Big Data is Not About the Data!

Gary King

Institute for Quantitative Social Science
Harvard University

Talk at the MIT Analytics Lab, 9/29/2015
The Data In Big Data (about people)

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
• Shrinking computers & the growing Internet: data everywhere
• The replication movement: data sharing (e.g., Dataverse)
• Governments encouraging data collection & experimentation
• Advances in statistical methods, informatics, & software
• The march of quantification: through academia, professions, government, & commerce (SuperCrunchers, The Numerati, MoneyBall, and innumerable "big data" articles)

• 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., etc.
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- The march of quantification: through academia, professions, government, & commerce (SuperCrunchers, The Numerati, MoneyBall, and innumerable "big data" articles)

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., etc.
The *Data* In Big Data (about people)

The Last 50 Years:

- Survey research
The Data In Big Data (about people)

The Last 50 Years:
- Survey research
- Aggregate government statistics

The Next 50 Years: Fast increases in new data sources, due to...
- Much more of the above — improved, expanded, and applied
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- The march of quantification: through academia, professions, government, & commerce (SuperCrunchers, The Numerati, MoneyBall, and innumerable "big data" articles)

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., etc.
The Data In Big Data (about people)

The Last 50 Years:

- Survey research
- Aggregate government statistics
- One off studies of individual places, people, or events
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- The march of quantification: through academia, professions, government, & commerce (SuperCrunchers, The Numerati, MoneyBall, and innumerable "big data" articles)

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., etc.
The Data In Big Data (about people)

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

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., etc.
The *Data In Big Data* (about people)

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
- Shrinking computers & the growing Internet: data everywhere

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., etc.
The *Data In Big Data* (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)

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., etc.
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation

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., etc.
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- 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., etc.
**The Data In Big Data (about people)**

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification:* through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)

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., etc
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)

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., etc.
The *Data* In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)

Impact:
The *Data* In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries
The *Data In Big Data* (about people)

**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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification:* through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks
The *Data In Big Data* (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- The march of quantification: through academia, professions, government, & commerce (SuperCrunchers, The Numerati, MoneyBall, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- The march of quantification: through academia, professions, government, & commerce (SuperCrunchers, The Numerati, MoneyBall, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification:* through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing
The **Data In Big Data** (about people)

**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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports
The Data In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- The march of quantification: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- Impact: changed most Fortune 500 firms; established new industries; altered friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy
The *Data In Big Data* (about people)

**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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification:* through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- 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 *Data* In Big Data (about people)

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
- Shrinking computers & the growing Internet: data everywhere
- The replication movement: data sharing (e.g., Dataverse)
- Governments encouraging data collection & experimentation
- Advances in statistical methods, informatics, & software
- *The march of quantification*: through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*, and innumerable “big data” articles)
- 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., etc
The Value in Big Data: the Analytics

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

- Where the Value is: the Analytics
- Output can be highly customized

- Moore’s Law (doubling speed/power every 18 months) v. Our Students (1000x speed increase in 1 day)
- $2M computer v. 2 hours of algorithm design
- Low cost; little infrastructure; mostly human capital needed
- Innovative analytics: enormously better than off-the-shelf
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) vs. Our Students (1000x speed increase in 1 day)
- $2M computer vs. 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
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

---

*Moore's Law (doubling speed/power every 18 months) v. Our Students (1000x speed increase in 1 day)*

- $2M computer v. 2 hours of algorithm design
- Low cost; little infrastructure; mostly human capital needed
- Innovative analytics: enormously better than off-the-shelf
The Value in Big Data: the Analytics

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

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

- **Where the Value is:** the Analytics
The Value in Big Data: the Analytics

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

• Where the Value is: the Analytics
  • Output can be highly customized
The Value in Big Data: the Analytics

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

- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore’s Law (doubling speed/power every 18 months)
The Value in Big Data: the Analytics

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

- **Where the Value is: the Analytics**
  - Output can be highly customized
  - Moore’s Law (doubling speed/power every 18 months) v. Our Students (1000x speed increase in 1 day)
The **Value** in Big Data: the Analytics

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

- **Where the Value is: the Analytics**
  - Output can be highly customized
  - Moore’s Law (doubling speed/power every 18 months) v. Our Students (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
The Value in Big Data: the Analytics

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

• Where the Value is: the Analytics
  • Output can be highly customized
  • Moore’s Law (doubling speed/power every 18 months) v. Our Students (1000x speed increase in 1 day)
  • $2M computer v. 2 hours of algorithm design
  • Low cost; little infrastructure; mostly human capital needed
The **Value** in Big Data: the Analytics

