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

## Gary King

#### Institute for Quantitative Social Science Harvard University

### Talk at BAE Systems, 9/9/2010

### Joint work with Justin Grimmer (Harvard ~> Stanford)

• Systematic method for computer-assisted conceptualization from text

ヨト イヨト

- Systematic method for computer-assisted conceptualization from text
- Conceptualization through 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).

- Systematic method for computer-assisted conceptualization from text
- Conceptualization through 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: simultaneously 1) invent categories and 2) assign documents to categories

- Systematic method for computer-assisted conceptualization from text
- Conceptualization through 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: simultaneously 1) invent categories and 2) assign documents to categories
- (We focus on texts, our methods apply more broadly)

# Why Johnny Can't Classify (Optimally)

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

• Clustering seems easy; its not!

- ★檀卜 ★恵卜 ★恵卜

- Clustering seems easy; its not!
- Bell(n) = number of ways of partitioning n objects

- ● 思 ▶ | ● 思 ▶

- Clustering seems easy; its not!
- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)

- \* 聞 ▶ - \* 恵 ▶ - \* 恵 ▶ -

- Clustering seems easy; its not!
- 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)

|(四) (日) (日)

- Clustering seems easy; its not!
- 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

- \* 聞 ▶ - \* 恵 ▶ - \* 恵 ▶ -

- Clustering seems easy; its not!
- 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)  $\approx$

/翻▶ 《图▶ 《图▶ /

- Clustering seems easy; its not!
- 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
- $\mathsf{Bell}(100) \approx 10^{28} \times$  Number of elementary particles in the universe

→ 圖▶ → 毛恵 ▶ →

- Clustering seems easy; its not!
- 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
- $\mathsf{Bell}(100)\approx 10^{28}\times$  Number of elementary particles in the universe
- Now imagine choosing the optimal classification scheme by hand!

- Clustering seems easy; its not!
- 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
- $\mathsf{Bell}(100)\approx 10^{28}\times$  Number of elementary particles in the universe
- Now imagine choosing the optimal classification scheme by hand!
- Its no surprise that automated algorithms can help, but which algorithms?

- 《聞》 《思》 《思》 -

Gary King (Harvard IQSS)

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

# Why HAL Can't Classify Either

• Large quantitative literature on cluster analysis

- \* 聞 > - \* 图 > - \* 图 >

# Why HAL Can't Classify Either

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

- 4 同 6 - 6 国 6 - 4 国 6

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

- (三) (三) (三)

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

- (三) (三) (三)

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

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

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

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

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

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

- ● 图 ▶ | ● 图 ▶

- Large quantitative literature on cluster analysis
- The Goal an optimal application-independent cluster analysis method is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graphbased, fuzzy *k*-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible
- Deep problem in cluster analysis literature: no way to know which method will work ex ante
- No surprise: everyone's tried cluster analysis; very few are satisfied

・ロト ・聞ト ・ヨト ・ヨト

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

Gary King (Harvard IQSS)

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

## • Methods and substance must be connected (no free lunch theorem)

高▶ 《高▶

- Methods and substance must be connected (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model

- 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:

- 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

- 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!

- 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

- 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

ヨト くまト

- 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
- The Question: How to organize all those clusterings?

