

# Quantitative Discovery of Qualitative Information: A General Purpose Document Clustering Methodology

Gary King

Institute for Quantitative Social Science  
Harvard University

Talk on 2/5/2010 for All-hands IQSS Staff Meeting

Joint work with Justin Grimmer, Harvard University

# Some context for related technology

- <http://ow.ly/14hDU>
- <http://ow.ly/14h36>

# A Method for Conceptualization

- Systematic method for computer-assisted conceptualization from text

# A Method for Conceptualization

- Systematic method for computer-assisted conceptualization from text
- 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 Conceptualization

- **Systematic method for computer-assisted conceptualization from text**
- 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).
- We focus on **Cluster Analysis**: simultaneously 1) invent categories and 2) assign documents to categories

# A Method for Conceptualization

- **Systematic method for computer-assisted conceptualization from text**
- 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).
- We focus on **Cluster Analysis**: simultaneously 1) invent categories and 2) assign documents to categories
- (We focus on texts, our methods apply more broadly)

# Why Johnny Can't Classify (Optimally)

# Why Johnny Can't Classify (Optimally)

- Clustering seems easy; its not!



# Why Johnny Can't Classify (Optimally)

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

# Why Johnny Can't Classify (Optimally)

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

# Why Johnny Can't Classify (Optimally)

- 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)

# Why Johnny Can't Classify (Optimally)

- 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)
- $Bell(5) = 52$

# Why Johnny Can't Classify (Optimally)

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

# Why Johnny Can't Classify (Optimally)

- 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

# Why Johnny Can't Classify (Optimally)

- 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!

# Why Johnny Can't Classify (Optimally)

- 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!
- Its no surprise that automated algorithms can help, but which algorithms?



# Why HAL Can't Classify Either

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **No free lunch theorem**: every possible clustering method performs equally well on average over all possible substantive applications

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **No free lunch theorem**: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **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,...

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **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

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **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**



# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **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**

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **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

# Why HAL Can't Classify Either

- Large quantitative literature on **cluster analysis**
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - **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 in cluster analysis literature: no way to know which method will work ex ante**

# If Ex Ante doesn't work, try Ex Post

# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)

# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model

# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model
- We do it **ex post** (by qualitative choice). For example:

# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model
- We do it **ex post** (by qualitative choice). For example:
  - Create long list of clusterings; choose the best



# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model
- We do it **ex post** (by qualitative choice). For example:
  - Create long list of clusterings; choose the best
  - Too hard for mere humans!

# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model
- We do it **ex post** (by qualitative choice). For example:
  - Create long list of clusterings; choose the best
  - Too hard for mere humans!
  - An **organized** list will make the search possible

# If Ex Ante doesn't work, try Ex Post

- **Methods and substance must be connected** (no free lunch theorem)
- The usual approach fails: hard to do it by understanding the model
- We do it **ex post** (by qualitative choice). For example:
  - Create long list of clusterings; choose the best
  - Too hard for mere humans!
  - An **organized** list will make the search possible
  - E.g.,: consider two clusterings that differ only because one document (of many) moves from category 5 to 6

