

# Computer-Assisted Clustering and Conceptualization from Unstructured Text

Gary King

Institute for Quantitative Social Science  
Harvard University

Talk at the Center for Research on Computation and Society, Harvard University, 3/7/2011

---

<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**

195

Car

C

<b>17 566-1282</b>	<b>Cartage New England Inc</b> 28 Allen Ln Ipswich 01938	978 356-9960	<b>Carter F</b> 24 Hibiscus Bn 02133	617 327-1105	<b>Carter Nella E</b> 323 Mainville Av Bn 02135	617 267-6483
<b>17 447-4101</b>	<b>Cartagena Lydia</b> 28 Sweet Box 02131	617 323-7639	<b>Faye &amp; Ricky</b> 20 Columbia Av Bn 02136	617 437-7331	<b>Nicholas S F</b> 115 Randolph Av Bn 02136	617 698-5307
<b>800 257-9961</b>	<b>Cartagena Avish</b> F Pleasant Bn 02139	617 442-9780	<b>Francis S</b> 134 Yankov W Av 02132	617 323-6781	<b>Nick 21 Farwell Bn 02136</b>	617 267-5222
<b>17 566-1282</b>	<b>B Hrd 02136</b>	617 361-5253	<b>Franklin &amp; Anne</b> 705 Mt Auburn Cn 02138	617 354-0798	<b>Nicole</b> 196 Hemlock Rd Newton 02459	617 527-0480
<b>17 364-5188</b>	<b>Justicia</b> 50 Decatur Cha 02129	617 241-0152	<b>Fred 41 Woodland Jn 02138</b>	617 524-3878	<b>Norman G</b> 38 Chickawhnd Dr 02125	617 822-1201
<b>361-0380</b>	<b>Luzmila</b> 174 Harvard Cn 02138	617 491-5621	<b>Fred 96 Holyoke Av Bn 02138</b>	617 698-1343	<b>P 40 Cranford Pl Bn 02135</b>	617 437-4754
<b>17 566-4548</b>	<b>M 95 Howe Bn 02132</b>	617 323-9713	<b>G &amp; B 8 Vardon Bn 02134</b>	617 434-8966	<b>P E 501 E South S Bn 02137</b>	617 268-8213
<b>17 628-8248</b>	<b>Melvin</b> 503 Green Cn 02129	617 576-1061	<b>Gayle</b> 25 Franklin Dr 02133	617 825-8232	<b>P L 44 Hutchings Bn 02131</b>	617 427-9170
<b>17 445-5116</b>	<b>Carte Nicholas</b> 18 Apollonia Boston 02116	617 695-6996	<b>Geo S</b> 115 Mass Mt Hill Rd Jn 02138	617 522-3215	<b>P R 81 Boyer Avn 02138</b>	617 968-8692
<b>17 822-2992</b>	<b>Carlton</b> 0 4 Bedford Av 02133	617 338-9219	<b>George</b> 125 Madison Bn 02134	617 367-9548	<b>Paul &amp; Constance</b> 114 Freeman St W Bn 02133	617 325-2036
<b>17 447-4101</b>	<b>Carter Thos Jr Sr &amp; Claire</b> 17 Lowell Rd Mt 02136	617 698-6163	<b>Carter Hillside Associa</b> 107 S Street Bn 02111	617 456-1689	<b>Paul M</b> 501 E South St S Bn 02137	617 268-4546
<b>17 822-2992</b>	<b>Thomas &amp; Kathleen</b> 50 Thompson Ln Mt 02136	617 696-6919	<b>Carter Harry F</b> 26 Bayne Rd Rt W Av 02132	617 325-5465	<b>Paul M</b> 27 Crown Bk 02139	617 787-2115
<b>17 427-5712</b>	<b>Carter A Av 02133</b>	617 297-2257	<b>Carter Hide Co Inc</b> 100 Franklin St 02148	617 542-7987	<b>Prudence</b> Prudence Dr 702	617 926-7063
<b>17 569-2698</b>	<b>A Nebra</b> A 23 Bethune Wy Rosbury 02139	617 442-5230	<b>Carter Hilary</b> 41 Harvey Cn 02148	617 876-2750	<b>Ronald</b> 100 Brookwood Dr 02122	617 541-2843
<b>17 667-5190</b>	<b>A 23 Bethune Wy Rosbury 02139</b>	617 442-1219	<b>Horace</b> 361 Walnut Av Rosbury 02139	617 442-5307	<b>Renee &amp; Andrew</b> 30 Walnut Bn 02138	617 720-3765
<b>17 569-1417</b>	<b>A M 203 Massachusetts Av Bn 02135</b>	617 266-7153	<b>Howard Jr</b> 28 Neve Drive Bn 02118	617 445-5532	<b>Rice Dore</b> Bullfinch Business Publishing 163 Main Wilmington 01887	800 638-1671
