Political Science 452: Text as Data

Justin Grimmer

Assistant Professor
Department of Political Science
Stanford University

May 11th, 2011

Where We've Been, Where We're Going

- Class 1: Finding Text Data
- Class 2: Representing Texts Quantitatively
- Class 3: Dictionary Methods for Classification
- Class 4: Comparing Language Across Groups
- Class 5: Texts in Space
- Class 6: Clustering
- Class 7: Topic models
- Class 8: Supervised methods for classification
- Class 9: Ensemble methods for classification
- Class 10: Scaling Speech

Entropy Explanation (Hanna Wallach Slides)

- Probability and Information are intimately related
- Less probable events $\rightarrow$ more information
- More certain something will occur, less information you gain knowing it occurred
- Focus on the occurrence of binary event A

Basic unit of information built around event A for which we are maximally uncertain:

$$\mathrm{P}[A] \quad = \quad P[\neg A] = 1/2$$

Entropy Explanation (Hanna Wallach Slides)

Desired properties

- Information, Probability: Inversely related
- Certain event will occur and it does: no information gained
- Certain event will not occur and it does: infinite information gained
- Maximally uncertain: we should gain one unit of information by learning that A or $\neg A$ occurred

Entropy Explanation (Hanna Wallach Slides)

Information in event A is then,

$$I(A) = \log_2 \frac{1}{\mathrm{P}[A]}$$

And for a series of disjoint events, the entropy is

$$H(A_1, \ldots, A_n) = \sum_{i=1}^{N} \mathrm{P}[A_i] \log_2 \frac{1}{\mathrm{P}[A_i]}$$

Recall: use entropy to describe how well words separate classes (Week 4):
Conditional entropy(w): Information that remains after condition on w

$$\text{Mutual information}(w) = \text{Entropy} - \text{Conditional Entropy}(w)$$

Measures the reduction in information$\rightsquigarrow$ greater reduction, less information, w is a better predictor.

Clustering

Last week: measures of similarity between documents.

- Place documents in space
- Measure similarity, dissimilarity of documents

This week: identify groups of similar documents

Fully Automated Clustering Algorithms:

- Task: partition documents
  - Mutual exclusive
  - Exhaustive
  - Set of Groupings
- Task name: Clustering
  - Estimate: categories
  - Estimate: each document's category
- Label Clusters in Clustering (Week 4 Methods!)
- How to use clustering methods: (THINK!)
  - Tune clustering methods to problem (Discuss more next week, virtue of statistical models)

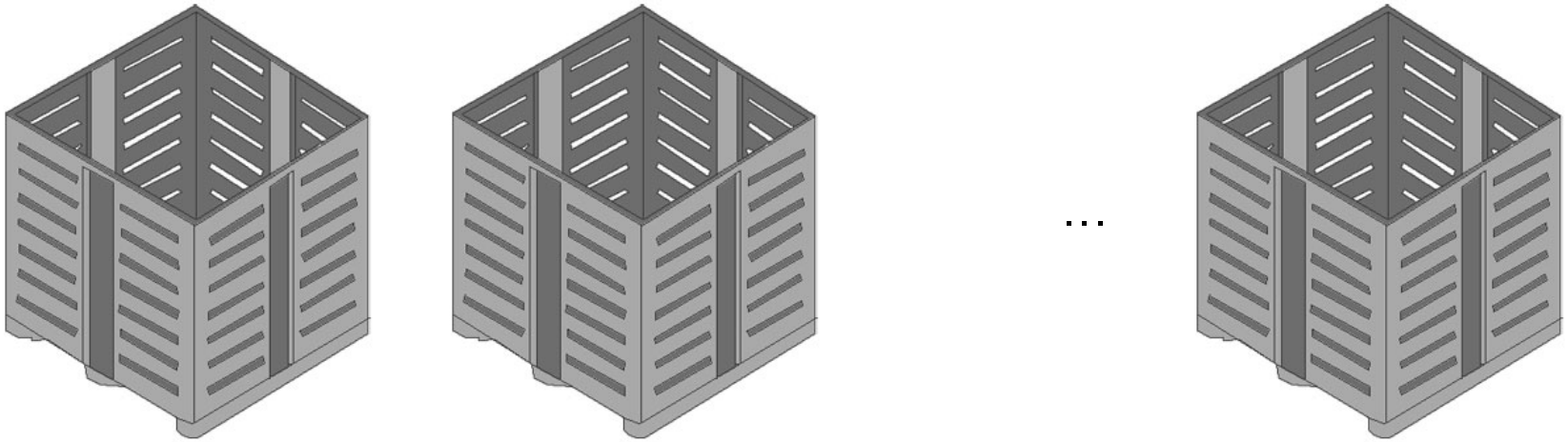
Computer Assisted Clustering Algorithms (Grimmer and King 2011)

