

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.

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

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).

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

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

Why Johnny Can't Classify (Optimally)

- $Bell(n)$ = number of ways of partitioning n objects
- $Bell(2) = 2$ (AB, A B)

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)

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)
- $Bell(5) = 52$

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)
- $Bell(5) = 52$
- $Bell(100) \approx$

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

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!

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

Why HAL Can't Classify Either

- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:

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

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

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:

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 propagation, self-organizing maps,...

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 propagation, self-organizing maps, . . .
 - **Well-defined** statistical, data analytic, or machine learning foundations

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 propagation, self-organizing maps, . . .
 - **Well-defined** statistical, data analytic, or machine learning foundations
 - How to add substantive knowledge:

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 propagation, self-organizing maps, . . .
 - **Well-defined** statistical, data analytic, or machine learning foundations
 - How to add substantive knowledge: With few exceptions, **who knows?!**

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
↔ 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 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**

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
↔ 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 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

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
↔ 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 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

If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)

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

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)

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

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

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:

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)

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)

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

Our Idea: Meaning Through Geography

wide at **SuperPages.com**

195

Car

C

87 566-1282	Cartage New England Inc 36 Allen St Ipswich 01938	978 356-9960
81 447-4101	Cartagena Lydia 28 Lowell St West 02155	617 323-7639
800 257-9981	Cartagena Aesth 9 Beacon St 02119	617 442-9780
87 566-1282	Carte Nicholas 30 Deane St 02129	617 241-0152
87 364-5188	Carte Nicholas 100 Tower St 02116	617 323-9713
361-0380	Carte Nicholas 100 Tower St 02116	617 576-1061
87 566-4548	Carte Nicholas 134 Appleton St 02114	617 695-6996
87 628-8248	Carten Thos Jr & Claire 4 Miller St 02118	617 338-8219
87 445-5116	Carten Thos Jr & Claire 1 Franklin St 02116	617 698-6163
87 822-2982	Carton & Kubitson 50 Thompson Ln 02126	617 696-6919
87 427-5712	Cartor A A Industry	617 442-5230
87 569-2608	Cartor A A 31 Redburn Wy 02119	617 442-1219
87 667-5190	Cartor A A 208 Putnam Av Cambridge 02139	617 492-4174
87 569-1417	Cartor A A 100 Essex St 02116	617 698-9074
87 338-9191	Cartor A A 45 Market Cambridge 02139	617 425-0193
87 825-1310	Cartor A A 222 W 4th St 02114	617 625-7623
87 296-1593	Cartor Arden 1101 Beacon St 02116	617 739-1022
87 670-2078	Cartor Barbara L MD 277 Rockwell Boston 02114	617 536-6329
87 623-9001	Cartor Barbara L MD B E 68 Gloucester Av Matt 02126	617 296-6911
87 296-4725	Cartor Beckey St 02114 Cartor Beckey St 02114	617 636-0051 617 523-4568
87 542-1521	Cartor Bernard J 117 Danvers St 02118	617 567-3430
87 364-8232	Cartor Bobbald 25 Middle St 02124	617 298-8713
