Quantitative Discovery of Qualitative Information: A General Purpose Document Clustering Methodology

Gary King
Institute for Quantitative Social Science
Harvard University

joint work with
Justin Grimmer (Harvard University)

(talk at the Northeast Methodology Program, New York University, 4/17/09)

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)

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- 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

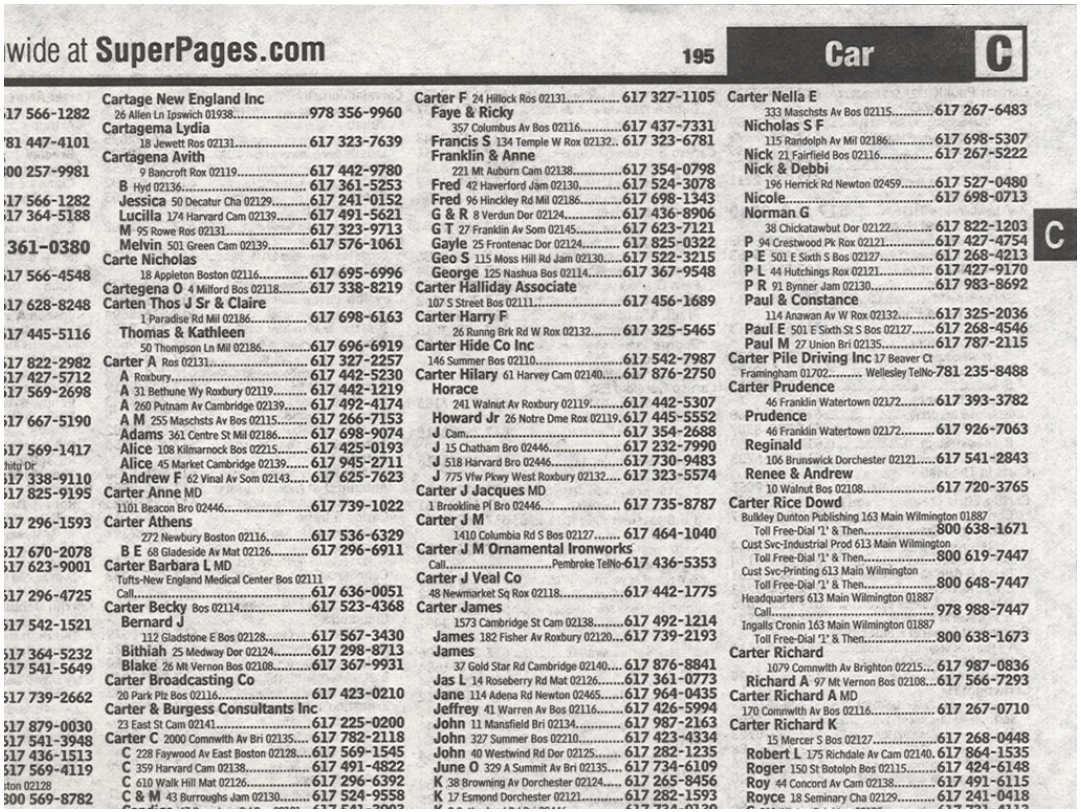
Why HAL Can't Classify Either

- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - Ugly duckling theorem: every pair of documents are equally similar $\rightsquigarrow$ every partition of documents is equally similar
  - 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 propogation, 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
  - Deriving such guidance: difficult or impossible
  - (Perhaps true by definition in unsupervised learning: If we knew the DGP, we wouldn't be at the discovery stage.)

If Ex Ante doesn't work, try Ex Post

- Methods and substance must be connected (no free lunch theorem)
- The usual approach fails: can't do it by understanding the model
- We do it ex post (by cherry-picking results)
  - For discovery (our goal): No problem
  - For estimation & confirmation: more difficult or biased
- Complicated concepts are easier to define ex post:
  - "I know it when I see it" (Justice Stewart's definition of obscenity)
  - Anchoring Vignettes (on defining concepts by example)
- But how to choose from an enormous list of clusterings?

