Automated Text Analysis in R

D-Lab INTENSIVE
Instructor: Laura Nelson
June 3-4, 2014

Theory

- Artificial Language

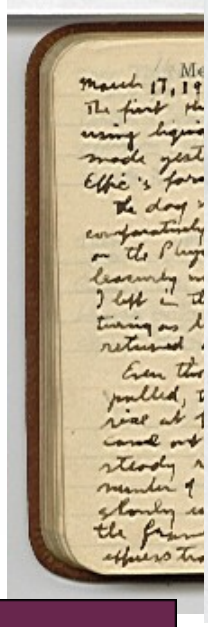
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

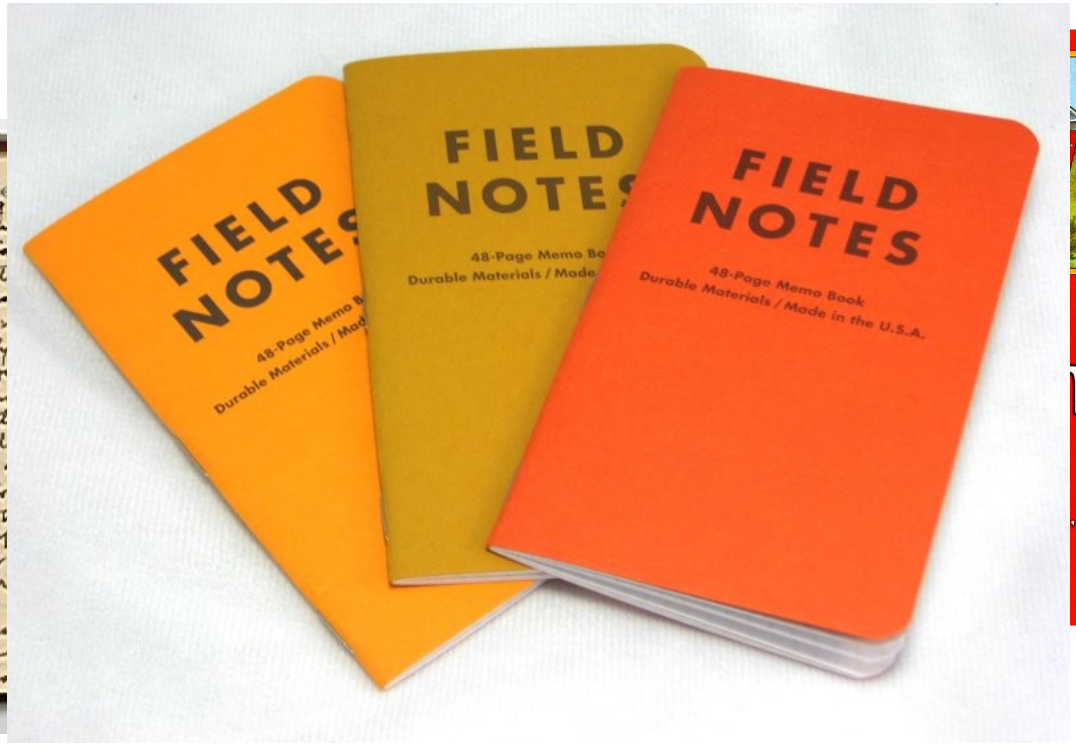
1.	$(x)(Q \supset Fx)$	
2.	$Q \supset Fx$	I,UI
3.	Q	
4.	Fx	2, 3, M.P.
5.	$(x)Fx$	4, UG
6.	$Q \supset (x)Fx$	3 – 5, C.P.
7.	$(x)(Q \supset (x)Fx) \supset [Q \supset (x)Fx]$	1 – 6, C.P.

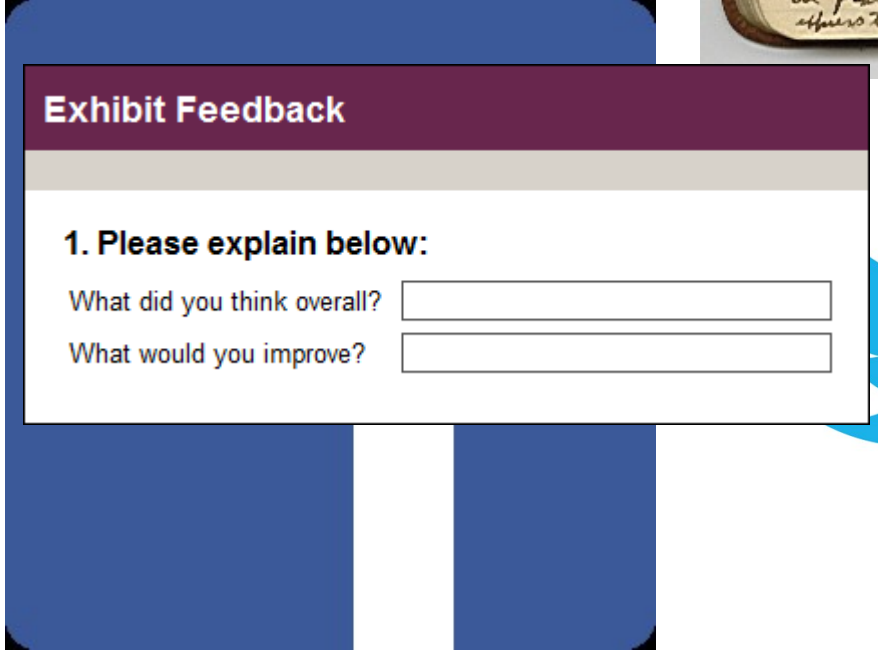
```
import scipy
from scipy import sparse

n = 200000
matrix = scipy.sparse.rand(n, n, density=.001)
print matrix
```

- Natural Language
  - “Time flies like an arrow. Fruit flies like a banana.”
- What does “cute” mean? Keen, or pretty?

