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

Gary King

Institute for Quantitative Social Science
Harvard University

Talk at BAE Systems, 9/9/2010

Joint work with Justin Grimmer (Harvard $\rightsquigarrow$ Stanford)

A Method for Conceptualization

- Systematic method for computer-assisted conceptualization from text

A Method for Conceptualization

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

A Method for Conceptualization

- Systematic method for computer-assisted conceptualization from text
- Conceptualization through Classification: "one of the most central and generic of all our conceptual exercises. ...the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis.... Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or,for that matter, social science research." (Bailey, 1994).
- We focus on Cluster Analysis: simultaneously 1) invent categories and 2) assign documents to categories

A Method for Conceptualization

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

Why Johnny Can't Classify (Optimally)

Why Johnny Can't Classify (Optimally)

- Clustering seems easy; its not!

Why Johnny Can't Classify (Optimally)

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

Why Johnny Can't Classify (Optimally)

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

Why Johnny Can't Classify (Optimally)

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

Why Johnny Can't Classify (Optimally)

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

Why Johnny Can't Classify (Optimally)

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

Why Johnny Can't Classify (Optimally)

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

Why Johnny Can't Classify (Optimally)

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

Why Johnny Can't Classify (Optimally)

- Clustering seems easy; its not!
- $\text{Bell}(n) =$ number of ways of partitioning n objects
- $\text{Bell}(2) = 2$ (AB, A B)
- $\text{Bell}(3) = 5$ (ABC, AB C, A BC, AC B, A B C)
- $\text{Bell}(5) = 52$
- $\text{Bell}(100) \approx 10^{28} \times$ Number of elementary particles in the universe
- Now imagine choosing the *optimal* classification scheme by hand!
- Its no surprise that automated algorithms can help, but which algorithms?

Why HAL Can't Classify Either

Why HAL Can't Classify Either

- Large quantitative literature on cluster analysis

Why HAL Can't Classify Either

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

Why HAL Can't Classify Either

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

Why HAL Can't Classify Either

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

Why HAL Can't Classify Either

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,. . .

Why HAL Can't Classify Either

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,. . .
  - Well-defined statistical, data analytic, or machine learning foundations

Why HAL Can't Classify Either

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

Why HAL Can't Classify Either

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

Why HAL Can't Classify Either

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

Why HAL Can't Classify Either

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible
- Deep problem in cluster analysis literature: no way to know which method will work ex ante

Why HAL Can't Classify Either

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible
- Deep problem in cluster analysis literature: no way to know which method will work ex ante
- No surprise: everyone's tried cluster analysis; very few are satisfied

If Ex Ante doesn't work, try Ex Post

If Ex Ante doesn't work, try Ex Post

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

If Ex Ante doesn't work, try Ex Post

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

If Ex Ante doesn't work, try Ex Post

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

If Ex Ante doesn't work, try Ex Post

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

If Ex Ante doesn't work, try Ex Post

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

If Ex Ante doesn't work, try Ex Post

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

If Ex Ante doesn't work, try Ex Post

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

If Ex Ante doesn't work, try Ex Post

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

Our Idea: Meaning Through Geography

Set of clusterings

Our Idea: Meaning Through Geography

Set of clusterings ≈
A list of unconnected addresses

Our Idea: Meaning Through Geography

Set of clusterings $\approx$

A list of unconnected addresses

wide at SuperPages.com 195 Car C

17 566-1282
81 447-4101
00 257-9981
17 566-1282
17 364-5188
361-0380
17 566-4548
17 628-8248
17 445-5116
17 822-2982
17 427-5712
17 569-2698
17 667-5190
17 569-1417
hitu Dr
17 338-9110
17 825-9195
17 296-1593
17 670-2078
17 623-9001
17 296-4725
17 542-1521
17 364-5232
17 541-5649
17 739-2662
17 879-0030
17 541-3948
17 436-1513
17 569-4119
ton 02128
300 569-8782

