

Computer-Assisted Clustering and Conceptualization from Unstructured Text

Gary King

Institute for Quantitative Social Science
Harvard University

(talk at University of Kentucky, 4/20/2012)

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

A Method for Computer Assisted Conceptualization

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
- We focus on unstructured text; methods apply more broadly.

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
- We focus on unstructured text; methods apply more broadly.
- Main goal: Switch from **Fully Automated** to **Computer Assisted**

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)

- Clustering seems easy; its not!

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- 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)

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- 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)
- $Bell(5) = 52$

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- 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)
- $Bell(5) = 52$
- $Bell(100) \approx 10^{28} \times$ Number of elementary particles in the universe

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- 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)
- $Bell(5) = 52$
- $Bell(100) \approx 10^{28} \times$ Number of elementary particles in the universe
- Now imagine choosing the *optimal* classification scheme by hand!

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

The Problem with Fully Automated Clustering

The Problem with Fully Automated Clustering

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

The Problem with Fully Automated 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

The Problem with Fully Automated 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
- Existing methods:

The Problem with Fully Automated 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
- Existing methods:
 - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy k -modes, affinity propagation, self-organizing maps,...

The Problem with Fully Automated 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
- Existing methods:
 - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy k -modes, affinity propagation, self-organizing maps,...
 - **Well-defined** statistical, data analytic, or machine learning foundations

The Problem with Fully Automated 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
- Existing methods:
 - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy k -modes, affinity propagation, self-organizing maps,...
 - **Well-defined** statistical, data analytic, or machine learning foundations
 - How to add substantive knowledge: With few exceptions, **unclear**

The Problem with Fully Automated 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
- Existing methods:
 - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy k -modes, affinity propagation, self-organizing maps,...
 - **Well-defined** statistical, data analytic, or machine learning foundations
 - How to add substantive knowledge: With few exceptions, **unclear**
 - The literature: **little guidance on when methods apply**

The Problem with Fully Automated 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
- Existing methods:
 - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy k -modes, affinity propagation, self-organizing maps,...
 - **Well-defined** statistical, data analytic, or machine learning foundations
 - How to add substantive knowledge: With few exceptions, **unclear**
 - The literature: **little guidance on when methods apply**
 - **Deriving such guidance:** difficult or impossible

The Problem with Fully Automated 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
- Existing methods:
 - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy k -modes, affinity propagation, self-organizing maps,...
 - **Well-defined** statistical, data analytic, or machine learning foundations
 - How to add substantive knowledge: With few exceptions, **unclear**
 - The literature: **little guidance on when methods apply**
 - **Deriving such guidance:** difficult or impossible
- **Deep problem:** full automation requires more information

The Problem with Fully Automated 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
- Existing methods:
 - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy k -modes, affinity propagation, self-organizing maps,...
 - **Well-defined** statistical, data analytic, or machine learning foundations
 - How to add substantive knowledge: With few exceptions, **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

Switch from Fully Automated to Computer Assisted

Switch from Fully Automated to Computer Assisted

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

Switch from Fully Automated to Computer Assisted

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

Switch from Fully Automated to Computer Assisted

- **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

Switch from Fully Automated to Computer Assisted

- **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!

