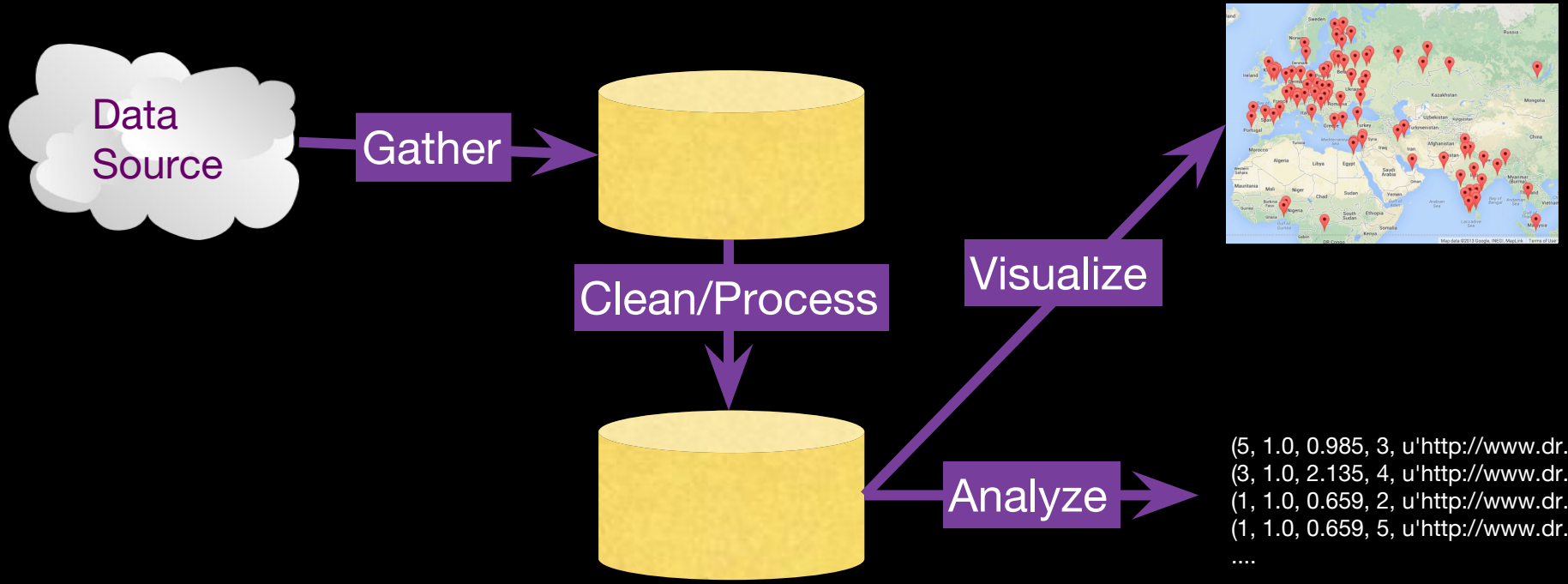
Retrieving and Visualizing Data

Charles Severance

Multi-Step Data Analysis



Many Data Mining Technologies

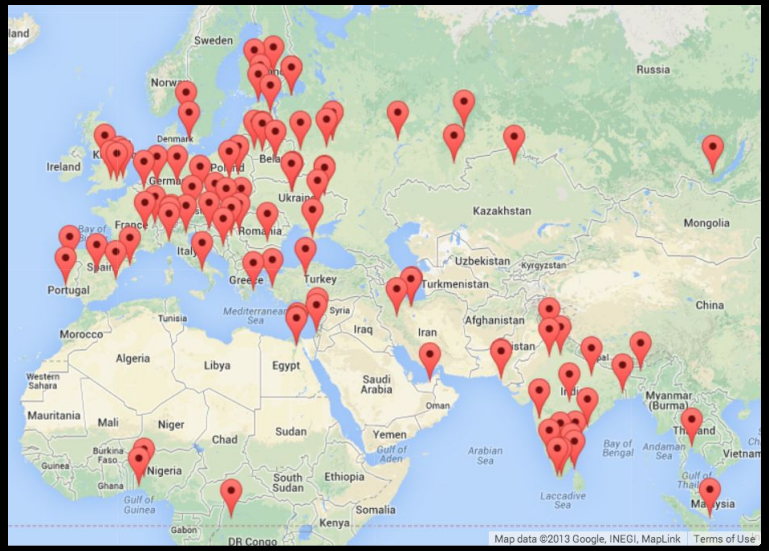
- https://hadoop.apache.org/
- http://spark.apache.org/
- https://aws.amazon.com/redshift/
- http://community.pentaho.com/
- ....

