Working with Data

Asst. Prof. Dr. Pisal Setthawong

BDM3301 – Data Analytic Fundamental

MIS4221 – Data Science

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

Topics

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

- CRISP-DM
- Working with Data
- Text vs Binary File
- Structured/Semi-Structured/Unstructured Data
- Data Types

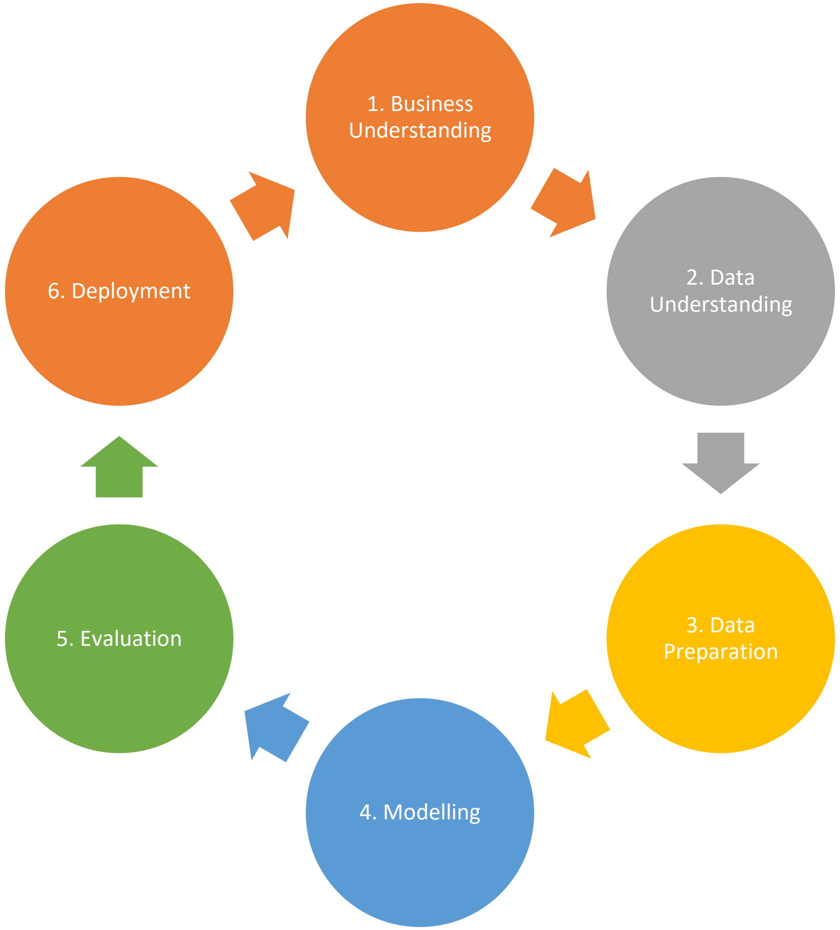
CRISP-DM

Cross-industry Standard Process for Data Mining (CRISP-DM)

- A structured approach to planning a data mining project
- Robust and well-proven methodology

CRISP-DM

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY



1. Business Understanding

- Understand what you want to accomplish from a business perspective
  - Set objectives
  - Produce project plan
  - Business success criteria
  - Data Mining problem definition (What Data to Collect?)

2. Data Understanding

- Acquire the data
- Explore the data and data quality

3. Data Preparation

- Select which of the data to use for analysis
  - Data Cleaning
  - Data Transformation
  - Data Integration

4. Modeling

- Select the most suitable modeling technique to apply to the data
- Test Design
- Parameter calibration for the modeling technique

5. Evaluation

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

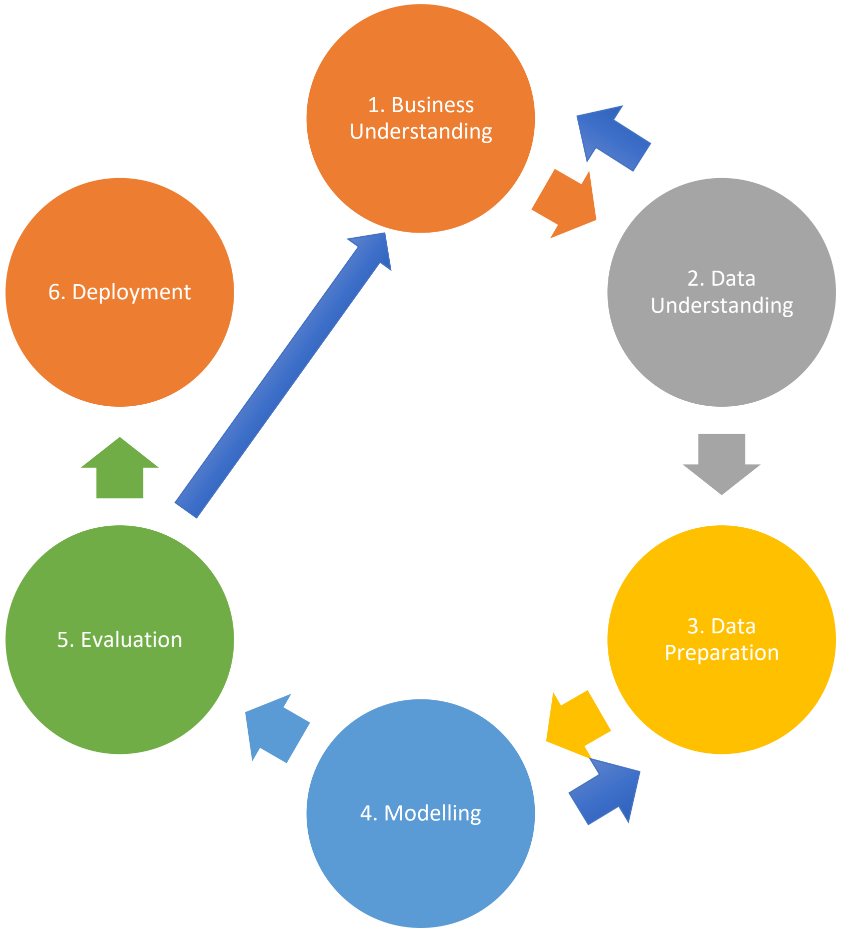
- Evaluate accuracy and generality of the model
- Does model meet business objectives?

6. Deployment

- Select the result model for deployment
- Repeatable implementation

CRISP-DM Flow

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY



Working with Data

Data

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

- One unit of data
- Can be a fact, symbol, or signal

- Plentiful
- Can contain both true and false data
- Not usually useful by itself

Data as New Economy

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

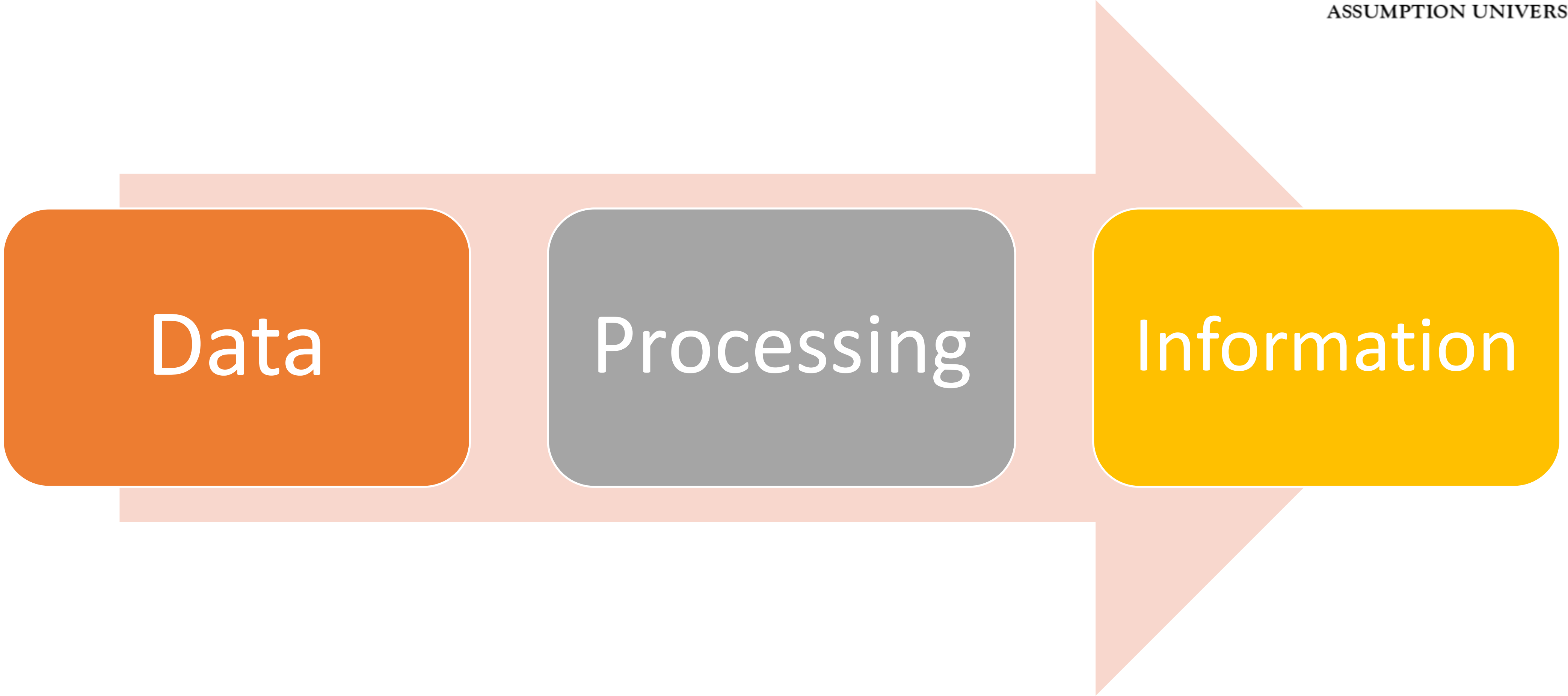
- With Cloud Technology, Increased Connectivity, it is easier to collect Data
- Data has been described as the new oil of the digital economy
  - Many business applications

Data Processing

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

- Data itself is useless
- Data need to be cleaned
  - Remove outliers
  - Instrument errors
  - Data Entry errors
- Data processing commonly needs to be done stage by stage
  - the output of one process is the input of another
- Continue until the data is useful

Data to Information



Data

Processing

Information

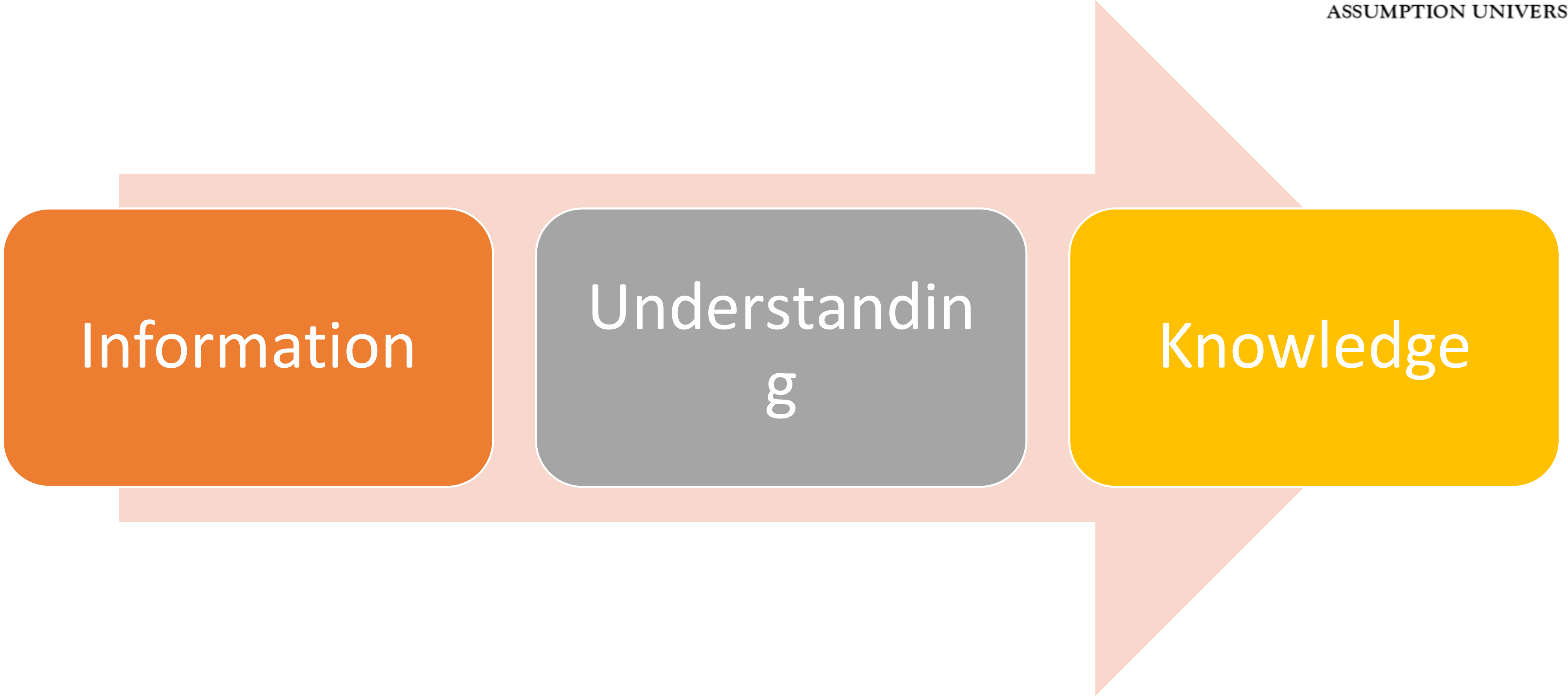
Information

- Processed data is usually referred to as Information
- Tracks the completeness, correctness, current, consistency and precision of the data

