

CWRU: DSCI353-353M-453 Syllabus
Data Science: Statistical Learning, Modeling and Prediction
Spring 2023 Tuesday, Thursday 11:30 am to 12:45 pm,
In-Person/Remote/Synchronous

Prof. Roger H. French, with Paul Leu (Pitt), Kris Davis, Mengjie Li (UCF), Sonya Cirlos (UTRGV), C

March 21, 2023

1 Joint Undergraduate and Graduate Course

1.1 DSCI353 and DSCI353M

DSCI353 is the 4th level class in the **Applied Data Science Undergraduate Minor**. The ADS Minor is available to CWRU students across all the schools of the University. For more information see

- Applied Data Science in the CWRU Bulletin
<https://bulletin.case.edu/schoolofengineering/materialsscienceengineering/#undergraduatetext>,
- And the CWRU Data Science Web Page
<https://case.edu/datascience/students/degree-programs/undergraduate-applied-data-science-minor>
page.

DSCI353 will introduce students to linear and beyond linear modeling, prediction and machine learning (including Kera/TensorFlow and Torch deep learning frameworks), the steps in a data analysis the following data cleaning, exploratory data analysis and introduction to linear modeling (the subject of DSCI351-451). This course will use an open data science tool chain consisting of R coding, Rmarkdown, Rstudio IDE and Git version control, and will be based in inferential statistical concepts, the stages of a data analysis and reproducible research.

DSCI353M section focuses specifically on Exploratory Data Science of Materials and Materials Systems.

1.2 DSCI453

DSCI453 is a graduate level introduction to data science modeling and prediction. Graduate students will, in addition to the coursework of DSCI353, develop a semester long data science project focused on a topic relevant to their graduate research area, for example time-series, spectral, or image data science problems.

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.

DSCI453 is one of the classes that satisfy the requirements for the CWRU **University Graduate Certificate in Applied Data Science**. More information on this transcriptable certificate is available.

- <https://case.edu/gradstudies/prospective-students/certificates>
- <https://engineering.case.edu/materials-science-and-engineering/graduate-certificate-in-applied-data-science>

1.3 DSCI353-353M and DSCI453 Prerequisites

For DSCI353/353m

1. ENGR131 Elementary Computer Programming or equivalent
2. STAT312R Basic Statistics for Engineering and Science or equivalent
3. DSCI351 Exploratory Data Analysis

For DSCI453,

1. DSCI451 Exploratory Data Analysis
2. Or comparable experience in data analysis with R or Python

2 Course Description

In this course, we will use an open data science tool chain to develop reproducible data analyses useful for inference and prediction, using modeling and machine learning, for the behavior of complex systems. In addition to the standard data cleaning, assembly and exploratory data analysis steps essential to all data analyses, we will identify statistically significant relationships from datasets derived from population samples, and infer the reliability of these findings. We will use regression methods to model a number of both real-world and lab-based systems producing predictive models applicable in comparable populations. We will assemble and explore real-world datasets, use pair-wise plots to explore correlations, perform clustering, self-similarity, and logistic regression develop both fixed-effect and mixed-effect predictive models. We will also introduce machine-learning approaches for regression and classification including Keras/TensorFlow for neural network and other modeling techniques. Results will be interpreted, visualized and discussed.

We will introduce the basic elements of data science and analytics using [R Project open source software](#). We will also use the Rstudio IDE (Integrated Development Environment), <https://www.rstudio.com/products/RStudio/>. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and assembly functions, and a rich selection of statistical packages, used for data analytics, model development, prediction, inference and clustering. We will also learn tidy principles for data analysis, pipes and ggplot vs base graphics approaches to data visualization.

For students with little prior R experience, we'll introduce resources to learn R data types, reading and writing data, looping, plotting and regular expressions. With this background, it becomes possible to start performing variable transformations for linear regression fitting and developing structural equation models, fixed-effects and mixed-effects models along with other statistical learning techniques, while exploring for statistically significant relationships.

Python version 3 (or Python3) is also commonly used for data science and data analyses, while at the same time Python is a general purpose programming language. Both R and Python are interpreted languages that do not require compiling the code prior to execution. Due to Python's broader spread of use cases (from data analysis to full applications to software engineering) it can be easier for a developing data scientist to find useful answers to questions, by learning R first. Python can then be learned as a second data analysis language. Students are welcome to use Python in the DSCI classes.

The class is taught using a "practicum" approach and will be structured to have a balance of theory and practice. We'll split class into Foundation and Practicum a) Foundation: lectures, presentations, discussion b) Practicum: coding, demonstrations and 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.

2.1 Outcomes

Capabilities

- Introduction to statistical and data science.

- Familiarity with R Statistics, scripting, functions, packages, automated data analysis.
- Familiarity with data assembly, exploratory data analysis and statistical modeling and learning.
- 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.

Predictive Modeling & Statistical and Machine Learning:

- Familiarity with inference and significance of sample results to populations
- Familiarity with regression and linear and non-linear statistical model building
Including training, testing and validating dataset strategies
- Applications of domain knowledge and statistical analytics
To identify important predictors and develop initial predictive models
- Familiarity with clustering, self-similarity methods
For categorization by different distance metrics
- Introduction to machine-learning approaches such as
 - Tree-based methods
 - Decision Trees
 - Random Forest
 - Support Vector Machines
 - Neural Networks such as
 - Multilayer Perceptron
 - Artificial Neural Networks
 - Convolutioanal Neural Networks
 - Recurrent Neural Networks
 - Deep Learning
 - Keras & TensorFlow2
 - Torch, rTorch and PyTorch

Data types include:

- Time-series,
- Spectral
- Images
- 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.

3 DSCI353-353M-453 Syllabus: Weekly Topics

DSCI353-353M-453 Syllabus: Data Science: Statistical Learning, Modeling, Prediction and Machine Learning

Day:Date	Foundation	Practicum	Readings(optional)	Due(optional)
w01a:Tu:1/17/23	Markov Cluster	R, Rstudio IDE, Git		(LE0)
w01b:Th:1/19/23	Stat. Learning, Approach	Bash, Git, Class Repo	ISLR1,2 (R4DS-1-3)	
w02a:Tu:1/24/23	Lin. Regr. Bias-Var.	SemProjs; Regr. Ovrw	ISLR3,(R4DS-4-6)	(LE0:Due) LE1
w02b:Th:1/26/23	Train/Test, Bias vs. Vari.	Tidyverse Review	DL01 DL02 (R4DS-7,8)	
w02Pr:Fr:1/27/23	ADD DROP	DEADLINE		453 Update 1
w03a:Tu:1/31/23	Logistic Regr. Classif	Pred. Analytics, Regr.	DL03,ISLR4	
w03b:Th:2/2/23	LDA/QDA	ggPlot2, Code Expect.	DL04, DL05	LE1:Due, LE2
w03Sa:2/4/23				LE1:Due
w04a:Tu:2/7/23	Resample Cross-Valid.	ggplot	ISLR5	
w04b:Th:2/9/23	DL, ML Overview	Multilevel Mod.	ISLR6 (R4DS9-16)	
w04Pr:Fr:2/10/23				453 Update 2
w05a:Tu:2/14/23	Resampling: Bootstrap	Bootstrap Mixed Effects	DL2R1, DL06,07	LE2:Due, LE3
w05b:Th:2/16/23	Subset Selec., Shrink.	Dim. Red. PCA	DLwR2	
w05Pr:Fr:2/17/23				453 Rep. Out 1
w06a:Tu:2/21/23	ML with NNs	ggplot, clustering	DLwR3	
w06b:Th:2/23/23	Beyond Linear Modls	Feature Select., Caret	ISLR7 (R4DS22-25)	LE3:Due, LE4
w06Pr:Fr:2/24/23				453 Update 3
w07a:Tu:2/28/23	Dec. Trees, Rand. Forest	Tidy Modeling	ISLR8, DL08,09	
w07b:Th:3/2/23	MidTerm Review, SVM	SVM, SVR, ROC	ISLR9 (R4DS26-30)	Peer Review 1
w08a:Tu:3/7/23	ML Overview	, Keras/TF2, Torch	ISLR10.1,10.2	
w08b:Th:3/9/23	MIDTERM EXAM		DL10,11	LE4:Due LE5
w08Pr:Fr:3/10/23				453 Update 4
Tu:3/14/23	SPRING	BREAK	ISLR10.3,10.4	
Th:3/16/23	SPRING	BREAK	ISLR10.5,10.6,	
w09a:Tu:3/21/23	Deep Learning	TF2 Keras Intro		ISLR10.7,10.8, DLwR3
w09b:Th:3/23/23	Computer Vision, CNN	CNN w/TF2, Overfit	DLwR4, DL12,13	
w09Pr:Fr:3/24/23				453 Rep. Out 2
w10a:Tu:3/28/23	Deep Learn Intro	NN Types	DLwR5 Hinton ImageNet	
w10b:Th:3/30/23	DL CNN,RNN ImageNet	NN Types, CNN wTF2		
w10Pr:Fr:3/31/23				453 Upd.5 & PrRev 2
Sa:4/1/23				LE5:Due LE6
w11a:Tu:4/4/23	Fitting NNs	AUC,Prec,Recall Fruit		
w11b:Th:4/6/23	NLP, Graphs & ML		LeCun DL Rev. 2015	
w12a:Tu:4/11/23	Graphs & ML	NLP with sequences	DLwR6	
w12b:Th:4/13/23	NLP w attention	Graph Repr Proc Wrk-flw		LE6:Due LE7
w13a:Tu:4/18/23	DL Frameworks	Explaining DL w Lime		
w13b:Th:4/20/23	Linux Distros XGBoost	Explain Preds	Deep Dream	
w13Pr:Fr:4/21/23				453 Rep. Out 3 Due
w14a:Tu:4/25/23	Tranformers			
w14b:Th:4/27/23	Final Exam Review	Torch NN & DeepLearn		LE7:Due
w14Pr:Fr:4/28/23				Peer Rev 3 Due
	FINAL EXAM	Th. 5/4/23, 12-3pm	Nord 356 & Zoom	
	453 Final PDF Report	Fr. 4/29, 11:59pm		

Table 1: DSCI353-353M-453 Weekly Syllabus. R4DS-x.y, OISx.y, ISLRx.y, DLwRx.y, DLGBx.y refers to chapters and sections assigned as reading in our textbooks. DLx are deep learning articles.
March 21, 2023

4 Homework, Project, Report-Presentation Grading

These classes are graded on a curve, not on a fixed point system.

