Computer-Assisted Conceptualization

Gary King

Institute for Quantitative Social Science Harvard University

Talk at Harvard Graduate School of Arts and Sciences, Alumni Day, 4/2/2011

 $^{^{1}\}mathsf{Based}$ on joint work with Justin Grimmer (Harvard \rightsquigarrow Stanford)

 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).

- 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).
- Cluster Analysis: simultaneously (1) invents categories and (2) assigns documents to categories

- 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).
- Cluster Analysis: simultaneously (1) invents categories and (2) assigns documents to categories
- We focus on unstructured text; methods apply more broadly.

- 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).
- Cluster Analysis: simultaneously (1) invents categories and (2) assigns documents to categories
- We focus on unstructured text; methods apply more broadly.
- Main goal: Switch from Fully Automated to Computer Assisted

What's Hard about Clustering?

Gary King (Harvard IQSS)

æ

▶ ★ 문 ▶ ★ 문 ▶

Image: Image:

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

What's Hard about Clustering? (aka Why Johnny Can't Classify)

• Clustering seems easy; its not!

3 / 21

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

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

- 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)

- 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

- 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

- 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

- 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!
- Fully automated algorithms can help, but which ones?

(ロト 《聞 と 《臣 と 《臣 と 三臣 … のへで …

• The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- 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,...

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- 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

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- 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 (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- 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

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- 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

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- 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: full automation requires more information

- The (Impossible) Goal: optimal, fully automated, application-independent cluster analysis
- 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: full automation requires more information
- No surprise: everyone's tried cluster analysis; very few are satisfied

< ロ > < 団 > < 団 > < 団 > < 団 > < 団 > < 団 > < 団 > < 団 < O < O</p>

• Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies

- E > - E >

- Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: Computer-Assisted Clustering

- Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: Computer-Assisted Clustering
 - Easy in theory: list all clusterings; choose the best

- Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: Computer-Assisted Clustering
 - Easy in theory: list all clusterings; choose the best
 - Impossible in practice: Too hard for us mere humans!

- Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: Computer-Assisted Clustering
 - Easy in theory: list all clusterings; choose the best
 - Impossible in practice: Too hard for us mere humans!
 - An organized list will make the search possible

- Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: Computer-Assisted Clustering
 - Easy in theory: list all clusterings; choose the best
 - Impossible in practice: Too hard for us mere humans!
 - An organized list will make the search possible
 - Insight: Many clusterings are perceptually identical

- Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: Computer-Assisted Clustering
 - Easy in theory: list all clusterings; choose the best
 - Impossible in practice: Too hard for us mere humans!
 - An organized list will make the search possible
 - Insight: Many clusterings are perceptually identical
 - E.g.,: consider two clusterings that differ only because one document (of 10,000) moves from category 5 to 6

通 ト イヨ ト イヨト

- Fully Automated Clustering may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: Computer-Assisted Clustering
 - Easy in theory: list all clusterings; choose the best
 - Impossible in practice: Too hard for us mere humans!
 - An organized list will make the search possible
 - Insight: Many clusterings are perceptually identical
 - E.g.,: consider two clusterings that differ only because one document (of 10,000) moves from category 5 to 6
- Question: How to organize clusterings so humans can understand?

