The Large Problem of Big Data:
Finding solutions to practical problems by combining Mathematics, Statistics, and Computer Science

Michael J. Higgins

mikehiggins@ksu.edu

Kansas State University

Oct. 09, 2017

A little about myself

- B.S. in Mathematics and Statistics from KSU in 2006.

A little about myself

- B.S. in Mathematics and Statistics from KSU in 2006.
- PhD in Statistics from UC Berkeley in 2013.

A little about myself

- B.S. in Mathematics and Statistics from KSU in 2006.
- PhD in Statistics from UC Berkeley in 2013.
- Post doc in Department of Politics at Princeton University.

A little about myself

- B.S. in Mathematics and Statistics from KSU in 2006.
- PhD in Statistics from UC Berkeley in 2013.
- Post doc in Department of Politics at Princeton University.
- Started as Asst. Prof. in Department of Statistics at KSU in 2015.

Big Data?

- What is Big Data?

Big Data?

- What is Big Data?
- Collecting information has never been cheaper.
- Ubiquity of *microdata*—disaggregated data—from increased ability to collect and store information.

Examples of Microdata

Retail:

Macro: Number of purchases per item per day.

Examples of Microdata

Retail:

Macro: Number of purchases per item per day.

Micro: What items each person purchases.
What day, hour, minute, second item was purchased
Where items are located, proximity to other items.

Examples of Microdata

Retail:

Macro: Number of purchases per item per day.

Micro: What items each person purchases.
What day, hour, minute, second item was purchased
Where items are located, proximity to other items.

Legislation:

Macro: Topic of bill, who introduced bill, on what day
Roll call votes on bill

Examples of Microdata

Retail:

Macro: Number of purchases per item per day.

Micro: What items each person purchases.
What day, hour, minute, second item was purchased
Where items are located, proximity to other items.

Legislation:

Macro: Topic of bill, who introduced bill, on what day
Roll call votes on bill

Micro: Wording of bill
What legislators say about the bill in a debate

Why microdata?

- Idea: Microdata contains details on how people or systems behave.

  Possible to gain much more insight on research questions

Why microdata?

- Idea: Microdata contains details on how people or systems behave.

  Possible to gain much more insight on research questions

- Problems: People behave in complicated ways. How can we model this additional complexity?

Why microdata?

- Idea: Microdata contains details on how people or systems behave.

  Possible to gain much more insight on research questions

- Problems: People behave in complicated ways. How can we model this additional complexity?

  What research questions should we now ask?

Why microdata?

- Idea: Microdata contains details on how people or systems behave.

  Possible to gain much more insight on research questions

- Problems: People behave in complicated ways. How can we model this additional complexity?

  What research questions should we now ask?

- Massive amounts of data! Provides more insight but may be "noisier."

  Methods of analysis need to be effective, but also efficient.

Research questions are changing

Retail:

Old Q: Are people more likely to buy certain items on specific days of the week?

How to order products to keep items in stock without being overstocked.

Research questions are changing

Retail:

Old Q: Are people more likely to buy certain items on specific days of the week?

How to order products to keep items in stock without being overstocked.

New Q: Given that a customer purchases certain items, what items are that customer likely to purchase next?

How to target advertising to make customer more likely to buy those products at store.

Research questions are changing

Retail:

Old Q: Are people more likely to buy certain items on specific days of the week?

How to order products to keep items in stock without being overstocked.

New Q: Given that a customer purchases certain items, what items are that customer likely to purchase next?

How to target advertising to make customer more likely to buy those products at store.

Example: Target predicting pregnancy of customer in 2012

Massive data questions

Google searches:

- 4.5 billion searches every day
- Given the search history of an individual, can we find most relevant search results, even if there are typos in the search query?

Massive data questions

Google searches:

- 4.5 billion searches every day
- Given the search history of an individual, can we find most relevant search results, even if there are typos in the search query?
- Multi-billion dollar question: Given a query and the search history of an individual, how to display advertisements to maximize ad clicks and purchases from sponsor?

Massive data questions

Google searches:

- 4.5 billion searches every day
- Given the search history of an individual, can we find most relevant search results, even if there are typos in the search query?
- Multi-billion dollar question: Given a query and the search history of an individual, how to display advertisements to maximize ad clicks and purchases from sponsor?
- Solution may involve more than past queries:
  E.g. click history, time spent on a page, location of ad on the page.
- A/B experiments help determine best settings.
  Performed on small percentage of all queries: Billions of observations.

Massive data questions

Facebook:

- 1.5 billion users log in each month.
- Additional "social network" structure: Information about friends of users, membership of groups, "liking" of posts, etc.

