

Computer-Assisted Conceptualization

Gary King

Institute for Quantitative Social Science
Harvard University

Talk at the Ethical Society of Boston, 10/16/2011

¹Based on joint work with Justin Grimmer (Harvard ↔ Stanford)

A Method for Computer Assisted Conceptualization

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

A Method for Computer Assisted Conceptualization

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

A Method for Computer Assisted Conceptualization

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

A Method for Computer Assisted Conceptualization

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

What's Hard about Clustering?

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

- Clustering seems easy; its not!

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

What's Hard about Clustering?

(aka Why Johnny Can't Classify)

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

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)
- $Bell(5) = 52$
- $Bell(100) \approx$ (Number of elementary particles in the universe) $\times 10^{28}$
- 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)
- $Bell(5) = 52$
- $Bell(100) \approx$ (Number of elementary particles in the universe) $\times 10^{28}$
- 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

| | | | | | | |
|--------------|-------------------------------------------------------------------------|--------------|-----------------------------------------------------------------|--------------|-----------------------------------------------------------|--------------|
| 17 566-1282 | Cartage New England Inc 28 Allen Ln Ipswich 01938 | 978 356-9960 | Carter F 34 Hibiscus Bldg 02133 | 617 327-1105 | Carter Nella E 323 Mainville Ave Box 02115 | 617 267-6483 |
| 17 447-4101 | Cartagena Lydia 28 Sweet Box 02131 | 617 323-7639 | Faye & Ricky 207 Columbia Ave Box 02136 | 617 437-7331 | Nicholas S F 115 Randolph Ave Box 02186 | 617 698-5307 |
| 100 257-9961 | Cartagena Avish F Pleasant Rd 02139 | 617 442-9780 | Franklin & Anne 705 Mt Auburn Cam 02138 | 617 354-0798 | Nick 21 Farwell Box 02114 | 617 267-5222 |
| 17 566-1282 | B Had 02134 | 617 361-5253 | Fred 41 Hawthorn Hill 02136 | 617 524-3078 | Nicole | 617 698-0713 |
| 17 364-5188 | Justica 50 Decatur Cha 02129 | 617 241-0152 | Fred 76 Newbury Ave 02138 | 617 698-1343 | Norman G 38 Chickawhatch Dr 02125 | 617 822-1201 |
| 361-0380 | Luzmila 124 Harvard Cam 02138 | 617 491-5621 | G & B 8 Verden Dcr 02134 | 617 434-8906 | P 40 Cranston Pl Box 02135 | 617 437-4754 |
| 17 566-4548 | M 90 Howe Box 02132 | 617 323-9713 | G T 27 Franklin St 02145 | 617 623-7121 | P E 501 E South S Box 02137 | 617 268-8213 |
| 17 628-8248 | Melvin 503 Green Cam 02139 | 617 576-1061 | Gayle 25 Franklin St 02134 | 617 823-0322 | P L 44 Hutchings Box 02131 | 617 427-9170 |
| 17 445-5116 | Carte Nicholas 18 Appleton Boston 02114 | 617 695-6996 | George 125 Madison Box 02134 | 617 367-9548 | P R 91 Boyer Ave 02138 | 617 968-8692 |
| 17 822-2962 | Cartagena O 4 Bradford Box 02138 | 617 338-0219 | Carter Hillside Assoc/AM 107 S Street Box 02111 | 617 456-1689 | Paul & Constance 114 Freeman St W Box 02131 | 617 325-2036 |
| 17 427-5712 | Carten Thos J Sr & Claire 1 Franklin Rd Mt 02136 | 617 698-6163 | Carter Harry F 26 Irving Rd Rt W Box 02132 | 617 325-5465 | Paul M 27 Union Rd 02135 | 617 787-2115 |
| 17 569-2698 | Thomas & Kathleen 50 Thompson Ln Mt 02136 | 617 696-6919 | Carter Hide Co Inc 160 Boston Rd 02132 | 617 542-7987 | Prudence 40 Franklin Waterlton 02127 | 617 393-3782 |
| 17 667-5190 | Carte A Box 02133 | 617 442-5230 | Carter Hilary 41 Harvey Cam 02148 | 617 876-2750 | Renee & Andrew 106 Emerald Dorchester 02122 | 617 541-2843 |
| 17 569-1417 | A 203 Newbury Ave Cambridge 02142 | 617 492-4174 | Horace 301 Walnut St Roxbury 02119 | 617 442-5307 | Renee & Andrew 106 Emerald Dorchester 02122 | 617 541-2843 |
| 17 338-9110 | A M 255 Massachusetts Ave 02115 | 617 266-7153 | Howard Jr 28 Netha Drive Box 02118 | 617 445-5532 | Rice David 10 Walnut Box 02118 | 617 720-3765 |
| 17 825-1993 | Adams 361 Centre St Mt 02136 | 617 698-7074 | J Dan | 617 354-2658 | Rice David 10 Walnut Box 02118 | 617 720-3765 |
| 17 296-1593 | Alice 108 Elmwood St 02134 | 617 423-0193 | J 31 Chatham Box 02144 | 617 232-7990 | Robbie Dennis Publishing 163 Main Wilmington 01887 | 800 638-1671 |
| 17 670-2078 | Alice O Market Cambridge 02139 | 617 945-2711 | J 538 Harvard Box 02146 | 617 730-9483 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 621-9001 | Andrew F 42 West St Box 02138 | 617 625-7623 | J 775 The Pines West Roxbury 02132 | 617 323-5374 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 296-4725 | Carte Anne MD 1161 Beacon St 02144 | 617 739-1022 | Jacques J Jacques MD 1 Ipswich Pl Box 02146 | 617 735-8787 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 542-1521 | Carte J M 371 Newbury Boston 02114 | 617 536-6329 | Carte J M 3410 Columbia Rd S Box 02137 | 617 464-1040 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 364-5232 | B E 18 Graduate Ave Mt 02136 | 617 296-6911 | Carte J M Ornamental Ironworks 40 Newbury Falls 02134 | 617 876-5353 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 541-5249 | Carte Barbara L MD Tufts-New England Medical Center Box 02111 | 617 636-0051 | Carte J Veal Co 40 Hawthorn Hill 02136 | 617 442-1775 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 739-2662 | Carte Becky Jo 02134 | 617 523-4368 | Carte James 157 Cambridge St Cam 02138 | 617 492-1214 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 879-0030 | Bernard J 122 Goodhue E Box 02136 | 617 567-9430 | James 412 Foster Ave Roxbury 02119 | 617 739-2193 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 364-5232 | Bibbiah 25 Midway Dcr 02134 | 617 298-8713 | James 311 Good Star Rd Cambridge 02141 | 617 876-8841 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 541-5249 | Bliss 26 Elmwood St 02134 | 617 367-9931 | Jane 14 Adams Rd Newton 02465 | 617 564-0435 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 739-2662 | Carte Broadcasting Co 50 Park Pl Box 02134 | 617 423-0210 | Jas L 34 Newbury Rd Mt 02136 | 617 361-0773 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 879-0030 | Carte & Business Consultants Inc 73 East St Cam 02141 | 617 225-0200 | Jane 14 Adams Rd Newton 02465 | 617 564-0435 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 541-3948 | Carte C 200 Commonwealth Ave 02135 | 617 782-2118 | John 107 Summer Box 02128 | 617 423-4334 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 436-1511 | C 218 Harvard Ave East Boston 02128 | 617 569-1545 | John 40 Hawthorn Hill 02136 | 617 282-1235 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 17 569-4119 | C 109 Harvard Cam 02138 | 617 491-4822 | John 107 Summer Box 02128 | 617 423-4334 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 100 257-9961 | C 218 Harvard Ave East Boston 02128 | 617 569-1545 | John 107 Summer Box 02128 | 617 423-4334 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 100 257-9961 | C 109 Harvard Cam 02138 | 617 491-4822 | John 107 Summer Box 02128 | 617 423-4334 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 100 257-9961 | C 218 Harvard Ave East Boston 02128 | 617 569-1545 | John 107 Summer Box 02128 | 617 423-4334 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 100 257-9961 | C & M 43 Bernhagen Jan 02138 | 617 524-9558 | John 107 Summer Box 02128 | 617 423-4334 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |
| 100 257-9961 | C & M 43 Bernhagen Jan 02138 | 617 524-9558 | John 107 Summer Box 02128 | 617 423-4334 | Carl Sec Industrial Prod 613 Main Wilmington | 800 616-7447 |