# 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
<b>Cartage New England Inc</b> 28 Allen St Ipswich 01938.....	978 356-9960		
<b>Cartagena Lydia</b> 38 Sweet St 02131.....	617 323-7639		
<b>Cartagena Aveth</b> F Blandish Box 02139.....	617 442-9780		
B Had 02136.....	617 361-5253		
<b>Justicia</b> 50 Decatur Cha 02129.....	617 241-0152		
<b>Lucilla</b> 124 Harvard Cam 02138.....	617 491-5621		
M 95 Howe Box 02133.....	617 323-9713		
<b>Melvin</b> 501 Green Cam 02139.....	617 576-1061		
<b>Carte Nicholas</b> 38 Appleton Boston 02114.....	617 695-6996		
<b>Cartegena O</b> 44 Bedford Box 02138.....	617 338-9219		
<b>Carten Thos J Sr &amp; Claire</b> 1 Ivesville Rd MA 02138.....	617 698-6163		
<b>Thomas &amp; Kathleen</b> 50 Thompson Ln MA 02136.....	617 696-6919		
<b>Carter A Box 02133.....</b>	617 359-2257		
<b>A Huber</b> A 31 Bethune Wy Roxbury 02119.....	617 442-5230		
A 200 Putnam Av Cambridge 02142.....	617 492-4174		
A M 255 Murchieson Av Box 02113.....	617 266-7153		
<b>Adams</b> 381 Centre St MA 02138.....	617 698-9074		
<b>Alice</b> 108 Elmwood Box 02133.....	617 425-0193		
<b>Alice</b> 40 Market Cambridge 02139.....	617 945-2711		
<b>Andrew F</b> 42 Wal St Som 02143.....	617 625-7635		
<b>Carter Anne MD</b> 1161 Beacon St 02144.....	617 739-1022		
<b>Carter Adhans</b> 272 Newbury Boston 02134.....	617 536-6329		
<b>B E C</b> 48 Graduate Av West 02154.....	617 296-6911		
<b>Carter Barbara L MD</b> Tufts New England Medical Center Box 02111 Cam.....	617 636-9051		
<b>Carter Becky</b> Box 02134.....	617 523-4368		
<b>Carter Bernad J</b> 132 Cambridge F Box 02136.....	617 567-3430		
<b>Bithiah</b> 25 Melrose Dr 02134.....	617 299-8713		
<b>Bithiah</b> 25 Melrose Dr 02134.....	617 367-9931		
<b>Carter Broadcasting Co</b> 26 Park Plt Box 02134.....	617 423-9210		
<b>Carter &amp; Burgess Consultants Inc</b> 73 East St Cam 02141.....	617 225-0200		
<b>Carter C</b> 2000 Cambridge Av Box 02135.....	617 782-2118		
<b>C</b> 219 Concord Av East Boston 02128.....	617 569-1545		
<b>C</b> 109 Harvard Cam 02138.....	617 491-4822		
<b>C</b> 108 Elmwood Box 02133.....	617 425-0193		
<b>C &amp; M</b> 43 Burroughs Jan 02138.....	617 524-9558		
<b>Carter F</b> 24 Hillock Box 02131.....	617 327-1105		
<b>Faye &amp; Ricky</b> 207 Cambridge Av Box 02136.....	617 437-7331		
<b>Francis S</b> 134 Temple W Av Box 02132.....	617 323-6781		
<b>Franklin &amp; Anne</b> 251 Mt Auburn Cam 02138.....	617 354-0798		
<b>Fred</b> 42 Hawthorn Jan 02136.....	617 524-3078		
<b>Fred</b> 96 Harvard Av MA 02138.....	617 698-1343		
<b>G &amp; R</b> 8 Harvard Der 02134.....	617 436-8906		
<b>G T</b> 27 Fyfield Av Som 02145.....	617 623-7121		
<b>Gayle</b> 25 Providence Der 02124.....	617 825-0322		
<b>Geo S</b> 115 Mount Hill Rd Jan 02138.....	617 522-3215		
<b>George</b> 125 Neponset Box 02114.....	617 367-9548		
<b>Carter Halliday Associate</b> 107 S Street Box 02111.....	617 456-1689		
<b>Carter Harry F</b> 26 Baring St Rd W Av Box 02132.....	617 325-5465		
<b>Carter Hide Co Inc</b> 144 Somers Box 02113.....	617 542-7987		
<b>Carter Hilary</b> 41 Harvey Cam 02140.....	617 876-2750		
<b>Horace</b> 361 Walnut Av Roxbury 02119.....	617 442-5307		
<b>Howard Jr</b> 28 Neponset Box 02116.....	617 445-5552		
<b>J Cam</b> 41 Chatham Box 02144.....	617 232-7990		
<b>J S</b> 118 Harvard Box 02144.....	617 730-9483		
<b>J 775</b> The Younger Wy Roxbury 02132.....	617 323-5374		
<b>Carter J Jacques MD</b> 1 Broadview Plt Box 02144.....	617 735-8787		
<b>Carter J M</b> 3410 Columbia St Box 02137.....	617 464-1040		
<b>Carter J M Ornamental Ironworks</b> Cambridge Falls 02137.....	617 436-5353		
<b>Carter J Neal Co</b> 40 Newmarket St 02138.....	617 442-1775		
<b>Carter James</b> 1573 Cambridge St Cam 02136.....	617 492-1214		
<b>James</b> 32 Fisher Av Roxbury 02130.....	617 739-2193		
<b>James</b> 31 Gold Star Rd Cambridge 02140.....	617 876-8841		
<b>Jan L</b> 34 Roslindale Rd MA 02134.....	617 361-0773		
<b>Janice</b> 14 Adams Rd Newton 02458.....	617 564-0435		
<b>Jeffrey</b> 40 Warren Av Box 02134.....	617 426-5994		
<b>John</b> 11 Mansfield Rd 02134.....	617 987-2163		
<b>John</b> 207 Summer Box 02133.....	617 423-4334		
<b>John</b> 40 Westwood Rd 02125.....	617 282-1235		
<b>June O</b> 129 A Summit Av Box 02138.....	617 734-6109		
<b>K</b> 280 University Av 02134.....	617 265-9456		
<b>K</b> 17 Exford Dorchester 02122.....	617 282-1593		
<b>Carter Nellie E</b> 323 Marshfield Av Box 02135.....	617 267-6483		
<b>Nicholas S F</b> 115 Randolph Av MA 02136.....	617 698-5307		
<b>Nick</b> 21 Fairfield Box 02134.....	617 267-5222		
<b>Nick &amp; Debbi</b> 156 Vernick Rd Newton 02459.....	617 527-0480		
<b>Nicole</b> 38 Chickadee Rd Der 02125.....	617 822-1203		
<b>P</b> 40 Cranston Plt Box 02133.....	617 427-4754		
<b>P E</b> 501 E South St Box 02137.....	617 268-8213		
<b>P L</b> 44 Hastings Box 02131.....	617 427-9170		
<b>P R</b> 91 Boyer Jan 02134.....	617 968-8692		
<b>Paul &amp; Constance</b> 114 Aspen Av W Box 02130.....	617 325-2036		
<b>Paul E</b> 501 E South St Box 02137.....	617 268-4546		
<b>Paul M</b> 27 Union Plt 02135.....	617 787-2115		
<b>Carter Pike Driving Inc</b> 27 Beaver Ct Framingham 01702.....	Wellesley Falls 781.235-0488		
<b>Carter Prudence</b> 40 Franklin Woburn 02172.....	617 393-3782		
<b>Prudence</b> 40 Franklin Woburn 02172.....	617 926-7063		
<b>Reginald</b> 106 Brookview Dorchester 02215.....	617 541-2843		
<b>Renee &amp; Andrew</b> 10 Walnut Box 02108.....	617 720-3765		
<b>Carter Rice David</b> Baker Dennis Publishing 163 Main Wilmington 01887 Toll Free 800 638-1671			
<b>Carl Rice Industrial Prod</b> 113 Main Wilmington Toll Free 800 619-7447			
<b>Toll Free 800 619-7447</b> Toll Free 800 619-7447			
<b>Carl</b> Headquarters 413 Main Wilmington 01887 Toll Free 800 619-7447			
<b>Carl</b> Ingalls Crane 163 Main Wilmington 01887 Toll Free 800 619-7447			
<b>Carter Richard</b> 207 Cambridge Av Brighton 02135.....	617 987-9836		
<b>Richard A</b> 87 Mt Vernon Box 02106.....	617 267-0710		
<b>Carter Richard A MD</b> 13 Market St Box 02137.....	617 268-9448		
<b>Richard L</b> 275 Melrose Av Cam 02142.....	617 864-1535		
<b>Roger</b> 150 St Pauls Box 02134.....	617 424-6148		
<b>Roy</b> 41 Concord Av 02134.....	617 491-6115		
<b>Royce</b> 18 Sumner Cha 02129.....	617 241-0418		

