Computer-Assisted Text Analysis: An Overview and Guide

Laura K. Nelson

Kellogg School of Management

Content Analysis PWD

Academy of Management Annual Conference

August 7, 2015

Vancouver, BC

Goals

- Describe the wide range of computer-assisted text analysis techniques available
  - Hint: more than dictionaries
- Provide some guidance about how to choose your methods
- Give empirical examples of these methods in practice

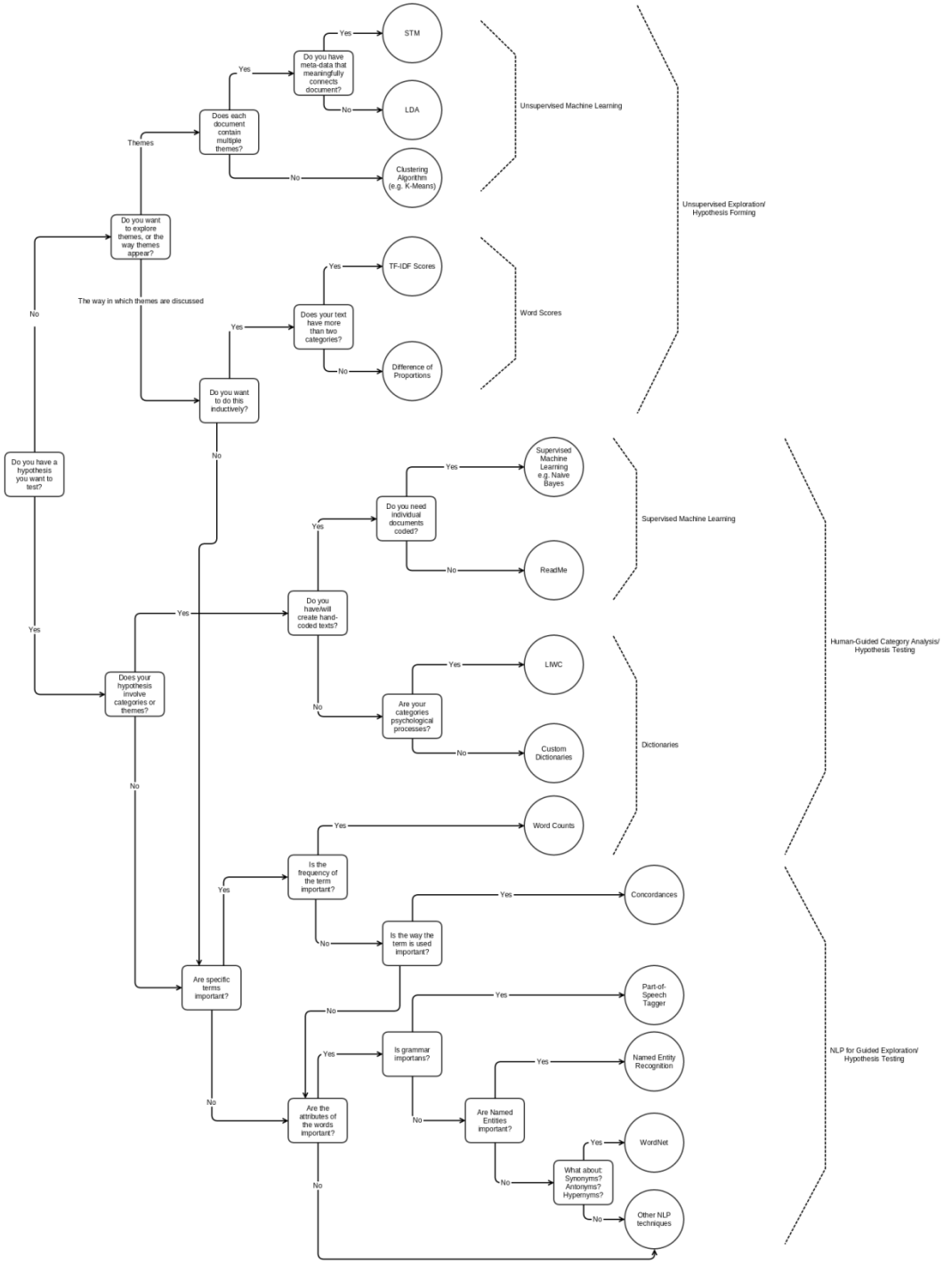
Why Use Computer-Assisted Techniques?

- Speed
  - Humans are slow
  - Text is becoming large
- Reliability / Reproducibility
- Validity (this is controversial)
  - Expanded memory
  - Unburdened by bias

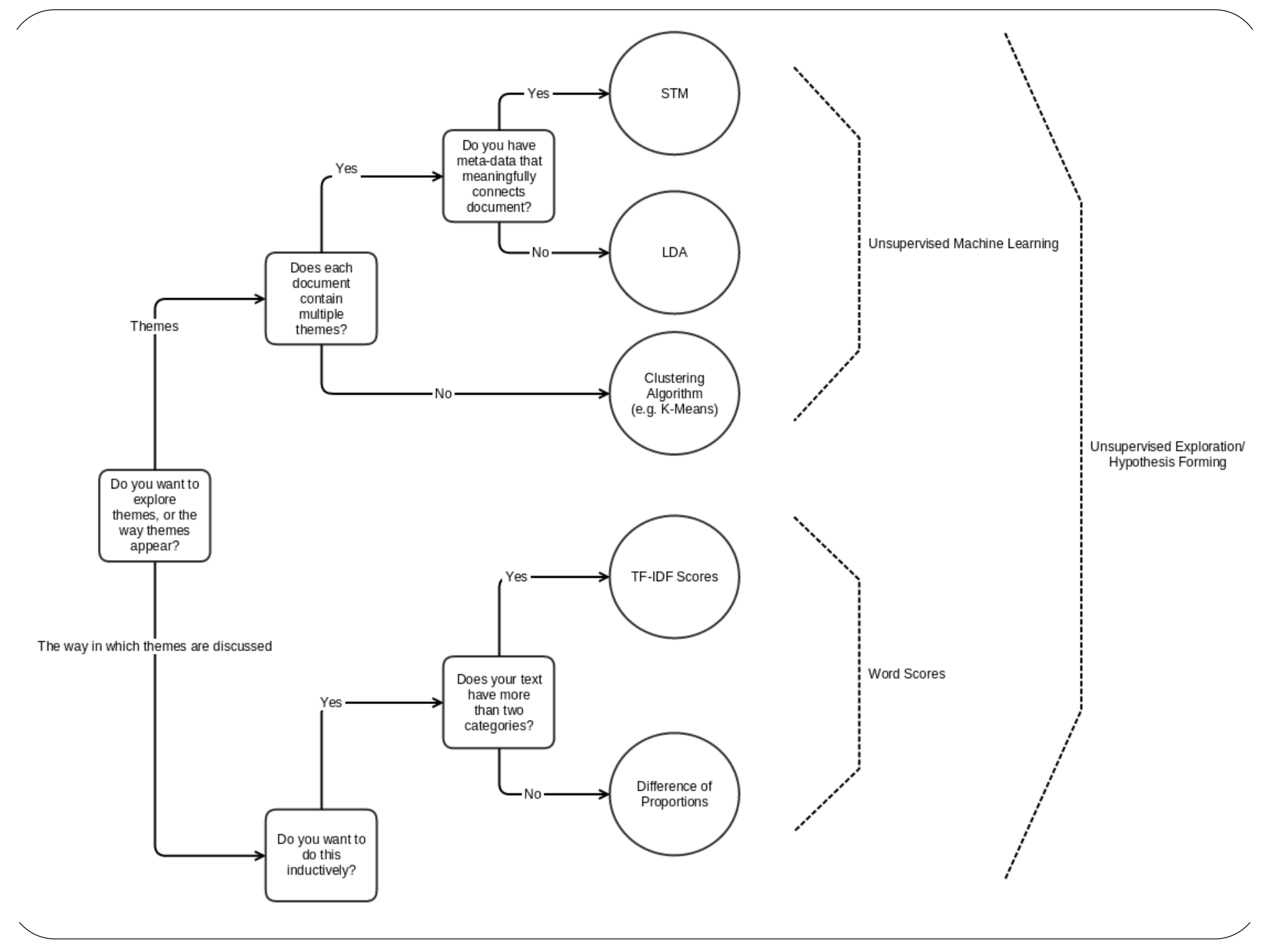
Does not remove the need for interpretation!

