

# Bioinformatics Introduction

Sebastian Schmeier

s.schmeier@gmail.com

<http://sschmeier.github.io/bioinf-workshop/>

03.08.2015



# Overview

- Bioinformatics
- Big data
- Command line interface
- Linux
- Virtual machines



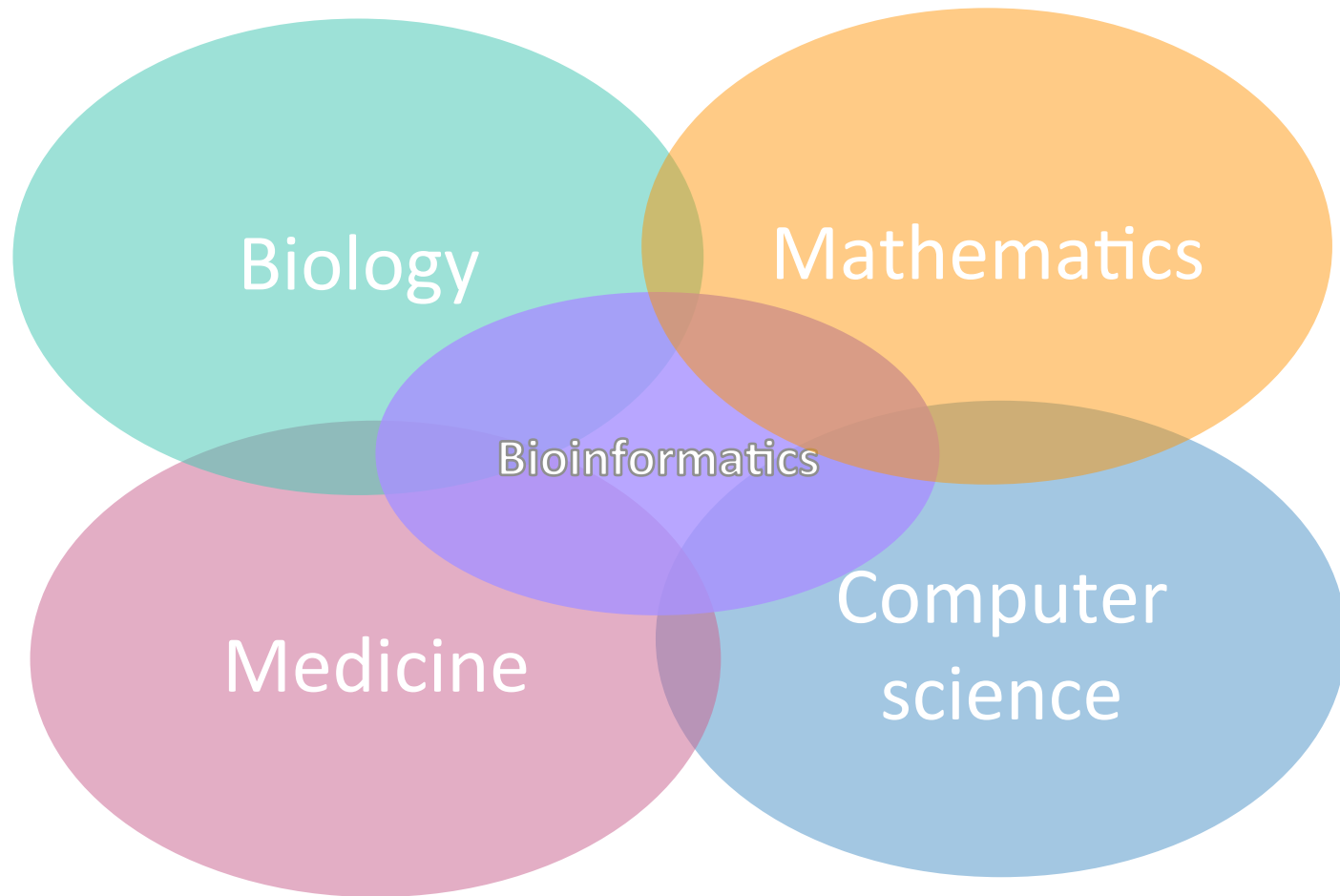
# Bioinformatics

- From Wikipedia

“Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to study and process biological data”



# Bioinformatics



A multifaceted discipline



# Bioinformatics

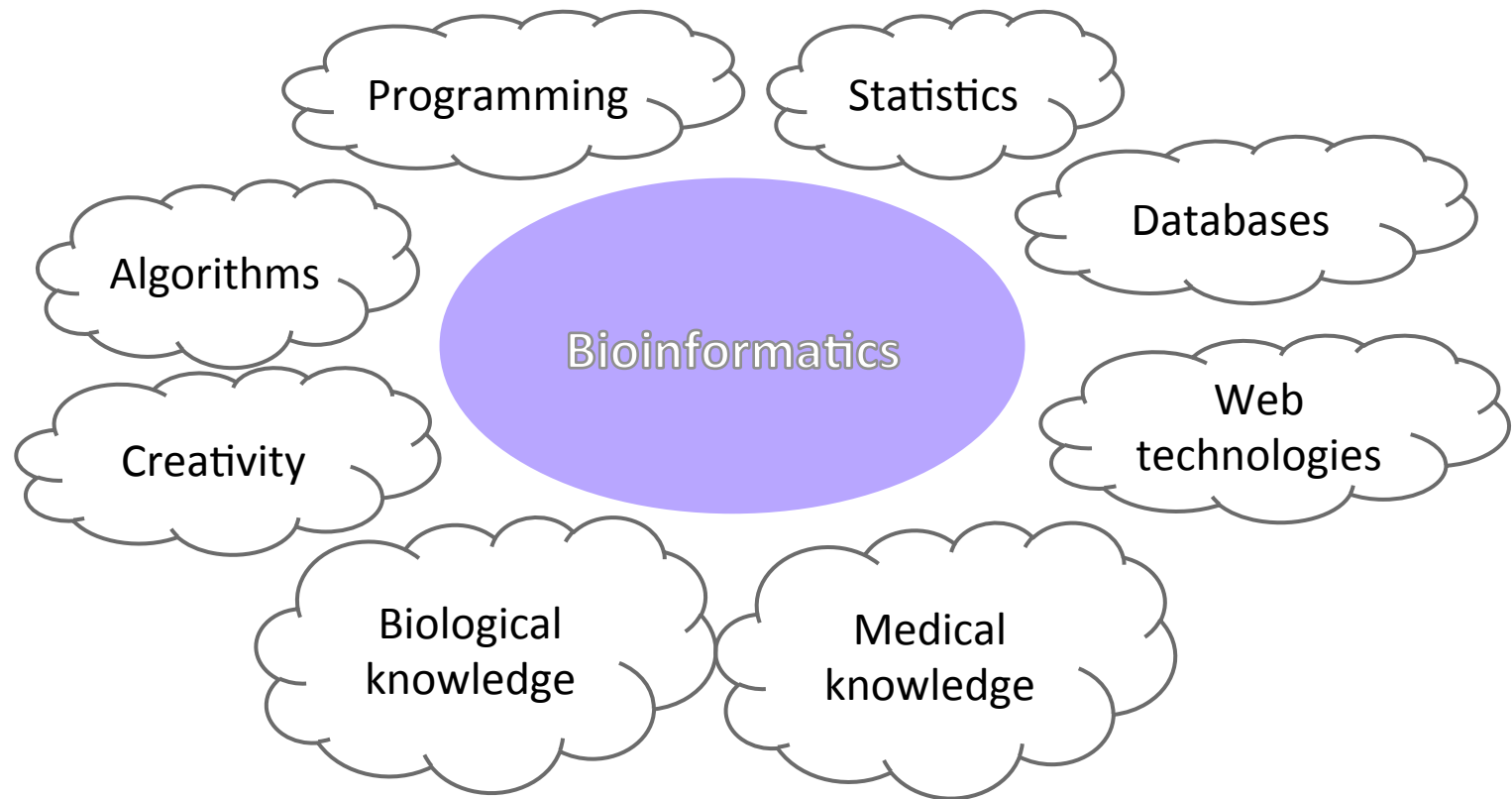
**Biologists + Clinicians**  
collect molecular data:  
DNA & Protein sequences,  
gene expression, mutations, etc

