

O'REILLY®
oscon
open source convention

JULY 16–20, 2012 **PORTLAND, OR**

#oscon

OPEN KNOWLEDGE

DIGGING INTO OPEN DATA

Kim Rees, Periscope

[@krees](#), [@periscope](#)

kim@periscope.com



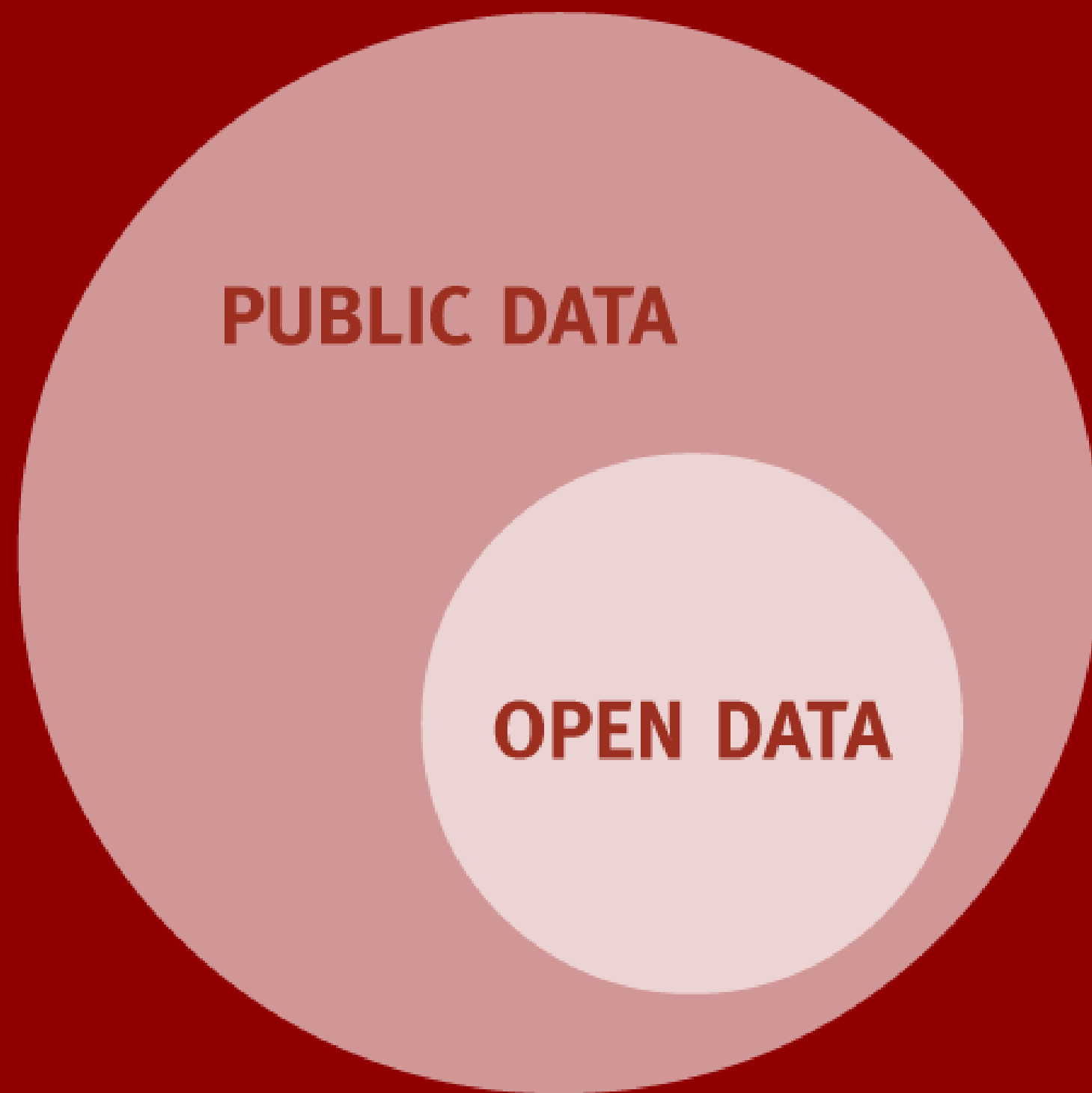


PUBLIC \neq OPEN

Copyrights, patents, trademarks,
restrictive licenses, etc.

OPEN DATA IS...

- Accessible without limitations on entity or intent
- In a digital, machine-readable format
- Free of restriction on use or redistribution in its licensing conditions



OPEN ≠ EXEMPT

Be sure to check the Data Use Policies of your sources.

- Citations
- Attributions

See

<http://opendefinition.org/licenses/>



OPEN/PUBLIC \neq GOVERNMENT

- Publications
 - The Guardian, WSJ, NYT, The Economist, etc.
- Companies
 - GE, Yahoo, Nike, Mint, Trulia, etc.
- Academia
 - [Carnegie Mellon DASL](#), [Berkeley Data Lab](#), [MIT Open Data Library](#), etc.

