HPHA 5330

Health Informatics and Data Analytics

Computers

Hardware

- First counting/calculating devices – fingers, pebbles, beads, knots on a string, etc.
- Next – abacus (put the beads on a frame)
- First electronic computers used vacuum tubes – large, expensive to build, expensive to operate

Hardware, cont'd

- Transistor technology replaced vacuum tubes (smaller, less expensive to build and operate, faster, more rugged)
- Integrated circuits (IC's) replaced transistors (smaller size, faster, less expensive)
- Advances -> more affordable, more portable, multiple applications

Definitions

- Server – a computer that shares resources (programs, applications, data) and communicates with other computers; multiple users “simultaneously”; often a mainframe
- Workstation – smaller, fewer functions than a server; fewer users (usually one at a time); specialized functions; usually connects to a server

More definitions

- Personal computer (PC) – often single user; home or office; email, word processing, personal spreadsheets, databases, personal files, education, games, internet connection, etc.
- Terminal (dumb terminal) – no processing capability; connects with a server; retrieves data from the server, inputs data to server; retrieves and displays the results of computations

More definitions

- Central Processing Unit (CPU) – performs operations and computations, coordinates operations, coordinates communication with peripheral devices (hard drive, external hard drive, printer, modem, scanner, etc.)
- Memory – stores data, stores programs
- Storage media – tape, disk (floppy, CD, DVD), external hard drive, memory chip, thumb drive, etc.

More definitions

- Input and output (I/O) devices –
  - Input devices – keyboard, mouse, input from external storage media (tape reader, disk reader, card reader, etc.)
  - Output devices – screen, printer, output to external storage media (tape writer, disk writer, card writer, speakers, etc.)

Communications devices

- Modem – converts data in digital format to analog format and vice versa
- Hub – simplest device for connecting computers with each other; all information is available to all computers on the network
- Switch – more discriminating than a hub; recognizes individual computers and directs information accordingly
- Router – can distinguish facts about the information and can choose alternate pathways for transmitting the information

Binary

- Data is manipulated and stored as digital information; two states – current is either "ON" or "OFF"; voltages are either "high" or "low"
- Represented as "1" or "0"

One state (either “0” or “1”) is called a “bit”

Sequences of bits (“0”s and “1”s) represent numbers, alphabetic characters, punctuation, or programming instructions

In order to understand the workings of a digital computer, you need to understand binary arithmetic (arithmetic using base 2 instead of base 10)

In base 10 arithmetic (what we normally use every day), we have ten numerals to work with:
0, 1, 2, 3, 4, 5, 6, 7, 8, 9

We can write any number using these ten digits, by assigning "place values", increasing from right to left by a factor of 10 for each place

1= 1 unit = 10^0*
10 = 10 units = 10^1
100 = 100 units = 10^2
1,000 = 1,000 units= 10^3
etc.

* remember, **any** number, other than zero, raised to the "0" power = 1

The numeral that is in a specific place tells us how many to multiply that place value by:
125,364 =
1x10^5 units + 2x10^4 units + 5x10^3 units + 3x10^2 units + 6x10^1 units + 4x10^0 units

Multiply and then add them up

- 1x10^5 units + 2x10^4 units + 5x10^3 units + 3x10^2 units + 6x10^1 units + 4x10^0 units
- = 100,000 + 20,000 + 5,000 + 300 + 60 + 4
- = 125,364

In binary arithmetic, we have only two digits:
0, 1

We can represent any number using just these two digits, by assigning "place values", increasing from right to left by a factor of 2 for each place location

1 = 1 unit
10 = 2 units
100 = 4 units
1000 = 8 units
10000 = 16 units
etc.

The numeral that is in a specific place tells us how many to multiply that place value by:

1011 in binary =
1x2^3 units + 0x2^2 units +
1x2^1 units + 1x2^0 units

The calculations

- 1011 in base 2
- 1x2^3 units + 0x2^2 units + 1x2^1 units + 1x2^0 units
- = 1 x 8 + 0 x 4 + 1 x 2 + 1 x 1
- = 8 + 0 + 2 + 1
- = 11 in base 10

"There are 10 kinds of people in the world. Those who understand binary arithmetic, and those who don't."