- More useful than data

Information to Knowledge

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY



Information

Understandin
g

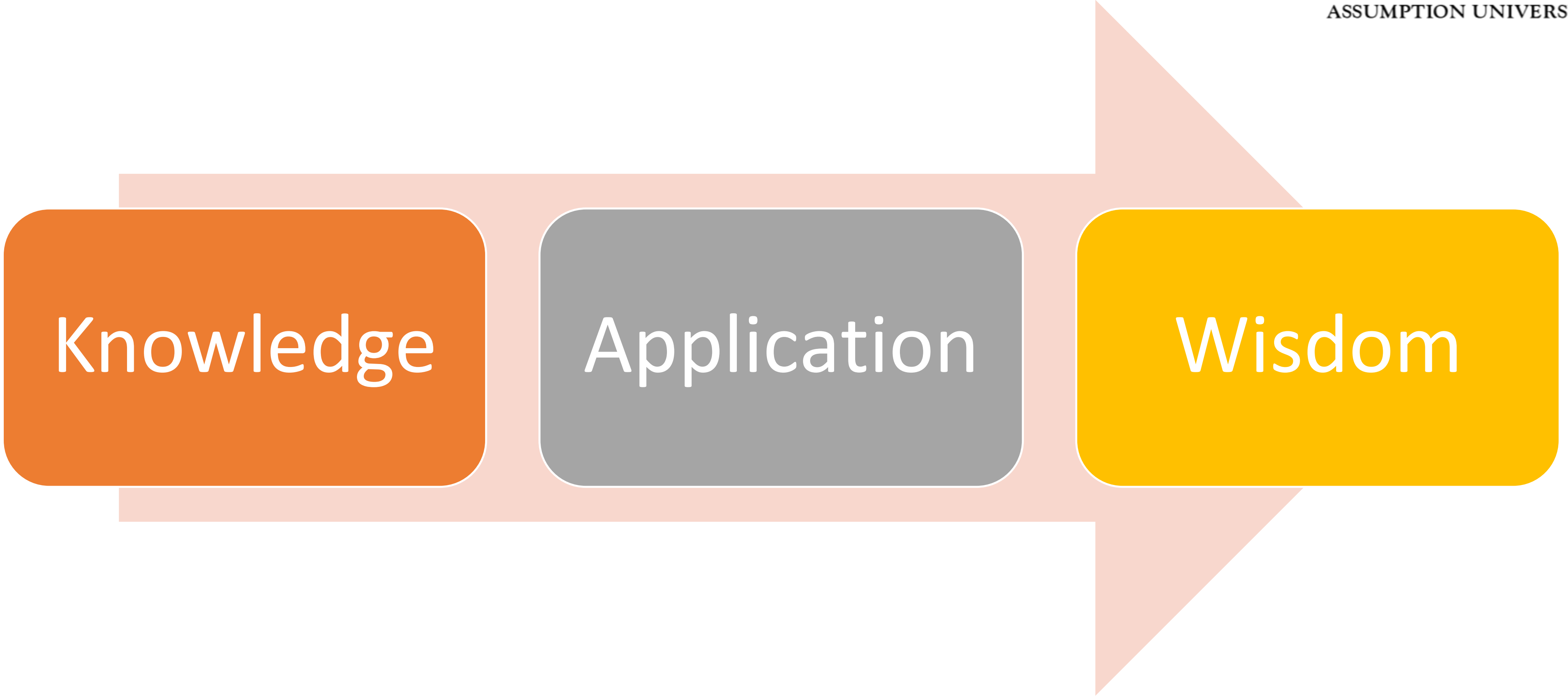
Knowledge

Knowledge

- Understanding the information creates knowledge
- Trends and patterns of information can be deduced

Knowledge to Wisdom

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY



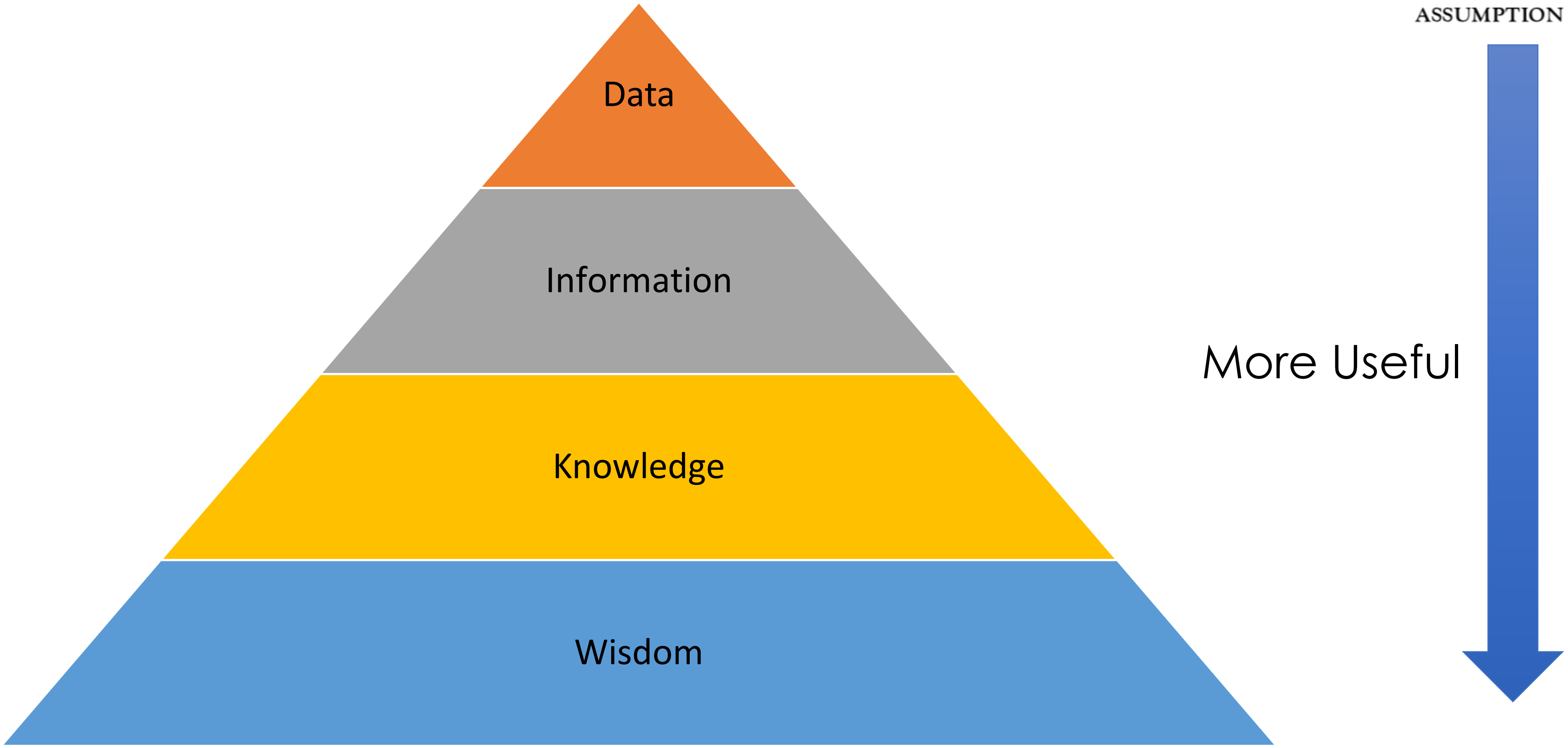
Wisdom

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

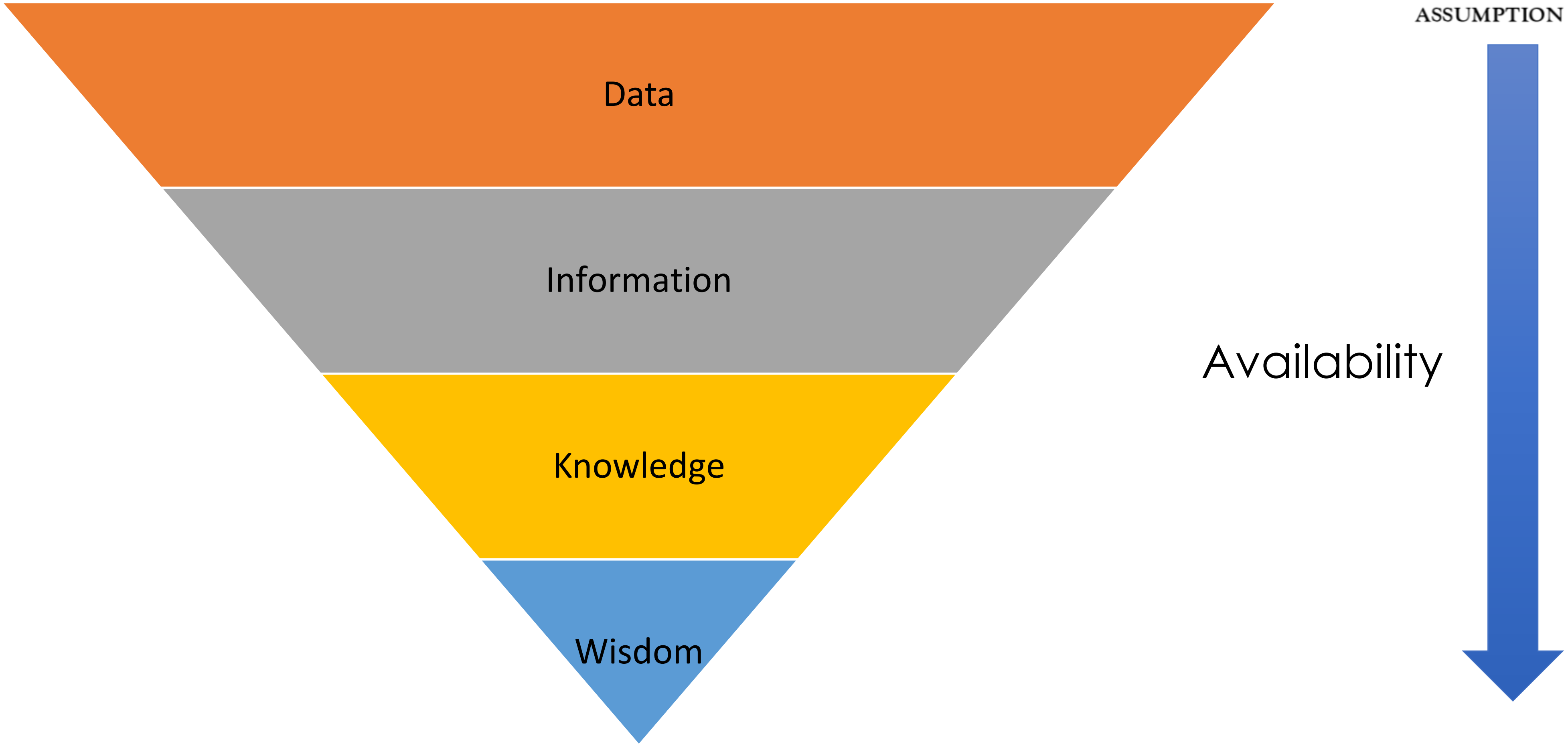
- Applications of knowledge to fit with the context of usage

Comparing Usefulness

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY



Comparing Availability



Text vs Binary Files

File Extensions

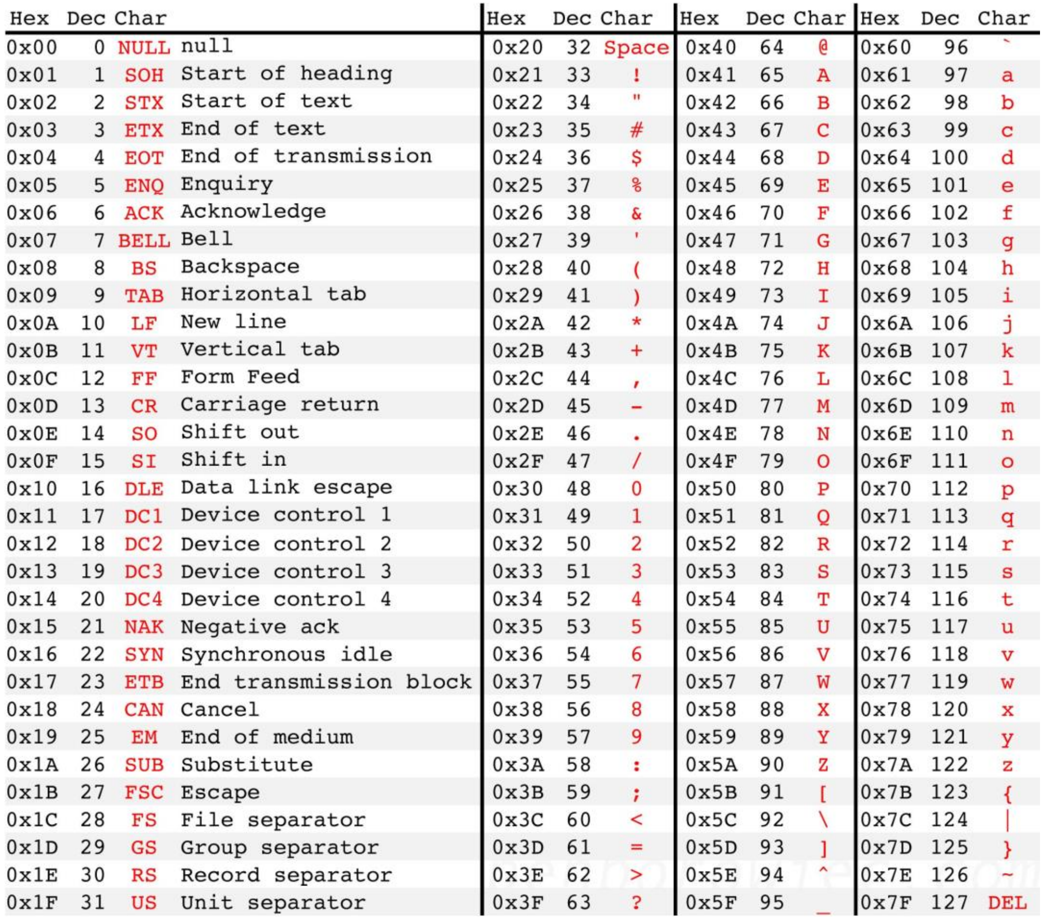
- File Extensions are the file suffix (letters after the .) which are used to help the computer identify what application should handle the file
- Files are either Text or Binary
  - By knowing the file extension, the correct application could be used to open the file

