CIS 635 Knowledge Discovery & Data Mining

Basic Data Types and Introduction to Data Encoding

Outline

- Basic Data Types
- A brief introduction to Image Data
- Data Encoding (categorical data)
- NumPy basics

Digital data

- In computing everything is digital and binary
- All data types we talked about
- Bit(0 / 1): Digital letter
- Byte (000 0011): Digital word
- Kilo (Byte), Mega(Byte), Giga (Byte): We are talking about Digital data and their sizes mainly

Basic Data types

Data: Most data, at lower level, fall into two groups:

- Numerical, or
- Categorical

Basic Data types

Data: Most data, at lower level, fall into two groups:

- **Numerical**, or
- Categorical

Numerical: Also known as “**Quantitative data**”

- **Measurements:** person’s height, weight, or blood pressure; or
- **Counts**: such as number of stock shares a person wons, number of teeth a dog has, or the number of pages your favourite book contains

Basic Data types

Data: Most data, at lower level, fall into two groups:

- **Numerical**, or
- Categorical

Numerical: Also known as “**Quantitative data**”

- **Measurements:** person’s height, weight, or blood pressure; or
- **Counts**: such as number of stock shares a person wons, number of teeth a dog has, or the number of pages your favourite book contains

Numerical data can be further broken into two types: discrete and continuous

- **Discrete:** items that can be counted; they can take on possible values that can be listed out.
- **Continuous**: Usually represents measurements; their possible values cannot be counted such as: a person’s height, weight, IQ, or blood pressure

Basic Data types

Data: Most data, at lower level, fall into two groups:

- Numerical, or
- **Categorical**

Numerical: Also known as “**Quantitative data**”

- **Measurements:** person’s height, weight, or blood pressure; or
- **Counts**: such as number of stock shares a person wons, number of teeth a dog has, or the number of pages your favourite book contains

Numerical data can be further broken into two types: discrete and continuous

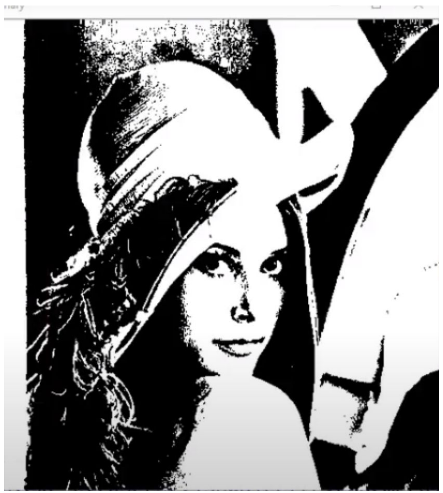
- **Discrete:** items that can be counted; they can take on possible values that can be listed out.
- **Continuous**: Usually represents measurements; their possible values cannot be counted such as: a person’s height, weight, IQ, or blood pressure

Categorical data: Categorical data represent characteristics such as a person’s gender, marital status, country of birth, or the types of movies they like.

- Can be ordinal (say, student grades A, B, C; days of week, moths of week)
- Non ordinal data (person’s gender, marital status, country of birth)

