Big Data is Not About the Data!

Gary King[1]

Institute for Quantitative Social Science
Harvard University

Talk at the *MIT Analytics Lab*, 9/29/2015

[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 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 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 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)

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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)

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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- 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 replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable "big data" articles)
- 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. Our Students (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. Our Students (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. Our Students (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. Our Students (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

Examples of what's now possible

Examples of what's now possible

- Opinions of activists:

Examples of what's now possible

- Opinions of activists: A few thousand interviews

Examples of what's now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)

Examples of what's now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise:

Examples of what's now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: "How many times did you exercise last week?

Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (650M/day)
- 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

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: "How many times did you exercise last week? ⇝ 500K people carrying cell phones with accelerometers
- Social contacts:

Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (650M/day)
- 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

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: “How many times did you exercise last week? ⇝ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: “Please tell me your 5 best friends” ⇝ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

Examples of what’s now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: “How many times did you exercise last week? ⇝ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: “Please tell me your 5 best friends” ⇝ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries:

Examples of what's now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: "How many times did you exercise last week? ⇝ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" ⇝ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics

Examples of what's now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: "How many times did you exercise last week? ⇝ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" ⇝ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics ⇝ satellite images of human-generated light at night, road networks, other infrastructure

Examples of what's now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: "How many times did you exercise last week? ⇝ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" ⇝ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics ⇝ satellite images of human-generated light at night, road networks, other infrastructure
- Many, **many**, more. . .

Examples of what's now possible

- Opinions of activists: A few thousand interviews ⇝ billions of political opinions in social media posts (650M/day)
- Exercise: A survey: "How many times did you exercise last week? ⇝ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends" ⇝ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics ⇝ satellite images of human-generated light at night, road networks, other infrastructure
- Many, **many**, more. . .
- In each: without new analytics, the data are useless

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 Read a Billion Blog Posts & Classify Deaths without Physicians

How to Read a Billion Blog Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:

How to Read a Billion Blog Posts & Classify Deaths without Physicians

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

How to Read a Billion Blog Posts & Classify Deaths without Physicians

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

How to Read a Billion Blog Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Different problems, Same Analytics Solution:

How to Read a Billion Blog Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Different problems, Same Analytics Solution:
  - Key to both methods: *classifying* (deaths, social media posts)

How to Read a Billion Blog Posts & Classify Deaths without Physicians

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

How to Read a Billion Blog Posts & Classify Deaths without Physicians

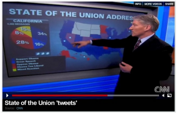
- Examples of Bad Analytics:
    - Physicians' "Verbal Autopsy" analysis
    - Sentiment analysis via word counts
- Different problems, Same Analytics Solution:
    - Key to both methods: *classifying* (deaths, social media posts)
    - Key to both goals: *estimating %'s*
- Modern Data Analytics: New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

How to Read a Billion Blog Posts & Classify Deaths without Physicians

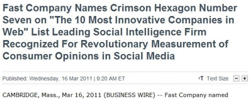
- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Different problems, Same Analytics Solution:
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating %'s*
- Modern Data Analytics: New method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

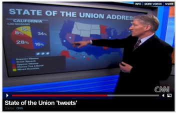
Published: Wednesday, 16 Mar 2011 | 9:20 AM ET

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

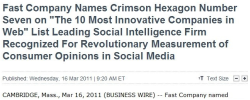
How to Read a Billion Blog Posts & Classify Deaths without Physicians

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

The Solvency of Social Security

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts:

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
  - Few statistical improvements for 75 years

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
    - Few statistical improvements for 75 years
    - Ignores risk factors (smoking, obesity)

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
    - Few statistical improvements for 75 years
    - Ignores risk factors (smoking, obesity)
    - Mostly informal (subject to error & political influence)
    - Forecasts: All systematically biased since 2000

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000
- New customized analytics we developed:

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000
- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000
- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
    - Few statistical improvements for 75 years
    - Ignores risk factors (smoking, obesity)
    - Mostly informal (subject to error & political influence)
    - Forecasts: All systematically biased since 2000
- New customized analytics we developed:
    - Logical consistency (e.g., older people have higher mortality)
    - More accurate forecasts
    - $\rightsquigarrow$ Trust fund needs $\approx$ \$800 billion more than SSA thought

The Solvency of Social Security

- Successful: single largest government program; lifted a whole generation out of poverty; extremely popular
- Solvency: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- SSA data: little change other than updates for 75 years
- SSA analytics:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000
- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts
  - $\rightsquigarrow$ Trust fund needs $\approx$ \$800 billion more than SSA thought
  - Other applications to insurance industry, public health, etc.

Following Conversations that Hide in Plain Sight

Following Conversations that Hide in Plain Sight

Example Substitution 1:

Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由

Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由	“Freedom”

Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由 “Freedom” CENSORED

Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由 “Freedom” CENSORED

目田

Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由	“Freedom”	CENSORED
目田	“Eye field”	

Following Conversations that Hide in Plain Sight

Example Substitution 1:

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

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐	“Harmonious [Society]”	(official slogan)

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐	"Harmonious [Society]" (official slogan)	CENSORED

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐	“Harmonious [Society]”	(official slogan)	CENSORED
河蟹			

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐	“Harmonious [Society]”	(official slogan)	CENSORED
河蟹	“River crab”		

Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

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

Following Conversations that Hide 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)	

Following Conversations that Hide 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; Our methods can!

Following Conversations that Hide 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; Our methods can!

The same task:

Following Conversations that Hide 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; Our methods can!

The same task: (1) Government and industry analyst’s job,

Following Conversations that Hide 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; Our methods can!

The same task: (1) Government and industry analyst’s job, (2) language drift (#BostonBombings ⇝ #BostonStrong),

Following Conversations that Hide 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; Our methods can!

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

Following Conversations that Hide 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; Our methods can!

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

Following Conversations that Hide 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; Our methods can!

The same task: (1) Government and industry analyst's job, (2) language drift (#BostonBombings ⇝ #BostonStrong), (3) Child pornographers, (4) Look-alike modeling,(5) Starting point for sophisticated automated text analysis

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 firms: impose fixed categorizations to tally customer complaints, sort reports, retrieve information

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

- To understand many documents, humans create categories to represent conceptualization, insight, etc.
- Most firms: impose fixed categorizations to tally customer 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 firms: impose fixed categorizations to tally customer 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 firms: impose fixed categorizations to tally customer 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 firms: impose fixed categorizations to tally customer 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 firms: impose fixed categorizations to tally customer 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 firms: impose fixed categorizations to tally customer 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 firms: impose fixed categorizations to tally customer 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
  - (Lots of technology, but it's behind the scenes)

Example Insights from Computer-Assisted Reading

Example Insights from Computer-Assisted Reading

What Members of Congress Do

Example Insights from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators' press releases

Example Insights 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 Insights 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 Insights 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 Insights 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 Insights 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' "
- How common is it?

Example Insights 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' "
- How common is it? 27% of all Senatorial press releases!

How did we come to study Chinese Censorship?

How did we come to study Chinese Censorship?

- We were working on methods of automated text analysis

How did we come to study Chinese Censorship?

- We were working on methods of automated text analysis
- How to stress test the methods?

How did we come to study Chinese Censorship?

- We were working on methods of automated text analysis
- How to stress test the methods? Do they work in Chinese?

How did we come to study Chinese Censorship?

