Chapter 3
Navigating the Choices in R and Cloud Computing

3.1 Which Version of R to Use?

You should always use the latest version of R unless it breaks one of your packages that is needed for critical usage. In that case you are advised to both write to package maintainer as well as keep multiple versions of R. Some other versions of R are pqR Renjin and Riposte but they are not as mature right now. To help you keep your software updated, consider the installr package (covered later in this chapter)

We also show you two other forks of R or non-mainstream versions of R. Maybe they suit your needs or you can contribute to them!

3.1.1 Renjin

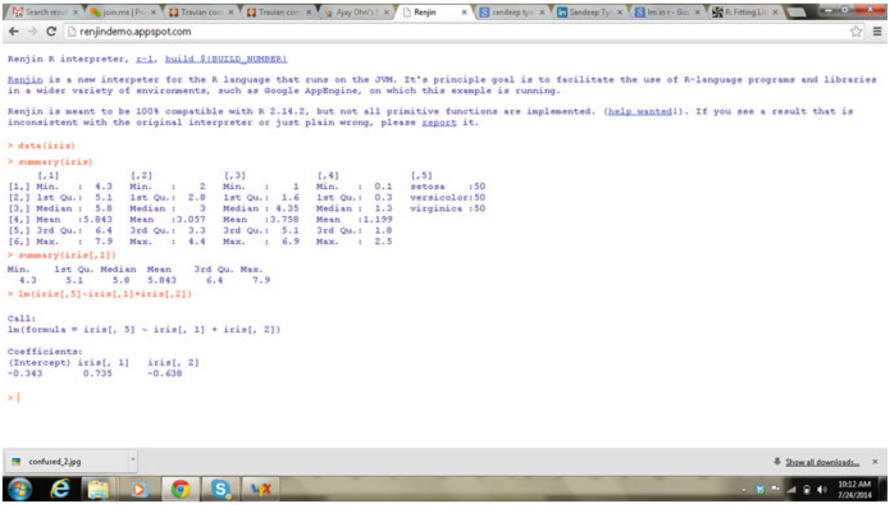
Upcoming is a project called Renjin. Renjin is a new implementation of the R language and environment for the Java Virtual Machine (JVM), whose goal is to enable transparent analysis of Big Data sets and seamless integration with other enterprise systems such as databases and application servers. See the demo at http://renjindemo.appspot.com/. However, Renjin is not yet production ready.

3.1.2 pqR

pqR is a new version of the R interpreter. It is based on R-2.15.0, distributed by the R Core Team (at r-project.org), but improves on it in many ways, mostly ways that speed it up, but also by implementing some new features and fixing some bugs.

One notable improvement is that pqR is able to do some numeric computations in parallel with each other, and with other operations of the interpreter, on systems with multiple processors or processor cores. It is created by Radford Neal (http://www.

A Ohri, *R for Cloud Computing*, DOI 10.1007/978-1-4939-1702-0_3



Renjin R interpreter, r-1, build ${BUILD_NUMBER}

Renjin is a new interpeter for the R language that runs on the JVM. It's principle goal is to facilitate the use of R-language programs and libraries in a wider variety of environments, such as Google AppEngine, on which this example is running.

Renjin is meant to be 100% compatible with R 2.14.2, but not all primitive functions are implemented. (help wanted!). If you see a result that is inconsistent with the original interpreter or just plain wrong, please report it.

```
> data(iris)
> summary(iris)
     [,1]            [,2]            [,3]            [,4]            [,5]
[1,] Min.   : 4.3    Min.   :  2     Min.   :  1     Min.   : 0.1    setosa    :50
[2,] 1st Qu.: 5.1    1st Qu.: 2.8    1st Qu.: 1.6    1st Qu.: 0.3    versicolor:50
[3,] Median : 5.8    Median :  3     Median : 4.35   Median : 1.3    virginica :50
[4,] Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
[5,] 3rd Qu.: 6.4    3rd Qu.: 3.3    3rd Qu.: 5.1    3rd Qu.: 1.8
[6,] Max.   : 7.9    Max.   : 4.4    Max.   : 6.9    Max.   : 2.5
> summary(iris[,1])
Min.    1st Qu. Median  Mean    3rd Qu. Max.
  4.3     5.1     5.8   5.843     6.4     7.9
> lm(iris[,5]~iris[,1]+iris[,2])

Call:
lm(formula = iris[, 5] ~ iris[, 1] + iris[, 2])

Coefficients:
(Intercept) iris[, 1]   iris[, 2]
-0.343        0.735      -0.638

> |
```

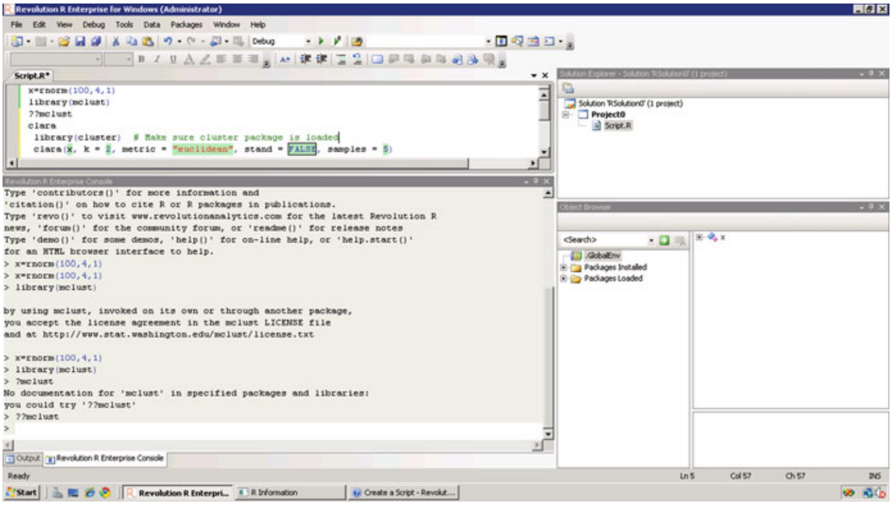
cs.utoronto.ca/~radford/homepage.html) and is available at http://www.pqr-project.org/

In addition to this there is the Riposte project (see Riposte: a trace-driven compiler and parallel VM for vector code in R http://dl.acm.org/citation.cfm?id=2370825). However, the github repository has not been updated in a year

3.2 Which Interface of R to Use

http://rforanalytics.wordpress.com/useful-links-for-r/code-enhancers-for-r/

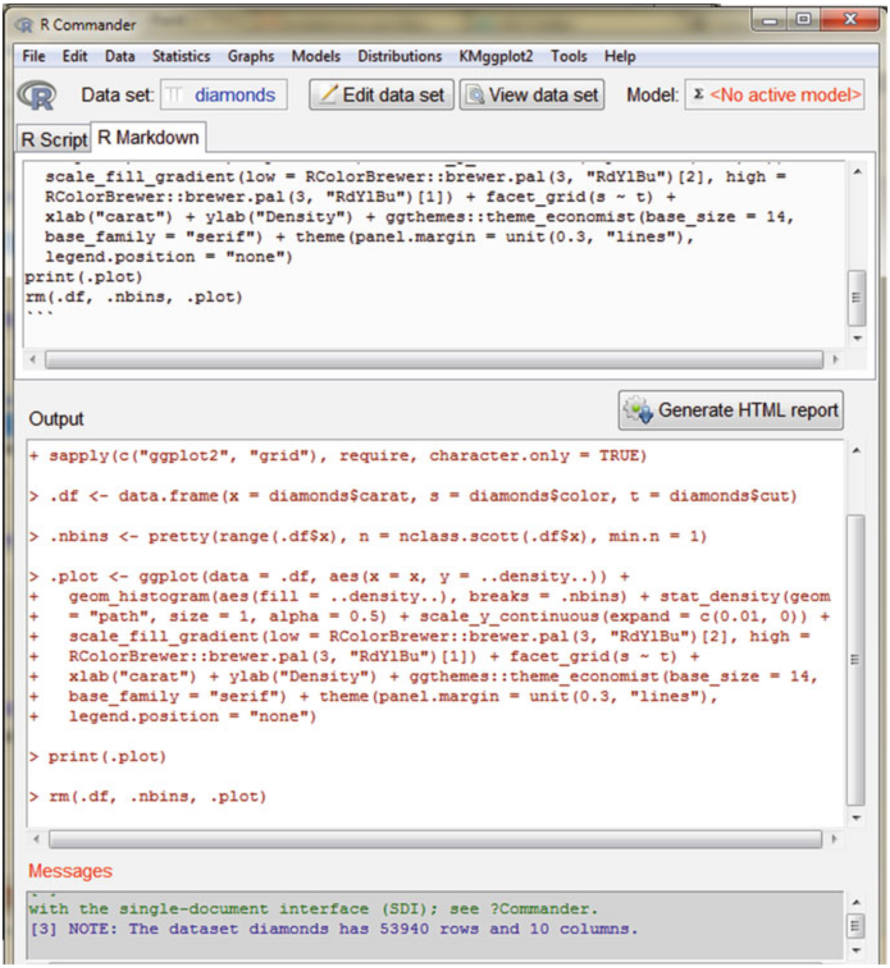
- Integrated Development Environment (IDE)
  - RStudio—One of the most popular IDEs used by developers in the R community, and it makes developing R code extremely easy and professional.
  - Eclipse with statet http://www.walware.de/goto/statet StatET is an Eclipse based IDE (integrated development environment) for R. It offers a set of mature tools for R coding and package building. This includes a fully integrated R Console, Object Browser, and R Help System, whereas multiple local and remote installations of R are supported. StatET is provided as plug-in for the Eclipse IDE.
  - Revolution Analytics—Revolution R Enterprise DevelopR (earlier R Productivity Environment) created by Revolution Analytics is a sturdy enterprise grade software environment. A notable innovation is the use of code snippets to help write code. A partial screenshot is shown here showing the snippets facility.


