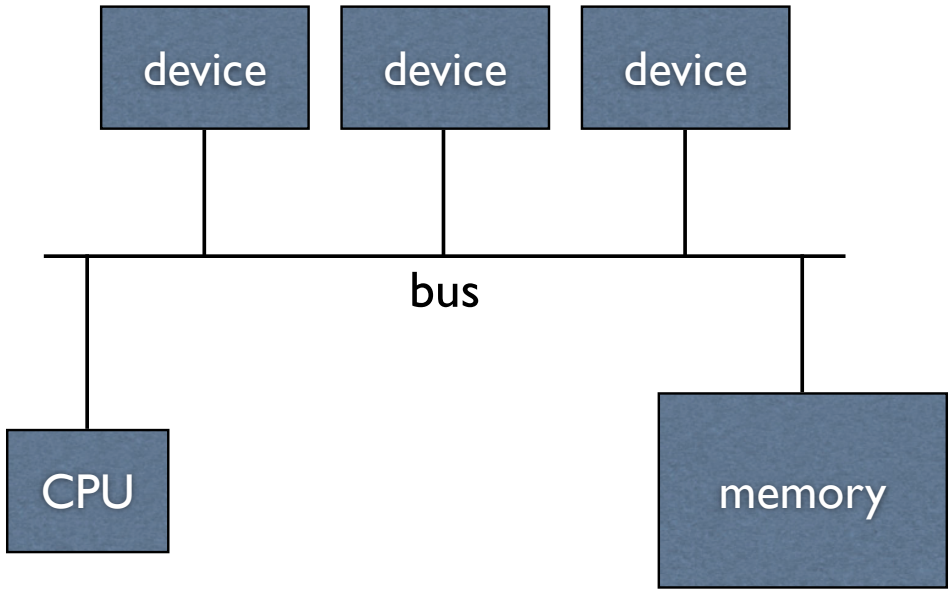
Where's My Stuff?

Computers, Networks, Files, and Data

A Computer "Model"

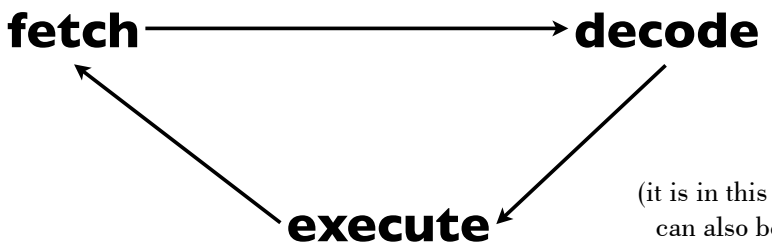


An Attempt at Analogy

Computer	Organism
CPU	brain (mostly), heart (clock)
memory	hippocampus, cerebellum, cerebral cortex
input devices	senses
output devices	limbs, voice, face

A Computer's "Pulse"

- Deep inside, all computing activity falls under a uniform, infinitely repeated structure — like a pulse:



(it is in this sense that a CPU can also be compared to a heart in addition to a brain)

- What we perceive as productive work gets done when these activities cause devices to behave in ways that we recognize as "useful"

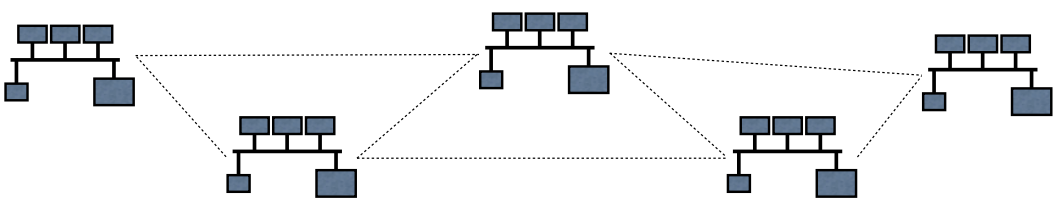
Two Types of Memory

Computing systems use two types of memory, analogous to our short- and long-term memory:

- Volatile memory needs power to stay around; cut the power, lose the information — this is typically the "main memory" or "RAM" in the computer, and constitutes the "memory" part of our computer model
- Non-volatile memory retains information without power — hard drives, discs, and flash drives fall under this category, and they are viewed as devices in our model

Networked Computers

- If we take our computer model and:
  - Make multiple copies of this model
  - Assume that these copies can communicate with each other in some way
- …then we have a network of computers