Natural Language as Data

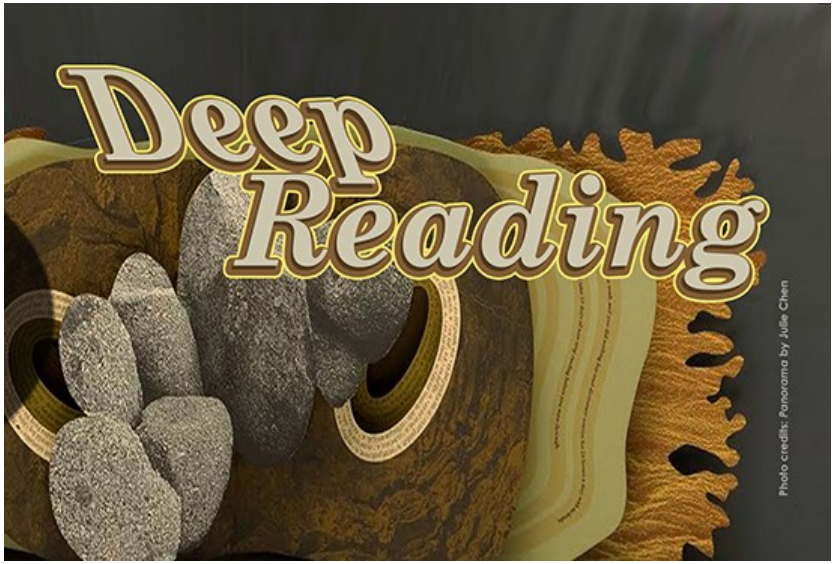






How Do We Analyze Text?

“We need to steer clear of this poverty of ambition, where people want to drive fancy cars and wear nice clothes and live in nice apartments but don't want to work hard to accomplish these things. Everyone should try to realize their full potential.” -Barack Obama



How Do We Analyze Text?

"We need to steer clear of this poverty of ambition, where people want to drive fancy cars and wear nice clothes and live in nice apartments but don't want to work hard to accomplish these things. Everyone should try to realize their full potential." -Barack Obama

How Do We Analyze Text?

"We need to steer clear of this poverty of ambition, where people want to drive fancy cars and wear nice clothes and live in nice apartments but don't want to work hard to accomplish these things. Everyone should try to realize their full potential." -Barack Obama

Why Use Computer-Assisted Methods?

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

Does not remove the need for interpretation!

Important Terms

- Corpus
  - Collection of texts/documents
- Lexical: fancy name for word
- N-gram

