

Computer-Assisted Conceptualization

Gary King

Institute for Quantitative Social Science
Harvard University

Talk at Harvard Law School, 11/17/2011

¹Based on joint work with Justin Grimmer (Harvard ↔ Stanford)

A Method for Computer Assisted Conceptualization

- **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).

A Method for Computer Assisted Conceptualization

- **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

A Method for Computer Assisted Conceptualization

- **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
- Focus on **unstructured text**; methods apply more broadly.

What's Hard about Clustering?

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = number of ways of partitioning n objects

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = number of ways of partitioning n objects
- $Bell(2) = 2$ (AB, A B)

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = 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)

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = 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)
- $Bell(5) = 52$

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = 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)
- $Bell(5) = 52$
- $Bell(100) \approx$

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = 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)
- $Bell(5) = 52$
- $Bell(100) \approx (\text{Number of elementary particles in the universe}) \times 10^{28}$

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = 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)
- $Bell(5) = 52$
- $Bell(100) \approx (\text{Number of elementary particles in the universe}) \times 10^{28}$
- Now imagine choosing the *optimal* classification scheme by hand!

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Choosing the best conceptualization (i.e., clustering) is hard!
- $Bell(n)$ = 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)
- $Bell(5) = 52$
- $Bell(100) \approx$ (Number of elementary particles in the universe) $\times 10^{28}$
- Now imagine choosing the *optimal* classification scheme by hand!
- Fully automated algorithms can help, but which ones?

From Fully Automated to Computer Assisted Clustering

From Fully Automated to Computer Assisted Clustering

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

From Fully Automated to Computer Assisted Clustering

- **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

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well
- Our alternative approach: **Computer-Assisted Clustering**

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well
- Our alternative approach: **Computer-Assisted Clustering**
 - **Easy in theory:** list all clusterings; choose the “best”

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well
- Our alternative approach: **Computer-Assisted Clustering**
 - **Easy in theory:** list all clusterings; choose the “best”
 - **Impossible in practice:** Too hard for us mere humans!

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well
- Our alternative approach: **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

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well
- Our alternative approach: **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

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well
- Our alternative approach: **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

From Fully Automated to Computer Assisted Clustering

- **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
- **Deep problem:** full automation requires more substance
- No surprise: cluster analysis rarely works well
- Our alternative approach: **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?

Our Idea: Meaning Through Geography

Set of clusterings

Our Idea: Meaning Through Geography

Set of clusterings ≈

A list of unconnected addresses

wide at SuperPages.com

	195	Car	C
Cartage New England Inc 28 Allen Ln Ipswich 01938..... 978 356-9960	Carter F 34 Hibiscus Bldg 02133..... 617 327-1105	Carter Nellie E 323 Montclair Ave Box 02115..... 617 267-6483	
Cartagena Lydia 28 Sweet Box 02131..... 617 323-7639	Faye & Ricky 207 Columbia Ave Box 02136..... 617 437-7331	Nicholas S F 115 Randolph Ave Box 02186..... 617 698-6307	
Cartagena Avish F Pleasant Box 02139..... 617 442-9780	Francis S 134 Yankov W Ave 02132..... 617 323-6781	Nick 21 Farwell Box 02114..... 617 267-5222	
B Hrd 02134..... 617 361-5253	Franklin & Anne 705 Mt Auburn Cam 02138..... 617 354-0798	Nick & Debbi 196 Herold Rd Newton 02459..... 617 527-0480	
17 566-1282 Jessica 50 Decatur Cha 02129..... 617 241-0152	Fred 40 Hawthorn Elm 02138..... 617 524-3078	Nicole..... 617 698-0713	
17 364-5188 Lucille 124 Harvard Cam 02136..... 617 491-5621	Fred 76 Rowley Rd Mt 02136..... 617 698-1343	Norman G 38 Chickawhoh Dr 02125..... 617 822-1201	
361-0380 Mahn 503 Green Cam 02139..... 617 576-1061	G & B 8 Vardon Dcr 02134..... 617 434-8906	P 40 Cranford Pl Box 02135..... 617 437-4754	
17 566-4548 Certe Nicholas..... 617 576-1061	G T 27 Franklin Ave Sun 02145..... 617 623-7121	P E 501 E South S Box 02137..... 617 268-8213	
17 628-8248 Carlton 204 4th Row 02134..... 617 695-6996	Gayle 25 Franklin Dcr 02134..... 617 823-0322	P E 14 Hutchings Box 02131..... 617 427-9170	
17 445-5116 Thomas & Kathleen..... 617 698-6163	George 1255 Mass Hill Rd 02138..... 617 522-3215	P E 81 Boyer Jan 02138..... 617 968-8692	
17 822-2962 Carter A Box 02133..... 617 229-2257	George 250 Madison Box 02134..... 617 367-9548	Paul & Constance 114 Freeman St W Box 02131..... 617 325-2036	
17 427-5712 A Weber..... 617 442-5230	Carter Hillside Assocn 107 S Street Box 02111..... 617 456-1689	Paul E 501 E South S S Box 02137..... 617 268-4546	
17 569-2698 A 201 Rowley Av Cambridge 02238..... 617 492-4174	Carter Hide Co Inc 26 Burm In Rd W Ave 02132..... 617 325-5465	Paul M 27 Union Br 02139..... 617 787-2115	
17 667-5190 Adams 361 Centre St Mt 02136..... 617 698-7074	Carter Hilary 41 Harvey Cam 02148..... 617 876-2750	Prangman 02102..... Wobley Tpk 781.235-0488	
17 338-9117 Alice 108 Elmwood Box 02134..... 617 423-0193	Horace 361 Walnut Av Rosbury 02139..... 617 442-5307	Carter Prudence 40 Franklin Waterlton 02127..... 617 393-3782	
17 825-9119 Carter Anne MD..... 617 739-1022	Howard Jr 28 Nctra Dco Box 02118..... 617 445-5552	Prudence 40 Franklin Waterlton 02127..... 617 926-7063	
17 296-1593 Carter J M..... 617 739-1022	J Dan..... 617 354-2658	Roginald 106 Bromwich Dorchester 02122..... 617 541-2843	
17 670-2078 B E 18 Graduate Av Mt 02136..... 617 296-6911	J 21 Chatham Box 02446..... 617 233-7990	Renee & Andrew 30 Walnut Box 02128..... 617 720-3765	
17 621-9001 Carter Barbara L MD..... 617 636-0051	J 538 Harvard Box 02446..... 617 730-9483	Carter Rice Dorel Bulfinch Denton Publishing 163 Main Wilmington 01887..... 800 638-1671	
17 296-4725 Carter Becky Box 02134..... 617 523-4368	J 775 The Pines West Rosbury 02135..... 617 323-5374	Ted Free-Dal '2' & Thom..... 800 616-7447	
17 542-1521 Bernard J..... 617 567-9430	J Freestone Pl Box 02146..... 617 735-8787	Ted Free-Dal '3' & Thom..... 800 648-7447	
17 364-5232 Bibbith 25 Midway Dcr 02136..... 617 298-8713	3410 Columbia Rd S Box 02136..... 617 464-1040	Cal..... 978 988-7447	
17 541-5249 Carter Broadcasting Co..... 617 367-9931	Carter J M Ornamental Ironworks 260 Walnut Falls 02136..... 617 876-5353	Ingala Stone 163 Main Wilmington 01887..... 800 638-1673	
17 739-2662 Carter & Business Consultants Inc..... 617 423-0210	Carter J Neal Co 40 Hawthorn Elm 02138..... 617 442-1775	Cal..... 978 988-7447	
17 879-0030 Carter C 200 Concord Av Mt 02135..... 617 782-2118	Carter James 157 Cambridge St Cam 02138..... 617 492-1214	Richard A M 207 Concord Av Brighton 02215..... 617 982-0836	
17 436-1511 C 218 Harvard Av East Boston 02128..... 617 569-1545	James 312 Foster Av Rosbury 02138..... 617 739-2193	Richard A MD 47 Mt Vernon Box 02106..... 617 566-7293	
17 569-4119 C 109 Harvard Cam 02136..... 617 491-8522	Janet 14 Adams Rd Newton 02465..... 617 964-0435	Richard R K 130 Concord Av Mt 02136..... 617 267-0710	
800 569-8782 C & M 43 Bernham Jan 02138..... 617 524-9558	Jeffrey 41 Morris Av Mt 02134..... 617 426-5994	Richard R K 23 Mallett S Box 02137..... 617 268-0448	
	John 107 Summer Box 02128..... 617 423-4134	Roger 130 St Brnagh Box 02131..... 617 424-6148	
	John 40 Hawthorn Elm 02138..... 617 282-1235	Roy 41 Concord Cam 02138..... 617 491-6115	
	June O 129 A Summit Av Br 02133..... 617 734-6109	Royce 18 Sanyday Cha 02129..... 617 241-0418	
	J 28 Hawthorn Elm 02138..... 617 265-8456		
	K 17 Concord Dorchester 02127..... 617 282-1593		