87 541-5649	Cartor Broadcasting Co 26 Blake St 02116	617 367-9931
87 739-2662	Cartor Burgess Consultants Inc 30 Park Pl 02116	617 423-0210
87 879-0300	Cartor C 71 East St 02114	617 225-0200
87 541-3948	Cartor C 200 Cambridge St 02116	617 782-2118
87 436-1513	Cartor C 208 Faywood Av East Boston 02128	617 569-1545
87 569-4119	Cartor C 619 Hook Rd 02118	617 491-4822
800 599-8782	Cartor C & M 41 Barnstable Ave 02116	617 526-9236
87 327-1105	Carter F & Ricky 57 Columbia Av 02116	617 437-7331
87 323-4763	Carter F & Ricky 57 Columbia Av 02116	617 323-4763
87 354-4798	Carter Franklin & Anne 271 Mt Auburn Car 02138	617 354-4798
87 524-3078	Carter Fred 42 Woodmont Av 02116	617 524-3078
87 698-1343	Carter G B R 60 Waverly St 02114	617 698-1343
87 436-8906	Carter G B R 60 Waverly St 02114	617 436-8906
87 623-7121	Carter G T Jr 27 Franklin Av 02116	617 623-7121
87 825-0322	Carter Gayle 25 Frontview Dr 02114	617 825-0322
87 522-3215	Carter G J 155 Main St 02116	617 522-3215
87 367-9548	Carter George 125 Hazlett St 02114	617 367-9548
87 456-1689	Carter Holiday Associate 180 S Street 02116	617 456-1689
87 325-5465	Carter Harry E 180 S Street 02116	617 325-5465
87 542-7987	Carter Hide Co Inc 146 Summer St 02116	617 542-7987
87 076-2750	Carter Hilary 61 Harvey Car 02186	617 076-2750
87 442-5307	Carter Horace 942 Walnut Av 02119	617 442-5307
87 445-5552	Carter Howard Jr 20 Main Dr 02119	617 445-5552
87 354-2688	Carter J Dan J Dan	617 354-2688
87 232-7990	Carter J S 15 Chatham St 02144	617 232-7990
87 323-5574	Carter J S 775 The Pines Wy 02116	617 323-5574
87 735-8757	Carter J Jacques MD 1 Brookline Pl 02146	617 735-8757
87 464-1040	Carter J M 1433 Columbia St 02117	617 464-1040
87 634-5353	Carter J H Ornamental Ironworks Portland Falls 02116	617 634-5353
87 442-1775	Carter J Veal Co 40 Newport St 02116	617 442-1775
87 492-1214	Carter James 1573 Cambridge St 02118	617 492-1214
87 739-2193	Carter James 162 Fisher Av 02116	617 739-2193
87 876-8841	Carter James 27 Gold Star Rd Cambridge 02140	617 876-8841
87 361-0773	Carter James 114 Waverly St 02116	617 361-0773
87 964-9435	Carter James 114 Waverly St 02116	617 964-9435
87 428-5994	Carter Jeffrey 41 Warren Av 02116	617 428-5994
87 957-2163	Carter John 15 Montrose St 02116	617 957-2163
87 423-4334	Carter John 127 Summer St 02116	617 423-4334
87 282-1235	Carter John 101 Waverly St 02116	617 282-1235
87 734-6109	Carter John 209 A Summit Av 02116	617 734-6109
87 265-9456	Carter John 38 Browning Av 02116	617 265-9456
87 282-1593	Carter John 27 Forest St 02116	617 282-1593
87 267-6483	Carter Nella E 321 Main St 02116	617 267-6483
87 498-5307	Carter Nicholas S F 125 Northrup Av 02186	617 498-5307
87 267-5222	Carter Nick & Debbie 21 Fairport St 02116	617 267-5222
87 527-0480	Carter Nicole 194 Harvard Rd Newton 02459	617 527-0480
87 698-0713	Carter Norman G Norman G	617 698-0713
87 822-1203	Carter P 38 Chatham St 02116	617 822-1203
87 427-4754	Carter P E 101 S South St 02116	617 427-4754
87 427-9170	Carter P L 44 Middlebury St 02116	617 427-9170
87 983-8692	Carter Paul & Constance Paul & Constance	617 983-8692
87 325-2036	Carter Paul 114 Ansonia Av W 02130	617 325-2036
87 268-4546	Carter Paul M 303 S Main St 02117	617 268-4546
87 787-2115	Carter Paul M 37 Union St 02116	617 787-2115
87 926-8488	Carter Pile Driving Inc Pile Driving Inc	87 926-8488
87 393-3782	Carter Prudence 48 Francis Waltham 02127	617 393-3782
87 926-7063	Carter Prudence 48 Francis Waltham 02127	617 926-7063
87 541-2843	Carter Renee & Andrew 336 Ironsides Dr 02112	617 541-2843
87 720-3765	Carter Renee & Andrew 336 Ironsides Dr 02112	617 720-3765
800 638-1671	Carter Rice David 1000 Main St 02116	800 638-1671
800 619-7447	Carter Richard 741 Free St 02116	800 619-7447
800 648-7447	Carter Richard 1000 Main St 02116	800 648-7447
87 908-7447	Carter Richard 741 Free St 02116	87 908-7447
87 987-0836	Carter Richard 1879 Commonwealth Av Brighton 02215	87 987-0836
87 566-7293	Carter Richard A MD 1000 Main St 02116	87 566-7293
87 267-7010	Carter Richard K 130 Commonwealth Av 02116	87 267-7010
87 268-0448	Carter Robert L 15 Morris St 02117	87 268-0448
87 864-1535	Carter Roger 175 Nichols Av 02116	87 864-1535
87 424-6148	Carter Roy 100 Waverly St 02116	87 424-6148
87 241-0418	Carter Royce 100 Waverly St 02116	87 241-0418