Perspective 1: Supervised Methods

Document 1 Document 2 $\cdots$ Document N

Perspective 1: Supervised Methods

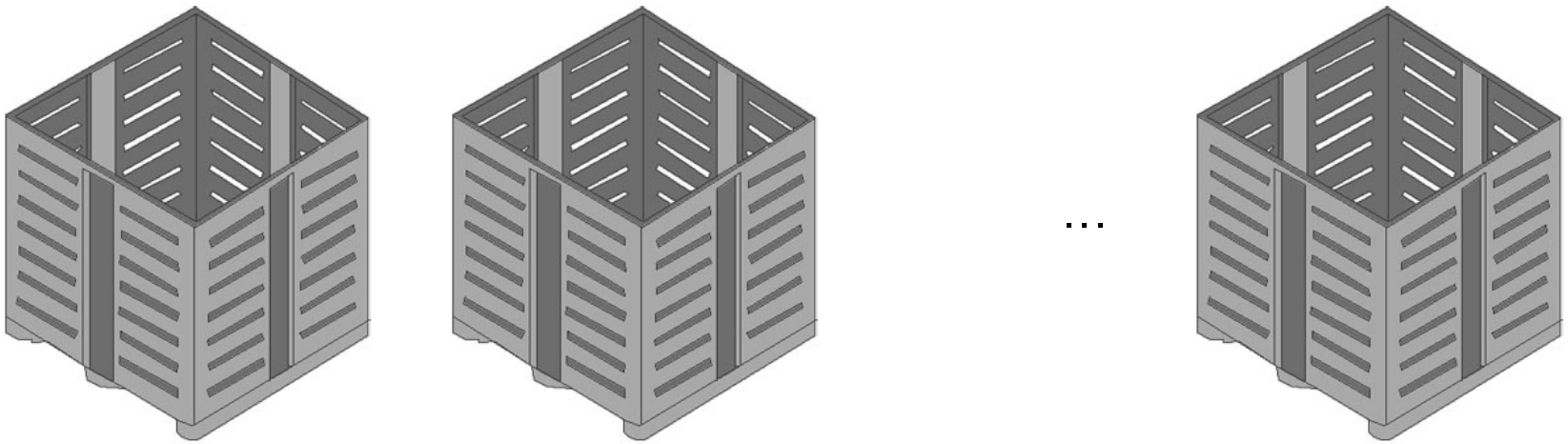
Document 1 Document 2 ··· Document N



Bins Known, Bin Assignment Estimated

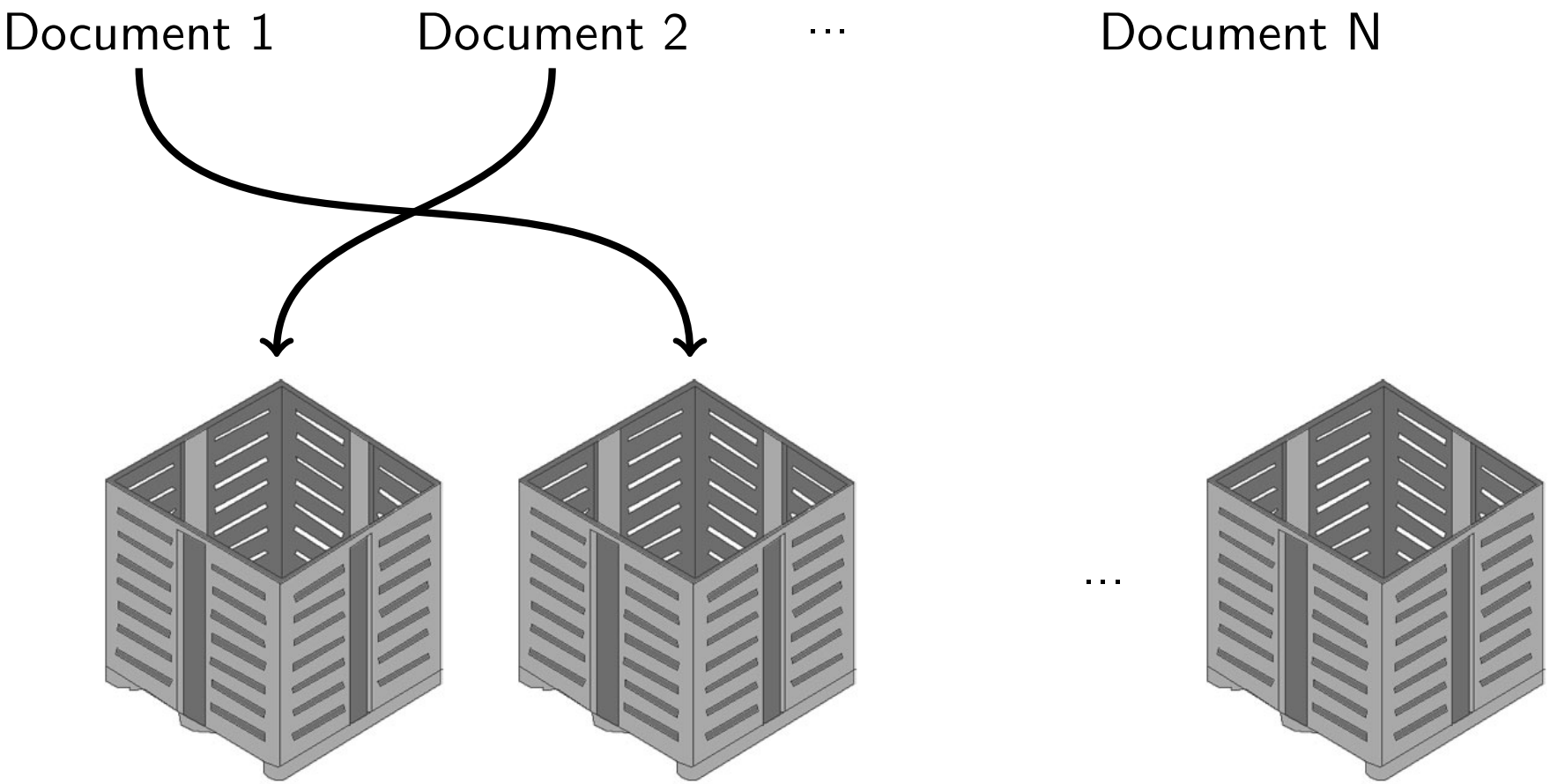
Perspective 1: Supervised Methods

Document 1 Document 2 ··· Document N



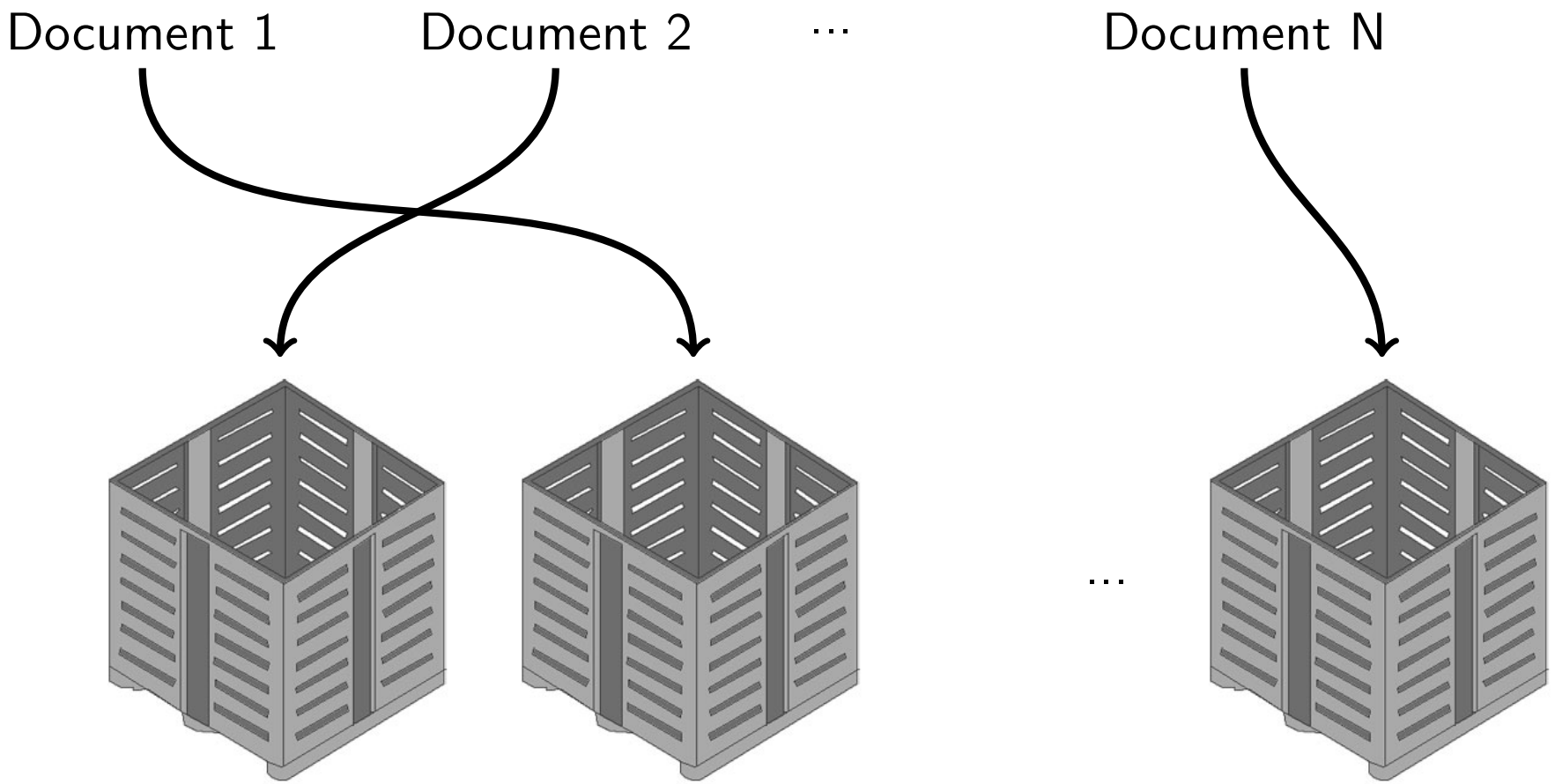
Bins Known, Bin Assignment Estimated

Perspective 1: Supervised Methods



Bins Known, Bin Assignment Estimated

Perspective 1: Supervised Methods



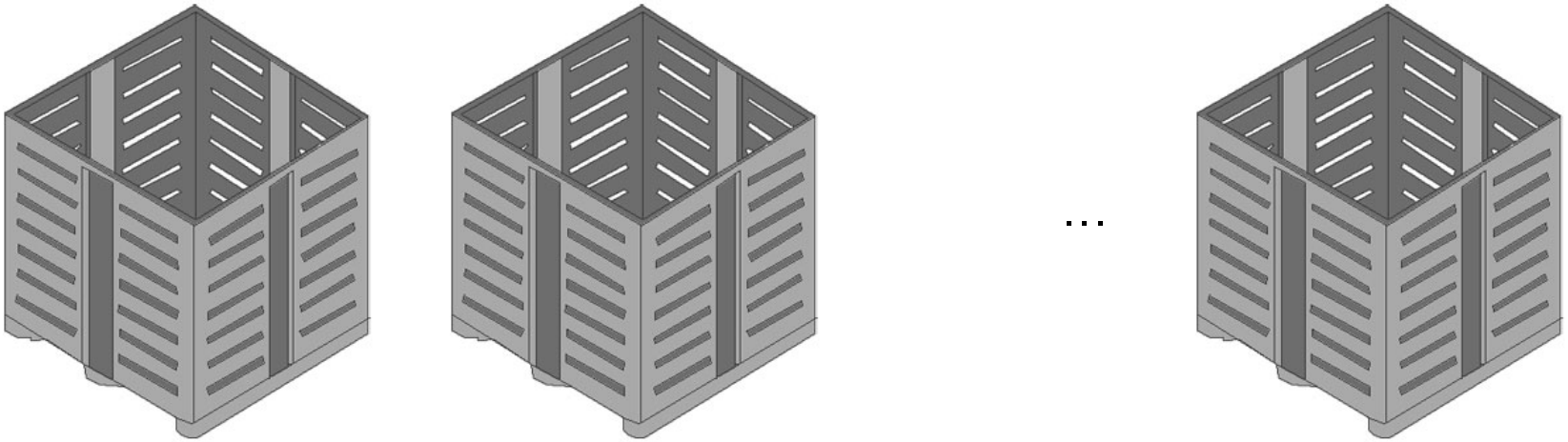
Bins Known, Bin Assignment Estimated

Perspective 1: Clustering

Document 1 Document 2 $\cdots$ Document N

Perspective 1: Clustering

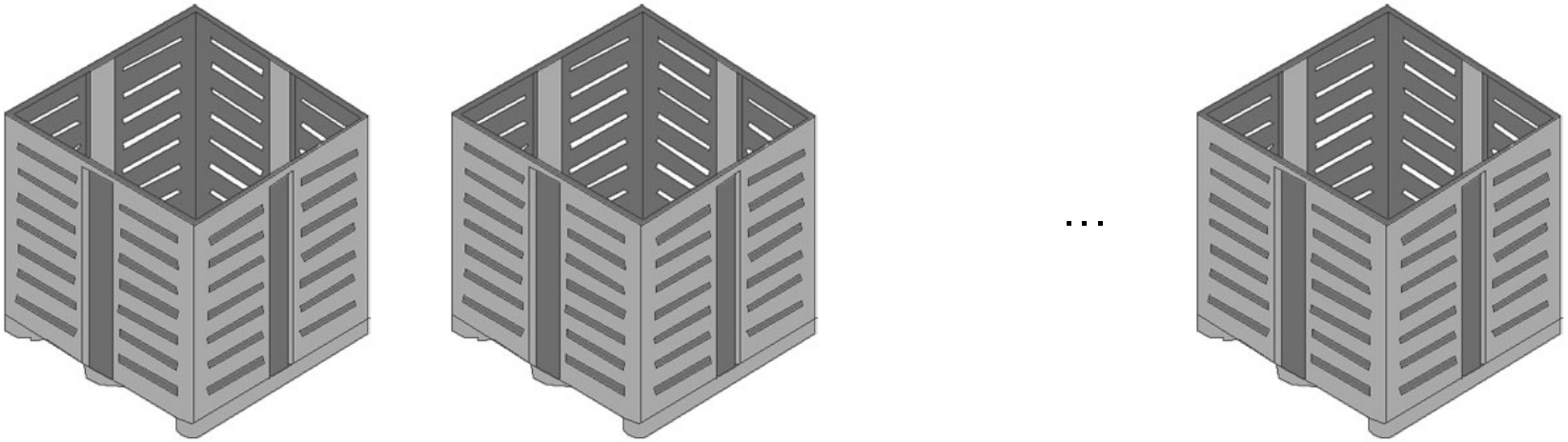
Document 1 Document 2 ··· Document N



Bins and Bin Assignments Estimated

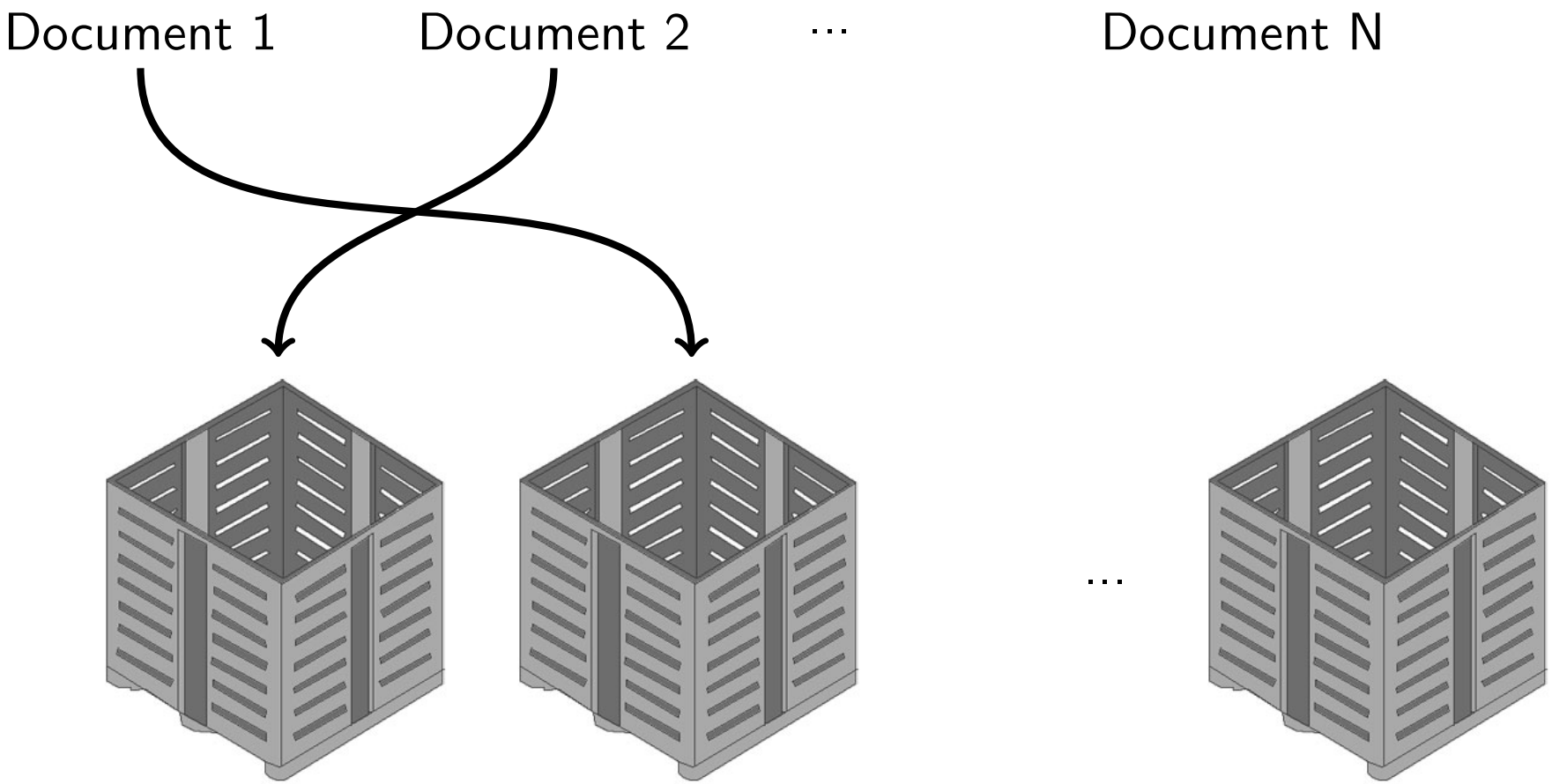
Perspective 1: Clustering

Document 1 Document 2 ··· Document N



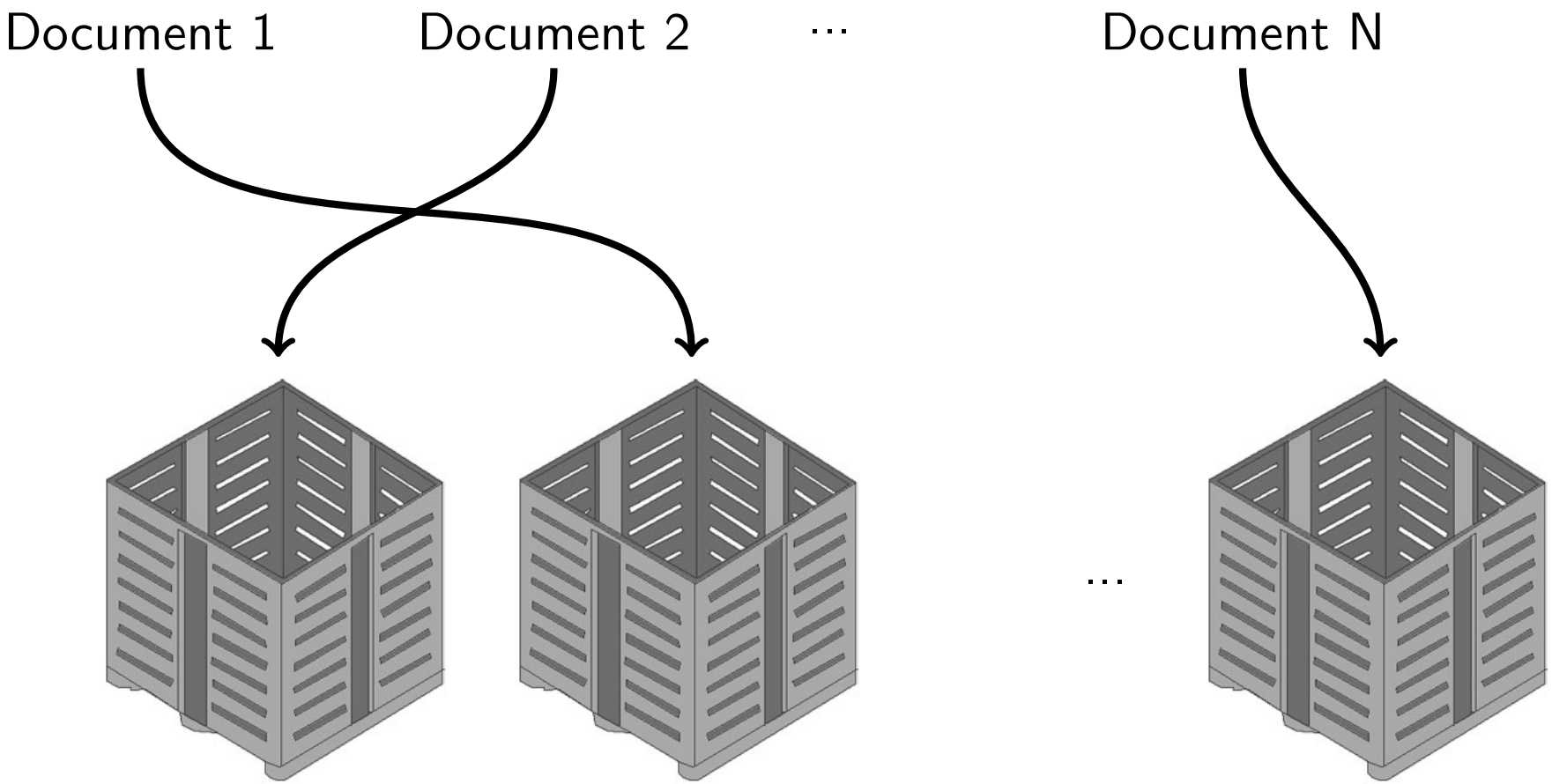
Bins and Bin Assignments Estimated

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

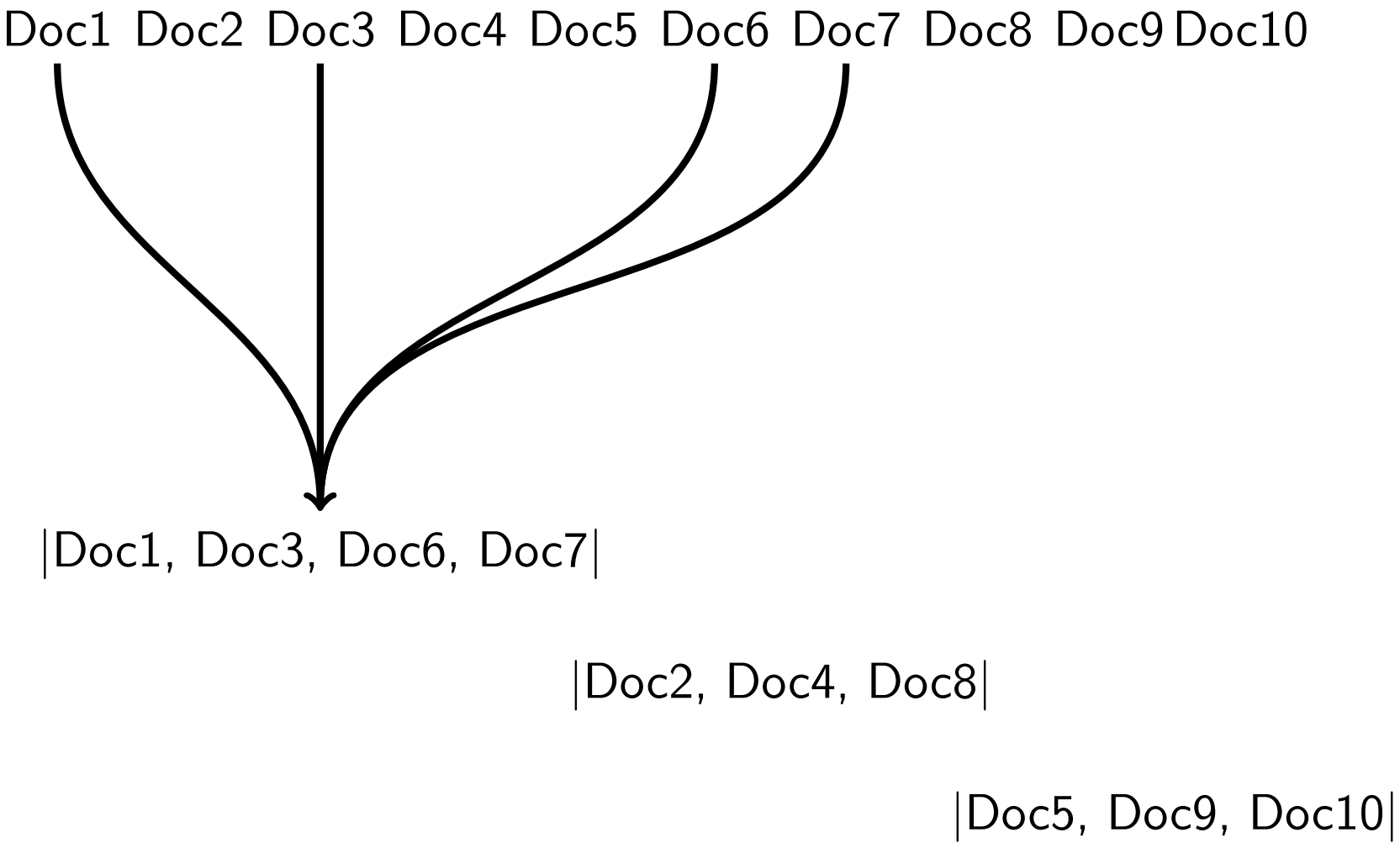
Doc1 Doc2 Doc3 Doc4 Doc5 Doc6 Doc7 Doc8 Doc9 Doc10

