DSCI 351-351M-451 Syllabus
Exploratory Data Science
Fall 2018 Tuesday, Thursday 11:30 am to 12:45 pm Olin 313

Prof. Roger H. French

September 4, 2018

1 Joint Undergraduate and Graduate Course

1.1 DSCI351 and DSCI351M

DSCI351 is the 3rd level class in the Applied Data Science UG Minor. The ADS Minor is available to CWRU students across all the schools in the University. For more information see http://bulletin.case.edu/schoolofengineering/datascience/.

DSCI351 will introduce students to the basic elements of a data analysis, including R coding, Rstudio IDE and Git version control, statistical concepts, the stages of a data analysis and reproducible research.

DSCI351M section focuses specifically on Exploratory Data Science for Materials and Materials' Systems.

1.2 DSCI451

DSCI 451 is a graduate level introduction to data science and analytics. Graduate students will develop a 10 week data science project focused on time-series, spectral, image, statistical process control or machine learning data science problems that are relevant to their research interests. These projects will include preparing datasets, code scripts and functions, a git repository for other students to use these codes as open source resources, and the preparation of reproducible data science analyses for these problems.

2 Course Description

In this course, we will learn data science and analysis approaches to identify statistically significance relationships and better model and predict the behavior of these systems. We will assembly and explore real-world datasets, perform clustering and pair plot analyses to investigate correlations, and logistic regression will be employed to develop associated predictive models. Results will be interpreted, visualized and discussed.

We will introduce basic elements of statistical analysis using R Project open source software for exploratory data analysis and model development. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and munging functions, and a rich selection of statistical packages, used for data analytics, model development and prediction. This will include an introduction to R data types, reading and writing data, looping, plotting and

regular expressions, so that one can start performing variable transformations for linear fitting and developing structural equation models, while exploring for statistically significant relationships. We will also learn tidy principles for data analysis, pipes and ggplot vs base graphics approaches to data visualization.

R Analytics will be applied to the case of time-series, spectral, image and categorical datasets and problems by analyzing system responses, combined with results of experiments to identify fundamental principles that are statistically significant in the observed system performance. We will learn about longitudinal and cross-sectional data science studies, along with retrospective and prospective studies.

The class will be structured to have a balance of theory and practice. We'll split class into Foundation and Practicum a) lectures, presentations, discussion b) coding, demonstrations and live hands-on data science work.

Every student will have access to their own pre-configured Open Data Science VDI computer, already configured for fast and easy adoption of good data science practices and tools. And the ODS VDIs are updated each month so as to have the latest versions of R packages and software for data science.

2.1 Class Repository Folder Structure

Please browse within each folder to learn more about the intended purpose of each folder in the standard structure.

This folder structure has been designed to accomodate each type of file you may need to create and modify - please do not create additional folders in the structure, and please pay attention to naming conventions when creating new files.

Course Material Folders in this Couse Repo

- 1-Assignments is where you will find the Lab Exercises and Exams
- 2-Readings folder contains textbooks and readings for the course
- 3-Class contains daily class notes as *.Rmd and *.pdf files
- These are split into a Foundations "f" and a Practicum "p" class notes
- 4-Syllabus contains the updated Course Syllabus

Your Working Data Analysis Folders

- Scripts is where to write you scripts for data analysis
- Data contains course datasets and your datasets
- Figs is the figures folder, accessible for both Scripts, Topics and Docs
- Topics is where to write reports and presentation for your data analysis
- Docs is where to write formal documentation as *.Rmd, *.tex files
- Packages is where to build R packages, if you project involves this

2.2 License applied to course materails and some datasets used for data analysis

Class materials

- License: This work is legally bound by the following software license: [CC-A-NS-SA-4.0][1]
- Please see the LICENSE.txt file, in the root of this repository, for further details.

Assessment materials

- Homework Assignments, Project Assignments and Exams are all rights reserved.
- They are NOT creative commons licensed, and can not be distributed.

Datasets derived from funded research projects

- During this class you may be working on a project that is part of a funded research award at Case Western Reserve University.
- Information or material made available to you in connection with this funded research project, and coursework, data, results or other intellectual property you may develop in conjunction with this project, will be subject to Case Western Reserve's Intellectual Property Policy as well as terms of the sponsored research agreement.
- You acknowledge that you understand that you will not have ownership of intellectual property created in conjunction with the project.
- Please sign the "1801-DSCI-Acknowledgement-of-IP.pdf" form.

2.3 Outcomes

Capabilities

- Familiarity with R Statistics, scripting, functions, packages, automated data analysis.
- Familiarity with exploratory data analysis, statistical model building.
- Familiarity with Principles of Tidy Data Science from the tidyverse package an use of pipes (% > %) from the magrittr package.
- Applications of domain knowledge and statistical analytics to identify important predictors and develop initial predictive models.
- Applications of domain knowledge and analytics to identify important predictors and develop initial predictive models.
- Introduction to methods of reproducible research, including markdown, LaTeX and Git.

Data types include:

- Time-series, spectral, image and higher order datatypes,

  And their assembly to produce augmented and derivative datasets.

Data set characteristics will include:

- Variety: Types of data and information, including both structured and unstructured data.
- Volume: Data from human sources (vendors, suppliers, distributors, customers, etc.) and sensor networks, both small and large data volumes.
- Velocity: Short time interval datasets.

2.4 Prerequisites

1. ENGR131 Elementary Computer Programming, EECS132, DSCI134 or equivalent
2. STAT312R Basic Statistics for Engineering and Science or equivalent

3 Homework, Project, Report-Presentation Grading

DSCI351 and DSCI351M is graded on 100 points basis

$$Six\,Homeworks,\ worth\,5\,points\,each = 30pts.$$
$$Four\,Data\,Analyses,\ worth\,10\,points\,each = 40pts.$$
$$Midterm\,Exam = 10pts.$$
$$Final\,Exam = 20pts.$$
$$\textbf{Total = 100pts}$$

DSCI451 is graded on a 140 point basis

$$Six\,Homeworks,\ worth\,5\,points\,each = 30pts.$$
$$Four\,Data\,Analyses,\ worth\,10\,points\,each = 40pts.$$
$$One\,Semester\,Project\,with\,3\,presentations\,in\,a\,relevant\,data\,science\,domain = 40pts.$$
$$Midterm\,Exam = 10pts.$$
$$Final\,Exam = 20pts.$$
$$\textbf{Total = 140pts}$$

Grading is done on a curve

No.	351-451 Homework	351-451 Projects	451 SemProjects
1	DataSci Question	Cleaning, Assembly, Summarize	Data Sources, Tidying & Question,
2	Stats Questions	Assembly, Pairs, EDA & Packages	Context, Hypothesis, Techniques & Approach
3	Get Data, Clean, Tidy, Assemble, DataViz	EDA and Linear Models & Insights	EDA, Visualization, Models & Variable Selection
4	Tidy vs. Non-Tidy, Pipes vs For loops, Functions & Insights	Tidy Analysis, EDA & Insights	Reprodicble Analysis, Answers & Insights
5	Linear & Nonlinear Models		
6	Multiple Regression and Variable Selection		

Table 1: DSCI351-451 Weekly Syllabus.

Figure 1: **PRP**: Roger Peng, **R Programming for Data Science**. 2014 [1]

Figure 2: **EDA**: Roger Peng, **Exploratory Data Analysis With R**. 2015 [2]

4 Textbooks and Readings

Required Texts and their Abbreviation, which is used in the syllabus:

Peng R Programming (PRP) and Peng Exploratory Data Analysis (EDA) are introductory books to R and Data Science and Analysis. These are Leanpub books, available from LeanPub for a "pay what you want" price.

R for Data Science (R4DS) is a new book teaching R for Tidy Data Science, and is available as a bookdown book on the web R for Data Science, you can buy an ebook for $18 from Google Play Store page for this book.

Open Intro Statistics version 3 (OIS) is an open source text book on Inferential Statistics, published under a Creative Commons license, for free distribution as a pdf. In addition a copy can be purchased from Amazon for $9.

Introduction to Statistical Learning with R (ISLR) is a Springer book which is also available for free as a pdf. ISLR is the text book used extensively in DSCI353/453 on Statistical Learning.

