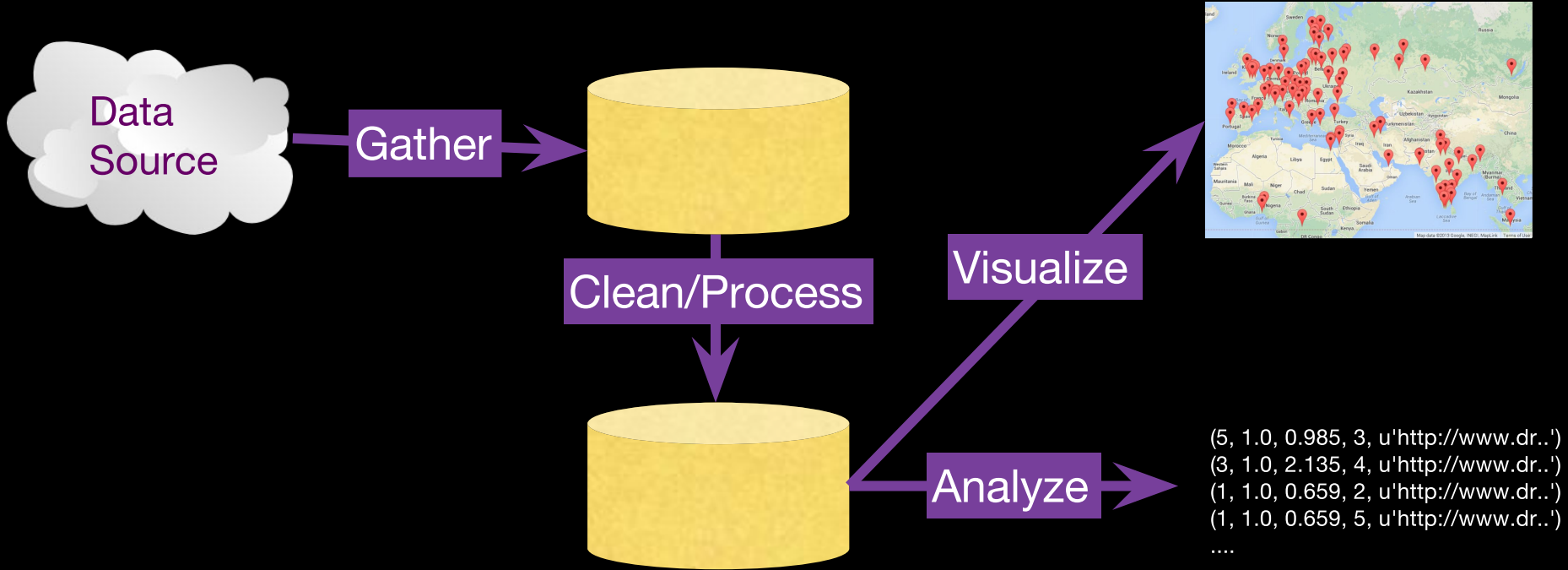


# Retrieving and Visualizing Data

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

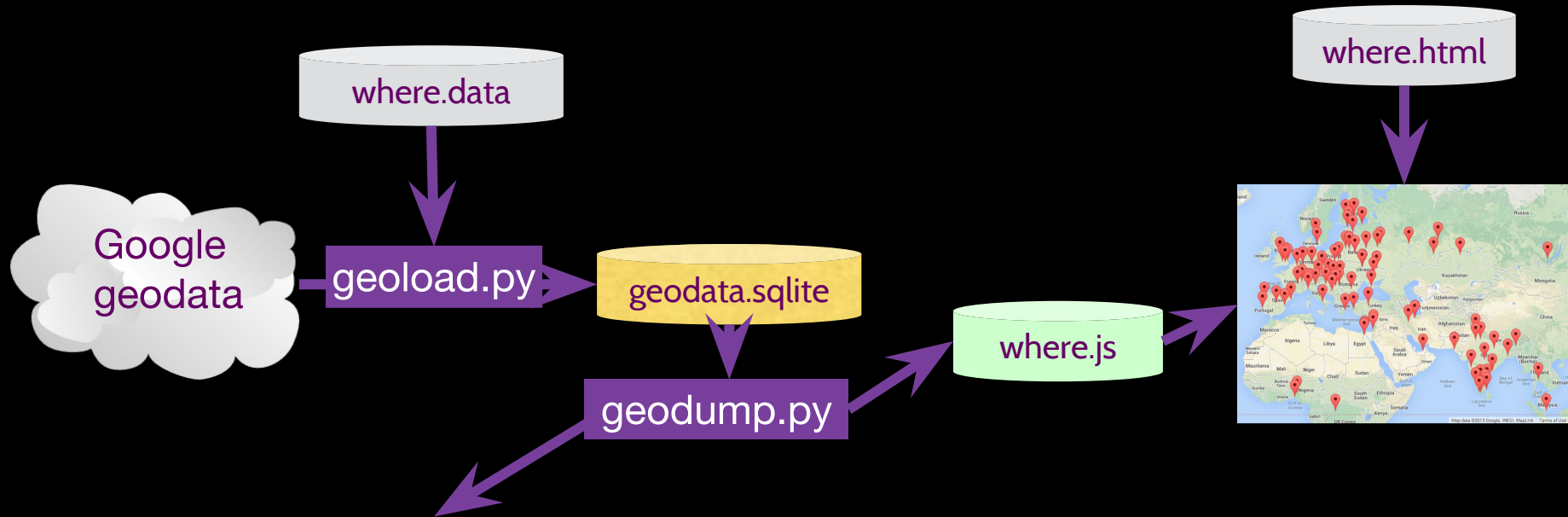
