Statistics and Big Data

Marie Davidian

Department of Statistics
North Carolina State University

http://www4.stat.ncsu.edu/~davidian
Aren’t We Data Science?

Last month, I shared this column with President-elect Nat Schenker and Past President Bob Rodriguez to announce an ASA strategic initiative to promote engagement of statisticians in Big Data (http://magazine.amstat.org/blog/2013/06/01/the-asa-and-big-data). I’m following that announcement with an account of some of my recent experiences regarding data science, which inspire my enthusiasm for this effort. One in particular serves as a metaphor for the disconnect between statistics and data science we noted last month.

Around the time we were finalizing that column, President Barack Obama launched the National Center for Data Science (NCDS) (aiming to make North Carolina a national hub for data-intensive business and data science research.” It went on to note that the NCDS had been launched at the Renaissance Computing Institute at The University of North Carolina at Chapel Hill (UNC-CH) and included among its founding members businesses, government organizations, and major research universities.

I highlight that last group because, upon locating the NCDS website (http://data2discovery.org), I was astonished to review the list of founding members and see that not only is my university (North Carolina State) a founding member, but so are...
Thanks to Terry Speed
Data

• **Big Data**
  - Massive # of records
  - Huge dimension
  - Intensive
  - All of the above

• Hysteria and panic =⇒ rational recognition that data can improve, inform, enhance just about everything

• Medium and Small Data
  - Still dominate
  - Clinical trials, regulatory
  - Industrial/agricultural/laboratory/etc experimentation
  - Surveys
  - . . .

TPSEMath 2015
• **Big Data**
  - Massive # of records
  - Huge dimension
  - Intensive
  - All of the above
  - “Nontraditional,” e.g., unstructured text

• **Hysteria** and **panic** $\implies$ **rational recognition** that data can improve, inform, enhance just about **everything**
• **Big Data**
  ▶ Massive # of records
  ▶ Huge dimension
  ▶ Intensive
  ▶ All of the above
  ▶ “**Nontraditional**,” e.g., unstructured text

• **Hysteria and panic → rational recognition** that data can improve, inform, enhance just about *everything*

• **Medium and Small Data**
  ▶ Still dominate
  ▶ Clinical trials, regulatory
  ▶ Industrial/agricultural/laboratory/etc experimentation
  ▶ Surveys
  ▶ ...
Data science:

- **2013**: Blind-sided; statistical principles, contributions ignored
Data science:

- **2013**: Blind-sided; statistical principles, contributions ignored
- **2015**: Statistics as a core discipline (uncertainty, inference, confounding, ...)

The new normal:

- Big Data will just be data
Data science:

- **2013**: Blind-sided; statistical principles, contributions ignored
- **2015**: Statistics as a core discipline (uncertainty, inference, confounding, ...)
- [http://www.mastersindatascience.org/](http://www.mastersindatascience.org/)
Data science:

- **2013**: Blind-sided; statistical principles, contributions ignored
- **2015**: Statistics as a core discipline (uncertainty, inference, confounding, ...)
- [http://www.mastersindatascience.org/](http://www.mastersindatascience.org/)

The new normal:

- Big Data will just be *data*
What should undergraduate/graduate statistics training look like?

- Traditional training: Foundations (probability, inference, methods), statistical software
- Messy data/data cleaning
- Data visualization/exploration
- Empirical methods
- Confounding/missingness/mismeasuredness/etc
- Programming/computation/databases
- Team projects, open-end problem-solving, presentations
- Engaging with other disciplines
- Communication/leadership/collaborative skills
- Challenge: Integrate/interweave
Statistical education

What should undergraduate/graduate statistics training look like?

- *Traditional training*: Foundations (probability, inference, methods), statistical software
Statistical education

What should undergraduate/graduate statistics training look like?

• Traditional training: Foundations (probability, inference, methods), statistical software
• Messy data/data cleaning
• Data visualization/exploration
• Empirical methods
• Confounding/missingess/mismeasuredness/etc
• Programming/computation/databases
• Team projects, open-end problem-solving, presentations
• Engaging with other disciplines
• Communication/leadership/collaborative skills
Statistical education

What should undergraduate/graduate statistics training look like?

- **Traditional training**: Foundations (probability, *inference*, methods), statistical software
- Messy data/data cleaning
- Data visualization/exploration
- Empirical methods
- Confounding/missingess/mismeasuredness/etc
- Programming/computation/databases
- Team projects, open-end problem-solving, presentations
- Engaging with other disciplines
- Communication/leadership/collaborative skills
- **Challenge**: Integrate/interweave
Implications for faculty

Disconnect:

- Research and training are intertwined

- More "out-of-the-mainstream" research
  - Risky, application-driven/focused, novel collaborations
  - Data procurement/preparation
  - Software, data products, blogs, . . .
  - Open access repositories

- Traditional reward culture
  (peer-reviewed publications, citations, grants, . . .)
Implications for faculty

Disconnect:

- Research and training are intertwined
- More “out-of-the-mainstream” research

Risky, application-driven/focused, novel collaborations

Data procurement/preparation

Software, data products, blogs,...

Open access repositories

Traditional reward culture (peer-reviewed publications, citations, grants,...)
Implications for faculty

Disconnect:

- Research and training are intertwined
- More “out-of-the-mainstream” research
  - Risky, application-driven/focused, novel collaborations
  - Data procurement/preparation
  - Software, data products, blogs,…
  - Open access repositories
Implications for faculty

Disconnect:

- Research and training are intertwined
- More “out-of-the-mainstream” research
  - Risky, application-driven/focused, novel collaborations
  - Data procurement/preparation
  - Software, data products, blogs,…
  - Open access repositories
- Traditional reward culture (peer-reviewed publications, citations, grants,…)

TPSEMath 2015