Methods in Open Science

Computers

Image from NASA: https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

WV VIEW

AmericaView
Empowering Earth Observation Education
AmericaView.org - Est. 2003

Now that we have discussed key characteristics of data, we can move on to explore key concepts relating to how computers work with a focus on topics important to data scientists. Since computers and computation are a key component of analyzing data, it is important to understand how computers function and how they perform calculations.

Computers in Data Science

- Crunch numbers
- Store and process data
- Collect data
- Visual data
- Run models and experiments

2

Computers allow us to perform calculations. We use computers to store, process, collect, and visualize data. They serve as environments to run models and perform computational experiments. Thus, computers are central to data science.

Components

3

We will begin with a discussion of computer components including both software and hardware.

Software

- Computer code
- Instructions that allow computer to perform certain tasks or processes
- Computers can perform a variety of tasks:
  - Display images
  - Play/process audio and/or video
  - Allow users to draw 2D and 3D art
  - Generate text documents and graphics
  - Edit video files
  - Process data
  - Etc.

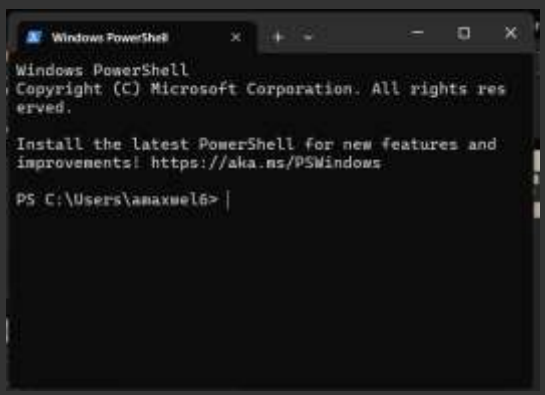
4

Software is effectively computer code that allows a computer to perform certain type of operations. Examples include creating text documents or presentations; surfing the web; viewing or editing images, videos, and/or audio; playing games; and making art.

In data science, we use specialized software to work with and analyze our data. We need specialized software because we undertake specific tasks, such as cleaning and preparing data, making graphs, performing statistical tests, and training machine learning algorithms.

Type of User Interface

- Graphical User Interface (GUI)
- Command Line Interface (CLI)



https://www.bbc.co.uk/bitesize/guides/z7qqmsg/revision/1

5

Most modern software provide a graphical user interface (GUI) in which you can graphically interact with the software tools using your mouse and keyboard. However, some software works only through a command line interface (CLI), in which the user must type commands into a terminal. Some software can be interacted with using a GUI or a CLI.

Although GUIs are generally easier to use, CLIs can be powerful for automating complex or repetitive tasks or executing code. Commands provided at the command line are often executed faster due to reduced overhead as the GUI does not need to be rendered and updated.

Operating System

- Control hardware
- Platform for software to run
- User interface
- Manage processes
- Manage memory
- Mange input and output devices
- Manage applications
- Manage security

Examples

Microsoft Windows

MacOS

Linux Distributions

Android

iOS

https://www.bbc.co.uk/bitesize/guides/z7qqmsg/revision/1

6

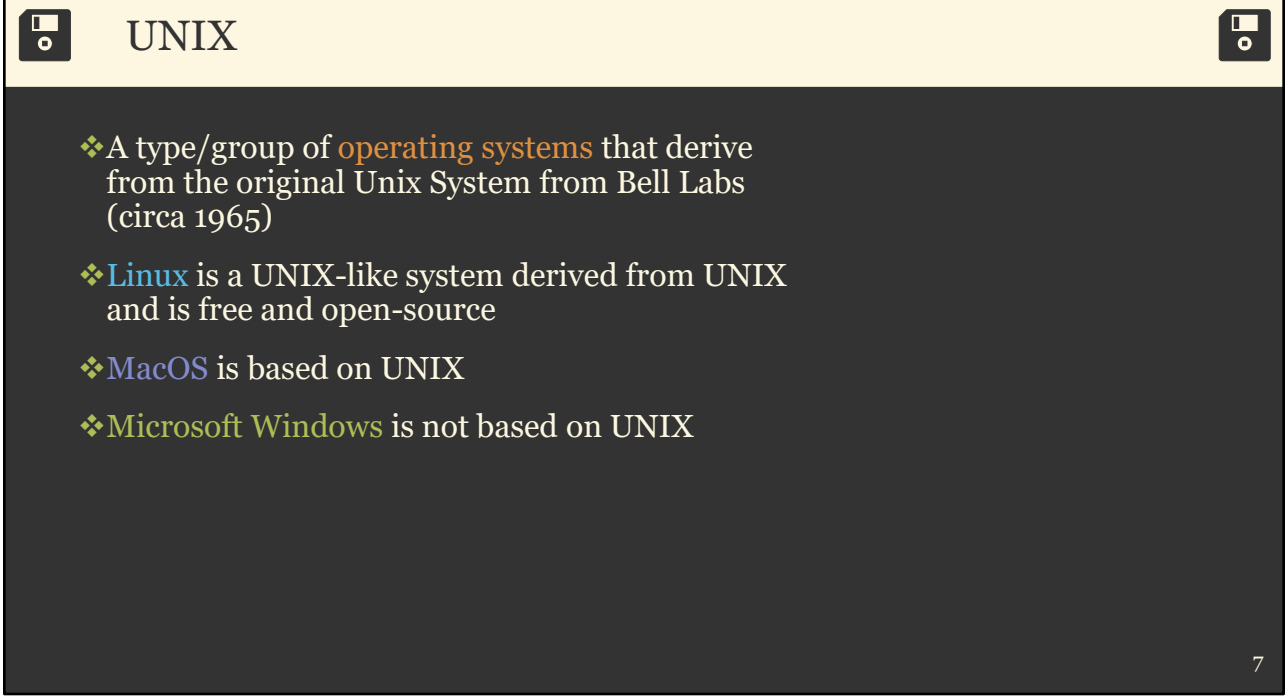
The computer's operating system controls and communicates with the hardware; acts as a platform for using software; provides a user interface to interact with the computer; and manages processes, memory, input devices, output devices, applications, and security.

Currently, the most commonly used commercial operating systems for laptops and desktop computers are Microsoft Windows and MacOS. There are also specific operating systems designed for servers (e.g., there are server distributions for Linux and Windows) and mobile devices (e.g., Android and iOS).

Microsoft Windows and MacOS are proprietary. The source code is not made available, and users must pay for a license to use the software. Linux offers an open-source operating system where the source code is made available. There are different versions of Linux desktop operating systems known as Linux distributions, or distros. Linux-based operating systems generally include the Linux kernel (kernels will be discussed in a later slide) and a package management system. Commonly used Linux distros include Ubuntu and Linux Mint. Both offer a window-based or GUI (graphical user interface) experience, which simplifies the use of the operating system.