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

- **Where the Value is: the Analytics**
  - Output can be highly customized
  - Moore’s Law (doubling speed/power every 18 months) v. Our Students (1000x speed increase in 1 day)
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - **Innovative analytics:** enormously better than off-the-shelf
Examples of what’s now possible

- Opinions of activists:
  - A few thousand interviews ➞ billions of political opinions in social media posts (650M/day)

- Exercise:
  - A survey: “How many times did you exercise last week?” ➞ 500K people carrying cell phones with accelerometers

- Social contacts:
  - A survey: “Please tell me your 5 best friends” ➞ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

- Economic development in developing countries:
  - Dubious or nonexistent governmental statistics ➞ satellite images of human-generated light at night, road networks, other infrastructure

- Many, many, more...

In each: without new analytics, the data are useless
Examples of what’s now possible

• Opinions of activists:

• Exercise:
  A survey: “How many times did you exercise last week?”
  ⇝ 500K people carrying cell phones with accelerometers

• Social contacts:
  A survey: “Please tell me your 5 best friends”
  ⇝ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

• Economic development in developing countries:
  Dubious or nonexistent governmental statistics
  ⇝ satellite images of human-generated light at night, road networks, other infrastructure

• Many, many, more...

• In each: without new analytics, the data are useless
Examples of what’s now possible

- Opinions of activists: A few thousand interviews

- Exercise: A survey: “How many times did you exercise last week?”
  - 500K people carrying cell phones with accelerometers

- Social contacts: A survey: “Please tell me your 5 best friends”
  - Continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

- Economic development in developing countries:
  - Dubious or nonexistent governmental statistics
  - Satellite images of human-generated light at night, road networks, other infrastructure

- Many, many, more...

In each: without new analytics, the data are useless
Examples of what’s now possible

- **Opinions of activists:** A few thousand interviews $\sim$ billions of political opinions in social media posts (650M/day)
Examples of what’s now possible

- **Opinions of activists:** A few thousand interviews $\sim$ billions of political opinions in social media posts (650M/day)
- **Exercise:**

\[ 4/23 \]
Examples of what’s now possible

- **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?”
Examples of what’s now possible

- **Opinions of activists**: A few thousand interviews $\leadsto$ billions of political opinions in social media posts (650M/day)
- **Exercise**: A survey: “How many times did you exercise last week? $\leadsto$ 500K people carrying cell phones with accelerometers

Many, many, more...
Examples of what’s now possible

• **Opinions of activists:** A few thousand interviews $\sim$ billions of political opinions in social media posts (650M/day)

• **Exercise:** A survey: “How many times did you exercise last week? $\sim$ 500K people carrying cell phones with accelerometers

• **Social contacts:**
Examples of what’s now possible

• Opinions of activists: A few thousand interviews $\sim\rightarrow$ billions of political opinions in social media posts (650M/day)

• Exercise: A survey: “How many times did you exercise last week? $\sim\rightarrow$ 500K people carrying cell phones with accelerometers

• Social contacts: A survey: “Please tell me your 5 best friends”

• Economic development in developing countries: Dubious or nonexistent governmental statistics $\sim\rightarrow$ satellite images of human-generated light at night, road networks, other infrastructure

• Many, many, more...

In each: without new analytics, the data are useless
Examples of what’s now possible

- **Opinions of activists:** A few thousand interviews $\leadsto$ billions of political opinions in social media posts (650M/day)
- **Exercise:** A survey: “How many times did you exercise last week? $\leadsto$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\leadsto$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

- Many, many, more...

In each: without new analytics, the data are useless
Examples of what’s now possible

- **Opinions of activists:** A few thousand interviews $\sim$ billions of political opinions in social media posts (650M/day)
- **Exercise:** A survey: “How many times did you exercise last week? $\sim$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\sim$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:**
Examples of what’s now possible

- **Opinions of activists:** A few thousand interviews $\leadsto$ billions of political opinions in social media posts (650M/day)
- **Exercise:** A survey: “How many times did you exercise last week? $\leadsto$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\leadsto$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:** Dubious or nonexistent governmental statistics

• Many, many... • In each: without new analytics, the data are useless
Examples of what’s now possible

• **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

• **Social contacts:** A survey: “Please tell me your 5 best friends” \(\rightsquigarrow\) continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

• **Economic development in developing countries:** Dubious or nonexistent governmental statistics \(\rightsquigarrow\) satellite images of human-generated light at night, road networks, other infrastructure
Examples of what’s now possible

