Big Data is Not About the Data!

Gary King

Institute for Quantitative Social Science
Harvard University

Abt Associates 9/28/2017
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to:
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- The march of quantification: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software

Impact:
- Changed most Fortune 500 firms
- Established new industries
- Altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature, etc., etc., etc.
The *Data In Big Data* (about people)

The Last 50 Years:

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

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

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

Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature, etc., etc., etc.
The *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 Next 50 Years: Fast increases in new data sources, due to...

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

**Impact:** changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature, etc., etc., etc.
The *Data In Big Data* (about people)

The Last 50 Years:

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

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

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

Impact:

- Changed most Fortune 500 firms
- Established new industries
- Altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature, etc., etc., etc.
The *Data In Big Data* (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to:
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- The march of quantification: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature, etc., etc., etc.
The *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

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 Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce

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 *Data* In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
The Data In Big Data (about people)

The Last 50 Years:

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

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

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

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact:
The *Data In Big Data* (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- The march of quantification: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries
The *Data In Big Data* (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks
The *Data In Big Data (about people)*

**The Last 50 Years:**
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

**The Next 50 Years:** Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification:* through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns
The Data In Big Data (about people)

The Last 50 Years:

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

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

- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health
The *Data* In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- The march of quantification: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics
The Data In Big Data (about people)

The Last 50 Years:
• Survey research
• Aggregate government statistics
• One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
• Much more of the above — improved, expanded, and applied
• Shrinking computers & the growing Internet: data everywhere
• *The march of quantification*: through academia, professions, government, & commerce
• The replication movement: data sharing (e.g., Dataverse)
• Governments encouraging data collection & experimentation
• Advances in statistical methods, informatics, & software
• Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports
The *Data In* Big Data (about people)

**The Last 50 Years:**
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

**The Next 50 Years:** Fast increases in new data sources, due to.
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification:* through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature,
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, literature, etc., etc., etc
The *Value* in Big Data: the Analytics
The *Value* in Big Data: the Analytics

- Data:

  - 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)
  - 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
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

Moore's Law (doubling speed/power every 18 months) vs. One good data scientist (1000x speed increase in 1 day)

$2M$ computer vs. 2 hours of algorithm design

Low cost; little infrastructure; mostly human capital needed

Innovative analytics: enormously better than off-the-shelf
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

- Opinions of activists:
  - A few thousand interviews
  - Billions of political opinions in social media posts (750M/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
Exciting Data, But Useless without Novel Analytics

- Opinions of activists:
  - A few thousand interviews
  - Billions of political opinions in social media posts (750M/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...
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 $\sim$ billions of political opinions in social media posts (750M/day)
Exciting Data, But Useless without Novel Analytics

- **Opinions of activists:** A few thousand interviews \(\leadsto\) billions of political opinions in social media posts (750M/day)
- **Exercise:**

...
Exciting Data, But Useless without Novel Analytics

- **Opinions of activists**: A few thousand interviews \(\leadsto\) billions of political opinions in social media posts (750M/day)

- **Exercise**: A survey: “How many times did you exercise last week?”

- **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.
Exciting Data, But Useless without Novel Analytics

- **Opinions of activists**: A few thousand interviews $\leadsto$ billions of political opinions in social media posts (750M/day)
- **Exercise**: A survey: “How many times did you exercise last week? $\leadsto$ 500K people carrying cell phones with accelerometers

• Many, many more. . .

In each: without new analytics, the data are useless
Exciting Data, But Useless without Novel Analytics

- **Opinions of activists**: A few thousand interviews \(\leadsto\) billions of political opinions in social media posts (750M/day)
- **Exercise**: A survey: “How many times did you exercise last week? \(\leadsto\) 500K people carrying cell phones with accelerometers
- **Social contacts**: 

... Many, many, more...

