

Computer-Assisted Clustering and Conceptualization from Unstructured Text

Gary King

Institute for Quantitative Social Science
Harvard University

talk at University of Chicago, Computation Institute, 5/9/2011

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
- 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: Meaning 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 28 Sweet Box 00331.....	617 323-7639		
Cartagena Avish F Pleasant Rd 02139.....	617 442-9780		
B Hrd 02134.....	617 361-5253		
Justica 50 Decatur Cha 02129.....	617 241-0152		
Luzmila 124 Harvard Cam 02138.....	617 491-5621		
M 95 Howe Box 02132.....	617 323-9713		
Melvin 503 Green Cam 02139.....	617 576-1061		
361-0380 Carte Nicholas 18 Appleton Boston 02114.....	617 695-6996		
Cartier 0 4 Bedford Box 02133.....	617 338-9219		
628-8248 Carten Thos Jr Sr & Claire 1 Franklin St Mt 02136.....	617 698-6163		
445-5116 Thomas & Kathleen 50 Thompson Ln Mt 02136.....	617 696-6919		
822-2962 Carte A Box 02133	617 329-2257		
427-5712 A Helen	617 442-5230		
569-2698 A 203 Beulah Wy Roxbury 02119.....	617 442-1219		
667-5190 A 203 Massachusetts Av Bos 02115.....	617 266-7153		
569-1412 Adams 361 Centre St Mt 02136.....	617 698-9074		
338-9110 Alice 108 Elmwood Box 02134.....	617 423-0193		
825-9119 Andrew F 42 West St Som 02143.....	617 945-2711		
825-9119 Carte Anne MD	617 625-7623		
296-1593 Carte Anne MD	617 739-1022		
670-2078 B E 18 Graduate Av Mt 02136.....	617 536-6329		
621-9001 Carte Barbara L MD Tufts New England Medical Center Box 02111 Cam.....	617 436-0051		
296-4725 Carte Becky Jo 02134	617 523-4368		
542-1521 Bernard J 122 Goodhue F Rd 02136.....	617 567-9430		
364-5232 Bibbiah 25 Midway Dr 02134.....	617 298-8713		
541-5429 Bill 20 Elmwood Box 02138.....	617 367-9031		
739-2662 Carte Broadcasting Co 58 Park Pl Box 02134.....	617 423-0210		
879-0030 Carte C 200 Commonwealth Av 02135.....	617 225-0200		
436-1511 C 218 Harvard Av East Boston 02128.....	617 762-2118		
569-4119 C 109 Harvard Cam 02138.....	617 491-4822		
869-8782 C & M 43 Bernham Jan 02136.....	617 524-9558		
327-1105 Carter F 34 Hibisc Box 02133.....	617 327-1105		
Faye & Ricky 20 Columbia Av Bos 02136.....	617 437-7331		
Francis S 134 Temple W Av 02132.....	617 323-6781		
Franklin & Anne 705 Mt Auburn Cam 02138.....	617 354-0798		
Fred 42 Hawthorn Elm 02136.....	617 524-3078		
Fred 16 Howland Av Mt 02136.....	617 698-1343		
G & B 8 Vardon Dr 02134.....	617 434-8966		
G T 27 Franklin Av Som 02145.....	617 623-7121		
Gayle 25 Franklin Dr 02134.....	617 823-0322		
Geo S 115 Mass Mt Hill Rd Jan 02138.....	617 522-3215		
George 125 Madison Box 02134.....	617 367-9548		
Carter Hillside Assoc 107 S Street Box 02111.....	617 456-1689		
Carter Harry F 26 Irving St Mt W Av 02132.....	617 325-5465		
Carter Hide Co Inc 100 Franklin St 02114.....	617 542-7987		
Carter Hilary 41 Harvey Cam 02148.....	617 876-2750		
Horace 301 Walnut Av Roxbury 02119.....	617 442-5307		
Howard Jr 28 Nona Ave Box 02118.....	617 445-5552		
J Dan	617 354-2658		
J S 13 Chatham Box 02146.....	617 232-7990		
J S 138 Harvard Box 02146.....	617 730-9483		
J 775 The Pines West Roxbury 02132.....	617 323-5374		
Carter J Jacques MD 1 Brookline Pl Bos 02146.....	617 735-8787		
Carter J M 3410 Columbia Rd S Bos 02136.....	617 464-1040		
Carter J M Ornamental Ironworks 200 Franklin Falls 02134.....	617 436-5353		
Carter J Veal Co 40 Hawthorn Elm 02136.....	617 442-1775		
Carter James 157 Cambridge St Cam 02138.....	617 492-1214		
James 412 Foster Av Roxbury 02132.....	617 739-2193		
James 31 Good Star Rd Cambridge 02141.....	617 876-8841		
Jas L 34 Roslindale Rd Mt 02134.....	617 361-0773		
Janice 134 Adams Rd Newton 02458.....	617 564-0435		
Jeffrey 41 Warren Av Mt 02134.....	617 424-5994		
John 11 Mountfort Bv 02134.....	617 987-2163		
John 307 Summer St 02135.....	617 423-4334		
John 40 Hawthorn Elm 02136.....	617 422-1235		
June O 129 A Summit Av Br 02133.....	617 734-6199		
K 29 Irving St Mt W Av 02132.....	617 265-8456		
K 17 Concord Dr 02127.....	617 282-1593		
Carter Nellie E 323 Main St Mt 02115.....	617 267-6483		
Nicholas S F 115 Randolph Av Mt 02136.....	617 698-6307		
Nick 21 Fyfehill Box 02114.....	617 267-5222		
Nick & Debbi 196 Vermont Rd Newton 02459.....	617 527-0480		
Nicole	617 698-0713		
Norman G 38 Chickawholl Dr 02125.....	617 822-1201		
P 40 Cranston Pl Bos 02135.....	617 437-4754		
P E 501 E South S Bos 02137.....	617 268-8213		
P L 44 Hutchings Box 02131.....	617 427-9170		
P R 91 Boyer Jan 02134.....	617 968-8692		
Paul & Constance 114 Franklin Av W Bos 02132.....	617 325-2036		
Paul E 501 E South S Bos 02137.....	617 268-4546		
Paul M 27 Union St 02135.....	617 787-2115		
Carter Pile Driving Inc 17 Avenue Ct Frankington 02102.....	Wellesley Falls 781.235-0488		
Carter Prudence 40 Franklin Waterlton 02127.....	617 393-3782		
Prudence 40 Franklin Waterlton 02127.....	617 926-7063		
Roginald 106 Brookside Drchester 02122.....	617 541-2843		
Renee & Andrew 10 Walnut Box 02138.....	617 720-3765		
Carter Rice David Baker Books Publishing 163 Main Wilmington 01887 Toll Free-Dial '9 & Then.....	800 638-1671		
Carl Eric Industrial Prod 113 Main Wilmington Toll Free-Dial '9 & Then.....	800 616-7447		
Carl Toll Free-Dial '9 & Then.....	800 648-7447		
Headquarters 113 Main Wilmington 01887 Cam.....	978 988-7447		
Ingala 163 Main Wilmington 01887 40 Franklin St & Thon.....	800 638-1673		
Carter Richard 2079 Commonwealth Av Brighton 02135.....	617 987-0836		
Richard A 97 Mt Vernon Box 02106.....	617 566-7293		
Carter Richard A MD 120 Commonwealth Av 02136.....	617 267-0710		
Carter Richard K 13 Mather S Bos 02127.....	617 268-0448		
Richard 175 Rockdale Av Cam 02141.....	617 864-1535		
Roger 130 St Braughn Box 02131.....	617 424-6148		
Roy 41 Concord Cam 02138.....	617 491-6115		
Royce 18 Sanyday Cha 02129.....	617 241-0418		

