

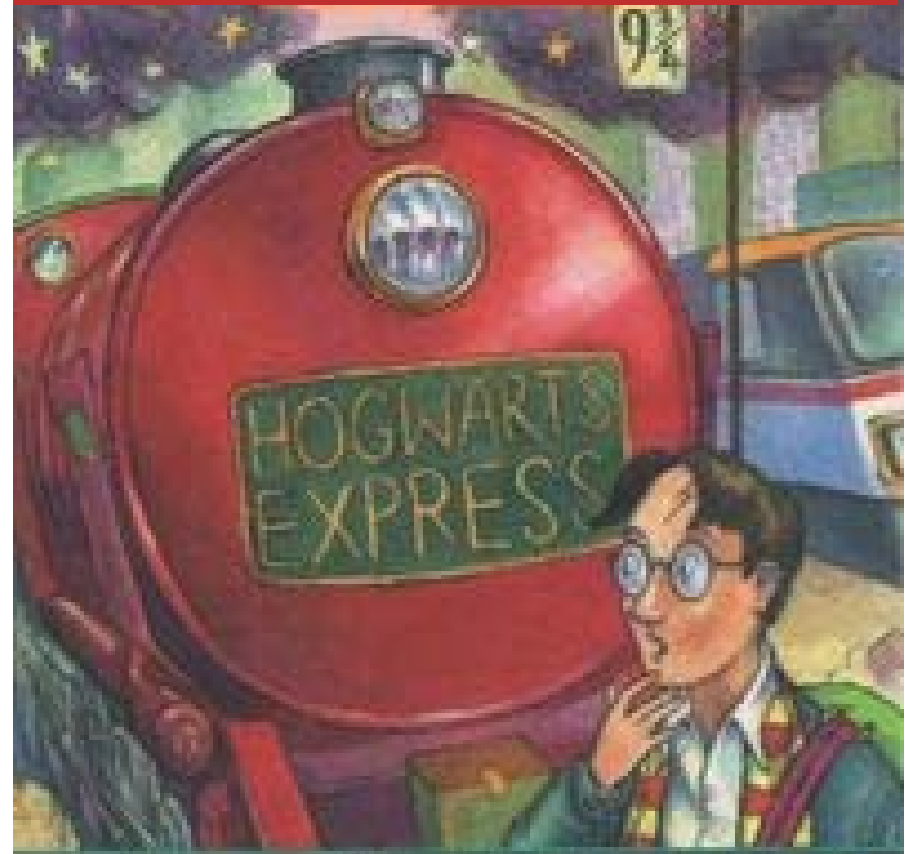
A very brief introduction to using R & MX

- Matthew Keller

Some material cribbed from: UCLA Academic Technology Services
Technical Report Series (by Patrick Burns) and presentations (found
online) by Bioconductor, Wolfgang Huber and Hung Chen, & various
Harry Potter websites

R programming language is a lot like magic...
except instead of spells you have functions.

R, And the Rise of the Best Software Money Can't Buy



"...this is a terrific book." *The Sunday Telegraph*

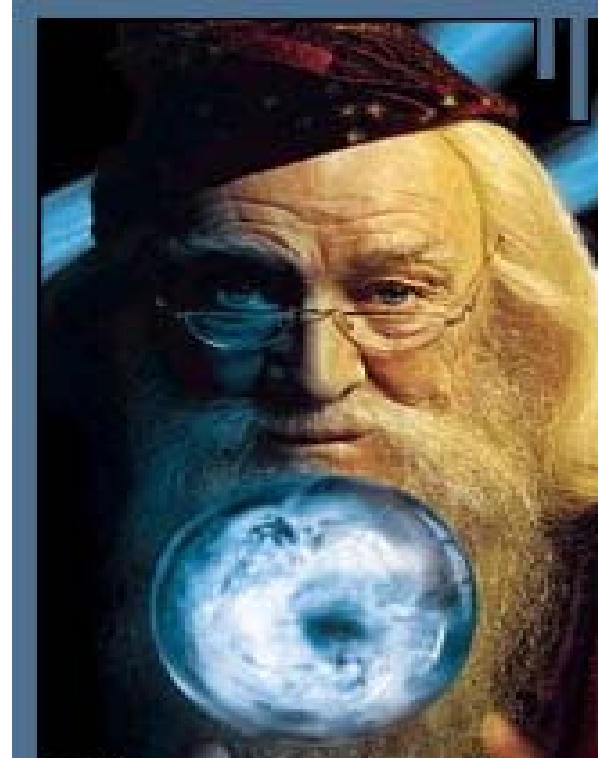


=



muggle

SPSS and SAS users are like muggles. They are limited in their ability to change their environment. They have to rely on algorithms that have been developed for them. The way they approach a problem is constrained by how SAS/SPSS employed programmers thought to approach them. And they have to pay money to use these constraining algorithms.



