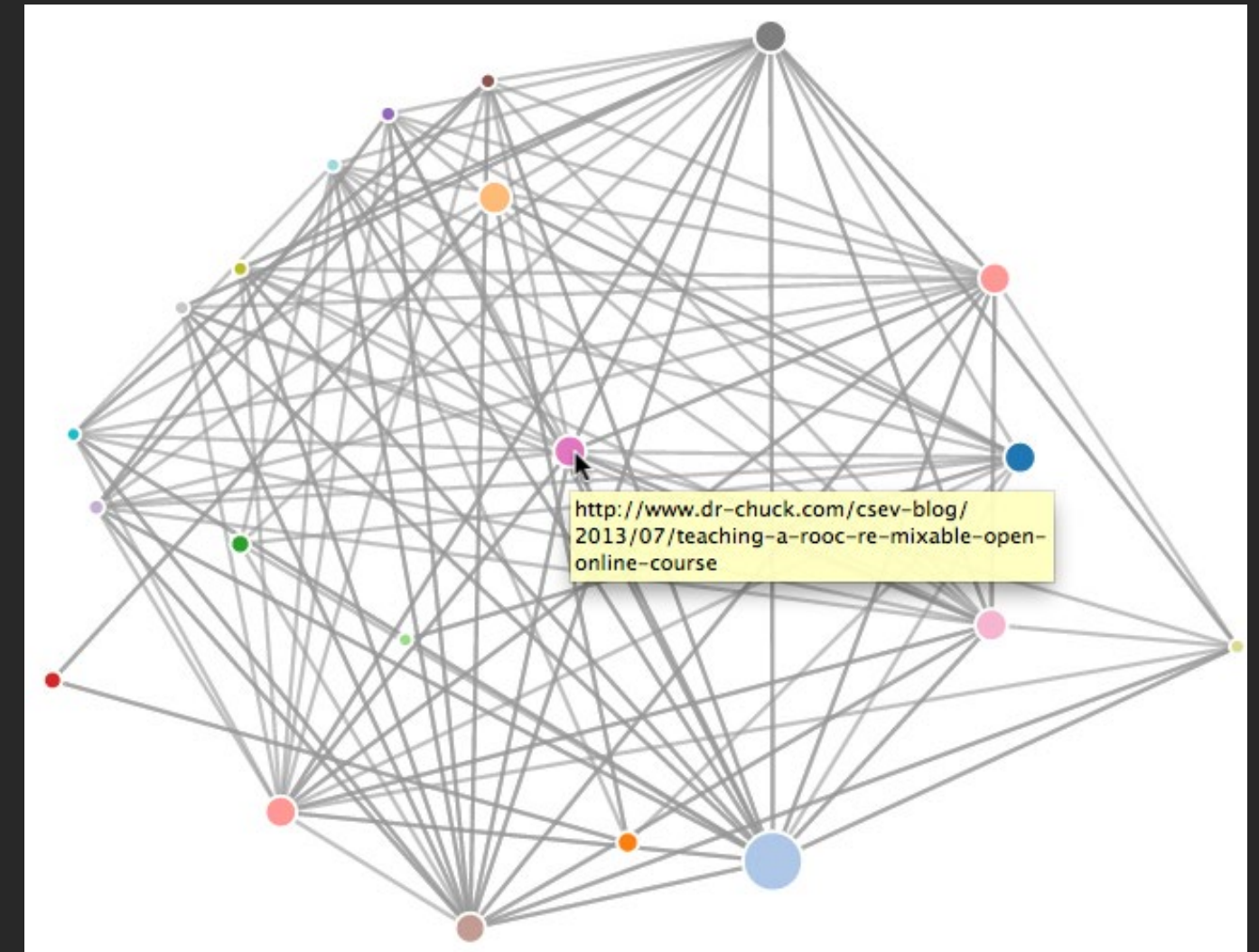# Page Rank

- Write a simple web page crawler

- Compute a simple version of Google's Page Rank algorithm

- Visualize the resulting network
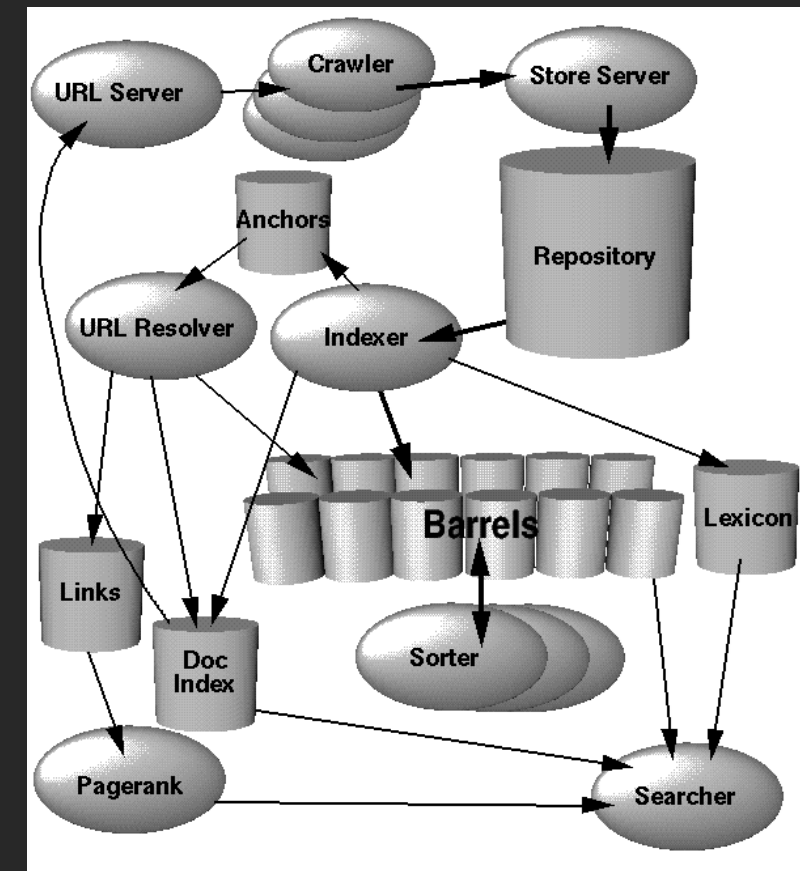


