

# Получение и визуализация данных

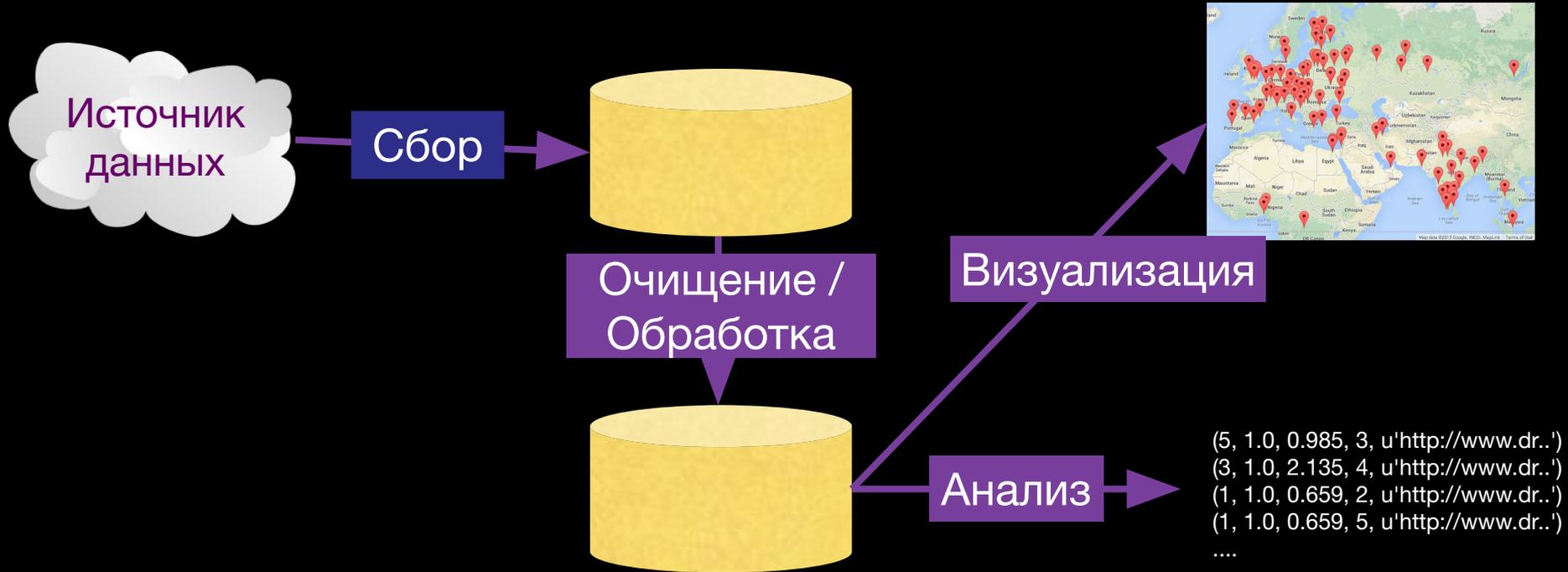
Чарльз Северанс



Пайтон для всех  
[www.py4e.com](http://www.py4e.com)



# Многоступенчатый анализ данных



# Технологии интеллектуального анализа данных

- <https://hadoop.apache.org/>
- <http://spark.apache.org/>
- <https://aws.amazon.com/redshift/>
- <http://community.pentaho.com/>
- ....

