Text as Data: Statistical Text Analysis for the Social Sciences, Discovery

Prof Molly Roberts

UC San Diego

September 7, 2022

Discovery

- Assumption: we arrive at data an exact hypothesis and test in mind.

Discovery

- Assumption: we arrive at data an exact hypothesis and test in mind.
- Sometimes: we don't know what is in the data.

Discovery

- Assumption: we arrive at data an exact hypothesis and test in mind.
- Sometimes: we don't know what is in the data.
- We want to discover new concepts, organizations of the data.

Discovery

- Assumption: we arrive at data an exact hypothesis and test in mind.
- Sometimes: we don't know what is in the data.
- We want to discover new concepts, organizations of the data.
- Conceptualization: defining new concepts

Discovery

- Assumption: we arrive at data an exact hypothesis and test in mind.
- Sometimes: we don't know what is in the data.
- We want to discover new concepts, organizations of the data.
- Conceptualization: defining new concepts
- Content analysis: discovering concepts and categorizing texts

Example: Conceptualizing U.S. Congress

- Scholars of American political institutions: study Congress

Example: Conceptualizing U.S. Congress

- Scholars of American political institutions: study Congress
- In order to do that, have to use concepts

Example: Conceptualizing U.S. Congress

- Scholars of American political institutions: study Congress
- In order to do that, have to use concepts
  - Geography

Example: Conceptualizing U.S. Congress

- Scholars of American political institutions: study Congress
- In order to do that, have to use concepts
  - Geography
  - Ideology

Example: Conceptualizing U.S. Congress

- Scholars of American political institutions: study Congress
- In order to do that, have to use concepts
  - Geography
  - Ideology
  - Senate vs. House

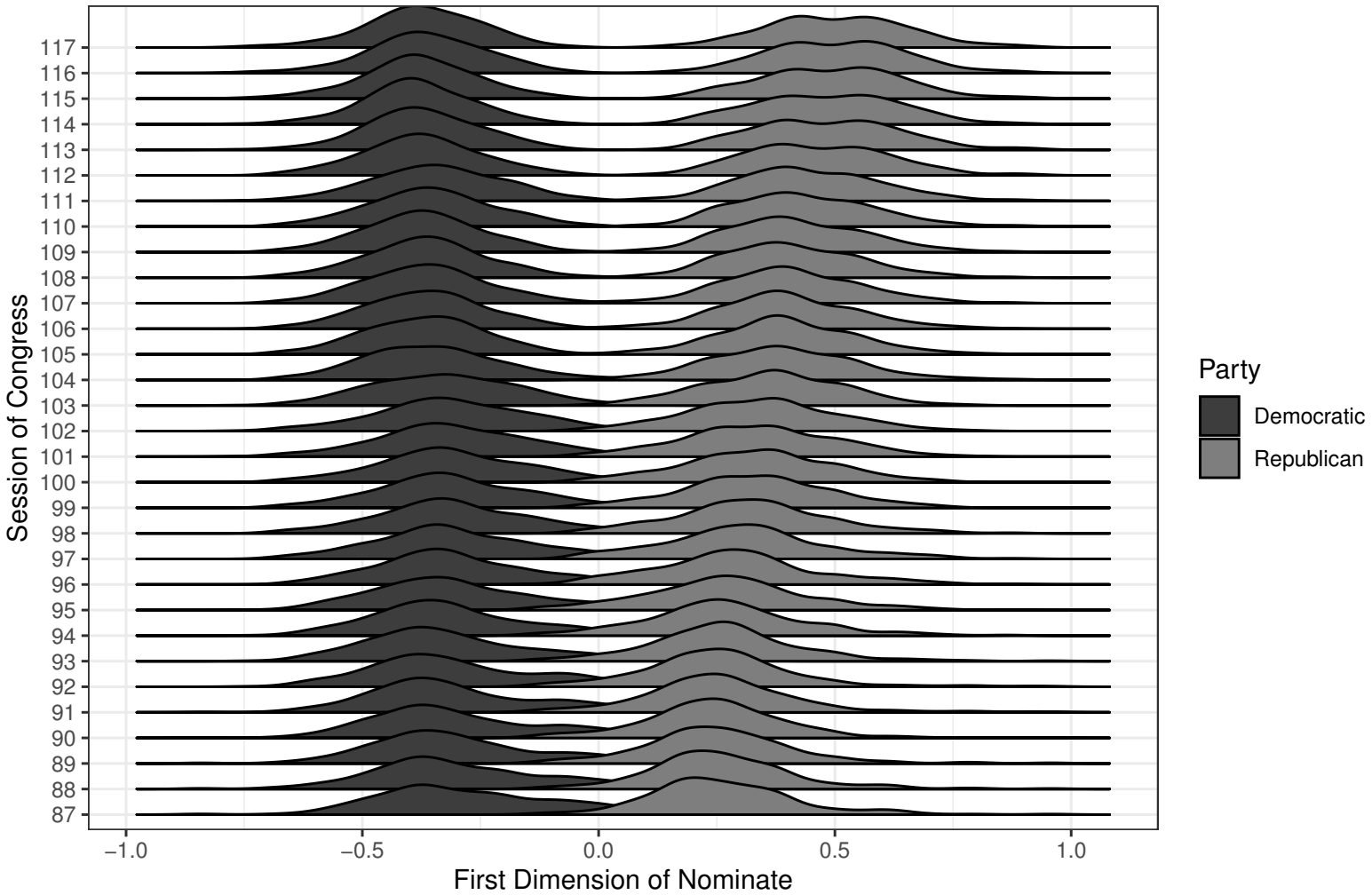
Example: Conceptualizing U.S. Congress

- Scholars of American political institutions: study Congress
- In order to do that, have to use concepts
  - Geography
  - Ideology
  - Senate vs. House
  - Demographics

Example: Conceptualizing U.S. Congress

- Scholars of American political institutions: study Congress
- In order to do that, have to use concepts
  - Geography
  - Ideology
  - Senate vs. House
  - Demographics
- Example: DW-Nominate

Example: Conceptualizing U.S. Congress



Is this the "right" conceptualization?

Example: Computational Grounded Theory (Nelson 2017)

- Grounded theory (Glaser and Strauss 1999, 2005)

Example: Computational Grounded Theory (Nelson 2017)

- Grounded theory (Glaser and Strauss 1999, 2005)
  - identify categories and themes inductively

Example: Computational Grounded Theory (Nelson 2017)

- Grounded theory (Glaser and Strauss 1999, 2005)
  - identify categories and themes inductively
  - theory-producing, hypothesis-generating

Example: Computational Grounded Theory (Nelson 2017)

- Grounded theory (Glaser and Strauss 1999, 2005)
  - identify categories and themes inductively
  - theory-producing, hypothesis-generating
- Problem: requires lots of reading, many judgement calls, difficult to reproduce

Example: Computational Grounded Theory (Nelson 2017)

- Computational Grounded Theory

Example: Computational Grounded Theory (Nelson 2017)

- Computational Grounded Theory
  - Use computational methods to help explore a corpus

Example: Computational Grounded Theory (Nelson 2017)

- Computational Grounded Theory
  - Use computational methods to help explore a corpus
  - Combine computational and interpretative approaches.

Example: Computational Grounded Theory (Nelson 2017)

- Computational Grounded Theory
  - Use computational methods to help explore a corpus
  - Combine computational and interpretative approaches.
- Advantages

Example: Computational Grounded Theory (Nelson 2017)

- Computational Grounded Theory
  - Use computational methods to help explore a corpus
  - Combine computational and interpretative approaches.
- Advantages
  - Efficiency

Example: Computational Grounded Theory (Nelson 2017)

- Computational Grounded Theory
  - Use computational methods to help explore a corpus
  - Combine computational and interpretative approaches.
- Advantages
  - Efficiency
  - Reproducability

Proposed Steps

1. Pattern detection with computational exploratory analysis
2. Hypothesis refinement using human-centered interpretation
3. Pattern confirmation

Example: Computational Grounded Theory (Nelson 2017)

- First and second wave feminist movements viewed as distinct

Example: Computational Grounded Theory (Nelson 2017)

- First and second wave feminist movements viewed as distinct
- But they had similar debates, why?

Example: Computational Grounded Theory (Nelson 2017)

- First and second wave feminist movements viewed as distinct
- But they had similar debates, why?
- Nelson: Geography mattered.

Example: Computational Grounded Theory (Nelson 2017)

- First and second wave feminist movements viewed as distinct
- But they had similar debates, why?
- Nelson: Geography mattered.
  - Chicago vs. New York

Example: Computational Grounded Theory (Nelson 2017)

- First and second wave feminist movements viewed as distinct
- But they had similar debates, why?
- Nelson: Geography mattered.
  - Chicago vs. New York
  - Different philosophies

Example: Computational Grounded Theory (Nelson 2017)

Table 2. Most Distinctive Words, Difference of Proportions.

First Wave[a]		Second Wave[b]	
Hull House (Chicago)	Heterodoxy (New York City)	CWLU (Chicago)	Redstockings (New York City)
hullhouse	woman	chicago	movement
club	man	children	women
miss	women	center	men
school	life	union	radical
given	know	school	feminist
year	world	work	male
members	like	cwlu	political
chicago	sanger	vietnam	history
mr	men	nixon	womens
classes	said	people	feminism
house	home	office	revolution
boys	just	day	love
work	say	health	feminists
years	don't	city	left
social	little	working	power
held	way	vietnamese	oppression
clubs	think	legal	class
mrs	things	war	female
residents	want	care	personal
room	sex	womankind	woman
children	right	government	really

Example: Computational Grounded Theory (Nelson 2017)

Table 5. Sample of Structured Dataframe Sorted by *Movement History* Topic Weights.

File Name	Text (Head)	Mo Histo W
nyc.redstockings.1973.sarachild.powerofhistory-28.txt	THE ARCS OF HISTORY tific and fearless writers of her day" and Elizabeth Cady Stanton, too, "the matchless writer." [. . .]	0.96
nyc.redstockings.1973.sarachild.powerofhistory-27.txt	Ten any idea of what that work was all about "it's purpose and the breadth of its contents and even its method- we [. . .]	0.96
nyc.redstockings.1973.sarachild.powerofhistory-29.txt	Paraging depiction of the History in the bibliography, it suddenly struck m_e that Stanton, Anthony and Gage's [. . .]	0.94
nyc.redstockings.1973.sarachild.powerofhistory-30.txt	Said it was. And those who did take action in the area of history "especially for the present record"did not [. . .]	0.94
nyc.redstockings.1973.sarachild.powerofhistory-26.txt	And so their absence from history books meant there weren't any. We had to discover the problem of [. . .]	0.91
nyc.redstockings.1973.sarachild.powerofhistory-18.txt	Opposite of Beauvoir's book which was disappearing from the lists. I soon learned, even without reading it, [. . .]	0.86
nyc.redstockings.1973.sarachild.powerofhistory-31.txt	POSTSCRIPT The history question had a lot to do with the leadership question. History, after all, is all about [. . .]	0.86
nyc.redstockings.1973.sarachild.powerofhistory-25.txt	Most of the theories citing history that we encountered from both the right and the left essentially counseled [. . .]	0.81
nyc.redstockings.1973.sarachild.powerofhistory-02.txt	The Power Of History I am obnoxious to each carping tongue Who says my hand a needle better fits, A poet's [. . .]	0.75
nyc.redstockings.1973.sarachild.powerofhistory-17.txt	Of the key elements of Beauvoir's analysis upon which the WLM later built its work included: 1. Women have [. . .]	0.70

Note: This is a sample of a structured dataframe (see Figures 1 and 2) structured via a 40-topic Structural Topic Model. The topic weights indicate the perce each document related to each topic (e.g., 97 percent of the total words in the first document are related to the *movement history* topic, as indicated by the Weight" column, while close to 0 are related to the *antiwar* topic). Doing a descending sort by the *movement history* topic quickly indicates which documents this topic. As the text is included in the dataframe, the researcher can quickly read these documents to better understand the content of each topic. In this pa

Example: Computational Grounded Theory (Nelson 2017)

Validation:

- Chicago organizations used more concrete words (dictionary)
- Chicago organizations named more organizations, NY more individuals (NER)

Methods of Discovery

- Discriminating Words

Methods of Discovery

- Discriminating Words
  - Known categories

Methods of Discovery

- Discriminating Words
  - Known categories
  - Discover the words associated with categories

Methods of Discovery

- Discriminating Words
  - Known categories
  - Discover the words associated with categories
  - Helpful in describing group or clusters
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)

Methods of Discovery

- Discriminating Words
  - Known categories
  - Discover the words associated with categories
  - Helpful in describing group or clusters
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)
- Topic Models (A form of clustering)

Methods of Discovery

- Discriminating Words
  - Known categories
  - Discover the words associated with categories
  - Helpful in describing group or clusters
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)
- Topic Models (A form of clustering)
- Embedding Methods

Methods of Discovery

- Discriminating Words
  - Known categories
  - Discover the words associated with categories
  - Helpful in describing group or clusters
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)
- Topic Models (A form of clustering)
- Embedding Methods
  - Placing documents or words in a low-dimensional space

Clustering

Clustering as Discovery

Clustering

Clustering as Discovery

- When we analyze text data we to organize and simplify

Clustering

Clustering as Discovery

- When we analyze text data we to organize and simplify
- How do we formulate new ways to organize texts?

Clustering

Clustering as Discovery

- When we analyze text data we to organize and simplify
- How do we formulate new ways to organize texts?
  - Clustering methods suggest new (model and data driven) ways to organize texts

Clustering

Clustering as Discovery

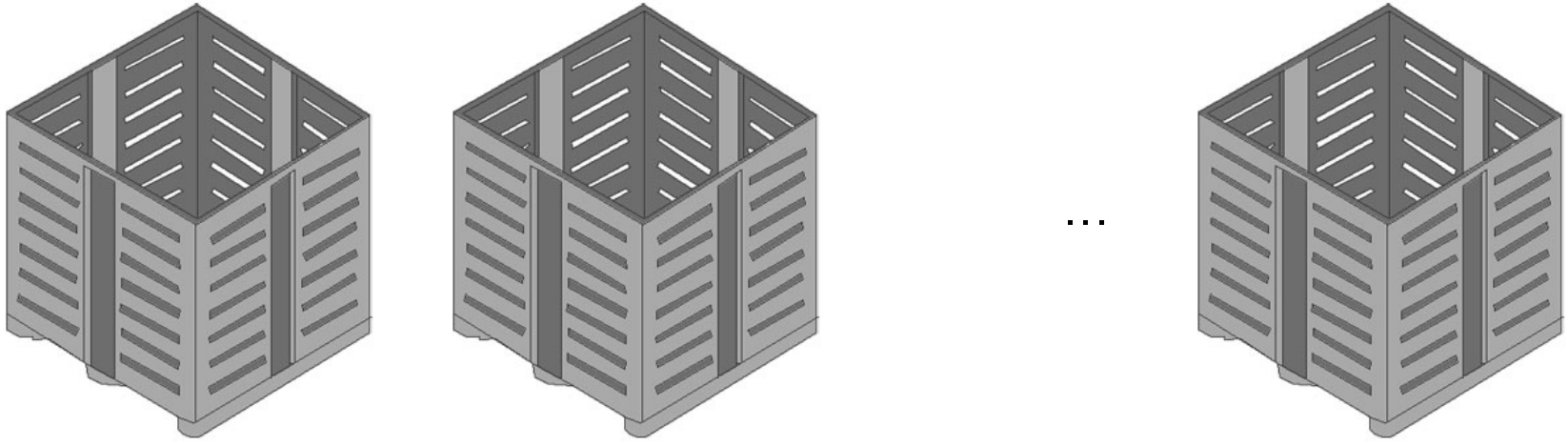
- When we analyze text data we to organize and simplify
- How do we formulate new ways to organize texts?
  - Clustering methods suggest new (model and data driven) ways to organize texts
  - Using new method, new lens to look at social interaction