Revolution R Enterprise for Windows (Administrator)
File Edit View Debug Tools Data Packages Window Help
Debug
Script.R*
Solution Explorer - Solution 'RSolution0' (1 project)
Solution 'RSolution0' (1 project)
Project0
Script.R
Object Browser
<Search>
.GlobalEnv
Packages Installed
Packages Loaded
Revolution R Enterprise Console
Output
Revolution R Enterprise Console
Ready
Ln 5
Col 57
Ch 57
INS
Start
Revolution R Enterpri...
R Information
Create a Script - Revolut...

```
x=rnorm(100,4,1)
library(mclust)
??mclust
clara
 library(cluster)  # Make sure cluster package is loaded
 clara(x, k = 2, metric = "euclidean", stand = FALSE, samples = 5)
```

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'revo()' to visit www.revolutionanalytics.com for the latest Revolution R
news, 'forum()' for the community forum, or 'readme()' for release notes
Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()'
for an HTML browser interface to help.
> x=rnorm(100,4,1)
> x=rnorm(100,4,1)
> library(mclust)

by using mclust, invoked on its own or through another package,
you accept the license agreement in the mclust LICENSE file
and at http://www.stat.washington.edu/mclust/license.txt

> x=rnorm(100,4,1)
> library(mclust)
> ?mclust
No documentation for 'mclust' in specified packages and libraries:
you could try '??mclust'
> ??mclust
>
```

- Console—This is the default interface.
- GUIs
  - Rattle—Rattle is a GUI focused at data mining. We cover an example of Rattle in Chap. 5.
  - R Commander—This is one of the most widely used GUIs in R project. It is excellent for beginners. In addition it has a huge number of plug-ins that extend its capability. Notable plug-ins are epack (for time series) and kmggplot2 (for enhanced data visualization).
  - Deducer—Deducer is a GUI focused at data visualization. We cover an example of Deducer in Chap. 5.
- Online Versions of R-
  - Statace
  - R Fiddle
  - OpenCPU
- Commercial Versions
  - Revolution R
  - Oracle R
  - Tibco R



```
  scale_fill_gradient(low = RColorBrewer::brewer.pal(3, "RdYlBu")[2], high =
  RColorBrewer::brewer.pal(3, "RdYlBu")[1]) + facet_grid(s ~ t) +
  xlab("carat") + ylab("Density") + ggthemes::theme_economist(base_size = 14,
  base_family = "serif") + theme(panel.margin = unit(0.3, "lines"),
  legend.position = "none")
print(.plot)
rm(.df, .nbins, .plot)
```
```

```
+ sapply(c("ggplot2", "grid"), require, character.only = TRUE)

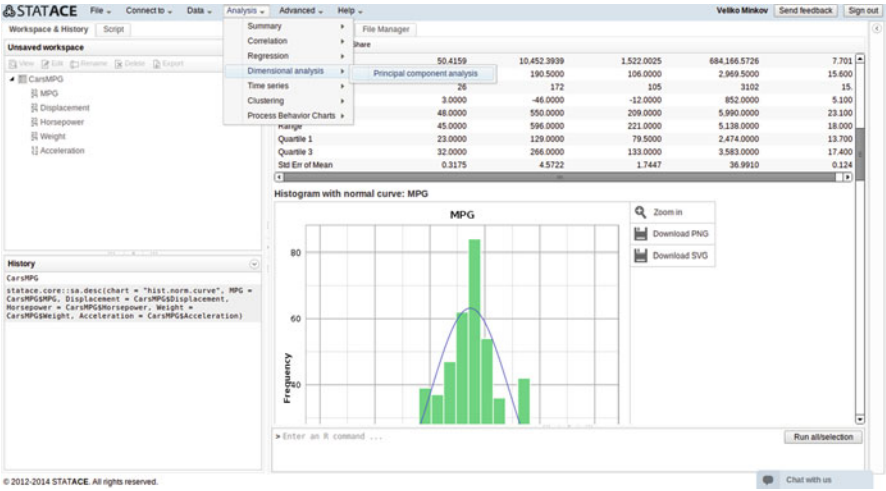
> .df <- data.frame(x = diamonds$carat, s = diamonds$color, t = diamonds$cut)

> .nbins <- pretty(range(.df$x), n = nclass.scott(.df$x), min.n = 1)

> .plot <- ggplot(data = .df, aes(x = x, y = ..density..)) +
+   geom_histogram(aes(fill = ..density..), breaks = .nbins) + stat_density(geom
+   = "path", size = 1, alpha = 0.5) + scale_y_continuous(expand = c(0.01, 0)) +
+   scale_fill_gradient(low = RColorBrewer::brewer.pal(3, "RdYlBu")[2], high =
+   RColorBrewer::brewer.pal(3, "RdYlBu")[1]) + facet_grid(s ~ t) +
+   xlab("carat") + ylab("Density") + ggthemes::theme_economist(base_size = 14,
+   base_family = "serif") + theme(panel.margin = unit(0.3, "lines"),
+   legend.position = "none")

> print(.plot)

> rm(.df, .nbins, .plot)
```

```
with the single-document interface (SDI); see ?Commander.
[3] NOTE: The dataset diamonds has 53940 rows and 10 columns.
```

3.3 Using R from the Browser

3.3.1 Statace

Statace is a web-based platform that empowers data scientists to collaborate with each other and build graphical self-service applications for people with no knowledge of statistics. Each application is automatically available through an API for easy integration with existing systems.

is an extract from an interview with its CEO, Christian Mladenov.

Ajay Ohri (AO)—What is the difference between using R by StatAce and using R by RStudio on a R Studio server hosted on Amazon EC2?
```



	50.4159	10,452.3939	1,522.0025	684,166.5726	7.701
		190.5000	106.0000	2,969.5000	15.600
	26	172	105	3102	15.
	3.0000	-46.0000	-12.0000	852.0000	5.100
	48.0000	550.0000	209.0000	5,990.0000	23.100
Range	45.0000	596.0000	221.0000	5,138.0000	18.000
Quartile 1	23.0000	129.0000	79.5000	2,474.0000	13.700
Quartile 3	32.0000	266.0000	133.0000	3,583.0000	17.400
Std Err of Mean	0.3175	4.5722	1.7447	36.9910	0.124