Cartage New England Inc
26 Allen Ln Ipswich 01938.........978 356-9960
Cartagema Lydia
18 Jewett Ros 02131.........617 323-7639
Cartagena Avith
9 Bancroft Rox 02119.........617 442-9780
B Hyd 02136.........617 361-5253
Jessica 50 Decatur Cha 02129.........617 241-0152
Lucilla 174 Harvard Cam 02139.........617 491-5621
M 95 Rowe Ros 02131.........617 323-9713
Melvin 501 Green Cam 02139.........617 576-1061
Carte Nicholas
18 Appleton Boston 02116.........617 695-6996
Cartegena O 4 Milford Bos 02118.........617 338-8219
Carten Thos J Sr & Claire
1 Paradise Rd Mil 02186.........617 698-6163
Thomas & Kathleen
50 Thompson Ln Mil 02186.........617 696-6919
Carter A Ros 02131.........617 327-2257
A Roxbury.........617 442-5230
A 31 Bethune Wy Roxbury 02119.........617 442-1219
A 260 Putnam Av Cambridge 02139.........617 492-4174
A M 255 Maschsts Av Bos 02115.........617 266-7153
Adams 361 Centre St Mil 02186.........617 698-9074
Alice 108 Kilmarnock Bos 02215.........617 425-0193
Alice 45 Market Cambridge 02139.........617 945-2711
Andrew F 62 Vinal Av Som 02143.........617 625-7623
Carter Anne MD
1101 Beacon Bro 02446.........617 739-1022
Carter Athens
272 Newbury Boston 02116.........617 536-6329
B E 68 Gladeside Av Mat 02126.........617 296-6911
Carter Barbara L MD
Tufts-New England Medical Center Bos 02111
Call.........617 636-0051
Carter Becky Bos 02114.........617 523-4368
Bernard J
112 Gladstone E Bos 02128.........617 567-3430
Bithiah 25 Medway Dor 02124.........617 298-8713
Blake 26 Mt Vernon Bos 02108.........617 367-9931
Carter Broadcasting Co
20 Park Plz Bos 02116.........617 423-0210
Carter & Burgess Consultants Inc
23 East St Cam 02141.........617 225-0200
Carter C 2000 Comnwlth Av Bri 02135.........617 782-2118
C 228 Faywood Av East Boston 02128.........617 569-1545
C 359 Harvard Cam 02138.........617 491-4822
C 610 Walk Hill Mat 02126.........617 296-6392
C & M 43 Burroughs Jam 02130.........617 524-9558

Carter F 24 Hillock Ros 02131.........617 327-1105
Faye & Ricky
357 Columbus Av Bos 02116.........617 437-7331
Francis S 134 Temple W Rox 02132.........617 323-6781
Franklin & Anne
221 Mt Auburn Cam 02138.........617 354-0798
Fred 42 Haverford Jam 02130.........617 524-3078
Fred 96 Hinckley Rd Mil 02186.........617 698-1343
G & R 8 Verdun Dor 02124.........617 436-8906
G T 27 Franklin Av Som 02145.........617 623-7121
Gayle 25 Frontenac Dor 02124.........617 825-0322
Geo S 115 Moss Hill Rd Jam 02130.........617 522-3215
George 125 Nashua Bos 02114.........617 367-9548
Carter Halliday Associate
107 S Street Bos 02111.........617 456-1689
Carter Harry F
26 Runng Brk Rd W Rox 02132.........617 325-5465
Carter Hide Co Inc
146 Summer Bos 02110.........617 542-7987
Carter Hilary 61 Harvey Cam 02140.........617 876-2750
Horace
241 Walnut Av Roxbury 02119.........617 442-5307
Howard Jr 26 Notre Dme Rox 02119.617 445-5552
J Cam.........617 354-2688
J 15 Chatham Bro 02446.........617 232-7990
J 518 Harvard Bro 02446.........617 730-9483
J 775 Vfw Pkwy West Roxbury 02132.........617 323-5574
Carter J Jacques MD
1 Brookline Pl Bro 02446.........617 735-8787
Carter J M
1410 Columbia Rd S Bos 02127.........617 464-1040
Carter J M Ornamental Ironworks
Call.........Pembroke TelNo-617 436-5353
Carter J Veal Co
48 Newmarket Sq Rox 02118.........617 442-1775
Carter James
1573 Cambridge St Cam 02138.........617 492-1214
James 182 Fisher Av Roxbury 02120.........617 739-2193
James
37 Gold Star Rd Cambridge 02140.........617 876-8841
Jas L 14 Roseberry Rd Mat 02126.........617 361-0773
Jane 114 Adena Rd Newton 02465.........617 964-0435
Jeffrey 41 Warren Av Bos 02116.........617 426-5994
John 11 Mansfield Bri 02134.........617 987-2163
John 327 Summer Bos 02210.........617 423-4334
John 40 Westwind Rd Dor 02125.........617 282-1235
June O 329 A Summit Av Bri 02135.........617 734-6109
K 38 Browning Av Dorchester 02124.........617 265-8456
K 17 Esmond Dorchester 02121.........617 282-1593