In each: without new analytics, the data are useless.
Exciting Data, But Useless without Novel Analytics

- **Opinions of activists:** A few thousand interviews $\leadsto$ billions of political opinions in social media posts (750M/day)
- **Exercise:** A survey: “How many times did you exercise last week? $\leadsto$ 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 $\implies$ billions of political opinions in social media posts (750M/day)

- **Exercise:** A survey: “How many times did you exercise last week? $\implies$ 500K people carrying cell phones with accelerometers

- **Social contacts:** A survey: “Please tell me your 5 best friends” $\implies$ 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 ⇾ billions of political opinions in social media posts (750M/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:**

---

4/15
Exciting Data, But Useless without Novel Analytics

• **Opinions of activists:** A few thousand interviews \(\sim\) billions of political opinions in social media posts (750M/day)

• **Exercise:** A survey: “How many times did you exercise last week? \(\sim\) 500K people carrying cell phones with accelerometers

• **Social contacts:** A survey: “Please tell me your 5 best friends” \(\sim\) 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 $\rightarrow$ billions of political opinions in social media posts (750M/day)
- **Exercise:** A survey: “How many times did you exercise last week? $\rightarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\rightarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:** Dubious or nonexistent governmental statistics $\rightarrow$ 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 $\leadsto$ billions of political opinions in social media posts (750M/day)

• **Exercise:** A survey: “How many times did you exercise last week?” $\leadsto$ 500K people carrying cell phones with accelerometers

• **Social contacts:** A survey: “Please tell me your 5 best friends” $\leadsto$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

• **Economic development in developing countries:** Dubious or nonexistent governmental statistics $\leadsto$ 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 (750M/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

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. Worldwide cause-of-death estimates

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. Worldwide cause-of-death estimates for
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:
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. Worldwide cause-of-death estimates for
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. Worldwide cause-of-death estimates for
  2. [Image of World Health Organization]
Bias in U.S. 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

⇝ Trust fund needs >$800 billion more than SSA thought

Many other applications to different types of forecasts
Bias in U.S. Social Security Administration Forecasts

- **Social Security**: single largest government program; lifted a whole generation out of poverty; extremely popular
Bias in U.S. 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 U.S. 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 U.S. Social Security Administration Forecasts

- **Social Security**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Forecasts**: used for programs comprising $\geq 50\%$ 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 U.S. 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 U.S. Social Security Administration Forecasts

- **Social Security:** single largest government program; lifted a whole generation out of poverty; extremely popular
- **Forecasts:** used for programs comprising $\geq 50\%$ 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 U.S. 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 U.S. 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 U.S. 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**: New customized analytics we developed:
    - Logical consistency (e.g., older people have higher mortality)
    - Far more accurate forecasts
      - Trust fund needs $\geq \$800$ billion more than SSA thought
- Many other applications to different types of forecasts
Bias in U.S. 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 U.S. 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 U.S. 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

New customized analytics we developed:

- Logical consistency (e.g., older people have higher mortality)
- Far more accurate forecasts

$\rightarrow$ Trust fund needs $>800$ billion more than SSA thought

Many other applications to different types of forecasts
Bias in U.S. 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 U.S. 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

New customized analytics we developed:
- Logical consistency (e.g., older people have higher mortality)
- Far more accurate forecasts
  - $\Rightarrow$ Trust fund needs \$800 billion more than SSA thought
- Many other applications to different types of forecasts
Bias in U.S. 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 U.S. Social Security Administration Forecasts

- **Social Security**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Forecasts**: used for programs comprising $\geq 50\%$ 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 U.S. Social Security Administration Forecasts

- **Social Security**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Forecasts**: used for programs comprising $\geq 50\%$ 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 U.S. 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

$\Rightarrow$ Trust fund needs $> $800 billion more than SSA thought

Many other applications to different types of forecasts
Bias in U.S. 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
  - \( \Rightarrow \) Trust fund needs \( > \$800 \text{ billion} \) more than SSA thought
Bias in U.S. 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  
  - Trust fund needs > $800 billion more than SSA thought  
- Many other applications to different types of forecasts
Thresher.io: 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])
- 河蟹 (River crab) (irrelevant)

They can't follow the conversation; Thresher can.