```
CarsMPG
statace.core::sa.desc(chart = "hist.norm.curve", MPG =
CarsMPG$MPG, Displacement = CarsMPG$Displacement,
Horsepower = CarsMPG$Horsepower, Weight =
CarsMPG$Weight, Acceleration = CarsMPG$Acceleration)
```

Christian Mladenov (CM)—There are a few ways in which I think StatAce is better:

You do not need the technical skills to set up a server. You can instead start straight away at the click of a button.

You can save the full results for later reference. With an RStudio server you need to manually save and organize the text output and the graphics.

We are aiming to develop a visual interface for all the standard stuff. Then you will not need to know R at all.

We are developing features for collaboration, so that you can access and track changes to data, scripts, and results in a team. With an RStudio server, you manage commits yourself, and Git is not suitable for large data files.

AO—How do you aim to differentiate yourself from other providers of R based software including Revolution, RStudio, Rapporter, and even Oracle R Enterprise?

CM—We aim to build a scalable, collaborative and easy to use environment. Pretty much everything else in the R ecosystem is lacking one, if not two of these. Most of the GUIs lack a visual way of doing the standard analyses. The ones that have it (e.g., Deducer) have a rather poor usability. Collaboration tools are hardly built in. RStudio has Git integration, but you need to set it up yourself, and you cannot really track large source data in Git.

Revolution Analytics have great technology, but you need to know R and you need to know how to maintain servers for large scale work. It is not very collaborative and can become quite expensive.

Rapporter is great for generating reports, but it is not very interactive—editing templates is a bit cumbersome if you just need to run a few commands. I think it wants to be the place to go to after you have finalized the development of the R code, so that you can share it. Right now, I also do not see the scalability.

With Oracle R Enterprise you again need to know R. It is targeted at large enterprises and I imagine it is quite expensive, considering it only works with Oracle's database. For that you need an IT team. Screenshot from 2013-11-18 21:31:08

AO—How do you see the space for using R on a cloud?

CM—I think this is an area that has not received enough quality attention—there are some great efforts (e.g. ElasticR), but they are targeted at experienced R users. I see a few factors that facilitate the migration to the cloud:

Statisticians collaborate more and more, which means they need to have a place to share data, scripts, and results.

The number of devices people use is increasing, and now frequently includes a tablet. Having things accessible through the web gives more freedom.

More and more data lives on servers. This is both because it is generated there (e.g., click streams) and because it is too big to fit on a user's PC (e.g., raw DNA data). Using it where it already is prevents slow download/upload.

Centralizing data, scripts, and results improves compliance (everybody knows where it is), reproducibility, and reliability (it is easily backed up).

For me, having R to the cloud is a great opportunity.

AO—What are some of the key technical challenges you currently face and are seeking to solve for R based cloud solutions?

CM—Our main challenge is CPU use, since cloud servers typically have multiple slow cores and R is mostly single-threaded. We have yet to fully address that and are actively following the projects that aim to improve R's interpreter—pqR, Renjin, Riposte, etc. One option is to move to bare metal servers, but then we will lose a lot of flexibility.

Another challenge is multi-server processing. This is also an area of progress where we do not yet have a stable solution.

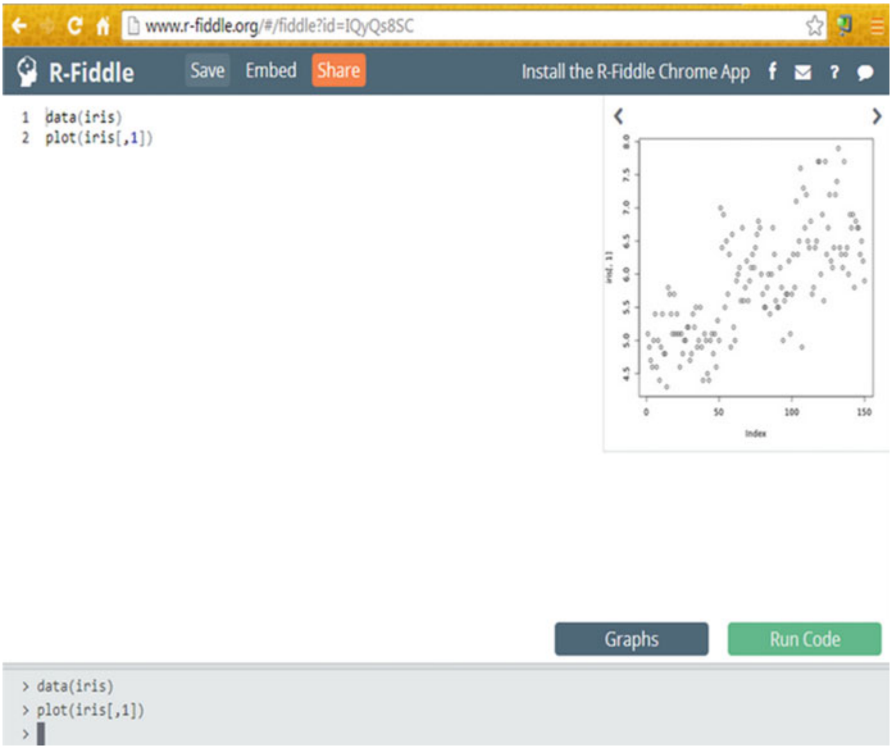
3.3.2 R Fiddle

R Fiddle is another project that enables people to start using R from within the browser. Here is a partial extract from an interview with Dr Jonathan Cornelissen, CEO of Datamind which also makes RDocumentation and R-fiddle.

Ajay—Congrats for making on the first page of hacker news with R-Fiddle. What were your motivations for making http://www.r-fiddle.org/?

Jonathan—Thank you. I must admit it was very exciting to be mentioned on Hacker News, since a lot of people were exposed to the R-fiddle project immediately. In addition, it was a first good test on how our servers would perform.

The motivation for building R-fiddle was simple; our CTO Dieter frequently uses the popular sitehttp://jsfiddle.net/ to prototype webpages, and to share his coding ideas with us. We were looking for something similar for R but it turned out a website allowing you to quickly write, run, and share R-code right inside your browser that did not exist yet. Since we were convinced a fiddle-like tool for R