A "byte" is a sequence of eight bits
One byte can represent up to 256 values:
0000 0000 to 1111 1111
(0 to 255 in base 10)

(it is customary when writing out bytes to separate the digits into groups of four for easier reading)

We can represent all of the alphanumeric characters, punctuation marks, symbols, and some simple formatting instructions by using just one byte of binary data for each

The ASCII (American Standard Code for Information Interchange) standard uses 7 bits of data to represent 95 characters

The eighth bit in a byte is used as a parity bit

There is an extended ASCII code for representing additional characters, e.g., characters of foreign alphabets

Memory and storage size

- Kilobyte – 1,000 bytes
- Megabyte – 1,000,000 bytes
- Gigabyte – 1,000,000,000 bytes
- Terabyte – 1,000,000,000,000 bytes

- Well, almost …..

The reality

- Kilobyte = 2^10 bytes = 1,024 bytes
- Megabyte = 2^20 bytes = 1,048,576 bytes
- Gigabyte = 2^30 bytes = 1,073,741,824 bytes
- Terabyte = 2^40 bytes = 1,099,511,627,776 bytes

Is this a problem?

- Discrepancy between advertised storage size and actual storage space
- Discrepancy between file size and storage space available
- Worse as memory storage devices have increased in capacity
  - 1,000 vs 1,024 = 2.4%
  - 1,000,000 vs 1,048,576 = ~4.86%
  - 1,000,000,000 vs 1,073,741,824 = ~7.37%
  - 1,000,000,000,000 vs 1,099,511,627,776 = ~10%

Kibi, Mebi, Gibi???

- In 2002, the Institute of Electrical and Electronics Engineers (IEEE) proposed a new naming system:
  - Kibibyte = 1,024 bytes
  - Mebibyte = 1,048,576 bytes
  - Gibibyte = 1,073,741,824 bytes
  - Tebibyte = 1,099,511,627,776 bytes

Leaving the Greek prefixes intact

- Kilobyte = 1,000 bytes
- Megabyte = 1,000,000 bytes
- Gigabyte = 1,000,000,000 bytes
- Terabyte = 1,000,000,000,000 bytes

The proposal of 2002 was given a trial period of several years and was formally adopted by the IEEE in 2005. It has also been adopted by the International Electrotechnical Commission (IEC)

The new naming system hasn't caught on in popularity, though, and many computer people still use the classical nomenclature. With the introduction of the Macintosh operating system, Snow Leopard, Apple Computer began to express disk capacity in terms of powers of 10 units

Until, or unless, the new nomenclature catches on, there will still be some confusion …

We'll use the traditional terminology in this course

1 kilobyte = 1,024 bytes, etc.

Data is entered into the computer via an input device, converted into binary code, and stored in memory

The CPU (central processing unit) extracts data items, one at a time, from memory and places the data item in a “register” (temporary storage area in the CPU) from which the data can be manipulated

The CPU can perform mathematical and logical functions on the data and can store the results of those operations in other storage registers, extracting, manipulating, and re-storing as needed

Separate registers in the CPU store the steps from a computer program that directs the CPU as to which register to extract a data item from, how to manipulate it, and what to do with the result (store it in another register, output it to an output device, etc.)

Each programming step is performed in sequence until the entire program is run, or some "glitch" occurs

The CPU has two sectors:

ROM (Read Only Memory)
RAM (Random Access Memory)

ROM (Read Only Memory)

- Permanent, can be read but not changed, added to, nor erased
- Stores essential programs, e.g. startup "bootstrap" programs and programs that need to be run quickly. ROM is built into the computer by the manufacturer

RAM (Random Access Memory)

- Can be read, changed, added to, erased
- Holds programs currently being used, data currently being used, intermediate results of computations and data manipulations

We look at the amount of RAM when we buy a computer. The larger the RAM, the more programming, and the more data, that can be stored on board the CPU at one time

Not to say you can't run a large program or manipulate large amounts of data on a machine with less RAM, but it will have to move the data back and forth between the CPU and the disk drive many times while it's working on it

That takes time. Therefore, computers with smaller amounts of RAM will perform more slowly than those with larger amounts of RAM, even if their CPUs run at the same rate of speed

We measure RAM in terms of the number of bytes of information the RAM memory can hold

Kilobytes, megabytes, gigabytes, terabytes, etc.

The size of an individually stored data item that can be accessed, as a unit, by the CPU is called a "word"

"Word" size is determined by the computer design and commonly ranges from 8 bits to 64 bits in length. The larger the word size, the more information the CPU can handle at any one time

"Word" is an outmoded term, and it's more common to hear terms like, "32-bit" or "64-bit" processor

Data in RAM is not held forever. It is "volatile". When we close a program or shut off the computer, that data is lost, unless there is an automatic back-up application running

Store it, often!!!