Set of clusterings

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

# Set of clusterings $\approx$ A list of unconnected addresses

wide at	SuperPages.com	195	Car C
NO.10, SHITT	Cartage New England Inc	Carter F 24 Hillock Res 00131	Carter Nella E
7 566-1282	26 Alles Ln Ipswich 01938	Faye & Ricky	333 Maschets Av Bos 02115
		357 Calumbus Av Bos 02116	Nicholas S F 115 Randolph Av Mil 02185
81 447-4101	13 Jewett Ros 02131	Francis S 134 Temple W Rox 02132. 617 323-6781	115 Randolph Av Mil 02186
	Cartagena Avith	Franklin & Anne	Nick 21 Fairfeld Bos 02115
0 257-9981	9 Baecroft Ros 02119	221 Mt Auborn Cam (0138	196 Herrick Rd Newton 00457
	B Hyd 02136	Fred 42 Haverland Jam 02130	Nicole
7 566-1282	Jessica 50 Decatar Cha 02229	Fred of Heckley Rd Mil 02185	Norman G
7 364-5188	Lucilla 1/4 Harvard Cam (2139 617 491-5621		38 Chickstawbet Der 02122
	M 75 Rowe Ros 00133	G T 27 Frankin Av Son 02145	P 96 Crestwood Pk Ray (21)21
861-0380	Melvin 500 Green Cam 02139		P # Crestwood Pk Rax (2121
State State	Carte Nicholas	Geo S 115 Moss Hill Rd Jam 02130617 522-3215 George 125 Nashua Bos 02114617 367-9548	P L 44 Hutchings Rox (2121
7 566-4548	18 Appleton Baston 02116	George 125 hashua 6xs 02114617 367-9548 Carter Halliday Associate	P R 91 Bunner Jan 02130
	Cartegena O 4 Millard Bos 02138 617 338-8219	Carter Halliday Associate	Paul & Constance
7 628-8248	Carten Thos J Sr & Claire	107 5 Street Bos 02111	114 Anaryan Ar W Rax (2)32
Participant	1 Peradise Rd Mil 02186 617 698-6163	Carter Harry F	Paul F stor F Sets ST S Res (27)27. 617 268-454
7 445-5116	Thomas & Kathleen	26 Runnig Brit Rd W Rox 02132 617 325-5465 Carter Hide Co Inc	Paul M 27 Union Bri (2135
	50 Thompson Ln Mil 02286		Carter Pile Driving Inc 17 Beaver Ct
7 822-2982	Carter A Res 02131	146 Summer Bos 02110	Framingham (0/102
7 427-5712	A Roebery	Carter Hilary 61 Harvey Can 02140617 876-2750	Carter Prudence
7 569-2698	A 31 Bethune Wy Roxbury 02119 617 442-1219	Horace (37 440 5367	46 Frankin Watertown (21/2
	A 260 Putnam Ar Cambridge 02239 617 492-4174	241 Walnut Av Roebury 02119	Prodence
7 667-5190	A M 255 Maschels Ar Bos 02115 617 266-7153	Howard Jr 25 Notes Dine Res 00119, 617 445-5552	di Frankin Watertman (0172
	Adams 361 Centre St Mil 02186 617 698-9074	J Can	Reginald
7 569-1417	Alice 108 Kilmarnock Bos 02215 617 425-0193	J 538 Harvard Bro 02446	106 Brunswick Dorchester (2121617 541-284)
to Dr	Alice 45 Market Cambridge 02139 617 945-2711	J 538 Harvard Bro 0246	Rence & Andrew
7 338-9110	Andrew F 42 Vinal &r Som 02143 617 625-7623		10 Walnut Bos 02108
7 825-9195	Carter Anne MD 1165 Beacon Bro 00446	Carter J Jacques MD 1 Brooking Pl Bra 62445	Carter Rice Dowd
			Builder Dunton Publishing 163 Main Wilmington 01887
7 296-1593	Carter Athens	Carter J M	Toll Free-Dial '1' & Then
	272 Newbury Boston 02116	1410 Columbia Rd S Bos 02127 617 464-1040 Carter J M Ornamental Ironworks	
7 670-2078	B E 68 Gladeside Av Mat 02126 617 296-6911	Call	Toll Free-Dial '1' & Then
7 623-9001	Carter Barbara L MD	CalPenerola 1840-017 430-3333	Cust Svc-Printing 613 Main Wilmington
	Tufts-New England Medical Center Bes 02111 Call	Carter J Veal Co 68 Meansacket So Res 62118	Toll Free-Dial '2' & Then
7 296-4725	Carter Becky 8x5 (21)4	Carter James	Headquarters 613 Main Wilnington 00887
ALL HINKS	Carter Becky 9/5 E2194		Call
7 542-1521		1573 Cambridge St Cam 02138617 492-1214	Ingaits Cronin 163 Main Wilmington 01887
	112 Gladstone E Bes 02128	James 182 Fisher Av Roebury 02120617 739-2193	Toll Free-Dial '2' & Then
7 364-5232	Bithiah 25 Medway Der 02124	James (17.07/ 001)	Carter Richard
7 541-5649	Blake 25 Mt Vernon Bos 02108	37 Gold Star Rd Cambridge 02140 617 876-8841	1079 Commetth Av Brighton 02215 617 987-0830
	Carter Broadcasting Co	Jas L 14 Reseberry Rd Mat 02126 617 361-0773	Richard A \$7 Mt Vernos Bos 02108617 566-729.
7 739-2662	20 Park Ptz Box 02116	Jane 114 Adena Rd Newton 02465617 964-0435 Jeffrey 41 Warren & Bos 02115617 426-5994	Carter Richard A MD 170 Committe Av Bos 02116
	Carter & Burgess Consultants Inc	Jeffrey 41 Warren Ar Bos 02116 617 426* 5994	170 Commwith Av Bos 02116 617 267-071
17 879-0030	23 East St Cam 02241	John 11 Manafield Bri 02134	Carter Richard K
7 541-3948	Carter C 2000 Committe Ar Bri 02135 617 782-2118	John 327 Summer Bos 02210	15 Mercer S Bos 02127
7 436-1513	C 228 Faywood Av East Sector 02128 617 569-1545	John 40 Westwind Rd Der 02125 617 282-1235	Robert L 175 Richdale Av Cam 02140. 617 864-153.
17 569-4119	C 359 Harvard Cam 02138	June O 229 A Summit Av Bri 02135 617 734-6109	Roger 150 St Botniph Bos (2115 617 424-614
on 02128	C 633 Walk Hill Mat 07126	K 38 Browning Av Dorchester 02124 617 265-8456	Roy 44 Cancord Av Cam 02138
00 569-8782	C & M 43 Burroughs Jam 02130 617 529-9558	K 17 Esmond Dorchester 02121	Royce 18 Seminary Cha 02129

