Data Science and the Undergraduate Curriculum

By Iain Carmichael
09/11/17
UNC Chapel Hill
STOR Department Colloquium

Data science is a fraught term

https://www.linkedin.com/pulse/putting-science-back-data-paul-dalen

Data science is a fraught term

Seems to marginalize statistics

Means different things to different people

Lot’s of hype

Data science vs. statistics

- The Statistics Identity Crisis: Are We Really Data Scientists? (JSM, 2015)
- Arguably data science = applied statistics

Term originated from statisticians

- **Data science**: an action plan for expanding the technical areas of the field of statistics, by William Cleveland (Cleveland, 2001)
- John Tukey (Tukey, 1962)

The ability to work with data is **empowering** and **in demand**

Large, unmet demand (Manyika, 2011)

- Industry
- Academia
- Government

High salaries, interesting work (Burtch, 2014)

Academic programs are changing to meet new demands

New and expanding academic programs

- http://datascience.community/colleges (over 500 college data science programs)
- http://midas.umich.edu/

Berkeley Foundations of Data Science

- "fastest growing program in the history of Berkeley" (Alivisatos, 2017)
- Data8 = 155 + 320 http://data8.org/
  - Broad target audience beyond traditional STEM majors

Online courses and bootcamps outside of traditional academia

- http://datascience.community/bootcamps
- https://www.coursera.org/specializations/jhu-data-science

Opportunities for STOR department

Increase opportunities for **students in STOR department**

- Otherwise may struggle to get jobs
- E.g. tech companies often favor programmers with some statistics over statisticians with little programming

Appeal to **other STEM students** outside of statistics

- Math, CS, Bio, INLS, Chem, etc

Appeal to **students outside of STEM**

- Journalism, English, etc

Data science can increase interest in “traditional” statistics courses

STOR 320: **Introduction to Data Science**

Previously STOR 390

https://idc9.github.io/stor390/

Outline

1. **Goals and background**
2. Course Overview
3. Undergraduate curriculum

Goal of data science: **use data to solve problems**

Use data to **understand** something

- Inference
- Ex: Associations between genetics and disease outcomes, consumer behavior

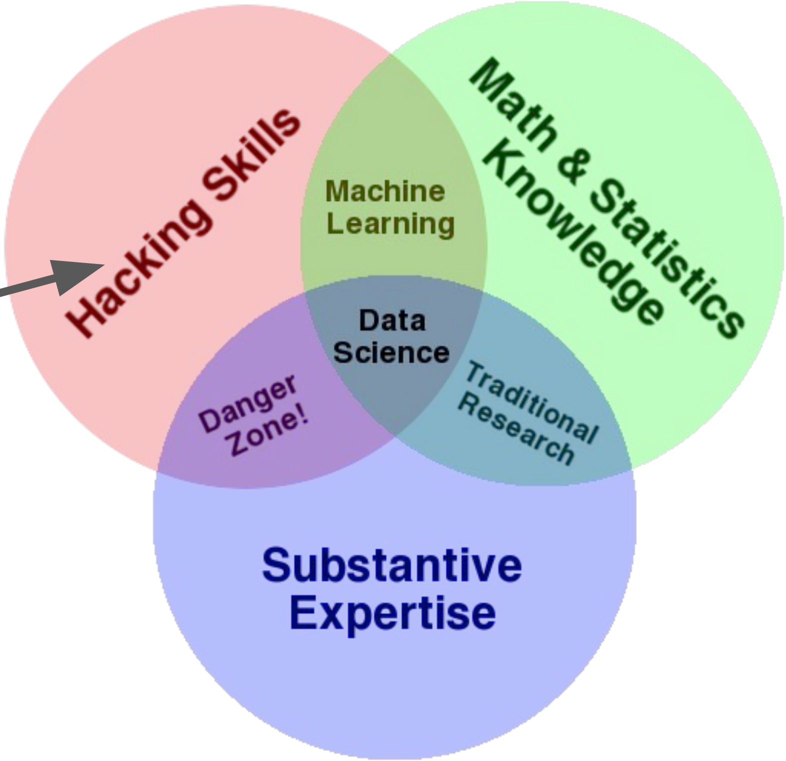
Use data to **do** something

- Prediction
- Ex: Stock market prediction, facial recognition, self driving car
- Machine learning/artificial intelligence