Carter Nella E
333 Maschsts Av Bos 02115.........617 267-6483
Nicholas S F
115 Randolph Av Mil 02186.........617 698-5307
Nick 21 Fairfield Bos 02116.........617 267-5222
Nick & Debbi
196 Herrick Rd Newton 02459.........617 527-0480
Nicole.........617 698-0713
Norman G
38 Chickatawbut Dor 02122.........617 822-1203
P 94 Crestwood Pk Rox 02121.........617 427-4754
P E 501 E Sixth S Bos 02127.........617 268-4213
P L 44 Hutchings Rox 02121.........617 427-9170
P R 91 Bynner Jam 02130.........617 983-8692
Paul & Constance
114 Anawan Av W Rox 02132.........617 325-2036
Paul E 501 E Sixth St S Bos 02127.........617 268-4546
Paul M 27 Union Bri 02135.........617 787-2115
Carter Pile Driving Inc 17 Beaver Ct
Framingham 01702......... Wellesley TelNo-781 235-8488
Carter Prudence
46 Franklin Watertown 02172.........617 393-3782
Prudence
46 Franklin Watertown 02172.........617 926-7063
Reginald
106 Brunswick Dorchester 02121.........617 541-2843
Renee & Andrew
10 Walnut Bos 02108.........617 720-3765
Carter Rice Dowd
Bulkley Dunton Publishing 163 Main Wilmington 01887
Toll Free-Dial '1' & Then.........800 638-1671
Cust Svc-Industrial Prod 613 Main Wilmington
Toll Free-Dial '1' & Then.........800 619-7447
Cust Svc-Printing 613 Main Wilmington
Toll Free-Dial '1' & Then.........800 648-7447
Headquarters 613 Main Wilmington 01887
Call.........978 988-7447
Ingalls Cronin 163 Main Wilmington 01887
Toll Free-Dial '1' & Then.........800 638-1673
Carter Richard
1079 Comnwlth Av Brighton 02215.........617 987-0836
Richard A 97 Mt Vernon Bos 02108.........617 566-7293
Carter Richard A MD
170 Comnwlth Av Bos 02116.........617 267-0710
Carter Richard K
15 Mercer S Bos 02127.........617 268-0448
Robert L 175 Richdale Av Cam 02140.617 864-1535
Roger 150 St Botolph Bos 02115.........617 424-6148
Roy 44 Concord Av Cam 02138.........617 491-6115
Royce 18 Seminary Cha 02129.........617 241-0418

C

$\rightsquigarrow$

Our Idea: Meaning Through Geography

Set of clusterings ≈
A list of unconnected addresses


wide at **SuperPages.com** 195 **Car** **C**

Cartage New England Inc
26 Allen Ln Ipswich 01938....978 356-9960
Cartagema Lydia
18 Jewett Ros 02131....617 323-7639
Cartagena Avith
9 Bancroft Rox 02119....617 442-9780
B Hyd 02136....617 361-5253
Jessica 50 Decatur Cha 02129....617 241-0152
Lucilla 174 Harvard Cam 02139....617 491-5621
M 95 Rowe Ros 02131....617 323-9713
Melvin 501 Green Cam 02139....617 576-1061
Carte Nicholas
18 Appleton Boston 02116....617 695-6996
Cartegena O 4 Milford Bos 02118....617 338-8219
Carten Thos J Sr & Claire
1 Paradise Rd Mil 02186....617 698-6163
Thomas & Kathleen
50 Thompson Ln Mil 02186....617 696-6919
Carter A Ros 02131....617 327-2257
A Roxbury....617 442-5230
A 31 Bethune Wy Roxbury 02119....617 442-1219
A 260 Putnam Av Cambridge 02139....617 492-4174
A M 255 Maschsts Av Bos 02115....617 266-7153
Adams 361 Centre St Mil 02186....617 698-9074
Alice 108 Kilmarnock Bos 02215....617 425-0193
Alice 45 Market Cambridge 02139....617 945-2711
Andrew F 62 Vinal Av Som 02143....617 625-7623
Carter Anne MD
1101 Beacon Bro 02446....617 739-1022
Carter Athens
272 Newbury Boston 02116....617 536-6329
B E 68 Gladeside Av Mat 02126....617 296-6911
Carter Barbara L MD
Tufts-New England Medical Center Bos 02111
Call....617 636-0051
Carter Becky Bos 02114....617 523-4368
Bernard J
112 Gladstone E Bos 02128....617 567-3430
Bithiah 25 Medway Dor 02124....617 298-8713
Blake 26 Mt Vernon Bos 02108....617 367-9931
Carter Broadcasting Co
20 Park Plz Bos 02116....617 423-0210
Carter & Burgess Consultants Inc
23 East St Cam 02141....617 225-0200
Carter C 2000 Comnwlth Av Bri 02135....617 782-2118
C 228 Faywood Av East Boston 02128....617 569-1545
C 359 Harvard Cam 02138....617 491-4822
C 610 Walk Hill Mat 02126....617 296-6392
C & M 43 Burroughs Jam 02130....617 524-9558