# GeoData

- 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>

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

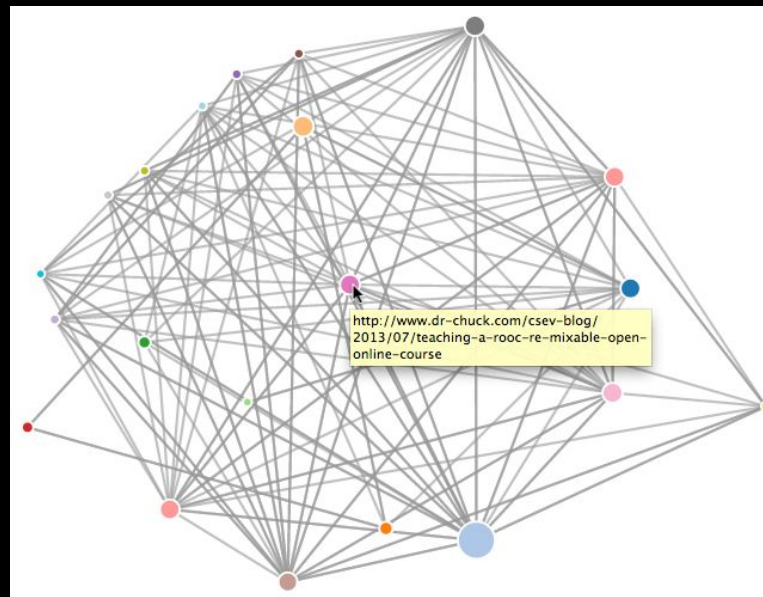
...

