"Computer-Assisted Reading" and other Discoveries from Quantitative Social Science

Gary King

Institute for Quantitative Social Science
Harvard University

(*Crimson Conversations* Talk, Riverside, CT, 4/12/12)

The Emergence of Quantitative Social Science

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact:

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity, political campaigns

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity, political campaigns, public health

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity, political campaigns, public health, legal analysis

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity, political campaigns, public health, legal analysis, policing

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity, political campaigns, public health, legal analysis, policing, economics

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity, political campaigns, public health, legal analysis, policing, economics, sports

The Emergence of Quantitative Social Science

(a.k.a. "Big Data," "Data analytics," "data science," etc)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- In depth studies of individual places, people, or events

The Next 50 Years: Spectacular increases in new data sources, due to...

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- Popular versions: *MoneyBall*, *SuperCrunchers*, *The Numerati*,
- The replication movement: academic data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, human expressive capacity, political campaigns, public health, legal analysis, policing, economics, sports, and public policy

Examples of what's now possible with QSS

Examples of what's now possible with QSS

- Opinions of activists:

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise:

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week?

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts:

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends"

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, electronic address books

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, electronic address books
- Economic development in developing countries:

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, electronic address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, electronic address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, or networks of roads and other infrastructure

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, electronic address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, or networks of roads and other infrastructure
- Expert-vs-Statistician contests: Whenever enough information is quantified (& a right answer exists), stats wins

Examples of what's now possible with QSS

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 3.3Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, electronic address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, or networks of roads and other infrastructure
- Expert-vs-Statistician contests: Whenever enough information is quantified (& a right answer exists), stats wins
- Many, **many**, more. . .

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:
  - Then: Quill tip pen & expensive paper and a few books and articles

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:
  - Then: Quill tip pen & expensive paper and a few books and articles
  - Now: Microsoft Word

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:
  - Then: Quill tip pen & expensive paper and a few books and articles
  - Now: Microsoft Word and Huge pile of books and articles

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:
  - Then: Quill tip pen & expensive paper and a few books and articles
  - Now: Microsoft Word and Huge pile of books and articles
- How has reading changed?

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:
  - Then: Quill tip pen & expensive paper and a few books and articles
  - Now: Microsoft Word and Huge pile of books and articles
- How has reading changed?
  - 100 years ago: Get book; read cover to cover

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:
  - Then: Quill tip pen & expensive paper and a few books and articles
  - Now: Microsoft Word and Huge pile of books and articles
- How has reading changed?
  - 100 years ago: Get book; read cover to cover
  - Now: When did you last read a book cover-to-cover (for work)?

Progress in Reading and Writing

- Improvements for knowledge workers over 200 years:
  - Then: Quill tip pen & expensive paper and a few books and articles
  - Now: Microsoft Word and Huge pile of books and articles
- How has reading changed?
  - 100 years ago: Get book; read cover to cover
  - Now: When did you last read a book cover-to-cover (for work)?
  - We now read a tiny fraction haphazardly, and delude ourselves into thinking we understand all we need

Computer-Assisted Reading

Computer-Assisted Reading

- To understand many documents, humans create categories

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help
- Computer-Assisted Clustering

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help
- Computer-Assisted Clustering
  - Easy in theory: list all clusterings; choose the best

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help
- Computer-Assisted Clustering
  - Easy in theory: list all clusterings; choose the best
  - Impossible in practice: Too hard for us mere humans!

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help
- 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

Computer-Assisted Reading

- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help
- 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

Computer-Assisted Reading

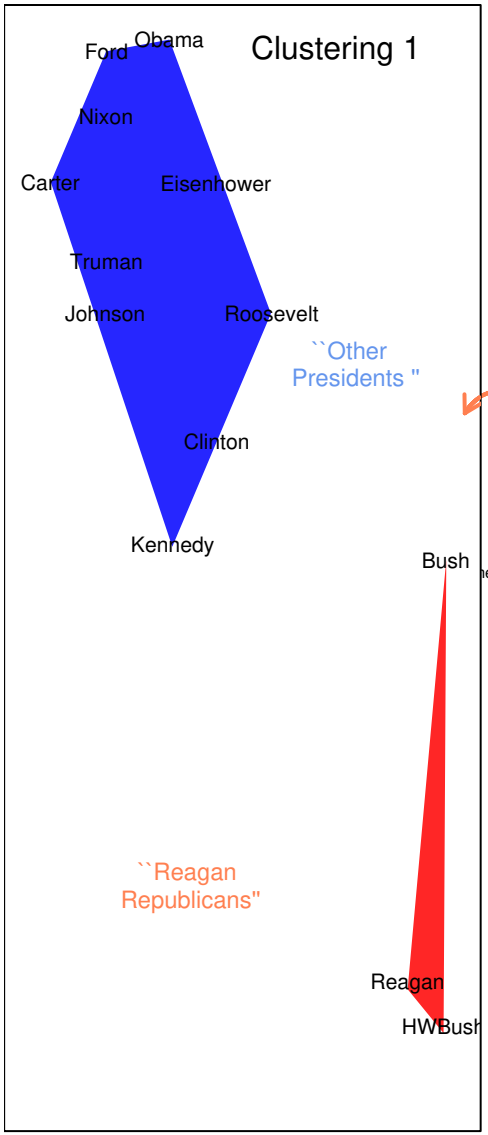
- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help
- 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