-78 ►

#### Set of clusterings $\approx$ A list of unconnected addresses

#### wide at SuperPages.com Cartage New England Inc Carter F 24 Hillock Ros 02131. 617 32 17 566-1282 978 356-9960 Faye & Ricky ma Lydia 357 Columbus Av Bos 02136. 617 43 617 323-7639 Francis S 134 Temple W Rax 02132. 617 32 \$1 447-4101 18 Jewett Ros 02131. Jeena Avith Franklin & Anne 00 257-9981 617 442-978 .617 361-5253 .617 241-0152 .617 491-5621 .617 323-9713 17 566-12 red 96 Hinckley Rd Mil 0218 wills the second Con 222 20 617 576-106 361-0380 Melvin S01 Green Cam 0223 617 695-6996 17 566-4548 18 Appieton Baston 0213 96 125 Nashua Bos 02114 ena O 4 Millard Bos 02138 Thos J Sr & Claire 617 338-8219 Carter I alliday Associate 617 45 17 628-8248 617 698-6163 Carter Harry F 617 32 17 445-5116 mas & Kathleen 25 Runna Brk Rd W Rox 0213 617 696-6919 Carter Hide Co Inc 17 822-2982 arter A Rus (013) 617 442-523 617 442-523 617 442-121 Carter Hilary 61 Harvey Can (22)40. 617 87 17 569-2698 Horace 617 492-4174 241 Walnut Av Roebury (2119) 5,617 44 17 667-5190 Howard Jr 25 Notre Die Rox 021 617 698-907 dams 341 Centre S2 MI 0218 17 569-1417 Alice 108 Kilmarnock Bos 02215 617 425-0193 lice 45 Market Cambridge 021 617 945-271 617 625-762 617 73 17 338-9110 Andrew F 62 Vinal Ar Som 0214 17 825-9195 Carter Anne MD 617 739-1022 617 73 17 296-1593 Carter Athens 617 46 617 536-6329 mbia Rd S Bos 02227 Carter J M Ornamental Ironworks 17 670-2078 Pembroka Tellio-617 43 Carter J Veal Co afts-New England Medical Center Bo 617 626-0051 17 296-4725 Carter Becky 805 82 617 523-4368 17 542-1521 Bernard J .617 567-3430 .617 298-8713 .617 367-9931 Tames 182 Fisher Av Roebury 02120...617 73 datone E Bes 02128 517 364-5232 517 541-5649 Bithiah 25 Medway Dor 02124. Blake 25 Mt Vernos Bes 00108 Carter Broadcasting Co 617 423-0210 617 96 517 739-2662 Jane 114 Adena Rd Newton 02465. Carter & Rurness Consultants offrey 41 Warren &r Bos (2116 617 225-0200 617 98 23 East St Cam 02240 ... hn 11 Manafield Bri 02134 Carter C 2000 Commeth Ar Bri 02135. 617 782-2118 517 436-1513 ohn 40 Westwind Rd Der 0212 617 491-4822 617 73 517 569-4119 C & M 43 Burroughs Jam 02230. 100 569-8782