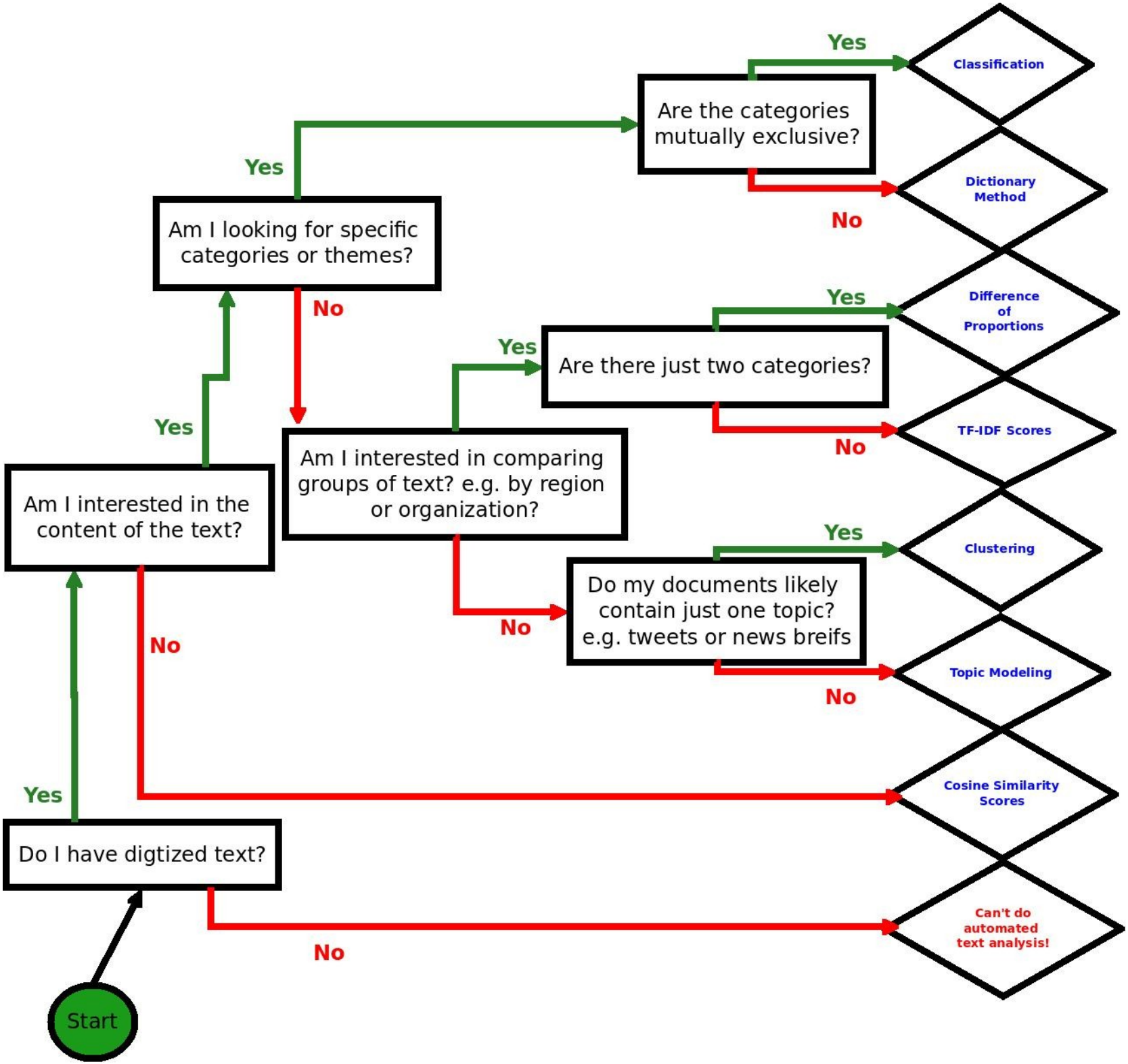
  poverty: uni-gram

  poverty of: bi-gram

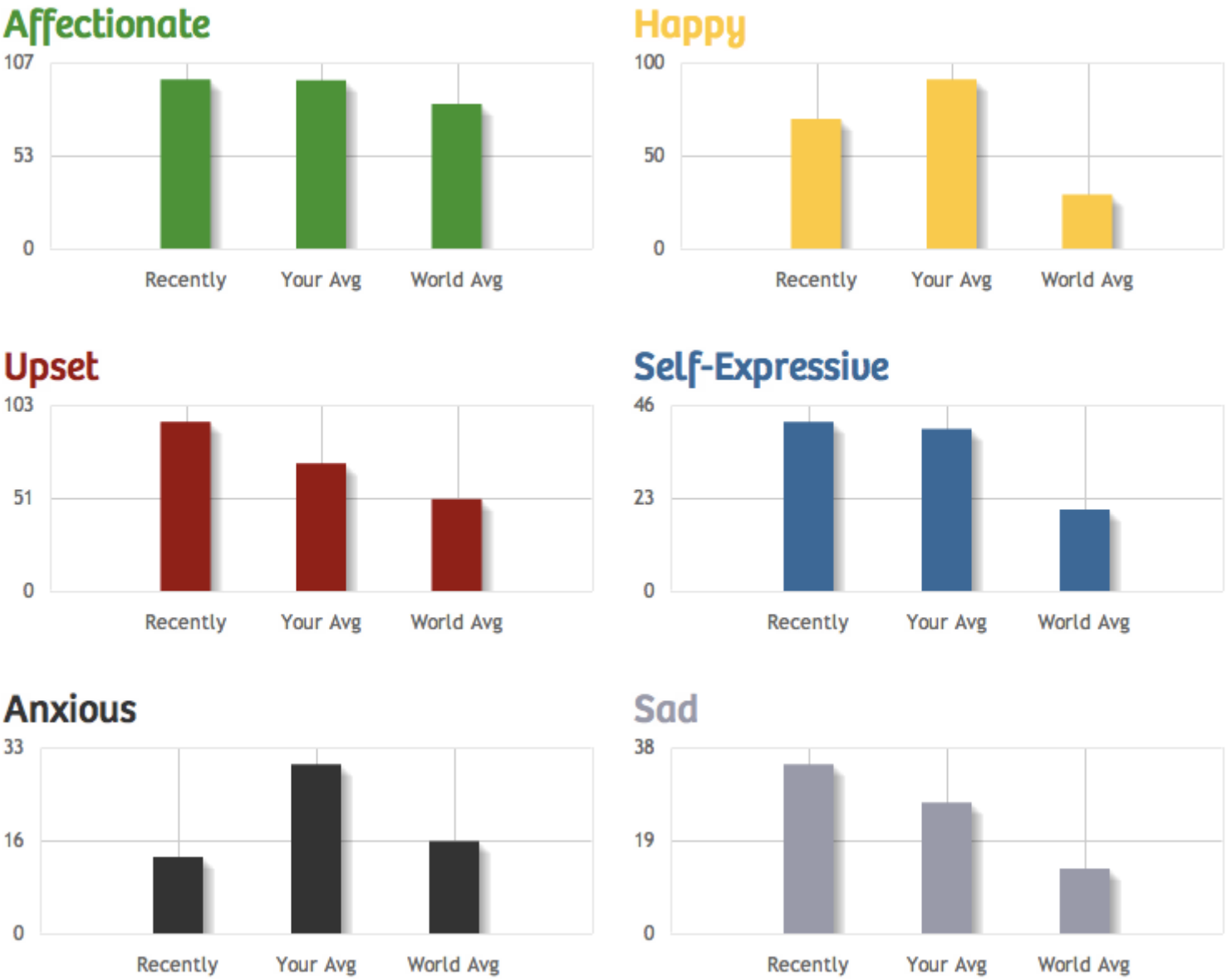
  poverty of ambition: tri-gram

Types of Automated Text Analysis

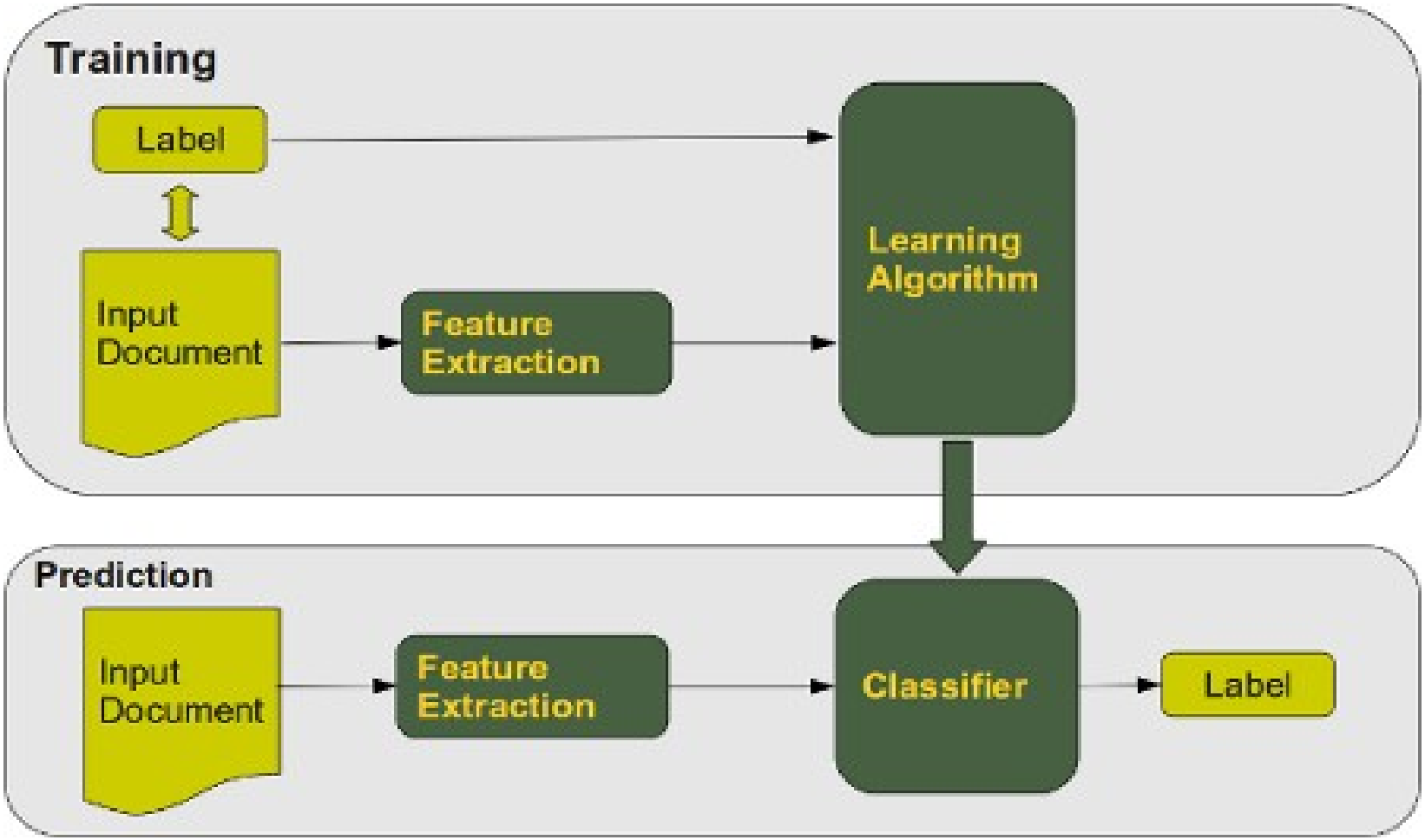
- Classification (deductive)
  - Dictionaries
  - Supervised Machine Learning
- Lexical selection (inductive)
  - Difference of proportions
  - Word scores, tf-idf
- Latent categorical analysis (inductive)
  - Clustering
  - Topic modeling


Start
Do I have digtized text?
Yes
No
Am I interested in the content of the text?
Yes
No
Am I looking for specific categories or themes?
Yes
No
Are the categories mutually exclusive?
Yes
No
Am I interested in comparing groups of text? e.g. by region or organization?
Yes
No
Are there just two categories?
Yes
No
Do my documents likely contain just one topic? e.g. tweets or news breifs
Yes
No
Classification
Dictionary Method
Difference of Proportions
TF-IDF Scores
Clustering
Topic Modeling
Cosine Similarity Scores
Can't do automated text analysis!

Classification: Dictionaries



Classification: Semi-Automated Machine Learning



Classification: Semi-Automated Machine Learning

Label	Document Text	Function
Anti-materialist	Document 1 Text	Training
Pro-hard-work	Document 2 Text	Training
other	Document 3 Text	Training
Pro-hard-work	Document 4 Text	Test
Pro-hard-work	Document 5 Text	Test
Anti-materialist	Document 6 Text	Test
?	Document 7 Text	Unknown
?	Document 8 Text	Unknown
?	Document 9 Text	Unknown
?	Document 10 Text	Unknown

Example: R

github.com/lknelson/D-Lab--Text-Analysis-Workshop

Analyzing Output

$$\text{precision} = \frac{\text{true positives}}{\text{true positives + false positives}}$$

