R for Bioinformatics

Introduction, Programming, Data Analysis and Visualization
R in Bigdata Era

Gang Chen
chengang@bgitechsolutions.com

November 30, 2013

Outline

Next

How to generate CUHK-R slides

Steps

1. R and LaTeX
2. knitr package
3. git and github.com
4. compile it to PDF

Reproducible Research

Reproducibility of Biological Research

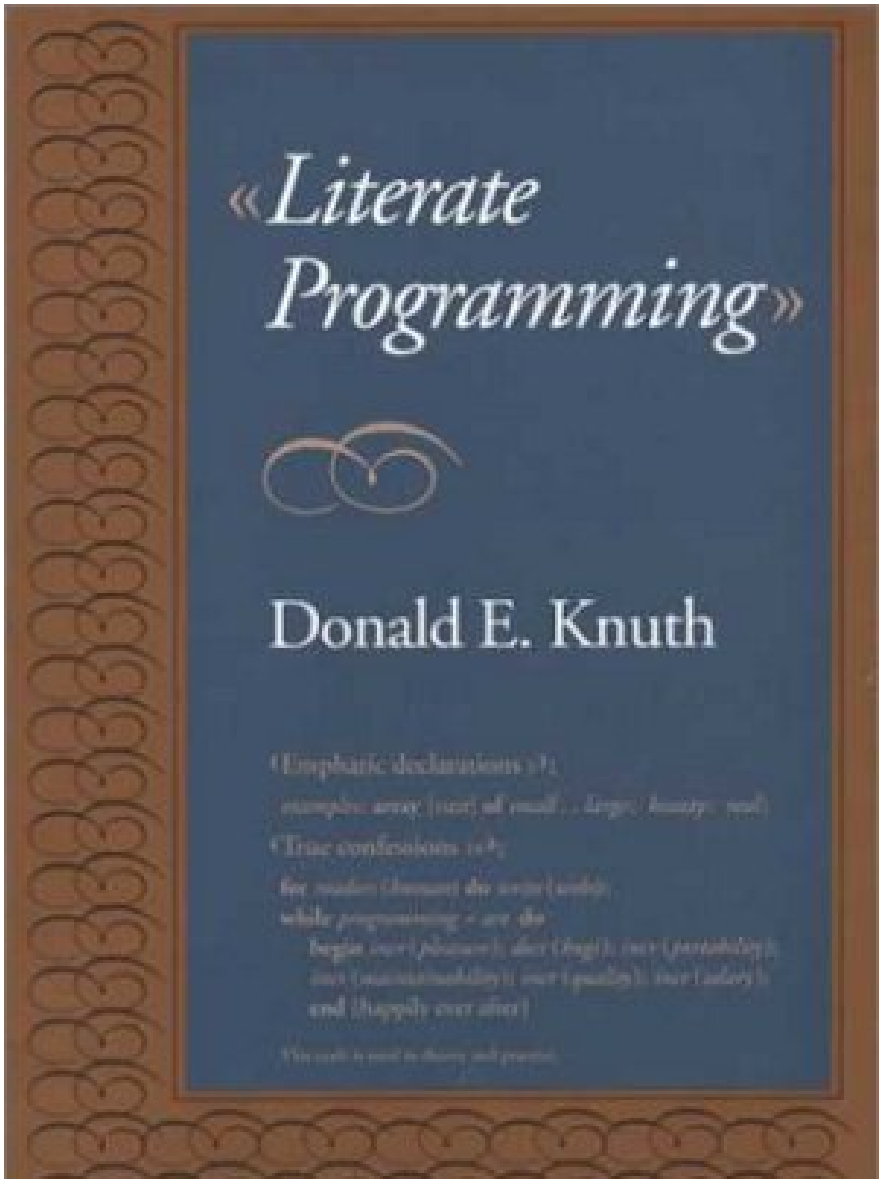
In 2012, a survey done for Nature found that 47 out of 53 medical research papers on the subject of cancer were irreproducible.
Begley, C. G.; Ellis, L. M. (2012). Nature

Reproducible Research

Reproducible Research

The term reproducible research refers to the idea that the ultimate product of academic research is the paper along with the full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research.

Literate Programming



Reproducible Research and R

http://cran.r-project.org/web/views/ReproducibleResearch.html

knitr

Overview

The knitr package was designed to be a transparent engine for dynamic report generation with R, solve some long-standing problems in Sweave, and combine features in other add-on packages into one package.

