Command-line interfaces and tools for biologists

Pierre Tocquin (ptocquin@uliege.be)

LIÈGE université

2021

Contents

List of Boxes

List of Tables

List of Figures

Chapter 1

Linux: Why and how ?

1.1 Why Linux ?

1.1.1 The operating system of the *Data Science* field.

By looking at the statistics of operating system popularity, we can legitimately question the interest of spending time and energy in learning to master an operating system such as Linux that does not equip more than 2% of desktop computers in Europe (Figure 1.1).

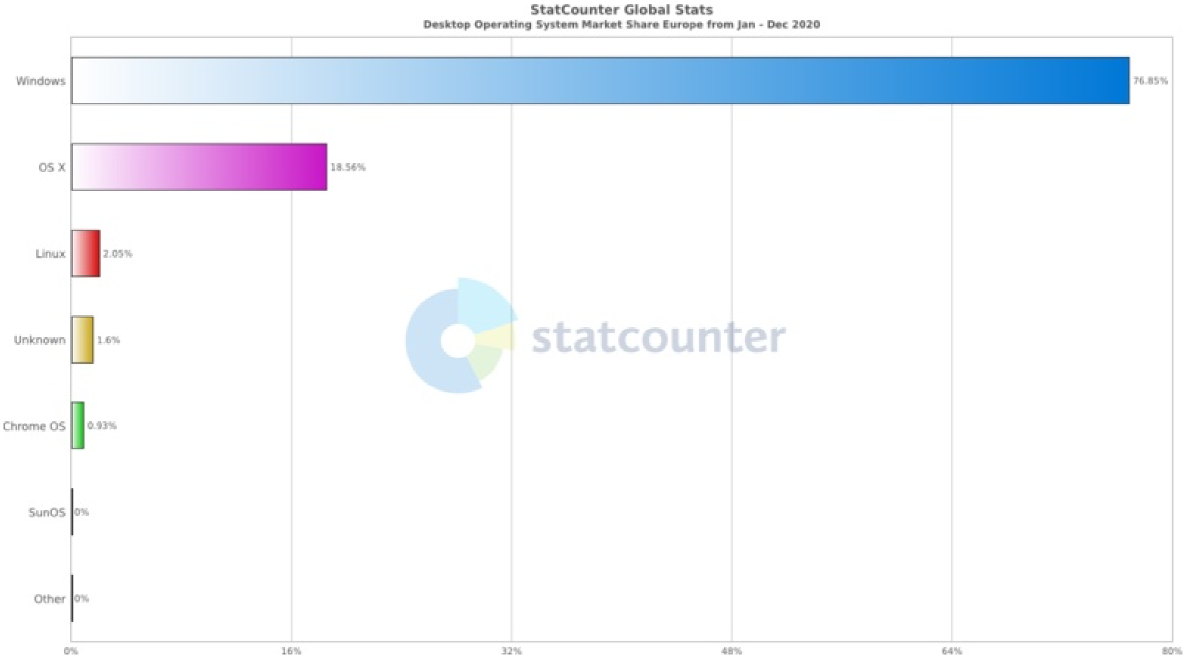

Figure 1.1: Usage statistics for operating systems in Europe in 2020. Source: StatCounter.com

These figures are obviously the result of an analysis carried out for use by the "general public" (those data are collected from websites visits). However, when excluding the domestic use of a computer, any more specific usage generally dictates in an important way the choice that it is made of the operating system. For example, it is remarkable that, among "gamers", Microsoft Windows is almost the only system used. (96,5% in June 2021. Source: Steam). Conversely, if we focus at the systems that equip the supercomputers used for data processing, we see that **all** of the 500 largest machines worldwide are running Linux. (Source: top500.org).

As we are not here to play but to learn the basics of biological data analysis, we could just stop argumenting in favor of Linux… But let us nevertheless take the time to discuss the origin of this operating system and to list some strong elements which explain why Linux prevails in the field of *Data Science*.

1.1.2 Unix-based and *open-source*.

Beyond Linux's monopoly on computing machines, this operating system is certainly the most popular among programmers, especially in the academic world which is involved in Linux from the beginning.

The history of Linux is closely linked to that of another operating system called Unix. Unix was born at the end of the '60s and developped in parallel with the C language. Its designers, Ken Thompson and Dennis Ritchie from the Bell Labs of the telecommunications company AT&T (now Nokia Bell Labs), wanted to create a **multi-user** and **multi-tasking** operating system which can be installed easily on different computer models. At that time, for legal reasons, AT&T could not market Unix and therefore decided to distribute it, in particular to universities, with a rather unrestrictive license. The consequence was a boost in the development of Unix with the University of California at Berkeley being one of the most important contributors to the development of an open-source version of Unix (known as *Berkeley Software Distribution* or BSD).

> A large number of operating systems used today, including MacOS but with the notable exception of Microsoft Windows, are derived from or inspired by Unix. The complete (and impressive!) Unix genealogy can be viewed at `https://www.levenez.com/unix/`.