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 St Ipswich 01938 | 978 356-9960 |
| 17 447-4101 | Cartagema Lydia 28 Sweet Briar 02131 | 617 323-7639 |
| 100 257-9961 | Cartagema Avish F Beach Rd 02139 | 617 442-9780 |
| 17 566-1282 | B Had 02136 | 617 361-5253 |
| 17 364-5188 | Lucilla 174 Harvard Cam 02136 | 617 241-0152 |
| 361-0380 | M 95 Howe St 02136 | 617 491-5621 |
| 17 566-4548 | Melvin 503 Green Cam 02139 | 617 576-1061 |
| 17 628-8248 | Carte Nicholas 18 Appleton Boston 02114 | 617 695-6996 |
| 17 445-5116 | Carlton 40 S 4thford Bay 02118 | 617 338-9219 |
| 17 822-2962 | Carlton 50 Thompson Ln Mt 02136 | 617 696-6919 |
| 17 427-5712 | A Heber 100 Main St 02111 | 617 225-2257 |
| 17 569-2698 | A 22 Beulah Wy Hoxbury 02119 | 617 442-1219 |
| 17 667-5190 | A 200 Pitman Av Cambridge 02142 | 617 492-4174 |
| 17 569-1417 | Adams 301 Centre St Mt 02136 | 617 698-7074 |
| 17 338-1107 | Alice 108 Elmwood Pl 02116 | 617 453-0193 |
| 17 825-9195 | Allice 40 Market Cambridge 02139 | 617 945-2711 |
| 17 296-1293 | Andrew F 42 West St 02135 | 617 625-7623 |
| 17 670-2078 | Arter Anne MD 1101 Beacon St 02144 | 617 739-1022 |
| 17 621-9001 | B E 108 Gladstone Av Mt 02116 | 617 296-6911 |
| 17 296-4725 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 542-1521 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 364-5232 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 541-5649 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 739-2662 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 879-0030 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 541-3948 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 436-1511 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 569-6119 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 100 802-0212 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 100 869-8782 | B 271 Newburgh Boston 02116 | 617 536-6329 |
| 17 327-1105 | Carter F 51 Hibiscus Bay 02131 | 617 327-1105 |
| 17 437-7331 | Faye & Ricky 20 Columbia Av Mt 02116 | 617 437-7331 |
| 17 323-6781 | Francis S 134 Temple W Av 02132 | 617 323-6781 |
| 17 354-0798 | Franklin & Anne 705 Mt Auburn Cam 02138 | 617 354-0798 |
| 17 524-3078 | Fred 40 Haverhill Av 02136 | 617 524-3078 |
| 17 698-1343 | Fred 16 Hoxbury Av Mt 02136 | 617 698-1343 |
| 17 436-8906 | G & B 8 Vardon Bay 02134 | 617 436-8906 |
| 17 623-7121 | G T 27 Fryden Av Sun 02145 | 617 623-7121 |
| 17 825-8322 | Gayle 25 Franklin St 02114 | 617 825-8322 |
| 17 522-3215 | Geo S 115 Meigs Hill Mt 02138 | 617 522-3215 |
| 17 367-9548 | George 120 Nones Bay 02114 | 617 367-9548 |
| 17 456-1689 | Carter Hillside Assoc 107 S Street Bay 02111 | 617 456-1689 |
| 17 325-5465 | Carter Harry F 100 Burns Rd W Av 02112 | 617 325-5465 |
| 17 542-7987 | Carter Hide Co Inc 140 Newbury St 02116 | 617 542-7987 |
| 17 876-2750 | Carter Hilary 41 Harvey Cam 02148 | 617 876-2750 |
| 17 442-5307 | Horace 301 Walnut Av Hoxbury 02119 | 617 442-5307 |
| 17 445-5532 | Howard Jr 28 New One Bay 02118 | 617 445-5532 |
| 17 354-2658 | J 400 Main St 02111 | 617 354-2658 |
| 17 232-7990 | J 15 Chatham St 02144 | 617 232-7990 |
| 17 730-9483 | J 538 Harvard St 02138 | 617 730-9483 |