# «Интеллектуальный анализ персональных данных»

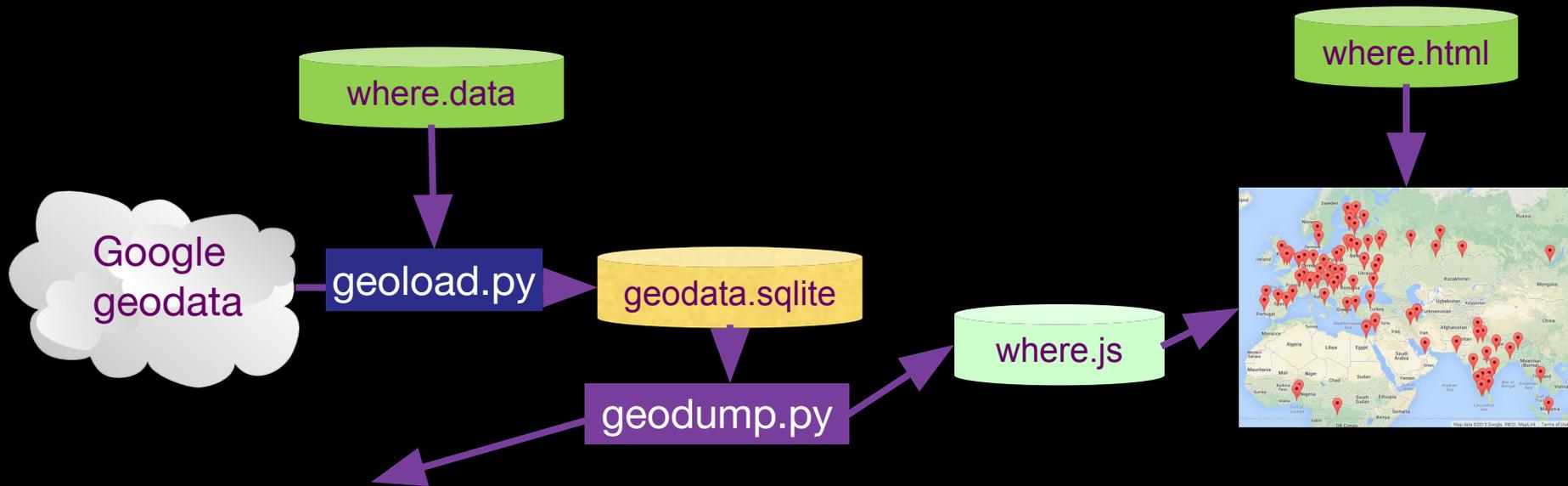
Наша цель — помочь вам стать лучше в программировании, а не сделать из вас экспертов по интеллектуальному анализу данных

# Геодата (Geodata)

- Создает Google-карту на основе введенных пользователем данных
- Использует Google Geodata API
- Кэширует данные в базе данных, чтобы избежать ограничения скорости обработки запросов и позволяет перезагрузку базы данных
- Отображается в браузере, используя Google Maps API



<http://www.py4e.com/code3/geodata.zip>



Северо-Восточный Университет, ... Бостон, Массачусетс 02115, США  
42.3396998 -71.08975

Университет Брэдли, 1501 ... Пеория, Иллинойс 61625, США 40.6963857  
-89.6160811

...

Technion, Viazman 87, Kesalsaba, 32000, Израиль 32.7775 35.0216667

Университет Монаша Клейтон... Виктория 3800, Австралия -37.9152113  
145.134682

Кокшетау, Казахстан 53.2833333 69.3833333

...

