

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: 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	Carter F 34 Hibiscus Bldg 02133..... 617 327-1105	Carter Nellie E 323 Main St Wob 02115..... 617 267-6483	
Cartagena Lydia 28 Sweet Box 02131..... 617 323-7639	Faye & Ricky 207 Columbia Ave Bos 02136..... 617 437-7331	Nicholas S F 115 Randolph Ave Mil 02186..... 617 698-5307	
Cartagena Avish F Pleasant Box 02139..... 617 442-9780	Francis S 134 Yankov W Ave 02132..... 617 323-6781	Nick 21 Farwell Box 02114..... 617 267-5222	
B Hed 02134..... 617 361-5253	Franklin & Anne 201 Mt Auburn Cam 02138..... 617 354-0798	Nick & Debbi 196 Herold Rd Newton 02459..... 617 527-0480	
Jessica 50 Decatur Cha 02129..... 617 241-0152	Fred 40 Hawthorn Elm 02136..... 617 524-3078	Nicole..... 617 698-0713	
Luzmila 124 Harvard Cam 02138..... 617 491-5621	Fred 16 Howland Ave Mil 02136..... 617 698-1343	Norman G 38 Chickawhoh Dr 02125..... 617 822-1201	
M 90 Howe Box 02132..... 617 323-9713	G & B 8 Vardon Bldg 02134..... 617 434-8966	P 40 Cranston Pl Bos 02135..... 617 457-4754	
Melvin 503 Green Cam 02129..... 617 576-1061	G T 27 Franklin Ave Som 02145..... 617 623-7121	P E 501 E South S Bos 02137..... 617 268-8213	
Carl Nicholas 18 Appleton Boston 02114..... 617 695-6996	Gayle 25 Franklin St 02133..... 617 823-0232	P L 44 Hutchings Box 02131..... 617 427-9170	
Carlton 0 4 Bradford Box 02133..... 617 338-0219	Geo S 115 Mass Mt Hill Rd Jam 02138..... 617 522-3215	P R 91 Boyer Ave 02138..... 617 968-8692	
Carten Thos J Sr & Claire 1 Franklin St Mil 02132..... 617 698-6163	George 125 Madison Box 02134..... 617 367-9548	Paul & Constance 114 Freeman St W Bos 02130..... 617 325-2036	
17 445-5116	Carter Holiday Assoc/Am 107 S Street Bos 02111..... 617 456-1689	Paul E 501 E South S Bos 02137..... 617 268-4546	
Thomson & Kullback 50 Thompson Ln Mil 02136..... 617 696-6919	Carter Harry F 26 Irving St W Ave 02132..... 617 325-5465	Paul M 27 Union St 02139..... 617 787-2115	
17 822-2992	Carter Hide Co Inc 160 Boston Ave 02131..... 617 542-7987	Prudence 40 Franklin Waterlton 02122..... 617 926-7063	
17 427-5712	A Weber 617 442-5230	Ronald 106 Bromfield Dorchester 02122..... 617 541-2843	
17 569-2698	Carter Hilary 41 Harvey Cam 02148..... 617 876-2750	Renee & Andrew 10 Walnut Box 02118..... 617 720-3765	
17 667-5190	Horace 301 Walnut St Roxbury 02119..... 617 442-5307	Carter Rice David Bulfinch Business Publishing 163 Main Winsten 01887 Toll Free-Dial '9 & Then..... 800 638-1671	
17 569-1417	Howard Jr 28 Nona Drive Box 02118..... 617 445-5532	Carl Eric Industrial Prod 613 Main Winsten..... 800 616-7447	
17 338-9110	J Dan..... 617 354-2658	Toll Free-Dial '9 & Then..... 800 648-7447	
17 825-1593	J 31 Chatham Ave 02146..... 617 232-7990	Carl..... 978 988-7447	
Carter Anne MD 1161 Beacon Ave 02144..... 617 739-1022	J 538 Harvard Bos 02148..... 617 730-9483	Carl..... 800 638-1673	
Carter J M 1 Ipswich Pl Bos 02146..... 617 335-5374	Jacques J 1 Ipswich Pl Bos 02146..... 617 735-8787	Carl..... 978 988-7447	
17 670-2078	Carter J D 3410 Columbia Rd S Bos 02137..... 617 464-1040	Carl..... 800 648-7447	
17 621-9001	Carter J M Ornamental Ironworks 200 Franklin Falls 02134..... 617 876-5353	Carl..... 978 988-7447	
17 296-4725	Carter J Neal Co 40 Hawthorn Elm 02136..... 617 442-1775	Carl..... 800 638-1673	
17 542-1521	Carter James 157 Cambridge St Cam 02138..... 617 492-1214	Carl..... 978 988-7447	
17 364-5232	James 312 Foster Ave Roxbury 02119..... 617 739-2193	Carl..... 800 638-1673	
17 541-5649	Bernad J 132 Good Star Rd Cambridge 02141..... 617 876-8841	Carl..... 978 988-7447	
17 739-2662	Carter Broadcasting Co 34 Broad St Mil 02134..... 617 361-0773	Carl..... 800 648-7447	
17 879-0030	Carter & Baines Consultants Inc 20 Park Pl Bos 02114..... 617 423-0210	Carl..... 978 988-7447	
17 436-1511	Carter C 200 Commonwealth Ave 02135..... 617 782-2118	Carl..... 800 638-1673	
17 569-4119	C 210 Harvard Ave East Boston 02128..... 617 569-1545	Carl..... 978 988-7447	
800 569-8782	C 109 Harvard Cam 02138..... 617 491-8522	Carl..... 800 638-1673	
	C 201 Irving St 02133..... 617 423-4392	Carl..... 978 988-7447	
	C & M 43 Bernham Jan 02138..... 617 524-9558	Carl..... 800 638-1673	