```
data(iris)
plot(iris[,1])
```

```
> data(iris)
> plot(iris[,1])
>
```

would be useful, we just started building it. Based on the positive reactions and the fast adoption of R-fiddle, I think we were right. That being said, this is a first version of R-fiddle, and we will definitely improve it over the coming months. Check out our blog for updates. (http://blog.datamind.org/)

Ajay—Why did you make http://www.rdocumentation.org/ given that there is so much #Rstats documentation all around the internet including http://www.inside-r.org/?

Jonathan—When we started working on the www.datamind.org platform, we did an online survey to find out whether there would be interest in an interactive learning platform for R and statistics. Although the survey was not on this topic, one of the most striking findings was that a large percentage of R users apparently is frustrated about the documentation of R and its packages. This is interesting since it not only frustrates current users, but it also increases the barrier to entry for new R users and hence puts a brake on the growth and adoption of R as a language. It is mainly for the latter reason we started building Rdocumentation. The whole focus is on usability and letting all users contribute to make the documentation stronger. By the end of next week, we will launch a new version of Rdocumentation, that introduces advanced search functionality for all R functions, shows the popularity of R packages and much more. So make sure to www.Rdocumentation.org for updates.

Ajay—What have been your responses to http://www.datamind.org/#/. Any potential content creation partners or even corporate partners like statistics.com, Revolution, RStudio, Mango etc.

Jonathan—The response to the beta version of DataMind has been great, thousands of learners signed up and took the demo course. We are talking to some of the leading companies in the space and some very well-known professors to develop courses together. It is too soon to disclose details, but we will put regular updates on www.datamind.org. Corporates interested in what we do should definitely get in contact with Martijn@datamind.org.

Ajay—Would it be accurate to call http://www.r-fiddle.org/#/ a browser based GUI for R on the cloud? What enhancements can we expect in the future?

Jonathan—R-fiddle is indeed a browser based GUI for R on the cloud. We have a lot of ideas to improve and extend it. Some of the ideas are: the ability for users to concurrently make changes to a fiddle (Google-docs-style), support for loading data sets, github integration, better security management, lists of popular fiddles, or fiddles from popular people, etc. However, the strong point about R-fiddle is that it is really simple and there is absolutely no friction to start using it. In that respect, we want to differentiate R-fiddle from more advanced solutions such as StatAce or Rstudio Server, which focus on more advanced R users or R usage.

Ajay—You described your architecture for datamind.org at http://blog.datamind.org/ which is very open and transparent of you. What is the architecture for http://www.r-fiddle.org/#/ and what is it based out of?

Jonathan—That is an easy one. Although some details differ obviously, from a high-level perspectiveDataMind.org and R-fiddle.org have exactly the same IT architecture.

Ajay—http://www.datamind.org/#/dashboard describes course creation. How many courses are in the pipeline and how many users and corporate training clients do you foresee in the next 12 months?

Jonathan—Since we launched DataMind, we were inundated by requests from teachers and industry experts eager to contribute their own coursework on the site. But up until last week, it was only possible to take courses instead of creating them yourself. We decided to change this since we do not want to be solely a content company, but also a platform for others to create courses.

Furthermore, by expanding DataMind with a content creation tool, we go beyond our naturally limited in-house ability to create courses. Now DataMind is ready to become a full on ecosystem to facilitate education between our users.

Ajay—What do you think of R in the cloud for teaching (http://blog.datamind.org/2013/07/23/how-to-run-r-in-the-cloud-for-teaching/) ?

Jonathan—We are convinced that cloud solutions are the future of teaching and learning in general. The main problem with the first wave of online education solutions (such as Coursera, EdX, Udacity, etc.) is that they "only" make a copy of the classroom online instead of leveraging technology to create a more engaging and efficient learning experience and interface. Therefore, I do not think the future is in generic learning solutions. Learning interfaces will differ from domain to domain.

Good examples are: Duolingo.com to learn languages, or Codeschool.com to learn web development. We are on a mission to build the best learning solutions for statistics and data science.

Ajay—What are some of the other ways we can help make R more popular on the cloud?

Jonathan—I really like the vision behind StatAce.com, and I think something like it will definitely increase further adoption of R. It is somewhat surprising that Rstudio is not offering something like that, but my assumption is they are working on it. That being said, what would be really cool is a very easy-to-use graphical user interface with R under the hood. Whether you like it or not, R has quite a steep learning curve for most people, and allowing them to analyse data with R through a graphical user interface on the web as a first step could start the adoption of R in less technical areas.

3.3.3 RShiny and RStudio

Shiny is a web application framework for R that helps turn analytics into interactive web apps (without HTML, CSS, or JavaScript knowledge). It is available at http://shiny.rstudio.com/. An example of an application from its gallery is below.

Here is an interview with Jeff Allen who works with R and the new package Shiny in his technology startup. His RShiny application can be seen at http://trestletechnology.net:3838/grn/.

Ajay—Describe how you started using R. What are some of the benefits you noticed on moving to R?

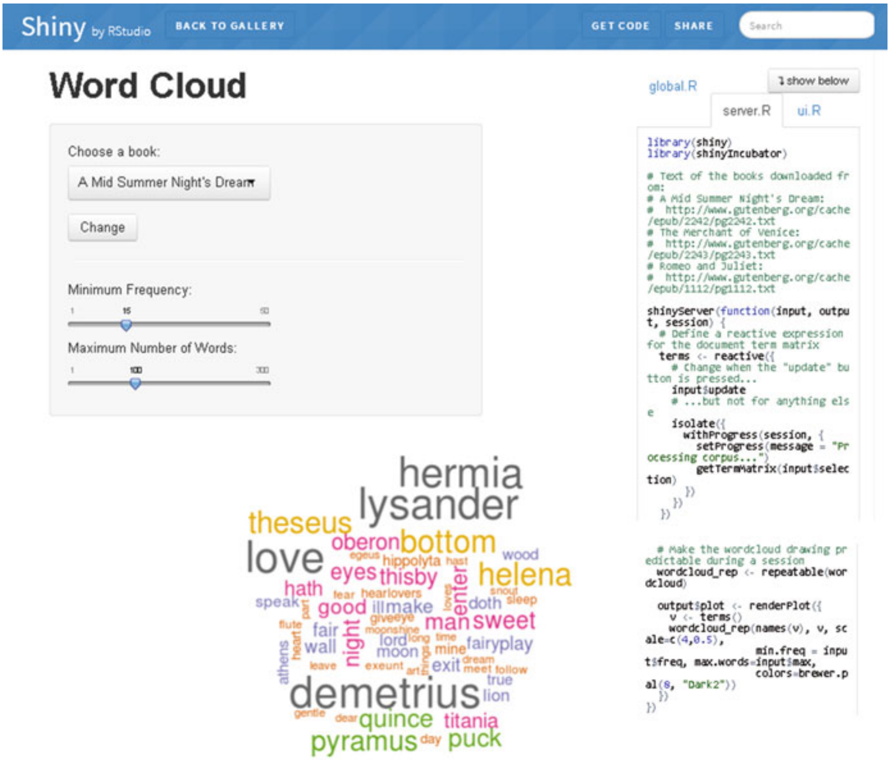
Jeff—I began using R in an internship while working on my undergraduate degree. I was provided with some unformatted R code and asked to modularize the code and then wrap it up into an R package for distribution alongside a publication.

To be honest, as a Computer Science student with training more heavily emphasizing the big high-level languages, R took some getting used to for me. It was not until after I concluded that initial project and began using R to do my own data analysis that I began to realize its potential and value. It was the first scripting language which really made interactive use appealing to me—the experience of exploring a dataset in R was unlike anything I would been taught in my University courses.

Upon gaining familiarity with the syntax and basics of the language, I began to see the immense value in the vast array of R packages which had already been created and made publicly available. I found repeatedly that many of the "niche" functions I would been coding myself in more traditional languages had already been developed and shared freely on CRAN or Bioconductor.

Ajay—Describe your work in computational biology using R

Jeff—I work in the Quantitative Biomedical Research Center (QBRC) at UT Southwestern Medical Center. My group is involved in analysing and mining massive biological datasets, much of which is now coming from different sequencing



Choose a book: A Mid Summer Night's Dream

Minimum Frequency: 15

Maximum Number of Words: 100

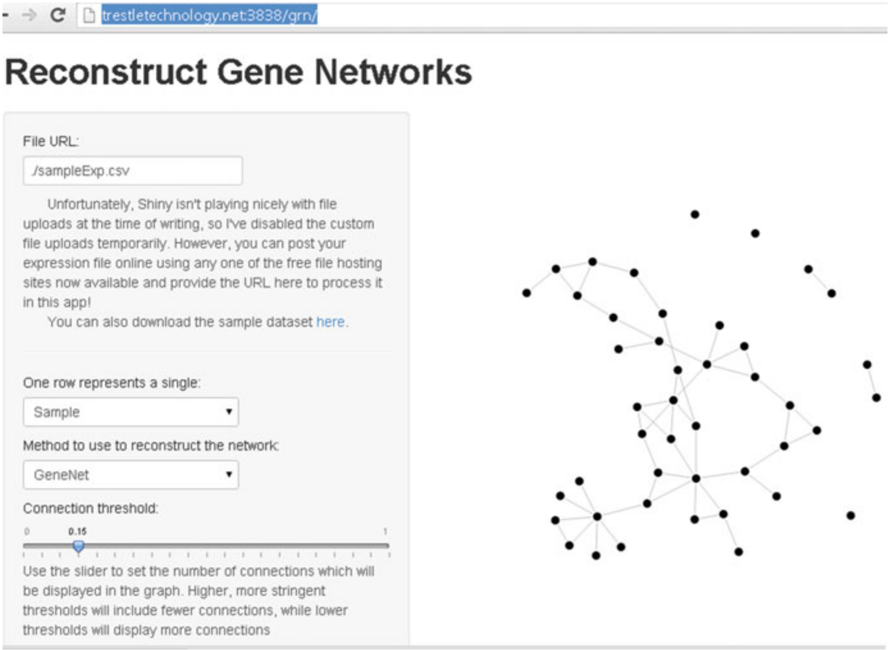
```
library(shiny)
library(shinyIncubator)

# Text of the books downloaded fr
om:
# A Mid Summer Night's Dream:
#  http://www.gutenberg.org/cache
/epub/2242/pg2242.txt
# The Merchant of Venice:
#  http://www.gutenberg.org/cache
/epub/2243/pg2243.txt
# Romeo and Juliet:
#  http://www.gutenberg.org/cache
/epub/1112/pg1112.txt

shinyServer(function(input, outpu
t, session) {
  # Define a reactive expression
for the document term matrix
  terms <- reactive({
    # Change when the "update" bu
tton is pressed...
    input$update
    # ...but not for anything els
e
    isolate({
      withProgress(session, {
        setProgress(message = "Pr
ocessing corpus...")
        getTermMatrix(input$selec
tion)
      })
    })
  })