Carter F 24 Hillock Ros 02131....617 327-1105
Faye & Ricky
357 Columbus Av Bos 02116....617 437-7331
Francis S 134 Temple W Rox 02132..617 323-6781
Franklin & Anne
221 Mt Auburn Cam 02138....617 354-0798
Fred 42 Haverford Jam 02130....617 524-3078
Fred 96 Hinckley Rd Mil 02186....617 698-1343
G & R 8 Verdun Dor 02124....617 436-8906
G T 27 Franklin Av Som 02145....617 623-7121
Gayle 25 Frontenac Dor 02124....617 825-0322
Geo S 115 Moss Hill Rd Jam 02130....617 522-3215
George 125 Nashua Bos 02114....617 367-9548
Carter Halliday Associate
107 S Street Bos 02111....617 456-1689
Carter Harry F
26 Runng Brk Rd W Rox 02132....617 325-5465
Carter Hide Co Inc
146 Summer Bos 02110....617 542-7987
Carter Hilary 61 Harvey Cam 02140....617 876-2750
Horace
241 Walnut Av Roxbury 02119....617 442-5307
Howard Jr 26 Notre Dme Rox 02119.617 445-5552
J Cam....617 354-2688
J 15 Chatham Bro 02446....617 232-7990
J 518 Harvard Bro 02446....617 730-9483
J 775 Vfw Pkwy West Roxbury 02132....617 323-5574
Carter J Jacques MD
1 Brookline Pl Bro 02446....617 735-8787
Carter J M
1410 Columbia Rd S Bos 02127....617 464-1040
Carter J M Ornamental Ironworks
Call....Pembroke TelNo-617 436-5353
Carter J Veal Co
48 Newmarket Sq Rox 02118....617 442-1775
Carter James
1573 Cambridge St Cam 02138....617 492-1214
James 182 Fisher Av Roxbury 02120...617 739-2193
James
37 Gold Star Rd Cambridge 02140....617 876-8841
Jas L 14 Roseberry Rd Mat 02126....617 361-0773
Jane 114 Adena Rd Newton 02465....617 964-0435
Jeffrey 41 Warren Av Bos 02116....617 426-5994
John 11 Mansfield Bri 02134....617 987-2163
John 327 Summer Bos 02210....617 423-4334
John 40 Westwind Rd Dor 02125....617 282-1235
June O 329 A Summit Av Bri 02135....617 734-6109
K 38 Browning Av Dorchester 02124....617 265-8456
K 17 Esmond Dorchester 02121....617 282-1593

Carter Nella E
333 Maschsts Av Bos 02115....617 267-6483
Nicholas S F
115 Randolph Av Mil 02186....617 698-5307
Nick 21 Fairfield Bos 02116....617 267-5222
Nick & Debbi
196 Herrick Rd Newton 02459....617 527-0480
Nicole....617 698-0713
Norman G
38 Chickatawbut Dor 02122....617 822-1203
P 94 Crestwood Pk Rox 02121....617 427-4754
P E 501 E Sixth S Bos 02127....617 268-4213
P L 44 Hutchings Rox 02121....617 427-9170
P R 91 Bynner Jam 02130....617 983-8692
Paul & Constance
114 Anawan Av W Rox 02132....617 325-2036
Paul E 501 E Sixth St S Bos 02127....617 268-4546
Paul M 27 Union Bri 02135....617 787-2115
Carter Pile Driving Inc 17 Beaver Ct
Framingham 01702.... Wellesley TelNo-781 235-8488
Carter Prudence
46 Franklin Watertown 02172....617 393-3782
Prudence
46 Franklin Watertown 02172....617 926-7063
Reginald
106 Brunswick Dorchester 02121....617 541-2843
Renee & Andrew
10 Walnut Bos 02108....617 720-3765
Carter Rice Dowd
Bulkley Dunton Publishing 163 Main Wilmington 01887
Toll Free-Dial '1' & Then....800 638-1671
Cust Svc-Industrial Prod 613 Main Wilmington
Toll Free-Dial '1' & Then....800 619-7447
Cust Svc-Printing 613 Main Wilmington
Toll Free-Dial '1' & Then....800 648-7447
Headquarters 613 Main Wilmington 01887
Call....978 988-7447
Ingalls Cronin 163 Main Wilmington 01887
Toll Free-Dial '1' & Then....800 638-1673
Carter Richard
1079 Comnwlth Av Brighton 02215...617 987-0836
Richard A 97 Mt Vernon Bos 02108...617 566-7293
Carter Richard A MD
170 Comnwlth Av Bos 02116....617 267-0710
Carter Richard K
15 Mercer S Bos 02127....617 268-0448
Robert L 175 Richdale Av Cam 02140. 617 864-1535
Roger 150 St Botolph Bos 02115....617 424-6148
Roy 44 Concord Av Cam 02138....617 491-6115
Royce 18 Seminary Cha 02129....617 241-0418

C

$\rightsquigarrow$

$\rightsquigarrow$ 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)

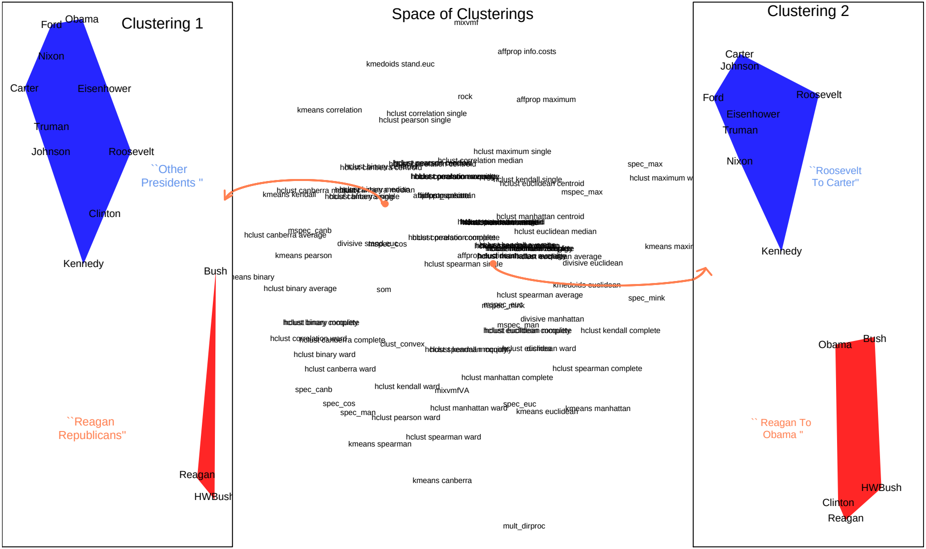
A New Strategy

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. $\rightsquigarrow$ Millions of clusterings, easily comprehended (takes about 10-15 minutes to choose a clustering with insight)