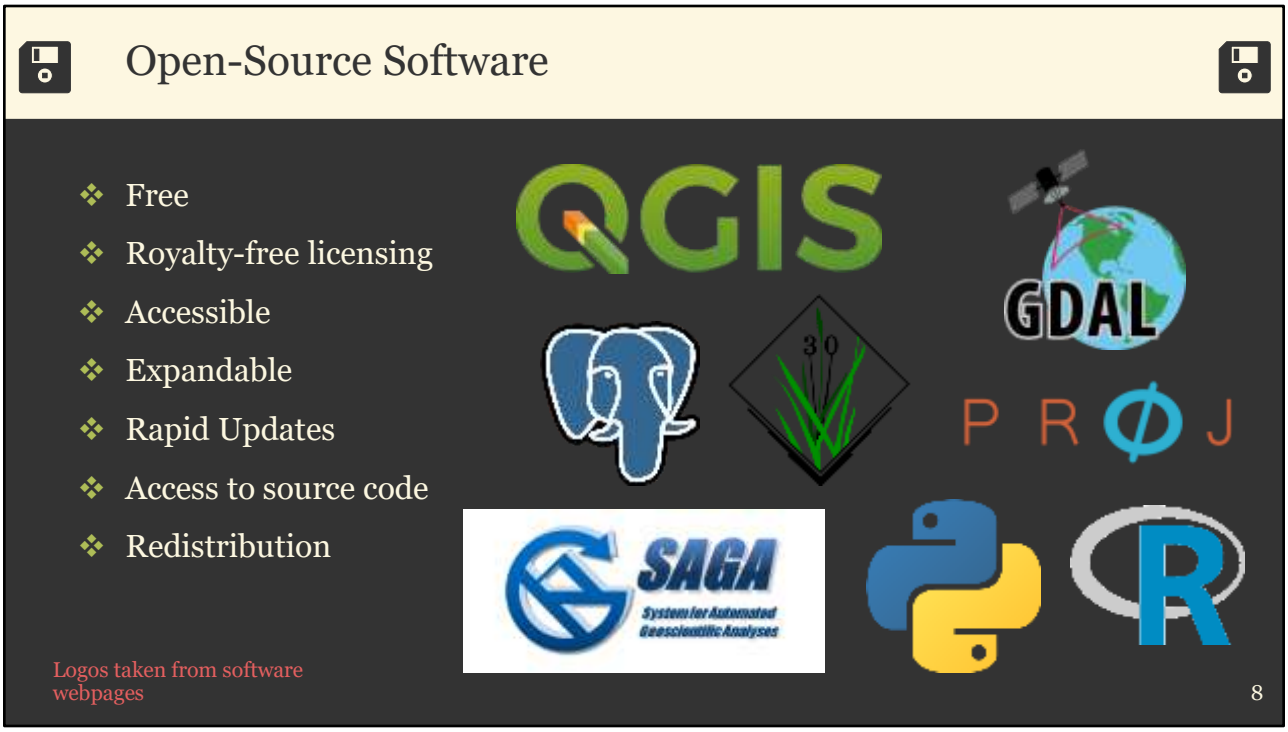
Most modern operating systems make use of a graphical user interface (GUI)

to allow easier interaction with the computer and installed software. However, there are also generally command line components that can be more efficient for certain tasks.



UNIX-based operating systems are derived from the original UNIX, or Unix, system developed by Bell Labs in the 1960s. UNIX was developed using a simple philosophy that its derivatives still follow to some degree. Linux distros are considered UNIX-like in that Linux was derived from an open-source version of UNIX. MacOS is based on UNIX. In contrast, Microsoft Windows is not based on UNIX.

UNIX or UNIX-like operating systems are commonly used in data science on laptops, desktops, and servers. However, many scientists use Windows-based systems without issue. Some scientists prefer UNIX or UNIX-like systems due to open-source options, speed and flexibility, and the availability of a powerful command line interface.



Why open-source? There are many reasons to consider working in an open-source environment. First, open-source software is free, which can greatly reduce the cost of implementing a project or operating a lab or consulting business. This also alleviates the need to manage and maintain commercial software licenses and increases the accessibility to the software.

Open-source software is often expanded by user-generated modules, libraries, packages, or plugins. As a result, open-source software tools are generally updated rapidly. If you are a coder or would like to contribute, the source code is made freely available and there are less redistribution restrictions.

However, there are some drawbacks. First, open-source software generally does not include customer support, which is commonly provided along with a commercial software license. This issue is partially alleviated by a large user base and online documentation. Using and troubleshooting the software will often involve searching or posting to help sites, such as Stack Exchange or Stack Overflow. Coding may be required to obtain the desired results.

In this course, we will rely on open-source software. Also, open-source software tools are the standard within the field of data science and help to foster reproduceable research. I am a big fan of open-source software. Without these tools, I would not be able to do my work. With that said, I do us a lot of

great commercial software products.

Open-Source Licensing

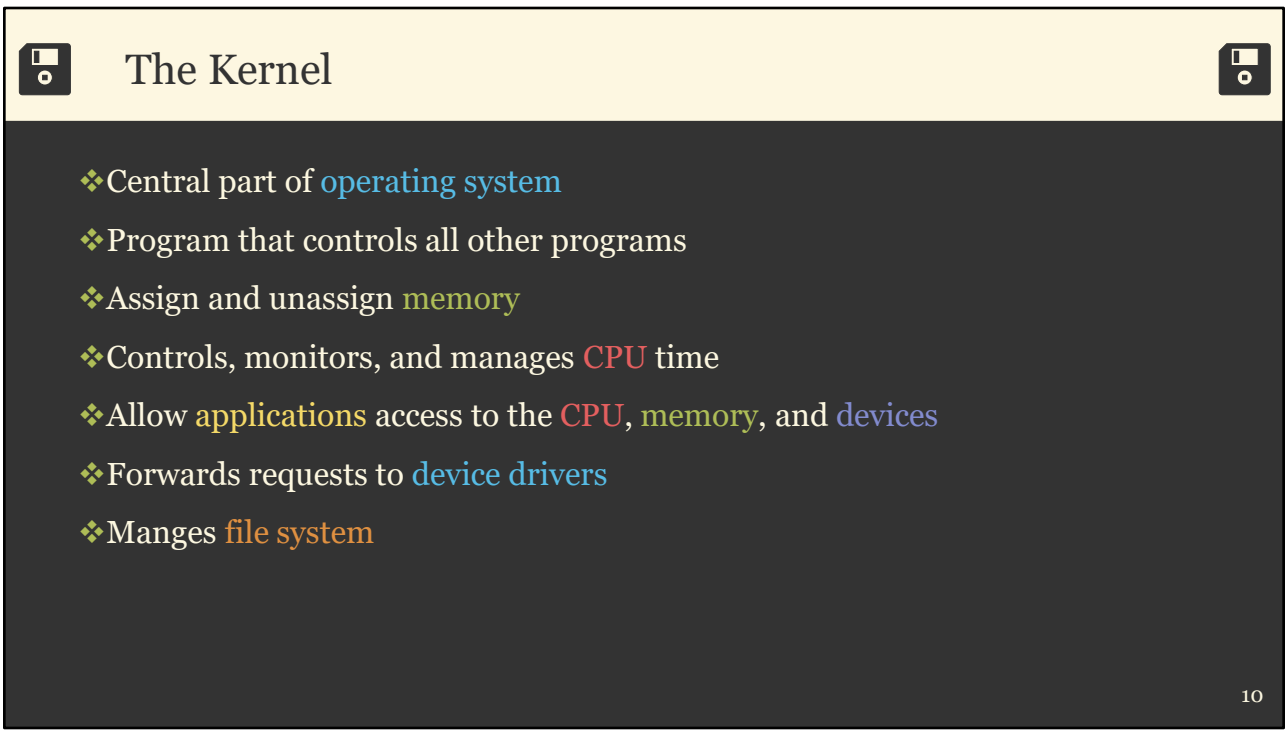
- Licensing with only limited restrictions on reuse and modification
- Examples:
  - MIT https://opensource.org/licenses/MIT
  - GNU GPL https://www.gnu.org/licenses/gpl-3.0.en.html