# Our Idea: Meaning Through Geography

Set of clusters  $\approx$

A list of unconnected addresses

wide at SuperPages.com

	195	Car	C
<b>Carterge New England Inc</b>			
37 566-1282	26 Allen Ln Ipswich MA 01938	978 356-9960	
<b>Carterge Lydia</b>			
81 447-4101	38 Sweet Rd 02131	617 323-7639	
<b>Carterge Aveth</b>			
90 257-9961	15 South St 02139	617 442-9780	
	8 Hrd 02136	617 361-5253	
37 566-1282	Justice 50 Decatur Cha 02129	617 241-0152	
37 364-5188	Lucilla 124 Harvard Cha 02139	617 491-5621	
	M 95 Howe Rd 02133	617 323-9713	
361-0380	Melvin 501 Green Cam 02139	617 576-1061	
<b>Carte Nicholas</b>			
37 566-4548	38 Appleton Boston 02114	617 695-6996	
	Carterge O 44 Harvard Cha 02131	617 338-9219	
37 628-8248	Carten Thos J Sr & Claire		
	1 Ivesdale Rd MA 02136	617 698-6163	
37 445-5116	Thomas & Kathleen		
	50 Thompson Ln MA 02136	617 696-6919	
37 822-9902	Carter A Sr 02133	617 339-2257	
37 422-5712	A Hubert	617 442-5230	
37 569-2698	A 31 Bethune Wy Rndwy 02131	617 442-1219	
	A 200 Putnam Av Cambridge 02139	617 492-4174	
37 667-5190	A M 255 Mchua Rd 02136	617 266-7153	
	Adams 381 Carter St MA 02136	617 698-9074	
37 569-1417	Alice 108 Elmwood Rch 02136	617 425-0193	
37 338-9110	Alice 40 Market Cambridge 02139	617 945-2711	
	Andrew F 42 Wal St 02133	617 625-6235	
37 825-1919	Carter Anne MD	617 739-1022	
	1101 Beacon Wn 02144		
37 296-1593	Carter Becky 02134	617 523-4368	
37 670-2078	B E 108 Graduate Av MA 02136	617 526-6329	
37 623-9001	Carter Barbara L MD	617 296-6911	
	Tuffs New England Medical Res 02131		
37 296-4725	Carter Becky 02134	617 523-4368	
37 542-1521	Bernard J		
	24 Cambridge St 02136	617 567-3430	
37 364-5232	Bethiah 25 Midway Av MA 02136	617 298-8713	
37 541-5649	Bishop 25 Midway Av MA 02136	617 367-9931	
37 739-2662	Carter Broadcasting Co		
	26 Park Pl 02136	617 423-0210	
37 879-0030	C 21 St 02131	617 225-0200	
37 541-3948	Carter C 200 Cambridge Av St 02135	617 782-2118	
37 936-1511	C 219 Harvard Av East Boston 02128	617 569-1545	
37 569-4119	C 259 Harvard Cam 02136	617 491-4822	
	C 259 Harvard Cam 02136	617 296-4392	
37 569-4782	C & M 41 Burroughs Jan 02136	617 524-5595	
<b>Carter F 24 Hickox Bx 02131</b>			
	617 327-1105		
<b>Faye &amp; Ricky</b>			
	207 Columbia Av Bx 02136	617 437-7331	
<b>Francis S 134 Temple W Av 02132</b>			
	617 323-6781		
<b>Franklin &amp; Anne</b>			
	251 Mt Auburn Cam 02136	617 354-0798	
	Fred 42 Harvard Cam 02136	617 524-3078	
	Fred 16 Harvard Av MA 02136	617 698-1343	
	G & E 8 Harvard Der 02134	617 436-8906	
	G T 27 Fyfield Av Stn 02145	617 623-7121	
	Gayle 25 Franklin Der 02134	617 825-0322	
	Geo S 115 Mystic Hill Rd Jan 02136	617 522-3215	
	George 125 Madison Bx 02131	617 367-9548	
<b>Carter Halliday Associate</b>			
	107 S Street Bx 02111	617 456-1689	
<b>Carter Harry F</b>			
	26 Baring Jct Rd W Av 02132	617 325-5465	
<b>Carter Hide Co Inc</b>			
	146 Sumner Bx 02131	617 542-7987	
	Carter Hilary 41 Harvey Cam 02140	617 876-2750	
<b>Horace</b>			
	381 Walnut Av Rndwy 02131	617 442-5307	
	Howard Jr 38 Nebe One Bx 02116	617 445-5552	
	J Cam	617 354-2658	
	J 35 Chatham Bx 02144	617 232-7990	
	J 35 Harvard Bx 02144	617 730-9483	
	J 775 Wy Wmly Westbury 02132	617 323-5574	
	Carter J Jacques MD		
	1 Broadview Pl Bx 02144	617 735-8787	
<b>Carter J M</b>			
	3410 Columbia Rd S Bx 02137	617 464-1040	
<b>Carter J M Ornamental Ironworks</b>			
	100 Franklin Der 02136	617 436-5353	
<b>Carter J Veal Co</b>			
	40 Newbury St 02138	617 442-1775	
<b>Carter James</b>			
	1573 Cambridge St Cam 02136	617 492-1214	
	James 302 Fisher Av Rndwy 02136	617 739-2193	
	James 312 Cedar St Rd Cambridge 02140	617 876-8841	
	Jas L 34 Broadway Rd MA 02136	617 361-0773	
	Jane 14 Adams Rd Newton 02458	617 564-0435	
	Jeffrey 41 Warner Av Bx 02136	617 424-5994	
	John 11 Mansfield Rd 02134	617 987-2163	
	John 207 Sumner Bx 02137	617 423-4334	
	John 40 Westwood Rd Der 02125	617 282-1235	
	June O 129 A Sumner Av St 02138	617 734-6109	
	K 10 Wemying Av Rndwy 02136	617 265-9456	
	K 17 Elwood Der 02123	617 282-1593	
<b>Carter Nellie E</b>			
	323 Marchette Av Bx 02135	617 267-6483	
<b>Nicholas S F</b>			
	115 Randolph Av MA 02136	617 698-5307	
	Nick 21 Fyfield Bx 02114	617 267-5222	
<b>Nick &amp; Debbi</b>			
	146 Vermont Rd Newton 02459	617 527-0480	
<b>Norman G</b>			
	38 Chickadee Rd 02125	617 822-1203	
	P 40 Cranston Pl Bx 02135	617 427-4754	
	P E 501 E South St Bx 02137	617 268-4213	
	P L 44 Matthews Bx 02131	617 427-9170	
	P R 91 Boyer Jan 02136	617 983-8692	
<b>Paul &amp; Constance</b>			
	114 Adams Av W Bx 02130	617 325-2036	
	Paul E 501 E South St Bx 02137	617 268-4546	
	Paul M 27 Union St 02135	617 787-2115	
<b>Carter Pile Driving Inc 17 Beaver Ct</b>			
	Franklin 02102	617 781-235-0488	
<b>Carter Prudence</b>			
	40 Franklin Westbury 02127	617 393-3782	
<b>Prudence</b>			
	40 Franklin Westbury 02127	617 393-3782	
<b>Reginald</b>			
	100 Broadview Der 02125	617 541-2843	
<b>Renee &amp; Andrew</b>			
	10 Walnut Bx 02138	617 720-3765	
<b>Carter Rice Dowd</b>			
	Baker Dennis Publishing 163 Main Wilmington 01887		
	Ted Free-Old 1 & Thon	800 638-1671	
	Carl Free-Old 1 & Thon 013 Main Wilmington	800 619-7447	
	Ted Free-Old 1 & Thon	800 619-7447	
	Ted Free-Old 1 & Thon	800 648-7447	
	Headquarters 412 Main Wilmington 01887		
	Cal	978 988-7447	
	Inga One 363 Main Wilmington 01887		
	Richard A 171 Thon 02137	800 638-1673	
<b>Carter Richard</b>			
	207 Carver Av Brighton 02137	617 987-0836	
	Richard A 9748 Vernon Bx 02136	617 556-7293	
<b>Carter Richard A MD</b>			
	170 Wmly Bx 02136	978 987-0710	
<b>Carter Richard K</b>			
	133 Merwin St 02137	617 268-9448	
	Robert L 175 Madison Av Cam 02140	617 864-1535	
	Roger 150 St Pauls Bx 02131	617 424-6148	
	Roy 41 Concord Rd 02138	617 491-6115	
	Royce 185 Salisbury Cha 02129	617 241-9418	