- 4 緑 ト - 4 三 ト - 4 三 ト

Set of clusterings

Set of clusterings \approx A list of unconnected addresses

wide at	SuperPages.com	195	Car C
and dentity	Cartage New England Inc	Carter F 24 Hilleck Res 02131	Carter Nella E 133 Maschets & Bos 02115
17 566-1282	26 Alles Ln Ipswich 01938	Faye & Ricky	133 Maschets & Bos 02115
	Cartagema Lydia 18 Jewett Ros 02131	357 Calumbus &r Bos 02116	115 Randolph Av Mil 02186
81 447-4101	13 Jewett Ros 02131 61/ 323-/639	Franklin & Anne	Nick 21 Fairfield Box 02116
10000-00	Cartagena Avith 9 Recent from 02119 617 442-9780	221 Mt Autorn Cen (2138	Nick & Debbi
00 257-9981		Fred 42 Haverland Jam 02130	146 Henrick Rd Newton 00459
17 566-1282	B Hy102136	Fred % Hinckley Rd Mil 02186	Nicole
17 364-5188	Jessica 50 Decatar Cha 02129	G & R # Verdun Der 00124	Norman G
1/ 304-5188	LUCINA 1/4 Harvard Can 02139	G T 27 Frankin by Sam 02145	38 Chickatzwbat Dor (0122
	M 25 Rowe Ras (0131	Gavie 25 Frantenac Dar 02134	P 64 Crestwood Pt Ray (2)31
361-0380	Carte Nicholas	Geo S 115 Most Hill Rd Jam 07130	P F 501 F Swith S Bas (0127
17 566-4548	Carte Nicholas 18 Assistes Barton 02116	George 125 Nashua Bos 02114617 367-9548	P L 44 Hutchings Rox 02121
17 300-4548	Cartegena 0 4 Millard Box 02114	Carter Halliday Associate	P R 91 Bynner Jan 02130
		107 5 Street Bos 02111	Paul & Constance
17 628-8248	Carten Thos J Sr & Claire	Carter Marry E	114 Anaryan Ar W Rax (2)32
	Thomas & Kathleen	26 Runne Brk Rd W Rex 02132 617 325-5465	Paul F 500 F Sub 97 S Res 02127
17 445-5116	50 Thompson Ln MI 02386	Carter Hide Co Inc	Paul M 27 Union Bri 02135
	Carter A Res (013)	146 Semmer Bos 02110	Carter Pile Driving Inc 17 Seaver Ct
17 822-2982	A Roberty	Carter Hilary 61 Harver Can 02140617 876-2750	Framingham (0/02 Wellesley TelNo-781 235-848
17 427-5712	A 31 Beflute Wy Rodury (2113	Horace	Carter Prudence
17 569-2698	A 31 Beflune Wy Rodory (0115	241 Walnut Av Roebery 02119	46 Frankin Watertown 02172
17 667-5190	A M 255 Maschits Ar Bes (2115 617 266-7153	Howard Jr 25 Note Dee Res 02119, 617 445-5552	Prodence
17 667-5190	A M 255 Mappings Ar Bos (2115 017 200*7153 Adams 381 Centre Stati (2184	J Can	44 Frankin Watertmen (0177
17 569-1417	Adams 31 Centre 52 Mil 02188 617 698-9074 Alice 108 Kimarneck Bei 02215 617 425-0193	J 15 Chatham Bro 02446	Reginald
	Alice 45 Market Cambridge 02139 617 945-2711	J 518 Harvard Bro 03445	106 Brumswick Dorchester 02121 617 541-284
17 338-9110	Andrew F #2 Vind & Som (014)	J 775 Why Plowy West Roebury 02132 617 323-5574	Rence & Andrew
17 825-9195	Carter Anne MD	Carter J Jacques MD	10 Walnut Bos 02108
11 852.4142	1100 Beacon Bro 00446	1 Brooking Pi Bro 02445	Carter Rice Dowd
17 296-1593	Carter Athens	Carter J M	Rulidev Dustree Publishing 163 Main Wilmington 01887
1/ 290-1593	772 Newbury Rodan 02116	100 Columbia Rd S Res 02227 617 464-1040	Toll Free-Dial '2' & Then
17 670-2078	B E 68 Gladeside Av Mat 02125	Carter J M Ornamental Ironworks	
17 623-9001	Carter Barbara L MD	CalPentroka Tellio-617 436-5353	Toll Free-Dial 12" & Then
11 052-2001		Carter J Veal Co	Cust Svo-Pvinting 613 Main Wilmington
17 296-4725	Cal	48 Newstartet 50 80x 07118	Toll Free-Dial '2' & Then
11 540.4152	Carter Becky 8xs (21)4	Carter James	Headquarters 613 Main Wilmington 00887 Call
17 542-1521	Bernard J	1573 Cambridge St Cam 02138	Installs Cronie 163 Main Wilmington (11887
17 245.1251	112 Gladatore E Bes 02128	James 182 Fisher Av Roebury 02120617 739-2193	Toll Free-Oial '2' & Then
17 364-5232	Bithiah 25 Median Der (0124 617 298-8713	James	Carter Richard
17 541-5649	Blake 25 Mt Vernos Bes 02108	37 Gold Star Rd Cambridge 02140 617 876-8841	M79 Connectin by Relation 02215 617 987-083
TL 247, 204A		Jas L 14 Roseberry Rd Mat 02126 617 361-0773	Richard A \$7 Mt Venos Bos (02108617 566-729
17 739-2662	20 Park Ptr Bes 00116	Jane 114 Adena Rd Newton 02465 617 964-0435	
1. 137 2002	Carter & Burgess Consultants Inc	Jeffrey 41 Warren & Bos (2116 617 426-5994	170 Committh Av Bos (221)6
17 879-0030	21 Last 52 Cam 02240	John 11 Manafield Bri 02134	Carter Richard K
17 541-3948	Carter C 2000 Commeth Ar Bri 02135 617 782-2118	John 327 Summer Box 02211	15 Merrer S Ros (27177
17 436-1513	C 228 Farmood Av East Boston 02228617 569-1545	John 40 Westwind Rd Der 02125 617 282-1235	Robert 1 125 Subdale & Cam 02140 617 864-153
17 569-4119		June O 129 A Summit Av Bri 02135 617 734-6109	Roger 150 St Botelin Bos 02115
17 309-4119 Ma 82228	C 430 West Mill Mer 07126 617 296-6392	K 38 Browning Av Dorchester 02124 617 265-8456	Row of Concord by Cam (2)138
00 569-8782	C & M 43 Burroughs Jan 02230 617 524-9558	K 17 Farmond Dorchester (22121	Royce 18 Seminary Cha (2229