Additional reading assignments will be distributed via the course git repository in the readings subdirectory.

Figure 3: **OIS**: David M. Diez, Christopher D. Barr, and Mine Cetinkaya-Rundel, **OpenIntro Statistics 3rd Ed.** 2015 [3]

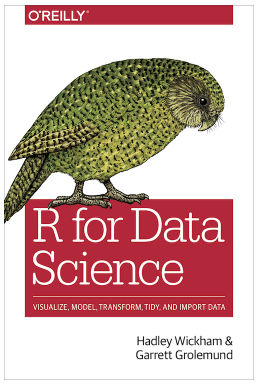

Figure 4: **R4DS**: Garrett Grolemund, Hadley Wickham **R for Data Science**. 2017 [4]

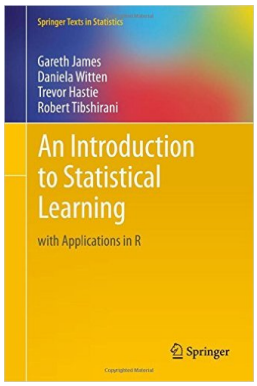

Figure 5: **ISLR**: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani **An Introduction to Statistical Learning: with Applications in R**, 2013 [5]

5 DSCI351-451 Syllabus: Weekly Topics

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1**,
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Table 2: DSCI351-451 Weekly Syllabus. Peng R Programming (PRPx.y), Peng Exploratory Data Analysis (EDAx.y), R for Data Science (R4DSx.y), Open Intro Statistics (OISx.y) and Introduction to Statistical Learning with R (ISLRx.y) refers to chapters and sections assigned as reading.

6 Contact Information

Prof. Roger H. French

- White 536, and electronically.
- Email is best: rxf131@case.edu, Use DSCI351 in the subject line
- @frenchrh on twitter
- Office Phone 216 368 3655, Cell Phone 302 468 6667

TA: JiQi Liu

- White 540, and electronically.
- Email is best: jxl1763@case.edu, Use DSCI351-451 in the subject line
- @liu_jiqi on twitter
- Lab Phone 216 368 0135, Cell Phone: (216) 303-0733

7 Course Mechanics

7.1 Lectures

Fall 2017 Tuesday, Thursday 11:30 am to 12:45 pm Olin 303

7.2 Consultations

After class or as needed. Contact Prof. French and Alan Curran, email or in person.

7.3 Homework Assignments

All homeworks assignments are submitted electronically through blackboard, uploading to the HW assignment page.

Filenames should contain DSCI351, -HW#, -YourLastName... e.g. DSCI351-HW2-French.R or DSCI351-HW2-French.R and DSCI351-HW2-French.pdf.

Homeworks need to be legible, organized and explain your thinking, process and results. Credit all resources you drew upon, including texts, papers, peers.

Homeworks are due by 11 am Tuesday, prior to the beginning of class. Homeworks will be graded on blackboard and reviewed in class.

8 Coding and Data Science Tools and Resources

Open Data Science (ODS) VDIs

You will not need to install software on your personal computers.
Instead you can install the Citrix Reciever [6] and then login to the CWRU CSE Portal. [7]
The CSE Portal is located at https://cseportal.cwru.edu/vpn/index.html

Scripting, Coding and Writing

And more resources for open science coding and scripting, including tools for code editing, code version control and languages.

R Statistics

We will be using R in this class for homeworks and projects. Its generally useful language for statistical analysis and data science.

- The R Project for Statistical Computing [8] main website
- R programming language R is a free software programming language and software environment for statistical computing and graphics.[9]
- RStudio provides popular open source and enterprise-ready professional software for the R statistical computing environment. [10]
- Google's R Style Guide

Rmarkdown as a path to open access and reproducible science

- R Markdown — Dynamic Documents for R. We will be doing all our work using Rmarkdown this semester. Class presentations, homeworks, projects, all done in Rmd, as reproducible science projects, including data, code, and final output.
- Introduction to R Markdown.
- R Markdown Cheat Sheets.
- An Rmarkdown Introduction slidedeck done from Rmarkdown and shared publicly on RPubs.

R Statistics, more resources

We will be using R in this class for homeworks and projects. Its generally useful language for statistical analysis and data science.

- The R Project for Statistical Computing [8] main website
- Roger Peng's Computing for Data Analysis introduction to R Statistics. These are from a Coursera course he does, with the same name. [11]
- A (very) short introduction to R [12]
- Google's R Style Guide
- Hadley Wickham's R Style Guide
- RStudio's R Cheatsheets for Rmarkdown and Data Wrangling
- An Rmd slideshow Intro to R

Open Source software and tools

- FOSS (Free and Open Source Software) is a copyleft approach to software which is hat is distributed in a manner that allows its users to run the software for any purpose, to redistribute copies of, and to examine, study, and modify, the source code. [13]

- vim (or Gvim the gui version) is a powerful text and code editor, that is universally available on all linux and mac computers.[14] NeoVim is a new Gvim fork.[15] It can be installed on windows computers, its available on the ODS VDIs.. [14]
- Git (Wikipedia) is a distributed content versioning system that is very popular. It enables collaborative code development and LaTeX writing projects.[16]
- Git server software is installed on each computer.[17]
- GitHub is a Git server website used for collaborative code development.[18]
- BitBucket is a Git server website used for collaborative code development. If you join with your case.edu email address, you get unlimited private repositories.[19]
- Stack Exchange [20] Code Question and Answer Websites: covering R, Python, Mathematica, LaTeX and many other things, such as English or Spanish etc.

Python (is also used for Data Science in many cases. But here we will focus on R first.

- Wikipedia: Python is a widely used general-purpose, high-level programming language. [21]
- The Python main website. [22]
- The Python Tutorial — Python v2.7.8 documentation [21]
- The Hitchhikers Guide to Python. This is an open access book being hosted on developed on GitHub and is located here https://github.com/vuvlab/python-guide. [23] [24]
- NumPy is the fundamental package [25] for scientific computing with Python.
- FiPy: Partial Differential Equations with Python [26]
- SciPy is a python-based ecosystem [27] of open-source software for mathematics, science, and engineering.
- PythonXY - Scientific-oriented Python Distribution based on Qt and Spyder that runs on Windows. [28]
- IPython Shell and Notebook [29]
- Spyder is the Scientific PYthon Development Environment [30]

LaTeX is used for publication quality writing. Its also the backend for Rmarkdown's pdf generation. It lets you write professional looking papers, theses and books, along with presentations.

- LaTeX is a program for writing documents, paper, journal articles, presentations and theses. [31]
- LaTeX - Wikibooks, open books for an open world. [32]
- Zotero Reference-Citation Manager, BibTeX Client [33]

9 Policies

9.1 Attendance

You attendance is expected. Some information is covered that is not in the text. Student participation is an important part of the class.

9.2 Readings

Readings must be done, BEFORE the class, where they are assigned. The reading assignment, is for the class with which it is listed.

9.3 Homework Assignments

Homeworks are due before noon on Monday after the week they are assigned. A 50% deduction will be assessed for submissions not received on Blackboard by noon on Monday.

9.4 Collaboration and Citation

Discussions and working together (except on exams) is acceptable and encouraged. It is not ethical to do someone else's work or to have someone do your work. You must cite all resources you used to work on your homework and projects. Citations should be done at the end of the document. These can be to books, Wikipedia and other web resources, and discussions with other students.

9.5 Academic Integrity Policy

All students in this course are expected to adhere to University standards of academic integrity. Cheating, plagiarism, misrepresentation, and other forms of academic dishonesty will not be tolerated. This includes, but is not limited to, consulting with another person during an exam, turning in written work that was prepared by someone other than you, making minor modifications to the work of someone else and turning it in as your own, or engaging in misrepresentation in seeking a postponement or extension. Ignorance will not be accepted as an excuse. If you are not sure whether something you plan to submit would be considered either cheating or plagiarism, it is your responsibility to ask for clarification.

For complete information, please go to https://students.case.edu/community/conduct/aiboard/policy.html.

9.6 Disability Resources

ESS Disability Resources is committed to assisting all CWRU students with disabilities by creating opportunities to take full advantage of the University's educational, academic, and residential programs.

