Big Data is Not About the Data!

Gary King[1]

Institute for Quantitative Social Science
Harvard University

(Talk at the Harvard FAS Campaign Launch, 10/26/2013)

[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; a free byproduct of IT improvements

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & your company will have more every year
  - Add a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Moore's Law (doubling speed/power every 18 months) v. Harvard 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; a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & your company will have more every year
  - Add a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Moore's Law (doubling speed/power every 18 months) v. Harvard Students (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Innovative analytics: enormously better than off-the-shelf

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

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

- Examples of Bad Analytics:

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

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

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

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

How to Read a Billion Social Media 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 Social Media 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 Social Media 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 Social Media 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 Social Media 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. Worldwide cause-of-death estimates for

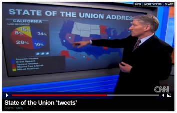
World Health Organization

How to Read a Billion Social Media 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. Worldwide cause-of-death estimates for World Health Organization

2. crimson hexagon ⇝



⇝

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

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

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

Reading and Writing Technology

Reading and Writing Technology

- Writing Technology: Big changes

Reading and Writing Technology

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

Reading and Writing Technology

- 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?

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?
    - We collect 100s of documents, read a few, delude ourselves into thinking we understand them all

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?
    - We collect 100s of documents, read a few, delude ourselves into thinking we understand them all
    - More data isn't helpful! Novel analytics needed.

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?
    - We collect 100s of documents, read a few, delude ourselves into thinking we understand them all
    - More data isn't helpful! Novel analytics needed.
- Our Approach: Computer-Assisted Reading & Insight

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?
    - We collect 100s of documents, read a few, delude ourselves into thinking we understand them all
    - More data isn't helpful! Novel analytics needed.
- Our Approach: Computer-Assisted Reading & Insight
  - Known as "Consilience"

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
   - Categories: (1) advertising, (2) position taking, (3) credit claiming

Example Insights from Computer-Assisted Reading

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

Example Insights from Computer-Assisted Reading

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

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Categories: (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!"

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Categories: (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' "

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do

    - Categories: (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' "
    - How common is it?

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Categories: (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' "
   - How common is it? 27% of all Senatorial press releases!

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Categories: (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' "
    - 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
    - Categories: (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' "
    - How common is it? 27% of all Senatorial press releases!
2. What is the Chinese Government Censoring?
    - Previous approach: look by hand

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Categories: (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' "
    - How common is it? 27% of all Senatorial press releases!
2. What is the Chinese Government Censoring?
    - Previous approach: look by hand
    - Data: download posts before the Chinese censor them

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Categories: (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' "
    - How common is it? 27% of all Senatorial press releases!
2. What is the Chinese Government Censoring?
    - Previous approach: look by hand
    - Data: download 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
   - Categories: (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' "
   - How common is it? 27% of all Senatorial press releases!
2. What is the Chinese Government Censoring?
   - Previous approach: look by hand
   - Data: download posts before the Chinese censor them
   - We analyzed 11 million posts, about 13% censored
   - Previous understanding: they censor criticism of the government

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Categories: (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' "
   - How common is it? 27% of all Senatorial press releases!
2. What is the Chinese Government Censoring?
   - Previous approach: look by hand
   - Data: download posts before the Chinese censor them
   - We analyzed 11 million posts, about 13% censored
   - Previous understanding: they censor criticism of the government
   - Results:

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
   - Categories: (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' "
   - How common is it? 27% of all Senatorial press releases!
2. What is the Chinese Government Censoring?
   - Previous approach: look by hand
   - Data: download posts before the Chinese censor them
   - We analyzed 11 million posts, about 13% censored
   - Previous understanding: they censor criticism of the government
   - Results:
     - Uncensored: criticism of the government

Example Insights from Computer-Assisted Reading

1. What Members of Congress Do
    - Categories: (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' "
    - How common is it? 27% of all Senatorial press releases!
2. What is the Chinese Government Censoring?
    - Previous approach: look by hand
    - Data: download posts before the Chinese censor them
    - We analyzed 11 million posts, about 13% censored
    - Previous understanding: they censor criticism of the government
    - Results:
        - Uncensored: criticism of the government
        - Censored: attempts at collective action

For more information

Gary King

Institute for Quantitative Social Science

King@Harvard.edu