Perspective 1: Clustering

Document 1 Document 2 ⋯ Document N

Perspective 1: Clustering

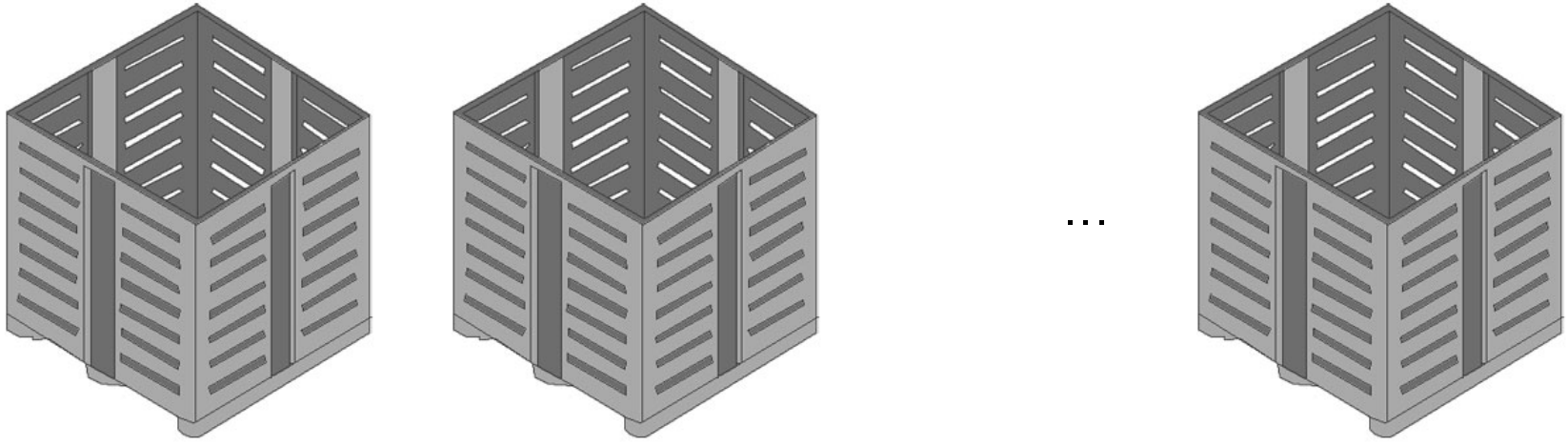
Document 1 Document 2 ⋯ Document N



Bins and Bin Assignments Estimated

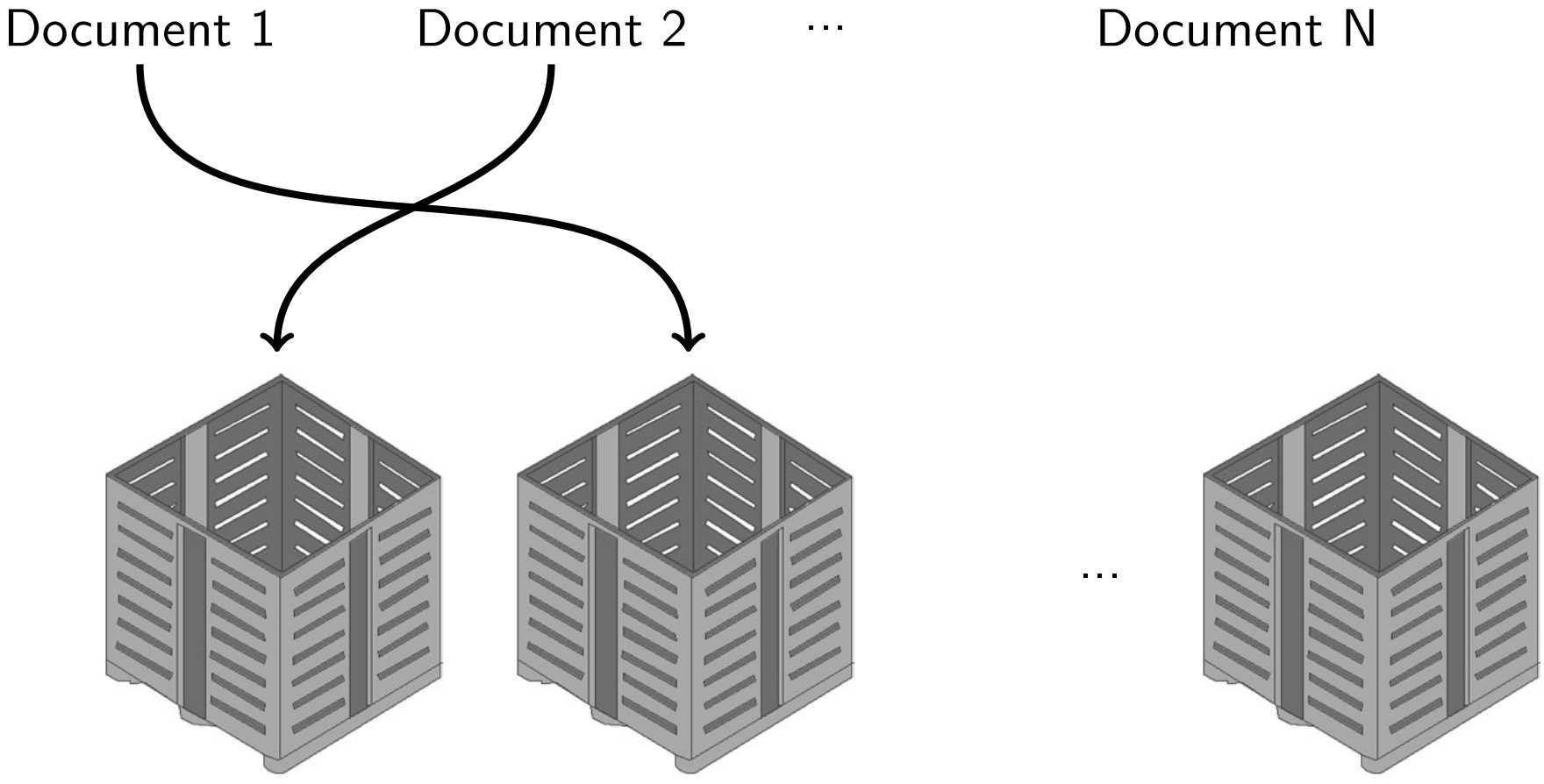
Perspective 1: Clustering

Document 1 Document 2 ⋯ Document N



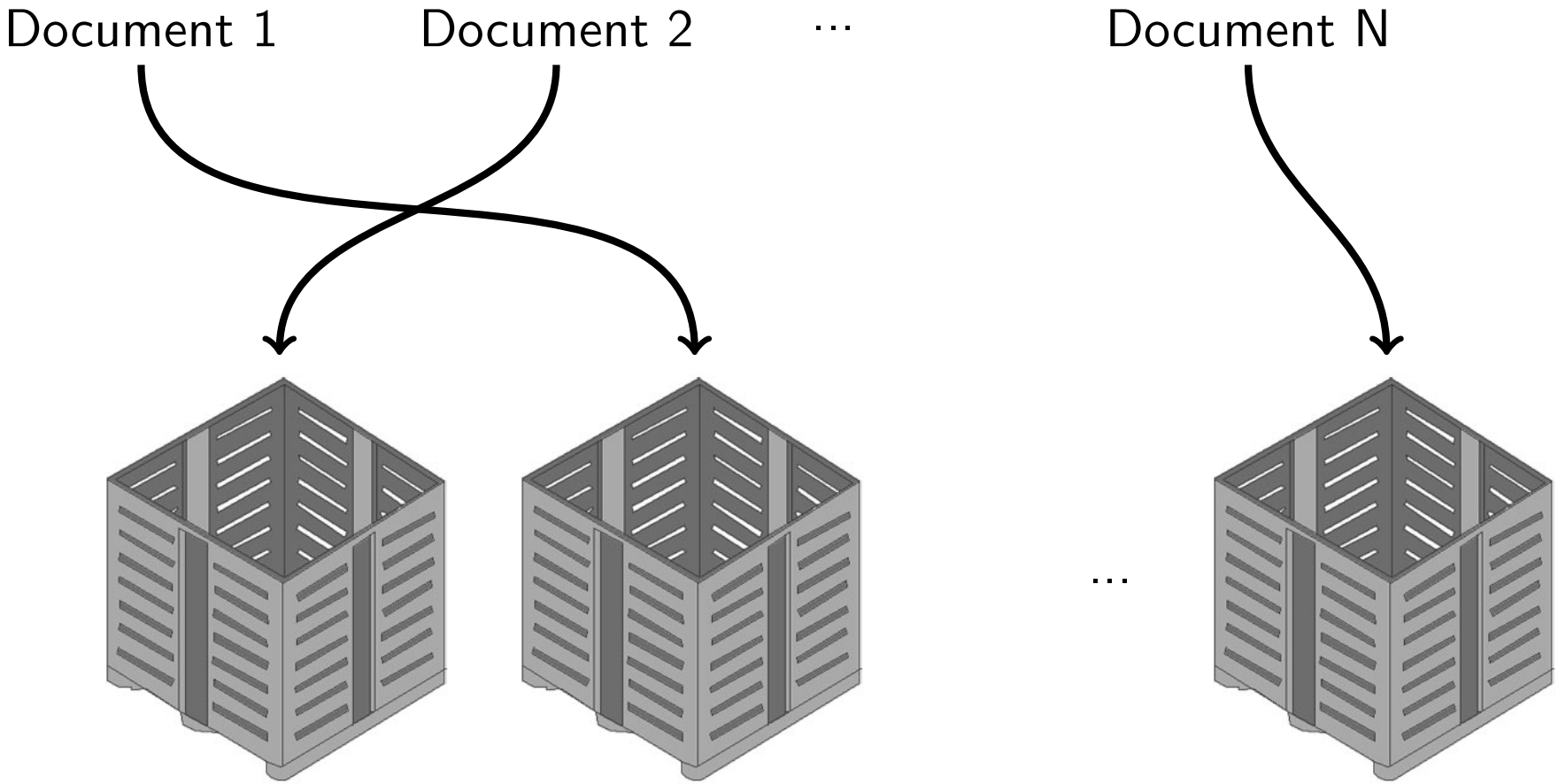
Bins and Bin Assignments Estimated

Perspective 1: Clustering



Bins and Bin Assignments Estimated

Perspective 1: Clustering



Bins and Bin Assignments Estimated

Perspective 2: Clustering

Doc1 Doc2 Doc3 Doc4 Doc5 Doc6 Doc7 Doc8 Doc9 Doc10

Perspective 2: Clustering

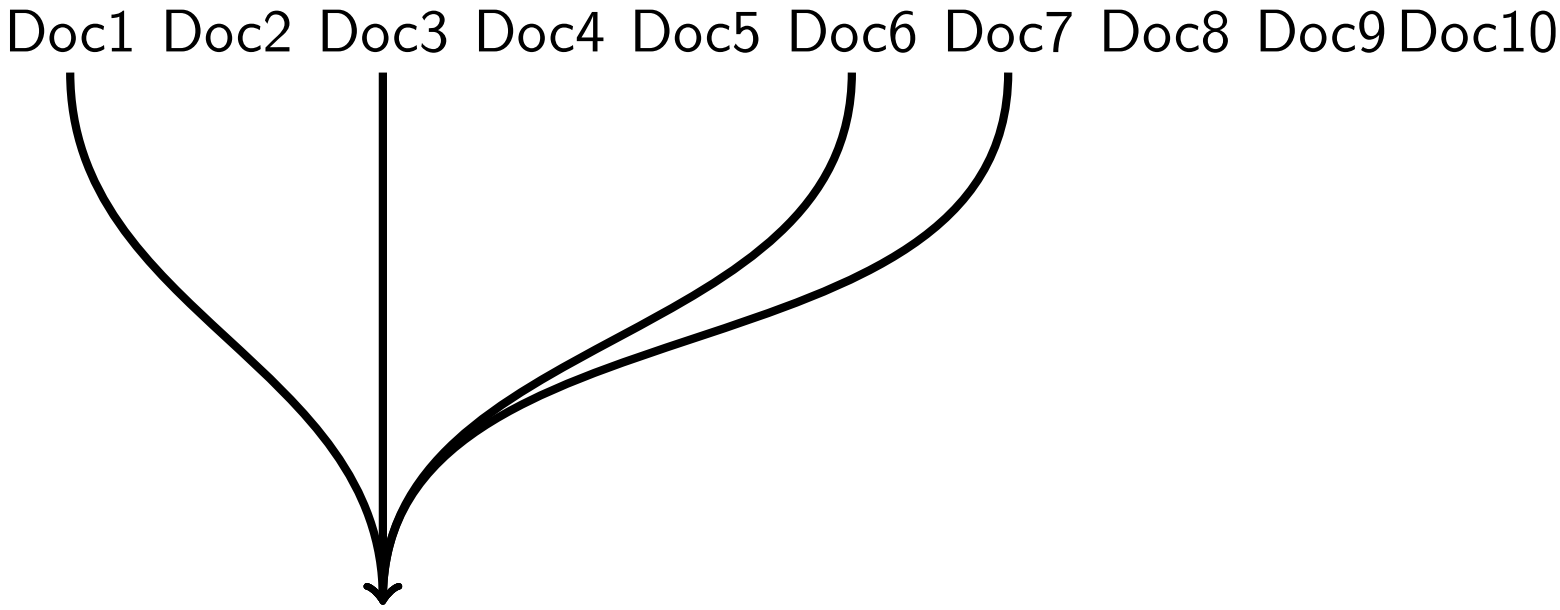
Doc1 Doc2 Doc3 Doc4 Doc5 Doc6 Doc7 Doc8 Doc9 Doc10

|Doc1, Doc3, Doc6, Doc7|

|Doc2, Doc4, Doc8|

|Doc5, Doc9, Doc10|

Perspective 2: Clustering

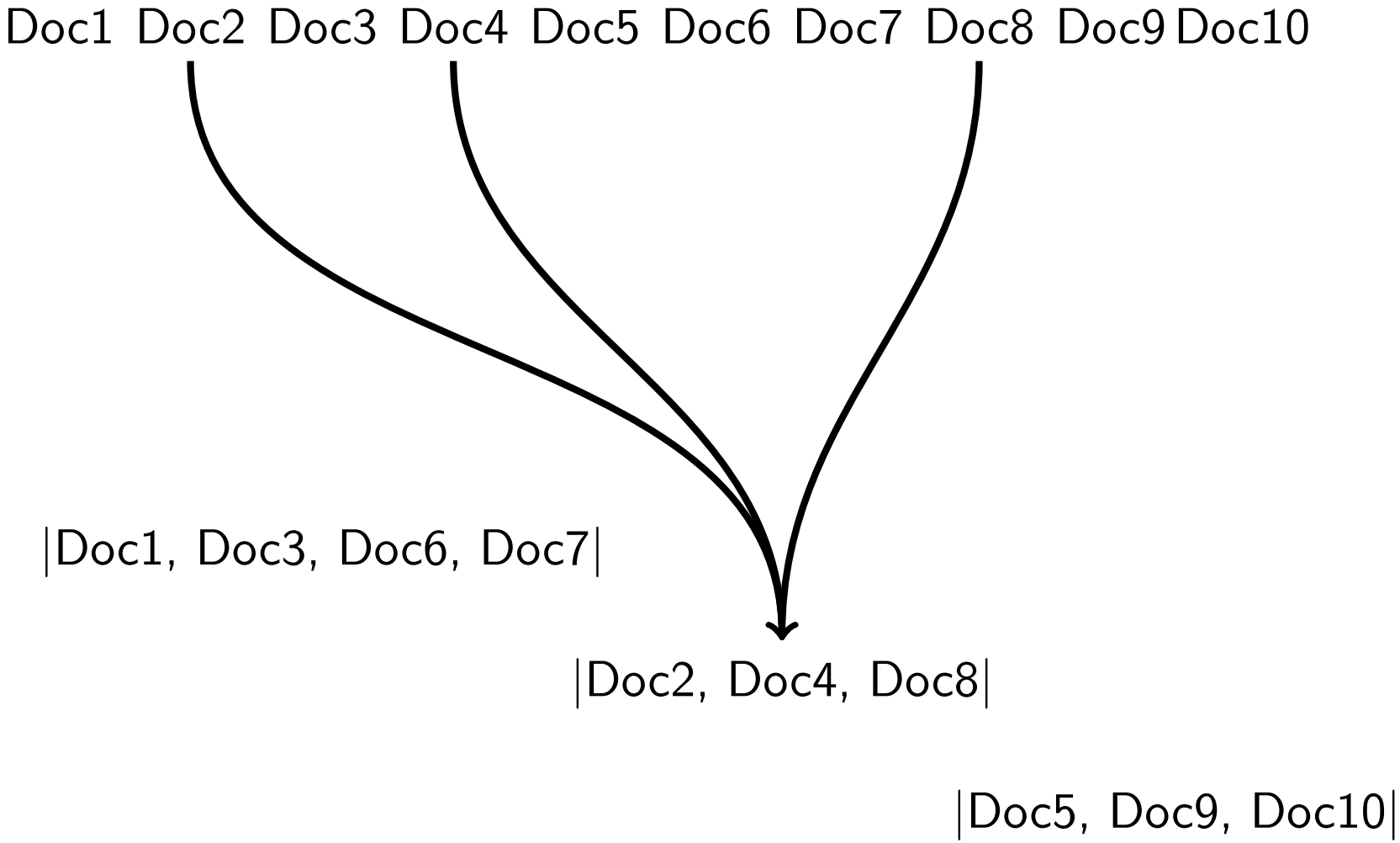


|Doc1, Doc3, Doc6, Doc7|

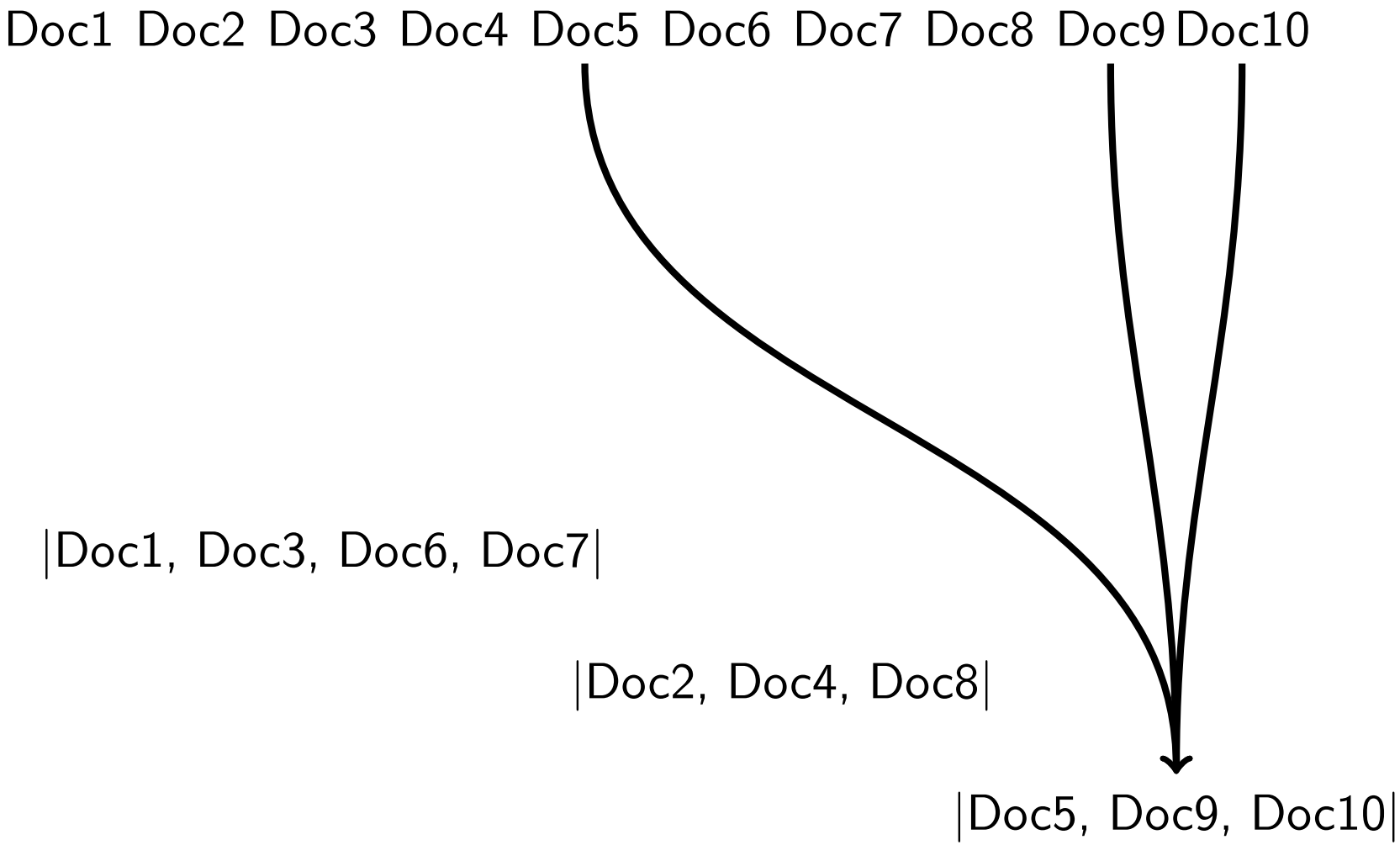
|Doc2, Doc4, Doc8|

|Doc5, Doc9, Doc10|

Perspective 2: Clustering

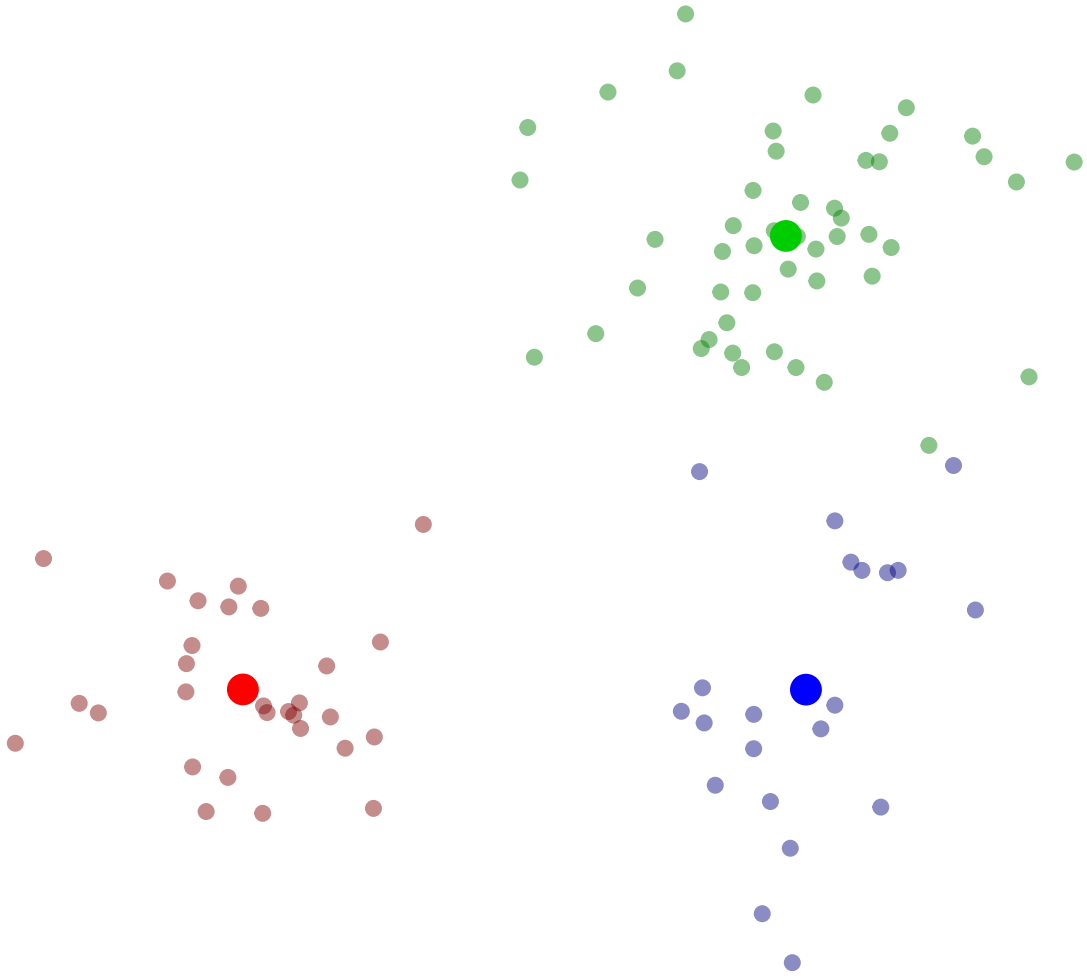


Perspective 2: Clustering



Perspective 3

Perspective 3



Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots, K$ clusters

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots, K$ clusters
Call c_i document i's cluster assignment

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots, K$ clusters
Call c_i document i's cluster assignment

- $c_i = 2 \rightsquigarrow$ Document i assigned to second cluster

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots,$ K clusters
Call c_i document i's cluster assignment

- $c_i = 2 \rightsquigarrow$ Document i assigned to second cluster
- $c_{10} = 4 \rightsquigarrow$ Document 10 assigned to fourth cluster

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots, K$ clusters
Call c_i document i's cluster assignment

- $c_i = 2 \rightsquigarrow$ Document i assigned to second cluster
- $c_{10} = 4 \rightsquigarrow$ Document 10 assigned to fourth cluster

Define clustering as a partition of observations. Mathematically:

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots, K$ clusters
Call c_i document i's cluster assignment

- $c_i = 2 \rightsquigarrow$ Document i assigned to second cluster
- $c_{10} = 4 \rightsquigarrow$ Document 10 assigned to fourth cluster

Define clustering as a partition of observations. Mathematically:

$$\boldsymbol{c} = (c_1, c_2, \ldots, c_N)$$

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots, K$ clusters
Call c_i document i's cluster assignment

- $c_i = 2 \rightsquigarrow$ Document i assigned to second cluster
- $c_{10} = 4 \rightsquigarrow$ Document 10 assigned to fourth cluster

Define clustering as a partition of observations. Mathematically:

$$\boldsymbol{c} = (c_1, c_2, \ldots, c_N)$$

$\boldsymbol{c}$ constitutes the clustering.

Clustering: Terms and Notation

Set of documents $i = 1, 2, \ldots, N$.
Partition documents into $j = 1, \ldots, K$ clusters
Call c_i document i's cluster assignment

- $c_i = 2 \rightsquigarrow$ Document i assigned to second cluster
- $c_{10} = 4 \rightsquigarrow$ Document 10 assigned to fourth cluster

Define clustering as a partition of observations. Mathematically:

$$\boldsymbol{c} = (c_1, c_2, \ldots, c_N)$$

$\boldsymbol{c}$ constitutes the clustering.
Two trivial clusterings

$$\begin{aligned} \boldsymbol{c} &= (1, 2, \ldots, N) \\ \boldsymbol{c} &= (1, 1, \ldots, 1) \end{aligned}$$

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly)

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly)
- Define objective function

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly)
- Define objective function
- Use approximate inference/optimization algorithm to identify optimal solution ← Huge search space, very difficult (and interesting!) problem, only hinted at here

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly)
- Define objective function
- Use approximate inference/optimization algorithm to identify optimal solution

An Example FAC Method

K-means: most commonly used clustering algorithm.

An Example FAC Method

K-means: most commonly used clustering algorithm.
Story: Data are grouped in K clusters and each cluster has a center or mean.

An Example FAC Method

K-means: most commonly used clustering algorithm.
Story: Data are grouped in K clusters and each cluster has a center or mean.
$\rightarrow$ Two types of parameters to estimate

An Example FAC Method

K-means: most commonly used clustering algorithm.
Story: Data are grouped in K clusters and each cluster has a center or mean.
$\rightarrow$ Two types of parameters to estimate

1) For each cluster j, $(j = 1, \ldots, K)$

   r_{ij} =Indicator, Document i assigned to cluster j

An Example FAC Method

K-means: most commonly used clustering algorithm.
Story: Data are grouped in K clusters and each cluster has a center or mean.
$\rightarrow$ Two types of parameters to estimate

1) For each cluster j, $(j = 1, \ldots, K)$

   r_{ij} =Indicator, Document i assigned to cluster j

2) For each cluster j

   $\boldsymbol{\mu}_j$ a cluster center for cluster j.

An Example FAC Method

K-means: most commonly used clustering algorithm.
Story: Data are grouped in K clusters and each cluster has a center or mean.
$\rightarrow$ Two types of parameters to estimate

1) For each cluster j, $(j = 1, \ldots, K)$

   r_{ij} =Indicator, Document i assigned to cluster j

2) For each cluster j

   $\boldsymbol{\mu}_j$ a cluster center for cluster j.

Specifying the Method

1) Assume Euclidean distance between objects.