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 & Debbi 21 Fynfield Box 02116	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	617 327-0480
17 364-5188	Justica 50 Decatur Cha 02129	617 241-0152	Fred 41 Haverhill Aven 02136	617 524-3078	Nick & Debbi 38 Chickadee Rd 02125	617 822-1203
361-0380	Luzella 174 Harvard Cam 02138	617 491-5621	Fred W. Haverhill Ave 02136	617 698-1343	P E 501 E South St Box 02137	617 267-4213
17 566-4548	M 90 Howe Box 02132	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 Der 02134	617 823-8322	P R 91 Brewer Ave 02138	617 968-8692
17 445-5116	Carte Nicholas 18 Appleton Boston 02114	617 695-6996	Geo S 115 Mount Hill Rd Box 02138	617 522-3215	Paul & Constance 114 Adams Ave W Box 02133	617 325-3034
17 822-2962	Cartagena O 4 Bradford Box 02133	617 338-9219	George 120 Naveson Box 02114	617 367-9548	Paul M 201 E South St Box 02137	617 268-4546
17 427-5712	Carten Thos J Sr & Claire 1 Furlow Ln Mt 02136	617 698-6163	Carter Holiday Assoc 107 S Street Box 02111	617 456-1689	Paul M 27 Crown Rd 02135	617 787-2115
17 569-2698	Carte Thos & Kathleen 50 Thompson Ln Mt 02136	617 696-6919	Carter Hide Co Inc 140 Bunker Hill W Ave 02132	617 325-5465	Paul M 201 E South St Box 02137	617 268-4546
17 667-5190	Carte A A 200 Riverside Av Cambridge 02142	617 492-4174	Carter Hilary 41 Harvey Cam 02148	617 876-2750	Prudence 40 Franklin Waterfront 02172	617 393-3782
17 569-1417	Carte Adams A 200 Riverside Av Cambridge 02142	617 492-4174	Horace 301 Walnut Av Roxbury 02119	617 442-5307	Prudence 40 Franklin Waterfront 02172	617 393-3782
17 338-9117	Carte Alice 301 Centre St Mt 02136	617 698-9074	Howard Jr 28 New One Box 02118	617 445-5532	Reginald 100 Brookview Circle 02123	617 541-2843
17 825-9195	Carte Andrew F 42 West St 02135	617 625-7623	J C 15 Chatham Ave 02144	617 233-7990	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 296-1293	Carte Anne MD 1101 Beacon Ave 02144	617 739-1022	J S 4775 The Pines West Roxbury 02132	617 323-5274	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 670-2078	Carte B E 371 Newbury Boston 02116	617 536-6229	Carter J 1 Crockett Pl Br 02144	617 735-8787	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 621-9001	Carte B E 371 Newbury Boston 02116	617 536-6229	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 296-4725	Carte Barbara L MD Tufts New England Medical Center Box 02111	617 636-0951	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 542-1521	Carte Becky Box 02114	617 523-4368	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 364-5232	Carte Bernard J 300 Cambridge St Cam 02138	617 567-9430	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 541-5649	Carte Bithiah 25 Midway Der 02124	617 298-8713	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 739-2662	Carte Broadcasting Co 50 Park Pl Box 02136	617 423-0210	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 879-0030	Carte Carter C 2000 Gesswally Av Br 02135	617 782-2118	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 436-1511	Carte C 210 Harvard Av East Boston 02128	617 569-1545	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
17 569-4119	Carte C 109 Harvard Cam 02138	617 491-4822	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843
100 869-8782	Carte C & M 41 Northgate Cha 02134	617 524-9558	Carter J M 3410 Columbia Rd S Box 02138	617 464-1040	Renee & Andrew 100 Brookview Circle 02123	617 541-2843