Massive data questions

Facebook:

- 1.5 billion users log in each month.
- Additional "social network" structure: Information about friends of users, membership of groups, "liking" of posts, etc.
- Likelihood of "taking action"—e.g. clicking on an ad—depends not only on user but the network of user.

Massive data questions

Facebook:

- 1.5 billion users log in each month.
- Additional "social network" structure: Information about friends of users, membership of groups, "liking" of posts, etc.
- Likelihood of "taking action"—e.g. clicking on an ad—depends not only on user but the network of user.
- Example: "A 61-million-person experiment in social influence and political mobilization." (Bond et. al.)

Problems with massive data

- With big data, computational power is limiting.

Problems with massive data

- With big data, computational power is limiting.
- E.g. many times, it's advantageous to compute a measure between every pair of data points:
- A billion data points $\Longrightarrow$ Roughly 1 quintillion operations.
  Would take the world's fastest computer at least 30 seconds.
  Much longer (or impossible) for most computers.

Problems with massive data

- With big data, computational power is limiting.
- E.g. many times, it's advantageous to compute a measure between every pair of data points:
- A billion data points $\Longrightarrow$ Roughly 1 quintillion operations.
  Would take the world's fastest computer at least 30 seconds.
  Much longer (or impossible) for most computers.
- New emphasis on fast statistical techniques instead of those that lead to "best tests."

Evolution of statistical research

- Big data has forced the discipline of statistics to evolve.

Evolution of statistical research

- Big data has forced the discipline of statistics to evolve.
- Traditionally: statistics has focused on experimental design, survey sampling, modeling and estimation, and statistical inference on population parameters.

  E.g. Gallop poll: Sample 1,000 people, predict next president

Evolution of statistical research

- Big data has forced the discipline of statistics to evolve.
- Traditionally: statistics has focused on experimental design, survey sampling, modeling and estimation, and statistical inference on population parameters.

  E.g. Gallop poll: Sample 1,000 people, predict next president
- Many methods assume a handful of variables of interest and either rely on large sample approximations or adjusts for small samples.

Evolution of statistical research

- Big data: Many data points—large sample theory is irrelevant.

Evolution of statistical research

- Big data: Many data points—large sample theory is irrelevant.
- Huge number of variables:

  Text: Each possible word or phrase is a variable.
  Genetics: Each gene is a variable ($\approx$ 25,000).
- Emergence of "data mining" techniques to find relevant variables and find "signal in the noise."

Evolution of statistical research

- Big data: Many data points—large sample theory is irrelevant.
- Huge number of variables:

  Text: Each possible word or phrase is a variable.
  Genetics: Each gene is a variable ($\approx$ 25,000).
- Emergence of "data mining" techniques to find relevant variables and find "signal in the noise."
- Computational complexity: Techniques to analyze data quickly while preserving effectiveness of methods.
- Increased emphasis of optimization, numerical approximation, etc.

  Many ideas from Computer Science: e.g. "Machine Learning"

The demand for statisticians

Big demand of big data = Big career opportunities.

The demand for statisticians

Big demand of big data = Big career opportunities.

- The burst of big data has created a tremendous demand for people with the ability to manipulate and analyze this data.

  Data Scientist: "The sexiest career of the 21st century"

- Supply <<<< Demand

The demand for statisticians

Big demand of big data = Big career opportunities.

- The burst of big data has created a tremendous demand for people with the ability to manipulate and analyze this data.

  Data Scientist: "The sexiest career of the 21st century"

- Supply $<<<<$ Demand
- Big data = Big dollars. Tremendous career opportunities for students who learn skills to tackle big data.

What do you need to become a "Data Scientist"?

Knowledge of:

What do you need to become a "Data Scientist"?

Knowledge of:

- Computational techniques and algorithms

What do you need to become a "Data Scientist"?

Knowledge of:

- Computational techniques and algorithms
- Mathematics and statistics

What do you need to become a "Data Scientist"?

Knowledge of:

- Computational techniques and algorithms
- Mathematics and statistics
- Most importantly: Substantive field of application

First necessary skill: Knowledge of application

- Statistics is driven by uncertainty found from data in substantive field
- Data is not enough! Need field-specific knowledge

  What data should be collected? What data would best answer research question?

  How can big data improve current methods?

First necessary skill: Knowledge of application

- Statistics is driven by uncertainty found from data in substantive field
- Data is not enough! Need field-specific knowledge

  What data should be collected? What data would best answer research question?

  How can big data improve current methods?
- Example: Why do legislators vote the way they do?

  Ideally want to know the thought process of a legislator.

First necessary skill: Knowledge of application

- Statistics is driven by uncertainty found from data in substantive field
- Data is not enough! Need field-specific knowledge

  What data should be collected? What data would best answer research question?