OPEN ≠ ACCESSIBLE



TOP ISSUES

TOP WORDS

CANDIDATES

EXPLORE BY DEBATE

02-22-12

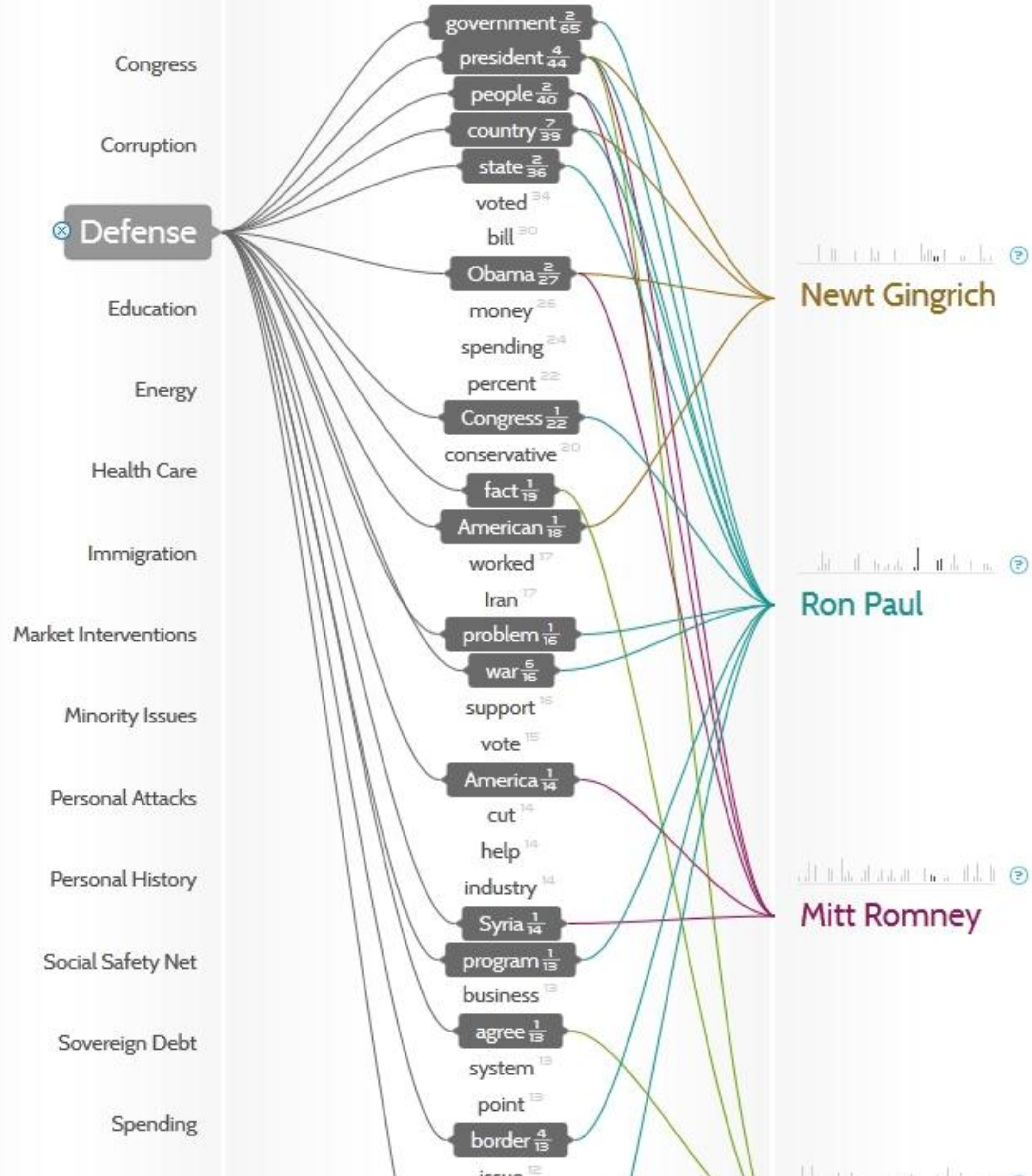
MESA, ARIZONA

Mesa Arts Center

Sponsors: CNN, and the Republican
Party of Arizona

- February 22nd, 2012
- January 26th, 2012
- January 23rd, 2012
- January 19th, 2012
- January 16th, 2012
- January 8th, 2012
- January 7th, 2012
- December 15th, 2011
- December 10th, 2011
- November 22nd, 2011
- November 9th, 2011
- October 18th, 2011
- October 11th, 2011
- September 22nd, 2011
- September 12th, 2011
- September 7th, 2011
- August 11th, 2011
- June 13th, 2011

+ Upcoming Debates



FINDING DATA

- Most government sites (some of these are rabbit holes)
- Commercial Data markets ([Infochimps](#), [DataMarket](#), [Azure Marketplace](#), [Kasabi](#))
- Locating free data
 - <http://thedatahub.org/>
 - Open Science Data: http://oad.simmons.edu/oadwiki/Data_repositories
- Ask! (often you can email researchers/journalists directly to request data you can't find online)
- Research time = liberal estimate * 5



John Woolley and Gerhard Peters

[HOME](#) [DATA](#) [DOCUMENTS](#) [ELECTIONS](#) [MEDIA](#) [LINKS](#)