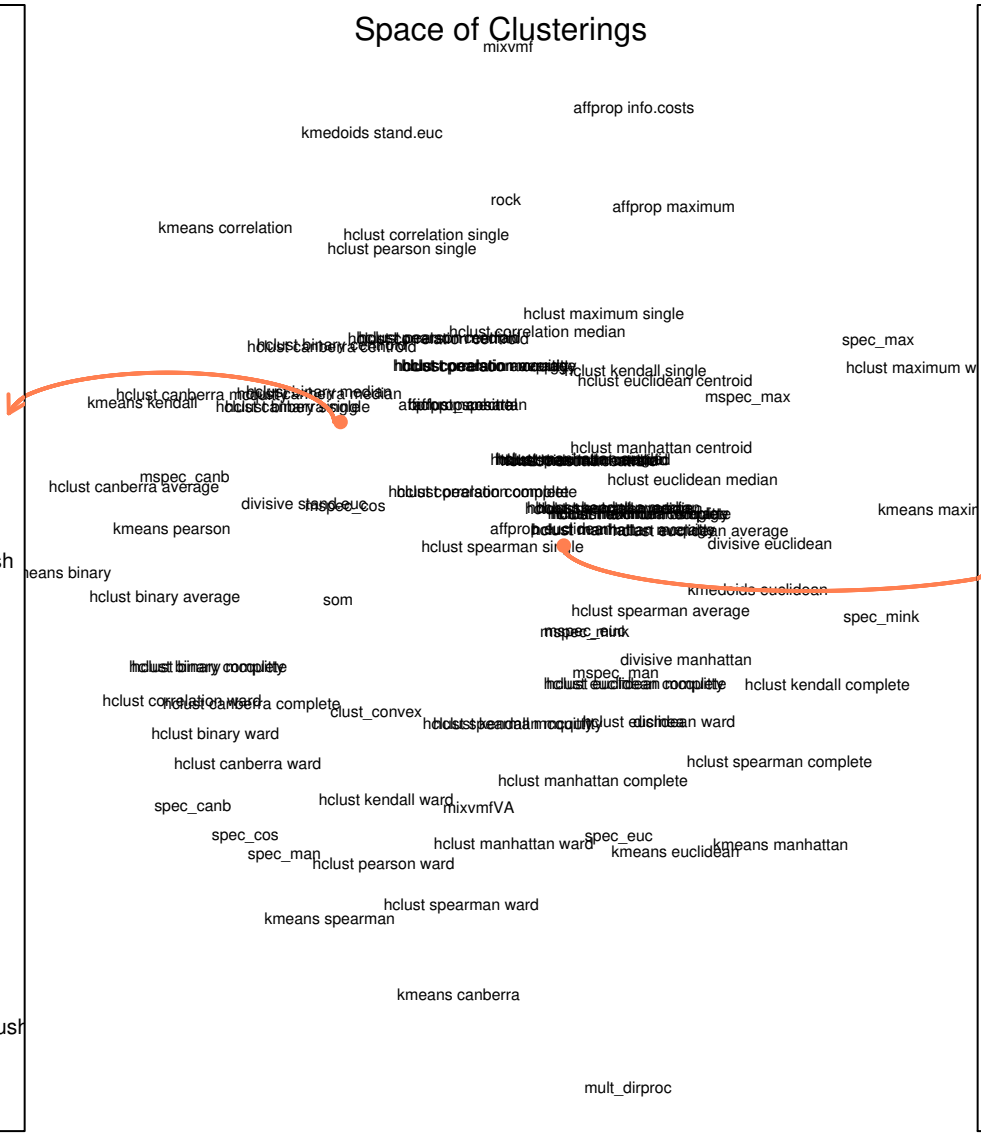
Computer-Assisted Reading

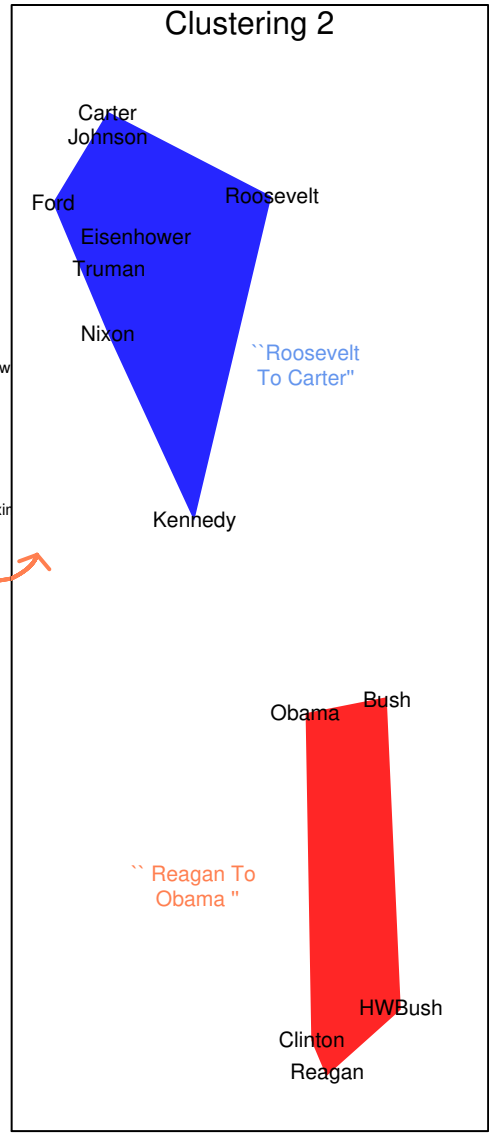
- To understand many documents, humans create categories
- Approaches
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated Cluster Analysis: no method works well in general; impossible to know which to apply!
  - Our Computer-assisted Methods: You, not some computer algorithm, decides what's important, but with help
- 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?

Humans Can Zoom in to Read; We Can Zoom Out

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







Evaluation: More Informative Discoveries

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for $>$ 1 year each

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:
  - human-generated (by these scholars, working for a year each)

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:
  - human-generated (by these scholars, working for a year each)
  - fully-automated computer-generated

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:
  - human-generated (by these scholars, working for a year each)
  - fully-automated computer-generated
  - computer-assisted generation (biased against us; took about an hour)

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:
  - human-generated (by these scholars, working for a year each)
  - fully-automated computer-generated
  - computer-assisted generation (biased against us; took about an hour)
- Conducted an evaluation; the scholar was the judge

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:
  - human-generated (by these scholars, working for a year each)
  - fully-automated computer-generated
  - computer-assisted generation (biased against us; took about an hour)
- Conducted an evaluation; the scholar was the judge
- Same result in each case:

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:
  - human-generated (by these scholars, working for a year each)
  - fully-automated computer-generated
  - computer-assisted generation (biased against us; took about an hour)
- Conducted an evaluation; the scholar was the judge
- Same result in each case:
  - Computer-assisted clustering won both competitions

Evaluation: More Informative Discoveries

- 2 scholars reading in archives for > 1 year each
- Separate Competitions among clusterings:
  - human-generated (by these scholars, working for a year each)
  - fully-automated computer-generated
  - computer-assisted generation (biased against us; took about an hour)
- Conducted an evaluation; the scholar was the judge
- Same result in each case:
  - Computer-assisted clustering won both competitions
  - Both scholars preferred our insight to their's

Evaluation: What Do Members of Congress Do?

Evaluation: What Do Members of Congress Do?

- David Mayhew's (1974) famous typology

Evaluation: What Do Members of Congress Do?

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

Evaluation: What Do Members of Congress Do?

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

Evaluation: What Do Members of Congress Do?