|Doc1, Doc3, Doc6, Doc7|

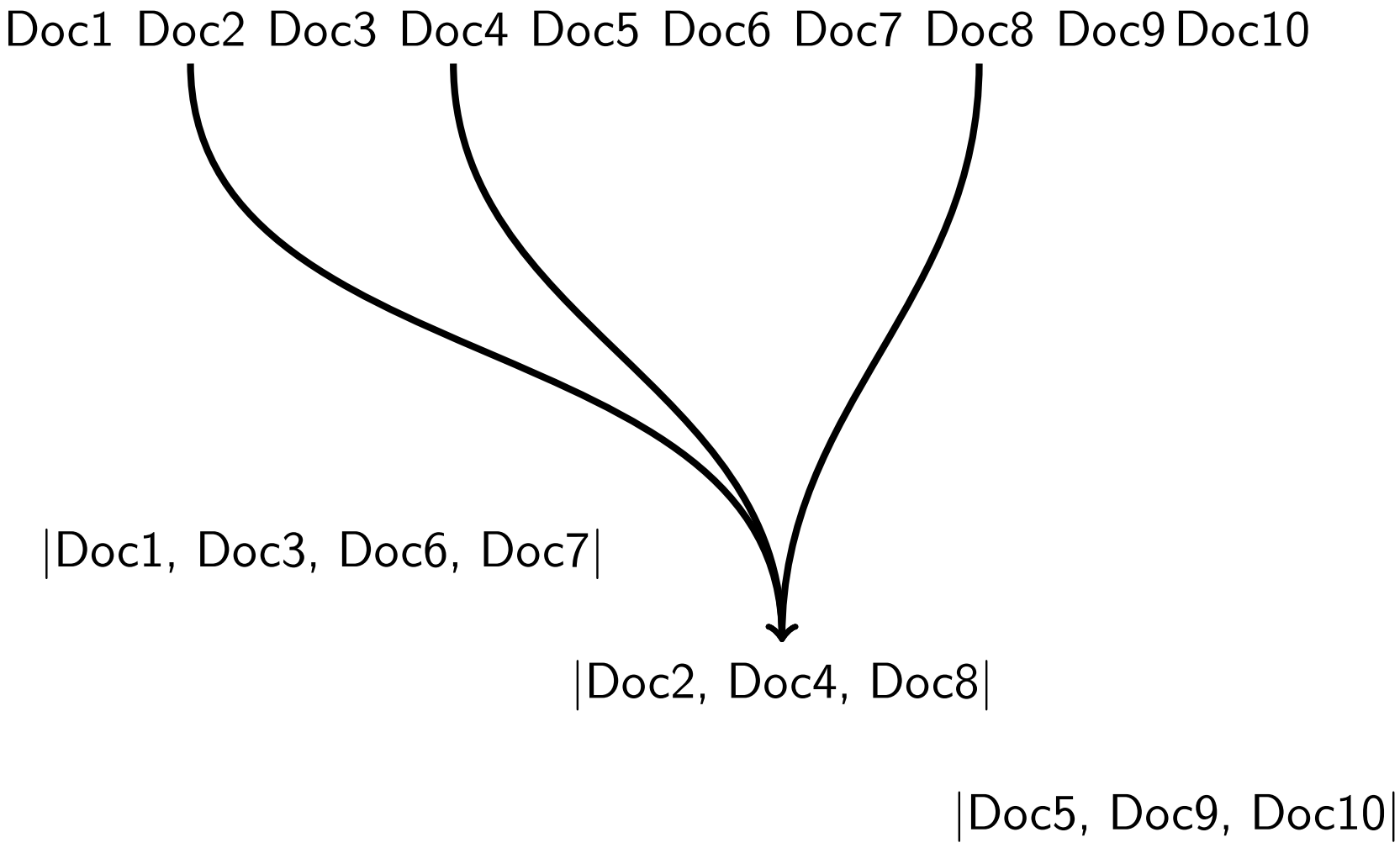
|Doc2, Doc4, Doc8|

|Doc5, Doc9, Doc10|

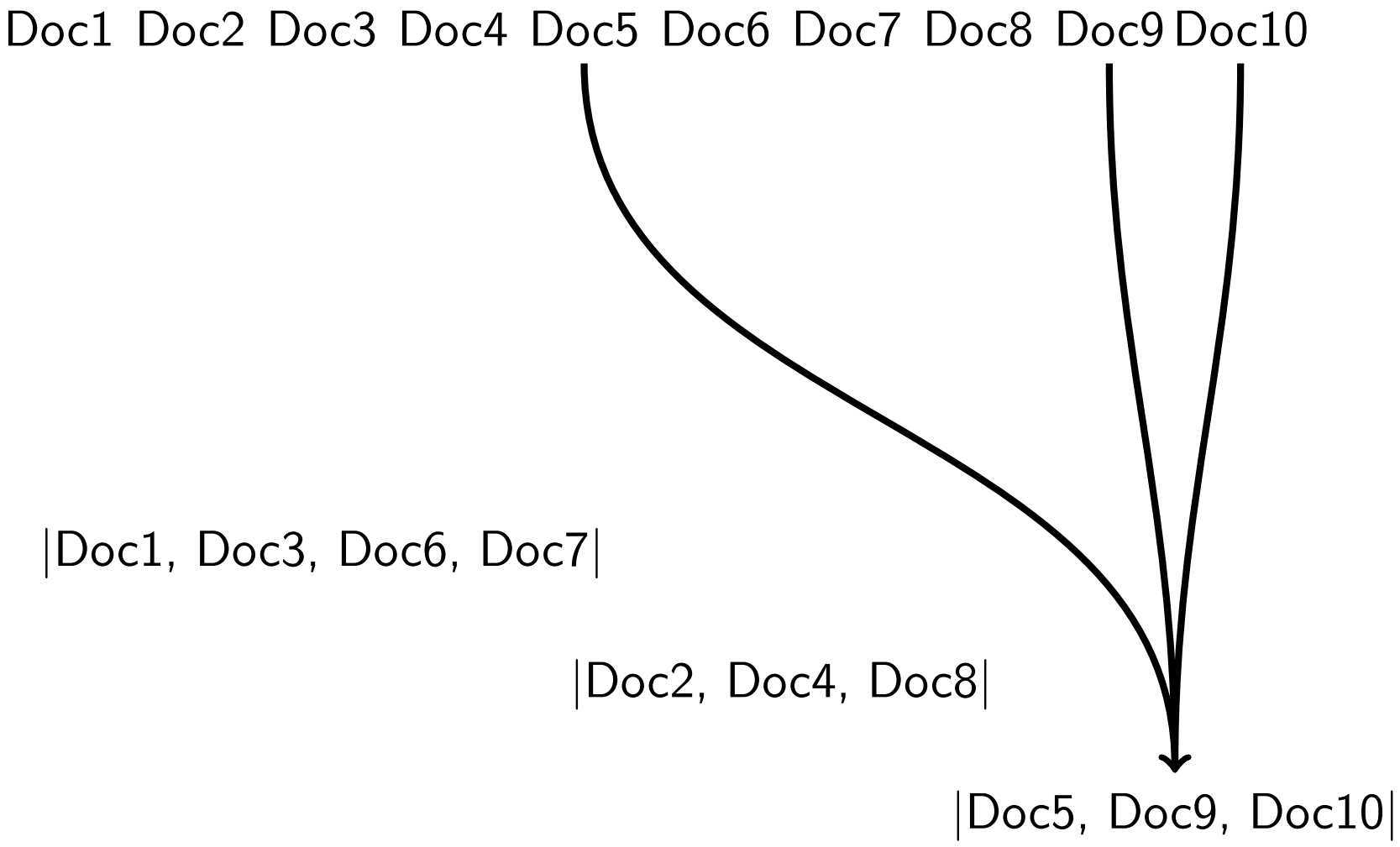
Perspective 2: Clustering



Perspective 2: Clustering

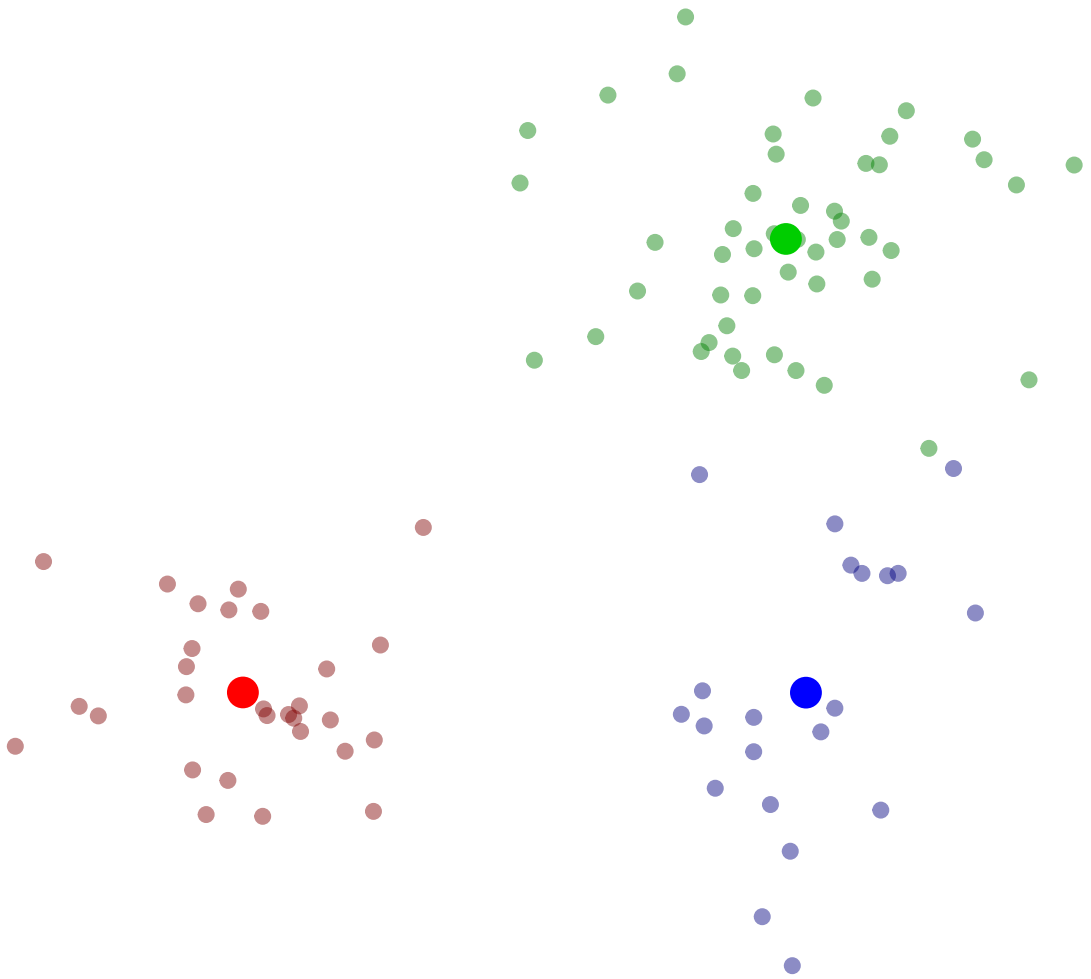


Perspective 2: Clustering



Perspective 3

Perspective 3



Perspective 4

Clustering as Compression:

Perspective 4

Clustering as Compression:
Identify groups of documents, clusters:

Perspective 4

Clustering as Compression:
Identify groups of documents, clusters:

1) Have high within group similarity

Perspective 4

Clustering as Compression:
Identify groups of documents, clusters:

1) Have high within group similarity
2) Have low across group similarity

Perspective 4

Clustering as Compression:
Identify groups of documents, clusters:

1) Have high within group similarity
2) Have low across group similarity

Compression: Retain only cluster label for documents in same group

Perspective 5

Clustering as Discovery

Perspective 5

Clustering as Discovery

- When we analyze texts (data) we have some idea how to organize them

Perspective 5

Clustering as Discovery

- When we analyze texts (data) we have some idea how to organize them
    - Liberals/Democrats

Perspective 5

Clustering as Discovery

- When we analyze texts (data) we have some idea how to organize them
    - Liberals/Democrats
    - Democracy/Autocracy

Perspective 5

Clustering as Discovery

- When we analyze texts (data) we have some idea how to organize them
    - Liberals/Democrats
    - Democracy/Autocracy
    - Assistant/Associate

Perspective 5

Clustering as Discovery

- When we analyze texts (data) we have some idea how to organize them
  - Liberals/Democrats
  - Democracy/Autocracy
  - Assistant/Associate
- How do we formulate new ways to organize texts?

Perspective 5

Clustering as Discovery

- When we analyze texts (data) we have some idea how to organize them
  - Liberals/Democrats
  - Democracy/Autocracy
  - Assistant/Associate
- How do we formulate new ways to organize texts?
  - Clustering methods suggest new (model and data driven) ways to organize texts

Perspective 5

Clustering as Discovery

- When we analyze texts (data) we have some idea how to organize them
  - Liberals/Democrats
  - Democracy/Autocracy
  - Assistant/Associate
- 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 politics

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:

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

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

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

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly)
- Define objective function
  - Domain (things you put in the function): space of clusterings
  - Range (thing that comes out of function): measure of clustering's performance
- Use approximate inference/optimization algorithm to identify optimal solution

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly) $\leftarrow$ Discussed extensively last week
- Define objective function
  - Domain (things you put in the function): space of clusterings
  - Range (thing that comes out of function): measure of clustering's performance
- Use approximate inference/optimization algorithm to identify optimal solution

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly)
- Define objective function $\leftarrow$ We will discuss extensively today
  - Domain (things you put in the function): space of clusterings
  - Range (thing that comes out of function): measure of clustering's performance
- Use approximate inference/optimization algorithm to identify optimal solution

Estimating Clustering: Data and Assumptions

Steps common across Fully Automated Clustering methods

- Assume similarity/dissimilarity between objects (Some methods assume implicitly)
- Define objective function
  - Domain (things you put in the function): space of clusterings
  - Range (thing that comes out of function): measure of clustering's performance
- Use approximate inference/optimization algorithm to identify optimal solution $\leftarrow$ Huge search space, very difficult (and interesting!) problem, only hinted at here

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

   $\mathbf{r}_j = (r_{1j}, r_{2j}, \ldots, r_{Nj})$

   $\mathbf{r} = (\mathbf{r}_1^{'}, \mathbf{r}_2^{'}, \ldots, \mathbf{r}_K^{'})$ ($N \times K$ matrix)

   Note: Same information in $\mathbf{r}$ and $\mathbf{c}$

2) For each cluster j

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

   $\boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{Mj})$

Notation. Representation of document i:

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iM})$$

Specifying the Method

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

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

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

- If $K = N$ $f(r^*, \boldsymbol{\mu}^*, \mathbf{y}) = 0$ (Minimum)
  - Each observation in own cluster
  - $\boldsymbol{\mu}_i = \mathbf{y}_i$
- If $K = 1$, $f(r^*, \boldsymbol{\mu}^*, \mathbf{y}) = N \times \sigma^2$
  - Each observation in one cluster
  - Center: average of documents

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 $\mathbf{r}^t$
- Conditional on $\mathbf{r}^t$, choose $\boldsymbol{\mu}^t$

Repeat until convergence, measured as change in f.

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

Specifying the Method

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

Algorithm for estimation:
Begin: initialize $\boldsymbol{\mu}_1^{t-1}, \boldsymbol{\mu}_2^{t-1}, \ldots, \boldsymbol{\mu}_K^{t-1}$
Choose $\mathbf{r}^t$

$$r_{ij}^t = \begin{cases} 1 \text{ if } j = \arg\min_k \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \\ 0 \text{ otherwise ,} \end{cases} .$$

In words: Assign each document $\mathbf{y}_i$ to the closest center $\boldsymbol{\mu}_k$

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

Conditional on $\mathbf{r}^t$, choose $\boldsymbol{\mu}^t$
Let's focus on $\boldsymbol{\mu}_k$

$$f(\mathbf{r}, \boldsymbol{\mu}_k, \mathbf{y})_k = \sum_{i=1}^{N} r_{ik} \left(\sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right)$$

Focus on just μ_{km}

$$f(\mathbf{r}, \mu_{km}, \mathbf{y})_{km} = \sum_{i=1}^{N} r_{ik}(y_{im} - \mu_{km})^2$$

Quadratic: take derivative, set equal to zero (second derivative test works)

$$\begin{aligned}
\frac{\partial f(\mathbf{r}, \mu_{km}, \mathbf{y})_{km}}{\partial \mu_{km}} &= -2\sum_{i=1}^{N} r_{ik}(y_{im} - \mu_{km}) \\
2\sum_{i=1}^{N} r_{ik}(y_{im} - \mu_{km}^t) &= 0 \\
\sum_{i=1}^{N} r_{ik} y_{im} - \mu_{km}^t \sum_{i=1}^{N} r_{ik} &= 0 \\
\frac{\sum_{i=1}^{N} r_{ik} y_{im}}{\sum_{i=1}^{N} r_{ik}} &= \mu_{km}^t
\end{aligned}$$

$$\boldsymbol{\mu}_k^t = \frac{\sum_{i=1}^N r_{ik}\mathbf{y}_i}{\sum_{i=1}^N r_{ik}}$$

In words:

- $\boldsymbol{\mu}_k^t$ is the average of documents assigned to the kth cluster

Algorithm, In Words

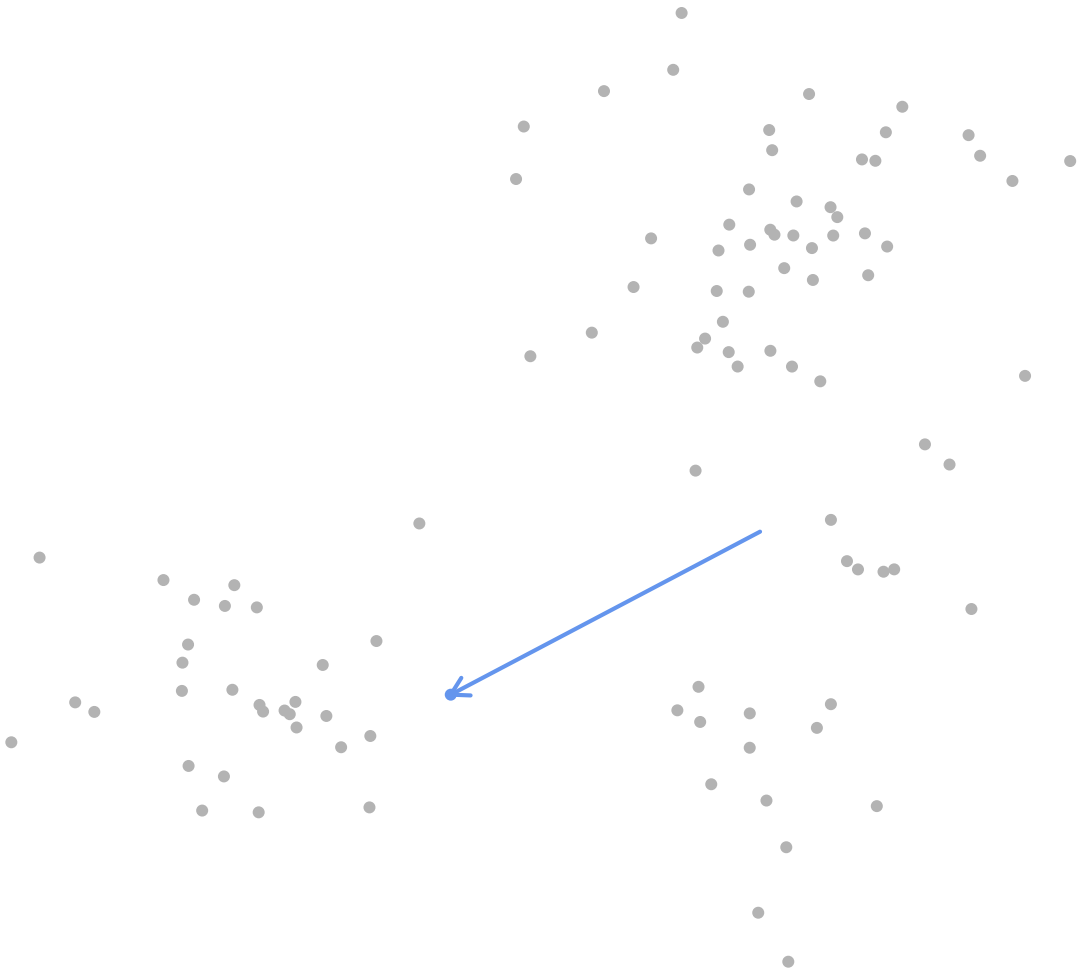
- Conditional on center estimates, assign documents to closest cluster centers
- Conditional on document assignments, cluster centers are averages of documents assigned to the cluster