Overview:
Types of Automated Text Analysis

- Unsupervised exploration (hypothesis forming/inductive)
  - Topic modeling
  - Lexical selection
- Human-Guided Categorical Analysis (traditional content analysis – deductive hypothesis testing)
  - Supervised machine learning
  - Dictionaries
- Natural Language Processing (guided inductive/hypothesis testing)
  - Part-of-Speech Tagging
  - Named Entity Recognition
  - Concordances
  - Sentiment analysis


Do you have a hypothesis you want to test?
Yes
No
Does your hypothesis involve categories or themes?
Yes
No
Do you have/will create hand-coded texts?
Yes
No
Do you need individual documents coded?
Yes
No
Supervised Machine Learning e.g. Naive Bayes
ReadMe
Supervised Machine Learning
Are your categories psychological processes?
Yes
No
LIWC
Custom Dictionaries
Dictionaries
Word Counts
Human-Guided Category Analysis/ Hypothesis Testing
Do you want to do this inductively?
Yes
No
Do you want to explore themes, or the way themes appear?
Themes
The way in which themes are discussed
Does each document contain multiple themes?
Yes
No
Do you have meta-data that meaningfully connects document?
Yes
No
STM
LDA
Clustering Algorithm (e.g. K-Means)
Unsupervised Machine Learning
Does your text have more than two categories?
Yes
No
TF-IDF Scores
Difference of Proportions
Word Scores
Unsupervised Exploration/ Hypothesis Forming
Are specific terms important?
Yes
No
Is the frequency of the term important?
Yes
No
Is the way the term is used important?
Yes
No
Concordances
Is grammar importans?
Yes
No
Part-of-Speech Tagger
Are the attributes of the words important?
Yes
No
Are Named Entities important?
Yes
No
Named Entity Recognition
What about: Synonyms? Antonyms? Hypernyms?
Yes
No
WordNet
Other NLP techniques
NLP for Guided Exploration/ Hypothesis Testing

Question 1:
Do you want to inductively explore the text?


Do you want to explore themes, or the way themes appear?
Themes
Does each document contain multiple themes?
Yes
Do you have meta-data that meaningfully connects document?
Yes
STM
No
LDA
No
Clustering Algorithm (e.g. K-Means)
Unsupervised Machine Learning
The way in which themes are discussed
Do you want to do this inductively?
Yes
Does your text have more than two categories?
Yes
TF-IDF Scores
No
Difference of Proportions
Word Scores
Unsupervised Exploration/ Hypothesis Forming

Unsupervised Exploration: The Goal

Informative
Groups of Words

Set-Up: Document-Term Matrix*

	ambit	poverti	people	full
Document1	4	2	0	0
Document2	1	3	7	0
Document3	2	0	0	1
Document4	9	1	4	2
Document5	0	0	2	6

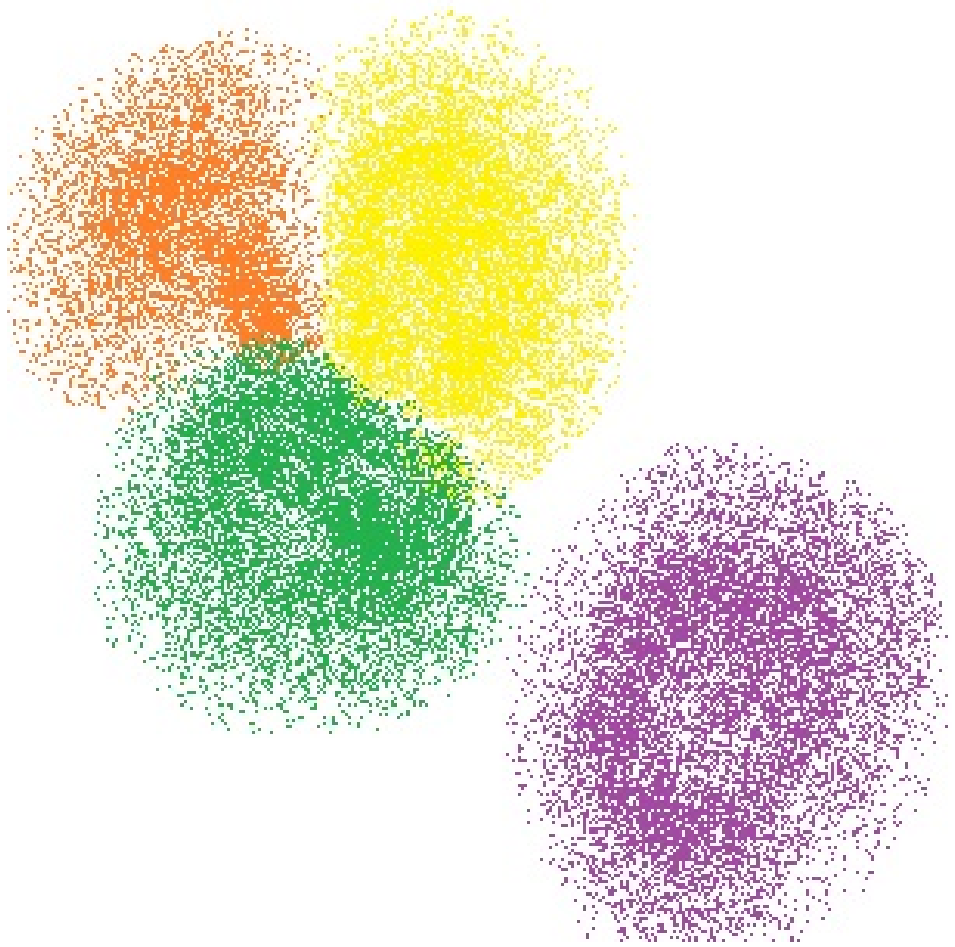
*Cells can be word frequencies or weighted word scores

Question 2: Themes or Style?

