A little philosophy, ramblings, and a preview of coming events

John W. Emerson

http://www.stat.yale.edu/~jay/
Associate Professor of Statistics, Yale University
(Professor Emerson prefers to be called “Jay”)

Please feel free to ask questions along the way!

http://www.stat.yale.edu/~jay/Brazil/Campinas/

Outline

Statistics? Computer Science? Bioinformatics? Sports? What's up with this guy?

- I love my job! The teaching, the research, the wide range of problems I see every week...
- Example: Yale's Statistical Clinics, `http://www.stat.yale.edu/clinic/Clinic.html`
- It's all about the data and real-world problems; statistics should be data-driven, or at least problem-driven
- Data analysis should not simply be an excuse to exercise new theory
- Você trabalha em que? Eu sou professor. Eu sou estudante. Ambos!

Citações favoritas

- “Para chamar o estatístico após o experimento é feito pode ser mais do que pedindo a ele para realizar um exame post-mortem: ele pode ser capaz de dizer que a experiência da morte.” - Sir Ronald A Fisher
- “O plural de anedota não é de dados.” - Roger Brinner
- “A combinação de alguns dados e um desejo doloroso de uma resposta não garante que uma resposta razoável pode ser extraído de uma determinada massa de dados.” - John Tukey

Favorite quotes

- “To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.” - Sir Ronald A Fisher
- “The plural of anecdote is not data.” - Roger Brinner
- “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.” - John Tukey

Monday Morning

A speedy introduction to R with some entertaining data analysis along the way.

An Introduction to R: por que R?

- R is the *lingua franca* of statistics.
- It is a language and environment for statistical programming that is ideal for *interactive* data analysis and graphics, and much, much more.
- It is extended by a large collection of *packages*.
- If you want a GUI, there are some options. But that misses point. *GUI* $\not\rightarrow$ *reproducible research*. Don't go there.

Preliminaries: available resources

- The R Project: `http://www.r-project.org/`
  - Official Documentation: `http://cran.r-project.org/manuals.html`
  - Contributed Documentation: `http://cran.r-project.org/other-docs.html`
  - Other resources linked on CRAN: Frequently Asked Questions (FAQs), the R Journal, a Wiki, Books, etc...
- Another R community site: `http://crantastic.org/`
- Sweave: `http://www.statistik.uni-muenchen.de/~leisch/Sweave/`
- Reproducible research: `http://cran.r-project.org/web/views/ReproducibleResearch.html`

About the talk

- Not for a “liberal arts” audience: no graphical user interface (GUI)
- Good for people newish to R who want to learn more (or hear a different perspective)
- Good for people who have never used R but who have a fairly solid programming/scripting background
- Will be about 30-40 minutes of formal introduction to the language fundamentals
- Will include about 30-40 minutes of data analysis on a real-world problem (judging bias in Olympic diving) to reinforce these fundamentals: engaging data with R

Monday Afternoon

The R package management system, the C/C++ interface, an introduction to parallel programing via **foreach**, all in the context of Bayesian change point analysis.

Why R?