# Our Idea: Meaning Through Geography

Set of clusterings  $\approx$   
A list of unconnected addresses

wide at SuperPages.com

195		Car		C	
37 566-1282	Cartage New England Inc	37 566-1282	Cartage F. J. Hillcock Inc 01213...	37 566-1282	Cartage F. J. Hillcock Inc 01213...
81 447-4101	Cartagena Lydia	81 447-4101	Faye & Ricky	81 447-4101	Faye & Ricky
90 257-9961	Cartagena Aveth	90 257-9961	Francis S. 134 Temple W Ave 01213...	90 257-9961	Francis S. 134 Temple W Ave 01213...
37 566-1282	Cartagena Aveth	37 566-1282	Franklin & Anne	37 566-1282	Franklin & Anne
37 564-5188	Cartagena Aveth	37 564-5188	Fred 42 Howard St 01213...	37 564-5188	Fred 42 Howard St 01213...
361-0380	Cartagena Aveth	361-0380	Fred 42 Howard St 01213...	361-0380	Fred 42 Howard St 01213...
37 566-4548	Cartagena Aveth	37 566-4548	Fred 42 Howard St 01213...	37 566-4548	Fred 42 Howard St 01213...
37 628-8248	Cartagena Aveth	37 628-8248	Fred 42 Howard St 01213...	37 628-8248	Fred 42 Howard St 01213...
37 445-5116	Cartagena Aveth	37 445-5116	Fred 42 Howard St 01213...	37 445-5116	Fred 42 Howard St 01213...
37 822-2992	Cartagena Aveth	37 822-2992	Fred 42 Howard St 01213...	37 822-2992	Fred 42 Howard St 01213...
37 422-5712	Cartagena Aveth	37 422-5712	Fred 42 Howard St 01213...	37 422-5712	Fred 42 Howard St 01213...
37 569-2698	Cartagena Aveth	37 569-2698	Fred 42 Howard St 01213...	37 569-2698	Fred 42 Howard St 01213...
37 667-5190	Cartagena Aveth	37 667-5190	Fred 42 Howard St 01213...	37 667-5190	Fred 42 Howard St 01213...
37 569-1417	Cartagena Aveth	37 569-1417	Fred 42 Howard St 01213...	37 569-1417	Fred 42 Howard St 01213...
37 822-1953	Cartagena Aveth	37 822-1953	Fred 42 Howard St 01213...	37 822-1953	Fred 42 Howard St 01213...
37 296-1593	Cartagena Aveth	37 296-1593	Fred 42 Howard St 01213...	37 296-1593	Fred 42 Howard St 01213...
37 670-2078	Cartagena Aveth	37 670-2078	Fred 42 Howard St 01213...	37 670-2078	Fred 42 Howard St 01213...
37 623-9001	Cartagena Aveth	37 623-9001	Fred 42 Howard St 01213...	37 623-9001	Fred 42 Howard St 01213...
37 296-4725	Cartagena Aveth	37 296-4725	Fred 42 Howard St 01213...	37 296-4725	Fred 42 Howard St 01213...
37 542-1521	Cartagena Aveth	37 542-1521	Fred 42 Howard St 01213...	37 542-1521	Fred 42 Howard St 01213...
37 364-5232	Cartagena Aveth	37 364-5232	Fred 42 Howard St 01213...	37 364-5232	Fred 42 Howard St 01213...
37 541-5649	Cartagena Aveth	37 541-5649	Fred 42 Howard St 01213...	37 541-5649	Fred 42 Howard St 01213...
37 739-2662	Cartagena Aveth	37 739-2662	Fred 42 Howard St 01213...	37 739-2662	Fred 42 Howard St 01213...
37 879-0030	Cartagena Aveth	37 879-0030	Fred 42 Howard St 01213...	37 879-0030	Fred 42 Howard St 01213...
37 936-1511	Cartagena Aveth	37 936-1511	Fred 42 Howard St 01213...	37 936-1511	Fred 42 Howard St 01213...
37 569-4119	Cartagena Aveth	37 569-4119	Fred 42 Howard St 01213...	37 569-4119	Fred 42 Howard St 01213...
37 569-4782	Cartagena Aveth	37 569-4782	Fred 42 Howard St 01213...	37 569-4782	Fred 42 Howard St 01213...



$\approx$  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)

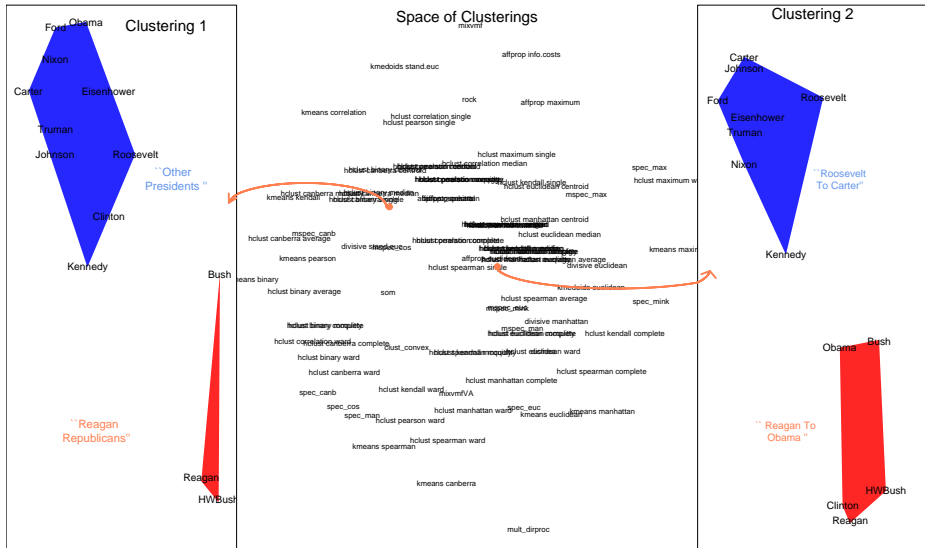
# A New Strategy

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** (takes about 10-15 minutes to choose a clustering with insight)