Our Idea: Meaning Through Geography



wide at **SuperPages.com** 195 **Car** C

Cartage New England Inc
26 Allen Ln Ipswich 01938 ..... 978 356-9960
Cartagema Lydia
18 Jewett Ros 02131 ..... 617 323-7639
Cartagena Avith
9 Bancroft Rox 02119 ..... 617 442-9780
B Hyd 02136 ..... 617 361-5253
Jessica 50 Decatur Cha 02129 ..... 617 241-0152
Lucilla 174 Harvard Cam 02139 ..... 617 491-5621
M 95 Rowe Ros 02131 ..... 617 323-9713
Melvin 501 Green Cam 02139 ..... 617 576-1061
Carte Nicholas
18 Appleton Boston 02116 ..... 617 695-6996
Cartegena O 4 Milford Bos 02118 ..... 617 338-8219
Carten Thos J Sr & Claire
1 Paradise Rd Mil 02186 ..... 617 698-6163
Thomas & Kathleen
50 Thompson Ln Mil 02186 ..... 617 696-6919
Carter A Ros 02131 ..... 617 327-2257
A Roxbury 617 442-5230
A 31 Bethune Wy Roxbury 02119 ..... 617 442-1219
A 260 Putnam Av Cambridge 02139 ..... 617 492-4174
A M 255 Maschsts Av Bos 02115 ..... 617 266-7153
Adams 361 Centre St Mil 02186 ..... 617 698-9074
Alice 108 Kilmarnock Bos 02215 ..... 617 425-0193
Alice 45 Market Cambridge 02139 ..... 617 945-2711
Andrew F 62 Vinal Av Som 02143 ..... 617 625-7623
Carter Anne MD
1101 Beacon Bro 02446 ..... 617 739-1022
Carter Athens
272 Newbury Boston 02116 ..... 617 536-6329
B E 68 Gladeside Av Mat 02126 ..... 617 296-6911
Carter Barbara L MD
Tufts-New England Medical Center Bos 02111
Call 617 636-0051
Carter Becky Bos 02114 ..... 617 523-4368
Bernard J
112 Gladstone E Bos 02128 ..... 617 567-3430
Bithiah 25 Medway Dor 02124 ..... 617 298-8713
Blake 26 Mt Vernon Bos 02108 ..... 617 367-9931
Carter Broadcasting Co
20 Park Plz Bos 02116 ..... 617 423-0210
Carter & Burgess Consultants Inc
23 East St Cam 02141 ..... 617 225-0200
Carter C 2000 Comnwlth Av Bri 02135 ..... 617 782-2118
C 228 Faywood Av East Boston 02128 ..... 617 569-1545
C 359 Harvard Cam 02138 ..... 617 491-4822
C 610 Walk Hill Mat 02126 ..... 617 296-6392
C & M 43 Burroughs Jam 02130 ..... 617 524-9558
Carter F 24 Hillock Ros 02131 ..... 617 327-1105
Faye & Ricky
357 Columbus Av Bos 02116 ..... 617 437-7331
Francis S 134 Temple W Rox 02132 ..... 617 323-6781
Franklin & Anne
221 Mt Auburn Cam 02138 ..... 617 354-0798
Fred 42 Haverford Jam 02130 ..... 617 524-3078
Fred 96 Hinckley Rd Mil 02186 ..... 617 698-1343
G & R 8 Verdun Dor 02124 ..... 617 436-8906
G T 27 Franklin Av Som 02145 ..... 617 623-7121
Gayle 25 Frontenac Dor 02124 ..... 617 825-0322
Geo S 115 Moss Hill Rd Jam 02130 ..... 617 522-3215
George 125 Nashua Bos 02114 ..... 617 367-9548
Carter Halliday Associate
107 S Street Bos 02111 ..... 617 456-1689
Carter Harry F
26 Running Brk Rd W Rox 02132 ..... 617 325-5465
Carter Hide Co Inc
146 Summer Bos 02110 ..... 617 542-7987
Carter Hilary 61 Harvey Cam 02140 ..... 617 876-2750
Horace
241 Walnut Av Roxbury 02119 ..... 617 442-5307
Howard Jr 26 Notre Dme Rox 02119 ..... 617 445-5552