| 17 323-5274 | J 775 The Pines West Hoxbury 02116 | 617 323-5274 |
| 17 735-8787 | J 1 Crockett Pl Mt 02144 | 617 735-8787 |
| 17 664-1040 | 3410 Columbia Rd S Cam 02136 | 617 664-1040 |
| 17 436-5353 | Carter J M Ornamental Ironworks Pondside Falls 01746 | 617 436-5353 |
| 17 442-1775 | Carter J Neal Co 40 Newbury St 02118 | 617 442-1775 |
| 17 492-1214 | 1573 Cambridge St Cam 02136 | 617 492-1214 |
| 17 739-2193 | James 102 Foster Av Hoxbury 02116 | 617 739-2193 |
| 17 876-8841 | James 31 East Star Rd Cambridge 02141 | 617 876-8841 |
| 17 361-0773 | Jane L 34 Rosbury Rd Mt 02136 | 617 361-0773 |
| 17 964-0435 | Janine 14 Adams Rd Newton 02458 | 617 964-0435 |
| 17 426-5994 | Jeffrey 41 Warren Av Sun 02145 | 617 426-5994 |
| 17 987-2163 | John 11 Mansfield St 02134 | 617 987-2163 |
| 17 423-4334 | John 207 Summer St 02116 | 617 423-4334 |
| 17 282-1235 | Jeffrey 41 Warren Av Sun 02145 | 617 282-1235 |
| 17 734-6109 | June O 129 A Summit Av Mt 02116 | 617 734-6109 |
| 17 265-8656 | K 179 Inverness St 02116 | 617 265-8656 |
| 17 282-1593 | K 17 Inverness St 02116 | 617 282-1593 |
| 17 267-6483 | Carter Nella E 323 Main St Av Mt 02115 | 617 267-6483 |
| 17 698-5307 | Nicholas S F 115 Randolph Av Mt 02136 | 617 698-5307 |
| 17 267-5222 | Nick 21 Farnham Bay 02114 | 617 267-5222 |
| 17 527-0480 | Nick & Debbi 136 Hermit Rd Newton 02459 | 617 527-0480 |
| 17 698-0713 | Norman G 38 Chickadee Dr 02116 | 617 698-0713 |
| 17 822-1201 | 38 Chickadee Dr 02116 | 617 822-1201 |
| 17 427-4754 | P 40 Cranston Pl Mt 02116 | 617 427-4754 |
| 17 268-4213 | P E 501 E South St Av 02137 | 617 268-4213 |
| 17 427-7170 | P L 44 Huxtings Bay 02116 | 617 427-7170 |
| 17 968-8692 | P R 91 Boyer Cam 02138 | 617 968-8692 |
| 17 325-2034 | Paul & Constance 114 Adams Av W Mt 02110 | 617 325-2034 |
| 17 268-4546 | Paul E 501 E South St Av 02137 | 617 268-4546 |
| 17 787-2115 | Paul M 27 Union St 02116 | 617 787-2115 |
| 17 235-8488 | Carter Pike Driving Inc 27 Beaver Ct Framingham 02719 | 617 235-8488 |
| 17 393-3782 | Carter Prudence 40 Franklin Waterbury 02172 | 617 393-3782 |
| 17 926-7063 | Prudence 40 Franklin Waterbury 02172 | 617 926-7063 |
| 17 541-2843 | Reginald 100 Broadview Center 02124 | 617 541-2843 |
| 17 720-3765 | Renee & Andrew 100 Broadview Center 02124 | 617 720-3765 |
| 17 800-638-1671 | Carter Rice Doan 163 Main Wilmington 01887 | 800 638-1671 |
| 17 744-7447 | Carl Eric Doan 163 Main Wilmington 01887 | 800 638-1671 |
| 17 648-7447 | Carl Eric Doan 163 Main Wilmington 01887 | 800 648-7447 |
| 17 978-988-7447 | Carl Eric Doan 163 Main Wilmington 01887 | 978 988-7447 |
| 17 638-1673 | Carrie 163 Main Wilmington 01887 | 800 638-1673 |
| 17 987-0836 | Carter Richard 2075 Carleton Av Brighton 02115 | 617 987-0836 |
| 17 566-7293 | Carter Richard A M 47 W Vernon St 02136 | 617 566-7293 |
| 17 967-0710 | Carter Richard A M 120 Centre St 02116 | 617 967-0710 |
| 17 268-0468 | Carter Richard R 120 Centre St 02116 | 617 268-0468 |
| 17 864-1535 | Robert L 175 Rockwood Av Cam 02141 | 617 864-1535 |
| 17 424-6148 | Roger 130 St Brains Bay 02116 | 617 424-6148 |
| 17 491-6115 | Royce 18 Salisbury Cir 02129 | 617 491-6115 |
| 17 241-9418 | Royce 18 Salisbury Cir 02129 | 617 241-9418 |



