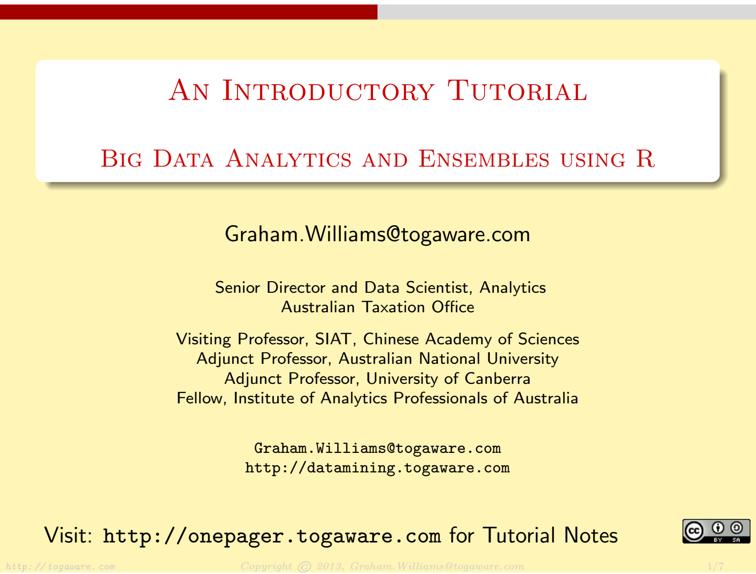


An Introductory Tutorial

Big Data Analytics and Ensembles using R

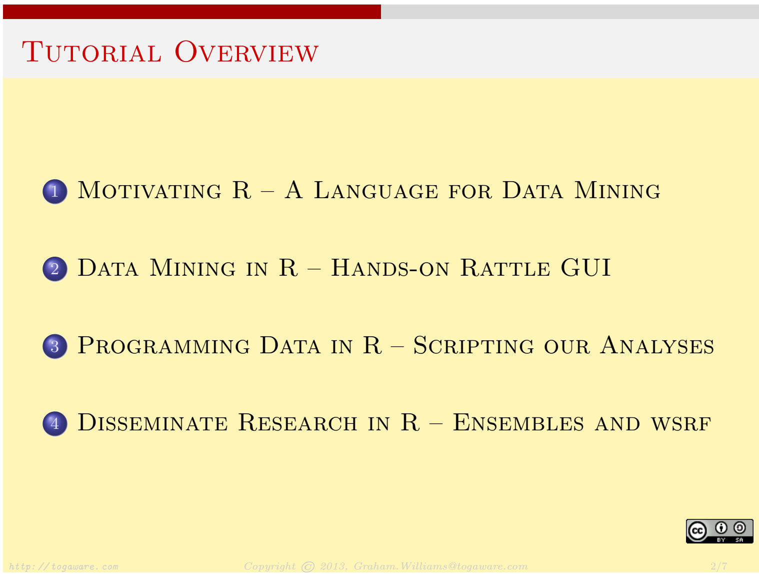
Graham.Williams@togaware.com

Senior Director and Data Scientist, Analytics
Australian Taxation Office

Visiting Professor, SIAT, Chinese Academy of Sciences
Adjunct Professor, Australian National University
Adjunct Professor, University of Canberra
Fellow, Institute of Analytics Professionals of Australia

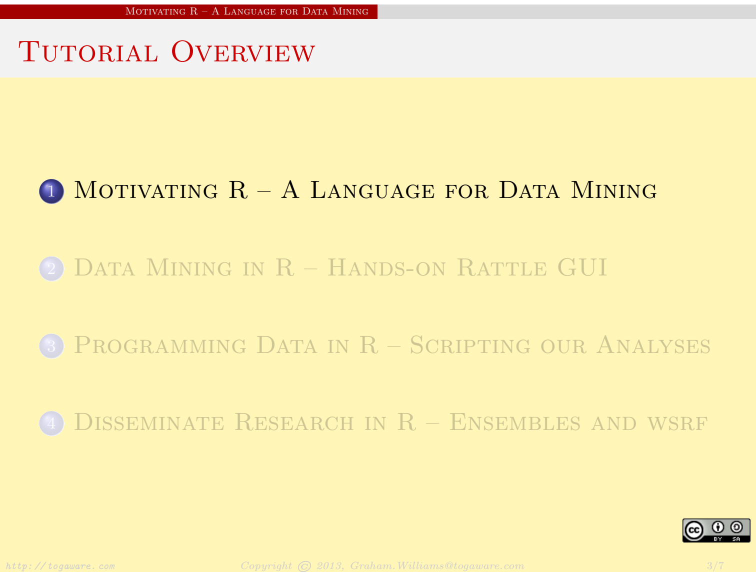
Graham.Williams@togaware.com
http://datamining.togaware.com

Visit: `http://onepager.togaware.com` for Tutorial Notes



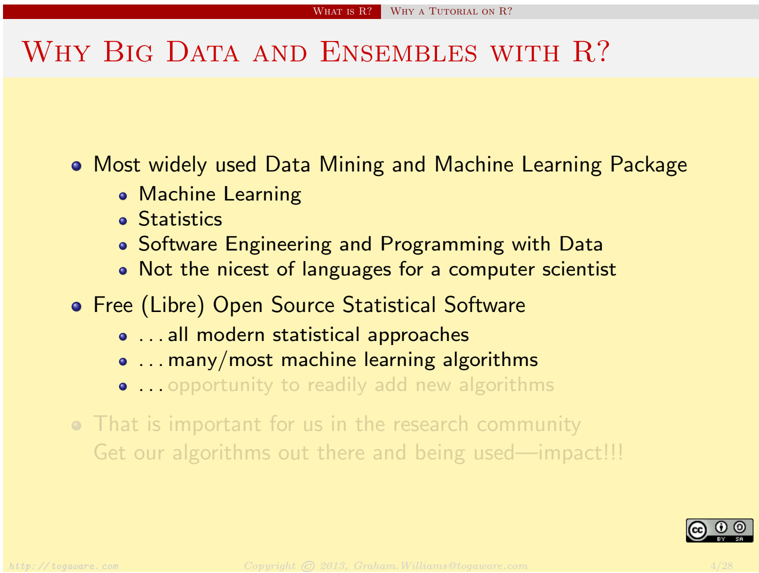
Tutorial Overview

1. Motivating R – A Language for Data Mining
2. Data Mining in R – Hands-on Rattle GUI
3. Programming Data in R – Scripting our Analyses
4. Disseminate Research in R – Ensembles and wsrf



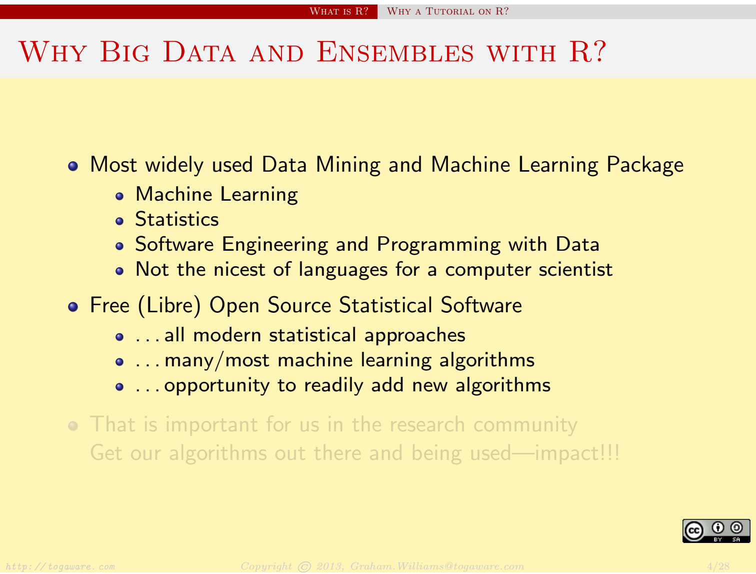
Tutorial Overview

1. Motivating R – A Language for Data Mining
2. Data Mining in R – Hands-on Rattle GUI
3. Programming Data in R – Scripting our Analyses
4. Disseminate Research in R – Ensembles and wsrf



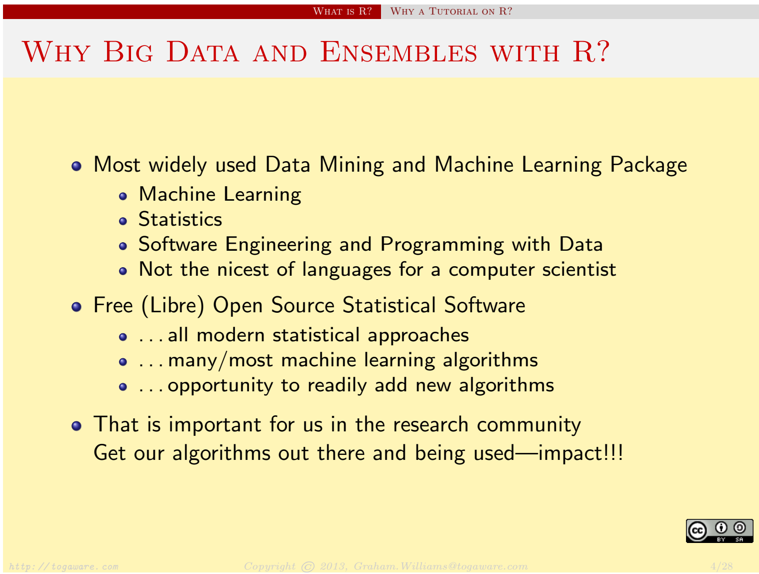
Why Big Data and Ensembles with R?

- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - Not the nicest of languages for a computer scientist
- Free (Libre) Open Source Statistical Software
  - ... all modern statistical approaches
  - ... many/most machine learning algorithms
  - ... opportunity to readily add new algorithms
- That is important for us in the research community
  Get our algorithms out there and being used—impact!!!



Why Big Data and Ensembles with R?

- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - Not the nicest of languages for a computer scientist
- Free (Libre) Open Source Statistical Software
  - ... all modern statistical approaches
  - ... many/most machine learning algorithms
  - ... opportunity to readily add new algorithms
- That is important for us in the research community
  Get our algorithms out there and being used—impact!!!



Why Big Data and Ensembles with R?

- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - Not the nicest of languages for a computer scientist
- Free (Libre) Open Source Statistical Software
  - ... all modern statistical approaches
  - ... many/most machine learning algorithms
  - ... opportunity to readily add new algorithms
- That is important for us in the research community
  Get our algorithms out there and being used—impact!!!

Why Big Data and Ensembles with R?

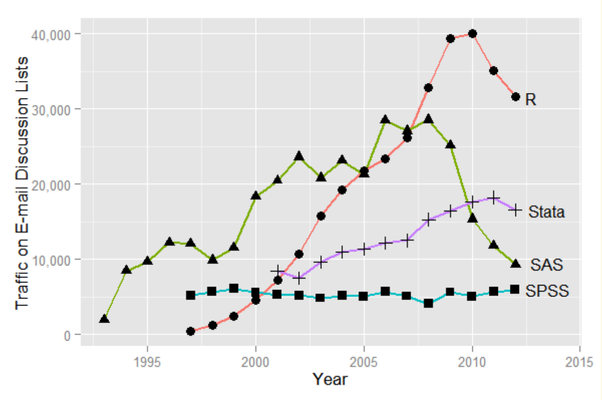
- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - Not the nicest of languages for a computer scientist
- Free (Libre) Open Source Statistical Software
  - . . . all modern statistical approaches
  - . . . many/most machine learning algorithms
  - . . . opportunity to readily add new algorithms
- That is important for us in the research community
  Get our algorithms out there and being used—impact!!!

Installing R

- Instructions on Togaware: `http://rattle.togaware.com`
- Visit CRAN: `http://mirrors.ustc.edu.cn/CRAN/`
- Linux: Install package for your distribution
  `$ wajig install r-recommended` (Debian/Ubuntu)
- Windows: Download and install from CRAN
- MacOSX: Download and install from CRAN

How Popular is R? Discussion List Traffic

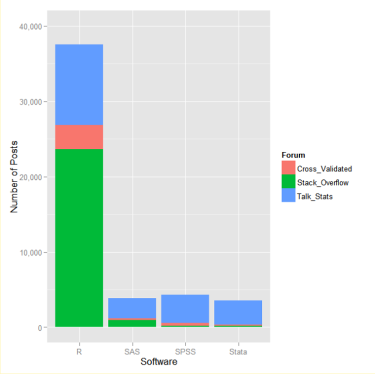
Monthly email traffic on software's main discussion list.



http://r4stats.com/articles/popularity/

How Popular is R? Discussion Topics

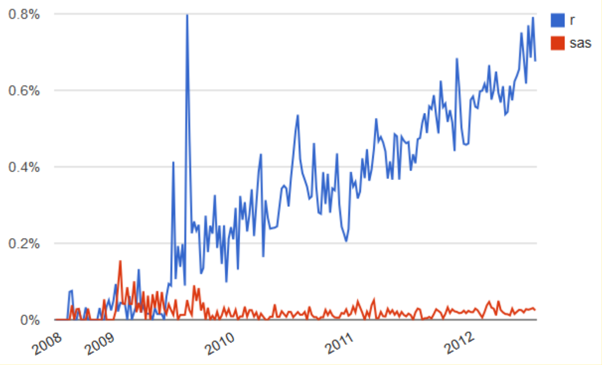
Number of discussions on popular QandA forums 2013.



http://r4stats.com/articles/popularity/

How Popular is R? R versus SAS

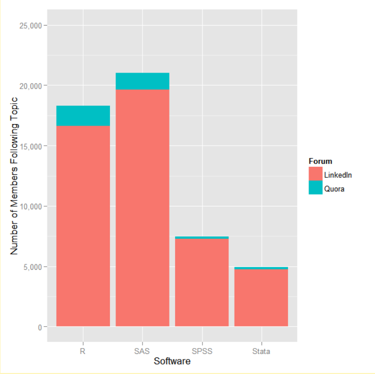
Number of R/SAS related posts to Stack Overflow by week.



http://r4stats.com/articles/popularity/

How Popular is R? Professional Forums

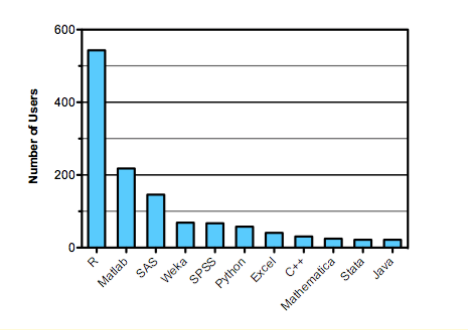
Registered for the main discussion group for each software.



http://r4stats.com/articles/popularity/

How Popular is R? Used in Analytics Competitions

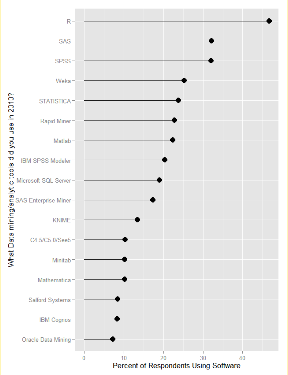
Software used in data analysis competitions in 2011.



http://r4stats.com/articles/popularity/

How Popular is R? User Survey

Rexer Analytics Survey 2010 results for data mining/analytic tools.



http://r4stats.com/articles/popularity/

What is R?

Video from Revolution Analytics - 90 seconds

`http://www.revolutionanalytics.com/what-is-open-source-r/`

Tutorial Overview

1. Motivating R – A Language for Data Mining
2. Data Mining in R – Hands-on Rattle GUI
3. Programming Data in R – Scripting our Analyses
4. Disseminate Research in R – Ensembles and wsrf

Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R

Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R

Data Mining and Big Data

- Application of
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - Intuition
- To Big Data — Volume, Velocity, Variety, Value, Veracity
- . . . to discover new knowledge
- . . . to improve business outcomes
- . . . to deliver better tailored services

The Business of Data Mining

- Australian Taxation Office
  - Lodgment ($110M)
  - Tax Havens ($150M)
  - Tax Fraud ($250M)
- Department of Immigration
- IBM Buys SPSS for $1.2B in 2009
- SAS has annual revenue approaching $3B
- Analytics is >$100B business and >$320B by 2020 (McKinsey)
- Amazon, eBay/PayPal, Google . . .

Basic Tools: Data Mining Algorithms

- Linear Discriminant Analysis (lda)
- Logistic Regression (glm)
- Decision Trees (rpart, wsrpart)
- Random Forests (randomForest, wsrf)
- Boosted Stumps (ada)
- Neural Networks (nnet)
- Support Vector Machines (kernlab)
- . . .

That's a lot of tools to learn in R!
Many with different interfaces and options.

Overview

1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R

Why a GUI?

- Statistics can be complex and traps await
- **So many** tools in R to deliver insights
- Effective analyses should be scripted
- Scripting also required for repeatability
- R is a language for **programming** with data

How to remember how to do all of this in R?
How to skill up 150 data analysts with Data Mining?

Users of Rattle

Today, Rattle is used world wide in many industries

- Health analytics
- Customer segmentation and marketing
- Fraud detection
- Government

It is used by

- Consultants and Analytics Teams across business
- Universities to teach Data Mining

It is and will remain freely available.

CRAN and `http://rattle.togaware.com`

Installation

- Rattle is built using R
- Need to download and install R from cran.r-project.org
- Recommend also install RStudio from www.rstudio.org
- Then start up RStudio and install Rattle:

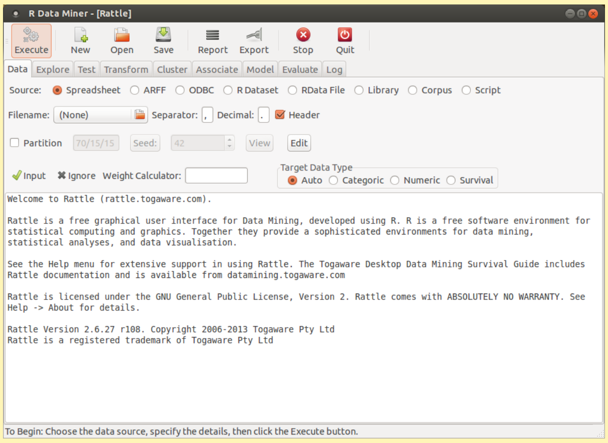
```
install.packages("rattle")
```

- Then we can start up Rattle:

```
rattle()
```

- Required packages are loaded as needed.