<b>17 338-9110</b>	<b>Adams</b> 361 Centre St Mt 02136	617 698-9074	<b>J Dan</b> 41 Chestnut Bn 02146	617 232-7990	<b>Richard A M</b> 2077 Cavendish Av Brighton 02115	617 987-8836
<b>17 825-1993</b>	<b>Alice</b> 108 Elmwood Bn 02133	617 425-0193	<b>J 23 Chatham Bn 02146</b>	617 232-7990	<b>Richard A M</b> 197 Mt Vernon Bn 02106	617 566-7293
<b>17 825-1993</b>	<b>Alice C</b> 40 Market Cambridge 02139	617 945-2711	<b>J 538 Harvard Bn 02146</b>	617 730-9483	<b>Richard R K M</b> 130 Conventry Pl Bn 02136	617 267-0710
<b>17 670-2078</b>	<b>Andrew F</b> 42 West St Bn 02138	617 625-7623	<b>J 775 The Pine Wood Rosbury 02132</b>	617 323-5374	<b>Richard R K M</b> 23 Mersey St Bn 02137	617 268-0448
<b>17 621-9001</b>	<b>Carter Anne MD</b> 1165 Beacon Bn 02146	617 739-1022	<b>J Jacques MD</b> 1 Breckin Pl Bn 02146	617 735-8787	<b>Roger</b> 130 St Bourgh Bn 02131	617 424-6148
<b>17 296-1593</b>	<b>Carter J M</b> 371 Newbury Boston 02116	617 536-6329	<b>Carter J D</b> 3410 Columbia Rd S Bn 02137	617 464-1040	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 670-2078</b>	<b>B E 18 Gladstone Av Mt 02136</b>	617 296-6911	<b>Carter J M Ornamental Interworks</b> 100 Franklin St 02148	617 876-5353	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 621-9001</b>	<b>Carter Barbara L MD</b> Tufts-New England Medical Center Bn 02111	617 636-0051	<b>Carter J Neal Co</b> 40 Woodland Bn 02138	617 442-1775	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 296-4725</b>	<b>Carter Becky Jo 02134</b>	617 523-4368	<b>Carter James</b> 157 Cambridge St Cam 02138	617 492-1214	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 542-1521</b>	<b>Bernard J</b> 122 Goodhue F Bn 02138	617 567-9430	<b>James</b> 412 Foster Av Rosbury 02138	617 739-2193	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 364-5232</b>	<b>Bethiah 25 Midway Dr 02136</b>	617 298-8713	<b>James L</b> 34 Rosbury Rd Mt 02134	617 876-8841	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 541-5249</b>	<b>Bible 28 Elmwood Bn 02133</b>	617 367-9051	<b>Jane 14 Adams Rd Newton 02465</b>	617 964-0435	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 739-2662</b>	<b>Carter Broadcasting Co</b> 28 Park Pl Bn 02136	617 423-0210	<b>Jeffrey C</b> 41 Barnes Av Bn 02136	617 426-5094	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 879-0830</b>	<b>Carter &amp; Business Consultants Inc</b> 73 East St Cam 02141	617 225-0200	<b>John 107 Summer Bn 02135</b>	617 423-4334	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 541-3948</b>	<b>Carter C 200 Cavendish Av Bn 02135</b>	617 782-2118	<b>John 40 Woodland Dr 02138</b>	617 282-1235	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 436-1511</b>	<b>C 218 Harvard Av East Boston 02128</b>	617 569-1545	<b>June O</b> 129 A Summit Av Bn 02133	617 734-6109	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>17 569-4119</b>	<b>C 109 Harvard Cn 02138</b>	617 491-8522	<b>June O</b> 129 A Summit Av Bn 02133	617 734-6109	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418
<b>800 569-8782</b>	<b>C 218 Harvard Av East Boston 02128</b>	617 569-1545	<b>K 17 Elwood Dr 02132</b>	617 282-1293	<b>Royce</b> 18 Sawbury Cha 02129	617 241-0418