Expectation-Maximization (EM) [connection guarantees convergence]

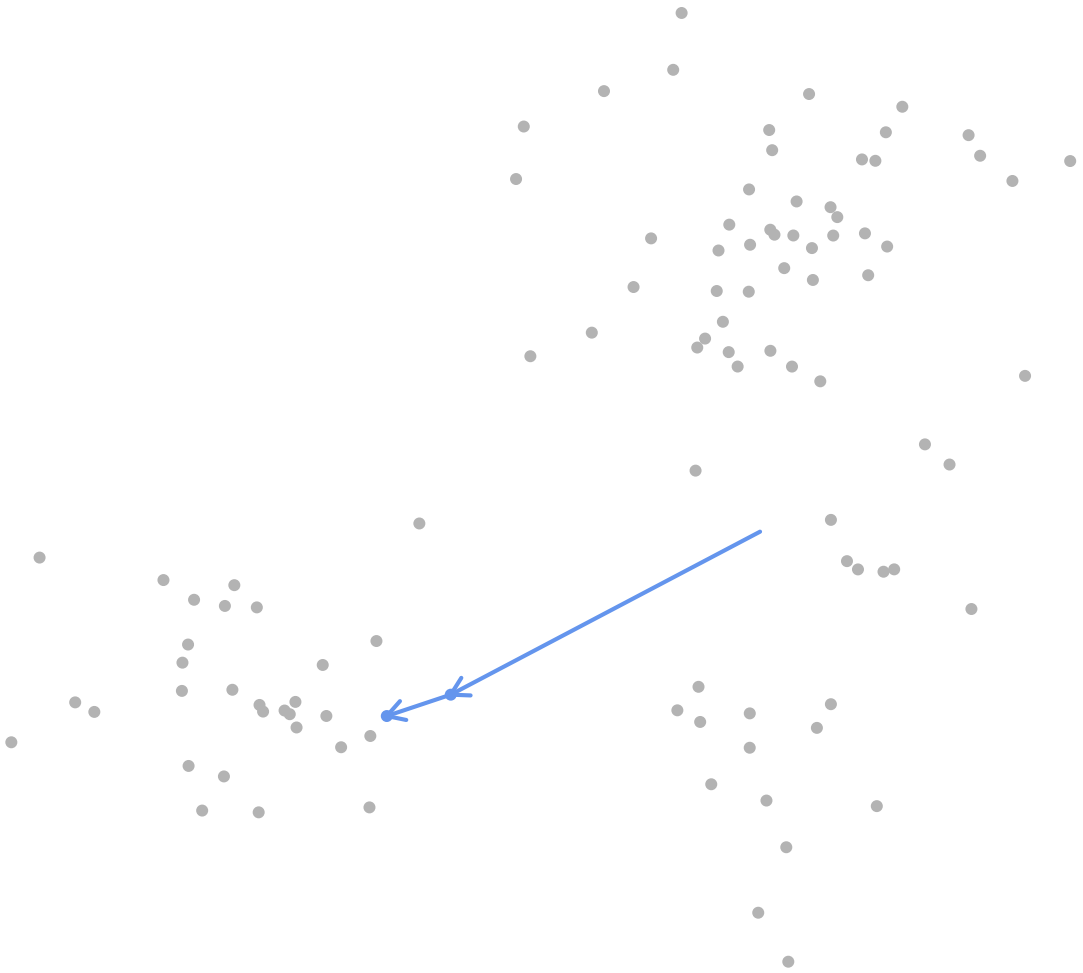
- Estimation of $r \rightsquigarrow$ Expectation step (data augmentation)
- Estimation of $\boldsymbol{\mu}_k \rightsquigarrow$ Maximization Step