- If we want to save the data we need to write it onto some form of non-volatile storage (hard disk, CD, DVD, thumb drive, etc.)

Input devices

- Keyboard
- Mouse w onscreen menu (text and/or icons)
- Joystick
- Touch screen
- Disk drive
- Connection ports (e.g., USB)
- Wireless connections (e.g., Bluetooth)

Output devices

- Screen
- Printer
- Speaker
- Hard disk
- Connectivity ports (e.g., USB)
- Slots
- Wi-fi (e.g., Bluetooth)

Storage media

- Punched tape
- Punch cards (Hollerith cards)
- Magnetic tape (reel-to-reel, cassette)
- Floppy disks
- Firm disks (hard floppies) and Zip disks
- CD (CD-R, CD-ROM, CD-RW)
- DVD (DVD–R, DVD+R, DVD–RDL, DVD+RDL, DVD–RW, DVD+RW)
- Flash memory (memory chips, sticks, thumb drives)
- “the cloud”

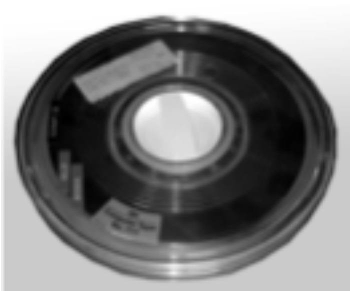
Punched (perforated) paper tape

http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Hollerith card

http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Magnetic tape

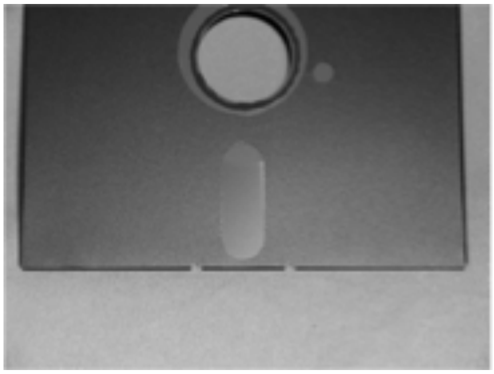


http://www.ricomputermuseum.org/Home/collections-gallery/the-history-of-computer-data-storage

Floppy disk – in its protective sleeve

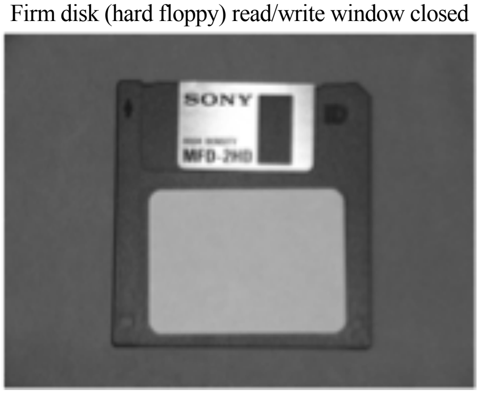
Floppy disk – and protective sleeve

Floppy disk – read/write opening

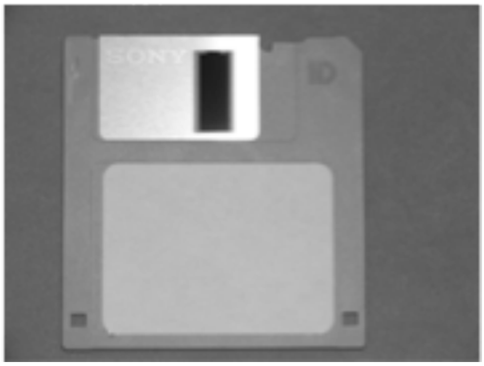


Firm disks (hard floppies)

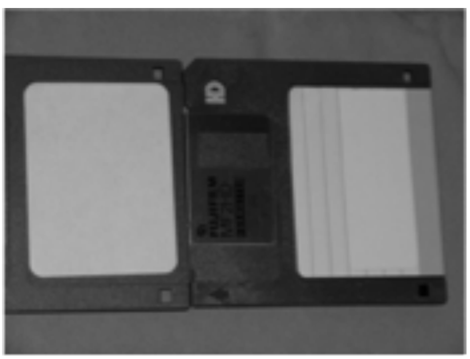
Firm disk (hard floppy) read/write window closed



