

# Advanced Quantitative Research Methodology, Lecture Notes: Text Analysis II: Unsupervised Learning via Cluster Analysis<sup>1</sup>

Gary King  
<http://GKing.Harvard.Edu>

December 23, 2011

---

<sup>1</sup>©Copyright 2010 Gary King, All Rights Reserved.

Justin Grimmer and Gary King. 2010. “Quantitative Discovery of Qualitative Information: A General Purpose Document Clustering Methodology”

<http://gking.harvard.edu/files/abs/discov-abs.shtml>.

# The Problem: Discovery from Unstructured Text

- Examples: scholarly literature, news stories, medical information, blog posts, comments, product reviews, emails, social media updates, audio-to-text summaries, speeches, press releases, legal decisions, etc.
- 10 minutes of worldwide email = 1 LOC equivalent
- An essential part of discovery is **classification**: “one of the most central and generic of all our conceptual exercises. . . the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis. . . Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research.” (Bailey, 1994).
- We focus on **cluster analysis**: discovery through (1) classification and (2) simultaneously inventing a classification scheme
- (We analyze text; our methods apply more generally)

# Why Johnny Can't Classify (Optimally)

- $\text{Bell}(n)$  = number of ways of partitioning  $n$  objects
- $\text{Bell}(2) = 2$  (AB, A B)
- $\text{Bell}(3) = 5$  (ABC, AB C, A BC, AC B, A B C)
- $\text{Bell}(5) = 52$
- $\text{Bell}(100) \approx 10^{28} \times$  Number of elementary particles in the universe
- Now imagine choosing the *optimal* classification scheme by hand!
- That we think of all this as astonishing ... is astonishing

# Why HAL Can't Classify Either

- 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, **who knows?!**
  - The literature: little guidance on when methods apply
  - **Deep problem in cluster analysis literature**: no way to know which method will work *ex ante*

# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model
- We do it **ex post** (by qualitative choice). For example:
  - Create long list of clusterings; choose the best
  - Too hard for mere humans!
  - An **organized** list will make the search possible
  - E.g.,: consider two clusterings that differ only because one document (of many) moves from category 5 to 6

## Our Idea: Meaning Through Geography

wide at **SuperPages.com**

195

Car

**C**

[illegible]

→ We develop a (conceptual) geography of clusterings

# A New Strategy

Make it easy to choose best clustering from millions of choices

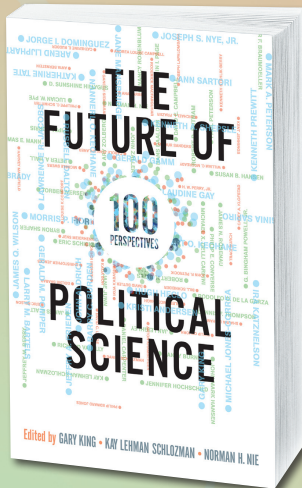
- 1 **Code text as numbers** (in one *or more* of several ways)
- 2 **Apply all clustering methods we can find** to the data — each representing different (unstated) substantive assumptions (<15 mins)
- 3 (Too much for a person to understand, but organization will help)
- 4 Develop an **application-independent distance metric** between clusterings, a **metric space of clusterings**, and a **2-D projection**
- 5 **“Local cluster ensemble”** creates a new clustering at any point, based on weighted average of nearby clusterings
- 6 A new **animated visualization** to explore the space of clusterings (smoothly morphing from one into others)
- 7 **↪ Millions of clusterings, easily comprehended** (takes about 10-15 minutes to choose a clustering with insight)

You choose one (or more), based on insight, discovery, useful information,...



# Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions
  - ① Distance between clusterings: a function of the **pairwise document agreements** (pairwise agreements  $\Rightarrow$  triples, quadruples, etc.)
  - ② **Invariance**: Distance is invariant to the number of documents (for any fixed number of clusters)
  - ③ **Scale**: the maximum distance is set to  $\log(\text{num clusters})$
- $\rightsquigarrow$  **Only one measure satisfies all three** (the “variation of information”)
- Meila (2007): derives same metric using different axioms (lattice theory)



Available March 2009: 304pp  
 Pb: 978-0-415-99701-0: **\$24.95**  
[www.routledge.com/politics](http://www.routledge.com/politics)

# THE FUTURE OF POLITICAL SCIENCE

## 100 Perspectives

Edited by Gary King, Harvard University, Kay Lehman Schlozman, Boston College  
 and Norman H. Nie, Stanford University

**"The list of authors in *The Future of Political Science* is a 'who's who' of political science. As I was reading it, I came to think of it as a platter of tasty hors d'oeuvres. It hooked me thoroughly."**

—Peter Kingstone, University of Connecticut

**"In this one-of-a-kind collection, an eclectic set of contributors offer short but forceful forecasts about the future of the discipline. The resulting assortment is captivating, consistently thought-provoking, often intriguing, and sure to spur discussion and debate."**

—Wendy K. Tam Cho, University of Illinois at Urbana-Champaign

**"King, Schlozman, and Nie have created a visionary and stimulating volume. The organization of the essays strikes me as nothing less than brilliant. . . It is truly a joy to read."**

—Lawrence C. Dodd, Manning J. Dauer Eminent Scholar in Political Science,  
 University of Florida

**Routledge**  
 Taylor & Francis Group  
 an informa business

# Evaluators' Rate Machine Choices Better Than Their Own

- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

| Pairs from          | Overall Mean | Evaluator 1 | Evaluator 2 |
|---------------------|--------------|-------------|-------------|
| Random Selection    | 1.38         | 1.16        | 1.60        |
| Hand-Coded Clusters | 1.58         | 1.48        | 1.68        |
| Hand-Coding         | 2.06         | 1.88        | 2.24        |
| Machine             | 2.24         | 2.08        | 2.40        |

p.s. The hand-coders did the evaluation!

# Evaluating Performance

- Goals:
  - **Validate Claim**: computer-assisted conceptualization outperforms human conceptualization
  - **Demonstrate**: new experimental designs for cluster evaluation
  - **Inject human judgement**: relying on insights from survey research
- We now present three evaluations
  - Cluster Quality  $\Rightarrow$  RA coders
  - Informative discoveries  $\Rightarrow$  Experienced scholars analyzing texts
  - Discovery  $\Rightarrow$  You're the judge

# Evaluation 1: Cluster Quality

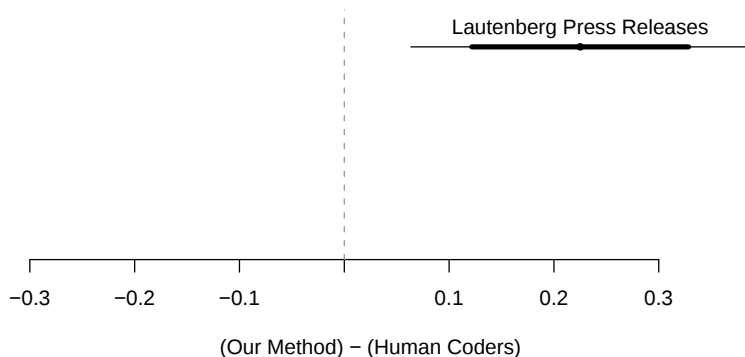
- What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time
- $\Rightarrow$  Cluster quality evaluation: human judgement of document pairs