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 Box 01386	617 698-5307
100 257-9961	Cartagena Avish F Beach Rd 02139	617 442-9780	Francis S. 134 Temple W Ave 01312	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	617 527-0480
17 364-5188	Justica 50 Decatur Cha 02129	617 241-0152	Fred 41 Haverhill Aven 02136	617 524-3078	Nick & Constance 38 Chickadee Dr 02125	617 822-1203
361-0380	Luzella 174 Harvard Cam 02138	617 491-5621	Fred W. Hovell Ave 02136	617 698-1343	P E 501 E South S Box 02137	617 268-4213
17 566-4548	M 95 Howe Box 02132	617 323-9713	G & B. 8 Vardon Ave 02134	617 436-8906	P L 44 Hatching Box 02131	617 427-9170
17 628-8248	Melvin 503 Green Cam 02139	617 576-1061	Gayle 25 Franklin St 02134	617 823-0322	P R 91 Brewer Ave 02138	617 968-8692
17 822-2962	Carte Nicholas 18 Appleton Boston 02114	617 695-6996	Geo S 115 Mount Hill Nat 02138	617 522-3215	Paul & Constance 114 Adams Ave W Box 02133	617 325-3034
17 427-5712	Cartagena O 4 Bradford Box 02133	617 338-0219	George 125 Boston Ave 02134	617 367-9548	Paul M 501 E South S Box 02137	617 268-4546
17 569-2698	Carten Thos J Sr & Claire 1 Furlow St Mt 02136	617 698-6163	Carter Holiday Assoc 107 S Street Box 02111	617 456-1689	Paul M 27 Union St 02135	617 787-2115
17 667-5190	Carte Thos & Kathleen 50 Thompson Ln Mt 02136	617 696-6919	Carter Harry F 30 Burns Rd W Box 02132	617 325-5465	Prudence 40 Franklin Waterbury 02172	617 393-3782
17 569-1417	Carte A A 202 Riverside Av Cambridge 02142	617 492-4174	Carter Hide Co Inc 161 Sycamore St 02148	617 542-7987	Prudence 40 Franklin Waterbury 02172	617 926-7063
17 338-9110	A M 255 Massachusetts Av 02115	617 266-7153	Carter Hilary 41 Harvey Cam 02148	617 876-2750	Ronald 100 Brookwood Circle 02124	617 541-2843
17 825-9195	Adams 301 Carter St Mt 02136	617 698-9074	Horace 301 Walnut Av Roxbury 02119	617 442-5307	Renee & Andrew 100 Brookwood Circle 02124	617 541-2843
17 296-1293	Alice 108 Elmwood Hs 02134	617 453-0193	Howard J 301 New One Box 02118	617 445-5552	Rice 301 Walnut St 02138	617 720-3765
17 670-2078	Alice 40 Market Cambridge 02139	617 945-2711	J 45 Canton St 02144	617 232-7990	Rice 301 Walnut St 02138	617 720-3765
17 621-9001	Andrew F 42 Mt St 02135	617 625-7623	J & Susan 4 775 The Pines West Roxbury 02132	617 323-5274	Rice 301 Walnut St 02138	617 720-3765
17 296-4725	Carter Anne MD 1101 Beacon Hs 02144	617 739-1022	Jacobson H B 02144	617 735-8787	Rice 301 Walnut St 02138	617 720-3765
17 542-1521	Carter Adhena 771 Newbury Boston 02116	617 536-6229	Carter J 3410 Columbia Rd S Box 02136	617 464-1040	Rice 301 Walnut St 02138	617 720-3765
17 364-5232	Carter Barbara L MD 100 State St Mt 02136	617 296-6911	Carter J M Ornamental Ironworks 100 State St Mt 02136	617 436-5353	Rice 301 Walnut St 02138	617 720-3765
17 739-2662	Carter Broadcasting Co 50 Park Pl Box 02134	617 423-0210	Carter J Neal Co 40 Newmarket St 02138	617 442-1775	Rice 301 Walnut St 02138	617 720-3765
17 879-0030	Carter C 2000 Cambridge St 73 East C Cam 02141	617 225-0200	Carter James 157 Cambridge St Cam 02138	617 492-1214	Rice 301 Walnut St 02138	617 720-3765
17 541-3948	Carter C 2000 Cambridge St 73 East C Cam 02141	617 225-0200	James 157 Cambridge St Cam 02138	617 492-1214	Rice 301 Walnut St 02138	617 720-3765
17 569-4119	Carter C 2000 Cambridge St 73 East C Cam 02141	617 225-0200	James 157 Cambridge St Cam 02138	617 492-1214	Rice 301 Walnut St 02138	617 720-3765
17 809-8782	Carter C 2000 Cambridge St 73 East C Cam 02141	617 225-0200	James 157 Cambridge St Cam 02138	617 492-1214	Rice 301 Walnut St 02138	617 720-3765