*for each class, how many assigned to that class were actually in that class?

$$\text{recall} = \frac{\text{true positives}}{\text{true positives + false negatives}}$$

*for each class, what percentage of the documents actually in that class assigned to that class?

Analyzing Output

F score = harmonic mean of precision and recall

$$= \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} * 2$$

****Tests accuracy

Problem of over-fitting: the model will perform poorly on unseen data.

Solution: Cross Validation

Recap

- The goal: analyze natural languages, effectively *and* efficiently
- The approach: Utilize computational tools to speed up the process, make it more reliable, and possibly more valid
- First method, used for deductive analysis: Classification
- Today:
  - Lexical Selection
  - Latent Categorical Analysis, or Automated Machine Learning

What is Machine Learning?

- Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed."
- Yesterday: computers *learn* how to recognize categories given by you (supervised machine learning, or *prediction*)
- Today: computers *learn* what categories arise from the text itself (unsupervised machine learning, or *discovery*)

Automated Inductive Text Analysis: The Goal

Informative
Groups of Words

Pre-Processing

“We need to steer clear of this povertyof ambition, where people want to drive fancy cars and wear nice clothes and live in nice apartments but don't want to work hard to accomplish these things. Everyone should try to realize their full potential.”

Pre-Processing

we need steer clear poverti ambit where peopl want drive fanci car wear nice cloth live nice apart want work hard accomplish thing everyon should tri realiz full potenti