Specifying the Method

1) Assume Euclidean distance between objects.
2) Objective function

Specifying the Method

1) Assume Euclidean distance between objects.
2) Objective function

$$f(\boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{N}\sum_{j=1}^{K} r_{ij} \left(\sum_{m=1}^{M} (w_{im} - \mu_{km})^2 \right)$$

Goal:
Choose $\boldsymbol{r}^*$ and $\boldsymbol{\mu}^*$ to minimize $f(\cdot, \cdot, \boldsymbol{w})$

Specifying the Method

1) Assume Euclidean distance between objects.
2) Objective function

$$f(\boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{N}\sum_{j=1}^{K} r_{ij} \left(\sum_{m=1}^{M} (w_{im} - \mu_{km})^2 \right)$$

Goal:
Choose $\boldsymbol{r}^*$ and $\boldsymbol{\mu}^*$ to minimize $f(\cdot, \cdot, \boldsymbol{w})$

Specifying the Method

1) Assume Euclidean distance between objects
2) Objective function

Specifying the Method

1) Assume Euclidean distance between objects
2) Objective function
3) Algorithm for optimization

Specifying the Method

1) Assume Euclidean distance between objects
2) Objective function
3) Algorithm for optimization

Iterative algorithm, Each Iteration t

Specifying the Method

1) Assume Euclidean distance between objects
2) Objective function
3) Algorithm for optimization

Iterative algorithm, Each Iteration t

- Conditional on $\boldsymbol{\mu}^{t-1}$ (from previous iteration), choose $\boldsymbol{r}^t$

Specifying the Method

1) Assume Euclidean distance between objects
2) Objective function
3) Algorithm for optimization

Iterative algorithm, Each Iteration t

- Conditional on $\boldsymbol{\mu}^{t-1}$ (from previous iteration), choose $\boldsymbol{r}^t$
- Conditional on $\boldsymbol{r}^t$, choose $\boldsymbol{\mu}^t$

Specifying the Method

1) Assume Euclidean distance between objects
2) Objective function
3) Algorithm for optimization

Iterative algorithm, Each Iteration t

- Conditional on $\boldsymbol{\mu}^{t-1}$ (from previous iteration), choose $\boldsymbol{r}^t$
- Conditional on $\boldsymbol{r}^t$, choose $\boldsymbol{\mu}^t$

Repeat until convergence, measured as change in f.

$$\text{Change} \quad = \quad f(\boldsymbol{\mu}^t, \boldsymbol{r}^t, \boldsymbol{w}) - f(\boldsymbol{\mu}^{t-1}, \boldsymbol{r}^{t-1}, \boldsymbol{w})$$

Algorithm, In Words

Algorithm, In Words

- Conditional on center estimates, assign documents to closest cluster centers

Algorithm, In Words

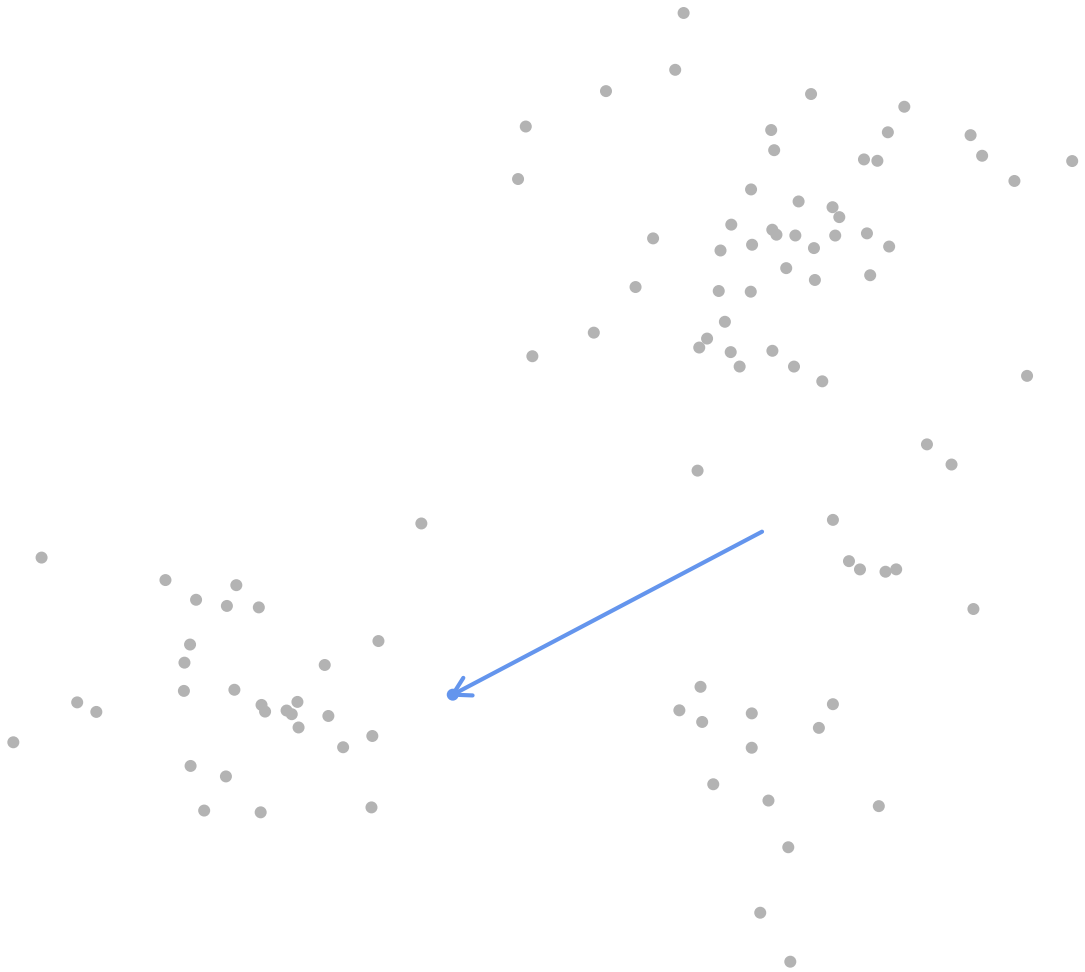
- Conditional on center estimates, assign documents to closest cluster centers
- Conditional on document assignments, cluster centers are averages of documents assigned to the cluster

Algorithm, In Words

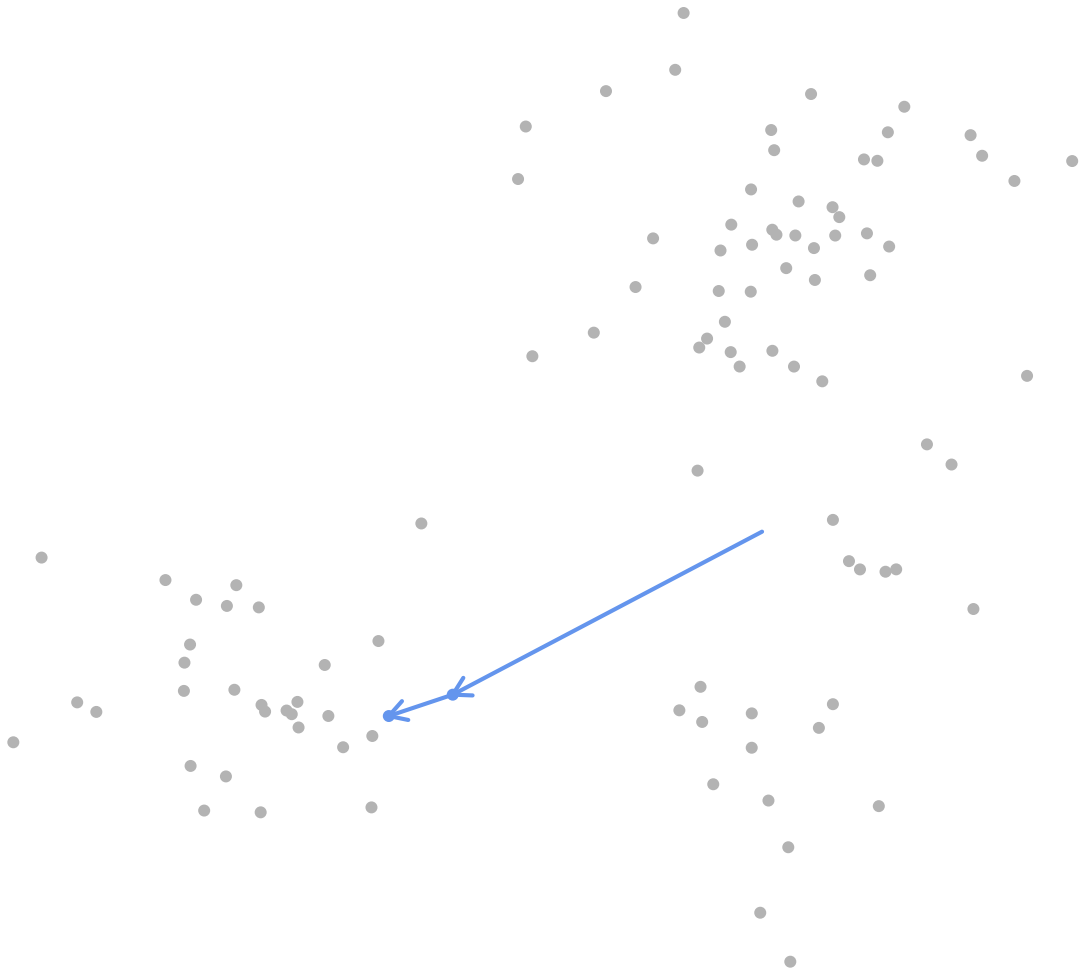
- Conditional on center estimates, assign documents to closest cluster centers
- Conditional on document assignments, cluster centers are averages of documents assigned to the cluster