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 28 Sweet Briar Rd 02331	617 323-7639
90 257-9961	Cartagena Avish F Beach Rd 02319	617 442-9780
	B Had 02334	617 361-5253
17 566-1282	Justicia 50 Decatur Cha 02329	617 241-0152
17 364-5188	Luzmila 124 Harvard Can 02334	617 491-5621
	M 95 Howe Box 02334	617 323-9713
361-0380	Melvin 503 Green Can 02329	617 576-1061
17 566-4548	Carte Nicholas 18 Appleton Boston 02314	617 695-6996
	Cartagena O 4 Bradford Box 02334	617 338-0219
17 628-8248	Carten Thos J Sr & Claire 1 Furlow Ln Mt 02336	617 698-6163
17 445-5116	Thos & Kathleen 50 Thompson Ln Mt 02336	617 696-6919
17 822-2962	Carter A Box 02334	617 259-2257
17 427-5712	A Heber A 200 Riverside Av Cambridge 02328	617 442-5230
17 569-2698	A 20 Riverside Av Cambridge 02328	617 442-5230
17 667-5190	A M 250 Massachusetts Av 02311	617 442-1219
	Adams 301 Carter St Mt 02336	617 698-7074
17 569-1417	Allice 200 Elmwood Pl 02334	617 423-0193
17 338-1107	Allice 40 Market Cambridge 02339	617 945-2711
	Andrew F 42 West St 02334	617 625-7623
17 822-1993	Carter Anne MD 1101 Beacon Bldg 02444	617 739-1022
17 296-1193	Carter Atlanta 971 Newbury Boston 02216	617 536-6229
17 670-2078	B E 100 Graduate Av Mt 02336	617 296-6911
17 621-9001	Carter Barbara L MD Tufts New England Medical Center Box 02311	617 636-0951
17 296-4725	Carter Becky 02314	617 523-4368
17 542-1521	Bernard J 3000 Ashburn E Rd 02336	617 567-3430
17 364-5232	Bibb 25 Midway Rd 02334	617 298-8713
17 541-5649	Bibb 25 Midway Rd 02334	617 367-9931
17 739-2662	Carter Broadcasting Co 50 Park Pl Box 02314	617 423-0210
	Carter Business Consultants Inc 73 East C Can 02341	617 225-0200
17 879-0030	Carter C 2000 Cambridge St 02335	617 782-2118
17 541-3948	C 210 Townsend Av East Boston 02338	617 569-1545
17 436-1511	C 109 Harvard Can 02336	617 491-4822
17 569-4119	C 109 Harvard Can 02336	617 491-4822
	C & M 41 Northgate Jct 02334	617 524-9558
900 569-8782	C 41 Northgate Jct 02334	617 524-9558
	Carter F J 514 Hickox Box 02334	617 327-1105
	Faye & Ricky 20 Columbia Av Box 02334	617 437-7331
	Francis S 134 Temple W Av 02334	617 323-6781
	Franklin & Anne 705 Mt Auburn Can 02336	617 354-0798
	Fred 41 Howard Av 02336	617 524-3078
	Fred 76 Howley Av Mt 02336	617 698-1343
	G & B 8 Yorker Box 02334	617 436-8906
	G T 27 Fossil Av Mt 02345	617 623-7121
	Gayle 25 Franklin St 02334	617 823-8322
	Geo S 115 Mount Mt Jct Box 02336	617 522-3215
	George 52 Hudson Box 02314	617 367-9548
	Carter Hillside Assoc 107 S Street Box 02311	617 456-1689
	Carter Harry F 30 Bayview Rd W Av 02334	617 325-5465
	Carter Hide Co Inc 140 Boston St 02334	617 542-7987
