Marijuana through the lens of the New York Times*

Peter Rasmussen
May 31, 2016

*Excluding editorials and other opinion pieces

Summary

Context: The legality of and public’s view towards marijuana is rapidly changing as more states decriminalize and legalize the drug

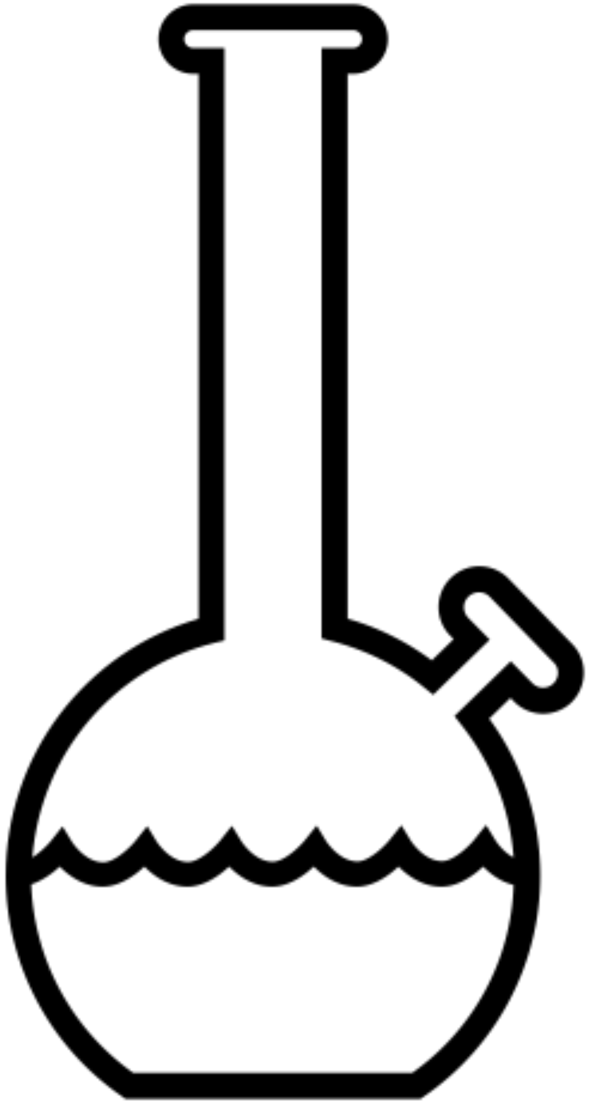
Research question: As such, how have the words associated with the topic of marijuana in news articles changed over time?

Objectives:

1. Identify distinct eras characterized by use of key words assoc. w/ marijuana
2. Assess whether words in each era represent larger themes regarding public’s view towards marijuana and drug’s legality

Conclusion: At a high level, articles’ focus has shifted from enforcement and crime to legalization and recreational use

Data pipeline



Source and acquisition

- Headline & article lead paragraphs from New York Times API
- Keyword search for ”marijuana” yielded ~5,600 articles from 1926 to 2016 and a total of 1.1 million n-grams

