What recedes as data science rises?

Cassandra Pattanayak
Wellesley College
eCOTS
May 25, 2018

Lots of great ideas this week!

Ideas for assignments, approaches, courses, curriculum

Ethics
Reproducibility
Web scraping
Real, messy data
Computational thinking
Data mining
R Shiny
Interactive visualization
...

But also:
“Is Central Limit Theorem still relevant?”
“How much of the data science tools should we teach in the first course?”
“Data science is not replacing statistics”

Have the goals of statistics education changed?

Yes.

A discussion at the New England Isolated Statisticians Meeting:

What topics should we add to intro stats?

What topics should we remove from intro stats to make room?

Let's be thoughtful, deliberate, purposeful about not only what we are adding, but also the consequences of adding and focusing on new material.

Questions:

What are we cutting or deemphasizing to make room for data science?

What should we be cutting or deemphasizing to make room for data science?

Contexts:

- Syllabus for intro, intermediate, or advanced course

  What students spend time on, even if syllabus unchanged

  Student skill set after leaving these courses

- Curriculum: courses offered, and *courses students actually take*
- Majors and minors: what's offered, *what do students choose*
- We are preparing students for modern career options. Lots of discussion about skills they need; what don't they know?

Case Study: STAT 260, Applied Data Analysis

Second-level applied stats course

Loosely uses *Statistical Sleuth*

Blended learning, videos for programming topics and also for some lectures

Prerequisite is any intro statistics course, or sometimes a Quantitative Reasoning course

Uses R

What should be removed from a syllabus when data science topics are added?
When we do something new in an existing course, what does it replace?

Sometimes doesn’t seem like a huge issue: replace a small, clean data set with a big, messy one.
Or, change in attitude, approach.

But still:
How are students spending their finite energy and attention? What skills are they leaving with?

Topics added to STAT 260 over past 5 years

Regression trees
Missing data
Large data sets and how to clean
Text analysis
Prediction contest (OkCupid data set, presidential election contest)
Web scraping
Interactive graphics
SQL
Various R packages

Used to let them figure out R on their own. Now, videos.

What did these topics replace?

Multiple lectures on causal inference
Calculus derivation of regression
Proofs that rely on expectation and variance
Anything parametric for categorical data
Introduction to logistic regression

Side effects: No longer prereqs in calculus or linear algebra or probability
To help with this: Blended learning materials include lecture video carrying out proofs omitted from course

Which existing courses should be offered less frequently to make room for data science, and which courses have relatively lower enrollments?

When we teach a new course, is it replacing a new course from the point of view of faculty? Of students?

Wellesley now has more than one **second-level modeling course**

Math-stats: Is it about simulation? About Bayes? About MLE theory?
Last year this was a new course for us, sometimes incongruous with current conversations in stats ed

What does a data science major/minor NOT know that a statistics major/minor knows?

How has the student population changed?

When we offer a new major/minor/program, is it replacing another field? Which one?

What don’t data science majors/minors have?

Probability? Impacts all other courses!

Math-stats

Upper-level courses in stats: Which ones? Maybe experimental design, or generalized linear models, or stochastic processes?

Depth of understanding: Can they explain a prediction model, a p-value, the way a particular algorithm works, how various regression quantities relate to each other, the components of software output...

What would the data science students have studied without data science?

Statistics? CS? Math? Econ? ...

What does the emphasis on data science remove from the training of students headed toward various professions?

The professions themselves have changed

For **graduate school in statistics**, shouldn't they have probability and math-stats?

Economics majors taking my class instead of stat theory or perhaps additional math courses, like linear algebra

Psychology majors learn more practical skills in the age of data science. Are they learning experimental design?

Those headed into **tech jobs** - do we teach them to recognize when their knowledge is not enough? (eg, are they running and over-interpreting hypothesis tests on giant data sets?)

Let’s be deliberate and try to anticipate consequences in order to make effective, long-lasting revisions to the field of statistics education.