Big Data is Not About the Data!

$\mathsf{Gary}\ \mathsf{King}^1$

Institute for Quantitative Social Science Harvard University

Conference on the Future of Science, Venice, Italy 9/23/2016

¹GaryKing.org

The *Data* In Big Data (about people) The Last 50 Years:

The Last 50 Years:

• Survey research

The Last 50 Years:

- Survey research
- Aggregate government statistics

The Last 50 Years:

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

The Last 50 Years:

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

The Last 50 Years:

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

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

• Much more of the above — improved, expanded, and applied

The Last 50 Years:

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

- Much more of the above improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere

The Last 50 Years:

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

- Much more of the above improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce

The Last 50 Years:

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

- Much more of the above improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

- Much more of the above improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- *The march of quantification*: through academia, professions, government, & commerce
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- Impact: changed most Fortune 500 firms

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

The Last 50 Years:

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

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

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

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

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

- 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

- 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

- 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

- 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)

- 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

• Examples of Bad Analytics:

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

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

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

- 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)

- 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

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

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



1.



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	T Text Size - +

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

• Examples of Bad Analytics:

1.

- 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:



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



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

2. Worldwide cause-of-death estimates for



Thresher: Finding Those Hiding in Plain Sight Example Substitution 1:

Example Substitution 1:

4		τ.
Н		\Box
	H	Ц

Example Substitution 1:

自由 "Freedom"

Example Substitution 1:





Example Substitution 1:







Example Substitution 1:

"Eye field"





Example Substitution 1:



"Freedom" CENSORED

"Eye field" (nonsensical)

Example Substitution 1: Homograph





"Eye field" (nonsensical)

Example Substitution 1: Homograph





"Eye field" (nonsensical)

Example Substitution 1: Homograph



"Eye field" (nonsensical)

Example Substitution 2:

Example Substitution 1: Homograph

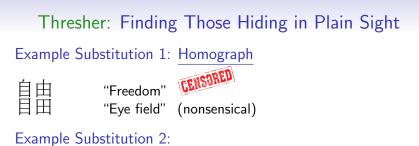
"Freedom"



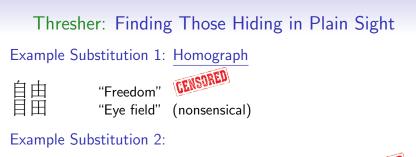
"Eye field" (nonsensical)

Example Substitution 2:

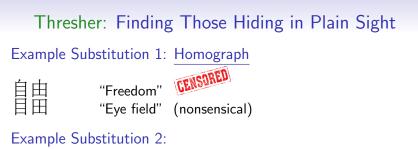
和谐



和谐 "Harmonious [Society]" (official slogan)



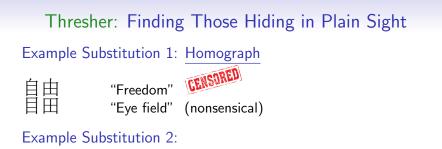
和谐 "Harmonious [Society]" (official slogan) Gensered



和谐

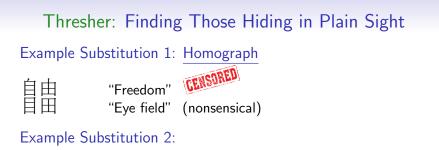


和 谐 河蟹





"Harmonious [Society]" (official slogan)



和	谐
泂	蟹

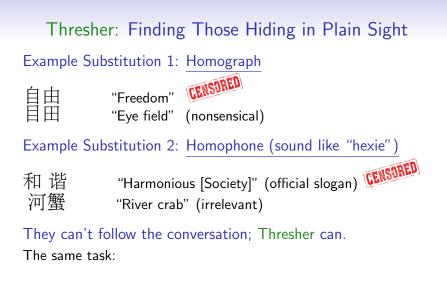
"Harmonious [Society]" (official slogan) GENSIAED "River crab" (irrelevant)

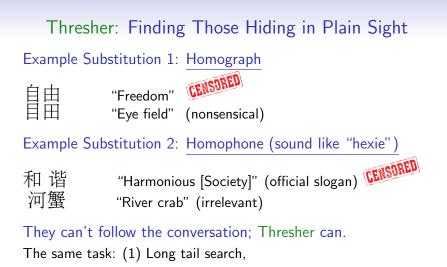




They can't follow the conversation;









The same task: (1) Long tail search, (2) Government and industry analyst's job,



The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings ↔ #BostonStrong),



The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings ↔ #BostonStrong), (4) Child pornographers,



The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings \rightsquigarrow #BostonStrong), (4) Child pornographers, (5) Look-alike modeling,



They can't follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings \rightsquigarrow #BostonStrong), (4) Child pornographers, (5) Look-alike modeling, (6) Starting point for other automated text methods,



They can't follow the conversation; Thresher can.