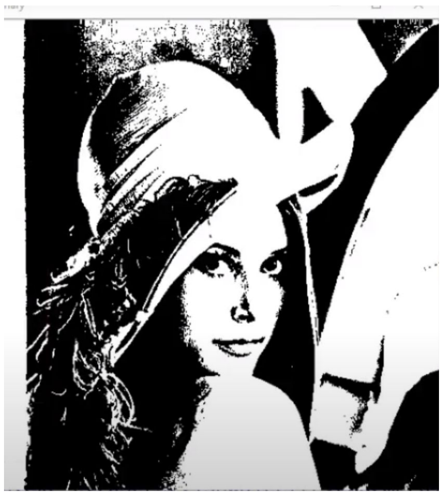
Image Data

Binary, gray-scale, and color images



Binary image

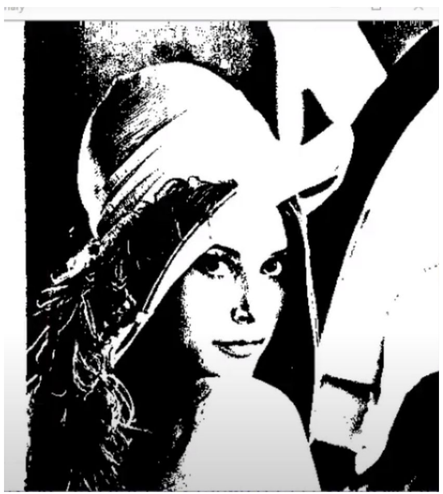
Binary, gray-scale, and color images



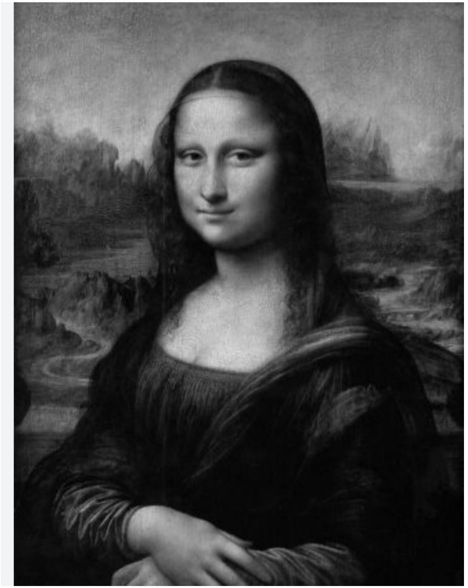
Binary image

B/W image

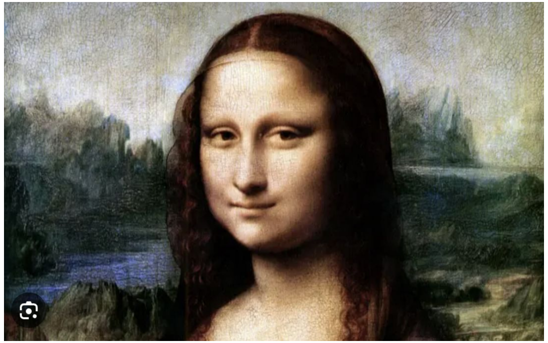
Binary, gray-scale, and color images



Binary image

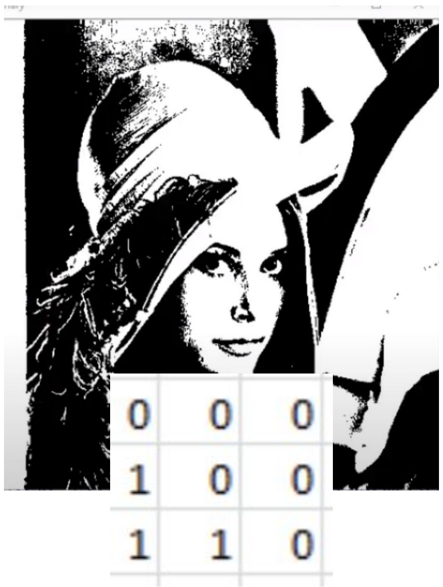


B/W image

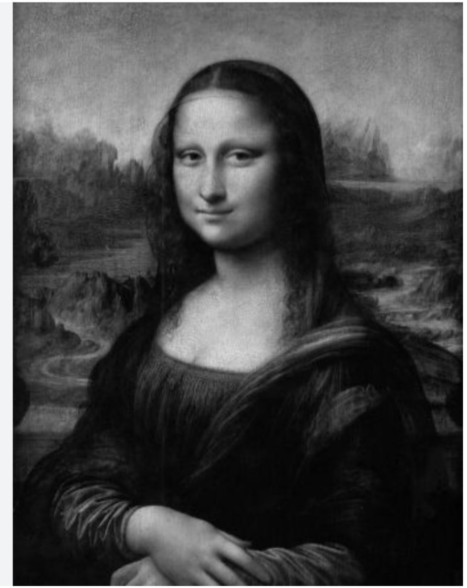


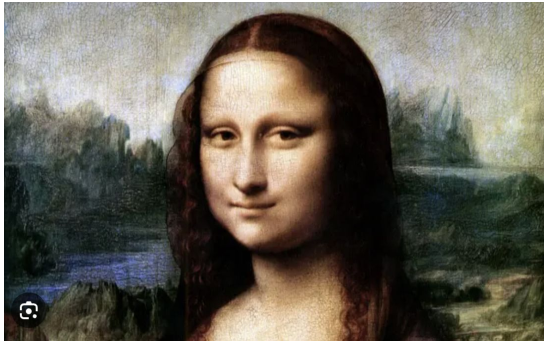
Color image

Binary, gray-scale, and color images

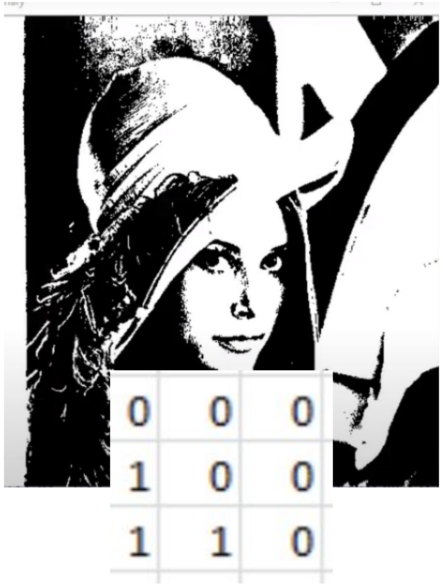


0	0	0
1	0	0
1	1	0

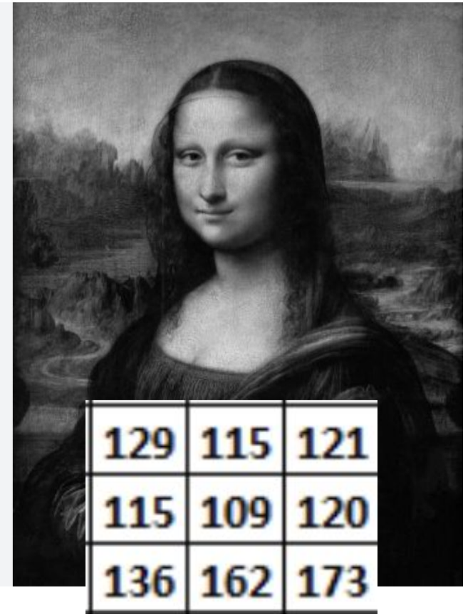