Text File

- Text are like letters that are stored using a specialized code (ASCII, ANSI, UNICODE)
- Text file is a file that contains only text
  - Images and Formatting is not possible with a text file
  - Small Size
  - Used mainly for source code, storing data, and documentation purposes
- To open and edit text files, a text editor is required (e.g. Notepad)

ASCII Format

- American Standard Code for Information Interchange
- The character encoding standard for electronic communication during the early days of computing
- Contains only 128 characters and 95 printable characters
    - Limited by early computers



Hex	Dec	Char		Hex	Dec	Char	Hex	Dec	Char	Hex	Dec	Char
0x00	0	NULL	null	0x20	32	Space	0x40	64	@	0x60	96	`
0x01	1	SOH	Start of heading	0x21	33	!	0x41	65	A	0x61	97	a
0x02	2	STX	Start of text	0x22	34	"	0x42	66	B	0x62	98	b
0x03	3	ETX	End of text	0x23	35	#	0x43	67	C	0x63	99	c
0x04	4	EOT	End of transmission	0x24	36	$	0x44	68	D	0x64	100	d
0x05	5	ENQ	Enquiry	0x25	37	%	0x45	69	E	0x65	101	e
0x06	6	ACK	Acknowledge	0x26	38	&	0x46	70	F	0x66	102	f
0x07	7	BELL	Bell	0x27	39	'	0x47	71	G	0x67	103	g
0x08	8	BS	Backspace	0x28	40	(	0x48	72	H	0x68	104	h
0x09	9	TAB	Horizontal tab	0x29	41	)	0x49	73	I	0x69	105	i
0x0A	10	LF	New line	0x2A	42	*	0x4A	74	J	0x6A	106	j
0x0B	11	VT	Vertical tab	0x2B	43	+	0x4B	75	K	0x6B	107	k
0x0C	12	FF	Form Feed	0x2C	44	,	0x4C	76	L	0x6C	108	l
0x0D	13	CR	Carriage return	0x2D	45	-	0x4D	77	M	0x6D	109	m
0x0E	14	SO	Shift out	0x2E	46	.	0x4E	78	N	0x6E	110	n
0x0F	15	SI	Shift in	0x2F	47	/	0x4F	79	O	0x6F	111	o
0x10	16	DLE	Data link escape	0x30	48	0	0x50	80	P	0x70	112	p
0x11	17	DC1	Device control 1	0x31	49	1	0x51	81	Q	0x71	113	q
0x12	18	DC2	Device control 2	0x32	50	2	0x52	82	R	0x72	114	r
0x13	19	DC3	Device control 3	0x33	51	3	0x53	83	S	0x73	115	s
0x14	20	DC4	Device control 4	0x34	52	4	0x54	84	T	0x74	116	t
0x15	21	NAK	Negative ack	0x35	53	5	0x55	85	U	0x75	117	u
0x16	22	SYN	Synchronous idle	0x36	54	6	0x56	86	V	0x76	118	v
0x17	23	ETB	End transmission block	0x37	55	7	0x57	87	W	0x77	119	w
0x18	24	CAN	Cancel	0x38	56	8	0x58	88	X	0x78	120	x
0x19	25	EM	End of medium	0x39	57	9	0x59	89	Y	0x79	121	y
0x1A	26	SUB	Substitute	0x3A	58	:	0x5A	90	Z	0x7A	122	z
0x1B	27	FSC	Escape	0x3B	59	;	0x5B	91	[	0x7B	123	{
0x1C	28	FS	File separator	0x3C	60	<	0x5C	92	\	0x7C	124	|
0x1D	29	GS	Group separator	0x3D	61	=	0x5D	93	]	0x7D	125	}
0x1E	30	RS	Record separator	0x3E	62	>	0x5E	94	^	0x7E	126	~
0x1F	31	US	Unit separator	0x3F	63	?	0x5F	95	_	0x7F	127	DEL

ANSI Format

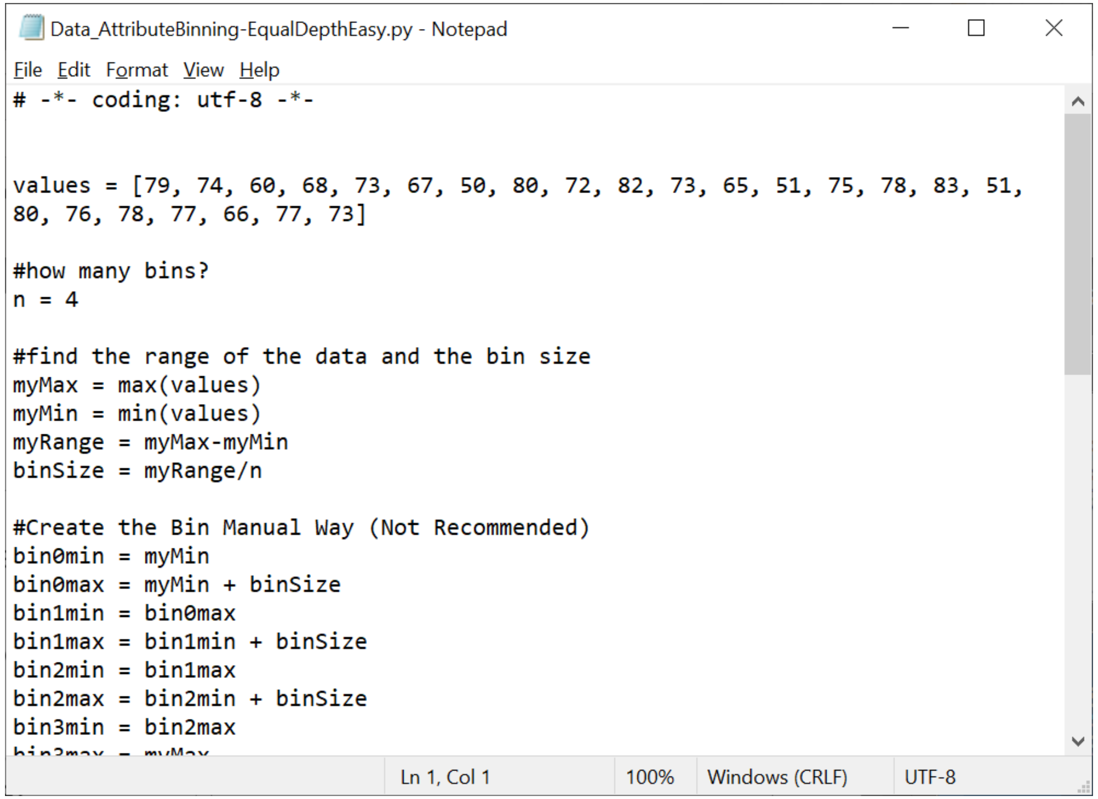
- Updated ASCII Format
- Increases from 7 to 8 bit (128->256 characters)
    - Allow for local variation of special symbols
- Microsoft-related standard for character set encoding
- Limited to English Only

Unicode Format

- The Unicode Standard
- Format aimed to provide consistent encoding, representation, and handling of text
    - Supports most of the writing systems
- Version 15 includes 149,186 characters that covers 161 scripts (including emoji and symbols and other formatting code)
- Most widespread use in the internationalization and localization of computer software
- Many variations exists offering different advantages/disadvantages

Text Editors

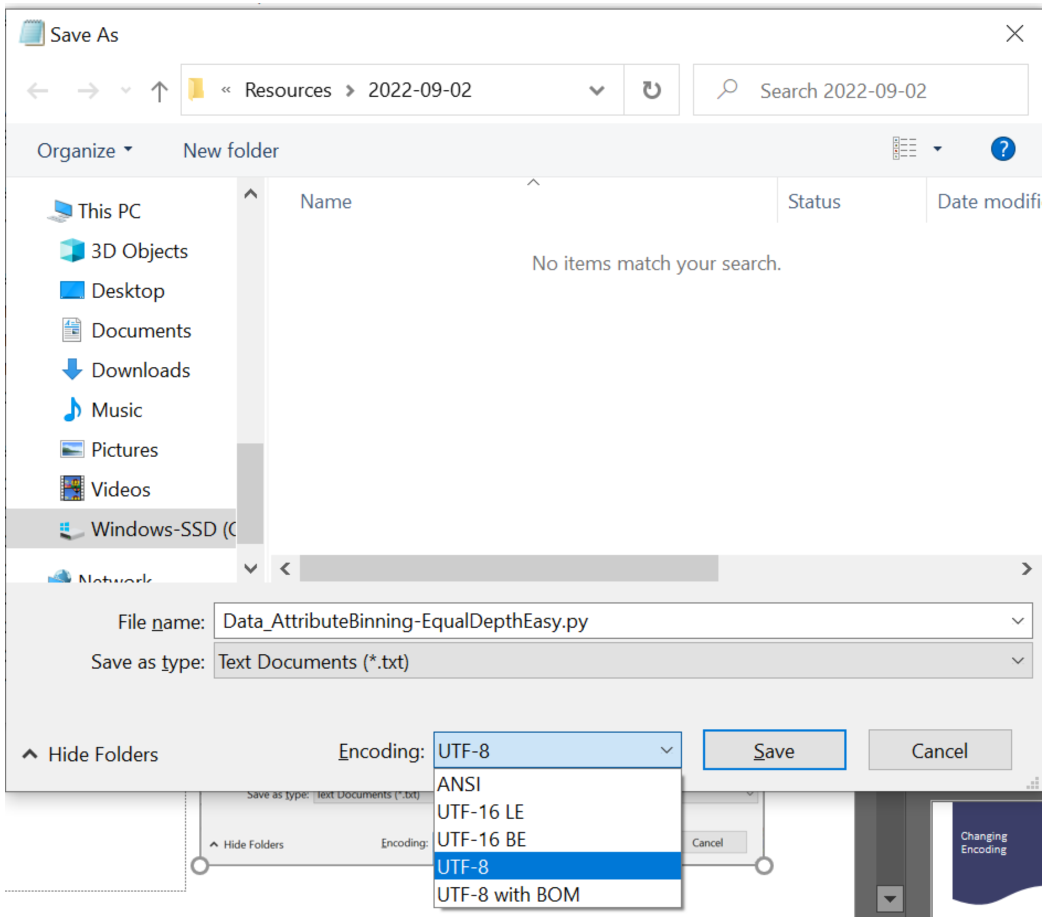
- Specialized programs that allow the viewing, editing, and creation of text files
- Different text editors contains different features
- Encoding formatting is important in cross language usage


Data_AttributeBinning-EqualDepthEasy.py - Notepad
File Edit Format View Help
Ln 1, Col 1
100%
Windows (CRLF)
UTF-8

```
# -*- coding: utf-8 -*-


values = [79, 74, 60, 68, 73, 67, 50, 80, 72, 82, 73, 65, 51, 75, 78, 83, 51,
80, 76, 78, 77, 66, 77, 73]

#how many bins?
n = 4

#find the range of the data and the bin size
myMax = max(values)
myMin = min(values)
myRange = myMax-myMin
binSize = myRange/n

#Create the Bin Manual Way (Not Recommended)
bin0min = myMin
bin0max = myMin + binSize
bin1min = bin0max
bin1max = bin1min + binSize
bin2min = bin1max
bin2max = bin2min + binSize
bin3min = bin2max
bin3max = myMax
```

Encoding Issues

- Text Encoding is usually defined by the Text Editor
- Changing the Encoding from the initial created file may cause some unintended side-effects



Save As

Resources › 2022-09-02

Search 2022-09-02

No items match your search.