Our Idea: Meaning Through Geography

wide at **SuperPages.com**

195

Car

C

17 566-1282
17 447-4101
800 257-9981
17 566-1282
17 364-5188
361-0380
17 566-4548
17 628-8248
17 445-5116
17 822-2982
17 427-5712
17 569-2608
17 667-5190
17 569-1417
17 338-8110
17 825-9195
17 296-1593
17 670-2078
17 623-9001
17 296-4725
17 542-1521
17 364-5232
17 541-5649
17 739-2612
17 879-0030
17 541-3948
17 436-1513
17 569-4119
800 569-8782

Cartage New England Inc
36 Abbott St Ipswich 01938... 978 536-9960
Cartagema Lydia
30 Liberty St Westfield 01191... 417 323-7639
Cartagema Aethi
9 Beacon St 02119... 617 442-9780
B Hill 02115... 617 361-5253
Jessica 100 Deaneur Cir 02129... 617 241-0152
Lucilla 174 Harvard Cn 02129... 617 491-5621
M 90 Tower St 02116... 617 323-9713
Melvin 301 Green Cn 02134... 617 576-1061
Carle Nicholas
18 Appleton Street 02114... 617 695-6996
Cartegena O Miller 80 St 02118... 617 338-8219
Carten Thos Jr Sr & Claire
1 Franklin St MA 02186... 617 698-6163
Thomas & Kathleen
50 Thompson Ln MA 02186... 617 696-6919
Carter A 801 02111... 617 327-2257
A Industry 02142-5230
A 31 Refuine Wy Roxbury 02119... 617 442-1219
A 208 Putnam Av Cambridge 02139... 617 492-4174
A M 255 Massachusetts Av 02115... 617 266-7153
Adams 301 Centre St MA 02186... 617 698-9074
Alice 108 Elmwood St 02115... 617 425-0193
Alice 45 Market Cambridge 02139... 617 945-2711
Andrew F 2244 Wy Ave 02143... 617 625-7623
Andrew D 02146... 617 739-1022
1101 Beacon St MA 02186... 617 739-1022
Carter B 02115
277 Beverly Boston 02114... 617 536-6329
B R & G Gloucester Av MA 02156... 617 296-6911
Carter Barbara L 02111
100 New England Medical Center Box 02111
Carter Becky 80 02114... 617 636-0051
Carter Bernard J
112 Grant St 02115... 617 567-3430
Bobbi 26 W Main St 02124... 617 290-8713
Blake 26 W Main St 02128... 617 367-9931
Carter Broadcasting Co
50 Park Pl 02114... 617 423-0210
Carter & Burgess Consultants Inc
71 East St 02114... 617 225-0200
Carter C 2000 Commonwealth Av 02116... 617 782-2118
C 208 Faywood St East Boston 02128... 617 269-1545
219 Harvard Cn 02129... 617 491-4822
C 618 Hook Mt 02118... 617 296-6392
C & M 41 Burroughs Ave 02116... 617 524-2324

