Big Data is Not About the Data!

Gary King[1]

Institute for Quantitative Social Science
Harvard University

Shanghai Jiao Tong University, 1/4/2017

[1] GaryKing.org

The Spectacular Success of Quantitative Social Science

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people):

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media)

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis; impacted crime and policing

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis; impacted crime and policing; reinvented economics

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis; impacted crime and policing; reinvented economics; transformed sports

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis; impacted crime and policing; reinvented economics; transformed sports; set standards for evaluating public policy

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis; impacted crime and policing; reinvented economics; transformed sports; set standards for evaluating public policy; etc.

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis; impacted crime and policing; reinvented economics; transformed sports; set standards for evaluating public policy; etc.; etc.

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling life expectancy in the last century?

⋮

- Quantitative social science (aka "big data," "data analytics," "data science" applied to people): transformed most Fortune 500 firms; established new industries; altered friendship networks; increased human expressive capacity (social media); changed political campaigns; transformed public health; changed legal analysis; impacted crime and policing; reinvented economics; transformed sports; set standards for evaluating public policy; etc.; etc., etc.

The *Value* in Big Data: the Analytics

The *Value* in Big Data: the Analytics

- Data:

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months) v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf
- Exciting data, useless without novel analytics

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months)
    v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf
- Exciting data, useless without novel analytics
  - Opinions of activists:

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months)
    v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf
- Exciting data, useless without novel analytics
  - Opinions of activists: A few thousand interviews

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months) v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf
- Exciting data, useless without novel analytics
  - Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (650M/day)

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months) v. One good data scientist (1000x speed increase in 1 day)
  - \$2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf
- Exciting data, useless without novel analytics
  - Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (650M/day)
  - Exercise:

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months) v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf
- Exciting data, useless without novel analytics
  - Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (650M/day)
  - Exercise: A survey: "How many times did you exercise last week?

The *Value* in Big Data: the Analytics

- Data:
  - easy to come by; often a free byproduct of IT improvements
  - becoming commoditized
  - Ignore it & every institution will have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore's Law (doubling speed/power every 18 months) v. One good data scientist (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf
- Exciting data, useless without novel analytics
  - Opinions of activists: A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (650M/day)
  - Exercise: A survey: "How many times did you exercise last week? $\rightsquigarrow$ 500K people carrying cell phones with accelerometers

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

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

- Examples of Bad Analytics:

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

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

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

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

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

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:

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

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:
  - Key to both methods: *classifying* (deaths, social media posts)

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

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

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

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

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

- Examples of Bad Analytics:
  - Physicians' "Verbal Autopsy" analysis
  - Sentiment analysis via word counts
- Unrelated substantive problems, same analytics solution:
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating %'s*
- Modern Data Analytics: New method led to:
  1.

crimson hexagon ⇝ 

⇝

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

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

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

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

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

1. crimson hexagon ⇝ 

⇝

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

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

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

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

Modern Analytics to Improve Student Learning

Modern Analytics to Improve Student Learning

- The problem:

Modern Analytics to Improve Student Learning

- The problem:
    - How many students buy the book?

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<50\%$

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments?

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%

Modern Analytics to Improve Student Learning

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

Modern Analytics to Improve Student Learning

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

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? <50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? <50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? <50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation:

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation:

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics:

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? <50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - <u>Novel data analytics:</u> keep students on track, with automated personal guidance, nudges,

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, non-adversarial grading

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? $<$50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - $>$90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, non-adversarial grading
  - Instructors save time, stay engaged: automated student confusion reports

Modern Analytics to Improve Student Learning

- The problem:
  - How many students buy the book? <50%
  - How many students do reading assignments? 20-30%
  - How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
  - A new type of collaborative e-reader, with novel data analytics, and cutting-edge behavioral research
  - >90% of students do the reading
  - Solitary reading assignments $\rightsquigarrow$ engaging collective activities
  - Intrinsic motivation: collaborative annotation in threads
  - Extrinsic motivation: automated grading of annotations & engagement (better than instructors can do on their own)
  - Novel data analytics: keep students on track, with automated personal guidance, nudges, non-adversarial grading
  - Instructors save time, stay engaged: automated student confusion reports
  - Want to try it at SJTU? see Perusall.com

Bias in U.S. Social Security Administration Forecasts

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Bias in U.S. Social Security Administration Forecasts

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

Humans are Terrible at Thinking of Keywords

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

- An experiment:

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples:

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

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev,

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings,

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

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying,

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, ...