Our Idea: Meaning Through Geography

Set of clusterings \approx

A list of unconnected addresses

wide at SuperPages.com

195

Car

C

17 566-1282	Cartage New England Inc 28 Allen Ln Ipswich 01938	978 356-9960	Carter F. 514 Hickox Ave 02131	617 327-1105	Carter Nella E 323 Marchant Ave Box 02115	617 267-6483	
17 447-4101	Cartagena Lydia 28 Sweet Briar 02131	617 323-7639	Faye & Ricky 20 Columbia Ave Box 02136	617 437-7331	Nicholas S F 115 Randolph Ave 02136	617 698-5307	
100 257-9961	Cartagena Avish F Beach Rd 02139	617 442-9780	Francis S. 134 Temple W Ave 02132	617 323-6781	Nick & Debbie 215 Fyfield Ave 02114	617 267-5222	
17 566-1282	B Had 02136	617 361-5253	Franklin & Anne 705 Mt Auburn Cam 02138	617 354-0798	Norman G 196 Hermit Rd Newton 02459	617 527-0480	
17 364-5188	Justica 50 Decatur Cha 02129	617 241-0152	Fred 40 Haverhill Aven 02136	617 524-3078	Nick & Constance 38 Chickadee Rd 02125	617 822-1203	
361-0380	Luzella 124 Harvard Cam 02139	617 491-5621	Fred W. Hovell Ave 02136	617 698-1343	P E 501 E South S Ave 02137	617 268-4213	
17 566-4548	M 90 Howe St 02139	617 323-9713	G & B. 8 Vardon Ave 02134	617 436-8906	P L. 44 Hutchings Box 02131	617 427-9170	
17 628-8248	Melvin 503 Green Cam 02139	617 576-1061	Gayle 25 Franklin St 02134	617 823-0322	P R 91 Brewer Ave 02138	617 968-8692	
17 822-2962	Carte Nicholas 18 Appleton Boston 02114	617 695-6996	Geo S. 115 Mount Hill Rd 02138	617 522-3215	Paul & Constance 114 Freeman St W 02131	617 325-3034	
17 427-5712	Cartagena O. 4 Bradford Box 02138	617 338-0219	George 120 Naveson Ave 02131	617 367-9548	Paul M. 501 E South S Ave 02137	617 268-4546	
17 569-2698	Carten Thos J Sr & Claire 1 Fyfield St Mt 02136	617 698-6163	Carter Holiday Assoc 107 S Street Box 02111	617 456-1689	Paul M. 27 Union St 02139	617 787-2115	
17 667-5190	Carte Thos & Kathleen 50 Thompson Ln Mt 02136	617 696-6919	Carter Harry F 26 Burns Rd Rt W Ave 02132	617 325-5465	Prudence 40 Franklin Waterfront 02172	617 393-3782	
17 569-1417	Carte A. 100 A 200 Pioneer Av Cambridge 02142	617 492-4174	Carter Hide Co Inc 117 542-7987	Horace 301 Walnut Av Rosbury 02139	617 626-4546	Prudence 40 Franklin Waterfront 02172	617 393-3782
17 825-9195	Carte Anne MD 1101 Beacon Ave 02144	617 739-1022	Carter Hilary 41 Harvey Cam 02148	617 876-2750	Roginald 100 Brookside Center 02124	617 541-2843	
17 670-2078	Carte B. E. 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 600 Brookside Center 02124	617 541-2843	
17 621-9001	Carte Barbara L MD 8 E. 100 Graduate Av Mt 02136	617 296-6911	Carter J M Ornamental Ironworks Pondside Falls 017 436-5353	Howard Jr 301 New One Box 02118	617 445-5552	Renee & Andrew 600 Brookside Center 02124	617 541-2843
17 296-4725	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J Neal Co 40 Newbury St 02138	617 442-1775	J. Dan 15 Chatham Ave 02144	617 232-7990	
17 542-1521	Carte B. E. 8 E. 100 Graduate Av Mt 02136	617 296-6911	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
17 364-5232	Carte B. E. 8 E. 100 Graduate Av Mt 02136	617 296-6911	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
17 541-5649	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
17 739-2662	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
17 879-0030	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
17 541-3948	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
17 436-1511	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
17 569-4119	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
100 257-9961	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	
100 257-9961	Carte Bernad J 371 Newbury Boston 02116	617 536-6239	Carter J. 300 3410 Columbia Rd S Box 02138	617 464-1040	J. S. 438 438 Harvard Ave 02138	617 730-9483	



