Bioinformatics analysis of NGS data: approaches and opportunities (command-line tools, commercial software, webservers)

Joakim Skarin

EURL Foodborne Viruses

Swedish Food Agency

- Analyzing millions of sequence reads usually requires several powerful software that performs different tasks (quality trimming, assembly, SNP-calling etc)

- When setting up NGS-analyzing capabilities in your lab, there are different approaches you can choose from

- Hire a bioinformatician and buy a Linux computer?

- Use free webservers?

- Buy expensive commerical software?

- Or do you use all of the above in combination?

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

A general whole genome sequencing workflow
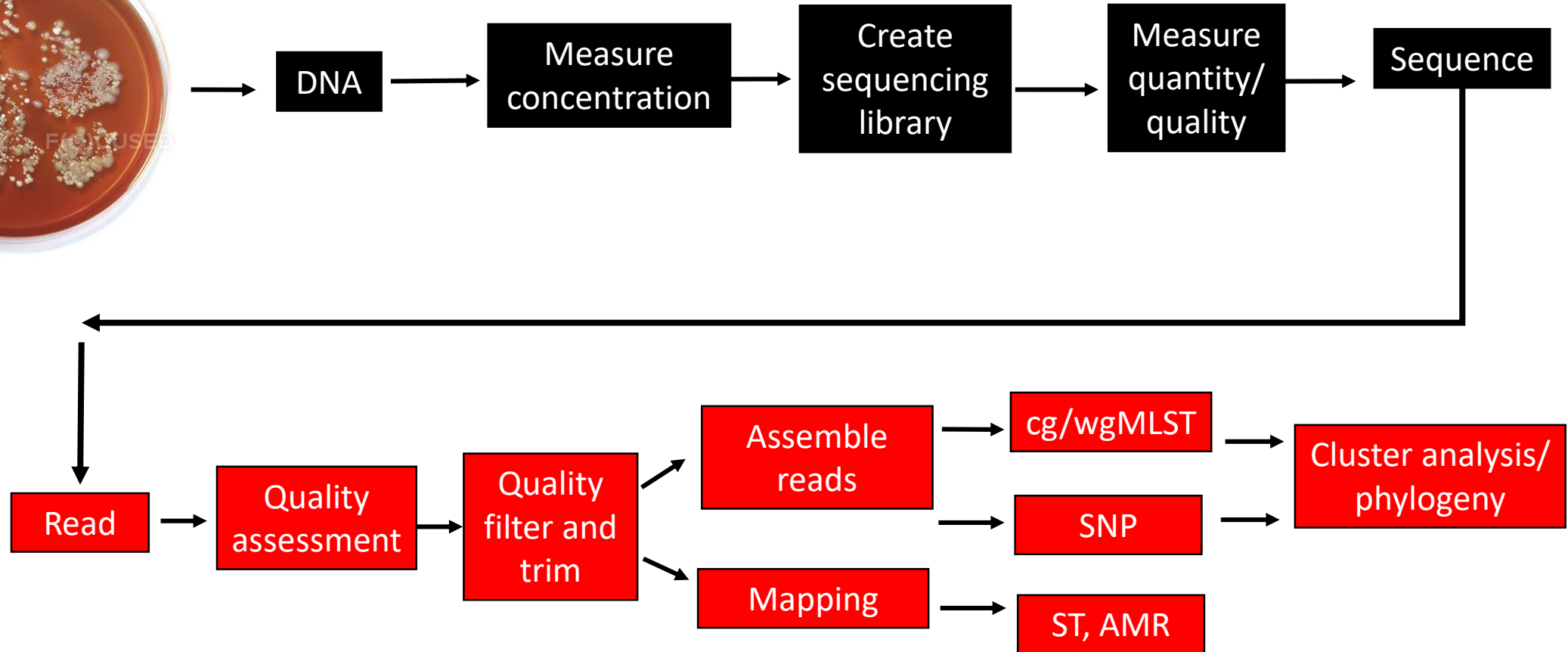
GUI vs CLI