	Carter Hilary 41 Harvey Can 02348	617 876-2750
	Horace 301 Walnut Av Rosbury 02319	617 442-5307
	Howard Jr 28 New One Box 02318	617 445-5552
	J Can 15 Chatham St 02444	617 232-7990
	J 538 Harvard St 02444	617 730-9483
	J 775 The Pine Way Westbury 02334	617 323-5374
	Carter J Jacques MD 1 Brockton Pl Box 02444	617 735-8787
	Carter J M 3410 Columbia Rd S Box 02337	617 464-1040
	Carter J M Ornamental Ironworks 200 Franklin Falls 02334	617 436-5353
	Carter J Veal Co 40 Newbury St 02318	617 442-1775
	Carter James 1573 Cambridge St Can 02336	617 492-1214
	James 422 Foster Av Rosbury 02318	617 739-2193
	James 31 East Star Rd Cambridge 02318	617 876-8841
	J 34 Howley Rd Mt 02336	617 361-0773
	Jane 14 Adams Rd Newton 02458	617 564-0435
	Jenny 1200 Cambridge St Mt 02336	617 426-9094
	John 11 Mansfield St 02334	617 987-2163
	John 207 Summer St 02334	617 423-4334
	John 40 Howard Av 02336	617 282-1235
	James O 129 A Summit Av 02334	617 734-6109
	J 29 Franklin St 02334	617 265-4856
	K 17 Concord Road 02323	617 282-1593
	Carter Nellie E 323 Marchant Av Box 02315	617 267-6483
	Nicholas S F 115 Randolph Av Mt 02336	617 698-5307
	Nick 21 Furlow Box 02314	617 267-5222
	Nick & Debbi 136 Hermit Rd Newton 02459	617 527-0480
	Norman G 38 Chickadee Dr 02326	617 822-1203
	P 41 Eastwood Pl Box 02315	617 427-4754
	P E 501 E South S Box 02337	617 268-4213
	P L 44 Hutchings Box 02315	617 427-9170
	P R 91 Boyer Can 02334	617 968-8692
	Paul & Constance 114 Adams Av W Mt 02333	617 325-3034
	Paul F 501 E South S Box 02337	617 268-4546
	Paul M 27 Union St 02319	617 787-2115
	Carter Pile Driving Inc 27 Beaver Ct Frankston 02302	Wellesley Tpk 781.235-0488
	Carter Prudence 40 Franklin Waterbury 02327	617 393-3782
	Prudence 40 Franklin Waterbury 02327	617 926-7063
	Reginald 100 Brookside Chester 02324	617 541-2843
	Reed & Andrew 30 Walnut Box 02318	617 720-3765
	Carter Rice David Building Division Publishing 163 Main Wilmington 01887 Toll Free 800 7 1 3 Then.....800 638-1671 Toll Free 800 7 1 3 Then.....800 619-7447 Toll Free 800 7 1 3 Then.....800 648-7447 Toll Free 800 7 1 3 Then.....978 988-7447 Ingalls Centre 163 Main Wilmington 01887 800 638-1673	
	Carter Richard 2079 Carleton Av Brighton 02321	617 987-0836
	Carter Richard A MD 1747 Waverley St 02336	617 566-7293
	Carter Richard A 1200 Cambridge St Mt 02336	617 267-0710
	Carter Richard K 123 Merwin St Box 02337	617 268-0468
	Robert L 175 Rockwood Av Can 02341	617 864-1535
	Royce 130 St Brantley Box 02311	617 424-6148
	Royce & Andrew 1800 Broad St 02329	617 491-6115
	Royce 18 Sandway Cha 02329	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**

