

Retrieving and Visualizing Data

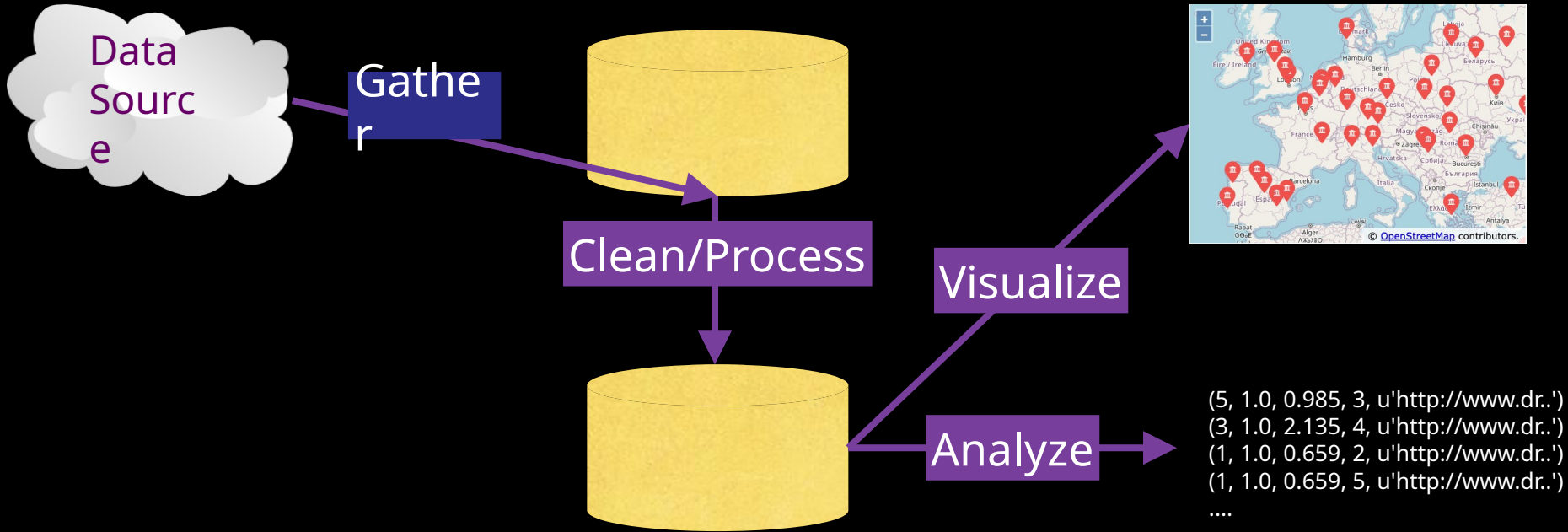
Charles Severance



Python for Everybody
www.py4e.com



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"

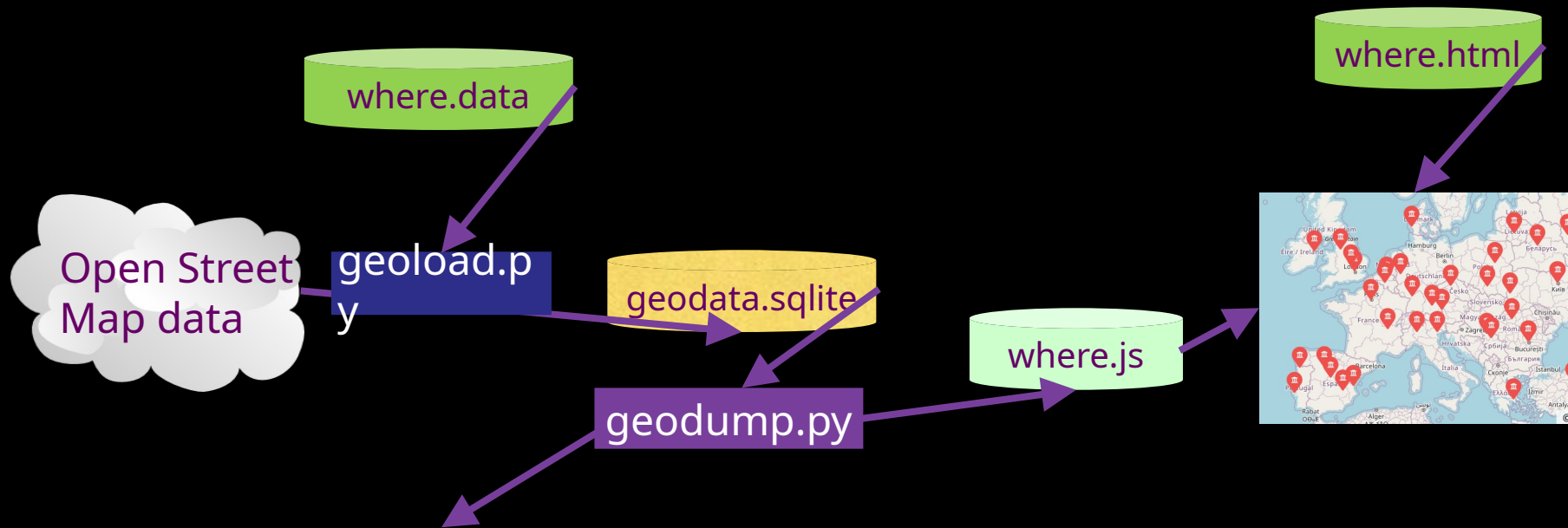
Our goal is to make you better programmers – not to make you data mining experts