190	Udi		U
-1105	Carter Nella E	1 X Mar	Contract,
	333 Maschets Av Bos 02115		67-6483
-7331	Nicholas S F	22008	
-6781	115 Randolph Av Mil 02186	617 6	98-5307
10015	Nick 21 Farteld Bos 02116	617 2	67-5ZZZ
-0798	196 Herrick Rd Newton 00/657	4175	0000-000
-1343	Nicole	(174	27-0480
-1343	Norman G		30-0/13
-7121	38 Chickstawbet Dor 02122	417 0	22-1202
-0322	P 94 Crestwood Pk Rox 02121	617.4	27-4754
-3215	P F St) F Sym S Bas (01127	617 2	68-4213
8429-1			
	P R 91 Dynner Jam 02130	617 9	83-8692
-1689	Paul & Constance		
	114 Anawan Ar W Rox 02132		25-2036
-5465	Paul E 501 E Sinth St S Bos 02127	617 2	68-4546
	Paul M 27 Usion Bri 02135		87-2115
-7987	Carter Pile Driving Inc 17 Beaver	α	
-2750	Framingham (0/02 Wellesley Te Carter Prudence	ap-781 2	35-8488
-5307	46 Frankin Watertown (21/2	117.7	
-5307	46 Frankin Waterlown 021/2 Prodence		93-3702
-2688	46 Frankin Watertown 021/2	617.0	26-7062
-7990	Reginald		20-7003
-9483	106 Brunswick Dorchester 02121.	617 5	41-2843
-5574			
	10 Walnut Bos 02108		20-3765
-8787	Carter Rice Dowd		
	Bulkley Duntoe Publishing 163 Main Wi	Imington 01	887
-1040	Toll Free-Dial '1' & Then		38-1671
	Cust Svc-Endustrial Prod 613 Main Wiln	sington .	10-7447
5-5353	Toll Free-Dial 12" & Then. Cust Svo-Pyinting 613 Main Wilmington		73-3443
200	Toll Error Find '2' & Then	800.6	48-7447
2-1775	Toll Free-Dial '2' & Then. Headquarters 613 Main Wilnington 00	87	
-1214			88-7447
-1214	Ingails Cronie 163 Main Wilmington 01	887	
	Toll Free-Dial '2' & Then		38-10/3
5-8841	Carter Richard 3079 Commetth Av Brighton 0222.		
-0773	Richard A \$7 Mt Wemon Bas 0210	01/ 7	67-0830
1-0435			
5-5994	170 Committh Av Bos 02116	617.2	67-0710
7-2163			
8-4334	15 Meerer S Ros 02327		68-0448
2-1235	Robert L 175 Richdale Av Cam 025	40.617 8	64-1535
1-6109			
5-8456			
2-1593	Royce 18 Seminary Cha 02129	617 2	41-0418

Car

C



イロト イポト イヨト イヨト

#### Set of clusterings $\approx$ A list of unconnected addresses

wide at	SuperPages.com	195	
YEAR SHOULD	Cartage New England Inc	Carter F 24 Hillock Res 00131	Carte
17 566-1282	36 Alles (n Ipswich 01938	Faye & Ricky	
81 447-4101	Cartagema Lydia 38 Jewett Ros 02231	357 Calambas &r Bos 02136	Nic
81 447-4101	Cartagena Avith	Franklin & Anne	Nie
00 257-9981	9 Rancraft Ros (2719	221 Mt Auburn Cam (0138	Nit
	B Hyd 02136	Fred 42 Haverland Jam 02130	
17 566-1282	Jessica 50 Decatar Cha 02229	Fred 96 Hinckley Rd Mil 02185	Nit
17 364-5188	Lucilla 1/4 Harvard Cam 02139 617 491-5621	G & R 3 Verdun Der 02124	NO
	M 75 Rowe Res 00131	Gave 25 Frontenac Dar 02134	P
361-0380	Carte Nicholas	Geo S 115 Mass Hill Rd Jam 00130 617 522-3215	PI
17 566-4548	18 Appieton Barton 00116	George 125 Nashia Bos 02114	PI
17 300 4540	Cartegena Q 4 Millard Box 02118 617 338-8219	Carter Halliday Associate	PI
17 628-8248	Carten Thos J Sr & Claire	107 5 Street Bos 02111	Pa
Antonio	1 Peradise Rd Mil 02186 617 698-6163	Carter Harry F	202
17 445-5116	Thomas & Kathleen	26 Runng Brk Rd W Rax 02132 617 325-5465	Pa
and the second	50 Thompson Ln MI 02286	Carter Hide Co Inc 146 Semmer Bos 02110	Carte
17 822-2982	Carter A 8us (013)	Carter Hilary 61 Harver Can 02140617 876-2750	Iran
17 427-5712	A Notbury	Horace	Carte
11, 203.5038	A 260 Putnam Ar Cambridge 02139 617 492-4174	241 Walnut Av Roebery 02119	
17 667-5190	A M 255 Maschets Ar Bos (2115 617 266-7153	Howard Jr 25 Notes Die Res (0119, 617 445-5552	Pr
11 001 0110	Adams 361 Centre 52 Mil 00186 617 698-9074	J Can	
17 569-1417	Alice 108 Kilmarnock Bes 02215 617 425-0193	J 15 Chatham Bro 02445	Re
NO DE	Alice 45 Market Cambridge 02139 617 945-2711	J 538 Harvard Bro 02445	1000
17 338-9110	Andrew F 62 Vinal Ar Som 02143 617 625-7623	J 775 Whe Plony West Roebury 02132 617 323-5574	Re
17 825-9195	Carter Anne MD 1165 Beacon Bro 00446	Carter J Jacques MD 1 Brooking Pl Bra 02445	Carte
17 296-1593	1100 Beacon Bro 00446	Carter J M	2.8.6
17 296-1593	772 Newbury Rodan 02116	1410 Columbia Rd S Res 02227 617 464-1040	T
17 670-2078	B E 48 Giadeside Av Mat 02125	Carter J M Ornamental Ironworks	C. CUST
17 623-9001	Carter Barbara L MD	Call	Ti Cust
	Tufts-New England Medical Center Bos 02111	Carter J Veal Co	Lusi
17 296-4725	Call	48 Newstarket Sq Rox (2138	Hea
and or Hilder	Carter Becky %% (21)4	Carter James 1573 Cambridge St Cam 02138617 492-1214	C
517 542-1521	Bernard J 112 Gadstore E Bes (2)28	James 182 Fisher Av Roebury 02120617 739-2193	Ing
17 364-5232	Bithiah 25 Medway Der 02124	James	Carte
517 541-5649	Blake 25 Mt Vernos Bos 02108	37 Gold Star Rd Cambridge 02140 617 876-8841	caro
11. 241-2041	Carter Broadcasting Co	Jas L 34 Roseberry Rd Mat 02226 617 361-0773	Ri
517 739-2662	20 Park Ptr Bes 02116	Jane 114 Adena Rd Newton 02465 617 964-0435	Carte
	Carter & Burgess Consultants Inc	Jeffrey 41 Warren & Bos 02116 617 426-5994	170
517 879-0030	23 East St Cam 02241	John 11 Manafield Bri 02134 617 987-2163	Carb
517 541-3948	Carter C 2000 Committe Ar Bri 02135 617 782-2118	John 327 Summer Bos 02218	122
517 436-1513	C 228 Faywood Av East Boston 0228617 569-1545 C 359 Harvard Can 02138	John 40 Westwind Rd Der 02125 617 282-1235 June O 129 A Sammit &r Bri 02135 617 734-6109	Re
517 569-4119	C 339 Harvard Cam 02138	K 38 Browning Av Dorchenter 02134	Ro
tos 02128	C & M 43 Barrosobs Jan 02230	K 17 Ermond Deerbaster (22)21 617 282-1593	Re