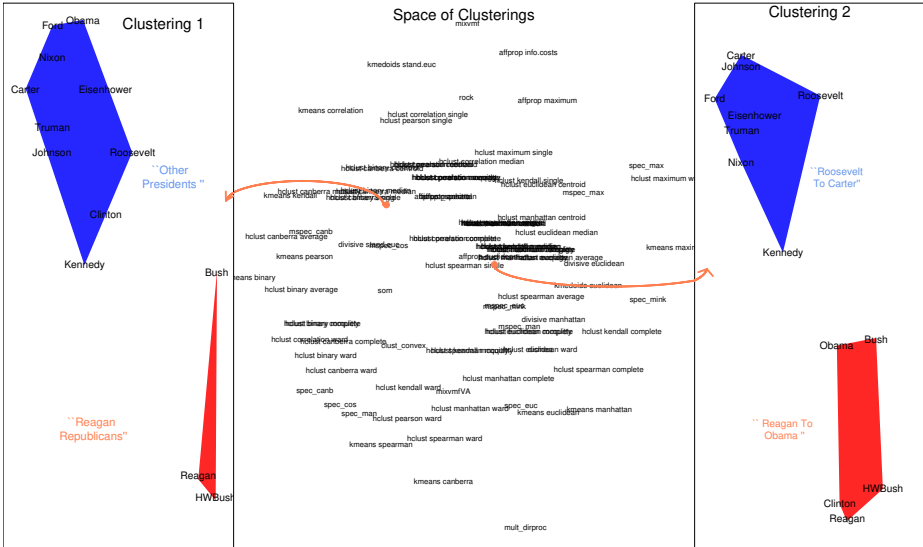
A New Strategy

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, . . .

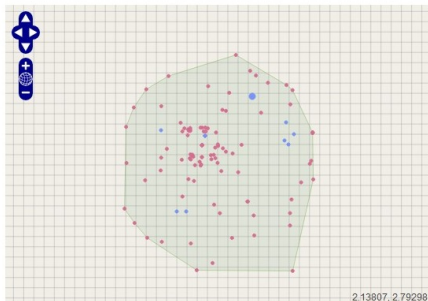


Software Screenshot

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)