195

Set of clusterings \approx A list of unconnected addresses

wide at SuperPages.com Cartage New England Inc Carter F 24 Hillock Res 02133 617 32 17 566-1282 978 356-9960 Faye & Ricky ma Lydia 357 Columbus Av Bos 02116. 617 43 617 323-7639 Francis S 134 Temple W Rox 02132. 617 32 \$1 447-4101 18 Jewett Ros 02131 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 socilla 134 strengt Con (2223) 617 576-106 617 82 361-0380 Melvin Stü Green Cam 0223 617 695-6996 17 566-4548 18 Appieton Baston 0213 98 125 Nashua Bos 02114 ena O 4 Millard Bos 02138 Thos J Sr & Claire 617 338-8219 **Carter Halliday Associate** 617 49 17 628-8248 617 698-6163 Carter Harry F 617 35 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) 617 442-5230 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 Dime Rox 0013 617 698-907 617 35 dams 341 Centre S2 MI 0218 17 569-1417 Alice 108 Kilmarnock Bos 02215 617 425-0193 lice 45 Market Cambridge 0213 617 945-2711 617 625-7623 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 4 617 536-6329 mbia Rd S Bos 02227 272 Newbury Boston 0211d Carter J M Ornamental Ironworks 17 670-2078 Pembroka Tellio-617 43 Carter Barbara L MD Carter J Veal Co afts-New England Medical Center Bo 617 626-0051 17 296-4725 Carter Becky 8xs 82 617 523-4368 1573 Cambridge St Cam 02138.......617 4 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 517 739-2662 Jane 114 Adena Rd Newton 02465. Carter & Burness Consultants leffrey 41 Warren &r Bos (2116 617 225-0200 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 7 517 569-4119 C & M 43 Burroughs Jam 02230. 100 569-8782