- Experimental Design to Assess Cluster Quality

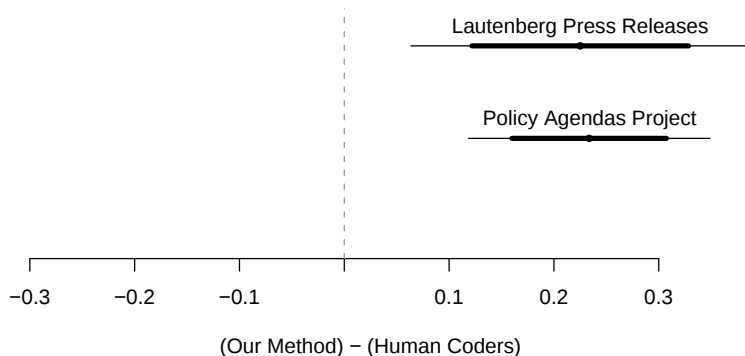
- automated visualization to choose one clustering
- many pairs of documents
- for coders: (1) unrelated, (2) loosely related, (3) closely related
- Quality = mean(within cluster) - mean(between clusters)
- Bias results against ourselves by not letting evaluators choose clustering

# Evaluation 1: Cluster Quality



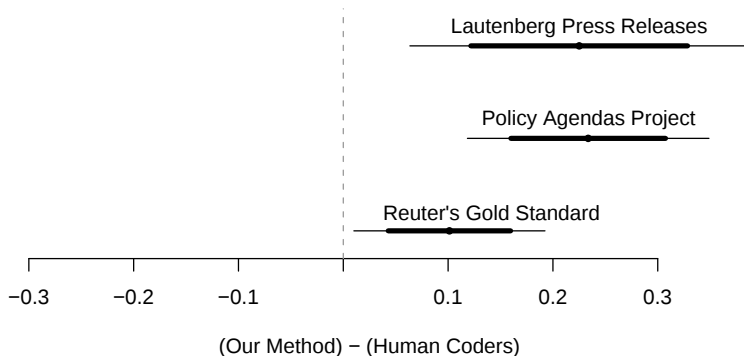
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

# Evaluation 1: Cluster Quality



Policy Agendas: 213 quasi-sentences from Bush's State of the Union (agriculture, banking & commerce, civil rights/liberties, defense, ...)

# Evaluation 1: Cluster Quality



Reuter's: financial news (trade, earnings, copper, gold, coffee, . . . ); “gold standard” for supervised learning studies

## Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for  $\binom{6}{2}=15$  pairwise comparisons
- User chooses  $\Rightarrow$  only care about the one clustering that wins
- Both cases a Condorcet winner:

“Immigration”:

Our Method 1  $\rightarrow$  vMF 1  $\rightarrow$  vMF 2  $\rightarrow$  Our Method 2  $\rightarrow$  K-Means 1  $\rightarrow$  K-Means 2

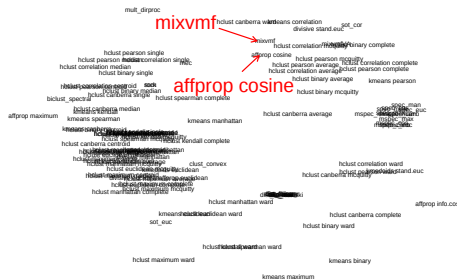
“Genetic testing”:

Our Method 1  $\rightarrow$  {Our Method 2, K-Means 1, K-means 2}  $\rightarrow$  Dir Proc. 1  $\rightarrow$  Dir Proc. 2

# Evaluation 3: What Do Members of Congress Do?

- 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

## Example Discovery



Red point: a **clustering** by Affinity Propagation-Cosine (Dueck and Frey 2007)

Close to:

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

## Example Discovery



Space between methods:  
local cluster ensemble



## Example Discovery



Mixture:

### 0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering  
Random Walk  
(Metrics 1-6)

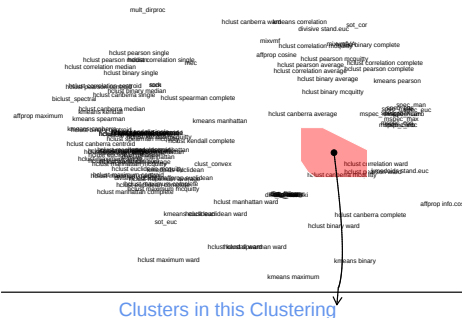
### 0.13 Hclust-Correlation-Ward

## 0.09 Hclust-Pearson-Ward

### 0.05 Kmediods-Cosine

0.04 Spectral clustering  
Symmetric  
(Metrics 1-6)

## Example Discovery

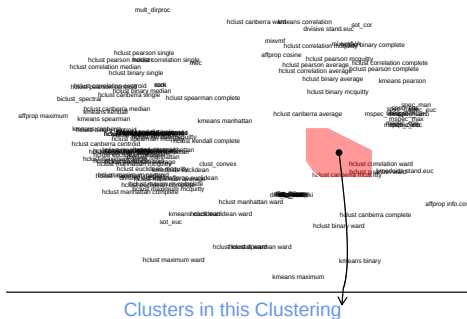


Credit Claiming  
Pork

**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

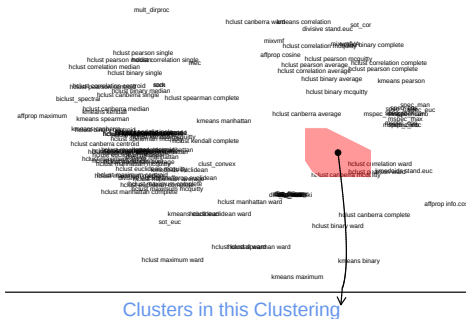


Mayhew Credit Claiming  
Legislation

## Credit Claiming, Legislation:

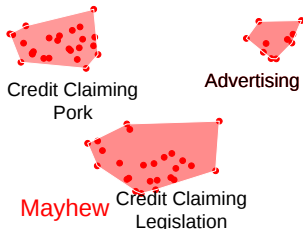
“As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period”

## Example Discovery

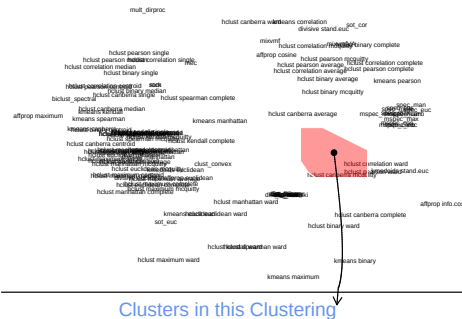


## Advertising:

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

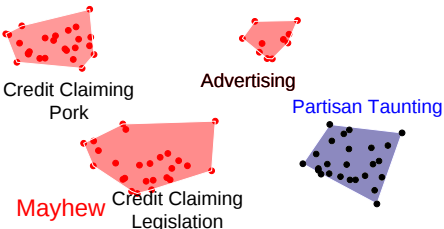


## Example Discovery: Partisan Taunting

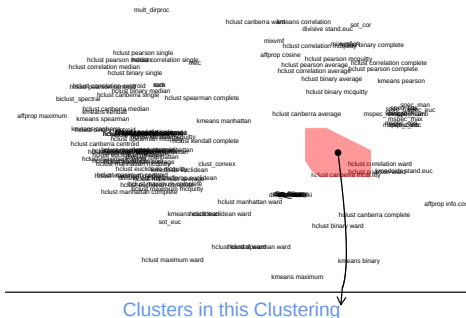


## Partisan Taunting:

## “Republicans Selling Out Nation on Chemical Plant Security”

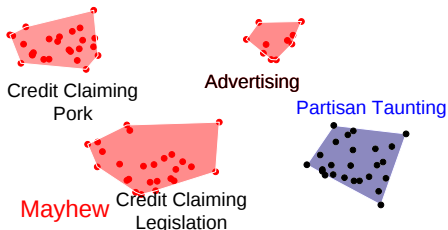


## Example Discovery: Partisan Taunting

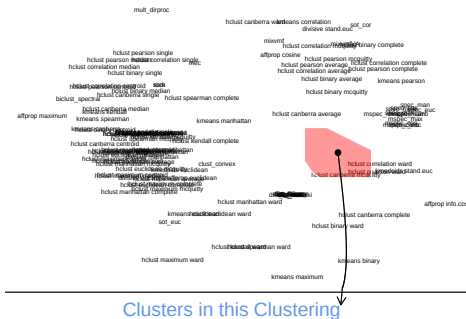


## Partisan Taunting:

“Senator Lautenberg’s amendment would change the name of ...the Republican bill...to ‘More Tax Breaks for the Rich and More Debt for Our Grandchildren Deficit Expansion Reconciliation Act of 2006’ ”