J Cam 617 354-2688
J 15 Chatham Bro 02446 ..... 617 232-7990
J 518 Harvard Bro 02446 ..... 617 730-9483
J 775 Vfw Pkwy West Roxbury 02132 ..... 617 323-5574
Carter J Jacques MD
1 Brookline Pl Bro 02446 ..... 617 735-8787
Carter J M
1410 Columbia Rd S Bos 02127 ..... 617 464-1040
Carter J M Ornamental Ironworks
Call ..... Pembroke TelNo 617 436-5353
Carter J Veal Co
48 Newmarket Sq Rox 02118 ..... 617 442-1775
Carter James
1573 Cambridge St Cam 02138 ..... 617 492-1214
James 182 Fisher Av Roxbury 02120 ..... 617 739-2193
James
37 Gold Star Rd Cambridge 02140 ..... 617 876-8841
Jas L 14 Roseberry Rd Mat 02126 ..... 617 361-0773
Jane 114 Adena Rd Newton 02465 ..... 617 964-0435
Jeffrey 41 Warren Av Bos 02116 ..... 617 426-5994
John 11 Mansfield Bri 02134 ..... 617 987-2163
John 327 Summer Bos 02210 ..... 617 423-4334
John 40 Westwind Rd Dor 02125 ..... 617 282-1235
June O 329 A Summit Av Bri 02135 ..... 617 734-6109
K 38 Browning Av Dorchester 02124 ..... 617 265-8456
K 17 Esmond Dorchester 02121 ..... 617 282-1593
Carter Nella E
333 Maschsts Av Bos 02115 ..... 617 267-6483
Nicholas S F
115 Randolph Av Mil 02186 ..... 617 698-5307
Nick 21 Fairfield Bos 02116 ..... 617 267-5222
Nick & Debbi
196 Herrick Rd Newton 02459 ..... 617 527-0480
Nicole 617 698-0713
Norman G
38 Chickatawbut Dor 02122 ..... 617 822-1203
P 94 Crestwood Pk Rox 02121 ..... 617 427-4754
P E 501 E Sixth S Bos 02127 ..... 617 268-4213
P L 44 Hutchings Rox 02121 ..... 617 427-9170
P R 91 Bynner Jam 02130 ..... 617 983-8692
Paul & Constance
114 Anawan Av W Rox 02132 ..... 617 325-2036
Paul E 501 E Sixth St S Bos 02127 ..... 617 268-4546
Paul M 27 Union Bri 02135 ..... 617 787-2115
Carter Pile Driving Inc 17 Beaver Ct
Framingham 01702 ..... Wellesley TelNo 781 235-8488
Carter Prudence
46 Franklin Watertown 02172 ..... 617 393-3782
Prudence
46 Franklin Watertown 02172 ..... 617 926-7063
Reginald
106 Brunswick Dorchester 02121 ..... 617 541-2843
Renee & Andrew
10 Walnut Bos 02108 ..... 617 720-3765
Carter Rice Dowd
Bulkley Dunton Publishing 163 Main Wilmington 01887
Toll Free-Dial '1' & Then 800 638-1671
Cust Svc-Industrial Prod 613 Main Wilmington
Toll Free-Dial '1' & Then 800 619-7447
Cust Svc-Printing 613 Main Wilmington
Toll Free-Dial '1' & Then 800 648-7447
Headquarters 613 Main Wilmington 01887
Call 978 988-7447
Ingalls Cronin 163 Main Wilmington 01887
Toll Free-Dial '1' & Then 800 638-1673
Carter Richard
1079 Comnwlth Av Brighton 02215 ..... 617 987-0836
Richard A 97 Mt Vernon Bos 02108 ..... 617 566-7293
Carter Richard A MD
170 Comnwlth Av Bos 02116 ..... 617 267-0710
Carter Richard K
15 Mercer S Bos 02127 ..... 617 268-0448
Robert L 175 Richdale Av Cam 02140 ..... 617 864-1535
Roger 150 St Botolph Bos 02115 ..... 617 424-6148
Roy 44 Concord Av Cam 02138 ..... 617 491-6115
Royce 18 Seminary Cha 02129 ..... 617 241-0418