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
17 447-4101	Cartagena Lydia 20 Sweet Briar Rd 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 02136	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 02114	617 695-6996
17 445-5116	Carlton D 4 Halford Box 02138	617 338-9219
17 822-2962	17 Franklin St Mt 02136	617 698-6163
17 427-5712	Carte A Box 02133	617 339-2257
17 569-2698	A 202 Riverside Av Cambridge 02142	617 492-4174
17 667-5190	A M 250 Massachusetts Av 02115	617 266-7153
17 569-1417	Adams 361 Carter St Mt 02136	617 698-9074
17 338-1101	Adams P 42 West St 02134	617 625-7623
17 822-1993	Carte Anne MD 1161 Beacon Bldg 02144	617 739-1022
17 296-1593	Carte Adhena 971 Newbury Boston 02116	617 536-6239
17 670-2078	B E 10 Gladstone Av Mt 02136	617 296-6911
17 621-9001	Carte Barbara L MD Tufts New England Medical Center Box 02111	617 436-0951
17 296-4725	Carte Becky Box 02114	617 523-4368
17 542-1521	Bernard J 301 Ashdown E Mt 02136	617 567-9430
17 364-5232	Bibb 25 Midway Rd 02136	617 298-8713
17 541-5649	Bibb 100 W Newbury St 02116	617 367-9931
17 739-2662	Carte Broadcasting Co 58 Park Pl Box 02116	617 423-0210
17 879-0030	Carte B 33 East Can 02141	617 225-0200
17 541-3948	Carte C 2000 Commonwealth St 02135	617 782-2118
17 436-1511	C 210 Townsend Av East Boston 02128	617 569-1545
17 569-4119	C 109 Harvard Can 02136	617 491-4822
909 569-8782	C & M 41 Northgate Jct 02134	617 524-9558
	Carter J F 157 Cambridge St Can 02136	617 492-1214
	Carter J M 402 Foster Av Andover 02186	617 739-2193
	James 31 East Star Rd Cambridge 02141	617 876-8841
	John 14 Boardwalk Rd Mt 02136	617 361-0773
	Jane 134 Adams Rd Newton 02459	617 564-0435
	John 11 Mansfield St 02134	617 987-2163
	John 207 Summer St 02129	617 423-4334
	John 41 Warren St Mt 02136	617 282-1235
	Jane 129 A Summit Av Br 02133	617 734-6109
	K 29 Franklin St 02134	617 265-4956
	K 17 Concord Street 02123	617 282-1593
	Carter Nellie E 323 Marchant Av Mt 02115	617 267-6483
	Nicholas S F 115 Randolph Av Mt 02136	617 698-5307
	Nick 21 Farwell 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 44 Woodland Pl Mt 02115	617 427-4754
	P E 501 E South St Box 02137	617 268-4213
	P L 44 Hutchings Box 02115	617 427-9170
	P R 91 Boyer Box 02138	617 968-8692
	Paul & Constance 114 Adams Av Mt 02110	617 325-3034
	Paul F 501 E South St Box 02137	617 268-4546
	Paul M 27 Union St 02139	617 787-2115
	Carte Pile Driving Inc 27 Beaver Ct Franklin 02102	Wellesley Tpk-781.235-0488
	Carte Prudence 40 Franklin Waterbury 02172	617 393-3782
	Prudence 40 Franklin Waterbury 02172	617 926-7063
	Reginald 100 Brookside Circle 02123	617 541-2843
	Carte R 30 Walnut St 02118	617 720-3765
	Carte Rice David Building Department 163 Main Wilmington 01887 Toll Free-Dial 7 & Then.....800 638-1671 Call Cartersville 617 313 Main Wilmington Toll Free-Dial 7 & Then.....800 616-7447 Toll Free-Dial 7 & Then.....800 648-7447 Travellers 413 Main Wilmington 01887 Call.....978 988-7447 Ingalls Centre 163 Main Wilmington 01887 Call.....800 638-1673	
	Carte Richard 2075 Carver Av Brighton 02111	617 987-0836
	Richard A 97 W Vernon St 02106	617 566-7293
	Carte Richard A 1200 Commonwealth St 02116	617 267-0710
	Carte Richard K 123 West St Box 02137	617 268-0468
	Robert L 175 Rockwood Av Can 02141	617 864-1535
	Renee & Andrew 130 St Brandy Box 02114	617 491-6115
	Royce 18 Sanderson Cir 02129	617 241-9418



