Comuter Science Concepts for Data Science - Notes

Table of contents

Introduction

How can Data Science improve Products & Services?

Data Science improves traditional productio nwith basic function by Digital services using Data Science driven functions. The result is a new poroduct type, for Example watch vs. smart watch. To do so, data has to be processec, visialized, tranferd storaged and used in a program.

Computer Network

Fundaments of Computer Networks

A computer network is build of several components:

- Connected devises: Hosts (end systems).
- Communications links: Wired and wireless.
- Protocols: Sending and receiving data.
- Packet switches: routers and switches.

Example Net: Enterprise Network (Ethernet)

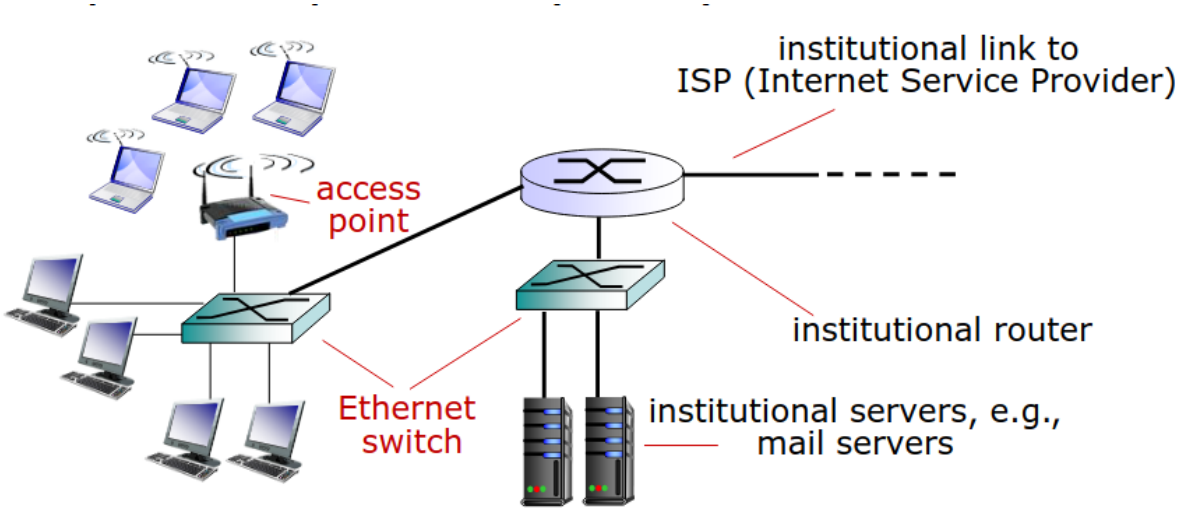


Figure 1: Enterprise Network

Important basic terms explained briefly

- LAN (Local Area Network): A local network, e.g., your home network or the network in an office building.
- WAN (Wide Area Network): A wide-ranging network that connects cities or countries. The internet is the largest WAN in the world.
- IP address: Your device's "address" on the network. Without this address, data packets would not know where to go.
- Router: The "traffic cop." It connects your local network (LAN) to the global network (Internet) and forwards the data to the correct device.

Packet Switching

The Host dates applications messages and breaks them into smaller chunks (aka. packets) of length L bits. The packets are transmitted into access network at rate R, called link capacity/bandwidth.

Packaet Transmission Delay

This is the time required by the transmitter to push all bits of a data packet onto the cable (or medium).

- L: Packet length (bits).
- R: Link bandwidth (dps).

$$d_{\text{trans}} = \frac{L \quad \text{(bits)}}{R \quad \text{(bits/sec)}}$$

> **Note**
>
> Also called link capacity or link bandwidth.

Propagation delay

This is the time it takes for a single bit to physically travel from point A to point B.

- d: Length of physical link.
- s: Propagation speed in medium.

$$d_{\text{prop}} = \frac{d}{s}$$

Rate

Rate (bits/time unit) at which bits transferred between sender/receiver. We disquinsh between:

- Instantaneous: Rate at given point in time.
- Average: Rate over longer period of time.

Protocol Layers

Protocols define format, order of msgs sent and received among network entities, and actions taken on msg transmission, receipt.

- Applications: FTP, SMTP, HTTP.
- Transport: TCP, UDP.
- Network: IP, routing protocols.
- Link: Ethernet, WiFi.
- Physical: Ethernet, WiFi.