Scientific method + problem solving/engineering

Data science is interdisciplinary

Computer science +
unglamorous problem solving



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Many statisticians have discussed shifting priorities of statistics

The Future of Data Analysis, John Tukey (Tukey, 1962)

Data science: an action plan for expanding the technical areas of the field of statistics, by William Cleveland (Cleveland, 2001)

Statistical modeling: The two cultures, by Leo Breiman (Breiman, 2001)

Rise of the Machines, by Larry Wasserman (Wasserman, 2014)

Curriculum guidelines for undergraduate programs in statistical science, by the American Statistical Association (ASA, 2014)

50 years of Data Science, by David Donoho (Donoho, 2015)

David Donoho on definition of data science (Donoho, 2015)

1. Data Exploration and Preparation
   a. Exploratory analysis
   b. Data cleaning
2. Data Representation and Transformation
   a. Databases (e.g. SQL)
   b. Mathematical representation (e.g. networks, images, etc)
3. Computing with Data
   a. Programming (R/Python)
   b. Technologies
4. Data Visualization and Presentation
5. Data Modeling
   a. Inferential
   b. Predictive
6. Science about Data Science
   a. Workflows
   b. Reproducibility

Current focus on **inference** and **theory**

1. Data Exploration and Preparation
   a. Exploratory analysis
   b. Data cleaning/carpentry/munging
2. Data Representation and Transformation
   a. Databases (e.g. SQL)
   b. Mathematical representation (e.g. networks, images, etc)
3. Computing with Data
   a. Programming (R/Python)
   b. Technologies
4. Data Visualization and Presentation
5. Data Modeling
   a. Inferential
   b. Predictive
6. Science about Data Science
   a. Workflows
   b. Reproducibility

Donoho (and others) argue statistics should be concerned with **all** of these areas*

1. **Data Exploration and Preparation**
   a. Exploratory analysis
   b. Data cleaning
2. **Data Representation and Transformation**
   a. Databases (e.g. SQL)
   b. Mathematical representation (e.g. networks, images, etc)
3. **Computing with Data**
   a. Programming (R/Python)
   b. Technologies
4. **Data Visualization and Presentation**
5. **Data Modeling**
   a. Inferential
   b. Predictive
6. **Science about Data Science**
   a. Workflows
   b. Reproducibility

*To a greater extent than it currently is

80/20 rule in data analysis

“first reasonable thing you can do to a set of data often is 80% of the way to the optimal solution” (Leek, 2014)

Corollary: to solve problems with data, the most bang for the buck come from

- Programming
- Exploratory analysis
- Basic inferential/predictive modeling
- Effective Communication

Explains why tech companies want data scientists to be programmers

Goal of 320: statistics majors should have the **skills** to analyze data and **experience** doing data analysis

Programming

Problem solving

Acquiring data

Working with real data and using statistical methodology

Communicating results

Outline

1. Goals and background

2. **Course Overview**
    a. Data analysis
    b. Communication
    c. Final Project
    d. Takeaways

3. Undergraduate curriculum

320 was developed with **a lot of help**

Data@Carolina

Shankar Bhamidi, Robin Cunningham, Brendan Brown, Dylan Glotzer, Marshal Markham, Varun Goel

Existing data science courses at other school

Books, blogs, podcasts

Data science education literature

Consulted 50+ people

See https://idc9.github.io/stor390/course_info/acknowledgments.html and references at end of slides

Topics breakdown of 320

Core R programming skills (11 lectures, 40%)

- Data manipulation, visualization, loops, if/else, etc

Data analysis (8 lectures, 30%)

- EDA, linear models, classification, etc

Getting data (3 lectures, 10%)

