Integrating R efficiently to allow secure, interactive analysis within a clinical data warehouse

Daniel W. Connolly, Bhargav Adagarla, John Keighley, Lemuel R. Waitman
University of Kansas Medical Center
Interactive R statistical visualization in HERON Clinical Data Repository
Interactive R statistical visualization in HERON Clinical Data Repository

Please don’t laugh if our R plots are crude and ugly. ;-) We're new to R and here to learn.
Overview

- R in HERON/I2B2:
  - What it looks like
  - Motivation: research support goals

- Background
  - I2B2
  - R Engine Cell

- Toward a general architecture for I2B2+R
  - Efficiency/Scalability
  - Separation of Concerns, Security
HERON Research Support Goals

Clinical Data Repository supports:

- Cohort Discovery
  - prospective trials: feasibility
  - retrospective studies: data use

- Hypothesis Generation
  - explore data
  - summarize
  - visualize

HERON System Architecture

- Data from Epic Clarity database (> 7,000 tables & 60,000 columns)
- Transformed into an I2B2-compatible schema. Then, de-identified, and loaded on a separate database server to be accessed by I2B2.
- De-identified data used by I2B2 is deemed non-human subjects research by our institutional review board.
Exploring Breast Cancer comorbidities: Obesity, Diabetes

HERON brings together diabetes diagnosis and BMI from hospital EMR with cancer staging from tumor registry and vital status from the U.S. SSA death index.
i2b2 Patient Data Query

Browse Concepts (ONT)
- Diagnosis
  - Asthma
  - Diabetes
- Test
  - Glucose
  - PFT
- Demographics
  - Age
  - Sex

Cohort Data Query (CDO QRY)
- Diabetes
- Asthma

Patient Set

Patient Data Query (PDO QRY)
- Age
- Sex

Patient Data
- Patient 1: 63 year Male
- Patient 2: 48 year Female
- Patient 3: 24 year Male

source: Murphy et. al. AMIA 2010
R Engine Cell

The web client plug-in sends the data to the RE Cell through dynamically created XML messages.

The RE Cell creates the dataset for the analysis by parsing the XML and runs the Kaplan–Meier jar application. This application, through the JRI libraries, uses the R statistical software.

Integrating the R Engine Cell with HERON for Cancer Research

Issues:

- Clinical Domain
  - cardio vs. cancer
  - start at birth vs start at diagnosis
  - stratification: gender vs. stage

- Version Skew
  - RE Cell: I2B2 version 1.4
  - HERON: I2B2 version 1.6

- Architecture...
Toward a General Architecture for R in I2B2

Python, SQL, HTML, JavaScript

Kaplan Meier Web Client Plug-in

abc

xyz

Web Client Plug-in

Cancer prevention, treatment

Biostatistics, R

Biostatistics, R

I2B2 Hive

PM cell

Rgate

Rpy libraries

Kaplan Meier Web Client Plug-in

Apache

I2B2 Hive

Patient privacy, institutional liability

I2B2

KU Medical Center
The University of Kansas
Efficiency, Scalability: R Engine Cell Data Path

Kaplan Meier Web Client Plug-in

CRC Cell sends back to the plug-in an XML response containing the requested data (extracted from the i2b2 datawarehouse).

I2B2 HIVE

R statistical software

725,000,000 facts incl. 60,000 cancer cases
Efficiency, Scalability: 
rgate connects R to Oracle directly

Like the CRC cell, rgate calls the PM cell to validate authorization.
R Engine Cell approach to R Integration: Kaplan Meier jar application

R Code Generation in KMAnalysis.java:

```java
Integer[] statusInteger = (Integer[])status.toArray(new Integer[status.size()]);
StringBuffer statusStr = new StringBuffer();
statusStr.append("status<-c(");
for(int i=0;i<statusInteger.length;i++){
    statusStr.append(statusInteger[i].intValue());
    if(i!=(statusInteger.length-1))
        statusStr.append(",");
    }
statusStr.append(")");

re.eval("data=data.frame(time,status,gender)");
re.eval("names(data)=c('time','status','gender')");
re.eval("setwd("+resultFolder+"")");
re.eval("library(survival)");
re.eval("fit <- survfit(Surv(data$time, data$status) ~ gender, data)");
```
R Engine Cell approach to R Integration: Kaplan Meier jar application

R Code Generation in KMAnalysis.java:

```java
Integer[] statusInteger = (Integer[]) status.toArray(new Integer[status.size()]);
StringBuffer statusStr = new StringBuffer();
statusStr.append("status<-c(");
for(int i=0; i<statusInteger.length; i++) {
    statusStr.append(statusInteger[i].intValue());
    if(i!=(statusInteger.length-1))
        statusStr.append(",");
}
statusStr.append(")");
```

```r
re.eval("data=data.frame(time,status,gender)\n"");
re.eval("names(data)=c(\"time\", \"status\", \"gender\")\n");
re.eval("setwd(\"+resultFolder+\")\n");
re.eval("library(survival)\n");
re.eval("fit <- survfit(Surv(data\$time, data\$status) ~ gender)\n");
```
Separation of Concerns in rgate: R code goes in .R files