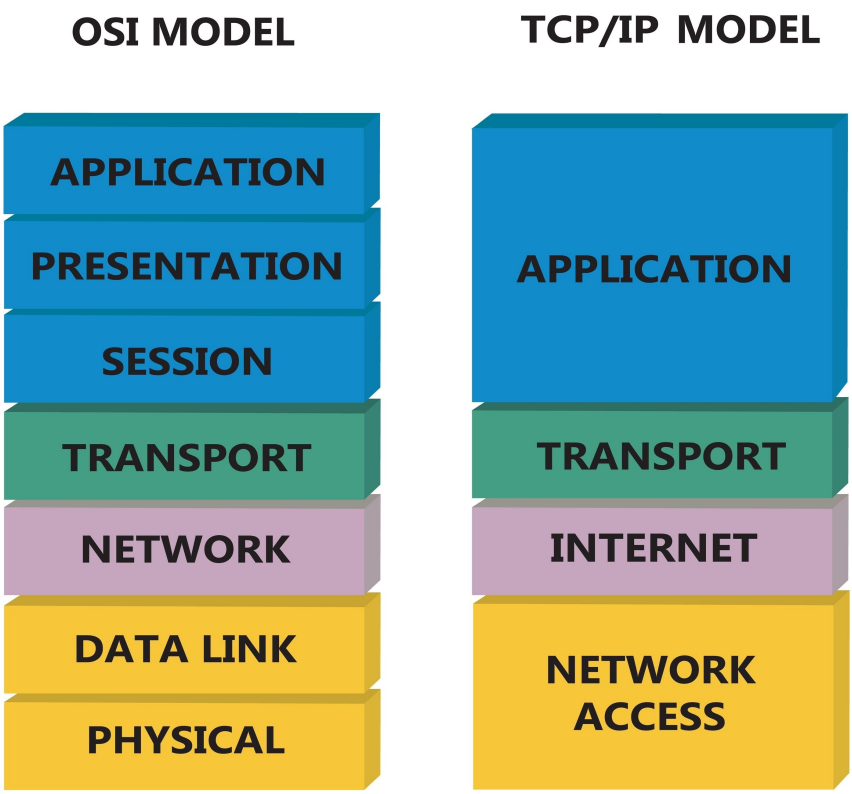


Figure 2: OSI Model & TCP/IP Model

Advantages and disadvantages of Protocol Layers

Advantage

- Flexibility: Individual parts are replaceable.
- Simplicity: Focus on a specific task.
- Interoperability: All devices understand each other.

Disadvantage

- Overhead: Data packets become larger due to multiple headers.
- Performance: “Packaging” and “unpackaging” takes time.
- Duplication: Some functions are performed multiple times.

Addressing in Computer Networks

Domain Name System

Translates domain names to numerical IP addresses.

- DNS Recursor (Client Side): Initiates the query. It’s the first stop and acts as a librarian trying to find the IP address.
- Root Nameserver (The Dot “.”): The starting point. It directs the Recursor to the correct Top-Level Domain (TLD) Nameserver.
- TLD Nameserver (e.g., .com, .de, .org): Manages all domains under its specific extension. It points the Recursor to the Authoritative Nameserver for the specific domain requested.
- Authoritative Nameserver: Holds the definitive DNS records (A, CNAME, MX, etc.) for the requested domain. It provides the actual IP address back to the Recursor.
- DNS Recursor / Client: Receives the IP address and finally connects to the web server to load the website.