DSCI353-353M is graded on 100 points basis

Seven LabExercises, worth 9 points each = 63pts.
353 – 353M – 453SemProjGradingFeedback, worth 2 points each = 6pts.
Midterm Exam = 10pts.
Final Exam = 21pts.
Total = 100pts

DSCI453 is graded on a 140 point basis

Seven LabExercises, worth 9 points each = 63pts.
353 – 353M – 453SemProjGradingFeedback, worth 2 points each = 6pts.
Midterm Exam = 10pts.
Final Exam = 21pts.
Five SemProj Updates worth (2pt/each) = 10 pts.
Three SemProj presentations(5pts/each) = 15 pts.
One SemProj Final Report = 15 pts.
Total = 140pts

5 Textbooks and Readings

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

OISv4 is an open source text book, published under a creative commons license, for free distribution as a pdf. In addition a copy can be purchased from Amazon for 20 dollars. OISv4 is the main textbook for DSCI351-351m-451, and is covered in that class. It is provided here for reference.

R4DS can be purchased as ebooks (pdf, epub, mobi formats) from [O'Reilly Media](#). It is also available as a web-readable book at [R For Data Science](#) and as a [Bookdown Code Repo on Bitbucket](#). They can also be purchased as physical books.

5.1 Background Data Science books from DSCI351

Peng R Programming (PRP) (Figure 1) and Peng Exploratory Data Analysis (EDA) (Figure 2) are introductory books to R and Data Science and Analysis. These are Leanpub books, available from LeanPub for a "pay what you want" price.

OpenIntro Statistics (Figure 3), is the open source Inferential Statistics Textbook that we use in DSCI 351/351m/451.

Also R for Data Science (Figure 4) teaches using R and the Tidyverse Functions for efficient and streamlined data analysis code. So this is a very important book to be familiar with.

R Programming for Data Science



Roger D. Peng

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

Exploratory Data Analysis with R



Roger D. Peng

Figure 2: **EDAwR:**
Roger Peng, **Ex-
ploratory Data
Analysis With R.**
2015 [\[2\]](#)

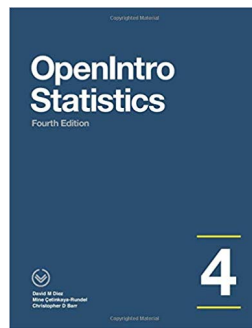


Figure 3: **OIS:** David
M. Diez, Christopher
D. Barr, and Mine
Cetinkaya-Rundel,
**OpenIntro Statis-
tics 4th Ed.** 2015
[\[3\]](#)

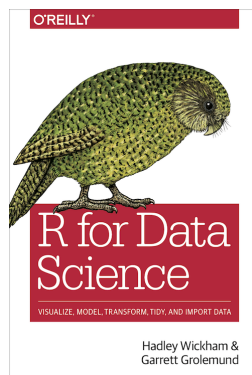


Figure 4: **R4DS:**
Garrett Grolemund,
Hadley Wickham, **R
for Data Science.**
2017 [\[4\]](#)

5.2 DSCI353-353M-453 Textbooks

The main textbooks for this semester are the following

1. **Introduction to Statistical Learning with R, 2nd Edition 2021** (ISLR, Figure 5) is a Springer book which is available for free as a pdf.
2. **R for Data Science** (R4DS, Figure 4), which teaches Tidyverse R coding, is used in both DSCI351/351m/451 and DSCI353/353m/453 courses.
3. **Deep Learning with R** (DLwR, Figure 7) is used in DSCI353-353m-453 for Deep Learning and TensorFlow.
4. **Deep Learning** by Goodfellow, Bengio, 2016 for training deep networks.

We will also use a second deep learning framework, Torch from R, and PyTorch for use from Python.

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

Note that ISLR 2nd Ed. is the introductory book, for which Elements of Statistical Learning (ESL, Figure 6) is the advanced book. ESL is widely considered the bible of Machine Learning, and is also in your repo in the 3-readings/1-textbooks folder.

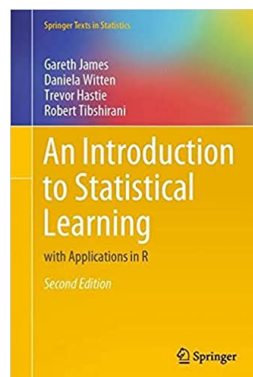


Figure 5: **ISLR:**
Gareth James,
Daniela Witten,
Trevor Hastie,
Robert Tibshirani
An Introduction to Statistical Learning: with Applications in R, Second Edition, 2021 [5]

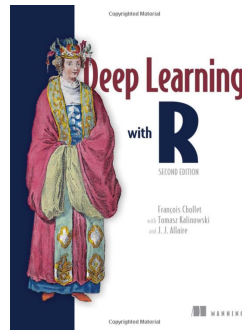


Figure 6: **DLwR**
2nd Ed.: François
Chollet, Tomasz
Kalinowski, and J.
J. Allaire, **Deep
Learning with R,**
2nd Ed.. 2022 [6]

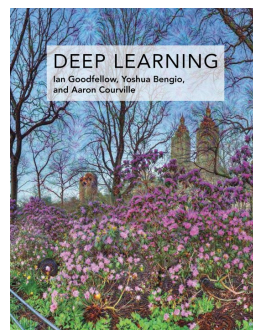


Figure 7: **DLGB:**
Ian Goodfellow,
Yoshua Bengio, and
Aaron Courville
Deep Learning,
2016 [7]

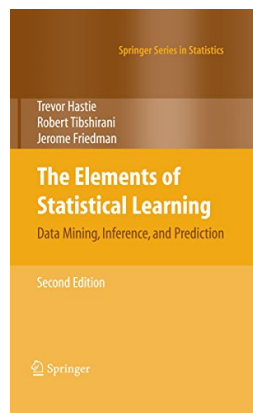


Figure 8: **ESL:**
Trevor Hastie,
Robert Tibshirani,
Jerome Friedman
**The Elements of
Statistical Learning: Data Mining,
Inference, and
Prediction, Second Edition,** 2009
[8]

6 Contact Information

Prof. Roger H. French

- Slack is best: CWRU-DSCI Slack, DSCI 353 Class Channel, <https://cwr-dsci.slack.com/>
- Email: rxfl31@case.edu,
Use DSCI353-353M-453 in the subject line
- @frenchrh on twitter
- White 536, and electronically.
- Office Phone 216 368 3655, Cell Phone 302 468 6667

TA: Kristen Hernandez

- DSCI Slack channel is best
- Email: kjh125@case.edu
Use DSCI353-353M-453 in the subject line
- White 615, and electronically.
- Cell Phone:

TA: Will Oltjen

- DSCI Slack channel is best
- Email: wco3@case.edu
Use DSCI353-353M-453 in the subject line
- White 614, and electronically.
- Cell Phone:

7 Course Mechanics

7.1 Lectures

Spring 2022: Tuesday, Thursday 11:30 am to 12:45 pm, In Nord 356 and On Zoom

7.2 DSCI Class Slack Channel

There is a Slack channel for this class

To join go to <https://cwr-dsci.slack.com> and sign up for an account, using your case.edu email address.

We use the Slack channel to share information, have discussions about topics from class, homework, etc.

If you have questions on homework, post them to Slack, and read other people's questions and answers, and answer questions you know how to.

7.3 Office Hours and Consultations

- Monday Office Hours: Mondays at 4pm on Zoom
- Afternoon Office Hours: Wednesday at 4 PM on Zoom
- Additional Consultations: After class or as needed.
Contact Prof. French, Kristen Hernandez, Will Oltjen
By Slack, email or in person.

7.4 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 accommodate 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 Course Repo

- 1-Assignments is where you will find the Lab Exercises and Exams
- 2-Class contains daily class notes as *.Rmd and *.pdf files
These are split into a Foundations "f" and a Practicum "p" class notes
- 3-Readings folder contains textbooks and readings for the course
- 4-Syllabus contains the updated Course Syllabus
- 5-SemProj contains information on the DSCI453 Semester Projects
The syllabus is updated throughout the course
The current syllabus is in 4-Syllabus folder

Your Working Data Analysis Folders

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

7.5 License applied to course materials 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 "2201-DSCI-Acknowledgement-of-IP.pdf" form.

7.6 Lab Exercise Assignments

All homework and Lab Exercise assignments are submitted electronically through canvas, uploading to the HW assignment page.

Filenames should contain DSCI353-353M-453, NAME, LE#... e.g. DSCI353-453FrenchLE4.Rmd

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

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

7.7 Data-science Semester Project Report

- Title
- Author
- Author Affiliation
- License: ideally CC-BY-SA 4.0 (but a license choice is yours)
- Abstract
- IntroductionModeling
- Data Science Methods
- Exploratory Data Analysis
- Statistical Learning: Modeling & Prediction(if appropriate)
- Discussion
- Conclusions
- Acknowledgements
- References, Citations

7.7.1 Abstract

Summary of the nature, finding and meaning of your data analysis project.

7.7.2 Introduction

Background and motivation of the Data Science question.

7.7.3 Data Science Methods

To be applied (such as image processing, time-series analysis, spectral analysis etc.

7.7.4 Exploratory Data Analysis

Results and steps in the data analysis.

7.7.5 Statistical Learning: Modeling & Prediction

If you analysis can accomplish some modeling, include it here.

7.7.6 Discussion

Discussion of the answers to the data science questions framed in the introduction.

7.7.7 Conclusions

7.7.8 Acknowledgments

7.7.9 References

7.7.10 How to make your report

The report is done as an Markdown document, which can be run/compiled to produce two versions of the report as a pdf. One shows your R code and figures, and the other doesn't show R code, just your figures.

You'll then turn in a zip file (and leave a copy in your repo), with the dataset (if its not to huge, if it is large, can you make a smaller dataset), Rmd file that works, and the two pdf reports. Just choose to do a pdf report, instead of a set of presentation slides.

The license choice of CC-BY-SA 4.0 is suggested so that others can use and build on your codes, in an open-source manner. With more restrictive licenses, others won't be able to use your code in the future.

8 Coding and Data Science Tools and Resources

Open Data Science (ODS) Desktops

You will not need to install software on your personal computers.

Instead you can install the Citrix Receiver [9] and then login to the CWRU CSE Portal. [10].

The CSE Portal is located at <https://cseportal.cwr.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 homework and projects. Its generally useful language for statistical analysis and data science.