Carter F 20 Inwood St 02117... 617 327-1105
Faye & Ricky
577 Chatham Av 02116... 617 437-7331
Francis S 154 Temple W 02128... 617 323-4763
Franklin & Anne
271 Mt Auburn Cn 02138... 617 354-4708
Fred 42 Woodmont Av 02138... 617 524-3078
Fred 50 Haverhill St MA 02186... 617 698-1343
G B R 177 Franklin St 02124... 617 436-8906
G T 27 Franklin Av 02116... 617 623-7121
Gayle 25 Frontenac Dr 02124... 617 825-0322
Geo S 155 Mount Mt St 02128... 617 522-3215
George 125 Haddon St 02114... 617 367-9548
Carter Halliday Associate
180 S Tower St 02111... 617 456-1689
Carter Harry P
25 Irving St 02118... 617 325-5465
Carter Hide Co Inc
146 Summer St 02115... 617 542-7987
Carter Hilary 61 Harvey Cn 02186... 617 676-2750
Horace
342 Walnut Av Roxbury 02119... 617 442-5307
Howard Jr 30 Main Dn St 02119... 617 445-5552
J Can 617 354-2688
J 25 Chatham St 02144... 617 232-7990
J S Harvard Av 02146... 617 730-9483
J 775 The Ferry Wood Roxbury 02111... 617 323-5574
Carter J Jacques MD
1 Brookline Pl 02146... 617 735-8787
Carter J M
1433 Columbia St S 02117... 617 464-1040
Carter J H Ornamental Ironworks
Portland Telco 617 436-3553
Carter J Veal Co
40 Newport St 02118... 617 442-1775
Carter James
573 Cambridge St Cam 02138... 617 492-1214
James 102 Fisher Av Roxbury 02119... 617 739-2193
James
27 Gold Star Rd Cambridge 02140... 617 876-8841
John 151 W Main St 02118... 617 361-0773
John 114 Adams Rd Newton 02459... 617 964-9435
Jeffrey 41 Warren Av 02116... 617 426-5994
John 11 Montpelier St 02118... 617 957-2163
John 127 Summer St 02118... 617 423-4334
John 41 Montpelier St 02118... 617 282-1235
John O 209 A Summit Av 02128... 617 734-6109
K 38 Browning Av Dorchester 02124... 617 265-9456
K 27 Forest St 02115... 617 282-1593

Carter Nella E
321 Main St 02115... 617 267-6483
Nicholas S F
125 Northrop Av MA 02186... 617 698-5307
Nick & Debbie
134 Harvard Mt Newton 02459... 617 267-5222
Nicole
146 Harvard Mt Newton 02459... 617 527-0480
Norman G
38 Chickadee Dr 02129... 617 822-1203
P H Croswell Pl 02128... 617 427-4754
P L 101 South St 02127... 617 248-0211
P L 44 Haddon St 02114... 617 427-9170
P R 21 Beacon Av 02114... 617 983-8692
Paul & Constantine
114 Anson Av W 02130... 617 325-2036
Paul 6 103 S South St 02127... 617 268-0546
Paul M 27 Union St 02115... 617 787-2115
Prudence
40 Francis Wadsworth 02127... 617 393-3782
Prudence
100 Franklin Wadsworth 02127... 617 986-7063
Reynold
336 Brookline Dorchester 02121... 617 541-2843
Renée & Andrew
30 Walnut St 02118... 617 720-3765
Carter Rice David
Berkley Station Parkside 013 Main Wilmington 01887
Toll Free 800 T & Thru... 800 638-1671
Call Cent. Industrial Prod 613 Main Wilmington 01887
Toll Free 800 T & Thru... 800 619-7447
Call Cent. Printing 613 Main Wilmington 01887
Toll Free 800 T & Thru... 800 648-7447
Headquarters 613 Main Wilmington 01887
Toll Free 800 T & Thru... 978 908-7447
Logistics Center 163 Main Wilmington 01887
Toll Free 800 T & Thru... 800 638-1673
Carter Richard
1879 Commonwealth Av Brighton 02117... 617 987-0836
Richard 101 Mt Vernon St 02128... 617 566-7293
Carter Richard A MD
130 Commonwealth Av 02116... 617 267-0710
35 Morris St 02137... 617 268-0408
Robert L 175 Rochdale Av 02126... 617 864-1535
Roy 100 St Andrews St 02115... 617 424-6148
Roy 44 Concord Av 02118... 617 491-6115
Boyce 18 Seabury Cir 02129... 617 241-6112