\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**

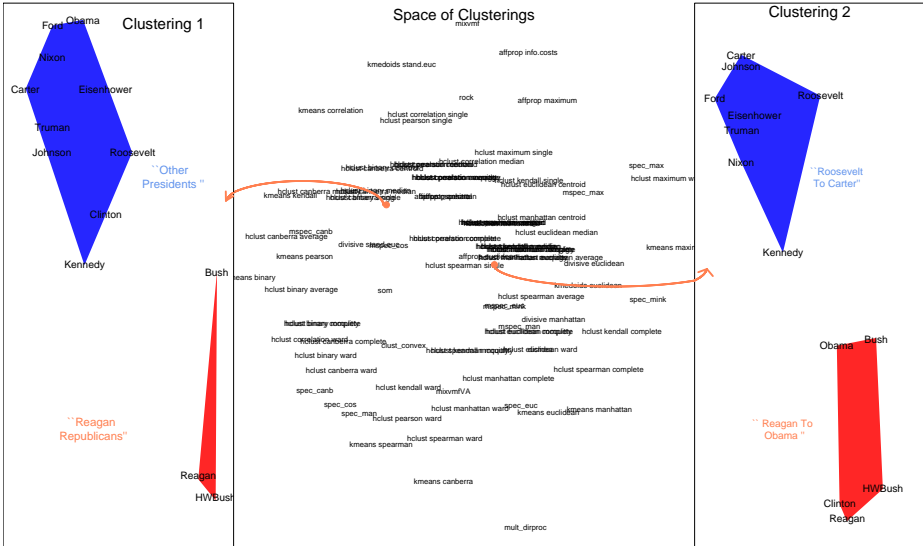
A New Strategy

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)

Many Thousands of Clusterings, Sorted & Organized

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

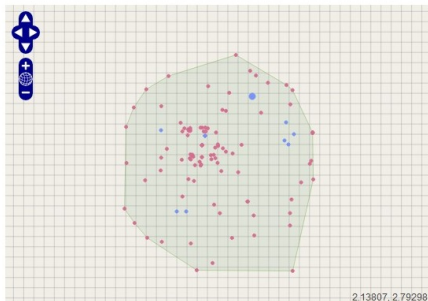


Software Screenshot

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

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

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

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

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

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
 - \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

