R and Bioconductor

Gang Chen
chengang@bgitechsolutions.com

November 22, 2014

华大科技 BGI·tech
香港中文大學 The Chinese University of Hong Kong

Outline

Next

R Package

- Hadley: In R, the fundamental unit of shareable code is the package.
- Hilary Parker: Seriously, it doesn’ t have to be about sharing your code (although that is an added benefit!). It is about saving yourself time.

References

- Writing R Extensions:
  http://cran.r-project.org/manuals.html
- R Packages from Hadley:
  http://r-pkgs.had.co.nz/
- Writing an R package from scratch:
  http://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/
- 开发R程序包之忍者篇:
  http://cos.name/2011/05/write-r-packages-like-a-ninja/

R Package from Scratch

see cgr directory

Why devtools?

- This book espouses my philosophy of package development:
- anything that can be automated, should be automated.
- Do as little as possible by hand.
- Do as much as possible with functions.

Next

Overview

Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 934 software packages, and an active user community.

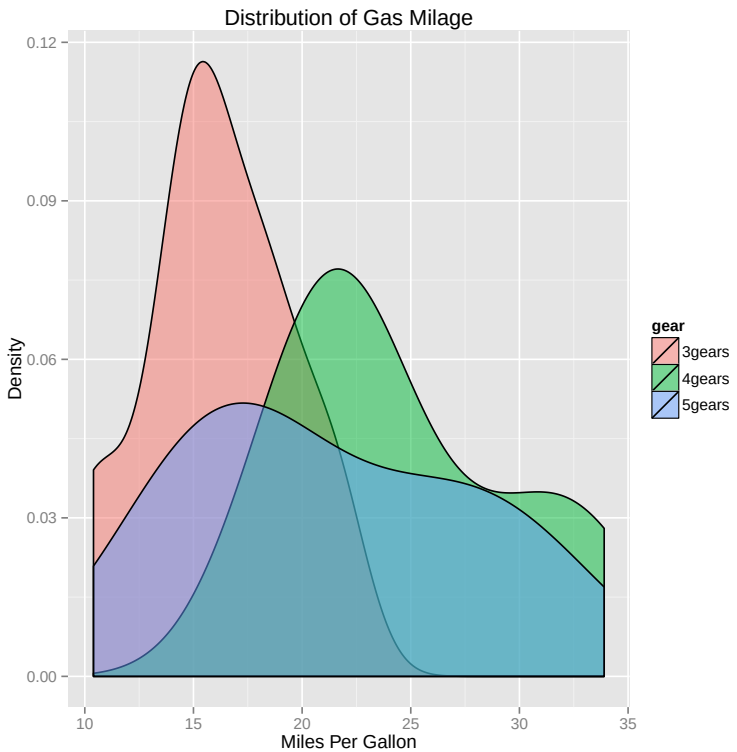
Goals

- To provide widespread access to a broad range of powerful statistical and graphical methods for the analysis of genomic data.
- To facilitate the inclusion of biological metadata in the analysis of genomic data, e.g. literature data from PubMed, annotation data from Entrez genes.
- To provide a common software platform that enables the rapid development and deployment of extensible, scalable, and interoperable software.
- To further scientific understanding by producing high-quality documentation and reproducible research.
- To train researchers on computational and statistical methods for the analysis of genomic data.

see GNBF5010-2013/4. Biological Data Analysis and Visualization in R/slides.pdf

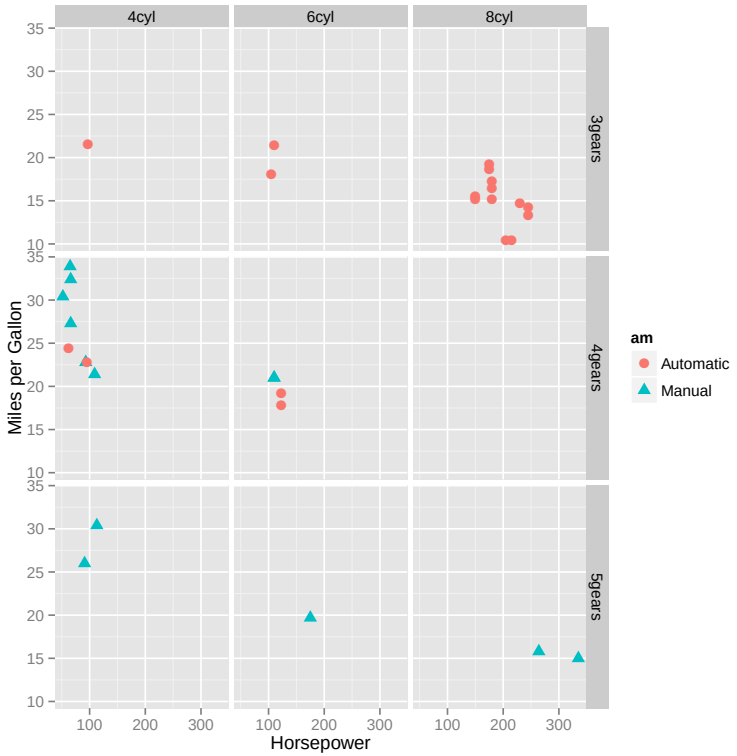
ggplot2

```
qplot(mpg, data=mtcars, geom="density", fill=gear, alpha=I(.
5), main="Distribution of Gas Milage", xlab="Miles Per
Gallon", ylab="Density")
```