  # Make the wordcloud drawing pr
edictable during a session
  wordcloud_rep <- repeatable(wor
dcloud)

  output$plot <- renderPlot({
    v <- terms()
    wordcloud_rep(names(v), v, sc
ale=c(4,0.5),
                  min.freq = inpu
t$freq, max.words=input$max,
                  colors=brewer.p
al(8, "Dark2"))
  })
})
```

technologies (DNA-seq, RNA-seq, etc.) which generate many Gigabytes of data with each experiment. Unfortunately, due to the sheer volume of that data, R is often unfit for the initial analysis and pre-processing of the data. However, once the data has been processed and reduced, we use R to identify statistically and (hopefully) biologically interesting trends or anomalies.

Personally, most of my research lately has focused on reconstructing the interactions between genes based on the patterns and behaviours we can observe. Thankfully, most of the data we work with here fits in memory, so I use R almost exclusively when doing analysis in this area. My most recent work was in “Ensemble Network Aggregation” (ENA), the package of which is now available in CRAN.

Ajay—Describe your work in web applications using R

Jeff—I was initially tasked with developing software packages which encapsulated the new statistical methodologies being developed within the group (which, at the time, were largely focused on microarray data). I continued developing R packages and began investigating how I might be able to integrate my prior experience with web development into these projects. We ended up developing a handful of different web applications which typically required that we use R to precompute any data which ultimately made its way (statically) into the application.



trestletechnology.net:3838/grn/

Reconstruct Gene Networks

File URL: ./sampleExp.csv

Unfortunately, Shiny isn't playing nicely with file uploads at the time of writing, so I've disabled the custom file uploads temporarily. However, you can post your expression file online using any one of the free file hosting sites now available and provide the URL here to process it in this app!

You can also download the sample dataset here.

One row represents a single: Sample

Method to use to reconstruct the network: GeneNet

Connection threshold: 0.15

Use the slider to set the number of connections which will be displayed in the graph. Higher, more stringent thresholds will include fewer connections, while lower thresholds will display more connections

More recently, we have been developing sites which take advantage of dynamic or real-time R analysis, such as our upcoming release of the Lung Cancer Explorer—a tool which allows for the interactive exploration of lung cancer data within a browser. We went to great lengths to develop the IT and software infrastructure that would allow us to interact with R remotely for these applications.

I have been taking some time on the side to play with RStudio's new Shiny application which, like most everything else that group has put out, represents a massive leap forward in this space. We have already begun looking at how we can supplement or replace some of our in-house systems with Shiny to start taking advantage of some of its capabilities.

Ajay—What is Trestle Technology focused on?

Jeff—I initially was doing a lot of web development, and helping small-medium businesses integrate and automate various software systems. Once R got added to my resume, however, I started finding more interesting work helping start-ups get their IT and analytics infrastructures off the ground.

My hope is to continue living at the intersection of data and software development and grow this company in that space. It is quite difficult to find groups doing good analysis and proper software development under one roof—especially in Academia. I thoroughly enjoy the process of enriching data analysis tools with more comprehensive, user-friendly interfaces which allow for more efficient exploration of the underlying datasets.

Ajay—Compare R and Python, What are some of the ways you think R can be further improved?

Jeff—I must confess that I am a fairly late-comer to the Python world. I had tinkered with Perl and Python a few years back and ended up feeling more comfortable with Perl for my needs at that point, so I have used it for most of my scripting. Only recently have I started revisiting Python and getting introduced to some of the neat tools available. To me, Python offers a more intuitive Object-Orienting framework than either Perl or R provides, which helps me stay more organized as my "scripts" inevitably grow into "software".

Aside from OO, I still feel there is much room for improvement in R's "Big Data" capabilities. The community has certainly been making huge strides in bypassing memory limitations (with packages like ff) and speeding up the code (with code compilation, etc). Nonetheless, I find that my motivation in leaving R for any part of my data analysis is typically motivated by performance concerns and a desire to avoid having to nest any C++ code in my R scripts (though the recent improvements in the Rcpp and devtools packages are making that a much less painful process).

About—

Jeffrey D Allen is a computational biologist at UT Southwestern Medical Center at Dallas.

To know more on the Shiny package and use it for interactive visualization please refer to the cheat sheet at http://shiny.rstudio.com/articles/cheatsheet.html.

3.4 The Cloud Computing Services Landscape

There are 3 main players in the infrastructure provider space for cloud computing world for data sciences. They are:

1. Amazon AWS
2. Microsoft Azure
3. Google Compute Cloud

Others are Oracle Java Cloud and IBM Softlayer

The easiest way to convince someone to start using R on the cloud remains using the Amazon Web Services with Windows Operating System and using a Remote Desktop Player to connect, and then use a graphical user interface (GUI) package of R like Deducer (data visualization), R Commander, or Rattle (Data Mining). This is the opinion of the author and it is based on his impression of the relative penetration of Linux in the world of business analytics as well as some of the issues that transitioning to Linux does on users. It is recommended that the user change one thing at a time, from hardware (PC or Cloud), operating system (Windows or Linux), and analytics software (R versus prior software). However the best option for a potential data science centre of excellence is to use Linux on the cloud albeit with the R Studio Server to make the transition easier.

3.4.1 *Choosing Infrastructure Providers*

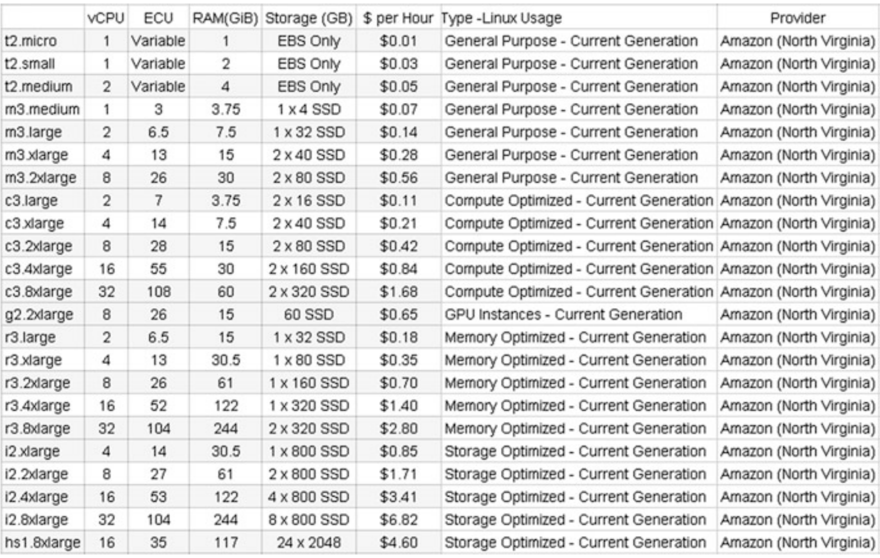
When it comes to choosing Infrastructure as a service, we can choose using the following options:

Cloud Provider	Strengths	Weakness
Amazon AWS (EC2)	Widest Range of Instances and Regions, Great Documentation	Loss of control
Microsoft Azure	Best Interface Design	Comparatively Expensive
Google Cloud (Compute Engine)	Secure, Cheap for Technically Advanced Users	No Windows Instances

If we do not want to use the cloud for our entire analytical needs, we can use it for computationally intensive things like running simulations or scoring models. This is where software that is offered as a service can help us. Note both Amazon, Microsoft, and Google offer more than just the cloud infrastructure as a service. It is quite useful to look at the complementary services that a cloud infrastructure provider provides, and choose based on suitability.

3.4.1.1 Amazon Cloud

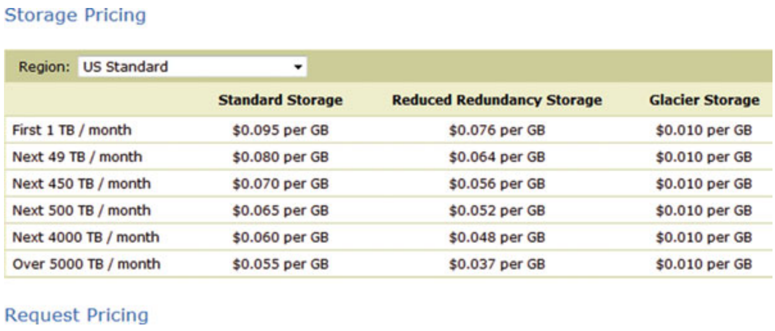
There are many kinds of instances in Amazon at http://aws.amazon.com/ec2/instance-types/. To remove confusion we have represented this in a mildly modified table.



	vCPU	ECU	RAM(GiB)	Storage (GB)	$ per Hour	Type -Linux Usage	Provider
t2.micro	1	Variable	1	EBS Only	$0.01	General Purpose - Current Generation	Amazon (North Virginia)
t2.small	1	Variable	2	EBS Only	$0.03	General Purpose - Current Generation	Amazon (North Virginia)
t2.medium	2	Variable	4	EBS Only	$0.05	General Purpose - Current Generation	Amazon (North Virginia)
m3.medium	1	3	3.75	1 x 4 SSD	$0.07	General Purpose - Current Generation	Amazon (North Virginia)
m3.large	2	6.5	7.5	1 x 32 SSD	$0.14	General Purpose - Current Generation	Amazon (North Virginia)
m3.xlarge	4	13	15	2 x 40 SSD	$0.28	General Purpose - Current Generation	Amazon (North Virginia)