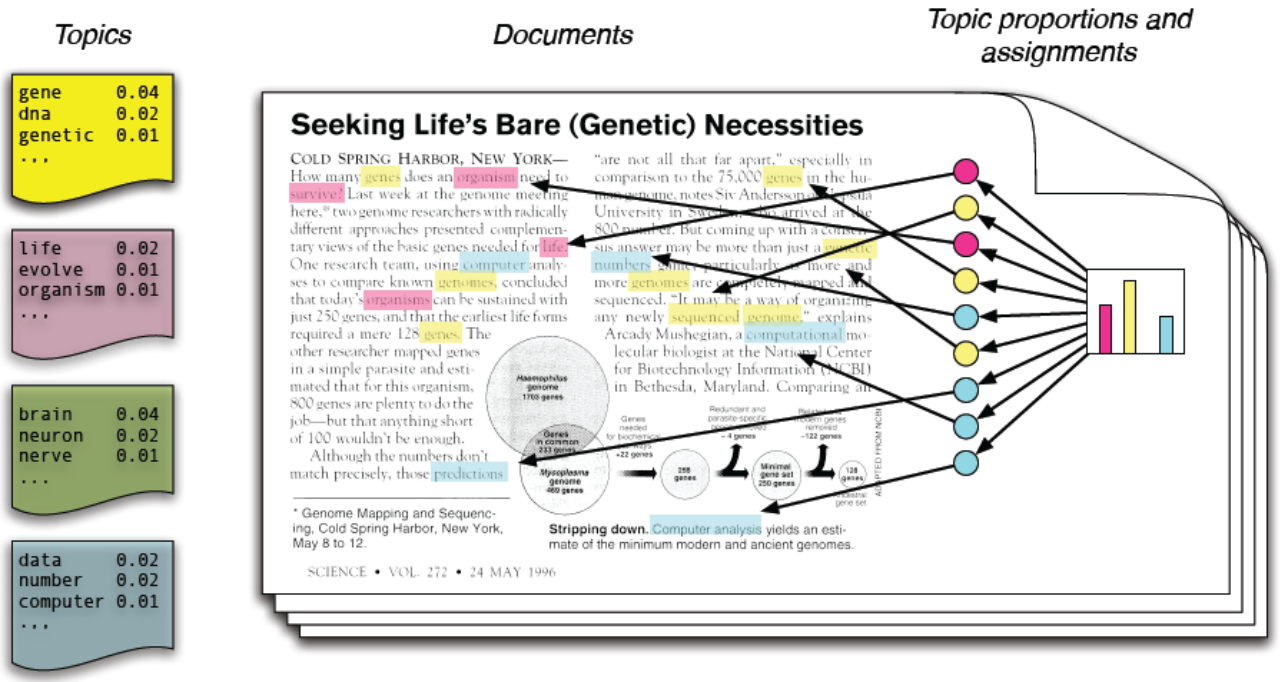
If themes:

Question 3: Multiple Categories?

Single Category per Text: Clustering



Multiple Categories: Topic Modeling



But Which Algorithm?

- Is the order of the documents important?
  - Yes? Structural Topic Modeling (STM)
- Are the topics correlated?
  - Yes? Correlated Topic Modeling (CTM)
- Order is relatively arbitrary, topics may not be related?
  - Latent Dirichlet Allocation (LDA)

music band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach	show film television movie series says life man character know
theater play production show stage street broadway director musical directed	clinton bush campaign gore political republican dole presidential senator house	stock market percent fund investors funds companies stocks investment trading	restaurant sauce menu food dishes street dining dinner chicken served	budget tax governor county mayor billion taxes plan legislature fiscal

Question 2: Themes or Style?

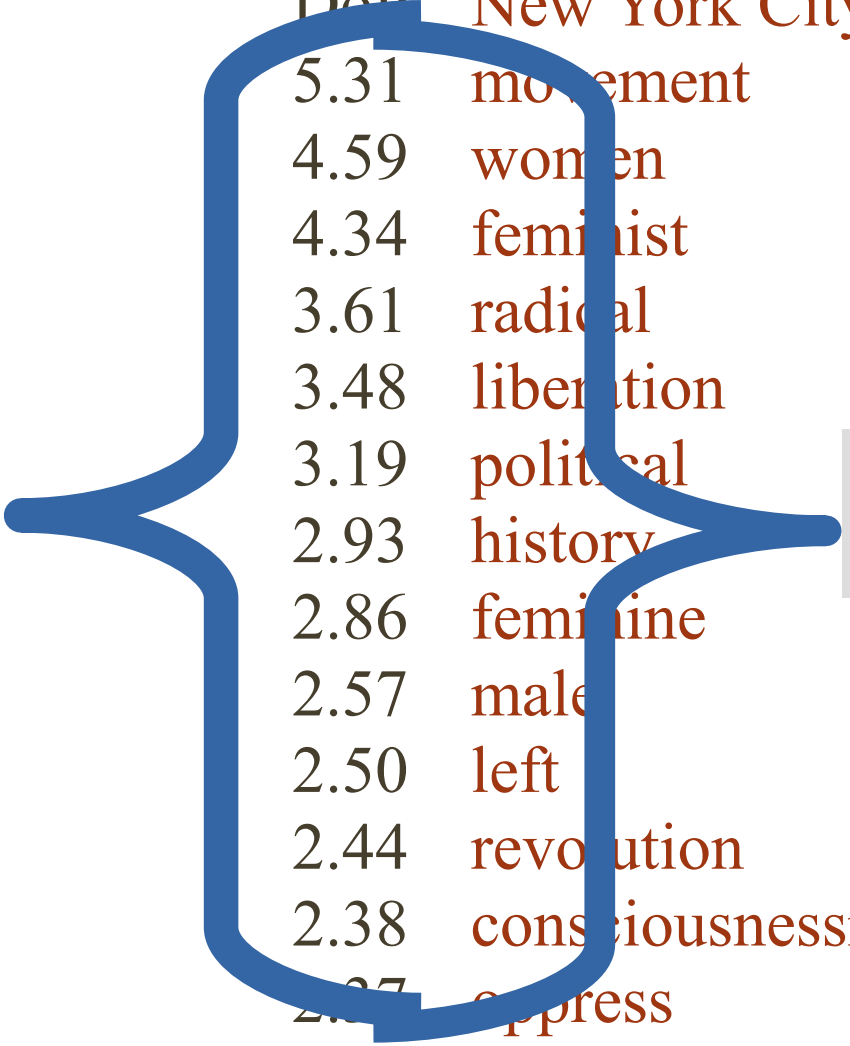
If style...