A Tour Thru Rattle: Startup



Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: (None) Separator: , Decimal: . ☑ Header

□ Partition 70/15/15 Seed: 42 View Edit

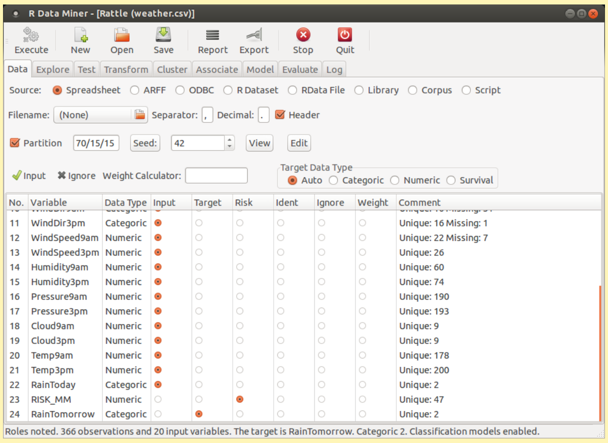
✓ Input ✖ Ignore Weight Calculator:

Target Data Type ● Auto ○ Categoric ○ Numeric ○ Survival

```
Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for
statistical computing and graphics. Together they provide a sophisticated environments for data mining,
statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The Togaware Desktop Data Mining Survival Guide includes
Rattle documentation and is available from datamining.togaware.com

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See
Help -> About for details.

Rattle Version 2.6.27 r108. Copyright 2006-2013 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd
```

To Begin: Choose the data source, specify the details, then click the Execute button.

A Tour Thru Rattle: Loading Data



Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: (None) Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42 View Edit

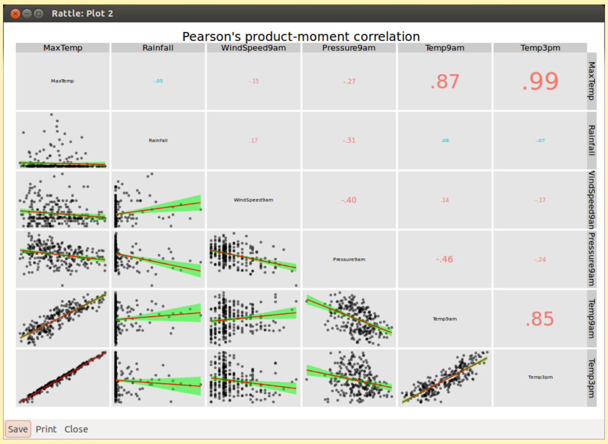
✓ Input ✖ Ignore Weight Calculator:

Target Data Type ● Auto ○ Categoric ○ Numeric ○ Survival

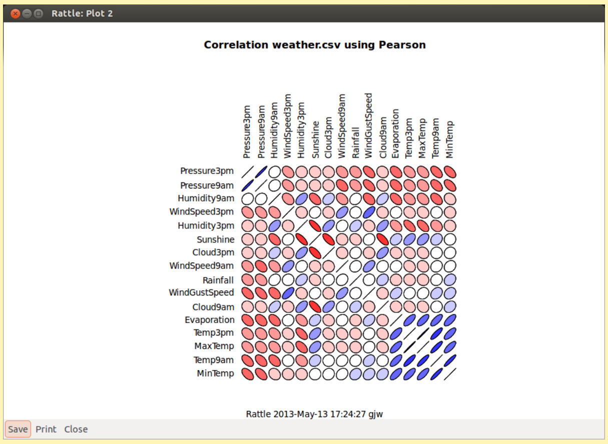
No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
11	WindDir3pm	Categoric	●	○	○	○	○	○	Unique: 16 Missing: 1
12	WindSpeed9am	Numeric	●	○	○	○	○	○	Unique: 22 Missing: 7
13	WindSpeed3pm	Numeric	●	○	○	○	○	○	Unique: 26
14	Humidity9am	Numeric	●	○	○	○	○	○	Unique: 60
15	Humidity3pm	Numeric	●	○	○	○	○	○	Unique: 74
16	Pressure9am	Numeric	●	○	○	○	○	○	Unique: 190
17	Pressure3pm	Numeric	●	○	○	○	○	○	Unique: 193
18	Cloud9am	Numeric	●	○	○	○	○	○	Unique: 9
19	Cloud3pm	Numeric	●	○	○	○	○	○	Unique: 9
20	Temp9am	Numeric	●	○	○	○	○	○	Unique: 178
21	Temp3pm	Numeric	●	○	○	○	○	○	Unique: 200
22	RainToday	Categoric	●	○	○	○	○	○	Unique: 2
23	RISK_MM	Numeric	○	○	●	○	○	○	Unique: 47
24	RainTomorrow	Categoric	○	●	○	○	○	○	Unique: 2

Roles noted. 366 observations and 20 input variables. The target is RainTomorrow. Categoric 2. Classification models enabled.

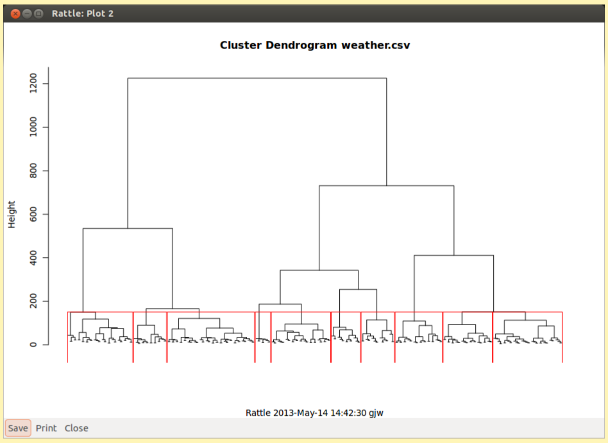
A Tour Thru Rattle: Explore Distribution



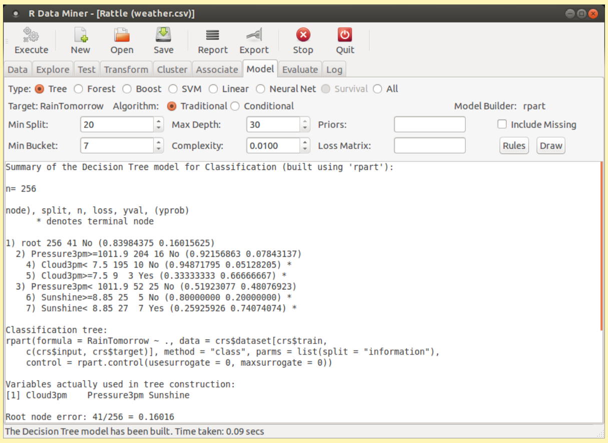
A Tour Thru Rattle: Explore Correlations



A Tour Thru Rattle: Hierarchical Cluster



A Tour Thru Rattle: Decision Tree



Type: ● Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: RainTomorrow Algorithm: ● Traditional ○ Conditional Model Builder: rpart

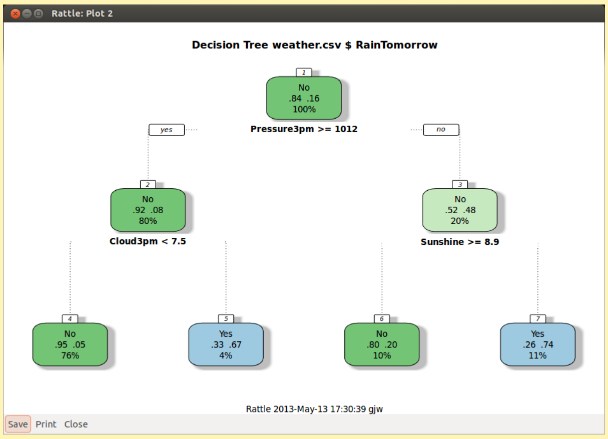
Min Split: 20 Max Depth: 30 Priors: □ Include Missing

Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

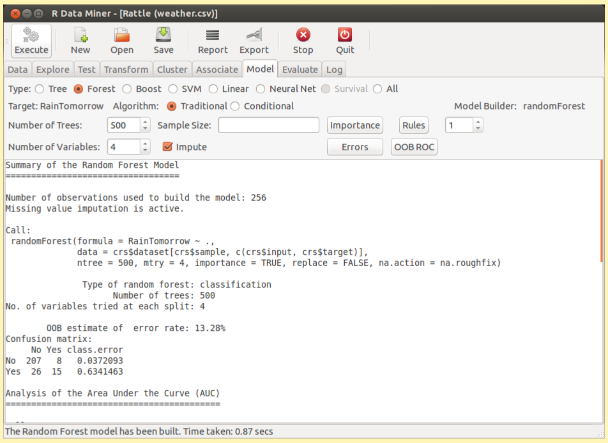
```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 256 41 No (0.83984375 0.16015625)
  2) Pressure3pm>=1011.9 204 16 No (0.92156863 0.07843137)
    4) Cloud3pm< 7.5 195 10 No (0.94871795 0.05128205) *
    5) Cloud3pm>=7.5 9  3 Yes (0.33333333 0.66666667) *
  3) Pressure3pm< 1011.9 52 25 No (0.51923077 0.48076923)
    6) Sunshine>=8.85 25  5 No (0.80000000 0.20000000) *
    7) Sunshine< 8.85 27  7 Yes (0.25925926 0.74074074) *

Classification tree:
rpart(formula = RainTomorrow ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)], method = "class", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] Cloud3pm    Pressure3pm Sunshine

Root node error: 41/256 = 0.16016
```