# Many Thousands of Clusterings, Sorted & Organized

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





# 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 the Performance of Our Method



# Evaluating the Performance of Our Method

- Goals:

# Evaluating the Performance of Our Method

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

# Evaluating the Performance of Our Method

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

# Evaluating the Performance of Our Method

- 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 the Performance of Our Method

- 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 the Performance of Our Method

- 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
  - Quality  $\Rightarrow$  RA coders

# Evaluating the Performance of Our Method

- 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
  - Quality  $\Rightarrow$  RA coders
  - Informative discoveries  $\Rightarrow$  Experienced scholars analyzing texts

# Evaluating the Performance of Our Method

- 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
  - Quality  $\Rightarrow$  RA coders
  - Informative discoveries  $\Rightarrow$  Experienced scholars analyzing texts
  - Discovery  $\Rightarrow$  You're the judge



# Evaluation 1: Cluster Quality

# Evaluation 1: Cluster Quality

- What Are Humans Good For?

# Evaluation 1: Cluster Quality

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

# Evaluation 1: Cluster Quality

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

# Evaluation 1: Cluster Quality

- What Are Humans Good For?

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

# Evaluation 1: Cluster Quality

- What Are Humans Good For?
  - They can't: keep many documents & clusters in their head
  - They can: compare two documents at a time
  - $\implies$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality

# Evaluation 1: Cluster Quality

- What Are Humans Good For?
  - They can't: keep many documents & clusters in their head
  - They can: compare two documents at a time
  - $\implies$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality
  - automated visualization to choose one clustering

# Evaluation 1: Cluster Quality

- What Are Humans Good For?
  - They can't: keep many documents & clusters in their head
  - They can: compare two documents at a time
  - $\Rightarrow$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality
  - automated visualization to choose one clustering
  - many pairs of documents



# Evaluation 1: Cluster Quality

- What Are Humans Good For?
  - They can't: keep many documents & clusters in their head
  - They can: compare two documents at a time
  - $\Rightarrow$  Cluster quality evaluation: human judgement of document pairs
- Experimental Design to Assess Cluster Quality
  - automated visualization to choose one clustering
  - many pairs of documents
  - for coders: (1) unrelated, (2) loosely related, (3) closely related

# Evaluation 1: Cluster Quality

- What Are Humans Good For?

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

- Experimental Design to Assess Cluster Quality

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

# Evaluation 1: Cluster Quality

- What Are Humans Good For?

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

- Experimental Design to Assess Cluster Quality

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

# Evaluation 1: Cluster Quality

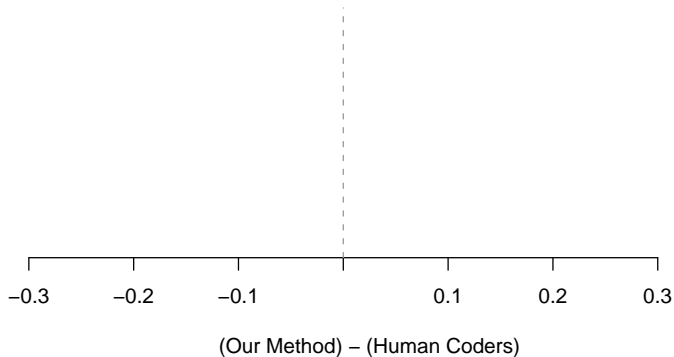
- What Are Humans Good For?

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

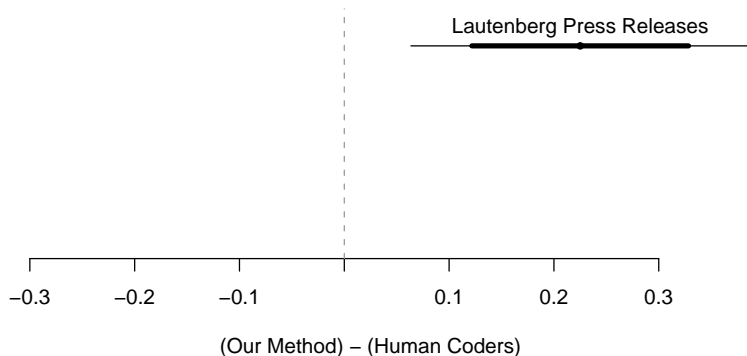
- Experimental Design to Assess Cluster Quality

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

# Evaluation 1: Cluster Quality

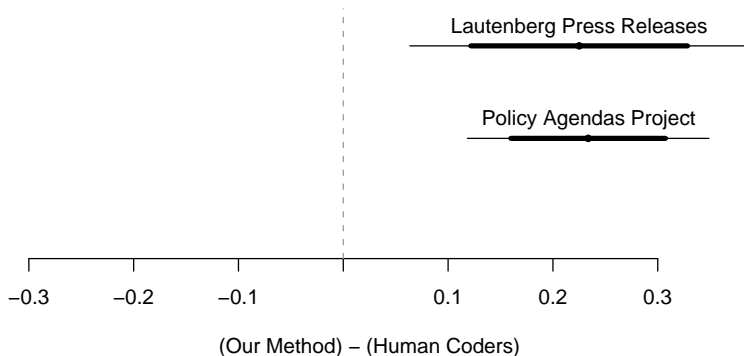


# Evaluation 1: Cluster Quality



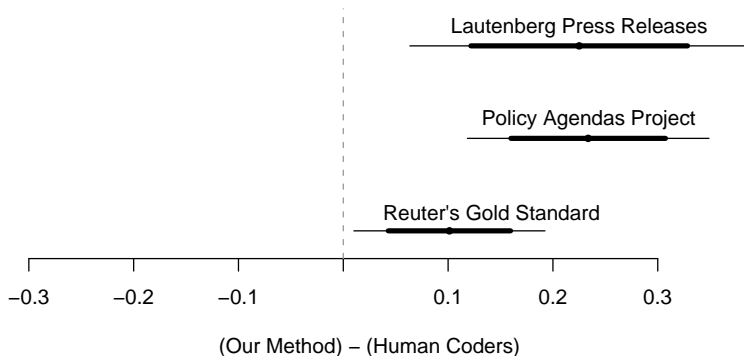
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