Switch from Fully Automated to Computer Assisted

- **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

Switch from Fully Automated to Computer Assisted

- **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

Switch from Fully Automated to Computer Assisted

- **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

Switch from Fully Automated to Computer Assisted

- **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?**

Our Idea: Meaning Through Geography

Set of clusterings

Our Idea: Measuring Through Geography

Set of clusterings \approx
A list of unconnected addresses

wide at **SuperPages.com**

	195	Car	C
Cartage New England Inc 28 Allen Ln Ipswich 01938.....	978 356-9960		
Cartagena Lydia 38 Sweet Box 00319.....	617 323-7639		
Cartagena Aveth F Blandish Box 00319.....	617 442-9780		
B Had 00319.....	617 361-5253		
Justicia 50 Decatur Cha 00229.....	617 241-0152		
Lucilla 124 Harvard Cam 00219.....	617 491-5621		
M 95 Howe Box 00313.....	617 323-9713		
Melvin 501 Green Cam 00219.....	617 576-1061		
Carte Nicholas 38 Appleton Boston 00219.....	617 695-6996		
Carterlena O 44 Bedford Box 00219.....	617 338-9219		
Cartern Thos J Sr & Claire 1 Plover Rd MA 00219.....	617 698-6163		
Thomas & Kathleen 50 Thompson Ln MA 00219.....	617 696-6919		
Carter A Box 00213.....	617 329-2257		
A Huber A 31 Bethune Wy Roxbury 00219.....	617 442-5230		
A 200 Putnam Av Cambridge 00219.....	617 492-4174		
A M 255 Murchies Av Box 00213.....	617 266-7153		
Adams 381 Centre St MA 00219.....	617 698-9074		
Alice 108 Elmwood Box 00213.....	617 425-0193		
Alice 40 Market Cambridge 00219.....	617 945-2711		
Andrew F 42 Wal St Box 00213.....	617 625-7532		
Carter Anne MD 1161 Beacon St 00446.....	617 739-1022		
Carter Adhans 272 Newbury Boston 00219.....	617 536-6329		
B E C 48 Graduate Av MA 00219.....	617 296-6911		
Carter Barbara L MD Tufts New England Medical Center Box 00211.....			
Cal.....	617 636-9051		
Carter Becky Box 00214.....	617 523-4368		
Carter Adhans Bernard J.....			
122 Cambridge F Box 00219.....	617 567-3430		
Bithiah 25 Melrose Dr 00219.....	617 299-8713		
Bithiah 25 Melrose Dr 00219.....	617 367-9931		
Carter Broadcasting Co 26 Park Plt Box 00219.....	617 423-9210		
Carter & Burgess Consultants Inc 73 East St Cam 00214.....	617 225-0200		
Carter C 200 Cambridge Av Box 00213.....	617 782-2118		
C 219 Concord Av East Boston 00218.....	617 569-1545		
C 109 Harvard Cam 00219.....	617 491-4822		
C 00219.....	617 491-4822		
C & M 41 Burroughs Jan 00219.....	617 524-9559		
Carter F 24 Hillock Box 00213.....	617 327-1105		
Faye & Ricky 207 Cambridge Av Box 00219.....	617 437-7331		
Francis S 134 Temple W Av Box 00212.....	617 323-6781		
Franklin & Anne 251 Mt Auburn Cam 00219.....	617 354-0798		
Fred 42 Harvard Jan 00219.....	617 524-3078		
Fred 96 Harvard Av MA 00219.....	617 698-1343		
G & R 8 Myer Den 00219.....	617 436-8906		
G T 27 Fyfield Av Sun 00215.....	617 623-7121		
Gayle 25 Providence Dr 00219.....	617 825-0322		
Geo S 115 Moss Hill Rd Jan 00218.....	617 522-3215		
George 125 Nathan Box 00214.....	617 367-9548		
Carter Halliday Associate 107 S Street Box 00211.....	617 456-1689		
Carter Harry F 26 Baum St Rd W Box 00212.....	617 325-5465		
Carter Hide Co Inc 144 Sumner Box 00219.....	617 542-7987		
Carter Hilary 41 Harvey Cam 00214.....	617 876-2750		
Horace 381 Walnut Av Roxbury 00219.....	617 442-5307		
Howard Jr 28 Notre Dame Box 00218.....	617 445-5552		
J Cam 41 Chatham Box 00446.....	617 334-2658		
J Cam 413 Harvard Box 00446.....	617 232-7990		
J 518 Harvard Box 00446.....	617 730-9483		
J 775 The Youngest Wy Roxbury 00213.....	617 323-5374		
Carter J Jacques MD 1 Broadbent Pl Box 00446.....	617 735-8787		
Carter J M 3410 Columbia St Box 00217.....	617 464-1040		
Carter J M Ornamental Ironworks Pembroke Falls 00217.....	617 436-5353		
Carter J Veal Co 40 Newbury St Box 00218.....	617 442-1775		
Carter James 1573 Cambridge St Cam 00219.....	617 492-1214		
James 352 Fisher Av Roxbury 00219.....	617 739-2193		
James 31 Gold Star Rd Cambridge 00214.....	617 876-8841		
Jan L 34 Roslindale Rd MA 00219.....	617 361-0773		
Jan L 34 Roslindale Rd MA 00219.....	617 364-0435		
Jeffrey 41 Warren Av Box 00219.....	617 426-5994		
John 11 Marlboro St 00219.....	617 987-2163		
John 107 Sumner Box 00213.....	617 423-4334		
John 40 Westwood Rd 00219.....	617 282-1235		
June O 129 A Summit Av Box 00219.....	617 734-6109		
K 280 Beverly Av Box 00219.....	617 265-9456		
K 17 Exford Dorchester 00212.....	617 282-1593		
Carter Nellie E 323 Marshfield Av Box 00215.....	617 267-6483		
Nicholas S F 115 Randolph Av MA 00219.....	617 698-5307		
Nick 21 Fairfield Box 00214.....	617 267-5222		
Nick & Debbi 156 Vermont Rd Newton 00245.....	617 527-0480		
Nicole 38 Chickadee Dr 00219.....	617 822-1201		
P 40 Cranston Pl Box 00213.....	617 427-4754		
P E 501 E South St Box 00217.....	617 268-8213		
P L 44 Hastings Box 00211.....	617 427-9170		
P R 91 Boyer Jan 00219.....	617 968-8692		
Paul & Constance 114 Aspen Av W Box 00210.....	617 325-2036		
Paul E 501 E South St Box 00217.....	617 268-4546		
Paul M 27 Union St 00219.....	617 787-2115		
Carter Pike Driving Inc 27 Beaver Ct Framingham 01702.....	Wellesley Falls 781.235-0488		
Carter Prudence 40 Franklin Woburn 00217.....	617 393-3782		
Prudence 40 Franklin Woburn 00217.....	617 926-7063		
Reginald 106 Brookwood Dorchester 00212.....	617 541-2843		
Renee & Andrew 10 Walnut Box 00218.....	617 720-3765		
Carter Rice David Baker Dennis Publishing 163 Main Wilmington 01887.....	Toll Free 800-638-1671		
Toll Free 800-638-1671 Cost Rec Industrial Prod 113 Main Wilmington.....	800 619-7447		
Toll Free 800-619-7447 Toll Free 800-619-7447.....	800 648-7447		
Headquarters 413 Main Wilmington 01887.....	978 988-7447		
Cal Ingalls Crane 163 Main Wilmington 01887.....	800 638-1673		
Toll Free 800-619-7447 Ingalls Crane 163 Main Wilmington 01887.....	978 988-7447		
Carter Richard 207 Cambridge Av Brighton 00215.....	617 987-0836		
Richard A 87 Mt Vernon Box 00216.....	617 566-7293		
Carter Richard A MD 120 Cambridge Av Box 00216.....	617 267-0710		
Carter Richard K 13 Merwin St Box 00217.....	617 268-9448		
Robert L 175 Melrose Av Cam 00214.....	617 864-1535		
Roger 150 St Pauls Box 00213.....	617 424-6148		
Roy 41 Concord St 00219.....	617 491-6115		
Royce 18 Sumner Cha 00219.....	617 241-0418		