OpenGeo

- Makes an annotated Open Street Map from user entered data
- Uses the proxied GeoAPI API
- Caches data in a database to avoid rate limiting and allow restarting
- Visualized in a browser using the Open Street Map



<http://www.py4e.com/code3/opengeo.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
 Kokshetau, Kazakhstan 53.2833333 69.3833333

...

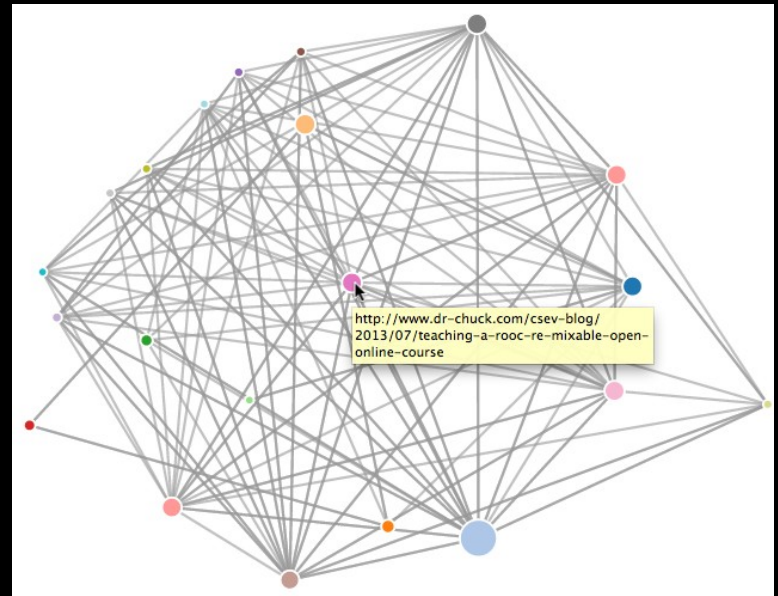
Technical University Munich 48.14907275 11.567444920339295
 178 records written to where.js

