

# Big Data is Not About the Data!

Gary King

Institute for Quantitative Social Science  
Harvard University

(Talk at the *Golden Seeds Innovation Summit*, New York City 1/30/2013)

# The *Data* in Big Data

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs

# The *Data* in Big Data

- ① **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- ② **Commerce:** credit cards, sales, real estate transactions, RFIDs
- ③ **Geographic location:** cell phones, Fastlane, garage cameras

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs
- 3 **Geographic location:** cell phones, Fastlane, garage cameras
- 4 **Health information:** digital medical records, hospital admittances, accelerometers & other devices in cell phones

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs
- 3 **Geographic location:** cell phones, Fastlane, garage cameras
- 4 **Health information:** digital medical records, hospital admittances, accelerometers & other devices in cell phones
- 5 **Biological sciences:** genomics, proteomics, metabolomics, brain imaging producing huge numbers of person-level variables

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs
- 3 **Geographic location:** cell phones, Fastlane, garage cameras
- 4 **Health information:** digital medical records, hospital admittances, accelerometers & other devices in cell phones
- 5 **Biological sciences:** genomics, proteomics, metabolomics, brain imaging producing huge numbers of person-level variables
- 6 **Satellite imagery:** increasing in scope, resolution, and availability.

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs
- 3 **Geographic location:** cell phones, Fastlane, garage cameras
- 4 **Health information:** digital medical records, hospital admittances, accelerometers & other devices in cell phones
- 5 **Biological sciences:** genomics, proteomics, metabolomics, brain imaging producing huge numbers of person-level variables
- 6 **Satellite imagery:** increasing in scope, resolution, and availability.
- 7 **Electoral activity:** ballot images, precinct-level results, individual-level registration, primary participation, campaign contributions

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs
- 3 **Geographic location:** cell phones, Fastlane, garage cameras
- 4 **Health information:** digital medical records, hospital admittances, accelerometers & other devices in cell phones
- 5 **Biological sciences:** genomics, proteomics, metabolomics, brain imaging producing huge numbers of person-level variables
- 6 **Satellite imagery:** increasing in scope, resolution, and availability.
- 7 **Electoral activity:** ballot images, precinct-level results, individual-level registration, primary participation, campaign contributions
- 8 **Web surfing artifacts:** clicks, searches, and advertising clickthroughs, multiplayer games, virtual worlds

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs
- 3 **Geographic location:** cell phones, Fastlane, garage cameras
- 4 **Health information:** digital medical records, hospital admittances, accelerometers & other devices in cell phones
- 5 **Biological sciences:** genomics, proteomics, metabolomics, brain imaging producing huge numbers of person-level variables
- 6 **Satellite imagery:** increasing in scope, resolution, and availability.
- 7 **Electoral activity:** ballot images, precinct-level results, individual-level registration, primary participation, campaign contributions
- 8 **Web surfing artifacts:** clicks, searches, and advertising clickthroughs, multiplayer games, virtual worlds
- 9 **> 90% of all data ever created was created last year**

# The *Data* in Big Data

- 1 **Unstructured text:** emails, speeches, reports, social media updates, web pages, newspapers, scholarly literature, product reviews
- 2 **Commerce:** credit cards, sales, real estate transactions, RFIDs
- 3 **Geographic location:** cell phones, Fastlane, garage cameras
- 4 **Health information:** digital medical records, hospital admittances, accelerometers & other devices in cell phones
- 5 **Biological sciences:** genomics, proteomics, metabolomics, brain imaging producing huge numbers of person-level variables
- 6 **Satellite imagery:** increasing in scope, resolution, and availability.
- 7 **Electoral activity:** ballot images, precinct-level results, individual-level registration, primary participation, campaign contributions
- 8 **Web surfing artifacts:** clicks, searches, and advertising clickthroughs, multiplayer games, virtual worlds
- 9 **> 90% of all data ever created was created last year**
- 10 **Popular versions:** *MoneyBall*, *SuperCrunchers*, *The Numerati*

# The *Value* in Big Data: the Analytics

- Data:

# The *Value* in Big Data: the Analytics

- **Data:**
  - becoming commoditized

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

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

# The *Value* in Big Data: the Analytics

- **Data:**
  - becoming commoditized
  - easy to come by; often a free byproduct of IT improvements
  - Ignore it & your company will still 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:**
  - becoming commoditized
  - easy to come by; often a free byproduct of IT improvements
  - Ignore it & your company will still 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:**
  - becoming commoditized
  - easy to come by; often a free byproduct of IT improvements
  - Ignore it & your company will still 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. 1000x increase with one algorithm

# The *Value* in Big Data: the Analytics

- **Data:**
  - becoming commoditized
  - easy to come by; often a free byproduct of IT improvements
  - Ignore it & your company will still 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. 1000x increase with one algorithm
  - \$2M computer v. 2 hours of algorithm design