Assimilation, cleaning, and pre-processing

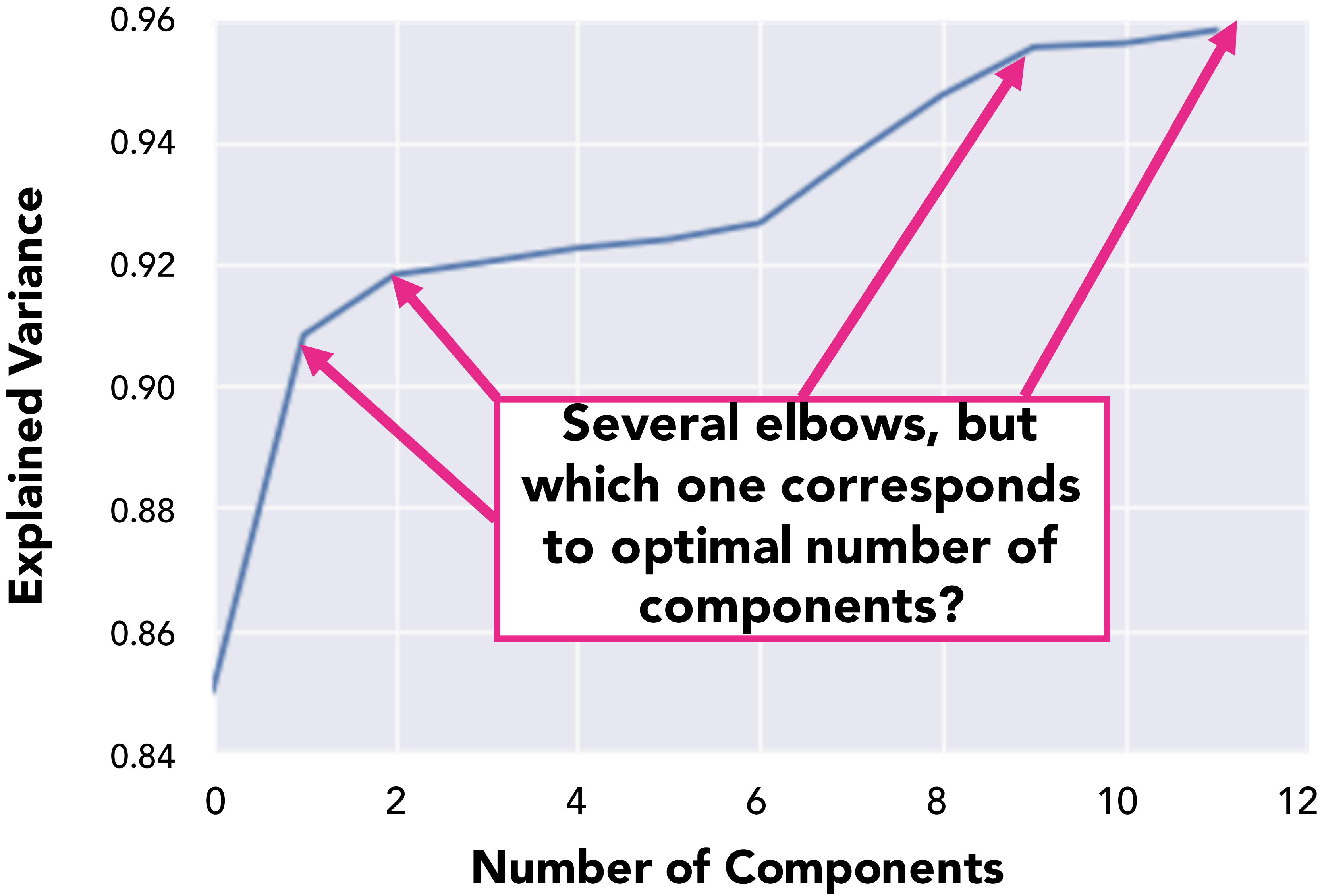
- Combine each headline and article and remove empty set articles, remove articles before 1935
- Group articles into 5-year chunks of text, lowercase words, rid punctuation, stem words, remove stop-words, n-grams of 4

Analysis

- Perform TF-IDF for each 5-year chunk across corpus
- Reduce dimensionality via PCA, select optimal # of components
- Generate cosine similarity matrix and visualize clusters
- Identify key words in each cluster based on TF-IDF results and compare cluster words to trends in public opinion and historical events

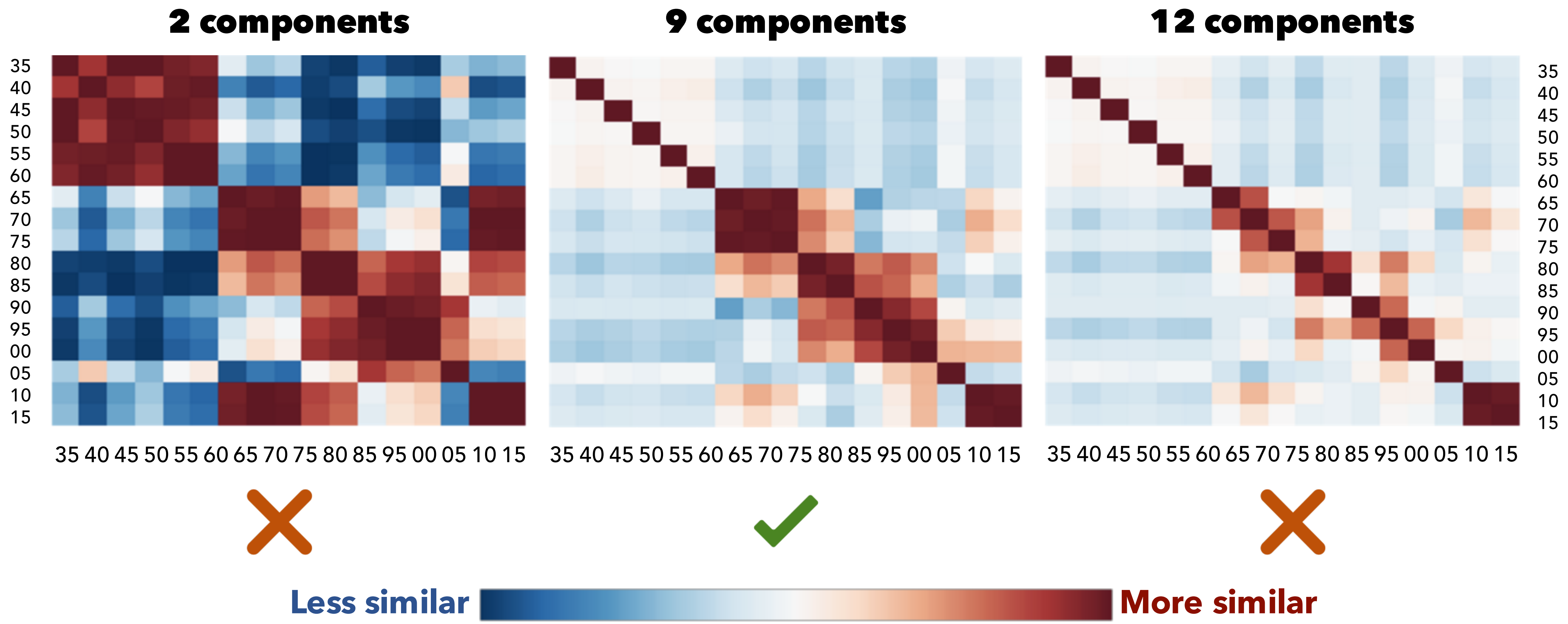
*Bong icon created by Cole M Johnstone from Noun project