Domains

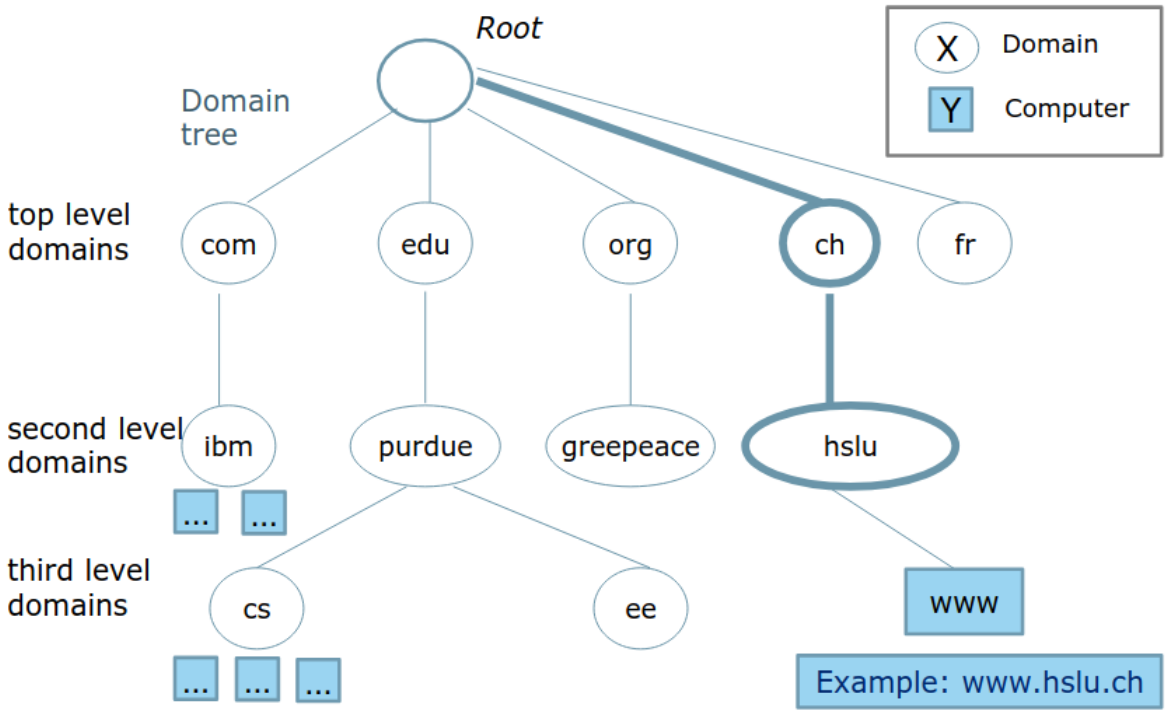


Figure 3: Domain Tree

IP Addressing

Each device is assigned a unique IP address build up by `Net ID` + `Host ID`

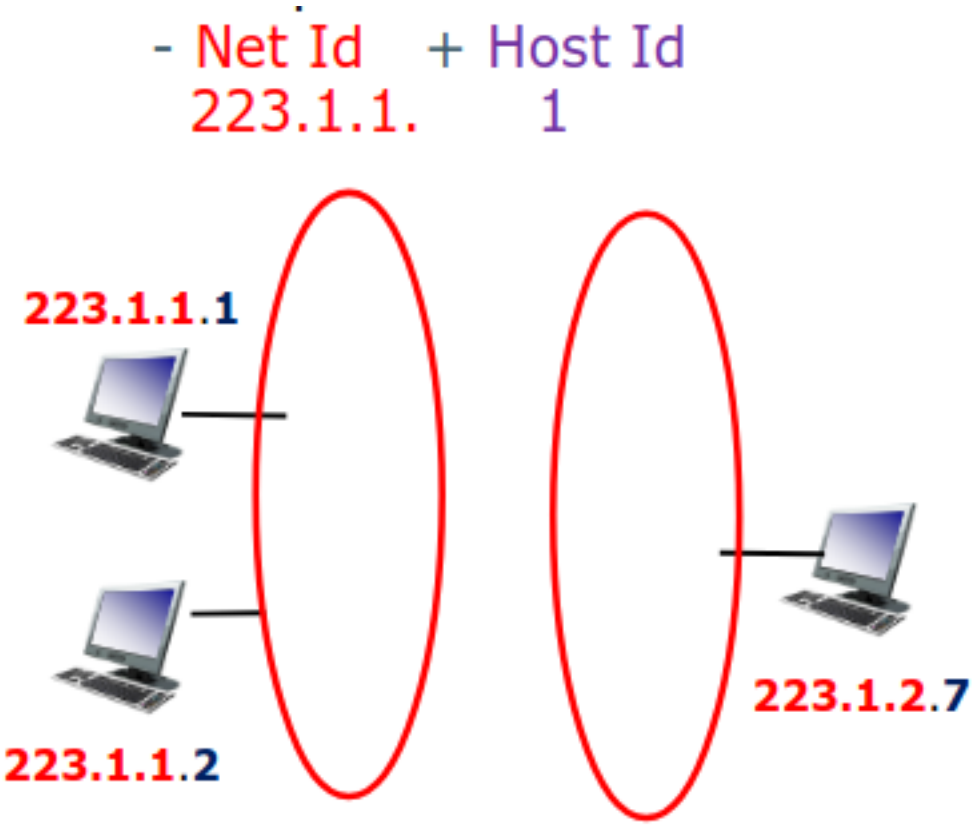
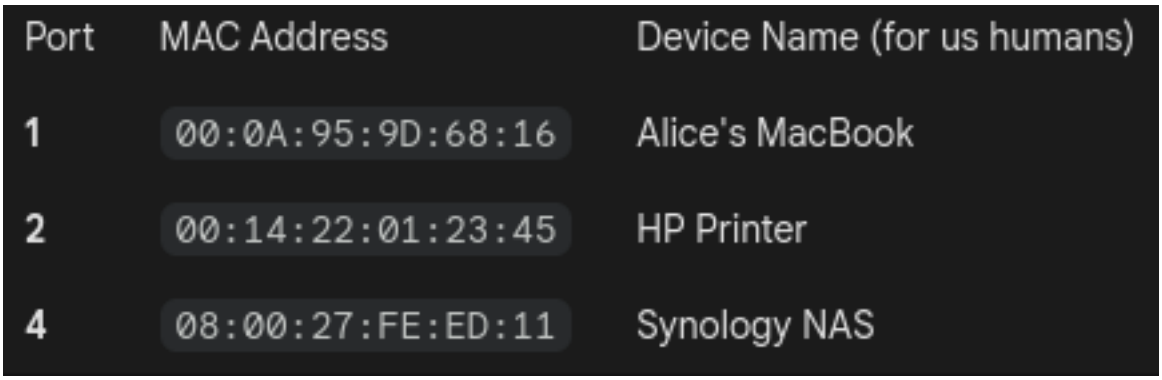


