The Devil You Know:
Threat Construction in the United States

DATA SOURCE:
INTERNET ARCHIVE'S .GOV COLLECTION

E-SCIENCE DATA INCUBATOR PARTICIPANT:
EMILY KALAH GADE

PROJECT LEADS:
DR. JOHN WILKERSON
DR. MAGDALENA BALAZINSKA

Project Background: Scope

- Political Science Side - NSF Grant re: Poli-Infomatics
  - Explore the wealth of textual resources available today
    - Measuring ideas in text… Using computers to do it
- Computer Science Side – Text in Myria

Project Background: Data Source

- Internet Archive .gov collection
  - 90 terabytes of data, 1.1 billion captures
  - Hosted by Altiscale on a Hadoop Cluster

- Incubator project:
  - Figure out what is "in there"
  - Make it useful for social science research
    - Focus: Language about threats – terrorism, climate change, and banking crisis

Incubator Goals

- Access the cluster, run scripts, get output
- Reliability and validity
- Results and write up
- Make it reproducible and accessible to other social scientists

Initial Hurdles

- Emily's coding abilities
- General big data issues
- Nit picky issues: unicode, parsers, carriage returns
- Best format for examining data
- Reliability and validity

- Reproducibility for anyone else (without the incubator!)

What a capture looks like once parsed

6

What word count data looks like

2009	5	dot\.gov	desertification	1
2009	5	hhs\.gov	(fresh\swater)	21
2009	5	state\.gov	desertification	1409
2009	5	state\.gov	(climate\schange)	28204
2009	5	energy\.gov	pollution	5812
2009	5	energy\.gov	anthropocene	1
2009	5	\.house\.gov	total	655546606
2009	5	defense\.gov	pollution	93
2009	6	ed\.gov pollution	732	
2009	6	va\.gov (global\swarming)	3	
2009	6	dod\.gov	(food\ssecurity)	3
2009	6	dod\.gov	(natural\sdisaster)	274
2009	6	dol\.gov	(forest\sconservation)	9
2009	6	dot\.gov	(food\ssecurity)	25
2009	6	dot\.gov	(greenhouse\sgas)	941
2009	6	hhs\.gov	pollution	1151
2009	6	state\.gov	(food\ssecurity)	4959
2009	6	state\.gov	(natural\sdisaster)	9621
2009	6	\.senate\.gov	anthropogenic	526
2009	6	\.senate\.gov	(ocean\sacidification)	48
2009	6	whitehouse\.gov	(global\swarming)	281
2009	7	other	(greenhouse\sgas)	138771
2009	7	va\.gov (climate\schange)	16	
2009	7	dod\.gov	(fresh\swater)	55
2009	7	state\.gov	(fresh\swater)	838
2009	7	energy\.gov	(forest\sconservation)	2
2009	7	\.house\.gov	(security\sof\sfood)	147
2009	7	\.senate\.gov	total	545452207
2009	7	whitehouse\.gov	desertification	1
2009	7	whitehouse\.gov	(climate\schange)	2157
2009	8	va\.gov total	100787717	
2009	8	va\.gov (natural\sdisaster)	492	

