

Big Data is Not About the Data!

Gary King¹

Institute for Quantitative Social Science
Harvard University

(Gov2001: Advanced Quantitative Research Methodology, 4/28/2013)

¹GaryKing.org

The *Data* In Big Data (about people)

The *Data* In Big Data (about people)

The Last 50 Years:

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 *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. . .

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

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 *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)

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

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 *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
- You are part of a tectonic movement: *The march of quantification* through academia, professions, government, & commerce (*SuperCrunchers*, *The Numerati*, *MoneyBall*)

The *Data* in Big Data: Examples

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Data* in Big Data: Examples

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

The *Value* in Big Data: the Analytics

The *Value* in Big Data: the Analytics

- Data:

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

The *Value* in Big Data: the Analytics

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

Examples of what's now possible

Examples of what's now possible

- Opinions of activists:

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:**

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: “How many times did you exercise last week?”

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: "How many times did you exercise last week?" $\rightsquigarrow$ 500K people carrying cell phones with accelerometers

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: "How many times did you exercise last week?" $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:**

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: "How many times did you exercise last week?" $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: "Please tell me your 5 best friends"

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: "How many times did you exercise last week?" $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: "Please tell me your 5 best friends" $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: “How many times did you exercise last week?” $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:**

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: “How many times did you exercise last week?” $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:** Dubious or nonexistent governmental statistics

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: “How many times did you exercise last week?” $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:** Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, road networks, other infrastructure

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: “How many times did you exercise last week?” $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:** Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, road networks, other infrastructure
- Many, **many**, more. . .

Examples of what's now possible

- **Opinions of activists:** A few thousand interviews $\rightsquigarrow$ billions of political opinions in social media posts (1B every 2 Days)
- **Exercise:** A survey: “How many times did you exercise last week?” $\rightsquigarrow$ 500K people carrying cell phones with accelerometers
- **Social contacts:** A survey: “Please tell me your 5 best friends” $\rightsquigarrow$ continuous record of phone calls, emails, text messages, bluetooth, social media connections, address books
- **Economic development in developing countries:** Dubious or nonexistent governmental statistics $\rightsquigarrow$ satellite images of human-generated light at night, road networks, other infrastructure
- Many, **many**, more. . .
- **In each: without new analytics, the data are useless**

The End of The Quantitative-Qualitative Divide

The End of The Quantitative-Qualitative Divide

- **Qualitative researchers:** overwhelmed by information; need help

The End of The Quantitative-Qualitative Divide

- **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 End of The Quantitative-Qualitative Divide

- **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 End of The Quantitative-Qualitative Divide

- **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 Spectacular Success of Quantitative Social Science

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling the human life span in the last century?
- $\vdots$
- Quantitative social science

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling the human life span in the last century?
- $\vdots$
- Quantitative social science (aka “big data,” “data analytics,” “data science”):

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling the human life span in the last century?
- $\vdots$
- Quantitative social science (aka “big data,” “data analytics,” “data science”): transformed most Fortune 500 firms

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

- The genetics revolution?
- The Higgs-like particle?
- Exoplanets? The Mars rovers?
- Doubling the human life span in the last century?
- $\vdots$
- Quantitative social science (aka “big data,” “data analytics,” “data science”): transformed most Fortune 500 firms; established new industries

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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

The Spectacular Success of Quantitative Social Science

What university research has had the biggest impact on you?

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

⋮

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