

# Computer-Assisted Clustering and Conceptualization from Unstructured Text

Gary King

Institute for Quantitative Social Science  
Harvard University

Machine Learning/Google Distinguished Lecture, Carnegie Mellon University, 3/17/2011

---

<sup>1</sup>Based on joint work with Justin Grimmer (Harvard ↔ Stanford)

# A Method for Computer Assisted Conceptualization

- Conceptualization through **Classification**: “one of the most central and generic of all our conceptual exercises. . . . the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis. . . . Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research.” (Bailey, 1994).

# A Method for Computer Assisted Conceptualization

- Conceptualization through **Classification**: “one of the most central and generic of all our conceptual exercises. . . . the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis. . . . Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research.” (Bailey, 1994).
- **Cluster Analysis**: simultaneously (1) invents categories and (2) assigns documents to categories

# A Method for Computer Assisted Conceptualization

- Conceptualization through **Classification**: “one of the most central and generic of all our conceptual exercises. . . . the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis. . . . Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research.” (Bailey, 1994).
- **Cluster Analysis**: simultaneously (1) invents categories and (2) assigns documents to categories
- We focus on unstructured text; methods apply more broadly.

# A Method for Computer Assisted Conceptualization

- Conceptualization through **Classification**: “one of the most central and generic of all our conceptual exercises. . . . the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis. . . . Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research.” (Bailey, 1994).
- **Cluster Analysis**: simultaneously (1) invents categories and (2) assigns documents to categories
- We focus on unstructured text; methods apply more broadly.
- Main goal: Switch from **Fully Automated** to **Computer Assisted**

# What's Hard about Clustering?

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!



# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!
- $Bell(n)$  = number of ways of partitioning  $n$  objects
- $Bell(2) = 2$  (AB, A B)
- $Bell(3) = 5$  (ABC, AB C, A BC, AC B, A B C)

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!
- $Bell(n)$  = number of ways of partitioning  $n$  objects
- $Bell(2) = 2$  (AB, A B)
- $Bell(3) = 5$  (ABC, AB C, A BC, AC B, A B C)
- $Bell(5) = 52$

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!
- $Bell(n)$  = number of ways of partitioning  $n$  objects
- $Bell(2) = 2$  (AB, A B)
- $Bell(3) = 5$  (ABC, AB C, A BC, AC B, A B C)
- $Bell(5) = 52$
- $Bell(100) \approx$

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!
- $Bell(n)$  = number of ways of partitioning  $n$  objects
- $Bell(2) = 2$  (AB, A B)
- $Bell(3) = 5$  (ABC, AB C, A BC, AC B, A B C)
- $Bell(5) = 52$
- $Bell(100) \approx 10^{28} \times$  Number of elementary particles in the universe

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!
- $Bell(n)$  = number of ways of partitioning  $n$  objects
- $Bell(2) = 2$  (AB, A B)
- $Bell(3) = 5$  (ABC, AB C, A BC, AC B, A B C)
- $Bell(5) = 52$
- $Bell(100) \approx 10^{28} \times$  Number of elementary particles in the universe
- Now imagine choosing the *optimal* classification scheme by hand!

# What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!
- $Bell(n)$  = number of ways of partitioning  $n$  objects
- $Bell(2) = 2$  (AB, A B)
- $Bell(3) = 5$  (ABC, AB C, A BC, AC B, A B C)
- $Bell(5) = 52$
- $Bell(100) \approx 10^{28} \times$  Number of elementary particles in the universe
- Now imagine choosing the *optimal* classification scheme by hand!
- Fully automated algorithms can help, but which ones?



# The Problem with Fully Automated Clustering

# The Problem with Fully Automated Clustering

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

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy  $k$ -modes, affinity propagation, self-organizing maps,...

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy  $k$ -modes, affinity propagation, self-organizing maps,...
  - **Well-defined** statistical, data analytic, or machine learning foundations

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy  $k$ -modes, affinity propagation, self-organizing maps, . . .
  - **Well-defined** statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, **unclear**

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy  $k$ -modes, affinity propagation, self-organizing maps, . . .
  - **Well-defined** statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, **unclear**
  - The literature: **little guidance on when methods apply**



# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy  $k$ -modes, affinity propagation, self-organizing maps, . . .
  - **Well-defined** statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, **unclear**
  - The literature: **little guidance on when methods apply**
  - **Deriving such guidance:** difficult or impossible

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy  $k$ -modes, affinity propagation, self-organizing maps,...
  - **Well-defined** statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, **unclear**
  - The literature: **little guidance on when methods apply**
  - **Deriving such guidance:** difficult or impossible
- **Deep problem:** full automation requires more information

# The Problem with Fully Automated Clustering

- **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
- **No free lunch theorem:** every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - **Many choices:** model-based, subspace, spectral, grid-based, graph-based, fuzzy  $k$ -modes, affinity propagation, self-organizing maps, . . .
  - **Well-defined** statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, **unclear**
  - The literature: **little guidance on when methods apply**
  - **Deriving such guidance:** difficult or impossible
- **Deep problem:** full automation requires more information
- No surprise: everyone's tried cluster analysis; very few are satisfied

# Switch from Fully Automated to Computer Assisted

# Switch from Fully Automated to Computer Assisted

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

# Switch from Fully Automated to Computer Assisted

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

# Switch from Fully Automated to Computer Assisted

- **Fully Automated Clustering** may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: **Computer-Assisted Clustering**
  - **Easy in theory:** list all clusterings; choose the best

# Switch from Fully Automated to Computer Assisted

- **Fully Automated Clustering** may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: **Computer-Assisted Clustering**
  - **Easy in theory:** list all clusterings; choose the best
  - **Impossible in practice:** Too hard for us mere humans!