**Computer scientists**  
(+Mathematicians, Statisticians, etc.)  
Develop tools, software, algorithms  
to store and analyze the data.

**Bioinformaticians**  
Study biological questions by  
analyzing molecular data with the  
help of computers



# Bioinformatics bag of tricks

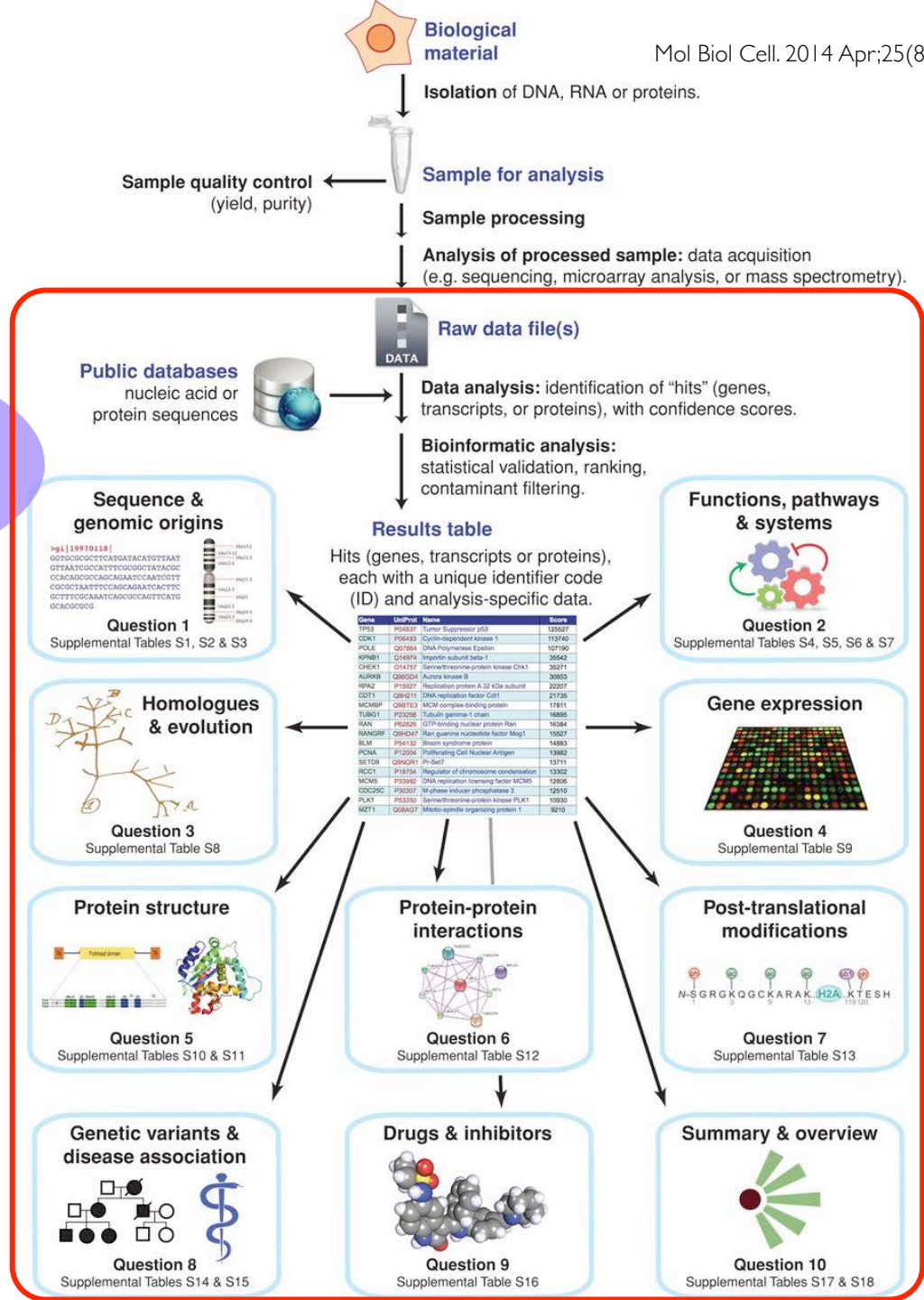




# Bioinformatics and genome research?

## Bioinformatics

Mol Biol Cell. 2014 Apr;25(8):1187-201





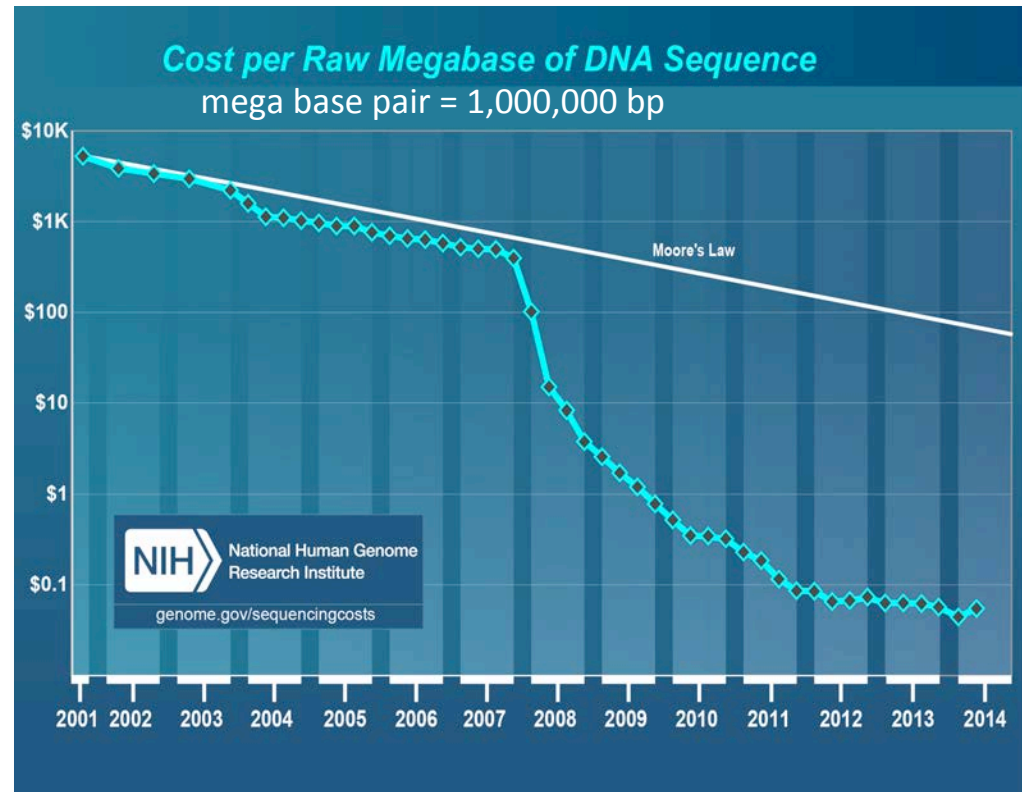
# Big data?

- Data production in biology and medicine has outpaced our ability to analyse the incoming data → e.g. sequencing
- Problems:
  - Data storage
  - Data transfer
  - Data security and privacy...and we did not talk about the actual analysis yet.



# Big data?

- Some examples:
  - 2000: 1 genome  
→ Human genome project
  - 2008~2012: 1,000 genomes  
→ 1,000 genome project
  - 2008~?: 25,000 genomes  
→ Int.l cancer genome project
  - 2012~2017: 100,000 genomes  
→ UK 100,000 genome project
    - 10 Petabyte of data  
(10,000,000 GB, at 100GB per human genome)





# Analyzing “big” data

- Currently Microsoft Excel has a row limit of about ~1 million rows

Version	Max. rows	Max. columns
Excel 2013	1,048,576	16,384
Excel 2010	1,048,576	16,384
Excel 2007	1,048,576	16,384
Excel 2003	65,536	256
Excel 2002 (XP)	65,536	256
Excel 2000	65,536	256
Excel 97	65,536	256
Excel 95	16,384	256
Excel 5	16,384	256

→ Excel is not suitable to work with large files