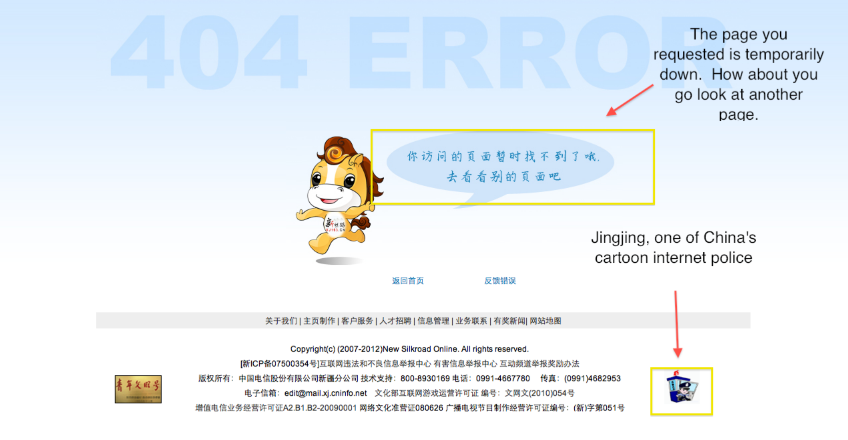
- We were working on methods of automated text analysis
- How to stress test the methods? Do they work in Chinese?

How did we come to study Chinese Censorship?

- We were working on methods of automated text analysis
- How to stress test the methods? Do they work in Chinese?

- We had the content of millions of censored Chinese posts!

Censorship is not Ambiguous: Example Error Page



404 ERROR

你访问的页面暂时找不到了哦,
去看看别的页面吧

The page you requested is temporarily down. How about you go look at another page.

Jingjing, one of China's cartoon internet police

返回首页 反馈错误

关于我们 | 主页制作 | 客户服务 | 人才招聘 | 信息管理 | 业务联系 | 有奖新闻| 网站地图

Copyright(c) (2007-2012)New Silkroad Online. All rights reserved.
[新ICP备07500354号]互联网违法和不良信息举报中心 有害信息举报中心 互动频道举报奖励办法
版权所有：中国电信股份有限公司新疆分公司 技术支持：800-8930169 电话：0991-4667780 传真：(0991)4682953
电子信箱：edit@mail.xj.cninfo.net 文化部互联网游戏运营许可证 编号：文网文(2010)054号
增值电信业务经营许可证A2.B1.B2-20090001 网络文化准营证080626 广播电视节目制作经营许可证编号：(新)字第051号

Chinese Censorship

Chinese Censorship

- The largest selective suppression of human expression in history

Chinese Censorship

- The largest selective suppression of human expression in history
    - implemented *manually* (within a few hours of posting),

Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by $\approx 200,000$ workers,

Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by $\approx 200,000$ workers,
  - located in government and inside social media firms

Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by $\approx 200,000$ workers,
  - located in government and inside social media firms
- A huge censorship organization:

Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by $\approx 200,000$ workers,
  - located in government and inside social media firms
- A huge censorship organization:
  - (obviously) designed to suppress information

Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by $\approx 200,000$ workers,
  - located in government and inside social media firms
- A huge censorship organization:
  - (obviously) designed to suppress information
  - (paradoxically) very revealing about the goals, intentions, and actions of the Chinese leadership

The Goals of Censorship make Social Media Actionable

The Goals of Censorship make Social Media Actionable

- Everyone knows the Goal:

The Goals of Censorship make Social Media Actionable

- Everyone knows the Goal:
  Stop criticism and protest about the state,
  its leaders, and their policies

The Goals of Censorship make Social Media Actionable

- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*

The Goals of Censorship make Social Media Actionable

- ~~Everyone knows the Goal:~~
  ~~Stop criticism and protest about the state,~~
  ~~its leaders, and their policies~~ *Wrong*
- What Could be the Goal?

The Goals of Censorship make Social Media Actionable

- ~~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

The Goals of Censorship make Social Media Actionable

- ~~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

The Goals of Censorship make Social Media Actionable

- ~~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

The Goals of Censorship make Social Media Actionable

- ~~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*

The Goals of Censorship make Social Media Actionable

- ~~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!

