How Big Data is Changing Economies

Larry Wasserman
Carnegie Mellon University
How Big Data is Changing Economies

Big Data: A Statistician’s Perspective

or

Big Data + Bad Analysis = Bad Decisions

Larry Wasserman
Carnegie Mellon University
www.stat.cmu.edu/~larry
WHO AM I?

• I am a professor at Carnegie Mellon.

• My main appointment is in the Department of Statistics.

• I also have an appointment in the Machine Learning Department in the School of Computer Science.

• I work on: statistical theory, machine learning, astrostatistics, biology
Main Points

• Statisticians are being left out
• This should worry everyone (not just statisticians)
• Big Data + Complex Models = Small Data
• The Future
Main Points

- Statisticians are being left out
Main Points

- Statisticians are being left out
- This should worry everyone (not just statisticians)
Main Points

• Statisticians are being left out

• This should worry everyone (not just statisticians)

• Big Data + Complex Models = Small Data
Main Points

- Statisticians are being left out
- This should worry everyone (not just statisticians)
- Big Data + Complex Models = Small Data
- The Future
Where are the Statisticians?

- President’s Council of Advisors on Science and Technology (PCAST) includes ... 0 statisticians!

- Chief Data Scientist of the United States Office of Science and Technology Policy. Not a statistician.

- Forbes: World’s 7 Most Powerful Data Scientists (0 statisticians).

- I have seen many Big Data/Data Science initiatives that include no statisticians.

- Why is this? Why should you care?
Where are the Statisticians?

- President’s Council of Advisors on Science and Technology (PCAST) includes ... 0 statisticians!
- Chief Data Scientist of the United States Office of Science and Technology Policy. Not a statistician.
Where are the Statisticians?

- President’s Council of Advisors on Science and Technology (PCAST) includes ... 0 statisticians!
- Chief Data Scientist of the United States Office of Science and Technology Policy. Not a statistician.
- Forbes: World’s 7 Most Powerful Data Scientists (0 statisticians).
Where are the Statisticians?

- President’s Council of Advisors on Science and Technology (PCAST) includes ... 0 statisticians!
- Chief Data Scientist of the United States Office of Science and Technology Policy. Not a statistician.
- Forbes: World’s 7 Most Powerful Data Scientists (0 statisticians).
- I have seen many Big Data/Data Science initiatives that include no statisticians.
Where are the Statisticians?

- President’s Council of Advisors on Science and Technology (PCAST) includes ... 0 statisticians!
- Chief Data Scientist of the United States Office of Science and Technology Policy. Not a statistician.
- Forbes: World’s 7 Most Powerful Data Scientists (0 statisticians).
- I have seen many Big Data/Data Science initiatives that include no statisticians.
- Why is this? Why should you care?
Why Should You Care?

- Big Data + Bad Analysis = Bad Decisions
Why Should You Care?

- Big Data + Bad Analysis = Bad Decisions

- Gary King: Big data is not about the data, it’s about the analytics.
Why Should You Care?

- Big Data + Bad Analysis = Bad Decisions

- Gary King: Big data is not about the data, it’s about the analytics.

- Google search: big data bad analytics = 10,700,000 hits
Why Should You Care?

- Big Data + Bad Analysis = Bad Decisions
- Gary King: Big data is not about the data, it’s about the analytics.
- Google search: big data bad analytics = 10,700,000 hits
- Statisticians have been doing data science for at least 100 years.
Why Should You Care?

- Big Data + Bad Analysis = Bad Decisions

- Gary King: Big data is not about the data, it’s about the analytics.

- Google search: big data bad analytics = 10,700,000 hits

- Statisticians have been doing data science for at least 100 years.

- You would not get brain surgery done by a cardiologist.
Interlude: The Four Questions

(1) Big Data and the Economy: ????
(2) Big Data and Theory: What Statistical Methods Apply?