Our Idea: Meaning Through Geography

Set of clusterings \approx

A list of unconnected addresses

wide at SuperPages.com

| 195 | Car | C |
|--------------|--------------------------------------------------------------------|--------------|
| 17 566-1282 | Cartage New England Inc 28 Allen Ln Ipswich 01938 | 978 356-9960 |
| 18 447-4101 | 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 02134 | 617 361-5253 |
| 17 364-5188 | Lucille 124 Harvard Can 02139 | 617 491-5621 |
| 361-0380 | M 95 Howe Box 02135 | 617 323-9713 |
| 17 566-4548 | Melvin 503 Green Can 02139 | 617 576-1061 |
| 17 628-8248 | Carte Nicholas 18 Appleton Boston 02114 | 617 695-6996 |
| 17 445-5116 | Carlton 4 4 Bradford Box 02138 | 617 338-0219 |
| 17 622-2962 | Carton 50 Thompson Ln Mt 02136 | 617 696-6919 |
| 17 427-5712 | A Weber A 200 Riverside Av Cambridge 02142 | 617 492-4174 |
| 17 569-2698 | A 21 Beetham Wy Haverhill 02119 | 617 442-1219 |
| 17 667-5190 | A M 250 Massachusetts Av 02115 | 617 266-7153 |
| 17 569-1417 | Adams 361 Carter St Mt 02136 | 617 698-9074 |
| 17 338-1101 | Adams P 42 West St 02134 | 617 945-2711 |
| 17 825-1993 | Carte Anne MD 1161 Beacon Bldg 02144 | 617 739-1022 |
| 17 296-1193 | Cartier Adhena 971 Newbury Boston 02116 | 617 536-6239 |
| 17 670-2078 | B E 18 Gladstone Av Mt 02136 | 617 296-6911 |
| 17 621-9001 | Cartier Barbara L MD Tufts New England Medical Center Box 02111 | 617 436-0951 |
| 17 296-4725 | Cartier Becky 90 02114 | 617 523-4368 |
| 17 542-1521 | Bernard J 301 Ashdown E Mt 02136 | 617 567-9430 |
| 17 364-5232 | Bibbith 25 Midway Dr 02134 | 617 298-8713 |
| 17 541-5649 | Billings 38 New Avenue 02138 | 617 367-9931 |
| 17 739-2662 | Cartier Broadcasting Co 58 Park Pl Box 02116 | 617 423-0210 |
| 17 879-0030 | Cartier C 2000 Cavendish 73 East C Can 02141 | 617 225-0200 |
| 17 541-3948 | Cartier C 200 Cavendish 73 East C Can 02141 | 617 782-2118 |
| 17 436-1511 | C 210 Townsend Av East Boston 02128 | 617 569-1545 |
| 17 569-4119 | C 109 Harvard Can 02138 | 617 491-4822 |
| 90 602-0211 | C 8 111 Cambridge 02139 | 617 492-4392 |
| 90 569-8782 | C & M 41 Northgate Jct 02134 | 617 524-9558 |
| 17 327-1105 | Carter F 54 Hillside Box 02131 | 617 327-1105 |
| 17 437-7331 | Faye & Ricky 20 Columbia Av Box 02136 | 617 437-7331 |
| 17 323-6781 | Francis S 134 Temple W Av 02132 | 617 323-6781 |
| 17 354-0798 | Franklin & Anne 705 Mt Auburn Can 02138 | 617 354-0798 |
| 17 524-3078 | Fred 41 Howard Av 02136 | 617 524-3078 |
| 617 698-1343 | Fred 96 Newbury Av Mt 02136 | 617 698-1343 |
| 617 436-8906 | G & B 8 Vardon Box 02134 | 617 436-8906 |
| 617 623-7121 | G T 27 Fossil Av Mt 02136 | 617 623-7121 |
| 617 823-8322 | Gayle 25 Franklin St 02134 | 617 823-8322 |
| 617 522-3215 | Geo S 115 Mount Mt Jct Box 02138 | 617 522-3215 |
| 617 367-9548 | George 25 Madison Box 02114 | 617 367-9548 |
| 617 456-1689 | Carter Hillside Assoc 107 S Street Box 02111 | 617 456-1689 |
| 617 325-5465 | Carter Harry F 30 Bayview Rd W Av 02132 | 617 325-5465 |
| 617 542-7987 | Carter Hide Co Inc 140 Newbury St 02116 | 617 542-7987 |
| 617 876-2750 | Carter Hilary 41 Harvey Can 02148 | 617 876-2750 |
| 617 442-5307 | Horace 361 Walnut Av Haverhill 02119 | 617 442-5307 |
| 617 445-5552 | Howard Jr 28 New One Box 02118 | 617 445-5552 |