The Decision Tree model has been built. Time taken: 0.09 secs

A Tour Thru Rattle: Decision Tree Plot



A Tour Thru Rattle: Random Forest



Type: ○ Tree ● Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: RainTomorrow Algorithm: ● Traditional ○ Conditional Model Builder: randomForest

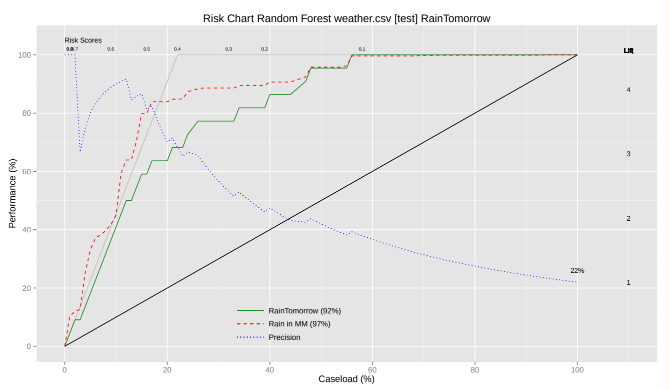
Number of Trees: 500 Sample Size: 1

Number of Variables: 4 ☑ Impute

```
Summary of the Random Forest Model
==================================

Number of observations used to build the model: 256
Missing value imputation is active.

Call:
 randomForest(formula = RainTomorrow ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 500, mtry = 4, importance = TRUE, replace = FALSE, na.action = na.roughfix)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 13.28%
Confusion matrix:
     No Yes class.error
No  207   8   0.0372093
Yes  26  15   0.6341463

Analysis of the Area Under the Curve (AUC)
==========================================
```

The Random Forest model has been built. Time taken: 0.87 secs

A Tour Thru Rattle: Risk Chart



Overview

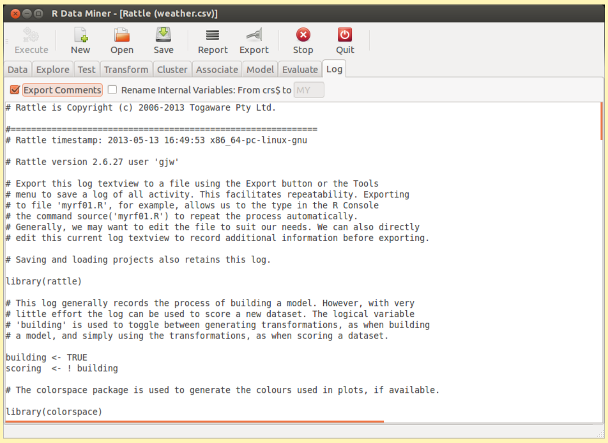
1. An Introduction to Data Mining
2. The Rattle Package for Data Mining
3. Moving Into R

Data Miners are Programmers of Data

- Data miners are programmers of data
- A GUI can only do so much
- R is a powerful statistical language

- Professional data mining
  - Scripting
  - Transparency
  - Repeatability

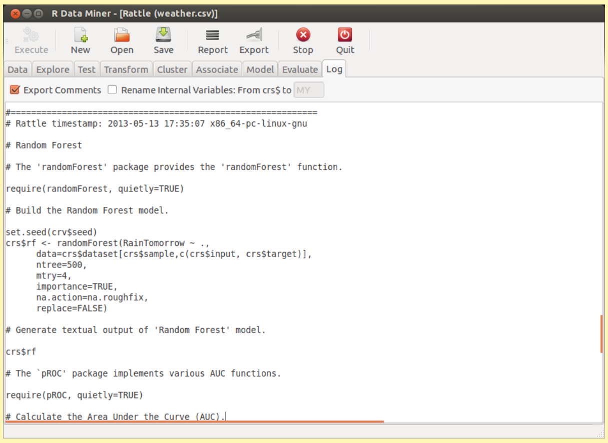
From GUI to CLI — Rattle's Log Tab



☑ Export Comments □ Rename Internal Variables: From crs$ to MY

```
# Rattle is Copyright (c) 2006-2013 Togaware Pty Ltd.

#============================================================
# Rattle timestamp: 2013-05-13 16:49:53 x86_64-pc-linux-gnu

# Rattle version 2.6.27 user 'gjw'

# Export this log textview to a file using the Export button or the Tools
# menu to save a log of all activity. This facilitates repeatability. Exporting
# to file 'myrf01.R', for example, allows us to the type in the R Console
# the command source('myrf01.R') to repeat the process automatically.
# Generally, we may want to edit the file to suit our needs. We can also directly
# edit this current log textview to record additional information before exporting.

# Saving and loading projects also retains this log.

library(rattle)

# This log generally records the process of building a model. However, with very
# little effort the log can be used to score a new dataset. The logical variable
# 'building' is used to toggle between generating transformations, as when building
# a model, and simply using the transformations, as when scoring a dataset.

building <- TRUE
scoring  <- ! building

# The colorspace package is used to generate the colours used in plots, if available.

library(colorspace)
```

From GUI to CLI — Rattle's Log Tab



☑ Export Comments □ Rename Internal Variables: From crs$ to MY

```
#============================================================
# Rattle timestamp: 2013-05-13 17:35:07 x86_64-pc-linux-gnu

# Random Forest

# The 'randomForest' package provides the 'randomForest' function.

require(randomForest, quietly=TRUE)

# Build the Random Forest model.

set.seed(crv$seed)
crs$rf <- randomForest(RainTomorrow ~ .,
      data=crs$dataset[crs$sample,c(crs$input, crs$target)],
      ntree=500,
      mtry=4,
      importance=TRUE,
      na.action=na.roughfix,
      replace=FALSE)

# Generate textual output of 'Random Forest' model.

crs$rf

# The `pROC' package implements various AUC functions.

require(pROC, quietly=TRUE)

# Calculate the Area Under the Curve (AUC).
```

Step 1: Load the Dataset

```
dsname <- "weather"
ds     <- get(dsname)
dim(ds)

## [1] 366  24

names(ds)

##  [1] "Date"          "Location"      "MinTemp"       "...
##  [5] "Rainfall"      "Evaporation"   "Sunshine"      "...
##  [9] "WindGustSpeed" "WindDir9am"    "WindDir3pm"    "...
## [13] "WindSpeed3pm"  "Humidity9am"   "Humidity3pm"   "...
....
```

Step 2: Observe the Data — Observations

```
head(ds)

##         Date Location MinTemp MaxTemp Rainfall Evapora...
## 1 2007-11-01 Canberra     8.0    24.3      0.0        ...
## 2 2007-11-02 Canberra    14.0    26.9      3.6        ...
## 3 2007-11-03 Canberra    13.7    23.4      3.6        ...
....

tail(ds)

##           Date Location MinTemp MaxTemp Rainfall Evapo...
## 361 2008-10-26 Canberra     7.9    26.1        0      ...
## 362 2008-10-27 Canberra     9.0    30.7        0      ...
## 363 2008-10-28 Canberra     7.1    28.4        0      ...
....
```

Step 2: Observe the Data — Structure

```
str(ds)

## 'data.frame': 366 obs. of  24 variables:
##  $ Date         : Date, format: "2007-11-01" "2007-11-...
##  $ Location     : Factor w/ 46 levels "Adelaide","Alba...
##  $ MinTemp      : num  8 14 13.7 13.3 7.6 6.2 6.1 8.3 ...
##  $ MaxTemp      : num  24.3 26.9 23.4 15.5 16.1 16.9 1...
##  $ Rainfall     : num  0 3.6 3.6 39.8 2.8 0 0.2 0 0 16...
##  $ Evaporation  : num  3.4 4.4 5.8 7.2 5.6 5.8 4.2 5.6...
##  $ Sunshine     : num  6.3 9.7 3.3 9.1 10.6 8.2 8.4 4....
##  $ WindGustDir  : Ord.factor w/ 16 levels "N"<"NNE"<"N...
##  $ WindGustSpeed: num  30 39 85 54 50 44 43 41 48 31 ...
##  $ WindDir9am   : Ord.factor w/ 16 levels "N"<"NNE"<"N...
##  $ WindDir3pm   : Ord.factor w/ 16 levels "N"<"NNE"<"N...
....
```

Step 2: Observe the Data — Summary

```
summary(ds)

##       Date                     Location      MinTemp   ...
##  Min.   :2007-11-01   Canberra     :366   Min.   :-5.3...
##  1st Qu.:2008-01-31   Adelaide     :  0   1st Qu.: 2.3...
##  Median :2008-05-01   Albany       :  0   Median : 7.4...
##  Mean   :2008-05-01   Albury       :  0   Mean   : 7.2...
##  3rd Qu.:2008-07-31   AliceSprings :  0   3rd Qu.:12.5...
##  Max.   :2008-10-31   BadgerysCreek:  0   Max.   :20.9...
##                       (Other)      :  0                ...
##     Rainfall      Evaporation      Sunshine       Wind...
##  Min.   : 0.00   Min.   : 0.20   Min.   : 0.00   NW    ...
##  1st Qu.: 0.00   1st Qu.: 2.20   1st Qu.: 5.95   NNW   ...
##  Median : 0.00   Median : 4.20   Median : 8.60   E     ...
....
```