# Switch from Fully Automated to Computer Assisted

- **Fully Automated Clustering** may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: **Computer-Assisted Clustering**
  - **Easy in theory:** list all clusterings; choose the best
  - **Impossible in practice:** Too hard for us mere humans!
  - An **organized list** will make the search possible

# Switch from Fully Automated to Computer Assisted

- **Fully Automated Clustering** may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: **Computer-Assisted Clustering**
  - **Easy in theory:** list all clusterings; choose the best
  - **Impossible in practice:** Too hard for us mere humans!
  - An **organized list** will make the search possible
  - **Insight:** Many clusterings are perceptually identical

# Switch from Fully Automated to Computer Assisted

- **Fully Automated Clustering** may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: **Computer-Assisted Clustering**
  - **Easy in theory:** list all clusterings; choose the best
  - **Impossible in practice:** Too hard for us mere humans!
  - An **organized list** will make the search possible
  - **Insight:** Many clusterings are perceptually identical
  - E.g.,: consider two clusterings that differ only because one document (of 10,000) moves from category 5 to 6

# Switch from Fully Automated to Computer Assisted

- **Fully Automated Clustering** may succeed sometimes, but fails in general: too hard to understand when each model applies
- An alternative: **Computer-Assisted Clustering**
  - **Easy in theory:** list all clusterings; choose the best
  - **Impossible in practice:** Too hard for us mere humans!
  - An **organized list** will make the search possible
  - **Insight:** Many clusterings are perceptually identical
  - E.g.,: consider two clusterings that differ only because one document (of 10,000) moves from category 5 to 6
- **Question: How to organize clusterings so humans can understand?**

# Our Idea: Meaning Through Geography

Set of clusterings

# Our Idea: Meaning Through Geography

Set of clusterings  $\approx$

A list of unconnected addresses

wide at [SuperPages.com](http://SuperPages.com)