GUI = graphical user interface

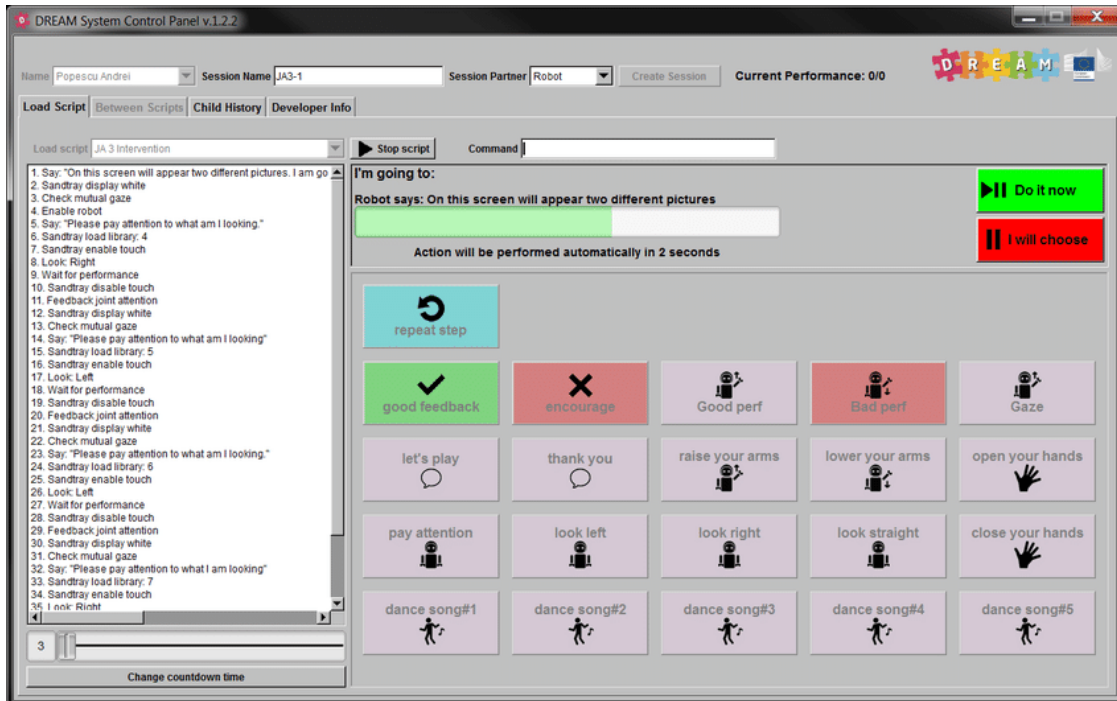GUI is a form of user interface that allows users to interact with electronic devices through graphical icons

CLI = command-line interface

CLI permits users to put in writing commands in terminal or console window to interact with an operating system

Command-line
interface ->>>>

Graphical user
interface
<<<<-

GUI

### Public server

+ No installation
+ Requires little bioinformatics skills
+ Up-to-date nomenclature/DB

- Dependent on service provider
- Downtime of server
- Internet connection
- Can take a long time to get results
- Difficult to analyse large datasets

Commercial software

+ Easy installation
+ Requires little bioinformatics skills
+ Database, easy to use, backup

- Expensive
- Dependent on a company
- Limited to functions in software
- Requires powerful computer

CLI

Local Linux or Mac-computer

+ No cost for software
+ Free to adjust analysis methods/settings
+ Easy to run once setup
+ Often faster
+ Automation

- Requires bioinformatics skills
- Requires computer knowledge to install and create pipelines
- Requires powerful computer