Same Thresher Technology solves a Very General Task:
- Coding documents into categories
Example Substitution 1:
Example Substitution 1:

自由
Example Substitution 1:

自由  “Freedom”
Example Substitution 1:

自由  “Freedom”

Example Substitution 2:

和  谐  “Harmonious [Society]” (official slogan)
河  蟹  “River crab” (irrelevant)
Thresher.io: Finding Those Hiding in Plain Sight

Example Substitution 1:

自由 “Freedom”

They can't follow the conversation; Thresher can.

Same Thresher Technology solves a Very General Task:

Coding documents into categories
Example Substitution 1:

自由 “Freedom”
目田 “Eye field”
Example Substitution 1:

自由
自由
“Freedom”  “Eye field” (nonsensical)
Example Substitution 1: Homograph

自由  “Freedom”
目田  “Eye field” (nonsensical)
Thresher.io: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom”
目田 “Eye field” (nonsensical)
Example Substitution 1: Homograph

自由 自由 “Freedom”
目 田 目 手 “Eye field” (nonsensical)

Example Substitution 2:
Example Substitution 1: Homograph

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

Example Substitution 2:

和谐
Example Substitution 1: Homograph

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

Example Substitution 2:

和谐 “Harmonious [Society]” (official slogan)
Thresher.io: 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)
Example Substitution 1: Homograph

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

Example Substitution 2:

和谐  “Harmonious [Society]” (official slogan)
Example Substitution 1: Homograph

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

Example Substitution 2:

和谐  和谐  “Harmonious [Society]” (official slogan)
河蟹  河蟹  “River crab”
Example Substitution 1: Homograph

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

Example Substitution 2:

和谐 “Harmonious [Society]” (official slogan)
河蟹 “River crab” (irrelevant)
Thresher.io: 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 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;
Example Substitution 1: Homograph

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

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

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

They can’t follow the conversation; Thresher can.
Thresher.io: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

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

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

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

They can’t follow the conversation; Thresher can.

Same Thresher Technology solves a Very General Task:
Thresher.io: Finding Those Hiding in Plain Sight

Example Substitution 1: Homograph

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

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

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

They can’t follow the conversation; Thresher can.

Same Thresher Technology solves a Very General Task:
Coding documents into categories
Computer-Assisted Reading (Consilience)

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

Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information.

Bad Analytics:

• Unassisted Human Categorization: time consuming; huge efforts trying not to innovate!

• Fully Automated "Cluster Analysis": Many widely available, but none work (computers don't know what you want!)

Our alternative: Computer-assisted Categorization

• You decide what's important, but with help

• Invert effort: you innovate; the computer categorizes

• Insights: easier, faster, better

• Technology: visualize the space of all possible clusterings

(Lots of technology, but it's behind the scenes)
Computer-Assisted Reading (Consilience)

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

Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information.
Computer-Assisted Reading (Consilience)

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

• Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information

• Bad Analytics:

  • 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)
Computer-Assisted Reading (Consilience)

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

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

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

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

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

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

• Most organizations: impose fixed categorizations to tally complaints, sort reports, retrieve information

• Bad Analytics:
  • Unassisted Human Categorization: time consuming; huge efforts trying not to innovate!
  • Fully Automated “Cluster Analysis”: Many widely available, but none work (computers don’t know what you want!)

• Our alternative: Computer-assisted Categorization
  • You decide what’s important, but with help
  • Invert effort: you innovate; the computer categorizes
  • Insights: easier, faster, better
Computer-Assisted Reading (Consilience)

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

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

• Categorization: (1) advertising, (2) position taking, (3) credit claiming
• Data: 64,000 Senators' press releases
• 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!
What Members of Congress Do
What Members of Congress Do

- Categorization: (1) advertising, (2) position taking, (3) credit claiming
What Members of Congress Do

- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- Data: 64,000 Senators’ press releases
Example Insight from Computer-Assisted Reading

What Members of Congress Do

- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- Data: 64,000 Senators’ press releases
- New Insight: partisan taunting
Example Insight from Computer-Assisted Reading