Our Idea: Meaning Through Geography

Set of clusterings \approx

A list of unconnected addresses

wide at SuperPages.com

	195	Car	C
Cartage New England Inc			
37 566-1282	26 Allen Ln Ipswich 01908	978 356-9960	
Cartagena Lydia			
81 447-4101	38 Sweet Rd 02131	617 323-7639	
Cartagena Avelis			
90 257-9961	1 Fairview Rd 02139	617 442-9780	
	8 Hrd 02136	617 361-5253	
37 566-1282	Justice 50 Decatur Cha 02129	617 241-0152	
37 364-5188	Lucilla 124 Harvard Cha 02139	617 491-5621	
	M 95 Howe Rd 02133	617 323-9713	
361-0380	Melvin 501 Green Cam 02139	617 576-1061	
Carte Nicholas			
37 566-4548	38 Appleton Boston 02114	617 695-6996	
	Cartagena O 4480rd Cha 02131	617 338-9219	
37 628-8248	Carten Thos J Sr & Claire		
	1 Ivesdale Rd 02136	617 698-6163	
37 445-5116	Thomas & Kathleen		
	50 Thompson Ln 02136	617 696-6919	
37 822-9902	Carte A R 02133	617 339-2257	
37 427-5712	A Huber	617 442-5230	
37 569-2698	A 31 Bethune Wy 02139	617 442-1219	
	A 200 Putnam Av Cambridge 02139	617 492-4174	
37 667-5190	A M 255 Murchies Av 02136	617 266-7153	
	Adams 381 Centre St 02136	617 698-9074	
37 569-1417	Alice 108 Elmwood Rd 02136	617 425-0193	
37 338-9110	Alice 40 Market Cambridge 02139	617 945-2711	
	Andrew F 42 Wal St 02133	617 625-7635	
37 825-1919	Carte Anne MD	617 739-1022	
	1101 Beacon W 02144		
37 296-1593	Carte Becky 02134	617 536-6329	
37 670-2078	B E 108 Graduate Av 02136	617 296-6911	
37 623-9001	Carte Barbara L MD		
	Tuffs New England Medical Res 02131	617 636-9051	
37 296-4725	Carte Becky 02134	617 523-4368	
37 542-1521	Bernard J		
	121 Cambridge St 02136	617 567-3430	
37 364-5232	Bethiah 25 Midway Cha 02136	617 298-8713	
37 541-5649	Bible 1015 W 02136	617 367-9931	
37 739-2662	Carte Broadcasting Co		
	26 Park Pl 02136	617 423-0210	
37 879-0030	C 21 2nd St 02131	617 225-0200	
37 541-3948	Carte C 2000 Cambridge Av 02135	617 782-2118	
37 936-1511	C 219 Harvard Av East Boston 02128	617 569-1545	
37 569-4119	C 109 Harvard Cha 02136	617 491-4822	
	C 109 Harvard Cha 02136	617 296-4392	
37 569-4782	C & M 41 Burroughs Jan 02136	617 524-5595	
	Carter F 24 Hilltop 02131	617 327-1105	
	Faye & Ricky		
	107 Columbia Av 02136	617 437-7331	
	Francis S 134 Temple W Av 02132	617 323-6781	
	Franklin & Anne		
	291 Mt Auburn Cam 02136	617 354-0798	
	Fred 42 Harvard Cha 02136	617 524-3078	
	Fred 16 Harvard Av 02136	617 698-1343	
	G & E 8 Harvard Cha 02136	617 436-8906	
	G T 27 Fyfield Av 02135	617 623-7121	
	Gayle 25 Franklin Cha 02136	617 825-0322	
	Geo S 115 Mount Mt Rd Jan 02136	617 522-3215	
	George 125 Madison 02131	617 367-9548	
	Carter Halliday Associate		
	107 S Street 02131	617 456-1689	
	Carter Harry F		
	140 Spring St 02131	617 325-5465	
	Carter Hide Co Inc		
	140 Spring St 02131	617 542-7987	
	Carter Hilary 41 Harvey Cam 02140	617 876-2750	
	Horace		
	301 Walnut Av 02136	617 442-5307	
	Howard Jr 38 Nebe One Box 02136	617 445-5552	
	J Cam	617 354-2658	
	J 31 Chatham St 02144	617 232-7990	
	J 31 Harvard St 02144	617 730-9483	
	J 775 Wy Weymouth 02155	617 323-5574	
	Carter J Jacques MD		
	1 Broadview Pl 02144	617 735-8787	
	Carter J M		
	3410 Columbia St 02137	617 464-1040	
	Carter J M Ornamental Ironworks		
	100 Franklin St 02136	617 436-5353	
	Carter J Veal Co		
	40 Newbury St 02136	617 442-1775	
	Carter James		
	1573 Cambridge St Cam 02136	617 492-1214	
	James 102 Federal Av 02136	617 739-2193	
	James 102 Federal Av 02136	617 876-8841	
	Jas L 34 Broadview Rd 02136	617 361-0773	
	Jane 14 Adams Rd 02146	617 964-0435	
	Jeffrey 41 Warren Av 02136	617 426-5994	
	John 11 Marshall Rd 02134	617 987-2163	
	John 207 Summer Box 02131	617 423-4334	
	John 40 Webster Rd 02135	617 282-1235	
	June O 129 A Sunset Av 02138	617 734-6109	
	K 109 Weymouth Av 02136	617 265-9456	
	K 17 Elwood Bedford 02123	617 282-1593	
	Carter Nellie E		
	323 Marchette Av 02135	617 267-6483	
	Nicholas S F		
	115 Randolph Av 02136	617 698-5307	
	Nick 21 Fairfield 02136	617 267-5222	
	Nick & Debbi		
	146 Vermont Rd 02139	617 527-0480	
	Norman G		
	38 Chickadee Dr 02135	617 822-1203	
	P 40 Cranston Pl 02135	617 427-4754	
	P E 501 E South St Box 02137	617 268-4213	
	P L 44 Matthews Box 02131	617 427-9170	
	P R 91 Boyer Jan 02136	617 968-8692	
	Paul & Constance		
	114 Adams Av W 02130	617 325-2036	
	Paul E 501 E South St Box 02137	617 268-4546	
	Paul M 27 Union St 02135	617 787-2115	
	Carter Pile Driving Inc 17 Beaver Ct		
	Franklin 02102	617 781-235-0488	
	Carter Prudence		
	40 Franklin Waterbury 02172	617 393-3782	
	Prudence		
	40 Franklin Waterbury 02172	617 926-7063	
	Reginald		
	100 Broadview Cambridge 02131	617 541-2843	
	Renee & Andrew		
	10 Walnut Box 02136	617 720-3765	
	Carter Rice David		
	Baker Dennis Publishing 163 Main Wilmington 01887		
	Ted Free-Old 1 & Thon	800 638-1671	
	Carl Free-Old 1st Fred 113 Main Wilmington	800 619-7447	
	Ted Free-Old 1 & Thon	800 648-7447	
	Headquarters 413 Main Wilmington 01887		
	Carl	978 988-7447	
	Ingala Crane 363 Main Wilmington 01887	800 638-1673	
	Carter Richard		
	207 Carver Av Brighton 02137	617 987-0836	
	Richard A 974 Vernon Box 02136	617 566-7293	
	Carter Richard A MD		
	170 Weymouth St 02136	617 267-0710	
	Carter Richard K		
	133 Merwin St 02131	617 268-9448	
	Robert L 175 Rockdale Av Cam 02140	617 864-1535	
	Roger 130 St Pauls Box 02131	617 424-6148	
	Roy 41 Concord 02136	617 491-6115	
	Royce 185 Salisbury Cha 02129	617 241-9418	