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
 - \implies Cluster quality evaluation: human judgement of document pairs
- **Experimental Design to Assess Cluster Quality**
 - automated visualization to choose one clustering
 - many pairs of documents
 - for coders: (1) unrelated, (2) loosely related, (3) closely related
 - Quality = mean(within cluster) - mean(between clusters)

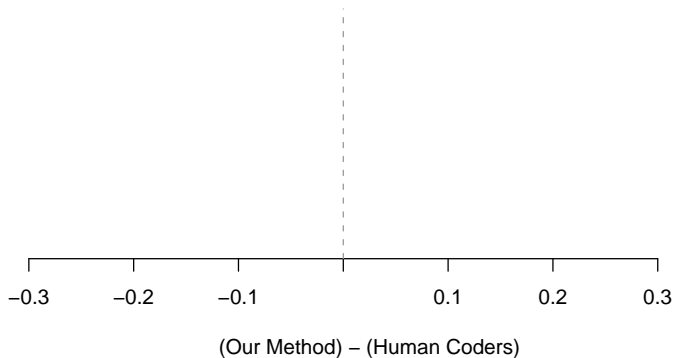
Evaluation 1: Cluster Quality

- **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
 - Quality = mean(within cluster) - mean(between clusters)
 - **Bias results against ourselves by not letting evaluators choose clustering**

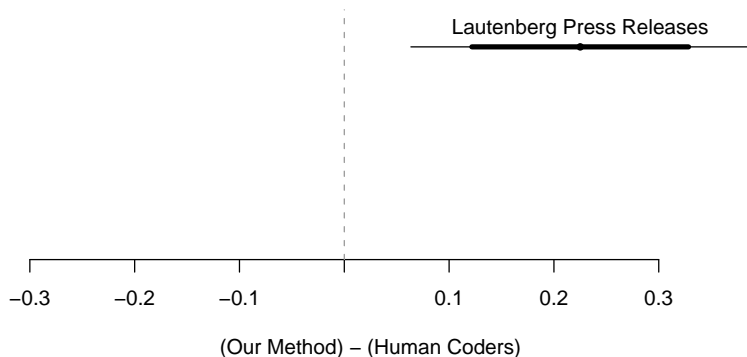
Evaluation 1: Cluster Quality

- **What Are Humans Good For?**
 - They can't: keep many documents & clusters in their head
 - They can: compare two documents at a time
 - \implies Cluster quality evaluation: human judgement of document pairs
- **Experimental Design to Assess Cluster Quality**
 - automated visualization to choose one clustering
 - many pairs of documents
 - for coders: (1) unrelated, (2) loosely related, (3) closely related
 - Quality = mean(within cluster) - mean(between clusters)
 - **Bias results against ourselves by not letting evaluators choose clustering**

Evaluation 1: Cluster Quality

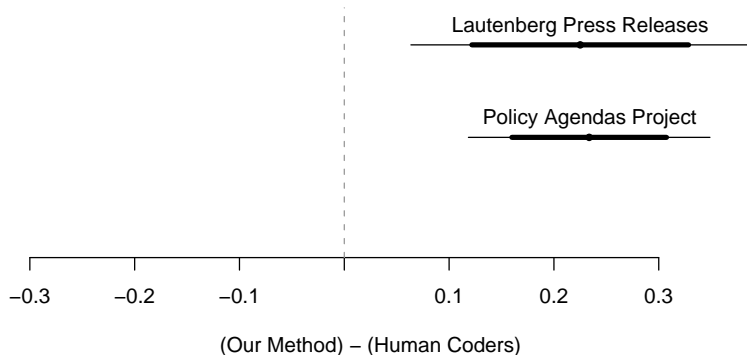


Evaluation 1: Cluster Quality



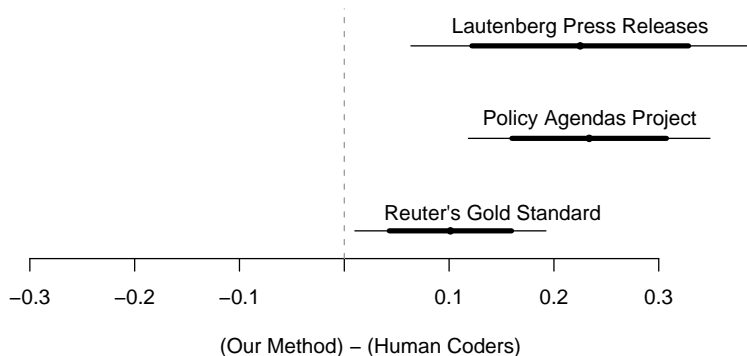
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

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

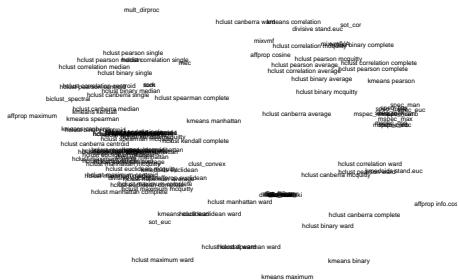
Evaluation 3: What Do Members of Congress Do?