- Web scraping, APIs, twitter

Communication (3 lectures, 10%)

- RMarkdown, general principles, Shiny

Additional topics (3 lectures, 10%)

- Text data (e.g. non-standard data)
- Data ethics/inequality, Weapons of Math Destruction (O'Neil, 2017)

Topics not mutually exclusive

Example data sets

Data.gov

UNC departments

Biodiversity in North Carolina

Museum of Modern Art

Movie ratings from IMDB

Bike Sharing

iPhone moment tracking

Beauty and the Beast (text script)

Harry Potter (text of the books)

All data sets can be found at:
https://github.com/idc9/stor390/tree/master/data

Homework breakdown

10 smaller labs

- Targeted to practice individual skills
- Sometimes real data, sometimes fake data
  - In future would like to avoid fake data

4 longer assignments

- Real data, some open ended questions
- (tried to be) problem driven

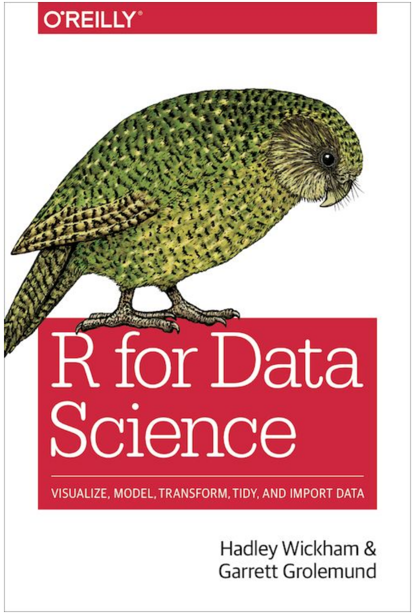
In class activities

- Practice individual skills
- Active learning (Brent and Felder, 2016)
- Sometimes interactive and/or discussion based

Final project

- Do a novel data analysis to answer a question and write about it

Technologies and textbooks: all **free**, state of the art



stackoverflow

Technologies and textbooks: all **free**, state of the art

R, RStudio, RMarkdown, Shiny

R for Data Science by Hadley Wickham (Wickham and Garrett, 2016)

Google/stack exchange

For some topics

- Blog posts (e.g. https://simplystatistics.org/)
- Introduction to Statistical Learning (Gareth et al, 2013)
- Text Mining with R (Silge and Robinson, 2016)

Outline

1. Goals and background

2. **Course Overview**
    a. **Data analysis**
    b. Final project
    c. Communication
    d. Takeaways

3. Undergraduate curriculum

First assignment: **get** a data set from data.gov, make some **visualizations** and **write** up results

https://idc9.github.io/stor390/labs/1/gov_data.html

Data analysis topics

Exploratory analysis

Linear regression

Clustering

- K-means

Classification

- KNN, Nearest centroid, SVM*
  - Should have done logistic regression instead of SVM

Data transformation, interactions, polynomial terms

Focused on **exploration** and **prediction**

Lot’s of visualizations

What does a model do?

How do you code the model?

Some of the underlying math?

Prediction is often **easier than inference**

Easier to know when model is “correct”

- Test set error
- Prediction : inference as physics : social sciences

Less background knowledge

Many statistical models can be **introduced** with only a little attention to randomness

- Linear regression, K nearest neighbors, Fisher’s linear discriminant, PCA/SVD
- Once students understand the models (how/what/why) *then* teach theory

Outline

1. Goals and background

2. **Course Overview**
    a. Data analysis
    b. **Final project**
    c. Communication
    d. Takeaways

3. Undergraduate curriculum

Final Project: **do a novel data analysis to answer a question and then write about it**

1. Ask a question
2. Find data set
3. Do analysis
4. Write a report
5. Write a blog post (less technical)

Teams

- Rules for teamwork (Brent and Felder, 2016)
    - Instructor assigned teams
    - Final grade weighted by peer review
    - Students can vote stragglers off the island
    - See https://idc9.github.io/stor390/final_project/description.html#grading
- Teamwork should be practiced