	195	Car	C
<b>Cartage New England Inc</b> 28 Allen Ln Ipswich 01938..... 978 356-9960	<b>Carter F</b> 34 Hibiscus Bldg 02133..... 617 327-1105	<b>Carter Nellie E</b> 323 Main St 02115..... 617 267-6483	
<b>Cartagena Lydia</b> 28 Sweet Box 02131..... 617 323-7639	<b>Faye &amp; Ricky</b> 20 Columbia Ave Box 02136..... 617 437-7331	<b>Nicholas S F</b> 115 Randolph Ave Box 02186..... 617 698-6307	
<b>Cartagena Avish</b> F Pleasant Box 02139..... 617 442-9780	<b>Francis S</b> 134 Yankov W Ave 02132..... 617 323-6781	<b>Nick 21 Farwell Box 02114</b> ..... 617 267-5222	
<b>B Hed 02134</b> ..... 617 361-5253	<b>Franklin &amp; Anne</b> 201 Mt Auburn Cam 02138..... 617 354-0798	<b>Nick &amp; Debbi</b> 196 Herold Rd Newton 02459..... 617 527-0480	
<b>Jessica</b> 50 Decatur Cha 02129..... 617 241-0152	<b>Fred 41 Hawthorn Elm 02138</b> ..... 617 524-3078	<b>Nicole</b> ..... 617 698-0713	
<b>Luzmila</b> 124 Harvard Cam 02138..... 617 491-5621	<b>Fred 16 Rowley Ave 02138</b> ..... 617 698-1343	<b>Norman G</b> 38 Chickawhoh Dr 02125..... 617 822-1201	
<b>M 95 Howe Box 02132</b> ..... 617 323-9713	<b>G &amp; B</b> 8 Vardon Dr 02134..... 617 436-8906	<b>P 40 Cranston Pl Box 02135</b> ..... 617 437-4754	
<b>Melvin</b> 503 Green Cam 02139..... 617 576-1061	<b>G T 27 Franklin Ave 02145</b> ..... 617 623-7121	<b>P E 501 E South S Box 02137</b> ..... 617 268-8213	
<b>Carte Nicholas</b> 18 Appleton Boston 02114..... 617 695-6996	<b>Gayle</b> 25 Franklin Dr 02133..... 617 823-0322	<b>P L 44 Hutchings Box 02135</b> ..... 617 427-9170	
<b>Cartier</b> 0 4 Bedford Box 02138..... 617 338-0219	<b>Geo S</b> 115 Mass Mt Hill Rd 02138..... 617 522-3215	<b>P R 91 Boyer Ave 02138</b> ..... 617 968-8692	
<b>Carten Thos Jr Sr &amp; Claire</b> 1 Franklin St Mt 02138..... 617 698-6163	<b>George</b> 125 Hudson Box 02134..... 617 367-9548	<b>Paul &amp; Constance</b> 114 Franklin St W Box 02133..... 617 325-2036	
<b>17 445-5116</b> <b>Thomas &amp; Kathleen</b> 50 Thompson Ln Mt 02136..... 617 696-6919	<b>Carter Hillside Assoc</b> 107 S Street Box 02111..... 617 456-1689	<b>Paul E 501 E South S Box 02137</b> ..... 617 268-4546	
<b>17 822-2962</b> <b>Carter A</b> Box 02133..... 617 229-2257	<b>Carter Harry F</b> 26 Bayne Rd Rt W Box 02132..... 617 325-5465	<b>Paul M 27 Union Rd 02139</b> ..... 617 787-2115	
<b>17 427-5712</b> <b>A Helen</b> 617 442-5230	<b>Carter Hide Co Inc</b> 161 Boston St 02131..... 617 542-7987	<b>Carter Pile Driving Inc 27 Avenue Ct</b> Frankington 02102..... Wobley Tpk 781.235-0488	
<b>17 569-2698</b> <b>A 33 Bethune Wy Roxbury 02119</b> ..... 617 442-1219	<b>Carter Hilary 41 Harvey Cam 02148</b> ..... 617 876-2750	<b>Carter Prudence</b> 34 Franklin Waterman 02127..... 617 393-3782	
<b>17 667-5190</b> <b>A M 255 Massachusetts Ave 02115</b> ..... 617 266-7153	<b>Horace</b> 301 Walnut St Roxbury 02119..... 617 442-5307	<b>Prudence</b> 40 Franklin Waterman 02127..... 617 926-7063	
<b>17 569-1417</b> <b>Adams 361 Centre St Mt 02138</b> ..... 617 698-9074	<b>Howard Jr 28 Nona Drive Box 02118</b> ..... 617 445-5532	<b>Roginald</b> 106 Brookview Dorchester 02122..... 617 541-2843	
<b>17 338-9110</b> <b>Alice 108 Elmwood Box 02134</b> ..... 617 423-0193	<b>J Dan</b> ..... 617 354-2658	<b>Renee &amp; Andrew</b> 10 Walnut Box 02138..... 617 720-3765	
<b>17 825-1919</b> <b>Alice 40 Market Cambridge 02139</b> ..... 617 945-2711	<b>J 31 Chatham Box 02146</b> ..... 617 232-7990	<b>Carter Rice David</b> 3400 Centre Publishing 163 Main Wilmington 01887 Toll Free-Dial '7 & Then..... 800 638-1671	
<b>17 296-1593</b> <b>Carter Anne MD</b> 1161 Beacon Bldg 02144..... 617 739-1022	<b>J 538 Harvard Box 02146</b> ..... 617 730-9483	<b>Carl Eric Industrial Prod 613 Main Wilmington</b> Toll Free-Dial '7 & Then..... 800 616-7447	
<b>17 670-2078</b> <b>Carter J M</b> 1 Ipswich Pl Box 02146..... 617 735-8787	<b>Carter J J</b> 310 Columbia Rd S Box 02137..... 617 464-1040	<b>Carl Free-Dial '7 &amp; Then</b> ..... 800 648-7447	
<b>17 621-9001</b> <b>B E 18 Graduate Ave Mt 02136</b> ..... 617 296-6911	<b>Carter J M Ornamental Ironworks</b> 201 Walnut St Roxbury 02119..... 617 442-5307	<b>Carl</b> 175 Franklin St 02131..... 617 988-7447	
<b>17 296-4725</b> <b>Carter Barbara L MD</b> Tufts-New England Medical Center Box 02111 Cam..... 617 636-0051	<b>Carter J Neal Co</b> 40 Hawthorn Elm 02138..... 617 442-1775	<b>Carl</b> Ingalls Centre 163 Main Wilmington 01887 ..... 800 638-1673	
<b>17 542-1521</b> <b>Carter Becky Jo 02134</b> ..... 617 523-4368	<b>Carter James</b> 157 Cambridge St Cam 02138..... 617 492-1214	<b>Carter Richard</b> 2079 Cambridge Ave Brighton 02215..... 617 982-0836	
<b>Bernard J</b> 122 Southside E Box 02136..... 617 567-9430	<b>James</b> 412 Foster St Roxbury 02119..... 617 739-2193	<b>Richard A MD</b> 47 Mt Vernon Box 02106..... 617 566-7293	
<b>17 364-5232</b> <b>Bibbiah 25 Midway Dr 02134</b> ..... 617 298-8713	<b>James L</b> 34 Roslindale Rd Cambridge 02142..... 617 876-8841	<b>Carter Richard A MD</b> 130 Canterbury St 02136..... 617 267-0710	
<b>17 541-5249</b> <b>Bilal 26 Elmwood Box 02138</b> ..... 617 367-9931	<b>Jane 14 Adams Rd Newton 02458</b> ..... 617 964-0435	<b>Carter Richard K</b> 23 Mather St 02127..... 617 268-0448	
<b>17 739-2662</b> <b>Carter Broadcasting Co</b> 58 Park Pl Box 02134..... 617 423-0210	<b>Jeffrey 41 Warren St 02134</b> ..... 617 426-5994	<b>Roger 130 St Braughn Box 02131</b> ..... 617 424-6148	
<b>17 879-0030</b> <b>Carter C 200 Commonwealth Ave 02135</b> ..... 617 782-2118	<b>John 11 Mansfield St 02134</b> ..... 617 987-2163	<b>Roy 41 Concord Cam 02138</b> ..... 617 491-6115	
<b>17 436-1511</b> <b>C 218 Harvard Ave East Boston 02128</b> ..... 617 569-1545	<b>John 207 Summer Box 02128</b> ..... 617 423-4334	<b>Royce 18 Sundry Cha 02129</b> ..... 617 241-0418	
<b>17 569-6119</b> <b>C 109 Harvard Cam 02138</b> ..... 617 491-4822	<b>John 40 Hawthorn Elm 02138</b> ..... 617 262-1235		
<b>800 569-4782</b> <b>C 109 Harvard Cam 02138</b> ..... 617 491-4822	<b>June O 129 A Summit Ave 02133</b> ..... 617 734-6109		
	<b>K 17 Irving Ave Cambridge 02142</b> ..... 617 265-8456		
	<b>K 17 Concord Dorchester 02127</b> ..... 617 282-1593		

# Our Idea: Meaning Through Geography

Set of clusterings  $\approx$

A list of unconnected addresses

wide at [SuperPages.com](http://SuperPages.com)