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

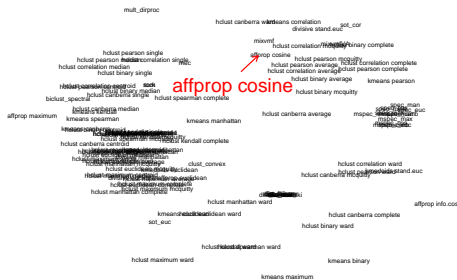
Evaluation 3: What Do Members of Congress Do?

- 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



Example Discovery



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

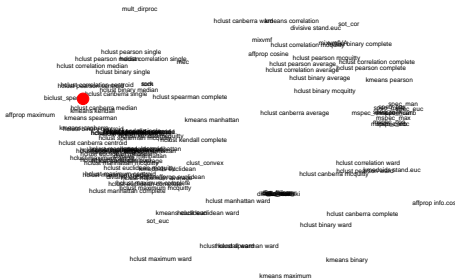
Example Discovery



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



Mixture:

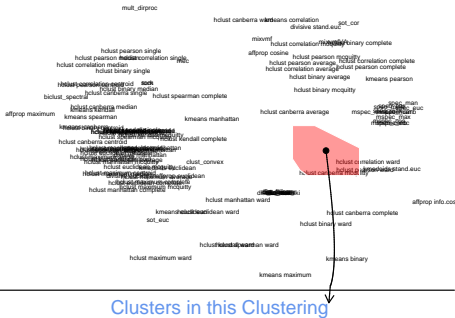
Example Discovery



Mixture:

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

Example Discovery



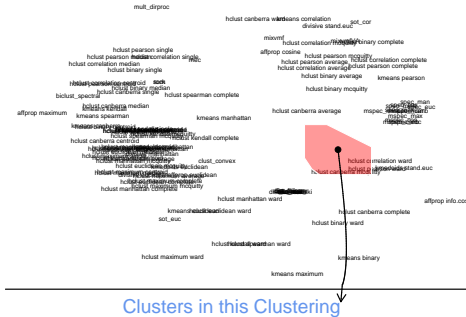
Credit Claiming, Legislation:
“As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period”

Credit Claiming
Pork

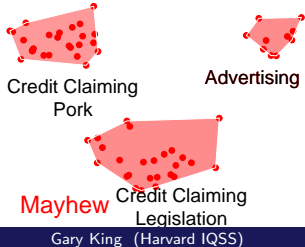


Mayhew
Credit Claiming
Legislation
Gary King (Harvard IQSS)

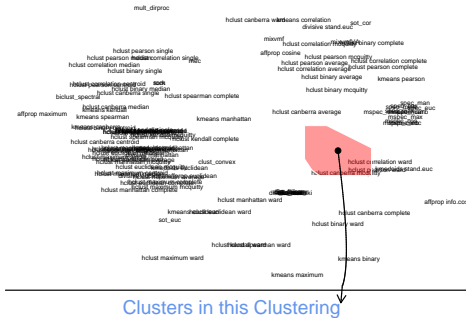
Example Discovery



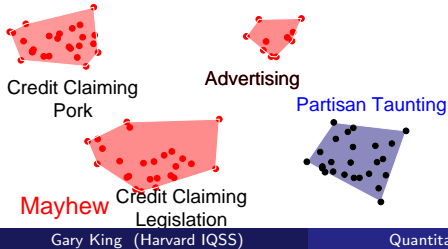
Advertising:
“Senate Adopts
Lautenberg/Menendez Resolution
Honoring Spelling Bee Champion
from New Jersey”