Step 2: Observe the Data — Variables

```
id     <- c("Date", "Location")
target <- "RainTomorrow"
risk   <- "RISK_MM"
(ignore <- union(id, risk))

## [1] "Date"     "Location" "RISK_MM"

(vars   <- setdiff(names(ds), ignore))

##  [1] "MinTemp"       "MaxTemp"       "Rainfall"      "...
##  [5] "Sunshine"      "WindGustDir"   "WindGustSpeed" "...
##  [9] "WindDir3pm"    "WindSpeed9am"  "WindSpeed3pm"  "...
## [13] "Humidity3pm"   "Pressure9am"   "Pressure3pm"   "...
....
```

Step 3: Clean the Data — Remove Missing

```
dim(ds)

## [1] 366  24

sum(is.na(ds[vars]))

## [1] 47

ds <- ds[-attr(na.omit(ds[vars]), "na.action"),]
```

Step 3: Clean the Data — Remove Missing

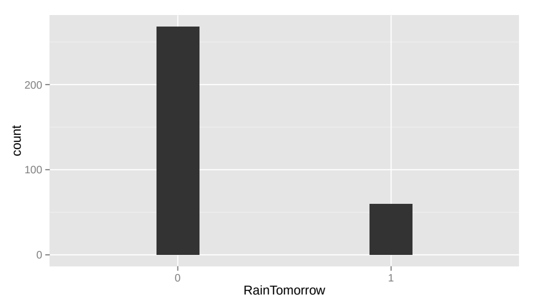
```
dim(ds)

## [1] 328  24

sum(is.na(ds[vars]))

## [1] 0
```

Step 3: Clean the Data—Target as Categoric

```
summary(ds[target])

##   RainTomorrow
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.183
##  3rd Qu.:0.000
##  Max.   :1.000
....

ds[target] <- as.factor(ds[[target]])
levels(ds[target]) <- c("No", "Yes")
```

Step 3: Clean the Data—Target as Categoric

```
summary(ds[target])

##  RainTomorrow
##  0:268
##  1: 60
```



Step 4: Prepare for Modelling

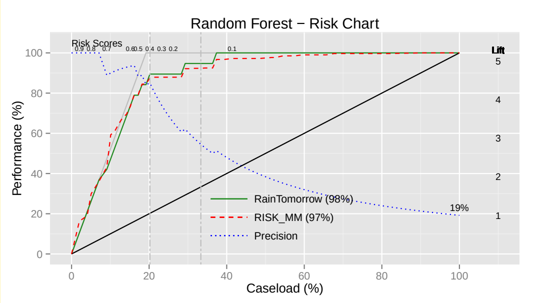
```
(form <- formula(paste(target, "~ .")))

## RainTomorrow ~ .

(nobs <- nrow(ds))

## [1] 328

train <- sample(nobs, 0.70*nobs)
length(train)

## [1] 229

test  <- setdiff(1:nobs, train)
length(test)

## [1] 99
```

Step 5: Build the Model—Random Forest

```
library(randomForest)
model <- randomForest(form, ds[train, vars], na.action=na.omit)
model

##
## Call:
##  randomForest(formula=form, data=ds[train, vars], ...
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
....
```

Step 6: Evaluate the Model—Risk Chart

```
pr <- predict(model, ds[test,], type="prob")[,2]
riskchart(pr, ds[test, target], ds[test, risk],
          title="Random Forest - Risk Chart",
          risk=risk, recall=target, thresholds=c(0.35, 0.15))
```



Tutorial Overview

1. Motivating R – A Language for Data Mining
2. Data Mining in R – Hands-on Rattle GUI
3. Programming Data in R – Scripting our Analyses
4. Disseminate Research in R – Ensembles and wsrf

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

Tools

- Ubuntu GNU/Linux operating system
  - Feature rich toolkit, up-to-date, easy to install, FLOSS
- RStudio
  - Easy to use integrated development environment, FLOSS
- R Statistical Software Language
  - Extensive, powerful, thousands of contributors, FLOSS
- KnitR
  - Produce beautiful documents, easily reproducible, FLOSS

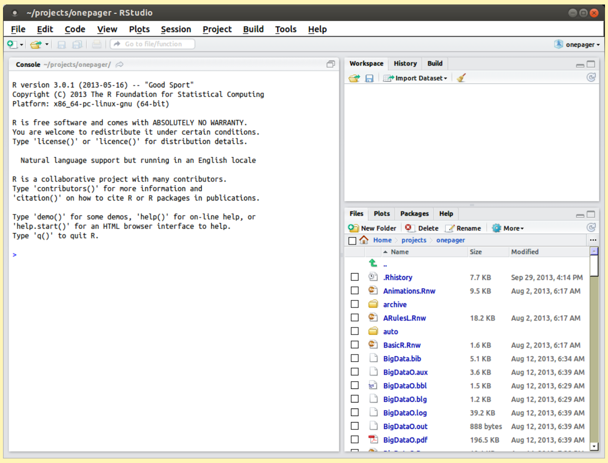
Using Ubuntu

- Desktop Ubuntu
- Connecting to Analytics Servers
  - Using XWin
  - Using VNC
- Start up RStudio from the Dash

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

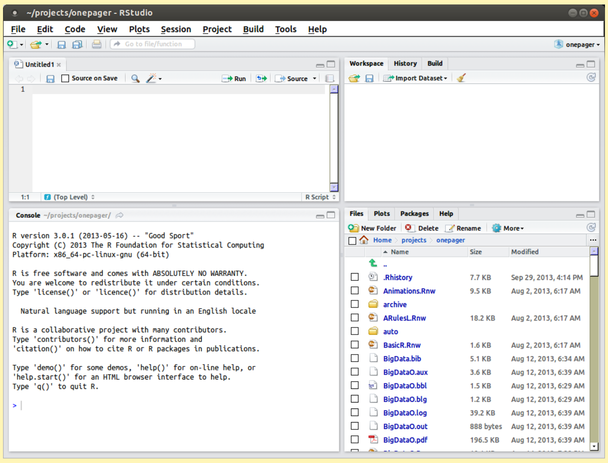
RStudio—The Default Three Panels


~/projects/onepager - RStudio
File Edit Code View Plots Session Project Build Tools Help
Go to file/function
onepager
Console ~/projects/onepager/
Workspace History Build
Import Dataset
Files Plots Packages Help
New Folder Delete Rename More
Home projects onepager

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Name	Size	Modified
..		
.Rhistory	7.7 KB	Sep 29, 2013, 4:14 PM
Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
archive		
ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
auto		
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM

RStudio—With R Script File—Editor Panel


~/projects/onepager - RStudio
File Edit Code View Plots Session Project Build Tools Help
Go to file/function
onepager
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/projects/onepager/
Workspace History Build
Import Dataset
Files Plots Packages Help
New Folder Delete Rename More
Home projects onepager

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Name	Size	Modified
..		
.Rhistory	7.7 KB	Sep 29, 2013, 4:14 PM
Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
archive		
ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
auto		
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

Scatterplot—R Code

Our first little bit of R code:

- Load a couple of *packages* into the R *library*

```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
```

- Then produce a quick plot using `qplot()`

```
ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
```

- Your turn: give it a go.

Scatterplot—R Code

Our first little bit of R code:

- Load a couple of *packages* into the R *library*

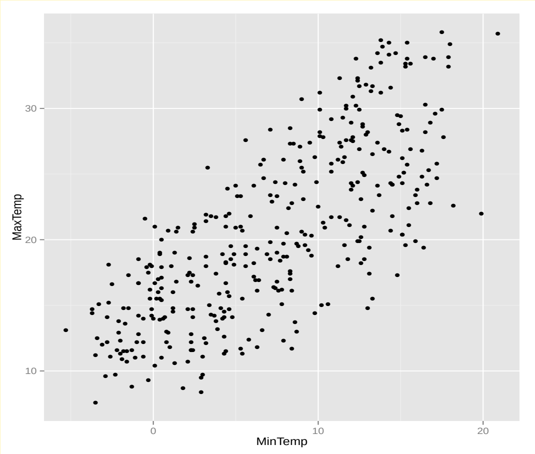
```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
```

- Then produce a quick plot using `qplot()`

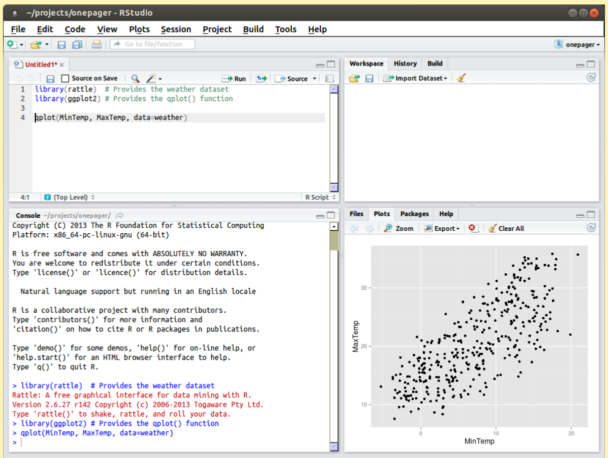
```
ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
```

- Your turn: give it a go.

