

Course syllabus

[Jump to today](#)

Big Data Management and Analysis in Linux

Date of course: Block 3 (July 30 to August 10)

ECTS credits: 3 credits

Contact hours: 45

Lecturers: Dr. Aysu Okbay, Richard K. Linnér

Contact info

- Dr Aysu Okbay: [a.okbay@vu.nl \(mailto:a.okbay@vu.nl\)](mailto:a.okbay@vu.nl),
- Richard K. Linnér: r.karlssonlinner@vu.nl
- Contact VU Amsterdam Summer School: [amsterdamsummerschool@vu.nl \(mailto:amsterdamsummerschool@vu.nl\)](mailto:amsterdamsummerschool@vu.nl), and telephone: +31 20 59 86429

Course Description

The growing availability of extremely large datasets requires scientists and analysts to use powerful supercomputers or computer clusters to store, manage, and analyze these data. These servers typically run on UNIX/LINUX, which requires some programming skills and insights into suitable software packages. Our course will introduce you to the UNIX command line utility, teach you how to manage very large datasets (e.g. using sed, awk, and grep commands) and create simple shell scripts to analyze your data (e.g. using a UNIX version of the freely available statistics program R). You will also learn how to present your data and results in customized plots and figures. These skills are extremely valuable for scientists from all disciplines as well as for business practitioners (e.g. consultants or financial analysts) who are planning to work with big data.

The format of the course is three-hour lectures in the morning, followed by two hours of supervised work in computer tutorials in the afternoon. Both the lectures and tutorials will be held in a computer room (**HG-1G28**). The lectures will be interactive, with short examples that allow students to apply the introduced concepts. In the tutorials, students will get more hands-on training in a supervised environment with exercises covering the day's topics. They will also have the opportunity to work on the assignments. The computer room will stay open to students for self-study after the tutorials.

Students are not required to bring their own laptops, but they are allowed to do so if they wish to work on their own computers.

Learning Objectives

By the end of this course, the student should understand and feel comfortable with:

- The Unix philosophy and environment; files, processes, pipes, filters and basic utilities

- Login and logout procedures, including remote login using SSH, and setting, protecting, and changing passwords.
- File transfer between systems.
- Text file manipulation with sed, awk, cut, paste, cat,
- Basic text editing using the vim editor
- Automation through shell scripts
- Version control with Git
- Working with R through the UNIX command line
- Plotting in R

Excursion(s)

On August 10 we will do an excursion to Amsterdam Science Park to visit the new server park that hosts the high-performance computing infrastructure of SURFsara. Please bring a valid photo ID, as this is a requirement to enter the facilities (student ID is not valid). The exact logistical details (where and when we meet, how we get there, etc.) will be announced at the start of the course.

Reading List

There is no compulsory reading material that needs to be purchased, but the following book by Mark G. Sobell is highly recommended:

Sobell, M. G. (2013). A Practical Guide to Commands, Editors, and Shell Programming, Third Edition. Upper Saddle River: Pearson Education (US).

Additional reading material will be provided prior to lectures.

Assignments

There will be four assignments in total, all of which have to be completed individually. The first three assignments will be completed within the first week. You will have the opportunity to work on the assignments in a supervised environment during the tutorials. The final assignment will cover all topics and you will have all of the second week to complete it. You are expected to work on the final assignment daily, focusing on questions on topics that are already covered.

The assignments will be given and collected on the following dates:

	Date given	Submission deadline
Assignment 1	July 30, Monday	August 1, Wednesday (end of day)
Assignment 2	August 1, Wednesday	August 3, Friday (end of day)
Assignment 3	August 3, Friday	August 6, Monday (end of day)
Final assignment	August 6, Monday	August 18, Friday (end of day)

Grading

Assignments 1-3 will constitute 60% of the final grade (20% each), and the final assignment will be 40%.

Daily course schedule

All lectures and tutorials will be held in room **HG-1G28**.

Week 1 (July 30 - August 3)

Day 1: Introduction to UNIX (Lectures by Okbay, tutorial by Okbay or Linnér)

We will start with some background on UNIX and its history. You will also learn the basics such as logging into a remote server, the UNIX file system, and some basic command line utilities.

16.30 – Preliminary, Pub quiz at The Basket.

Day 2: The Shell (Lectures by Okbay, tutorial by Linnér): This lecture will introduce you to the Shell, and teach you how to work with files using utilities that allow you to display text files, search files for strings, sort files, etc.

Day 3: Text editing and data manipulation (Lecturer: Okbay): In this lecture, you will learn how to create and edit text files using the vim editor. You will also get introduced to the AWK pattern processing language and the sed editor which are useful tools for editing and manipulating big data files.