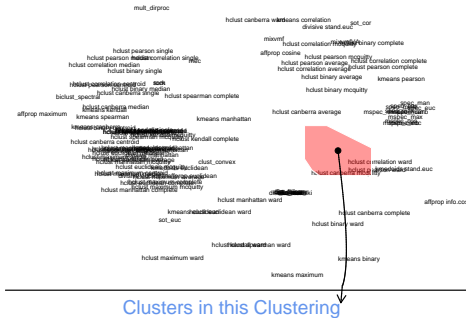
Example Discovery: Partisan Taunting



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



Example Discovery: Partisan Taunting



Credit Claiming
Pork

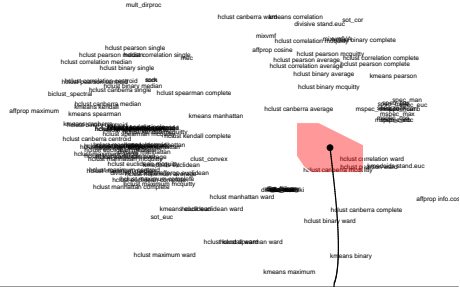
Advertising

Partisan Taunting

Mayhew
Credit Claiming
Legislation
Gary King (Harvard IQSS)

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



Clusters in this Clustering



Credit Claiming
Pork

Advertising

Partisan Taunting

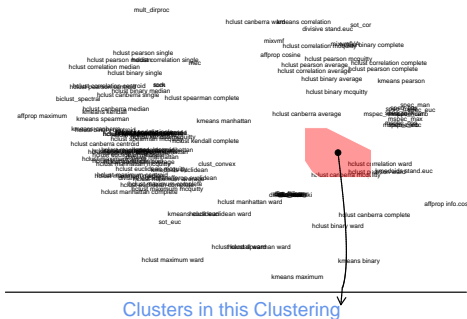


Mayhew Credit Claiming
Legislation

Gary King (Harvard IQSS)

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



Credit Claiming
Pork



Advertising

Partisan Taunting



Mayhew
Credit Claiming
Legislation

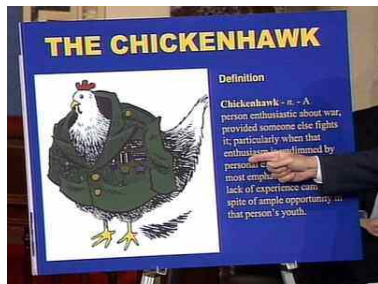


Gary King (Harvard IQSS)

Quantitative Discovery

In Sample Illustration of Partisan Taunting

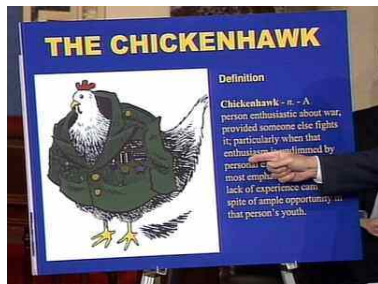
Taunting ruins deliberation



Sen. Lautenberg
on Senate Floor
4/29/04

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

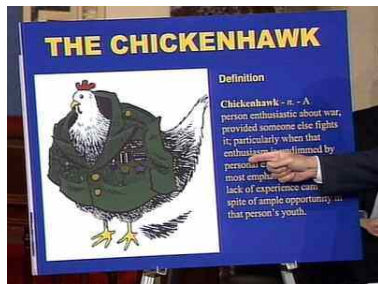
Taunting ruins deliberation



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

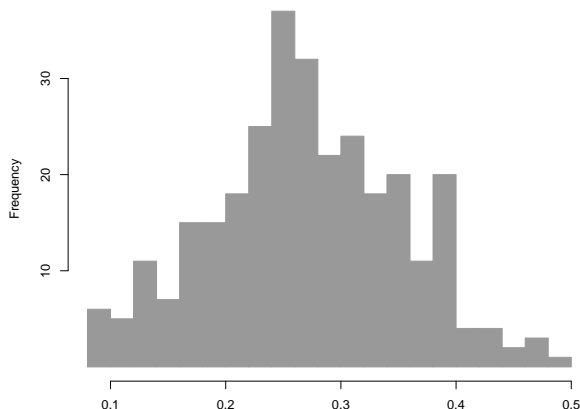
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

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

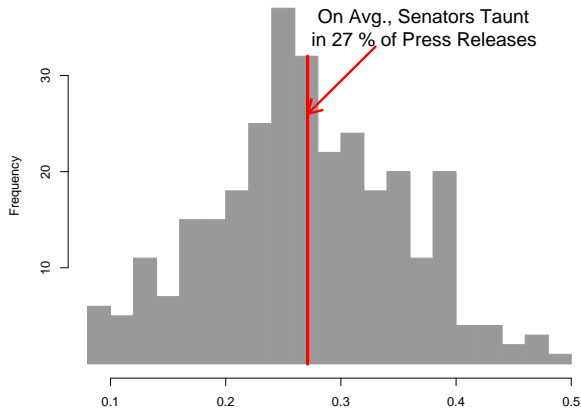
Out of Sample Confirmation of Partisan Taunting