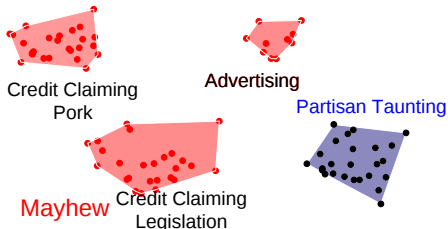


## Example Discovery: Partisan Taunting



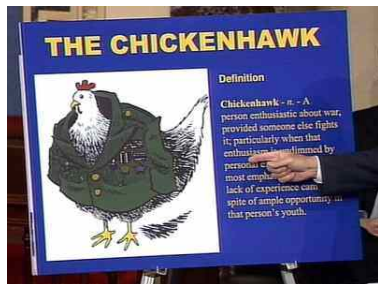
**Definition:** Explicit, public, and negative attacks on another political party or its members

**Taunting ruins deliberation**



# In Sample Illustration of Partisan Taunting

## Taunting ruins deliberation



Sen. Lautenberg  
on Senate Floor  
4/29/04

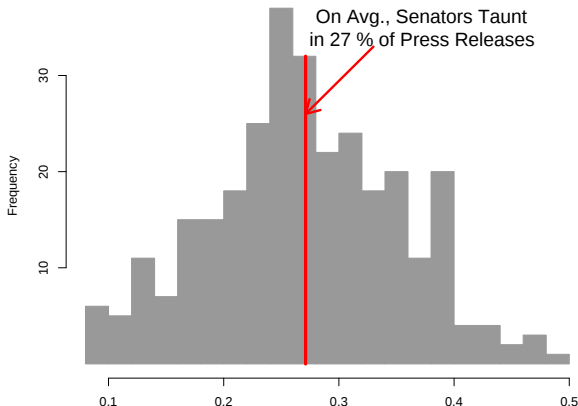
- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities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

# Out of Sample Confirmation of Partisan Taunting

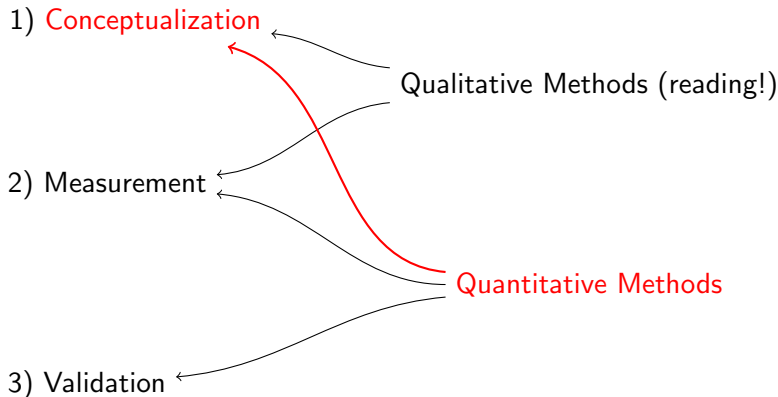
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure **proportion of press releases** a senator taunts other party

# Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure **proportion of press releases** a senator taunts other party



# Advancing the Objective of Discovery



Quantitative methods for conceptualization: aiding **discovery**

- Few formal methods designed explicitly for conceptualization
- **Belittled**: “Tom Swift and His Electric Factor Analysis Machine” (Armstrong 1967)
- Evaluation methods measure progress in discovery

For more information:

<http://GKing.Harvard.edu>