

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)

(talk at IQSS Mindich Text Analysis Conference, 5/30/09)

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
- **Qualitative-only approaches are hopeless**
- 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**
 - **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 qualitative choice)
 - 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

17 566-1282	Carriage New England Inc 36 Main St Ipswich 01938.....	978 356-9960
17 447-4101	Cartagena Lydia 19 Jordan Ave 02111.....	617 323-7639
800 257-9981	Cartagena Avelth F Pleasant Hill 02119.....	617 442-0780
17 566-1282	B Hef 02119.....	617 361-5253
17 364-5188	Jessica 50 Decatur Cir 02129.....	617 241-0152
17 364-5188	Luzilia 134 Harvard Cam 02139.....	617 491-5621
361-0380	M 95 Stone Rd 02114.....	617 323-9713
17 566-4548	Melvin 91 Green Can 02139.....	617 576-1061
17 628-8248	Carte Nicholas 18 Apollon Boston 02114.....	617 695-6996
17 445-5116	Cartesma O 4 Willow St 02118.....	617 338-8219
17 822-2982	Carton Tho & S Clave 1 Poydras Rd MO 02188.....	617 698-6163
17 627-5712	Thomas S Kuthbier 50 Thompson Ln 02186.....	617 696-6919
17 569-2698	Cartier A Ben 02111.....	617 327-2257
17 667-5190	A M 250 Main St 02119.....	617 442-5230
17 569-1417	Adams 381 Centre St 02116.....	617 492-4174
17 338-9110	Alice 108 Elmwood St 02115.....	617 698-9074
17 825-9158	Alice 45 Clark Cambridge 02139.....	617 425-0193
17 338-9110	Andrew F 22 Visual Ave 02143.....	617 945-2711
17 825-9158	Andrew F 22 Visual Ave 02143.....	617 625-7623
17 296-1593	Cartier Athens 777 Broadway Boston 02114.....	617 739-1022
17 670-2078	Cartier B Boston 02114.....	617 536-6329
17 623-7001	Cartier B Boston 02114.....	617 296-6911
17 296-4725	Carti 100 New England Medical Center Box 02111.....	617 636-0051
17 542-1521	Carti 20 Park Pl 02114.....	617 523-4368
17 364-5232	Carti 121 State St 02119.....	617 567-3430
17 541-5649	Carti 100 Main St 02114.....	617 298-8713
17 739-2662	Carti 20 Park Pl 02114.....	617 523-4368
17 679-8030	Cartier & Burgess Consultants Inc 71 East St 02114.....	617 225-0200
17 541-3948	Cartier C 2800 Commonwealth Ave 02116.....	617 782-2118
17 436-1513	Cartier C 2800 Commonwealth Ave 02116.....	617 569-1545
17 569-4119	C 359 Harvard Cam 02139.....	617 491-8222
17 569-4119	C 359 Harvard Cam 02139.....	617 296-6392
800 569-8782	C & B 41 Burroughs Ave 02118.....	617 328-9238
17 327-1105	Carter J 50 Woburn St 02116.....	617 327-1105
17 437-7331	Faye & Ricky 107 Columbus Ave 02116.....	617 437-7331
17 323-6963	Francis S 114 Temple Wn 02116.....	617 323-6963
17 354-4798	Franklin & Anne 221 Mt Auburn Cam 02138.....	617 354-4798
17 524-3078	Fred 45 Waverhill Ave 02139.....	617 524-3078
17 698-1343	Fred 50 Hinchey Rd MO 02136.....	617 698-1343
17 436-6906	G & R 100 Waverhill Ave 02139.....	617 436-6906
17 623-7121	G T 27 Franklin Ave 02114.....	617 623-7121
17 825-0322	Gayle 25 Promenade Dr 02139.....	617 825-0322
17 522-2215	Geo S 115 Waverhill Rd 02139.....	617 522-2215
17 367-9548	George 125 Nantua St 02114.....	617 367-9548
17 456-1689	Carter Halliday Associate 180 S Deyou St 02111.....	617 456-1689