13.00 – Preliminary, Social activities.

Day 4: Programming the Bourne Again Shell – I (Lectures by Okbay, tutorial by Linnér): This lecture will introduce you to the Bourne Again Shell (bash) and teach you how to write bash scripts to automate the analysis of your data. You will learn how to write functions, define parameters and variables, and submit jobs to a remote server.

Day 5: Programming the Bourne Again Shell – II (Lectures by Okbay, tutorial by Linnér): Picking up from where Lecture 4 left, this lecture will teach you how to use control structures such as if/else conditions, or for and while loops in your scripts. You will learn how to define parameters and variables, and evaluate arithmetical expressions.

Week 2 (August 6-10)

Day 1: Version Control with Git (Lectures by Okbay, tutorial by Linnér): Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later. Especially in

collaborative projects where multiple users can modify the code, version control is a must since it allows you to see who made which modifications and when, and revert your project back to a previous state if needed. In this lecture, you will learn how to use the free and open source distributed version control system Git to manage your projects.

Day 2: Introduction to R (Lectures by Okbay, tutorial by Linnér): This lecture will introduce you to the open source statistical computing environment R. You will learn how to run R from the command line terminal, install packages, and run R scripts, as well as some basic commands to read in and analyze your data.

Day 3: Plotting in R – Basics (Lectures by Linnér): The lecture will introduce the massive plotting capabilities of R. We will start with the basic plotting functions in standard R, discuss how to customize the layout of your plots, and give a quick introduction to the commonly used ggplot package. You will get hands-on experience of data visualization.

13.00 – Preliminary, social activities.

Day 4: Plotting in R – UNIX integration (Lectures and tutorial by Linnér): Simple plots are easily created on your laptop, while plots across multiple analyses each covering millions of data points quickly becomes infeasible for normal workstations. We will introduce how to integrate batch plotting on UNIX servers and how to set up simple control structures in your bash script to simplify user control of your scripts.

Day 5: Review and excursion to the SURFsara facilities (Lectures by Okbay): We will review the covered topics and discuss the final assignment in the morning. After the lecture, we will go on an excursion to the SURFsara facilities at Amsterdam Science Park. This is a great opportunity to put what you have learnt in perspective by seeing what the servers that we work on look like and hearing from the experts about how they function.

15.00 – Preliminary, goodbye drinks at The Basket.

Detailed schedule

The schedule of the course is laid out below. Note that it might be subject to minor changes as the course proceeds.

Week 1

Monday	Tuesday	Wednesday	Thursday	Friday
--------	---------	-----------	----------	--------





10.00 – 11.00	Basic overview and history of UNIX	Basic utilities	vim editor	Writing a shell script	Control structures
	Logging in from a terminal	Working with files	AWK pattern processing language	Job control	
11.15 – 12.30	Working with the Shell	Standard input and output		Processes	Parameters and variables
	The filesystem and file utilities	Redirection, pipes	sed editor	Functions	expressions
12.30 – 13.00	LUNCH				
		Running a command in the background			
13.00 – 15.00	Computer tutorial	Computer tutorial	Social activities	Computer tutorial	Computer tutorial

Week 2

	Monday	Tuesday	Wednesday	Thursday	Friday
10.00 – 11.00		Installing and running R from command line	Introduction to plots	R batch plotting on UNIX server	
	Version control	Installing packages	Overview of plotting capabilities	Introduction to getopts	Review, and Q&A for final assignment
	Git	Reading data into R and basic R commands	Basic plot functions, plot(), hist(), barplot(),	Importing UNIX arguments into R code	

			qqplot()		
			Basic object functions, segments(), abline()		
11.15 – 12.30	Git tutorial	Efficient routines for handling big data in R Running R scripts from terminal	Customization of plot layout Overview ggplot package	Overview of advanced plotting	Review, and Q&A for final assignment
12.30 – 13.00	LUNCH				
13.00 – 15.00	Computer tutorial	Computer tutorial	Social activities	Computer tutorial	Excursion to the SURFsara supercomputer facilities - Exact time will be communicated during the course.

Course summary:

Date	Details	
Wed, 1 Aug 2018	 Assignment 1 (https://canvas.vu.nl/courses/37154/assignments/34247)	due by 23:59
Mon, 6 Aug 2018	 Assignment 2 (https://canvas.vu.nl/courses/37154/assignments/34411)	due by 17:00
Wed, 8 Aug 2018	 Assignment 3 (https://canvas.vu.nl/courses/37154/assignments/34474)	due by 23:00
Sun, 12 Aug 2018	 Assignment 4 (https://canvas.vu.nl/courses/37154/assignments/34823)	due by 23:59