- **Opinions of activists:** A few thousand interviews $\sim\sim$ billions of political opinions in social media posts (650M/day)
- **Exercise:** A survey: “How many times did you exercise last week?” $\sim\sim$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\sim\sim$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:** Dubious or nonexistent governmental statistics $\sim\sim$ satellite images of human-generated light at night, road networks, other infrastructure
- **Many, many, more...**
Examples of what’s now possible

- **Opinions of activists**: A few thousand interviews \(\leadsto\) billions of political opinions in social media posts (650M/day)

- **Exercise**: A survey: “How many times did you exercise last week? \(\leadsto\) 500K people carrying cell phones with accelerometers

- **Social contacts**: A survey: “Please tell me your 5 best friends” \(\leadsto\) continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

- **Economic development in developing countries**: Dubious or nonexistent governmental statistics \(\leadsto\) satellite images of human-generated light at night, road networks, other infrastructure

- Many, *many*, more…

- In each: without new analytics, the data are useless
The End of The Quantitative-Qualitative Divide

• The 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)
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

Moral of the story:
- Fully human is inadequate
- Fully automated fails
- We need computer assisted, human controlled technology
  (Technically correct, & politically much easier)
The End of The Quantitative-Qualitative Divide

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

Moral of the story:
- Fully human is inadequate
- Fully automated fails
- We need computer assisted, human controlled technology

(Technically correct, & politically much easier)
The End of The Quantitative-Qualitative Divide

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

Moral of the story:
• Fully human is inadequate
• Fully automated fails
• We need computer assisted, human controlled technology
  (Technically correct, & politically much easier)
The End of The Quantitative-Qualitative Divide

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

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

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

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

- The Quant-Qual divide exists in every field.
- Qualitative researchers: overwhelmed by information; need help
- Quantitative researchers: recognize the huge amounts of information in qualitative analyses, starting to analyze unstructured text, video, audio as data
- Expert-vs-analytics contests: Whenever enough information is quantified, a right answer exists, and good analytics are applied: analytics wins
- Moral of the story:
  - Fully human is inadequate
  - Fully automated fails
  - We need computer assisted, human controlled technology
  - (Technically correct, & politically much easier)
How to Read a Billion Blog Posts & Classify Deaths without Physicians

Examples of Bad Analytics:
- Physicians’ “Verbal Autopsy” analysis
- Sentiment analysis via word counts

Different problems, Same Analytics Solution:
- Key to both methods: classifying (deaths, social media posts)
- Key to both goals: estimating %’s

Modern Data Analytics: New method led to:
How to Read a Billion Blog Posts
& Classify Deaths without Physicians

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

Different problems, Same Analytics Solution:
- Key to both methods: classifying (deaths, social media posts)
- Key to both goals: estimating %'s

Modern Data Analytics: New method led to:
How to Read a Billion Blog Posts & Classify Deaths without Physicians

• Examples of Bad Analytics:
  • Physicians’ “Verbal Autopsy” analysis
How to Read a Billion Blog Posts & Classify Deaths without Physicians

• Examples of Bad Analytics:
  • Physicians’ “Verbal Autopsy” analysis
  • Sentiment analysis via word counts
How to Read a Billion Blog Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians’ “Verbal Autopsy” analysis
  - Sentiment analysis via word counts

- Different problems, Same Analytics Solution:
How to Read a Billion Blog Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians’ “Verbal Autopsy” analysis
  - Sentiment analysis via word counts

- Different problems, Same Analytics Solution:
  - Key to both methods: *classifying* (deaths, social media posts)
How to Read a Billion Blog Posts & Classify Deaths without Physicians

- **Examples of Bad Analytics:**
  - Physicians’ “Verbal Autopsy” analysis
  - Sentiment analysis via word counts

- **Different problems, Same Analytics Solution:**
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating* %’s
How to Read a Billion Blog Posts
& Classify Deaths without Physicians

• Examples of Bad Analytics:
  • Physicians’ “Verbal Autopsy” analysis
  • Sentiment analysis via word counts

• Different problems, Same Analytics Solution:
  • Key to both methods: *classifying* (deaths, social media posts)
  • Key to both goals: *estimating* %’s

• Modern Data Analytics: New method led to:
How to Read a Billion Blog Posts & Classify Deaths without Physicians

- Examples of Bad Analytics:
  - Physicians’ “Verbal Autopsy” analysis
  - Sentiment analysis via word counts