Our Idea: Meaning Through Geography

Set of clusterings \approx

A list of unconnected addresses

wide at SuperPages.com

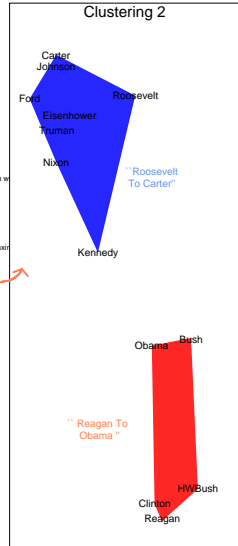
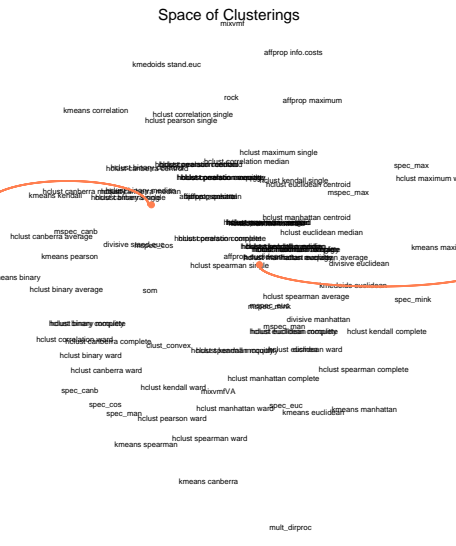
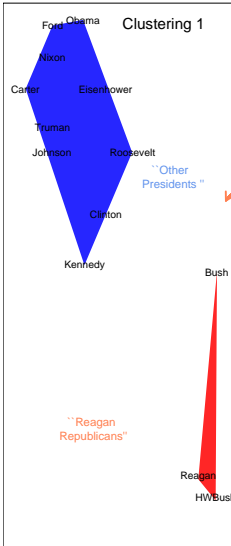
	195	Car	C
17 566-1282	Cartage New England Inc 20 Allen Ln Ipswich 01938	978 356-9960	
17 447-4101	Cartagena Lydia 28 Sweet Briar 02331	617 323-7639	
90 257-9961	Cartagena Avish F Beach Rd 02319	617 442-9780	
17 566-1282	B Had 02336	617 361-5253	
17 364-5188	Lucille 124 Harvard Can 02139	617 491-5621	
361-0380	M 95 Howe Box 02336	617 323-9713	
17 566-4548	Melvin 503 Green Can 02139	617 576-1061	
17 628-8248	Carte Nicholas 18 Appleton Boston 02134	617 695-6996	
17 445-5116	Carlton D 4 Bradford Box 02138	617 338-9219	
17 822-2962	Carlton S 10 Thompson Ln Mt 02336	617 696-6919	
17 427-5712	A Heber A 200 Pitman Av Cambridge 02142	617 492-4174	
17 569-2698	A 31 Beane Wy Hudson 02119	617 442-1219	
17 667-5190	A M 250 Main St Av 02115	617 266-7153	
17 569-1417	Adams 301 Carter St Mt 02336	617 698-9074	
17 338-1101	Adams P 42 West St 02336	617 945-2711	
17 822-1993	Carte Anne MD 1101 Beacon Bn 02444	617 739-1022	
17 296-1193	Carte Adhena 971 Newbury Boston 02116	617 536-6229	
17 670-2078	B C 10 Gladstone Av Mt 02336	617 296-6911	
17 621-9001	Carte Barbara L MD Turks New England Medical Center 02111	617 463-0951	
17 296-4725	Carte Becky 02134	617 523-4368	
17 542-1521	Bernard J 301 Ashdown E Rn 02336	617 567-9430	
17 364-5232	Bibb 25 Midway Rd 02336	617 298-8713	
17 541-5649	Bibb 100 W Newbury St 02336	617 367-9931	
17 739-2662	Carte Broadcasting Co 50 Park Pl Bn 02134	617 423-0210	
17 879-0030	Carte B C 73 East Can 02441	617 225-0200	
17 541-3948	Carte C 2000 Cambridge St 02135	617 782-2118	
17 436-1511	C 210 Fenwick Av East Boston 02336	617 569-1545	
17 569-4119	C 109 Harvard Can 02336	617 491-4822	
909 569-8782	C & M 41 Northgate 02336	617 526-4392	
	C & M 41 Northgate 02336	617 526-9558	
	Carter F 514 Hicks Box 02135	617 327-1105	
	Faye & Ricky 20 Columbia Av Bn 02136	617 437-7331	
	Francis S 134 Temple W Av 02132	617 323-6781	
	Franklin & Anne 701 Mt Auburn Can 02136	617 354-0798	
	Fred 40 Harvard Av 02136	617 524-3078	
	Fred 76 Newbury Av Mt 02336	617 698-1343	
	G & B 8 Vardon Box 02134	617 436-8906	
	G T 27 Fenwick Av Mt 02336	617 623-7121	
	Gayle 25 Franklin St 02134	617 823-8322	
	Geo S 115 Mount Mt Av 02136	617 522-3215	
	George 52 Madison Box 02134	617 367-9548	
	Carter Hillside Assoc 107 S Street Box 02111	617 456-1689	
	Carter Harry F 30 Bayne Rd W Av 02132	617 325-5465	
	Carter Hide Co Inc 140 Boston St W Av 02132	617 542-7987	
	Carter Hilary 41 Harvey Can 02348	617 876-2750	
	Horace 301 Walnut Av Hudson 02119	617 442-5307	
	Howard Jr 28 New One Box 02118	617 445-5532	
	J Can 15 Chatham Ln 02444	617 232-7990	
	J 538 Harvard St 02444	617 730-9483	
	J 775 The Pine Way Hudson 02119	617 323-5574	
	Carter J Jacques MD 1 Brookline Pl Bn 02444	617 735-8787	
	Carter J M 3410 Columbia Rd S Box 02137	617 464-1040	
	Carter J M Ornamental Ironworks 1000 Cambridge St 02136	617 436-5353	
	Carter J Neal Co 40 Newbury St 02118	617 442-1775	
	Carle James 1573 Cambridge St Can 02336	617 492-1214	
	James 622 Foster Av Hudson 02336	617 739-2193	
	John 31 East Star Rd Cambridge 02141	617 876-8841	
	John 14 Newbury Rd Mt 02336	617 361-0773	
	Jane 14 Adams Rd Newton 02459	617 564-0435	