Copyright <YEAR> <COPYRIGHT HOLDER>
Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:
The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.
THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

9

Most open-source software is licensed under the MIT license or the GNU General Public License.

These licenses generally provide free access to the software and allow the user to use, copy, modify, merge, publish, distribute, and/or sublicense the software under certain conditions. This slide provides example language from the MIT license specifically.



A term that often arises in data science and when troubleshooting computational issues is the concept of the kernel. The kernel is the central part of the operating system that controls all other programs; assigns and unassigns memory; controls, monitors, and manages the central processing unit (CPU); interacts with device drivers; and manages the operating system's file system.

The kernel lives in the computer's memory and has control over hardware and software. Applications must make request to the kernel to use the computer's hardware.

Note that there are different types of kernels (monolithic, microkernel, hybrid, nano, and exokernel), which we will not covered or differentiated here.

32- vs. 64-Bit

- 64-bit allow for working with more data at once
- 64-bit systems can access more than 4 billion times the physical memory of 32-bit systems
- 32-bit software/systems are limited in the amount of data they can handle and RAM they can use
- Software can be designed for 32-bit or 64-bit architecture

https://www.digitaltrends.com/computing/32-bit-vs-64-bit-operating-systems/

https://cran.r-project.org/bin/windows/base/

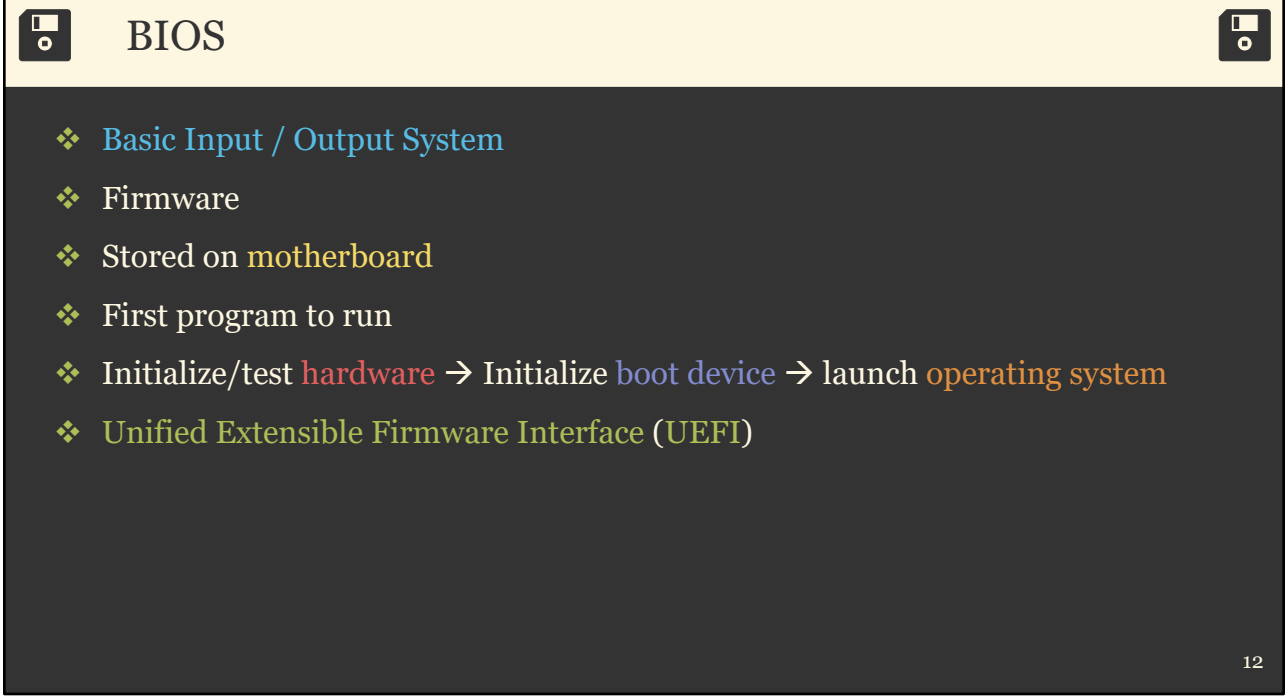
11

Another concept associated with the operating system and specific software applications in the concept of 32-bit vs. 64-bit. Practically, this relates to the amount of information or data that can be processed at once and how much memory is available to the application.

Most modern systems are 64-bit. Some software applications will have both 32- and 64-bit versions. If you have a 64-bit system, you should use 64-bit software if available.

Generally, 32-bit software can only use up to 4GB of RAM. So, if you have a computer with a larger amount of RAM (e.g., 64GB), the software will not be able to make full use of this memory. In contrast, 64-bit software will be able to access more of the available memory.

In short, you should use 64-bit software when possible.



Another common term is the BIOS. BIOS stands for Basic Input/Output System. The BIOS is usually stored in the read-only memory (ROM) on the computer's motherboard. It is the first program to execute when the computer boots or the power is turned on.