- [The R Project for Statistical Computing](#) [11] main website
- [R programming language](#) R is a free software programming language and software environment for statistical computing and graphics.[12]
- [Rstudio](#) provides popular open source and enterprise-ready professional software for the R statistical computing environment. [13]
- [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, homework, 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 slide deck done from Rmarkdown and shared publicly on RPubS](#).

R Statistics, more resources

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

- [The R Project for Statistical Computing](#) [11] 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. [14]
- [A \(very\) short introduction to R](#) [15]
- [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. [16]
- [vim \(or Gvim the GUI version\)](#) is a powerful text and code editor, that is universally available on all Linux and mac computers.[17] [NeoVim](#) is a new Gvim fork.[18] It can be installed on windows computers, its available on the ODS VDIs.. [17]
- [Git \(Wikipedia\)](#) is a distributed content versioning system that is very popular. It enables collaborative code development and LaTeX writing projects.[19]
- [Git server software](#) is installed on each computer.[20]
- [GitHub](#) is a Git server website used for collaborative code development.[21]
- [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.[22]
- [Stack Exchange](#) [23] 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.](#) [24]
- [The Python main website.](#) [25]
- [The Python Tutorial — Python v2.7.8 documentation](#) [24]
- [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>. [26] [27]
- [NumPy is the fundamental package](#) [28] for scientific computing with Python.
- [FiPy: Partial Differential Equations with Python](#) [29]
- [SciPy is a python-based ecosystem](#) [30] of open-source software for mathematics, science, and engineering.
- [PythonXY - Scientific-oriented Python Distribution](#) based on Qt and Spyder that runs on Windows. [31]
- [IPython Shell and Notebook](#) [32]
- [Spyder is the Scientific PYthon Development Environment](#) [33]

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[34].
- [LaTeX - Wikibooks](#), open books for an open world[35].
- [Zotero Reference-Citation Manager, BibTeX Client](#) [36].

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

<http://bulletin.case.edu/undergraduatestudies/academicintegrity/>.

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 DSCI class [37] is covered by a [Creative Commons](#) [38] 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. [39]

More information on licensing open works, can be found on Wikipedia. [40]

[GNU](#) [41] is the developer of the [GPL License](#) [42] that is used for many open source software projects, such as Linux.

10.2 Software Licenses

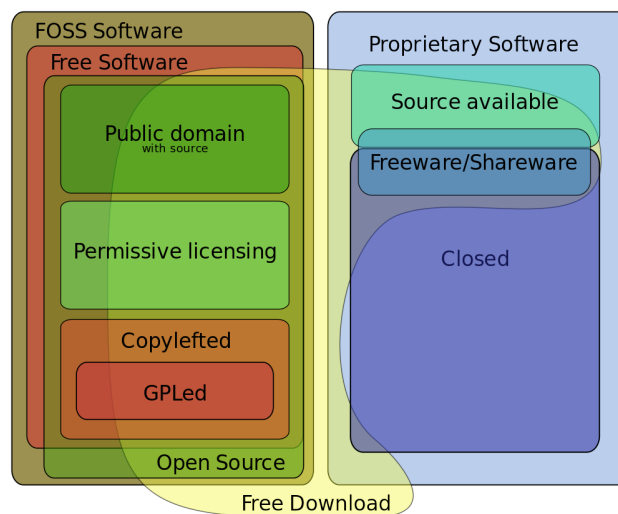


Figure 9: Diagram of software under various licenses according to the FSF and their The Free Software Definition: on the left side "free software", on the right side "proprietary software". On both sides, and therefore mostly orthogonal, "free download" (Freeware). CC0, <https://commons.wikimedia.org/w/index.php?curid=46544815>

Good discussion of software licenses is available at [this Wikipedia article](#).

And good comparison of different open-source software licenses is in [this Wikipedia article](#).

Typically the [Apache License of the Apache Software Foundation](#) or [Python Software Foundation](#) license are good choices for a "permissive" license.

But the Gnu General Public License [GPLv2](#) and [GPLv3](#) are "stronger" free and open source software licenses. This is the original free software license written by [Richard Stallman](#) of the [Free Software Foundation](#) for the [GNU Project](#).



Figure 10: GNU mascot, by Aurelio A. Heckert

10.3 Scoring Rubric for Oral Presentation

Presenter Name:

Date:

Evaluator Name:

10.3.1 Scientific/Technical Content (25 points), And Data Science Content (25 points)

Introduction:

- Defines background and importance of research.
- States data analytics objective
- Able to identify relevant questions.

Body:

- Presenter has a data analytic goal (EDA, Modeling, Classification).
- Addresses audience at an appropriate level (rigorous, but generally understandable to a scientifically-minded group).
- Offers evidence of methods tried, what worked.
- Describes methodology and implementation.
- The talk is logical well organized.

Conclusion:

- Summarizes major points of talk.
- Summarizes potential weaknesses (if any) in findings.
- Provides you with a “take-home” message.

10.3.2 Coding Elements & Style, .R, .Rmd (25 points)

- Code author, license, versioning.
- Code styling, indenting, commenting.
- Is it reproducible code, and well structured data.
- Cross-platform, cross-computer code: relative paths, or absolute paths
- Making and using functions.
- Use of interesting packages

10.3.3 Presentation Quality, Clarity, Style (25 pts)

- Graphs/figures are clear and understandable.
- The text is readable and clear.
- Audio/Visual components support the main points of the talk.
- Appropriate referencing of data that is/was not generated by presenter

10.3.4 General Comments

Final Score: / 100

11 Setting up your R data science computer

To setup and install the software for an R Open Data Science workstation, here's how.

Note that the “R standard packages are listed under the “For Windows” section in subsection [11.1.4](#).

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.

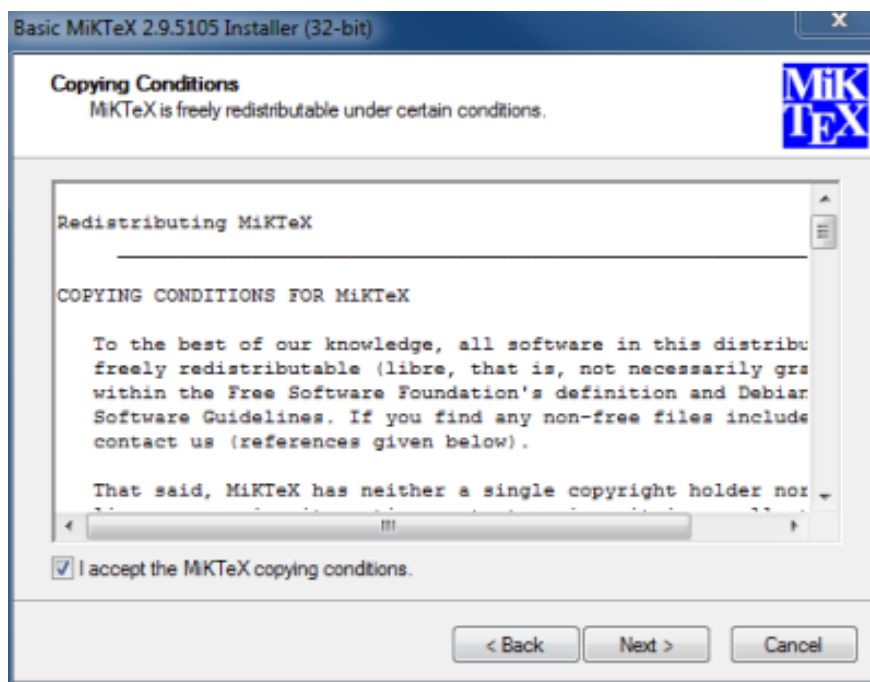


Figure 11

- 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.

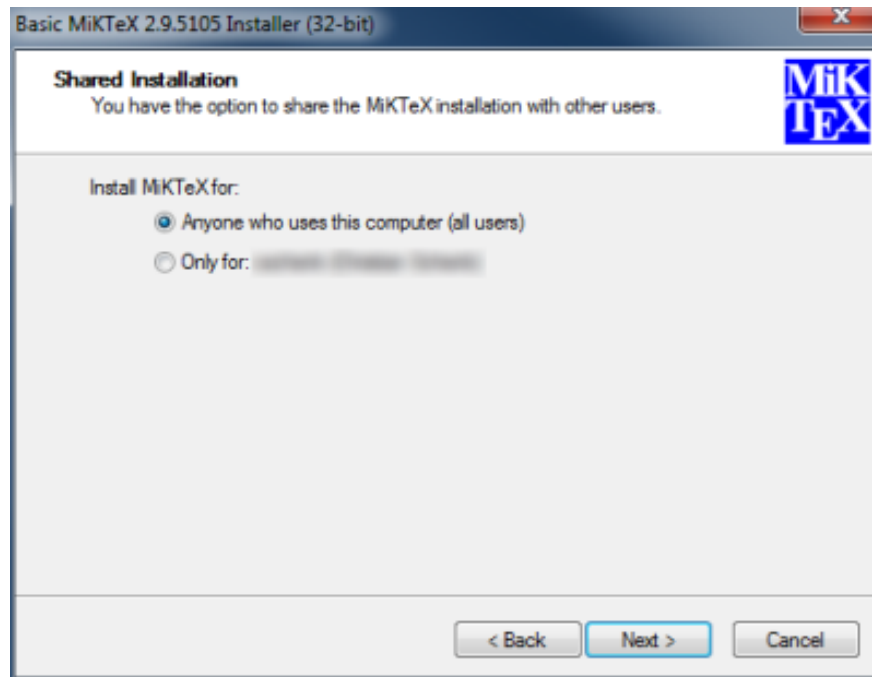


Figure 12

- 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.