The Goals of Censorship make Social Media Actionable

- ~~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:

The Goals of Censorship make Social Media Actionable

- ~~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

The Goals of Censorship make Social Media Actionable

- ~~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

The Goals of Censorship make Social Media Actionable

- ~~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:

The Goals of Censorship make Social Media Actionable

- ~~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

The Goals of Censorship make Social Media Actionable

- ~~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
    - Policies that generate dissent

The Goals of Censorship make Social Media Actionable

- ~~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
    - Policies that generate dissent
    - Dissidents to be arrested; peace treaties to sign; emerging scandals

The Goals of Censorship make Social Media Actionable

- ~~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
    - Policies that generate dissent
    - Dissidents to be arrested; peace treaties to sign; emerging scandals
    - Disagreements between central and local leaders

Classification of Events Generating Bursts of Social Media

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. Collective Action Potential

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. Collective Action Potential
2. Pornography

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. ~~Collective Action Potential~~ CENSORED
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. ~~Collective Action Potential~~ CENSORED
2. ~~Pornography~~ CENSORED
3. Criticism of censors
4. (Other) News
5. Government Policies

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. ~~Collective Action Potential~~ CENSORED
2. ~~Pornography~~ CENSORED
3. ~~Criticism of censors~~ CENSORED
4. (Other) News
5. Government Policies

Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

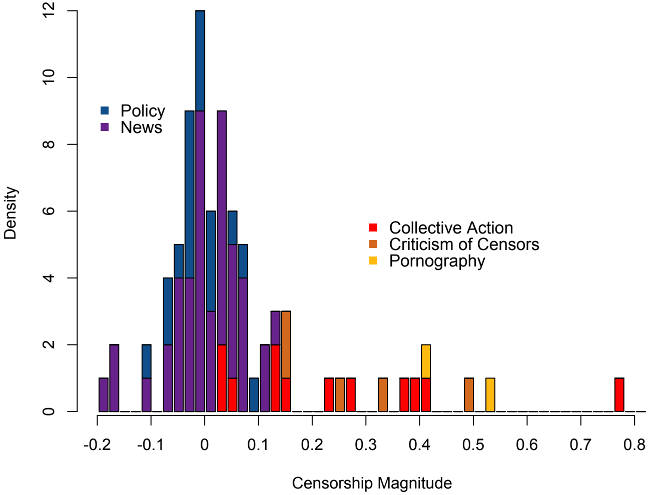
1. ~~Collective Action Potential~~ CENSORED
2. ~~Pornography~~ CENSORED
3. ~~Criticism of censors~~ CENSORED
4. (Other) News 👌
5. Government Policies

Classification of Events Generating Bursts of Social Media

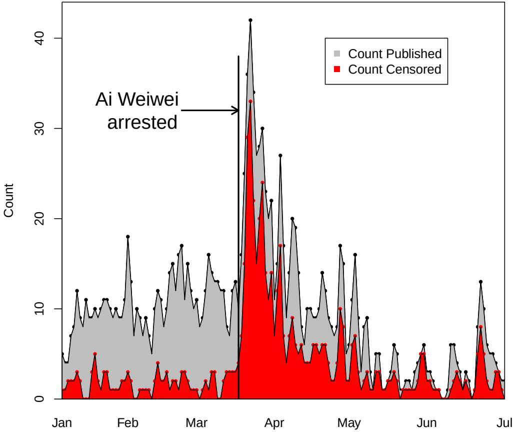
Each including $+$, $-$, or neutral comments about the state

1. ~~Collective Action Potential~~ CENSORED
2. ~~Pornography~~ CENSORED
3. ~~Criticism of censors~~ CENSORED
4. (Other) News 👌
5. Government Policies 👌

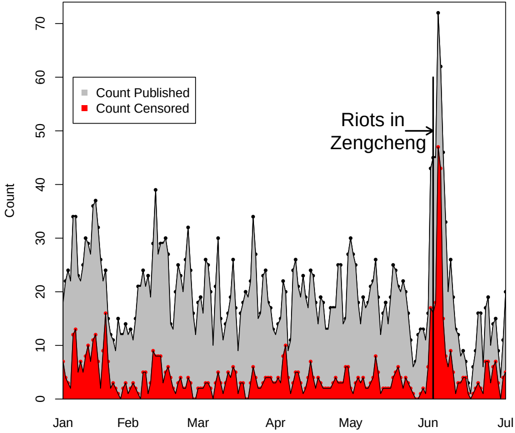
What Types of Events Are Censored?