Figure 4: IP Addressing in Computer Networks

Switch

A network switch is a central component in almost every wired network, especially in local area networks (LANs). It acts as an intelligent distribution station for data traffic. Based on incoming frames switch "learns" location of sender. Records sender/location pair in switch table.



Port	MAC Address	Device Name (for us humans)
1	00:0A:95:9D:68:16	Alice's MacBook
2	00:14:22:01:23:45	HP Printer
4	08:00:27:FE:ED:11	Synology NAS

Figure 5: Example Switch Table

Note: Layer 2 – Data Link Layer

Router

A router is a network device that forwards data packets between different networks. It ensures that information (e.g. a website, video or email) reaches the recipient from the sender – even if it has to pass through many intermediate stations.

Note: Layer 3 – Network Layer

Week 3: Case Study – Web Applications & HTML

Client Server Communication Pattern

Clients (e.g. Browser) send service requests. Servers (e.g. Web, Database) wait for requests to arrive from clients and then respond to. Client receive service responses from centralized server.

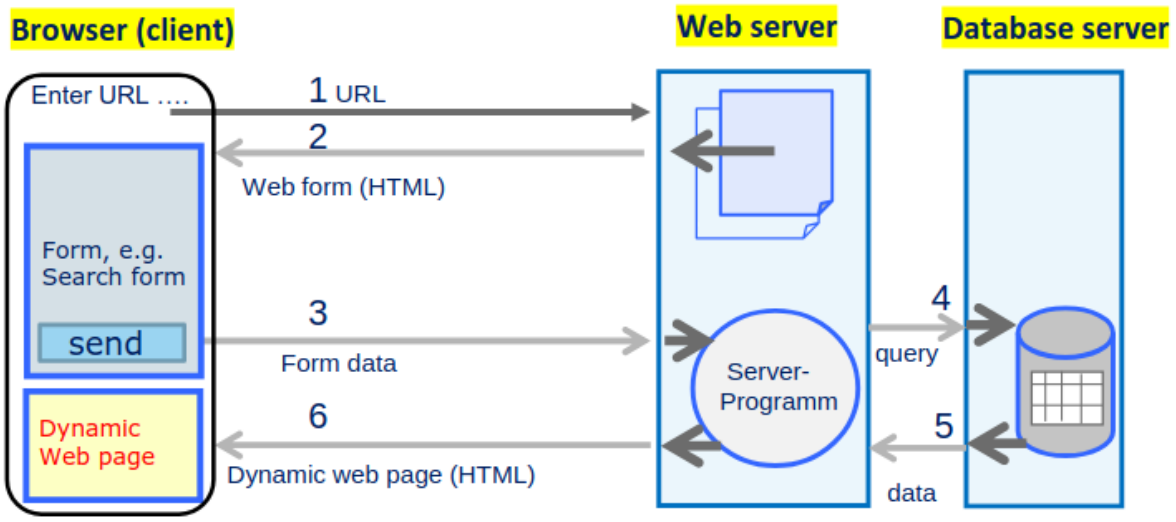


Figure 6: Client - Server Communication

Uniform Resource Locator (URL)

Reference to a web resource that specifies its location in a computer network (e.g. web page, video, image, etc.)