Windows
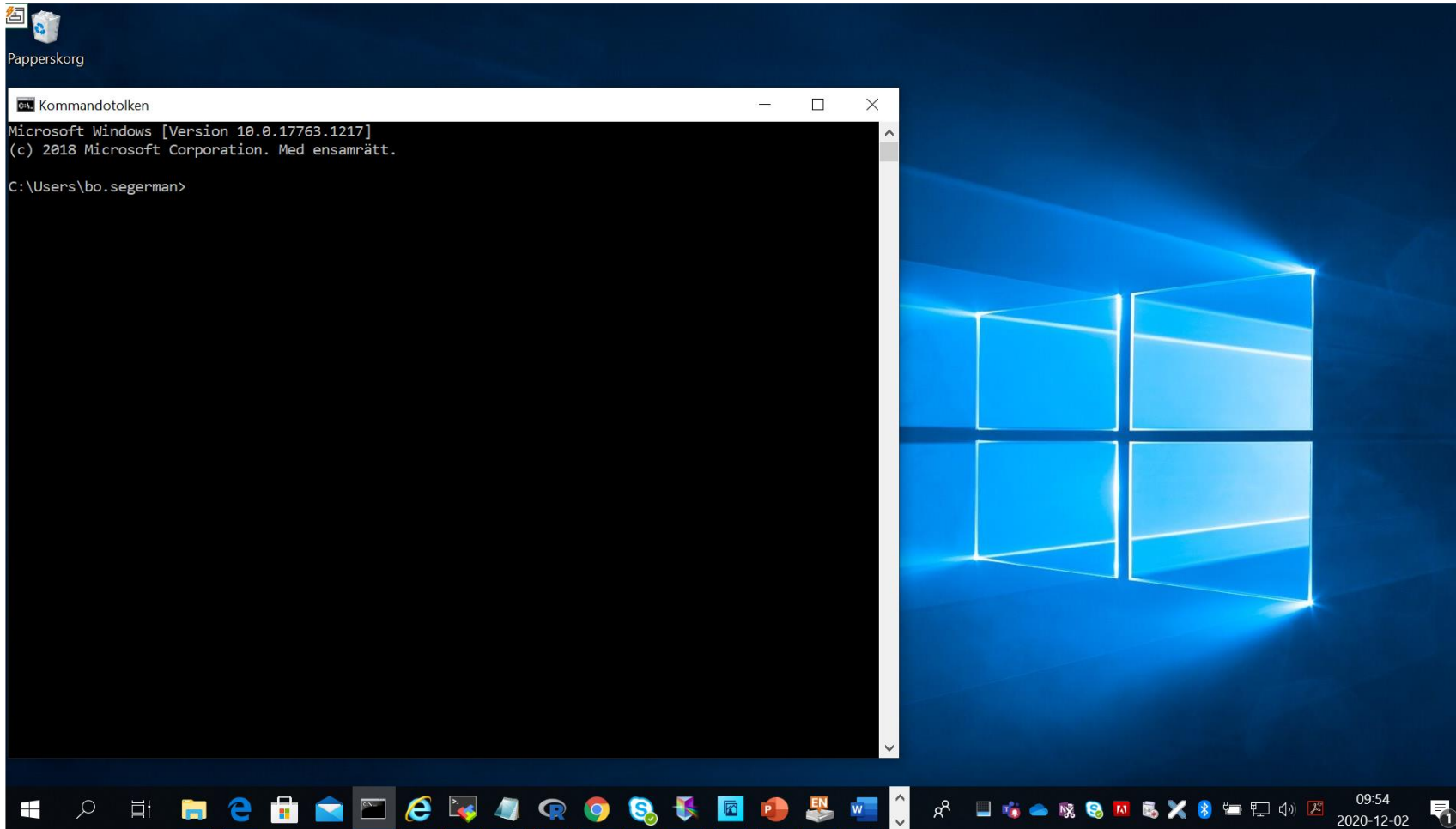
GUI operating system

Command Line interface available
(CMD or powershell)