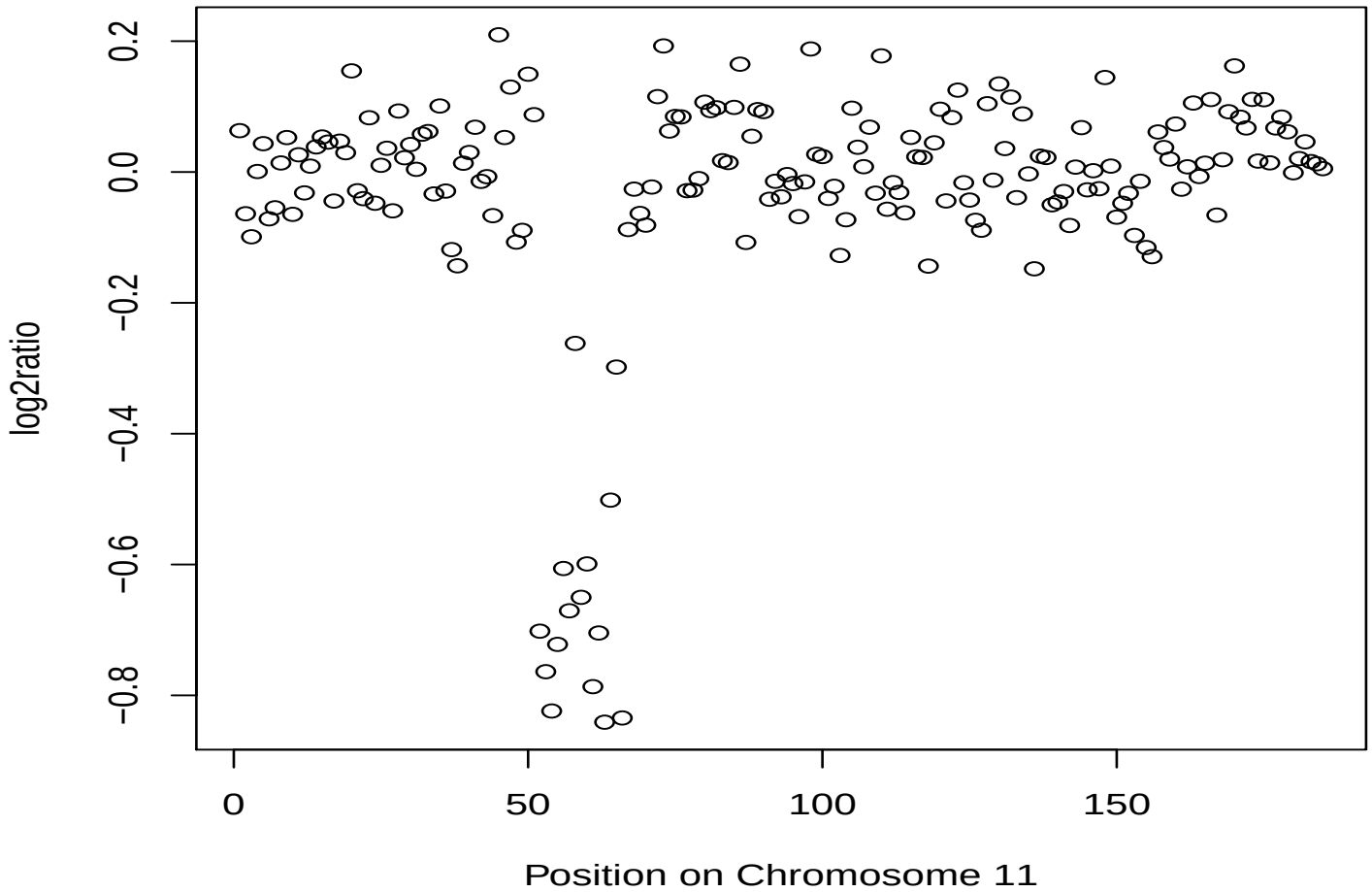
R is the *lingua franca* of statistics:

- The syntax is simple and well-suited for data exploration and analysis.
- It has excellent graphical capabilities.
- It is extensible, with over 2500 packages available on CRAN alone.
- It is open source and freely available for Windows/MacOS/Linux platforms.

This talk emphasizes the importance of the package management system. Much of the success of R should be attributed to:

- Ross & Robert's early decision to go open-source and encourage collaboration, and
- the growth of CRAN and the success of the package management system.

Example: Coriell cell lines (raw data)



foreach

The user may register any one of several “parallel backends” like **doMC** or **doSNOW**, or none at all. The code will either run sequentially or will make use of the parallel backend, if specified, without code modification.

```
> library(foreach)
> library(doMC)
> registerDoMC(2)
>
> a <- 10
> ans <- foreach(i = 1:5, .combine = c) %dopar%
+     {
+         a + i^2
+     }
>
> ans
[1]  11  14  19  26  35
```

Tuesday Morning

An introduction to the Bigmemory Project, covering pitfalls and solutions for working with massive data.

A new era

The analysis of very large data sets has recently become an active area of research in statistics and machine learning. Many new computational challenges arise when managing, exploring, and analyzing these data sets, challenges that effectively put the data beyond the reach of researchers who lack specialized software development skills of expensive hardware.

- “Entramos em uma era de enorme coleção de dados científicos, com a procura de respostas para os problemas de inferência em grande escala que estão além o âmbito das estatísticas clássicas.” – Efron (2005)
- “classical statistics” should include “mainstream computational statistics.” – Kane, Emerson, and Weston (in preparation, in reference to Efron’s quote)