wizard

R users are like wizards. They can rely on functions (spells) that have been developed for them by statistical researchers, but they can also create their own. They don't have to pay for the use of them, and once experienced enough (like Dumbledore), they are almost unlimited in their ability to change their environment.

History of R

- S: language for data analysis developed at Bell Labs circa 1976
- Licensed by *AT&T/Lucent* to *Insightful Corp.*
Product name: *S-plus*.
- R: initially written & released as an open source software by Ross Ihaka and Robert Gentleman at U Auckland during 90s (R plays on name “S”)
- Since 1997: international R-core team ~15 people

“Open source”... that just means I don't have to pay for it, right?

- No. Much more:

- Provides full access to algorithms and their implementation
- Gives you the ability to fix bugs and extend software
- Provides a forum allowing researchers to explore and expand the methods used to analyze data
- Ensures that scientists around the world - and not just ones in rich countries - are the co-owners to the software tools needed to carry out research
- Promotes reproducible research by providing open and accessible tools
- Most of R is written in... R! This makes it quite easy to see what functions are actually doing.

R

Advantages

- Fast and free.
- State of the art: Statistical researchers provide their methods as R packages. SPSS and SAS are years behind R!
- 2nd only to MATLAB for graphics.
- Mx, WinBugs, and other programs use or will use R.
- Active user community
- Excellent for simulation, programming, computer intensive analyses, etc.
- Forces you to *think* about your analysis.
- Interfaces with database storage software (SQL)

Disadvantages

R

Advantages

- Fast and free.
- State of the art: Statistical researchers provide their methods as R packages. SPSS and SAS are years behind R!
- 2nd only to MATLAB for graphics.
- Mx, WinBugs, and other programs use or will use R.
- Active user community
- Excellent for simulation, programming, computer intensive analyses, etc.
- Forces you to *think* about your analysis.
- Interfaces with database storage software (SQL)

Disadvantages

- Not user friendly @ start - steep learning curve, minimal GUI.
- No commercial support; figuring out correct methods or how to use a function on your own can be frustrating.
- Easy to make mistakes and not know.
- Working with large datasets is limited by RAM
- Data prep & cleaning can be messier & more mistake prone in R vs. SPSS or SAS
- Some users complain about hostility on the R listserve

Learning R....



R-help listserve....