133	Uai		U	
7-1105	Carter Nella E	1. 19. 16. 10	President	
	222 Maschety &r Ros 02115		67-6483	
7-7331	Nicholas S F	1174		
3.0191	115 Randolph Av Mil 02186	617.2	67-5222	
4-0798	Nick & Dehhi			
4-3078	196 Herrick Rd Newton 00459	617 5	27-0490	
8-1343	Nicole		98-0/13	
3-7121	38 Chickstawbet Dor 02122	617.8	22-1203	
5-0322				
2-3215	P E 501 E Soth S Bos 02127	617 2	68-4213	ļ
7-9548	P L 44 Hutchings Rox (2121 P R 91 Bynner Jan (2130		27-9170	
6-1689	P K 91 Bynner Jan 02130		03-0092	
0.1001	114 Annuan & W Boy 20132	617 3	25-2036	
5-5465	Paul F strip Sate St S Res (22)27.	617 2	68-4546	
1000	Paul M 27 Usion Bri 02135		87-2115	
2-7987	Carter Pile Driving Inc 17 Beaver Framingham (12/02	CT	25.0400	
0-2/30				
2-5307	46 Frankin Watertown 02172		93-3782	
5-5552	Prudence		1000	
4-2688	46 Franklin Watertown 02172 Reginald	617 9	26-7063	
2-7990	106 Brumswick Dorchester (212).	617 5	41-2843	
3-5574				
2.5	10 Walnut Bos 02108	617 7	20-3765	
5-8787	Carter Rice Dowd Builder Duttee Publishing 163 Main Wi	122	140000	
4-1040			38-1671	
4-1040				
6-5353	Toll Free-Dial '1' & Then		19-7447	
	Cust Svo-Pywbeg 613 Main Weinington	000.6	49-7447	
2-1775	Cust Svc Integral Prod 0.3 Main Will Toll Free-Dial '3' & Then. Cust Svc Privileg 0.3 Main Wilmington Toll Free-Dial '2' & Then. Headquarters 0.3 Main Wilmington 03	107		
2-1214			88-7447	
9-2193	Ingails Cronie 163 Main Wilmington 01 Toll Free-Okal '2' & Then	800 6	38-1673	
	Carter Richard			
6-8841	3879 Commette Av Brighton 0222	5_617 9	67-0836	
1-0773	Richard A \$7 Mt Vernos Bos 0210	s617 5	66-7293	
6-5994	Carter Richard A MD 170 Commette Av Ros 02116	617.2	67-0710	
7-2163				
3-4334	15 Meerer S Ros 02327		68-0448	
2-1235				
4-6109	Roger 150 St Botelph Bos 02115 Roy 44 Cencord Av Cam 02138	617 4	24-6148	
2-1593	Roy 64 Cancord Av Cam 02138 Royce 18 Seminary Cha 02129	617 2	41-0418	
4.0130	Ruyce is semiary the utility	(377	4113	

