NTNU

Norwegian University of Science and Technology

Data science vs. statistics: two cultures?

Iain Carmichael & J. S. Marron

Overview

- Terminology
- The stand taken in the article
- Critiques of statistics
- What is Data Science?
- Principle components of data science
- Going forward
- Conclusion

- Greater statistics = everything related to learning from data
- Lesser statistics = mathematical techniques performed by professors in statistics departments
  - Closer to the view of the general public

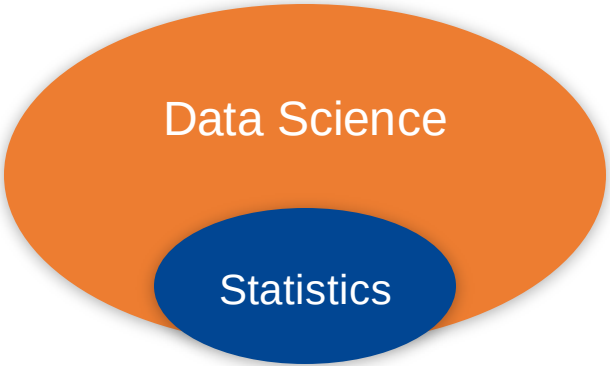
- Lack of common ground is often the root for disagreements

The stand taken in the article

Statistics = Lesser statistics

Data Science ≈ Greater statistics

"…providing value to society by broadening the discipline in technical ways"



Critiques of statistics

- is summarized as "too much theory, not enough computation"
- Focused on pre-computer problems
  - e.g. Poor programming skills among undergrades
- Hypothesis testing in Statistics 101

Redemption

- The need for theortical thinking is greater than ever
- R-Project, computational statistics and ggplot
- A way of thinking, *scientific method*

What is data science?

- Solving vs. understanding
- "[results in] data science should be judged by the extent to which they enable the analyst to learn from data..."

Data science: broader view of statistics

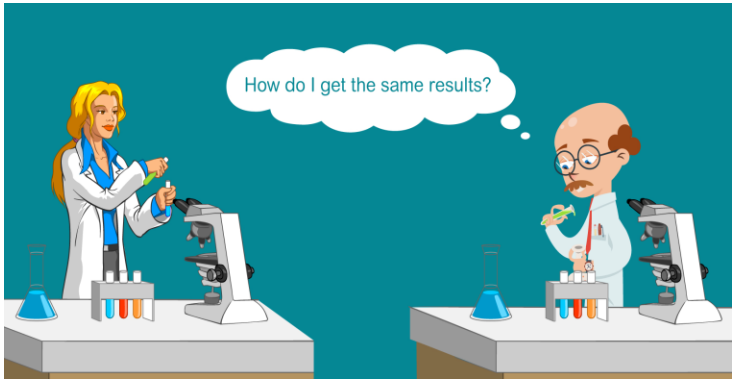
Greater statistics 1-6:

1. Data gathering, preparation, and exploration
2. Data representation and transformation
3. Computing with data
4. Data modeling
5. Data visualization and presentation
6. Science about data science

Reproducibility

- "[a]n article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship."

- Literate programming – improved our ability to do reproducible science



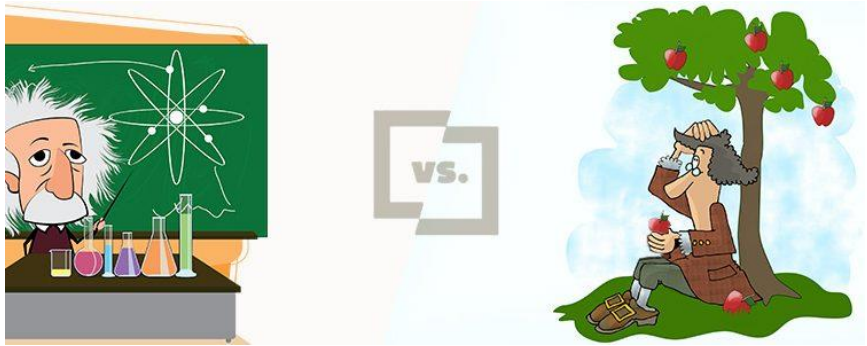
Do vs. understand

- Engineering is the business of creating a thing that does something
- Focus on predictive results
- Why computation have become more popular in recent years