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, ...
- Median keywords recalled by 43 undergrads:

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, ...
- Median keywords recalled by 43 undergrads: 8

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, ...
- Median keywords recalled by 43 undergrads: 8
- Unique keywords recalled:

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, . . .
- Median keywords recalled by 43 undergrads: 8
- Unique keywords recalled: 149

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, . . .
- Median keywords recalled by 43 undergrads: 8
- Unique keywords recalled: 149
- Keywords 42 of 43 failed to recall:

Humans are Terrible at Thinking of Keywords

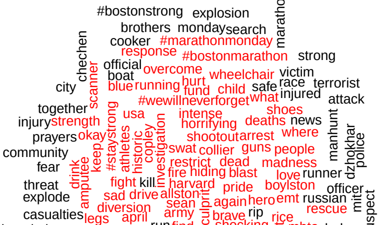
- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, ...
- Median keywords recalled by 43 undergrads: 8
- Unique keywords recalled: 149
- Keywords 42 of 43 failed to recall: 98 (66%)

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, . . .
- Median keywords recalled by 43 undergrads: 8
- Unique keywords recalled: 149
- Keywords 42 of 43 failed to recall: 98 (66%)
- ⇝ Humans recognize keywords well, recall them poorly

Humans are Terrible at Thinking of Keywords

- An experiment: We have 10,000 twitter posts, each containing the word "Boston," from the time period surrounding the Boston Marathon bombings. Please list any keywords which come to mind that will select posts in this set related to the bombings and will not select posts unrelated to the bombings.
- Examples: Tsarnaev, #BostonBombings, horrifying, . . .
- Median keywords recalled by 43 undergrads: 8
- Unique keywords recalled: 149
- Keywords 42 of 43 failed to recall: 98 (66%)
- $\rightsquigarrow$ Humans recognize keywords well, recall them poorly
- Thresher: New technology to discover the right keywords



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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

Computer-Assisted Reading (Consilience)

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

What Members of Congress Do

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:

What Members of Congress Do
(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising*,

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising*,
  2. *position taking*,

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising,*
  2. *position taking,*
  3. *credit claiming*

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising,*
  2. *position taking,*
  3. *credit claiming*
- Data: 64,000 Senators' press releases

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising,*
  2. *position taking,*
  3. *credit claiming*
- Data: 64,000 Senators' press releases
- New Insight: *partisan taunting*

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

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

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

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

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising,*
  2. *position taking,*
  3. *credit claiming*
- Data: 64,000 Senators' press releases
- New Insight: *partisan taunting*
  - Joe Wilson during Obama's State of the Union: "You lie!"
  - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
  - Basically anything said by a 2016 presidential candidate!

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising,*
  2. *position taking,*
  3. *credit claiming*
- Data: 64,000 Senators' press releases
- New Insight: *partisan taunting*
  - Joe Wilson during Obama's State of the Union: "You lie!"
  - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
  - Basically anything said by a 2016 presidential candidate!
- How common is it?

What Members of Congress Do

(Example Insight from Computer-Assisted Reading)

- Categorization from prior research:
  1. *advertising,*
  2. *position taking,*
  3. *credit claiming*
- Data: 64,000 Senators' press releases
- New Insight: *partisan taunting*
  - Joe Wilson during Obama's State of the Union: "You lie!"
  - "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' "
  - Basically anything said by a 2016 presidential candidate!
- How common is it? 27% of all Senatorial press releases!

Reverse Engineering China's "50c Party"

Reverse Engineering China's "50c Party"

- Prevailing view of scholars, activists, journalists, social media participants:

Reverse Engineering China's "50c Party"

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

Reverse Engineering China's "50c Party"

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

Existing evidence?