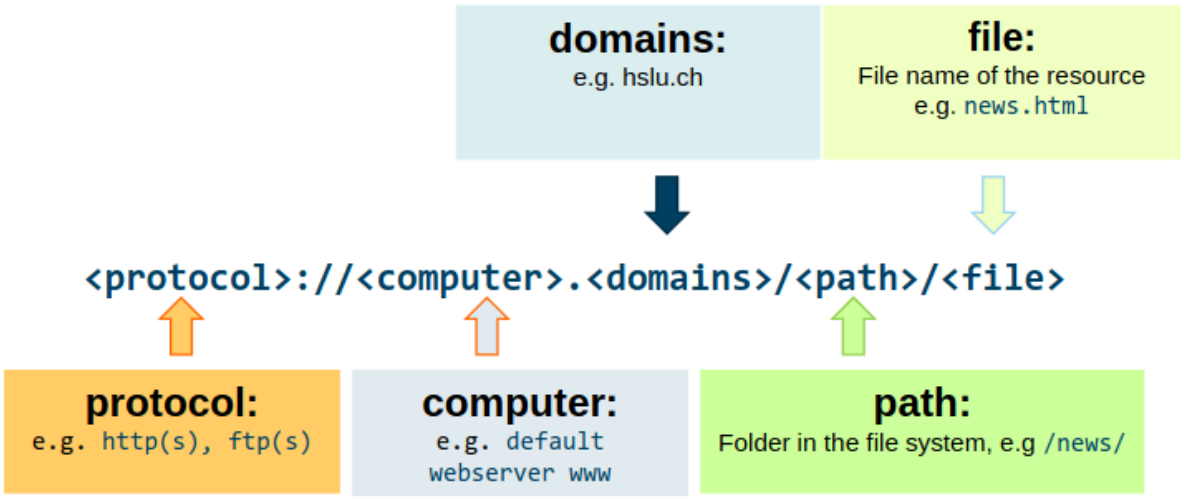


Figure 7: URL Structure

HTML

Descriptive Language HTML, with a set of HTML elements/tags, is used to create Websites and publish content on the Internet in a simple and standardized way. The Hypertext Transfer Protocol (HTTP) ensures that content on a Web server can be accessed and interpreted by a Web client and thus displayed within a browser.

Elements & Tags

Most HTML Elements are characterized by having a start and end tag.

```
<h1>Header 1</h1>
<p>Paragraph</p>
```

- Headings: Headings are text size pre-sets ranging from `<h1>` to `<h6>`
- Paragraphs: Paragraphs are used to create text blocks with `<p>content</p>`
- Bold: With `<b> text </b>` text can be set to bold.
- Italic: With `<i> text </i>` text can be set to italic.
- Underline: With `<u> text </u>` text can be underlined.
- Hyperlinks: Using the anchor tag with an link as attribute: `<a href="Reference target"> Text </a>`
- Images: Image tag with at least a source attribute: `<img src = "path/file" >`

Nested Elements

It is possible to include elements within elements. Such constructs are being referred to as nested elements.

```
<p>Star<b>Wars</b></p>
```

Structural Elements

The structural elements can be considered as the core setup of any HTML Website.