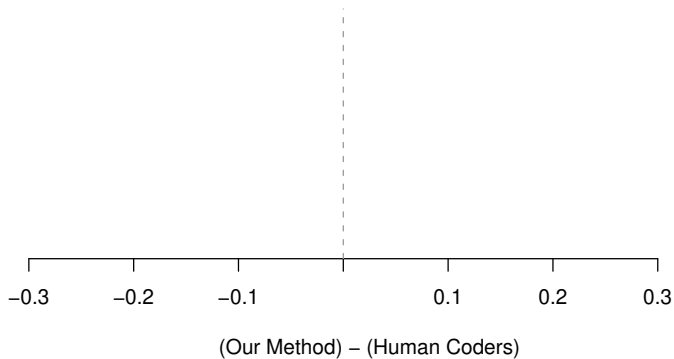
Evaluation 1: Cluster Quality

- **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**

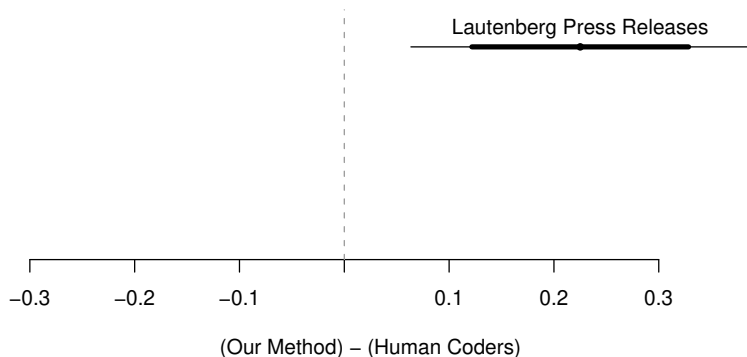
Evaluation 1: Cluster Quality

- **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

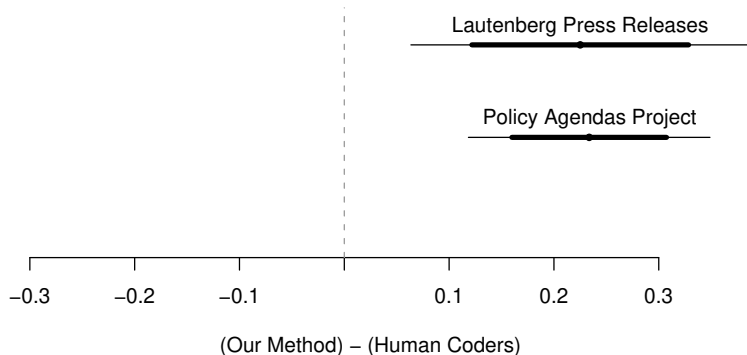


Evaluation 1: Cluster Quality



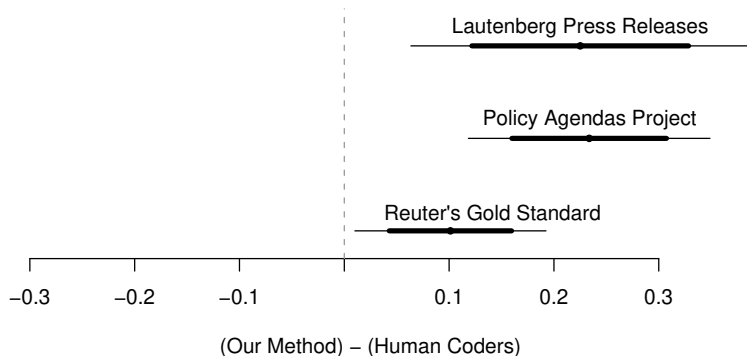
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

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

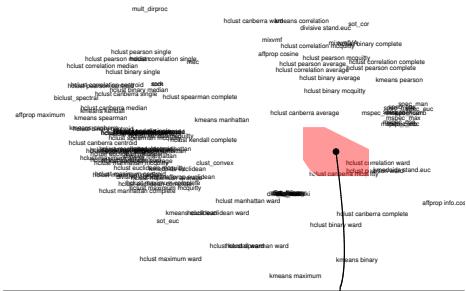
Evaluation 3: What Do Members of Congress Do?

