DSCI353-353M-453 Syllabus
Data Science Modeling, Prediction and Statistical and Machine Learning
Spring 2020 Tuesday, Thursday 11:30 am to 12:45 pm, Nord 356

Prof. Roger H. French, TA: Peitian Wang

January 16, 2020

1 Joint Undergraduate and Graduate Course

1.1 DSCI353 and DSCI353M

DSCI353 is the 4th 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 Applied Data Science in the CWRU Bulletin https://bulletin.case.edu/schoolofengineering/datascience/#minortext, and also the https://case.edu/datascience/students/degree-programs/undergraduate-applied-data-science-minor.

DSCI353 will introduce students to linear and beyond linear modeling, prediction and machine learning (including Kera/TensorFlow), the steps in a data analysis the following data cleaning, exploratory data analysis and introduction to linear modleing (the subject of DSCI351-451). This course will use an open data science tool chain consisting of R coding, 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.

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 assembly 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/proc 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 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

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 you project involves this
- Scripts is where to write you scripts for data analysis
- Topics is where to write reports and presentation for your data analysis

2.2 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 "2001-DSCI-Acknowledgement-of-IP.pdf" form.

2.3 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

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 DSCI353-353M Prerequisites

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

For DSCI453 it is assumed a student can start without explicit prerequisites.

3 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

$$Seven\ LabExercises,\ worth\ 8\ points\ each = 56pts.$$
$$353 - 353M - 453SemProjGradingFeedback,\ worth\ 4\ points\ each = 12pts.$$
$$Midterm\ Exam = 10pts.$$
$$Final\ Exam = 22pts.$$
$$\textbf{Total = 100pts}$$

DSCI453 is graded on a 140 point basis

$$Seven\ LabExercises,\ worth\ 8\ points\ each = 56pts.$$
$$353 - 353M - 453SemProjGradingFeedback,\ worth\ 4\ points\ each = 12pts.$$
$$Midterm\ Exam = 10pts.$$
$$Final\ Exam = 22pts.$$
$$12\ Weekly\ SemProj\ Updates\ worth\ 1\ point\ each\ = 12pts.$$
$$1\ SemProj\ with\ 3\ presentations\ \&\ 1\ Final\ Report = 28pts.$$
$$\textbf{Total = 140pts}$$

4 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.

ISLRv6 is an Springer book which is available for free as a pdf. In addition a copy can be purchased. ISLR is the introductory book, for which Elements of Statistical Learning (ESL) 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.

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. R4DS is used in both DSCI351... and DSCI353 courses

DLwR is used in DSCI353-353m-453 for Deep Learning and TensorFlow

4.1 Background Data Science books from DSCI351

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.

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

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

4.2 DSCI353-353M-453 Textbooks

See Figure 3, 4 and 5.

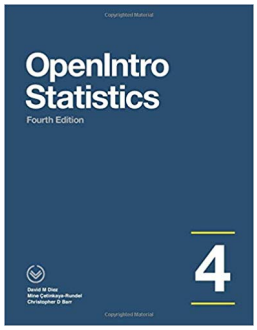

Figure 3: **OIS**: David M. Diez, Christopher D. Barr, and Mine Cetinkaya-Rundel, **OpenIntro Statistics 4th Ed.** 2015 [3]

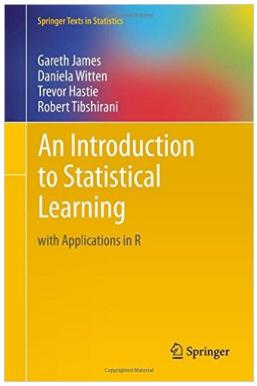

Figure 4: **ISLRv6**: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani **An Introduction to Statistical Learning: with Applications in R**, 2013 [4]

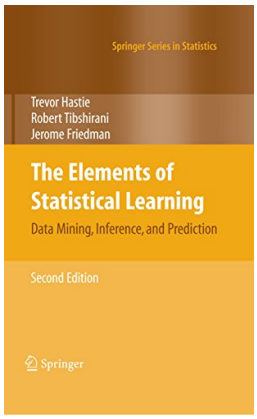

Figure 5: **ESL**: Trevor Hastie, Robert Tibshirani, Jerome Friedman **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2009** [5]

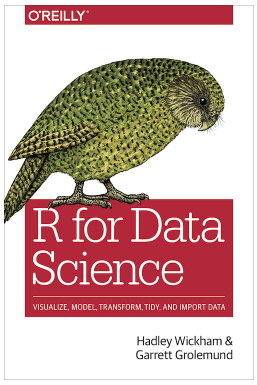

Figure 6: **R4DS**: Garrett Grolemund, Hadley Wickham, **R for Data Science**. 2017 [6]

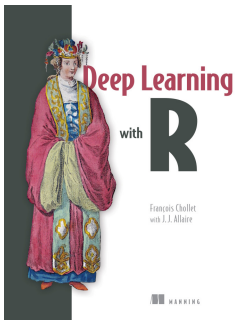

Figure 7: **DLwR**: François Chollet with J. J. Allaire, **Deep Learning with R**. 2019 [7]