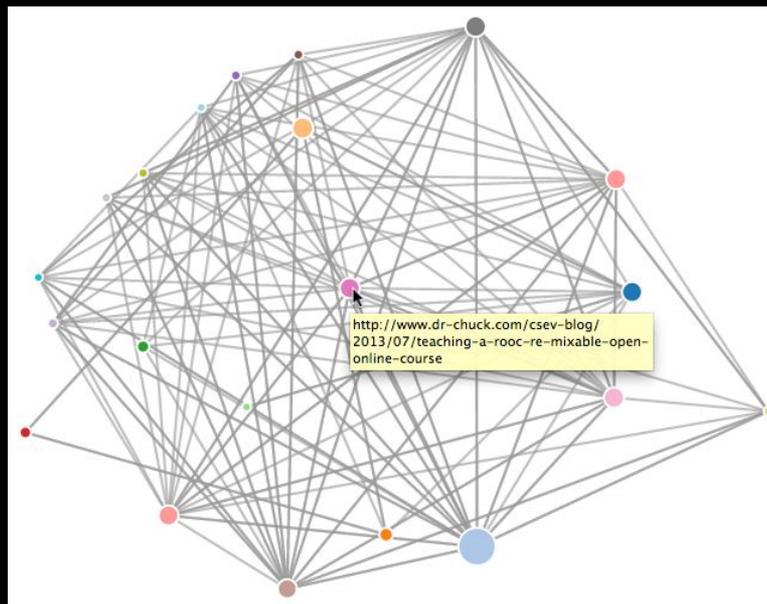
12 записей в файле where.js

Откройте файл where.html, чтобы посмотреть данные в окне браузера

<http://www.py4e.com/code3/geodata.zip>

# Пэйдж-ранк

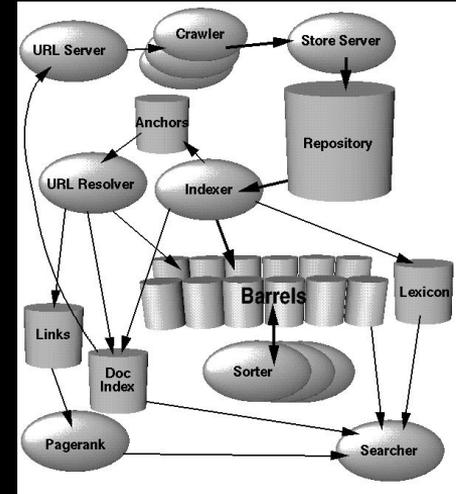
- Пишет простой поисковый робот для веб-страниц
- Вычисляет простую версию алгоритма ранжирования Google
- Отображает получившуюся сеть



<http://www.py4e.com/code3/pagerank.zip>

# Архитектура поисковой системы

- Поисковый робот
- Индексирование
- Поиск



<http://infolab.stanford.edu/~backrub/google.html>

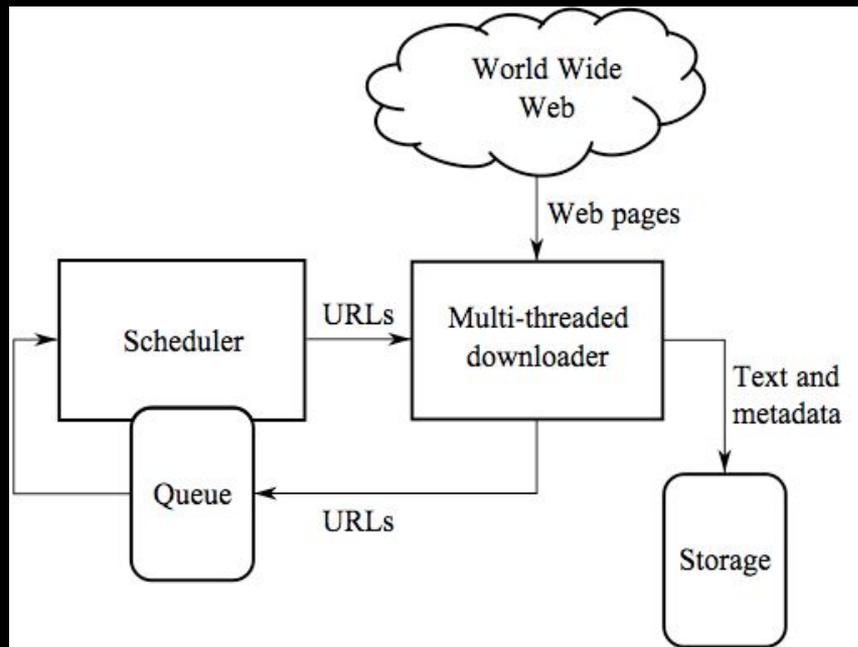
# Поисковый робот

Поисковый робот («веб-паук») — автоматизированная компьютерная программа, которая систематически просматривает Интернет. Поисковые роботы обычно используются для создания копий всех посещенных страниц, которые затем будут обработаны поисковой системой. Она проиндексирует загруженные страницы, чтобы обеспечить быстрый поиск результатов.