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

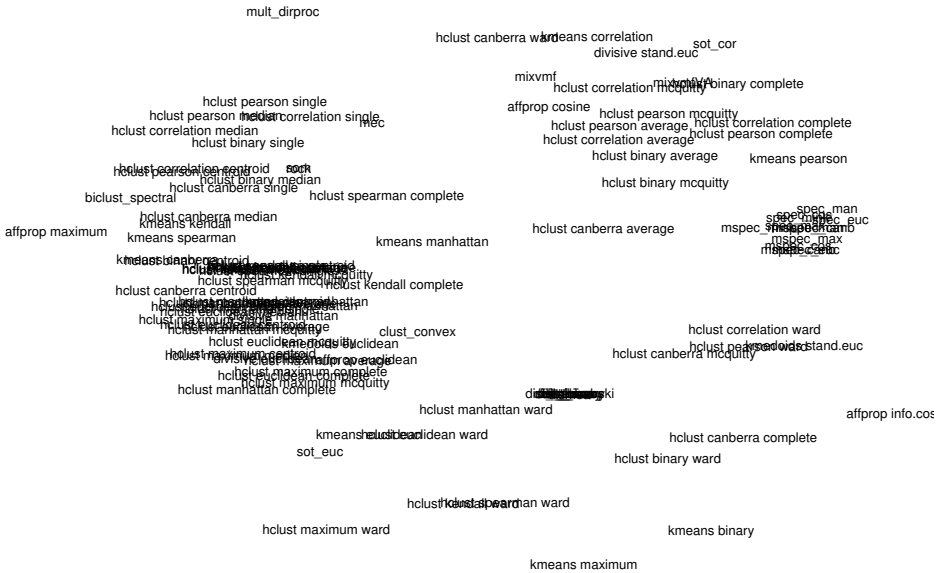
Evaluation: What Do Members of Congress Do?

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

Evaluation: 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



The space of clusterings

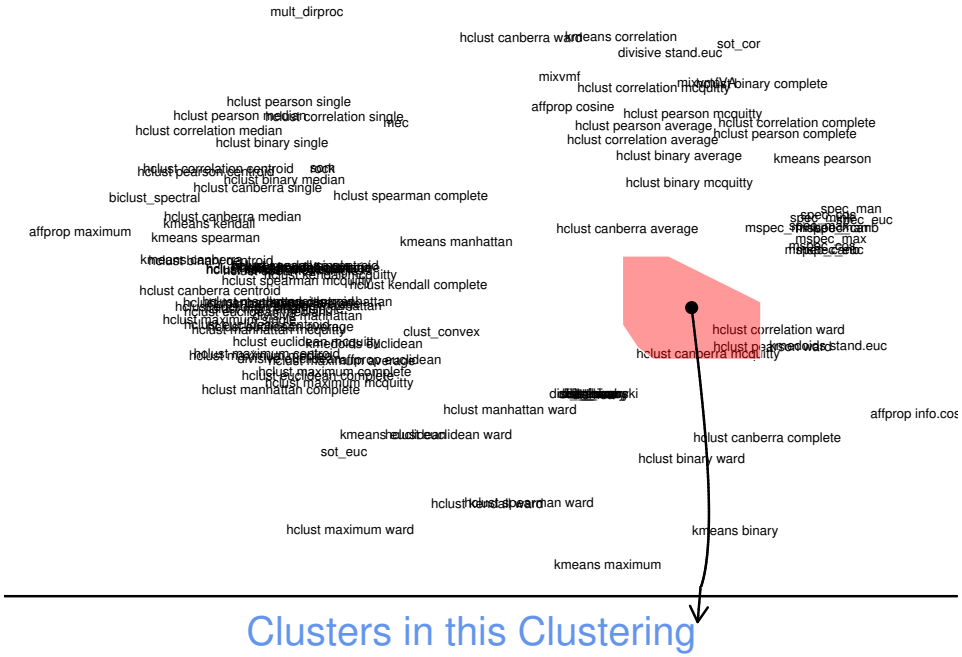
Example Discovery

Found a region with particularly insightful clusterings

Example Discovery

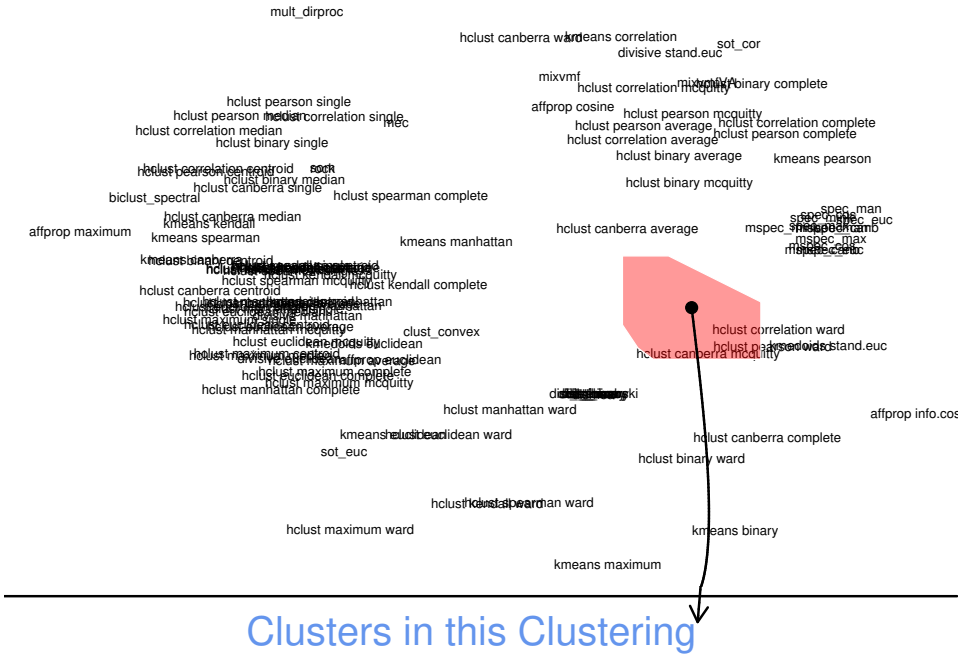
Let's look at **one clustering** in this region