- 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

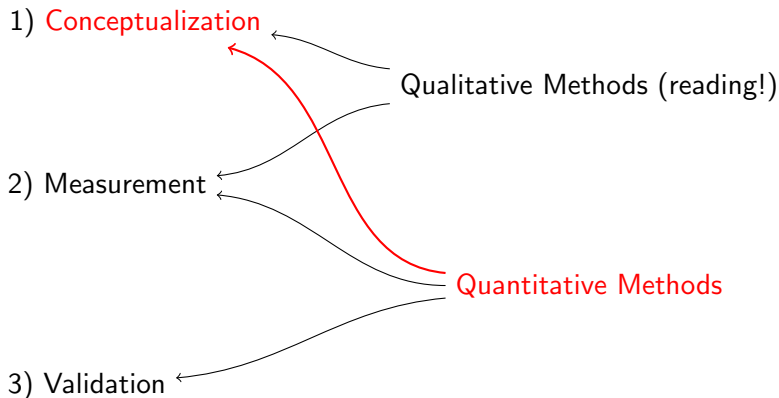


Out of Sample Confirmation of Partisan Taunting

- 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

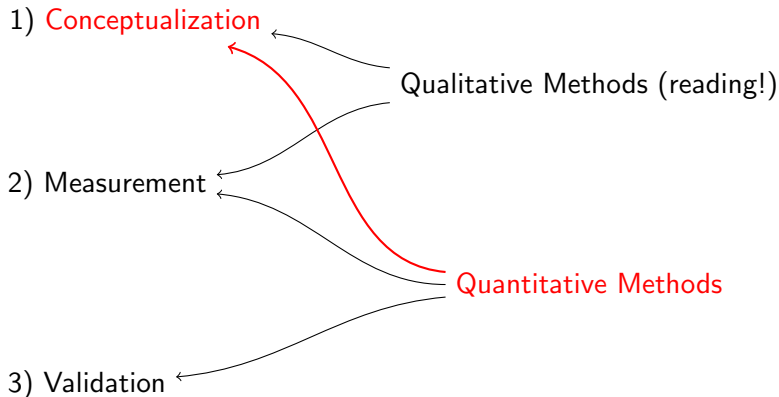


Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

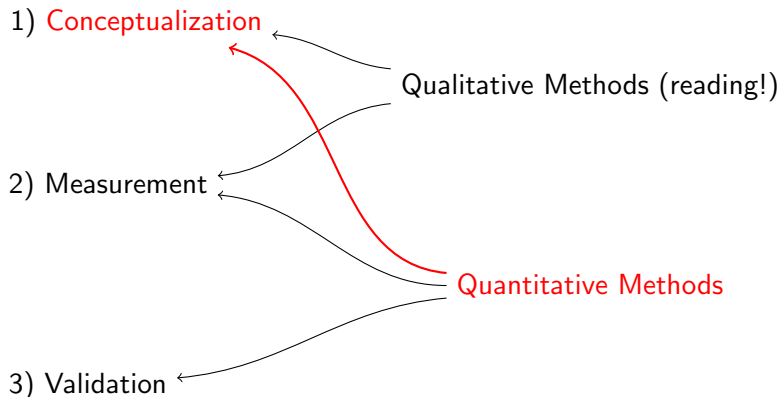
Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

- Few formal methods designed explicitly for conceptualization

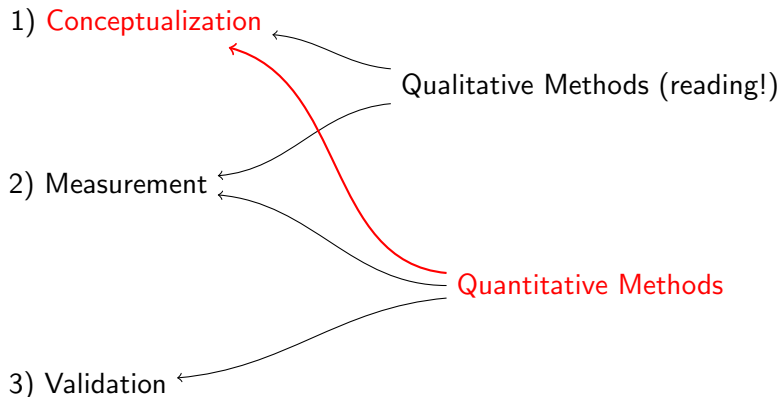
Quantitative Methods for Qualitative Conceptualization



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>