Our Idea: Meaning Through Geography

wide at SuperPages.com

195

Car

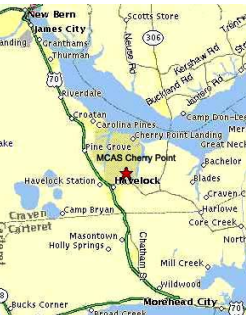
C

87 566-1282
81 447-4101
800 257-9981
87 566-1282
87 364-5188
361-0380
87 566-4548
87 628-8248
87 445-5116
87 822-2982
87 427-5712
87 569-2608
87 667-5190
87 569-1417
87 338-8110
87 825-9195
87 296-1593
87 670-2078
87 623-9001
87 296-4752
87 424-1521
87 364-8232
87 541-5649
87 739-2612
87 879-0030
87 541-3948
87 436-1513
87 569-4119
800 569-8782

Cartage New England Inc
36 Abbott St Ipswich 01938... 978 356-9960
Cartagena Lydia
30 Liberty St West 02155... 617 323-7639
Cartagena Aethi
9 Beach St Boston 02119... 617 442-9780
B Hill 02130... 617 361-3253
Jessica 50 Decatur Cir Boston 02129... 617 241-0152
Lucilla 174 Harvard Ave 02139... 617 491-5621
M 90 Tower St 02138... 617 303-9713
Melvin 321 Green Can 02134... 617 576-1061
Carle Nicholas
18 Appleton Street 02116... 617 695-6996
Cartena G 4 Miller St 02118... 617 338-8219
Carten Thos Jr & Claire
1 Franklin St MA 02188... 617 698-6163
Thomas & Kathleen
50 Thompson Ln MA 02186... 617 696-6919
Carter A 801... 617 327-2257
A Industry... 617 442-5230
A 31 Refuse Wy Roxbury 02119... 617 442-1219
A 208 Putnam Ave Cambridge 02139... 617 492-4174
A M 255 Massachusetts Ave 02115... 617 266-7153
Adams 381 Centre St MA 02188... 617 698-9074
Alice 108 Elmwood Ave 02125... 617 425-0193
Alice 45 Market Cambridge 02139... 617 945-2711
Andrew F 42 West Ave 02143... 617 625-7623
Andrew D 02144... 617 739-1022
1181 Beacon St 02188... 617 739-1022
Carter Arthens
277 Riverside Boston 02114... 617 536-6329
B E de 60 Spaulding Ave MA 02126... 617 296-6911
Carter Barbara L 02116
100 New England Medical Center Box 02115
Carter Becky 50 02134... 617 636-0051
C... 617 523-4568
Bernard J
112 Glen St E Boston 02128... 617 567-3430
Bobbi 25 Winthrop Dr 02124... 617 290-8713
Blake 26 Mt Vernon St 02188... 617 367-9931
Carter Broadcasting Co
50 Park Pl St 02116... 617 423-0210
Carter & Burgess Consultants Inc
71 East St 02089... 617 225-0200
Carter C 200 Commonwealth Ave 02116... 617 782-2118
C 209 Fenway Ave East Boston 02128... 617 569-1545
219 Harvard Ave 02139... 617 491-4822
C 619 Brook Hill MA 02138... 617 296-6392
C & M 41 Burroughs Ave 02118... 617 524-2524