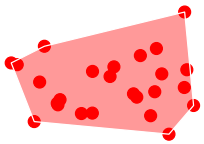
Example Discovery



Mayhew

Example Discovery



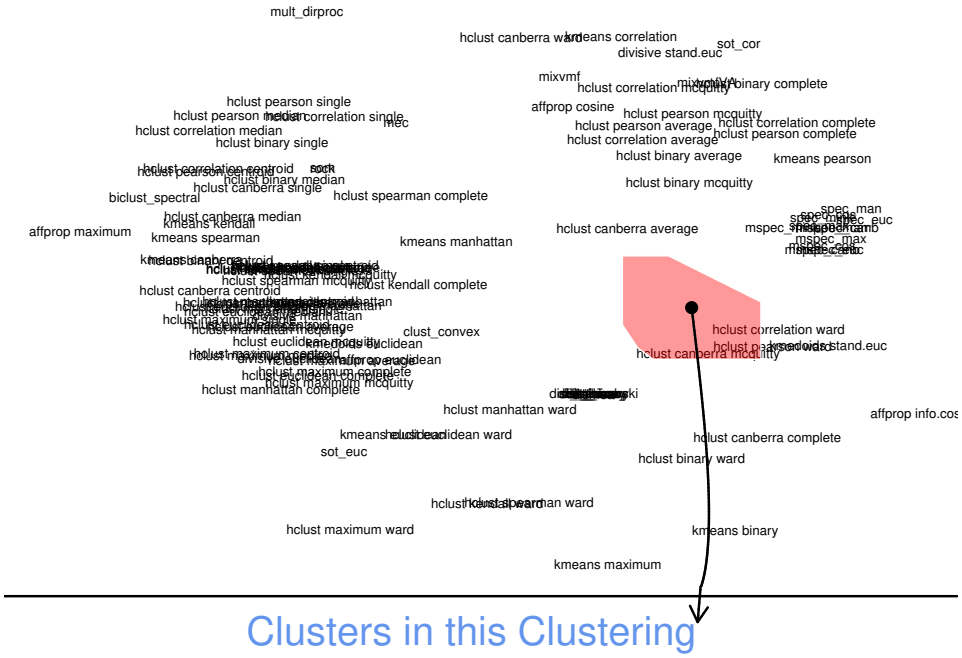


Credit Claiming
Pork

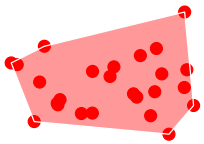
Credit Claiming, Pork:
"Sens. Frank R. Lautenberg (D-NJ) and Robert Menendez (D-NJ) announced that the U.S. Department of Commerce has awarded a $100,000 grant to the South Jersey Economic Development District"

Mayhew

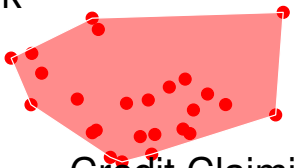
Example Discovery



Clusters in this Clustering



Credit Claiming
Pork

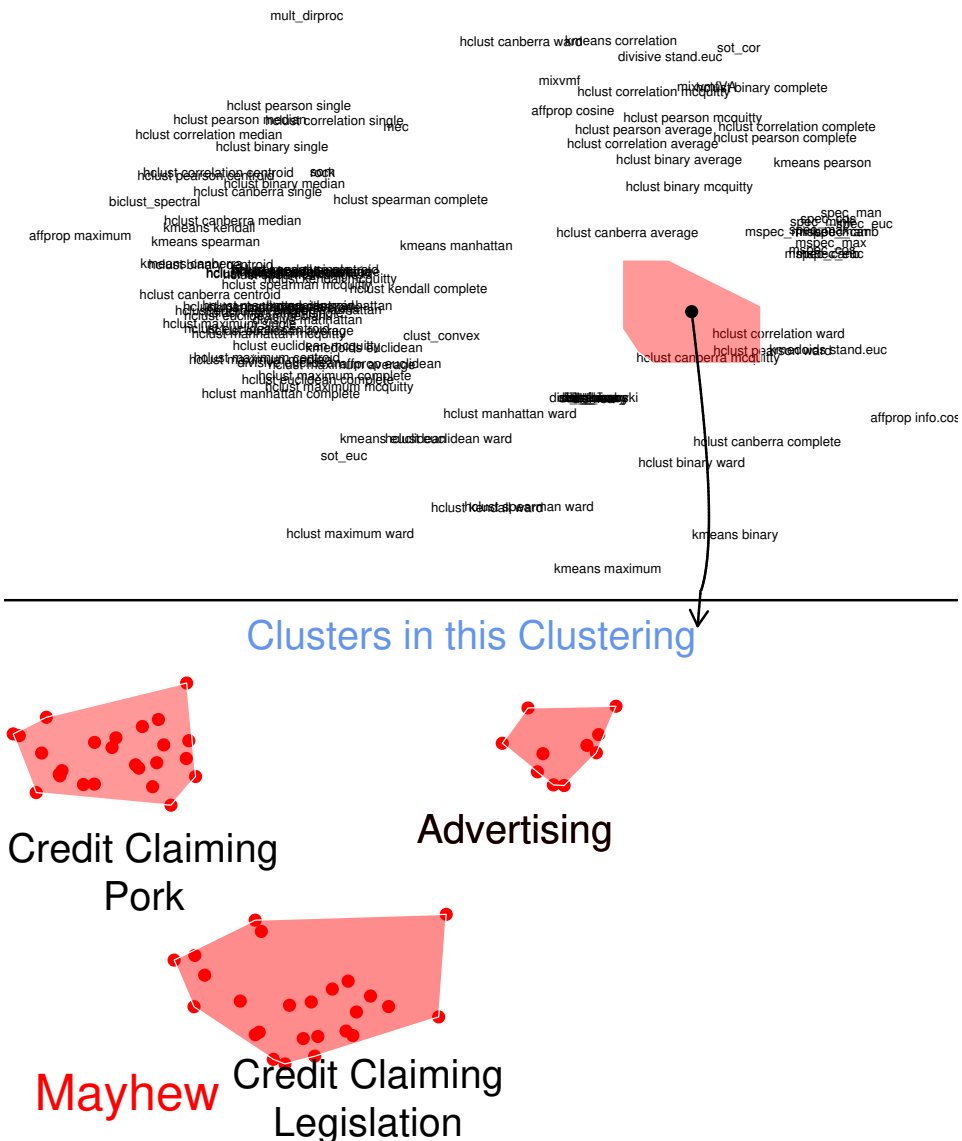


Mayhew

Credit Claiming
Legislation

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"