Visual Example

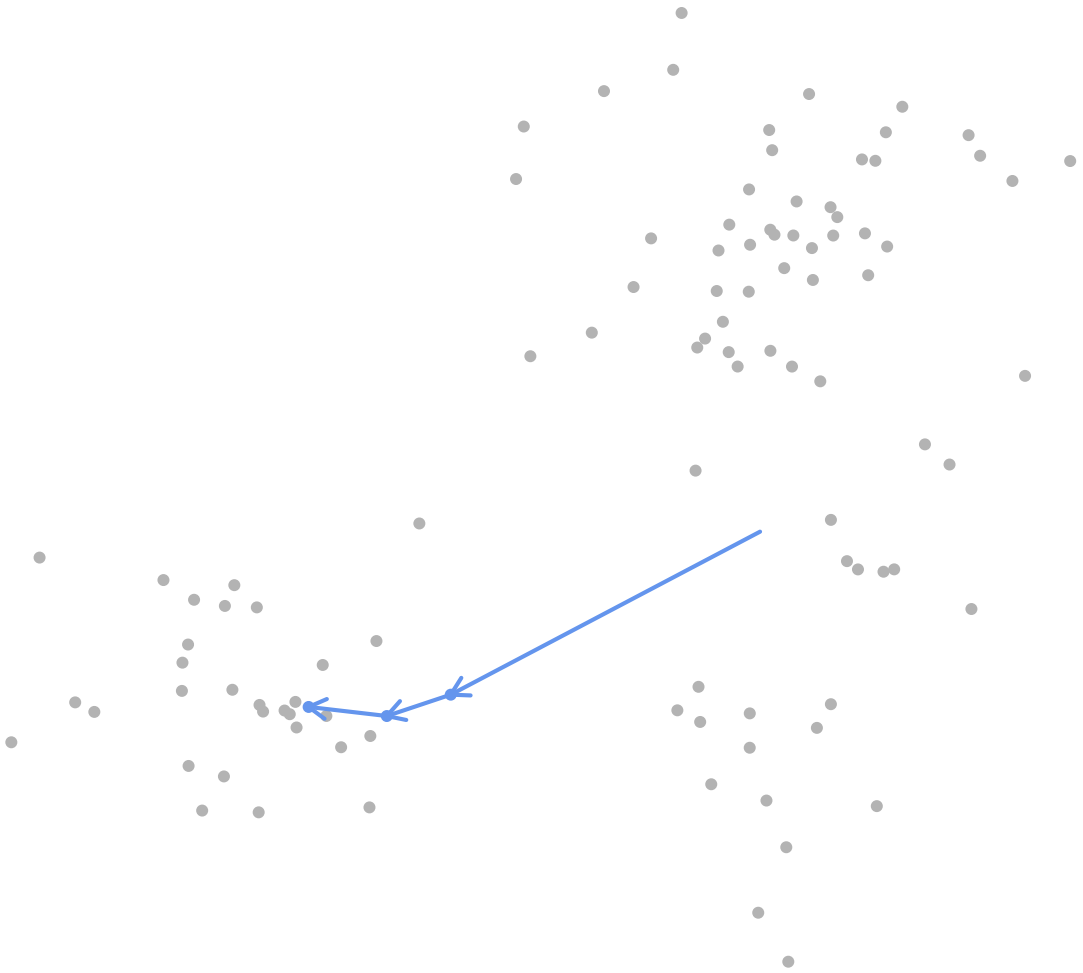
Visual Example



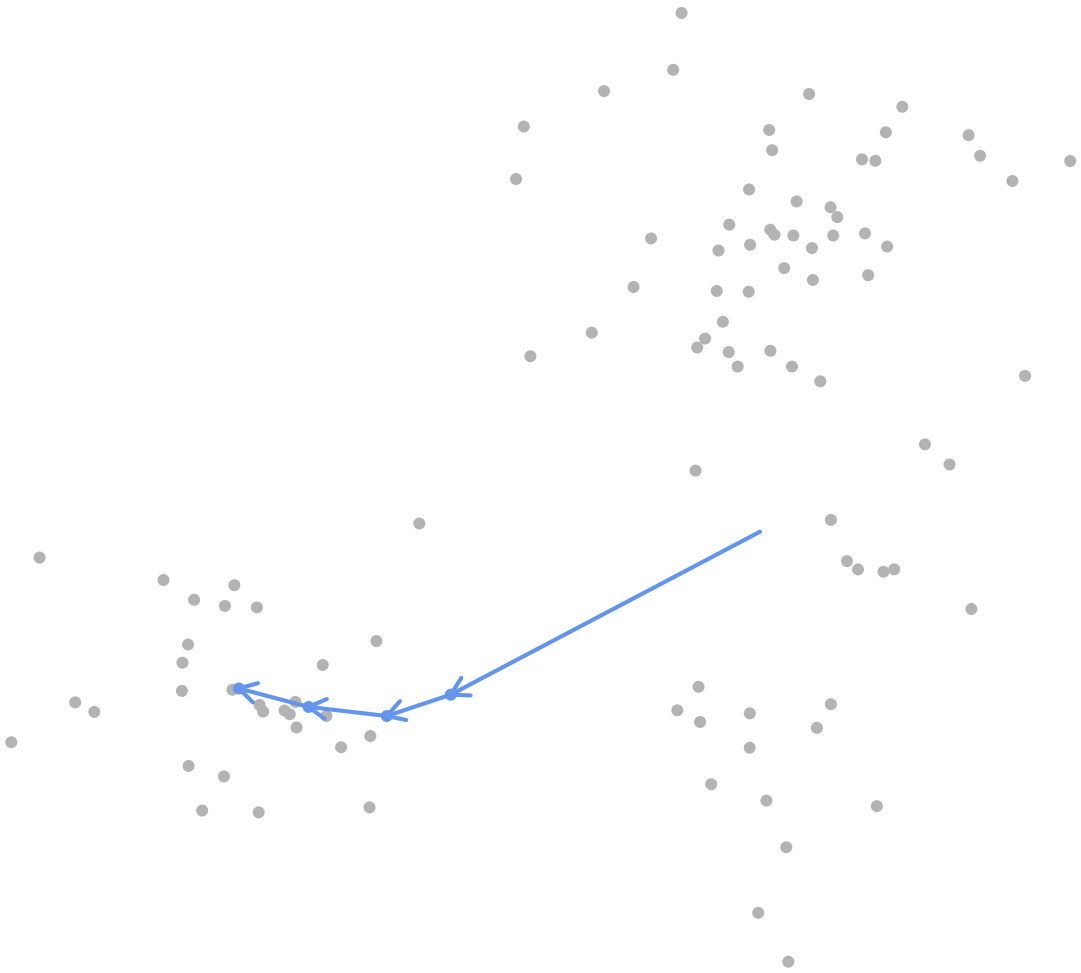
Visual Example



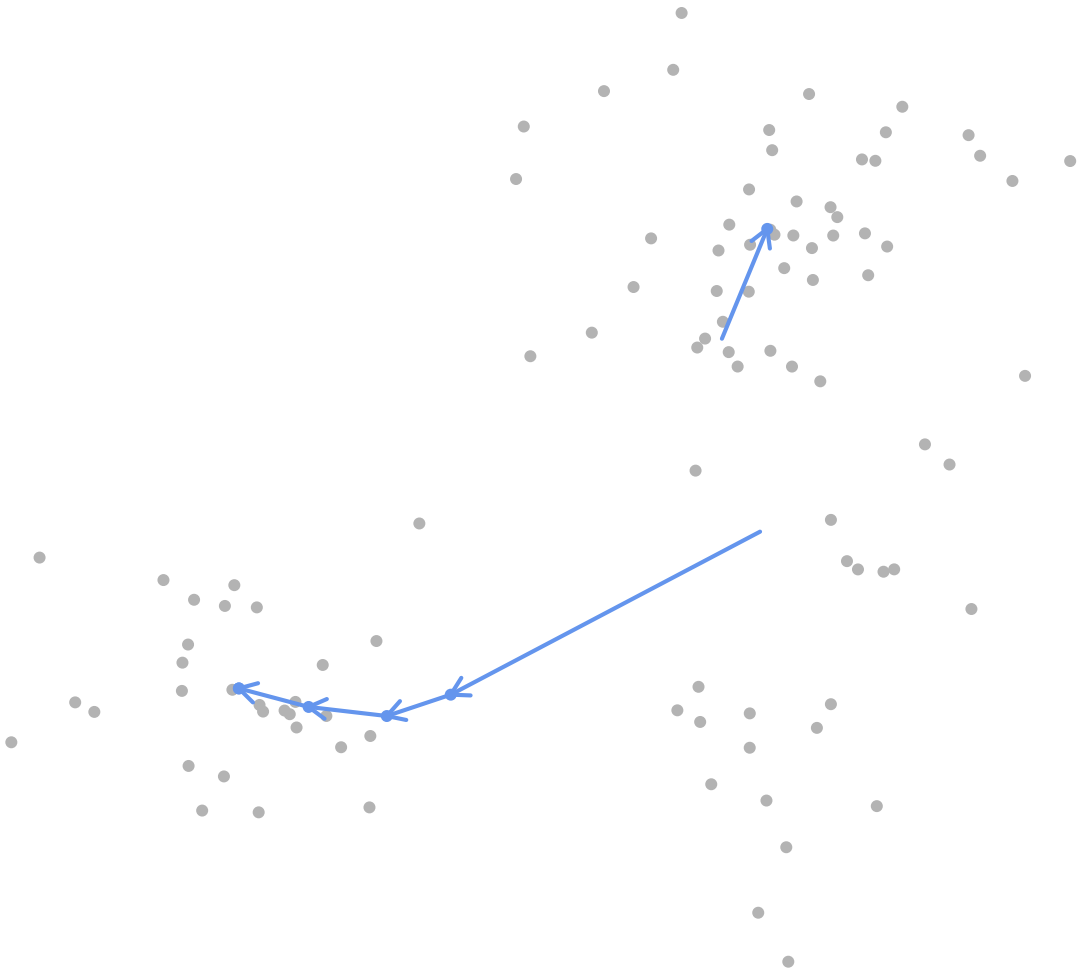
Visual Example



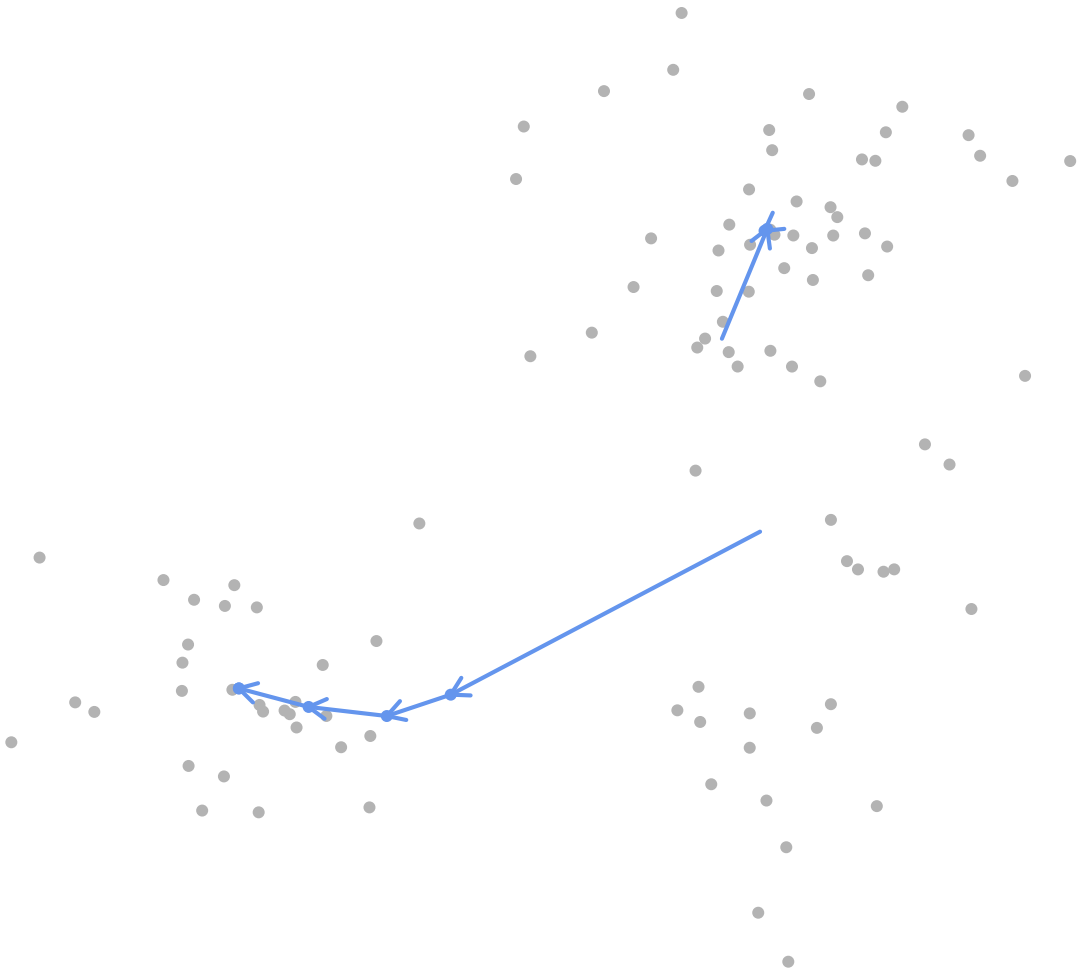
Visual Example



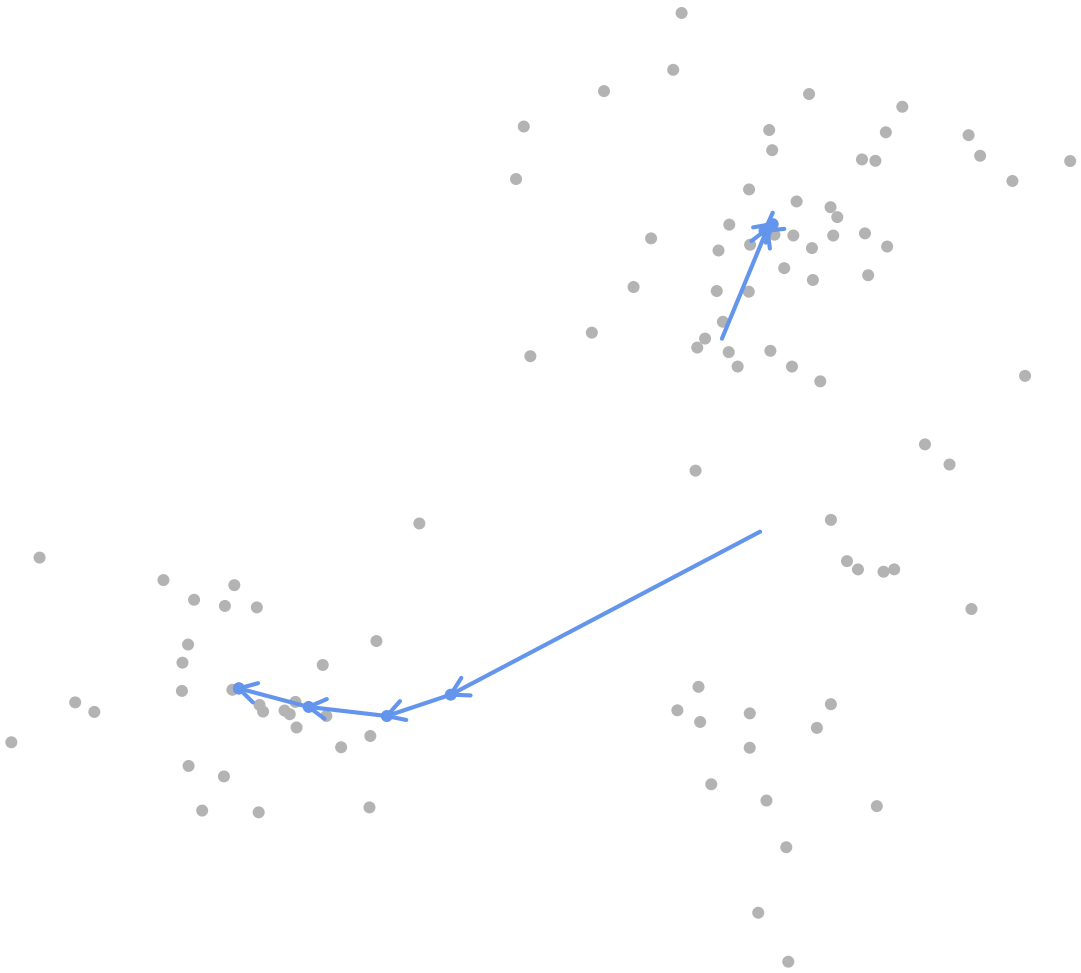
Visual Example



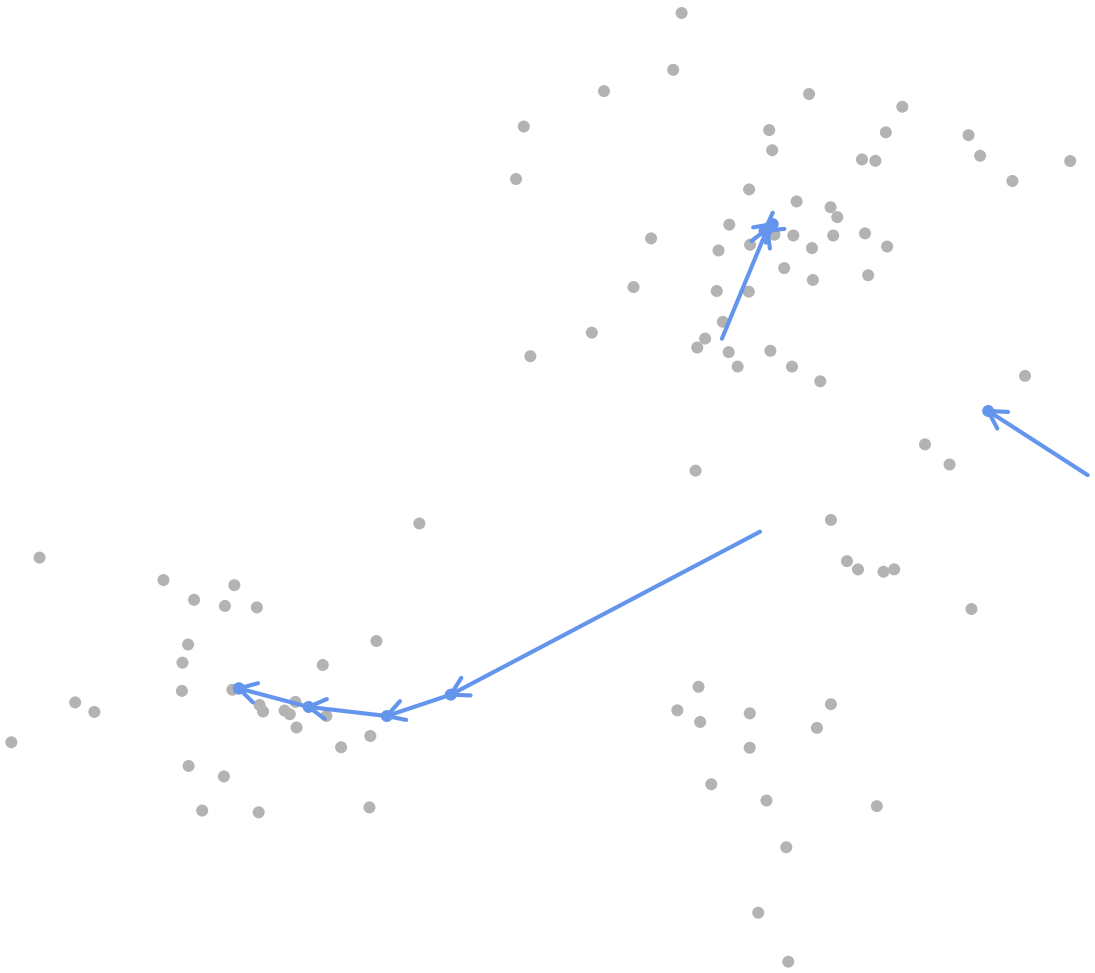
Visual Example



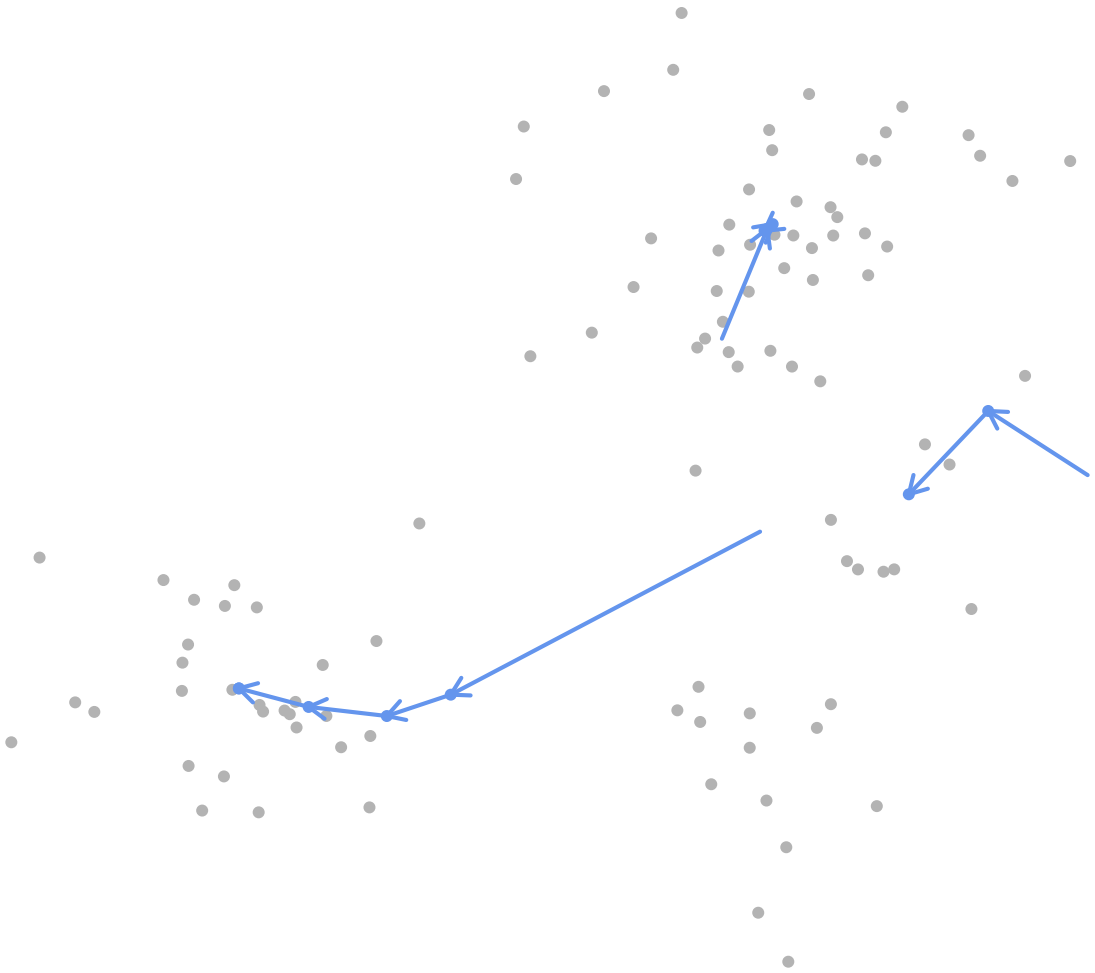
Visual Example



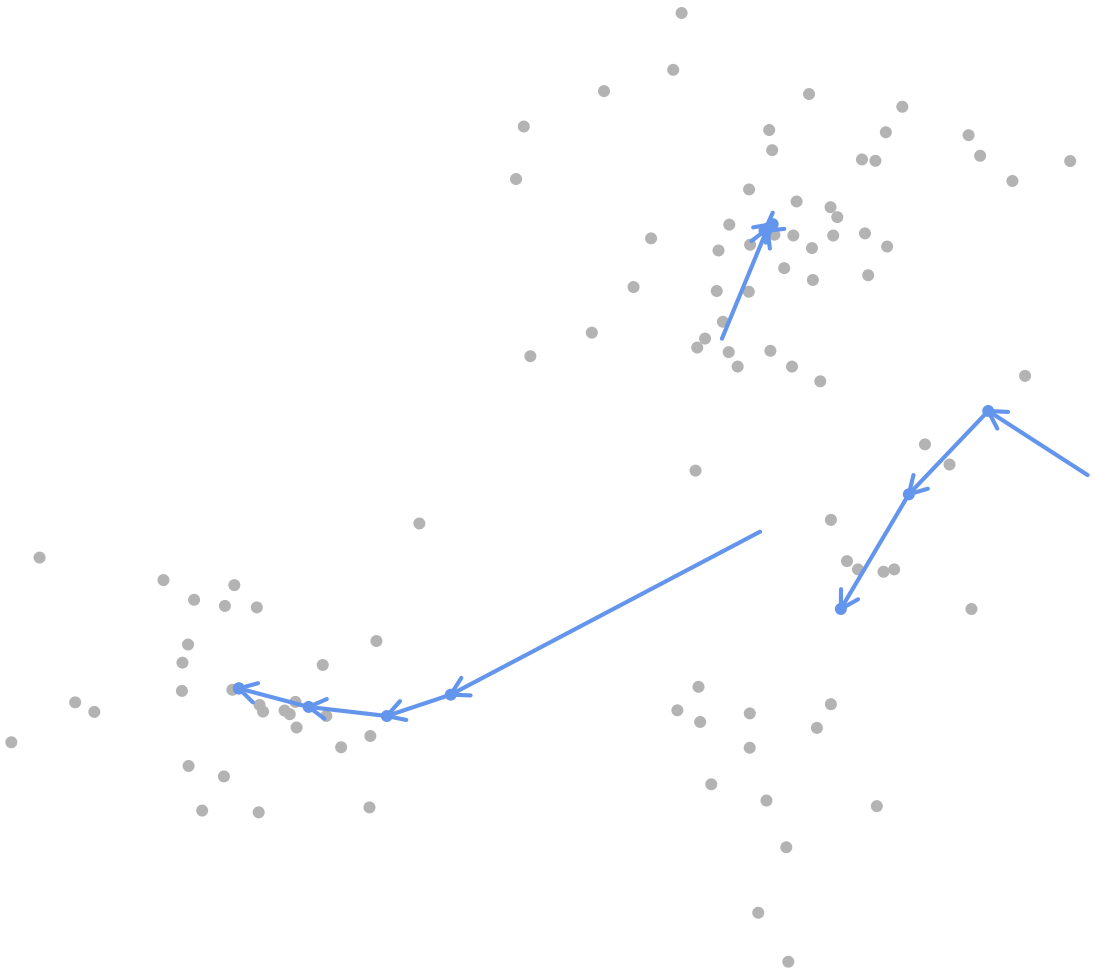
Visual Example



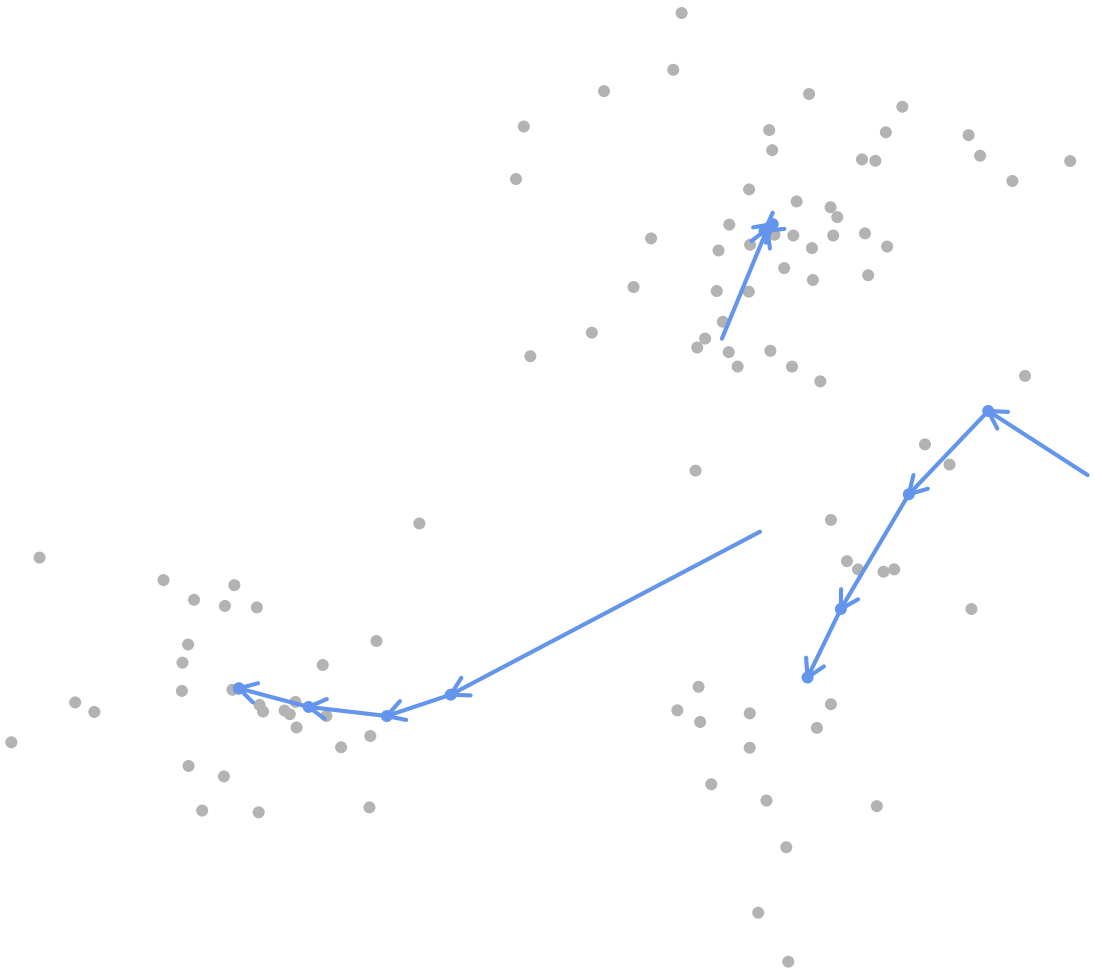
Visual Example



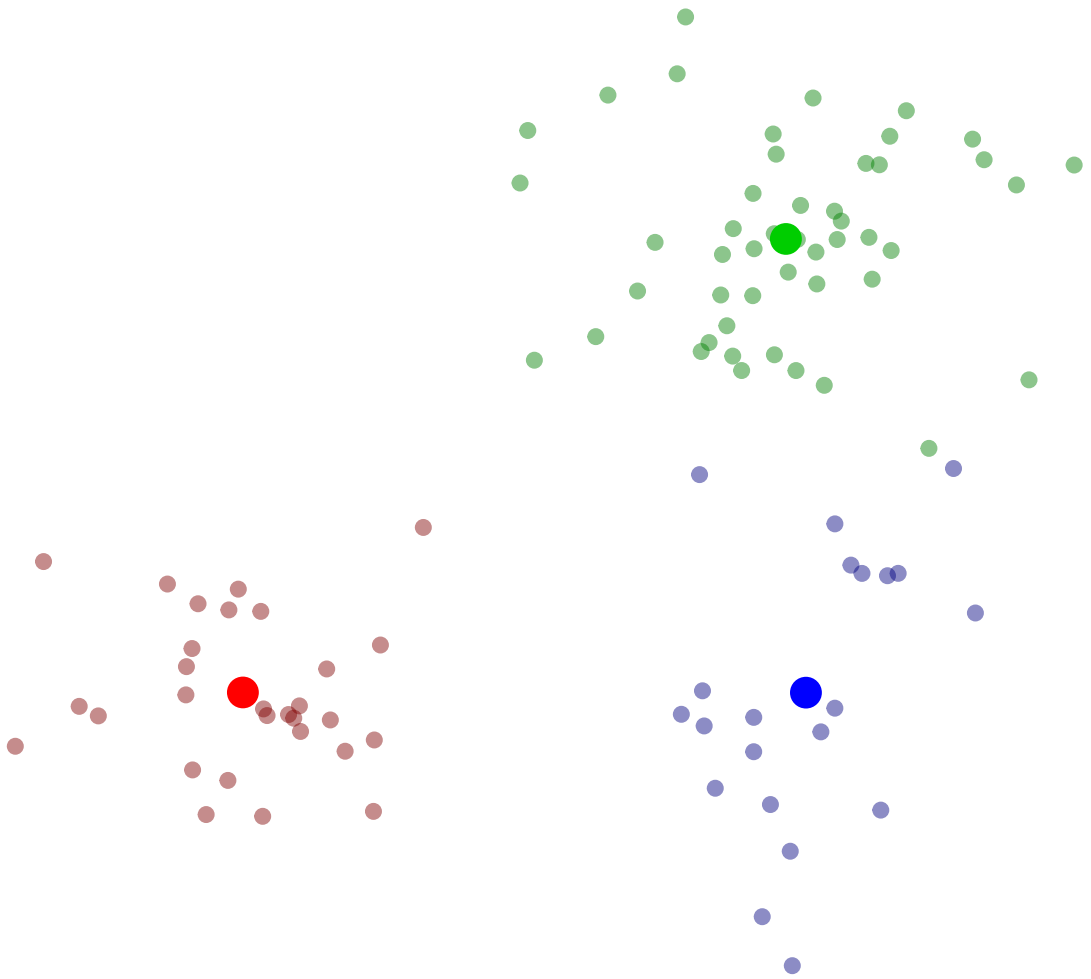
Visual Example



Visual Example



Visual Example



Interpreting Cluster Components

- Apply clustering methods, we have groups of documents
- How to interpret groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification (Week 4 methods)
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Best Validation:
  - Clustering methods suggest organization structure
  - Conditional on output, write coding rules
  - Humans code some documents
  - Use Week 8, 9 methods to classify
  - Correlation: strong evidence that grouping captures meaning you think

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)
  - Modelling 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)
  - Modelling 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)
  - Modelling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

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)
  - Modelling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- 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)
  - Modelling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- 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)
  - Modelling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistic + 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)
  - Modelling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

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

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Design to assess cluster quality

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Design to assess cluster quality

- Estimate clusterings

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)
- Select clustering with highest cluster quality

Cluster Quality (Grimmer and King 2011)

More general problem: model selection through humans

What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time

$\rightsquigarrow$ Cluster quality evaluation: using human judgement on pairs

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)
- Select clustering with highest cluster quality
- Can be used to compare any clusterings, regardless of source

Mixture Models

- Statistical models: make extensions/generalizations easier
- Mixture models: workhorse model for statistical clustering of data

Single distribution DGP:

$$\mathbf{y}_i \sim \text{Distribution(parameters)}$$

Mixture model DGP:

$$\begin{aligned} \mathbf{r}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \mathbf{y}_i|r_{ik} = 1 &\sim \text{Distribution}(\text{param}_k) \end{aligned}$$

DGP in Words

- Draw a cluster label
- Go to distribution, draw contents

A Mixture of Multinomial Distributions

Recall Multinomial Distribution:

$$\mathbf{y}_i|\boldsymbol{\theta} \sim \text{Multinomial}(\overbrace{n_i}^{\text{Number of Words}}, \underbrace{\boldsymbol{\theta}}_{\text{Rate Words are Used}})$$

$$p(\mathbf{y}_i|\boldsymbol{\theta}) \propto \prod_{m=1}^{M} \theta_m^{y_{im}}$$

A Mixture of Multinomial Distributions

$$\begin{aligned} \mathbf{r}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \mathbf{y}_i|r_{ij} = k, \boldsymbol{\theta}_k &\sim \text{Multinomial}(n_i, \boldsymbol{\theta}_k) \end{aligned}$$

where θ_{km} describes the rate word m is used in topic k.
Note: $\sum_{m=1}^{M} \theta_{km} = 1$.

A Mixture of Multinomial Distributions: Specifying Model

- Distance metric: Implicit, normalized Euclidean distance
- Objective function
- Optimization: EM Algorithm

$$p(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\pi} | \mathbf{y}) \quad \propto \quad \prod_{i=1}^{N} \prod_{j=1}^{K} \left[\pi_j \prod_{m=1}^{M} \theta_{km}^{y_{im}} \right]^{r_{ij}}$$

A Mixture of Multinomial Distributions: Specifying Model

- Distance metric: Implicit, normalized Euclidean distance
- Objective function
- Optimization: EM Algorithm

A Mixture of Multinomial Distributions: Specifying Model

- Distance metric: Implicit, normalized Euclidean distance
- Objective function
- Optimization: EM Algorithm

Proceeds in three steps: Initialize $\boldsymbol{\pi}^{t-1}$, $\boldsymbol{\theta}^{t-1}$

A Mixture of Multinomial Distributions: Specifying Model

- Distance metric: Implicit, normalized Euclidean distance
- Objective function
- Optimization: EM Algorithm

Proceeds in three steps: Initialize $\boldsymbol{\pi}^{t-1}$, $\boldsymbol{\theta}^{t-1}$

$$r_{ik}^{t} = \frac{\pi_k^{t-1} \prod_{m=1}^{M} \theta_{km}^{y_{im},t-1}}{\sum_{j=1}^{K} \pi_j^{t-1} \prod_{m=1}^{M} \theta_{jm}^{y_{im},t-1}}$$

A Mixture of Multinomial Distributions: Specifying Model

- Distance metric: Implicit, normalized Euclidean distance
- Objective function