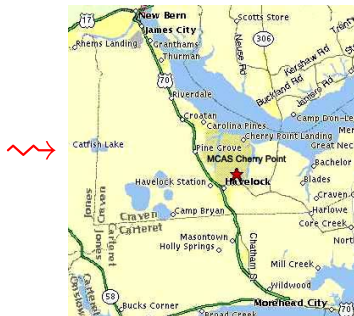
Our Idea: Meaning Through Geography

Set of clusterings \approx

A list of unconnected addresses

wide at SuperPages.com

		195	Car	C
37 566-1282	Cartage New England Inc			
81 447-4101	Cartagena Lydia			
90 257-9961	Cartagena Aveth			
37 566-1282	Cartagena Aveth			
37 566-1282	Cartagena Aveth			
361-0380	Carte Nicholas			
37 566-4548	Carte Nicholas			
37 628-8248	Carten Thos J Sr & Claire			
37 445-5116	Carten Thos J Sr & Claire			
37 822-9992	Carte A Inc			
37 442-5712	Carte A Inc			
37 569-2698	Carte A Inc			
37 667-5190	Carte A Inc			
37 569-1417	Carte A Inc			
37 822-9992	Carte A Inc			
37 296-1593	Carte A Inc			
37 670-2078	Carte A Inc			
37 623-9001	Carte A Inc			
37 296-4725	Carte A Inc			
37 542-1521	Carte A Inc			
37 364-5232	Carte A Inc			
37 541-5649	Carte A Inc			
37 739-2662	Carte A Inc			
37 879-0030	Carte A Inc			
37 936-1511	Carte A Inc			
37 569-4119	Carte A Inc			
37 569-4782	Carte A Inc			



\approx We develop a (conceptual) geography of clusterings

A New Strategy

Make it easy to choose best clustering from millions of choices

A New Strategy

Make it easy to choose best clustering from millions of choices

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

A New Strategy

Make it easy to choose best clustering from millions of choices

- ① **Code text as numbers** (in one *or more* of several ways)
- ② **Apply all clustering methods we can find** to the data — each representing different (unstated) substantive assumptions (<15 mins)

A New Strategy

Make it easy to choose best clustering from millions of choices

- ① **Code text as numbers** (in one *or more* of several ways)
- ② **Apply all clustering methods we can find** to the data — each representing different (unstated) substantive assumptions (<15 mins)
- ③ (Too much for a person to understand, but organization will help)

A New Strategy

Make it easy to choose best clustering from millions of choices

- ① **Code text as numbers** (in one *or more* of several ways)
- ② **Apply all clustering methods we can find** to the data — each representing different (unstated) substantive assumptions (<15 mins)
- ③ (Too much for a person to understand, but organization will help)
- ④ Develop an **application-independent distance metric** between clusterings, a **metric space of clusterings**, and a **2-D projection**

A New Strategy

Make it easy to choose best clustering from millions of choices

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

A New Strategy

Make it easy to choose best clustering from millions of choices

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

A New Strategy

Make it easy to choose best clustering from millions of choices

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

A New Strategy

Make it easy to choose best clustering from millions of choices

- 1 **Code text as numbers** (in one *or more* of several ways)
- 2 **Apply all clustering methods we can find** to the data — each representing different (unstated) substantive assumptions (<15 mins)
- 3 (Too much for a person to understand, but organization will help)
- 4 Develop an **application-independent distance metric** between clusterings, a **metric space of clusterings**, and a **2-D projection**
- 5 “**Local cluster ensemble**” creates a new clustering at any point, based on weighted average of nearby clusterings
- 6 A new **animated visualization** to explore the space of clusterings (smoothly morphing from one into others)
- 7 **↪ Millions of clusterings, easily comprehended**
- 8 (Or, our new strategy: represent the entire bell space directly; no need to examine document contents)