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

5 DSCI353-353M-453 Syllabus: Weekly Topics

Day:Date	Foundation	Practicum	Readings (optional)	Due (optional)
w01a:Tu:1/14/20	Open Data Science	R, Rstudio IDE, Git		(LE0)
w01b:Th:1/16/20	What is Data Science	Bash Git, Class Repo	(R4DS-1,2,3)	
w02a:Tu:1/21/20	Statistical Learning	Pred. Analytics	ISLR1,2	(LE0)
w02b:Th:1/23/20	Data Analytic Style	Tidy Data Manip.	(R4DS-4,5,6)	LE1
w03a:Tu:1/28/20	Lin. Regr.	Pairs Plots	(OIS7)	
w03b:Th:1/30/20	Mult. Lin. Regr.	Test Stats	ISLR3 (R4DS-7,8)	
w04a:Tu:2/4/20	Logistic Regr.	Interaction Terms	ISLR4	**LE1:Due**, LE2
w04b:Th:2/6/20	Classification		(OIS8)	
w05a:Tu:2/11/20*	Resample Cross-Valid.	Cluster Analysis	ISLR5	
w05b:Th:2/13/20	Bootstrap	Steps of Data Analysis	DL1,2 (R4DS9-16)	**LE2:Due**
w06a:Tu:2/18/20	LMS: Subset		ISLR6	LE3
w06b:Th:2/20/20	LMS: Feature Selec.	Coeff. Uncertainties	DL3,4 (R4DS17-21)	
w07a:Tu:2/25/20*	BeyondL: Spline, GAM	**SemProj1-453**	ISLR7	
w07b:Th:2/27/20	MidTerm Review	**SemProj1-3/452**	DL5,6 (R4DS22-25)	**LE3:Due**
w08a:Tu:3/3/20	**MIDTERM EXAM**			
w08b:Th:3/5/20	Dim. Reduc. & PCA		ISLR8 (R4DS26-30)	
Tu:Th:3/9-13/20	**SPRING BREAK**			
w09a:Tu:3/17/20	Regr. Trees	Dec. Trees	ISLR9	LE4
w09b:Th:3/19/20	Bagging, Boosting	How DT Work	ISLR10	
w10a:Tu:3/24/20*	Support Vector Mach.	**SemProj2-453**	ESL11	**LE4:Due**
w10b:Th:3/26/20*	ML Overview, Caret	**SemProj2-3/452**	DLR1	LE5
w11a:Tu:3/31/20*	Neural Networks	MNIST digits	DLR2	
w11b:Th:4/2/20*	NN Topo., Types, Train	ImageNet	DLR3	**LE5:Due**
w12a:Tu:4/7/20*	R-Keras/TensorFlow	CNN w TF	DLR4	LE6
w12b:Th:4/9/20	CNN w/TF	EL Image Sup. ML		
w13a:Tu:4/14/20	CNNs w/small data	DLwR 2.1		**LE6:Due** LE7
w13b:Th:4/16/20	Tboard, TFestimators	pretrained CNNS		
w14a:Tu:4/21/20	R Packaging	**SemProj3-453**		
w14b:Th:4/23/20	Final Exam Review	**SemProj3-3/452**		**LE7:Due**
	FINAL EXAM	**Th. 3/30/2020, 12-3pm**	Nord 356	

Table 1: DSCI353-353M-453 Weekly Syllabus. R4DS-x.y, OISx.y, ISLRx.y refers to chapters and sections assigned as reading. DLx is deep learning articles

6 Contact Information

Prof. Roger H. French

- White 536, and electronically.
- Email is best: rxf131@case.edu, Use DSCI353-353M-453 in the subject line
- CWRU-DSCI Slack, Class Channel
- @frenchrh on twitter
- Office Phone 216 368 3655, Cell Phone 302 468 6667

TA: Peitian Wang

- White 615, and electronically.
- Email is best: pxw223@case.edu

  Use DSCI353-353M-453 in the subject line
- on twitter
- Cell Phone: 216-703-8251

7 Course Mechanics

7.1 Lectures

Spring 2017 Tuesday, Thursday 11:30 am to 12:45 pm Nord 356

7.2 DSCI Class Slack Channel

There is a Slack channel for this class

To join go to https://cwru-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

Office Hours: Monday's and Wednesdays, 4pm to 5pm, in White 540 Consultations: After class or as needed. Contact Prof. French and Peitian Wang, by Slack, email or in person.

7.4 Homework / 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, YourName, HW#... e.g. DSCI353-453FrenchHW4.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.

7.5 Extra Credit DSCI353-353M Data-science Semester Project Report

For DSCI353-353M, the final data science research report should be written like a scientific paper and have the following types of sections.

- 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

7.5.1 Abstract

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

7.5.2 Introduction

Background and motivation of the Data Science question

7.5.3 Data Science Methods

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

7.5.4 Exploratory Data Analysis

Results and steps in the data analysis

7.5.5 Statistical Learning: Modeling & Prediction

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

7.5.6 Discussion

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

7.5.7 Conclusions

7.5.8 Acknowledgments

7.5.9 References

7.5.10 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) VDIs

You will not need to install software on your personal computers.
Instead you can install the Citrix Receiver [8] and then login to the CWRU CSE Portal. [9]
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 homework and projects. Its generally useful language for statistical analysis and data science.

- The R Project for Statistical Computing [10] main website
- R programming language R is a free software programming language and software environment for statistical computing and graphics.[11]
- RStudio provides popular open source and enterprise-ready professional software for the R statistical computing environment. [12]

- 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 [10] 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. [13]
- A (very) short introduction to R [14]
- 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. [15]
- vim (or Gvim the GUI version) is a powerful text and code editor, that is universally available on all Linux and mac computers.[16] NeoVim is a new Gvim fork.[17] It can be installed on windows computers, its available on the ODS VDIs.. [16]
- Git (Wikipedia) is a distributed content versioning system that is very popular. It enables collaborative code development and LaTeX writing projects.[18]
- Git server software is installed on each computer.[19]
- GitHub is a Git server website used for collaborative code development.[20]
- 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.[21]
- Stack Exchange [22] 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. [23]
- The Python main website. [24]
- The Python Tutorial — Python v2.7.8 documentation [23]
- 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. [25] [26]
- NumPy is the fundamental package [27] for scientific computing with Python.
- FiPy: Partial Differential Equations with Python [28]
- SciPy is a python-based ecosystem [29] of open-source software for mathematics, science, and engineering.
- PythonXY - Scientific-oriented Python Distribution based on Qt and Spyder that runs on Windows. [30]
- IPython Shell and Notebook [31]
- Spyder is the Scientific PYthon Development Environment [32]

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. [33]
- LaTeX - Wikibooks, open books for an open world. [34]
- Zotero Reference-Citation Manager, BibTeX Client [35]

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 [36] is covered by a Creative Commons [37] 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. [38]

More information on licensing open works, can be found on Wikipedia. [39]

GNU [40] is the developer of the GPL License [41] that is used for many open source software projects, such as Linux.

10.2 Software Licenses

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.