Final project brought the course together

Outline

1. Goals and background
2. **Course Overview**
   a. Data analysis
   b. Final project
   c. **Communication**
   d. Takeaways
3. Undergraduate curriculum

Communication in different **contexts** with different **mediums**

Lecture on principles of effective communication

- https://idc9.github.io/stor390/notes/communication/communication.html

Writing and in class discussion

How to ask questions

- https://stackoverflow.com/help/how-to-ask
- http://adv-r.had.co.nz/Reproducibility.html (reproducible example)

RMarkdown and Shiny

RMarkdown enables literate programming for data analysis

Literate programming: write one document with **code, text and images together**

Better communication of technical results

Reproducibility

Lot's of capabilities

- Websites, data analysis reports, books, resume, slideshow, dashboards
- http://rmarkdown.rstudio.com/gallery.html

Jupyter notebooks roughly equivalent for Python

Shiny for interactive applications

https://shiny.rstudio.com/gallery/

Guest lecture by Frances Tong

Outline

1. Goals and background

2. **Course Overview**
   a. Data analysis
   b. Final project
   c. Communication
   d. **Takeaways**

3. Undergraduate curriculum

Concrete learning outcomes

Programming in R

Ability to acquire and work with data

- The modern data analyst is expected to be able to actively get data for themselves

Solve problems with data

- Classify modeling problems (e.g. classification, clustering, regression)
- Basic understanding of some of the canonical models
- A taste of non-standard data

Problem solving skills

Communication

Creating your own data analysis instills **skepticism of poor data driven arguments**

See how the sausage gets made

Lot’s of things can go wrong other than insignificant p-values

- Finding a representative sample
- Coding error
- Choices/garden of forking paths

Challenges

Bi-modal class

- Some had programming experience, some didn't

Choosing topics to cover

Teaching programming

"Know enough to be dangerous"

Some takeaways from teaching 320

Teaching coding

- Modify existing code makes learning easier
- What is plagiarism?
    - For details see https://idc9.github.io/stor390/course_info/syllabus.html#honor_code

Problem/question oriented (tried to be)

- Hard, can get better as course progresses

Outline

1. Goals and background
2. Course Overview
3. **Undergraduate curriculum**

Decisions in resource constrained environment

Many things we might want to teach, only so many courses

Focus of this talk on what we might teach more of, not on trade-offs

Many have called to update the statistics curriculum

Statistics programs should provide majors with sufficient background in the following areas (ASA, 2014)

- Statistical methods and theory
- Data manipulation and computation
- Mathematical foundations
- Statistical practice (teamwork + communication)
- Discipline-specific knowledge

(Tukey, 1962), (Cleveland, 2001), (Nolan, 2010), (Hardin et al, 2015), (Baumer, 2015), (Cobb, 2015), (Donoho, 2015), (Hicks and Irizarry, 2017), (De Veaux et al, 2017)

STOR 320 is a start

Course design choices

Modular

- Topics
- Data sets
- Code

Open source (Leek, 2017)

- Textbook, programming language, data sets
- Many other online resources available (swirl, Coursera, etc)
- The entire course: https://idc9.github.io/stor390/

Visualization was the first topic

Guest speakers

Prioritize **simple, useful** and **generalizable** topics

Many possible topics, little time

- Did not cover: SQL, github, hadoop, more advanced modeling/computation, etc

Ex: did not cover SQL since we covered dplyr

Teach R or Python since they provide the most value

R and Python

- Open source and **free**
- Easy to learn and many **free, quality resources**
- High fixed cost, low marginal cost
- Lots of data analysis capabilities including advanced capabilities
    - e.g. RCPP, Cython
- **Flexible and generalizable**
    - General purpose programming languages
    - RMarkdown/Jupyter notebook, web scraping, data cleaning, communications

Other options do not meet all of the above

R and Python are roughly equivalent with some trade-offs

- http://makemeanalyst.com/most-popular-languages-for-data-science-and-analytics-2017/
- Maybe in a few years Julia will win...