# Evaluation 1: Cluster Quality



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

# Evaluation 1: Cluster Quality



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



# Evaluation 2: More Informative Discoveries

## Evaluation 2: More Informative Discoveries

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

## Evaluation 2: More Informative Discoveries

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

## Evaluation 2: More Informative Discoveries

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

## Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)

## Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)

## Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for  $\binom{6}{2}=15$  pairwise comparisons

## Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for  $\binom{6}{2}=15$  pairwise comparisons
- User chooses  $\Rightarrow$  only care about the one clustering that wins



## Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for  $\binom{6}{2}=15$  pairwise comparisons
- User chooses  $\Rightarrow$  only care about the one clustering that wins
- Both cases a Condorcet winner:

# Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for  $\binom{6}{2}=15$  pairwise comparisons
- User chooses  $\Rightarrow$  only care about the one clustering that wins
- Both cases a Condorcet winner:

“Immigration”:

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

## Evaluation 2: More Informative Discoveries

- Found 2 scholars analyzing lots of textual data for their work
- Created 6 clusterings:
  - 2 clusterings selected with our method (**biased** against us)
  - 2 clusterings from each of 2 other methods (varying tuning parameters)
- Created info packet on each clustering (for each cluster: exemplar document, automated content summary)
- Asked for  $\binom{6}{2}=15$  pairwise comparisons
- User chooses  $\Rightarrow$  only care about the one clustering that wins
- Both cases a Condorcet winner:

“Immigration”:

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

“Genetic testing”:

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

# Evaluation 3: What Do Members of Congress Do?

# Evaluation 3: What Do Members of Congress Do?

- David Mayhew's (1974) famous typology

# Evaluation 3: What Do Members of Congress Do?

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

# Evaluation 3: What Do Members of Congress Do?

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

# Evaluation 3: What Do Members of Congress Do?

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



# Evaluation 3: What Do Members of Congress Do?

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

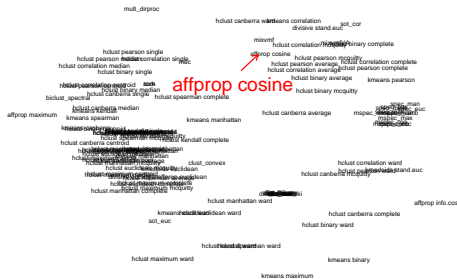
# Evaluation 3: What Do Members of Congress Do?

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

## / 21

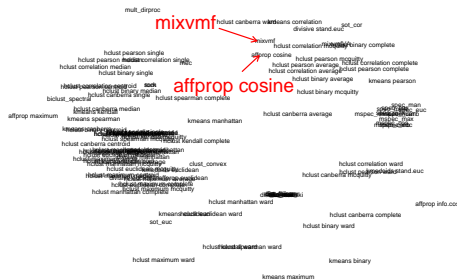


## Example Discovery



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)



between methods:

## Example Discovery

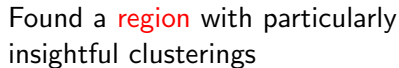


Space between methods:  
local cluster ensemble



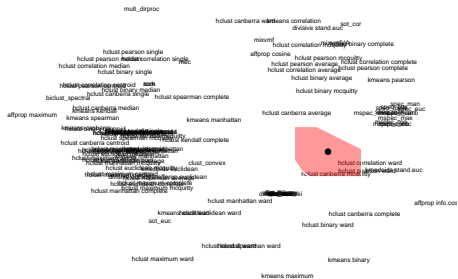
A set of small navigation icons typically found in Beamer presentations, including symbols for back, forward, search, and other slide controls.







## Example Discovery



Mixture:

### 0.39 Hclust-Canberra-McQuitty

## Example Discovery

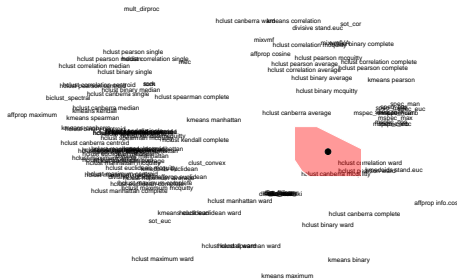


Mixture:

### 0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering  
Random Walk  
(Metrics 1-6)

## Example Discovery



Mixture:

### 0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering  
Random Walk  
(Metrics 1-6)

### 0.13 Hclust-Correlation-Ward

# Example Discovery



## Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering  
Random Walk  
(Metrics 1-6)

0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

## Example Discovery



Mixture:

### 0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering  
Random Walk  
(Metrics 1-6)

### 0.13 Hclust-Correlation-Ward

## 0.09 Hclust-Pearson-Ward

### 0.05 Kmediods-Cosine



## Example Discovery



Mixture:

### 0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering  
Random Walk  
(Metrics 1-6)

### 0.13 Hclust-Correlation-Ward

## 0.09 Hclust-Pearson-Ward

### 0.05 Kmediods-Cosine

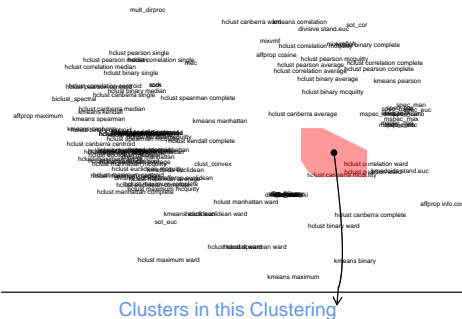
#### 0.04 Spectral clustering

Symmetric  
(Metrics 1-6)

Mayhew



## Example Discovery

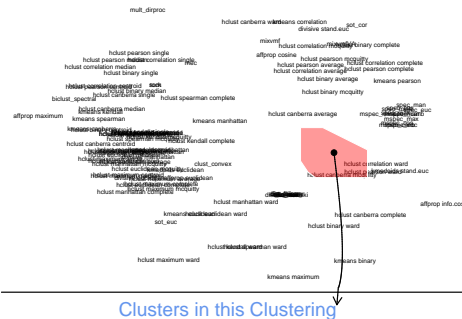


Credit Claiming  
Pork

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

## Mayhew

## Example Discovery



## Credit Claiming, Legislation:

"As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period"