| 617 354-2658 | J Can 15 Chatham St 02144 | 617 232-7990 |
| 617 730-9983 | J 538 Harvard St 02144 | 617 730-9983 |
| 617 323-5274 | J 775 The Pines West Haverhill 02132 | 617 323-5274 |
| 617 735-8787 | Carter J Jacques MD 1 Brookline Pl Bldg 02144 | 617 735-8787 |
| 617 464-1040 | 3410 Columbia Rd S Box 02137 | 617 464-1040 |
| 617 436-5353 | Carter J M Ornamental Ironworks 200 Franklin Falls 02137 | 617 436-5353 |
| 617 442-1775 | Carter J Veal Co 40 Newbury St 02138 | 617 442-1775 |
| 617 492-1214 | 1573 Cambridge St Can 02136 | 617 492-1214 |
| 617 739-2193 | James 62 Foster Av Haverhill 02118 | 617 739-2193 |
| 617 876-8841 | J 31 East Star Rd Cambridge 02141 | 617 876-8841 |
| 617 361-0773 | J 34 Newbury Rd Mt 02136 | 617 361-0773 |
| 617 964-0435 | Jane 14 Adams Rd Newton 02458 | 617 964-0435 |
| 617 426-9094 | John 1200 Cambridge St 02138 | 617 426-9094 |
| 617 987-2163 | John 11 Mansfield St 02134 | 617 987-2163 |
| 617 423-4134 | John 207 Summer St 02135 | 617 423-4134 |
| 617 282-1235 | John 40 Howard St 02139 | 617 282-1235 |
| 617 734-6109 | James O 129 A Summit Av 02131 | 617 734-6109 |
| 617 265-8656 | J 29 Inverness Dr 02134 | 617 265-8656 |
| 617 282-1593 | K 17 Concord Road 02123 | 617 282-1593 |
| 617 267-6483 | 323 Marchant Av Box 02115 | 617 267-6483 |
| 617 698-5307 | Nicholas S F 115 Randolph Av Mt 02136 | 617 698-5307 |
| 617 267-5222 | Nick 21 Fossil Box 02116 | 617 267-5222 |
| 617 527-0480 | 136 Hermit Rd Newton 02459 | 617 527-0480 |
| 617 698-0713 | Norman G 38 Chickadee Dr 02126 | 617 698-0713 |
| 617 822-1201 | 38 Chickadee Dr Box 02126 | 617 822-1201 |
| 617 427-4754 | P 41 Woodland Pl Box 02135 | 617 427-4754 |
| 617 268-8213 | P E 501 E South S Box 02137 | 617 268-8213 |
| 617 427-9170 | P L 44 Huddings Box 02131 | 617 427-9170 |
| 617 968-8692 | P R 91 Bayview Box 02138 | 617 968-8692 |
| 617 325-3034 | 114 Adams Av W Mt 02131 | 617 325-3034 |
| 617 268-4546 | Paul F 501 E South S Box 02137 | 617 268-4546 |
| 617 787-2115 | Paul M 27 Crown St 02139 | 617 787-2115 |
| 617 235-0488 | Carter Pike Driving Inc 27 Beaver Ct Franklin 02130 | 617 235-0488 |
| 617 393-3782 | Carter Prudence 40 Franklin Waterbury 02172 | 617 393-3782 |
| 617 926-7063 | Prudence 40 Franklin Waterbury 02172 | 617 926-7063 |
| 617 541-2843 | Reginald 100 Brookside Circle 02124 | 617 541-2843 |
| 617 720-3765 | Carter Rice 100 Brookside Circle 02124 | 617 720-3765 |
| 800 638-1671 | Red Free-Deal 'I' & 'Th' 100 Brookside Circle 02124 | 800 638-1671 |
| 800 619-7447 | Red Free-Deal 'I' & 'Th' 100 Brookside Circle 02124 | 800 619-7447 |
| 800 648-7447 | Red Free-Deal 'I' & 'Th' 100 Brookside Circle 02124 | 800 648-7447 |
| 978 988-7447 | Reginald 100 Brookside Circle 02124 | 978 988-7447 |
| 800 638-1673 | Richard 2075 Cambridge St 02138 | 800 638-1673 |
| 617 987-0836 | Richard A 974 Vermont St 02136 | 617 987-0836 |
| 617 267-0710 | Richard A 100 Brookside Circle 02124 | 617 267-0710 |
| 617 268-0468 | Richard R 123 Mount S Box 02137 | 617 268-0468 |
| 617 864-1535 | Robert L 175 Newbury Av Can 02141 | 617 864-1535 |
| 617 424-6148 | Royce 130 St Brnagh Box 02131 | 617 424-6148 |
| 617 491-6115 | Royce 130 St Brnagh Box 02131 | 617 491-6115 |
| 617 241-9418 | Royce 18 Sanderson Cir 02129 | 617 241-9418 |