Bridging the Gap

In all cases, the trick with "getting stuff done" involves connecting what computers manipulate in memory or across networks with what we consider to be useful

What You See	What Computers See
documents	**0**s and **1**s
pictures	**0**s and **1**s
movies	**0**s and **1**s
statistics	**0**s and **1**s
everything else	**0**s and **1**s

Names and Places

- In all cultures, names are powerful and fundamental constructs — they bestow identity and eliminate confusion (most of the time)
- Places are also powerful: they provide context, direction, and relationships
- Fortunately, names and places can be expressed in ways that computers can also interpret — they bridge the gap between the way we see information and the way computers see it

On the Network

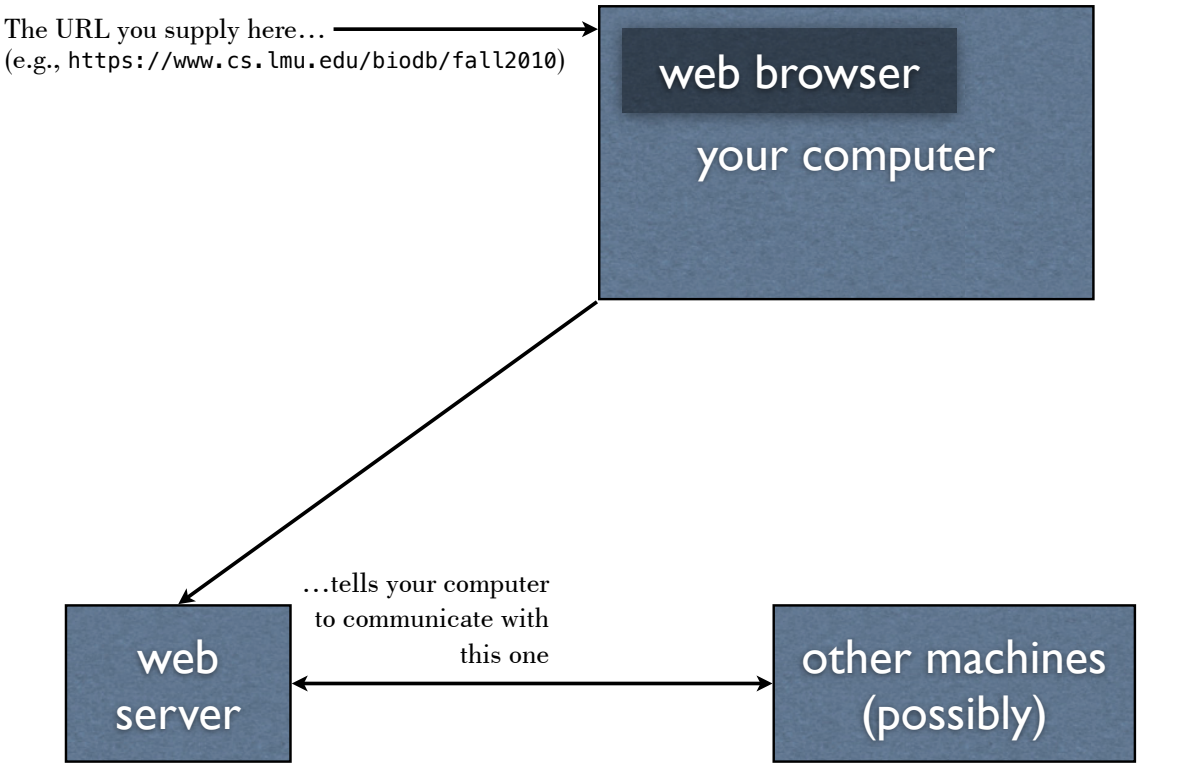
- When you are working on information within a web browser (e.g., Internet Explorer, Firefox, Safari, Google Chrome, etc.), you are most likely working with information that resides on a computer other than yours
- This information is identified using a URL, short for Uniform Resource Locator — something that combines name, place, and mechanism:

`https://www.cs.lmu.edu/biodb/fall2010`

protocol: how to communicate

host: the machine's name

path: where to go "within" the machine



That computer, in turn, may rely on other machines to do or store your work

On Your Computer

- Information that resides on your computer is typically stored on a device that is connected to it, such as a hard drive, a disc, a server, or a USB/flash drive
- This information is named using files
- Files are organized (i.e., put in place) using folders — these can contain both files and further folders
- The "top" of the chain of folders is the device itself (i.e., hard drive, disc, server, USB/flash drive, etc.) — this may be explicit or implicit, depending on the OS