  How can big data improve current methods?
- Example: Why do legislators vote the way they do?

  Ideally want to know the thought process of a legislator.
- Macro: Use topics of bills and roll call votes.
  Micro: Use text from debates about bills.

Second necessary skill: Computation

- Big data requires big computation—even loading and manipulating data may require sophisticated skills.
- Strong Computer Science background is absolutely necessary.

Second necessary skill: Computation

- Big data requires big computation—even loading and manipulating data may require sophisticated skills.
- Strong Computer Science background is absolutely necessary.
- Improved algorithms can reduce computation time of a statistical procedure by hours, days, years, etc.
- Data analysis requires efficient access to data

Second necessary skill: Computation

- Big data requires big computation—even loading and manipulating data may require sophisticated skills.
- Strong Computer Science background is absolutely necessary.
- Improved algorithms can reduce computation time of a statistical procedure by hours, days, years, etc.
- Data analysis requires efficient access to data
- Field-specific knowledge avoids blindly programming

Computation requirements

- Strong background in algorithms and optimization/approximation
- Knowledge of a statistical programming language (e.g. R) and other high-performance languages (e.g. C/C++, Python).
- Work with databases (e.g. SQL) a plus.

Third necessary skill: Mathematical/statistical intuition

- Mathematical/statistical intuition can improve efficacy of statistical methods and may motivate new methods.
- Best design for a study on a research question?
- Which method for analyzing data is best?

Third necessary skill: Mathematical/statistical intuition

- Mathematical/statistical intuition can improve efficacy of statistical methods and may motivate new methods.
- Best design for a study on a research question?
- Which method for analyzing data is best?
- Many unanswered questions:

  For many commonly used methods: What do we need to assume to ensure they work well?

  Computation vs. power and precision tradeoff

Third necessary skill: Mathematical/statistical intuition

- Mathematical/statistical intuition can improve efficacy of statistical methods and may motivate new methods.
- Best design for a study on a research question?
- Which method for analyzing data is best?
- Many unanswered questions:

  For many commonly used methods: What do we need to assume to ensure they work well?

  Computation vs. power and precision tradeoff
- Exploit links between well-solved math problems and real-life data problems.

  Example: Social network data $\Longleftrightarrow$ Graph Theory

Mathematics/Statistics requirements

- Strong knowledge of Mathematical Statistics and Linear Algebra. Mathematical analysis is a plus.
- Build a Statistics toolbelt: Courses in Machine Learning, High-Dimensional Data methods, Spatial Statistics, Bayesian Statistics, Sampling, Experimental Design
- More math can only help

How much education is necessary to get a job?

- Data Scientist jobs possible with undergraduate degree and solid background in Mathematics, Statistics, and Computer Science, preferably with field-specific knowledge.
- Most have at least a Masters degree: Popular choices include Statistics or Computer Science.
- PhD $\Longrightarrow$ Higher starting salary, but is not required for most jobs.

How much education is necessary to get a job?

- Data Scientist jobs possible with undergraduate degree and solid background in Mathematics, Statistics, and Computer Science, preferably with field-specific knowledge.
- Most have at least a Masters degree: Popular choices include Statistics or Computer Science.
- PhD $\Longrightarrow$ Higher starting salary, but is not required for most jobs.
- Median salary for Data Scientist—$112,000.

Conclusion

- Big data boom has reshaped the field of Statistics.
- Microdata has potential for understanding field-specific phenomena in great detail

  Tremendous demand for new methods—and more manpower—to make good use of big data.

Conclusion

- Big data boom has reshaped the field of Statistics.
- Microdata has potential for understanding field-specific phenomena in great detail

  Tremendous demand for new methods—and more manpower—to make good use of big data.
- Combination of Mathematics, Statistics, and Computer Science is necessary to tackle big data problems

Conclusion

- Big data boom has reshaped the field of Statistics.
- Microdata has potential for understanding field-specific phenomena in great detail

  Tremendous demand for new methods—and more manpower—to make good use of big data.
- Combination of Mathematics, Statistics, and Computer Science is necessary to tackle big data problems
- You have the opportunity *NOW* to build this skill set and begin to tackle these challenges.

Conclusion

- Big data boom has reshaped the field of Statistics.
- Microdata has potential for understanding field-specific phenomena in great detail

  Tremendous demand for new methods—and more manpower—to make good use of big data.
- Combination of Mathematics, Statistics, and Computer Science is necessary to tackle big data problems
- You have the opportunity *NOW* to build this skill set and begin to tackle these challenges.

$$\text{Supply} \; <<<< \; \text{Demand} \quad \Longrightarrow \quad \text{Skill set} \; = \$\$\$\$$$

Thank you