- 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



Clusters in this Clustering



Credit Claiming
Pork

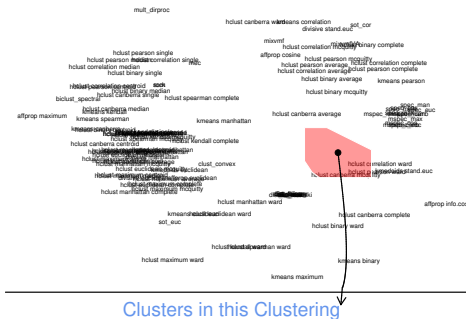
Advertising



Mayhew
Credit Claiming
Legislation
Gary King (Harvard IQSS)

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

Example Discovery: Partisan Taunting



Credit Claiming
Pork

Advertising

Partisan Taunting

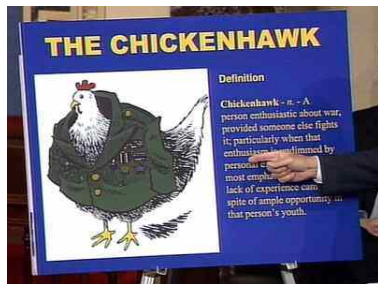
Mayhew
Credit Claiming
Legislation

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’”

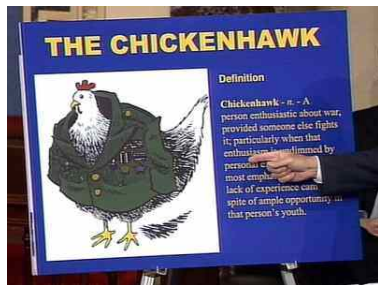
Taunting ruins deliberation



Sen. Lautenberg
on Senate Floor
4/29/04

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

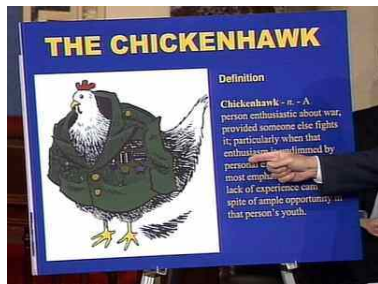
Taunting ruins deliberation



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

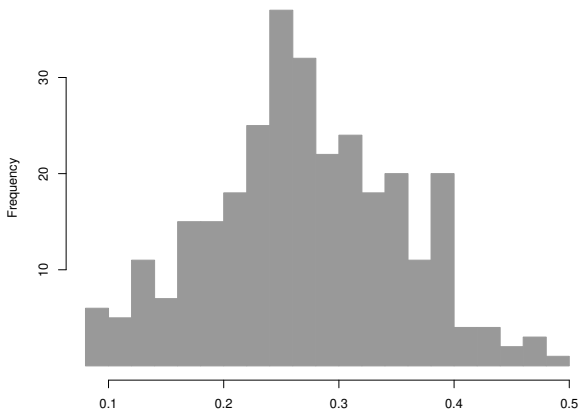
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

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

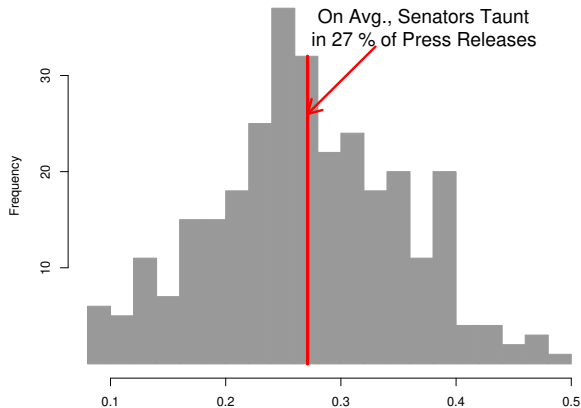
Out of Sample Confirmation of Partisan Taunting