For further information, please go to https://students.case.edu/academic/disability/.

10 Copyleft, References, Citations & Rubrics

10.1 CopyLeft

Creative Commons plays an important role in openness and open science, open data, open source efforts.

This DSCI351-451 class [34] is covered by a Creative Commons [35] copyleft licenses.

The license we'll use for class materials, code and presentations is covered by the "Attribution-ShareAlike 4.0 International" license, which is commonly called the CC BY-SA 4.0 license. [36]

More information on licensing open works, can be found on Wikipedia. [37]

GNU [38] is the developer of the GPL License [39] that is used for many open source software projects, such as Linux.

11 Setting up your R data science computer

If you do want to install the softwares on your personal computer, here's how.

11.1 For Windows

In Windows we are allowed to use spaces in filenames, however, most other systems does not support that. To avoid conflicts or troubles, we suggest using camelBack naming convention or use "-" or "_" to replace spaces.

11.1.1 LaTeX

LaTeX is a document preparation system that is widely used in the academia for producing scientific documents. You will need to install two softwares, Miktex and TeXstudio.

- Download and run the Basic MiKTeX Installer. MiKTeX has the ability to install missing packages automatically, i.e., this installer is suitable for computers connected to the Internet. Before you run the installer, you can check the prerequisites. The installer is available on the download page. You start it with a double-click on the downloaded file.
- Read the Copying Conditions carefully and click "I accept the MiKTeX copying conditions", the click "Next", as demonstrated below.

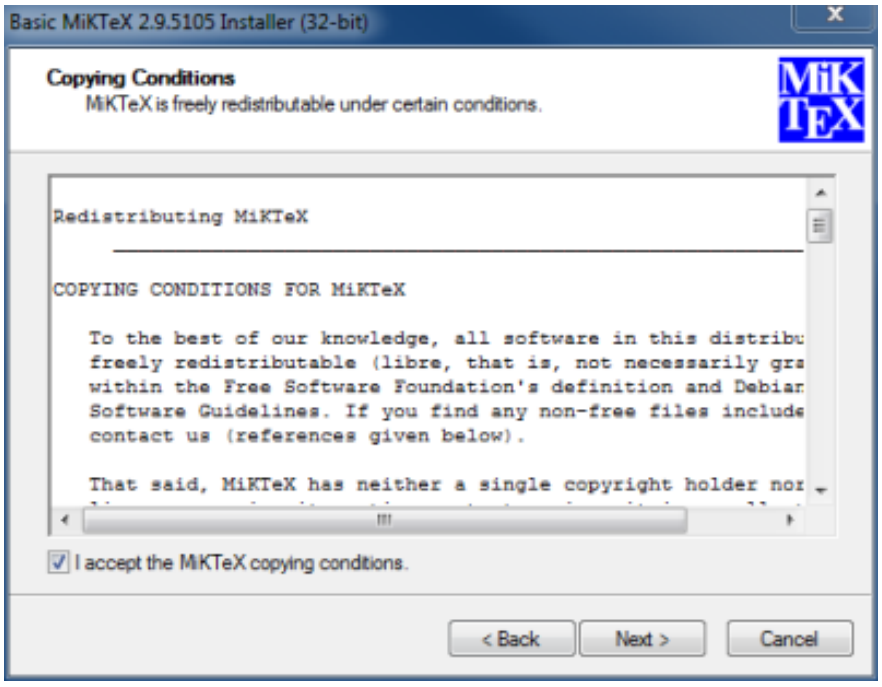

Figure 6

- You have the Option to create a shared MiKTeX installation. Click "Anyone who uses this Computer (all users), if you want to install MiKTeX for all users. Click "Only for ...", if you want to install MiKTeX for yourself only. When you have made your decision, click "Next" to go to the next page.

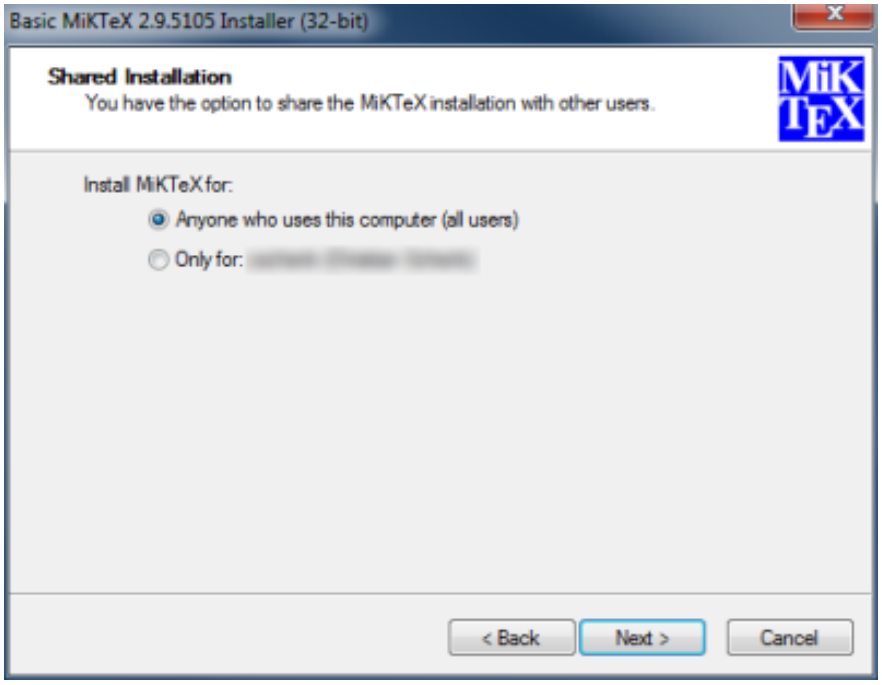

Figure 7

- You can specify the directory where you want to install your Miktex. Click "Browse", if you want to specify another (than the default) directory location. Click "Next", to go to the next page.

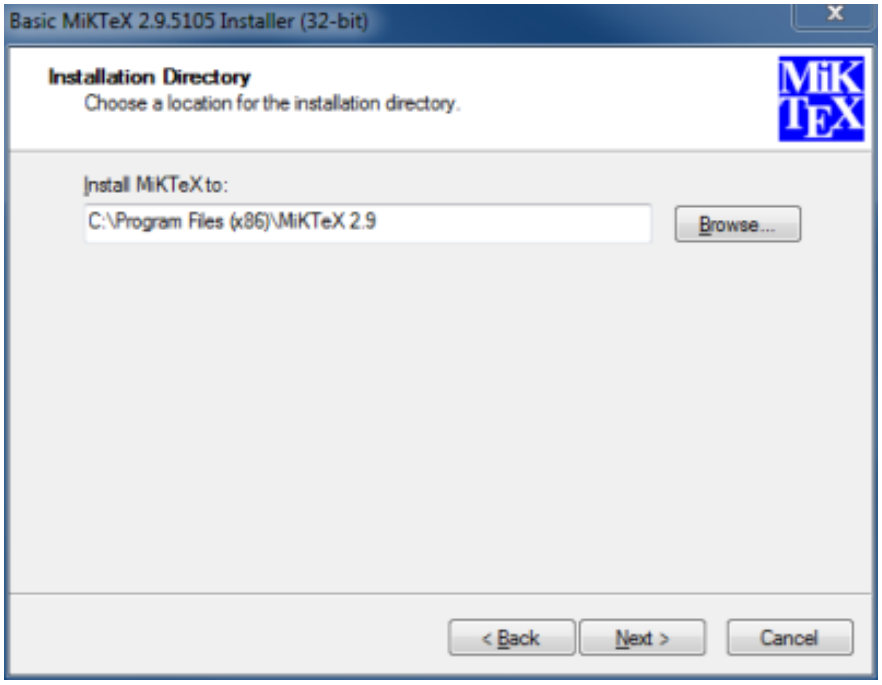

Figure 8

- The installer allows you to set the preferred paper size(usually it's A4 in China and letter size in the US). You also have the option to change the default behavior of the integrated package manager for the case where a required package is missing. Select "Yes", to make the package manager is always allowed to install missing packages. All these configurations can be changed later.