Style: Lexical Selection

- Goal: find words that are distinctive to different groups of text
- One solution: Difference of Proportions

Difference of Proportions

Chicago	DoP	New York City	DoP
chicago	5.31	mo ement	12.54
children	4.59	wom en	11.34
center	4.34	femi ist	8.91
union	3.61	radic al	8.56
school	3.48	liber tion	7.69
	3.19	polit al	
	2.93	history	
day	2.86	femi ine	3.85
vietnam	2.57	male	3.52
people	2.50	left	2.96
city	2.44	revo ution	2.58
hospital	2.38	cons iousnessraising	2.45
cwlu	2.37	o press	2.41



Abstract

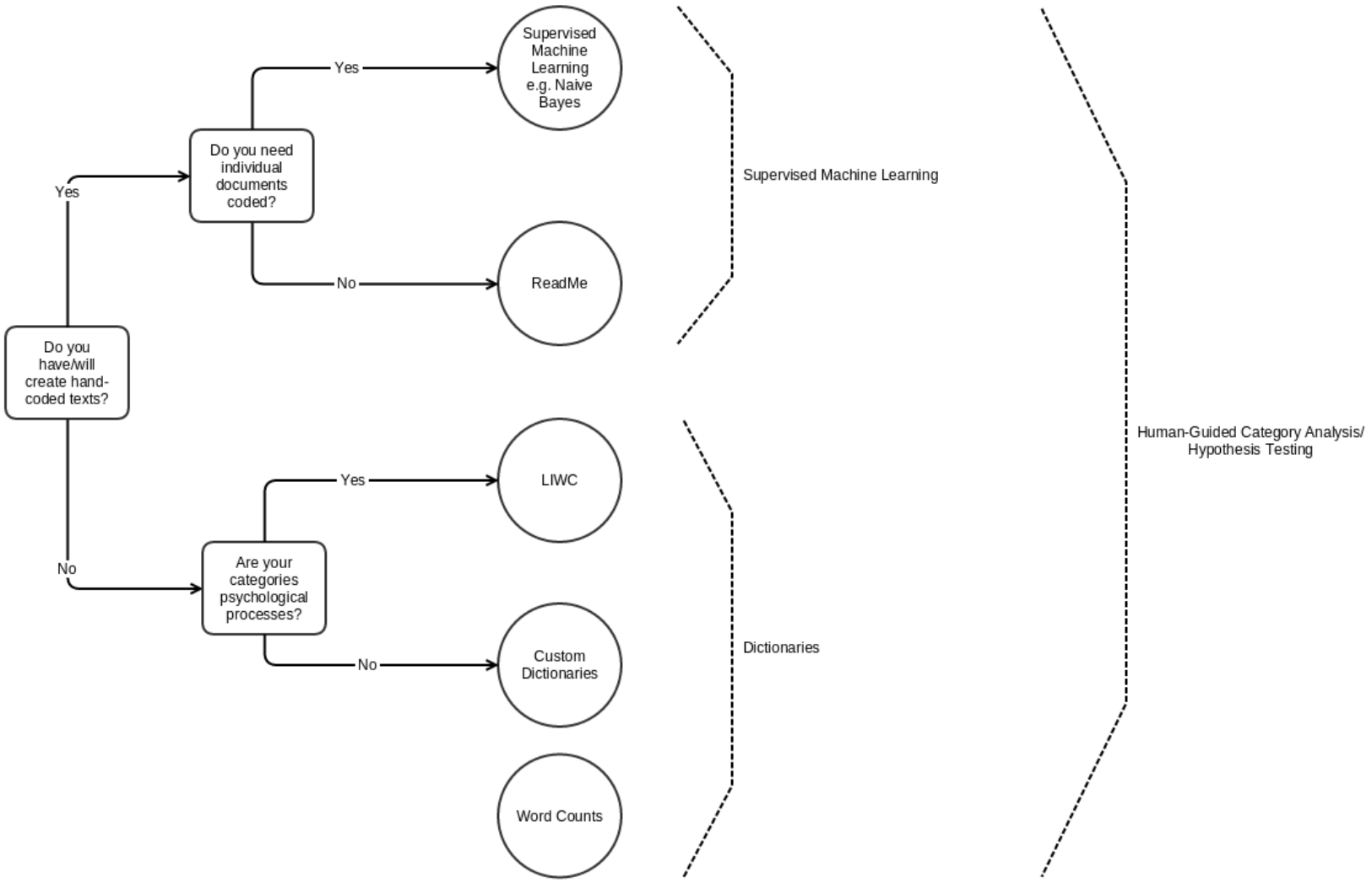
Concrete

Question 1:
Do you want to test a hypothesis?

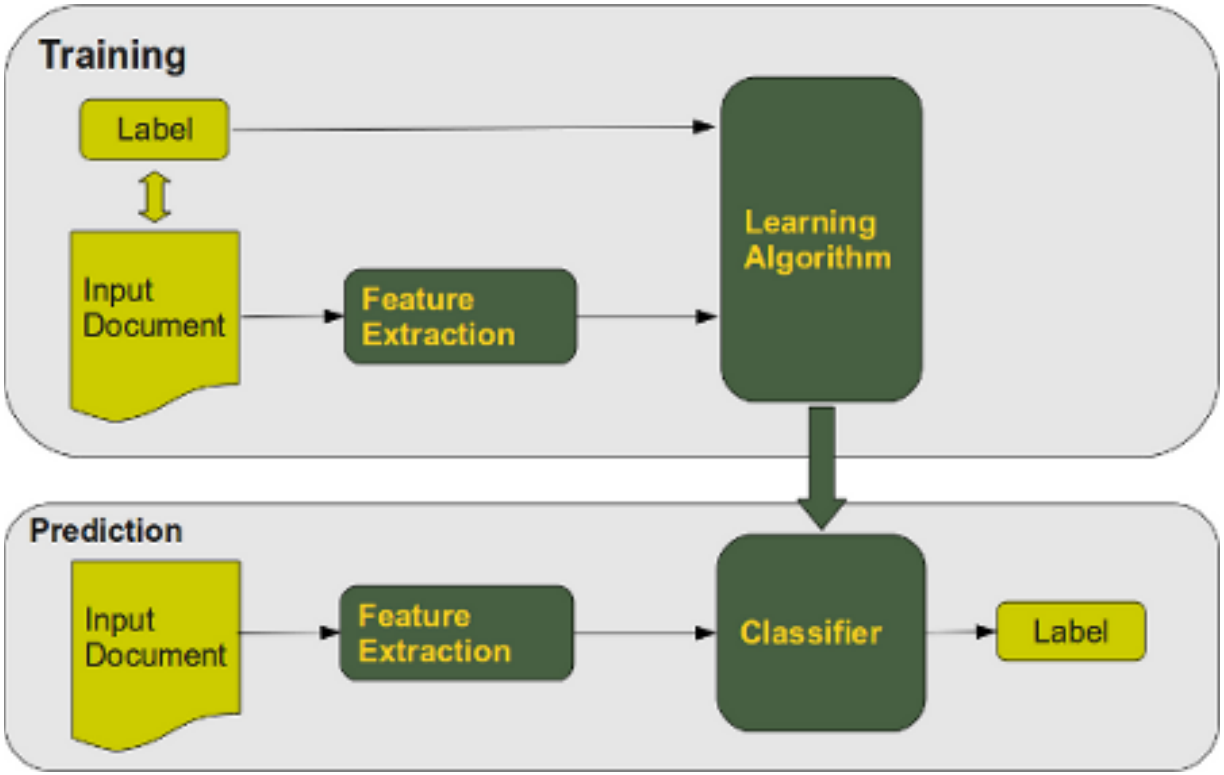
If yes:

Question 2: Themes or Styles?

If themes…


Do you have/will create hand-coded texts?
Yes
Do you need individual documents coded?
Yes
Supervised Machine Learning e.g. Naive Bayes
No
ReadMe
No
Are your categories psychological processes?
Yes
LIWC
No
Custom Dictionaries
Word Counts
Supervised Machine Learning
Dictionaries
Human-Guided Category Analysis/ Hypothesis Testing

Supervised Machine Learning



Which Algorithm?

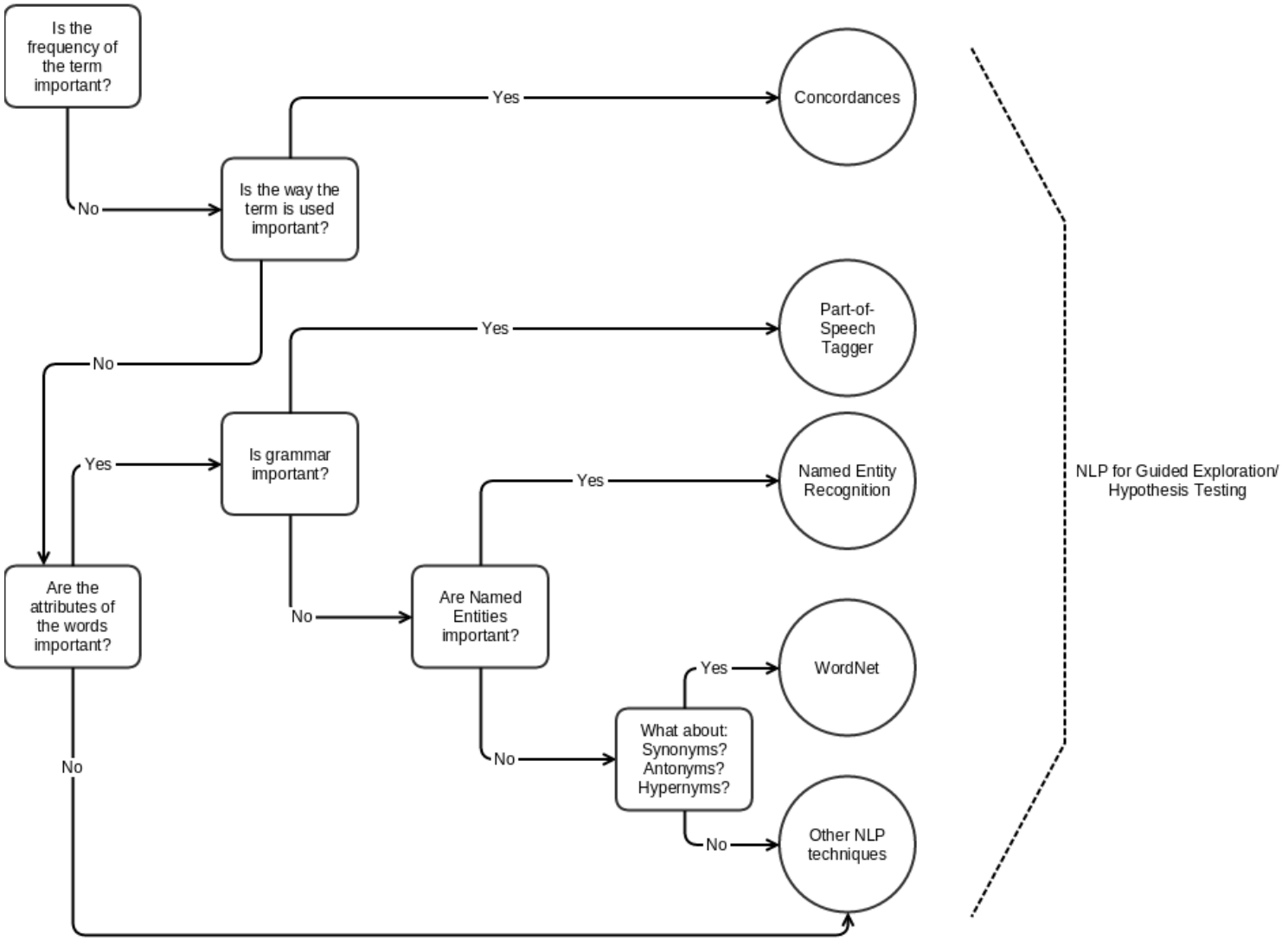
- You want individual documents coded:
  - Document Classification (e.g. SVM, Naive Bayes)
- You want proportion of documents in each category:
  - ReadMe (R package)

Dictionary Methods

- Standardized Dictionaries
  - LIWC (can be used for sentiment analysis)
- Custom Dictionary

Question 2: Themes or Styles?

If styles...


Is the frequency of the term important?
Yes
Concordances
No
Is the way the term is used important?
Yes
Part-of-Speech Tagger
No
Yes
Is grammar important?
Yes
Named Entity Recognition
Are the attributes of the words important?
No
Are Named Entities important?
Yes
WordNet
No
What about: Synonyms? Antonyms? Hypernyms?
No
Other NLP techniques
NLP for Guided Exploration/ Hypothesis Testing

Natural Language Processing

- Takes into account features of words, relationships between words, grammatical structures, etc.

