Big Data is Not About the Data!

Gary King[1]

Institute for Quantitative Social Science
Harvard University

University of Florida, 3/17/2016

[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 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)
- 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. 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

Exciting Data, But Useless without Novel Analytics

Exciting Data, But Useless without Novel Analytics

- Opinions of activists:

Exciting Data, But Useless without Novel Analytics

- Opinions of activists: A few thousand interviews

Exciting Data, But Useless without Novel Analytics

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

Exciting Data, But Useless without Novel Analytics

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

Exciting Data, But Useless without Novel Analytics

- 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?

Exciting Data, But Useless without Novel Analytics

- 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

Exciting Data, But Useless without Novel Analytics

- 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:

Exciting Data, But Useless without Novel Analytics

- 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"

Exciting Data, But Useless without Novel Analytics

- 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" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

Exciting Data, But Useless without Novel Analytics

- 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" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries:

Exciting Data, But Useless without Novel Analytics

- 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" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics

Exciting Data, But Useless without Novel Analytics

- 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" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, road networks, other infrastructure

Exciting Data, But Useless without Novel Analytics

- 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" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, road networks, other infrastructure
- Many, **many**, more. . .

Exciting Data, But Useless without Novel Analytics

- 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" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- Economic development in developing countries: Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, road networks, other infrastructure
- Many, **many**, more. . .
- In each: without new analytics, the data are useless

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*

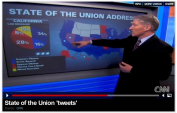
How to Read a Trillion Social Media Posts & Classify Deaths without Physicians

- 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 ⇝ 

⇝

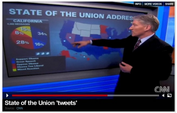
Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

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 Text Size

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

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

Bias in Social Security Administration Forecasts

Bias in Social Security Administration Forecasts

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

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures;

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
    - Methods:

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed;

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative;

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results:

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000;

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after
  - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after
  - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
  - In the process, they also insulated themselves from the facts:

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after
  - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
  - In the process, they also insulated themselves from the facts: Especially since 2000, Americans started living unexpectedly longer lives

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after
  - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
  - In the process, they also insulated themselves from the facts: Especially since 2000, Americans started living unexpectedly longer lives (due to statins, early cancer detection, etc.)

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after
  - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
  - In the process, they also insulated themselves from the facts: Especially since 2000, Americans started living unexpectedly longer lives (due to statins, early cancer detection, etc.)
- New customized analytics we developed:

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after
  - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
  - In the process, they also insulated themselves from the facts: Especially since 2000, Americans started living unexpectedly longer lives (due to statins, early cancer detection, etc.)
- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
    - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
    - Results: unbiased until 2000; systematically biased after
    - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
    - In the process, they also insulated themselves from the facts: Especially since 2000, Americans started living unexpectedly longer lives (due to statins, early cancer detection, etc.)
- New customized analytics we developed:
    - Logical consistency (e.g., older people have higher mortality)
    - Far more accurate forecasts

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
  - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
  - Results: unbiased until 2000; systematically biased after
  - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
  - In the process, they also insulated themselves from the facts: Especially since 2000, Americans started living unexpectedly longer lives (due to statins, early cancer detection, etc.)
- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)
  - Far more accurate forecasts
  - $\rightsquigarrow$ Trust fund needs $>$ \$800 billion more than SSA thought

Bias in Social Security Administration Forecasts

- Social Security: single largest government program; lifted a whole generation out of poverty; extremely popular
- Forecasts: used for programs comprising $> 50\%$ of the US expenditures; e.g., if retirees draw benefits longer than expected, the Trust Fund runs out
- First evaluation of SSA forecasts in 85 years:
    - Methods: little changed; mostly qualitative; a time when we've learned more about forecasting than at any time in history
    - Results: unbiased until 2000; systematically biased after
    - Actuaries hunkered down, insulated themselves, refused to budge when Democrats & Republicans pushed hard for changes
    - In the process, they also insulated themselves from the facts: Especially since 2000, Americans started living unexpectedly longer lives (due to statins, early cancer detection, etc.)
- New customized analytics we developed:
    - Logical consistency (e.g., older people have higher mortality)
    - Far more accurate forecasts
    - $\rightsquigarrow$ Trust fund needs $>$ \$800 billion more than SSA thought
    - Many other applications to different types of forecasts

Humans are Horrible at Thinking of Keywords

Humans are <u>Horrible</u> at Thinking of Keywords

- An experiment:

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples:

Humans are <u>Horrible</u> at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional,

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage,

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama,

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...

