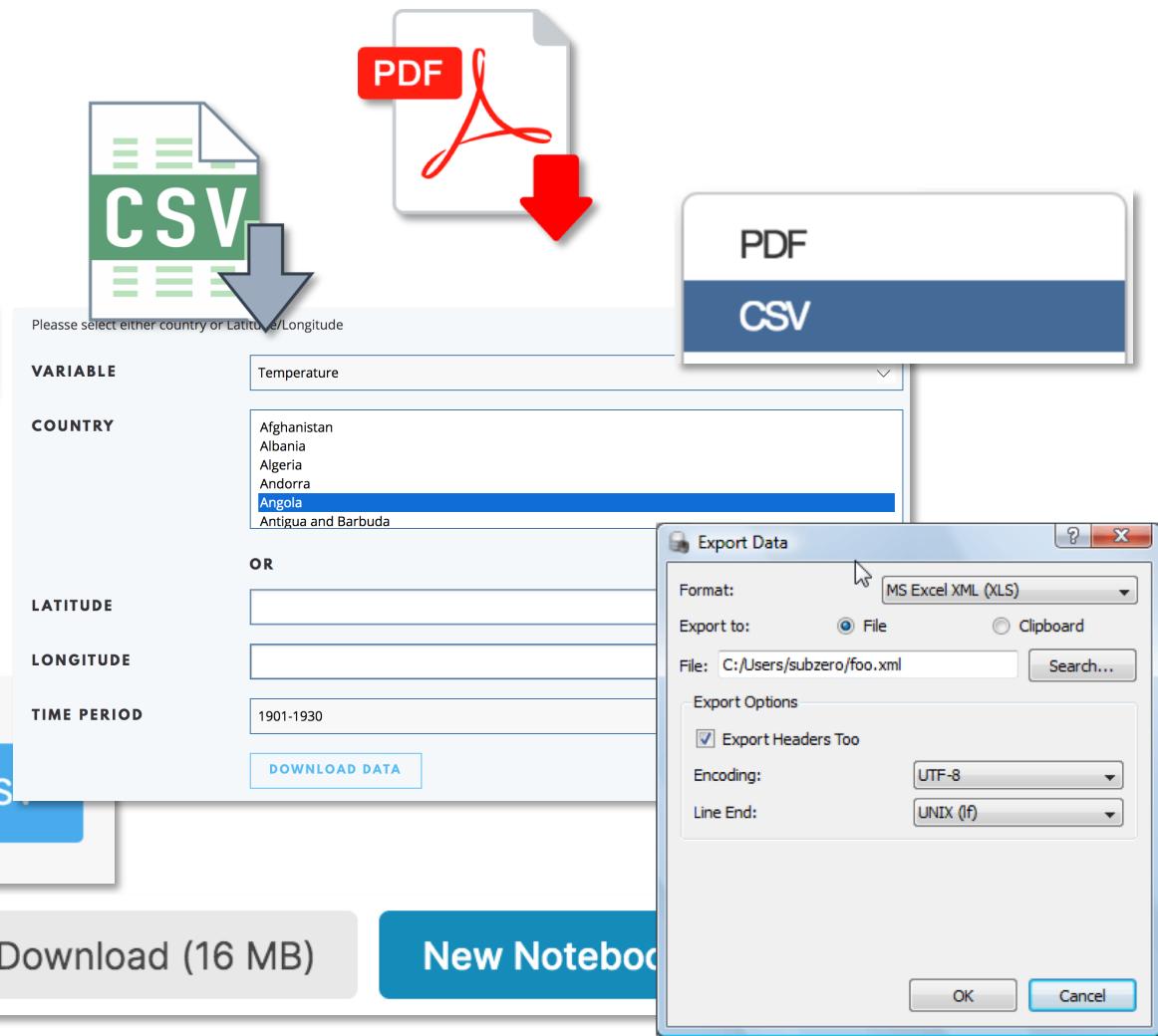
Зачем парсить?

Зачем парсить?