Gar	U	
Nella E		
133 Maschets Av Bos 02115		
115 Randolph Av Mil 02186	617 698-5307	
k 21 Fairleid Bos 02116 k & Debbi		
196 Herrick Rd Newton 00459	617 527-0490	
ole man G		
88 Chickatawbut Dor 02122	617 822-1203	C
Crestwood Pk Rax 02121	617 427-4754	C
501 E South S Bos 02127	617 268-4213	
44 Hutchings Rox 02121 91 Byrner Jam 02130	617 427-9170	
114 Anawan As W Rox 02132		
E 501 E Sinth St S Bos 02127	617 268-4546	
M 27 Union Bri 02135 Pile Driving Inc 17 Beaver	C	
ngham 00,702 Wellesky Te Prudence		
6 Franklin Watertown 02172 dence		
66 Frankin Watertown 02172 inald		
106 Brunswick Dorchester 02121.		$\sim$
0 Watnut Bos 02108	617 720-3765	
w Duriton Publishing 163 Main Wi	Iminoton 01887	
Free-Dial "1" & Then.		
Free-Dial '1' & Then		
Free-Dial '2' & Then. puarters 613 Main Wilmington 03		
Consist 147 Main Wilminster /17	978 988-7447	
Free-Olal '2' & Then Richard		
3579 Commelth Av Brighton 0221	617 987-0836	
hard A \$7 Mt Vernos Bos 0210 Richard A MD		
Richard K		
15 Meeter S Ros 02127		
sert L 125 Sichdale Av Cam 025	40.617 864-1535	
peter 150 St Rotelinh Ros 02115		
44 Cancord Av Cam 02138		

C-



.≣ ⊳

#### $\rightsquigarrow$ We develop a (conceptual) geography of clusterings

Gary King (Harvard IQSS)

## A New Strategy

Make it easy to choose best clustering from millions of choices

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

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

(王)▶ (王)▶

- Code text as numbers (in one *or more* of several ways)
- Apply all clustering methods we can find to the data each representing different (unstated) substantive assumptions (<15 mins)</p>

- Code text as numbers (in one *or more* of several ways)
- Apply all clustering methods we can find to the data each representing different (unstated) substantive assumptions (<15 mins)</p>
- (Too much for a person to understand, but organization will help)

- Code text as numbers (in one *or more* of several ways)
- Apply all clustering methods we can find to the data each representing different (unstated) substantive assumptions (<15 mins)</p>
- (Too much for a person to understand, but organization will help)
- Develop an application-independent distance metric between clusterings, a metric space of clusterings, and a 2-D projection

- Code text as numbers (in one *or more* of several ways)
- Apply all clustering methods we can find to the data each representing different (unstated) substantive assumptions (<15 mins)</p>
- (Too much for a person to understand, but organization will help)
- Develop an application-independent distance metric between clusterings, a metric space of clusterings, and a 2-D projection
- "Local cluster ensemble" creates a new clustering at any point, based on weighted average of nearby clusterings