# Our Idea: Meaning Through Geography

Set of clusterings  $\approx$

A list of unconnected addresses

wide at SuperPages.com

195

Car

C

17 566-1282 Cortage New England Inc 28 Allen Ln Ipswich 01938..... 978 356-9960	17 327-1105 Carter F. 514 Hickox Ave 02131..... 617 327-1105	17 267-6483 Carter Nella E 323 Marchant Ave Box 02115..... 617 267-6483
17 447-4101 Cortage Lydia 28 Sweet Briar 02131..... 617 323-7639	17 437-7331 Faye & Ricky 20 Columbia Ave Box 02136..... 617 437-7331	617 698-5307 Nicholas S F 115 Randolph Ave 02136..... 617 698-5307
17 257-9961 Cortageva Avish F Beach Rd Box 02139..... 617 442-9780	617 354-0798 Franklin & Anne 705 Mt Auburn Cam 02138..... 617 354-0798	617 267-5222 Nick & Debbi 215 Fyfield Box 02114..... 617 267-5222
17 566-1282 Lucille 174 Harvard Cam 02139..... 617 491-5621	617 524-3078 Fred 41 Harvard Cam 02138..... 617 524-3078	617 698-0713 Norman G 196 Hermit Rd Newton 02459..... 617 698-0713
17 364-5188 M 90 Howe Box 02139..... 617 323-9713	617 698-1343 Fred 96 Harvard Ave Box 02138..... 617 698-1343	617 822-1203 38 Chickadee Rd Der 02125..... 617 822-1203
361-0380 Mehin 503 Green Cam 02139..... 617 576-1061	617 825-0322 Gayle 25 Franklin Der 02124..... 617 825-0322	617 427-4754 P E 501 E South St Box 02137..... 617 427-4754
17 566-4548 Carte Nicholas 18 Appleton Boston 02114..... 617 695-6996	617 522-3215 Geo S 115 Mount Hill Rd Box 02134..... 617 522-3215	617 268-4813 P L 44 Hutchings Box 02115..... 617 268-4813
17 628-8248 Carten Thos J Sr & Claire 174 Appleton Ave 02114..... 617 338-9219	617 367-9548 George 125 Boston Ave 02114..... 617 367-9548	617 968-8692 Paul & Constance 124 Appaman Ave Box 02110..... 617 968-8692
17 445-5116 Thos & Kathleen 50 Thompson Ln Mt 02136..... 617 698-6163	617 456-1689 Carter Holiday Assoc 107 S Street Box 02111..... 617 456-1689	617 225-3034 Paul E 501 E South St Box 02137..... 617 225-3034
17 822-2962 Carter A Box 02131..... 617 229-2257	617 325-5465 Carter Hide Co Inc 140 Bayview Rd W Box 02112..... 617 325-5465	617 268-4546 Paul M 27 Crown St 02135..... 617 268-4546
17 427-5712 A Heber 17 442-5230 A 22 Bethune Wy Redbury 02119..... 617 442-5230	617 542-7987 Carter Hilary 41 Harvey Cam 02148..... 617 876-2750	617 787-2115 Carter Pike Driving Inc 27 Beaver Ct Framingham 02170..... 617 787-2115
17 569-2698 A 200 Pioneer Av Cambridge 02142..... 617 492-4174	Horace 301 Walnut Av Redbury 02119..... 617 442-5307	617 393-3782 Carter Prudence 40 Franklin Waterbury 02172..... 617 393-3782
17 667-5190 A M 255 Main St Av Box 02115..... 617 266-7153	617 445-5532 Howard Jr 28 New One Box 02118..... 617 445-5532	617 926-7063 Prudence 40 Franklin Waterbury 02172..... 617 926-7063
17 569-1417 Adams 31 Carter St Mt 02136..... 617 698-9074	617 232-2668 J C 15 Chatham Ave 02144..... 617 232-2668	617 541-2843 Reginald 100 Broadview Center 02121..... 617 541-2843
17 338-1141 Alice 40 Market Cambridge 02139..... 617 945-2711	617 730-9483 J S 4775 The Pines West Redbury 02125..... 617 730-9483	617 720-3765 Renée & Andrew 100 Walnut Box 02118..... 617 720-3765
17 825-9195 Andrea F 42 West St Box 02135..... 617 625-7623	617 735-8787 Carter J Jacques MD 1 Crockett Pl Br 02144..... 617 735-8787	617 800-6371 Carter Rice Doan 154 Dunton Publishing 163 Main Wilmington 01887 Tud Free-Old 'J' & Thom..... 800 638-1671
17 296-1293 1101 Beacon Ave 02144..... 617 739-1022	617 464-1040 Carter J M 3410 Columbia Rd Box 02138..... 617 464-1040	617 744-7447 Carter J M 101 Franklin Waterbury 02172..... 617 744-7447
17 670-2078 B E 10 Gladstone Ave Mt 02136..... 617 296-6911	617 436-5353 Carter J M Ornamental Ironworks Pondville Falls 02176..... 617 436-5353	617 648-7447 Tud Free-Old 'J' & Thom..... 600 648-7447
17 621-9001 Carter Barbara L MD Tufts New England Medical Center Box 02111 777 State St Boston 02114..... 617 636-0951	617 442-1775 Carter J Neal Co 40 Newmarket Box 02118..... 617 442-1775	617 978-7447 Carter Ingalls Circle 163 Main Wilmington 01887 978 978-7447
17 296-4725 Carter Becky Box 02114..... 617 523-4368	617 492-1214 James 1573 Cambridge St Cam 02138..... 617 492-1214	617 638-1673 Carter Richard 2079 Carver Ave Brighton 02111..... 617 638-1673
17 542-1521 Bernard J 3000 Ashburne Rd 02138..... 617 567-3430	617 876-8841 James 31 East Star Rd Cambridge 02141..... 617 876-8841	617 566-7293 Carter Richard A MD 2079 Carver Ave Brighton 02111..... 617 566-7293
17 364-5232 Bibb 25 Midway Der 02124..... 617 298-8713	617 361-0773 Jane L 34 Rosbury Rd Mt 02136..... 617 361-0773	617 267-0710 Carter Richard R 1200 Carver Ave Brighton 02111..... 617 267-0710
17 541-5649 Bill 30 W Ames Ave 02138..... 617 367-9931	617 964-0435 Jane 14 Adams Rd Newton 02459..... 617 964-0435	617 268-0448 Carter Richard R MD 1200 Carver Ave Brighton 02111..... 617 268-0448
17 739-2662 Carter Broadcasting Co 50 Park Pl Box 02114..... 617 423-0210	617 426-5094 John 11 Mansfield Dr 02134..... 617 426-5094	617 864-1535 Robert L 175 Rockwood Ave Cam 02141..... 617 864-1535
17 879-0030 Carter C 2000 Gesswilt Ave 02135..... 617 225-0200	617 423-4134 John 207 Summer St 02125..... 617 423-4134	617 424-6148 Roger 130 St Branhg Box 02111..... 617 424-6148
17 541-3948 C 210 Harvard Ave East Boston 02128..... 617 569-1545	617 282-1275 John 40 Harvard Ave 02128..... 617 282-1275	617 491-6115 Robert L 175 Rockwood Ave Cam 02141..... 617 491-6115
17 569-4119 C 109 Harvard Cam 02138..... 617 491-4822	617 265-4956 James O 129 A Summit Av Br 02113..... 617 265-4956	617 241-9418 Royce 18 Sapparyn Cir 02129..... 617 241-9418
17 569-8782 C & M 41 Northgate Box 02114..... 617 524-9558	617 282-1593 K 17 Concord Der 02127..... 617 282-1593	