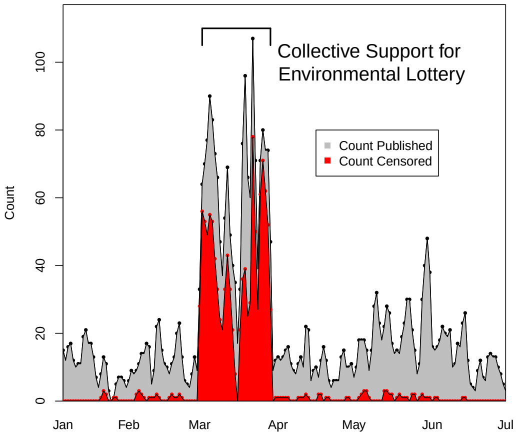
Censoring Collective Action: Ai Weiwei's Arrest



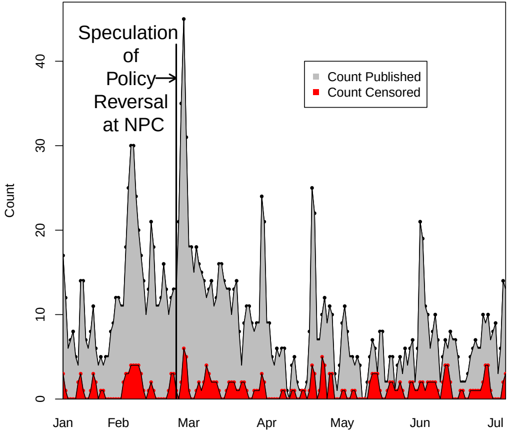
Censoring Collective Action: Riots in Zencheng



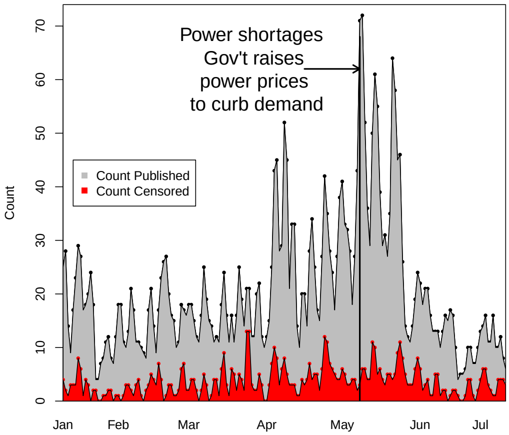
Censoring Collective Action: Environmental Lottery Rally



Low Censorship on Policy: One Child



Low Censorship on News: Power Prices



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 ⇝ big advances
  - Innovative analytics ⇝ immensely better than off-the-shelf

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics ⇝ big advances
  - Innovative analytics ⇝ immensely better than off-the-shelf
  - (Much harder to hire for innovative analytics;
    so consider a mix of in house hires and outside experts)

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
  - (Much harder to hire for innovative analytics;
    so consider a mix of in house hires and outside experts)
- Save it for 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
  - (Much harder to hire for innovative analytics;
    so consider a mix of in house hires and outside experts)
- Save it for 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 ⇝ big advances
  - Innovative analytics ⇝ immensely better than off-the-shelf
  - (Much harder to hire for innovative analytics;
    so consider a mix of in house hires and outside experts)
- Save it for 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
  - (Much harder to hire for innovative analytics;
    so consider a mix of in house hires and outside experts)
- Save it for 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 ⇝ big advances
  - Innovative analytics ⇝ immensely better than off-the-shelf
  - (Much harder to hire for innovative analytics;
    so consider a mix of in house hires and outside experts)
- Save it for 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 ⇝ big advances
  - Innovative analytics ⇝ immensely better than off-the-shelf
  - (Much harder to hire for innovative analytics;
    so consider a mix of in house hires and outside experts)
- Save it for 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 ⇝ big advances
    - Innovative analytics ⇝ immensely better than off-the-shelf
    - (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)
- Save it for 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

For more information

GaryKing.org