Analysis is written in the language of statisticians:

```r
##' km_analysis -- Kaplan Meyer analysis from i2b2 observations
library(ROracle)
acct = db_config()

patient.set.survival <- function(concept.paths, patient.set.id,
                                 web.folder, filename) {
  conn <- dbConnect(Oracle(), acct$username, acct$password, access)
  sql <- paste("select ", concept.paths$event, " panel
               , to_char(f.start_date, 'YYYY-MM-DD HH24:MI:SS') start_date
               , pset.patient_num
               , cd.name_char
               , cd.concept_cd
               from blueherondata.observation_fact f, ...")
  data = transform.observations(dbGetQuery(conn, sql))
  fit <- survfit(Surv(data$time, data$status) ~ concept.paths$stratum, data)
  png(paste(web.folder, filename, sep='/'))
  plot(fit, xlab="Time (Years)", ylab="Survival probability")
  dev.off()
}
```
Separation of Concerns in rgate: R code goes in .R files, but...

How well does the R code behave when the author is not there?:

```r
##' km_analysis -- Kaplan Meyer analysis from i2b2 observations
library(ROracle)
acct = db_config()
patient.set.survival <- function(concept.paths, patient.set.id,
                                 web.folder, filename) {
  conn <- dbConnect(Oracle(), acct$username, acct$password, access)
  sql <- paste("select '", concept.paths$event, ", to_char(f.start_date, 'YYYY-MM-DD HH24:MI:SS') start_date,
                pset.patient_num
                , cd.name_char
                , cd.concept_cd
  from blueherondata.observation_fact f, ...")
data = transform.observations(dbGetQuery(conn, sql))
fit <- survfit(Surv(data$time, data$status) ~ concept.paths$stratum, data)
png(paste(web.folder, filename, sep='/'))
plot(fit, xlab="Time (Years)", ylab="Survival probability")
dev.off()
}
```

python, SQL, HTML, JavaScript

what the R author needs

patient privacy, institutional liability
Object Capability Discipline supports the Principle of Least Authority

- Memory safety and encapsulation
  - Effects only by using held references
  - No powerful references by default

Reference graph ≡ Access graph
Only connectivity begets connectivity
Natural Least Authority
OO expressiveness for security patterns

1. closure inspection is not safe: `environment(function), as.list (function)`
2. `plot(fit)` implicitly uses results of `png(paste(web.folder, filename))`
3. R global environment most likely includes lots of powerful references

M. Miller, C. Morningstar, B. Frantz; "Capability-based Financial Instruments"; Proceedings of Financial Cryptography (Springer-Verlag); 2000
rgate Security Architecture: Authority in the System Context

Kaplan Meier Web Client Plug-in

POST

1

get user config

2

rgate

km_analysis.R

rpy libraries

3

R statistical software

select

4

I2B2 HIVE

I2B2 DW

PM cell
rgate Security Architecture:

km_analysis.R starts with no authority

"The principle of least authority requires one to design interfaces such that authority is handed out only on a need-to-do basis." - Miller et. al.
rgate Security Architecture:
deid.R attenuates DW access with patient set facet

"facets are objects that act as intermediaries between powerful objects and users that do not need (and should not be granted) its full power." - Miller et. al.
rgate Security Architecture:
km_analysis.R can only read patient set, write results

POST

Python

rgate.py

patients

deid.R

patient set #7..
rOracle
c

con #xf..

rpy2

run

observations

SELECT ...

I2B2 DW

POST

R

results

plot #2..

km_analysis.R

R

apache
Attenuated patient data access:

```r
# km_analysis -- Kaplan Meyer analysis from i2b2 observations
library(survival)
run_analysis <- function(patient.set, folder, filename, progress, 
paths, title, xmax) {
  obs.db = observations(patient.set, unlist(paths))
  progress(paste("query returned", nrow(obs.db), " observations."))

  data <- db2km(obs.db, paths)
  progress(paste("db2km resulted in ", nrow(data), " data points for plotting."))

  survplot(data, title, folder, xmax, filename)
  progress(paste("KM plot stored in", filename, " in", folder))
}
```

patient privacy, institutional liability

biostatistics, R
Efficient, Secure Interactive R statistical visualization in HERON/I2B2

cancer prevention, treatment

python, SQL, HTML, JavaScript

biostatistics, R

patient privacy, institutional liability

informatics.kumc.edu