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

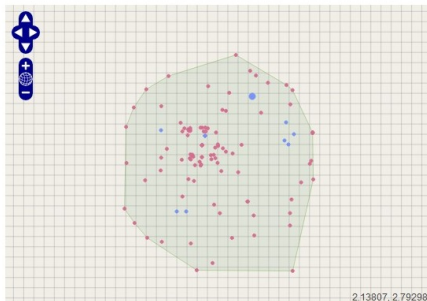


Software Screenshot

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% [\[Link\]](#)
72 research community health science public practice global political national urban
Label [\[+\]](#)

[View Detail](#)

27.46% [\[Link\]](#)
67 data economic markets policy survey models financial use not risk
Label [\[+\]](#)

[View Detail](#)

21.72% [\[Link\]](#)
53 human social science systems behavioral networks brain spatial complex dynamics
Label [\[+\]](#)

[View Detail](#)

15.16% [\[Link\]](#)
37 education students school learning creative skills teaching cognitive college teachers
Label [\[+\]](#)

[View Detail](#)

6.15% [\[Link\]](#)
15 language linguistic speech data speakers computer semantic cultural variation
documentation
Label [\[+\]](#)

[View Detail](#)

Application-Independent Distance Metric: Axioms

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions
 - ① Distance between clusterings: a function of the **pairwise document agreements** (pairwise agreements \Rightarrow triples, quadruples, etc.)

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions
 - ① Distance between clusterings: a function of the **pairwise document agreements** (pairwise agreements \Rightarrow triples, quadruples, etc.)
 - ② **Invariance**: Distance is invariant to the number of documents (for any fixed number of clusters)

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions
 - ① Distance between clusterings: a function of the **pairwise document agreements** (pairwise agreements \Rightarrow triples, quadruples, etc.)
 - ② **Invariance**: Distance is invariant to the number of documents (for any fixed number of clusters)
 - ③ **Scale**: the maximum distance is set to $\log(\text{num clusters})$

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions
 - ① Distance between clusterings: a function of the **pairwise document agreements** (pairwise agreements \Rightarrow triples, quadruples, etc.)
 - ② **Invariance**: Distance is invariant to the number of documents (for any fixed number of clusters)
 - ③ **Scale**: the maximum distance is set to $\log(\text{num clusters})$
- \rightsquigarrow **Only one measure satisfies all three** (the “variation of information”)

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions
 - ① Distance between clusterings: a function of the **pairwise document agreements** (pairwise agreements \Rightarrow triples, quadruples, etc.)
 - ② **Invariance**: Distance is invariant to the number of documents (for any fixed number of clusters)
 - ③ **Scale**: the maximum distance is set to $\log(\text{num clusters})$
- \rightsquigarrow **Only one measure satisfies all three** (the “variation of information”)
- (Meila, 2007, derives same metric using different axioms & lattice theory)

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
 - **Inject human judgement**: relying on insights from survey research

Evaluating Performance

- Goals:
 - **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

Evaluating Performance

- Goals:
 - **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

Evaluating Performance

- Goals:
 - **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
 - 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
 - **Inject human judgement**: relying on insights from survey research
- We now present 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

- What Are Humans Good For?

Evaluation 1: Cluster Quality

- What Are Humans Good For?
 - They can't: keep many documents & clusters in their head

Evaluation 1: Cluster Quality

- What Are Humans Good For?
 - They can't: keep many documents & clusters in their head
 - They can: compare two documents at a time

Evaluation 1: Cluster Quality

- What Are Humans Good For?

- 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

Evaluation 1: Cluster Quality

- What Are Humans Good For?
 - 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

Evaluation 1: Cluster Quality

- What Are Humans Good For?
 - 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

Evaluation 1: Cluster Quality

- What Are Humans Good For?

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

- Experimental Design to Assess Cluster Quality

- automated visualization to choose one clustering
- many pairs of documents

Evaluation 1: Cluster Quality

- What Are Humans Good For?

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

- Experimental Design to Assess Cluster Quality

- automated visualization to choose one clustering
- many pairs of documents
- for coders: (1) unrelated, (2) loosely related, (3) closely related

Evaluation 1: Cluster Quality

- What Are Humans Good For?

- They can't: keep many documents & clusters in their head
- They can: compare two documents at a time
- \Rightarrow 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
- $\text{Quality} = \text{mean}(\text{within cluster}) - \text{mean}(\text{between clusters})$

Evaluation 1: Cluster Quality

- What Are Humans Good For?

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

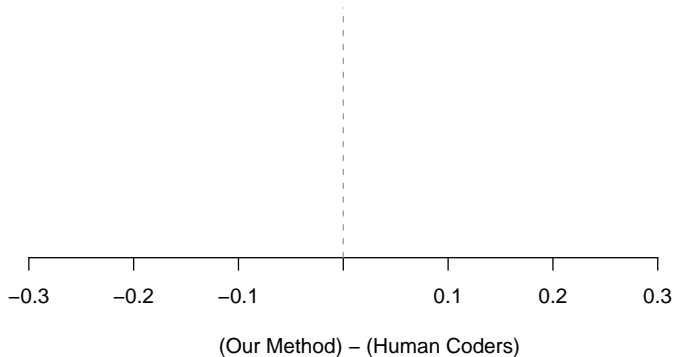
- What Are Humans Good For?

