Technologies for Data Sciences

Informal Seminar

Spring 2012

Aims

- Become familiar with existing and emerging technologies used in different areas of "data sciences"
  - Programming languages, data management systems, computational paradigms, modern visualization systems
- Know what people are using in industry and other disciplines for working with "big data".
- Learn to learn about new technologies.
- Foster culture of computing and network of "weakly linked collaborators".

Introductions

- Introduce yourself
  - Name & Program/research field,
  - what you work on,
  - what you want out of this seminar

Format

- You are the ones that I expect to provide much of the content and activity in this seminar.
- Hopefully, each of you will research and explore a topic that is of interest to you (and checked with myself and Jerome) and give a presentation on that.
- Work in pairs (probably best)
- The presentations can be short (15 or 30 minutes) or the whole hour depending on the topic.

- I'll try to help you with anything you want.
- You can ask me about difficulties you are having getting things to work, comparing technologies, better approaches.
- Best to ask on the mailing list to get input from other additional people
  Sign up to the mailing list

  statscicomp@ucdavis.edu
- Please try to send me an outline or the slides for a talk a day or so before the presentation so I can give some input.

- Create resource that we all can use and refer to.
- How-to's, help pages, etc. that we all edit.
  - A wikie, or via the github shared resource.
- Link to other pages and documents on the Web
- Create screen captures (i.e. videos of on-screen interaction)
- Dynamic documents.
- What do you use?

Presentations

- The presentations hopefully will be a resource that others can follow to
  - get started, and also
  - understand the benefits, deficiencies and limitations of the technology
  - show some advanced details and additional things to explore.
- Think of it as a conference talk - covering the big points that lead others to want to know more, but enough to get the main ideas.
  - We can provide additional slides, etc. with extra details.

Topics

- What topics do you want to know about?
- Don't want to focus on topics you can learn in an existing course. (But they are okay too.)
- Technologies for working with Data
- Statistical Methods for Big Data
- What do you want to learn about?
  - day-to-day topics
  - new topics you have heard about and are curious.

R Performance

- Hopefully somebody is interested in talking about some R packages and approaches for working with big data
  - e.g. bigmemory, biglm;
    using external SQL databases;
    compiling R functions (compile)
- Look at the Task Views

Fundamentals of Performance

- Vectorization
- Avoiding redundant computations
- Memory management & removing unnecessary objects
- Time Profiling via Rprof()
- Memory profiling