	John 1200 Cambridge St 02136	617 426-9094	
	John 11 Mansfield St 02336	617 987-2163	
	John 207 Summer St 02136	617 423-4334	
	John 40 Harvard St 02136	617 282-1235	
	Jane O 129 A Summit Av Bn 02135	617 734-6109	
	J 29 Harvard St 02136	617 265-8656	
	K 17 Concord Street 02132	617 282-1593	
	Carter Nellie E 323 Main St Av Bn 02115	617 267-6483	
	Nicholas S F 115 Randolph Av Mt 02336	617 698-5307	
	Nick 21 Fenwick Box 02116	617 267-5222	
	Nick & Debbi 136 Hermit Rd Newton 02459	617 527-0480	
	Norman G 38 Chickadee Dr 02126	617 822-1203	
	P 40 Cambridge Pl Bn 02115	617 427-4754	
	P E 501 E South S Box 02337	617 268-4213	
	P L 44 Hutchings Box 02115	617 427-9170	
	P R 91 Boyer Can 02136	617 968-8692	
	Paul & Constance 114 Adams Av W Mt 02110	617 325-3034	
	Paul E 501 E South S Box 02337	617 268-4546	
	Paul M 27 Union St 02139	617 787-2115	
	Carter Pile Driving Inc 27 Beaver Ct Framingham 02702	Wellesley Tpk-781.235-0488	
	Carter Prudence 40 Franklin Waterbury 02172	617 393-3782	
	Prudence 40 Franklin Waterbury 02172	617 926-7063	
	Reginald 100 Brookside Circle 02124	617 541-2843	
	Renée & Andrew 30 Walnut Box 02118	617 720-3765	
	Carter Rice David Building Division 163 Main Wilmington 01887 Toll Free-Dial '7' & Then.....800 638-1671 Toll Free-Dial '7' & Then.....800 619-7447 Toll Free-Dial '7' & Then.....800 648-7447 Toll Free-Dial '7' & Then.....800 648-7447 Inquire 613 Main Wilmington 01887 978 988-7447		
	Richard Ingalls Centre 163 Main Wilmington 01887 Toll Free-Dial '7' & Then.....800 638-1673		
	Carter Richard 2075 Carleton Av Brighton 02111	617 987-0836	
	Richard A 97 W Vernon St 02336	617 566-7293	
	Carter Richard A 1200 Cambridge St 02136	617 267-0710	
	Carter Richard K 123 Merwin St Box 02337	617 268-0468	
	Robert L 175 Newbury Av Can 02141	617 864-1535	
	Royce 130 St Braughn Box 02111	617 424-6148	
	Royce & Andrew 1800 Cambridge St 02136	617 491-6115	
	Royce 18 Sandway Cir 02129	617 241-9418	



\rightsquigarrow We develop a (conceptual) geography of clusterings

Many Thousands of Clusterings, Sorted & Organized

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



Evaluating Performance

Evaluating Performance

- Goals:

Evaluating Performance

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

Evaluating Performance

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

Evaluating Performance

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

Evaluating Performance

- Goals:
 - **Validate Claim:** computer-assisted conceptualization outperforms human conceptualization
 - **Demonstrate:** new experimental designs for cluster evaluation
- Three evaluations:
 - Cluster Quality \Rightarrow RA coders

Evaluating Performance

- Goals:
 - **Validate Claim:** computer-assisted conceptualization outperforms human conceptualization
 - **Demonstrate:** new experimental designs for cluster evaluation
- Three evaluations:
 - Cluster Quality \Rightarrow RA coders
 - Informative discoveries \Rightarrow Experienced scholars analyzing texts

Evaluating Performance

- Goals:
 - **Validate Claim:** computer-assisted conceptualization outperforms human conceptualization
 - **Demonstrate:** new experimental designs for cluster evaluation
- Three evaluations:
 - Cluster Quality \Rightarrow RA coders
 - Informative discoveries \Rightarrow Experienced scholars analyzing texts
 - Discovery \Rightarrow You're the judge