- Different problems, Same Analytics Solution:
  - Key to both methods: *classifying* (deaths, social media posts)
  - Key to both goals: *estimating %’s*

- Modern Data Analytics: New method led to:
  1. [Crimson Hexagon](https://www.crimsonhexagon.com)
How to Read a Billion Blog Posts & Classify Deaths without Physicians

• Examples of Bad Analytics:
  • Physicians’ “Verbal Autopsy” analysis
  • Sentiment analysis via word counts

• Different problems, Same Analytics Solution:
  • Key to both methods: classifying (deaths, social media posts)
  • Key to both goals: estimating %’s

• Modern Data Analytics: New method led to:
  1. 
  2. Worldwide cause-of-death estimates for
The Solvency of Social Security

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

- Solvency: depends on mortality forecasts:
  - If retirees receive benefits longer than expected, the Trust Fund runs out

- SSA data: little change other than updates for 75 years

- SSA analytics:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000

- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts

- Trust fund needs ≈ $800 billion more than SSA thought

- Other applications to insurance industry, public health, etc.
The Solvency of Social Security

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

- Solvency: depends on mortality forecasts:
  - If retirees receive benefits longer than expected, the Trust Fund runs out

- SSA data: little change other than updates for 75 years

- SSA analytics:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000

- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts

  $\Rightarrow$ Trust fund needs $\approx$ $800$ billion more than SSA thought

- Other applications to insurance industry, public health, etc.

7/23
The Solvency of Social Security

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

  - SSA data: little change other than updates for 75 years
  - SSA analytics: few statistical improvements for 75 years
    - Ignores risk factors (smoking, obesity)
    - Mostly informal (subject to error & political influence)
    - Forecasts: All systematically biased since 2000
  - New customized analytics we developed:
    - Logical consistency (e.g., older people have higher mortality)
    - More accurate forecasts
  - Trust fund needs ≈ $800 billion more than SSA thought
  - Other applications to insurance industry, public health, etc.
The Solvency of Social Security

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

SSA data: little change other than updates for 75 years

SSA analytics:
- Few statistical improvements for 75 years
- Ignores risk factors (smoking, obesity)
- Mostly informal (subject to error & political influence)
- Forecasts: All systematically biased since 2000

New customized analytics we developed:
- Logical consistency (e.g., older people have higher mortality)
- More accurate forecasts

⇒ Trust fund needs ≈ $800 billion more than SSA thought

Other applications to insurance industry, public health, etc.
The Solvency of Social Security

- **Successful**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- **SSA data**: little change other than updates for 75 years
• **Successful:** single largest government program; lifted a whole generation out of poverty; extremely popular
• **Solvency:** depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
• **SSA data:** little change other than updates for 75 years
• **SSA analytics:**
  • Few statistical improvements for 75 years
  • Ignores risk factors (smoking, obesity)
  • Mostly informal (subject to error & political influence)
  • Forecasts: All systematically biased since 2000
• New customized analytics we developed:
  • Logical consistency (e.g., older people have higher mortality)
  • More accurate forecasts
• \(\Rightarrow\) Trust fund needs \(\approx \$800\) billion more than SSA thought
• Other applications to insurance industry, public health, etc.
The Solvency of Social Security

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

⇝ Trust fund needs ≈ $800 billion more than SSA thought

Other applications to insurance industry, public health, etc.
The Solvency of Social Security

- **Successful**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- **SSA data**: little change other than updates for 75 years
- **SSA analytics**:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
- New customized analytics we developed:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts
- Trust fund needs ≈ $800 billion more than SSA thought
- Other applications to insurance industry, public health, etc.
The Solvency of Social Security

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

• **Solvency:** depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out

• **SSA data:** little change other than updates for 75 years

• **SSA analytics:**
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)

⇝ Trust fund needs $800 billion more than SSA thought

Other applications to insurance industry, public health, etc.
The Solvency of Social Security

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

• **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out

• **SSA data**: little change other than updates for 75 years

• **SSA analytics**:
  • Few statistical improvements for 75 years
  • Ignores risk factors (smoking, obesity)
  • Mostly informal (subject to error & political influence)
  • Forecasts: All systematically biased since 2000

• New customized analytics we developed:
  • Logical consistency (e.g., older people have higher mortality)
  • More accurate forecasts

• Trust fund needs $≈ 800 billion more than SSA thought

• Other applications to insurance industry, public health, etc.