[https://ru.wikipedia.org/wiki/Поисковый\\_робот](https://ru.wikipedia.org/wiki/Поисковый_робот)

# Поисковый робот

- Извлекает информацию со страницы
- Просматривает страницу на предмет ссылок на другие страницы
- Добавляет ссылки в список, чтобы затем извлечь информацию с этих страниц
- Повторяет процесс...



[https://ru.wikipedia.org/wiki/Поисковый\\_робот](https://ru.wikipedia.org/wiki/Поисковый_робот)

# Политика сканирования

- **политика выбора** указывает страницы для загрузки
- **политика повторного посещения** указывает, когда проверять наличие изменений на страницах
- **политика вежливости** указывает, как избежать перегрузки веб-сайта
- **политика параллелизации** определяет, как координировать распределенные поисковые роботы

# Протокол robots.txt

- Способ взаимодействия сайта с поисковыми роботами
- Неформальный добровольный стандарт
- Иногда администраторы сайта делают «Ловушку для пауков», чтобы отловить «плохих» пауков

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /images/  
Disallow: /tmp/  
Disallow: /private/
```

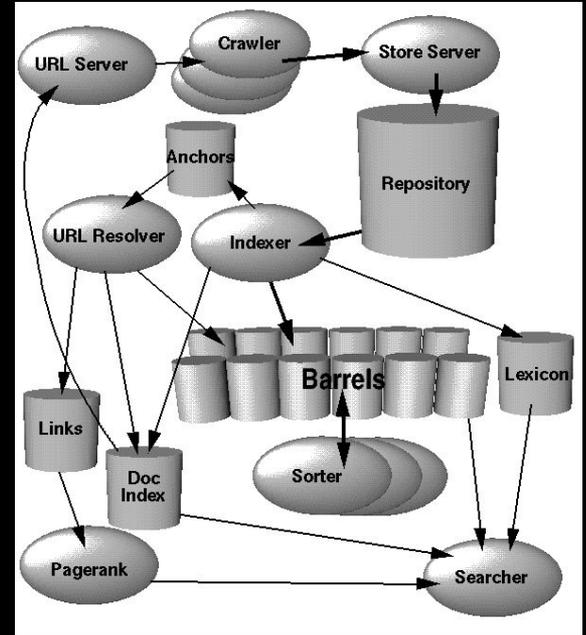
\*Запретить

[https://ru.wikipedia.org/wiki/Стандарт\\_исключений\\_для\\_роботов](https://ru.wikipedia.org/wiki/Стандарт_исключений_для_роботов)