What Members of Congress Do

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

- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- Data: 64,000 Senators’ press releases
- New Insight: partisan taunting
  - Joe Wilson during Obama’s State of the Union: “You lie!”
  - “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’”
What Members of Congress Do

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

- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- Data: 64,000 Senators’ press releases
- 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?
What Members of Congress Do

- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- Data: 64,000 Senators’ press releases
- New Insight: \textit{partisan taunting}
  - Joe Wilson during Obama’s State of the Union: “You lie!”
  - “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’ ”
  - Basically anything said by a 2016 presidential candidate!
- How common is it? 27% of all Senatorial press releases!
Modern Analytics to Improve Student Learning

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

- Our solution:
  - Perusall
    - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
    - >90% of students do the reading
  - Solitary reading assignments ⇝ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading
  - Instructors save time, stay engaged: automated student confusion reports

- Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

- The problem:

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

Our solution:

- Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments ⇝ engaging collective activities
    - Intrinsic motivation: collaborative annotation in threads
    - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading
  - Instructors save time, stay engaged: automated student confusion reports
  - Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book?
Modern Analytics to Improve Student Learning

• The problem:
  • How many students buy the book? <50%
Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? <50%
  - How many students do reading assignments?
Modern Analytics to Improve Student Learning

• The problem:
  • How many students buy the book? <50%
  • How many students do reading assignments? 20-30%
Modern Analytics to Improve Student Learning

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

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

- Our solution: Perusall
Modern Analytics to Improve Student Learning

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

- **Our solution:** *Perusall*
  - A new type of collaborative e-reader
Modern Analytics to Improve Student Learning

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

- **Our solution: **Perusall
  - A new type of collaborative e-reader, with novel data analytics
Modern Analytics to Improve Student Learning

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

• Our solution: Perusall
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
Modern Analytics to Improve Student Learning

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

• Our solution: Perusall
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
Modern Analytics to Improve Student Learning

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

• Our solution: Perusall
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
  • Solitary reading assignments → engaging collective activities

Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

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

- Our solution: **Perusall**
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments ↾ engaging collective activities
  - Intrinsic motivation:
Modern Analytics to Improve Student Learning

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

• Our solution: Perusall
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
  • Solitary reading assignments ⇝ engaging collective activities
  • Intrinsic motivation: collaborative annotation in threads

Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

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

• Our solution: Perusall
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
  • Solitary reading assignments ⇝ engaging collective activities
  • Intrinsic motivation: collaborative annotation in threads
  • Extrinsic motivation:

Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

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

• Our solution: Perusall
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
  • Solitary reading assignments ⇝ engaging collective activities
  • Intrinsic motivation: collaborative annotation in threads
  • Extrinsic motivation: automated grading of annotations & engagement

Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

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

• Our solution: Perusall
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
  • Solitary reading assignments \(\leadsto\) engaging collective activities
  • Intrinsic motivation: collaborative annotation in threads
  • Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
Modern Analytics to Improve Student Learning

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

• **Our solution:** *Perusall*
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
  • Solitary reading assignments ⟷ engaging collective activities
  • Intrinsic motivation: collaborative annotation in threads
  • Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  • Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading

Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

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

• Our solution: **Perusall**
  • A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  • >90% of students do the reading
  • Solitary reading assignments \(\rightsquigarrow\) engaging collective activities
  • Intrinsic motivation: collaborative annotation in threads
  • Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  • Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading
  • Instructors save time, stay engaged: automated student confusion reports

Want to try it here? see Perusall.com
Modern Analytics to Improve Student Learning

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

- **Our solution: Perusall**
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments \(\rightarrow\) engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, nonadversarial grading
  - Instructors save time, stay engaged: automated student confusion reports
  - Want to try it here? see Perusall.com
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- What could be the goal?
  - Stop collective action
  - Implications: Social Media is Actionable!
  - Chinese leaders:
    - Measure criticism: to judge local officials
    - Censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:
    - Officials in trouble, likely to be replaced
    - Dissident arrests; new peace treaties; emerging scandals
    - Disagreements between central and local leaders
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
Reverse Engineering Censorship in China