# Alternative: plain text-files

- What is a plain text-file
  - A file that is human readable
  - Readable as textual material without much processing → can be opened in any ordinary text editor
  - NO style information is included
  - NOT "binary files" in which some portions must be interpreted as binary objects (encoded integers, real numbers, images, etc.).
  - Its size is only limited by the operating systems file-system.



# Common file-systems and file-size limits

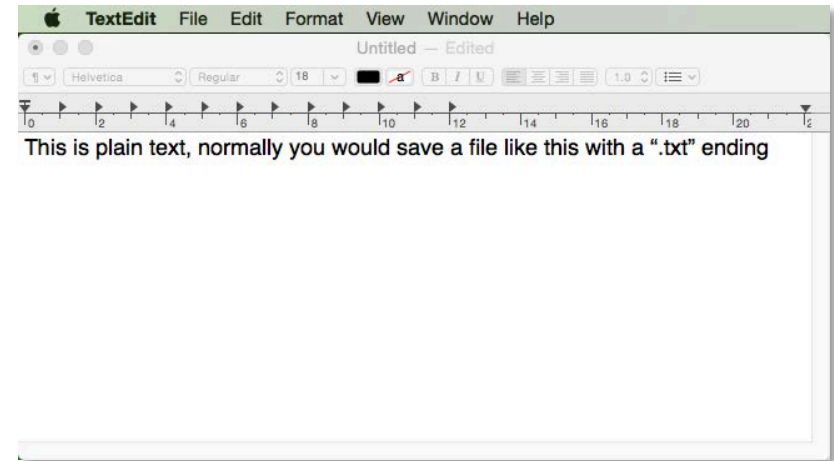
File-system	Max. file-size	Supported OS
FAT32	4 GB (Gigabyte)	Windows, Linux, Mac OSX
NTFS	~16 EB (Exabyte)	Windows, Mac OSX
exFAT	~16 EB (Exabyte)	Windows, Linux, Mac OSX
ext4	~16 TB (Terabyte)	Linux
HFS+	8 EB (Exabyte)	Mac OSX

Value	Name	
1000	kB	kilobyte
1000 <sup>2</sup>	MB	megabyte
1000 <sup>3</sup>	GB	gigabyte
1000 <sup>4</sup>	TB	terabyte
1000 <sup>5</sup>	PB	petabyte
1000 <sup>6</sup>	EB	exabyte



# Working with plain text files

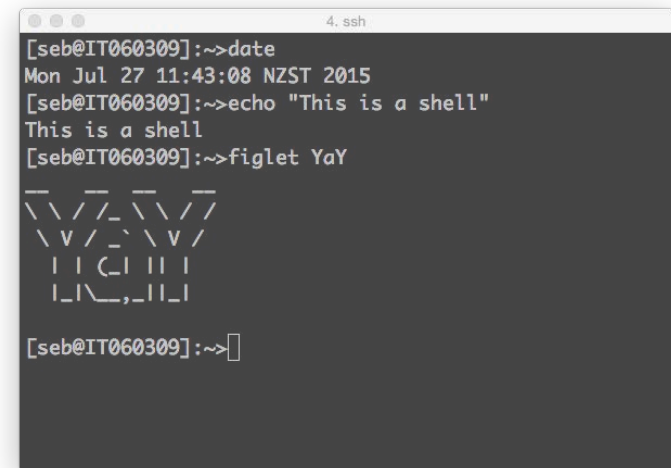
- Generally speaking you can open plain text in any text editor, e.g. TextEdit (Mac)
- However, you are limited in the amount of memory your computer has
- Thus, large files **cannot** be opened in a text editor





# Command-line interface (CLI)

- Many data scientist tend to work with a command-line interface on a terminal window
- Here, one can interact with the operating system by issuing commands in form of successive lines of text (**command lines**)