7/23
The Solvency of Social Security

- **Successful**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- **SSA data**: little change other than updates for 75 years
- **SSA analytics**:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000
- **New customized analytics we developed**:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts
  - Trust fund needs $\approx 800$ billion more than SSA thought

Other applications to insurance industry, public health, etc.
The Solvency of Social Security

- **Successful**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- **SSA data**: little change other than updates for 75 years
- **SSA analytics**:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000
- **New customized analytics we developed**:
  - Logical consistency (e.g., older people have higher mortality)
The Solvency of Social Security

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

- **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out

- **SSA data**: little change other than updates for 75 years

- **SSA analytics**:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000

- **New customized analytics we developed**:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts

- Trust fund needs \( \approx \$800 \) billion more than SSA thought

- Other applications to insurance industry, public health, etc.
The Solvency of Social Security

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

• **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out

• **SSA data**: little change other than updates for 75 years

• **SSA analytics**:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000

• **New customized analytics we developed**:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts
  - Trust fund needs $\approx 800 \text{ billion}$ more than SSA thought
The Solvency of Social Security

- **Successful**: single largest government program; lifted a whole generation out of poverty; extremely popular
- **Solvency**: depends on mortality forecasts: If retirees receive benefits longer than expected, the Trust Fund runs out
- **SSA data**: little change other than updates for 75 years
- **SSA analytics**:
  - Few statistical improvements for 75 years
  - Ignores risk factors (smoking, obesity)
  - Mostly informal (subject to error & political influence)
  - Forecasts: All systematically biased since 2000
- **New customized analytics we developed**:
  - Logical consistency (e.g., older people have higher mortality)
  - More accurate forecasts
  - Trust fund needs $\approx 800$ billion more than SSA thought
  - Other applications to insurance industry, public health, etc.
Following Conversations that Hide in Plain Sight

Example Substitution 1:
- Homograph
  - 自
  - 由
  - "Freedom"
  - 目
  - 田
  - "Eye field"

Example Substitution 2:
- Homophone (sound like “hexie”)
  - 和
  - 谐
  - "Harmonious [Society]"
  - 河
  - 蟹
  - "River crab"

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

The same task:
1. Government and industry analyst’s job,
2. Language drift (#BostonBombings ⇝ #BostonStrong),
3. Child pornographers,
4. Look-alike modeling,
5. Starting point for sophisticated automated text analysis
Following Conversations that Hide in Plain Sight

Example Substitution 1:
Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由
Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由 “Freedom”
Following Conversations that Hide in Plain Sight

Example Substitution 1:
自由  “Freedom”
Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由 “Freedom”

自由

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

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

Example Substitution 1:

<table>
<thead>
<tr>
<th>自由</th>
<th>“Freedom”</th>
</tr>
</thead>
<tbody>
<tr>
<td>目田</td>
<td>“Eye field”</td>
</tr>
</tbody>
</table>

They can't follow the conversation; Our methods can!
Following Conversations that Hide in Plain Sight

Example Substitution 1:

自由 自由
“Freedom” “Freedom”
目田 目田
“Eye field” (nonsensical)
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由  “Freedom”
自由  “Eye field” (nonsensical)
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由 自由
目田 目田
“Freedom” “Eye field” (nonsensical)
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和 谐 “Harmonious [Society]” (official slogan)
河 蟹 “River crab” (irrelevant)

They can't follow the conversation; Our methods can!
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由 自 由
自由 自由
“Freedom” “Eye field” (nonsensical)

Example Substitution 2:

和谐
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐  “Harmonious [Society]” (official slogan)
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由  “Freedom”
自由  “Eye field” (nonsensical)

Example Substitution 2:

和谐  “Harmonious [Society]” (official slogan)
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐  “Harmonious [Society]” (official slogan)
河蟹
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐 和谐
河蟹 “Harmonious [Society]” (official slogan)
河蟹 “River crab”
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

Example Substitution 2:

和谐 和谐  “Harmonious [Society]” (official slogan)
河蟹 河蟹  “River crab” (irrelevant)
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

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

和谐
河蟹
“Harmonious [Society]” (official slogan)
“River crab” (irrelevant)
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

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

和谐
“Harmonious [Society]” (official slogan)
河蟹
“River crab” (irrelevant)

They can’t follow the conversation; Our methods can!
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由  自由
“Freedom”  “Eye field” (nonsensical)

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

和谐 和谐
“Harmonious [Society]” (official slogan)  “River crab” (irrelevant)

They can’t follow the conversation; Our methods can!
The same task:
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

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

和谐  “Harmonious [Society]” (official slogan)
河蟹  “River crab” (irrelevant)