"Personal Data Mining"

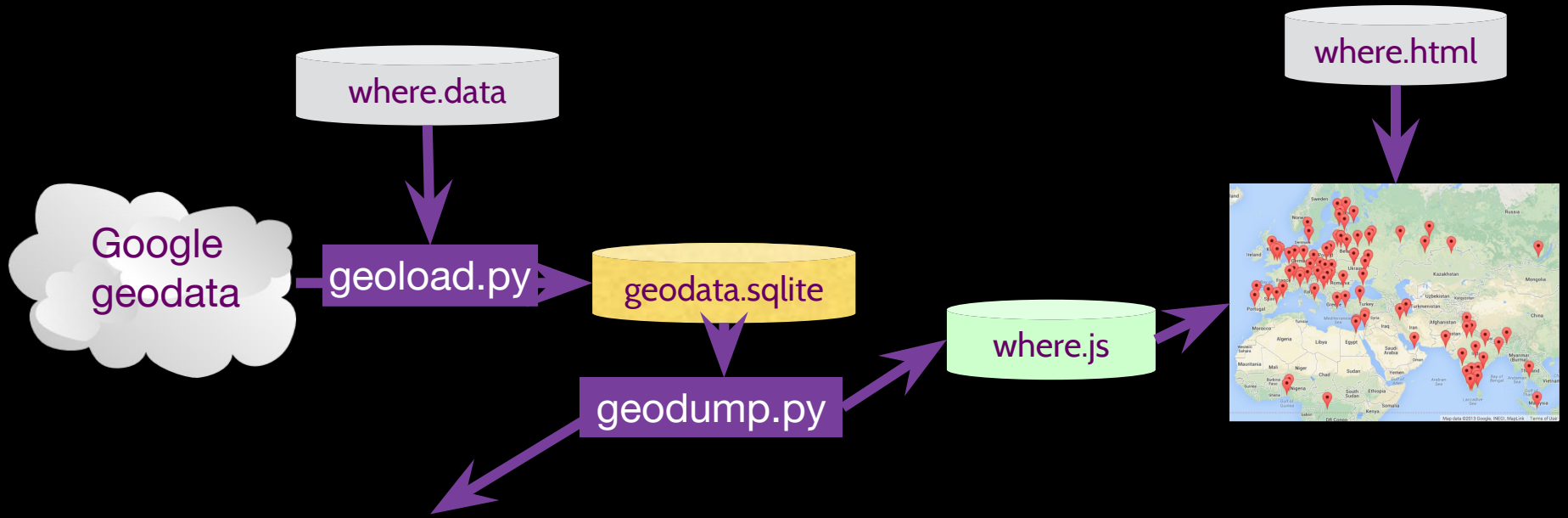
- Our goal is to make you better programmers – not to make you data mining experts

GeoData

- Makes a Google Map from user entered data
- Uses the Google Geodata API
- Caches data in a database to avoid rate limiting and allow restarting
- Visualized in a browser using the Google Maps API