# 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 02331	617 323-7639
100 257-9961	Cartagena Avish F Beach Rd 02319	617 442-9780
	B Had 02334	617 361-5253
17 566-1282	Justicia 50 Decatur Cha 02329	617 241-0152
17 364-5188	Luzmila 124 Harvard Can 02334	617 491-5621
	M 95 Howe Box 02334	617 323-9713
361-0380	Melvin 503 Green Can 02329	617 576-1061
17 566-4548	Carte Nicholas 18 Appleton Boston 02314	617 695-6996
	Cartagena O 4 Harvard Box 02334	617 338-9219
17 628-8248	Carten Thos J Sr & Claire 1 Furlow St Mt 02336	617 698-6163
17 445-5116	Thomas & Kathleen 50 Thompson Ln Mt 02336	617 696-6919
17 822-2962	Carter A Box 02334	617 339-2257
17 427-5712	A Heber A 200 Pinetree Av Cambridge 02328	617 442-5230
17 569-2698	A 200 Pinetree Av Cambridge 02328	617 442-5230
17 667-5190	A M 250 Massachusetts Av Box 02311	617 442-1219
	Adams 301 Carter St Mt 02336	617 698-7074
17 569-1417	Allice 108 Elmwood Av Box 02311	617 423-0193
17 338-9110	Allice 40 Market Cambridge 02334	617 945-2711
17 825-1993	Andrew F 42 West St Box 02334	617 625-7623
17 825-1993	Carter Anne MD 1101 Beacon St 02444	617 739-1022
17 296-1193	Carter Atlanta 971 Newbury Boston 02318	617 536-6239
17 670-2078	B E 10 Gladstone Av Mt 02316	617 296-6911
17 621-9001	Carter Barbara L MD Tufts New England Medical Center Box 02331	
17 296-4725	Cal Carter Becky 02314	617 636-0951
	Call Carter Becky 02314	617 523-4368
17 542-1521	Bernard J 3000 Ashburn E Rd 02336	617 567-9430
17 364-5232	Bibbith 25 Midway Dr 02334	617 298-8713
17 541-5649	Bill 30 W Newbury St 02336	617 367-9931
17 739-2662	Carter Broadcasting Co 50 Park Pl Box 02316	617 423-0210
	Carter Business Consultants Inc 73 East C St 02341	617 225-0200
17 879-0030	Carter C 2000 Cavendish Av St 02335	617 782-2118
17 541-3948	C 210 Townsend Av East Boston 02338	617 569-1545
17 436-1511	C 109 Harvard Can 02336	617 491-4822
17 569-4119	C 109 Harvard Can 02336	617 491-4822
100 02311	C & M 41 Northgate Jct 02334	617 524-9392
100 869-8782	C & M 41 Northgate Jct 02334	617 524-9392
	Carter F 24 Hibiscus Box 02334	617 327-1105
	Faye & Ricky 20 Columbia Av Box 02334	617 437-7331
	Francis S 134 Temple W Av Box 02334	617 323-6781
	Franklin & Anne 705 Mt Auburn Can 02336	617 354-0798
	Fred 40 Harvard Av 02336	617 524-3078
	Fred 76 Newbury Av Mt 02336	617 698-1343
	G & B 8 Harvard Box 02334	617 436-8906
	G T 27 Fossil Hill Av Box 02345	617 623-7121
	Gayle 25 Franklin St 02334	617 825-8322
	Geo S 115 Mount Mt Jct Box 02336	617 522-3215
	George 52 Madison Box 02314	617 367-9548
	Carter Hillside Assoc 107 S Street Box 02311	617 456-1689
	Carter Harry F 30 Bayview Rd W Av Box 02311	617 325-5465
	Carter Hide Co Inc 167 Essex St 02334	617 542-7987
	Carter Hilary 41 Harvey Can 02348	617 876-2750
	Horace 301 Walnut Av Newbury 02334	617 442-5307
	Howard Jr 28 New One Box 02316	617 445-5552
	J Can 15 Chatham St 02444	617 232-7990
	J 538 Harvard St 02444	617 730-9483
	J 775 The Pine Way Westbury 02334	617 323-5374
	Carter J Jacques MD 1 Brookline Pl Box 02444	617 735-8787
	Carter J M 3410 Columbia Rd S Box 02337	617 464-1040
	Carter J M Ornamental Ironworks 300 Franklin St 02334	617 436-5353
	Carter J Veal Co 40 Newbury St 02336	617 442-1775
	Carte James 1573 Cambridge St Can 02336	617 492-1214
	James 422 Foster Av Newbury 02336	617 739-2193
	James 31 East Star Rd Cambridge 02318	617 876-8841
	J 34 Newbury Rd Mt 02336	617 361-0773
	Jane 14 Adams Rd Newbury 02445	617 964-0435
	Janey 120 Cambridge St Mt 02336	617 426-9094
	John 11 Mansfield St 02334	617 987-2163
	John 207 Summer St 02334	617 423-4334
	John 40 Newbury St 02334	617 282-1235
	James O 129 A Summit Av Box 02334	617 734-6109
	J 29 Newbury St 02334	617 265-4956
	K 17 Concord Road 02334	617 282-1593
	Carter Nellie E 323 Main St Box 02315	617 267-6483
	Nicholas S F 115 Randolph Av Mt 02336	617 698-5307
	Nick 21 Furlow Box 02316	617 267-5222
	Nick & Debbi 136 Hermit Rd Newbury 02449	617 527-0480
	Norman G 38 Chickadee Dr 02326	617 822-1201
	P 40 Cambridge Pl Box 02334	617 427-4754
	P E 501 E South S Box 02337	617 268-4213
	P L 44 Hutchings Box 02311	617 427-9170
	P R 91 Bayview Can 02336	617 968-8692
	Paul & Constance 114 Beacon St W Mt 02333	617 325-3034
	Paul F 501 E South S Box 02337	617 268-4546
	Paul M 27 Union St 02336	617 787-2115
	Carter Pike Driving Inc 27 Beaver Ct Franklin 02334	Wellesley Tpk-781.235-0488
	Carter Prudence 40 Franklin Waterbury 02327	617 393-3782
	Prudence 40 Franklin Waterbury 02327	617 926-7063
	Reginald 100 Brookside Circle 02324	617 541-2843
	Renée & Andrew 30 Walnut St 02338	617 720-3765
	Carter Rice David Building Division Publishing 163 Main Wilmington 01887 Toll Free-Dial '7 & Then.....800.638-1671 Toll Free-Dial '7 & Then.....800.619-7447 Toll Free-Dial '7 & Then.....800.648-7447 Toll Free-Dial '7 & Then.....978.988-7447 Ingalls Centre 163 Main Wilmington 01887 800.638-1673	
	Carter Richard 2079 Cavendish Av Brighton 02321	617 987-0836
	Richard A 97 W Vernon St 02336	617 566-7293
	Carter Richard A 120 Cambridge St Mt 02336	617 267-0710
	Carter Richard K 123 Mount S Box 02337	617 268-0468
	Robert L 175 Newbury Av Can 02341	617 864-1535
	Royce 130 St Brnagh Box 02311	617 424-6148
	Royce & Andrew 18 Springdale Cha 02329	617 491-6115
	Royce 18 Springdale Cha 02329	617 241-9418