- Science is the business of understanding how something works
- Focus on interpretability, understanding of model and what the data look like
- XAI

Theoretically vs. Empirically driven

- "Data science is exploratory data analysis gone mad"
- Exploratory data analysis (EDA) vs. confirmatory analysis
- "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" - end of confirmatory analysis



Problem first vs. hammer looking for a nail

The 80/20 rule

- More emphasize on previously undervalued areas: data visualization, exploratory data analysis, data mining, programming, data storage/processing, computation with large datasets and communication

Going forward

- Statistics research
  - Complex data and representation
  - Robustness to unknown heterogeneity
  - Scalability of robust models
  - Automation and interperability
  - M.L. and data processing
- Communication
  - Reproducibility, readable code, open source
- Education
  - More focus on the hypothesis searching, not testing

Complex data

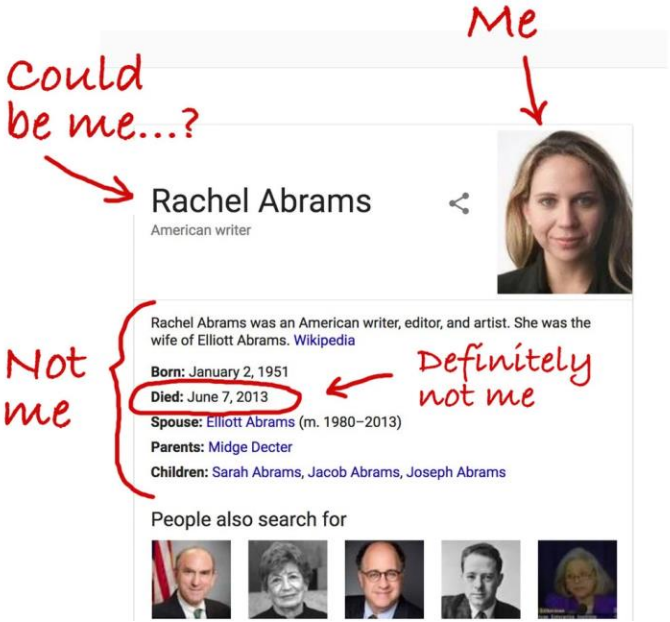
- Object oriented data analysis (OODA)
  - Connections between OODA and representation learning?
    - Resolving what should be the data objects?
    - Deep neural nets automatically find "good" representations?
  - Multidisciplinary:
    - differential geometry
    - Topology
    - Optimization, etc..

Robustness

- Big datasets are often made by collaboration
  - Example: cancer research
- The Gaussian standard model insufficient
  - More research into for example Gaussian mixture
- Make robust models more scalable

Automation and interpretability

- Negative consequences
  - "Weapons of Math Destruction"
- Interpretability
  - XAI, LIME,
- Raw data preprocessed using deep NN
  - Another black box before the black box?
  - If google thinks you're dead, the world also does



Communication

- Literate programming
  - Write code for humans to read, not machines
  - R Markdown, Knitr, Jupyter notebook
  - Important when demonstrating programming examples
- Open-source
  - Not inventing the wheel every time
  - Suspects most modern statistics research are not available in open-source software packages

Education

- Is the statistics curriculum up to date?
- More computation
  - Infeasible to know everything
  - Probably a rough set of knowledge every statistician should have
- Pedagogy
  - Focus on real world datasets
  - Exploratory analysis first, inferential thereafter

Conclussion

- It is important to increase diversity
- Is it two different cultures?
  - Given that statistics = Lesser statistics => YES
  - Given that statistcs = Greater statistics => NO

Thoughts

- Not enough focus on commercialisation? e.g.
  - INLA
  - Bayesian Optimisation

- What is "providing value to society"?
  - Why should we get paid?
  - What do the tax-payers get back?


That's all Folks!