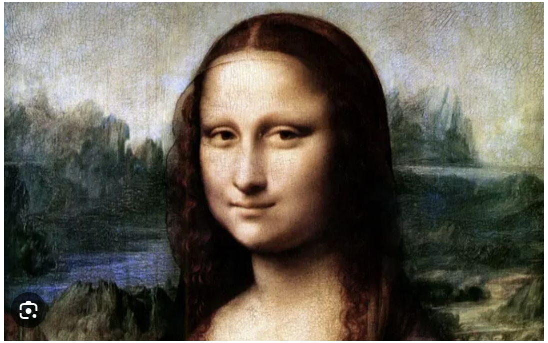
Binary, gray-scale, and color images



0	0	0
1	0	0
1	1	0

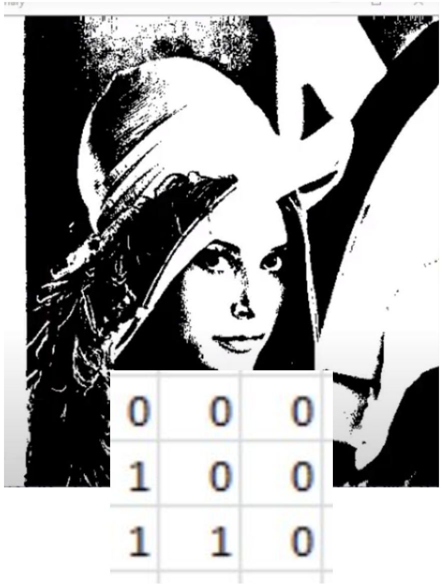


129	115	121
115	109	120
136	162	173

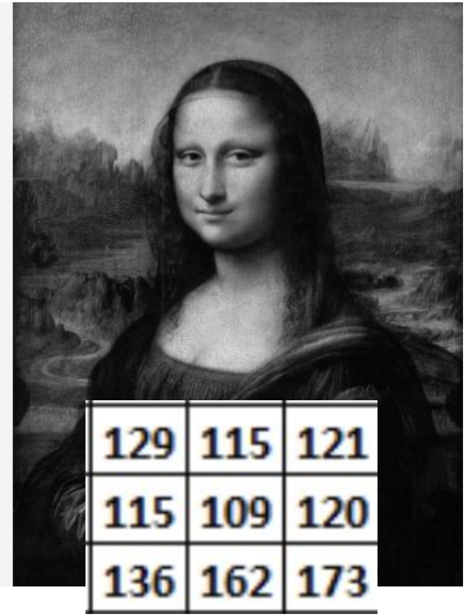


Value range: [0 - 255]

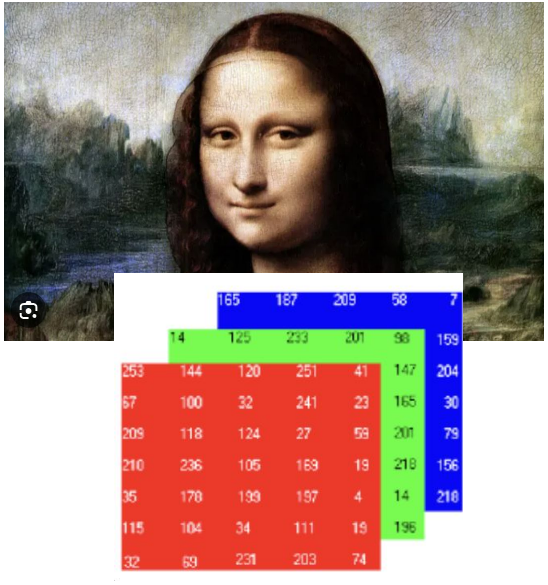
Binary, gray-scale, and color images



0	0	0
1	0	0
1	1	0



129	115	121
115	109	120
136	162	173



165	187	209	58	7
				159
				204
				30
				79
				156
				218

14	125	233	201	98
				147
				165
				201
				218
				14
				196

253	144	120	251	41
67	100	32	241	23
209	118	124	27	59
210	236	105	169	19
35	178	199	197	4
115	104	34	111	19