A screenshot of a terminal window titled "4. ssh". The prompt is "[seb@IT060309]:~>". The user enters "date", and the output is "Mon Jul 27 11:43:08 NZST 2015". The user enters "echo \"This is a shell\"", and the output is "This is a shell". The user enters "figlet YaY", and the output is a stylized ASCII art of the text "YaY". The prompt is now "[seb@IT060309]:~>".

```
[seb@IT060309]:~>date
Mon Jul 27 11:43:08 NZST 2015
[seb@IT060309]:~>echo "This is a shell"
This is a shell
[seb@IT060309]:~>figlet YaY
  _/ _/ _/ _/
 \v / _ \v /
  | | C | | |
  | | \, _ | |

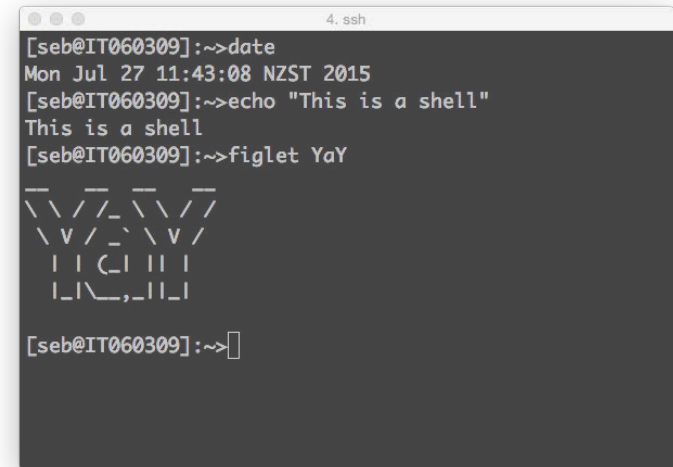
[seb@IT060309]:~>
```

Who has heard of the command-line?  
Anyone tried to work with the command-line?



# Advantages of the CLI

- **Control**
  - Possibility to access the operating system file-system and programs
- **Speed**
  - Processing data only with the keyboard
  - No interaction with menus, etc. necessary
- **Saves resources**
  - Terminal does not require much computer resources
  - able to run on many kinds of hardware
- **Able to script**
  - Able to write most work processes in forms of small scripts
  - This makes it even speedier and also
- **Remote access**
  - Processing data from different computers over the network is simple
- **Reproducibility**
  - Placing the commands in a script or text-file lets one see/reproduce all steps taken to get to the result
- **Less strain**
  - Less switching between keyboard and mouse

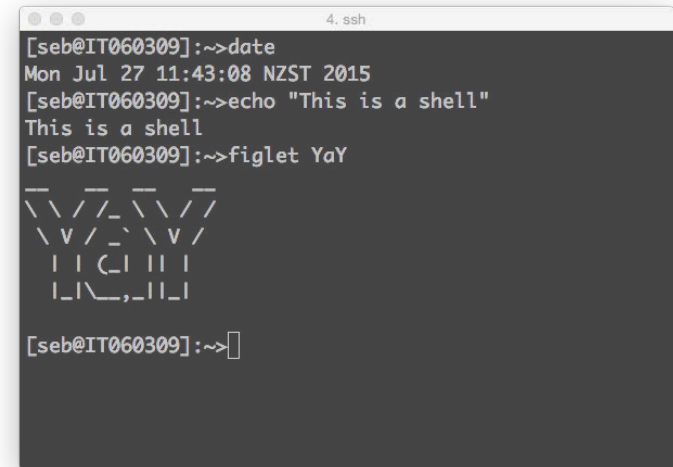


```
4. ssh
[seb@IT060309]:~>date
Mon Jul 27 11:43:08 NZST 2015
[seb@IT060309]:~>echo "This is a shell"
This is a shell
[seb@IT060309]:~>figlet YaY
  _/ _/ _/ _/ _/
  \V / _\ \V /
  | | C | | |
  | | \, _ | |
```



# Disadvantages of the CLI

- No fancy user interface
  - Everything happens in the terminal window
- Not able to navigate with mouse
  - One needs to input command-lines one by one
- Needs to remember important commands
  - No menus available to hunt through to find the right commands



```
4. ssh
[seb@IT060309]:~>date
Mon Jul 27 11:43:08 NZST 2015
[seb@IT060309]:~>echo "This is a shell"
This is a shell
[seb@IT060309]:~>figlet YaY
  _/ _/ _/ _/
  \V / _\ \V /
  | | C | | |
  | | \, _ | |

[seb@IT060309]:~>
```



# Where can I use the CLI

- **Windows** has a very rudimentary CLI (shell)
  - not very user friendly, many essential programs are missing
- **Mac OSX** has a Terminal program
  - needs some tinkering to make it fully usable
- **Linux** in-built terminal
  - Most comprehensive set of CLI programs available

Who is working on Windows?

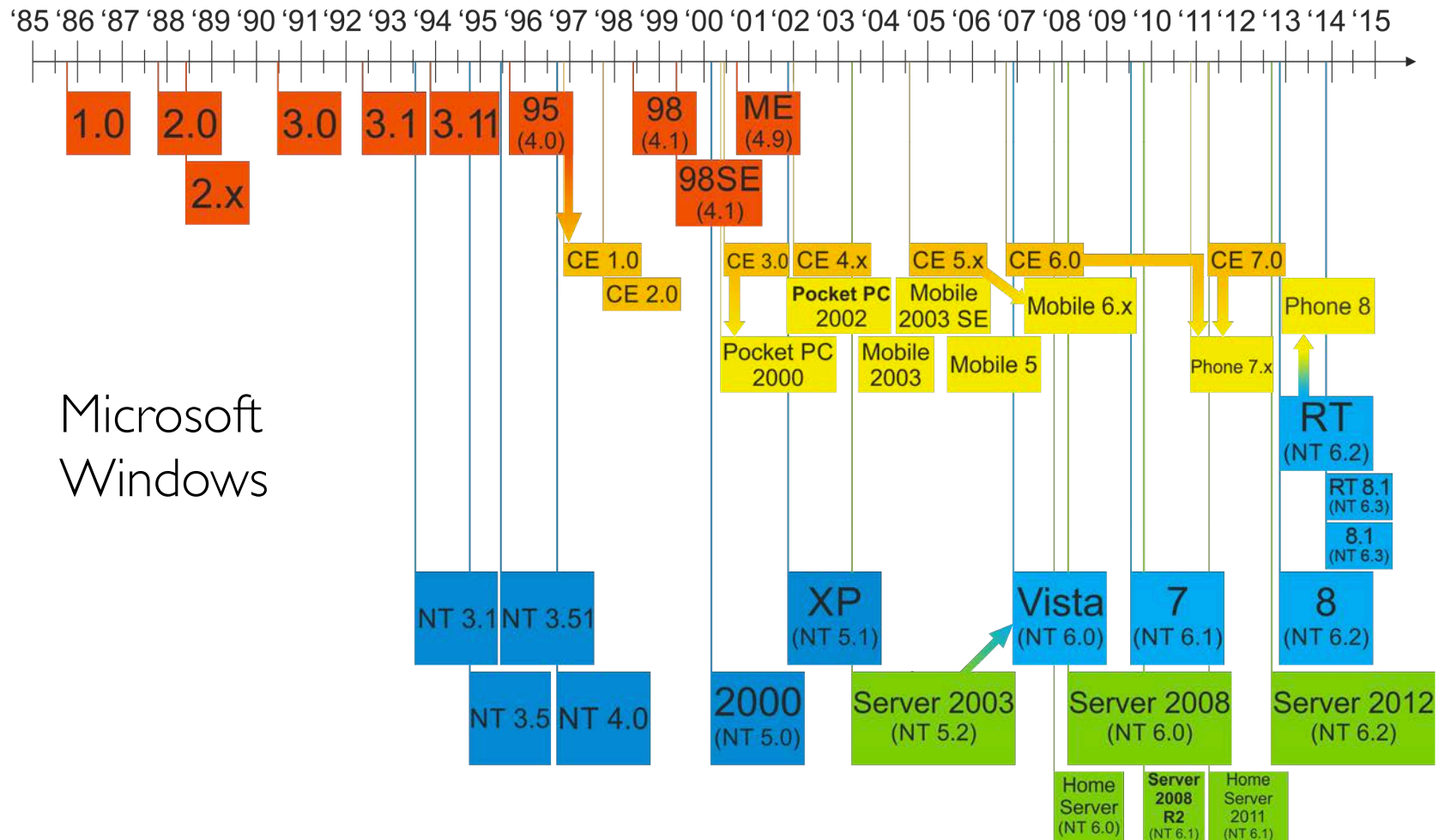
Who is working on Mac OSX?