File name: Data_AttributeBinning-EqualDepthEasy.py

Save as type: Text Documents (*.txt)

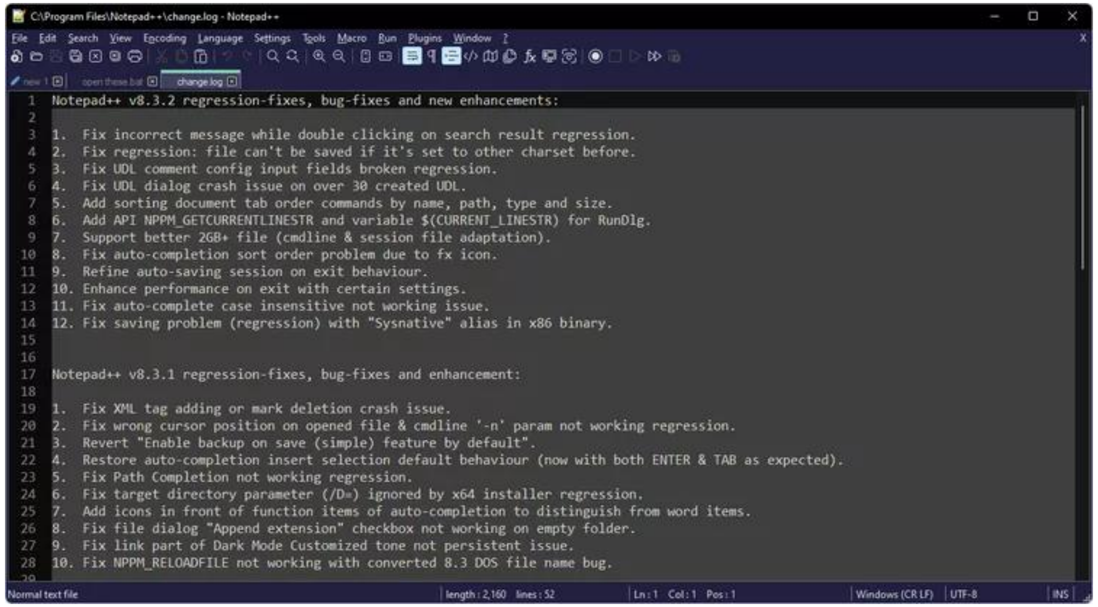
Encoding: UTF-8

File Formats that are Text Files

- Text Files (e.g.)
  - .txt
  - .rtf
- Data Interchange Format (e.g.)
  - .json
  - .xml
  - .csv
- Source Code (e.g.)
  - .py
  - .c / .cpp
  - .java

Notepad++

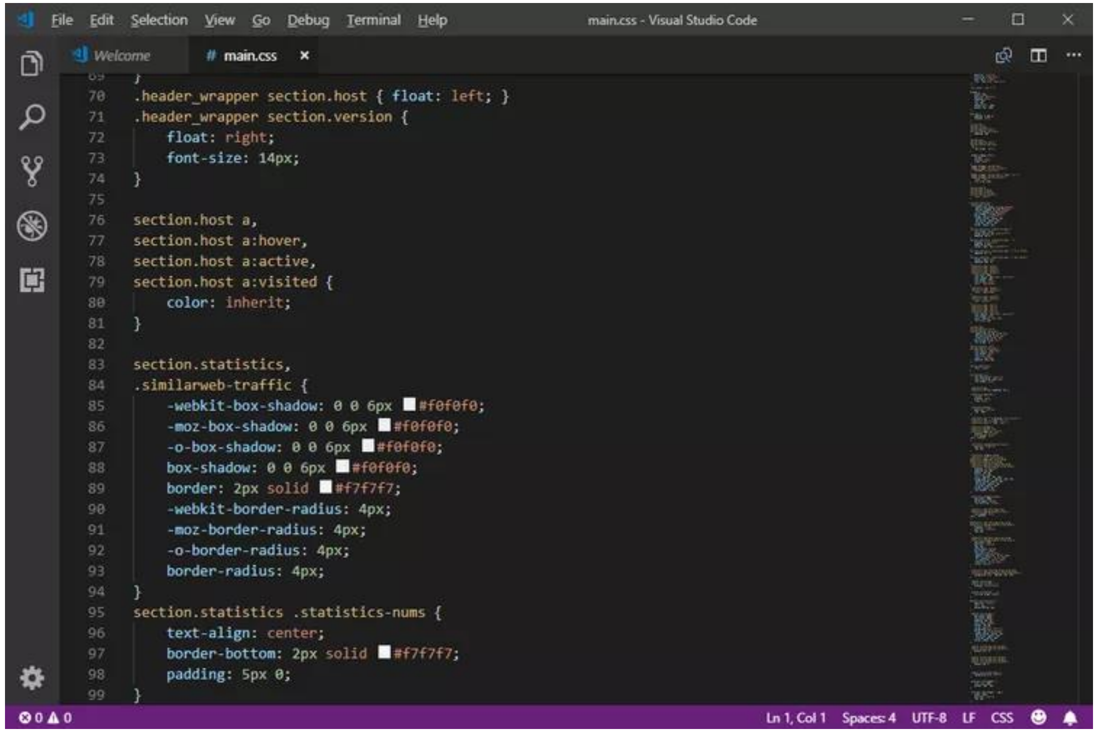
- Versatile Feature Heavy but light-weight text editor
- Only available in Windows
- https://notepad-plus-plus.org


C:\Program Files\Notepad++\change.log - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
new 1
open these.bat
change.log

```
Notepad++ v8.3.2 regression-fixes, bug-fixes and new enhancements:

1.  Fix incorrect message while double clicking on search result regression.
2.  Fix regression: file can't be saved if it's set to other charset before.
3.  Fix UDL comment config input fields broken regression.
4.  Fix UDL dialog crash issue on over 30 created UDL.
5.  Add sorting document tab order commands by name, path, type and size.
6.  Add API NPPM_GETCURRENTLINESTR and variable $(CURRENT_LINESTR) for RunDlg.
7.  Support better 2GB+ file (cmdline & session file adaptation).
8.  Fix auto-completion sort order problem due to fx icon.
9.  Refine auto-saving session on exit behaviour.
10. Enhance performance on exit with certain settings.
11. Fix auto-complete case insensitive not working issue.
12. Fix saving problem (regression) with "Sysnative" alias in x86 binary.


Notepad++ v8.3.1 regression-fixes, bug-fixes and enhancement:

1.  Fix XML tag adding or mark deletion crash issue.
2.  Fix wrong cursor position on opened file & cmdline '-n' param not working regression.
3.  Revert "Enable backup on save (simple) feature by default".
4.  Restore auto-completion insert selection default behaviour (now with both ENTER & TAB as expected).
5.  Fix Path Completion not working regression.
6.  Fix target directory parameter (/D=) ignored by x64 installer regression.
7.  Add icons in front of function items of auto-completion to distinguish from word items.
8.  Fix file dialog "Append extension" checkbox not working on empty folder.
9.  Fix link part of Dark Mode Customized tone not persistent issue.
10. Fix NPPM_RELOADFILE not working with converted 8.3 DOS file name bug.
```

Visual Studio Code

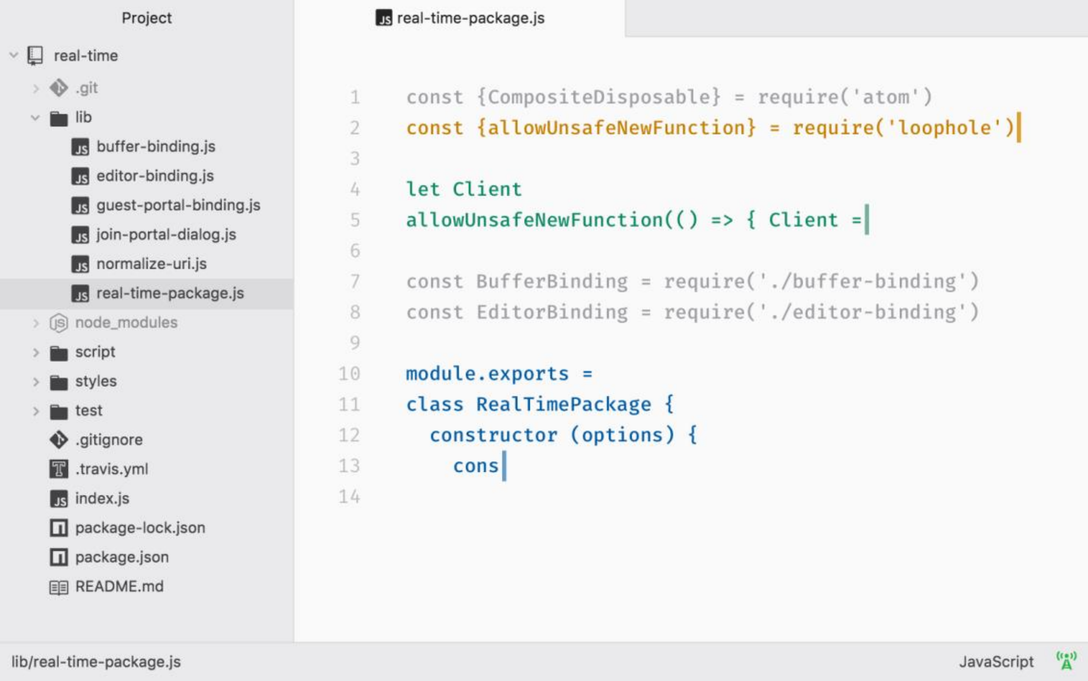
- Code Centric Text Editor
- Power features and addons
- Difficult to change the settings and be overwhelmed by the features
- https://code.visualstudio.com/



```css
}
.header_wrapper section.host { float: left; }
.header_wrapper section.version {
    float: right;
    font-size: 14px;
}

section.host a,
section.host a:hover,
section.host a:active,
section.host a:visited {
    color: inherit;
}

section.statistics,
.similarweb-traffic {
    -webkit-box-shadow: 0 0 6px ■#f0f0f0;
    -moz-box-shadow: 0 0 6px ■#f0f0f0;
    -o-box-shadow: 0 0 6px ■#f0f0f0;
    box-shadow: 0 0 6px ■#f0f0f0;
    border: 2px solid ■#f7f7f7;
    -webkit-border-radius: 4px;
    -moz-border-radius: 4px;
    -o-border-radius: 4px;
    border-radius: 4px;
}
section.statistics .statistics-nums {
    text-align: center;
    border-bottom: 2px solid ■#f7f7f7;
    padding: 5px 0;
}
```

Atom

- Highly Customizable Text Editor
- Multi-platform
- https://atom.io/



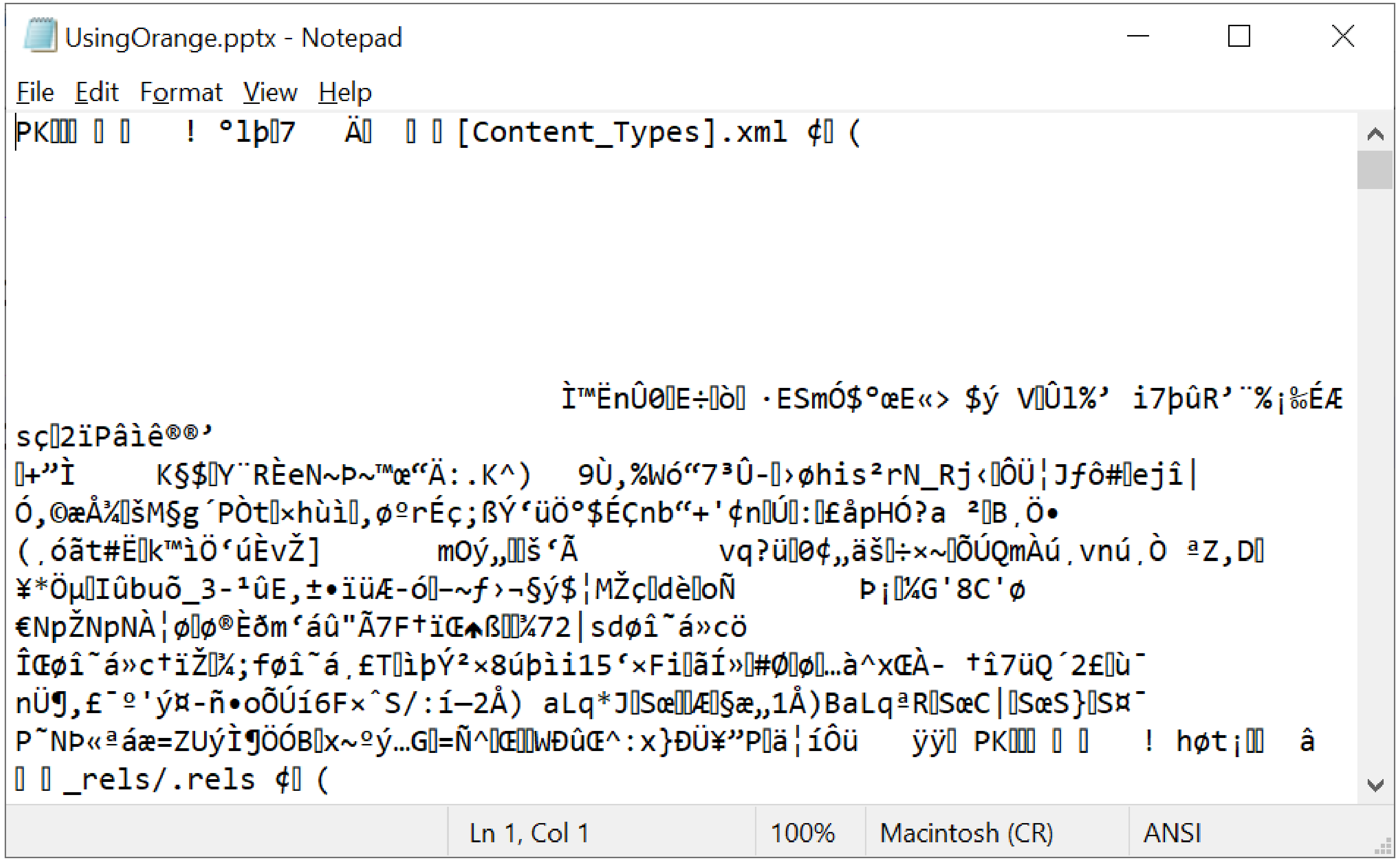
```
const {CompositeDisposable} = require('atom')
const {allowUnsafeNewFunction} = require('loophole')

let Client
allowUnsafeNewFunction(() => { Client =

const BufferBinding = require('./buffer-binding')
const EditorBinding = require('./editor-binding')

module.exports =
class RealTimePackage {
  constructor (options) {
    cons
```