Analysis: PCA results



Several elbows, but which one corresponds to optimal number of components?

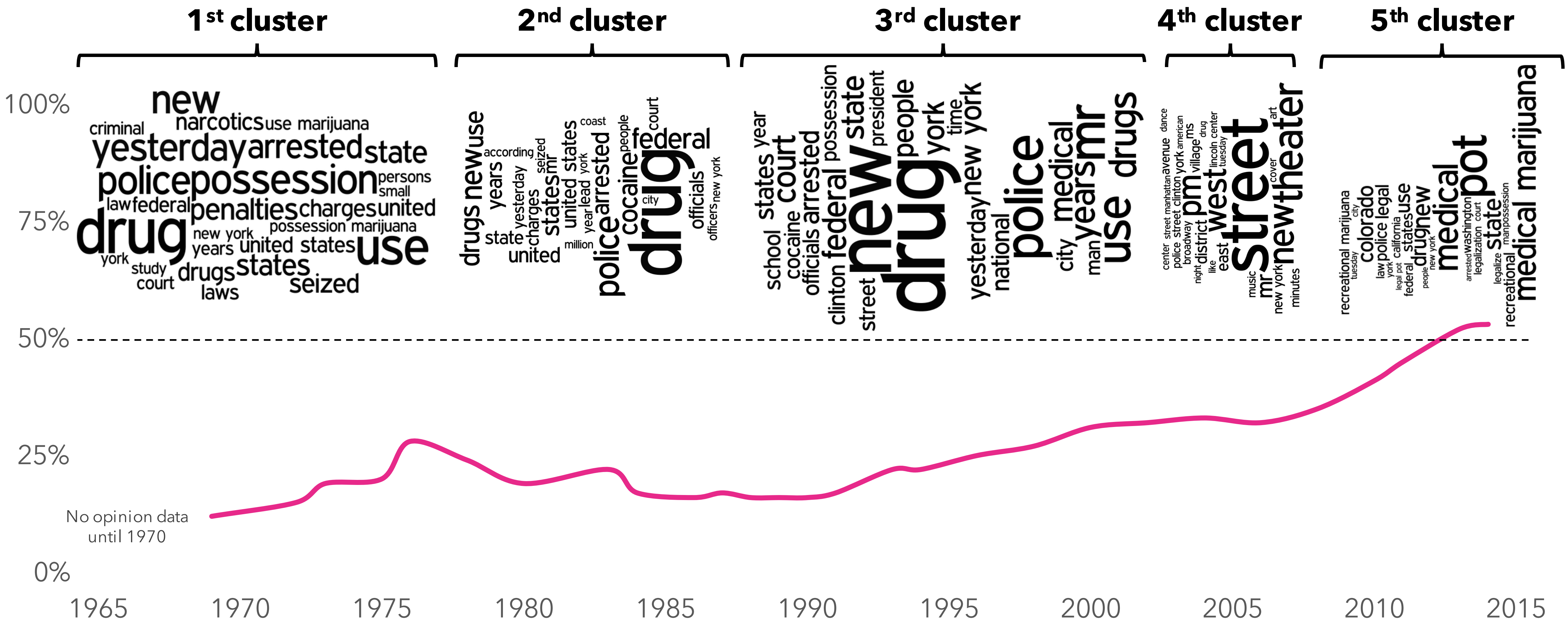
Analysis: Cosine similarity matrix



*Check icon created by Nathan Thomson from Noun Project, dismiss icon created by Brian Dys Sahagun from Noun Project

Analysis: Mapping key words to public opinion on marijuana

% of Americans who think marijuana should be legalized*



*Recreation of Pew Research Center chart that used data from Gallup, General Social Survey and Pew Research Center

Conclusions

- Identified five distinct clusters collectively spanning from 1965 to 2016 using this pre-processing and clustering approach
- Words associated with marijuana for most of the 20th century related to crime, enforcement, and prevention
- Then, stories on marijuana shifted toward decriminalization, legalization, and recreational marijuana use as public support for legalization increased
- Further investigation needed to better understand why key words circa 2005 did not seem to pertain to usual themes associated with marijuana