Open where.html to view the data in a browser

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

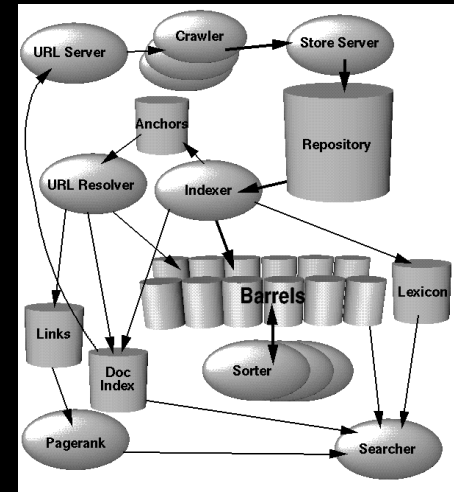
Page Rank

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



Search Engine Architecture

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



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

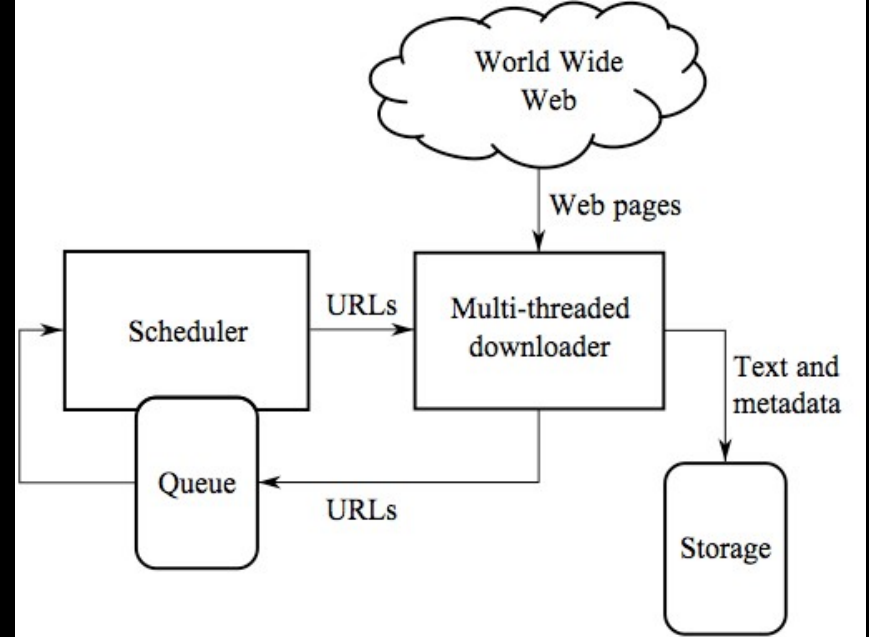
Web Crawler

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

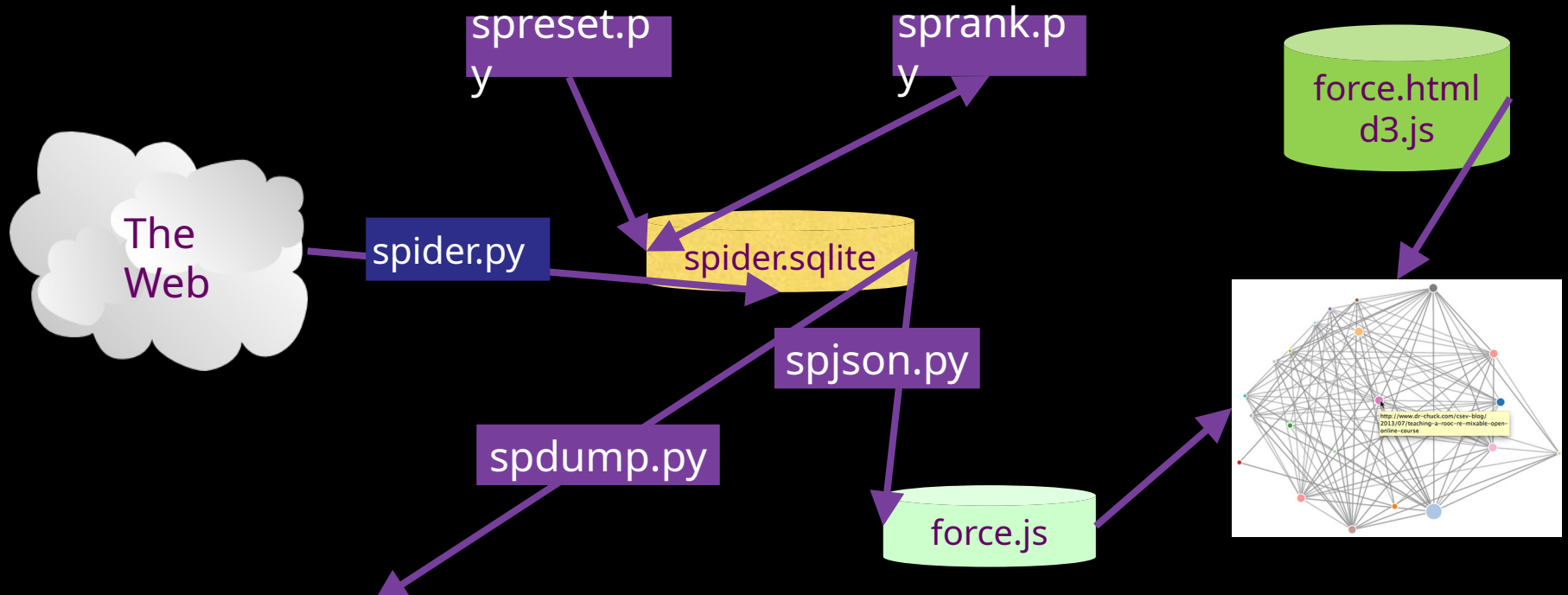
```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /images/  
Disallow: /tmp/  
Disallow: /private/
```

[http://en.wikipedia.org/wiki/
Robots_Exclusion_Standard](http://en.wikipedia.org/wiki/Robots_Exclusion_Standard)
http://en.wikipedia.org/wiki/Spider_trap

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.

