RCloud: Social, Exploratory Data Analysis on the Web

Carlos Scheidegger
AT&T Labs – Research

NYU-Poly Reproducibility Workshop
May 30th, 2013

R at AT&T Labs

- Lots of ad-hoc use (S was invented at the Bell Labs)
- Interactive sessions are the primary form of exploration
    - Statisticians don’t know what they’re looking for when doing EDA
        - They don’t want to become software developers
            - And they shouldn’t!
- Not all data is for all eyes

We're not looking for watertight reproducibility

- A lot of "reproducibility" is lost simply by misplacing scripts, or executing them in different computers and interactive sessions, and losing the history

- This is a user-experience problem

    - So it requires a user-experience solution

    - Current environments make it easy to do the wrong thing

    - Let's make the right thing easy

- In our case, reproducibility cannot exist at the expense of generality

What if every R session was transparently and automatically versioned?

- We call each script a notebook, a la Mathematica and iPython
- A notebook is a collection of R commands and Markdown documents
- User interface is over a webpage, talking to a server R process over Websockets
- Notebooks are always executed on the server, and metadata is captured there*
- No notebook execution happens unversioned

Github gists

- Every notebook is a git repository backed on Github via their gists API
- Users don't need to know this, but they can if they want to
    - So offline access to git repository is possible in case notebooks need to packaged and exported
- Much of the "Social Web" comes for free: likes, comments, shares, forks

Let's kill share by screenshot (or print, or save_figure)

- This is similar to the analog hole problem in media piracy. Screenshotting is bootlegging :)

- The only effective way to prevent it is to make it easier to do something else

- Every shared notebook result has a link to the notebook which created it, so changes can be made

Integrated searching; think site-wide .bash_history*

- "What are the popular scripts that use the $some_library library?"
    - Should work for results as well: "which of those scripts generate reports with $some_keyword?"

- Complication: different users at AT&T have different execution privileges
    - I should be able to see everyone's scripts, but **not** everyone's results
    - Example: employee survey text analysis!

Provenance of UI interaction*

- “Interaction [...] can be [...] slavery” - Jon Claerbout
- RCloud brings modern HTML5 plotting to R
    - currently, DC.js (D3) bindings, and some very basic WebGL plots
- We must save the state of the interaction in these plots!
    - Every possible UI state has to be serializable, capturable and replayable

Data provenance*

- API support
- Rcloud will provide an R library that lets users store data products, and associates them with notebooks and execution sessions
- “Elephants”: git-backed R data frames
  - don’t underestimate the speed of your computer

Downsides

- Execution is centralized; must be connected to the web
- No guarantees of any sort
    - For example, we expect package upgrades to be a problem in practice
- Dependence on github
    - Scripts are either public, or you pay for Github Enterprise

Current State

- In active development right now (read: mostly broken! :))

- “Open source”: http://github.com/cscheid/rcloud

- We have a handful of internal users; by end-of-year plan is to push out deployment to IT organization (from 5-50 to ~5000 developers)

Acknowledgments

- Simon Urbanek, Gordon Woodhull, Mike Kane, Stephen North, Chris Volinsky

- Many of the open-source libraries we use, including but not limited to

    - ACE, underscore.js, jquery, d3, dc.js, knitr