Computing in the Statistics Curricula

Background

- NSF (DUE) funded grant on "Integrating Computing into the Statistics Curricula"
- Goals:
  - map syllabi and sequences of courses for modern stat. programs
  - Facilitate instructors to introduce and teach such classes, providing resources (lecture notes, syllabi, assignments, data sets) and support network.
- 4th of 4 workshops
  - Second one for teaching instructors to teach stat. computing.
  - 1st discussed model syllabi for different types of students
  - 3rd develop case studies in stat. computing for use by the stat. community. (In preparation.)

State of affairs

- Computing is a very large part of a data analysts daily task, yet remarkably small part of our educational programs
- We teach only enough computing to enable doing assignments, leaving students to learn on their own by mimicking and adjusting existing templates.
- Results are not good.

Importance of Computing

- Computing is as fundamental to statistics education as mathematics.
- We need to make the students facile with computing so they can transparently use it to express ideas.
  - Compute correctly and efficiently
- As long as computing is a difficulty, it is a distraction from one's primary focus.
  - We must reduce it from being an obstacle or focus when performing analyses.
  - Need to enable them to reason about computations.

Critical point in statistics.

- Statistical computing is much broader than traditional material we have taught
  - Need different topics for different types of students.
- Computing is becoming increasingly vital part of statistical education in this era of
  - the Web
  - Ubiquitous data availability & sources.
  - Increased volume and complexity of data.
  - New and ever-evolving Web technologies.
  - Increased relevance of data analysis in all fields, done by non-statisticians
  - Communicating results in new ways (e.g. Web graphics)
- If we can't compute the right answer, others will be involved to compute another answer.
- And computing the answer means dealing with increasingly complex computations and technologies.

- Computing is typically just one missing dimension of our programs
  - Exposure to modern statistical methods.
  - Actual solving of scientific/data problems with statistical approaches.
- We tend to use a computing class to address both of these, as well as computing/ programming topics.

- We need to teach it early in the students program.
- Integrate it properly into a program
  - Teach it an intellectually rich level so students learn to reason about computation, not just "trial and error".
  - Leverage it as an alternative medium for learning statistical concepts.

Focus

- Our focus here is on
  - our upper-division and graduate classes.
  - Teaching computational skills for data analysis and statistical research.
  - Equipping our students with knowledge for their careers to be able to use computers easily and to learn and evaluate new technologies as they continue to emerge.
- We are less focused on "using computing to teach statistics"
  - But are interested in how to integrate computing into such classes.

Challenges

- Students today are very familiar with computers, phones, etc.
- But not with scientific computing
  - Vocabulary, computational reasoning, programming
- We have to get them past "this is too hard to use", "why can't I click on a button", …
  - Need to learn grammar and computational model to express themselves, both to the computer and to humans.
  - Compare with how we teach stat. concepts, or to write essays
- GUIs are convenient, but imply that the ability to compute is of secondary importance and that the computations are generic/templates.

Organizational Challenges

- We also have to get our colleagues on board
  - Difficult since few of us have been trained in computing.
- Fit more into already crammed curricula, and room for few additional classes.
- Computing requires regular, repeated exposure to internalize concepts, vocabulary and "grammar".

Format

- Hopefully, very interactive with questions, comments, etc. from everybody.
- Please let us know if you want us to schedule discussion on a topic not in the "schedule".
- Notes will be available on the Web site after the conference and will be updated as comments/ questions are received.
- Goal is to make these materials available to you and others to facilitate teaching stat. computing classes.

"Rules of engagement"

- When speaking, please identify yourself to everyone and let us know your institution.
- We hope in the coffee/food breaks you'll interact and build connections.

Schedule

- During the day, we'll have a mix of discussions and tutorials
- We have asked some people to give short presentations on what they are doing.
- Everyone else, please participate throughout.
- Each topic will be a mix of
  - Big picture reason about why the topic is important
  - Technical details about the material – tutorial
  - Ways we teach this material
- Introduce some material about R that is less familiar to most people, with the hope of it helping you to teach computing.

Emphasis on R

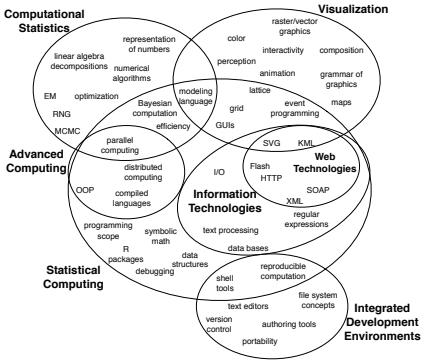
- We will be talking a great deal about R and numerous "modern" technologies.
- We hope to be discussing computing & programming at a level that is general and abstracts to other programming languages, e.g. MATLAB, Python.
- Some will expect material on SAS perfectly reasonable.

SAS, etc.

- Firstly, we are much more familiar with R and other technologies.
- Focus is on general, complete programming languages and modern technologies students are likely to encounter over the years.
- Fundamentals of programming are transferable to other languages.
- R is free, easy to deploy and widely used in stat. departments
  - For both teaching and research.
- Increasingly widely used in industry.

- In our courses, we teach 5 languages
  - R and 4 DSLs – Domain Specific Languages
    - Regular expressions for patterns in text
    - XML/HTML/KML/SVG & XPath
    - SQL – database query language
    - (Unix) shell
- Key is to learn how to learn new languages and technologies.

What to teach - Themes & Topics



Leave it to Computer Science?

- It may seem reasonable to have stat. students take computer science classes. Is this good?
- Students would benefit from such classes, but only in addition to an solid statistical computing class.
- The skills are different –
  programming for data analysis versus
  many data structures, programming languages, software design, efficiency in run-time
- Vectorized computations on sample observations versus focus on individual elements.
- Text manipulation & data input
- Graphics & Information visualization versus low-level graphics.

- If we want statistics to thrive and play an important role in the data-centric world, we need to
  - Teach computing
  - Do research in the next generation of computing technologies for statistics.
  - And the only way to do the latter, is to do the former!

Summary

- We have to stop thinking of computing as an optional add-on within the study of statistics that students pick up in an ad hoc manner by themselves.
- Computing is an essential, foundational skill for modern data analysis and without such skills, our students are missing at least half of the essential knowledge.
- It is at least time for stat. instructors to learn modern stat. computing & to enhance their programs with computing & data analysis classes.