http://www.pythonlearn.com/code/geodata.zip

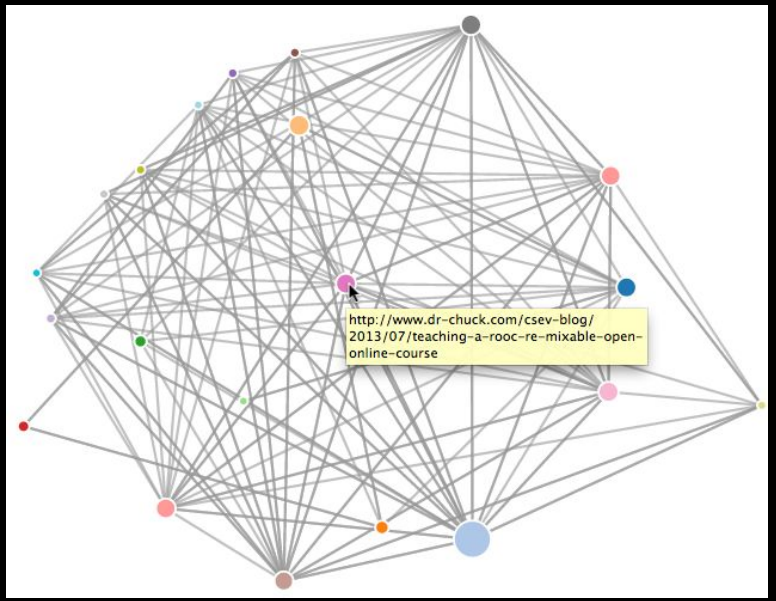



```
Northeastern University, ... Boston, MA 02115, USA 42.3396998 -71.08975
Bradley University, 1501 ... Peoria, IL 61625, USA 40.6963857 -89.6160811
...
Technion, Viazman 87, Kesalsaba, 32000, Israel 32.7775 35.0216667
Monash University Clayton ... VIC 3800, Australia -37.9152113 145.134682
Kokshetau, Kazakhstan 53.2833333 69.3833333
...
12 records written to where.js
Open where.html to view the data in a browser
```

http://www.pythonlearn.com/code/geodata.zip

Page Rank

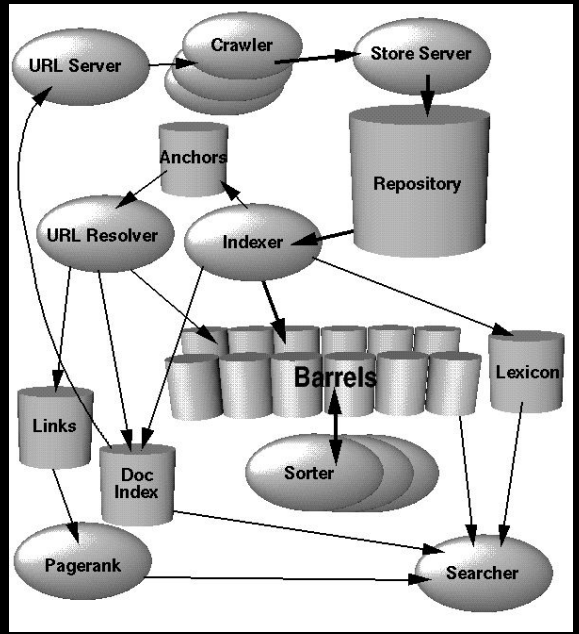
- Write a simple web page crawler
- Compute a simple version of Google's Page Rank algorithm
- Visualize the resulting network



http://www.pythonlearn.com/code/pagerank.zip

Search Engine Architecture

- Web Crawling
- Index Building
- Searching



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

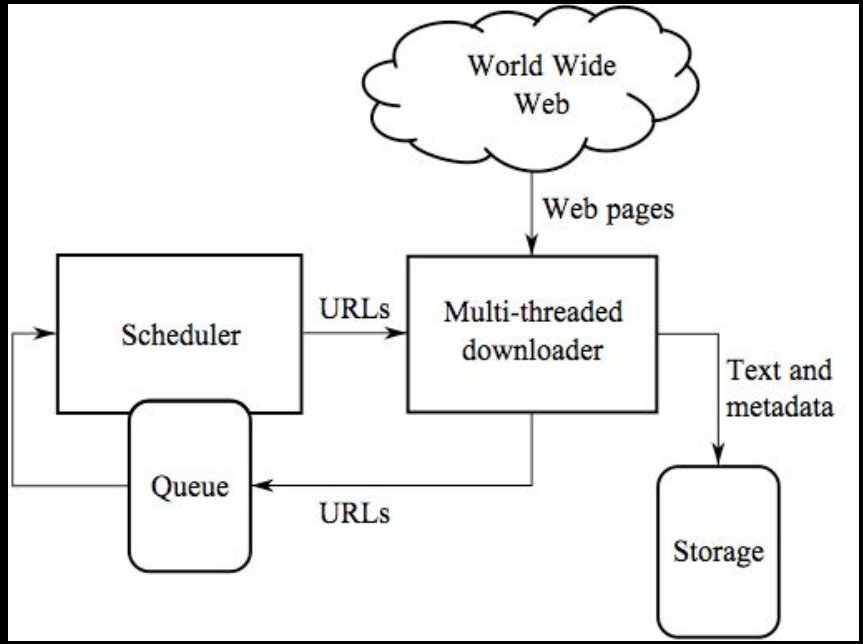
Web Crawler

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

http://en.wikipedia.org/wiki/Web_crawler

Web Crawler

- Retrieve a page
- Look through the page for links
- Add the links to a list of “to be retrieved” sites
- Repeat...