Firm disk (hard floppy) read/write window open



Firm disks showing write protect tabs



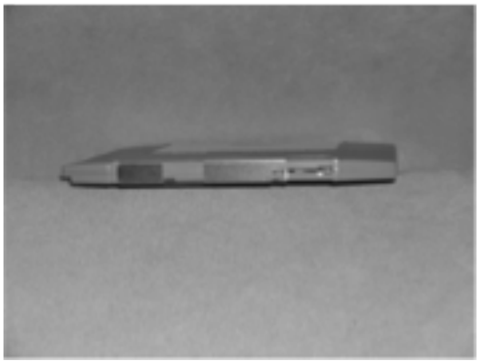
Firm disk write protect tab open

Firm disk write protect tab closed

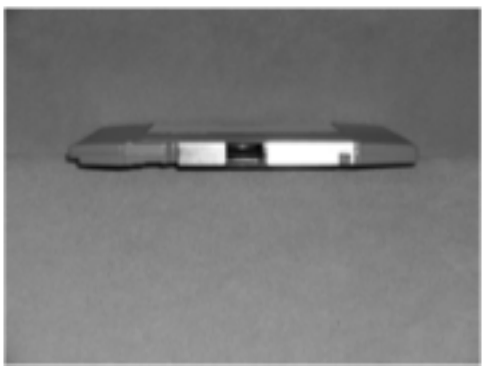
Zip Disk

Zip Disk

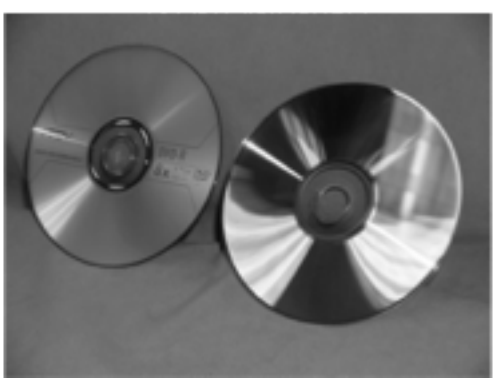
Zip Disk read/write window closed



Zip Disk read/write window open

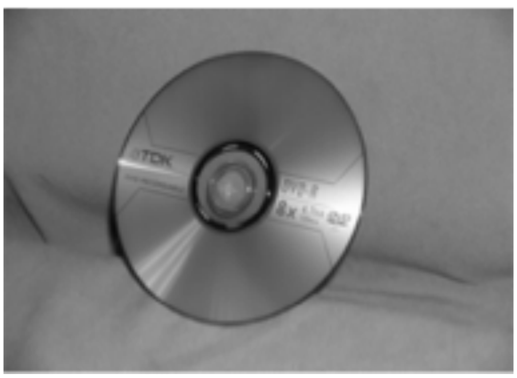


DVD (CD's look the same)

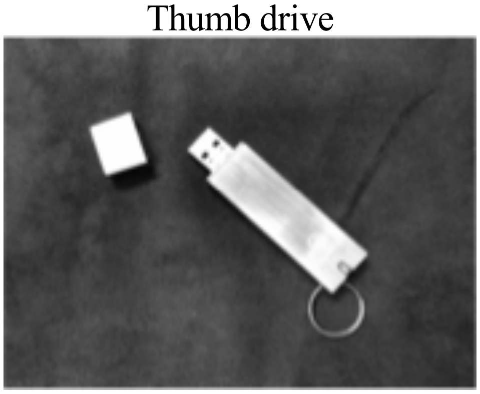


CD – 700 MB capacity

DVD – 4.7 GB capacity



Thumb drive



The Cloud

Programming

- Machine language - instructions to the CPU, written as sequences of binary digits (0's, 1's)
- Assembly language - acronyms and abbreviations that correspond to machine code but are easier to understand:
- LDA = load
- ADD = add
- STA = store

Assembly language commands need to be **very** specific

Assembler - translates assembly code into machine code for the CPU; works at the level of the CPU and deals with things like stacks and registers

Higher level languages

- Easier for humans to understand, compared to machine or assembly language
- Allow more complex operations in one or a few words (e.g., “PRINT”, “GO TO”, “SAVE”, “END”). Each of those commands may represent hundreds of machine language instructions

Compiler

- Translates a high level language program into machine code
- Translates all of the programming statements, creates a machine language, “compiled”, version of the program, then runs the program
- The compiled (machine language) version can be saved and run directly. The original program does not have to be “re-compiled” each time you want to run it.