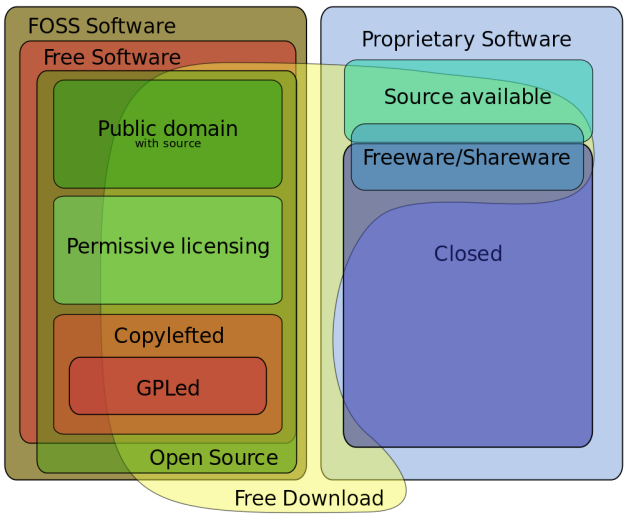

Figure 8: Diagram of software under various licenses according to the FSF and their The Free Software Definition: on the left side "free software", on the right side "proprietary software". On both sides, and therefore mostly orthogonal, "free download" (Freeware). CC0, https://commons.wikimedia.org/w/index.php?curid=46544815

Figure 9: 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, identing, 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

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.

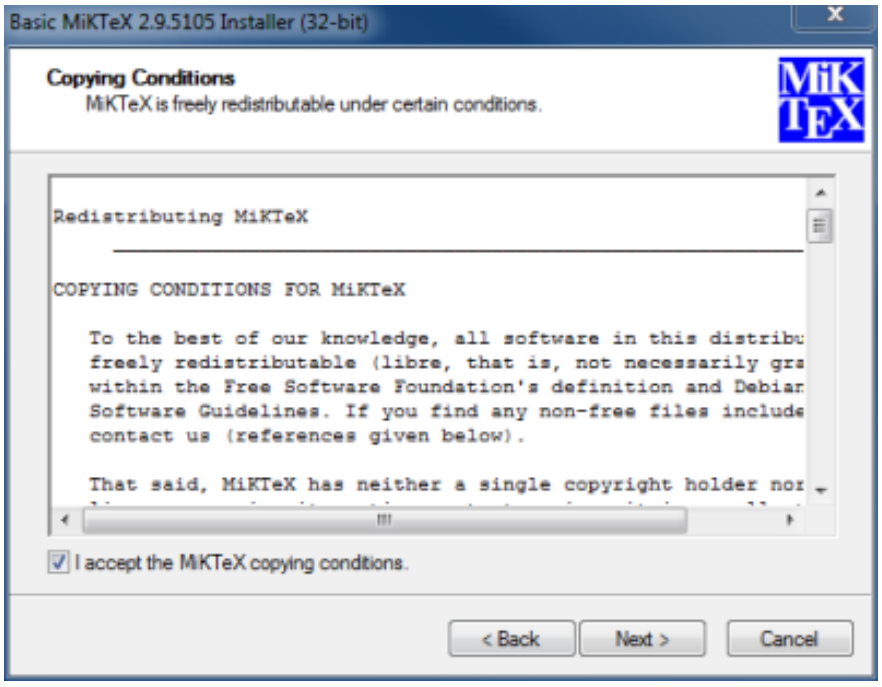

Figure 10

- 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.

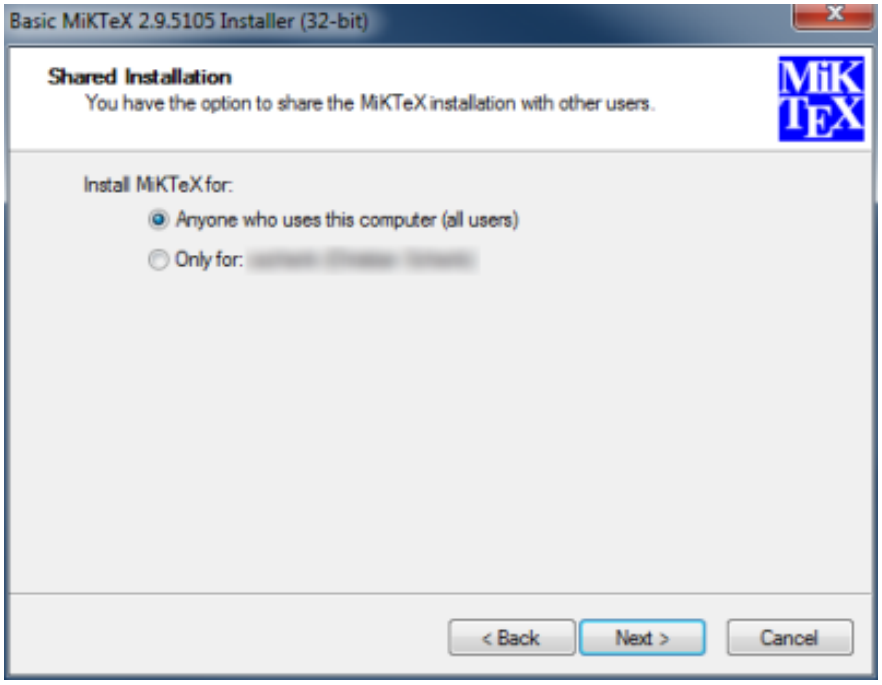

Figure 11

- 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.

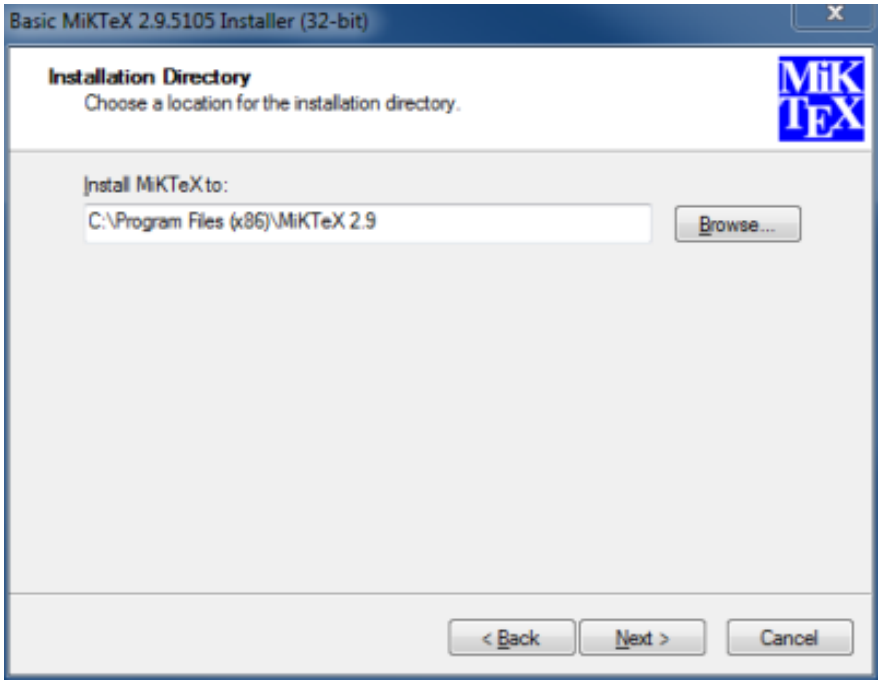

Figure 12

- 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.