http://www.py4e.com/code3/pagerank.zip

# 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...
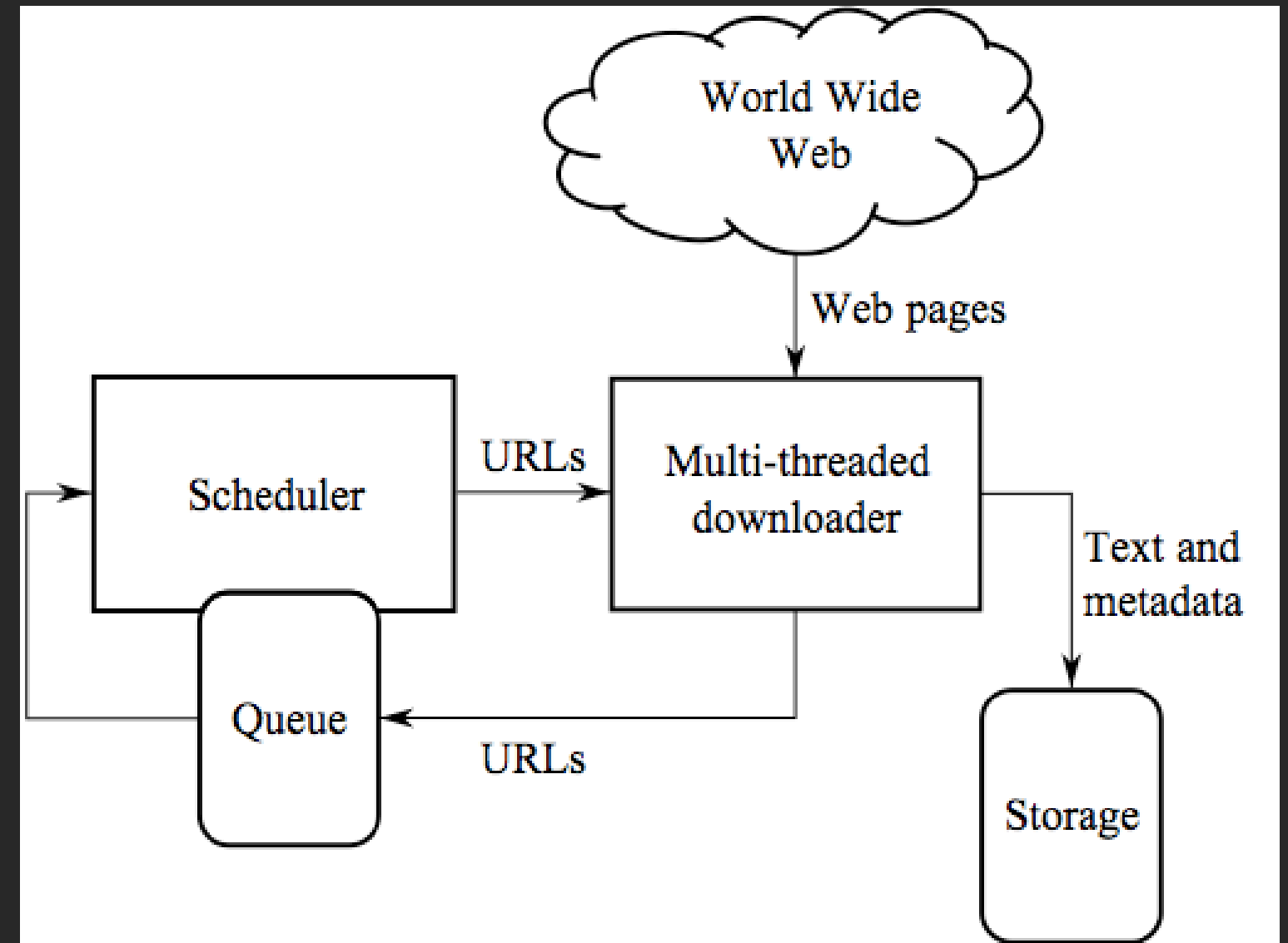


http://en.wikipedia.org/wiki/Web_crawler