Binary Files

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

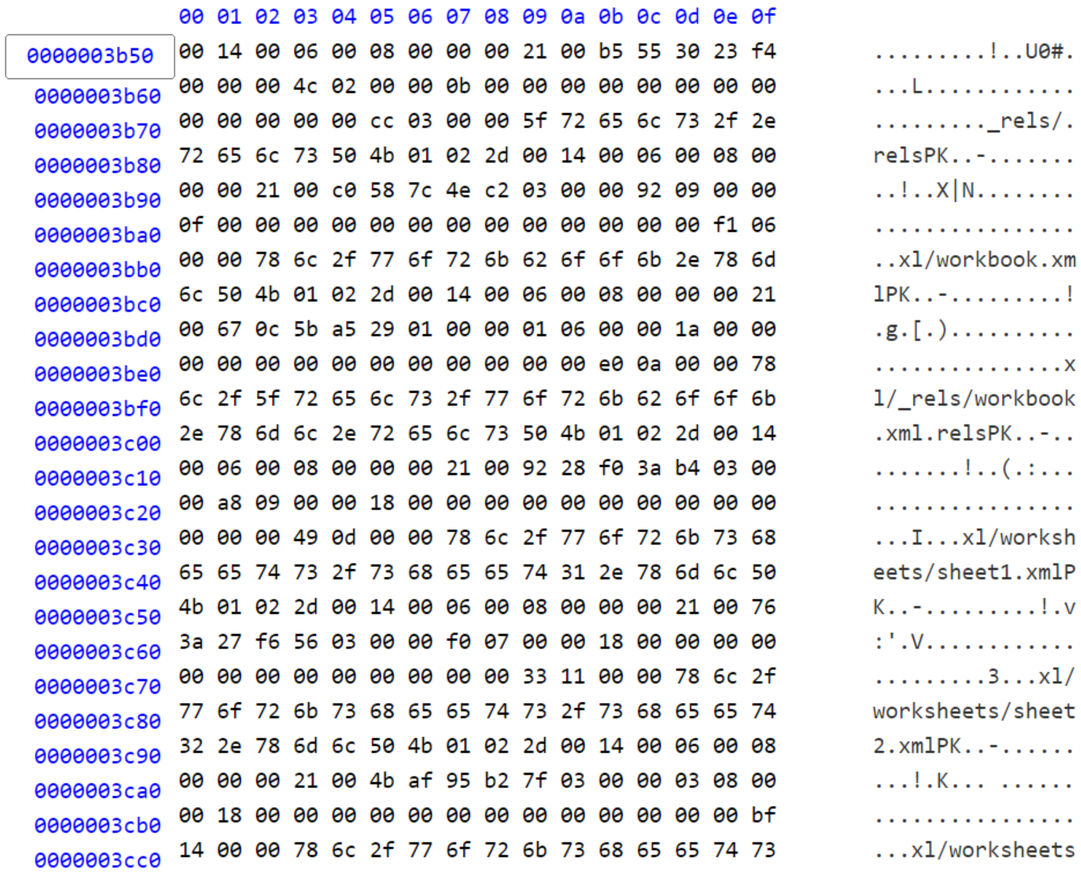
- Non-Text Files written as sequence of bits and bytes
- Usually referred to executable files that can be run by the computer
- Content such as images, video, data, and other files can be saved as a binary file
  - Binary files are more efficient in saving data
  - The content is not readable with a text editor and requires specialized software to understand the content
- Hex Editors are specialized program that can be used to help understand the content of Binary Files

Opening Binary File with Text Editor



```
PK    ! °1þ7   Ä   [Content_Types].xml ¢ (




                              Ì™ËnÛ0E÷òÂ ·ESmÓ$°œE«> $ý VÛ1%’ i7þûR’¨%¡‰ÉÆ
sç2ïPâìê®®’
+”Ì    K§$Y¨RÈeN~Þ~™œ“Ä:.K^)  9Ù,%Wó“7³Û->øhis²rN_Rj‹ÔÜ¦Jƒô#ejî|
Ó,©æÅ¾šM§g´PÒtxhùì,øºrÉç;ßÝ‘üÖ°$ÉÇnb“+'¢nÚ:£åpHÓ?a ²B¸Ö•
(¸óãt#Ëk™ìÖ‘úÈvŽ]      mÓý„š‘Ã        vq?ü0¢„äš÷×~ÕÚQmÀú¸vnú¸Ò ªZ,D
¥*ÖµIûbuõ_3-¹ûE,±•ïüÆ-ó–~ƒ›¬§ý$¦MŽçdèoÑ      Þ¡¼G'8C'ø
€NpŽNpNÀ¦øø®Èðm‘áû"Ã7F†ïŒßß¼72|sdøî˜á»cö
ÎŒøî˜á»c†ïŽ¼;føî˜á¸£TìþÝ²×8úþìi15‘×FiãÍ»#Øø…à^xŒÀ- †î7üQ´2£ù¯
nÜ¶,£¯º'ý¤-ñ•oÕÚí6F×ˆS/:í–2Å) aLq*JSœÆ§æ„1Å)BaLqªRSœC|SœS}S¤¯
P˜NÞ«ªáæ=ZUýÌ¶ÖÓBx~ºý…G=Ñ^ŒWÐûŒ^:x}ÐÜ¥”Pä¦íÔü   ÿÿ PK    ! høt¡  â
  _rels/.rels ¢ (
```

Opening Binary File with Hex Editor



```
            00 01 02 03 04 05 06 07 08 09 0a 0b 0c 0d 0e 0f
0000003b50  00 14 00 06 00 08 00 00 00 21 00 b5 55 30 23 f4     .........!..U0#.
0000003b60  00 00 00 4c 02 00 00 0b 00 00 00 00 00 00 00 00     ...L............
0000003b70  00 00 00 00 00 cc 03 00 00 5f 72 65 6c 73 2f 2e     ........._rels/.
0000003b80  72 65 6c 73 50 4b 01 02 2d 00 14 00 06 00 08 00     relsPK..-.......
0000003b90  00 00 21 00 c0 58 7c 4e c2 03 00 00 92 09 00 00     ..!..X|N........
0000003ba0  0f 00 00 00 00 00 00 00 00 00 00 00 00 00 f1 06     ................
0000003bb0  00 00 78 6c 2f 77 6f 72 6b 62 6f 6f 6b 2e 78 6d     ..xl/workbook.xm
0000003bc0  6c 50 4b 01 02 2d 00 14 00 06 00 08 00 00 00 21     lPK..-.........!
0000003bd0  00 67 0c 5b a5 29 01 00 00 01 06 00 00 1a 00 00     .g.[.)..........
0000003be0  00 00 00 00 00 00 00 00 00 00 00 e0 0a 00 00 78     ...............x
0000003bf0  6c 2f 5f 72 65 6c 73 2f 77 6f 72 6b 62 6f 6f 6b     l/_rels/workbook
0000003c00  2e 78 6d 6c 2e 72 65 6c 73 50 4b 01 02 2d 00 14     .xml.relsPK..-..
0000003c10  00 06 00 08 00 00 00 21 00 92 28 f0 3a b4 03 00     .......!..(.:...
0000003c20  00 a8 09 00 00 18 00 00 00 00 00 00 00 00 00 00     ................
0000003c30  00 00 00 49 0d 00 00 78 6c 2f 77 6f 72 6b 73 68     ...I...xl/worksh
0000003c40  65 65 74 73 2f 73 68 65 65 74 31 2e 78 6d 6c 50     eets/sheet1.xmlP
0000003c50  4b 01 02 2d 00 14 00 06 00 08 00 00 00 21 00 76     K..-.........!.v
0000003c60  3a 27 f6 56 03 00 00 f0 07 00 00 18 00 00 00 00     :'.V............
0000003c70  00 00 00 00 00 00 00 00 00 33 11 00 00 78 6c 2f     .........3...xl/
0000003c80  77 6f 72 6b 73 68 65 65 74 73 2f 73 68 65 65 74     worksheets/sheet
0000003c90  32 2e 78 6d 6c 50 4b 01 02 2d 00 14 00 06 00 08     2.xmlPK..-......
0000003ca0  00 00 00 21 00 4b af 95 b2 7f 03 00 00 03 08 00     ...!.K... ......
0000003cb0  00 18 00 00 00 00 00 00 00 00 00 00 00 00 00 bf     ................
0000003cc0  14 00 00 78 6c 2f 77 6f 72 6b 73 68 65 65 74 73     ...xl/worksheets
```

Structured/Semi-Structured/ Unstructured Data

Unstructured Data

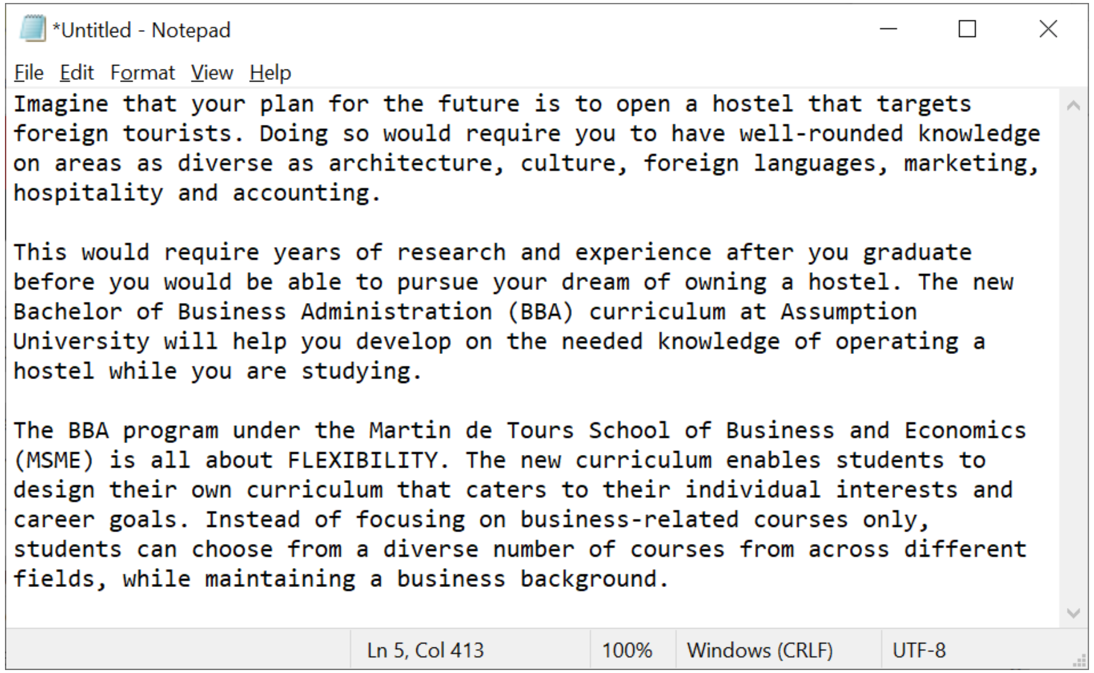
- Data stored with no specific format or organization
- Difficult to use
- Examples of Unstructured Data
  - Text Files, PDF Files, Images, Sound, Word Documents

Using Unstructured Data

- Difficult to use for analysis
- Formless and has no structure
- Required extensive pre-processing process before it can be analyzed by computer systems
  - Humans have higher cognitive level than computers!
- Most data created is unstructured

Unstructured Data Examples

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY



```
Imagine that your plan for the future is to open a hostel that targets
foreign tourists. Doing so would require you to have well-rounded knowledge
on areas as diverse as architecture, culture, foreign languages, marketing,
hospitality and accounting.

This would require years of research and experience after you graduate
before you would be able to pursue your dream of owning a hostel. The new
Bachelor of Business Administration (BBA) curriculum at Assumption
University will help you develop on the needed knowledge of operating a
hostel while you are studying.

The BBA program under the Martin de Tours School of Business and Economics
(MSME) is all about FLEXIBILITY. The new curriculum enables students to
design their own curriculum that caters to their individual interests and
career goals. Instead of focusing on business-related courses only,
students can choose from a diverse number of courses from across different
fields, while maintaining a business background.
```

MSME BUSINESS SCHOOL
50 Anniversary
มหาวิทยาลัยอัสสัมชัญ
ASSUMPTION UNIVERSITY of THAILAND
Product Design
Art and Design
Data Science
BBA
Music Business
Real Estate Business
Digital Marketing
Interior design
Architecture
Technology
DESIGN IT YOURSELF
NEW CURRICULUM 2020
MSME BUSINESS SCHOOL
Apply Now !
www.msme.au.edu | E-mail: au_admission@au.edu