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

The same task: (1) Government and industry analyst’s job,
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由 自由
“Freedom” (censored)
“Eye field” (nonsensical)

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

和谐 和谐
“Harmonious [Society]” (official slogan) (censored)
“River crab” (irrelevant)

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

The same task: (1) Government and industry analyst’s job, (2) language drift (#BostonBombings ⇝ #BostonStrong),
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph
自 由
“Freedom”
目 田
“Eye field” (nonsensical)

Example Substitution 2: Homophone (sound like “hexie”)
和 谐
“Harmonious [Society]” (official slogan)
河 蟹
“River crab” (irrelevant)

They can’t follow the conversation; **Our methods can!**
The same task: (1) Government and industry analyst’s job, (2) language drift (♯BostonBombings ⇝ ♫BostonStrong), (3) Child pornographers,
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

自由
“Freedom”
自由
“Eye field” (nonsensical)

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

和谐
“Harmonious [Society]” (official slogan)
和谐
“River crab” (irrelevant)

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

The same task: (1) Government and industry analyst’s job, (2) language drift (#BostonBombings ⇝ #BostonStrong), (3) Child pornographers, (4) Look-alike modeling,
Following Conversations that Hide in Plain Sight

Example Substitution 1: Homograph

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

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

和谐  “Harmonious [Society]” (official slogan)
河蟹  “River crab” (irrelevant)

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

The same task: (1) Government and industry analyst’s job, (2) language drift (#BostonBombings ⇝ #BostonStrong), (3) Child pornographers, (4) Look-alike modeling, (5) Starting point for sophisticated automated text analysis
To understand many documents, humans create categories to represent conceptualization, insight, etc. Most firms impose fixed categorizations to tally customer complaints, sort reports, retrieve information. Bad Analytics: Unassisted Human Categorization: time consuming; huge efforts trying not to innovate! Fully Automated "Cluster Analysis": Many widely available, but none work (computers don’t know what you want!). Our alternative: Computer-assisted Categorization. You decide what’s important, but with help. Invert effort: you innovate; the computer categorizes. Insights: easier, faster, better. (Lots of technology, but it’s behind the scenes.)
Computer-Assisted Reading (Consilience)

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

Most firms: impose fixed categorizations to tally customer complaints, sort reports, retrieve information

Bad Analytics:

Our alternative: Computer-assisted Categorization

You decide what's important, but with help

Invert effort: you innovate; the computer categorizes

Insights: easier, faster, better

(Lots of technology, but it's behind the scenes)
Computer-Assisted Reading (Consilience)

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

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

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

Most firms: impose fixed categorizations to tally customer complaints, sort reports, retrieve information

Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying not to innovate!
  - Fully Automated “Cluster Analysis”: Many widely available, but none work (computers don’t know what you want!)

Our alternative: Computer-assisted Categorization
  - You decide what’s important, but with help
To understand many documents, humans create categories to represent conceptualization, insight, etc.

Most firms: impose fixed categorizations to tally customer complaints, sort reports, retrieve information

Bad Analytics:
  - Unassisted Human Categorization: time consuming; huge efforts trying not to innovate!
  - Fully Automated “Cluster Analysis”: Many widely available, but none work (computers don’t know what you want!)

Our alternative: Computer-assisted Categorization
  - You decide what’s important, but with help
  - Invert effort: you innovate; the computer categorizes
Computer-Assisted Reading (Consilience)

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

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

• Most firms: impose fixed categorizations to tally customer complaints, sort reports, retrieve information

• Bad Analytics:
  • Unassisted Human Categorization: time consuming; huge efforts trying not to innovate!
  • Fully Automated “Cluster Analysis”: Many widely available, but none work (computers don’t know what you want!)

• Our alternative: Computer-assisted Categorization
  • You decide what’s important, but with help
  • Invert effort: you innovate; the computer categorizes
  • Insights: easier, faster, better
  • (Lots of technology, but it’s behind the scenes)
Example Insights from Computer-Assisted Reading

Data: 64,000 Senators' press releases

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

New Insight: partisan taunting

Joe Wilson during Obama's State of the Union: "You lie!"

"Senator Lautenberg Blasts Republicans as 'Chicken Hawks'"

How common is it?