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

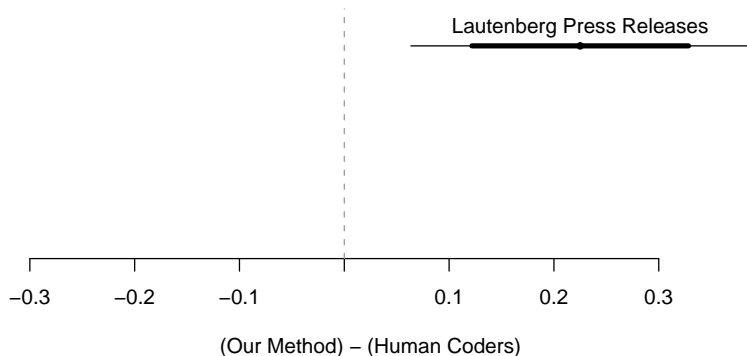
- Experimental Design to Assess Cluster Quality

- 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

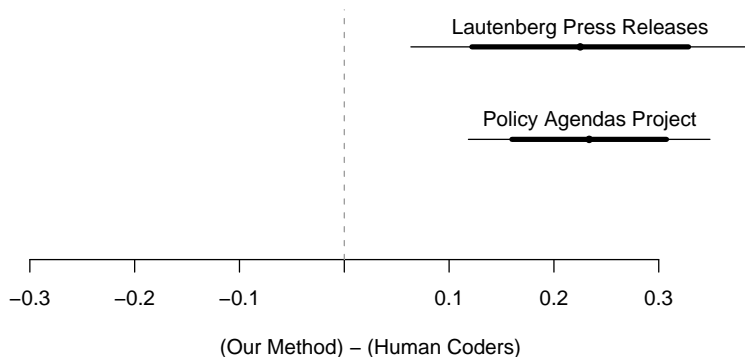


Evaluation 1: Cluster Quality



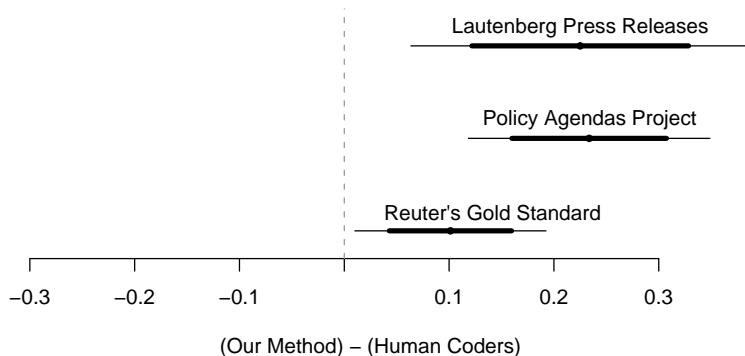
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

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

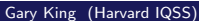
Evaluation 3: What Do Members of Congress Do?

- 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

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ 🔍 ↺

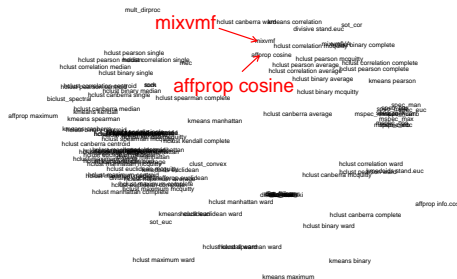


Example Discovery



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

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:

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻

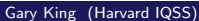


Example Discovery

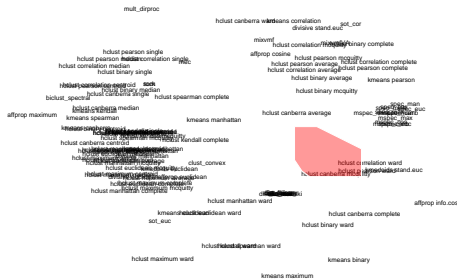


Space between methods:
local cluster ensemble

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ≡ ≡ ≡ ↺ 🔍 ↻



Example Discovery

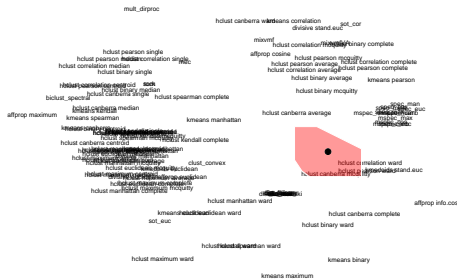


Found a **region** with particularly insightful clusterings

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻



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

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 Kmediods-Cosine

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering
Random Walk
(Metrics 1-6)

0.13 Hclust-Correlation-Ward

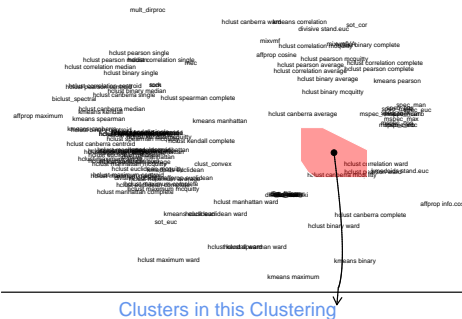
0.09 Hclust-Pearson-Ward

0.05 Kmediods-Cosine

0.04 Spectral clustering
Symmetric
(Metrics 1-6)

Clusters in this Clustering

Example Discovery

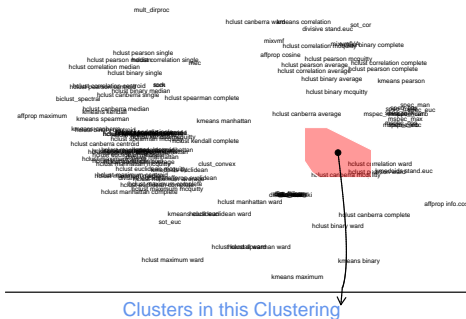


Credit Claiming
Pork

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”