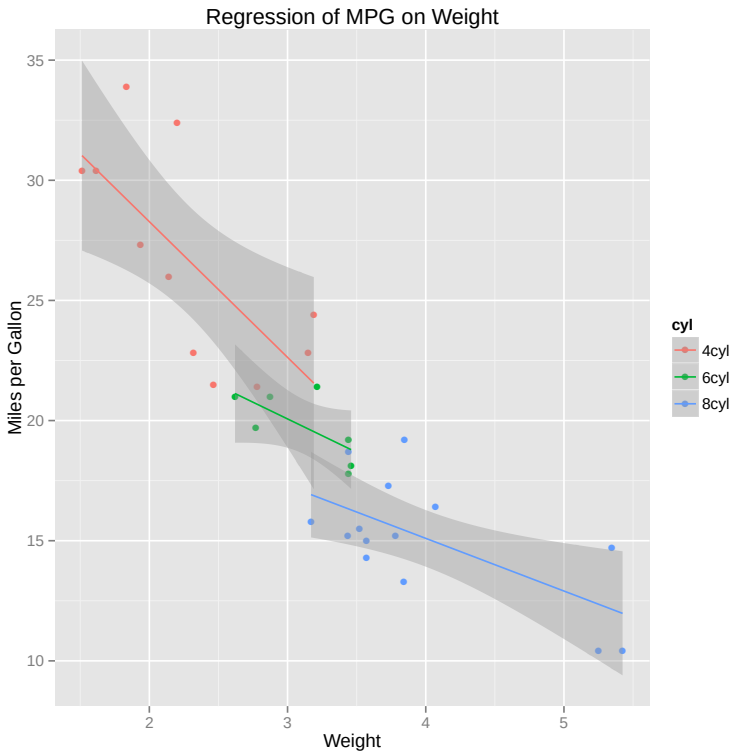
ggplot2

```
qplot(hp, mpg, data=mtcars, shape=am, color=am, facets=gear ~
size=I(3), xlab="Horsepower", ylab="Miles per Gallon")
```



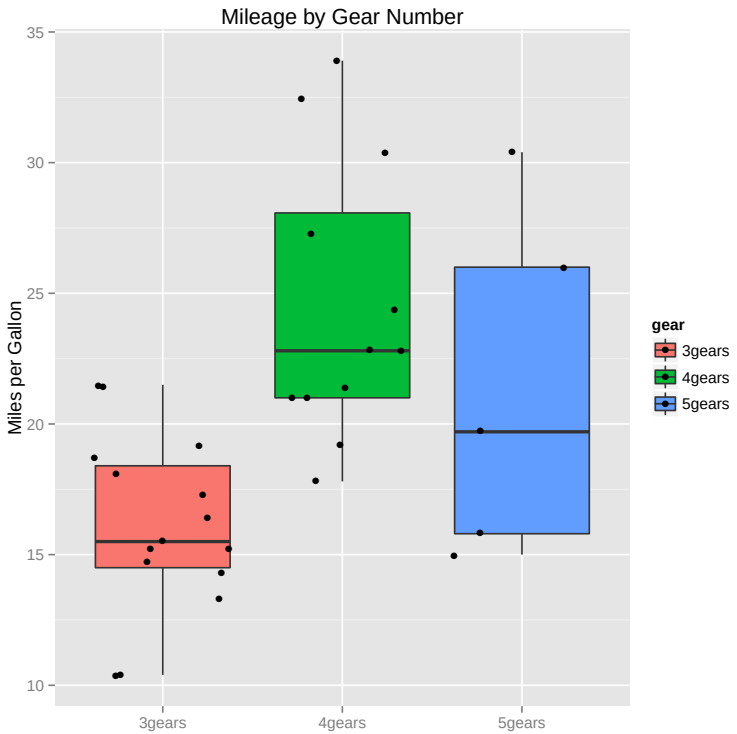
ggplot2

```
qplot(wt, mpg, data=mtcars, geom=c("point", "smooth"),
method="lm", formula=y x, color=cyl, main="Regression
of MPG on Weight", xlab="Weight", ylab="Miles per Gallon")
```



ggplot2

```
qplot(gear, mpg, data=mtcars, geom=c("boxplot", "jitter"),
fill=gear, main="Mileage by Gear Number", xlab="", ylab="Mile
per Gallon")
```



ggbio

- ggplot2 + bioconductor = ggbio
- Website: http://www.tengfei.name/ggbio/
- Author: Tengfei Yin at Seven Bridges Genomics

ggbio Examples

```
source("http://bioconductor.org/biocLite.R")
biocLite("ggbio")
library(ggbio)
example(autoplot)
```

Next

Overview

- Official website: http://yihui.name/knitr/
- Reference: Dynamic Documents with R and knitr
- Author: Yihui Xie

Examples

- the slides of the R lectures are generated by knitr
  - Knitr
  - XeLaTeX
- see knitr directory for the Knitr example in Markdown

Interactive Report

- Google Analytics
- 百度统计 from Baidu.com
- 数据魔方 from Taobao.com

Shiny Overview

- A web application framework for R
- Turn your analyses into interactive web applications
- No HTML, CSS, or JavaScript knowledge required
- http://shiny.rstudio.com/

Example

see shinyApp directory

Next

Machine Learning

- Deep learning
- Support Vector Machine
- Decision Tree
- Recommendation System
- Exper System: Computer-aided diagnosis

Machine Learning in R

- Task View: Machine Learning & Statistical Learning http://cran.r-project.org/web/views/MachineLearning.html
  - RWeka
  - Rattle
  - e1071, C50, randomForest
  - ...

High Performance Computing

- Task View: High-Performance and Parallel Computing with R
  http://cran.r-project.org/web/views/HighPerformanceComputing.html
  - Revolution R Enterprise
  - Rcpp
  - Multi-core
  - GPU
  - MPI

Big Data Framework

- Hadoop:
- Spark: SparkR, AMPLab UC BERKELEY
  http://amplab-extras.github.io/SparkR-pkg/
- Storm:
  - https://github.com/allenday/R-Storm
  - https://github.com/quintona/storm-r