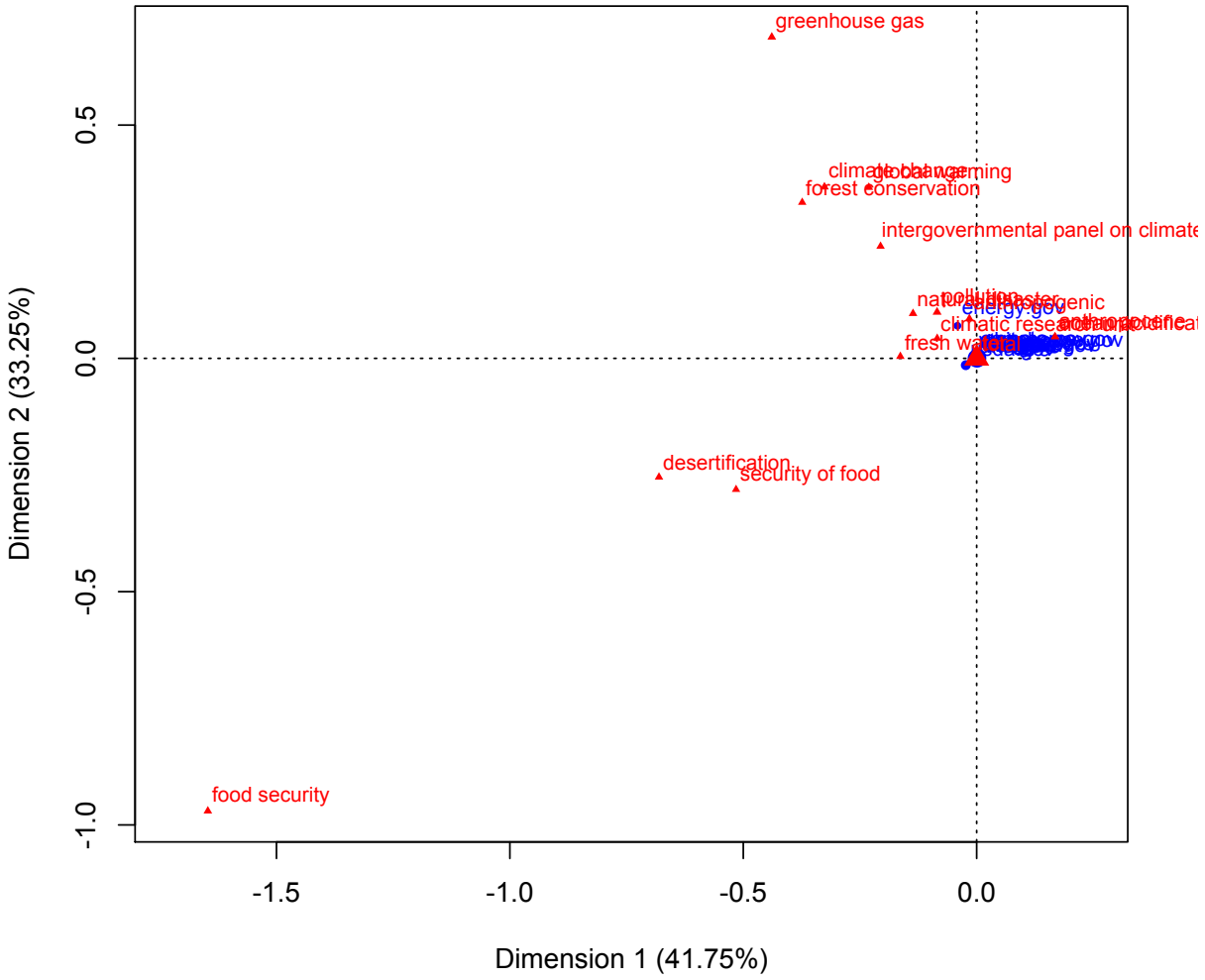
Data Quality

- Complete Crawls Taken from Nov to Jan around elections from 2004 to present
  - This becomes Gold Standard
- Early data crawled irregularly
  - Early data sensitive to terms