Many Thousands of Clusterings, Sorted & Organized

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



Application-Independent Distance Metric: Axioms

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions

Application-Independent Distance Metric: Axioms

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

Application-Independent Distance Metric: Axioms

- Metric based on 3 assumptions
  1. Distance between clusterings: a function of the pairwise document agreements (pairwise agreements ⇒ 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 ⇒ 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(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(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 ⇒ 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(num clusters)
- ⇝ 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 ⇒ 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 ⇒ RA coders
  - Informative discoveries ⇒ 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 ⇒ RA coders
  - Informative discoveries ⇒ Experienced scholars analyzing texts
  - Discovery ⇒ 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
  - $\Longrightarrow$ 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
  - $\Longrightarrow$ 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
  - $\Longrightarrow$ 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
  - $\Longrightarrow$ 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
  - $\Longrightarrow$ 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
  - $\Longrightarrow$ 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)

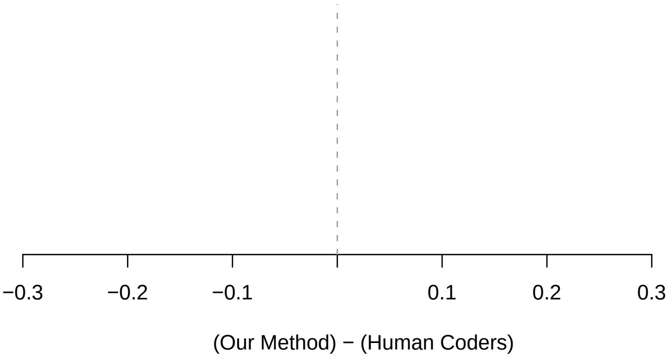
Evaluation 1: Cluster Quality

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

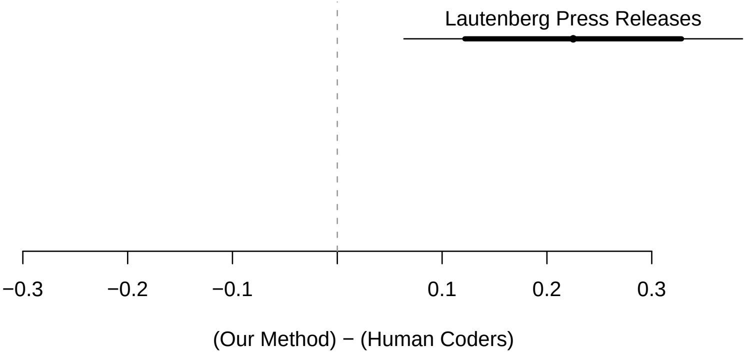
Evaluation 1: Cluster Quality

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

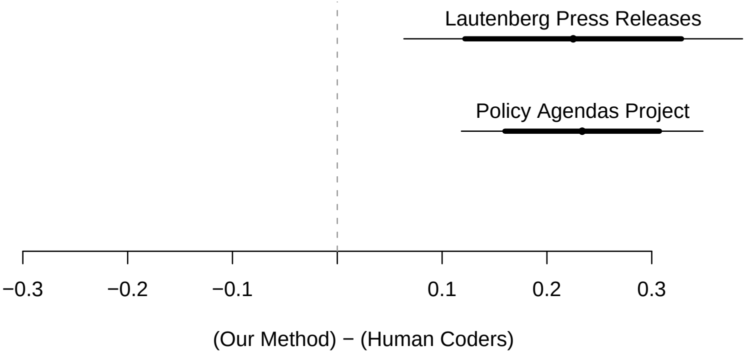


Evaluation 1: Cluster Quality



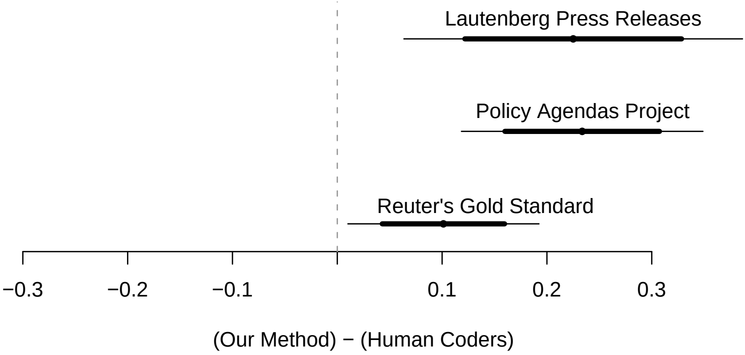
Lautenberg: 200 Senate Press Releases (appropriations, economy, education, tax, veterans, . . .)