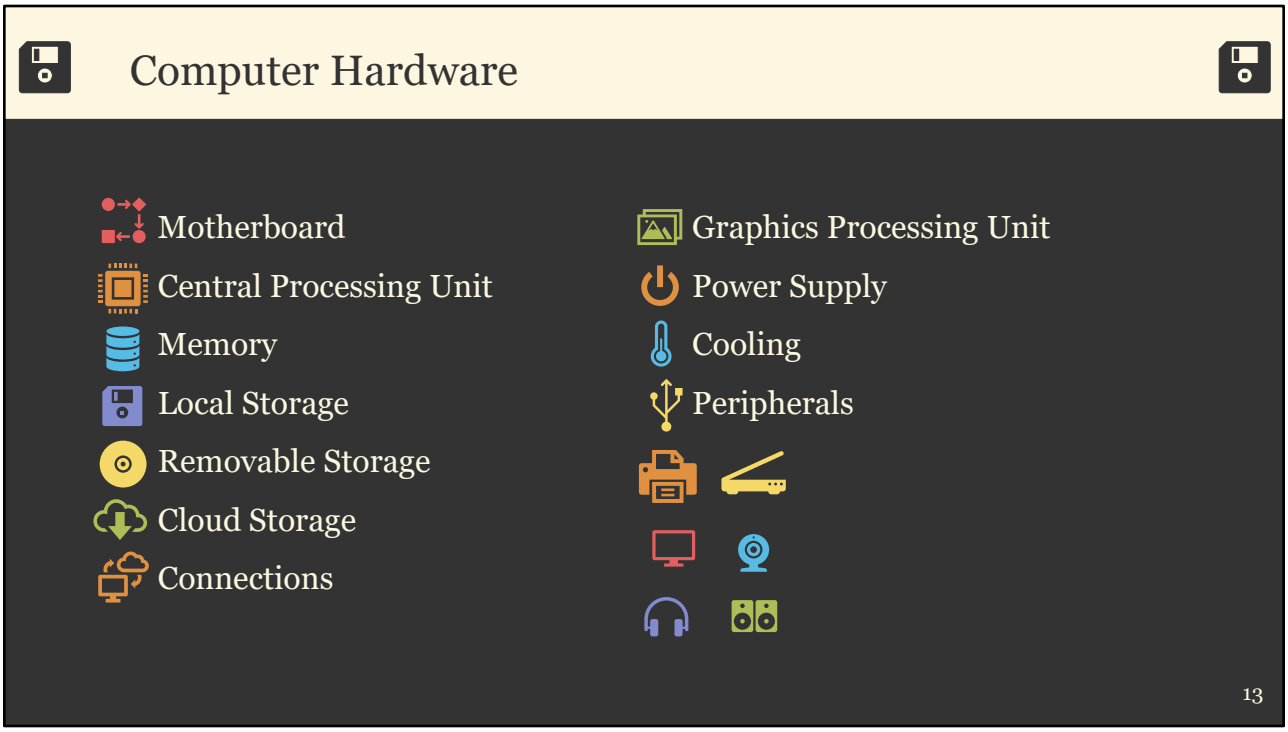
During the boot process, the BIOS will initialize and test the computer hardware. If there is an issue with the hardware, for example if the hard drive cannot be detected, an error will be raised, and the operating system may fail to boot. If all hardware initializes and tests without error, the BIOS will initialize the boot device, which stores the operating system. This is generally one of the hard drives in the computer. Lastly, the operating system will launch from the boot device.

You can think of the BIOS as the initial firmware that gets the computer ready for the operating system to launch.

Many modern computers now use a Unified Extensible Firmware Interface (UEFI) that provides more functionality than a basic BIOS. UEFIs often provide a GUI as opposed to a command line interface and offer the ability to boot the operating system from large drives.

The BIOS or UEFI can be used to change the boot order of devices. For

example, you can have the computer boot from a USB drive as opposed to the internal hard drive when installing or reinstalling the operating system. You an also use the BIOS or UEFI for power management and to change fan speeds.



We will now discuss key hardware components of a computer. Later slides will discuss the motherboard, central processing unit, memory, and graphical processing unit.

Other key hardware components include a power supply that provides power and an electrical current to all components in your computer that require them. The wires within the computer are generally either for communication or to provide power. Power supplies vary based on the wattage provided (energy per unit time = power). When computers are built, it is important that the power supply be able to provide enough power to run all the components. Some computer components draw more power than others. For example, graphics processing units tend to need a lot of power, especially those designed for gaming and high performance.

Connections include internal connections between the computer's components, such as connections between the motherboard and hard drive, connections with external devices, such as computer monitors, and connections with other computers within a local network or via the internet. Connections to the internet can be made using wires, such as broadband/ethernet, or wirelessly. The motherboard often includes hardware to allow internet connections via ethernet cables. Laptops generally have integrated wireless hardware. However, it is common for desktops to require

an external card or device to use wireless.

Cooling devices are very important to maintain the computer's health. The most common means to cool a computer is by using fans. There are generally fans that blow air into the computer and a separate set of fans that blow air out in order to maintain airflow. Hardware that generates a lot of heat commonly has dedicated fans. For example, the CPU will have a dedicated fan (i.e., the CPU fan). Graphics processing units also often have dedicated fans. An alternative to fans is to water cool a computer. A water-cooling system circulated water in tubs to carry away heat. It is also possible to use a combination of water cooling and fan cooling. For example, the CPU fan may be replaced with a water-cooling system while all other system fans are maintained. Note that the BIOS or UEFI can be used to control the fan speeds. Fan speeds can also adjust based on the amount of heat produced.

Common peripherals include monitors, printers, scanners, webcams, audio interfaces, and speakers.

Motherboard

- Ties computer components together and allows them to communicate
- Data is transferred through a series of wires connected to each component on the motherboard. These wires are called the bus.
- Buses
  - High speed connections
    - Address bus
    - Data bus
    - Control bus