m3.2xlarge	8	26	30	2 x 80 SSD	$0.56	General Purpose - Current Generation	Amazon (North Virginia)
c3.large	2	7	3.75	2 x 16 SSD	$0.11	Compute Optimized - Current Generation	Amazon (North Virginia)
c3.xlarge	4	14	7.5	2 x 40 SSD	$0.21	Compute Optimized - Current Generation	Amazon (North Virginia)
c3.2xlarge	8	28	15	2 x 80 SSD	$0.42	Compute Optimized - Current Generation	Amazon (North Virginia)
c3.4xlarge	16	55	30	2 x 160 SSD	$0.84	Compute Optimized - Current Generation	Amazon (North Virginia)
c3.8xlarge	32	108	60	2 x 320 SSD	$1.68	Compute Optimized - Current Generation	Amazon (North Virginia)
g2.2xlarge	8	26	15	60 SSD	$0.65	GPU Instances - Current Generation	Amazon (North Virginia)
r3.large	2	6.5	15	1 x 32 SSD	$0.18	Memory Optimized - Current Generation	Amazon (North Virginia)
r3.xlarge	4	13	30.5	1 x 80 SSD	$0.35	Memory Optimized - Current Generation	Amazon (North Virginia)
r3.2xlarge	8	26	61	1 x 160 SSD	$0.70	Memory Optimized - Current Generation	Amazon (North Virginia)
r3.4xlarge	16	52	122	1 x 320 SSD	$1.40	Memory Optimized - Current Generation	Amazon (North Virginia)
r3.8xlarge	32	104	244	2 x 320 SSD	$2.80	Memory Optimized - Current Generation	Amazon (North Virginia)
i2.xlarge	4	14	30.5	1 x 800 SSD	$0.85	Storage Optimized - Current Generation	Amazon (North Virginia)
i2.2xlarge	8	27	61	2 x 800 SSD	$1.71	Storage Optimized - Current Generation	Amazon (North Virginia)
i2.4xlarge	16	53	122	4 x 800 SSD	$3.41	Storage Optimized - Current Generation	Amazon (North Virginia)
i2.8xlarge	32	104	244	8 x 800 SSD	$6.82	Storage Optimized - Current Generation	Amazon (North Virginia)
hs1.8xlarge	16	35	117	24 x 2048	$4.60	Storage Optimized - Current Generation	Amazon (North Virginia)

AWS Terminology:

- Instances—AMIs—An instance is like a virtual computer dedicated to you running remotely. Think of it as a combination of hardware choices (like how much RAM) and operating system choices (Windows, Red Hat Linux, or Ubuntu Linux). Each instance has a unique id called AMAZON INSTANCE ID.
- Elastic Block Storage (EBS)—Volumes—A volume is like a hard disk, or ROM Storage. Once you terminate an instance everything on the volume would be deleted. You can however stop and restart an EBS backed Instance
- Elastic Block Storage—Snapshots—Snapshots are like stored or saved volumes for further reference.
- Network and Security—Security Groups—Security groups are used for giving access to certain Ports, to certain access types (SSH, or HTTP, or HTTPS, or RDP). We will discuss these in further detail later.
- Network and Security—Elastic IPs—You can choose to refer to your amazon instance by its IP Address, by attaching an Elastic IP. This is free if attached to a running instance, but costs a small sum if reserved without a running instance. Once you release a certain IP address, it goes back to the pool of available IP addresses for Elastic IP on AWS.
- Amazon has s3 storage (http://aws.amazon.com/s3/) and glacier storage (http://aws.amazon.com/glacier/)



Storage Pricing

Region: US Standard

	Standard Storage	Reduced Redundancy Storage	Glacier Storage
First 1 TB / month	$0.095 per GB	$0.076 per GB	$0.010 per GB
Next 49 TB / month	$0.080 per GB	$0.064 per GB	$0.010 per GB
Next 450 TB / month	$0.070 per GB	$0.056 per GB	$0.010 per GB
Next 500 TB / month	$0.065 per GB	$0.052 per GB	$0.010 per GB
Next 4000 TB / month	$0.060 per GB	$0.048 per GB	$0.010 per GB
Over 5000 TB / month	$0.055 per GB	$0.037 per GB	$0.010 per GB

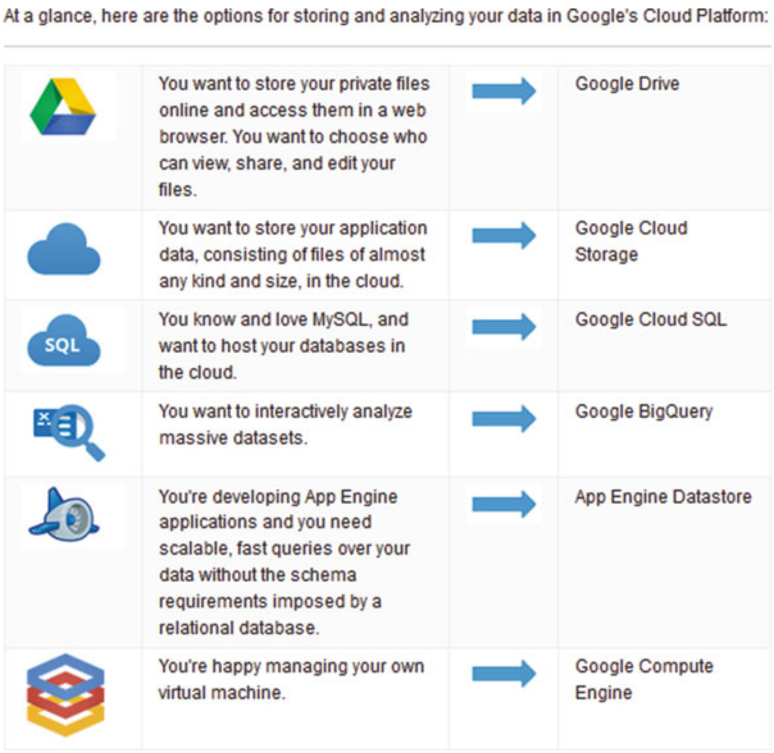
Request Pricing

3.4.1.2 Other Components of Amazon Cloud

Amazon Simple Storage deals with data storage. The vast offerings by Amazon can be summed up in this webpage http://aws.amazon.com/products/.

3.4.1.3 Google Cloud Services

A lucid infographic from https://cloud.google.com/developers/articles/storage-overview helps you understand various components of Google 's cloud offerings.



At a glance, here are the options for storing and analyzing your data in Google's Cloud Platform:

	You want to store your private files online and access them in a web browser. You want to choose who can view, share, and edit your files.	→	Google Drive
	You want to store your application data, consisting of files of almost any kind and size, in the cloud.	→	Google Cloud Storage
	You know and love MySQL, and want to host your databases in the cloud.	→	Google Cloud SQL
	You want to interactively analyze massive datasets.	→	Google BigQuery
	You're developing App Engine applications and you need scalable, fast queries over your data without the schema requirements imposed by a relational database.	→	App Engine Datastore
	You're happy managing your own virtual machine.	→	Google Compute Engine

This can be seen at https://cloud.google.com/. It has the following components.

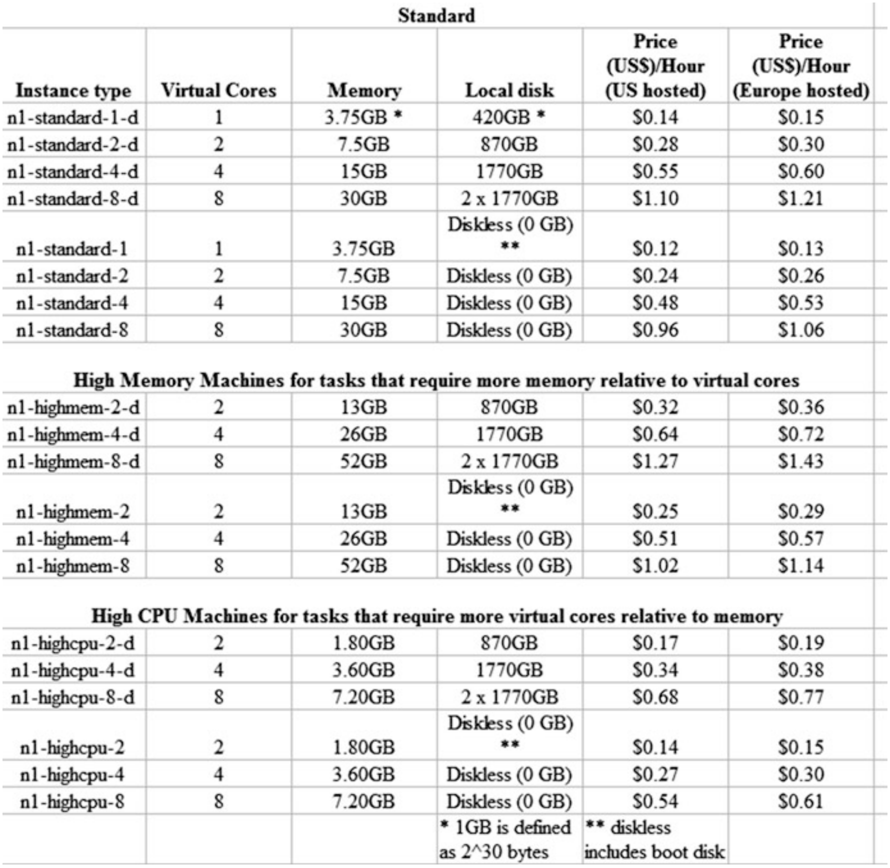
1. Application Platforms

   a. Google App Engine—Used for building web applications, mobile applications, gaming, as well as websites. You can see your existing Google Cloud Projects at https://cloud.google.com/console.

2. Computing Infrastructure

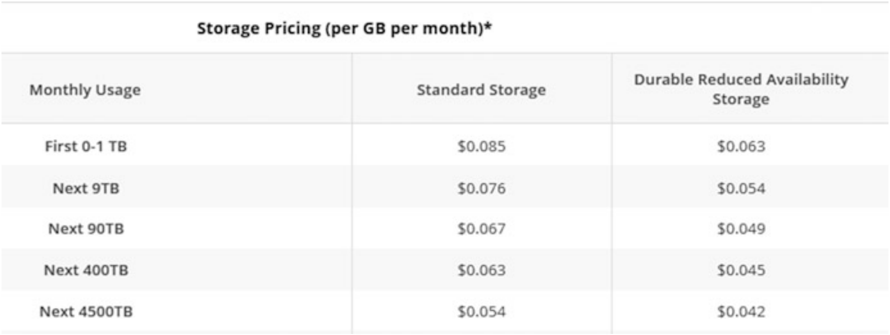
   a. Google Compute Engine—This is primary cloud infrastructure offering by Google. It was in limited preview at the time of writing this book, though the author is grateful to the team for the free access. However it is likely to change based on competition as well as feedback from users once it comes out of preview. One notable disadvantage is that the Windows operating system is completely ignored by Google Compute Engine. The pricing of Google

Compute Engine remains one of the most cost effective—as this chart shows for pricing dated 27 February 2013. Note that prices of cloud instances change after every few months, usually becoming cheaper.



Standard					
Instance type	Virtual Cores	Memory	Local disk	Price (US$)/Hour (US hosted)	Price (US$)/Hour (Europe hosted)
n1-standard-1-d	1	3.75GB *	420GB *	$0.14	$0.15
n1-standard-2-d	2	7.5GB	870GB	$0.28	$0.30
n1-standard-4-d	4	15GB	1770GB	$0.55	$0.60
n1-standard-8-d	8	30GB	2 x 1770GB	$1.10	$1.21
n1-standard-1	1	3.75GB	Diskless (0 GB) **	$0.12	$0.13
n1-standard-2	2	7.5GB	Diskless (0 GB)	$0.24	$0.26
n1-standard-4	4	15GB	Diskless (0 GB)	$0.48	$0.53
n1-standard-8	8	30GB	Diskless (0 GB)	$0.96	$1.06
High Memory Machines for tasks that require more memory relative to virtual cores					
n1-highmem-2-d	2	13GB	870GB	$0.32	$0.36
n1-highmem-4-d	4	26GB	1770GB	$0.64	$0.72
n1-highmem-8-d	8	52GB	2 x 1770GB	$1.27	$1.43
n1-highmem-2	2	13GB	Diskless (0 GB) **	$0.25	$0.29
n1-highmem-4	4	26GB	Diskless (0 GB)	$0.51	$0.57
n1-highmem-8	8	52GB	Diskless (0 GB)	$1.02	$1.14
High CPU Machines for tasks that require more virtual cores relative to memory					
n1-highcpu-2-d	2	1.80GB	870GB	$0.17	$0.19
n1-highcpu-4-d	4	3.60GB	1770GB	$0.34	$0.38
n1-highcpu-8-d	8	7.20GB	2 x 1770GB	$0.68	$0.77
n1-highcpu-2	2	1.80GB	Diskless (0 GB) **	$0.14	$0.15
n1-highcpu-4	4	3.60GB	Diskless (0 GB)	$0.27	$0.30
n1-highcpu-8	8	7.20GB	Diskless (0 GB)	$0.54	$0.61
			* 1GB is defined as 2^30 bytes	** diskless includes boot disk	

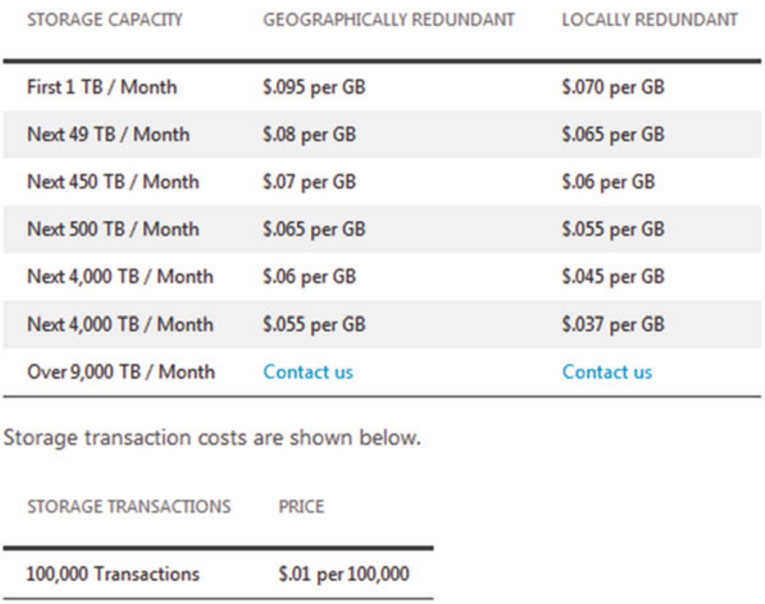
Google Cloud Storage is available at https://cloud.google.com/pricing/cloud-storage. It has two types of pricing including durable reduced availability storage and standard storage. It is Google's offering for cloud storage, but is different from the file storage offered by Google Drive. Durable Reduced Availability Storage enables you to store data at lower cost, with the tradeoff of lower availability than standard Google Cloud Storage.



Storage Pricing (per GB per month)*		
Monthly Usage	Standard Storage	Durable Reduced Availability Storage
First 0-1 TB	$0.085	$0.063
Next 9TB	$0.076	$0.054
Next 90TB	$0.067	$0.049
Next 400TB	$0.063	$0.045
Next 4500TB	$0.054	$0.042

3.4.1.4 Windows Azure

1. Windows Azure storage provides two levels of redundancy: Locally Redundant Storage (LRS)—provides highly durable and available storage within a single sub region. Geo Redundant Storage (GRS)—provides highest level of durability by additionally storing data in a second sub-region within the same region.



STORAGE CAPACITY	GEOGRAPHICALLY REDUNDANT	LOCALLY REDUNDANT
First 1 TB / Month	$.095 per GB	$.070 per GB
Next 49 TB / Month	$.08 per GB	$.065 per GB
Next 450 TB / Month	$.07 per GB	$.06 per GB
Next 500 TB / Month	$.065 per GB	$.055 per GB
Next 4,000 TB / Month	$.06 per GB	$.045 per GB