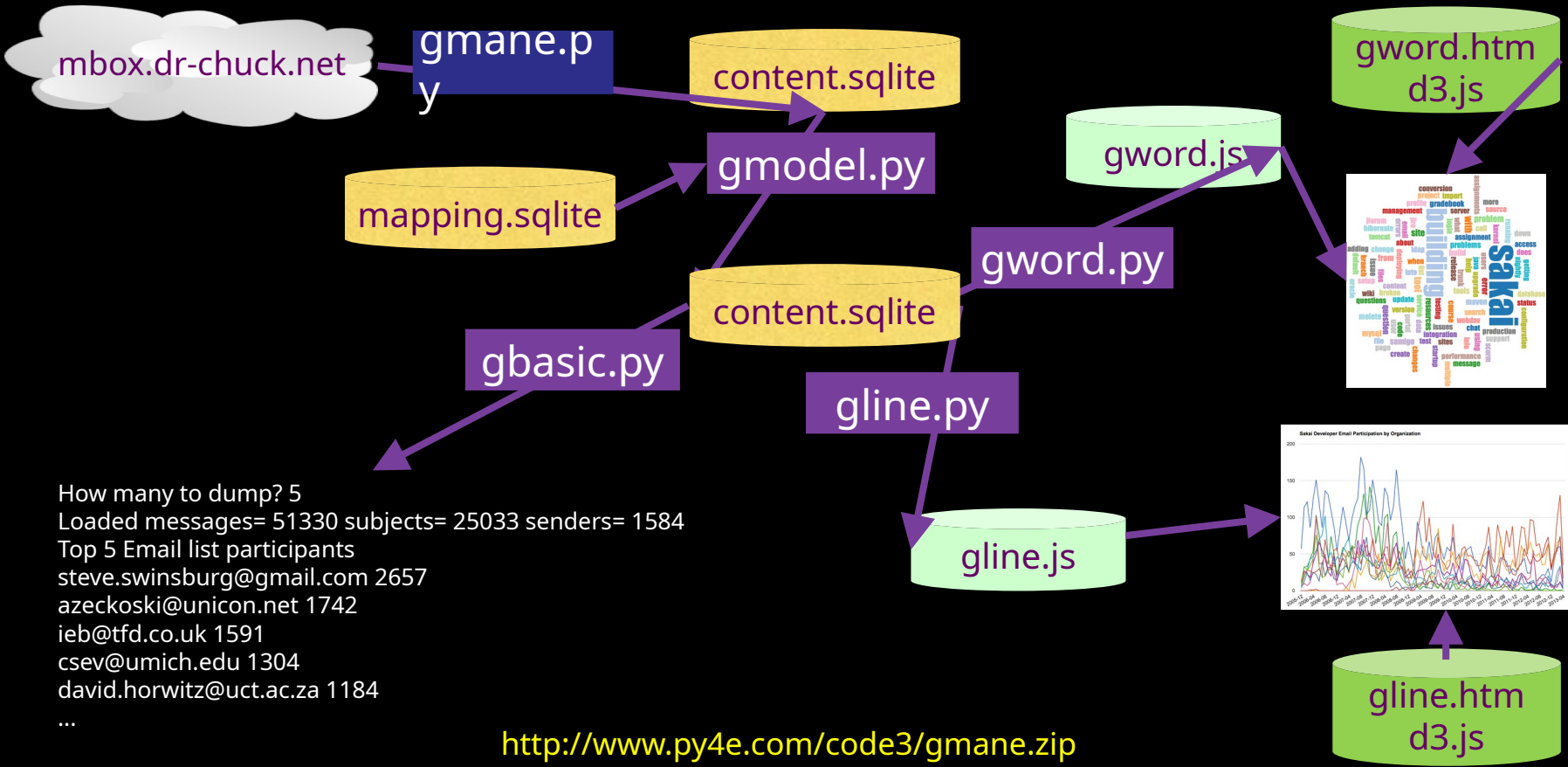
<http://www.py4e.com/code3/pagerank.zip>

Warning: This Dataset is > 1GB

- Sadly the original source of this data (gmane.org) has been shut down
- We made a copy of a subset of the data before it was shut down

Use this for your testing:

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



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.py4e.com/code3/gmane.zip>



Acknowledgements / Contributions



These slides are Copyright 2010- Charles R. Severance (www.dr-chuck.com) of the University of Michigan School of Information and 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