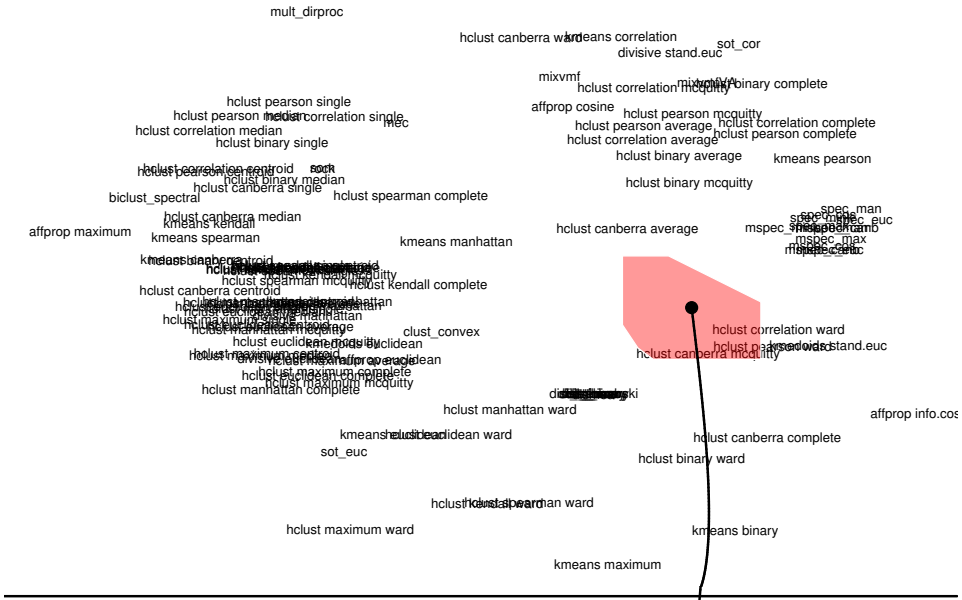
Example Discovery



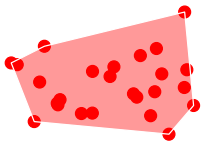
Advertising:

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

Example Discovery



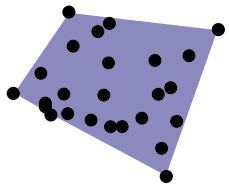
Clusters in this Clustering

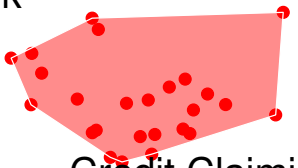


Credit Claiming
Pork

Advertising

Partisan Taunting



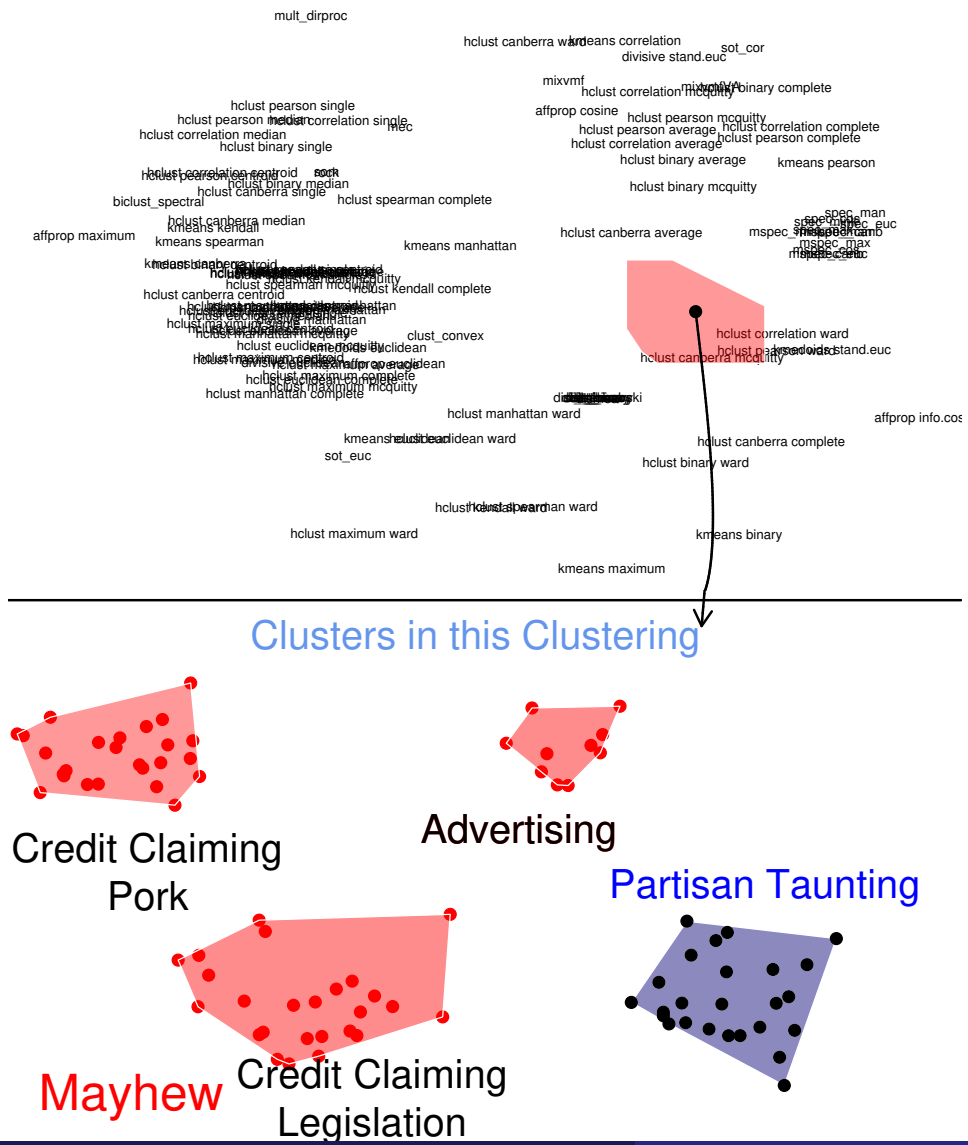


Mayhew

Credit Claiming
Legislation

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

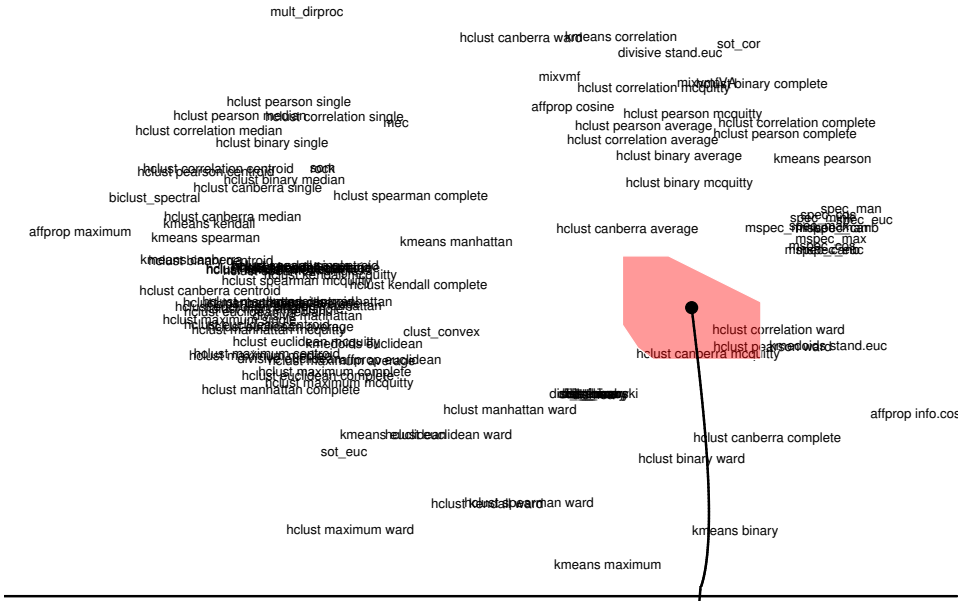
Example Discovery



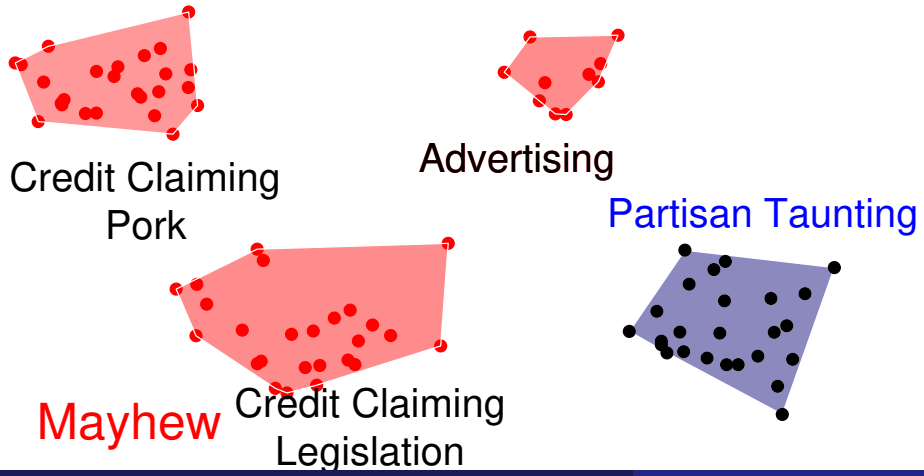
Partisan Taunting:

"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

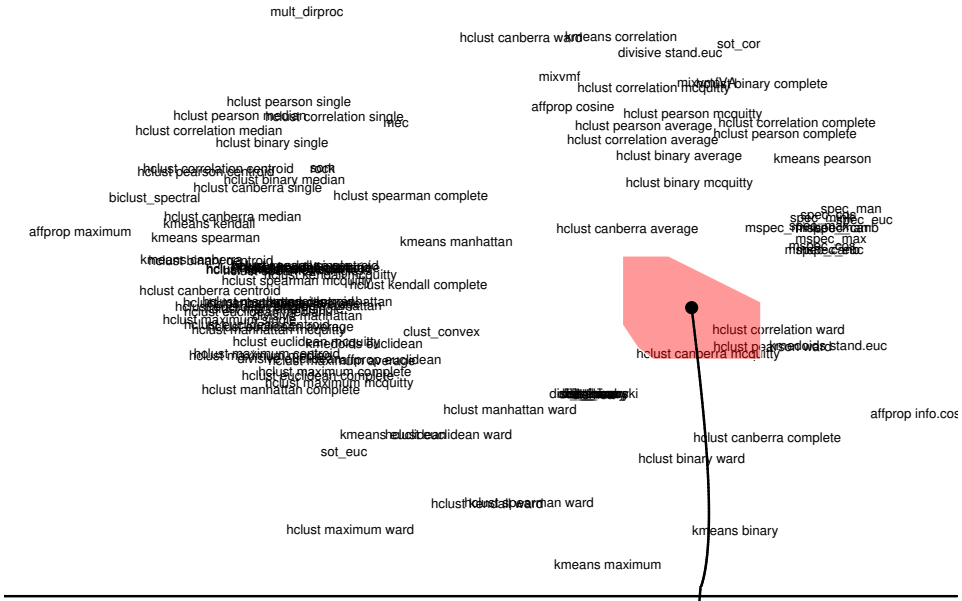


Clusters in this Clustering

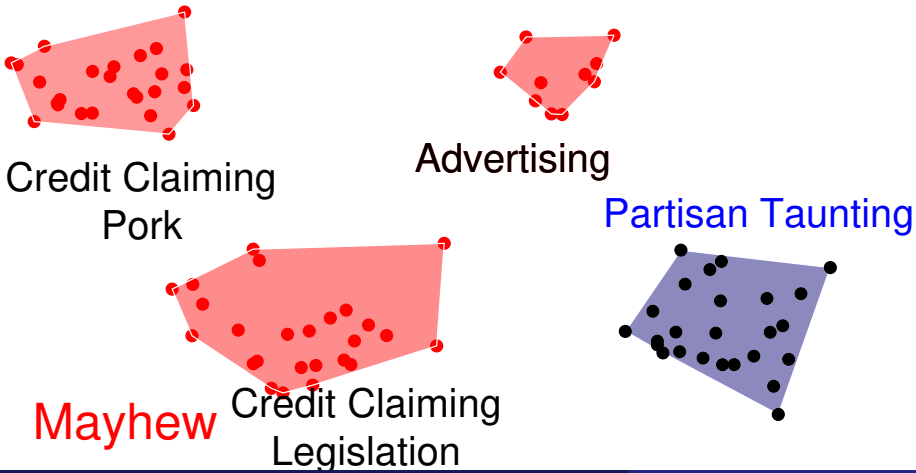


Definition of Partisan Taunting:
Explicit, public, and negative attacks on another political party or its members