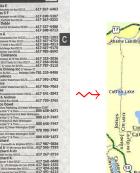
Car

C



Set of clusterings \approx A list of unconnected addresses

wide at SuperPages.com 195				
NO. ID. SHITT	Cartage New England Inc	Carter F 24 Hillock Res 02131	Carter Ne	
17 566-1282	26 Allen Ln Ipswich 01938	Fave & Ricky	333 N	
		357 Columbus Ar Bos 02116	Nichola	
81 447-4101	13 Jewett Ros 02131	Francis S 134 Temple W Rox (2132., 617 323-6781	115 8	
		Franklin & Anne	Nick 21	
00 257-9981	9 Rancraft Ros (2019	221 Mt Auborn Cam (0138	Nick &	
	B thd 02136	Fred 42 Haverland Jam 02130	2961	
17 566-1282	Jessica 50 Deceter Cha 02729	Fred 96 Hinckley Rd Mil 02186	Nicole.	
17 364-5188	Lucilla 174 Harvard Can (2139 617 491-5621	G & R # Verdan Der 00124	Norma	
	M 25 Rowe Res (0213)	G T 27 Franklin Av Som 02145	38 0	
361-0380	Melvin 500 Green Cam 02139	Gavle 25 Frontenar Dor 07134	P 54 Ore	
301-0300	Carte Nicholas	Geo S 115 Most Hill Rd Jam 00130 617 522-3215	PE 501	
17 566-4548	18 Appleton Baston 00116	George 125 Nashua Bos 02114	PL 441	
11 200 4240	Cartegena Q 4 Millard Box 02118	Carter Halliday Associate	PRMF	
17 628-8248	Carten Thos J Sr & Claire	107 5 Street Bes 02111	Paul &	
11 079-9549	1 Paradise Rd Mil (0186	Carter Harry F	114.0	
17 445-5116	Thomas & Kathleen	26 Runno Brk Rd W Rox 02132 617 325-5465	Paul F	
11 442.2110	50 Thompson Ln Mil 02386	Carter Hide Co Inc	Paul M	
17 822-2982	Carter A Ras (013)	146 Summer Bos 02110	Carter Pil	
17 427-5712	A Robery	Carter Hilary 61 Harver Can (2)40617 876-2750	Transrebo	
17 569-2698	A 31 Bethune Wy Rosbury (2113	Horace	Carter Pr	
11 203.5038	A 260 Putnam Ar Cambridge 02139 617 492-4174	241 Walnut Av Roebery (0119	46 Fr	
	A 260 Patham Ar Cambridge 02139 017 492-4174 A M 255 Maschets Ar Bos 02115 617 266-7153	Howard Jr 26 Note Die Res 02119, 617 445-5552	Prodec	
17 667-5190	A M 255 Maschels & Bes (2115	J Cam	40 Fr	
	Adams 31 Centre 5 Mil 0218 617 698-9074 Alice 108 Kilmarneck Bes 02215 617 425-0193	J 15 Dathan Bro 02446	Regina	
17 569-1417	Allice 108 Kilmarrock Ses (02215 617 425-0193	J 538 Harvard Bro 02445	1061	
ity Dr	Alice 45 Market Cambridge 02139 617 945-2711	J 775 Whe Pleny West Rosbury 02132 617 323-5574	Rence	
17 338-9110	Andrew F #2 Vinal Ar Som 02143 617 625-7623	Carter J Jacoues MD	10 %	
17 825-9195	Carter Anne MD	1 Brookine Pi Bro 02446	Carter Ri	
	1104 Beacon Bro 00446	Carter J M	Builder Dr	
17 296-1593	Carter Athens		Toll Fre	
	272 Newbury Boston 02116	1410 Columbia Rd S Bos 02227 617 464-1040	CHI Sect	
17 670-2078	B E 68 Gladeside Av Mat 02135	Carter J M Ornamental Ironworks	Toll Fre	
17 623-9001	Carter Barbara L MD	CalPembroka Tellio-617 436-5353	Orst Serve	
	Tufts-New England Medical Center Bos 02111	Carter J Veal Co	Toll Fre	
17 296-4725	Call	48 Newmarket 5q Rox 02138	Headourt	
	Carter Becky 8xs 82114	Carter James	Collere	
17 542-1521	Bernard J	1573 Cambridge St Cam 02138617 492-1214	Inquits On	
	112 Gladstone E Bes 02128	James 182 Fisher Av Roebury 02120617 739-2193	Toll Fre	
17 364-5232	Bithiah 25 Medway Der 02124	James	Carter Ri	
17 541-5649	Blake 35 Mt Vernos Bos 02108	37 Gold Star Rd Cambridge 02140 617 876-8841	1075	
	Carter Broadcasting Co	Jas L 14 Reseberry Rd Mat 02126617 361-0773	Richar	
17 739-2662	20 Park Ptr Bes 02116	Jane 114 Adena Rd Newton 02465 617 964-0435	Carter Rie	
	Carter & Burgess Consultants Inc	Jeffrey 41 Warren Ar Bos 02116 617 426-5994	170 Comm	
17 879-0030	23 East St Cam 023-0	John 11 Manafield Bri 02134	Carter Rie	
17 541-3948	Carter C 2000 Commeth Ar Bri (2135 617 782-2118	John 327 Summer Bos 02218	15 M	
17 416-1513	C 228 Farwood Av East Beston 02228617 569-1545	John 40 Westwind Rd Der 02125 617 282-1235	Robert	
17 569-4119	C 357 Harvard Cars 02138	June O 329 A Summit Av Bri 02135 617 734-6109	Roger	
itva 02128	C 630 Walk Hill Mat 07126	K 38 Browning Av Dorchester 02124 617 265-8456	Roy 44	
300 569-8782	C & M 43 Berroube Jan 02230 617 524-9558	K 17 Ermont Deerbaster (22)21 617 282-1593	Royce	
		APTA APT 711		