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

1. Stop collective action
   - Implications: Social Media is Actionable!
   - Chinese leaders:
     - measure criticism: to judge local officials
     - censor: to stop events with collective action potential
     - Thus, we can use criticism & censorship to predict:
       - Officials in trouble, likely to be replaced
       - Dissident arrests; new peace treaties; emerging scandals
       - Disagreements between central and local leaders
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns
• Everyone knows the Goal:
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns
• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies Wrong
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- Everyone knows the Goal:
  - Stop criticism and protest about the state, its leaders, and their policies *Wrong*
- What Could be the Goal?
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- **Everyone knows the Goal:**
  - Stop criticism and protest about the state, its leaders, and their policies *Wrong*
- **What Could be the Goal?**
  1. Stop criticism of the state
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns
• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*
• What Could be the Goal?
  1. Stop criticism of the state
  2. Stop collective action
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies **Wrong**
- What Could be the Goal?
  1. Stop criticism of the state **Wrong**
  2. Stop collective action
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies _Wrong_
- What Could be the Goal?
  1. Stop criticism of the state _Wrong_
  2. Stop collective action _Right_
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*
- **What Could be the Goal?**
  1. Stop criticism of the state *Wrong*
  2. Stop collective action *Right*
- **Implications:** Social Media is Actionable!
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns

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

• What Could be the Goal?
  1. Stop criticism of the state  **Wrong**
  2. Stop collective action  **Right**

• Implications: Social Media is Actionable!
  • Chinese leaders:
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns

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

What Could be the Goal?
1. Stop criticism of the state Wrong
2. Stop collective action Right

Implications: Social Media is Actionable!
• Chinese leaders:
  • measure criticism: to judge local officials
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies Wrong
- What Could be the Goal?
  1. Stop criticism of the state Wrong
  2. Stop collective action Right
- Implications: Social Media is Actionable!
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns
• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies Wrong
• What Could be the Goal?
  1. Stop criticism of the state Wrong
  2. Stop collective action Right
• Implications: Social Media is Actionable!
  • Chinese leaders:
    • measure criticism: to judge local officials
    • censor: to stop events with collective action potential
  • Thus, we can use criticism & censorship to predict:
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns
• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies Wrong
• What Could be the Goal?
  1. Stop criticism of the state Wrong
  2. Stop collective action Right
• Implications: Social Media is Actionable!
  • Chinese leaders:
    • measure criticism: to judge local officials
    • censor: to stop events with collective action potential
  • Thus, we can use criticism & censorship to predict:
    • Officials in trouble, likely to be replaced
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns
• Everyone knows the Goal: Stop criticism and protest about the state, its leaders, and their policies Wrong

• What Could be the Goal?
  1. Stop criticism of the state Wrong
  2. Stop collective action Right

• Implications: Social Media is Actionable!
  • Chinese leaders:
    • measure criticism: to judge local officials
    • censor: to stop events with collective action potential
  • Thus, we can use criticism & censorship to predict:
    • Officials in trouble, likely to be replaced
    • Dissident arrests;
Reverse Engineering Censorship in China

- Previous approach: watch a few posts; see what’s removed
- Data: Download all posts before the Chinese censor them
- Novel methods of automated text analysis to discover patterns
- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*
- **What Could be the Goal?**
  1. Stop criticism of the state *Wrong*
  2. Stop collective action *Right*
- **Implications: Social Media is Actionable!**
  - Chinese leaders:
    - measure criticism: to judge local officials
    - censor: to stop events with collective action potential
  - Thus, we can use criticism & censorship to predict:
    - Officials in trouble, likely to be replaced
    - Dissident arrests; new peace treaties;
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns

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

What Could be the Goal?
1. Stop criticism of the state **Wrong**
2. Stop collective action **Right**