Next 4,000 TB / Month	$.055 per GB	$.037 per GB
Over 9,000 TB / Month	Contact us	Contact us

Storage transaction costs are shown below.

STORAGE TRANSACTIONS	PRICE
100,000 Transactions	$.01 per 100,000

2. Querying or Data Access on the Cloud:

   a. Google Big Query is available at https://cloud.google.com/products/big-query. BigQuery supports analysis of datasets up to hundreds of terabytes. BigQuery uses a columnar data structure and is a cloud enabled massively parallel query engine. You can read more on this at https://cloud.google.com/files/BigQueryTechnicalWP.pdf.

3.5 Interview Ian Fellows Deducer

Here is an interview with Dr Ian Fellows, creator of acclaimed packages in R like Deducer and the Founder and President of Fellstat.com

Ajay—Describe your involvement with the Deducer Project and the various plug-ins associated with it. What has been the usage and response for Deducer from R Community.

Ian—Deducer is a graphical user interface for data analysis built on R. It sprung out of a disconnect between the tool chain used by myself and the tool chain of the psychologists that I worked with at the University of California, San Diego. They were primarily SPSS user, whereas I liked to use R, especially for anything that was not a standard analysis. I felt that there was a big gap in the audience that R serves. Not all consumers or producers of statistics can be expected to have the computational background (command-line programming) that R requires. I think it is important to recognize and work with the areas of expertise that statistical users have. I am not an expert in psychology, and they did not expect me to be one. They are not experts in computation, and I do not think that we should expect them to be in order to be a part of the R tool chain community.ian. This was the impetus behind Deducer, so it is fundamentally designed to be a familiar experience for users coming from an SPSS background and provides a full implementation of the standard methods in statistics, and data manipulation from descriptives to generalized linear models. Additionally, it has an advanced GUI for creating visualizations which has been well received, and won the John Chambers award for statistical software in 2011.

Uptake of the system is difficult to measure as CRAN does not track package downloads, but from what I can tell there has been a steadily increasing user base. The online manual has been accessed by over 75,000 unique users, with over 400,000 page views. There is a small, active group of developers creating add-on packages supporting various sub-disciplines of statistics. There are 8 packages on CRAN extending/using Deducer, and quite a few more on r-forge. Ajay—Do you see any potential for Deducer as an enterprise software product (like R Studio et al) Ian Like R Studio, Deducer is used in enterprise environments but is not specifically geared towards that environment. I do see potential in that realm, but do not have any particular plan to make an enterprise version of Deducer.

Ajay—Describe your work in Texas Hold'em Poker. Do you see any potential for R for diversifying into the casino analytics—which has hitherto been served exclusively by non open source analytics vendors?

Ian—As a Statistician, I am very much interested in problems of inference under uncertainty, especially when the problem space is huge. Creating an Artificial Intelligence that can play (heads-up limit) Texas Hold'em Poker at a high level is a perfect example of this. There is uncertainty created by the random drawing of cards, the problem space is 10^{18}, and our opponent can adapt to any strategy that we employ. While high level chess A.I.s have existed for decades, the first viable program to tackle full scale poker was introduced in 2003 by the incomparable

Computer Poker Research group at the University of Alberta. Thus poker represents a significant challenge which can be used as a test bed to break new ground in applied game theory. In 2007 and 2008 I submitted entries to the AAA's annual computer poker competition, which pits A.I.s from universities across the world against each other. My program, which was based on an approximate game theoretic equilibrium calculated using a co-evolutionary process called fictitious play, came in second behind the Alberta team.

Ajay—Describe your work in social media analytics for R. What potential do you see for Social Network Analysis given the current usage of it in business analytics and business intelligence tools for enterprise?

Ian—My dissertation focused on new model classes for social network analysis (http://arxiv.org/pdf/1208.0121v1.pdf and http://arxiv.org/pdf/1303.1219.pdf). R has a great collection of tools for social network analysis in the statnet suite of packages, which represents the forefront of the literature on the statistical modeling of social networks. I think that if the analytics data is small enough for the models to be fit, these tools can represent a qualitative leap in the understanding and prediction of user behaviour. Most uses of social networks in enterprise analytics that I have seen are limited to descriptive statistics (what is a user's centrality; what is the degree distribution), and the use of these descriptive statistics as fixed predictors in a model. I believe that this approach is an important first step, but ignores the stochastic nature of the network, and the dynamics of tie formation and dissolution. Realistic modeling of the network can lead to more principled, and more accurate predictions of the quantities that enterprise users care about. The rub is that the Markov Chain Monte Carlo Maximum Likelihood algorithms used to fit modern generative social network models (such as exponential-family random graph models) do not scale well at all. These models are typically limited to fitting networks with fewer than 50,000 vertices, which is clearly insufficient for most analytics customers who have networks more on the order of 50,000,000. This problem is not insoluble though. Part of my ongoing research involves scalable algorithms for fitting social network models. Ajay—You decided to go from your Phd into consulting (www.fellstat.com). What were some of the options you considered in this career choice? Ian I have been working in the role of a statistical consultant for the last 7 years, starting as an in-house consultant at UCSD after obtaining my MS. Fellows Statistics has been operating for the last 3 years, though not fulltime until January of this year. As I had already been consulting, it was a natural progression to transition to consulting fulltime once I graduated with my Phd. This has allowed me to both work on interesting corporate projects and continue research related to my dissertation via sub-awards from various universities.

Ajay—What does Fellstat.com offer in its consulting practice?

Ian—Fellows Statistics offers personalized analytics services to both corporate and academic clients. We are a boutique company that can scale from a single statistician to a small team of analysts chosen specifically with the client's needs in mind. I believe that by being small, we can provide better, close-to-the-ground responsive service to our clients. As a practice, we live at the intersection of mathematical sophistication and computational skill, with a hint of UI design thrown into the mix. Corporate clients can expect a diverse range of analytic skills from the development of novel algorithms to the design and presentation of data for a general audience. We have worked with Revolution Analytics developing algorithms for their ScaleR product, the Center for Disease Control developing graphical user interfaces set to be deployed for world-wide HIV surveillance, and Prospectus analysing clinical trial data for retinal surgery. With access to the cutting edge research taking place in the academic community, and the skills to implement them in corporate environments, Fellows Statistics is able to offer clients world-class analytics services.

Ajay—How does Big Data affect the practice of statistics in business decisions?