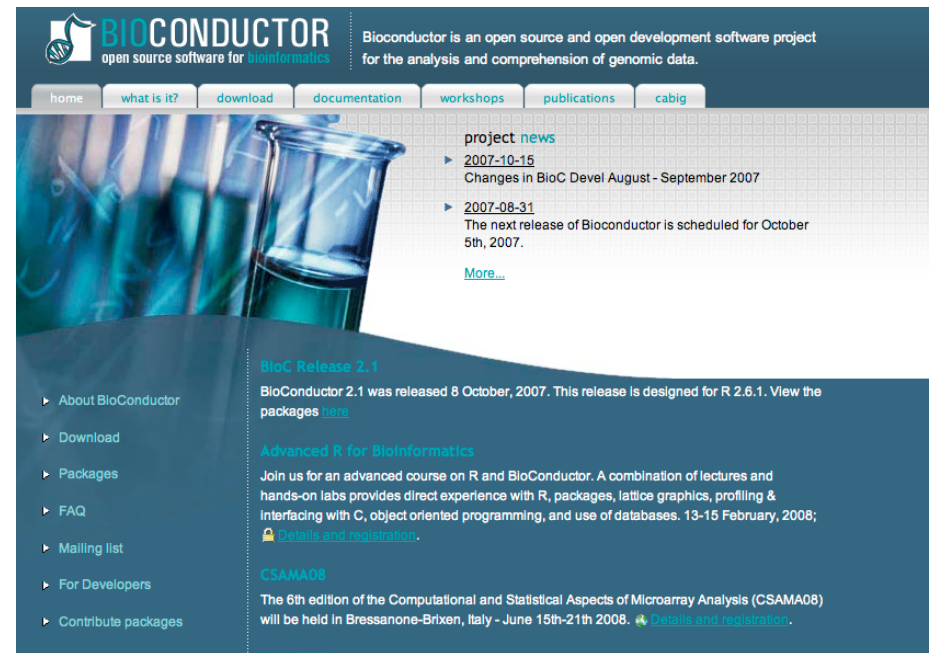
There are over 800 add-on packages

(<http://cran.r-project.org/src/contrib/PACKAGES.html>)

- This is an enormous advantage - new techniques available without delay, and they can be performed using the R language you already know.
- Allows you to build a customized statistical program suited to your own needs.
- Downside = as the number of packages grows, it is becoming difficult to choose the best package for your needs, & QC is an issue.

A particular R strength: genetics

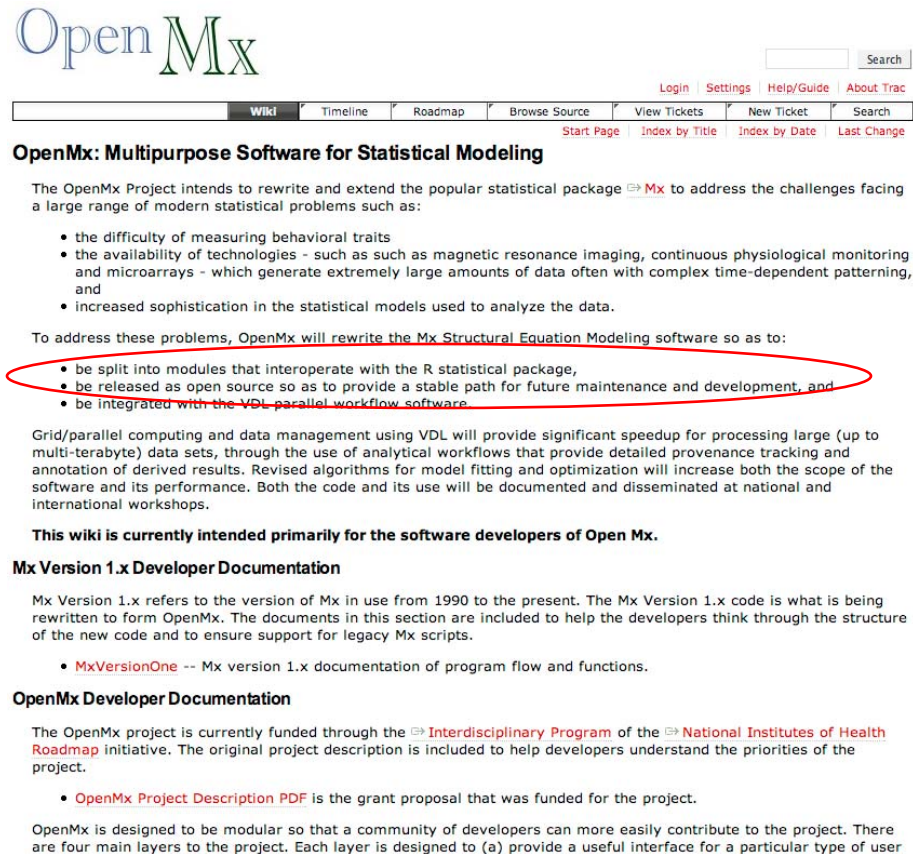
- Bioconductor is a suite of additional functions and some 200 packages dedicated to analysis, visualization, and management of genetic data
- Much more functionality than software released by Affy or Illumina



The screenshot shows the Bioconductor website homepage. At the top, the Bioconductor logo is displayed with the tagline "open source software for bioinformatics". To the right, a description states: "Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data." Below this is a navigation menu with links for "home", "what is it?", "download", "documentation", "workshops", "publications", and "cabig". The main content area features a background image of laboratory glassware. On the right side, there is a "project news" section with two entries: "2007-10-15 Changes in BioC Devel August - September 2007" and "2007-08-31 The next release of Bioconductor is scheduled for October 5th, 2007." Below the news section, there are three highlighted announcements: "BioC Release 2.1" (released 8 October, 2007), "Advanced R for Bioinformatics" (an advanced course on R and Bioconductor), and "CSAMA08" (the 6th edition of the Computational and Statistical Aspects of Microarray Analysis).