[http://en.wikipedia.org/wiki/Spider\\_trap](http://en.wikipedia.org/wiki/Spider_trap)

# Архитектура Google

- Веб-сканирование
- Индексация
- Поиск



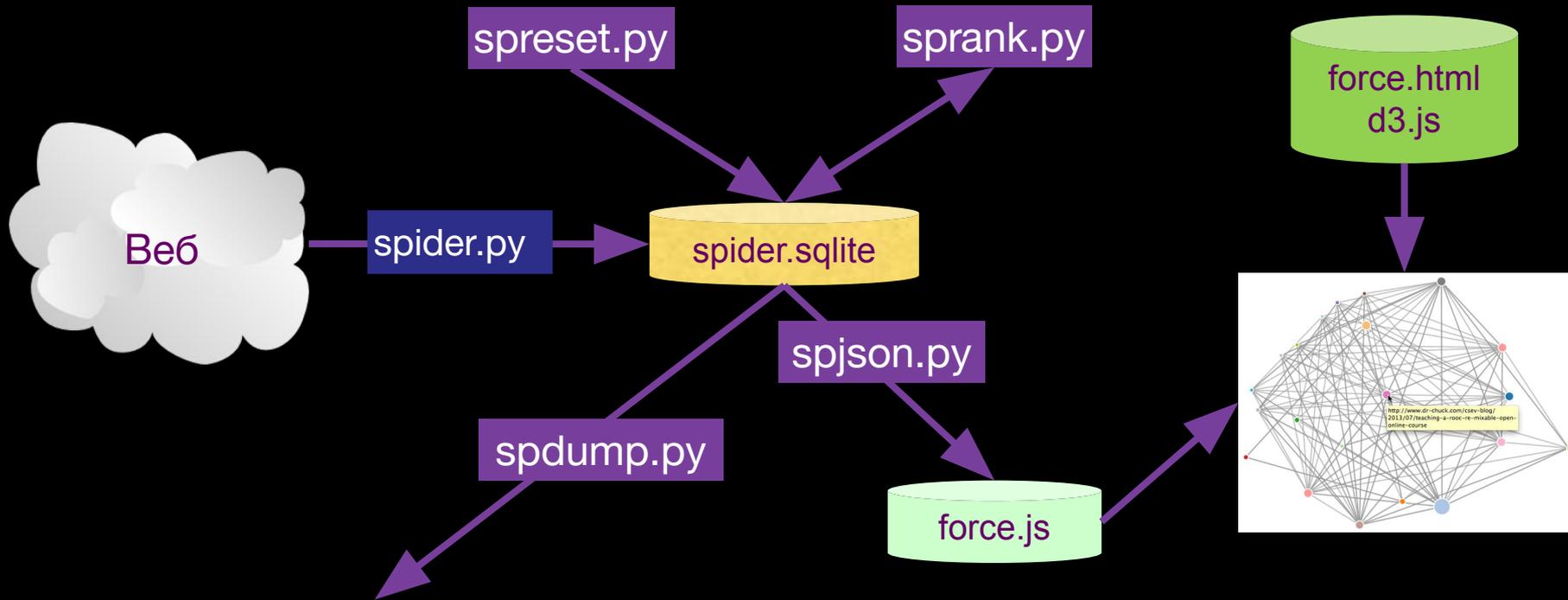
<http://infolab.stanford.edu/~backrub/google.html>

# Поисковый индекс

Поисковая машина индексирует, обрабатывает и хранит данные для обеспечения быстрого и точного поиска информации.

Целью хранения индекса является повышение скорости и производительности поиска релевантных документов по поисковому запросу. Без индекса поисковая машина была бы вынуждена сканировать каждый документ в корпусе, что потребовало бы большого количества времени и вычислительной мощности.

[https://ru.wikipedia.org/wiki/Поисковый\\_индекс](https://ru.wikipedia.org/wiki/Поисковый_индекс)



(5, None, 1.0, 3, u'http://www.dr-chuck.com/csev-blog')  
 (3, None, 1.0, 4, u'http://www.dr-chuck.com/dr-chuck/resume/speaking.htm')  
 (1, None, 1.0, 2, u'http://www.dr-chuck.com/csev-blog/')  
 (1, None, 1.0, 5, u'http://www.dr-chuck.com/dr-chuck/resume/index.htm')  
 4 строки.

<http://www.py4e.com/code3/pagerank.zip>

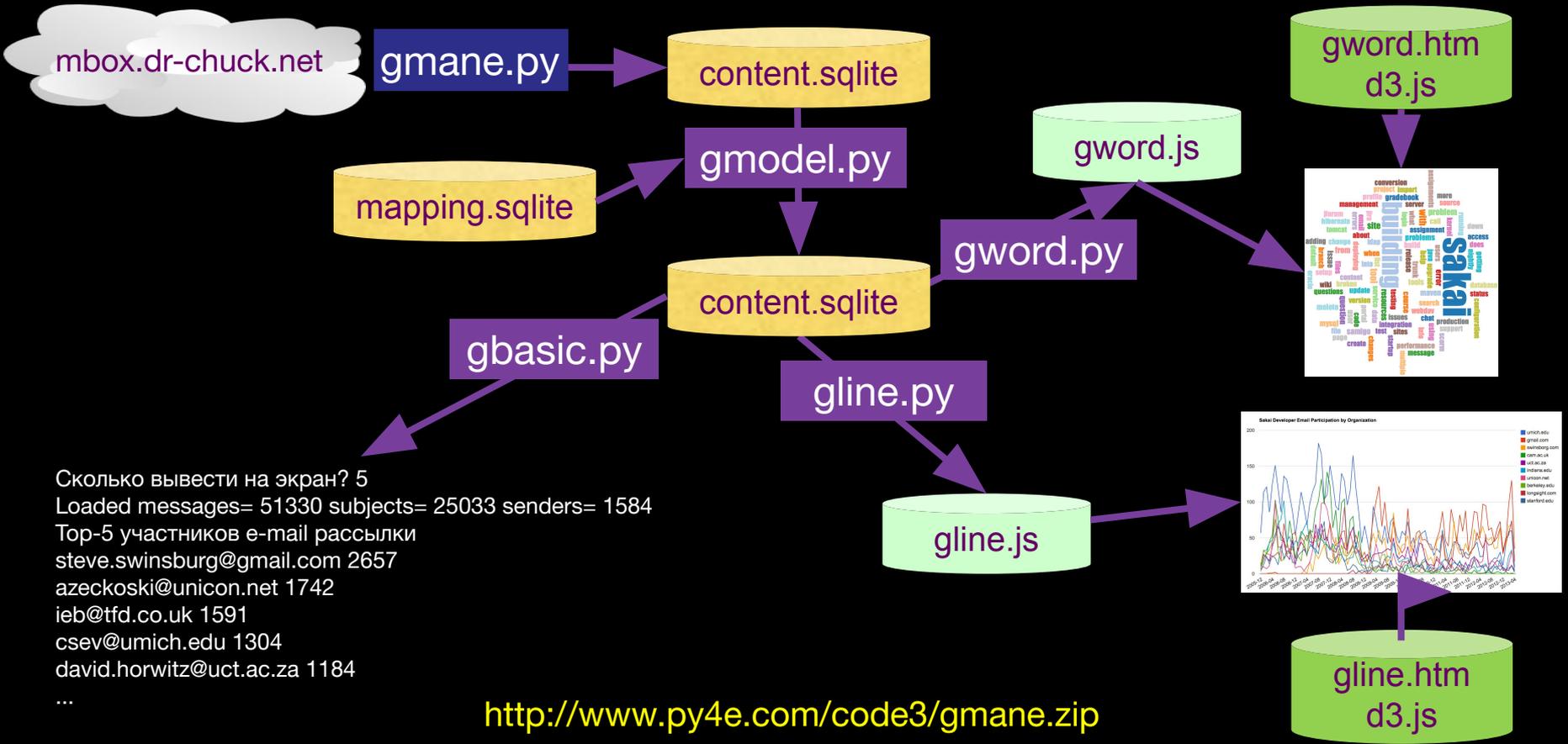