Humans are <u>Horrible</u> at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA. . .
- Median keywords recalled:

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...
- Median keywords recalled: 8

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...
- Median keywords recalled: 8
- Unique keywords recalled by 43 undergrads:

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...
- Median keywords recalled: 8
- Unique keywords recalled by 43 undergrads: 149

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA. . .
- Median keywords recalled: 8
- Unique keywords recalled by 43 undergrads: 149
- Keywords 42 of 43 failed to recall:

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...
- Median keywords recalled: 8
- Unique keywords recalled by 43 undergrads: 149
- Keywords 42 of 43 failed to recall: 98 (66%)

Humans are Horrible at Thinking of Keywords

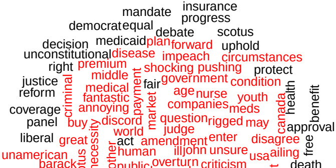
- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...
- Median keywords recalled: 8
- Unique keywords recalled by 43 undergrads: 149
- Keywords 42 of 43 failed to recall: 98 (66%)
- $\rightsquigarrow$ Humans recognize keywords well, recall them poorly

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...
- Median keywords recalled: 8
- Unique keywords recalled by 43 undergrads: 149
- Keywords 42 of 43 failed to recall: 98 (66%)
- $\rightsquigarrow$ Humans recognize keywords well, recall them poorly
- Thresher: New technology to discover the right keywords

Humans are Horrible at Thinking of Keywords

- An experiment: "We have 10,000 twitter posts, each containing the word 'healthcare', from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obama care."
- Examples: unconstitutional, coverage, obama, ACA...
- Median keywords recalled: 8
- Unique keywords recalled by 43 undergrads: 149
- Keywords 42 of 43 failed to recall: 98 (66%)
- $\rightsquigarrow$ Humans recognize keywords well, recall them poorly
- Thresher: New technology to discover the right keywords



Finding Those Hiding in Plain Sight

Finding Those Hiding in Plain Sight

Example Substitution 1:

Finding Those Hiding in Plain Sight

Example Substitution 1:

自由

Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom”

Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom” 

Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom” 

目田

Finding Those Hiding in Plain Sight

Example Substitution 1:

自由	“Freedom”	
目田	“Eye field”	

Finding Those Hiding in Plain Sight

Example Substitution 1:

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

Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

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

Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

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

Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和 谐

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)

Finding Those Hiding 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

Finding Those Hiding in Plain Sight

Example Substitution 1: <u>Homograph</u>

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

Example Substitution 2:

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

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"	

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)

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)	

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)

Example Substitution 3:

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)

Example Substitution 3:

薄瓜瓜

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)

Example Substitution 3:

薄瓜瓜 Bo Guagua (son of Bo Xilai)

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)	

Example Substitution 3:

薄瓜瓜	Bo Guagua (son of Bo Xilai)	CENSORED

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)	

Example Substitution 3:

薄瓜瓜	Bo Guagua (son of Bo Xilai)	CENSORED
b瓜瓜,		

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)	

Example Substitution 3:

薄瓜瓜	Bo Guagua (son of Bo Xilai)	CENSORED

b瓜瓜, b瓜,

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)

Example Substitution 3:

薄瓜瓜 Bo Guagua (son of Bo Xilai) 

b瓜瓜, b瓜, 瓜瓜,

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)	

Example Substitution 3:

薄瓜瓜	Bo Guagua (son of Bo Xilai)	CENSORED

b瓜瓜, b瓜, 瓜瓜, bmelon,

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)

Example Substitution 3:

薄瓜瓜 Bo Guagua (son of Bo Xilai) CENSORED

b瓜瓜, b瓜, 瓜瓜, bmelon, abb

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)	

Example Substitution 3: Slang

薄瓜瓜	Bo Guagua (son of Bo Xilai)	CENSORED

b瓜瓜, b瓜, 瓜瓜, bmelon, abb

Finding Those Hiding in Plain Site

Finding Those Hiding in Plain Site

- Substitutions:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot.

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search.

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;
Humans excel at recognition

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;
Humans excel at recognition
  - Many applications:

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;
Humans excel at recognition
  - Many applications:
    - Language drift: #BostonBombings $\rightsquigarrow$ #BostonStrong

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;
Humans excel at recognition
  - Many applications:
    - Language drift: #BostonBombings $\rightsquigarrow$ #BostonStrong
    - Political advantage: pro-life/choice $\rightsquigarrow$ reproductive rights

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;
Humans excel at recognition
  - Many applications:
    - Language drift: #BostonBombings $\rightsquigarrow$ #BostonStrong
    - Political advantage: pro-life/choice $\rightsquigarrow$ reproductive rights
    - Criminals hiding in plain sight

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;
    Humans excel at recognition
  - Many applications:
    - Language drift: #BostonBombings $\rightsquigarrow$ #BostonStrong
    - Political advantage: pro-life/choice $\rightsquigarrow$ reproductive rights
    - Criminals hiding in plain sight
    - Look-alike modeling

Finding Those Hiding in Plain Site

- Substitutions:
  - An effective strategy for evading censors
  - Invisible to Chinese government filters
  - Not in any thesaurus
  - No other structured data can help
  - The Theory: Humans are more creative than computers
- The Government or Industry Analyst:
  - The Method: Read a lot. Search. Repeat.
  - The Theory: Our humans are more creative than yours
- Thresher:
  - The Method: Computer-assisted human led technology
  - The Theory: Computers excel at discovery;
    Humans excel at recognition
  - Many applications:
    - Language drift: #BostonBombings $\rightsquigarrow$ #BostonStrong
    - Political advantage: pro-life/choice $\rightsquigarrow$ reproductive rights
    - Criminals hiding in plain sight
    - Look-alike modeling
    - Starting point for most 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
  - 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 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
  - 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!

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: We get posts before the Chinese censor them

Reverse Engineering Censorship in China

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

Reverse Engineering Censorship in China

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

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