Carter F 24 Inwood Ave 02115... 617 327-1105
Faye & Ricky
577 Chatham Ave MA 02116... 617 437-7331
Francis S 154 Temple W 02128... 617 323-6763
Franklin & Anne
271 Mt Auburn Ave 02138... 617 354-4708
Fred 47 Westwood Ave 02136... 617 434-3078
Fred 16 Haverhill St MA 02188... 617 698-1343
G & R 100 West St 02124... 617 436-8906
G T 27 Franklin Ave 02136... 617 623-7121
Gayle 25 Frontenac Dr 02134... 617 825-0322
Geo S 155 Mount Hill St MA 02128... 617 522-3215
George 125 Hudson Ave 02114... 617 367-9548
Carter Holiday Associate
189 S Street St 02111... 617 456-1689
Carter Harry F
315 Irving St MA 02116... 617 325-5465
Carter Hide Co Inc
146 Summer St 02113... 617 542-7987
Carter Hilary 61 Harvey Can 02188... 617 076-2750
Horace
362 Walnut Ave Roxbury 02119... 617 442-5307
Howard Jr 30 Main Drive West 02124... 617 445-5552
J Can
4 25 Chatham St 02144... 617 232-7990
J S 518 Harvard Ave 02146... 617 730-9483
J 775 The Pinery West Roxbury 02131... 617 323-5574
Carter J Jacques MD
1181 Beacon St 02188... 617 739-1022
Carter J M
1433 Columbia St S 02127... 617 464-1040
Carter J M Ornamental Ironworks
Portland Telco 617 434-3553
Carter J Veal Co
40 Newmarket St 02118... 617 442-1775
Carter James
1573 Cambridge St Can 02138... 617 492-1214
James
102 Fisher Ave 02138... 617 739-2193
James
27 Gold Star Rd Cambridge 02140... 617 876-8841
John 124 Woodland St MA 02118... 617 361-0773
Jane 114 Adams Rd Newton 02459... 617 964-9435
Jeffrey 41 Warren Ave 02116... 617 426-5994
John 11 Montpelier St 02116... 617 957-3163
John 127 Summer St 02118... 617 423-4334
John 41 Westfield Rd 02126... 617 282-1235
John O 209 A Summit Ave 02126... 617 734-6109
K 38 Browning Ave Dorchester 02216... 617 265-9456
K 27 Forest Road 02115... 617 442-2142

Carter Nella E
321 Main St 02115... 617 267-6483
Nicholas S F
125 Northrop Ave MA 02186... 617 698-5307
Nick & Debbie
124 Harvard Rd Newton 02459... 617 267-5222
Nicole
194 Harvard Rd Newton 02459... 617 527-0480
Norman G
38 Chickadee Dr 02129... 617 822-1203
P 34 Croswell Pl R 02128... 617 427-4754
P E 161 E South St 02127... 617 248-0213
P L 44 Holliston Ave 02124... 617 427-9170
P R 21 Boyer Ave 02118... 617 983-8692
Paul & Constance
114 Ansonia Av W 02130... 617 325-2036
Paul E 303 E South St 02127... 617 268-0546
Paul M 27 Union St 02135... 617 787-2115
Prudence
40 Francis Wadsworth 02127... 617 393-3782
Prudence
100 Franklin Wadsworth 02127... 617 986-7063
Reginald
336 Waverland Dr 02121... 617 541-2843
Renée & Andrew
30 Walnut St 02108... 617 720-3765
Carter Rice David
Berkley Station Parkside 015 Main Wilmington 01887
Tall Free-Deal 'T & Th... 800 638-1671
Cust Sec-Industrial Prod 613 Main Wilmington 01887
Tall Free-Deal 'T & Th... 800 619-7447
Cust Sec-Printing 613 Main Wilmington 01887
Tall Free-Deal 'T & Th... 800 648-7447
Headquarters 613 Main Wilmington 01887
Tel 978 908-7447
Logistics 163 Main Wilmington 01887
Tel Free-Deal 'T & Th... 800 638-1673
Carter Richard
1879 Commonwealth Ave Brighton 02111... 617 987-0836
Richard R 101 Main Avenue St 02124... 617 566-7293
Carter Richard A MD
130 Commonwealth Ave 02116... 617 267-0710
Carter Richard K
15 Morser S 02137... 617 268-0408
Robert L 175 Rochdale Ave Can 02136... 617 864-1535
Roger 124 St Nicholas Ave 02115... 617 424-6148
Roy 44 Concord Ave 02138... 617 491-6115
Boyce 18 Seabury Cir 02129... 617 241-6112