# Предупреждение: если набор данных превышает 1Гб,

- не настраивайте использование [gmane.org](https://gmane.org) из своего приложения
- Нет ограничения частоты запросов – это круто!

Для тестирования используйте:

<http://mbox.dr-chuck.net/sakai.devel/4/5>



Сколько вывести на экран? 5  
 Loaded messages= 51330 subjects= 25033 senders= 1584  
 Top-5 участников e-mail рассылки  
 steve.swinsburg@gmail.com 2657  
 azeckoski@unicon.net 1742  
 ieb@tfd.co.uk 1591  
 csev@umich.edu 1304  
 david.horwitz@uct.ac.za 1184  
 ...

<http://www.py4e.com/code3/gmane.zip>



## Авторы / Благодарности



... Insert new Contributors and Translations here

Авторские права на эти слайды принадлежат Чарльзу Р. Северансу ([www.dr-chuck.com](http://www.dr-chuck.com)), 2010 г., Школе Информации Мичиганского Университета и [open.umich.edu](http://open.umich.edu), и доступны по лицензии Creative Commons Attribution 4.0 License.

Пожалуйста, сохраняйте этот слайд во всех копиях этого документа, в соответствии с требованиями Лицензии. Если вы внесли изменения, добавьте свое имя или организацию в список участников на этой странице.

Исходная разработка: Чарльз Северанс, Школа Информации Мичиганского Университета.

Перевод выполнила Фомкина Виолетта.

... Добавьте сюда новых авторов и переводчиков