17 325-5465	Carter Harry E 180 S Deyou St 02111.....	617 325-5465
17 542-7067	Carter Hide Co Inc 140 Summer St 02111.....	617 542-7067
17 876-2750	Carter Hilary 41 Newry Cam 02116.....	617 876-2750
17 442-5307	Horace 342 Woburn Ave 02119.....	617 442-5307
17 354-2680	Howard Jr 20 Nebra Dns 02119.....	617 354-2680
17 232-7990	J 15 Chatham St 02146.....	617 232-7990
17 730-9483	J 518 Harvard Bro 02146.....	617 730-9483
17 323-5574	J 775 Wey Plaz Wobury 02132.....	617 323-5574
17 739-1022	Carter J Jacques MD 1181 Beacon Hill 02116.....	617 739-1022
17 464-1040	Carter J M 1410 Columbia St S 02117.....	617 464-1040
17 343-3353	Carter J M Ornamental Ironworks 100 North St 02114.....	617 343-3353
17 442-1775	Carter J Veal Co 40 Newmarket St 02116.....	617 442-1775
17 492-1214	Carter James 1573 Cambridge St Cam 02138.....	617 492-1214
17 739-2193	James 1102 Fisher Ave Wobury 02138.....	617 739-2193
17 876-8941	James 1102 Fisher Ave Wobury 02138.....	617 876-8941
17 363-0773	James 1102 Fisher Ave Wobury 02138.....	617 363-0773
17 968-8435	James 114 Adams Rd Haverhill 02470.....	617 968-8435
17 426-5999	Jeffrey 41 Warren Ave 02116.....	617 987-2163
17 423-4334	John 127 Summer St 02111.....	617 423-4334
17 282-1235	John 127 Summer St 02111.....	617 282-1235
17 734-6199	June O 329 A Summit Ave 02116.....	617 734-6199
17 265-9456	K 38 Browning Ave Dorchester 02122.....	617 265-9456
17 282-1232	K 38 Browning Ave Dorchester 02122.....	617 282-1232
17 267-6483	Carter Nella E 323 Manchester Ave 02115.....	617 267-6483
17 498-5307	Nicholas S F 115 Northgate Ave 02186.....	617 498-5307
17 267-5222	Nick 21 Fairfield St 02116.....	617 267-5222
17 698-0713	Nick & Dobb 104 Harvard Rd Newton 02459.....	617 698-0713
17 822-1203	Noni 38 Chittenden Dr 02135.....	617 822-1203
17 427-4754	P 96 Croswell Pk 02119.....	617 427-4754
17 266-4213	P E 101 E South St 02116.....	617 266-4213
17 427-9170	P L 44 Huntington Box 02114.....	617 427-9170
17 983-5692	P R 25 Forest Cam 02138.....	617 983-5692
17 325-2036	Paul & Constance 114 Anson Ave W 02138.....	617 325-2036
17 268-4546	Paul F 303 E South St 02116.....	617 268-4546
17 787-2115	Paul M 27 Union St 02115.....	617 787-2115
281-235-8488	Prudence 40 Franklin Woburn 02179.....	617 393-3782
17 926-7063	Prudence 40 Franklin Woburn 02179.....	617 926-7063
17 541-2843	Reginald 386 Braintree Dorchester 02122.....	617 541-2843
17 720-3765	Renee & Andrew 30 Walnut St 02108.....	617 720-3765
800 638-1671	Carter Rice David Buckley Dutton Publishing 513 Main Wilmington 01887.....	800 638-1671
800 619-7447	Carl S 140 Free St T & Thon.....	800 619-7447
800 648-7447	Carl S 140 Free St T & Thon.....	800 648-7447
800 638-1673	Carl S 140 Free St T & Thon.....	800 638-1673
17 987-0836	Carl S 189 Convent Ave Brighton 02111.....	617 987-0836
17 566-7293	Carl S 189 Convent Ave Brighton 02111.....	617 566-7293
17 267-0710	Carl S 176 Convent Ave 02116.....	617 267-0710
17 268-0448	Carl S 176 Convent Ave 02116.....	617 268-0448
17 424-6148	Carl S 176 Convent Ave 02116.....	617 424-6148
17 491-6115	Carl S 176 Convent Ave 02116.....	617 491-6115
17 241-9418	Carl S 176 Convent Ave 02116.....	617 241-9418