However, few bioinformatic programs support these

If you use Windows - use a GUI solution (typically commercial) OR
Use a method to access a UNIX compatible system in Windows
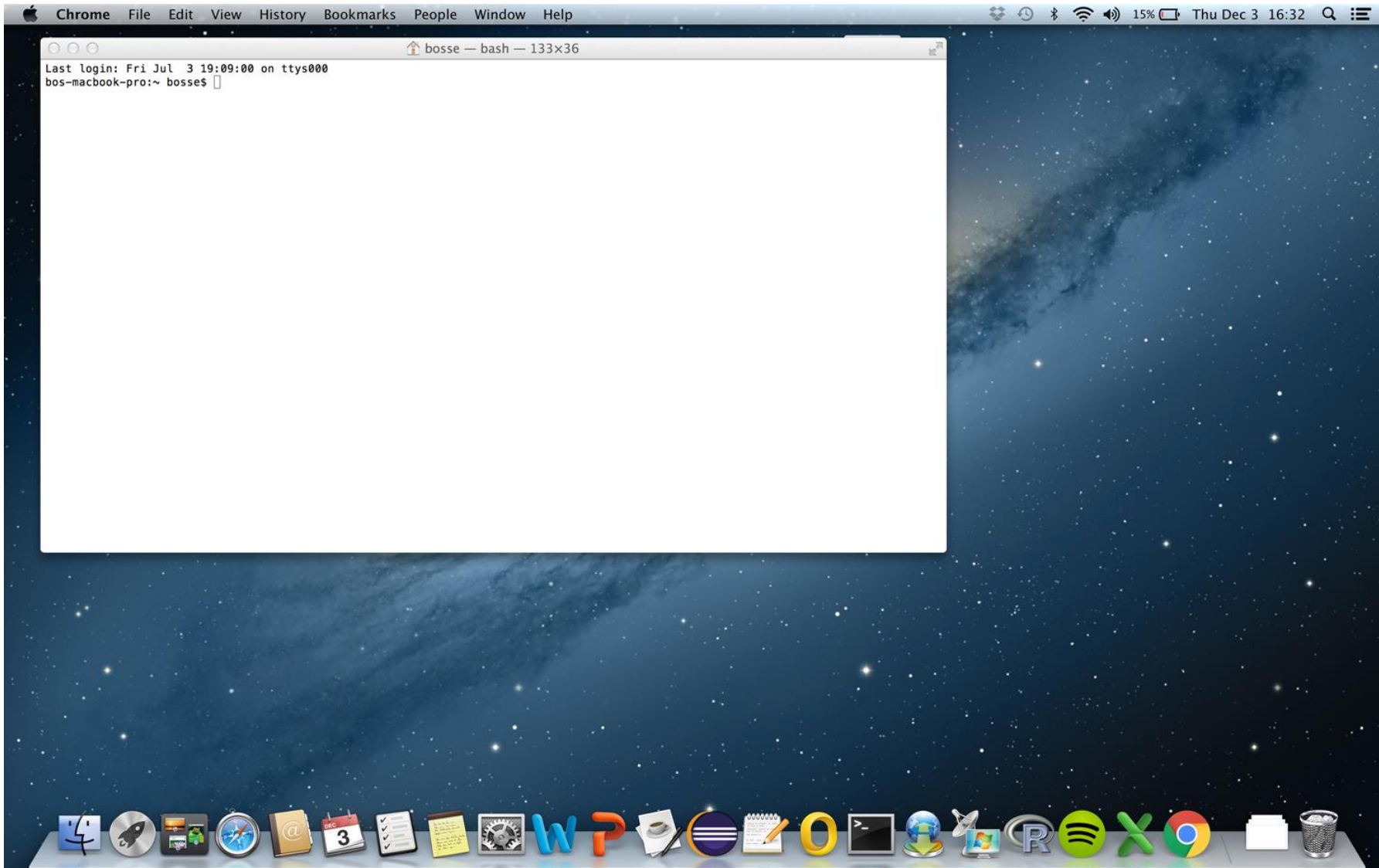
Dual boot (choose Windows or Linux when starting up)

