Data analysis on unix/linux systems

Barbera van Schaik

KEBB, Bioinformatics Laboratory
Amsterdam UMC (AMC)

b.d.vanschaik@amsterdamumc.nl

5 March 2020

At the end of today you are able to...

- Download and install a command line program
- Download a dataset
- Run the program
- Investigate and change parameters
- Make data summaries using common linux commands

Linux and bioinformatics

- Powerful software development environment
- Flexibility of command line tools
- Ability to create pipelines
- Many open source software/code available for (re)use

Open source

Video

`http://youtu.be/P_mS4CIXcLY`
Stephen Fry for the 25th birthday of GNU

Free as in freedom

You can use, change, integrate, and review the code
Open source allows sharing and promotes collaboration
No vendor lock-in

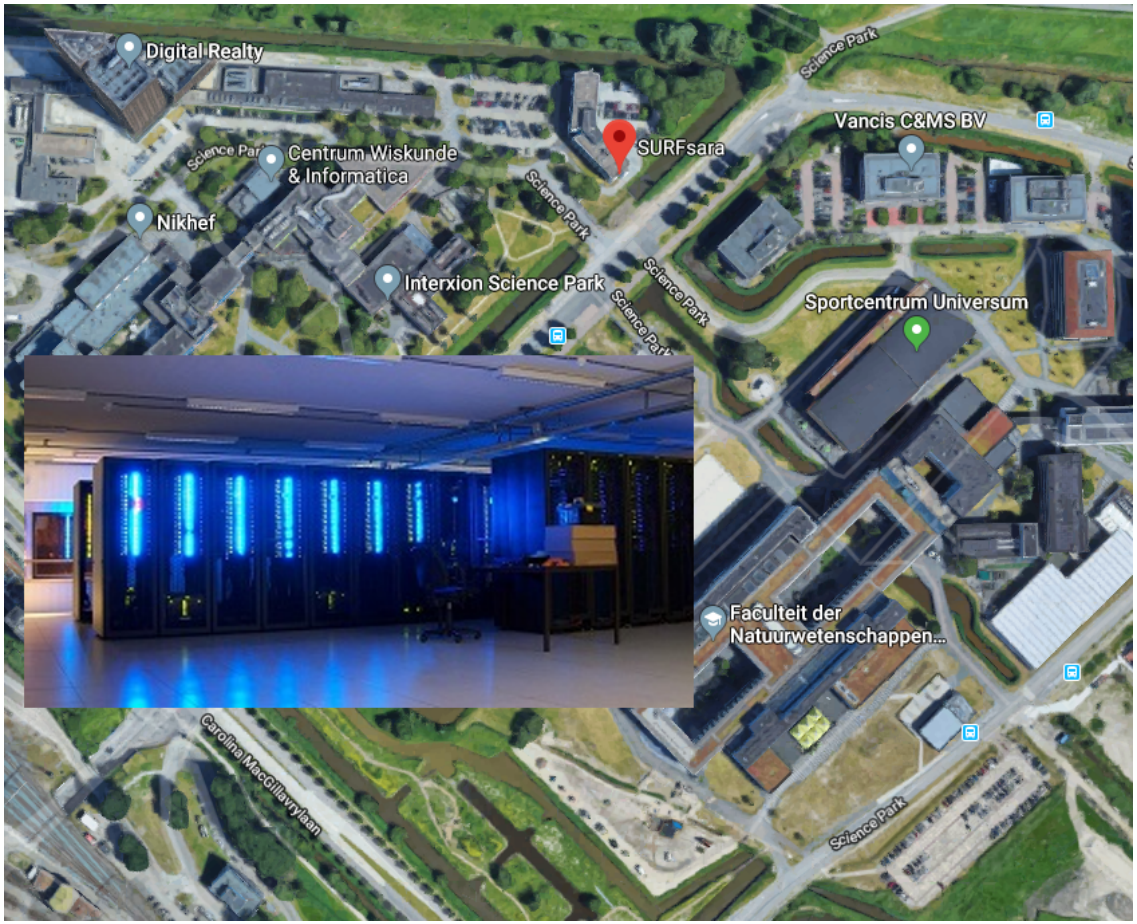
Open source

- Software
- Databases
- Journals
- Standards

- Hardware
- Art
- Money
- Drinks
- Medicine
- Fashion
- Education