Evaluation 1: Cluster Quality

Evaluation 1: Cluster Quality

- Experimental Design to Assess Cluster Quality

Evaluation 1: Cluster Quality

- Experimental Design to Assess Cluster Quality
 - automated visualization to choose one clustering

Evaluation 1: Cluster Quality

- **Experimental Design to Assess Cluster Quality**
 - automated visualization to choose one clustering
 - Evaluate many pairs of documents:
(1) unrelated, (2) loosely related, (3) closely related

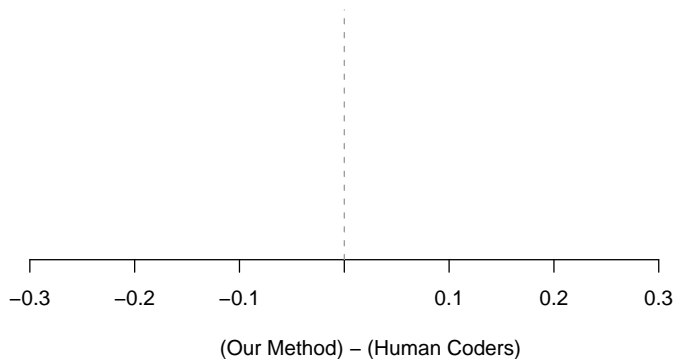
Evaluation 1: Cluster Quality

- **Experimental Design to Assess Cluster Quality**
 - automated visualization to choose one clustering
 - Evaluate many pairs of documents:
 - (1) unrelated, (2) loosely related, (3) closely related
 - Cluster Quality = $\text{mean}(\text{within cluster}) - \text{mean}(\text{between clusters})$

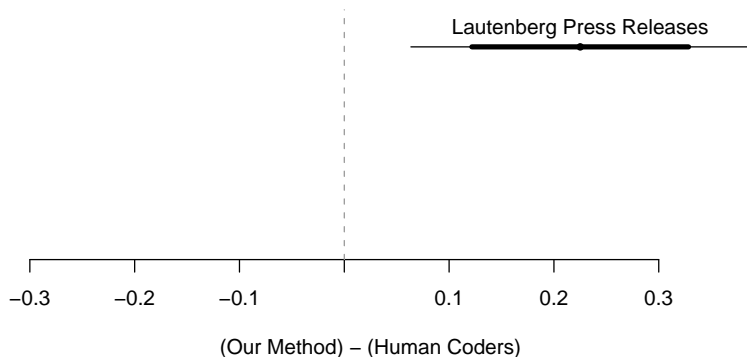
Evaluation 1: Cluster Quality

- **Experimental Design to Assess Cluster Quality**
 - automated visualization to choose one clustering
 - Evaluate many pairs of documents:
 - (1) unrelated, (2) loosely related, (3) closely related
 - Cluster Quality = mean(within cluster) - mean(between clusters)
 - Bias results against ourselves by not letting evaluators choose clustering

Evaluation 1: Cluster Quality

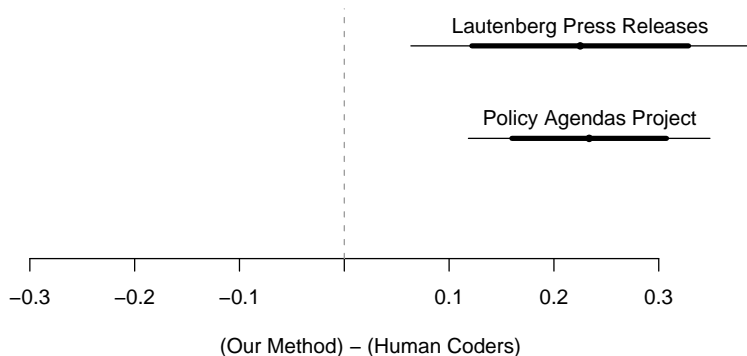


Evaluation 1: Cluster Quality



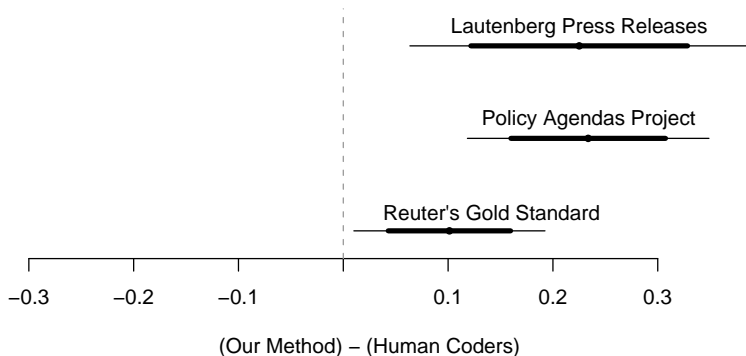
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

Evaluation 1: Cluster Quality



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

Evaluation 1: Cluster Quality



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

Evaluation 2: More Informative Discoveries

Evaluation 2: More Informative Discoveries

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

Evaluation 2: More Informative Discoveries

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

Evaluation 2: More Informative Discoveries

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

Evaluation 2: More Informative Discoveries

- 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)

Evaluation 2: More Informative Discoveries

- 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)

Evaluation 2: More Informative Discoveries

- 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

Evaluation 2: More Informative Discoveries

- 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

Evaluation 2: More Informative Discoveries

- 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:

Evaluation 2: More Informative Discoveries

- 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”:

Our Method 1 \rightarrow vMF 1 \rightarrow vMF 2 \rightarrow Our Method 2 \rightarrow K-Means 1 \rightarrow K-Means 2

Evaluation 2: More Informative Discoveries

- 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”:

Our Method 1 \rightarrow vMF 1 \rightarrow vMF 2 \rightarrow Our Method 2 \rightarrow K-Means 1 \rightarrow K-Means 2

“Genetic testing”:

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

Evaluation 3: What Do Members of Congress Do?