ヨト イヨト

- Code text as numbers (in one *or more* of several ways)
- Apply all clustering methods we can find to the data each representing different (unstated) substantive assumptions (<15 mins)</p>
- (Too much for a person to understand, but organization will help)
- Develop an application-independent distance metric between clusterings, a metric space of clusterings, and a 2-D projection
- "Local cluster ensemble" creates a new clustering at any point, based on weighted average of nearby clusterings
- A new animated visualization to explore the space of clusterings (smoothly morphing from one into others)

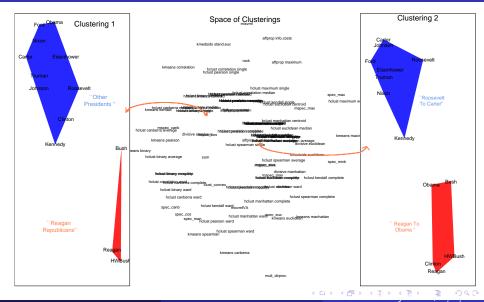
(温)▶ | ◆ 温)▶

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

- ▲圖▶ ▲圖▶ ▲圖▶

## Many Thousands of Clusterings, Sorted & Organized

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



Gary King (Harvard IQSS)

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

• Metric based on 3 assumptions

(王)▶ (王)▶

#### • Metric based on 3 assumptions

Oistance between clusterings: a function of the pairwise document agreements (pairwise agreements ⇒ triples, quadruples, etc.)

- Oistance between clusterings: a function of the pairwise document agreements (pairwise agreements ⇒ triples, quadruples, etc.)
- Invariance: Distance is invariant to the number of documents (for any fixed number of clusters)

- Oistance between clusterings: a function of the pairwise document agreements (pairwise agreements ⇒ triples, quadruples, etc.)
- Invariance: Distance is invariant to the number of documents (for any fixed number of clusters)
- Scale: the maximum distance is set to log(num clusters)

- Oistance between clusterings: a function of the pairwise document agreements (pairwise agreements ⇒ triples, quadruples, etc.)
- Invariance: Distance is invariant to the number of documents (for any fixed number of clusters)
- Scale: the maximum distance is set to log(num clusters)
- ~ Only <u>one</u> measure satisfies all three (the "variation of information")

- Oistance between clusterings: a function of the pairwise document agreements (pairwise agreements ⇒ triples, quadruples, etc.)
- Invariance: Distance is invariant to the number of documents (for any fixed number of clusters)
- Scale: the maximum distance is set to log(num clusters)
- ~ Only <u>one</u> measure satisfies all three (the "variation of information")
- Meila (2007): derives same metric using different axioms (lattice theory)

## **Evaluating Performance**

Gary King (Harvard IQSS)

2

▲口▶ ▲圖▶ ▲国▶ ▲国▶

#### • Goals:

2

▲口▶ ▲圖▶ ▲国▶ ▲国▶

- Goals:
  - Validate Claim: computer-assisted conceptualization outperforms human conceptualization

- (三) (三) (三)

\_\_ ▶

- Validate Claim: computer-assisted conceptualization outperforms human conceptualization
- Demonstrate: new experimental designs for cluster evaluation

ヨト イヨト

- Validate Claim: computer-assisted conceptualization outperforms human conceptualization
- Demonstrate: new experimental designs for cluster evaluation
- Inject human judgement: relying on insights from survey research

ヨト くまト

- 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

高▶ 《高▶

- 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

- ◆ 唐 ▶

- 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
  - $\bullet~$  Informative discoveries  $\Rightarrow~$  Experienced scholars analyzing texts

- 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
  - $\bullet~$  Informative discoveries  $\Rightarrow$  Experienced scholars analyzing texts
  - Discovery  $\Rightarrow$  You're the judge

Gary King (Harvard IQSS)

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

3

(王) (王)

\_\_ ▶

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

ヨト イヨト

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

ヨト イヨト

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

高▶ 《高▶

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time
- ullet  $\implies$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time
- $\implies$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality
  - automated visualization to choose one clustering

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time
- ullet  $\implies$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality
  - automated visualization to choose one clustering
  - many pairs of documents

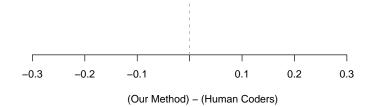
- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time
- $\implies$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality
  - automated visualization to choose one clustering
  - many pairs of documents
  - for coders: (1) unrelated, (2) loosely related, (3) closely related

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time
- ullet  $\implies$  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)

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

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

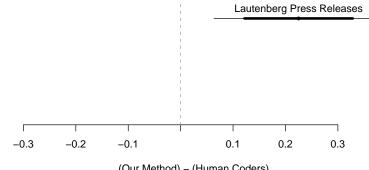
## Evaluation 1: Cluster Quality



2

(田) → (田) →

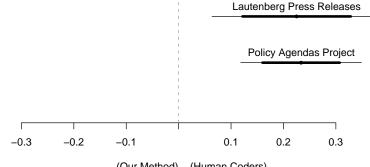
< C > < 🗗 > .