Scatterplot—Plot



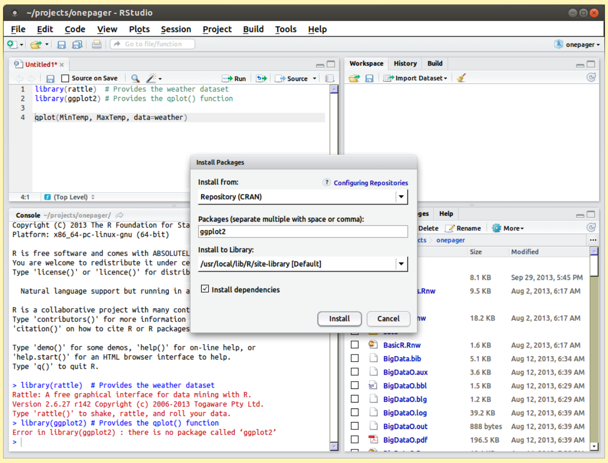
Scatterplot—RStudio


~/projects/onepager - RStudio
File Edit Code View Plots Session Project Build Tools Help
Untitled1*
Source on Save
Run
Source
Workspace History Build
Import Dataset
4:1 (Top Level) R Script
Console ~/projects/onepager/
Files Plots Packages Help
Zoom Export Clear All
MaxTemp
MinTemp

```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

qplot(MinTemp, MaxTemp, data=weather)
```

```
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(rattle)  # Provides the weather dataset
Rattle: A free graphical interface for data mining with R.
Version 2.6.27 r142 Copyright (c) 2006-2013 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> library(ggplot2) # Provides the qplot() function
> qplot(MinTemp, MaxTemp, data=weather)
>
```

Missing Packages–Tools→Install Packages. . .



Install from: Repository (CRAN)

Packages (separate multiple with space or comma): ggplot2

Install to Library: /usr/local/lib/R/site-library [Default]

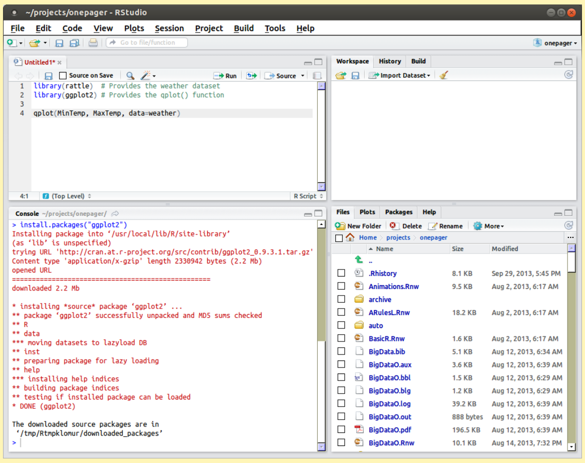
☑ Install dependencies

```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

qplot(MinTemp, MaxTemp, data=weather)
```

```
Copyright (C) 2013 The R Foundation for Sta
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTEL
You are welcome to redistribute it under ce
Type 'license()' or 'licence()' for distrib

  Natural language support but running in a

R is a collaborative project with many cont
Type 'contributors()' for more information
'citation()' on how to cite R or R packages

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(rattle)  # Provides the weather dataset
Rattle: A free graphical interface for data mining with R.
Version 2.6.27 r142 Copyright (c) 2006-2013 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> library(ggplot2) # Provides the qplot() function
Error in library(ggplot2) : there is no package called 'ggplot2'
>
```

	Size	Modified
	8.1 KB	Sep 29, 2013, 5:45 PM
s.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
nw	18.2 KB	Aug 2, 2013, 6:17 AM
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM

RStudio—Installing ggplot2


~/projects/onepager - RStudio
File Edit Code View Plots Session Project Build Tools Help
Untitled1*
Source on Save
Run
Source
Workspace History Build
Import Dataset
4:1 (Top Level) R Script
Console ~/projects/onepager/
Files Plots Packages Help
New Folder Delete Rename More
Home projects onepager

```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

qplot(MinTemp, MaxTemp, data=weather)
```

```
> install.packages("ggplot2")
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
trying URL 'http://cran.at.r-project.org/src/contrib/ggplot2_0.9.3.1.tar.gz'
Content type 'application/x-gzip' length 2330942 bytes (2.2 Mb)
opened URL
==================================================
downloaded 2.2 Mb

* installing *source* package 'ggplot2' ...
** package 'ggplot2' successfully unpacked and MD5 sums checked
** R
** data
*** moving datasets to lazyload DB
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (ggplot2)

The downloaded source packages are in
 '/tmp/Rtmpklomur/downloaded_packages'
>
```

Name	Size	Modified
..		
.Rhistory	8.1 KB	Sep 29, 2013, 5:45 PM
Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
archive		
ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
auto		
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM
BigDataO.Rnw	10.1 KB	Aug 14, 2013, 7:32 PM

RStudio—Keyboard Shortcuts

These will become very useful!

- Editor:
  - Ctrl-Enter will send the line of code to the R console
  - Ctrl-2 will move the cursor to the Console
- Console:
  - UpArrow will cycle through previous commands
  - Ctrl-UpArrow will search previous commands
  - Tab will complete function names and list the arguments
  - Ctrl-1 will move the cursor to the Editor

Your turn: try them out.

RStudio—Keyboard Shortcuts

These will become very useful!

- Editor:
  - Ctrl-Enter will send the line of code to the R console
  - Ctrl-2 will move the cursor to the Console

- Console:
  - UpArrow will cycle through previous commands
  - Ctrl-UpArrow will search previous commands
  - Tab will complete function names and list the arguments
  - Ctrl-1 will move the cursor to the Editor

Your turn: try them out.

Basic R

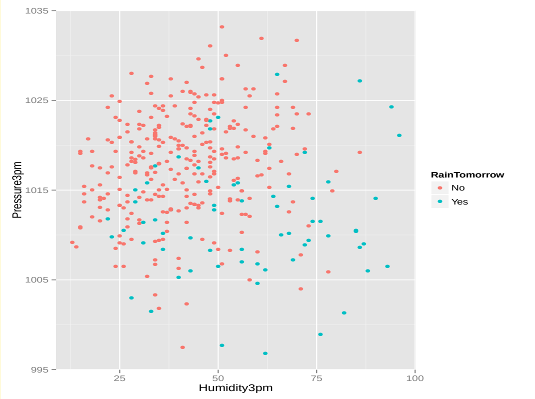
```
library(rattle)   # Load the weather dataset.
head(weather)     # First 6 observations of the dataset.

##         Date Location MinTemp MaxTemp Rainfall Evapora...
## 1 2007-11-01 Canberra     8.0    24.3      0.0         ...
## 2 2007-11-02 Canberra    14.0    26.9      3.6         ...
## 3 2007-11-03 Canberra    13.7    23.4      3.6         ...
....

str(weather)      # Strunctur of the variables in the dataset.

## 'data.frame': 366 obs. of  24 variables:
##  $ Date          : Date, format: "2007-11-01" "2007-11-...
##  $ Location      : Factor w/ 46 levels "Adelaide","Alba...
##  $ MinTemp       : num  8 14 13.7 13.3 7.6 6.2 6.1 8.3 ...
....
```

Basic R

```
summary(weather)  # Univariate summary of the variables.

##       Date                   Location      MinTemp       ...
##  Min.    :2007-11-01   Canberra     :366   Min.   :-5.30   ...
##  1st Qu.:2008-01-31   Adelaide     :  0   1st Qu.: 2.30   ...
##  Median :2008-05-01   Albany       :  0   Median : 7.45   ...
##  Mean    :2008-05-01   Albury       :  0   Mean   : 7.27   ...
##  3rd Qu.:2008-07-31   AliceSprings :  0   3rd Qu.:12.50   ...
##  Max.    :2008-10-31   BadgerysCreek:  0   Max.   :20.90   ...
##                        (Other)      :  0                  ...
##     Rainfall       Evaporation        Sunshine       WindGust...
##  Min.   : 0.00   Min.   : 0.20   Min.   : 0.00   NW      : ...
##  1st Qu.: 0.00   1st Qu.: 2.20   1st Qu.: 5.95   NNW     : ...
##  Median : 0.00   Median : 4.20   Median : 8.60   E       : ...
##  Mean   : 1.43   Mean   : 4.52   Mean   : 7.91   WNW     : ...
##  3rd Qu.: 0.20   3rd Qu.: 6.40   3rd Qu.:10.50   ENE     : ...
....
```

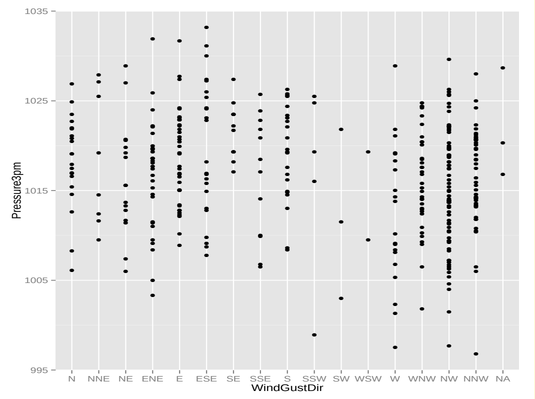
Visual Summaries—Add A Little Colour

```
qplot(Humidity3pm, Pressure3pm, colour=RainTomorrow, data=ds)
```