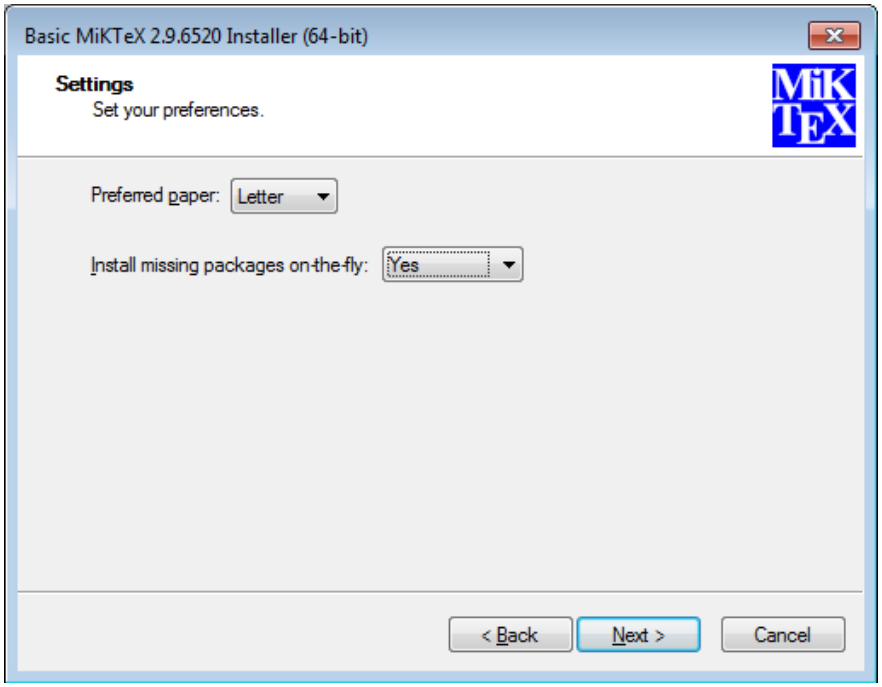

Figure 13

- 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 14

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

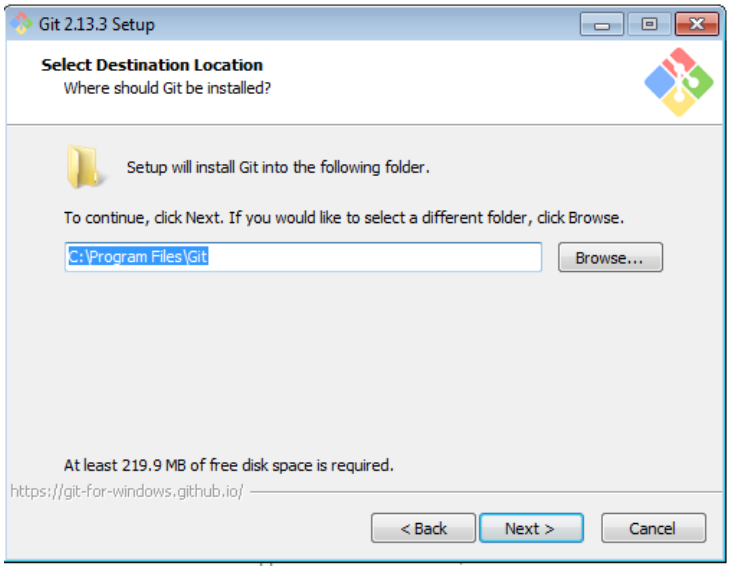

Figure 15

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

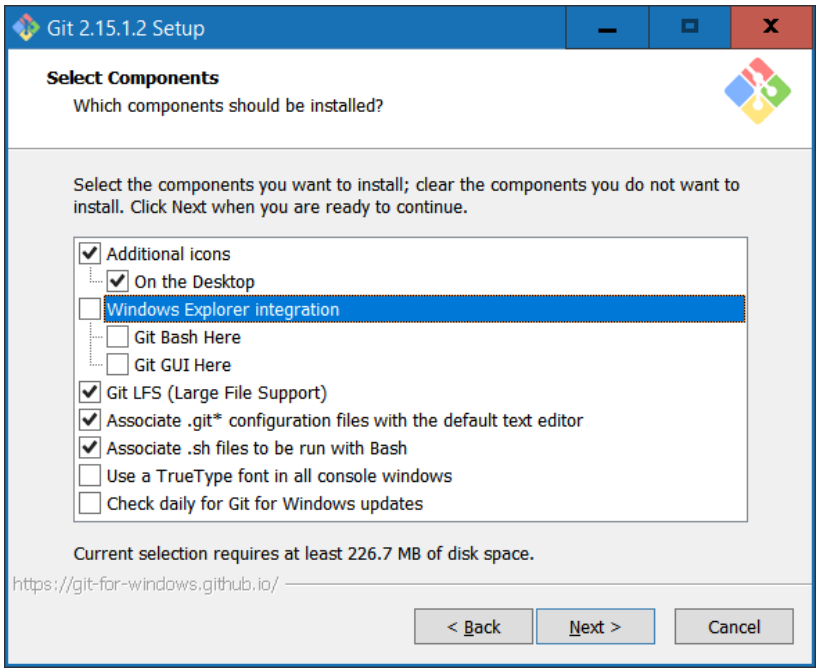

Figure 16

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