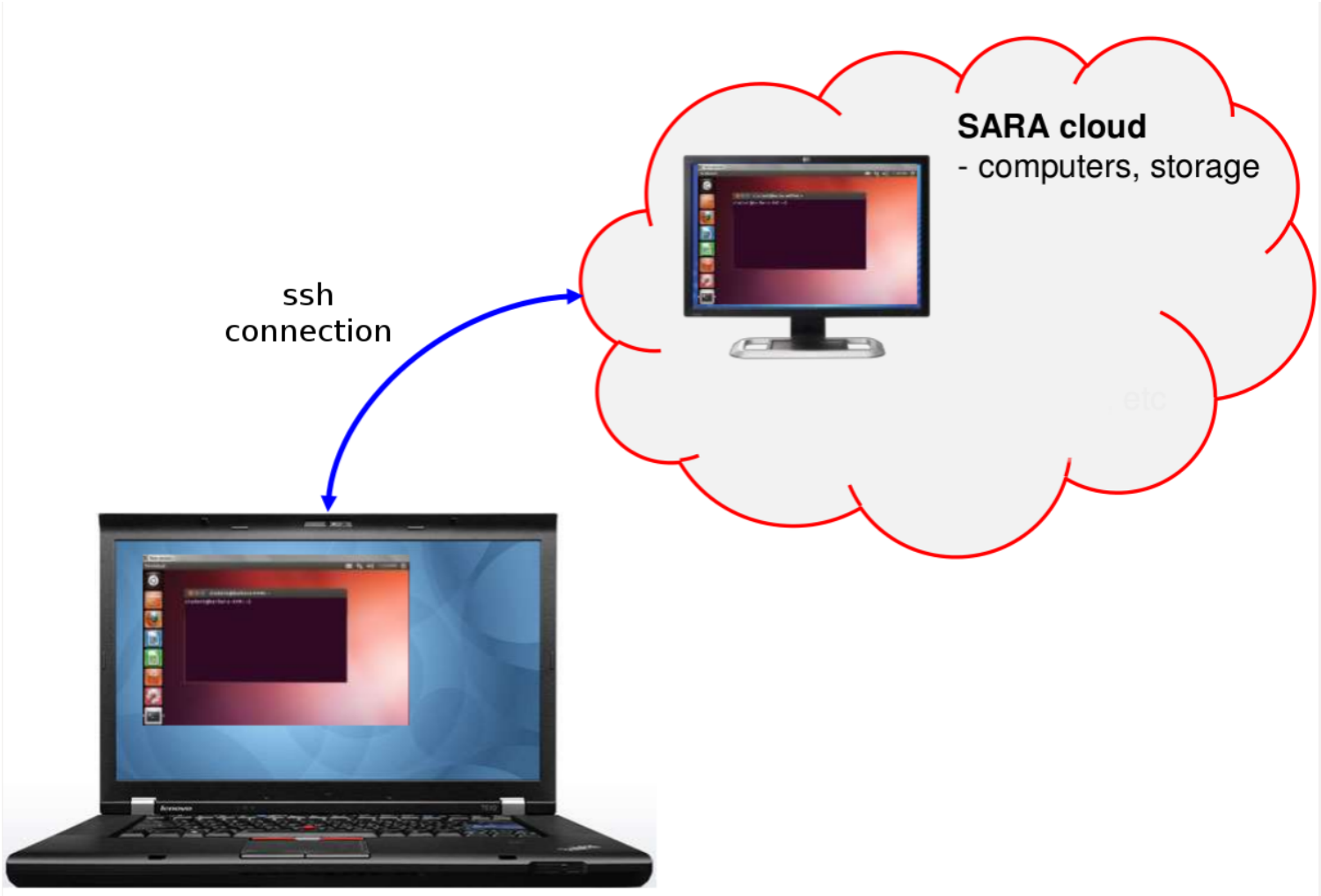
`https://en.wikipedia.org/wiki/Open_source`

Work environment



`http://surfsara.nl/`

SurfSara HPC cloud



`http://surfsara.nl/`

Exercises

Dataset

Metagenomics sample from glacier ice in Germany
16S RNA has been sequenced
Roche 454 dataset

Software

Linux tools to download data and software
Your own program to convert between data formats
Basic Local Alignment Search Tool
Use standard linux tools for summarizing data

A few "warnings"

To take into account when you analyse data

- Run time issues
- Memory limits
- Disk space
- Long run times

HELP

I don't know how this program works

RTM
man command
command –help
command -h
command[enter]

My program gives an error

File names with spaces
Typo
And thousands of other reasons

The internet can help you

Duck duck go is your friend - `https://duckduckgo.com/`
Stack overflow - `https://stackoverflow.com/`
SEQanswers forum - `http://seqanswers.com/`