Big Data is Not About the Data!

Gary King[1]

Institute for Quantitative Social Science
Harvard University

Conference on the Future of Science, Venice, Italy 9/23/2016

[1] GaryKing.org

The *Data* In Big Data (about people)

The *Data* In Big Data (about people)

The Last 50 Years:

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

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

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

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

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact:

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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 *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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 *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health, legal analysis

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health, legal analysis, policing

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health, legal analysis, policing, economics

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health, legal analysis, policing, economics, sports

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health, legal analysis, policing, economics, sports, public policy

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature,

The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: 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, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature, etc., etc., etc

The *Value* in Big Data: the Analytics

The *Value* in Big Data: the Analytics

- Data:

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases

The *Value* in Big Data: the Analytics

- Data:
    - easy to come by; often a free byproduct of IT improvements
    - becoming commoditized
    - Ignore it & every institution will have more every year
    - With a bit of effort: huge data production increases
- Where the Value is: the Analytics

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months)

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months)
    v. One good data scientist (1000x speed increase in 1 day)

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months)
    v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months) v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months) v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:
  - Key to both methods: *classifying* (deaths, social media posts)

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating %'s*

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating %'s*
- Modern Data Analytics: New method led to:

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating %'s*
- Modern Data Analytics: New method led to:

1. crimson hexagon ⇝

⇝

Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET Text Size

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named

How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating %'s*
- Modern Data Analytics: New method led to:

1. crimson hexagon ⇝ 

⇝ 

Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named

2. Worldwide cause-of-death estimates for World Health Organization

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1:

Example Substitution 1:

自由

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom”

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom” 

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom”

目田

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom”

目田 “Eye field”

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1:

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 

目田 “Eye field” (nonsensical)

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	“Freedom”	
目田	“Eye field”	(nonsensical)

Example Substitution 2:

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2:

和 谐

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 

目田 “Eye field” (nonsensical)

Example Substitution 2:

和 谐 “Harmonious [Society]” (official slogan)

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 

目田 “Eye field” (nonsensical)

Example Substitution 2:

和 谐 “Harmonious [Society]” (official slogan) 

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 

目田 “Eye field” (nonsensical)

Example Substitution 2:

和谐 “Harmonious [Society]” (official slogan) 

河蟹

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2:

和 谐	"Harmonious [Society]" (official slogan)	
河蟹	"River crab"	

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2:

和 谐	"Harmonious [Society]" (official slogan)	
河蟹	"River crab" (irrelevant)	

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 

目田 “Eye field” (nonsensical)

Example Substitution 2: Homophone (sound like “hexie”)

和 谐 “Harmonious [Society]” (official slogan) 

河蟹 “River crab” (irrelevant)

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	CENSORED
目田	"Eye field"	(nonsensical)

Example Substitution 2: Homophone (sound like "hexie")

和 谐	"Harmonious [Society]" (official slogan)	CENSORED
河蟹	"River crab" (irrelevant)	

They can't follow the conversation;

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2: Homophone (sound like "hexie")

和 谐	"Harmonious [Society]" (official slogan)	
河蟹	"River crab" (irrelevant)	

They can't follow the conversation; Thresher can.

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 

目田 “Eye field” (nonsensical)

Example Substitution 2: Homophone (sound like “hexie”)

和谐 “Harmonious [Society]” (official slogan) 

河蟹 “River crab” (irrelevant)

They can’t follow the conversation; Thresher can.

The same task:

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 "Freedom" 

目田 "Eye field" (nonsensical)

Example Substitution 2: Homophone (sound like "hexie")

和 谐 "Harmonious [Society]" (official slogan) 

河蟹 "River crab" (irrelevant)

They can't follow the conversation; Thresher can.