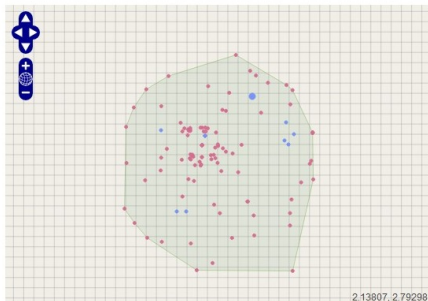
\approx We develop a (conceptual) geography of clusterings

Software Screenshot

Size: 244 Files

Description: NSF - Updated Set

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



Display History Display Method Points

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

(*) Discoverable

Coordinates: 0.84753, 1.42538

Clusters: 5

Label [+] methods-oriented clustering

29.51%
72 research community health science public practice global political national urban

[View Detail](#)

27.46%
67 data economic markets policy survey models financial use not risk

[View Detail](#)

21.72%
53 human social science systems behavioral networks brain spatial complex dynamics

[View Detail](#)

15.16%
37 education students school learning creative skills teaching cognitive college teachers

[View Detail](#)

6.15%
15 language linguistic speech data speakers computer semantic cultural variation documentation

[View Detail](#)

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
 - Cluster Quality = $\text{mean}(\text{within cluster}) - \text{mean}(\text{between clusters})$

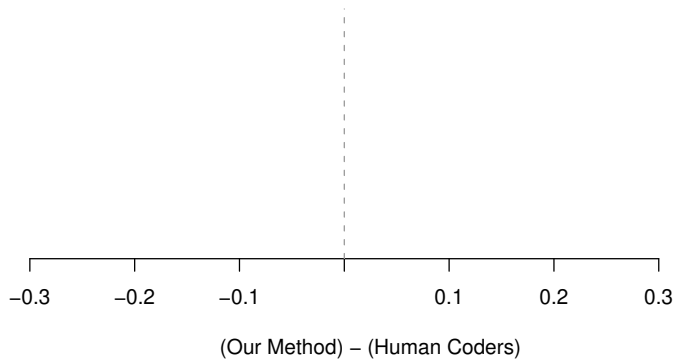
Evaluation 1: Cluster Quality

- **What Are Humans Good For?**
 - They can't: keep many documents & clusters in their head
 - They can: compare two documents at a time
 - \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
 - Cluster Quality = $\text{mean}(\text{within cluster}) - \text{mean}(\text{between clusters})$
 - **Bias results against ourselves by not letting evaluators choose clustering**

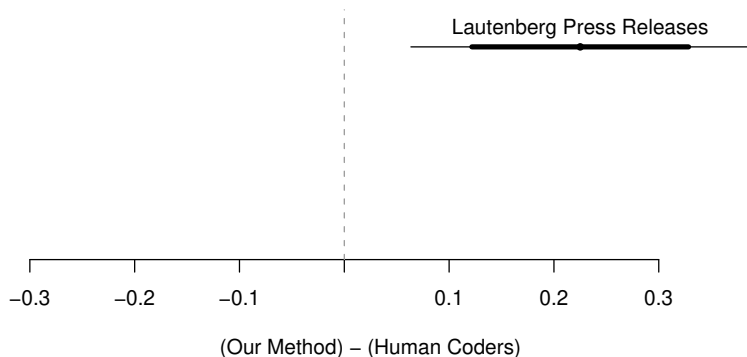
Evaluation 1: Cluster Quality

- **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
 - Cluster Quality = $\text{mean}(\text{within cluster}) - \text{mean}(\text{between clusters})$
 - **Bias results against ourselves by not letting evaluators choose clustering**