https://commons.wikimedia.org/wiki/File:Mainboard_asus_pbh_67-v_IMGP9326_wp.jpg

14

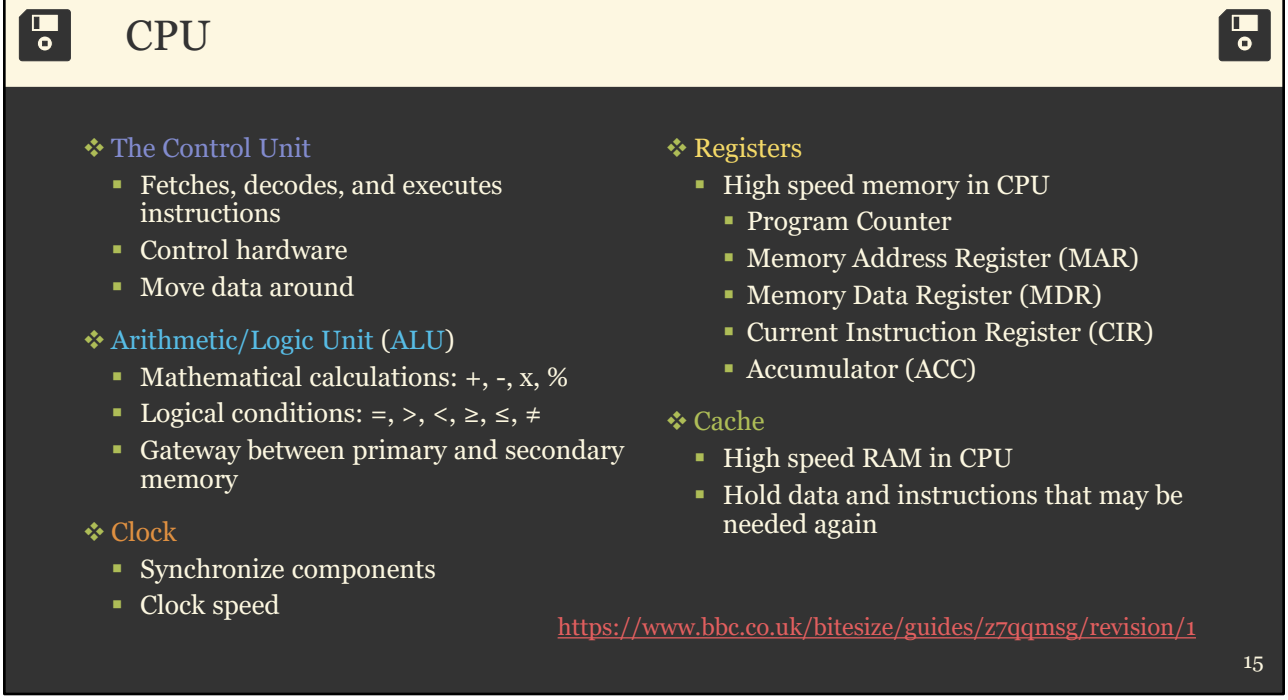
The motherboard acts as the central hub of the computer to which all other components and devices are connected. Computer components communicate with one another via the motherboard. The BIOS or UEFI firmware is installed on the motherboard in the ROM. The central processing unit is seated on the motherboard.

Motherboards often contain integrated chips to support internet connections, audio connections, and video output to monitors. However, additional hardware can be connected to expand functionality. For example, the PCI Express (Peripheral Component Interconnection Express) serial bus slots allow for a variety of cards to be connected to the motherboard including graphics processing units, audio cards, Wi-Fi cards, USB cards, etc.

Buses are used to send data and control signals from the CPU to other computer components. The address bus deals with memory addresses while the data bus is used to move around data. The control bus deals with control signals. Think of the motherboard as a small city, with each of the components being a building. Just like roads between buildings in a city, the buses are the roads that your data travel on to move from one component to another.

Memory and storage devices are connected to the motherboard so that data can be sent throughout the system. All the fans and cooling devices are also

attached to and controlled by components on the motherboard.



Please see the link on the bottom of this slide for additional information about the CPU.

The central processing unit (CPU) is the brain of the computer and performs computational and logical tasks. It is seated or connected to the motherboard and commonly has a dedicated fan. In order to more efficiently move heat away from the CPU, thermal paste is often applied between the CPU and CPU fan.

The CPU has several components. The control unit fetches, decodes, and executes instructions; controls hardware; and moves data around the system.

The Arithmetic/Logical Unit (ALU) performs the arithmetic (addition, subtraction, multiplication, and division) and logical operations (equal to, less-than, greater-than, less-than or equal to, greater-than or equal to, not equal to) that are central to how a computer “thinks”, processes input data, and generates output. Input data comes from the primary memory and can be saved to storage.

The CPU contains a clock that is used to synchronize all computational activities via regularly transmitted electrical pulses. The number of pulses sent per second (Hertz) is the clock speed of the CPU, which relates to the number

of computations that can be performed per unit time.

Registers are high-speed memory that is integrated into the CPU that store small amounts of data needed during processing. Registers speed up computation by making necessary data readily available. Different processors will have different numbers of registers for different purposes.

Cache is also memory, or RAM, stored within the CPU that is used to hold data temporarily and speed up calculation and computation.

CPU Performance

- Clock Speed = number of CPU clock pulses per second
- Cache Size = amount of available cache. Caches are temporary memory that store data close to the CPU. CPUs can pull data directly from the cache instead of having to pull data from memory or storage
- Number of Cores = separate cores to fetch, decode, and execute instructions

https://www.bbc.co.uk/bitesize/guides/z7qqmsg/revision/1

16

Several factors impact the speed and performance of a CPU. As discussed on the prior slide, the clock speed relates to the number of electrical pulses sent by the CPU per unit time in order to keep the computer's components coordinated. A higher clock speed generally indicates that more processes or computations can be performed per unit time. Thus, higher clock speeds correlate with faster computers.

The amount of cache can also impact performance. A larger cache means that more information can be stored within the CPU to support faster computation.

Many modern processors have multiple cores (2-, 4-, 6-, 8-, 10-, and 12-core CPUs are common). This allows for parallel processing of data by spreading the computation across multiple cores. Many modern CPUs can also make use of multithreading or hyper-threading to allow a single core to execute multiple threads of computation, thus mimicking parallel processing within a single core.

It is important to note that not all software will be able to make full use of a CPUs architecture (i.e., clock speed, cache size, and number of cores). This depends on how the software code was written. For example, software developers generally must develop code for a specific software that supports CPUs with multiple cores. If you are building or buying a computer for a

certain task, you should review the hardware requirements for the specific software to determine what is required and what might be overkill.

RAM

- Random Access Memory
- Primary memory
- Holds data and information for processing
- Only holds information as needed, not permanent.
- Data is volatile = needs electric power to be maintained. If power is lost or interrupted, data stored in RAM is lost

https://commons.wikimedia.org/wiki/File:SDRAM_memory_module.jpg

17

RAM, or Random Access Memory, is the primary memory used to store information during computations. The data and information stored in memory are available for the CPU to process.

Note that data and information are not permanently stored in RAM. Instead, data are added and removed as needed to support the computations and software applications being used at any given time. When a computer shuts down, the RAM is emptied. Data stored in the RAM is considered volatile. If power is interrupted, data stored in the RAM is lost.

Having a larger amount of RAM means that more data can be temporarily stored to support computations and processes. More RAM can support multitasking, such as running multiple pieces of software at once and/or performing multiple computations at once. Note that the operating system controls the amount of RAM that can be used by all software that are running and require memory resources.

Although more RAM generally supports the processing of more data and/or larger files, software must be written to take full advantage of the available RAM. For example, 32-bit software only has access to up to 4GB of RAM regardless of the amount of RAM installed in the system.

HDDs vs. SSDs

- Secondary memory
- Long-term storage

- Hard Disk Drives (HDD)
  - Magnetic disk
  - Moving parts
  - Generally slower
  - More power consumption
  - Not silent
  - Currently cheaper
- Solid-State Drives (SSD)
  - Flash memory
  - No moving parts
  - Generally faster
  - Less power consumption
  - Silent
  - Currently more expensive

https://commons.wikimedia.org/wiki/File:IBM_IDE_Hard_Drive04.jpg

https://commons.wikimedia.org/wiki/File:Samsung_SSD_840_120GB_MZ-7TD120--2_BOTTOM_SCREWS_REMOVED.JPG

18

Data are stored long term or more permanently in the secondary memory. It is also possible to store data on a more mobile or external storage devices, such as an external hard drive or USB flash drive. Data can also be written to media, such as optical disks (i.e., CDs, DVDs, or Blu-ray). It is also possible to store or back up data remotely to a network drive, server, or cloud storage device.