knitr

http://yihui.name/knitr/

Installation

- Stable version is shipped with R-core
- Develop Version:

```
update.packages(ask = FALSE,
    repos = 'http://cran.rstudio.org')
install.packages('knitr',
    repos = c('http://rforge.net',
              'http://cran.rstudio.org'),
 type = 'source')
```

knitr: LaTeXexample

LaTeXexample

- Source file: sample.Rtex
- ```
  library(knitr)
  knit("sample.Rtex")
  ```

  Output: sample.tex
- Compile sample.tex by using pdfLaTeXor XeLaTeX
  ```

Integrating Codes, Data and Report

R Package

Packages provide a mechanism for loading optional code, data and documentation as needed.

Example Package

- package.skeleton function
- DESCRIPTION file
- package structrue
- Compile and Install

Next

Why we need Interactive Report?

Static Report

Dynamic Report

Interactive Report

Go and R

http://115.29.195.56/iTRAQ/

What you need to know?

- Go (or PHP, Python, Perl ...)
- HTML
- Javascript
- R
- CSS
- SQL
- designing

Shiny Package

http://www.rstudio.com/shiny/

shiny package

Shiny makes it super simple for R users like you to turn analyses into interactive web applications that anyone can use. Let your users choose input parameters using friendly controls like sliders, dropdowns, and text fields. Easily incorporate any number of outputs like plots, tables, and summaries.
No HTML or JavaScript knowledge is necessary. If you have some experience with R, you're just minutes away from combining the statistical power of R with the simplicity of a web page.

Shiny Example

What you need to know?

- R

Next

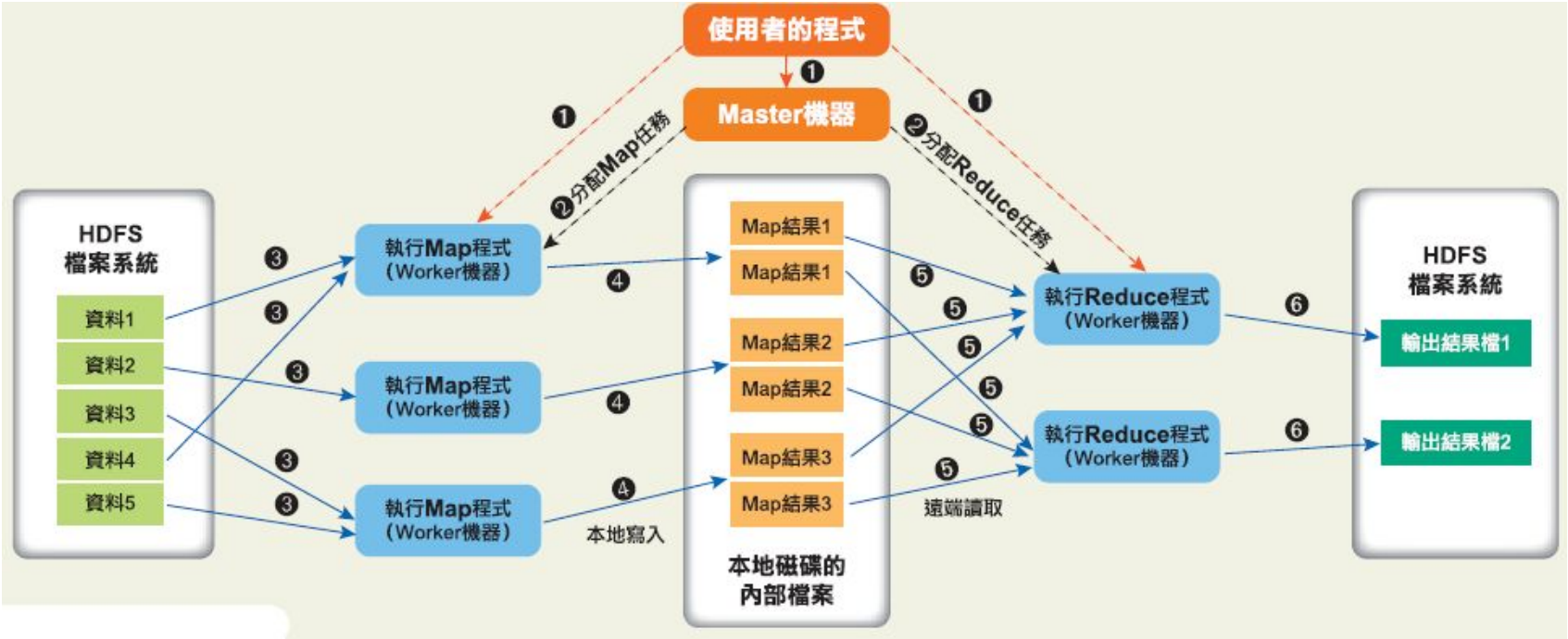
Packages for Bigdata

- bigmemory: www.bigmemory.org
- biganalytics
- parallel
- Rmpi
- snow
- gputools: http://www.r-tutor.com/gpu-computing

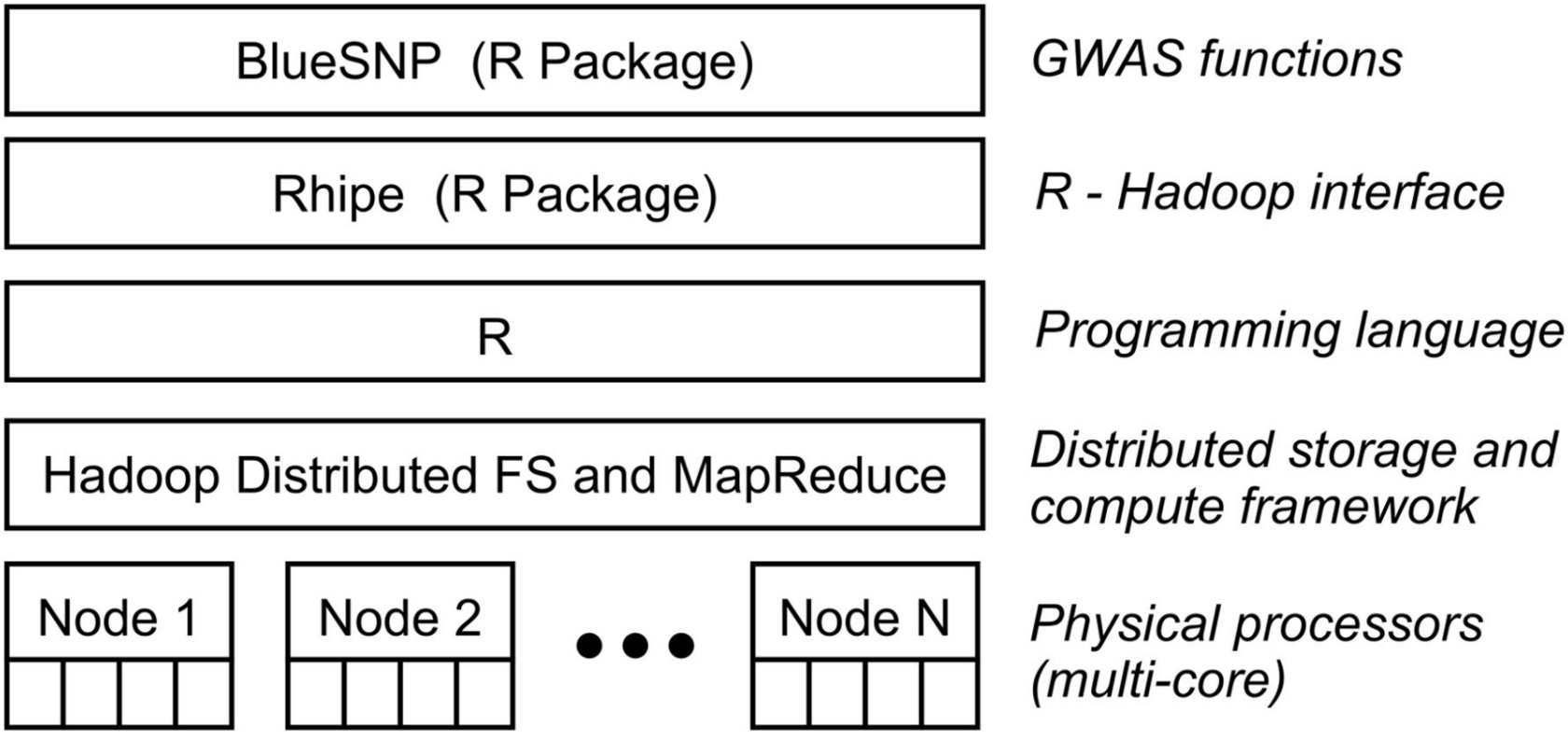
parallel example