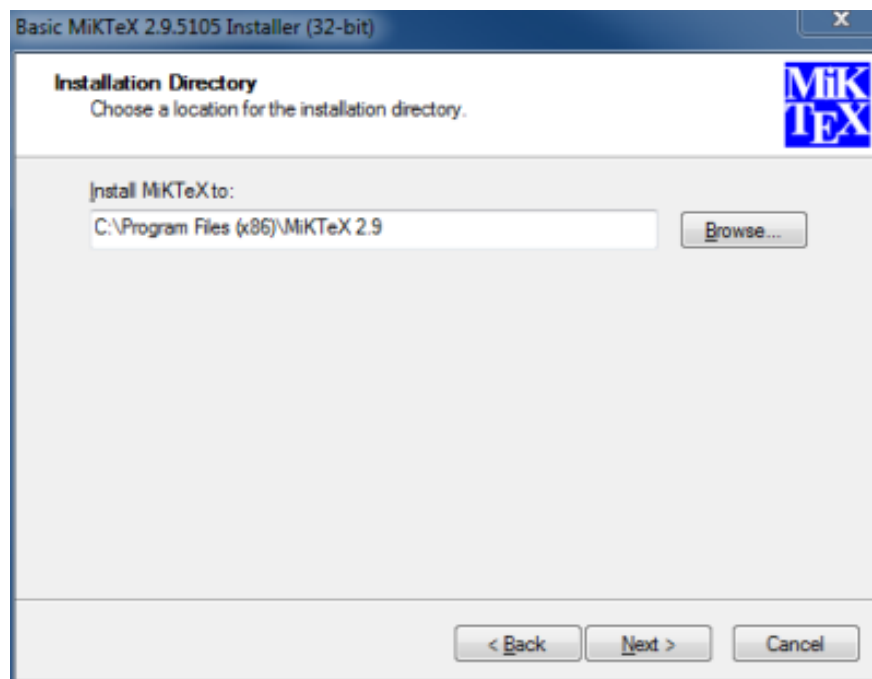


Figure 13

- 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.

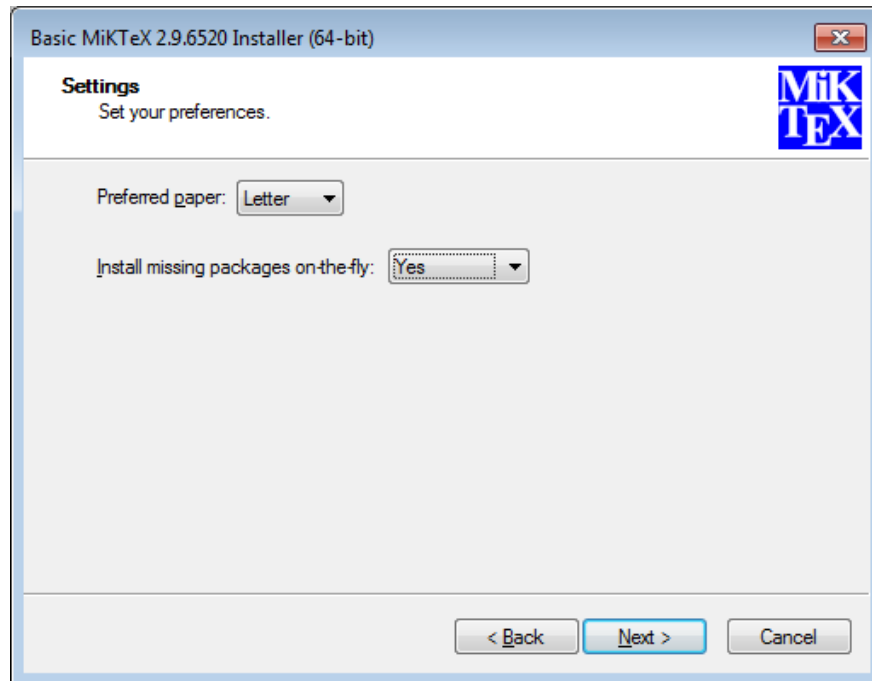


Figure 14

- 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 15

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

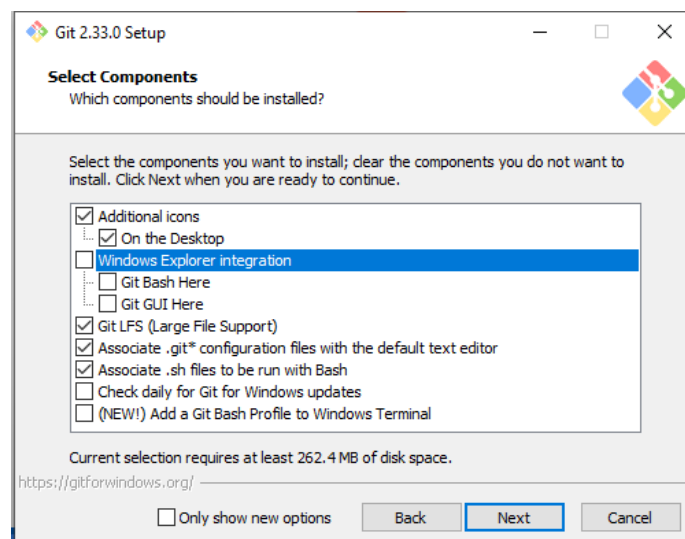


Figure 16

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

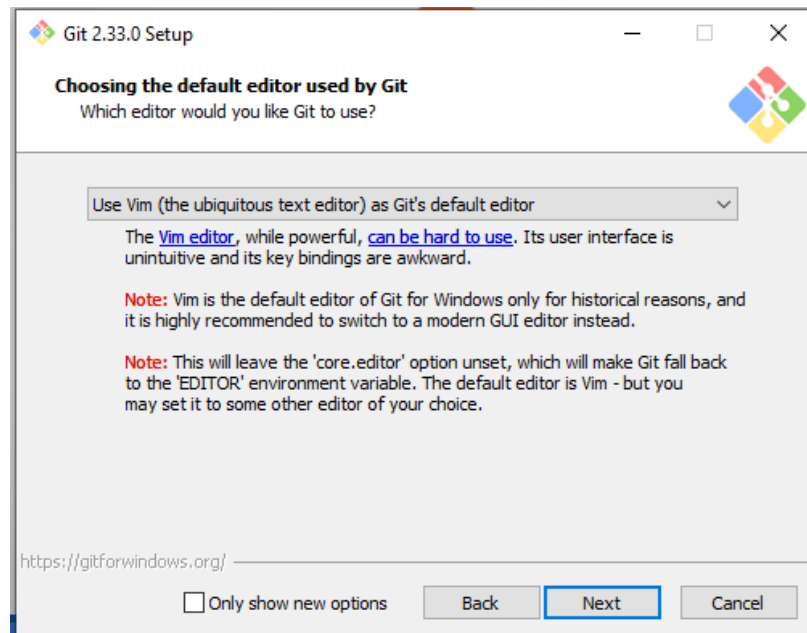


Figure 17

- Choose default initial branch name.

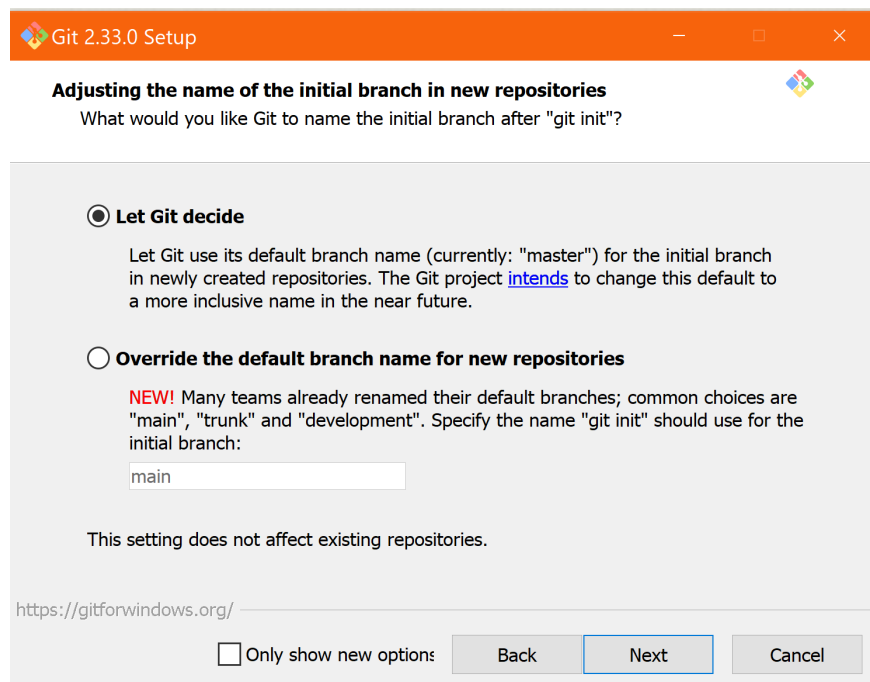


Figure 18

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

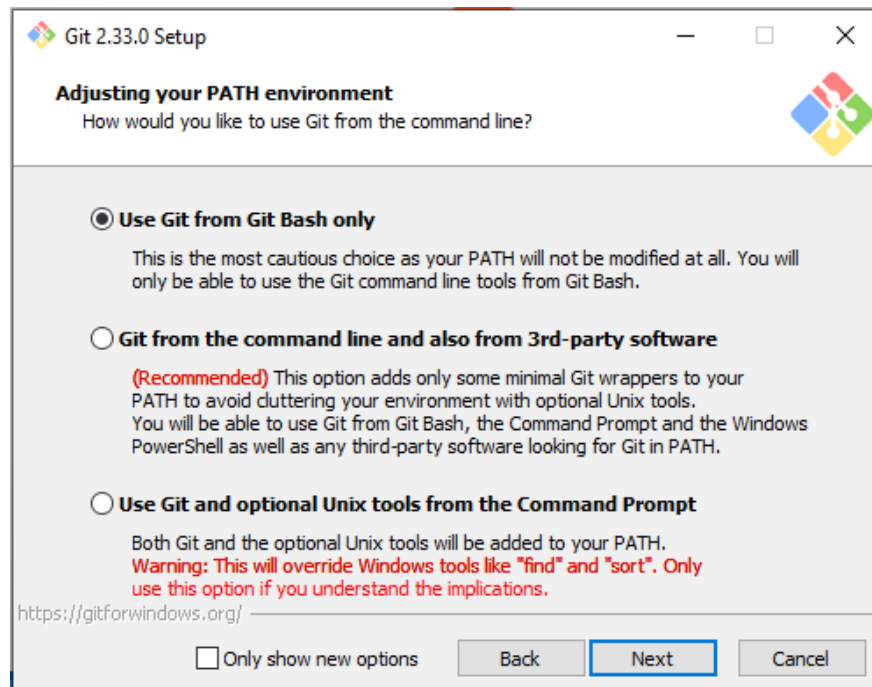


Figure 19

- Select which SSH client program to use.