Run a "Virtual" Linux as an app in Windows

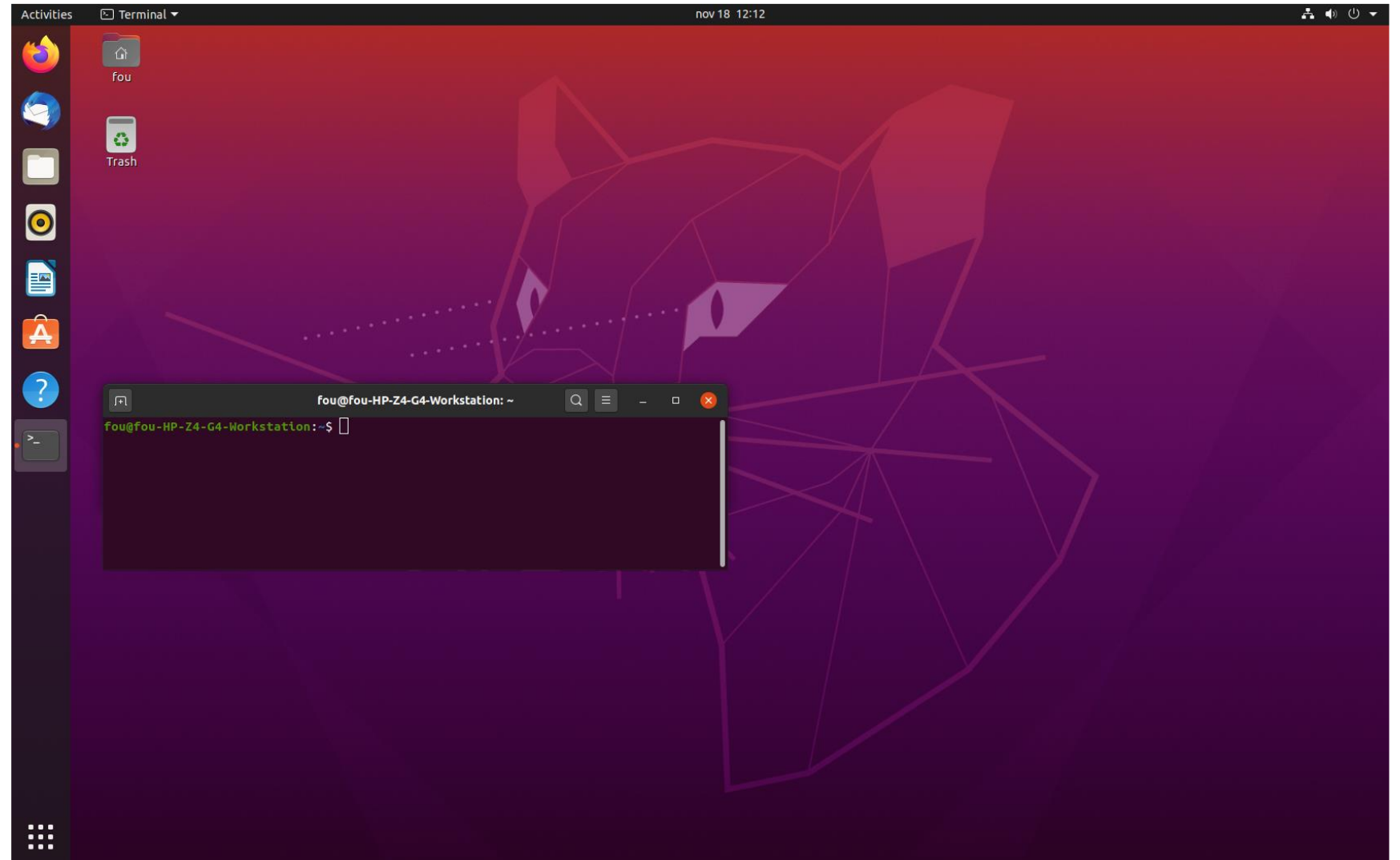**Connect to a Linux computer (server) and run remotely**
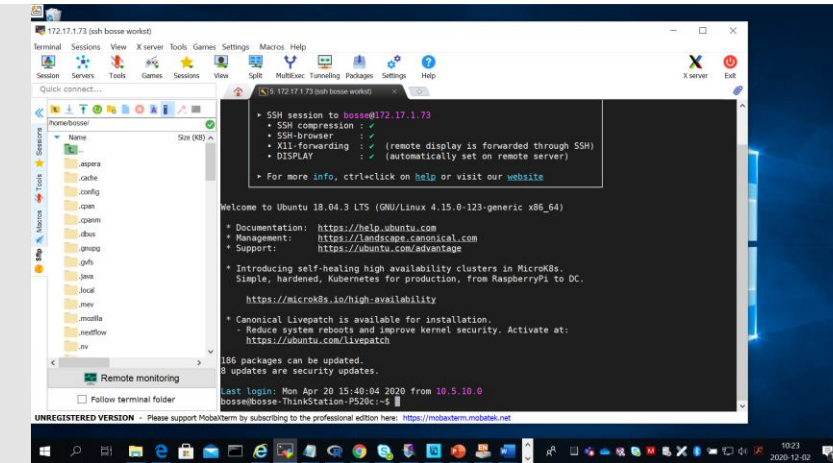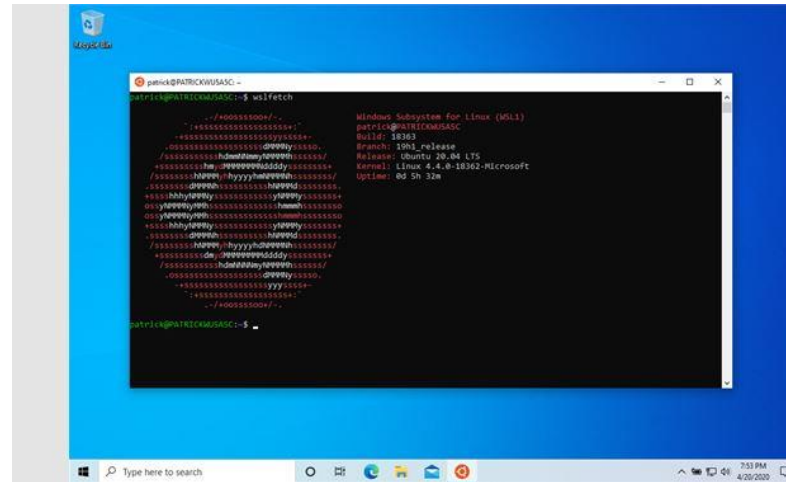
Macintosh