  Click "Next", to go to the next page.

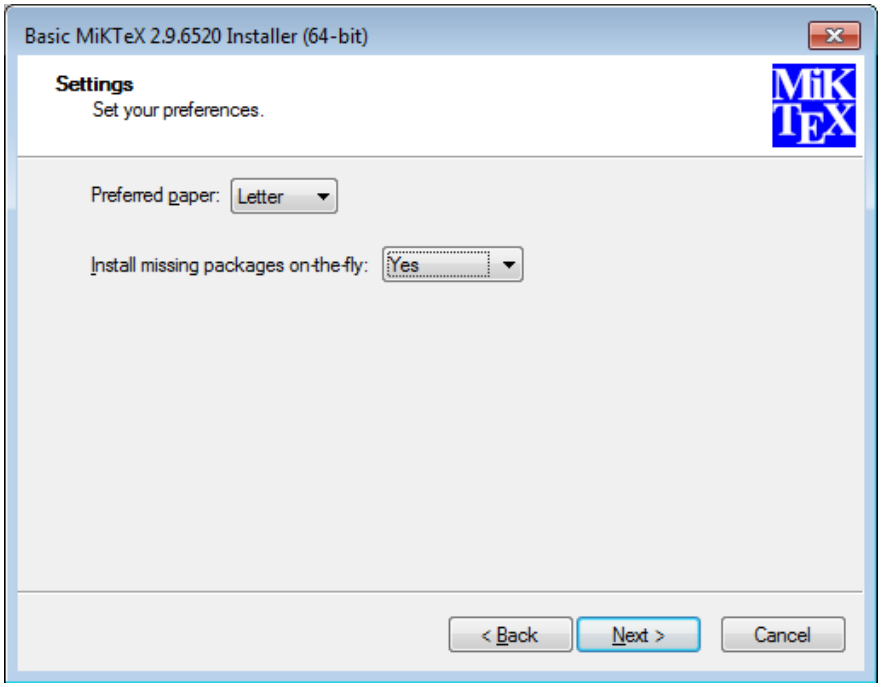

Figure 9

- Before the actual installation process begins, you get a chance to review your decisions. If you are satisfied with the settings, then click "Start" to start the actual installation.
- The installation will take a few minutes. The progress bar shows an approximate percentage of completion. When the installation has finished, you can click "Next" to open the last page.
- MiKTeX is now installed. Click "Close", to close the installer.
- In order to make use of latex the easiest way is to use a integrated development editor (IDE). **TeXstudio** is an free package that allows you to edit tex documents, compile and view them, it has syntax highlighting, auto completion, in line spell and grammar checker and much more. You can find the downloads page here and click on **download now**.
- Once downloaded, run and start the installer.
- Accept all the default conditions, and start up TeXstudio to finish.
- If you need instructions on how to start using LaTeX, here are some tutorials.

11.1.2 Git

Git Bash is command line programs which allow you to interface with the underlying git program. Bash is a Linux-based command line, which has been ported over to Windows.

- Download latest version of Git Bash on the official website.
- Once Git Bash Windows installer is downloaded, run the executable file and follow the setups:
- Agree to the GNU General Public License and click "Next".

Figure 10

- Select the location where you want to install the Git Bash.

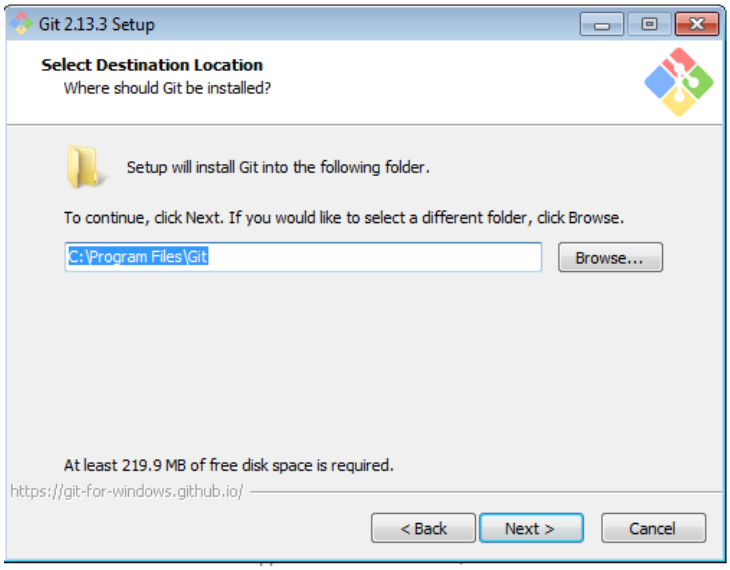

Figure 11

- Select the components you want to install and click Next. We suggest that you should unselect Windows Explorer integration.

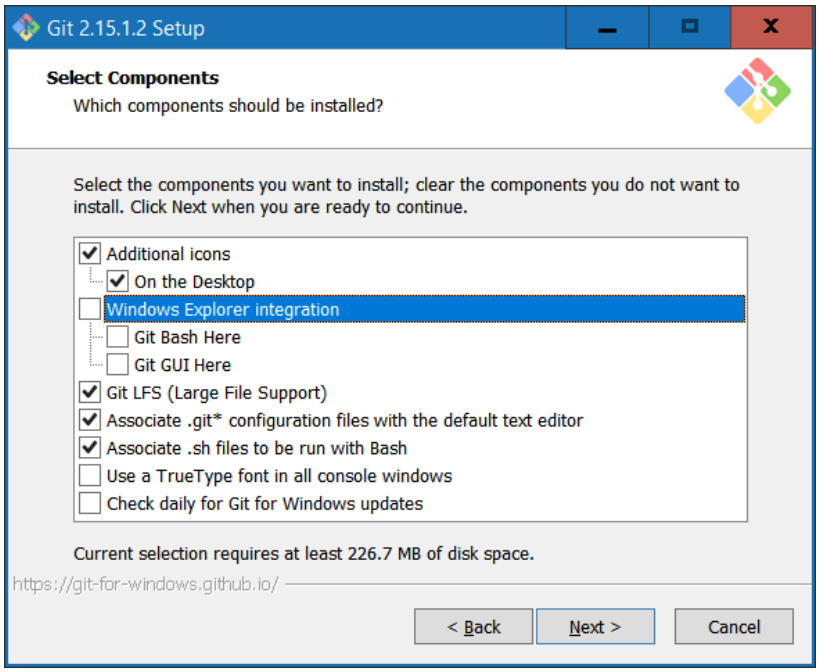

Figure 12

- Set default editor to Vim(which is the default option).

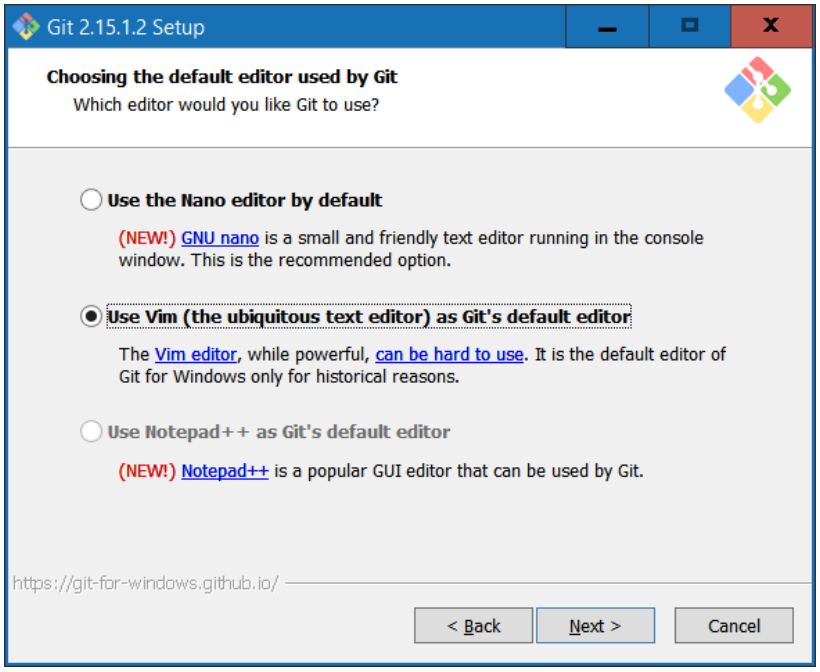

Figure 13

- We suggest that you use the default option, which is "Use Git from Git Bash only".