195

Car

C

17 566-1282	Cartage New England Inc 28 Allen St 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	Cartagema Lydia 28 Sweet Briar 02131	617 323-7639	Faye & Ricky 20 Columbia Ave Box 02136	617 437-7331	Nicholas S F 115 Randolph Ave 02136	617 698-5307	
100 257-9961	Cartagema Avish F Beach Rd 02139	617 442-9780	Francis S. 134 Temple W Ave 02132	617 323-6781	Nick & Debbi 21 Farnham 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 Jam 02136	617 524-3078	Nick & Debbi 38 Chickadee Rd 02126	617 822-1203	
361-0380	Luzella 124 Harvard Cam 02136	617 491-5621	Fred W. 96 Valley St 02136	617 698-1343	P E 501 E South St Box 02137	617 268-8213	
17 566-4548	M 95 Howe Box 02132	617 323-9713	G & B. 8 Vardon Ave 02134	617 436-8906	P L 44 Hutchings Box 02131	617 427-9170	
17 628-8248	Melvin 503 Green Cam 02139	617 576-1061	Gayle 25 Franklin Der 02134	617 823-0322	P R 91 Brewer Jan 02138	617 968-8692	
17 445-5116	Carte Nicholas 18 Appleton Boston 02114	617 695-6996	Geo S. 115 Mount Hill Jan 02138	617 522-3215	Paul & Constance 124 Adams Ave W Box 02132	617 325-3034	
17 822-2962	Cartagena O 4 Bradford Box 02133	617 338-0219	George 120 Naves St 02134	617 367-9548	Paul M. 501 E South St Box 02137	617 268-4546	
17 427-5712	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	
17 569-2698	Carte Thos & Kathleen 50 Thompson Ln Mt 02136	617 696-6919	Carter Harry F 30 Bayview Rd W Box 02132	617 325-5465	Patricia Pile Driving Inc 27 Beaver Ct Framingham 02702	978 235-0488	
17 667-5190	A 202 Pines Ave Cambridge 02142	617 492-4174	Carter Hide Co Inc 117 542-7987	Horace 301 Walnut Av Roxbury 02119	617 442-5307	Prudence 40 Franklin Waterbury 02172	617 393-3782
17 569-1417	A M 255 Massachusetts Ave 02115	617 266-7153	Carter Hilary 41 Harvey Cam 02148	617 876-2750	Reginald 100 Brookwood Center 02124	617 541-2843	
17 338-9117	Adams 301 Carter St Mt 02136	617 698-9074	Horace 301 Walnut Av Roxbury 02119	617 442-5307	Renee & Andrew 100 Brookwood Center 02124	617 541-2843	
17 825-9195	Alice 108 Elmwood Ave 02136	617 453-0193	Howard Jr 28 New One Box 02118	617 445-5532	Rice David 30 Walnut Box 02138	617 720-3765	
17 296-1293	Allice 40 Market Cambridge 02139	617 945-2711	J Cam 41 Canton St 02144	617 232-7990	Carl Rice Doon Building Division 163 Main Wilmington 01887	800 638-1671	
17 670-2078	Andrew F 42 Mt St Box 02138	617 625-7623	J Chanen Ne 02146	617 336-2658	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 619-7447	
17 621-9001	Carte Anne MD 1161 Beacon Ave 02144	617 739-1022	J Chanen Ne 02146	617 336-2658	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 648-7447	
17 296-4725	Carte Anne MD 1161 Beacon Ave 02144	617 739-1022	Carte J 1 Brookline Pl Br 02144	617 735-8787	Carl Rice Doon Full Free-Stat 'I' & Then.....	978 988-7447	
17 542-1521	Carte Anne MD 1161 Beacon Ave 02144	617 739-1022	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
17 364-5232	Carte Barbara L MD Tufts New England Medical Center Box 02111	617 296-6911	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
17 541-5649	Carte Beckey Box 02114	617 636-0951	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
17 739-2662	Carte Bernad J 371 Newbury Boston 02116	617 536-6229	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
17 879-0030	Carte Bernad J 371 Newbury Boston 02116	617 536-6229	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
17 541-3948	Carte Bernad J 371 Newbury Boston 02116	617 536-6229	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
17 436-1511	Carte Bernad J 371 Newbury Boston 02116	617 536-6229	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
17 569-4119	Carte Bernad J 371 Newbury Boston 02116	617 536-6229	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	800 638-1673	
100 569-8782	Carte Bernad J 371 Newbury Boston 02116	617 536-6229	Carte J M 3410 Columbia Rd S Box 02136	617 464-1040	Carl Rice Doon Full Free-Stat 'I' & Then.....	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			
17 447-4101	Cartagena Lydia 28 Sweet Briar Rd 02131	617 323-7639			
90 257-9961	Cartagena Avish F Beach Rd 02139	617 442-9780			
17 566-1282	B Had 02136	617 361-5253			
17 364-5188	Lucille 174 Harvard Can 02139	617 491-5621			
361-0380	M 95 Howe Bus 02136	617 323-9713			
17 566-4548	Melvin 503 Green Can 02139	617 576-1061			
17 628-8248	Carte Nicholas 18 Appleton Boston 02114	617 695-6996			
17 445-5116	Carters D 4 Highland Way 02133	617 338-0219			
17 822-2962	Cartes Thos & Sr & Claire 1 Franklin St 02136	617 698-6163			