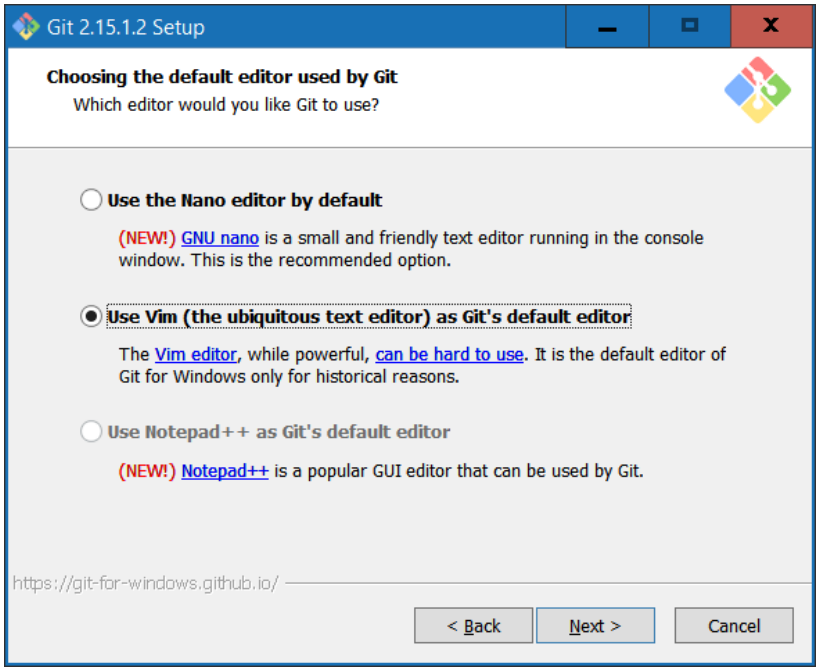

Figure 17

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

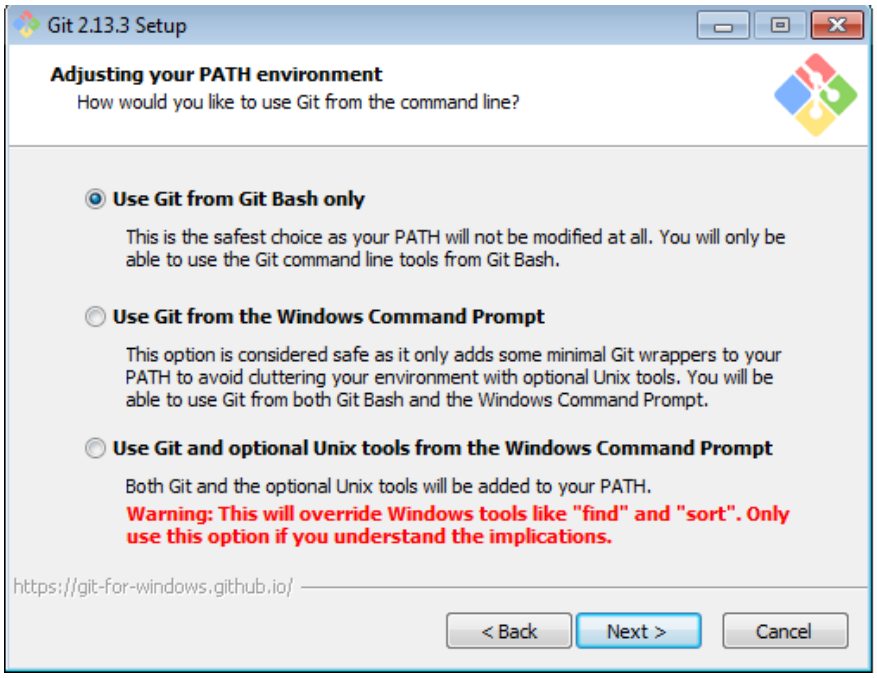

Figure 18

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

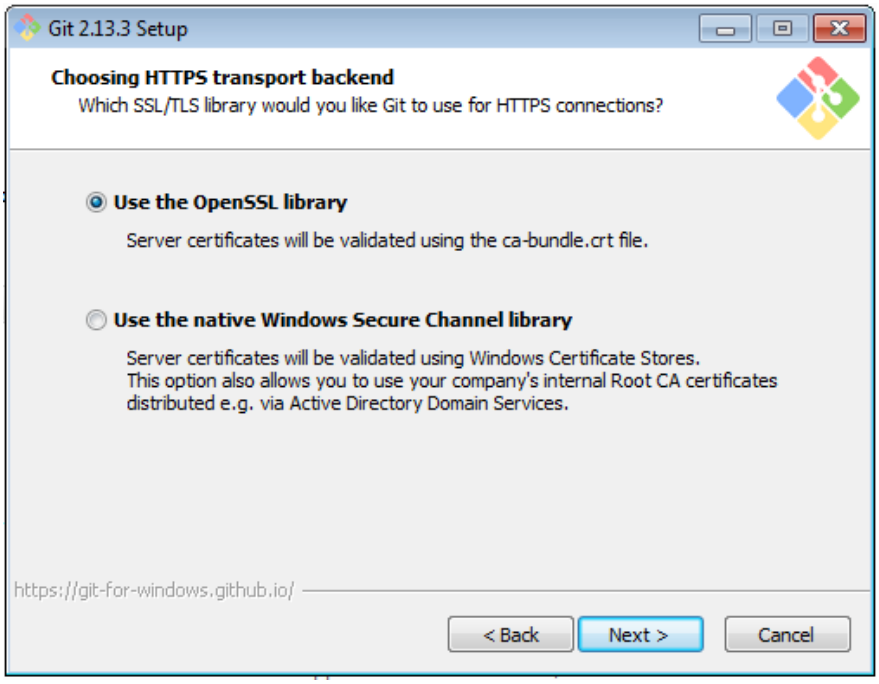

Figure 19

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

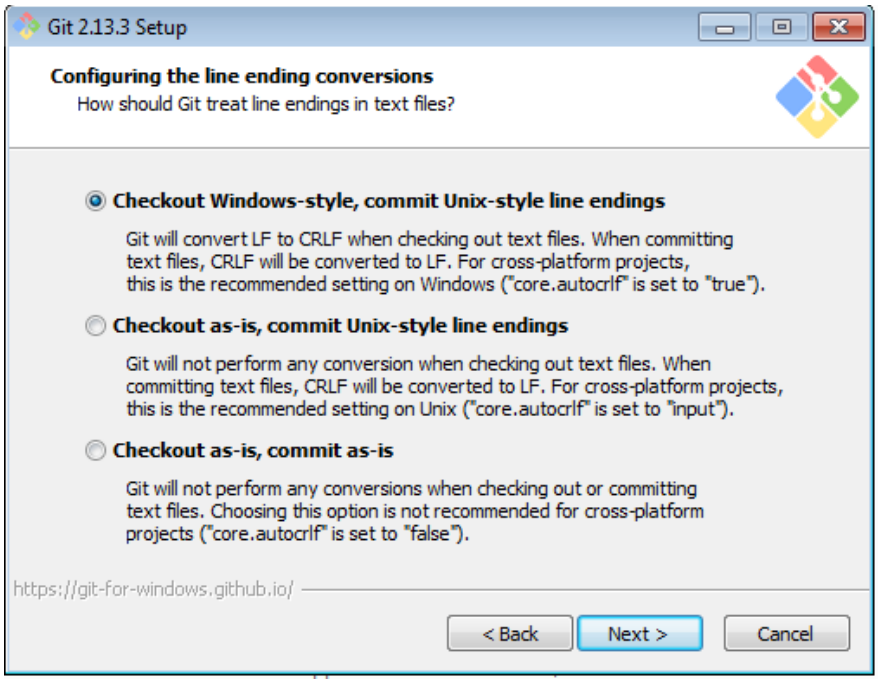

Figure 20