Example Discovery



Clusters in this Clustering



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

Taunting ruins deliberation

In Sample Illustration of Partisan Taunting

Taunting ruins deliberation

Sen. Lautenberg
on Senate Floor
4/29/04

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

In Sample Illustration of Partisan Taunting

Taunting ruins deliberation

Sen. Lautenberg
on Senate Floor
4/29/04

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

In Sample Illustration of Partisan Taunting

Taunting ruins deliberation

Sen. Lautenberg
on Senate Floor
4/29/04

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

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

Out of Sample Confirmation of Partisan Taunting

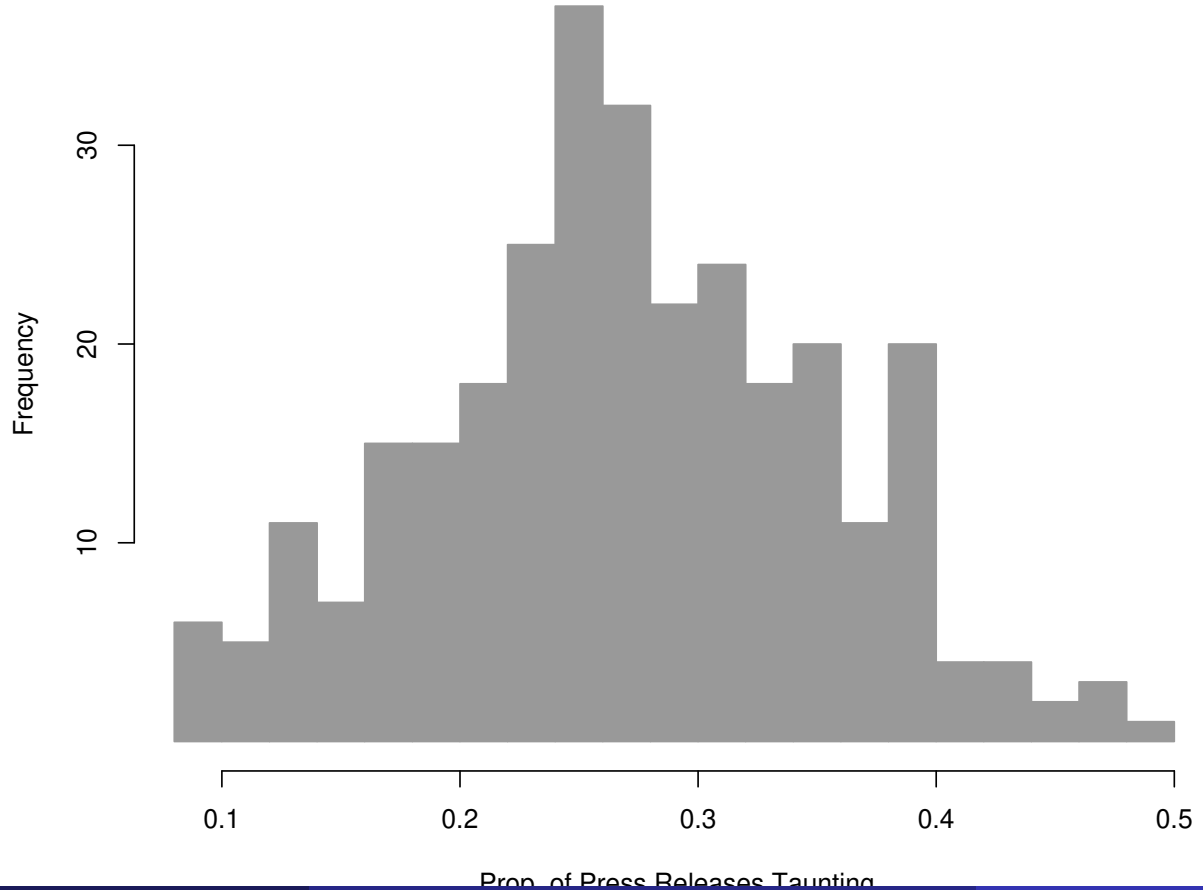
- Discovered using 200 press releases; 1 senator.
- Confirmed using 64,033 press releases; 301 senator-years.

Out of Sample Confirmation of Partisan Taunting

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

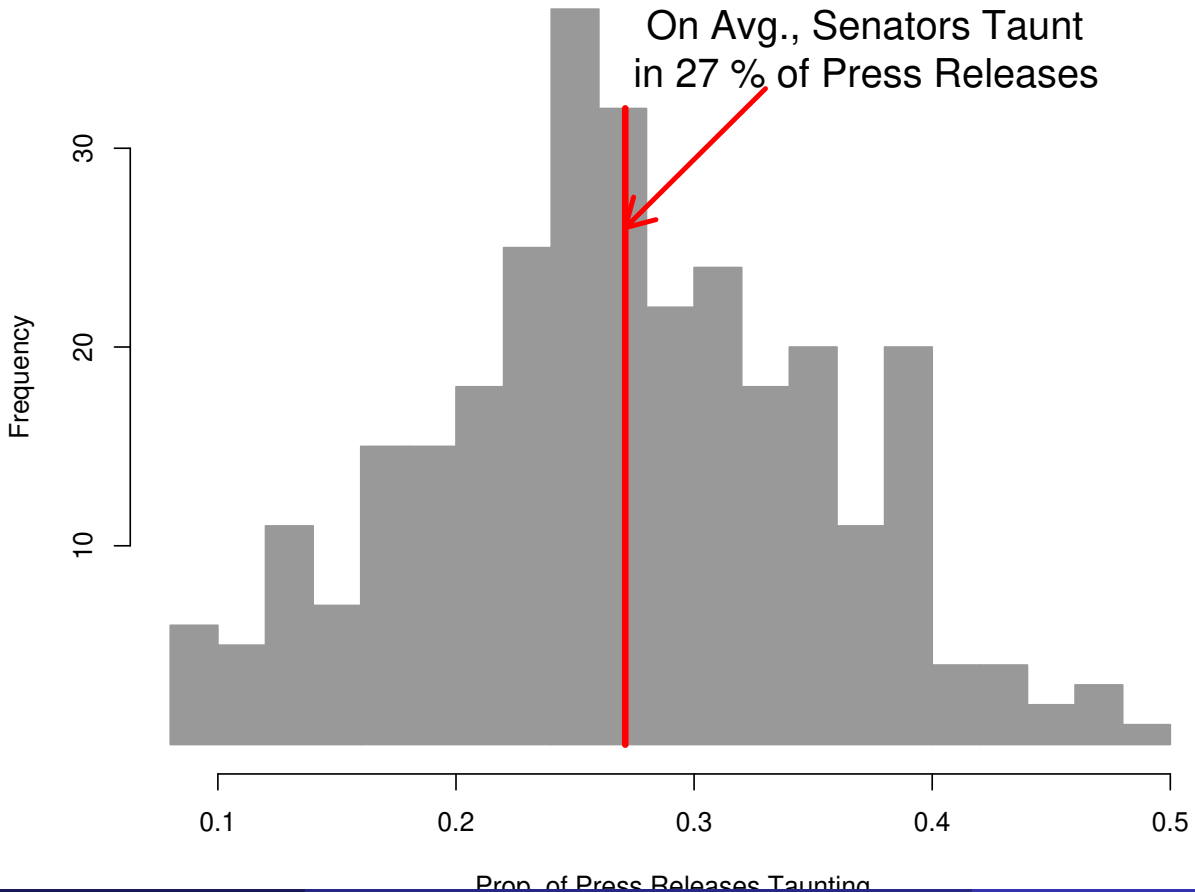
Out of Sample Confirmation of Partisan Taunting

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



Out of Sample Confirmation of Partisan Taunting

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



Some New Data Types

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature
2. Commercial activity: credit cards, sales data, and real estate transactions, product RFIDs

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature
2. Commercial activity: credit cards, sales data, and real estate transactions, product RFIDs
3. Geographic location: cell phones, Fastlane or EZPass transponders, garage cameras

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature
2. Commercial activity: credit cards, sales data, and real estate transactions, product RFIDs
3. Geographic location: cell phones, Fastlane or EZPass transponders, garage cameras
4. Health information: digital medical records, hospital admittances, google/MS health, and accelerometers and other devices being included in cell phones

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature
2. Commercial activity: credit cards, sales data, and real estate transactions, product RFIDs
3. Geographic location: cell phones, Fastlane or EZPass transponders, garage cameras
4. Health information: digital medical records, hospital admittances, google/MS health, and accelerometers and other devices being included in cell phones
5. Biological sciences: effectively becoming social sciences as genomics, proteomics, metabolomics, and brain imaging produce huge numbers of *person-level variables*.