Until the 1970s, the development of computers and informatics was the prerogative of universities. Ideas and computer codes circulate and are exchanged freely. But this situation evolved negatively, computer programs being seen more and more as marketable products which became protected by more restrictive and prohibitive licenses. Unix is no exception to this trend. In reaction, Richard Stallman decided to create a new and completely open operating system which he named GNU (**G**NU's **N**ot **U** NIX, which is pronounced “gnou”). He makes it similar and compatible with Unix. He therefore adopted the same philosophy which is to use small programs, each doing only one thing (but doing it well) and able to collaborate with each other [@ mcilroy_unix_1978]. GNU only missed one thing: a kernel which is the part of the operating system that manages the hardware. This gap will be filled in 1991 when a Finnish student undertook to develop an operating system for the use of his new computer based on an Intel processor. He developed a new kernel, Linux, and added the GNU tools: GNU/Linux was born.

> As such, Linux is thus not a full featured operating system. It is only a *kernel* containing all the programs used to manage the hardware: memory management, access to disks, to peripherals, etc. It is therefore more accurate to consider the combination of Linux and GNU as the true operating system. This is why we speak of “GNU/Linux”.

1.1.3 The favorite playground of bioinformaticians

Linux supports the majority of programming languages, which makes it an ideal development environment. This is one of the reasons why most bioinformatic programs are available on Linux (sometimes exclusively) and are distributed under an open-source license.

Another important element is that the Linux command line (or Shell), is much superior to that of Windows. We will discover it through this course of which it is the main objective. This was not lost on Microsoft, who decided to integrate a GNU/Linux environment into Windows 10, the *Windows Subsystem for Linux* (WSL) which allows "… to run native Linux command line tools directly on Windows, along with your desktop and traditional Windows applications" (See also Barnes (2021)).

Finally, Linux supports natively the *ssh* communication protocol which makes working in client/server mode, which is the rule in bioinformatics, very simple. Your desktop computer should generally only serve as an access door (terminal) to powerful machines designed for the processing of big data involving resource-intensive calculations, both in storage and in computing power. It is not uncommon for genomic analyzes to take hours, days or weeks for being processed on these machines! No chance about making them on your laptop.

1.2 Installation of the Ubuntu distribution

The open-source and modular nature of GNU/Linux has led to the emergence of a multitude of variants, called distributions, each being a coherent set of softwares, most often open-source, running on top of the Linux kernel. These distributions differ by their user experience, their graphical interface, the system management tools, their update cycle, etc… Some are "general public" oriented while others are specialized for enterprise specific applications.

Ubuntu is one of the most popular GNU/Linux distributions. It was born in 2005 from another distribution, Debian, with the aim of making it an easy to install, user-friendly distribution with a large catalog of regularly updated softwares.

This distribution exists either as a Desktop or a Server, meaning no graphical user interface, version. It is updated twice a year in April and October. In addition to this short life-cycle, every 2 years a version receives the LTS (Long Term Support) label, which has a recent Linux kernel and, as the name suggests, is maintained for an extended period of 5 years. These LTS versions are therefore ideal for a professional use of GNU/Linux.

1.2.1 Downloading of the latest LTS version of Ubuntu and creation of a *bootable* USB drive

An ISO file (disk image) can be downloaded from the Ubuntu website: `https://ubuntu.com/download/desktop`.

1.2.2 Try/Install Ubuntu on your hard drive.

BOX 1: It's your turn> create an installation USB drive

Take a minimum 4GB USB key that can be completely erased. Then follow the procedures described for Windows and for MacOS at the bottom of the download page ("Easy ways to switch to Ubuntu") to create an Ubuntu *bootable* system on this drive.

Configure your computer to use the USB drive as a bootable media. This is usually done by pressing a key during the first few seconds after starting your computer (F12, ESC, F2, F10 or some other key that is usually mentioned briefly on the screen at startup).

When the system boots on the USB key, you can first choose either to install or to try Ubuntu. This last possibility is interesting, especially if you want to test a few distributions before choosing the one that suits you the best. In this case, you will access an Ubuntu session identical to the one you would have after installation except that no changes are made to your system: everything happens on the USB key.

When you choose to install the distribution, you will be guided through a simple and intuitive installation process. Be careful, however, this operation will necessarily modify the structure of your hard drive. **This is not without risk and you should always make sure that you have an up-to-date backup of your data**.

The next screen let you choose the type of installation and other options. I suggest you choose minimal installation and uncheck all other options.

The next step is the most critical since you will decide either to completely erase your disk to install Ubuntu on it, or to install Ubuntu "alongside" a pre-existing operating system. As already mentionned, even in the latter case, your data may be damaged or lost in the process! However, it generally goes very well and allows you to choose, when you start your computer, which system you want to use. In this configuration, called "dual-boot", it is therefore not possible to switch from one system to another without restarting your machine. The advantage is that each system can use 100% of your computer's resources (except the hard drive which is shared during the installation process).

References

Barnes, H. (2021). Pro Windows Subsystem for Linux (WSL): Powerful Tools and Practices for Cross-Platform Development and Collaboration (Berkeley, CA: Apress).