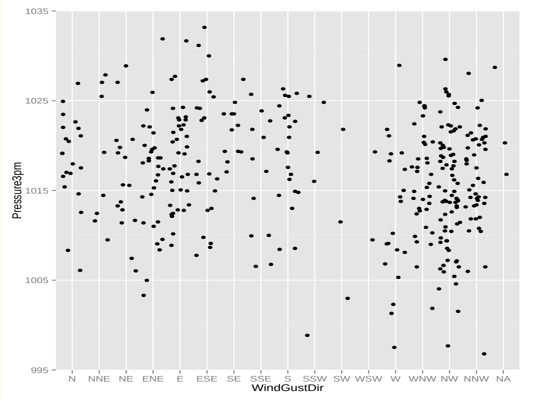
Visual Summaries—Careful with Categorics

```
qplot(WindGustDir, Pressure3pm, data=ds)
```



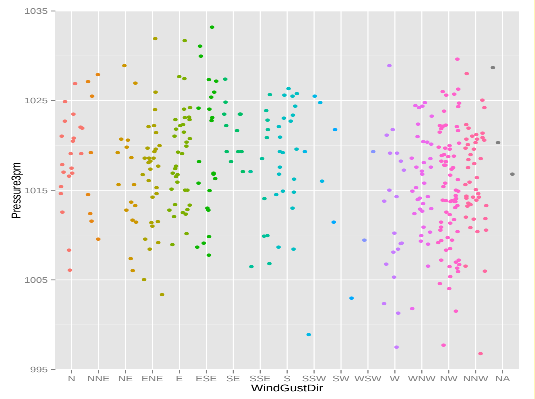
Visual Summaries—Add A Little Jitter

```
qplot(WindGustDir, Pressure3pm, data=ds, geom="jitter")
```

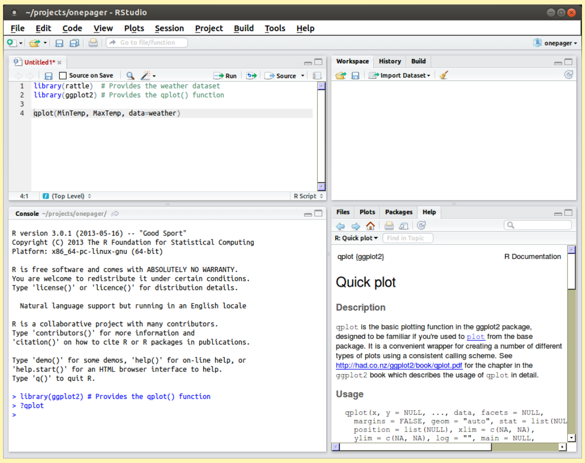


Visual Summaries—And Some Colour

```
qplot(WindGustDir, Pressure3pm, data=ds, colour=WindGustDir, geom="jitter")
```



Getting Help—Precede Command with ?



```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

qplot(MinTemp, MaxTemp, data=weather)
```

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(ggplot2) # Provides the qplot() function
> ?qplot
>
```

qplot {ggplot2} R Documentation

Quick plot

Description

qplot is the basic plotting function in the ggplot2 package, designed to be familiar if you're used to plot from the base package. It is a convenient wrapper for creating a number of different types of plots using a consistent calling scheme. See http://had.co.nz/ggplot2/book/qplot.pdf for the chapter in the ggplot2 book which describes the usage of qplot in detail.

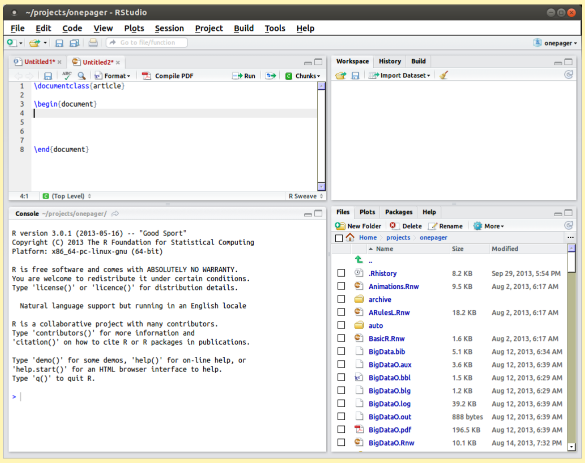
Usage

```
qplot(x, y = NULL, ..., data, facets = NULL,
  margins = FALSE, geom = "auto", stat = list(NUL
  position = list(NULL), xlim = c(NA, NA),
  ylim = c(NA, NA), log = "", main = NULL,
```

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

Create a KnitR Document: New→R Sweave


~/projects/onepager - RStudio
File Edit Code View Plots Session Project Build Tools Help
Go to file/function
onepager
Untitled1*
Untitled2*
Format
Compile PDF
Run
Chunks
4:1
(Top Level)
R Sweave
Workspace History Build
Import Dataset
Console ~/projects/onepager/
Files Plots Packages Help
New Folder
Delete
Rename
More
Home › projects › onepager

```
\documentclass{article}

\begin{document}




\end{document}
```

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Name	Size	Modified
..		
.Rhistory	8.2 KB	Sep 29, 2013, 5:54 PM
Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
archive		
ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
auto		
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM
BigDataO.Rnw	10.1 KB	Aug 14, 2013, 7:32 PM

Setup KnitR

We wish to use KnitR rather than the older Sweave processor

In RStudio we can configure the options to use knitr:

- Select Tools→Options
- Choose the Sweave group
- Choose **knitr** for *Weave Rnw files using:*
- The remaining defaults should be okay
- Click **Apply** and then**OK**

Simple KnitR Document

Insert the following into your new KnitR document:

```
\title{Sample KnitR Document}
\author{Graham Williams}
\maketitle

\section*{My First Section}

This is some text that is automatically typeset
by the LaTeX processor to produce well formatted
quality output as PDF.
```

Your turn—Click **Compile PDF** to view the result.

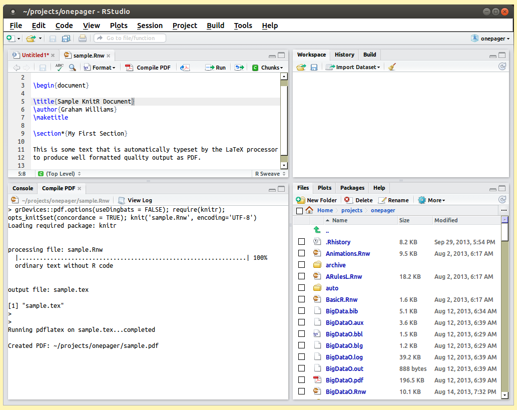
Simple KnitR Document

Insert the following into your new KnitR document:

```
\title{Sample KnitR Document}
\author{Graham Williams}
\maketitle

\section*{My First Section}

This is some text that is automatically typeset
by the LaTeX processor to produce well formatted
quality output as PDF.
```

Your turn—Click **Compile PDF** to view the result.

Simple KnitR Document



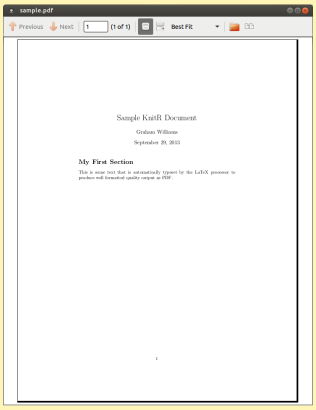
```

\begin{document}

\title{Sample KnitR Document}
\author{Graham Williams}
\maketitle

\section*{My First Section}

This is some text that is automatically typeset by the LaTeX processor
to produce well formatted quality output as PDF.

```

```
> grDevices::pdf.options(useDingbats = FALSE); require(knitr);
opts_knit$set(concordance = TRUE); knit('sample.Rnw', encoding='UTF-8')
Loading required package: knitr


processing file: sample.Rnw
  |.................................................................| 100%
  ordinary text without R code


output file: sample.tex

[1] "sample.tex"
>
>
Running pdflatex on sample.tex...completed

Created PDF: ~/projects/onepager/sample.pdf
```

Name	Size	Modified
..		
.Rhistory	8.2 KB	Sep 29, 2013, 5:54 PM
Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
archive		
ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
auto		
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM
BigDataO.Rnw	10.1 KB	Aug 14, 2013, 7:32 PM

Simple KnitR Document—Resulting PDF

Result of **Compile PDF**



Sample KnitR Document

Graham Williams

September 29, 2013

My First Section

This is some text that is automatically typeset by the LaTeX processor to produce well formatted quality output as PDF.

1

KnitR: Add R Commands

R code can be used to generate results into the document:

```
<<echo=FALSE, message=FALSE>>=
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
@
```

Your turn—Click **Compile PDF** to view the result.

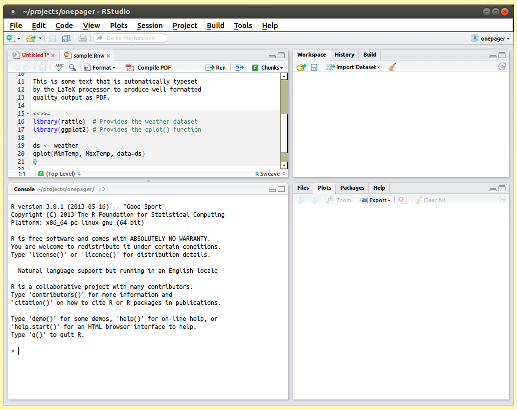
KnitR: Add R Commands

R code can be used to generate results into the document:

```
<<echo=FALSE, message=FALSE>>=
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
@
```

Your turn—Click **Compile PDF** to view the result.

KnitR Document With R Code

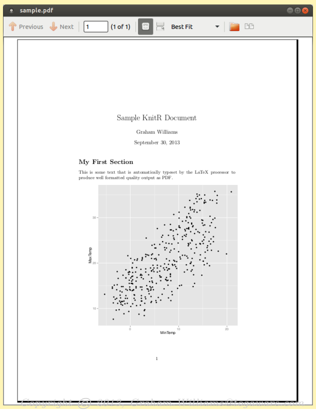


```
This is some text that is automatically typeset
by the LaTeX processor to produce well formatted
quality output as PDF.