- 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

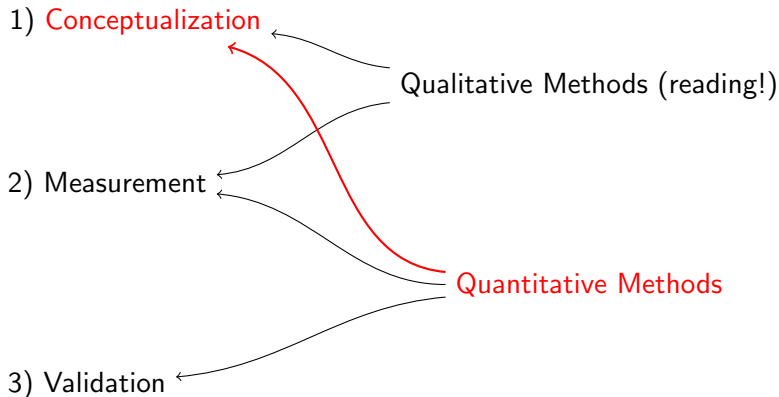


Out of Sample Confirmation of Partisan Taunting

- 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

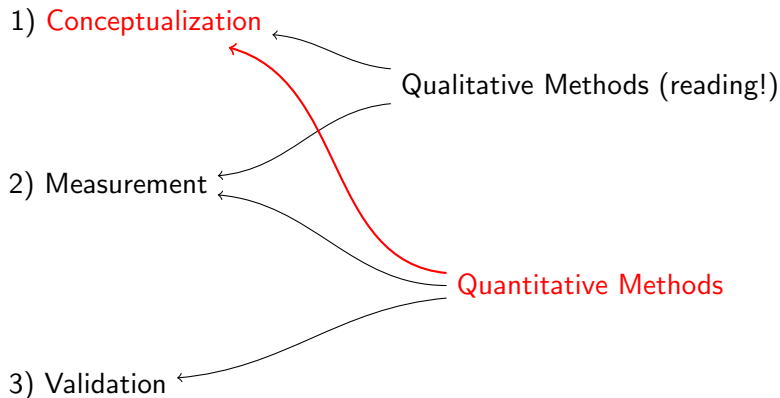


Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

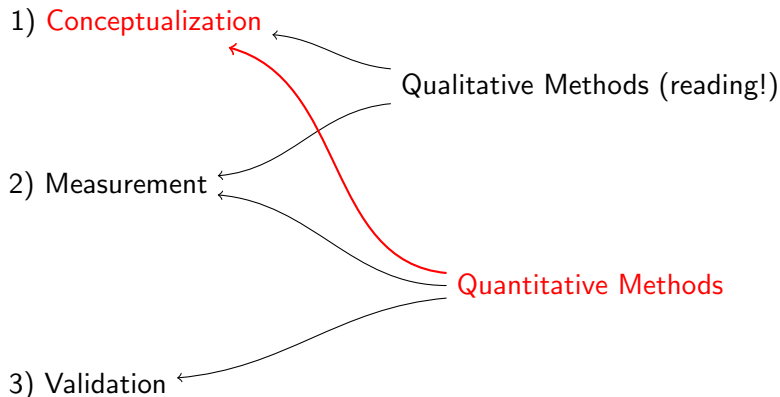
Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

- Few formal methods designed explicitly for conceptualization

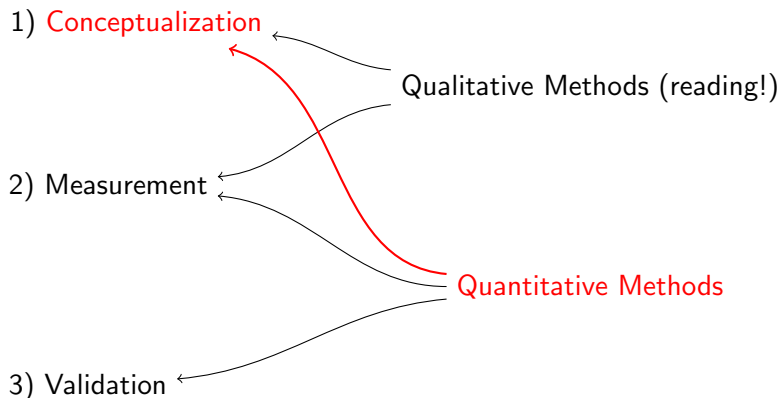
Quantitative Methods for Qualitative Conceptualization



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>