Pre-Processing

potenti apart
need steer
tri should
we ambit
wear full
nice

Document-Term Matrix

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

Lexical Selection: Difference of Proportions

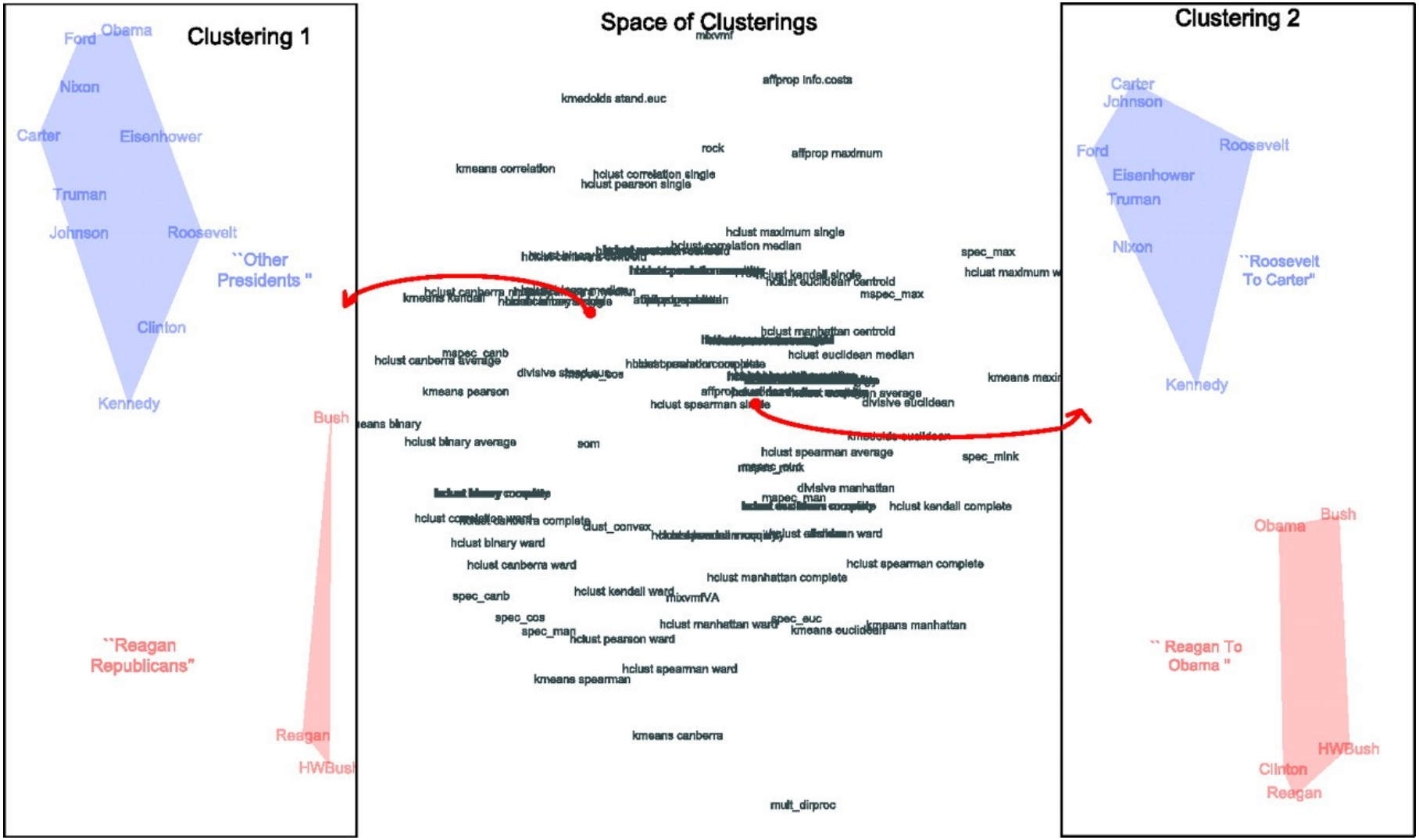
	ambit	poverti	peopl	full	*Total*
Document1	.57	.29	0	.14	*7*
Document2	.09	.27	.64	0	*11*
Diff of Prop	.48	.02	-.64	.14	