GUI operating system

Modern Mac OS contains a true UNIX compatible terminal

Linux comes in different distributions

- RedHat
- CentOS
- Fedora
- openSUSE
- Debian
- **Ubuntu**

……

The terminal (where you put in commands)
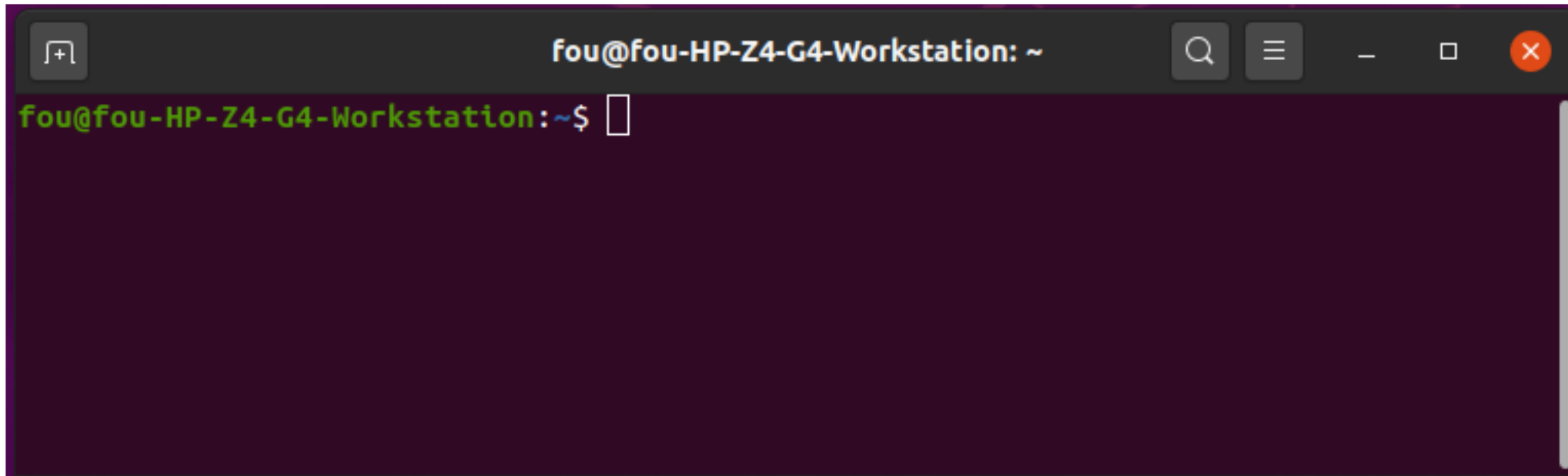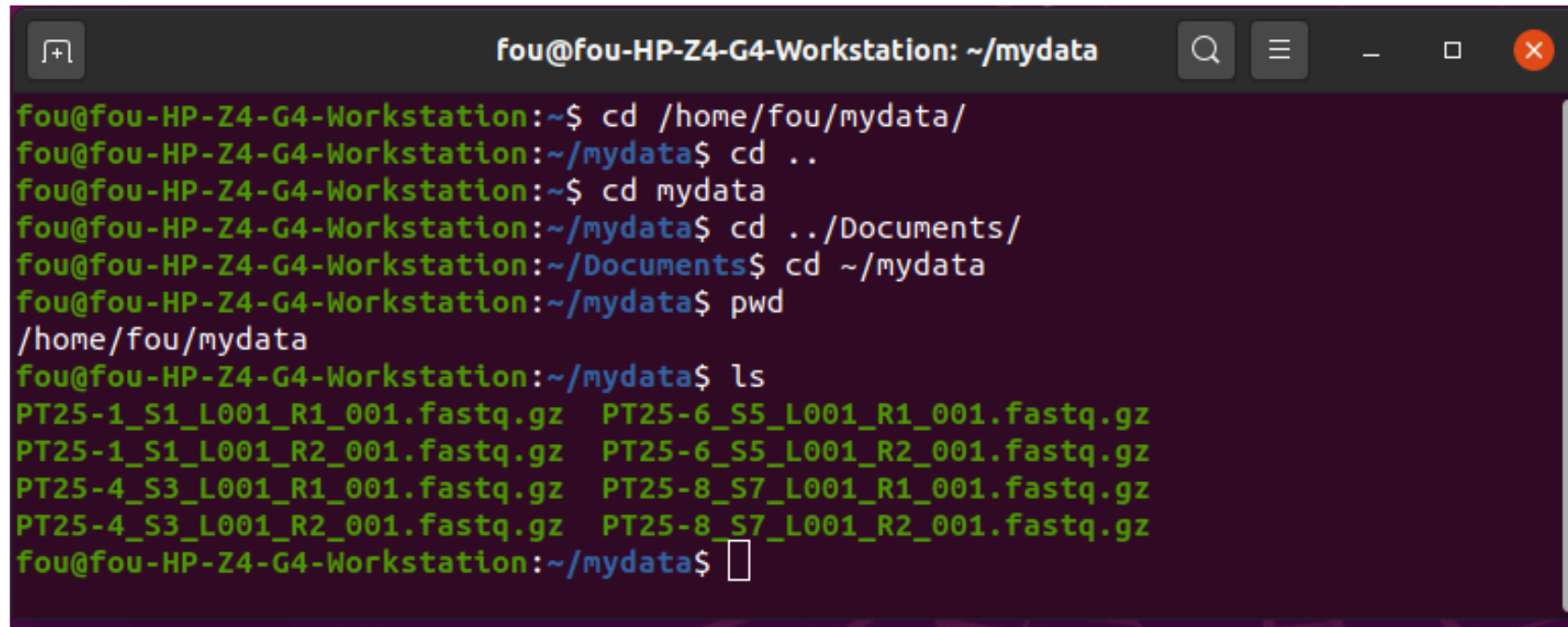
Navigating the filesystem