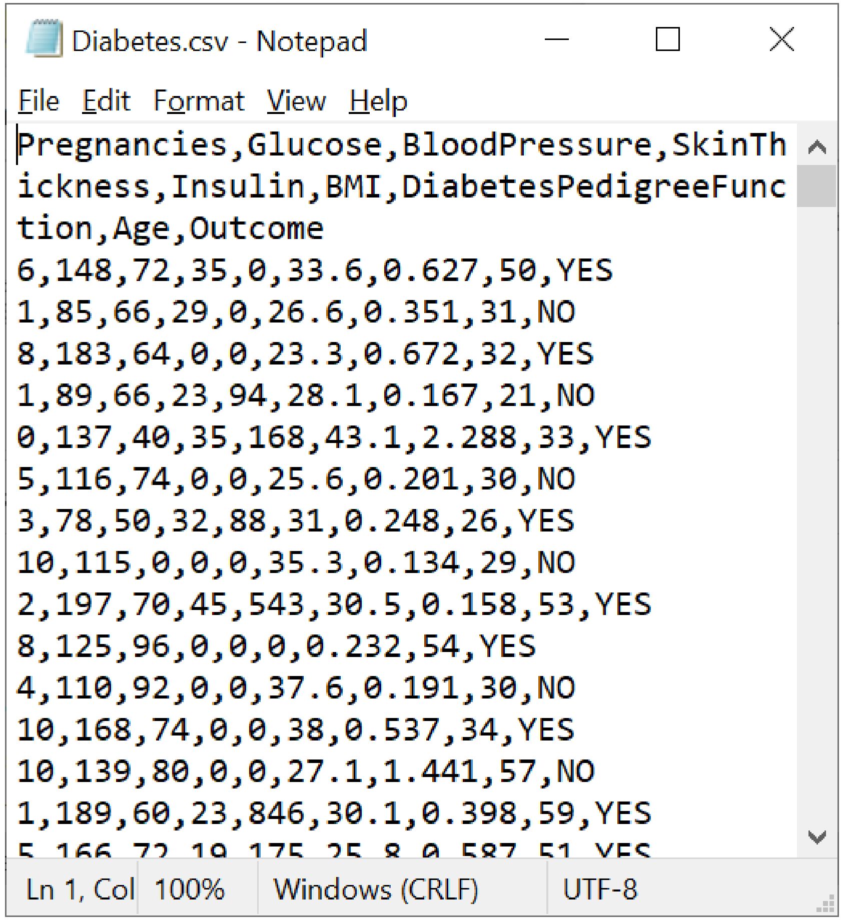
Semi-Structured Data

- Data with some degree of structure and organization
- Usually contains some form of markup code to help in the organization of the data
- Popular data-interchange formats such as XML/JSON are Semi-Structured and widely used in computer applications

Using Semi-Structured Data

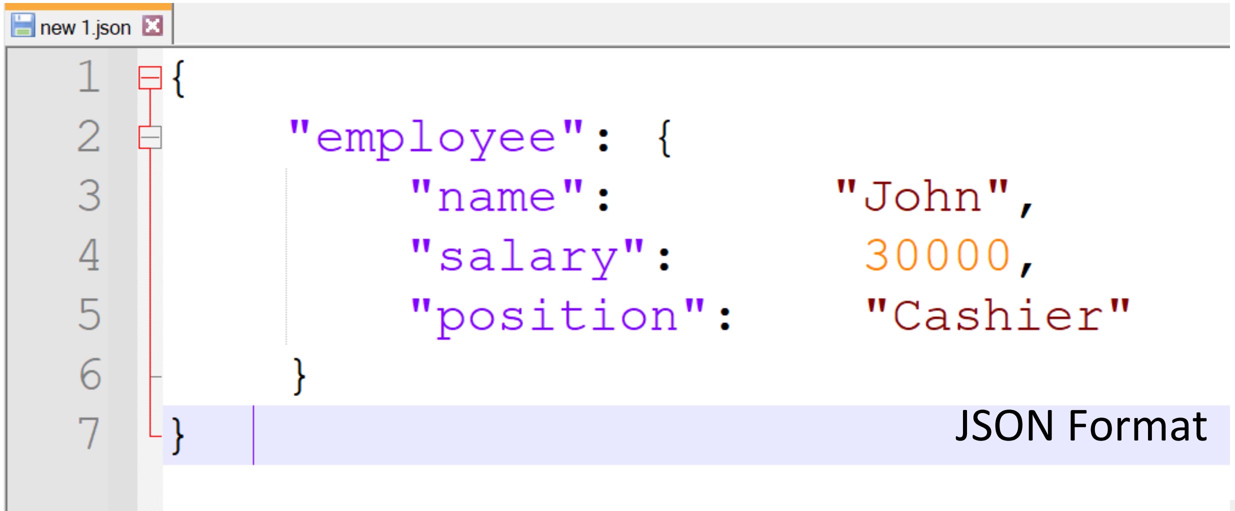
- Markup/Tags/Markers are used to provide structure to data
  - Separate the elements
  - Enforce Hierarchies
  - Provide additional information to the data
- Allow dynamic element size and flexible order
- Easier to grow and manage the data
- Requires some pre-processing before can be used by computers
  - Popular formats (XML/JSON) have an assortment of tools for usage
- Web 2.0 utilizes semi-structured data to help drive growth

Semi-Structured Data Examples



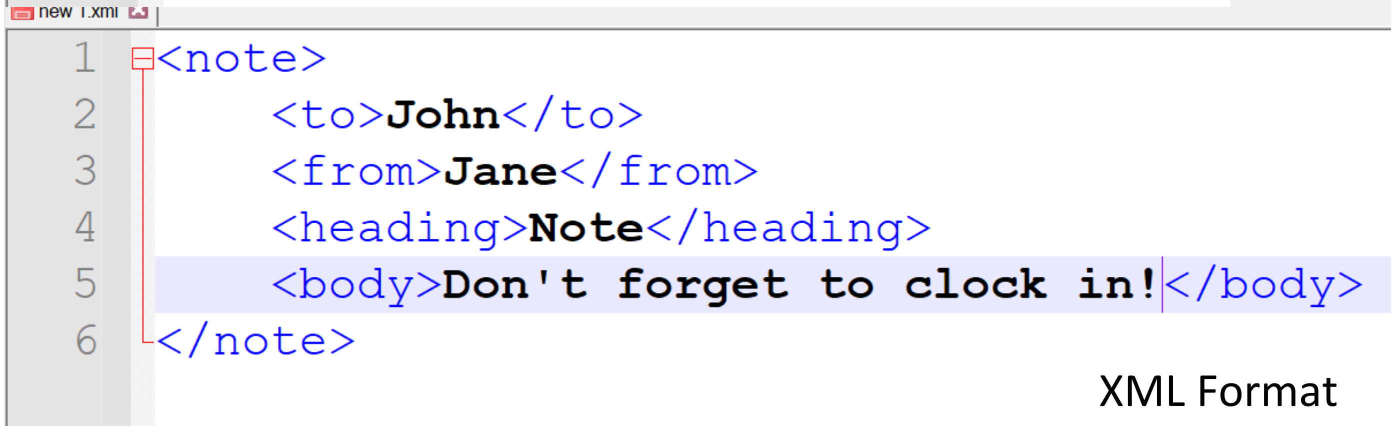
```
Pregnancies,Glucose,BloodPressure,SkinTh
ickness,Insulin,BMI,DiabetesPedigreeFunc
tion,Age,Outcome
6,148,72,35,0,33.6,0.627,50,YES
1,85,66,29,0,26.6,0.351,31,NO
8,183,64,0,0,23.3,0.672,32,YES
1,89,66,23,94,28.1,0.167,21,NO
0,137,40,35,168,43.1,2.288,33,YES
5,116,74,0,0,25.6,0.201,30,NO
3,78,50,32,88,31,0.248,26,YES
10,115,0,0,0,35.3,0.134,29,NO
2,197,70,45,543,30.5,0.158,53,YES
8,125,96,0,0,0,0.232,54,YES
4,110,92,0,0,37.6,0.191,30,NO
10,168,74,0,0,38,0.537,34,YES
10,139,80,0,0,27.1,1.441,57,NO
1,189,60,23,846,30.1,0.398,59,YES
5,166,72,19,175,25.8,0.587,51,YES
```

CSV format (comma separated values)



```
{
	"employee":  {
		"name":		"John",
		"salary":	 30000,
		"position":	 "Cashier"
	}
}
```

JSON Format



```
<note>
	<to>John</to>
	<from>Jane</from>
	<heading>Note</heading>
	<body>Don't forget to clock in!</body>
</note>
```

XML Format

CSV Format

- Text file - Comma Separated Value file format
- Text File format that is usually used to store data
  - A comma separates the field
  - A new line is used to specify a new record
  - Quotation pairs are used to specify between text and number data
- Spreadsheets and other data storage formats used this format
- Easy to export from Spreadsheets (EXCEL) to data mining software with this format
- Cannot contain hierarchical data like XML/JSON but used to store tabular data

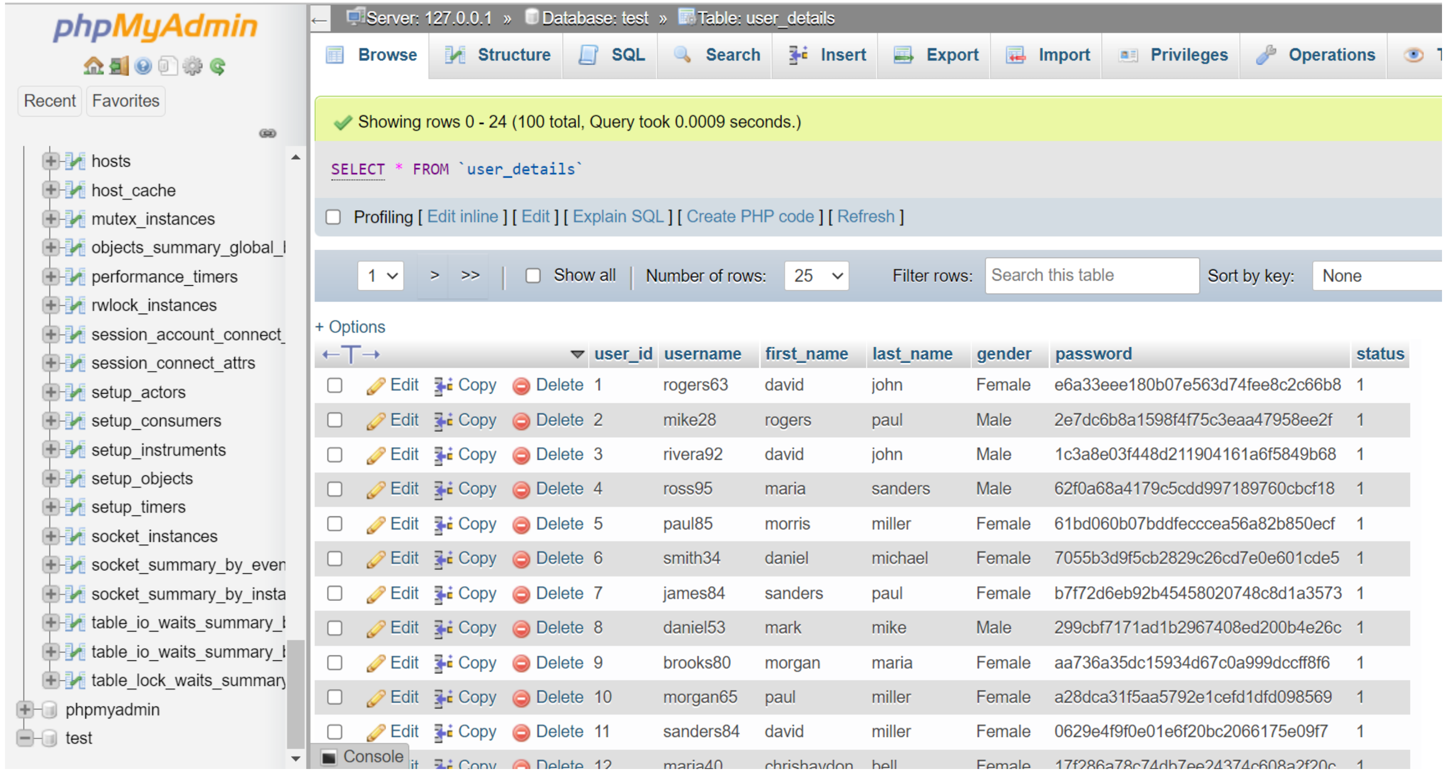
Structured Data

- Data with a high degree of structure and organization
- Usually have specialized software to maintain and use the data
- Database Management Systems are examples of systems that work with structured data

Using Structured Data

- Can be analyzed without any preprocessing
- Requires the usage of specialized tools for extracting the structured data
    - Structure Query Language (SQL) and Query by Example (QBE) are two popular approaches used to extract the structured data
    - Popular and robust tools allows for complex queries and applications

Structured Data Example



hosts
host_cache
mutex_instances
objects_summary_global_
performance_timers
rwlock_instances
session_account_connect_
session_connect_attrs
setup_actors
setup_consumers
setup_instruments
setup_objects
setup_timers
socket_instances
socket_summary_by_even
socket_summary_by_insta
table_io_waits_summary_
table_io_waits_summary_
table_lock_waits_summar
phpmyadmin
test

✓ Showing rows 0 - 24 (100 total, Query took 0.0009 seconds.)

```
SELECT * FROM `user_details`
```

□ Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

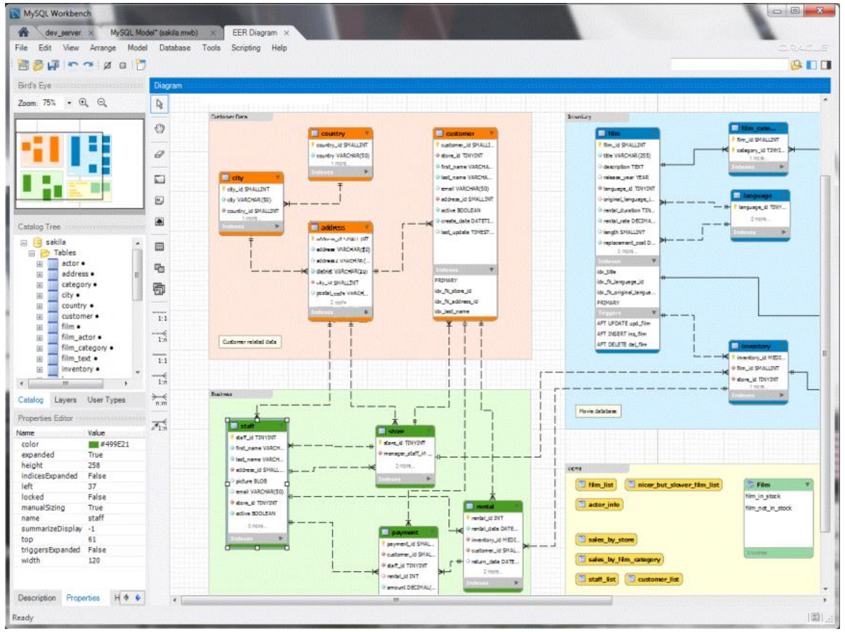
1 > >> □ Show all Number of rows: 25 Filter rows: Search this table Sort by key: None

+ Options

	user_id	username	first_name	last_name	gender	password	status
□ Edit Copy Delete	1	rogers63	david	john	Female	e6a33eee180b07e563d74fee8c2c66b8	1
□ Edit Copy Delete	2	mike28	rogers	paul	Male	2e7dc6b8a1598f4f75c3eaa47958ee2f	1
□ Edit Copy Delete	3	rivera92	david	john	Male	1c3a8e03f448d211904161a6f5849b68	1
□ Edit Copy Delete	4	ross95	maria	sanders	Male	62f0a68a4179c5cdd997189760cbcf18	1
□ Edit Copy Delete	5	paul85	morris	miller	Female	61bd060b07bddfecccea56a82b850ecf	1
□ Edit Copy Delete	6	smith34	daniel	michael	Female	7055b3d9f5cb2829c26cd7e0e601cde5	1
□ Edit Copy Delete	7	james84	sanders	paul	Female	b7f72d6eb92b45458020748c8d1a3573	1
□ Edit Copy Delete	8	daniel53	mark	mike	Male	299cbf7171ad1b2967408ed200b4e26c	1
□ Edit Copy Delete	9	brooks80	morgan	maria	Female	aa736a35dc15934d67c0a999dccff8f6	1
□ Edit Copy Delete	10	morgan65	paul	miller	Female	a28dca31f5aa5792e1cefd1dfd098569	1
□ Edit Copy Delete	11	sanders84	david	miller	Female	0629e4f9f0e01e6f20bc2066175e09f7	1
Copy Delete	12	maria40	chrishaydon	bell	Female	17f286a78c74db7ee24374c608a2f20c	1

Console

Interfacing with MySQL with PHPMyAdmin



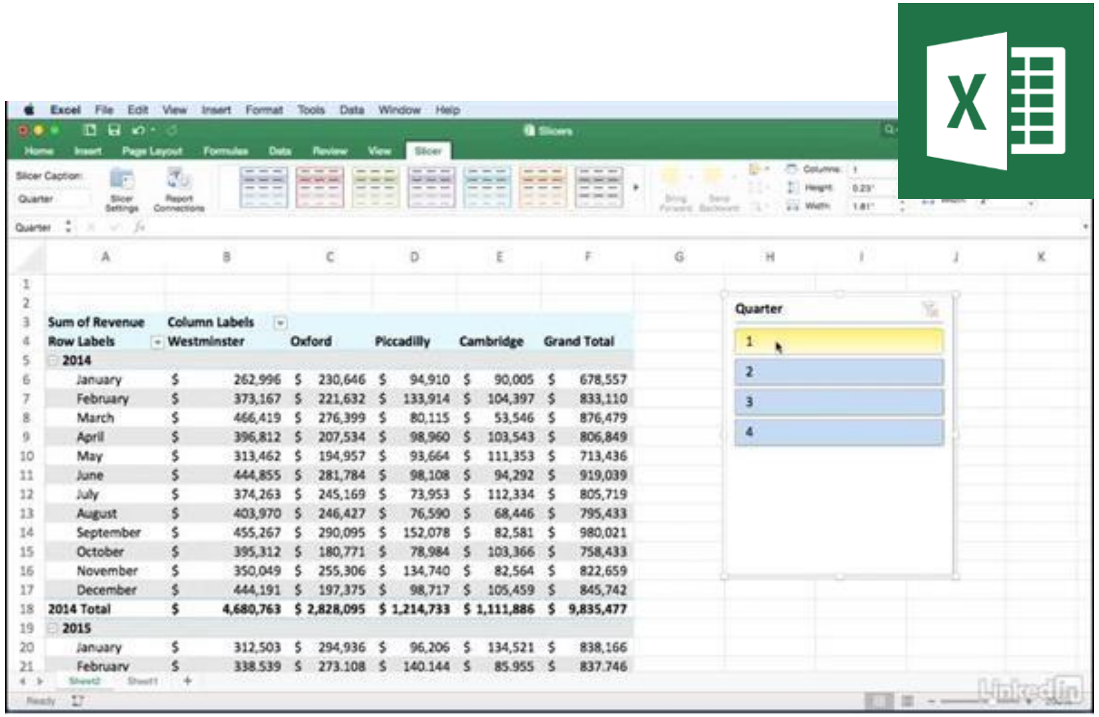
MySQL DBMS

Spreadsheets vs Database

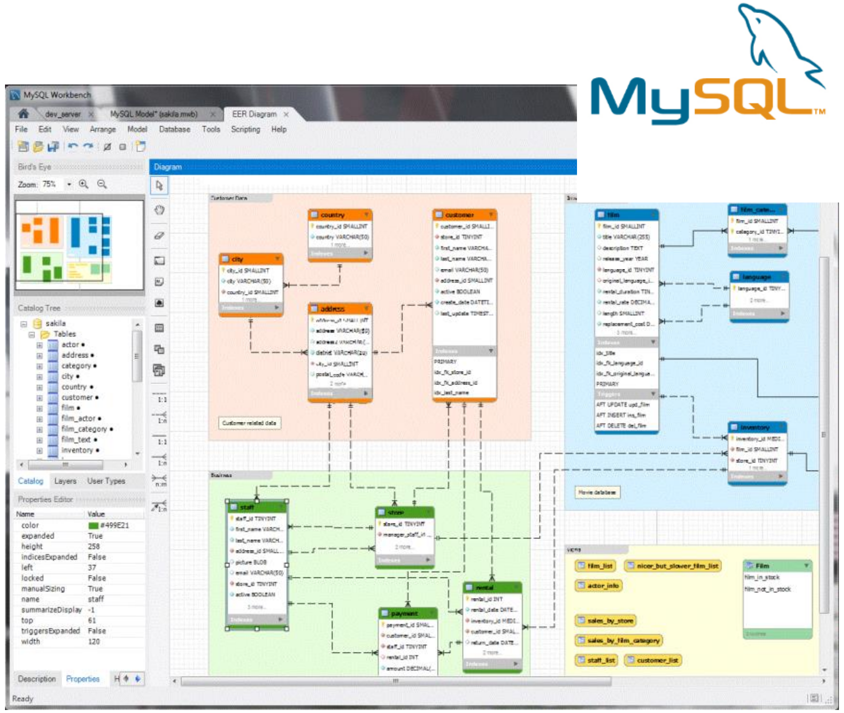
- Unstructured
- General Purpose
- Able to easily Structure, Analyze, and Organize Data
- Relatively Easy to Use
  - Virtually Every Professional can Use

- Structured
- Powerful Tools
- Requires Knowledge of Database
  - Database Design
  - SQL
  - Platform Basics
- Only IT Professionals can understand and use

Spreadsheets vs Relational Database


Excel File Edit View Insert Format Tools Data Window Help
Slicers
Home Insert Page Layout Formulas Data Review View Slicer
Slicer Caption:
Quarter
Slicer Settings
Report Connections
Quarter
1
2
3
4
Sheet2 Sheet1
Ready

	A	B	C	D	E	F
3	Sum of Revenue	Column Labels				
4	Row Labels	Westminster	Oxford	Piccadilly	Cambridge	Grand Total
5	2014					
6	January	$ 262,996	$ 230,646	$ 94,910	$ 90,005	$ 678,557
7	February	$ 373,167	$ 221,632	$ 133,914	$ 104,397	$ 833,110
8	March	$ 466,419	$ 276,399	$ 80,115	$ 53,546	$ 876,479
9	April	$ 396,812	$ 207,534	$ 98,960	$ 103,543	$ 806,849
10	May	$ 313,462	$ 194,957	$ 93,664	$ 111,353	$ 713,436
11	June	$ 444,855	$ 281,784	$ 98,108	$ 94,292	$ 919,039
12	July	$ 374,263	$ 245,169	$ 73,953	$ 112,334	$ 805,719
13	August	$ 403,970	$ 246,427	$ 76,590	$ 68,446	$ 795,433
14	September	$ 455,267	$ 290,095	$ 152,078	$ 82,581	$ 980,021
15	October	$ 395,312	$ 180,771	$ 78,984	$ 103,366	$ 758,433
16	November	$ 350,049	$ 255,306	$ 134,740	$ 82,564	$ 822,659
17	December	$ 444,191	$ 197,375	$ 98,717	$ 105,459	$ 845,742
18	2014 Total	$ 4,680,763	$ 2,828,095	$ 1,214,733	$ 1,111,886	$ 9,835,477
19	2015					
20	January	$ 312,503	$ 294,936	$ 96,206	$ 134,521	$ 838,166
21	February	$ 338.539	$ 273.108	$ 140.144	$ 85.955	$ 837.746



Properties Editor

Name	Value
color	🟩 #499E21
expanded	True
height	258
indicesExpanded	False
left	37
locked	False
manualSizing	True
name	staff
summarizeDisplay	-1
top	61
triggersExpanded	False
width	120

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

Data Types

Data

- Data is usually referred to as a collection of (data) objects and their attributes

Attributes

- Objects contain many attributes
  - E.g. height, weight, age, etc.
- Attribute is similar to the following terms
  - Variable
  - Field
  - Property
  - Feature
  - Characteristics

Objects

- Collection of attributes used to describe an object
  - A student can be defined by many attributes such as id, name, gender, major, minor, dob, etc.
- Object is similar to the following terms
  - Record
  - Sample
  - Entity
  - Instance
  - Case

Object/Attributes

Attributes

Objects

ID	FirstName	LastName	Major
6115555	John	Smith	IBM
6115556	Jane	Doe	MIS
6115559	Peter	Parker	HTM
6115823	Rachel	Ingram	ACT
.	.	.	.
.	.	.	.

Types of Attributes

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

- Many types of attributes
- Need to distinguish the difference between the attributes
  - Data Type (computer format)
  - Measurement Scales

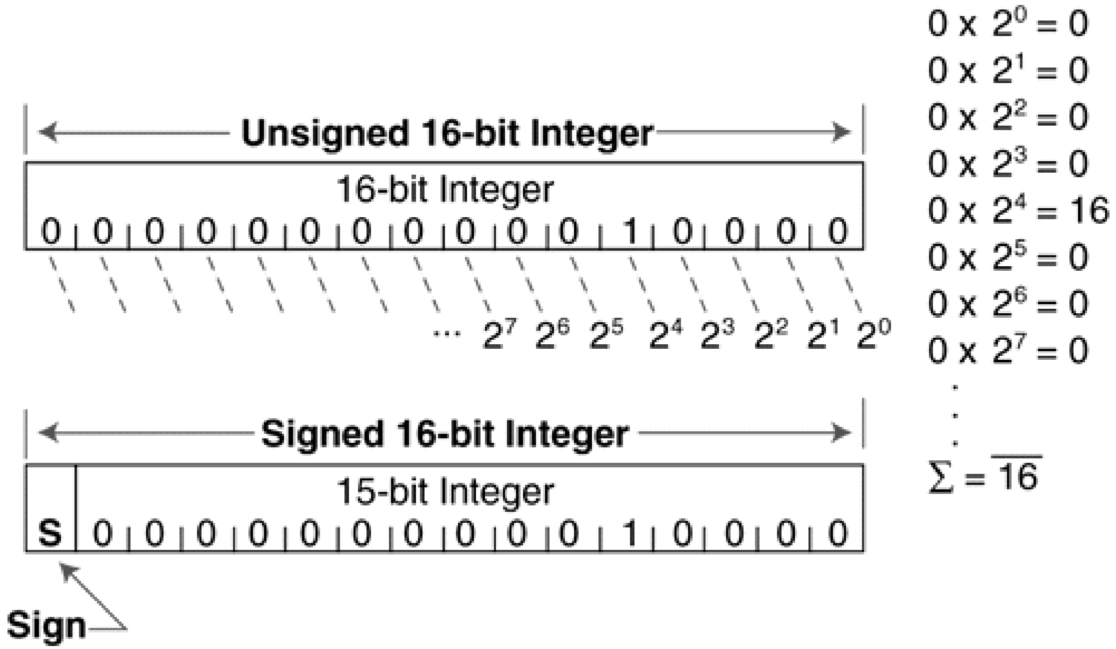
Data Type

- Computer systems can keep data of certain types
- Need to know all the data types the system can keep
- Common data types include
  - Integer
  - Floating Point
  - Fixed Point Number
  - Character
  - String
  - Boolean