⇝ We develop a geography of clusterings

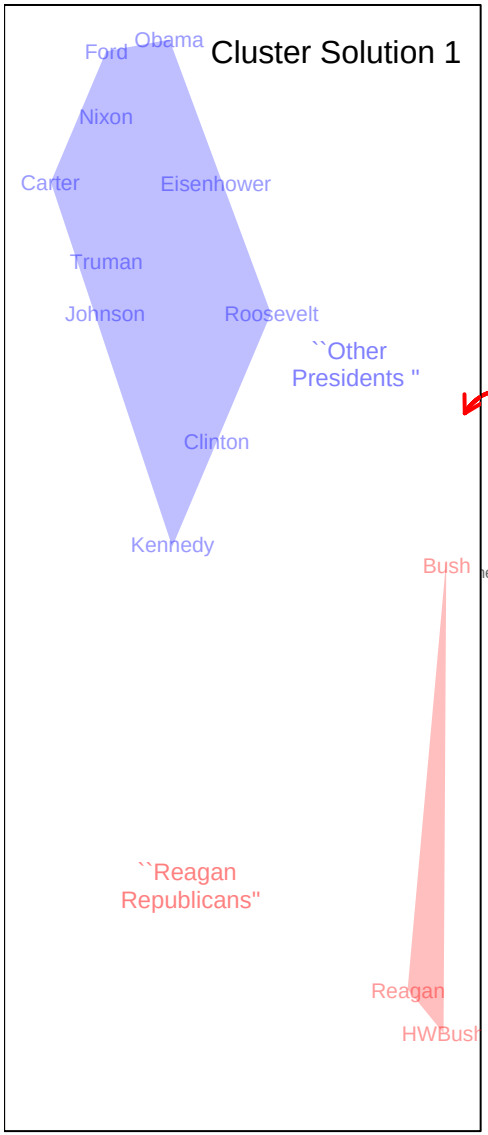
A New Strategy

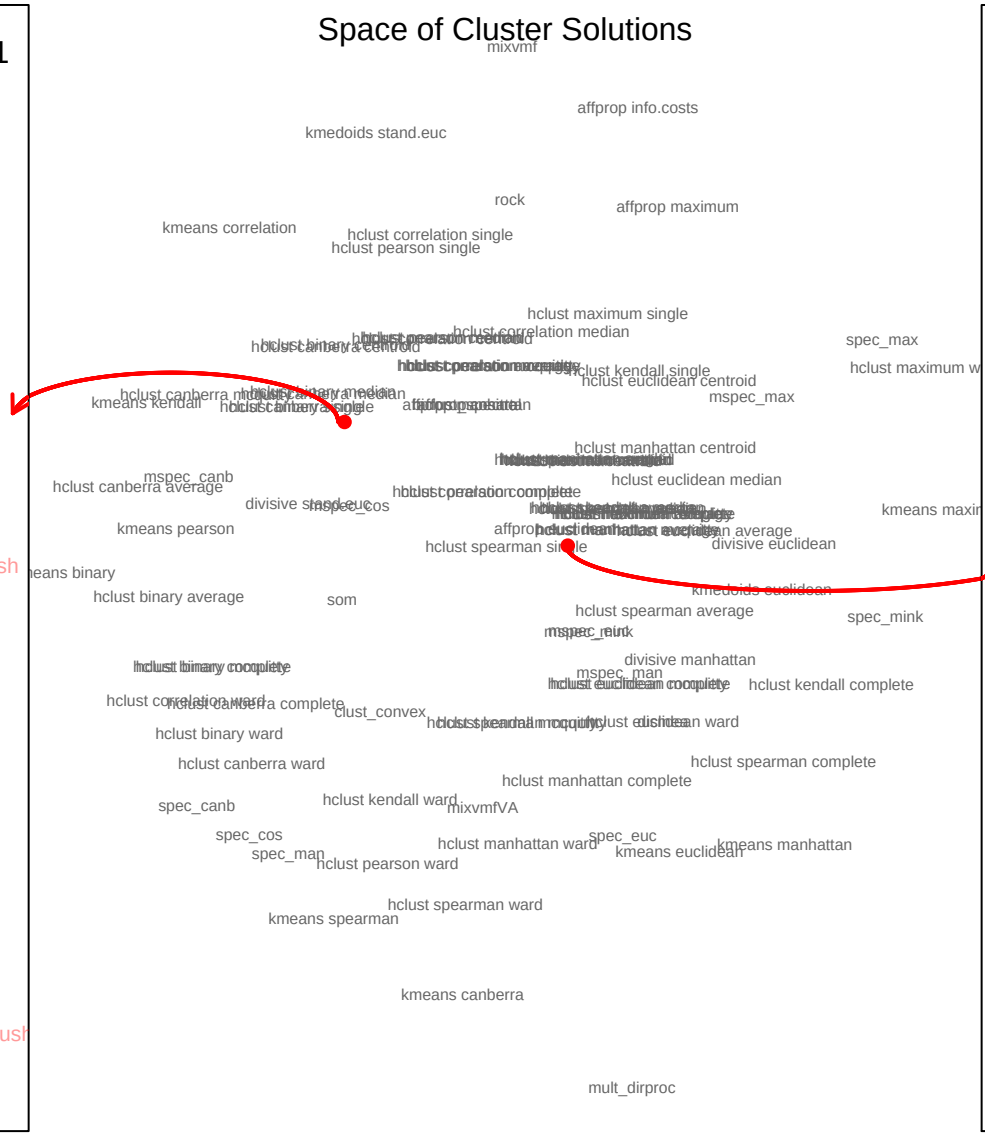
1. Code text as numbers (in one *or more* of several ways)
2. Apply all existing clustering methods (that have been used by at least one person other than the author) to the data — each representing different substantive assumptions ($<$15 mins)