```
library(parallel)
cl <- makeCluster(getOption("cl.cores", 4))
parApply(cl, d, 1, function(x) {
    while (1) {
    }
})
```

Hadoop


使用者的程式
Master機器
❶
❶
❷分配Map任務
❶
❷分配Reduce任務
HDFS
檔案系統
資料1
資料2
資料3
資料4
資料5
❸
❸
❸
❸
執行Map程式
(Worker機器)
執行Map程式
(Worker機器)
執行Map程式
(Worker機器)
❹
❹
❹
本地寫入
Map結果1
Map結果1
Map結果2
Map結果2
Map結果3
Map結果3
本地磁碟的
內部檔案
❺
❺
❺
❺
❺
遠端讀取
執行Reduce程式
(Worker機器)
執行Reduce程式
(Worker機器)
❻
❻
HDFS
檔案系統
輸出結果檔1
輸出結果檔2

Package: RHIPE



Package: RHadoop

https://github.com/RevolutionAnalytics/RHadoop/wiki

RHadoop

plyrmr	higher level plyr-like data processing for structured data, powered by rmr
rmr	functions providing Hadoop MapReduce functionality in R
rhdfs	functions providing file management of the HDFS from within R
rhbase	functions providing database management for the HBase distributed database from within R

Task View

http://cran.r-project.org/web/views/HighPerformanceComputing.html

Next

R Studio Server

http://www.rstudio.com/ide/download/server

Next

R is not the only choice