~ We develop a geography of clusterings

A New Strategy

A New Strategy

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

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)

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

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

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

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)

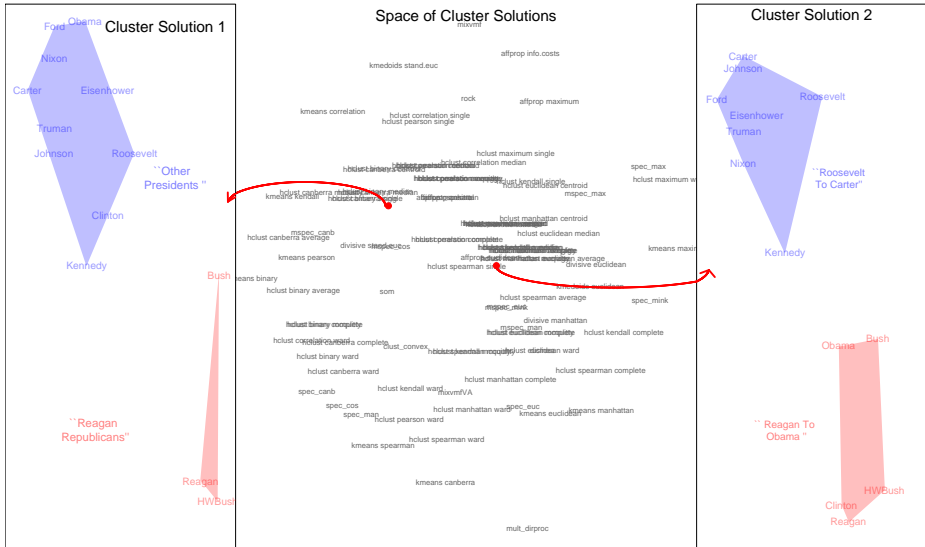
A New Strategy

- 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

Many Thousands of Clusterings, Sorted & Organized

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



Application-Independent Distance Metric: Axioms

Application-Independent Distance Metric: Axioms

- 1 Clusterings with more **pairwise document agreements** are closer (we prove: pairwise agreements encompass triples, quadruples, etc.)

Application-Independent Distance Metric: Axioms

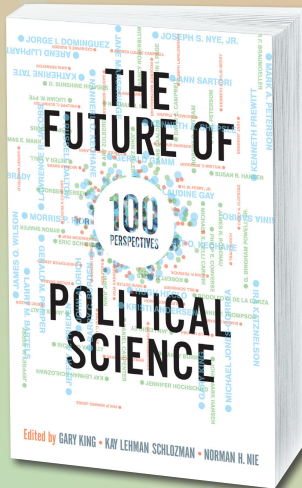
- 1 Clusterings with more **pairwise document agreements** are closer (we prove: pairwise agreements encompass triples, quadruples, etc.)
- 2 **Invariance**: Distance is invariant to the number of documents (for any fixed number of clusters)

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 $\log(\text{num clusters})$

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

Evaluators' Rate Machine Choices Better Than Their Own

- Scale: (1) unrelated, (2) loosely related, or (3) closely related

Evaluators' Rate Machine Choices Better Than Their Own

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

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

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

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

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

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

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!