Document Archive

- Public Papers of the Presidents
- State of the Union Addresses & Messages
- Inaugural Addresses
- [Weekly Addresses](#)
- Fireside Chats
- News Conferences
- Executive Orders
- Proclamations
- Signing Statements
- Press Briefings
- Statements of Administration Policy
- Economic Report of the President
- [Debates](#)
- Convention Speeches
- Party Platforms
- 2012 Election Documents
- 2008 Election Documents
- 2004 Election Documents
- 1960 Election Documents
- 2009 Transition
- 2001 Transition

Data Archive

[Data Index](#)

Media Archive

[Audio/Video Index](#)

Elections

[Election Index](#)

[Florida 2000](#)

Links

[Presidential Libraries](#)

View Public Papers by Month and Year

Month Year



PRESIDENTIAL DEBATES

1960 and 1976 - 2012

Republican Candidates Debate in Mesa, Arizona

February 22, 2012

Like 2.5k

Tweet 194

+1

PARTICIPANTS:

Former Speaker of the House Newt Gingrich (GA);
Representative Ron Paul (TX);
Former Governor Mitt Romney (MA); and
Former Senator Rick Santorum (PA)

MODERATOR:

John King (CNN)

KING: Gentlemen, I want to ask you to take your seats. I'll take a moment now to explain to you how our debate will work.

I'll question the candidates, as well as we'll also take some questions from members of our audience. I'll follow up and guide tonight's discussion.

Candidates, we're going to try to make sure each of you get your fair amount of questions. And you'll have a minute to answer and 30 seconds for rebuttal and follow-ups. And if you're singled out for a particular criticism, I'll make sure you get a chance to respond.

Now we're going to have each of the candidates introduce themselves. And so we have more time to debate tonight, we're going to ask them to keep it short.

Here's an example. I'm John King from CNN. I'm honored to be your moderator tonight and I'm thrilled to be in a state that reminds us baseball season is just around the corner. *[applause]*

KING: Congressman Paul, we begin with you, sir.

PAUL: I'm Congressman Ron Paul, a congressman from Texas.

I am the defender of the Constitution. I'm the champion of liberty. This shows the roadmap to peace and prosperity. *[applause]*

SANTORUM: I'm Rick Santorum.

And we have a lot of troubles around the world, as you see, the Middle East in flames, and what's going on in this country with gas prices and the economy. And I'm here to talk about a positive solutions that confront this country that include everybody

COLLECTION:
*Campaign
2012*

Location:



United States
Arizona

Font Size:

[A](#) [A](#) [A](#) [A](#)

Print

Report Typo

Share

The American Presidency
Project

facebook

Name:
The American
Presidency Project



SCRAPING DATA

- Needlebase, RIP!!!!

ALTERNATIVES

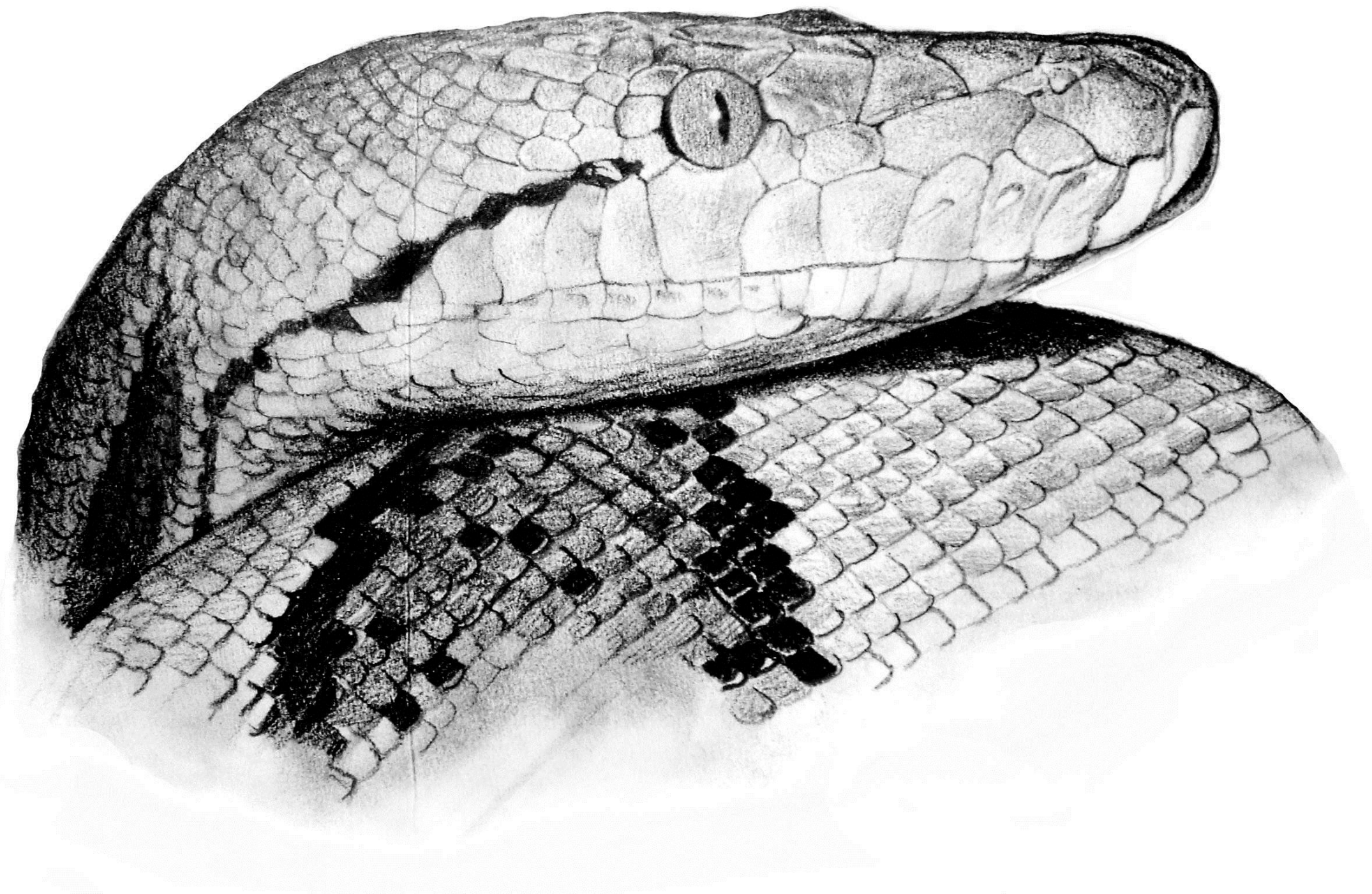
- [WebHarvy](#) (\$\$, robust)
- [Dapper](#) (free, but limited)
- [Google](#) (free, but limited)
- [OutWit Hub](#) (\$\$, free limited version)
- [Mozenda](#) (\$\$\$\$ subscription based)
- [Able2Extract](#) (\$\$, for PDFs)
- [ScraperWiki](#) (free, but programming required)