- all methods apply but: all methods have: bias + variance
- Big Data reduces variance. It has not effect on bias (possibly negative effect)
- a non-identifiable model is non-identifiable even with infinite data
- Big Data can be small (more later if I have time)
- what to study? standard statistical theory, nonparametrics, distribution free methods, optimization, **online methods**
(3) Differential Privacy:

- good idea.
- Doesn’t work.
- Need to add huge amounts of noise.
- Query-response model = bogus

(4) Implementation:

- develop streaming versions of statistical, ML methods.
- Distributed approaches.
- Nonparametric/distribution free (don’t assume linear model)
Statistical Issues

• sampling bias
Statistical Issues

- sampling bias
- correlation is not causation (oldie but goodie)
Statistical Issues

- sampling bias
- correlation is not causation (oldie but goodie)
- effects of mining the data (seek and ye shall find)
Statistical Issues

- sampling bias
- correlation is not causation (oldie but goodie)
- effects of mining the data (seek and ye shall find)
- rigor: what assumptions are you making? what is the best you can do under those assumptions?
Why Are Statisticians Left Out?

Statisticians are:

- conservative
- stubborn
- inflexible
- bad at selling themselves
- afraid
- experts at saying what you can't do
Why Are Statisticians Left Out?

Statisticians are:

- conservative
- stubborn
- inflexible
- bad at selling themselves
- afraid
- experts at saying what you can't do
Why Are Statisticians Left Out?

Statisticians are:

- conservative
- stubborn
Why Are Statisticians Left Out?

Statisticians are:

conservative
stubborn
inflexible
Why Are Statisticians Left Out?

Statisticians are:

- conservative
- stubborn
- inflexible
- bad at selling themselves
Why Are Statisticians Left Out?

Statisticians are:

- conservative
- stubborn
- inflexible
- bad at selling themselves
- afraid
Why Are Statisticians Left Out?

Statisticians are:

- conservative
- stubborn
- inflexible
- bad at selling themselves
- afraid
- experts at saying what you can’t do
A True Story

- Astronomer asks us for help.
A True Story

- Astronomer asks us for help.
- We spend months learning the science, cleaning the data and carefully analyzing the data.
A True Story

- Astronomer asks us for help.
- We spend months learning the science, cleaning the data and carefully analyzing the data.
- Some careful, modest results after one year.
A True Story

- Astronomer asks us for help.
- We spend months learning the science, cleaning the data and carefully analyzing the data.
- Some careful, modest results after one year.
- In the meantime...
A True Story

- Astronomer asks us for help.
- We spend months learning the science, cleaning the data and carefully analyzing the data.
- Some careful, modest results after one year.
- In the meantime...
  ... my astronomer friend went to see my friends in ML.
A True Story

- Astronomer asks us for help.
- We spend months learning the science, cleaning the data and carefully analyzing the data.
- Some careful, modest results after one year.
- In the meantime...
  ... my astronomer friend went to see my friends in ML.
- Two days later the ML people produced fancy plots, analyses etc.
A True Story

- Astronomer asks us for help.
- We spend months learning the science, cleaning the data and carefully analyzing the data.
- Some careful, modest results after one year.
- In the meantime...
  - my astronomer friend went to see my friends in ML.
- Two days later the ML people produced fancy plots, analyses etc.
- We complain that their analysis was not rigorous.
A True Story

- Astronomer asks us for help.
- We spend months learning the science, cleaning the data and carefully analyzing the data.
- Some careful, modest results after one year.
- In the meantime...
  ... my astronomer friend went to see my friends in ML.
- Two days later the ML people produced fancy plots, analyses etc.
- We complain that their analysis was not rigorous.
- Who will the astronomer go to in the future?
What to Do?

**Statisticians**

- Statisticians need to be more nimble and flexible

Users/Consumers/Data Scientists: need to be aware that:

- Careful analysis matters.
- Blindly running fancy algorithms on big data does not always lead to good outcomes.
- Invite statisticians to panels so we can **communicate** to reach out (thanks!)
What to Do?

**Statisticians**

- Statisticians need to be more nimble and flexible
- Statisticians need a Carl Sagan.
What to Do?