The same task: (1) Long tail search,

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2: Homophone (sound like "hexie")

和 谐	"Harmonious [Society]" (official slogan)	
河蟹	"River crab" (irrelevant)	

They can't follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst's job,

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2: Homophone (sound like "hexie")

和 谐	"Harmonious [Society]" (official slogan)	
河蟹	"River crab" (irrelevant)	

They can't follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings $\rightsquigarrow$ #BostonStrong),

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2: Homophone (sound like "hexie")

和 谐	"Harmonious [Society]"	(official slogan)
河蟹	"River crab"	(irrelevant)

They can't follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings ⇝ #BostonStrong), (4) Child pornographers,

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	“Freedom”	CENSORED
目田	“Eye field”	(nonsensical)

Example Substitution 2: Homophone (sound like “hexie”)

和 谐	“Harmonious [Society]” (official slogan)	CENSORED
河蟹	“River crab” (irrelevant)	

They can’t follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst’s job, (3) language drift (#BostonBombings ⇝ #BostonStrong), (4) Child pornographers, (5) Look-alike modeling,

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由	"Freedom"	
目田	"Eye field"	(nonsensical)

Example Substitution 2: Homophone (sound like "hexie")

和 谐	"Harmonious [Society]" (official slogan)	
河蟹	"River crab" (irrelevant)	

They can't follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings $\rightsquigarrow$ #BostonStrong), (4) Child pornographers, (5) Look-alike modeling, (6) Starting point for other automated text methods,

Thresher: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom”

目田 “Eye field” (nonsensical)

Example Substitution 2: Homophone (sound like “hexie”)

和 谐 “Harmonious [Society]” (official slogan)

河蟹 “River crab” (irrelevant)

They can’t follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst’s job, (3) language drift (#BostonBombings $\rightsquigarrow$ #BostonStrong), (4) Child pornographers, (5) Look-alike modeling, (6) Starting point for other automated text methods, (7) Infinitely improvable classification, eDiscovery, etc., etc.

Computer-Assisted Reading (Consilience)

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
    - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)
- Our alternative: Computer-assisted Categorization

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)
- Our alternative: Computer-assisted Categorization
  - You decide what's important, but *with help*

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)
- Our alternative: Computer-assisted Categorization
  - You decide what's important, but *with help*
  - Invert effort: you innovate; the computer categorizes

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)
- Our alternative: Computer-assisted Categorization
  - You decide what's important, but *with help*
  - Invert effort: you innovate; the computer categorizes
  - Insights: easier, faster, better

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)
- Our alternative: Computer-assisted Categorization
  - You decide what's important, but *with help*
  - Invert effort: you innovate; the computer categorizes
  - Insights: easier, faster, better
  - Technology: visualize the space of all possible clusterings

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information
- Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying *not* to innovate!
  - Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)
- Our alternative: Computer-assisted Categorization
  - You decide what's important, but *with help*
  - Invert effort: you innovate; the computer categorizes
  - Insights: easier, faster, better
  - Technology: visualize the space of all possible clusterings
  - (Lots of technology, but it's behind the scenes)

Example Insight from Computer-Assisted Reading

Example Insight from Computer-Assisted Reading

What Members of Congress Do

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: *partisan taunting*

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: *partisan taunting*
    - Joe Wilson during Obama's State of the Union: "You lie!"

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: *partisan taunting*
    - Joe Wilson during Obama's State of the Union: "You lie!"
    - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: *partisan taunting*
    - Joe Wilson during Obama's State of the Union: "You lie!"
    - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
    - Basically anything said by a 2016 presidential candidate!

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: *partisan taunting*
  - Joe Wilson during Obama's State of the Union: "You lie!"
  - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
  - Basically anything said by a 2016 presidential candidate!
- How common is it?

Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: *partisan taunting*
  - Joe Wilson during Obama's State of the Union: "You lie!"
  - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
  - Basically anything said by a 2016 presidential candidate!
- How common is it? 27% of all Senatorial press releases!

Modern Analytics to Improve Student Learning

Modern Analytics to Improve Student Learning

- The problem:

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments?

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book?