Visual Example

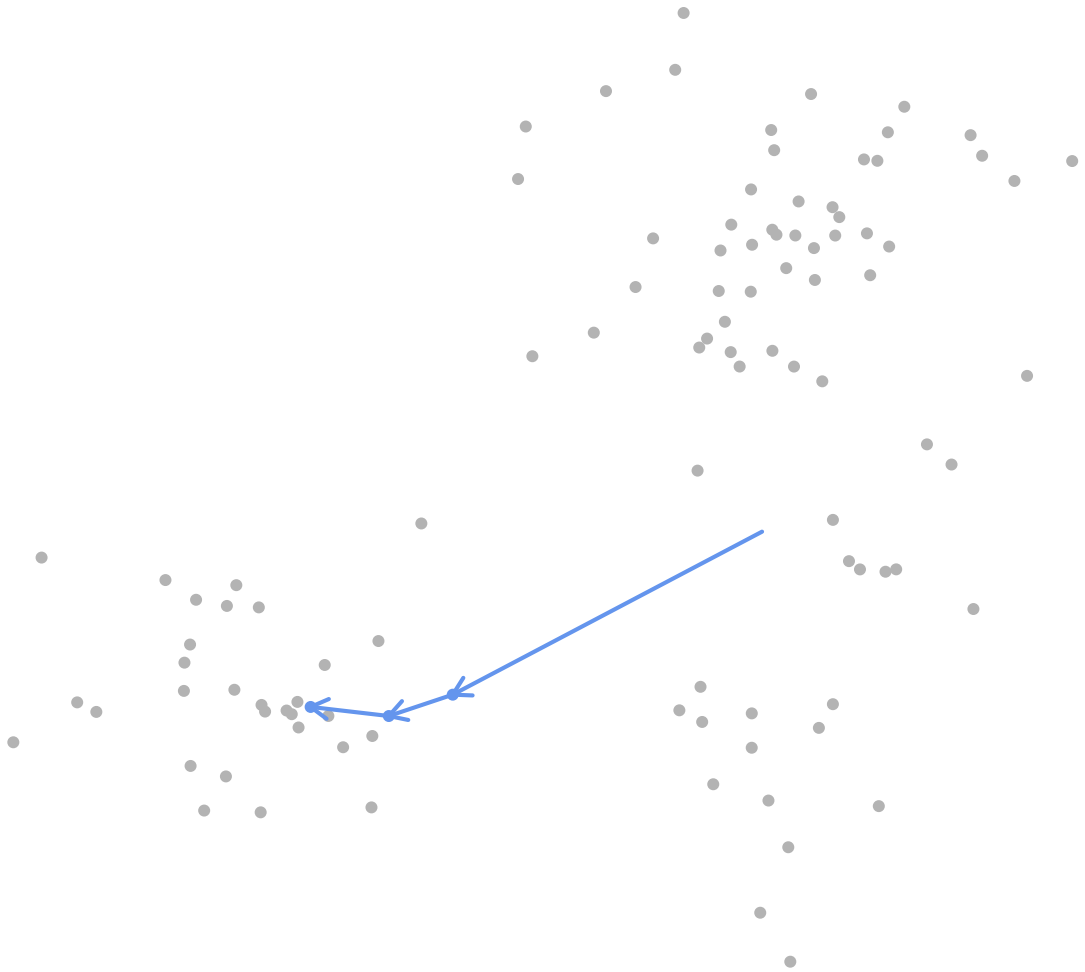
Visual Example



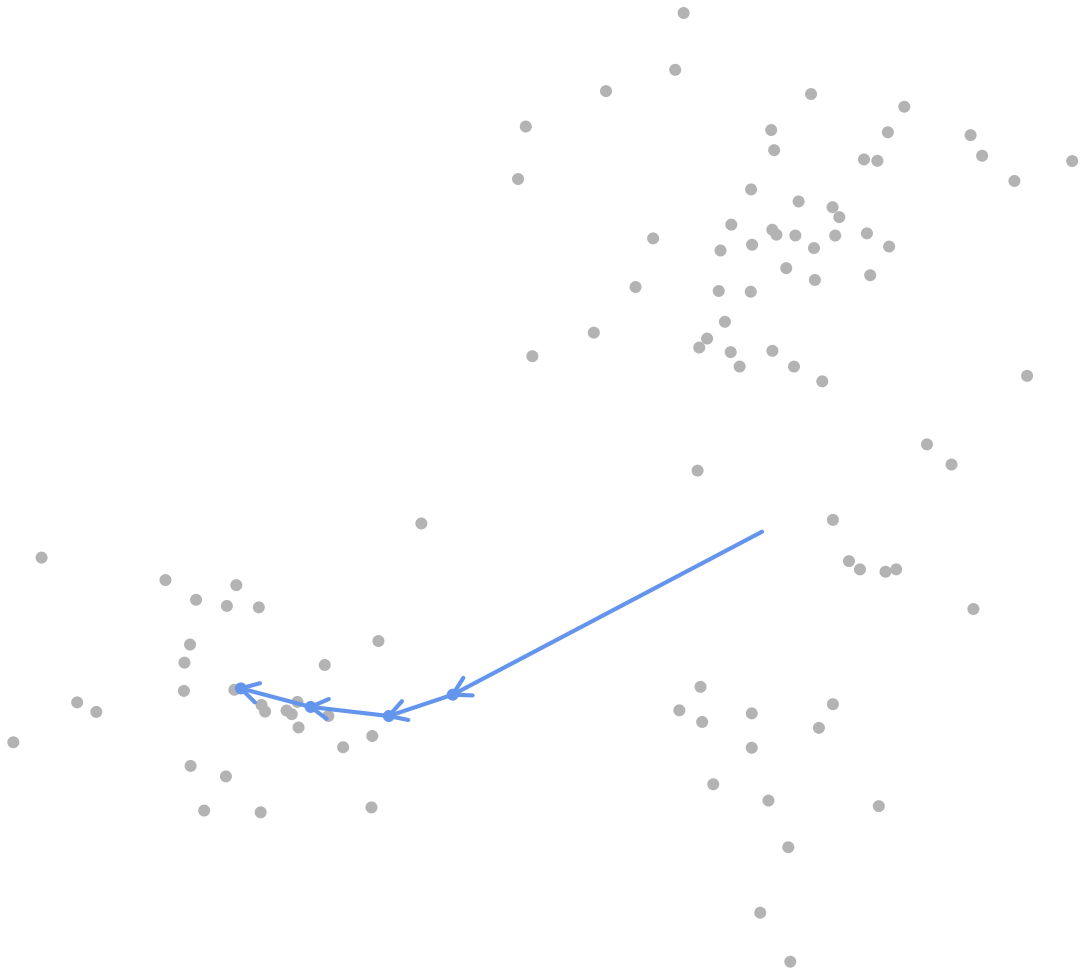
Visual Example



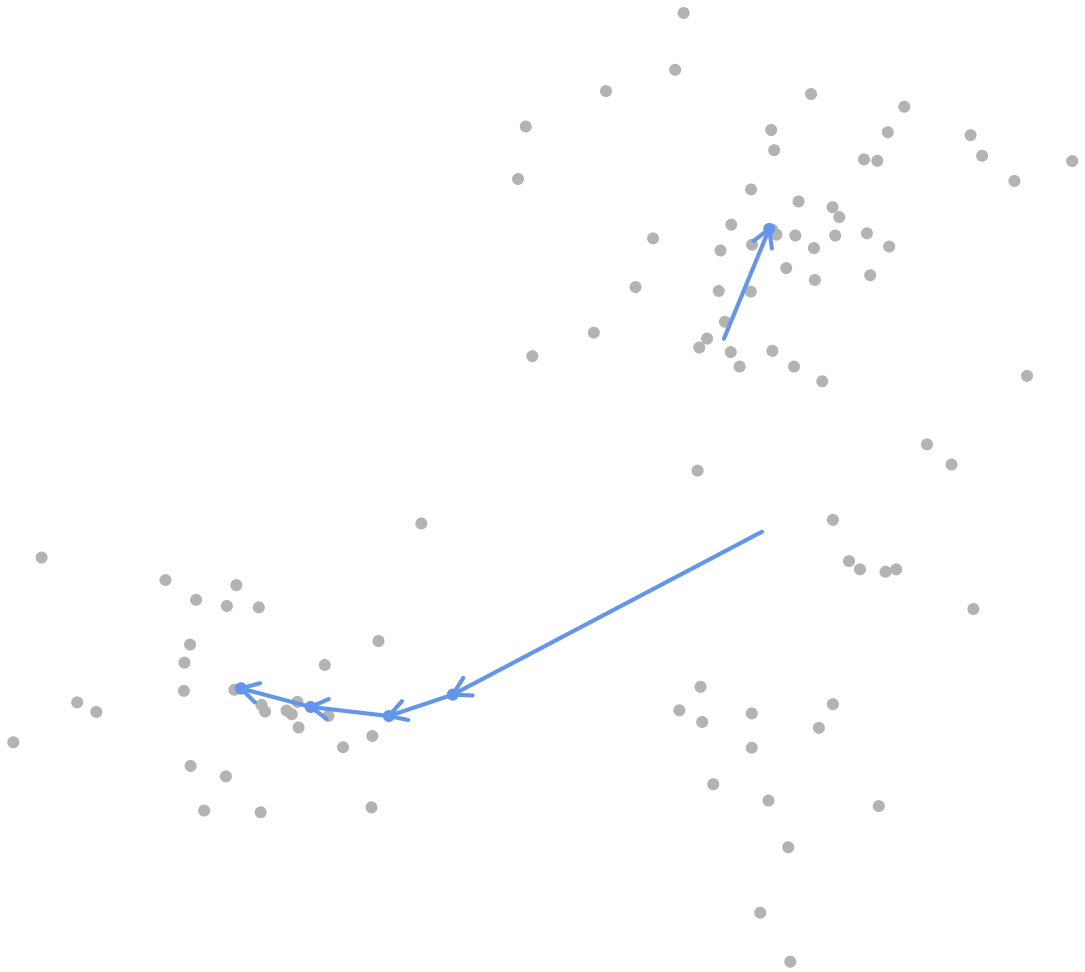
Visual Example



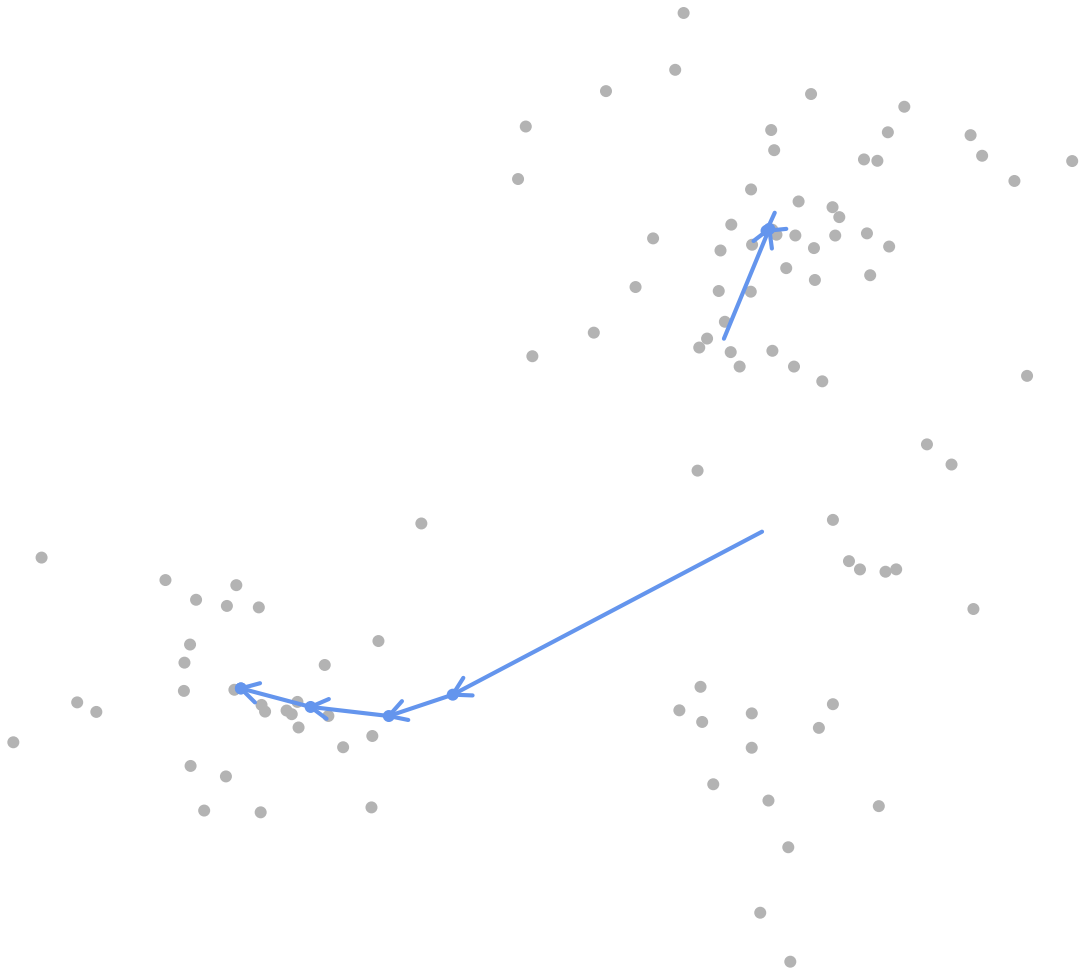
Visual Example



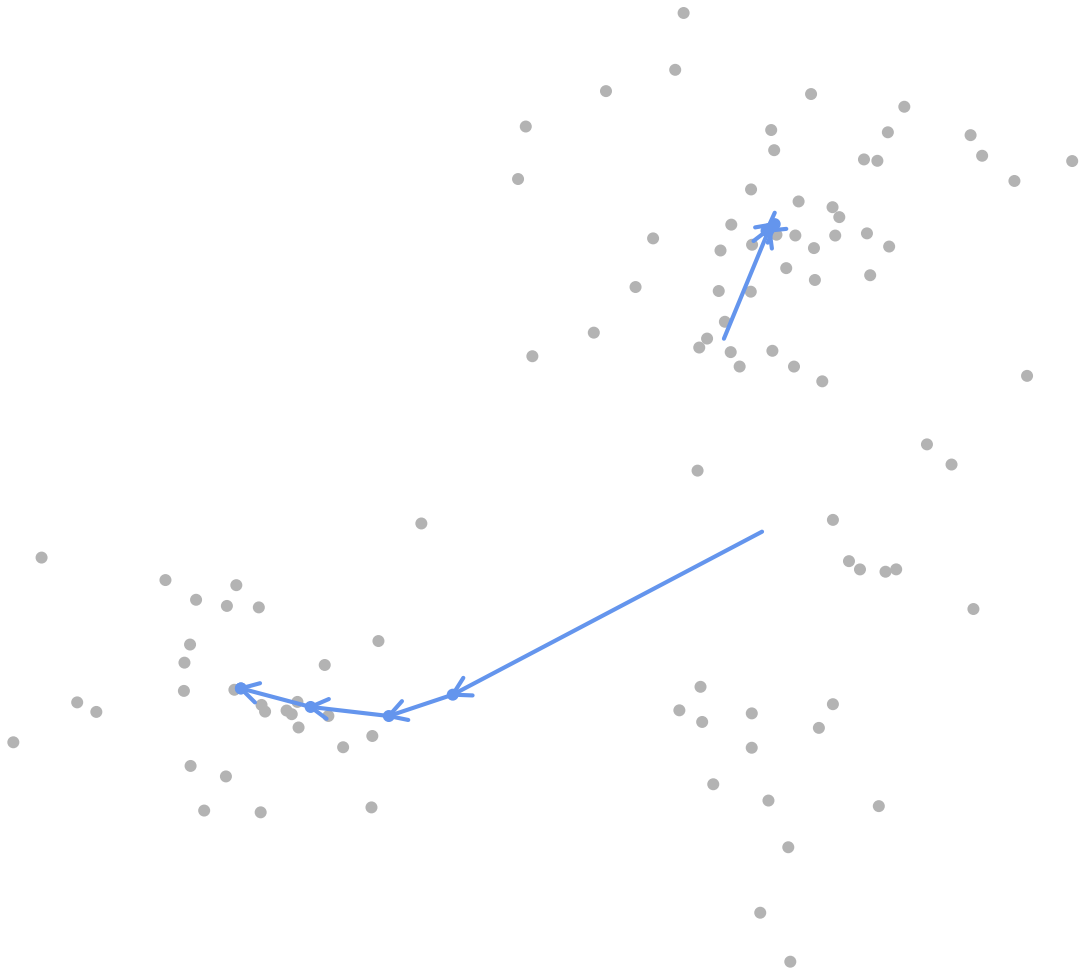
Visual Example



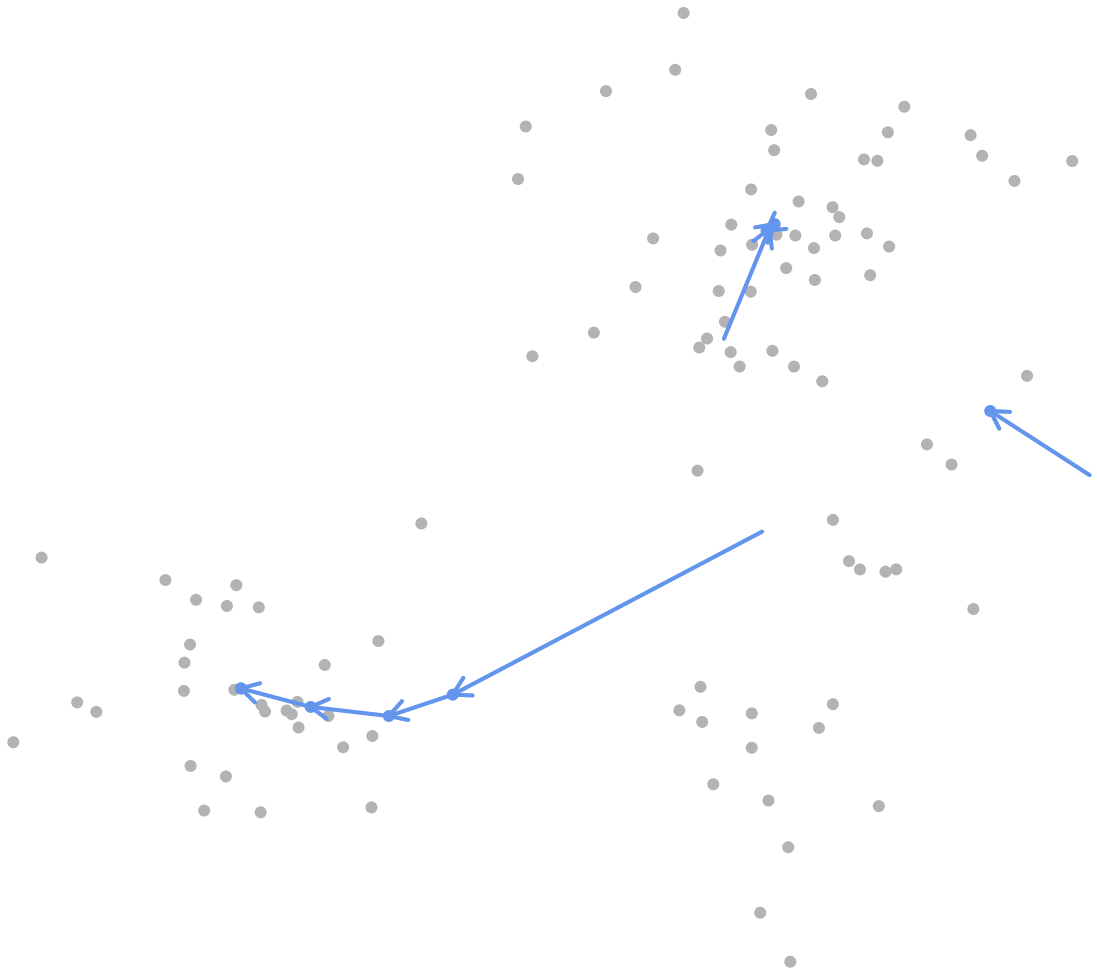
Visual Example



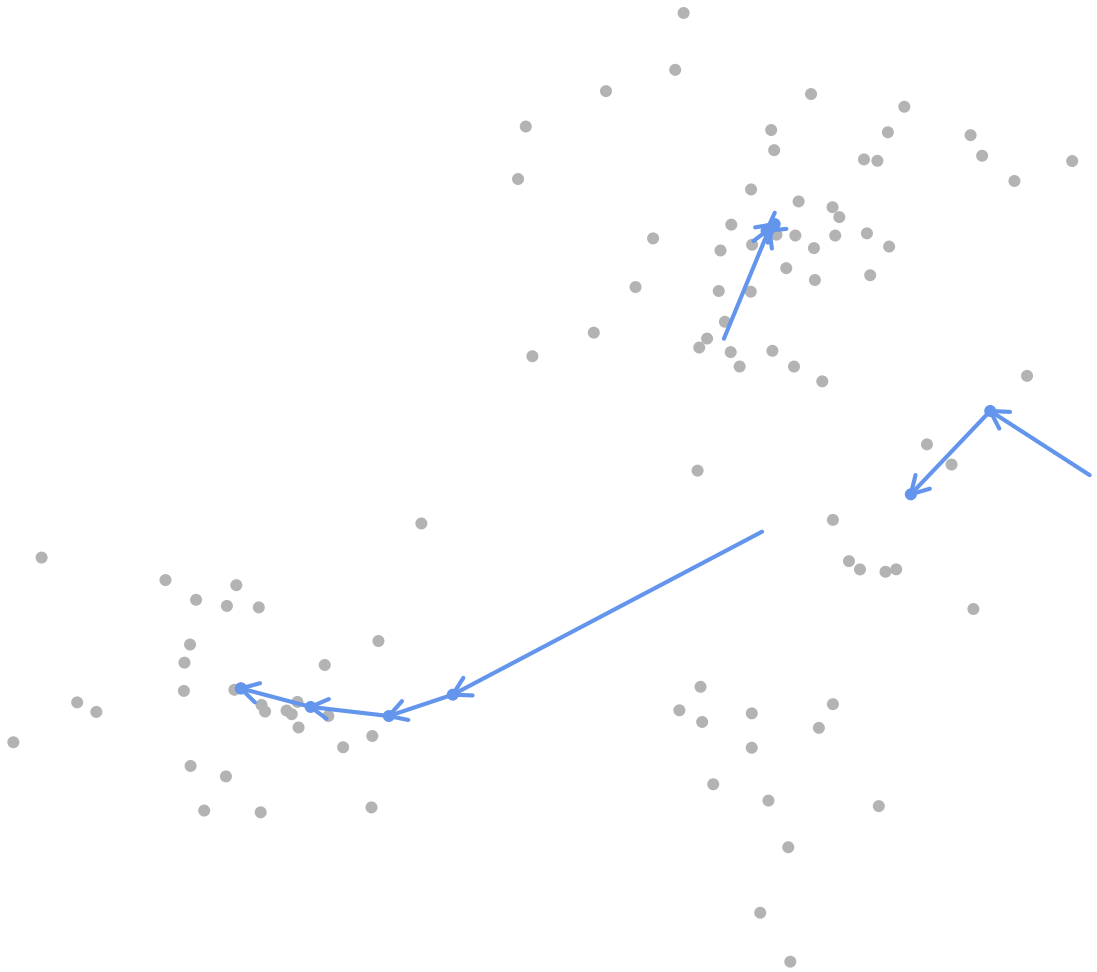
Visual Example



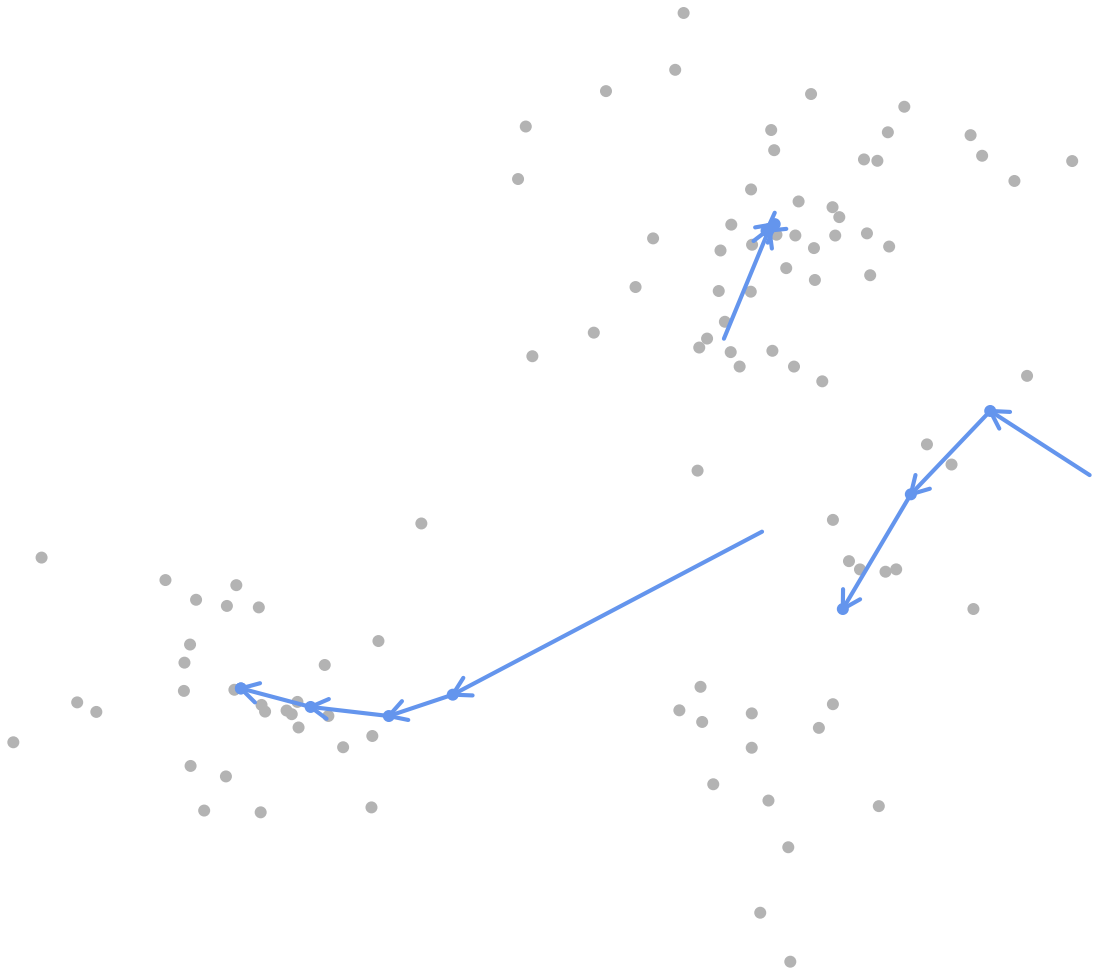
Visual Example



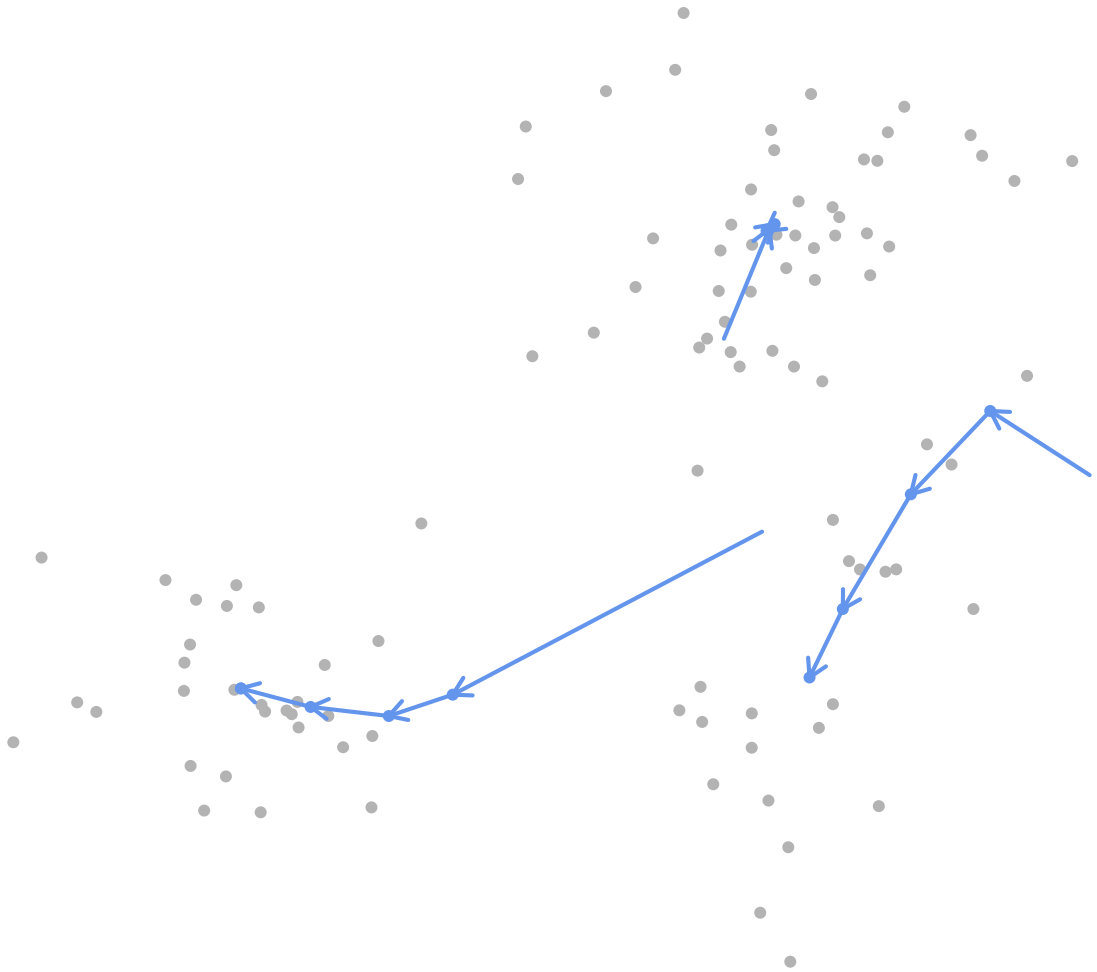
Visual Example



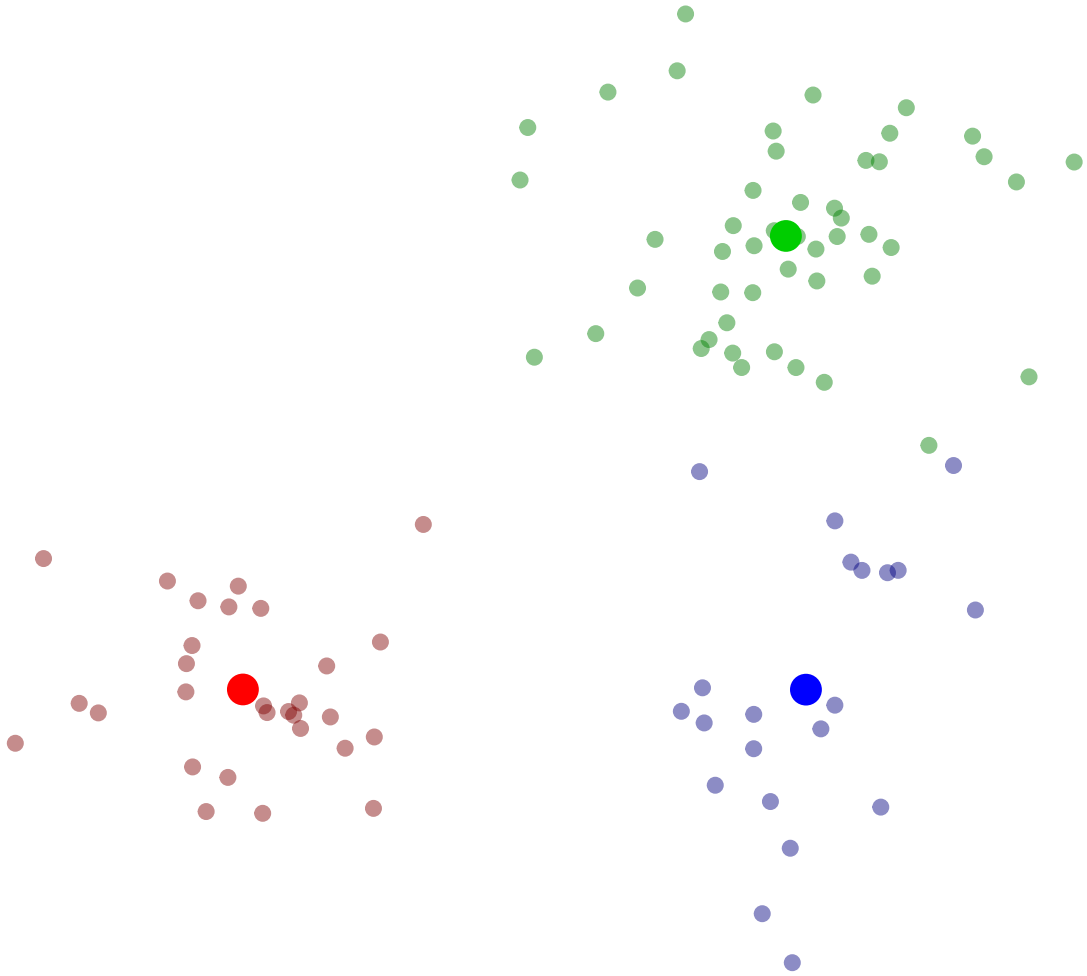
Visual Example



Visual Example



Visual Example



Example: Kmeans on Lautenberg Press Releases

Table 12.1: Applying k-means to the Press Releases

	Proportion	Words
1	0.17	s, presid, bush, administr, u, elect, compani, opec, tax, mr
2	0.16	senat, lautenberg, frank, r, statement, press, washington, comment, d, question
3	0.28	new, jersey, menendez, secur, million, fund, grant, chemic, nj, feder
4	0.39	sen, epa, bill, act, legisl, safeti, famili, protect, requir, victim

Example: Kmeans on Lautenberg Press Releases

Table 12.1: Applying k-means to the Press Releases

	Proportion	Words
1	0.17	s, presid, bush, administr, u, elect, compani, opec, tax, mr
2	0.16	senat, lautenberg, frank, r, statement, press, washington, comment, d, question
3	0.28	new, jersey, menendez, secur, million, fund, grant, chemic, nj, feder
4	0.39	sen, epa, bill, act, legisl, safeti, famili, protect, requir, victim

	Proportion	Words
1	0.28	new, jersey, menendez, fund, project, feder, million, program, counti, nj
2	0.21	senat, lautenberg, frank, r, statement, s, presid, comment, press, follow
3	0.42	american, famili, victim, legisl, epa, act, libya, terror, sponsor, report
4	0.08	secur, chemic, homeland, port, law, risk, protect, administr, dhs, bill

Table 12.3: Applying k-means to the press releases, with different starting values than Table 12.1.

Example: New Representation

Cluster	Words
1	passenger, amtrak, rail, security, gas, lott, stronger, preempt, approved, billion
2	court, she, demint, burr, mi, nc, nomination, r, was, medal
3	iraq, dc, american, president, united, taxpayer, issued, following, statement, senator
4	funds, for, announce, county, million, will, river, grants, project, hudson
5	income, care, drug, children, health, coverage, programs, beneficiaries, medicare, deductibles
6	epa, pollution, residents, dangerous, near, environmental, data, safety, agency, chemicals

Table 12.2: k-means clusters using word embeddings representation.

Example: Mixtures of von Mises-Fisher Distributions

Table 12.4: Applying Mixtures of von Mises-Fisher Distributions to Lautenberg Press Releases

Cluster	Top Words
1	sen, d, bill, victim, famili, terror, militari, libya, r, american
2	secur, chemic, homeland, port, risk, law, dhs, protect, tsa, rail
3	new, jersey, fund, menendez, feder, project, nj, million, program, counti
4	senat, s, lautenberg, presid, bush, frank, releas, statement, unit, iraq

Example: Hierarchical Clustering

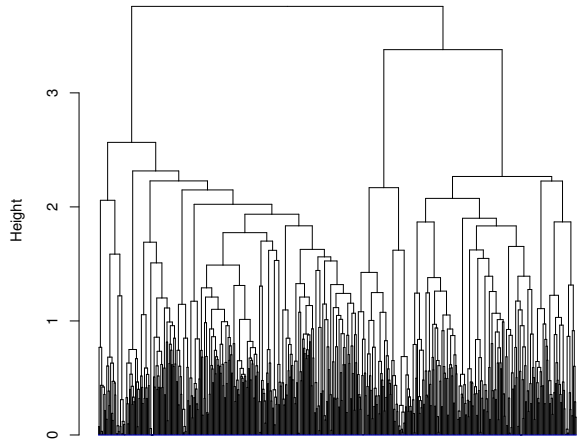


Figure 12.3: Hierarchical Clustering of the Lautenberg Press Releases using Ward's Minimum Variance Method.