27% of all Senatorial press releases!
Example Insights from Computer-Assisted Reading

What Members of Congress Do

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

What Members of Congress Do

- Data: 64,000 Senators’ press releases

  - New Insight: partisan taunting
  - Joe Wilson during Obama’s State of the Union: “You lie!”
  - “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’”

  - How common is it? 27% of all Senatorial press releases!
Example Insights from Computer-Assisted Reading

What Members of Congress Do

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

New Insight:
- Partisan taunting
- Joe Wilson during Obama’s State of the Union: “You lie!”
- “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’”

How common is it?
- 27% of all Senatorial press releases!
What Members of Congress Do

- Data: 64,000 Senators’ press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: partisan taunting
Example Insights from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators’ press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: partisan taunting
  - Joe Wilson during Obama’s State of the Union: “You lie!”
Example Insights from Computer-Assisted Reading

What Members of Congress Do

- Data: 64,000 Senators’ press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: partisan taunting
  - Joe Wilson during Obama’s State of the Union: “You lie!”
  - “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’ ”
Example Insights from Computer-Assisted Reading

What Members of Congress Do

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

- Data: 64,000 Senators’ press releases
- Categorization: (1) advertising, (2) position taking, (3) credit claiming
- New Insight: *partisan taunting*
  - Joe Wilson during Obama’s State of the Union: “You lie!”
  - “Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’ ”
- How common is it? 27% of all Senatorial press releases!
How did we come to study Chinese Censorship?
How did we come to study Chinese Censorship?

- We were working on methods of automated text analysis.
How did we come to study Chinese Censorship?

- We were working on methods of **automated text analysis**
- How to stress test the methods?
How did we come to study Chinese Censorship?

- We were working on methods of **automated text analysis**
- How to stress test the methods? **Do they work in Chinese?**
How did we come to study Chinese Censorship?

- We were working on methods of automated text analysis.
- How to stress test the methods? Do they work in Chinese?
How did we come to study Chinese Censorship?

- We were working on methods of automated text analysis
- How to stress test the methods? Do they work in Chinese?
- We had the content of millions of censored Chinese posts!
Censorship is not Ambiguous: Example Error Page

The page you requested is temporarily down. How about you go look at another page.

Jingjing, one of China’s cartoon internet police
Chinese Censorship

The largest selective suppression of human expression in history
implemented manually (within a few hours of posting),
by ≈ 200,000 workers,
located in government and inside social media firms

A huge censorship organization:
(obviously) designed to suppress information
(paradoxically) very revealing about the goals, intentions, and actions of the Chinese leadership
Chinese Censorship

- The largest selective suppression of human expression in history
Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by \( \approx 200,000 \) workers,
Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by \( \approx 200,000 \) workers,
  - located in government and inside social media firms
Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by \( \approx 200,000 \) workers,
  - located in government and inside social media firms
- A huge censorship organization:
Chinese Censorship

- The largest selective suppression of human expression in history
  - implemented *manually* (within a few hours of posting),
  - by $\approx 200,000$ workers,
  - located in government and inside social media firms
- A huge censorship organization:
  - (obviously) designed to suppress information
Chinese Censorship

• The largest selective suppression of human expression in history
  • implemented *manually* (within a few hours of posting),
  • by $\approx 200,000$ workers,
  • located in government and inside social media firms

• A huge censorship organization:
  • (obviously) designed to suppress information
  • (paradoxically) very revealing about the goals, intentions, and actions of the Chinese leadership
The Goals of Censorship make Social Media Actionable

• What Could be the Goal?
  1. Stop collective action

• 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
  • Policies that generate dissent
  • Dissidents to be arrested; peace treaties to sign; emerging scandals
  • Disagreements between central and local leaders
The Goals of Censorship make Social Media Actionable

- Everyone knows the Goal:
The Goals of Censorship make Social Media Actionable

• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies
The Goals of Censorship make Social Media Actionable

• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies Wrong
The Goals of Censorship make Social Media Actionable

- Everyone knows the Goal: Stop criticism and protest about the state, its leaders, and their policies  

Wrong

- What Could be the Goal?
The Goals of Censorship make Social Media Actionable

• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies \textit{Wrong}

• What Could be the Goal?
  1. Stop criticism of the state
The Goals of Censorship make Social Media Actionable

- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*

- **What Could be the Goal?**
  1. Stop criticism of the state
  2. Stop collective action
The Goals of Censorship make Social Media Actionable

- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*

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

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 and censorship to predict:
  - Officials in trouble, likely to be replaced
  - Policies that generate dissent
  - Dissidents to be arrested; peace treaties to sign; emerging scandals
  - Disagreements between central and local leaders
The Goals of Censorship make Social Media Actionable

- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*

- **What Could be the Goal?**
  1. Stop criticism of the state *Wrong*
  2. Stop collective action *Right*
The Goals of Censorship make Social Media Actionable