Reverse Engineering China's "50c Party"

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

Existing evidence? A few anecdotes;

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

Existing evidence? A few anecdotes; "no ground truth";

Reverse Engineering China's "50c Party"

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

Existing evidence? A few anecdotes; "no ground truth"; "no successful attempts to quantify" 50c party activity;

Reverse Engineering China's "50c Party"

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

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

Reverse Engineering China's "50c Party"

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

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

Our evidence:

Reverse Engineering China's "50c Party"

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

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

Our evidence: (1) Used a leaked archive of 50c posts too hard to analyze,

Reverse Engineering China's "50c Party"

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

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

Our evidence: (1) Used a leaked archive of 50c posts too hard to analyze, (2) developed methods of automated text analysis to decipher,

Reverse Engineering China's "50c Party"

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

Existing evidence? A few anecdotes; "no ground truth"; "no successful attempts to quantify" 50c party activity; even several analyses with made up dependent variables!
Our evidence: (1) Used a leaked archive of 50c posts too hard to analyze, (2) developed methods of automated text analysis to decipher, (3) discovered patterns and extrapolated to all of China,

Reverse Engineering China's "50c Party"

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

Existing evidence? A few anecdotes; "no ground truth"; "no successful attempts to quantify" 50c party activity; even several analyses with made up dependent variables!
Our evidence: (1) Used a leaked archive of 50c posts too hard to analyze, (2) developed methods of automated text analysis to decipher, (3) discovered patterns and extrapolated to all of China, (4) did a poll(!) and predicted 50c members acknowledged their behavior

Reverse Engineering China's "50c Party"

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

Reverse Engineering China's "50c Party"

- ~~Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies~~ *Wrong*
- Fabricates 450M social media posts a year!

Reverse Engineering China's "50c Party"

- ~~Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies~~ *Wrong*
- Fabricates 450M social media posts a year!
- Does not argue; does not engage on controversial issues

Reverse Engineering China's "50c Party"

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

The End of The Quantitative-Qualitative Divide

The 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, now analyzing as data unstructured text, video, audio, location, transactions, conversations, etc.

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, now analyzing as data unstructured text, video, audio, location, transactions, conversations, etc.
- 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, now analyzing as data unstructured text, video, audio, location, transactions, conversations, etc.
- 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, now analyzing as data unstructured text, video, audio, location, transactions, conversations, etc.
- 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, now analyzing as data unstructured text, video, audio, location, transactions, conversations, etc.
- 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, now analyzing as data unstructured text, video, audio, location, transactions, conversations, etc.
- 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, now analyzing as data unstructured text, video, audio, location, transactions, conversations, etc.
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- Moral of the story:
  - Fully human is inadequate
  - Fully automated fails
  - We need computer assisted, human controlled technology
  - (Technically correct, & politically much easier)

How To Take Advantage of Big Analytics

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ <u>immensely</u> better than off-the-shelf
- Save it for ~~last~~ first!

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need
    (most statistics are designed solely to overcome data problems)

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need (most statistics are designed solely to overcome data problems)
  - Building analytics during design:

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need (most statistics are designed solely to overcome data problems)
  - Building analytics during design:
    - avoids problems before they occur

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need
    (most statistics are designed solely to overcome data problems)
  - Building analytics during design:
    - avoids problems before they occur
    - saves a fortune,

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need
    (most statistics are designed solely to overcome data problems)
  - Building analytics during design:
    - avoids problems before they occur
    - saves a fortune,
    - opens many more possibilities

How To Take Advantage of Big Analytics

- Its cheap and powerful; don't skimp!
  - Off-the-shelf analytics $\rightsquigarrow$ big advances
  - Innovative analytics $\rightsquigarrow$ immensely better than off-the-shelf
- Save it for ~~last~~ first!
  - The goal is "inference":
    using facts you know to learn about facts you don't know
  - The uncertainties in inference: not having the facts you need (most statistics are designed solely to overcome data problems)
  - Building analytics during design:
    - avoids problems before they occur
    - saves a fortune,
    - opens many more possibilities
- Build a new discipline of data science

For more information

GaryKing.org

Perusall.com

Institute for Quantitative Social Science
Harvard University