17 427-5712	Carte A 50 Thompson Ln 02136	617 696-6919			
17 569-2698	A 200 Pitman Av Cambridge 02142	617 492-4174			
17 667-5190	A M 250 Massachusetts Av 02115	617 266-7153			
17 569-1417	Adams 301 Carter St 02138	617 698-9074			
17 338-1101	Adams P 42 West St 02135	617 625-7623			
17 825-1193	Carte Anne MD 1101 Beacon St 02144	617 739-1022			
17 296-1295	Cartier Adeline 971 Newbury Boston 02116	617 536-6229			
17 670-2078	B E 10 Gladstone Av 02136	617 296-6911			
17 621-9001	Cartier Barbara L MD Tufts New England Medical Center 02111	617 432-9001			
17 296-4725	Cartier Becky 02134	617 436-0951			
17 542-1521	Bernard J 1000 Ashburn E 02136	617 523-4368			
17 364-5232	Bibb 25 Midway Rd 02136	617 567-9430			
17 541-5649	Bibb 25 Midway Rd 02136	617 298-8713			
17 739-2662	Cartier Broadcasting Co 50 Park Pl 02136	617 367-9931			
17 879-0030	Cartier C 31 East Can 02141	617 423-0210			
17 541-3948	Cartier C 2000 Commonwealth Av 02135	617 225-0200			
17 436-1511	C 210 Townsend Av East Boston 02128	617 782-2118			
17 569-4119	C 109 Harvard Can 02136	617 569-1545			
909 569-8782	C & M 41 Northgate Jan 02134	617 491-8822			
	C & M 41 Northgate Jan 02134	617 524-9558			
	Carter F 514 Hillside St 02131	617 327-1105			
	Faye & Ricky 20 Columbia Av 02136	617 437-7331			
	Francis S 134 Temple W Av 02132	617 323-6781			
	Franklin & Anne 705 Mt Auburn Can 02138	617 354-0798			
	Fred 41 Howard Jan 02136	617 524-3078			
	Fred 76 Howley Av 02136	617 698-1343			
	G & B 8 Verden Ave 02134	617 436-8906			
	G T 27 Franklin St 02145	617 623-7121			
	Gayle 25 Franklin St 02134	617 823-8322			
	Geo S 115 Mount Mt Jan 02136	617 522-3215			
	George 52 Madison St 02134	617 367-9548			
	Carter Hillside Assoc 107 S Street St 02111	617 456-1689			
	Carter Harry F 30 Bayview Rd W Av 02132	617 325-5465			
	Carter Hide Co Inc 140 Boston St 02131	617 542-7987			
	Carter Hilary 41 Harvey Can 02148	617 876-2750			
	Horace 301 Walnut Av 02139	617 442-5307			
	Howard Jr 28 New One Bus 02118	617 445-5552			
	J Can 15 Chatham St 02144	617 232-7990			
	J 538 Harvard St 02146	617 730-9483			
	J 775 The Pine Way West 02132	617 323-5574			
	Carter J Jacques MD 1 Brookline Pl 02144	617 735-8787			
	Carter J M 3410 Columbia Rd S 02137	617 464-1040			
	Carter J M Ornamental Ironworks 1000 Franklin St 02131	617 436-5353			
	Carter J Neal Co 40 Newbury St 02138	617 442-1775			
	Carter James 1573 Cambridge St Can 02136	617 492-1214			
	James 422 Foster Av 02136	617 739-2193			
	James 31 East Star Rd Cambridge 02141	617 876-8841			
	James 14 Boardwalk Rd Mt 02136	617 361-0773			
	Jan 14 Adams Rd Newton 02458	617 564-0435			
	Jan 1200 Cambridge St 02136	617 426-9094			
	John 11 Mansfield St 02134	617 987-2163			
	John 207 Summer St 02135	617 423-4134			
	John 40 Harvard St 02139	617 282-1235			
	James O 129 A Summit Av 02131	617 734-6109			
	J 29 Harvard St 02134	617 265-8656			
	K 17 Concord Street 02123	617 282-1593			
	Carter Nellie E 323 Marchant Av 02115	617 267-6483			
	Nicholas S F 115 Randolph Av 02136	617 698-5307			
	Nick 21 Fyfield Bus 02116	617 267-5222			
	Nick & Debbi 136 Hermit Rd Newton 02459	617 527-0480			
	Norman G 38 Chickadee Dr 02126	617 822-1201			
	P 41 Eastwood Pl 02135	617 427-4754			
	P E 501 E South S 02137	617 268-8213			
	P L 44 Hutchings Bus 02131	617 427-9170			
	P R 91 Boyer Jan 02138	617 968-8692			
	Paul & Constance 114 Adams Av W 02133	617 325-3034			
	Paul F 501 E South S 02137	617 268-4546			
	Paul M 27 Union St 02139	617 787-2115			
	Carter Pike Driving Inc 27 Beacon St Framingham 02170	Wellesley Tpk-781.235-0488			
	Carter Prudence 40 Franklin Waterbury 02172	617 393-3782			
	Prudence 40 Franklin Waterbury 02172	617 926-7063			
	Reginald 100 Broadview Center 02124	617 541-2843			
	Reed & Andrew 30 Walnut St 02138	617 720-3765			
	Carter Rice David Building Division 163 Main Wilmington 01887 Toll Free 800 713 7386	800 638-1671			
	Toll Free 800 713 7386	800 619-7447			
	Toll Free 800 713 7386	800 648-7447			
	Frederick 413 Main Wilmington 01887	978 988-7447			
	Ingalls Crane 163 Main Wilmington 01887	800 638-1673			
	Carter Richard 2077 Carver Av Brighton 02131	617 987-0836			
	Carter Richard A MD 41 W Vernon St 02136	617 566-7293			
	Richard A 1200 Cambridge St 02136	617 267-0710			
	Carter Richard K 123 Merwin St 02137	617 268-0468			
	Robert L 175 Rockwood Av Can 02141	617 864-1535			
	Royce 130 Brattle St 02131	617 424-6148			
	Royce & Andrew 18 Salisbury Cir 02129	617 491-6115			
	Royce 18 Salisbury Cir 02129	617 241-9418			