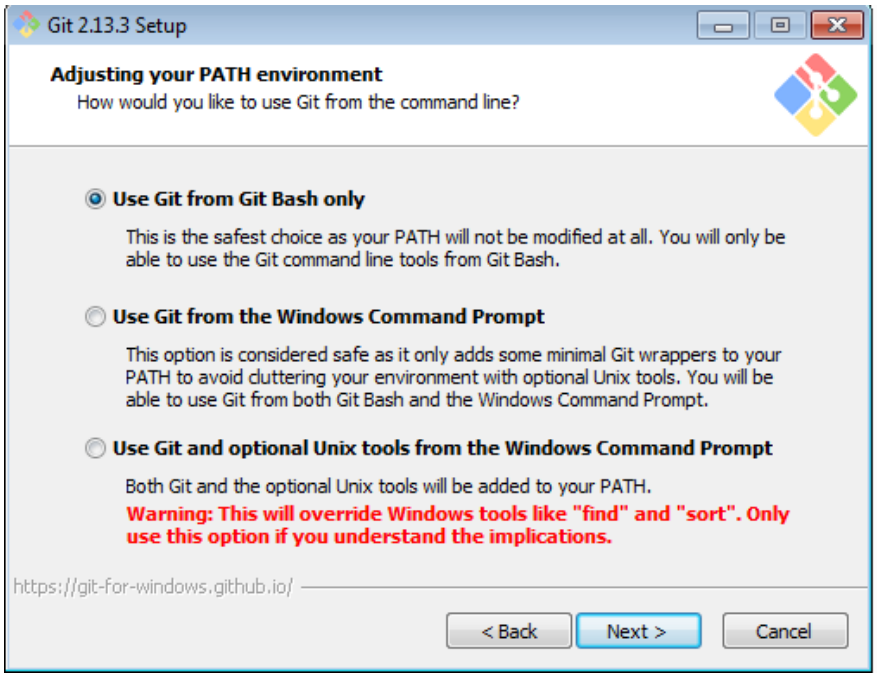

Figure 14

- Select which SSL/TLS library would you like to use for HTTPS connection and click Next.

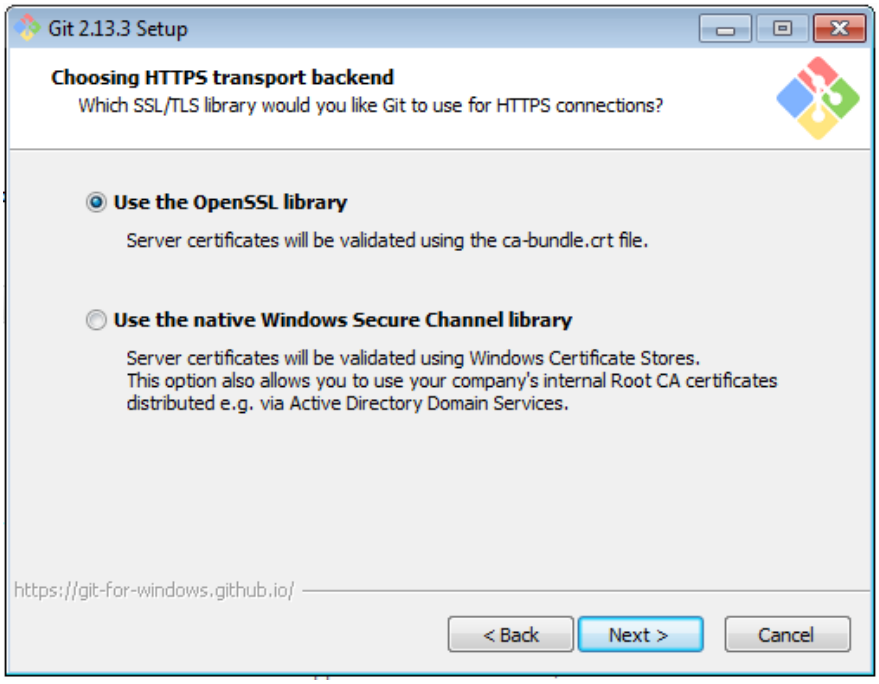

Figure 15

- Select, how should Git treat line endings in text files and click Next.

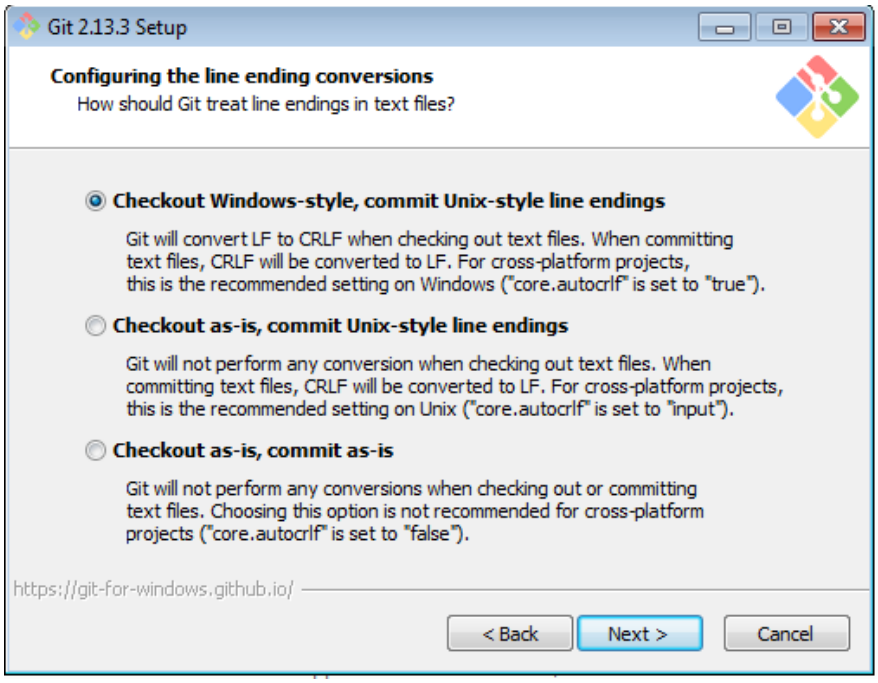

Figure 16

- Select the terminal you want to use for Git Bash.

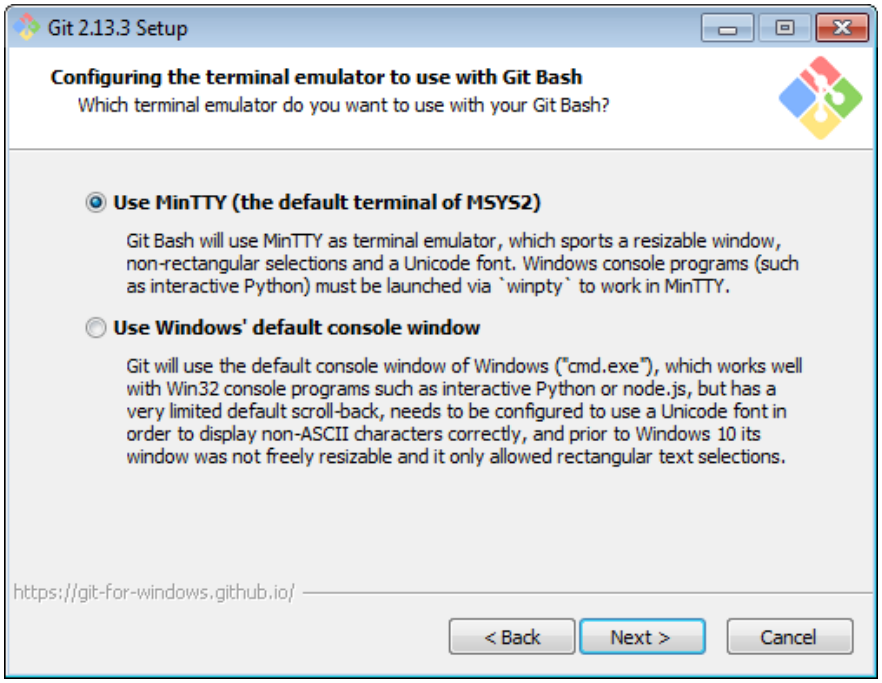

Figure 17

- Select the features you want to enable and click "Next". We suggest that you unselect "Enable Git Credential Manager".