- Optimization: EM Algorithm

Proceeds in three steps: Initialize $\boldsymbol{\pi}^{t-1}$, $\boldsymbol{\theta}^{t-1}$

$$
\begin{aligned}
r_{ik}^{t} &= \frac{\pi_k^{t-1} \prod_{m=1}^{M} \theta_{km}^{y_{im},t-1}}{\sum_{j=1}^{K} \pi_j^{t-1} \prod_{m=1}^{M} \theta_{jm}^{y_{im},t-1}} \\
\pi_k &= \sum_{i=1}^{N} r_{ik}
\end{aligned}
$$

A Mixture of Multinomial Distributions: Specifying Model

- Distance metric: Implicit, normalized Euclidean distance
- Objective function
- Optimization: EM Algorithm

Proceeds in three steps: Initialize $\boldsymbol{\pi}^{t-1}$, $\boldsymbol{\theta}^{t-1}$

$$
\begin{aligned}
r_{ik}^{t} &= \frac{\pi_k^{t-1} \prod_{m=1}^{M} \theta_{km}^{y_{im},t-1}}{\sum_{j=1}^{K} \pi_j^{t-1} \prod_{m=1}^{M} \theta_{jm}^{y_{im},t-1}} \\
\pi_k &= \sum_{i=1}^{N} r_{ik} \\
\theta_k &\propto \sum_{i=1}^{N} r_{ik} \mathbf{y}_i
\end{aligned}
$$

(Non-parametric) Clustering of Press Releases (Grimmer 2011)

Apply version of mixture of multinomials to 64,033 Senate press releases
Model fit with approximately 45 topics

Label	Identifying Stems	% Press Rele
Appropriations/Grants	fund,project,000,million,water	8.6
Honorary	honor,servic,school,serv,american	8.2
Iraq War	iraq,troop,war,iraqi,american	6.6
Health Grants	health,program,educ,children,school	6.3
Homeland Security	secur,homeland,port,border,depart	5.3
Judicial Nominations	court,vote,justic,american,judg	4.8
Hurricanes/Disasters	disast,assist,hurrican,fema,flood	4.5
Taxes	tax,american,budget,social,secur	4.4
Defense Projects	million,defens,fund,air,militari	4.2
Health Policy	health,care,drug,medicar,senior	3.8

An Overview of Clustering Models

There are a lot of different clustering models (and many variations within each):
k-means

An Overview of Clustering Models

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials

An Overview of Clustering Models

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids

An Overview of Clustering Models

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation

An Overview of Clustering Models

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation , agglomerative Hierarchical

An Overview of Clustering Models

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation , agglomerative Hierarchical fuzzy k-means, trimmed k-means, k-Harmonic means, fuzzy k-medoids, fuzzy k modes, maximum entropy clustering, model based hierarchical (agglomerative), proximus, ROCK, divisive hierarchical, DISMEA, Fuzzy, QTClust, self-organizing map, self-organizing tree, unnormalized spectral, MS spectral, NJW Spectral, SM Spectral, Dirichlet Process Multinomial, Dirichlet Process Normal, Dirichlet Process von-mises Fisher, Mixture of von mises-Fisher (EM), Mixture of von Mises Fisher (VA), Mixture of normals, co-clustering mutual information, co-clustering SVD, LLAhclust, CLUES, bclust, c-shell, qtClustering, LDA, Express Agenda Model, Hierarchical Dirichlet process prior, multinomial, uniform process mulitinomial, Chinese Restaurant Distance Dirichlet process multinomial, Pitmann-Yor Process multinomial, LSA, ...

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method —

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible

Deep problem in cluster analysis literature: full automation requires more information

Fully Automated → Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
  - Easy (if you don't think about it): list all clustering, choose best
  - Impossible in Practice
  - Solution: Organized list
  - Insight: Many clusterings are perceptually identical
  - Consider two clusterings of 10,000 documents, we move one document from 5 to 6.
- How to organize clusterings so humans can undestand?
- Our answer: a geography of clusterings

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
   - New Clustering: weighted average of clusterings from methods

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
   - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
   - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)
7) $\rightsquigarrow$ Millions of clusterings easily comprehended

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
   - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)
7) $\rightsquigarrow$ Millions of clusterings easily comprehended
8) (Or, our new strategy: represent entire Bell space directly; no need to examine document contents)

Grimmer, King, and Stewart, In Progress

A brief live demonstration of α software (time permitting)

Example Discovery: What Do Members of Congress Do?

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

Example Discovery: What Do Members of Congress Do?

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

Example Discovery: What Do Members of Congress Do?

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

Example Discovery: What Do Members of Congress Do?

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

Example Discovery: What Do Members of Congress Do?

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

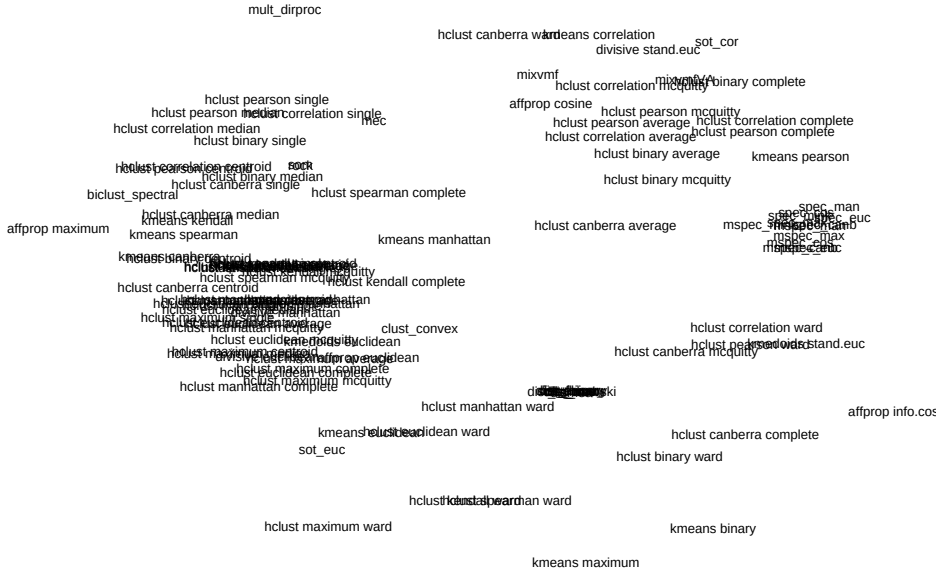
Example Discovery: What Do Members of Congress Do?

- 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: What Do Members of Congress Do?