Document1: ambit
Document2: peopl

Example: R

github.com/lknelson/D-Lab--Text-Analysis-Workshop

Latent Categorical Analysis



Clustering

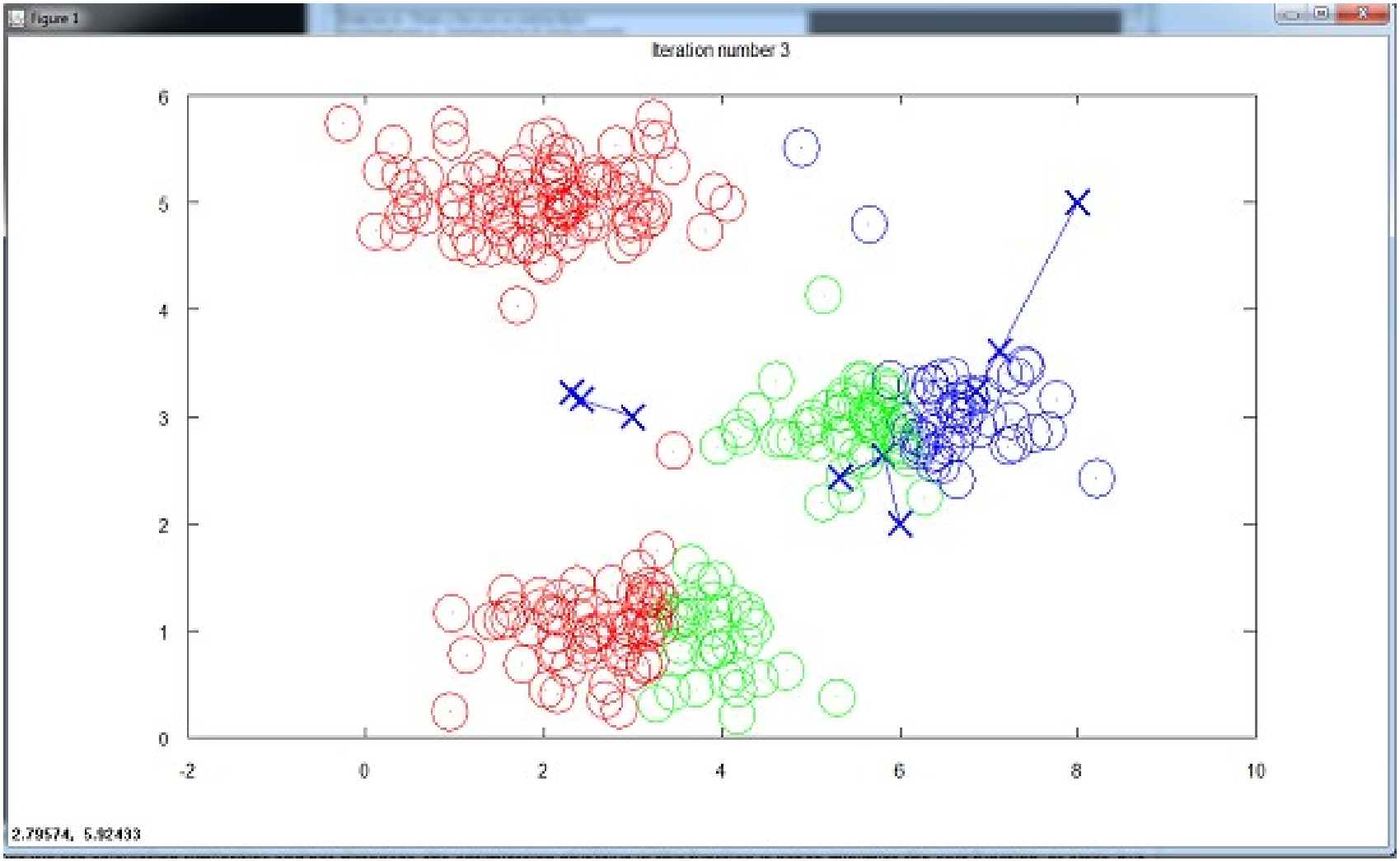


Figure 1
Iteration number 3
6
5
4
3
2
1
0
-2
0
2
4
6
8
10
2.79574, 5.92433

Topic Modeling

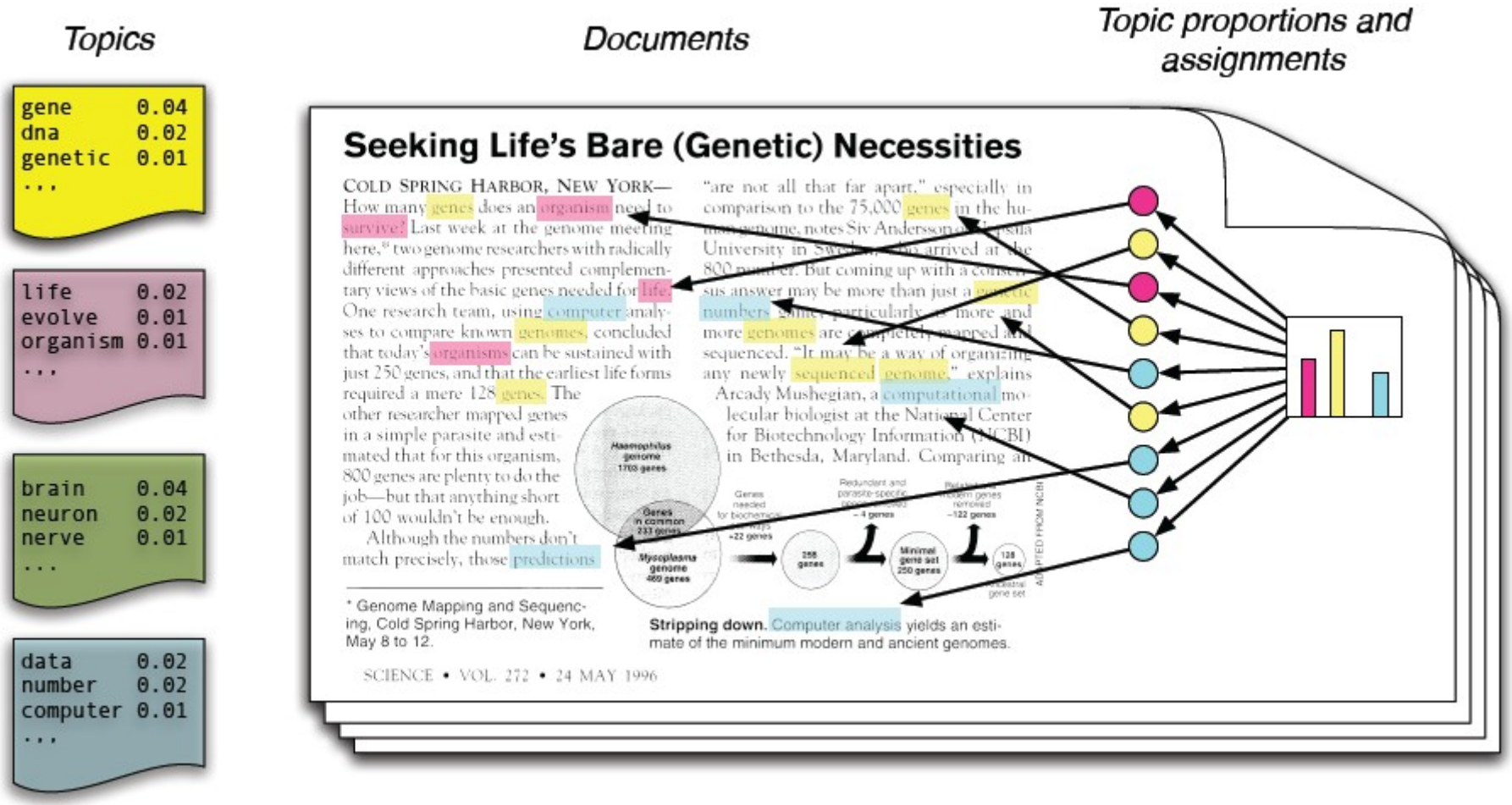
It does:

- Allow categories to arise inductively
- Find latent categories
- Find patterns across text
- Handle large and diverse corpora
- Find key differences between categories

It does not:

- Find the “one” best way to categorize text
- Capture the categories *you* want
- Tell you who does what to whom
- *Magically* reveal meaning

Latent Dirichlet Allocation



Latent Dirichlet Allocation

Each *topic* is a distribution over ALL words.