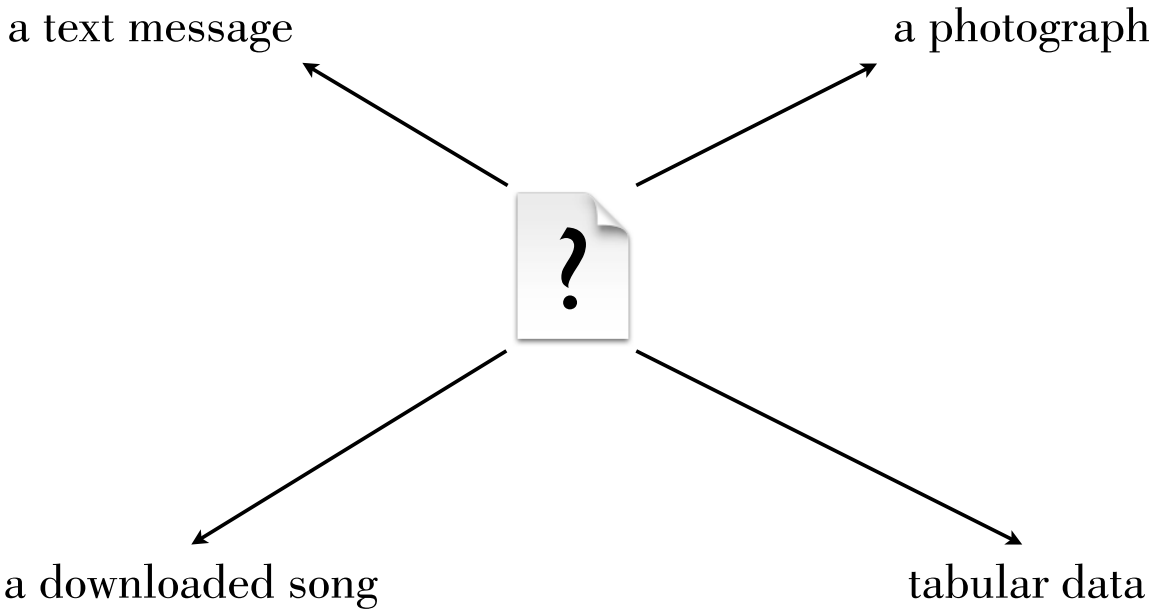
Paths Lead to Files

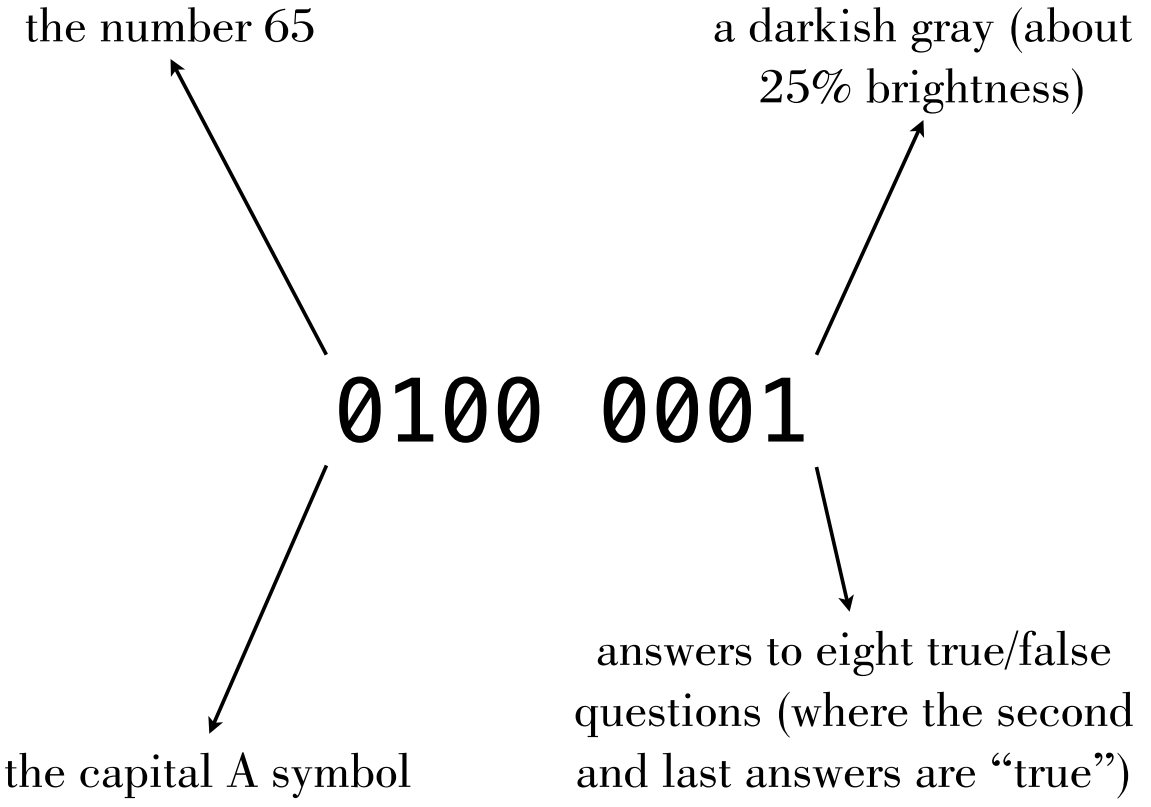
- Since files and folders can go into additional folders, there is also a notion of place, e.g., "The file called paper.txt is in the Documents folder, which is in the joe folder, which is in the Users folder on my hard drive"
- This sequence is called a path, and is expressed succinctly by just writing the whole thing out, sometimes including the overall storage device:

```
/Users/joe/Documents/paper.txt
```

```
C:\Users\joe\Documents\paper.txt
```