Internal secondary or long-term memory generally makes use of hard disk drives or solid-state drives. Hard disk drives make use of magnetic disks and have moving parts required to read the data. In contrast, solid-state drives use flash memory as opposed to magnetic disks and have no moving parts. Solid-state drives tend to require less energy to operate and are faster than hard disk drives. However, they are currently more expensive.

In modern computers, it is common to use a solid-state drive as the primary drive from which the operation system is stored and boots. This generally allows for faster start-up and responsiveness. A secondary data drive can then be used to store large volumes of data, which could be a hard disk drive or a solid-state drive. Currently, it is cheaper to use hard disk drives if a large volume of data storage is needed (e.g., > 2TB).

Some software tools will require or suggest the use of solid-state drives. Also, some data science methods will work more efficiently using solid-state storage.

For example, when performing convolutional neural network (CNN)-based deep learning, it is generally recommended that all the training and testing data be placed on a solid-state drive. This is because large volumes of data must be moved around, so transferring data from a hard disk drive could drastically slow the process.

Currently, using a solid-state drive as your primary system drive as opposed to a hard disk drive is often the first suggestion for speeding up a slow computer unless the slowness is resulting from a software as opposed to hardware issue.

GPUs

- Graphics Processing Unit
- Renders graphics
- Chip on your video card (discrete), or on the CPU (integrated)
- Can be used for computation
- Thousands of cores to process tasks in parallel

https://commons.wikimedia.org/wiki/File:NvidiaTesla.jpg

https://www.notebookcheck.net/NVIDIA-T600-Laptop-GPU-GPU-Benchmarks-and-Specs.532552.0.html

19

Graphics processing units (GPUs) are used to render graphics. Although most motherboards have integrated graphics, a more powerful graphics card may be needed for certain tasks such as gaming, 3D animation, virtual reality, video editing, and data science. GPUs integrated into a video card or graphics card are known as discrete while those integrated into the CPU are known as integrated.

GPUs generally have thousands of cores to process data in parallel. They are generally very efficient at processing and rendering 3D and 2D graphics.

For data scientists, graphics cards can also be used to support computations. For example, CNN-based deep learning models are often trained using graphics cards, which are much faster than CPUs for this task.

GPUs are often some of the more expensive components of computers and often consume a lot of power and produce a lot of heat. They tend to have dedicated fans or cooling systems. They are generally connected to the motherboard using Peripheral Component Interconnected Express (PCIe) slots.

Drivers

- Software that allows the operating system to communicate with devices, hardware, and some software
- Drivers are required for:
  - Controllers
  - Motherboard chipsets
  - Network/Wi-Fi cards
  - Sound cards
  - GPUs
  - USB devices
  - Printers and scanners
  - And more

20

Drivers are pieces of software that allow the operating system to communicate with hardware devices including controllers, the motherboard chipset, network/Wi-Fi cards, sound cards, GPUs, USB devices, printers, and scanners.

The operating system uses the drivers to allow software to accept input from and/or send output to devices.

I generally think of drivers as the software instructions that allow the operating system and software to communicate with and use your hardware.

Most modern operating systems can automatically find, install, and update drivers. However, specialized hardware, such as a GPU, may require the manual installation or updating of drivers. Custom-built systems may also require more manual management of drivers.

Drivers are often downloaded from the hardware or software vendor's website.

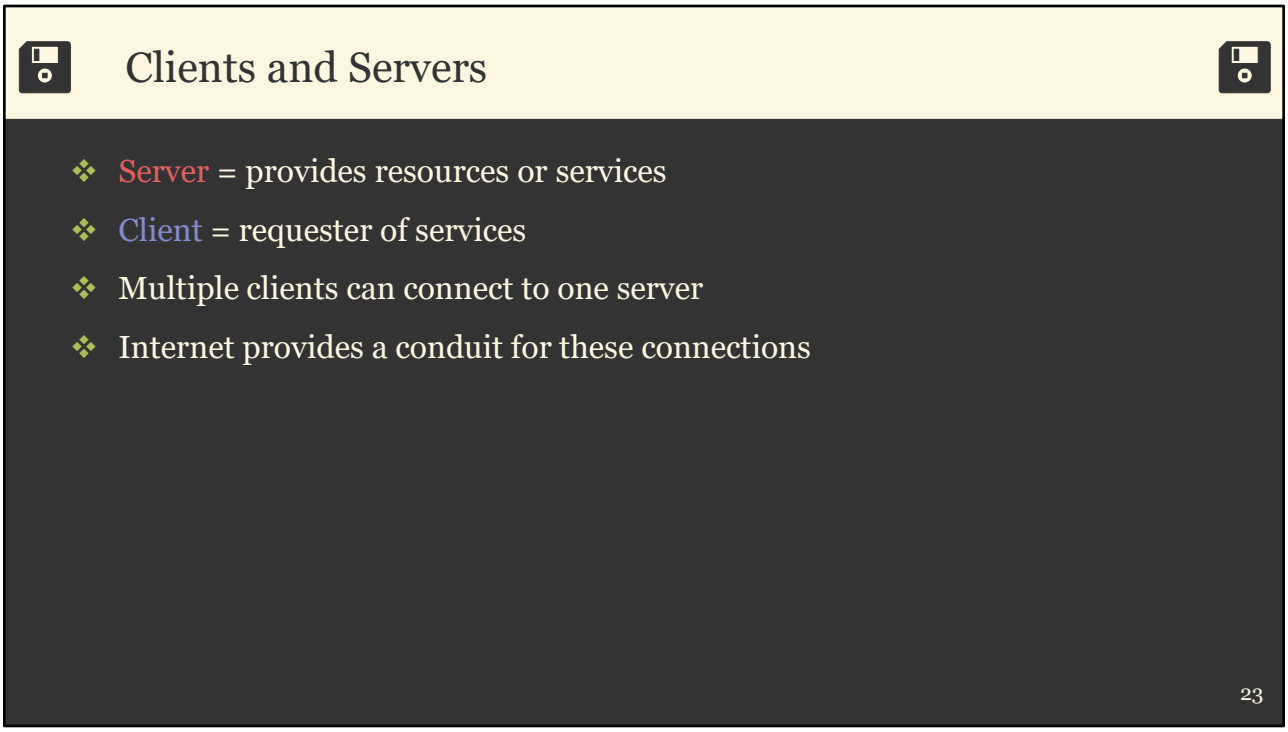
Summary of Computer Components

- Computers are a key tool for working with and analyzing data
- Having a basic understanding of how computers work is beneficial to data scientists
- Computers that will be used to analyze data should be built with appropriate software and hardware specifications

Connections

22

In this last section of the module, we will discuss the web and how computers are connected to allow for communication.