(Our Method) – (Human Coders)

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

Gary King (Harvard IQSS)

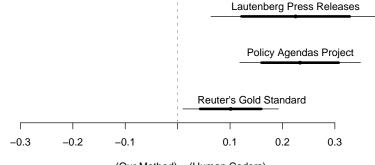


(Our Method) – (Human Coders)

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

Gary King (Harvard IQSS)

→ @ ト → 毛 ト → 石ト



(Our Method) – (Human Coders)

▲口▶ ▲圖▶ ▲国▶ ▲国▶

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

Gary King (Harvard IQSS)

Gary King (Harvard IQSS)

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

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

ヨト くまト

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:

高▶ 《高▶

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (biased against us)

.≣- ►

- 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)

- 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)

- 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

- 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
- $\bullet~$  User chooses  $\Rightarrow~$  only care about the one clustering that wins

- 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
- $\bullet~$  User chooses  $\Rightarrow~$  only care about the one clustering that wins
- Both cases a Condorcet winner:

- 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" :

 $\underline{\text{Our Method 1}} \rightarrow \text{vMF 1} \rightarrow \text{vMF 2} \rightarrow \underline{\text{Our Method 2}} \rightarrow \text{K-Means 1} \rightarrow \text{K-Means 2}$ 

・ロト ・聞 ト ・ヨト ・ ヨトー

- 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" :

```
\underline{\text{Our Method 1}} \rightarrow \text{vMF 1} \rightarrow \text{vMF 2} \rightarrow \underline{\text{Our Method 2}} \rightarrow \text{K-Means 1} \rightarrow \text{K-Means 2}
```

"Genetic testing":

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

æ

▲口▶ ▲圖▶ ▲国▶ ▲国▶

- David Mayhew's (1974) famous typology

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

- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming

- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking

- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking

#### - Data: 200 press releases from Frank Lautenberg's office (D-NJ)

ヨト くまト

- 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

ヨト くまト



/ 20

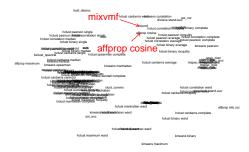
★白▶ ★圖▶ ★国▶ ★因▶ → 臣



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

イロト イポト イヨト イヨト

3



Red point: a clustering by Affinity Propagation-Cosine (Dueck and Frey 2007) Close to: Mixture of von Mises-Fisher distributions (Banerjee et. al.

イロト イポト イヨト イヨト

2005)



#### Space between methods:

▲口▶ ▲圖▶ ▲国▶ ▲国▶ -

Gary King (Harvard IQSS)

æ



#### Space between methods:

▲口▶ ▲圖▶ ▲国▶ ▲国▶ -

Gary King (Harvard IQSS)

/ 20

æ

<text><text><text><text><text><text>

holust maximum ward loneans binary

kmeans maximum

Space between methods: local cluster ensemble

・ロト ・団ト ・ヨト ・ヨト … ヨ



/ 20

★白▶ ★圖▶ ★国▶ ★因▶ → 臣



Found a region with particularly insightful clusterings

イロト イポト イヨト イヨト

Gary King (Harvard IQSS)

æ



#### Gary King (Harvard IQSS)

Mixture:

★白▶ ★御▶ ★恵▶ ★恵▶ 三臣



### Mixture:

#### 0.39 Hclust-Canberra-McQuitty

▲口▶ ▲圖▶ ▲国▶ ▲国▶ -



### Mixture:

0.39 Hclust-Canberra-McQuitty

・ロト ・同ト ・ヨト ・ヨト

0.30 Spectral clustering Random Walk (Metrics 1-6)



kmeans maximum

### Mixture:

- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward

イロト イポト イヨト イヨト



#### Mixture:

- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward

・ロト ・同ト ・ヨト ・ヨト

0.09 Hclust-Pearson-Ward



### Mixture:

- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward

・ロト ・同ト ・ヨト ・ヨト

- 0.09 Hclust-Pearson-Ward
- 0.05 Kmediods-Cosine



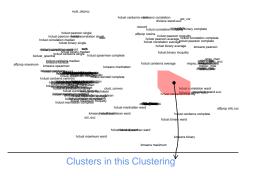
kmeans maximum

#### 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)

イロト イポト イヨト イヨト

э

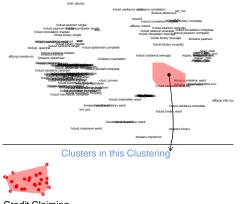


#### Mayhew

Gary King (Harvard IQSS)

2

・ロト ・聞 ト ・ヨト ・ヨトー



Credit Claiming Pork

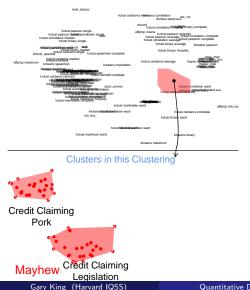
#### Credit Claiming, Pork:

"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"

・ロト ・団ト ・ヨト ・ヨト

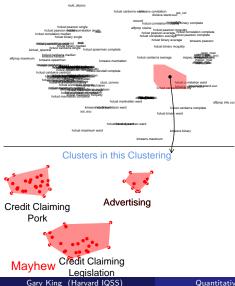
#### Mayhew

Gary King (Harvard IQSS)



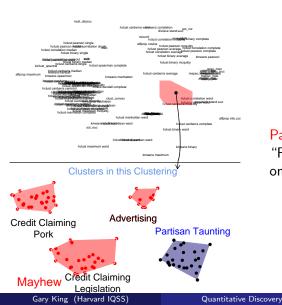
Credit Claiming, Legislation:

"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"



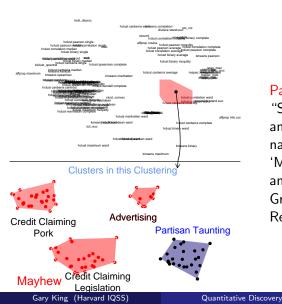
Advertising: "Senate Adopts Lautenberg/Menendez Resolution Honoring Spelling Bee Champion from New Jersey"

▲口▶ ▲圖▶ ▲国▶ ▲国▶



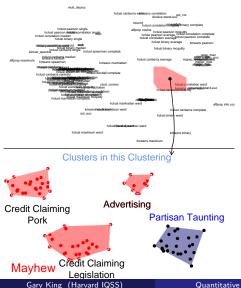
#### Partisan Taunting:

"Republicans Selling Out Nation on Chemical Plant Security"

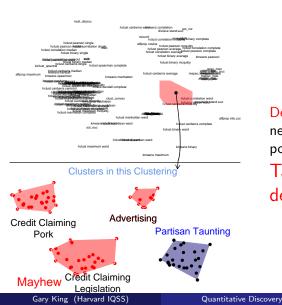


### Partisan Taunting:

"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' "



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



Definition: Explicit, public, and negative attacks on another political party or its members Taunting ruins deliberation

# Taunting ruins deliberation



Sen. Lautenberg on Senate Floor 4/29/04  "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]

ヨト イヨト

# 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]

# 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]

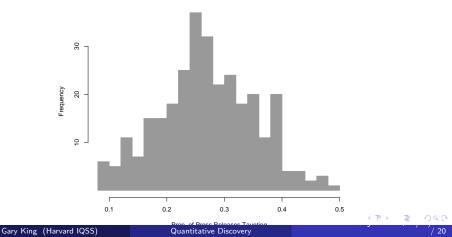
- Discovered using 200 press releases; 1 senator.

ヨト ★ ヨト

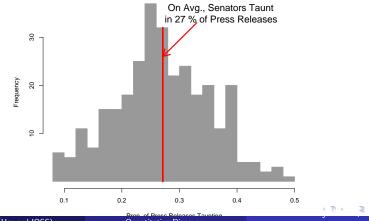
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

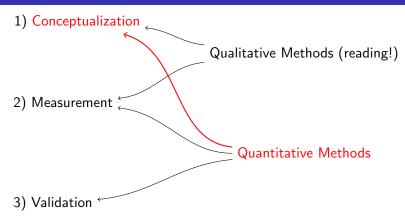
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party

- 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

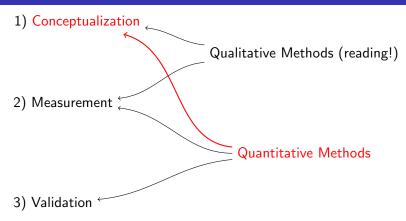


- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party



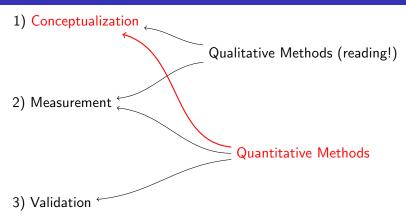


Quantitative methods for conceptualization: aiding discovery



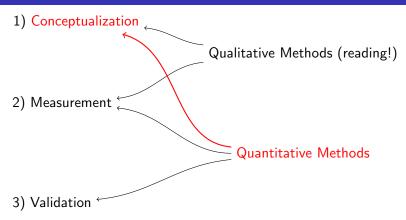
Quantitative methods for conceptualization: aiding discovery

- Few formal methods designed explicitly for conceptualization



Quantitative methods for conceptualization: aiding discovery

- Few formal methods designed explicitly for conceptualization
- Belittled: "Tom Swift and His Electric Factor Analysis Machine" (Armstrong 1967)



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 (on adding zooming out to the human ability to zoom in)

# http://GKing.Harvard.edu

伺下 くヨト くまト