Sta 771S: Data Science / Statistics

- Discuss the questions assigned to your team: 20 mins
- Prepare a brief presentation that summarizes the key points of your discussion: 2-3 mins
- Lead a discussion on your topic: 5 mins

Tentative teams are noted below. Team assignments might be updated if there are absences.

Team 1: Michael, Kyle, Matt

1. Speed plays a significant role in data science because some decisions need to be made as soon as possible. Is it appropriate to give the students an in-class data analysis or programming exam in an introductory data science / statistics course?
2. How useful are the various described courses in data science in addition to statistics (learning to analyze data) or CS (learning to program) for undergraduate's future careers say 20 years from now, considering that technology and the nature of new data evolves so quickly?
3. If the interest is in decision making using data, shouldn't we only be teaching Bayesian statistics? Incorporating decision making is what sets apart Bayesian approaches from frequentist (a fact often overlooked by even Bayesians; see Lindley (1992)). It also integrates well with computation. Why would we waste time teaching students frequentist statistics (beyond the basics)? (Note there are 0 mentions of Bayesian approaches in both papers.)
4. How important is math or theoretical statistics for an industry or government employee? This also ties in with the previous readings, and removing math barriers for students. But the previous readings took more of a position that you don't need all the math that is typically presented, and these ones seem to think that most of it is a waste of time, and that it doesn't apply to real data anymore.
5. Of the jobs that exist in the market today, which statistical principles do you think are most important in preparing students for the current jobs? (i.e. dealing with a million observations instead of just n $\rightarrow \infty$)

Team 2: Willem, Mao, Lu, Shaobo

1. At what level should the course introduced in Baumer (2015) be to have the most impact since it appears that for students to really appreciate the course, they should be "reasonably" comfortable with introductory statistics and some level of programming? Especially since we would not want students to spend all their time learning basic stats and programming syntax instead of actually learning the art of data science.
2. The Data Science course discussed in Baumer (2015) has Statistics as a prerequisite. Given our previous discussion on how to make intro Stats more attractive, doesn't it

make more sense to make Data Science a prerequisite for Stats? How should the Data Science course change then?

3. An introductory course covering the major pipelines of data science suffices to prepare students thinking roughly with data. However, many more statistical methodology courses are necessary to cultivate students really being able to think rigorously with statistics. In teaching data science, are the necessity and importance of focusing on data engineering techniques (e.g. scrape data, clean data, or visualization) overemphasized?
4. In your opinion, which of the six data science courses that Hardin et al. (2015) describes would be best for an advanced undergraduate statistics major?
5. Which aspects of the curriculum should we really consider primary. For example, some of the courses covered efficiency in programming but should that be a primary learning objective?

Team 3: Ken, Christine, Princeton

1. Which is better to emphasize as a department: proficiency with one programming language (such as R) or exposure to many programming languages (such as SQL, Python, R, SAS, Stata, etc.)?
2. How to enhance the programing skills of students with statistical backgrounds? What aspects would a programming course taught in a statistic department look different from that in a computer science department?
3. Is it necessary to replace the neat and tidy datasets with complex and unstructured alternatives in teaching statistics? Although the toy datasets used in statistics lectures are often too simple/unrealistic for real world applications, the statistical thinking trained through those neat datasets will still be useful in the future career of students as data analysts.
4. Common topics in the courses covered in this paper is the heterogeneity of computational skills of the students. But, what level of computation is needed to be a data scientist? This vastly ranges between, for example, a tech company or a financial firm. While data cleaning is stressed, in some cases, there are in house tech people to assist with this process. I think the focus should be on knowing what kind of data to collect and how to analyze/visualize it rather than learning how to clean data, which often is specific to that data set.
5. Also, I completely agree with the statement in Hardin et al. (2015) that students are often faced with messier and more complex data in real applications than they do in introductory stats classes. However, should we also be wary of overwhelming students with too much complex data at the beginning? Should we consider data somewhere between the two extremes instead?

Additional questions:

Baumer (2015)

- 14 weeks is short in my opinion for such a dense class. For effective learning, shouldn't the content of the course be split into multiple classes (Baumer 2015 does agree that the content could be split into multiple classes)? If yes, how many?
- How is data science at Duke taught differently than data science taught at a liberal arts college, such as Smith?
- Why is data visualization recommended to be taught in an introductory statistics course? Can this be covered in advanced workshops instead?
- Why and why not should we separate the (sub?)field Data Science from Statistics and Computer Science, like how CS has been separated from Math and Physics and AI from CS and Psychology? Note that this is not about the relevance of Data Science for Statistics (or vice versa). e.g., programming is relevant to Statistics but belongs to the field CS.
- If the paper gives an accurate description of the future of data science, and data is no longer going to fit classical statistical models, what is the use of theoretical statisticians? Is it enough to justify its own department and major, separate from data science?
- Is it in the best interest of the schools to define their undergraduate stats curriculums as academia or industry based, and let students choose which school they want to attend (i.e. MBA vs. JD vs. PhD programs)? Or does it make more sense for a school to have a curriculum that seeks to fit both populations at the risk of fulfilling none?
- In the class discussed in this paper, the professor finds the goal and data, and the students are expected to clean, analyze/infer, and graph the data. While these skills are useful, isn't the purpose of education to train students to come up with their own goals, find their own data, and solve the problem (teach them how to fish)? If the goals are well thought out, the other skills should follow naturally.
- The risk of setting a goal is that the student might be uninterested in the topic and thus loose motivation. How can we keep the students motivated? What are the advantages of this class over doing a research project class, where students go through the whole process of Fig1 on their own, and the professor acting more as a mentor?

Hardin et al. (2015)

- Is the ability to work with textual data necessary for today's statistician? Should we teach this in a data science course?
- What is the difference between “think with data” and “think in statistics”?
- In terms of efficiency, how do massive online open courses (MOOC) compare with face-to-face data science and statistics courses?
- In MOOCs, how can we keep students with different backgrounds, experiences, and motivations from dropping out? Challenging assignments may help keep advanced students excited, but may also scare or frustrate novice students. It seems difficult to strike a balance.
- How much of the traditional statistics curriculum could (or should) be included in a first semester introductory data analysis course? What about a year-long course?
- Should statistics departments try to incorporate data science? Should data science be its own department? Should it be a joint venture of statistics and computer science? Baumer (2015) states that the CS majors with weaker stats backgrounds usually do better than the stats students with weaker programming backgrounds.
- How relevant is Data Science, excluding the Statistics, to Duke StatSci graduate students? Do they already know the material? Should they know the material?