$\approx$  We develop a (conceptual) geography of clusterings



# A New Strategy

Make it easy to choose best clustering from millions of choices

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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



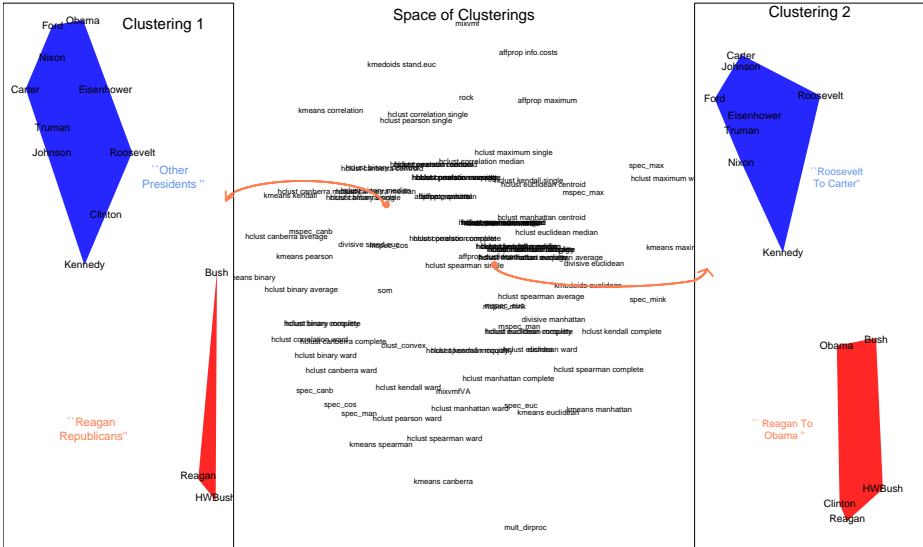
# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# Many Thousands of Clusterings, Sorted & Organized

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

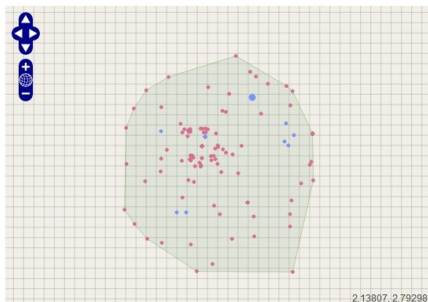


# Software Screenshot

Size: 244 Files

Description: NSF - Updated Set

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



Display History   Display Method Points

Label	Coordinates	Clusters
<a href="#">an interesting clustering [Link]</a>	-0.30819, 0.46229	5
<a href="#">methods-oriented clustering [Link]</a>	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)

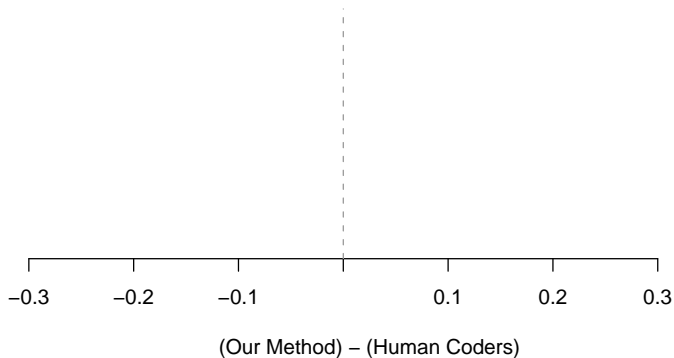
# Evaluation 1: Cluster Quality

- **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

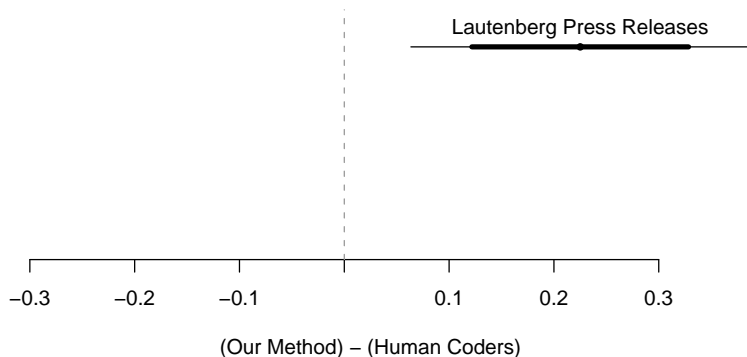
- **What Are Humans Good For?**
  - They can't: keep many documents & clusters in their head
  - They can: compare two documents at a time
  - $\implies$  Cluster quality evaluation: human judgement of document pairs
- **Experimental Design to Assess Cluster Quality**
  - automated visualization to choose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



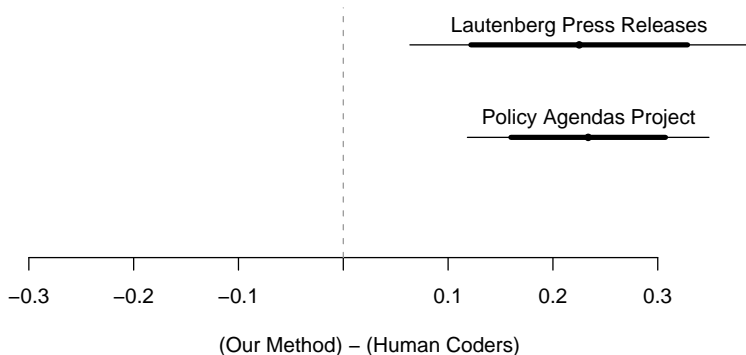


# Evaluation 1: Cluster Quality



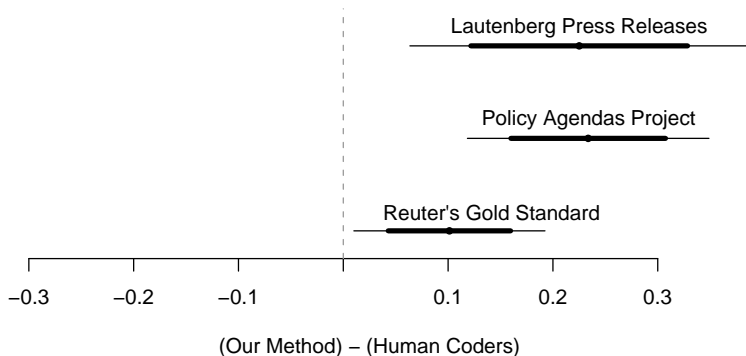
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