Integer

- Full Numbers are written without a fractional component
- Can be positive or negative
- E.g. 2, -7, 315, 999, 1337

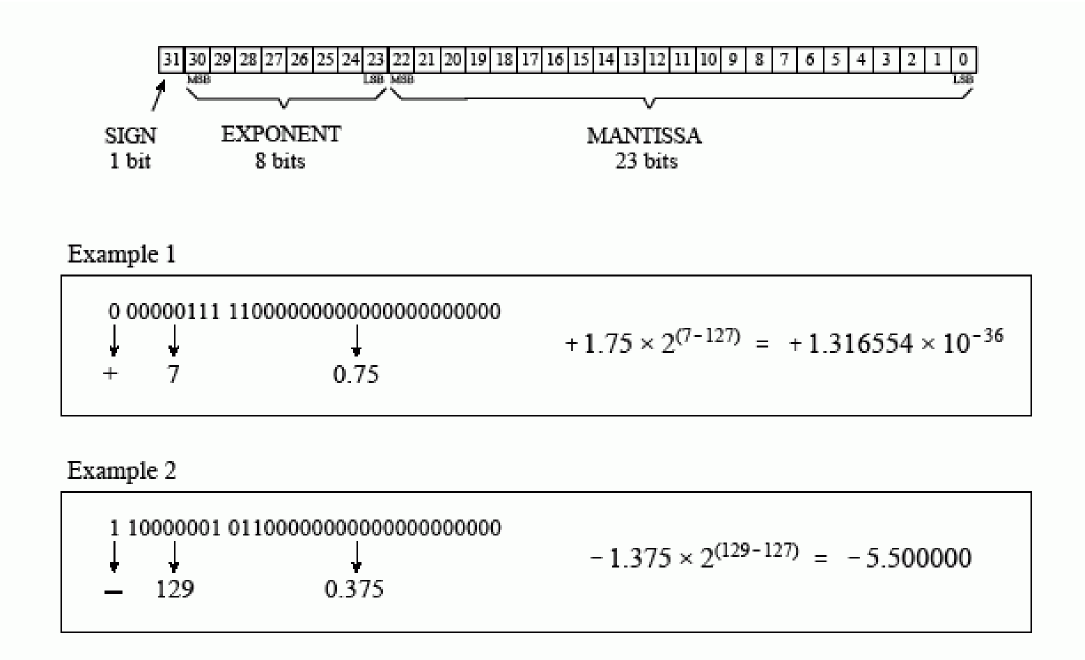
Integer Number



Floating Point

- A computer representation of real numbers that include the fractional part
- Modern computers use the IEEE 754 Standard for Floating-Point Arithmetic
- Many GPUs can do very fast Floating Point Operations

Floating Point Number



Example 1

0 00000111 11000000000000000000000

$+$ 7 0.75

$+1.75 \times 2^{(7-127)} = +1.316554 \times 10^{-36}$

Example 2

1 10000001 01100000000000000000000

$-$ 129 0.375

$-1.375 \times 2^{(129-127)} = -5.500000$

Single Precision Floating Point Data

Fixed Point Number

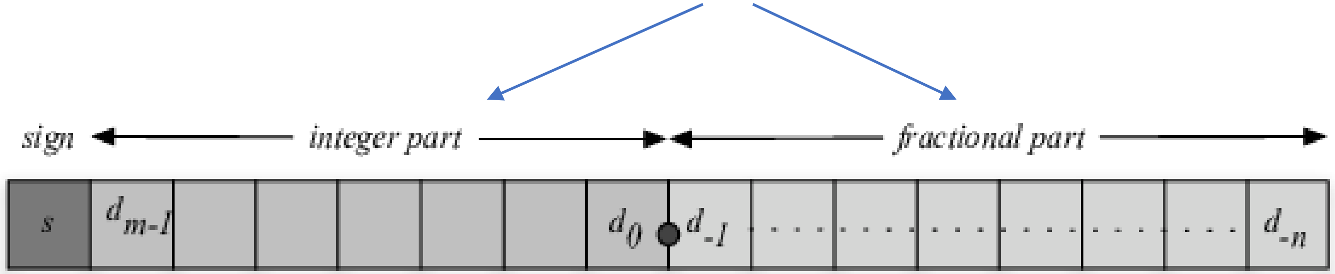
- Another number system that allows representation of real numbers that include the fractional part
- Higher level of control and precision with specific numbers if used
- Not commonly implemented in many programming language

Fixed Point Number

MSME
Martin de Tours School of Management and Economics
BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

Number of bits for integer and fractional part can be defined

Both use Integer Styled Calculation



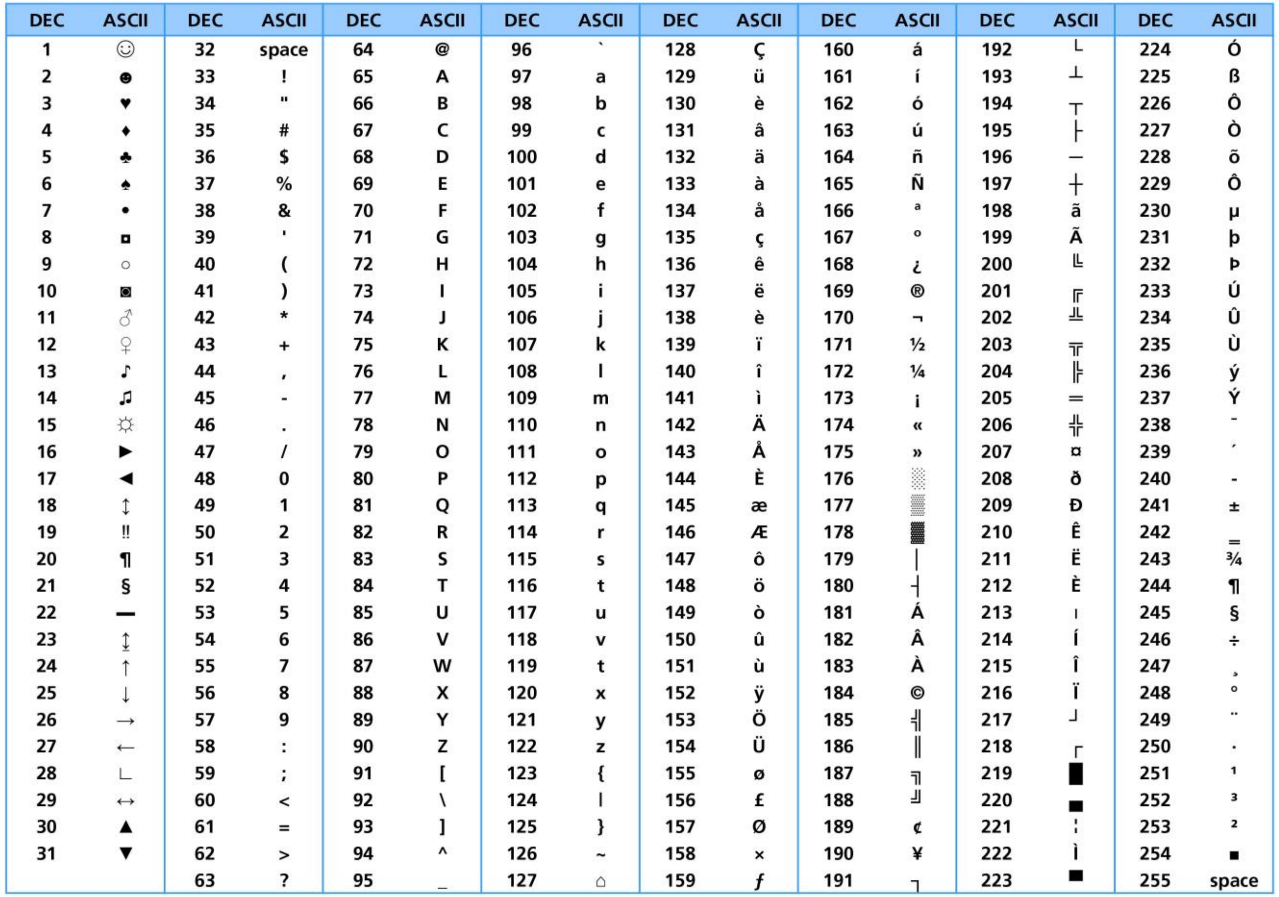
Fixed-Point Format

Character

- Used to keep one character, or one character in a text
- There are character encoding for characters
- The most popular is UTF-8 as it allows for multi-language encoding

Character (ASCII) 8 Bit

American Standard Code for Information Interchange (ASCII)
Uses 7 Bit Data which allows 128 different characters
Only English Script



DEC	ASCII	DEC	ASCII	DEC	ASCII	DEC	ASCII	DEC	ASCII	DEC	ASCII	DEC	ASCII	DEC	ASCII
1	☺	32	space	64	@	96	`	128	Ç	160	á	192	└	224	Ó
2	☻	33	!	65	A	97	a	129	ü	161	í	193	┴	225	ß
3	♥	34	"	66	B	98	b	130	è	162	ó	194	┬	226	Ô
4	♦	35	#	67	C	99	c	131	â	163	ú	195	├	227	Ò
5	♣	36	$	68	D	100	d	132	ä	164	ñ	196	─	228	õ
6	♠	37	%	69	E	101	e	133	à	165	Ñ	197	┼	229	Ô
7	•	38	&	70	F	102	f	134	å	166	ª	198	ã	230	µ
8	◘	39	'	71	G	103	g	135	ç	167	º	199	Ã	231	þ
9	○	40	(	72	H	104	h	136	ê	168	¿	200	╚	232	Þ
10	◙	41	)	73	I	105	i	137	ë	169	®	201	╔	233	Ú
11	♂	42	*	74	J	106	j	138	è	170	¬	202	╩	234	Û
12	♀	43	+	75	K	107	k	139	ï	171	½	203	╦	235	Ù
13	♪	44	,	76	L	108	l	140	î	172	¼	204	╠	236	ý
14	♫	45	-	77	M	109	m	141	ì	173	¡	205	═	237	Ý
15	☼	46	.	78	N	110	n	142	Ä	174	«	206	╬	238	¯
16	►	47	/	79	O	111	o	143	Å	175	»	207	¤	239	´
17	◄	48	0	80	P	112	p	144	È	176	░	208	ð	240	-
18	↕	49	1	81	Q	113	q	145	æ	177	▒	209	Đ	241	±
19	‼	50	2	82	R	114	r	146	Æ	178	▓	210	Ê	242	‗
20	¶	51	3	83	S	115	s	147	ô	179	│	211	Ë	243	¾
21	§	52	4	84	T	116	t	148	ö	180	┤	212	È	244	¶
22	▬	53	5	85	U	117	u	149	ò	181	Á	213	ı	245	§
23	↨	54	6	86	V	118	v	150	û	182	Â	214	Í	246	÷
24	↑	55	7	87	W	119	t	151	ù	183	À	215	Î	247	¸
25	↓	56	8	88	X	120	x	152	ÿ	184	©	216	Ï	248	°
26	→	57	9	89	Y	121	y	153	Ö	185	╣	217	┘	249	¨
27	←	58	:	90	Z	122	z	154	Ü	186	║	218	┌	250	·
28	∟	59	;	91	[	123	{	155	ø	187	╗	219	█	251	¹
29	↔	60	<	92	\	124	|	156	£	188	╝	220	▄	252	³
30	▲	61	=	93	]	125	}	157	Ø	189	¢	221	¦	253	²
31	▼	62	>	94	^	126	~	158	×	190	¥	222	Ì	254	■
		63	?	95	_	127	⌂	159	ƒ	191	┐	223	▀	255	space

Character (Unicode)

- Unicode is computing industry standard for the consistent encoding, representation, and handling of text
- Most recent version, *Unicode 12.1*, contains 137,994 characters, covering modern/history scripts, symbols and emoji
- Allows representation of multiple languages

String

- Multiple characters are contained as a string
- Label and Text data are string data type by nature
- There is usually a prefix and suffix of double quotation (“) to signify the content inside is a string data
  - “Hello World”

Boolean

- Value that could be either True or False
- Requires only 1 bit

Date and Time

- Contain data for date and time
- Different systems use different approaches
  - Dates in excel is saved as a number and counts the number of days from Jan 1, 1900
    - Use display as Date to convert number to date format (e.g. 2 => Jan 2, 1900
  - MySQL uses timestamp format ('YYYY-MM-DD hh:mm:ss’) to store date and time data
    - Use functions to convert or extract required data from timestamp

Limitation of Data Type

- Data type is important if you are working with computer systems
- Without knowing data type, the data scientist would not be able to understand what is used in the computer system

Data Types in Data Mining Tool

- Data Mining Tools simplify the data types use to make it easier to use
- Orange provides the following data types
  - Categorical
  - Numeric
  - Text
  - Date/Time
- Possible to set the target of attributes (feature/target/meta)