3. Develop an application-independent distance metric between clusterings
4. Create a metric space of clusterings, and a 2D projection
5. Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
6. Propose a new animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)

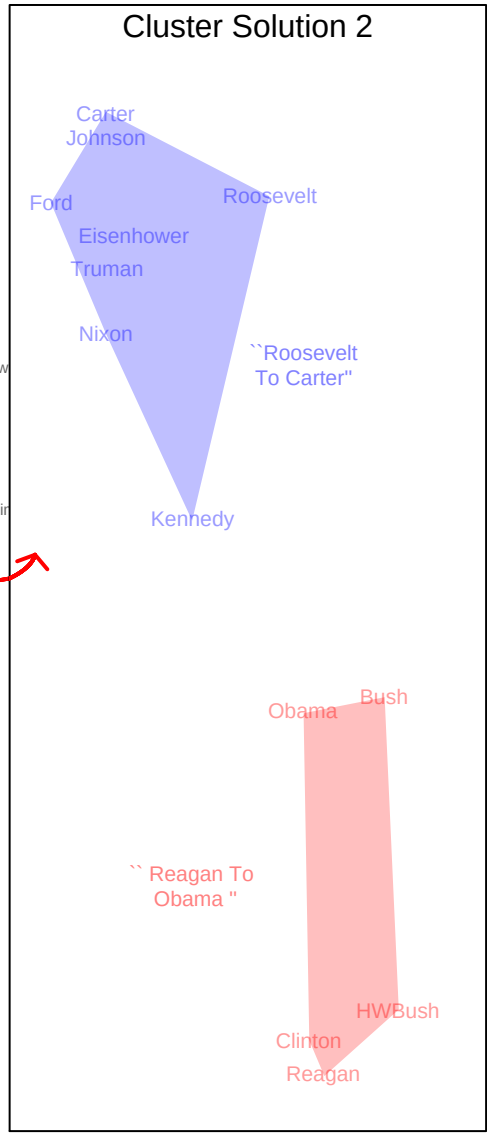
$\rightsquigarrow$ meaning revealed through a *geography* of clusterings