- `<!DOCTYPE html>`: Declares using HTML5
- `<html> </html>`: Defines the root of an HTML document
- `<head> </head>`: Meta data about the HTML document
- `<body> </body>`: All content of the HTML document

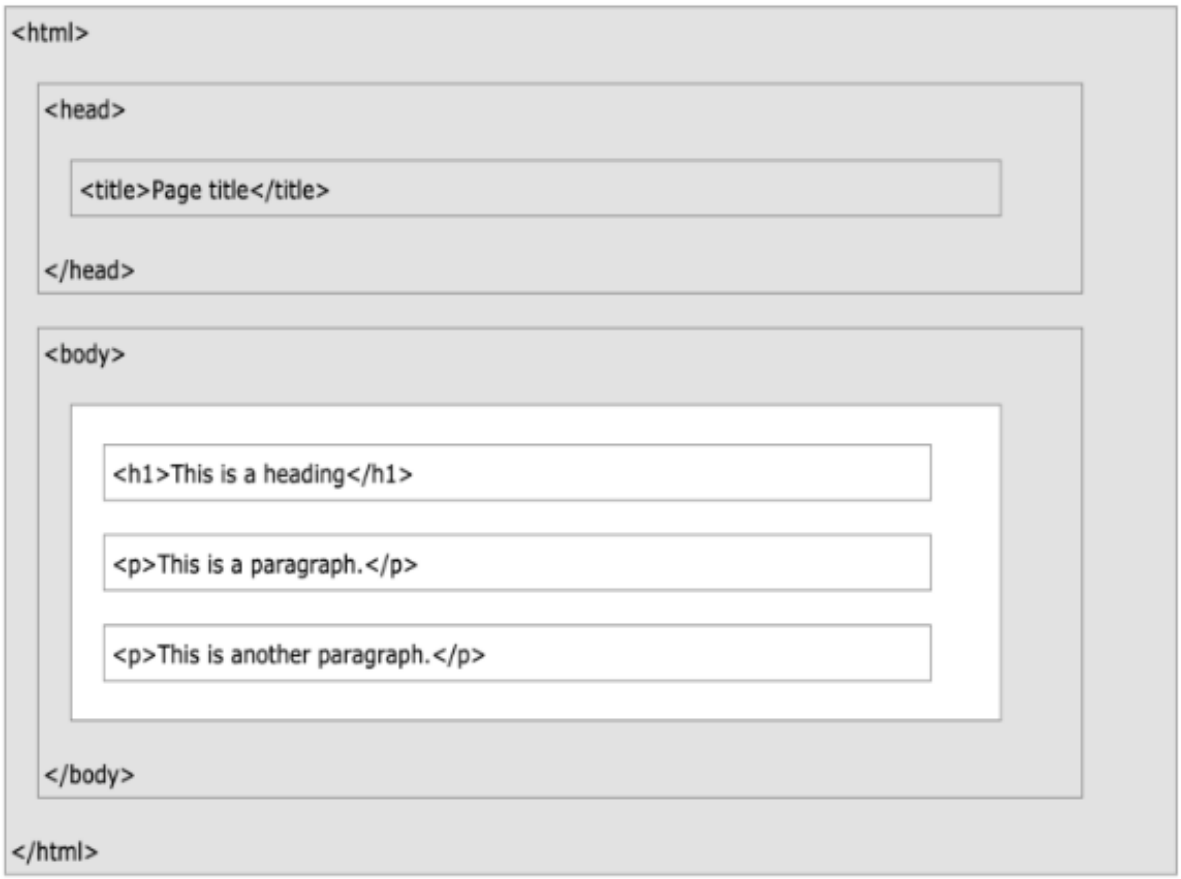


Figure 8: Example HTML Site

CSS

CSS is a stylesheet language that is primarily used to describe the appearance and formatting of a document written in a markup language such as HTML.

```
header {
    color: #0000FF;
    font-size: 26px;
    text-align: center;
}

#image_medium {
    height: 331px;
}

footer {
    text-align: center;

}
```

Figure 9: CSS Code

```
<html>
    <head>
        <title>Star Wars Coffee</title>
        <link rel="stylesheet" href="starwars.css" type="text/css">
    </head>
    <body>
            <header>The force is strong with this coffee!</header>
            <img id="image_medium"
              src="http://3.bp.blogspot.com/_2O5gTwhJBps/TLOUxfFfZA
            <footer>Come to the dark side, we have cookies!</footer

    </body>
</html>
```

Figure 10: CSS Usage

Week 4: Algorithms & Programming Language Concepts

Algorithms

A computational problem specifies an input-output relationship. An algorithm is an exact specification of how to solve a computational problem. Algorithms must be:

- Correct: For each input produce an appropriate output
- Efficient: Run as quickly as possible, and use as little memory as possible

Designing Algorithms

Break problem up into smaller (easier) sub-problems. Think only about how to use the smaller solution to get the larger one Do not worry about how to solve to smaller problem (it will be solved using an even smaller one).