C

Car



(日) (周) (三) (三)

→ We develop a (conceptual) geography of clusterings

A New Strategy

Make it easy to choose best clustering from millions of choices

3

メロト メポト メヨト メヨト

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

Image: Image:

- 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)
- Oevelop 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)
- 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

イロト イ理ト イヨト 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)
- 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)

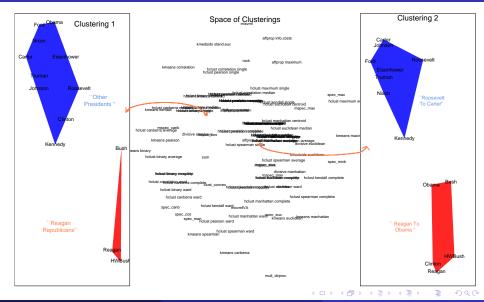
- Code text as numbers (in one *or more* of several ways)
- Apply all clustering methods we can find to the data each representing different (unstated) substantive assumptions (<15 mins)</p>
- (Too much for a person to understand, but organization will help)
- Oevelop an application-independent distance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

- 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
- Or, our new strategy: represent the entire bell space directly; no need to examine document contents)

イロト イ理ト イヨト イヨト

Many Thousands of Clusterings, Sorted & Organized

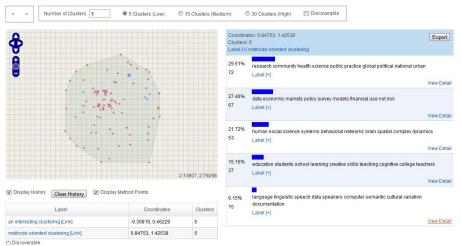
You choose one (or more), based on insight, discovery, useful information,...



Gary King (Harvard IQSS)

Software Screenshot

Size: 244 Files Description: NSF - Updated Set



æ

(日) (周) (三) (三)

▲ロ ▶ ▲圖 ▶ ▲ 臣 ▶ ▲ 臣 ▶ ▲ 臣 → りんで !

• Metric based on 3 assumptions

3 K K 3 K

• 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

▲ロ ▶ ▲ 聞 ▶ ▲ 臣 ▶ ▲ 臣 ▶ ● ○ ○ ○ ○

• Goals:

▲口と▲聞と▲臣と▲臣と 臣 のへの

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

Image: Image:

- - E + - E +

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

- E > - E >

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

- 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

B ▶ < B ▶

- 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

- 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

- ・ロト・(個)ト・(目)ト・目、 の(

Gary King (Harvard IQSS)

12 / 21

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

∃ ▶ ∢ ∃ ▶

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

∃ ► < ∃ ►</p>

- 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

- 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
- 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

- 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

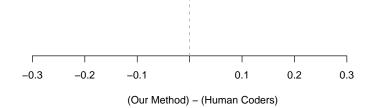
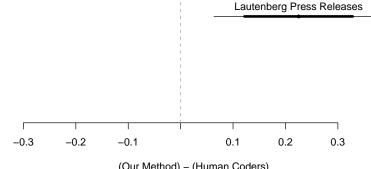


Image: Image:

æ

-∢∃>

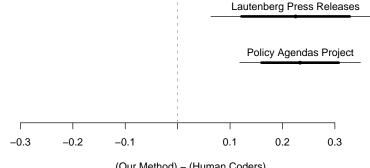
Evaluation 1: Cluster Quality



(Our Method) - (Human Coders)

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

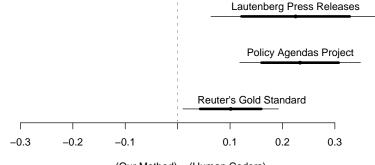
- ∢ ∃ →



(Our Method) – (Human Coders)

Image: Image:

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



(Our Method) – (Human Coders)

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

< ロ > < 同 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ >

Gary King (Harvard IQSS)

14 / 21

æ

- ∢ 臣 ► ∢ 臣 ►

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

14 / 21

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

14 / 21

- 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
- 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
- ${\scriptstyle \bullet}$ 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" :

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

ヘロト 人間 とくほ とくほ とう

- 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":

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

▲□▶ ▲圖▶ ▲필▶ ▲필▶ - ヨー わえぐ!