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



# A New Strategy

Make it easy to choose best clustering from millions of choices

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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

# A New Strategy

Make it easy to choose best clustering from millions of choices

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



# A New Strategy

Make it easy to choose best clustering from millions of choices

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

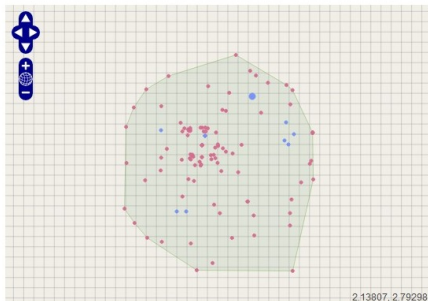


# Software Screenshot

Size: 244 Files

Description: NSF - Updated Set

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



Display History   Display Method Points

Label	Coordinates	Clusters
an interesting clustering [Link]	-0.30819, 0.46229	5
methods-oriented clustering [Link]	0.84753, 1.42538	5

(\*) Discoverable

Coordinates: 0.84753, 1.42538

Clusters: 5

Label [+] methods-oriented clustering

29.51%  research community health science public practice global political national urban  
72  
Label [+]

[View Detail](#)

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

[View Detail](#)

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

[View Detail](#)

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

[View Detail](#)

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

[View Detail](#)

# Application-Independent Distance Metric: Axioms

# Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions

# Application-Independent Distance Metric: Axioms

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

# Application-Independent Distance Metric: Axioms

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

# Application-Independent Distance Metric: Axioms

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



# Application-Independent Distance Metric: Axioms

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

# Application-Independent Distance Metric: Axioms

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

# Evaluating Performance

# Evaluating Performance

- Goals:

# Evaluating Performance

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

# Evaluating Performance

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

# Evaluating Performance

- Goals:
  - **Validate Claim**: computer-assisted conceptualization outperforms human conceptualization
  - **Demonstrate**: new experimental designs for cluster evaluation
  - **Inject human judgement**: relying on insights from survey research

# Evaluating Performance

- Goals:
  - **Validate Claim**: computer-assisted conceptualization outperforms human conceptualization
  - **Demonstrate**: new experimental designs for cluster evaluation
  - **Inject human judgement**: relying on insights from survey research
- We now present three evaluations