http://en.wikipedia.org/wiki/Web_crawler

Web Crawling Policy

a selection policy that states which pages to download,
a re-visit policy that states when to check for changes to the pages,
a politeness policy that states how to avoid overloading Web sites, and
a parallelization policy that states how to coordinate distributed Web crawlers

http://en.wikipedia.org/wiki/Web_crawler

robots.txt

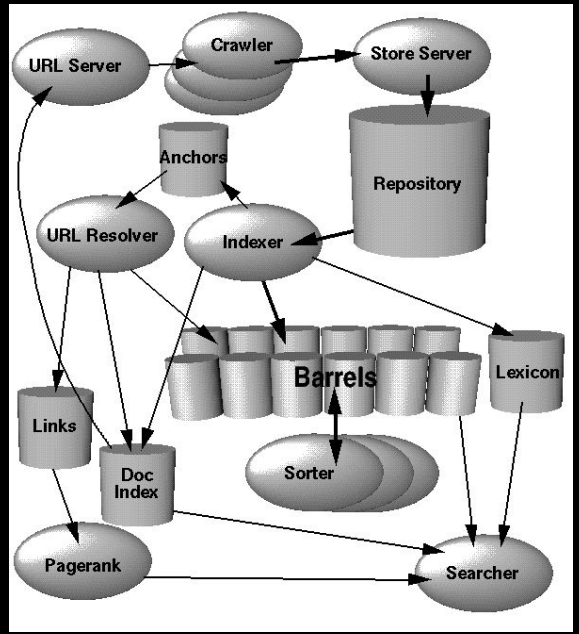
- A way for a web site to communicate with web crawlers
- An informal and voluntary standard
- Sometimes folks make a “Spider Trap” to catch “bad” spiders

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

http://en.wikipedia.org/wiki/Robots_Exclusion_Standard
http://en.wikipedia.org/wiki/Spider_trap