Interpreter

- Converts high level language to machine code
- Converts and executes each statement individually before continuing
- The interpreter has to be run each time the program is run
- Does not create a separate, stand-alone, program
- Does not require as much pre-processing compared to using a compiler
- Some programming languages are compiled, others are interpreted

Programming languages, like verbal languages, have vocabulary, syntax, "grammar", and semantics

Different programming languages were developed in response to needs, perceived "problems" with previous languages, or to handle specific types of problems or situations (math, engineering, accounting, medicine, graphing, communications, database creation and manipulation, etc.)

Some programming languages

- COBOL (**Co**mmon **B**usiness **O**riented **L**anguage)
- FORTRAN (**For**mula **Tran**slation/**Tran**slator)
- BASIC (**B**eginner's **A**ll-purpose **S**ymbolic **I**nstruction **C**ode)
- PASCAL (named for the 17th century French mathematician, Blaise Pascal)
- PERL (**P**ractical **E**xtraction and **R**eporting **L**anguage)
- C
- C++
- Java (not an acronym)
- UNIX
- LINUX

Operating system (OS)

- The operating system is the program that controls the computer, other hardware, and controls the execution of programs
- Windows OS
- Mac OS

Programs written for one operating system sometimes can't be run on computers that use a different operating system. In the early days they definitely couldn't; now it's somewhat better. E.g., documents created with Microsoft (MS) Word for Windows can be read and worked on using MS Word for Mac

Cross platform programming

MS Office
Java

Communications

- Direct connection (e.g., Ethernet)
- Phone line
- T1 line
- Fiberoptic line
- Wireless; uses electromagnetic (radio) waves (e.g., Wi-fi, Bluetooth)

Digital phone lines exist

- ISDN (Integrated Services Digital Network)
- DSL (Digital Subscriber Line)
- T1 line
- ADSL (Asymmetric Digital Subscriber Line)
- T3 line
- ATM (Asynchronous Transfer Mode)
- Cable

- These provide higher transmission speeds than the standard analog phone lines and with a lower error rate

Rate of transmission

- Data is transmitted at rates of speed measured in “bits / second” (bps)

- Analog phone = 56 kbps
- ISDN = 1.5 Mbps
- DSL = up to 52 Mbps downstream and 128 kbps upstream
- ADSL = up to 8 Mbps downstream and up to 1 Mbps upstream
- T1 = 1.544 Mbps
- T3 = 44.736 Mbps
- ATM = 600 Mbps to 10 Gbps
- Cable = 1.5 Mbps to 35 Mbps downstream

Networking

- Ethernet – directly connected computers in close proximity
- LAN – Local Area Network
- WLAN – Wireless Local Area Network
- WAN – Wide Area Network
- MAN – Metropolitan Area Network
- CAN – Campus Area Network

Internet

- The internet is a WAN that connects over backbone links to multiple regional networks. The internet is an evolution of ARPANET
- The regional networks are now commercial and connect to local networks

Internet cont'd

- Other WANs exist and are national and governmental. They connect to the internet and to each other via gateways
- ISPs (Internet Service Providers) - e.g., AOL and Earthlink, as well as those that are provided by other commercial companies, organizations, and academic institutions, that serve the individual user and connect to the regional networks

Internet communication standards

- TCP/IP - Transmission Control Protocol / Internet Protocol; structured message packets; each machine or device on the Internet has its own, unique, address
- The IPv4 (Internet Protocol, version 4) address system was launched on January 1, 1983

IPv4

- IPv4 addresses – each address is a 32-bit number, written as four 8-bit numbers, each separated by "." Each 8-bit number is usually written as its base-10 (decimal) equivalent (e.g., 215.35.69.3). Sometimes you may see an IPv4 address written in its binary format: 11010111.00100011.01000101.00000011 is the binary representation of 215.35.69.3
- The maximum number of addresses that you can have with a 32-bit number is 2^32, or 4,294,967,296
- The Internet Assigned Numbers Authority (IANA) coordinates the internet address system and assigns unallocated numbers to each of the five Regional Internet Registries (RIR)

Regional Internet Registries

- AfrinNIC (African Network Information Centre): Africa
- APNIC (Asia-Pacific Network Information Centre): Asia / Pacific
- ARIN (American Registry for Internet Numbers): North America
- LACNIC (Latin American and Caribbean Network Information Centre): Latin America and some Caribbean islands
- RIPE NCC (Réseaux IP Européens Network Coordination Centre): Europe / Middle East / Central Asia

Unallocated internet addresses

- As of January 2010, 90% of the IPv4 addresses had been assigned, leaving only 10% (approximately 400 million) unallocated
- While that seems like a large number, remember, that means that 90% of the addresses had been assigned in less than 30 years (since 1983 when the IPv4 addresses were first launched)
- Enter, IPv6!