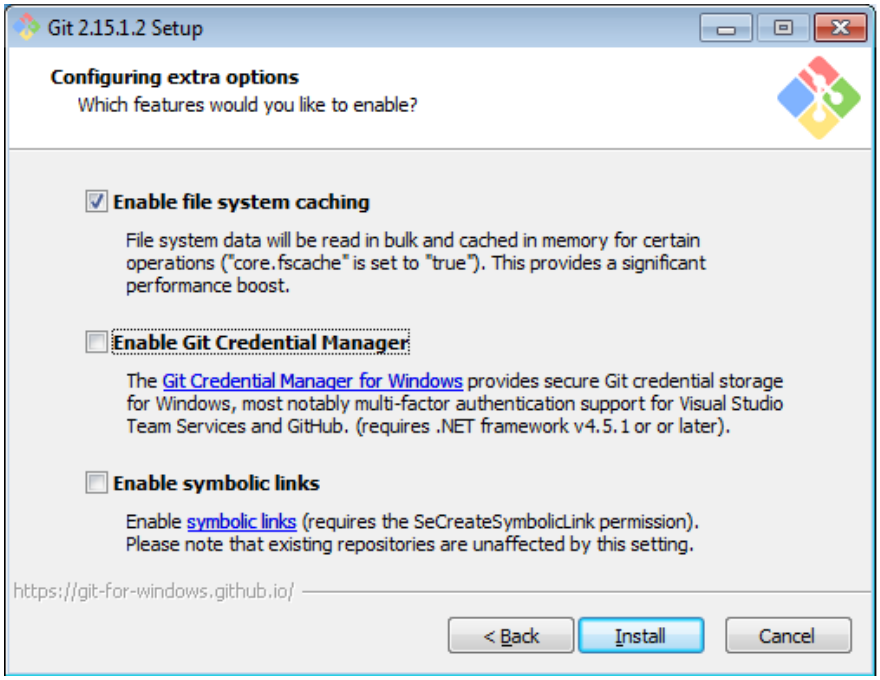

Figure 18

- Please wait while Setup wizard installs Git on your computer and click "Finish" to exit the Setup wizard.
- After Git Bash installation finishes you will ready to use the Linux command on a windows machine. Double click on below icon to start the Git Bash.

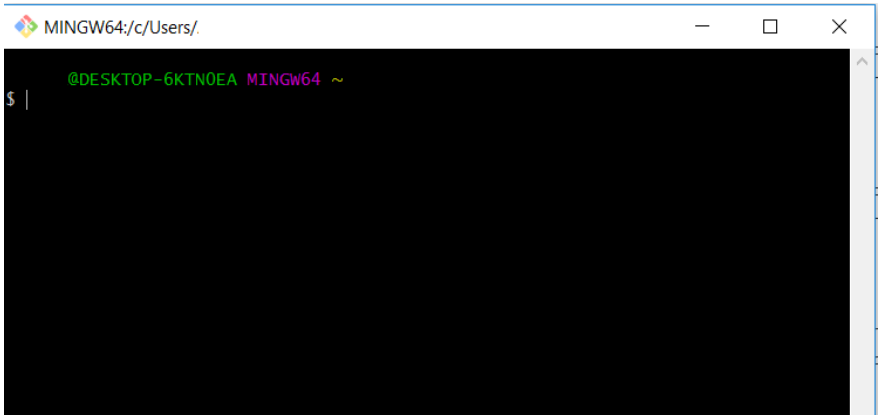

Figure 19

- Set up user name and email in Git.

```
git config --global user.name "yourusername"
git config --global user.email "youremail@website.com"
```

```
git config --global color.ui auto
```

- Here are some common commands you can use in git:
  pwd – present working directory
  cd – change directory
  ls – list files in current working directory
  mkdir – make new directory

- If you want to learn more about how git works (pull request, merge and more), you can read some tutorials.

11.1.3 R

R is a free programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing, while RStudio is a free and open-source integrated development environment for R.

- To download R, please choose your "install R for Windows" and then choose base R for a complete installation.
- Double click on the installer, and follow the instructions.
- Users of Vista/Windows 7/8/Server 2008/2012 installing for a single user using an account with administrator rights should consider installing into a non-system area (such as C:\R).
- Please try to avoid spaces or any special characters other than English letters and numbers in your installation directory, which may cause error later.
- After installing R, you can download **Rstudio** here, and choose the RStudio Desktop Open Source License version (the left most one).
- Run the installer and follow the installation instructions.
- Again, please try to avoid spaces or any special characters other than English letters and numbers in your installation directory.
- Rstudio have some built-in packages such as tidyverse and ggplot2, but if you are interested in building your own R packages, you can download **Rtools**. Please choose the latest version, as the older versions are not compatible with latest release of R.
- Run the installer, and accept the defaults throughout.
- Confirm and finish the installation.
- Once the Rtools installation completes, open RStudio and go to Profile–Global options–Code and change the code editing options as follows:

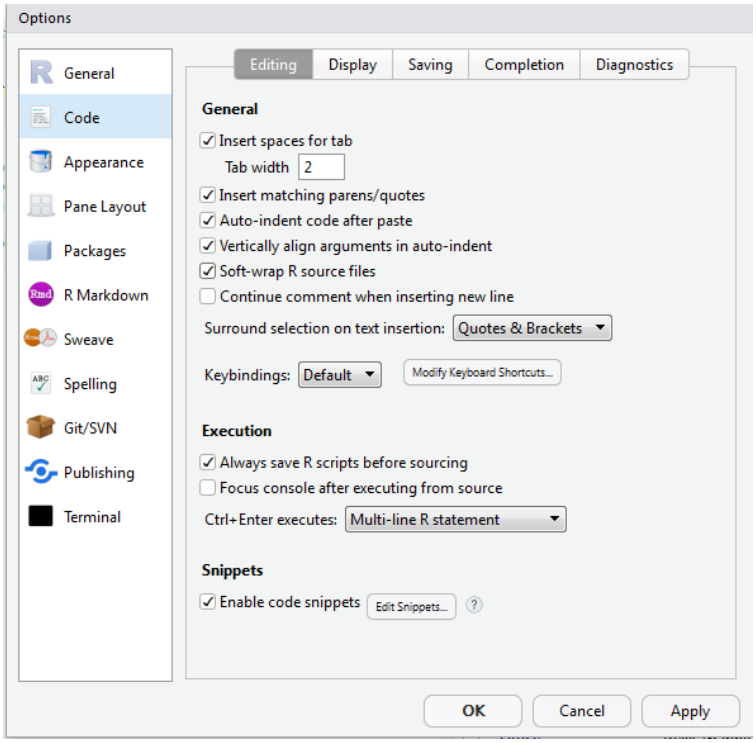

Figure 20

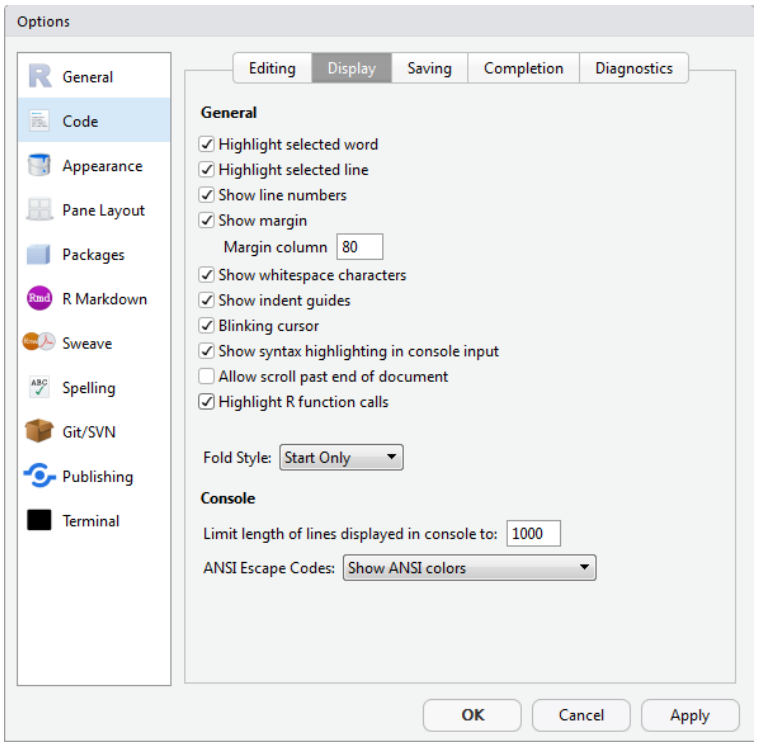

Figure 21

- You can also change your appearance style in Global Options:

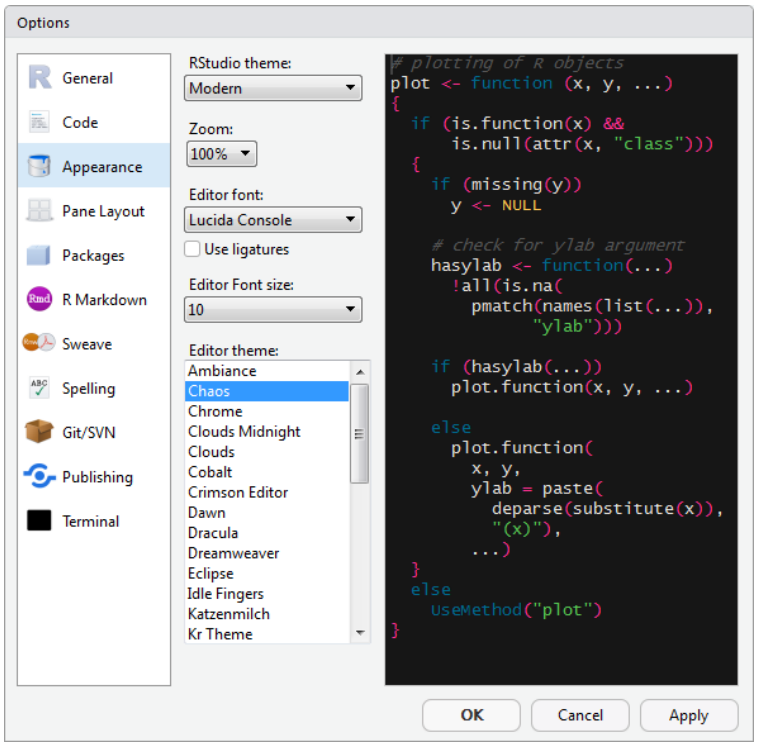

Figure 22

- Here is the list of standard packages that we suggest you install. If a warning comes up asking whether you want to install packages from the source, answer "y" for yes.

  Alphabetical packages:
  acepack akima animation anomalyDetection anytime arules aspace astsa bagRboostR baseline bcpa bda birk bit64 blogdown bookdown bookdownplus BoomSpikeSlab boot breakDown breakpoint brms broom bsts C50 car caret caretEnsemble Causalimpact changepoint changepoint.np ChemoSpec cgwtools class ClimClass cluster.datasets CORElearn corrgram cpca data.table data.tree dataMaid DBI DBItest dbscan devtools DiagrammeR DiagrammeRsvg digest doParallel dplyr dummies dtwclust e1071 eemR ElemStatLearn factoextra fastcluster feather flexclust forecast foreach gam gapminder gbm gclus GGally ggbiplot ggmap ggplot2 ggRandomForest s ggraph ggridges ggthemes ggvis glmnet gmodels googleVis gridBase gridExtra h2o HadoopStreaming HarmonicRegression hcp hdpca hexbin HH httr htmlwidgets hyperSpec igraph ipred IQCC ISLR itertools jsonlite kableExtra keras kernlab keyring kgc klaR knitcitations knitr Lahman lars lavaan lavaan.survey leaps learningr learNN learnBayes lme4 logitnorm magick magrittr Make Mapmate maptools MASS Matrix MatrixModels matrixStats markovchain mcmc MCMCglmm metRology Metrics mgcv minpack.lm MTS multiway NbClust netSEM neural neuralnet NeuralNetTools nnet nycflights13 odbc OIdata olsrr OIsurv onehot onlineCPD openintro optimx packrat pacman parallelSVM pca3d PerformanceAnalytics pipeR plot3D plotmo plotKML plotly pls Plumber plyr plyrmr png pool pROC prophet profvis propagate proxy pryr psych purrr qcc qtlmt qualityTools quantmod r2d3 randomForest randomForestSRC ranger raster rasterVis RColorBrewer Rcpp RCurl Rd-

pack readr RefManageR relaimpo reshape reshape2 reticulate rgdal rgeos rggobi rgl rJava rjson RJSONIO rlist rmarkdown Rmisc Rmpi RMySQL RNiftyReg rNMF roxygen2 rpart rprojroot rPython RSNNS rstan rstanarm rsvg RTextTools rticles Rtsne rtweet RUnit rvest rworldmap rworldxtra scatterplot3d scrypt segmented sem shiny shinydashboard shinystan shinytest shinythemes signal simpleNeural SixSigma sp sparklyr spc sqldf sqliter sqlutils stationaRy stlplus stockPortfolio StreamMetabolism stringi svglite svUnit SwarmSVM TeachingDemos TeachingSampling tensorflow testthat tfdatasets tfestimators tfruns tibble tictoc tidyr tidytext tidyverse tidygraph timeDate tinytex tm tree tweezer validate vtreat WaveletComp wavelets wavethresh wmtsa WGCNA WDI wordcloud XLConnect XML xtable xts zipcode zoo

You can go the the highlighted tab in below picture and install/upgrade you packages here. To install, simply paste the list of packages in the window.

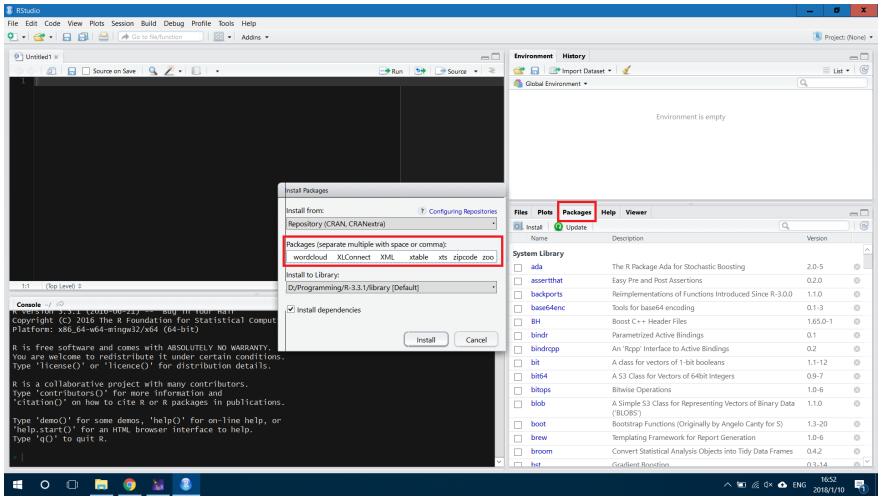

Figure 23

To upgrade your packages, select all packages and press "Install Updates"

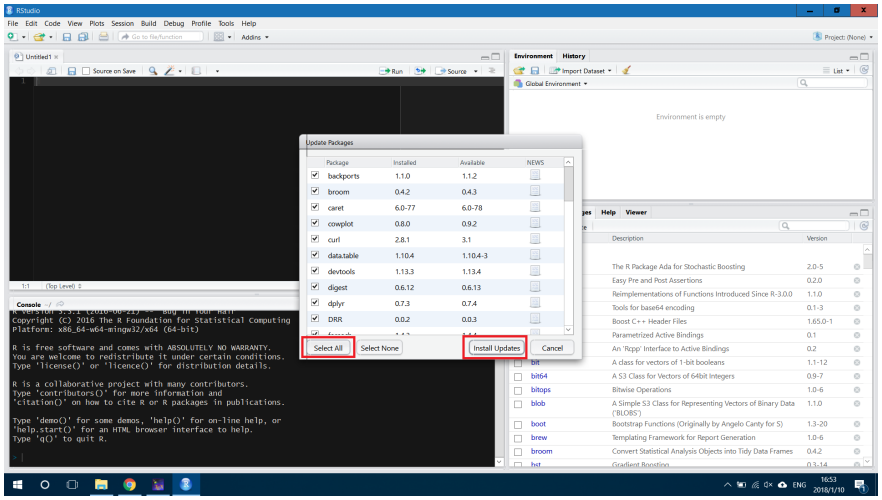

Figure 24

11.2 GVim

GVim offers a graphic user interface for the editor **Vim**. This is a powerful editor but could be a little bit hard to use.

