# Big Data is Not About the Data!

Gary King[1]

Institute for Quantitative Social Science
Harvard University

Talk at the *History of Evidence* class, Harvard Law School, 11/17/2014

[1]GaryKing.org

# The *Value* in Big Data: the Analytics

# The *Value* in Big Data: the Analytics

- Data:

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

## The *Value* in Big Data: the Analytics

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

# Examples of what's now possible

# Examples of what's now possible

- Opinions of activists:

# Examples of what's now possible

- Opinions of activists: A few thousand interviews

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- Exercise:

## Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- Exercise: A survey: "How many times did you exercise last week?

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers

## Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts:

## Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- Social contacts: A survey: "Please tell me your 5 best friends"

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- 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

## Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- 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:

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- 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

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- 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

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- 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...

# Examples of what's now possible

- Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- 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

# The End of The Quantitative-Qualitative Divide

## The End of The Quantitative-Qualitative Divide

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

# The End of The Quantitative-Qualitative Divide

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

# The End of The Quantitative-Qualitative Divide

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

## The End of The Quantitative-Qualitative Divide

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

## The End of The Quantitative-Qualitative Divide

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

## The End of The Quantitative-Qualitative Divide

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

## The End of The Quantitative-Qualitative Divide

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

# The End of The Quantitative-Qualitative Divide

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

## The End of The Quantitative-Qualitative Divide

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

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

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

- Examples of Bad Analytics:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.

crimson hexagon ⇝


⇝

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

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

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

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

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

  1. crimson hexagon $\rightsquigarrow$ 

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

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

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

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

# The Solvency of Social Security

# The Solvency of Social Security

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

## The Solvency of Social Security

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

# The Solvency of Social Security

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

# The Solvency of Social Security

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

# The Solvency of Social Security

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

## The Solvency of Social Security

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

# The Solvency of Social Security

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

# The Solvency of Social Security

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

## The Solvency of Social Security

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

# The Solvency of Social Security

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

# The Solvency of Social Security

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

# The Solvency of Social Security

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

# The Solvency of Social Security

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

# The Solvency of Social Security

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

# Following Conversations that Hide in Plain Sight

## Following Conversations that Hide in Plain Sight

Example Substitution 1:

## Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由

# Following Conversations that Hide in Plain Sight

Example Substitution 1:

| | |
|---|---|
| 自由 | “Freedom” |

# Following Conversations that Hide in Plain Sight

## Example Substitution 1:

自由 “Freedom” 

## Following Conversations that Hide in Plain Sight

### Example Substitution 1:

自由 “Freedom” 


目田

## Following Conversations that Hide in Plain Sight

### Example Substitution 1:

| | |
|---|---|
| 自由 | "Freedom" |
| 目田 | "Eye field" |

## Following Conversations that Hide in Plain Sight

### Example Substitution 1:

自由 "Freedom" 


目田 "Eye field" (nonsensical)

# Following Conversations that Hide in Plain Sight

## Example Substitution 1: Homograph

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

## Following Conversations that Hide in Plain Sight

### Example Substitution 1: Homograph

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

## Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

## Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐

## Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

| | | |
|---|---|---|
| 和谐 | “Harmonious [Society]” | (official slogan) |

# Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 


目田 “Eye field” (nonsensical)

Example Substitution 2:

和谐 “Harmonious [Society]” (official slogan) 

# Following Conversations that Hide in Plain Sight

## Example Substitution 1: Homograph

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

## Example Substitution 2:

| | | |
|---|---|---|
| 和谐 | "Harmonious [Society]" (official slogan) | CENSORED |
| 河蟹 | | |

# Following Conversations that Hide in Plain Sight

## Example Substitution 1: Homograph

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

## Example Substitution 2:

| | | |
|---|---|---|
| 和谐 | "Harmonious [Society]" (official slogan) |  |
| 河蟹 | "River crab" | |

# Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

| | | |
|---|---|---|
| 和谐 | "Harmonious [Society]" (official slogan) |  |
| 河蟹 | "River crab" (irrelevant) | |

# Following Conversations that Hide 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) |

# Following Conversations that Hide 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; Our methods can!

# Following Conversations that Hide 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; Our methods can!

The same task:

# Following Conversations that Hide 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; Our methods can!

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

## Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” 


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; Our methods can!

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

# Following Conversations that Hide 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; Our methods can!

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

## Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由 “Freedom” CENSORED<br>
目田 “Eye field” (nonsensical)

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

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

They can't follow the conversation; Our methods can!

