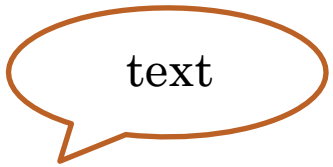
Week 12: Data/information storage formats

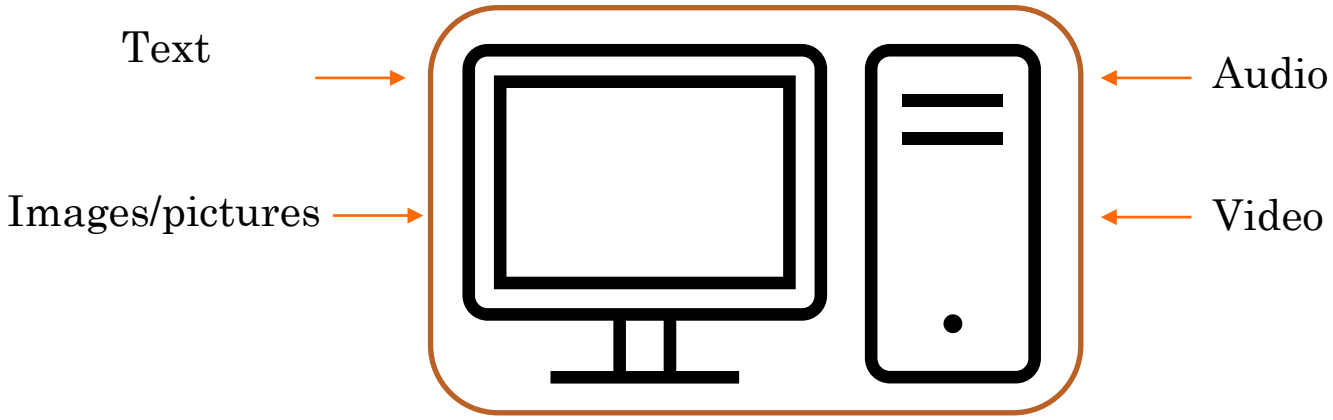
Learning objectives

- To know different types of data storage formats
- ASCII Vs Unicode

At the conclusion of this lecture, students will be able to understand the fundamental data storage formats that have been widely deployed in computers for data storage/retrieval