Google Architecture

- Web Crawling
- Index Building
- Searching

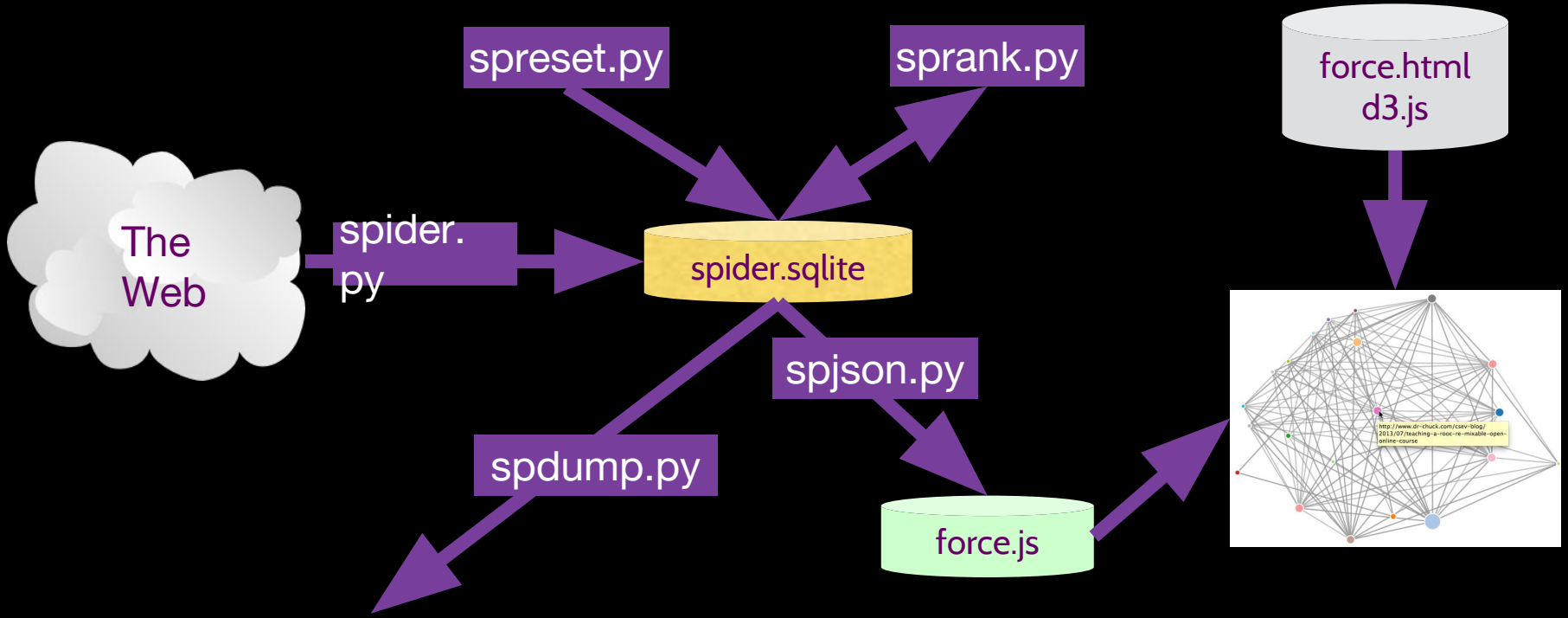


http://infolab.stanford.edu/~backrub/google.html

Search Indexing

Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power.

http://en.wikipedia.org/wiki/Index_(search_engine)



(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 rows.

http://www.pythonlearn.com/code/pagerank.zip

Mailing Lists - Gmane

- Crawl the archive of a mailing list
- Do some analysis / cleanup
- Visualize the data as word cloud and lines

http://www.pythonlearn.com/code/gmane.zip

Warning: This Dataset is > 1GB

- Do not just point this application at gmane.org and let it run all night
- There is no rate limits – these are cool folks
- Don't ruin it for the rest of us
- Please use my non-rate-limited copy of this data for your testing

`http://gmane.dr-chuck.net//gmane.comp.cms.sakai.devel/4/5`

Exporting

Gmane is primarily an archival site, and as such, it's important that it's easy to extract the contents again. If a list admin wants to use a different archival method, or just wants a copy for herself, then there has to be a procedure for extracting the contents of a group.

Now, Gmane is a news server, so it's easy enough to just point a news grabber at a group and just say "go". The main problem with this is that you're not getting a pristine copy of what was appeared on the list. Gmane does protocol conversions between mail and news, and you probably don't want those in a non-news-based archive.

So, a different method is provided, too. And it's scriptable.

Gmane has a "download" interface that returns a Un*x mbox file. To download messages 851 up to (but not including) 855 in the group gmane.discuss, just say

```
http://download.gmane.org/gmane.discuss/851/855
```

You could, for instance, set up a cron job to fetch the new messages in that group on a monthly basis, but you'd have to keep track of which article numbers you want to get.

This download interface tries to reverse all mail-to-news protocol transformations. It does not decrypt encrypted email addresses at present. Perhaps there could be some special interface for list admins where they could supply a password or the like and get a decrypted archive, if the list is encrypted. But that's not implemented yet.

This interface is a slight CPU and bandwidth hog, so if it's abused, it will be shut down. (List admins will then have to get a user name/password thing going.)

http://gmane.org/export.php

gmane.py

content.sqlite

gmodel.py

content.sqlite

gbasic.py

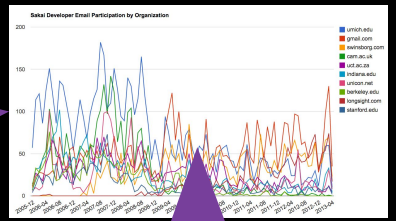
gword.py

gword.js

gword.htm
d3.js

gline.py

gline.js



gline.htm
d3.js

How many to dump? 5
Loaded messages= 51330 subjects= 25033 senders= 1584
Top 5 Email list participants
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.pythonlearn.com/code/gmane.zip

Acknowledgements / Contributions

Initial Development: Charles Severance, University of Michigan School of Information

… Insert new Contributors here

...