Who has heard of Linux?

Who has worked with Linux?



# Evolution of computer operating systems



[https://upload.wikimedia.org/wikipedia/commons/6/6d/Windows\\_Updated\\_Family\\_Tree.png](https://upload.wikimedia.org/wikipedia/commons/6/6d/Windows_Updated_Family_Tree.png)



C:\>command

Microsoft(R) MS-DOS(R) Version 4.01  
(C)Copyright Microsoft Corp 1981-1988

C:\>ver

MS-DOS Version 4.01

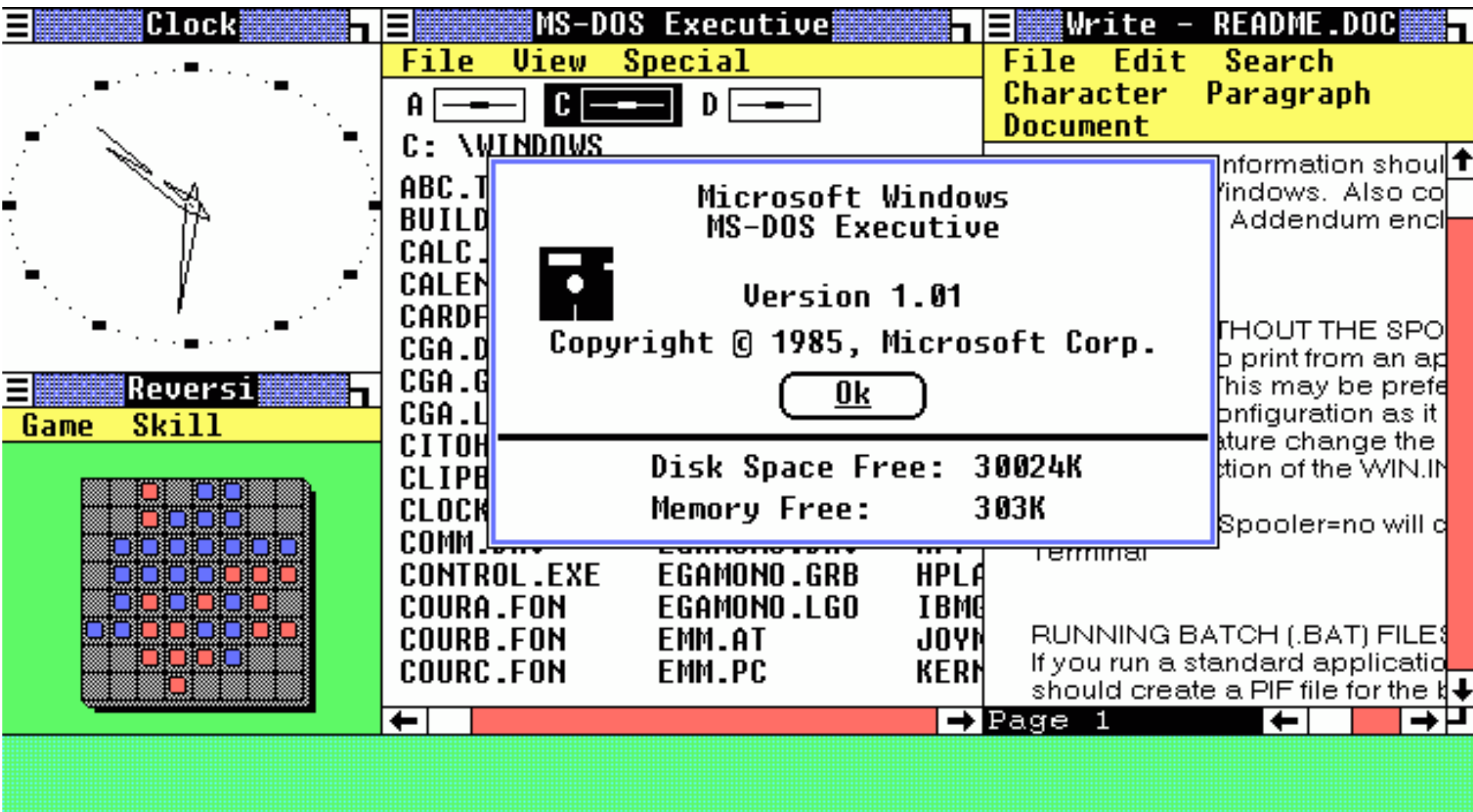
C:\>dir command.com

Volume in drive C is DOS  
Volume Serial Number is 2432-07DC  
Directory of C:\

COMMAND	COM	37557	12-19-88	12:00a
1 File(s)		495624192 bytes free		

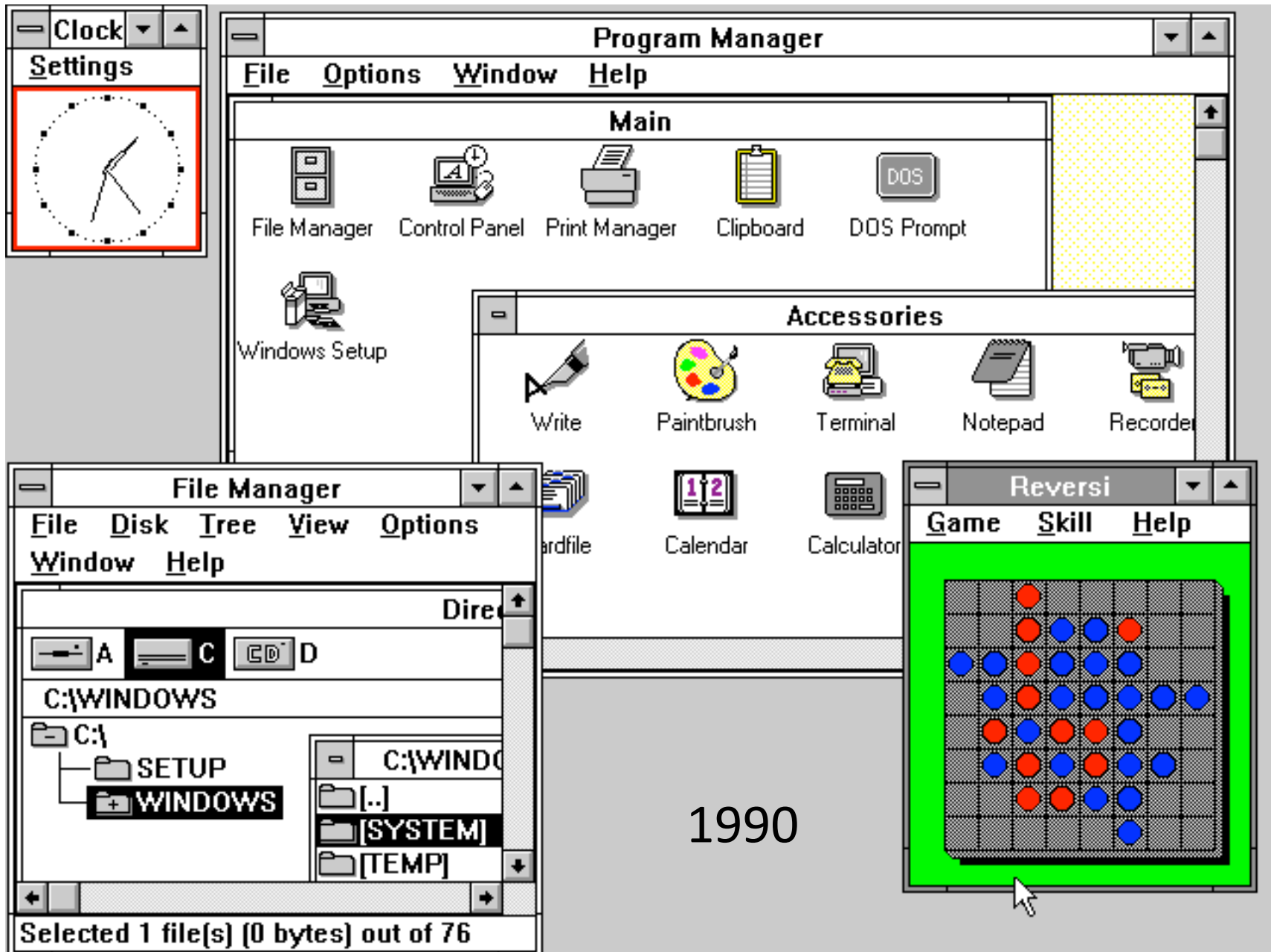
C:\>\_





[https://en.wikipedia.org/wiki/Windows\\_1.0](https://en.wikipedia.org/wiki/Windows_1.0)