Many Thousands of Clusterings, Sorted & Organized

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



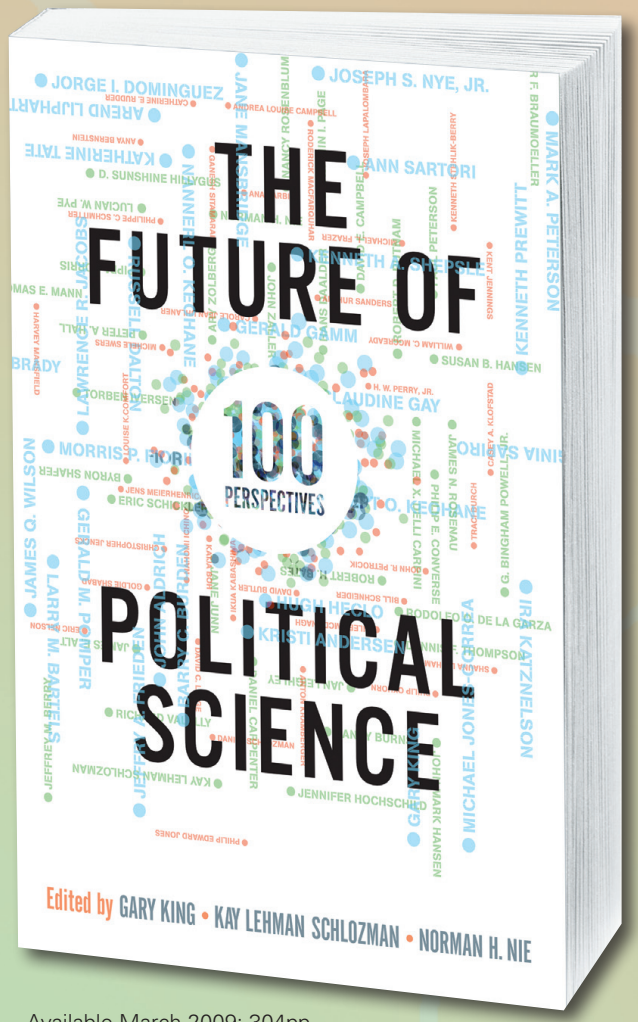




Application-Independent Distance Metric: Axioms

1. Clusterings with more pairwise document agreements are closer (we prove: pairwise agreements encompass 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(num clusters)

$\rightsquigarrow$ Only one measure satisfies all three (the "variation of information")



Available March 2009: 304pp
Pb: 978-0-415-99701-0: **$24.95**
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