- У вас есть идея и есть данные, но их мало?
- У вас есть идея, но нет данных?
- У вас нет идеи, и нет данных?

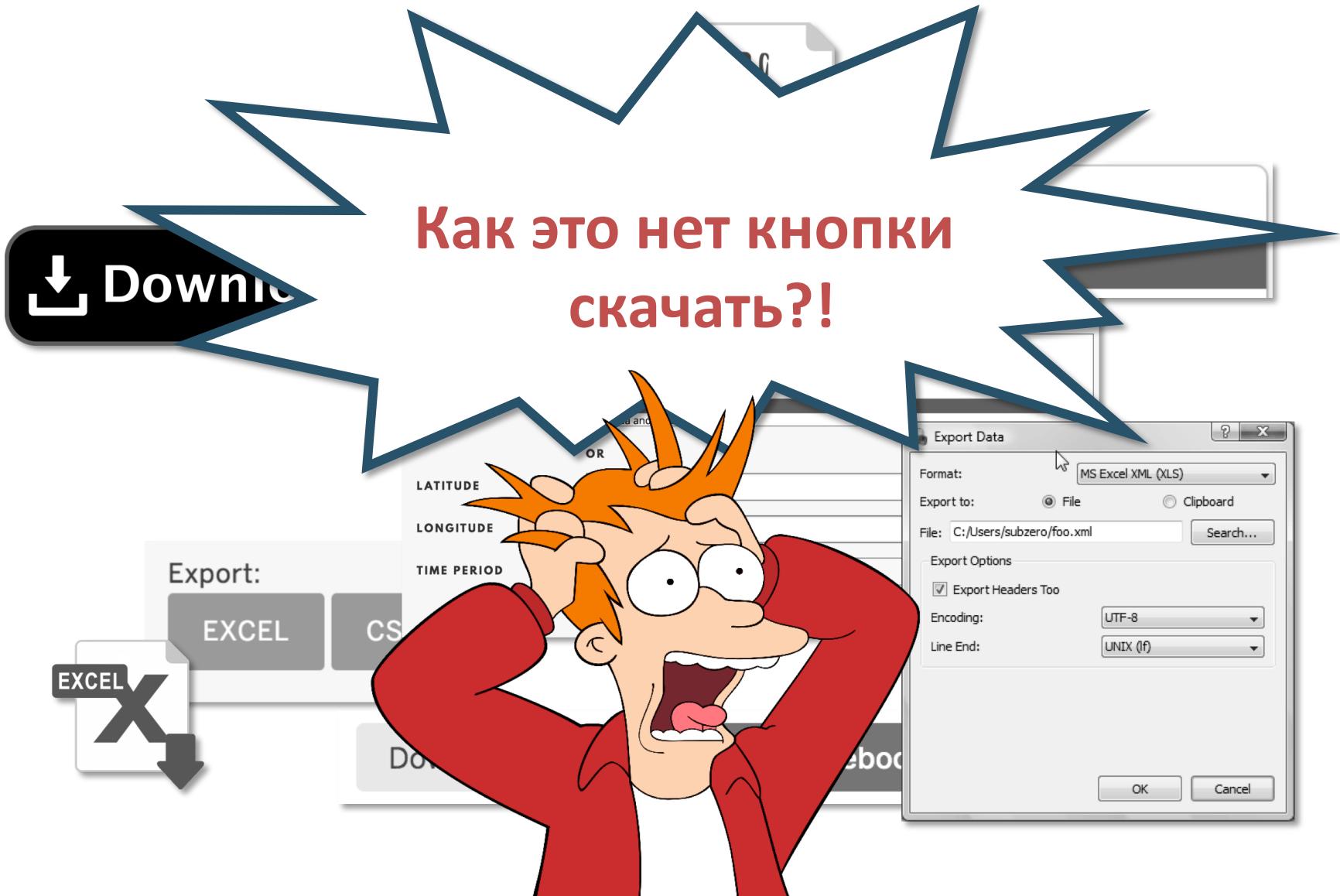
Зачем парсить?