Evaluation 1: Cluster Quality

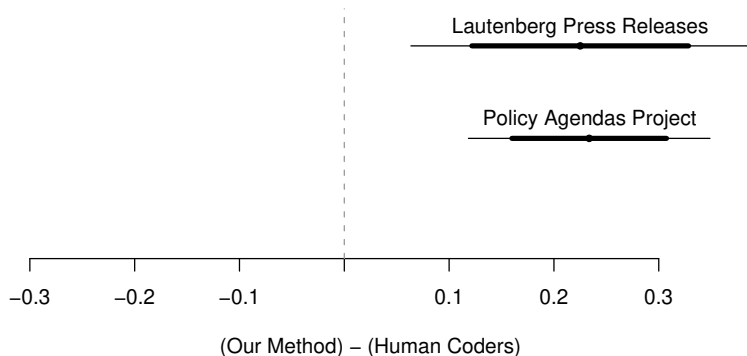


Evaluation 1: Cluster Quality



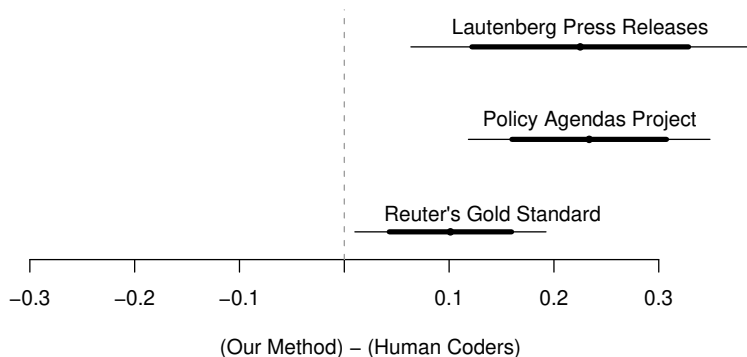
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, ...)

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

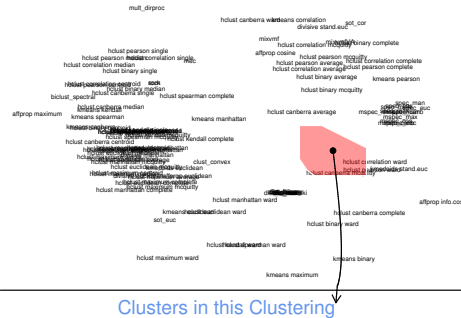
Evaluation 3: What Do Members of Congress Do?

- 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



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”



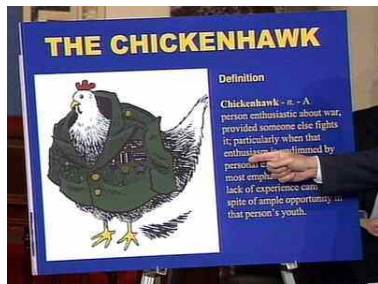
Credit Claiming
Pork



Mayhew Credit Claiming
Legislation

Gary King (Harvard IQSS)

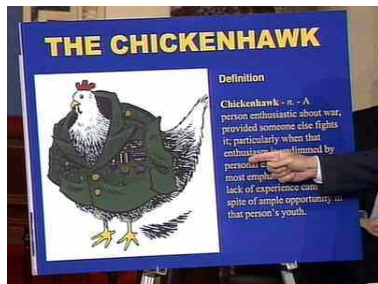
Taunting ruins deliberation



Sen. Lautenberg
on Senate Floor
4/29/04

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

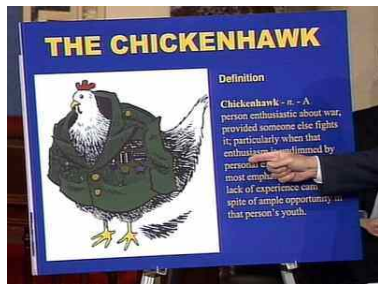
Taunting ruins deliberation



Sen. Lautenberg
on Senate Floor
4/29/04

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

Taunting ruins deliberation



Sen. Lautenberg
on Senate Floor
4/29/04

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

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

Out of Sample Confirmation of Partisan Taunting

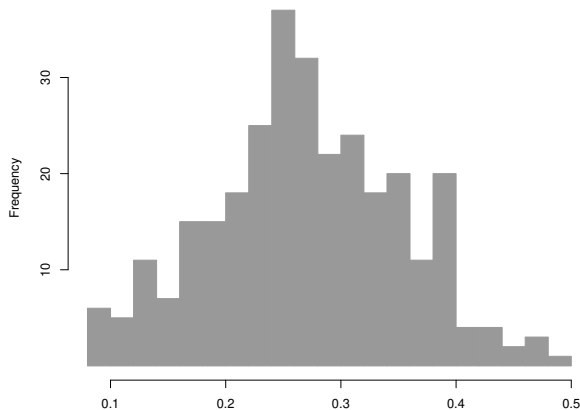
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

Out of Sample Confirmation of Partisan Taunting

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

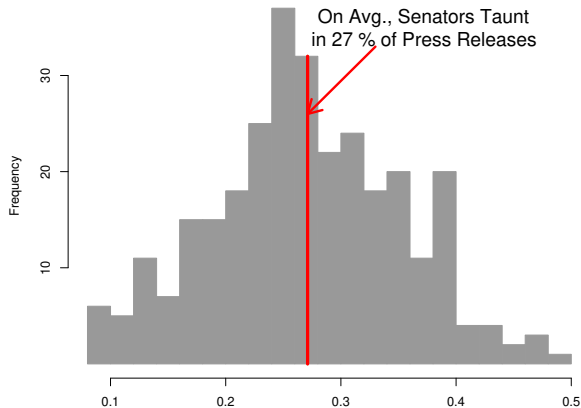
Out of Sample Confirmation of Partisan Taunting