Cluster	Top Words
1	american, dc, kill, unit, compens, bipartisan, john, legal, r, md
2	jersey, menendez, new, nj, robert, resid, communiti, project, garden, contact
3	secur, chemic, homeland, stronger, dhs, preempt, port, adopt, deem, attack

Table 12.6: Applying Hierarchical Clustering to the Lautenberg Press Releases

Description from Wikipedia

Back in the Senate, Lautenberg was once again considered one of the chamber's most liberal members. He was pro-choice, supported gun control, introduced many bills increasing penalti and car theft, and criticized the Bush administration on national security issues.[10] He was heavily involved in various anti-smoking and airline safety legislation. He also co-sponsored legisl drunk driving penalties. He was probably best known as the author of the legislation that banned smoking from most commercial airline flights.[19] He also is known for authoring the Ryan W which provides services to AIDS patients. **Upon his return to the Senate, Lautenberg was the first U.S. senator to introduce legislation calling for homeland security funds to be distributed so of risk and vulnerability.**[20]

In 2005, he became a leading voice within the Senate in calling for an investigation into the Bush administration payment of columnists.[21]

When Jon Corzine resigned from the Senate to become Governor of New Jersey, Lautenberg became the senior senator again in 2006. This also made him the only person to have been b senior senator from New Jersey twice each.[*citation needed*] Lautenberg received an "A" on the Drum Major Institute's 2005 Congressional Scorecard on middle-class issues.[22]

In 2007, Lautenberg proposed the Denying Firearms and Explosives to Dangerous Terrorists Act of 2007, designed to deny weapons purchases by persons that the government has placed watchlist. On June 21, 2007, Lautenberg passed Clifford Case for the most votes on the Senate floor of any United States Senator in New Jersey history.[23]

Interpreting Cluster Components

Unsupervised methods

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify (separating) words

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify (separating) words
    - Use these to help infer differences across clusters

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify (separating) words
    - Use these to help infer differences across clusters
- Transparency

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify (separating) words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify (separating) words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify (separating) words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean
  - Provide documents + organizations

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify (separating) words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean
  - Provide documents + organizations

How Do We Choose K?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
    - Sum squared errors decreases as K increases

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

- No one statistic captures how you want to use your data

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

- No one statistic captures how you want to use your data
- But, can help guide your selection

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistics + manual search

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistics + manual search
- Humans should be the final judge

Lautenberg Press Releases with Many Clusters

Cluster	Top Words
1	hous, author, deadlin, system, manag, asset, new, hud, public, extend
2	epa, children, librari, test, mercuri, protect, agenc, chemic, product, health
3	s, u, compani, iran, foreign, busi, halliburton, control, airlin, sanction
4	secur, million, fund, grant, jersey, nj, homeland, port, new, dhs
5	program, educ, food, gang, train, youth, provid, job, abstin, fund
6	new, jersey, state, declar, menendez, feder, counti, disast, faa, administr
7	militari, increas, sen, famili, tricar, d, afghanistan, co, retire, fee
8	budget, new, million, cut, jersey, presid, highland, s, bush, educ
9	fire, safeti, grant, depart, firefight, colleg, campus, assist, nj, student
10	airport, grant, receiv, feder, jersey, menendez, fund, improv, announc, new
11	site, epa, ringwood, superfund, contamin, cleanup, report, resid, environment, agenc
12	presid, bush, s, senat, iraq, statement, administr, white, social, iraqi
13	rang, safeti, guard, grove, warren, air, fire, resid, communiti, account
14	senat, lautenberg, committe, monmouth, iraq, frank, commiss, fort, republican, r
15	american, women, honor, gao, iraq, news, live, mr, vaccin, propaganda
16	cigarett, tobacco, light, tar, compani, nicotin, market, smoker, low, d
17	broadband, municip, access, offer, provid, communiti, state, afford, public, prohibit
18	corp, park, armi, natur, beach, palmyra, cove, engin, dredg, munit
19	build, energi, emiss, bill, global, govern, warm, greenhous, green, feder
20	fund, new, jersey, project, menendez, transport, million, program, counti, tunnel
21	contract, presid, t, said, secretari, hud, get, jackson, select, don
22	ocean, water, state, protect, bill, pet, beach, whistleblow, legisl, nation
23	secur, port, homeland, risk, tsa, screener, rail, senat, chertoff, cap
24	safeti, airport, flight, faa, delay, truck, administr, runway, new, transport
25	judg, court, s, suprem, right, nomin, decis, senat, wigenton, attorney
26	lautenberg, press, statement, ambassador, sen, farm, issu, frank, offic, d
27	senat, lautenberg, statement, frank, r, follow, issu, unit, question, press

Example Discovery: Congressional Communication to Constituents

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)

Example Discovery: Congressional Communication to Constituents

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology

Example Discovery: Congressional Communication to Constituents

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising

Example Discovery: Congressional Communication to Constituents

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming

Example Discovery: Congressional Communication to Constituents

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking

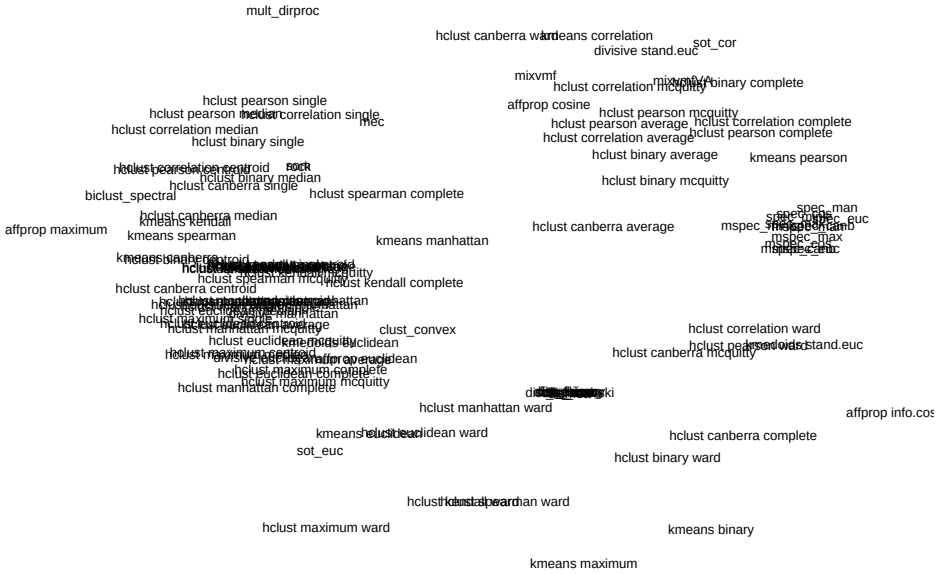
Example Discovery: Congressional Communication to Constituents

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)

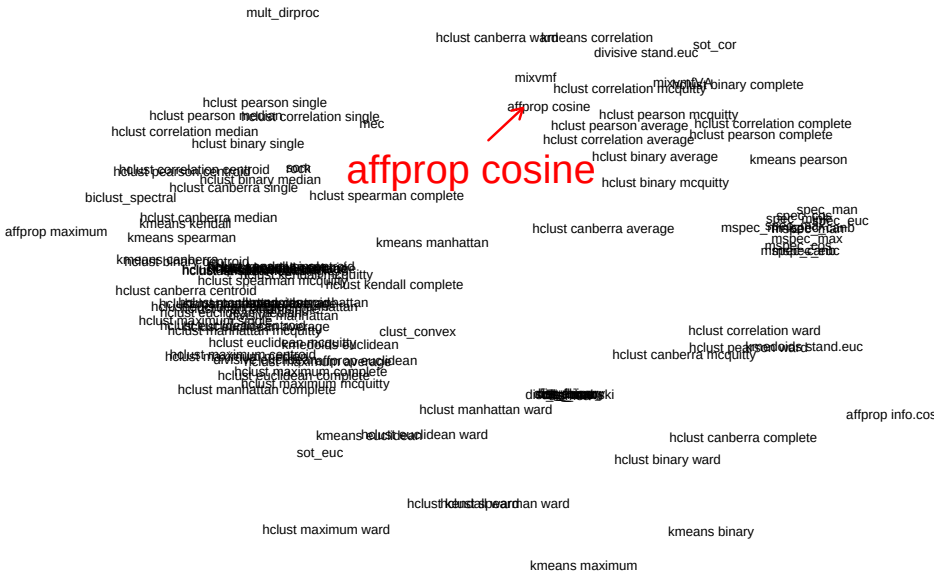
Example Discovery: Congressional Communication to Constituents

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)
- Apply method (relying on many clustering algorithms)

Example Discovery

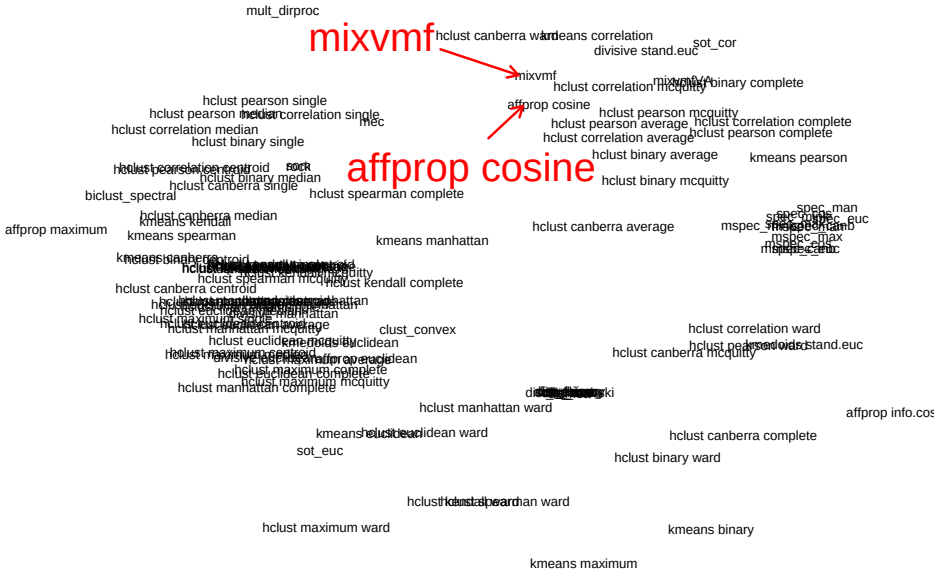


Example Discovery



Each point is a clustering
Affinity Propagation-Cosine
(Dueck and Frey 2007)

Example Discovery



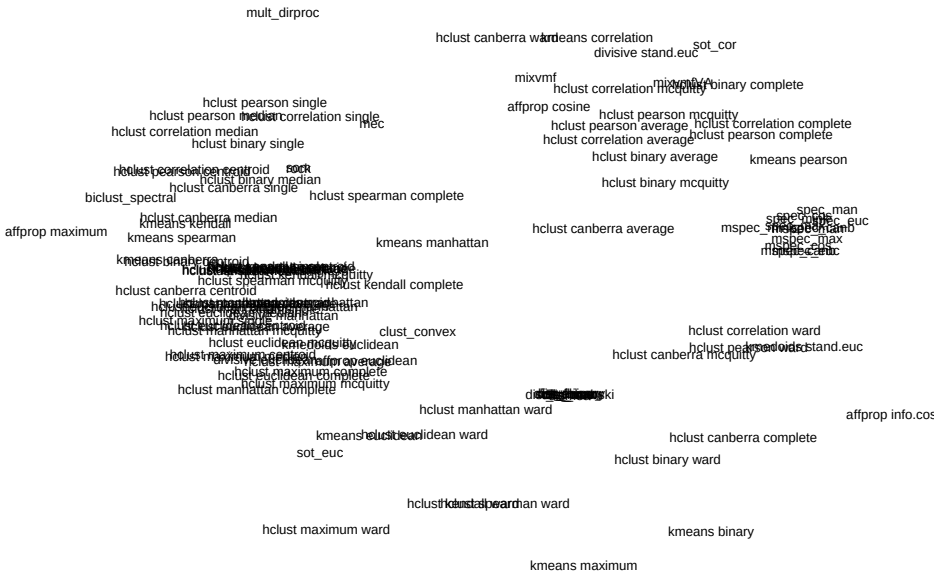
Each point is a clustering Affinity Propagation-Cosine (Dueck and Frey 2007)

Close to:

Mixture of von Mises-Fisher distributions (Banerjee et. al. 2005)

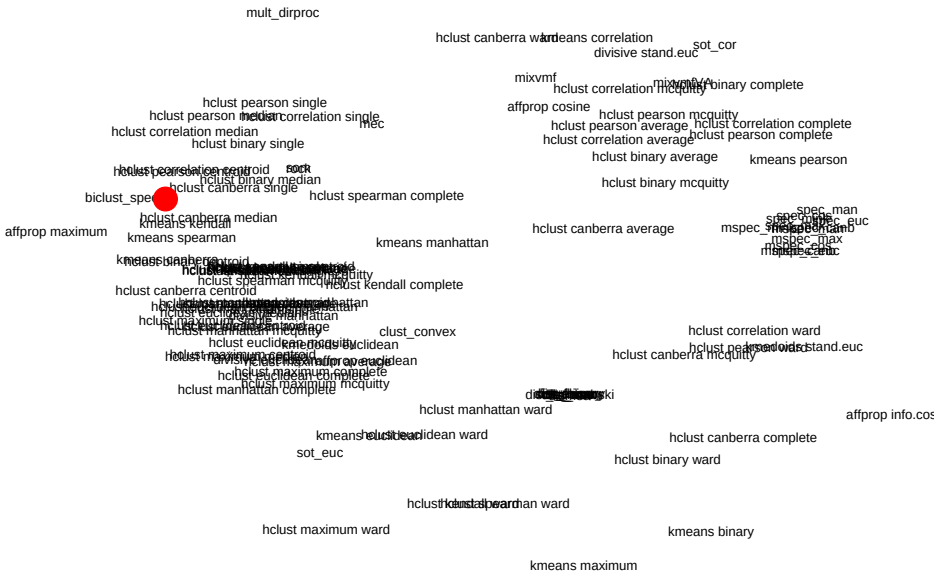
$\Rightarrow$ Similar clustering of documents

Example Discovery



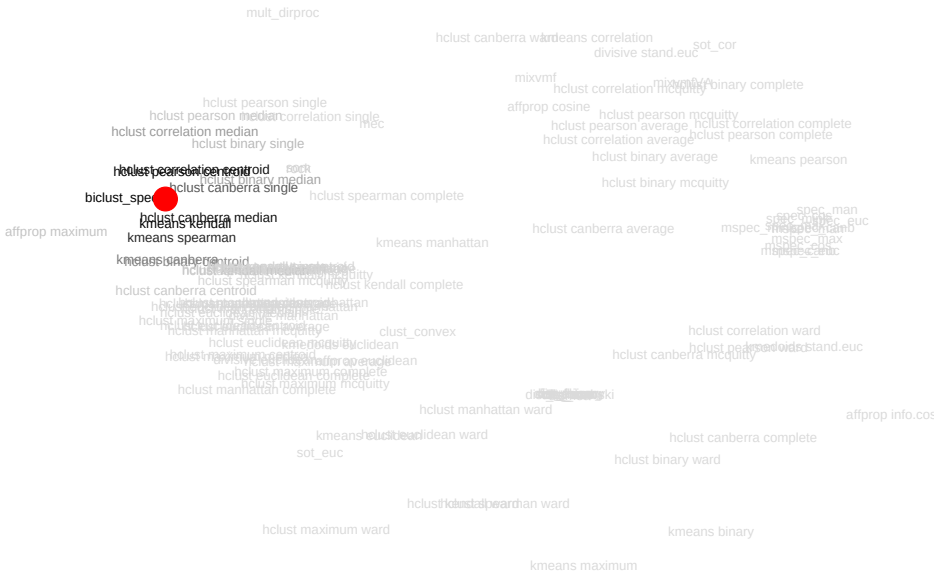
Space between methods:

Example Discovery



Space between methods:

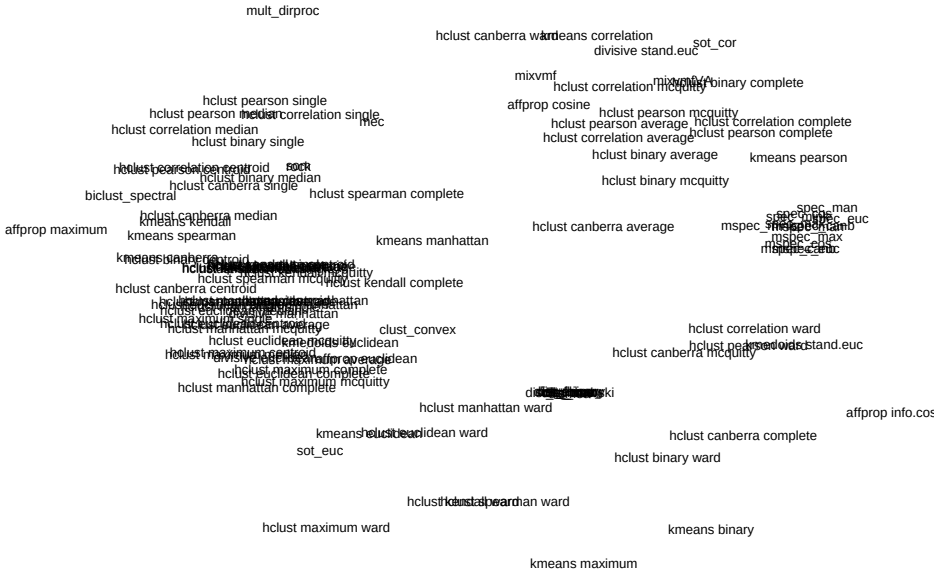
Example Discovery



Space between methods:

local cluster ensemble

Example Discovery



Example Discovery

Found a region with clusterings that all reveal the same important insight

Example Discovery

Mixture:

Example Discovery

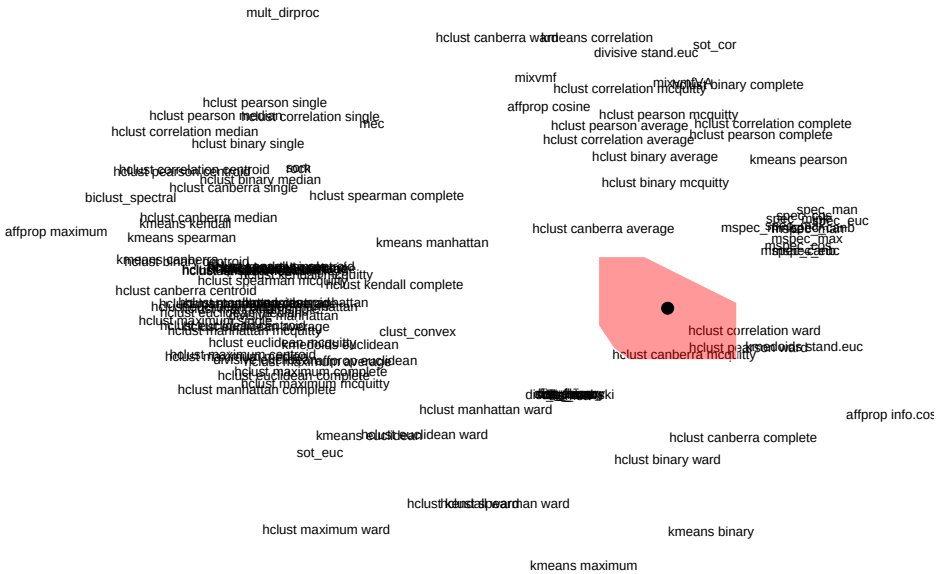
Mixture:

0.39 Hclust-Canberra-McQuitty

0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering
Random Walk
(Metrics 1-6)

0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

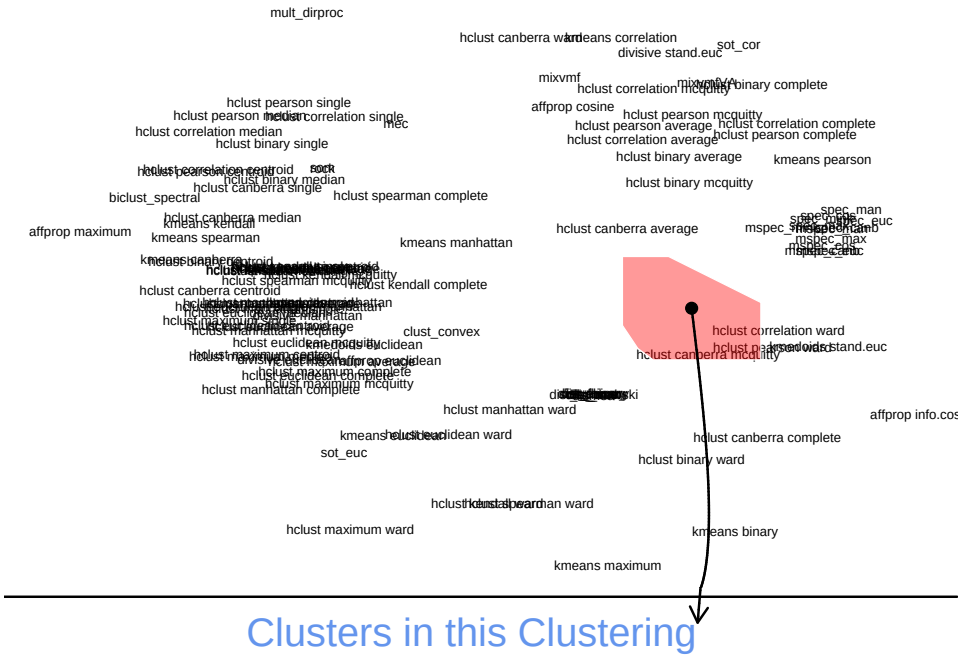
0.04 Spectral clustering
Symmetric
(Metrics 1-6)

Example Discovery

Mixture:

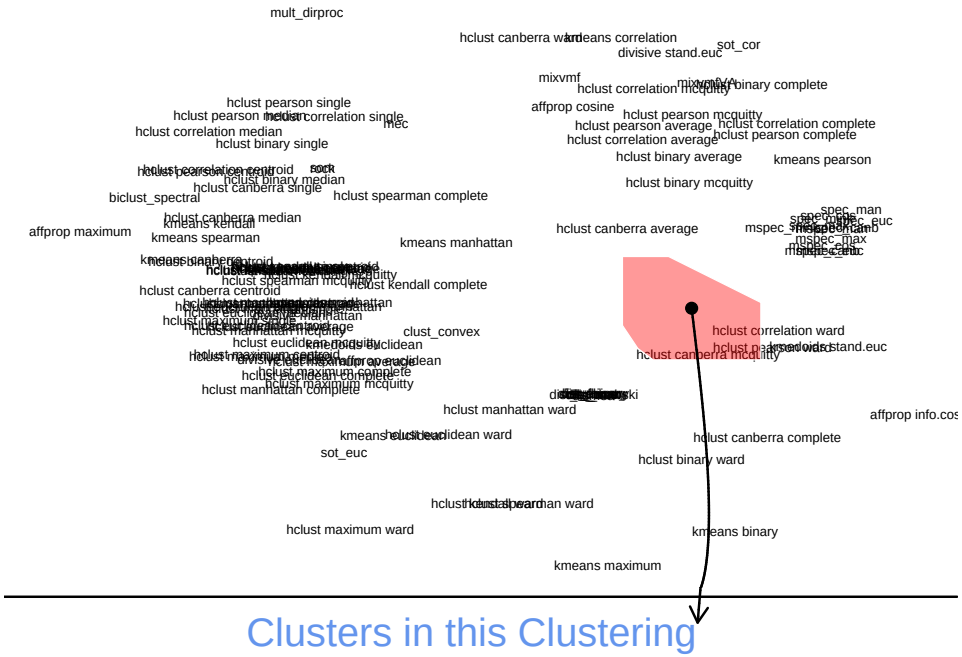
- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward
- 0.09 Hclust-Pearson-Ward
- 0.05 Kmediods-Cosine
- 0.04 Spectral clustering Symmetric (Metrics 1-6)

