Do It Yourself Cloud Computing with R and Apache

Code4Lib 2010, Asheville NC

Harrison Dekker
Head of Data Lab
UC Berkeley Libraries
data@library.berkeley.edu

Intro

- Observations of a Data Librarian
  - discovery **and** analysis involves special software (e.g. Excel, SAS, SPSS, Stata, ArcGIS (and R, of course)
    - codebooks and metadata (when available) are an important start...
    - descriptive statistics and visualizations help
    - file format problems not going away soon
  - many opportunities exist for Libraries wrt data
    - (relatively) few experts
    - involvement in data-related user communities
    - use more data analysis in our own work

Overview

- The Cloud
- What is R
- What is rapache
- Library Relevance
- Getting Started

The Role of the Cloud

- Right choice when it makes us work smarter.
- Our data is often "in the cloud" already, why move it to analyze it.
    - Make things as easy as possible for decision makers
    - Many of their needs predictable
    - Browser has become desktop (most of the time)
- Don't rule out existing cloud applications
    - e.g. Google Analytics, Google Charts are great for many purposes.
    - Need guidelines for how/when to use

What is R

- Open source alternative to SAS, SPSS, Stata
- Supports analysis (i.e. statistical algorithms), visualization, data retrieval/management/munging
- Cross-platform
- Interpreted language
- Created by statisticians for statisticians
- Huge user community with large (and growing) library of add-on packages
- Formal infrastructure (CRAN) for finding/installing packages

R: Pros and Cons

Pros (just a few)

- de facto standard
- huge user community across disciplines/industries
- better graphics than Excel
- powerful, extensible...

Cons

- learning curve
- limited gui
- problems with very large datasets

What is Rapache

- Apache module developed at Vanderbilt U.
- puts an instance of R in each Apache process
- works similarly to PHP
  - embed R script in web pages
  - provides interface to GET/POST data
- Demo
  - baseball scores: http://data.vanderbilt.edu/rapache/bbplot
  - stock plot: http://rweb.stat.ucla.edu/stockplot/
  - more...
  - Local example

Relevance to Code4Lib and library community

- Roll-your-own interactive web, ILS, or e-journal usage visualizations
- Real-time user survey results
- Data visualization instruction tools
- Network analysis
- You tell me!

Getting started with R

- Learn some R
  - For a free resource, google: Verzani simpleR
  - Better yet, buy R in a Nutshell (best book I've found and it's in O'Reilly Safari if you subscribe)
  - Important packages: ggplot2, lattice
- Install rapache
  - source install instructions available at rapache website
    - (but I had problems on OS X)
  - rapache site has a vmware image if you don't want to compile
  - my VirtualBox image (Ubuntu server)