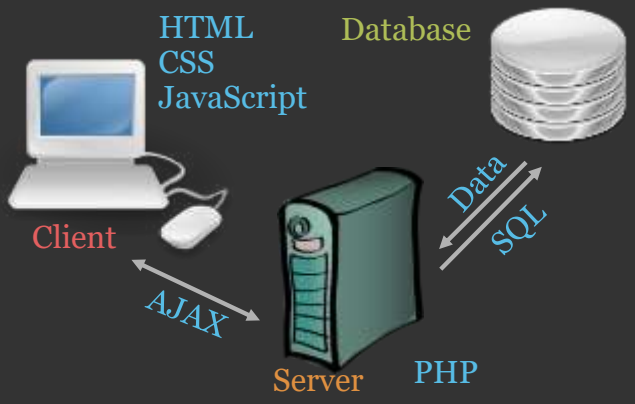


The internet consists of computers and other devices that are connected remotely to form a network. Within this network, servers provide resources and services while clients request these services.

It is possible for multiple clients to access or connect to one server using the internet. In order to transfer information and perform tasks, protocols have been established for standardization.

Client-Server Architecture

- Client = computers, tablets, phones, etc.
- Multiple clients can interact with one server
- Client requests data from server
- Server processes request and returns result
- Client does something with result



By OpenClipart - OpenClipart, CC0, https://commons.wikimedia.org/w/index.php?curid=721320
https://commons.wikimedia.org/wiki/File:Scribley_PC_clip_art.png
https://commons.wikimedia.org/wiki/File:Oxygen480-places-server-database.svg

24

Client devices are those that are requesting information from a server and can then do something with this information, such as render a webpage or view data from a database. Client devices include personal computers, tablets, phones, etc. Multiple clients can request data from a server simultaneously.

Servers process requests from clients and return information, such as the information needed to render a webpage or data from a database.

A variety of computer languages are used on the client, generally termed client-side technologies. These include hypertext markup language (HTML), cascading style sheets (CSS), and JavaScript.

It is common for different technologies and languages to be used on the server-side. For example, structured query language (SQL) can be used to access data from a database.

Although we commonly think of clients and servers as computers, they are really software, as a traditional desktop computer can act as a server with the correct software installed.

Client-Side Web Languages

HTML (Hypertext Markup Language) = used to create the basic structure and content of a webpage

https://www.w3schools.com/html/default.asp

CSS (Cascading Style Sheets) = used to design the webpage (where everything is placed and how it looks)

https://www.w3schools.com/css/default.asp

JavaScript = define the interactive elements of a webpage that helps to engage the users

https://www.w3schools.com/js/default.asp

25

Hypertext Markup Language (HTML) is used to define or build the content of a webpage. For example, it is used to define text, images, lists, links, and additional content on a webpage. Currently, we are using HTML5, the most recent iteration of the language.

Cascading Style Sheets (CSS) are used to define the styling of the webpage, such as fonts, font size, colors, and layout. It is also used for responsive web design, which entails the layout of the page changing based on the screen size. This is how a website can function on screens of variable sizes. Given the wide variety of devices we currently use to access web content, this is extremely important.

JavaScript is a programming language that allows for interactive content. For example, clicking a button on a webpage causes something to happen, such as data from a form being uploaded or the page content changing.

Book Recommendation

https://www.oreilly.com/library/view/html-css/9781118206911/

https://www.oreilly.com/library/view/javascript-and-jquery/9781118531648/

https://www.wiley.com/en-us/PHP+%26+MySQL%3A+Server+side+Web+Development-p-9781119149217

26

This book series is a great resource if you want to learn web development. This is not a topic covered in this course; however, I wanted to point out these great resources.

Anatomy of URL

- URL (Uniform Resource Locator) = address for things on the web

  http://www.wvgis.wvu.edu/lidar/

- Protocol = A set of shared instructions which specifies how two systems communicate
  - HTTP (Hypertext Transfer Protocol)
  - FTP (File Transfer Protocol)
  - Etc.
- Domain Name = identifies a website
- Path = file or directory on the web server

27

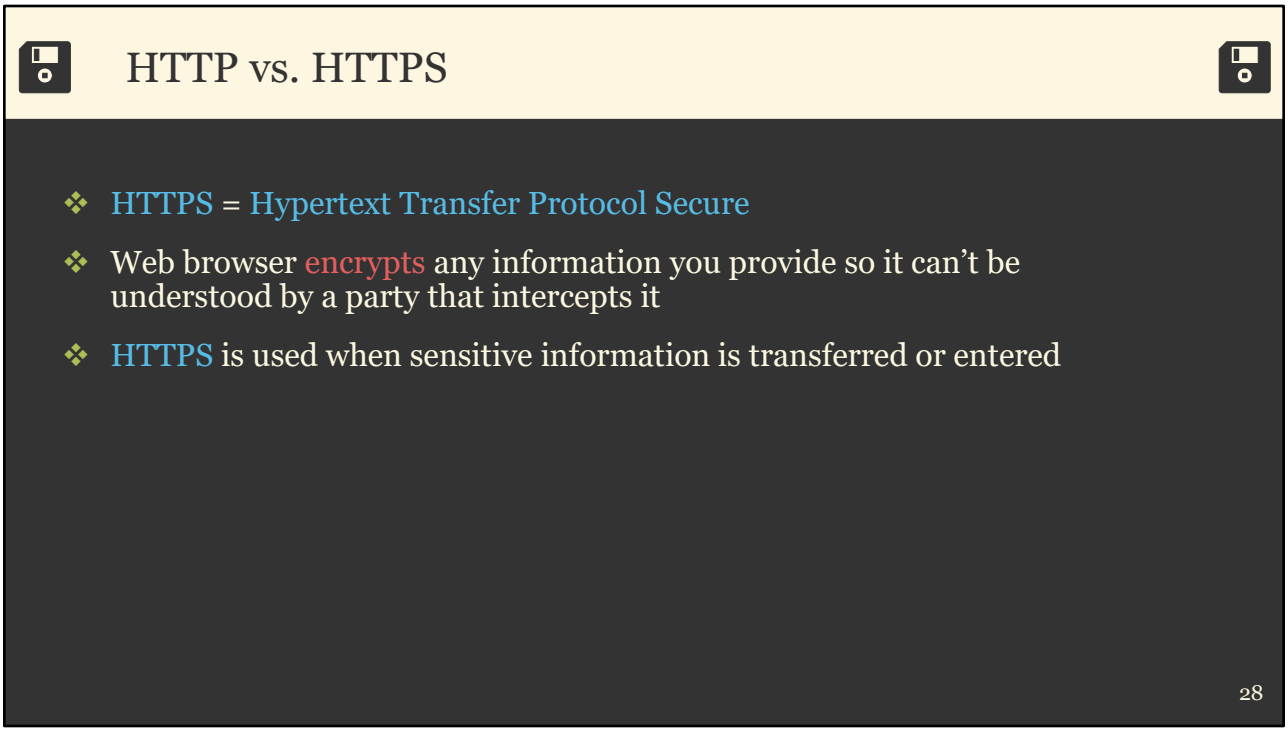
A URL (Uniform Resource Locator) acts as an address for a specific page or content available on the internet, and it consists of several components.