32	69	231	203	74

Question

- What will be the **vector** size of a 40x50 **RGB** color image?

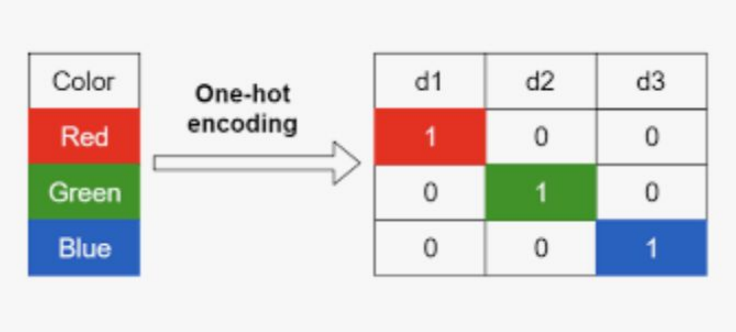
Question

- What will be the **vector** size of a 40x50 **RGB** color image?
- Answer: **6000**

Data Encoding - Categorical Data

One hot encoding

- Only one bit is 1
- A vector representation of categorical values



Color
Red
Green
Blue

One-hot encoding

d1	d2	d3
1	0	0
0	1	0
0	0	1

One hot encoding (cont.)

Classification task:

- Binary example {Cat vs Dog}
- Set size is 2
  - Cat (0, 1)
  - Dog (1, 0)
  - Or vice versa
- Same rule applies every categorical data

Numpy

[Let’s practice]