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

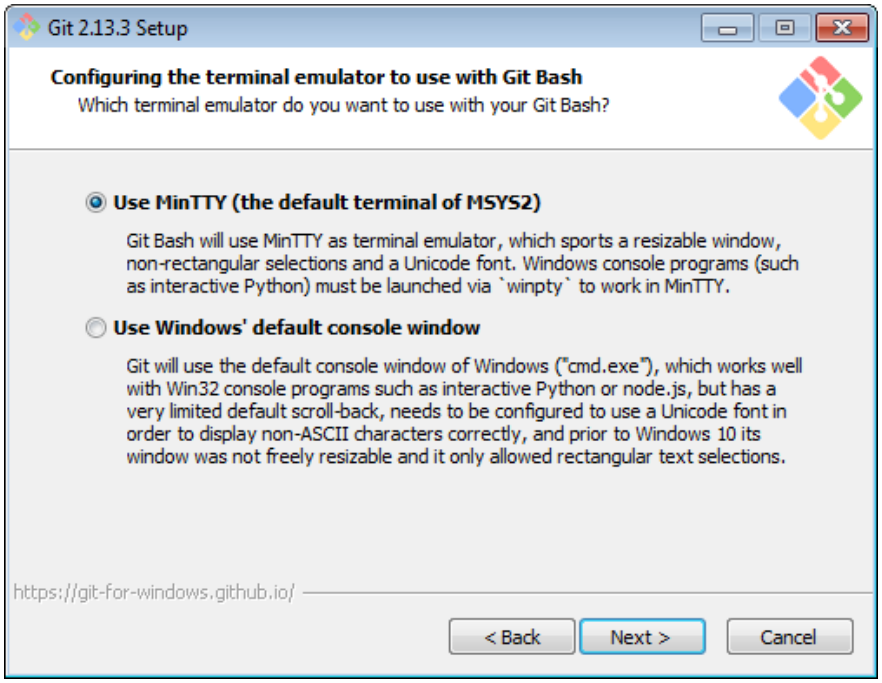

Figure 21

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

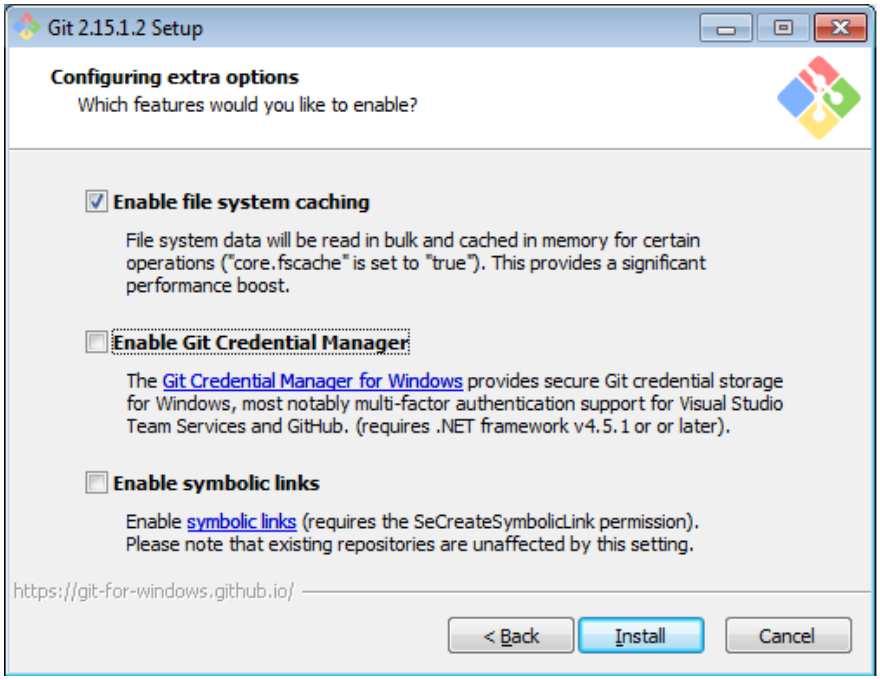

Figure 22

- 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.

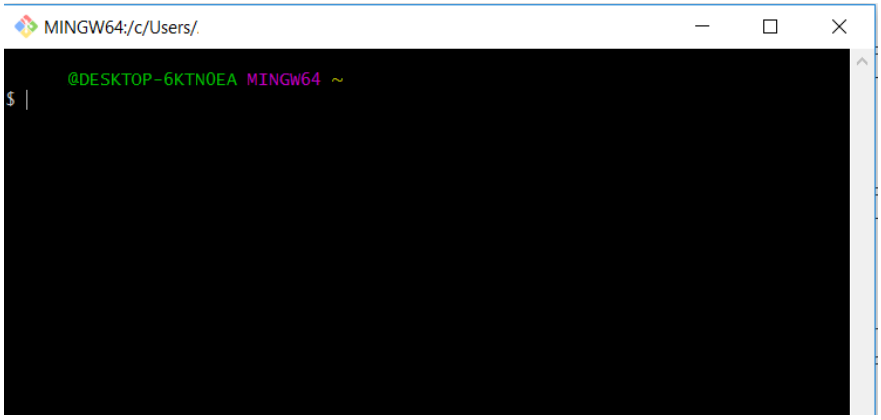

Figure 23

- 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:

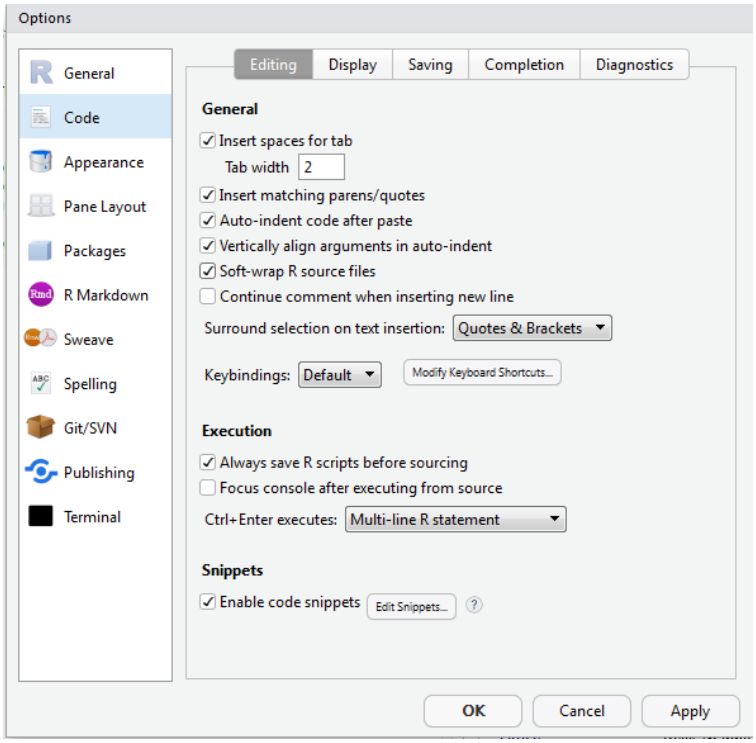

Figure 24

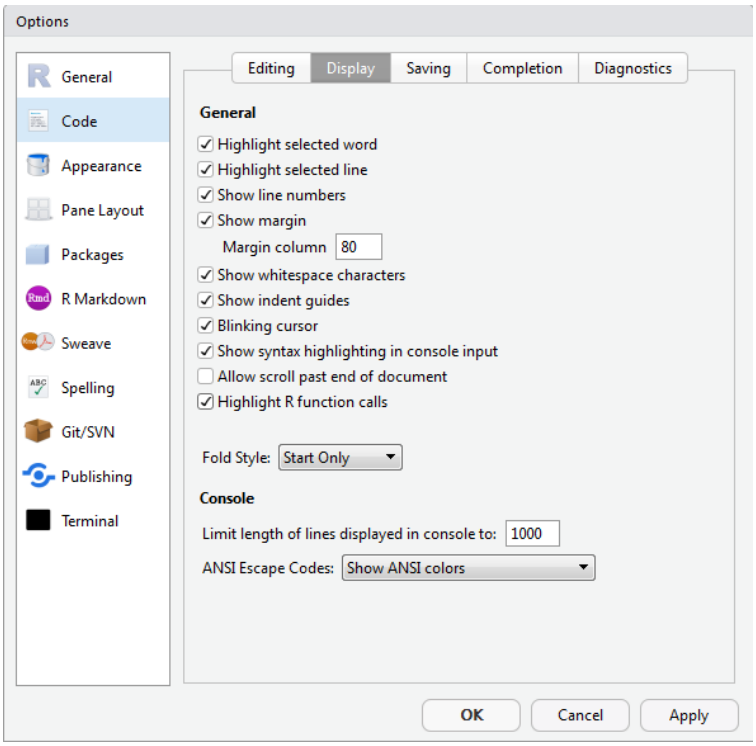

Figure 25

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

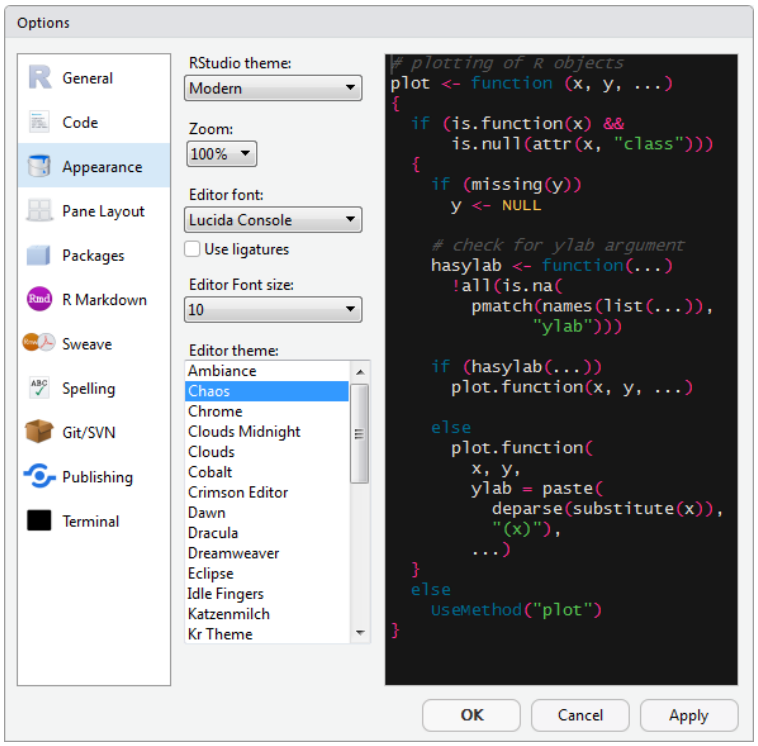

Figure 26

- 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 adabag akima AmesHousing animation anytime arules aspace astsa available bagR-boostR baseline BayesFactor bayesSurv BayHaz bcpa bda birk bit64 blogdown BMA book-down bookdownplus BoomSpikeSlab boot bootstrap breakDown breakpoint brms broom bsts C50 Cairo car caret caretEnsemble CausalImpact centiserve changepoint changepoint.np ChemoSpec cgwtools class ClimClass cloudml cluster.datasets coda CORElearn corrgram corrplot cowplot cpca ctv CVXR data.table data.tree dataMaid DBI DBItest dbscan ddiv devtools DiagrammeR DiagrammeRsvg dice digest DMwR doParallel dplyr DT dummies dtwclust e1071 eemR ElemStatLearn epiDisplay factoextra fastcluster feather flexclust fore-cast foreach gam gapminder gbm gclus GGally gganimate ggbiplot ggforce ggmap ggplot2 ggpubr ggQC ggRandomForest ggraph ggridges ggthemes ggvis glmnet gmodels googleVis graphlayouts gridBase gridExtra gsl gstat gWidgets h2o HadoopStreaming HarmonicRegres-sion hcp hdpca hexbin HH httr htmlwidgets htmltools hyperSpec igraph infer ipred IQCC IRk-ernel ISLR itertools jsonlite kableExtra keras kerasformula kerasR kernlab keyring kgc klaR knitcitations knitr Lahman lars lavaan lavaan.survey leaps learningr learNN learnBayes lime lintr lme4 lobstr logitnorm magick magrittr Make mapdata Mapmate mapproj maps mark-down maptools MASS Matrix MatrixModels matrixStats markovchain mcmc MCMCglmm metRology Metrics mgcv minpack.lm MTS multiway NbClust netSEM neural neuralnet Neu-