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation:

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation:

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? <50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading
  - Instructors save time, stay engaged: automated student confusion reports

Modern Analytics to Improve Student Learning

- The problem:
  - How many students do reading assignments? 20-30%
  - How many students buy the book? $<$50%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading
  - Instructors save time, stay engaged: automated student confusion reports
  - Want to try it here? see Perusall.com

Reverse Engineering Censorship in China

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download <u>all posts</u> before the Chinese censor them
- $\approx$ 13% censored overall

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx 13\%$ censored overall
- Everyone knows the Goal:

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx 13\%$ censored overall
- Everyone knows the Goal:
  Stop criticism and protest about the state,
  its leaders, and their policies

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. Stop criticism of the state

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. Stop criticism of the state
  2. Stop collective action

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx 13\%$ censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx 13\%$ censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:
    - Officials in trouble, likely to be replaced

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:
    - Officials in trouble, likely to be replaced
    - Dissident arrests;

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:
    - Officials in trouble, likely to be replaced
    - Dissident arrests; new peace treaties;

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download <u>all posts</u> before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:
    - Officials in trouble, likely to be replaced
    - Dissident arrests; new peace treaties; emerging scandals

Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- $\approx$ 13% censored overall
- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?
  1. ~~Stop criticism of the state~~ *Wrong*
  2. Stop collective action *Right*
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:
    - Officials in trouble, likely to be replaced
    - Dissident arrests; new peace treaties; emerging scandals
    - Disagreements between central and local leaders

Reverse Engineering China's 50c Party

Reverse Engineering China's 50c Party

- Prevailing view of scholars, activists, journalists, social media participants:

Reverse Engineering China's 50c Party

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies

Reverse Engineering China's 50c Party

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies

Evidence?

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies

Evidence? A few anecdotes;

Reverse Engineering China's 50c Party

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies

Evidence? A few anecdotes; "no ground truth";

Reverse Engineering China's 50c Party

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies

Evidence? A few anecdotes; "no ground truth"; "no successful attempts to quantify" 50c party activity;

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies

Evidence? A few anecdotes; "no ground truth"; "no successful attempts to quantify" 50c party activity; even several analyses with made up dependent variables!

Reverse Engineering China's 50c Party

- ~~Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies~~ *Wrong*

Reverse Engineering China's 50c Party

- ~~Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies~~ *Wrong*
- Does not argue; does not engage on controversial issues

Reverse Engineering China's 50c Party

- ~~Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies~~ *Wrong*
- Does not argue; does not engage on controversial issues
- Distracts; redirects public attention from criticism and central issues to cheerleading and positive discussions of valence issues

The End of The Quantitative-Qualitative Divide

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- Moral of the story:

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- Moral of the story:
    - Fully human is inadequate

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- Moral of the story:
  - Fully human is inadequate
  - Fully automated fails

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- Moral of the story:
  - Fully human is inadequate
  - Fully automated fails
  - We need computer assisted, human controlled technology

The End of The Quantitative-Qualitative Divide

- The Quant-Qual divide exists in *every* field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- Moral of the story:
    - Fully human is inadequate
    - Fully automated fails
    - We need computer assisted, human controlled technology
    - (Technically correct, & politically much easier)

How To Take Advantage of Big Analytics

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need
    (most statistics are designed solely to overcome data problems)

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need
    (most statistics are designed solely to overcome data problems)
  - Building analytics during design:

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need
    (most statistics are designed solely to overcome data problems)
  - Building analytics during design:
    - avoids problems before they occur

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need (most statistics are designed solely to overcome data problems)
  - Building analytics during design:
    - avoids problems before they occur
    - saves a fortune,

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need
    (most statistics are designed solely to overcome data problems)
  - Building analytics during design:
    - avoids problems before they occur
    - saves a fortune,
    - opens many more possibilities

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
    - Off-the-shelf analytics $\rightsquigarrow$ big advances
    - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
    - The goal is "inference":
      using facts you know to learn about facts you don't know
    - The uncertainties in inference: not having the facts you need (most statistics are designed solely to overcome data problems)
    - Building analytics during design:
        - avoids problems before they occur
        - saves a fortune,
        - opens many more possibilities
- Build a new discipline of data science

For more information

GaryKing.org

Institute for Quantitative Social Science
Harvard University