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

# Following Conversations that Hide 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; Our methods can!

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

# Computer-Assisted Reading (Consilience)

## Computer-Assisted Reading (Consilience)

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

# Computer-Assisted Reading (Consilience)

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

# Computer-Assisted Reading (Consilience)

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

## Computer-Assisted Reading (Consilience)

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

## Computer-Assisted Reading (Consilience)

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

# Computer-Assisted Reading (Consilience)

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

# Computer-Assisted Reading (Consilience)

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

# Computer-Assisted Reading (Consilience)

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

## Computer-Assisted Reading (Consilience)

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

## Computer-Assisted Reading (Consilience)

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

# Example Insights from Computer-Assisted Reading

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Data: 64,000 Senators' press releases

# Example Insights from Computer-Assisted Reading

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

## Example Insights from Computer-Assisted Reading

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

# Example Insights from Computer-Assisted Reading

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

# Example Insights from Computer-Assisted Reading

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

# Example Insights from Computer-Assisted Reading

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

## Example Insights from Computer-Assisted Reading

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

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Data: 64,000 Senators' press releases
   - Categorization: (1) advertising, (2) position taking, (3) credit claiming
   - New Insight: *partisan taunting*
     - Joe Wilson during Obama's State of the Union: "You lie!"
     - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
   - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Data: 64,000 Senators' press releases
   - Categorization: (1) advertising, (2) position taking, (3) credit claiming
   - New Insight: *partisan taunting*
     - Joe Wilson during Obama's State of the Union: "You lie!"
     - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
   - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship
   - Previous approach: manual effort to see what is taken down

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Data: 64,000 Senators' press releases
    - Categorization: (1) advertising, (2) position taking, (3) credit claiming
    - New Insight: *partisan taunting*
        - Joe Wilson during Obama's State of the Union: "You lie!"
        - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
    - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship
    - Previous approach: manual effort to see what is taken down
    - Data: We get posts before the Chinese censor them

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Data: 64,000 Senators' press releases
    - Categorization: (1) advertising, (2) position taking, (3) credit claiming
    - New Insight: *partisan taunting*
        - Joe Wilson during Obama's State of the Union: "You lie!"
        - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
    - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship
    - Previous approach: manual effort to see what is taken down
    - Data: We get posts before the Chinese censor them
    - We analyzed 11 million posts, about 13% censored

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Data: 64,000 Senators' press releases
   - Categorization: (1) advertising, (2) position taking, (3) credit claiming
   - New Insight: *partisan taunting*
     - Joe Wilson during Obama's State of the Union: "You lie!"
     - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
   - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship
   - Previous approach: manual effort to see what is taken down
   - Data: We get posts before the Chinese censor them
   - We analyzed 11 million posts, about 13% censored
   - Previous understanding: they censor criticisms of the government

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Data: 64,000 Senators' press releases
   - Categorization: (1) advertising, (2) position taking, (3) credit claiming
   - New Insight: *partisan taunting*
     - Joe Wilson during Obama's State of the Union: "You lie!"
     - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
   - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship
   - Previous approach: manual effort to see what is taken down
   - Data: We get posts before the Chinese censor them
   - We analyzed 11 million posts, about 13% censored
   - Previous understanding: they censor criticisms of the government
   - Results:

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Data: 64,000 Senators' press releases
   - Categorization: (1) advertising, (2) position taking, (3) credit claiming
   - New Insight: *partisan taunting*
     - Joe Wilson during Obama's State of the Union: "You lie!"
     - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
   - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship
   - Previous approach: manual effort to see what is taken down
   - Data: We get posts before the Chinese censor them
   - We analyzed 11 million posts, about 13% censored
   - Previous understanding: they censor criticisms of the government
   - Results:
     - Uncensored: criticism of the government

# Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Data: 64,000 Senators' press releases
    - Categorization: (1) advertising, (2) position taking, (3) credit claiming
    - New Insight: *partisan taunting*
        - Joe Wilson during Obama's State of the Union: "You lie!"
        - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
    - How common is it? 27% of all Senatorial press releases!
2. Reverse Engineering Chinese Censorship
    - Previous approach: manual effort to see what is taken down
    - Data: We get posts before the Chinese censor them
    - We analyzed 11 million posts, about 13% censored
    - Previous understanding: they censor criticisms of the government
    - Results:
        - Uncensored: criticism of the government
        - Censored: attempts at collective action

# For more information

GaryKing.org