**Statisticians**

- Statisticians need to be more nimble and flexible
- Statisticians need a Carl Sagan.
- Statisticians need to take some chances.
What to Do?

**Statisticians**

- Statisticians need to be more nimble and flexible
- Statisticians need a Carl Sagan.
- Statisticians need to take some chances.

BUT ...
What to Do?

**Statisticians**

- Statisticians need to be more nimble and flexible
- Statisticians need a Carl Sagan.
- Statisticians need to take some chances.

**BUT ...**

- **Users/Consumers/Data Scientists**: need to be aware that:
  - Careful analysis matters.
  - Blindly running fancy algorithms on big data does not always lead to good outcomes.
  - Invite statisticians to panels so we can more fully reach out (thanks!)
What to Do?

**Statisticians**
- Statisticians need to be more nimble and flexible.
- Statisticians need a Carl Sagan.
- Statisticians need to take some chances.

**BUT ...**
- **Users/Consumers/Data Scientists:** need to be aware that:
  - Careful analysis matters.
What to Do?

**Statisticians**

- Statisticians need to be more nimble and flexible
- Statisticians need a Carl Sagan.
- Statisticians need to take some chances.

**BUT ...**

- **Users/Consumers/Data Scientists**: need to be aware that:
  - Careful analysis matters.
  - Blindly running fancy algorithms on big data does not always lead to good outcomes.
What to Do?

Statisticians
- Statisticians need to be more nimble and flexible
- Statisticians need a Carl Sagan.
- Statisticians need to take some chances.

BUT ...
- **Users/Consumers/Data Scientists**: need to be aware that:
- Careful analysis matters.
- Blindly running fancy algorithms on big data does not always lead to good outcomes.
- Invite statisticians to panels so we can reach out (thanks!)
Big Data Sets Are Not Necessarily Big

Big Data $\rightarrow$ Complex/Numerous Questions $\rightarrow$ Small Effective Sample Size
Big Data Sets Are Not Necessarily Big

Big Data $\rightarrow$ Complex/Numerous Questions $\rightarrow$ Small Effective Sample Size

- Old Days (Paleozoic era): $n = 100$ people. Measures $d = 5$ different things on each person.

\[ \frac{100}{5} = 20 \text{ observations per parameter. Good.} \]
Big Data Sets Are Not Necessarily Big

Big Data $\rightarrow$ Complex/Numerous Questions $\rightarrow$ Small Effective Sample Size

- Old Days (Paleozoic era): $n = 100$ people. Measures $d = 5$ different things on each person.
  $\frac{100}{5} = 20$ observations per parameter. Good.

- More recently (Mesozoic era): $n = 100$ people. Measures $d = 5000$ genes per person.
  $\frac{100}{5000} = .02$ observations per parameter. Bad.
Big Data Sets Are Not Necessarily Big

Big Data $\rightarrow$ Complex/Numerous Questions $\rightarrow$ Small Effective Sample Size

- **Old Days (Paleozoic era):** $n = 100$ people. Measures $d = 5$ different things on each person.

  \[
  \frac{100}{5} = 20 \text{ observations per parameter. Good.}
  \]

- **More recently (Mesozoic era):** $n = 100$ people. Measures $d = 5000$ genes per person.

  \[
  \frac{100}{5000} = .02 \text{ observations per parameter. Bad.}
  \]

- **Big Data comes along: (Cenozoic era)** $n = 100,000$ people. Measures $d = 5000$ things.

  \[
  \frac{100,000}{5000} = 20 \text{ observations per parameter. We're good again.}
  \]
Bigger Data Leads to Bigger Questions

- Predict disease from $d$ genes. $X_1, \ldots, X_d$. 

- With Big Data we can ask harder questions: also include interactions into the predictions:
  - two-way interactions $X_j X_k$
  - three-way interactions $X_j X_k X_\ell$ etc.

- $10^3, 10^6 = 20$ observations per parameter but $10^3, 10^6 = 0.00000001$ observations per parameter.