Ian—There is a big gap in terms of how the basic practice of statistics is taught in most universities, and the types of analyses that are useful when data sizes become large. Back when I was at UCSD, I remember a researcher there jokingly saying that everything is correlated rho=.2. He was joking, but there is a lot of truth to that statement. As data sizes get larger everything becomes significant if a hypothesis test is done, because the test has the power to detect even trivial relationships.

Ajay—How is the R community including developers coping with the Big Data era? What do you think R can do more for Big Data?

Ian—On the open source side, there has been a lot of movement to improve R's handling of Big Data. The bigmemory project and the ff package both serve to extend R's reach beyond in-memory data structures. Revolution Analytics also has the ScaleR package, which costs money, but is lightening fast and has an ever growing list of analytic techniques implemented. There are also several packages integrating R with hadoop.

Ajay—Describe your research into data visualization including word cloud and other packages. What do you think of Shiny, D3.Js and online data visualization?

Ian—I recently had the opportunity to delve into d3.js for a client project and absolutely love it. Combined with Shiny, d3, and R one can very quickly create a web visualization of an R modeling technique. One limitation of d3 is that it does not work well with internet explorer 6–8. Once these browsers finally leave the ecosystem, I expect an explosion of sites using d3.

Ajay—Do you think word cloud is an overused data visualization type and how can it be refined?

Ian—I would say yes, but not for the reasons you would think. A lot of people criticize word clouds because they convey the same information as a bar chart, but with less specificity. With a bar chart you can actually see the frequency, whereas you only get a relative idea with word clouds based on the size of the word. I think this is both an absolutely correct statement, and misses the point completely. Visualizations are about communicating with the reader. If your readers are statisticians, then they will happily consume the bar chart, following the bar heights to their point on the y-axis to find the frequencies. A statistician will spend time with a graph, will mull it over, and consider what deeper truths are found there. Statisticians are weird though. Most people care as much about how pretty the graph looks as its content. To communicate to these people (i.e., everyone else) it is appropriate and right to sacrifice statistical specificity to design considerations. After all, if the user stops reading you have not conveyed anything. But back to the question I would say that they are over used because they represent a very superficial analysis of a text or corpus. The word counts do convey an aspect of a text, but not a very nuanced one. The next step in looking at a corpus of texts would be to ask how are they different and how are they the same. The word cloud package has the comparison and commonality word clouds, which attempt to extend the basic word cloud to answer these questions (see: http://blog.fellstat.com/?p=101).

About—Dr. Ian Fellows is a professional statistician based out of the University of California, Los Angeles. His research interests range over many sub-disciplines of statistics. His work in statistical visualization won the prestigious John Chambers Award in 2011, and in 2007–2008 his Texas Hold'em AI programs were ranked second in the world.

Applied data analysis has been a passion for him, and he is accustomed to providing accurate, timely analysis for a wide range of projects, and assisting in the interpretation and communication of statistical results. He can be contacted at info@fellstat.com.

3.6 Notable R Projects

Some of the notable R projects have contributed greatly in the past and present to making the R environment more friendly for web applications. Some other projects which are of note:

3.6.1 Installr Package

R is great for installing software. Through the "installr" package you can automate the updating of R (on Windows, using updateR()) and install new software. Software installation is initiated through a gui (just run installr()), or through functions such as: install.Rtools(), install.pandoc(), install.git(), and many more. The updateR() command performs the following: finding the latest R version, downloading it, running the installer, deleting the installation file, copy and updating old packages to the new R installation. It was created by Tal Gallili, who is also the creator of the R-Bloggers website that aggregates 500 plus blogs on R language through RSS feeds.

3.6.2 Rserve

Rserve is a TCP/IP server which allows other programs to use facilities of R (see www.r-project.org) from various languages without the need to initialize R or link against R library. Every connection has a separate workspace and working directory. Client-side implementations are available for popular languages such as C/C + +, PHP and Java. Rserve supports remote connection, authentication, and file transfer. Typical use is to integrate R back end for computation of statistical models, plots, etc. in other applications. Here, Rserve acts as a socket server (TCP/IP or local sockets) which allows binary requests to be sent to R.

http://cran.r-project.org/web/packages/Rserve/index.html

3.6.3 RApache

rApache is a project supporting web application development using the R statistical language and environment and the Apache web server. The current release runs on UNIX/Linux and Mac OS X operating systems. It is available at http://rapache.net/. Please do read the fine manual at http://rapache.net/manual.html.

3.6.4 Rook

This package contains the Rook specification and convenience software for building and running Rook applications. It borrows heavily from Ruby's Rack project A Rook application is literally an R function that takes an R environment as input, and returns a list of HTTP-relevant items as output. You can write and run Rook apps within R using the built-in Rhttpd web server and deploy them on rApache.

Rook is both a web server interface and an R package. The idea behind the former is to separate application development from server implementation. Thus, when a web server supports a web server interface, an application written to its specifications is guaranteed to run on that server.

3.6.5 RJ and Rservi

Note RServi is a confusing name (from Rserve) but it is a distinct project. RJ is an open source library providing tools and interfaces to integrate R in Java applications. RJ project also provides a pool for R engines, easy to setup and manage by a web-interface or JMX. One or multiple client can borrow the R engines (called RServi) see http://www.walware.de/it/rj/ and https://github.com/walware/rj-servi.

3.6.6 *R and Java*

- rJava is a simple R-to-Java interface. It is comparable to the .C/.Call C interface. rJava provides a low-level bridge between R and Java (via JNI). It allows to create objects, call methods, and access fields of Java objects from R.
- JRI is a Java/R Interface, which allows to run R inside Java applications as a single thread. Basically it loads R dynamic library into Java and provides a Java API to R functionality. It supports both simple calls to R functions and a full running REPL. JRI is now part of rJava.
- The JGR project makes the full use of both JRI and rJava to provide a full Java GUI for R.http://cran.r-project.org/web/packages/JGR/.
- iPlots is a package for the R statistical environment which provides high interaction statistical graphics, written in Java. http://cran.r-project.org/web/packages/iplots/index.html.

3.7 Creating Your Desktop on the Cloud

If you have been using the Windows Operating System on your desktop, when moving to the cloud you may choose the following software to minimize any issues in transition to cloud computing. Download and install the following

1. Chrome Browser www.google.com/chrome
2. Adobe Acrobat Reader http://get.adobe.com/reader/
3. Libre Office www.libreoffice.org/download or Open Office http://www.openoffice.org/download/
4. 7 zip http://www.7-zip.org/download.html
5. Java http://www.java.com/en/download/index.jsp

Additional Software for your statistical computing environment

1. Python http://www.python.org/download/
2. R http://cran.stat.ucla.edu/bin/windows/ or other locations at http://cran.r-project.org/mirrors.html
3. R Studio IDE http://www.rstudio.com/ide/download/desktop