Example Discovery

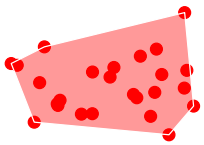


Mayhew

Example Discovery



Clusters in this Clustering

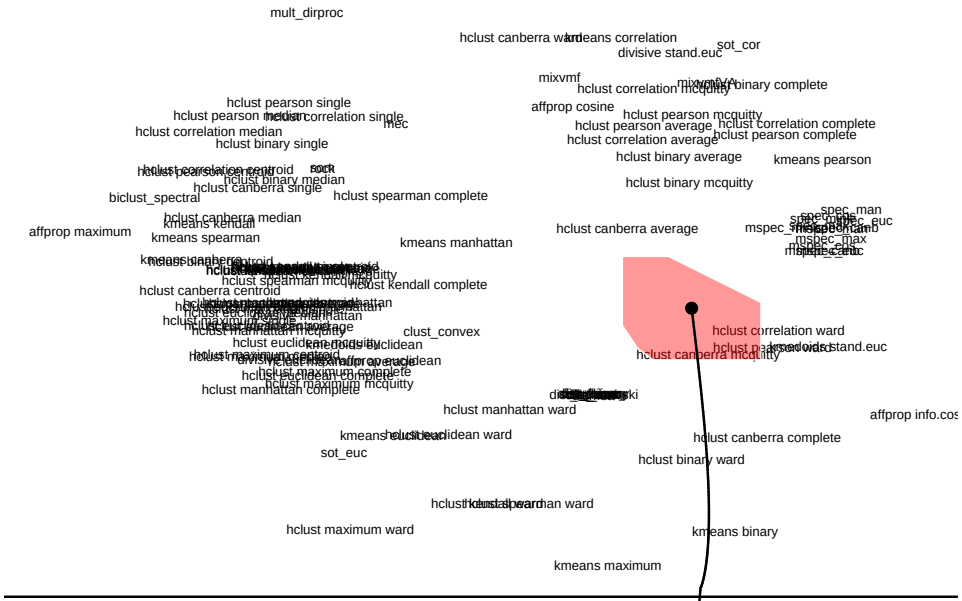


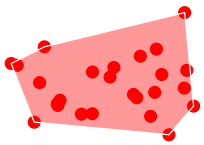
Credit Claiming
Pork

Credit Claiming, Pork:
"Sens. Frank R. Lautenberg (D-NJ) and Robert Menendez (D-NJ) announced that the U.S. Department of Commerce has awarded a $100,000 grant to the South Jersey Economic Development District"

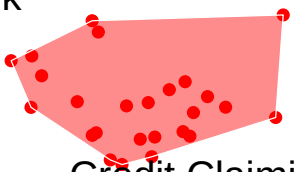
Mayhew

Example Discovery





Credit Claiming
Pork

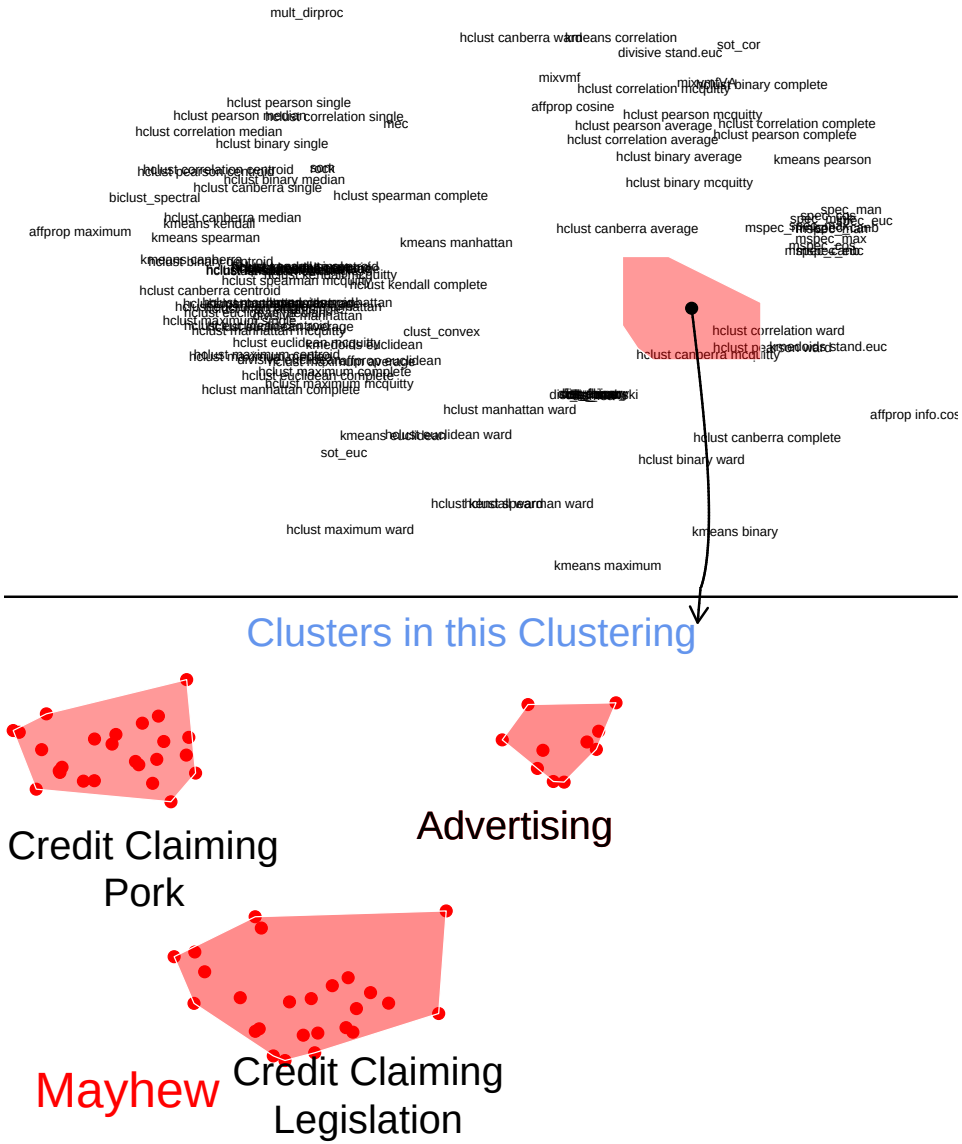


Mayhew

Credit Claiming
Legislation

Credit Claiming, Legislation:
"As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period"

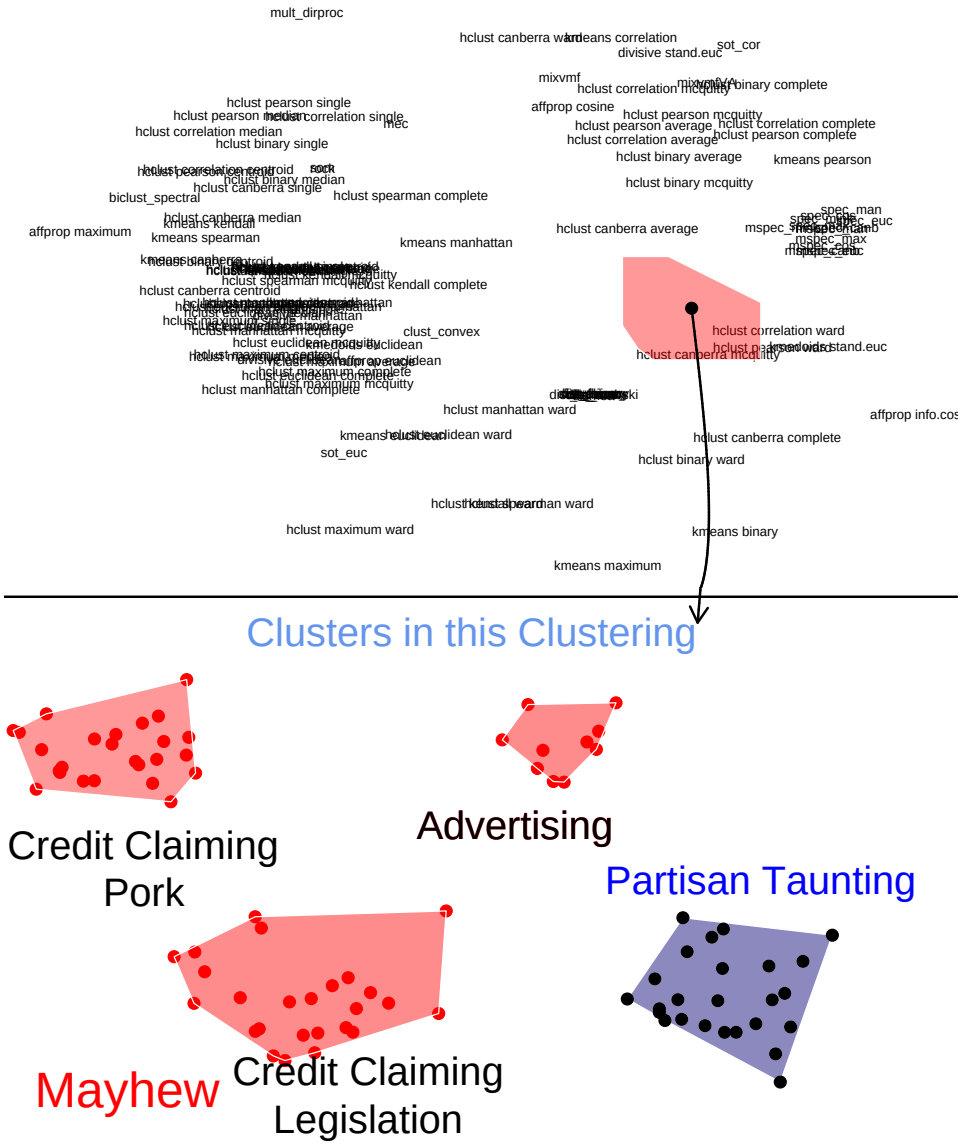
Example Discovery



Advertising:

"Senate Adopts Lautenberg/Menendez Resolution Honoring Spelling Bee Champion from New Jersey"

Example Discovery: Partisan Taunting



Clusters in this Clustering

Partisan Taunting:
"Republicans Selling Out Nation on Chemical Plant Security"

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Sen. Lautenberg
on Senate Floor
4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Sen. Lautenberg
on Senate Floor
4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Sen. Lautenberg
on Senate Floor
4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities less safe." [Homeland Security]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Definition: Explicit, public, and negative attacks on another political party or its members

Sen. Lautenberg on Senate Floor 4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities less safe." [Homeland Security]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Definition: Explicit, public, and negative attacks on another political party or its members

Consequences for representation: Deliberative, Polarization, Policy

Sen. Lautenberg on Senate Floor 4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities less safe." [Homeland Security]

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

Out of Sample Confirmation of Partisan Taunting

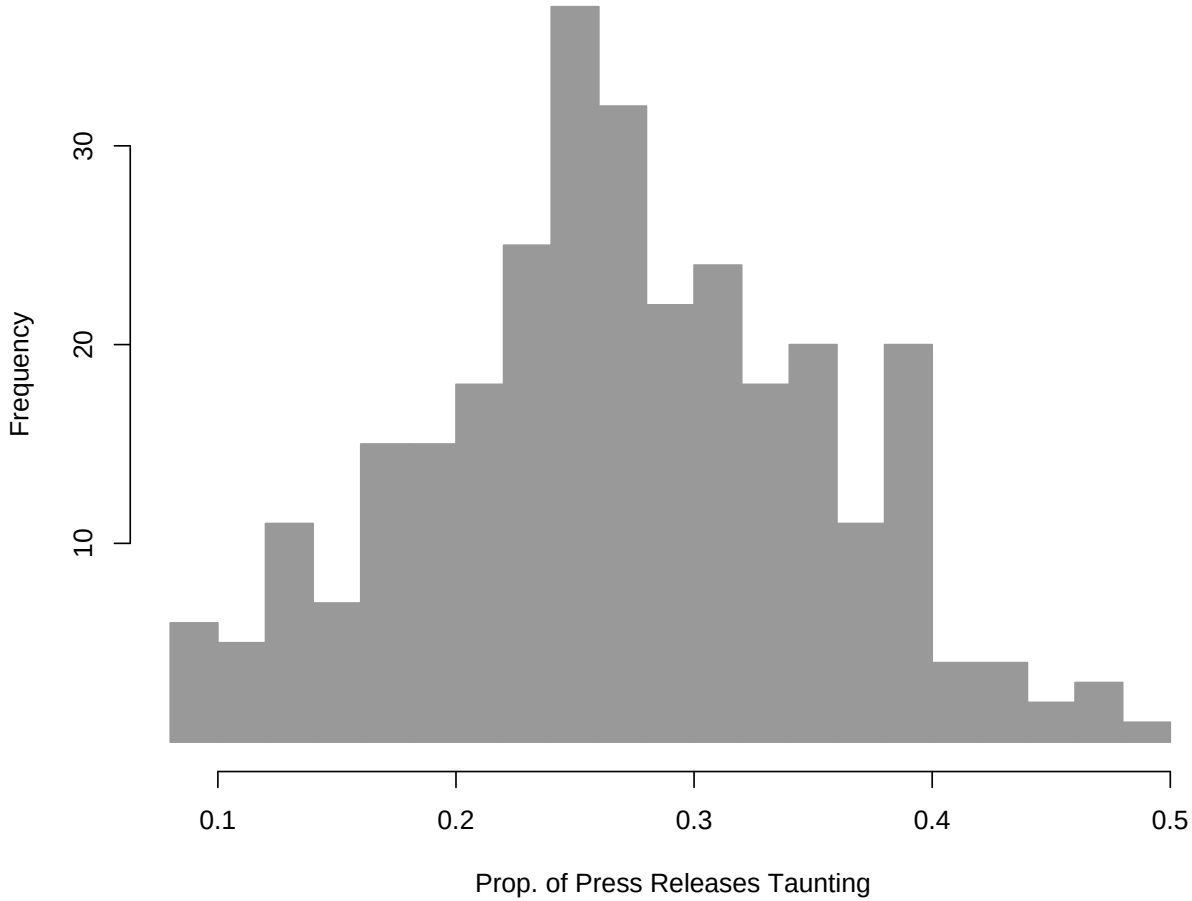
- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party

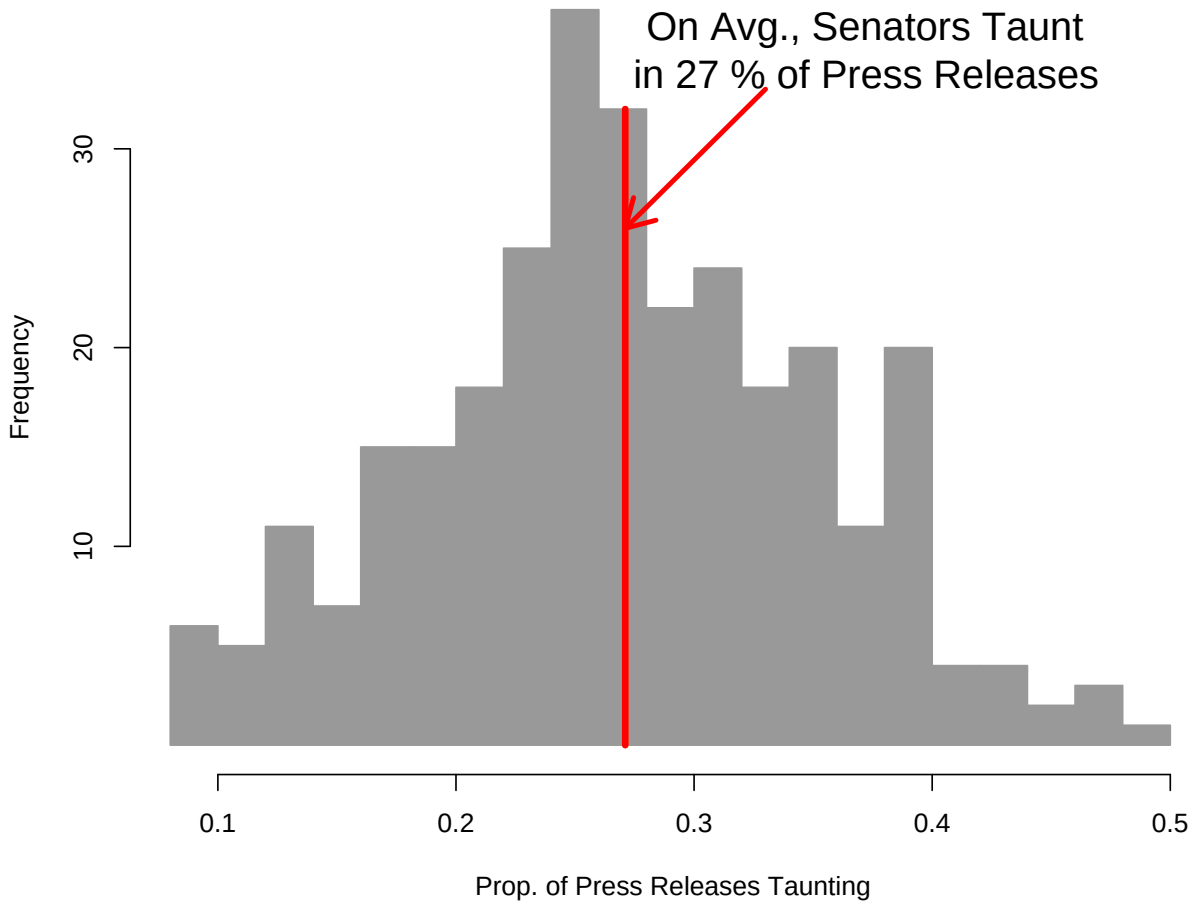
Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party

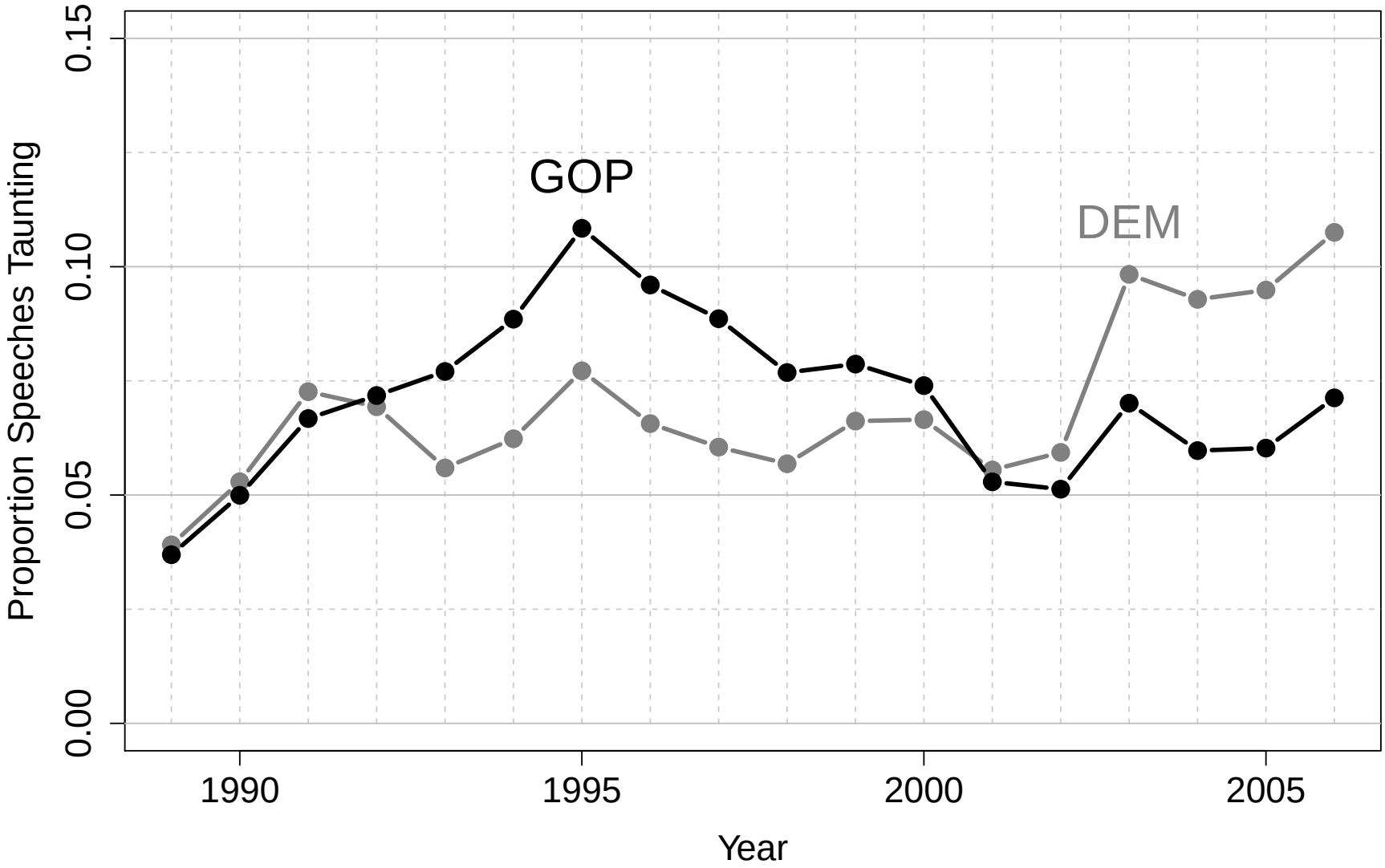


Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party



Over Time Taunting Rates in Speeches



Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

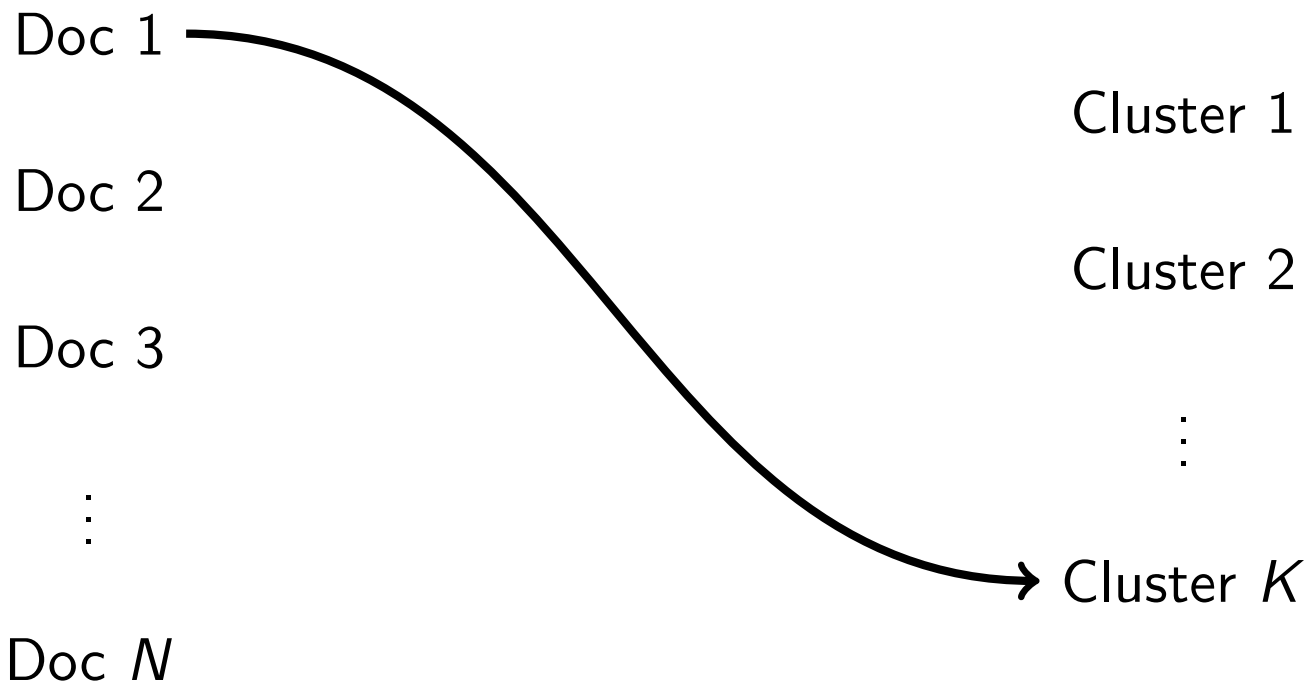
$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering

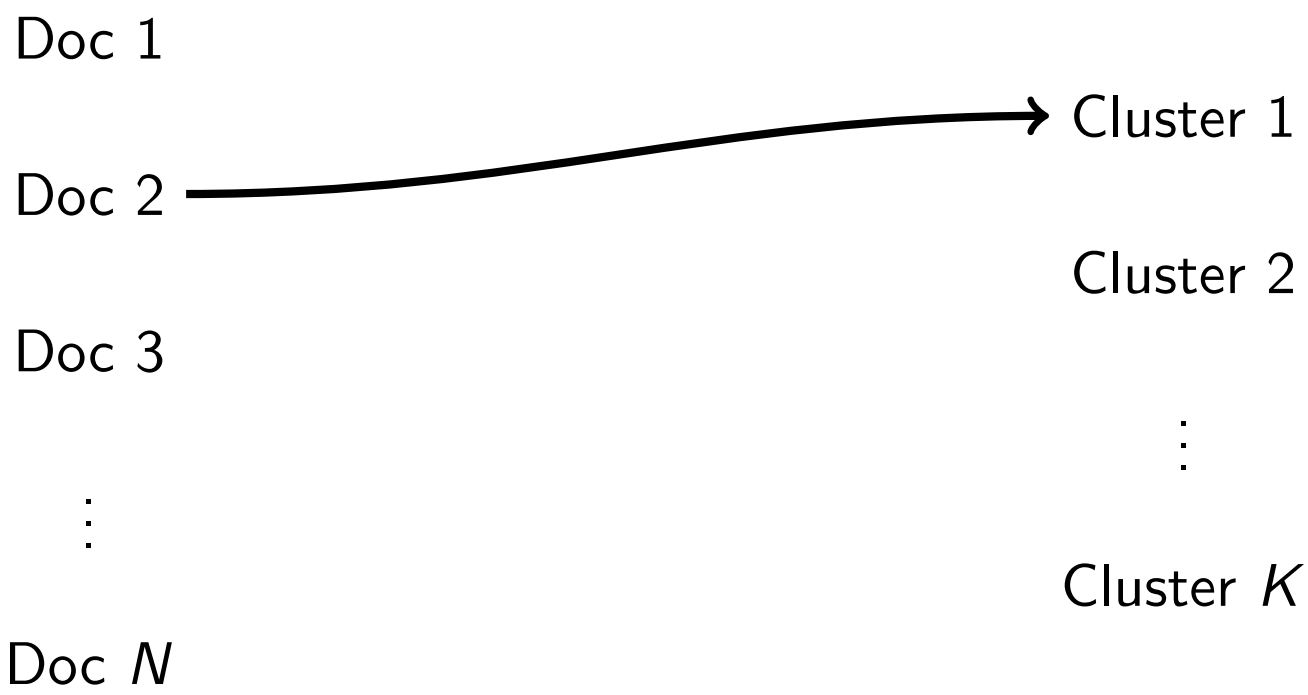
Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

Clustering

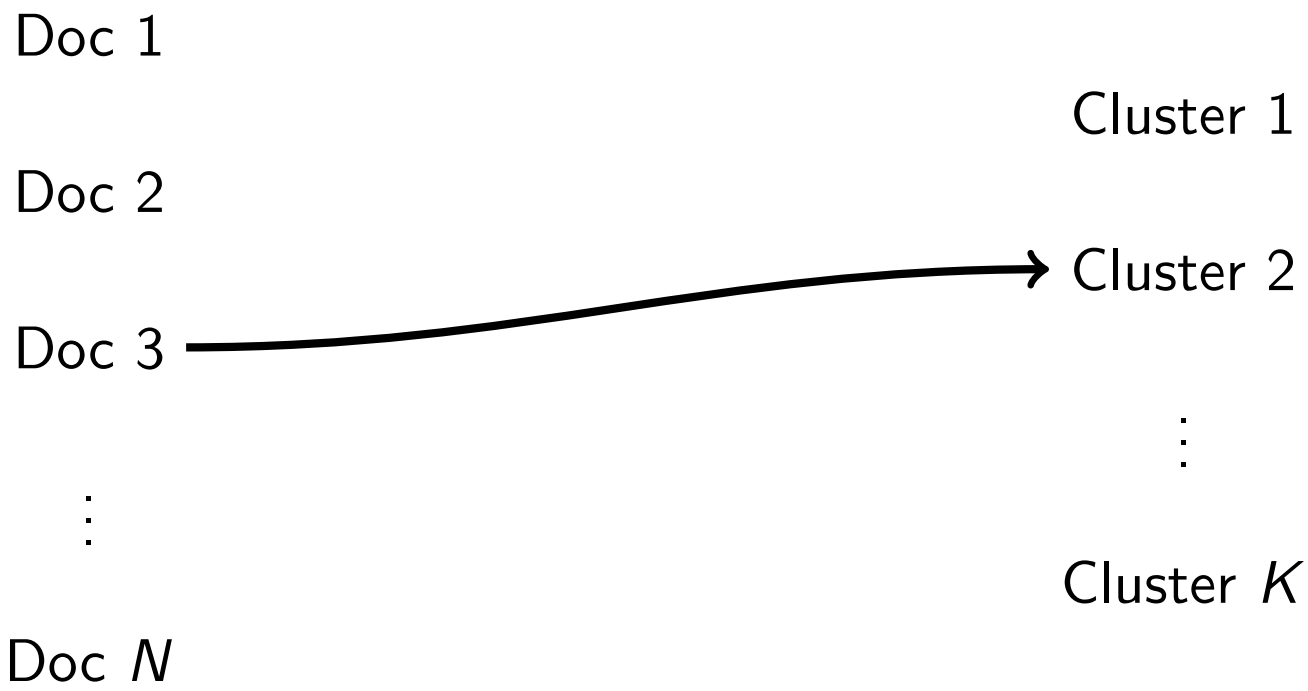
Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

Clustering

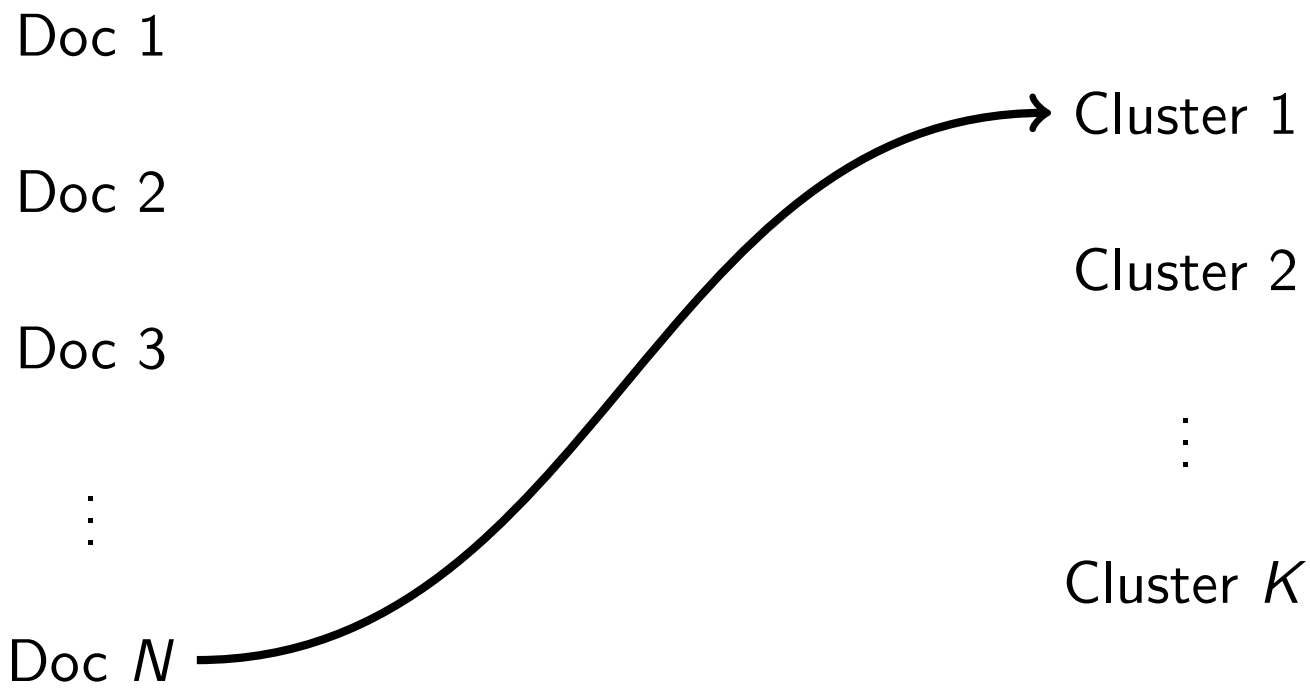
Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster



Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

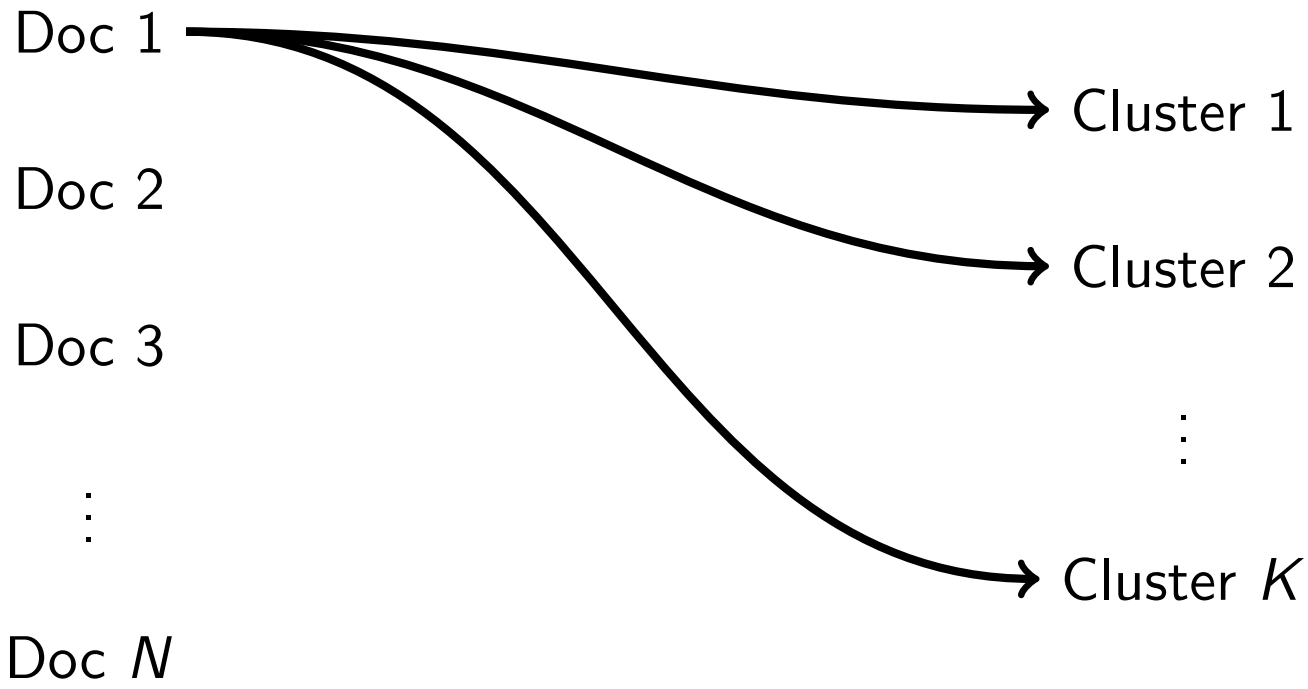
$\vdots$

Cluster K

Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters



Topic Models

Topic Models

Two matrices estimated:

Topic Models

Two matrices estimated:

1) Topical Prevalence Matrix (DxK)

Topic Models

Two matrices estimated:

1) Topical Prevalence Matrix (DxK)

$$\boldsymbol{\pi} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

Topic Models

Two matrices estimated:

1) Topical Prevalence Matrix (DxK)

$$\boldsymbol{\pi} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

2) Topical Content Matrix (JxK)

Topic Models

Two matrices estimated: $W \approx \pi\mu$

1) Topical Prevalence Matrix (DxK)

$$\boldsymbol{\pi} = \left[\begin{array}{c|ccc|c} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

2) Topical Content Matrix (JxK)

$$\boldsymbol{\mu} = \left[\begin{array}{c|ccc|c} & Topic1 & Topic2 & \dots & TopicK \\ \hline \text{“}text\text{”} & .02 & .001 & \dots & 0.001 \\ \text{“}data\text{”} & .001 & .02 & \dots & 0.001 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{“}analysis\text{”} & .01 & .01 & \dots & 0.0005 \end{array}\right]$$

Latent Dirichlet Allocation

- Latent Dirichlet Allocation estimates:

Latent Dirichlet Allocation

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.

Latent Dirichlet Allocation

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Latent Dirichlet Allocation

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Latent Dirichlet Allocation

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Maintained assumptions: Bag of words/fix number of topics ex ante.

What this means in pictures

Say you have
a lot of people.

What this means in pictures

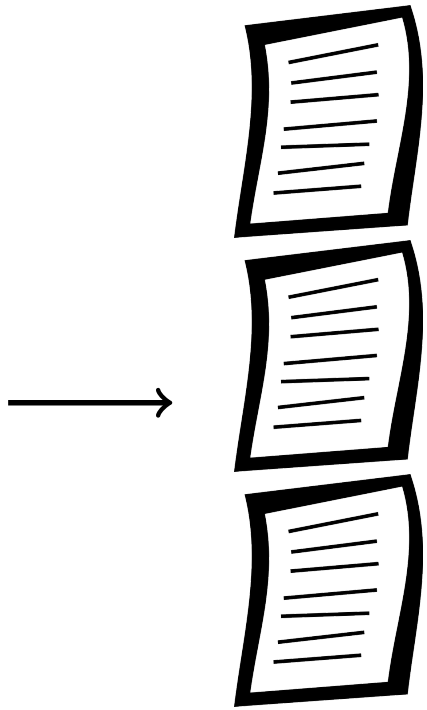
Say you have
a lot of people.

Each writes
some texts

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

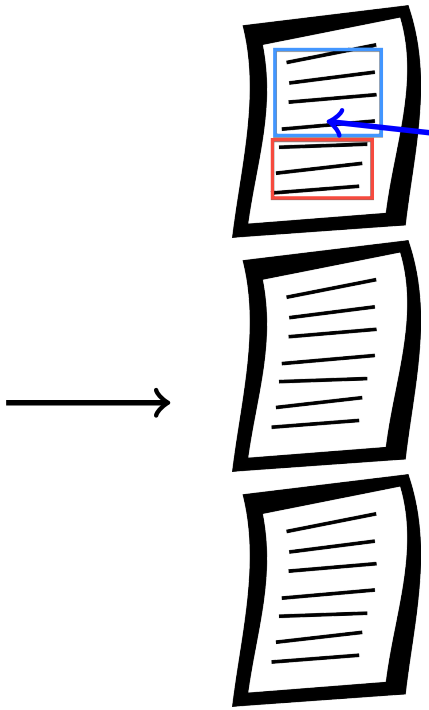


that discuss a few
different topics

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



that discuss a few
different topics

Topic 1

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

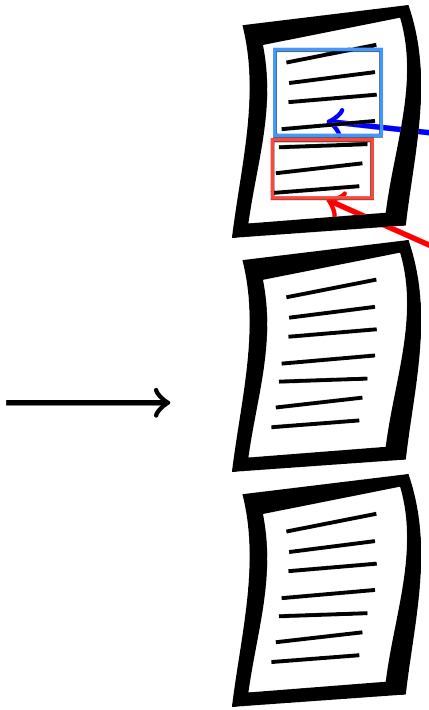
Topic 1

Topic 2

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



that discuss a few
different topics

Topic 1

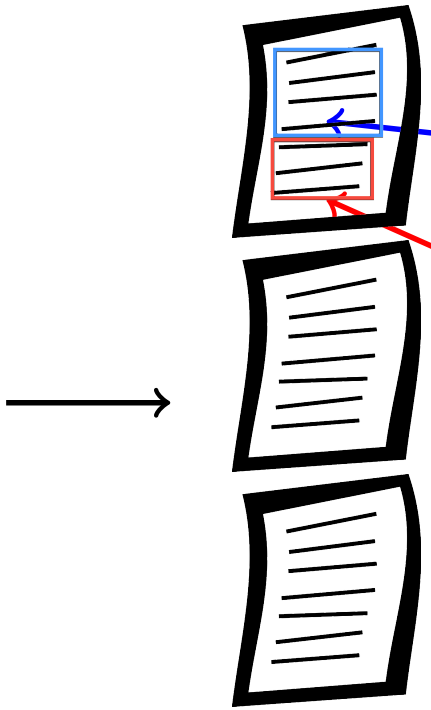
Topic 2

The Latent Dirichlet Allocation estimates:

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



that discuss a few
different topics

Topic 1

Topic 2

The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words

What this means in pictures

Say you have a lot of people.

Each writes some texts

that discuss a few different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words

What this means in pictures

Say you have a lot of people.

Each writes some texts

that discuss a few different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words
2. The proportion of each document in each topic

STM = LDA + Contextual Information

STM = LDA + Contextual Information

- STM provides two ways to include contextual information

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - ★ e.g. Democrats talk more about education than Republicans

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - ★ e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - ★ e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - ★ e.g. Democrats are less likely to use the word "life" when talking about abortion than Republicans

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - ★ e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - ★ e.g. Democrats are less likely to use the word "life" when talking about abortion than Republicans
- Including context improves the model:

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - ★ e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - ★ e.g. Democrats are less likely to use the word "life" when talking about abortion than Republicans
- Including context improves the model:
  - more accurate estimation

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - ★ e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - ★ e.g. Democrats are less likely to use the word "life" when talking about abortion than Republicans
- Including context improves the model:
  - more accurate estimation
  - better qualitative interpretability

STM: What this means in pictures

Say you have a lot of people.

Each writes some text

that discuss a few different topics

.7

.3

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

STM: What this means in pictures

Say you have
a lot of people.

Each writes
some text

that discuss a few
different topics

.7

.3

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender

STM: What this means in pictures

Say you have a lot of people.

Each writes some text

that discuss a few different topics

.7

.3

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender

STM: What this means in pictures

Say you have a lot of people.

Each writes some text

that discuss a few different topics

Group A

Group B

Group A

.7

.3

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender
2. The topic proportions to vary by group

STM: What this means in pictures

Say you have a lot of people.

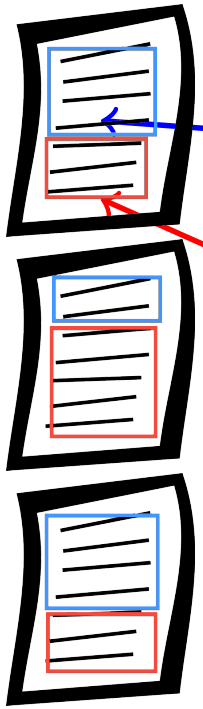
Each writes some text

that discuss a few different topics

Group A

Group B

Group A



Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender
2. The topic proportions to vary by group

Example: Japanese Campaign Manifestos (Catalinac 2011)

- IR question: why is Japan now willing to engage militaristic foreign action?