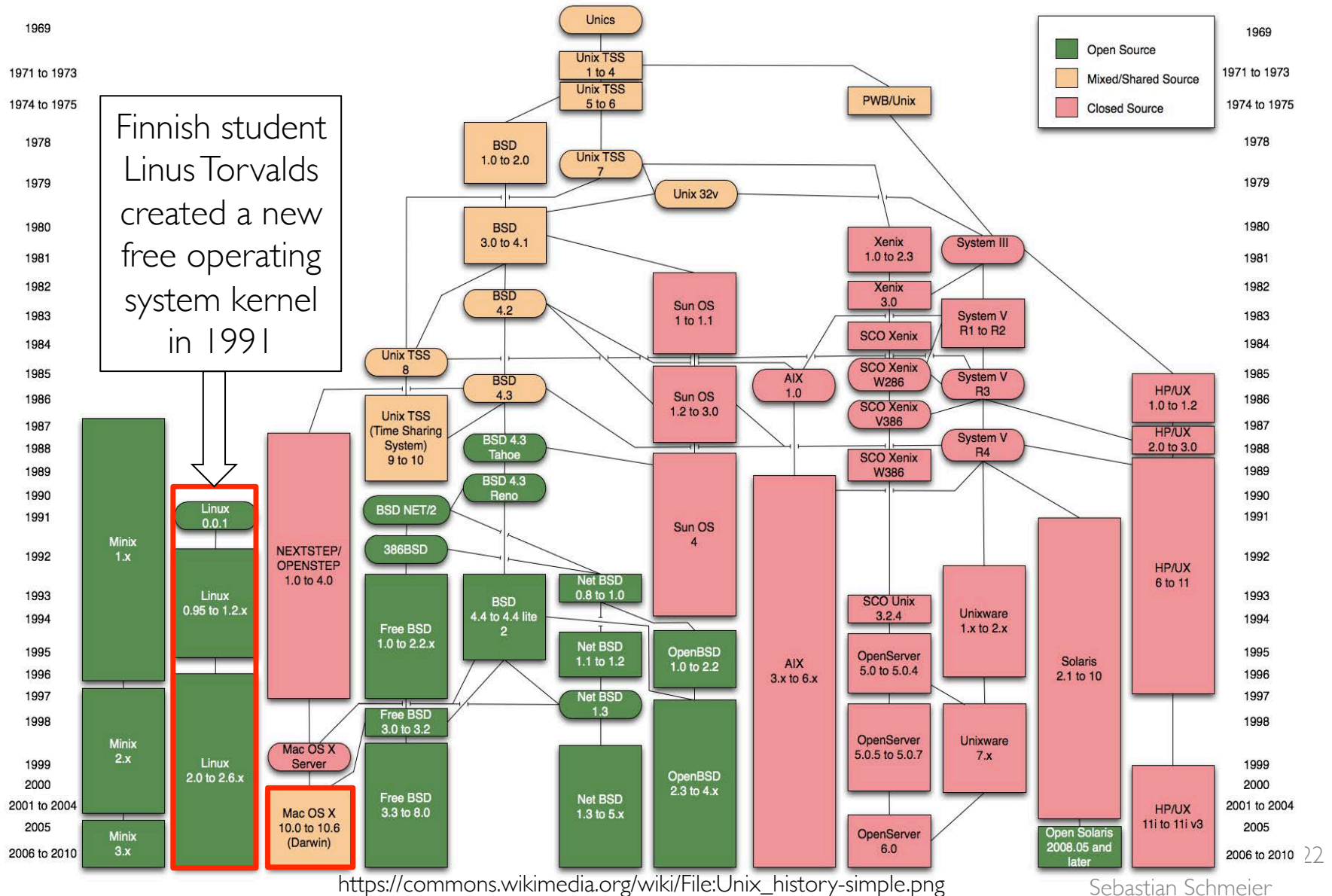




1990



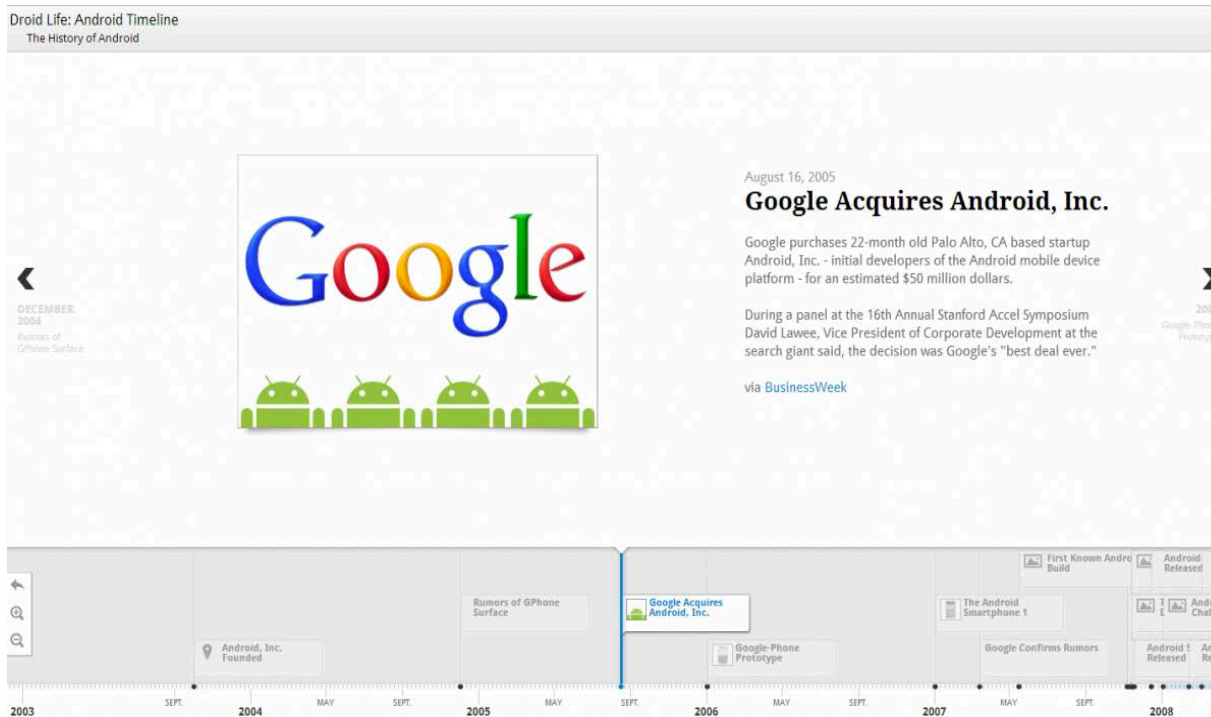
# Evolution of computer operating systems





# Yes! Android as well

- 2005: Google acquired Android Inc. for at least \$50 million
- 2013: Google's **Linux-based Android** claims **75%** of the smartphone market share, in terms of the #phones shipped.







# Advantages of Linux-based computers

- Free
- Open-source
- Customizable
- Alternatives/Choice
- Security
- Huge community
- Being able to contribute
- Low on resource / also good for old hardware
- ...



# Disadvantages of Linux-based computers



- Won't run any Windows programs natively e.g. Word, Excel, Powerpoint, Photoshop etc.
- For some peripheral hardware there are no Linux drivers available
- Requires some getting used to for first time users
- There is no “standard Linux” edition
- Support is sometimes harder to come by and some technical background is sometimes required to understand solutions
- Gaming is not YET as easily done on a Linux computer





# Most popular Linux distributions

Ubuntu 

openSUSE 

Mint 

ArchLinux 

Debian 

Fedora 

Mageia 

Many more distributions available.  
Some distributions build on other  
distributions and enhance them  
with certain feature sets.