# 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



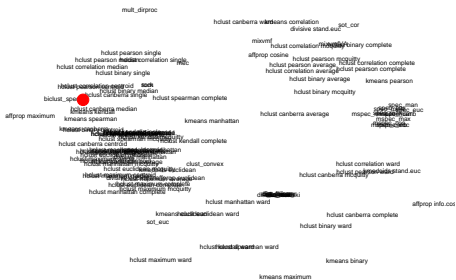
Red point: a **clustering** by Affinity Propagation-Cosine (Dueck and Frey 2007)

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





# Example Discovery



Space between methods:

# Example Discovery



Space between methods:  
local cluster ensemble















# Example Discovery



## Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering  
Random Walk  
(Metrics 1-6)

0.13 Hclust-Correlation-Ward

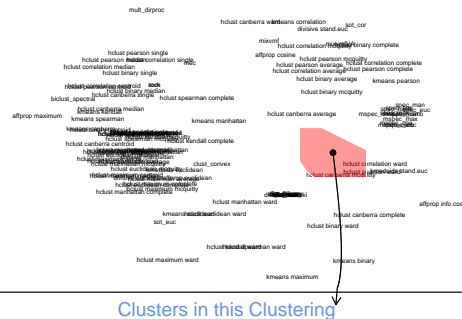
0.09 Hclust-Pearson-Ward







# Example Discovery

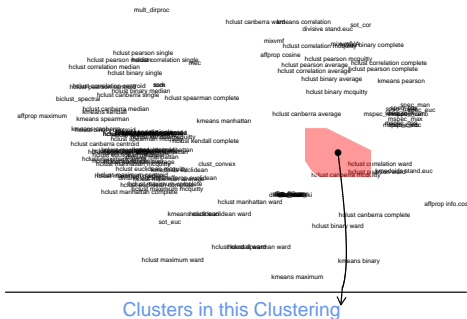


Credit Claiming  
Pork

## Credit Claiming, Pork:

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

# Example Discovery



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



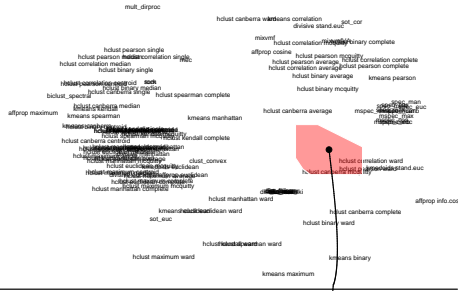
Credit Claiming  
Pork



Mayhew Credit Claiming  
Legislation

Gary King (Harvard IQSS)

# Example Discovery



Clusters in this Clustering



Credit Claiming  
Pork

Advertising



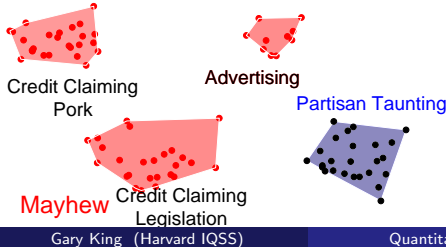
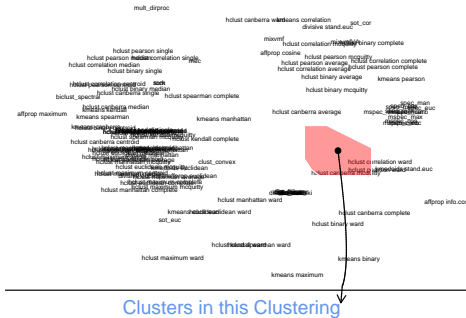
Mayhew  
Credit Claiming  
Legislation  
Gary King (Harvard IQSS)

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





# Example Discovery: Partisan Taunting



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



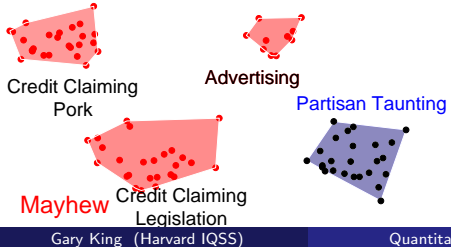
# Example Discovery: Partisan Taunting



Clusters in this Clustering

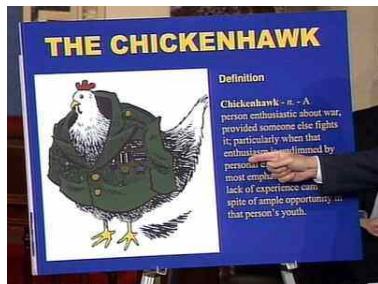
**Definition:** Explicit, public, and negative attacks on another political party or its members

**Taunting ruins deliberation**



# In Sample Illustration of Partisan Taunting

## Taunting ruins deliberation

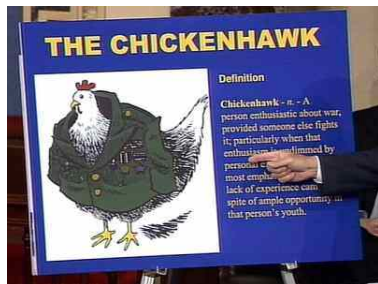


Sen. Lautenberg  
on Senate Floor  
4/29/04

- "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]