Gary King (Harvard IQSS)

- David Mayhew's (1974) famous typology

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

∃ ▶ ∢ ∃ ▶

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

B ▶ < B ▶

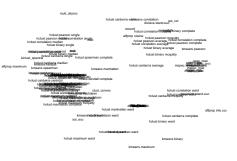
- David Mayhew's (1974) famous typology
 - Advertising
 - Credit Claiming
 - Position Taking

B ▶ < B ▶

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

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

- 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



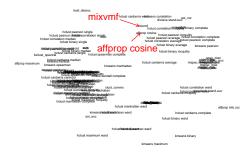
▲ロト ▲圖ト ▲画ト ▲画ト 三回 - 釣ん(で)



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

< ロ > < 同 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ >

э



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

distributions (Banerjee et. al. 2005)

イロト イポト イヨト イヨト

э



Space between methods:

イロト イ理ト イヨト イヨト

Gary King (Harvard IQSS)

æ



Space between methods:

イロト イ理ト イヨト イヨト

Gary King (Harvard IQSS)

16 / 21

æ

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

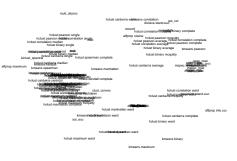
holust maximum ward kimeans binary

kmeans maximum

Space between methods: local cluster ensemble

イロト イ理ト イヨト イヨト

2



▲ロト ▲圖ト ▲画ト ▲画ト 三回 - 釣ん(で)



Found a region with particularly insightful clusterings

イロト イ理ト イヨト イヨトー

Gary King (Harvard IQSS)

æ



Mixture:

Gary King (Harvard IQSS)

3

メロト メロト メヨト メヨト



Mixture:

0.39 Hclust-Canberra-McQuitty

・ロト ・聞 ト ・ヨト ・ヨト ・ヨー りへの



Mixture:

0.39 Hclust-Canberra-McQuitty

イロト イヨト イヨト イヨト

0.30 Spectral clustering Random Walk (Metrics 1-6)

æ



kmeans maximum

Mixture:

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

< ロ > < 同 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ >

3



Mixture:

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

(日) (周) (三) (三)

0.09 Hclust-Pearson-Ward

3



Mixture:

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

(日) (周) (三) (三)

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

э



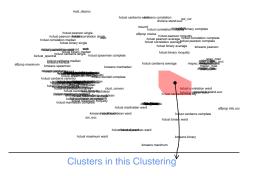
kmeans maximum

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

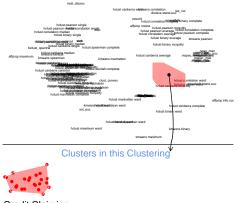
э



Mayhew

Gary King (Harvard IQSS)

▲□▶ ▲圖▶ ▲ 圖▶ ▲ 圖▶ ▲ 圖 - めるの



Credit Claiming Pork

Credit Claiming, Pork:

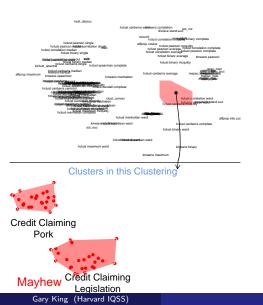
"Sens. Frank R. Lautenberg (D-NJ) and Robert Menendez (D-NJ) announced that the U.S. Department of Commerce has awarded a \$100,000 grant to the South Jersey Economic Development District"

イロト イポト イヨト イヨト

Mayhew

Gary King (Harvard IQSS)

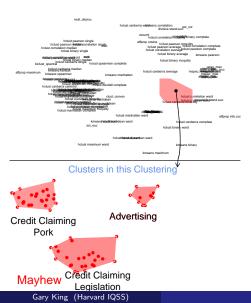
э



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"