Examples: Use NLP to test hypotheses

- Hypothesis: Author A is more descriptive than Author B.
  - Test: Part-of-Speech tagger, extract adjectives, count and compare.
- Hypothesis: Organizations in New York City are more internationally focused than organizations in Silicon Valley.
  - Test: Named Entity Recognition, compare against lists of corporations, places, and people.

Example: Use NLP to test hypotheses

- Hypothesis: the word "disruptive" is used in a positive way, and has a different meaning, for Silicon Valley organizations compared to Wall Street organizations.
- Test: Concordances

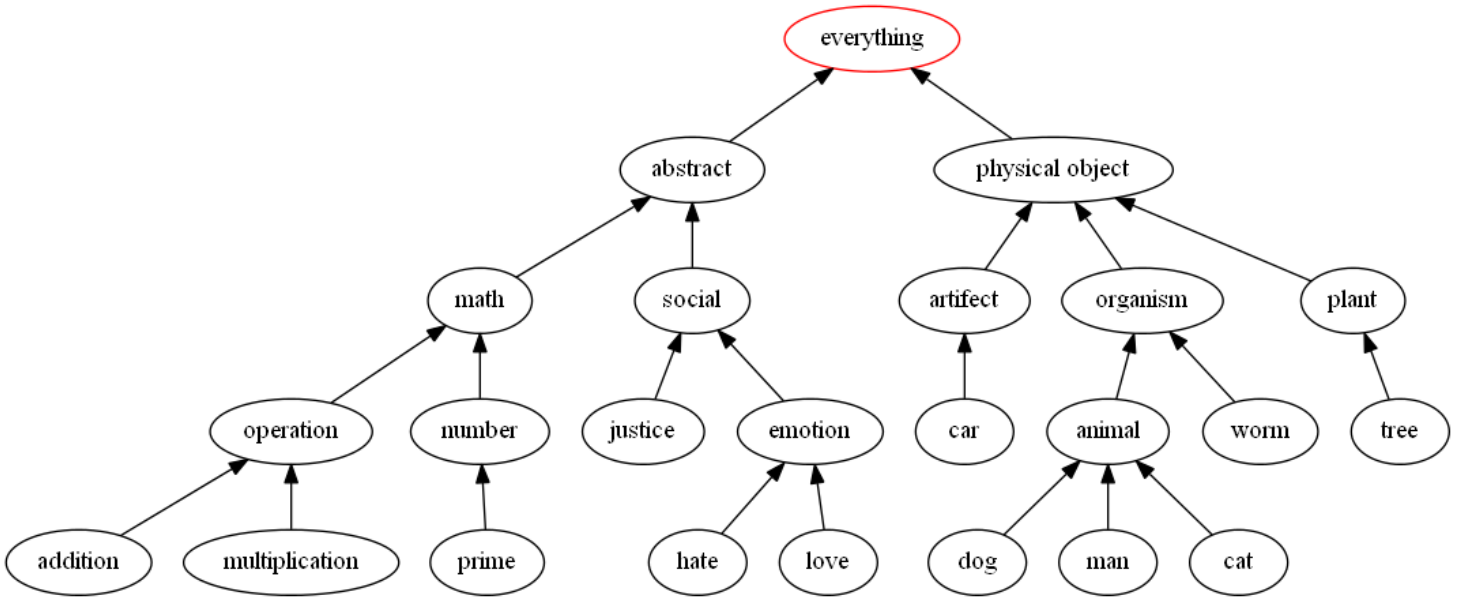
NLP: Concordances

```
ong the former , one was of a most monstrous size . ... This came towards us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork we have r
ll over with a heathenish array of monstrous clubs and spears . Some were thick
d as you gazed , and wondered what monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himmal
they might scout at Moby Dick as a monstrous fable , or still worse and more de
th of Radney .'" CHAPTER 55 Of the monstrous Pictures of Whales . I shall ere l
ing Scenes . In connexion with the monstrous pictures of whales , I am strongly
ere to enter upon those still more monstrous stories of them which are to be fo
ght have been rummaged out of this monstrous cabinet there is no telling . But
```

- very heartily so exceedingly remarkably as vast a great amazingly extremely good sweet
  - Mostly positive
- mean part maddens doleful gamesome subtly uncommon careful untoward exasperate loving passing mouldy christian few true mystifying imperial modifies contemptible
  - Mostly negative

Example: NLP and WordNet

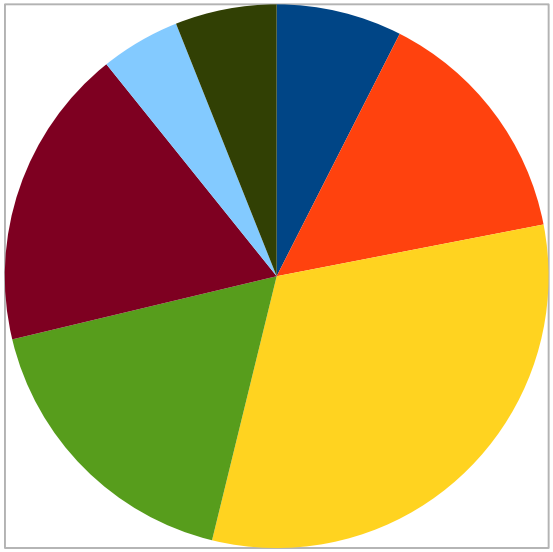
- Hypothesis: Women's movement organizations in New York City approach politics more abstractly compared to those in Chicago, who have a more concrete approach to politics.



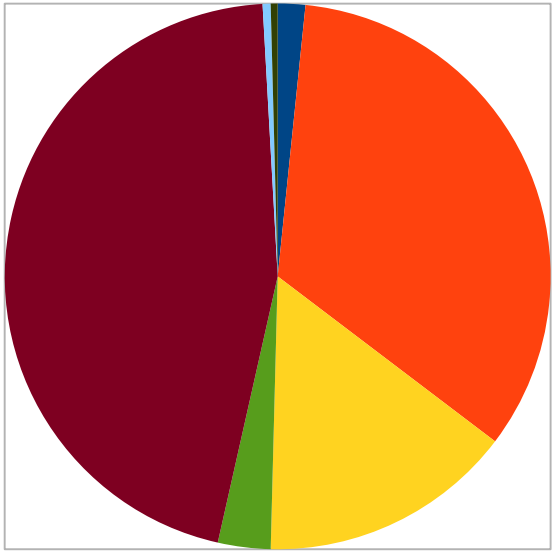
Tactics and Issues Over Time

- Structural Topic Models (structured on year)
  - Used R package stm (Roberts, Stewart, and Tingly)
- Further grouped the 40 topics into 7 *topic categories*
- Python NLTK, extracted verbs/verb phrases
- Hand identified tactics, created dictionaries of tactical categories