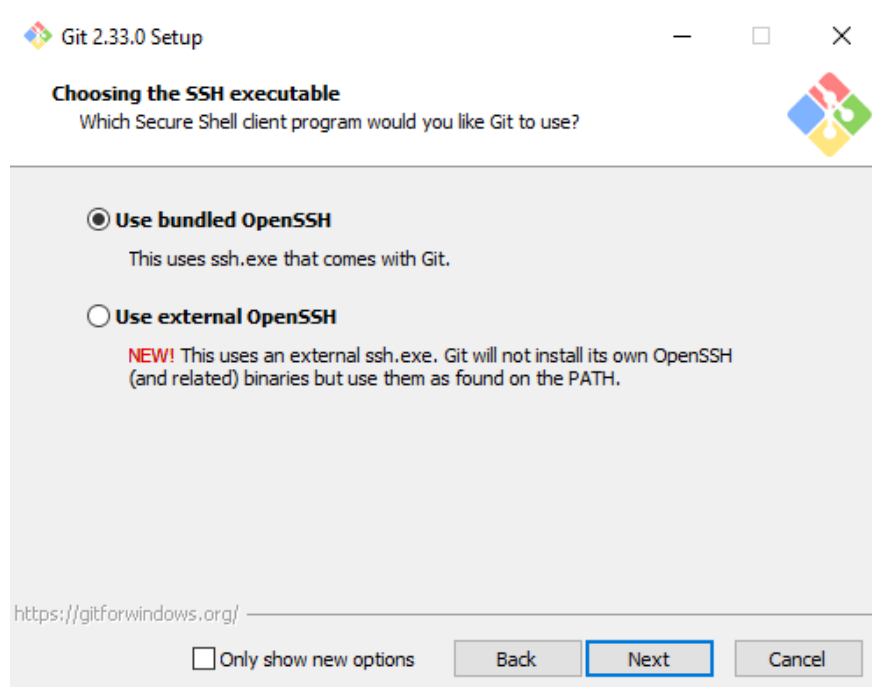


Figure 20

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

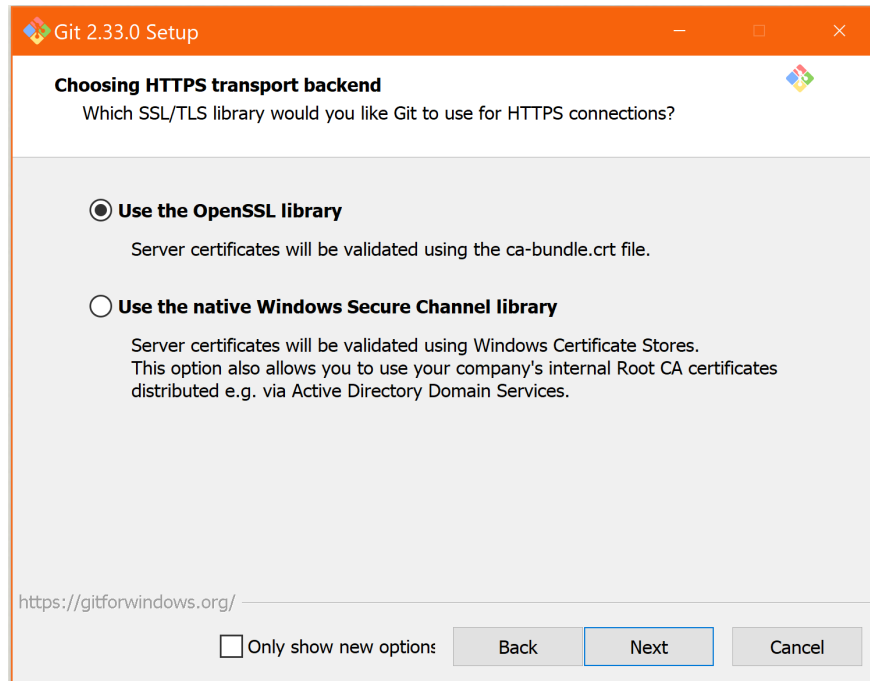


Figure 21

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

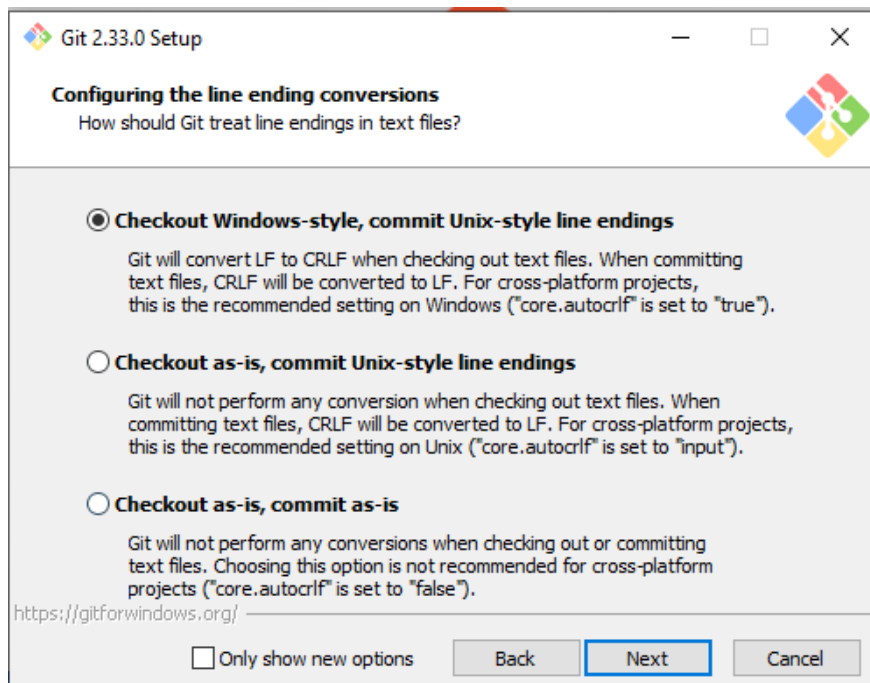


Figure 22

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

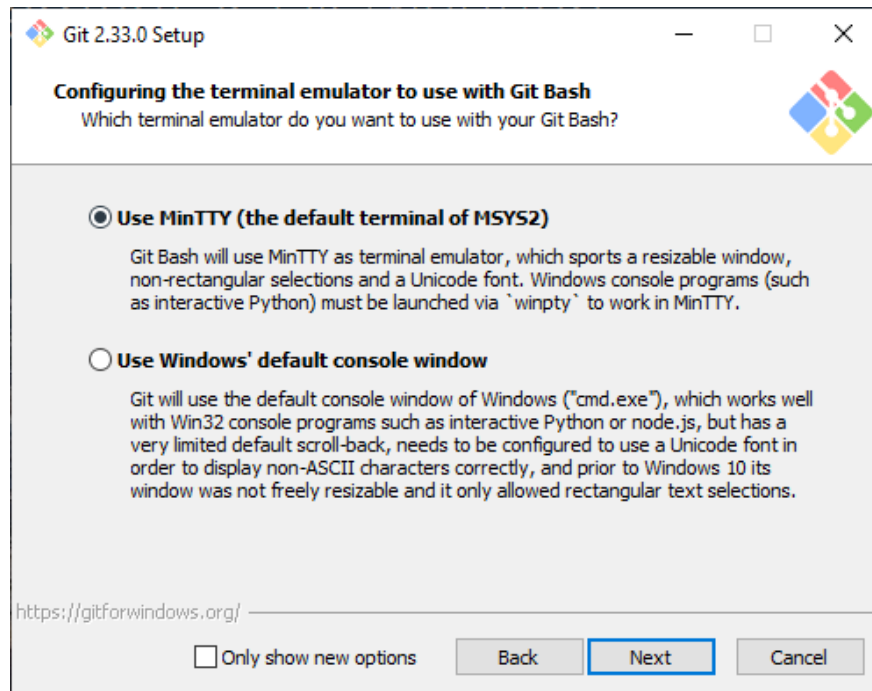


Figure 23

- Select the default behavior of 'git pull'

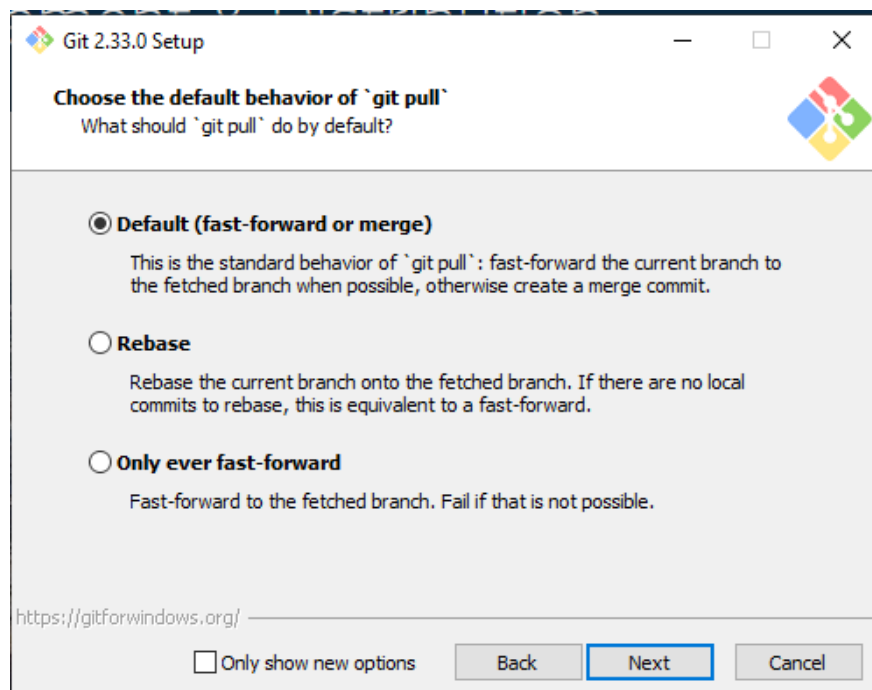


Figure 24

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

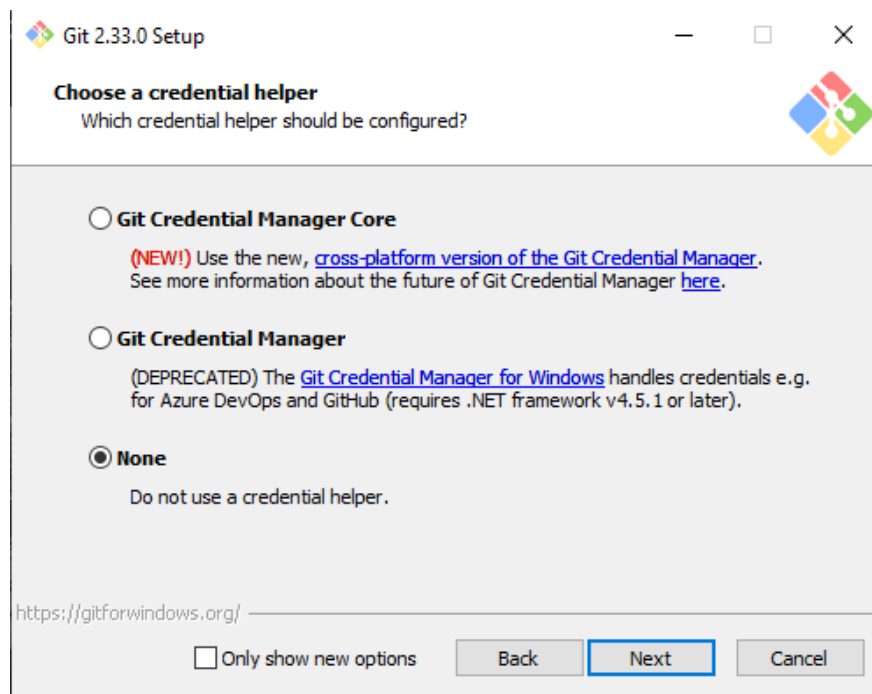


Figure 25

- Unselect the experimental options.