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 ⇒ 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 → vMF 1 → vMF 2 → Our Method 2 → K-Means 1 → K-Means 2

"Genetic testing":

Our Method 1 → {Our Method 2, K-Means 1, K-means 2} → Dir Proc. 1 → 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?

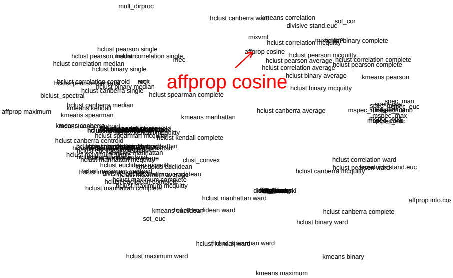
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)

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

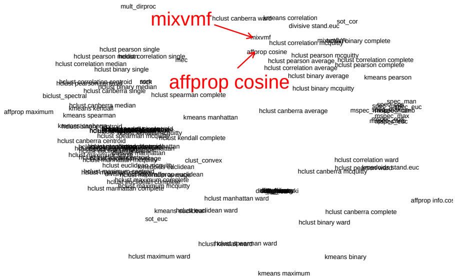
Example Discovery

Example Discovery



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

Example Discovery



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

Close to:

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

Example Discovery

Space between methods:

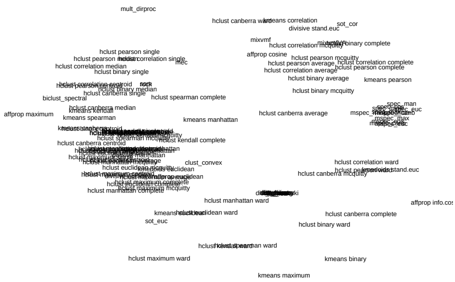
Example Discovery

Space between methods:

Example Discovery

Space between methods:
local cluster ensemble

Example Discovery



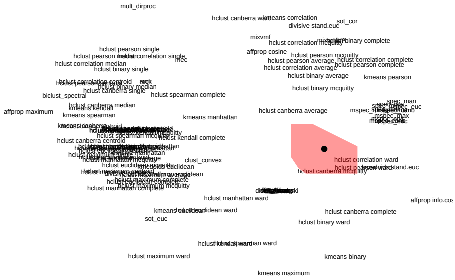
Example Discovery

Found a region with particularly insightful clusterings

Example Discovery

Mixture:

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

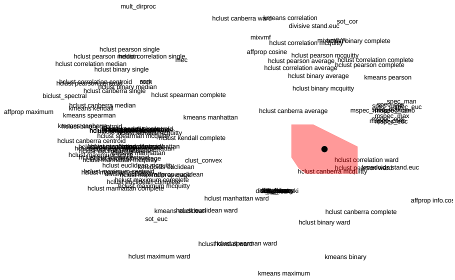
Example Discovery

Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering Random Walk (Metrics 1-6)

Example Discovery



Mixture:

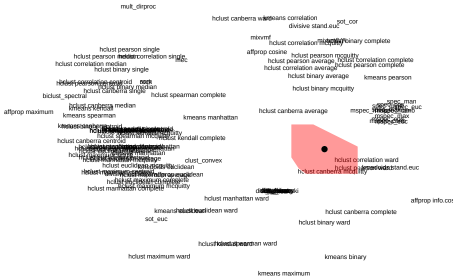
0.39	Hclust-Canberra-McQuitty
0.30	Spectral clustering Random Walk (Metrics 1-6)
0.13	Hclust-Correlation-Ward

Example Discovery

Mixture:

- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward
- 0.09 Hclust-Pearson-Ward

Example Discovery



Mixture:

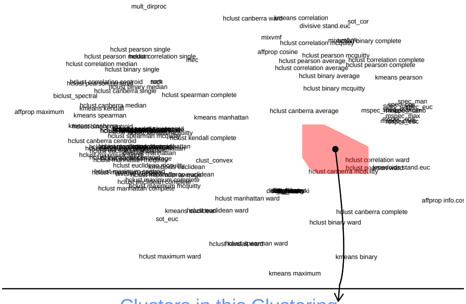
- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward
- 0.09 Hclust-Pearson-Ward
- 0.05 Kmediods-Cosine

Example Discovery

Mixture:

- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward
- 0.09 Hclust-Pearson-Ward
- 0.05 Kmediods-Cosine
- 0.04 Spectral clustering Symmetric (Metrics 1-6)