Example data sets

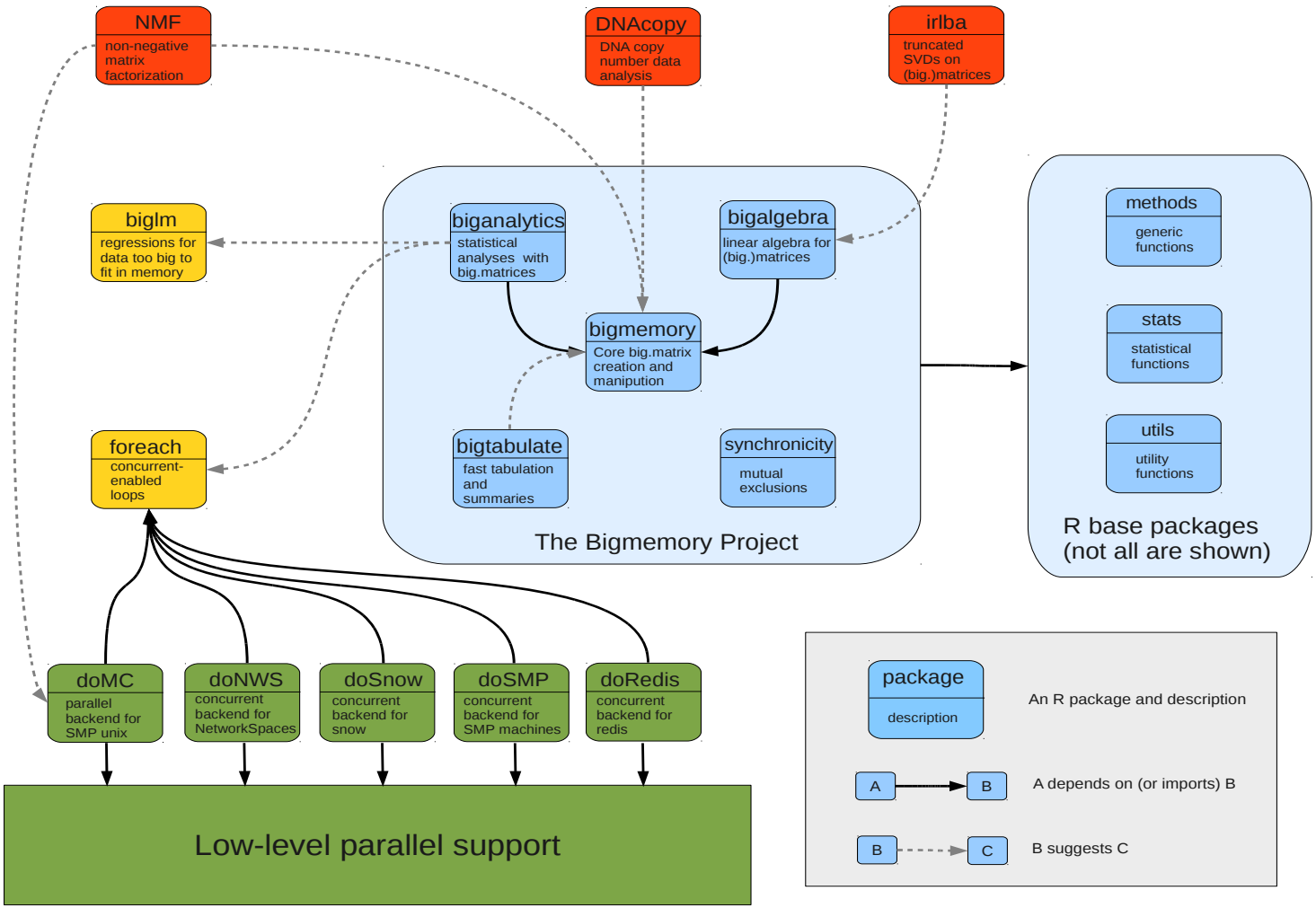
- Airline on-time data
  - 2009 JSM Data Expo (thanks, Hadley!)
  - About 120 million commercial US airline flights over 20 years
  - 29 variables, integer-valued or categorical (recoded as integer)
  - About 12 gigabytes (GB)
  - `http://stat-computing.org/dataexpo/2009/`
- Netflix data
  - About 100 million ratings from 500,000 customers for 17,000 movies
  - About 2 GB stored as integers
  - No statisticians on the winning team; hard to find statisticians on the leaderboard
  - Top teams: access to expensive hardware; professional computer science and programming expertise
  - `http://www.netflixprize.com/`

Why R?

R is the *lingua franca* of statistics! (Did I say that earlier?)

Currently, the Bigmemory Project is designed to extend the R programming environment through a set of packages (**bigmemory**, **bigtabulate**, **biganalytics**, **synchronicity**, and **bigalgebra**), but it could also be used as a standalone C++ library or with other languages and programming environments.

The Bigmemory Project: http://www.bigmemory.org/



In a nutshell...

- The approaches adopted by statisticians in analyzing small data sets don't scale to massive ones.
- Statisticians who want to explore massive data must
  - be aware of the various pitfalls;
  - adopt new approaches to avoid them.
- We will
  - illustrate common challenges for dealing with massive data;
  - provide general solutions for avoiding the pitfalls.

Examples

Some examples, time permitting.

Conclusion

Espero que alguns de vocês podem desfrutar de uma discussão mais aprofundada de alguns dos temas que vou falar. Por favor não se acanhe em pedir perguntas, durante ou antes ou depois de qualquer das conversações. Isto é particularmente verdadeiro para o Projeto Bigmemory, onde alguns de vocês podem já estar a usá-lo e tiver perguntas específicas.