# The *Value* in Big Data: the Analytics

- **Data:**
  - becoming commoditized
  - easy to come by; often a free byproduct of IT improvements
  - Ignore it & your company will still 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. 1000x increase with one algorithm
  - \$2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed

# The *Value* in Big Data: the Analytics

- **Data:**
  - becoming commoditized
  - easy to come by; often a free byproduct of IT improvements
  - Ignore it & your company will still 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. 1000x increase with one algorithm
  - \$2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - **Innovative analytics:** enormously better than off-the-shelf approaches

# 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 3 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 3 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 3 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 3 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 3 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 3 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 3 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 3 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 3 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 3 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 3 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
- **Expert-vs-analytics contests:** Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins

# Examples of what's now possible

- **Opinions of activists:** A few thousand interviews  $\rightsquigarrow$  billions of political opinions in social media posts (1B every 3 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
- **Expert-vs-analytics contests:** Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- **In each: without new analytics, the data are useless**

# How to Read a Billion Blog Posts & Classify Deaths w/o Physicians

- Examples of Bad Analytics:

# How to Read a Billion Blog Posts & Classify Deaths w/o Physicians

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

# How to Read a Billion Blog Posts & Classify Deaths w/o Physicians

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

# How to Read a Billion Blog Posts & Classify Deaths w/o 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 w/o 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 w/o 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 w/o 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 for estimating %'s led to:

# How to Read a Billion Blog Posts & Classify Deaths w/o 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 for estimating %'s led to:

1



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 Test Size

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

# How to Read a Billion Blog Posts & Classify Deaths w/o 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 for estimating %'s led to:

1



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  Test 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

- **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  $\approx$  **\$1 trillion** 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  $\approx$  **\$1 trillion** more than SSA thought
  - Other applications to insurance industry, public health, etc.

- Writing Technology: Big changes

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc

# Reading and Writing Technology

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc
- **Reading Technology: Little change (ripe for disruption)**

# Reading and Writing Technology

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc
- **Reading Technology: Little change (ripe for disruption)**
  - **Then:** 50, 100, 300 years ago: Get book; read cover to cover

# Reading and Writing Technology

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc
- **Reading Technology: Little change (ripe for disruption)**
  - **Then:** 50, 100, 300 years ago: Get book; read cover to cover
  - **Now:**

# Reading and Writing Technology

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc
- **Reading Technology: Little change (ripe for disruption)**
  - **Then:** 50, 100, 300 years ago: Get book; read cover to cover
  - **Now:**
    - How often do you read a book cover-to-cover for work?

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc
- **Reading Technology: Little change (ripe for disruption)**
  - **Then:** 50, 100, 300 years ago: Get book; read cover to cover
  - **Now:**
    - How often do you read a book cover-to-cover for work?
    - We collect 100s of documents, read a few, delude ourselves into thinking we understand them all

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc
- **Reading Technology: Little change (ripe for disruption)**
  - **Then:** 50, 100, 300 years ago: Get book; read cover to cover
  - **Now:**
    - How often do you read a book cover-to-cover for work?
    - We collect 100s of documents, read a few, delude ourselves into thinking we understand them all
    - Goal: understanding from unstructured data (hardest part of big data)

- **Writing Technology: Big changes**
  - **Then:** Quill tip pen & expensive paper
  - **Now:** Microsoft Word, Google docs, etc
- **Reading Technology: Little change (ripe for disruption)**
  - **Then:** 50, 100, 300 years ago: Get book; read cover to cover
  - **Now:**
    - How often do you read a book cover-to-cover for work?
    - We collect 100s of documents, read a few, delude ourselves into thinking we understand them all
    - Goal: understanding from unstructured data (hardest part of big data)
    - More data isn't helpful! Novel analytics needed.

# 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

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

## ① What Members of Congress Do

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

## 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*

## 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!"

## 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' "

## 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 What is the Chinese Government Censoring?

# 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 What is the Chinese Government Censoring?

- 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 What is the Chinese Government Censoring?

- Previous approach: manual effort to see what is taken down
- Data: Crimson Hexagon gets posts before the Chinese censor them

## 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 What is the Chinese Government Censoring?

- Previous approach: manual effort to see what is taken down
- Data: Crimson Hexagon gets 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 What is the Chinese Government Censoring?

- Previous approach: manual effort to see what is taken down
- Data: Crimson Hexagon gets 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 What is the Chinese Government Censoring?

- Previous approach: manual effort to see what is taken down
- Data: Crimson Hexagon gets 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 What is the Chinese Government Censoring?

- Previous approach: manual effort to see what is taken down
- Data: Crimson Hexagon gets 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

## 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 What is the Chinese Government Censoring?

- Previous approach: manual effort to see what is taken down
- Data: Crimson Hexagon gets 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