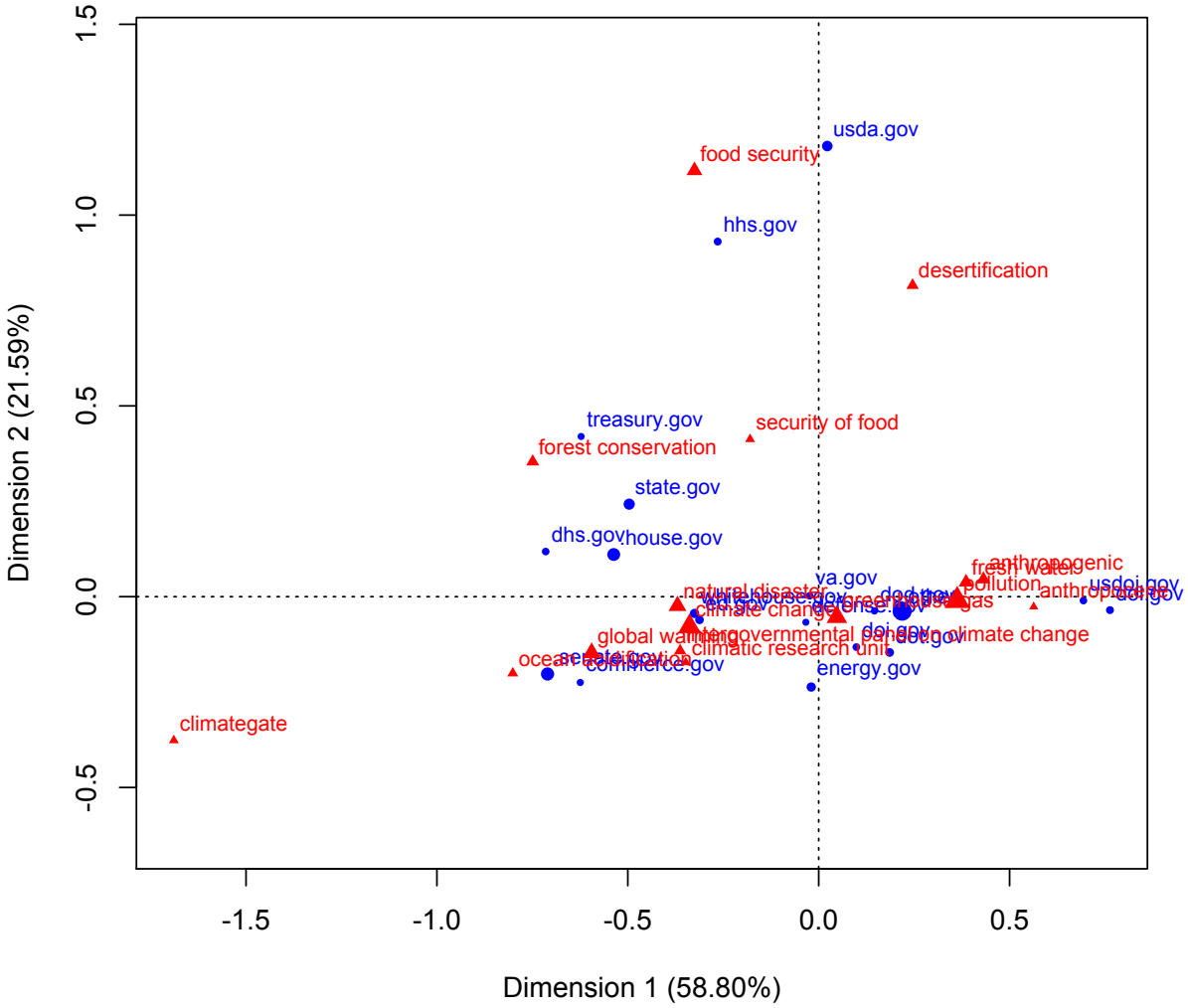
Preliminary Results

- Correspondence Analysis
  - Angles between topics and agency indicate more emphasis of that topic by that agency (smaller angels = more emphasis)
  - Interagency distances and inter topic distances are interpretable (but not topic-agency distances)