An R weakness

- Structural Equation Modeling - the sem package is quite limited.
- But this may not be a weakness for long...



The screenshot shows the OpenMx website. At the top, there is a navigation bar with links for 'Login', 'Settings', 'Help/Guide', and 'About Trac'. Below this is a search bar and a menu with options like 'Wiki', 'Timeline', 'Roadmap', 'Browse Source', 'View Tickets', 'New Ticket', and 'Search'. The main heading is 'OpenMx: Multipurpose Software for Statistical Modeling'. The text below explains the project's goal to rewrite and extend the Mx package. A list of challenges is provided, followed by a list of goals for the OpenMx project. The third goal, 'be released as open source so as to provide a stable path for future maintenance and development, and be integrated with the VDL parallel workflow software', is circled in red. The page also includes sections for 'Mx Version 1.x Developer Documentation' and 'OpenMx Developer Documentation'.

OpenMx: Multipurpose Software for Statistical Modeling

The OpenMx Project intends to rewrite and extend the popular statistical package [Mx](#) to address the challenges facing a large range of modern statistical problems such as:

- the difficulty of measuring behavioral traits
- the availability of technologies - such as such as magnetic resonance imaging, continuous physiological monitoring and microarrays - which generate extremely large amounts of data often with complex time-dependent patterning, and
- increased sophistication in the statistical models used to analyze the data.

To address these problems, OpenMx will rewrite the Mx Structural Equation Modeling software so as to:

- be split into modules that interoperate with the R statistical package,
- be released as open source so as to provide a stable path for future maintenance and development, and
- be integrated with the VDL parallel workflow software.

Grid/parallel computing and data management using VDL will provide significant speedup for processing large (up to multi-terabyte) data sets, through the use of analytical workflows that provide detailed provenance tracking and annotation of derived results. Revised algorithms for model fitting and optimization will increase both the scope of the software and its performance. Both the code and its use will be documented and disseminated at national and international workshops.

This wiki is currently intended primarily for the software developers of Open Mx.

Mx Version 1.x Developer Documentation

Mx Version 1.x refers to the version of Mx in use from 1990 to the present. The Mx Version 1.x code is what is being rewritten to form OpenMx. The documents in this section are included to help the developers think through the structure of the new code and to ensure support for legacy Mx scripts.

- [MxVersionOne](#) -- Mx version 1.x documentation of program flow and functions.

OpenMx Developer Documentation

The OpenMx project is currently funded through the [Interdisciplinary Program](#) of the [National Institutes of Health Roadmap](#) initiative. The original project description is included to help developers understand the priorities of the project.

- [OpenMx Project Description PDF](#) is the grant proposal that was funded for the project.

OpenMx is designed to be modular so that a community of developers can more easily contribute to the project. There are four main layers to the project. Each layer is designed to (a) provide a useful interface for a particular type of user

How does R tie into what you've done this week?

- MX will soon become one of those add on packages in R
- “runmx”: You can run MX from within R (easier to find & manipulate matrices, save aspects of them, compare $-2LL$, etc)
- “GeneEvolve”: You can use R to simulate genetically informative designs.

Why use GeneEvolve? Modeling aid

- Check bias & identification:
 - ◆ Feed PE parameters you are modeling, simulate data, & see if your model recovers the parameters
- Check model's sensitivity to assumptions:
 - ◆ Simulate violations of assumptions & note its effects on estimates
- Estimate power & multivariate sampling dist's of estimates under very general conditions:
 - ◆ Run PE multiple times given whatever condition you want

Why use it? Predictor of population / evolutionary genetics dynamics

- Find changes in variance parameters & relative covariances under different modes of AM, VT, & genetic effects:
- Simulate random genetic drift by varying population size
- Introduce selection (coming) to test theories on maintenance of genetic variation

Final Words of Warning

- “Using R is a bit akin to smoking. The beginning is difficult, one may get headaches and even gag the first few times. But in the long run, it becomes pleasurable and even addictive. Yet, deep down, for those willing to be honest, there is something not fully healthy in it.” --Francois Pinard