- You can download Vim from their download page. For Windows system, click on "PC: MS-DOS and MS-Windows", and download "gvim80.exe".
- Open the installer and accept the default conditions.

11.3 For Linux

11.3.1 LaTeX

TeX Live is an easy way to get up and running with the TeX document production system, it is available on most Unix-like systems, but it is recommended to use MacTeX if you are using MacOSX. To install TeXLive and TeXstudio, run the following code:

```
sudo apt-get install texlive-full texworks texstudio
```

11.3.2 Git

To install Git, run the following code:

```
sudo apt-get install git
```

11.3.3 R

- Install from CRAN:

```
## This sets up the CRAN repository in your Linux Package Manager
sudo echo "deb http://cran.rstudio.com/bin/linux/ubuntu xenial/" |
sudo tee -a /etc/apt/sources.list
gpg --keyserver keyserver.ubuntu.com --recv-key E084DAB9
gpg -a --export E084DAB9 | sudo apt-key add -
sudo apt-get update
sudo apt-get install r-base r-base-dev
## extra linux packages needed by
sudo apt-get install r-cran-xml  pkg-config libxml2-dev
libtiff5-dev fftw3 fftw3-dev  tmux libav-tools
cifs-utils openssh-server openssh-client tree htop
gdebi curl libcurl4-openssl-dev libssl-dev
```

- Before installing, you should check the latest version of RStudio, and change the version number in the code below accordingly. Install RStudio:

```
## Update to the latest version number in the lines below
wget https://download1.rstudio.org/rstudio-1.1.383-amd64.deb
sudo gdebi -n rstudio-1.1.383-amd64.deb
rm rstudio-1.1.383-amd64.deb
```

11.4 For Mac

11.4.1 Homebrew

Homebrew is a package manager for Mac OS. To install Homebrew, paste and run the following command in terminal:

```
/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

You can read more about Homebrew here.

11.4.2 XQuartz

To correctly set up your linux environment, you should also install XQuartz. XQuartz is Apple Inc.'s version of the X server, a component of the X Window System for macOS. You can download and install the latest version of XQuartz.

11.4.3 LaTeX

To install LaTeX on Mac, you need to install MacTeX and TeXstudio.

- The current distribution as of today (September 4, 2018)is MacTeX-2017. This distribution requires Mac OS 10.10, Yosemite, or higher and runs on Intel processors. To download, click MacTeX Download.
- After downloading, double click on the MacTeX.pkg to install. Follow the straightforward instructions. Installation on a recent Macintosh takes four to six minutes.
- At the end of installation, the installer will report "Success." But sometimes, the installer puts up a dialog saying "Verifying..." and then the install hangs. In all cases known to us, rebooting the Macintosh fixes this problem. After the reboot, install again.
- Now you can start installing TeXstudio. You can find the corresponding installer on the TeXstudio website.
- Because the developers of TeXstudio do not have an Apple Developer Account, OS X may complain about an unidentified developer and deny opening TXS. In that case, open the context menu on the TXS icon (Ctrl + Click) and select open.

11.4.4 Git

There are several ways to install Git on a Mac. In fact, if you've installed XCode (or it's Command Line Tools), Git may already be installed. To find out, open a terminal and enter git –version.

Apple actually maintain and ship their own fork of Git, but it tends to lag behind mainstream Git by several major versions. You may want to install a newer version of Git using the method below:

- Download the latest Git for Mac installer.
- Follow the prompts to install Git.
- Open a terminal and verify the installation was successful by typing git –version.

- Configure your Git username and email using the following commands, replacing "yourusername" with your own. These details will be associated with any commits that you create:

```
$ git config --global user.name "yourusername"
$ git config --global user.email "youremail@website.com"
```

11.4.5 R

- Download R from CRAN and click "Download R for (Mac) OS X".
- Follow the instructions and install R.
- Download the latest RStudio from their website. Open the installer and follow the instructions.

References

[1] R. D. Peng, *R Programming for Data Science*. Leanpub, Feb. 2014.

[2] R. D. Peng, *Exploratory Data Analysis with R*. Leanpub, Apr. 2015.

[3] David M. Diez, Christopher D. Barr, and Mine Çetinkaya-Rundel, *OpenIntro Statistics: Third Edition*. S.l.: OpenIntro, Inc., 3 edition ed., July 2015.

[4] H. Wickham and G. Grolemund, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 1 edition ed., Jan. 2017.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics, New York: Springer, 1st ed. 2013, corr. 5th printing 2015 edition ed., Aug. 2013.

[6] Citrix, "Citrix receiver for xen VDIs and xen apps," 2014. 00000.

[7] C. Portal, "CWRU CSE portal for VDIs and XenApps," 2014.

[8] R, "R (programming language)," Aug. 2014. 00000 Page Version ID: 621330268.

[9] R. Project, "The r project for statistical computing," 2014.

[10] RStudio, "RStudio," 2014. 00000.

[11] R. Peng, "Computing for data analysis: Week 1 - YouTube," 2014.

[12] P. Torfs and C. Brauer, "A (very) short introduction to r," 2014.

[13] "Portal:free software," Sept. 2014. 00000 Page Version ID: 581465934.

[14] "Gvim online," 2014. 00000.

[15] "Neovim," 2014. 00000.

[16] "Git (software) - wikipedia, the free encyclopedia," 2014. 00000.

[17] "Git," 2014. 00027.

[18] "GitHub," 2014. 00004.

[19] "Bitbucket: Free source code hosting for git," 2014. 00000.

[20] S. Exchange, "Stack exchange," 2014.

[21] T. Python, "The python tutorial — python v2.7.8 documentation," 2014. 00000.

[22] Python, "Python.org," 2013.

[23] "The hitchhiker's guide to python! — the hitchhiker's guide to python," 2014. 00000.

[24] "kennethreitz/python-guide," 2014. 00000.

[25] NumPy, "NumPy — numpy," 2014.

[26] J. E. Guyer, D. Wheeler, and J. A. Warren, "FiPy: Partial differential equations with python," *Computing in Science & Engineering*, vol. 11, pp. 6–15, May 2009.

[27] SciPy, "SciPy.org — SciPy.org," 2014.

[28] PythonXY, "pythonxy - scientific-oriented python distribution based on qt and spyder - google project hosting," 2014. 00000.

[29] IPython, "IPython shell and notebook," 2014.

[30] Spyder, "Spyder is the scientific PYthon development environment," 2014. 00000.

[31] "TeX users group (TUG)," 2014. 00000.

[32] LaTeX, "LaTeX - wikibooks, open books for an open world," 2014. 00000.

[33] Zotero, "Zotero reference/citation manager, BibTeX client," 2014.

[34] R. H. French, "DSCI351-451: Exploratory data analysis for energy & manufacturing," 2016.

[35] C. Commons, "Creative commons - about the licenses," 2014.

[36] "Creative commons — attribution-ShareAlike 4.0 international — CC BY-SA 4.0," 2015.

[37] C. Commons, "Creative commons license," Aug. 2014. 00000 Page Version ID: 618703231.

[38] Gnu, "gnu.org," 2014. 00007.

[39] G. GPL, "GNU general public license," Aug. 2014. 00000 Page Version ID: 622300724.