Mayhew

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

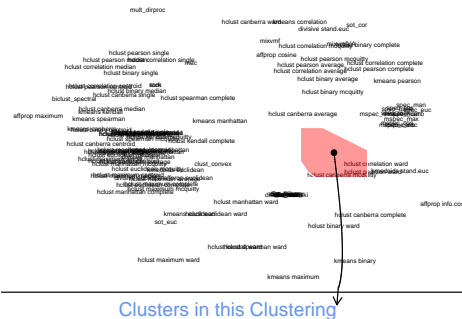
Credit Claiming Legislation

Mayhew

Gary King (Harvard IQSS)

Quantitative Discovery

Example Discovery



Advertising:

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



Credit Claiming
Pork

Advertising

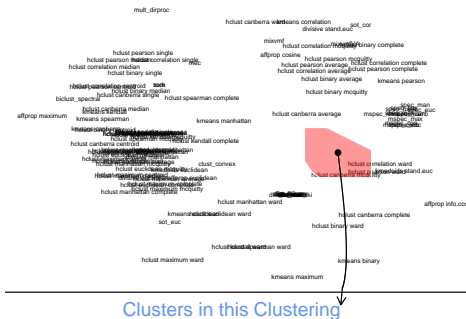
Mayhew

Credit Claiming Legislation

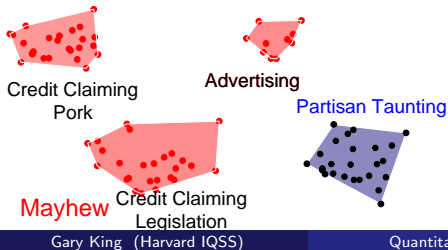
Gary King (Harvard IQSS)

Quantitative Discovery

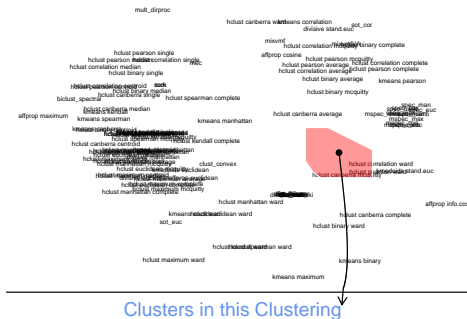
Example Discovery: Partisan Taunting



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’”



Credit Claiming
Pork

Advertising

Partisan Taunting

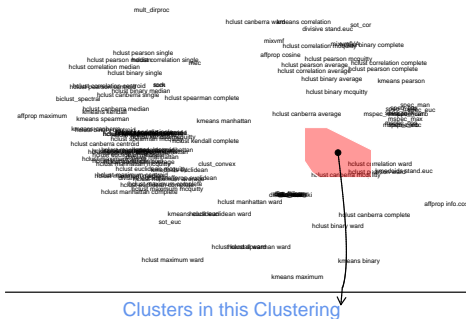
Mayhew

Credit Claiming Legislation

Gary King (Harvard IQSS)

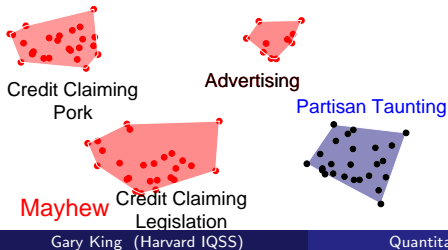
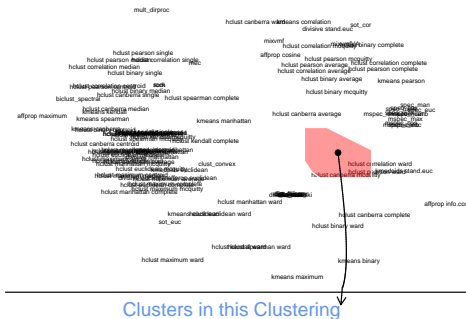
Quantitative Discovery

Example Discovery: Partisan Taunting



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

Example Discovery: Partisan Taunting

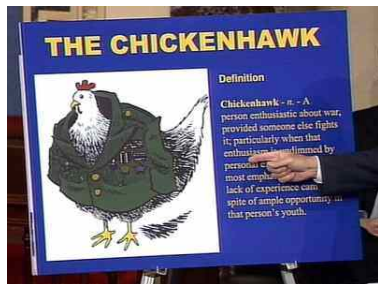


Definition: Explicit, public, and negative attacks on another political party or its members

Taunting ruins deliberation

In Sample Illustration of Partisan Taunting

Taunting ruins deliberation



Sen. Lautenberg
on Senate Floor
4/29/04

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

In Sample Illustration of Partisan Taunting

Taunting ruins deliberation

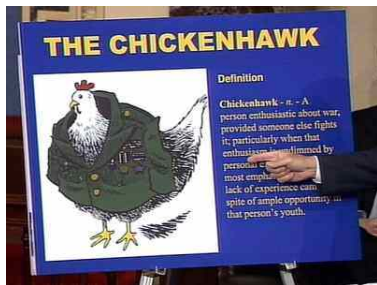


Sen. Lautenberg
on Senate Floor
4/29/04

- “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]

In Sample Illustration of Partisan Taunting

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

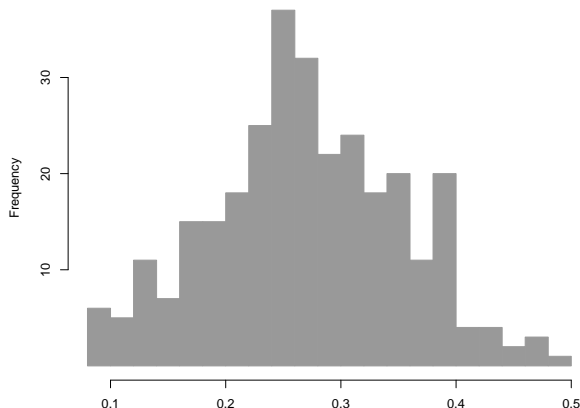
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