 Download



The screenshot shows a user interface for data download. At the top right, there's a red button labeled "PDF" with a red arrow pointing down. Below it is a "CSV" button. A large blue arrow points from the "CSV" button to a dropdown menu titled "Please select either country or Latitude/Longitude". The dropdown contains several country names: Afghanistan, Albania, Algeria, Andorra, **Angola**, and Antigua and Barbuda. The word "Angola" is highlighted with a blue selection bar. To the left of the dropdown, there are input fields for "VARIABLE" (set to "Temperature"), "COUNTRY" (with the dropdown open), "LATITUDE", "LONGITUDE", and "TIME PERIOD" (set to "1901-1930"). Below these is a "DOWNLOAD DATA" button. On the far left, there's a "Export:" section with "EXCEL" and "CSV" buttons, and a green "EXCEL" icon with a green arrow pointing down. At the bottom, there are buttons for "Download (16 MB)" and "New Notebook". A modal dialog box titled "Export Data" is open on the right, showing settings for "Format: MS Excel XML (XLS)", "Export to: File" (selected), "File: C:/Users/subzero/foo.xml", "Export Options" (checkbox checked for "Export Headers Too"), "Encoding: UTF-8", and "Line End: UNIX (lf)". There are "OK" and "Cancel" buttons at the bottom of the dialog.

Зачем парсить?



Кадр из мультипликационного сериала «Футурама».
Автор Мэтт Грейнинг, Дэвид Коэн. Производство 20th Century Fox

Чем парсить?

- Конечно же python



Чем парсить?

- И ещё парой библиотек



Requests



BeautifulSoup

Как парсить?

```
►<span class="post-meta__item post-meta__published-at">...</span>
</div>
▼<h1 class="post-title post-header__title post-article__title article-feature-title">
  "
    Правда или ложь? "
    <span class="thin">Насколько правдива документальная фотография?</span> == $0
</h1>
▼<p class="post-excerpt post-header__excerpt post-article__excerpt">
  "
    Принято считать, что документальный снимок передает реальность достоверно, но т
  "
</p>
►<div class="post-authors post-header__authors post-article__authors">...</div>
</header>
```

Как парсить?

HyperText Markup Language (HTML)

Это язык разметки для создания веб страниц и приложений

- **<a>** используют для создания ссылок на другие странички
- **<h1> ... <h6>** используют для создания заголовков разных уровней
- **** задаёт полужирное очертание текста
- **<div>** выделяет любой отдельный блок веб-страницы
- И многие другие тэги ...

Как парсить?

- Открываем сайт с интересующими нас данными
- Находим интересующий нас элемент и переходим в браузере в режим просмотра кода (inspect element code)
- Находим соответствующие тэги, окружающие наш элемент
- Дальше библиотека Beautiful Soup всё сделает сама (она превращает html код страницы в дерево, и мы можем по нему перемещаться)

Что может пойти не так?

- Какие наиболее распространённые проблемы могут возникнуть при парсинге?
- Страница может быть не найдена
- А может быть найдена, но доступа к ней у вас нет
- А может быть и есть, но через некоторое время сервер его отбирает

Серверные ошибки

- Как понять, что что-то пошло не так?
- При помощи кода серверной ошибки!
- Он будет возвращен в случае неуспешного запроса

Серверные ошибки



403 – доступ
запрещен

404



404 – страница
не найдена



401 – неавторизованный
доступ



504 – сервер
не дождался ответа

Почему возникают ошибки?

- Сервер не любит, когда его бомбардируют запросами
- Серверу может не понравиться, что вы не человек
- Серверу надоедает, когда по нему бродят парсеры

Как это исправить?

- Ограничить число запросов в секунду/минуту/...
- Притвориться человеком
- Иногда даже скрываться и менять своё местоположение
(достаточно просто IP адреса)

Как это исправить?

- Ограничить число запросов в секунду/минуту/...
- Притвориться человеком
- Иногда даже скрываться и менять своё местоположение (достаточно просто IP адреса)



Если в результате получаем серверный код 200 – всё получилось!