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature
2. Commercial activity: credit cards, sales data, and real estate transactions, product RFIDs
3. Geographic location: cell phones, Fastlane or EZPass transponders, garage cameras
4. Health information: digital medical records, hospital admittances, google/MS health, and accelerometers and other devices being included in cell phones
5. Biological sciences: effectively becoming social sciences as genomics, proteomics, metabolomics, and brain imaging produce huge numbers of *person-level variables*.
6. Satellite imagery: increasing in scope, resolution, and availability.

Some New Data Types

1. Unstructured text: emails (1 LOC every 10 minutes), speeches, government reports, blogs, social media updates, web pages, newspapers, scholarly literature
2. Commercial activity: credit cards, sales data, and real estate transactions, product RFIDs
3. Geographic location: cell phones, Fastlane or EZPass transponders, garage cameras
4. Health information: digital medical records, hospital admittances, google/MS health, and accelerometers and other devices being included in cell phones
5. Biological sciences: effectively becoming social sciences as genomics, proteomics, metabolomics, and brain imaging produce huge numbers of *person-level variables*.
6. Satellite imagery: increasing in scope, resolution, and availability.
7. Electoral activity: ballot images, precinct-level results, individual-level registration, primary participation, and campaign contributions

Some More New Data Examples

Some More New Data Examples

8. Social media: facebook, twitter, social bookmarking, blog comments, product reviews, virtual worlds, game behavior, crowd sourcing

Some More New Data Examples

8. Social media: facebook, twitter, social bookmarking, blog comments, product reviews, virtual worlds, game behavior, crowd sourcing
9. Web surfing artifacts: clicks, searches, and advertising clickthroughs. (Google collects 1 petabyte/72 minutes on human behavior!)

Some More New Data Examples

8. Social media: facebook, twitter, social bookmarking, blog comments, product reviews, virtual worlds, game behavior, crowd sourcing
9. Web surfing artifacts: clicks, searches, and advertising clickthroughs. (Google collects 1 petabyte/72 minutes on human behavior!)
10. Multiplayer web games and virtual worlds: Billions of highly controlled experiments on human behavior

Some More New Data Examples

8. Social media: facebook, twitter, social bookmarking, blog comments, product reviews, virtual worlds, game behavior, crowd sourcing
9. Web surfing artifacts: clicks, searches, and advertising clickthroughs. (Google collects 1 petabyte/72 minutes on human behavior!)
10. Multiplayer web games and virtual worlds: Billions of highly controlled experiments on human behavior
11. Government bureaucracies: moving from paper to electronic data bases, increasing availability

Some More New Data Examples

8. Social media: facebook, twitter, social bookmarking, blog comments, product reviews, virtual worlds, game behavior, crowd sourcing
9. Web surfing artifacts: clicks, searches, and advertising clickthroughs. (Google collects 1 petabyte/72 minutes on human behavior!)
10. Multiplayer web games and virtual worlds: Billions of highly controlled experiments on human behavior
11. Government bureaucracies: moving from paper to electronic data bases, increasing availability
12. Governmental policies: requiring more data collection, such e.g., "No Child Left Behind Act"; allowing randomized policy experiments; Obama pushing data distribution

Some More New Data Examples

8. Social media: facebook, twitter, social bookmarking, blog comments, product reviews, virtual worlds, game behavior, crowd sourcing
9. Web surfing artifacts: clicks, searches, and advertising clickthroughs. (Google collects 1 petabyte/72 minutes on human behavior!)
10. Multiplayer web games and virtual worlds: Billions of highly controlled experiments on human behavior
11. Government bureaucracies: moving from paper to electronic data bases, increasing availability
12. Governmental policies: requiring more data collection, such e.g., "No Child Left Behind Act"; allowing randomized policy experiments; Obama pushing data distribution
13. Scholarly data: the replication movement in academia, led in part by political science, is massively increasing data sharing

For more information

http://GKing.Harvard.edu

What's Hard about Clustering?

(Why Johnny Can't Classify)

What's Hard about Clustering?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering

What's Hard about Clustering?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering
- Bell(n) = number of ways of partitioning n objects

What's Hard about Clustering?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering
- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)

What's Hard about Clustering?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering
- 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?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering
- 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?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering
- 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?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering
- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100) $\approx 10^{28} \times$ Number of elementary particles in the universe

What's Hard about Clustering?

(Why Johnny Can't Classify)

- Goal: Computer-assisted conceptualization & clustering
- 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!

Normative Implications of Taunting

- Partisan taunting:

Normative Implications of Taunting

- Partisan taunting:
  - Very common

Normative Implications of Taunting

- Partisan taunting:
  - Very common
  - Makes deliberation less likely

Normative Implications of Taunting

- Partisan taunting:
  - Very common
  - Makes deliberation less likely
  - Occurs more often in homogeneously partisan districts (i.e., when preaching to the choir)

Normative Implications of Taunting

- Partisan taunting:
  - Very common
  - Makes deliberation less likely
  - Occurs more often in homogeneously partisan districts (i.e., when preaching to the choir)
- Incompatibility of the principles of representative democracy

Normative Implications of Taunting

- Partisan taunting:
  - Very common
  - Makes deliberation less likely
  - Occurs more often in homogeneously partisan districts (i.e., when preaching to the choir)
- Incompatibility of the principles of representative democracy
  - To get reflection: Homogeneous (noncompetitive) districts

Normative Implications of Taunting

- Partisan taunting:
  - Very common
  - Makes deliberation less likely
  - Occurs more often in homogeneously partisan districts (i.e., when preaching to the choir)
- Incompatibility of the principles of representative democracy
  - To get reflection: Homogeneous (noncompetitive) districts
  - To get deliberation (no taunting): Heterogeneous (competitive) districts

Normative Implications of Taunting

- Partisan taunting:
  - Very common
  - Makes deliberation less likely
  - Occurs more often in homogeneously partisan districts (i.e., when preaching to the choir)
- Incompatibility of the principles of representative democracy
  - To get reflection: Homogeneous (noncompetitive) districts
  - To get deliberation (no taunting): Heterogeneous (competitive) districts
  - $\rightsquigarrow$ you can't have both!