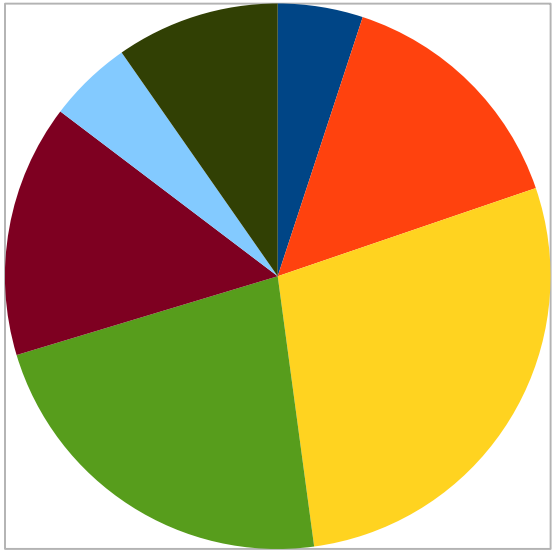
Percent Total Words



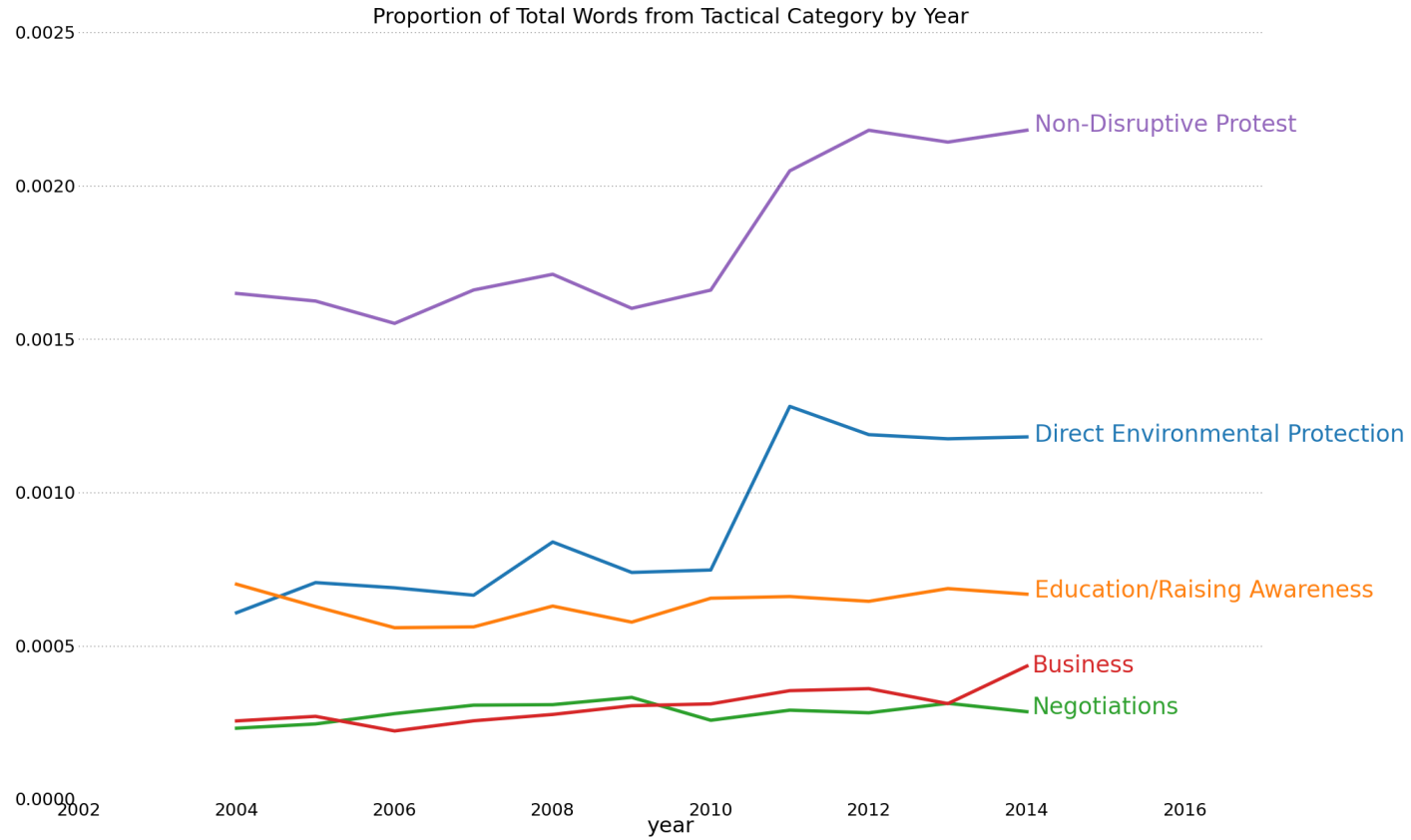
Percent Total Words in 2000



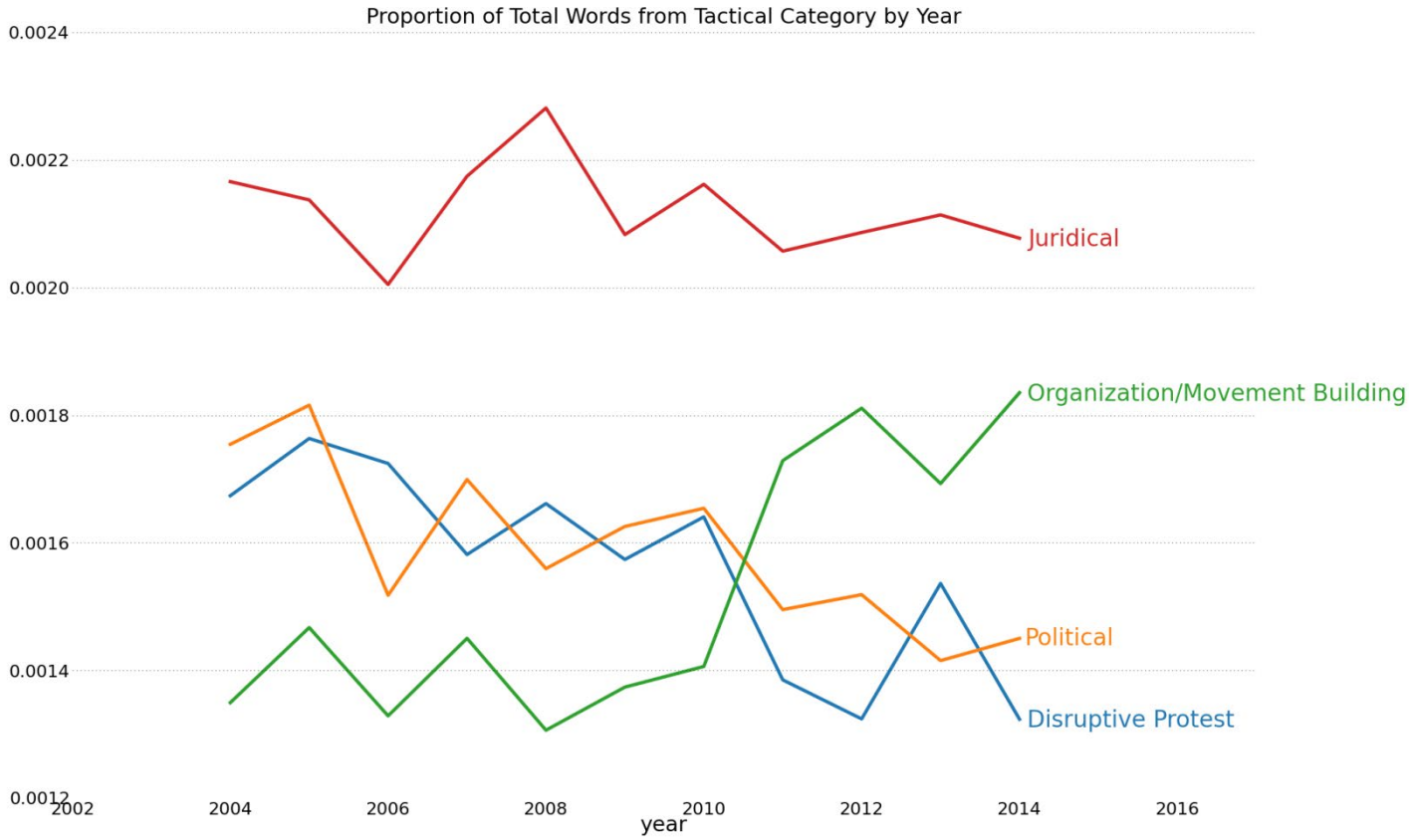
Percent Total Words in 2012



Tactics by Year



Tactics by Year



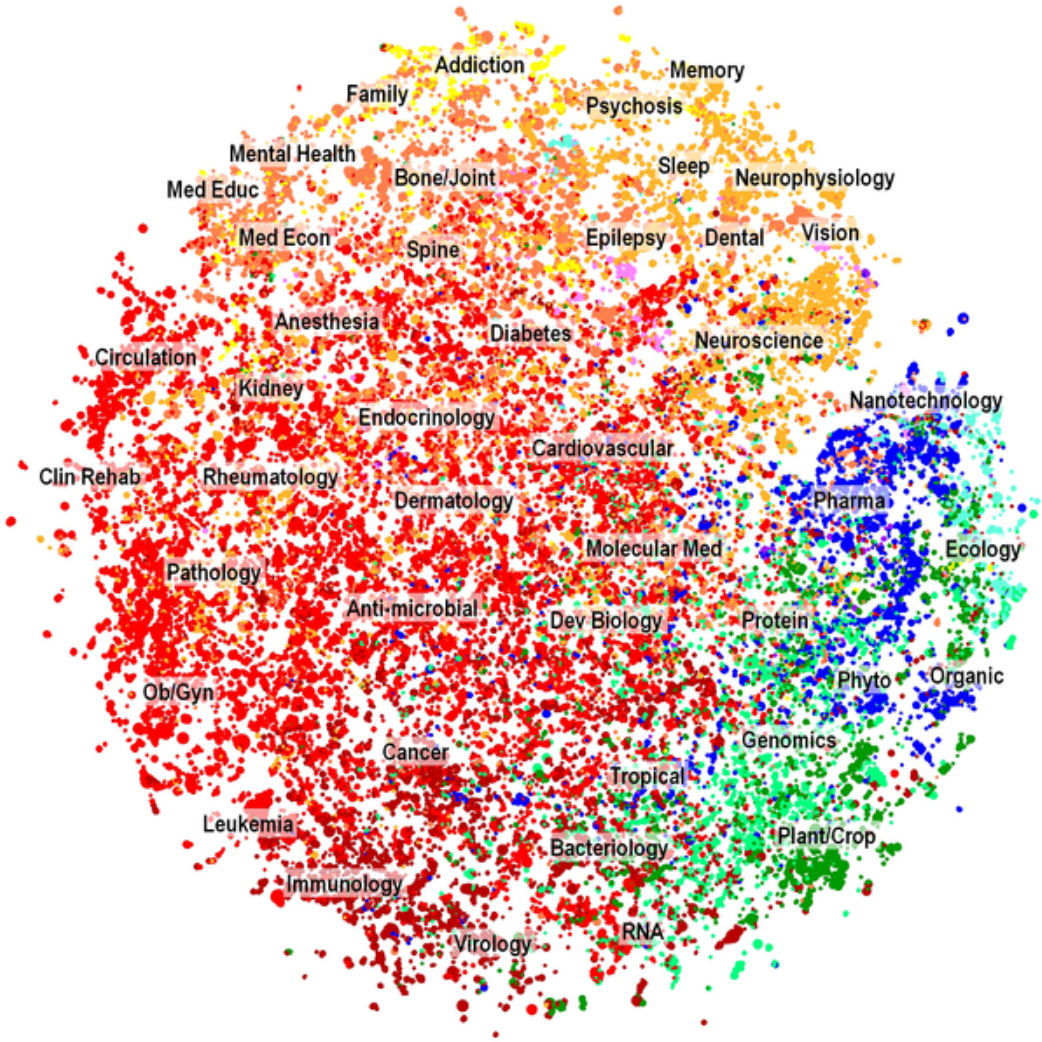

BUSH
BYE BYE BUSH


SUBARU AND THE ENVIRONMENT:
a Partnership
FOR A
Healthier Planet

Conclusion:

- Research design is key! Good data is critical!
- Match your method to your question and data. Be purposeful, not trendy
- Use multiple methods, including qualitative, to verify the analysis
- Learn a programming language
  - Off-the-shelf tools box you in (see point 2).
  - I recommend Python, R is also good
- Read NLP and machine learning literature

Happy text analyzing!



Laura K. Nelson

laura.nelson@kellogg.northwestern.edu

Tactical Categories*

Direct Environmental Protection: build, improve, protect, recycle

Non-Disruptive Protest: chant, demonstrate, organize, petition, protest

Disruptive Protest: blockade, chain, prevent, damage, sabotage

Political: campaign, donate, elect, endorse, regulate

Juridical: audit, enforce, inspect, represent, testify

Verbal Statements: advocate, comment, criticize, explain, refute

Business: boycott, buy, invest, purchase, sponsor

Education/Raising Awareness: editorial, outreach, publish, report, tweet

Organization/Movement Building: fund-raise, initiate, launch, participate

Negotiations: deal, discuss, engage, listen, persuade

**Categories are not mutually exclusive*