The same task: (1) Long tail search, (2) Government and industry analyst's job, (3) language drift (#BostonBombings ~~ #BostonStrong), (4) Child pornographers, (5) Look-alike modeling, (6) Starting point for other automated text methods, (7) Infinitely improvable classification, eDiscovery, etc., etc.

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

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

- 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:

- 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!

- 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!)

- 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

- 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

- 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

- 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

- 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

- 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

• Data: 64,000 Senators' press releases

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

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

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

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

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

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

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

• The problem:

- The problem:
 - How many students do reading assignments?

- The problem:
 - How many students do reading assignments? 20-30%

• The problem:

- How many students do reading assignments? 20-30%
- How many students buy the book?

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

• The problem:

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

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

• The problem:

- How many students do reading assignments? 20-30%
- How many students buy the book? <50%
- How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
 - A new type of (award-winning, patent pending) collaborative e-reader

- How many students do reading assignments? 20-30%
- How many students buy the book? <50%
- How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
 - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics

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

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

- How many students do reading assignments? 20-30%
- How many students buy the book? <50%
- How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
 - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
 - >90% of students do the reading
 - Solitary reading assignments → engaging collective activities

- How many students do reading assignments? 20-30%
- How many students buy the book? <50%
- How much time do instructors have to write detailed quizzes?
- Our solution: Perusall
 - A new type of (award-winning, patent pending) collaborative e-reader, using novel data analytics, and cutting-edge behavioral research
 - $\bullet~>90\%$ of students do the reading
 - Solitary reading assignments → engaging collective activities
 - Intrinsic motivation:

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

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

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

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

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

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

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

• Previous approach: watch a few posts; see what's removed

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

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

Stop criticism and protest about the state, its leaders, and their policies *Wrong*

• What Could be the Goal?

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

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

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

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

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

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

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials
 - censor: to stop events with collective action potential

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials
 - censor: to stop events with collective action potential
 - Thus, we can use criticism & censorship to predict:

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials
 - censor: to stop events with collective action potential
 - Thus, we can use criticism & censorship to predict:
 - Officials in trouble, likely to be replaced

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials
 - censor: to stop events with collective action potential
 - Thus, we can use criticism & censorship to predict:
 - Officials in trouble, likely to be replaced
 - Dissident arrests;

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials
 - censor: to stop events with collective action potential
 - Thus, we can use criticism & censorship to predict:
 - Officials in trouble, likely to be replaced
 - Dissident arrests; new peace treaties;

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials
 - censor: to stop events with collective action potential
 - Thus, we can use criticism & censorship to predict:
 - Officials in trouble, likely to be replaced
 - Dissident arrests; new peace treaties; emerging scandals

- Previous approach: watch a few posts; see what's removed
- Data: Download all posts before the Chinese censor them
- pprox 13% censored overall
- Everyone knows the Goal:

- What Could be the Goal?
 - 1. Stop criticism of the state Wrong
 - 2. Stop collective action *Right*
- Implications: Social Media is Actionable!
 - Chinese leaders:
 - measure criticism: to judge local officials
 - censor: to stop events with collective action potential
 - Thus, we can use criticism & censorship to predict:
 - Officials in trouble, likely to be replaced
 - Dissident arrests; new peace treaties; emerging scandals
 - Disagreements between central and local leaders

Reverse Engineering China's 50c Party

Reverse Engineering China's 50c Party

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

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

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

Evidence?

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

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

Evidence? A few anecdotes; "no ground truth";

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

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

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

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

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

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

- Prevailing view of scholars, activists, journalists, social media participants: 50c party argues against those who criticize the government, its leaders, and their policies Wrong
- 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 Quant-Qual divide exists in *every* field.

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

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

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

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

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

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

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

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

• Its cheap and powerful; don't skimp!

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

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

- 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!

- 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

- 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)

- 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:

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

- 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

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

- 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,

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

- 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

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

- The goal is "inference": using facts you know to learn about facts you don't know
- The uncertainties in inference: not having the facts you need (most statistics are designed solely to overcome data problems)
- Building analytics during design:
 - avoids problems before they occur
 - saves a fortune,
 - opens many more possibilities
- Build a new discipline of data science

For more information

GaryKing.org

Institute for Quantitative Social Science Harvard University