Evaluation 3: What Do Members of Congress Do?

- David Mayhew's (1974) famous typology

Evaluation 3: What Do Members of Congress Do?

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

Evaluation 3: What Do Members of Congress Do?

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

Evaluation 3: What Do Members of Congress Do?

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

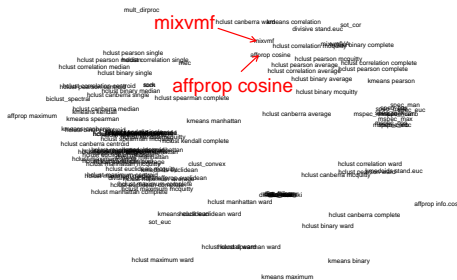
Evaluation 3: What Do Members of Congress Do?

- David Mayhew's (1974) famous typology
 - Advertising
 - Credit Claiming
 - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)

Evaluation 3: What Do Members of Congress Do?

- 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

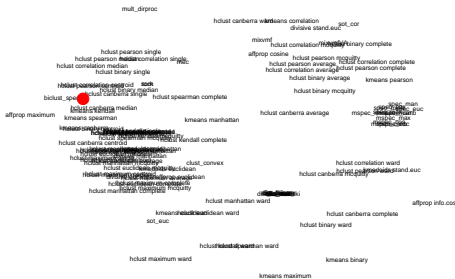
Example Discovery



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

Close to: Mixture of von Mises-Fisher distributions (Banerjee et. al. 2005)

Example Discovery



Space between methods:

Example Discovery



Space between methods:
local cluster ensemble

Example Discovery



Mixture:

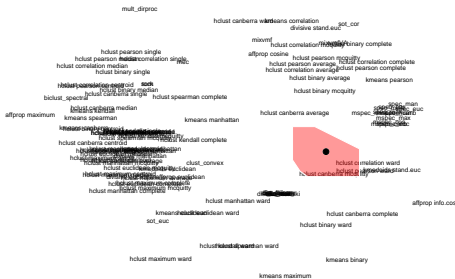
Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering
Random Walk
(Metrics 1-6)

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering
Random Walk
(Metrics 1-6)

0.13 Hclust-Correlation-Ward

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

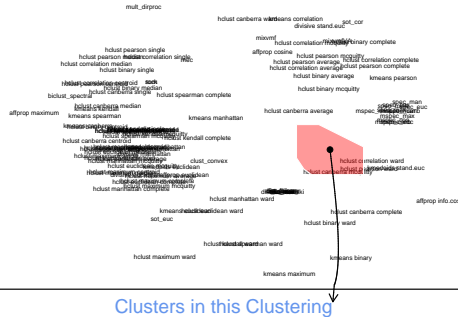
0.30 Spectral clustering
Random Walk
(Metrics 1-6)

0.13 Hclust-Correlation-Ward

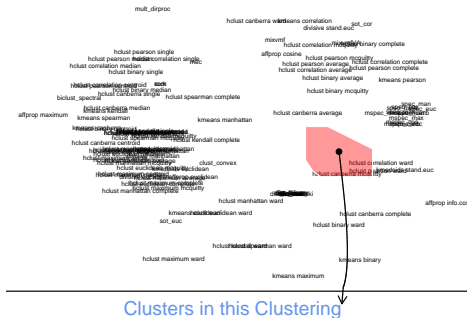
0.09 Hclust-Pearson-Ward

0.05 Kmedioids-Cosine

Example Discovery



Example Discovery

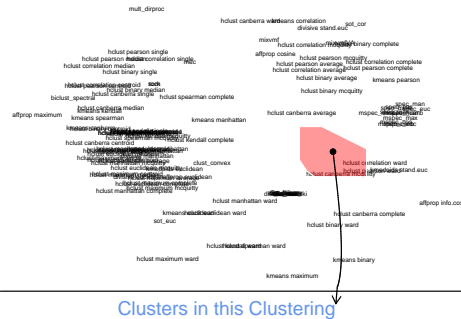


Credit Claiming
Pork

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”

Example Discovery



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”



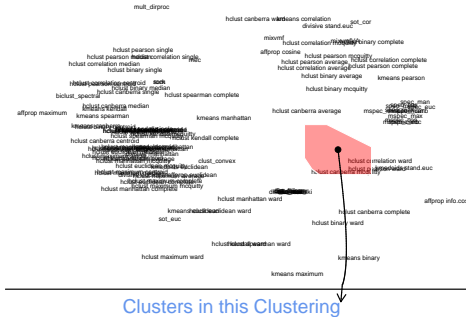
Credit Claiming
Pork



Mayhew Credit Claiming
Legislation

Gary King (Harvard IQSS)

Example Discovery



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



Credit Claiming
Pork



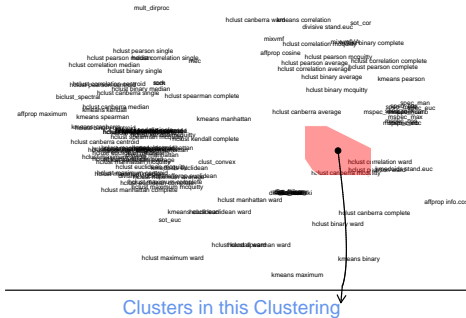
Advertising



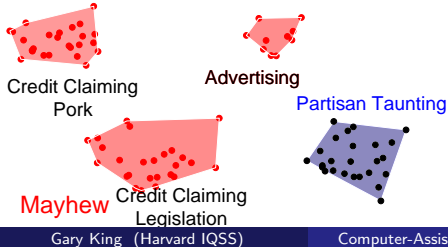
Mayhew Credit Claiming
Legislation

Gary King (Harvard IQSS)