SCRAPING DATA PROGRAMMATICALLY

You can use any programming language, but Python is the language of choice.

Libraries for getting web pages:

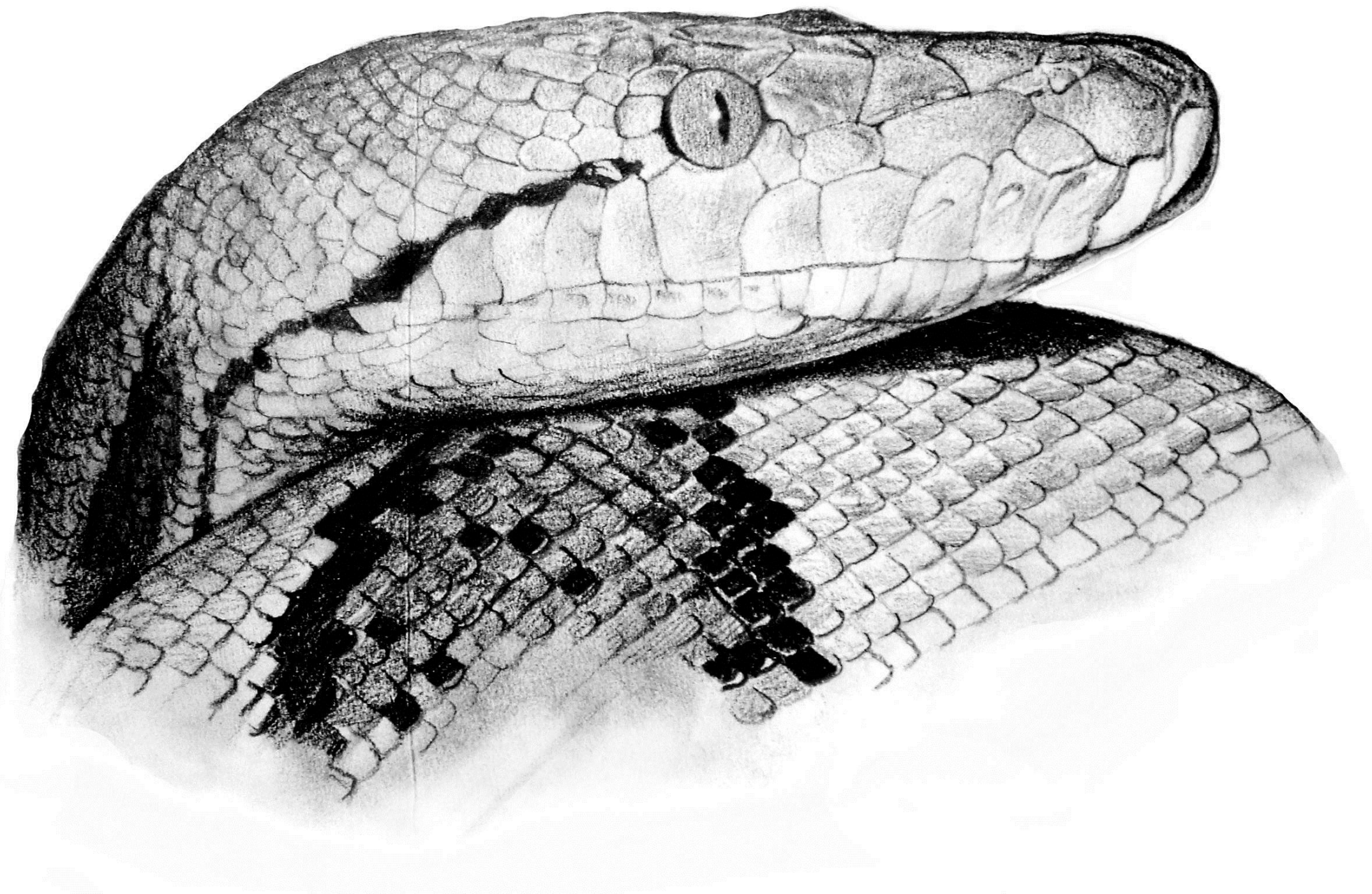
- urllib2
- requests
- mechanize



SCRAPING DATA PROGRAMMATICALLY

Libraries for parsing web pages:

- `html5lib`
- `lxml`
- `BeautifulSoup`




```
import mechanize

url = "http://www.presidency.ucsb.edu/ws/index.php?pid=99556"
b = mechanize.Browser()
b.set_handle_robots(False)
ob = b.open(url)
page = ob.read()
b.close()
```

```
import mechanize
import re

url = "http://www.presidency.ucsb.edu/ws/index.php?pid=99001"
b = mechanize.Browser()
b.set_handle_robots(False)
ob = b.open(url)
html = ob.read()
b.close()

bold = re.compile(' ( (?<=<b>) .*? (?=</b>) ) ')
full = re.compile(' (?s) (?<=<span class="displaytext">) .*? (?=</span>) ')
t = full.search(html).group()
s = list(set(
    [x.replace(":", "") for x in bold.findall(t)]
))
print s
```


C:\Python27>python oscon_ex1.py

```
['SMITH', 'REINHARD', 'SANTORUM', 'ROMNEY', 'MODERATOR', 'PARTICIPANTS',  
 'WILLIAMS', 'GINGRICH']
```

C:\Python27>_



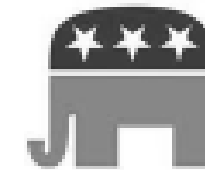
John Woolley and Gerhard Peters

[HOME](#) [DATA](#) [DOCUMENTS](#) [ELECTIONS](#) [MEDIA](#) [LINKS](#)
**Document Archive**

- [Public Papers of the Presidents](#)
- [State of the Union](#)
- [Addresses & Messages](#)
- [Inaugural Addresses](#)
- [Weekly Addresses](#)
- [Fireside Chats](#)
- [News Conferences](#)
- [Executive Orders](#)
- [Proclamations](#)
- [Signing Statements](#)
- [Press Briefings](#)
- [Statements of Administration Policy](#)
- [Economic Report of the President](#)
- [Debates](#)
- [Convention Speeches](#)
- [Party Platforms](#)
- [2012 Election Documents](#)
- [2008 Election Documents](#)
- [2004 Election Documents](#)
- [1960 Election Documents](#)
- [2009 Transition](#)
- [2001 Transition](#)

Data Archive[Data Index](#)**Media Archive**[Audio/Video Index](#)**Elections**[Election Index](#)[Florida 2000](#)**Links**[Presidential Libraries](#)**Presidential Debates • 1960 - 2012****General Election**

<i>October 22nd, 2012</i>	Presidential Debate at Lynn University in Boca Raton, Florida
<i>October 16th, 2012</i>	Presidential Debate at Hofstra University in Hempstead, New York
<i>October 3rd, 2012</i>	Presidential Debate at the University of Denver
<i>October 11th, 2012</i>	Vice-Presidential Debate at Centre College in Danville, Kentucky

Primary Election**Republican Party****2012**

<i>February 22nd, 2012</i>	Republican Candidates Debate in Mesa, Arizona
<i>January 26th, 2012</i>	Republican Candidates Debate in Jacksonville, Florida
<i>January 23rd, 2012</i>	Republican Candidates Debate in Tampa, Florida
<i>January 19th, 2012</i>	Republican Candidates Debate in Charleston, South Carolina
<i>January 16th, 2012</i>	Republican Candidates Debate in Myrtle Beach, South Carolina
<i>January 8th, 2012</i>	Republican Candidates Debate in Concord, New Hampshire
<i>January 7th, 2012</i>	Republican Candidates Debate in Manchester, New Hampshire
<i>December 15th, 2011</i>	Republican Candidates Debate in Sioux City, Iowa
<i>December 10th, 2011</i>	Republican Candidates Debate in Des Moines, Iowa
<i>November 22nd, 2011</i>	Republican Candidates Debate in Washington, DC
<i>November 12th, 2011</i>	Republican Candidates Debate in Spartanburg, South Carolina
<i>November 9th, 2011</i>	Republican Candidates Debate in Rochester, Michigan
<i>October 18th, 2011</i>	Republican Candidates Debate in Las Vegas, Nevada
<i>October 11th, 2011</i>	Republican Candidates Debate in Hanover, New Hampshire
<i>September 22nd, 2011</i>	Republican Candidates Debate in Orlando, Florida
<i>September 12th, 2011</i>	Republican Candidates Debate in Tampa, Florida
<i>September 7th, 2011</i>	Republican Candidates Debate in Simi Valley, California
<i>September 5th, 2011</i>	Palmetto Freedom Forum in Columbia, South Carolina
<i>August 11, 2011</i>	Republican Candidates Debate in Ames, Iowa
<i>June 13th, 2011</i>	Republican Candidates Debate in Manchester, New Hampshire