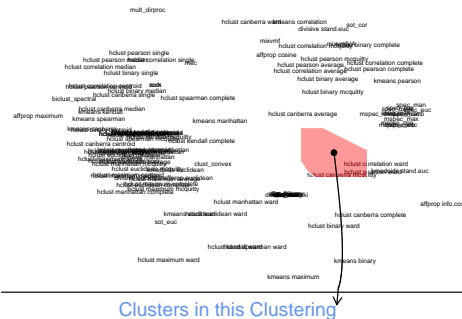
## Credit Claiming Pork

# Mayhew Credit Claiming Legislation

Gary King (Harvard IQSS)

## Quantitative Discovery

## Example Discovery



## Advertising:

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



Credit Claiming  
Pork

## Advertising

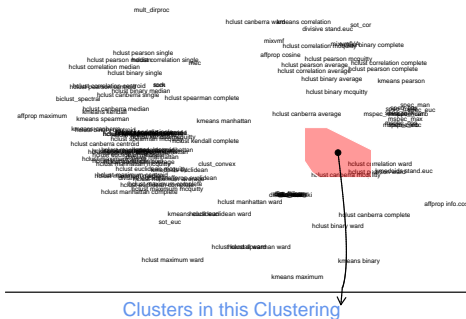
## Mayhew

## Credit Claiming Legislation

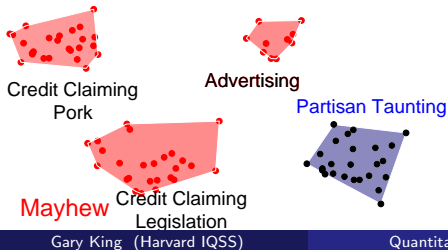
Gary King (Harvard IQSS)

## Quantitative Discovery

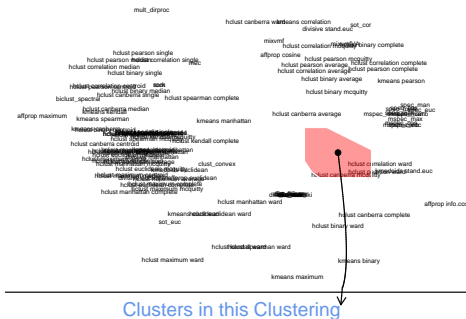
## Example Discovery: Partisan Taunting



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



## Example Discovery: Partisan Taunting



## Partisan Taunting:

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



Credit Claiming  
Pork

## Advertising

## Partisan Taunting

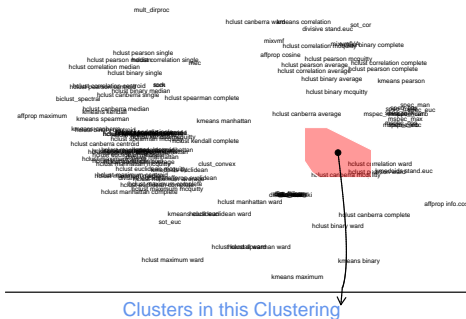
## Mayhew

## Credit Claiming Legislation

Gary King (Harvard IQSS)

## Quantitative Discovery

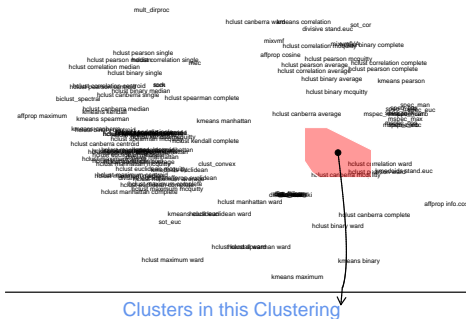
## Example Discovery: Partisan Taunting



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

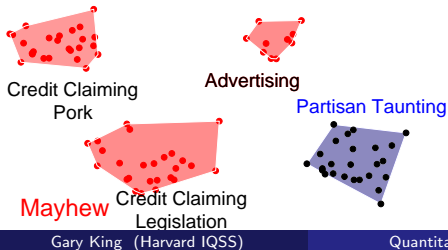


## Example Discovery: Partisan Taunting



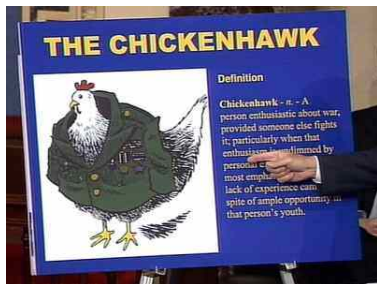
**Definition:** Explicit, public, and negative attacks on another political party or its members

## Taunting ruins deliberation



# In Sample Illustration of Partisan Taunting

## Taunting ruins deliberation

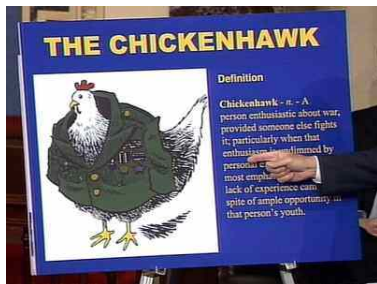


Sen. Lautenberg  
on Senate Floor  
4/29/04

- “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’ ”  
[Government Oversight]

# In Sample Illustration of Partisan Taunting

## Taunting ruins deliberation

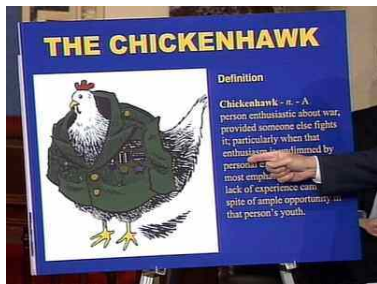


Sen. Lautenberg  
on Senate Floor  
4/29/04

- “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’ ” [Government Oversight]
- “The scopes trial took place in 1925. Sadly, President Bush’s veto today shows that we haven’t progressed much since then” [Healthcare]

# In Sample Illustration of Partisan Taunting

## Taunting ruins deliberation



Sen. Lautenberg  
on Senate Floor  
4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities less safe." [Homeland Security]

# Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

# Out of Sample Confirmation of Partisan Taunting

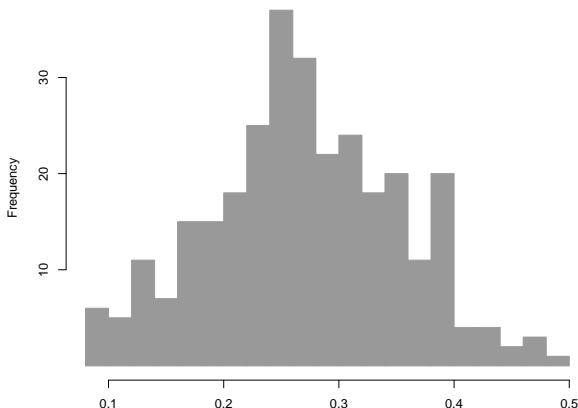
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

# Out of Sample Confirmation of Partisan Taunting

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

# Out of Sample Confirmation of Partisan Taunting

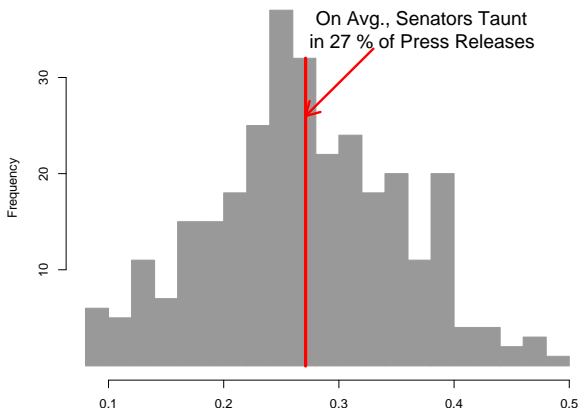
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.
- Apply supervised learning method: measure **proportion of press releases** a senator taunts other party



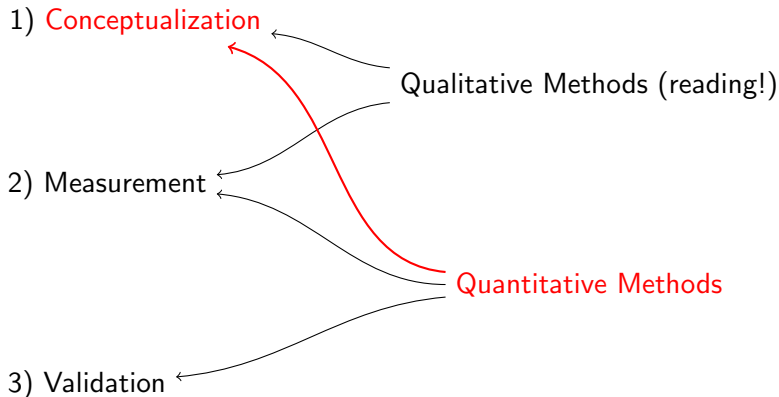


# Out of Sample Confirmation of Partisan Taunting

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

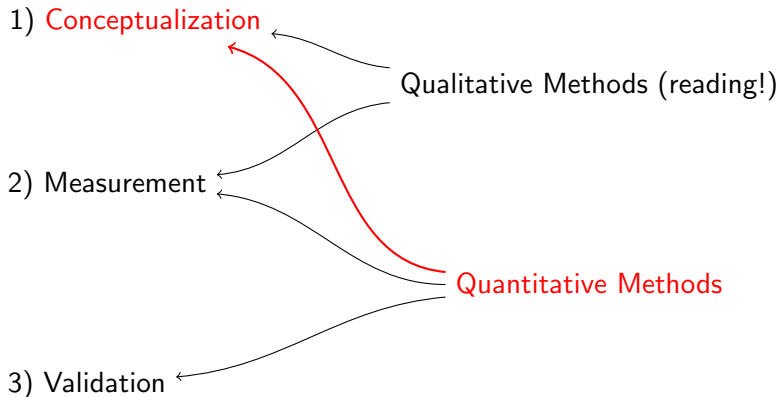


# Advancing the Objective of Discovery



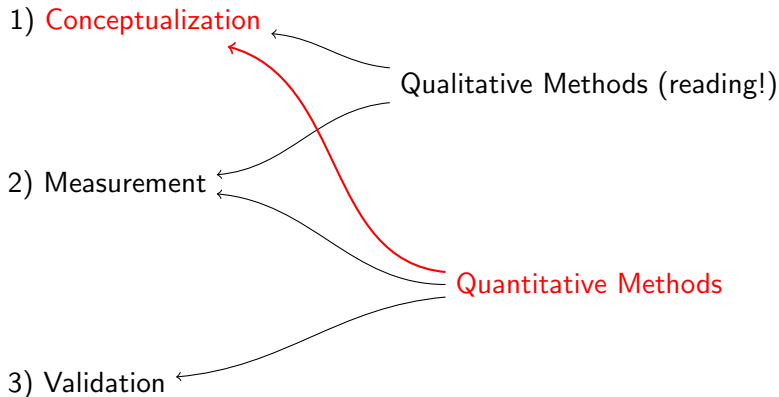
Quantitative methods for conceptualization: aiding **discovery**

# Advancing the Objective of Discovery



- Quantitative methods for conceptualization: aiding **discovery**
- Few formal methods designed explicitly for conceptualization

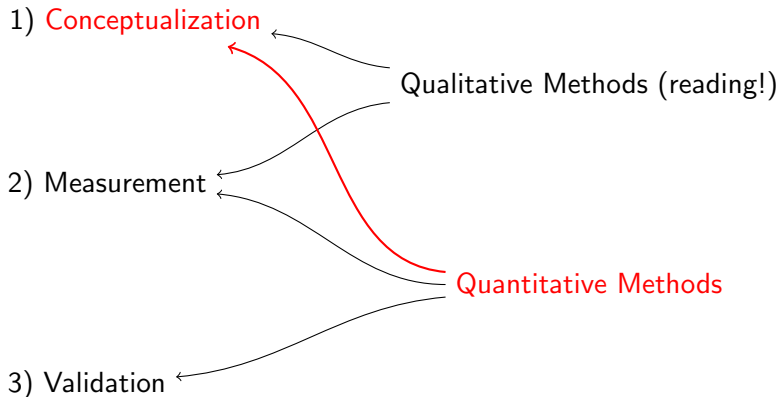
# Advancing the Objective of Discovery



Quantitative methods for conceptualization: aiding **discovery**

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

# Advancing the Objective of Discovery



Quantitative methods for conceptualization: aiding **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 (on adding zooming out to the human ability to zoom in)

<http://GKing.Harvard.edu>