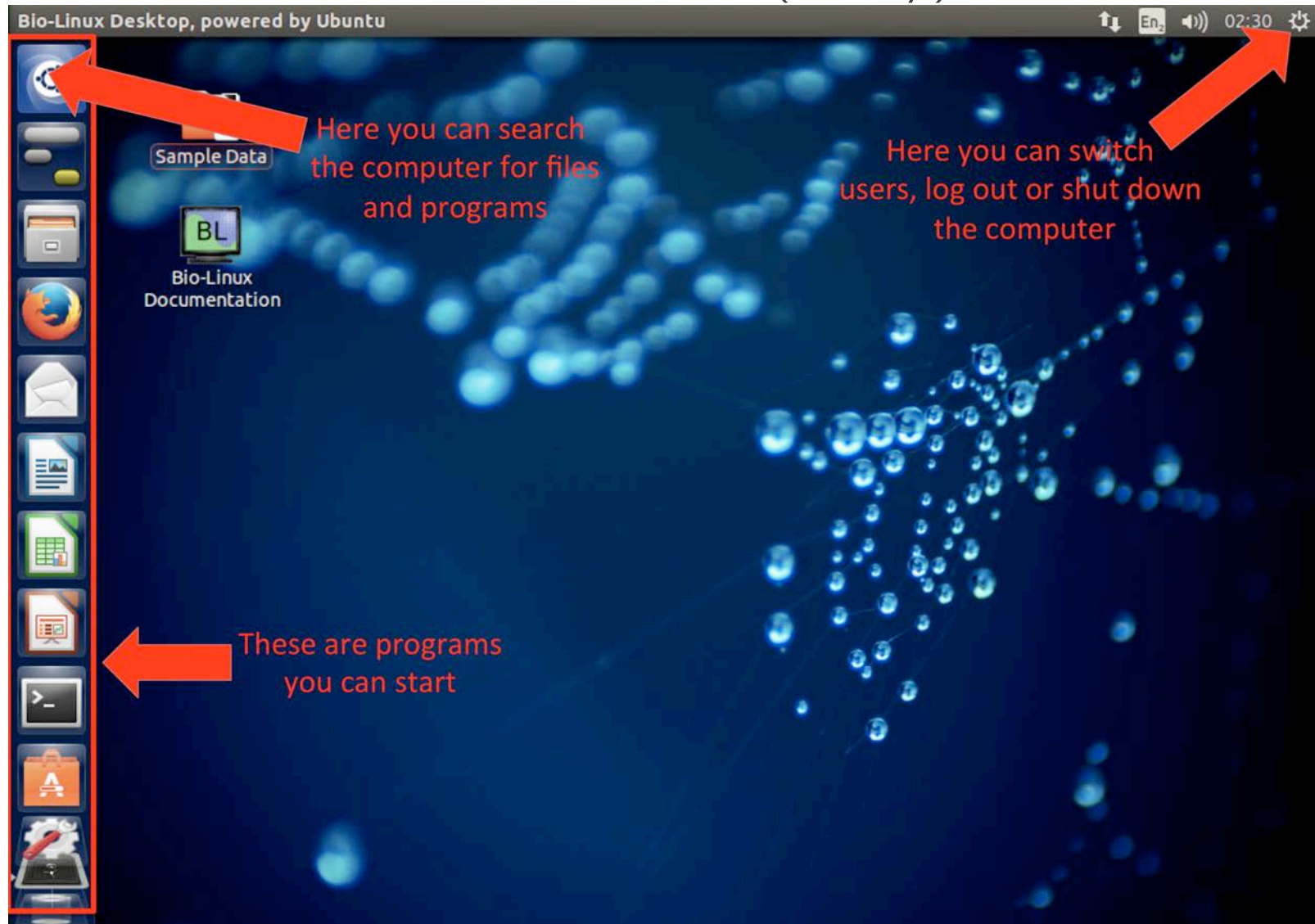


- Based on Ubuntu Linux
- Powerful, free bioinformatics workstation platform
- can be installed on anything from a laptop to a large server, or run as a virtual machine.
- Adds more than 250 bioinformatics packages to an Ubuntu Linux 14.04 LTS base, providing around 50 graphical applications and several hundred command line tools.
- Incorporates the Galaxy environment for browser-based data analysis and workflow construction

Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N. and Thurston, M. 2006. Open Software for biologists: from famine to feast. *Nature Biotechnology* 24, 801–803.



# The BioLinux user interface (Unity)





# How to run Linux

- There are several options available:

1. **Install it on a computer:**

Either instead of your original OS or along side your OS (called dual boot)

2. **Run Linux of a LiveCD:**

This requires you to boot from a DVD that contains the Linux Live version. However, once you shutdown the Live version of Linux all your saved data is lost. This is a choice if you want to try out a Linux distro without an actual install.

3. **Run Linux inside a Virtual machine:**

This a great option if you have a powerful pc/laptop. You can install a virtualisation software for free, e.g. VirtualBox and install Linux from a disk image into a new virtual machine. The advantage is that if you break your Linux machine (which is generally hard to do), you easily can reinstall it, as the virtual machine is basically only a file on your computer (<https://www.virtualbox.org/>)