General Election

<i>October 15th, 2008</i>	Presidential Debate in Hempstead, New York
<i>October 7th, 2008</i>	Presidential Debate in Nashville, Tennessee
<i>September 26th, 2008</i>	Presidential Debate in Oxford, Mississippi
<i>October 2nd, 2008</i>	Vice-Presidential Debate in St. Louis, Missouri

Primary Election**Democratic Party**


```
import mechanize
import re

page_ids = [98936, 99001, 98929]          #page id's of interest
b = mechanize.Browser()
base_url = "http://www.presidency.ucsb.edu/ws/index.php?pid="

html = {}

for pid in page_ids:
    page = b.open(base_url + str(pid))
    print ("processing: " +b.title())
    html[pid] = parseit(page.read())      #our previous script
    page.close()
b.close()
```

```
from nltk import WordNetLemmatizer
```

```
WordNetLemmatizer().lemmatize(token)
```


CLEANING DATA

- Google Refine
- Data Wrangler
- ParseNIP
- Python
- SQL

VISUALIZING DATA

- Tableau (\$\$)
- Spotfire (\$\$)
- Many Eyes, Gephi
- R
- D3, Protovis, etc.

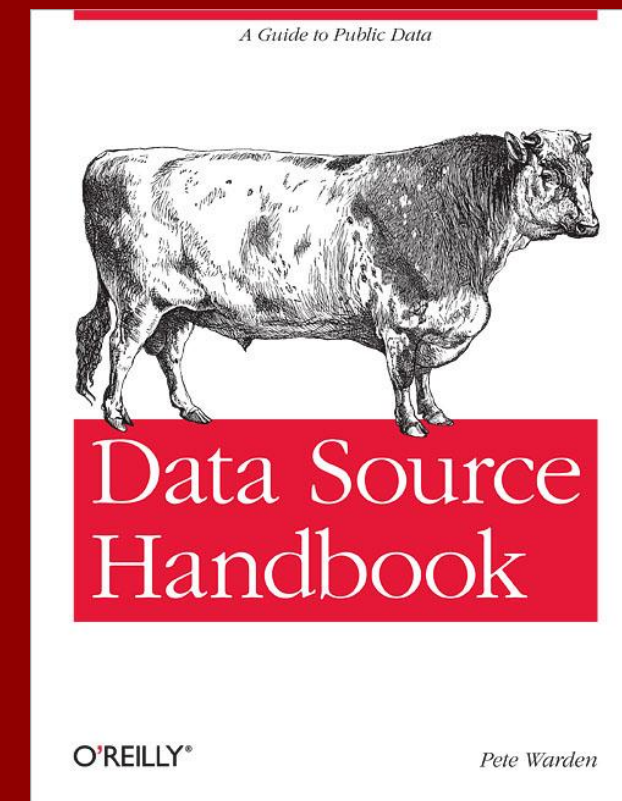
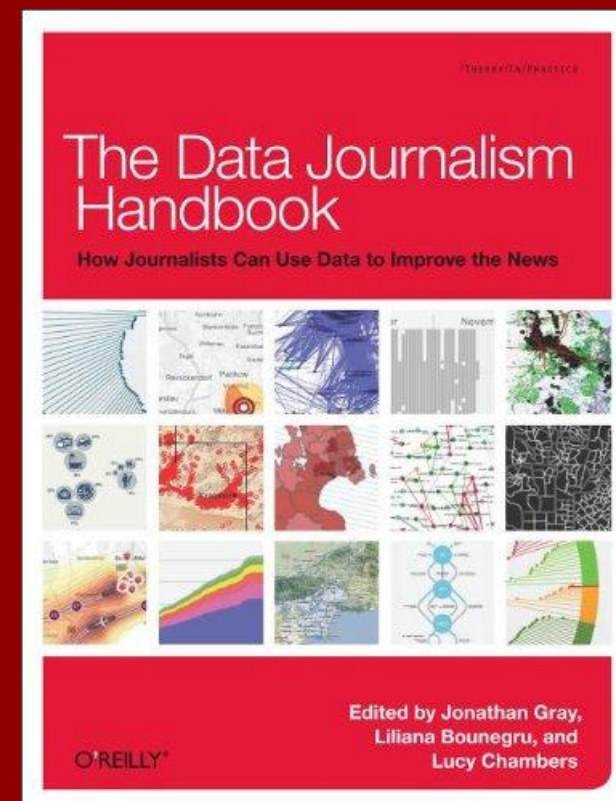
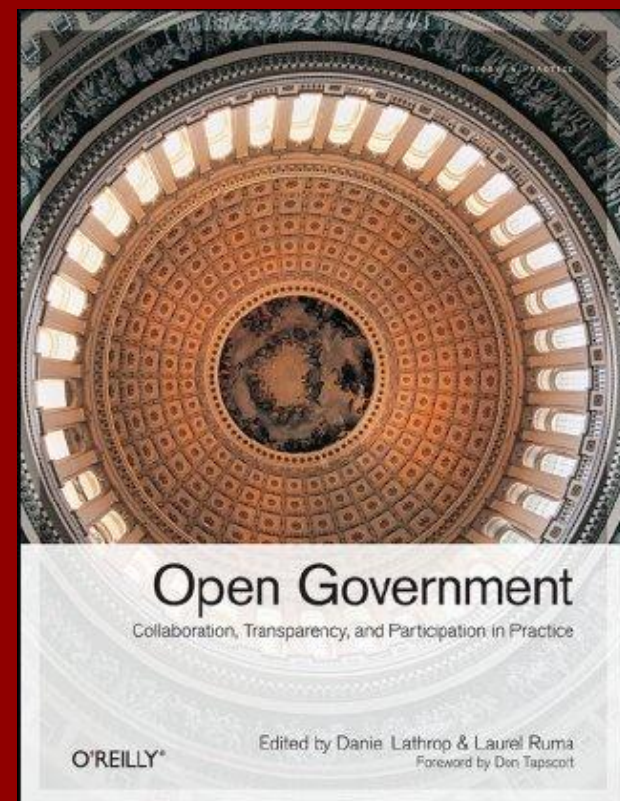
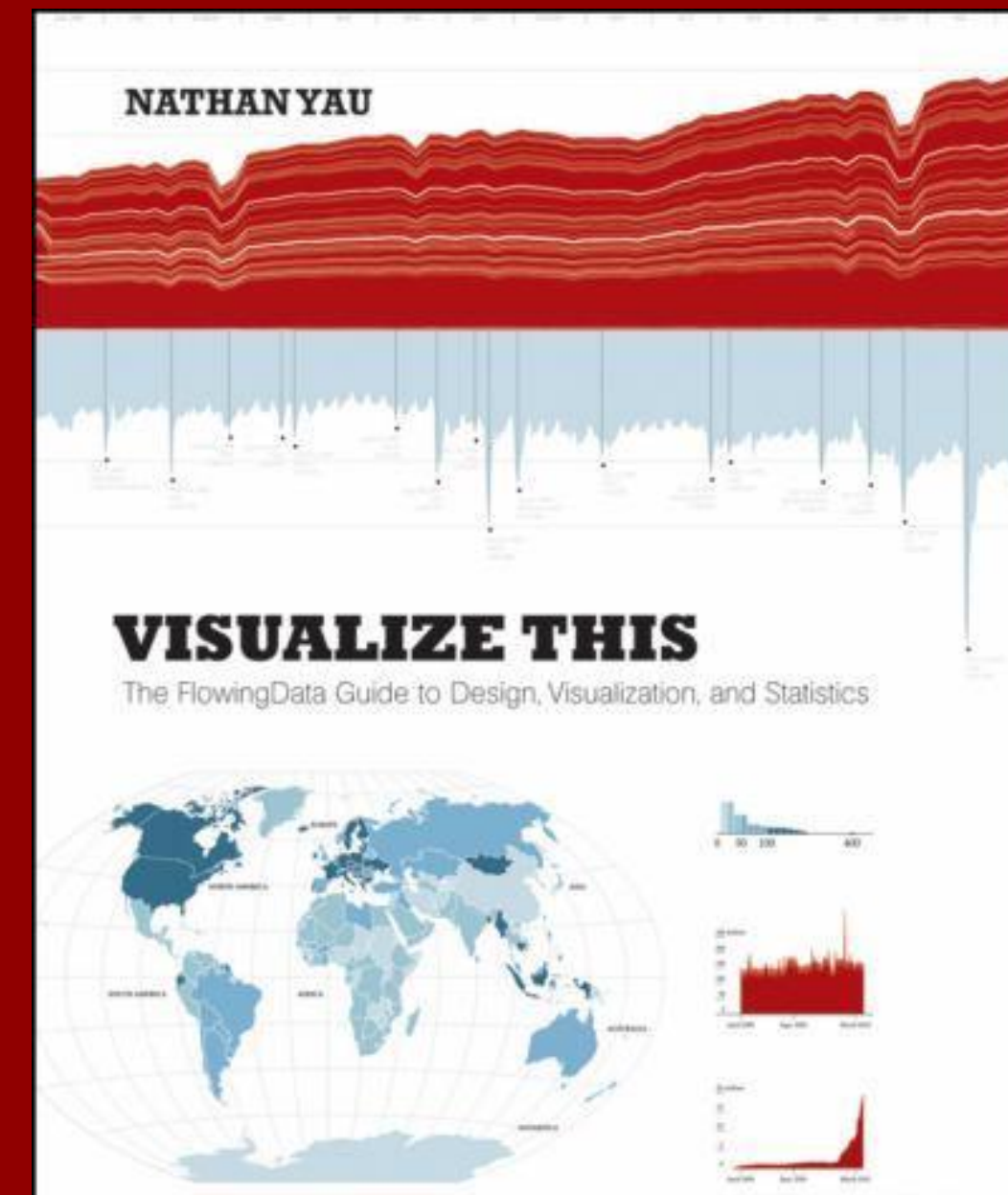
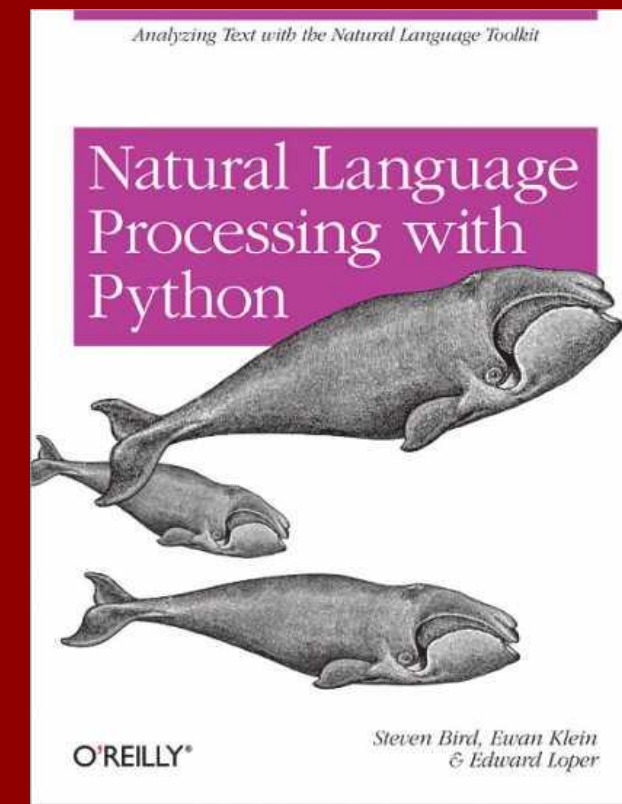
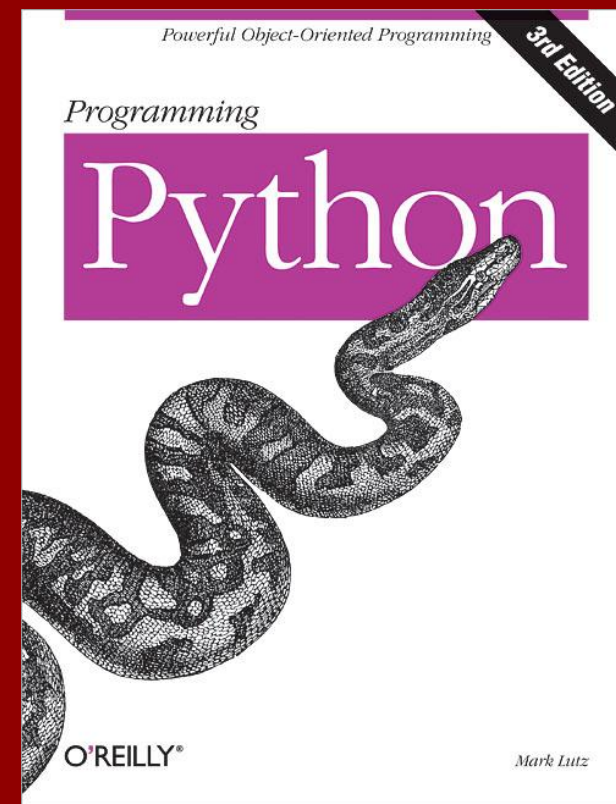
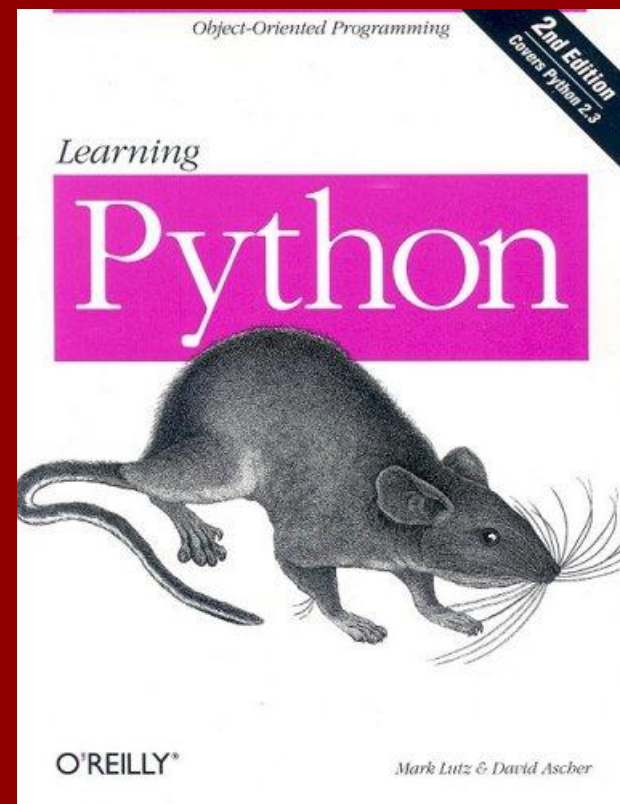
BUSINESS CONSIDERATIONS

- The ins and outs of using existing tools or rolling your own data parsing scripts
- Thinking ahead – the stability of open data
- Data timeliness
- When screen scraping, no one will tell you when the format of the page is going to change. ScraperWiki can help this a bit if it's an option for you.

FUTURE...

- Linked data
- More adoption (keeping up appearances)
- More adoption in private industry
 - Better anonymized data
- Better discovery methods
- Better tools

RESOURCES



O'REILLY®
oscon
open source convention

JULY 16–20, 2012 **PORTLAND, OR**

#oscon

OPEN KNOWLEDGE

DIGGING INTO OPEN DATA

Kim Rees, Periscope

[@krees](#), [@periscope](#)
kim@periscope.com