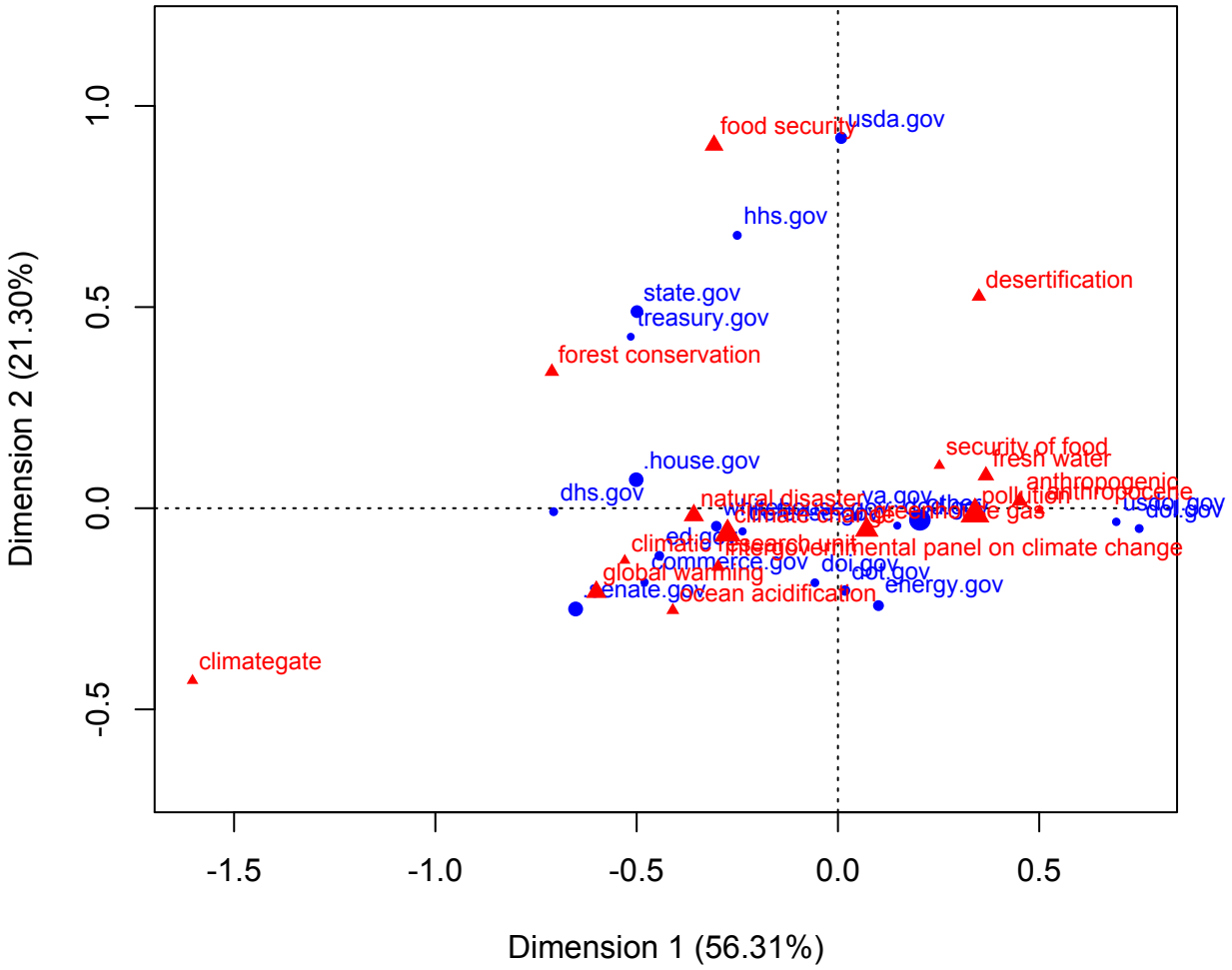
2004 Agency Relative Emphasis Biplot



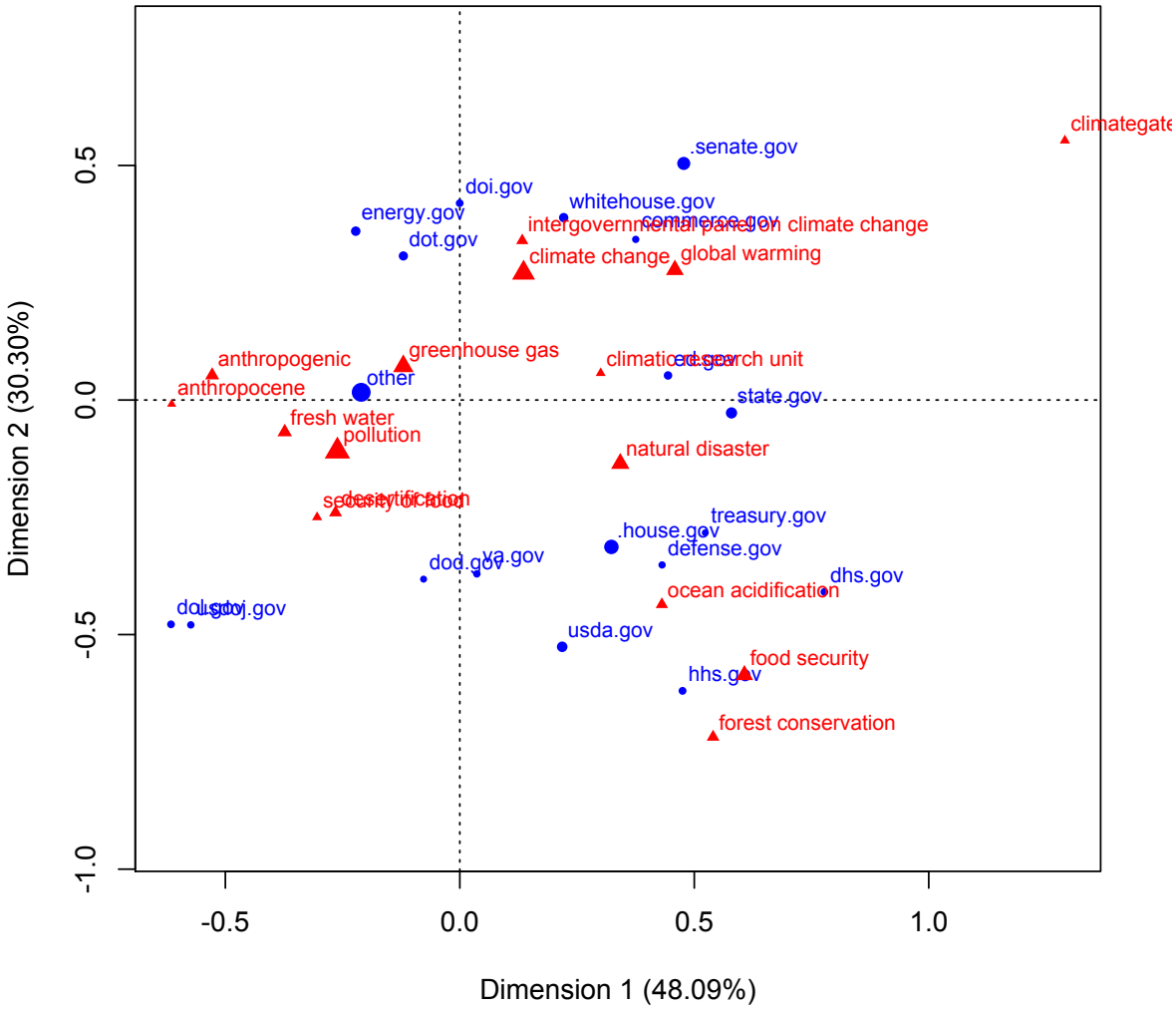
2008 Agency Relative Emphasis Biplot



2012 Agency Relative Emphasis Biplot



2013 Agency Relative Emphasis Biplot



Creating Institutional Memory

- Taking data to PoliInfomatics conference
- Publishing code on Github, publishing data on PoliInfomatics
- Contact with Internet Archive and two other groups known to be exploring this data
  - One at Harvard, one in Germany
- Article re: .gov collection as a data source

Proposed Publications

- Methodological Paper re: .gov data as a data source for social science research
  - Including preliminary results and scripts

- The Devil You Know: Threat Construction in the United States
  - Tracking threat construction in the US; Agenda analysis

Thank You!