- Solution: need statisticians.
Bigger Data Leads to Bigger Questions

• Predict disease from $d$ genes. $X_1, \ldots, X_d$.

• With Big Data we can ask harder questions: also include interactions into the predictions:
Bigger Data Leads to Bigger Questions

- Predict disease from $d$ genes. $X_1, \ldots, X_d$.
- With Big Data we can ask harder questions: also include interactions into the predictions:
  - two-way interactions $X_j X_k$
  - three-way interactions $X_j X_k X_\ell$ etc.

100,000 5000 = 20 observations per parameter

but 100,000 5000 3 = 0.00000001 observations per parameter

- Solution: need statisticians
Bigger Data Leads to Bigger Questions

- Predict disease from $d$ genes. $X_1, \ldots, X_d$.
- With Big Data we can ask harder questions: also include interactions into the predictions:
  - two-way interactions $X_j X_k$
  - three-way interactions $X_j X_k X_\ell$ etc.

$100,000 = 20$ observations per parameter
$\frac{100,000}{5000} = 0.00000001$ observations per parameter

Solution: need statisticians
Bigger Data Leads to Bigger Questions

• Predict disease from $d$ genes. $X_1, \ldots, X_d$.

• With Big Data we can ask harder questions: also include interactions into the predictions:
  
  • two-way interactions $X_j X_k$
  
  • three-way interactions $X_j X_k X_\ell$ etc.

\[
\frac{100,000}{5000} = 20 \text{ observations per parameter}
\]
Bigger Data Leads to Bigger Questions

- Predict disease from $d$ genes. $X_1, \ldots, X_d$.
- With Big Data we can ask harder questions: also include interactions into the predictions:
  - two-way interactions $X_j X_k$
  - three-way interactions $X_j X_k X_\ell$ etc.

\[
\frac{100,000}{5000} = 20 \text{ observations per parameter}
\]

but
Bigger Data Leads to Bigger Questions

- Predict disease from $d$ genes. $X_1, \ldots, X_d$.
- With Big Data we can ask harder questions: also include interactions into the predictions:
  - two-way interactions $X_jX_k$
  - three-way interactions $X_jX_kX_\ell$ etc.

$$\frac{100,000}{5000} = 20 \text{ observations per parameter}$$

but

$$\frac{100,000}{5000^3} = 0.00000001 \text{ observations per parameter}$$
Bigger Data Leads to Bigger Questions

• Predict disease from $d$ genes. $X_1, \ldots, X_d$.

• With Big Data we can ask harder questions: also include interactions into the predictions:

  • two-way interactions $X_j X_k$
  • three-way interactions $X_j X_k X_\ell$ etc.

\[
\frac{100,000}{5000} = 20 \text{ observations per parameter}
\]

but

\[
\frac{100,000}{5000^3} = 0.000000001 \text{ observations per parameter}
\]

• Solution: need statisticians
Conclusion: The Future

- “Big Data” is one of the most exciting things to happen.
Conclusion: The Future

- “Big Data” is one of the most exciting things to happen.
- Statisticians should be more assertive.
Conclusion: The Future

- "Big Data" is one of the most exciting things to happen.
- Statisticians should be more assertive.
- Big Data users should talk to statisticians.
“Big Data” is one of the most exciting things to happen.

Statisticians should be more assertive.

Big Data users should talk to statisticians.

Statistics departments are poor. We need money!
Conclusion: The Future

- “Big Data” is one of the most exciting things to happen.
- Statisticians should be more assertive.
- Big Data users should talk to statisticians.
- Statistics departments are poor. We need money!
- Big Datasets are not magic: drawing conclusions requires assumptions and careful analysis
Conclusion: The Future

- “Big Data” is one of the most exciting things to happen.
- Statisticians should be more assertive.
- Big Data users should talk to statisticians.
- Statistics departments are poor. We need money!
- Big Datasets are not magic: drawing conclusions requires assumptions and careful analysis.
- Big Data methods (data science) work best when we work in teams: statisticians + computer scientists + economists + ...
THE END