# 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

# 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/
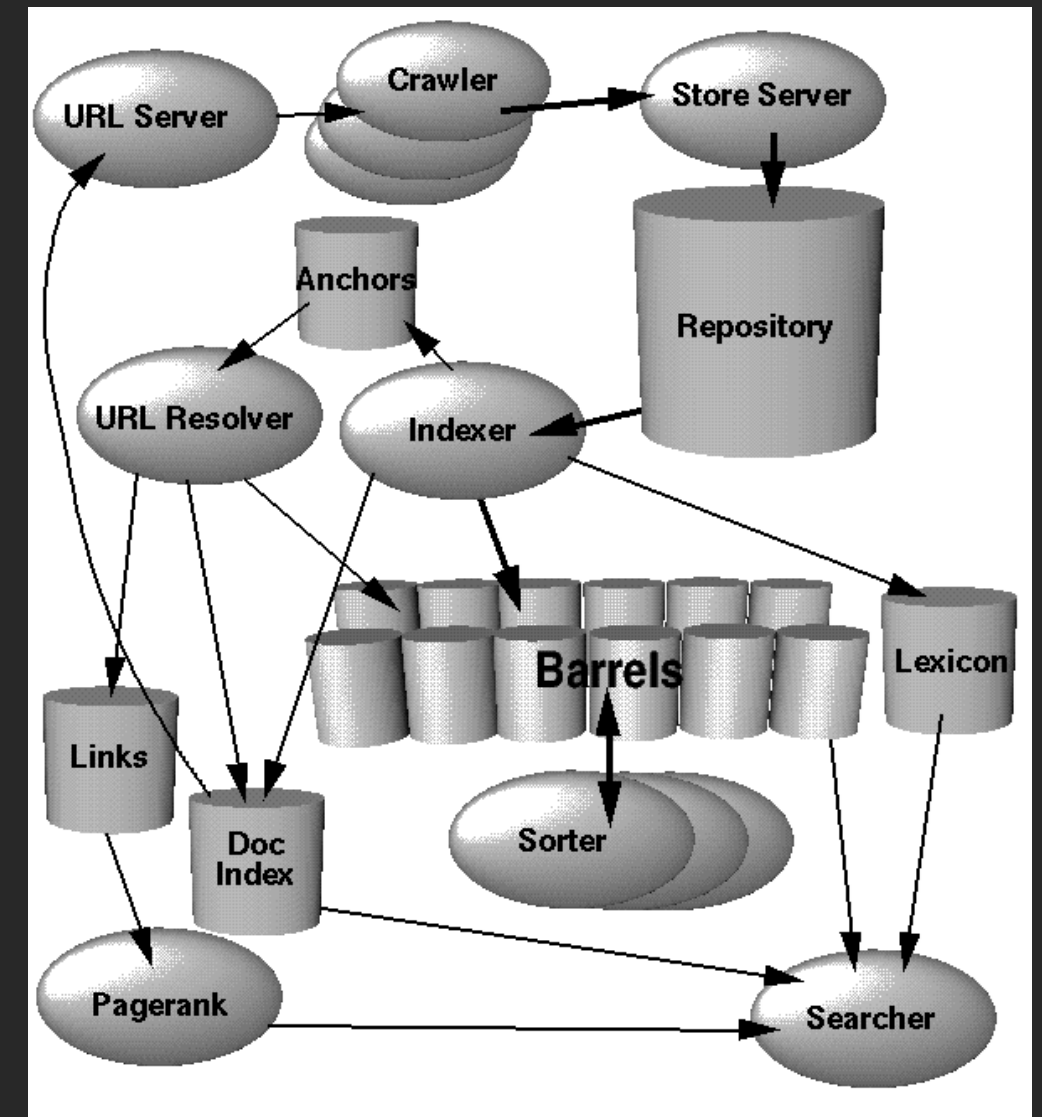Disallow: /images/
Disallow: /tmp/
Disallow: /private/