ralNetTools nnet nycflights13 odbc OIdata olsrr OIsurv onehot OneR onlineCPD openintro optimx optiRum packHV packrat pacman parallelSVM patchwork pca3d PerformanceAnalytics pipeR plot3D plotmo plotKML plotly pls Plumber plyr plyrmr png pool prodlim pROC prophet profvis propagate proxy pryr psych purrr qcc qtlmt qualityTools quantmod r2d3 randomForest randomForestSRC ranger raster rasterVis rCharts RColorBrewer Rcpp RCurl Rdice Rdpack readr recipes RefManageR relaimpo reshape reshape2 reticulate rgdal rgexf rgeos rggobi rgl rJava rjson RJSONIO jsonvalidate rlist RLRsim rmarkdown Rmisc Rmpi rms RMySQL RNiftyReg rNMF roxygen2 rpart rprojroot rPython rsample RSNNS rstan rstanarm rsvg RTest RTextTools rticles Rtsne rtweet RUnit rvest rworldmap rworldxtra scatterplot3d scrypt segmented sem shiny shinydashboard shinyjs shinystan shinytest shinythemes signal simpleNeural SixSigma sp sparklyr sparktf spc spelling sqldf sqliter sqlutils stationaRy statsr stlplus stockPortfolio StreamMetabolism stringi stringr styler survival survivAll survivalAnalysis survivalMPL survivalROC survivalsvm survminer svglite svUnit SwarmSVM synthpop TeachingDemos TeachingSampling tensorflow testthat tfdatasets tfdeploy tfestimators tfruns tibble tictoc tidyGraph tidymodels tidyposterior tidyr tidytext tidyverse tidygraph timeDate timevs tinytex tm transformr tree TSclust TSstudio tweenr tweezer validate vtreat WaveletComp wavelets wavethresh wmtsa WGCNA WDI wordcloud XLConnect XML xtable xts yardstick zipcode zoo

You can go to the highlighted tab in the below picture and install-upgrade your packages here.

To install, simply paste the list of packages in the window.

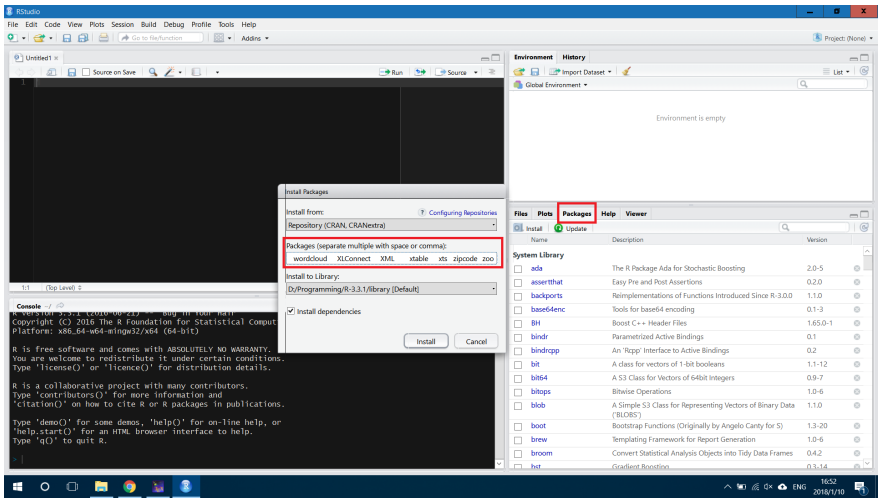

Figure 27

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

11.1.4 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.

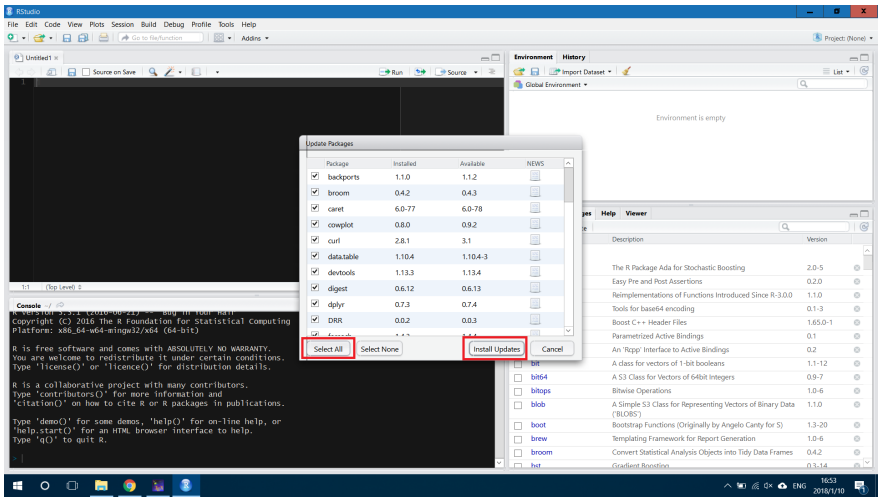

Figure 28

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  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
```

11.3 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 (January 16, 2020)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 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"
```

11.3.5 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.l.: OpenIntro, Inc., 4th edition ed., 2019.

[4] 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.

[5] 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.

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

[7] Francois Chollet and J. J. Allaire, *Deep Learning with R*. Manning Publications, Jan. 2018.

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

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

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

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

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

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

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

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

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

[17] "Neovim," 2014. 00000.

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

[19] "Git," 2014. 00027.

[20] "GitHub," 2014. 00004.

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

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

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

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

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

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

[27] NumPy, "NumPy — numpy," 2014.

[28] 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.

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

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

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

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

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

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

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

[36] R. H. French, "DSCI351-4511: Exploratory data analysis for energy & manufacturing," 2015.

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

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

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

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

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