What level should “intro to data science” be taught?

Freshman

- no pre-req
- ex: Berkeley’s Data8 = 155 + 320 http://data8.org/

Junior-senior (current version)

- some stats/prog pre-reqs

Senior-masters

- more stats/math/prog pre-reqs
- ex: Harvard’s CS109 http://cs109.github.io/2015/

Where does 320 go from here (microscale)?

Prerequisites

- **Require a programming class before 320** (e.g. comp 110)
- Should 320 come before, during or after 455?
  - Affects how we cover modeling (linear/logistic regression, etc)

Improve data sets, examples, explanations, etc

More instructional staff

- e.g undergrad TAs

How does 320 fit in with the rest of the curriculum?

Should we teach data before statistics?

Lot’s of visualization

Demonstrate kinds of questions one might ask

Once students comfortable with data, then teach statistics/probability

Co-teach data science with other departments

Different options

- 1 stat + 1 CS
- 1 stat + 1 topical dept (e.g. journalism)
- 1 stat + 1 CS + 1 topical dept

Large course for general audience

- Berkeley’s Data8

Bureaucratic barriers, but there is enthusiasm around UNC for this idea

General recommendations

Open up the black box: teach computations that are most used in statistics

- Eigen-decomposition/numerical linear algebra
- Gradient descent family

Experience with real problems

- **Standard practice in engineering and natural sciences**

Introduce students to data before statistical theory

Teach **real applications and code** along with theory

What about the data science and the graduate curriculum?

For additional information

Course website: **https://idc9.github.io/stor390/**

- Syllabus
- Notes/slides
- Homeworks
- Readings
- Final project
- Other resources
- All code for the course on https://github.com/idc9/stor390

Contact Iain

- Website: https://idc9.github.io/
- Email: iain@unc.edu

Helpful literature and courses on next two slides

Other data science courses relevant to 320

Data8 at Berkeley
http://data8.org/

Johns Hopkins data science specialization on Coursera
https://www.coursera.org/specializations/jhu-data-science

Introduction to Data Analysis by Hadley Wickham
http://stat405.had.co.nz/

Data Science in Statistics Curricula: Preparing Students to “Think with Data”
http://www.stat.purdue.edu/~mdw/papers/paper032.pdf

STAT 545 by Jenny Bryan at UBC
http://stat545.com/

CS109 at Harvard
http://cs109.github.io/2015/

Machine Learning by Emily Fox and Carlos Guestrin on Coursera
https://www.coursera.org/specializations/machine-learning

Computational Statistics and Statistical Computing at Duke (graduate level)
http://people.duke.edu/~ccc14/sta-663-2017/

Teaching data science literature

Data Science in Statistics Curricula: Preparing Students to “Think with Data”
http://www.stat.purdue.edu/~mdw/papers/paper032.pdf

Curriculum Guidelines for Undergraduate Programs in Data Science
https://www.stat.berkeley.edu/~nolan/Papers/Data.Science.Guidelines.16.9.25.pdf

Curriculum Guidelines for Undergraduate Programs in Statistical Science
http://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf

Computing in the Statistics Curricula
https://www.stat.berkeley.edu/~statcur/Preprints/ComputingCurric3.pdf

A Guide to Teaching Data Science
https://arxiv.org/pdf/1612.07140.pdf

Biased sample of references I found helpful

Bibliography

Alivisatos, Paul. “STEM and Computer Science Education: Preparing the 21st Century Workforce.” Research and Technology Subcommittee House Committee on Science, Space, and Technology. (2017)
http://docs.house.gov/meetings/SY/SY15/20170726/106330/HHRG-115-SY15-Wstate-AlivisatosA-20170726.pdf
(Alivisatos, 2017)

American Statistical Association. "Curriculum guidelines for undergraduate programs in statistical science." *Retrieved March 3, 2009, from http://www. amstat. org/education/curriculumguidelines. cfm* (2014).
http://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf
(ASA, 2014)