↪ We provide a (conceptual) geography of clusterings

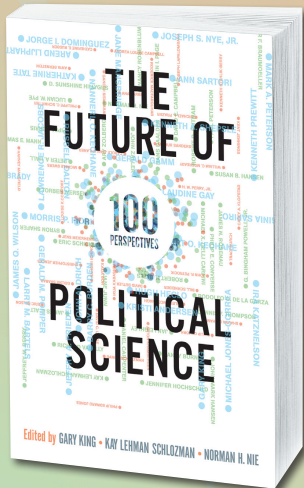
A New Strategy

- 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)

↪ meaning revealed through a geography of clusterings

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 $\log(\text{num clusters})$
- ↪ **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

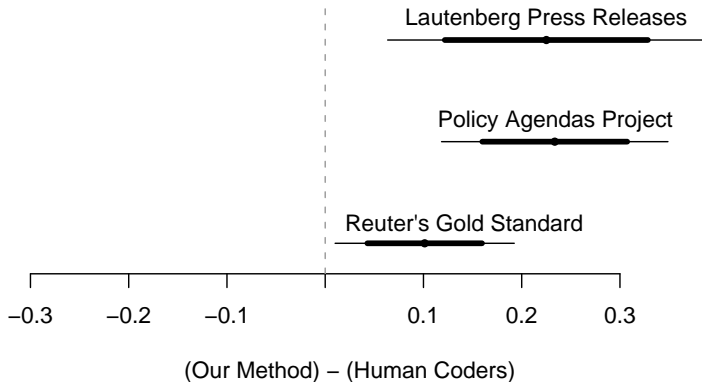
- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

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

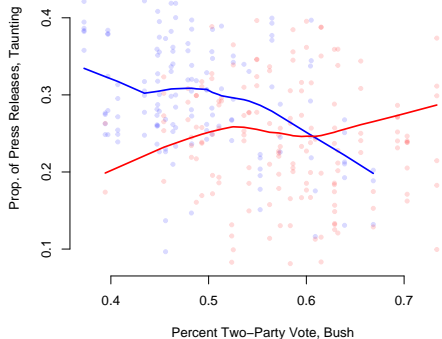
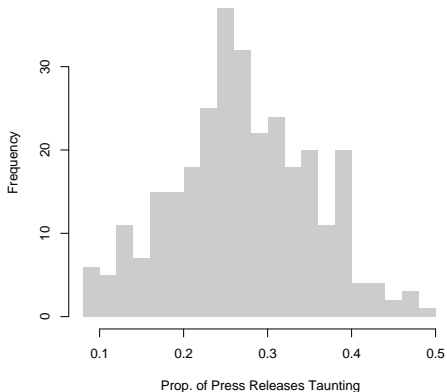
What do Members of Congress Do?

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 scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then." [Healthcare]
 - "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]
 - ↪ **Is this what "party in government" means?**

Partisan Taunting Hypothesis Verification

Discovered in $n = 200$ (1 senator); confirmed in $n = 64,033$ (301 senator-years)



Prevalence of partisan taunting (it happens a lot!)

Partisan taunting occurs more in lopsided districts

↪ Conflict: deliberation (no taunting) v. reflection (lopsided districts)

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 → vMF 1 → vMF 2 → Our Method 2 → K-Means 1 → K-Means 2

“Genetic testing”:

Our Method 1 → {Our Method 2, K-Means 1, K-means 2} → Dir Proc. 1 → Dir Proc. 2

- **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>