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

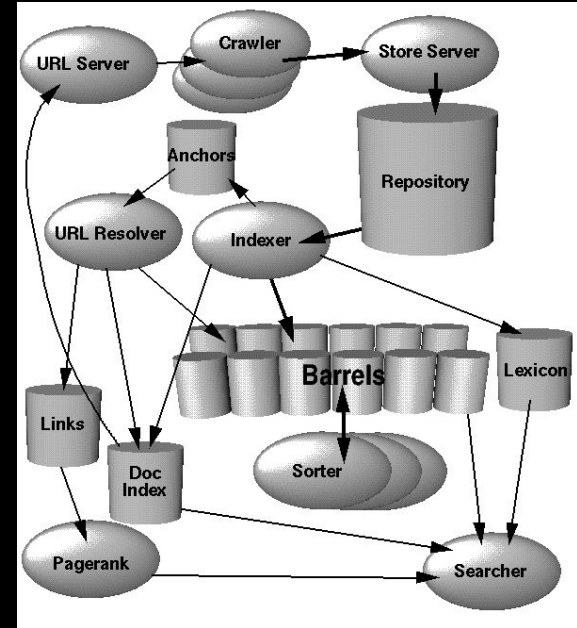


<http://www.pythonlearn.com/code/pagerank.zip>

zip

# Search Engine Architecture

- Web Crawling
- Index Building
- Searching



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



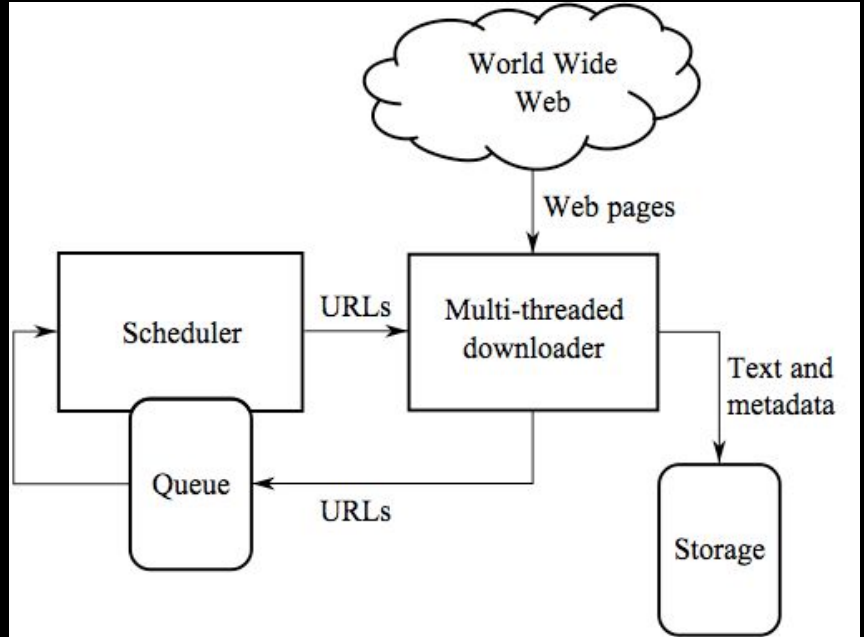
# Web Crawler

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](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](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](http://en.wikipedia.org/wiki/Web_crawler)

# robots.txt

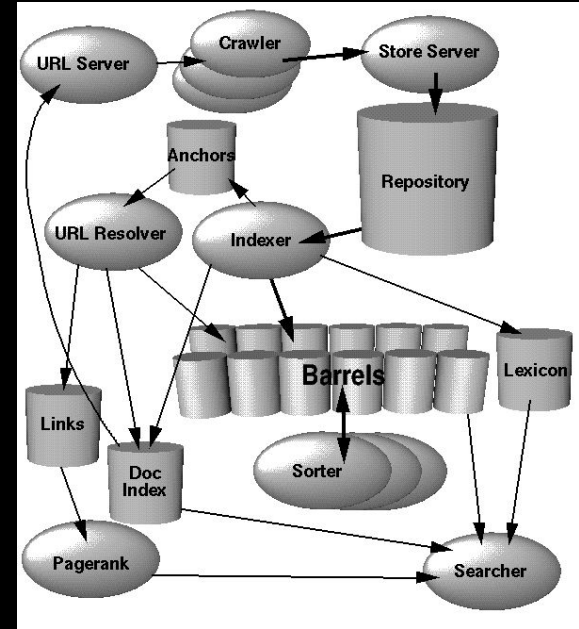
- A way for a web site to communicate with web crawlers
- An informal and voluntary standard
- Sometimes folks make a “Spider Trap” to catch “bad” spiders