- 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 our method (relying on many clustering algorithms)

Example Discovery



Example Discovery

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

Example Discovery

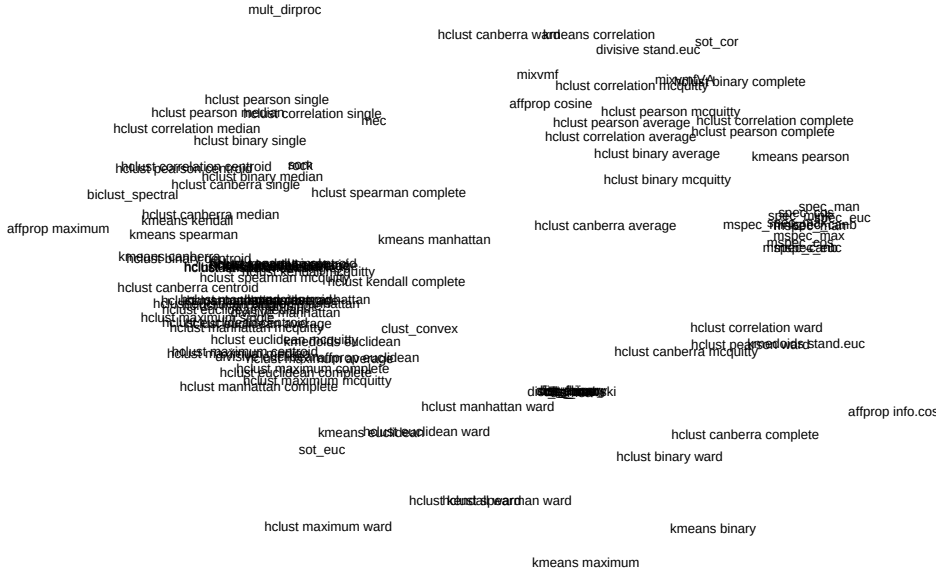
Each point is a clustering Affinity Propagation-Cosine (Dueck and Frey 2007)

Close to:

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

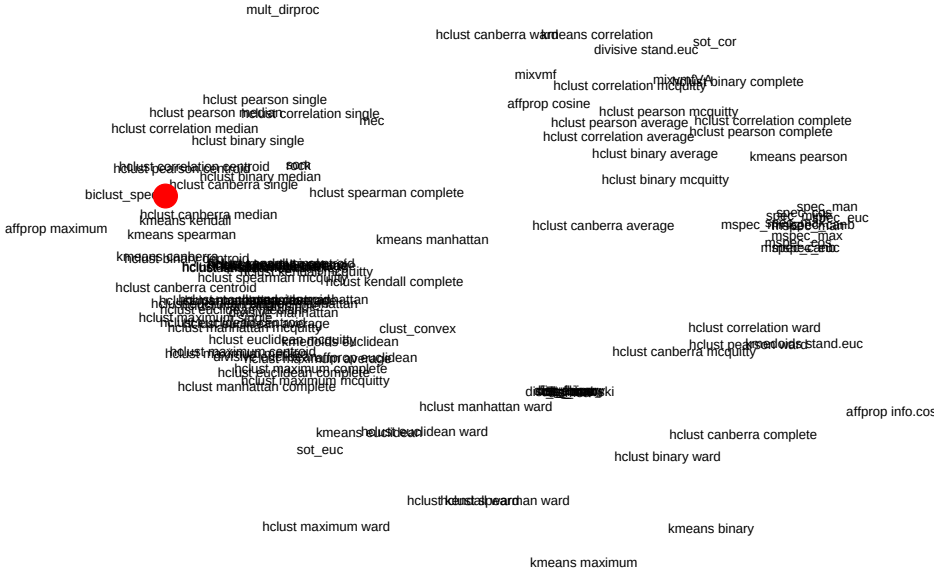
$\Rightarrow$ Similar clustering of documents

Example Discovery



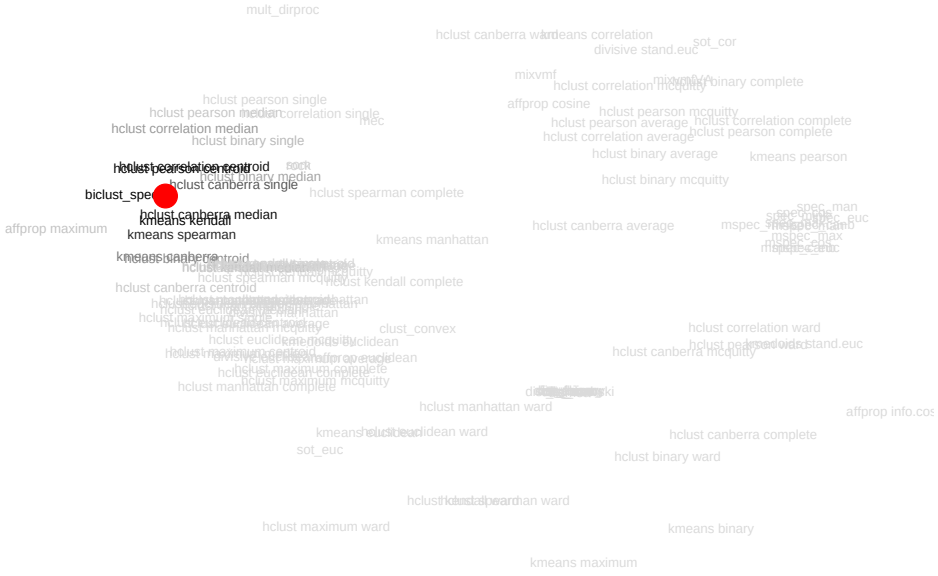
Space between methods:

Example Discovery



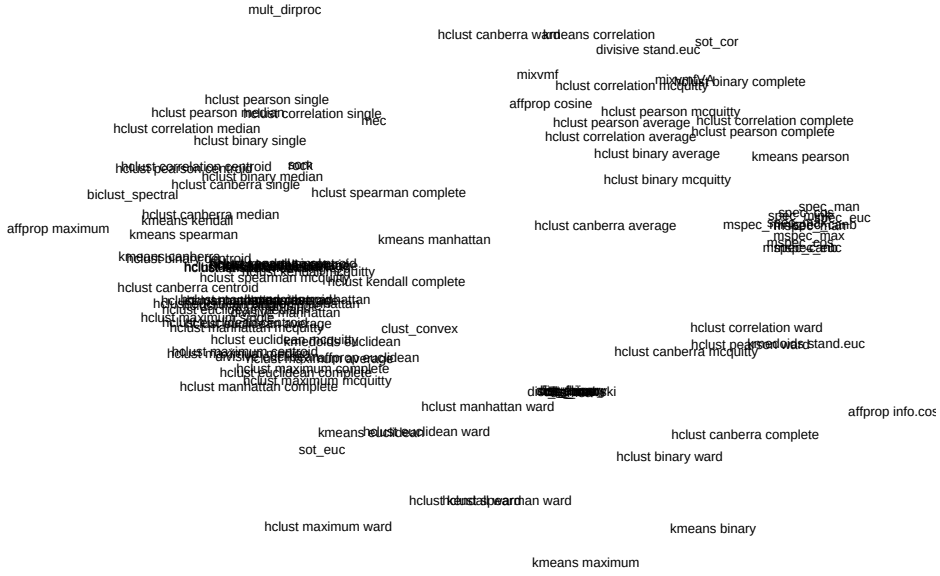
Space between methods:

Example Discovery

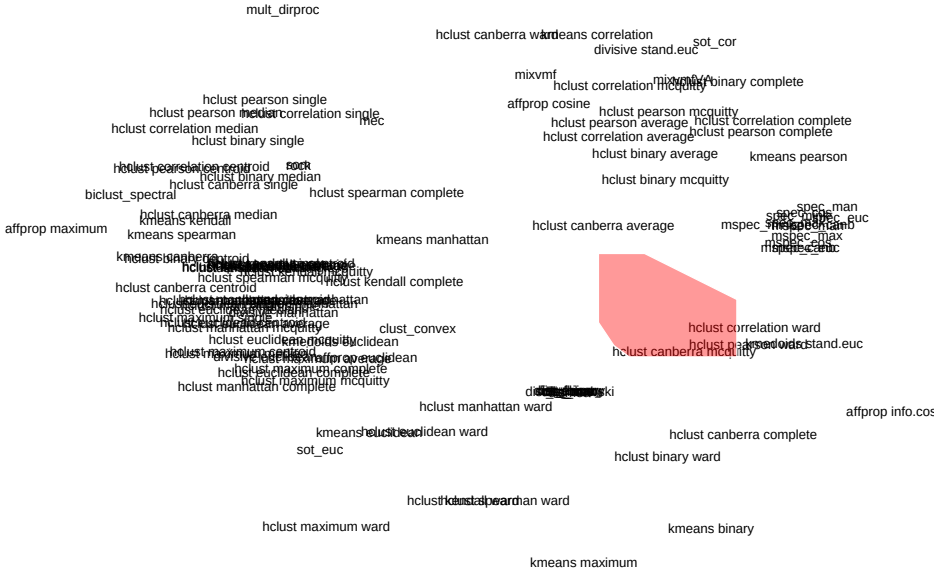


Space between methods:
local cluster ensemble

Example Discovery

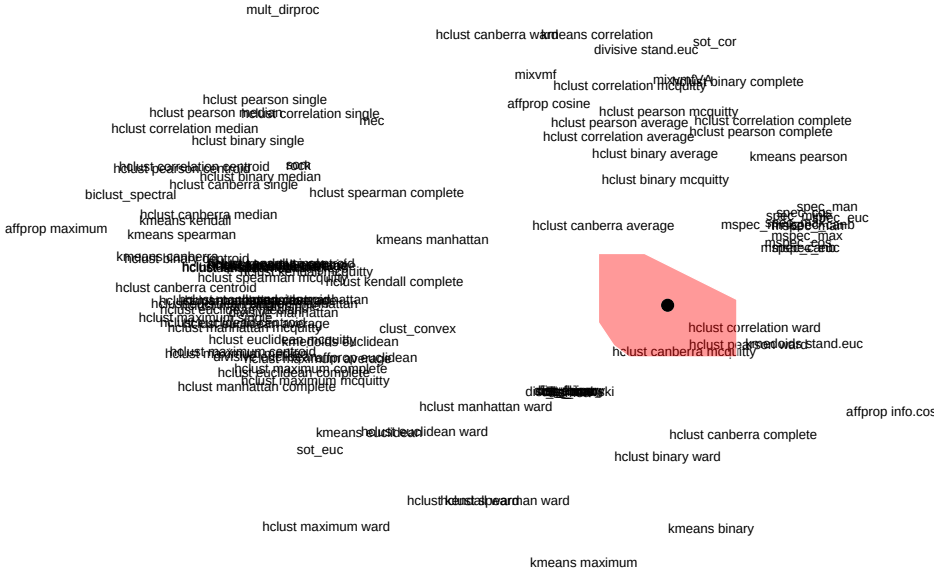


Example Discovery



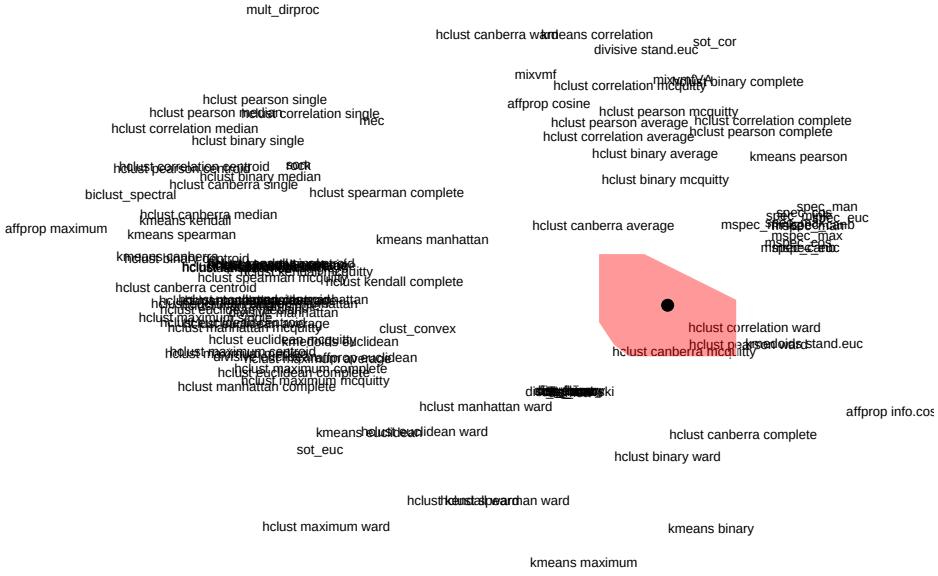
Found a region with clusterings that all reveal the same important insight

Example Discovery



Mixture:

Example Discovery



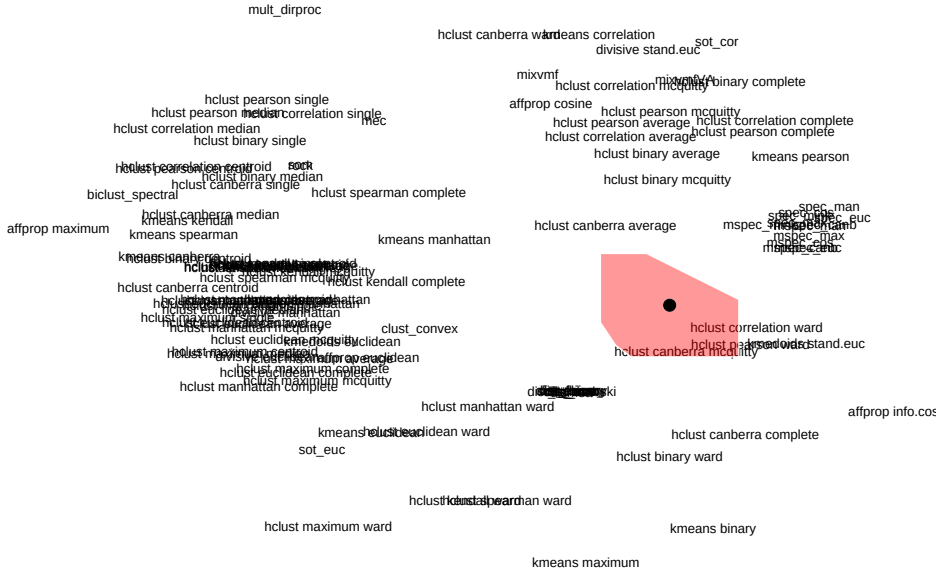
Mixture:

0.39 Hclust-Canberra-McQuitty

0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

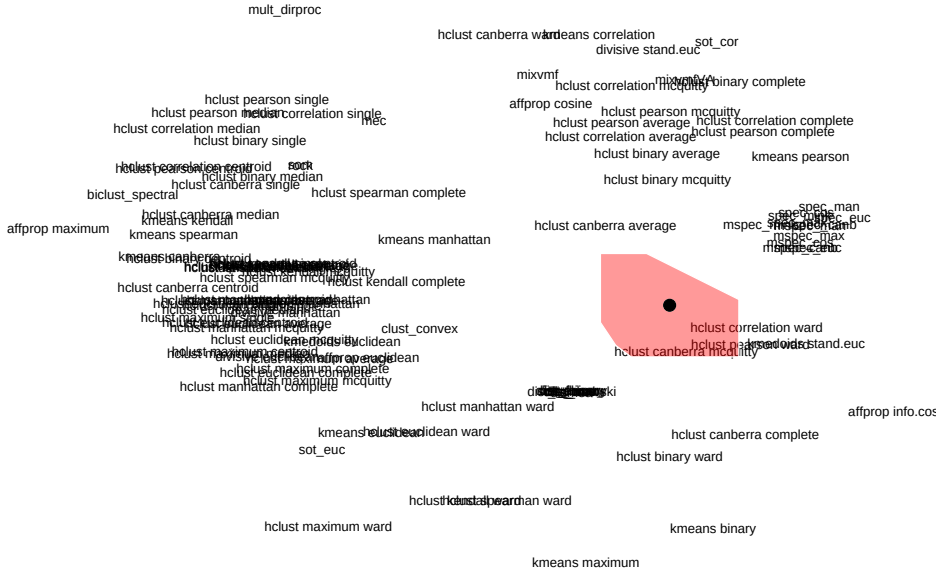
0.30 Spectral clustering
Random Walk
(Metrics 1-6)

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)

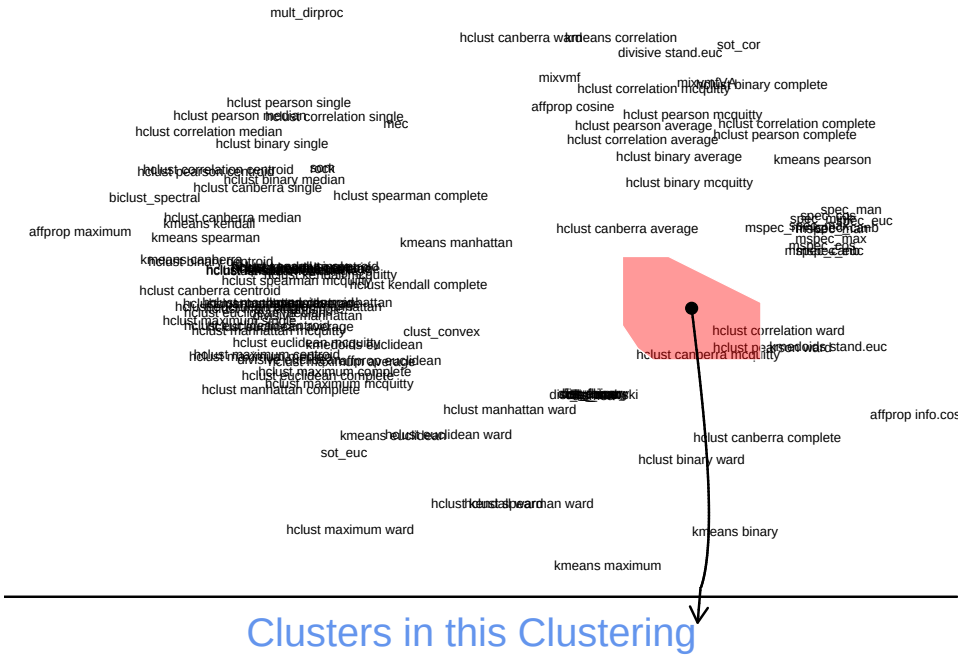
0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