# Evaluating Performance

- Goals:
  - **Validate Claim**: computer-assisted conceptualization outperforms human conceptualization
  - **Demonstrate**: new experimental designs for cluster evaluation
  - **Inject human judgement**: relying on insights from survey research
- We now present three evaluations
  - Cluster Quality  $\Rightarrow$  RA coders

# Evaluating Performance

- Goals:
  - **Validate Claim**: computer-assisted conceptualization outperforms human conceptualization
  - **Demonstrate**: new experimental designs for cluster evaluation
  - **Inject human judgement**: relying on insights from survey research
- We now present three evaluations
  - Cluster Quality  $\Rightarrow$  RA coders
  - Informative discoveries  $\Rightarrow$  Experienced scholars analyzing texts

# Evaluating Performance

- Goals:
  - **Validate Claim**: computer-assisted conceptualization outperforms human conceptualization
  - **Demonstrate**: new experimental designs for cluster evaluation
  - **Inject human judgement**: relying on insights from survey research
- We now present three evaluations
  - Cluster Quality  $\Rightarrow$  RA coders
  - Informative discoveries  $\Rightarrow$  Experienced scholars analyzing texts
  - Discovery  $\Rightarrow$  You're the judge

# Evaluation 1: Cluster Quality

# Evaluation 1: Cluster Quality

- What Are Humans Good For?

# Evaluation 1: Cluster Quality

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

# Evaluation 1: Cluster Quality

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

# Evaluation 1: Cluster Quality

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



# Evaluation 1: Cluster Quality

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

# Evaluation 1: Cluster Quality

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

# Evaluation 1: Cluster Quality

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

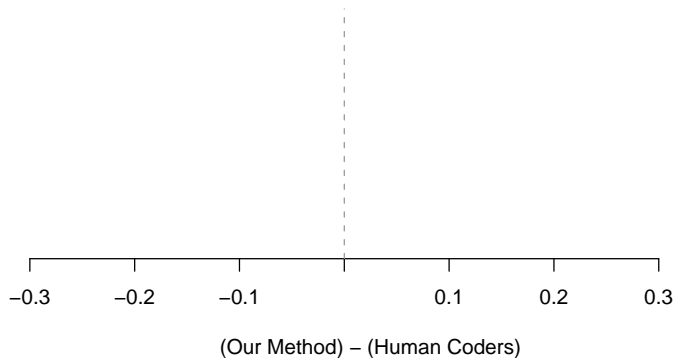
# Evaluation 1: Cluster Quality

- **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

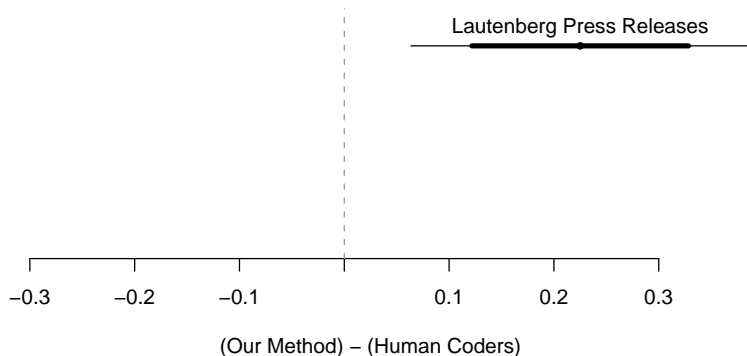
- **What Are Humans Good For?**
  - They can't: keep many documents & clusters in their head
  - They can: compare two documents at a time
  - $\implies$  Cluster quality evaluation: human judgement of document pairs
- **Experimental Design to Assess Cluster Quality**
  - automated visualization to choose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



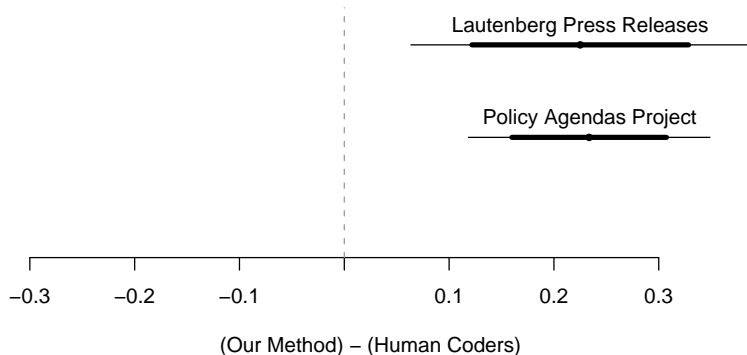


# Evaluation 1: Cluster Quality



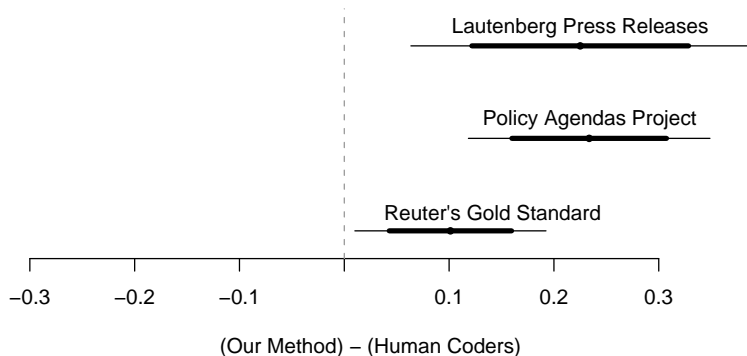
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