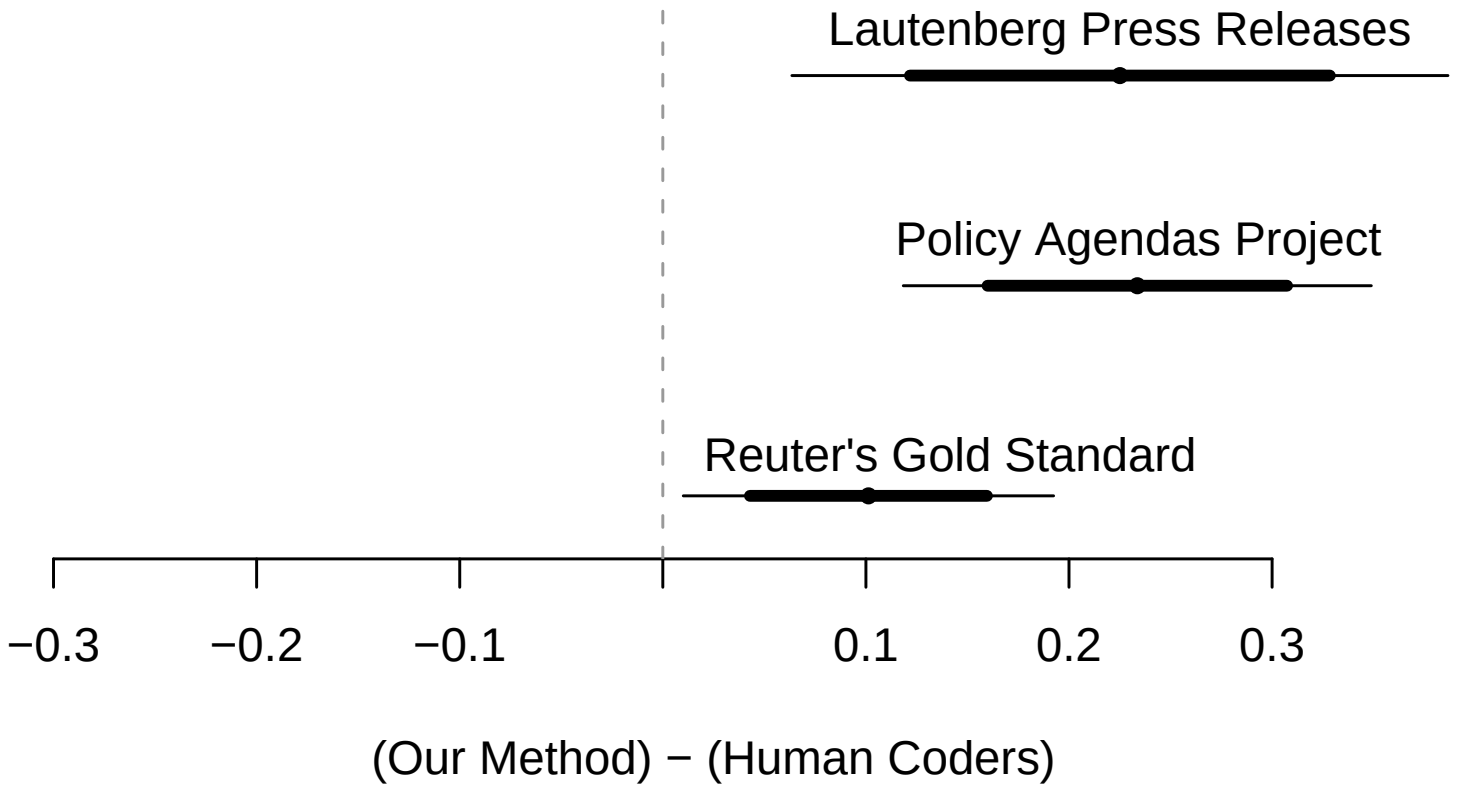
- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

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!

Cluster Quality Experiments

Scale: mean(within clusters) − mean(between clusters)



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

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

Reuter's: financial news (trade, earnings, copper, gold, coffee, . . .): "gold

What do Members of Congress Do?

Substantive example of a finding, using our approach

- David Mayhew's (1974) famous typology
  1. Advertising
  2. Credit Claiming
  3. Position Taking
- We find one more: Partisan Taunting
  - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
  - "The Intolerance and discrimination from the Bush administration against gay and lesbian Americans is astounding" [Civil Rights]
  - "John Kerry had enough conviction to sign up for the military during wartime, unlike the Vice President, who had a deep conviction to avoid military service" [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]
  - $\rightsquigarrow$ Perhaps this is what it means to be a member of a political party in the U.S.?

More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- For each: created 2 clusterings from each of 3 methods, including ours
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for $\binom{6}{2}$=15 pairwise comparisons
- 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

Last Points

- Intended contributions:
  - An encompassing cluster analytic approach for discovery
  - A new approach to evaluating results in unsupervised learning
  - Especially useful for the ongoing spectacular increase in the production and availability of unstructured text
- Future research:
  - Advancing our approach: (1) >2D exploration, (2) alternative visualizations of the space of clusterings, (3) including more methods
  - Evaluating new individual methods: (1) distance from existing methods and their averages, (2) usefulness of discoveries in given data sets.

For more information:

http://GKing.Harvard.edu