Example: Japanese Campaign Manifestos (Catalinac 2011)

- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives

Example: Japanese Campaign Manifestos (Catalinac 2011)

- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years

Example: Japanese Campaign Manifestos (Catalinac 2011)

- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
    - That sounds hard

Example: Japanese Campaign Manifestos (Catalinac 2011)

- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible

Example: Japanese Campaign Manifestos (Catalinac 2011)

- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection

Example: Japanese Campaign Manifestos (Catalinac 2011)

- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection
- Latent Dirichlet Allocation (on japanese texts)

Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level

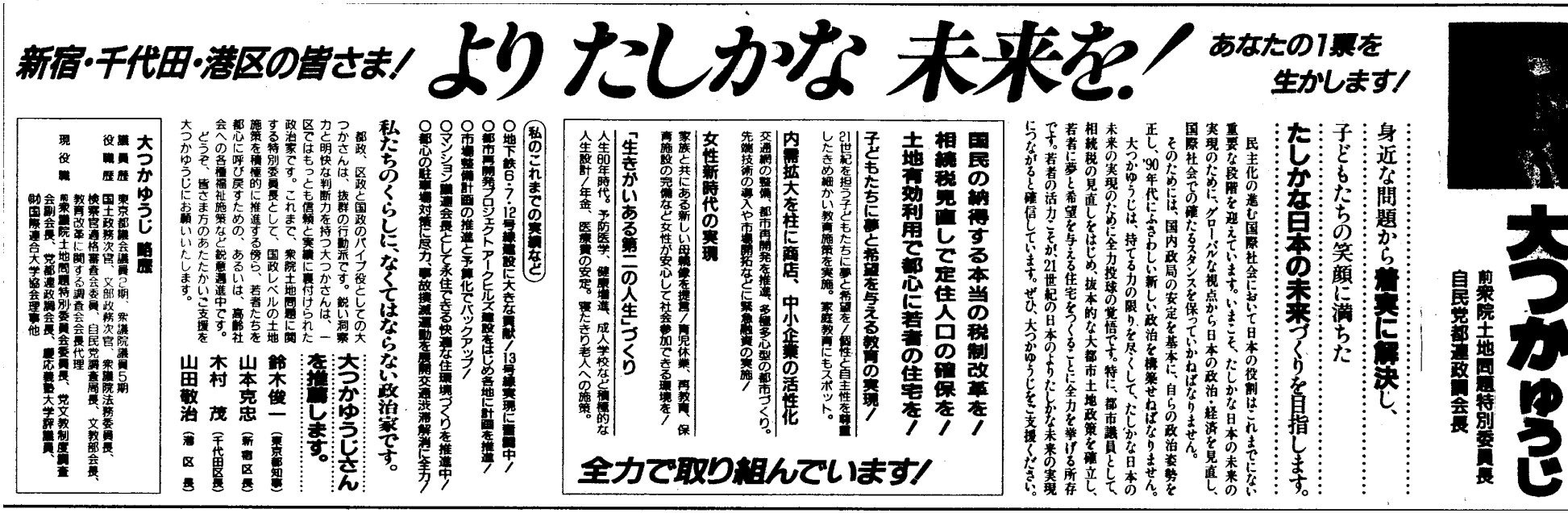
Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet

Japanese Campaign Manifestos (Catalinac 2011)

Typical Manifesto:



新宿・千代田・港区の皆さま！

よりたしかな未来を！

あなたの1票を生かします！

大つかゆうじ

前衆院土地問題特別委員長
自民党都連政調会長

身近な問題から着実に解決し、
子どもたちの笑顔に満ちた
たしかな日本の未来づくりを目指します。

民主化の進む国際社会において日本の役割はこれまでにない重要な段階を迎えています。いまこそ、たしかな日本の未来の実現のために、グローバルな視点から日本の政治・経済を見直し、国際社会での確たるスタンスを保っていかねばなりません。

そのためには、国内政局の安定を基本に、自らの政治姿勢を正し、'90年代にふさわしい新しい政治を構築せねばなりません。

大つかゆうじは、持てる力の限りを尽くして、たしかな日本の未来の実現のために全力投球の覚悟です。特に、都市議員として相続税の見直しをはじめ、抜本的な大都市土地政策を確立し、若者に夢と希望を与える住宅をつくることに全力を挙げる所存です。若者の活力こそが、21世紀の日本のよりたしかな未来の実現につながると確信しています。ぜひ、大つかゆうじをご支援ください。

国民の納得する本当の税制改革を！
相続税見直しで定住人口の確保を！
土地有効利用で都心に若者の住宅を！

子どもたちに夢と希望を与える教育の実現！
21世紀を担う子どもたちに夢と希望を／個性と自主性を尊重したきめ細かい教育施策を実施。家庭教育にもスポット。

内需拡大を柱に商店、中小企業の活性化
交通網の整備、都市再開発を推進、多極多心型の都市づくり。先端技術の導入や市場開拓などに緊急融資の実施。

女性新時代の実現
家族と共にある新しい母親像を提言／育児休業、再教育、保育施設の完備など女性が安心して社会参加できる環境を！

「生きがいある第二の人生」づくり
人生80年時代。予防医学、健康増進、成人学校など積極的な人生設計／年金、医療費の安定。寝たきり老人への施策。

全力で取り組んでいます！

（私のこれまでの実績など）

○地下鉄6・7・12号線建設に大きな貢献／13号線実現に奮闘中！
○都市再開発プロジェクトアークヒルズ建設をはじめ各地に計画を推進！
○市場整備計画の推進と予算化でバックアップ！
○マンション建設促進委員長として永住できる快適な住環境づくりを推進中！
○都心の駐車場対策に応力、事故撲滅運動を展開交通洪滞解消に全力！

私たちのくらしに、なくてはならない政治家です。

都政、区政と国政のパイプ役としての大つかさんは、抜群の行動派です。鋭い洞察力と明快な判断力を持つ大つかさんは、一区ではもっとも信頼と実績に裏付けられた政治家です。これまで、衆院土地問題に関する特別委員長として、国政レベルの土地施策を積極的に推進する傍ら、若者たちを都心に呼び戻すための、あるいは、高齢社会への各種福祉施策など鋭意邁進中です。

どうぞ、皆さま方のあたたかいご支援を大つかゆうじにお願いいたします。

大つかゆうじさんを推薦します。

鈴木俊一（東京都知事）
山本克忠（新宿区長）
木村　茂（千代田区長）
山田敬治（港区長）

大つかゆうじ 略歴

議員歴　東京都議会議員3期、衆議院議員5期
役職歴　国土政務次官、文部政務次官、衆議院法務委員長、検察官適格審査会委員、自民党調査会長、文教部会長、教育改革に関する調査会会長代理
現役職　前衆議院土地問題特別委員会委員長、党文教制度調査会副会長、党都連政調会長、慶応義塾大学評議員、㈱国際連合大学協会理事他

Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009

Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
    - Available only at district level

Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm

Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM

Japanese Campaign Manifestos (Catalinac 2011)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM
- Harder for Japanese

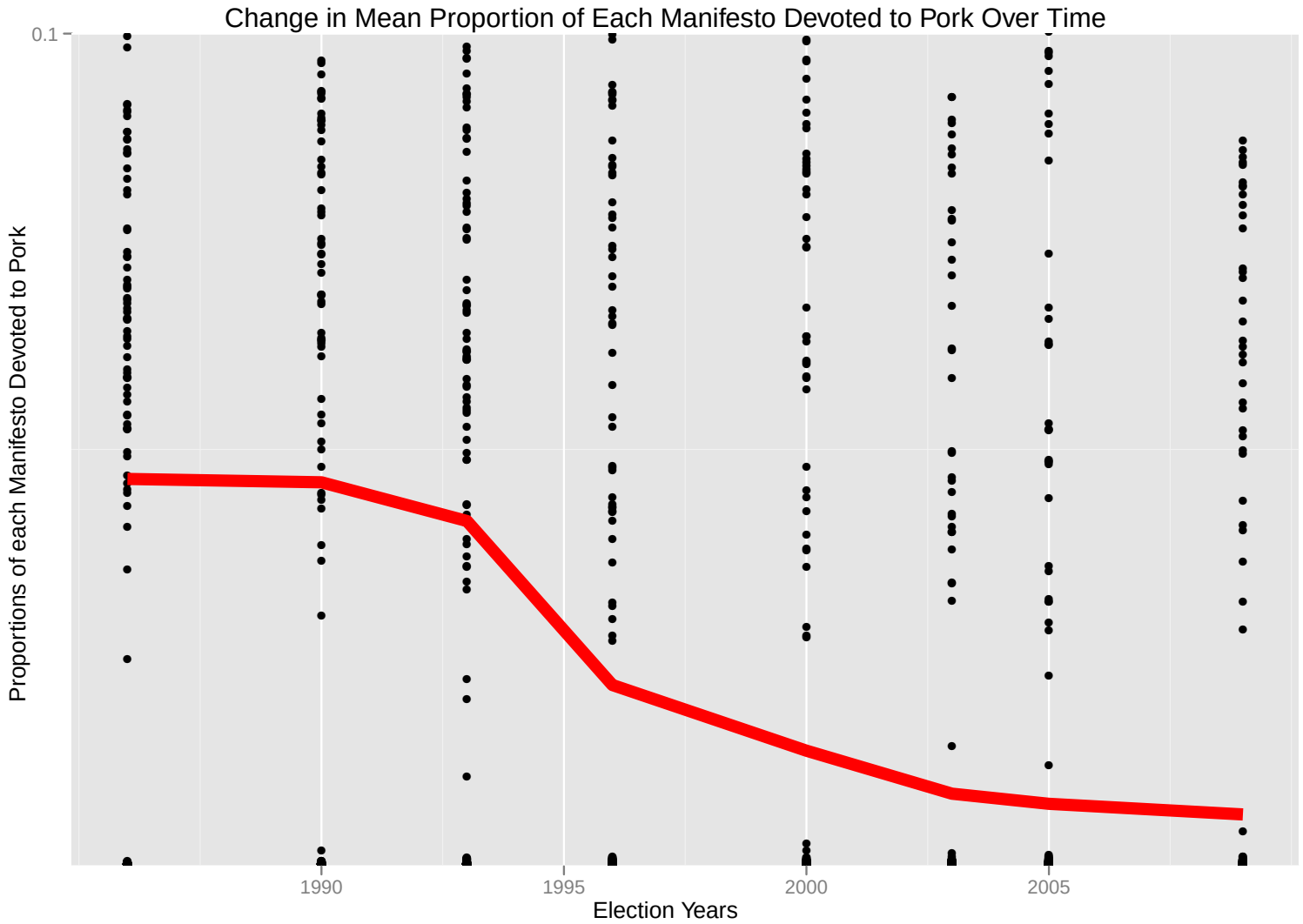
Japanese Campaign Manifestos (Catalinac 2014)

- Applies Vanilla LDA
- Output: topics (with Japanese characters)

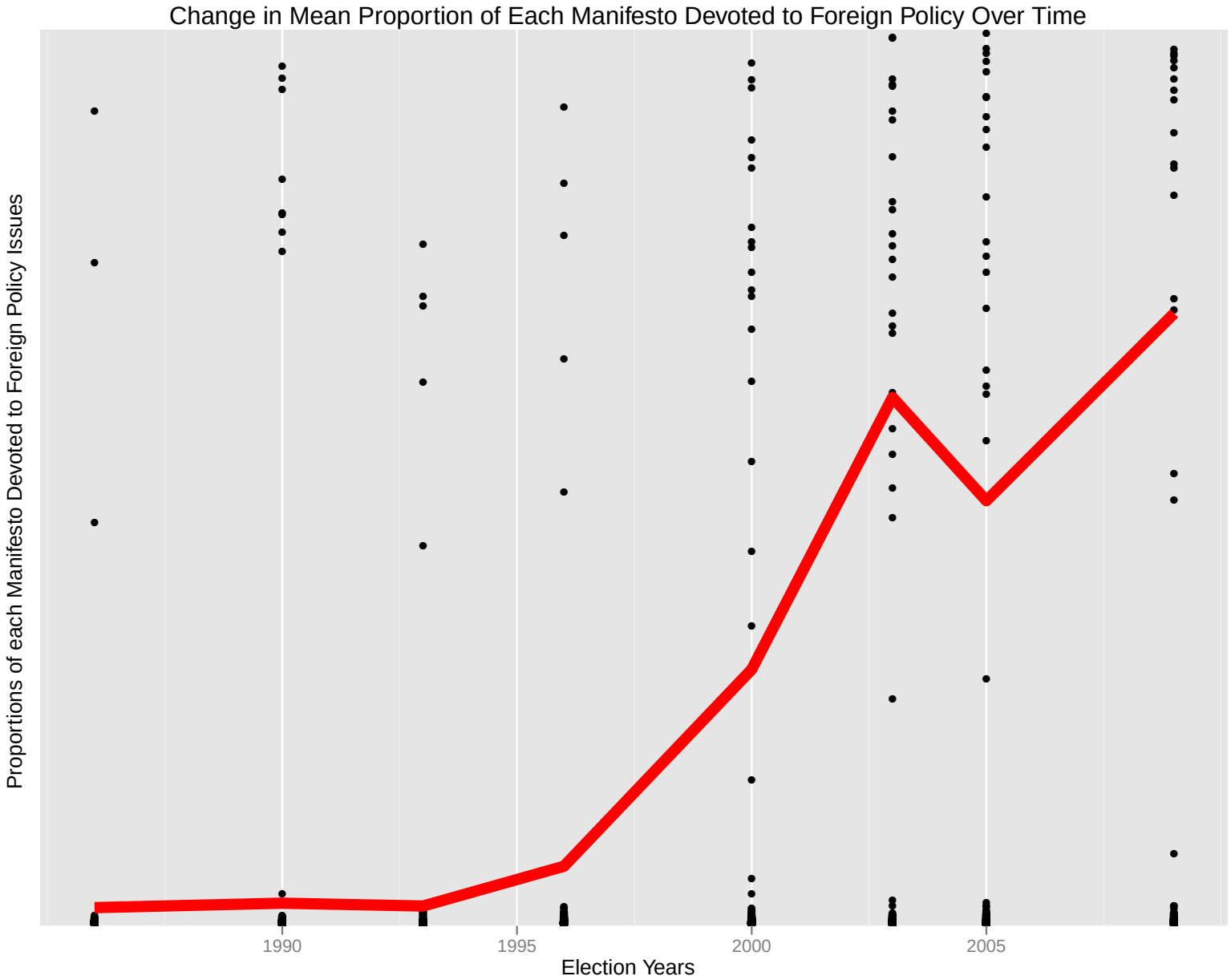
Japanese Campaign Manifestos (Catalinac 2011)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic
改革	年金	推進	区	政治	日本
郵政	円	整備	政策	改革	国
民営	廃止	図る	地域	国民	外交
小泉	改革	つとめる	まち	企業	国家
構造	兆	社会	鹿児島	自民党	社会
政府	実現	対策	全力	日本	国民
官	無駄	振興	選挙	共産党	保障
推進	日本	充実	国政	献金	安全
民	増税	促進	作り	金権	地域
自民党	削減	安定	横浜	党	拉致
日本	一元化	確立	対策	選挙	経済
制度	政権	企業	中小	禁止	守る
民間	子供	実現	発電	憲法	問題
年金	地域	中小	推進	腐敗	北朝鮮
実現	ひと	育成	エネルギー	団体	教育
進める	サラリーマン	制度	企業	区	責任
断行	制度	政治	声	ソ連	力
地方	議員	地域	実現	守る	創る
止める	金	福祉	活性	平和	安心
保障	民主党	事業	自民党	円	目指す
財政	年間	改革	地方	反対	誇り
作る	一掃	確保	尽くす	真	憲法
賛成	郵政	強化	商店	是正	可能
社会	道路	教育	いかす	一掃	道
国民	交代	施設	全国	悪政	未来
公務員	社会保険庁	生活	政党	抜本	ひと
力	月額	支援	ひと	定数	再生
経済	手当	環境	支援	政党	将来
国	談合	発展	経済	金丸	解決
安心	支援	施策	福祉	改悪	基本
Postal privatization	Reducing Wasteful Public Spending	Pork for the District	Policies for the district	Political Reform	Nation

Japanese Campaign Manifestos (Catalinac 2011)



Japanese Campaign Manifestos (Catalinac 2011)



Albertson and Gadarian: Anxiety and Immigration

Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- "... When you think about immigration, what makes you worried?..."

Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- "... When you think about immigration, what makes you worried?..."
- "... When you think about immigration, what do you think of?..."

Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- "... When you think about immigration, what makes you worried?..."
- "... When you think about immigration, what do you think of?..."

Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- "... When you think about immigration, what makes you worried?..."
- "... When you think about immigration, what do you think of?..."

Original analysis:

Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- "... When you think about immigration, what makes you worried?..."
- "... When you think about immigration, what do you think of?..."

Original analysis:

- Human coders using pre-established coding categories (Fear, Anger, Enthusiasm)

Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- "... When you think about immigration, what makes you worried?..."
- "... When you think about immigration, what do you think of?..."

Original analysis:

- Human coders using pre-established coding categories (Fear, Anger, Enthusiasm)
- Treatment had impact on Fear and Anger.

Topics

- Topic 1

Topics

- Topic 1

Topics

- Topic 1
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag

Topics

- Topic 1
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - "problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals."

Topics

- Topic 1
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - "problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals."
- Topic 2

Topics

- Topic 1
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - "problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals."
- Topic 2
  - immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker

Topics

- Topic 1
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - "problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals."
- Topic 2
  - immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker
  - "i worry about the republican party doing something very stupid. this country was built on immigration, to deny anyone access to citizenship is unconstitutional. what happened to give me your poor, sick, and tired?"

Effects on Topic 1

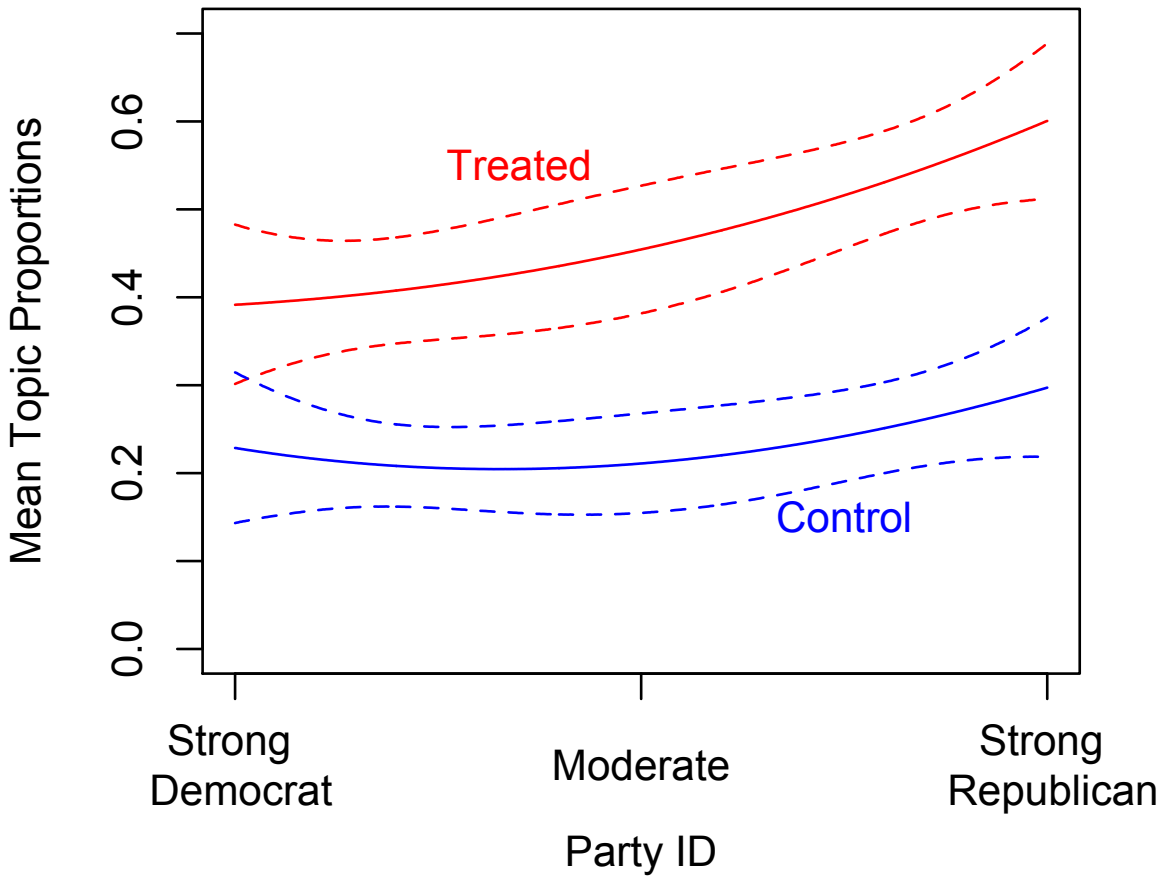
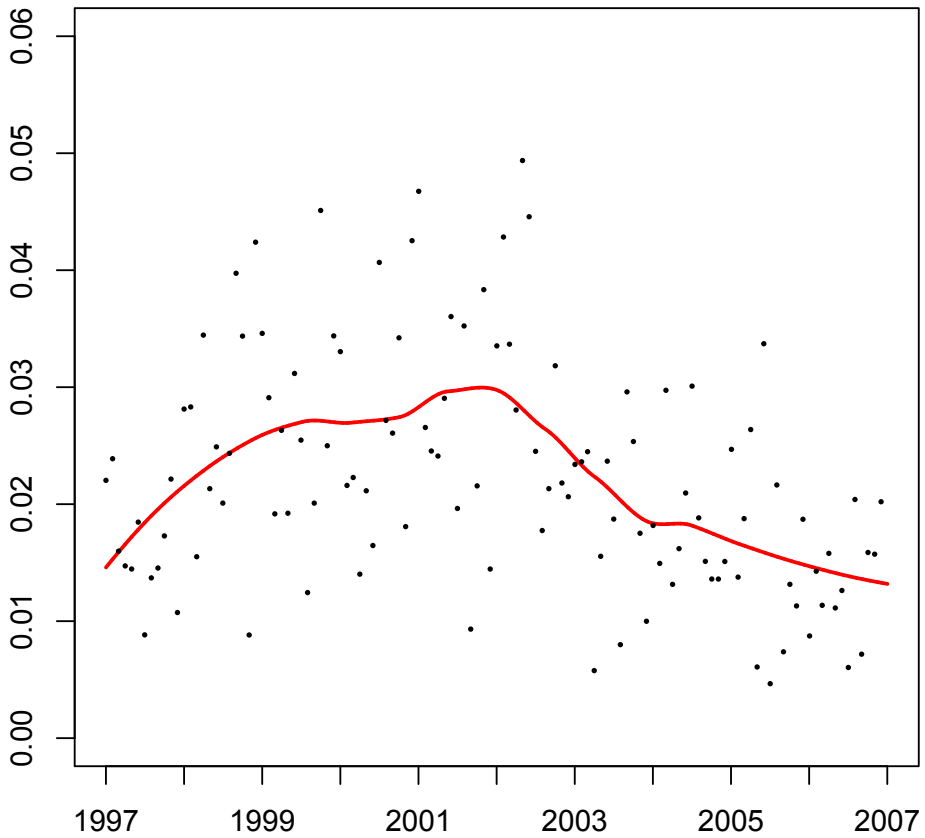


Figure: Topic 1.

Different Newspapers, Different Perspectives



Associated Press

polic protest gong arrest
falun detain author releas
follow prison inform offic
wang crackdown activist
movement sentenc center squar
china demonstr zhang
tiananmen dissid investig

Xinhua

polic illeg smuggl public
gong investig arrest offic
crimin cult falun immigr
suspect case custom organ
author depart order china
properti told accord sentenc
terrorist

Different Newspapers, Different Perspectives