<<>>=
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
@
```

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Simple KnitR Document—Resulting PDF with Plot

Result of **Compile PDF**



Sample KnitR Document

Graham Williams

September 30, 2013

My First Section

This is some text that is automatically typeset by the LaTeX processor to produce well formatted quality output as PDF.

1

LaTeX Basics

```
\subsection*{...}         % Introduce a Sub Section

\subsubsection*{...}      % Introduce a Sub Sub Section

\textbf{...}              % Bold font
\textit{...}              % Italic font

\begin{itemize}           % A bullet list
  \item ...
  \item ...
\end{itemize}
```

Plus an extensive collection of other markup and capabilities.

KnitR Basics

```
echo=FALSE        # Do not display the R code
eval=TRUE         # Evaluate the R code

results="hide"    # Hide the results of the R commands

fig.width=10      # Extend figure width from 7 to 10 inches
fig.height=8      # Extend figure height from 7 to 8 inches

out.width="0.8\\textwidth"    # Fit figure 80% page width
out.height="0.5\\textheight"  # Fit figure 50% page height
```

Plus an extensive collection of other options.

Tutorial Overview

1. Motivating R – A Language for Data Mining
2. Data Mining in R – Hands-on Rattle GUI
3. Programming Data in R – Scripting our Analyses
4. Disseminate Research in R – Ensembles and wsrf

Case Study – Ensembles in R

Major advances in Data Mining

- The best off-the-shelf technology includes random forests, boosting and support vector machines?
- Available for investigation now through open source solutions, with closed source tools catching up.

Case Study – Ensembles in R

Major advances in Data Mining

- The best off-the-shelf technology includes random forests, boosting and support vector machines?
- Available for investigation now through open source solutions, with closed source tools catching up.

Case Study – Ensembles in R

Major advances in Data Mining

- The best off-the-shelf technology includes random forests, boosting and support vector machines?
- Available for investigation now through open source solutions, with closed source tools catching up.

Introducing Random Forests

Research with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

- Random forests are a popular classification method building an ensemble of a single type of decision tree.
- It is unsurpassed in accuracy among current algorithms.
- Algorithmically intuitive and simple.
- It is used widely in numerous research domains including bioinformatics, image classification, text classification.

Random Forests Algorithm

- Build many decision trees (e.g., 500).
- For each tree:
  - Select a random subset of the training set (N);
  - Choose different subsets of features for each node of the decision tree ($m << M$);
  - Build the tree without pruning (i.e., overfit)
- Classify a new entity using every decision tree:
  - Each tree "votes" for the entity.
  - The decision with the largest number of votes wins!
  - The proportion of votes is the resulting score.

Using Weighted Variable Subspaces

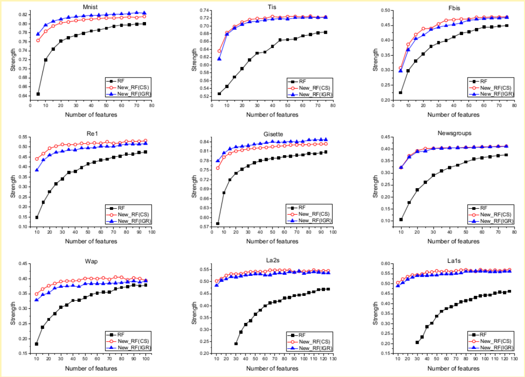
- Performance of a random forest is improved by
  - **Strengthening** each tree
  - Reducing **correlation** between each tree
- Problem of large number of variables:
  - Random selection means too many irrelevant variables
- Introduce the concept of weighted subspace random forests
  - Bias the selection of variables toward most important variables

Datasets

Name	# Features	# Train Set	# Test Set	# Classes
Mnist	780	60,000	10,000	2
Tis	927	5200	6875	2
Fbis	2000	1711	752	17
Re1	3758	1147	510	25
Gisette	5000	5000	1000	2
Newsgroups	5000	11,268	7504	20
Wap	8460	1104	456	20
La2s	12,432	1855	845	6
La1s	13,195	1963	887	6

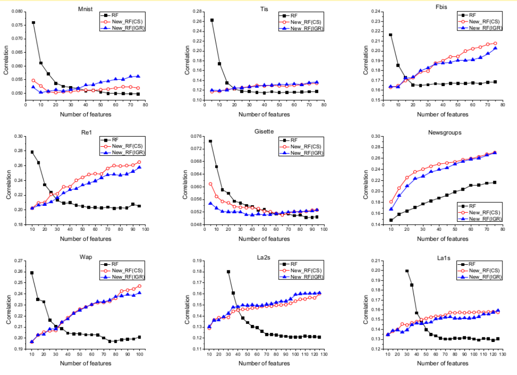
(From International Journal of Data Warehousing and Mining 2012)

Comparison of Strength vs Features



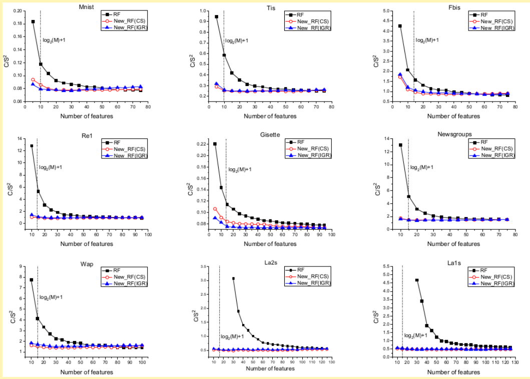
(From International Journal of Data Warehousing and Mining 2012)

Comparison of Correlation vs Features



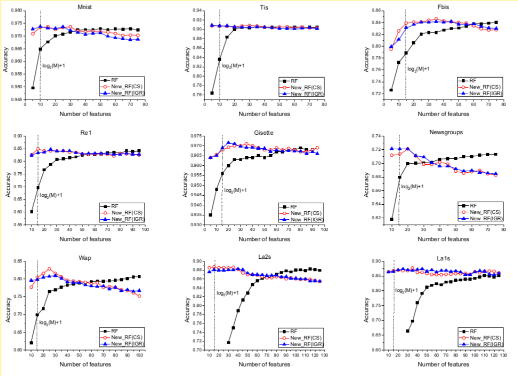
(From International Journal of Data Warehousing and Mining 2012)

c/s^2 vs Features



(From International Journal of Data Warehousing and Mining 2012)

Accuracy vs Features



(From International Journal of Data Warehousing and Mining 2012)

Repeatable and Transparent Research

- SIAT have implemented the research as C++ code
- Enhanced for parallel environment - multi-core and multi-node
- At least as good as random forest but always very much quicker.
- Integrated into R using the Rcpp package of R
- Now openly available for use and peer review:
  `install.packages("wsrf", repos="http://rattle.togaware.com")`
- Similarly wskm, wsrpart, eqrf.
- Will be published to CRAN shortly.
- **Publish-by-Example** — use wsrf as template
- C, C++, Fortran, Java (RWeka)

Using the Package

```
library(wsrf)
model <- wsrf(form, ds[train, vars])
pr    <- predict(model, na.omit(ds[test, vars]))
```

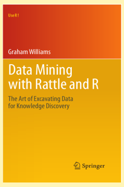
Tutorial Overview

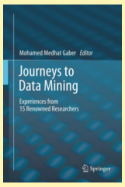
1. Motivating R – A Language for Data Mining
2. Data Mining in R – Hands-on Rattle GUI
3. Programming Data in R – Scripting our Analyses
4. Disseminate Research in R – Ensembles and wsrf

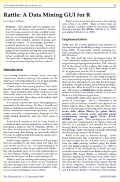
Resources and References

- **OnePageR**: `http://onepager.togaware.com` – Tutorial Notes
- Rattle: `http://rattle.togaware.com`
- Guides: `http://datamining.togaware.com`
- Practise: `http://analystfirst.com`

- Book: Data Mining using Rattle/R
- Chapter: Rattle and Other Tales
- Paper: A Data Mining GUI for R — R Journal, Volume 1(2)


Graham Williams
Data Mining with Rattle and R
The Art of Excavating Data for Knowledge Discovery
Springer


Mohamed Medhat Gaber Editor
Journeys to Data Mining
Experiences from 15 Renowned Researchers
Springer



Thank You

Question Time

This document, sourced from HefeiTutorial.Rnw revision 239, was processed by KnitR version 1.2 of 2013-04-10 and took 0.1 seconds to process. It was generated by gjw on theano running Ubuntu 13.04 with Intel(R) Core(TM) i7-3517U CPU @ 1.90GHz having 4 cores and 3.9GB of RAM. It completed the processing 2013-10-19 20:14:44.