

Mailing Lists - Gmane

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



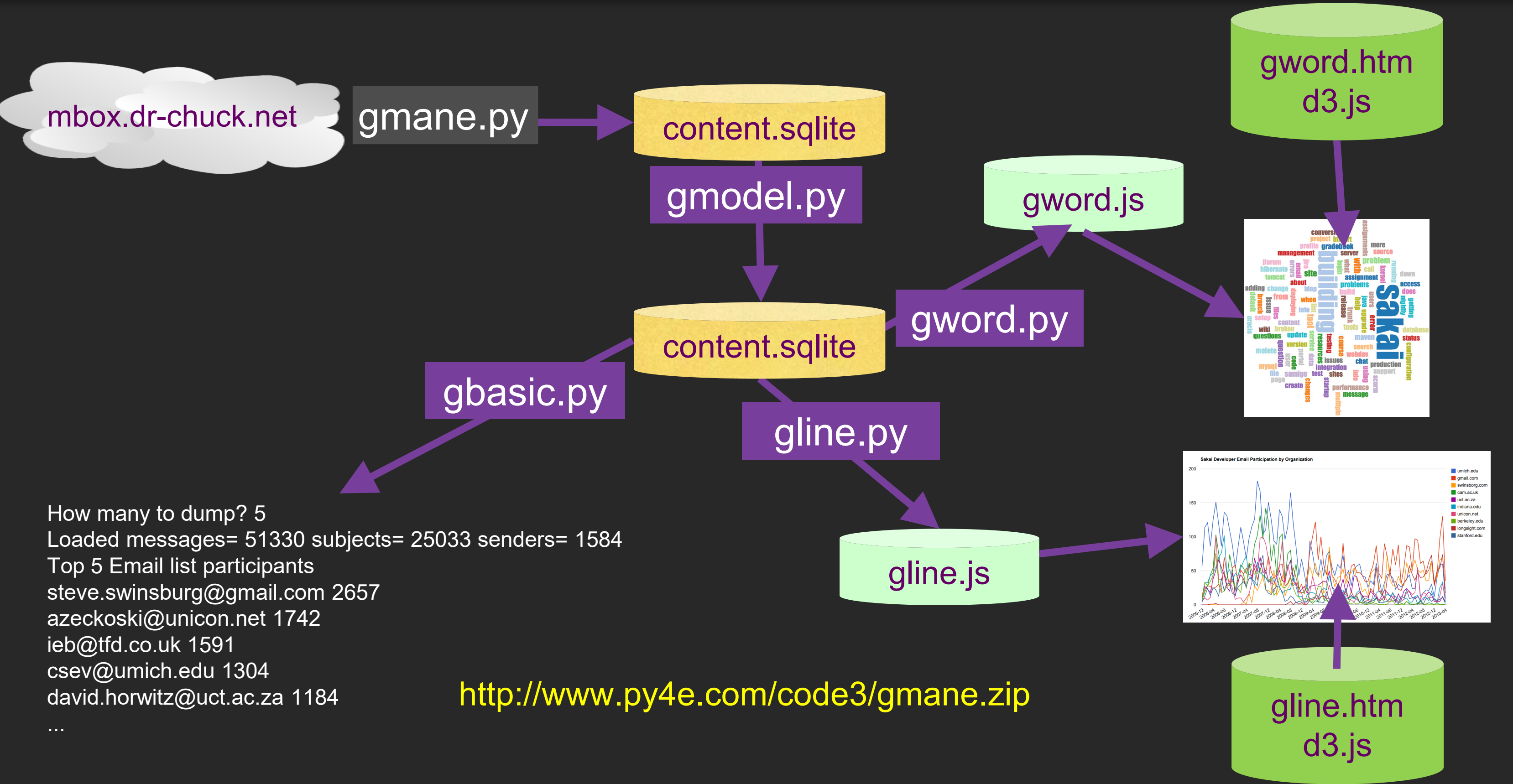
<http://www.py4e.com/code3/gmane.zip>

Warning: This Dataset is > 1GB

- Sadly the original source of this data (gmane.org) has 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>

Summary

- Combining all of the techniques in the class, we can make data management system that can pull data (restarting as necessary), clean it up, and visualize it



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

...