Topic1	topic1_weight	Topic2	topic2_weight	Topic3	topic3_weight
gene	0.5	genetic	0.4	dna	0.7
dna	0.3	dna	0.2	genetic	0.2
genetic	0.2	gene	0.2	gene	0.1
Sum	**1.0**		**1.0**		**1.0**

Each *document* is a distribution over ALL topics.

Doc1	doc1_weight	Doc2	doc2_weight	Doc3	doc3_weight
Topic1	0.6	Topic2	0.8	Topic3	0.5
Topic2	0.3	Topic3	0.1	Topic1	0.3
Topic3	0.1	Topic1	0.1	Topic2	0.2
Sum	**1.0**		**1.0**		**1.0**

Example: R

github.com/lknelson/D-Lab--Text-Analysis-Workshop

Example

- Four women's organizations in two cities, looking at difference between cities.
- Method: automated text analysis--LDA and difference of proportions
- Interested in differences between organizations, within cities, and over time.
- Combined methods to come to conclusion.

Results: Difference of Proportions

CWLU	DoP
chicago	5.31
children	4.59
center	4.34
union	3.61
school	3.48
abort	3.19
nixon	2.93
day	2.86
vietnam	2.57
people	2.50
city	2.44
hospital	2.38
cwlu	2.37

Redstockings	DoP
movement	12.54
women	11.34
feminist	8.91
radical	8.56
liberation	7.69
political	5.81
history	5.68
feminine	3.85
male	3.52
left	2.96
revolution	2.58
consciousnessraising	2.45
oppress	2.41

Results: Difference of Proportions

CWLU	DoP
chicago	5.31
children	4.59
center	4.34
union	3.61
school	3.48
[illegible]	3.19
[illegible]	2.93
day	2.86
vietnam	2.57
people	2.50
city	2.44
hospital	2.38
cwlu	2.37

Abstract

Redstockings	DoP
movement	12.54
women	11.34
feminist	8.91
radical	8.56
liberation	7.69
political	5.81
history	5.68
feminine	3.85
male	3.52
left	2.96
revolution	2.58
consciousnessraising	2.45
oppress	2.41

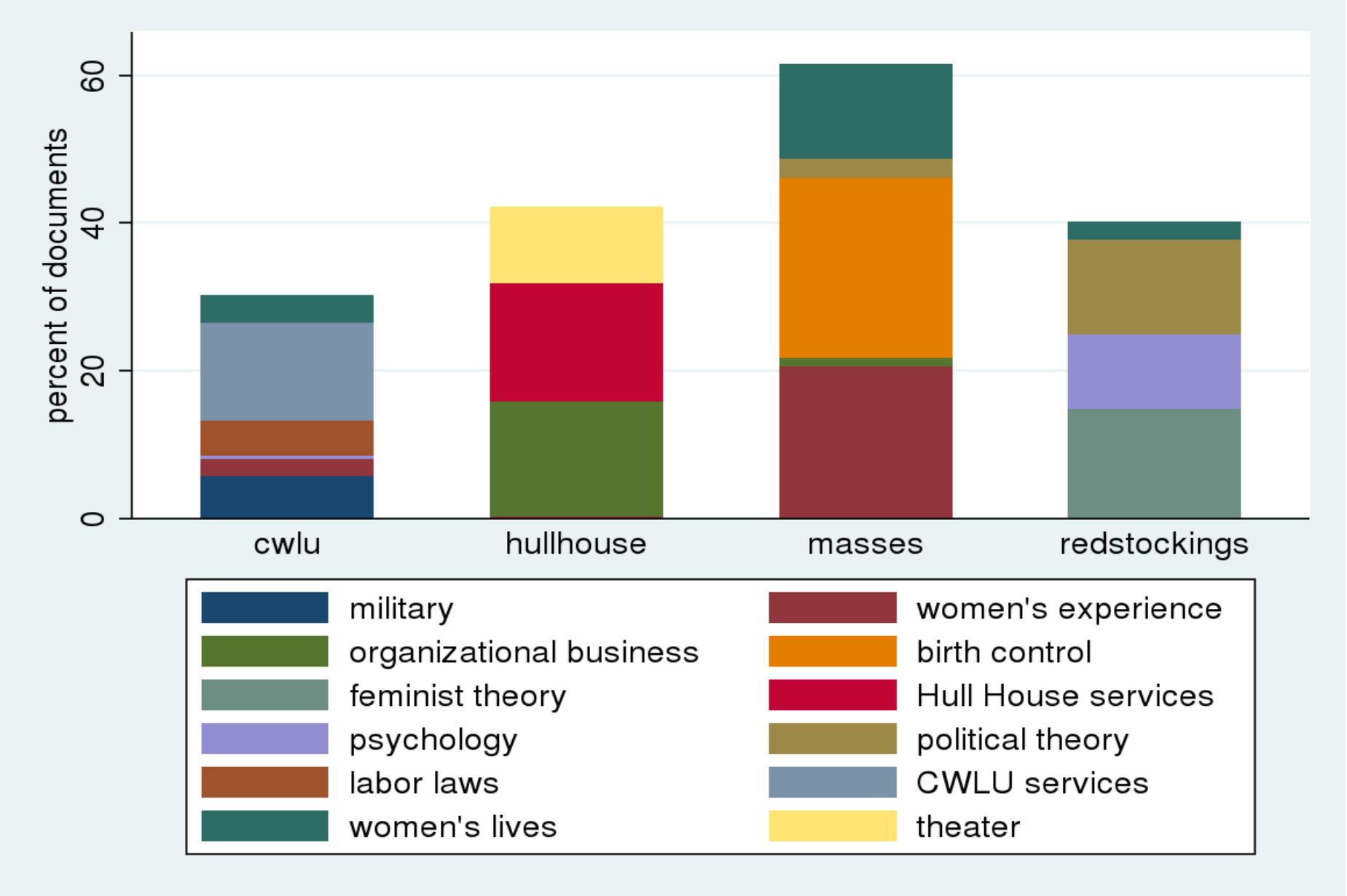
Results: Difference of Proportions

CWLU	DoP
chicago	5.31
children	4.59
center	4.34
union	3.61
school	3.48
abort	3.19
nixon	2.93
day	2.86
vietnam	2.57
people	2.50
city	2.44
hospital	2.38
cwlu	2.37