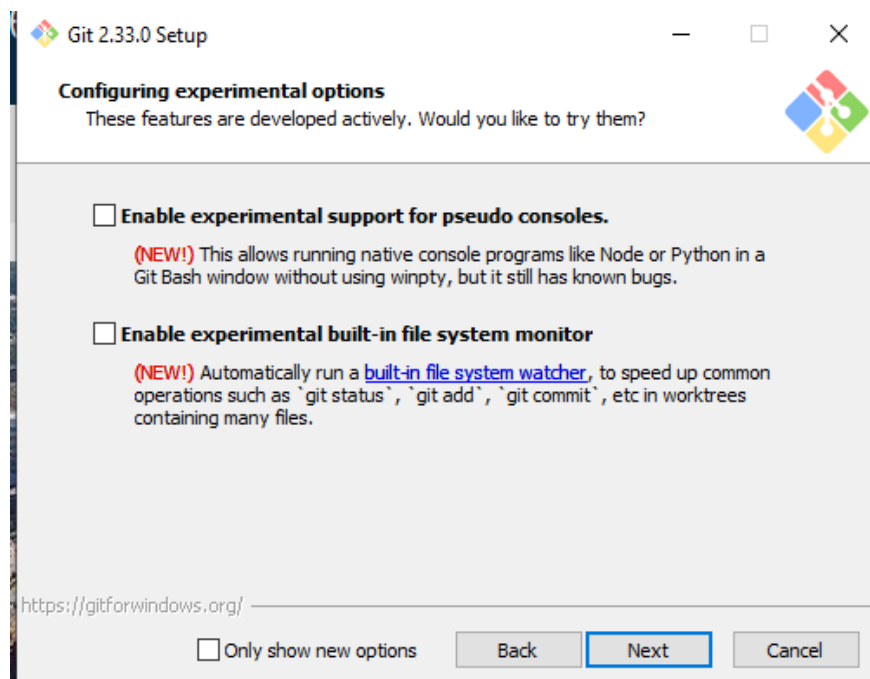


Figure 26

- 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 be ready to use the Linux command on a windows machine. Double click on below icon to start the Git Bash.

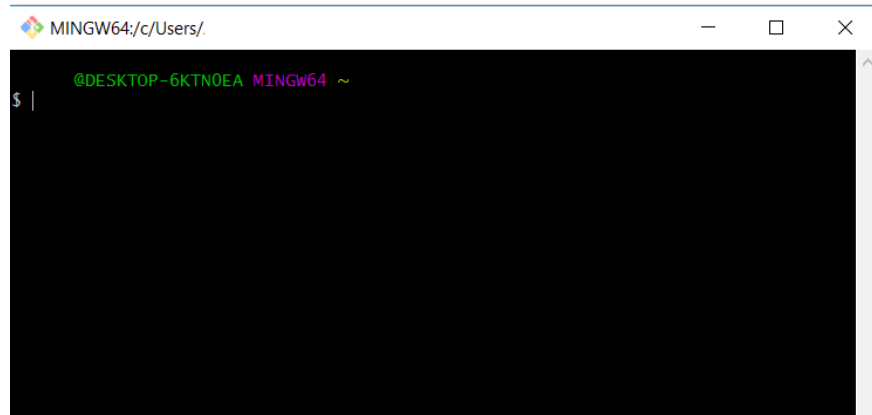


Figure 27

- 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:

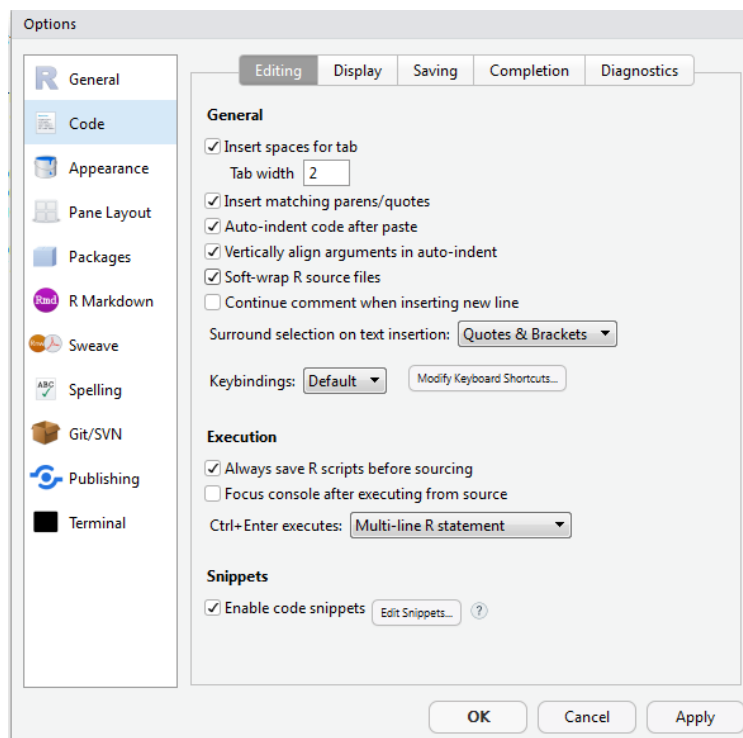


Figure 28

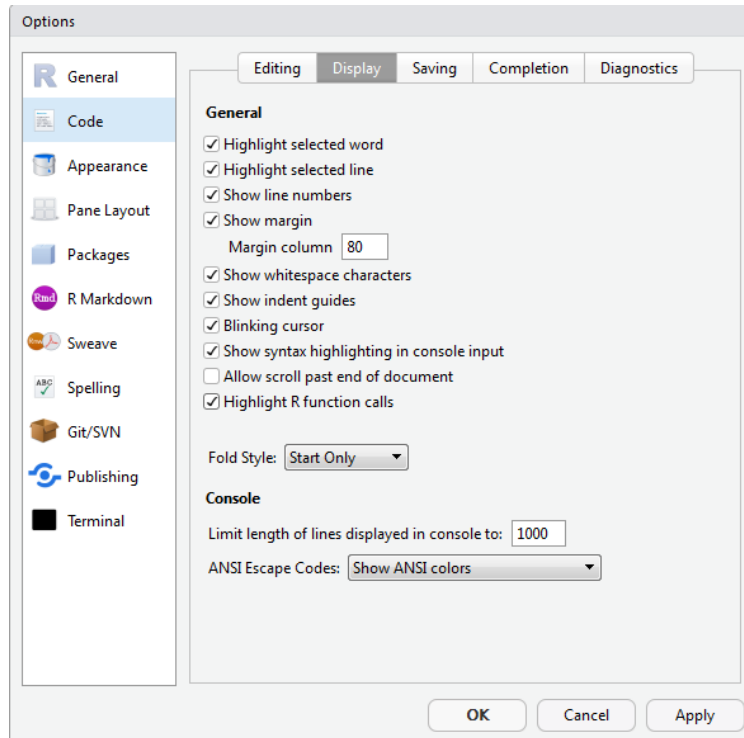


Figure 29

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

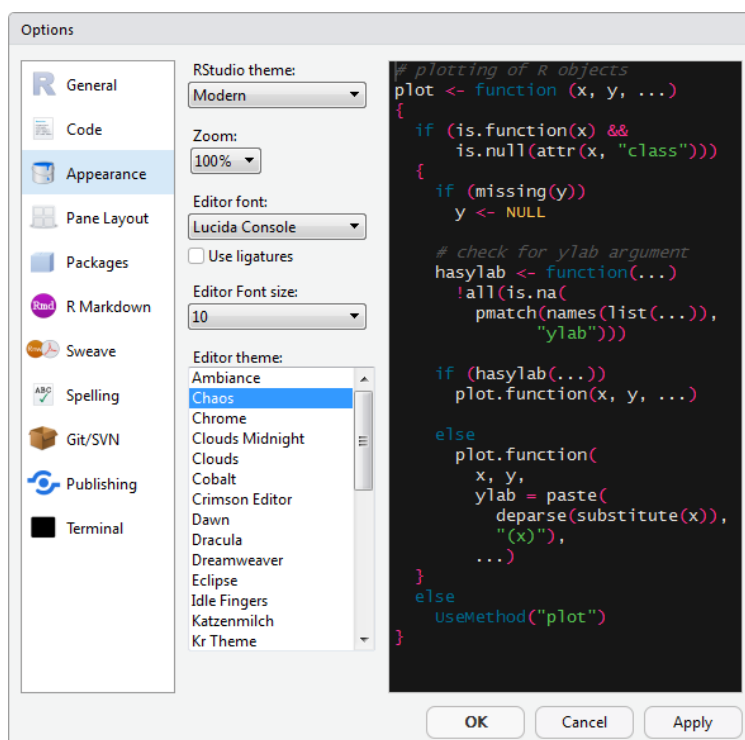


Figure 30

11.1.4 R Standard Packages for Win,Lin,Mac

Here are the standard packages you should install.