Recursion

Recursion is a method of solving a problem where the solution depends on solutions to smaller instances of the same problem. This allows programming in a style that reflects divide-n-conquer algorithmic thinking.

Search Algorithms

Linear search or sequential search is a simple method for finding an element within a sorted list. It sequentially checks each element of the list until a match is found or the whole list has been searched. A linear search runs in at worst linear time and makes at most n comparisons.

Binary Search

Binary search is a search algorithm that finds the position of a target value within a sorted list of values. Binary search compares the target value to the middle element of the list. If they are not equal, the half in which the target cannot lie is eliminated and the search continues on the remaining half, again taking the middle element to compare to the target value, and repeating this until the target value is found.

Week 5:Algorithms & Programming Language Concepts

Low Code Environments

Combines visual program elements like variables, loops or conditional statements rather than specifying them textually.

Week 6: Visualization

Data Types

Operation	Nominal	Ordinal	Interval	Ratio
Equality	X	X	X	X
Order		X	X	X
Add / Subtract			X	X
Multiply / Divide				X
Mode	X	X	X	X
Median		X	X	X
Arithmetic Mean			X	X
Geometric Mean				X

Figure 11: Data Types

Visual Variables

- Mark / Shape: Used shapes to distinguish between different types in a plot
- Position: Position of an object Used to encode data
- Size (Length, Area and Volume): Sizes to encode data
- Brightness: Mapping values to the plot
- Color: Colormap that can be used to encode a data variable
- Orientation: Used to encode data
- Texture: Mapping values to the plot
- Motion: Associated with any of the other visual variables

Design Principles

> **Note**
>
> The purpose of visualization is insight, not pictures.

- Do not expect the user to spot changes in interactive diagrams

- Overview first, zoom and filter, then details on demand
- Similarity in feature space corresponds to similarity in visual space
- Order between data items corresponds to visual order
- Choose your coloring wisely
- The importance of data items corresponds to their visual saliency
- A single visualization can only reveal one data facet
- Visualize only the interesting information

Week 8: Cloud Computing & Virtualization

Central Processing Unit (CPU)

The CPU is the primary component of a computer that acts as its "brain," responsible for interpreting and carrying out most of the commands from the hardware and software. It performs basic arithmetic, logic, and input/output operations to process data and manage the system's overall performance.

Fetch

In this initial stage, the CPU retrieves a program instruction from the computer's memory (RAM). The instruction's location is identified by a program counter, which then prepares the processor for the next step in the cycle.

Decode

Once the instruction is fetched, the control unit breaks it down into signals that the CPU's internal components can understand. This process determines what action is required, such as performing a math calculation or moving data between registers.

Execute

During the final stage, the CPU carries out the decoded instruction by performing the necessary operations. The result is then stored in a register or written back to the memory before the cycle begins again.

Random Access Memory (RAM)

RAM is a form of high-speed, volatile storage that holds the data and instructions currently being used by the CPU. Unlike a hard drive, it allows for nearly instantaneous access to information, though all stored data is lost once the computer is powered off.

Cloud Computing Definition

- Gossmann (2009): Clouds, or clusters of distributed computers, provide on-demand resources and services over a network, usually the Internet, with the scale and reliability of a data center.
- NIST Definition: Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or serviceprovider interaction.

Business Model

- Client: Pay by use instead of provisioning for peak.
- Provider: Share capabilities.

Cloud Properties

- Self-managing services: A consumer can provision cloud services as needed and automatically without requiring human interaction with each service's provider.
- Accessible and highly available: Cloud resources are available over the network anytime and anywhere.

Deployment Model

- Public clouds: Owned by cloud service providers who charge for the use of cloud resources (AWS (Amazon), Azure (Microsoft))
- Private clouds: Cloud infrastructure belongs to and is operated by only one organization
- Hybrid clouds: Cloud infrastructure is a composition of two or more clouds (private or public). Bound together by standardized or proprietary technology that enables data and application portability.

Infrastructure as a Service (IaaS)

IaaS provides the fundamental building blocks of cloud IT, offering on-demand access to virtualized servers, storage, and networking. It gives you the highest level of flexibility and control over your IT resources, as you are responsible for managing the operating system, middleware, and applications.

Platform as a Service (PaaS)

PaaS removes the need for you to manage underlying infrastructure (hardware and OS), allowing you to focus solely on the deployment and management of your applications. It typically includes the application stack, development tools, and database management systems, which helps increase developer efficiency.