http://www.py4e.com/robots.txt
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)
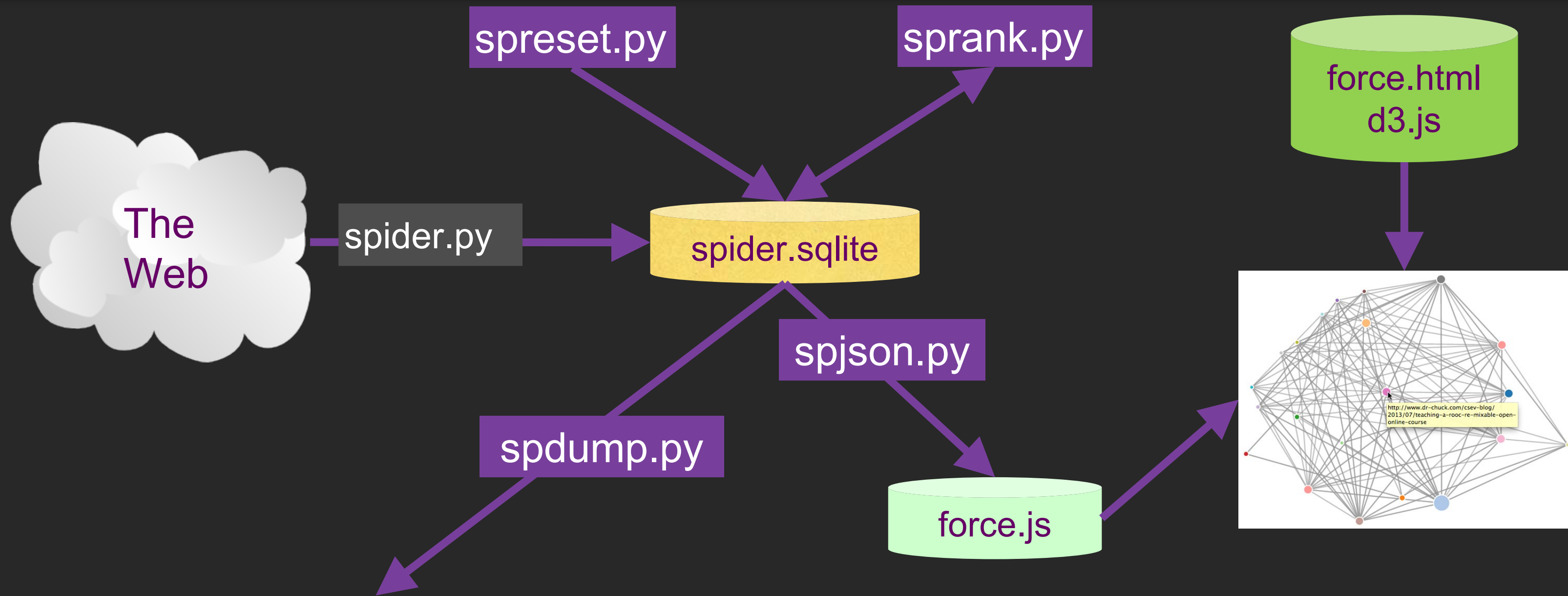
(5, None, 1.0, 3, u'http://www.dr-chuck.com/csev-blog')
(3, None, 1.0, 4, u'http://www.dr-chuck.com/dr-chuck/resume/speaking.htm')
(1, None, 1.0, 2, u'http://www.dr-chuck.com/csev-blog/')
(1, None, 1.0, 5, u'http://www.dr-chuck.com/dr-chuck/resume/index.htm')
4 rows.

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

# Visualizing Mail Data

# Acknowledgements / Contributions

...