Data storage formats

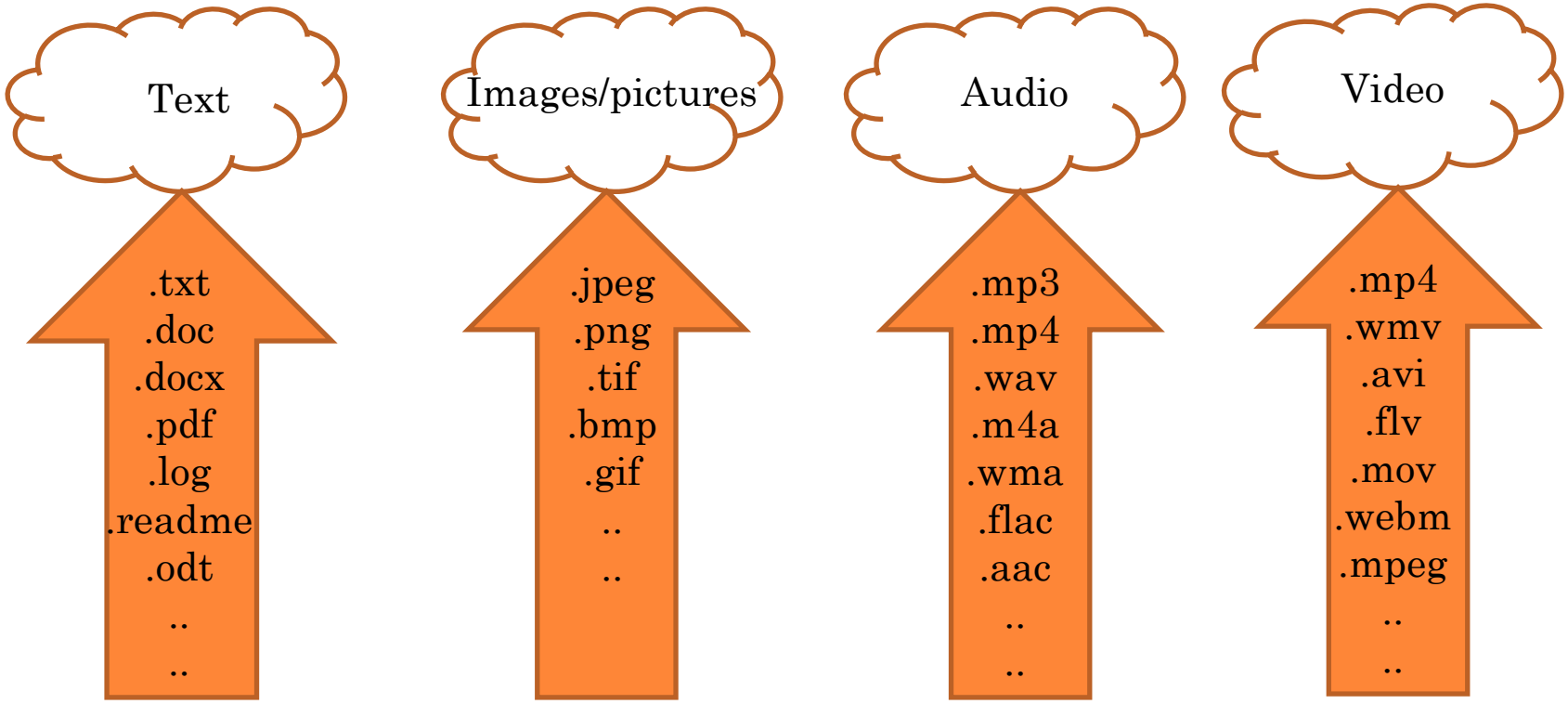


Encoding to binary / machine understandable format

The type of data stored in the memory are distinguished by its extension name at source level. Example: .txt, .png, .mp3, mp4… (file formats)

File formats defines how a binary file will be structured. It is normally depending on OS that contains programs/modules for easy access: **.docx**– Windows / **.pages** -Mac

How are different type of data handled in real world applications and for storage and retrieval purposes?



Structured data : data are organized in the form of row and columns that are related with each other . Example: relational database, Excel files

Semi-structured data: The data that does not follow strict structured database framework but still have some structural link between its properties. Example: XML files

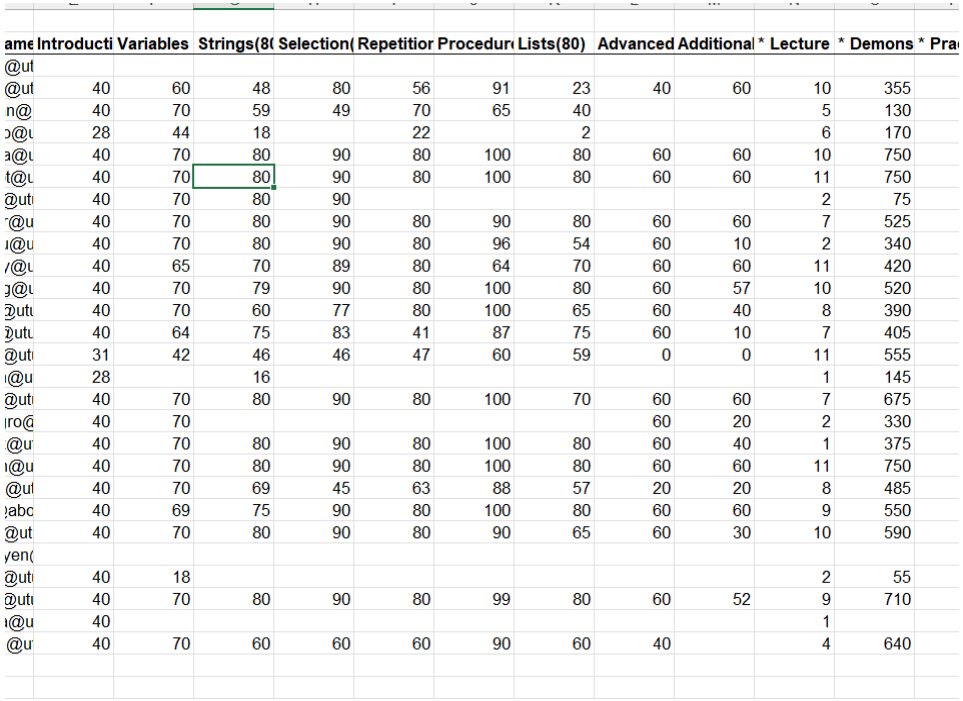
Unstructured data: The data that can't have any structural database framework or tabular format standard. Example: video, audio files, emails, social media contents..