Implications: Social Media is Actionable!
• Chinese leaders:
  • measure criticism: to judge local officials
  • censor: to stop events with collective action potential
• Thus, we can use criticism & censorship to predict:
  • Officials in trouble, likely to be replaced
  • Dissident arrests; new peace treaties; emerging scandals
Reverse Engineering Censorship in China

• Previous approach: watch a few posts; see what’s removed
• Data: Download all posts before the Chinese censor them
• Novel methods of automated text analysis to discover patterns

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

What Could be the Goal?
1. Stop criticism of the state Wrong
2. Stop collective action Right

Implications: Social Media is Actionable!
• Chinese leaders:
  • measure criticism: to judge local officials
  • censor: to stop events with collective action potential
• Thus, we can use criticism & censorship to predict:
  • Officials in trouble, likely to be replaced
  • Dissident arrests; new peace treaties; emerging scandals
  • Disagreements between central and local leaders
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

- Prevailing view of scholars, activists, journalists, social media participants:
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

Existing evidence?
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

Existing evidence? A few anecdotes;
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

Existing evidence? A few anecdotes; “no ground truth”;

•Fabricates 450M social media posts a year!
•Does not argue; does not engage on controversial issues
•Distracts; redirects public attention from criticism and central issues to cheerleading and positive discussions of valence issues
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

  Existing evidence? A few anecdotes; “no ground truth”; “no successful attempts to quantify” 50c party activity;
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

Existing evidence? A few anecdotes; “no ground truth”; “no successful attempts to quantify” 50c party activity; even several analyses with made up dependent variables!
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

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

Our evidence:
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

Existing evidence? A few anecdotes; “no ground truth”; “no successful attempts to quantify” 50c party activity; even several analyses with made up dependent variables!
Our evidence: (1) Found a leaked archive of 50c posts too hard to analyze,
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

Existing evidence? A few anecdotes; “no ground truth”; “no successful attempts to quantify” 50c party activity; even several analyses with made up dependent variables!
Our evidence: (1) Found a leaked archive of 50c posts too hard to analyze, (2) developed methods of automated text analysis to decipher,
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

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

Our evidence: (1) Found a leaked archive of 50c posts too hard to analyze, (2) developed methods of automated text analysis to decipher, (3) discovered patterns and extrapolated to all of China,
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

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

Our evidence: (1) Found a leaked archive of 50c posts too hard to analyze, (2) developed methods of automated text analysis to decipher, (3) discovered patterns and extrapolated to all of China, (4) did a poll(!) and predicted 50c members acknowledged their behavior
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies *Wrong*
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

- Prevailing view of scholars, activists, journalists, social media participants: 50c Party argues against those who criticize the government, its leaders, and their policies Wrong
- Fabricates 450M social media posts a year!
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

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

• Fabricates 450M social media posts a year!

• Does not argue; does not engage on controversial issues
Reverse Engineering China’s 50c Party
The Government Surreptitiously Fabricating Social Media Posts

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies Wrong
- Fabricates 450M social media posts a year!
- Does not argue; does not engage on controversial issues
- Distracts; redirects public attention from criticism and central issues to cheerleading and positive discussions of valence issues
The End of The Quantitative-Qualitative Divide

- The 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)
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

Moral of the story:
- Fully human is inadequate
- Fully automated fails
- We need computer assisted, human controlled technology
  - (Technically correct, & politically much easier)
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

- It's cheap and powerful; don't skimp!
- Off-the-shelf analytics ⇝ big advances
- Innovative analytics ⇝ 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
How To Take Advantage of Big Analytics

• Its cheap and powerful; don’t skimp!
How To Take Advantage of Big Analytics

- It's 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 $\leadsto$ big advances
  - Innovative analytics $\leadsto$ 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 $\leadsto$ big advances
  • Innovative analytics $\leadsto$ 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 $\leadsto$ big advances
  • Innovative analytics $\leadsto$ 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 $\leadsto$ big advances
  • Innovative analytics $\leadsto$ 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 ➝ big advances
  - Innovative analytics ➝ 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 $\leadsto$ big advances
  • Innovative analytics $\leadsto$ 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