pwd   print the current working directory
ls   list content of current directory

cd Music   goto directory Music
cd ..   Up one level
cd -   goto previous directory

Why use CLI?

If you use the exact same method all the time, according to a set method available in a commercial software – use that.
However, if you do any kind of development, research and try new methods – use Linux and CLI as well.

Most research articles involving new bioinformatics methods contains a link at the end to a git-repository. This can be accessed via the terminal window and you then have access to their software/method on your computer

- Working with files and text with built-in functions in Linux:

  ➤ If you want to extract the first 1000 reads from a file containing millions of reads:
     head –n 4000 large.fastq > small.fastq

  ➤ Finding a sequence motif in a reads-file:
     grep "ATCGGGC" reads.fastq

Why use CLI?

Pipelines

Start one script/pipeline that performs a series of operations on your data. Saves time and minimizes human errors and hands-on time. Pipeline can be optimized for your specific needs (sequencing technology, analysis settings, reporting etc)

Example of a pipeline structure:

For all samples of a sequencing run, perform:

"Extract 100,000 reads | quality-trim | assemble | calculate assembly metrics | typing analysis | write report"

Since they are difficult to create – we should all decide on and share pipelines to harmonize our capabilities!
(Still requires computer know-how to get them going and maintaining them)

Commercial software:
CLC Genomics Workbench

An all-round NGS-analysis software.
Does "everything"

Commercial software:
Ridom SeqSphere+

Focus on typing of
bacterial genomes
cgMLST and wgMLST

Assembly, mapping
and quality control of
data

Commercial software:
Bionumerics

Specializes in typing applications. Will however be discontinued. A totally new software is planned

Commercial software: **Geneious Prime**

Not specific for typing. One of the more affordable solutions for working with sequence data in general

Public server:
Galaxy

User-friendly service where many (most) bioinformatics operations and programs can be used and put into workflows.

Start with raw reads and design own workflow

You can share your workflow with other users

Public server: CGE

- Species identification
- VirulenceFinder
- ResFinder
- MLST
- Etc

Public server:
Via EURLs
- **Aries (E. coli)**
- **Starflow (Listeria)**

Organism-specific
WGS analysis servers

Istituto Superiore di Sanita'

ARIES - Advanced Research Infrastructure for Experimentation in Genomics - Galaxy Instance at ISS

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology

Public server:
Illumina BaseSpace
Cloud-based NGS
analysis service.