Common data formats that used for storage in big data analytics

houseLahti.txt - Notepad

File Edit Format View Help

```
h101,59,60000.00,apartment
h204,70,100200.00,row house
h131,245,324800.00,house
h124,24,34000.00,shop
h401,40,72000.00,apartment
h345,32,62400.00,shop
h123,54,134000.00,row house
h783,140,220500.00,house
h200,101,160000.00,house
h145,71,120400.00,apartment
h112,34,65100.00,apartment
```



ame	Introducti	Variables	Strings(8(	Selection(	Repetitior	Procedur	Lists(80)	Advanced	Additional	* Lecture	* Demons	* Pra
@ut												
@ut	40	60	48	80	56	91	23	40	60	10	355	
n@	40	70	59	49	70	65	40			5	130	
o@u	28	44	18		22		2			6	170	
a@u	40	70	80	90	80	100	80	60	60	10	750	
t@u	40	70	80	90	80	100	80	60	60	11	750	
@uti	40	70	80	90						2	75	
r@u	40	70	80	90	80	90	80	60	60	7	525	
u@u	40	70	80	90	80	96	54	60	10	2	340	
y@u	40	65	70	89	80	64	70	60	60	11	420	
g@u	40	70	79	90	80	100	80	60	57	10	520	
2uti	40	70	60	77	80	100	65	60	40	8	390	
2utu	40	64	75	83	41	87	75	60	10	7	405	
@uti	31	42	46	46	47	60	59	0	0	11	555	
u@u	28		16							1	145	
@uti	40	70	80	90	80	100	70	60	60	7	675	
iro@	40	70						60	20	2	330	
@u	40	70	80	90	80	100	80	60	40	1	375	
u@u	40	70	80	90	80	100	80	60	60	11	750	
@ut	40	70	69	45	63	88	57	20	20	8	485	
abc	40	69	75	90	80	100	80	60	60	9	550	
@ut	40	70	80	90	80	90	65	60	30	10	590	
yen(												
@uti	40	18								2	55	
@uti	40	70	80	90	80	99	80	60	52	9	710	
u@u	40									1		
@u	40	70	60	60	60	90	60	40		4	640	

.csv : Comma separated values format
.txt : text file
.tsv : tab separated file format
.xml: Extensible markup language
.sql – structured query language database table

```
<?xml version="1.0" encoding="UTF-8"?>
<note>
  <from>Jani</from>
  <to>Tove</to>
  <message>Remember me this weekend</message>
</note>
```

Movie

mvID	Title	Rating	Rel_date	Length	Studio
1	Angels & Demons	M	14-05-2009	138	Sony Pictures
2	Coco Avant Chanel	PG	25-06-2009	108	Roadshow
3	Harry Potter and the Half-Blood Prince	M	15-07-2009	153	Roadshow
4	The Proposal	PG	18-06-2009	107	Disney
5	Ice Age: Dawn of the Dinosaurs	PG	01-07-2009	94	20th Century Fox

ASCII Vs Unicode

- ASCII – American Standard code for Information Interchange
- These are character sets used for encoding documents on computers
- This code represent 256 characters as numbers – 0 to 255
- The ASCII value for capital letter **A** is 65 – binary form → 01000001
- Small letter **a** is 97-binary form→ 01100001
- It uses 8-bits to represent a character
- **When ASCII is there why Unicode developed?**
- ASCII can represent at the maximum of 255 characters and meant for English based keyboards and limited to other language characters
- Unicode represents characters which allows more characters than ASCII. **8/16/32-bit**
- This lets Unicode to have code for every character and symbol in every language.
- Unicode standard defines values for over 128,000 characters and can be seen at the Unicode Consortium.