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

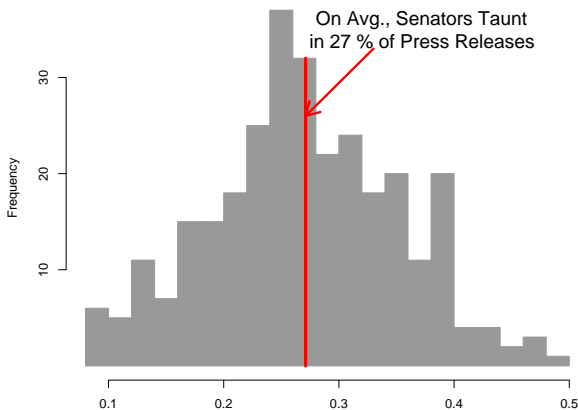
Out of Sample Confirmation of Partisan Taunting

- 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

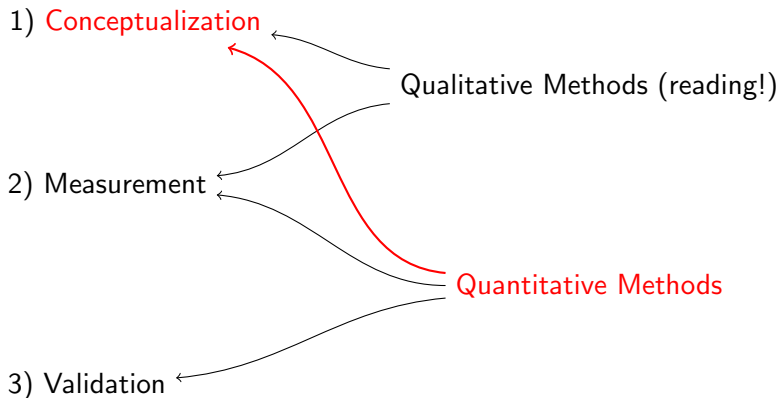


Out of Sample Confirmation of Partisan Taunting

- 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

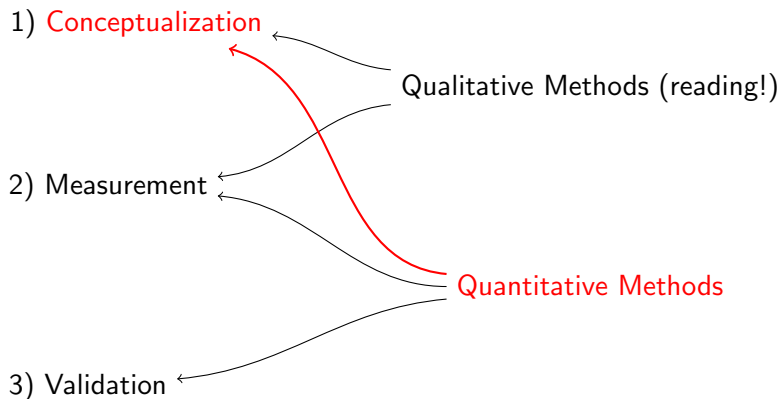


Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

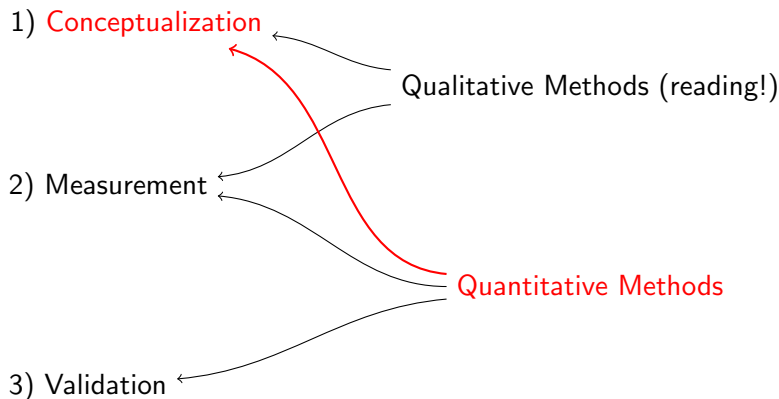
Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

- Few formal methods designed explicitly for conceptualization

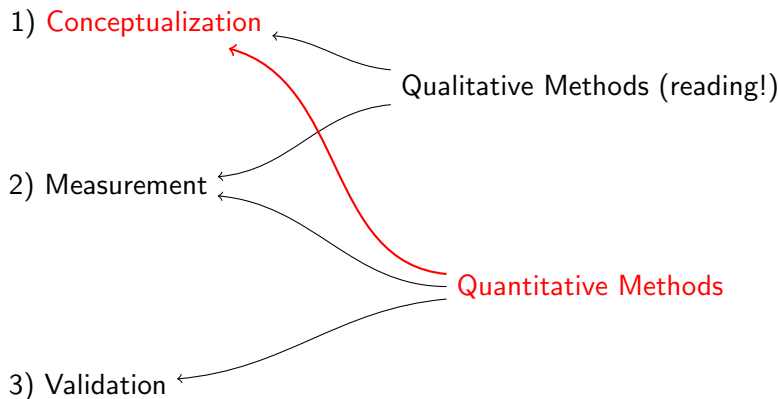
Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

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

Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

- Few formal methods designed explicitly for conceptualization
- Belittled: “Tom Swift and His Electric Factor Analysis Machine” (Armstrong 1967)
- Evaluation methods measure progress in discovery

For more information



<http://GKing.Harvard.edu>