- 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:
 - netSEM
 - acepack
 - adabag
 - akima
 - Amelia
 - AmesHousing
 - animation
 - anytime
 - apache.sedona
 - arrow
 - arsenal
 - arules
 - askpass

astsa
available
babynames
backports
baseline
BayesFactor
bayesSurv
bcpa
bda
BiocManager
birk
bit64
blavaan
blogdown
BMA
bookdown
bookdownplus
BoomSpikeSlab
boot
bootstrap
breakDown
breakpoint
brms
broom
bsts
C50
Cairo
callr
calendR
car
carData
caret
cartogram
caretEnsemble
CausalImpact
cdlTools
centiserve
cgwtools
changeoint
changeoint.np
ChemoSpec

circulize
Ckmeans.1d.dp
class
ClimClass
cloudml
cluster.datasets
coda
CodeDepends
colormap
CORElearn
corrgram
corrplot
cowplot
cowsay
crayon
cpca
cranlogs
ctv
CVXR
datapasta
dataRetrieval
data.table
data.tree
dataMaid
DBI
DBItest
dbplyr
dbscan
ddiv
DescTools
devtools
DiagrammeR
DiagrammeRsvg
dice
dichromat
digest
directlabels
dlstats
doFuture
doParallel
dplyr

drake
drumr
lidR
DT
dtwclust
e1071
easyNCDF
Ecdat
eemR
effects
elevatr
epiDisplay
equationomatic
exifr
extrafont
fable
fabletools
factoextra
FAIRmaterials
fastcluster
fastDummies
feasts
feather
flexclust
flextable
flipbookr
forecast
foreach
formattable
fpp2
fpp3
fmsb
fun
fuzzyjoin
gam
gapminder
gbm
gclus
geojson
geojsonio
geoR

GGally
gganimate
ggalluvial
ggbump
ggcorrplot
ggdag
ggdark
ggdendro
ggdist
ggeffects
ggExtra
ggenealogy
ggforce
ggfortify
gghighlight
ggplot2movies
ggsci
ggsoccer
ggstatsplot
ggiraph
gglasso
ggm
ggmap
ggnetwork
ggnewscale
ggplot2
ggpubr
ggQC
ggRandomForests
ggraph
ggraph
ggrepel
ggridges
ggrepel
ggseas
ggsignif
ggspectra
ggstance
ggstream
ggtern
ggthemes

ggTimeSeries
ggvis
ggvenn
ggvoronoi
ggdark
ggalt
ggraph
ggpmisc
ggnetwork
ggTimeSeries
ggstance
ggbeeswarm
ggridges
glmnet
glmnetUtils
gmodels
googlesheets4
googleVis
GPArotation
graphlayouts
gridBase
gridExtra
gridGraphics
gsl
gstat
gt
gtrendsR
gWidgets2
hdf5r
hdpca
heatmaply
here
hexbin
highcharter
HH
Hmisc
hrbrthemes
httr
htmlwidgets
htmltools
huxtable

hyperSpec
igraph
igraphdata
igraphinshiny
imager
infer
ipred
IQCC
IRdisplay
IRkernel
ISLR
ISLR2
itertools
janitor
jsonld
jsonlite
jsonvalidate
jtools
kableExtra
keras
kerasR
kernlab
keyring
kgc
klaR
knitcitations
knitr
Lahman
lars
latex2exp
lavaan
lavaan.survey
leaps
learningr
learnr
learnrbook
learNN
lime
lintr
lme4
lmerTest

lobstr
logitnorm
magick
magickGUI
magrittr
mapdata
mapproj
maps
markdown
maptools
mapSpain
mapview
MASS
Matrix
MatrixModels
matrixStats
markovchain
mcmc
MCMCglmm
meltr
metRology
meme
Metrics
mgcv
mice
minpack.lm
mixtools
mlbench
MLmetrics
mlr
mnormt
moments
mosaicData
MTS
multiway
multicolor
naniar
NbClust
ncdf4
networkD3
neuralnet

NeuralNetTools
nFactors
NLP
NMF
nnet
nycflights13
odbc
olsrr
onehot
OneR
openair
openintro
OpenMx
OpenStreetMap
operator.tools
optimx
ordinal
osmdata
OSMscale
oysteR
ozmaps
packageRank
packHV
packrat
pacman
palmerpenguins
packcircles
pastecs
patchwork
pca3d
PCAmixdata
performance
PerformanceAnalytics
pins
pipeR
pixiedust
paletteer
plot3D
plotmo
plotKML
plotly

plotrix
plotROC
pls
plumber
plyr
png
pool
pracma
prodlim
pROC
processx
prophet
profvis
propagate
proxy
pryr
psych
purrr
PVplr
pwr
qcc
qqplotr
qtlmt
quantmod
r2d3
ragg
randomcoloR
randomForest
randomForestExplainer
randomForestSRC
randomNames
ranger
raster
rasterVis
rayshader
RColorBrewer
rcompanion
Rcpp
RCurl
Rdice
Rdpack

readr
recipes
RefManageR
regclass
relaimpo
remotes
repr
reshape
reshape2
reactable
reticulate
revealjs
Rfast
rgdal
rgexf
rgeos
rgl
RgoogleMaps
rhub
rJava
rjson
RJSONIO
rlang
rlist
RLRsim
rmapshaper
rmarkdown
Rmisc
Rmpi
rms
RMySQL
RNiftyReg
RNHANES
roxygen2
rockchalk
ROCR
rpart
rpart.plot
rpf
rprojroot
rsample

RSNNS
rstan
rstanarm
rsvg
RTextTools
rticles
rTorch
Rtsne
rtweet
RUnit
rvest
rworldmap
rworldxtra
scatterplot3d
scico
scrypt
seasonal
segmented
sem
semPlot
sf
showtext
shiny
shinydashboard
shinyjs
shinystan
shinytest
shinythemes
signal
simpleNeural
SixSigma
sjstats
skimr
slackr
sloop
sp
spacetime
spacyr
sparklyr
sparkline
sparktf

spc
spectacles
spelling
sqldf
sqliter
staRdom
stargazer
stars
stationaRy
statsr
stlplus
stopwords
StreamMetabolism
stringi
stringr
styler
suncalc
SunsVoc
sunburstR
survival
survivalAnalysis
survivalMPL
survivalROC
survivalsvm
survminer
svglite
svUnit
SwarmSVM
synthpop
targets
TeachingDemos
TeachingSampling
tensorflow
terra
testthat
textdata
tfdatasets
tfdeploy
tfestimators
tfruns
tibble

tictoc
tidygraph
tidymodels
tidyposterior
tidyr
tidytext
tidyverse
timeDate
tinytex
tipr
tm
tmap
tmaptools
torch
torchdatasets
torchvision
transformr
tree
treemap
treemapify
tsibble
TSclust
TSstudio
tweenr
usmap
V8
validate
vcd
vioplot
viridis
visNetwork
vtreat
waterfalls
WaveletComp
wavelets
wavethresh
weathermetrics
webshot
WikidataQueryServiceR
WikidataR
withr

WGCNA
WDI
wordcloud
xaringan
xkcd
xgboost
XLConnect
XML
xtable
xts
yardstick
zeallot
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.

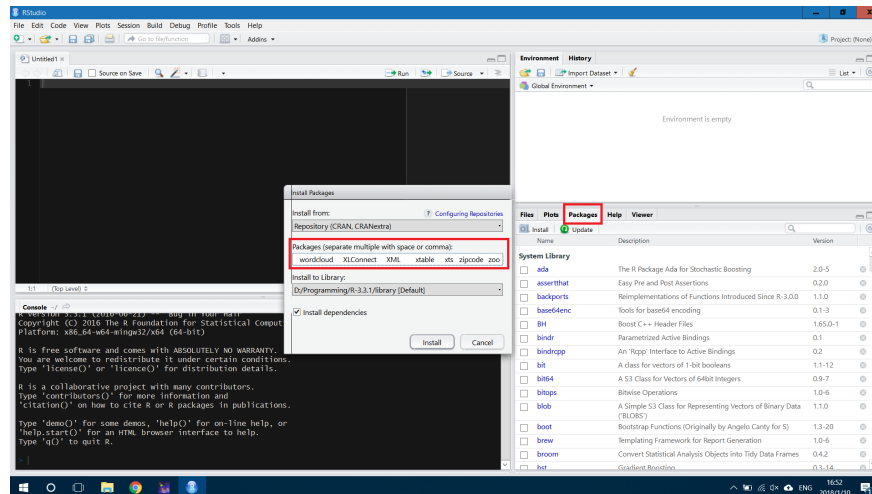


Figure 31

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

11.1.5 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.

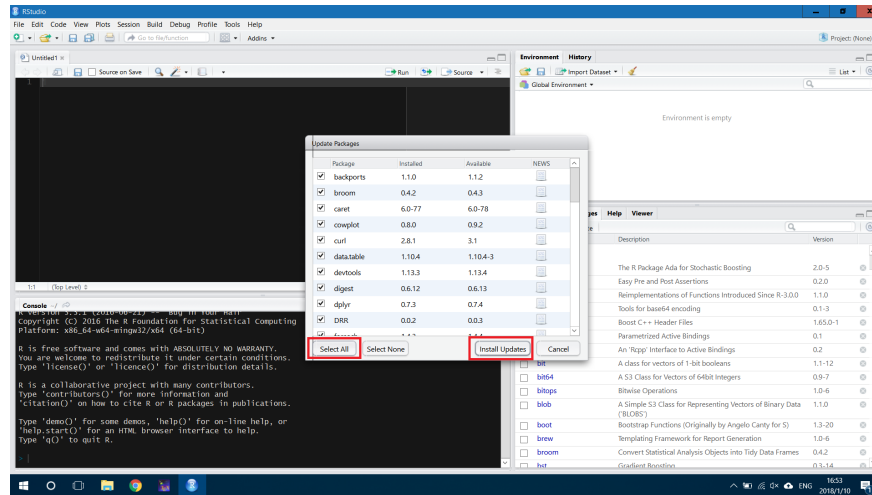


Figure 32

11.2 For Linux

11.2.1 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
```

11.2.2 Git

To install Git, run the following code:

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

11.2.3 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 tmu
cifs-utils openssh-server openssh-client tree http
gdebi curl libcurl4-openssl-dev libssl-dev ffmpeg
```

```
sdle@vuv54:
/$ gpg --keyserver keyserver.ubuntu.com --recv-key E084DAB9
gpg: key 4359ED62E084DAB9: public key "Totally Legit Signing Key <mallory@example.org>" imported
gpg: Total number processed: 1
gpg: imported: 1
sdle@vuv54:
/$ gpg -a --export E084DAB9 | sudo apt-key add -
OK
sdle@vuv54:
/$ sudo apt-get update
Get:2 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran35/ InRelease [3,626 B]
Hit:3 http://us.archive.ubuntu.com/ubuntu bionic InRelease
Hit:4 http://archive.canonical.com/ubuntu bionic InRelease
Get:5 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Ign:6 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64 InRelease
Get:7 http://cran.rstudio.com/bin/linux/ubuntu xenial/ InRelease [3,607 B]
```