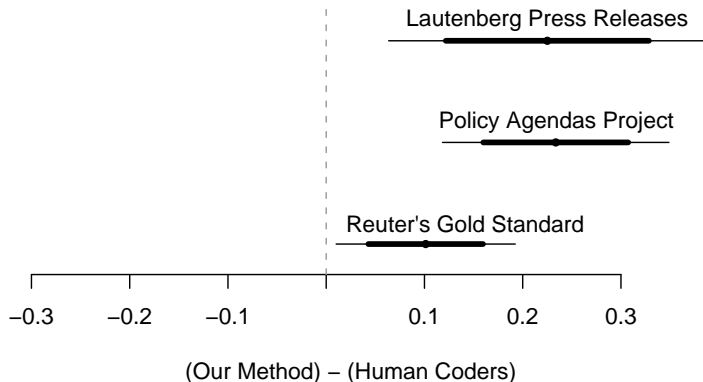
Cluster Quality Experiments

Cluster Quality Experiments

Scale: $\text{mean}(\text{within clusters}) - \text{mean}(\text{between clusters})$

Cluster Quality Experiments

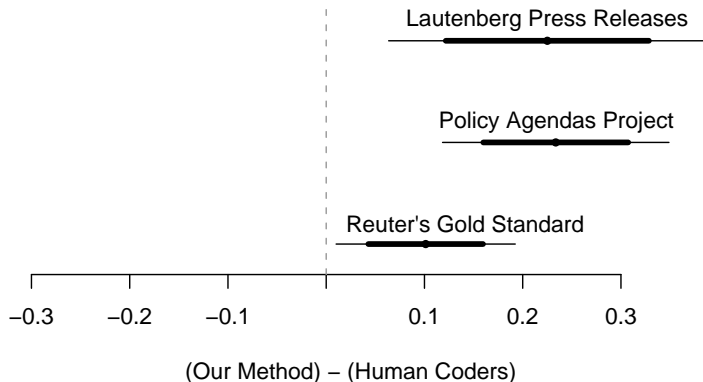
Scale: mean(within clusters) – mean(between clusters)



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

Cluster Quality Experiments

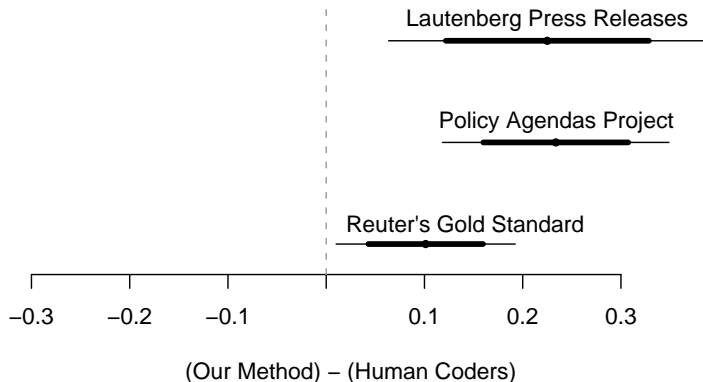
Scale: $\text{mean}(\text{within clusters}) - \text{mean}(\text{between clusters})$



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

Cluster Quality Experiments

Scale: $\text{mean}(\text{within clusters}) - \text{mean}(\text{between clusters})$



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

What do Members of Congress Do?

What do Members of Congress Do?

Substantive example of a finding, using our approach

What do Members of Congress Do?

Substantive example of a finding, using our approach

- David Mayhew's (1974) famous typology

What do Members of Congress Do?

Substantive example of a finding, using our approach

- David Mayhew's (1974) famous typology
 - ① Advertising

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

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

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**

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]

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]

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]

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]

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]
 - ↪ **Perhaps this is what it means to be a member of a political party in the U.S.?**

More Informative Discoveries

More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work

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

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)

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

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:

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

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

Last Points

- Intended contributions:

- **Intended contributions:**
 - An encompassing cluster analytic approach for discovery

- **Intended contributions:**
 - An encompassing cluster analytic approach for discovery
 - A new approach to evaluating results in unsupervised learning

- **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

- **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:**

- **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:

- **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,

- **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,

- **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

- **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:

- **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,

- **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>