- Everyone knows the Goal: Stop criticism and protest about the state, its leaders, and their policies **Wrong**
- What Could be the Goal?
  1. Stop criticism of the state **Wrong**
  2. Stop collective action **Right**
- Implications: Social Media is Actionable!
The Goals of Censorship make Social Media Actionable

- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*

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

- **Implications: Social Media is Actionable!**
  - Chinese leaders:
The Goals of Censorship make Social Media Actionable

• Everyone knows the Goal:
  Stop criticism and protest about the state, its leaders, and their policies \textit{Wrong}

• What Could be the Goal?
  1. Stop criticism of the state \textit{Wrong}
  2. Stop collective action \textit{Right}

• Implications: Social Media is Actionable!
  • Chinese leaders:
    • measure criticism: to judge local officials
The Goals of Censorship make Social Media Actionable

• **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*

• **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
The Goals of Censorship make Social Media Actionable

- **Everyone knows the Goal**: Stop criticism and protest about the state, its leaders, and their policies **Wrong**

- **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:
The Goals of Censorship make Social Media Actionable

- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*

- **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
The Goals of Censorship make Social Media Actionable

- **Everyone knows the Goal:**
  Stop criticism and protest about the state, its leaders, and their policies *Wrong*

- **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
    - Policies that generate dissent
The Goals of Censorship make Social Media Actionable

- Everyone knows the Goal: Stop criticism and protest about the state, its leaders, and their policies *Wrong*
- 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  
    - Policies that generate dissent  
    - Dissidents to be arrested; peace treaties to sign; emerging scandals
The Goals of Censorship make Social Media Actionable

- Everyone knows the Goal: Stop criticism and protest about the state, its leaders, and their policies *Wrong*
- 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
    - Policies that generate dissent
    - Dissidents to be arrested; peace treaties to sign; emerging scandals
    - Disagreements between central and local leaders
Classification of Events Generating Bursts of Social Media
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies
Classification of Events Generating Bursts of Social Media
Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies
Classification of Events Generating Bursts of Social Media

Each including $+$, $-$, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies
Classification of Events Generating Bursts of Social Media

Each including +, −, or neutral comments about the state

1. Collective Action Potential
2. Pornography
3. Criticism of censors
4. (Other) News
5. Government Policies
What Types of Events Are Censored?

![Graph showing the distribution of censorship magnitude and density across different event types.]

- Policy
- News
- Collective Action
- Criticism of Censors
- Pornography
Censoring Collective Action: Ai Weiwei’s Arrest

Ai Weiwei arrested
Censoring Collective Action: Riots in Zengcheng

![Graph showing count published vs count censored over months from January to July. The graph highlights the increase in censored content in June, labeled as "Riots in Zengcheng."](image)
Censoring Collective Action: Environmental Lottery Rally
Low Censorship on Policy: One Child

Speculation of Policy Reversal at NPC

Count Published

Count Censored
Low Censorship on News: Power Prices

Power shortages
Gov't raises power prices to curb demand

Count Published
Count Censored
How To Take Advantage of Big Analytics

- It's cheap and powerful; don't skimp!
- Off-the-shelf analytics ⇝ big advances
- Innovative analytics ⇝ immensely better than off-the-shelf
  (Much harder to hire for innovative analytics; so consider a mix of in-house hires and outside experts)

- Save it for 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!
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 ⟷ big advances
  • Innovative analytics ⟷ immensely better than off-the-shelf
How To Take Advantage of Big Analytics

- Its cheap and powerful; don’t skimp!
  - Off-the-shelf analytics $\rightsimeq$ big advances
  - Innovative analytics $\rightsimeq$ immensely better than off-the-shelf
  - (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)
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
  - (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)

- **Save it for first!**
How To Take Advantage of Big Analytics

- Its cheap and powerful; don’t skimp!
  - Off-the-shelf analytics $\leadsto$ big advances
  - Innovative analytics $\leadsto$ immensely better than off-the-shelf
  - (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)

- Save it for 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 $\leadsto$ big advances
  • Innovative analytics $\leadsto$ immensely better than off-the-shelf
    • (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)

• Save it for 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
  • (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)

• Save it for 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
  • (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)

• Save it for 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 ⟷ big advances
  • Innovative analytics ⟷ immensely better than off-the-shelf
  • (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)

• Save it for 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 $\leadsto$ big advances
  • Innovative analytics $\leadsto$ immensely better than off-the-shelf
  • (Much harder to hire for innovative analytics; so consider a mix of in house hires and outside experts)

• Save it for 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