Figure 33

```
sdle@vuv54:
/$ sudo apt-get install r-base r-base-dev
Reading package lists... Done
Building dependency tree
Reading state information... Done
r-base-dev is already the newest version (3.6.3-1bionic).
r-base-dev set to manually installed.
r-base is already the newest version (3.6.3-1bionic).
0 upgraded, 0 newly installed, 0 to remove and 2 not upgraded.
```

Figure 34

```
sdle@vuv54:~$ sudo apt-get install r-cran-xml pkg-config libxml2-dev libtiff5-dev fftw3 fftw3-dev tmux cifs-utils openssh-server openssh-client tree htop gdebi curl libcurl4-openssl-dev libssl-dev ffmpeg
Reading package lists... Done
Building dependency tree
Reading state information... Done
Note, selecting 'libfftw3-3' instead of 'fftw3'
Note, selecting 'libfftw3-dev' instead of 'fftw3-dev'
htop is already the newest version (2.1.0-3).
libfftw3-dev is already the newest version (3.3.7-1).
pkg-config is already the newest version (0.29.1-0ubuntu2).
gdebi is already the newest version (0.9.5.7+nmu2).
libfftw3-3 is already the newest version (3.3.7-1).
tree is already the newest version (1.7.0-5).
cifs-utils is already the newest version (2:6.8-1ubuntu1.1).
curl is already the newest version (7.58.0-2ubuntu3.14).
libcurl4-openssl-dev is already the newest version (7.58.0-2ubuntu3.14).
libssl-dev is already the newest version (1.1.1-1ubuntu2.1~18.04.10).
libtiff5-dev is already the newest version (4.0.9-5ubuntu0.4).
```

Figure 35

- 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-1.4.1717-amd64.deb
sudo gdebi -n rstudio-1.4.1717-amd64.deb
rm rstudio-1.4.1717-amd64.deb
```

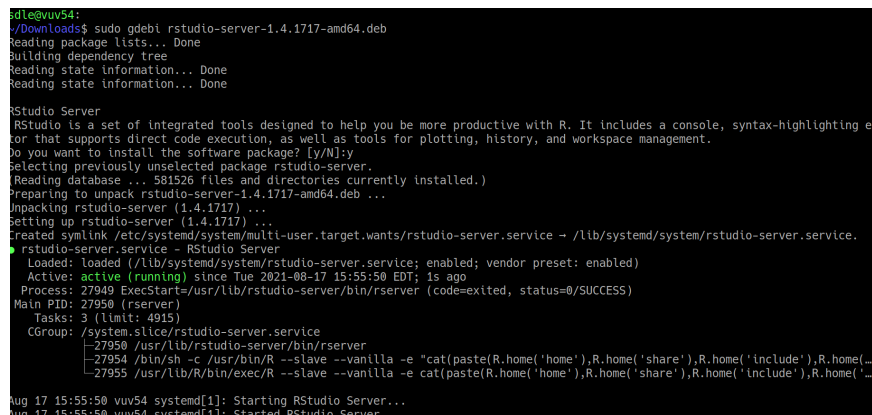


```
sdl@vuv54:~$ cd /home/sdl/Downloads/
sdl@vuv54:~/Downloads$ sudo apt-get install gdebi-core
Reading package lists... Done
Building dependency tree
Reading state information... Done
gdebi-core is already the newest version (0.9.5.7+nmu2).
0 upgraded, 0 newly installed, 0 to remove and 2 not upgraded.
sdl@vuv54:~/Downloads$ wget https://download2.rstudio.org/server/bionic/amd64/rstudio-server-1.4.1717-amd64.deb
--2021-08-17 15:55:23-- https://download2.rstudio.org/server/bionic/amd64/rstudio-server-1.4.1717-amd64.deb
Resolving download2.rstudio.org (download2.rstudio.org)... 99.84.189.44, 99.84.189.81, 99.84.189.124, ...
Connecting to download2.rstudio.org (download2.rstudio.org)|99.84.189.44|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 57613138 (55M) [application/x-deb]
Saving to: 'rstudio-server-1.4.1717-amd64.deb'

rstudio-server-1.4.1717-amd64.de 100%[=====] 54.94M 38.6MB/s in 1.4s

2021-08-17 15:55:24 (38.6 MB/s) - 'rstudio-server-1.4.1717-amd64.deb' saved [57613138/57613138]
```

Figure 36



```
sdl@vuv54:~/Downloads$ sudo gdebi rstudio-server-1.4.1717-amd64.deb
Reading package lists... Done
Building dependency tree
Reading state information... Done
Reading state information... Done

RStudio Server
RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, and workspace management.
Do you want to install the software package? [y/N]: y
Selecting previously unselected package rstudio-server.
Reading database ... 581526 files and directories currently installed.)
Preparing to unpack rstudio-server-1.4.1717-amd64.deb ...
Unpacking rstudio-server (1.4.1717) ...
Setting up rstudio-server (1.4.1717) ...
Created symlink /etc/systemd/system/multi-user.target.wants/rstudio-server.service → /lib/systemd/system/rstudio-server.service.
● rstudio-server.service - RStudio Server
   Loaded: loaded (/lib/systemd/system/rstudio-server.service; enabled; vendor preset: enabled)
   Active: active (running) since Tue 2021-08-17 15:55:50 EDT; 1s ago
     Process: 27949 ExecStart=/usr/lib/rstudio-server/bin/rsrver (code=exited, status=0/SUCCESS)
    Main PID: 27950 (rsrver)
      Tasks: 3 (limit: 4915)
   CGroup: /system.slice/rstudio-server.service
           └─27950 /usr/lib/rstudio-server/bin/rsrver
             └─27954 /bin/sh -c /usr/bin/R --slave --vanilla -e "cat(paste(R.home('home'),R.home('share'),R.home('include'),R.home('...'))"
               └─27955 /usr/lib/R/bin/exec/R --slave --vanilla -e cat(paste(R.home('home'),R.home('share'),R.home('include'),R.home('...'))"

Aug 17 15:55:50 vuv54 systemd[1]: Starting RStudio Server...
Aug 17 15:55:50 vuv54 systemd[1]: Started RStudio Server.
```

Figure 37

11.3 For Mac

11.3.1 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](#).

11.3.2 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.

11.3.3 LaTeX

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

- The current distribution as of today (March 21, 2023) 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.

11.3.4 Git

There are several ways to install Git on a Mac. In fact, if you've installed XCode (or its 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"
```

11.3.5 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, Mine Çetinkaya-Rundel, and Christopher D. Barr, *OpenIntro Statistics: Fourth Edition*. S.I.: OpenIntro, Inc., 4th edition ed., 2019.
- [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: 2nd Ed., with Applications in R*. New York: Springer, 2nd ed. 2021 edition ed., July 2021.
- [6] Francois Chollet and J. J. Allaire, *Deep Learning with R*. Manning Publications, Jan. 2018.
- [7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Adaptive Computation and Machine Learning Series, Cambridge, MA, USA: MIT Press, Nov. 2016.
- [8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics, New York: Springer-Verlag, 2 ed., 2009.
- [9] Citrix, “Citrix receiver for xen VDIs and xen apps,” 2014. 00000.
- [10] C. Portal, “CWRU CSE portal for VDIs and XenApps,” 2014.
- [11] R, “R (programming language),” Aug. 2014. 00000 Page Version ID: 621330268.
- [12] R. Project, “The r project for statistical computing,” 2014.
- [13] RStudio, “RStudio,” 2014. 00000.
- [14] R. Peng, “Computing for data analysis: Week 1 - YouTube,” 2014.
- [15] P. Torfs and C. Brauer, “A (very) short introduction to r,” 2014.
- [16] “Portal:free software,” Sept. 2014. 00000 Page Version ID: 581465934.
- [17] “Gvim online,” 2014. 00000.
- [18] “Neovim,” 2014. 00000.
- [19] “Git (software) - wikipedia, the free encyclopedia,” 2014. 00000.
- [20] “Git,” 2014. 00027.
- [21] “GitHub,” 2014. 00004.
- [22] “Bitbucket: Free source code hosting for git,” 2014. 00000.
- [23] S. Exchange, “Stack exchange,” 2014.
- [24] T. Python, “The python tutorial — python v2.7.8 documentation,” 2014. 00000.
- [25] Python, “Python.org,” 2013.
- [26] “The hitchhiker’s guide to python! — the hitchhiker’s guide to python,” 2014. 00000.
- [27] “kennethreitz/python-guide,” 2014. 00000.
- [28] NumPy, “NumPy — numpy,” 2014.
- [29] 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.
- [30] SciPy, “SciPy.org — SciPy.org,” 2014.
- [31] PythonXY, “pythonxy - scientific-oriented python distribution based on qt and spyder - google project hosting,” 2014. 00000.
- [32] IPython, “IPython shell and notebook,” 2014.
- [33] Spyder, “Spyder is the scientific PYTHON development environment,” 2014. 00000.
- [34] “TeX users group (TUG),” 2014. 00000.

- [35] LaTeX, “LaTeX - wikibooks, open books for an open world,” 2014. 00000.
- [36] Zotero, “Zotero reference/citation manager, BibTeX client,” 2014.
- [37] R. H. French, “DSCI351-4511: Exploratory data analysis for energy & manufacturing,” 2015.
- [38] C. Commons, “Creative commons - about the licenses,” 2014.
- [39] “Creative commons — attribution-ShareAlike 4.0 international — CC BY-SA 4.0,” 2015.
- [40] C. Commons, “Creative commons license,” Aug. 2014. 00000 Page Version ID: 618703231.
- [41] Gnu, “gnu.org,” 2014. 00007.
- [42] G. GPL, “GNU general public license,” Aug. 2014. 00000 Page Version ID: 622300724.