Internet communication standards, IPv6

- IPv6 (Internet Protocol, version 6)
- Developed in 1995; started being used in 1999; its world-wide launch date was June 6, 2012
- Increased each address from 32 bits to 128 bits, allowing for 2^128, or about 3.4028 x 10^38 individual addresses

IPv6 addresses

- IPv6 addresses are written as eight groups of four hexadecimal numbers, with each group separated from the others by a colon (:)
- A group that consists of all “0”s, can be omitted from the address, with a double colon (::) indicating the missing group, or consecutive groups of all zeros
- 2073:adef:ecda:0000:0000:0000:28ce:843d can be written as 2073:adef:ecda::28ce:843d
- However, 2073:0000:adef:ecda:0000:0000:28ce:843d cannot be written as 2073::adef:ecda::28ce:843d, with two “double colon” indicators, because you wouldn’t be able to tell how many groups of zeros each double colon replaced

Hexadecimal (a short diversion)

- Just as binary is a numbering system using base-2, hexadecimal is another numbering system that uses a base other than 10. In hexadecimal, the base is 16. That is, there are sixteen individual symbols for the numbers from 0 to 15 (compare the base-10 system that has 10 individual symbols for the numbers from 0 to 9)

Hexadecimal

- We can use the first ten symbols: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 to represent their customary values
- We use the first six letters of the alphabet as symbols to represent the remaining digits in the hexadecimal system:
- a = 10, b = 11, c = 12, d = 13, e = 14, f = 15
- The letters can be written either in upper or lower case

Hexadecimal

- Converting from hexadecimal to base-10 is similar to converting from binary to base-10, except that instead of raising '2' to the appropriate power, we raise '16' to that power
- E.g.: 4be2 = (4 x 16^3) + (11 x 16^2) + (14 x 16^1) + (2 x 16^0)
- = (4 x 4,096) + (11 x 256) + (14 x 16) + (2 x 1)
- = 16,384 + 2,816 + 224 + 2
- = 19,426

Back to internet communication

- Devices that are permanently connected to the internet have fixed IP addresses. Those that go through an ISP have temporary addresses assigned for that connection session

IP addresses are converted to named addresses for convenience

Name servers - computers that convert the user name to an IP address

Routing takes place based on the numeric IP address

You can connect to an address via the numeric IP address **or** the name

IP names
(e.g., elearn.ttuhsc.edu)
are read from right to left, from the more general to the more specific

Top level domain names

- .com - commercial
- .gov - governmental
- .edu - educational
- .net - network
- .mil - military
- .org - non-profit organization

Geographical designations

- .uk – United Kingdom
- .it – Italy
- .fr – France
- .nz – New Zealand
- .ru – Russia
- Etc.

Network communications

- Computers differ
- Operating systems differ
- Transmission lines differ
- Communication protocols differ

- There needs to be some standardization if data is to be shared between computers

Network service stacks and network protocols handle the compatibility issues

Network service stacks - organize software within a machine. Network stacks are machine and OS dependent

Network protocols - handle the communications between machines

Network protocols

- Standardize communications between machines
- Standardize ways of representing data, requesting an action, replying to a request
- Different protocols for different network functions

Examples of network protocols

- SMTP - Simple Mail Transport Protocol
- FTP - File Transfer Protocol
- HTTP - Hypertext Transfer Protocol (web pages)
- ICMP - Internet Control Message Protocol
- MIME - Multipurpose Internet Mail Extension (for attachment of multimedia to email)

We've come a long way, but definitely NOT the end of computers!

References

- Abacus
  - http://www.youtube.com/watch?v=wIiDomlEjJw&feature=related
  - http://www.youtube.com/watch?v=8jsn9ccJ-5U&feature=related
- Character codes (more than you'd ever want to know)
  - http://tronweb.super-nova.co.jp/characcodehist.html

More references

- Bytes and more
  - http://www.t1shopper.com/tools/calculate/
- IP addresses
  - http://www.nro.net/news/less-than-10-of-ipv4-addresses-remain-unallocated-says-number-resource-organization
  - https://www.arin.net/knowledge/ipv6_info_center.html
- Domain names
  - www.iana.org/domain-names.htm

End of the lesson

Computers