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

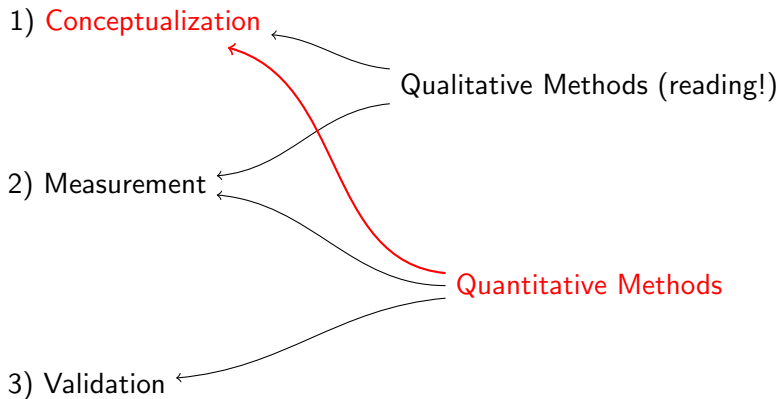


Out of Sample Confirmation of Partisan Taunting

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

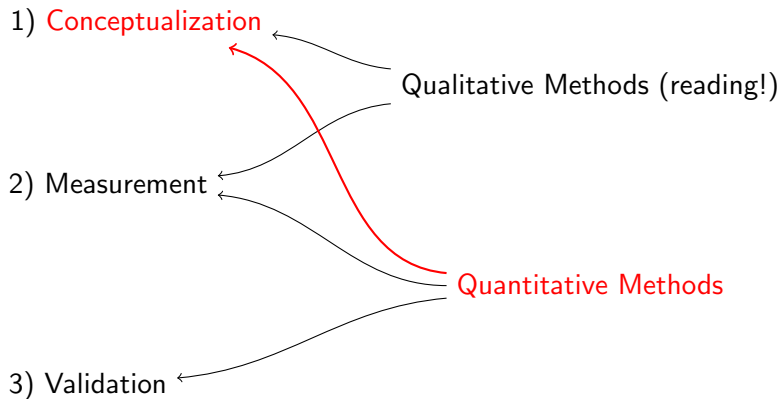


Quantitative Methods for Qualitative Conceptualization



Computer-Assisted Methods for conceptualization and discovery

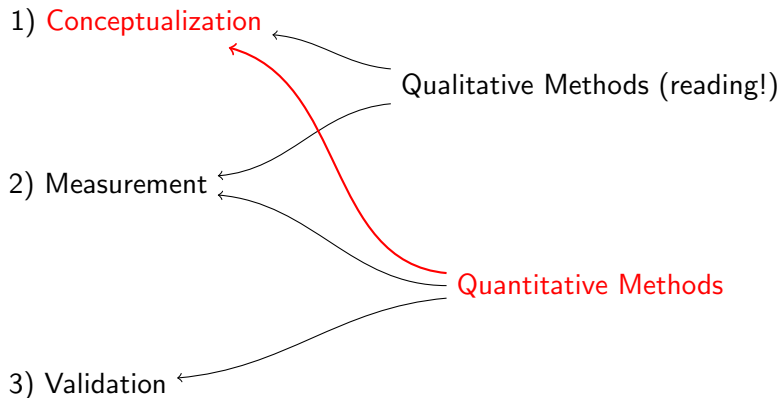
Quantitative Methods for Qualitative Conceptualization



Computer-Assisted Methods for conceptualization and discovery

- Methods designed explicitly for conceptualization

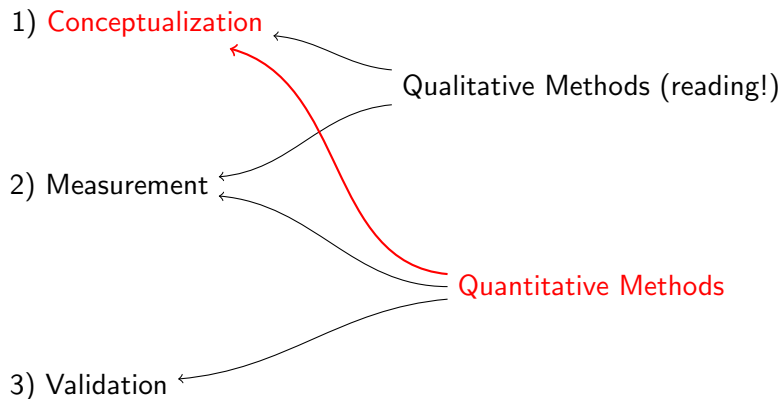
Quantitative Methods for Qualitative Conceptualization



Computer-Assisted Methods for conceptualization and discovery

- Methods designed explicitly for conceptualization
- The end of quantitative v qualitative debates

Quantitative Methods for Qualitative Conceptualization



Computer-Assisted Methods for conceptualization and discovery

- Methods designed explicitly for conceptualization
- The end of quantitative v qualitative debates
- Evaluation methods measure progress in discovery

For more information



<http://GKing.Harvard.edu>