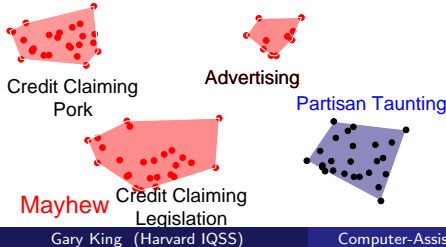
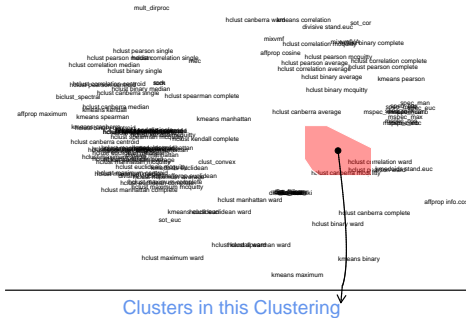
Example Discovery: Partisan Taunting



Partisan Taunting:
“Republicans Selling Out Nation
on Chemical Plant Security”

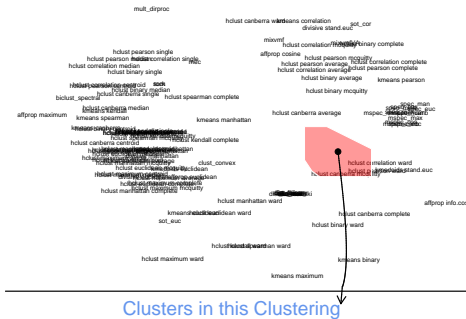


Example Discovery: Partisan Taunting

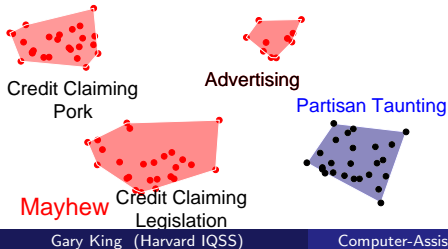


Partisan Taunting:
 “Senator Lautenberg’s amendment would change the name of . . . the Republican bill. . . to ‘More Tax Breaks for the Rich and More Debt for Our Grandchildren Deficit Expansion Reconciliation Act of 2006’”

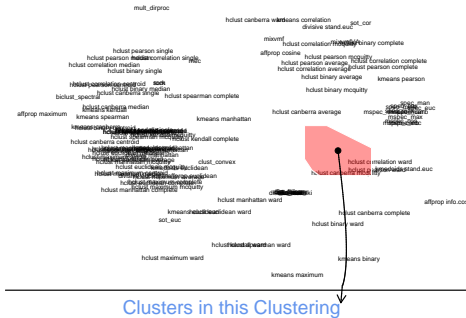
Example Discovery: Partisan Taunting



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

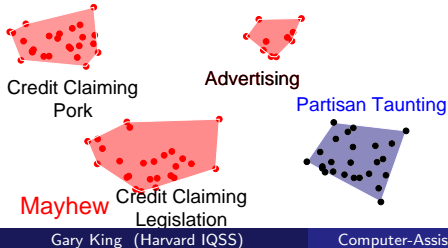


Example Discovery: Partisan Taunting

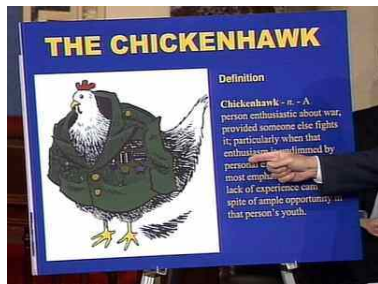


Definition: Explicit, public, and negative attacks on another political party or its members

Taunting ruins deliberation



Taunting ruins deliberation



Sen. Lautenberg
on Senate Floor
4/29/04

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

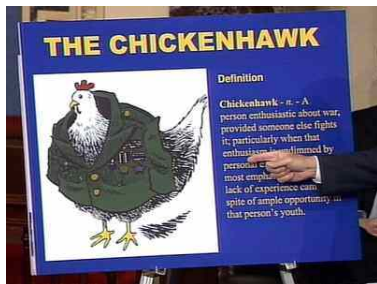
Taunting ruins deliberation



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]

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

Out of Sample Confirmation of Partisan Taunting

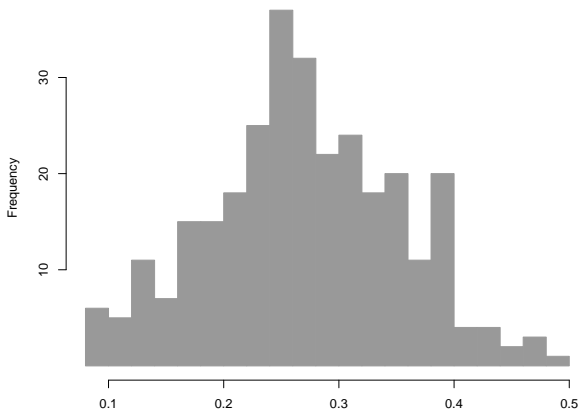
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

Out of Sample Confirmation of Partisan Taunting

- 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

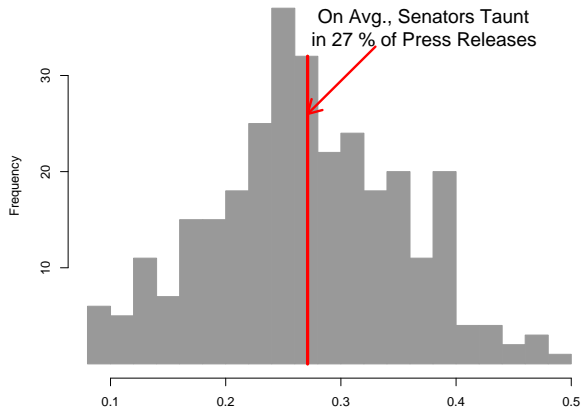
Out of Sample Confirmation of Partisan Taunting