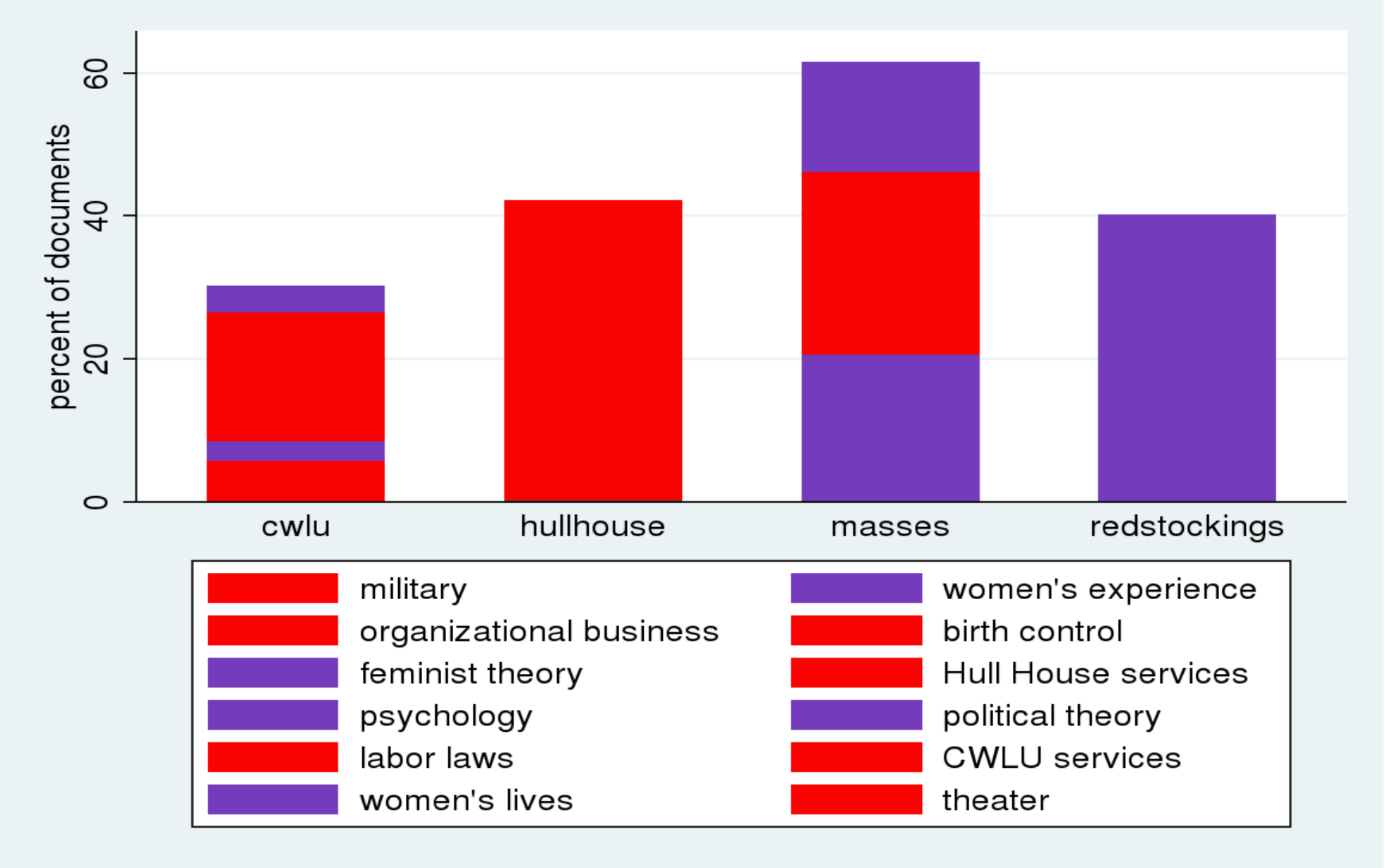
Redstockings	DoP
movement	12.54
women	11.34
feminist	8.91
radical	8.56
liberation	7.69
political	
history	
feminine	3.85
male	3.52
left	2.96
revolution	2.58
consciousnessraising	2.45
oppress	2.41

Concrete

Results: LDA, Top 12 Topics



Results: LDA, Aggregated Topics



Qualitative Check

Women attorneys in private practice who are members of the Chicago Women's Liberation Union legal clinic will counsel at the YWCA-Loop Center on everything from domestic relations to criminal law. Among the most frequently asked questions at these legal clinics, said one of the attorneys, are those concerning a woman's rights in marriage, ownership rights, property rights, rights in business, and labor union problems.
---*Womankind,* 1973

It was 1969 when she became pregnant again. This time she wanted the baby. The birth in the middle of the night at a public hospital ward affected her badly. And when the doctor finally arrived at 9:30 a.m., the local anesthesia had worn off so that she was stitched up without it. When she got over the exhaustion following all this, she had to go to work in a mens clothing factory (which today is under investigation for unsanitary working conditions).
---*Feminist Revolution*, 1979

Summary of Findings

	Quantitative Results	Qualitative Analysis	Political Logic
New York City	Words and categories that are abstract and general dominate.	The authors used stories and narratives to make generalizable claims about the social world that are applicable to all women, abstracting these stories to claims about social structure.	Social change happens through individuals Strategy: Change individual consciousness
Chicago	Words and categories that are concrete and particular dominate.	The documents outline each organization's attempts to identify concrete needs of the community and their efforts to take practical steps toward meeting those needs.	Social change happens through institutions Strategy: short-term goals winning concrete changes

Conclusion

- Computers can be helpful, but...
- Still requires qualitative choices
  - Stem words? Remove stop words? Which stop words? Remove frequent/infrequent words? How many clusters? How many topics?
- They will not do the interpretive work for you!
- Choose methods based on your question
- Combine different methods
- Validate, validate, validate!