Baumer, Ben. "A data science course for undergraduates: Thinking with data." *The American Statistician* 69.4 (2015): 334-342.
https://arxiv.org/pdf/1503.05570.pdf
(Baumer, 2015)

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical science* 16.3 (2001): 199-231.
https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726
(Breiman, 2001)

Felder, Richard M., and Rebecca Brent. *Teaching and learning STEM: A practical guide*. John Wiley & Sons, 2016.
http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118925815.html
(Brent and Felder, 2016)

Burtch, Linda. "The Burtch Works Study: Salaries of Data Scientists.(April 2014)." (2014).
http://www.burtchworks.com/files/2014/07/Burtch-Works-Study_DS_final.pdf
(Burtch Works, 2014)

Cleveland, William S. "Data science: an action plan for expanding the technical areas of the field of statistics." *International statistical review* 69.1 (2001): 21-26.
https://pdfs.semanticscholar.org/915c/d8e2b39eb02723553913d592b2237d4d9960.pdf
(Cleveland, 2001)

Cobb, George. "Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up." *The American Statistician* 69.4 (2015): 266-282.
https://arxiv.org/pdf/1507.05346.pdf
(Cobb, 2015)

De Veaux, Richard D., et al. "Curriculum guidelines for undergraduate programs in data science." *Annual Review of Statistics and Its Application* 4 (2017): 15-30.
http://www.annualreviews.org/doi/full/10.1146/annurev-statistics-060116-053930
(De Veaux et al, 2017)

Donoho, David. "50 years of Data Science." *Princeton NJ, Tukey Centennial Workshop*. 2015.
http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
(Donoho, 2015)

Hardin, Johanna, et al. "Data science in statistics curricula: Preparing students to “think with data”." *The American Statistician* 69.4 (2015): 343-353.
http://www.stat.purdue.edu/~mdw/papers/paper032.pdf
(Hardin et al, 2015)

Hicks, Stephanie C., and Rafael A. Irizarry. "A Guide to Teaching Data Science." The American Statistician just-accepted (2017): 00-00.
https://arxiv.org/pdf/1612.07140.pdf
(Hicks and Irizarry, 2017)

James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.
http://www-bcf.usc.edu/~gareth/ISL/
(Gareth et al, 2013)

Joint Statistical Meetings. “The Statistics Identity Crisis: Are We Really Data Scientists?” (2015)
https://ww2.amstat.org/meetings/JSM/2015/onlineprogram/ActivityDetails.cfm?SessionID=211266
(JSM, 2015)

Leek, Jeff. “The 80/20 rule of statistical methods development” Simply Statistics. (2014)
https://simplystatistics.org/2014/03/20/the-8020-rule-of-statistical-methods-development/
(Leek, 2014)

Leef, Jeff. “The future of education is plain text” Simply Statistics. (2017)
https://simplystatistics.org/2017/06/13/the-future-of-education-is-plain-text/
(Leek, 2017)

Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).
http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
(Manyika, 2011)

Nolan, Deborah, and Duncan Temple Lang. "Computing in the statistics curricula." *The American Statistician* 64.2 (2010): 97-107.
https://www.stat.berkeley.edu/~statcur/Preprints/ComputingCurric3.pdf
(Nolan, 2010)

O'Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.
https://weaponsofmathdestructionbook.com/
(O’Neil, 2017)

Silge, Julia, and David Robinson. "tidytext: Text mining and analysis using tidy data principles in R." (2016).
http://tidytextmining.com/
(Silge and Robinson, 2016)

Tukey, John W. "The future of data analysis." The annals of mathematical statistics 33.1 (1962): 1-67.
https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711
(Tukey, 1962)

Wasserman, Larry. "Rise of the Machines." *Past, present, and future of statistical science* (2014): 1-12.
http://www.stat.cmu.edu/~larry/Wasserman.pdf
(Wasserman, 2014)

Wickham, Hadley, and Garrett Grolemund. "R for data science." (2016).
http://r4ds.had.co.nz/
(Wickham and Garrett, 2016)