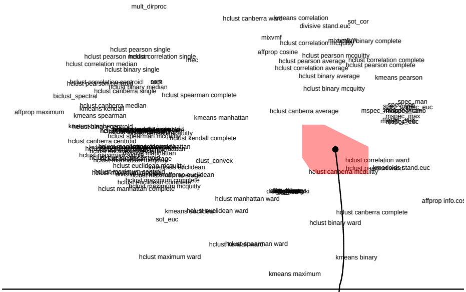
Example Discovery



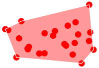
Clusters in this Clustering

Mayhew

Example Discovery



Clusters in this Clustering

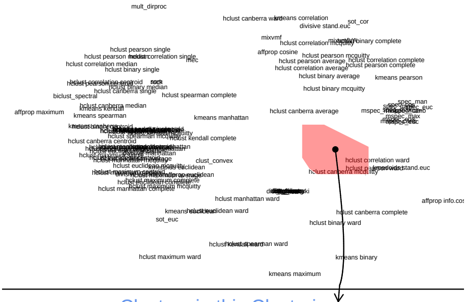


Credit Claiming
Pork

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

Mayhew

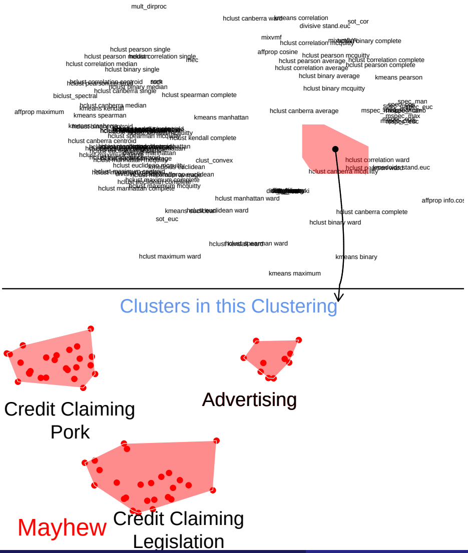
Example Discovery



Clusters in this Clustering

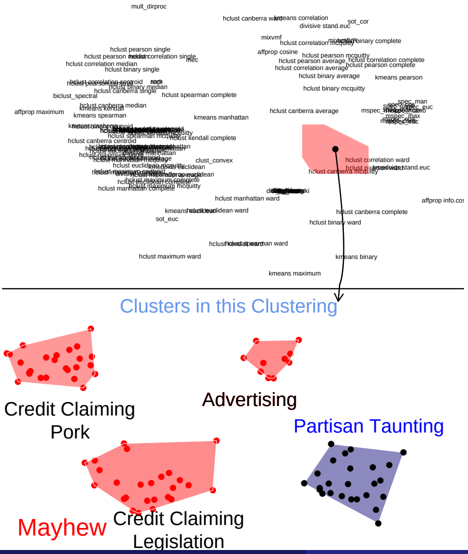
Credit Claiming, Legislation:
"As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period"

Example Discovery



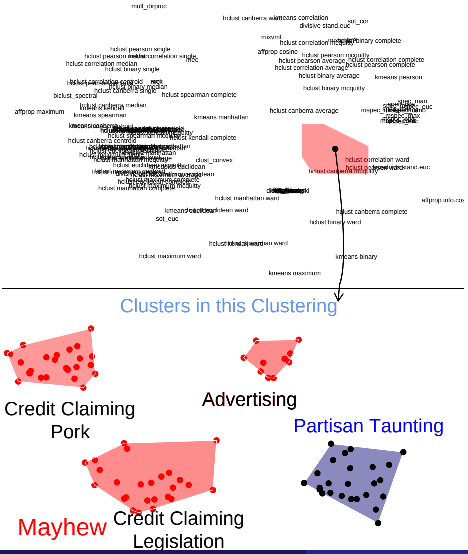
Advertising:
"Senate Adopts Lautenberg/Menendez Resolution Honoring Spelling Bee Champion from New Jersey"

Example Discovery: Partisan Taunting



Partisan Taunting:
"Republicans Selling Out Nation on Chemical Plant Security"

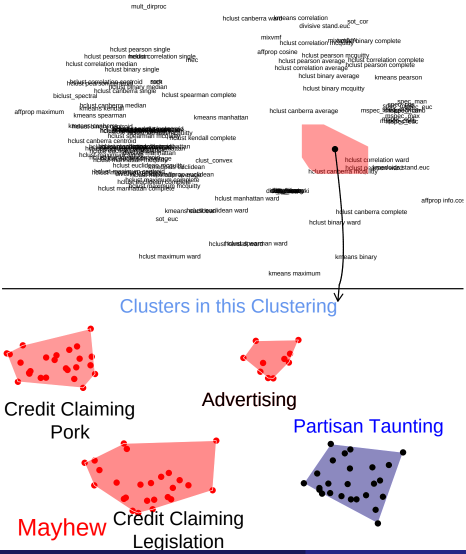
Example Discovery: Partisan Taunting



Partisan Taunting:

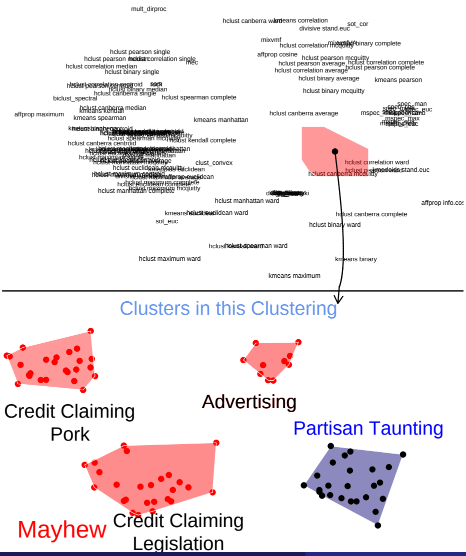
"Senator Lautenberg's amendment would change the name of ...the Republican bill...to 'More Tax Breaks for the Rich and More Debt for Our Grandchildren Deficit Expansion Reconciliation Act of 2006' "

Example Discovery: Partisan Taunting



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

Example Discovery: Partisan Taunting



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

Taunting ruins deliberation

In Sample Illustration of Partisan Taunting

Taunting ruins deliberation

Sen. Lautenberg on Senate Floor 4/29/04

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

In Sample Illustration of Partisan Taunting

Taunting ruins deliberation

Sen. Lautenberg
on Senate Floor
4/29/04

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

Taunting ruins deliberation

Sen. Lautenberg on Senate Floor 4/29/04

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

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

Out of Sample Confirmation of Partisan Taunting

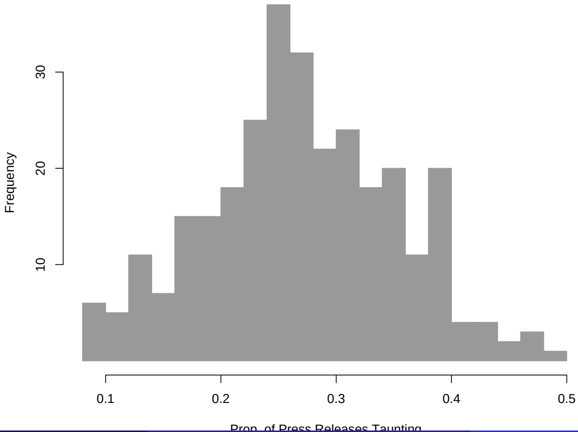
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

Out of Sample Confirmation of Partisan Taunting

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

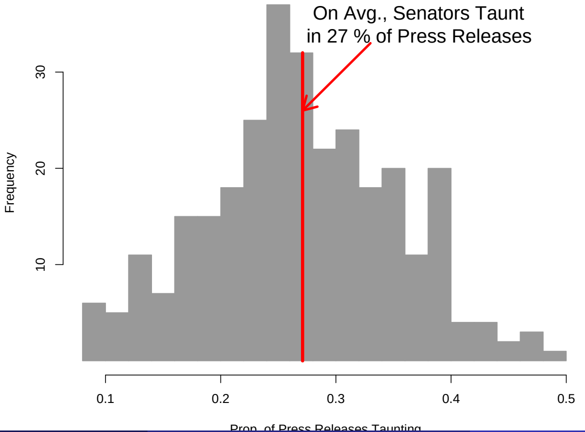
Out of Sample Confirmation of Partisan Taunting

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



Out of Sample Confirmation of Partisan Taunting

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



Advancing the Objective of Discovery

1) Conceptualization

Qualitative Methods (reading!)

2) Measurement

Quantitative Methods

3) Validation

Quantitative methods for conceptualization: aiding discovery

Advancing the Objective of Discovery

1) Conceptualization

Qualitative Methods (reading!)

2) Measurement

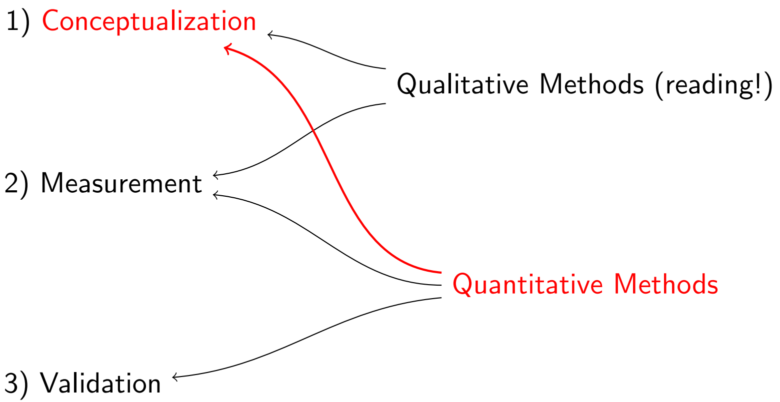
Quantitative Methods

3) Validation

Quantitative methods for conceptualization: aiding discovery

- Few formal methods designed explicitly for conceptualization

Advancing the Objective of Discovery



Quantitative methods for conceptualization: aiding discovery

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

Advancing the Objective of Discovery

1) Conceptualization

Qualitative Methods (reading!)

2) Measurement

Quantitative Methods

3) Validation

Quantitative methods for conceptualization: aiding discovery

- Few formal methods designed explicitly for conceptualization
- Belittled: "Tom Swift and His Electric Factor Analysis Machine" (Armstrong 1967)
- Evaluation methods measure progress in discovery

For more information (on adding zooming out to the human ability to zoom in)

http://GKing.Harvard.edu