0.05 Kmediods-Cosine

0.04 Spectral clustering
Symmetric
(Metrics 1-6)

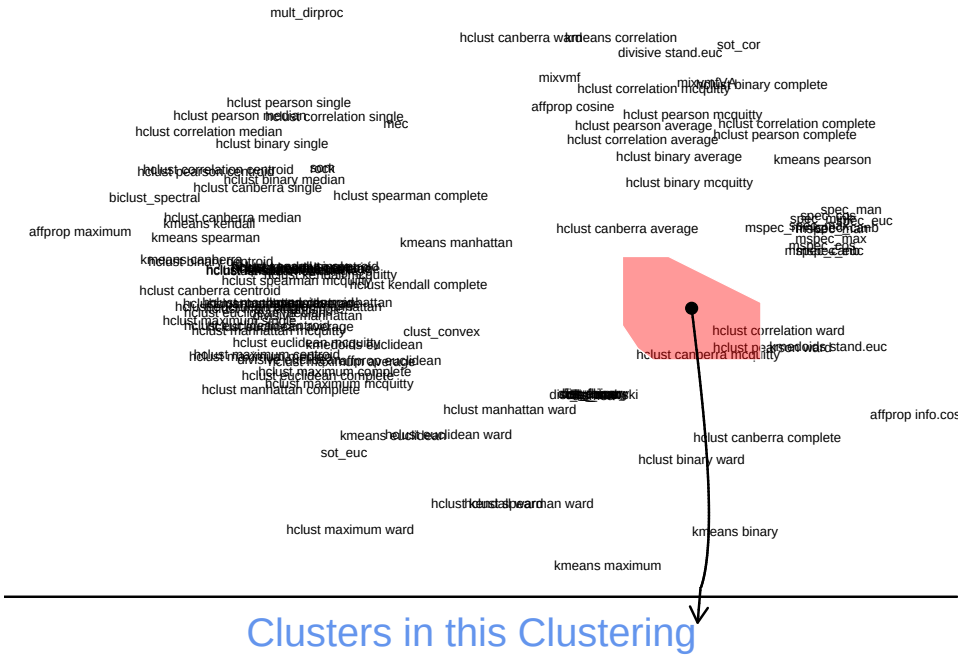
Example Discovery

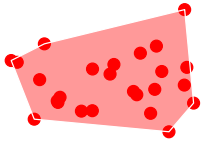


Clusters in this Clustering

Mayhew

Example Discovery





Credit Claiming
Pork

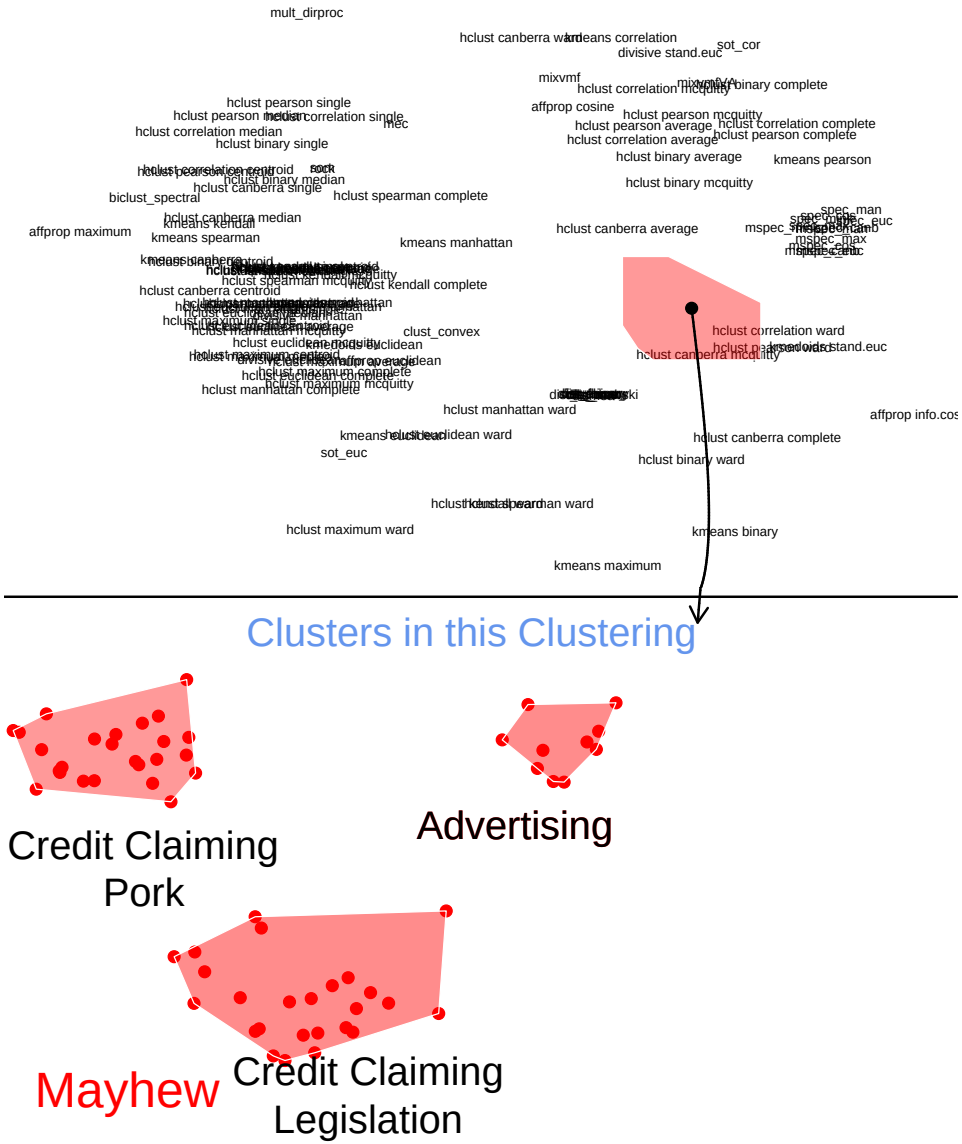
Credit Claiming, Pork:
"Sens. Frank R. Lautenberg (D-NJ) and Robert Menendez (D-NJ) announced that the U.S. Department of Commerce has awarded a $100,000 grant to the South Jersey Economic Development District"

Mayhew

Example Discovery

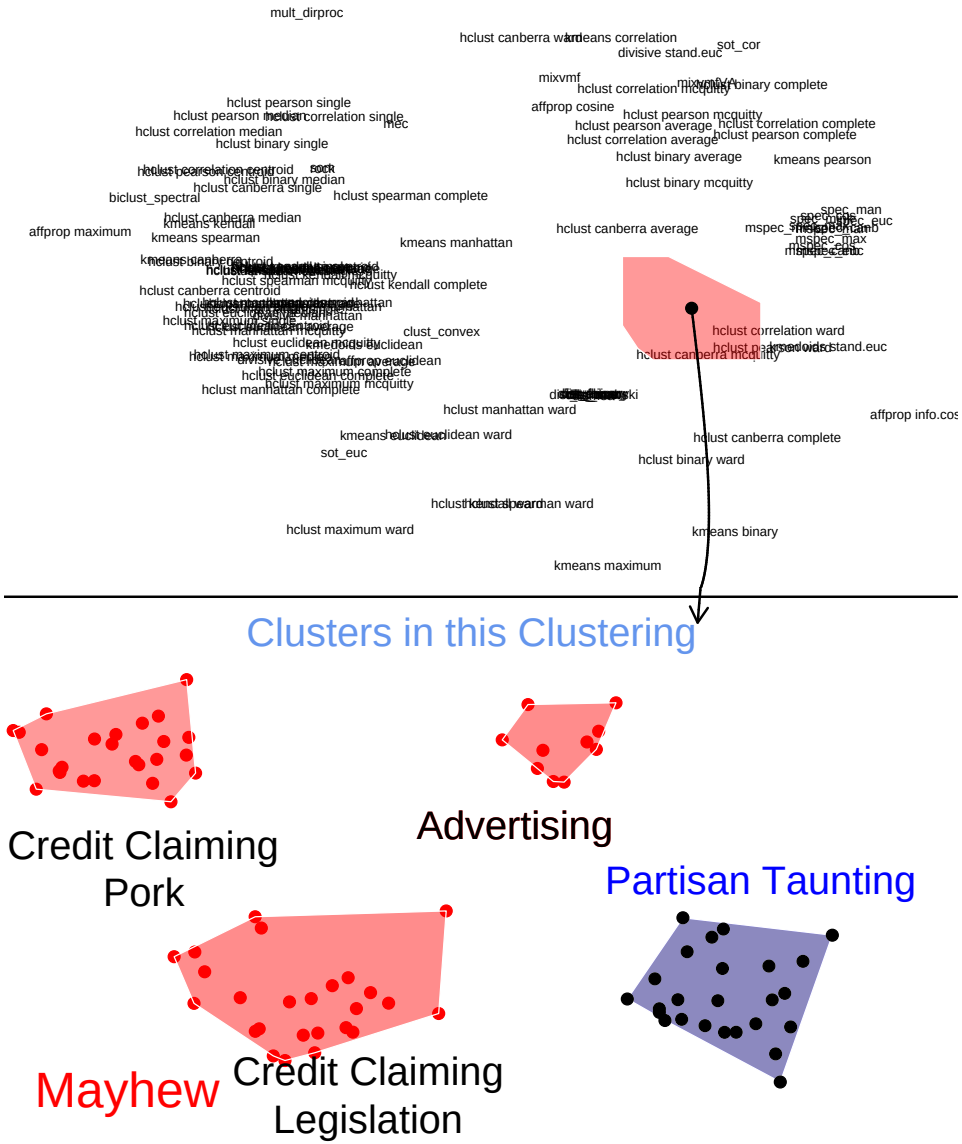
Credit Claiming, Legislation:
"As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period"

Example Discovery



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

Example Discovery: Partisan Taunting



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

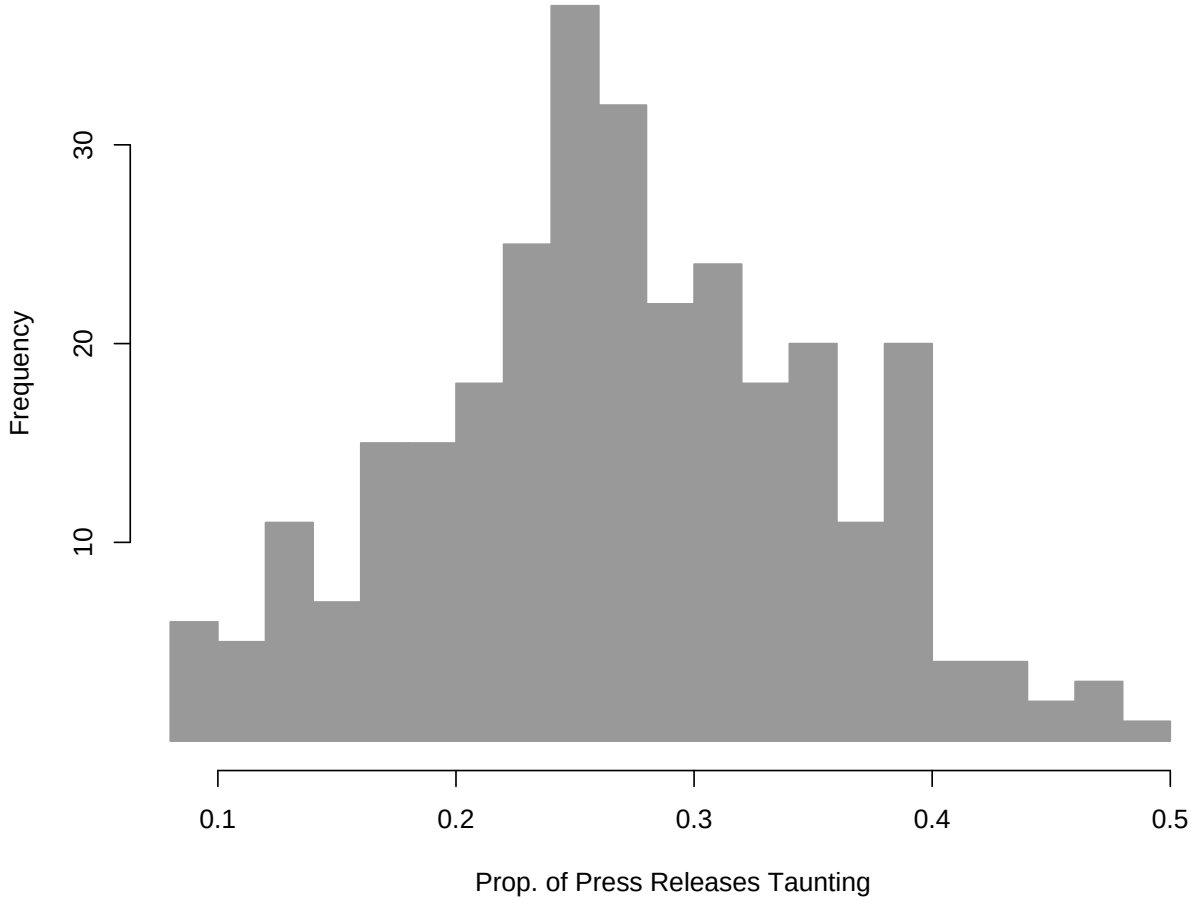
- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.

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

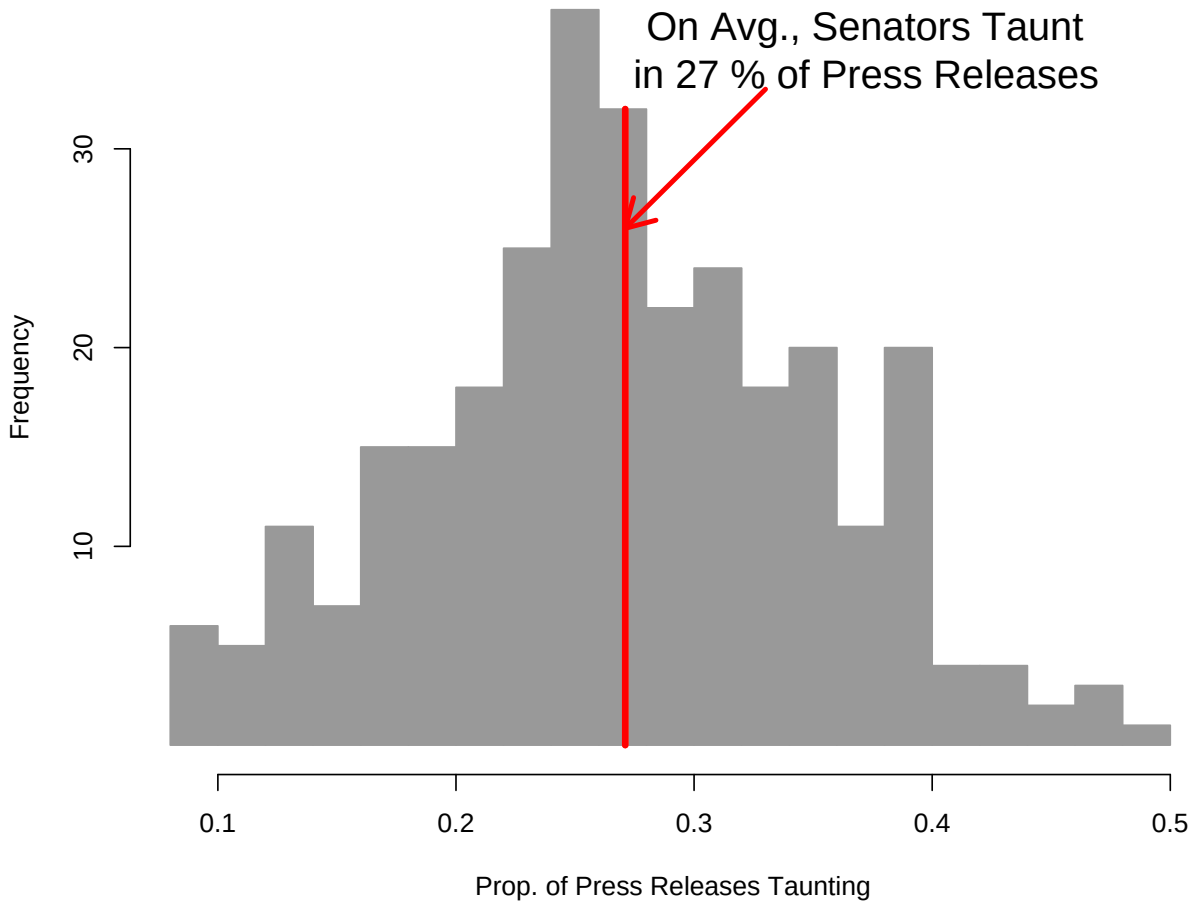
Out of Sample Confirmation of Partisan Taunting

- 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



Clustering, FAC and CAC

This week

- Introduction to clustering
- Fully automated clustering algorithms
- Introduction to computer assisted clustering

Next week:

- Topic models
- Discover underlying issues in texts