What's in a File?





File Extensions

- The most prevalent solution to the challenge of telling a computer "what's in the file" is the file extension
- A file extension is a suffix that is appended to its name, preceded by a period: `.txt`, `.html`, `.jpg`, `.doc`, `.xls`, `.zip` — the list goes on and on
- A file extension is part of a file's name, so it can be changed — but it does not change the file's contents (which is why most user interfaces hide or restrict their display/manipulation)

- Any single type of file, particularly common ones like text, tabular data, or images, can be interpreted and manipulated by more than one computer program
- As a shortcut, many operating systems allow us to specify a default program to start whenever we double-click on a file, determined by the file extension
- While this is sufficient for basic file use, advanced use goes beyond this behavior very quickly, resulting in less double-clicking and more:
  - Starting a program first and then opening from there
  - Invocation of Open With… or similar commands
  - Use of drag-and-drop

“Text” File Types

Some file types are considered to be “text” types — this means that most human beings can read their contents without much difficulty, and they can be successfully opened by any text editor program, in addition to the specialized programs for which they are intended

- Tabular data (`.txt`, `.csv`, `.dat`, `.tab`, etc.): These files have internal structure, such as separate columns, individual records, or specific expected values (a specific symbol to “separate” columns, numbers for a particular column, etc.)

- Source code (`.java`, `.js`, `.c`, `.sql`, etc.): Most computer programs start in the form of “source code” — human-readable text that specifies what the program is supposed to do — and these are also “just” text files written in a particular language
- XML files (`.xml`): These store data like their tabular equivalents, but are more outline-oriented
- Web pages (`.html`): If you’ve ever tried out the View Source command on a web browser, you’d have seen that web pages are really human-readable text — when a text editing program opens an `.html` file, you see this code; when a web browser opens it, you see its intended layout and content

“Binary” File Types

In contrast to text files, binary files cannot be easily read by humans — they appear as junk when opened in a text editor, but remain meaningful to computers and their associated programs

- Image files (`.jpg`, `.png`, `.gif`, `.psd`): Internally, they hold grids of color values in addition to “metadata”
- Document files (`.doc`, `.xls`, `.ppt`, `.pdf`): When a file needs formatted text or visual layouts, they typically use a binary representation
- Media files (`.mp3`, `.m4p`, `.mov`, `.wmv`, `.mpg`): These include music and movie files
- Executable files (`.exe`, `.app`, `.ipa`): Yes, programs are files too — and so they also have their own extensions
- Compressed files (`.zip`, `.gz`, `.tar`, `.rar`, etc.): These are files containing other files (text or binary), processed so that they take less space — compressed files cannot be used as is; they need to be decompressed first
  - Some file formats use compression intrinsically — image and media files, in particular, have done this for a long time
  - More recently, even document files have started to use compression as part of their format (case in point: the latest Office’s `.docx`, `.xlsx`, `.pptx`, etc. formats)