# 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

































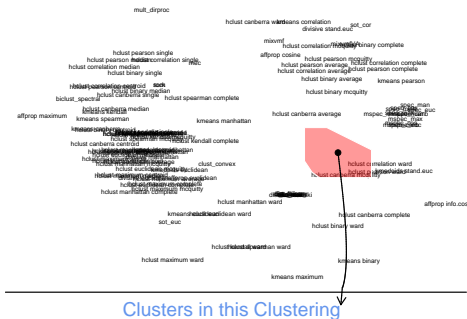








# Example Discovery



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



Credit Claiming  
Pork

Advertising

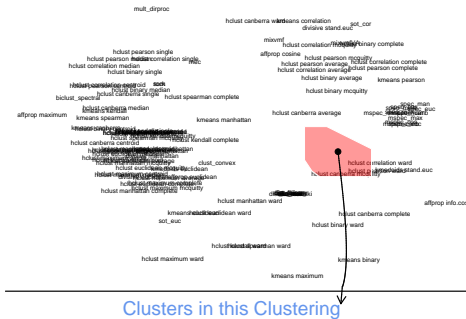


Mayhew  
Credit Claiming  
Legislation

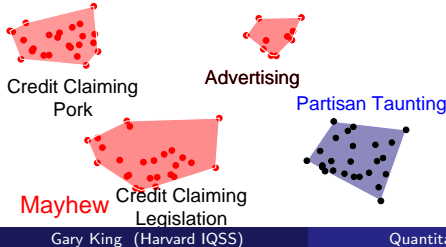
Gary King (Harvard IQSS)



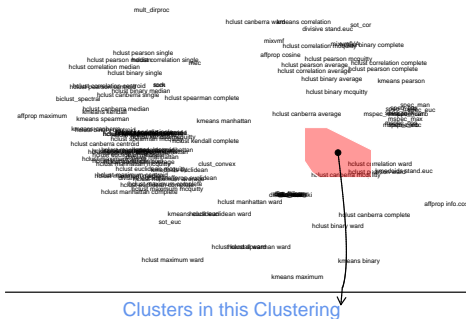
# Example Discovery: Partisan Taunting



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



# Example Discovery: Partisan Taunting



Credit Claiming  
Pork

Advertising

Partisan Taunting

Mayhew  
Credit Claiming  
Legislation

Gary King (Harvard IQSS)

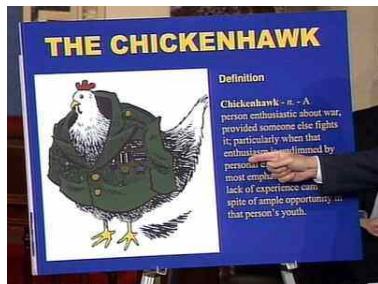
## Partisan Taunting:

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





## Taunting ruins deliberation

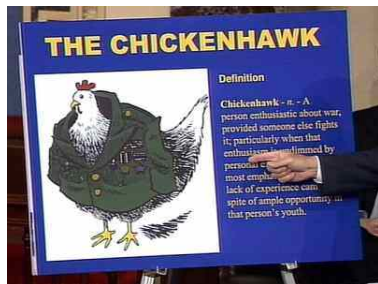


Sen. Lautenberg  
on Senate Floor  
4/29/04

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

# In Sample Illustration of Partisan Taunting

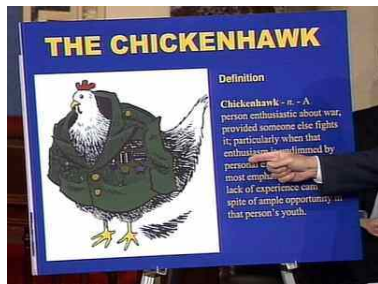
## Taunting ruins deliberation



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

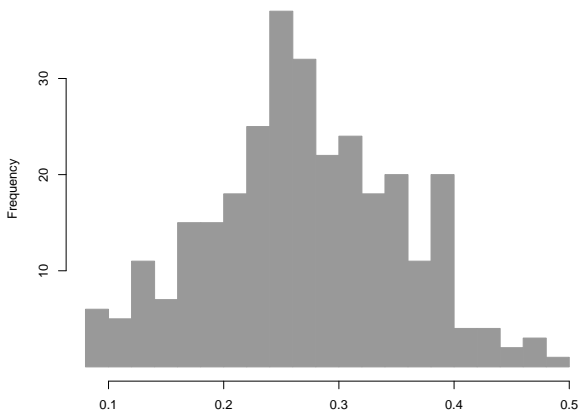
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