- 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

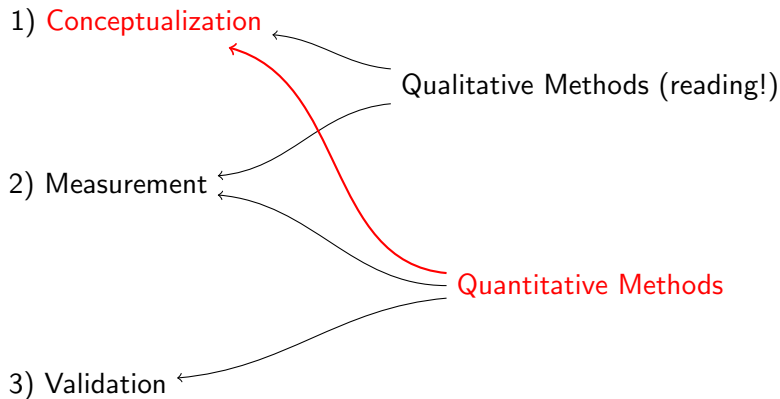


Out of Sample Confirmation of Partisan Taunting

- 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

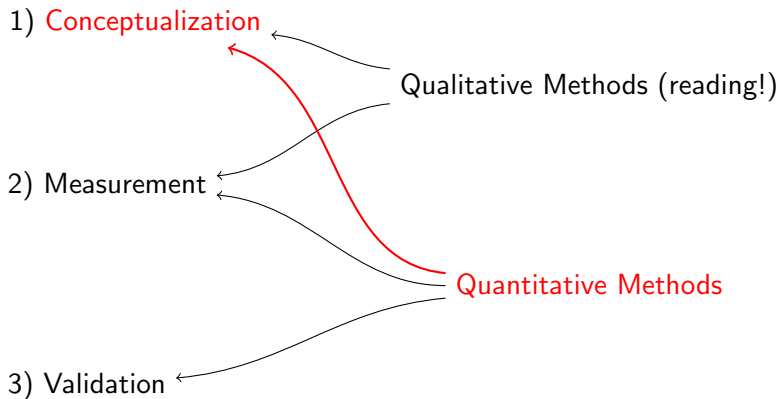


Quantitative Methods for Qualitative Conceptualization



Computer-Assisted Methods for conceptualization and discovery

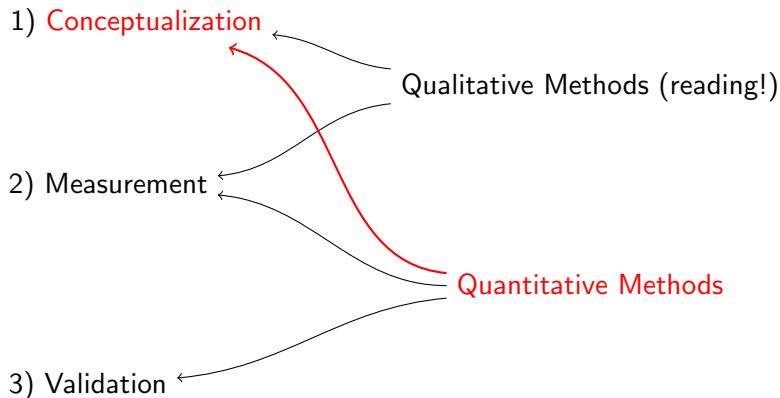
Quantitative Methods for Qualitative Conceptualization



Computer-Assisted Methods for conceptualization and discovery

- Methods designed explicitly for conceptualization

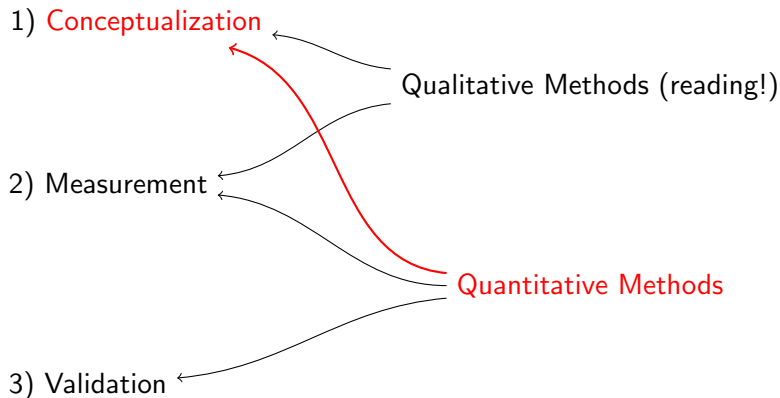
Quantitative Methods for Qualitative Conceptualization



Computer-Assisted Methods for conceptualization and discovery

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

Quantitative Methods for Qualitative Conceptualization



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



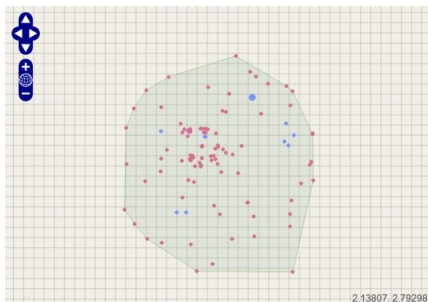
<http://GKing.Harvard.edu>

Software Screenshot

Size: 244 Files

Description: NSF - Updated Set

< > Number of Clusters 5 Clusters (Low) 15 Clusters (Medium) 30 Clusters (High) Discoverable



Display History Display Method Points


Label	Coordinates	Clusters
an interesting clustering [Link]	-0.30819, 0.46229	5
methods-oriented clustering [Link]	0.84753, 1.42538	5

(*) Discoverable

Coordinates: 0.84753, 1.42538

Clusters: 5


Label [+] methods-oriented clustering

29.51% 
72 research community health science public practice global political national urban
Label [+]

[View Detail](#)

27.46% 
67 data economic markets policy survey models financial use not risk
Label [+]


[View Detail](#)

21.72% 
53 human social science systems behavioral networks brain spatial complex dynamics
Label [+]

[View Detail](#)

15.16% 
37 education students school learning creative skills teaching cognitive college teachers
Label [+]

[View Detail](#)

6.15% 
15 language linguistic speech data speakers computer semantic cultural variation
documentation
Label [+]

[View Detail](#)