**We will be running BioLinux as a virtual machine**



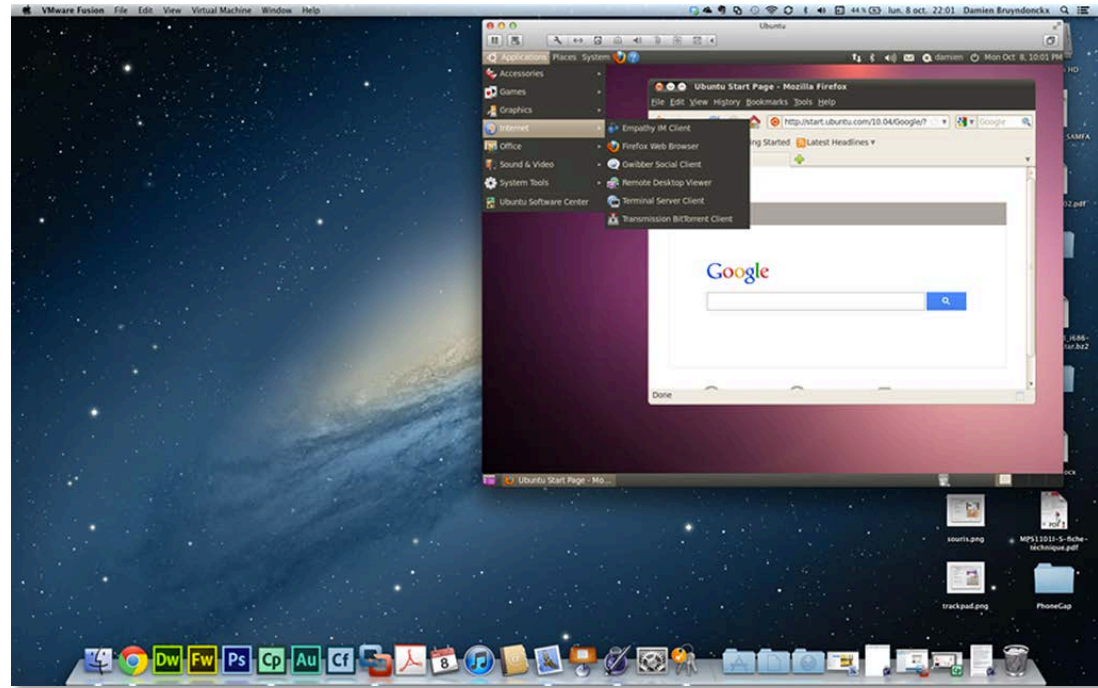
# Virtual machines (VMs)

- Basically the idea is to install a OS into a container, which is a single file.
- The virtualization software will start/run the OS from that file, e.g. uses the file as a hard-drive.
- An window within your host OS will open that shows the “virtual” OS running in the container.





## Virtual machines (2)



- Once you go over the window with your mouse you will be moving around inside the “virtual” OS.
- You can exchange files with the host OS through specifically created folders (this is needs configuration)
- The “virtual” OS will use as many host OS computer resources as you have allocated during the installation



# Computing philosophy

Unlike your Science...

- Be lazy.
- Copy others.
- Don't invent anything you don't have to.
- Re-USE, re-CYCLE, DON'T re-invent.
- Don't be afraid to ask others.
- Resort to new code **only when absolutely necessary.**
- Add comments to your code - **ALWAYS**





# Computing philosophy (2)

- You're not a CS, not a programmers
- Don't try to be them
- **But! Try to think like them**, at least a bit
- Google is your friend





# Questions?

Sebastian Schmeier  
s.schmeier@gmail.com  
<http://sschmeier.github.io/bioinf-workshop/>