First, a protocol must be defined. For example, HTTP (Hypertext Transfer Protocol) is used to transfer HTML content while FTP (File Transfer Protocol) can be used to access and transfer files.

Next, a domain name is provided. For example, wvview.org or wvgis.wvu.edu.

Finally, a path can be provided for a specific page within the domain. For example, http://www.wvview.org/maxwell.html.

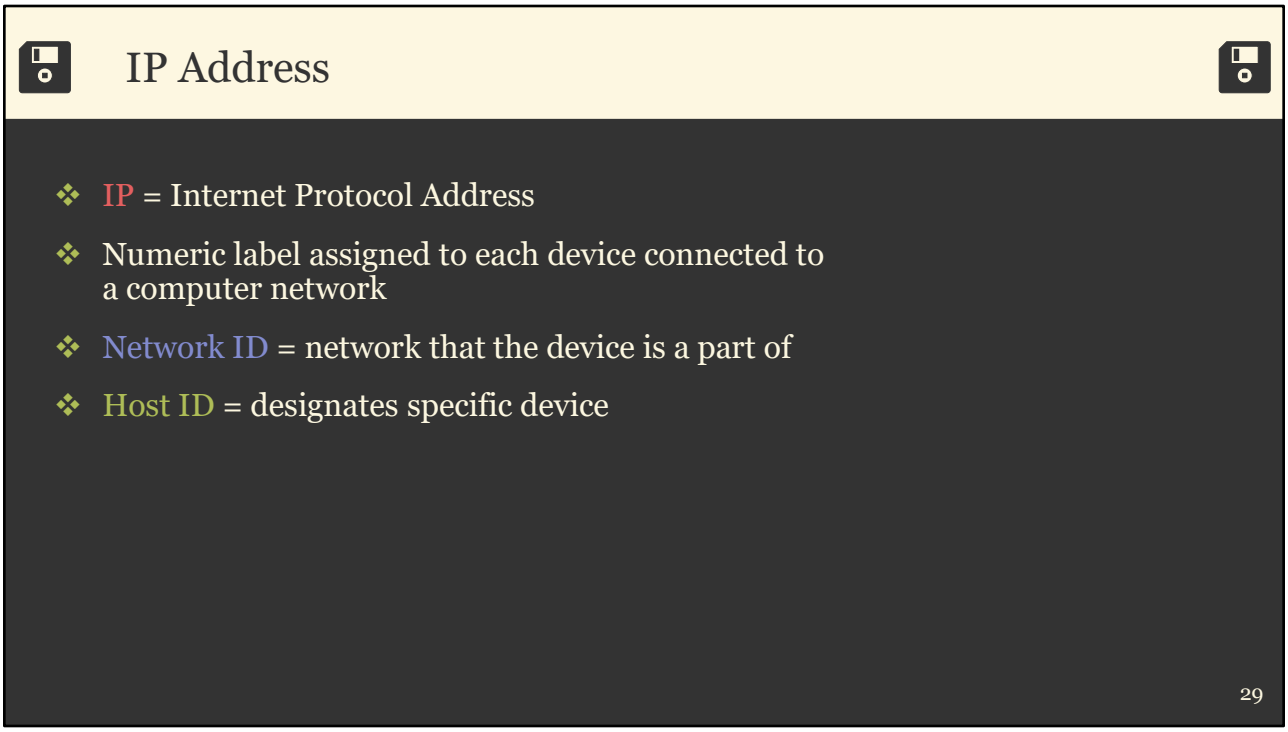
On a server, a webpage is generally defined by a folder with the main page or landing page named index.html. The subsequent pages will occur within the folder. For example, on the West Virginia View webpage the main landing page is named index.html while my webpage is named maxwell.html.



In contrast to HTTP, HTTPS (Hypertext Transfer Protocol Secure) allows for data encryption to protect sensitive data.

Data are encrypted for transfer as opposed to being submitted as plain text that can be interpreted if intercepted.

In order to make a site secure using HTTPS, you will need to purchase a certificate and set up an SSL (Secure Sockets Layer).



An IP (Internet Protocol) address provides a unique, numeric address for each client or server connected to the internet. They can also be used to identify a device on a local network, such as a university domain, business domain, or home network.

The IP address consists of the Network ID, which indicates the network that the device is a part of, and the host ID, which designates the specific computer or device.

For example, all devices on your home network will have the same Network ID but different host IDs.

Using a Domain Name System (DNS) allows you to provide a URL as opposed to an IP address to access a website.

A dynamic IP address is often used. This entails assigning host IDs to devices dynamically within the network when they are requested. So, each device in the network will not maintain a consistent host ID.

LAN (Local Area Network) is used to define a local network, such as that for a single business or home, while WAN (Wide Area Network) is used to define larger networks, such as those associated with internet service providers.

Ports

- Servers make services available through the Internet or network using numbered ports
- Separate ports will be used for different services and protocols
  - HTTP commonly uses TCP port 80
  - HTTPS commonly uses TCP port 443
  - FTP commonly uses TCP ports 20, 21
- Clients connect to a service at a specific IP address and a specific port.
- IP Address + Port defines address of a particular service on a particular system

30

Ports identify specific addresses for applications and services provided by a server.

For example, a port can be defined to provide a license for software in the case of a license server. A client computer that needs to access the license would need to do so using the appropriate IP address and port for the server that is hosting it.

The server will listen for requests coming in through the port.

Note that it is generally not a good idea or impossible to use the same port number for multiple tasks, such as to license multiple pieces of software. Fortunately, multiple port numbers are generally available for a single server.

What is an API?

- API = Application Programming Interface
- Allow software to interact with each other
- Provide code, functions, and classes that can be used to complete a task on the web

31

An Application Programming Interface (API) is a set of rules, functions, and tools that allows one piece of software to interact with another.

APIs provide predefined code and procedures that developers can use to perform specific tasks without needing to write the underlying functionality from scratch.

APIs have many uses and can be thought of as libraries of code and tools that allow programs to access data, process information, or generate content.

Summary of Computer Connections

- The internet is a network of interconnected computers that allows for the sharing of information and is central to the collection and dissemination of data.
- Client computers make requests to servers and servers respond to these requests.
- A variety of client-side and server-side technologies allow computers to communicate and information to be requested and delivered.
- APIs allow code to talk to other code and are essential in many data science tasks.

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

Thanks! Hope you found this useful.