Free to use basic
functions but requires
(?) paid subscription
to access everything
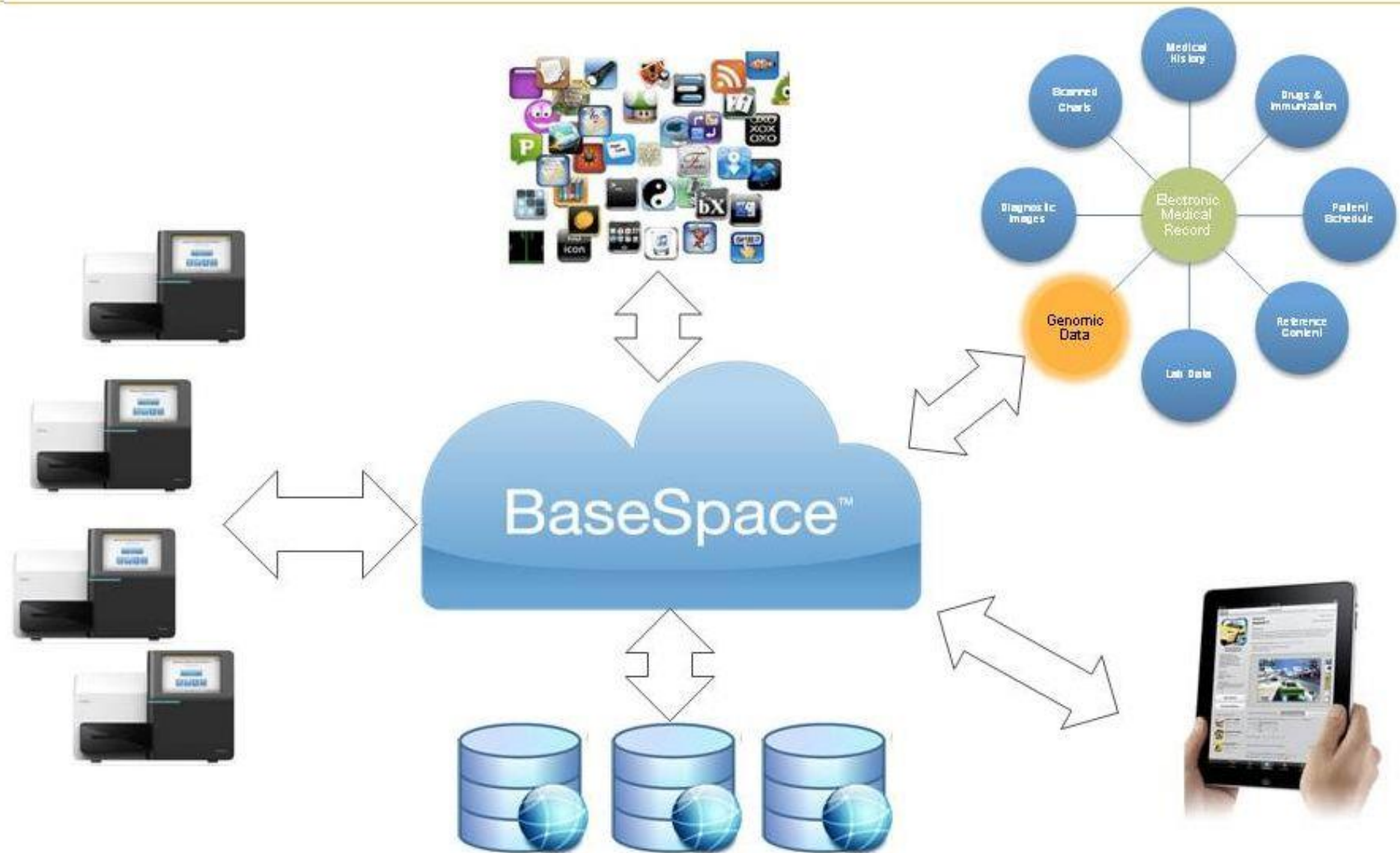
Example – what we use at the Swedish Food Agency

Windows computer
- Commercial software for cgMLST, used for outbreak investigations and surveillance

- Linux app that controls a pipeline for: QC, trim, assembly, serotyping, SNP-analysis etc.

- Online servers used to quickly determine a sequence type or AMR or species

Linux computer
- CLI-workflows written in Python for metagenomics, assembly, QC, contamination checks and SNP-typing etc

- Powerful enough to handle large metagenomics datasets

Windows laptops
- Commercial software for general NGS analysis

Questions?