Software as a Service (SaaS)

SaaS provides a complete, ready-to-use software product that is run and managed by the service provider. You don't have to worry about how the service is maintained or how the underlying infrastructure is managed; you simply access the application via a web browser or API.

Comparison

Feature	IaaS	PaaS	SaaS
You Manage	OS, Apps, Data	Applications, Data	Nothing (just usage)
Provider Manages	Hardware, Network	Hardware, OS, Stack	Everything
Example	AWS (EC2), Azure	Google App Engine	Gmail, Salesforce

Figure 12: Comparison Summary of Cloud Services

Virtualization

Virtualization is essentially a way to trick a computer into thinking it is running multiple separate machines, all on the same physical hardware. It lets you run many independent operating systems (called Virtual Machines or VMs) side-by-side. The piece of software that makes this possible is called a Hypervisor.

There are two main types of Hypervisors:

```
- Type 1 (Bare-Metal): This is the most powerful type. It is installed directly onto the phys

- Type 2 (Hosted): This type is installed like a regular application on top of a host operati
```

Key Advantages of Virtualization

- Efficiency & Consolidation: Run multiple virtual servers on a single physical machine to maximize hardware utilization.
- Cost Savings: Significantly reduces expenses for hardware, power consumption, and cooling infrastructure.
- Isolation & Security: Enhances security by isolating VMs; if one system is compromised, the others remain unaffected.
- Disaster Recovery: Simplifies backups and recovery by allowing entire VMs to be moved or restored across different hardware instantly.

Key Disadvantages of Virtualization

- Single Point of Failure: If the underlying physical server (the host) fails, all virtual machines running on it will crash simultaneously.
- Performance Overhead: Because the hypervisor requires resources to manage the VMs, there is a slight performance loss compared to running on "bare metal."
- Complexity: Managing a virtualized environment requires specialized software and expertise, which can increase administrative overhead.
- Resource Contention: If too many VMs are running, they may compete for the same CPU or RAM, leading to "bottlenecks" and slowed performance.

Week 9: Big Data Processing

Processing Data

> **Note**
>
> All Linux commands consist of three parts:
>
> - command (e.g., grep)
> - options: to modify the behavior of the command
> - arguments: objects upon which the command acts

grep

Take a stream of text or data, perform operations on it, and produce a modified version of that stream as output.

```
grep -e <REGEX_PATTERN> <FILE>
```

cut

Selects portions of each line (as specified by list) from each file and writes them to the output.

```
cut -d <DELIMITER> -f <FIELD_SELECTION> <FILE>
```

wget

Wget is a free utility for non-interactive download of files and whole websites (follows links) from the Internet.

Pipes & Filters Pattern

A pipe is used in Linux to send the output of one command/process to another command/process for further processing. Decompose a task that performs complex processing into a series of separate processing steps that can be reused.

```
grep -e <REGEX_PATTERN> <FILE> | cut -d <DELIMITER> -f <FIELD_SELECTION> <FILE>
```

5V-Model

The 5 V's of big data are Volume, Velocity, Variety, Veracity, and Value. These characteristics help define the challenges and opportunities associated with managing and analyzing large data sets.

Regular Expressions

A regular expression, often shortened to regex, is a sequence of characters that defines a search pattern. It uses a specific syntax combining literal characters (e.g., a, 1) and metacharacters (e.g., . , *, []) which have special meaning.

Essetials

- Metacharacters: These are the building blocks, defining rules like matching any character (.), the start of a line (^), or the end of a line ($).
- Quantifiers: Symbols like * (zero or more), + (one or more), and ? (zero or one) specify how many times the preceding element must occur.
- Character Classes/Sets: Square brackets [] define a set of characters, matching any one character within the set (e.g., [0-9] matches any digit).
- Anchors: Characters like ^, $, and \b (word boundary) match a position within the text, not an actual character.
- Grouping: Parentheses () are used to combine parts of the pattern and create capturing groups, allowing the extraction of specific matched sub-strings.
- Applications: Primarily used for data validation (e.g., emails, phone numbers), complex search and replace operations, and parsing/extracting data from text.

```
def f(n1, n2, delta):
    if n1 <= n2:
        return n2
    return n1 * f(n1 - delta, n2, delta)

print(f(35, 15, 5))
```

```
7875000
```