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

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

# Google Architecture

- Web Crawling
- **Index Building**
- Searching

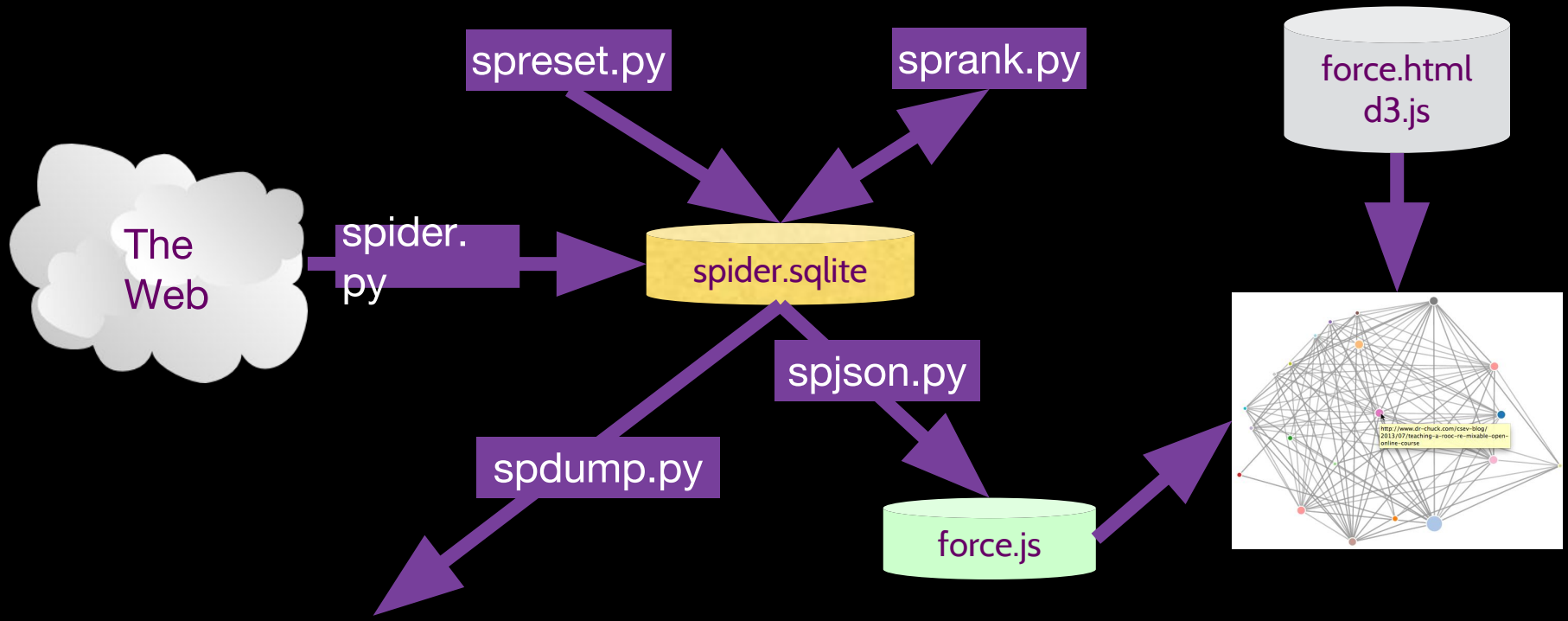


<http://infolab.stanford.edu/~backrub/google.html>

# 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\\_\(search\\_engine\)](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>

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.](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://mbox.dr-chuck.net/sakai.devel/4/5>



# Acknowledgements / Contributions



These slides are Copyright 2010- Charles R. Severance ([www.dr-chuck.com](http://www.dr-chuck.com)) of the University of Michigan School of Information and [open.umich.edu](http://open.umich.edu) and made available under a Creative Commons Attribution 4.0 License. Please maintain this last slide in all copies of the document to comply with the attribution requirements of the license. If you make a change, feel free to add your name and organization to the list of contributors on this page as you republish the materials.

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

... Insert new Contributors here