## 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

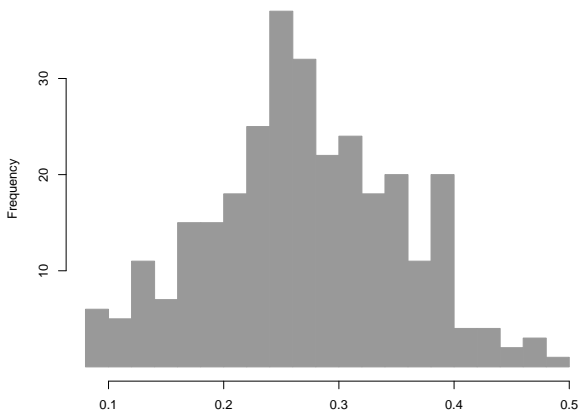
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

# 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

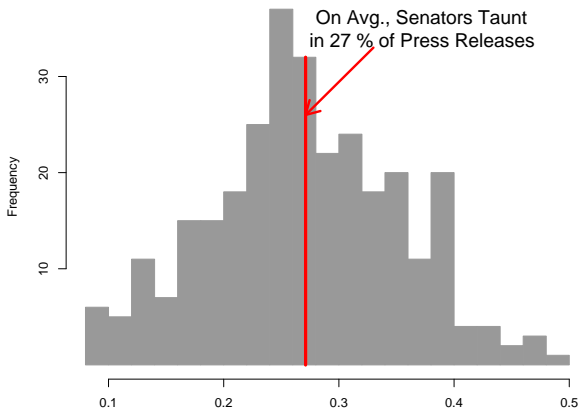
# Out of Sample Confirmation of Partisan Taunting

- 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

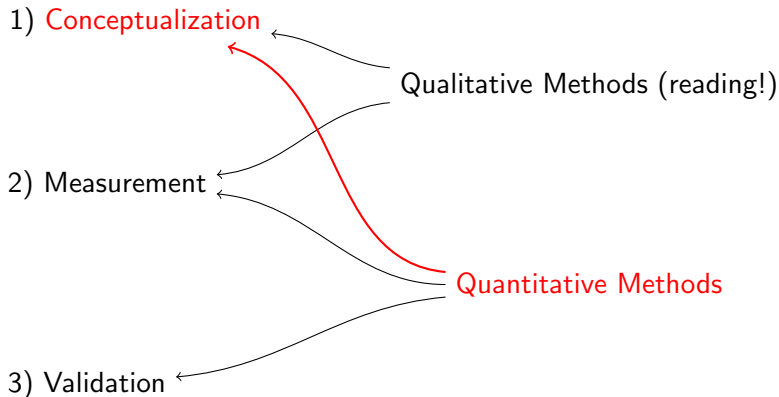


# Out of Sample Confirmation of Partisan Taunting

- 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

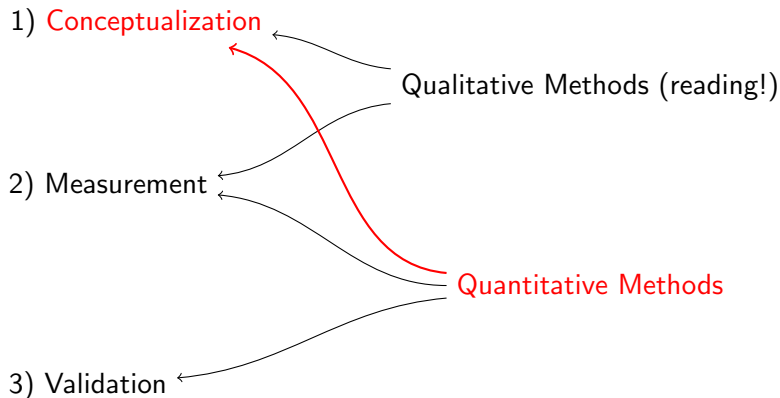


# Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

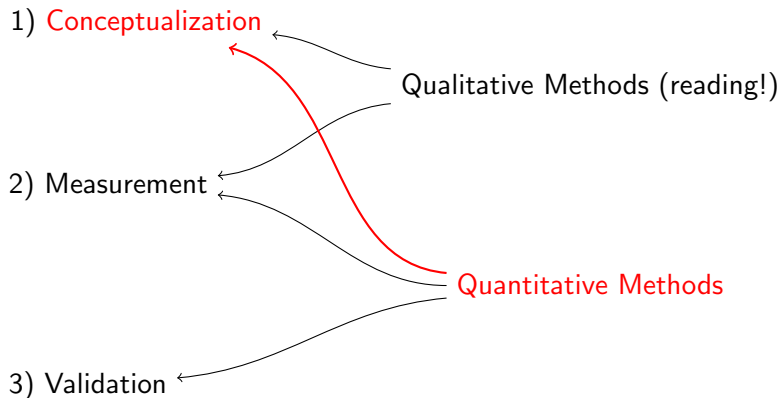
# Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

- Few formal methods designed explicitly for conceptualization

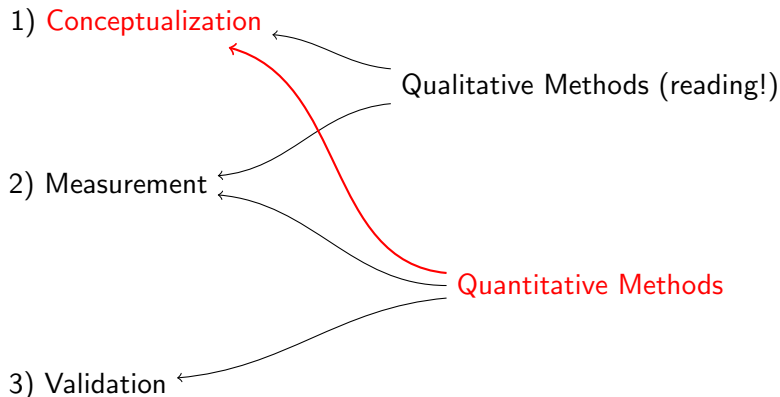
# Quantitative Methods for Qualitative Conceptualization



## 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>