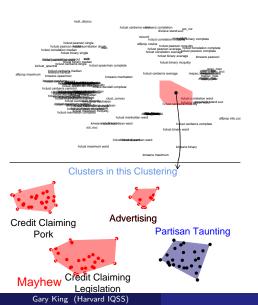
イロト イポト イヨト イヨト



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

イロト イポト イヨト イヨト

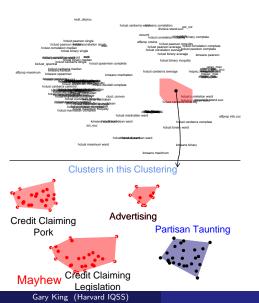
э



Partisan Taunting:

"Republicans Selling Out Nation on Chemical Plant Security"

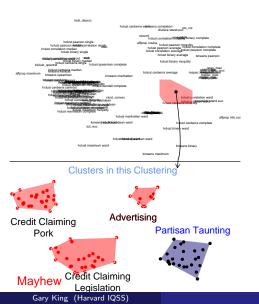
イロト イポト イヨト イヨト



Partisan Taunting:

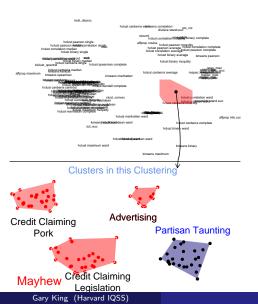
"Senator Lautenberg's amendment would change the name of...the Republican bill...to 'More Tax Breaks for the Rich and More Debt for Our Grandchildren Deficit Expansion Reconciliation Act of 2006"'

(日) (同) (三) (三)



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

(日) (同) (三) (三)



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

(日) (同) (三) (三)

Taunting ruins deliberation



Sen. Lautenberg on Senate Floor 4/29/04

 "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]

Image: Image:

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]

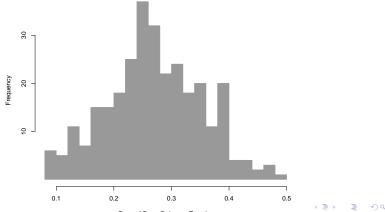
イロト イ押ト イヨト イヨト

- Discovered using 200 press releases; 1 senator.

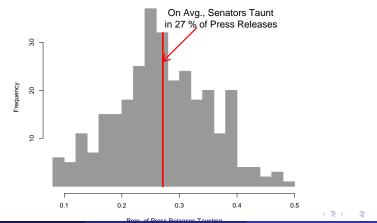
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

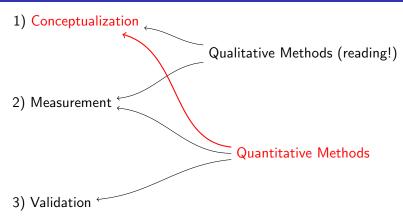
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party

- 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

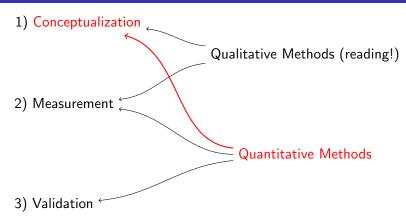


- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party



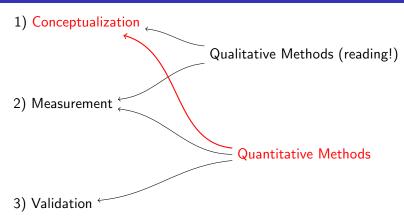


Computer-Assisted Methods for conceptualization and discovery



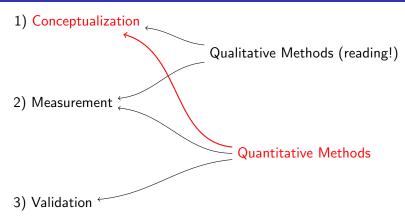
Computer-Assisted Methods for conceptualization and discovery

- Methods designed explicitly for conceptualization



Computer-Assisted Methods for conceptualization and discovery

- Methods designed explicitly for conceptualization
- The end of quantitative v qualitative debates



Computer-Assisted Methods for conceptualization and discovery

- Methods designed explicitly for conceptualization
- The end of quantitative v qualitative debates
- Evaluation methods measure progress in discovery

For more information



http://GKing.Harvard.edu

A B M A B M