# 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

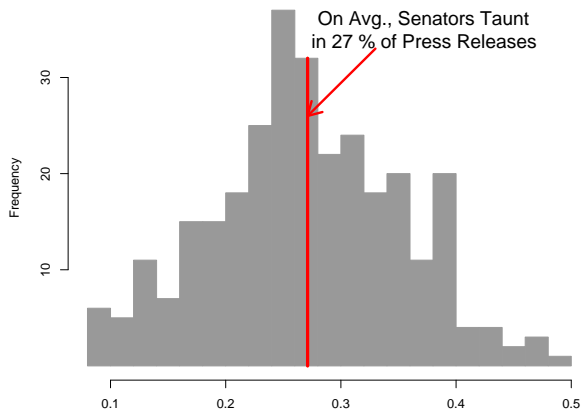
# Out of Sample Confirmation of Partisan Taunting

- 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

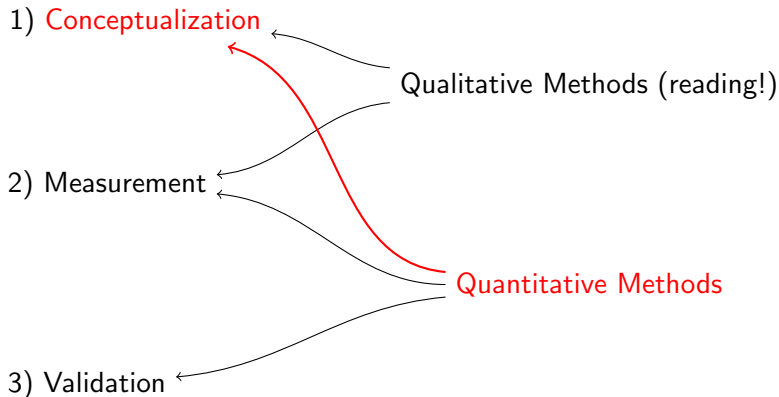


# Out of Sample Confirmation of Partisan Taunting

- 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

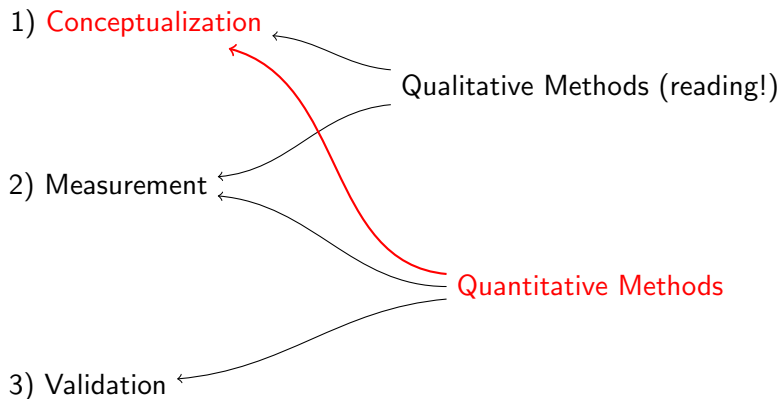


# Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

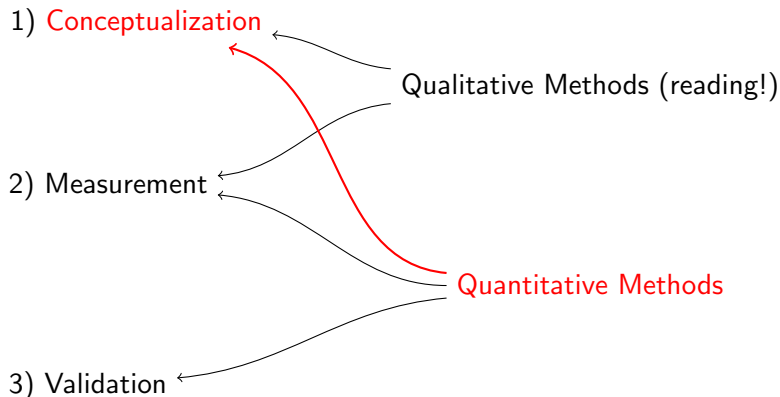
# Quantitative Methods for Qualitative Conceptualization



Quantitative methods for conceptualization and discovery

- Few formal methods designed explicitly for conceptualization

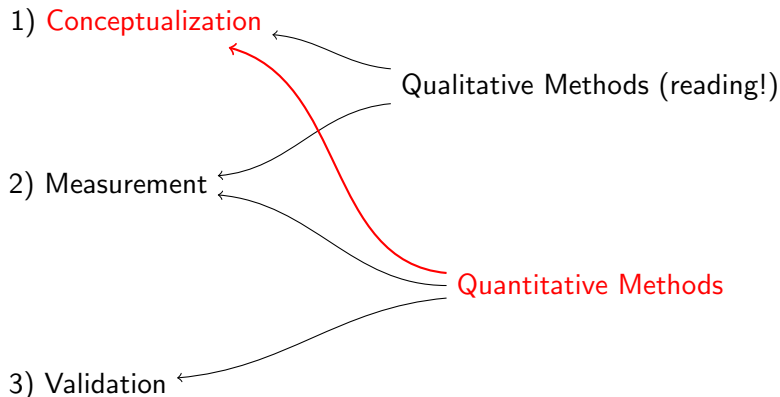
# Quantitative Methods for Qualitative Conceptualization



## 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>